nature machine intelligence

Article

https://doi.org/10.1038/s42256-024-00855-1

Multiscale topology-enabled structure-to-sequence transformer for protein–ligand interaction predictions

Received: 20 November 2023

Accepted: 15 May 2024

Published online: 21 June 2024

Check for updates

Dong Chen [1], Jian Liu [2] ✉ & Guo-Wei Wei [1,3,4] ✉

Despite the success of pretrained natural language processing (NLP) models in various fields, their application in computational biology has been hindered by their reliance on biological sequences, which ignores vital three-dimensional (3D) structural information incompatible with the sequential architecture of NLP models. Here we present a topological transformer (TopoFormer), which is built by integrating NLP models and a multiscale topology technique, the persistent topological hyperdigraph Laplacian (PTHL), which systematically converts intricate 3D protein–ligand complexes at various spatial scales into an NLP-admissible sequence of topological invariants and homotopic shapes. PTHL systematically transforms intricate 3D protein–ligand complexes into NLP-compatible sequences of topological invariants and shapes, capturing essential interactions across spatial scales. TopoFormer gives rise to exemplary scoring accuracy and excellent performance in ranking, docking and screening tasks in several benchmark datasets. This approach can be utilized to convert general high-dimensional structured data into NLP-compatible sequences, paving the way for broader NLP based research.

Drug discovery is crucial for modern healthcare, profoundly affecting our lives. Traditional drug development methods are laborious and expensive, taking over a decade and billions of dollars to bring a single drug to market[1]. These methods, including molecular docking[2–5], free energy perturbation[6] and empirical modelling[7], have advanced drug discovery but have limitations. They often lack accuracy, are computationally intensive for large-scale screenings and may miss unconventional binding sites or interaction kinetics, potentially overlooking therapeutic opportunities.

Deep learning models are emerging as promising tools in drug design[8–11], celebrated for their ability to predict protein structures and identify complex patterns for superior predictions[12]. The transition to deep learning, leveraging chemoinformatics and bioinformatics[13], signifies a pivotal shift towards data-driven approaches in drug design and discovery[14–16]. However, challenges such as the need for frequent retraining and dependence on labelled data remain substantial obstacles.

Groundbreaking transformer-based models like ChatGPT, which leverage large-scale pretraining and unlabelled data, highlight the untapped potential of self-supervised learning[17–19]. These models provide a powerful glimpse into potential solutions, particularly in the field of drug discovery where an insufficiency of labelled data can be a limiting factor[20,21]. 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 protein–ligand complex modelling, raises pertinent questions because the method neglects important stereochemical information of structures. 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.

[1]Department of Mathematics, Michigan State University, East Lansing, MI, USA. [2]Mathematical Science Research Center, Chongqing University of Technology, Chongqing, China. [3]Department of Electrical and Computer Engineering, Michigan State University, East Lansing, MI, USA. [4]Department of Biochemistry and Molecular Biology, Michigan State University, East Lansing, MI, USA. ✉e-mail: liujian@cqut.edu.cn; weig@msu.edu

https://doi.org/10.1038/s42256-024-00855-1

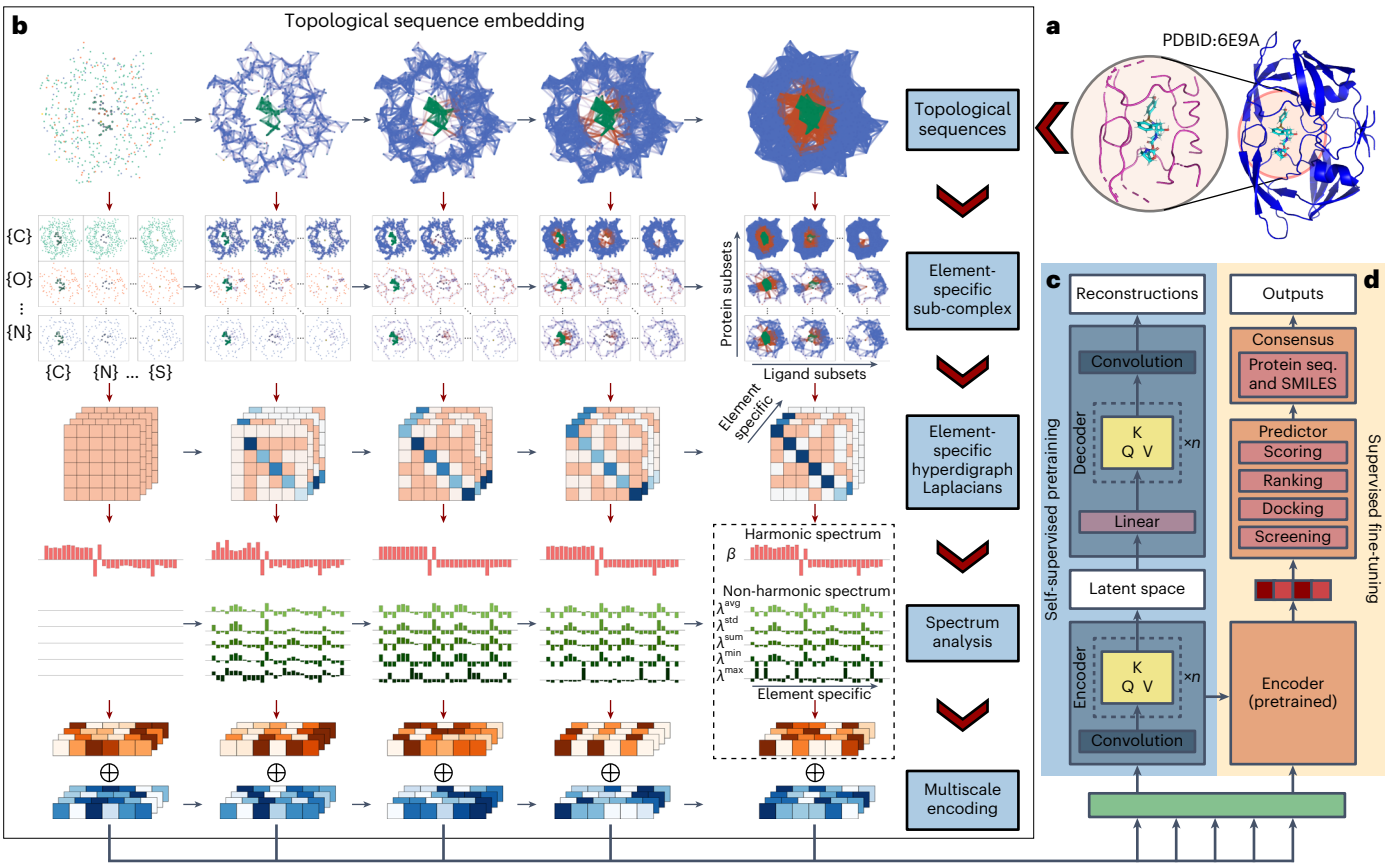


Fig. 1 | Schematic illustration of the overall TopoFormer model. a, A 3D protein–ligand complex (PDBID 6E9A) and its interactive domain. **b**, The topological sequence embedding of a 3D protein–ligand complex. Initially, the complex is split into a topological sequence, known as a chain complex in algebraic topology. Then, element-specific subcomplexes are created to encode physical interactions on a variety of scales controlled by a filtration parameter. Subsequently, element-specific 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 that 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 assembled into a topological sequence as the input to the transformer. **c**, Self-supervised learning is applied to unlabelled 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. seq., sequence.

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 notable successes[22–25]. Drawing upon insights from advanced mathematics, we unveil our topological transformer model: TopoFormer. TopoFormer is built upon persistent topological hyperdigraph Laplacian (PTHL)[26], an advanced algebraic topological method. While intrinsically mirroring foundational topological invariants akin to traditional persistent homology[27], this multiscale technique introduced the 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) intricacies of protein–ligand complexes. In a nutshell, PTHL utilizes its multiscale topology and multiscale spectrum to convert intricate 3D protein–ligand complexes into one-dimensional topological sequences that are ideally suitable for the sequential architecture of transformers (Fig. 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.

The following text introduces the topological transformer (TopoFormer) model and evaluates its performance in key tasks such as scoring, ranking, docking and screening. The analysis highlights TopoFormer's strengths and advantages over traditional methods.

Overview of TopoFormer

The transformer architecture[17] introduced a new technique using attention mechanisms for sequential data analysis across domains[18,28,29]. Inspired by this, we developed a topological transformer model, TopoFormer, integrating our PTHL[26] with the transformer framework, as depicted in Fig. 1. Unlike traditional transformers that process protein and ligand sequences, TopoFormer inputs 3D protein–ligand complexes. It transforms these complexes into sequences of topological invariants and homotopic shapes through PTHL, capturing their physical, chemical and biological interactions at multiple scales. Pretraining on a diverse dataset enables TopoFormer to understand complex

https://doi.org/10.1038/s42256-024-00855-1

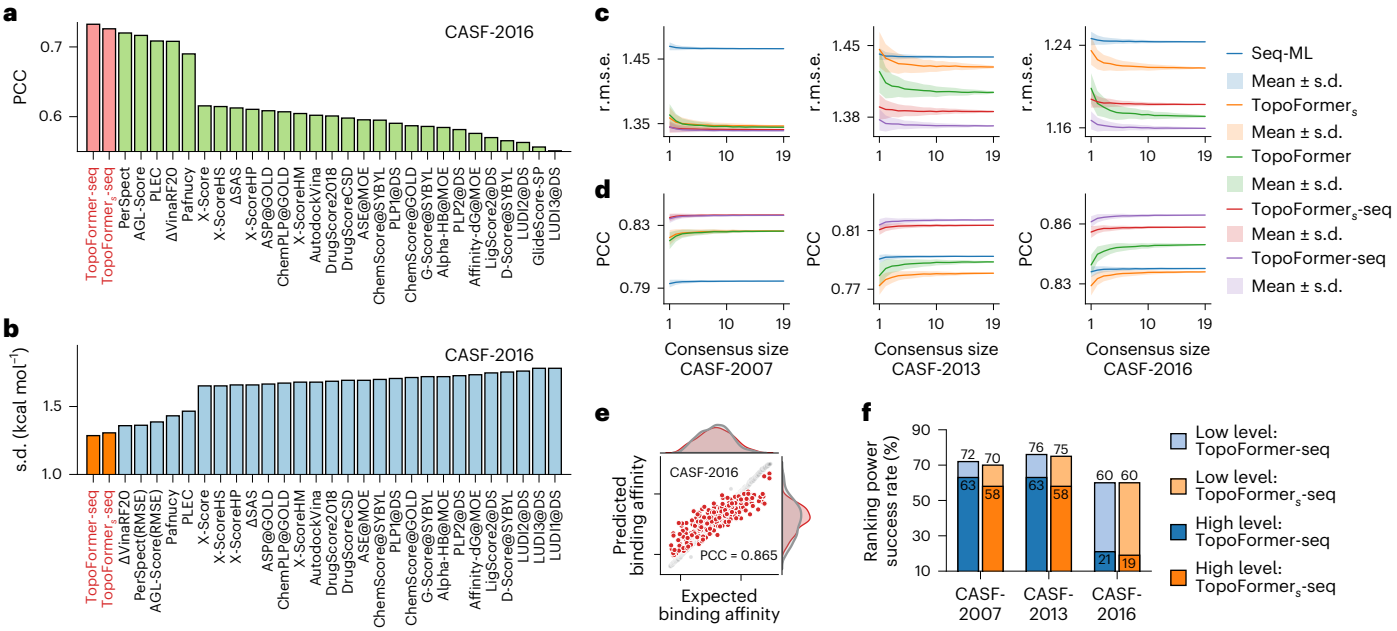


Fig. 2 | Performance of TopoFormer on scoring and ranking tasks. **a**, Comparison of PCCs of various models for protein–ligand complex binding affinity scoring on the CASF-2016 benchmark. The results from other methods are in the green colour. **b**, Comparison of s.d. of different models for protein–ligand complex binding affinity scoring on the CASF-2016 benchmark. The results from other methods are in the blue colour. The quantitative results of **a** and **b** are listed in Supplementary Table 4, taking from refs. 22,25,30–32, 42,48,69,70. **c**, Comparison of the r.m.s.e. (in log K_a) 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 r.m.s.e. and the shaded background provides the error bar for these 400 r.m.s.e. values. **d**, Comparison of the PCC 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 and the shaded background provides the error bar for the 400 PCCs at each consensus size. **e**, 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 and red dots denote the test data. **f**, 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 and those from TopoFormer$_s$-Seq are illustrated in orange.

molecular interactions, including stereochemical effects not evident in molecular sequences. Fine-tuning on specific datasets allows it to capture detailed interactions within complexes and their characteristics relative to the entire dataset, enhancing downstream deep learning applications.

To focus our analysis, we identify heavy ligand and nearby protein atoms within set distances, using a 20 Å or a more precise 12 Å cutoff, as depicted in Fig. 1a. TopoFormer then transforms 3D molecular structures into topological sequences through its topological sequence embedding module (Fig. 1b), utilizing PTHLs for a multiscale analysis. This process embeds various physical, chemical and biological interactions into sequences of vectors.

TopoFormer undergoes self-supervised pretraining with unlabelled protein–ligand complexes, as shown in Fig. 1c, using a transformer encoder–decoder to reconstruct topological sequences. This phase, which measures the accuracy by comparing output and input embeddings, prepares the model to understand protein–ligand dynamics without labelled data. Following pretraining, TopoFormer enters a supervised fine-tuning stage with labelled complexes (Fig. 1d), where the initial embedded vector becomes a key feature for downstream tasks like scoring, ranking, docking and screening. Each task has a dedicated head in the predictor module. To ensure accuracy and reduce biases, TopoFormer integrates multiple topological transformer deep learning models initialized with different seeds and complements them with sequence-based models. The final output is a consensus of these diverse predictions, making TopoFormer a comprehensive model for analysing protein–ligand interactions, leveraging both topological insights and deep learning.

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 (refs. 30–32). The Pearson correlation coefficient (PCC), the standard deviation (s.d.) and the root mean squared error (r.m.s.e.) are used to measure the performance of the scoring function. In this task, we consider two TopoFormer models: a large model (TopoFormer) with longer topological sequence 100; a smaller model (TopoFormer$_s$) with topological sequence of length 50.

To enhance robustness, we train 20 topological transformers (TopoFormer) with unique random seeds for each dataset, minimizing initialization errors. Predictions from smaller models are labelled TopoFormer$_s$. To mitigate biases from using a single model type, we also employ sequence-based models, incorporating protein features from the experience sampling method (ESM) model[33] and Simplified Molecular Input Line Entry System (SMILES) features from the Transformer-CPZ model[28]. Additionally, 20 gradient boosting regressor tree models are trained on these sequence-based features, with their collective predictions termed Seq-ML. The final output, a blend of TopoFormer and Seq-ML predictions, is represented as TopoFormer-Seq and TopoFormer$_s$-Seq for the smaller models. Figure 2c,d illustrates how consensus size affects performance, with 400 trials per size showing that larger consensus sizes yield better performance (higher PCC, lower r.m.s.e.) and more stability (less error variation). A consensus size of 10 is chosen for further analysis, where TopoFormer-Seq consistently outperforms other models, closely followed by TopoFormer$_s$-Seq.

Table 1 | The PCC and r.m.s.e. (in kcal mol^{-1}) of our TopoFormer models on the three benchmarks of CASF-2007, CASF-2013 and CASF-2016

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)

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 gradient boosting regressor tree parameters can be found in Supplementary Information Section 2.

Our TopoFormer-based models consistently outperform others in terms of PCC scores across three benchmark datasets, as illustrated in Fig. 2a and Supplementary Fig. 3a,b, with the lowest s.d. compared to methods with reported s.d. or r.m.s.e., as detailed in Supplementary Table 4. By averaging results from 400 repetitions, TopoFormer-Seq achieves an average PCC of approximately 0.84 across these datasets, as detailed in Table 1. Notably, on the PDBbind v.2016 dataset, TopoFormer-Seq excels with a PCC of 0.866 and an r.m.s.e. of 1.561 kcal mol^{-1}, surpassing the previous leader, TopBP[22]. These benchmarks are summarized in Table 3, with Fig. 2e and Supplementary Fig. 3c,d showcasing the comparison between predicted and experimental binding affinities.

To assess TopoFormer's performance on structurally similar proteins, we employed the CASF-2016 core set, which can be grouped into 57 clusters based on protein sequence similarity. As illustrated in Supplementary Fig. 4, TopoFormer-Seq yielded the lowest mean r.m.s.e. (1.504 kcal mol^{-1}) across all clusters. DeltaVinaRF20 followed closely with an r.m.s.e. of 1.563 kcal mol^{-1}. These findings indicate TopoFormer's superior performance compared to other commonly used methods. For a comprehensive overview, the quantitative r.m.s.e. values of all methods across all clusters are presented in Supplementary Table 5.

Recently, several deep learning models have been reported for the prediction of protein–ligand binding affinity. Notable examples include the graphDelta model[34], ECIF model[35], OnionNet-2 model[36], DeepAtom model[37] and others[38–40]. 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 Supplementary Table 3. For the latest PDBbind v.2020 (ref. 41), 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 r.m.s.e. of 1.295 (equivalent to 1.769 kcal mol^{-1}) on the core set of CASF-2007. For the CASF-2013 core set, the PCC of 0.832 and an r.m.s.e. of 1.301 (equivalent to 1.777 kcal mol^{-1}) are obtained. Similarly, on the CASF-2016 core set, we obtained a PCC of 0.881 with an r.m.s.e. of 1.095 (equivalent to 1.496 kcal mol^{-1}). For the PDBbind v.2016 core set, we achieved a PCC of 0.883 with an r.m.s.e. of 1.086 (equivalent to 1.483 kcal mol^{-1}). 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.

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. 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 Information Section 1.

Figure 2f 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 Supplementary Fig. 5 and low-level measurement Supplementary Fig. 6. Similarly, for the CASF-2013, the TopoFormer-Seq model achieved 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[32], 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 (ref. 42).

Evaluating TopoFormer on docking tasks

In the present study, we harnessed the capabilities of TopoFormer$_s$ to assess its docking proficiency, particularly its ability to distinguish native binding poses from those generated by established docking software packages. (Due to computational resource constraints, we employed TopoFormer$_s$ for both docking and screening tasks.) Our evaluation centred on benchmark datasets CASF-2007 and CASF-2013 (refs. 30,31). A pose was considered native if its root-mean-square deviation (r.m.s.d.) 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. The detailed assessment of docking success rates is available in Supplementary Information Section 1.

In molecular docking, deep learning methods have been applied effectively, leading to notable advancements[43]. Notable approaches include DeepDock[44] (62.11% success), OnionNet-SFCT[45] (76.84%), DeepBSP[46] (79.7%) and RTMScore[47] (80.7%) on the PDBbind core set. However, direct comparisons are difficult due to training on diverse datasets. For a fair evaluation, we trained TopoFormer$_s$ on publicly available data and compared it on the CASF-2007 and CASF-2013 datasets[42,48,49], as detailed in Methods. As depicted in Fig. 3f,g, TopoFormer$_s$ achieved success rates of 93.3% on CASF-2007 and 91.3% on CASF-2013, outperforming existing models and demonstrating the efficacy of our topological approach. This underscores the diversity and potential of new methodologies in improving docking accuracy, offering a comprehensive and innovative solution to the docking challenge.

https://doi.org/10.1038/s42256-024-00855-1

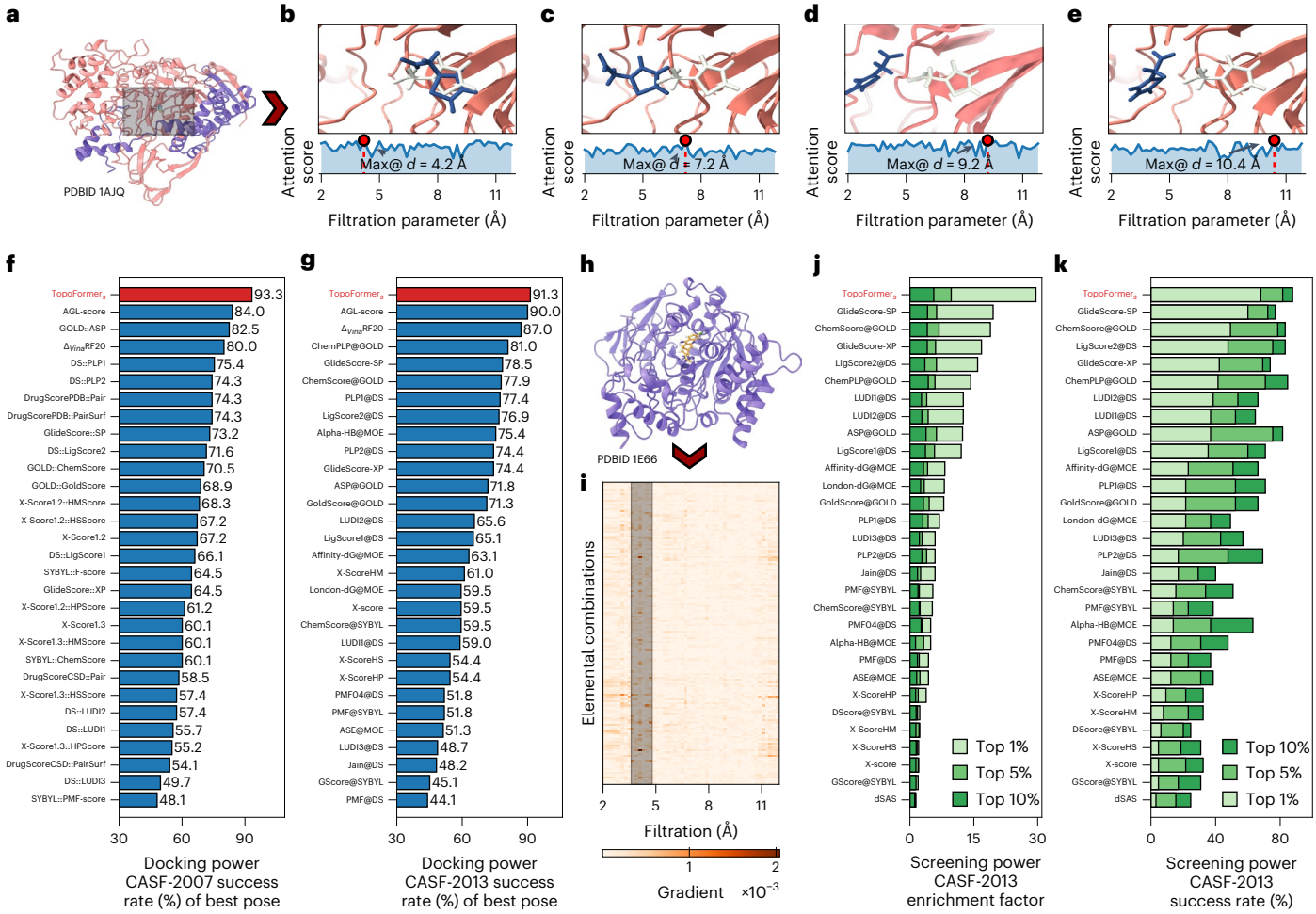


Fig. 3 | Performance of TopoFormer on docking and screening tasks. **a**, Visualization of the protein–ligand complex PDBID 1AJQ. The highlighted rectangle shows the protein's pocket area. **b**–**e**, Four distinct ligand poses within the protein 1AJQ. The molecule in light grey represents the true pose, while the blue molecules depict alternative poses with r.m.s.d. values of 0 Å (**b**), 1.6 Å (**c**), 5.8 Å (**d**) and 7.5 Å (**e**). The light blue curve represents the attention score generated by TopoFormer, varying with the filtration parameter (that is, 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 in **b**, **c**, **d** and **e**, respectively. **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 colour bar 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. Max@, maximum at.

To understand what TopoFormer$_s$ learned in post-fine-tuning, we analysed the impact of spatial scale on protein–ligand interactions using attention scores. Figure 3b–e illustrates four ligand poses near the protein pocket (Protein Data Bank Identifier (PDBID) 1AJQ) highlighted in Fig. 3a. The real experimental pose in Fig. 3b has an r.m.s.d. of 0 Å. We calculated TopoFormer$_s$'s attention scores for all spatial scales, reflecting the impact of interaction ranges on the docking score. The highest attention score at $d = 4.2$ Å suggests this scale most influences binding affinity. Figure 3c–e displays poses with r.m.s.d.s of 1.6 Å, 5.9 Å and 7.5 Å, respectively, with their maximum attention scores at scales $d = 7.2$ Å, $d = 9.2$ Å and $d = 10.4$ Å. This indicates a positive correlation between pose deviation from the true position and the scale at which interactions most affect the docking score.

Evaluating TopoFormer on screening tasks

Machine learning transforms the screening task by making it more accurate, efficient and cost-effective, which is vital for accelerating the pace of drug discovery[50]. To assess the screening capabilities of our TopoFormer method, we employ the CASF-2013 core set. 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 pretrained 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[28] and the ESM model[33]. This yields probabilities for the given ligands, referred to as S_2. Ligands with high multiplied scores ($S = S_1 \times S_2$) are identified as predicted binders. 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), specifically $EF_{1\%}$, $EF_{5\%}$ and $EF_{10\%}$, are used in the virtual screening for drug discovery. 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 EF are provided in Supplementary Information Section 1.

Figure 3j,k shows that TopoFormer outperforms previous methods in success rate and EF. TopoFormer achieves a 68% success rate and

a 29.6% EF for the top 1%-ranked molecules, surpassing GlideScore-SP's 60% success rate and 19% EF. For the top 5% and 10% ranked molecules, TopoFormer's success rates are 81.5% and 87.8%, with EFs of 9.7 and 5.6, respectively, the highest among tested methods (Fig. 3k). AGL-score[48] and ΔVineRF20 (ref. 42) show comparable results but only for the top 1% on the CASF-2013 dataset. Deep learning models like RTMScore[47], DeepDock[44] and PIGNet[51] also show notable results but were evaluated on the CASF-2016 set and trained on different datasets, limiting direct comparison with TopoFormer.

To identify the most influential scales of protein–ligand interactions on TopoFormer$_s$ predictions, we generated a saliency map for a specific complex (PDBID 1E66), as shown in Fig. 3h. The analysis considers protein atoms within 12 Å of the ligand. In Fig. 3i, the *y* axis represents different element-specific combinations, and the *x* axis shows the filtration parameter from 2 Å to 12 Å. The colour bar indicates the gradient intensity for each topological feature, with large gradients marked in black, especially around the 4 Å scale. This saliency map highlights the decision-making process of TopoFormer$_s$, showing that heavy-atom interactions around 4 Å substantially impact the model's screening output, given the absence of hydrogen atoms in the PDBbind database and our models.

We also evaluated our proposed method using the LIT-PCBA dataset, which is characterized by an extreme imbalance between experimentally verified actives and inactives, reflecting the challenging conditions of real screening tasks. We included all 15 targets from the LIT-PCBA dataset in our evaluation, measuring performance using the $EF_{1\%}$ across these targets. As shown in Table 2, our model demonstrated competitive performance, achieving an average $EF_{1\%}$ of 7.29, which surpasses most score function-based screening methods, except for the Interaction Fingerprint (IFP) method, which reported an $EF_{1\%}$ of 7.46 (ref. 52). It is important to note that IFP and GRIM methods, while not strictly score function-based, resemble fingerprint similarity search approaches, and their generalizability may be limited in some cases[53,54]. Our model relies on 3D poses generated by AutoDock Vina, which itself achieved a screening efficacy of $EF_{1\%} = 4.74$. To provide a comprehensive understanding of our model's performance, we conducted detailed evaluations for each target within the dataset and compared our results with the most recent published work[55] (Supplementary Table 6). Despite comparing our model against the best results from models trained with multiple parameters, our approach outperformed others on 8 out of the 15 targets. It is also important to clarify that our evaluation did not involve overfitting our model; the reported results are the average outcomes from 20 TopoFormer-Seq models.

However, it must be acknowledged that comparisons may not be entirely fair due to the use of different docking software across methods, which can substantially impact performance. Despite these challenges, our findings indicate that our model maintains excellent screening capabilities across large virtual screening benchmarks. Furthermore, we will make all 3D poses generated during this study publicly available, contributing to the transparency and reproducibility of our research.

Table 2 | Comparison of the screening powers on LIT-PCBA dataset

Groups	Docking programs	Scoring function	Average $EF_{1\%}$
Ref. [52]	Surflex	Surflex	2.51
		Pafnucy	5.32
		ΔVinaRF20	5.38
		IFP	7.46[a]
		GRIM	6.87[a]
Ref. [67]	Smina	RFScore-4	1.28
		RFScore-VS	0.73
		Vina	1.1
		Dense (affinity)	2.58
	Smina + Vinardo	Vinardo	0.99
Ref. [68]	Smina + Lin_F9	Vina	2.78
		ΔVinaRF20	3.18
		Lin_F9	2.21
		Δ_{Lin_F9}XGB	5.55
Ref. [55]	Glide SP	Glide SP	4.06
		GT	6.51[b]
		GatedGCN	6.8[b]
This work	AutoDock Vina	Vina	4.74
		TopoFormer$_s$-Seq	7.29

[a]Similarity searching approach. [b]The best score among models trained with different hyperparameters is shown.

Discussion

In our study, we utilize the PTHL for a detailed representation of 3D protein–ligand complexes, offering advantages over conventional graphs, simplicial complexes and hypergraphs (see Supplementary Fig. 9). As shown in Fig. 4c, the topological hyperdigraph captures complex higher-order relationships through directed hyperedges that connect vertices in specific sequences, covering dimensions from 0 to 3. This approach allows for modelling complex interactions beyond simple pairwise connections by using directed hyperedges of various dimensions. Moreover, the orientation of these edges incorporates physical and chemical properties, such as electronegativity and ionization energy, providing a more nuanced representation than traditional methods. Supplementary Fig. 10g,h demonstrates this capability by differentiating two $B_7C_2H_9$ isomers with directed hyperedges, showcasing the method's ability to effectively distinguish elemental configurations.

In the investigation of protein–ligand complexes, we employ topological hyperdigraphs for initial representation, further enhanced by PTHL theory[26] to analyse their geometric and topological features. Drawing inspiration from physical systems like molecular structures, where the zeroth-dimensional Hodge Laplacian operator has the connection with the kinetic energy operator of the Hamiltonian for well-defined quantum systems, we extend a discrete analogy to topological hyperdigraphs. These eigenvalues of Laplacian matrix provide insights into the topological object's properties, akin to a physical system's energy spectrum, offering a detailed view of the structural and energetic aspects of complex systems.

Compared to traditional persistent homology, our PTHL method marks a substantial advancement by analysing a broader range of structures beyond simplicial complexes. It captures fundamental homology information and geometric insights, including Betti numbers and homotopic shape evolution, through the non-harmonic spectra of persistent Laplacians. Supplementary Fig. 7a–e shows our method's analysis results, offering a more comprehensive characterization than traditional homology, which is illustrated in Supplementary Fig. 7f. The multiplicity of zero eigenvalues of the Laplacians, corresponding to Betti numbers, confirms that our approach encompasses barcode information, as shown in Supplementary Fig. 7e, providing a robust framework for understanding protein–ligand complexes.

To capture the complex range of atomic interactions in protein–ligand complexes, including covalent, ionic and van der Waals forces, we utilize the PTHL for a multiscale analysis. This method allows for the examination of interactions across scales by evolving topological sequences based on filtration parameters, aiding transformer models in recognizing the contributions of each scale to properties like binding affinity. Figure 3b–e illustrate how different scales contribute to protein–ligand complex formation through attention scores.

Elemental interactions, including hydrogen bonding, van der Waals forces and pi-stacking, are fundamental to the stability and specificity of protein–ligand complexes. To analyse these interactions at the elemental level, we introduce an element-specific analysis within the topological sequence embedding, as shown in Fig. 1b. This approach constructs sub-hyperdigraphs based on common heavy elements in proteins and ligands, generating element-specific Laplacian matrices to encode interactions within the complex. This technique extracts detailed physical and chemical features, enhancing the transformer model's understanding of the complex dynamics in protein–ligand interactions. Further details on this element-specific analysis are provided in Methods.

Methods

Datasets

The dataset utilized for pretraining 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 v.2020 (ref. 41). 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 pretraining. 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 v.2007), CASF-2013 (equivalent to PDBbind v.2013), CASF-2016 and PDBbind v.2016 datasets. The test set encompasses the respective core sets of these datasets. Given the absence of a core set in PDBbind v.2020, the general set (19,443), excluding the all core sets from CASF-2007, CASF-2013, CASF-2016 and PDBbind v.2016, 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 3.

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[30,31]. In preparation for the docking task training set, a set of 1,000 training poses are generated for each given target ligand–receptor pair within the test set. These training poses were generated using GOLD v.5.6.33 (ref. 56). 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[30]. 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 v.2015 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[48,57]. 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. Additionally, to ensure an unbiased evaluation, we employed the LIT-PCBA benchmark dataset[58], which comprises 15 targets with a total of 7,955 true actives and 2,644,022 inactives. This dataset's active-to-inactive ratio of approximately 1:1000 closely mirrors real-world virtual screening scenarios. Following established practices and to optimize computational efficiency[55], we selected the most representative PDB template for each target as the docking target. Autodock Vina[57] was used to generate up to ten docking poses per compound, with an energy range of three and exhaustiveness of ten. The pose with the strongest binding affinity score was selected for prediction by our model. This resulted in a total of 2,651,977 protein–ligand complexes. Detailed dataset information is summarized in Supplementary Table 7. All posed protein–ligand complexes and associated scores are publicly available at https://github.com/WeilabMSU/TopoFormer.

Table 3 | Detailed information of the used datasets

	Datasets	Training set	Test set (core set)
Pretraining (self-supervised learning)	Combined PDBbind (CASF-2007, 2013, 2016, PDBbind v.2015, v.2020)	19,513	/
Fine-tuning (supervised learning)	CASF-2007	1,105	195
	CASF-2013	2,764	195
	CASF-2016	3,772	285
	PDBbind v.2016	3,767	290
	PDBbind v.2020	18,904	195 (CASF-2007 core set)
			195 (CASF-2013 core set)
			285 (CASF-2016 core set)

Topological sequence embedding

Topological hyperdigraph. Topological hyperdigraphs offer a powerful generalization, encompassing graphs, digraphs, simplicial complexes and hypergraphs. They excel at representing intricate relationships, including multi-source to multi-target mappings and asymmetric connections, which pose challenges for traditional graphs or simplicial complexes[26]. Essentially, a topological hyperdigraph consists of sequences of distinct elements from a finite set, known as directed hyperedges. Figure 4c provides examples of directed hyperedges of varying dimensions. These sequences share similarities with simplices in a simplicial complex (Fig. 4b). For detailed definitions of common graph, simplicial complex and hypergraph concepts, refer to Supplementary Information Section 3. A hyperdigraph $\mathcal{H}$ comprises a vertex set V and a collection of sequences with distinct elements in V. A sequence of length $k+1$ is called a k-directed hyperedge, mathematically represented as an inclusion map $e:[k]\to V$, where $[k]=\{0,1,\ldots,k\}$. A hyperdigraph is essentially a collection of directed hyperedges on V, sometimes denoted as $\overrightarrow{\mathcal{H}}=(V,\mathbf{E})$, with $\mathbf{E}$ representing the set of directed hyperedges. Notably, hyperdigraphs can be reduced to hypergraphs when the set V and all directed hyperedges are ordered, and to directed graphs when all directed edges are restricted to one dimension. This versatility positions hyperdigraphs as powerful aggregators, enabling flexible and diverse data representation.

More formally, let G be an abelian group, and 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

$$\partial_k(x_0,x_1,\ldots,x_k)=\sum_{i=0}^{k}(-1)^k(x_0,\ldots,\hat{x}_i,\ldots,x_k). \quad (1)$$

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

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

Then, $\Omega_k(\overrightarrow{\mathcal{H}};G)$ is also a chain complex, specifically tailored for exploring the topology of hyperdigraphs. It is essential to highlight that the chain complex $\Omega_k(\overrightarrow{\mathcal{H}};G)$ undergoes simplification when the

https://doi.org/10.1038/s42256-024-00855-1

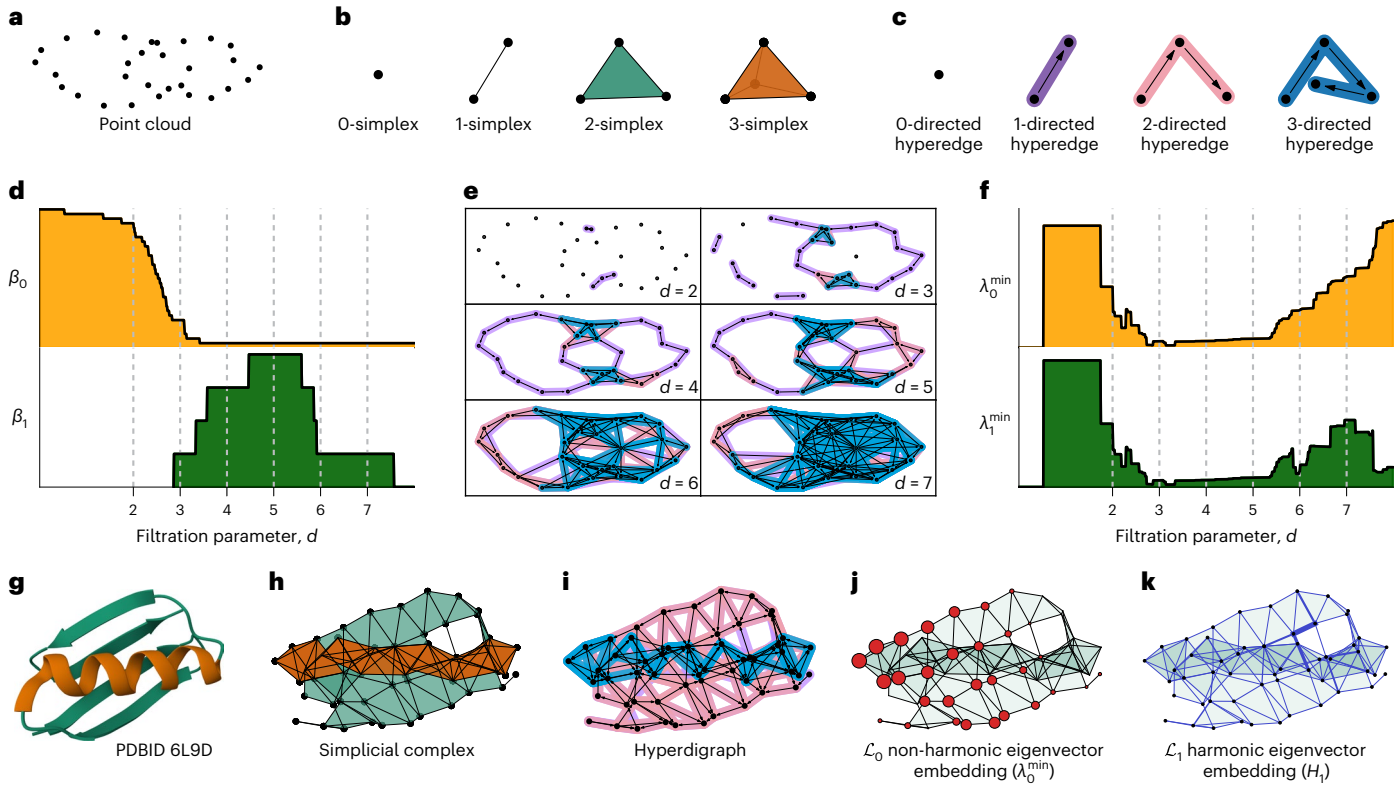


Fig. 4 | Illustration of the concepts related to topological sequence embedding. a, Representation of structural data as a point cloud. **b**, Depiction of 0-simplex (node), 1-simplex (edge), 2-simplex (triangle) and 3-simplex (tetrahedron), which serve as the fundamental building blocks of a simplicial 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, that is, 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 non-zero minimum non-harmonic spectra of the PTHL 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 Cα atoms. The alpha helix is highlighted in orange, while the beta helix is shown in green. **h**, Illustrations of simplicial complex representation for the Cα atoms of protein 6L9D at a cutoff distance of $d = 5$ Å. The 2-simplices are filled by green, 3-simplices are coloured by orange. **i**, Visualizations of hyperdigraph representations for the Cα 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 $\mathcal{L}_0$ non-zero smallest non-harmonic eigenvector embedding for the Cα atoms of protein 6L9D at a cutoff distance of $d = 5$ Å. **k**, Explanation of the $\mathcal{L}_1$ harmonic eigenvector embedding for the edges between the Cα atoms of protein 6L9D at a cutoff distance of $d = 5$ Å.

hyperdigraph is transformed back into a simplicial complex or hypergraph. The corresponding simplicial complex representation of Cα atoms in protein 6L9D is depicted in Fig. 4h. Here, blue triangles represent the 2-simplices, while orange highlights designate the 3-simplices, providing a rough visualization of the alpha helix structures. Additionally, Fig. 4i illustrates the 3-directed hyperedges within the hyperdigraph, highlighted in blue, serving as an alternative representation of the alpha helix in the structure. Supplementary Fig. 9 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 Section 3 and the original paper[26].

Furthermore, to enable the practical application of hyperdigraphs in protein–ligand complex analysis, we introduce Vietoris–Rips (VR) and alpha hyperdigraphs. These hyperdigraphs are inspired by the widely used Vietoris–Rips complex and alpha complex topological models, respectively. All analyses in this work utilize the VR hyperdigraph unless otherwise specified. For illustrative purposes, Supplementary Figs. 1 and 2 depict VR and alpha hyperdigraphs, respectively. Detailed construction methods and definitions are provided in Supplementary Information Section 4.

Topological Laplacians and spectrum analysis. The combinatorial Laplacian is a cornerstone tool in discrete geometry and algebraic topology, offering insights into the structure of topological systems like simplicial complexes, hypergraphs and hyperdigraphs. Just as the graph Laplacian analyses graph properties (considering graphs as 1-simplices), the combinatorial Laplacian extends this analysis to higher-dimensional structures. Eigenvalues of the graph Laplacian encode connectivity information. For instance, the second smallest eigenvalue (Fiedler vector) reflects algebraic connectivity, while the smallest positive eigenvalue (spectral gap) relates to the Cheeger constant. The collection of eigenvalues forms the Laplacian spectrum. Interestingly, the graph Laplacian matrix ($\mathcal{L} = D - A$, where D is the degree matrix and A is the adjacency matrix) can be expressed as $\mathcal{L} = B_1 B_1^{\mathrm{T}}$ when considering the graph as a 1-dimensional simplicial complex and B_1 as the one-dimensional boundary operator matrix. This observation inspires the generalization of the Laplacian operator to higher dimensions using boundary operators, 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

$$\mathcal{L}_k = B_{k+1}B_{k+1}^{\mathrm{T}} + B_k^{\mathrm{T}}B_k. \tag{3}$$

Here, B_k^{T} denotes the transpose matrix of B_k. The term $B_k^{\mathrm{T}}B_k$ indicates the connectivity arising from the intersections of k-simplices at $(k-1)$-simplices, while the term $B_{k+1}B_{k+1}^{\mathrm{T}}$ 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 $\Omega_\cdot$ 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

$$\Delta_k = \partial_{k+1} \circ \partial_{k+1}^* + \partial_k^* \circ \partial_k. \tag{4}$$

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

$$\mathcal{L}_k = B_{k+1}B_{k+1}^{\mathrm{T}} + B_k^{\mathrm{T}}B_k, \tag{5}$$

where B_k is the representation matrix of boundary operator ∂_k by left multiplication[59]. This combinatorial Laplacian is a generalization of the graph Laplacian, which is just a carve-out of the properties of graphs (that is, 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, that is, the harmonic spectrum, encode the topological information. While the non-zero eigenvalues (the non-harmonic spectrum) encode the geometric information about the system. Figure 4j shows the $\mathcal{L}_0$ non-zero smallest non-harmonic eigenvector embedding for the Cα atoms (that is, 0-simplices in the simplicial complex) of protein 6L9D at a cutoff distance of $d = 5$ Å. And Fig. 4k shows the $\mathcal{L}_1$ harmonic eigenvector embedding for the edges (that is, 1-simplices in the simplicial complex) between the Cα atoms of protein 6L9D at a cutoff distance of $d = 5$ Å. Specifically, for $\mathcal{L}_k$, the multiplicity of the zero eigenvalue (that is, the number of times 0 appears as an eigenvalue) equals the number of independent cycles; it also equals the topological invariant (β_k) in the k-dimensional space[60]. For example, multiplicity of zero for $\mathcal{L}_0$ (that is, β_0) is the number of connected components in the graph (1-simplicial complex), the multiplicity of zero for $\mathcal{L}_1$ (that is, β_1) is the number of circles, and it means the number of cavities for $\mathcal{L}_2$. The largest eigenvalue $\lambda_k^{\max}$ 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, $0 \le \lambda_k^{\max} \le 2d_k$. The smallest non-zero eigenvalue for $\mathcal{L}_k$, also known as spectral gap, denoted as $\lambda_k^{\min}$, reflects the geometric structure of the system. In this work, the multiplicity of zero, the average value, the s.d., 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 Supplementary Fig. 10. 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[61] and a discrete point cloud setting[24] in 2019. A filtration process is essential to achieving the multiscale representation in persistent Laplacians[24,26,62] as well as in persistent homology[27,63]. The choice of the filtration (scale) parameter, denoted as d, varies based on the data structure in question: for point cloud data (Fig. 4a), it is often the sphere radius (or diameter). By systematically adjusting d, one can derive a sequence of hierarchical representations, illustrated in Fig. 1a. 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 VR complex, Čech complex and alpha complex. The VR complex is used in this work. Mathematically, the nested simplicial complexes can be written as:

$$\varnothing \subseteq K_{d_0} \subseteq K_{d_1} \subseteq \cdots \subseteq K_{d_n} = K \tag{6}$$

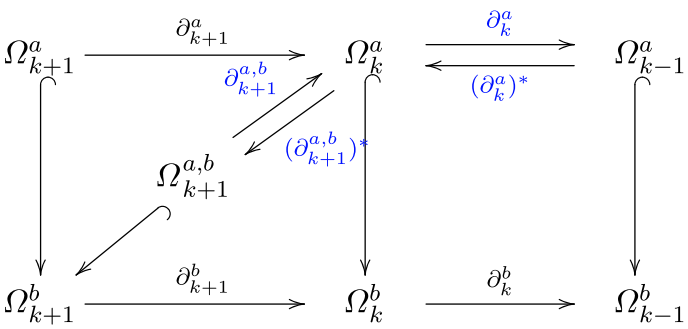
Here, for any two $d_i < d_j$, we have $K_{d_i} \subseteq K_{d_j}$. The concept extends to hyperdigraphs as well, namely the VR hyperdigraph: one can form nested hyperdigraphs by properly defining directed hyperedges[26]. To visualize the effects of changing filtration parameters, Fig. 4e depicts alterations in point cloud connectivity from Fig. 4a, leading to a sequence of hyperdigraphs. Additionally, Supplementary Fig. 8a showcases the simplicial complex produced at different filtration parameters and Supplementary Fig. 8b illustrates hyperdigraphs generated at different filtration parameters. The details about the construction of a VR hyperdigraph can be seen in Supplementary Fig. 1. In addition, inspired by the alpha complex, the alpha hyperdigraph is also introduced in this work, as shown in Supplementary Fig. 2.

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:

$$\cdots \to C_{k+1}(K_{d_i}; G) \xrightarrow{\partial_{k+1}^{d_i}} C_k(K_{d_i}; G) \xrightarrow{\partial_k^{d_i}} C_{k-1}(K_{d_i}; G) \to \cdots \tag{7}$$

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 $a \le b$, 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 obtained 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 $x \in \Omega_{k+1}^{a,b}$.



The kth persistent Laplacian is defined as

$$\Delta_k^{a,b} = \partial_{k+1}^{a,b} \circ \left(\partial_{k+1}^{a,b}\right)^* + \left(\partial_k^a\right)^* \circ \partial_k^a. \quad (9)$$

Here, $\left(\partial_{k+1}^{a,b}\right)^*$ and $\left(\partial_k^a\right)^*$ are adjoint operators of $\partial_{k+1}^{a,b}$ and ∂_k^a, respectively. It is worth noting that the harmonic part of $\Delta_k^{a,b}$, that is, $\ker \Delta_k^{a,b} = \{x \in \Omega_k^a | \Delta_k^{a,b} x = 0\}$, is naturally isomorphic to the (a, b)-persistent homology $H_k^{a,b} = \mathrm{im}(H_k(\Omega_*^a) \to H_k(\Omega_*^b))$ (ref. 64). 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 Laplacian 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 4d illustrates the evolution of zero eigenvalue multiplicities in the associated Laplacian matrix as the filtration (scale) parameters change, while Fig. 4f depicts the variation in the smallest positive eigenvalue with changing filtration (scale) parameters. Additional persistent attributes are presented in Supplementary Fig. 7.

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[22] is used to characterize protein–ligand interactions.

When analysing ligands, the focus is on heavy elements such as carbon (C), nitrogen (N), oxygen (O), sulfur (S), phosphorus (P), fluorine (F), chlorine (Cl), bromine (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 $\mathcal{E}_{\mathrm{protein}}$ = {{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 $\mathcal{E}_{\mathrm{ligand}}$ = {{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 $K_{\{C,N\},\{S\}}$ 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:

$$D(i,j) = \begin{cases} \left\|\mathbf{r}_i - \mathbf{r}_j\right\|, & \text{if } \mathbf{r}_i \in \mathcal{E}_{\mathrm{protein}}, \mathbf{r}_j \in \mathcal{E}_{\mathrm{ligand}} \text{ or } \mathbf{r}_i \in \mathcal{E}_{\mathrm{ligand}}, \mathbf{r}_j \in \mathcal{E}_{\mathrm{protein}} \\ \infty, & \text{other} \end{cases} \quad (10)$$

where the $\mathbf{r}_i$ and $\mathbf{r}_j$ are coordinates for the ith and jth atoms in the set, and $\left\|\mathbf{r}_i - \mathbf{r}_j\right\|$ 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 PTHL approach.

TopoFormer model

TopoFormer utilizes a topological embedding model to transform 3D protein–ligand complexes into topological sequences characterized by multiscale features. The larger TopoFormer variant employs a scale range of 0 to 10 Å (0.1 Å increments), generating a 100-unit sequence. At each scale, embedded features are represented by a 143 × 6 matrix (6 attributes per $\mathcal{L}_0$). Topological embeddings are combined with trainable multiscale embeddings to produce the final output (Fig. 1a). Convolutional layers within the transformer's encoder and decoder convert these matrices into 1-dimensional vectors (Fig. 1c). TopoFormer's attention mechanism utilizes encoded representations (queries, keys and values) for each filtration increment, similar to conventional transformers. An asymmetric design, inspired by the Masked Autoencoders (MAE) model in computer vision[65], is applied to the encoder and decoder. Detailed model settings are provided in Supplementary Information Section 2. Training involves two phases: (1) Self-supervised learning: 19,513 unlabelled protein–ligand complexes from PDBbind are used to pretrain TopoFormer. Topological embeddings are reconstructed, and the mean squared error serves as the reconstruction loss. This approach allows the model to learn generalized representations of protein–ligand interactions from vast amounts of unlabelled data. (2) Supervised learning: For scoring, ranking, docking and screening tasks, TopoFormer is fine-tuned to predict specific scores for protein–ligand complexes, again using MAE as the loss function.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

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 v.2020. 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. All data used in this study can be downloaded from the official PDBbind website: http://www.pdbbind.org.cn/index.php. We also provide a comprehensive set of resources at https://github.com/WeilabMSU/TopoFormer. This includes topological embedded features used in both TopoFormer and TopoFormer$_s$, sequence-based features derived from the Transformer-CPZ[28] and ESM[33] models and all additional generated poses with their associated scores, which were crucial for the docking and screening tasks. Instructions for accessing the poses are also available via Zenodo at https://doi.org/10.5281/zenodo.10892799 (ref. 66).

Code availability

All source code and models are publicly available via Zenodo at https://doi.org/10.5281/zenodo.10892799 (ref. 66).

References

1. Fleming, N. How artificial intelligence is changing drug discovery. *Nature* **557**, S55–S57 (2018).
2. Lyu, J. et al. Ultra-large library docking for discovering new chemotypes. *Nature* **566**, 224–229 (2019).
3. Kitchen, D. B., Decornez, H., Furr, J. R. & Bajorath, J. Docking and scoring in virtual screening for drug discovery: methods and applications. *Nat. Rev. Drug Discov.* **3**, 935–949 (2004).
4. Pinzi, L. & Rastelli, G. Molecular docking: shifting paradigms in drug discovery. *Int. J. Mol. Sci.* **20**, 4331 (2019).
5. Pagadala, N. S., Syed, K. & Tuszynski, J. Software for molecular docking: a review. *Biophys. Rev.* **9**, 91–102 (2017).
6. Wang, L. 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. *J. Am. Chem. Soc.* **137**, 2695–2703 (2015).

7. Sliwoski, G., Kothiwale, S., Meiler, J. & Lowe, E. W. Computational methods in drug discovery. *Pharmacol. Rev.* **66**, 334–395 (2014).
8. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
9. Baek, M. et al. Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876 (2021).
10. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
11. Song, Y. & Wang, L. Multiobjective tree-based reinforcement learning for estimating tolerant dynamic treatment regimes. *Biometrics* **80**, ujad017 (2024).
12. Luo, J., Wei, W., Waldispühl, J. & Moitessier, N. Challenges and current status of computational methods for docking small molecules to nucleic acids. *Eur. J. Med. Chem.* **168**, 414–425 (2019).
13. Lo, Yu-Chen, Rensi, S. E., Torng, W. & Altman, R. B. Machine learning in chemoinformatics and drug discovery. *Drug Discov. Today* **23**, 1538–1546 (2018).
14. The Atomwise AIMS Program. AI is a viable alternative to high throughput screening: a 318-target study. *Sci. Rep.* **14**, 7526 (2024).
15. Gómez-Sacristán, P., Simeon, S., Tran-Nguyen, V.-K., Patil, S. & Ballester, P. J. Inactive-enriched machine-learning models exploiting patent data improve structure-based virtual screening for PDL1 dimerizers. *J. Adv. Res.* (in the press); https://doi.org/10.1016/j.jare.2024.01.024
16. Hu, X. et al. Discovery of novel non-steroidal selective glucocorticoid receptor modulators by structure-and IGN-based virtual screening, structural optimization, and biological evaluation. *Eur. J. Med. Chem.* **237**, 114382 (2022).
17. Vaswani, A. et al. Attention is all you need. In *NIPS'17: Proc. 31st International Conference on Neural Information Processing Systems* (eds von Luxburg, U. et al.) 6000–6010 (Curran Associates, 2017).
18. Devlin, J., Chang, M. W., Lee, K. & Toutanova, K. B. BERT: pre-training of deep bidirectional transformers for language understanding. In *Proc. 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies* **1**, 4171–4186 (Association for Computational Linguistics, 2019).
19. Ouyang, L. et al. Training language models to follow instructions with human feedback. *Adv. Neural Inf. Process. Syst.* **35**, 27730–27744 (2022).
20. Singh, R., Sledzieski, S., Bryson, B., Cowen, L. & Berger, B. Contrastive learning in protein language space predicts interactions between drugs and protein targets. *Proc. Natl Acad. Sci. USA* **120**, e2220778120 (2023).
21. Saar, K. L. et al. Turning high-throughput structural biology into predictive inhibitor design. *Proc. Natl Acad. Sci. USA* **120**, e2214168120 (2023).
22. Cang, Z., Mu, L. & Wei, G.-W. Representability of algebraic topology for biomolecules in machine learning based scoring and virtual screening. *PLoS Comput. Biol.* **14**, e1005929 (2018).
23. Nguyen, D. D., Cang, Z. & Wei, G.-W. A review of mathematical representations of biomolecular data. *Phys. Chem. Chem. Phys.* **22**, 4343–4367 (2020).
24. Wang, R., Nguyen, D. D. & Wei, G.-W. Persistent spectral graph. *Int. J. Numer. Methods Biomed. Eng.* **36**, e3376 (2020).
25. Meng, Z. & Xia, K. Persistent spectral–based machine learning (PerSpect ML) for protein-ligand binding affinity prediction. *Sci. Adv.* **7**, eabc5329 (2021).
26. Chen, D., Liu, J., Wu, J. & Wei, G.-W. Persistent hyperdigraph homology and persistent hyperdigraph Laplacians. *Found. Data Sci.* **5**, 558–588 (2023).
27. Zomorodian, A. & Carlsson, G. Computing persistent homology. *Discrete Comput. Geom.* **33**, 249–274 (2005).
28. Chen, D., Zheng, J., Wei, G.-W. & Pan, F. Extracting predictive representations from hundreds of millions of molecules. *J. Phys. Chem. Lett.* **12**, 10793–10801 (2021).
29. Ruff, K. M. & Pappu, R. V. AlphaFold and implications for intrinsically disordered proteins. *J. Mol. Biol.* **433**, 167208 (2021).
30. Li, Y., Han, L., Liu, Z. & Wang, R. Comparative assessment of scoring functions on an updated benchmark: 2. Evaluation methods and general results. *J. Chem. Inf. Model.* **54**, 1717–1736 (2014).
31. Cheng, T., Li, X., Li, Y., Liu, Z. & Wang, R. Comparative assessment of scoring functions on a diverse test set. *J. Chem. Inf. Model.* **49**, 1079–1093 (2009).
32. Su, M. et al. Comparative assessment of scoring functions: the CASF-2016 update. *J. Chem. Inf. Model.* **59**, 895–913 (2018).
33. Trull, T. J. & Ebner-Priemer, U. W. Using experience sampling methods/ecological momentary assessment (ESM/EMA) in clinical assessment and clinical research: introduction to the special section. *Psychol. Assess.* **21**, 457–462 (2009).
34. Karlov, D. S., Sosnin, S., Fedorov, M. V. & Popov, P. graphDelta: MPNN scoring function for the affinity prediction of protein–ligand complexes. *ACS Omega* **5**, 5150–5159 (2020).
35. Sánchez-Cruz, N., Medina-Franco, J., Mestres, J. & Barril, X. Extended connectivity interaction features: improving binding affinity prediction through chemical description. *Bioinformatics* **37**, 1376–1382 (2021).
36. Wang, Z. et al. Onionnet-2: a convolutional neural network model for predicting protein-ligand binding affinity based on residue-atom contacting shells. *Front. Chem.* **9**, 753002 (2021).
37. Rezaei, M. A., Li, Y., Wu, D., Li, X. & Li, C. Deep learning in drug design: protein-ligand binding affinity prediction. *IEEE/ACM Trans. Comput. Biol. Bioinform.* **19**, 407–417 (2020).
38. Wang, S. et al. Se-onionnet: a convolution neural network for protein–ligand binding affinity prediction. *Front. Genet.* **11**, 607824 (2021).
39. Jones, D. et al. Improved protein–ligand binding affinity prediction with structure-based deep fusion inference. *J. Chem. Inf. Model.* **61**, 1583–1592 (2021).
40. Boyles, F., Deane, C. M. & Morris, G. M. Learning from the ligand: using ligand-based features to improve binding affinity prediction. *Bioinformatics* **36**, 758–764 (2020).
41. Liu, Z. et al. PDB-wide collection of binding data: current status of the PDBbind database. *Bioinformatics* **31**, 405–412 (2015).
42. Wang, C. & Zhang, Y. Improving scoring-docking-screening powers of protein–ligand scoring functions using random forest. *J. Comput. Chem.* **38**, 169–177 (2017).
43. Gentile, F. et al. Automated discovery of noncovalent inhibitors of SARS-Cov-2 main protease by consensus deep docking of 40 billion small molecules. *Chem. Sci.* **12**, 15960–15974 (2021).
44. Méndez-Lucio, O., Ahmad, M., del Rio-Chanona, E. A. & Wegner, J. K. A geometric deep learning approach to predict binding conformations of bioactive molecules. *Nat. Mach. Intell.* **3**, 1033–1039 (2021).
45. Zheng, L. et al. Improving protein–ligand docking and screening accuracies by incorporating a scoring function correction term. *Brief. Bioinform.* **23**, bbac051 (2022).
46. Bao, J., He, X. & Zhang, J. Z. H. DeepBSP—a machine learning method for accurate prediction of protein–ligand docking structures. *J. Chem. Inf. Model.* **61**, 2231–2240 (2021).
47. Shen, C. et al. Boosting protein–ligand binding pose prediction and virtual screening based on residue–atom distance likelihood potential and graph transformer. *J. Med. Chem.* **65**, 10691–10706 (2022).

48. Nguyen, D. D. & Wei, G.-W. AGL-Score: algebraic graph learning score for protein–ligand binding scoring, ranking, docking, and screening. *J. Chem. Inf. Model.* **59**, 3291–3304 (2019).
49. Liu, X., Feng, H., Wu, J. & Xia, K. Dowker complex based machine learning (DCML) models for protein-ligand binding affinity prediction. *PLoS Comput. Biol.* **18**, e1009943 (2022).
50. Tran-Nguyen, V.-K., Junaid, M., Simeon, S. & Ballester, P. J. A practical guide to machine-learning scoring for structure-based virtual screening. *Nat. Protoc.* **18**, 3460–3511 (2023).
51. Moon, S., Zhung, W., Yang, S., Lim, J. & Kim, W. Y. PIGNet: a physics-informed deep learning model toward generalized drug–target interaction predictions. *Chem. Sci.* **13**, 3661–3673 (2022).
52. Tran-Nguyen, V.-K., Bret, G. & Rognan, D. True accuracy of fast scoring functions to predict high-throughput screening data from docking poses: the simpler the better. *J. Chem. Inf. Model.* **61**, 2788–2797 (2021).
53. Tran-Nguyen, V.-K. & Ballester, P. J. Beware of simple methods for structure-based virtual screening: the critical importance of broader comparisons. *J. Chem. Inf. Model.* **63**, 1401–1405 (2023).
54. Tran-Nguyen, V.-K., Simeon, S., Junaid, M. & Ballester, P. J. Structure-based virtual screening for PDL1 dimerizers: evaluating generic scoring functions. *Curr. Res. Struct. Biol.* **4**, 206–210 (2022).
55. Shen, C. et al. A generalized protein–ligand scoring framework with balanced scoring, docking, ranking and screening powers. *Chem. Sci.* **14**, 8129–8146 (2023).
56. Jones, G., Willett, P., Glen, R. C., Leach, A. R. & Taylor, R. Development and validation of a genetic algorithm for flexible docking. *J. Mol. Biol.* **267**, 727–748 (1997).
57. Trott, O. & Olson, A. J. AutoDock Vina: improving the speed and accuracy of docking with a new scoring function, efficient optimization, and multithreading. *J. Comput. Chem.* **31**, 455–461 (2010).
58. Tran-Nguyen, V.-K., Jacquemard, C. & Rognan, D. LIT-PCBA: an unbiased data set for machine learning and virtual screening. *J. Chem. Inf. Model.* **60**, 4263–4273 (2020).
59. Horak, D. & Jost, J. Spectra of combinatorial Laplace operators on simplicial complexes. *Adv. Math.* **244**, 303–336 (2013).
60. Eckmann, B. Harmonische funktionen und randwertaufgaben in einem komplex. *Comment. Math. Helv.* **17**, 240–255 (1944).
61. Chen, J., Zhao, R., Tong, Y. & Wei, G.-W. Evolutionary de Rham-Hodge method. *Discrete Continuous Dyn. Syst. Ser. B.* **26**, 3785–3821 (2021).
62. Mémoli, F., Wan, Z. & Wang, Y. Persistent Laplacians: properties, algorithms and implications. *SIAM J. Math. Data Sci.* **4**, 858–884 (2022).
63. Edelsbrunner, H., Letscher, D. & Zomorodian, A. Topological persistence and simplification. *Discrete Comput. Geom.* **28**, 511–533 (2002).
64. Liu, J., Li, J. & Wu, J. The algebraic stability for persistent Laplacians. Preprint at *arXiv* https://doi.org/10.48550/arXiv.2302.03902 (2023).
65. He, K. et al. Masked autoencoders are scalable vision learners. In *Proc. 2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 15979–15988 (IEEE, 2022).
66. Chen, D. WeilabMSU/TopoFormer: TopoFormer. *Zenodo* https://doi.org/10.5281/zenodo.10892799 (2024).
67. Sunseri, J. & Koes, D. R. Virtual screening with Gnina 1.0. *Molecules* **26**, 7369 (2021).
68. Yang, C. & Zhang, Y. Delta machine learning to improve scoring-ranking-screening performances of protein–ligand scoring functions. *J. Chem. Inf. Model.* **62**, 2696–2712 (2022).
69. Wójcikowski, M., Kukiełka, M., Stepniewska-Dziubinska, M. M. & Siedlecki, P. Development of a protein–ligand extended connectivity (PLEC) fingerprint and its application for binding affinity predictions. *Bioinformatics* **35**, 1334–1341 (2019).
70. Stepniewska-Dziubinska, M. M., Zielenkiewicz, P. & Siedlecki, P. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. *Bioinformatics* **34**, 3666–3674 (2018).

Acknowledgements

This work was supported in part by NIH grant nos. R01GM126189, R01AI164266 and R35GM148196, National Science Foundation grant nos. DMS2052983 and IIS-1900473, Michigan State University Research Foundation, and Bristol-Myers Squibb grant no. 65109. The work of J.L. was performed while visiting Michigan State University.

Author contributions

D.C. designed the project, modified the method, wrote the code, performed computational studies, wrote the first draft and revised the manuscript. J.L. wrote the methods section and revised the manuscript. G.-W.W. conceptualized and supervised the project, acquired funding and revised the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s42256-024-00855-1.

Correspondence and requests for materials should be addressed to Jian Liu or Guo-Wei Wei.

Peer review information *Nature Machine Intelligence* thanks Emil Alexov, Pedro Ballester and the other, anonymous, reviewer(s) for their contribution to the peer review of this work.

nature portfolio

Corresponding author(s): Jian Liu, Guowei Wei

Last updated by author(s): May 11, 2024

Reporting Summary

Nature Portfolio wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Portfolio policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☐	☒	The exact sample size (*n*) for each experimental group/condition, given as a discrete number and unit of measurement
☐	☒	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	☐	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	☐	A description of all covariates tested
☒	☐	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	☐	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	☐	For null hypothesis testing, the test statistic (e.g. *F*, *t*, *r*) with confidence intervals, effect sizes, degrees of freedom and *P* value noted *Give P values as exact values whenever suitable.*
☒	☐	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	☐	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	☐	Estimates of effect sizes (e.g. Cohen's *d*, Pearson's *r*), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection

1. No sofeware used for this data. This study utilized the PDBbind databases, including CASF-2007, CASF-2013, PDBbind v2015, PDBbind 2016, CASF-2016, and PDBbind v2020. All of the data mentioned in this study are publicly accessible on the official PDBbind website: http://www.pdbbind.org.cn/index.php.

2. No sofeware used for this data. The additional poses generated and their corresponding scores, crucial for the docking and screening tasks, are available from the following source: https://weilab.math.msu.edu/AGL-Score.

3. Autodock Vina generated the docking poses for the LIT-PCBA benchmark dataset, producing up to 10 poses per compound with an energy range of 3 and an exhaustiveness of 10. The pose with the highest binding affinity score was chosen for prediction by our model, resulting in a total of 2,651,977 protein-ligand complexes. The generated data can be downloaded from https://github.com/WeilabMSU/TopoFormer

Data analysis	1. This study utilized the PDBbind databases, including CASF-2007, CASF-2013, PDBbind v2015, PDBbind 2016, CASF-2016, and PDBbind v2020. All of the data mentioned in this study are publicly accessible on the official PDBbind website: http://www.pdbbind.org.cn/index.php. 2. The additional poses generated and their corresponding scores, crucial for the docking and screening tasks, are available from the following source: https://weilab.math.msu.edu/AGL-Score. 3. Autodock Vina generated the docking poses for the LIT-PCBA benchmark dataset, producing up to 10 poses per compound with an energy range of 3 and an exhaustiveness of 10. The pose with the highest binding affinity score was chosen for prediction by our model, resulting in a total of 2,651,977 protein-ligand complexes. The generated data can be downloaded from https://github.com/WeilabMSU/TopoFormer

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Portfolio guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A description of any restrictions on data availability
- For clinical datasets or third party data, please ensure that the statement adheres to our policy

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: \linebreak http://www.pdbbind.org.cn/index.php. We also provide a comprehensive set of resources at https://github.com/WeilabMSU/TopoFormer. This includes topological embedded features used in both TopoFormer and TopoFormer$_s$, sequence-based features derived from the Transformer-CPZ \cite{chen2021extracting} and ESM \cite{trull2009using} models, and all additional generated poses with their associated scores, which were crucial for the docking and screening tasks.

No restrictions on data availability. No clinical datasets were used in this work.

Human research participants

Policy information about studies involving human research participants and Sex and Gender in Research.

Reporting on sex and gender	Not applicable to this work
Population characteristics	Not applicable to this work
Recruitment	Not applicable to this work
Ethics oversight	Not applicable to this work

Note that full information on the approval of the study protocol must also be provided in the manuscript.

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

☒ Life sciences ☐ Behavioural & social sciences ☐ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Life sciences study design

All studies must disclose on these points even when the disclosure is negative.

Sample size	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. 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.

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. 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. Consequently, for both CASF-2007 and CASF-2013, there was a total of 365,000 training poses available for fine-tuning purposes.

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. 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. 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.

Additionally, to ensure an unbiased evaluation, we employed the LIT-PCBA benchmark dataset \cite{tran2020lit}, which comprises 15 targets with a total of 7,955 true actives and 2,644,022 inactives. This dataset's active-to-inactive ratio of approximately 1:1000 closely mirrors real-world virtual screening scenarios. Following established practices and to optimize computational efficiency\cite{shen2023generalized}, we selected the most representative PDB template for each target as the docking target. AutodockVina\cite{trott2010autodock} was used to generate up to 10 docking poses per compound, with an energy range of 3 and exhaustiveness of 10. The pose with the strongest binding affinity score was selected for prediction by our model. This resulted in a total of 2,651,977 protein-ligand complexes. Detailed dataset information is summarized in Supplementary Table \ref{stable:lit_pcba_datainfo}. All posed protein-ligand complexes and associated scores are publicly available at https://github.com/WeilabMSU/TopoFormer.

Data exclusions

No data excluded from the analysis.

Replication

All the attempts at replication were successful.

1. To enhance robustness, we trained 20 distinct topological transformers for each dataset, each initialized with different random seeds to mitigate initialization-related errors.

2. Subsequently, we developed 20 Gradient Boosting Regressor Tree (GBRT) models exclusively utilizing sequence-based features.

3. We randomly selected 10 models from the set described in statements 1 and 2. The experimental prediction is derived from the consensus of these 20 predictions. The final result discussed in the paper is obtained by averaging the outcomes over 400 repetitions.

Randomization

Different random seeds were used during the replications for training the models, including the GBRT models and TopoFormer models.

Blinding

This work does not include wet experiments for binding tests. Predictions of binding affinities by the proposed model (for all replications) are available at https://github.com/WeilabMSU/TopoFormer.

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	☐ Antibodies
☒	☐ Eukaryotic cell lines
☒	☐ Palaeontology and archaeology
☒	☐ Animals and other organisms
☒	☐ Clinical data
☒	☐ Dual use research of concern

Methods

n/a	Involved in the study
☒	☐ ChIP-seq
☒	☐ Flow cytometry
☒	☐ MRI-based neuroimaging