ProInterVal-BioXtal: A Web Server to Distinguish Biological from Crystallographic Interfaces Using Representation Learning and Graph Neural Networks

Defne Alnıgeniş[1], Damla Övek Baydar [2,3], Attila Gürsoy [3], and Özlem Keskin[1]

1 Chemical and Biological Engineering, Koç University, Istanbul, Turkey
2 Norwegian Centre for Molecular Biosciences and Medicine, University of Oslo
3 Computer Engineering, Koç University, Istanbul, Turkey

1- INTRODUCTION

- Proteins are **the fundamental building blocks** of cells as they regulate several biological processes through sophisticated interactions.
- They physically interact through the specific regions called **Protein-Protein Interfaces (PPIs)** to form functional complexes.
- Understanding the characteristics of these interactions is essential for **advancing drug development**[1,2].

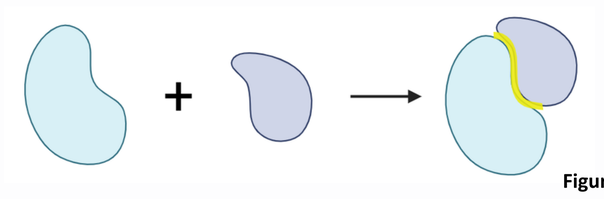


Figure 1. PPIs

- **X-ray crystallography** is used to retrieve 3D protein structures which resulted in 80.2% of all structures in the Protein Data Bank (PDB)[3].
- These structures are mostly reliable but some problems exist causing artifacts and misleading conclusions[2].

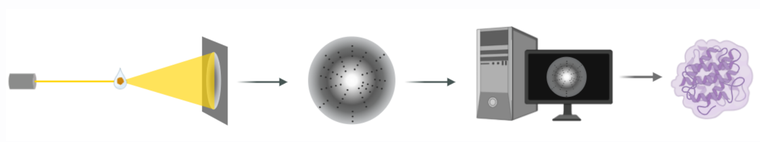
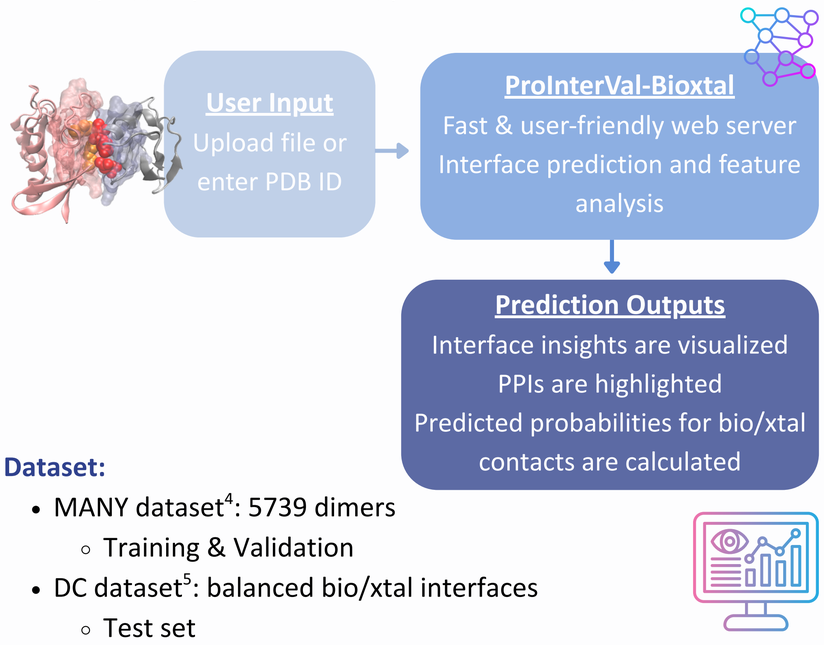


Figure 2. X-ray crystallography

2- AIM

- The purpose of this study is to present fast and user-friendly web server, **ProInterVal-Bioxtal**, which provides enhanced insights into protein complexes.
- The server implements the novel **graph-based deep learning** methodology based on **learned representations.**
- Consequently, **discrimination of biological interfaces from crystallographic artifacts** will be possible.
- The ProInterVal-BioXtal server is freely **available** at https://3dpath.ku.edu.tr/prointerval-bioxtal/.



Dataset:

- MANY dataset[4]: 5739 dimers
  - Training & Validation
- DC dataset[5]: balanced bio/xtal interfaces
  - Test set
- Benchmark dataset[7]: 1677 homodimers
  - Test set

5- DISCUSSION

- ProInterVal-Bioxtal addresses a fundamental challenge in structural biology by utilizing graph-based protein representation which **can detect complex interactions and dependencies, with a promising accuracy level**.
- Sequential and spatial relationships between the residues on the interface are determined.
- Essential features of the PPIs are captured.
- Overcomes the limitations of handcrafted calculations.

3- METHODOLOGY

ProInterVal- Bioxtal Web Server Framework

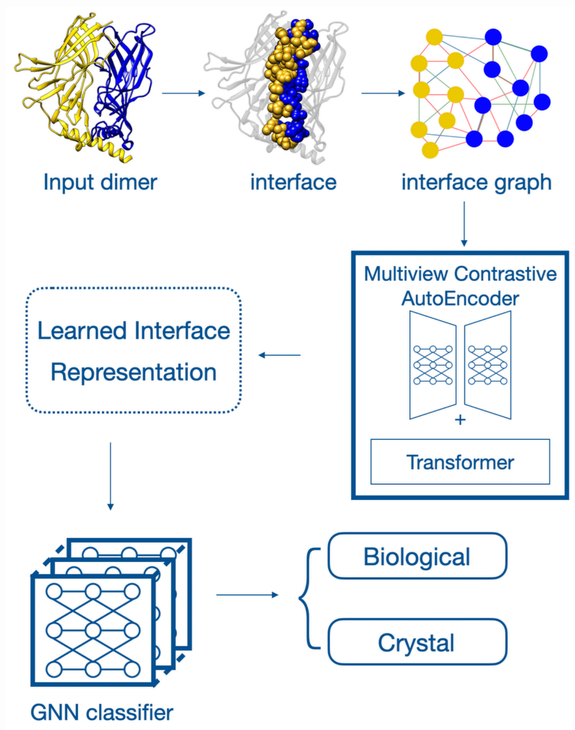


Figure 3. Input dimer complex undergoes the explained graph-based deep learning methodology. The contacting residues on the interface are represented as a graph where nodes represent amino acids and edges indicate their interaction.

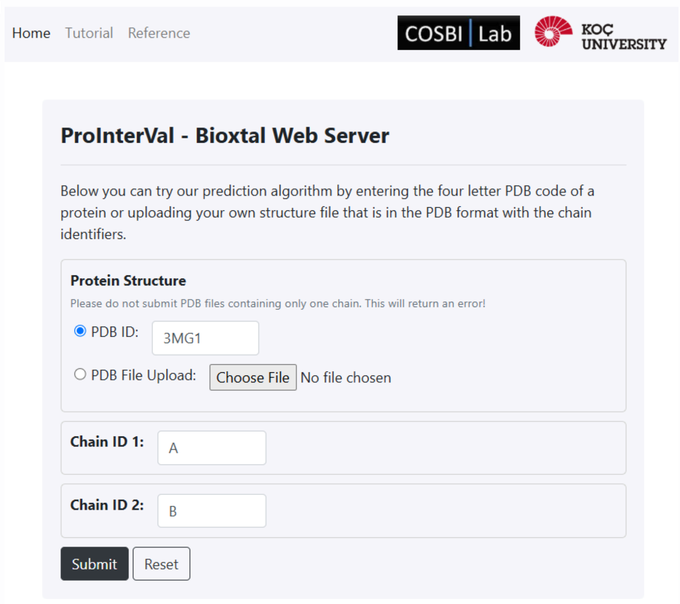
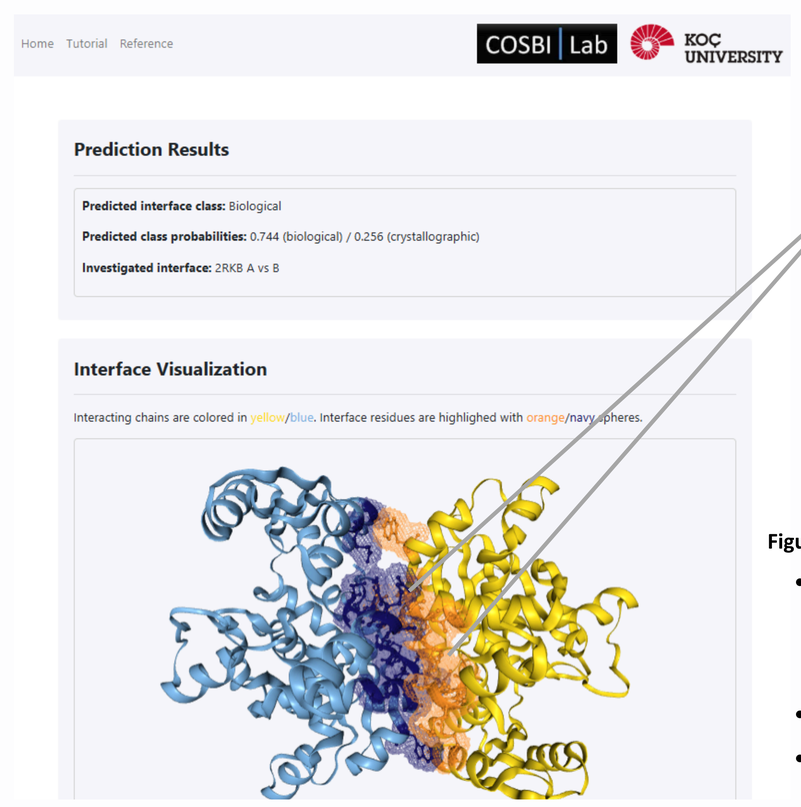


Figure 4. The server accepts a PDB file or PDB ID to fetch the structure as an input. If it is a multimer, the user should specify two chain identifiers of interest.



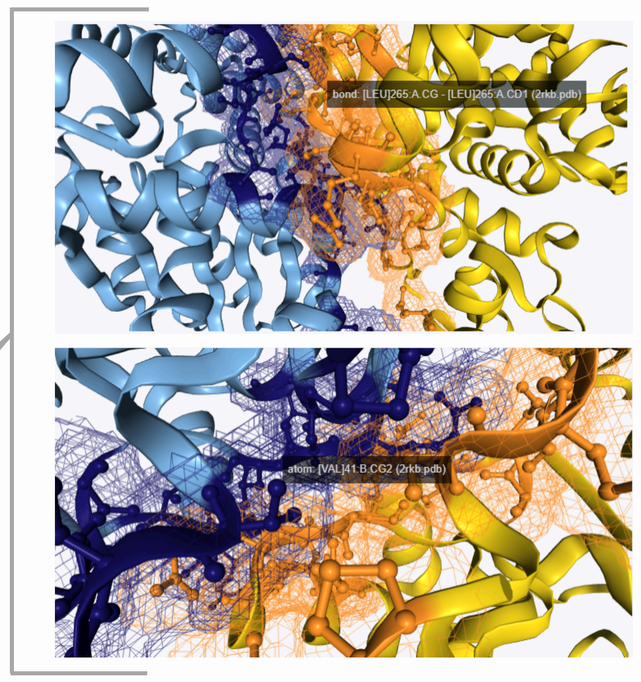
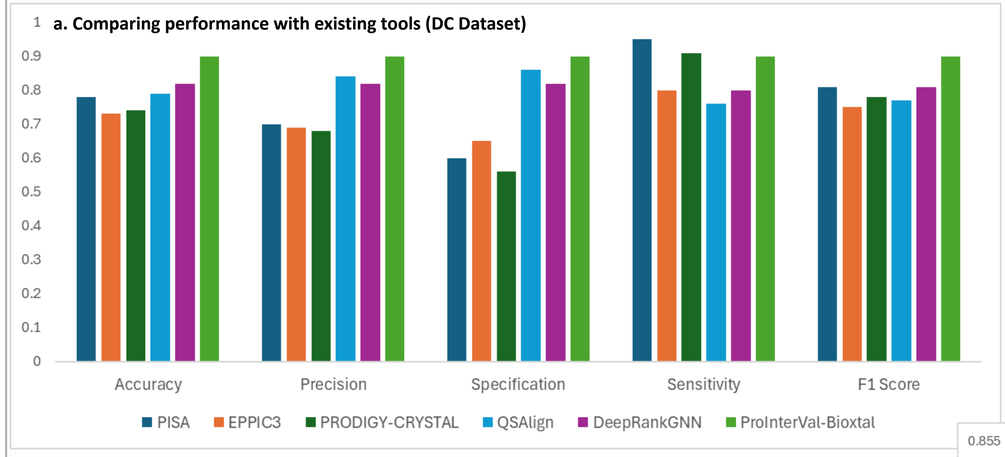


Figure 5. Results page.

- Two predicted probabilities are displayed both for the biological and crystal contacts; the higher probability determines the label.
- An informative and fast tool.
- Interface residues highlighted using distinct sphere colors.

4- RESULTS



- The server outperforms existing classifiers tested on the DC dataset[5].
- 90% accuracy, 90% precision and 90% F1 score on the test set.
- 83% accuracy on benchmark dataset[7] as the secondary evaluation.

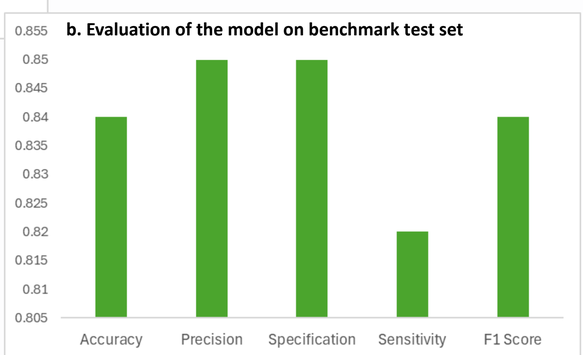


Figure 6. Evaluation of ProInterVal-Bioxtal on the test set. a) The performance of ProInterVal-Bioxtal on the test set, DC[5], compared with 5 state-of-the-art existing tools. b) The method is evaluated on benchmark[7] dataset.

REFERENCES

1. Keskin, O., Tuncbag, N., & Gursoy, A. (2016). Predicting Protein–Protein Interactions from the Molecular to the Proteome Level. *Chemical Reviews, 116*(8), 4884–4909. https://doi.org/10.1021/acs.chemrev.5b00683
2. Ovek, D., Keskin, O., & Gursoy, A. (2024). ProInterVal: Validation of Protein–Protein Interfaces through Learned Interface Representations. *Journal of Chemical Information and Modeling, 64*(8), 2979–2987. https://doi.org/10.1021/acs.jcim.3c01788
3. RCSB Protein Data Bank. PDB Data Distribution by Experimental Method and Molecular Type. RCSB PDB Statistics. Accessed 17 June 2026
4. Baskaran, K., Duarte, J. M., Biyani, N., Bliven, S., & Capitani, G. (2014). A PDB-wide, evolution-based assessment of protein-protein interfaces. *BMC Structural Biology, 14*(1). https://doi.org/10.1186/s12900-014-0022-0
5. Duarte, J. M., Srebniak, A., Schärer, M. A., & Capitani, G. (2012). Protein interface classification by evolutionary analysis. *BMC Bioinformatics, 13*(1). https://doi.org/10.1186/1471-2105-13-334
6. Zhang, Z., Xu, M., Jamasb, A., Chenthamarakshan, V., Lozano, A., Das, P., & Tang, J. (2022). Protein Representation learning by Geometric Structure pretraining. *arXiv (Cornell University).* https://doi.org/10.48550/arxiv.2203.06125
7. Schweke, H., Xu, Q., Tauriello, G., Pantolini, L., Schwede, T., Cazals, F., Lhéritier, A., Fernandez-Recio, J., Rodríguez-Lumbreras, L. A., Schueler-Furman, O., Varga, J. K., Jiménez-García, B., Réau, M. F., Bonvin, A. M. J. J., Savojardo, C., Martelli, P., Casadio, R., Tubiana, J., Wolfson, H. J., . . . Wodak, S. J. (2023). Discriminating physiological from non-physiological interfaces in structures of protein complexes: A community-wide study. *PROTEOMICS, 23*(17). https://doi.org/10.1002/pmic.202200323