

# DeepRank2: Mining 3D Protein Structures with Geometric Deep Learning

Giulia Crocioni<sup>1\*</sup>, Dani L. Bodor<sup>1\*</sup>, Coos Baakman<sup>2\*</sup>, Farzaneh M. Parizi<sup>2</sup>, Daniel T. Rademaker<sup>2</sup>, Gayatri Ramakrishnan<sup>2</sup>, Sven van der Burg<sup>1</sup>, Dario F. Marzella<sup>2</sup>, João M. C. Teixeira<sup>3</sup>, and Li C. Xue<sup>2</sup>

<sup>1</sup> Netherlands eScience Center, Amsterdam, The Netherlands <sup>2</sup> Department of Medical BioSciences, Radboud University Medical Center, Nijmegen, The Netherlands <sup>3</sup> Independent Researcher ¶ Corresponding author \* These authors contributed equally.

DOI: 10.xxxxxx/draft

## Software

- [Review](#)
- [Repository](#)
- [Archive](#)

Editor: Richard Gowers

## Reviewers:

- [@jnwei](#)
- [@abhishektiware](#)

Submitted: 22 September 2023

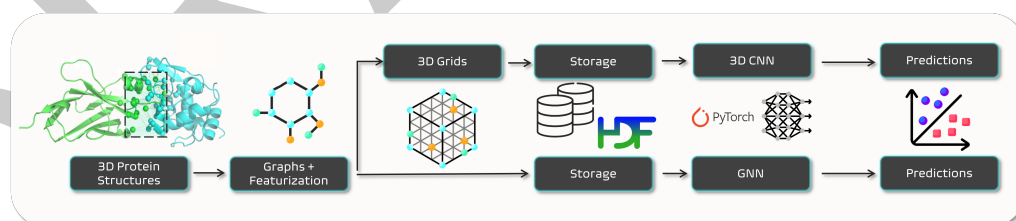
Published: unpublished

## License

Authors of papers retain copyright and release the work under a Creative Commons Attribution 4.0 International License (CC BY 4.0).

## Summary

We present DeepRank2, a deep learning (DL) framework geared towards making predictions on 3D protein structures for variety of biologically relevant applications. Our software can be used for predicting structural properties in drug design, immunotherapy, or designing novel proteins, among other fields. DeepRank2 allows for transformation and storage of 3D representations of both protein-protein interfaces (PPIs) and protein single-residue variants (SRVs) into either graphs or volumetric grids containing structural and physico-chemical information. These can be used for training neural networks for a variety of patterns of interest, using either our pre-implemented training pipeline for graph neural networks (GNNs) or convolutional neural networks (CNNs) or external pipelines. The entire framework flowchart is visualized in Figure 1. The package is fully open-source, follows the community-endorsed FAIR principles for research software, provides user-friendly APIs, publicly available documentation, and in-depth tutorials.



**Figure 1:** DeepRank2 framework overview. 3D coordinates of protein structures are extracted from PDB files and converted into graphs, using either an atomic or a residue level, depending on the user's requirements. Then, the data are enriched with geometrical and physicochemical information, and eventually mapped to 3D grids, before finally storing them into HDF5 files. The processed data can be used in the pre-implemented DL pipeline for training PyTorch networks and computing predictions.

## State of the field

The 3D structure of proteins and protein complexes provides fundamental information to understand biological processes at the molecular scale. Exploiting or engineering these molecules is key for many biomedical applications such as drug design (Gane & Dean, 2000), immunotherapy (Sadelain et al., 2013), or designing novel proteins (Liu et al., 2007). For example, PPI data can be harnessed to address critical challenges in the computational prediction of peptides presented on the major histocompatibility complex (MHC) protein, which play a key role in T-cell immunity. Protein structures can also be exploited in molecular diagnostics for the

29 identification of SRVs, that can be pathogenic sequence alterations in patients with inherited  
30 diseases (B. Li et al., 2020; Shroff et al., 2020).

31 In the past decades, a variety of experimental methods (e.g., X-ray crystallography, nuclear  
32 magnetic resonance, cryogenic electron microscopy) have determined and accumulated a  
33 large number of atomic-resolution 3D structures of proteins and protein-protein complexes  
34 (Schwede, 2013). Since experimental determination of structures is a tedious and expensive  
35 process, several computational prediction methods have been developed over the past decades,  
36 exploiting classical molecular modelling (Baek et al., 2021; Dominguez et al., 2003; Sanchez  
37 & Sali, 1997), and, more recently, DL (Jumper et al., 2021; Richard Evans, 2021). The large  
38 amount of data available makes it possible to use DL to leverage 3D structures and learn their  
39 complex patterns. Unlike other machine learning (ML) techniques, deep neural networks hold  
40 the promise of learning from millions of data points without reaching a performance plateau  
41 quickly, which is made computationally feasible by hardware accelerators (i.e., GPUs, TPUs)  
42 and parallel file system technologies.

43 The main types of data structures in vogue for representing 3D structures are 3D grids, graphs,  
44 and surfaces. 3D CNNs have been trained on 3D grids for the classification of biological  
45 vs. crystallographic PPIs (Renaud et al., 2021), and for the scoring of models of protein-protein  
46 complexes generated by computational docking (Renaud et al., 2021; Wang et al., 2020). Gaiza  
47 et al. have applied geodesic CNNs to extract protein interaction fingerprints by applying 2D  
48 CNNs on spread-out protein surface patches (Gainza et al., 2023). 3D CNNs have been used  
49 for exploiting protein structure data for predicting mutation-induced changes in protein stability  
50 (B. Li et al., 2020; Ramakrishnan et al., 2023) and identifying novel gain-of-function mutations  
51 (Shroff et al., 2020). Contrary to CNNs, in GNNs the convolution operations on graphs can  
52 rely on the relative local connectivity between nodes and not on the data orientation, making  
53 graphs rotationally invariant. Additionally, GNNs can accept any size of graph, while in a CNN  
54 the size of the 3D grid for all input data needs to be the same, which may be problematic  
55 for datasets containing highly variable in size structures. Based on these arguments, different  
56 GNN-based tools have been designed to predict patterns from PPIs (Fout et al., 2017; Réau  
57 et al., 2022; Wang et al., 2021). Eisman et al. developed a rotation-equivariant neural network  
58 trained on point-based representation of the protein atomic structure to classify PPIs (Eismann  
59 et al., 2021).

## 60 Statement of need

61 Data mining 3D structures of proteins presents several challenges. These include complex  
62 physico-chemical rules governing structural features, the possibility of characterization at  
63 different scales (e.g., atom-level, residue level, and secondary structure level), and the large  
64 diversity in shape and size. Furthermore, because a structure can easily comprise of hundreds  
65 to thousands of residues (and ~15 times as many atoms), efficient processing and featurization  
66 of many structures is critical to handle the computational cost and file storage requirements.  
67 Existing software solutions are often highly specialized and not developed as reusable and  
68 flexible frameworks, and cannot be easily adapted to diverse applications and predictive tasks.  
69 Examples include DeepAtom (Y. Li et al., 2019) for protein-ligand binding affinity prediction  
70 only, and MaSIF (Gainza et al., 2023) for deciphering patterns in protein surfaces. While some  
71 frameworks, such as TorchProtein and TorchDrug (Zhu et al., 2022), configure themselves  
72 as general-purpose ML libraries for both molecular sequences and 3D structures, they only  
73 implement geometric-related features and do not incorporate fundamental physico-chemical  
74 information in the 3D representation of molecules.

75 These limitations create a growing demand for a generic and flexible DL framework that  
76 researchers can readily utilize for their specific research questions while cutting down the  
77 tedious data preprocessing stages. Generic DL frameworks have already emerged in diverse  
78 scientific fields, such as computational chemistry (e.g., DeepChem (Ramsundar et al., 2019))  
79 and condensed matter physics (e.g., NetKet (Vicentini et al., 2022)), which have promoted

80 collaborative efforts, facilitated novel insights, and benefited from continuous improvements  
81 and maintenance by engaged user communities.

## 82 Key features

83 DeepRank2 allows to transform and store 3D representations of both PPIs and SRVs into 3D  
84 grids or graphs containing both geometric and physico-chemical information, and provides a  
85 DL pipeline that can be used for training pre-implemented neural networks for a given pattern  
86 of interest to the user. DeepRank2 is an improved and unified version of three previously  
87 developed packages: [DeepRank](#), [DeepRank-GNN](#), and [DeepRank-Mut](#).

88 As input, DeepRank2 takes [PDB-formatted](#) atomic structures, which is one of the standard and  
89 most widely used formats in the field of structural biology. These are mapped to graphs, where  
90 nodes can represent either residues or atoms, as chosen by the user, and edges represent the  
91 interactions between them. The user can configure two types of 3D structures as input for the  
92 featurization phase: - PPIs, for mining interaction patterns within protein-protein complexes; -  
93 SRVs, for mining mutation phenotypes within protein structures.

94 The physico-chemical and geometrical features are then computed and assigned to each node  
95 and edge. The user can choose which features to generate from several pre-existing options  
96 defined in the package, or define custom features modules, as explained in the documentation.  
97 Examples of pre-defined node features are the type of the amino acid, its size and polarity, as  
98 well as more complex features such as its buried surface area and secondary structure features.  
99 Examples of pre-defined edge features are distance, covalency, and potential energy. A detailed  
100 list of predefined features can be found in the [documentation's features page](#). Graphs can  
101 either be used directly or mapped to volumetric grids (i.e., 3D image-like representations),  
102 together with their features. Multiple CPUs can be used to parallelize and speed up the  
103 featurization process. The processed data are saved into HDF5 files, designed to efficiently  
104 store and organize big data. Users can then use the data for any ML or DL framework suited  
105 for the application. Specifically, graphs can be used for the training of GNNs, and 3D grids  
106 can be used for the training of CNNs.

107 DeepRank2 also provides convenient pre-implemented modules for training simple [PyTorch](#)-  
108 based GNNs and CNNs using the data generated in the previous step. Alternatively, users can  
109 implement custom PyTorch networks in the DeepRank package (or export the data to external  
110 software). Data can be loaded across multiple CPUs, and the training can be run on GPUs.  
111 The data stored within the HDF5 files are read into customized datasets, and the user-friendly  
112 API allows for selection of individual features (from those generated above), definition of the  
113 targets, and the predictive task (classification or regression), among other settings. Then the  
114 datasets can be used for training, validating, and testing the chosen neural network. The final  
115 model and results can be saved using built-in data exporter modules.

116 DeepRank2 embraces the best practices of open-source development by utilizing platforms like  
117 GitHub and Git, unit testing (as of August 2023 coverage is 83%), continuous integration,  
118 automatic documentation, and Findable, Accessible, Interoperable, and Reusable (FAIR)  
119 principles. Detailed [documentation](#) and [tutorials](#) for getting started with the package are  
120 publicly available. The project aims to create high-quality software that can be easily accessed,  
121 used, and contributed to by a wide range of researchers.

122 We believe this project will have a positive impact across the all of structural bioinformatics,  
123 enabling advancements that rely on molecular complex analysis, such as structural biology,  
124 protein engineering, and rational drug design. The target community includes researchers  
125 working with molecular complexes data, such as computational biologists, immunologists,  
126 and structural bioinformaticians. The existing features, as well as the sustainable package  
127 formatting and its modular design make DeepRank2 an excellent framework to build upon.  
128 Taken together, DeepRank2 provides all the requirements to become the all-purpose DL tool  
129 that is currently lacking in the field of biomolecular interactions.

## Acknowledgements

This work was supported by the [Netherlands eScience Center](#) under grant number NLESC.OEC.2021.008, and [SURF](#) infrastructure, and was developed in collaboration with the [Department of Medical BioSciences](#) at RadboudUMC (Hypatia Fellowship, Rv819.52706). This work was also supported from NVIDIA Academic Award.

## References

- Baek, M., DiMaio, F., Anishchenko, I., Dauparas, J., Ovchinnikov, S., Lee, G. R., Wang, J., Cong, Q., Kinch, L. N., Schaeffer, R. D., Millán, C., Park, H., Adams, C., Glassman, C. R., DeGiovanni, A., Pereira, J. H., Rodrigues, A. V., Dijk, A. A. van, Ebrecht, A. C., ... Baker, D. (2021). *Accurate prediction of protein structures and interactions using a three-track neural network*.
- Dominguez, C., Boelens, R., & Bonvin, A. M. J. J. (2003). HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *Journal of the American Chemical Society*, 125(7), 1731–1737. <https://doi.org/10.1021/ja026939x>
- Eismann, S., Townshend, R. J. L., Thomas, N., Jagota, M., Jing, B., & Dror, R. O. (2021). Hierarchical, rotation-equivariant neural networks to select structural models of protein complexes. *Proteins: Structure, Function, and Bioinformatics*, 89(5), 493–501. <https://doi.org/10.1002/prot.26033>
- Fout, A., Byrd, J., Shariat, B., & Ben-Hur, A. (2017). Protein interface prediction using graph convolutional networks. *Advances in Neural Information Processing Systems*, 30.
- Gainza, P., Wehrle, S., Van Hall-Beauvais, A., Marchand, A., Scheck, A., Hartevelde, Z., Buckley, S., Ni, D., Tan, S., Sverrisson, F., & others. (2023). De novo design of protein interactions with learned surface fingerprints. *Nature*, 1–9.
- Gane, P. J., & Dean, P. M. (2000). Recent advances in structure-based rational drug design. *Current Opinion in Structural Biology*, 10(4), 401–404. [https://doi.org/10.1016/S0959-440X\(00\)00105-6](https://doi.org/10.1016/S0959-440X(00)00105-6)
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., ... Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873), 583–589. <https://doi.org/10.1038/s41586-021-03819-2>
- Li, B., Yang, Y. T., Capra, J. A., & Gerstein, M. B. (2020). Predicting changes in protein thermodynamic stability upon point mutation with deep 3D convolutional neural networks. *PLoS Comput Biol*. <https://doi.org/10.1371/journal.pcbi.1008291>
- Li, Y., Rezaei, M. A., Li, C., Li, X., & Wu, D. (2019). DeepAtom: A framework for protein-ligand binding affinity prediction. *2019 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, 303–310. <https://doi.org/10.1109/BIBM47256.2019.8982964>
- Liu, S., Liu, S., Zhu, X., Liang, H., Cao, A., Chang, Z., & Lai, L. (2007). Nonnatural protein-protein interaction-pair design by key residues grafting. *Proceedings of the National Academy of Sciences*, 104(13), 5330–5335. <https://doi.org/10.1073/pnas.0606198104>
- Ramakrishnan, G., Baakman, C., Heijl, S., Vroling, B., Horck, R. van, Hiraki, J., Xue, L. C., & Huynen, M. A. (2023). Understanding structure-guided variant effect predictions using 3D convolutional neural networks. *Frontiers in Molecular Biosciences*, 10.
- Ramsundar, B., Eastman, P., Walters, P., Pande, V., Leswing, K., & Wu, Z. (2019). *Deep learning for the life sciences*. O'Reilly Media.

- 175 Réau, M., Renaud, N., Xue, L. C., & Bonvin, A. M. J. J. (2022). DeepRank-GNN: A graph  
176 neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*,  
177 39(1). <https://doi.org/10.1093/bioinformatics/btac759>
- 178 Renaud, N., Geng, C., Georgievskaya, S., Ambrosetti, F., Ridder, L., Marzella, D. F., Réau, M.  
179 F., Bonvin, A. M. J. J., & Xue, L. C. (2021). DeepRank: A deep learning framework  
180 for data mining 3D protein–protein interfaces. *Nature Communications*, 12(1), 7068.  
181 <https://doi.org/10.1038/s41467-021-27396-0>
- 182 Richard Evans, A. P., Michael O'Neill. (2021). Protein complex prediction with AlphaFold-  
183 multimer. *bioRxiv*. <https://doi.org/10.1101/2021.10.04.463034>
- 184 Sadelain, M., Brentjens, R., & Rivière, I. (2013). The basic principles of chimeric antigen  
185 receptor design. *Cancer Discovery*, 3(4), 388–398. [https://doi.org/10.1158/2159-8290.](https://doi.org/10.1158/2159-8290.CD-12-0548)  
186 [CD-12-0548](https://doi.org/10.1158/2159-8290.CD-12-0548)
- 187 Sanchez, R., & Sali, A. (1997). Evaluation of comparative protein structure modeling by  
188 MODELLER-3, proteins suppl. 1, 50– 58. *Google Scholar There Is No Corresponding*  
189 *Record for This Reference.*
- 190 Schwede, T. (2013). Protein modeling: What happened to the “protein structure gap”?  
191 *Structure*, 21(9), 1531–1540. <https://doi.org/https://doi.org/10.1016/j.str.2013.08.007>
- 192 Shroff, R., Cole, A. W., Diaz, D. J., Morrow, B. R., Donnell, I., Annapareddy, A., Gollihar,  
193 J., Ellington, A. D., & Thyer, R. (2020). Discovery of novel gain-of-function mutations  
194 guided by structure-based deep learning. *ACS Synthetic Biology*, 9(11), 2927–2935.  
195 <https://doi.org/10.1021/acssynbio.0c00345>
- 196 Vicentini, F., Hofmann, D., Szabó, A., Wu, D., Roth, C., Giuliani, C., Pescia, G., Nys,  
197 J., Vargas-Calderón, V., Astrakhantsev, N., & Carleo, G. (2022). NetKet 3: Machine  
198 learning toolbox for many-body quantum systems. *SciPost Physics Codebases*. <https://doi.org/10.21468/scipostphyscodeb.7>  
199 <https://doi.org/10.21468/scipostphyscodeb.7>
- 200 Wang, X., Flannery, S. T., & Kihara, D. (2021). Protein docking model evaluation by graph  
201 neural networks. *Frontiers in Molecular Biosciences*, 8. [https://doi.org/10.3389/fmolb.](https://doi.org/10.3389/fmolb.2021.647915)  
202 [2021.647915](https://doi.org/10.3389/fmolb.2021.647915)
- 203 Wang, X., Terashi, G., Christoffer, C. W., Zhu, M., & Kihara, D. (2020). Protein docking model  
204 evaluation by 3D deep convolutional neural networks. *Bioinformatics*, 36(7), 2113–2118.
- 205 Zhu, Z., Shi, C., Zhang, Z., Liu, S., Xu, M., Yuan, X., Zhang, Y., Chen, J., Cai, H., Lu, J.,  
206 Ma, C., Liu, R., Xhonneux, L.-P., Qu, M., & Tang, J. (2022). TorchDrug: A powerful and  
207 flexible machine learning platform for drug discovery. *arXiv Preprint arXiv:2202.08320*.