nature methods

Brief Communication

https://doi.org/10.1038/s41592-026-03086-x

eSIG-Net: an interaction language model that decodes the protein code of single mutations

Received: 19 November 2024

Accepted: 27 March 2026

Published online: 29 April 2026

Check for updates

Xingxin Pan[1,2,3,20] ✉, Aditya Shrawat[4,20], Sidharth Raghavan[1,2,20], Chuanpeng Dong[5,6], Yuntao Yang[7], Zhao Li[7], W. Jim Zheng[7], S. Gail Eckhardt[8], Erxi Wu[1,2,3,9,10,11], Juan I. Fuxman Bass[12,13], Daniel F. Jarosz[14], Sidi Chen[5,6], Daniel J. McGrail[15,16], Gloria M. Sheynkman[17], Jason H. Huang[1,2] ✉, Nidhi Sahni[18,19] ✉ & S. Stephen Yi[1,2,3,8,11] ✉

Most proteins act through interactions with other molecules, yet predicting how single mutations perturb these interactions—defined as 'protein codes'—remains a central challenge in computational biology. Here we introduce eSIG-Net, the edgetic mutation sequence-based interaction grammar network, a language model that integrates protein sequence embeddings with syntax-aware and evolution-aware mutation encoding and contrastive learning to predict mutation-driven interaction changes. eSIG-Net outperforms state-of-the-art sequence-based and structure-based methods, nominates causal variants and provides mechanistic insights. Together, eSIG-Net is a mutation-centric interaction language model that accurately predicts interaction-specific network rewiring from sequence information alone and generalizes across biological contexts.

Substantial improvements in genome and exome sequencing technology in the past 15 years have identified a surfeit of human genetic variation orders of magnitude more extensive than what was previously appreciated. However, how most variants influence the molecular properties and functions of molecules they encode, as well as their impacts on disease initiation and progression remain largely unknown[1]. Among these genetic variants, missense variants are the most common type of protein-coding mutations. Even single missense variants can drastically change protein–protein interactions or PPIs[2,3] (referred to as 'protein code' for interactions), and therefore rewire protein signaling[4]. Similar to the 'activity cliff'[5] problem in chemistry machine learning, where small structural changes often lead to large or unpredictable changes in activity, single mutations pose an 'interaction cliff' grand challenge, causing computational models to mispredict mutation-mediated PPIs (Supplementary Note 1).

Applying protein language models is a potential solution to these limitations and has been implemented in methods such as ESM1b[6], ESM-2[7], ProtT5[8], ESM3[9], D-SCRIPT[10] and AlphaMissense[11]. However, these methods also face at least two substantial challenges. First, they do not explicitly learn the sequence distinctions between mutant proteins and their corresponding wild-type (WT) counterparts. Second, they fail to capture the inherent complexity of PPIs, which are critical for PPI-related tasks.

Here we introduce a mutation-centric interaction language model named eSIG-Net (edgetic mutation sequence-based interaction grammar network). In contrast to conventional PPI prediction methods (Extended Data Fig. 1a), eSIG-Net focuses on the discrepancy between WT and mutant proteins, as well as their PPI profiles with a specific interaction partner. As shown in Fig. 1a, the framework of eSIG-Net consists of two encoder modules: (1) the first is a PPI 'protein encoder' module, which is commonly employed in classical PPI prediction tasks. It typically involves separately obtaining the encodings of a protein and its interactor and then merging them to predict PPIs. By constrast, in our PPI perturbation prediction pipeline, we obtained the merged encodings of both the WT with its interactor and the mutant with its interactor. These merged encodings were then fed into a constrained discrepancy module to attempt to discern the differences between them. (2) The second module is a mutant 'protein

A full list of affiliations appears at the end of the paper. ✉e-mail: xingxin.pan@bswhealth.org; jason.huang@bswhealth.org; nidhi.sahni.2025@gmail.com; song.yi@bcm.edu

https://doi.org/10.1038/s41592-026-03086-x

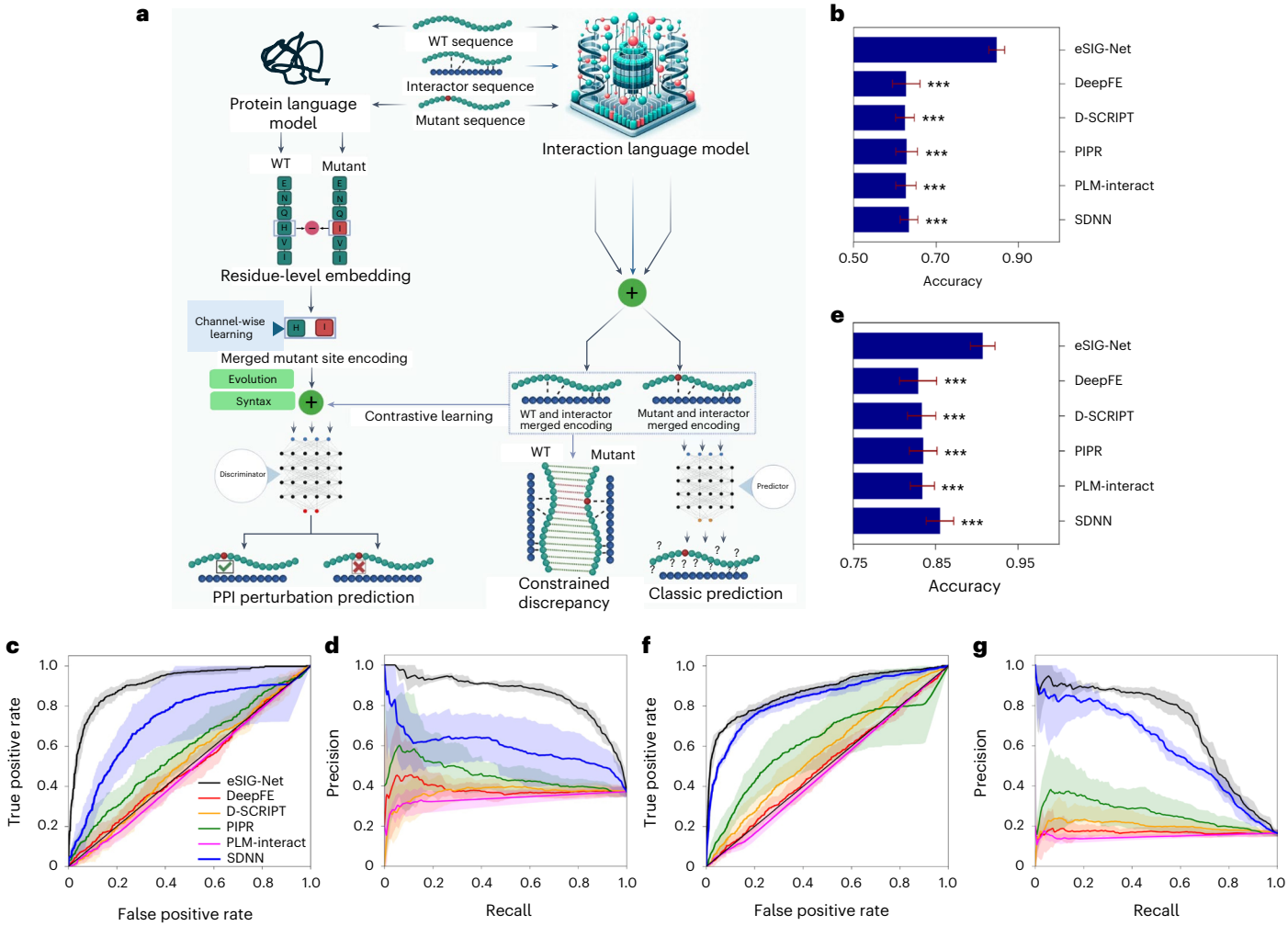


Fig. 1 | Overview of the eSIG-Net model, and benchmarking with state-of-the-art sequence-based prediction methods. a, The prediction framework of eSIG-Net: WT and mutant sequences are processed by a protein language model to obtain residue-level embeddings. These embeddings are then merged through mutation encoding and passed through a channel-wise mutation-site learning module. Concurrently, WT–interactor PPI and mutant–interactor PPI pairs are encoded by protein encoder, and their merged encodings are used for both constrained discrepancy assessment and traditional PPI prediction. Finally, the combined encodings—mutation-site, WT–interactor PPI and mutant–interactor PPI—are input into a discriminator to predict potential PPI perturbations caused by the mutation. **b**, Prediction accuracy comparison for the disease mutation PPI dataset, showcasing the performance of eSIG-Net against other PPI prediction models, with statistical significance denoted by asterisks. Error bars denote the standard deviations ($n = 1{,}633$ PPIs for each prediction model plotted). When compared to eSIG-Net, $P = 1.3 \times 10^{-6}$ for DeepFE, $P = 7.9 \times 10^{-7}$ for D-SCRIPT, PIPR, PLM-interact and SDNN. **c**, ROC curves for the disease mutation dataset, comparing the AUC metrics for eSIG-Net and other models, highlighting eSIG-Net's superior performance. **d**, Precision–recall curves for the disease mutation PPI dataset, with eSIG-Net outperforming other models in terms of both precision and recall. Line color scheme is the same as **c**. **e**, Prediction accuracy comparison for the gnomAD–ExAC population variant PPI dataset, with eSIG-Net achieving the highest accuracy. Error bars denote the standard deviations ($n = 4{,}020$ PPIs for eSIG-Net, DeepFE, PIPR, PLM-interact, SDNN; $n = 4{,}002$ PPIs for D-SCRIPT). When compared to eSIG-Net, the $P = 3.8 \times 10^{-4}$ for DeepFE, $P = 3.6 \times 10^{-4}$ for D-SCRIPT, $P = 3.6 \times 10^{-4}$ for PIPR, $P = 2.6 \times 10^{-4}$ for PLM-interact and $P = 8.7 \times 10^{-4}$ for SDNN. **f**, ROC curves for the population variant PPI dataset, with AUC values for each model, indicating that eSIG-Net maintains a high performance on this dataset as well. **g**, Precision–recall curves for the ExAC population variant PPI dataset, detailing the precision and recall performance of each model, with eSIG-Net providing a competitive precision–recall balance. Line color scheme is the same as **f**. *P* values are calculated by two-sided paired *t*-tests, with Holm–Bonferroni correction. ***$P < 0.001$. The error bars indicate ±s.d. and the centers of the error bars indicate mean accuracy ($n = 3$ independent experiments). For **c**,**d**,**f**,**g**, shading indicates s.d. The centers for the error bands indicate mean true or false positive rate (in **c**,**f**), and mean precision or recall (in **d**,**g**), respectively.

language model' encoder. Extensive empirical evidence has demonstrated that leveraging protein language models could capture the evolutionary information of proteins, thereby facilitating various downstream protein-related tasks. To accentuate the differences between mutant and WT proteins, we exclusively used the residue-level embeddings of the mutation sites. This was processed through channel-wise learning to obtain a merged mutation-site encoding (Fig. 1a). Finally, the two merged encodings were integrated and fed into a discriminator for the prediction of PPI perturbations. Compared with the conventional PPI prediction methods, eSIG-Net thus uses an innovative discrepancy strategy to effectively discern the effects of single amino acid changes on proteins and predict ensuing perturbations in their interaction profiles.

We first benchmarked the performance of eSIG-Net against other methods using two independent datasets: the disease mutation PPI dataset[2] and the population variant PPI dataset[12] (see Methods for details). For each of the datasets, we applied a fivefold cross-validation strategy to avoid the influence of random samples on the performance.

Most existing methods are not specifically tailored to predict PPI perturbations caused by missense mutations, therefore we compared

our eSIG-Net model against five state-of-the-art sequence-based PPI prediction methods: SDNN[13], D-SCRIPT[10], DeepFE[14], PIPR[15] and PLM-interact[16] (see Methods for details). Using disease mutation PPI dataset[2], eSIG-Net significantly outperformed all the benchmarking methods, and achieved an accuracy improvement of more than 20% compared with the other methods on the disease mutation PPI dataset (eSIG-Net accuracy of 0.85 ± 0.02; best accuracy by other methods of 0.63 ± 0.02) (Fig. 1b, Extended Data Fig. 1b–i and Supplementary Note 2). In addition, eSIG-Net outperformed the other existing methods in receiver operating characteristic (ROC) curve analysis. eSIG-Net achieved an area under curve (AUC) of 0.91 ± 0.02, while SDNN had an AUC of 0.73 ± 0.15, D-SCRIPT had an AUC of 0.51 ± 0.05, DeepFE had an AUC of 0.50 ± 0.04, PIPR had an AUC of 0.58 ± 0.07 and PLM-interact had an AUC of 0.48 ± 0.02 (Fig. 1c). Similarly, eSIG-Net also exhibited a better performance in precision–recall curve analysis. eSIG-Net achieved an average precision of 0.86 ± 0.01, whereas SDNN had an average precision of 0.61 ± 0.10, D-SCRIPT had an average precision of 0.37 ± 0.05, DeepFE had an average precision of 0.39 ± 0.04, PIPR had an average precision of 0.46 ± 0.09 and PLM-interact had an average precision of 0.37 ± 0.02 (Fig. 1d). Similar prediction performance was achieved using population variant PPI dataset (Fig. 1e–g and Supplementary Note 3). These results demonstrate eSIG-Net's superiority in learning the distinctions between WT and mutant proteins, compared with traditional PPI prediction methods that heavily rely on sequence learning.

To validate the effectiveness of two main modules in the eSIG-Net framework, we designed and executed an ablation study. As a baseline control, we performed traditional ESM (Evolutionary Scale Modeling)[7] pooling (esm2_t33_650M_UR50D) embeddings ('standard models', Fig. 2a), which yielded a limited improvement in accuracy (0.69 ± 0.03; Fig. 2a) compared with other benchmarking methods (Fig. 1b) on the disease mutation PPI dataset. Moreover, the accuracy on the imbalanced population variant PPI dataset (0.72 ± 0.02; Fig. 2b) was even lower than the worst-performing benchmarking method (Fig. 1e). However, incorporating our mutation-site encoding module led to accuracy improvement on both datasets, reaching 0.75 ± 0.03 and 0.78 ± 0.02, respectively (Fig. 2a,b). Finally, with the introduction of our constrained discrepancy learning module, the model's performance observed further enhancement (accuracy on disease mutation dataset of 0.85 ± 0.02, accuracy on population variant dataset of 0.90 ± 0.01) (Fig. 2a,b and Supplementary Note 4).

It is important to note that, current state-of-the-art structure-based method AlphaFold[17]-derived FoldDock[18] is limited, and fails to predict interaction alterations by select disease mutations (Fig. 2c–e,i, Extended Data Table 1 and Supplementary Note 5). Other structure-based prediction tools require the input of protein complex structures, we first subjected the protein sequences to the AlphaFold-Multimer model[19] to predict their structures, which then served as input for five structure-based prediction methods (MutaBind2[20], BeAtMuSiC[21], GeoPPI[22], TopNetTree[23] and PIONEER[24]) (Supplementary Note 6). For comparative purposes, mutations classified by these methods as deleterious were considered to be interaction-perturbing (see Methods for details). All five benchmarking prediction tools only had an around or below 60% accuracy rate (Fig. 2f and Extended Data Fig. 2a), much lower than eSIG-Net. In ROC curve analysis, eSIG-Net achieved an AUC of 0.91 ± 0.02, while MutaBind2 had an AUC of 0.60 ± 0.06, BeAtMuSiC had an AUC of 0.63 ± 0.02, GeoPPI had an AUC of 0.52 ± 0.03, TopNetTree had an AUC of 0.49 ± 0.06 and PIONEER had an AUC of 0.49 ± 0.01 (Fig. 2g). Similarly, eSIG-Net also exhibited a better performance in precision–recall curve analysis. eSIG-Net achieved an average precision of 0.86 ± 0.01, whereas MutaBind2 had an average precision of 0.44 ± 0.03, BeAtMuSiC had an average precision of 0.48 ± 0.05, GeoPPI had an average precision of 0.41 ± 0.06, TopNetTree had an average precision of 0.39 ± 0.06 and PIONEER had an average precision of 0.37 ± 0.03 (Fig. 2h). Together, we found that eSIG-Net significantly outperformed all the mutation-centric structure-based PPI prediction benchmarking tools.

Although millions of coding variants have been identified in the human genome, most of them remain classified as variants of unknown significance (VUS). To address this challenge, eSIG-Net offers a generalizable framework that can be applied across diverse biological contexts and adapted to predict interaction-specific variant effects directly from a protein sequence (Fig. 2j–k, Extended Data Fig. 2b,c and Supplementary Note 7). A good example is pleiotropism, where different mutations in the same gene cause different diseases. In the pleiotropic gene *TPM3*, two mutations, L100M and M9R, cause fiber-type disproportion myopathy[25] and nemaline myopathy[26], respectively (Fig. 2j). eSIG-Net predicted the mutation L100M to selectively perturb (that is, edgetic) the interaction with HSF2, which was known to be expressed in muscle and involved in myotube regeneration[27]. In contrast, eSIG-Net predicted the mutation M9R to retain the interaction with HSF2 (Fig. 2j). Together, eSIG-Net provided possible mechanistic insights into pleiotropic phenotypic outcomes through accurate prediction of distinct interaction profiles.

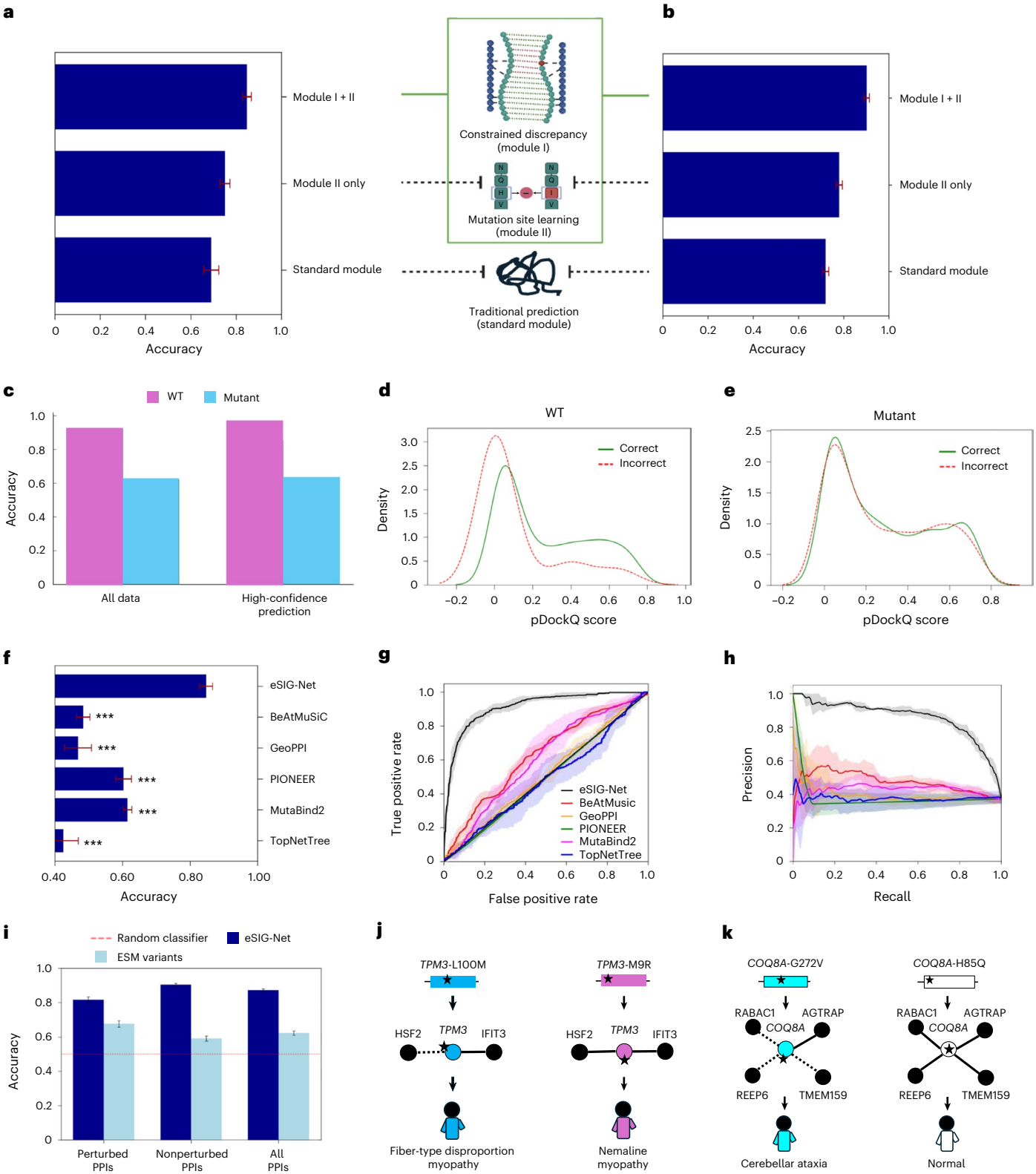
At present, large-scale studies of mutational impact on protein activities are extremely challenging, which are primarily measured by high-throughput wet-laboratory experimental platforms, such as functional variomics[3] and deep mutational scanning[28]. Although these methods have made enormous strides in characterizing large numbers of protein variants, they remain time-consuming and labor-intensive. The eSIG-Net method is designed to exactly tackle this problem, and can serve as an accurate and alternative functional characterization of variants at large scale through deep in silico mutagenesis. This effort will greatly facilitate the annotation and analysis of many protein

Fig. 2 | Benchmarking eSIG-Net with mutation-centric, structure-based prediction tools and application across diverse biological contexts. **a**,**b**, Our ablation study showcasing mean accuracy on the disease mutation PPI dataset (**a**) and the population variant PPI dataset (**b**). Error bars denote the standard deviations (n = 1,633 PPIs for each module configuration plotted). **c**, Bar chart summarizing the mean accuracy of FoldDock for predicting WT versus mutant protein interactions using all data (left) or high-confidence predictions (pDockQ > 0.5; right). **d**, Density distribution plot of pDockQ scores with correct (green) and incorrect (red) cases, for WT protein-mediated interactions. **e**, Density distribution plot of pDockQ scores with correct (green) and incorrect (red) cases, for mutant protein-mediated interactions. **f**–**h**, Prediction evaluation using mutation-centric, structure-based methods. The disease mutation PPI dataset is used here. **f**, Bar chart summarizing the mean accuracy of different structural algorithms, with the length of each bar representing the mean accuracy and the error bars denoting the standard deviations (n = 1,633 PPIs for eSIG-Net, BeAtMuSiC, GeoPPI, PIONEER; n = 1,612 PPIs for MutaBind2; n = 1,157 PPIs for TopNetTree). When compared to eSIG-Net, the $P = 1.3 \times 10^{-8}$ for BeAtMuSiC, $P = 2.3 \times 10^{-7}$ for GeoPPI, $P = 2.7 \times 10^{-7}$ for PIONEER, $P = 9.5 \times 10^{-8}$ for MutaBind2 and $P = 2.3 \times 10^{-7}$ for TopNetTree. **g**, ROC curves displaying the comparative AUC values for various structural algorithms. Shading indicates standard deviations. **h**, Precision–recall curves for structural algorithms. Shading indicates standard deviations. The centers for the error bands indicate mean true or false positive rate (in **g**), and mean precision or recall (in **h**), respectively. **i**, Bar chart showing the mean accuracy of ESM variants (cyan; a mutation-centric disease-causing prediction tool), compared with eSIG-Net (blue). Bar length represents the mean accuracy and error bars denote the standard deviations. Dashed line indicates a random classifier (n = 1,027 'Perturbed PPIs'; n = 606 'Nonperturbed PPIs'). **j**, eSIG-Net-predicted interaction profiles of two disease mutations in the pleiotropic gene *TPM3*. **k**, eSIG-Net-predicted interaction profiles of a disease mutation and a population variant in the gene *COQ8A* (also known as *ADCK3*). P values are calculated by two-sided paired t-tests, with Holm–Bonferroni correction. ***$P < 0.001$. The error bars indicate ±s.d. and the centers of the error bars indicate mean accuracy (n = 3 independent experiments).

https://doi.org/10.1038/s41592-026-03086-x

variants of unknown significance thus far, potentially contributing to discovery of new disease-relevant biomarkers and therapeutic strategies (Supplementary Note 8).

Similar to other state-of-the-art methods, eSIG-Net also has potential limitations. First, methods that use multiple sequence alignment (MSA) to extract information typically yield embeddings that capture valuable mutation and evolutionary information[11]. In the eSIG-Net framework, we use sequence-based biostatistical embeddings and protein language model embeddings for both the input to the PPI prediction module and the mutation-site encoding module to expedite embedding extraction. Nevertheless, this approach sacrifices some coevolutionary information under specific biological contexts. While the current version of eSIG-Net primarily predicts mutational effects on the energetic or biophysical favorability of interactions between a pair of proteins, there is room for future development considering that many disease-causing mutations lead to disease in a tissue-specific manner. Nor does a change



of PPI directly reveal causation of disease, let alone the druggable target identification. Nevertheless, we believe that eSIG-Net has the potential to revolutionize our comprehension of the mechanistic effects caused by mutations in molecular networks and to catalyze substantial advancements in therapeutic interventions for genetic disorders.

Online content

Any methods, additional references, Nature Portfolio reporting summaries, source data, extended data, supplementary information, acknowledgements, peer review information; details of author contributions and competing interests; and statements of data and code availability are available at https://doi.org/10.1038/s41592-026-03086-x.

References

1. Ng, P. K. et al. Systematic functional annotation of somatic mutations in cancer. *Cancer Cell* **33**, 450–462 (2018).
2. Sahni, N. et al. Widespread macromolecular interaction perturbations in human genetic disorders. *Cell* **161**, 647–660 (2015).
3. Yi, S. et al. Functional variomics and network perturbation: connecting genotype to phenotype in cancer. *Nat. Rev. Genet.* **18**, 395–410 (2017).
4. Li, Y. et al. e-MutPath: computational modeling reveals the functional landscape of genetic mutations rewiring interactome networks. *Nucleic Acids Res.* **49**, e2 (2021).
5. van Tilborg, D., Alenicheva, A. & Grisoni, F. Exposing the limitations of molecular machine learning with activity cliffs. *J. Chem. Inf. Model.* **62**, 5938–5951 (2022).
6. Brandes, N., Goldman, G., Wang, C. H., Ye, C. J. & Ntranos, V. Genome-wide prediction of disease variant effects with a deep protein language model. *Nat. Genet.* **55**, 1512–1522 (2023).
7. Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* **379**, 1123–1130 (2023).
8. Pokharel, S., Pratyush, P., Heinzinger, M., Newman, R. H. & Kc, D. B. Improving protein succinylation sites prediction using embeddings from protein language model. *Sci. Rep.* **12**, 16933 (2022).
9. Hayes, T. et al. Simulating 500 million years of evolution with a language model. *Science* **387**, 850–858 (2025).
10. Sledzieski, S., Singh, R., Cowen, L. & Berger, B. D-SCRIPT translates genome to phenome with sequence-based, structure-aware, genome-scale predictions of protein–protein interactions. *Cell Syst.* **12**, 969–982 (2021).
11. Cheng, J. et al. Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
12. Fragoza, R. et al. Extensive disruption of protein interactions by genetic variants across the allele frequency spectrum in human populations. *Nat. Commun.* **10**, 4141 (2019).
13. Li, X. et al. SDNN-PPI: self-attention with deep neural network effect on protein–protein interaction prediction. *BMC Genomics* **23**, 474 (2022).
14. Yao, Y., Du, X., Diao, Y. & Zhu, H. An integration of deep learning with feature embedding for protein–protein interaction prediction. *PeerJ* **7**, e7126 (2019).
15. Chen, M. et al. Multifaceted protein–protein interaction prediction based on Siamese residual RCNN. *Bioinformatics* **35**, i305–i314 (2019).
16. Liu, D. et al. PLM-interact: extending protein language models to predict protein–protein interactions. *Nat. Commun.* **16**, 9012 (2025).
17. Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
18. Bryant, P., Pozzati, G. & Elofsson, A. Improved prediction of protein–protein interactions using AlphaFold2. *Nat. Commun.* **13**, 1265 (2022).
19. Evans, R. et al. Protein complex prediction with AlphaFold-Multimer. Preprint at *bioRxiv* https://doi.org/10.1101/2021.10.04.463034 (2022).
20. Zhang, N. et al. MutaBind2: predicting the impacts of single and multiple mutations on protein–protein interactions. *iScience* **23**, 100939 (2020).
21. Dehouck, Y., Kwasigroch, J. M., Rooman, M. & Gilis, D. BeAtMuSiC: prediction of changes in protein–protein binding affinity on mutations. *Nucleic Acids Res.* **41**, W333–W339 (2013).
22. Liu, X., Luo, Y., Li, P., Song, S. & Peng, J. Deep geometric representations for modeling effects of mutations on protein–protein binding affinity. *PLoS Comput. Biol.* **17**, e1009284 (2021).
23. Wang, M., Cang, Z. & Wei, G. W. A topology-based network tree for the prediction of protein–protein binding affinity changes following mutation. *Nat. Mach. Intell.* **2**, 116–123 (2020).
24. Xiong, D. et al. A structurally informed human protein–protein interactome reveals proteome-wide perturbations caused by disease mutations. *Nat. Biotechnol.* **43**, 1510–1524 (2025).
25. Clarke, N. F. et al. Mutations in *TPM3* are a common cause of congenital fiber type disproportion. *Ann. Neurol.* **63**, 329–337 (2008).
26. Laing, N. G. et al. A mutation in the alpha tropomyosin gene *TPM3* associated with autosomal dominant nemaline myopathy. *Nat. Genet.* **9**, 75–79 (1995).
27. McArdle, A. et al. HSF expression in skeletal muscle during myogenesis: implications for failed regeneration in old mice. *Exp. Gerontol.* **41**, 497–500 (2006).
28. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein science. *Nat. Methods* **11**, 801–807 (2014).

[1]Department of Neurosurgery, Baylor College of Medicine, Temple, TX, USA. [2]Department of Neurosurgery and Neuroscience Institute, Baylor Research Institute, Temple, TX, USA. [3]Dell Medical School, The University of Texas at Austin, Austin, TX, USA. [4]Department of Physiology and Biophysics, Case Western Reserve University, Cleveland, OH, USA. [5]Department of Genetics, and Yale Comprehensive Cancer Center, Yale University School of Medicine, New Haven, CT, USA. [6]Systems Biology Institute, Integrated Science and Technology Center, West Haven, CT, USA. [7]McWilliams School of Biomedical Informatics, University of Texas Health Science Center at Houston, Houston, TX, USA. [8]Dan L Duncan Comprehensive Cancer Center, Baylor College of Medicine, Houston, TX, USA. [9]Naresh K. Vashisht College of Medicine and Irma Lerma Rangel College of Pharmacy, Texas A&M University, College Station, TX, USA. [10]Baylor Research Institute, Baylor Scott & White Health, Temple, TX, USA. [11]Dan L Duncan Institute for Clinical and Translational Research, Baylor College of Medicine, Houston, TX, USA. [12]Department of Biology, Boston University, Boston, MA, USA. [13]Center for Cancer Systems Biology (CCSB),

and Department of Cancer Biology, Dana-Farber Cancer Institute, Boston, MA, USA. [14]Department of Chemical and Systems Biology, and Department of Developmental Biology, Stanford University School of Medicine, Stanford, CA, USA. [15]Center for Immunotherapy and Precision Immuno-Oncology, Cleveland Clinic, Cleveland, OH, USA. [16]Lerner Research Institute, Cleveland Clinic, Cleveland, OH, USA. [17]Department of Molecular Physiology and Biological Physics, Department of Biochemistry and Molecular Genetics, and UVA Comprehensive Cancer Center, University of Virginia, Charlottesville, VA, USA. [18]Department of Neurosurgery, Baylor Research Institute and Baylor College of Medicine, Temple, TX, USA. [19]Program in Quantitative and Computational Biosciences (QCB), Baylor College of Medicine, Houston, TX, USA. [20]These authors contributed equally: Xingxin Pan, Aditya Shrawat, Sidharth Raghavan. ✉e-mail: xingxin.pan@bswhealth.org; jason.huang@bswhealth.org; nidhi.sahni.2025@gmail.com; song.yi@bcm.edu

Methods

eSIG-Net model

We first assembled multiple effective feature generation methods for PPI prediction to enhance the representation of proteins. Then, we optimized the PPI prediction framework by introducing a constrained discrepancy learning module. This module was specifically designed to differentiate the merged encoding of 'mutant–interactor PPI' and 'WT–interactor PPI' pairs, as their differences were subtle yet crucial for capturing the effects of missense mutations on PPI states. In addition, we harnessed the capabilities of protein language models by incorporating mutation-site encoding into the embeddings obtained from these models. We also employed a discriminator to predict the impact of missense mutations on PPI states.

It is essential to note that eSIG-Net primarily focused on learning the differences between WT and mutant protein embeddings, in conjunction with interactors, to predict the occurrence of PPI perturbations. Rather than directly predicting PPI states, our approach took a distinct strategy. In comparison to traditional PPI prediction methods, eSIG-Net incorporated several innovative designs to effectively distinguish between similar samples (that is, protein sequences that differ by a single mutation) and predict perturbations in their PPI states.

Datasets

Two mutation-mediated PPI datasets were used in this study: the disease mutation PPI dataset from ref. 2, and the population variant PPI dataset from ref. 12. The ref. 2 dataset consisted of 1,633 samples, with each sample composed of three proteins ('triplet': the WT protein, the mutant protein, the interactor) and the binding states of the WT–interactor PPI (WT–interactor) and mutant–interactor PPI (MT–interactor) pairs (0 or 1). This dataset contained 527 disease mutations in 220 genes, associated with 606 perturbed PPIs and 1,027 nonperturbed PPIs. On the other hand, the ref. 12 dataset carried one of the largest compilations of variants observed in the general population (from gnomAD database), serving as a baseline for neutral or nonpathogenic variation. The ref. 12 dataset contained 1,650 population variants in 772 genes, with a total of 663 perturbed PPIs and 3,357 nonperturbed PPIs. We excluded synonymous mutations from this analysis. We further defined a positive sample as one where a PPI state change (that is, perturbation) occurred, while a negative sample represented cases where no PPI perturbation took place. Note that the population variants dataset was used to validate the robustness of our model and other methods. This dataset contained only approximately 16% positive samples, making it an imbalanced dataset.

To obtain the gene sequences corresponding to the genotypes, we retrieved them from the hORFeome[29] V9.1 Library (http://horfdb.dfci.harvard.edu/) and converted the nucleotide sequence into amino acid sequences.

Model input features

In the experiments, the protein sequences were transformed into fixed-length feature vectors before being input into the neural network due to variations in sequence length. The feature fusion strategy was employed to convert protein sequences into feature vectors using three methods: amino acid composition, conjoint triad and auto covariance.

(1) The amino acid composition method[30] normalized the frequency of occurrence of each amino acid in the protein sequence. It counted the frequency of the 20 amino acids, resulting in a 20-dimensional feature vector for each protein sequence.

(2) The conjoint triad method[31] divided the 20 amino acids into 7 different clusters based on the volume of amino acid side chains and dipoles. Each cluster grouped together amino acids with similar characteristics. This resulted in a 343-dimensional feature vector that represented the normalized triples ($7 \times 7 \times 7$) of amino acids.

(3) The auto covariance method was used to capture interactions between amino acids that were separated by a specific number of residues within a protein sequence. The process started by converting the amino acid residues into numerical values that represented their physicochemical properties, such as hydrophobicity, polarity or molecular weight. After encoding the sequence numerically, the auto covariance calculation was carried out, which involved computing the covariance between the properties of amino acids separated by a fixed distance. The resulting 210-dimensional auto covariance feature vector was normalized to zero mean and unit standard deviation (s.d.). The experimental setup followed refs. 13,32 for obtaining the auto covariance feature vector.

In summary, these three feature vectors were concatenated to form a 573-dimensional feature vector, which was used for discrepancy learning and prediction. In addition, as shown in Extended Data Fig. 1, our model used ESM-2[7] to obtain additional embeddings for the mutation-site encoding module. Specifically, the 'esm2_t33_650M_UR50D' version of the model[7] was used to obtain residue-level embeddings. We extracted the embedding at the mutation site to obtain a 1,280-dimensional ESM feature vector.

Constrained discrepancy learning

In traditional machine learning, a model is trained by minimizing a loss function that measures the discrepancy between the model's predictions and the true labels. A constrained discrepancy loss augments this standard loss function with additional conditions that represent known knowledge (that is, single mutations can perturb PPIs). Constrained discrepancy loss is especially useful here when edgetic training data is limited. The goal of our training is to find the set of model parameters (weights and biases) that results in the lowest possible loss. For example, the loss is expected to be lower when there is a PPI perturbation. This is because learning PPI perturbation makes the model more accurate. Therefore, by leveraging previous knowledge, constrained discrepancy learning helps to improve the model's performance and interpretability.

To achieve the objective of learning a precise amount of discrepancy across mutated proteins with different binding state changes, we defined the distance between the merged encodings of missense mutations before and after as d_i and indicated whether the binding state changed as c_i ($c_i = 1$ if there is a perturbation, otherwise $c_i = 0$). We formalized our constrained discrepancy loss in equation (1):

$$\mathcal{L}_{cd} = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j=1, j\neq i}^{n} \left(\frac{d_i}{1+\lambda \times c_i} - \frac{d_j}{1+\lambda \times c_j} \right)^2 \quad (1)$$

where n is the number of the samples in one training batch, and d_i is the embedding-level differences before (W_i, WT) and after (M_i, mutant) the mutation, which were measured by L2 norm: $d_i = |W_i - Mi|_2$. λ is a hyperparameter that balances the penalty on perturbed versus nonperturbed samples. In the current study, we fixed $\lambda = 1$ across experiments, given the limited availability of large-scale experimental edgetic datasets and the observation that performance was stable under this setting. It is important to note that using equation (1) alone is insufficient, as the trivial solution would be to set all the embedding-level distances to zero (that is, $d_i = 0$ for all i). However, learning the accurate amount of discrepancy while avoiding the trivial solution is achievable by jointly training with the objective of the original PPI prediction change. In this joint training approach, the model was devised to embed mutant–interactor PPI and WT–interactor PPI differently to discriminate between them.

Mutation-site encoding

As illustrated in Extended Data Fig. 1, we harnessed the capabilities of the protein language model (specifically ESM-2[7]) to acquire

residue-level embeddings for both the WT and mutant (MT) proteins. These embeddings captured contextual information from the protein sequence. Subsequently, to reduce redundancy, we isolated the embedding specific to the mutation site and directed it through the mutation-site channel-wise self-attention module.

The pipeline of our mutation-site encoding module was as follows:

$$f(\mathrm{MT}, \mathrm{WT}) = \mathrm{MLP}(|\mathrm{esm}(\mathrm{MT})_s - \mathrm{esm}(\mathrm{WT})_s|)$$

$\mathrm{esm}(\cdot)_s \in R^d$ represents the vector corresponding to the mutation site extracted from the residue-level embedding ($\mathrm{esm}(\cdot) \in R^{L\times d}$), where ($L$) is the length of the protein and (d) is the feature dimension of ESM. A multi-layer perceptron (MLP) was implemented to obtain the final mutation-site encoding.

Subsequently, we concatenated the mutant–interactor PPI merged, WT–interactor PPI merged and mutation-site encoding, and fed them into the discriminator to predict whether the missense mutation led to a perturbation in the interaction state. Our overall learning objective is given as follows:

$$\mathcal{L} = \mathcal{L}_{\mathrm{discrim}} + \alpha_1 \mathcal{L}_{\mathrm{pred}} + \alpha_2 \mathcal{L}_{cd}$$

where both the discriminator loss L_{discrim} and predictor loss L_{pred} use the cross-entropy loss function. The discriminator's labels are based on whether the interaction is perturbed, with 1 indicating a perturbation and 0 indicating no change. The predictor's labels represent the interaction itself, where 1 denotes a binding (interaction) and 0 denotes no binding.

The hyperparameters α_1 and α_2 (and other loss weights) were selected exclusively based on the validation set within each fold, never on the test set. Our pipeline was a strict fivefold cross-validation: in each run, one fold was held out as the test set, while the remaining four folds were split into training and validation. Hyperparameters were tuned only on the validation set, and the final performance was reported on the held-out test set.

Training strategy

The mutant or WT-encoding module and the interaction-encoding module shared the same network structure, which consisted of seven linear layers forming an MLP. The merge layer was a linear layer with a size of 32. The mutation-site encoding module comprised two linear layers. Both the predictor and discriminator consisted of two MLP layers serving as classifiers. Batch normalization and a dropout rate of 0.3 were applied after each linear layer in the model. Each of our linear layers (except the output layer) was followed by a rectified linear unit nonlinearity, to allow complexity of functions that our eSIG-Net model could learn.

For optimization, we employed multi-step learning rate descent with epoch_index = [30] for epoch indices. The learning rate decay factor was set to 0.1. Adam was used as the optimizer for our method, with an initial learning rate of 0.005. The weights α_1 and α_2 for the loss function were set to 0.9 and 0.1, respectively. After training 50 epochs, we selected the optimal checkpoint based on the validation accuracy.

All loss weights and other hyperparameters were tuned on validation sets only. Specifically, in each fold of the fivefold cross-validation, one fold was held out as the test set and the remaining four folds were split into training and validation subsets. Hyperparameters were selected based on validation performance, and the final reported results were obtained on the unseen test fold. No test set information was ever used for model selection.

Fivefold cross-validation

We applied a fivefold cross-validation strategy to avoid the influence of random samples on the performance. The eSIG-Net model trained on PPI datasets was employed to predict the impact of new (previously 'unseen') mutations on PPI perturbation. The data were partitioned into five distinct subsets, each serving as a fold in cross-validation. Hyperparameters were tuned only on the validation subset, and the test fold remained unseen until final evaluation. Reported performance metrics were obtained by averaging across the five test folds, with no fold used simultaneously for model selection and for evaluation, avoiding data leakage. To ensure a stringent assessment, test set proteins were completely different from those in the training sets. This model demonstrates consistently robust performance in prediction across different sequence groups. A use of similarity filter to define the training set and test set was found to make no significant difference, further demonstrating that our model performance was not driven by sequence similarity alone.

Benchmarking with other sequence-based methods

Due to the absence of existing methods specifically tailored to predict changes in the original PPI states caused by missense mutations, we sought to compare our framework against five state-of-the-art PPI prediction methods. While both our framework and the benchmarking methods addressed PPI-related tasks, they differed in their inputs. Our model for predicting changes in PPI states by missense mutations involved triplets composed of the WT protein, the mutant protein and the interactor, with the output indicating whether the binding states of WT–interactor and mutant–interactor changed. In contrast, conventional PPI methods typically take protein–interactor pairs as input, predicting the binding state of the protein–interactor pair. To align inputs for a fair comparison, we split the triplets into WT–interactor and mutant–interactor pairs, treating them as two separate samples for conventional PPI tasks.

The logit output of these sequence-based methods is a continuous value between 0 and 1 (with '1' representing the probability of PPI or binding). The binding state of the mutant–interactor was determined by binarizing the logit value. For mutation-perturbed PPIs, the predicted interaction probabilities of the WT PPI were above 0.5, whereas the probabilities of the mutant PPI fell below 0.5. For nonperturbed PPIs, the predicted interaction probabilities of the WT and mutant PPIs were either 'both above 0.5' (that is, both having PPIs) or 'both below 0.5' (that is, both exhibiting no PPIs). For all benchmark models that required retraining, we used the same model selection strategy as for eSIG-Net. Specifically, a held-out validation set was used to fine-tune training, and the model parameters corresponding to the best epoch on the validation set were selected for testing. This ensured a fair comparison across all methods. For performance evaluation, we compared the predicted binding state of mutant–interactor and WT–interactor pairs, to infer whether the mutation led to a change in interactions.

SDNN. SDNN[13] evaluates diverse protein feature extraction combinations and identifies the most effective feature sets. It further enhances performance by using attention-based networks, achieving results in PPI prediction tasks. In this study, we adopted the optimized feature combinations as reported in ref. 13 and employed their PyTorch codes to train the PPI prediction model.

D-SCRIPT. D-SCRIPT[10] introduces a method for embedding proteins based on their amino acid sequences, aiming to bring proteins with similar structures closer in the embedded space. It uses a stacked three-layer bidirectional LSTM (long short-term memory) for protein embedding, yielding results in PPI tasks. We followed the same pipeline[10] and used D-SCRIPT package to train the prediction model.

DeepFE. DeepFE[14] integrates handcrafted features with Word2vec technology, enabling the creation of protein sequence embeddings that capture intricate semantic relationships among amino acids. These embeddings are then harnessed within deep neural networks, proficiently extracting features, reducing dimensionality and making

predictions of PPIs. In this study, we adopted the original codes[14] to generate the input embeddings and train the prediction model.

PIPR. PIPR[15] incorporates a Siamese neural network featuring a deep residual recurrent convolutional architecture. This design effectively combines local features and contextual information, enabling the capture of mutual influences within protein sequences. Moreover, PIPR simplifies data preprocessing and demonstrates adaptability across various applications. We used the PIPR codes[15] to train the PPI model.

PLM-interact. PLM-interact[16] leverages joint encoding of protein pairs, enabling the model to directly learn relational patterns between sequences. By fine-tuning pretrained transformer-based protein language models with contrastive learning objectives, PLM-interact effectively captures both individual sequence features and inter-protein dependencies. In this study, we used the PLM-interact[16] implementation to fine-tune the model and predict PPI outcomes.

Computation of ROC and AUPR curves

For each mutation–interactor pair, the model output two raw logits corresponding to the 'no-change' and 'change' classes (denoted as pred_logic_0 and pred_logic_1, respectively). To obtain continuous prediction scores suitable for curve-based evaluation, we computed a confidence score as confidence = σ(pred_logic_1 − pred_logic_0), where $\sigma(x)$ is the sigmoid function. These continuous confidence values (ranging from 0 to 1) represent the model's probability of predicting a change, and were directly used to compute the ROC-AUC and average precision (AUPR) metrics using the sklearn.metrics implementation. Binary predictions (pred_change) were only used for discrete comparison, whereas all AUPR and ROC analyses were based on the continuous confidence values.

Sequence-based baseline methods output only an interaction probability $p(\text{PPI}) \in [0,1]$ for a given protein pair and do not provide explicit change or no-change logits. To enable a fair comparison on our mutation-induced PPI perturbation task, we converted each baseline's outputs into a continuous 'confidence-of-change' score using paired predictions for the WT and mutant pairs. Specifically, for each mutation–interactor sample, we computed $p_{\text{WT}} = p(\text{PPI}|\text{WT, interactor})$ and $p_{\text{Mut}} = p(\text{PPI}|\text{Mut, interactor})$, where values >0.5 indicated the presence of an interaction according to the baseline method's definition. We defined PPI state change to include both loss of interaction ($p_{\text{WT}} \geq 0.5$ and $p_{\text{Mut}} < 0.5$) and gain of interaction ($p_{\text{WT}} < 0.5$ and $p_{\text{Mut}} \geq 0.5$). For curve-based evaluation, we used a continuous change score

$$s = |p_{\text{WT}} - p_{\text{Mut}}|,$$

which quantified the magnitude of the mutation-induced change in predicted interaction likelihood. ROC-AUC and AUPR curves for baseline methods were computed using these continuous scores s against the ground-truth change labels (rather than hard-thresholded binary predictions). For our model, a continuous confidence score was directly derived from the change and/or no-change logits (described above), and all curve-based metrics were computed from continuous scores for consistency across methods.

Visualization and quantification of model interpretability

A normalized confusion matrix was used for displaying the prediction, and a t-distributed stochastic neighbor embedding plot was used for the visualization of model interpretability. We used two metrics to compute distances in the original embedding space and to quantify the degree to which different classes of data points (perturbed versus unperturbed PPIs) were separated in the visualization: (1) the separation ratio, which was computed as the between-centroids distance divided by the root-mean-square within-cluster distance. A higher separation ratio generally indicates that the clusters are more distinct and well-separated in the t-stochastic neighbor embedding visualization, suggesting that the data points within each cluster are more similar to each other than to points in other clusters. (2) The silhouette score of each point was computed using the formula: $s = (b - a)/\max(a, b)$, where a represents the average distance to all other points within its cluster, and b represents the average distance to all points in the nearest cluster. The overall silhouette score is the average of all individual silhouette scores. A high silhouette score indicates well-defined and separated clusters.

AlphaFold-based PPI prediction

DeepMind's AlphaFold[17] offers regular atomic accuracy in protein structure prediction through a MSA encoder and builds pairwise representations, a three-dimensional rotation-equivariant network to build structure and an iterative recycling mechanism to optimize structure prediction. This end-to-end structure prediction framework can benefit the analysis of various protein structures, properties and functions[16]. FoldDock[18] leverages the AlphaFold[17] pipeline for protein interaction prediction. To predict mutation-mediated PPI changes, we fed all WT or mutant and interactor sequence samples into the FoldDock predictor. As MSA was executed, the quantification of interface contacts was obtained. A count of fewer than one interface contact was interpreted as a noninteraction or a negative PPI prediction. Given the intensive time requirements for MSA extraction and the substantial graphical processing unit resources demanded by the sophisticated AlphaFold2 model, our study's performance metrics were confined to the disease mutation PPI dataset.

Benchmarking with mutation-centric, structure-based methods

To compare eSIG-Net with other mutation-centric, structure-based methods, we first subjected the WT and interactor sequences to the AlphaFold-Multimer model[19]. This model includes a search for MSAs and predicts the structure of the WT–interactor PPI complex. With the structures predicted, we then incorporated the mutation information into the prediction model to estimate the changes in binding affinity ($\Delta\Delta G$) caused by the mutation. Due to these methods classifying a mutation as deleterious if $\Delta\Delta G \geq 1.5$ or ≤ -1.5 kcal mol^{-1}, we defined such deleterious mutations as perturbing PPI profiles. To illustrate the AUC curves, we converted ($\Delta\Delta G$) into a logit score by performing the following operation: score = sigma ($|\Delta\Delta G| - 1.5$), where (sigma) represents the sigmoid function. The disease mutation PPI dataset was used for benchmarking with all structure-based methods.

Population-based PPI function validation in the context of cancer immunotherapy

To validate the potential impact of mutations on eSIG-net predicted PPI pairs, we obtained transcriptomic and somatic mutation data from The Cancer Genome Atlas and MMRF-COMPASS cohorts via the UCSC Xena platform[33], encompassing 34 cancer types and more than 11,000 patients. For analysis on the response to immunotherapy, patients were stratified into a 'both-high' group (expression of both PPI genes ≥median) versus all others. Statistical significance was assessed using Fisher's exact test.

Statistical analysis

To compare the mean accuracies of eSIG-Net against the other models, pairwise independent t-tests were conducted. Given the multiple comparisons being made (each algorithm against eSIG-Net), it was necessary to adjust for the increased probability of type I errors. To this end, we used the Holm–Bonferroni method for adjusting P values.

Reporting summary

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

Data availability

Two mutation-mediated PPI datasets used in this study were: the disease mutation PPI dataset[2] and the population variant PPI dataset[12]. Protein sequences were obtained from the UniProt database at https://www.uniprot.org/. Source data are provided with this paper.

Code availability

All codes and documentation are available at https://github.com/Stephen-Yi-Laboratory/eSIG-Net.

References

29. Yang, X. et al. A public genome-scale lentiviral expression library of human ORFs. *Nat. Methods* **8**, 659–661 (2011).
30. Du, X. et al. DeepPPI: boosting prediction of protein–protein interactions with deep neural networks. *J. Chem. Inf. Model.* **57**, 1499–1510 (2017).
31. Chen, C., Zhang, Q., Ma, Q. & Yu, B. LightGBM-PPI: predicting protein–protein interactions through LightGBM with multi-information fusion. *Chemometr. Intell. Lab. Syst.* **191**, 54–64 (2019).
32. Zhang, L., Yu, G., Xia, D. & Wang, J. Protein–protein interactions prediction based on ensemble deep neural networks. *Neurocomputing* **324**, 10–19 (2019).
33. Goldman, M. J. et al. Visualizing and interpreting cancer genomics data via the Xena platform. *Nat. Biotechnol.* **38**, 675–678 (2020).

Acknowledgements

We acknowledge assistance by B. Hitz, Z. Shen, K. Lin and S. Dong from the DACC of IGVF (Impact of Genomic Variation on Function) Consortium. The work was supported by NIH grants R35GM133658 (to S.S.Y.), R33CA281919 (to S.S.Y. and G.M.S.), R00CA240689 (to D.J.M.), R35GM137836 (to N.S.), R01HG012366 (to D.F.J.), R35GM128625 (to J.I.F.B.) and U24MH130988 (to W.J.Z.). We also acknowledge the Department of Defense grant W81XWH-22-1-0164 (to W.J.Z.), and the CPRIT (Cancer Prevention & Research Institute of Texas) grant RP240537 (to E.W.). S.S.Y. is an Affiliate Member of the NHGRI IGVF Consortium, a Partner Member of the NHGRI GREGoR (Genomics Research to Elucidate the Genetics of Rare diseases) Consortium and an Associate Member of the NCI Cancer Systems Biology Consortium (CSBC). N.S. was supported by Alfred P. Sloan Scholar Research Fellowship (FG-2018–10723). S.C. was supported by Cancer Research Institute Lloyd Old STAR Award (CRI 4964). S.G.E. acknowledges the Dan L. Duncan Comprehensive Cancer Center grant no. P30CA125123.

Author contributions

S.S.Y., X.P. and N.S. conceived of the study. X.P., A.S. and S.R. conducted most of the computational modeling and analyses, with input from C.D., Y.Y., Z.L., D.J.M., S.S.Y. and N.S. E.W., J.H.H., S.G.E., W.J.Z., J.I.F.B., G.M.S., S.C. and D.F.J. provided intellectual input and constructive feedback. S.S.Y. and N.S. provided supervision throughout the course of the study. X.P., A.S., N.S. and S.S.Y. wrote the paper with input from S.C., D.J.M., J.I.F.B. and D.F.J. All authors read and approved the paper.

Competing interests

The authors declare no competing interests.

Additional information

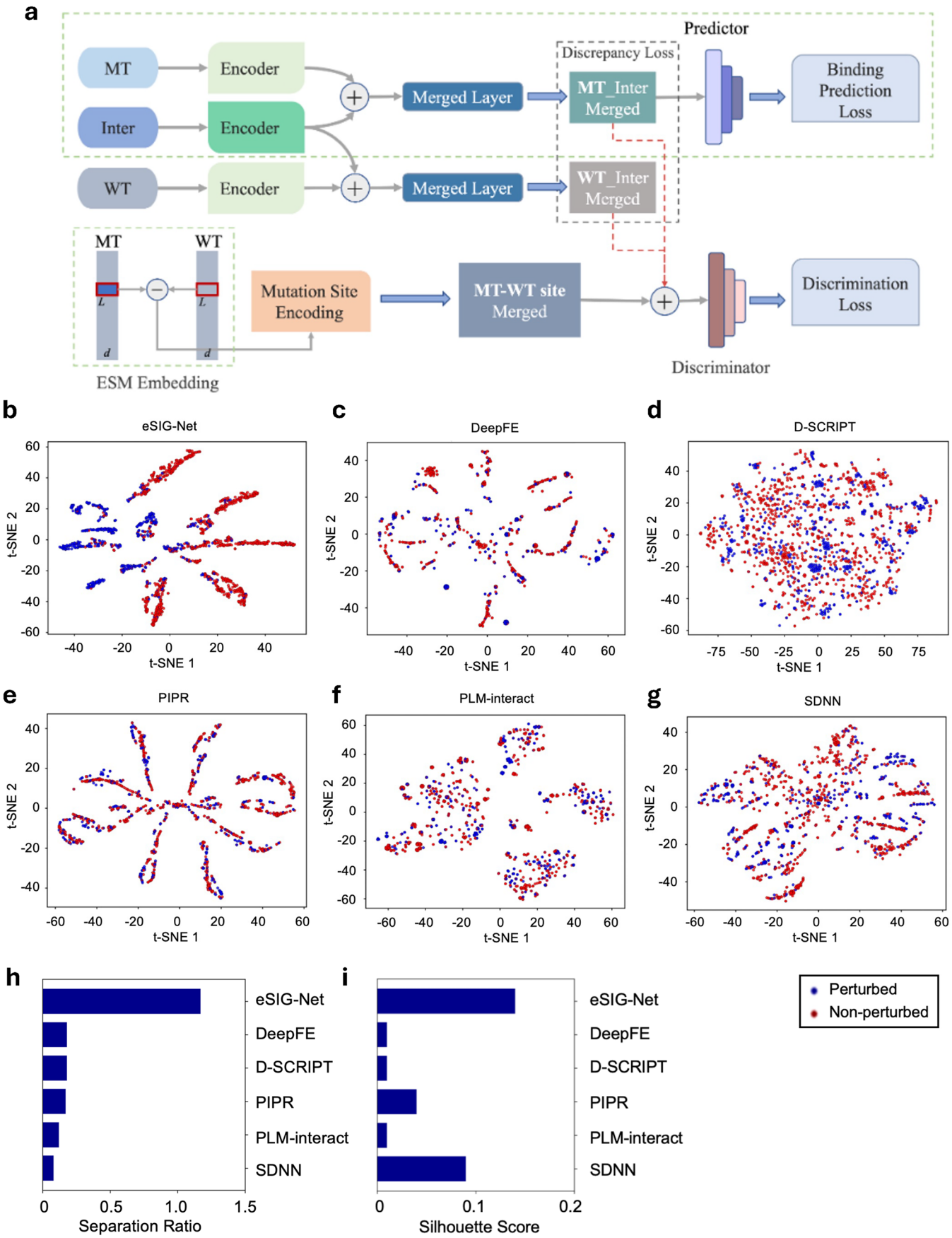
Extended data is available for this paper at https://doi.org/10.1038/s41592-026-03086-x.

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41592-026-03086-x.

Correspondence and requests for materials should be addressed to Xingxin Pan, Jason H. Huang, Nidhi Sahni or S. Stephen Yi.

Peer review information *Nature Methods* thanks Ulrich Stelzl and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Madhura Mukhopadhyay, in collaboration with the *Nature Methods* team.

https://doi.org/10.1038/s41592-026-03086-x

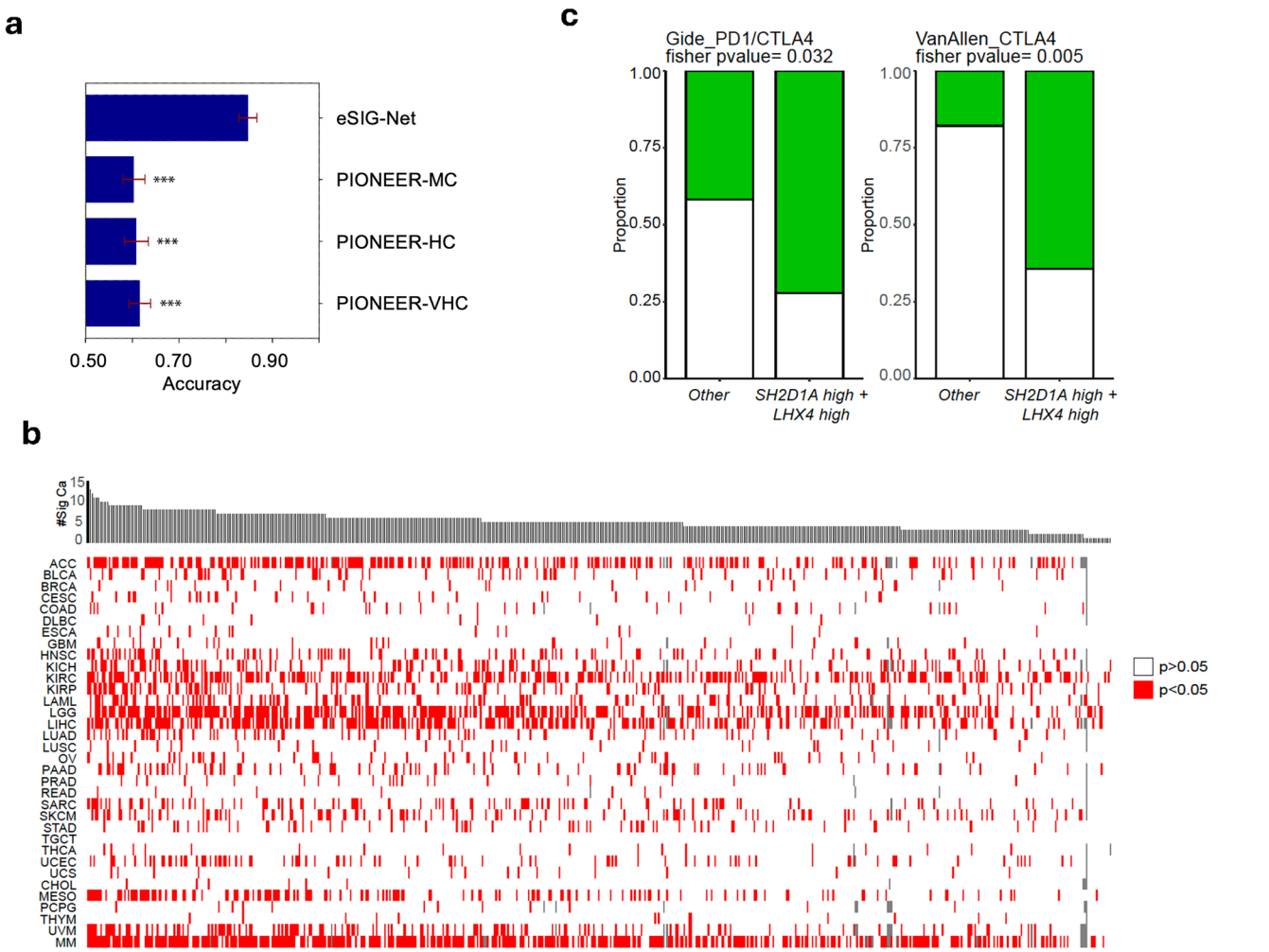


Extended Data Fig. 1 | See next page for caption.

https://doi.org/10.1038/s41592-026-03086-x

Extended Data Fig. 1 | Architecture and interpretability of the eSIG-Net model. (**a**) Two protein encoders are used to encode Mutant (MT)/Wild-type (WT) and Interactor (Inter), with MT and WT encoders sharing weights. The resulting encodings are concatenated and input into the Merged layer, producing MT-Inter merged encoding and WT-Inter merged encoding. Similar to traditional PPI prediction methods, the MT-Inter merged encoding is fed into the predictor for interaction/binding state prediction. Additionally, we calculate the discrepancy loss between MT-Inter merged encoding and WT-Inter merged encoding to constrain the distance between positive and negative samples at the encoding level. We then extract MT and WT embeddings from ESM-2 and subtract the mutation site embeddings to input into the mutation site encoding module to learn differences at the mutation site. Finally, we concatenate MT-Inter merged encoding, WT-Inter merged encoding, and MT-WT site merged encoding and input them into the discriminator for the ultimate PPI perturbation prediction. For details, refer to the *Methods* section. (**b**-**g**) t-SNE dimensional reduction visualization comparing eSIG-Net (**b**) with other state-of-the-art sequence-based methods, including DeepFE (**c**), D-SCRIPT (**d**), PIPR (**e**), PLM-interact (**f**) and SDNN (**g**). Predicted perturbed and unperturbed PPIs are shown in blue and red color, respectively. Normalized confusion matrix is used for displaying the prediction, and t-distributed Stochastic Neighbor Embedding (t-SNE) plot is used for the visualization of model interpretability. The embedding of the output layer's previous hidden layer output from all methods is extracted to perform the visualization. For quantification, two metrics (separation ratio and silhouette score) are computed to quantify the degree to which distinct classes of data points are separated in the visualization. (**h**) Barcharts showing the separation ratio metric to compare methods for their prediction performance on separating perturbed vs non-perturbed PPIs, using the disease mutation dataset. (**i**) Barcharts showing the silhouette score metric to compare methods for their prediction performance, using the disease mutation dataset.

https://doi.org/10.1038/s41592-026-03086-x



Extended Data Fig. 2 | Benchmarking with the state-of-the-art structure-based prediction method PIONEER, and independent validation across different biological contexts. (**a**) Bar chart shows the mean accuracy of the structure-based prediction method PIONEER across different confidence levels (MC: Medium confidence; HC: High confidence; VHC: Very high confidence), using the disease mutation dataset. The length of each bar represents the mean accuracy and the error bars denote the standard deviations ($n = 1{,}633$ PPIs for each prediction model plotted). *P* values are calculated by two-sided paired *t*-tests, with Holm-Bonferroni correction. ***, $P < 1.0 \times 10^{-3}$. The error bars indicate ±s.d. and the centers of the error bars indicate mean accuracy ($n = 3$ independent experiments). (**b**) PPI-paired genes as predicted by eSIG-Net may jointly influence cancer patient survival in TCGA-MMRF cohorts. Patients are stratified into a "both-high" group (expression of both PPI genes ≥ median) versus all other patients. P values are calculated by Cox proportional hazards regression Wald test (two-sided). (**c**) Expression of PPI gene pairs as predicted by eSIG-Net is associated with immunotherapy response in melanoma patients. Gide PD-1/CTLA-4 cohort ($n = 74$); Van Allen CTLA-4 cohort ($n = 43$).

 https://doi.org/10.1038/s41592-026-03086-x

Extended Data Table 1 | AlphaFold-Multimer predictions for WT and mutant proteins are almost indistinguishable at the global structural level

	TMscore_ref1	TMscore_ref2	TMscore_avg	GDT_TS	RMSD	Aligned_length	CA_RMSD_BioPDB
Fold 1	0.8481333	0.8481333	0.8481333	NaN	0.9271938	283.025	6.8795437
Fold 2	0.8317494	0.8317494	0.8317494	NaN	0.9198365	281.8553459	7.1898467
Fold 3	0.8469792	0.8469792	0.8469792	NaN	0.9229367	289.1075949	7.0784096
Fold 4	0.815529	0.815529	0.815529	NaN	0.9194286	274.8571429	8.7055335
Fold 5	0.8647881	0.8647881	0.8647881	NaN	0.9323377	282.192053	5.5984687

The predicted structures are nearly identical, with an average TM-score of 0.84 and a global RMSD of 0.92 Å.

nature portfolio

Corresponding author(s): S. Stephen Yi

Last updated by author(s): Jan 29, 2026

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	No software was used for data collection.
Data analysis	Python (3.7), SDNN, DSCRIPT, DeepFE, PIPR, PLM-interact, FoldDock, MutaBind2, BeAtMuSiC, GeoPPI, TopNetTree, PIONEER, ESM1b, and AlphaFold-Multimer. The source code of eSIG-Net is available at https://github.com/Stephen-Yi-Laboratory/eSIG-Net.

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

All datasets, code and documentation are available at https://github.com/Stephen-Yi-Laboratory/eSIG-Net.

Human research participants

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

Reporting on sex and gender	Not applicable.
Population characteristics	Not applicable.
Recruitment	Not applicable.
Ethics oversight	Not applicable.

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	No statistical method was used to predetermine sample size. Sample size was determined by data availability. All available samples and data were used in our experiments and analyses.
Data exclusions	No data were excluded for data analysis and modeling.
Replication	All attempts at replication were performed independently and were successful.
Randomization	Training, validation and benchmark tests were split randomly following the standard machine learning practice.
Blinding	Blinding was used during benchmark tests following the standard machine learning practice.

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

Eukaryotic cell lines

Policy information about cell lines and Sex and Gender in Research

Cell line source(s)	Cells were purchased from ATCC.
Authentication	Cell lines have been thoroughly authenticated by ATCC.
Mycoplasma contamination	Cells were free of mycoplasma contamination.
Commonly misidentified lines (See ICLAC register)	No commonly misidentified cell lines were used.