TopoFormer: Multiscale Topology-enabled Structure-to-Sequence Transformer for Protein-Ligand Interaction Predictions

Guo-Wei Wei (weig@msu.edu)
Michigan State University https://orcid.org/0000-0001-8132-5998
Dong Chen
School of Advanced Materials, Peking University, Shenzhen Graduate School https://orcid.org/0000-0001-5397-0447
Jian Liu
Michigan State University

Article

Keywords: Drug design, Topological sequences, Topological Transformer, Multiscale Topology, Hyperdigraph Laplacian.

Posted Date: February 9th, 2024

DOI: https://doi.org/10.21203/rs.3.rs-3640878/v1

Additional Declarations: There is **NO** Competing Interest.

TopoFormer: Multiscale Topology-enabled Structure-to-Sequence Transformer for Protein-Ligand Interaction Predictions

Dong Chen[1], Jian Liu[2,1], and Guo-Wei Wei [†1,3,4]

[1]*Department of Mathematics, Michigan State University, MI, 48824, USA*
[2]*Mathematical Science Research Center, Chongqing University of Technology, Chongqing 400054, China*
[3]*Department of Electrical and Computer Engineering, Michigan State University, MI 48824, USA*
[4]*Department of Biochemistry and Molecular Biology, Michigan State University, MI 48824, USA*

Abstract Pre-trained deep Transformers have had tremendous success in a wide variety of disciplines. However, in computational biology, essentially all Transformers are built upon the biological sequences, which ignores vital stereochemical information and may result in crucial errors in downstream predictions. On the other hand, three-dimensional (3D) molecular structures are incompatible with the sequential architecture of Transformer and natural language processing (NLP) models in general. This work addresses this foundational challenge by a topological Transformer (TopoFormer). TopoFormer is built by integrating NLP and a multiscale topology techniques, the persistent topological hyperdigraph Laplacian (PTHL), which systematically converts intricate 3D protein-ligand complexes at various spatial scales into a NLP-admissible sequence of topological invariants and homotopic shapes. Element-specific PTHLs are further developed to embed crucial physical, chemical, and biological interactions into topological sequences. TopoFormer surges ahead of conventional algorithms and recent deep learning variants and gives rise to exemplary scoring accuracy and superior performance in ranking, docking, and screening tasks in a number of benchmark datasets. The proposed topological sequences can be extracted from all kinds of structural data in data science to facilitate various NLP models, heralding a new era in AI-driven discovery.

Keywords Drug design, Topological sequences, Topological Transformer, Multiscale Topology, Hyperdigraph Laplacian.

[†]Corresponding author: weig@msu.edu

Contents

1 Introduction **3**

2 Results **4**
2.1 Overveiw of TopoFormer for protein-ligand binding analysis 4
2.2 Evaluating TopoFormer on scoring tasks 6
2.3 Evaluating TopoFormer on ranking tasks 8
2.4 Evaluating TopoFormer on docking tasks 9
2.5 Evaluating TopoFormer on screening tasks 11

3 Discussion **13**

4 Methods **15**
4.1 Datasets 15
4.2 Topological sequence embedding 16
4.3 TopoFormer model 22

A Supplementary Information **24**
A.1 Evaluation metrics 24
A.2 Hyperparameter selection and optimization 26
A.3 Topological objects 28
A.4 Vietoris-Rips hyperdigraph and alpha hyperdigraph 29
A.5 Supplementary tables 32
A.6 Supplementary figures 35

1 Introduction

The importance of discovery in modern healthcare cannot be overemphasized, as it profoundly impacts our daily lives. However, traditional methods of drug development are notably labor-intensive, consuming over a decade and costing billions of dollars for a single prescription medicine to reach the market [1]. Historically, this domain has been anchored by traditional methods such as molecular docking [2, 3, 4, 5], free energy perturbation [6], and empirically based modeling [7]. While these techniques have provided insights into drug discovery, they come with their share of limitations. Their predictive abilities also often waver in accuracy and reliability, and the computational intensity of these methods further renders them suboptimal for large-scale or swift screening endeavors. Additionally, they may overlook non-traditional binding sites or novel interaction kinetics, leading to missed therapeutic opportunities or misjudged drug efficacy.

In the evolving landscape of drug design, the deep learning models are becoming an attractive options [8, 9, 10], they have shown great capacity to predict protein structures. It was celebrated for their unmatched capability to unravel intricate patterns and deliver superior predictive outcomes. [11] This shift towards deep learning, built on the successes of chemoinformatics and bioinformatics [12], embodies the modern era's tilt towards data-driven methodologies. However, challenges like the necessity for frequent retraining and an overwhelming reliance on labeled data have been persistent roadblocks.

The groundbreaking Transformer framework and models like ChatGPT, which owe their triumphs to large-scale pre-training and the adept use of unlabeled data, point towards the untapped potential of self-supervised learning [13, 14, 15]. These models offer a glimpse of powerful solutions, especially when traditional labeled data is a limiting factor. While the success of the Transformer framework in the realm of natural language processing is undeniable, its direct application to the domain of drug discovery, especially for the protein-ligand complex modeling, raises pertinent questions because of its neglecting important stereochemical relations. One pivotal quandary is tailoring a model, intrinsically designed for serialized language translations, to suit the study of protein-ligand complexes, which inherently defy serialized representation.

In response to the existing challenges, we leverage advanced mathematical models from algebraic topology, differential geometry, and combinatorial graph theory. These models, previously applied to represent biomolecular systems, have achieved significant successes [16, 17, 18, 19]. Drawing upon unique insights from advanced mathematics, we unveil our topological transformer model: TopoFormer. TopoFormer is built upon persistent topological hyperdigraph Laplacian (PTHL) [20], a tarnsformative algebraic topological model. While intrinsically mirroring foundational topological invariants akin to traditional persistent homology [21], this multiscale technique introduced the novel topological hyperdigraph to capture intrinsic physical, chemical, and biological interactions in protein-ligand binding, and uniquely delivers a non-harmonic spectrum, shedding light on the three-dimensional (3D) shape intricacies of protein-ligand complexes. In a nutshell, PTHL utilizes its multiscale topology and multiscale spectrum to convert intricate 3D protein-ligand complexes into 1D topological sequences that are ideally suitable for the sequential architecture of Transformers (Figure 1). This innovative fusion not only melds topological insights with cutting-edge machine learning but also heralds a paradigm shift in our grasp of protein-ligand relationships. Capitalizing on its deep-rooted topological framework, TopoFormer redefines performance benchmarks in drug research tasks like scoring, ranking, docking, and screening. Its nuanced design

ensures that unconventional interactions are not overlooked but are instead spotlighted. As shown in the results, TopoFormer consistently outshines its peers, achieving state-of-the-art outcomes across diverse benchmark datasets in drug discovery.

2 Results

In this section, an overview of the proposed topological transformer (TopoFormer) model is provided, followed by a comprehensive evaluation of its performance across crucial tasks, including scoring, ranking, docking, and screening. The analysis contextualizes TopoFormer's capabilities within the framework of existing methodologies, thus revealing both the strengths and advantages of this novel model when compared to established techniques.

2.1 Overveiw of TopoFormer for protein-ligand binding analysis

Transformer [13] architecture offered a groundbreaking technique that leverages attention mechanisms to understand sequential data in various domains [14, 22, 23]. Drawing inspiration from the Transformer's design and capabilities, we have conceived a topological transformer model named TopoFormer, as shown in Figure 1. TopoFormer integrates our new persistent topological hyperdigraph Laplacian (PTHL) [20] and transformer for the first time. Unlike other Transformers that are based on protein and ligand sequence information, TopoFormer takes 3D protein-ligand complexes as inputs. This is made possible through the unique transformation of intricate 3D protein-ligand complexes into sequences of topological invariants and homotopic shape and stereochemical evolution by PTHL. The PTHL technique sequentially embeds the topological invariants, the homotopic shape, and the physical, chemical, and biological interactions of 3D protein-ligand complexes at various scales into a topological sequence admissible to the transformer architecture. Pretraining on a diverse set of protein-ligand complexes empowers the model to grasp the broad characteristics and nuances of molecular interactions, including various sterochemical effects that cannot be captured by traditional molecular sequences. Subsequent fine-tuning on specific datasets ensures that the output embeddings for each complex not only capture the intrinsically intricate interactions within the complex but also represents the traits of the complex in contact with the whole dataset, which facilitates the downstream deep learning.

To define a specific domain for our analysis, we firstly pinpoint all heavy ligand atoms and the protein atoms within a predetermined distance, as shown in Figure 1**a**. And two versions of the model are available: one with a generous 20 Å cutoff and another with a 12 Å cutoff (suited for a more focused analysis). Next, in order to convert 3D molecular structures to admissible format, TopoFormer applies its unique topological sequence embedding module, as shown in Figure 1**b**. By employing a multiscale analysis, also known as a filtration process in algebraic topology, the 3D structures are transformed into topological sequences using our newly developed persistent topological hyperdigraph Laplacians (PTHLs). We further embed various physical, chemical, and biological interactions element-specific PTHLs. The outcome is a sequence of embedding vectors, enabled through the multiscale analysis of PTHLs. A more detail description of the topological sequence embedding module can be found in the methods section 4.2.

To take the advantages of a vast variety of unlabeled protein-ligand complexes, TopoFormer utilizes a self-supervised pretraining phase as depicted in Figure 1**c**. At its core lies the Transformer

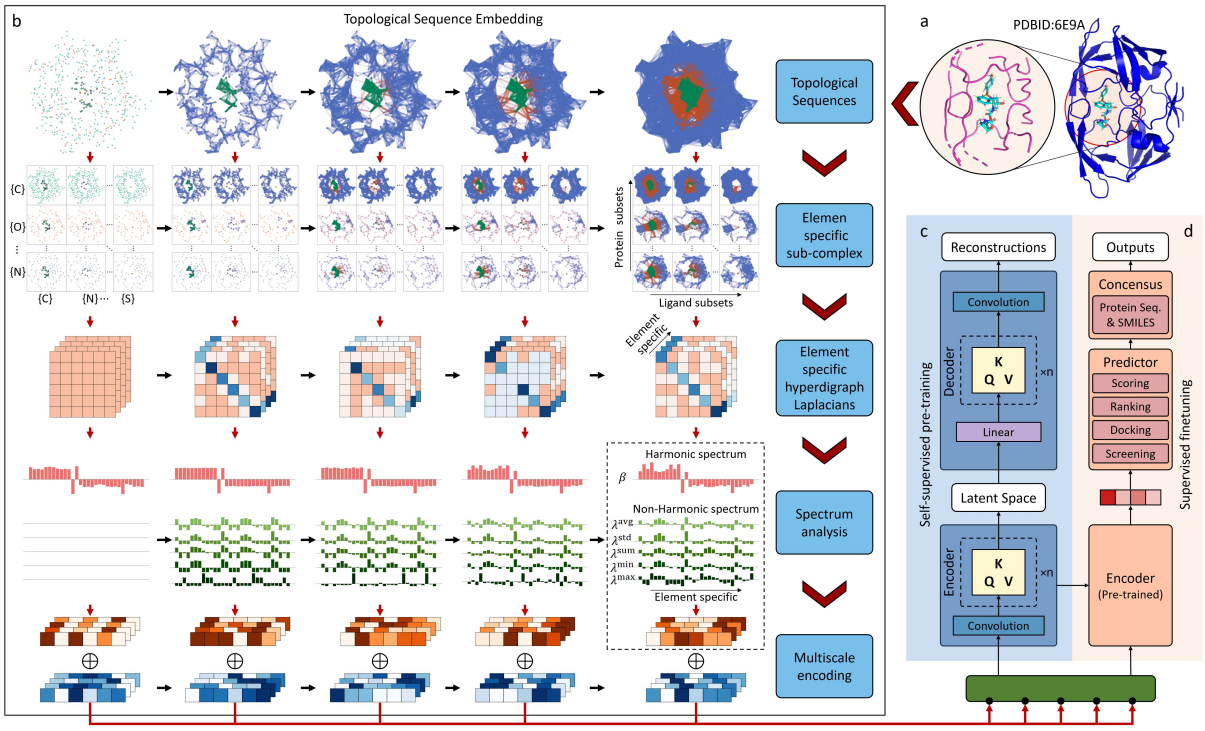

Figure 1: Schematic illustration of the overall TopoFormer model. **a**, A 3D protein-ligand complex (PDBID: 6E9A) and its interactive domain. **b**, The topological sequence embedding of a 3D protein-ligand complex. Initially, the complex is split into a topological sequence, known as a chain complex in algebraic topology. Then, element-specific sub-complexes are created to encode physical interactions a variety of scales controlled by a filtration parameter. Subsequently, element-specific persistent topological hyperdigraph Laplacians (PTHLs) are utilized to extract the topological invariant and capture the shape and stereochemistry of the subcomplexes. For these subcomplexes, their topological invariant changes over scales are retained in the harmonic spectrum of the hyperdigraph Laplacians, while their homotopic shape evolution over scales are manifested in the non-harmonic spectrum. Finally, the multiscale topological invariant changes and homotopic shape (stereochemical) evolution are resembled into a topological sequence as the input the Transformer. **c**, Self-supervised learning is applied to unlabeled topological sequences for both Transformer Encoders and Transformer Decoders. The outputs from the reconstructed topological sequences are used to calculate the reconstruction loss. **d**, At the supervised fine-tuning stage, task-specific protein-ligand complex data are fed into the pretrained encoder, which is equipped with specific predictor heads, such as the Scoring head, Ranking head, Docking head, and Screening head. Subsequently, except for the docking task, the remaining predictions are consolidated with sequence-based predictions to produce the final result.

encoder-decoder architecture, wherein the decoder diligently aims to reconstruct the topological sequence embedding from its encoded version. The precision of this phase is quantified by measuring the disparity between the output and input embeddings. Absent of labeled annotations, this step equips the model with an innate ability to decipher the intricate dynamics of protein-ligand interactions. Subsequent to pretraining, as illustrated in Figure 1**d**, the model acquaints itself with labeled protein-ligand complexes, transitioning to a supervised fine-tuning stage. Leveraging the pretrained encoder, the foremost embedded vector evolves into a pivotal latent feature, guiding a plethora of downstream tasks. Among TopoFormer's distinguishing attributes is its proficiency in executing multiple tasks, encompassing scoring, ranking, docking, and screening. Each task is equipped with its specialized head within the predictor module. To enhance precision, several topo-

logical transformer deep learning models (TF-DL) are initiated, each with a unique random seed, to mitigate initialization-related inaccuracies. Additionally, to temper the inherent biases of relying solely on one modeling approach, sequence-based models are also incorporated. Consequently, the conclusive output of TopoFormer is derived as an amalgamation of these varied predictions. The consensus methodology for each task will be elaborated upon in the subsequent task-specific results. In essence, TopoFormer is a holistic model tailored for a myriad of tasks in protein-ligand interaction analysis, bringing together topological insights and deep learning.

2.2 Evaluating TopoFormer on scoring tasks

The prediction of protein-ligand binding affinity plays a pivotal role in drug design and discovery. To assess the scoring capability of our models, we have evaluated them using the three most widely recognized protein-ligand datasets from the PDBbind database: CASF-2007, CASF-2013, and CASF-2016 [24, 25, 26]. The Pearson correlation coefficient (PCC) and the root mean squared error (RMSE) are used to measure the performance of the scoring function. For the kcal/mol unit conversion, we multiply the predicted values by 1.3633 in the predictions. In this task, we consider two TopoFormer models: a large model (TopoFormer) with an input topological sequence of length 100, employing filtration parameters at 0.1 intervals spanning from 0 Å to 10 Å. The topological analysis encompasses a domain extending up to 20 angstroms, centered at the ligand. Additionally, we employ a smaller model (TopoFormer$_s$) with an input topological sequence of length 50, using filtration parameters ranging from 2 angstroms to 12 angstroms, and with filtration parameter increments of 0.2 angstroms for constructing the corresponding simplex complex. The topological analysis covers a domain up to 12 angstroms, centered at the ligand.

To ensure robustness, 20 topological transformers are trained for each dataset with distinct random seeds to address initialization-related errors. Here, the predictions only from small topological transformers are denoted as TopoFormer$_s$. In addition, to attenuate systematic discrepancies inherent in a singular model approach, we deploy sequence-based models. Specifically, we harness embedded protein features from the ESM model [31] and the SMILES features from the Transformer-CPZ model [22]. And 20 gradient boosting regressor tree (GBRT) models are subsequently trained one these sequence-based features. The aggregated predictions from these models, denoted as Seq-ML, render a more holistic prediction. Thus, the final verdict results from a balanced average of TopoFormer and Seq-ML predictions, denoted as TopoFormer-Seq and TopoFormer$_s$-Seq for small TopoFormer model. Figure 2**b** and Figure 2**c** show the effect of consensus size (i.e., the number of randomly selected models) on performance. We performed 400 repetitions for each consensus size, taking the average result (solid line) and showing error variation (lighter-colored regions). It can be seen that the increasing in the consensus size improved performance metrics (higher PCC, lower RMSE) and stability (reduced error fluctuation). Ultimately, the consensus size is fixed as 10 for the subsequent comparisons. It can also be noticed that the TopoFormer-Seq performs best on almost all datasets, closely followed by the TopoFormer$_s$-Seq model.

To gain a comprehensive understanding of our models' performance, we benchmark our PCC results against representative results from the literature, as visualized in Figure 2**a** and Figure 7**a-b**. Remarkably, our TopoFormer-based models consistently achieved the highest PCC scores across all three benchmark datasets. The RMSE of our model is also the lowest in all three benchmark datasets when compared to methods with accessible RMSE (1). In this work, the TopoFormer-

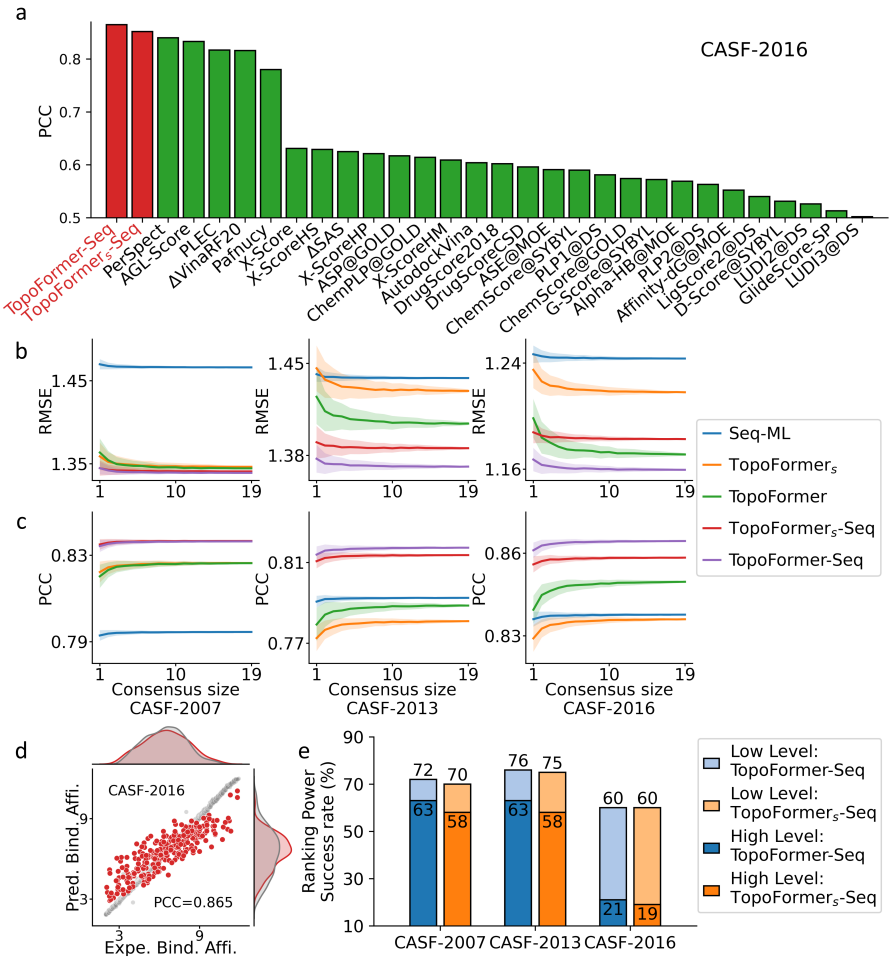

Figure 2: Performance of TopoFormer on scoring and ranking tasks. **a** Comparison of Pearson correlation coefficients (PCCs) of various models for protein-ligand complex binding affinity scoring on the CASF-2016 benchmark. The results from other methods are in the green color, taking from Refs [25, 24, 26, 16, 27, 19, 28, 29, 30]. **b** Comparison of the RMSEs of predictions for the CASF-2007, CASF-2013, and CASF-2016 datasets from the Seq-ML model, TopoFormer model, TopoFormer$_s$ model, TopoFormer$_s$-Seq, and TopoFormer-Seq. The horizontal axis is the number of models in the consensus (consensus size). The solid line represents the median RMSE, while the shaded background provides the error bar for these 400 RMSE values. **c** Comparison of the PCCs of predictions for the CASF-2007, CASF-2013, and CASF-2016 datasets from the Seq-ML model, TopoFormer model, TopoFormer$_s$ model, TopoFormer$_s$-Seq, and TopoFormer-Seq. The horizontal axis is the consensus size. The solid line represents the averages, while the shaded background provides the error bar for 400 PCCs at each consensus size. **d** The correlation between predicted protein-ligand binding affinities (TopoFormer PCC=0.865) and experimental results for the CASF-2016 benchmarks. Grey dots represent the training data, while red dots denote the test data. **e**, Comparison of the ranking power assessed using both high-level success measurements (depicted in dark shades) and low-level success measurements (shown in lighter shades) across three benchmarks. Results from TopoFormer-Seq are represented in blue, while those from TopoFormer$_s$-Seq are illustrated in orange.

based model's performance is quantified by calculating averages from 400 repetitions, and the results are tabulated in Table 1. Across all three datasets, Transformer-Seq achieves an average PCC of approximately 0.84. For a detailed comparison of various models trained on the same dataset, please refer to Table S1. Notably, in the case of the PDBbind v2016 dataset, [26] which has five

more components (290) in its test set compared to the CASF-2016 core set (285), our TopoFormer-Seq model also demonstrated state-of-the-art performance with a PCC of 0.866 and a low RMSE of 1.561 kcal/mol. The detailed information of these three benchmarks can be found in Table S 2. Figure 2**d**, and Figure 7**c-d** visualized the comparisons of predicted protein-ligand binding affinities and experimental results for the test set of CASF-2007, CASF-2013, and CASF-2016 benchmarks.

Table 1: The PCCs(RMSE in kcal/mol) of our TopoFormer models on the three benchmarks of CASF-2007, CASF-2013, and CASF-2016. TopoFormer and TopoFormer$_s$ are considered. The averages of 400 repetitions are computed as the performance of the model. The detailed setting of two TopoFormers and GBRT parameters can be found in Supplementary Information Section A.2.

Dataset	CASF-2007	CASF-2013	CASF-2016	Average
TopoFormer-Seq	0.837(1.807)	0.816(1.859)	0.864(1.568)	0.839(1.745)
TopoFormer$_s$-Seq	0.839(1.798)	0.809(1.886)	0.855(1.609)	0.834(1.764)
TopoFormer	0.826(1.830)	0.788(1.910)	0.849(1.595)	0.821(1.778)
TopoFormer$_s$	0.826(1.832)	0.781(1.944)	0.836(1.657)	0.814(1.811)
Seq-ML	0.798(1.974)	0.790(1.960)	0.837(1.693)	0.808(1.876)

Recently, several deep learning models have been reported for the prediction of protein-ligand binding affinity. Notable examples include the graphDelta model[32], ECIF model[33], OnionNet-2 model[34], DeepAtom model[35], and others[36, 37, 38]. These new models typically leverage on large training datasets that incorporate additional data from the general sets of the PDBbind database and thus are not comparable with other models that were trained on different training datasets. The details regarding the composition of training sets, testing sets, and their corresponding performance, are tabulated in Table S 5. In this study, the proposed TopoFormer model was trained strictly on the refine dataset following the standard procedure [24, 25, 26], representing a smaller subset compared to the general set. The results obtained with TopoFormer, trained on the refine set (which is notably smaller), outperform the majority of these models. Furthermore, for the latest PDBbind v2020 [39], we consider a total of 18,904 protein-ligand complexes for training, which has no overlap with the core sets of CASF-2007, CASF-2013 and CASF-2016. Our model achieved a commendable final PCC of 0.853 and an RMSE of 1.295 (equivalent to 1.769 kcal/mol) on the core set of CASF-2007. For the CASF-2013 core set, the PCC of 0.832 and an RMSE of 1.301 (equivalent to 1.777 kcal/mol) are obtained. Similarly, on the CASF-2016 core set, we obtained a PCC of 0.881 with an RMSE of 1.095 (equivalent to 1.496 kcal/mol). For the PDBbind v2016 core set, we achieved a PCC of 0.883 with an RMSE of 1.086 (equivalent to 1.483 kcal/mol). Here, all the results are the average of 400 repeated experiments. These results underscore the robustness and predictive power of the TopoFormer model in the realm of protein-ligand binding affinity predictions.

2.3 Evaluating TopoFormer on ranking tasks

The efficacy of a scoring function is critically assessed by its aptitude to accurately rank the binding affinities of protein-ligand complexes within distinct clusters. The benchmarks CASF-2007 and CASF-2013 comprise 65 clusters, with each cluster containing three complexes formed by an identical protein partnered with varied ligands [25, 24]. On the other hand, the CASF-

2016 benchmark encompasses 57 clusters, each having five distinct complexes [26]. In this work, two evaluative approaches are employed: the high-level and the low-level success measurements. In the high-level success metric, the objective is to perfectly rank the binding affinities of the complexes within each cluster. Conversely, the low-level success criterion requires the scoring function to merely identify the complex with the pinnacle binding affinity. The assessment of ranking efficacy termed "ranking power" is gauged by the proportion of correctly identified affinities across a specified benchmark. The mathematical formulations of the high-level and low-level success measurements can be found in the Supplementary Materials Section A.1.

Figure 2**e** illustrates the ranking power of TopoFormer-based models. For the CASF-2007, the TopoFormer-Seq model achieved outstanding success rates, with 72% for low-level measurement and 63% for high-level measurement. In comparison, the TopoFormer$_s$-Seq model achieved success rates of 70% for low-level and 58% for high-level measurement. Both models outperformed previous approaches, as demonstrated in high-level measurement Figure S8 and low-level measurement Figure S9. Similarly, for the CASF-2013, the TopoFormer-Seq model achieved remarkable success rates of 76% for low-level and 63% for high-level measurement, surpassing the performance of earlier models. The challenges intensified in CASF-2016, comprising 57 clusters, each containing five distinct complexes [26], making ranking tasks notably more demanding. In this context, the TopoFormer-Seq model achieved a success rate of 60% for low-level measurement and 21% for high-level measurement. The best-performing models for low-level (68%) and high-level (29%) success were ΔVinaRF20 [30].

2.4 Evaluating TopoFormer on docking tasks

Molecular docking stands as a formidable computational tool, essential in the fields of drug discovery, structural biology, and the elucidation of molecular intricacies underlying biological processes. The pivotal role of a robust scoring function becomes evident when selecting the most promising binding poses and predicting binding affinities. In the present study, we harnessed the capabilities of TopoFormer$_s$ (Due to computational resource constraints, we only employed TopoFormer$_s$ for both docking and screening tasks.) to assess its docking proficiency, particularly its ability to distinguish native binding poses from those generated by established docking software packages. Our evaluation centered on benchmark datasets CASF-2007 and CASF-2013 [25, 24]. Each dataset comprises a total of 195 test ligands, with each ligand accompanied by 100 poses generated by various docking programs. A pose was considered native if its root mean square deviation (RMSD) with respect to the true binding pose was less than the 2 Å threshold. Successful prediction occurred when the pose with the highest predicted binding energy matched a native pose. Following this comprehensive evaluation encompassing all 195 test ligands, an overall success rate was computed for the employed scoring function. Additional information detailing the assessment of docking success rates is available in the Supplementary Information Section S A.1.

In the field of molecular docking, several noteworthy approaches have made significant strides, each contributing uniquely to our understanding of protein-ligand interactions. Notable among these are DeepDock [40], which achieved a commendable success rate of 62.11%, OnionNet-SFCT [41] further enhanced performance to an impressive 76.84%, followed by DeepBSP [42] at 79.7%, and RTMScore [43] reaching a remarkable 80.7% success rate on the PDBbind core set. It is noteworthy that these methodologies were trained on diverse datasets, making direct comparisons

challenging. In the pursuit of a comprehensive evaluation, we utilized the publicly available training data to train the TopoFormer$_s$. We then conducted a rigorous comparison on the CASF-2007 and CASF-2013 datasets, fostering a fair and unbiased assessment of our methodology [27, 30, 44]. Detailed pose data and labels are provided in Section 4.1. Impressively, as depicted in Figures 3**f** and 3**g**, TopoFormer$_s$ attained an exceptional success rate of 93.3% on the CASF-2007 core set and 91.3% on the CASF-2013 core set. TopoFormer$_s$ outperformed other established docking tools and models, highlighting the effectiveness of our topological approach. Our methodology stands as a testament to the richness of approaches in the field, harnessing innovative techniques and a meticulously curated dataset to achieve remarkable success rates in docking tasks, while ensuring fairness in comparison. It offers a fresh perspective and a robust toolkit for the docking challenge.

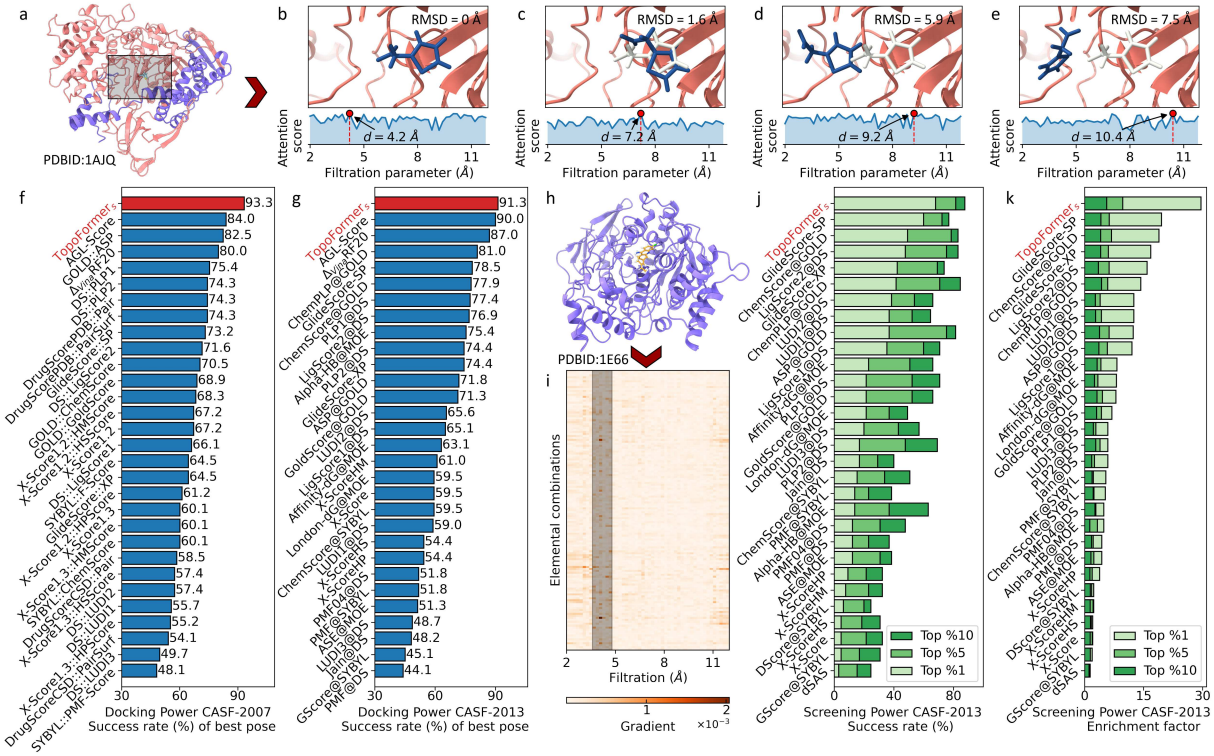

Figure 3: Performance of TopoFormer on docking and screening tasks. **a**, Visualization of the protein-ligand complex PDBID: 1AJQ. The highlighted rectangle shows the protein's pocket area. **b-e**, Four distinct ligand poses within the protein 1AJQ. The molecule in light gray represents the true pose, while the blue molecules depict alternative poses with RMSD values of 0 Å, 1.6 Å, 5.8 Å, and 7.5 Å, respectively. The light blue curve represents the attention score generated by TopoFormer, varying with the filtration parameter (i.e., the scale) of the topological embedding. The highest attention scores are observed at scales of d=4.2 Å, d=7.2 Å, d=9.2 Å, and d=10.4 Å for poses from **b** to **e**. **f-g**, Comparison of docking success rates between TopoFormer$_s$ and traditional docking tools on the CASF-2007 core set (**f**) and the CASF-2013 core set (**g**). **h**, Visualization of the protein-ligand complex PDBID: 1E66. **i**, The saliency map of the topological embedding for complex 1E66. The colorbar represents the gradient weights of each feature relative to the prediction. **j**, Comparison of screening success rates for the top 1%, top 5%, and top 10% selected ligands between TopoFormer$_s$ and docking tools on the CASF-2013 core set. **k**, Comparison of average enhancement factors for the top 1%, top 5%, and top 10% selected ligands between TopoFormer$_s$ and docking tools on the CASF-2013 core set.

In order to better understand what Topoformer$_s$ learned from training after fine-tuning, we explored which filtration parameter, i.e., the spatial scale, had the greatest impact on protein-

ligand interactions through attention scores. Figures 3**b-e** show the four poses of the ligand in the vicinity of the protein (PDBID: A1JQ) pocket (black boxed portion in Figure 3**a**). Where Figure 3**b** is the real pose measured by experiment, which has an RMSD of 0 Å. After training, we obtain the TopoFormer$_s$'s attention score for all filtration parameters, i.e., the average of the attentional weights of all heads in all TopoFormer layers. This attention score indicates the magnitude of the impact of the protein-ligand interaction of each range on the final docking score. From Figure 3**b**, it can be noticed that the largest attention score occurs at $d = 4.2$ Å, which generally indicates that the interactions at ranges with the scale of 4.2 Å have the largest impact on the binding affinity of this pose. Similarly, Figure 3**c-e** show poses with RMSDs of 1.6 Å, 5.9 Å, and 7.5 Å, respectively, where the light gray compounds refer to ligand's pose when the RMSD is zero. Their corresponding maximum attention score, on the other hand, occur at scales $d = 7.2$ Å, $d = 9.2$ Å, and $d = 10.4$ Å, respectively, which are positively correlated with RMSDs. It indicates that the more a pose deviates from the true pose position, the greater the scale at which the interactions have the greatest impact on the docking score is.

2.5 Evaluating TopoFormer on screening tasks

The screening task in biology is of paramount importance in identifying potential drug candidates and advancing drug development endeavors. To assess the screening capabilities of our TopoFormer method, we employ the CASF-2013 core set in this study. Given that the evaluation of screening power necessitates the identification of three true binders for each of the 65 proteins in the core set, we take the crucial step of fine-tuning the pre-trained TopoFormer$_s$ model. For this purpose, we assemble a training dataset encompassing both ligand poses and energy labels, customizing TopoFormer$_s$ for each protein target. Our screening task comprises two key steps. First, we generate poses for the 195 ligands through a docking procedure and predict their scores using TopoFormer$_s$, denoted as S_1. Subsequently, we employ a sequence-based classification gradient boosting decision tree model, leveraging combined features from the Transformer-CPZ model [22] and the ESM model [31] for these 195 ligands and the respective target proteins. This yields probabilities for the given ligands, referred to as S_2. Ligands with high multiplied scores ($S = S_1 * S_2$) are identified as predicted binders. Consistent with prior research, the training set for each target protein comprises all complex structures and their associated energy labels from the PDBbind v2015 refine set, excluding the core (test) set complexes. Furthermore, for each target protein, additional poses and their corresponding labels in the training set are generated [45, 27]. Comprehensive pose data and labels for the screening task can be found in Section 4.3. Here, due to computational resource constraints, we only utilize TopoFormer$_s$ for virtual screening. Additionally, in this work, the success rate and enrichment factor (EF) are used in the virtual screening for drug discovery. The success rate measures the proportion of true positive predictions among the top-ranked compounds. And the enrichment factor is a measure of how well a screening method enriches the dataset with active compounds (true binders) at the top of the ranked list, specifically 1%, 5%, and 10%, compared to a random selection. It provides insight into the ability of the method to prioritize active compounds over non-active ones. The detailed definitions for both success rate and enrichment factor are provided in Supplementary Information Section SA.1.

As suggested by Figures 3**j** and 3**k**, the proposed TopoFormer model outperformed the previous methods in all two metrics, e.g., success rate and enrichment (EF), compared with popular

conventional methods. Concretely, for the task of identifying true binders for a certain protein, TopoFormer attains a success rate of 68% and an average enhancement factor (EF) of 29.6% on top 1%-ranked molecules. It was better than AGL-score [27] (success rate=68%, EF=25.6) and ΔVineRF20 [30] (success rate=60%, EF=20.9), whose were validated only for the top 1% ranked molecules for the CASF-2013 dataset. In addition, the results of proposed method are greatly higher than those of the second best performing method GlideScore-SP (with the success rate of 60% and EF of 19%). Additionally, the TopoFormer reaches higher success rates of 81.5% and 87.8% on the top 5% and top 10% ranked molecules. The averaged enrichment factor are 9.7 and 5.6 on top 5% and top 10%, respectively. The AGL-score (success rate=68%, EF=25.6) and ΔVineRF20 (success rate=60%, EF=20.9), were validated only for the top 1% ranked molecules for the CASF-2013 dataset. It also worth to note that some recent deep learning-based models, such as RTMScore [43] (with EF of 28 and success rate of 66.7%), DeepDock [40] (with EF of 16.4 and success rate of 43.9%), and PIGNet [46] (with EF of 19.36 and success rate of 55.4%), but all these model are evaluated on the CASF-2016 core set and trained on different training data, so there is no direct comparison with these method.

As depicted in Figures 3**j** and 3**k**, the proposed TopoFormer model surpasses previous methods across both key metrics: success rate and enrichment factor (EF). When tasked with identifying true binders for specific proteins, TopoFormer demonstrates a remarkable success rate of 68% and an average enhancement factor (EF) of 29.6% for the top 1%-ranked molecules. TopoFormer's results significantly outshine those of the second-best performing method, GlideScore-SP (success rate of 60% and EF of 19%). Furthermore, TopoFormer exhibits high success rates of 81.5% and 87.8% for the top 5% and top 10% ranked molecules, respectively. The corresponding averaged enrichment factors are 9.7 and 5.6 for the top 5% and top 10%, which are the highest performance as shown in Figure 3**k**. Notably, AGL-score [27] (success rate=68%, EF=25.6) and ΔVineRF20 [30] (success rate=60%, EF=20.9) were assessed solely for the top 1% ranked molecules on CASF-2013 dataset core set. It's worth highlighting some recent deep learning-based models, including RTMScore [43] (success rate of 66.7% and EF of 28), DeepDock [40] (success rate of 43.9% and EF of 16.4), and PIGNet [46] (success rate of 55.4% and EF of 19.36). However, it is important to note that these models were evaluated on the CASF-2016 core set and trained on different datasets, making direct comparisons with our method impractical.

To understand which scales of the protein-ligand interactions have the most significant impact on the model's predictions, the saliency map is generated from finetuned TopoFormer$_s$ for a given protein-ligand complex (PDBID:1E66), as shown in Figure 3**h**. Specifically, the protein atom within 12 Å around the ligand is considered in the analysis. As suggested from the Figure 3**i**, the y-axis corresponding to different element-specific combination of the given complex, and the x-axis is the filtration parameter from 2Å to 12Å. The color bar indicates the gradient on each feature of the topological embedding. The gradients which are significantly higher than the others have been marked with black area, and it denotes the filtration parameter around 4 Å. The saliency map provides insight into the model's decision-making process by highlighting the relative importance of the topological embedding features at various scales. That means the heavy-atom protein-ligand interaction around scale 4 Å has a stronger influence on the TopoFormer$_s$ output in the screening task, which is reasonable as hydrogen atoms are not presented in the PDBbind database and our models.

In order to discern the critical facets of protein-ligand interactions that wield the most profound influence on model's predictions, we have employed the generation of a saliency map with the fine-tuned TopoFormer$_s$ for a specific protein-ligand complex (PDBID: 1E66), as illustrated in Figure 3h. Our analysis focuses specifically on protein atoms located within a 12 Å radius around the ligand. As depicted in Figure 3**i**, the y-axis corresponds to different element-specific combinations within the given complex, while the x-axis represents the filtration parameter ranging from 2 Å to 12 Å. The color bar visually signifies the gradient assigned to each feature within the topological embedding. Distinctive gradients, significantly elevated compared to others, are demarcated by black regions on the map, specifically around the scale of 4 Å in the filtration parameter. The saliency map serves as an invaluable tool for gaining insights into the decision-making process of our model. It accomplishes this by accentuating the relative importance of topological sequence embeddings at various scales. Consequently, it becomes evident that the heavy-atom protein-ligand interactions occurring at approximately 4 Å radius exert a more substantial influence on the output of TopoFormer$_s$ in the screening task.

3 Discussion

In this section, we aim to unravel the intricate web of insights that the TopoFormer brings to the realm of protien-ligand interactions. At its core, the model harnesses the power of persistent topological hyperdigraph Laplacian features, a strategic choice that imbues our framework with a unique prowess in deciphering interaction landscapes.

In this study, we employ the persistent topological hyperdigraph Laplacian to give a comprehensive representation for 3D protein-ligand complexes, surpassing traditional graph, simplicial complex, and hypergraph structures (refer to Figure 13). The topological hyperdigraph naturally captures higher-order relationships by allowing directed hyperedges to connect vertices with specific orders, as illustrated in Figure 4**c**. These directed hyperedges, spanning 0 to 3 dimensions, offer a flexible framework for modeling intricate interactions in protein-ligand complexes, accommodating relationships beyond pairwise connections. By employing directed hyperedges of varying dimensions, our approach provides a nuanced representation of the system's underlying structure. Additionally, introducing orientations enables encoding of physical/chemical knowledge into directed hyperedges, reflecting differences in electronegativity, atomic radius, weights, and ionization energy for distinct elements. This enhancement serves as an improvement over traditional graph, simplicial complex, and hypergraph representations. Figures 14 **g** and **h** showcase hyperdigraph representations for a multi-elemental system, specifically two $B_7C_2H_9$ isomers, highlighting the capacity to capture different elemental configurations through the directionality of corresponding directed hyperedges.

In the investigation of protein-ligand complexes, we introduce the use of topological hyperdigraphs as an initial step to represent these intricate molecular systems. Subsequently, we incorporate the persistent topological hyperdigraph Laplacian theory [20], to establish a robust and comprehensive framework for analyzing the geometric and topological characteristics of protein-ligand complex systems. Drawing inspiration from physical systems like molecular structures, where the zeroth-dimensional Laplacian matrix is linked to the kinetic energy operator in the Hamiltonian in quantum mechanics [47], we extend this analogy to topological hyperdigraphs. The Laplacian

energy, associated with the eigenvalues of the Laplacian matrix in a hyperdigraph context, becomes a valuable tool. These Laplacian eigenvalues offer insights into various properties of the topological object, bearing connections to the energy spectrum of physical systems. Notably, our proposed topological Laplacian analysis in this work provides a means to elucidate the structural and energetic characteristics of complex systems, aligning with fundamental principles in physical systems.

Moreover, in comparison to traditional persistent homology theory, the proposed persistent topological hyperdigraph Laplacian presents significant advancements on multiple fronts. Firstly, it has been demonstrated to effectively analyze the topological hyperdigraph, a high-level generalization encompassing traditional graphs, digraphs, simplicial complexes, and hypergraphs, surpassing the limited applicability of traditional persistent homology theory, which is confined to simplicial complexes. Secondly, the persistent topological hyperdigraph Laplacian provides a more comprehensive approach to characterize protein-ligand complexes. It not only encapsulates the fundamental homology information, such as Betti numbers representing connected components, loops, voids, and higher-dimensional features, but also incorporates additional geometric insights and homotopic shape evolution derived from the non-harmonic spectra of the persistent Laplacians. Illustrated in Figure 10, panels **a-e** depict the results of persistent topological hyperdigraph Laplacian analysis for the protein-ligand complex, contrasting with traditional homology analysis in panel **f**. Importantly, it has been confirmed that the multiplicity of zero eigenvalues of the Laplacians corresponds to the Betti numbers, indicating that the barcode information in Figure 10**f** is encompassed by persistent topological hyperdigraph Laplacians [20], exemplified in Figure 10**e**.

Considering the diverse scales at which atomic interactions unfold—encompassing phenomena like covalent, ionic, dipole-dipole, and Van der Waals interactions-it becomes apparent that a comprehensive analysis is vital. The proposed persistent topological hyperdigraph Laplacian introduces persistence, offering a multiscale examination of the system. This manifests as a topological sequence evolving with changing in scales, i.e., filtration parameters in the algebraic topology sense, effectively capturing interactions across various scales. This approach proves invaluable in guiding transformer models to discern the distinct contributions of each scale to the desired property, such as binding affinity in protein-ligand complexes, throughout the fine-tuning process. Illustrated in Figures 3 **b** to **e** are visualizations depicting the contribution of different scales, i.e., attention scores, for diverse protein-ligand complexes.

Within physical systems, such as the protein-ligand complexes explored in this study [16, 48, 49], a myriad of elemental interactions intricately governs molecular stability and specificity. Hydrogen bonding, van der Waals forces, ionic and polar interactions, nonpolar hydrophobic forces, as well as pi-stacking and dipole-dipole interactions collaboratively mold the structural integrity of the complex. These diverse interactions play pivotal roles in substrate recognition, stability, and the overall specificity of binding events. Recognizing the significance of elemental-level interactions is crucial for deciphering molecular recognition mechanisms, shaping drug design strategies, and advancing our understanding of complex biological processes. To incorporate the elemental interactions between proteins and ligands, we introduce an element-specific analysis, as illustrated in the element-specific hyperdigraph Laplacians module within the topological sequence embedding (Figure 1**b**). Specifically, interactions between proteins and ligands are considered by constructing sets of common heavy elements in proteins (4 types) and ligands (9 types). Sub-hyperdigraphs

of the overall protein-ligand hyperdigraph are generated based on different combinations of these elemental sets, leading to the construction of element-specific Laplacian matrices for each sub-hyperdigraph. The analysis of these matrices encodes the elemental-level interactions within the protein-ligand complex. This element-specific technique enhances the extraction of richer physical and chemical features, aiding the transformer model in comprehending the intricate internal dynamics of protein-ligand complexes under both self-supervised and supervised learning paradigms. More details about the element-specific analysis can be found in the Method Section 4.

4 Methods

4.1 Datasets

The dataset utilized for pre-training in this study is a comprehensive compilation of protein-ligand complexes (without the labels) sourced from the diverse PDBbind database, including CASF-2007, CASF-2013, CASF-2016, and PDBbind v2020 [39]. To ensure the dataset's integrity and to eliminate redundancies, a rigorous curation process was meticulously conducted, resulting in a total of 19,513 non-overlapping complexes for pre-training. Rigorous training-test splitting is employed and advocated in this work. For the standard scoring and ranking tasks, the training set comprises the defined refine set, excluding the core set, from PDBbind CASF-2007 (equivalent to PDBbind v2007), CASF-2013 (equivalent to PDBbind v2013), CASF-2016, and PDBbind v2016 datasets. The test set encompasses the respective core sets of these datasets. Given the absence of a core set in PDBbind v2020, the general set (19443), excluding the all core sets from CASF-2007, CASF-2013, CASF-2016, and PDBbind v2016, is employed as the training set (18,904) for the large TopoFormer model. This approach enables a meaningful comparison with recently developed models that have been trained using different data sources. Further details regarding the datasets can be found in Table 2.

Table 2: Detailed information of the used datasets.

	Datasets	Training set	Test set (core set)
Pretraining (Self-supervised Learning)	Combind PDBbind (CASF-2007, CASF-2013, PDBbind v2015, CASF-2016,v2020)	19513	/
Finetuning (Supervised Learning)	CASF-2007	1105	195
	CASF-2013	2764	195
	CASF-2016	3772	285
	PDBbind v2016	3767	290
	PDBbind v2020	18904	195(CASF-2007 core set)
			195(CASF-2013 core set)
			285(CASF-2016 core set)

For the docking task, the test sets were sourced from the benchmark datasets CASF-2007 and CASF-2013. Each of these datasets consists of 195 test ligands, and for each ligand, 100 poses are generated using various docking programs [25, 24]. In preparation for the docking task training

set, a set of 1000 training poses are generated for each given target ligand-receptor pair within the test set. These training poses were generated using GOLD v5.6.33 [50]. Consequently, for both CASF-2007 and CASF-2013, there was a total of 365,000 training poses available for fine-tuning purposes. The pose structures and their corresponding scores, as reported by GOLD, are accessible at https://weilab.math.msu.edu/AGL-Score.

For the screening task, the core set of CASF-2013 was utilized as the test dataset. This set comprises 65 proteins, and each protein interacts with three true binders selected from the 195 ligands within the core set [24]. Regarding the training set, for each target protein present in the test set, the training dataset was constructed using all complex structures and their associated energy labels from the PDBbind v2015 refine set. Notably, the core (test) set complexes were excluded from this training dataset. To augment the training dataset, additional poses and their corresponding labels were generated [45, 27]. It is worth mentioning that the list of true binders for each protein is available in the CASF 2013 benchmark dataset. For each ligand, the pose with the highest energy was used as the upper bound for the training set. All pose structures and their scores can be accessed at https://weilab.math.msu.edu/AGL-Score.

4.2 Topological sequence embedding

Topological Hyperdigraph. The topological hyperdigraph serves as a versatile generalization, encompassing digraphs, simplicial complexes, and hypergraphs. It excels in representing intricate relationships such as multi-source to multi-target mappings and asymmetric connections, which are challenging to convey within traditional graphs or simplicial complexes [20]. In essence, the topological hyperdigraph consists of sequences of distinct elements within a finite set, known as directed hyperedges, acting as the fundamental building blocks. Figure 4**c** provides examples of 0-directed, 1-directed, 2-directed, and 3-directed hyperedges. Notably, these sequences bear a resemblance to the simplices in a simplicial complex. Figure 4**b** illustrates the 0-simplex (a node), 1-simplex (line segment), 2-simplex (a triangle), and 3-simplex (a tetrahedron) for comparison. For a more in-depth understanding of commonly used graph, simplicial complex, and hypergraph definitions, refer to the Supplementary Information in Section A.3.

A *hyperdigraph* $\vec{\mathcal{H}}$ consists of a vertex set V and a collection of sequences with distinct elements in V. An sequence of length $k+1$ in $\vec{\mathcal{H}}$ is called a k-directed hyperedge. Mathematically, a k-directed hyperedge is an inclusion map $e : [k] \to V$, here $[k] = \{0, 1, \ldots, k\}$. A hyperdigraph is a collection of directed hyperedges on V. Sometimes, we denote $\vec{\mathcal{H}} = (V, \vec{E})$, where $\vec{E}$ is the set of directed hyperedges. In particular, if the set V is an ordered set, and all directed hyperedges are ordered, then the hyperdigraph can be reduced to a hypergraph. If all directed edges are restricted to be one-dimensional, hyperdigraphs can be simplified to the usual directed graphs. In this sense, hyperdigraphs act more like a versatile aggregator, offering a more flexible and diverse portrayal of data.

More formally, let $C_k(V; G)$ be the abelian group generated by the sequences with $(k+1)$ distinct elements in V. Then $C_*(V; G)$ is a chain complex with the boundary operator $\partial_k : C_k(V; G) \to C_{k-1}(V; G)$ given by

$$d_k(x_0, x_1, \ldots, x_k) = \sum_{i=0}^{k} (-1)^k (x_0, \ldots, \hat{x_i}, \ldots, x_k). \tag{1}$$

Here, $\widehat{x_i}$ means omission of the term x_i. Let $F_k(\vec{\mathcal{H}}; G)$ be the abelian group generated by the k-directed hyperedges on $\vec{\mathcal{H}}$. It follows that $F_k(\vec{\mathcal{H}}; G)$ is a graded subgroup of $C_*(V; G)$. We denote

$$\Omega_k(\vec{\mathcal{H}}; G) = \{x \in F_k(\vec{\mathcal{H}}; G) | \partial_k x \in F_{k-1}(\vec{\mathcal{H}}; G)\}. \tag{2}$$

Then, $\Omega_k(\vec{\mathcal{H}}; G)$ is also a chain complex, specifically tailored for exploring the topology of hyperdigraphs. It is essential to highlight that the chain complex $\Omega_k(\vec{\mathcal{H}}; G)$ undergoes simplification when the hyperdigraph is transformed back into a simplicial complex or hypergraph. The corresponding simplicial complex representation of C_α atoms in protein 6L9D is depicted in Figure 4**h**. Here, blue triangles represent the 2-simplices, while orange highlights designate the 3-simplices, providing a rough visualization of the alpha helix structures. Additionally, Figure 4**i** illustrates the 3-directed hyperedges within the hyperdigraph, highlighted in blue, serving as an alternative representation of the alpha helix in the structure. Figure 13 further presents diverse topological representations, encompassing graphs, simplicial complexes, hypergraphs, and hyperdigraphs. More detailed descriptions and definitions of graphs, simplicial complexes, and hypergraphs are available in the Supplementary Information (see Section A.3) and the original paper [20].

Vietoris-Rips hyperdigraph and alpha hyperdigraph. The Vietoris-Rips (VR) complex and the alpha complex stand out as the most popular topological models for characterizing sets of data points. In the case of $\mathcal{K}$ being a VR complex or an alpha complex, the points forming a simplex in $\mathcal{K}$ inherently carry geometric information, encompassing both the magnitude and orientation of the point set. Motivated by the definitions of the VR complex and alpha complex, we introduce the Vietoris-Rips (VR) hyperdigraph and alpha hyperdigraph to capture such geometric information. We employ a weight function $w : \mathcal{K} \to \mathbb{R}$ and a graded orientation function $\varrho_n : \mathcal{K}_n \to S_{n+1}$ for $n \geq 1$ to articulate the geometry of simplices. Here, S_n denotes the permutation group of n elements. The VR/alpha hyperdigraph is defined as

$$\vec{\mathcal{H}}_\eta := \{S \times \varrho_*(S) | w(S) \leq \eta, S \in \mathcal{K}\}. \tag{3}$$

The VR/alpha hyperdigraph can be regarded as a generalization of the VR/alpha complex. It simplifies to the VR/alpha complex when assuming the functions $w : \mathcal{K} \to \mathbb{R}$ and $\varrho_n : \mathcal{K}_n \to S_{n+1}$ are constant.

In this work, all analyses are derived from the VR hyperdigraph without specified instructions. The demonstrations of VR/alpha hyperdigraphs can be found in Figures 5 and 6. The detailed constructions of VR hyperdigraphs and alpha hyperdigraphs are provided in the Supplementary Information A.4.

Topological Laplacians and spectrum analysis The combinatorial Laplacian is a fundamental tool in discrete geometry and algebraic topology. It offers a way to understand the structure of the topological system, such as simplicial complexes, hypergraphs, and hyperdigraphs. Just as the graph Laplacian can be used to study properties of graphs (the graph can be regard as 1-simplices), the combinatorial Laplacian can be used to study properties of simplicial complexes and hyperdigraphs. The eigenvalues of the graph Laplacian can translate the connectivity information of the graph. For example, the second smallest eigenvalue, also known as the Fiedler vector, reflects the algebraic connectivity of the graph, and the smallest positive eigenvalue, also known as the spectral

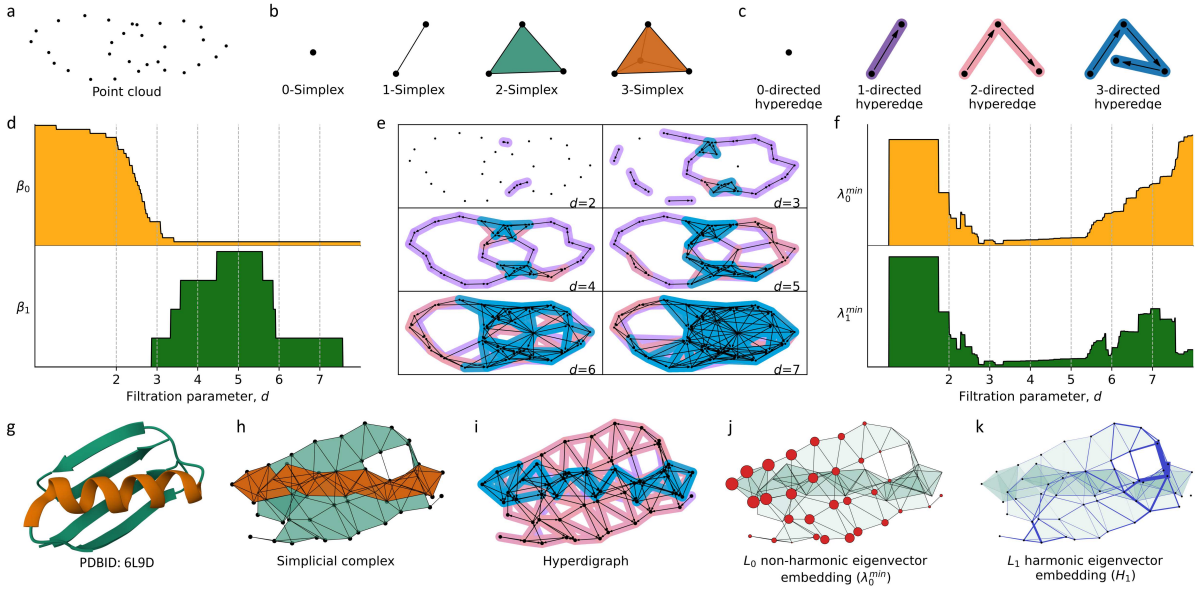

Figure 4: Illustration of the concepts related to topological sequence embedding. **a**, Representation of structural data as a point cloud. **b**, Depiction of 0-simplex (node), 1-simplex (edge), 2-simplex (triangle), and 3-simplex (tetrahedron), which serve as the fundamental building blocks of a simplicial complex. **c**, Illustration of 0-directed hyperedge, 1-directed hyperedge, 2-directed hyperedge, and 3-directed hyperedge, which form the basic building blocks of a hyperdigraph. **d**, Visualization of the multiplicity of zero spectra, i.e., topological invariants, of the persistent topological hyperdigraph at the 0th (β_0) and 1st (β_1) dimensions, respectively, showcasing their variations with respect to the filtration (scale)parameter d. **e**, Illustration of the impact of varying the filtration parameter on multiscale analysis, resulting in changes in the connectivity of the point cloud and the creation of a sequence of hyperdigraphs, representing a series of topological structures. **f**, Representation of nonzero minimum non-harmonic spectra of the persistent topological hyperdigraph Laplacian at the 0th and 1st dimensions (λ_0^{min} and λ_1^{min}), highlighting their dependence on the filtration parameter d. **g**, Visualization of protein 6L9D with a representation featuring only C_α atoms. The alpha helix is highlighted in orange, while the beta helix is shown in green. **h**, Illustrations of simplicial complex representation for the C_α atoms of protein 6L9D at a cutoff distance of $d = 5$ Å. The 2-simplices are filled by green, 3-simplices are colored by orange. **i**, Visualizations of hyperdigraph representations for the C_α atoms of protein 6L9D at a cutoff distance of $d = 5$ Å. The 1-directed hyperedges are depicted as purple edges with arrows, the 2-directed hyperedges are represented by pink edges with arrows, and the 3-directed hyperedges are illustrated as blue edges with arrows. **j**, Description of the L_0 nonzero minimum non-harmonic eigenvector embedding for the C_α atoms of protein 6L9D at a cutoff distance of $d = 5$ Å. **k**, Explanation of the L_1 harmonic eigenvector embedding for the edges between the C_α atoms of protein 6L9D at a cutoff distance of $d = 5$ Å.

gap, is closely related to the Cheeger constant. The collection of eigenvalues for the Laplacian operator is the spectrum.

Recall that the Laplacian matrix of a graph is given by $\mathcal{L} = D - A$, where D is the degree matrix and A is the adjacency matrix. On the other hand, if the graph is regard as a 1-dimensional simplicial complex, and denote the matrix representing the one-dimensional boundary operator as B_1, we observe that the Laplacian matrix of the graph can be precisely expressed as $\mathcal{L} = B_1 B_1^T$. This inspires the generalization of the Laplacian operator to higher dimensions using the boundary operator, leading to the Laplacian operator on simplicial complexes. Let K be a simplicial complex, and let B_k be the representation matrix of its k-dimensional boundary operator. The Laplacian matrix is defined as

$$\mathcal{L}_k = B_{k+1} B_{k+1}^T + B_k^T B_k. \tag{4}$$

Here, B_k^T denotes the transpose matrix of B_k. The term $B_k^T B_k$ indicates the connectivity arising from the intersections of k-simplices at $(k-1)$-simplices, while the term $B_{k+1}B_{k+1}^T$ implies the interactions resulting from the inclusions of k-simplices into $(k+1)$-simplices.

Recall that the topological information for simplicial complexes, hypergraphs, or hyperdigraphs is derived from their respective chain complexes. From now on, we will define the Laplacian operator starting from the perspective of chain complexes. Let Ω_* be a chain complex with the differential $\partial_k : \Omega_k \to \Omega_{k-1}$. Assume that, for each k, there is always an inner product structure on Ω_k. Consequently, the boundary operator ∂_k has its adjoint operator ∂_k^*. The *combinatorial Laplacian* $\Delta_k : \Omega_k \to \Omega_k$ is defined by

$$\Delta_k = \partial_{k+1} \circ \partial_{k+1}^* + \partial_k^* \circ \partial_k. \tag{5}$$

In particular, $\Delta_0 = \partial_1 \circ \partial_1^*$. For each k, choose a standard orthonormal basis for Ω_k, then representation matrix L_k of the Laplacian operator Δ_k with respect to the standard orthonormal basis is given by

$$\mathcal{L}_k = B_{k+1}B_{k+1}^T + B_k^T B_k, \tag{6}$$

where B_k is the representation matrix of boundary operator ∂_k by left multiplication [51]. This combinatorial Laplacian is a generalization of the graph Lapalcian, which is just a carve-out of the properties of graphs (i.e., 1-simplical complex). The combinatorial Laplacian, on the other hand, extends the analysis to higher dimensions. Its eigenvectors and eigenvalues encode geometric and topological information about the simplicial complex or hyperdigraph. Because the Laplacian matrix is positive semidefinite, all eigenvalues of the Laplacian matrix are non-negative. Particularly, the zero eigenvalues, i.e., the harmonic spectrum, encode the topological information. While the non-zero eigenvalues (the non-harmonic spectrum) encode the geometric information about the system. Figure 4**j** shows the L_0 nonzero minimum non-harmonic eigenvector embedding for the C_α atoms (i.e., 0-simplex in simplicial complex) of protein 6L9D at a cutoff distance of $d = 5$ Å. And Figure 4**k** shows the L_1 harmonic eigenvector embedding for the edges (i.e., 1-simplex in simplcial complex) between the C_α atoms of protein 6L9D at a cutoff distance of $d = 5$ Å. Specifically, for $\mathcal{L}_k$, the multiplicity of the zero eigenvalue (i.e., the number of times 0 appears as an eigenvalue) equals the number of independent components, it also equals the topological invariants (β_k) in the k-dimensional space [52]. For example, multiplicity of zero for $\mathcal{L}_0$ (i.e., β_0) is the number of connected components in the graph (1-simplcial complex), the multiplicity of zero for $\mathcal{L}_1$ (i.e., β_1) is the number of cycles, and it means the number of cavities for $\mathcal{L}_2$. The largest eigenvalue λ_k^{max} of $\mathcal{L}_k$ is less than or equal to the maximum number d_k of $k+1$-simplex shared one k-simplex (maximum degree of the graph for $\mathcal{L}_0$). Specifically, $0 \leq \lambda_k^{max} \leq 2d_k$. The smallest non-zero eigenvalue for $\mathcal{L}_k$, also know as spectral gap, denoted as λ_k^{min}, reflect the geometric structure of the system. In this work, the multiplicity of zero, the average value, the standard deviation, the minimum, the maximum, and the summation of the positive eigenvalue for $\mathcal{L}_0$ are used to embed the given topological Laplacians. In addition, to validate the power of topological hyperdigraph Laplacian, two $B_7C_2H_9$ isomers with identical geometric structures, differing only in the positions of carbon atoms are constructed in the validation, as shown in Figure 14. The findings indicate that the hyperdigraph Laplacian possesses the capacity to encode more information compared to standard Laplacians.

Persistent Laplacians Persistent Laplacians or multiscale topological Laplacians, were introduced in a series of papers on a differential manifold setting [53] and a discrete point cloud setting [18, 54] in 2019. A filtration process is essential to achieving the multiscale representation in persistent Laplacians [18, 20, 55] as well as in persistent homology [21, 56]. The choice of the filtration (scale) parameter, denoted as d, varies based on the data structure in question: for point cloud data (Figure 4**a**), it is often the sphere radius (or diameter). By systematically adjusting d, one can derive a sequence of hierarchical representations, illustrated in Figure 1**a**. Notably, these representations are not limited to simplicial complexes, but can also be realized with hyperdigraphs. As an example, consider a filtration operation applied to a distance matrix, where the matrix elements represent distances between vertices. One could define a cutoff value as the scale parameter; if the distance between two vertices falls below this cutoff, they are connected. By progressively increasing this cutoff, one obtains a sequence of nested graphs. Each graph in this sequence, derived from a smaller cutoff value, is a subset of the graph generated with a higher cutoff.

In a similar vein, nested simplicial complexes can be formed based on different complex definitions like the Vietoris-Rips complex, Čech complex, and alpha complex. The Vietoris-Rips complex is used in this work. Mathematically, the nested simplicial complexes can be written as:

$$\emptyset \subseteq K_{d_0} \subseteq K_{d_1} \subseteq \cdots \subseteq K_{d_n} = K \tag{7}$$

Here, for any two $d_i < d_j$, we have $K_{d_i} \subseteq K_{d_j}$. The concept extends to hyperdigraphs as well, namely Vietoris-Rips hyperdigraph: one can form nested hyperdigraphs by properly defining directed hyperedges [20]. To visualize the effects of changing filtration parameters, Figure 4**e** depicts alterations in point cloud connectivity from Figure 4**a**, leading to a sequence of hyperdigraphs. Additionally, Figure 11 showcases simplicial complex produced at different filtration parameters and Figure 12 illustrates hyperdigraphs generated at different filtration parameters. The details about the construction of Vietoris-Rips hyperdigraph can be seen in Figure 5. In addition, inspired by the alpha complex, the alpha hyperdigraph is also introduced in this work, as shown in Figure 6.

As a filtration process unfolds, it naturally gives rise to a family of chain complexes. For each filtration step d_i (with i indexing the steps), a chain complex $C(K_{d_i}; G)$ is constructed. Mathematically, a chain complex for a particular filtration step is a sequence of Abelian groups (or modules) and boundary homomorphisms:

$$\cdots \to C_{k+1}(K_{d_i}; G) \xrightarrow{\partial_{k+1}^{d_i}} C_k(K_{d_i}; G) \xrightarrow{\partial_k^{d_i}} C_{k-1}(K_{d_i}; G) \to \cdots \tag{8}$$

where $C_k(K_{d_i}; G)$ is the k-dimensional chain group at filtration step d_i.

For a more general exposition, we now introduce the Laplacian in a mathematical formalism. For real numbers $a \leq b$, let Ω_*^a and Ω_*^b be chain complexes. Suppose that $\Omega_*^a \subseteq \Omega_*^b$. The chain complexes considered can be the chain complexes obtain from a filtration of simplicial complexes, hypergraphs, or hyperdigraphs, among other possibilities. Moreover, the chain complexes Ω_*^a and Ω_*^b are endowed with the compatible inner product structures. Let $\Omega_{k+1}^{a,b} = \{x \in \Omega_{k+1}^b | \partial_{k+1}^b x \in \Omega_k^a\}$.

The persistent boundary operator $\partial_{k+1}^{a,b} : \Omega_{k+1}^{a,b} \to \Omega_k^a$ is defined by $\partial_{k+1}^{a,b} x = \partial_{k+1}^b x$ for $x \in \Omega_{k+1}^{a,b}$.

$$\begin{array}{ccccc} \Omega_{k+1}^a & \xrightarrow{\partial_{k+1}^a} & \Omega_k^a & \underset{(\partial_k^a)^*}{\overset{\partial_k^a}{\rightleftarrows}} & \Omega_{k-1}^a \\ \downarrow & \overset{\partial_{k+1}^{a,b}}{\nearrow} \; \underset{(\partial_{k+1}^{a,b})^*}{\swarrow} & \downarrow & & \downarrow \\ & \Omega_{k+1}^{a,b} & & & \\ \downarrow & \searrow & \downarrow & & \downarrow \\ \Omega_{k+1}^b & \xrightarrow{\partial_{k+1}^b} & \Omega_k^b & \xrightarrow{\partial_k^b} & \Omega_{k-1}^b \end{array} \tag{9}$$

The *k-th persistent Laplacian* is defined as

$$\Delta_k^{a,b} = \partial_{k+1}^{a,b} \circ (\partial_{k+1}^{a,b})^* + (\partial_k^a)^* \circ \partial_k^a. \tag{10}$$

It is worth noting that the harmonic part of $\Delta_k^{a,b}$, i.e., $\ker \Delta_k^{a,b}$, is naturally isomorphic to the (a, b)-persistent homology $H_k^{a,b} = \mathrm{im}(H_k(\Omega_*^a) \to H_k(\Omega^b)_*)$ [57]. In a broad sense, the harmonic part of the persistent Laplacian contains information about persistent homology. To glean insights from each chain complex, one can resort to spectrum analysis. By constructing the Lapalcian matrices corresponding to each ∂_k and ∂_{k+1} and examining their spectra (eigenvalues and eigenvectors), one can uncover rich structural information about the topological and geometric properties inherent in the data at that particular scale of the filtration. This spectral information often provides a compact and informative summary of the data, allowing for efficient comparison and analysis across different scales. Figure 4**d** illustrates the evolution of zero eigenvalue multiplicities in the associated Laplacian matrix as the filtration (scale) parameters change, while Figure 4**f** depicts the variation in the minimum positive eigenvalue with changing filtration (scale) parameters. Additional persistent attributes are presented in Figure 10.

Element-specific embedding In this work, the topological embedding method is applied to encoding the protein-ligand complex. An accurate prediction requires a better representation of the interactions between proteins and ligands at the molecular level. Here, the element-specific topological embedding [16] is used to characterize protein-ligand interactions.

When analyzing ligands, the focus is on heavy elements such as carbon (C), nitrogen (N), oxygen (O), sulfur (S), phosphorus (P), fluoride (F), chloride (Cl), bromide (Br), and iodine (I). Conversely, for proteins, only carbon (C), nitrogen (N), oxygen (O), and sulfur (S) are considered. Subsequently, a range of element combinations, arranged in a specific sequence, will represent the interactions between the protein and the ligand. For proteins, the combinations are denoted as $\mathcal{E}_{\text{protein}}$={{C}, {N}, {O}, {S}, {C, N}, {C, O}, {C, S}, {N, O}, {N, S}, {O, S}, {C, N, O, S}}. Meanwhile, the ligand combinations are $\mathcal{E}_{\text{ligand}}$={{C}, {N}, {O}, {S}, {C, N}, {C, O}, {C, S}, {N, O}, {N, S}, {O, S}, {N, P}, {F, Cl, Br, I}, {C, O, N, S, F, P, Cl, Br, I}}. Within the Element-specific embedding approach, the interactions between proteins and ligands are defined by the topological links between two sets of atoms: one from the protein and the other from the ligand. For example, a representation like $K_{\{\text{C,N}\},\{\text{S}\}}$ indicates the topological hyperdigraph representation where the C and N atoms are derived from the protein, while the S atom comes from the ligand. The Element-specific embeddings detail interactions based on their spatial relationships. It can be

characterized by distance matrix D as follows,

$$D(i,j) = \begin{cases} \|\mathbf{r}_i - \mathbf{r}_j\|, & \text{if } \mathbf{r_i} \in \mathcal{E}_{\text{protein}}, \mathbf{r_j} \in \mathcal{E}_{\text{ligand}} \text{ or } \mathbf{r_i} \in \mathcal{E}_{\text{ligand}}, \mathbf{r_j} \in \mathcal{E}_{\text{protein}} \\ \infty, & \text{other} \end{cases} \tag{11}$$

where the $\mathbf{r}_i$ and $\mathbf{r}_j$ are coordinates for the ith and jth atoms in the set, and $\|\mathbf{r}_i - \mathbf{r}_j\|$ is their Euclidean distance. In the TopoFormer model, protein atoms located within 20 Å of ligand atoms are taken into account. For the TopoFormer$_s$ model, the range is reduced to protein atoms within 12 Å of the ligand atoms. In this study, emphasis is placed on the protein-ligand interactions by assigning an infinite value to the distance between atoms either within the protein or the ligand. For a specific protein-ligand complex, there are 143 potential combinations (derived from 11 protein sets multiplied by 13 ligand sets). Each of these combinations functions as a simplicial complex and is further examined using the persistent topological hyperdigraph Laplacian approach.

4.3 TopoFormer model

Model architecture. The TopoFormer model introduced in our work incorporates a Topological embedding model. This model transforms the 3D protein-ligand complex into a topological sequence characterized by topological features at various scales. Specifically, in the larger version of the TopoFormer model, the scale range extends from 0 Å to 10 Å in increments of 0.1 Å, resulting in a topological sequence of 100 units in length. At each filtration (scale) increment, the embedded features possesses a matrix of 143 by 6 (6 attributes associated with each $\mathcal{L}_0$). The combined outputs from the topological embedding module are obtained by summing the topological embeddings with the trainable multiscale embeddings, as depicted in Figure 1**a**. To convert the 143 by 6 matrix for every filtration increment into a 1-dimensional vector, we have incorporated a convolutional layer into both the Transformer's original encoder and decoder, as shown in Figure 1**c**. Subsequently, the conventional dot-product attention mechanism in the Transformer utilizes encoded representations of the input in the form of queries (Q), keys (K), and values (V) designated for each filtration increment. This attention can be mathematically represented as,

$$\text{Attention}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}\right)V \tag{12}$$

Here, the $\sqrt{d_k}$ is the scalar defined by the root of embedding dimension($d_k = 512$ in this work). The resulting bidirectional attention matrix is then derived from this formula. In addition, similar with the MAE model [58] in computer vision, an asymmetric design is applied for TopoFormer's encoder and decoder. Detailed settings of the TopoFormer are provided in Supplementary Information Section SA.2. The training process of the model encompasses two phases: initially, self-supervised learning is applied to unlabeled data to obtain a pre-trained model. Subsequently, supervised learning is employed on specific benchmarks tailored to various tasks, resulting in a fine-tuned model.

Self-supervised and supervised learning in TopoFormer. In this study, we utilized 19,513 unlabeled protein-ligand complexes from the PDBbind database for the pretraining of TopoFormer. The topological embeddings derived from these complexes were reconstructed and subsequently employed to compute the reconstruction loss. For this purpose, the mean square error (MAE) was

adopted as the metric for reconstruction loss. This self-supervised approach enables the model to discern deep, generalized representations of protein-ligand complex patterns using a vast amount of unlabeled data. Such an approach potentially simplifies the downstream fine-tuning process. In this study, a dataset of only nearly 20,000 unlabeled complexes yielded exceptional performance across most tasks. Moving forward, we envisage incorporating even more protein-ligand complexes into the pretraining workflow, without the necessity for experimental data. In this study, all tasks encompassing scoring, ranking, docking, and screening involve fine-tuning the TopoFormer model to predict a specific score for a given protein-ligand complex. Consequently, the mean square error was selected as the loss function for these tasks.

Data availability

The training dataset employed in this study comprises a comprehensive collection of protein-ligand complexes sourced from various PDBbind databases, specifically CASF-2007, CASF-2013, CASF-2016, and PDBbind v2020. To ensure the dataset's reliability and eliminate redundancies, a meticulous curation process was undertaken, resulting in a total of 19,513 non-overlapping complexes. And all data used in this study can be downloaded from the official PDBbind website: http://www.pdbbind.org.cn/index.php. Additionally, the topological embedded features utilized in both TopoFormer and TopoFormer$_s$, as well as the sequence-based features derived from the Transformer-CPZ model [22] and the ESM model [31], are readily available for download at https://github.com/WeilabMSU/TopoFormer. The additional generated poses and their associated scores, which were instrumental in the docking and screening tasks, can be obtained from the following source: https://weilab.math.msu.edu/AGL-Score.

Code availability

All source codes and models are publicly available at https://github.com/WeilabMSU/TopoFormer

Acknowledgments

This work was supported in part by NIH grants R01GM126189, R01AI164266, and R35GM148196, National Science Foundation grants DMS2052983 and IIS-1900473, Michigan State University Research Foundation, and Bristol-Myers Squibb 65109.

A Supplementary Information

This document provides additional details not essential to the main body of the paper but potentially of interest to readers.

A.1 Evaluation metrics

Evaluation of scoring power. In this study, the Pearson correlation coefficient (PCC) is used in the evaluation of scoring power, and it is defined as below:

$$\mathrm{PCC} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum (x_i - \bar{x})^2 \sum (y_i - \bar{y})^2}} \tag{13}$$

where x_i is the value of the x variable in ith sample, $\bar{x}$ is the mean of the values of the x variable, y_i is the value of the y variable in the ith sample, $\bar{y}$ is mean of the values of the y variable. The Pearson correlation coefficient (PCC) explains the relationship between the x variable and y variable.

The root mean squared error (RMSE) is defined as below:

$$\mathrm{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2} \tag{14}$$

where y_i and $\hat{y}_i$ are predicted value and true value of ith sample respectively.

Evaluation of ranking power. In this work, two evaluative approaches are employed: the high-level and the low-level success measurements. In the high-level success metric, the objective is to perfectly rank the binding affinities of the complexes within each cluster. Conversely, the low-level success criterion requires the scoring function to merely identify the complex with the pinnacle binding affinity. The assessment of ranking efficacy termed "ranking power" is gauged by the proportion of correctly identified affinities across a specified benchmark.

Let us denote the set of protein-ligand complexes in a given cluster as C, and let A_i be the binding affinity of the i^{th} complex in C, where smaller i indicates smaller binding affinity. For a cluster C with n complexes ($n = 3$ for benchmarks CASF-2007 and CASF-2013, $n = 5$ for CASF-2016): The scoring function f is successful in the sense of high-level if and only if:

$$f(A_i) \geq f(A_j), \text{if } i \geq j \text{ and } A_i, A_j \in C \tag{15}$$

The low-level success measurement is defined as:

$$f(A_{max}) > f(A_i), \forall A_i \in C \tag{16}$$

where A_{max} is the complex with the highest binding affinity in C.

The "Ranking Power" of a scoring function across a benchmark can then be calculated as:

$$\text{Ranking Power} = \frac{\text{Number of successful clusters by the function}}{\text{Total number of clusters in the benchmark}} \times 100\% \tag{17}$$

This provides a percentage-based assessment of how often the scoring function correctly ranks the binding affinities within the given clusters. Notably, the current iteration of the ranking power

metric can be optimized. Presently, it is restricted to determining the accurate binding affinity sequence of three native ligands for each target receptor in the core set. This might not adequately mirror the complexities of an authentic virtual screening scenario where a multitude of ligands might vie for the same target receptor. Incorporating more comprehensive evaluation metrics, like Kendall's tau or the Spearman correlation coefficient, could enhance accuracy. In our study, the results of high level success measurement are shown in Figure S 8, and the low level success measurement are shown in Figure S 9.

Evaluation of docking power. The present assessment evaluates a scoring function's proficiency in distinguishing the "native" pose from an array of poses generated by docking software. Within the benchmark parameters, a pose is deemed "native" if its root-mean-square deviation (RMSD) relative to the genuine binding pose is less than 2 Å. To ensure alignment with prior research, we have anchored our validation efforts to both the CASF-2007 and CASF-2013 datasets, adhering to training and test sets as delineated in the extant literature [27, 25, 24]. In the CASF-2007 benchmark, each ligand was supplied with 100 distinctive poses, all generated using specific docking software packages. Meanwhile, the CASF-2013 benchmark produced 100 poses for each ligand, courtesy of three eminent docking applications: GOLD v5.1, Surflex-Dock (integrated within SYBYL v8.1), and MOE v2011. For researchers seeking accessibility, the curated poses can be procured from https://weilab.math.msu.edu/AGL-Score/. It is worth noting that in both benchmarks, owing to structures that exhibit certain symmetries, a given ligand may possess multiple "native" poses within the dataset. As such, if a method successfully discerns any of these native poses, it is adjudged as successful for that ligand. The ultimate measure of efficacy, termed "docking power", is gauged by the tally of ligands for which "native" poses are accurately pinpointed. It can be calculated as:

$$\text{Docking Power} = \frac{\text{Number of complexes that successfully identified "native" poses}}{\text{Total number of complexes in the benchmark}} \times 100\% \quad (18)$$

In the docking task, the Root Mean Square Deviation (RMSD) is a measure used to assess the similarity between the predicted or generated molecular structure (usually a ligand) and a reference structure (often the experimentally determined or known structure). RMSD is often used to evaluate the accuracy of how a docking program predicts the binding mode of a ligand within a protein's binding site. It is defined as:

$$\text{RMSD} = \sqrt{\frac{\sum_i (A_i - B_i)^2}{n}}, \quad (19)$$

where A_i is the coordinates of the i-th atom in the docked structure, B_i is the coordinate of the i-th atom in the experimental structure, and n means the total number of atoms being compared in both structures. RMSD help to determine how well a docking program can reproduce known binding poses or predict the binding mode of a ligand within a protein's active site. Lower RMSD values are generally desirable, indicating more accurate predictions.

Evaluation of screening power. There are two kinds of screening power measurements. The first one the enrichment factor in protein-ligand screening, which is often used in the field of computational chemistry and drug discovery. The enrichment factor (EF) is a measure of how

effectively a virtual screening or docking method enriches active or potent compounds (ligands) within a larger library of compounds. It is used to assess the performance of these methods in identifying potential drug candidates. The enrichment factor is typically calculated as follows

$$\text{EF} = \frac{\text{Number of true positives}}{\text{Number of total hits}} \cdot \frac{\text{Total number of compounds}}{\text{Number of active compounds}}, \tag{20}$$

where the "Number of true positives" is the number of active compounds correctly identified as hits by the TopoFormer model. Number of total hits is the total number of compounds identified as hits by the screening method. Number of active compounds means the total number of active or potent compounds in the entire test set. The total number of compounds means the total number of compounds in the test set. The objective of the second screening power measurement is to pinpoint the optimal true binder. The success rate is determined by the $x\%$ of the top-ranked candidates, among which the best binders from a pool of 65 receptors are discovered.

Loss function for training. In this work, the mean squared error (MSE) is applied as the loss function during the pre-training and fine-tuning stages. MSE is a widely used metric to quantify the difference between predicted and actual values in statistical modeling and machine learning. It measures the average squared differences between predictions and actual observations. The mathematical definition of MSE is:

$$\text{MSE} = \frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2 \tag{21}$$

where y_i is the actual value for the i-th data point. $\hat{y}_i$ is the predicted value for the i-th data point. And n is the total number of data points.

A.2 Hyperparameter selection and optimization

In the Seq-ML approach, we employ the Gradient Boosted Decision Trees (GBDT) algorithm to predict protein-ligand binding affinity. The parameters are set as follows: 'n_estimators' to 10,000, 'max_depth' to 7, 'min_samples_split' to 2, with a subsample size of 0.4, and a learning rate set at 0.005. All other parameters retain their default values as defined in the algorithm [59]. For the classification task within the screening process, the GBDT parameters remain consistent with those described for the regression task.

In the TopoFormer models, we utilize a self-supervised learning approach in the pre-training phase, followed by a supervised learning strategy during the fine-tuning stage. Relevant parameters are detailed in Table 3.

In the pre-training stage of our proposed model, we diligently selected training hyperparameters to encourage robust learning and convergence. A batch size of 64 and a maximum of 30,000 training steps were utilized, alongside an initial learning rate of 0.001, ensuring a smooth and steady journey towards optimal weight adjustments. A warm-up period, comprising 5% of the maximum training steps, was integrated to stabilize the initial training phase, incrementally increasing the learning rate from 0 to the set initial rate. Following this, the fine-tuning stage implemented a supervised learning strategy, ensuring task-specific model refinement without overfitting the hyperparameters. The batch size was reduced to 32 and the initial learning rate was slightly diminished to 0.0008. Distinct

Table 3: The parameter settings for TopoFormer

Parameters	Pre-training stage	Finetuning stage
attention_probs_dropout_prob	0.1	0.1
decoder_hidden_size	768	/
decoder_intermediate_size	3072	/
decoder_num_attention_heads	12	/
decoder_num_hidden_layers	8	/
hidden_act	gelu	gelu
hidden_dropout_prob	0.1	0.1
hidden_size	1024	1024
image_size(large)	(100, 143)	(100, 143)
image_size(small)	(50, 143)	(50, 143)
initializer_range	0.02	0.02
intermediate_size	4096	4096
num_attention_heads	4096	4096
num_channels	6	6
num_hidden_layers	12	12
patch_size	(1, 143)	(1, 143)

maximum training steps were employed for varying tasks: 10,000 steps for scoring tasks, and a more succinct 5,000 steps for both the docking and screening tasks. It is noteworthy that specific parameters, such as the warm-up steps and optimizer, were consistently held across both pre-training and fine-tuning stages, ensuring a coherent model development. Furthermore, for the fine-tuning of the scoring task, additional parameter combinations proximate to the pre-defined settings were tested to validate the robustness of the proposed model. Specifically, combinations of batch size 64 with a learning rate of 0.0008, and batch size 32 with a learning rate of 0.001 were examined. The results, delineated in Table 4, reveal closely tied performances across the different settings, underscoring the model's stability and robustness amidst variations in the hyperparameters.

A.3 Topological objects

Graph. Graph is the most fundamental object for describing relationships among entities and is one of the most common data types. It consists of nodes and edges, capturing the relationships between nodes. Common extensions of graphs include directed graphs, weighted graphs, and geometric graphs, among others. These graph-based models often provide an effective representation of relationships and characteristics within various contexts. Strictly speaking, a *graph* is a pair (V, E), where V is a vertex set and $E \subseteq V \times V$ is the edge set. Vertices and edges are the fundamental objects of a graph. Various tools are employed to characterize the relationships between points and edges, such as adjacency matrices, degree matrices, and Laplacian matrices. These matrices play a crucial role in graph theory and network analysis, effectively capturing the topological structure of the graph. Given that a graph inherently has a 1-dimensional structure, certain models from simplicial complexes are also employed to capture the higher-dimensional structures of the graph. Examples include the clique complex, neighborhood complex, and Hom complex [60, 61].

Simplicial complex. A simplicial complex is a topological space that is built up from simple pieces called simplices. A simplex is a generalization of the concept of a triangle or tetrahedron to arbitrary dimensions. Given a vertex set V, a k-simplex σ is often represented by a $(k+1)$-element subset of vertices in V, denoted as $\sigma = \langle v_0, v_1, \ldots, v_k \rangle$. And a subset of σ is a face of σ.

A *simplicial complex* K on a vertex set V is a collection of simplices satisfying the following two conditions: (1) If a simplex σ is in K, then so is each face of σ, including the individual vertices; (2) The intersection of any two simplices in K is either an empty set or a face (subset) of both simplices. Using the above properties, it is clear that a graph can be viewed as a 1-dimensional simplicial complex, as its simplices are its vertices (0-simplices) and edges (1-simplices).

For a given k-simplex, the boundary is essentially the collection of its $(k-1)$-dimensional faces. Mathematically, the *boundary operator*, denoted by ∂_k, acts on a k-simplex $\langle v_0, v_1, \ldots, v_k \rangle$ as:

$$\partial_k \langle v_0, v_1, \ldots, v_k \rangle = \sum_{i=0}^{k} (-1)^i \langle v_0, \ldots, \widehat{v_i}, \ldots, v_k \rangle, \tag{22}$$

where $\widehat{v_i}$ means that vertex v_i is omitted. A chain complex is a sequence of Abelian groups (or modules) connected by boundary operators. Let G be an abelian group. The k-th group, denoted as $C_k(K; G)$, in the chain complex consists of formal sums of k-simplices, and the boundary operator $\partial_k : C_k(K; G) \to C_{k-1}(K; G)$ maps a k-simplex to its $(k-1)$-dimensional boundary. The chain complex can be represented as a sequence like this:

$$\cdots \xrightarrow{\partial_{k+1}} C_k(K; G) \xrightarrow{\partial_k} C_{k-1}(K; G) \xrightarrow{\partial_{k-1}} \cdots \xrightarrow{\partial_2} C_1(K; G) \xrightarrow{\partial_1} C_0(K; G). \tag{23}$$

An essential property of the boundary operator is that the composition of two successive boundary operators is zero, i.e., $\partial_{k-1} \circ \partial_k = 0$. It means that the boundary of a boundary is always zero, which has topological implications. The chain complex structure provides a framework to understand how the boundaries fit together.

While simplicial complexes serve as topological models to depict relationships in most data, there are instances where they remain somewhat restrictive. In such cases, topological hypergraphs, as a more general model and combinatorial object, exhibit significant potential in applications.

Topological hypergraph. Topological hypergraph, as a relatively new combinatorial object, can be considered as a generalization of the concepts of graphs and simplicial complexes. From a graph perspective, topological hypergraphs can be seen as an extension of edges in graphs, where edges are not limited to pairs of vertices but can include multiple vertices. From a simplicial complex perspective, topological hypergraphs can be viewed as relaxing the condition that the faces of simplices must be simplices.

A *topological hypergraph* $\mathcal{H}$ on a vertex V is a collection of subsets of V. The $(k+1)$-element subsets of V are the *k-hyperedges*. The *simplicial closure* of a topologcal hypergraph $\mathcal{H}$ is given by

$$\Delta\mathcal{H} = \{\sigma | \sigma \subseteq \tau \text{for some hyperedge } \tau \in \mathcal{H}\}. \tag{24}$$

The simplicial $\Delta\mathcal{H}$ closure is the minimal simplicial complex containing $\mathcal{H}$. In light of the close connection between topological hypergraphs and simplicial complexes, topological hypergraphs can always be constructed based on simplicial complexes, which inspires the study of the topological structures of topological hypergraphs. Recently, embedded homology for topological hypergraphs has been introduced to investigate their topological features [62]. Let $D_k(\mathcal{H}; G)$ be the abelian group generated by the k-hyperedges. Then $D_*(\mathcal{H}; G)$ is a graded subgroup of the chain complex $C_*(\Delta\mathcal{H}; G)$ of the simplicial complex $\Delta\mathcal{H}$. Thus, one can obtain the infimum complex

$$\mathrm{Inf}_*(\mathcal{H}; G) = \{x \in D_*(\mathcal{H}; G) | \partial x \in D_*(\mathcal{H}; G)\}. \tag{25}$$

Here, ∂ is the boundary operator on $C_*(\Delta\mathcal{H}; G)$. The name "infimum complex" primarily stems from the fact that $\mathrm{Inf}_*(\mathcal{H}; G)$ is the minimal sub chain complex of $C_*(\Delta\mathcal{H}; G)$ containing $D_*(\mathcal{H}; G)$. The topological information on hypergraphs is based on the infimum complex $\mathrm{Inf}_*(\mathcal{H}; G)$.

Topological hypergraphs have become a very general research object for studying interactions in complex systems. However, when exploring complex systems and structures involving directional and asymmetric relationships, topological hypergraphs may not be sufficiently inclusive. In such cases, topological hyperdigraphs, as objects incorporating higher-dimensional structures, multi-faceted interactions, and directional information, become our new focus.

A.4 Vietoris-Rips hyperdigraph and alpha hyperdigraph

The Vietoris-Rips (VR) hyperdigraph is constructed based on the VR complex. Let (M, d) be a metric space. Let X be a finite point set in M. For a given parameter d, the VR complex $\mathcal{VR}_d$ is defined by

$$\mathcal{VR}_d = \{S \subseteq X | \text{every two points } x, y \text{ in } S \text{ has the distance } d(x, y) \leq d\}. \tag{26}$$

The VR complex is always regarded as an abstract simplicial complex; that is, a simplex S is considered only as a set, without considering its geometric structure. This provides us with the motivation to study more general structures. If we take into account geometric properties such as angles, volumes, or even their manifestations in biology or materials, then the hyperdigraph becomes a more versatile topological model. Specifically, for any simplex S, we assign to it both weight information and orientation information. Mathematically, for a VR complex $\mathcal{VR}_d$, there is a weight function $w : \mathcal{VR}_d \to \mathbb{R}$ and a graded orientation function $\varrho_n : (\mathcal{VR}_d)_n \to S_{n+1}$ for

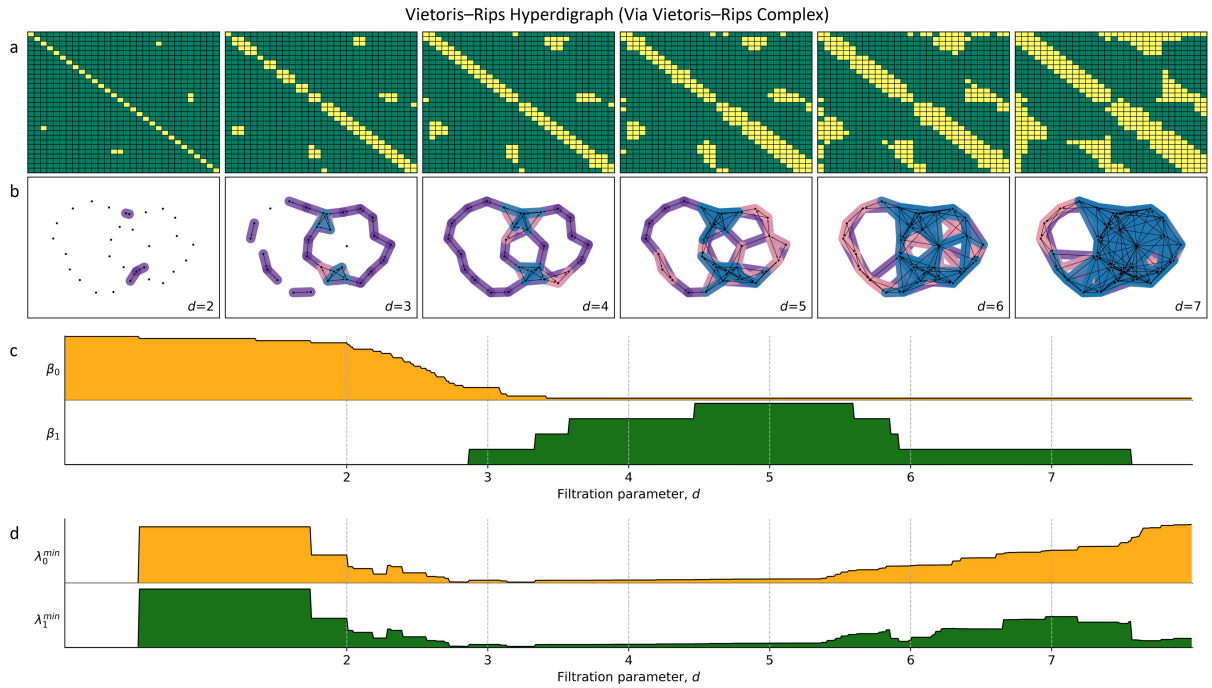

Figure 5: Illustration of Vietoris-Rips hyperdigraph construction with over scales for the point cloud in Figure 4**a**. **a** Illustration of the adjacency matrices of the point cloud at various scales (i.e., filtration parameter d values). Yellow entries in the matrices represent connections between points with distances smaller than the threshold, while green entries indicate points that are not connected. **b** The constructed Vietoris-Rips hyperdigraphs at various scales, including $d = 2$, $d = 3$, $d = 4$, $d = 5$, $d = 6$, and $d = 7$. **c** A display of the persistent Betti numbers, denoted as β_i with $i = 0$ and $i = 1$. Vertical dashed lines mark the Betti numbers corresponding to specific scales. **d** The nonzero minimum non-harmonic spectra of the persistent topological hyperdigraph Laplacian at the 0th and 1st dimensions (λ_0^{min} and λ_1^{min}), highlighting their dependence on the scale parameter d.

$n \geq 1$. Here, S_n is the permutation group of n elements. Then for any $\eta \in \mathbb{R}$, the *Vietoris-Rips hyperdigraph* is defined by

$$\mathcal{VR}_d\vec{\mathcal{H}}_\eta := \{S \times \varrho_*(S) | w(S) \leq \eta, S \in \mathcal{VR}_d\}. \tag{27}$$

Note that there is a one-one corresponding between the sequences and the permutation group for a fixed length [20]. Thus the element $S \times \varrho_*(S)$ is essentially a sequence. The homology and Laplacians of hyperdigraphs can be computed to detect topological and geometric features of point set. In this work, the weight function $w : \mathcal{VR}_d \to \mathbb{R}$ and the graded orientation function $\varrho_n : (\mathcal{VR}_d)_n \to S_{n+1}$ are taken to be the trivial functions. Consequently, the hyperdigraph construction can be simplified to coincide with the Vietoris-Rips complex.

Informally, a Vietoris-Rips (VR) complex is a simplicial complex whose simplices are formed by finite sets of points, with the condition that any two points in the set are no larger than a specified threshold parameter, known as the filtration parameter. Subsequently, directed hyperedges are constructed on all simplices in the complex. In this process, we take into account the geometric information of the simplices, considering both their direction and magnitude. By selecting certain simplices and endowing them with orientations, we form a collection of selected simplices, constituting a hyperdigraph. If the direction for each directed hyperedge is determined by a pre-

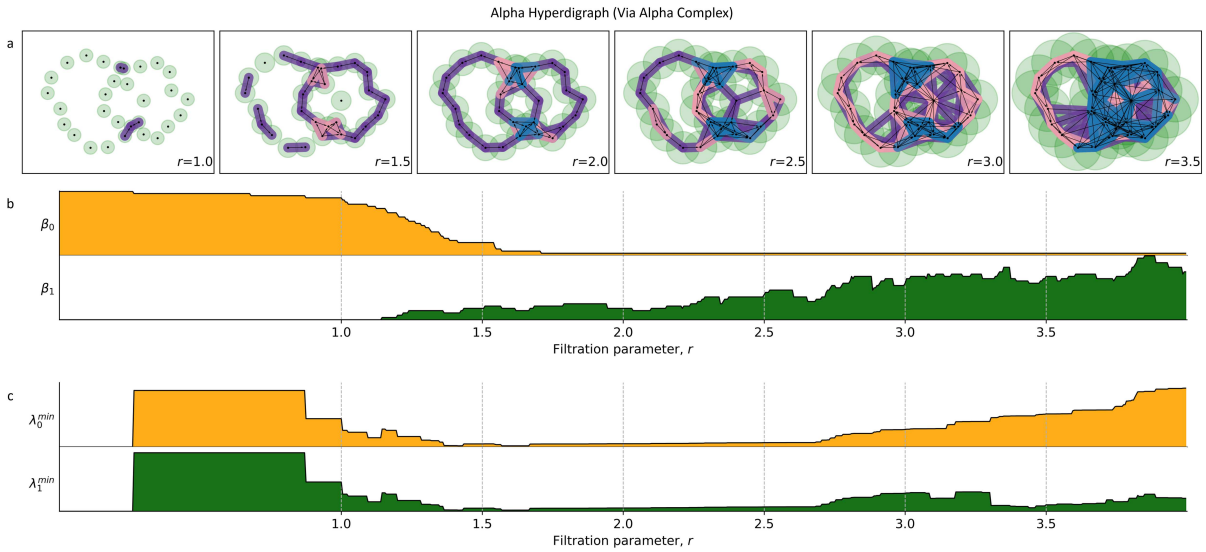

Figure 6: Illustration of the construction of alpha hyperdigraph via alpha complex with changing scale parameter for the point cloud in Figure 4**a**. **a** The constructed alpha hyperdigraphs for given scale parameter, i.e., $r = 1.0$, $r = 1.5$, $d = 2.0$, $d = 2.5$, $d = 3.0$, and $d = 3.5$. **c** The persistent Betti numbers of alpha hyperdigraphs, $\beta_i, i = 0, 1$. The verdical desh lines indicate the Betti numbers for given scale (filtration) parameters. **d** Representation of nonzero minimum non-harmonic spectra of the persistent topological hyperdigraph Laplacian at the 0th and 1st dimensions (λ_0^{min} and λ_1^{min}) for alpha hyperdigraph, highlighting their dependence on the filtration (scale) parameter d.

defined order of the points, the resulting topological hyperdigraph is constructed as a subset of the collection of these directed hyperedges. In this scenario, the VR hyperdigraph can be reduced to a hypergraph. Additionally, if we disregard the magnitude information of simplices, meaning all simplices are selected, then the VR hyperdigraph coincides with the VR complex. In this case, we denote the VR hyperdigraph by $\mathcal{VR}\vec{\mathcal{H}}_d(X)$.

The VR hyperdigraph $\mathcal{VR}\vec{\mathcal{H}}_d(X)$ captures the topological features of the underlying space at a scale determined by the parameter d. As d increases, as shown in Figure 5**a** and **b**, more simplices are added, resulting in the construction of more directed hyperedges. This provides information about the connectivity and holes in the space at different scales.

Similarly, the alpha hyperdigraph can be constructed on the alpha complex. For a metric space (M, d), let X be a finite point set in M. For a given parameter r, the alpha complex $\mathcal{A}_r$ is defined by

$$\mathcal{A}_r = \{S \subseteq X | \text{there is a disk of radius } r \text{ that covers } S\}. \tag{28}$$

Usually, people tend to regard the alpha complex as an abstract simplicial complex for computing its homology. However, the alpha complex itself possesses geometric structure. Considering functions $w : \mathcal{A}_r \to \mathbb{R}$ and $\varrho_n : (\mathcal{A}_r)_n \to S_{n+1}$, for any real number η, we can obtain the alpha hyperdigraph:

$$\mathcal{A}_r\vec{\mathcal{H}}_\eta = \{S \times \varrho_*(S) | w(S) \leq \eta, S \in \mathcal{A}_r\}. \tag{29}$$

Similar to the relationship between the alpha complex and the VR complex, alpha hyperdigraphs can capture distinct information compared to VR hyperdigraphs. If the maps $w : \mathcal{A}_r \to \mathbb{R}$ and $\varrho_n : (\mathcal{A}_r)_n \to S_{n+1}$ are chosen as trivial maps, the alpha hyperdigraph can also be reduced to the alpha complex.

In informal terms, the alpha complex involves the simplices that are close enough, meaning that one can find a disk of a given radius containing all the points in the simplex. The construction can also be derived from the 3-dimensional Voronoi diagram [63] or the Delaunay triangulation [64]. The alpha hyperdigraph is a collection of simplices in the alpha complex which are endowed with the corresponding orientation. If the orientation is chosen to follow a given order, the alpha hyperdigraph can be reduced to a hypergraph. Besides, if we choose all the simplices in the alpha complex as the collection, the alpha hyperdigraph can also be reduced to the alpha complex. In such case, the construction is denoted by $\mathcal{A}_r\vec{\mathcal{H}}(X)$.

For a given parameter $r > 0$, the alpha complex and alpha hyperdigraph, denoted as $\mathcal{A}_r(X)$ and $\mathcal{A}_r\vec{\mathcal{H}}(X)$, are constructed step by step as follows: 1. Include a vertex for each point in X. 2. For each subset S of X such that the maximum pairwise distance between points in S is less than or equal to r, include a simplex in the complex with vertices corresponding to the points in S. 3. Directed hyperedges are defined on all simplices using the predefined order of the set X, and the hyperdigraph $\mathcal{A}_r\vec{\mathcal{H}}(X)$ is then generated as the collection of these directed hyperedges.

The alpha hyperdigraph includes directed hyperedges for subsets of points that are in close proximity within the specified radius. As r increases, as shown in Figure 6**a**, more directed hyperedges are added to the hyperdigraph, capturing different levels of connectivity and features in the dataset. As illustrated in Figures 5**c**, **d** and 6**b**, **c**, the persistent attributes in higher dimensions of VR hyperdigraph Laplacians and alpha hyperdigraph Laplacians exhibit notable differences. However, for the 0-dimensional information, their persistent patterns remain the same.

Figures 5 and 6 illustrate the construction of the VR hyperdigraph and alpha hyperdigraph, respectively, with varying filtration parameters. Notably, both of these hyperdigraphs in this study are constructed based on the simplicial complex, incorporating the Vietoris-Rips (VR) complex and the alpha complex.

A.5 Supplementary tables

In the following section, we provide supplementary tables that offer additional data and insights pertinent to our study. Readers are encouraged to refer to these tables for a more detailed exploration of the topics covered in the main text.

In the finetuning stage of the Transformer model in TopoFormer-seq, Table 4 outlines three sets of hyperparameters, while keeping other settings constant at 10,000 training steps. According to the table, the optimal performance on the CASF-2007 dataset is achieved by TopoFormer$_s$-seq, with a batch size of 32 and a learning rate of 0.0008 during the finetuning stage. However, the performance remains nearly identical for the other two hyperparameter settings. For the CASF-2013 dataset, the best performance is observed with TopoFormer-seq, employing a batch size of 32 and a learning rate of 0.0008, resulting in a PCC of 0.816 and an RMSE of 1.367. Remarkably, even the least favorable hyperparameter setting, with a batch size of 32 and a learning rate of 0.001, yields a PCC of 0.815 and an RMSE of 1.373, a result very close to the optimal performance. For the CASF-2016 dataset, the superior performance is achieved by TopoFormer-seq with a batch size of 32 and a learning rate of 0.0008. All other hyperparameters result in the same PCC (0.864), albeit with slightly higher RMSE values of 1.160 and 1.157. The results indicate that the TopoFormer-seq and TopoFormer$_s$-seq models exhibit remarkable stability across different hyperparameter settings. To mitigate overfitting, a batch size of 32 and a learning rate of 0.0008 are consistently employed

Table 4: The PCCs and RMSEs of our TopoFormer-seq and TopoFormer$_s$-seq models on the three benchmarks of CASF-2007, CASF-2013, and CASF-2016 with different hyperparameter settings. The average of 400 experiments are reported in the table.

		TopoFormer-seq		TopoFormer$_s$-seq	
Hyperparameters	Datasets	PCC	RMSE	PCC	RMSE
Batch size 32 Learning rate 0.0008	CASF-2007	0.836	1.329	0.839	1.322
	CASF-2013	0.816	1.367	0.810	1.392
	CASF-2016	0.864	1.153	0.855	1.183
Batch size 64 Learning rate 0.0008	CASF-2007	0.837	1.333	0.837	1.335
	CASF-2013	0.816	1.373	0.812	1.387
	CASF-2016	0.864	1.160	0.858	1.184
Batch size 32 Learning rate 0.001	CASF-2007	0.837	1.331	0.838	1.329
	CASF-2013	0.815	1.373	0.810	1.390
	CASF-2016	0.864	1.157	0.856	1.183

for all scoring tasks in this study during validation and comparisons in the main text.

Table 5: The performance of recently proposed models is assessed through the evaluation of their PCCs(RMSEs) using various training datasets. To convert to the unit kcal/mol, a conversion factor of 1.3633 should be multiplied with the RMSEs in the table. Footnote [a] indicates variations in test dataset sizes, involving the PDBbind-v2013 core set (N = 180) and the PDBbind-v2016 core set (N = 276). Footnote [b] signifies the utilization of the PDBbind-v2016 core set (N = 290) as the testing dataset. We conducted performance testing on the CASF-2007 dataset using a training set comprising 18,904 protein-ligand complexes from the v2020 general set, excluding all core sets. The results show the best performance, with a Pearson correlation coefficient (PCC) of 0.853 and a root mean square error (RMSE) of 1.295.

Model	Training set	Core set (CASF-2013)	Core set (CASF-2016)
Ligand-based [38]	PDBbind-v2018(11663)	0.780[a]	0.821[a]
graphDelta [32]	PDBbind-v2018(8766)		0.87(1.05)[b]
ECIF [33]	PDBbind-v2019(9299)		0.866(1.169)
OnionNet-2 [34]	PDBbind-v2019(>9000)	0.821(1.357)	0.864(1.164)
DeepAtom [35]	PDBbind-v2018(9383)		0.831(1.232)[b]
SE-OnionNet [36]	PDBbind-v2018(11663)	0.812(1.692)	0.83
Deep Fusion [37]	PDBbind-v2016(9226)		0.803(1.327)[b]
TopoFormer-Seq	PDBbind-v2020(18904)	0.832(1.301)	0.881(1.095) 0.883(1.086)[b]

The performance of recently proposed models is presented in Table 5. Due to variations in the training sets utilized, direct comparisons among these models are not fair. The majority of these models employ a general set (or preprocessed general set) for training to enhance their performance on benchmarks. In this study, we introduced the TopoFormer-Seq model, trained on the PDBbind-v2020 general set, with exclusion of the core sets used for evaluation from the training process. As demonstrated in Table 5, the TopoFormer-Seq consistently exhibits the best performance across all benchmarks. For CASF-2007, the model achieves a PCC of 0.853 with an RMSE of 1.295, and for CASF-2013, the PCC is 0.832 with an RMSE of 1.301. Similarly, for CASF-2016, the TopoFormer-

Seq attains a PCC of 0.881 with a corresponding RMSE of 1.095. The model's performance on PDBbind-v2016 is also assessed, with a PCC of 0.883 and an RMSE of 1.086. It is important to note that, in this work, the performance of TopoFormer-Seq-2020 (trained on PDBbind-v2020 general set) is solely utilized to showcase the capability of the proposed model. The reported best performance in the main text adheres to the standard pipeline.

A.6 Supplementary figures

In this section, we present a series of supplementary figures that further elucidate and complement the findings discussed in the main text. Readers are encouraged to consult these figures for a richer understanding and visual representation of the concepts and results introduced in the primary manuscript.

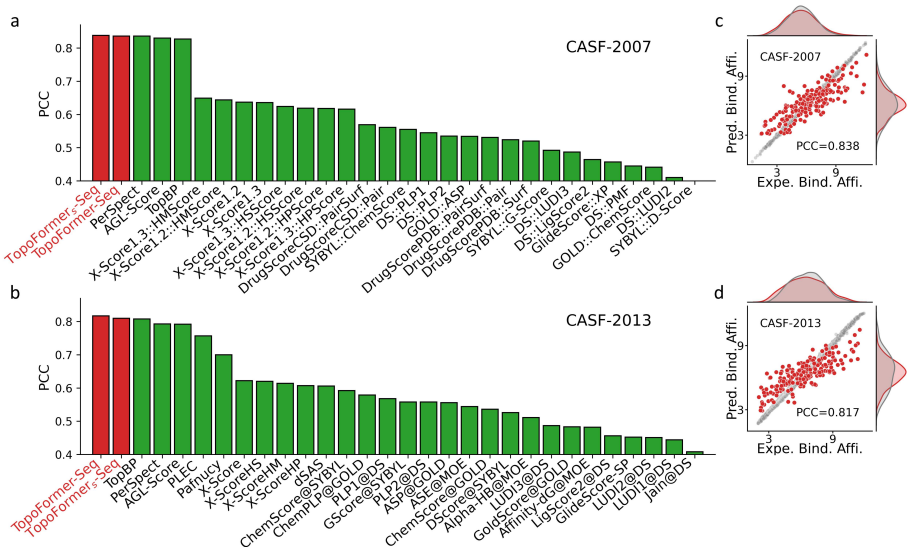

Figure 7: Performance of TopoFormer on scoring tasks for CASF-2007 and CASF-2013. **a**-**b** display the Pearson correlation coefficients (PCCs) of TopoFormers for protein-ligand complex binding affinity scoring. These are compared with models on the CASF-2007 (**a**), and CASF-2013 (**b**) benchmarks. The results from other methods are taking from refs ([25, 24, 26, 16, 27, 19, 28, 29, 30]), are in the green color. **c** to **d** showcase comparisons of predicted protein-ligand binding affinities and experimental results for the CASF-2007 (**c**, PCC=0.838), and CASF-2013 (**d**, PCC=0.817) benchmarks. Grey dots represent the training data, while red dots denote the test data.

Figure 10 presents the persistent attributes derived from the non-harmonic spectra of a persistent topological hypergraph at 0th and 1st dimensions. Specifically, the nonzero maximum spectra are denoted as $\lambda^{max}0$ and $\lambda^{max}1$ for the first and second dimensions, respectively. The persistent standard deviation of nonzero spectra is represented by λ_0^{std} and λ_1^{std} for the first two dimensions. Additionally, the values λ_0^{avg} and λ_1^{avg} correspond to the 0th and 1st dimensional persistent average values of nonzero spectra. The λ_0^{sum} and λ_1^{sum}, as shown in Figure 10**e**, indicate the persistent summation values of nonzero spectra of the persistent topological hyperdigraph Laplacian. Furthermore, we generated the persistent topological hypergraph homology for the Vietoris-Rips hypergraph, as depicted in Figure 10**f**. This figure displays persistent barcodes for homology groups H_0 and H_1 at the 0th and 1st dimensions, respectively. The topological invariants, specifically Betti numbers, correspond to the multiplicity of the zero eigenvalue of the topological hypergraph Laplacian for a given filtration parameter d. Notably, these persistent attributes vary with different filtration parameters, indicating the recording of both topological and geometric information in a multiscale manner.

The Figure 13 illustrates diverse topological representations for the C_α atoms in protein PDBID: 6L9D, with a cutoff distance set to 5 Å. In Figure 13**a**, the commonly used graph representation of the structure is depicted. Generalizing from the graph representation, the 0-simplices in

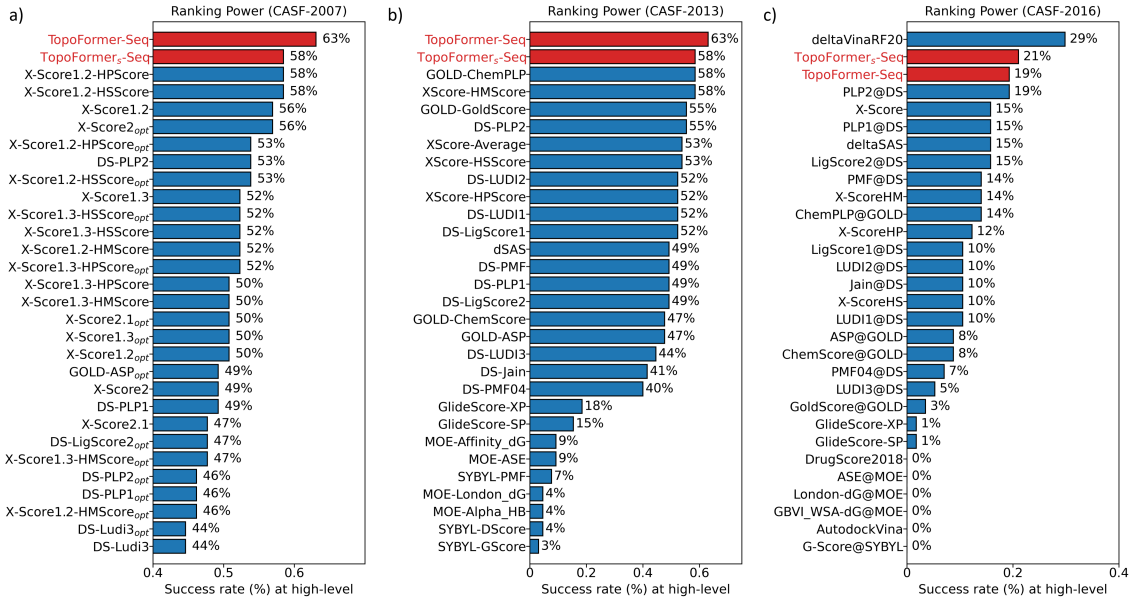

Figure 8: Performance of ranking power evaluated by the high-level success measurement compares with different scoring functions on CASF-2007, CASF-2013, and CASF-2016 benchmarks. The proposed TopoFormer-based models are plotted in the red color. The results of other methods, taken from refs ([25, 24, 26, 16, 27, 19, 30, 65]), are in the blue color

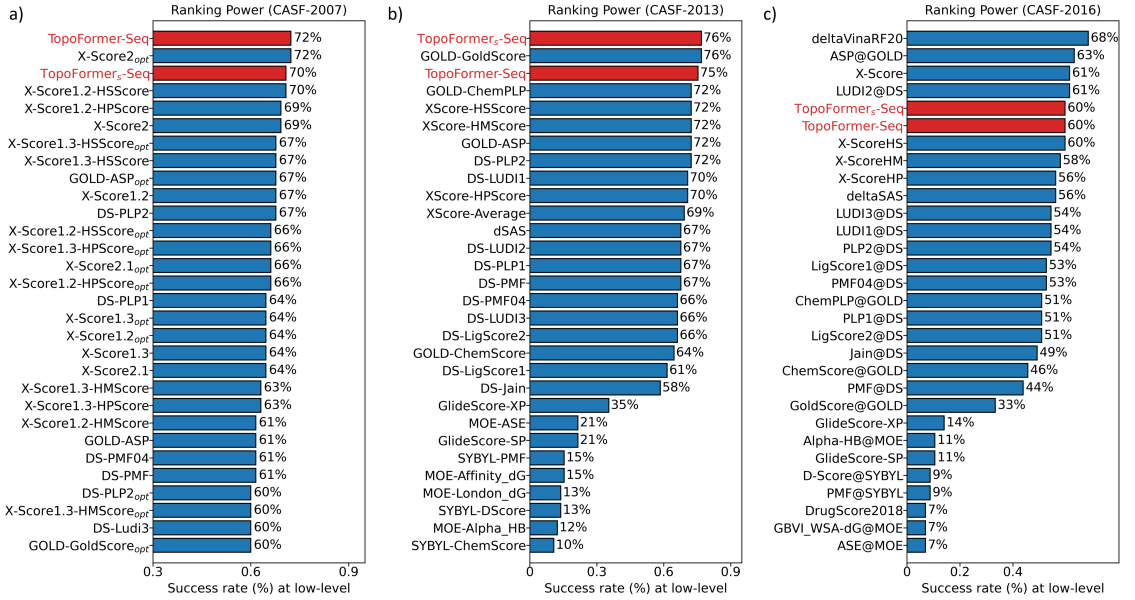

Figure 9: Performance of ranking power evaluated by the low-level success measurement compares with different scoring functions on CASF-2007, CASF-2013, and CASF-2016 benchmarks. The proposed TopoFormer-based models are plotted in the red color. The results of other methods, taken from refs ([25, 24, 26, 16, 27, 19, 30, 65]), are in the blue color.

the simplicial complex correspond to the vertices in the graph, while the 1-simplices represent edges with vertices from the graph, as shown in the second and third rows of Figure 13**a** and **b**. Additionally, higher-dimensional simplices in the simplicial complex provide more intricate information

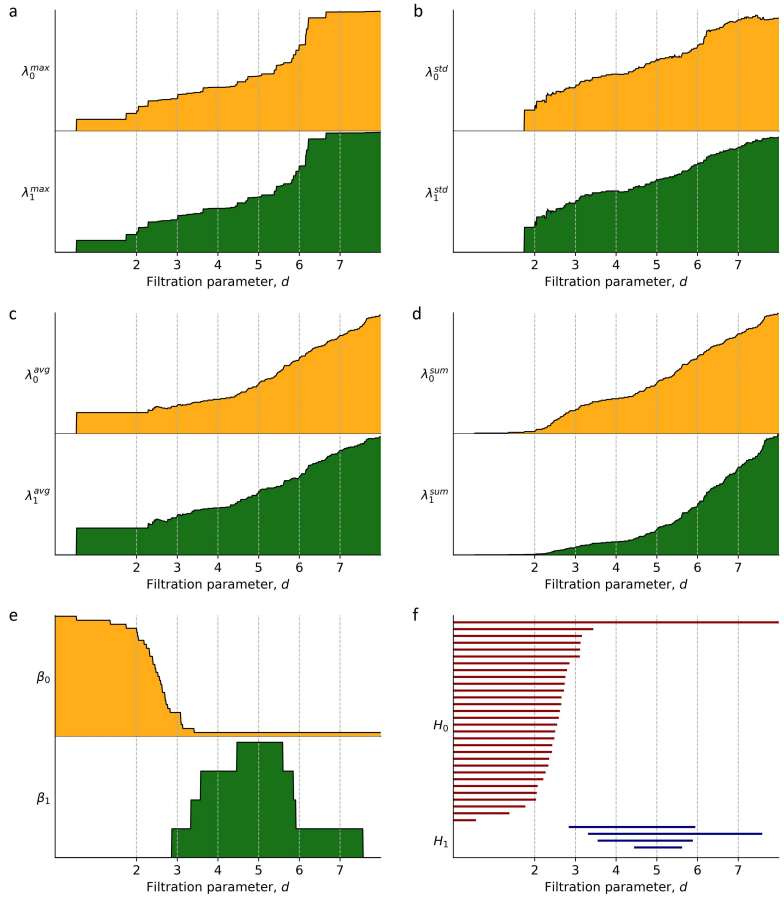

Figure 10: Comparison of persistent topological hyperdigraph Laplacian and persistent homology for the point cloud in Figure 4**a**. **a** Representation of nonzero maximum non-harmonic spectra of the persistent topological hyperdigraph Laplacian at the 0th and 1st dimensions (λ_0^{max} and λ_1^{max}), highlighting their dependence on the filtration (scale) parameter d. **b** Representation of standard deviation of nonzero spectra of the persistent topological hyperdigraph Laplacian at the 0th and 1st dimensions (λ_0^{std} and λ_1^{std}). **c** Representation of average value of nonzero spectra of the persistent topological hyperdigraph Laplacian at the 0th and 1st dimensions (λ_0^{avg} and λ_1^{avg}). **d** Representation of summation value of nonzero spectra of the persistent topological hyperdigraph Laplacian at the 0th and 1st dimensions (λ_0^{sum} and λ_1^{sum}). **e** Representation of multiplicity of zero of the persistent topological hyperdigraph Laplacian at the 0th and 1st dimensions (β_0 and β_1). **f** Visualization of barcodes for persistent homology groups H_0 and H_1 at the 0th and 1st dimensions, respectively, showcasing their variations with respect to the filtration parameter d.

about the structure; for instance, alpha helices can be roughly represented by 3-simplices.

Moving beyond simplicial complexes, hypergraphs offer a more generalized representation of the structure, as demonstrated in Figure 13**c**. Furthermore, with directional information, hyperdigraphs present an even more generalized view compared to simplicial complexes and hypergraphs. The hyperdigraph representation, along with different dimensional directed hyperedges, captures information at various levels, as illustrated in figure 13**d**. Notably, the representation via 3-directed hyperedges can also unveil the presence of an alpha helix.

To illustrate the power of proposed topological hyperdigraph and its Laplacian, two $B_7C_2H_9$ isomers with identical geometric structures, differing only in the positions of carbon atoms are used in the validation. Figures 14**a** and **b** show the molecular structures of these two $B_7C_2H_9$

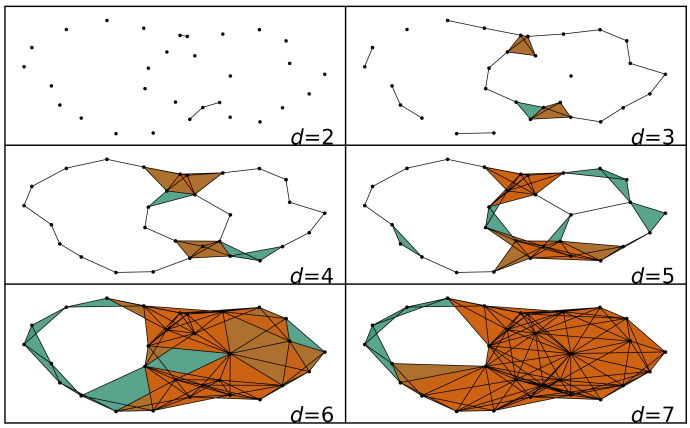

Figure 11: Illustration of how the changing filtration parameter leads to alterations in the connectivity of the point cloud in Figure 4**a**, resulting in the generation of a series of simplicial complex. The 2-simplices are triangles colored by the green. The 3-simplices are tetrahedrons colored by orange.

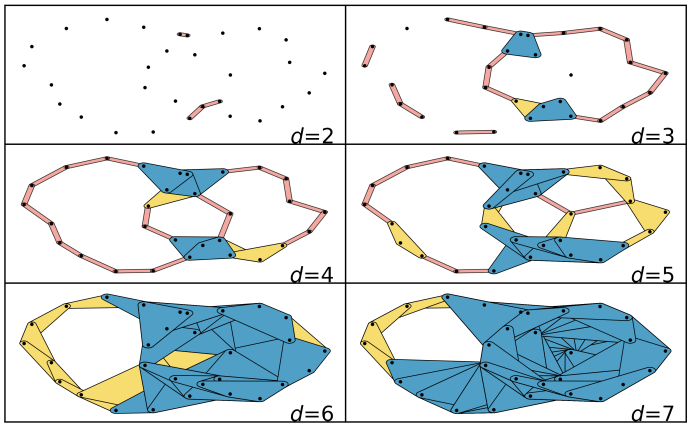

Figure 12: Illustration of how the changing filtration parameter leads to alterations in the connectivity of the point cloud Figure 4**a**, resulting in the generation of a series of hypergraph. The 1-hyperedge is represented by the light red area. The 2-hyperedge is represented by yellow area. The 3-hyperedge is represented by blue area.

isomers. Here, the structure without hydrogen atoms are considered in the analysis, as shown in Figure 14**c** and **d**. Figures 14**e**, **g**, and **i** are the simplicial complex, hypergraph, and hyperdigraph representations and their Laplacians' analysis results for structure 14**c**. Figures 14**f**, **h**, and **j** are the simplicial complex, hypergraph, and hyperdigraph representations and their Laplacians' analysis results for other structure in Figure 14**d**. While only carbon atoms are changed in the structures 14**c** and **d**, the Laplacians analysis for simplicial complex and hypergraph can not classify these two structures. For hyperdigraph, because the directed hyperedge can be used to encode the non-symmetry and non-balance relations, the changing position of carbon atoms can be captured by the directed hyperedge, which result different topological hyperdigraph Laplacians. So that either the multiplicity of zero eigenvalue of these Laplacians (β_0, β_1, and β_2) or the minimum nonzero

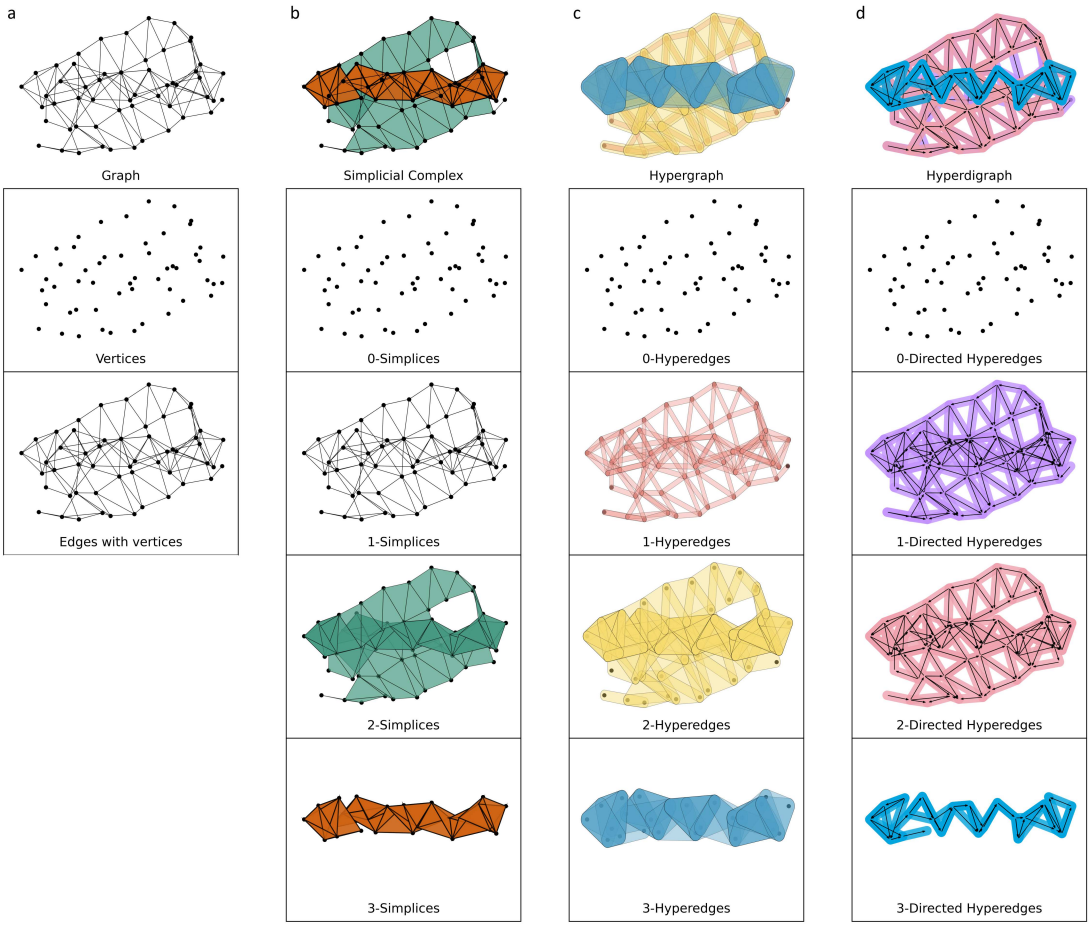

Figure 13: Illustration of Different Representations for C_α Atoms in Protein PDBID: 6L9D. **a** The graph representation of the structure. **b** The simplicial complex representation of the structure, including a list of 0-simplices, 1-simplices, 2-simplices, and 3-simplices in the complex. **c** The hypergraph representation of the structure, along with a list of 0, 1, 2, and 3-hyperedges within the hypergraph. **d** The hyperdigraph representation of the structure, providing a listing of 0, 1, 2, and 3-directed hyperedges in the hyperdigraph.

spectra of there Laplacians (λ_0^{min}, λ_1^{min}, and λ_2^{min}) can distinct these two structures.

To assess the efficacy of the proposed topological hyperdigraph and its Laplacian, we employ two $B_7C_2H_9$ isomers sharing identical geometric structures but differing solely in the positions of carbon atoms. Figures 14**a** and **b** illustrate the molecular structures of these isomers. In the analysis, we consider the structures without hydrogen atoms, as depicted in Figures 14**c** and **d**. Figures 14**e**, **g**, and **i** present the analysis results for the simplicial complex, hypergraph, and hyperdigraph representations, along with their Laplacians, for the structure in Figure 14**c**. Similarly, Figures 14**f**, **h**, and **j** display the analysis results for the other structure in Figure 14**d**. Despite the alteration being limited to the carbon atoms in structures shown in Figures 14**c** and **d**, the Laplacian analysis of simplicial complex and hypergraph representations fails to differentiate these two structures. In contrast, the hyperdigraph, leveraging directed hyperedges to encode non-symmetry and non-balance relations, effectively captures the changing positions of carbon atoms. This results in distinct topological hyperdigraph Laplacians, reflected in either the multiplicity of zero eigenvalues (β_0, β_1, and β_2) or the minimum nonzero spectra of these Laplacians (λ_0^{min}, λ_1^{min},

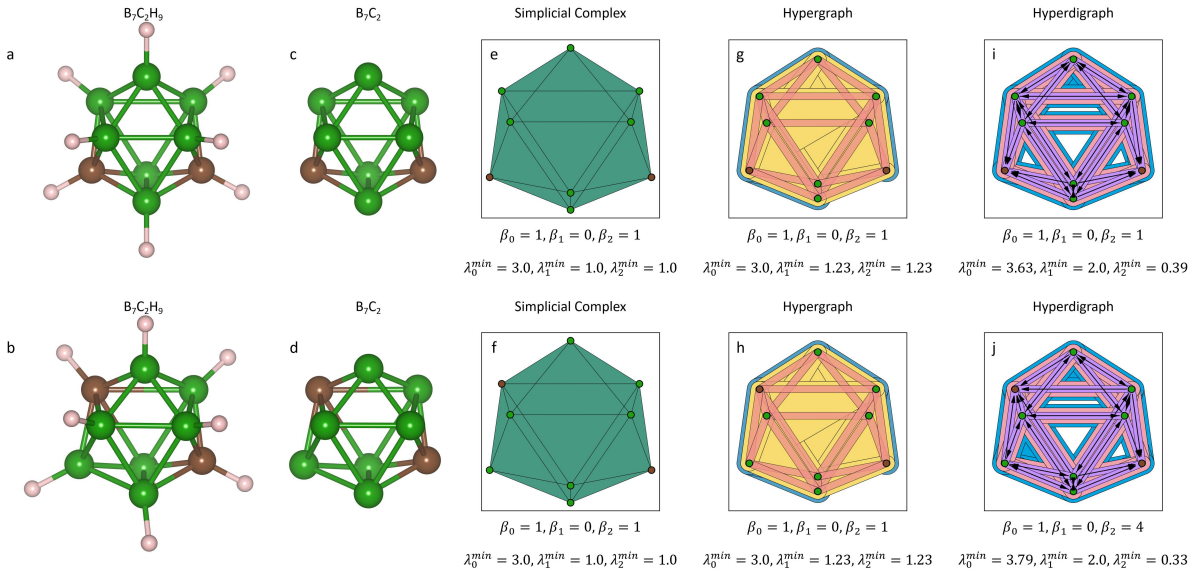

Figure 14: Illustration of the different Laplacian analysis for two $B_7C_2H_9$ isomers. **a** and **b** Two $B_7C_2H_9$ isomers with identical geometric structures, differing only in the positions of carbon atoms. **c** and **d** The structural representations of $B_7C_2H_9$ after the removal of hydrogen atoms. **e** and **f** The simplicial complex representations of structures **c** and **d**. Carbon atoms are highlighted in coffee color, while boron atoms are shown in light green. The 2-simplices are shaded in green. Notably, the corresponding topological invariants and non-harmonic spectra are consistent, with $\beta_0 = 1$, $\beta_1 = 0$, and $\beta_2 = 1$, as well as $\lambda_0^{min} = 3.0$, $\lambda_1^{min} = 1.0$, and $\lambda_2^{min} = 1.0$. **g** and **h** The hypergraph representations of structures **c** and **d**. The 1-hyperedges are colored in light red, and 2-hyperedges are colored yellow, while 3-hyperedges are shaded in blue. The corresponding topological invariants and non-harmonic spectra are consistent, with $\beta_0 = 1$, $\beta_1 = 0$, and $\beta_2 = 1$, as well as $\lambda_0^{min} = 3.0$, $\lambda_1^{min} = 1.23$, and $\lambda_2^{min} = 1.23$. **i** The hyperdigraph representation of structure **c**. The 1-directed hyperedges are colored in purple, the 2-directed hyperedges are pink, and the 3-directed hyperedges are blue. The corresponding topological invariants are $\beta_0 = 1$, $\beta_1 = 0$, and $\beta_2 = 1$. The minimum non-zero non-harmonic spectra are $\lambda_0^{min} = 3.63$, $\lambda_1^{min} = 2.0$, and $\lambda_2^{min} = 0.39$. **j** The hyperdigraph representation of structure **c**. The 1-directed hyperedges are colored in purple, the 2-directed hyperedges are pink, and the 3-directed hyperedges are blue. The corresponding topological invariants are $\beta_0 = 1$, $\beta_1 = 0$, and $\beta_2 = 4$. The minimum non-zero non-harmonic spectra are $\lambda_0^{min} = 3.79$, $\lambda_1^{min} = 2.0$, and $\lambda_2^{min} = 0.33$. The difference in Charts **i** and **j** demonstrates that topological hyperdigraph Laplacians can distinguish these two isomers.

and λ_2^{min}). Consequently, these topological features successfully distinguish two structures.

References

[1] Nic Fleming. Computer-calculated compounds. *Nature*, 557(7707):S55–7, 2018.

[2] Jiankun Lyu, Sheng Wang, Trent E Balius, Isha Singh, Anat Levit, Yurii S Moroz, Matthew J O'Meara, Tao Che, Enkhjargal Algaa, Kateryna Tolmachova, et al. Ultra-large library docking for discovering new chemotypes. *Nature*, 566(7743):224–229, 2019.

[3] Douglas B Kitchen, Hélène Decornez, John R Furr, and Jürgen Bajorath. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nature reviews Drug discovery*, 3(11):935–949, 2004.

[4] Luca Pinzi and Giulio Rastelli. Molecular docking: shifting paradigms in drug discovery. *International journal of molecular sciences*, 20(18):4331, 2019.

[5] Nataraj S Pagadala, Khajamohiddin Syed, and Jack Tuszynski. Software for molecular docking: a review. *Biophysical reviews*, 9:91–102, 2017.

[6] Lingle Wang, Yujie Wu, Yuqing Deng, Byungchan Kim, Levi Pierce, Goran Krilov, Dmitry Lupyan, Shaughnessy Robinson, Markus K Dahlgren, Jeremy Greenwood, et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *Journal of the American Chemical Society*, 137(7):2695–2703, 2015.

[7] Gregory Sliwoski, Sandeepkumar Kothiwale, Jens Meiler, and Edward W Lowe. Computational methods in drug discovery. *Pharmacological reviews*, 66(1):334–395, 2014.

[8] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[9] Minkyung Baek, Frank DiMaio, Ivan Anishchenko, Justas Dauparas, Sergey Ovchinnikov, Gyu Rie Lee, Jue Wang, Qian Cong, Lisa N Kinch, R Dustin Schaeffer, et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science*, 373(6557):871–876, 2021.

[10] Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

[11] Jiaying Luo, Wanlei Wei, Jérôme Waldispühl, and Nicolas Moitessier. Challenges and current status of computational methods for docking small molecules to nucleic acids. *European journal of medicinal chemistry*, 168:414–425, 2019.

[12] Yu-Chen Lo, Stefano E Rensi, Wen Torng, and Russ B Altman. Machine learning in chemoinformatics and drug discovery. *Drug discovery today*, 23(8):1538–1546, 2018.

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. *Advances in neural information processing systems*, 30, 2017.

[14] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[16] Zixuan Cang, Lin Mu, and Guo-Wei Wei. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS computational biology*, 14(1):e1005929, 2018.

[17] Duc Duy Nguyen, Zixuan Cang, and Guo-Wei Wei. A review of mathematical representations of biomolecular data. *Physical Chemistry Chemical Physics*, 22(8):4343–4367, 2020.

[18] Rui Wang, Duc Duy Nguyen, and Guo-Wei Wei. Persistent spectral graph. *International journal for numerical methods in biomedical engineering*, 36(9):e3376, 2020.

[19] Zhenyu Meng and Kelin Xia. Persistent spectral–based machine learning (perspect ml) for protein-ligand binding affinity prediction. *Science advances*, 7(19):eabc5329, 2021.

[20] Dong Chen, Jian Liu, Jie Wu, and Guo-Wei Wei. Persistent hyperdigraph homology and persistent hyperdigraph laplacians. *Foundations of Data Science*, 5:558–588, 2023.

[21] Afra Zomorodian and Gunnar Carlsson. Computing persistent homology. In *Proceedings of the twentieth annual symposium on Computational geometry*, pages 347–356, 2004.

[22] Dong Chen, Jiaxin Zheng, Guo-Wei Wei, and Feng Pan. Extracting predictive representations from hundreds of millions of molecules. *The journal of physical chemistry letters*, 12(44): 10793–10801, 2021.

[23] Kiersten M Ruff and Rohit V Pappu. Alphafold and implications for intrinsically disordered proteins. *Journal of Molecular Biology*, 433(20):167208, 2021.

[24] Yan Li, Li Han, Zhihai Liu, and Renxiao Wang. Comparative assessment of scoring functions on an updated benchmark: 2. evaluation methods and general results. *Journal of chemical information and modeling*, 54(6):1717–1736, 2014.

[25] Tiejun Cheng, Xun Li, Yan Li, Zhihai Liu, and Renxiao Wang. Comparative assessment of scoring functions on a diverse test set. *Journal of chemical information and modeling*, 49(4): 1079–1093, 2009.

[26] Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: the casf-2016 update. *Journal of chemical information and modeling*, 59(2):895–913, 2018.

[27] Duc Duy Nguyen and Guo-Wei Wei. Agl-score: algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. *Journal of chemical information and modeling*, 59(7):3291–3304, 2019.

[28] Maciej Wójcikowski, Michał Kukiełka, Marta M Stepniewska-Dziubinska, and Pawel Siedlecki. Development of a protein–ligand extended connectivity (plec) fingerprint and its application for binding affinity predictions. *Bioinformatics*, 35(8):1334–1341, 2019.

[29] Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics*, 34(21):3666–3674, 2018.

[30] Cheng Wang and Yingkai Zhang. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *Journal of computational chemistry*, 38(3):169–177, 2017.

[31] Timothy J Trull and Ulrich W Ebner-Priemer. Using experience sampling methods/ecological momentary assessment (esm/ema) in clinical assessment and clinical research: introduction to the special section. 2009.

[32] Dmitry S Karlov, Sergey Sosnin, Maxim V Fedorov, and Petr Popov. graphdelta: Mpnn scoring function for the affinity prediction of protein–ligand complexes. *ACS omega*, 5(10): 5150–5159, 2020.

[33] Norberto Sánchez-Cruz, José L Medina-Franco, Jordi Mestres, and Xavier Barril. Extended connectivity interaction features: improving binding affinity prediction through chemical description. *Bioinformatics*, 37(10):1376–1382, 2021.

[34] Zechen Wang, Liangzhen Zheng, Yang Liu, Yuanyuan Qu, Yong-Qiang Li, Mingwen Zhao, Yuguang Mu, and Weifeng Li. Onionnet-2: a convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. *Frontiers in chemistry*, 9:753002, 2021.

[35] Mohammad A Rezaei, Yanjun Li, Dapeng Wu, Xiaolin Li, and Chenglong Li. Deep learning in drug design: protein-ligand binding affinity prediction. *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, 19(1):407–417, 2020.

[36] Shudong Wang, Dayan Liu, Mao Ding, Zhenzhen Du, Yue Zhong, Tao Song, Jinfu Zhu, and Renteng Zhao. Se-onionnet: a convolution neural network for protein–ligand binding affinity prediction. *Frontiers in Genetics*, 11:607824, 2021.

[37] Derek Jones, Hyojin Kim, Xiaohua Zhang, Adam Zemla, Garrett Stevenson, WF Drew Bennett, Daniel Kirshner, Sergio E Wong, Felice C Lightstone, and Jonathan E Allen. Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *Journal of chemical information and modeling*, 61(4):1583–1592, 2021.

[38] Fergus Boyles, Charlotte M Deane, and Garrett M Morris. Learning from the ligand: using ligand-based features to improve binding affinity prediction. *Bioinformatics*, 36(3):758–764, 2020.

[39] Zhihai Liu, Yan Li, Li Han, Jie Li, Jie Liu, Zhixiong Zhao, Wei Nie, Yuchen Liu, and Renxiao Wang. Pdb-wide collection of binding data: current status of the pdbbind database. *Bioinformatics*, 31(3):405–412, 2015.

[40] Oscar Méndez-Lucio, Mazen Ahmad, Ehecatl Antonio del Rio-Chanona, and Jörg Kurt Wegner. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nature Machine Intelligence*, 3(12):1033–1039, 2021.

[41] Liangzhen Zheng, Jintao Meng, Kai Jiang, Haidong Lan, Zechen Wang, Mingzhi Lin, Weifeng Li, Hongwei Guo, Yanjie Wei, and Yuguang Mu. Improving protein–ligand docking and screening accuracies by incorporating a scoring function correction term. *Briefings in Bioinformatics*, 23(3):bbac051, 2022.

[42] Jingxiao Bao, Xiao He, and John ZH Zhang. Deepbsp—a machine learning method for accurate prediction of protein–ligand docking structures. *Journal of chemical information and modeling*, 61(5):2231–2240, 2021.

[43] Chao Shen, Xujun Zhang, Yafeng Deng, Junbo Gao, Dong Wang, Lei Xu, Peichen Pan, Tingjun Hou, and Yu Kang. Boosting protein–ligand binding pose prediction and virtual screening based on residue–atom distance likelihood potential and graph transformer. *Journal of Medicinal Chemistry*, 65(15):10691–10706, 2022.

[44] Xiang Liu, Huitao Feng, Jie Wu, and Kelin Xia. Dowker complex based machine learning (dcml) models for protein-ligand binding affinity prediction. *PLoS Computational Biology*, 18 (4):e1009943, 2022.

[45] Oleg Trott and Arthur J Olson. Autodock vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *Journal of computational chemistry*, 31(2):455–461, 2010.

[46] Seokhyun Moon, Wonho Zhung, Soojung Yang, Jaechang Lim, and Woo Youn Kim. Pignet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chemical Science*, 13(13):3661–3673, 2022.

[47] Ivan Gutman, Xueliang Li, and Jianbin Zhang. Graph energy. *Analysis of Complex Networks: From Biology to Linguistics*, pages 145–174, 2009.

[48] Jianjun Su, Jun-Jie Zhang, Jiacheng Chen, Yun Song, Libei Huang, Minghui Zhu, Boris I Yakobson, Ben Zhong Tang, and Ruquan Ye. Building a stable cationic molecule/electrode interface for highly efficient and durable co 2 reduction at an industrially relevant current. *Energy & Environmental Science*, 14(1):483–492, 2021.

[49] Hui Ren, Shidong Yu, Lingfeng Chao, Yingdong Xia, Yuanhui Sun, Shouwei Zuo, Fan Li, Tingting Niu, Yingguo Yang, Huanxin Ju, et al. Efficient and stable ruddlesden–popper perovskite solar cell with tailored interlayer molecular interaction. *Nature Photonics*, 14(3): 154–163, 2020.

[50] Gareth Jones, Peter Willett, Robert C Glen, Andrew R Leach, and Robin Taylor. Development and validation of a genetic algorithm for flexible docking. *Journal of molecular biology*, 267 (3):727–748, 1997.

[51] Danijela Horak and Jürgen Jost. Spectra of combinatorial laplace operators on simplicial complexes. *Advances in Mathematics*, 244:303–336, 2013.

[52] Beno Eckmann. Harmonische funktionen und randwertaufgaben in einem komplex. *Commentarii Mathematici Helvetici*, 17(1):240–255, 1944.

[53] Jiahui Chen, Rundong Zhao, Yiying Tong, and Guo-Wei Wei. Evolutionary de rham-hodge method. *arXiv preprint arXiv:1912.12388*, 2019.

[54] Rui Wang, Rundong Zhao, Emily Ribando-Gros, Jiahui Chen, Yiying Tong, and Guo-Wei Wei. Hermes: Persistent spectral graph software. *Foundations of data science (Springfield, Mo.)*, 3 (1):67, 2021.

[55] Facundo Mémoli, Zhengchao Wan, and Yusu Wang. Persistent laplacians: Properties, algorithms and implications. *SIAM Journal on Mathematics of Data Science*, 4(2):858–884, 2022.

[56] Edelsbrunner, Letscher, and Zomorodian. Topological persistence and simplification. *Discrete & Computational Geometry*, 28:511–533, 2002.

[57] Jian Liu, Jingyan Li, and Jie Wu. The algebraic stability for persistent laplacians. *arXiv preprint arXiv:2302.03902*, 2023.

[58] Kaiming He, Xinlei Chen, Saining Xie, Yanghao Li, Piotr Dollár, and Ross Girshick. Masked autoencoders are scalable vision learners. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 16000–16009, 2022.

[59] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in python. *the Journal of machine Learning research*, 12:2825–2830, 2011.

[60] Dimitry Kozlov. *Combinatorial algebraic topology*, volume 21. Springer Science & Business Media, 2008.

[61] László Lovász. Kneser's conjecture, chromatic number, and homotopy. *Journal of Combinatorial Theory, Series A*, 25(3):319–324, 1978.

[62] Stephane Bressan, Jingyan Li, Shiquan Ren, and Jie Wu. The embedded homology of hypergraphs and applications. *Asian Journal of Mathematics*, 23(3):479–500, 2019.

[63] Franz Aurenhammer. Voronoi diagrams—a survey of a fundamental geometric data structure. *ACM Computing Surveys (CSUR)*, 23(3):345–405, 1991.

[64] S Geršgorin. Bulletin de l'académie des sciences de l'urss. *Classe des sciences mathématiques et naturelles*, 6:749, 1931.

[65] Hongjian Li, Kwong-Sak Leung, Man-Hon Wong, and Pedro J Ballester. Low-quality structural and interaction data improves binding affinity prediction via random forest. *Molecules*, 20(6): 10947–10962, 2015.