Supporting Information

PPI-Affinity: A web tool for the prediction and optimization of protein – peptide and protein – protein binding affinity

Sandra Romero-Molina[1], Yasser B. Ruiz-Blanco [1], Joel Mieres-Perez[1], Mirja Harms[2], Jan Münch[2,3], Michael Ehrmann[4] and Elsa Sanchez-Garcia[1]*

[1]Computational Biochemistry, Center of Medical Biotechnology, University of Duisburg-Essen, Essen, Germany

[2]Institute of Molecular Virology, Ulm University Medical Center, Ulm, Germany

[3]Core Facility Functional Peptidomics, Ulm University Medical Center, Ulm, Germany

[4]Faculty of Biology, Center of Medical Biotechnology, University of Duisburg-Essen, Essen, Germany

*Email: elsa.sanchez-garcia@uni-due.de

Table of Contents

Table SI-1. Description of PPI-Affinity and state-of-the-art PPI prediction tools.

Method	*Training size*	*Database*	*Type of problem*	*Validation technique*	*Domain of application*	*Input information*	*Type of molecular descriptors*	*ML algorithms*	*Availability*
Prodigy[1]	81	Benchmark dataset (Kastritis et al., 2011)[2]	Regression	4-fold cross validation (repeated 10 times)	Protein-Protein	3D	Descriptors based in the network of inter-residue contacts (ICs) and the non-interacting surface (NIS)	Linear Regression	Webserver
DFIRE[3]	-	-	-	Test set	Protein-Protein/Peptide	3D	-	-	Webserver / Standalone program
CP_PIE[4]	540	PDB[5, 6]	Regression	Test set	Protein-Protein	3D	Descriptors based on residue contacts and the overlapping area	Linear programming formulation	Standalone program
Kdeep[7]	3767	PDBbind (v.2016)[8]	Regression	Test set	Protein-Ligand	3D	Pharmacophoric-like properties in the	Convolutional Neural Networks	Webserver

							binding site		
RF-Score[9]	1105	PDBbind (v.2007)[8]	Regression	Out-Of-Bag data/Test set/y-scrambling	Protein-Ligand	3D	Occurrence counts of atom type pairs within a distance range	Random Forest	Standalone program
Trypano-PPI[10]	5872	PDB[5, 6] / dataset reported by Dobson and Doig[11]	Classification	Validation set	Protein-Protein interactions in Trypanosome	3D	Markov Chain numerical descriptors based on average electrostatic potentials (MARCH-INSIDE 2.0 package)	Linear Neural Network	Webserver
Plasmod-PPI[12]	5257	PDB[5, 6] / dataset reported by Dobson and Doig[11]	Classification	Validation set	Protein-Protein interactions in Plasmodium	3D	Markov Chain numerical descriptors based on electrostatic entropies calculations (MARCH-INSIDE 2.0 package)	Classifier Trees (CT) / Linear discriminant analysis (LDA) for feature selection	Webserver

GO-PseAA[13]	6323	STRING[14]	Classification	Jack-knife test (LOOCV)	Protein-Protein	1D	Pseudo-amino acid composition (PseAA)/ Gene Ontology Consortium (GO)	Intimate Sorting (ISort) Classifier	-
ALT-IN[15]	Dataset 1: 1831 Dataset 2: 5460	Dataset 1:[16] Dataset 2:[16-20]	Classification	Nested leave-group-out cross validation (CV) /Case of study	Protein-Protein (predicting disruptions in PPIs induced by alternative splicing)	1D	Biochemical features of the reference isoform and its interaction partner/ Domain interaction knowledge-based statistical potentials/ Selected characteristics of alternative splicing events	Random Forest-driven supervised and semi-supervised learning	Standalone program
iPPBS-Opt[21]	13771 surface-residue / 27442	Dataset of 99 proteins[22]	Classification	10-fold CV	Protein-Protein/Peptide binding sites	1D	Pseudo Amino Acid Composition (PseAAC)	Random Forest	Webserver

	all-residue								
PPA-Pred[23]	135	Benchmark dataset (Kastritis *et al.*, 2011)[2]	Regression	LOOCV	Protein-Protein	1D	Descriptors based in sequence and structure properties	Multiple regression technique	Webserver
ISLAND[24]	135/39 test	Benchmark dataset (Kastritis *et al.* , 2011)[2]	Regression	LOOCV	Protein-Protein	1D	Descriptors based on biophysical amino acids properties and structural properties derived from the sequence (Propy package)[25]	Support Vector Machine (SVM) for regression	Webserver
SSIPe[26]	1470 training/ 734 test/ 888 test/ 190/152	NIL[27] /STRING[14]/ SKEMPI 2.0[28]/CAPRI[29]	Regression	5-fold CV/Test set	Protein-Protein	3D	Descriptors based in sequence and structure	Monte Carlo procedure/ Linear regression	Webserver/ Standalone

							interface evolutionary profiles		
Wang et al.[30]	158 wild-type protein complexes and 3205 mutants	SKEMPI 1.0[31]	Regression	LOOCV	Protein-Protein	3D	Numerical descriptors based on energetic contributions and calculated with the Monte-Carlo algorithm	Knowledge-based potential (Monte-Carlo simulations for weights optimization)	-
LUPIA[32]	128 (training set) /39 protein-(validation set)	Benchmark dataset (Kastritis *et al.*, 2011)[2]/ Chen et al.[33]	Classification	LOOCV/ validation set	Protein-Protein	1D	k-mer composition, BLOSUM-62 features,[3] [4] number of interacting residue pairs, Moal[35] and Dias	Learning Using Privileged Information (LUPI) /SVM	Webserver / Standalone program

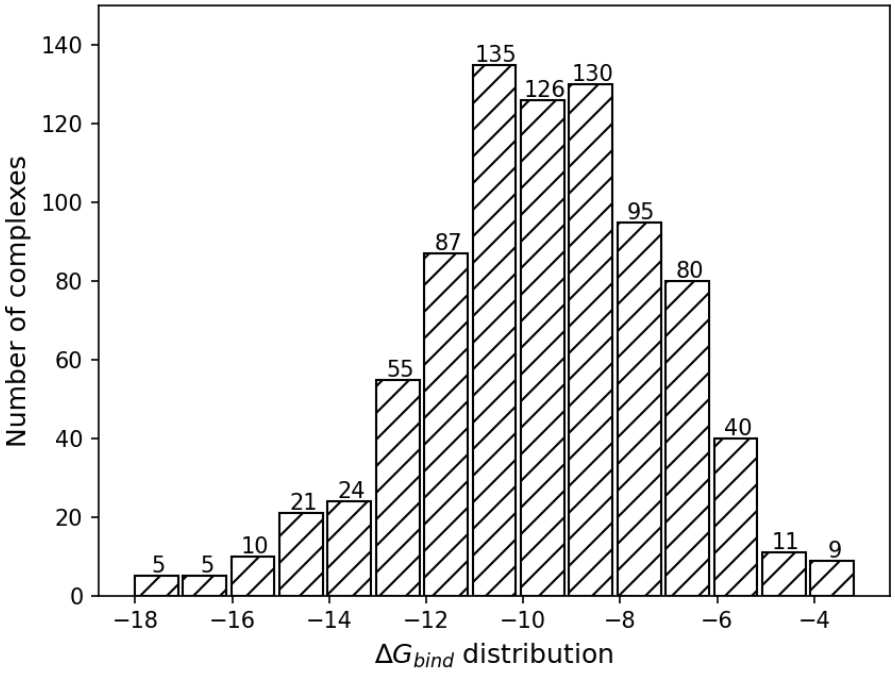
							descriptors[36]		
PPI-Affinity	648 protein-protein / 922 protein-peptide	PDBbind (v.2020)[8] / Biolip[37]	Regression	10-fold CV/Validation set/Test set	Protein-Protein/Peptide	3D	Structural numerical descriptors calculated with ProtDCal software	Ensemble method / SVM	Webserver

Table SI-2. Correlation coefficient (R) of protein-protein and protein-ligand BA predictors on the test set of 100 protein-peptide complexes.

Method	*R*	MAE (kcal/mol)
Prodigy[1, 38]	0.13	1.9
DFIRE[3]	0.29	8.7
PIE[4]	-0.28	9.0
Kdeep*[7]	0.32	10.7
RF-Score*[39]	0.23	1.8

Protein-ligand methods are marked with *

Figure SI-1. Distribution of ΔG_{bind} values in the dataset of 833 protein-protein complexes.



Section SI-1. Parameter setup for the calculation of the structural descriptors in ProtDCal.[40, 41]

Sections	Description
directory:	
Datasets/PDB_Protein_Format	Path to directory with the input PDB files
indices:	
wNc, wFLC, wNLC, wCO, wLCO, wRWCO, wCTP, wCLQ wCO: ECI, IP, ISA, Z1, Z2, Z3 wRWCO: ECI, IP, ISA, Z1, Z2, Z3 wNc: ECI, IP, ISA, Z1, Z2, Z3 wCLQ: ECI, IP, ISA, Z1, Z2, Z3 wNLC: ECI, IP, ISA, Z1, Z2, Z3 wLCO: ECI, IP, ISA, Z1, Z2, Z3 wCTP: ECI, IP, ISA, Z1, Z2, Z3 wFLC: ECI, IP, ISA, Z1, Z2, Z3	List of used Topographic Indices of Folded Protein States List of used weighting coefficients for each Topographic Indices of Folded Protein States
groups:	
ALA, ARG, ASN, ASP, CYS, GLU, GLN, GLY, HIS, ILE, LEU, LYS, MET, PHE, PRO, SER, THR, TRP, TYR, VAL, RTR, BSR, AHR, ALR, NPR, ARM, PLR, PCR, NCR, UCR, UFR, PRT	List of grouping operators
invariants:	
N1, N2, Ar, P2, G, V, CV, S, RA, K, DE, I50, SI, MI, TI	List of vicinity operators
parameters(t_cont,s_cont,A%,HydGroup,n,bins,K,SubG):	
4000.0, 10.0, 5.0, 9.4, 3.0, 30, 5, 3	Parameters values for internal options of the program
options(decimals,harmonicMeanType,geometricMeanType,windexID,datasetType,outputOrder):	
3, 0, 0, -1, pdb, true	Parameters values for internal options of the program

Section SI-2 Preliminary study conducted to the identify the Machine Learning (ML) technique to use in the development of the models.

First, we randomly divided the 833 protein-protein complexes into datasets of 743 and 90 structures to train and test the models respectively. Initially, ProtDCal generated 23 040 molecular descriptors for each protein-protein complex. We reduced this high number of dimensions through an attribute selection process. First, we applied a filter to remove descriptors with non or little variation in the training set (*RemoveUseless* filter of Weka). A descriptor that cannot be calculated for an instance represents a missing value in the final vector. For a regression problem, this means that missing values need to be replaced by a number (mean, median, ...). Since this could add more noise to the dataset, we handled this issue by deleting all the attributes that contained at least one instance with a missing value. These steps reduced the dataset to 1 477 attributes. Next, we ordered the attributes by their Pearson's correlation coefficient with the class (*CorrelationAttributeEval* implementation of Weka). The highest correlation of an attribute with the class was 0.24, while the minimum was -0.32. We selected those with correlation values between $0.1 <= R <= -0.1$, reducing the data to 476 descriptors.
Next, we applied the *WrapperSubsetEval* technique implemented in Weka for obtaining the best subset of attributes to predict the class value. This is a supervised technique that evaluates subsets of features by training and evaluating, directly employing the classifier. The selection of the subsets is held by a search method. Here, we employed a genetic algorithm with a population of 20 individuals, crossover and mutation probabilities of 0.6 and 0.033 respectively and 20 generations maximum. Each subset was evaluated in 5-fold cross-validation, and the selection of the best subset was carried out attending to the correlation coefficient achieved by the classifier. This process was performed by each method individually (Table SI-2.1). We employed the Weka implementations of Linear Regression, Multilayer Perceptron: with zero (Linear Neural Network, LNN), one (ANN-1H) and two (ANN-2H) hidden layers, Random Forest, SVM for regression with the polynomial of degree one (SVM-PK-D1), degree two (SVM-PK-D2), and radial basis function (SVM-RBF) kernels. All the classifiers were executed with the default parameters values provided by Weka. In the case of Neural Networks, the amount of nodes (a) per hidden layer is defined by default in Weka as: $a = (d + c) / 2$, where d is the number of descriptors and c is the number of classes (c=1 for regression). For SVM-PK-D2, the Wrapper method was not applied, and the evaluation was performed on the attributes selected by SVM-PK-D1.

Table SI-2.1. Summary of the performance of different classifiers after the application of feature selection steps on the protein-protein BA data. The column "Descriptors" contains the number of attributes selected by each method with the Wrapper technique. The performance is expressed as the Pearson's Correlation coefficient (R) between experimental and predicted BA. Each model was evaluated on the training set, in 10-fold cross-validation (10-fold CV), and on the test set of 90 data points taken from PDBbind (v.2020).[8] The methods tested were Linear Regression, Multilayer Perceptron with zero (Linear Neural Network, LNN), one (ANN-1H) and two (ANN-2H) hidden layers, Random Forest, Support Vector Machine for regression with the polynomial (SVM-PK-D1, SVM-PK-D2) and the Radial Basis Function (SVM-RBF) kernels.

	Descriptors	Training set		10-fold CV		Test set	
		R	MAE	R	MAE	R	MAE
Linear Regression	74	0.68	1.4	0.60	1.6	0.39	2.0
LNN	12	-0.05	3.4	0.20	2.3	-0.01	3.1
ANN-1H	11	0.58	1.6	0.40	1.9	0.41	1.8
ANN-2H	10	0.55	1.7	0.39	1.9	0.40	1.9
Random Forest	27	0.98	0.6	0.67	1.5	0.61	1.6
SVM-PK-D1	36	0.57	1.5	0.54	1.6	0.40	1.9
SVM-PK-D2	36	0.77	1.0	0.42	2.0	0.38	2.0
SVM-RBF	171	0.63	1.4	0.58	1.6	0.45	1.8

The smallest subset of descriptors was obtained by the Neural Networks methods. However, under these architectures the models suffered from both: under-fitting in the case of LNN, with non or small correlation values among the model evaluations, as well as over-fitting in the case of ANN-1H and ANN-2H, with a difference greater than the 20% between the correlation on the training set, and 10-fold CV, and test set. In the case of Random Forest, the results on the training set (R = 0.98, MAE = 0.6), along with the performance in CV (R = 0.67, MAE = 1.5) and on the test set (R = 0.61, MAE = 1.6) clearly denoted high over-fitting.

Linear Regression showed close performance on the training set (R = 0.68, MAE = 1.4) and in 10-fold CV (R = 0.6, MAE = 1.6). However, the number of descriptors equal to 74 could harm the generalization power of the model, which may explain the fall in performance on the test set.

SVM exhibited slightly better results with the RBF kernel than with the polynomial kernel of degree one (SVM-PK-D1). However, this performance was achieved with almost five times more descriptors. In addition, we must consider that the polynomial kernel of first order is a linear equation. Thus, we can conclude that SVM with the polynomial kernel of first order provides the simplest and best-performing solutions. Therefore, we selected this framework to develop our models.

Section SI-3. Description of the feature selection process for the protein-protein BA modeling.

Initially, each instance in the dataset was represented as a vector of 23 040 molecular descriptors generated with ProtDCal, creating an initial matrix of 653 instances x 23.040 descriptors. For identifying the subset of attributes that best approximates the binding free energy value, we applied unsupervised and supervised features selection techniques.

First, we reduced the matrix dimensionality to 9 004 descriptors by applying the *RemoveUseless* filter that eliminates the attributes with none or too much variation for all the instances in the training set. A descriptor that cannot be calculated for an instance represents a missing value in the final vector. For a regression problem, this means that missing values need to be replaced by a number (mean, median, ...). Since this could add more noise to the dataset, we handled this issue by deleting all the attributes that contained at least one instance with a missing value, reducing the dataset to 1 477 attributes.

Then, we made use of the supervised method *CorrelationAttributeEval*, which orders all the attributes by their Pearson's correlation coefficient with the class. The highest correlation of an attribute with the class was 0.25, while the minimum was -0.32. We selected those attributes with correlation values between $0.1 <= R <= -0.1$, which were 447 attributes.

Next, we applied the filter *InterquartileRange* implemented in Weka to identify instances with extreme values. This method flags a descriptor of an instance as extreme if its value is greater than the 75th quartile or if it is minor than the 25th quartile, by the product of an extreme value factor and the interquartile range. We kept the default value of the filter, extremeValuesFactor = 6, and we removed those instances with more than the 5% of the amount of attributes flagged as extreme cases. This way we reduced the training set to 648 instances with 447 attributes.

Finally, for obtaining the best subset of attributes to train the final model, we applied the *WrapperSubsetEval* technique for the selection of attributes. This supervised method evaluates subsets of attributes by training a classifier and assessing its performance in cross-validation. The classifier we used was Support Vector Machine for regression, SMOreg package of Weka, with the polynomial kernel. The selection of the subsets is held by a search method. Here, we employed a genetic algorithm with a population of 20 individuals, crossover and mutation probabilities of 0.6 and 0.033 respectively and 20 generations maximum. Each subset was evaluated in 5-folds cross-validation, and the selection of the best one was carried out attending to the correlation coefficient achieved by the classifier. This step reduced the dataset to 26 structural features to train and test our model. The list of final descriptors can be found in the Supporting Information ppro_project.idl file. The file can be directly uploaded in ProtDCal-Suite[41] to calculate the descriptors of protein-protein complexes.

Table SI-3.1. Descriptors of the protein-protein model.

The evaluation measures are the Pearson's correlation coefficient (R), the Spearman's rank correlation coefficient (R_S), and the Kendall's (R_K) rank correlation coefficient between each descriptor and the binding affinity values.

Descriptor	Description	R	R_S	R_K
wNc(ECI)_NO_AHR_G	Geometric mean (G) of the weighted number of contacts (wNc) of the common residues in Alfa Helix structure (AHR)	-0.20	-0.18	-0.12

wNc(ECI)_NO_ALR_G	Geometric mean (G) of the weighted number of contacts (wNc) of the aliphatic residues (ALR)	-0.17	-0.16	-0.11
wNc(Z2)_NO_PLR_V	Variance (V) of the weighted number of contacts (wNc) of the polar residues (PLR)	-0.11	-0.05	-0.03
wNc(Z2)_NO_PCR_G	Geometric mean (G) of the weighted number of contacts (wNc) of the positive charged residues (PCR)	-0.11	-0.11	-0.07
wNc(Z3)_NO_GLU_V	Variance (V) of the weighted number of contacts (wNc) of the glutamic acid residues (GLU)	0.12	0.13	0.09
wNc(Z3)_NO_PLR_P2	Potential mean (P2) of the weighted number of contacts (wNc) of the polar residues (PLR)	0.15	0.19	0.13
wNc(Z3)_NO_PRT_P2	Potential mean (P2) of the weighted number of contacts (wNc) of the whole protein (PRT)	0.16	0.20	0.14
wFLC(ECI)_NO_ILE_P2	Potential mean (P2) of the weighted fraction of local contacts (wFLC) of the isoleucine residues (ILE)	0.16	0.19	0.14
wFLC(IP)_NO_PCR_Ar	Arithmetic mean (Ar) of the weighted fraction of local contacts (wFLC) of the positive charged residues (PCR)	0.18	0.22	0.15
wFLC(IP)_NO_PCR_V	Variance (V) of the weighted fraction of local contacts (wFLC) of the positive charged residues (PCR)	0.12	0.14	0.11
wFLC(ISA)_NO_PLR_Ar	Arithmetic mean (Ar) of the weighted fraction of local contacts (wFLC) of the polar residues (PLR)	0.18	0.22	0.17
wNLC(ECI)_NO_AHR_V	Variance (V) of the weighted number of local contacts (wNLC) of the common residues in Alfa Helix structure (AHR)	0.15	0.16	0.11
wNLC(ECI)_NO_NPR_N1	Manhattan distance (N1) of the weighted number of local contacts (wNLC) of the nonpolar residues (NPR)	-0.15	-0.20	-0.13
wNLC(ECI)_NO_NPR_DE	Standard deviation (DE) of the weighted number of local contacts (wNLC) of the nonpolar residues (NPR)	-0.15	-0.16	-0.10
wNLC(IP)_NO_BSR_N1	Manhattan distance (N1) of the weighted number of local contacts (wNLC) of the common residues in Beta Sheet structure (BSR)	-0.14	-0.20	-0.13
wNLC(IP)_NO_PLR_N1	Manhattan distance (N1) of the weighted number of local contacts (wNLC) of the polar residues (PLR)	-0.12	-0.18	-0.12
wNLC(ISA)_NO_BSR_N2	Euclidean distance (N2) of the weighted number of local contacts (wNLC) of the common residues in Beta Sheet structure (BSR)	-0.13	-0.15	-0.10
wNLC(ISA)_NO_PLR_G	Geometric mean (G) of the weighted number of local contacts (wNLC) of the polar residues (PLR)	0.15	0.16	0.11
wNLC(Z1)_NO_AHR_DE	Standard deviation (DE) of the weighted number of local contacts (wNLC) of the common residues in Alfa Helix structure (AHR)	0.13	0.15	0.10

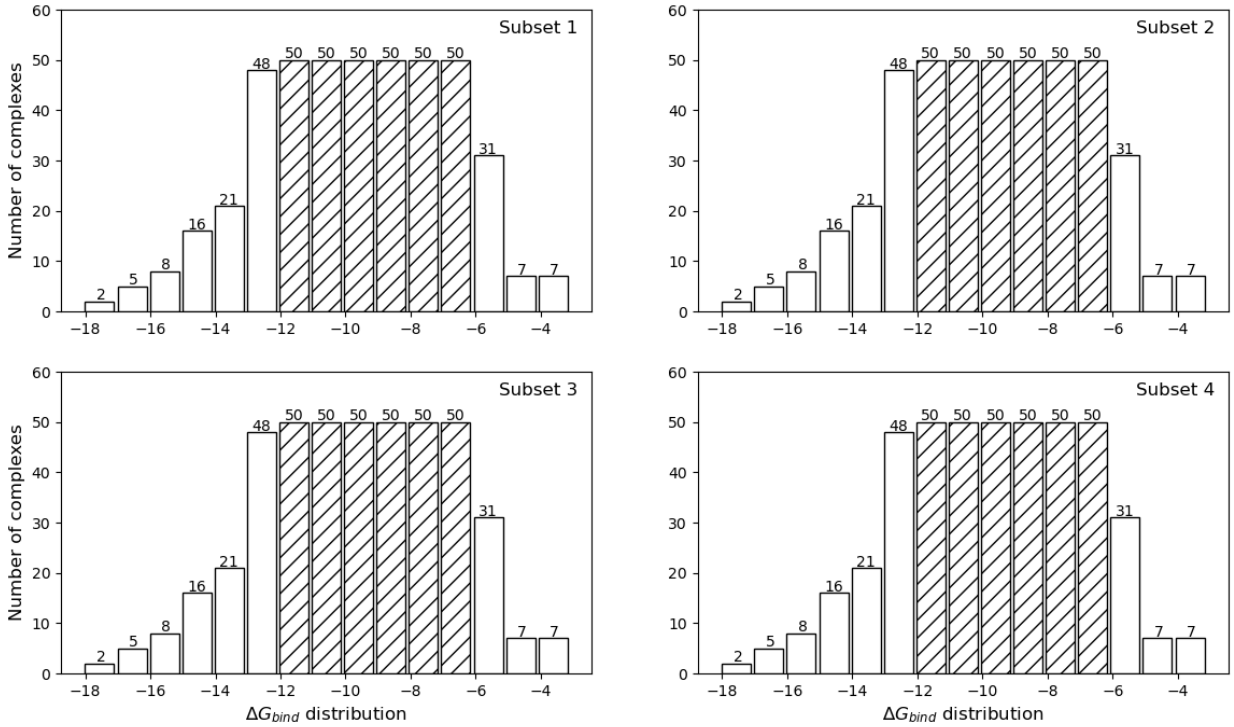
wNLC(Z1)_NO_ALR_N2	Euclidean distance (N2) of the weighted number of local contacts (wNLC) of the aliphatic residues (ALR)	-0.13	-0.13	-0.09
wNLC(Z1)_NO_NCR_P2	Potential mean (P2) of the weighted number of local contacts (wNLC) of the negative charged residues (NCR)	-0.17	-0.15	-0.10
wNLC(Z1)_NO_PRT_Ar	Arithmetic mean (Ar) of the weighted number of local contacts (wNLC) of the whole protein (PRT)	-0.16	-0.15	-0.10
wNLC(Z2)_NO_BSR_G	Geometric mean (G) of the weighted number of local contacts (wNLC) of the common residues in Beta Sheet structure (BSR)	-0.13	-0.12	-0.08
wNLC(Z2)_NO_ALR_N1	Manhattan distance (N1) of the weighted number of local contacts (wNLC) of the aliphatic residues (ALR)	-0.32	-0.29	-0.20
wNLC(Z3)_NO_AHR_Ar	Arithmetic mean (Ar) of the weighted number of local contacts (wNLC) of the common residues in Alfa Helix structure (AHR)	0.25	0.25	0.16
wNLC(Z3)_NO_ALR_Ar	Arithmetic mean (Ar) of the weighted number of local contacts (wNLC) of the aliphatic residues (ALR)	0.14	0.13	0.09

The weights of the descriptors are:

- IP: Isoelectric Point
- ECI: Electronic Charge Index
- ISA: Isotropic Surface Area
- Z1: Combined measure of hydrophobicity related properties
- Z2: Combined measure of bulkiness related properties
- Z3: Combined measure of electron related properties

File SI-1 (separated). Configuration file for ProtDCal to compute the 26 structural descriptors.

Figure SI-2. Distribution of ΔG_{bind} values in the four subsets of the training data for the protein-protein ensemble learning protocol.



We defined 15 intervals in the scale of ΔG_{bind} values, according to the range in the entire data. We filled the intervals with a maximum of 50 instances by sampling (without replacement) the entire dataset. We iterated this procedure to create four subsets, each one containing 445 complexes distributed along the complete range of affinity values. The white bars denote the intervals that were repeated among the four datasets, while the stripped bars denote those intervals where the sampling was performed.

Table SI-3. Summary of the intermediate models and performance measures for the protein-protein BA modeling during the hyperparameters tuning process.

	Training subset 1				*Training subset 2*				*Training subset 3*				*Training subset 4*			
Model	**TS**	**CV**	**DEV**	**Score**	**TS**	**CV**	**DEV**	**Score**	**TS**	**CV**	**DEV**	**Score**	**TS**	**CV**	**DEV**	**Score**
PK_C0.03125_D1	0,514	0,476	0,343	0,167	0,494	0,441	0,331	0,148	0,517	0,479	0,406	0,205	0,498	0,434	0,372	0,169
PK_C0.03125_D2	0,582	0,476	0,429	0,211	0,586	0,455	0,468	0,222	0,588	0,509	0,493	0,265	0,578	0,461	0,459	0,222
PK_C0.03125_D3	0,702	0,460	0,426	0,145	0,696	0,418	0,484	0,161	0,709	0,490	0,497	0,227	0,684	0,385	0,464	0,124
PK_C0.04419_D1	0,525	0,486	0,369	0,186	0,503	0,446	0,342	0,156	0,527	0,493	0,426	0,221	0,508	0,450	0,395	0,187
PK_C0.04419_D2	0,597	0,483	0,427	0,211	0,599	0,464	0,486	0,235	**0,602**	**0,508**	**0,495**	**0,266**	0,588	0,468	0,470	0,230
PK_C0.04419_D3	0,720	0,461	0,440	0,146	0,713	0,409	0,479	0,138	0,725	0,486	0,492	0,211	0,699	0,374	0,446	0,087
PK_C0.0625_D1	0,534	0,495	0,385	0,198	0,512	0,452	0,372	0,175	0,534	0,501	0,433	0,228	0,514	0,462	0,400	0,195
PK_C0.0625_D2	0,611	0,490	0,419	0,206	0,614	0,463	0,492	0,236	0,618	0,507	0,488	0,260	0,598	0,470	0,476	0,233
PK_C0.0625_D3	0,739	0,462	0,455	0,147	0,731	0,400	0,471	0,108	0,745	0,480	0,469	0,173	0,715	0,365	0,442	0,060
PK_C0.08838_D1	0,542	0,492	0,411	0,212	0,522	0,458	0,392	0,188	0,542	0,505	0,442	0,235	0,520	0,470	0,405	0,201
PK_C0.08838_D2	0,627	0,494	0,419	0,202	0,632	0,457	0,501	0,233	0,633	0,514	0,481	0,257	0,611	0,461	0,478	0,227
PK_C0.08838_D3	0,759	0,465	0,462	0,142	0,751	0,390	0,473	0,082	0,764	0,470	0,453	0,133	0,732	0,358	0,445	0,041
PK_C0.125_D1	0,551	0,493	0,417	0,216	0,531	0,463	0,409	0,199	0,547	0,508	0,442	0,237	0,524	0,474	0,406	0,202
PK_C0.125_D2	0,645	0,494	0,416	0,193	0,649	0,455	0,495	0,223	0,649	0,516	0,482	0,256	0,627	0,453	0,479	0,217
PK_C0.125_D3	0,778	0,466	0,464	0,129	0,774	0,383	0,462	0,041	0,784	0,461	0,428	0,080	0,750	0,354	0,444	0,016
PK_C0.17677_D1	0,556	0,501	0,410	0,215	0,539	0,473	0,417	0,207	0,551	0,511	0,448	0,241	0,530	0,470	0,423	0,210
PK_C0.17677_D2	0,665	0,494	0,411	0,180	0,661	0,458	0,493	0,219	0,664	0,514	0,486	0,254	0,641	0,444	0,471	0,201
PK_C0.17677_D3	0,796	0,462	0,452	0,094	0,794	0,377	0,457	0,007	0,805	0,450	0,409	0,026	0,770	0,342	0,450	-0,015
PK_C0.25_D1	0,562	0,506	0,409	0,216	0,547	0,476	0,422	0,212	0,555	0,512	0,455	0,246	0,533	0,468	0,438	0,217
PK_C0.25_D2	0,682	0,496	0,408	0,169	0,672	0,458	0,483	0,208	0,675	0,512	0,485	0,248	0,655	0,435	0,465	0,185
PK_C0.25_D3	0,814	0,454	0,442	0,057	0,815	0,357	0,452	-0,049	0,823	0,434	0,398	-0,028	0,792	0,327	0,435	-0,075
PK_C0.35355_D1	0,565	0,513	0,427	0,230	0,554	0,487	0,432	0,222	0,561	0,514	0,468	0,253	0,536	0,465	0,452	0,222
PK_C0.35355_D2	0,697	0,495	0,411	0,163	0,684	0,450	0,475	0,190	0,686	0,505	0,483	0,237	0,663	0,415	0,459	0,159
PK_C0.35355_D3	0,833	0,444	0,438	0,020	0,834	0,337	0,439	-0,115	0,842	0,411	0,394	-0,085	0,817	0,318	0,430	-0,126
PK_C0.5_D1	0,570	0,521	0,438	0,240	0,560	0,493	0,454	0,236	0,567	0,519	0,470	0,257	0,540	0,462	0,461	0,226
PK_C0.5_D2	0,710	0,490	0,424	0,163	0,695	0,436	0,475	0,171	0,698	0,501	0,483	0,229	0,672	0,394	0,452	0,130

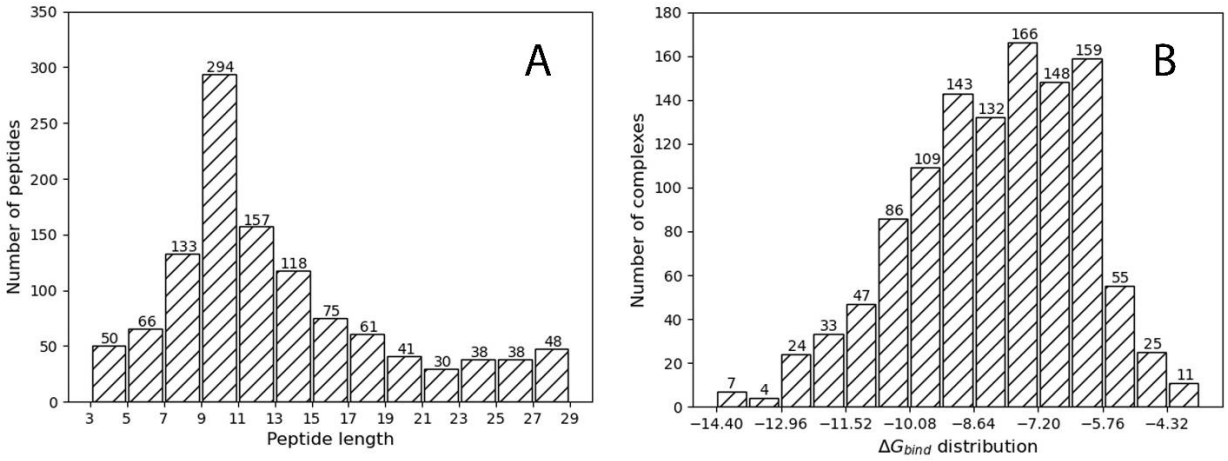
PK_C0.5_D3	0,852	0,439	0,422	-0,031	0,854	0,326	0,420	-0,182	0,859	0,393	0,376	-0,156	0,843	0,313	0,424	-0,179
PK_C0.7071_D1	0,574	0,525	0,449	0,248	0,566	0,494	0,456	0,238	0,573	0,522	0,471	0,259	0,544	0,463	0,451	0,221
PK_C0.7071_D2	0,723	0,481	0,437	0,158	0,705	0,424	0,487	0,163	0,710	0,495	0,467	0,205	0,683	0,374	0,450	0,102
PK_C0.7071_D3	0,870	0,434	0,407	-0,079	0,872	0,316	0,394	-0,260	0,877	0,377	0,369	-0,216	0,866	0,293	0,415	-0,256
PK_C1.41421_D1	0,581	0,526	0,471	0,262	0,574	0,505	0,475	0,254	0,580	0,524	0,471	0,261	0,553	0,469	0,467	0,232
PK_C1.41421_D2	0,751	0,472	0,448	0,140	0,733	0,394	0,493	0,119	0,739	0,476	0,416	0,122	0,702	0,354	0,427	0,048
PK_C1.41421_D3	0,901	0,412	0,384	-0,188	0,906	0,297	0,352	-0,409	0,912	0,357	0,338	-0,352	0,901	0,255	0,370	-0,441
PK_C1_D1	0,577	0,527	0,466	0,259	0,570	0,500	0,460	0,243	0,577	0,524	0,472	0,261	0,550	0,466	0,462	0,228
PK_C1_D2	0,737	0,476	0,448	0,155	0,719	0,408	0,501	0,150	0,725	0,488	0,442	0,168	0,691	0,363	0,436	0,075
PK_C1_D3	0,886	0,429	0,399	-0,120	0,889	0,310	0,373	-0,326	0,894	0,371	0,357	-0,270	0,885	0,275	0,394	-0,346
PK_C11.3137_D1	0,588	0,532	0,486	0,273	**0,581**	**0,507**	**0,500**	**0,268**	0,586	0,523	0,472	0,261	**0,560**	**0,466**	**0,483**	**0,238**
PK_C11.3137_D2	0,810	0,403	0,365	-0,087	0,811	0,268	0,406	-0,214	0,804	0,387	0,240	-0,266	0,788	0,298	0,396	-0,149
PK_C11.3137_D3	0,981	0,309	0,224	-0,770	0,984	0,235	0,091	-1,167	0,991	0,238	0,184	-0,995	0,981	0,180	0,271	-0,917
PK_C16_D1	0,589	0,529	0,491	0,274	0,581	0,505	0,500	0,267	0,585	0,524	0,472	0,261	0,560	0,464	0,483	0,237
PK_C16_D2	0,818	0,382	0,360	-0,130	0,820	0,256	0,391	-0,264	0,812	0,363	0,199	-0,368	0,796	0,286	0,381	-0,195
PK_C16_D3	0,989	0,296	0,218	-0,826	0,991	0,225	0,059	-1,274	0,996	0,224	0,159	-1,085	0,989	0,163	0,231	-1,045
PK_C2.82842_D1	0,586	0,528	0,487	0,271	0,579	0,511	0,489	0,264	0,587	0,524	0,475	0,263	0,558	0,467	0,473	0,234
PK_C2.82842_D2	0,772	0,451	0,410	0,062	0,755	0,367	0,454	0,035	0,764	0,450	0,375	0,030	0,729	0,336	0,443	0,016
PK_C2.82842_D3	0,933	0,368	0,330	-0,387	0,938	0,272	0,295	-0,604	0,941	0,335	0,283	-0,532	0,932	0,215	0,346	-0,609
PK_C2_D1	0,584	0,529	0,486	0,271	0,574	0,509	0,480	0,258	0,586	0,525	0,473	0,262	0,555	0,469	0,467	0,232
PK_C2_D2	0,762	0,466	0,435	0,113	0,745	0,380	0,469	0,073	0,752	0,463	0,390	0,072	0,715	0,347	0,432	0,032
PK_C2_D3	0,918	0,387	0,361	-0,283	0,923	0,282	0,326	-0,507	0,929	0,343	0,305	-0,457	0,916	0,235	0,357	-0,524
PK_C4_D1	0,587	0,533	0,488	0,275	0,582	0,511	0,492	0,266	0,586	0,527	0,470	0,262	0,559	0,465	0,472	0,232
PK_C4_D2	0,784	0,440	0,382	0,007	0,767	0,344	0,442	-0,017	0,776	0,436	0,349	-0,027	0,746	0,330	0,443	-0,009
PK_C4_D3	0,946	0,350	0,305	-0,481	0,951	0,258	0,246	-0,742	0,955	0,315	0,266	-0,622	0,946	0,208	0,339	-0,665
PK_C5.65685_D1	**0,588**	**0,533**	**0,488**	**0,275**	0,580	0,506	0,495	0,265	0,588	0,525	0,474	0,263	0,560	0,468	0,474	0,235
PK_C5.65685_D2	0,796	0,428	0,364	-0,041	0,781	0,313	0,428	-0,087	0,787	0,421	0,323	-0,089	0,761	0,321	0,439	-0,040
PK_C5.65685_D3	0,958	0,342	0,268	-0,583	0,964	0,249	0,199	-0,875	0,968	0,291	0,238	-0,743	0,959	0,198	0,329	-0,731
PK_C8_D1	0,588	0,533	0,486	0,274	0,581	0,507	0,495	0,266	0,587	0,525	0,478	0,265	0,560	0,468	0,480	0,238
PK_C8_D2	0,803	0,414	0,368	-0,061	0,798	0,284	0,424	-0,152	0,796	0,403	0,288	-0,168	0,774	0,309	0,419	-0,091
PK_C8_D3	0,970	0,327	0,237	-0,692	0,974	0,242	0,138	-1,032	0,981	0,260	0,214	-0,874	0,970	0,192	0,308	-0,804

Table SI-4. Summary of the performance of the individual and the ensemble models for protein-protein models.

The models M1, M2, M3 and M4 correspond to the best predictors obtained from the training subsets 1, 2, 3 and 4 respectively. Then, the correlation coefficients (R) of the estimations in the development set were calculated for each model (R_IND), as well as all possible combinations of the models. The combination rules were the average (V_AVG), maximum (V_MAX), and minimum, (V_MIN) predictions. The optimal ensemble model corresponds to the model that outputs the binding affinity based on the minimum predicted value between the models obtained from the training subsets 2 and 3.

Ensemble	*Model*	*R_IND*	*V_AVG*	*V_MAX*	*V_MIN*
1	M1	0,488	0,501	0,508	0,489
	M2	0,500			
2	M1	0,488	0,505	0,479	0,520
	M3	0,495			
3	M1	0,488	0,491	0,482	0,497
	M4	0,482			
4	M2	0,500	0,508	0,480	**0,526**
	M3	0,495			
5	M2	0,500	0,496	0,496	0,493
	M4	0,482			
6	M3	0,495	0,500	0,483	0,508
	M4	0,482			
7	M1	0,488	0,508	0,482	0,517
	M2	0,500			
	M3	0,495			
8	M1	0,488	0,498	0,496	0,492
	M2	0,500			
	M4	0,482			
9	M1	0,488	0,502	0,479	
	M3	0,495			
	M4	0,482			
10	M2	0,500	0,505	0,473	0,507
	M3	0,495			
	M4	0,482			
11	M1	0,488	0,505	0,479	0,508
	M2	0,500			
	M3	0,495			
	M4	0,482			

Figure SI-3. Characterization of the dataset of protein-peptide complexes used in this work.



The distribution of the peptide lengths (number of residues) in the dataset is presented in panel A, while panel B shows the distribution of ΔG values across all the protein-peptide complexes contained in the dataset.

Section SI-4. Description of the process of feature selection for creating the protein-peptide model.

Here, we followed the pipeline described in section SI-2 for selecting the attributes of the final model. Starting with a matrix of 949 instances x 23 040 descriptors generated with ProtDCal, we reduced the matrix dimensionality to 8 999 attributes by applying the filter *RemoveUseless*. Then, we deleted all the attributes that contained at least one instance with a missing value, reducing the dataset to 2 358 attributes. In a third step, we made use of the supervised method *CorrelationAttributeEval*, for correlating each attribute with the class. The highest correlation obtained was 0.24 and the minimum -0.23. We selected those attributes with correlation values between $0.1 <= R <= -0.1$ that were 631 attributes.

At this point, after decreasing the dimensionality of the problem by more than 95%, we applied the filter *InterquartileRange* implemented in Weka to identify instances with extreme values. We kept the default value of the filter, which is the extremeValuesFactor = 6 and we removed those instances with more than the 5% of number of attributes flagged as extreme cases. This way we reduced the training set to 922 instances with 631 attributes.

Finally, for obtaining the best subset of attributes to train the final model we applied the *WrapperSubsetEval* attribute selection technique. The classifier we used was support vector machine for regression with the lineal kernel and the search method a genetic algorithm with a population of 20 individuals, crossover and mutation probabilities of 0.6 and 0.033 respectively and 20 generations maximum. Each subset was evaluated in 5-folds cross-validation, and the selection of the best subset was attending to the correlation coefficient achieved by the classifier. This step reduced the dataset to 37 structural features to train and test the models.

Table SI-4.1. Descriptors of the protein-peptide model.

The evaluation measures are the Pearson's correlation coefficient (R), the Spearman's rank correlation coefficient (R_S), and the Kendall's (R_K) rank correlation coefficient between each descriptor and the binding affinity values.

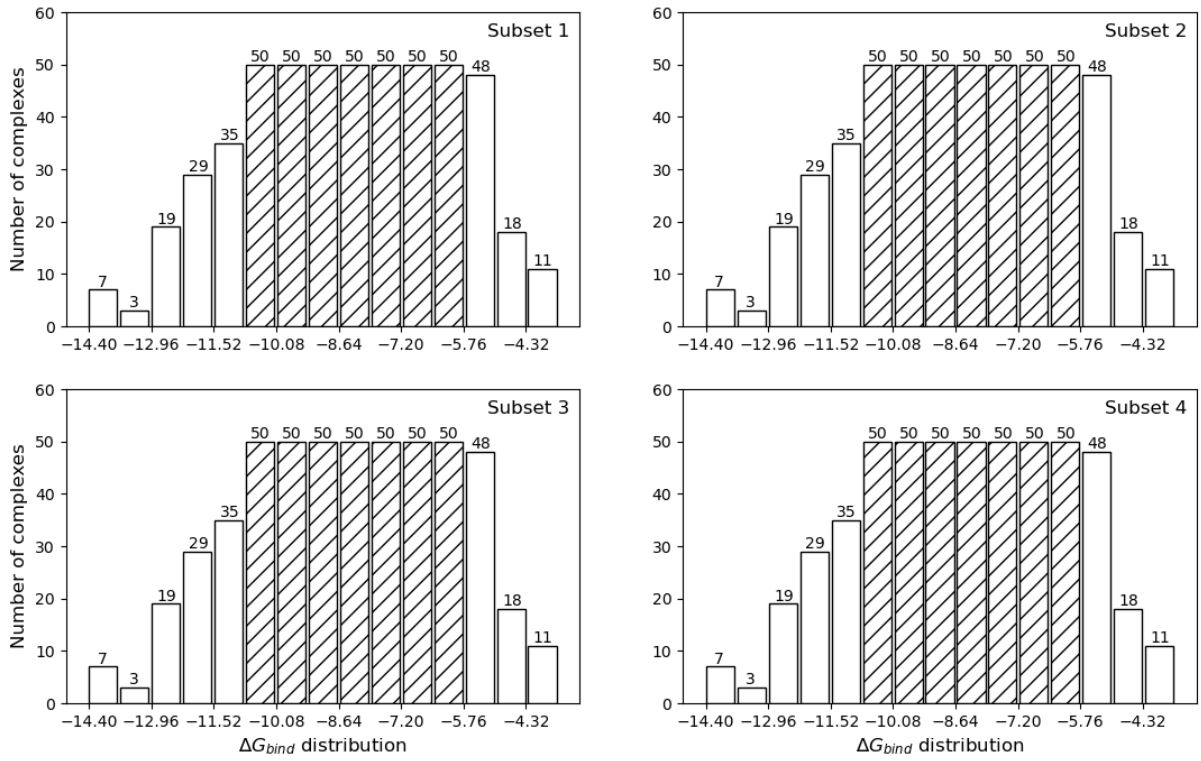
Descriptor	Description	R	R_S	R_K
wNc(ECI)_NO_AHR_N1	Manhattan distance (N1) of the weighted number of contacts (wNc) of the common residues in Alfa Helix structure (AHR)	-0.20	-0.17	-0.11
wNc(ECI)_NO_ALR_N1	Manhattan distance (N1) of the weighted number of contacts (wNc) of the aliphatic residues (ALR)	-0.13	-0.05	-0.04
wNc(IP)_NO_PLR_N1	Manhattan distance (N1) of the weighted number of contacts (wNc) of the polar residues (PLR)	-0.13	-0.12	-0.08
wNc(ISA)_NO_PLR_V	Variance (V) of the weighted number of contacts (wNc) of the polar residues (PLR)	0.10	0.10	0.06
wNc(Z1)_NO_NPR_N2	Euclidean distance (N2) of the weighted number of contacts (wNc) of the nonpolar residues (NPR)	-0.19	-0.17	-0.12

wNc(Z1)_NO_ARM_V	Variance (V) of the weighted number of contacts (wNc) of the aromatic residues (ARM)	-0.11	-0.05	-0.04
wNc(Z1)_NO_PRT_N1	Manhattan distance (N1) of the weighted number of contacts (wNc) of the whole protein (PRT)	-0.12	-0.11	-0.07
wNc(Z3)_NO_AHR_N1	Manhattan distance (N1) of the weighted number of contacts (wNc) of the common residues in Alfa Helix structure (AHR)	-0.17	-0.11	-0.07
wNc(Z3)_NO_UCR_N1	Manhattan distance (N1) of the weighted number of contacts (wNc) of the uncharged residues (UCR)	-0.13	-0.07	-0.04
wFLC(ECI)_NO_AHR_SI30	Standardized Information Content (SI30) of the weighted fraction of local contacts (wFLC) of the common residues in Alfa Helix structure (AHR)	0.16	0.15	0.10
wFLC(ECI)_NO_PCR_V	Variance (V) of the weighted fraction of local contacts (wFLC) of the positive charged residues (PCR)	0.11	0.09	0.07
wFLC(IP)_NO_RTR_MI30	Mean Information Content (MI30) of the weighted fraction of local contacts (wFLC) of the common residues in reverse turn (RTR)	0.12	0.07	0.05
wFLC(IP)_NO_BSR_Ar	Arithmetic mean (Ar) of the weighted fraction of local contacts (wFLC) of the common residues in Beta Sheet structure (BSR)	0.11	0.17	0.12
wFLC(IP)_NO_UCR_RA	Range (RA) of the weighted fraction of local contacts (wFLC) of the uncharged residues (UCR)	0.13	0.12	0.09
wFLC(IP)_NO_UCR_K	Kurtosis (K) of the weighted fraction of local contacts (wFLC) of the uncharged residues (UCR)	-0.17	-0.21	-0.14
wFLC(IP)_NO_PRT_TI30	Total content (TI30) of the weighted fraction of local contacts (wFLC) of the whole protein (PRT)	-0.20	-0.22	-0.15
wFLC(ISA)_NO_NPR_P2	Potential mean (P2) of the weighted fraction of local contacts (wFLC) of the nonpolar residues (NPR)	0.24	0.26	0.18
wFLC(ISA)_NO_NPR_DE	Standard deviation (DE) of the weighted fraction of local contacts (wFLC) of the nonpolar residues (NPR)	0.24	0.27	0.18
wFLC(ISA)_NO_PRT_TI30	Total content (TI30) of the weighted fraction of local contacts (wFLC) of the whole protein (PRT)	-0.19	-0.20	-0.13
wFLC(Z1)_NO_BSR_SI30	Standardized Information Content (SI30) of the weighted fraction of local contacts (wFLC) of the common residues in Beta Sheet structure (BSR)	0.20	0.22	0.14
wFLC(Z2)_NO_PLR_K	Kurtosis (K) of the weighted fraction of local contacts (wFLC) of the polar residues (PLR)	-0.12	-0.15	-0.10
wNLC(ECI)_NO_BSR_I50	Interquartile range (I50) of the weighted number of local contacts (wNLC) of the	-0.10	-0.09	-0.06

	common residues in Beta Sheet structure (BSR)			
wNLC(ECI)_NO_PLR_K	Kurtosis (K) of the weighted number of local contacts (wNLC) of the polar residues (PLR)	-0.12	-0.17	-0.11
wNLC(ECI)_NO_UCR_S	Skewness (S) of the weighted number of local contacts (wNLC) of the uncharged residues (UCR)	-0.14	-0.13	-0.09
wNLC(IP)_NO_RTR_V	Variance (V) of the weighted number of local contacts (wNLC) of the common residues in reverse turn (RTR)	0.11	0.09	0.06
wNLC(IP)_NO_UCR_I50	Interquartile range (I50) of the weighted number of local contacts (wNLC) of the uncharged residues (UCR)	-0.12	-0.09	-0.06
wNLC(ISA)_NO_UCR_S	Skewness (S) of the weighted number of local contacts (wNLC) of the uncharged residues (UCR)	-0.20	-0.22	-0.14
wNLC(Z1)_NO_RTR_SI30	Standardized Information Content (SI30) of the weighted number of local contacts (wNLC) of the common residues in reverse turn (RTR)	0.12	0.13	0.09
wNLC(Z1)_NO_BSR_K	Kurtosis (K) of the weighted number of local contacts (wNLC) of the common residues in Beta Sheet structure (BSR)	-0.10	-0.17	-0.11
wNLC(Z1)_NO_UCR_K	Kurtosis (K) of the weighted number of local contacts (wNLC) of the uncharged residues (UCR)	-0.13	-0.17	-0.11
wNLC(Z2)_NO_AHR_P2	Potential mean (P2) of the weighted number of local contacts (wNLC) of the common residues in Alfa Helix structure (AHR)	-0.17	-0.19	-0.13
wNLC(Z2)_NO_PCR_N2	Euclidean distance (N2) of the weighted number of local contacts (wNLC) of the positive charged residues (PCR)	-0.15	-0.15	-0.10
wNLC(Z2)_NO_NCR_P2	Potential mean (P2) of the weighted number of local contacts (wNLC) of the negative charged residues (NCR)	-0.11	-0.09	-0.06
wNLC(Z2)_NO_UCR_K	Kurtosis (K) of the weighted number of local contacts (wNLC) of the uncharged residues (UCR)	-0.11	-0.17	-0.12
wNLC(Z3)_NO_PLR_SI30	Standardized Information Content (SI30) of the weighted number of local contacts (wNLC) of the polar residues (PLR)	0.12	0.13	0.08
wNLC(Z3)_NO_NCR_RA	Range (RA) of the weighted number of local contacts (wNLC) of the negative charged residues (NCR)	-0.13	-0.15	-0.10
wNLC(Z3)_NO_UCR_TI30	Total content (TI) of the weighted number of local contacts (wNLC) of the uncharged residues (UCR)	-0.19	-0.20	-0.14

File SI-2 (separated). Configuration file for ProtDCal to compute the 37 structural descriptors.

Figure SI-4. Distribution of ΔG_{bind} values in the four subsets of the training data for the protein-peptide ensemble learning protocol.



We defined 15 intervals in the scale of ΔG_{bind} values, according to the rage in the entire data. Then, we filled the intervals with a maximum of 50 instances by sampling (without replacement) the entire dataset. We iterated this procedure to create four subsets, each one containing 520 complexes distributed along the complete range of affinity values. The white bars denote the intervals that were repeated among the four datasets, while the stripped bars denote those intervals where the sampling was performed.

Table SI-5. Summary of the intermediate models and performance measures for the protein-peptide modeling during the hyperparameters tuning process.

	Training 1				***Training 2***				***Training 3***				***Training 4***			
Model	**TS**	**CV**	**DEV**	**Score**	**TS**	**CV**	**DEV**	**Score**	**TS**	**CV**	**DEV**	**Score**	**TS**	**CV**	**DEV**	**Score**
PK_C0.0078125_D1	0,503	0,463	0,481	0,230	0,480	0,441	0,406	0,189	0,494	0,448	0,430	0,203	0,494	0,464	0,412	0,201
PK_C0.0078125_D2	0,636	0,574	0,529	0,321	0,625	0,549	0,539	0,313	0,630	0,550	0,519	0,302	0,639	0,577	0,505	0,307
PK_C0.0078125_D3	0,735	0,586	0,465	0,259	0,728	0,535	0,506	0,261	0,726	0,549	0,556	0,313	0,722	0,539	0,555	0,305
PK_C0.0110485_D1	0,529	0,486	0,495	0,250	0,507	0,468	0,437	0,215	0,522	0,468	0,449	0,222	0,517	0,484	0,440	0,223
PK_C0.0110485_D2	0,644	0,578	0,527	0,322	0,637	0,557	0,546	0,322	0,642	0,557	0,525	0,309	0,648	0,578	0,518	0,316
PK_C0.0110485_D3	0,755	0,586	0,467	0,251	0,746	0,528	0,490	0,233	0,748	0,538	0,558	0,297	0,740	0,529	0,550	0,287
PK_C0.015625_D1	0,555	0,510	0,499	0,267	0,531	0,489	0,463	0,238	0,549	0,492	0,452	0,235	0,545	0,502	0,462	0,244
PK_C0.015625_D2	**0,654**	**0,581**	**0,527**	**0,324**	**0,650**	**0,564**	**0,539**	**0,322**	0,653	0,562	0,533	0,317	0,657	0,578	0,524	0,320
PK_C0.015625_D3	0,777	0,581	0,468	0,237	0,764	0,524	0,481	0,209	0,770	0,526	0,565	0,283	0,760	0,521	0,550	0,271
PK_C0.022097_D1	0,575	0,532	0,503	0,281	0,555	0,504	0,485	0,257	0,573	0,509	0,453	0,243	0,569	0,523	0,479	0,264
PK_C0.022097_D2	0,664	0,583	0,521	0,320	0,661	0,564	0,537	0,320	0,666	0,563	0,538	0,320	0,665	0,576	0,526	0,320
PK_C0.022097_D3	0,798	0,571	0,472	0,218	0,786	0,521	0,467	0,178	0,792	0,512	0,563	0,256	0,782	0,513	0,565	0,264
PK_C0.03125_D1	0,592	0,542	0,507	0,290	0,574	0,523	0,492	0,271	0,586	0,526	0,467	0,259	0,587	0,539	0,478	0,272
PK_C0.03125_D2	0,678	0,585	0,521	0,320	0,671	0,561	0,535	0,316	**0,678**	**0,562**	**0,543**	**0,322**	**0,675**	**0,570**	**0,532**	**0,320**
PK_C0.03125_D3	0,818	0,562	0,470	0,193	0,807	0,517	0,465	0,156	0,813	0,497	0,565	0,229	0,802	0,502	0,575	0,250
PK_C0.0441941_D1	0,601	0,551	0,504	0,292	0,584	0,535	0,502	0,282	0,596	0,537	0,465	0,263	0,595	0,548	0,477	0,275
PK_C0.0441941_D2	0,694	0,588	0,517	0,317	0,685	0,559	0,540	0,317	0,692	0,561	0,546	0,321	0,685	0,563	0,537	0,317
PK_C0.0441941_D3	0,837	0,554	0,448	0,144	0,830	0,512	0,475	0,139	0,835	0,492	0,551	0,194	0,821	0,487	0,587	0,234
PK_C0.0625_D1	0,605	0,556	0,500	0,293	0,592	0,543	0,515	0,294	0,600	0,543	0,455	0,259	0,602	0,551	0,490	0,285
PK_C0.0625_D2	0,713	0,591	0,516	0,314	0,701	0,555	0,536	0,308	0,707	0,555	0,543	0,312	0,698	0,553	0,546	0,315
PK_C0.0625_D3	0,856	0,542	0,423	0,082	0,853	0,507	0,466	0,102	0,854	0,493	0,544	0,170	0,840	0,471	0,585	0,199
PK_C0.0883883_D1	0,607	0,558	0,505	0,297	0,597	0,547	0,515	0,297	0,604	0,545	0,460	0,263	0,612	0,556	0,484	0,284
PK_C0.0883883_D2	0,732	0,598	0,511	0,310	0,718	0,547	0,505	0,274	0,725	0,547	0,537	0,297	0,714	0,543	0,550	0,306
PK_C0.0883883_D3	0,874	0,525	0,436	0,060	0,874	0,498	0,439	0,033	0,873	0,493	0,508	0,112	0,858	0,460	0,579	0,163
PK_C0.125_D1	0,610	0,561	0,505	0,299	0,605	0,557	0,504	0,296	0,607	0,545	0,476	0,274	0,617	0,563	0,471	0,279
PK_C0.125_D2	0,749	0,602	0,502	0,299	0,734	0,540	0,472	0,232	0,742	0,541	0,526	0,277	0,730	0,528	0,537	0,280

PK_C0.125_D3	0,893	0,512	0,488	0,088	0,895	0,486	0,413	-0,041	0,891	0,492	0,479	0,056	0,877	0,456	0,553	0,113
PK_C0.1767766_D1	0,613	0,564	0,508	0,302	0,610	0,562	0,500	0,296	0,614	0,548	0,489	0,282	0,621	0,575	0,465	0,280
PK_C0.1767766_D2	0,767	0,596	0,488	0,273	0,751	0,537	0,450	0,199	0,759	0,535	0,503	0,244	0,746	0,515	0,537	0,263
PK_C0.1767766_D3	0,911	0,494	0,537	0,106	0,914	0,473	0,404	-0,098	0,907	0,484	0,480	0,028	0,895	0,454	0,466	-0,013
PK_C0.25_D1	0,615	0,562	0,508	0,301	0,613	0,565	0,500	0,297	0,617	0,550	0,487	0,283	0,623	0,577	0,467	0,282
PK_C0.25_D2	0,786	0,590	0,480	0,251	0,766	0,535	0,435	0,171	0,778	0,524	0,501	0,221	0,763	0,501	0,540	0,244
PK_C0.25_D3	0,926	0,477	0,525	0,049	0,931	0,460	0,382	-0,174	0,925	0,469	0,524	0,039	0,911	0,446	0,413	-0,116
PK_C0.3535533_D1	0,616	0,561	0,503	0,298	0,613	0,565	0,505	0,301	0,620	0,554	0,488	0,285	0,624	0,573	0,472	0,284
PK_C0.3535533_D2	0,805	0,582	0,464	0,214	0,784	0,527	0,432	0,148	0,800	0,511	0,525	0,216	0,781	0,485	0,554	0,229
PK_C0.3535533_D3	0,940	0,463	0,513	-0,002	0,944	0,454	0,366	-0,229	0,943	0,447	0,618	0,095	0,927	0,437	0,374	-0,210
PK_C0.5_D1	0,619	0,564	0,501	0,298	0,612	0,562	0,501	0,297	0,622	0,554	0,493	0,289	0,624	0,569	0,476	0,285
PK_C0.5_D2	0,821	0,571	0,459	0,187	0,803	0,518	0,439	0,130	0,817	0,499	0,556	0,220	0,798	0,469	0,557	0,204
PK_C0.5_D3	0,951	0,451	0,526	-0,018	0,954	0,453	0,340	-0,290	0,956	0,428	0,660	0,098	0,941	0,428	0,350	-0,285
PK_C0.7071067_D1	0,620	0,569	0,498	0,298	0,613	0,560	0,502	0,297	0,623	0,554	0,494	0,289	0,626	0,566	0,490	0,292
PK_C0.7071067_D2	0,835	0,562	0,447	0,154	0,822	0,506	0,447	0,108	0,833	0,493	0,579	0,223	0,813	0,451	0,567	0,181
PK_C0.7071067_D3	0,962	0,434	0,500	-0,093	0,963	0,447	0,298	-0,385	0,967	0,408	0,612	-0,001	0,955	0,410	0,322	-0,381
PK_C1.4142135_D1	0,622	0,568	0,497	0,297	0,614	0,554	0,496	0,290	0,625	0,549	0,503	0,292	0,630	0,563	0,505	0,300
PK_C1.4142135_D2	0,864	0,532	0,441	0,086	0,858	0,487	0,421	0,018	0,860	0,486	0,559	0,174	0,843	0,418	0,577	0,125
PK_C1.4142135_D3	0,980	0,390	0,507	-0,179	0,974	0,430	0,261	-0,498	0,983	0,370	0,586	-0,115	0,973	0,372	0,298	-0,517
PK_C1_D1	0,622	0,569	0,498	0,299	0,612	0,557	0,501	0,295	0,624	0,549	0,503	0,292	0,628	0,564	0,497	0,296
PK_C1_D2	0,850	0,549	0,436	0,113	0,841	0,497	0,445	0,078	0,847	0,494	0,580	0,215	0,827	0,430	0,580	0,157
PK_C1_D3	0,972	0,413	0,516	-0,119	0,969	0,444	0,270	-0,450	0,976	0,386	0,578	-0,088	0,965	0,387	0,305	-0,463
PK_C11.3137084_D1	0,623	0,559	0,480	0,282	0,615	0,552	0,481	0,280	0,624	0,551	0,502	0,292	0,633	0,563	0,503	0,299
PK_C11.3137084_D2	0,935	0,382	0,504	-0,124	0,944	0,391	0,417	-0,243	0,939	0,372	0,453	-0,214	0,931	0,361	0,482	-0,178
PK_C11.3137084_D3	0,994	0,336	0,431	-0,406	0,991	0,336	0,236	-0,728	0,999	0,288	0,542	-0,341	0,992	0,297	0,480	-0,397
PK_C16_D1	0,624	0,558	0,477	0,280	0,615	0,551	0,482	0,280	0,624	0,552	0,505	0,294	0,633	0,563	0,503	0,299
PK_C16_D2	0,947	0,349	0,493	-0,208	0,953	0,387	0,402	-0,286	0,951	0,331	0,441	-0,315	0,944	0,343	0,440	-0,284
PK_C16_D3	0,997	0,329	0,423	-0,436	0,994	0,310	0,235	-0,781	1,000	0,277	0,514	-0,402	0,994	0,289	0,491	-0,401
PK_C2.8284271_D1	0,623	0,564	0,485	0,288	0,615	0,554	0,487	0,285	0,624	0,548	0,502	0,291	0,634	0,562	0,504	0,299
PK_C2.8284271_D2	0,890	0,491	0,475	0,050	0,892	0,445	0,398	-0,109	0,890	0,460	0,466	0,002	0,874	0,410	0,578	0,082
PK_C2.8284271_D3	0,986	0,373	0,474	-0,265	0,982	0,408	0,251	-0,564	0,991	0,337	0,560	-0,217	0,984	0,341	0,347	-0,508
PK_C2_D1	0,623	0,567	0,489	0,292	0,614	0,555	0,495	0,290	0,625	0,547	0,500	0,289	0,633	0,562	0,508	0,301

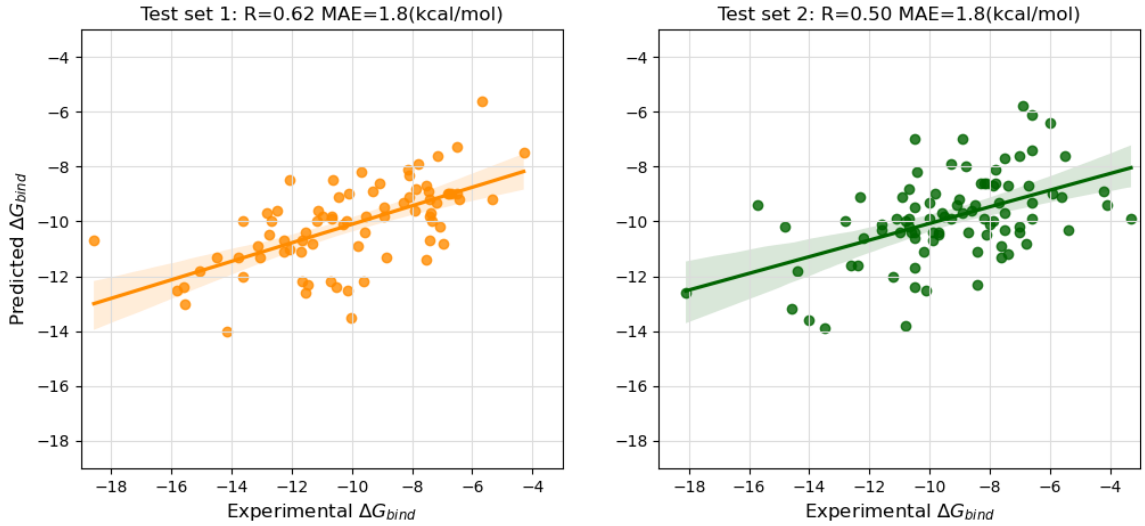
PK_C2_D2	0,877	0,513	0,455	0,067	0,876	0,466	0,397	-0,062	0,875	0,474	0,518	0,100	0,859	0,414	0,576	0,102
PK_C2_D3	0,983	0,382	0,495	-0,216	0,979	0,418	0,257	-0,531	0,988	0,355	0,592	-0,141	0,979	0,358	0,318	-0,517
PK_C4_D1	0,624	0,563	0,483	0,286	0,616	0,553	0,486	0,284	0,624	0,551	0,507	0,295	0,634	0,561	0,503	0,298
PK_C4_D2	0,902	0,468	0,499	0,037	0,907	0,432	0,410	-0,134	0,903	0,438	0,424	-0,100	0,890	0,396	0,569	0,035
PK_C4_D3	0,988	0,361	0,461	-0,306	0,984	0,400	0,247	-0,589	0,994	0,326	0,555	-0,247	0,986	0,328	0,396	-0,456
PK_C5.6568542_D1	0,624	0,562	0,483	0,285	0,615	0,554	0,484	0,283	0,623	0,549	0,505	0,293	0,633	0,562	0,502	0,298
PK_C5.6568542_D2	0,914	0,444	0,518	0,013	0,922	0,421	0,423	-0,155	0,917	0,425	0,435	-0,125	0,905	0,383	0,547	-0,027
PK_C5.6568542_D3	0,990	0,353	0,426	-0,376	0,986	0,386	0,249	-0,610	0,996	0,314	0,560	-0,265	0,988	0,316	0,436	-0,421
PK_C8_D1	0,624	0,560	0,481	0,283	0,614	0,553	0,481	0,280	0,623	0,549	0,505	0,293	0,634	0,562	0,504	0,299
PK_C8_D2	0,925	0,417	0,513	-0,046	0,933	0,403	0,417	-0,206	0,928	0,405	0,453	-0,145	0,919	0,375	0,524	-0,084
PK_C8_D3	0,992	0,344	0,414	-0,414	0,988	0,363	0,245	-0,660	0,997	0,303	0,551	-0,300	0,990	0,305	0,464	-0,403

Table SI-6. Summary of the performance of the individual and the ensemble models for protein-peptide models.

The models M1, M2, M3 and M4 correspond to the best predictors obtained from the training subsets 1, 2, 3 and 4 respectively. Then, the correlation coefficients (R) of the estimations in the development set were estimated for each model (R_IND), as well as for each possible combination of models. The combination rules were average (V_AVG), maximum (V_MAX), and minimum, (V_MIN) probabilities. The optimal ensemble model corresponds, from the group of ensembles two, to the model that outputs the binding affinity based on the maximum predicted value between the models obtained from the training subsets 1 and 3.

Ensembles	***Model***	***R_IND***	***V_AVG***	***V_MAX***	***V_MIN***
1	M1	0,527	0,537	0,552	0,519
	M2	0,539			
2	M1	0,527	0,540	**0,556**	0,519
	M3	0,543			
3	M1	0,527	0,536	0,532	0,533
	M4	0,532			
4	M2	0,539	0,545	0,548	0,539
	M3	0,543			
5	M2	0,539	0,542	0,532	0,546
	M4	0,532			
6	M3	0,543	0,544	0,543	0,540
	M4				
7	M1	0,527	0,542	0,559	0,520
	M2	0,539			
	M3	0,543			
8	M1	0,527	0,540	0,538	0,530
	M2	0,539			
	M4	0,532			
9	M1	0,527	0,542	0,548	0,530
	M3	0,543			
	M4	0,532			
10	M2	0,539	0,545	0,546	0,547
	M3	0,543			
	M4	0,532			
11	M1	0,527	0,543	0,553	0,529
	M2	0,539			
	M3	0,543			
	M4	0,532			

Figure SI-5. Plots of experimental vs. predicted BA values of PPI-Affinity on the test sets of protein – protein affinity data.



The performance is reported as the Pearson's Correlation coefficient (R) and the Mean Absolute Error (MAE) between experimental and predicted BA. Test set 1 corresponds to the benchmark of 79 complexes taken from Vangone and Bonvin.[1] Test set 2 corresponds to the hold-out set of 90 data points extracted from PDBbind (v.2020).[8]

Section SI-5. Performance of PPI-Affinity vs. a state-of-the-art biding affinity classifier.

Recently, Abbasi, W. A. et al.[32] developed a method using *Learning Using Privileged Information* (LUPI) paradigm and Support Vector Machines, that classifies as "high" or "low" the binding affinity of protein-protein complexes. The predictor, called LUPIA,[32] was trained using sequence and structure information. Nevertheless, in production, only the sequence information of protein pairs is required. To build the model, the authors discretized the BA values into two classes. For this, they used as threshold -10.86, which is the median value of the BA of the training dataset.
Here we used this value to discretize the output of PPI-Affinity and compare our protein-protein model with LUPIA.[32] The assessment was performed in two test sets (Table SI-5.1). Test set 1 corresponds to the hold-out set of 90 data points taken from PDBbind (v.2020).[8] We removed four cases that are in common with the training set of LUPIA,[32] and the evaluation was performed on the remaining 86 data points. The test set has protein-protein complexes with BA values between -18.1 and -3.3 kcal/mol. After discretizing such values, 15 and 71 data points were classified as "high" and "low" BA, respectively.
The Test set 2 corresponds to 26 wild-type and 151 mutants of protein-protein complexes taken from the SKEMPI v2.0[28] database. This data was employed to assess the performance of PPI-Affinity (R = 0.78 and MAE = 1.4 kcal/mol). When only considering the wild-types, the performance of the model was R = 0.77 and MAE = 1.1 kcal/mol.
Here, we employed the 26 wild-type complexes to compare our method to the LUPIA[32] classifier. The experimental binding affinity values of the protein pairs ranged between -16.3 and -7.0 kcal/mol. We used as reference the threshold value -10.86 and divided the data into 14 cases classified as "high" and 12 classified as complexes with "low" binding affinity.
The benchmark of 79 protein-protein complexes employed by Vangone and Bonvin[1] was not used in this comparison, as most of the cases were found in to be common with the training set of the LUPIA[32] predictor.

The performance measures used to evaluate the models were Sensitivity (Sn) and Specificity (Sp), formulated as:

$$Sn = TP \,/\, (TP + FN)$$

$$Sp = TN/(TN + FP)$$

where:
TP: number of protein-protein complexes correctly predicted as presenting "high" BA,
TN: number of protein-protein complexes correctly predicted as presenting "low" BA,
FP: number of protein-protein complexes incorrectly predicted as presenting "high" BA,
FN: number of protein-protein complexes incorrectly predicted as presenting "low" BA,

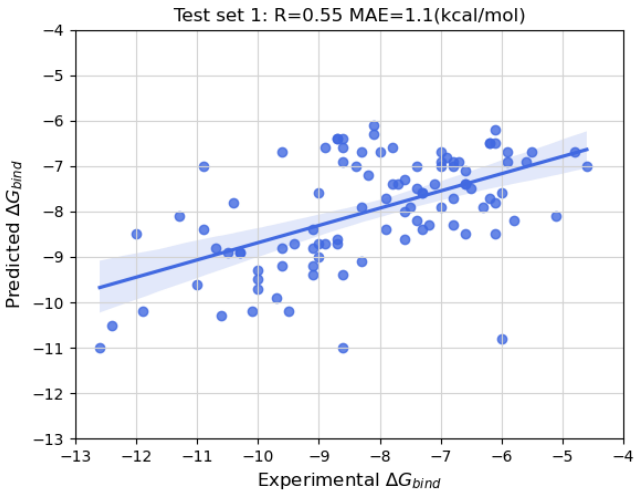
Table SI-5.1. Summary of the evaluation of PPI-Affinity and the LUPIA[32] classifier on two sets of protein – protein affinity data.

	Test set 1		Test set 2	
	LUPIA[32]	PPI-Affinity	LUPIA[32]	PPI-Affinity
Sn	**0.87**	0.47	**1.0**	0.86
Sp	0.32	**0.86**	0.17	**0.67**

The performance measures are Sensitivity (Sn) and Specificity (Sp). Test set 1 corresponds to a test set of 86 data points taken from PDBbind (v.2020),[8] while Test set 2 corresponds to 26 wild-type structures taken from the SKEMPI 2.0 dataset.[28]

The sensitivity of PPI-Affinity in Test set 1 decreased (Sn = 0.47). Nevertheless, of the 15 cases labeled as "High" in this data set, the eight misclassified differ by less than 1.7 kcal/mol from the threshold used, which is within the margin of error of the PPI-Affinity model (MAE = 1.8 kcal/mol). From them, six BA values differ by less than 1 kcal/mol. This reflects the downside of using a threshold to classify BA values. Furthermore, in both test sets, it can be seen that LUPIA[32] suffered from been too optimistic, as the method ranked most of the cases as with "High" affinity.

Figure SI-6. Plot of experimental vs. predicted by PPI-Affinity BA values on the test set of protein – peptide affinity data.



The performance is reported as the Pearson's Correlation coefficient (R) and the Mean Absolute Error (MAE) between experimental and predicted BA. The test set 1 corresponds to the hold-out set of 100 data points extracted from Biolip.[37]

Section SI-6. Description of the assays used to determine the binding affinities of EPI-X4 derivatives against the CRCX4 receptor**.**

To determine the affinity of the peptides for CXCR4, an antibody competition assay was used. The assay is based on the competitive binding of a fluorescently labelled anti-CXCR4 antibody (clone 12G5) with CXCR4 ligands.[42] For this 50,000 SupT1 cells/well were seeded in 96-well V-bottom microtiter plates in PBS supplemented with 1% FCS. The buffer was removed by centrifugation and cells precooled at 4°C for 15 min. The compound was serially diluted in PBS and added to the cells together with a 12G5 antibody at a constant concentration. Cells were incubated at 4°C for 2 hours before the unbound antibody and compounds were removed by 2 washing steps followed by fixation in 2% PFA. Mean fluorescence (MFI) of cells was determined by flow cytometry using FACS CytoFLEX. The isotype control was subtracted and values normalized to the 12G5-APC stained PBS control. IC_{50} values were determined by non-linear regression using GraphPad Prism.

Table SI-6.1 Summary of the binding affinities of EPI-X4 derivatives against CRCX4.

Derivative	**Sequence**	**IC_{50}(nM)**
JM#21	ILRWSRKLPCVS	77
JM#122	ILRWSRKLPSVS	130
JM#151	IVRWSKKVPSVS	131
JM#23	ILRWSRKVPSVS	173
JM#19	ILRWSRKMPCFS	262
WSC02	IVRWSKKVPCVS	268
JM#13	ILRWSRKMPCVS	271
JM#20	ILRWSRKMPCMS	272,5
JM#10	IFRWSRKVPCVS	315
JM#18	ILRWSRKMPCLS	385
408-414	LVRYTKK	482
JM#4	IVRWSHKVPCVS	535
408-415	LVRYTKKV	562
JM#146	IYRWSRKMPCLS	584
JM#9	ILRWSHKVPCVS	615
JM#1	ILRWSKKVPCVS	732
408-421	LVRYTKKVPQVSTP	825
JM#105	IIRWSKKVPCVS	966
JM#113	IIRWSRKLPCVS	1099
408-420	LVRYTKKVPQVST	1102
408-417	LVRYTKKVPQ	1103
408-418	LVRYTKKVPQV	1364
408-422	LVRYTKKVPQVSTPT	1407
JM#133	IVRWSKYVS	1584

JM#111	FLRWSRKLPCVS	3353
JM#112	PLRWSRKLPCVS	3371
EPI-X4	**LVRYTKKVPQVSTPTL**	**3709**
JM#94	LIRYTKKVPQVSTPTL	6623
JM#103	FVRWSKKVPCVS	8748
408-413	LVRYTK	9295
JM#123	ILKSSKLPCLS	>10000
JM#125	ILRHSRGPS	>10000
JM#126	IPKWSRGVS	>10000
JM#127	ILKQSRKAPL	>10000
JM#128	ILRTSRFISS	>10000
JM#129	IVRSRKGGTVS	>10000
JM#130	IVRWSPPCVS	>10000
JM#131	IVKSKKAPCVS	>10000
JM#132	IVRKKVPCPS	>10000
JM#134	IVKSHKAPCVS	>10000
JM#135	IVRSSRKVVS	>10000
JM#136	IARSKRGPCAN	>10000
JM#137	IVKNQRKVPV	>10000
JM#138	VVRNSKAAFH	>10000
JM#152	CLKLPGGSCM	>10000
JM#153	CLRLPGGSC	>10000
JM#154	NIRVGGTGMF	>10000
JM#155	QKVVAGVANAL	>10000
JM#156	SRVLNLGPI	>10000
JM#157	MRRAPAFLSA	>10000
JM#158	AGRGKLIAV	>10000
JM#159	NEKRFYLK	>10000
JM#160	SRDKALLRL	>10000
JM#161	GKHVPRAVFV	>10000
JM#162	SKSGRLLLAGY	>10000
JM#92	FVRYTKKVPQVSTPTL	>10000
JM#93	PVRYTKKVPQVSTPTL	>10000

Section SI-7. Description of the models used for the generation of the data related to peptide binders to the PDZ domain of HTRA1 or HTRA3 (Tables 3 and 4).

The template used for the predictions of the binding affinity of peptides with HTRA1 was based on the high-resolution structure corresponding to the PDB code 2JOA.[43] The first model of the ensemble was used for the predictions. In the case of the HTRA3 protein-peptide complexes, the structure with the PDB code 2PW3 was employed as template.[43] Chain A from the protein dimeric structure was used since it lacks only two residues in contrast to chain B which lacks 8 residues. Conformer A of Ser 446 was used in our model. Residues missing in the model were given to the server by using the sequence option.

Table SI-7.1. Ranking of BA of HTRA1-peptide complexes as predicted by state-of-the-art models.

Kdeep		DFIRE		CP_PIE		RF-Score		Prodigy	
Ranking	**BA**	**Ranking**	**BA**	**Ranking**	**BA**	**Ranking**	**BA**	**Ranking**	**BA**
1) WDKIWHV	-13,5	1) ASRIWWV	-16,6	1) DSAIWWV	1,25491	1) DSRAWWV	7,11	1) GWKTWIL	-8,1
2) ASRIWWV	-13,4	2) DIETWLL	-16,5	2) DIETWLL	1,22589	**2) DSRIWWV**	**7,03**	2) DIETWLL	-7,4
3) DIETWLL	-12,6	3) DARIWWV	-16,2	3) DARIWWV	1,16295	3) DARIWWV	7,01	3) WDKIWHV	-7,2
4) DARIWWV	-12,5	3) DSAIWWV	-16,2	4) ASRIWWV	1,11223	4) WDKIWHV	7	4) ASRIWWV	-6,9
4) DSRAWWV	-12,5	**4) DSRIWWV**	**-16,1**	5) DSRIWWA	1,06328	5) ASRIWWV	6,96	5) DARIWWV	-6,7
4) GWKTWIL	-12,5	4) GWKTWIL	-16,1	**6) DSRIWWV**	**1,05627**	6) DSRIWAV	6,95	5) DSAIWWV	-6,7
5) DSRIWWA	-12,0	5) DSRAWWV	-15,7	7) DSRIAWV	1,04117	7) DSAIWWV	6,91	6) DSRAWWV	-6,5
6) DSAIWWV	-11,9	6) WDKIWHV	-15,6	8) DIGPVCFL[a]	1,00710	7) GWKTWIL	6,91	7) DSRIWWA	-6,2
7) DSRIWWV	**-11,7**	7) DIGPVCFL[a]	-15,3	9) DSRAWWV	1,00222	8) DIETWLL	6,9	**7) DSRIWWV**	**-6,2**
8) DSRIAWV	-11,1	8) DSRIWWA	-14,9	10) WDKIWHV	0,91622	8) DSRIWWA	6,9	8) DSRIAWV	-6,1
9) DIGPVCFL[a]	-11,0	9) EVKIMVV[a]	-14,1	11) GWKTWIL	0,88501	9) DSRIAWV	6,79	8) DSRIWAV	-6,1
10 EVKIMVV[a]	-10,7	10) DSRIWAV	-14,0	12) DSRIWAV	0,83207	10) DIGPVCFL[a]	6,64	9) EVKIMVV[a]	-5,8
11) DSRIWAV	-10,1	11) DSRIAWV	-13,4	13) EVKIMVV[a]	0,82242	11) EVKIMVV[a]	6,47	10) DIGPVCFL[a]	-5,6

[a]These protein-peptide structures are at the border of the applicability domain of PPI-Affinity

Table SI-7.2. Ranking of BA of HTRA3-peptide complexes as predicted by state-of-the-art models.

Kdeep		DFIRE		CP_PIE		RF-Score		Prodigy	
Ranking	**BA**	**Ranking**	**BA**	**Ranking**	**BA**	**Ranking**	**BA**	**Ranking**	**BA**
1) FGRWF[b]	-10,0	1) FGRWA[a]	-6,6	1) FGRWA[a]	0,357316	1) FARWV[b]	4,24	1) RSWWV	-4,9
2) FGRWV[a]	**-8,9**	2) FARWV[b]	-6,4	**2) FGRWV[a]**	**0,356035**	**2) FGRWV[a]**	**4,08**	2) RWV[a]	-3,7
3) FGRWI[b]	-8,8	2) FGRWL[a]	-6,4	3) FGRWF[b]	0,35465	3) FGRWI[b]	3,97	3) WG[b]	-3,6
4) RSWWV	-8,6	3) FGAWV[b]	-6,2	4) FGRWL[a]	0,352168	4) FGRWA[a]	3,9	4) FARWV[b]	-3,5
5) FARWV[b]	-8,5	3) FGRWI[b]	-6,2	5) FGAWV[b]	0,348097	5) FGRWL[a]	3,88	4) WA[b]	-3,5
6) FGRWL[a]	-8,2	4) FGRWF[b]	-6,1	6) FGRWI[b]	0,345033	6) FGRWF[b]	3,84	5) FGRWA[a]	-3,3
7) FGRWA[a]	-7,9	**5) FGRWV[a]**	**-6,0**	7) FARWV[b]	0,333353	7) FGAWV[b]	3,81	5) FGRWF[b]	-3,3
8) FGAWV[b]	-7,8	6) RSWWV	-5,3	8) FGRAV[a]	0,248552	7) RSWWV	3,81	**5) FGRWV[a]**	**-3,3**
9) FGRAV[a]	-7,0	7) FGRAV[a]	-4,1	9) WG[b]	0,163576	8) FGRAV[a]	3,66	5) GRWV[a]	-3,3

10) GRWV[a]	-6,8	8) RWV[a]	-3,9	10) RSWWV	0,159256	9) WG[b]	3,47	5) WV[b]	-3,3
11) RWV[a]	-5,7	9) WV[b]	-3,7	11) WA[b]	0,150201	10) WV[b]	3,45	6) FGAWV[b]	-3,2
12) WA[b]	-5,3	10) WA[b]	-3,5	12) RWV[a]	0,14853	11) GRWV[a]	3,44	6) FGRWI[b]	-3,2
13) WV[b]	-5,1	11) WG[b]	-3,3	13) GRWV[a]	0,101262	12) RWV[a]	3,42	6) FGRWL[a]	-3,2
14) WG[b]	-4,2	12) GRWV[a]	-2,9	14) WV[b]	0,0838	13) WA[b]	3,4	7) FGRAV[a]	-3

[a]These protein-peptide structures are at the border of the applicability domain of PPI-Affinity. [b]These protein-peptide structures are outside the applicability domain of PPI-Affinity

Table SI-7. Summary of values that define the applicability domain of the protein-protein model.

MODEL 2

DESCRIPTOR	MINIMUM	MAXIMUM	1st PERCENTILE	99TH PERCENTILE
WNC(ECI)_NO_AHR_G	0	1	0,13	1
WNC(ECI)_NO_ALR_G	0	1	0,05	1
WNC(Z2)_NO_PLR_V	0	46,59	0	27,418
WNC(Z2)_NO_PCR_G	-3,128	2,512	-2,237	1,857
WNC(Z3)_NO_GLU_V	0	6,545	0	5,892
WNC(Z3)_NO_PLR_P2	0	5,317	0	5,195
WNC(Z3)_NO_PRT_P2	0	4,133	0	3,833
WFLC(ECI)_NO_ILE_P2	0	0,014	0	0,006
WFLC(IP)_NO_PCR_AR	0	0,035	0	0,027
WFLC(IP)_NO_PCR_V	0	0,002	0	0,001
WFLC(ISA)_NO_PLR_AR	0	0,011	0	0,008
WNLC(ECI)_NO_AHR_V	0	2,31	0,005	2,15
WNLC(ECI)_NO_NPR_N1	0	81,424	0,407	67,557
WNLC(ECI)_NO_NPR_DE	0	0,599	0,016	0,568
WNLC(IP)_NO_BSR_N1	0	28896,816	304,609	25512,865
WNLC(IP)_NO_PLR_N1	0	56383,04	695,219	51326,103
WNLC(ISA)_NO_BSR_N2	0	607045,942	46153,136	568135,465
WNLC(ISA)_NO_PLR_G	1	10458,877	1,671	9535,444
WNLC(Z1)_NO_AHR_DE	0	12,626	1,257	12,044
WNLC(Z1)_NO_ALR_N2	0	269,091	17,584	212,189
WNLC(Z1)_NO_NCR_P2	0	15,07	1,402	12,439
WNLC(Z1)_NO_PRT_AR	-2,423	3,379	-2,075	2,05
WNLC(Z2)_NO_BSR_G	-4,237	4,778	-4,157	4,315
WNLC(Z2)_NO_ALR_N1	-112,241	2145,601	-51,79	1292,54
WNLC(Z3)_NO_AHR_AR	-0,777	3,884	-0,666	2,621
WNLC(Z3)_NO_ALR_AR	-0,581	1,985	-0,318	1,324

MODEL 3

DESCRIPTOR	MINIMUM	MAXIMUM	1st PERCENTILE	99TH PERCENTILE
WNC(ECI)_NO_AHR_G	0	1	0,175	1
WNC(ECI)_NO_ALR_G	0	1	0,05	1
WNC(Z2)_NO_PLR_V	0	46,59	0	27,418
WNC(Z2)_NO_PCR_G	-2,014	2,512	-1,888	1,615
WNC(Z3)_NO_GLU_V	0	6,457	0	5,216
WNC(Z3)_NO_PLR_P2	0	5,413	0	5,28
WNC(Z3)_NO_PRT_P2	0	4,133	0	3,904
WFLC(ECI)_NO_ILE_P2	0	0,014	0	0,007
WFLC(IP)_NO_PCR_AR	0	0,035	0,001	0,02
WFLC(IP)_NO_PCR_V	0	0,002	0	0,001
WFLC(ISA)_NO_PLR_AR	0	0,009	0	0,008
WNLC(ECI)_NO_AHR_V	0	2,31	0,142	2,177
WNLC(ECI)_NO_NPR_N1	0	75,842	3,807	66,742
WNLC(ECI)_NO_NPR_DE	0	0,599	0,064	0,582
WNLC(IP)_NO_BSR_N1	0	25889,035	458,054	22440,66
WNLC(IP)_NO_PLR_N1	0	52027,76	790,424	47683,177
WNLC(ISA)_NO_BSR_N2	0	569128,79	77888,907	497878,815
WNLC(ISA)_NO_PLR_G	1	10502,058	4,237	9636,287
WNLC(Z1)_NO_AHR_DE	0	12,628	3,082	12,373
WNLC(Z1)_NO_ALR_N2	0	269,091	27,865	186,411
WNLC(Z1)_NO_NCR_P2	0	12,857	3,593	11,926
WNLC(Z1)_NO_PRT_AR	-4,954	2,927	-2,312	2,048
WNLC(Z2)_NO_BSR_G	-4,237	4,778	-4,157	4,315
WNLC(Z2)_NO_ALR_N1	-112,241	2145,601	-51,79	1292,54
WNLC(Z3)_NO_AHR_AR	-1,035	2,876	-0,706	2,59
WNLC(Z3)_NO_ALR_AR	-0,581	1,65	-0,385	1,288

Table SI-8. Summary of values that define the applicability domain of the protein-peptide model.

MODEL 1

DESCRIPTOR	MINIMUM	MAXIMUM	1st PERCENTILE	99TH PERCENTILE
WNC(ECI)_NO_AHR_N1	0	44,171	0,086	28,189
WNC(ECI)_NO_ALR_N1	0	6,367	0,037	4,032
WNC(IP)_NO_PLR_N1	0	5391,296	96,607	3672,786
WNC(ISA)_NO_PLR_V	0	318853829,8	543962,387	250318617,3
WNC(Z1)_NO_NPR_N2	0	132,965	4,695	111,105
WNC(Z1)_NO_ARM_V	0	692,98	0	323,717
WNC(Z1)_NO_PRT_N1	-241,412	418,803	-169,452	331,683
WNC(Z3)_NO_AHR_N1	-57,396	149,198	-37,179	58,761
WNC(Z3)_NO_UCR_N1	-34,347	33,82	-32,638	25,764
WFLC(ECI)_NO_AHR_SI30	0,007	1	0,011	0,764
WFLC(ECI)_NO_PCR_V	0	0,026	0	0,015
WFLC(IP)_NO_RTR_MI30	0,062	4,74	0,074	4,523
WFLC(IP)_NO_BSR_AR	0,001	0,037	0,001	0,024
WFLC(IP)_NO_UCR_RA	0,001	0,5	0,001	0,314
WFLC(IP)_NO_UCR_K	46,226	12840,122	70,881	11654,217
WFLC(IP)_NO_PRT_TI30	14,202	573,443	18,237	454,996
WFLC(ISA)_NO_NPR_P2	0,006	0,094	0,007	0,082
WFLC(ISA)_NO_NPR_DE	0,005	0,089	0,007	0,071
WFLC(ISA)_NO_PRT_TI30	14,202	582,373	18,237	461,122
WFLC(Z1)_NO_BSR_SI30	0,009	1	0,02	0,914
WFLC(Z2)_NO_PLR_K	113,123	49106,737	364,597	40834,695
WNLC(ECI)_NO_BSR_I50	0,124	1,531	0,13	1,323
WNLC(ECI)_NO_PLR_K	110,951	25195,103	240,618	18518,305
WNLC(ECI)_NO_UCR_S	-0,109	2,746	0,071	2,166
WNLC(IP)_NO_RTR_V	339,368	1437,317	375,712	1305,349
WNLC(IP)_NO_UCR_I50	11,749	62,265	18,864	56,295
WNLC(ISA)_NO_UCR_S	-0,009	4,109	0,53	3,201
WNLC(Z1)_NO_RTR_SI30	0,547	1	0,562	0,95
WNLC(Z1)_NO_BSR_K	80,492	12661,731	113,741	10133,441
WNLC(Z1)_NO_UCR_K	38,097	8637,703	57,65	7834,548
WNLC(Z2)_NO_AHR_P2	1,46	4,799	1,647	4,391
WNLC(Z2)_NO_PCR_N2	4,845	56,314	6,671	54,94
WNLC(Z2)_NO_NCR_P2	0,714	4,332	0,905	3,913
WNLC(Z2)_NO_UCR_K	56,544	7155,916	69,33	6747,321
WNLC(Z3)_NO_PLR_SI30	0,419	0,827	0,423	0,775
WNLC(Z3)_NO_NCR_RA	4,361	27,671	5,534	22,597
WNLC(Z3)_NO_UCR_TI30	20,529	504,745	29,219	462,806

MODEL 3

DESCRIPTOR	MINIMUM	MAXIMUM	1st PERCENTILE	99TH PERCENTILE
WNC(ECI)_NO_AHR_N1	0	44,171	0,056	27,473
WNC(ECI)_NO_ALR_N1	0	6,367	0,023	4,239
WNC(IP)_NO_PLR_N1	0	5554,6	66,773	4292,97
WNC(ISA)_NO_PLR_V	0	290472538,7	543962,387	236857692,4
WNC(Z1)_NO_NPR_N2	0	132,965	4,695	115,217
WNC(Z1)_NO_ARM_V	0	692,98	0	323,717
WNC(Z1)_NO_PRT_N1	-241,412	418,803	-167,234	331,683
WNC(Z3)_NO_AHR_N1	-63,649	149,198	-40,292	58,582
WNC(Z3)_NO_UCR_N1	-35,227	33,82	-32,638	25,745
WFLC(ECI)_NO_AHR_SI30	0,01	1	0,013	0,764
WFLC(ECI)_NO_PCR_V	0	0,026	0	0,014
WFLC(IP)_NO_RTR_MI30	0,071	4,657	0,079	4,501
WFLC(IP)_NO_BSR_AR	0,001	0,035	0,001	0,024
WFLC(IP)_NO_UCR_RA	0,001	0,5	0,001	0,322
WFLC(IP)_NO_UCR_K	23,224	20147,154	79,781	11654,217
WFLC(IP)_NO_PRT_TI30	13,456	566,217	17,456	453,162
WFLC(ISA)_NO_NPR_P2	0,006	0,094	0,007	0,081
WFLC(ISA)_NO_NPR_DE	0,004	0,089	0,005	0,071
WFLC(ISA)_NO_PRT_TI30	13,456	577,568	17,456	456,221
WFLC(Z1)_NO_BSR_SI30	0,009	0,969	0,02	0,906
WFLC(Z2)_NO_PLR_K	113,123	47098,384	407,834	35890,325
WNLC(ECI)_NO_BSR_I50	0,118	1,396	0,129	1,253
WNLC(ECI)_NO_PLR_K	110,951	22599,466	244,172	18485,829
WNLC(ECI)_NO_UCR_S	-0,109	2,746	0,106	2,346
WNLC(IP)_NO_RTR_V	305,401	1463,52	375,712	1200,763
WNLC(IP)_NO_UCR_I50	13,728	62,265	19,17	57,883
WNLC(ISA)_NO_UCR_S	0,398	4,109	0,64	2,989
WNLC(Z1)_NO_RTR_SI30	0,552	1	0,569	0,932
WNLC(Z1)_NO_BSR_K	83,132	12260,904	113,637	10124,995
WNLC(Z1)_NO_UCR_K	38,097	10241,145	63,865	7305,526
WNLC(Z2)_NO_AHR_P2	1,575	4,856	1,723	4,396
WNLC(Z2)_NO_PCR_N2	4,845	56,314	6,671	54,92
WNLC(Z2)_NO_NCR_P2	0,714	4,332	0,905	3,968
WNLC(Z2)_NO_UCR_K	58,983	7155,916	72,516	6550,209
WNLC(Z3)_NO_PLR_SI30	0,419	0,833	0,423	0,76
WNLC(Z3)_NO_NCR_RA	4,361	27,671	5,36	22,538
WNLC(Z3)_NO_UCR_TI30	16	504,745	31,299	462,593

Table SI-9. Summary of the minimum and maximum values of the sequences' length of the peptides and proteins in each dataset.

		Protein size (aa)			
		Peptide		Receptor	
	Data size	Min	Max	Min	Max
Training	922	3	29	31	957
Development	100	4	29	51	559
Test	100	4	29	51	496
EPI-X4	57	6	16	319	319
HTRA1	13	7	8	105	105
HTRA3	14	2	5	105	105

The described subsets are the training, development and test sets with data points taken from the Biolip[37] database, as well as those used to test the protein-peptide model on experimentally measured BA data: EPI-X4, HTRA1 and HTRA3.

Table SI-10. Descriptive statistics of the different data sets used in the modeling and test of the protein-protein BA predictor.

		Binding Affinity (ΔG)			
	Data size	Min	Max	Mean	StdDev
Training set	648	-18.1	-3.1	-9.7	2.5
Development set	90	-18.1	-4.3	-9.5	2.6
Test set 1[1]	79	-18.6	-4.3	-10.1	2.8
Test set 2	90	-18.1	-3.3	-9.3	2.6
Test set 3[28]	177	-16.3	-5.5	-10.4	2.6

The reported statistics are the minimum (Min), maximum (Max), mean and standard deviation (StdDev) of the binding free energy (ΔG) values in each dataset. The described subsets are: the training, development, and test (Test set 2) sets with data taken from the PDBbind (v.2020)[8] dataset, Test set 1 corresponding to the benchmark employed by Vangone and Bonvin,[1] and Test set 3 corresponding to the set of 26 wild-types and 151 mutants taken from the SKEMPI[28] dataset. All the training protein-protein complexes contained two protein sequences with individual sequence length ranging from 20 to 958 amino acids.

Table SI-11. Descriptive statistics of the training, development and test sets used in the modeling of the protein-peptide BA predictor.

		Binding Affinity (ΔG)			
	Data size	Min	Max	Mean	StdDev
Training set	922	-14.4	-3.6	-8.2	2.1
Development set	100	-13.6	-4.8	-8.5	2.0
Test set	100	-12.6	-4.6	-8.1	1.7

The reported statistics are the minimum (Min), maximum (Max), mean and standard deviation (StdDev) of the binding free energy (ΔG) values in each dataset.

References

1. Vangone, A.; Bonvin, A. M., Contacts-based prediction of binding affinity in protein-protein complexes. *Elife* **2015**, 4, e07454-e07454.
2. Kastritis, P. L.; Moal, I. H.; Hwang, H.; Weng, Z.; Bates, P. A.; Bonvin, A. M. J. J.; Janin, J., A structure-based benchmark for protein–protein binding affinity. *Protein Science* **2011**, 20, 482-491.
3. Liu, S.; Zhang, C.; Zhou, H.; Zhou, Y., A physical reference state unifies the structure-derived potential of mean force for protein folding and binding. *Proteins: Structure, Function, and Bioinformatics* **2004**, 56, 93-101.
4. Ravikant, D. V. S.; Elber, R., PIE-efficient filters and coarse grained potentials for unbound protein-protein docking. *Proteins* **2010**, 78, 400-419.
5. Berman, H. M.; Westbrook, J.; Feng, Z.; Gilliland, G.; Bhat, T. N.; Weissig, H.; Shindyalov, I. N.; Bourne, P. E., The Protein Data Bank. *Nucleic Acids Research* **2000**, 28, 235-242.
6. RCSB PDB. https://www.rcsb.org/
7. Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G., KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *Journal of Chemical Information and Modeling* **2018**, 58, 287-296.
8. Wang, R.; Fang, X.; Lu, Y.; Wang, S., The PDBbind Database: Collection of Binding Affinities for Protein−Ligand Complexes with Known Three-Dimensional Structures. *Journal of Medicinal Chemistry* **2004**, 47, 2977-2980.
9. Ballester, P. J.; Mitchell, J. B. O., A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics* **2010**, 26, 1169-1175.
10. Rodriguez-Soca, Y.; Munteanu, C. R.; Dorado, J.; Pazos, A.; Prado-Prado, F. J.; González-Díaz, H., Trypano-PPI: A Web Server for Prediction of Unique Targets in Trypanosome Proteome by using Electrostatic Parameters of Protein−protein Interactions. *Journal of Proteome Research* **2010**, 9, 1182-1190.
11. Dobson, P. D.; Doig, A. J., Distinguishing Enzyme Structures from Non-enzymes Without Alignments. *Journal of Molecular Biology* **2003**, 330, 771-783.
12. Rodriguez-Soca, Y.; Munteanu, C. R.; Dorado, J.; Rabuñal, J.; Pazos, A.; González-Díaz, H., Plasmod-PPI: A web-server predicting complex biopolymer targets in plasmodium with entropy measures of protein–protein interactions. *Polymer* **2010**, 51, 264-273.
13. Chou, K.-C.; Cai, Y.-D., Predicting Protein−Protein Interactions from Sequences in a Hybridization Space. *Journal of Proteome Research* **2006**, 5, 316-322.
14. von Mering, C.; Jensen, L. J.; Snel, B.; Hooper, S. D.; Krupp, M.; Foglierini, M.; Jouffre, N.; Huynen, M. A.; Bork, P., STRING: known and predicted protein–protein associations, integrated and transferred across organisms. *Nucleic Acids Research* **2005**, 33, D433-D437.
15. Narykov, O.; Johnson, N. T.; Korkin, D., Predicting protein interaction network perturbation by alternative splicing with semi-supervised learning. *Cell Reports* **2021**, 37.
16. Yang, X.; Coulombe-Huntington, J.; Kang, S.; Sheynkman, Gloria M.; Hao, T.; Richardson, A.; Sun, S.; Yang, F.; Shen, Yun A.; Murray, Ryan R.; Spirohn, K.; Begg, Bridget E.; Duran-Frigola, M.; MacWilliams, A.; Pevzner, Samuel J.; Zhong, Q.; Trigg, Shelly A.; Tam, S.; Ghamsari, L.; Sahni, N.; Yi, S.; Rodriguez, Maria D.; Balcha, D.; Tan, G.; Costanzo, M.; Andrews, B.; Boone, C.; Zhou, Xianghong J.; Salehi-Ashtiani, K.; Charloteaux, B.; Chen, Alyce A.; Calderwood, Michael A.; Aloy, P.; Roth, Frederick P.; Hill, David E.; Iakoucheva, Lilia M.; Xia, Y.; Vidal, M., Widespread Expansion of Protein Interaction Capabilities by Alternative Splicing. *Cell* **2016**, 164, 805-817.
17. Rolland, T.; Taşan, M.; Charloteaux, B.; Pevzner, Samuel J.; Zhong, Q.; Sahni, N.; Yi, S.; Lemmens, I.; Fontanillo, C.; Mosca, R.; Kamburov, A.; Ghiassian, Susan D.; Yang, X.; Ghamsari, L.; Balcha, D.; Begg, Bridget E.; Braun, P.; Brehme, M.; Broly, Martin P.; Carvunis, A.-R.; Convery-Zupan, D.; Corominas, R.; Coulombe-Huntington, J.; Dann, E.; Dreze, M.; Dricot, A.; Fan, C.; Franzosa, E.; Gebreab, F.; Gutierrez,

Bryan J.; Hardy, Madeleine F.; Jin, M.; Kang, S.; Kiros, R.; Lin, Guan N.; Luck, K.; MacWilliams, A.; Menche, J.; Murray, Ryan R.; Palagi, A.; Poulin, Matthew M.; Rambout, X.; Rasla, J.; Reichert, P.; Romero, V.; Ruyssinck, E.; Sahalie, Julie M.; Scholz, A.; Shah, Akash A.; Sharma, A.; Shen, Y.; Spirohn, K.; Tam, S.; Tejeda, Alexander O.; Trigg, Shelly A.; Twizere, J.-C.; Vega, K.; Walsh, J.; Cusick, Michael E.; Xia, Y.; Barabási, A.-L.; Iakoucheva, Lilia M.; Aloy, P.; De Las Rivas, J.; Tavernier, J.; Calderwood, Michael A.; Hill, David E.; Hao, T.; Roth, Frederick P.; Vidal, M., A Proteome-Scale Map of the Human Interactome Network. *Cell* **2014**, 159, 1212-1226.

18. Rual, J.-F.; Venkatesan, K.; Hao, T.; Hirozane-Kishikawa, T.; Dricot, A.; Li, N.; Berriz, G. F.; Gibbons, F. D.; Dreze, M.; Ayivi-Guedehoussou, N.; Klitgord, N.; Simon, C.; Boxem, M.; Milstein, S.; Rosenberg, J.; Goldberg, D. S.; Zhang, L. V.; Wong, S. L.; Franklin, G.; Li, S.; Albala, J. S.; Lim, J.; Fraughton, C.; Llamosas, E.; Cevik, S.; Bex, C.; Lamesch, P.; Sikorski, R. S.; Vandenhaute, J.; Zoghbi, H. Y.; Smolyar, A.; Bosak, S.; Sequerra, R.; Doucette-Stamm, L.; Cusick, M. E.; Hill, D. E.; Roth, F. P.; Vidal, M., Towards a proteome-scale map of the human protein–protein interaction network. *Nature* **2005**, 437, 1173-1178.

19. Venkatesan, K.; Rual, J.-F.; Vazquez, A.; Stelzl, U.; Lemmens, I.; Hirozane-Kishikawa, T.; Hao, T.; Zenkner, M.; Xin, X.; Goh, K.-I.; Yildirim, M. A.; Simonis, N.; Heinzmann, K.; Gebreab, F.; Sahalie, J. M.; Cevik, S.; Simon, C.; de Smet, A.-S.; Dann, E.; Smolyar, A.; Vinayagam, A.; Yu, H.; Szeto, D.; Borick, H.; Dricot, A.; Klitgord, N.; Murray, R. R.; Lin, C.; Lalowski, M.; Timm, J.; Rau, K.; Boone, C.; Braun, P.; Cusick, M. E.; Roth, F. P.; Hill, D. E.; Tavernier, J.; Wanker, E. E.; Barabási, A.-L.; Vidal, M., An empirical framework for binary interactome mapping. *Nature Methods* **2009**, 6, 83-90.

20. Yu, H.; Tardivo, L.; Tam, S.; Weiner, E.; Gebreab, F.; Fan, C.; Svrzikapa, N.; Hirozane-Kishikawa, T.; Rietman, E.; Yang, X.; Sahalie, J.; Salehi-Ashtiani, K.; Hao, T.; Cusick, M. E.; Hill, D. E.; Roth, F. P.; Braun, P.; Vidal, M., Next-generation sequencing to generate interactome datasets. *Nature Methods* **2011**, 8, 478-480.

21. Jia, J.; Liu, Z.; Xiao, X.; Liu, B.; Chou, K.-C., iPPBS-Opt: A Sequence-Based Ensemble Classifier for Identifying Protein-Protein Binding Sites by Optimizing Imbalanced Training Datasets. *Molecules (Basel, Switzerland)* **2016**, 21, E95-E95.

22. Deng, L.; Guan, J.; Dong, Q.; Zhou, S., Prediction of protein-protein interaction sites using an ensemble method. *BMC Bioinformatics* **2009**, 10, 426.

23. Yugandhar, K.; Gromiha, M. M., Protein–protein binding affinity prediction from amino acid sequence. *Bioinformatics* **2014**, 30, 3583-3589.

24. Abbasi, W. A.; Yaseen, A.; Hassan, F. U.; Andleeb, S.; Minhas, F. U. A. A., ISLAND: in-silico proteins binding affinity prediction using sequence information. *BioData Mining* **2020**, 13, 20.

25. Cao, D.-S.; Xu, Q.-S.; Liang, Y.-Z., propy: a tool to generate various modes of Chou's PseAAC. *Bioinformatics* **2013**, 29, 960-962.

26. Huang, X.; Zheng, W.; Pearce, R.; Zhang, Y., SSIPe: accurately estimating protein–protein binding affinity change upon mutations using evolutionary profiles in combination with an optimized physical energy function. *Bioinformatics* **2020**, 36, 2429-2437.

27. Xiong, P.; Zhang, C.; Zheng, W.; Zhang, Y., BindProfX: Assessing Mutation-Induced Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts. *Journal of Molecular Biology* **2017**, 429, 426-434.

28. Jankauskaitė, J.; Jiménez-García, B.; Dapkūnas, J.; Fernández-Recio, J.; Moal, I. H., SKEMPI 2.0: an updated benchmark of changes in protein–protein binding energy, kinetics and thermodynamics upon mutation. *Bioinformatics* **2019**, 35, 462-469.

29. Janin, J.; Henrick, K.; Moult, J.; Eyck, L. T.; Sternberg, M. J. E.; Vajda, S.; Vakser, I.; Wodak, S. J., CAPRI: A Critical Assessment of PRedicted Interactions. *Proteins: Structure, Function, and Bioinformatics* **2003**, 52, 2-9.

30. Wang, B.; Su, Z.; Wu, Y., Computational Assessment of Protein–Protein Binding Affinity by Reverse Engineering the Energetics in Protein Complexes. *Genomics, Proteomics & Bioinformatics* **2021**.

31. Moal, I. H.; Fernández-Recio, J., SKEMPI: a Structural Kinetic and Energetic database of Mutant Protein Interactions and its use in empirical models. *Bioinformatics* **2012**, 28, 2600-2607.
32. Abbasi, W. A.; Asif, A.; Ben-Hur, A.; Minhas, F. u. A. A., Learning protein binding affinity using privileged information. *BMC Bioinformatics* **2018**, 19, 425.
33. Chen, J.; Sawyer, N.; Regan, L., Protein–protein interactions: General trends in the relationship between binding affinity and interfacial buried surface area. *Protein Science* **2013**, 22, 510-515.
34. Eddy, S. R., Where did the BLOSUM62 alignment score matrix come from? *Nature Biotechnology* **2004**, 22, 1035-1036.
35. Moal, I. H.; Agius, R.; Bates, P. A., Protein–protein binding affinity prediction on a diverse set of structures. *Bioinformatics* **2011**, 27, 3002-3009.
36. Dias, R.; Kolaczkowski, B., Improving the accuracy of high-throughput protein-protein affinity prediction may require better training data. *BMC Bioinformatics* **2017**, 18, 102.
37. Yang, J.; Roy, A.; Zhang, Y., BioLiP: a semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic acids research* **2013**, 41, D1096-103.
38. Xue, L. C.; Rodrigues, J. P.; Kastritis, P. L.; Bonvin, A. M.; Vangone, A., PRODIGY: a web server for predicting the binding affinity of protein–protein complexes. *Bioinformatics* **2016**, 32, 3676-3678.
39. Ballester, P. J.; Mitchell, J. B. O., A machine learning approach to predicting protein-ligand binding affinity with applications to molecular docking. *Bioinformatics (Oxford, England)* **2010**, 26, 1169-1175.
40. Ruiz-Blanco, Y. B.; Paz, W.; Green, J.; Marrero-Ponce, Y., ProtDCal: A program to compute general-purpose-numerical descriptors for sequences and 3D-structures of proteins. *BMC Bioinformatics* **2015**, 16, 162.
41. Romero-Molina, S.; Ruiz-Blanco, Y. B.; Green, J. R.; Sanchez-Garcia, E., ProtDCal-Suite: A web server for the numerical codification and functional analysis of proteins. *Protein Science* **2019**, 28, 1734-1743.
42. Harms, M.; Gilg, A.; Ständker, L.; Beer, A. J.; Mayer, B.; Rasche, V.; Gruber, C. W.; Münch, J., Microtiter plate-based antibody-competition assay to determine binding affinities and plasma/blood stability of CXCR4 ligands. *Scientific Reports* **2020**, 10, 16036.
43. Runyon, S. T.; Zhang, Y.; Appleton, B. A.; Sazinsky, S. L.; Wu, P.; Pan, B.; Wiesmann, C.; Skelton, N. J.; Sidhu, S. S., Structural and functional analysis of the PDZ domains of human HtrA1 and HtrA3. *Protein Science* **2007**, 16, 2454-2471.