PLOS COMPUTATIONAL BIOLOGY

RESEARCH ARTICLE

E(3) equivariant graph neural networks for robust and accurate protein-protein interaction site prediction

Rahmatullah Roche, Bernard Moussad, Md Hossain Shuvo, Debswapna Bhattacharya

Department of Computer Science, Virginia Tech, Blacksburg, Virginia, United States of America

* dbhattacharya@vt.edu

Citation: Roche R, Moussad B, Shuvo MH, Bhattacharya D (2023) E(3) equivariant graph neural networks for robust and accurate protein-protein interaction site prediction. PLoS Comput Biol 19(8): e1011435. https://doi.org/10.1371/journal.pcbi.1011435

Editor: Jinyan Li, University of Technology Sydney, AUSTRALIA

Received: March 10, 2023

Accepted: August 15, 2023

Published: August 31, 2023

Data Availability Statement: Data availability: The raw data used in this study, including the datasets for train, test and validation are collected from publicly available sources and freely available at https://github.com/biomed-AI/GraphPPIS. Code availability: An open-source software implementation of EquiPPIS, licensed under the GNU General Public License v3, is freely available at https://github.com/Bhattacharya-Lab/EquiPPIS.

Funding: This work was partially supported by the National Institute of General Medical Sciences

Abstract

Artificial intelligence-powered protein structure prediction methods have led to a paradigm-shift in computational structural biology, yet contemporary approaches for predicting the interfacial residues (i.e., sites) of protein-protein interaction (PPI) still rely on experimental structures. Recent studies have demonstrated benefits of employing graph convolution for PPI site prediction, but ignore symmetries naturally occurring in 3-dimensional space and act only on experimental coordinates. Here we present EquiPPIS, an E(3) equivariant graph neural network approach for PPI site prediction. EquiPPIS employs symmetry-aware graph convolutions that transform equivariantly with translation, rotation, and reflection in 3D space, providing richer representations for molecular data compared to invariant convolutions. EquiPPIS substantially outperforms state-of-the-art approaches based on the same experimental input, and exhibits remarkable robustness by attaining better accuracy with predicted structural models from AlphaFold2 than what existing methods can achieve even with experimental structures. Freely available at https://github.com/Bhattacharya-Lab/EquiPPIS, EquiPPIS enables accurate PPI site prediction at scale.

Author summary

Predicting how proteins interact and characterizing the interacting residues at the protein-protein interaction interface (i.e., sites) is of central importance to understanding various biological processes actuated by protein-protein interactions (PPI). Despite the remarkable recent progress in protein structure prediction driven by artificial intelligence, existing approaches for PPI site prediction still rely on experimental input. This paper presents an E(3) equivariant graph neural network approach for PPI site prediction that takes into account symmetries naturally occurring in 3-dimensional space and transforms equivariantly with translation, rotation, and reflection. Rigorous experimental validation shows that our method attains substantially improved accuracy and robustness over the existing approaches. Moving beyond what is currently possible with only experimental input, our method enables large-scale PPI site prediction using predicted structural models from AlphaFold2 without compromising on accuracy. An open-source software

(R35GM138146 to D.B.) and the National Science Foundation (DBI2208679 to D.B.). The funders had no role in study design, data collection and analysis, decision to publish, or preparation of the manuscript.

Competing interests: The authors have declared that no competing interests exist.

implementation of EquiPPIS, licensed under the GNU General Public License v3, is freely available at https://github.com/Bhattacharya-Lab/EquiPPIS.

Introduction

Protein-protein interactions (PPI) underpin numerous biological processes [1, 2]. Despite their importance, experimental characterization of PPI remains challenging due to the costly and time-consuming nature of the experimental assays [3]. Computational methods offer a cheaper and high-throughput alternative by predicting the bound complex structures of interacting proteins from the sequences and/or the unbound structures of individual protein chains. A closely related problem—and the one addressed in this study—is the prediction of the PPI sites, which are the interfacial residues of the interacting protein chains.

Accurately predicting the interface of interacting proteins and the identification of the PPI sites remain challenging even after decades of research [4–6]. Various methods have been proposed, but with limited success. Partner-independent PPI site prediction [7–12], which involves the prediction of putative interaction sites based only upon the surface of an isolated protein, without any knowledge of the partner or complex, is even more challenging compared to partner-aware PPI site prediction [13–17] due to the absence of any information about the partner protein and auxiliary information on the complex interfaces. In this work, we focus on partner-independent PPI site prediction.

Predicting how proteins interact, and in particular, predicting the PPI sites, has a long history [12, 15, 18–25]. While initial models focused on feature engineering with machine learning [7, 19, 26, 27], subsequent work sought to capture more complex patterns using deep learning [8, 10, 13, 14, 17]. The vast majority of the existing methods rely on readily available protein sequence information, but their predictive accuracies are often quite limited [28]. Structure-based methods that integrate known structural information from the Protein Data Bank (PDB [29]) are usually more accurate. However, these approaches are limited by the paucity of experimentally solved protein structures in the PDB. In the 14th edition of the Critical Assessment of Structure Prediction (CASP14) experiment, AlphaFold2 [30] attained an unprecedented performance level, enabling highly accurate prediction of single-chain protein structural models at proteome-wide scale [31, 32]. Given the recent progress, a natural question arises: can we leverage the predicted structural information by AlphaFold2 for accurate partner-independent PPI site prediction at scale?

In the recent past, representation learning with graph structured data has been prevailing in different applications. In particular, graph neural networks (GNNs) have surged as the major choice for deep graph learning [33–35]. GNNs are permutation equivariant networks that operate on graph structured data, with numerous applications ranging from dynamical systems to conformational energy estimation [36, 37]. However, off-the-shelf GNNs do not take into account symmetries naturally occurring in 3-dimensional space. That is, they ignore the effects of invariance and equivariance with respect to the E(3) symmetry group, i.e., the group of rotations, reflections, and translations in 3D space. The recent E(n) equivariant graph neural networks [38] address this problem by being translation, rotation, and reflection equivariant in 3D space that can be scaled to higher dimensional spaces (E(n)), while preserving permutation equivariance. SE(3) equivariant neural networks [39] are another recent graph-based models that can deal with the absolute coordinate systems in 3D space, but SE(3) equivariant models do not commute with reflections of the input. E(3) equivariant neural networks, on the other hand, transform equivariantly with translation, rotation and reflections, which make

them suitable for molecular data where chirality of the molecules is often important, such as proteins, particularly when predicted protein structures are used as input that may contain mirror images. As such, E(3) equivariant graph neural networks offer an elegant choice for partner-independent PPI site prediction, where the input consists of a 3D structure of an isolated protein, known to be involved in PPIs, but where the structure of the partner or complex is not known. Being designed from geometric first-principles, symmetry-aware models such as E(3) equivariant graph neural networks are highly suitable for 3D molecular data, providing richer representation while avoiding expensive data augmentation strategies.

The contribution of the present work is the introduction of a symmetry-aware PPI site prediction method, EquiPPIS, built on E(3) equivariant graph neural networks that yields state-of-the-art accuracy by substantially outperforming existing approaches based on the same experimental input. What is more striking is that EquiPPIS attains better accuracy with AlphaFold2-predicted structural models as input than what existing methods can achieve even with experimental input. We directly verify that the performance gains are connected to the unique E(3) equivariant architecture of EquiPPIS. The robustness and performance resilience of our method enable large-scale PPI site prediction without compromising on accuracy.

Results

Overview of the E(3) equivariant graph neural network architecture for protein–protein interaction site prediction

Fig 1 illustrates our E(3) equivariant graph neural network model for partner-independent PPI site prediction. Different from the recent structure-aware graph learning approaches for PPI site prediction [10] that only exploit pairwise distances between all residue pairs (i.e., distance maps) as the spatial information, our E(3) equivariant graph neural network model directly leverages the C coordinates extracted from the input monomer together with sequence- and structure-based node and edge features. By using an E(3) equivariant architecture, our symmetry-aware model can learn to preserve the known transformation properties of 3D coordinates under translation, rotation, and reflection, improving PPI site prediction. The EquiPPIS method consists of three major modules. The first module (Fig 1A) converts the

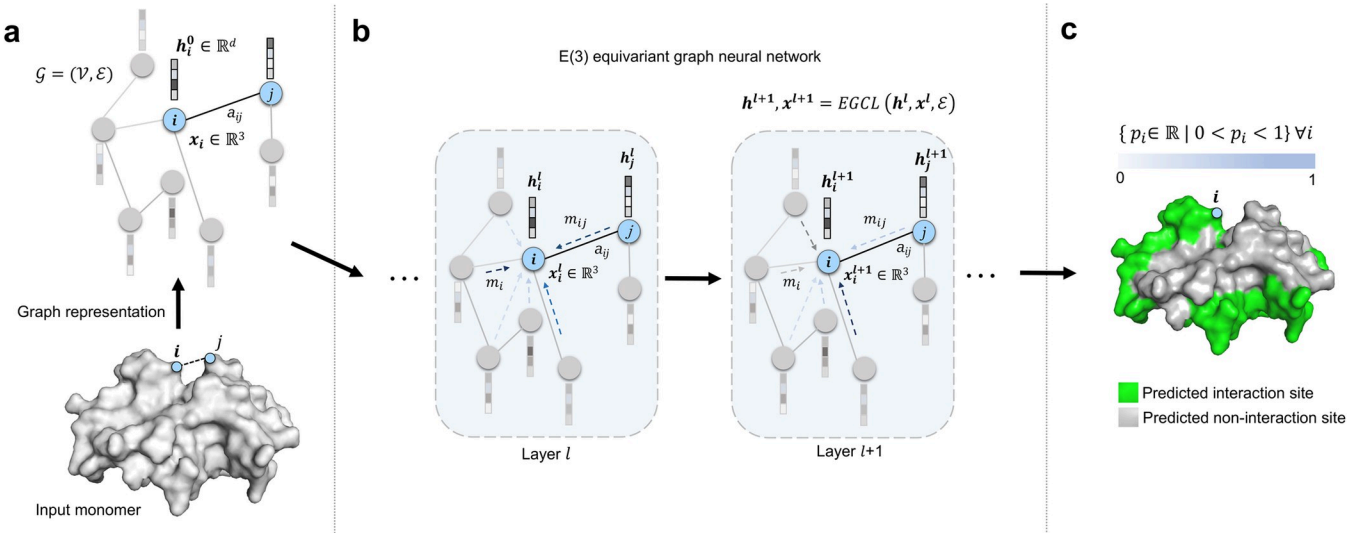

Fig 1. Illustration of the EquiPPIS method for protein–protein interaction site prediction. (**a**) The input protein monomer is converted into an undirected graph. (**b**) Equivariant graph convolutions are then employed on the input graph. (**c**) PPI sites are finally predicted as a graph node classification task.

https://doi.org/10.1371/journal.pcbi.1011435.g001

input protein monomer into an undirected graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, with $\mathcal{V}$ denoting the residues (nodes) and $\mathcal{E}$ denoting the interaction between nonsequential residue pairs according to their pairwise spatial proximity (edges). The spatial proximity between nonsequential residue pairs (i.e., having sequence separation greater or equal to 6) is determined by calculating the Euclidean distances between the C atom of all residue pairs and then setting a cutoff distance of 14Å to obtain the interacting pairs, determined through ablation experiments using an independent validation set. The sequence- and structure-based node and edge features (see the Methods section) are then fed into the second module together with the C coordinates extracted from the input monomer. The second module (**Fig 1B**) is a deep E(3) equivariant graph neural network that conducts a series of transformations of its input through a stack of equivariant graph convolution layer (EGCL) [38], each updating the coordinate and node embeddings using the edge information and the coordinate and node embeddings from the previous layer. Finally, a sigmoidal function is applied to the last EGCL node embedding to predict the probability of every residue in the input monomer to be a PPI site, thereby converting the PPI site prediction into a graph node classification task (**Fig 1C**).

Experiments

For training and performance evaluation, we use a combination of three widely used and publicly available benchmark datasets: Dset_186 [7], Dset_72 [7], and Dset_164 [40], named by the number of proteins in the datasets. We follow the same train and test splits from the recent PPI site prediction method GraphPPIS [10], which combines the aforementioned three datasets and subsequently removes redundancy, resulting in 335 targets as the training set (Train_335) and 60 targets as the test set (Test_60). Details of our training procedure are provided in the Methods section. We evaluate our proposed method on a diverse series of challenging test scenarios using standard performance evaluation metrics including accuracy, precision, recall, F1-score (F1), Matthews correlation coefficient (MCC), area under the receiver operating characteristic curve (ROC-AUC), and area under the precision-recall curve (PR-AUC). First, we demonstrate that EquiPPIS improves upon state-of-the-art accuracy on Test_60 dataset by comparing it directly against a wide variety of existing approaches including PSIVER [7] (based on Naïve Bayes), ProNA2020 [41] (based on neural network), SCRIBER [42] (based on two layers of logistic regression), DLPred [43] (based on long short-term memory), DELPHI [9] (based on an ensemble of convolutional and recurrent neural networks), DeepPPISP [8] (based on convolutional neural network), SPPIDER [11] (based on support vector machine, neural network, and linear discriminant analysis), MaSIF-site [44] (based on geometric deep learning), and GraphPPIS (based on deep graph convolutional network). Among the competing methods, PSIVER, ProNA2020, SCRIBER, DLPred, and DELPHI are purely sequence-based methods; whereas DeepPPIS, SPPIDER, MaSIF-site, GraphPPIS, and our new method EquiPPIS additionally integrate structural information. Next, we examine the reasons for such high performance attained by EquiPPIS and verify that it is indeed connected to the equivariant nature of the model used. To broaden the applicability of our method beyond predicting PPI sites based only on experimentally-solved input monomers extracted from the bound complex structures, we explore a number of extensions including using unbound experimental structures and computationally predicted structural models as input. Specifically, using a subset of 31 proteins from the Test_60 dataset with known unbound monomeric structures in the PDB as an additional unbound test set (hereafter called UBtest_31), we show that EquiPPIS exhibits remarkable robustness and performance resilience compared to the existing approaches. Furthermore, by replacing the experimental input monomers with AlphaFold2 predicted structural models for the Test_60 dataset, we

demonstrate that EquiPPIS attains state-of-the-art accuracy, which is better than what the top performing competing method can achieve even with experimental structures. The superior performance of EquiPPIS even when using predicted structural models as input dramatically enhances the scalability of partner-independent PPI site prediction without compromising on accuracy. Finally, we examine the relative importance of each feature we adopted by conducting feature ablation experiments using an independent validation set consisting of 42 targets (hereafter called Validation_42) collected from the Test_315 dataset of the published work of GraphPPIS after filtering out proteins with >25% pairwise sequence identity with our test sets. We also use this validation set for hyperparameter selection.

Test set performance

We compare EquiPPIS with five sequence-based (PSIVER, ProNA2020, SCRIBER, DLPred and DELPHI) and four structure-based (DeepPPISP, SPPIDER, MaSIF-site, and GraphPPIS) predictors on the Test_60 set. As shown in **Table 1**, in addition to outperforming the sequence-based methods (PR-AUC ranging from 0.190 to 0.319) by a large margin, EquiPPIS significantly improves upon state-of-the-art accuracy by outperforming the structure-based methods. Remarkably, EquiPPIS is the only method attaining ROC-AUC of more than 0.8, which is noticeably better than the closest competing method GraphPPIS. Interestingly, the published work of GraphPPIS sets the goal of achieving ROC-AUC of 0.8 as a motivation for future work, while acknowledging it as one of the current impediments. In summary, EquiPPIS is a leap forward for partner independent PPI site prediction.

Fig 2 presents two representative examples from the Test_60 dataset comparing the PPI site predictions using EquiPPIS and GraphPPIS. For the first example of a sugar binding protein of Trichosanthes kirilowii (PDB ID: 1GGP, chain A) having length 234 (**Fig 2A**), EquiPPIS correctly predicts majority of the observed PPI sites, attaining Precision, Recall, F1, and MCC of 0.8, 0.545, 0.649, and 0.601, respectively; whereas GraphPPIS fails to predict any correct PPI sites with Precision, Recall, F1 and MCC of 0, 0, 0, -0.185, respectively. The second example is a Hydrolase inhibitor of Triticum aestivum in complex with Bacillus subtilis (PDB ID: 2B42, chain A) having length 364 (**Fig 2B**), where GraphPPIS predicts many false positive PPI sites, resulting in low Precision, Recall, F1 and MCC of 0.105, 0.231, 0.144, and -0.004, respectively. EquiPPIS on the other hand attains reasonably accurate predictive performance having Precision, Recall, F1, and MCC of 0.595, 0.564, 0.579, and 0.53, respectively. In both cases, EquiPPIS predictions are strikingly similar to the experimentally observed PPI sites.

Table 1. PPI site prediction performance on the Test_60 dataset for various methods.

Method	Accuracy	Precision	Recall	F1	MCC	ROC-AUC	PR-AUC
PSIVER	0.561	0.188	0.534	0.278	0.074	0.573	0.190
ProNA2020	0.738	0.275	0.402	0.326	0.176	N/A	N/A
SCRIBER	0.667	0.253	0.568	0.350	0.193	0.665	0.278
DLPred	0.682	0.264	0.565	0.360	0.208	0.677	0.294
DELPHI	0.697	0.276	0.568	0.372	0.225	0.699	0.319
DeepPPISP	0.657	0.243	0.539	0.335	0.167	0.653	0.276
SPPIDER	0.752	0.331	0.557	0.415	0.285	0.755	0.373
MaSIF-site	0.780	0.370	0.561	0.446	0.326	0.775	0.439
GraphPPIS	0.776	0.368	0.584	0.451	0.333	0.786	0.429
EquiPPIS	**0.787**	**0.389**	**0.615**	**0.477**	**0.366**	**0.805**	**0.467**

Note: Except EquiPPS, results for the other methods are obtained directly from the published work of GraphPPIS; values in bold represent the best performance.

https://doi.org/10.1371/journal.pcbi.1011435.t001

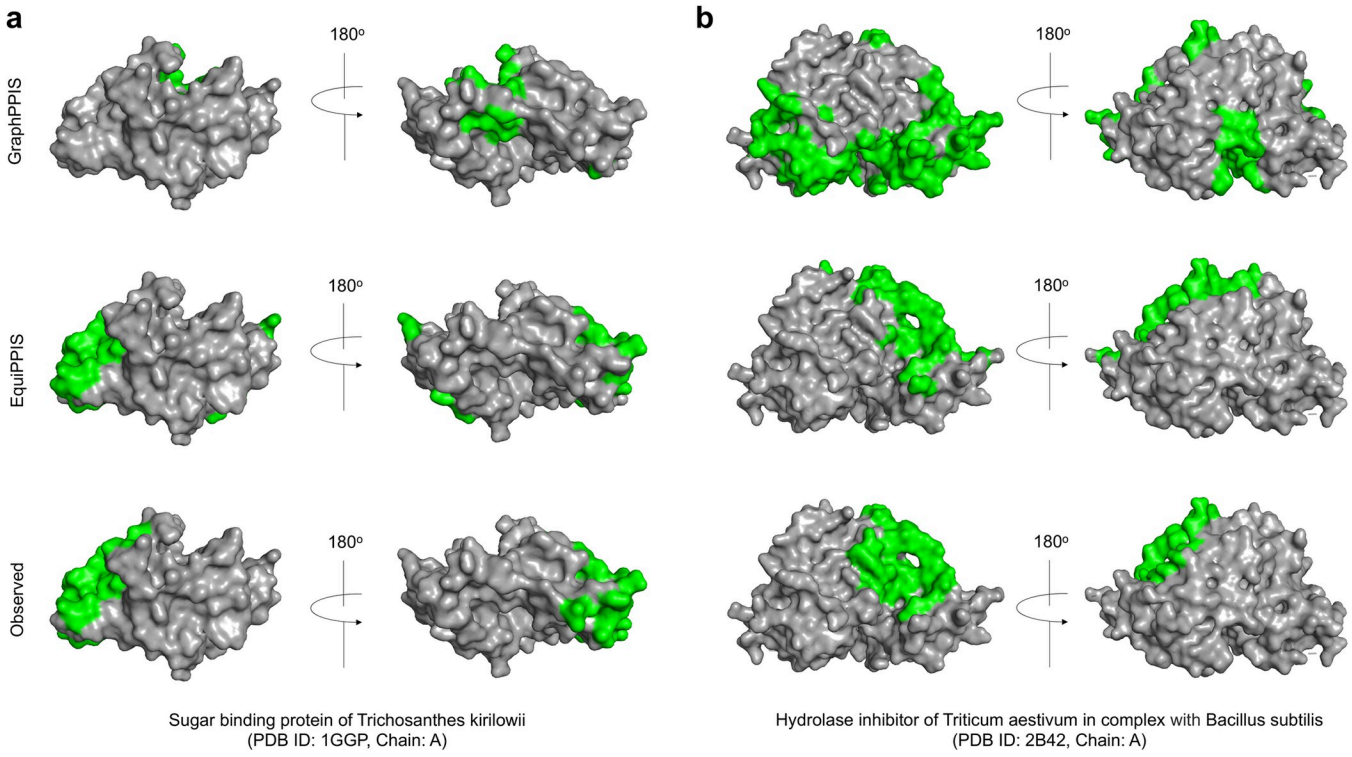

Fig 2. GraphPPIS and EquiPPIS predictions compared to the experimental observation. (**a**) Sugar binding protein of Trichosanthes kirilowii. (**b**) Hydrolase inhibitor of Triticum aestivum in complex with Bacillus subtilis. The regions highlighted in green represent PPI sites.

https://doi.org/10.1371/journal.pcbi.1011435.g002

Analyzing the importance of equivariance

In the above experiments, EquiPPIS exhibits significantly improved performance. In order to gain insight into the reasons behind such high performance and verify that it is connected to the equivariant nature of the model, we perform a series of experiments by gradually isolating the effect of the equivariant graph convolutions used in EquiPPIS. In particular, we train several baseline models and compare them head-to-head with the full-fledged version of EquiPPIS. First, we train a baseline network by turning off the coordinate updates of the equivariant graph convolution layers, thus making it an invariant network (hereafter called 'EquiPPIS invariant'). Since the full-fledged version of EquiPPIS employs attention operations for aggregated embedding as part of the equivariant message passing, we train another baseline network where attention operation is turned off during equivariant message passing, resulting in an equivariant network but without attention (hereafter called 'EquiPPIS w/o attention'). Additionally, we train two off-the-shelf GNNs for PPI site prediction: graph convolution network (GCN) [35] and graph attention network (GAT) [45]. All baseline networks are trained on the same Train_335 dataset using the same set of input features and hyperparameters as the full-fledged version of EquiPPIS (see the Methods section). **Fig 3A to 3D** show the performance of EquiPPIS compared to the baseline networks on the Test_60 set. The results demonstrate that the full-fledged version of EquiPPIS outperforms all baseline models. For example, we observe that the full-fledged version of EquiPPIS attains an ROC-AUC of more than 0.8, which is the best accuracy compared to all baseline models. The 'EquiPPIS invariant' baseline, however, falls short of achieving an ROC-AUC of 0.8, suggesting that it is the equivariant nature of EquiPPIS that is responsible for the accuracy gain. Turning off the attention operation as done

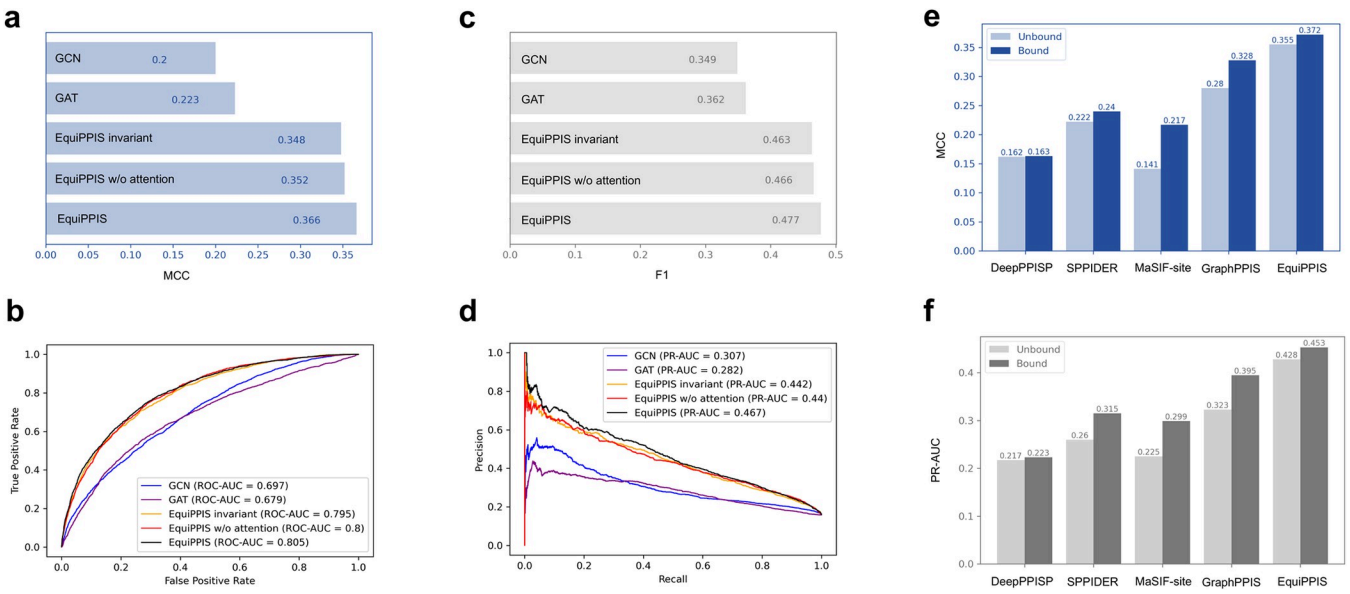

Fig 3. Performance analysis on Test_60 set (**a**-**d**) and on UBtest_31 set (**e**-**f**). (**a**) MCC, (**b**) ROC-AUC, (**c**) F1, and (**d**) PR-AUC of EquiPPIS on Test_60 set compared to the baseline models 'EquiPPIS invariant', 'EquiPPIS w/o attention', graph convolution network (GCN), and graph attention network (GAT). (**e**) MCC and (**f**) PR-AUC of EquiPPIS on UBtest_31 set compared to other structure- based methods GraphPPIS, MaSIF-site, SPPIDER, and DeepPPISP.

https://doi.org/10.1371/journal.pcbi.1011435.g003

in the 'EquiPPIS w/o attention' baseline leads to an accuracy decline (an ROC-AUC of 0.8) compared to the full-fledged version of EquiPPIS, but still better than the invariant network. That is, attention operation during equivariant message passing contributes to an improvement in accuracy. It is worth noting that despite the accuracy drop from the full-fledged version of EquiPPIS, both 'EquiPPIS invariant' and 'EquiPPIS w/o attention' baselines outperform GraphPPIS. On the other hand, off-the-shelf GCN- and GAT-based baselines exhibit much lower accuracies compared to GraphPPIS, let alone EquiPPIS. Overall, the results underscore the importance of equivariance in particular and symmetry-aware nature of the new EquiPPIS model in general for improved predictive accuracy.

In addition to prediction accuracy, robustness of model is another key aspect to consider. While experimentally-solved bound complex structures are used during EquiPPIS training, protein–protein binding often leads to conformational changes by "induced fit" mechanism (binding first) or "conformational selection" (conformational change first) [46]. To evaluate the robustness of EquiPPIS and the effect of conformational changes, we examine the impact on the accuracy when unbound structures are used during prediction instead of their bound states for EquiPPIS as well as the other structure-based PPI site predictors (DeepPPISP, SPPIDER, MaSIF-site, and GraphPPIS) using the unbound test set (UBtest_31) of 31 proteins. As shown in **Fig 3E and 3F**, EquiPPIS outperforms all other methods by a large margin, while having the least impact on accuracy when unbound structures are used during prediction. For example, the closest competing methods MaSIF-site and GraphPPIS suffer from significant accuracy drop both in terms of MCC (35%, and 14.6% drop, respectively) and PR-AUC (24.7%, and 18.2% drop, respectively), whereas EquiPPIS experiences only 4.6%, and 5.5% drop in MCC and PR-AUC, respectively. What is most striking is that the accuracy gap between EquiPPIS and the competing methods is so large that EquiPPIS using unbound structures attains much better accuracy even when the competing methods are using the bound structures. That is, EquiPPIS exhibits remarkable robustness and performance resilience compared to existing approaches.

Table 2. Performance comparison between EquiPPIS and GraphPPIS with experimental input and AlphaFold2 predicted structural models for the Test_60 dataset.

Method	Input type	Accuracy	Precision	Recall	F1	MCC	ROC-AUC	PR-AUC
EquiPPIS	Experimental	**0.787**	**0.389**	**0.615**	**0.477**	**0.366**	**0.805**	**0.467**
	AlphaFold2	0.780	0.379	0.615	0.469	0.356	0.795	0.451
GraphPPIS	Experimental*	0.776	0.368	0.584	0.451	0.333	0.786	0.429
	AlphaFold2	0.767	0.357	0.590	0.445	0.324	0.772	0.399

*Results obtained directly from the published work of GraphPPIS; values in bold represent the best performance.

https://doi.org/10.1371/journal.pcbi.1011435.t002

Beyond experimental input: state-of-the-art performance with AlphaFold2

EquiPPIS achieves state-of-the-art accuracy with experimental structures as input in both bound and unbound states. A natural question to ask is can we achieve similar predictive accuracy when computationally predicted structural models are used as input instead of experimental structures? Given the exceptional performance of AlphaFold2 in the CASP14 experiment and the open availability of the AlphaFold2 protocol, it is now possible to predict single-chain protein structural models from the amino acid sequence with high degree of accuracy. In principle, a robust method such as EquiPPIS should be able to generalize when predicted structural models are used without significant drop in accuracy, thereby broadening its applicability beyond experimental input. Motivated by the prospect, we examine the impact on the accuracy by replacing the experimental input with AlphaFold2 predicted structural models for the Test_60 dataset. **Table 2** shows the performance of EquiPPIS compared to the closest competing method GraphPPIS. Remarkably, EquiPPIS using AlphaFold2-predicted structural models attains better accuracy (PR-AUC = 0.451) than GraphPPIS using experimental structures (PR-AUC = 0.429), let alone GraphPPIS using predicted structural models (PR-AUC = 0.399). While there is a performance decline for both methods when switching from experimental input to prediction, the performance drop for EquiPPIS is lower (PR-AUC = 0.016) than that of GraphPPIS (PR-AUC = 0.03). The results demonstrate the generalizability of EquiPPIS, thus opening the possibility of large-scale PPI site prediction by utilizing high-throughput computational prediction without compromising on accuracy.

Significance test

To investigate if our performance improvement is significant or due to chance, we perform significance tests, following the procedures of previous studies [47–49]. Specifically, we randomly sample 70% of the test set (Test_60), and calculate the F1, MCC, ROC-AUC, and PR-AUC scores of EquiPPIS and the closest competing method GraphPPIS with both experimental and AlphaFold2 predicted structures, where the same structures are used as inputs to both the prediction methods. We repeat this 10 times and obtain 10 pairs of scores. If the measurement is normal, which is determined through Anderson-Darling test [50], then paired t-test is used to calculate significance of the measurement. If the measurement is not normal, then we use Wilcoxon rank sum test [51]. As reported in **Table 3**, EquiPPIS is statistically significantly better than GraphPPIS on both experimental and predicted structures at 95% confidence level with p-values < 0.05 for all four metrics.

Impact of secondary structure content on prediction accuracy

To examine the impact of physical characteristics of the input structures on the prediction accuracy of EquiPPIS, we analyze the secondary structure content of the proteins for the

Table 3. Significance test between EquiPPIS and GraphPPIS using experimental and AlphaFold2 predicted structures.

Input type	Method	F1	MCC	ROC-AUC	PR-AUC
Experimental	EquiPPIS	**0.4793** ± 0.000302678	**0.3664** ± 0.000287822	**0.8032** ± 9.50667E-05	**0.4684** ± 0.000536933
	GraphPPIS	0.4539 ± 0.0004801	0.3334 ± 0.000352489	0.7818 ± 0.000117733	0.4395 ± 0.000764056
	p-value	3.18683E-05	2.41981E-05	3.97085E-05	0.000373678
AlphaFold2	EquiPPIS	**0.4712** ± 0.000345067	**0.356** ± 0.000401778	**0.7912** ± 0.000189733	**0.4516** ± 0.0008396
	GraphPPIS	0.4434 ± 0.000515378	0.3199 ± 0.000447433	0.7667 ± 0.000159567	0.3956 ± 0.0007036
	p-value	7.71066E-06	0.002827272	0.001939728	9.92622E-08

***** For all four metrics, the two numbers reported are mean and variance respectively; values in bold represent the best performance in terms of mean.

https://doi.org/10.1371/journal.pcbi.1011435.t003

Test_60 set. The targets are divided into three groups: (1) helices (referred to as 'Primarily helix'), (2) beta strands (referred to as 'Primarily beta'), and (3) mixture of helices and beta strands (referred to as 'Mix'). We calculate the ROC-AUC scores for each group and compare them with the results obtained from the full set. S1 Fig presents a comparison of our prediction accuracy in terms of ROC-AUC scores across the groups with different physical characteristics based on secondary structure. EquiPPIS achieves a higher ROC-AUC of 0.855 in the 'Primarily helix' group compared to the full set's ROC-AUC of 0.805. However, in the 'Primarily beta' group, EquiPPIS attains a somewhat lower ROC-AUC of 0.783. As for the 'Mix' group, EquiPPIS achieves the ROC-AUC of 0.799, which is comparable to that obtained on the full set. The results demonstrate that secondary structure content has a minor impact on the prediction accuracy with primarily helical proteins yielding the best performance, whereas there is still room for improvement for the proteins that are primarily made of beta strands.

Running time analysis

We analyze the running time of EquiPPIS and the closest competing method GraphPPIS for all targets in the Test_60 set on the same Linux machine with identical hardware environment. Fig 4 presents a target-wise running time comparison between EquiPPIS and GraphPPIS. Not surprisingly, free from time-consuming MSA-searching, EquiPPIS demonstrates noticeably lower running time compared to GraphPPIS. Overall, EquiPPIS is considerably efficient in terms of the running time.

Ablation studies and choice of hyperparameters

To examine the relative importance of the features adopted in EquiPPIS, we conduct feature ablation experiments by gradually isolating the contribution of individual feature or groups of features during model training and evaluating the accuracy on the independent validation set Validation_42. Fig 5A shows the accuracy decline measured in terms of PR-AUC and ROC-AUC when various features are isolated from the full-fledged version of EquiPPIS. The results demonstrate that all features contribute to the overall accuracy achieved by EquiPPIS. For example, we notice accuracy decline when we isolate the sequence-based features one by one including amino acid residue type (No AA), position specific scoring matrix (No PSSM), and protein language model ESM2 (No ESM2). Not surprisingly, the evolutionarily feature

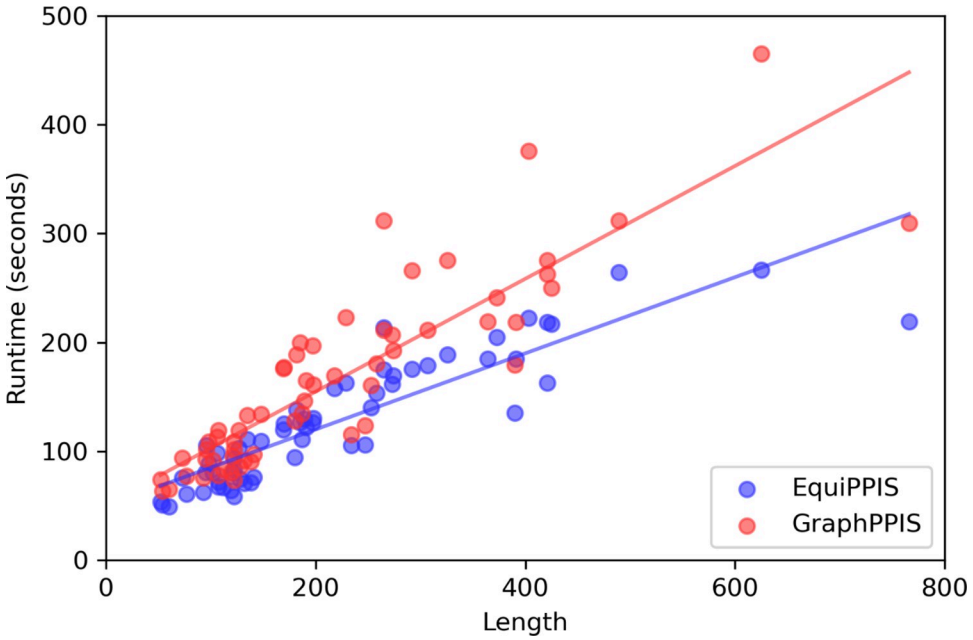

Fig 4. The running time of EquiPPIS and GraphPPIS on Test_60 set. For each target, input protein length versus runtime (in seconds) of EquiPPIS (blue) and GraphPPIS (red) are shown.

https://doi.org/10.1371/journal.pcbi.1011435.g004

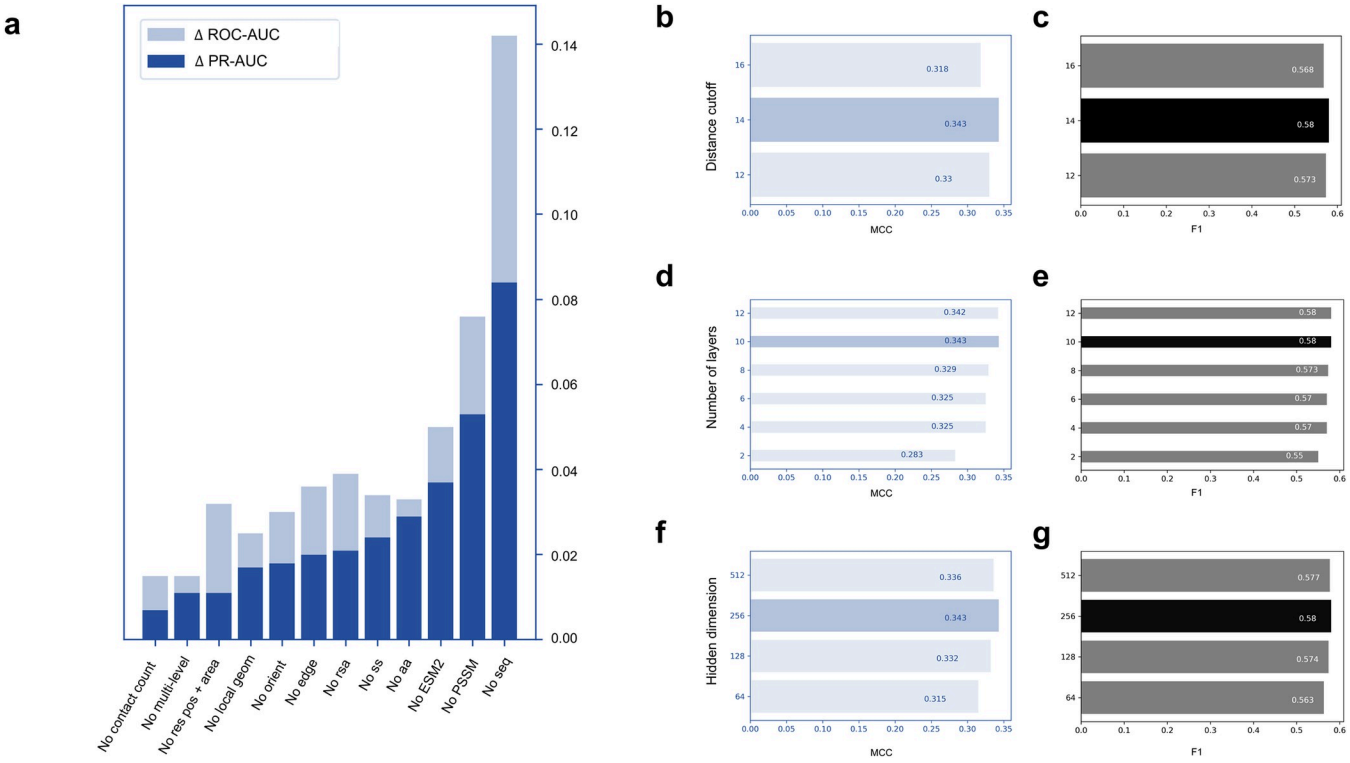

Fig 5. Ablation studies and hyperparameter selection. (**a**) Validation set accuracy decline after feature ablation. Dark and light blue bars indicate the PR-AUC decline (PR-AUC) and the ROC-AUC decline (ROC-AUC), respectively. Validation set accuracy in terms of MCC and F1 for various choices of hyperparameters including (**b**, **c**) distance cutoff, (**d**, **e**) number of layers, and (**f**, **g**) hidden dimension. The selected hyperparameters yielding the best accuracy are highlighted in darker shade.

https://doi.org/10.1371/journal.pcbi.1011435.g005

PSSM and protein language model-based ESM2 feature contribute more than just the residue type features. We notice a significant performance drop when all three sequence-based features are isolated (No seq). Similarly, we notice consistent accuracy decline when we discard the structure-based features individually including secondary structure (No ss), relative solvent accessibility (No rsa), local geometry (No local geom), residue orientation (No orient), contact count (No contact count) as well as relative residue positioning and residue virtual surface area (No res pos + area). Because we use multi-level discretization of secondary structure (e.g., 3-state and 8-state) and relative solvent accessibility (e.g., 2-state and 8-state) as well as backbone torsion angles, which are closely related to the secondary structure, we conduct feature ablation experiments by discarding the 8-state secondary structure, 8- state relative solvent accessibility, and backbone torsion angles. The resulting model (No multi-level) shows accuracy decline compared to the full-fledged version of EquiPPIS, indicating the effectiveness of combining multi-granular information. Finally, we also notice an accuracy drop when we isolate the edge feature (No edge) that takes into account the contributions of sequence separation and spatial interaction.

We also use the Validation_42 set to select the hyperparameters. Based on the results of the grid search as shown in **Fig 5B to 5G**, we select a 10-layer EGCL framework with 256 hidden units and the cutoff distance used to obtain the interacting residue pairs is set to 14Å. We use the hyperparameters selected in the independent validation set during training and testing.

Discussion

This work introduces EquiPPIS, a symmetry-aware deep graph learning model for protein–protein interaction site prediction based on E(3) equivariant graph neural networks. We demonstrate that EquiPPIS outperforms existing methods and despite being trained on experimental structures, it generalizes extremely well to predicted structural models from AlphaFold2 to the extent that EquiPPIS attains better accuracy with predicted structural models than what existing approaches can achieve even with experimental structures. Through controlled experiments, we verify the importance of equivariance as one of the major driving forces behind the improved performance. In addition to questions around the effect of equivariance on accuracy, our ablation study on an independent validation set confirms the contribution of various features adopted in EquiPPIS. Our study leads to a series of interesting questions to consider: of particular interest is the possibility of broadening the applicability of our method beyond experimental input for large-scale PPI site predictions with high accuracy by utilizing rapid computational prediction. In this regard, considering the diversity of the predictive modeling ensemble and accounting for the conformational states of the interacting proteins having multi-state conformational dynamics may help broaden the horizon of computational PPI site prediction. Further, a promising direction for future work is to investigate the potential benefits of explicitly including multiple sequence alignment (MSA) information and measure the extent to which it may influence the accuracy. While an MSA-free method such as EquiPPIS offers some unique advantages by being broadly applicable even for proteins that do not have homologous sequences in the current sequence databases and bypasses the computational overhead of MSA searching, MSA may still provide a rich source of additional information for further improving the accuracy of PPI site prediction that might be worth exploring. Finally, while we find that EquiPPIS exhibits excellent predictive accuracy and remarkable robustness, an open challenge that remains is the interpretability of our deep learning model. The evolutionary and functional significance of the residues predicted to be in PPI site by means of the latent representation underlying the neural architecture of EquiPPIS still need to be systematically explored. We expect our proposed method can be easily extended to other biomolecular

interaction site prediction tasks, including predicting protein-binding sites with other molecules, such as DNA, RNA and small ligands, as well as predicting gene-gene interaction and gene-networks with improved accuracy and robustness.

Materials and Methods

Graph representation and featurization.

Graph representation

We represent the input protein monomer as a graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, in which a node $v \in \mathcal{V}$ represents a residue and an edge $e \in \mathcal{E}$ represents an interacting residue pair. We consider two residues to be interacting if their C Euclidean distance is no more than 14Å. The cutoff 14Å is chosen on an independent validation set as presented in Fig 5. To focus on longer-rage interactions, we only consider interacting residue pairs having a minimum sequence separation of 6.

Node features

We use three types of sequence-based node features: (1) one-hot encoding of the residue (i.e., a binary vector of 20 entries indicating its amino acid type), resulting in L×20 feature set, where L is the length of the input monomer; (2) position specific scoring matrix (PSSM) obtained by running PSI-BLAST [52] to obtain L×20 feature set by considering the first 20 columns from the PSSM and normalizing the values by applying a sigmoidal function; and (3) features from ESM2 [53], which is a recent protein language model trained on 15 billion parameters, leading to L×33 feature set, after normalizing the values using sigmoidal function.

Additionally, we extract a total of L×45 structure-based node features from the structure of the input monomer by either calculating various structural information directly from the 3D coordinates or by running the DSSP [54] program. We describe them below.

Secondary structure and relative solvent accessibility. We use one-hot encoding of both 3-state and 8-state secondary structures (SS), leading to L×3 and L×8 feature sets, respectively. Additionally, we use one-hot encoding of 2-state relative solvent accessibility (RSA) by adopting an RSA cutoff of 50 (L×2 feature set), and use finer-grained RSA binning by discretizing into 8 bins as 0–30, 30–60, 60–90, 90–120, 120–150, 150–180, 180–210, and >210, represented by one-hot encoding (L×8 feature set). The rationale of using multiple discretization of SS (e.g., 3-state and 8-state) and RSA (e.g., 2-state and 8-state) is to combine multi-level information, and ablation studies guiding these decisions are presented in Fig 5.

Local geometry. We use a total of L×11 feature set from the local geometries by calculating various planar and torsion angles of the polypeptide chain, including (1) the cosine angle between the consecutive residues of the C = O bond; (2) sine and cosine of the virtual bond and torsion angles formed between the consecutive C atoms; (3) normalized values of the backbone torsion angles.

Relative residue positioning. To capture the relative positional information for each residue, we extract two types of features: (1) for i^{th} residue, we use the inverse of i to capture the relative sequence position (L×1 feature set); and (2) we use the inverse of the Euclidean distance from the centroid of the input protein monomer to the C atom of the i^{th} residue to capture the spatial positioning of a residue relative to the overall structure (L×1 feature set).

Residue orientation. To define the orientation of each amino acid residue, we adopted features from a recent work [55] including (1) the forward and reverse unit vectors in the directions of $C^{(i+1)} - C^{i}$ and $C^{(i-1)} - C^{i}$, respectively (L×6 feature set); and (2) the unit vector in the imputed direction of $C^{i} - C^{i}$, computed by assuming tetrahedral geometries and normalization (L×3 feature set).

Residue virtual surface area. An amino acid residue can be perceived as a virtual convex hull constructed by its atoms. We calculate the virtual surface area of the convex hull and use its inverse as a feature (L×1 feature set).

Contact count. If the Euclidean distance between the C atoms of a residue pair is within a cutoff of 8Å, the two residues can be considered to be in contact. We calculate the contact-count by calculating the number of spatial neighbors of a given residue (i.e., residues which are in contact) and use the normalized number of contact count per residue as a feature (L×1 feature set).

The sequence-based amino acid (L×20 feature set), PSSM (L×20 feature set), and ESM2 (L×33 feature set) features are concatenated with the structure-based node features (L×45 feature set), leading to a total of L×118 features, which serves as an input to our E(3) equivariant graph neural network model (S2 Fig).

Edge features

As the edge feature for the graph $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, we calculate the ratio of the logarithmic sequential separation of two residues (i.e., logarithm of the absolute difference between the two residue indices) corresponding to two nodes in the graph and the Euclidean distance between them, defined as:

$$\boldsymbol{a}_{ij} = \frac{\mathbf{log}(\boldsymbol{abs}(\boldsymbol{i} - \boldsymbol{j}))}{\|\boldsymbol{x}_i - \boldsymbol{x}_j\|}$$

Here, the numerator captures how the two residues are separated in the primary sequence while the denominator captures their spatial interactions.

Network architecture

We formulate the PPI site prediction into a graph node classification task and predict the probability of every residue in the input monomer to be a PPI site using a deep E(3) equivariant graph neural network. The network architecture consists of a stack of equivariant graph convolution layer (EGCL) [38], performing a series of transformations of its input by updating the coordinate and node embeddings using the edge information and the coordinate and node embeddings from the previous layer. The EGCL operation attains equivariance primarily by changing the standard message passing ($\boldsymbol{m}_{ij} = \ _e(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, \boldsymbol{a}_{ij})$) [56] to equivariant message passing and by introducing coordinate updates in the graph neural network, as follows:

$$\boldsymbol{m}_{ij} = \ _e(\boldsymbol{h}_i^l, \boldsymbol{h}_j^l, \|\boldsymbol{x}_i^l - \boldsymbol{x}_j^l\|^2, \boldsymbol{a}_{ij})$$

$$\boldsymbol{x}_i^{l+1} = \boldsymbol{x}_i^l + \boldsymbol{C}\sum\nolimits_{j \neq i}(\boldsymbol{x}_i^l - \boldsymbol{x}_j^l) \ _x(\boldsymbol{m}_{ij})$$

where $\boldsymbol{a}_{ij}$ denotes edge features; $\boldsymbol{h}_i^l$, and $\boldsymbol{h}_j^l$ are node embeddings at layer $\boldsymbol{l}$ for nodes $\boldsymbol{i}$ and $\boldsymbol{j}$, respectively; $\boldsymbol{\phi_e}$, $\boldsymbol{\phi_h}$, and $\boldsymbol{\phi_x}$ are multilayer perceptrons (MLP) for edge, node, and coordinate operations, respectively. Equivariant message passing for an edge $(\boldsymbol{i}, \boldsymbol{j})$ is attained by considering the squared distance between node $\boldsymbol{i}$ and $\boldsymbol{j}$: $(\|\boldsymbol{x}_i^l - \boldsymbol{x}_j^l\|^2)$ in the edge operation. The coordinate update for node $\boldsymbol{i}$ is obtained by the weighted sum of coordinate embedding difference from the previous layer, normalizing with a factor $\boldsymbol{C} = 1/(\boldsymbol{M}-1)$, where $\boldsymbol{M}$ is the number of nodes in the graph. The weights for the sum are generated through the multilayer perceptron (MLP) of coordinate operation applied on the equivariant message passing ($\boldsymbol{m}_{ij}$) for each edge $(\boldsymbol{i}, \boldsymbol{j})$.

Unlike off-the-shelf graph neural networks that aggregate messages only from the neighboring nodes, equivariant graph neural networks aggregate messages from the whole graph. Additionally, an attention embedding ($\boldsymbol{m_a}$) can be employed through an attention operation ($\boldsymbol{\phi_a}$), which is a linear transformation on the aggregated message embedding ($\boldsymbol{m_i}$), followed by a sigmoidal non-linear transformation. The 'attended' aggregated message embedding is subsequently obtained through a scalar multiplication with the attention embedding. The node embedding is updated by applying an MLP ($\boldsymbol{\phi_h}$) on the aggregated message and the node embeddings of the previous layer, as follows:

$$\boldsymbol{m_i} = \sum\nolimits_{j \neq i} \boldsymbol{m_{ij}}$$

$$\boldsymbol{m_a} = \ _{\boldsymbol{a}}(\boldsymbol{m_i})$$

$$\boldsymbol{m_i} = \boldsymbol{m_a} \times \boldsymbol{m_i}$$

$$\boldsymbol{h_i^{l+1}} = \ _{\boldsymbol{h}}(\boldsymbol{h_i^l}, \boldsymbol{m_i})$$

Finally, A linear transformation ($\boldsymbol{\phi_0}$) is applied to squeeze the hidden dimension ($\boldsymbol{h_i^L}$) of the last EGCL, followed by a sigmoidal function to obtain the node-level classification ($\boldsymbol{p_i}$) for PPI site prediction:

$$\boldsymbol{h_i} = \ _{0}(\boldsymbol{h_i^L})$$

$$\boldsymbol{p_i} = \frac{1}{1 + \boldsymbol{e}^{-h_i}}$$

The network architecture of EquiPPIS consists of 10 layers of EGCL with a hidden dimension of 256, where the hyperparameters are chosen on an independent validation set, and empirical results guiding these decisions are presented in **Fig 5**. EquiPPIS is implemented on Pytorch 1.12.0 [57] and Deep Graph Library (DGL) 0.9.0 [58]. We use binary cross entropy loss function between the node-level prediction and the ground truth, and we utilize cosine annealing scheduler from SGDR [59], ADAM optimizer [60] with a learning rate of 1e-4, and weight decay of 1e-16. The training process consists of at most 50 epochs on an NVIDIA A40 GPU. In addition to the full-fledged version of EquiPPIS, we train several baseline models on the same Train_335 dataset using the same set of input features and hyperparameters including off-the-shelf graph convolution network (GCN) [35] and graph attention network (GAT) [45], both implemented using the DGL [58], as well as two variants of EquiPPIS: (1) 'EquiPPIS invariant', an invariant network with the coordinate updates of the equivariant graph convolution layers turned off; and (2) 'EquiPPIS w/o attention', an equivariant network with the attention operation turned off during equivariant message passing.

Datasets, benchmarking, and performance evaluation

We use a combination of three widely used and publicly available benchmark datasets: Dset_186 [7], Dset_72 [7], and Dset_164 [40]. Dset_72 is created based on protein-protein benchmark version 3.0 [61], Dset_186 is constructed through a six-step filtering process which involves the exclusion of structures containing over 30% missing residues, identical UniprotKB/Swiss-Prot accessions, interface polarity and buried surface accessibility below specific thresholds, as well as oligomeric structures, transmembrane, and redundant protein structures, where 186 targets are collected from known protein complexes, and Dset_164 consists of 164

targets obtained from known heterodimers. While Dset_186, Dset_72, and Dset_164 are non-redundant data sets independently, the combined dataset is further reduced to 395 targets by filtering out redundant chains among the datasets. Following the same train-test split as GraphPPIS [10], we use a train set (Train_335) having 10,374 and 55,992 interacting and non-interacting residues, respectively; and a test set (Test_60) having 2,075 and 11,069 interacting and noninteracting residues, respectively. In the Train_335 set, the average length of protein is ~198 residues ranging from 44 to 869 residues. In the Test_60 set, the average length of protein is ~219 residues ranging from 52 to 766 residues, with no homodimeric protein-protein interaction present within this set. To assess the robustness of EquiPPIS and examine the effect of conformational changes on its performance, we analyze a subset of 31 proteins from the Test_60 set with known unbound monomeric conformations. This additional unbound test set (UBtest_31) having 841 and 5813 interacting and non-interacting residues, respectively, is adopted from the published work on GraphPPIS. Additionally, we adopt a dataset named Test_315 from the published work on GraphPPIS consisting of newly solved protein complexes that are non-redundant to the train set. We filter out 42 targets from the Test_315 set by discarding protein chains having more than 25% pairwise sequence identity with our test set and create an independent validation set (Validation_42) to perform feature ablation and hyperparameter selection.

During prediction, we use both experimentally-solved structures as well as on AlphaFold2--predicted structural models as input. We run AlphaFold2 with default parameter settings by locally installing the officially released version [30] to generate five predicted structural models and then select the model with the highest pLDDT confidence score. For target 4cdgA that failed during the MSA generation stage of the AlphaFold2 pipeline, we run Colabfold [62] that uses MMSeqs2 [63] for MSA generation and subsequently employs AlphaFold2 protocol for structure prediction.

EquiPPIS is compared against both sequence-based (PSIVER [7], ProNA2020 [41], SCRIBER [42], DLPred [43], and DELPHI [9]) and structure-aware (DeepPPIS [8], SPPIDER [11], MaSIF-site [44], and GraphPPIS [10]) PPI site prediction methods. PSIVER employs a Naïve Bayes classifier along with kernel density estimation by utilizing sequence-based features. ProNA2020 combines homology modeling with a neural network for residue-level PPI site prediction. SCRIBER employs two layers of logistic regression, where the first layer utilizes sequence-based features while the second layer combines the output from the first layer for the final prediction. DLPred employs a simplified long-short-term memory model for PPI site prediction. DELPHI uses an ensemble of convolutional and recurrent neural networks architectures with a large feature set. DeepPPISP combines local contextual features with global features and employs convolutional neural networks to predict PPI sites. SPPIDER leverages support vector machine, neural network, and linear discriminant analysis with an extensive feature search and extraction process. MaSIF-site predicts PPI sites by learning protein structural fingerprints through geometric deep learning. GraphPPIS employs structure-aware deep residual neural networks for PPI site prediction.

For benchmarking and performance assessment, we use standard performance evaluation metrics including accuracy, precision, recall, F1-score (F1), and Matthews correlation coefficient (MCC), defined as:

$$\textit{Accuracy} = \frac{TP + TN}{TP + TN + FP + FN}$$

$$\textit{Precision} = \frac{TP}{TP + FP}$$

$$Recall = \frac{TP}{TP + FN}$$

$$F1 = 2\frac{Precision \times Recall}{Precision + Recall}$$

$$MCC = \frac{TP \times TN - FN \times FP}{\sqrt{(TP + FP) \times (TP + FN) \times (TN + FP) \times (TN + FN)}}$$

where, TP denotes the number of true PPI site residues that are correctly predicted, FP denotes the number of non-PPI site residues that are incorrectly predicted to be in PPI sites, TN denotes the number of non-PPI site residues that are correctly predicted, and FN denotes the number of PPI site residues that are incorrectly predicted as non-PPI site. We additionally use area under the receiver operating characteristic curve (ROC-AUC) and area under the precision-recall curve (PR-AUC) for performance evaluation.

Supporting information

S1 Fig. Impact of secondary structure content on prediction accuracy. ROC-AUC scores achieved by EquiPPIS grouped by secondary structure content ('Primarily helix', 'Primarily beta', and 'Mix') as well as the overall ROC-AUC ('All') in the Test_60 set.
(TIF)

S2 Fig. Input node feature generation. The sequence-based amino acid (L×20 feature set), PSSM (L×20 feature set), and ESM2 (L×33 feature set) features are concatenated with the structure-based node features (L×45 feature set), leading to a total of L×118 features, which serves as an input to the E(3) equivariant graph neural networks.
(TIF)

Author Contributions

Conceptualization: Debswapna Bhattacharya.

Data curation: Rahmatullah Roche.

Formal analysis: Rahmatullah Roche, Debswapna Bhattacharya.

Funding acquisition: Debswapna Bhattacharya.

Investigation: Rahmatullah Roche, Debswapna Bhattacharya.

Methodology: Debswapna Bhattacharya.

Project administration: Debswapna Bhattacharya.

Resources: Debswapna Bhattacharya.

Software: Rahmatullah Roche, Bernard Moussad, Md Hossain Shuvo, Debswapna Bhattacharya.

Supervision: Debswapna Bhattacharya.

Validation: Rahmatullah Roche, Bernard Moussad, Md Hossain Shuvo, Debswapna Bhattacharya.

Visualization: Rahmatullah Roche, Debswapna Bhattacharya.

Writing – original draft: Rahmatullah Roche, Debswapna Bhattacharya.

Writing – review & editing: Debswapna Bhattacharya.

References

1. Jones S, Thornton JM. Principles of protein-protein interactions. Proceedings of the National Academy of Sciences. 1996; 93(1):13–20. https://doi.org/10.1073/pnas.93.1.13 PMID: 8552589
2. Sharan R, Suthram S, Kelley RM, Kuhn T, McCuine S, Uetz P, et al. Conserved patterns of protein interaction in multiple species. Proceedings of the National Academy of Sciences. 2005; 102(6):1974–9. https://doi.org/10.1073/pnas.0409522102 PMID: 15687504
3. Shoemaker BA, Panchenko AR. Deciphering protein–protein interactions. Part I. Experimental techniques and databases. PLoS computational biology. 2007; 3(3):e42.
4. Keskin O, Gursoy A, Ma B, Nussinov R. Principles of protein%protein interactions: what are the preferred ways for proteins to interact? Chemical reviews. 2008; 108(4):1225–44. https://doi.org/10.1021/cr040409x PMID: 18355092
5. Nooren IM, Thornton JM. Diversity of protein–protein interactions. The EMBO journal. 2003; 22 (14):3486–92. https://doi.org/10.1093/emboj/cdg359 PMID: 12853464
6. Chatrabgoun O, Daneshkhah A, Esmaeilbeigi M, Safa NS, Alenezi AH, Rahman A. Predicting Primary Sequence-Based Protein-Protein Interactions Using a Mercer Series Representation of Nonlinear Support Vector Machine. IEEE Access. 2022; 10:124345–54.
7. Murakami Y, Mizuguchi K. Applying the Na &e Bayes classifier with kernel density estimation to the prediction of protein–protein interaction sites. Bioinformatics. 2010; 26(15):1841–8.
8. Zeng M, Zhang F, Wu F-X, Li Y, Wang J, Li M. Protein–protein interaction site prediction through combining local and global features with deep neural networks. Bioinformatics. 2020; 36(4):1114–20. https://doi.org/10.1093/bioinformatics/btz699 PMID: 31593229
9. Li Y, Golding GB, Ilie L. DELPHI: accurate deep ensemble model for protein interaction sites prediction. Bioinformatics. 2021; 37(7):896–904. https://doi.org/10.1093/bioinformatics/btaa750 PMID: 32840562
10. Yuan Q, Chen J, Zhao H, Zhou Y, Yang Y. Structure-aware protein–protein interaction site prediction using deep graph convolutional network. Bioinformatics. 2022; 38(1):125–32.
11. Porollo A, Meller J. Prediction-based fingerprints of protein–protein interactions. Proteins: Structure, Function, and Bioinformatics. 2007; 66(3):630–45. https://doi.org/10.1002/prot.21248 PMID: 17152079
12. Li M-H, Lin L, Wang X-L, Liu T. Protein–protein interaction site prediction based on conditional random fields. Bioinformatics. 2007; 23(5):597–604. https://doi.org/10.1093/bioinformatics/btl660 PMID: 17234636
13. Fout A, Byrd J, Shariat B, Ben-Hur A. Protein interface prediction using graph convolutional networks. Advances in neural information processing systems. 2017;30.
14. Townshend R, Bedi R, Suriana P, Dror R. End-to-end learning on 3d protein structure for interface prediction. Advances in Neural Information Processing Systems. 2019;32.
15. Afsar Minhas FuA Geiss BJ, Ben-Hur A. PAIRpred: partner-specific prediction of interacting residues from sequence and structure. Proteins: Structure, Function, and Bioinformatics. 2014; 82(7):1142–55.
16. Sanchez-Garcia R, Sorzano COS, Carazo JM, Segura J. BIPSPI: a method for the prediction of partner-specific protein–protein interfaces. Bioinformatics. 2019; 35(3):470–7. https://doi.org/10.1093/bioinformatics/bty647 PMID: 30020406
17. Dai B, Bailey-Kellogg C. Protein interaction interface region prediction by geometric deep learning. Bioinformatics. 2021; 37(17):2580–8. https://doi.org/10.1093/bioinformatics/btab154 PMID: 33693581
18. Li N, Sun Z, Jiang F. Prediction of protein-protein binding site by using core interface residue and support vector machine. BMC bioinformatics. 2008; 9:1–13.
19. Northey TC, Bareši ' A, Martin AC. IntPred: a structure-based predictor of protein–protein interaction sites. Bioinformatics. 2018; 34(2):223–9. https://doi.org/10.1093/bioinformatics/btx585 PMID: 28968673
20. Hou Q, De Geest PF, Vranken WF, Heringa J, Feenstra KA. Seeing the trees through the forest: sequence-based homo-and heteromeric protein-protein interaction sites prediction using random forest. Bioinformatics. 2017; 33(10):1479–87. https://doi.org/10.1093/bioinformatics/btx005 PMID: 28073761
21. Sriwastava BK, Basu S, Maulik U. Protein–protein interaction site prediction in Homo sapiens and E. coli using an interaction-affinity based membership function in fuzzy SVM. Journal of biosciences. 2015; 40:809–18. https://doi.org/10.1007/s12038-015-9564-y PMID: 26564981

22. Lin X, Chen Xw. Heterogeneous data integration by tree-augmented na &e B ayes for protein–protein interactions prediction. Proteomics. 2013; 13(2):261–8.
23. Fariselli P, Pazos F, Valencia A, Casadio R. Prediction of protein–protein interaction sites in heterocomplexes with neural networks. European Journal of Biochemistry. 2002; 269(5):1356–61. https://doi.org/10.1046/j.1432-1033.2002.02767.x PMID: 11874449
24. Chen H, Zhou HX. Prediction of interface residues in protein–protein complexes by a consensus neural network method: test against NMR data. Proteins: Structure, Function, and Bioinformatics. 2005; 61 (1):21–35. https://doi.org/10.1002/prot.20514 PMID: 16080151
25. Liang S, Zhang C, Liu S, Zhou Y. Protein binding site prediction using an empirical scoring function. Nucleic acids research. 2006; 34(13):3698–707. https://doi.org/10.1093/nar/gkl454 PMID: 16893954
26. Deng A, Zhang H, Wang W, Zhang J, Fan D, Chen P, et al. Developing computational model to predict protein-protein interaction sites based on the XGBoost algorithm. International journal of molecular sciences. 2020; 21(7):2274. https://doi.org/10.3390/ijms21072274 PMID: 32218345
27. Wei Z-S, Han K, Yang J-Y, Shen H-B, Yu D-J. Protein–protein interaction sites prediction by ensembling SVM and sample-weighted random forests. Neurocomputing. 2016; 193:201–12.
28. Zhang J, Kurgan L. Review and comparative assessment of sequence-based predictors of protein-binding residues. Briefings in bioinformatics. 2018; 19(5):821–37. https://doi.org/10.1093/bib/bbx022 PMID: 28334258
29. Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The protein data bank. Nucleic acids research. 2000; 28(1):235–42. https://doi.org/10.1093/nar/28.1.235 PMID: 10592235
30. Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, et al. Highly accurate protein structure prediction with AlphaFold. Nature. 2021; 596(7873):583–9. https://doi.org/10.1038/s41586-021-03819-2 PMID: 34265844
31. Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, et al. Highly accurate protein structure prediction for the human proteome. Nature. 2021; 596(7873):590–6. https://doi.org/10.1038/s41586-021-03828-1 PMID: 34293799
32. Varadi M, Anyango S, Deshpande M, Nair S, Natassia C, Yordanova G, et al. AlphaFold Protein Structure Database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. Nucleic Acids Research. 2022; 50(D1):D439–D44. https://doi.org/10.1093/nar/gkab1061 PMID: 34791371
33. Bruna J, Zaremba W, Szlam A, LeCun Y. Spectral networks and locally connected networks on graphs. arXiv preprint arXiv:13126203. 2013.
34. Defferrard M, Bresson X, Vandergheynst P. Convolutional neural networks on graphs with fast localized spectral filtering. Advances in neural information processing systems. 2016;29.
35. Kipf TN, Welling M. Semi-supervised classification with graph convolutional networks. arXiv preprint arXiv:160902907. 2016.
36. Weiler M, Cesa G. General e (2)-equivariant steerable cnns. Advances in Neural Information Processing Systems. 2019;32.
37. Rezende DJ, Racani (re S, Higgins I, Toth P. Equivariant hamiltonian flows. arXiv preprint arXiv:190913739. 2019.
38. Satorras VcG, Hoogeboom E, Welling M. E(n) Equivariant Graph Neural Networks. In: Marina M, Tong Z, editors. Proceedings of the 38th International Conference on Machine Learning; Proceedings of Machine Learning Research: PMLR; 2021. p. 9323–32.
39. Thomas N, Smidt T, Kearnes S, Yang L, Li L, Kohlhoff K, et al. Tensor field networks: Rotation-and translation-equivariant neural networks for 3d point clouds. arXiv preprint arXiv:180208219. 2018.
40. Dhole K, Singh G, Pai PP, Mondal S. Sequence-based prediction of protein–protein interaction sites with L1-logreg classifier. Journal of Theoretical Biology. 2014; 348:47–54. https://doi.org/10.1016/j.jtbi.2014.01.028 PMID: 24486250
41. Qiu J, Bernhofer M, Heinzinger M, Kemper S, Norambuena T, Melo F, et al. ProNA2020 predicts protein–DNA, protein–RNA, and protein–protein binding proteins and residues from sequence. Journal of Molecular Biology. 2020; 432(7):2428–43. https://doi.org/10.1016/j.jmb.2020.02.026 PMID: 32142788
42. Zhang J, Kurgan L. SCRIBER: accurate and partner type-specific prediction of protein-binding residues from proteins sequences. Bioinformatics. 2019; 35(14):i343–i53. https://doi.org/10.1093/bioinformatics/btz324 PMID: 31510679
43. Zhang B, Li J, Quan L, Chen Y, Lü Q. Sequence-based prediction of protein-protein interaction sites by simplified long short-term memory network. Neurocomputing. 2019; 357:86–100. https://doi.org/10.1016/j.neucom.2019.05.013

44. Gainza P, Sverrisson F, Monti F, Rodol) E, Boscaini D, Bronstein MM, et al. Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. Nature Methods. 2020; 17 (2):184–92. https://doi.org/10.1038/s41592-019-0666-6 PMID: 31819266

45. Veli *kovi ' P, Cucurull G, Casanova A, Romero A, Lio P, Bengio Y. Graph attention networks. arXiv preprint arXiv:171010903. 2017.

46. Hammes GG, Chang Y-C, Oas TG. Conformational selection or induced fit: A flux description of reaction mechanism. Proceedings of the National Academy of Sciences. 2009; 106(33):13737–41. https://doi.org/10.1073/pnas.0907195106 PMID: 19666553

47. Peng Z, Kurgan L. High-throughput prediction of RNA, DNA and protein binding regions mediated by intrinsic disorder. Nucleic acids research. 2015; 43(18):e121–e. https://doi.org/10.1093/nar/gkv585 PMID: 26109352

48. Xia Y, Xia C-Q, Pan X, Shen H-B. GraphBind: protein structural context embedded rules learned by hierarchical graph neural networks for recognizing nucleic-acid-binding residues. Nucleic acids research. 2021; 49(9):e51–e.

49. Yuan Q, Chen S, Rao J, Zheng S, Zhao H, Yang Y. AlphaFold2-aware protein–DNA binding site prediction using graph transformer. Briefings in Bioinformatics. 2022; 23(2):bbab564. https://doi.org/10.1093/bib/bbab564 PMID: 35039821

50. Anderson TW, Darling DA. Asymptotic theory of certain" goodness of fit" criteria based on stochastic processes. The annals of mathematical statistics. 1952:193–212.

51. Wilcoxon F. Individual comparisons by ranking methods. In: Kotz S, Johnson NL, editors. Breakthroughs in Statistics: Methodology and Distribution. New York, NY: Springer New York; 1992. p. 196–202.

52. Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, et al. Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. Nucleic Acids Research. 1997; 25 (17):3389–402. https://doi.org/10.1093/nar/25.17.3389 PMID: 9254694

53. Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. Science. 2023; 379(6637):1123–30. https://doi.org/10.1126/science.ade2574 PMID: 36927031

54. Kabsch W, Sander C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. Biopolymers. 1983; 22(12):2577–637. https://doi.org/10.1002/bip.360221211 PMID: 6667333

55. Jing B, Eismann S, Suriana P, Townshend RJ, Dror R. Learning from protein structure with geometric vector perceptrons. arXiv preprint arXiv:200901411. 2020.

56. Gilmer J, Schoenholz SS, Riley PF, Vinyals O, Dahl GE, editors. Neural message passing for quantum chemistry. International conference on machine learning; 2017: PMLR.

57. Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, et al. Pytorch: An imperative style, high-performance deep learning library. Advances in neural information processing systems. 2019;32.

58. Wang M, Zheng D, Ye Z, Gan Q, Li M, Song X, et al. Deep graph library: A graph-centric, highly-performant package for graph neural networks. arXiv preprint arXiv:190901315. 2019.

59. Loshchilov I, Hutter F. Sgdr: Stochastic gradient descent with warm restarts. arXiv preprint arXiv:160803983. 2016.

60. Kingma DP, Ba J. Adam: A method for stochastic optimization. arXiv preprint arXiv:14126980. 2014.

61. Hwang H, Pierce B, Mintseris J, Janin J, Weng Z. Protein–protein docking benchmark version 3.0. Proteins: Structure, Function, and Bioinformatics. 2008; 73(3):705–9. https://doi.org/10.1002/prot.22106 PMID: 18491384

62. Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. ColabFold: making protein folding accessible to all. Nature Methods. 2022; 19(6):679–82. https://doi.org/10.1038/s41592-022-01488-1 PMID: 35637307

63. Steinegger M, Söding J. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. Nature Biotechnology. 2017; 35(11):1026–8. https://doi.org/10.1038/nbt.3988 PMID: 29035372