DEELIG: A Deep Learning-based approach to predict protein-ligand binding affinity.

Asad Ahmed[1¶], Bhavika Mam[2,3¶], Ramanathan Sowdhamini[2]*

1. National Institute of Technology, Warangal, Telangana, India
2. National Centre for Biological Sciences, Tata Institute for Fundamental Research, GKVK Campus, Bangalore, Karnataka, India
3. The University of Trans-Disciplinary Health Sciences and Technology (TDU), Bangalore, Karnataka, India

¶ Co-first author

*** Corresponding author:**

Email: mini@ncbs.res.in (RS)

Abstract

Protein-ligand binding prediction has extensive biological significance. Binding affinity helps in understanding the degree of protein-ligand interactions and has wide protein applications. Protein-ligand docking using virtual screening and molecular dynamic simulations are required to predict the binding affinity of a ligand to its cognate receptor. In order to perform such analyses, it requires intense computational power and it becomes impossible to cover the entire chemical space of small molecules. It has been aided by a shift towards using Machine Learning-based methodologies that aids in binding prediction using regression. Recent developments using deep learning has enabled us to make sense of massive amounts of complex datasets. Herein, the ability of the model to "learn" intrinsic patterns in a complex plane of data is the strength of the approach. Here, we have incorporated Convolutional Neural Networks that find spatial relationships among data to help us predict affinity of binding of proteins in whole superfamilies towards a diverse set of ligands. The models were trained and validated using a detailed methodology for feature extraction. We have also tested DEELIG on protein complexes relevant to the current public health scenario. Our approach to network construction and training on protein-ligand dataset prepared in-house has provided significantly better results than previously existing methods in the field.

Introduction

Proteins are a diverse class of dynamic macromolecular structures in living organisms and are essential for the biochemistry and physiology of the organism. Depending on their functional role (s), proteins may bind to other proteins, peptides, nucleic acids and non-peptide ligands with varying affinities. Determining protein-ligand affinity helps in understanding the reaction mechanism and kinetics of the reaction and has applications in drug development and pharmacology (1).

Protein-ligand interaction is measured in terms of Binding affinity. The stronger the readout for binding affinity, the stronger the interaction between protein and ligand may be inferred. It is quantified in terms of Inhibition constant (Ki), dissociation constant (Kd), changes in free energy measures (delta G, delta H) and (IC50) (2). Predicting binding affinity between a protein and ligand complements experimental approaches and is usually used as a start-point for the latter. Prediction-based approaches are also useful in cases where experimental determination of binding affinity may not be feasible.

Classical prediction methods to score free binding energies of small ligands to biological macromolecules such as MM/GBSA and MM/PBSA typically rely on molecular dynamic simulations for calculations and aid *in silico* docking and virtual screening as well as experimental approaches. However, there is a trade-off between computational resources and accuracy (3).

With a recent shift towards the use of machine learning and deep-learning based methods in the field of structural biology, making biologically significant predictions using regression and 'learning' intrinsic patterns in a complex plane of available data has led to resource-optimal predictions without compromising on accuracy. Deep learning has been known to learn representations and patterns in complex data forms. Our aim was to apply deep learning to predict binding affinity of protein-ligand interaction.

Convolutional Neural Networks (CNN) are deep neural networks that use an input layer, output later as well as convolutional hidden layer(s). The first CNN was incorporated by LeCunn in 1998 (4) the connectivity pattern of which was inspired by the elegant experiments of Hubert and Weisel on the mammalian visual cortex in the 1960s (5). With the growing technical advancements and massive amounts of data, CNNs have emerged popular in biological fields in the recent decade with various applications (6).

In our study, we have used CNNs to provide a quantitative estimate of protein-ligand binding using various sets of features corresponding to protein and ligand respectively by finding

spatial relationships amongst the data. Our approach was validated using ligand-bound complexes from kinases superfamily in the PDB. Kinases belong to a class of enzymes required for substrate dependent phosphorylation. They are represented across diverse cellular functions like signaling, differentiation, glycolysis (7). We have also tested our model on COVID-19 main protease (8) of the novel coronavirus strain complexed with various inhibitors of which binding affinities have not been predicted or experimentally determined so far.

Materials and Method

Novel Dataset: Raw Data

The raw data for our novel database was obtained from RCSB PDB (9) database, where following were selected as the query parameters.

- **Chain Type:** Protein Chain, No DNA or RNA or DNA/RNA Hybrid.
- **Binding Affinity:** Kd or Ki value present.
- **Chemical Components:** Has ligand (s)
- **X-ray crystallography method:** Resolution upto 2.5 A.

These criteria resulted in a list of 5464 protein PDB IDs, 2568 complexed ligand (s) and corresponding binding affinity values. The search results include the structures present in PDBdatabase, PDBBind (10, 11, 12), PDBMoad (13, 14) and scPDB (15) for its results. Initial raw data database created contained protein structures in PDB format, protein sequences in FASTA format, ligand in SDF format and binding affinity values of corresponding protein-ligand pairs for *5464* complexes.

Dataset Refinement

The PDB, FASTA and SDF files filtered were further processed to refine our novel dataset, as shown in Figure 1. Protein-ligand complexes were 5,464 in number and corresponded to 29,650 complex **unique** chain-ligand pairs. Binding affinity values were obtained from the RCSB database and protein chain-ligand pairs with corresponding binding affinity as 0 were discarded to reduce statistical errors. This narrowed down the total complexes to 4,750 protein-ligand pairs.

Pocket information was extracted from the protein using Ghecom (16) and converted to MOL2 format using Chimera (17), which narrowed our results to 4699 pocket-ligand pairs. It narrowed down the size of the dataset to 4286 pocket-ligand pairs.
We discarded other protein-ligand pairs with missing PSSM profiles, secondary structure or dihedral angle information.

It resulted in a total of 4041 pocket-ligand pairs, which corresponds to 7414 pocket-ligand pairs containing unique chains.

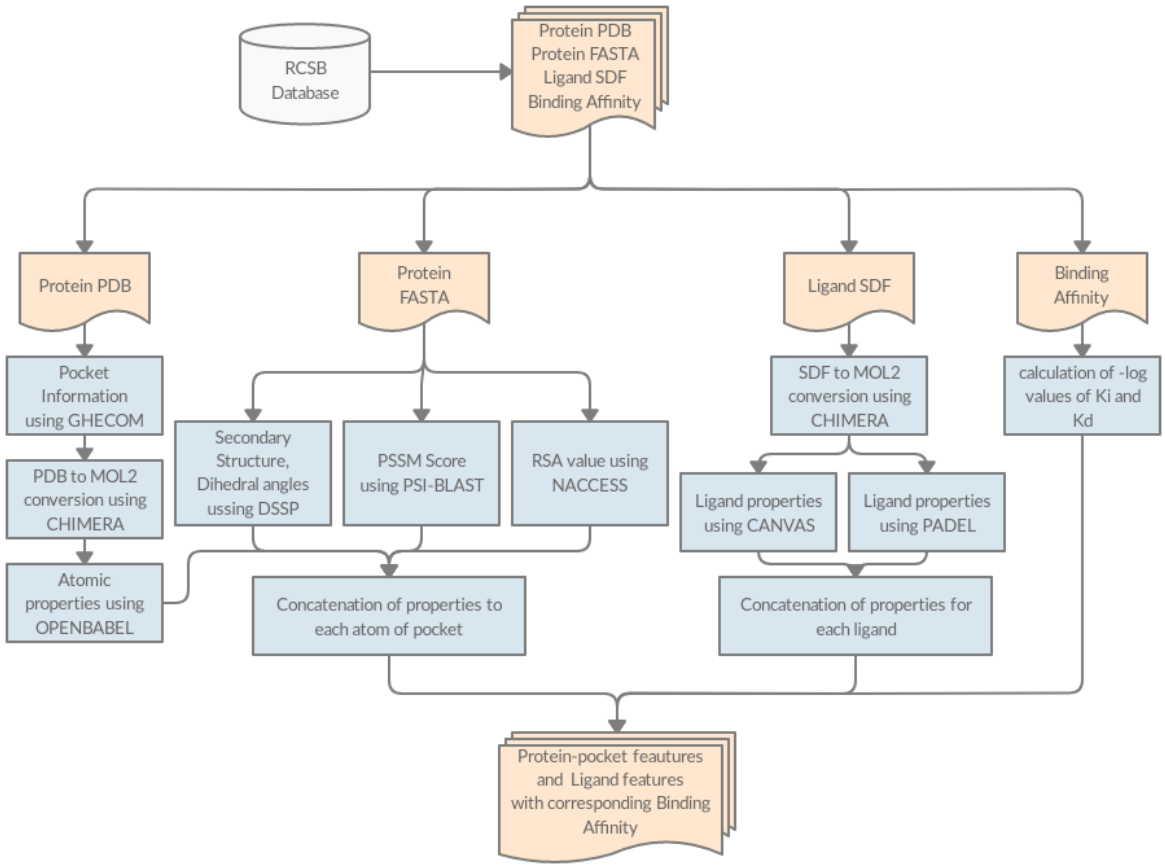


Figure 1: Feature Extraction pipeline

Feature Extraction

Training the deep learning network on raw information is known to result in longer time for convergence and less accuracy. We followed a conventional methodology for feature extraction and used the deep learning framework to learn the interaction between the protein-pocket and ligand for their affinity prediction.

Protein-Pocket features

A comprehensive two-level feature extraction methodology, one at the atomic level and the other at the level of amino acids utilizing structural information and protein sequence respectively.

Atomic Level (19 Bits)

- 9 Bit 1 hot or all null hot encoding for atom types: B, C, N, O, P, S, Se, halogen and metal.
- 1 integer for hybridization
- 1 integer representing the number of bonds with heavy atoms
- 1 integer representing the number of bonds with hetero atoms
- 5 bits (1 if present) encoding properties defined with SMARTS patterns: hydrophobic, aromatic, acceptor, donor and ring
- 1 float for partial charges
- 1 integer to distinguish between ligand as -1 and protein as 1

Amino Acid level (25 Bits)

We utilized the sequence information of protein to get more features about the protein pocket-ligand interaction.

- Position-Specific Scoring Matrix (PSSM): PSSM is a matrix that represents the probability of mutation at each point of the sequence. It gives a 20 bit- probability for each amino acid at each location. PSSM profiles were obtained using PSI-BLAST (18) with SwissProt as subject database and E-value threshold as 0.001. Chains with less than 50 amino acids were removed from the input dataset.
- Relative Solvent Accessibility (RSA): It is encoded by 1 bit of information for each amino acid that provides whether it is buried or exposed to the solvent. We set a threshold of 25% in RSA values. RSA was obtained using NACCESS (19).
- Secondary Structure: It is encoded by 1 bit of information about the structure as coil, helix or plate and was predicted using the DSSP (20, 21).
- Dihedral Angles: It is encoded by 2 bits of information with phi / psi angles of each of the amino acids and was predicted using DSSP (20, 21) for obtaining dihedral angles.

Ligand Features

Standard ligand features were calculated for ligands in our dataset using PADEL (22) and 1D, 2D and chemical fingerprints, which includes hybridisation, atom pair interaction, counts of various functional group.

We also used QikProp (34) and CANVAS (23) to derive ADMET (Absorption, Distribution, Metabolism, Excretion, and Toxicity) properties, which includes the physical properties,

solubility and partition coefficients. The exhaustive list of every property calculated is given in the appendix.
It results in a 1D array of 14,716 dimensions containing the various properties of a given ligand. This is used as a feature vector representing the ligand represented in MOL2 format.

Grid Formation

The three-dimensional co-ordinates of atoms were converted into a 3D grid of resolution 10Å with 1Å spacing between the two axes centered along the centroid of the ligand. Atoms outside each such grid were discarded. The atoms lying inside the grid were rounded up to the nearest coordinate of the grid where features of corresponding atoms that lay in the same coordinates were added up.
This resulted in projecting ligand-interacting residues into a three-dimensional cube with features representing the atomic as well as protein-based properties of each atom of the protein pocket.

Strategies

Atomic Model

Preprocessing

Features were calculated at the atomic level (Section 4.1.1) corresponding to each atom of an amino acid and ligand. A 19-bit vector was calculated that uniquely identified each of the atoms in the 3D co-ordinates of a given protein-pocket and ligand complex. A 4D tensor each of size m x m x m x 19, i.e. the 3 coordinates (x, y, z) and the features, where m represents the number of atoms present in a complex was constructed as the feature vector representing the given protein pocket-ligand.

The 4D vector contains the protein-pocket features and was converted to a 3D grid using grid featurization (Section 4.3). The 3D- featurized grid is essentially a 4D tensor, where the coordinates are approximated to the points on the grid.

The dataset is converted to vectors and is divided into training:validation:test sets in ratio 80:10:10.

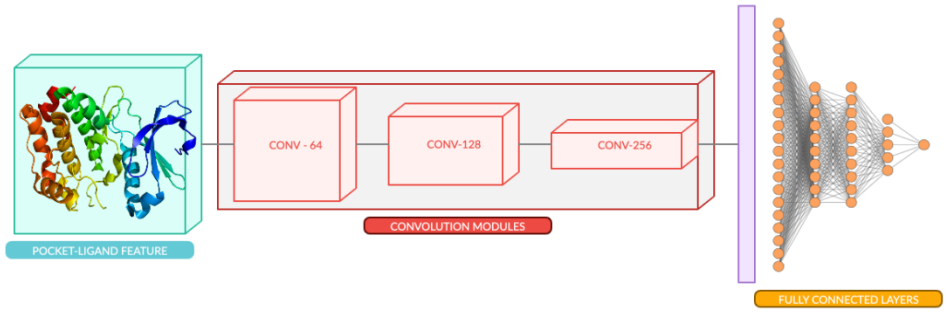


Figure 2: Training framework for Atomic Model. The framework is trained on 19 bits features each of protein-pocket and ligand together as input.

Architecture

Convolutional Neural Networks (24) have been used to capture spatial features in an image. We use CNNs to capture the interaction between ligand and protein atoms in three-dimensional space. A network was constructed (Figure 2) with a 3D CNN of varying channel sizes of [64, 128, 256] with non-linear activation ReLU after each layer, each 3D CNN had a filter of 5Å cube which was used to perform convolution operations. MaxPool (25) layer that acts in three dimensions to lower the dimension with a pool size of 2Å cube and Batch Normalization (26) layer is added after each CNN layer, this in turn decreases the training time and helps in faster convergence.

The latent features learnt from the above CNN layers were then flattened and used for calculating the binding affinity of the protein pocket-ligand pair. The CNN derives the relation among the 3D coordinates and their features, which would correspond well to the binding affinities of complexes.

The features from the last CNN layer are then flattened out, and passed through a fully connected neural network having the number of neurons as [1000, 500, 250] with ReLU as non-linearity after each layer. Dropout (25) is added after each layer to prevent overfitting by forcing the neural network to learn various other pathways by randomly assigning neurons to zero, 0.50 as Dropout threshold. Dense network predicts a regressive value of Binding Affinity, corresponding to a single neuron output.
Training framework is shown in Figure 2 and a detailed layer network is shown in Figure 4 (a).

Training

The featurized protein-pocket grid formed was rotated to all 24 combinations possible, such that the network is able learn in an orientation invariant form.

The network was trained by taking Mean Square error between the predicted and actual values as a loss function. The network was optimized using Adam (27) as the optimizer with a learning rate of 1e-5 and weight decay of 0.001 for 20 epochs. Network was trained on an Nvidia Pascal GPU using Pytorch (28) as the framework.

Composite Model

Preprocessing

Features were calculated at the amino acid level (Section 4.1.2) and were concatenated alongside the atomic level features (Section 4.1.1) to each atom of amino acid. It results in a 44-bit vector uniquely identifying each of the atoms in the 3D co-ordinates of a given protein. A 4D tensor each of sizes m x m x m x 44, i.e. the 3 coordinates (x, y, z) and the features, where m represents the number of atoms present in a complex is constructed as the feature vector of protein pocket.

The 4D vector contains the protein-pocket features, it was converted to a 3D grid using grid featurization (Section 4.3). The 3D featurized grid is essentially a 4D tensor, where the coordinates are approximated to the points on the grid.

The ligands were separately featurized by calculating the ligand properties (Section 4.2), which results in a 1D tensor.

The dataset is converted to vectors and is divided into training:validation:test sets in ratio 80:10:10.

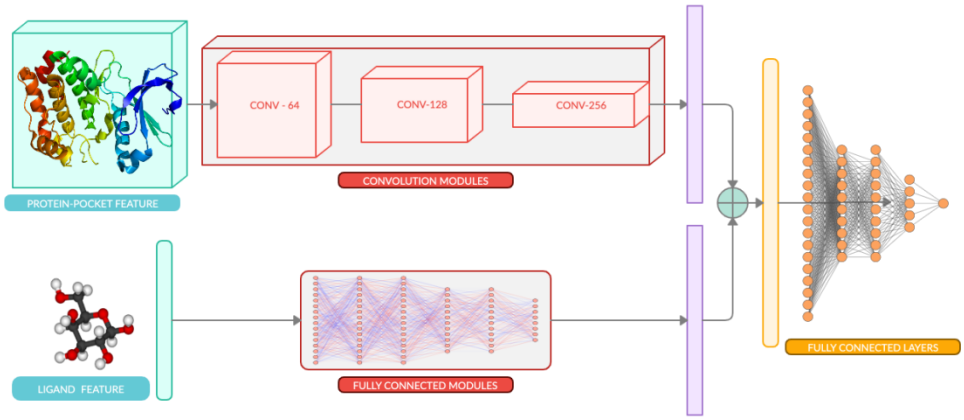


Figure 3: Training framework for Composite Model. The framework is trained on 44 bits features of protein-pocket and 14716 bits of ligand as separate inputs.

Architecture

A multi-input network was constructed (29) with a 3D CNN (24) of varying channel sizes of [64, 128, 256] with non-linear activation ReLU after each layer, each 3D CNN had a filter of 5Å cube which was used to perform convolution operations. We also added MaxPool (25) layer that acts *three-dimensionally* to lower *dimensionality* while retraining features learnt after each CNN layer. It has a filter size of 2Å cube. Batch Normalization (26) layer was added after each CNN module for faster convergence.

The ligand features were passed through the dense layers of sizes [7000, 5000, 2000] with ReLU as non-linearity after each layer and we also perform dropout operations after each dense layer to prevent it from overfitting (30). This results in a latent vector representing the relevant features for each ligand.

The latent output from the CNN layers is flattened and concatenated with the latent feature vector of ligand, to create one single feature vector of protein pocket-ligand interactions. This vector is passed through a densely connected neural network having the number of neurons as [7000, 2000, 500, 200] with ReLU as non-linearity after each layer and we used Dropout after each layer also to prevent overfitting forcing the neural network to learn various other pathways by randomly assigning weights of neurons to zero, with 0.50 as Dropout threshold. This dense network finally predicts a regressive value of Binding Affinity, corresponding to a single neuron output.
Training framework is shown in Figure 3 and a detailed layer network is shown in Figure 4 (b)

Training

The featurized protein-pocket grid formed was rotated to all 24 combinations possible, such that the network is able to learn in an orientation invariant form.

The featurized protein pocket-ligand pair of training set was passed through corresponding the network and trained by taking Mean Square error between the predicted and actual values as a loss function. The network was optimized using Adam (27) as the optimizer with a learning rate of 1e-5 and weight decay of 0.001. The network was trained on an Nvidia Pascal GPU using Pytorch (28) as the framework.

Performance Evaluation

The performance of the models was quantified using Mean Absolute error (MAE) and Root mean square error (RMSE). It was tested on validation and testing sets which were initially divided from our dataset as mentioned in the training section. Lower error corresponds to better learning capacity of the model. Standard deviation among the real and predicted values was also calculated.
The MAE, RMSE and SD values are shown in Table 1.

Table 1: Predictions accuracy on test set of our novel dataset

Method	**MAE**	**RMSE**	**SD**
Atomic Model	2.84	3.93	2.62
Composite Model	**2.27**	**3.07**	**2.06**

For the purpose of training and testing models, one NVIDIA Tesla P100 GPU cluster was used. Computational time taken for featurization of dataset, training and testing were 52 hours, 22 hours and 8 minutes respectively.

Results and Discussion

Two modules were trained. The first module was trained using a small set of features for protein and ligand, which were represented together in a 3D grid space. This approach has also been part of a previous study (29). However, the previous study uses a restricted ligand set that does not involve larger ligands. Here we have used a diverse set of ligands as one of our inputs. With training of Atomic Model for *35 epochs*, ***MAE score of 2.84 was achieved*** (Table 1).

We constructed another module that enabled us to improve on the ligand and protein based information. To this purpose, we used an increased feature vector size which amounted to 14716 bits in size for ligand and 44 bits for each atom of protein. With training of Composite Model for only *4 epochs*, **MAE score of 2.27** was achieved (Table 1).

The performance of our model was further evaluated using ligand-bound complexes from the kinase superfamily from PDB. The composite model outperformed the atomic model significantly and with lower standard deviation.

Table 2 : Predictions accuracy on kinases

Method	**MAE**	**RMSE**	**SD**
Atomic Model	2.48	3.24	3.11
Composite Model	**2.24**	**2.71**	**2.67**

In light of the ongoing coronavirus pandemic, we tested protein-ligand complexes from the coronavirus (CoV) family. The COVID-19 main protease is a key enzyme for the novel strain of coronavirus that is being implicated in the pandemic. A recent study involved testing of in-vitro binding efficacy of coronavirus COVID-19 virus main protease (Mpro) with a potent reversible synthetic inhibitor, N3 (31). However, the highly potent inhibition by N3 rendered the experimental determination of binding affinity not achievable. Using the structure of Mpro at high resolution (7BQY: 1.7 Angstrom), we have been able to predict the binding affinity of N3 to 3.1e+4 nanomolar (Table 3). This value agrees with the observed high affinity in the course of recent experiments (31).

The crystal structure complexes (PDB IDs: 5R7Y, 5R7Z, 5R82, 5R84) of the COVID-19 main protease with inhibitors ((Z45617795: N-[(5-methylisoxazol-3-yl)carbonyl]alanyl-L-valyl-N~1~- ((1R,2Z)-4- (benzyloxy)-4-oxo-1-{[(3R)-2-oxopyrrolidin-3-yl]methyl)but-2-enyl)-L-leucinamide); Z1220452176: (~{N)-[2- (5-fluoranyl-1~{H)-indol-3-yl)ethyl]ethanamide); Z219104216: 6- (ethylamino)pyridine-3-carbonitrile; Z31792168: 2-cyclohexyl-~{N)-pyridin-3-yl-ethanamide)) respectively has been recently deposited in PDB (2020; unpublished).

Another study has deposited the complex of the COVID-19 main protease with a broad-spectrum inhibitor X77 (N- (4-tert-butylphenyl)-N-[(1R)-2- (cyclohexylamino)-2-oxo-1-(pyridin-3-yl)ethyl]-1H-imidazole-4-carboxamide) (2020; unpublished). We used these complexes to predict their respective binding affinities as they have not been made available. Based on our model-based predictions, broad spectrum inhibitor X77 scores for highest affinity followed by ligands Z45617795, N3, Z31792168, Z1220452176 and Z219104216 in the order of decreasing binding affinity (Table 3) strengthening the suitability of X77 as a potential candidate against COVID-19 virus protease.

Table3: Predictions of Binding Affinity on COVID-19 complexes

PDB	Ligand	-Log (Kd/Ki)	[Kd] or [Ki] (nM)
5R7Y	Z45617795	11.96	6.39e+3
5R7Z	Z1220452176	7.69	4.57e+5
5R82	Z219104216	6.12	2.18e+6
5R84	Z31792168	8.32	2.43e+5
6W63	X77	15.34	2.17e+2
7BQY	N3	10.38	3.10e+4

Conclusion

We propose a deep-learning based approach to predict ligand (eg., drug)–target binding affinity using only structures of target protein (PDB format) and ligand (SDF format) as inputs. Convolutional Neural Networks (CNN) were used to learn representations from the features extracted from these inputs and hidden layers in the affinity prediction task. We used two approaches to feature extraction- atomic level as well as composite level and compared their performance using the same network.

Deep-learning based approaches have been implemented for prediction of binding affinity. One of the studies used atomic level features of complex in a CNN based framework for binding affinity prediction (32), while another study used protein sequence level features in a CNN based framework for prediction (33). Another approach used as been to use feature learning along with gradient boosting algorithms to predict binding affinity (34). Here, we provide a composite model that incorporates tripartite structural, sequence and atomic level features with those of the atomic and other chemical features of the ligand to predict binding affinity of a putative complex.

We have trained two models to predict the binding affinity between protein and ligand in a given complex. This would help existing databases like RSCB PDB, PDBBind in filling missing binding affinity data for complexes.
We have constructed a novel dataset that represents a diverse set of ligands and using a novel deep learning based approach we have achieved significant improvement in prediction of binding affinity of protein-ligand complexes. Interestingly, our approach performed better without ligand coordinates as input. To counter filtering or noise reduction in our dataset, our dataset constructed is smaller than PDBBind (35) but we have overcome the constraints on ligand selection part of a previous study (29). Although our dataset contains 5464 complexes

compared to 16,151 complexes found in PDBBind, the ligands used as part of our training include 452 unique ligands absent in PDBBind. This helps in achieving ligand diversity during training the CNN model. The similarity matrix constructed from the binary fingerprints of ligands used in the dataset supports our claim of improved ligand diversity in our dataset (Supplementary File S1).
We have also eliminated the need of providing ligands in a complex form with protein. Thus a given protein pocket may be tested for the degree of binding for any given ligand. This can be extended to predicting potential binding partners for proteins in other superfamilies as well. It is also important to consider that docking score and pose is not a reliable correlation with MM/GBSA poses (36). **DEELIG can be used for a member of any protein superfamily and a non-peptide ligand, the docking pose of which may or may not be known.**
The code repository for the project is publicly available at :

https://github.com/asadahmedtech/DEELIG

Future Direction

Binding affinity predictions through DEELIG can be extended to protein-ligand complexes of protein superfamilies where the affinity is quantitatively unknown due to experimental limitations or where the potential for binding is yet to be explored *in vitro*. A webserver to implement DEELIG for easy online access would be useful for the general scientific community and this will also be in the pipeline. A later version of DEELIG which is trained on peptide ligand dataset will also be worked on.

References

[1] Ahmed Aqeel, Smith Richard D, Clark Jordan J, Dunbar Jr. James B, and Carlson Heather A. Recent improvements to binding moad: a resource for protein ligand binding affinities and structures. Nucl. Acids Res., 43 (D):465–469, 2014.

[2] Stephen F Altschul, Thomas L Madden, Alejandro A Schäffer, Jinghui Zhang, Zheng Zhang, Webb Miller, and David J Lipman. Gapped blast and psi-blast: a new generation of protein database search programs. Nucleic acids research, 25 (17):3389–3402, 1997.

[3] Jianxin Duan, Steven L Dixon, Jeffrey F Lowrie, and Woody Sherman. Analysis and comparison of 2d fingerprints: insights into database screening performance using eight fingerprint methods. Journal of Molecular Graphics and Modelling, 29 (2):157–170, 2010.

[4] Samuel Genheden and Ulf Ryde. The mm/pbsa and mm/gbsa methods to estimate ligand-binding affinities Expert opinion on drug discovery, 10 (5):449–461, 2015.

[5] Tong He, Marten Heidemeyer, Fuqiang Ban, Artem Cherkasov, and Martin Ester. Simboost: a read-across approach for predicting drug–target binding affinities using gradient boosting machines. Journal of cheminformatics, 9 (1):1– 14, 2017.

[6] Simon J Hubbard and Janet M Thornton. Naccess. Computer Program, Department of Biochemistry and Molecular Biology, University College London. 2 (1), 1993.

[7] David H Hubel and Torsten N Wiesel. Receptive fields, binocular interaction and functional architecture in the cat's visual cortex. The Journal of Physiology, 160 (1):106–154, 1962.

[8] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. arXiv preprint arXiv:1502.03167 , 2015.

[9] Jin, Z., Du, X., Xu,Y. et al. Structure of M^{pro} from SARS-CoV-2 and discovery of its inhibitors. *Nature* **582,** 289–293 (2020)

[10] Robbie P Joosten, Tim AH Te Beek, Elmar Krieger, Maarten L Hekkelman, Rob WW Hooft, Reinhard Schneider, Chris Sander, and Gert Vriend. A series of pdb related databases for everyday needs. Nucleic acids research, 39 (suppl_1):D411–D419, 2010.

[11] Desaphy Jérémy, Bret Guillaume, Rognan Didier, and Kellenberger Esther. sc-pdb: a 3d-database of ligandable binding sites—10 years on. Nucleic Acids Research, 43 (D1):399–404, 2015.

[12] Wolfgang Kabsch and Christian Sander. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. Biopolymers: Original Research on Biomolecules, 22 (12):2577–2637, 1983.

[13] Panagiotis L Kastritis and Alexandre MJJ Bonvin. On the binding affinity of macromolecular interactions: daring to ask why proteins interact. Journal of The Royal Society Interface, 10 (79):20120835, 2013.8

[14] Takeshi Kawabata. Detection of multiscale pockets on protein surfaces using mathematical morphology. Proteins: Structure, Function, and Bioinformatics , 78 (5):1195–1211, 2010.

[15] Diederik P Kingma and Jimmy Ba. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980,2014.

[16] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In Advances in neural information processing systems, pages 1097–1105, 2012.

[17] Hu L and RD Smith MG Lerner HA Carlson ML, Benson. Binding moad (mother of all databases. Proteins, 60:333–40, 2005.

[18] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 86 (11):2278–2324, 1998.

[19] Yanjun Li, Mohammad A Rezaei, Chenglong Li, Xiaolin Li, and Dapeng Wu. Deepatom: A framework for protein-ligand binding affinity prediction. arXiv preprint arXiv:1912.00318, 2019.

[20] Seonwoo Min, Byunghan Lee, and Sungroh Yoon. Deep learning in bioinformatics. Briefings in bioinformatics, 18 (5):851–869, 2017.

[21] Benson ML, Smith RD, Khazanov NA, Dimcheff B, Beaver J, Dresslar P, Nerothin J, and Carlson HA. Bindingmoad, a high-quality protein-ligand database. Nucleic Acids Research, 36 (D):674–678, 2008.

[22] Hakime Öztürk, Arzucan Özgür, and Elif Ozkirimli. Deepdta: deep drug–target binding affinity prediction. Bioinformatics, 34 (17):i821–i829, 2018.

[23] Adam Paszke, Sam Gross, Soumith Chintala, and Gregory Chanan. Pytorch: Tensors and dynamic neural networks in python with strong gpu acceleration. PyTorch: Tensors and dynamic neural networks in Python with strong GPU acceleration, 6, 2017.

[24] Eric F Pettersen, Thomas D Goddard, Conrad C Huang, Gregory S Couch, Daniel M Greenblatt, Elaine C Meng, and Thomas E Ferrin. Ucsf chimera—a visualization system for exploratory research and analysis. Journal of computational chemistry, 25 (13):1605–1612, 2004.

[25] Peter W Rose, Bojan Beran, Chunxiao Bi, Wolfgang F Bluhm, Dimitris Dimitropoulos, David S Goodsell, Andreas Prlic, Martha Quesada, Gregory B Quinn, John D Westbrook, et al. The rcsb protein data bank: redesigned website and web services. Nucleic acids research, 39 (suppl_1):D392–D401, 2010.

[26] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. arXiv preprint arXiv:1409.1556, 2014.

[27] Marta M Stepniewska-Dziubinska, Piotr Zielenkiewicz, and Pawel Siedlecki. Development and evaluation of a deep learning model for protein–ligand binding affinity prediction. Bioinformatics, 34 (21):3666–3674, 05 2018.

[28] R Wang, X Fang, Y Lu, and S Wang. The pdbbind database: collection of binding affinities for protein-ligand complexes with known three-dimensional structures.J. Med. Chem., 47 (12):2977–80, 2004.

[29] R Wang, X Fang, Y Lu, CY Yang, and S Wang. The pdbbind database: methodologies and updates. J. MedChem., 48 (12):4111–9, 2005.

[30] Renxiao Wang, Xueliang Fang, Yipin Lu, Chao-Yie Yang, and Shaomeng Wang. The pdbbind database: methodologies and updates. Journal of medicinal chemistry, 48 (12):4111–4119, 2005.

[31] Chun Wei Yap. Padel-descriptor: An open source software to calculate molecular descriptors and fingerprints.Journal of computational chemistry, 32 (7):1466–1474, 2011.

[32] Liu Zhihai, Li Yan, Han Li, Li Jie, Liu Jie, Zhao Zhixiong, Nie Wei, Liu Yuchen, and Wang Renxiao. Pdb-wide collection of binding data: current status of the pdbbind database. Bioinformatics, 31 (3):405–12, 2014.

[33] Binding of nicotinoids and the related compounds to the insect nicotinic acetyicholine receptor. Journal of Pesticide Science, 17 (4):231–236, 1992.

[34] Schrödinger Release 2020-3: QikProp, Schrödinger, LLC, New York, NY, 2020

Supplementary Files

a. Similarity matrix of ligands used in dataset (Supplementary File S1)
b. Dataset details (Supplementary File S2)
c. Dataset distribution (Supplementary File S3)

Appendix

A.1: Property list for ligand features

Following properties of ligand were calculated using PADEL (22),

- Basic Group Count
- Carbon Type
- Hybridization Ratio
- Manhold LogP (The Ratio of carbon to hetero atoms)
- Number of Aromatic bonds
- MACCSS Key
- Klehotaroth fingerprints (Types and Counts)
- AtomPair2D fingerprints (Types and Counts)

Following are ADMET and present in PADEL

- *donorHB*
- *accptHB*
- *Constitutional (Electronegativity)*
- *rotatableBondCounts (#ringatoms)*
- *RuleofFive*
- *VABC (Volume)*
- *Weight (mol_MW)*

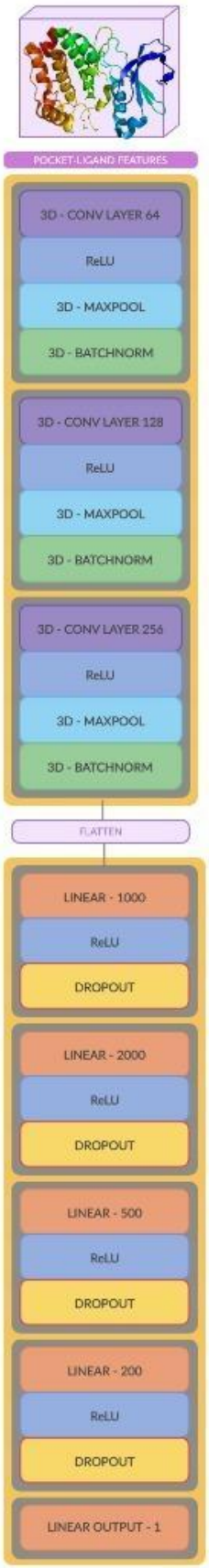
Following {*ADMET*) properties of ligand were calculated using QikProp (34) and CANVAS (16, 21),

- Amine
- Amidine
- Acid
- Amide
- Rotor
- rtvFG (reactive functional groups)

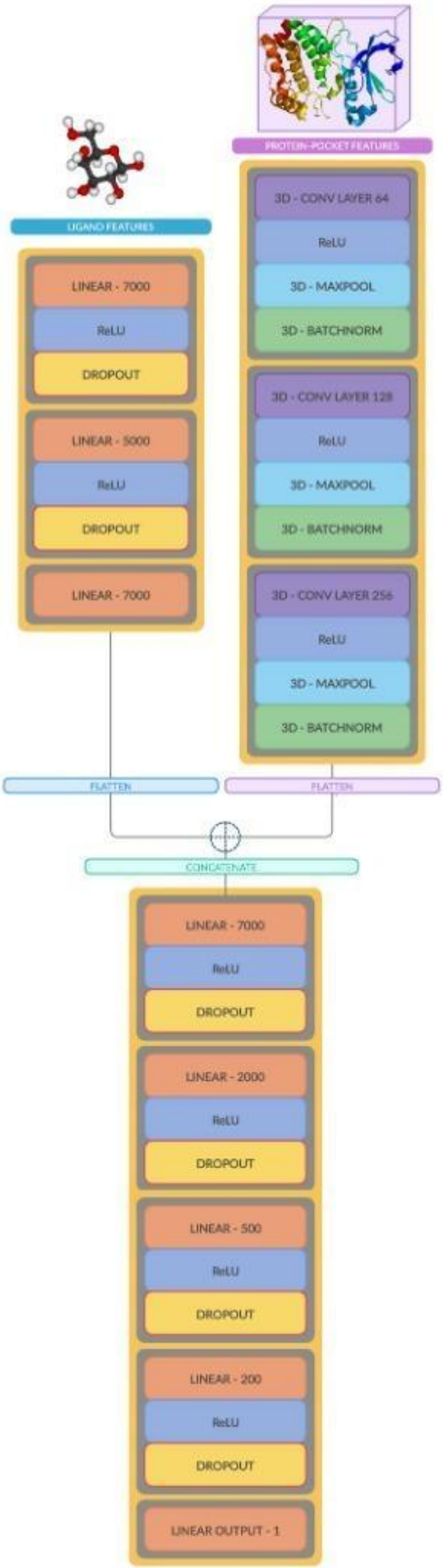
- mol_MW, dipole
- Volume
- donorHB
- accptHB
- QPpolrz (polarizability)

- SASA
    - SASA (probe of 1.4A)
    - FOSA (hydrophobic component of SASA)
    - FISA (hydrophilic component of SASA)
    - PISA (pi of SASA)
    - WPSA (polar of SASA)
    - SAFluorine
    - SAamideO
- **Partition coefficients =>**
    - QPlogPC16
    - QPlogPoct
    - QPlogPw
    - QPlogPo/w
- CIQPlogS (**Conformation indie aqueous solubility**)
- IP (ev) (ionization potential)
- EA (eV) (electron affinity)
- #metab (likely metabolic reactions)
- PSA (van der waals SA of polar N and O atoms)
- #NandO, #ringatoms (number of atoms in rings)
- #in34 (number of atoms in 3 or 4 membered rings)
- #in56 (number of atoms in 5 or 6 membered rings)
- #noncon (ring atoms cannot form conjugated aromatic bonds)
- #nonHatm (heavy atoms- nonhydrogen atoms)
- RuleOfThree
- RuleOfFive (lipinski violations)
- QPlogKhsa (binding to human serum albumin)
- PercentHuman-OralAbsorption

- Globular nature index



(a) Module 1 network



(b) Module 3 network

B.1 Kinases prediction dataset

PDBID	GroundTruth (-log (kd/ki))	Predicted (-log (kd/ki))	Set
1ATP_ATP_E_355	14.33	10.162511	training
1B38_ATP_A_381	12.89	11.011607	training
1B39_ATP_A_381	13.64	10.672279	training
1BX6_BA1_A_351	16.88	17.481277	training
1KV1_BMU_A_391	11.34	11.811321	training
1PXJ_CK2_A_500	8.11	8.140871	training
1Q8T_Y27_A_930	9.22	8.202146	training
1Q8U_H52_A_961	12.6	14.963462	training
1Q8U_H52_A_962	12.6	12.837863	training
1Q8W_M77_A_960	11.05	13.436903	training
1R0E_DFN_A_702	20.73	18.322815	training
1R0E_DFN_B_501	20.73	20.813835	training
1TVO_FRZ_A_1001	11.68	14.313412	training
1UNL_RRC_A_1293	3.97	5.2938104	training
1UU3_LY4_A_1374	4.97	12.5554905	training
1XH4_R69_A_351	14.9	10.410368	training
1XH5_R68_A_1001	10.29	11.608167	training
1XWS_BI1_A_1001	16.12	15.265529	training
1YDT_IQB_E_351	14.51	11.634923	training
2BAK_AQZ_A_401	14.81	13.099378	training
2C5O_CK2_A_1297	8.11	4.862343	training
2C5O_CK2_C_1298	8.11	5.131049	training
2EWA_SB2_A_361	15.98	13.587485	training
2F2U_M77_A_501	14.07	10.874512	training
2F2U_M77_B_1501	14.07	12.677324	training
2J2I_LY4_B_1307	5.92	9.034196	training
2NPQ_BOG_A_1000	10.39	10.77333	training

2NPQ_BOG_A_2000	10.39	9.395628	training
2O3P_QUE_A_501	15.21	11.99402	training
2O63_MYC_A_501	11.77	10.327488	training
2O64_MYU_A_501	12.6	15.132906	training
2QHM_7CS_A_500	11.93	15.419877	training
2RIO_ADP_A_1101	8.5	8.839855	training
2RIO_ADP_B_2101	8.5	8.12286	training
2RKU_R78_A_500	17.33	17.202715	training
2UZT_SS3_A_1351	13.93	10.50961	training
2VU3_LZE_A_1299	15.02	11.712693	training
2WTV_ZZL_A_1390	16.48	14.356284	training
2WTV_ZZL_B_1390	16.48	14.618113	training
2XJ2_985_A_1001	14.24	13.576728	training
2Y7J_B49_A_1294	9.74	8.980873	training
2Y7J_B49_B_1294	9.74	10.673288	training
2Y7J_B49_C_1294	9.74	8.560607	training
2Y7J_B49_D_1294	9.74	5.335432	training
2YIW_YIW_A_1353	19.34	12.520471	training
2YIX_YIX_A_1355	17.23	14.783937	training
2ZB1_GK4_A_361	12.25	11.635828	training
3AMA_SKE_A_351	9.22	6.7474203	training
3AMB_VX6_A_351	9.22	11.450221	training
3AT4_CCK_A_336	17.61	16.972866	training
3BWJ_ARX_A_352	12.66	10.589522	training
3D0E_G93_A_1	17.04	12.532833	training
3D0E_G93_B_2	17.04	17.183178	training
3DDQ_RRC_A_299	12.9	8.297859	training
3DDQ_RRC_C_299	12.9	7.178257	training
3E5A_VX6_A_500	17.73	12.699175	training
3EQG_4BM_A_1	14.99	9.964328	training

3FC1_52P_X_362	10.49	10.373807	training
3FLS_FLS_A_361	18.24	13.88539	training
3FLW_FLW_A_361	18.16	16.465137	training
3FSK_RO6_A_450	14.96	13.170166	training
3GCP_SB2_A_361	15.72	12.1016035	training
3GCS_BAX_A_401	5.6	6.2856894	training
3GCU_R48_B_401	13.32	12.8222475	training
3GCV_SS6_A_361	14.12	12.847186	training
3GNI_ATP_B_1	15.43	9.653784	training
3GP0_NIL_A_1	14.84	18.384321	training
3HEC_STI_A_1	9.22	8.287979	training
3HEG_BAX_A_1	5.6	7.008605	training
3HMO_STU_A_1	15.01	12.163776	training
3HP5_52P_A_401	10.49	10.918285	training
3HV6_R39_A_361	12.72	9.466206	training
3IW5_DF3_A_362	11.69	9.562438	training
3IW6_PP0_A_361	10.33	7.345064	training
3IW8_HIZ_A_361	8.92	9.450232	training
3JVS_AGY_A_900	13.44	14.485944	training
3L8S_BFF_A_361	13.83	12.68417	training
3L8X_N4D_A_361	16.12	14.4477	training
3LFA_1N1_A_361	8.22	9.860175	training
3MYG_EML_A_1	22.34	17.777912	training
3NPC_B96_A_365	9.53	10.554703	training
3NPC_B96_B_365	9.53	8.5307665	training
3O8P_BMU_A_361	11.34	10.370246	training
3O8U_BMU_A_361	11.34	9.472248	training
3OBJ_BMU_A_361	11.34	11.657911	training
3PG3_DG7_A_362	11.01	13.260027	training
3PXF_2AN_A_304	7.91	8.43764	training

3PXQ_2AN_A_300	7.91	8.458857	training
3PXQ_2AN_A_301	7.91	7.9644737	training
3PXQ_2AN_A_302	7.91	7.834199	training
3PXZ_2AN_A_299	7.91	6.5781374	training
3PXZ_JWS_A_301	9.74	7.590602	training
3PY1_2AN_A_301	7.91	6.6630545	training
3PY1_2AN_A_302	7.91	7.0281353	training
3PY1_SU9_A_300	13.56	13.19573	training
3RGF_BAX_A_465	5.78	8.478868	training
3SW7_19K_A_299	9.76	12.033982	training
3TZM_085_A_1	6.27	13.889732	training
3UBD_SL0_A_400	10.45	11.020773	training
3UO4_0C0_A_1	12.73	12.905641	training
3UOL_0C7_A_2	15.54	13.293035	training
3UOL_0C7_B_1	15.54	14.602809	training
3VQH_IQB_A_401	15.29	15.187998	training
3VVH_4BM_B_503	14.99	13.505534	training
3VVH_4BM_C_503	14.99	15.487566	training
3ZSH_469_A_400	19.12		training
3ZSI_52P_A_1000	10.49	7.871111	training
4BCQ_TJF_C_1295	13.44	12.84407	training
4BTK_DTQ_A_1337	12.95	9.336148	training
4CRL_C1I_A_1360	20.06	15.915939	training
4DLI_IRG_A_401	10.64	9.909659	training
4DLI_IRG_A_402	10.64	11.62527	training
4EK6_10K_A_301	9.92	10.227652	training
4EZ7_2AN_A_302	7.91	7.93563	training
4EZ7_2AN_A_303	7.91	6.879238	training
4F9Y_GG5_A_401	12.03	14.077505	training
4F9Y_GG5_A_402	12.03	12.441417	training

4F9Y_LM3_A_403	13.21	10.422124	training
4FKI_09K_A_301	7.45	10.119772	training
4FKL_CK2_A_300	8.11	7.9371223	training
4FKO_20K_A_301	7.84	8.178918	training
4FKU_60K_A_301	12.39	9.98849	training
4FKU_60K_A_303	12.39	14.65224	training
4GUE_QCT_A_401	9.76	10.720871	training
4I3Z_ADP_A_301	0.67	5.0668316	training
4I3Z_ADP_C_301	0.67	6.001342	training
4I5M_R78_A_401	18.65	21.034546	training
4JBQ_VX6_A_501	17.73	14.951972	training
4KS8_B49_A_701	3.73	7.098278	training
4L9I_8PR_A_601	8.95	12.205049	training
4L9I_8PR_B_601	8.95	10.86388	training
4LOO_SB4_A_401	16.48	16.590876	training
4LOP_SB4_A_401	16.48	16.209667	training
4LOP_SB4_B_401	16.48	14.673704	training
4LOP_SB4_C_401	16.48	14.650061	training
4LOP_SB4_D_401	16.48	14.583004	training
4LOQ_SB4_A_401	16.48	15.305688	training
4LOQ_SB4_B_401	16.48	12.504713	training
4LOQ_SB4_C_401	16.48	13.820618	training
4OTI_MI1_A_1001	13.13	12.5323515	training
4QMN_DB8_A_401	11.62	14.84667	training
4QMZ_B49_A_401	5.69	6.55925	training
4QP2_36R_A_401	4.67	7.273883	training
4QTA_38Z_A_411	15.86	13.754637	training
4QTB_38Z_A_418	13.99	12.003158	training
4QTB_38Z_B_412	13.99	11.510162	training
4QTE_390_A_430	17.73	13.258313	training

4QYY_3G7_A_401	14.51	10.219946	training
4TXC_38G_A_301	12.95	11.588561	training
4U43_3D8_A_401	7.65	5.3566923	training
4X21_3WH_A_501	14.84	11.492774	training
4X21_3WH_B_501	14.84	13.035936	training
4XX9_RF4_A_402	9.44	8.770052	training
4Y8D_49J_A_401	16.24	11.525299	training
4Y8D_49J_B_401	16.24	15.099169	training
4ZJI_4OQ_B_601	7.83	7.283536	training
4ZJI_4OQ_C_601	7.83	9.089074	training
4ZJI_4OQ_D_601	7.83	6.8545666	training
4ZJJ_4OR_A_601	7.83	7.0933084	training
4ZJJ_4OR_B_601	7.83	8.350812	training
4ZJJ_4OR_C_601	7.83	5.1687207	training
4ZJJ_4OR_D_601	7.83	7.2560215	training
5AJQ_DB8_A_800	16.48	14.70757	training
5AJQ_DB8_B_800	16.48	14.7908	training
5AUT_2AN_A_301	9.62	8.227689	training
5CS6_K82_A_404	5.07	4.468193	training
5CS6_K82_A_405	5.07	5.1155643	training
5CS6_K82_A_406	5.07	4.588729	training
5CS6_K82_A_407	5.07	4.937118	training
5CSH_54E_A_401	5.92	5.2921076	training
5CSH_54E_A_402	5.92	7.947792	training
5CSH_54E_B_403	5.92	5.49813	training
5CSH_54E_B_404	5.92	6.6865587	training
5CSP_54G_A_401	7.46	6.192384	training
5CU3_54S_A_404	12.66	16.073565	training
5CU3_54S_B_403	12.66	9.69426	training
5CU4_54S_A_404	12.66	12.544091	training

5DN3_5DN_A_402	10.19	7.8209085	training
5DR9_SKE_A_401	6	7.5168114	training
5DRB_5FJ_A_501	17.11	14.781893	training
5DT0_SKE_A_401	6	5.988515	training
5JQ5_I74_A_302	10.67	13.649733	training
5L4Q_LKB_A_401	14.46	12.602347	training
5L4Q_LKB_B_401	14.46	14.451378	training
5MO8_C98_A_404	11.02	11.826718	training
5MO8_C98_B_401	11.02	11.957595	training
5MOD_86L_A_404	5.81	5.3589926	training
5MOE_OQC_A_409	5.3	4.6797047	training
5MOE_OQC_A_410	5.3	5.384391	training
5MOE_OQC_A_411	5.3	6.2392187	training
5MOE_OQC_B_409	5.3	6.606739	training
5MRB_C5N_A_901	12.27	16.349714	training
5MTX_FJI_A_401	15.55	13.412746	training
5MTY_HB9_A_401	18.16	13.793211	training
5TBE_78L_A_401	16.94	19.41852	training
5TE0_XIN_A_401	14.28	14.470296	training
5TF9_7AV_A_501	10.04	7.720195	training
5VC3_DB8_A_601	14.08	16.318983	training
5VC4_XZN_A_601	14.65	10.742605	training
5VC5_96M_A_601	16.03	12.171665	training
5VC6_P48_A_601	15.82	14.553338	training
5VCV_1N1_A_404	15.72	9.767446	training
5VCW_93J_A_401	4.51	5.9813185	training
5VCW_93J_B_401	4.51	4.4235907	training
5VCZ_XZN_A_401	12.33	13.311826	training
5VD0_8X7_A_401	12.64	13.869585	training
5VD1_P48_A_401	11.53	17.269943	training

5VD3_H8H_A_401	10.48	12.441893	training
1PY5_PY1_A_700	16.82	14.969334	validation
1XH7_R96_A_351	13.42	14.521075	validation
1XH9_R69_A_351	15.59	11.105695	validation
2A4L_RRC_A_300	3.39	6.7256346	validation
2FVD_LIA_A_299	19.12	15.857263	validation
2UZW_SS4_E_1351	16.82	13.931561	validation
2WTV_ZZL_D_1390	16.48	16.501717	validation
3BWJ_ARX_A_351	12.66	10.008118	validation
3GCQ_1BU_A_401	12.59	17.14584	validation
3GCU_R48_A_401	13.32	11.107234	validation
3GI3_B10_A_391	17.84	15.133845	validation
3HMP_CX4_A_1	10.34	7.4939575	validation
3HUB_469_A_361	19.12	14.843432	validation
3HUC_G97_A_362	11.49	10.609373	validation
3LFF_Z83_A_362	13.15	12.004912	validation
3O8T_BMU_A_361	11.34	10.189346	validation
3PXF_2AN_A_305	7.91	9.076647	validation
3PXZ_2AN_A_300	7.91	6.3745623	validation
3SW4_18K_A_299	9.95	14.080001	validation
3U9N_09H_A_301	13.82	11.017376	validation
3UVQ_FS8_A_361	15.59	11.264186	validation
3VVH_4BM_A_703	14.99	12.857493	validation
4BCQ_TJF_A_1296	13.44	12.502231	validation
4BTJ_ATP_B_1338	8.3	10.083336	validation
4KKH_1RQ_A_501	14.6	17.964321	validation
4LOQ_SB4_D_401	16.48	22.440546	validation
4NJ3_2KD_A_301	12.72	11.144922	validation
4QMS_1N1_A_401	3.97	8.428536	validation
4QMU_SKE_A_401	5.63	5.1778917	validation

4ZJI_4OQ_A_601	7.83	5.5979853	validation
5D1J_56H_A_4000	14.79	12.7819	validation
5DPV_SKE_A_402	6	7.86001	validation
5LVL_537_A_401	11.29	10.439565	validation
5MOE_OQC_B_408	5.3	6.949047	validation
5V5Y_8X7_A_601	15.83	15.258941	validation
5VCY_DB8_A_401	12.83	14.903077	validation
1KE9_LS5_A_299	10.95	10.35851	test
1XH6_R94_A_351	14.55	14.093109	test
2BAJ_1PP_A_401	17.04	12.874502	test
2BAL_PQA_A_401	12.23	12.644074	test
2WTV_ZZL_C_1392	16.48	13.081798	test
2XJ1_XJ1_A_1307	15.09	12.160637	test
3FLN_3FN_C_361	20.04	13.301807	test
3HRF_P47_A_1374	9.19	9.049356	test
3HV7_1AU_A_361	15.94	10.493131	test
3LFE_Z84_A_361	12.05	12.960303	test
3TI1_B49_A_299	9.22	9.492598	test
4EK8_16K_A_301	10.04	11.950938	test
4FKP_LS5_A_301	10.95	11.706519	test
4FKW_62K_A_301	13.95	14.523321	test
5TCO_79Q_A_401	17.28	12.303295	test
5TF9_7AV_B_501	10.04	10.4065275	test

Acknowledgements

AA acknowledges funding awarded by the Indian Academy of Sciences, Bangalore (2019). BM would like to acknowledge Tata Trusts-TDU Fellowship for PhD awarded to her from 2017 to 2019. All authors acknowledge NCBS for infrastructural support.

Conflict of interest

The authors declare no conflict of interest.

Author Contributions

Conceptualization: Bhavika Mam, Asad Ahmed
Data curation: Asad Ahmed, Bhavika Mam
Formal analysis: Asad Ahmed, Bhavika Mam
Funding acquisition: Ramanathan Sowdhamini.
Investigation: Asad Ahmed, Bhavika Mam
Methodology: Asad Ahmed, Bhavika Mam
Project administration: Ramanathan Sowdhamini
Resources: Ramanathan Sowdhamini
Supervision: Ramanathan Sowdhamini
Validation: Asad Ahmed, Bhavika Mam
Visualization: Asad Ahmed
Writing – Asad Ahmed, Bhavika Mam
Writing – review & editing: Ramanathan Sowdhamini

Author ORCIDs

1. Asad Ahmed : https://orcid.org/0000-0003-3775-9320
2. Bhavika Mam : https://orcid.org/0000-0002-3130-0925
3. Prof. Ramanathan Sowdhamini : https://orcid.org/0000-0002-6642-2367