The Journal of Open Source Software

DeepRank2: Mining 3D Protein Structures with Geometric Deep Learning

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

1 Netherlands eScience Center, Amsterdam, The Netherlands **2** Radboud University Medical Center, Nijmegen, The Netherlands **3** Independent Researcher ¶ Corresponding author * These authors contributed equally.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

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.

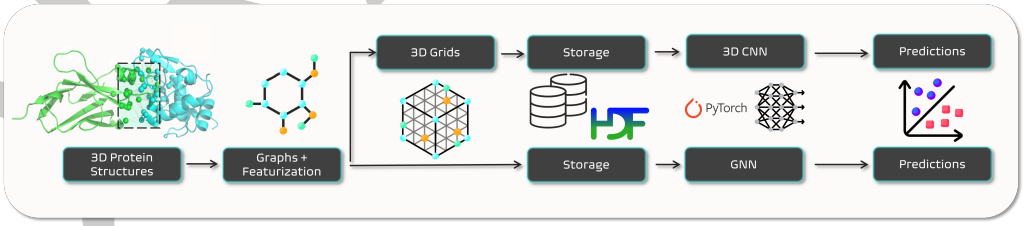

Figure 1: DeepRank2 framework overview. 3D coordinates of protein structures are extracted from PDB files and converted into graphs, using either an atomic or a residue level, depending on the user's requirements. Then, the data are enriched with geometrical and physicochemical information, and eventually 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

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

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

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

Statement of need

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

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

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

Key features

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

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

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

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

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

We believe this project will have a positive impact across the all of structural bioinformatics, enabling advancements that rely on molecular complex analysis, such as structural biology, protein engineering, and rational drug design. The target community includes researchers working with molecular complexes data, such as computational biologists, immunologists, and structural bioinformaticians. The existing features, as well as the sustainable package formatting 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.

The Journal of Open Source Software

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 Acamedic 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/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., Harteveld, 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/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., 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.

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

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

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

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

Sanchez, R., & Sali, A. (1997). Evaluation of comparative protein structure modeling by MODELLER-3, proteins suppl. 1, 50- 58. *Google Scholar There Is No Corresponding Record for This Reference*.

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

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

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

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

Wang, X., Terashi, G., Christoffer, C. W., Zhu, M., & Kihara, D. (2020). Protein docking model evaluation by 3D deep convolutional neural networks. *Bioinformatics*, *36*(7), 2113–2118.

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