



BIPSPI+: Mining Type-Specific Datasets of Protein Complexes to Improve Protein Binding Site Prediction

R. Sanchez-Garcia^{1,2*}, J. R. Macias¹, C. O. S. Sorzano¹, J. M. Carazo^{1*} and J. Segura³

1 - Biocomputing Unit, National Center for Biotechnology (CSIC), Darwin 3, Campus Univ. Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain

2 - Oxford Protein Informatics Group, Department of Statistics, University of Oxford, 29 St Giles' Oxford OX1 3LB, UK

3 - Research Collaboratory for Structural Bioinformatics, Protein Data Bank, San Diego Supercomputer Center, University of California San Diego, La Jolla, CA 92093, USA

Correspondence to R. Sanchez-Garcia and J.M. Carazo: Oxford Protein Informatics Group, Department of Statistics, University of Oxford, Oxford OX1 3LB, UK, Biocomputing Unit, National Center for Biotechnology (CSIC), Darwin 3, Campus Univ. Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain. ruben.sanchez-garcia@stats.ox.ac.uk (R. Sanchez-Garcia), carazo@cnb.csic.es (J.M. Carazo) @cossStock (C.O.S. Sorzano), @JM_Carazo (J.M. Carazo)

<https://doi.org/10.1016/j.jmb.2022.167556>

Edited by Michael Sternberg

Abstract

Computational approaches for predicting protein-protein interfaces are extremely useful for understanding and modelling the quaternary structure of protein assemblies. In particular, partner-specific binding site prediction methods allow delineating the specific residues that compose the interface of protein complexes. In recent years, new machine learning and other algorithmic approaches have been proposed to solve this problem. However, little effort has been made in finding better training datasets to improve the performance of these methods. With the aim of vindicating the importance of the training set compilation procedure, in this work we present BIPSPI+, a new version of our original server trained on carefully curated datasets that outperforms our original predictor. We show how prediction performance can be improved by selecting specific datasets that better describe particular types of protein interactions and interfaces (e.g. homo/hetero). In addition, our upgraded web server offers a new set of functionalities such as the sequence-structure prediction mode, hetero- or homo-complex specialization and the guided docking tool that allows to compute 3D quaternary structure poses using the predicted interfaces. BIPSPI+ is freely available at <https://bipspi.cnb.csic.es>.

© 2022 The Authors. Published by Elsevier Ltd. This is an open access article under the CC BY license (<http://creativecommons.org/licenses/by/4.0/>).

Introduction

Protein-protein interactions (PPIs) play a pivotal role in most biological processes and thus, understanding how PPIs occur is an important step towards elucidating how these processes take place in cells and organisms. Studying the biochemical underpinnings behind PPIs can be better approached from a structural perspective.

Experimental techniques such as X-ray crystallography, nuclear magnetic resonance or cryo-electron microscopy are capable of solving the 3D structure of PPIs, in many cases, reaching atomic resolutions. However, these techniques are expensive, time-consuming, and they cannot keep pace with the amount of interactomic data that every year is generated. As a result, many computational approaches have been developed

to complement experimental methods and provide PPIs details at different levels of granularity.

In recent years, many computational methods have been designed to characterize PPIs when different levels of molecular information are available. For instance, protein docking methods can predict the full 3D structure of the PPI conformation when structural data of the interacting participants are available.^{1–5} When protein atomic models are not available, new deep learning methods have been highly successful predicting the tertiary structure of proteins.^{6,7} However quaternary structure prediction is more challenging and although initial steps have been conducted in that direction, they are computationally demanding and still require from manual intervention.⁸ As an alternative, lower granularity predictions can be computed fully automatically with less computational requirements. For instance, some methods can predict what protein regions or amino acids pairs might be involved in the interaction using sequence information.^{9–12} Another family of approaches predicts protein binding sites using sequence or structural information.^{13–18} One of such approaches is partner-specific binding site prediction.^{19–22} Contrary to conventional binding site prediction (non-partner specific), which aims to predict all the residues of a given protein that participate in any interaction, partner-specific methods seek to identify those residues that are involved in a particular PPI. Since proteins tend to interact with many distinct partners²³ and the involved interfaces can be quite different, partner-specificity is a convenient feature when studying a particular PPI.

Partner-specific predictors were firstly proposed by Ahmad and Mizaguchi¹⁹ and, since then, many more have been developed.^{20,24–30} Most of these methods aim to predict pairs of interacting residues, each belonging to a different protein partner, using machine learning algorithms trained over datasets derived from atomic models of protein complexes. Although several algorithmic approaches have been proposed, little emphasis has been made on the dataset used for training and developing these approaches. Thus, most, and especially the recently published partner-specific predictors, have been limited to small datasets, mainly the different versions of the Protein-Protein Docking Benchmark.^{16,19,20,25} Indeed, to the best of our knowledge, only the works of Meyer et al. and Townshend et al. tried to build datasets for this particular problem, yet their impact on performance was not analysed in detail.^{24,27} More importantly, only a single strategy for dataset compilation was considered.

In this work, we present BIPSPI+, a new version of our partner-specific binding site predictor that illustrates how a carefully selected training dataset can severely improve machine learning-based methods performance. BIPSPI+, as the original version,²¹ can be employed to predict the binding

sites of two interacting proteins given either their sequences or their structure. The new version offers a novel mode that can be used in those cases in which only the structure of one of the partners is known, exhibiting better performance than the sequence-only version. Additionally, the new approach was trained independently to predict binding sites for hetero- and homo-dimer cases. Overall, BIPSPI+ outperforms the original version in all studied datasets irrespectively of the input type, being especially worth noting the improvements for homo-complexes predictions.

In addition to offering better performance, the BIPSPI+ web server has been upgraded to include a new guided docking option that employs PatchDock² on BIPSPI+ predictions used as restraints. As a result, BIPSPI+ can now provide both binding site prediction and atomic models for the PPIs. To our knowledge, our method and the Ahmad and Mizaguchi one are the only partner-specific predictors available through web servers, and only ours allows the users to directly perform guided docking from the predictions.

BIPSPI+ web app is publicly available at <https://bipspi.cnb.csic.es> and as a stand-alone tool at <https://github.com/rsanchezgarc/BIPSPI>.

Methods

BIPSPI is a machine learning-based partner-specific binding site predictor trained on structurally solved protein assemblies deposited in the PDB.^{31,32} The training set consists of interacting and non-interacting residue pairs obtained from the 3D structure of protein complexes using a distance threshold criterion. BIPSPI+ is an upgraded version of the BIPSPI v1 web platform that implements three new major features: a new input mode (sequence & structure), complex-type stratification (homo-complex vs hetero-complex mode), and an optional step of guided Protein-Protein Docking (PP-docking). The following section briefly presents these new features, summarized in [Figure 1](#). For a complete description of the method, we refer the reader to the [Supplementary Material section 1](#).

Input: sequence-sequence, structure-structure and sequence-structure modes

BIPSPI v1 could be employed to predict the interacting residues of two protein structures or two sequences. BIPSPI+ has been redesigned to work also for cases in which only the structure of one of the interacting partners is known. This new input mode, which we have termed as the “sequence-structure mode”, employs sequence-only features to describe the sequence amino acids of the partner with no atomic model whereas residues of the other partner of the complex are described employing all features as in structure-structure mode. Consequently, the result page for this mode ([Figure 1\(j\)](#)) is a hybrid of the structure-structure ([Figure 1\(i\)](#)) and the sequence-sequence mode ([Figure 1\(h\)](#)) viewers, consisting of a 3D-

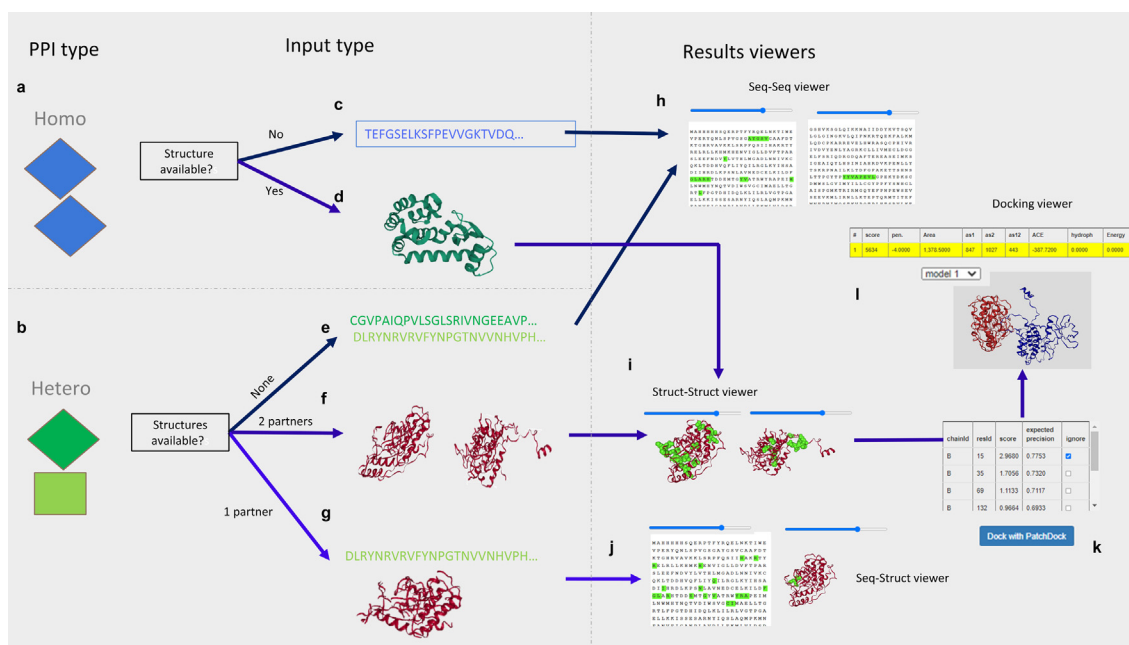


Figure 1. BIPSPI+ execution options. The protein complex to be predicted can be either a homo-complex (a) or a hetero-complex (b). Depending on the availability of atomic models, BIPSPI+ can be executed under different modes. The sequence-only mode is used if none of the structures is known (c and e), results being displayed in the Seq-Seq viewer (h). For the case of heterocomplexes in which only one of the structures is available (g), the sequence vs structure mode is used and the results are displayed in the Seq-Struct viewer (j). Finally, if the structure of the monomer is either known for homo-complexes (d) or the structure of the two interacting partners is known for hetero-complexes (f), the full structure mode is used instead. In this case, results are displayed in the Struct-Struct viewer (i), in which the structure of the two interacting partners, or two copies of the monomer, are shown. From this viewer, it is possible to execute guided docking, selecting the subset of residues to be employed as constraints (k). Docking results are displayed in the Docking viewer (l).

viewer for the partner with structure and a sequence panel for the partner with unknown structure.

Guided docking

BIPSPI+ web platform has been upgraded to perform an optional step of guided protein–protein docking using PatchDock². Thus, after computing binding site predictions for protein structures, the user can select, using a threshold slider and a table with checkboxes, which are the residues that will be used as restraints for guided docking. Then, PatchDock is executed with default parameters. After execution, a results page allows for interactive visualization of the highest-score predicted poses as well as downloading the atomic models and raw files generated during the docking step (see Figure 1(k–l)).

Homo-complexes and hetero-complexes datasets

BIPSPI+ was trained on two different datasets, one dataset consisting only of hetero-complexes (HEMt) and another dataset containing only of homo-complexes (HODt), both being more than one order of magnitude larger than the original

BIPSPI v1 training dataset. [Supplementary Material section 1.1, 1.2 and 2.5](#) describe these and other studied datasets. Similarly, the performance of our method was also assessed against two testing datasets representing the two possible types of complexes. Particularly, we employed the Protein-Protein Docking Benchmark v5 (Bv5), composed of 230 hetero-complexes and a custom evaluation benchmark, which we termed HOe (Homo-complexes evaluation), composed of 223 homodimers. Since the HOe dataset contains only complexes in bound state, performance could be overestimated when using structural features, but since comparison against BIPSPI v1 was also carried out on this dataset, improvement conclusions could be considered robust. Moreover, it is important to notice that the model trained on sequence-based features only is not affected by this problem and thus, its performance estimation is reliable.

Results

BIPSPI+ usage

BIPSPI+ can be employed to obtain partner-specific binding site predictions for protein

complexes. First, the user needs to select the oligomerization state of the PPI as homo-complex or hetero-complex (Figure 1(a and b), Supplementary Figure 1). Then, the user needs to provide either a sequence or an atomic model for the monomer in the homo-complex case (Figure 1(c and d), Supplementary Figure 1) or for both interacting partner in the hetero-complex case (Figure 1(e–g), Supplementary Figure 1). The 5 different oligomerization types and input types combinations (homo-sequence, homo-structure, hetero-sequence-sequence, hetero-sequence-structure, and hetero-structure-structure) are processed by 5 different models trained on the same types of data as the input.

After calculations, binding site predictions are displayed in one of the three different types of viewers depending on the input type (Figure 1(h–j)). In each of the viewers, the predicted interface residues with a score greater than the selected threshold are highlighted on the input sequences or structures (Figure 1(h–j)). Thresholds can be changed using a slider that displays the expected precision for the predictions given the current value of the threshold.²¹ For easiness of visualization, homo-complexes results are displayed using the same graphical interface in which two exact copies of the input monomer and the predictions are displayed as independent partners.

Finally, for the case of homo-complexes with structure or heterocomplexes with structures for both partners, it is possible to launch a guided docking job using as restraints the binding site residues predicted by BIPSPI at different thresholds (Figure 1(k), Supplementary Figure 3) or a custom subset of them, by checking the ignore checkbox of some of the residues with scores above the selected threshold. Once the residues to be used as restraints are selected, the docking calculations are carried out, and the highest score docking results are displayed in the Docking viewer, in which the user can visually inspect or download the proposed models (Figure 1(l), Supplementary Figure 4).

Better training data enhances performance

Since the performance of machine learning methods is severely influenced by the amounts and quality of the available data, it seems reasonable to believe that partner-specific binding site prediction can also benefit from this strategy. However, obtaining PPI complexes for a training dataset is challenging. First of all, the total number of solved complexes represents only a small fraction of the interactome. For instance, in humans, less than 10% of the binary interactions have been structurally solved.³³ Second, there are very few examples for which we know the structure of both the bound and unbound structure, most of them contained in the Bv5. While the former problem cannot be directly tackled until more experi-

mental data is obtained, the importance of the latter could be not so critical for methods like BIPSPI, which integrates both structural and sequence-based features.²¹ Consequently, for the second version of our method, we constructed larger training datasets that, for the majority of the complexes, do not contain the unbound version of the interacting partners. Despite this limitation, as it is shown in Figure 2 blue and red curves and described in Supplementary Material section 2.2, the inclusion of more bound complexes in the dataset was able to significantly improve results over BIPSPI v1. Thus, for the Bv5 using structural information, we measured a mean ROC AUC for residue-residue pairs interactions of 0.927 (median ROC AUC of 0.951) and a ROC AUC of 0.848 for binding site prediction. The new version increased both metrics with respect to our original method (0.905 and 0.823, respectively), achieving state-of-the-art performance (see Supplementary Table 5). For a detailed description of the evaluation approach see Supplementary Material section 1.3.

In addition to the size of the dataset, we also studied some other parameters that affect the quality of the data. For instance, we showed (see Supplementary Material section 2.3 and Supplementary Figure 8) that the inclusion of multimers, despite multiple caveats such as automatic receptor/ligand definition, enhances the performance of the predictions for predicting both dimers and multimers. Other studied parameters are discussed in Supplementary Material section 2.4–5.

Another important challenge when increasing the size of the dataset is the fact that most of the protein complexes contained in the PDB correspond to homo-complexes while the standard testing dataset, the Bv5, only contains hetero-complexes. Although the physics behind homo-complexes is the same that in hetero-complexes, statistical analysis show that physicochemical features of hetero- and homo-complex interfaces differ in many aspects such as contact preference, composition or hydrophobicity.¹¹ Consequently, some difference in performance could be expected depending on the oligomerization state. However, when we first studied the impact of the oligomerization type, the observed difference in performance for BIPSPI v1 was beyond our expectations, with a difference in MCC of 0.15 (see Supplementary Table 1) and important precision drops in the high-threshold region (high precision and low recall), the most interesting one for experimental validation (see Figure 2 left vs right panel). For BIPSPI+ we included homo-complexes in the training dataset using two strategies: first, training using two different datasets, one for each complex type (HEMt and HODt) and second, combining HEMt and HODt into one single training dataset (HEHODt). Supplementary Table 1 and Supplementary Figures 5–6 show that the first strategy offers comparable or bet-

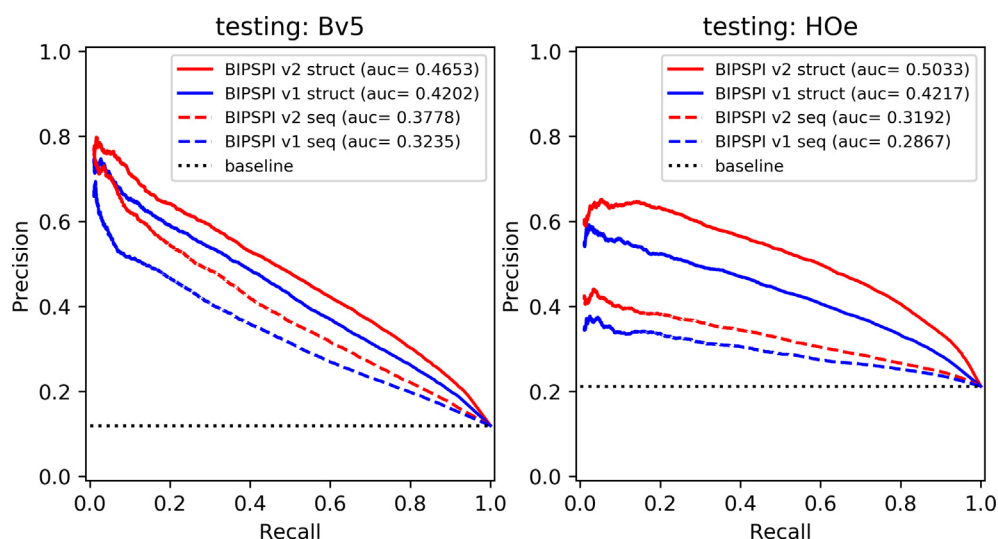


Figure 2. BIPSPI v1 and BIPSPI+ performance comparison. BIPSPI v1 (blue) and BIPSPI+ (red) precision-recall curves evaluated on hetero-complexes (Bv5, left) and homo-complexes (HOe, right) using as input either the sequences of the interacting partners (dashed lines) or their structures (solid lines).

ter results for all the analysed benchmarks, suggesting that the information extracted by our method from one oligomerization state is of little value, if not harmful for the other type. Consequently, in BIPSPI+, the users are required to select the oligomerization state and the two trained models are applied accordingly.

Sequence-structure mode

Partner-specific binding site predictors require sequence information and/or structural data for both interacting partners as input. So far, existing methods only consider the symmetric cases in which either the two structures or the two sequences are present. However, it is quite common that only the structure of one of the interacting protein partners is available (e.g., modelling low-resolution regions in cryo-EM maps, synthetic designs, etc.). Given the fact that structural data allows for better prediction performance, the common alternative of approaching those cases as if only the sequences were available is not compelling. In order to overcome this shortcoming, we have developed for BIPSPI+ the sequence-structure (seq-struct) mode.

We evaluated the performance of the seq-struct mode using as evaluation benchmark Bv5 and we studied how the new mode performed on both the input provided as sequence and the one provided as structure (see [Supplementary Material Table 3](#)). As expected, the quality of the predictions for the seq-struct mode, with an MCC value of 0.331, lies between the performance of the model that only employs sequences (MCC of 0.311) and the model that employs structures from

the two partners (MCC of 0.403). For more details, see [Supplementary Results section 2.6](#).

[Figure 3\(a\)](#) illustrates the benefits of this new execution mode on 2OZA, one of the protein complexes of the Bv5 for which we computed the predictions providing as input either the two sequences of chains A and B (X in unbound) or the sequence of the chain B and the structure of chain A. From direct inspection, it could be noticed that, when the structure of the studied protein partner is employed, the quality of the predictions largely improves. Thus, for chain A, the accuracy at threshold 0.5 is 0.60 when only the sequences are employed. However, when the structure of chain A is employed, accuracy gets boosted to 0.89.

Guided docking

While binding site predictions are invaluable sources of hypothesis for multiple experimental scenarios (e.g. mutagenesis experiments), when possible, 3D atomic models of the protein complexes offer a much richer description of PPIs. BIPSPI predictions have been successfully used as guided PP-docking constraints,^{34–36} improving the quality of 3D models. However, guided docking pipelines tend to be complicated, involving several computational steps and requiring a good understanding of the different tools.³⁷

With the aim of facilitating the generation of 3D models, we have included a simple guided docking pipeline based on PatchDock, a rigid body docking algorithm based on geometric hashing. Our pipeline simply requires the users to select a threshold for the binding site predictions so that the selected residues will be provided to PatchDock as binding site restraints, limiting the

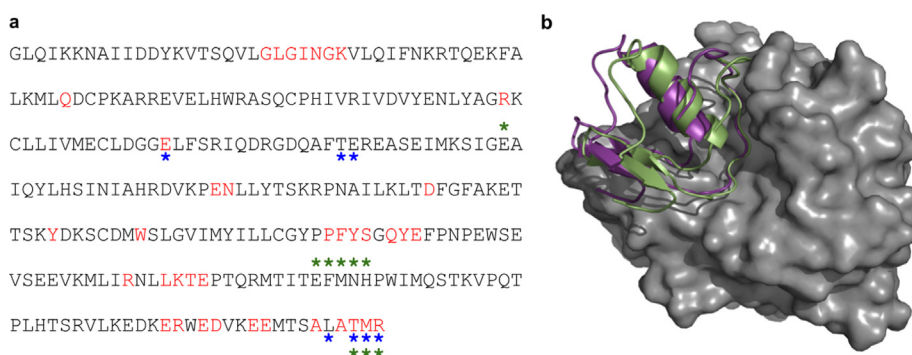


Figure 3. BIPSPI+ use cases. a) Sequence-structure mode improvement example. Sequence-only predictions (blue asterisks) and sequence-structure (green asterisks) predictions on structurally solved residues for Bv5 unbound complex 20ZA chain X (B in bound). Residues in contact with chain A are marked in red. Predictions above 0.5 score are marked with stars in blue when using only sequence information for both chain A and X and green when the structure of chain X is employed alongside the sequence of chain A. b) PatchDock docking model for the Subtilisin Carlsberg-OMTKY3 Complex (PDB code 1YU6, chains A and C respectively) obtained from BIPSPI+ web server. The crystallographic structures are depicted in grey for Chain A and green for Chain C whereas the docked model is depicted in purple.

search space to those poses compatible with the selected restraints. We acknowledge that our pipeline is simple and, consequently, better results could be easily obtained using more complicated pipelines and/or algorithms. However, our intention was to develop a user-friendly solution to retrieve fast initial structural models that could help the users to understand the binding site predictions.

Despite our pipeline's simplicity, accurate models can be obtained in many cases, providing conformational changes are not severe. Thus, Figure 3(b), illustrates an example of a 3D model for the protein complex Subtilisin Carlsberg-OMTKY3 Complex (PDB code 1YU6,³⁸ chains A and C respectively) computed using the BIPSPI+ web application. From direct inspection of the figure, it can be noticed that binding site predictions for this complex were of high quality, with an important part of the binding site accurately predicted. These accurate predictions ultimately allowed the docking algorithm to propose a high-ranked solution (3rd) of medium quality (iRMS 1.5 Å, DockQ = 0.64³⁹) in a totally automatic fashion.

Conclusion

Partner-specific binding site predictions have proven to be a useful resource in several contexts, especially for guiding protein-protein docking. Consequently, new approaches have been developed in recent times. However, while most of the new methods make special emphasis on algorithmic aspects, the crucial impact that datasets have on performance was not deeply studied. With the aim of addressing this issue, we developed BIPSPI+, an improved version of our original method, trained on carefully selected

datasets of complexes, that exhibit enhanced performance. While BIPSPI+ outperforms BIPSPI v1 in all the evaluated benchmarks, it is especially for the case of homo-complexes when performance is largely boosted. In addition to enhanced performance, the BIPSPI+ web application, freely available at <https://bipspi.cnb.csic.es>, has been updated to easily deal with homo-complexes and also for hetero-complexes in which only one of the interacting partners is structurally solved. Finally, the BIPSPI+ web application offers an optional step of guided protein-protein docking that can provide users with complete structural models of the protein interaction.

Data availability

Precomputed models and results are available at <https://zenodo.org/record/5574182#>. YYhiOrvLfmH. BIPSPI+ web server is available at <https://bipspi.cnb.csic.es/> the predictor code is available at <https://github.com/rsanchezgarc/BIPSPI>.

CRedit authorship contribution statement

R. Sanchez-Garcia: Conceptualization, Methodology, Software, Writing – original draft. **J. R. Macias:** Software, Validation. **C.O.S. Sorzano:** Validation, Writing – review & editing. **J.M. Carazo:** Funding acquisition, Writing – review & editing. **J. Segura:** Conceptualization, Data curation, Supervision, Writing – review & editing.

Acknowledgments

We acknowledge the Centro de Supercomputación de Galicia (CESGA) for their computational resources that were kindly provided.

Funding

Grant PID2019-104757RB-I00 funded by MCIN/AEI/10.13039/501100011033/ and “ERDF A way of making Europe”, by the “European Union”; “Comunidad Autónoma de Madrid” through Grant: S2017/BMD-3817; HighResCells (ERC - 2018 - SyG, Proposal: 810057); SEV-2017-0712 funded by MCIN/AEI/10.13039/501100011033; EOSC Life (INFRAEOSC-04-2018, Proposal: 824087); Grant DBI-1832184 by the National Science Foundation; Grant DE-SC0019749 by the US Department of Energy; and Grant R01GM133198 (Principal Investigator: Stephen K. Burley) by the National Cancer Institute, National Institute of Allergy and Infectious Diseases, and National Institute of General Medical Sciences of the National Institutes of Health.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supplementary data

Supplementary data to this article can be found online at <https://doi.org/10.1016/j.jmb.2022.167556>.

Received 27 November 2021;
Accepted 16 March 2022;
Available online 21 March 2022

Keywords:

protein interactions;
binding site;
web server;
machine learning

References

- Dominguez, C., Boelens, R., Bonvin, A.M.J.J., (2003). HADDOCK: A protein-protein docking approach based on biochemical or biophysical information. *J Am Chem Soc* **125**, 1731–1737.
- Schneidman-Duhovny, D., Inbar, Y., Nussinov, R., Wolfson, H.J., (2005). PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Res* **33**, W363–W367.
- Cheng, T.M.K., Blundell, T.L., Fernandez-Recio, J., (2007). pyDock: electrostatics and desolvation for effective scoring of rigid-body protein-protein docking. *Proteins* **68**, 503–515.
- Ghoorah, A.W., Devignes, M.D., Smaïl-Tabbone, M., Ritchie, D.W., (2013). Protein docking using case-based reasoning. *Proteins Struct Funct Bioinf* **81**, 2150–2158.
- Zhang, Q. et al, (2016). Recent advances in protein-protein docking. *Curr Drug Targets* **17**, 1586–1594.
- Baek, M. et al, (2021). Accurate prediction of protein structures and interactions using a three-track neural network. *Science* **373**, 871–876.
- Jumper, J. et al, (2021). Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589.
- Humphreys, I.R. et al, (2021). Computed structures of core eukaryotic protein complexes. *Science*. <https://doi.org/10.1126/science.abm4805>.
- Segura, J. et al, (2016). 3DIANA: 3D domain interaction analysis: A toolbox for quaternary structure modeling. *Biophys J* **110**, 766–775.
- Segura, J., Sorzano, C.O.S.S., Cuenca-Alba, J., Aloy, P., Carazo, J.M., (2015). Using neighborhood cohesiveness to infer interactions between protein domains. *Bioinformatics* **31**, 2545–2552.
- Ofran, Y., Rost, B., (2003). Predicted protein-protein interaction sites from local sequence information. *FEBS Lett* **544**, 236–239.
- Zhang, J., Kurgan, L., (2019). SCRIBER: Accurate and partner type-specific prediction of protein-binding residues from proteins sequences. *Bioinformatics*. Oxford Academic.
- Segura, J., Jones, P.F., Fernandez-Fuentes, N., (2012). A holistic in silico approach to predict functional sites in protein structures. *Bioinformatics* **28**, 1845–1850.
- Segura, J., Jones, P.F., Fernandez-Fuentes, N., (2011). Improving the prediction of protein binding sites by combining heterogeneous data and Voronoi diagrams. *BMC Bioinf* **12**, 352.
- Porollo, A., Meller, J., (2007). Prediction-based fingerprints of protein-protein interactions. *In: Proteins: Structure, Function and Genetics*, pp. 630–645.
- Savojardo, C., Fariselli, P., Martelli, P.L., Casadio, R., (2017). ISPRED4: Interaction sites PREDiction in protein structures with a refining grammar model. *Bioinformatics* **33**, 1656–1663.
- Šikić, M., Tomić, S., Vlahoviček, K., (2009). Prediction of protein-protein interaction sites in sequences and 3D structures by random forests. *PLoS Comput Biol* **5**, e1000278.
- Andreani, J., Quignot, C., Guerois, R., (2020). Structural prediction of protein interactions and docking using conservation and coevolution. *Wiley Interdiscipl Rev: Comput Mol Sci* **10**, e1470.
- Ahmad, S., Mizuguchi, K., (2011). Partner-aware prediction of interacting residues in protein-protein complexes from sequence data. *PLoS ONE* **6**, e29104.
- Minhas, F.ul A.A., Geiss, B.J., Ben-Hur, A., (2014). PAIRpred: partner-specific prediction of interacting residues from sequence and structure. *Proteins* **82**, 1142–1155.

21. Sanchez-Garcia, R., Sorzano, C.O.S., Carazo, J.M., Segura, J., (2019). BIPSPI: A method for the prediction of partner-specific protein-protein interfaces. *Bioinformatics* **35**, 470–477.
22. Xue, L.C., Dobbs, D., Bonvin, A.M.J.J., Honavar, V., (2015). Computational prediction of protein interfaces: A review of data driven methods. *FEBS Lett* **589**, 3516–3526.
23. Grigoriev, A., (2003). On the number of protein-protein interactions in the yeast proteome. *Nucleic Acids Res* **31**, 4157–4161.
24. Meyer, M.J. et al, (2018). Interactome INSIDER: a structural interactome browser for genomic studies. *Nat Methods* **15**, 107–114.
25. Fout, A., Shariat, B., Byrd, J., Ben-Hur, A., (2017). Protein interface prediction using graph convolutional networks. *Adv Neural Inform Process Syst* **30**, 6512–6521.
26. Chen, T., Guestrin, C., (2016). XGBoost: A Scalable Tree Boosting System. In: *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining - KDD '16*, pp. 785–794. <https://doi.org/10.1145/2939672.2939785>.
27. Townshend, R.J.L., Bedi, R., Suriana, P.A., Dror, R.O., (2019). End-to-end learning on 3D protein structure for interface prediction. *Advances in Neural Information Processing Systems*. Neural Information Processing Systems Foundation.
28. Dai, B., Bailey-Kellogg, C., (2021). Protein interaction interface region prediction by geometric deep learning. *Bioinformatics* **37**, 2580–2588.
29. Xue, L.C., Dobbs, D., Honavar, V., (2011). HomPPI: a class of sequence homology based protein-protein interface prediction methods. *BMC Bioinf* **12**, 244.
30. Yan, Y., Huang, S.Y., (2021). Accurate prediction of inter-protein residue-residue contacts for homo-oligomeric protein complexes. *Briefings Bioinf* **22**, 1–13.
31. Berman, H.M. et al, (2000). The protein data bank. *Nucleic Acids Res* **28**, 235–242.
32. Burley, S.K. et al, (2021). RCSB Protein Data Bank: Powerful new tools for exploring 3D structures of biological macromolecules for basic and applied research and education in fundamental biology, biomedicine, biotechnology, bioengineering and energy sciences. *Nucleic Acids Res* **49**, D437–D451.
33. Mosca, R., Céol, A., Aloy, P., (2012). Interactome3D: adding structural details to protein networks. *Nat Methods* **10**, 47–53.
34. Pozzati, G., Kundrotas, P. & Elofsson, A. Improved protein docking by predicted interface residues. *bioRxiv* 2021.08.25.457642 (2021) 10.1101/2021.08.25.457642.
35. Id, A.L.A., Barnes Id, A.B., Martin Id, A.T., Wang, L., Koid, D.C., (2021). Variation in Leishmania chemokine suppression driven by diversification of the GP63 virulence factor. *PLOS Neglected Tropical Dis* **15**, e0009224.
36. Sharma, A. et al, (2021). Ppar-responsive elements enriched with alu repeats may contribute to distinctive ppar γ -dnmt1 interactions in the genome. *Cancers* **13**, 3993.
37. Segura, J., Marín-López, M.A., Jones, P.F., Oliva, B., Fernandez-Fuentes, N., (2015). VORFFIP-Driven Dock: V-D2OCK, a fast and accurate protein docking strategy. *PLoS ONE* **10**, e0118107.
38. Maynes, J.T., Cherney, M.M., Qasim, M.A., Laskowski, M., James, M.N.G., (2005). Structure of the subtilisin Carlsberg-OMTKY3 complex reveals two different ovomucoid conformations. *Acta Crystallogr. Section D, Biol Crystallogr* **61**, 580–588.
39. Basu, S., Wallner, B., (2016). DockQ: A quality measure for protein-protein docking models. *PLoS ONE* **11**, e0161879

BIPSPI+: Mining type-specific datasets of protein complexes to improve protein binding site prediction

R Sanchez-Garcia^{1,2*}([orcid 0000-0001-6156-3542](https://orcid.org/0000-0001-6156-3542)), JR Macias¹([orcid 0000-0003-2621-6806](https://orcid.org/0000-0003-2621-6806)), COS Sorzano¹, ([orcid 0000-0002-9473-283X](https://orcid.org/0000-0002-9473-283X)), JM Carazo¹, ([orcid 0000-0003-0788-8447](https://orcid.org/0000-0003-0788-8447)) J Segura³ ([orcid 0000-0003-0788-8447](https://orcid.org/0000-0003-0788-8447))

1. Biocomputing Unit, National Center for Biotechnology (CSIC), Darwin 3, Campus Univ. Autónoma de Madrid, Cantoblanco, 28049 Madrid, Spain
 2. Oxford Protein Informatics Group, Department of Statistics, University of Oxford, Oxford OX1 3LB, UK
 3. Research Collaboratory for Structural Bioinformatics Protein Data Bank, San Diego Supercomputer Center, University of California, La Jolla, CA 92093, USA
- * Corresponding ruben.sanchez-garcia@stats.ox.ac.uk

1. Supplementary Methods

1.1. Training datasets

The training datasets employed in this work can be broadly classified into three categories: only heterocomplexes datasets; only homocomplexes datasets; and a mixture of the previous two types. As only heterocomplexes representative, we compiled a new dataset, termed HEDt, consisting of 2,401 bound heterodimers collected from the PDB. Additionally, we extended the HEDt dataset to include 1,571 hetero-multimers resulting in the HEMt dataset. It is important to notice that, contrary to the case of dimers, when multiple chains are included in the atomic models, splitting the complex into its ligand and receptor components is not trivial and, in many cases, subjective. At the risk of making mistakes (and introducing noise during training but not in evaluation), we simply extracted each of the interacting chains together with the other chains that are in contact with each of them. In the case of a shared interacting chain, we assign it to one of the partners randomly. Similarly, for the homo-complexes dataset, we have collected from PDB 1,981 bound homodimers that constitute the HODt dataset. Finally, the HEHODt dataset is composed of the heterodimers included in HEDt and the homodimers contained in HODt, thus exhibiting a balanced proportion between homo-complexes and heterocomplexes (~8:11).

For the sequence-structure mode, we generate two training instances for each complex, one in which the sequence is available for the ligand and the structure for the receptor and another in which the roles are reversed.

1.2. Dataset compilation

All the protein complexes included in this study, except for the ones comprising Bv5¹, were collected from the PDB database according to the following criteria: 1) resolution better than 3.5 Å; 2) number of residues structurally determined >50%; 3) sequence length of each chain >30 residues; and 4) number of interacting residue pairs >10.

Due to the fact that some protein families are overrepresented in the PDB, using all available protein complexes would result in biased datasets and thus, potentially leading to poor performance in machine learning models. For this reason, in order to preserve diversity while reducing redundancy, we used a combination of pairs of SCOPe families and sequence-based clustering as sampling criteria. In particular, we have grouped protein complexes based on their SCOPe families, and within each SCOPe pair group, we have further divided it into groups according to sequence identity (95% threshold). Then, we have selected as a representative for each of the groups the structure with the best resolution. Protein chains lacking SCOPe family classification were grouped based only on sequence identity-based clusters using a 30% identity threshold. This threshold may be considered as the limit to observe 3D structural similarity between proteins and thus, avoids including a large amount of structural redundancy from the null class². Proceeding in this way, we reduce the redundancy introduced by highly-populated SCOPe families and non-classified proteins while preserving most of its diversity.

We employed the same definition as in BIPSPI, SASnet, and many other previous publications. Consequently, a pair of residues is labelled as interacting if the distance between any of their heavy atoms is $<6.0 \text{ \AA}$. Due to symmetry reasons, when training and/or evaluating with homo-complexes, we have corrected the list of interacting pairs to include as positive pairs those that are not directly in contact in the structures but that are equivalent to others that are in contact. For example, given the homodimer A-B and the interacting residue pair A:i-B:j, the pair A:i-B:j should also be considered as positive since residues A:j and B:j and A:i and B:i are equivalents. Such correction could be also useful for the case of hetero-complexes in which one of the partners is a homo-complex, but it was not considered in this work in order to simplify comparisons.

1.3. Evaluation

The performance for Residue-Residue Interaction (RRI) prediction was measured by computing the mean Area Under the ROC Curve (mRAUC) as in many other works. Binding site prediction performance was evaluated using several metrics, including the Area Under the ROC Curve (RAUC), Matthew's Correlation Coefficient (MCC), True and False Positive Rates (TPR, FPR), and Positive Predictive Value (PPV) (see Sanchez-Garcia et al.³ for more details). Violin plots displaying the distribution of RRI ROC-AUC and binding site AUC were also computed.

Hetero-complexes predictions were evaluated using the Protein-Protein Docking Benchmark v5 (Bv5), a dataset of 230 hetero-complexes for which both the bound and unbound structures are available Bv5. Since we are interested in the performance under different oligomerization states, we have also computed the same metrics for two subsets of the Bv5: BM90C⁴, the subset of heterodimers contained in Bv5 and Bv5Mul, the subset of multimeric proteins contained in Bv5.

For the case of homo-complexes, we compiled an evaluation benchmark following the same principles as in Bv5 except for the fact that all the complexes employed are dimers in bound state. This evaluation benchmark, which we termed HOe, is composed of 223 homodimers. Notice that HOe comprises only bound complexes and thus, performance using structural information could be overestimated. On the contrary, the performance estimations measured using sequence-only features are not affected by this caveat.

For the sequence-structure mode, since we generated two training instances for each complex, evaluation is performed also on the two instances, reporting them independently for each partner and also as averages when considering the whole complexes.

The evaluation process consisted of a 10-fold cross-validation approach between testing and training set, i.e. the testing set was divided into 10 subsets of equal size and then, for each subset a model was trained removing from the training set any protein that shared a SCOPe domain with the particular testing subset. This approach guarantees that the training does not contain any information on the tested data.

1.4. Algorithm

BIPSPI+ employs the same algorithm that BIPSPI v1 uses, including the same features, model and hyperparameters. The main differences with respect to the original version are related to the different new input types. First, in version 2 we always apply two stacked models (feedback model) independently of the input type provided, whereas originally, this strategy was only employed for structural input. Second, for the sequence-structure mode, we employ sequence-only features for one of the partners and both structural and sequence-only features for the other. Last, for homocomplexes prediction, the algorithm is executed with two copies of the same monomer as inputs and the final predictions for the pairs are computed averaging the predictions for the same.

In addition to the aforementioned modifications, the procedure to ensure the independence between the training and testing set is also different, since now there are more than one complex with the same pair of SCOPe families. Consequently, a grouped ten-fold cross-validation strategy, using as groups pairs of SCOPe families, was used to prevent cross-contamination. Proteins with no SCOPe defined are assigned to virtual families according to sequence clustering at 30% identity.

1.5. Data augmentation

We generated simulated conformations, obtained from the atomic models contained in the training sets, as a novel type of data augmentation. Particularly, we randomly sampled poses from trajectories generated with the “imc” program of the iMod package⁵, using default parameters. This program performs a Monte Carlo simulation guided by Normal Modes Analysis on Internal Coordinates and generates plausible trajectories that begin at the atomic model provided. Other alternatives such as Molecular Dynamics or Flexible Docking were not considered due to computational limitations but could produce similar results.

1.6. Method comparison

For comparison with other methods, we employed the SASNet neural network⁶ as described in the original publication and we trained it on the HEMt dataset since SASNet original publication reported performance using Bv5 and our best dataset for Bv5 is HEMt. Moreover, due to the fact that SASNet was trained on both the Bv5 and DIPs, a custom dataset, a direct comparison between DIPs and HEMt can be conducted.

2. Supplementary Results

2.1. BIPSPI+ Usage

HOME PREDICT binding sites HELP ABOUT US

Hetero complex Homo complex

partner 1 seq partner 1 structure

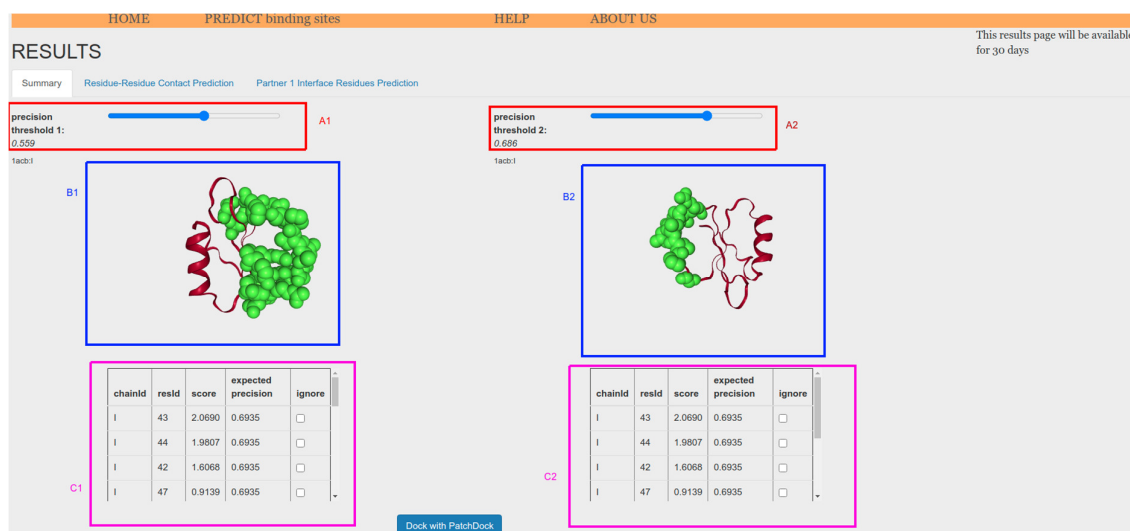
partner 1 sequence: or upload your fasta sequence 1: No file chosen

partner 2 seq partner 2 structure

partner 2 PDB id: or upload your model 2: No file chosen

Examples
4ov6:E-4ov6:D

Supplementary Figure 1. BIPSPI+ input page. The user needs to select whether the PPI is a homo-complex or a hetero-complex (red box). Then, for each interacting partner (or for the monomer in the case of homo-complexes), the user needs to provide either the sequence or the structure of the protein after selecting the input type using the radio buttons (blue box). Structures can be provided (orange box) as either “.pdb” files or be automatically downloaded if a PDB id is provided instead. Sequences can be provided (green box) as either “.fasta” files or by directly pasting them into textboxes.



Supplementary Figure 2. BIPSPI+ results page for structure-structure mode. The structures of the two interacting partners are displayed in the blue boxes. Residues with scores higher than the threshold are displayed in green in the structure and their scores are displayed in tables (pink box). The threshold, that represents the expected precision, can be set using the sliders at the top of the tab (red boxes).

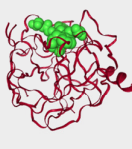
HOME PREDICT binding sites HELP ABOUT US

RESULTS This results page will be available for 30 days

Summary Residue-Residue Contact Prediction Partner 1 Interface Residues Prediction Partner 2 Interface Residues Prediction

precision threshold 1: 0.652
1acb:I

precision threshold 2: 0.651
1acb:E

chainid	resid	score	expected precision	ignore
I	45	4.0794	0.8279	<input type="checkbox"/>
I	46	3.0840	0.7753	<input type="checkbox"/>
I	47	2.0788	0.7386	<input type="checkbox"/>
I	44	1.7410	0.7320	<input type="checkbox"/>

Dock with PatchDock

chainid	resid	score	expected precision	ignore
E	214	1.4683	0.7248	<input type="checkbox"/>
E	216	1.4351	0.7247	<input type="checkbox"/>
E	193	1.2894	0.7201	<input checked="" type="checkbox"/>
E	192	1.1137	0.7119	<input checked="" type="checkbox"/>

Supplementary Figure 3. BIPSPI+ interface for launching guided docking. By default, all residues with scores above the threshold will be used as constraints for guided docking. Thresholds can be modified using the sliders on top of the structures (red boxes). Residues above the threshold can be excluded by ticking their associated checkbox in the tables (blue boxes). Guided docking will be launched as soon as the “Dock with PatchDock”⁷ button (violet box) is pressed.

HOME PREDICT binding sites HELP ABOUT US

DOCKING SOLUTIONS for: 1acb:I ~ 1acb:E This results page will be available for 30 days

model number
model 1 A

Download all solutions E

Download Ligand C1

Download Receptor C2

B



#	score	pen.	Area	as1	as2	as12	ACE	hydroph	Energy	cluster
1	9562	-2.9100	1,250.6000	2248	3523	4241	-521.3200	0.0000	0.0000	0
2	9474	-2.3500	1,121.1000	1759	1688	1864	-102.2100	0.0000	0.0000	0
3	9134	-2.6000	1,166.0000	1779	1604	713	-158.5600	0.0000	0.0000	0
4	9094	-2.8800	1,306.5000	1010	1294	0	-356.9600	0.0000	0.0000	0
5	9076	-2.4200	1,192.0000	1674	2170	2172	-101.3600	0.0000	0.0000	0
6	9056	-3.1400	1,656.5000	1396	990	236	-433.0900	0.0000	0.0000	0
7	8966	-2.7300	1,012.6000	1213	1838	384	-179.9700	0.0000	0.0000	0
8	8888	-3.0400	1,367.1000	1505	2075	1038	-298.6000	0.0000	0.0000	0
9	8838	-3.2100	1,231.2000	1520	788	231	-298.7100	0.0000	0.0000	0
10	8776	-2.6400	1,101.3000	1687	1432	565	84.9900	0.0000	0.0000	0

D

Supplementary Figure 4. Docking viewer displays the top-10 highest score models. The model currently displayed in the 3D viewer (blue box) and highlighted in the scores table (green box) can be selected in the drop-down list (red box). The displayed model can be individually downloaded (pink box) or jointly downloaded with the other solutions (cyan box).

2.2. BIPSPI+ overall performance analysis

Supplementary Table 1 summarizes the results obtained with BIPSPI+ trained on several datasets and evaluated in both Bv5 and HOe. From direct inspection, several conclusions can be drawn.

First, the training datasets proposed in this work significantly outperform original BIPSPI v1 results. Supplementary Table 2 shows the p-values obtained from multiple statistical tests that

compare the performance of both versions. Additionally, in Supplementary Figure 7, BIPSPI+ exhibits far better performance than BIPSPI v1 when sequences are used as input. In that case, the improvements are so important that, for the residue-residue interaction (RRI) problem, the first and second quartile in BIPSPI+ approximately matched the second and third quartile in BIPSPI v1. More importantly, the improvement in RRI results also translate to an important improvement in binding site prediction, in which the BIPSPI+ distribution is shifted by $\sim 1/4$ of the interquartile range. For the case of homocomplexes, independently of whether the input is a sequence or a structure, a similar improvement is observed. The differences in performance for heterocomplexes using structural features are less striking, but still statistically significant in all computed tests (see Supplementary Table 2) and visually noticeable. Since for all cases the first quartile is the one that varies more between versions, this implies that the worse performing examples tend to be better predicted in our new version, although improvements are observed for all the range of values.

Second, the stratification of the training data into homo- or hetero-complexes leads to better results, as it can be concluded from the facts that 1) when the opposite type of oligomerisation dataset is used for training, results severely worsen, and 2) the performance measured when using HEHODt is similar or slightly worse than when using their specific counterparts (HEMt/HEDt and HODt, see Supplementary Figure 5 and Supplementary Figure 6). Additionally, it can also be concluded that the addition of multimers to the dataset has an overall positive effect even when the inputs are two sequences and thus, the concept of multimer is not naturally modelled.

In the following subsections, some particular aspects of the dataset will be studied in more detail.

Supplementary Table 1. Ten-fold cross-validation performance evaluated on Bv5 and HOe for several training datasets.

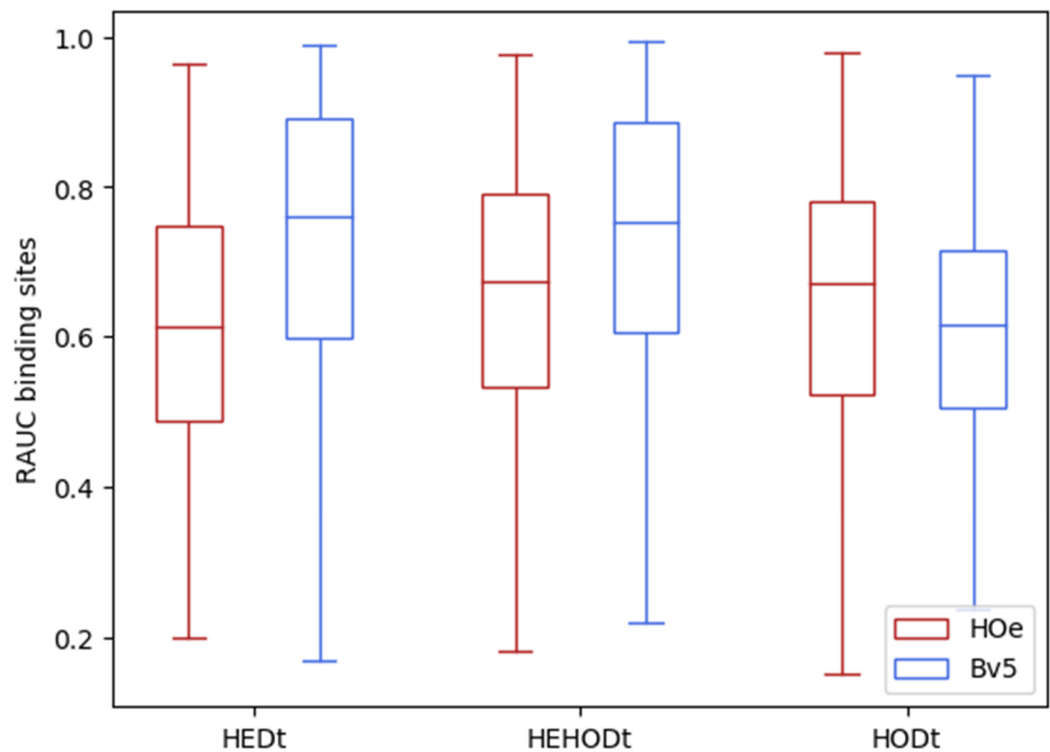
Testset	Trainset	RRI	Binding site prediction				
		mRAUC	RAUC	MCC	PPV	TPR	FPR
Input type: Sequence							
Bv5	HODt	0.681	0.654	0.157	0.213	0.415	0.210
	HEDt	0.844	0.751	0.310	0.380	0.415	0.093
	HEHODt	0.836	0.750	0.300	0.370	0.409	0.095
	HEMt	0.868	0.786	0.336	0.380	0.473	0.104
	Bv5 (BIPSPI v1)	0.802	0.753	0.279	0.300	0.482	0.113
DOCKGR OUND4	HEMt	0.831	0.749	0.285	0.325	0.437	0.113
	Bv5 (BIPSPI v1)	0.814	0.730	0.261	0.304	0.420	0.121
HOe	HODt	0.758	0.695	0.244	0.340	0.607	0.318
	HEDt	0.706	0.642	0.171	0.284	0.654	0.445

	HEHODt	0.760	0.695	0.245	0.363	0.525	0.248
	HEMt	0.721	0.658	0.191	0.325	0.500	0.280
	Bv5 (BIPSPI v1)	0.656	0.623	0.145	0.285	0.536	0.362
Input type: Structure							
Bv5	HODt	0.837	0.742	0.276	0.311	0.471	0.143
	HEDt	0.917	0.826	0.403	0.432	0.541	0.097
	HEHODt	0.914	0.824	0.396	0.420	0.543	0.102
	HEMt	0.927	0.848	0.422	0.438	0.573	0.100
	Bv5 (BIPSPI v1)	0.905	0.823	0.386	0.391	0.5585	0.089
HOe	HODt	0.898	0.830	0.447	0.512	0.655	0.167
	HEDt	0.856	0.797	0.391	0.446	0.671	0.224
	HEHODt	0.889	0.825	0.441	0.485	0.693	0.198
	HEMt	0.870	0.810	0.415	0.454	0.707	0.229
	Bv5 (BIPSPI v1)	0.805	0.750	0.322	0.409	0.595	0.231
<p>Notes: Bv5 230 hetero-multimers (including dimers) HOe 223 homodimers HEDt 2401 heterodimers HEMt 3972 hetero-multimers (HEDt + Higher order) HODt 1981 homodimers HEHODt 4382 heterodimers + homodimers (HEDt + HODt)</p> <p>Metrics: mRAUC: mean ROC AUC RAUC: ROC AUC, all predictions pooled together MCC: Matthews correlation coefficient PPV: Positive Predictive Value TPR: True Positive Rate FPR: False Positive Rate</p>							

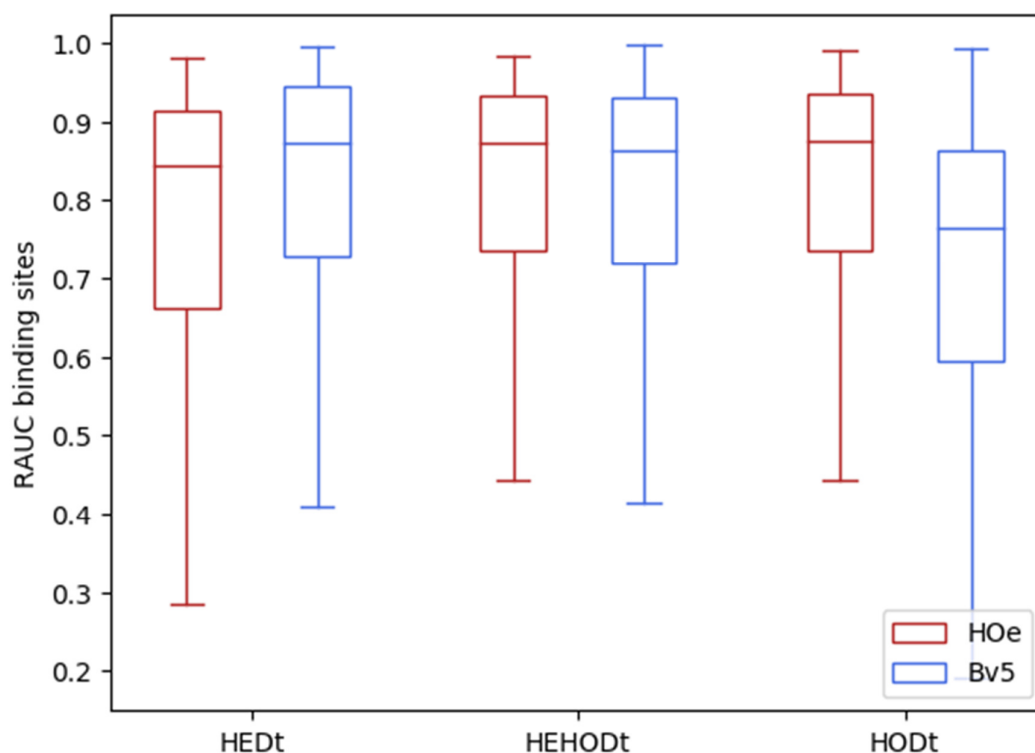
Supplementary Table 2. Statistical tests results comparing BIPSPI v1 and BIPSPI+ ROC-AUCs per complex for the residue-residue interaction (RRI) and the binding site prediction problem.

		Heterocomplexes		Homocomplexes	
One-sided Wilcoxon test					
Prediction target	Input type	Statistic	p-value	statistic	p-value
RRI	Sequence	22490.5	3.938E-20	20688.5	9.326E-18
RRI	Structure	20152	5.230E-12	23589	5.942E-31

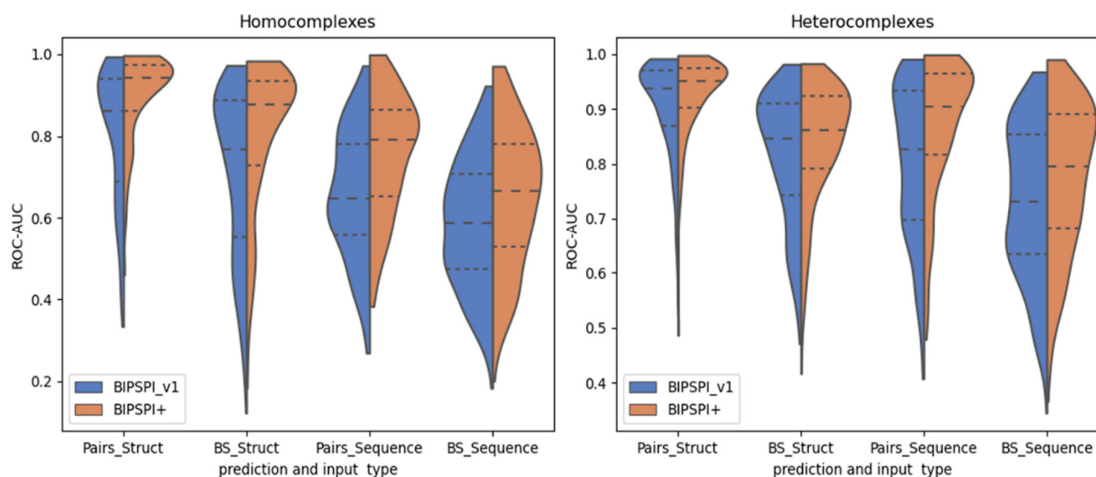
Binding Site	Sequence	20011	1.365E-11	17914	9.253E-09
Binding Site	Structure	18540	9.740E-08	21652	1.041E-21
One-sided paired t-test					
RRI	Sequence	9.647263	5.327E-19	9.834859	1.770E-19
RRI	Structure	6.612167	1.319E-10	12.307265	3.510E-27
Binding Site	Sequence	7.322938	2.048E-12	5.788818	1.200E-08
Binding Site	Structure	5.998334	3.856E-09	9.834859	1.770E-19



Supplementary Figure 5. ROC-AUC (RAUC) distribution for binding site predictions for sequence-sequence mode, trained on HEDt, HEHODt, and HODt and evaluated on H0e (red) and Bv5 dataset (blue).



Supplementary Figure 6. ROC-AUC (RAUC) distribution for binding site prediction for structure-structure mode, trained on HEDt, HEHODt, and HODt and evaluated on HOe (red) and Bv5 dataset (blue).

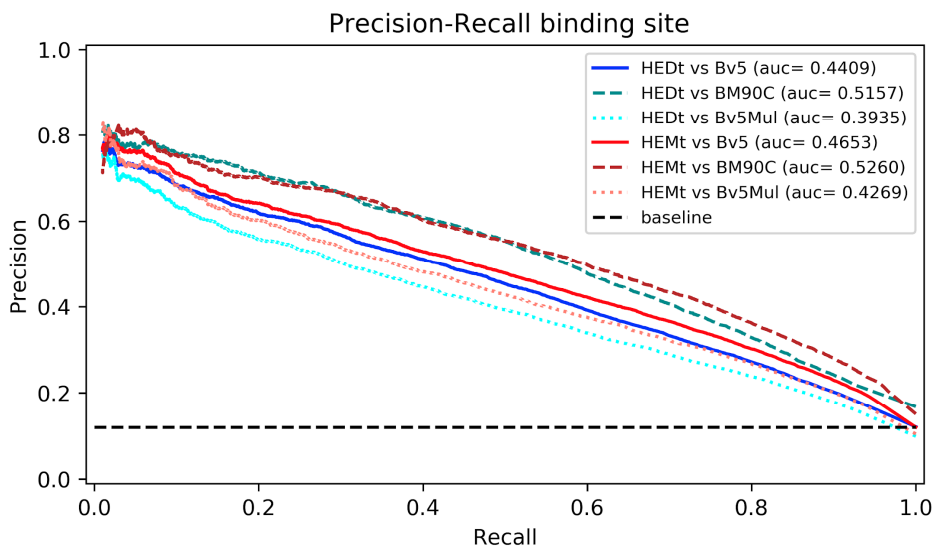


Supplementary Figure 7. ROC-AUC distributions for the residue-residue interaction problem (Pairs) or the binding site prediction problem (BS) at the complex level, for the complexes contained in HOe (Homocomplexes) and Bv5 (Heterocomplexes) predicted using BIPSPI v1 and BIPSPI+ using as input sequences (*_Sequence) or structures (*_Struct)

2.3. Dimers and multimers

Automatic compilation of large training sets including only biologically feasible complexes is not a simple task. Serve as an example the case of an atomic model of a complex that contains 4 protein chains. If no additional information is available, it might not be trivial to determine whether the complex is composed of two dimers interacting or one trimer interacting with a monomer. On the contrary, limiting to atomic models of two chains makes automatic

compilation much simpler, since most of them will represent actual dimeric interactions. With the aim of determining the impact of oligomerization number on binding site prediction performance, we trained BIPSPI on hetero-dimers only (HEDt) and also on a dataset that includes both heterodimers and automatically sampled multimers (HEMt). The first conclusion that can be derived from this comparison (see Supplementary Figure 8, dotted vs dashed lines) is that, in both cases, the performance predicting dimers (BM90C) is superior to the performance predicting hetero-multimers (Bv5Mul). This result is not surprising since the number of potential interacting residue pairs in heterocomplexes tends to be much larger and also because determining which protein chains are in contact supposes an additional challenge. Secondly, we can also observe that the addition of multimers in the training set improved the performance of the method for both dimers and multimers, yet the improvement was larger in the latter case (HEDt, blue tones vs HEMt, red tones). Those results suggest that, despite the fact that automatic multimer processing is not perfect, on average, including this data, despite its noise, positively contributes to the predictions.



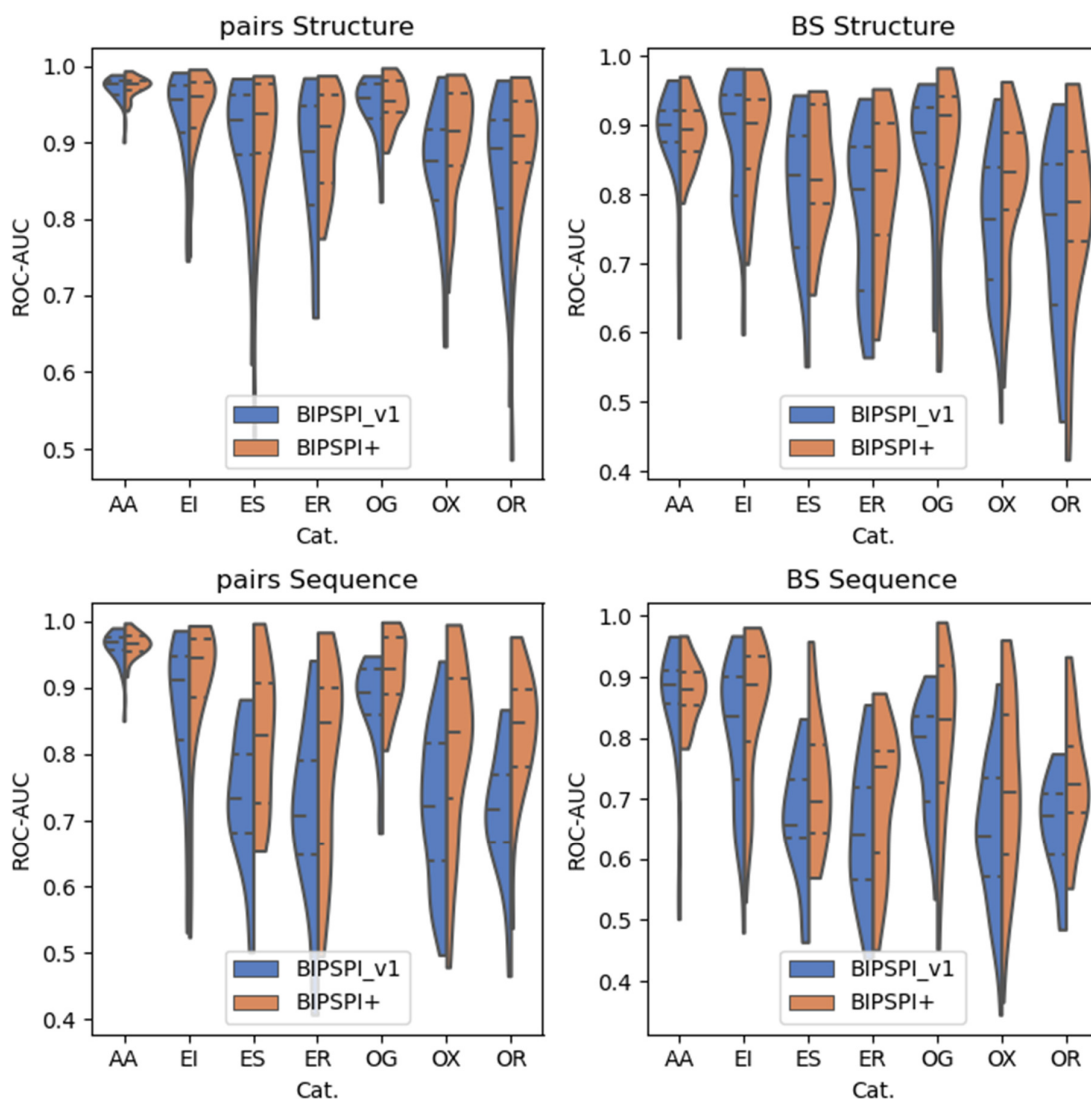
Supplementary Figure 8. Precision-recall curves for heterodimers and hetero-multimer evaluation. The training was performed on HEDt (blue tones) or HEMt (red tones), and performance was recorded for the whole Bv5 (solid lines), the dimers subset (BM90C, dashed lines) or the multimers subset (Bv5Mul, dotted lines).

2.4. Performance per complex type

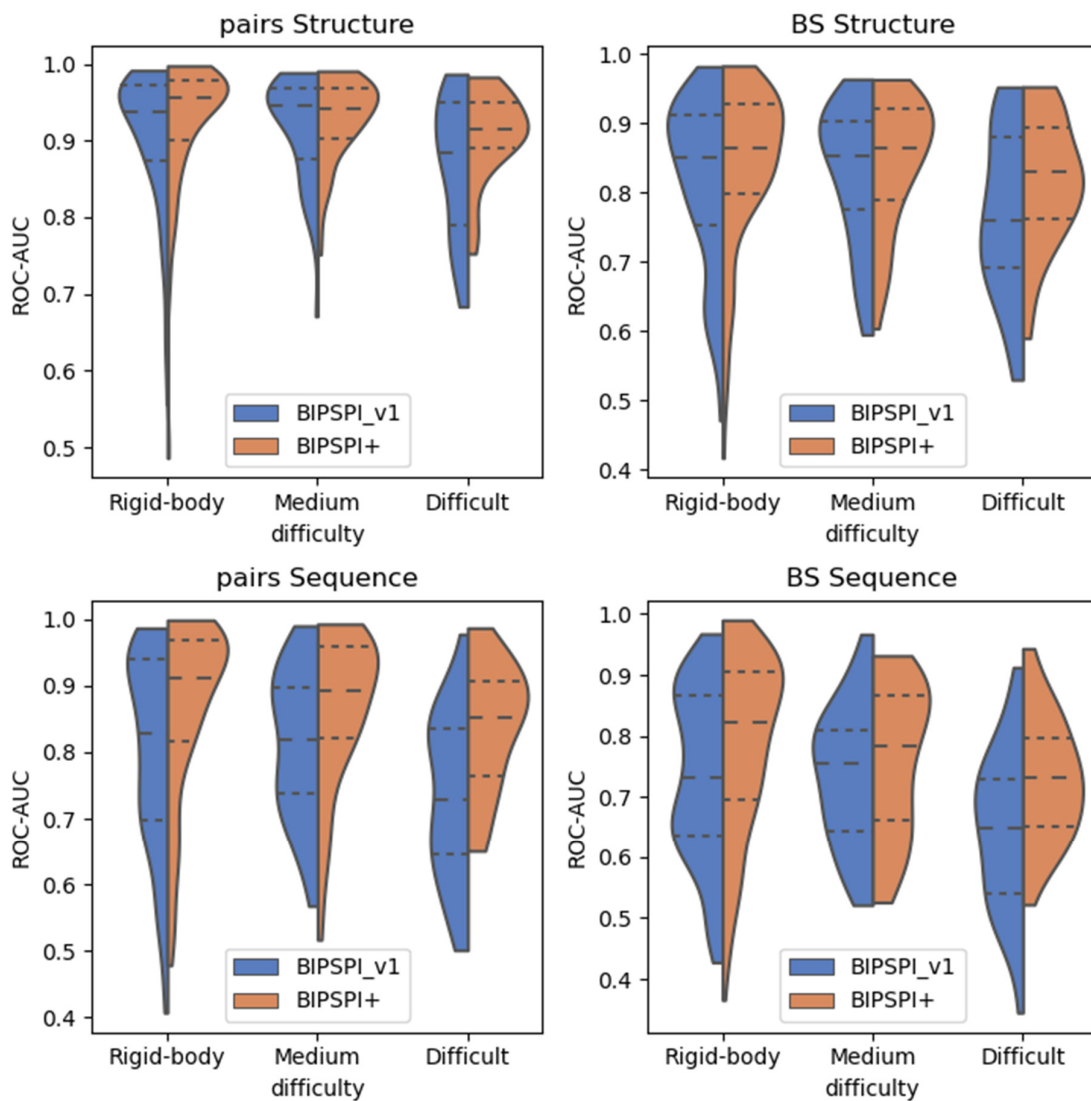
We examined the performance of BIPSPI+ considering the different types of complexes contained on Bv5.

With respect to the Bv5 difficulty levels (see Supplementary Figure 9), BIPSPI+ exhibits similar performance for the easy (rigid-body) and medium difficulty cases when using structures as input (median ROC-UAC of 0.86 for binding site prediction in both cases). The performance for the difficult cases is considerably worse (median ROC-UAC of 0.83) but is still remarkable for an important amount of cases as more than 25% of the difficult complexes obtained a ROC-AUC for binding site prediction above 0.9. On the other hand, when using sequences as inputs, the degree of variability is much larger and the quality of the predictions is more influenced by the difficulty of the complex with median ROC-UAC values of 0.82, 0.78 and 0.73 for binding site prediction of easy, medium and difficult complexes respectively. It is also worth noting that BIPSPI+ predictions are considerably better than BIPSPI v1 predictions, especially for the difficult complexes where the first quartile in BIPSPI+ ROC-AUC approximately corresponds to the median value in BIPSPI v1.

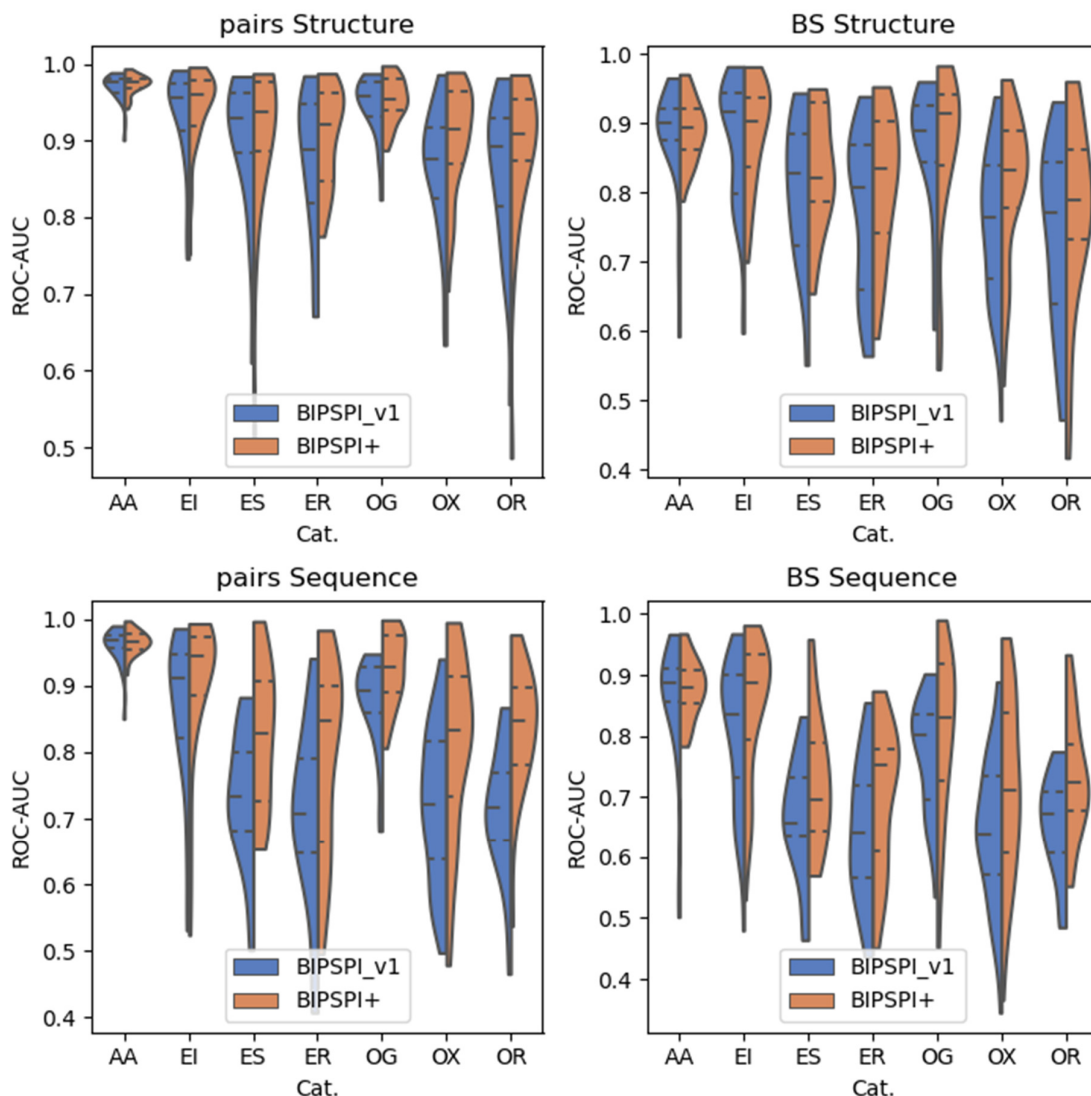
Next, we break down the performance analysis according to the protein-complex type. As it is shown in



Supplementary Figure 10, both BIPSPI v1 and BIPSPI+ exhibit their best performance for complexes of type antigen-antibody and enzyme-inhibitors, independently of the input type. For these two types of protein complexes, BIPSPI+ results (median ROC-AUCs for binding site prediction of 0.89 y 0.90 using structures as input) do not tend to outperform the already highly accurate BIPSPI v1 results (median ROC-AUCs for binding site prediction of 0.90 y 0.91 using structures as input). For the rest of complex types, BIPSPI+ outperforms our original version, especially, as stated before, when using sequences as input. Despite BIPSPI+ improvements, enzyme-substrate and enzyme-regulator complexes together with interactions involving receptors are the worst performing categories by a large margin.



Supplementary Figure 9. ROC-AUC distributions for the complexes contained in Bv5 according to difficulty levels predicted with BIPSPI v1 and BIPSPI+ using as input either sequences or structures. The term Pairs refers to the problem of residue-residue interaction prediction while BS refers to the problem of binding site prediction.



Supplementary Figure 10. ROC-AUC distributions for the complexes contained in Bv5 according complex types predicted using BIPSPI v1 and BIPSPI+. The term Pairs refers to the problem of residue-residue interaction prediction while BS refers to the problem of binding site prediction. AA: Antigen-antibody; EI: Enzyme-inhibitor; ES: Enzyme-substrate; ER: Enzyme-regulatory/accessory chain; OG: Complex containing G-protein; OX: Other Complexes; OR: Complex containing Receptor

2.5. Dataset selection criteria

Since the new datasets contained protein complexes in bound state, in order to determine which should be the properties of the best dataset, we evaluated the performance of different datasets using models trained with sequence features and using as the validation dataset the subset of dimers contained in the Bv5. For this experiment, we labelled the datasets using a 6-letter code. The two initial characters indicate the type of interaction: (HE) heterodimers, (HO) homodimers or (DM) heterodimers sampled from multimers. The next two letters indicate if the dataset redundancy was filtered using SCOP families and group clusters (CL) or only SCOP families were considered (NC). Finally, the two last letters indicate if a pair of PDB chains could contain the SCOP null family (NF) or not (NN). For example, the dataset 'HECLNF' is composed of heterodimer pairs of chains where each pair has a unique

combination of SCOP families and group clusters and considers the SCOP null family members.

Supplementary Table 3 summaries the measured performance for the distinct combinations of features. The most promising results were further benchmarked in this work.

Supplementary Table 3. Benchmark BMC90 multiple conditions over sequence features.

SET ID	Dataset name	#Comple xes	RAUC	MCC	PPV	TPR	ACC	FPR	SPC	NPV
HECLNF	HEDt	2401	0.78	0.38	0.47	0.49	0.83	0.10	0.90	0.90
HECLNN		875	0.77	0.35	0.43	0.50	0.82	0.12	0.88	0.90
HENCNF		2069	0.79	0.38	0.44	0.53	0.82	0.12	0.88	0.91
HENCNN		533	0.76	0.34	0.44	0.44	0.83	0.10	0.90	0.90
HONCNN	HODt	1981	0.67	0.19	0.26	0.48	0.71	0.24	0.76	0.89
HECLNF+ HONCNN	HEHODt	4382	0.78	0.37	0.45	0.49	0.83	0.11	0.89	0.90
DMNCNN		2359	0.77	0.36	0.44	0.50	0.82	0.12	0.88	0.90
DMCLNF	HEMt	3972	0.783	0.37	0.44	0.51	0.82	0.12	0.88	0.91
DMCLNF+ HONCNN		5953	0.783	0.37	0.46	0.47	0.83	0.10	0.90	0.90
Metrics: RAUC: ROC AUC, all predictions pooled together MCC: Matthews correlation coefficient PPV: Positive Predictive Value TPR: True Positive Rate ACC: Accuracy FPR: False Positive Rate SPC: Specificity NPV: Negative Predictive Value										

2.6. Sequence-structure mode

We evaluated the performance of the sequence-structure mode using as evaluation benchmark Bv5 and we studied how the new mode performed on both the input provided as sequence and the one provided as structure (Supplementary Table 4, Partner type seq and struct respectively). As expected, the performance for the sequence-structure mode (MCC of 0.331 for HEDt and 0.279 for Bv5-leave-one-out) lies between the performance of the model that employs sequences only (MCC of 0.311 for HEDt and 0.307 for Bv5-leave-one-out) and the model that employs structures from the two partners (MCC of 0.403 for HEDt and 0.386 for Bv5-leave-one-out). More interestingly, although we observed that the improvement in overall performance is driven by the enhancement of Residue-Residue Interaction predictions (RRI RAUC of 0.874 vs 0.844 for HEDt against Bv5 or 0.844 vs 0.874 in Bv5-leave-one-out), it is the partner for which structural information is available the one that benefits more from the

improvement, while the sequence-only partner predictions remain comparable to the ones obtained in the sequence-only mode (partner type seq-struct vs seq-seq).

Supplementary Table 4. Sequence-structure mode performance summary.

Trainset	Partner type	RRI	Binding site prediction				
			mRAUC	RAUC	MCC	PPV	TPR
Bv5	mean(seq-struct, struct-seq)	0.851	0.760	0.307	0.363	0.434	0.103
	seq-struct		0.747	0.284	0.350	0.403	0.102
	struct-seq		0.784	0.338	0.368	0.503	0.117
	seq-seq	0.802	0.753	0.279	0.300	0.482	0.113
	struct-struct	0.905	0.823	0.386	0.391	0.559	0.089
HEDt	mean(seq-struct, struct-seq)	0.874	0.773	0.331	0.403	0.426	0.086
	seq-struct		0.758	0.309	0.376	0.419	0.095
	struct-seq		0.790	0.366	0.383	0.543	0.120
	seq-seq	0.844	0.751	0.311	0.380	0.415	0.093
	struct-struct	0.917	0.826	0.403	0.432	0.541	0.097
Notes:							
Partner type refers to which partner is represented as sequence and which is represented as structure. Seq-struct refers to the cases in which the evaluated partner is provided as sequence while the other partner is provided as structure. The opposite applies to struct-seq							
Bv5 230 hetero-multimers (including dimers) HEDt 2401 heterodimers							
Metrics: mRAUC: mean ROC AUC RAUC: ROC AUC, all predictions pooled together MCC: Matthews correlation coefficient PPV: Positive Predictive Value TPR: True Positive Rate FPR: False Positive Rate							

2.7. Comparison to other methods

Whereas the conclusions so far presented in this work referred only to BIPSPI+, our dataset could be used to train other models, possibly improving their performance. As an illustration, we have retrained the SASNet⁶ method using our dataset for hetero-complexes and evaluated

in Bv5 and we have compared the results obtained using our new proposed dataset compared to their proposed dataset DIPs and the original Bv5 leave-one-out. Thus, we measured a median ROC AUC for Residue-Residue pair prediction of 0.896, compared to 0.892 for the DIPs dataset and 0.876 for Dv5-leave-one-out.

Since we detected from the learning curves that even with our larger dataset the SASNet model would be able to benefit from an even larger dataset, we tried a simple data augmentation strategy in which, for each protein complex at training time, we included some additional computational conformation for each of the partners. The virtual conformations were obtained applying Normal Modes Analysis to the bound crystallographic structures. This strategy turned out to be successful, increasing the performance of the SASNet model up to a median ROC AUC of 0.935 when 3 computational poses were included for each experimental one. Unfortunately, BIPSPI+, which relies heavily on sequence-based information and coarse-grained structural information, was not able to benefit from this sort of data augmentation.

Finally, for completeness, we have collected from the literature the reported performance on Bv5 of other partner-specific residue-residue prediction methods published before and after BIPSPI v1. As displayed in Supplementary Table 5, BIPSPI+ outperforms all them.

Supplementary Table 5. Reported median ROC AUC for residue-residue interaction prediction using Bv5 as test set.

Method	Release year	median ROC AUC
PAIRpred ⁸ (struct)	2014	0.863
BIPSPI v1 (seq)	2018	0.827
BIPSPI v1 (struct)	2018	0.937
BIPSPI+ (seq)	2021	0.905
BIPSPI+ (struct)	2021	0.952
SASNet /DIPS ⁶	2018	0.885
(Liu et al., 2020) ⁹	2020	0.908
DIPS-Plus ¹⁰	2021	0.947

References

1. Vreven, T. *et al.* Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2. *Journal of Molecular Biology* **427**, 3031–3041 (2015).
2. Khor, B. Y., Tye, G. J., Lim, T. S. & Choong, Y. S. General overview on structure prediction of twilight-zone proteins. *Theoretical Biology and Medical Modelling* **12**, 1–11 (2015).
3. Sanchez-Garcia, R., Sorzano, C. O. S., Carazo, J. M. & Segura, J. BIPSPI: A method for the prediction of partner-specific protein-protein interfaces. *Bioinformatics* **35**, 470–477 (2019).

4. Savojardo, C., Fariselli, P., Martelli, P. L. & Casadio, R. ISPRED4: Interaction sites PREDiction in protein structures with a refining grammar model. *Bioinformatics* **33**, 1656–1663 (2017).
5. López-Blanco, J. R., Garzón, J. I. & Chacón, P. iMod: Multipurpose normal mode analysis in internal coordinates. *Bioinformatics* **27**, 2843–2850 (2011).
6. Townshend, R. J. L., Bedi, R., Suriana, P. A. & Dror, R. O. End-to-end learning on 3D protein structure for interface prediction. in *Advances in Neural Information Processing Systems* vol. 32 (Neural information processing systems foundation, 2019).
7. Schneidman-Duhovny, D., Inbar, Y., Nussinov, R. & Wolfson, H. J. PatchDock and SymmDock: servers for rigid and symmetric docking. *Nucleic Acids Research* **33**, W363–W367 (2005).
8. Minhas, F. ul A. A., Geiss, B. J. & Ben-Hur, A. PAIRpred: partner-specific prediction of interacting residues from sequence and structure. *Proteins* **82**, 1142–55 (2014).
9. Liu, Y., Yuan, H., Cai, L. & Ji, S. Deep Learning of High-Order Interactions for Protein Interface Prediction. *Proceedings of the ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* **20**, 679–687 (2020).
10. Morehead, A., Chen, C., Sedova, A. & Cheng, J. DIPS-Plus: The Enhanced Database of Interacting Protein Structures for Interface Prediction. (2021).