EnCPdock: a web-interface for direct conjoint comparative analyses of complementarity and binding energetics in inter-protein associations

Gargi Biswas[1], Debasish Mukherjee[2], Nalok Dutta[3], Prithwi Ghosh[4], Sankar Basu[5]*

[1]Department of Chemistry and Structural Biology, Weizmann Institute of Science, Rehovot 76100 Isreal; gargi.biswas@weizmann.ac.il
[2]Institute of Molecular Biology gGmbH (IMB), Ackermannweg 4, 55128 Mainz, Germany; d.mukherjee@imb-mainz.de
[3]Bioproducts, sciences and engineering laboratory, Washington State University, WA 99354, USA; nalok.dutta@wsu.edu
[4]Department of Botany, Narajole Raj College, Vidyasagar University, Midnapore 721211, India; prithwi11@gmail.com
[5]Department of Microbiology, Asutosh College, University of Calcutta, 92, Shyama Prasad Mukherjee Rd, Bhowanipore, Kolkata 700026, India; sankarchandra.basu@asutoshcollege.in

* Correspondence: nemo8130@gmail.com, sankarchandra.basu@asutoshcollege.in

Abstract
Protein-protein interaction (PPI) is a key component linked to virtually all cellular processes. Be it an enzyme catalysis (*'classic type functions'* of proteins) or a signal transduction (*'non-classic'*), proteins generally function involving stable or quasi-stable multi-protein associations. The physical basis for such associations is inherent in the combined effect of shape and electrostatic complementarities (Sc, EC) of the interacting protein partners at their interface. While Sc is a necessary criterion for inter-protein associations, EC can be favorable as well as disfavored (e.g., in transient interactions). Estimating equilibrium thermodynamic parameters ($\Delta G_{binding}$, K_d) by experimental means is costly and time consuming, thereby opening windows for computational structural interventions. Attempts to empirically probe $\Delta G_{binding}$ from coarse-grain structural descriptors (primarily, surface area based terms) have lately been overtaken by physics-based, knowledge-based and their hybrid approaches (MM/PBSA, FoldX etc.) that directly compute $\Delta G_{binding}$ without involving intermediate structural descriptors. Here we present EnCPdock (www.scinetmol.in/EnCPdock/), a user-friendly web-interface for the direct conjoint comparative analyses of complementarity and binding energetics in proteins. EnCPdock returns an AI-predicted $\Delta G_{binding}$ computed by combining complementarity (Sc, EC) and other high-level structural terms, and, renders a prediction accuracy comparable to the *state-of-the-art*. EnCPdock further locates a PPI complex in terms of its {Sc, EC} values (taken as an ordered pair) in the two-dimensional Complementarity Plot. In addition, it also generates mobile molecular graphics of the interfacial atomic contact network. Combining all its features and applications, EnCPdock presents a unique online tool that should be beneficial to structural biologists and researchers across related fraternities.

Keywords: protein-protein interactions (PPI), complementarity, Complementarity Plot (CP), free energy of binding ($\Delta G_{binding}$), support vector regression (SVM) machines, EnCPdock

1. Introduction

A host of biological processes thrives on the efficacy of protein-protein interactions (PPI) ranging from agglutination reactions to aggregation-based processes initiating the immune response cascade [1]. PPI plays a key role in cellular function. Hence, understanding the bio-physical chemistry underlying protein-protein association is helpful to decipher cellular functions. Till date, a significant amount of progress has been made towards generating PPI networks by utilizing high-throughput proteomic techniques (yeast two-hybrid assays and mass spectroscopy). Many biologically important interactions occur in transient complexes, and are difficult to determine experimentally, even when the structures of the component proteins are known. This has prompted an increasing interest in computational methods, especially protein-protein docking, for the structural prediction of macromolecular assemblies, to complement the experimental approaches in several ways. A deep understanding of the integral protein structure is quintessential in deciphering the underlying forces for structural analysis of the macromolecular complex. Even, to date, in spite of the efforts taken by the researchers all across the scientific community, a large chunk of the protein moieties are not enlisted in the PDB [2]. Associations among proteins are determined by several physical forces involving (for example) van der Waals interactions, electrostatic interactions, hydrogen bonds, exclusion of solvent, and entropic changes. The dynamic association between interacting partners also involve both enthalpic (ΔH) and entropic component ($-T\Delta S$) thus the recognition is accompanied by structural and dynamic changes in both the interacting partners. Protein–protein binding occurs, like any spontaneous process, only when the change in Gibbs free energy (in the direction of association, i.e., $A+B \rightarrow AB$) of the system ($\Delta G_{binding}$) at constant pressure and temperature is negative.

Various studies have linked protein-ligand and protein-protein associations with changes in heat capacity based on high resolution crystal structures. Hydrophobic interaction causes large changes in heat capacity, which is one of the most notable observations. Molecular interactions known as hydrophobic interactions involve desolvation of non-polar protein atoms and packing of those atoms within proteins or protein complexes isolated from the aqueous environment. In this way, the hydrophobic contribution to ΔCp can be efficiently estimated by observing the buried non-polar solvent-accessible area during complexation or folding [3]. The dissolution of hydrophobic model compounds in water has led to several empirical formulations. Structural biology has been grappling with the problem of building models for protein complexes for years, but the problem remains unsolved. The fundamental agenda can be split into sampling and scoring to identify the correct docking pattern [4,5]. Each approach provides a unique aspect of docking model quality, and when combined (into hybrid scoring functions), can lead to more accurate docking results [6]. The complexity of the pattern of scoring protein docking models have led to a number of viable approaches with shape complementarity being a determining factor for analyzing complex formation [7].

Hydrogen bonds, electrostatic interactions, van der Waals forces constitute the non-covalent interactions between PPI complexes for stabilizing them. For determining binding constants, surface plasmon resonance is commonly used [8,9]. A metal film supports the surface plasma in this method, and the refractive index of the contact surface is measured to detect interactions between bio-molecules. Another route to determine the binding free energy exchange between ligands and proteins is via FRET and isothermal titration calorimetry [10,11]. PPI binding free energies can also be evaluated using computational approaches. Compared to free energy perturbation methods spanning thermodynamic integration, it is much more efficient to use molecular mechanics combined with their Poisson–Boltzmann electrostatics and accessible surface area estimates (MM/PBSA) and generalized Born surface area (MM/GBSA) because of the lowered costs incurred

in the processes [12]. MM/PBSA and MM/GBSA can be used to decompose energy contributions originating from different types of interactions and molecular groups; therefore, these methods are widely employed in protein-ligand / protein-drug interactions and PPIs [13]. MM/GBSA performs better than the tested scoring functions not only at predicting binding affinities but also at identifying near-native binding modes for PPIs [14].

One of the primary determinant of biological assemblies (such as protein-protein complexes) is shape complementarity of the interacting molecular partners. There are reports of many computational docking approaches using shape complementarity (Sc) [7] as a primary guide. In parallel, the use of electrostatic complementarity (EC) [15] has also found its scope in several computational endevours, e.g., in docking, homology modeling, structure validation etc [16]. Together, Sc and EC render a dual-nature of (short and long range) complementarity in (inter and intra) protein associations, wherein, binding and folding can be discussed from a common conceptual platform [17]. The current paper presents a unique user-friendly web-interface, namely, EnCPdock (www.scinetmol.in/EnCPdock/) that allows direct concurrent comparative analyses of complementarity as well as binding free energetics of PPI complexes. An AI-predicted $\Delta G_{binding}$ is returned by EnCPdock, determined by combining complementarity (Sc, EC), surface area estimates and other high-level structural descriptors taken as input feature vectors. EnCPdock is effectively an extension of CPdock (available only as a standalone github package https://github.com/nemo8130/CPdock) [18] – which was developed to work primarily as an initial filter in protein docking scoring pipelines. From the Complementarity Plot (CP), CPdock measures indirect probabilistic stability estimates of an input (docked / bound) PPI complex based on its {Sc, EC} values (taken as an ordered pair). Over and above CPdock, the current web-interface (EnCPdock) goes deeper into protein binding energetics and predicts the actual free energy of interaction ($\Delta G_{binding}$) using a non-linear support vector regression (SVM) machine trained on a collection of wisely chosen high-level structural descriptors. As a free energy predictor, EnCPdock performs comparable to the *state-of-the-art* – as revealed from both cross- and independent validations. In addition, EnCPdock returns the usual mapping of the binary PPI complex in the Complementarity Plot (as in CPdock) and also returns mobile molecular graphics of the atomic contact network probed at the protein-protein interface. Hence, combining all different features, it presents an unique online tool for structural biologists and researchers to explore across related fraternities.

2. Materials and Methods

2.1. Datasets

The primary database was build from scratch for training and cross-validation of EnCPdock. To that end, first, protein binary complexes were collected from the Protein Data Bank (PDB) using its advanced search options with the following culling criteria:

i) Each target molecule must be a protein-protein dimer complex (homo- and hetero-dimers) and not a single protein.

ii) All atomic models must be determined either by X-ray crystallography or by cryo Electron Microscopy.

iii) All structures had to have a resolution of ≤ 2Å.

iv) None of the structures had RNA/DNA chains.

v) For X-ray structures, they had to have a R-observed ≤ 0.2.

vi) None of the structures contained large prosthetic groups or co-factors (i.e., a non-protein chemical component with MW > 200 Da), e.g., proto-porphyrin rings, NADP, acetyl CoA, etc.

vi) Each complex should have had at least one residue coming from each partner at their interface buried upon association.

This led to 6658 structures (asymmetric units) which got further reduced to 3200 non-redundant binary PPI complexes upon implementation of a sequence identity cutoff (in NW-align: https://zhanggroup.org/NW-align/) of no more than 30% between any two receptor chains and separately for any two ligand chains. The PDB IDs along with the corresponding chain IDs of receptors and ligands of PPI complexes pertaining to the final training dataset are to be found enlisted in **Dataset S1, Supplementary Materials**.

For independent validation purposes, all binary PPI complexes (entirely non-overlapping to the training dataset) were culled from the updated version of the Affinity benchmark (v.2) [19]. All these structures had annotated experimental binding affinities and/or free energies (K_d or $\Delta G_{binding}$) associated with each entry. Structures with reported suboptimal accuracy in terms of binding energetics (https://github.com/haddocking/binding-affinity-benchmark) were further discarded. This led to a final size of 106 binary PPI complexes (**Dataset S2**, **Supplementary Materials**) for the dataset used in independent validation.

2.2. Built of EnCPdock

2.2.1. The input feature vectors

Six input feature vectors that are effectively high-level (fine-grained) structural descriptors of the overall protein complex or the protein-protein interface have been utilized for training of EnCPdock. The structural features are described as below.

2.2.1.1. Surface or shape complementarity (Sc)

Qualitatively, the shape complementarity (Sc) function measures the goodness of fit between two surfaces, and in case of PPI complexes this function serves as an essential condition for binding. For calculation of Sc, van der Waal's surfaces of two interacting partners (e.g., protein-protein, protein-DNA or residue-residue) is generated, and for each pair of points from the respective surfaces, the distance between points and the dot product between the corresponding normals are taken into account [20]. The elaborate mathematical description of shape complementarity function can be obtained from the work of Lawrence and Colman [7] whose value ranges from -1 to +1 for an anti- and perfectly correlated surfaces respectively. The Sc value of a protein-protein interface was computed by the original program 'sc' (© Lawrence), distributed as a part of the CCP4 package [21].

2.2.1.2. Electrostatic complementarity (EC)

The complementarity of charge distribution upon two interacting surfaces is measured by electrostatic complementarity (EC) function. Similar to Sc, EC is also a necessary determinant for binding. For the calculation of EC, the molecular surface is generated first using the software EDTSurf [22], and then using the DelPhi software [23] the electrostatic potential on each surface point at the protein-protein interface was determined. The atoms who have undergone a change (non-zero) in the solvent accessible surface area (ASA) upon binding, have been identified as the atoms at the protein-protein interface (the difference in ASA is calculated using NACCESS [24]. For the calculation of electrostatic potential, the partial charges and atomic radii were assigned from AMBER94 molecular mechanics force field [25], and the solution was obtained by iteratively solving a linearized version of Poission-Boltzman equation (employing DelPhi). Similar to Sc, EC value also ranges from -1 to +1, where approaching the value 1 signifies the anti-correlation of electrostatic potentials between two interacting surfaces, and negative EC implies unbalanced electric fields at the interface.

2.2.1.3. Normalized buried surface area (nBSA)

Although Sc and EC are essential measures for binding, but they do not consider the area or size of the interface of PPI complexes. So, in order to incorporate that feature, two additional measures have been taken into account. For any atom (say, i^{th} atom) at the interface, for which there is a non-zero change in the ASA (as mentioned previously), the buried surface area has been calculated using the following formula,

$$BSA(i) = \Delta ASA(i) = |\Delta ASA_{bound}(i) - \Delta ASA_{unbound}(i)| \quad (1)$$

- where ΔASA_{bound} stands for the accessible surface area (ASA) at the bound state and $\Delta ASA_{unbound}$ is the ASA at the unbound state. Now, for the calculation of nBSA, for a protein complex consisting of two proteins (say, protein A and protein B) the following formula have been used,

$$nBSA = \frac{\sum_{i=1}^{A} \Delta ASA(i) + \sum_{i=1}^{B} \Delta ASA(i)}{\sum_{j=1}^{A+B} ASA(j)} \quad (2)$$

- where $\sum_{i=1}^{A} \Delta ASA(i)$ and $\sum_{i=1}^{B} \Delta ASA(i)$ are sum over the ΔASA of all the interface atoms of protein A and protein B, and $\sum_{j=1}^{A+B} ASA(j)$ is the total ASA of all atoms of protein A and protein B. For calculation of nBSA, all the interface atoms, irrespective of their polarity have been taken into account.

Further, considering the polarity of the interface atoms, nBSA has been calculated separately for polar (N and O) and non-polar (C and H) atoms [3] and the corresponding nBSA have been called as $nBSA_p$ and $nBSA_{np}$ respectively.

2.2.1.4. F_{intres}

Similar to nBSA, the fraction of residues buried at the interface of a PPI complex (consisting of two proteins, say protein A and protein B) have been defined by the following formula,

$$F_{intres} = \frac{N_{intres}}{N_{tot}} \quad (3)$$

- where N_{intres} and N_{tot} are respectively the number of interfacial residues and total number of residues pertaining to the protein-protein complex.

In addition to the complementarity and accessibility based structural descriptors, several precise network parameters have also been included as input feature vectors. Given these are essentially binary (receptor – ligand) protein-protein interactions, the resultant PPI networks at the interconnected interface would necessarily map to bipartite graphs. Importantly, two residues at the interfaces from each of the partner molecules (receptor and ligand) are connected with a link if at least one non hydrogen atom from both the residues are within 4 Å from each other. Notably, the distance cut-off for C-C van der Waal's contact is 3.8 Å, thus, threshold for formation of a link (i.e.,

4 Å) applied here is reasonably stringent [26,27]. However, after defining the criteria for formation of an inter-residue inter-chain link, subsequent construction of bipartite network adjacency matrices was derived. In connection with these, a few network-based features were designed as described as below.

2.2.1.5. Link density (Ld)

Let the receptor and the ligand have N1 and N2 interfacial residues in physical contact with one or more residues coming from the partner molecules for each. The link density for such a bipartite network could then be defined as follows,

$$Ld = \frac{N_{icnt}}{N1 \times N2} \tag{4}$$

- where N_{icnt} is the number of inter-chain inter-residue contacts to have formed at the said receptor – ligand interface. In short, Ld can be defined as the ratio of the actual number of links to the theoretical maximum number of links between two chains and, it's value will range from 0 to 1, where a complete bipartite graph (i.e., every node of the receptor is connected to every node of ligand and vice-versa) would hit a value of 1.

2.2.1.6. Average contact intensity (ACI)

As defined in an earlier work from this laboratory in context to salt-bridges [27], ACI accounts for the individual intensities or atomic contact densities of each inter-residue interchain link to have formed in a PPI bipartite network. To that end, the actual number of inter-atomic contacts [atcon(i)] involved in each (i^{th}) inter-residue interchain link was averaged over all links [28].

$$ACI = \frac{\sum_{i=1}^{N_{icnt}} atcon(i)}{N_{icnt}} \tag{5}$$

2.2.1.7. Degree distribution profile-based features

Degrees of each node constituting these PPI bipartite networks were computed from their corresponding (unweighted) 1-0 adjacency matrices and their normalized frequency distributions were plotted in a $\log_{10}$-scale for both axes. Given their characteristic power law distributions, the obtained points in these log-log plots for each graph would follow linear relationships. To that end, a linear least-squares fitting was performed on the obtained points (for each graph) from which the slope (m), the Y-intercept (c) and the Pearson's correlation coefficient (r) between the observed (Y_{obs}) and the expected abscissa (as in, $Y_{exp} = mX_{obs} + c$) coming from the plotted points were analytically determined. The r value (let's call it CCp_{dd}) between Y_{exp} and Y_{obs} [r (Y_{exp},Y_{obs})] were determined using the well-known relation [29],

$$CCp_{dd} = r(Y_{\exp}, Y_{obs}) = \frac{cov(Y_{\exp}, Y_{obs})}{\sigma(Y_{\exp})\sigma(Y_{obs})} \tag{6}$$

- where cov is the co-variance between the mentioned parameters within the bracket and σ is the variance of each parameter.

2.2.1.8. Features based on comparative and absolute size of the interaction partners

The chain length of each interacting molecular partner (lenR: receptor, lenL: ligand) was taken as the estimate of their size. Two features were then defined based on these chain lengths as follows,

$$logN = \log_{10}(lenR + lenL); \log_{asp} = \log_{10}\left(\frac{lenR}{lenL}\right) \quad (7)$$

The second of these terms can be considered analogous to the aspect ratio (in log scale) of a crystal or a nano particle. For identical sizes of receptor and ligand (e.g., rare homo-dimeric biological assemblies) $\log_{asp}$ would hit a value of 0. The primary purpose of taking logarithm for these (comparative and absolute) chain length features was to render values in a desired (low numeric) range comparable to that of the other SVM input feature vectors.

2.2.2. Support Vector Regression (SVM) training & parametrization

For the training of EnCPdock, the Support Vector Regression module in the SVM^{light} package [30] have been utilized using 10-fold cross validation with a radial basis function (RBF) kernel.

2.2.2.1. Cross validation test-set and parameter optimization

All the 3200 PPI complexes were divided into ten equal parts, nine of which were utilized for each round of training, while the remaining ones were used as the test set in that round. This was performed for each of the ten parts to get predictions for the whole set. Three parameters, namely, C, γ and ε have been optimized using a grid search method to obtain a maximum correlation (determined by Pearson's correlation coefficient, r) between target function and the predicted output. Among the three parameters, C is the penalty value associated to the training error (for large and small value of C, a smaller and larger margin were accepted respectively if all the training points were correctly predicted); RBF γ-value defines how far the influence of a single training example reaches (low value indicates far and high value indicates close); and ε is the distance of the target function from the actual value within which no penalty is associated in the training loss function [31]. Optimization of C, γ and ε was performed within the ranges 1.0 to 5.0 (in steps of 0.5), 0.01 to 2.0 (in steps of 0.01) and 0.1 to 1.0 (in steps of 0.05) respectively.

2.2.2.2. Target function and predicted output

During both the training and cross-validation procedure, $\Delta G_{binding}$ was predicted by FoldX [32] which is a well-accomplished empirical force-field based method parameterized by empirical data from actual protein engineering experiments [33] and is known to return near-native binding/affinity estimates. The FoldX platform is based on a 'fragment-based strategy' which uses fragment libraries in an analogous way to the most fascinating 'fragment assembly simulated annealing' approach for protein structure prediction credited to David Baker and Rosetta [34]. The standalone (C++ with boost library) version (v.4) of FoldX (http://foldxsuite.crg.eu/) was used. The FoldX-derived $\Delta G_{binding}$ (let's call it ΔG_{FoldX}) was normalized by the number of interfacial residues, N_{intres} (let's call it $\Delta G_{FoldX_per\text{-}res}$ denoting per-residue contribution to binding free energy) for all the PPI complexes. This $\Delta G_{FoldX_per\text{-}res}$ was suitable numerically to be used as the target function in the training of the free-energy predictor (in the EnCPdock web-interface) which naturally returned the predicted output interpretable in terms of the per-residue contribution (i.e., normalized) to $\Delta G_{binding}$ (let's call it, $\Delta G_{EnCPdock_per\text{-}res}$). The normalization (of the target function) also ensured similar numerical ranges for both the target function (TF) and the predicted output (PO). In the final output of the EnCPdock web-interface, both per-residue (i.e., normalized) and total binding free energies were returned.

2.2.2. Support Vector Regression (SVM) training & parametrization

For the training of EnCPdock, the Support Vector Regression module in the SVM^{light} package [30] have been implemented using a 10-fold cross validation scheme with a radial basis function (RBF) kernel. The RBF kernel ensured a non-linear relationship between the input feature vectors and the target function ($\Delta G_{EnCPdock_per\text{-}res}$). In a SVM classification, for a given set of n observations given by (x1, y1), …, (xn, yn), predictions are made from the following formula [35].

$$f(x) = sign\left(\sum_{i=1}^{n} y_i \alpha_i K(x_i, x - b)\right) \tag{8}$$

- where b is a numeric offset threshold and α_i denote the weight of each observation, known as Kühn-Tucker coefficient. The Kernel function K (x_i, x) defines a dot product to transform the observations from input space to the feature space. Here, we have used a Gaussian Kernel function, popularly known as an RBF Kernel, given by,

$$K(x, x') = e^{-\gamma \|(x - x')\|^2} \tag{9}$$

- where γ is a parameter (>0) controlling the width of the Gaussian Kernel. γ is a free parameter whose value can be optimized. Apart from γ, two other free parameters have been optimized during the SVM training, namely, C and ε. C is a cost parameter which controls the trade-off between minimizing the number of misclassified samples in the training set and maximizing the margin width; while ε can be defined as the sum of classification error and the complexity penalty given by,

$$\varepsilon = \frac{1}{2n} \sum_{i=1}^{n} |y_i - f(x_i)| + \lambda \left(\frac{n_{sv}}{n}\right)^{\Gamma} \tag{10}$$

- where the first term is the classification error and, n_{sv}, λ and Γ are number of input feature vectors, regularization parameter and the penalty respectively. During the SVM training and cross validation, these three parameters c, γ and ε have been optimized using a grid search method to obtain a maximum correlation (determined by the Pearson's correlation coefficient, r, eqn. 11) between target function and the predicted output. As detailed earlier (see section **2.2.2.2**, **Materials and Methods**), for each individual training cycle, the target function was taken to be $\Delta G_{FoldX_per\text{-}res}$ while the predicted output was the $\Delta G_{EnCPdock_per\text{-}res}$ for the test set predicted by the rest of the nine subsets. Thus, for a well-trained predictor, a high correlation should be expected in the performed cross-validation between the predicted values of $\Delta G_{binding_per\text{-}res}$ by EnCPdock with reference to FoldX predictions in the same test set. However, a more authentic indicator of the same would be the recovery of performance of the trained predictor in an independent validation performed on a separate dataset of PPI complexes (completely non-overlapping to that of the training set) with their annotated experimental binding free energies taken as references.

$$r(\Delta G_{FoldX_per-res}, \Delta G_{EnCPdock_per-res}) = \frac{cov(\Delta G_{FoldX_per-res}, \Delta G_{EnCPdock_per-res})}{\sigma(\Delta G_{FoldX_per-res})\sigma(\Delta G_{EnCPdock_per-res})} \tag{11}$$

To obtain a maximum r value during the training the c, γ and ε were optimized within the ranges 1.0 to 5.0 (in steps of 0.5), 0.01 to 2.0 (in steps of 0.01) and 0.1 to 1.0 (in steps of 0.05) respectively.

2.2.3. Evaluation metrics

To evaluate the performance of an SVM model on test predictions, firstly, the number of true positives (TP), true negatives (TN), false positives (FP) and false negatives (FN) were quantified. Based on these measures, the performance of the predictor (EnCPdock) was estimated by the following evaluation metrices.

First, the actual positives and actual negatives that were correctly identified out of all predicted positives and predicted negatives respectively were determined by the true positive rate (TPR) and false positive rate (FPR), as defined below.

$$TPR = \frac{TP}{P} = \frac{TP}{TP + FN} \tag{12}$$

$$FPR = \frac{FP}{N} = \frac{FP}{FP + TN} \tag{13}$$

- where P is the total number of positives, consisting of true positives and false negatives while N is the total number of negatives consisting of false positives and true negatives. The positive and negative cases were identified based on various cut-offs spanning the entire range of the target function (i.e., $\Delta G_{FoldX_per\text{-}res}$). Further, the receiver operating characteristics curves (ROC) were plotted (TPR vs. FPR) and the Area Under the Curve (ROC-AUC) were computed. For a random predictor, the AUC will exhibit a value of 0.5 and a perfect prediction model would hit a value of 1.0 [36]. In parallel, the balanced accuracy (BACC) score and Matthew's correlation coefficient (MCC) were also computed, defined as follows.

$$BACC = \frac{1}{2}(TPR + TNR) \tag{14}$$

$$MCC = \frac{(TP \times TN) - (FP \times FN)}{\sqrt{(TP + FP)(TP + FN)(TN + FP)(TN + FN)}} \tag{15}$$

- where TPR and TNR are the true positive rate and true negative rates respectively. TNR can be defined similar to TPR by the following formula,

$$TNR = \frac{TN}{N} = \frac{TN}{FP + TN} \tag{16}$$

To note, since BACC score is the arithmetic mean of TPR and TNR, swapping the definition of TPR and TNR does not alter the BACC score.

3. Results and Discussion

The objective of the current study was to develop an integrated web-interface (EnCPdock) to analyze complementarity and energetics of protein – protein interactions, serving as a common conceptual platform, that could be of practical use to researchers across the modern biology fraternity. The potential beneficiaries include experimental as well as computational structural biologists, and, also other members of the broad biological community working with specific protein – protein interactions involved in enzyme catalysis, signal transductions and/or other protein functionalities. The energetics part of EnCPdock has been based on supervised learning, hitting

comparable (if not better) accuracy to predict binding free-energies ($\Delta G_{binding}$) of PPI complexes from the knowledge of fine-grained high-level structural descriptors. PPI complexes are probed by EnCPdock in terms of its {Sc, EC} values (taken as an ordered pair) in the two-dimensional Complementarity Plot (as in CPdock [18]) evaluating indirect probabilistic estimates of stability. The uniqueness of our approach lies in the operational credibility of the EnCPdock system, which is a comprehensible web-interface (www.scinetmol.in/EnCPdock/) that permits direct, synchronized comparisons of complementarity and binding energies of PPI complexes. The performance of EnCPdock has been evaluated and authenticated in a two-step validation pipeline: (i) firstly, by a 10-fold cross validation performed on a large non-redundant training set consisting of binary PPI complexes with structural coordinates alone (irrespective of having experimental references to their $\Delta G_{binding}$ values) and (ii) backed up by an independent validation performed on one of the latest non-redundant datasets of binary PPI complexes [19] with recorded experimental $\Delta G_{binding}$ values. The independent validation clearly shows recovery of performance with respect to the cross-validated results – which is comparable (if not better) to the *state-of-the-art*. In addition, EnCPdock not only returns the mapping of the binary PPI complex in the Complementarity Plot but also further returns mobile molecular graphics of the atomic contact network probed at the protein-protein interface, along with the detailed lists of weighted inter-residue contacts for further analyses for interested parties.

3.1. SVM Training and Cross-validation

As described in the **Materials and Methods** (section **2.2.1**), thirteen high-level structural descriptors were taken as input feature vectors to build and train the binding free-energy predictor in EnCPdock. These features together may be considered to be harboring necessary and sufficient fine-grained structural details of the PPI complexes. The first and foremost criteria for macro-molecular binding lie in the complementarity in shape and surface electrostatic potential {Sc, EC} of the interacting molecular surfaces [37] which were the first two features in EnCPdock. As a matter of fact, both of these measures are surface-derived [7,15] offering high-level of sophistication in the structural description of the (target) complexes. The shape of both the molecular partners should be such that the side-chain atoms are engaged in complementary interlocking at the interface for a feasible interaction (Lawrence and Colman, 1993; Basu et al., 2012). Thus, a high value of shape complementarity (Sc) serves as an essential criteria for macro-molecular binding [20,38]. Similarly, an optimal electrostatic interaction between two interacting surfaces plays a significant role in binding of two molecular partners, hence the correlation in electrostatic surface potentials (EC) between the interacting surfaces of partner molecules should also be taken into consideration [39,40]. However, there are instances of sub-optimal or even negative EC values between the surfaces of a statistically significant fraction of PPI complexes [~ 20% cases as described in (Basu & Wallner, 2016b) which is complemented by an elevated geometric fit between the corresponding surfaces. Consequently, EC serve as a secondary or sufficient criteria for protein-protein interactions while Sc acts as a primary or necessary criteria for binding (Basu et al., 2021b). Moreover, the choice of both Sc and EC as input feature vectors accounts for the dual nature (i.e., both local (Sc) and non-local (EC)) of complementarity of the interacting molecular partners.

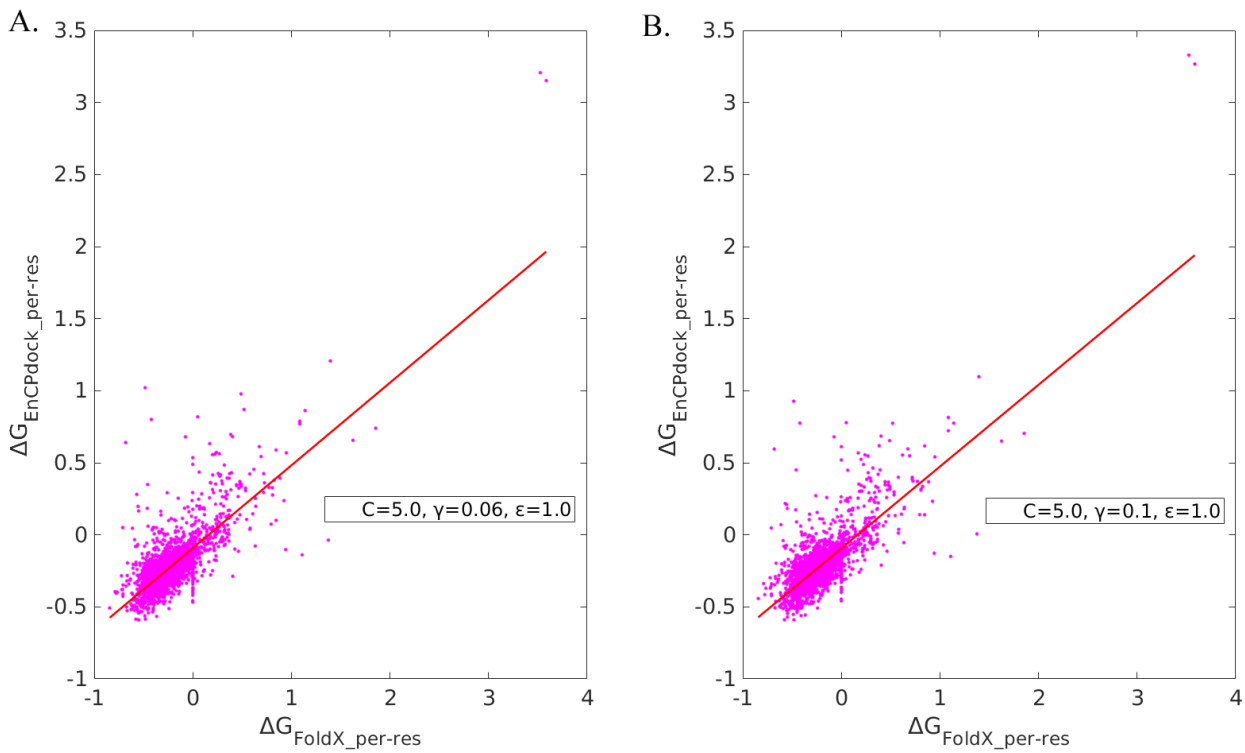
In addition to these, the structural features describing size of the interacting molecules (i.e., receptor and ligand chain lengths) and that of their interface, the amount of change in the solvent accessible surface area upon binding and properties attributed by inter-residue interaction networks formed at the interface also play a determining role (Basu and Wallner, 2016a, 2016b) in the stability of PPI complexes. All these properties were accounted for by administrating eleven other descriptors, discussed in detail in the (see section **2.2.1**, **Materials and Methods**) sections. Due to the scarcity of sufficient high quality experimental binding or affinity data, the training of

EnCPdock had to be performed on FoldX-derived binding free energies ($\Delta G_{binding}$) (van Durme et al., 2011), appropriately normalized to $\Delta G_{binding_per\text{-}res}$ (by the number of interfacial residues (Basu et al., 2021a)) to be taken as target functions for the training set consisting of 3200 non-redundant targets (high resolution PPI complexes) (see section **2.1**, **Materials and Methods**). Consequently, the normalized binding free-energy values could be interpreted as per-residue contribution to the overall binding free energy of the interactions which also matched the numerical ranges to be effectively used as a target function in SVM regression. Given the fact that FoldX is a well accomplished near-native method [33] its utilization ensured a sufficiently large dataset for training, scaling to those of the modern-day supervised learning methods (rather than having to rely on ~100 experimental datapoints).

In order to implement a rigorous training procedure, a large dataset was needed and chosen (as described above) and a ten-fold cross validation was performed during the EnCPdock training – which is a standard technique for training SVM regressors [41–43]. To that end, the whole dataset was first divided into ten equal parts (subsets) yielding a statistically significant number of datapoints (320 PPI complexes) in each subset. During training, each time, the regressor was trained on $9/10^{th}$ of the training dataset and was tested on the leftover ($1/10^{th}$) fraction and this process was repeated for all combinations of the different subsets. One round of complete training in this way results in the attainment of 'predicted output' (PO) for each (target) entry in the training dataset. The predicted outputs (predicted $\Delta G_{binding_per\text{-}res}$) were then compared with their corresponding target functions (FoldX-derived $\Delta G_{binding_per\text{-}res}$) by means of evaluating the Pearson's correlation coefficient (r) on the entire training set (i.e., 3200 PPI complexes). In addition, a threshold (i.e., cutoff) spanning the entire range of the target function was sampled at an interval of 0.025 from its minimum to maximum obtained values for each round of training to convert both the continuous functions (TF, PO) to binary (1/0) variables (see section **2.2.2.2**, **Materials and Methods**) each time leading to an optimized BACC score for that round of training (see section **2.2.3**, **Materials and Methods**). In order to find a suitable model that effectively describes the $\Delta G_{binding_per\text{-}res}$ as a robust function of the structural descriptors, a Gaussian Kernel function (Radial Basis Function or RBF Kernel) was used. The SVM-training (10-fold cross-validation) was performed for a judiciously chosen range of values for the three main parameters part of svm_learn [30] guided by strong literature support: (i) the γ-parameter controlling the width of the RBF Kernel, (ii) the C-parameter presiding over the training error and margin width and (iii) the cost-function (ε). Initially, the SVM training was performed for a broad range of C and γ-values in a grid-search keeping the cost-function fixed (set to 0.1); the performance was tested for C-values between 2^{-15} to 2^{10}, and the γ within the range of 2^{-10} to 2^{10} in $\log_2$ steps [6,35]. The eventual aim was obviously to maximize the cross-validated r. The 'course-grained' search gave us an initial feel of the numbers based on which a more stringent search was invoked around those C and γ values which lead to the maximum r values in the initial search. Thus, in the subsequent fine-grained search, the values of the C, γ and ε were kept in the ranges of 1.0 to 5.0 (in steps of 0.5), 0.01 to 2.0 (in steps of 0.01) and 0.1 to 1.0 (in steps of 0.05) respectively (see section **2.2.2.1**, **Materials and Methods**). Finally, a maximum r value of 0.745 (**Fig. 1**) was obtained with a corresponding maximum BACC score of 0.833. The values of C, γ and cost-function exhibiting the highest BACC score and maximum r values (**Table S1** in **Supplementary Materials**) were then accumulated (closely ranged between 0.744 to 0.745) and considered for independent validations. The corresponding ROC-AUC of TPR vs. FPR plots (see section **2.2.3.1**, **Materials and Methods**) were found to be ~0.75 (**Fig. S1** in **Supplementary Materials**). For each of the combinations of C, γ and ε in **Table S1**, ten models were obtained (pertaining to each cross-validation subset) resulting in a total of 90 models that demonstrated excellent correlations between the target function and the predicted outputs. Thus, a target test PPI complex undergoing EnCPdock predictions will have 90 predicted outputs. The central tendencies (mean, median, mode) along with maximum and minimum values of these

predictions were studied in great detail. Most distributions were not really symmetric, rather than varying from left-skewed, bi-modal, multi-modal to right-skewed (**Fig. S2** in **Supplementary Materials**) which led us to opt for the 'median' as a better measure of central tendency than mean or mode – which was set as the final measure for the predicted $\Delta G_{binding_per\text{-}res}$ in EnCPdock. The finally selected 90 models were then taken forward for subsequent independent validations to eliminate the effect of over-fitting that might have occurred during the course of cross-validation.

Figure 1. Cross-validated correlation (scatter) plots. The predicted output of $\Delta G_{binding_per\text{-}res}$ from EnCPdock ($\Delta G_{EnCPdock_per\text{-}res}$) for 3200 PPI complexes (during cross-validation) plotted against their corresponding per-residue binding free energies predicted by FoldX (or $\Delta G_{FoldX_per\text{-}res}$) for two top-performing training sets trained at particular sets of values for of C, γ and ε (values provided in insets) both of which exhibited BACC scores of 0.833 and r values of 0.745.



3.2. Feature Trends and Contribution

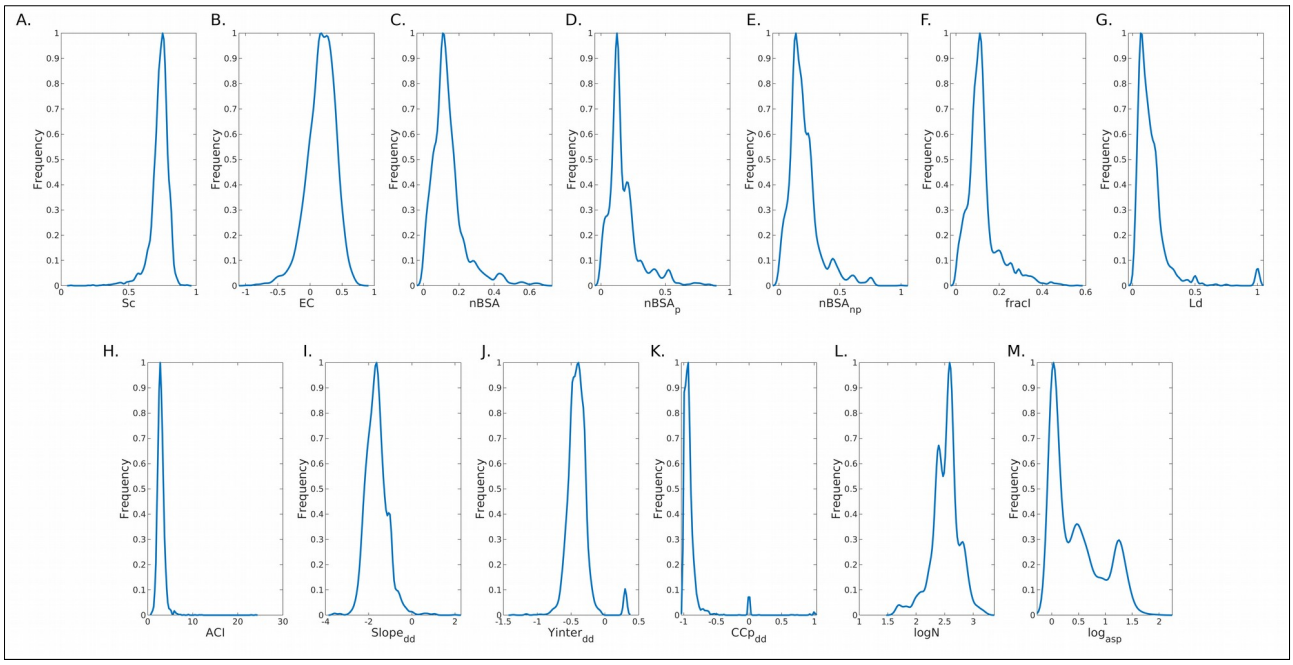
Numeric trends and contributions of individual features in the overall cross-validated performance of EnCPdock were investigated in the training dataset by (i) estimating their descriptive statistics (mean, standard deviations, frequency distributions) followed by (ii) enumerating their (Pearson's) correlations (r) with the predicted $\Delta G_{per\text{-}res}$. Shape Complementarity (Sc), being a necessary criteria (acting like a threshold filter) for inter-protein associations, gave rise to a sharp, narrowly ranged (mildly left-skewed) distribution with a high mean value (0.73; SD: 0.07, **Table 1**) – which, as could be anticipated, is consistent with the short range nature of (van der Waals') force responsible for it. In distinct contrast, electrostatic complementarity (EC), being resulted from a long range (Coulomb's) force has a much broader width with only a modest average (0.18; SD: 0.22, **Table 1**) in the training dataset. This again is consistent with the earlier findings that EC can be favorable as well as disfavored in bound protein complexes [6,18], wherein, the later, less frequent event (in ~20% cases) often signals for transient interactions [44]. Other features gave characteristic distributions, mostly tight and uni-modal, sometimes, with a few low-frequency (outlier-like) disjoint borderline modes (**Fig. 2**). Among all features, ACI has arguably the thinnest

width (or, the sharpest peak) while on the other hand, the chain-length features, logN and $\log_{asp}$ account for large variability in the PPI complexes.

Table 1. Feature trends in terms of descriptive statistics. Descriptive statistics are furnished in terms of mean and standard deviations for all features. Alongside, the statistics for the predicted output ($\Delta G_{per\text{-}res}$) are also tabulated.

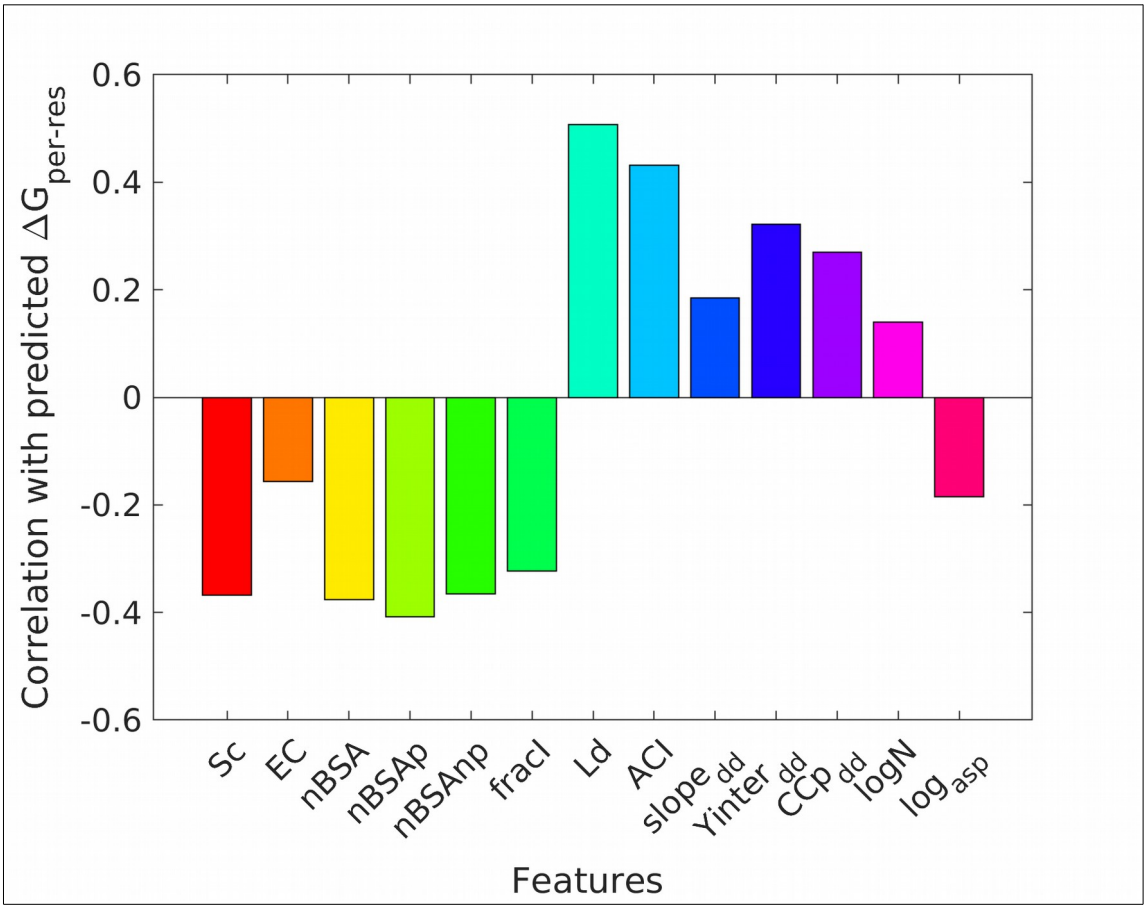
Features	**Mean**	**Standard Deviations**
Sc	0.729	0.069
EC	0.176	0.219
nBSA	0.140	0.098
$nBSA_p$	0.175	0.128
$nBSA_{np}$	0.204	0.128
fracI	0.122	0.076
Ld	0.151	0.148
ACI	2.922	0.926
$slope_{dd}$	-1.633	0.494
$Yinter_{dd}$	-0.405	0.155
CCp_{dd}	-0.895	0.203
logN	2.514	0.244
$\log_{asp}$	0.456	0.476
$\Delta G_{per\text{-}res}$	-0.215	0.233

Figure 2. Trends of Features in terms of their frequency distribution patterns. Normalized frequencies were computed in the observed ranges (X-axes) individually for all features and kernel density distribution plots were constructed based on these normalized frequencies.



Relative contributions of the different features to the overall (cross-validated) performance of EnCPdock were estimated by drawing correlation (scatter) plots of the features (on X-axis) and the predicted output ($\Delta G_{per\text{-}res}$ on Y-axis) for each feature followed by computing the corresponding Pearson's correlations (r) (**Fig. 3**). These correlations are not additive and the associated signs (+/-) in the correlations simply refer to the directionality of the feature with respect to the predicted output, wherein, the canceling effect of terms in training is nullified during their non-linear combinatorial optimization of the Support Vector Regression machine. Thus, higher the magnitude of the correlation (irrespective of the sign), higher is the feature contribution. Let's recall that the overall cross-validated performance of EnCPdock has an r-value of ~0.75. Interestingly, buried surface area based terms as well as most network parameters performed particularly well wherein average contact intensity (ACI) and Link density (Ld) had the best individual correlations (0.51, 0.43 respectively) (**Fig. 3**). Complementarity based terms showed an opposite directionality to the predicted output, wherein, Sc, being the threshold filter also attained a fairly high correlation (-0.37), while, the low correlation in EC (-0.16) accounted for the variability in terms of electrostatic match or miss-match at the interface of the complexes.

Figure 3. Feature Contributions in terms of correlation with the predicted output. The Y-axis represents the Pearson's correlation (r) value for each feature against the predicted output ($\Delta G_{per\text{-}res}$), displayed as colored bars. Similar features have been put with nearer colors in the color bars.



3.3. Independent Validation

To check for any bias (and eliminate) that might have occurred from the training dataset during cross validation, or, in other words, to take care of over-fitting issues in training, it is customary to perform independent validations on datasets, strictly non-overlapping with the training set. To that end, we attempted to compile a database consisting of high-resolution PPI complexes with annotated experimental affinity (K_d and/or $\Delta G_{binding}$) values. We encountered a great deal of difficulty to accumulate such complexes with enough number and the desired accuracy as an unmistakable scarcity of such data (in the form of readily available annotated datasets) was evident from repeated painstaking web-search. One of the prime sources, an Affinity benchmark dataset (v.1), once made available was recently removed from the public domain (https://bmm.crick.ac.uk/~bmmadmin/Affinity/). One of the significant findings in our study was the discovery of 3 new targets as mentioned in the updated benchmark, v.2.114 of them were binary and were already taken into account. Subsequent declarations were made with regard to detected inaccuracies in the dataset (https://github.com/haddocking/binding-affinity-benchmark) in terms of the reported experimental thermodynamic parameters ($\Delta G_{binding}$ and/or K_d) and an upgraded version of the dataset was made available by incorporating fresh new entries (though accounting for comparatively a lower number of targets) having the desired accuracy. Together from earlier [45] and updated versions [46] of the Affinity benchmark (v.2), 225 PPI complexes were first assembled out of which 46 complexes were new entries that could be considered fully accurate (https://github.com/haddocking/binding-affinity-benchmark).

Barring these experimental inaccuracies and updates, these affinity benchmark datasets (both versions) however ensured sampling of a diverse spectrum of PPI complexes in terms of biological functions (covering G-proteins, receptor extracellular domains, antigen/antibody, enzyme/inhibitor, and enzyme/substrate complexes), as well as in terms of affinities (from low to high) of the binding partners with K_d ranging from 10^{-5} to 10^{-14} M [45,46] and was well categorized into different classes of comparative difficulty of affinity prediction (rigid body, medium, difficult) etc.

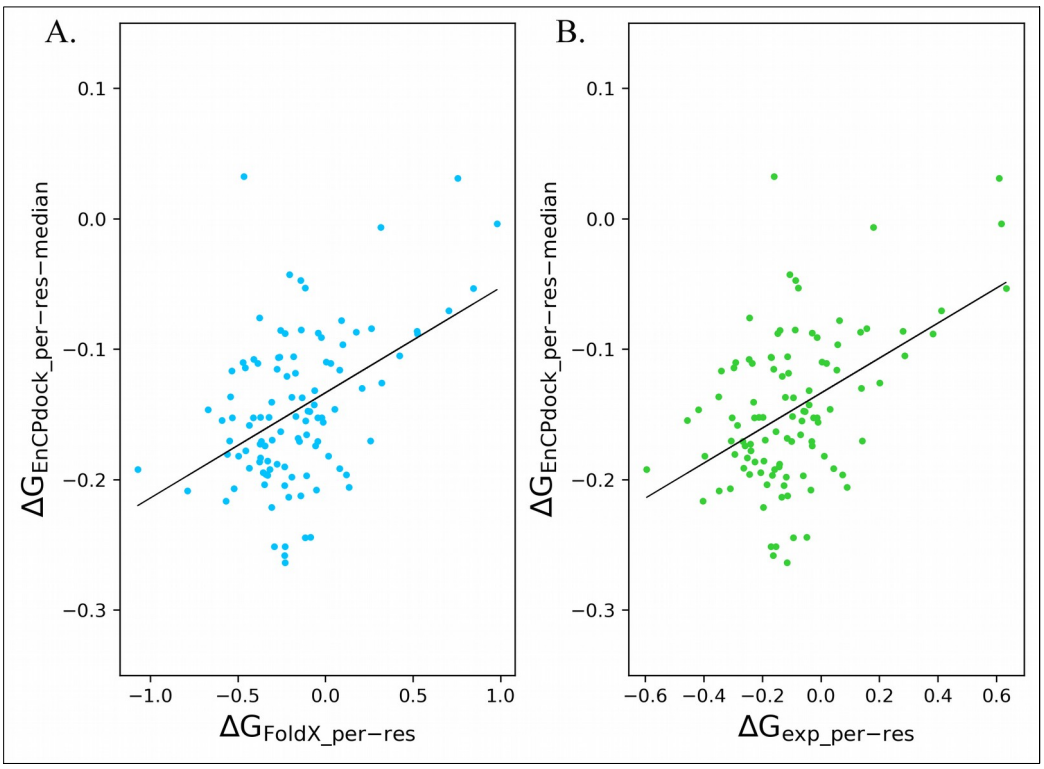
As a methodology, we had chosen to restrict our calculations strictly to binary complexes (i.e., receptor and ligand each consisting of a single polypeptide chain) which was also compelled by the factor that EnCPdock was trained on binary PPI complexes alone (see section **2.1**, **Materials and Methods**). This reduced the number of complexes to 106 (see **dataset S2** in **Supplementary Materials**) which have crystallographic resolutions in the range of 1.1-3.5 Å, and R-factors (working set) within the range of 0.125 to 0.252. This accumulated set contained 31 binary complexes out of the new 46 entries. Complete non-overlap of the training and test sets were ensured during culling of these (disjoint) datasets. To that end, any common entry found both in the training and the test sets was removed from the training set and the predictor was retrained afresh on the filtered training set consisting of 3200 complexes - which had no common entry to that of the test set(s). This ensured not losing out on any valuable 'test target' having experimental binding data.

To start with the post-prediction (performance evaluation) analyses, $\Delta G_{binding_per\text{-}res}$ values predicted by EnCPdock for each of the 106 targets compiling the filtered independent dataset (let's call them $\Delta G_{EnCPdock_per\text{-}res}$) were plotted against their corresponding experimental values ($\Delta G_{exp_per\text{-}res}$) and FoldX-derived values ($\Delta G_{FoldX_per\text{-}res}$) [47,48] into two separate 2D scatter plots (as a mean to compare) (**Fig. 4** and **Fig. S3** in **Supplementary Materials**). Note, each of the 90 models predicts $\Delta G_{EnCPdock_per\text{-}res}$ for each of the 106 targets. Thus, for each of the targets we have 90 $\Delta G_{EnCPdock_per\text{-}res}$ values. Here in the comparative analyses, we considered the central tendencies (mean, median and mode: $\Delta G_{EnCPdock_per\text{-}res\text{-}mean}$, $\Delta G_{EnCPdock_per\text{-}res\text{-}median}$, $\Delta G_{EnCPdock_per\text{-}res\text{-}mode}$) and the extreme values (minimum and maximum: $\Delta G_{EnCPdock_per\text{-}res\text{-}minimum}$, $\Delta G_{EnCPdock_per\text{-}res\text{-}maximum}$) for each target to compare with their corresponding $\Delta G_{exp_per\text{-}res}$ and $\Delta G_{FoldX_per\text{-}res}$ values. As reasoned earlier, 'median' was chosen as the preferred statistic over the other measures. In line with that, the primary comparison was drawn between the corresponding 'median' values (**Fig. 4**) while comparison between other measures were plotted alongside separately (**Fig. S3**). The main purpose of carrying out these plots (side-by-side) was to decipher if the performance of EnCPdock (in terms of the predicted $\Delta G_{EnCPdock_per\text{-}res}$ values) obtained against the target function taken in training (i.e., $\Delta G_{FoldX_per\text{-}res}$) is retained (or, if there's any improvement) when compared to the corresponding experimental $\Delta G_{binding_per\text{-}res}$ values ($\Delta G_{exp_per\text{-}res}$). To that end, in the left, right panels of these plots (**Fig. 4, Fig. S3**), $\Delta G_{FoldX_per\text{-}res}$, $\Delta G_{exp_per\text{-}res}$ values (both serving as the benchmarks) were taken along the X-axis (independent variables) respectively while the corresponding predicted $\Delta G_{EnCPdock_per\text{-}res}$ were taken along the Y-axis (dependent variable).

Linear least squares fitting was performed (in MATLAB, version R2018b) to each of these plots using perpendicular offsets. Any empirical correlations between these methods were first surveyed by visually analyzing these 2D scatter plots (and by the linear least squares fitting performed on them) followed by quantifying the obtained correlations by means of computing the Pearson's correlation coefficients (r) between them (**Table S2** in **Supplementary Materials**). EnCPdock performed reasonably well on the filtered independent dataset of 106 binary complexes giving rise to a correlation of r = 0.448 between the predicted $\Delta G_{EnCPdock_per\text{-}res\text{-}median}$ and the

corresponding FoldX-derived values taken as the benchmark. When compared to the actual experimental values (taken as the benchmark), the performance was not only retained but in fact was slightly improved leading to a correlation of r = 0.474 with $\Delta G_{EnCPdock_per\text{-}res\text{-}median}$. The other central tendencies and extreme values exhibited correlation slightly lesser than the $\Delta G_{EnCPdock_per\text{-}res\text{-}median}$ in both the cases ($\Delta G_{exp_per\text{-}res}$ and $\Delta G_{FoldX_per\text{-}res}$). The comparative results clearly show the apt of sophisticated surface-derived features like complementarity (Sc, EC), network based and other high-level structural descriptors in an advanced predictor like EnCPdock. The results may be envisaged as a performance booster of FoldX (or similar semi-empirical force-field based methods) when used as a target function in an advanced predictor compared to when used as the exclusive predictor itself. In future, a plausible hybrid method can be envisaged compiling FoldX and EnCPdock to raise the performance even more.

Figure 4. Correlation (scatter) plots in independent validations. The predicted output of $\Delta G_{binding_per\text{-}res}$ from EnCPdock ($\Delta G_{EnCPdock_per\text{-}res}$) for 106 PPI complexes (during independent-validation) plotted against their corresponding per-residue binding free energy predicted by FoldX ($\Delta G_{FoldX_per\text{-}res}$) and per-residue binding free energy estimated by experimental (calorimetric, spectroscopic etc.) means ($\Delta G_{exp_per\text{-}res}$), plotted in panels A and B respectively (with their least-squares fitted lines).



In order to further validate the EnCPdock performance on this filtered independent dataset, the central tendencies and the extreme values of $\Delta G_{binding_per\text{-}res}$ predicted by EnCPdock, for each of the 106 targets were converted to binaries. This conversion to binaries were based on a set of cut-offs spanning the ranges of $\Delta G_{exp_per\text{-}res}$ and $\Delta G_{FoldX_per\text{-}res}$. Further, upon this conversion, for each cut-off of $\Delta G_{exp_per\text{-}res}$ and $\Delta G_{FoldX_per\text{-}res}$, a set of TPR, TNR, BACC and MCC (see section **2.2.3**, **Materials and Methods**) were computed, which have been listed in **Table S3** in **Supplementary Materials**. For both $\Delta G_{exp_per\text{-}res}$ and $\Delta G_{FoldX_per\text{-}res}$, it could be seen that the BACC and MCC exhibit a value more than 0.75 and 0.50 respectively, for almost all the cases when the cut-off is set to a positive value, which indicates that the prediction of $\Delta G_{binding_per\text{-}res}$ by EnCPdock cannot be a mere

coincidence. Also, interestingly, all the central tendencies and the extreme values did not show much deviation from each other if their performance is compared based on these evaluation matrices (**Table S3**). As these central tendencies and extreme values for each target depend on the predicted $\Delta G_{binding_per\text{-}res}$ values from 90 different models, it seems all of the models perform almost similarly as the central tendencies are not much different from the extremes. Further, among different cut-offs, we obtain, a maximum of BACC score 0.892 and 0.885 when compared to their corresponding $\Delta G_{FoldX_per\text{-}res}$ and $\Delta G_{exp_per\text{-}res}$ values (**Table S3**) respectively with associated MCC values 0.782 and 0.779. This high BACC and MCC score indicate that this new method EnCPdock that is based on structural descriptors can be a useful tool in predicting the $\Delta G_{binding_per\text{-}res}$ of PPI complexes.

3.4. Comparison with the *state-of-the-art* in affinity predictions.

Decades long effort have been devoted in structure-based predictions of equilibrium thermodynamic parameters of protein-protein binding with gradual (linear) amplification of available experimental data of the desired quality. The history goes back to the early days of the trivial use of accessible surface area measures (ΔASA) which never (systematically) correlated well (even for the rigid complexes) with binding free energy ($r = -0.16$) barring a few case studies with much of empirical (*ad-hoc*) formalism [3]. It was found that both large and not-so-large ΔASA-complexes could give rise to high affinities [46]. To that end, it was clear that change in solvent accessible surface area upon binding was definitely not the sole determinant of binding affinity in proteins. Consequently, root-mean-square deviation measures defined on the interface residues alone (I-RMSD) were attempted as an affinity-predictor which also gave only a modest correlation with $\Delta G_{binding}$ ($r = -0.24$), and only a nominal improvement was found using a minimal linear model combining ΔASA and I-RMSD ($r = 0.31$) [49]. In the latter years, a coulomb based model (defined over interfacial distances) was used to determine the effect of electrostatic force on PPI complexes which also could not lead to much higher a correlation [50]. Also, since, protein structure and function affect the interaction and binding affinity of proteins, sequence-based prediction is challenging.

We are now in the era of the sophisticated ‘affinity scores’ - many of which (like that of our method) are AI-trained. Machine learning is certainly preferred among computational methods for predicting the binding affinity of proteins due to its implicit consideration of any factors affecting PPIs and the flexibility of using empirical data instead of a fixed or predetermined function. The correlation (with $\Delta G_{binding}$) was further improved by prediction methods (affinity scores) including the specific geometry and composition of the interaction, raising the overall correlations of up to $r = 0.53$ over a collection of flexible and rigid targets [46]. The best of these methods were trained either using an earlier version of the affinity benchmark [45] or using changes in affinity upon mutation(s) [51]. Interestingly, in spite of this elevated performance in the earlier benchmark (reported best correlation of $r = 0.63$), these methods [52–54] could not better their performance in the updated benchmark [46]. Consequently, performance of 12 commonly used algorithms compiling affinity scores as well as free energy predictors were found to give poor correlations on the earlier version of the benchmark consisting of 81 PPI complexes (https://github.com/haddocking/binding-affinity-benchmark) which slightly improved (up to $r=0.53$) on the subset of 46 fully accurate complexes [46].

In this background, here, in this paper, we tested the operational credibility of the EnCPdock system (www.scinetmol.in/EnCPdock/) to probe binding energetics in protein interactions. Let’s recall that unlike other predictors, EnCPdock is a comprehensive web-interface for direct conjoint comparative analyses of complementarity and binding energetics in inter-protein associations. Trained by non-linear support vector regression machines on thirteen judiciously chosen high-level

structural descriptors (and on FoldX-derived free energies as its target function), EnCPdock returns a performance of r=0.745 (with associated BACC score of 0.833, **Table S1**) in cross-validation over 3200 PPI complexes and r=0.474 against actual experimental data ($\Delta G_{binding_per\text{-}res}$) in an independent validation performed on a filtered set of 106 binary complexes, culled from the updated benchmark. Thus, considerable retention of performance is reflected in the independent validation on the updated affinity benchmark, comparable to the *state-of-the-art*. In addition, EnCPdock returns the mapping of the binary PPI complex in the Complementarity Plot and further returns mobile molecular graphics of the atomic contact network probed at the protein-protein interface for further analyses.

3.5. Case studies demonstrating the utilities of EnCPdock in probing inter-protein associations of various types and biological origins

EnCPdock has been built to exercise complementarity and energetics from a common interactive platform to probe protein – protein interactions. Currently the web-server runs only for binary protein (receptor – ligand) complexes, with a future plan to cover even more complex oligomeric protein associations. To demonstrate its utility in protein energetics, individual cases were studied in finer details, sampling a diverse plethora of biological origin and function of protein binary complexes. These included antigen – antibody interactions, protein – inhibitor complexes, interactions involved in signal transduction pathways, enzyme – substrate complexes and others.

One such example was that of a RAB GTPase bound to Rabenosyn-5, its cognate ligand (an effector molecule) (PDB ID: 1Z0K, [55]) involved in different stages of membrane trafficking, serving as a transport protein. The complex was found to raise an Sc value of 0.7 with an EC value of 0.12 which together made it fall into the 'probable' region of the Complementarity plot (CP_{dock}) (**Fig. 5.A, Table 2**). The inter-chain contact network spanned across a densely connected extended interface involving two α-helices of the ligand and primarily β-sheets coming from the receptor (**Fig. 5.B**). The Link density (Ld) of the interfacial network was found to be 0.11 with an average contact intensity (ACI) of 2.89 (see section **2.2.1.6**, **Materials and Methods**). Scale-freeness was evident from the degree distribution profile based features (combined) with a Pearsons correlation (r) of -0.987 (see section **2.2.2**, **Materials and Methods,**). The interface involved 22 residues combining both partners which eventually gave rise to a predicted normalized (per-residue) binding free energy ($\Delta G_{binding_per\text{-}res}$) of -0.183 $kcal.mol^{-1}$ – which was much closer to the experimental (Calorimetric [56]) value for the same (-0.196 $kcal.mol^{-1}$) as compared to FoldX predictions (-0.330 $kcal.mol^{-1}$).

One other example was that of a protein – inhibitor complex (PDB ID: 2SNI [57]) involving Subtilisin Novo, a Serine-Protease and its inhibitor. The complementarity values were much similar to the earlier complex (Sc: 0.72, EC: 0.12) again mapping to the 'probable' region of CP_{dock} (**Fig. 5.C, Table 2**). The inter-chain contact network was more extended (involving 32 residues) also mapped to a similar (link) densely (Ld: 0.13) however mostly involving β-sheets and loops coming from both binding partners (**Fig. 5.D**) with an even higher ACI value of 3.2 (see **Materials and Methods**). Scale-freeness was also evident here (r: -0.979) as with most bipartite PPI networks (see **Materials and Methods**). The complex lead to a similar predicted $\Delta G_{binding_per\text{-}res}$ value of -0.182 $kcal.mol^{-1}$ – which again was nearer to the experimental (Calorimetric [58]) reference (-0.268 $kcal.mol^{-1}$) as compared to FoldX predictions (-0.365 $kcal.mol^{-1}$).

Yet another example was also that of a protein – inhibitor complex from the Hydrolase family involving bovine trypsin and a tick tryptase inhibitor (PDB ID: 2UUY [59]). This complex presented a repulsive electrostatic interaction (EC: -0.052) compensated by stronger shape constraints (0.76) in terms of complementarity. Such compensation has been found in native PPI

complexes at a statistically significant extent (~20% of all native inter-protein interactions), often representing transient interactions [6,18,44]. Consistent with these (statistical) estimates, here, the {Sc, EC} point fell into the 'less probable' region of CP_{dock} (**Fig. 5.E, Table 2**). Here the interface was less-extended (involving 19 residues), albeit, more dense (Ld: 0.16, ACI: 2.69), again involving primarily β-sheets and loops coming from both binding partners (**Fig. 5.F**). As with the other examples, scale-freeness was also evident here (r: -0.98). The complex here led to a similar predicted $\Delta G_{binding_per\text{-}res}$ value of -0.17 kcal.mol^{-1} – which was comparable to FoldX predictions (-0.158 kcal.mol^{-1}) with respect to the experimental [60] benchmark (-0.115 kcal.mol^{-1}).

Together these examples adequately demonstrate the apt usage of EnCPdock as a combined graphical and analytical platform to survey complementarity and energetics attributed to protein – protein interactions.

Figure 5. EnCPdock outputs for three different PPI complexes, taken as demonstrative case studies. The Complementarity Plots (CP_{dock}) (panels A, C, E) and the interfacial interaction networks (panels B, D, F) returned by the EnCPdock web-server for the three chosen complexes with PDB IDs: 1Z0K (A and B respectively), 2SNI (C and D respectively) and 2UUY (E and F respectively) are portrayed.

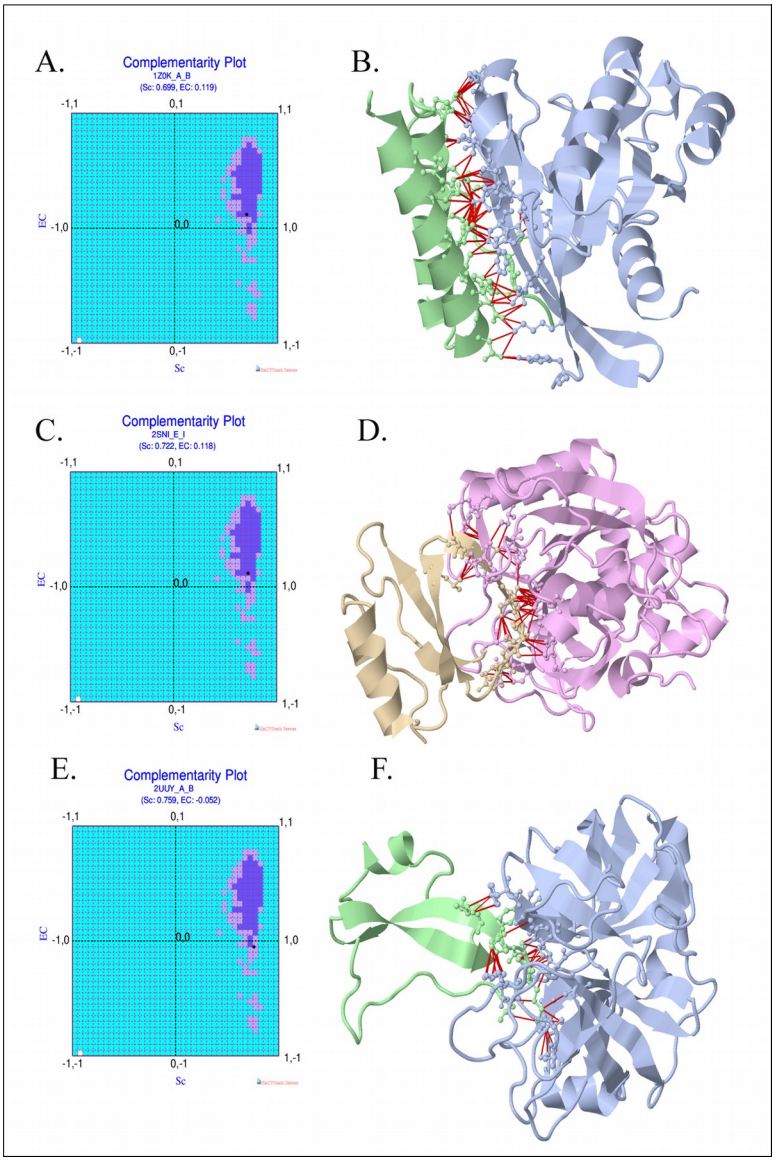


Table 2. Feature scores and predicted free energies returned by EnCPdock for the demonstrative examples chosen as case studies.

Features	PDB ID		
	1Z0K	2SNI	2UUY
N_{intres}	32	22	19
Sc	0.699	0.722	0.759
EC	0.119	0.118	-0.052
nBSA	0.166	0.128	0.109
$nBSA_p$	0.155	0.101	0.095
$nBSA_{np}$	0.175	0.149	0.120
fracI	0.217	0.142	0.156
Ld	0.105	0.134	0.155
ACI	2.889	3.200	2.692
$Slope_{dd}$	-1.545	-2.179	-1.657
$Yinter_{dd}$	-0.434	-0.299	-0.312
CCp_{dd}	-0.987	-0.979	-0.982
logN	2.361	2.530	2.439
log_{asp}	0.443	0.633	0.632
$\Delta G_{binding_per\text{-}res}$ (EnCPdock) (kcal.mol^{-1})	-0.183	-0.183	-0.170
$\Delta G_{binding_per\text{-}res}$ (FoldX) (kcal.mol^{-1})	-0.330	-0.365	-0.158
$\Delta G_{binding_per\text{-}res}$ (Experimental) (kcal.mol^{-1})	-0.196	-0.268	-0.115

3.6. Application of EnCPdock in probing peptide binding specificity and mutational effects

In principle, EnCPdock can also be used effectively to probe protein – peptide interactions and mutational effects. In order to test this, the human protein T-cell Lymphoma Invasion and Metastasis-1 (PDZ domain) was chosen in its native and Quadruple Mutant (QM) forms in complex with similar classes of peptides [61]. The QM is defined to be the simultaneous four mutations: L911M, K912E, L915F and L920V in the native receptor chain. Both forms having crystal structures (PDB ID: 4GVD, 4NXQ for native, mutant respectively) enabled us to carry out their direct comparison in EnCPdock. The native protein (4GVD) has two symmetry-related homologous bio-assemblies (chains: A-D; B-C), while the mutant form has three (chains: A-D, B-E, C-F). In effect, the asymmetric units of the native and the mutant forms contain (homo-) dimeric and trimeric complexes. All bio-assemblies of the native and the mutant were individually surveyed in EnCPdock, giving similar estimates in terms of most high-level structural descriptors (**Table 3**), yet, their cumulative difference adding to effective measures of (binding) affinity and stability (mutant

vs. native) could be probed unambiguously in terms of the predicted energetics (**Table 3**). The interaction of these classes of proteins (taken as receptors) and peptides (ligands) seem to have a characteristic suboptimal electrostatic complementarity (EC), often associated with transient (or, quasi-stable) complexes [44] – wherein the repulsive electrostatic effect are usually compensated by stronger shape constraints [6,18]. Indeed, all bio-assemblies pertaining to both the native and the mutant, unambiguously gave rise to negative EC values (EC_{native_4GVD}: -0.515 (A-D), -0.277 (B-C); EC_{mutant_4NXQ}: -0.354 (A-D), -0.608 (B-E), -0.282 (C-F)) indicative of their quasi-stability. As is characteristic with such metastable interactions, the corresponding shape complementarities (Sc) were found to be appreciably high, closely ranging between 0.69 to 0.73 among all complexes. Consequently, the {Sc, EC} ordered pair points (corresponding to all complexes) had fallen to the 4th quadrant of the Complementarity Plot (CP_{dock}) only managing to hit 'less probable' / 'improbable' regions of the plot (**Fig. 6**). The combined drop in stability and affinity (in terms of binding energetics) in the mutant compared to the native complexes, however, was unambiguous across the different bio-assemblies for both (**Table 3**). A destabilizing effect was evident in the mutant compared to the native complexes ($\Delta G_{EnCPdock_total}$ = -2.019 ± 0.054 kcal.mol^{-1} (native); -0.881 ± 1.471 kcal.mol^{-1} (mutant)), as also was revealed from indirect experimental estimates of $\Delta G_{binding}$ pertaining to analogous complexes, probed by fluorescence binding assays [62].

Figure 6. EnCPdock outputs to probe peptide binding specificity and mutational effects. EnCPdock outputs for the native human protein T-cell Lymphoma Invasion and Metastasis-1 protein (PDZ) domain (PDB ID: 4GVD) and its quadruple (QM) mutant (PDB ID: 4NXQ) bound to peptides of similar classes. The Complementarity Plots (CP_{dock}) and the interfacial contact networks of the native PPI complex have been analyzed and displayed for its two symmetry-related bio-assemblies (4GVD, chains: A-D, and B-C respectively), and for three symmetry-related bio-assemblies pertaining to the mutant (4NXQ, chains: A-D, B-E and C-F respectively).

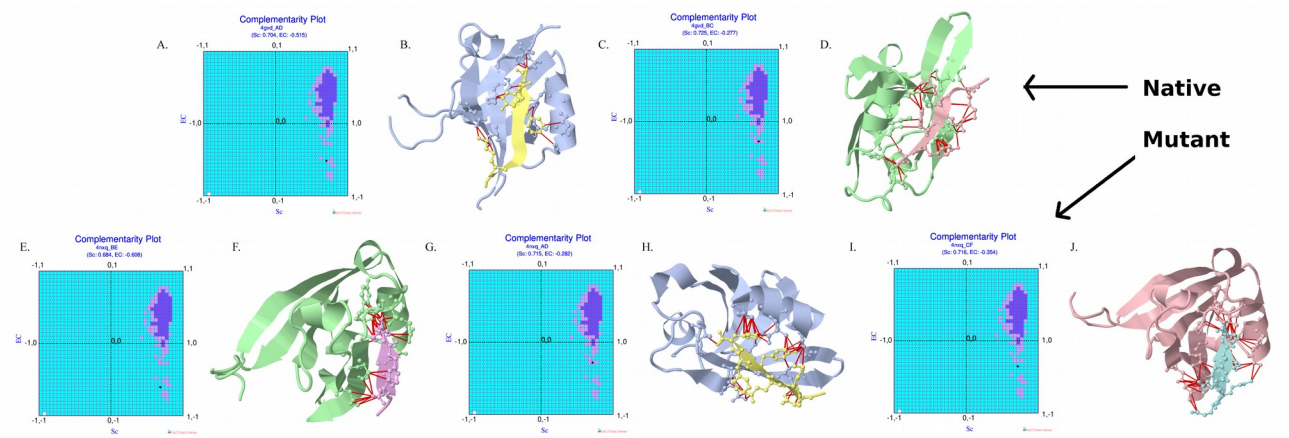


Table 3. Feature scores and predicted free energies returned by EnCPdock for the peptide-protein complexes used in probing peptide binding specificity and mutational effects. BA stands for Bio-assemblies present in a PDB structure. The native and mutant protein consists of two and three bio-assemblies respectively.

Descriptions, Features and predicted Free energies	Native		Mutant		
	BA-1	BA-2	BA-1	BA-2	BA-3
<PDB ID>_<Chain-ID>	4GVD_AD	4GVD_BC	4NXQ_AD	4NXQ_BE	4NXQ_CF
N_{intres}	14	18	17	12	17
Sc	0.704	0.725	0.715	0.684	0.716
EC	-0.515	-0.277	-0.282	-0.608	-0.354
nBSA	0.161	0.188	0.182	0.138	0.155
$nBSA_p$	0.149	0.192	0.172	0.120	0.148
$nBSA_{np}$	0.170	0.184	0.190	0.152	0.161
fracI	0.274	0.309	0.265	0.230	0.250
Ld	0.250	0.236	0.212	0.281	0.197
ACI	1.900	2.588	3.428	3.667	3.308
$Slope_{dd}$	-1.757	-1.261	-1.589	-1.333	-1.649
$Yinter_{dd}$	-0.434	-0.564	-0.399	-0.392	-0.417
CCp_{dd}	-0.881	-0.813	-0.991	-0.958	-0.950
logN	1.978	1.987	1.991	2.000	2.000
log_{asp}	1.036	1.046	1.113	1.194	1.194
$\Delta G_{EnCPdock_per\text{-}res}$ ($kcal.mol^{-1}$)	-0.147	-0.110	-0.110	0.067	-0.093
$\Delta G_{EnCPdock_total}$ ($kcal.mol^{-1}$)	-2.057	-1.980	-1.582	0.809	-1.871

4. Conclusion

EnCPdock was envisaged and built with the objective of developing a comprehensive web-interface for the conjoint comparative analyses of physico-chemically relevant high-level structural descriptors (complementarity in particular) and binding energetics of interacting protein partners. With that broad objective in mind, the current version of the web-sever provides fine-grained interface properties of binary PPI complexes inclusive of complementarity and other high-level structural features, and, concomitantly predicts their free energies of binding (per-residue contribution as well as total) from atomic coordinates. In addition, the web-server also generates mobile molecular graphics (in JSmol) of the interfacial atomic contact network and returns the contact map of the interface. A non-linear support vector regression machine has been implemented to train and build EnCPdock, elegantly combining fine-grained structural descriptors of size, shape,

electrostatics, accessible and buried surface areas, network parameters of the interfacial contact network etc. from an input protein complex. Trained on this knowledge-base, EnCPdock predicts the corresponding binding free energy with the desired accuracy and reasonable speed (~2 mins for each average-sized input). A detailed documentation of the web-interface can be found in the associated help page (https://scinetmol.in/EnCPdock/help.php) with description of all its inputs and outputs. In spite of being trained on near-native FoldX-derived estimates (due to lack of enough experimental data for training), EnCPdock performs comparable to (if not better than) the *state-of-the-art* in prediction accuracy of binding energetics – *as* revealed by cross-validations on a large dataset. The performance is well retained in an independent validation performed on a smaller dataset with actual experimental estimates of binding free energies. Demonstrative examples presented as case studies further show the robustness of EnCPdock across protein complexes coming from a variety of biological origin and types. Lastly, the utility of EnCPdock in probing peptide binding specificity has also been exemplified, coupled with its use in mutational analyses. The applications can naturally be extended to rank and re-rank docked poses of protein – protein / peptides. While the current version of EnCPdock is built to handle binary complexes alone, we plan to extend the web-server in the future enabled with the functionality to deal with higher order protein assemblies.

Acknowledgment and Funding The project was self funded.

Author's Contributions: SB conceptualized the idea, designed the calculations and wrote the main manuscript with help from GB, ND. DM developed the web-server with assistance from SB. ND and PG did the required literature survey. ND wrote the Introduction. GB performed all training, validations and actively participated in drafting the results and discussion. All authors read and approved the final manuscript.

References

[1] S. Jones, J.M. Thornton, Principles of protein-protein interactions, Proc Natl Acad Sci U S A. 93 (1996) 13–20. https://doi.org/10.1073/pnas.93.1.13.

[2] H.M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T.N. Bhat, H. Weissig, I.N. Shindyalov, P.E. Bourne, The Protein Data Bank, Nucl. Acids Res. 28 (2000) 235–242. https://doi.org/10.1093/nar/28.1.235.

[3] V. Venugopal, A.K. Datta, D. Bhattacharyya, D. Dasgupta, R. Banerjee, Structure of cyclophilin from Leishmania donovani bound to cyclosporin at 2.6 A resolution: correlation between structure and thermodynamic data, Acta Crystallogr D Biol Crystallogr. 65 (2009) 1187–1195. https://doi.org/10.1107/S0907444909034234.

[4] B.G. Pierce, K. Wiehe, H. Hwang, B.-H. Kim, T. Vreven, Z. Weng, ZDOCK server: interactive docking prediction of protein-protein complexes and symmetric multimers, Bioinformatics. 30 (2014) 1771–1773. https://doi.org/10.1093/bioinformatics/btu097.

[5] S. Basu, B. Wallner, DockQ: A Quality Measure for Protein-Protein Docking Models, PLOS ONE. 11 (2016) e0161879. https://doi.org/10.1371/journal.pone.0161879.

[6] S. Basu, B. Wallner, Finding correct protein-protein docking models using ProQDock, Bioinformatics. 32 (2016) i262–i270. https://doi.org/10.1093/bioinformatics/btw257.

[7] M.C. Lawrence, P.M. Colman, Shape complementarity at protein/protein interfaces, J. Mol. Biol. 234 (1993) 946–950. https://doi.org/10.1006/jmbi.1993.1648.

[8] J. Homola, Surface Plasmon Resonance Sensors for Detection of Chemical and Biological Species, Chem. Rev. 108 (2008) 462–493. https://doi.org/10.1021/cr068107d.
[9] I. Navratilova, A.L. Hopkins, Emerging role of surface plasmon resonance in fragment-based drug discovery, Future Med Chem. 3 (2011) 1809–1820. https://doi.org/10.4155/fmc.11.128.
[10] I. Jelesarov, H.R. Bosshard, Isothermal titration calorimetry and differential scanning calorimetry as complementary tools to investigate the energetics of biomolecular recognition, J Mol Recognit. 12 (1999) 3–18. https://doi.org/10.1002/(SICI)1099-1352(199901/02)12:1<3::AID-JMR441>3.0.CO;2-6.
[11] V. Fernández-Dueñas, J. Llorente, J. Gandía, D.O. Borroto-Escuela, L.F. Agnati, C.I. Tasca, K. Fuxe, F. Ciruela, Fluorescence resonance energy transfer-based technologies in the study of protein-protein interactions at the cell surface, Methods. 57 (2012) 467–472. https://doi.org/10.1016/j.ymeth.2012.05.007.
[12] G. Schreiber, Kinetic studies of protein-protein interactions, Curr Opin Struct Biol. 12 (2002) 41–47. https://doi.org/10.1016/s0959-440x(02)00287-7.
[13] E. Wang, H. Sun, J. Wang, Z. Wang, H. Liu, J.Z.H. Zhang, T. Hou, End-Point Binding Free Energy Calculation with MM/PBSA and MM/GBSA: Strategies and Applications in Drug Design, Chem. Rev. 119 (2019) 9478–9508. https://doi.org/10.1021/acs.chemrev.9b00055.
[14] F. Chen, H. Liu, H. Sun, P. Pan, Y. Li, D. Li, T. Hou, Assessing the performance of the MM/PBSA and MM/GBSA methods. 6. Capability to predict protein-protein binding free energies and re-rank binding poses generated by protein-protein docking, Phys Chem Chem Phys. 18 (2016) 22129–22139. https://doi.org/10.1039/c6cp03670h.
[15] A.J. McCoy, V. Chandana Epa, P.M. Colman, Electrostatic complementarity at protein/protein interfaces, J. Mol. Biol. 268 (1997) 570–584. https://doi.org/10.1006/jmbi.1997.0987.
[16] S. Basu, D. Bhattacharyya, R. Banerjee, Applications of complementarity plot in error detection and structure validation of proteins, Indian J. Biochem. Biophys. 51 (2014) 188–200.
[17] S. Basu, D. Bhattacharyya, R. Banerjee, Self-Complementarity within Proteins: Bridging the Gap between Binding and Folding, Biophys J. 102 (2012) 2605–2614. https://doi.org/10.1016/j.bpj.2012.04.029.
[18] S. Basu, CPdock: the complementarity plot for docking of proteins: implementing multi-dielectric continuum electrostatics, J Mol Model. 24 (2017) 8. https://doi.org/10.1007/s00894-017-3546-y.
[19] T. Vreven, I.H. Moal, A. Vangone, B.G. Pierce, P.L. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P.A. Bates, J. Fernandez-Recio, A.M.J.J. Bonvin, Z. Weng, Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2, Journal of Molecular Biology. 427 (2015) 3031–3041. https://doi.org/10.1016/j.jmb.2015.07.016.
[20] R. Banerjee, M. Sen, D. Bhattacharya, P. Saha, The jigsaw puzzle model: search for conformational specificity in protein interiors, J. Mol. Biol. 333 (2003) 211–226.
[21] M.D. Winn, C.C. Ballard, K.D. Cowtan, E.J. Dodson, P. Emsley, P.R. Evans, R.M. Keegan, E.B. Krissinel, A.G.W. Leslie, A. McCoy, S.J. McNicholas, G.N. Murshudov, N.S. Pannu, E.A. Potterton, H.R. Powell, R.J. Read, A. Vagin, K.S. Wilson, Overview of the CCP4 suite and current developments, Acta Crystallogr D Biol Crystallogr. 67 (2011) 235–242. https://doi.org/10.1107/S0907444910045749.
[22] D. Xu, Y. Zhang, Generating Triangulated Macromolecular Surfaces by Euclidean Distance Transform, PLoS ONE. 4 (2009) e8140. https://doi.org/10.1371/journal.pone.0008140.
[23] L. Li, C. Li, S. Sarkar, J. Zhang, S. Witham, Z. Zhang, L. Wang, N. Smith, M. Petukh, E. Alexov, DelPhi: a comprehensive suite for DelPhi software and associated resources, BMC Biophys. 5 (2012) 9. https://doi.org/10.1186/2046-1682-5-9.
[24] Naccess homepage, (n.d.). http://www.bioinf.manchester.ac.uk/naccess/ (accessed April 6, 2022).

[25] W.D. Cornell, P. Cieplak, C.I. Bayly, I.R. Gould, K.M. Merz, D.M. Ferguson, D.C. Spellmeyer, T. Fox, J.W. Caldwell, P.A. Kollman, A Second Generation Force Field for the Simulation of Proteins, Nucleic Acids, and Organic Molecules, J. Am. Chem. Soc. 117 (1995) 5179–5197. https://doi.org/10.1021/ja00124a002.
[26] S. Basu, D. Bhattacharyya, R. Banerjee, Mapping the distribution of packing topologies within protein interiors shows predominant preference for specific packing motifs, BMC Bioinformatics. 12 (2011) 195. https://doi.org/10.1186/1471-2105-12-195.
[27] S. Basu, P. Biswas, Salt-bridge dynamics in intrinsically disordered proteins: A trade-off between electrostatic interactions and structural flexibility, Biochimica et Biophysica Acta (BBA) - Proteins and Proteomics. 1866 (2018) 624–641. https://doi.org/10.1016/j.bbapap.2018.03.002.
[28] S. Roy, P. Ghosh, A. Bandyopadhyay, S. Basu, Capturing a Crucial 'Disorder-to-Order Transition' at the Heart of the Coronavirus Molecular Pathology—Triggered by Highly Persistent, Interchangeable Salt-Bridges, Vaccines. 10 (2022) 301. https://doi.org/10.3390/vaccines10020301.
[29] C. Geng, L.C. Xue, J. Roel-Touris, A.M.J.J. Bonvin, Finding the ΔΔG spot: Are predictors of binding affinity changes upon mutations in protein–protein interactions ready for it?, WIREs Computational Molecular Science. 9 (2019) e1410. https://doi.org/10.1002/wcms.1410.
[30] T. Joachims, Learning to Classify Text Using Support Vector Machines, Springer US, Boston, MA, 2002. http://link.springer.com/10.1007/978-1-4615-0907-3 (accessed November 14, 2015).
[31] V.N. Vapnik, The Nature of Statistical Learning Theory, Springer New York, New York, NY, 2000. https://doi.org/10.1007/978-1-4757-3264-1.
[32] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, L. Serrano, The FoldX web server: An online force field, Nucleic Acids Research. 33 (2005) 382–388. https://doi.org/10.1093/nar/gki387.
[33] O. Buß, J. Rudat, K. Ochsenreither, FoldX as Protein Engineering Tool: Better Than Random Based Approaches?, Comput Struct Biotechnol J. 16 (2018) 25–33. https://doi.org/10.1016/j.csbj.2018.01.002.
[34] R.F. Alford, A. Leaver-Fay, J.R. Jeliazkov, M.J. O'Meara, F.P. DiMaio, H. Park, M.V. Shapovalov, P.D. Renfrew, V.K. Mulligan, K. Kappel, J.W. Labonte, M.S. Pacella, R. Bonneau, P. Bradley, R.L. Dunbrack, R. Das, D. Baker, B. Kuhlman, T. Kortemme, J.J. Gray, The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design, J. Chem. Theory Comput. 13 (2017) 3031–3048. https://doi.org/10.1021/acs.jctc.7b00125.
[35] M. Boardman, T. Trappenberg, A Heuristic for Free Parameter Optimization with Support Vector Machines, in: The 2006 IEEE International Joint Conference on Neural Network Proceedings, 2006: pp. 610–617. https://doi.org/10.1109/IJCNN.2006.246739.
[36] Q. Wei, R.L.D. Jr, The Role of Balanced Training and Testing Data Sets for Binary Classifiers in Bioinformatics, PLOS ONE. 8 (2013) e67863. https://doi.org/10.1371/journal.pone.0067863.
[37] H.A. Gabb, R.M. Jackson, M.J.E. Sternberg, Modelling protein docking using shape complementarity, electrostatics and biochemical information11Edited by J. Thornton, Journal of Molecular Biology. 272 (1997) 106–120. https://doi.org/10.1006/jmbi.1997.1203.
[38] Y. Yan, S.-Y. Huang, Pushing the accuracy limit of shape complementarity for protein-protein docking, BMC Bioinformatics. 20 (2019) 696. https://doi.org/10.1186/s12859-019-3270-y.
[39] D. Dell'Orco, W.-F. Xue, E. Thulin, S. Linse, Electrostatic Contributions to the Kinetics and Thermodynamics of Protein Assembly, Biophysical Journal. 88 (2005) 1991–2002. https://doi.org/10.1529/biophysj.104.049189.

[40] H.-X. Zhou, X. Pang, Electrostatic Interactions in Protein Structure, Folding, Binding, and Condensation, Chem Rev. 118 (2018) 1691–1741. https://doi.org/10.1021/acs.chemrev.7b00305.
[41] D. Basak, S. Pal, D.C. Patranabis, Support Vector Regression, Neural Information Processing. 11 (2007).
[42] A. Koike, T. Takagi, Prediction of protein-protein interaction sites using support vector machines, Protein Engineering Design and Selection. 17 (2004) 165–173. https://doi.org/10.1093/protein/gzh020.
[43] E. Aybey, Ö. Gümüş, SENSDeep: An Ensemble Deep Learning Method for Protein–Protein Interaction Sites Prediction, Interdiscip Sci Comput Life Sci. (2022). https://doi.org/10.1007/s12539-022-00543-x.
[44] S. Basu, D. Chakravarty, D. Bhattacharyya, P. Saha, H.K. Patra, Plausible blockers of Spike RBD in SARS-CoV2—molecular design and underlying interaction dynamics from high-level structural descriptors, J Mol Model. 27 (2021) 191. https://doi.org/10.1007/s00894-021-04779-0.
[45] P.L. Kastritis, I.H. Moal, H. Hwang, Z. Weng, P.A. Bates, A.M.J.J. Bonvin, J. Janin, A structure-based benchmark for protein–protein binding affinity, Protein Science. 20 (2011) 482–491. https://doi.org/10.1002/pro.580.
[46] T. Vreven, I.H. Moal, A. Vangone, B.G. Pierce, P.L. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P.A. Bates, J. Fernandez-Recio, A.M.J.J. Bonvin, Z. Weng, Updates to the Integrated Protein-Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2, J. Mol. Biol. 427 (2015) 3031–3041. https://doi.org/10.1016/j.jmb.2015.07.016.
[47] R. Guerois, J.E. Nielsen, L. Serrano, Predicting changes in the stability of proteins and protein complexes: a study of more than 1000 mutations, J Mol Biol. 320 (2002) 369–387. https://doi.org/10.1016/S0022-2836(02)00442-4.
[48] J. Schymkowitz, J. Borg, F. Stricher, R. Nys, F. Rousseau, L. Serrano, The FoldX web server: an online force field, Nucleic Acids Res. 33 (2005) W382-388. https://doi.org/10.1093/nar/gki387.
[49] J. Janin, A minimal model of protein–protein binding affinities, Protein Science. 23 (2014) 1813–1817. https://doi.org/10.1002/pro.2560.
[50] A.V. Morozov, T. Kortemme, D. Baker, Evaluation of Models of Electrostatic Interactions in Proteins, J. Phys. Chem. B. 107 (2003) 2075–2090. https://doi.org/10.1021/jp0267555.
[51] I.H. Moal, J. Fernandez-Recio, Intermolecular Contact Potentials for Protein–Protein Interactions Extracted from Binding Free Energy Changes upon Mutation, J. Chem. Theory Comput. 9 (2013) 3715–3727. https://doi.org/10.1021/ct400295z.
[52] I.H. Moal, R. Agius, P.A. Bates, Protein–protein binding affinity prediction on a diverse set of structures, Bioinformatics. 27 (2011) 3002–3009. https://doi.org/10.1093/bioinformatics/btr513.
[53] T. Vreven, H. Hwang, B.G. Pierce, Z. Weng, Prediction of protein–protein binding free energies, Protein Science. 21 (2012) 396–404. https://doi.org/10.1002/pro.2027.
[54] P.L. Kastritis, J.P.G.L.M. Rodrigues, G.E. Folkers, R. Boelens, A.M.J.J. Bonvin, Proteins Feel More Than They See: Fine-Tuning of Binding Affinity by Properties of the Non-Interacting Surface, Journal of Molecular Biology. 426 (2014) 2632–2652. https://doi.org/10.1016/j.jmb.2014.04.017.
[55] S. Eathiraj, X. Pan, C. Ritacco, D.G. Lambright, Structural basis of family-wide Rab GTPase recognition by rabenosyn-5, Nature. 436 (2005) 415–419. https://doi.org/10.1038/nature03798.

[56] G.D. Henry, D.J. Corrigan, J.V. Dineen, J.D. Baleja, Charge Effects in the Selection of NPF Motifs by the EH Domain of EHD1, Biochemistry. 49 (2010) 3381–3392. https://doi.org/10.1021/bi100065r.
[57] C.A. McPhalen, M.N. James, Structural comparison of two serine proteinase-protein inhibitor complexes: eglin-c-subtilisin Carlsberg and CI-2-subtilisin Novo, Biochemistry. 27 (1988) 6582–6598.
[58] S. Krystek, T. Stouch, J. Novotny, Affinity and specificity of serine endopeptidase-protein inhibitor interactions. Empirical free energy calculations based on X-ray crystallographic structures, J Mol Biol. 234 (1993) 661–679. https://doi.org/10.1006/jmbi.1993.1619.
[59] G.C. Paesen, C. Siebold, K. Harlos, M.F. Peacey, P.A. Nuttall, D.I. Stuart, A Tick Protein with a Modified Kunitz Fold Inhibits Human Tryptase, Journal of Molecular Biology. 368 (2007) 1172–1186. https://doi.org/10.1016/j.jmb.2007.03.011.
[60] I. Maffucci, A. Contini, Improved Computation of Protein–Protein Relative Binding Energies with the Nwat-MMGBSA Method, ACS Publications. (2016). https://doi.org/10.1021/acs.jcim.6b00196.
[61] N. Panel, F. Villa, E.J. Fuentes, T. Simonson, Accurate PDZ/Peptide Binding Specificity with Additive and Polarizable Free Energy Simulations, Biophys J. 114 (2018) 1091–1102. https://doi.org/10.1016/j.bpj.2018.01.008.
[62] T.R. Shepherd, R.L. Hard, A.M. Murray, D. Pei, E.J. Fuentes, Distinct Ligand Specificity of the Tiam1 and Tiam2 PDZ Domains, Biochemistry. 50 (2011) 1296–1308. https://doi.org/10.1021/bi1013613.