THE JOURNAL OF PHYSICAL CHEMISTRY LETTERS
A JOURNAL OF THE AMERICAN CHEMICAL SOCIETY

pubs.acs.org/JPCL

Perspective

Structural Feature Extraction via Topological Data Analysis

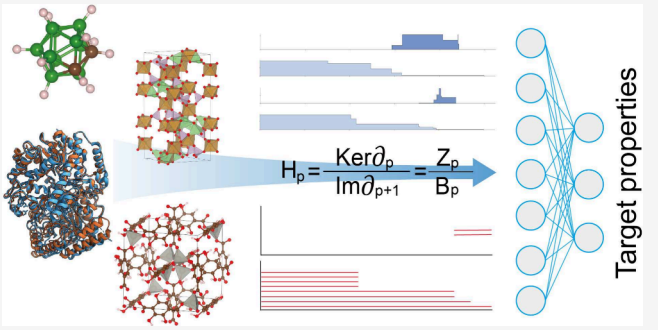
Bingxu Wang, Bin Feng, Linpeng Lv, Shunning Li,* and Feng Pan*

Cite This: *J. Phys. Chem. Lett.* 2025, 16, 8056−8067

Read Online

ACCESS | Metrics & More | Article Recommendations

ABSTRACT: With the advancement of artificial intelligence models, the development of scientifically grounded and structurally appropriate feature extraction methods has become critical for machine learning-based structure prediction and materials design. In recent years, there has been growing dissatisfaction with inefficient empirical descriptors and black-box feature extraction processes that require extensive training. This article introduces a topological data analysis-based framework for extracting structural features of materials, offering an informative perspective on structure−property relationships and predictive strategies. Emphasis is placed on the predictive power and interpretability of topological features, highlighting their advantages in uncovering structure−property correlations and providing physical insights into material behavior. This approach establishes a mathematically rigorous and computationally efficient paradigm for the discovery and design of advanced materials, achieving up to 55% reduction in prediction error for defect-sensitive properties and a notable improvement in MOF gas uptake prediction accuracy (e.g., R^2 from 0.74 to 0.85), thus demonstrating both theoretical clarity and practical performance.

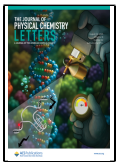


In the realm of physical chemistry, the primary objective of materials investigation lies in understanding their microscopic structural characteristics, which dictate their physical and chemical properties in various applications.[1−3] Traditionally, researchers have relied on experimental techniques and theoretical calculations to acquire the structural information. With the rapid progress of computer science, especially the growing influence of artificial intelligence and machine learning technologies, the digital representation of materials structures has become one of the central topics in materials science.[4,5] The extraction of structural features is not only fundamental for revealing the intrinsic nature of materials, but also plays a critical role in enabling tasks such as property prediction, inverse design, and the discovery of novel materials.[6] In recent years, numerous machine learning models have been constructed, and their performance is closely linked to feature selection.[7,8] As a result, developing structural descriptors that are representative, discriminative, and physically interpretable is essential for the improvement in predictive performance, which could help uncover the hidden relationships between structures and properties, and ultimately accelerate the discovery and design of advanced materials.

Current approaches for structural feature representation in materials science can be categorized into two groups. The first involves handcrafted, empirically derived descriptors such as coordination numbers.[9−11] The second leverages deep learning techniques, with graph neural networks (GNNs) being a prominent example, where crystal structures are encoded as graphs and learned representations are integrated into end-to-end predictive models.[12−14] While empirical descriptors are intuitive and physically interpretable, they often suffer from incomplete representation of structural information and are highly dependent on subjective criteria, which may introduce bias or overlook critical features. In contrast, GNN-based models automatically extract high-dimensional features optimized for specific tasks, and typically achieve superior performance in predicting materials properties. However, these data-driven representations lack physical interpretability, making it difficult to uncover the underlying structure−property relationships.[15] Moreover, the computational cost associated with the training of these models is substantial, which limits their scaling to large materials data sets. The black-box nature of these models also poses challenges for downstream analysis and experimental guidance, thereby constraining their application in the broader context of materials design.

While empirical descriptors are often limited by their fixed functional forms and deep learning-based models may suffer from interpretability and data efficiency issues, hybrid approaches have recently emerged. Notably, the DeePMD

Received: June 15, 2025
Revised: July 28, 2025
Accepted: July 28, 2025
Published: July 31, 2025



https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067

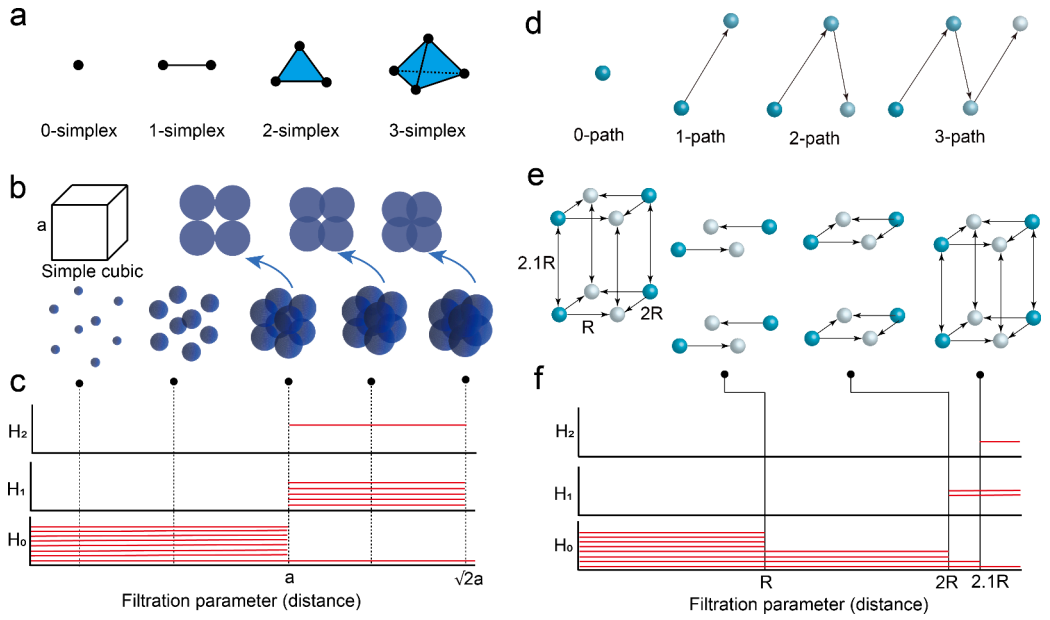


Figure 1. Schematic illustration of topological data analysis for structural feature extraction. (a) Basic elements constituting a simplicial complex across different dimensions. (b) Graph representation of a cubic structure constructed from eight vertices. (c) Evolution of Betti numbers β_0, β_1, and β_2 for the cubic system under increasing filtration values. (d) Fundamental components of a path complex, including vertices, directed edges, and higher-dimensional directed paths. (e) Graph representation of a cubic system with distinct vertex weights and directionally encoded connections. (f) Variation in β_0, β_1, and β_2 for the directed cubic system under different filtration thresholds, capturing structural features associated with weighted connectivity and directional topology.

framework introduces a method known as Deep Potential,[16] which constructs physically inspired intermediate descriptors using local symmetry functions based on neighbor lists. These are then passed through a deep neural network to obtain system-specific representations. This approach alleviates some limitations of handcrafted descriptors while maintaining local chemical interpretability. However, as Deep Potential still relies fundamentally on atom-centered local environments, it does not capture long-range, global, or topological relationships.

Algebraic topology-based data analysis offers a promising solution to the limitations of current structural feature extraction methods in materials science. Unlike empirical descriptors or opaque deep learning representations, this approach extracts structural features from rigorous mathematical theorems, enabling a reasonable and efficient description of materials structures. Building upon the concepts from algebraic topology, these structural features could establish a direct and interpretable link between the atomic structures of materials and the information encoded in the latent space of the deep learning models. In this work, we focus on structural feature extraction through topological data analysis (TDA). We begin with an introductory overview of the mathematical foundation in this method, with discussions kept at a level comprehensible to a broad audience. We then provide some representative examples in the application of TDA-based structural representation in both crystalline solids and molecular systems. Particular emphasis is placed on the predictive power and interpretability of these topological features, highlighting their advantages in uncovering the structure−property relationships and providing physical insights into how geometric arrangements of atoms could influence the thermodynamic and kinetic properties of different materials. Finally, we summarize the key challenges and emerging opportunities associated with the integration of algebraic topology into structural data analysis, pointing toward a mathematically grounded and computationally efficient paradigm for materials discovery and design.

TDA for structural feature extraction is centered on identifying topological invariants that characterize the intrinsic connectivity and shape of a structure, regardless of geometric deformations such as stretching or bending.[17] These invariants serve as robust descriptors of the topology of materials structures and are crucial for representing essential features such as connectivity, loops, and voids, which are closely linked to material properties. To rigorously define and compute these invariants, TDA relies on algebraic topology, particularly the mathematical framework of homology theory.[18] Within this framework, homology groups are introduced to systematically identify and classify "holes" or missing structures across different dimensions, thereby capturing the most fundamental topological characteristics that remain invariant under continuous transformations. Homology groups provide an algebraic representation of the topological structure by encoding the number and types of cycles in each dimension. The ranks of these homology groups, referred to as Betti numbers, offer a concise summary of the topological structures. Specifically, β_0, β_1, and β_2 quantify the number of connected components, 1-dimensional critical structural features, and 2-dimensional critical structural features, respectively.[19] To retain geometric information while extracting these invariants, TDA employs a filtration process, in which a scale or threshold parameter varies to generate a sequence of nested complexes.[20] By calculating the topological invariants across different filtration values, TDA reveals how structural features persist or disappear over scales. This multiscale analysis enables a more nuanced and informative representation of materials structures, particularly in distinguishing structures with similar global shapes but different local geometries. The process of extracting topological invariants is illustrated below using two of the simplest types of complexes as representative examples.

https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067

When a structure is modeled as a simplicial complex for topological analysis, each element of the structure corresponds to a simplex. As shown in Figure 1a, simplices of different dimensions represent the simplest topological units: a 0-dimensional simplex is a vertex, a 1-dimensional simplex is an edge, a 2-dimensional simplex is a triangle, and a 3-dimensional simplex is a tetrahedron. Each represents the most basic configuration for its respective dimension. We use a cube as the representative example to illustrate the process of computing the topological invariants. In Figure 1b, a filtration is applied by using pairwise distances from a reference vertex as the filtration parameter. As the parameter increases, more edges are formed among vertices, progressively constructing higher-order simplices and generating topological features. Figure 1c depicts the changes in the ranks of the homology groups H_0, H_1, and H_2, which correspond to the Betti numbers β_0, β_1, and β_2, respectively. The value of β_0, representing the number of connected components, decreases as connections form, eventually reaching 1 when the structure becomes fully connected. The value of β_1, representing 1-dimensional holes, reaches 6 as the cube configuration forms closed loops. The value of β_2, representing 2-dimensional voids, becomes 1 when an enclosed cavity is formed. These topological invariants across filtration scales serve as concise and informative descriptors of the structural geometry and connectivity of the cubic configuration. The above process is referred to as feature extraction based on persistent homology.[21,22]

When a structure is modeled as a simplicial complex for topological analysis, all vertices are treated as indistinguishable points, which leads to a loss of information intrinsic to the vertices themselves. This abstraction may be insufficient for accurately capturing the structural characteristics of certain materials systems. To address this limitation, materials structures can instead be represented as a path complex, in which each component element is a path. As illustrated in Figure 1d, paths of different dimensions are defined in a topological sense: a 0-dimensional path corresponds to a vertex, a 1-dimensional path is a directed edge whose orientation is determined based on its underlying physico-chemical meaning, a 2-dimensional path consists of two connected directed edges, and a 3-dimensional path consists of three sequentially connected directed edges. To demonstrate this representation, a cube with eight vertices that are assigned with distinct weights is used as an example. For any two vertices with the same weight, a bidirectional connection is defined, while vertices with different weights are connected by a directed edge from the lower to the higher weight. Based on this representation, topological invariants of the path complex can be computed. As shown in Figure 1e, the filtration is constructed using pairwise distances as the filtration parameter. As the filtration value increases, more directed connections (paths) are formed, giving rise to new topological features. Figure 1f shows the evolution of the ranks of the homology groups H_0, H_1, and H_2, corresponding to the Betti numbers β_0, β_1, and β_2, respectively. Here, β_0 indicates the number of connected components, which reaches 1 when the structure is fully connected. β_1 represents the number of directed cycles, which reaches 2 when two loops are formed within the structure. β_2 denotes the number of higher-order directed cavities, with $\beta_2 = 1$ indicating the formation of an enclosed directional cavity. These topological invariants computed at different filtration scales provide informative and robust descriptors for characterizing the structure of the cube with weighted vertices. The above process is referred to as feature extraction based on GLMY homology.[23,24]

In addition, for systems with varying structural complexities, advanced topological methods such as persistent hypergraph homology and persistent directed hypergraph homology can be introduced to extract the topological invariants from different perspectives.[25,26] These approaches enable the characterization of higher-order and directed interactions within a structure. However, increasing the complexity of the topological representation does not necessarily lead to superior predictive power of the models. The effectiveness of topological descriptors depends on the specific characteristics of the system under investigation. Therefore, careful system-specific analysis is essential to determine the appropriate representation. Moreover, constructing multiscale topological features for a single system can provide a richer and more nuanced understanding of its structure. The following examples highlight successful instances of such feature constructions in physically meaningful systems.

Several TDA techniques are employed in this work to extract structural features of materials. Each method is grounded in algebraic topology and tailored for applications in materials modeling. Persistent homology quantifies topological features such as connected components, loops, and voids across multiple spatial scales. It builds a filtration over the structure by progressively adding geometric elements, tracks the appearance and disappearance of topological features, and summarizes their persistence as a concise signature. This approach captures essential shape information with robustness to noise and geometric deformation. GLMY homology, named after Grigor'yan, Lin, Muranov, and Yau, extends classical homology to directed systems by defining topological elements as directed paths rather than simplices. It is particularly effective for analyzing structures with directional relationships, such as molecular graphs or reaction networks. Its persistent formulation allows multiscale tracking of asymmetry and directional connectivity. Hypergraph and directed hypergraph homology generalize topological representations beyond pairwise interactions. In a hypergraph, hyperedges connect multiple nodes simultaneously, making it suitable for modeling many-body interactions in materials and biomolecules. Directed hypergraphs further encode the directionality of interactions, such as donor and acceptor roles in hydrogen bonding, enabling more detailed analysis of structural and functional organization. Together with filtration strategies based on geometric or chemical parameters, these techniques provide interpretable and multiscale structural representations that support predictive modeling and structure−property analysis in materials science.

Persistent homology was initially introduced for the exploration of molecules, with one of the earliest successful examples being the prediction of the formation energy and stability of fullerene molecules.[29] It was observed that the heat of formation is related to the presence of local hexagonal cavities in small fullerenes, while the total curvature energies of fullerene isomers are associated with their sphericities, which are quantified by the lengths of long-persisting β_2 bars. Following this pioneering study, persistent homology has been widely applied in molecular feature extraction.[30,31] In crystalline materials, persistent homology enables systematic sampling across diverse coordination environments and morphologies, allowing for automatic and efficient identification of active phases.[32] This approach addresses the challenge

https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067

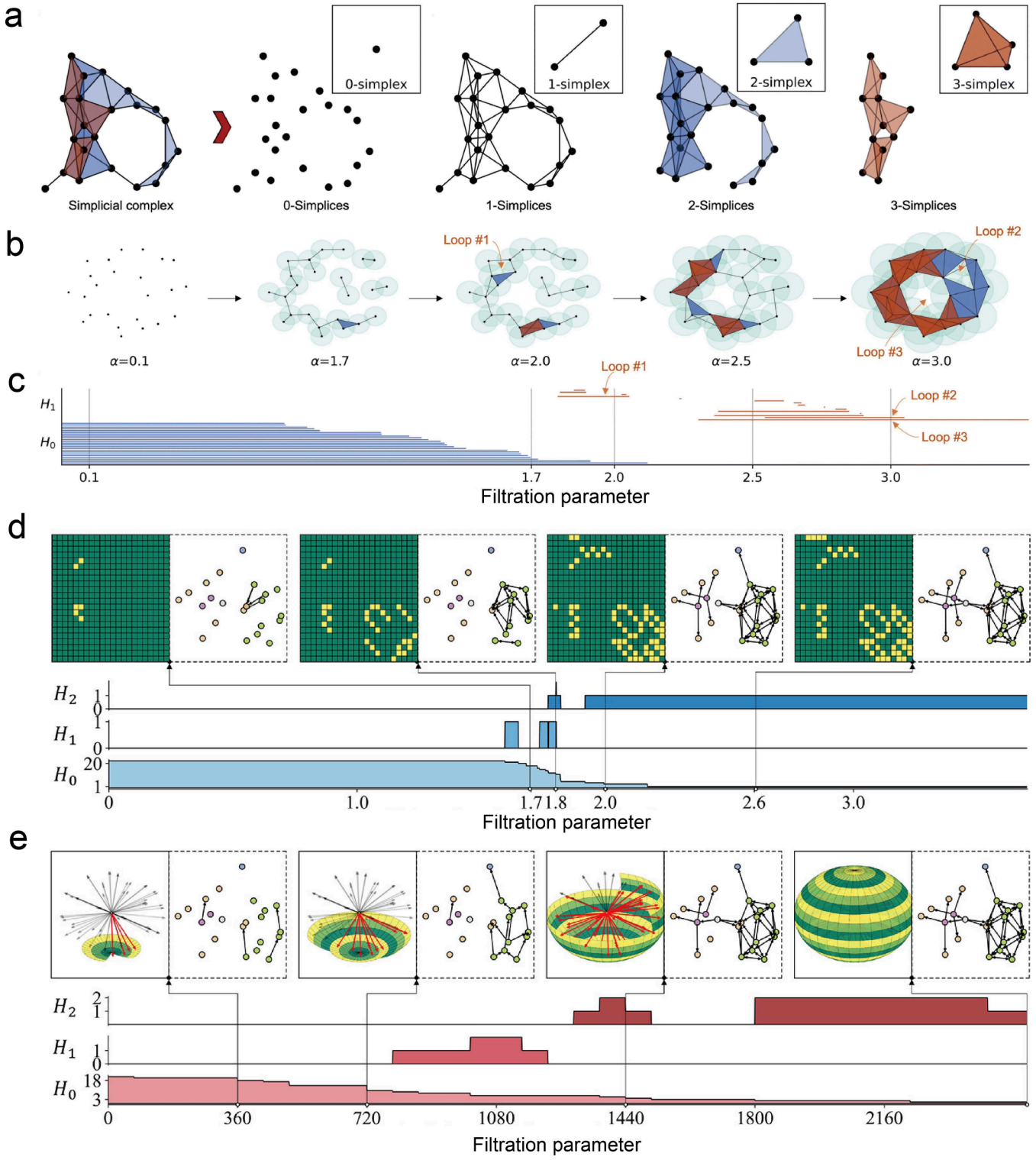


Figure 2. Schematic illustration of feature extraction in real physical space. (a) Representation of simplices within a MOF structure, highlighting the correspondence between molecular geometry and topological elements.[27] Copyright 2025 Royal Society of Chemistry. Licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License. (b) Depiction of a complex porous structure and the associated loops captured by the corresponding simplicial complex.[27] Copyright 2025 Royal Society of Chemistry. Licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License. (c) Persistent homology-based feature extraction process, illustrating the correspondence between topological invariants and structural characteristics.[27] Copyright 2025 Royal Society of Chemistry. Licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License. (d) Topological features extracted using interatomic distance as the filtration parameter.[28] Copyright 2023 American Chemical Society. (e) Topological features extracted using bond angle as the filtration parameter.[28] Copyright 2023 American Chemical Society.

https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067

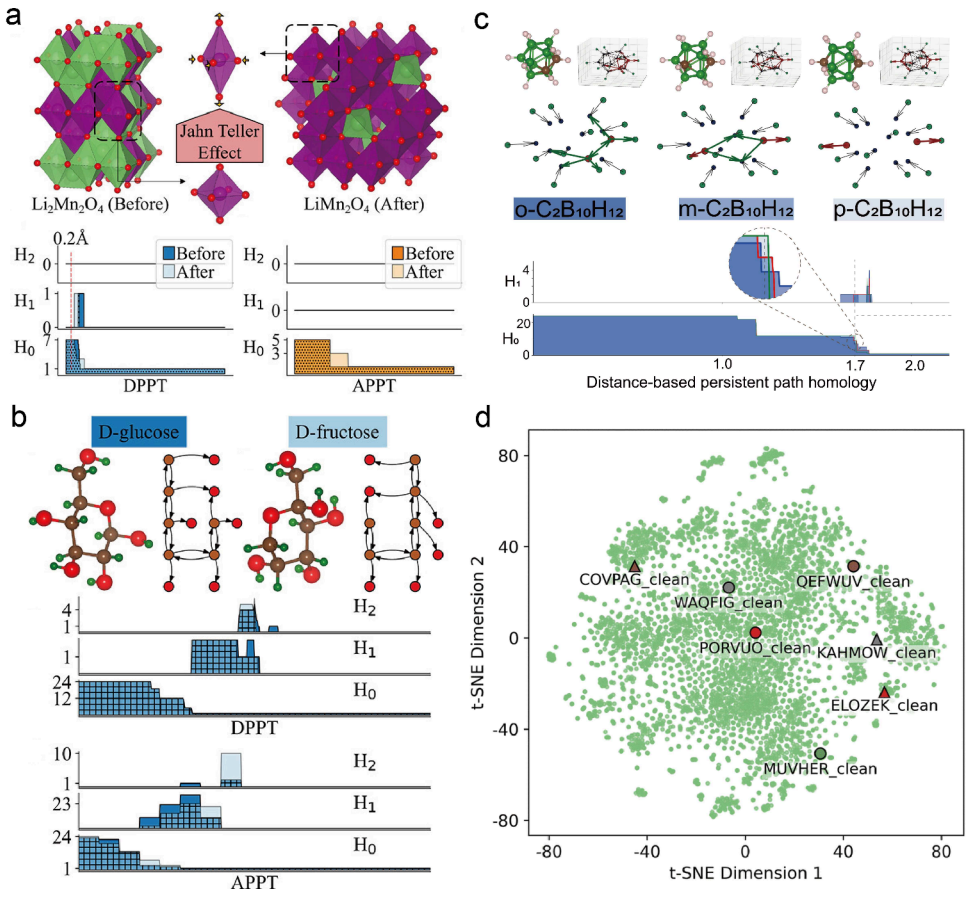


Figure 3. Schematic illustration of structure−property relationships revealed by topological features. (a) Illustration of the Jahn−Teller effect, demonstrating how electronic configurations influence molecular geometries and their Angle-based Persistent Path Topology (APPT) and Distance-based Persistent Path Topology (DPPT).[28] Copyright 2023 American Chemical Society. (b) Molecular structure of D-glucose (left) and its associated directed graph representation (right), highlighting directional atomic interactions.[28] Copyright 2023 American Chemical Society. (c) The structural diagrams and the digraphs of o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$, and p-$C_2B_{10}H_{12}$ and their topological features.[35] Copyright 2024 World Scientific Publishing Company. (d) t-SNE dimensionality reduction of category-specific topological features extracted from MOF materials. Each green point represents a distinct MOF, while highlighted circles and triangles mark materials with maximum and minimum values, respectively, for four key properties: N_2 uptake (mol kg^{-1}), O_2 uptake (mol kg^{-1}), self-diffusivity of N_2 at infinite dilution (cm^2 s^{-1}), and self-diffusivity of O_2 at infinite dilution (cm^2 s^{-1}).[27] Copyright 2025 Royal Society of Chemistry. Licensed under a Creative Commons Attribution-NonCommercial 3.0 Unported License.

of generating and evaluating large numbers of atomic configurations in heterogeneous catalysis. While traditional persistent homology does not differentiate between atomic species, multiscale topological learning frameworks can resolve this limitation by separating the structure into chemically meaningful subsystems. For instance, lithium-only and lithium-free substructures can be analyzed separately, and topological features such as cycle density and minimum connectivity distance can be used to ensure both structural integrity and ion transport compatibility.[33] This strategy has proven effective in accelerating the discovery of fast lithium-ion conductors. For hybrid systems such as metal−organic frameworks (MOFs), which combine both crystalline and molecular characteristics, persistent homology has been demonstrated to be a successful approach.[27] As shown in Figure 2a, the simplices representing a porous structure are constructed. Figure 2b visualizes the emergence and disappearance of loops under varying filtration thresholds. Figure 2c highlights the corresponding features extracted from this structure, demonstrating the validity of this method in encoding the structures of MOFs. Furthermore, in more complex applications such as protein−ligand binding prediction, persistent homology has also exhibited excellent predictive power.[34]

When analyzing systems composed of multiple atomic species, persistent GLMY homology demonstrates improved effectiveness in feature extraction.[35] As illustrated in Figure 2d, interaction patterns among atoms in a molecule evolve as the filtration distance increases, and the corresponding adjacency matrices clearly reveal that such interactions are inherently asymmetric and nonequilibrium. This asymmetry enhances the sensitivity of feature extraction in systems with heterogeneous atomic types.[28] The construction of filtration parameters can be versatile. Typically, filtration values based on relative distances between nodes within the system are employed, as they are particularly effective in capturing intrinsic structural features. Nevertheless, in certain contexts, it is advantageous to define a fixed external reference frame to extract more interpretable macroscopic features. Figure 2e presents an example in which a spherical coordinate system is defined with the South Pole as the origin and angular values serve as the filtration parameter.[28] In this setting, structural interactions are incrementally included with increasing angles, allowing for more effective differentiation between chiral molecules.

https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067

Similarly, in the context of heterogeneous catalysis, where both ligand and coordination effects must be considered to identify potential catalytic configurations, persistent GLMY homology offers notable advantages in encoding spatial configurations and node weights. In the design of high-entropy alloy catalysts, conventional feature extraction algorithms often fail to distinguish the subtle influences of chemically similar metal species. By incorporating atomic weights, persistent GLMY homology enables refined topological partitioning of samples across the complex and extensive chemical space of high-entropy alloys, thereby providing a robust feature foundation for subsequent catalytic property prediction and structure generation.

For systems featured in more complex topological relationships, persistent hypergraph homology offers distinct advantages. In the structural representation of proteins, hypergraphs demonstrate a high capability in modeling higher-order and multivalent interactions, which contrasts with conventional graph-based methods that are limited to capturing pairwise residue interactions.[36,37] Hypergraphs allow a single hyperedge to connect multiple nodes simultaneously, making it possible to naturally represent cooperative multiresidue features commonly observed in proteins, such as hydrophobic cores, hydrogen-bonding networks, and structural motifs.[38] These higher-order associations play critical roles in stabilizing global conformations and facilitating biological function during processes such as protein folding, domain organization, and molecular recognition.[39] By employing hypergraph-based representations, it becomes feasible to accurately capture the combinations of spatially distant residues that are functionally coupled, thereby improving the ability to model nonlocal dependencies within protein structures. Furthermore, recent studies have demonstrated that directed hypergraphs exhibit superior performance in protein−ligand recognition tasks, primarily due to their ability to finely encode the interaction directionality and multibody causal dependencies. Unlike undirected hypergraphs, directed hypergraphs explicitly represent source and target nodes of molecular interactions, such as donor and acceptor roles in hydrogen bonding or the ligand-induced conformational responses in proteins.[40,41] This allows for more precise modeling of authentic binding mechanisms. Additionally, directed hypergraphs can represent synergistic interaction pathways involving multiple atoms or residues, thereby improving the accuracy of functional site identification and deepening the understanding of conformation-dependent regulation.

Topological data analysis demonstrates high sensitivity to structure−property relationships.[30,35,42−44] As illustrated in Figure 3a, the $Li_2Mn_2O_4/LiMn_2O_4$ cathode system exhibits a Jahn−Teller distortion, where the oxidation of Mn to a +3 oxidation state during lithiation induces elongation of the MnO_6 octahedra along the *z*-axis. Persistent GLMY homology with interatomic distance as the filtration parameter captures the topological changes caused by this distortion, with noticeable divergence in H_0 and H_1 observed beyond a filtration threshold of 0.2 Å. Further analysis using angle-based filtration reveals that the spherical filter initially omits parts of the directed edges, leading to a reduced number of connected components in H_0 compared to the number of atoms. Simultaneously, variations in the positions of oxygen atoms result in the formation of additional directed cycles (H_1) and cavities (H_2) throughout the filtration process, reflecting pronounced topological changes induced by structural distortion. Figure 3b presents the molecular structures of D-fructose and D-glucose, which share the same molecular formula ($C_6H_{12}O_6$) yet differ in spatial configuration, resulting in distinct optical activity and biological functionality. D-glucose exhibits physiological activity through enzymatic recognition in the human body, whereas d-fructose does not. This functional divergence arises from geometric differences that are captured in their topological representations of the molecules. By incorporating directional information into the molecular structures, both distance-based and angle-based persistent GLMY homology effectively extract topological features that differentiate the stereoisomers, providing insight into structure−activity relationships.[28] o-$C_2B_{10}H_{12}$, m-$C_2B_{10}H_{12}$ and p-$C_2B_{10}H_{12}$ are three isomers with distinct carbon atom arrangements, which significantly influence their electron distributions, thermodynamic stabilities, and luminescent properties. Among them, the p-type is the most thermodynamically stable, while the o-type is the least, due to the varying spatial separations of carbon atoms and the presence or absence of intervening boron atoms. As shown in Figure 3c, the topological fingerprints extracted via persistent GLMY homology capture the intrinsic structural differences among the three isomers. At a filtration parameter of 1.7, the o-type exhibits the lowest β_0 due to denser directed connections between carbon atoms, while the p-type shows the highest β_0 as its carbon atoms are disconnected from boron. Furthermore, both o- and m-types form directed cycles, resulting in $\beta_1 = 1$, whereas the p-type lacks such cycles, yielding $\beta_1 = 0$. These topological invariants are strongly correlated with the relative positions of carbon atoms, serving as a powerful tool for multielement cluster property prediction.

Topological features can effectively reveal structure−property relationships, even for complex systems.[45−47] A study on SARS-CoV-2 combined 3-dimensional structural data with mutation information to establish a topological model of the Mpro protein structure−activity relationship.[40] The research began by calculating the minimum atomic distance between Mpro residues and the drug nirmatrelvir based on their cocrystal structure, defining their topological proximity within the structure. The mutation frequency of residues over time was then correlated with their distance to the drug, revealing that residues located near the drug-binding site (within 15 Å) exhibited a significant increase in mutation frequency after widespread use of PAXLOVID, forming potential hotspots for drug resistance. This method, combining spatial topological information with dynamic mutation trends, uncovered the evolutionary pressure exerted on target-adjacent regions by drug use, thus providing a structure−mutation−function framework that offers a topological perspective for predicting and monitoring drug resistance risks. Figure 3d presents a 2D t-SNE dimensionality reduction, where each green point represents a different MOF material, with clustering reflecting the influence of topological features.[27] Key properties such as N_2 uptake, O_2 uptake, and self-diffusivity are mapped, with materials marked with maximum and minimum values. Even without predictive modeling, topological features effectively distinguish structures with significant performance differences, indicating that the model itself captures key structure−property relationships. For example, the MOF material labeled ELOZEK_clean, with the lowest N_2/O_2 uptake and Henry constant (8.64×10^{-3} mol kg^{-1}), shows poor gas adsorption capacity. Similarly, COVPAG_clean exhibits the lowest N_2 self-diffusivity (4.15 ×

https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067

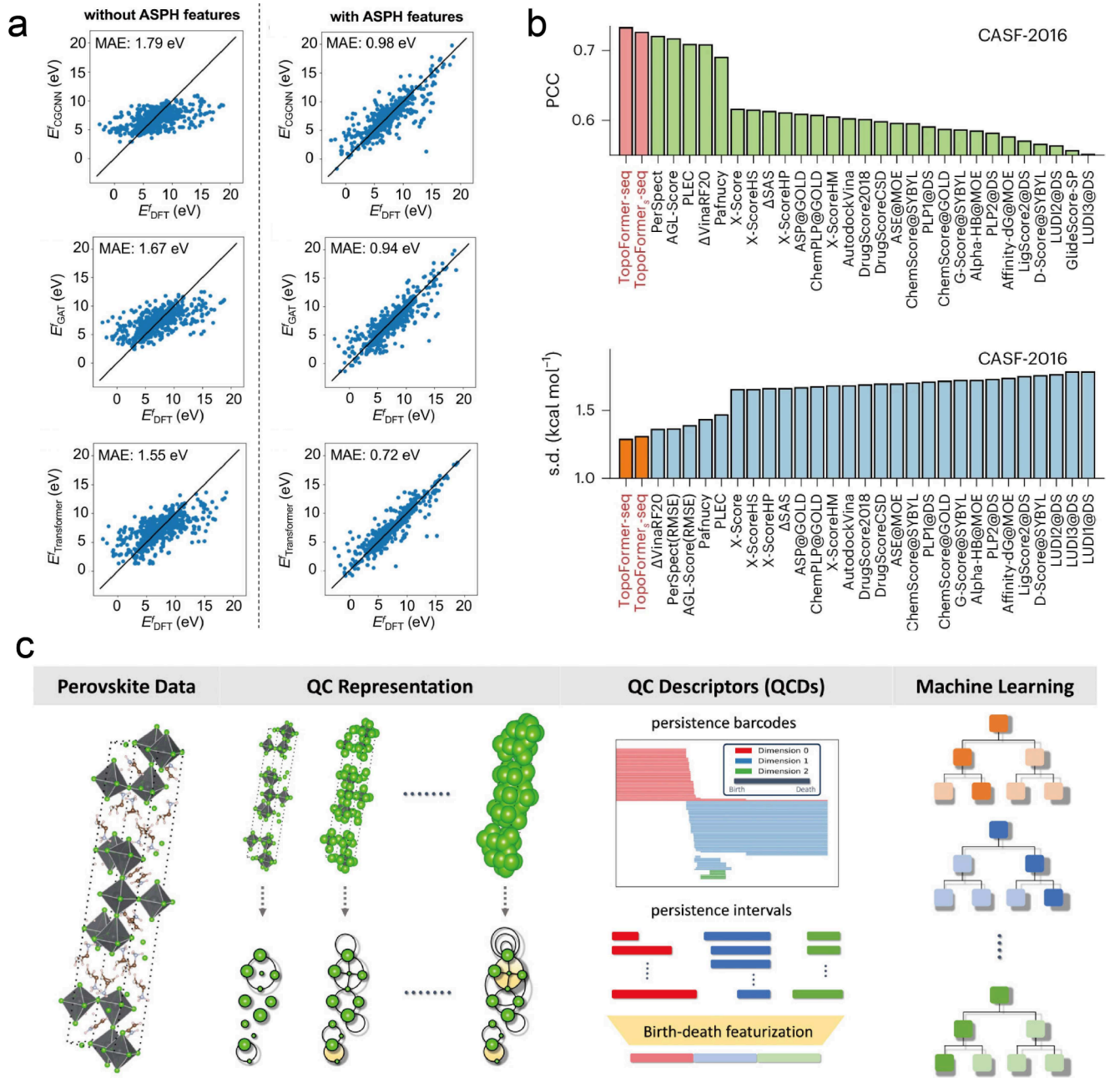


Figure 4. Schematic illustration of the advantages of topological descriptors in property prediction and structural design. (a) Performance comparison of graph neural network (GNN) models with and without Angular Spherical Persistent Homology (ASPH) on an oxygen-based perovskite data set, using a global max pooling layer.[46] Copyright 2025 American Chemical Society. (b) Comparison of Pearson correlation coefficients (PCCs) and standard deviations (s.d.) across various models for protein−ligand binding affinity prediction on the CASF-2016 benchmark data set.[41] Copyright 2024 Springer Nature. (c) Workflow of a machine learning model for band gap prediction in 2-dimensional perovskite structures, utilizing topological descriptors (QC).[48] Copyright 2025 American Chemical Society.

10^{-7} cm^2 s^{-1}), reflecting its limited diffusion ability. These differences highlight the power of the category-specific topological learning method in directly revealing key structural changes through category-specific topological embeddings, efficiently distinguishing materials with extreme performance values in the MOF data set.

Topological features have demonstrated clear advantages over conventional descriptors in property prediction tasks across diverse material and molecular systems.[47,49−51] As shown in Figure 4a, GNN models, including crystal graph neural network, graph attention neural network, and Transformer architectures, were evaluated on a curated data set of oxygen-containing perovskites.[46] The incorporation of topological descriptors led to a significant reduction in mean absolute error (MAE) across all models, indicating a consistent enhancement in the prediction of defect-sensitive properties. Among the tested models, the Transformer network achieved the most substantial improvement, with its MAE decreasing from 1.55 to 0.72 eV, representing a 55% relative reduction. This improvement can be attributed to the attention mechanism in Transformer networks, which encodes interactions between central and neighboring nodes using query-key similarity. Such similarity is informed not only by elemental properties but also by chemical context captured

https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067

Table 1. Comparison of Topological Feature Extraction and Conventional Structural Descriptors

Method	Applicable Systems	Captures Complex Structures	Structural Sensitivity	Interpretability	Scientific Rigor and Efficiency
Topological Data Analysis	Crystals, MOFs, biomolecules	Excellent for both global and hierarchical topology	Highly sensitive to defects and geometric distortions	Strong (mathematically grounded and interpretable)	Very efficient; rigorous and nonparametric
SOAP[52]	Molecules, solids	Good for local symmetry; limited global expressiveness	Moderate; focused on local atom overlaps	Moderate; local environment visualizable	Efficient and popular; less expressive in complex cases
Coordination Number[53]	Solids, MOFs, catalysts	Simple local structure only	Sensitive to coordination but misses long-range order	High; intuitive for basic chemistry	Very efficient; lacks depth for advanced systems
GNNs[13]	Molecules, proteins, crystals	Good multiscale potential via message passing	Good, but training dependent	Medium; partial via attention/gradient maps	Expensive; often black-box and data-dependent
Atomic Environment Descriptors[54]	Crystals, surfaces	Focused on local geometry	Local sensitivity only	Moderate; interpretable via geometry	Fast; limited structural depth
MD-derived Features[55]	Liquids, biomolecules	Captures dynamics, but weak on topology	Sensitive to atomic motion, not shape	Low; dynamic stats are indirect	Computationally costly; weak in structure property mapping

through topological features, such as defect counts and spatial proximities, thereby enhancing the ability to extract defect-related information.

The effectiveness of topology-enriched Transformer models extends to biomolecular domains. As illustrated in Figure 4b, TopoFormer-Seq consistently outperformed existing approaches in protein−ligand binding affinity prediction on the PDBbind v.2016 data set, achieving a Pearson correlation coefficient (PCC) of 0.866 and a root-mean-square error (RMSE) of 1.561 kcal/mol, surpassing the previous best-performing model, TopBP22. Compared to other recent deep learning frameworks such as graphDelta, ECIF, and DeepAtom, TopoFormer demonstrated superior predictive stability and accuracy, regardless of training set variations.[41]

In periodic crystalline systems, a topological learning model based on the Quotient Complex (QC) has been developed for accurate prediction of band gaps in 2-dimensional perovskite materials.[48] As illustrated in Figure 4c, the modeling framework consists of three primary stages. First, atomic types such as organic atoms (e.g., C), inorganic metal atoms (e.g., Sn), and halogen atoms (e.g., Cl), along with their associated configurational sites (including organic sites A, inorganic sites B, halogen sites X, and composite types such as ACBX), are systematically extracted from the crystal unit cell. Second, a multiscale filtration process is constructed using atomic radii as the filtration parameters, enabling the generation of a sequence of quotient complexes that encode periodicity and higher-order structural relationships. From this filtration, topological feature vectors, referred to as Quotient Complex Descriptors (QCDs), are derived to represent the intrinsic topological properties of the structure. Third, these descriptors are integrated with a gradient boosting tree (GBT) model to achieve highly accurate band gap predictions. Notably, the quotient complex framework offers a distinct advantage over conventional material representations by naturally encoding periodic structure and capturing local topological interactions that are critical for structure−property relationships.

Similar improvements in model performance were demonstrated in property prediction tasks for MOFs, where topologically informed model outperformed state-of-the-art baselines, including descriptor-based models, MOFTransformer, and PMTransformer.[27] Specifically, in tasks involving Henry's constants for N_2 and O_2, gas uptake, and self-diffusivity under both 1 bar and infinite dilution conditions, the proposed model consistently achieved higher coefficients of determination (R^2) and lower MAE and RMSE values. For instance, in O_2 uptake prediction, topologically informed model attained an R^2 of 0.85 and an RMSE of 6.82×10^{-2}, outperforming MOFTransformer's R^2 of 0.74 and RMSE of 9.28×10^{-2}. Similarly, in predicting the self-diffusivity of O_2 at 1 bar, the proposed model achieved an MAE of 3.21×10^{-5} and an RMSE of 4.45×10^{-5}, exceeding the accuracy of PMTransformer and other baseline methods. These results underscore the efficacy of category-specific topological features in capturing structure−property relationships within MOFs, thereby enabling models with improved generalizability and predictive performance.

Despite the promising successes of topological data analysis in capturing multiscale structural features, it is important to acknowledge its limitations and potential failure scenarios. First, while persistent homology is sensitive to geometric and topological variations, its expressiveness is not uniformly optimal across all structure−property correlation tasks. In systems where properties depend on very subtle electronic effects, such as fine energy-level splitting or orbital hybridization, purely topological features may lack the resolution required to capture such distinctions. Additionally, feature instability can arise near topological tipping points, which refer to regions in the filtration space where small geometric perturbations, such as those caused by thermal noise or numerical precision, can result in abrupt changes in the birth or death of topological features. This issue is particularly relevant in molecular dynamics simulations with thermal fluctuations or in low-symmetry crystalline systems, where minor variations may significantly alter the resulting persistence diagrams. Furthermore, for highly disordered or amorphous materials, defining a meaningful filtration can be challenging, often leading to nonreproducible or ambiguous topological features. Empirical studies have observed that TDA performs less effectively in tasks where global geometry is less influential than local chemical environments or specific functional groups. For instance, in the prediction of reaction yields for certain organic molecules or in property modeling of noncrystalline polymers, topological features tend to underperform compared to descriptors that explicitly encode functional groups or electronic properties. This limitation arises because TDA focuses on capturing geometric and relational invariants, but not energetic or quantum-level interactions. As a result, its performance is highly dependent on the extent to which structure governs function in the targeted application. These considerations highlight the

https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067

necessity of integrating topological descriptors with chemically informed or domain-specific features. Rather than serving as a universal replacement, topological data analysis is best applied as a complementary representation that enhances the geometric understanding of structure in cases where spatial organization plays a central role.

Table 1 presents a comparative overview of various feature extraction methods commonly used in materials and molecular modeling, highlighting their applicability, structural sensitivity, interpretability, computational efficiency, and theoretical foundation. As shown in the table, topological data analysis (TDA) stands out across multiple criteria, especially in its ability to capture complex, hierarchical structures and provide strong interpretability grounded in rigorous mathematics. Based on the aforementioned studies, topological data analysis has demonstrated its critical role in elucidating structure−property relationships in materials. Owing to its foundation in rigorous mathematical theory, topological feature extraction offers a more scientifically grounded alternative to empirically driven approaches. In contrast to trainable models such as graph neural networks, topological descriptors can capture higher-dimensional structural features while requiring significantly less computational time, rendering their cost negligible in typical systems. Most importantly, topological features have already achieved superior predictive performance compared to state-of-the-art models across a variety of domains. This indicates a promising new paradigm for computational materials research, where structure-informed, theory-driven representations can enhance both the interpretability and accuracy of property prediction.

As summarized in Table 1, TDA maintains both rigorous mathematical grounding and high computational efficiency compared to conventional descriptors and deep learning-based models. The computational cost of persistent homology, a core TDA method, depends primarily on the number of simplices generated from atomic interactions. For a structure with n atoms, the worst-case time complexity of Vietoris−Rips filtration is $O(n^3)$, but in practical settings involving sparse molecular or crystalline graphs and filtration restricted to low-dimensional homology (e.g., H_0, H_1, H_2), the effective runtime is substantially lower.[56,57] Benchmark studies on biomolecular systems have reported wall-clock runtimes of approximately 0.1−0.3 s per structure using libraries such as Ripser on a single CPU.[42] These results are consistent with typical structure sizes found in porous materials and molecular crystals. In contrast, deep learning approaches such as graph neural networks (GNNs) or Transformer-based models, require significantly higher computational overhead, including GPU acceleration and extensive model training, which may span several hours for large data sets. This contrast highlights the advantage of TDA descriptors in rapid, large-scale screening tasks and supports their integration as lightweight yet expressive representations in hybrid machine learning workflows.

The selection of a suitable TDA representation hinges on three key factors: (1) the nature of the structure (e.g., crystalline, molecular, disordered), (2) the type of physical property being predicted (e.g., electronic, transport, binding), and (3) the presence or absence of directional, periodic, or many-body interactions. Persistent Homology with distance-based filtration is most effective in periodic or rigid systems where connectivity and voids at different length scales matter, such as in porous materials, ionic crystals, or framework structures. GLMY Homology, which encodes directionality, is especially suited for systems with intrinsic asymmetry or nonequilibrium behavior, such as chemical reaction networks, charged molecular graphs, or transition states. When angle-based filtrations are applied, it can capture stereochemistry or field-driven anisotropy (e.g., in chiral molecules or Jahn−Teller-distorted lattices). Directed Hypergraph Homology becomes advantageous in biomolecular or soft-matter systems where cooperative, many-body, and causally directional interactions dominate. Protein−ligand binding, allosteric regulation, or multisite adsorption are typical examples. Furthermore, we emphasize that feature validation remains an empirical process. For most applications, we recommend cross-comparing the predictive relevance of different topological features (e.g., Betti numbers, cycle density, lifetime statistics) using model-agnostic feature importance analysis such as SHAP or permutation tests. By introducing these guidelines, we aim to make TDA more accessible and operational for a broader materials research audience. We agree that the method's generalizability is best realized when its customization is clearly connected to physical intuition and domain-specific modeling goals.

Beyond geometry-based analysis, an important future direction lies in extending topological data analysis to incorporate quantum mechanical information that is essential for ab initio simulations. While current TDA frameworks primarily operate on atomic coordinates or interatomic distances, many properties predicted by quantum chemical methods, such as energy levels, charge distributions, and orbital hybridizations, originate from electronic structure features that are not directly represented by spatial geometry. One promising strategy is to apply persistent homology to continuous scalar fields generated by quantum calculations, including electron density, electrostatic potential, or orbital isosurfaces. These fields can be discretized on a grid and analyzed through sublevel set filtrations in order to extract topological invariants that capture aspects of electronic organization.[58] Nevertheless, this extension introduces several computational challenges. These include increased data dimensionality, strong sensitivity to grid resolution and threshold values, and the difficulty of aligning topological features with physically meaningful quantum observables. Addressing these issues will be crucial for developing topology-informed representations that connect structural geometry with electronic structure, ultimately enabling broader applications of TDA in quantum mechanical modeling.

The superior interpretability of topological features makes them a promising direction for novel materials design.[59] As structure-aware encoders, topological descriptors can systematically encode atomic and geometric configurations into mathematically meaningful representations. These representations offer a scientifically grounded encoding scheme that can serve as robust inputs to downstream models with structure reconstruction capabilities. By integrating topological features into generative frameworks, several persistent challenges in materials design can be effectively addressed, including difficulties in encoding periodic hierarchical structures, capturing long-range correlations, and ensuring chemical validity during inverse design. This approach enables the incorporation of domain knowledge into data-driven pipelines, thereby enhancing both the fidelity and controllability of structure generation in complex materials systems.

https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067

Topological data analysis does not function as an isolated methodology. A promising direction lies in its integration with existing machine learning models, where the structural sensitivity of topological descriptors to physical configurations, along with their compatibility with neural network architectures, offers substantial potential for enhancing predictive accuracy.[60−62] Recent implementations have demonstrated the feasibility of integrating topological descriptors into neural network architectures. For example, the TopoQA framework incorporates persistent homology-based features into GNNs networks for protein interface quality assessment, showcasing a practical pipeline for combining topological and learned structural representations.[63] By embedding topological features into learning pipelines, it becomes possible to enrich the model's understanding of spatial and relational complexities inherent in crystal structures. This integration facilitates a more comprehensive exploration of the chemical space, enabling more informed predictions and the discovery of novel materials with targeted properties.

Topological data analysis shares conceptual elements with traditional structural descriptors, such as interpreting β_0 as connected components analogous to isolated atom clusters, and β_1, β_2 as higher-order interaction motifs. However, TDA provides a fundamentally different perspective by constructing a continuous filtration over the structure, thus capturing multiscale topological features that evolve smoothly with the filtration parameter. Unlike fixed-cutoff descriptors, persistent homology summarizes connectivity patterns and voids across all scales, yielding hierarchical, global invariants that encode complex geometric correlations beyond local pairwise or three-body interactions. Moreover, TDA descriptors are mathematically grounded with stability guarantees and robustness to noise, enabling more reliable representation of structural complexity. This multiscale and topologically global viewpoint allows TDA to implicitly encode an extensive range of interactions, including subtle long-range correlations and higher-dimensional cycles, which are typically not accessible by conventional local descriptors limited to specific neighbor shells or handcrafted functions. Therefore, TDA should be viewed as a complementary and enriched feature space that systematically extends the information content of classical descriptors. This extension underpins the consistent improvement in predictive modeling observed across diverse materials and molecular systems, as shown by our results.

Overall, topological data analysis provides a powerful and underexplored framework for representing and understanding structural information in materials science. In contrast to conventional descriptors and deep learning representations that focus primarily on predictive performance, topology-based methods capture fundamental and scale-invariant features that reflect intrinsic geometric and connectivity patterns. As the field advances toward more interpretable, generalizable, and automated materials discovery, algebraic topology through persistent homology, GLMY homology, and hypergraph-based approaches offers a principled foundation for encoding structural information in a physically meaningful way. Although further work is needed to integrate these descriptors seamlessly into predictive and generative pipelines, the conceptual clarity and representational strength of topological methods make them a compelling direction for the future of materials informatics.

AUTHOR INFORMATION

Corresponding Authors

Shunning Li − *School of Energy and Environment, City University of Hong Kong, Kowloon, Hong Kong 999077, P.R. China*; Email: shunnili@cityu.edu.hk

Feng Pan − *School of Advanced Materials, Peking University, Shenzhen 518055, P.R. China*; orcid.org/0000-0002-8216-1339; Email: panfeng@pkusz.edu.cn

Authors

Bingxu Wang − *School of Advanced Materials, Peking University, Shenzhen 518055, P.R. China*

Bin Feng − *School of Advanced Materials, Peking University, Shenzhen 518055, P.R. China*

Linpeng Lv − *School of Advanced Materials, Peking University, Shenzhen 518055, P.R. China*

Complete contact information is available at: https://pubs.acs.org/10.1021/acs.jpclett.5c01831

Author Contributions

B.W. and B.F. contributed equally to this work.

Notes

The authors declare no competing financial interest.

Biographies

Bingxu Wang received his B.E. degree in 2021 from School of mathematical science, Jiangsu University, China. He is now a Ph.D. student in the School of Advanced Materials, Peking University, Shenzhen Graduate School, China. His research interest focuses on computational design of energy storage materials via topological data analysis, first-principles calculations and machine learning techniques.

Bin Feng received his B.E. degree in 2023 from School of Materials Science and Engineering, South China and University of Technology. He is now a M.E. student in the School of Advanced Materials, Peking University, Shenzhen Graduate School, China. His research interest focuses on characterization technique via combined first-principles calculations and machine learning techniques.

Linpeng Lv received his B.E. degree in 2022 from the School of Mechanical Engineering, Sichuan University, and received his M.E. degree in 2025 from the School of Advanced Materials, Peking University, Shenzhen Graduate School. His research interests focus on structural analysis and design of energy materials by combining first-principles calculations with machine learning techniques.

Shunning Li received his B.E. degree in 2013 and Ph.D. degree in 2018 from the School of Materials Science and Engineering, Tsinghua University, China. After working at the School of Advanced Materials, Peking University, Shenzhen Graduate School, he joined City University of Hong Kong as a research fellow at 2025. His research interest focuses on first-principles and machine learning studies of energy storage materials and heterogeneous catalysts.

Feng Pan is chair professor of Peking University, Vice President of Shenzhen Graduate School of Peking University, Founding Dean of the School of New Materials, Member of the Chinese Chemical Society, and Executive editor of Structural Chemistry. He has long been committed to the development of structural chemistry methodology and its application in the research and development of new materials, as evidenced by his creation of the structural chemistry theory based on graph theory; establishment of the in situ dynamic structure characterization system based on large scientific devices, such as neutron and synchrotron radiation; and exploration and revelation of the material genes and structure−function relationship. His breakthroughs have been made in solving scientific

https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067

problems, such as lithium battery energy storage density, power density, and stability. He has published ~500 SCI articles in well-known journals, such as Nature (2), Nature Energy (1), and Nature Nanotechnology (3), as the corresponding author. He has been awarded with the China Electrochemical Contribution Award and the Battery Technology Award of the American Electrochemical Society.

■ ACKNOWLEDGMENTS

The authors acknowledge financial support from Guangdong Key Laboratory of Design and Calculation of New Energy Materials (No. 2017B030301013), the Basic and Applied Basic Research Foundation of Guangdong Province (2021B1515130002 and 2023A1515011391), and the Major Science and Technology Infrastructure Project of Material Genome Big-science Facilities Platform supported by Municipal Development and Reform Commission of Shenzhen.

■ REFERENCES

(1) van de Walle, A. A complete representation of structure−property relationships in crystals. *Nature materials* **2008**, *7* (6), 455−458.
(2) Schorr, S.; Weidenthaler, C. *Crystallography in Materials Science: From Structure-Property Relationships to Engineering*; Walter de Gruyter GmbH & Co KG: 2021.
(3) Katritzky, A. R.; Fara, D. C. How chemical structure determines physical, chemical, and technological properties: An overview illustrating the potential of quantitative structure- property relationships for fuels science. *Energy Fuels* **2005**, *19* (3), 922−935.
(4) Musil, F.; Grisafi, A.; Bartok, A. P.; Ortner, C.; Csanyi, G.; Ceriotti, M. Physics-inspired structural representations for molecules and materials. *Chem. Rev.* **2021**, *121* (16), 9759−9815.
(5) Reyes, K. G.; Maruyama, B. The machine learning revolution in materials? *MRS Bull.* **2019**, *44* (7), 530−537.
(6) Lee, J.; Park, D.; Lee, M.; Lee, H.; Park, K.; Lee, I.; Ryu, S. Machine learning-based inverse design methods considering data characteristics and design space size in materials design and manufacturing: a review. *Materials Horizons* **2023**, *10* (12), 5436−5456.
(7) Balachandran, P. V.; Xue, D.; Theiler, J.; Hogden, J.; Gubernatis, J. E.; Lookman, T. Importance of feature selection in machine learning and adaptive design for materials. In *Materials discovery and design: By means of data science and optimal learning*; Springer: 2018; pp 59−79.
(8) Fu, Z.; Liu, W.; Huang, C.; Mei, T. A review of performance prediction based on machine learning in materials science. *Nanomaterials* **2022**, *12* (17), 2957.
(9) Usuga, A. F.; Praveen, C. S.; Comas-Vives, A. Local descriptors-based machine learning model refined by cluster analysis for accurately predicting adsorption energies on bimetallic alloys. *Journal of Materials Chemistry A* **2024**, *12* (5), 2708−2721.
(10) Wu, D.; Xi, C.; Dong, C.; Liu, H.; Du, X.-W. Bond-energy-integrated coordination number: An accurate descriptor for transition-metal catalysts. *J. Phys. Chem. C* **2019**, *123* (46), 28248−28254.
(11) Nanba, Y.; Koyama, M. An element-based generalized coordination number for predicting the oxygen binding energy on pt3m (m= co, ni, or cu) alloy nanoparticles. *ACS omega* **2021**, *6* (4), 3218−3226.
(12) Chandrasekaran Selvaraj, S. *Graph neural networks based deep learning for predicting structural and electronic properties*; arXiv preprint arXiv:2411.02331, 18 Dec 2024; accessed 2025-07-15.
(13) Xie, T.; Grossman, J. C. Crystal graph convolutional neural networks for an accurate and interpretable prediction of material properties. *Physical review letters* **2018**, *120* (14), 145301.
(14) Reiser, P.; Neubert, M.; Eberhard, A.; Torresi, L.; Zhou, C.; Shao, C.; Metni, H.; van Hoesel, C.; Schopmans, H.; Sommer, T.; Friederich, P.; et al. Graph neural networks for materials science and chemistry. *Communications Materials* **2022**, *3* (1), 93.
(15) Rao, J.; Zheng, S.; Lu, Y.; Yang, Y. Quantitative evaluation of explainable graph neural networks for molecular property prediction. *Patterns* **2022**, *3* (12), 100628.
(16) Wang, H.; Zhang, L.; Han, J.; E, W.; et al. Deepmd-kit: A deep learning package for many-body potential energy representation and molecular dynamics. *Comput. Phys. Commun.* **2018**, *228*, 178−184.
(17) Hatcher, A. *Algebraic Topology*; Cambridge University Press: 2002.
(18) Hilton, P. J.; Wylie, S. *Homology theory: An introduction to algebraic topology*; CUP Archive: 1967.
(19) Carlsson, G. Topology and data. *Bulletin of the American Mathematical Society* **2009**, *46* (2), 255−308.
(20) Edelsbrunner, H.; Harer, J. *Computational topology: an introduction*; American Mathematical Soc.: 2009.
(21) Edelsbrunner; Letscher; Zomorodian. Zomorodian. Topological persistence and simplification. *Discrete computational geometry* **2002**, *28*, 511−533.
(22) Zomorodian, A.; Carlsson, G. Computing persistent homology. In *Proceedings of the twentieth annual symposium on Computational geometry*, 2004; pp 347−356.
(23) Grigor'yan, A. A.; Lin, Y.; Muranov, Y. V.; Yau, S.-T. Path complexes and their homologies. *Journal of Mathematical Sciences* **2020**, *248*, 564−599.
(24) Chowdhury, S.; Mémoli, F. Persistent path homology of directed networks. In *Proceedings of the Twenty-Ninth Annual ACM-SIAM Symposium on Discrete Algorithms*; SIAM: 2018; pp 1152−1169.
(25) Chen, D.; Liu, J.; Wu, J.; Wei, G.-W. Persistent hyperdigraph homology and persistent hyperdigraph laplacians. *Foundations of data science (Springfield, Mo.)* **2023**, *5* (4), 558.
(26) Suwayyid, F.; Wei, G.-W. Persistent dirac of paths on digraphs and hypergraphs. *Foundations of data science (Springfield, Mo.)* **2024**, *6* (2), 124.
(27) Chen, D.; Chen, C.-L.; Wei, G.-W. Category-specific topological learning of metal−organic frameworks. *Journal of Materials Chemistry A* **2025**, *13* (13), 9292−9303.
(28) Chen, D.; Liu, J.; Wu, J.; Wei, G.-W.; Pan, F.; Yau, S.-T. Path topology in molecular and materials sciences. *journal of physical chemistry letters* **2023**, *14* (4), 954−964.
(29) Xia, K.; Feng, X.; Tong, Y.; Wei, G. W. Persistent homology for the quantitative prediction of fullerene stability. *Journal of computational chemistry* **2015**, *36* (6), 408−422.
(30) Townsend, J.; Micucci, C. P.; Hymel, J. H.; Maroulas, V.; Vogiatzis, K. D. Representation of molecular structures with persistent homology for machine learning applications in chemistry. *Nat. Commun.* **2020**, *11* (1), 3230.
(31) Ehiro, T. Descriptor generation from morgan fingerprint using persistent homology. *SAR and QSAR in Environmental Research* **2024**, *35* (1), 31−51.
(32) Zheng, S.; Zhang, X.-M.; Liu, H.-S.; Liang, G.-H.; Zhang, S.-W.; Zhang, W.; Wang, B.; Yang, J.; Jin, X.; Pan, F.; Li, J.-F.; et al. Active phase discovery in heterogeneous catalysis via topology-guided sampling and machine learning. *Nat. Commun.* **2025**, *16* (1), 2542.
(33) Chen, D.; Wang, B.; Li, S.; Zhang, W.; Yang, K.; Song, Y.; Wei, G.-W.; Pan, F. Superionic ionic conductor discovery via multiscale topological learning. *J. Am. Chem. Soc.* **2025**, *147* (24), 20888−20898.
(34) Cang, Z.; Mu, L.; Wei, G.-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS computational biology* **2018**, *14* (1), No. e1005929.
(35) Wang, B.; Zhang, M.; Pan, F. Persistent path homology for quantitative analysis of carboranes. *Journal of Computational Biophysics and Chemistry* **2025**, *24* (01), 1−11.
(36) Liu, X.; Wang, X.; Wu, J.; Xia, K. Hypergraph-based persistent cohomology (hpc) for molecular representations in drug design. *Briefings in Bioinformatics* **2021**, *22* (5), bbaa411.

https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067

(37) Liu, X.; Feng, H.; Wu, J.; Xia, K. Persistent spectral hypergraph based machine learning (psh-ml) for protein-ligand binding affinity prediction. *Briefings in Bioinformatics* **2021**, *22* (5), bbab127.
(38) Murgas, K. A.; Saucan, E.; Sandhu, R. Hypergraph geometry reflects higher-order dynamics in protein interaction networks. *Sci. Rep.* **2022**, *12* (1), 20879.
(39) Gaudelet, T.; Malod-Dognin, N.; Przulj, N. Higher-order molecular organization as a source of biological function. *Bioinformatics* **2018**, *34* (17), i944−i953.
(40) Chen, D.; Liu, G.; Du, H.; Jones, B.; Wee, J.; Wang, R.; Chen, J.; Shen, J.; Wei, G.-W. Drug resistance predictions based on a directed flag transformer. *Advanced Science* **2025**, No. e02756.
(41) Chen, D.; Liu, J.; Wei, G.-W. Multiscale topology-enabled structure-to-sequence transformer for protein−ligand interaction predictions. *Nature Machine Intelligence* **2024**, *6* (7), 799−810.
(42) Meng, Z.; Anand, D. V.; Lu, Y.; Wu, J.; Xia, K. Weighted persistent homology for biomolecular data analysis. *Sci. Rep.* **2020**, *10* (1), 2079.
(43) Minamitani, E., Nakamura, T.; Obayashi, I.; Mizuno, H. *Persistent homology elucidates hierarchical structures responsible for mechanical properties in covalent amorphous solids*. arXiv preprint arXiv:2407.17707, 31 Mar 2025, accessed 2025-07-15.
(44) Liu, J.; Chen, D.; Pan, F.; Wu, J. Neighborhood path complex for the quantitative analysis of the structure and stability of carboranes. *Journal of Computational Biophysics and Chemistry* **2023**, *22* (04), 503−511.
(45) Krishnapriyan, A. S.; Montoya, J.; Haranczyk, M.; Hummelshøj, J.; Morozov, D. Machine learning with persistent homology and chemical word embeddings improves prediction accuracy and interpretability in metal-organic frameworks. *Sci. Rep.* **2021**, *11* (1), 8888.
(46) Fang, Z.; Yan, Q. Leveraging persistent homology features for accurate defect formation energy predictions via graph neural networks. *Chemistry of materials* **2025**, *37* (4), 1531−1540.
(47) Krishnapriyan, A. S.; Haranczyk, M.; Morozov, D. Topological descriptors help predict guest adsorption in nanoporous materials. *J. Phys. Chem. C* **2020**, *124* (17), 9360−9368.
(48) Hu, C.-S.; Mayengbam, R.; Xia, K.; Sum, T. C. Quotient complex (qc)-based machine learning for 2d hybrid perovskite design. *J. Chem. Inf. Model.* **2025**, *65* (2), 660−671.
(49) Wang, Z.-L.; Ogawa, T.; Adachi, Y. Property predictions for dual-phase steels using persistent homology and machine learning. *Advanced Theory and Simulations* **2020**, *3* (3), 1900227.
(50) Szymanski, N. J.; Smith, A.; Daoutidis, P.; Bartel, C. J. Topological descriptors for the electron density of inorganic solids. *ACS Materials Letters* **2025**, *7*, 2158−2164.
(51) Shen, C.; Zhang, Y.; Han, F.; Xia, K. Molecular topological deep learning for polymer property prediction. arXiv preprint arXiv:2410.04765, 7 Oct 2024, accessed 2025-07-15.
(52) Bartok, A. P.; Kondor, R.; Csanyi, G. On representing chemical environments. *Physical Review B Condensed Matter and Materials Physics* **2013**, *87* (18), 184115.
(53) Pauling, L. The principles determining the structure of complex ionic crystals. *Journal of the American chemical society* **1929**, *51* (4), 1010−1026.
(54) Behler, J.; Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical review letters* **2007**, *98* (14), 146401.
(55) Rupp, M.; Tkatchenko, A.; Muller, K.-R.; von Lilienfeld, O. A. Fast and accurate modeling of molecular atomization energies with machine learning. *Physical review letters* **2012**, *108* (5), 058301.
(56) Otter, N.; Porter, M. A; Tillmann, U.; Grindrod, P.; Harrington, H. A A roadmap for the computation of persistent homology. *EPJ. Data Science* **2017**, *6*, 1−38.
(57) Bauer, U. Ripser: efficient computation of vietoris−rips persistence barcodes. *Journal of Applied and Computational Topology* **2021**, *5* (3), 391−423.
(58) Carlsson, E.; Carlsson, J. *A new construction for sublevel set persistence*. arXiv preprint arXiv:2106.04020, 15 Jun 2021, accessed 2025-07-15.
(59) Wang, B.; Zheng, S.; Wu, J.; Li, J.; Pan, F. Inverse design of catalytic active sites via interpretable topology-based deep generative models. *npj Computational Materials* **2025**, *11*, 147.
(60) Wang, Y.; Liu, X.; Zhang, Y.; Wang, X.; Xia, K. Join persistent homology (jph)-based machine learning for metalloprotein−ligand binding affinity prediction. *J. Chem. Inf. Model.* **2025**, *65* (6), 2785−2793.
(61) Meng, Z.; Xia, K. Persistent spectral−based machine learning (perspect ml) for protein-ligand binding affinity prediction. *Science advances* **2021**, *7* (19), No. eabc5329.
(62) Pun, C. S.; Lee, S. X.; Xia, K. Persistent-homology-based machine learning: a survey and a comparative study. *Artificial Intelligence Review* **2022**, *55* (7), 5169−5213.
(63) Han, B.; Zhang, Y.; Li, L.; Gong, X.; Xia, K. Topoqa: a topological deep learning-based approach for protein complex structure interface quality assessment. *Briefings in Bioinformatics* **2025**, *26* (2), bbaf083.

https://doi.org/10.1021/acs.jpclett.5c01831
J. Phys. Chem. Lett. 2025, 16, 8056−8067