Classifying Protein–Protein Binding Affinity with Free-Energy Calculations and Machine Learning Approaches

Emma Goulard Coderc de Lacam, Benoît Roux, Christophe Chipot

Classifying Protein-Protein Binding Affinity with Free-Energy Calculations and Machine-Learning Approaches

Emma Goulard Coderc de Lacam [†], Benoît Roux [‡,¶] and Christophe Chipot [*,†,§,‡,‖]

†*Laboratoire International Associé Centre National de la Recherche Scientifique et University of Illinois at Urbana-Champaign, Unité Mixte de Recherche n°7019, Université de Lorraine, B.P. 70239, 54506 Vandœuvre-lès-Nancy cedex, France*

‡*Department of Biochemistry and Molecular Biology, The University of Chicago, 929 E. 57th Street W225, Chicago, Illinois 60637, USA*

¶*Department of Chemistry, The University of Chicago, 5735 S Ellis Ave, Chicago, IL 60637, Chicago, IL 60637, USA*

§*Theoretical and Computational Biophysics Group, Beckman Institute, and Department of Physics, University of Illinois at Urbana-Champaign, Urbana, USA*

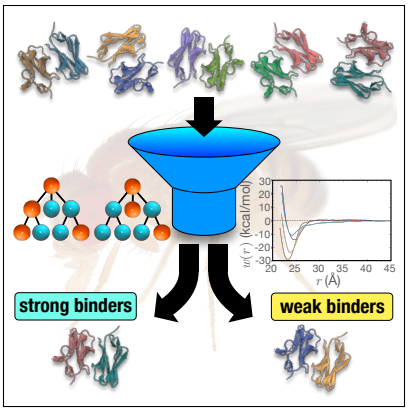
‖*Department of Chemistry, The University of Hawai'i at Mānoa, 2545 McCarthy Mall, Honolulu, Hawaii 96822, USA*

E-mail: chipot@illinois.edu

Abstract

Understanding the intricate phenomenon of neuronal wiring in the brain is of topical interest in neuroscience. In the fruit fly, *Drosophila Melanogaster*, the Dpr-DIP interactome has been identified to play an important role in this process. However, experimental data suggest that merely a limited subset of complexes, essentially 57 out of a total of 231, exhibit strong binding affinity. In this work, we sought to identify the residue-level molecular basis underlying the difference in binding affinity using a state-of-the-art methodology consisting of standard binding free-energy calculations with the geometrical route and machine learning (ML) techniques. We determined the binding affinity for two complexes using statistical mechanics simulations, achieving an excellent reproduction of the experimental data. Moreover, we predicted the binding free energy for two additional low-affinity complexes, devoid of experimental estimation while simultaneously identifying key residues for the binding. Furthermore, through the use of ML algorithms, linear discriminant analysis and random forest, we achieved remarkable accuracy, as high as 0.99, in discerning between strong (cognate) and weak (non-cognate) binders. The presented ML approach encompasses easily transferable input features, enabling its broad application to any interactome, while facilitating the identification of pivotal residues critical for binding interactions. The predictive power of the generated model was probed on similar protein families from thirteen diverse species. Our ML model exhibited commendable performance on these additional datasets, showcasing its reliability and robustness across the species barrier.

TOC Graphic


strong binders
weak binders
30
20
10
0
-10
-20
-30
(kcal/mol)
20
25
30
35
40
45
r (Å)

Introduction

The brain comprises a vast array of neurons linked by a complex synaptic connection network. Understanding how neurites distinguish synaptic partners in densely packed neuronal tissue remains a central question in neurobiology.[1,2] Roger Sperry's chemoaffinity hypothesis proposed in the 1960s stated that neurons make specific connections based on the affinity of cell-surface labels.[3] This concept was further refined into the idea that cell adhesion-like cell-surface proteins expressed on neuron surfaces bind to each other and trigger downstream events leading to synapse formation between appropriate partners.[4] The realm of cell recognition encompasses cell adhesion molecules known as cell-surface proteins (CSPs), which frequently arise from gene duplication events occurring throughout evolutionary processes. Such duplications give rise to multiple members within the CSP family, with each protein exhibiting a characteristic canonical binding interface indicative of the particular family.[5] In this case, the binding strength is solely responsible for the different functions of each protein.[5–7] In the fruit fly, the Defective proboscis extension response (Dpr)-Dpr Interacting Proteins(DIP) interactome has been identified by Ozkan et al.[8] to play that role. Out of the 231 heterodimers formed by the 11 DIP and 21 Dprs, respectively, only 57 displayed high affinity by surface plasmon resonance (SPR)[9] assay[5], assuming that the strong binding correspond to cognate complexes leaving the remaining 174 complexes as non-cognate (weak binders). Due to the high sequence similarity of these complexes, special efforts are required to understand the mechanism underlying the specific association of the binding partners.

The availability of comprehensive structural information regarding our protein–protein complexes has opened avenues for the use of empirical structure-based scoring methods that leverage force fields and thermodynamics properties.[10–12] In the MM/PBSA approximation, the binding affinity is computed as the sum of its gas-phase energy (MM), the solvation free energy using Poisson-Boltzmann continuum electrostatics and the solvent-exposed surface area (PBSA), and a contribution due to the configurational entropy of the solute extracted from molecular dynamics (MD) simulations.[11,13–16] The MM/PBSA was tested for the Dpr-DIP interaction, resulting in a

maximum distinguishability of 30 % between cognate and non-cognate complexes.[17] However, the numerous underlying approximations (implicit solvation and constant dielectric medium)[18–20], as well as the poor resulting discrimination obtained, call into question its general applicability for protein–protein complexes. Since implicit solvation approaches cannot provide reliable estimates for the Dpr-DIP interactome, the idea is to use an explicit solvent model with enhanced sampling to favor rare events such as binding or unbinding. Steered MD involves one protein being pulled away from the other.[21] This method allows the recovery of the binding free energy through the Jarzynski equation at the hefty cost of multiple realizations in a near-equilibrium regime,[22,23] under the questionable assumption that binding depends solely on the physical separation.[24] The utilization of this strategy necessitates a significant allocation of computational resources, rendering it incompatible with the processing of large datasets. The geometrical route,[20] conceptualized by Woo and Roux[25] and generalized for protein–protein complexes by Gumbart et al.,[20] offers a framework with potential-of-mean force (PMF) calculations, in which the slow degrees of freedom of the reversible association are gradually restrained to enhance convergence and reduce the computational cost while preserving estimates reaching chemical accuracy.

Alternative approaches using machine learning (ML) to predict, or classify binding partners at a lower computational cost have emerged in the last decade, and are mostly directed at protein–ligand complexes due to the interest of the pharmaceutical industry in the drug design field.[26–31] Only a few applications have been dedicated to predicting the binding of protein–protein complexes.[17,32–34] The methodologies employed to address this challenge can be broadly categorized into two distinct categories, leveraging either sequence or structural characteristics.[35] In the case of sequence-based approaches, features rely on evolutionarily conserved residues, hypothesized to be essential for the function, and, therefore, for binding.[36] The various types of input features include (i) residue encoding, (ii) evolutionary information, (iii) residue physicochemical properties, and (iv) predicted structural features.[35] Residue encoding, using traditional techniques such as the one-hot representation (vector of length twenty containing zeros representing the twenty amino acids. The position in that vector representing the encoded residue will have a value of one to differentiate

from the surrounding zeros), is insufficient to predict the protein interactions efficiently.[35] Consequently, evolutionary descriptors, like mutual information,[36] are preferred. Inasmuch as structural features are concerned, they usually consist of geometrical descriptors extracted from the structure, such as the surface accessible to the solvent, the curvature of a set of atoms, or a set of invariant geometrical fingerprints.[35,37] An inherent limitation of structure-based approaches is the quality and availability of reliable structures linked to the difficulties to solve structures either experimentally or theoretically.

In this work, we aimed at predicting amid a series of homologous protein–protein complexes which ones are cognate, and at identifying the molecular basis of their formation at the residue level. Towards this end, we first used the geometrical route[20] as a precise method based on first principles, and applied to a few complexes. Then, we turned to ML schemes using sequence-based features to treat the whole Dpr-DIP dataset and identify the key residues responsible for differentiating cognate and non-cognate complexes.

Methods

The following subsections briefly recap the methodology employed in this work, its theoretical underpinnings, and describe the protocols of the calculations reported herein.

Binding Free-Energy Calculations

The formation of a protein–protein complex involves significant conformational changes, hardening the ergodic sampling of configurational space within the simulation time amenable to MD. In the realm of computational techniques for protein binding, importance sampling algorithms[38] emerge as powerful tools. These algorithms operate by introducing external forces onto collective variables (CVs) to accelerate the sampling of rare events, such as protein binding. These CVs

are essential degrees of freedom involved in the reversible association, and can be controlled and monitored during the course of the simulation.[39] One of the straightforward CVs to control the reversible binding of two proteins is the Euclidean distance between their centers of mass (COMs). However, it does not prevent random tumbling of the two molecular objects at play, nor conformational changes upon association, slowing down the convergence of the separation potential of mean force (PMF) calculation.[38,40] To alleviate this shortcoming, a set of restraints using additional CVs, representing the minimum information to define one partner with respect to the other one, are applied in the course of the separation as prescribed in the geometrical route introduced by Gumbart et al.[20] These CVs are the backbone distance root-mean-square deviations (RMSDs) of the two proteins with respect to the reference, native conformation—i.e., in the bound state, the three Euler angles describing their relative orientation, and two additional angles (polar and azimuth) for their relative position. Applying geometrical restraints in the form of harmonic potentials onto these CVs results in an effective loss in configurational entropy, corresponding to conformational ($\Delta G_c^{\mathrm{site}}$), orientational ($\Delta G_o^{\mathrm{site}}$), and positional ($\Delta G_a^{\mathrm{site}}$) free-energy contributions, which must be accounted for in the computation of the standard binding free energy, both in the "bulk" (unbound state) and "site" (bound state). Therefore, the geometrical route involves a series of independent PMF calculations determined sequentially with the progressive introduction of restraints, prefacing the separation PMF calculation restrained along the specific axis "a". The binding free energy can then be expressed as a sum of these different free-energy contributions, namely,

$$\Delta G_{\mathrm{b}}^{\circ} = \Delta G_c^{\mathrm{bulk}} - \Delta G_c^{\mathrm{site}} - \Delta G_o^{\mathrm{site}} - \Delta G_a^{\mathrm{site}}(\theta,\phi) + \Delta G_o^{\mathrm{bulk}} - \frac{1}{\beta}\ln(S^* I^* C^{\circ}) \qquad (1)$$

where $\beta = (k_{\mathrm{B}}T)^{-1}$, with k_{B}, the Boltzmann constant, and T, the temperature. C° represents the standard concentration of 1 M, which corresponds to 1/1661 (Å^3).[41] I^* is the separation term, and S^* is a surface term, which represents the fraction of a sphere of radius r^*, centered at the binding site of the reference protein, accessible to its partner, that is,

$$\begin{cases} I^* &= \displaystyle\int_{\text{site}} \mathrm{d}r\ \mathrm{e}^{-\beta\left(w(r)-w(r^*)\right)} \\ S^* &= r^{*2}\displaystyle\int_0^{\pi}\mathrm{d}\theta\ \sin\theta\int_0^{2\pi}\mathrm{d}\phi\ \mathrm{e}^{-\beta u_a} \end{cases} \tag{2}$$

Here, r^* is a point located far away from the binding site, where the proteins no longer interact with each other, and u_a is the sum of the harmonic restraint potentials imposed on polar angles θ and ϕ to restrain the separation along the axis a.

In some cases, additional RMSD restraints acting on protein–protein interfacial side chains are required due to their exposure to the solvent and the possibility of their local conformational change in the course of the separation,[20] For the four selected complexes, specific side chains from the residues displayed in Table 1 were restrained based on visual inspection of the starting structures. The numbering of residues defined from the position in multiple sequence alignment (MSA) to facilitate direct comparison between complexes. ARG150 is, however, present in a region that lacks conservation between Dprs, hence, it is not assigned a number within the MSA.

Table 1. Side chains restrained residue numbers during the geometrical route using their position in the MSA.

Complexes	Dpr restrained residues	DIP restrained residues
Dpr6-DIPα	ARG150 (not in MSA),30,36	8, 9, 42, 44
Dpr6-DIPβ	10, 11, 22	9, 42, 44
Dpr6-DIPγ	20, 21, 32, 34	42
Dpr6-DIPθ	7, 10, 11, 20, 21, 30, 32	9, 40, 42

Computational assays

All simulations were performed using the NAMD3 program.[42] The all-atoms macromolecular CHARMM36[43] force field and the TIP3P model[44] were used to describe the proteins and the water, respectively. All computational assays corresponded to a physiological concentration of

NaCl of 0.15 M. The temperature (300 K) and the pressure (1 atm) were kept constant employing a Langevin thermostat[45] and the Langevin piston algorithm,[46] respectively. Long-range electrostatic interactions were handled by the particle-mesh Ewald (PME) algorithm.[47] Van der Waals and short-range electrostatic interactions were truncated with a smoothed 12-Å spherical cutoff. The equations of motion were integrated every 2 fs.

Theoretical models of each interacting moiety, namely their first immunoglobulin domain, were generated using homology modeling and simulated for 200 ns at equilibrium.[17] The starting structures for the binding free energy calculations were the resulting equilibrated complexes.[17] They were piped into the binding free-energy estimator 2 (BFEE2),[48] a tool designed to help set up binding free-energy calculations for protein–ligand complexes,[49] and expanded to protein–protein complexes by including RMSD calculations of the backbone of each protein, both in the bulk and at the binding site. The well-tempered extended adaptive biasing force algorithm (WTM-eABF)[50] as implemented in the collective variable module (Colvars) of NAMD[39] was chosen as the importance-sampling algorithm. The PMFs were run sequentially for all complexes, starting from the distance RMSD of the proteins with respect to the native conformation up to the physical separation of the two proteins. Once all the PMF calculations were completed, BFEE2[48] was invoked again at the post-processing stage, to extract the individual contributions to the binding affinity from each PMF, and to infer the final binding free-energy estimate.

Machine Learning strategies

Input features generation

One of the most demanding and pivotal aspects of ML lies in the generation of the input features. These features serve to describe our protein–protein complex interacting regions (interfaces) for every complex in a suitable format for ML algorithms[35] We chose to rely on evolutionary features, better at describing the binding interface at the residue level.[35] The mutual information (MI) was

obtained based on a multi-sequence alignment of the first immunoglobulin domain for each protein family[17] performed with CLUSTAL-W.[51] The MI is defined as the difference between individual entropy at the residue position in the MSA, H_i and H_j, and the joint entropy H_{ij}.

$$\mathrm{MI} = -H_{ij} + H_i + H_j = \sum_{x,y} p_{ij}(x,y) \log_2 \frac{p_{ij}(x,y)}{p_i(x) p_j(y)} \tag{3}$$

where $p_i(x)$ and $p_j(y)$ stand for the occurrence probability of a given amino acid (x or y) at position i or j, $p_{ij}(x,y)$ is the joint probability of the amino acids at the positions i and j in the multiple sequence alignment. This metric is utilized to estimate how much information is gained by accounting for the covariance of residues of DIP and Dpr rather than by treating them separately. The MI between non-cognate pairs is treated as noise, since they are not supposed to have co-evolved, and are, therefore, generated by simple combinatorial. Consequently, the MI generated by non-cognate pairs is subtracted from the one obtained for cognate pairs. Only residue pairs with a non-zero MI value were kept for the input features. To account solely for interacting residues at the interface, the MI is then weighted by an inverse distance, either from $\alpha-$carbon atoms extracted from the crystal structure of the Dpr10-DIPα[52] since every complex shares the same fold, or from the inverse distance between the COM of the side chains extracted from MD equilibrium trajectories taken from reference 17. To consider the energetics associated with residue interactions, a pairwise amino-acid (AA) potential is added to the MI and distance for every residue pair of the complex. The sum of these features for all residues forms a score for the complex.

$$\mathrm{score} = \sum_{\mathrm{AApairs}} (\mathrm{MI} \times \mathrm{distance} \times \mathrm{potential}) \tag{4}$$

ML algorithms and parameters

To distinguish between cognate and non-cognate complexes, two different ML algorithms were selected for their simplicity and their interpretability properties, namely linear discriminant analysis (LDA)[53] and random forest (RF).[54] The scikit-learn library[55] and the specific functions *LinearDiscriminantAnalysis* and *RandomForestClassifier* in Python 3.8.5[56]were employed. The selection criteria selected for RF was the entropy with no limit on the entropy value or tree length to limit tree growth. The training test consisted of 80% of the full dataset picked randomly, leaving the remaining 20% as the validation set, and was repeated a thousand times to ascertain the robustness of our results.

Extraction of sequences from the database

To test out the trained models on similar proteins from different *Drosophila* species, we searched in the Uniprot database,[57] for each DIP and Dpr, the most similar proteins based on their amino-acids sequences, using BLAST (basic local alignment searching tool) version 2.9.0+.[58] Each recovered sequence was assigned to a specific DIP and Dpr based on similarity scores, the maximum being 100%, and the minimum 47.6%. In case of a conflict of assignment between two recovered sequences, the choice was made towards the one with the highest identity score. To define this score, BLAST was used in combination with the BLOSSUM62 substitution matrix,[59] recommended for queries longer than 85 amino acids with a threshold of 10. Not all the DIPs and Dprs were recovered in the different selected *Drosophila* species, and the names of the missing proteins are reported in Table 2.

Table 2. Summary of the Blast search reporting the name of proteins for which the sequence was not recovered in Uniprot[57] for all 14 *Drosophila* species considered in this study

Species (*Drosophila...*)	Missing DIPs	Missing Dprs	Cognate complexes	Non − cognate complexes
... *Melanogaster*	-	-	57	174
... *Busckii*	κ	19, 21	51	139
... *Sechellia*	β, λ	7	42	138
... *Persimillis*	β, ι, λ, κ	7,8,12,21	34	85
... *Simulans*	β, δ, λ	5,7,12,21	37	99
... *Rhopaloa*	β, η	8,11,21	42	120
... *Guanche*	β, λ	-	46	143
... *Pseudoobscura Pseudoobscura*	-	-	57	174
... *Ananassae*	-	-	57	174
... *Kikkawai*	λ	-	53	157
... *Virilis*	β	3	47	153
... *Grimshawi*	β, λ	-	46	143
... *Willistoni*	α	8,12,13,21	49	121
... *Mojavensis*	-	3	54	166

Results and Discussion

Binding Free-Energy Calculations

The binding free-energy estimates, $\Delta G^{\circ}_{\mathrm{b}}$, of four Dpr-DIP complexes, were determined using the geometrical route,[20] and are reported in Table 3. Dpr6-DIPβ and Dpr6-DIPα are both cognate complexes, for which the exact binding free energies are known experimentally.[5] Conversely, Dpr-6DIPθ and Dpr6-DIPγ are both non-cognate complexes, for which only a rough estimation of their binding affinity is available.

Table 3. Binding free energy estimates using the geometrical route

Complex	$\Delta G^{\circ}_{\mathrm{b}}$ (kcal/mol)	$\Delta G^{\circ}_{\mathrm{b}}$, exp (kcal/mol)	Simulation time (μs)
Dpr6-DIPα	-6.7 ± 0.1	-7.7[5]	2.1
Dpr6-DIPβ	-5.4 ± 0.4	-6.5[5]	2.3
Dpr6-DIPγ	-4.5 ± 0.4	< -4.1[5]	2.2
Dpr6-DIPθ	-2.3 ± 0.6	< -4.5[5]	1.5

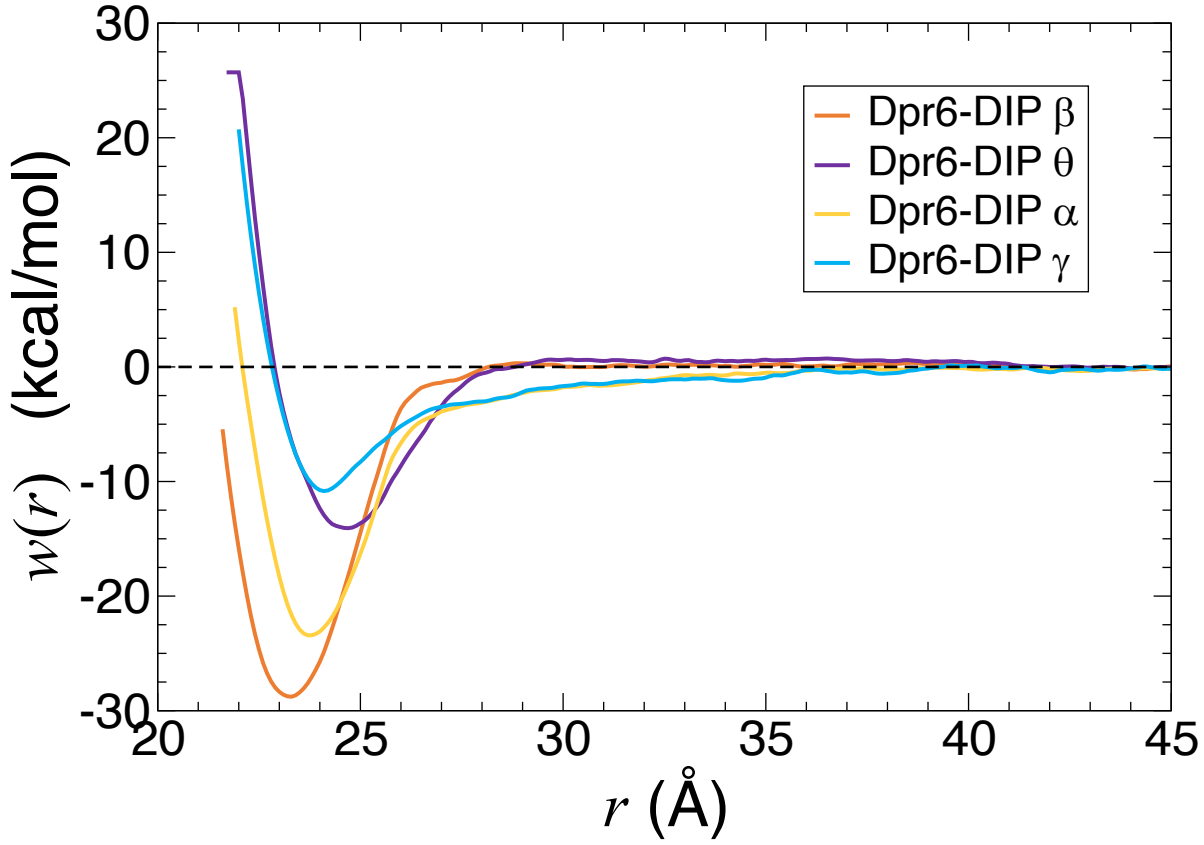


Figure 1. Separation PMFs of all Dpr6 complexes examined with the geometrical route. r stands for the Euclidean distance between the protein COMs.

During the physical separation simulation, a progressive decay of the PMF in the shape of successive plateaus was observed, mirroring a loss of interaction at the interface, most likely due to side-chain isomerization in response to solvent exposure. As stated in the Methods section, introducing additional restraints on the RMSD of those interacting side chains was sufficient to correct this artificial behavior, and recover the experimental value.[5] The same procedure was applied to the non-cognate complexes, and, as anticipated, binding free energies lower than those of the cognate ones were obtained.

In the SPR assays,[5] complexes with a dissociation constant (K_d) above 200 μM, which corresponds to binding affinities lower than -5.1 kcal/mol, were considered as non-cognate. Nevertheless, Cosmanescu et al. provided an estimation of the binding affinity, namely lower than -4.1

and -4.5 for Dpr6-DIPγ and Dpr6-DIPθ, respectively. Our $\Delta G^{\circ}_{\mathrm{b}}$ for Dpr6-DIPγ is slightly above the experimental threshold, while remaining below the limit separating cognate from non-cognate complexes (-5.1 kcal/mol). Dpr6-DIPθ computed binding affinity, on the other hand, is under the expected value of -4.5 kcal/mol. The lesser accuracy of the dissociation constant measurements due to less sensitive sensor chips[5] precludes a more precise validation of our predictions for the non-cognate complexes.

Comparing the different physical separation PMFs (see Figure 1), the depth of the wells reveals a major difference between the cognate and non-cognate complexes. The non-cognate complexes are close to -10 kcal/mol, and the cognates are in the -24 to -30 kcal/mol region, demonstrating a stark contrast in their separation behavior, which could help roughly categorize the complexes without the necessity to follow the entire geometrical route.

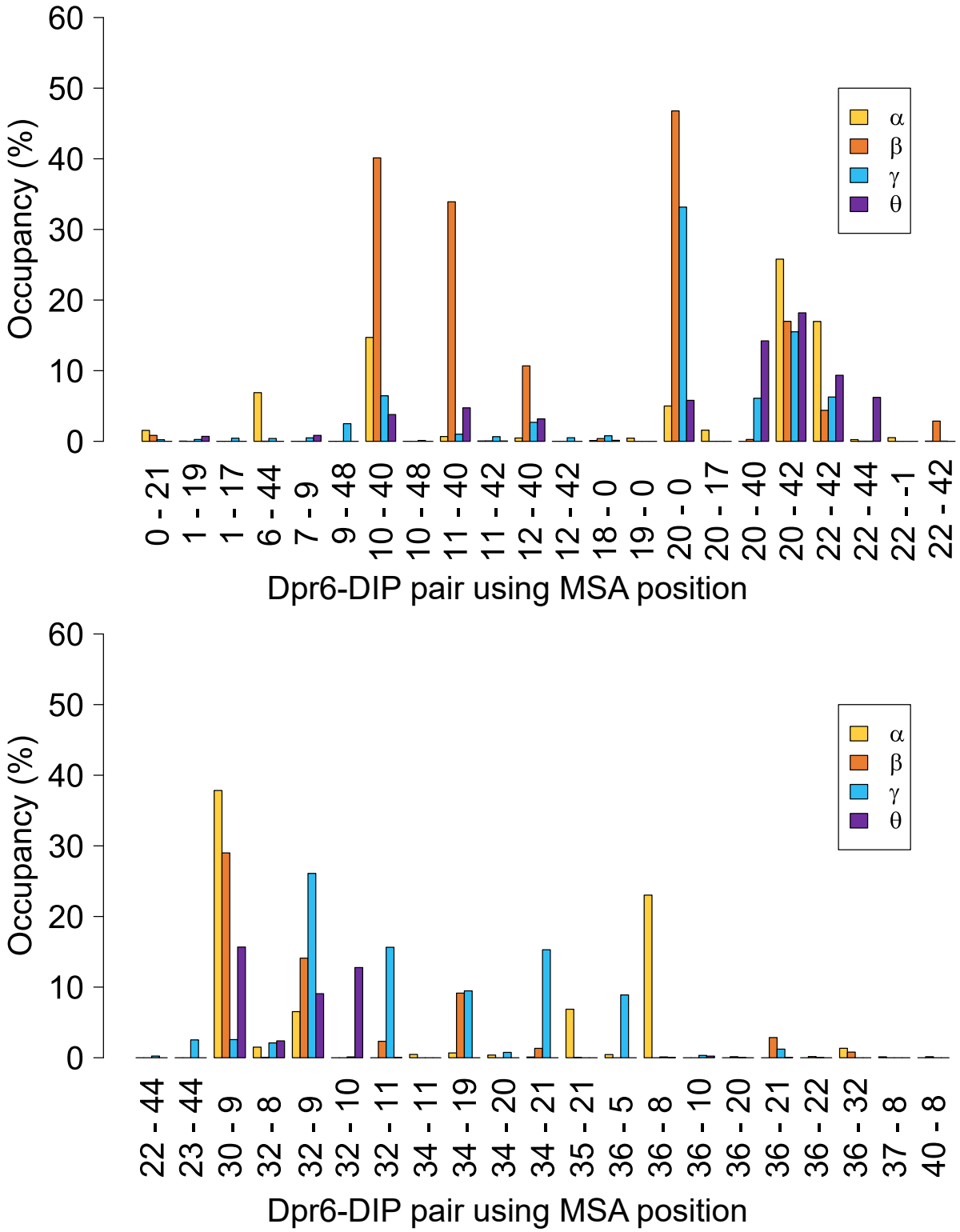


Figure 2. Hydrogen-bonds occupancies obtained during the separation trajectories of the geometrical route when the proteins are in close contact (in the 4 Å interval of the distance corresponding to their PMF minima).

Looking into the protein–protein interactions of simulated complexes, the hydrogen-bond occupancies (see Figure 2) for residue pairs detected by VMD[60] were examined. A clear difference arises between cognate (α, β) and non-cognate (γ, θ) complexes for the 10-40 and 30-9 pairs (using the MSA Dpr-DIP residue positions). Furthermore, RMSD side-chain restraints are present for β and θ in pair 10-40 for α and θ in pair 30-9, and yet a clear difference with a lower occupancy in the case of the non-cognate complexes is observed in the separation simulations. This result confirms that cognate complexes form more stable interactions than non-cognate ones. The close binding affinity estimates for β (cognate) and γ (non-cognate) might stem from similar high occupancies for several pairs (20-0, 34-19). Overall, the remarkable accord observed between the experimental and the predicted binding affinities obtained using the geometrical route offers a compelling evidence for the reliability and robustness of this methodology in effectively classifying complexes and providing a more precise binding affinity than the experimental assay. Nevertheless, the specific tuning required to perform these calculations cannot be easily implemented for all Dpr-DIP complexes within a reasonable time and bereft of human intervention.Table 3 shows that in an ideal world, MD could be able to discriminate the complexes, but it is computationally prohibiting to do all of the complexes now. Leveraging the documented affinities and a pre-establish classification inferred from SPR assays,[5] we turned next to supervised ML techniques to gain additional insight into the distinction between the two types of complexes.

Machine Learning Strategies

Nandigrami et al.[17] established that an LDA-weighted AIMS algorithm was able to discriminate between cognate and non-cognate partners in a set of 231 complexes, achieving an accuracy of approximately 0.88. In their work, they applied a pairwise amino-acids elementary potential, assigning a value of $+2$ to strong interactions (e.g., R and D) while highly repulsive interactions (e.g., K and K), were given a -2 value. However, this initial approach has certain limitations. Specifically, it disregards the contribution of several amino acids (e.g., alanine, threonine, serine, and

glycine) by attributing them a value of 0, thereby treating them as non-interacting residues. Due to the inherent simplifications in this elementary pairwise potential, we sought a more accurate scoring metric to describe the interaction of residue pairs. In particular, we considered four different potentials obtained with different approaches, all presented in the Supporting Information (SI). Dill and Thomas proposed a pairwise potential for amino acids, which was constructed iteratively until it could effectively distinguish between native and decoy conformations.[61] Following the same aim, Dima et al.[62] proposed a pairwise amino-acids potential using a coarse-grained representation of the amino acids, therefore reducing their complexity while still capturing the essential interactions. Similarly, Dosztányi et al.[63] developed their pairwise interaction potential by employing a quadratic form of the amino acids composition. This potential was derived through statistical analysis of a comprehensive database of globular proteins, allowing residue-residue interactions that contribute to the stability and folding of proteins to be identified. Furthermore, Betancourt et al. conducted all-atom simulations in a water environment to extract the interaction potential of all amino acids using radial density functions, which were then utilized to build a coarse-grained model.[64]

Exchanging the elementary potential with those based either on calculations or experiments[61–64] merely improved the accuracy by 0.04 at best, as reported in Table 4. Interestingly, the exclusion of the four amino acids discussed previously did not adversely affect the final classification, hence indicating that their contribution to the binding distinguishability was minimal and suggesting that modification of the pairwise potential was not the crucial factor in enhancing the performance of the ML algorithm. However, when using a neutral value of 1 as the interacting potential for all amino acids as a control, the accuracy dropped to 0.5 when using LDA, emphasizing the importance of incorporating an interaction potential in input features for the classification algorithm. Besides, this observation underscores that the elementary potential, despite its simplicity, contains sufficient information for an accurate classification. Since the affinity between proteins is non-linear, we replaced LDA with a non-linear classification algorithm, namely RF (see Methods section for details). The RF model exhibited remarkable performance by achieving an accuracy of

up to 0.98 when utilizing the elementary potential, surpassing the LDA's accuracy by 0.1. Furthermore, testing each potential using the RF model revealed comparable accuracies across different potential conditions, reaffirming our earlier assertion regarding the limited impact of distinct potentials on classification accuracy.

Table 4. Accuracy results using different pairs amino acids potentials for LDA and RF

Potential	LDA		RF	
	training	validation	training	validation
Elementary[17]	0.8667	0.7707	0.9814	0.9252
Dill[61]	0.9241	0.8283	0.9923	0.9302
Dima[62]	0.9109	0.7849	0.9925	0.9339
Dosztányi[63]	0.8901	0.7471	0.9924	0.9311
Betancourt[64]	0.9164	0.8127	0.9922	0.9251

A thorough examination of the feature importance mapped back to specific residues in the RF model provided insights into the key residues contributing to distinguishing between cognate and non-cognate protein–protein complexes (Figure 3 B and C). However, upon further investigations of the interactions mediated by these residues using the available MD equilibrium trajectories for randomly selected complexes, no direct distinction could be established between cognate and non-cognate complexes.

In addition to the potential, another parameter that can be fine-tuned to improve the description of the complexes at the residue level is the distance embedded in the input features. To reflect the actual interacting distance more accurately in the score than with a simple $\alpha-$carbon atom distance between pairs identical for all complexes, we took advantage of the equilibrium MD simulations. Specifically, we extracted the distances separating the side-chains COMS from the MD trajectories and calculated an average distance over the last 100 ns of the simulation. The updated ML results are gathered in Table 5 below.

Incorporating the COM mean distance instead of that of $\alpha-$carbon atoms, lead to similar results for the training set when using both LDA and RF. However, a slight decrease in performance was observed for RF, with a validation set loss of approximately 0.1, indicating a possible overfitting

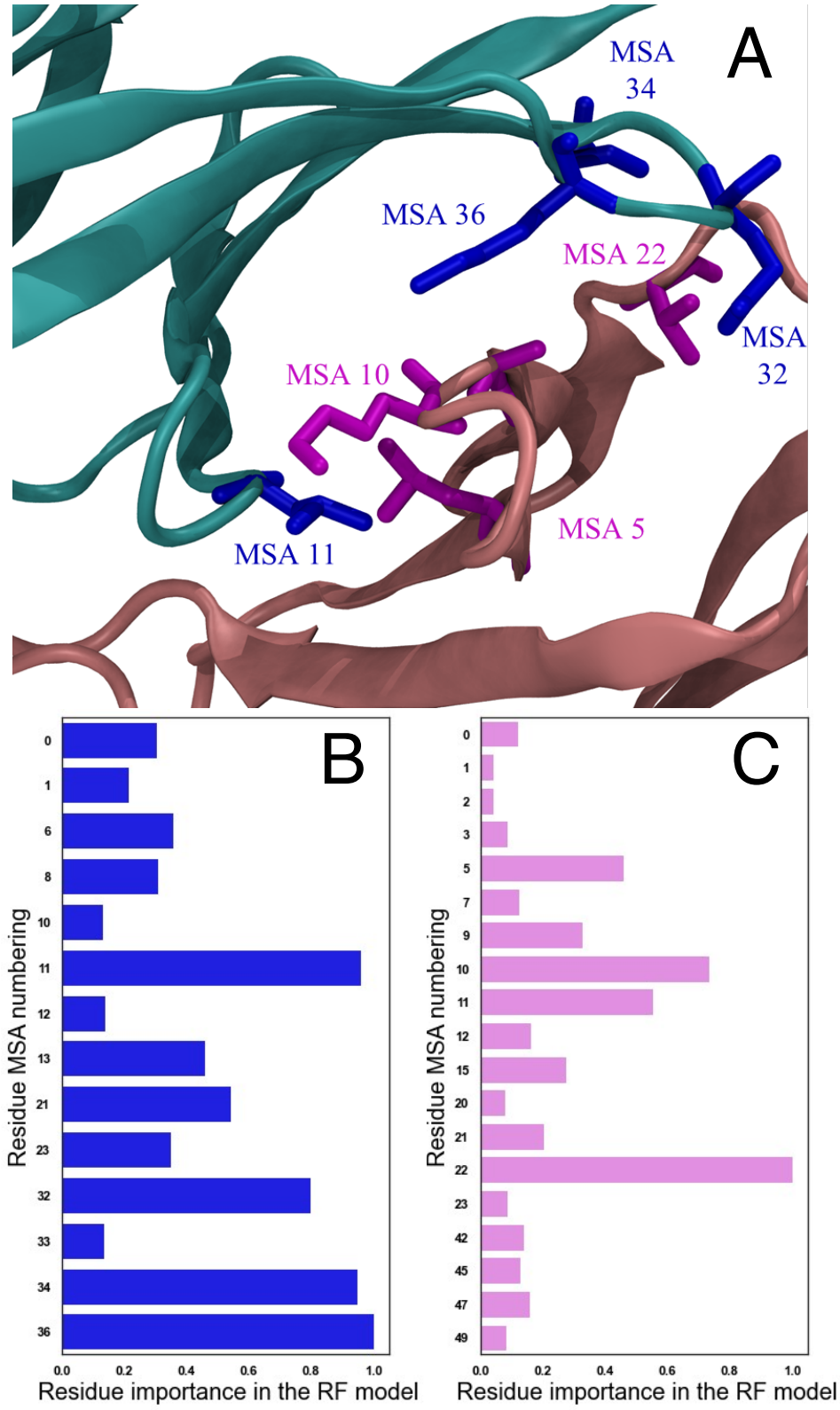


Figure 3. A) Interface pattern of Dpr1-DIPα with MSA important residue B) Dpr per-residue importance according to the RF model and MSA position C) DIP per-residue importance according to the RF model with backbone distance and MSA position.

Table 5. Accuracy results for LDA or RF with side chains COM mean distance

Potential	LDA		RF	
	training	validation	training	validation
Elementary[17]	0.8230	0.6956	0.9998	0.8364
Dill[61]	0.9333	0.8082	0.9999	0.8650
Dima[62]	0.9267	0.7857	0.9999	0.8849
Dosztányi[63]	0.8886	0.7493	0.9999	0.8627
Betancourt[64]	0.8805	0.7200	0.9999	0.8390

problem, meaning that the model has overlearned some patterns from the training that are not transferable to new data and can cause significant generalization problems. Surprisingly, the COM mean distance from MD simulations did not provide any significant information to distinguish between cognate and non-cognate complexes. Moreover, the identification of important residues was found to be identical to that obtained using the $\alpha-$carbon atom distance, neither showing gain of additional structural information nor contributing to a better understanding of the distinctive characteristics between cognate and non-cognate complexes.

Table 6. Accuracy results for LDA or RF with side chains COM mean distance and all pairs at the interface

Potential	LDA		RF	
	training	validation	training	validation
Elementary[17]	0.9888	0.7596	1.0	0.8364
Dill[61]	0.9751	0.7811	1.0	0.873
Dima[62]	0.9682	0.7873	0.9999	0.8753
Dosztányi[63]	0.9677	0.7921	0.9999	0.8747
Betancourt[64]	0.9626	0.7450	0.9999	0.8671

In an effort to obtain a more comprehensive description of the binding interface and identify critical residues, we extended our ML approach to include all residue pairs rather than focusing on MI pre-selected residue pairs. This broader strategy resulted in an improved accuracy for the LDA model, with values similar to those for the RF model with $\alpha-$carbon distance. For the RF model, the results are similar to the previous RF with MD distance showing no improvement. LDA, with a more detailed and tailored description of the complexes, was able to perform on par with RF, incorporating less information, emphasizing the importance of feature engineering when

developing ML models for the classification of protein–protein complexes.

Analyzing the two per-residue profiles (see Figure 3B and 4A), we still identify identical critical residues as when using MI interacting pairs, such as 5, 10, and 22 for the Dprs, with high importance scores above 0.5 in both cases. Residue 10 was identified as well through the hydrogen-bond analysis with high occupancies for the cognate complexes in the geometrical route, demonstrating its specific importance. In the case of the DIPs, residues 32 and 34 exhibited both a relative importance above 0.5 for both the MI and all pairs. Interestingly, residue 32 was observed to form hydrogen bonds with weak occupancy in the cognate complexes examined in the geometrical route (see Figure 2)

While extending the analysis to include all residue pairs validated some critical residues identified through MI pairs, our investigation revealed that direct classification relying on these critical residues interactions in the modeled complex structures is impossible. This finding is in line with the study of Sergeeva et al., which demonstrates that the compatibility between complexes is predominantly governed by negative constraints rather than by the formation of new and stronger interactions. These constraints were found to be specific to each binding subgroup defined by Sergeeva et al.,[65] emphasizing the need for careful consideration of subgroup-specific constraints and, highlighting the challenge of generalizing the findings to other complexes.

Expanding to Drosophila Family Homodimers and to Other Species

To assess the robustness and further validate our ML approach differently, we applied the trained ML algorithms (LDA or RF) to other Dpr-DIP complexes from other close species. Thirteen species from the Drosophila family with the most DIP and Dpr matches in the Uniprot database[57] were considered, namely *Drosophila buskii*, *Drosophila sechellia*, *Drosophila persimillis*, *Drosophila simulans*, *Drosophila rhopaloa*, *Drosophila guanche*, *Drosophila pseudoobscura pseudoobscura*, *Drosophila ananassae*, *Drosophila kikkawai*, *Drosophila virilis*, *Drosophila grimshawi*, *Drosophila*

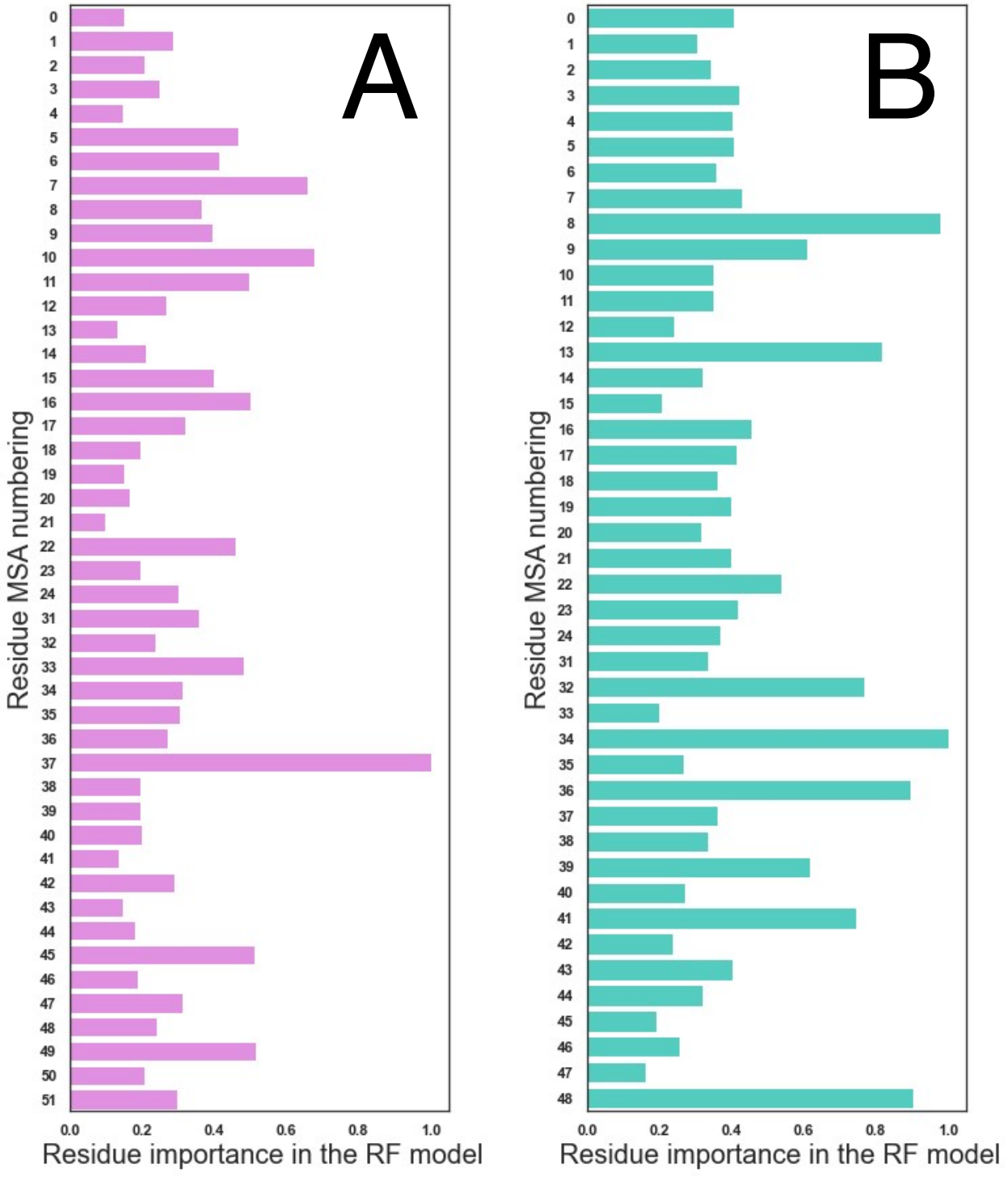


Figure 4. Per-residues importance according to MSA position for the RF model using molecular dynamics distance, all pairs at the interface, and the Dima potential[62] for A) the Dprs and B) the DIPs.

Willistoni, and *Drosophila mojavensis* (see the detailed procedure in the Methods section).The level of annotation in the Uniprot database for the recovered protein sequences of these thirteen species is thin (score of one or two out of five in the majority), and several were only referred to as their genetic loci (genetic position). This result can be explained by the focus on *Drosophila Melanogaster* as a biologically relevant model, given the short lifetime, the high numbers of larvas, and the low cost associated with their breeding.[66,67]

Table 7. Accuracy results for trained LDA and RF using the different species of *Drosophila*.

Species (*Drososphila ...*)	LDA accuracy	RF accuracy
... *Busckii*	0.92	1.0
... *Sechellia*	0.95	0.98
... *Persimillis*	0.99	1.0
... *Simulans*	0.98	0.99
... *Rhopaloa*	0.94	0.99
... *Guanche*	0.96	0.99
... *Pseudoobscura Pseudoobscura*	0.92	0.995
... *Ananassae*	0.92	0.99
... *Kikkawai*	0.93	0.98
... *Virilis*	0.94	1.0
... *Grimshawi*	0.95	0.96
... *Willistoni*	0.96	0.99
... *Mojavensis*	0.93	0.98

Due to the high percentage of identity for the found protein matches (the lowest being 44.4% for Dpr12 in *Drosophila Grimshawi*, and 99% for the highest, see the SI), we assume that cognate complexes in *Drosophila Melanogaster* were conserved in the different *Drosophila* species, and that the fold was identical to reuse the Dpr10-DIPα $\alpha-$carbon atoms distance. We chose the Betancourt potential[64], as it provided the best results with these conditions for both RF and LDA

(see Table 4).

We obtained an accuracy in the range of $0.92-0.96$ and $0.98-1.0$, for LDA and RF, respectively. It is noteworthy that the accuracies observed in our study align closely with the training phase conducted on *Drosophila Melanogaster*, which demonstrates the remarkable consistency and robustness of our strategy. We, therefore, anticipate that our ML models will exhibit a similar level of proficiency when presented with novel Dpr-DIP complexes. Thus, the models will be able to effectively distinguish and accurately classify these complexes as either cognate or non-cognate.

The Dpr-DIP interactome comprises also a few homodimers formed by DIPα, DIPη, DIPθ, DIPζ, Dpr8, Dpr12 and Dpr21. We generated the input features for all possible homodimers using a distance taken from one specific DIP homodimer (DIPα, PDB: 6EFY). Using the already trained LDA model, we successfully predicted the correct classes of such homodimers, except for DIPζ, which corresponds to the weakest binding affinity amongst the three available measurements, namely for DIPα, DIPη and DIPζ. This result underscores the predictive power of the present methodology generalized from hetero– to homodimers, yet training only on a single type of complexes. It is noteworthy, however, that RF fails to predict with the expected high accuracy (0.54) the strongest candidates amongst the DIPs homodimers, which may stem from a potential overfitting, as has been detected in previous analysis (see section Machine Learning Strategies). The binding affinities of the homodimers are also weaker than those characteristic of the heterodimers formed between the proper partners, which can explain the difficulty of RF to accurately identify the strong binders. The homophilic interactions are hypothesized to play a role in rescuing a default in heterophilic interactions.[68]

Members of the Dpr and DIP families are homologous to those of the IgLONs family in humans, which is composed of the opioid-binding protein/cell adhesion molecule-like (OPCML/OBCAM/IgLON1), the neurotrimin (NTM/IgLON2), the limbic system associated membrane protein (LSAMP/IgLON3), the neuronal growth regulator 1 (NEGR1/IgLON4), and the IGLON5. They can form hetero– and homodimers with a high affinity, i.e., -9.9 kcal/mol for the weakest complex,

the IGLON4 homodimer.[69] We tested our model on the 25 interaction combinations, considering all complexes as cognate, given their affinity for one another.[69] The distance feature was extracted from the IgLON5/NEGR1 heterodimer structure with PDB accession number 6DLD.[69] The model was modified to use the MI directly in lieu of the MI difference between cognate and non-cognate, since we do not have any non-cognate IgLON complex. The LDA model predicted all the complexes as cognate, whereas RF failed to classify them accurately. To further verify the LDA results, we assigned IgLON complexes with affinity below -11 kcal/mol as non-cognate complexes to use the MI difference in the input feature. This LDA model still classifies all complexes as cognate, which is expected given the greater affinity of IgLONs compared to that of the strongest Dpr-DIP complex (-8.2 kcal/mol for Dpr9-DIPλ).[5] The success of the LDA model for the IgLONS further validates our findings that the algorithm is applicable to different species, or complexes. In stark contrast, RF fails to predict on a different dataset, thus highlighting the limitations of the algorithm. However, when aiming at identifying the key residues for binding in a particular dataset, with no interest in transferability, RF outperforms LDA by learning better the specific patterns and leveraging the RF ensemble properties to obtain the relative residue importance in the classification.

Conclusion

In this work, we sought to elucidate the molecular basis differentiating strong (cognate) and weak (non-cognate) binders of the Dpr-DIP interactome with state-of-the-art approaches resting on a restraint-based method, the so-called geometrical route,[20] and ML with RF[54] and LDA.[53] The excellent agreement between experimental and computational binding free energy estimates obtained with the geometrical route[20] for two cognate complexes underscores the predictive power of the methodology. The estimation of low binding affinities for non-cognate complexes further validated the method by providing low estimates that fit within the expected range. We emphasize the need for specific side chain restraints in this particular family of proteins due to the exposure of the inter-

face to water, although it may only sometimes be necessary.[20] However, the geometrical route has limitations of its own when dealing with a large number of congeneric complexes due to its computational cost and unique tuning requirements, incompatible with high-throughput predictions. As an alternative, we turned to ML methods, which are well-suited for handling large datasets. The difficulty resides in choosing an appropriate description of the binding interface in a proper format for the algorithm.[35] We used input features composed of a multiple sequence alignment (sequence-based approach) with a structure-based distance and a pre-established pairwise interaction potential (physicochemical compatibility), proving that combining different types of information is vital for accurate predictions. These three components are non-specific, so that the input features can be seamlessly transferred to any similar biological interacting problem, like the binding specificity of cell-adhesion proteins, such as cadherins,[6] as we demonstrated the transferability of the trained LDA model to homologous proteins, namely IgLONs and Dpr-DIP homodimers. Furthermore, using per-residue importance extracted from the RF model, we were able to identify key residues necessary for Dpr-DIP binding. However, they are insufficient to use directly as a criterion to separate randomly selected cognate and non-cognate complexes. It should be noted that this observation is specific to the Dpr-DIP interactome,[65] and should not hinder the identification of essential residues for a different set of protein complexes. Additionally, we acknowledge that our classification of Dpr-DIP complexes was based solely on in-vitro assays[5], overlooking crucial biological factors such as cell expression and localization, the importance of which has been investigated previously.[70,71] These factors should be taken into account for a more comprehensive understanding of the Dpr-DIP interactome at a biological level. Overall, our study provides insights into the molecular basis of binding specificity in the Dpr-DIP interactome and highlights the potential of combining physics-based and ML approaches to analyze protein–protein interactions and quickly separate the wheat from the chaff—i.e., the strong binders from the weaker ones, acting like a selection filter for high throughput. Our method is primarily aimed at protein families with evolutionarily linked members due to the MSA requirement for the MI in the input features. Furthermore, supervised ML requires prior knowledge of the "target" value to learn how to predict

in the training process. Datasets with robust structural and thermodynamic information are scarce, and mainly directed at protein–ligand complexes, e.g., PDBbind,[72] which contains 19,433 protein–ligand complexes versus 2,852 protein–protein complexes in the current version (2020), thereby limiting the use of supervised ML algorithms for protein–protein standard binding free-energy predictions.

Data and Software Availability

The input files required for calculating the binding free energies of all molecular assemblies included in this study can be found in the Supporting Information. The geometric approach is incorporated into BFEE2, accessible at https://github.com/fhh2626/BFEE2. A concise tutorial on utilizing BFEE2 for protein?protein standard binding free-energy calculations is also available in the Supporting Information.

Supporting Information Available

The supporting information contains the free-energy contributions of the $\Delta G^{\circ}_{\mathrm{b}}$ computations for the four selected Dpr-DIP complexes, the pairwise amino-acid potentials, as well as the parameter files, the initial configuration files, and the input files for the simulations described herein (ZIP).

Acknowledgement

The authors are grateful to Christopher Boughter, Florence Sczepaniak and Engin Ozkan for numerous helpful discussions and for their support. This investigation was supported by the University of Lorraine through the LUE initiative and by the Agence Nationale de la Recherche (Protea-

seInAction and LOR-AI) and by the France and Chicago Collaborating in The Sciences (FACCTS) program. BR was supported the National Science Foundation (grant MCB-1517221).

References

(1) Winding, M.; Pedigo, B. D.; Barnes, C. L.; Patsolic, H. G.; Park, Y.; Kazimiers, T.; Fushiki, A.; Andrade, I. V.; Khandelwal, A.; Valdes-Aleman, J. et al. The connectome of an insect brain. *Science* **2023**, *379*, eadd9330.

(2) Chen, B. L.; Hall, D. H.; Chklovskii, D. B. Wiring optimization can relate neuronal structure and function. *Proc. Natl. Acad. Sci. USA* **2006**, *103*, 4723–4728.

(3) Sperry, R. W. Chemoaffinity in the orderly growth of nerve fiber patterns and connections. *Proc. Natl. Acad. Sci. USA* **1963**, *50*, 703–710.

(4) Xu, S.; Sergeeva, A. P.; Katsamba, P. S.; Mannepalli, S.; Bahna, F.; Bimela, J.; Zipursky, S. L.; Shapiro, L.; Honig, B.; Zinn, K. Affinity requirements for control of synaptic targeting and neuronal cell survival by heterophilic IgSF cell adhesion molecules. *Cell Rep.* **2022**, *39*, 110618.

(5) Cosmanescu, F.; Katsamba, P. S.; Sergeeva, A. P.; Ahlsen, G.; Patel, S. D.; Brewer, J. J.; Tan, L.; Xu, S.; Xiao, Q.; Nagarkar-Jaiswal, S. et al. Neuron-subtype-specific Expression, interaction affinities, and specificity determinants of DIP/Dpr cell recognition proteins. *Neuron* **2018**, *100*, 1385–1400.e6.

(6) Brasch, J.; Katsamba, P. S.; Harrison, O. J.; Ahlsén, G.; Troyanovsky, R. B.; Indra, I.; Kaczynska, A.; Kaeser, B.; Troyanovsky, S.; Honig, B. et al. Homophilic and heterophilic interactions of type II cadherins identify specificity groups underlying cell-adhesive behavior. *Cell Rep.* **2018**, *23*, 1840–1852.

(7) Goodman, K. M.; Yamagata, M.; Jin, X.; Mannepalli, S.; Katsamba, P. S.; Ahlsén, G.; Sergeeva, A. P.; Honig, B.; Sanes, J. R.; Shapiro, L. Molecular basis of sidekick-mediated cell-cell adhesion and specificity. *eLife* **2016**, *5*, e19058.

(8) Özkan, E.; Carrillo, R. A.; Eastman, C. L.; Weiszmann, R.; Waghray, D.; Johnson, K. G.; Zinn, K.; Celniker, S. E.; Garcia, K. C. An extracellular interactome of immunoglobulin and LRR proteins reveals receptor-ligand etworks. *Cell* **2013**, *154*, 228–239.

(9) Singh, P. SPR Biosensors: Historical Perspectives and Current Challenges. *Sens. Actuators B: Chem.* **2016**, *229*, 110–130.

(10) Alford, R. F.; Leaver-Fay, A.; Jeliazkov, J. R.; O'Meara, M. J.; DiMaio, F. P.; Park, H.; Shapovalov, M. V.; Renfrew, P. D.; Mulligan, V. K.; Kappel, K. et al. The Rosetta All-Atom Energy Function for Macromolecular Modeling and Design. *J. Chem. Theory Comput.* **2017**, *13*, 3031–3048.

(11) Homeyer, N.; Gohlke, H. Free Energy Calculations by the Molecular Mechanics Poisson-Boltzmann Surface Area Method. *Mol. Inform.* **2012**, *31*, 114–122.

(12) Chen, R.; Li, L.; Weng, Z. ZDOCK: an initial-stage protein-docking algorithm. *Proteins* **2003**, *52*, 80–87.

(13) Onufriev, A. V.; Case, D. A. Generalized Born Implicit Solvent Models for Biomolecules. *Annu. Rev. Biophys.* **2019**, *48*, 275–296.

(14) Srinivasan, J.; Cheatham, T. E.; Cieplak, P.; Kollman, P. A.; Case, D. A. Continuum Solvent Studies of the Stability of DNA, RNA, and Phosphoramidate-DNA Helices. *J. Am. Chem. Soc.* **1998**, *120*, 9401–9409.

(15) Im, W.; Beglov, D.; Roux, B. Continuum solvation model: Computation of electrostatic forces from numerical solutions to the Poisson-Boltzmann equation. *Comput. Phys. Commun.* **1998**, *111*, 59–75.

(16) Lee, M. S.; Olson, M. A. Calculation of Absolute Protein-Ligand Binding Affinity Using Path and Endpoint Approaches. *Biophys. J.* **2006**, *90*, 864–877.

(17) Nandigrami, P.; Szczepaniak, F.; Boughter, C. T.; Dehez, F.; Chipot, C.; Roux, B. Computational Assessment of Protein–Protein Binding Specificity within a Family of Synaptic Surface Receptors. *J. Phys. Chem. B* **2022**, *126*, 7510–7527.

(18) Swanson, J. M. J.; Henchman, R. H.; McCammon, J. A. Revisiting free energy calculations: a theoretical connection to MM/PBSA and direct calculation of the association free energy. *Biophys. J.* **2004**, *86*, 67–74.

(19) Pearlman, D. A. Evaluating the Molecular Mechanics Poisson-Boltzmann Surface Area Free Energy Method Using a Congeneric Series of Ligands to p38 MAP Kinase. *J. Med. Chem.* **2005**, *48*, 7796–7807.

(20) Gumbart, J. C.; Roux, B.; Chipot, C. Efficient determination of protein-protein standard binding free energies from first principles. *J. Chem. Theory Comput.* **2013**, *9*, 10.1021/ct400273t.

(21) Izrailev, S.; Stepaniants, S.; Isralewitz, B.; Kosztin, D.; Lu, H.; Molnar, F.; Wriggers, W.; Schulten, K. Steered Molecular Dynamics. Computational Molecular Dynamics: Challenges, Methods, Ideas. Berlin, Heidelberg, 1999; pp 39–65.

(22) Jarzynski, C. Nonequilibrium Equality for Free Energy Differences. *Phys. Rev. Lett.* **1997**, *78*, 2690–2693.

(23) Park, S.; Khalili-Araghi, F.; Tajkhorshid, E.; Schulten, K. Free energy calculation from steered molecular dynamics simulations using Jarzynski's equality. *J. Chem. Phys.* **2003**, *119*, 3559–3566.

(24) Chipot, C. Free energy methods for the description of molecular processes. *Ann. Rev. Biophys.* **2023**, *52*, 113–138.

(25) Woo, H.-J.; Roux, B. Calculation of absolute protein–ligand binding free energy from computer simulations. *Proc. Natl. Acad. Sci. USA* **2005**, *102*, 6825–6830.

(26) Jiménez, J.; Škalič, M.; Martínez-Rosell, G.; De Fabritiis, G. KDEEP: Protein–Ligand Absolute Binding Affinity Prediction via 3D-Convolutional Neural Networks. *J. Chem. Inf. Model.* **2018**, *58*, 287–296.

(27) Jiménez, J.; Doerr, S.; Martínez-Rosell, G.; Rose, A. S.; De Fabritiis, G. DeepSite: protein-binding site predictor using 3D-convolutional neural networks. *Bioinformatics* **2017**, *33*, 3036–3042.

(28) Hassan-Harrirou, H.; Zhang, C.; Lemmin, T. RosENet: Improving Binding Affinity Prediction by Leveraging Molecular Mechanics Energies with an Ensemble of 3D Convolutional Neural Networks. *J. Chem. Inf. Model.* **2020**, *60*, 2791–2802.

(29) Bitencourt-Ferreira, G.; de Azevedo, W. F. Development of a machine-learning model to predict Gibbs free energy of binding for protein-ligand complexes. *Biophys. Chem.* **2018**, *240*, 63–69.

(30) Gebhardt, J.; Kiesel, M.; Riniker, S.; Hansen, N. Combining Molecular Dynamics and Machine Learning to Predict Self-Solvation Free Energies and Limiting Activity Coefficients. *J. Chem. Inf. Model.* **2020**, *60*, 5319–5330.

(31) Riniker, S. Molecular Dynamics Fingerprints (MDFP): Machine Learning from MD Data To Predict Free-Energy Differences. *J. Chem. Inf. Model.* **2017**, *57*, 726–741.

(32) Hashemifar, S.; Neyshabur, B.; Khan, A. A.; Xu, J. Predicting protein–protein interactions through sequence-based deep learning. *Bioinformatics* **2018**, *34*, i802–i810.

(33) Conti, S.; Ovchinnikov, V.; Karplus, M. ppdx: Automated modeling of protein–protein interaction descriptors for use with machine learning. *J. Comput. Chem.* **2022**, *43*, 1747–1757.

(34) Das, S.; Chakrabarti, S. Classification and prediction of protein–protein interaction interface using machine learning algorithm. *Sci Rep* **2021**, *11*, 1761.

(35) Casadio, R.; Martelli, P. L.; Savojardo, C. Machine learning solutions for predicting protein–protein interactions. *WIREs Comput. Mol. Sci* **2022**, *12*, e1618.

(36) Ovchinnikov, S.; Kamisetty, H.; Baker, D. Robust and accurate prediction of residue–residue interactions across protein interfaces using evolutionary information. *eLife* **2014**, *3*, e02030.

(37) Yin, S.; Proctor, E. A.; Lugovskoy, A. A.; Dokholyan, N. V. Fast screening of protein surfaces using geometric invariant fingerprints. *Proc. Natl. Acad. Sci. USA* **2009**, *106*, 16622–16626.

(38) Chipot, C., Pohorille, A., Eds. *Free energy calculations. Theory and applications in chemistry and biology*; Springer Verlag, 2007.

(39) Fiorin, G.; Klein, M. L.; Hénin, J. Using collective variables to drive molecular dynamics simulations. *Mol. Phys.* **2013**, *111*, 3345–3362.

(40) Blazhynska, M.; Goulard Coderc de Lacam, E.; Chen, H.; Roux, B.; Chipot, C. Hazardous Shortcuts in Standard Binding Free Energy Calculations. *J. Phys. Chem. Lett.* **2022**, 6250–6258.

(41) Deng, Y.; Roux, B. Calculation of Standard Binding Free Energies: Aromatic Molecules in the T4 Lysozyme L99A Mutant. *J. Chem. Theory Comput.* **2006**, *2*, 1255–1273.

(42) Phillips, J. C.; Hardy, D. J.; Maia, J. D. C.; Stone, J. E.; Ribeiro, J. V.; Bernardi, R. C.; Buch, R.; Fiorin, G.; Hénin, J.; Jiang, W. et al. Scalable molecular dynamics on CPU and GPU architectures with NAMD. *J. Chem. Phys.* **2020**, *153*, 044130.

(43) Huang, J.; MacKerell, A. D. CHARMM36 all-atom additive protein force field: Validation based on comparison to NMR data. *J. Comput. Chem.* **2013**, *34*, 2135–2145.

(44) Mark, P.; Nilsson, L. Structure and Dynamics of the TIP3P, SPC, and SPC/E Water Models at 298 K. *J. Phys. Chem. A* **2001**, *105*, 9954–9960.

(45) Uhlenbeck, G. E.; Ornstein, L. S. On the Theory of the Brownian Motion. *Phys. Rev.* **1930**, *36*, 823–841.

(46) Feller, S. E.; Zhang, Y.; Pastor, R. W.; Brooks, B. R. Constant pressure molecular dynamics simulation: The Langevin piston method. *J. Chem. Phys.* **1995**, *103*, 4613–4621.

(47) Darden, T.; York, D.; Pedersen, L. Particle mesh Ewald: An *N* -log(*N*) method for Ewald sums in large systems. *J. Chem. Phys.* **1993**, *98*, 10089–10092.

(48) Fu, H.; Chen, H.; Cai, W.; Shao, X.; Chipot, C. BFEE2: Automated, Streamlined, and Accurate Absolute Binding Free-Energy Calculations. *J. Chem. Inf. Model.* **2021**, *61*, 2116–2123.

(49) Fu, H.; Chen, H.; Blazhynska, M.; Goulard Coderc de Lacam, E.; Szczepaniak, F.; Pavlova, A.; Shao, X.; Gumbart, J. C.; Dehez, F.; Roux, B. et al. Accurate determination of protein:ligand standard binding free energies from molecular dynamics simulations. *Nat. Protoc.* **2022**, *17*, 1114–1141.

(50) Fu, H.; Shao, X.; Cai, W.; Chipot, C. Taming Rugged Free Energy Landscapes Using an Average Force. *Acc. Chem. Res.* **2019**, *52*, 3254–3264.

(51) Thompson, J. D.; Higgins, D. G.; Gibson, T. J. CLUSTAL W: improving the sensitivity of progressive multiple sequence alignment through sequence weighting, position-specific gap penalties and weight matrix choice. *Nucleic Acids Res.* **1994**, *22*, 4673–4680.

(52) Cheng, S.; Park, Y.; Kurleto, J. D.; Jeon, M.; Zinn, K.; Thornton, J. W.; Özkan, E. Family of neural wiring receptors in bilaterians defined by phylogenetic, biochemical, and structural evidence. *Proc. Natl. Acad. Sci. USA* **2019**, *116*, 9837–9842.

(53) Barker, M.; Rayens, W. Partial least squares for discrimination. *J. Chemom.* **2003**, *17*, 166–173.

(54) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(55) Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V. et al. Scikit-learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.

(56) van Rossum, G. *Python tutorial*; 1995.

(57) The UniProt Consortium, UniProt: the universal protein knowledgebase in 2021. *Nucleic Acids Res.* **2021**, *49*, D480–D489.

(58) Altschul, S. F.; Gish, W.; Miller, W.; Myers, E. W.; Lipman, D. J. Basic local alignment search tool. *J. Mol. Biol.* **1990**, *215*, 403–410.

(59) Henikoff, S.; Henikoff, J. G. Amino acid substitution matrices from protein blocks. *Proc. Natl. Acad. Sci. USA* **1992**, *89*, 10915–10919.

(60) Humphrey, W.; Dalke, A.; Schulten, K. VMD – Visual Molecular Dynamics. *Journal of Molecular Graphics* **1996**, *14*, 33–38.

(61) Thomas, P. D.; Dill, K. A. An iterative method for extracting energy-like quantities from protein structures. *Proc. Natl. Acad. Sci. USA* **1996**, *93*, 11628–11633.

(62) Dima, R. I.; Settanni, G.; Micheletti, C.; Banavar, J. R.; Maritan, A. Extraction of interaction potentials between amino acids from native protein structures. *J. Chem. Phys.* **2000**, *112*, 9151–9166.

(63) Dosztányi, Z.; Csizmók, V.; Tompa, P.; Simon, I. The Pairwise Energy Content Estimated from Amino Acid Composition Discriminates between Folded and Intrinsically Unstructured Proteins. *J. Mol. Biol.* **2005**, *347*, 827–839.

(64) Betancourt, M. R.; Omovie, S. J. Pairwise energies for polypeptide coarse-grained models derived from atomic force fields. *J. Chem. Phys.* **2009**, *130*, 195103.

(65) Sergeeva, A. P.; Katsamba, P. S.; Cosmanescu, F.; Brewer, J. J.; Ahlsen, G.; Mannepalli, S.; Shapiro, L.; Honig, B. DIP/Dpr interactions and the evolutionary design of specificity in protein families. *Nat Commun* **2020**, *11*, 2125.

(66) Tolwinski, N. S. Introduction: Drosophila—A Model System for Developmental Biology. *J Dev Biol* **2017**, *5*, 9.

(67) Jennings, B. H. Drosophila – a versatile model in biology & medicine. *Mater Today* **2011**, *14*, 190–195.

(68) Xu, S.; Xiao, Q.; Cosmanescu, F.; Sergeeva, A. P.; Yoo, J.; Lin, Y.; Katsamba, P. S.; Ahlsen, G.; Kaufman, J.; Linaval, N. T. et al. Interactions between the Ig-Superfamily Proteins DIP-α and Dpr6/10 Regulate Assembly of Neural Circuits. *Neuron* **2018**, *100*, 1369–1384.e6.

(69) Ranaivoson, F. M.; Turk, L. S.; Ozgul, S.; Kakehi, S.; von Daake, S.; Lopez, N.; Trobiani, L.; De Jaco, A.; Denissova, N.; Demeler, B. et al. A Proteomic Screen of Neuronal Cell-Surface Molecules Reveals IgLONs as Structurally Conserved Interaction Modules at the Synapse. *Structure* **2019**, *27*, 893–906.e9.

(70) Morey, M. Dpr-DIP matching expression in Drosophila synaptic pairs. *Fly (Austin)* **2016**, *11*, 19–26.

(71) Wang, Y.; Lobb-Rabe, M.; Ashley, J.; Chatterjee, P.; Anand, V.; Bellen, H. J.; Kanca, O.; Carrillo, R. A. Systematic expression profiling of Dpr and DIP genes reveals cell surface codes in Drosophila larval motor and sensory neurons. *J. Dev.* **2022**, *149*, dev200355.

(72) Wang, R.; Fang, X.; Lu, Y.; Yang, C.-Y.; Wang, S. The PDBbind Database Methodologies and Updates. *J. Med. Chem.* **2005**, *48*, 4111–4119.