The Journal of Open Source Software

DeepRank-Core: Mining 3D Protein Structures with Geometric Deep Learning

Giulia Crocioni [1*¶], **Dani 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.

DOI: 10.xxxxxx/draft

Software

- Review
- Repository
- Archive

Editor: Open Journals

Reviewers:

- @openjournals

Submitted: 01 January 1970
Published: unpublished

Summary

We present DeepRank-Core, an open-source deep learning (DL) framework that offers researchers unified and user-friendly APIs to accelerate development of software solutions allowing biologically relevant predictions to gain knowledge on protein 3D structures for a wide variety of purposes, such as drug design, immunotherapy, or designing novel proteins. DeepRank-Core allows to transform and store 3D representations of both protein-protein interfaces (PPIs) and individual proteins' variants into grids or graphs containing structural and physico-chemical information, which can then be used for training neural networks for whatever specific pattern of interest for the user. DeepRank-Core also offers a pre-implemented training pipeline which can use either convolutional neural networks (CNNs) or graph neural networks (GNNs), as well as handy output exporters for evaluating performances. The entire framework flowchart is visualized in Figure 1. DeepRank-Core software aims at unifying previously developed DL frameworks for data mining PPIs (DeepRank (Renaud et al., 2021), DeepRank-GNN (Réau et al., 2022)), and proteins' variants (DeepRank-Mut [@]). The package follows the community-endorsed FAIR principles for Research Software, provides user-friendly APIs, publicly available documentation and in depth tutorials. Additionally, the software allows for much greater flexibility, allowing users to easily tailor the framework to specific patterns of interest and features, and select the pipeline's steps that best suits their requirements.

State of the field

Individual proteins' and protein complexes' 3D structures provide fundamental information to decipher biological processes at the molecular scale. Gaining knowledge on how those biomolecules interact in 3D space is key for understanding their functions and exploiting or engineering these molecules for a wide variety of purposes 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 missense variants, that are pathogenic sequence alterations in patients with inherited diseases.

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. Since the experimental structure determination is a tedious and expensive process, several computational methods have also been developed over the past few years, such as Alphafold for protein

structures, and PANDORA (Marzella et al., 2022), HADDOCK (Dominguez et al., 2003), and Alphafold-Multimer (Richard Evans, 2021) for protein complexes. 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 without reaching a performance plateau quickly, which is made computationally feasible by hardware accelerators (i.e., GPUs, TPUs) and parallel file system technologies.

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 Wang et al. (n.d.). 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., 2020). 3D CNNs have been used for exploiting protein structure data for predicting mutation-induced changes in protein stability (Li B, 2020) 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, making graphs rotational invariant, and such networks can accept any size of graph, making them more representative of the PPIs diversity. Based on these arguments, different GNN-based tools have been designed to predict patterns from PPIs Réau et al. (2022). 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 proteins structures still presents several unique challenges because of the different physico-chemical rules that govern them, the possibility of characterizartion at different levels (atom-atom level, residue-residue level, and secondary structure level), and the highly diversity in terms of shapes and sizes. The efficient processing and featurization of a large number of atomic coordinates files of proteins is also critical in terms of computational cost and file storage requirements. Existing software solutions are often highly specialized, developed for analysis and visualization of specific research projects' results, and cannot be easily adapted to diverse applications and predictive tasks, not being developed as reusable and flexible frameworks. Examples include DeepAtom (Li et al., 2019), for protein-ligand binding affinity prediction only, MaSIF (Gainza et al., 2020), for deciphering patterns in protein surfaces, and MHCFlurry 2.0 (O'Donnell et al., 2020), for predicting binding affinity for a specific type of protein-protein complex, the peptide-major histocompatibility complex (MHC). Other frameworks instead, such as TorchProtein and TorchDrug (Zhu et al., 2022), configure themselves as general-purpose ML libraries for both molecular sequences and 3D structures. However, they only implement geometric-related features and do not incorporate fundamental physico-chemical information in the molecules' 3D representations, which may be crucial for accurate predictions.

These limitations create a growing demand for generic and flexible DL frameworks 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

DeepRank-Core allows to transform and store 3D representations of both PPIs and individual proteins' variants into 3D grids or graphs containing geometric and physico-chemical information,

and provides a DL pipeline which can be used for training pre-implemented neural networks for whatever specific pattern of interest for the user.

The 3D protein structures provided in the form of PDB files are mapped to graphs in which nodes represent residues or atoms, according to the resolution chosen by the user, and edges 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; - missense variants, for mining pathologic mutations within protein structures.

The graphs can also be mapped to volumetric grids (i.e., 3D image-like representations). Then the physico-chemical and geometrical features for the grids and/or graphs are computed. The user can choose which features to generate from several ones already defined in the package, and can also define custom features modules, as explained in the documentation. Examples of pre-defined node features are the type of the amino acid, its polarity, the solvent-accessible surface area. Examples of pre-defined edge features are distance, covalent bond, electrostatic potential. The full and detailed list of features can be found in the documentation's features page. Multiple CPUs can be used to parallelize and speed up the featurization process. The mapped data are finally saved into HDF5 files, designed to efficiently store and organize big data. Users can then use the data saved into HDF5 files for whatever architecture and DL framework is more suited for the application. In particular, graphs can be used for the training of GNNs, and 3D grids can be used for the training of CNNs.

DeepRank-Core provides also handy modules for training Pytorch neural networks with the data stored into the HDF5 files. A few simple GNNs and CNNs are pre-implemented within the package, and can be trained on the generated data. The neural networks have been developed using PyTorch, and the user is also free to implement custom networks using the same PyTorch framework. The 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 the selection of specific features across the ones generated, the definition of the targets and the predictive tasks. Then the datasets can be used for training, validating, and testing the chosen neural network, and the results together with the trained model can be saved using the DeepRank-Core exporters' module.

The package 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 by a wide range of researchers.

The project is expected to have an impact across structural bioinformatics domains, enabling advancements in the disciplines 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 bioinformatics scientists. The existing features, as well as the sustainable package formatting and its great modularity make DeepRank-Core an excellent framework to build upon, to generate the all-purpose deep learning tool that is currently lacking in the field of biomolecular interactions.

The Journal of Open Source Software

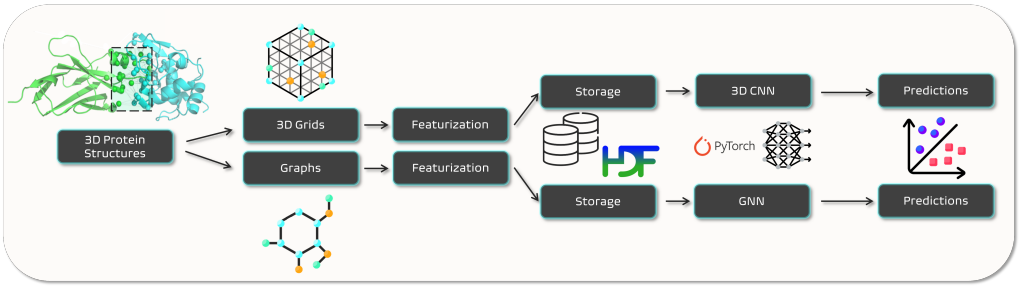

Figure 1: DeepRank-Core framework overview. 3D coordinates of protein structures are extracted from PDB files and converted into graphs and grids, using either an atomic or a residual level, depending on the user's requirements. The data are enriched with geometrical and physicochemical information and are stored into HDF5 files, and can then be used in the pre-implemented DL pipeline for training PyTorch networks and computing predictions.

Acknowledgements

This work was supported by the Netherlands eScience Center, Amsterdam, and SURF infrastructure, and was developed in collaboration with the Department of Medical BioSciences at RadboudUMC, Nijmegen.

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

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.*

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

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. (n.d.). Protein docking model evaluation by 3D deep convolutional neural networks. *Bioinformatics*.

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*.