PLAS-20k: Extended Dataset of Protein-Ligand Affinities from MD Simulations for Machine Learning Applications

Divya B. Korlepara[1,3,†], C. S. Vasavi[1,†], Rakesh Srivastava[2], Pradeep Kumar Pal[2], Saalim H. Raza[1], Vishal Kumar[2], Shivam Pandit[1], Aathira G Nair[1], Sanjana Pandey[1], Shubham Sharma[1], Shruti Jeurkar[2], Kavita Thakran[1], Reena Jaglan[1], Shivangi Verma[1], Indhu Ramachandran[1], Prathit Chatterjee[1], Divya Nayar[4,*], and U. Deva Priyakumar[2,*]

[1]IHub-Data, International Institute of Information Technology, Hyderabad, 500032, India.
[2]Centre for Computational Natural Sciences and Bioinformatics, International Institute of Information Technology, Hyderabad, 500032, India.
[3]Divison of Physics, School of Advanced Sciences, Vellore Institute of Technology, Chennai, 600127, India.
[4]Department of Materials Science and Engineering, Indian Institute of Technology Delhi, Hauz Khas, New Delhi, 110016, India.
[†]These authors contributed equally to this work: Divya B. Korlepara, C.S. Vasavi.
[*]corresponding author(s): Divya Nayar, U. Deva Priyakumar (divyanayar@mse.iitd.ac.in, deva@iiit.ac.in)

ABSTRACT

Computing binding affinities is of great importance in drug discovery pipeline and its prediction using advanced machine learning methods still remains a major challenge as the existing datasets and models do not consider the dynamic features of protein-ligand interactions. To this end, we have developed PLAS-20k dataset, an extension of previously developed PLAS-5k, with 97,500 independent simulations on a total of 19,500 different protein-ligand complexes. Our results show good correlation with the available experimental values, performing better than docking scores. This holds true even for a subset of ligands that follows Lipinski's rule, and for diverse clusters of complex structures, thereby highlighting the importance of PLAS-20k dataset in developing new ML models. Along with this, our dataset is also beneficial in classifying strong and weak binders compared to docking. Further, OnionNet model has been retrained on PLAS-20k dataset and is provided as a baseline for the prediction of binding affinities. We believe that large-scale MD-based datasets along with trajectories will form new synergy, paving the way for accelerating drug discovery.

Background & Summary

High-throughput screening plays a crucial role in the drug discovery process. However, this approach to identifying lead molecules is time-consuming and labour-intensive. On the other hand, computational methods offer a promising solution by significantly reducing the cost, time, and resources required for physical experiments in screening potential hit molecules. High-throughput docking and molecular dynamics (MD) simulations provide an appealing virtual screening approach to expedite the discovery of biologically active hit compounds[1]. Despite the advantages of these methods, certain limitations and drawbacks still exist in docking. These include a restricted sampling of both protein and ligand conformation during pose prediction and the use of approximated scoring functions that often yield docking scores with poor correlation to experimental binding affinities[2]. On the other hand, MD simulations offer several benefits for investigating the structural and dynamical properties of a Protein-Ligand (PL) system and accurately predicting binding affinities. However, screening of umpteen molecules consumes prohibitively expensive computational resources rendering the prediction of binding affinity (MD based) on a large scale infeasible[3].

In recent years, machine learning (ML) has emerged as a powerful tool to accelerate various aspects of drug development[4]. ML has already shown to be successful in the hunt for antibiotics,[5] drug re-purposing for emerging diseases[6,7], virtual screening[8,9], bio-molecular interactions, prediction of binding site and protein folding[10–14]. Notably, enormous ML models have been developed to predict PL binding affinity[15]. These data-driven approaches have been successful in attaining a high level of accuracy by learning the binding modes directly from rapidly growing experimental three-dimensional (3D) PL structural data deposited in Protein Data Bank (PDB)[16,17]. Numerous attempts have been made to enhance the performance of machine learning (ML) models through different types of encoding, topology, spectral sequence, and atom pairs. These

approaches have predominantly relied on feature engineering from static 3D structures[18]. However, this static picture of PL interactions often lacks dynamic features. Incorporating dynamic properties can provide crucial insights into bio-molecular processes such as protein folding, conformational changes, and ligand binding. In addition, considering dynamic features can help address fundamental questions related to binding affinity and specificity[19,20]. The greatest strength of MD simulations lies in their ability to reveal dynamic effects of the bio-molecules that go beyond the experimentally determined structures available in PDB[21,22]. Furthermore, MD simulations capture the interactions and energy exchanges between the protein, ligand (solute), and solvent(water, buffer ions) to dictate the binding event through both long-range and short-range interactions[23–26]. While existing ML models have shown promise in predicting binding affinity, they often rely on training datasets composed of only a few hundred static binding poses of PL complexes. With the continuous growth in the number of ligands and proteins, there is an increasing demand for massive and dynamic data to improve the ML model's accuracy in predicting binding affinities.

By integrating MD simulations with ML techniques, researchers can leverage the dynamic nature of biomolecular systems and incorporate a broader range of data, leading to more accurate and reliable predictions of binding affinities. The combination of MD simulations and ML holds great potential for accelerating drug discovery efforts in an ever-expanding chemical space. To this end, in our previous work, we developed a MD-based dataset called PLAS-5k[27]. This dataset included binding affinities averaged over conformations of each of 5000 PL complexes, representing various classes of enzymes. In addition to the binding affinities, the dataset also included energy components contributing to the binding free energy.

When attempting to accurate prediction of PL interactions through ML models, a labyrinth of interactions needs to be accounted for. In continuation of our previous dataset, the current work focuses on expanding heterogeneous proteins and a large spectrum of ligand types, including small organic molecules and peptides. The extended dataset, encompasses 19,500 PL structures, providing protein-ligand affinities and non-covalent interaction components, along with accompanying trajectories suitable for machine learning applications.

The creation of the PLAS dataset was primarily motivated by the need for high-quality datasets that can support the development of advanced algorithms and drive significant advancements in drug development. The PLAS-20k dataset comprises a diverse collection of protein-ligand (PL) complexes, providing a valuable resource for researchers in the field. To assess the performance of calculated binding affinities, we conducted comparisons by calculating correlation coefficients between experimentally determined values and the affinities obtained through molecular mechanics/Poisson-Boltzmann surface area (MM-PBSA) and docking methods. This evaluation allowed us to validate the accuracy and reliability of the computational approaches employed. Based on the experimental binding affinities within the PLAS-20k dataset, we categorized the complexes into strong binders (SB) and weak binders (WB). This classification helps to differentiate between PL complexes with high and low affinities, providing valuable insights into the range of binding strengths within the dataset. Furthermore, we assessed the ligand's adherence to Lipinski's Rule of 5, which offers insights into their drug-like properties. As a baseline for comparison, we retrained the OnionNet framework using our dataset. The availability of large datasets is often considered essential for successful deep learning applications. Thus, we believe that the PLAS-20k dataset will serve as a catalyst for the development of data-driven methods in various drug design tasks, including hit identification, lead optimization, and de novo molecular design. By providing a comprehensive and diverse dataset, the PLAS-20k dataset empowers researchers to more effectively explore and apply data-driven approaches, leading to advancements in drug discovery and design processes. The dataset's availability will drive further innovation and contribute to significant progress in the field of drug development.

Methods

Data Curation

In this article, we have chosen a set of 14,500 complexes from the Protein Data Bank (PDB)[17], expanding upon our previous PLAS-5k[27] dataset. The selection criteria for these complexes focused on proteins that are complex with small molecules (ligands) or peptides.

Dataset Preparation

In this study, we followed the preprocessing and calculation protocol similar to our previous work[27]. A brief account of the methods is given here. The initial structure of the complexes was taken from PDB[17]. Protein chains with missing residues were modelled as loop regions using UCSF Chimera[?,28]. Further, the protein chains were protonated at a physiological pH, 7.4 using H++ server[29]. The tleap program of ambertools[30,31] was used to build the input files of each complex system (protein-ligand, cofactors and crystal water molecules) files required for MD simulations. The crystal waters were modelled using a TIP3P force field[32] The proteins were modelled using Amber ff14SB force field[33] in the all-atom model, and parameters of the ligand and cofactors were taken from General AMBER force field (GAFF2)[34] using antechamber program[35]. Each complex was solvated in an orthorhombic TIP3P water box with a 10 Å extension from the protein surface. More detailed information on the dataset preparation is discussed in our earlier work with 5k complexes[27] and the flowchart for data preparation is shown in Figure 1. The counter ions were added to maintain the charge neutrality of the system.

MD simulations were performed using OpenMM 7.2.0 program[36]. The simulation protocol involved several steps as described below. To initiate the simulations, we performed a minimization process using the L-BFGS minimizer with a harmonic potential applied to the atoms of the protein backbone. The force constant for this potential was set to 10 kcal/mol/Å^2. The minimization consisted of 1000 steps, and after every 10 steps, the restraint force on the backbone atoms was reduced by half. Subsequently, an additional 1000 steps of minimization were conducted after removing the harmonic potential entirely.

During the simulation, a time step of 2 fs was used, and constraints were applied to the bonds involving hydrogen atoms. We implemented a Langevin thermostat with a friction coefficient of 5 ps^{-1} to maintain the temperature. The system was gradually heated from an initial temperature of 50 K to the target temperature of 300 K, increasing by 1 K every 100 steps (200 fs). The backbone atoms of the protein were restrained using harmonic potentials during this heating process. Once the target temperature was reached, the simulations were performed for 1 ns in the NVT ensemble.

In the next step, the systems were equilibrated in NPT ensemble at 300 K and 1 atm using a Langevin thermostat and Monte Carlo barostat for 2 ns. Finally, a production run of 4 ns in NPT ensemble is performed and the trajectory is saved every 100 ps for post-processing analysis. The final coordinates of the systems were subjected to minimization for 4000 steps. The coordinates at every 1000 steps were saved and used as the initial structures to start the four more independent simulations.

MD trajectories from five independent simulations were used to calculate the binding affinity using MMPBSA (Molecular-Mechanics Poisson Boltzmann Surface Area) method. Here we used a single trajectory approach to estimate the contribution of the complex, ligand, and receptors separately. We considered two explicit water molecules near the active site. The binding affinity is calculated as follows:

$$\Delta G_{MM-PBSA} = \Delta E_{MM} + \Delta G_{Sol} \tag{1}$$

Electrostatic interaction energy ΔE_{ele}, and Van der Waals interaction energy ΔE_{vdw} contributs to ΔE_{MM} (equation (2)) and ΔG_{Sol}, is defined as sum of polar ΔG_{pol}, and non-polar contributions ΔG_{np} (equation (3))

$$\Delta E_{MM} = \Delta E_{ele} + \Delta E_{vdw} \tag{2}$$

$$\Delta G_{Sol} = \Delta G_{pol} + \Delta G_{np} \tag{3}$$

Data Records

The PLAS-20k dataset is available publicly and can be accessed at (https://healthcare.iiit.ac.in/d4/plas20k/plas20k.html). The list of PDB ids that are part of PLAS-20k is provided and can be downloaded from the website. The PDB id search icon in the database opens a specific 3D structure along with energy components (Van der Waals interaction energy, electrostatic energy, polar and non-polar solvation free energies in conjunction with binding affinity) from the MD trajectories using the MM-PBSA method. An example of HIV-1 protease complex (PDB id: 1hxw) is shown in Supplementary Figure S1. The binding affinity and energy components for all the complexes can be accessed through `https://figshare.com/s/05a562608b47d1682b8f` in csv format.

Technical Validation

Overall Structures of the Protein-Ligand Complexes

Though there are a lot of advances in predicting PL binding affinity through machine learning methods, the incorporation of receptor flexibility remains a major bottleneck. In the present work, we propose a novel dataset based on binding affinities of PL complexes retrieved from MD simulations. The binding affinities were calculated by considering the flexibility of both protein and ligand. The simulated complexes were validated by calculating the RMSD with respect to the experimental structure. The protein structures were superimposed to calculate RMSDs of protein and ligand. These calculations have been performed over 200 frames (40 from each simulation trajectory) and the corresponding distributions are shown in Supplementary Figure S2. The long tails of RMSD distributions of protein and ligand are evident due to the flexibility of the complex during the simulations.

Comparison of experimental vs computed binding affinities

Experimentally, the binding affinity of a protein-ligand complex is expressed in terms of dissociation constant (K_d) or inhibition constant (K_i). This experimentally determined binding equilibrium constant is related to binding free energy as,

$$\Delta G_{expt} = -k_B T ln K_i = -k_B T ln(1/K_d) \tag{4}$$

In this work, for a comparison study, we selected a subset of 6842 complexes of the PLAS-20k dataset, whose experimental binding affinities are available. To assess the performance of our dataset, the Pearson correlation coefficient (R_p) and Spearman rank correlation coefficient (R_s) were calculated. Both these correlation coefficients showed that, studies based on MM-PBSA have superior performance with (R_p) of 0.50 and (R_s) of 0.56 compared to docking studies whose (R_p)&(R_s) are 0.39 and 0.41 respectively. The corresponding plots are shown in Figure 2. The results highlight the importance of considering both protein and ligand flexibility. We expect that ML-based scoring functions developed using the PLAS-20k dataset could be more reliable than classical scoring functions. The distribution of the calculated binding affinity is shown in Supplementary Figure S3.

Classification of Binders

Drug discovery is the process by which lead molecules are identified by screening chemical space based on binding affinity. The existing ML models or scoring functions were formulated based on several assumptions but they still have certain limitations. Mostly, researchers are interested in identifying only strong binders (SB), and one of the major reasons for neglecting weak binding molecules in drug discovery is because of its cross reactivity[37,38]. However, these weak binders (WB) are also equally important as they play a key role in fragment-based drug design[39] and they serve as a foundation towards the development of more potent and selective drug candidates with improved therapeutic efficacy.

In our dataset, 4343 PL complexes with experimental $K_{i/d}$ fall into SB and WB categories. This subset is used to classify SB and WB based on experimental vs MMPBSA and experimental vs docking binding affinities. For experimental binding affinities, the strong and weak binders were classified with a predefined cut-off value of -8.18 kcal/mol. The corresponding MMPBSA and docking cut-offs are -38.70 kcal/mol and -6.35 kcal/mol respectively. A brief discussion of the binding affinity cutoff values is given in detail in Supplementary Information.

The classification based on MMPBSA and Docking is shown in Figure 3 and the qualitative performance was evaluated using the metrics given in Tables 1-2. In Figure 3, the diagonal elements of the confusion matrix represent the number of correct predictions, while the off-diagonal elements represent incorrect predictions. Based on the evaluation metrics, given in Tables 1-2 and correlation coefficients (Supplementary Figure S4) it can be observed that MMPBSA classification is performing better compared to docking scores. Also, the confusion matrix revealed that the majority of SB (true positives) and WB (true negatives) were correctly identified with respect to MMPBSA, indicating the dataset is good enough to distinguish SB and WB. The definitions of the evaluation metrics are provided in SI.

Performance of Diverse Protein Sequences

The central goal of any machine learning (ML) model is to get the best model, and its performance depends on training data. More diverse the training data, one can expect a better model. We have collected a humongous number of complex structures for this dataset preparation. Our dataset covers 1856 protein families which are of functional significance and a pie chart of the highly populated family is shown in supplementary Figure S5. Proteins with sequence similarity of $\leq 40\%$ are grouped and the correlation coefficients are shown in Supplementary Figure S6. The results highlight the importance of the PLAS-20k dataset as it shows a good correlation for a diverse set of proteins.

Performance Based on Ligand Structural Properties

In the field of drug discovery, prediction of bio-active molecules are based on several rules such as Lipinski,[40] MDDR-like rule,[41] Veber rule,[42] and Ghose filter[43]. The physicochemical properties like molecular weight and hydrogen bonding capacity are important to design drug-like molecules. For a comparison study, we chose a set of ligands with drug-like properties (Molecular weight ≤ 500, number of hydrogen bond donors ≤ 5, number of hydrogen bond acceptors ≤ 10) and evaluated the performance of those complexes based on docking and MMPBSA calculations.

As seen in Figure 4, MMPBSA calculations showed good correlation with (R_p) of 0.55 and (R_p) of 0.57 compared to docking with (R_p),(R_s) 0.41 and 0.43 respectively. Also, for each of the individual components of drug-like properties, MMPBSA showed a good correlation compared to docking and the results are shown in Supplementary Figure S7-S9. Further, as seen in Supplementary Figure S10 our dataset holds diverse ligands highlighting a few molecular descriptors, as they play an important role in drug discovery.

Components of the Binding Free Energies

Binding free energy is the most important initial indicator of drug potency and remains a major challenge in predicting affinities. In this work, we have provided binding energies for 19,500 PL complexes along with energy components (ΔE_{ele}, ΔE_{vdw}, and ΔG_{Sol}). This PLAS-20k dataset could be helpful in training ML models for predicting the binding affinities and energy components. The knowledge of these components can help in lead optimization. The distribution of the energy components is shown in Supplementary Figure S11. Moreover, the availability of dynamic binding poses from the PLAS-20k dataset can help in building ML models that can screen lead compounds in a more efficient manner compared to existing methods.

Machine Learning Baseline

PLAS-20k data was also trained and tested using a deep Convolutional Neural Network (CNN) based model, OnionNet. As ML and deep learning methods have begun to make significant contributions in predicting the binding affinity of a PL complex. The OnionNet model extracts various features from the 3D molecular structure of each PL complex and corresponding binding affinities as input, it then predicts the binding affinity of unknown complexes using deep CNN. The model trained on PLAS-20k data gave an R_p of 0.91 with an RMSE of 8.15 kcal/mol as shown in Figure 5. This further shows that the PLAS-20k dataset can be used effectively for training various ML and deep learning models.

Code availability

There is no in-house code used for ML model. We used OnionNet[44] http://github.com/zhenglz/onionnet/ ML model to train on PLAS-20k dataset.

References

1. Shim, H., Kim, H., Allen, J. E. & Wulff, H. Pose classification using three-dimensional atomic structure-based neural networks applied to ion channel–ligand docking. *J. Chem. Inf. Model.* **62**, 2301–2315 (2022).
2. Gilson, M. K. & Zhou, H.-X. Calculation of protein-ligand binding affinities. *Annu. review biophysics biomolecular structure* **36**, 21–42 (2007).
3. Osaki, K., Ekimoto, T., Yamane, T. & Ikeguchi, M. 3d-rism-ai: A machine learning approach to predict protein–ligand binding affinity using 3d-rism. *The J. Phys. Chem. B* **126**, 6148–6158 (2022).
4. Karthikeyan, A. & Priyakumar, U. D. Artificial intelligence: machine learning for chemical sciences. *J. Chem. Sci.* **134**, 1–20 (2022).
5. Stokes, J. M. *et al.* A deep learning approach to antibiotic discovery. *Cell* **180**, 688–702 (2020).
6. Morselli Gysi, D. *et al.* Network medicine framework for identifying drug-repurposing opportunities for covid-19. *Proc. Natl. Acad. Sci.* **118**, e2025581118 (2021).
7. Choudhury, C., Murugan, N. A. & Priyakumar, U. D. Structure-based drug repurposing: Traditional and advanced ai/ml-aided methods. *Drug Discov. Today* (2022).
8. Goel, M., Aggarwal, R., Sridharan, B., Pal, P. K. & Priyakumar, U. D. Efficient and enhanced sampling of drug-like chemical space for virtual screening and molecular design using modern machine learning methods. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **13**, e1637 (2023).
9. Mehta, S., Goel, M. & Priyakumar, U. D. Mo-memes: A method for accelerating virtual screening using multi-objective bayesian optimization. *Front. Medicine* **9** (2022).
10. Chelur, V. R. & Priyakumar, U. D. Birds-binding residue detection from protein sequences using deep resnets. *J. Chem. Inf. Model.* **62**, 1809–1818 (2022).
11. Aggarwal, R., Gupta, A., Chelur, V., Jawahar, C. & Priyakumar, U. D. Deeppocket: ligand binding site detection and segmentation using 3d convolutional neural networks. *J. Chem. Inf. Model.* **62**, 5069–5079 (2021).
12. Huang, K., Xiao, C., Glass, L. M., Zitnik, M. & Sun, J. Skipgnn: predicting molecular interactions with skip-graph networks. *Sci. reports* **10**, 1–16 (2020).
13. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**, 184–192 (2020).
14. Žitnik, M. *et al.* Gene prioritization by compressive data fusion and chaining. *PLoS computational biology* **11**, e1004552 (2015).
15. Ashtawy, H. M. *Data-Driven and Task-Specific Scoring Functions for Predicting Ligand Binding Poses and Affinity and for Screening Enrichment* (Michigan State University, 2017).
16. Avery, C., Patterson, J., Grear, T., Frater, T. & Jacobs, D. J. Protein function analysis through machine learning. *Biomolecules* **12**, 1246 (2022).
17. Berman, H. M. *et al.* The protein data bank. *Nucleic acids research* **28**, 235–242 (2000).
18. Yang, J., Shen, C. & Huang, N. Predicting or pretending: artificial intelligence for protein-ligand interactions lack of sufficiently large and unbiased datasets. *Front. pharmacology* **11**, 69 (2020).

19. Sinha, S., Tam, B. & Wang, S. M. Applications of molecular dynamics simulation in protein study. *Membranes* **12**, 844 (2022).

20. Du, X. *et al.* Insights into protein–ligand interactions: mechanisms, models, and methods. *Int. journal molecular sciences* **17**, 144 (2016).

21. Childers, M. C. & Daggett, V. Insights from molecular dynamics simulations for computational protein design. *Mol. systems design & engineering* **2**, 9–33 (2017).

22. Kanakala, G. C., Aggarwal, R., Nayar, D. & Priyakumar, U. D. Latent biases in machine learning models for predicting binding affinities using popular data sets. *ACS Omega* (2023).

23. Defelipe, L. A. *et al.* Solvents to fragments to drugs: Md applications in drug design. *Molecules* **23**, 3269 (2018).

24. Seo, M.-H., Park, J., Kim, E., Hohng, S. & Kim, H.-S. Protein conformational dynamics dictate the binding affinity for a ligand. *Nat. communications* **5**, 1–7 (2014).

25. Bronowska, A. K. Thermodynamics of ligand-protein interactions: implications for molecular design. In *Thermodynamics-Interaction Studies-Solids, Liquids and Gases* (IntechOpen, 2011).

26. Wang, C., Greene, D., Xiao, L., Qi, R. & Luo, R. Recent developments and applications of the mmpbsa method. *Front. molecular biosciences* **4**, 87 (2018).

27. Korlepara, D. B. *et al.* Plas-5k: Dataset of protein-ligand affinities from molecular dynamics for machine learning applications. *Sci. data* **9**, 1–10 (2022).

28. Pettersen, E. F. *et al.* Ucsf chimera—a visualization system for exploratory research and analysis. *J. computational chemistry* **25**, 1605–1612 (2004).

29. Gordon, J. C. *et al.* H++: a server for estimating p k as and adding missing hydrogens to macromolecules. *Nucleic acids research* **33**, W368–W371 (2005).

30. Case, D. A. *et al.* The amber biomolecular simulation programs. *J. computational chemistry* **26**, 1668–1688 (2005).

31. Salomon-Ferrer, R., Case, D. A. & Walker, R. C. An overview of the amber biomolecular simulation package. *Wiley Interdiscip. Rev. Comput. Mol. Sci.* **3**, 198–210 (2013).

32. Jorgensen, W. L., Chandrasekhar, J., Madura, J. D., Impey, R. W. & Klein, M. L. Comparison of simple potential functions for simulating liquid water. *The J. chemical physics* **79**, 926–935 (1983).

33. Maier, J. A., Martinez, C., Kasavajhala, L., Koushik a nd Wickstrom, Hauser, K. E. & Simmerling, C. ff14sb: improving the accuracy of protein side chain and backbone parameters from ff99sb. *J. chemical theory computation* **11**, 3696–3713 (2015).

34. Wang, J., Wolf, R. M., Caldwell, J. W., Kollman, P. A. & Case, D. A. Development and testing of a general amber force field. *J. computational chemistry* **25**, 1157–1174 (2004).

35. Wang, J., Wang, W., Kollman, P. A. & Case, D. A. Automatic atom type and bond type perception in molecular mechanical calculations. *J. molecular graphics modelling* **25**, 247–260 (2006).

36. Eastman, P. *et al.* Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology* **13**, e1005659 (2017).

37. Wang, J. *et al.* Weak-binding molecules are not drugs?—toward a systematic strategy for finding effective weak-binding drugs. *Briefings Bioinforma.* **18**, 321–332 (2017).

38. Buratto, R., Mammoli, D., Canet, E. & Bodenhausen, G. Ligand–protein affinity studies using long-lived states of fluorine-19 nuclei. *J. medicinal chemistry* **59**, 1960–1966 (2016).

39. Ohlson, S. Designing transient binding drugs: a new concept for drug discovery. *Drug Discov. Today* **13**, 433–439 (2008).

40. Ivanović, V., Rančić, M., Arsić, B. & Pavlović, A. Lipinski's rule of five, famous extensions and famous exceptions. *Pop. Sci. Article* **3**, 171–177 (2020).

41. Oprea, T. I. Property distribution of drug-related chemical databases. *J. computer-aided molecular design* **14**, 251–264 (2000).

42. Veber, D. F. *et al.* Molecular properties that influence the oral bioavailability of drug candidates. *J. medicinal chemistry* **45**, 2615–2623 (2002).

43. Ghose, A. K., Viswanadhan, V. N. & Wendoloski, J. J. A knowledge-based approach in designing combinatorial or medicinal chemistry libraries for drug discovery. 1. a qualitative and quantitative characterization of known drug databases. *J. combinatorial chemistry* **1**, 55–68 (1999).

44. Zheng, L., Fan, J. & Mu, Y. Onionnet: a multiple-layer intermolecular-contact-based convolutional neural network for protein–ligand binding affinity prediction. *ACS omega* **4**, 15956–15965 (2019).

Acknowledgements

We thank Akash Ranjan for the website development. We thank IHub-Data for its support. The authors thank IIT Delhi and IIIT Hyderabad HPC facilities for computational resources. DN acknowledges financial support by INSPIRE faculty research grant (DST/INSPIRE/04/2018/000455) provided by the Department of Science and Technology, India. UDP thanks DST-SERB (CRG/2021/008036) and Kohli Center on Intelligent Systems, IIIT Hyderabad for support.

Author contributions statement

UDP conceived the study, D.B.K. and S.H.R wrote the codes and analyzed the data. C.S.V. and D.B.K. contributed to the writing of the manuscript. S.H.R trained ML model. D.B.K., C.S.V., and S.P. performed docking studies. D.B.K., C.S.V., R.S., P.K.P., S.H.R., V.K., S.P., S.S., S.J., S.P., K.T., R.J., S.V., A.G.N., contributed to the preparation of dataset and simulation. D.N., and UDP supervised the project. Indhu Ramachandran for coordinating this project and P.C contributed in checking data.

Competing interests

The authors declare no competing interests.

Figures & Tables

Exp. vs MMPBSA	Precision	Recall	f1-score	support
Strong Binders	0.86	0.78	0.82	2579
Weak Binders	0.72	0.81	0.76	1764
Accuracy			0.79	4343
Macro Average	0.79	0.79	0.79	4343
Weighted Average	0.80	0.79	0.79	4343

Table 1. Performance metrics from confusion matrix to evaluate the classification models performance in distinguishing strong and weak binders based on MMPBSA calculations.

Exp. vs Docking	Precision	Recall	f1-score	support
Strong Binders	0.79	0.74	0.76	2579
Weak Binders	0.65	0.72	0.68	1764
Accuracy			0.73	4343
Macro Average	0.72	0.73	0.72	4343
Weighted Average	0.74	0.73	0.73	4343

Table 2. Performance metrics from confusion matrix to evaluate the classification models performance in distinguishing strong and weak binders based on docking simulations.

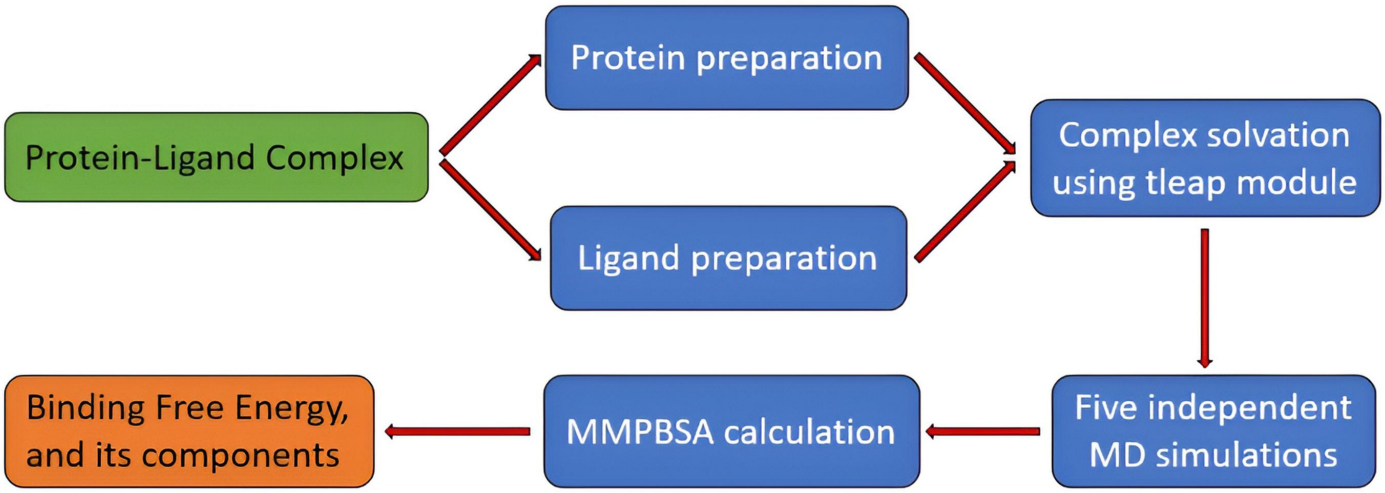

Figure 1. Protocol for input preparation and simulations. A similar approach to our earlier work has been followed.[27]

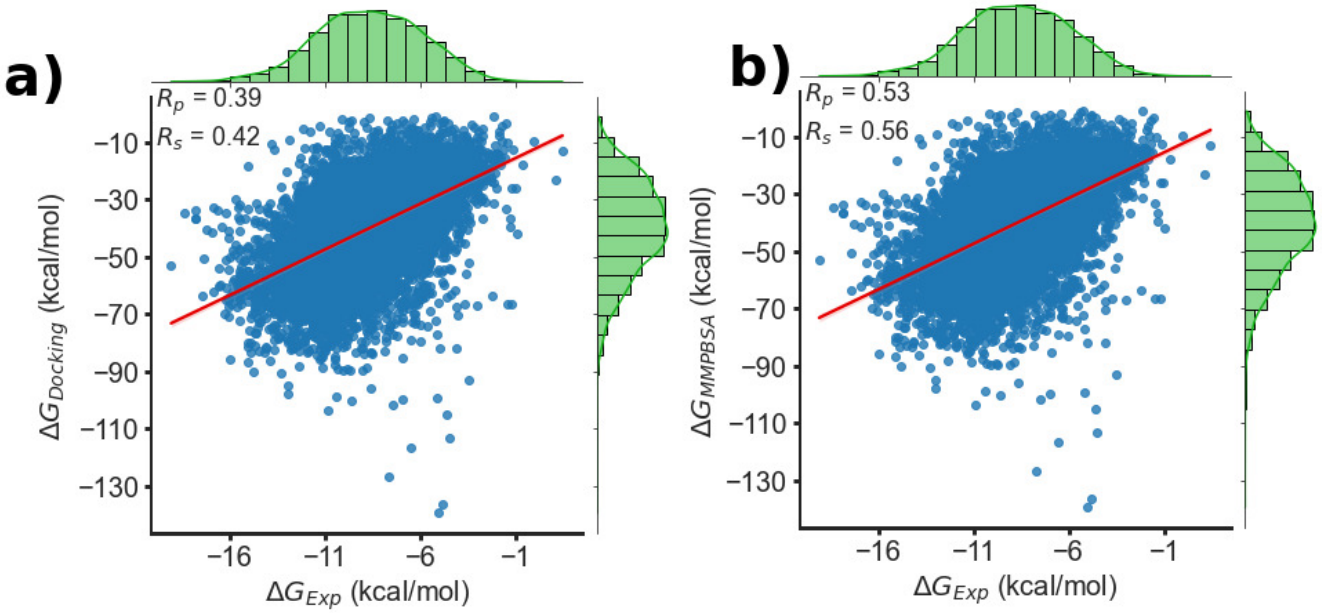

Figure 2. Correlation plots between the experimental and calculated binding affinities for a subset with 6842 (includes 2000 data points from PLAS-5k dataset[27]) pdbids. The calculated binding affinities are calculated (a) using Auto-dock Vina, and (b) using MM-PBSA.

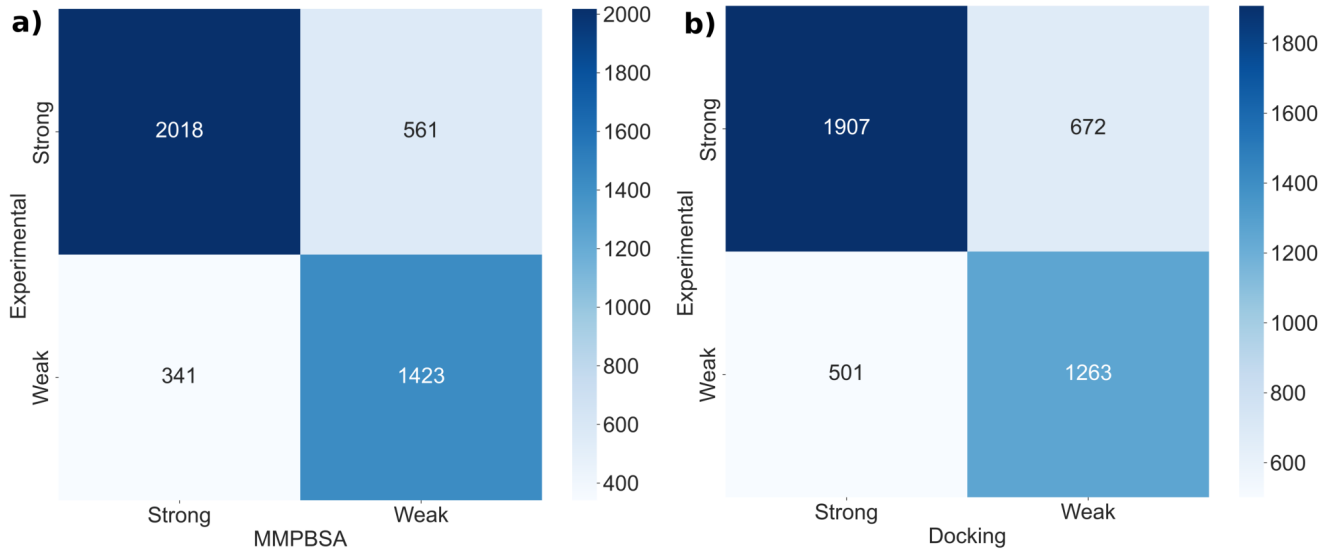

Figure 3. Confusion matrix to distinguish strong and weak binders (a) Experimental vs MMPBSA, (b) Experimental vs Docking.

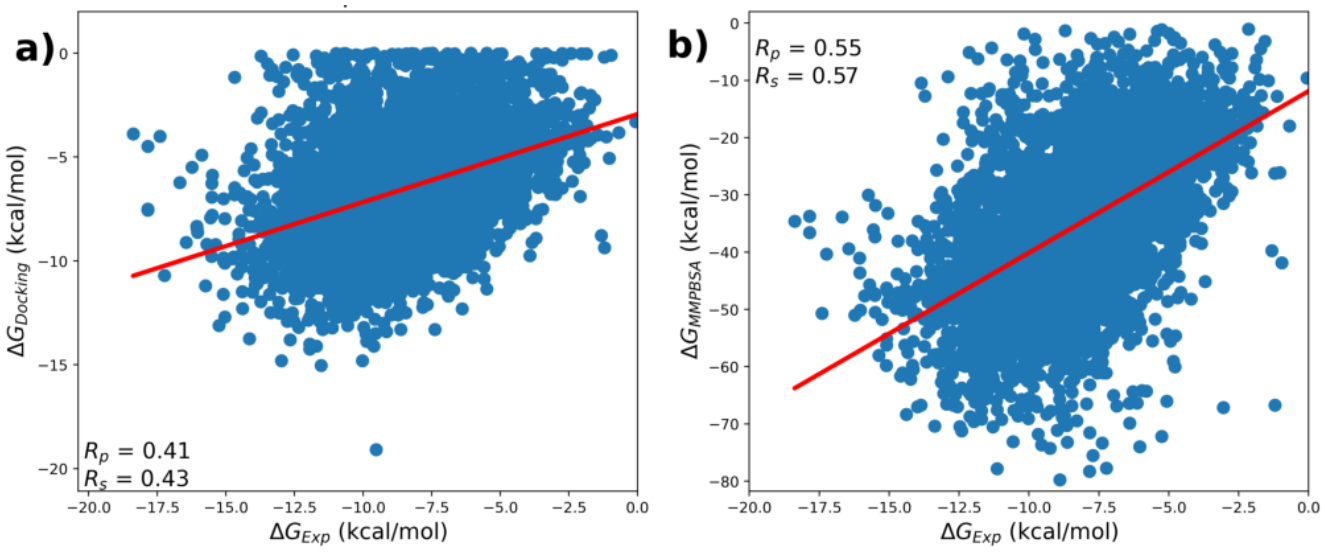

Figure 4. Correlation plots for a set of PDB ids from PLAS-20k (which follows Lipinski rule of five - Molecular weight, number of donors and number of acceptors of the ligand) for which experimental binding affinities are known - (a) Experimental vs Docking, (b) Experimental vs MM-PBSA

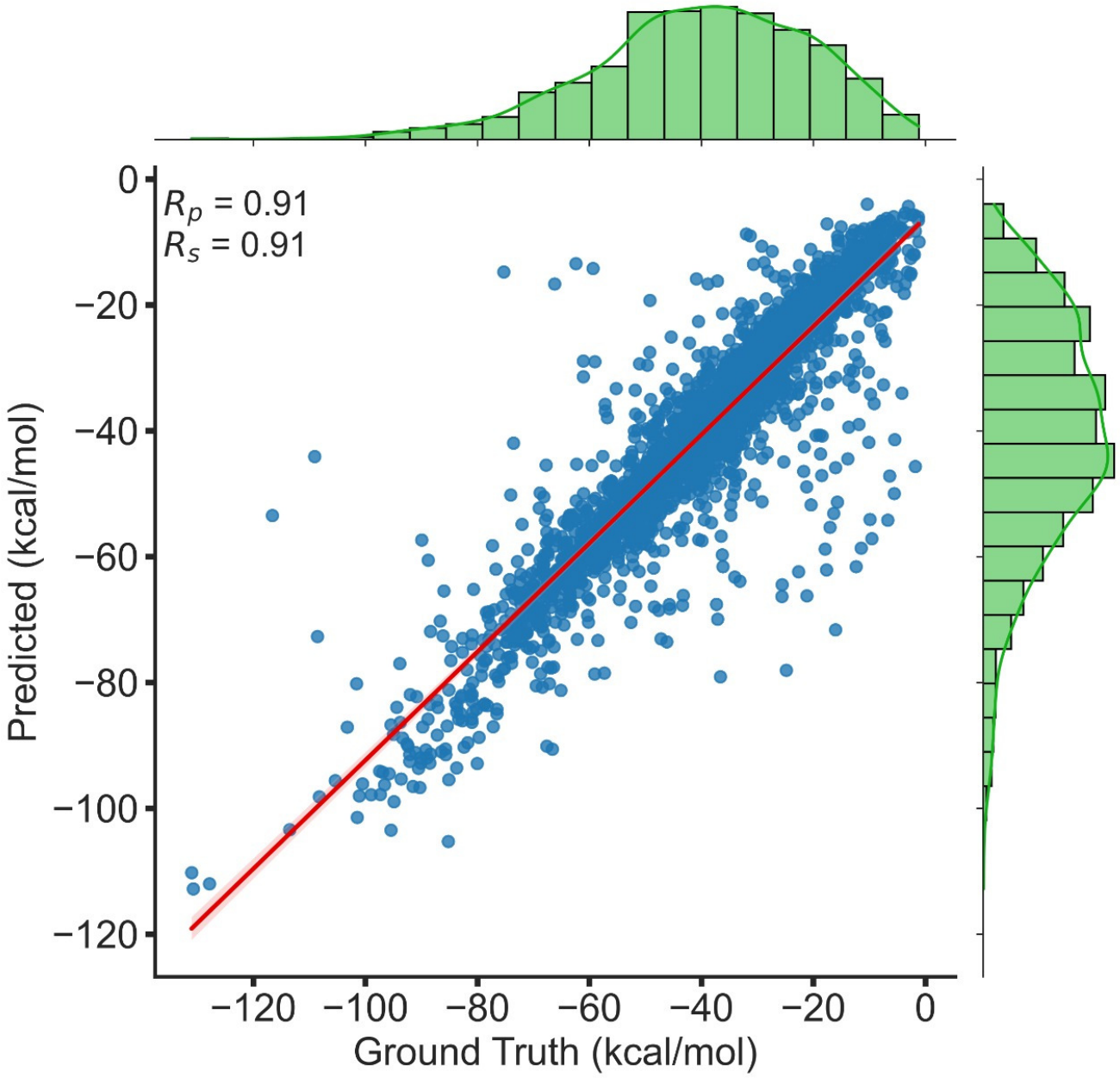

Figure 5. Pearson correlation coefficient of OnionNet trained on PLAS-20k dataset.