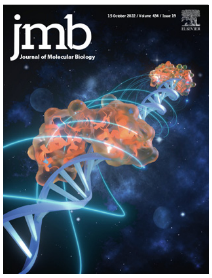


ScanNet: A Web Server for Structure-based Prediction of Protein Binding Sites with Geometric Deep Learning

Jérôme Tubiana [1]*, Dina Schneidman-Duhovny [2] and Haim J. Wolfson [1]

*1 - **Blavatnik School of Computer Science,** Tel Aviv University, Israel*

*2 - **School of Computer Science and Engineering,** The Hebrew University of Jerusalem, Israel*

***Correspondence to Jérôme Tubiana:** jertubiana@gmail.com, jeromet@mail.tau.ac.il (J. Tubiana)@TubianaJerome (J. Tubiana), @DinaSchneidman (D. Schneidman-Duhovny)*

https://doi.org/10.1016/j.jmb.2022.167758

Edited by Michael Sternberg

Abstract

Predicting the various binding sites of a protein from its structure sheds light on its function and paves the way towards design of interaction inhibitors. Here, we report ScanNet, a freely available web server for prediction of protein–protein, protein - disordered protein and protein - antibody binding sites from structure. ScanNet (Spatio-Chemical Arrangement of Neighbors Network) is an end-to-end, interpretable geometric deep learning model that learns spatio-chemical patterns directly from 3D structures. ScanNet consistently outperforms Machine Learning models based on handcrafted features and comparative modeling approaches. The web server is linked to both the PDB and AlphaFoldDB, and supports user-provided structure files. Predictions can be readily visualized on the website via the Molstar web app and locally via ChimeraX. ScanNet is available at http://bioinfo3d.cs.tau.ac.il/ScanNet/.

© 2022 Elsevier Ltd. All rights reserved.

Background

Recent progress in experimental[1] and ML-based methods of protein structure determination[2–7] has led to a spectacular rise in the number of available protein structures. The recently released AlphaFold database[8] contains about 600 K models of protein structures as of January 2022, and is expected to grow to millions of entries. However, the overwhelming majority of these proteins lack any experimental functional annotation, raising the question of how to best leverage this wealth of structural data to yield biological insights. Identifying the functional sites of a protein, such as its catalytic sites, various binding sites or sites of post-translational modification is a first step towards elucidation of its mechanism of action *in vivo* and can guide rational design of inhibitors.

The most common approach for structure-based functional site annotation is comparative modelling[9–14,–17]: given a query structure, proteins with known annotations and similar folds and/or local structural motifs are identified, and their functional sites mapped onto the query. Comparative modelling has however limited applicability in the era of structural bioinformatics: first, it does not scale well to large databases since it involves pairwise comparisons. Second, its coverage is limited to proteins sharing similar fold or structural motifs with experimentally annotated proteins. In contrast, large-scale structure predictions have already led to the discovery of hundreds to thousands of novel protein folds.[7] Last, by construction, comparative modelling has limited sequence sensitivity and therefore cannot be used to monitor the evolution of functional sites within gene families (*e.g.*, antigenic drift in viral proteins).

ML-based models are an appealing alternative to comparative modeling, owing to their speed, sequence sensitivity and ability to generalize to unseen protein families. Early works were based on handcrafted features pipelines: for each amino acid of a given protein, various features of geometrical, physicochemical and evolutionary nature are first extracted and then combined into functional site predictions via a Machine Learning model for tabular data (*e.g.*, decision trees).[18–21,16,17,22] However, ML-based models have been historically limited by the expressivity of the features employed. Indeed, mathematically defined features such as solvent accessibility, molecular surface curvature or hydrophobicity scale cannot capture function-bearing motifs defined by specific arrangements of atoms of amino acids, such as Zinc fingers, or catalytic triads of serine proteases. Recently, several works have explored with varying degrees of success end-to-end learning architectures for protein structures as a mean to directly learn the relevant structural features from raw data.[23–26,24,27–32]

We recently introduced ScanNet (Spatio-Chemical Arrangement of Neighbors Neural Network), a geometric deep learning model for structure-based prediction of binding sites.[32] Given the raw structure file and, optionally, a position-weight matrix, ScanNet iteratively builds representations of atoms and amino acids based on the spatio-chemical arrangement of their neighbors, and exploits them to predict amino acid-wise binding site probabilities. ScanNet consistently outperformed (accuracy-wise and speed-wise) other approaches based on comparative modelling and handcrafted features for prediction of protein–protein binding sites, B-cell epitopes (protein-antibody binding sites).[32] We report here similar findings for protein - disordered protein binding sites. Importantly, ScanNet generalized well to unseen protein folds, unlike comparative modelling. We further analyzed the learned, pharmacophore-like spatio-chemical patterns and corresponding representations that underpin the network predictions. The learned atomic patterns included classical structural motifs, such as backbone-backbone or side chain-backbone hydrogen bonds, or bundled helical fragments. Conversely, some patterns featured a prescribed absence of atoms at certain locations, allowing the identification of *e.g.* solvent-exposed side chain atoms or backbone nitrogens/oxygens available for hydrogen bonding. At the amino acid scale, ScanNet learned complex patterns spanning over variable number of residues, and including ”O-ring” architectures or transmembrane helical patterns. Importantly, the learned representations encompassed numerous known functionally-relevant features of structures, of geometrical (*e.g.*, solvent accessibility, convexity of the molecular surface, secondary structure), physico-chemical (*e.g.*, electrostatic potential), or evolutionary (*e.g.*, conservation) nature. Taken together, the performance benchmark and pattern analysis suggest that ScanNet successfully learned some of the physico-chemical principles of protein–protein interactions.

Here, we report a web server for running ScanNet without local installation. The web server supports prediction of protein–protein, protein-antibody and protein-disordered protein binding sites, for both single-chain and multi-chain assemblies. We first briefly describe the architecture of ScanNet, its use cases and its expected performance. We next describe and illustrate the main options for running the web server.

The ScanNet Model

Overview of the model architecture

We first briefly sketch the architecture of the model, additional details being available from the article describing the method.[32] The architecture of ScanNet is depicted in Figure 1 (reproduced from.[32]) ScanNet takes as input the raw structure file of the protein, and, optionally, a position-weight matrix derived from a multiple sequence alignment of evolutionary-related sequences. ScanNet first extracts a neighborhood around each heavy atom (fixed number of $K = 16$ neighbors, corresponding to a ball of about 4 Å radius), and calculates their coordinates in a local frame centered around the atom and oriented using its covalent bonds. The neighborhood (upper left panel), formally a point cloud with attributes (atom group type) is then passed through a set of trainable spatio-chemical filters. Each filter computes a matching score between the neighborhood and a trainable spatio-chemical pattern (upper middle panel). For instance, the pattern of filter 1 is defined by presence of a *NH* group in the center and an oxygen a few angstroms away in the opposite direction from the two covalent bonds; it corresponds to a hydrogen bond. The pattern of filter 2 is defined by a side chain carbon in the center, in the vicinity of an aromatic ring and a nitrogen group; importantly, it also features a prescribed absence of atoms in the opposite direction from the covalent bonds (gray ellipsoid). Prescribed absence of atoms implies reactivity, and hence is critical for binding site prediction. The resulting atom-wise embeddings (upper right panel) are next pooled at the amino acid level, additional residue-wise information is concatenated (position weight matrix or one-hot encoded sequence) and the process is repeated at the amino acid level (lower panels). Finally, the resulting amino acid-wise embeddings (lower right) are converted to propensity scores via a neighborhood attention module (not shown), which projects the embeddings to scalar values and smoothes them (in a learned fashion) across a neighborhood.

Supported classes of binding sites and expected performance

Currently, ScanNet supports three classes of binding sites: protein–protein binding sites

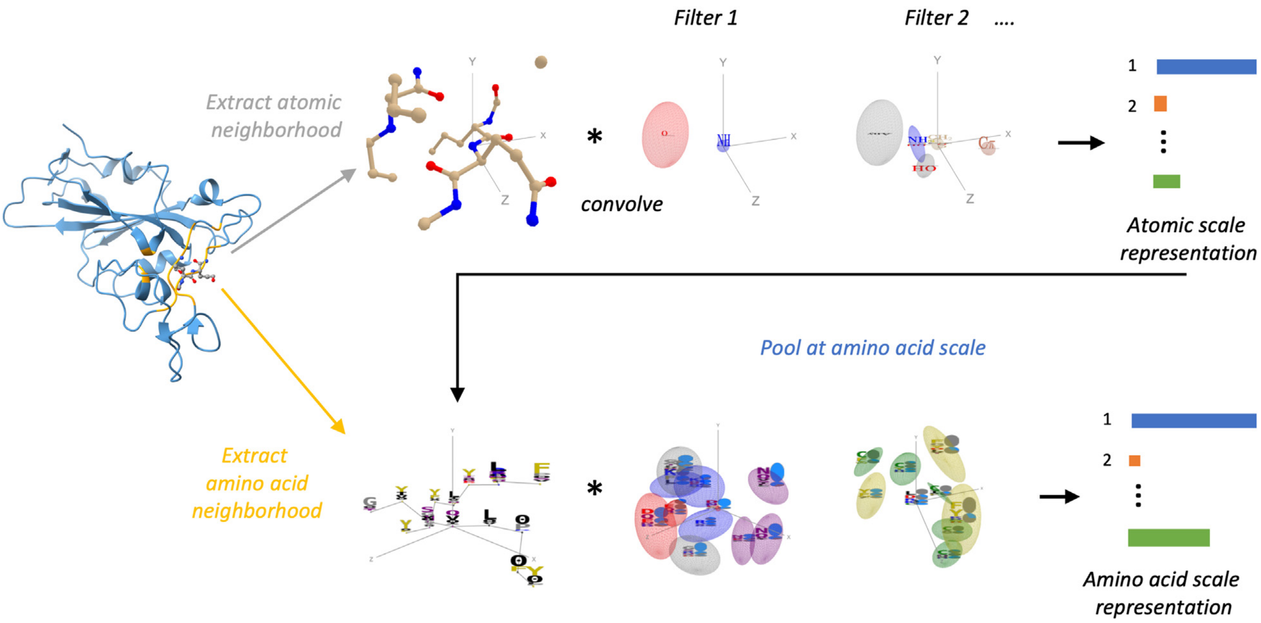

Extract atomic neighborhood
Filter 1
Filter 2 ....
*
convolve
1
2
Atomic scale representation
Pool at amino acid scale
Extract amino acid neighborhood
*
1
2
Amino acid scale representation

Figure 1. Overview of the ScanNet architecture.

(PPBS), protein-disordered proteins binding sites (PIDPBS) and B-cell epitopes (BCE). PPBS are defined as the residues directly involved in one or more native, high affinity protein–protein interaction (4Å or less between at least one of its atoms and the partner). PIDPBS are similarly defined, but the binding region of the partner is prescribed to be disordered. B-cell epitopes (BCE) are defined as residues directly involved in an antibody-antigen complex. Although a priori any surface residue can be targeted by an appropriately matured antibody, some regions elicit stronger humoral response than others, likely because they are easier to bind to with high affinity and specificity. For instance, the anti-spike protein antibodies extracted from sera of patients vaccinated against or recovered from SARS-CoV-2 mainly target the receptor binding domain; and within the receptor binding domain, most antibodies target one of the five main epitope regions.[33] Interactions with both globular and disordered proteins are typically mediated by hydrophobic interactions and involve conserved residues, whereas interactions with antibodies are more heavily based on electrostatic interactions. We used the same network architecture for the three classes; ScanNet was initially trained for PPBS prediction on a large and diverse set of 20 K representative protein chains (95% sequence identity clusters), then fine-tuned for PIDPBS and BCE on respectively 4600 and 800 representative chains. The PPBS and BCE datasets were previously described[32]; for the PIDPBS, we gathered from the PDB all complexes involving one globular and one disordered protein. The later is identified either by its presence in the Disprot database[34] (ensuring at least 50% overlap between the crystallized fragment and the annotated disordered regions) or by its length (between 10 and 30, excluding small peptides which may bind tiny binding sites and long, potentially globular proteins). The pdb identifiers, sample weights, data partitioning and label files for all three datasets are available from https://github.com/jertubiana/ScanNet/tree/main/datasets.

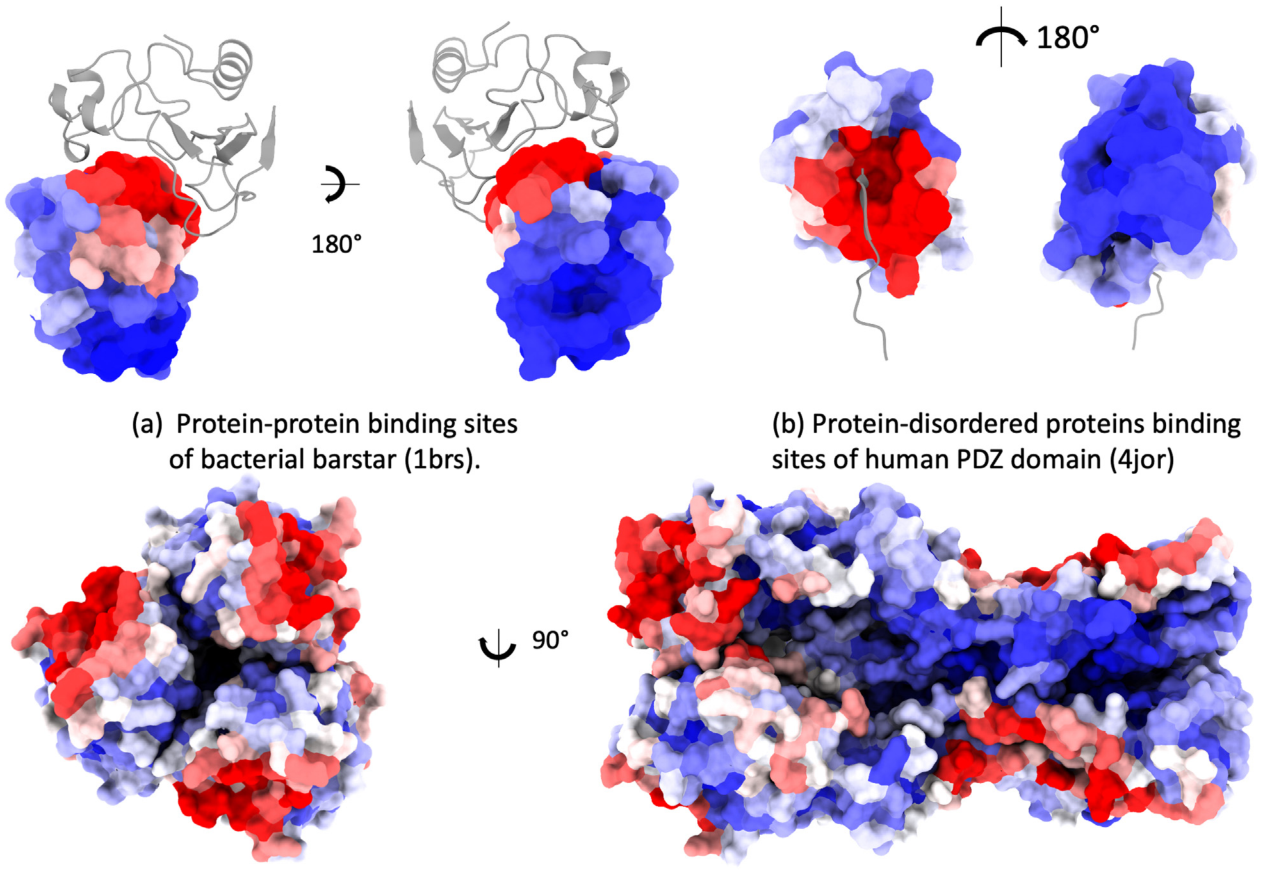
For PPBS prediction, ScanNet reached an average test set accuracy of 87.7% corresponding to 73.5% precision at 50% recall. The model performed equally across phylas, sequence lengths, interaction types (homomer, heteromer or both) and protein types. Performance was however dependent on the degree of homology between the test example and the train set. More specifically, we stratified the test set into four subsets of similar such with decreasing homology to the train set: the first set consisted of examples having at least one homolog with 70–95% sequence identity in the train set, whereas examples in the last test did not have any protein with similar fold topology in the train set (T level of CATH classification). Performance decreased with the degree of homology (from 78% to 63% precision at 50% recall) – indicating that the network recognizes previously encountered protein folds - but much less rapidly than a structural homology baseline method (from 90% to 33% precision at 50% recall) - confirming that it learned general physio-chemical principles underlying protein binding. For PIDPBS prediction, training and performance was evaluated in 5-fold cross-validation setup, where the five sets were such that proteins belonging to the same protein family (H level of CATH classification) were assigned to the same set. In this setting, ScanNet achieved cross-validated accuracy of 91.6%, corresponding to 24.8% precision at 50% recall. The relatively lower precision was due to confusion with regular PPBS, in particular the ones found in homomultimer interfaces. We however found that in a test setting, such errors can be easily avoided by providing the native homomultimer complex as input and performing multi-chain prediction rather than providing a single chain (see next section). For BCE prediction, a similar cross-validation setup was used, where partitioning was done using a 70% sequence identity cut-off since epitopes are weakly conserved throughout evolution. The network achieved a cross-validated positive predicted value at L/10 of 27.5% (*i.e.*, the fraction of correct epitopes in the top-10% highest scoring residues). This number is likely an underestimation of the true one as for most antigens, not all the epitopes are known. For the well-studied receptor binding domain of the SARS-CoV-2 spike protein, we found a Spearman correlation of 0.75 between the predicted BCE probability and the empirical antibody hit rate as computed from available PDB structures.[32] Selected examples of predictions for the three classes of binding sites are shown in Figure 2.

The ScanNet Web Server

Inputs description

Input structure The model takes as input a structure for a protein or a multimeric assembly. Three formats are supported; if applicable, the first two formats are recommended, as the results will be sent directly if the same query was already submitted previously.

- A PDB ID (*e.g.*, 2okj). The experimental structure file is fetched from the Protein Data Bank. The user can specify whether to use the asymmetric unit file or the first biological assembly file in the advanced options



(a) Protein-protein binding sites of bacterial barstar (1brs).

(b) Protein-disordered proteins binding sites of human PDZ domain (4jor)

(c) B-cell epitopes of Influenza H3 Hemagglutinin (1rvx)

Figure 2. Selected examples of ScanNet binding site predictions overlaid on the structures Proteins are represented in molecular surface representation using ChimeraX,[35] with colors indicating binding propensity, from low (blue) to high (red). Predictions were generated using the web server as follows: (a) Input structure: 1brs; Chain Identifiers: D; Binding site type: Protein–protein; default value elsewhere. (b) Input structure: 4jor; Chain Identifiers: A; Binding site type: Protein-disordered proteins; default value elsewhere. (c) Input structure: 1rvx; Chain Identifiers: All; Binding site type: Protein-antibody; Multiple Sequence Alignment: no; default value elsewhere.

panel (see https://pdb101.rcsb.org/learn/guide-to-understanding-pdb-data/biological-assemblies for a tutorial on the difference between both). By default, the biological assembly file is used, if it exists.
- A Uniprot ID (*e.g.*, P38398). The corresponding AlphaFold model is fetched from the AlphaFold Database, if it exists. Although we have not thoroughly benchmarked the performance of ScanNet on AlphaFold models, we did not find qualitative changes in predictions for high confidence regions. For regions that are predicted to be disordered and/or with low confidence, the network predictions are not expected to be well calibrated. Indeed, in the crystal structures used for training, unfolded linear regions appear mostly only if they are in complex with a partner; otherwise, their disorder level is too high and they are typically missing from the file. Therefore, disordered regions systematically have higher binding propensity than ordered ones. Notwithstanding this bias, ScanNet predictions still exhibit a potentially significant variability within disordered regions. The later variability mainly arises from the chemical properties of individual amino acids and the detection of linear patterns in the sequence (see [32]).
- A local file. Input file must be in either PDB or mmCIF file format; the correct extension must be appended to the file name (.pdb or.cif). The current implementation was tested for assemblies with up to about 3000 residues.

Of note, the current implementation discards all solvent molecules and non-protein cofactors. If provided, the coordinates of the hydrogen atoms are ignored. Post-translational modifications are ignored and exotic residues are treated as regular ones.

Chain identifiers For multi-chain structure files, all the chains in the file are processed by default. To specify the chain(s) of interest, two formats are supported:

- A specific chain or semicolon separated list of chains (*e.g.*, A or A;B;C). If the file has multiple models, the first model is used.
- A pair model:chain or semicolon separated list of model:chain pairs (*e.g.* 2:A or 0:A;1:A;2:B). Models are indexed starting from 0.

Binding site type Three types of binding sites are currently supported, see description above.

Email address ScanNet results are sent via email once the run is completed. We recommend checking the spam folder if the results are not send after a few minutes.

Job ID An optional job name.

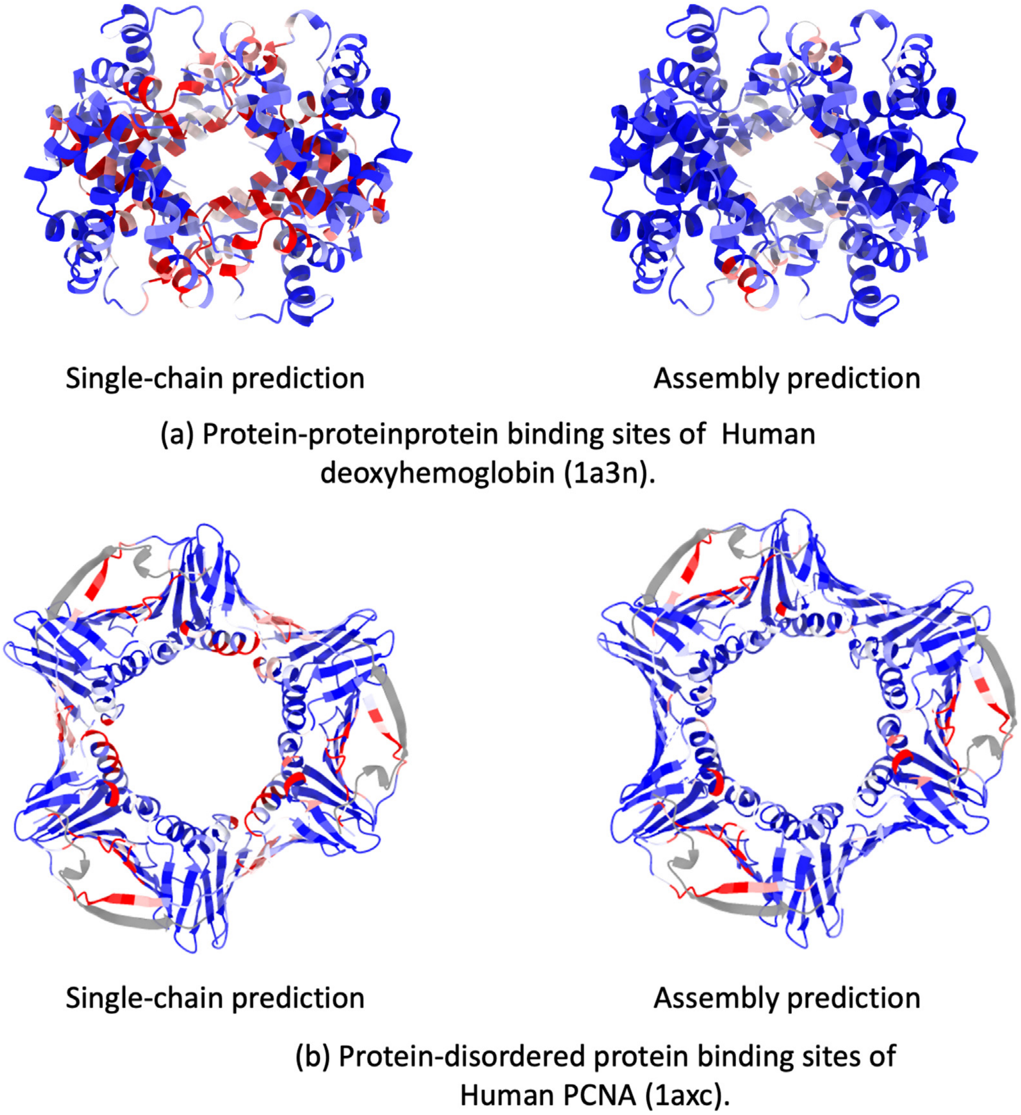
Multi-chain or Single-chain prediction This option is only relevant if binding sites are predicted for several chains at once, and specifies whether the binding sites should be computed for the chains taken together as a single biological assembly (default) or independently from one another. In the first case, the atomic and amino acid neighborhoods include all the atoms/residues present in the file (or the selection of chains) whereas in the second case, the neighborhood of a given atom (resp. residue) is only constituted by atoms (resp. residues) belonging to the same chain. The first option is preferred if the protein of interest is assembled with other proteins in its native state. Residues that are exposed for isolated chains but buried within a biological assembly will not be predicted as binding sites. Example of use case include prediction of the B-cell conformational epitopes of the SARS-CoV-2 spike protein trimer, and of the disordered binding sites of Human calcineurin (which contains two sub-units). The second option should be preferred if i) multiple conformations of the same protein are provided in a single file (e.g. NMR ensembles) ii) the complex depicted in the structure file is crystal-induced, and not representative of the native biological state, iii) one wants to investigate a posteriori the interfaces of a protein–protein complex structure or model or iv) the complex is very large. Figure 3 illustrates the difference between multi-chain and single-chain predictions on two examples.

Use Multiple Sequence Alignment or not Whether to use evolutionary information to perform prediction or not (default is yes). If yes, a search for homologous sequences over UniRef is performed using HH-blits,[36] a Multiple Sequence Alignment (MSA) is constructed and a Position Weight Matrix (PWM) is derived and provided to the network. Usage of evolutionary information substantially improves performance for protein–protein and protein-disordered protein binding sites as they are often more conserved than other surface residues. Conversely, evolutionary information does not significantly improve performance for prediction of B-cell epitopes. MSA should not be used if fast results are favored (as this is the current computational bottleneck) or if the input protein is designed and not natural (as evolution does not reflect functionality anymore).

Outputs description and runtime

After computation, the user is redirected to a user-friendly web page featuring an interactive 3D visualization of the query structure colored by protein binding propensity. The protein structure is rendered using Molstar.[37] In addition, a zip archive containing the result files is sent over via email. Four files are created:

- The pdb file with binding site probabilities provided in the B-factor field.
- A csv file containing the binding site probabilities in comma-separated value (csv) format.
- A cxc script for visualizing the results with ChimeraX.[35]



Single-chain prediction

Assembly prediction

(a) Protein-proteinprotein binding sites of Human deoxyhemoglobin (1a3n).

Single-chain prediction

Assembly prediction

(b) Protein-disordered protein binding sites of Human PCNA (1axc).

Figure 3. Difference between single-chain and multi-chain prediction Structures are depicted in cartoon representation, with colors indicating binding propensity. In single-chain mode (left panels), the chains of the structure file are processed independently from one another; the binding sites involved in internal interfaces, *e.g.* between different components of the hemoglobin tetramer are identified. In multi-chain mode (right panels), the selected chains are processed jointly; the internal binding sites are buried and hence not identified, whereas the other sites stand out. The panels were generated using the following parameters. Top: Input structure: 1a3n; Chain IDs: all; Binding site type: protein–protein; single-chain (left panel) and multi-chain (right panel) prediction. Bottom: Input structure: 1axc; Chain IDs: “A;C;E”; Binding site type: protein-disordered protein; single-chain (left panel) and multi-chain (right panel) prediction.

- A python script for visualizing the results with Chimera, the predecessor of ChimeraX.

Each ScanNet run typically takes a few minutes depending on the size and nature of the query. The runtime is dominated, in order, by i) the construction of the multiple sequence alignment ii) loading and compilation of the network iii) the downloading and parsing of the structure file and iv) actual inference itself.

Concluding Remarks

A technical breakthrough was recently achieved in computational protein structure prediction. Turning this unprecedented wealth of structural data into meaningful biological discoveries requires the development of appropriate tools for function prediction from structure. ScanNet represents one small step towards this direction, and we hereby make it easily accessible to the scientific community. Future prospects include extension to different types of binding sites (RNA, DNA, small molecules), partner-specific predictions and protein-level functional annotations. Historically, structure-based methods have consistently outperformed sequence-based methods, as knowledge of the geometry of the protein facilitates annotation; for instance catalytic sites are usually easily identified as concave regions of the molecular surface.[16] However, the emergence of increasingly complex sequence models trained via self-supervised learning and that implicitly encode structure could disrupt this hierarchy.[38–45,45,45,46] We nonetheless anticipate that relatively simple structure-based models should prove more interpretable and explainable. Beyond accuracy, future efforts should also be directed towards providing comprehensible explanations of the predictions.

Data availability

No data was used for the research described in the article.

DECLARATION OF COMPETING INTEREST

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Acknowledgements

J.T. acknowledges financial support from the Edmond J. Safra Center for Bioinformatics at Tel Aviv University and from the Human Frontier Science Program (cross-disciplinary postdoctoral fellowship LT001058/2019-C). D.S. was supported by ISF 1466/18, Israel ministry of Science and Technology and HUJI-CIDR. This work was supported by Len Blavatnik and the Blavatnik Family Foundation. We are grateful to Sonia Lichtenzveig Sela and the TAU-CS system team for their technical support.

Received 1 February 2022;
Accepted 19 July 2022;
Available online xxxx

References

1. Kühlbrandt, W., (2014). The resolution revolution. *Science* **343**, 1443.
2. Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., et al., (2021). Highly accurate protein structure prediction with alphafold. *Nature* **1**
3. Tunyasuvunakool, K., Adler, J., Wu, Z., Green, T., Zielinski, M., Žídek, A., Bridgland, A., Cowie, A., et al., (2021). Highly accurate protein structure prediction for the human proteome. *Nature* **1**
4. Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G.R., Wang, J., Cong, Q., et al., (2021). Accurate prediction of protein structures and interactions using a 3-track network. *Science*.
5. Chowdhury, R., Bouatta, N., Biswas, S., Rochereau, C., Church, G.M., Sorger, P.K., AlQuraishi, M.N., (2021). Single-sequence protein structure prediction using language models from deep learning. *bioRxiv*.
6. Weißenow, K., Heinzinger, M., Rost, B., (2021). Protein language model embeddings for fast, accurate, alignment-free protein structure prediction. *bioRxiv*.
7. Kandathil, S.M., Greener, J.G., Lau, A.M., Jones, D.T., (2022). Ultrafast end-to-end protein structure prediction enables high-throughput exploration of uncharacterized proteins. *Proc. Nat. Acad. Sci.* **119**
8. Varadi, M., Anyango, S., Deshpande, M., Nair, S., Natassia, C., Yordanova, G., et al., (2022). Alphafold protein structure database: Massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic acids research* **50**, D439.
9. Shulman-Peleg, A., Nussinov, R., Wolfson, H.J., (2005). Siteengines: recognition and comparison of binding sites and protein–protein interfaces. *Nucleic Acids Res.* **33**, W337.
10. Carl, N., Konc, J., Vehar, B., Janezic, D., (2010). Protein-protein binding site prediction by local structural alignment. *J. Chem. Informat. Model.* **50**, 1906.
11. Zhang, Q.C., Petrey, D., Norel, R., Honig, B.H., (2010). Protein interface conservation across structure space. *Proc. Nat. Acad. Sci.* **107**, 10896.
12. Xue, L.C., Dobbs, D., Honavar, V., (2011). Homppi: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinformat.* **12**, 1.
13. Shoemaker, B.A., Zhang, D., Tyagi, M., Thangudu, R.R., Fong, J.H., Marchler-Bauer, A., Bryant, S.H., Madej, T., et al., (2012). Ibis (inferred biomolecular interaction server) reports, predicts and integrates multiple types of conserved interactions for proteins. *Nucleic Acids Res.* **40**, D834.
14. Jordan, R.A., Yasser, E.-M., Dobbs, D., Honavar, V., (2012). Predicting protein-protein interface residues using local surface structural similarity. *BMC Bioinformat.* **13**, 1.
15. Esmaielbeiki, R., Nebel, J.-C., (2012). Unbiased protein interface prediction based on ligand diversity quantification. *Noop*.
16. Xue, L.C., Dobbs, D., Bonvin, A.M., Honavar, V., (2015). Computational prediction of protein interfaces: A review of data driven methods. *FEBS Lett.* **589**, 3516.
17. Esmaielbeiki, R., Krawczyk, K., Knapp, B., Nebel, J.-C., Deane, C.M., (2016). Progress and challenges in predicting protein interfaces. *Briefings Bioinformat.* **17**, 117.
18. Neuvirth, H., Raz, R., Schreiber, G., (2004). Promate: a structure based prediction program to identify the location of protein–protein binding sites. *J. Mol. Biol.* **338**, 181.
19. Chung, J.-L., Wang, W., Bourne, P.E., (2006). Exploiting sequence and structure homologs to identify protein–protein binding sites, Proteins: Structure. *Funct. Bioinformat.* **62**, 630.
20. Porollo, A., Meller, J., (2007). Prediction-based fingerprints of protein–protein interactions, Proteins: Structure. *Funct. Bioinformat.* **66**, 630.
21. Sweredoski, M.J., Baldi, P., (2008). Pepito: improved discontinuous b-cell epitope prediction using multiple

distance thresholds and half sphere exposure. *Bioinformatics* **24**, 1459.

22. Mishra, S.K., Kandoi, G., Jernigan, R.L., (2019). Coupling dynamics and evolutionary information with structure to identify protein regulatory and functional binding sites, Proteins: Structure. *Funct. Bioinformat.* **87**, 850.
23. Townshend, R., Bedi, R., Suriana, P., Dror, R., (2019). End-to-end learning on 3d protein structure for interface prediction. *Adv. Neural Informat. Process. Syst.* **32**, 15642.
24. Wang, X., Terashi, G., Christoffer, C.W., Zhu, M., Kihara, D., (2020). Protein docking model evaluation by 3d deep convolutional neural networks. *Bioinformatics* **36**, 2113.
25. Gainza, P., Sverrisson, F., Monti, F., Rodola, E., Boscaini, D., Bronstein, M., Correia, B., (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184.
26. Jing, B., Eismann, S., Suriana, P., Townshend, R.J. & Dror. R. (2020). Learning from protein structure with geometric vector perceptrons, arXiv preprint arXiv:2009.01411 (2020).
27. Wang, X., Flannery, S.T., Kihara, D., (2021). Protein docking model evaluation by graph neural networks. *Front. Mol. Biosci.* **8**, 402.
28. Renaud, N., Geng, C., Georgievska, S., Ambrosetti, F., Ridder, L., Marzella, D.F., Réau, M.F., Bonvin, A.M., et al., (2021). Deeprank: a deep learning framework for data mining 3d protein-protein interfaces. *Nat. Commun.* **12**, 1.
29. Réau, M., Renaud, N., Xue, L.C., Bonvin, A.M., (2021). Deeprank-gnn: A graph neural network framework to learn patterns in protein-protein interfaces. *bioRxiv*.
30. Sverrisson, F., Feydy, J., Correia, B.E., Bronstein, M.M., (2021). Fast end-to-end learning on protein surfaces. *In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 15272–15281.
31. Eismann, S., Townshend, R.J., Thomas, N., Jagota, M., Jing, B., Dror, R.O., (2021). Hierarchical, rotation-equivariant neural networks to select structural models of protein complexes, Proteins: Structure. *Funct. Bioinformat.* **89**, 493.
32. Tubiana, J., Schneidman-Duhovny, D., Wolfson, H.J., (2021). Scannet: An interpretable geometric deep learning model for structure-based protein binding site prediction. *bioRxiv*.
33. Yuan, M., Huang, D., Lee, C.-C.D., Wu, N.C., Jackson, A. M., Zhu, X., Liu, H., Peng, L., et al., (2021). Structural and functional ramifications of antigenic drift in recent sars-cov-2 variants. *Science*.
34. Hatos, A., Hajdu-Soltész, B., Monzon, A.M., Palopoli, N., Álvarez, L., Aykac-Fas, B., Bassot, C., Benítez, G.I., et al., (2020). Disprot: intrinsic protein disorder annotation in 2020. *Nucleic Acids Res.* **48**, D269.
35. Goddard, T.D., Huang, C.C., Meng, E.C., Pettersen, E.F., Couch, G.S., Morris, J.H., Ferrin, T.E., (2018). Ucsf chimerax: Meeting modern challenges in visualization and analysis. *Protein Sci.* **27**, 14.
36. Remmert, M., Biegert, A., Hauser, A., Söding, J., (2012). Hhblits: lightning-fast iterative protein sequence searching by hmm-hmm alignment. *Nat. Methods* **9**, 173.
37. Sehnal, D., Bittrich, S., Deshpande, M., Svobodová, R., Berka, K., Bazgier, V., Velankar, S., Burley, S.K., et al., (2021). Mol* viewer: modern web app for 3d visualization and analysis of large biomolecular structures. *Nucleic Acids Res.* **49**, W431.
38. Rives, A., Meier, J., Sercu, T., Goyal, S., Lin, Z., Liu, J., Guo, D., Ott, M., et al., (2021). Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proc. Nat. Acad. Sci.* **118**
39. Rao, R., Meier, J., Sercu, T., Ovchinnikov, S., Rives, A., (2020). Transformer protein language models are unsupervised structure learners. *In: International Conference on Learning Representations*.
40. Meier, J., Rao, R., Verkuil, R., Liu, J., Sercu, T., Rives, A., (2021). Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv. Neural Informat. Process. Syst.* **34**
41. Vig, J., Madani, A., Varshney, L.R., Xiong, C., Socher, R. & Rajani, N.F. (2020). Bertology meets biology: Interpreting attention in protein language models, arXiv preprint arXiv:2006.15222.
42. Nambiar, A., Heflin, M., Liu, S., Maslov, S., Hopkins, M., Ritz, A., (2020). Transforming the language of life: transformer neural networks for protein prediction tasks. *In: Proceedings of the 11th ACM International Conference on Bioinformatics* Computational Biology and Health Informatics*, pp. 1–8.
43. Heinzinger, M., Elnaggar, A., Wang, Y., Dallago, C., Nechaev, D., Matthes, F., Rost, B., (2019). Modeling aspects of the language of life through transfer-learning protein sequences. *BMC Bioinformat.* **20**, 1.
44. Elnaggar, A., Heinzinger, M., Dallago, C., Rihawi, G., Wang, Y., Jones, L., Gibbs, T. & Feher, T. et al. (2020). Prottrans: towards cracking the language of life's code through self-supervised deep learning and high performance computing, arXiv preprint arXiv:2007.06225.
45. Littmann, M., Heinzinger, M., Dallago, C., Weissenow, K., Rost, B., (2021). Protein embeddings and deep learning predict binding residues for various ligand classes. *Sci. Rep.* **11**, 1.
46. Littmann, M., Heinzinger, M., Dallago, C., Olenyi, T., Rost, B., (2021). Embeddings from deep learning transfer go annotations beyond homology. *Sci. Rep.* **11**, 1.