DeepRank2: Mining 3D Protein Structures with Geometric Deep Learning

Giulia Crocioni [1][*][¶], **Dani L. Bodor** [1][*], **Coos Baakman** [2][*], **Daniel Rademaker** [2], **Dario Marzella** [2], **Gayatri Ramakrishnan** [2], **Sven van der Burg** [1], **Farzaneh Meimandi Parizi** [2], and **Li C. Xue** [2]

1 Netherlands eScience Center, Amsterdam, The Netherlands **2** Radboud University Medical Center, Nijmegen, The Netherlands ¶ Corresponding author * These authors contributed equally.

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) and 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.
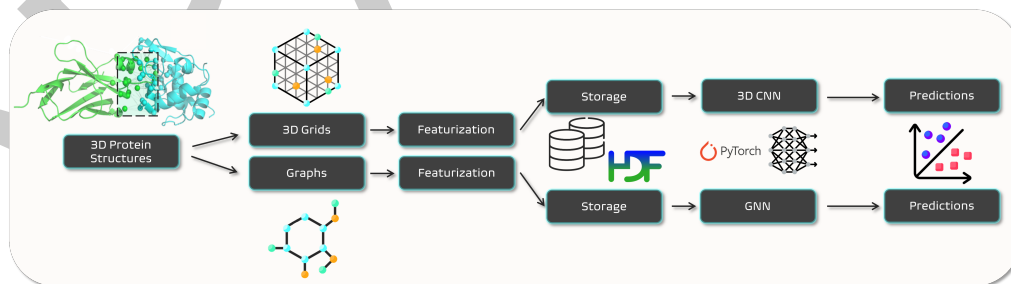
Figure 1: DeepRank2 framework overview. 3D coordinates of protein structures are extracted from PDB files and converted into graphs and grids, using either an atomic or a residual level, depending on the user's requirements. The data are enriched with geometrical and physicochemical information and are stored into HDF5 files, and can then 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

²⁶ presented on the major histocompatibility complex (MHC) protein, which play a key role in
²⁷ T-cell immunity. Protein structures can also be exploited in molecular diagnostics for the
²⁸ identification of SRVs, that can be pathogenic sequence alterations in patients with inherited
²⁹ diseases (Li B, 2020; Shroff et al., 2020).

³⁰ In the past decades, a variety of experimental methods (e.g., X-ray crystallography, nuclear
³¹ magnetic resonance, cryogenic electron microscopy) have determined and accumulated a large
³² number of atomic-resolution 3D structures of proteins and protein-protein complexes (Schwede,
³³ 2013). Because experimental determination of structures is a tedious and expensive process,
³⁴ several computational prediction methods have also been developed over the past few years,
³⁵ such as Alphafold (Jumper et al., 2021) for single proteins, and PANDORA (Marzella et al.,
³⁶ 2022), HADDOCK (Dominguez et al., 2003), and Alphafold-Multimer (Richard Evans, 2021)
³⁷ for protein complexes. The large amount of data available makes it possible to use DL to
³⁸ leverage 3D structures and learn their complex patterns. Unlike other machine learning (ML)
³⁹ techniques, deep neural networks hold the promise of learning from millions of data without
⁴⁰ reaching a performance plateau quickly, which is made computationally feasible by hardware
⁴¹ accelerators (i.e., GPUs, TPUs) and parallel file system technologies.

⁴² 3D CNNs have been trained on 3D grids for the classification of biological vs. crystallographic
⁴³ PPIs (Renaud et al., 2021), and for the scoring of models of protein-protein complexes generated
⁴⁴ by computational docking (Renaud et al., 2021; Wang et al., n.d.). Gaiza et al. have applied
⁴⁵ geodesic CNNs to extract protein interaction fingerprints by applying 2D CNNs on spread-out
⁴⁶ protein surface patches (Gainza et al., 2020). 3D CNNs have been used for exploiting protein
⁴⁷ structure data for predicting mutation-induced changes in protein stability (Li B, 2020) and
⁴⁸ identifying novel gain-of-function mutations (Shroff et al., 2020). Contrary to CNNs, in GNNs
⁴⁹ the convolution operations on graphs can rely on the relative local connectivity between nodes
⁵⁰ and not on the data orientation, making graphs rotational invariant. Additionally, GNNs can
⁵¹ accept any size of graph, while in a CNN the size of the 3D grid for all input data needs to be
⁵² the same, which may be problematic for datasets containing highly variable in size structures.
⁵³ Based on these arguments, different GNN-based tools have been designed to predict patterns
⁵⁴ from PPIs (Fout et al., n.d.; Réau et al., 2022; Wang et al., 2021). Eisman et al. developed a
⁵⁵ rotation-equivariant neural network trained on point-based representation of the protein atomic
⁵⁶ structure to classify PPIs (Eismann et al., 2021).

⁵⁷ Statement of need

⁵⁸ Data mining 3D structures of proteins presents several challenges. These include complex
⁵⁹ physico-chemical rules governing structural features, the possibility of characterizartion at
⁶⁰ different scales (e.g., atom-level, residue level, and secondary structure level), and the large
⁶¹ diversity in shape and size. Furthermore, because a structures can easily comprise of hundreds
⁶² to thousands of residues (and ~15 times as many atoms), efficient processing and featurization
⁶³ of many structures is critical to handle the computational cost and file storage requirements.
⁶⁴ Existing software solutions are often highly specialized and not developed as reusable and
⁶⁵ flexible frameworks, and cannot be easily adapted to diverse applications and predictive tasks.
⁶⁶ Examples include DeepAtom (Li et al., 2019), for protein-ligand binding affinity prediction only,
⁶⁷ MaSIF (Gainza et al., 2020), for deciphering patterns in protein surfaces, and MHCFlurry 2.0
⁶⁸ (O'Donnell et al., 2020), for predicting binding affinity for a specific type of protein-protein
⁶⁹ complex, the peptide-major histocompatibility complex (MHC). While some frameworks, such
⁷⁰ as TorchProtein and TorchDrug (Zhu et al., 2022), configure themselves as general-purpose
⁷¹ ML libraries for both molecular sequences and 3D structures, they only implement geometric-
⁷² related features and do not incorporate fundamental physico-chemical information in the 3D
⁷³ representation of molecules.

⁷⁴ These limitations create a growing demand for a generic and flexible DL framework that
⁷⁵ researchers can readily utilize for their specific research questions while cutting down the
⁷⁶ tedious data preprocessing stages. Generic DL frameworks have already emerged in diverse

77 scientific fields, such as computational chemistry (e.g., DeepChem (Ramsundar et al., 2019))
78 and condensed matter physics (e.g., NetKet (Vicentini et al., 2022)), which have promoted
79 collaborative efforts, facilitated novel insights, and benefited from continuous improvements
80 and maintenance by engaged user communities.

Key features

82 DeepRank2 allows to transform and store 3D representations of both PPIs and SRVs into 3D
83 grids or graphs containing both geometric and physico-chemical information, and provides
84 a DL pipeline which can be used for training pre-implemented neural networks for a given
85 pattern of interest to the user.

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

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

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

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

120 This project is expected to have an impact across the all of structural bioinformatics, enabling
121 advancements that rely on molecular complex analysis, such as structural biology, protein
122 engineering, and rational drug design. The target community includes researchers working with
123 molecular complexes data, such as computational biologists, immunologists, and structural
124 bioinformatics scientists. The existing features, as well as the sustainable package formatting
125 and its modular design make DeepRank2 an excellent framework to build upon. Taken together,

DeepRank2 provides all the requirements to become the all-purpose DL tool that is currently lacking in the field of biomolecular interactions.

Acknowledgements

References

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/https://doi.org/10.1002/prot.26033

Fout, A., Byrd, J., Shariat, B., & Ben-Hur, A. (n.d.). *Protein interface prediction using graph convolutional networks*.

Gainza, P., Sverrisson, F., Monti, F., Rodolà, E., Boscaini, D., Bronstein, M., & Correia, B. (2020). Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nature Methods*, *17*(2), 184–192.

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/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., Žídek, 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, C. J., Yang YT. (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. (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

Marzella, D. F., Parizi, F. M., Tilborg, D. van, Renaud, N., Sybrandi, D., Buzatu, R., Rademaker, D. T., Hoen, P. A. C. 't, & Xue, L. C. (2022). PANDORA: A fast, anchor-restrained modelling protocol for peptide: MHC complexes. *Frontiers in Immunology*, *13*. https://doi.org/10.3389/fimmu.2022.878762

O'Donnell, T., Rubinsteyn, A., & Laserson, U. (2020). MHCflurry 2.0: Improved pan-allele prediction of MHC class i-presented peptides by incorporating antigen processing, cell syst. 11 (2020) 42-48. e7. *P42-P48. E7*.

169 Ramsundar, B., Eastman, P., Walters, P., Pande, V., Leswing, K., & Wu, Z. (2019). *Deep*
170   *learning for the life sciences*. O'Reilly Media.

171 Réau, M., Renaud, N., Xue, L. C., & Bonvin, A. M. J. J. (2022). DeepRank-GNN: A graph
172   neural network framework to learn patterns in protein–protein interfaces. *Bioinformatics*,
173   *39*(1). https://doi.org/10.1093/bioinformatics/btac759

174 Renaud, N., Geng, C., Georgievska, S., Ambrosetti, F., Ridder, L., Marzella, D. F., Réau, M.
175   F., Bonvin, A. M. J. J., & Xue, L. C. (2021). DeepRank: A deep learning framework
176   for data mining 3D protein-protein interfaces. *Nature Communications*, *12*(1), 7068.
177   https://doi.org/10.1038/s41467-021-27396-0

178 Richard Evans, A. P., Michael O'Neill. (2021). Protein complex prediction with AlphaFold-
179   multimer. *bioRxiv*. https://doi.org/10.1101/2021.10.04.463034

180 Sadelain, M., Brentjens, R., & Rivière, I. (2013). The basic principles of chimeric antigen
181   receptor design. *Cancer Discovery*, *3*(4), 388–398. https://doi.org/10.1158/2159-8290.
182   CD-12-0548

183 Schwede, T. (2013). Protein modeling: What happened to the "protein structure gap"?
184   *Structure*, *21*(9), 1531–1540. https://doi.org/https://doi.org/10.1016/j.str.2013.08.007

185 Shroff, R., Cole, A. W., Diaz, D. J., Morrow, B. R., Donnell, I., Annapareddy, A., Gollihar,
186   J., Ellington, A. D., & Thyer, R. (2020). Discovery of novel gain-of-function mutations
187   guided by structure-based deep learning. *ACS Synthetic Biology*, *9*(11), 2927–2935.
188   https://doi.org/10.1021/acssynbio.0c00345

189 Vicentini, F., Hofmann, D., Szabó, A., Wu, D., Roth, C., Giuliani, C., Pescia, G., Nys,
190   J., Vargas-Calderón, V., Astrakhantsev, N., & Carleo, G. (2022). NetKet 3: Machine
191   learning toolbox for many-body quantum systems. *SciPost Physics Codebases*. https:
192   //doi.org/10.21468/scipostphyscodeb.7

193 Wang, X., Flannery, S. T., & Kihara, D. (2021). Protein docking model evaluation by graph
194   neural networks. *Frontiers in Molecular Biosciences*, *8*. https://doi.org/10.3389/fmolb.
195   2021.647915

196 Wang, X., Terashi, G., Christoffer, C. W., Zhu, M., & Kihara, D. (n.d.). Protein docking
197   model evaluation by 3D deep convolutional neural networks. *Bioinformatics*.

198 Zhu, Z., Shi, C., Zhang, Z., Liu, S., Xu, M., Yuan, X., Zhang, Y., Chen, J., Cai, H., Lu, J.,
199   Ma, C., Liu, R., Xhonneux, L.-P., Qu, M., & Tang, J. (2022). TorchDrug: A powerful and
200   flexible machine learning platform for drug discovery. *arXiv Preprint arXiv:2202.08320*.