Towards Fast Binding Affinity Scoring in Protein–Protein Complexes via Interatomic Contacts and Linear Regression

Ana Luísa Araújo Bastos[1,*[0009−0009−4875−1333]], Rafael Pereira Lemos[1,*[0000−0002−5894−2354]], Diego Mariano[1[0000−0002−5899−2052]], Camila A. O. Yamada[2[0000−0001−8248−4640]], Milenna M. Pirovani[1[0000−0002−5060−9418]], and Raquel Cardoso de Melo-Minardi[1[0000−0001−5190−100X]]

[1] Laboratory of Bioinformatics and Systems (LBS), Department of Computer Science, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
[2] Laboratory for Macromolecular Biophysics - LBM, Department of Biochemistry and Immunology, Universidade Federal de Minas Gerais, Belo Horizonte, Brazil
∗ These authors contributed equally to this work.
rafaellemos@ufmg.br, raquelcm@dcc.ufmg.br

Abstract. Protein-protein interactions (PPIs) are essential to understanding how proteins work together to perform most of the molecular processes that underlie life. However, capturing the structural and functional complexity of proteins, as well as their interactions, using computational representations remains an open challenge. In this study, we investigated the individual impact of contact types on binding affinity by constructing linear regression models based on interatomic contacts calculated with the COCαDA tool. A curated dataset comprising 81 protein–protein complexes was employed to systematically analyze ten categories of contacts, including both specific and nonspecific interactions. Our results emphasize the significance of polar-apolar contacts (r=-0.55), evaluated at the atomic level, in reducing the binding strength between protein-protein complexes, which corroborates previous findings in the literature. Additionally, hydrogen bonds demonstrated a notable contribution to binding stability, with a Pearson correlation of r = –0.42. The results presented here are an initial step towards establishing a scoring function based solely on contact calculations.

Keywords: Protein-protein Interactions · Interatomic Contacts · Binding Affinity · Linear Regression.

1 Introduction

Weak interactions are non-covalent forces that stabilize the three-dimensional structure of proteins and mediate interactions with other molecules [1]. Contacts, typically detected using computational methods based on distance (*e.g.*, Euclidean distance) or geometric criteria (*e.g.*, Delaunay and Voronoi tessellations) [2], represent spatial relationships between atoms or residues within or between molecules [3].

It is essential to distinguish between "contacts," which are purely spatial proximities, and "interactions", which involve energetic contributions such as hydrophobic or electrostatic forces [3, 4]. Although not every contact leads to an interaction, they are often prerequisites for biologically meaningful interactions. With this, physicochemical characteristics can be incorporated alongside spatial proximity, reducing the incidence of false positives in computational predictions. Some tools employ coarse-grained approaches, treating interactions at the residue level to minimize computational cost (reviewed in [3, 5]).

Recently, the command-line tool COCαDA (Contact Optimization by Cα Distance Analysis) proposed a new efficient algorithm for calculating contacts at the atomic level [5]. The first version of COCαDA calculates seven types of contacts: Hydrogen and Disulfide Bond (HB and DB, respectively), Hydrophobic (HY), Repulsive (RE), Attractive (AT), Salt Bridge (SB), and Aromatic Stacking (AS). It also classifies contacts in protein structures into two types: INTRA, intramolecular contacts made between residues of the same protein chain, and INTER, intermolecular contacts made between different polypeptide chains.

Intermolecular contacts are crucial for characterizing protein-protein interactions (PPIs), which are fundamental to most biological processes, ranging from the regulation of essential functions such as cell signaling and immune response to DNA replication [6]. These interactions enable proteins and other macromolecules to act in a coordinated manner, forming functional networks that enhance the complexity and specificity of cellular responses.

Due to the significant importance of these kinds of interaction, several studies have sought to better understand and characterize these interactions using computational methods [7]. One example is the PRODIGY (PROtein binDIng enerGY prediction) tool [8, 9], which predicts the binding affinity (BA) between two protein subunits based on their three-dimensional structures, typically obtained from Protein Data Bank (PDB) files. It estimates the binding free energy (ΔG) using models based on simple structural features of the protein-protein interface, such as the number of contacts calculated using a coarse-grained method, the binding energy, and the surface area of the directly interacting proteins.

In this context, the following questions were raised: 1) To what extent can we better understand the individual contribution of specific interatomic contacts to the overall BA of protein–protein complexes? 2) Can the prediction of BA be improved by using more accurate contact calculation methods, such as atomic-level contacts computed with COCαDA?

In this work, we used the interatomic contacts predicted by the COCαDA tool to predict the binding energy between protein-protein chains, along with three new types based on polarity and charge alone. We built a linear regression model using Orange Data Mining [10], in-house Python scripts, and the Scikit-learn library [11]. To evaluate our method, we collected protein complex data from the PDB and compared them with the results obtained in [8]. Our results highlight the importance of attractive contacts and hydrogen bonds in chain interactions, while also emphasizing the role of polar-apolar atom contacts in reducing the attraction force between structures.

2 Methods

2.1 Data Collection

We utilized the same list of 81 'reliable' protein-protein complexes from [8], which is derived from [12]. The data were downloaded from the PDB website in the legacy '.pdb' format, and the complex-forming chains were then filtered according to those described in [12] to avoid spurious contacts from other chain pairs.

2.2 Contact Definition and Calculation

COCαDA [5] uses definitions from [13, 14] for distance-based interatomic contact calculations. Seven contact types are defined natively: HB, DB, HY, RE, AT, SB and AS. We then used an in-house modified version that includes three new types discussed in [8] (Table 1): a) Polar-apolar (PA), where polar atoms are those that perform hydrogen bonds (either donors or acceptors), and apolar atoms are those that perform hydrophobic interactions; b) Positive-apolar (PosA), where positive atoms are those that possess positive charges in neutral pH; and c) Negative-apolar (NegA), where negative atoms are those that possess negative charges in neutral pH. All three types use the 5.5Å distance cutoff previously defined.

Table 1. Summary of types, range and conditions for polar-apolar, positive-apolar and negative-apolar interatomic contacts. D_a = Euclidean distance between the atom pair.

Contact Type	**Range (Å)**	**Condition (other than range)**
Polar-Apolar	$0 \leq D_{\mathrm{a}} \leq 5.5$	Polar + Apolar atoms
Positive-Apolar	$0 \leq D_{\mathrm{a}} \leq 5.5$	Positive + Apolar atoms
Negative-Apolar	$0 \leq D_{\mathrm{a}} \leq 5.5$	Negative + Apolar atoms

2.3 Model construction

Linear regression models were initially built with Orange Data Mining [10]. No regularization parameters were used. The models were trained using a 10-fold cross-validation approach. To generate the scatter plots, the models were trained using each variable as an individual input. The following contact types were used as parameters for each input, regarding their absolute number in the interface:

HB, AT, RE, SB, HY, PA, PosA, and NegA. The input PDB ID and the experimentally obtained ΔG value were also included. Furthermore, the models were trained using different combinations of variables calculated by COCαDA. To validate the parameters used in the model presented here, correlation calculation experiments using inter-residue contacts proposed in the Prodigy paper [8] were reproduced using the parameters trained for the model proposed in this work. The same Pearson correlation was obtained for the combination of all inter-residue contacts (r=0.59).

3 Results and Discussion

We first built a regressor to evaluate the representation capacity of the binding affinity of the contact types calculated by COCαDA. In the future, we intend to develop a scoring function to predict the binding energy between protein-protein complexes. Thus, we evaluated the Pearson correlation using each contact type individually.

Our results indicate that polar-apolar contacts yielded the highest Pearson correlation (r = -0.55). Next, the model built using only hydrogen bonds as input obtained an r = -0.42. The model built only with attractive contacts obtained an r = -0.22, while the model with repulsive contacts also obtained an r = -0.22. The salt bridge model obtained an r = -0.13, and the model with hydrophobic contacts r = -0.25. Finally, the model built with pairs of positively charged and nonpolar atoms obtained an r = -0.28, while the model built with pairs of negatively charged atoms obtained an r = -0.14 (Figure 1).

The study by Vangone and Bonvin [8] considered contacts made at the residue level, obtaining the following results for each pair: charged/charged (r = -0.17), charged/polar (r = -0.26), charged/apolar (r = -0.45), polar/polar (r = -0.13), polar/apolar (r = -0.56), apolar/apolar (r = -0.34), hydrophilic/hydrophilic (r = -0.53), hydrophilic/hydrophobic (r = -0.34), and hydrophobic/hydrophilic (r = -0.05). It is important to emphasize that, considering only the impact of each variable, our results are comparable to the results obtained in [8]. However, while they combined multiple attributes related to contacts, interface characteristics, and free energy to achieve a higher overall correlation (r = -0.73), the focus of the present work is distinct. Rather than attempting to surpass the performance reported in their work, our primary goal is to evaluate the predictive potential of atomic-level contact calculations. The results presented here represent an initial step toward the development of a scoring function based solely on detailed interatomic contact analysis.

3.1 Evaluation of the Contribution of Individual Interatomic Contact Types on BA Values

In their study, Vangone and Bonvin [8] considered charges regardless of their sign, meaning that both positively and negatively charged residues were treated equally. To further detail our analysis, we decided to split the “charged-apolar"

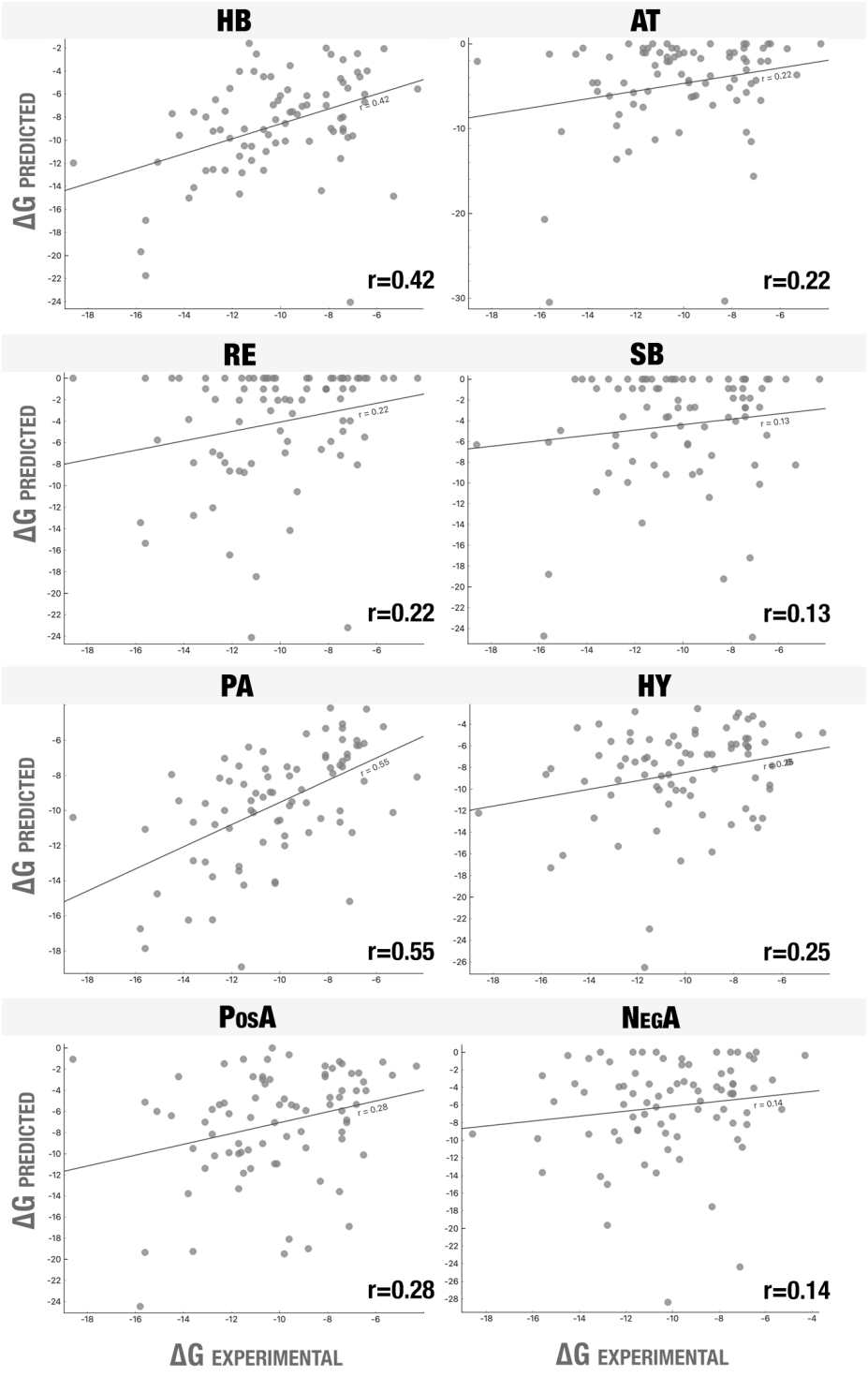

Fig. 1. Scatter plot of predicted vs. experimental binding affinities for each individual contact type. HB: Hydrogen Bond; AT: Attractive; RE: Repulsive; SB: Salt Bridge; HY: Hydrophobic; PA: Polar-Apolar; PosA: Positively Charged-Polar; and NegA: Negatively Charged-Polar. Pearson correlation (r) values are shown on the bottom right, and binding affinities are reported as absolute values in kcal/mol^{-1}.

contacts into positively charged-PosA and negatively charged-NegA. With this, we analyzed 10 different contact types, divided into two groups: specific contacts and nonspecific contacts. Specific contacts are those that pose significant energetic contributions, and are natively calculated by COCαDA: HB, DB, AT, RE, AS, SB, and HY. Nonspecific contacts, on the other hand, typically make no significant energetic contributions or effects, such as PA, PosA, and NegA.

No DB contacts were identified in the 81 complexes present in the dataset, and only 13 AS contacts were observed, so they were removed from further analysis. We then compared the eight remaining contact types according to their individual correlation with the experimental data (Figure 1).

PA, PosA and NegA are not, in a strict sense, *bona-fide* interatomic interactions. Even when atoms are in close proximity (which defines a contact), no significant energetic contributions or effects take place. Rather, they arise primarily from dense steric packing and surface shape complementarity. In the context of PPIs, such contacts are largely governed by principles of maximal atomic close-packing, as previously described [15, 16].

Consequently, these contacts are best classified as "nonspecific contacts" or "crystal-packing interactions," as they typically lack functional roles and are likely subject to negative selective pressures [17, 18]. Nonetheless, despite their lack of direct energetic contribution, these contacts have been demonstrated to serve as important structural markers of BA. Their distribution and frequency can correlate well with the experimental free energy measurements of protein complexes, reinforcing their relevance in computational and structural analyses of PPIs [8, 9, 19].

When viewing the surface of protein-protein complexes, we can see how specific and nonspecific interatomic contacts take part on their interacting portions. In an example using PDB ID 1E6E (Figure 2), the chain A (receptor protein) forms an extensive interacting patch with chain B (ligand protein), containing several contacts pertaining to the eight different types analyzed, resulting in an experimental ΔG value of -8.3 kcal/mol^{-1}.

To provide an in-depth view of how specific and nonspecific contacts manifest at protein-protein interfaces, Figure 3 showcases examples extracted from the 1E6E protein–protein complex. In A), we can see the five specific contact types: AT, SB, RE, HB and HY; in B), nonspecific contacts are represented: NegA, PA and PosA. Details for all contacts are shown in Table 2.

The presence of hydrophobic (apolar) patches within protein interfaces further contributes to complex stability. This stabilization effect arises from the burial of hydrophobic surfaces away from the aqueous environment, a principle fundamental to PPI energetics. However, this hydrophobic stabilization must be carefully balanced to avoid compromising the stability of the individual, unbound proteins [19]. The introduction of polar atoms into these hydrophobic patches can disrupt compact packing and diminish interface stability, a phenomenon critical to the modulation of BAs [20]. Notably, PA contacts—which reflect the presence of polar atoms—have been found to correlate inversely with the BA [8]. Their occurrence appears largely incidental, a byproduct of steric

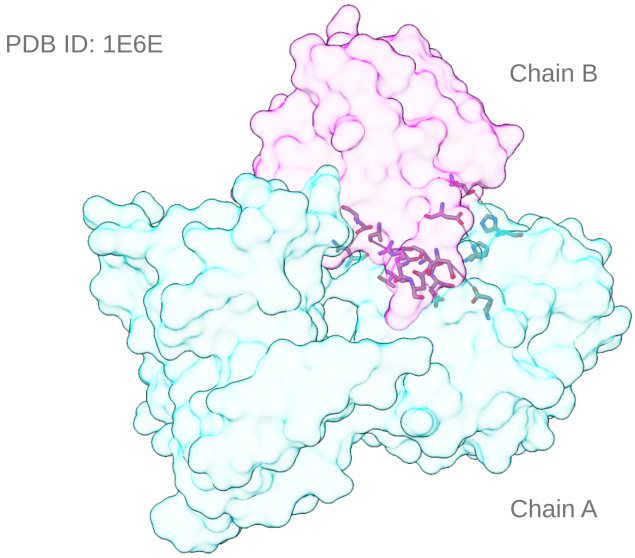

Fig. 2. Surface representation of the interface region with examples of specific and nonspecific interatomic contacts. The receptor (chain A) and the ligand (chain B) of the PBD ID 1E6E are shown as surfaces in cyan and magenta, respectively. The interacting residues are represented as sticks.

Table 2. Details of specific and nonspecific interatomic contacts identified in the interfacing region of the PDB ID 1E6E protein-protein complex. Atom nomenclature is Chain:Residue-Atom, and atom names follow the ones used in the PDB.

	Type	Atom 1	Atom 2	Distance (Å)
Specific	AT	A:28H-ND1	B:39D-OD2	5.27
	SB	A:24H-ND1	B:41D-OD2	3.84
	RE	A:382D-OD1	B:116E-OE2	5.55
	HB	A:378T-OG1	B:115R-NE	3.82
	HY	A:60N-CB	B:47E-CG	4.22
Nonspecific	NegA	A:353E-OE1	B:113D-CB	5.15
	PA	A:64T-CG2	B:45A-O	4.44
	PosA	A:381T-CG2	B:115R-CZ	4.50

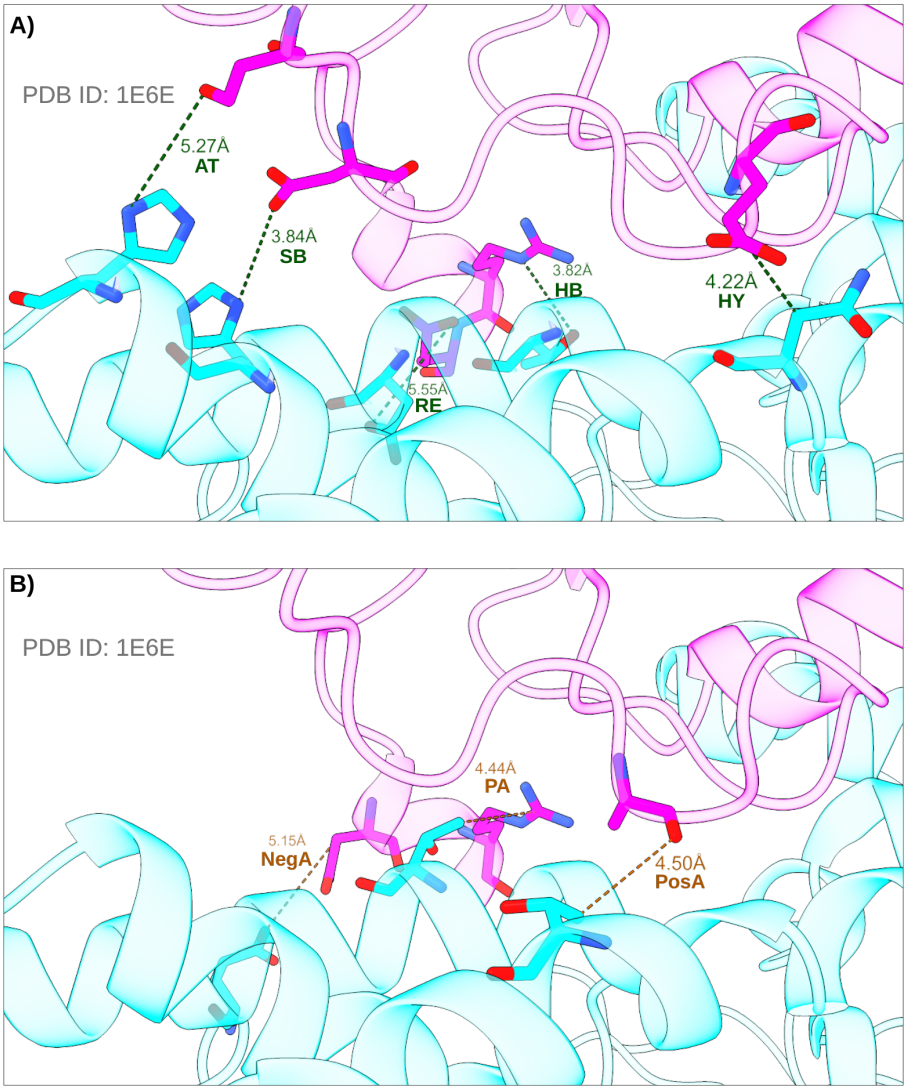

Fig. 3. Examples of interatomic contacts identified in the interfacing region of a protein-protein complex. For both figures, the same region of PDB ID 1E6E is shown. The receptor (chain A) and the ligand (chain B) are shown as cartoons in cyan and magenta, respectively. A) Specific contacts. AT: attractive contact between A:28H-ND1 and B:39D-OD2; SB: salt bridge between A:24H-ND1 and B:41D-OD2; RE: repulsive contact between A:382D-OD1 and B:116E-OE2; HB: hydrogen bond between A:378T-OG1 and B:115R-NE; HY: hydrophobic contact between A:60N-CB and B:47E-CG. B) Nonspecific contacts. NegA: negative-apolar contact between A:353E-OE1 and B:113D-CB; PA: positive-apolar contact between A:64T-CG2 and B:45A-O; PosA: positive-apolar contact between A:381T-CG2 and B:115R-CZ. Nomenclature is Chain:Residue-Atom, and atom names follow the ones used in the PDB. Distances between the atom pair, in angstroms, are highlighted for each contact.

surface packing rather than an optimized energetic feature, which may explain the evolutionary pressure to minimize such contacts.

Despite the general destabilizing role of polar contacts within hydrophobic interfaces, specific polar interactions, particularly hydrogen bonds (which are formed by polar atoms), are frequently observed even in tightly packed protein-protein interfaces [21, 22]. Larsen and colleagues [23], in a study of 137 homodimeric protein complexes, demonstrated that most of the interfaces contained numerous small hydrophobic patches interspersed with polar interactions and water molecules. This pattern underscores a complex architectural strategy: the overall interface maximizes hydrophobic stabilization while maintaining partial hydration, relying on a combination of localized hydrophobic contacts and hydrogen bonds [23, 24]. This circumvents the need for complete desolvation, which is energetically costly, while preserving the structural integrity and flexibility of the proteins.

Such findings suggest that, while nonspecific contacts dominate the gross features of PPI interfaces, specific polar interactions are crucial in refining and stabilizing the binding landscape at a finer scale. Besides, structural and mutational analyses have revealed that networks of polar and charged residues — including hydrogen bonds and salt bridges — can contribute significantly to the stability of protein complexes, partially compensating for the energetic cost of desolvation [25].

4 Conclusions

In this study, we developed and evaluated a linear regression model based on interatomic contact types to predict the binding affinity of protein–protein complexes. Using the COCαDA tool, we systematically calculated a wide variety of contact types, including five types of specific contacts: hydrogen bonds, hydrophobic, attractive, repulsive, and salt bridge interactions; and three nonspecific types: polar–apolar, negative-apolar, and positive-apolar.

Our results showed that polar–apolar contacts exhibited the strongest individual correlation with experimental binding affinities, followed by hydrogen bonds. These findings suggest that not only energetically significant interactions but also steric and surface packing effects play key roles in determining the stability of protein–protein interfaces.

We adopted linear regression for its interpretability and computational simplicity. However, we acknowledge that this approach assumes linearity and independence between variables, which may not fully capture the underlying complexity of protein–protein interactions. As a next step, we plan to explore machine learning approaches such as Random Forest and Neural Networks to better capture complex relationships and improve predictive performance.

In addition, future efforts will focus on incorporating additional structural features such as Buried Surface Area (BSA) and Non-Interacting Surface (NIS), which can provide valuable contributions. We also plan to expand the dataset to increase the robustness and generalizability of the model.

Overall, although this work represents an initial step, it provides new insights into the specific and nonspecific contributions to protein–protein binding and establishes a basis for the development of faster and more accurate scoring functions based on atomic-level contact analysis.

Acknowledgments. The authors thank the agencies: Coordenação de Aperfeiçoamento de Pessoal de Nível Superior—CAPES, Brazil; Fundação de Amparo à Pesquisa do Estado de Minas Gerais—FAPEMIG; and Conselho Nacional de Pesquisa e Desenvolvimento Científico e Tecnológico—CNPq. This study was financed in part by the Coordenação de Aperfeiçoamento de Pessoal de Nível Superior - Brasil (CAPES) - Finance Code 001.

Disclosure of Interests. The authors have no competing interests to declare that are relevant to the content of this article.

Data Availability Statement. The dataset used in this study was obtained in the Prodigy web tool, available at https://rascar.science.uu.nl/prodigy/dataset. PDB files were downloaded from the RCSB PDB website, available at https://www.rcsb.org. Supplementary files, containing the Orange session, input and output tables, are available at https://github.com/LBS-UFMG/binding_affinity_score_2025.

References

1. Khatri, B., Majumder, P., Nagesh, J., Penmatsa, A., Chatterjee, J.: Increasing protein stability by engineering the n $\rightarrow \pi^*$ interaction at the β-turn. Chem. Sci. (2020)
2. Martins, P.M., Mayrink, V.D., de A. Silveira, S., da Silveira, C.H., de Lima, L.H.F., de Melo-Minardi, R.C.: How to compute protein residue contacts more accurately? In: Proceedings of the 33rd Annual ACM Symposium on Applied Computing. ACM, New York, NY, USA (Apr 2018)
3. da Silveira, C.H., Pires, D.E.V., Minardi, R.C., Ribeiro, C., Veloso, C.J.M., Lopes, J.C.D., Meira, Jr, W., Neshich, G., Ramos, C.H.I., Habesch, R., Santoro, M.M.: Protein cutoff scanning: A comparative analysis of cutoff dependent and cutoff free methods for prospecting contacts in proteins. Proteins **74**(3), 727–743 (Feb 2009)
4. Godzik, A., Kolinski, A., Skolnick, J.: Topology fingerprint approach to the inverse protein folding problem. J. Mol. Biol. **227**(1), 227–238 (Sep 1992)
5. Lemos, R.P., Mariano, D., Silveira, S.A., de Melo-Minardi, R.C.: COCαDA - large-scale protein interatomic contact cutoff optimization by Cα distance matrices. In: Annals of the XVII Brazilian Symposium on Bioinformatics (BSB 2024). pp. 59–70. Brazilian Computer Society (SBC) (Dec 2024)
6. Greenblatt, J.F., Alberts, B.M., Krogan, N.J.: Discovery and significance of protein-protein interactions in health and disease. Cell **187**(23), 6501–6517 (Nov 2024)
7. Keskin, O., Tuncbag, N., Gursoy, A.: Predicting protein-protein interactions from the molecular to the proteome level. Chem. Rev. **116**(8), 4884–4909 (Apr 2016)
8. Vangone, A., Bonvin, A.M.: Contacts-based prediction of binding affinity in protein-protein complexes. Elife **4**, e07454 (Jul 2015)

9. Xue, L.C., Rodrigues, J.P., Kastritis, P.L., Bonvin, A.M., Vangone, A.: PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. Bioinformatics **32**(23), 3676–3678 (Dec 2016)
10. Demšar, J., Zupan, B., Leban, G., Curk, T.: Orange: From experimental machine learning to interactive data mining. In: Lecture Notes in Computer Science, pp. 537–539. Lecture notes in computer science, Springer Berlin Heidelberg, Berlin, Heidelberg (2004)
11. Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., Duchesnay, E.: Scikit-learn: Machine learning in Python. Journal of Machine Learning Research **12**, 2825–2830 (2011)
12. Kastritis, P.L., Moal, I.H., Hwang, H., Weng, Z., Bates, P.A., Bonvin, A.M.J.J., Janin, J.: A structure-based benchmark for protein-protein binding affinity. Protein Sci. **20**(3), 482–491 (Mar 2011)
13. Sobolev, V., Sorokine, A., Prilusky, J., Abola, E.E., Edelman, M.: Automated analysis of interatomic contacts in proteins. Bioinformatics **15**(4), 327–332 (Apr 1999)
14. Fassio, A.V., Santos, L.H., Silveira, S.A., Ferreira, R.S., de Melo-Minardi, R.C.: napoli: A graph-based strategy to detect and visualize conserved protein-ligand interactions in large-scale. IEEE/ACM Trans. Comp. Biol. Bioinf. **17**(4), 1317–1328 (Jul 2020). https://doi.org/10.1109/tcbb.2019.2892099, http://dx.doi.org/10.1109/TCBB.2019.2892099
15. Chothia, C., Janin, J.: Principles of protein-protein recognition. Nature **256**(5520), 705–708 (Aug 1975)
16. Norel, R., Lin, S.L., Wolfson, H.J., Nussinov, R.: Shape complementarity at protein-protein interfaces. Biopolymers **34**(7), 933–940 (Jul 1994)
17. Janin, J.: Specific versus non-specific contacts in protein crystals. Nat. Struct. Biol. **4**(12), 973–974 (Dec 1997)
18. Prasad Bahadur, R., Chakrabarti, P., Rodier, F., Janin, J.: A dissection of specific and non-specific protein–protein interfaces. J. Mol. Biol. **336**(4), 943–955 (Feb 2004)
19. Tsai, C.J., Xu, D., Nussinov, R.: Protein folding via binding and vice versa. Fold. Des. **3**(4), R71–80 (1998)
20. Tsai, C.J., Nussinov, R.: Hydrophobic folding units at protein-protein interfaces: implications to protein folding and to protein-protein association. Protein Sci. **6**(7), 1426–1437 (Jul 1997)
21. Xu, D., Tsai, C.J., Nussinov, R.: Hydrogen bonds and salt bridges across protein-protein interfaces. Protein Eng. Des. Sel. **10**(9), 999–1012 (Sep 1997)
22. Tam, J.Z., Palumbo, T., Miwa, J.M., Chen, B.Y.: Analysis of protein-protein interactions for intermolecular bond prediction. Molecules **27**(19), 6178 (Sep 2022)
23. Larsen, T.A., Olson, A.J., Goodsell, D.S.: Morphology of protein-protein interfaces. Structure **6**(4), 421–427 (Apr 1998)
24. Rego, N.B., Xi, E., Patel, A.J.: Identifying hydrophobic protein patches to inform protein interaction interfaces. Proc. Natl. Acad. Sci. U. S. A. **118**(6), e2018234118 (Feb 2021)
25. Sheinerman, F.B., Norel, R., Honig, B.: Electrostatic aspects of protein-protein interactions. Curr. Opin. Struct. Biol. **10**(2), 153–159 (Apr 2000)