scientific reports

OPEN

Electrostatic complementarity at the interface drives transient protein-protein interactions

Greta Grassmann[1,2], Lorenzo Di Rienzo[2], Giorgio Gosti[2,3], Marco Leonetti[2,3], Giancarlo Ruocco[2,4], Mattia Miotto[2,5]✉ & Edoardo Milanetti[2,4,5]✉

Understanding the mechanisms driving bio-molecules binding and determining the resulting complexes' stability is fundamental for the prediction of binding regions, which is the starting point for drug-ability and design. Characteristics like the preferentially hydrophobic composition of the binding interfaces, the role of van der Waals interactions, and the consequent shape complementarity between the interacting molecular surfaces are well established. However, no consensus has yet been reached on the role of electrostatic. Here, we perform extensive analyses on a large dataset of protein complexes for which both experimental binding affinity and pH data were available. Probing the amino acid composition, the disposition of the charges, and the electrostatic potential they generated on the protein molecular surfaces, we found that (i) although different classes of dimers do not present marked differences in the amino acid composition and charges disposition in the binding region, (ii) homodimers with identical binding region show higher electrostatic compatibility with respect to both homodimers with non-identical binding region and heterodimers. Interestingly, (iii) shape and electrostatic complementarity, for patches defined on short-range interactions, behave oppositely when one stratifies the complexes by their binding affinity: complexes with higher binding affinity present high values of shape complementarity (the role of the Lennard-Jones potential predominates) while electrostatic tends to be randomly distributed. Conversely, complexes with low values of binding affinity exploit Coulombic complementarity to acquire specificity, suggesting that electrostatic complementarity may play a greater role in transient (or less stable) complexes. In light of these results, (iv) we provide a novel, fast, and efficient method, based on the 2D Zernike polynomial formalism, to measure electrostatic complementarity without the need of knowing the complex structure. Expanding the electrostatic potential on a basis of 2D orthogonal polynomials, we can discriminate between transient and permanent protein complexes with an AUC of the ROC of ∼ 0.8. Ultimately, our work helps shedding light on the non-trivial relationship between the hydrophobic and electrostatic contributions in the binding interfaces, thus favoring the development of new predictive methods for binding affinity characterization.

Interactions among proteins constitute the molecular basis of most processes in living organisms, and their deregulation or disruption often leads to disease[1–3]. Among other things, such interactions may differ in the number (dimers, tetrameters, etc) and kind of involved proteins (homo or hetero complexes), the stability of the binding (transient/permanent bindings), and the type of the binding process, i.e. lock and key, induced fit and conformational selection. While it has been estimated that over 80% of proteins operate in molecular complexes[4], detailed comprehension of the mechanism behind the protein binding process and the stability of the resulting protein complexes is still incomplete. At a qualitative level, binding involves a recognition phase where distant molecules have to recognize themselves in the crowded cellular environment, followed by a docking process where the two molecules reorient/adapt to binding in specific regions. Despite this, complex formation is often highly specific: a binding partner could be recognized by only one of the members in a protein family even if they all have the same folds[5]. This compatibility is determined by an interplay between various contributions on the

[1]Department of Biochemical Sciences "Alessandro Rossi Fanelli", Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy. [2]Center for Life Nano & Neuro Science, Istituto Italiano di Tecnologia, Viale Regina Elena 291, 00161 Rome, Italy. [3]Soft and Living Matter Laboratory, Institute of Nanotechnology, Consiglio Nazionale delle Ricerche, 00185 Rome, Italy. [4]Department of Physics, Sapienza University of Rome, Piazzale Aldo Moro 5, 00185 Rome, Italy. [5]These authors contributed equally: Mattia Miotto and Edoardo Milanetti. ✉email: mattia.miotto@roma1.infn.it; edoardo.milanetti@uniroma1.it

molecular surface and can either (i) be present from the beginning, when the two proteins are far apart (lock and key model), or (ii) be assumed by the proteins while exploring their conformational landscape (conformational selection model) or be gained while interacting with the partner (induced fit model)[6–8]. Once the proteins are bound, their binding regions are known to display a combination of geometrical and chemical complementarities, which ultimately reflect on the binding stability[9–14].

At the level of amino acid composition, it is widely known that the composition of binding regions differs with respect to the rest of the solvent-exposed region: while the latter is preferentially populated by hydrophilic residues, binding regions have a higher number of hydrophobic residues, like Val and Leu, that tend to establish stronger van der Waals interactions[10,15]. From a geometrical point of view, the optimization of short-ranged interactions between atoms at the interface leads to a local shape complementarity of the proteins' molecular surfaces. Indeed, the side chain rearrangements minimize the van der Waals interaction, thus determining shape complementarity at the interfaces, which is typically evaluated by geometrical approaches[16–21].

Conversely, there is still no full consensus on the role played by electrostatic interactions, including hydrogen bonding, ionic/Coulombic, cation$-\pi, \pi-\pi$, lone-pair sigma hole, and orthogonal multipolar interactions[5,22–24]. In fact, acting at longer distances, it is unanimously understood that electrostatic compatibility plays a role at the beginning of the recognition process when partners are far away from each other[25]; indeed, proteins move in a very crowded environment and since electrostatic interactions are the most long-ranged ones, they can produce a drift in the Brownian motion of the two binding proteins. However, while this could be true for heterodimers (that may possess opposite charges), homodimers have the same net charge, thus attractive interactions can only take place between parts of the proteins (at the most)[26].

Therefore, many studies are focusing on assessing the electrostatic match of protein complexes, to better understand why and how binding happens[9,26–31]. In particular, McCoy *et al.*[32] found that binding sites are characterized by significant electrostatic complementarity, if defined as the correlation of surface electrostatic potential at binding sites on a small number of protein complexes. Another study discussed the importance of electrostatic interactions in the binding adaptation[33]. Shashikala and coworkers[25] investigated the role of electrostatic interactions in diseases, finding that disease-causing mutations frequently alter wild-type electrostatic interactions. Moreover, electrostatic turned out to be a key feature even for machine learning methods that look at the identification of protein-protein binding sites[18]. Similarly, electrostatic and shape complementarity turned out to be sufficient to predict the DNA-binding sites on proteins with 80% accuracy[29].

Since electrostatic interactions can act both at short and long distances[34], the interaction region that should be considered is of non-trivial definition. In fact, smaller regions are able to capture the binding properties due to van der Waals forces, which lead to a shape complementarity between the two interacting molecular surfaces. On the other hand, larger regions lose shape complementarity but involve more charged residues, which are typically excluded from the binding regions but are widely present in the other exposed regions, playing a crucial role in the thermal stability of the protein structure[35]. It is in fact known that electrostatic interactions between 9 and 12 Å are of crucial importance to distinguish obligate from non-obligate complexes[36], thus playing a key role in the characterization of the binding. For a quantitative description of the Lennard-Jones potential at the interfaces of protein complexes, we have recently shown that shape complementarity is maximized for interacting patches less than 9 Å[16], highlighting thus the effect of van der Waals interactions at the binding interface[10]. For a patch on the interface of this size, where the involvement of long-range electrostatic interactions is reduced, the quantification of the contribution of electrostatic complementarity remains unclear.

Here, we characterize the role of electrostatic interaction in protein binding and quantitatively measure the electrostatic complementarity at the interface of the molecular complexes, defining the binding interfaces in a 9 Å radius sphere. With this aim, we collect a dataset of protein complexes (see Methods for details) and characterize the complexes in terms of interface type, amino acid composition, and charge properties. To study the relationship between binding affinity and electrostatic complementarity we consider a second dataset, the 'Affinity' dataset (see Methods for details). Next, we analyze the contribution to the binding of electrostatic, by comparing the potential values of mirroring points on binding regions. We show that taking into account properly rescaled values of the electrostatic potential, we obtain a negative correlation between the complexes binding affinity and electrostatic complementarity. Following these results, we propose a new method able to quickly distinguish between interacting and non-interacting patches by describing their electrostatic potential projections with vectors and looking at the difference between these descriptors. This computational approach has been developed starting from the 2D Zernike method, that we proposed to quickly evaluate the shape complementarity at interfaces[16,37–41]; for what concerns shape complementarity, our method has already been demonstrated to be able to efficiently identify interacting regions by measuring the shape complementarity in terms of the Euclidean distance between the Zernike invariant descriptors associated with the projections of the molecular surfaces patches (see Methods for more details).

Here, the Zernike invariant vectors describe the electrostatic potential by considering in the same function both positive and negative values. As a final step, we show that the Zernike descriptions of the electrostatic allow for fast and superposition-free discrimination between transient and permanent interactions.

Results

Charges distribution and compatibility.

To characterize the role of electrostatic complementarity in protein binding, we collected a balanced dataset of human protein complexes for which structural data were available ('Human' dataset). The dataset is composed of 164 homodimers, which can be divided into 44 dimers with an Identical Binding Region (IBR-hom), 66 Shifted Binding Region (SBR-hom), and 54 non-Identical Binding Region (nIBR-hom), depending on the similarity of the interacting patches; finally, the dataset includes 35 heterodimers (nIBR-het). In addition to this dimer classification, the same complexes can also be structurally classified looking at the prevailing secondary structure of the proteins: the same dataset includes 133 complexes where both binding partners have a prevalence of helices residues (HH), 57 where both proteins have mostly strands residues (SS) and 9 complexes where one of the partner has more strand residues while the other one has more helices (SH). See Figs. 1a, S2a and Methods section for more details. To begin with, we investigated the charge distribution of the proteins in the 'Human' dataset and their amino acid composition with respect to their dimer classification. In Fig. 1b, the percentage of complexes that have a total sum of the charges (considering either all the residues of each protein or only the interacting ones) with opposite signs is displayed. According to this analysis, heterodimers are the only class whose interacting patches ($\sim$ 4%) have a discordant sum of the charges among binding partners. Looking more specifically at the charge of interacting residues, as shown in Figure S1a of the Supplementary, we observe that only $\sim$ 1.5% of the negative interacting residues are close to other negative residues on the other protein in the complex. Positive residues instead can be found in proximity more frequently, between 1.6% and 3.2% of the time, depending on the complex category and on radius defining the surrounding of a residue. This condition is particularly common for nIBR homodimers. Oppositely charged residues can be rarely found (0.15-0.25% of the time) among the interacting patches of SBR homodimers as well. On the other hand, heterodimers and IBR homodimers have a higher percentage, up to 3.6% and 3.2% respectively, of opposite charges facing each other in binding regions. Figure S1b in the Supplementary characterizes the residues surrounding non-charged amino acids as well. It can be seen that, in general, all complexes tend to have non-charged residue facing each other, confirming the predominantly hydrophobic nature of the interacting regions of the protein-protein complexes[42].

Figure 1c shows a general overview of the amino acid abundances, computed considering (i) all the residues in a protein, (ii) only the solvent-exposed ones, and (iii) only the ones included in the binding regions (see Methods for details about the definition of solvent-exposed residues and binding regions). The analysis of the amino acid composition confirms the (well-known) result[10] that hydrophobic amino acids, such as Ile or Met, are uncommon in the solvent-exposed surface of proteins. However, when one of them is present in the exposed regions, it is more likely to find it in a binding site rather than on the rest of the surface. On the contrary, charged amino acids, such as Lys or Glu, are more present on the surface, but the fraction taking part in the binding is relatively small. Figure 1c also shows that IBR homodimers tend to have less charged amino acid on the interacting regions, compared to heterodimers and nIBR homodimers.

Finally, we performed the same analyses for the three classes in which the dataset is divided when considering the secondary structures. The results are shown in Figure S2. In particular, Figure S2b shows that the heterodimers whose interacting patches have a discordant sum of the charges are classified as SS or HH. However, as shown in Figure S2c, when considering neighboring charges, SH complexes tend to have more negative interacting residues close to positive residues on other protein's surface, even reaching 3.3% for a small patch radius. For all three classes the hydrophobic nature and the amino acid composition of the binding sites are confirmed by Figure S2d and S2e.

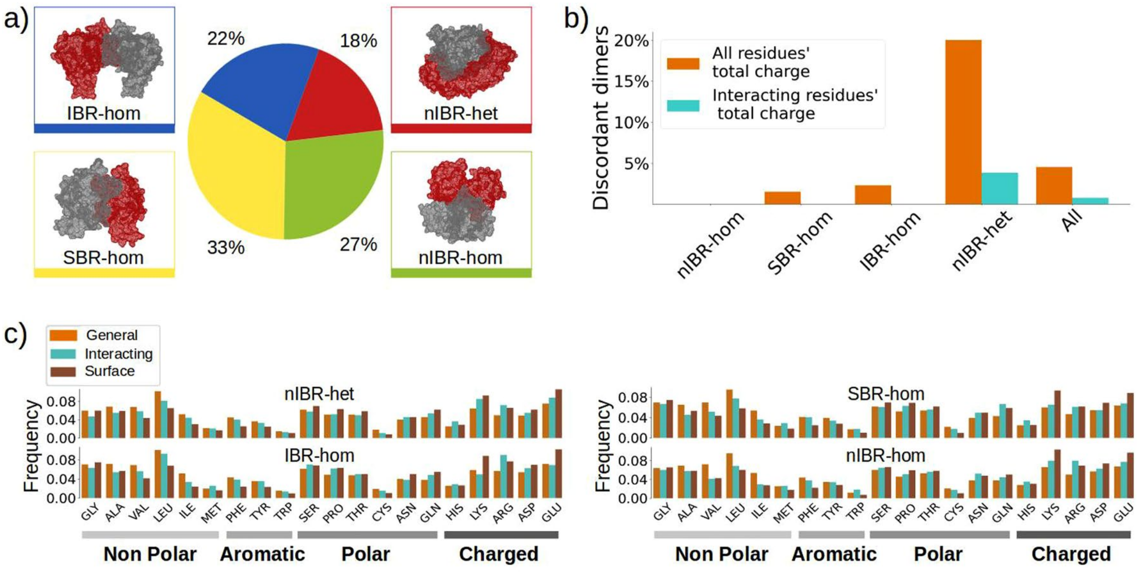

Figure 1. Amino acid composition, charge properties, and classification of the dataset. **(a)** The complexes in the dataset are divided into heterodimers and IBR, SBR and nIBR homodimers. The colored boxes report an example for each category. The same colors are used to indicate in the pie chart each class abundance in the dataset. **(b)** For each protein, the sum of the charges of all its residues and only the interacting residues on the surface is computed. For each complex, these total and interacting charges from the two interacting partners are multiplied. The bar plot shows, for the whole dataset and each class, the percentage of complexes whose total (in orange) and interacting (in blue) products are negative. **(c)** The relative abundances of each of the twenty natural amino acids considering all the residues (orange), only the interacting ones (in green), and only the solvent-exposed residues (brown) are shown. The results are divided into the four classes.

The specificity of electrostatic complementarity depends on both complex kinds and environmental factors. Next, we moved to analyze the spatial disposition of the electrostatic interactions. To do so, we evaluated the electrostatic potential generated by the protein charges on the molecular surface[43]. Being a high-level representation of the external protein structure, such representation allows for an efficient evaluation of both the geometrical and chemical complementarities. In fact, the amino acids composition analysis provides information only about the chemical and physical properties of the residues belonging to the two binding sites, while the resolution of the Poisson-Boltzmann equation provides information on the electrostatic potential distribution at the interfaces, due to all the atoms of the protein, even those not in close proximity with the binding site.

To analyze the contribution of electrostatic to the binding between two proteins, we define a quantitative measure of electrostatic complementarity able to differentiate between interacting and non-interacting regions. We start by analyzing the 'Human' dataset and comparing the electrostatic potential values of the binding regions with those of non-interacting regions. To measure the complementarity, we describe each patch with a Simplified Electrostatic Matrix (*SEM*). The latter is computed starting from the Electrostatic Matrix (*EM*), obtained by projecting the considered region of the electrostatic potential surface on the x-y plane, which is defined as the best fit of the surface points belonging to the specific patch. The matrix is then built on the plane and each pixel is associated with the average value of the electrostatic potential of the points projected into that pixel (see Fig. 2a and Methods section for more details). To obtain the *SEM*, we assign +1 and -1 to all the positive and negative pixels respectively. Figure 2a shows on the right the *SEM*s of two interacting patches, and in the center the corresponding *EM*s. To evaluate the electrostatic complementarity between two patches, we compare each pixel of the *SEM* describing the first patch with the pixel at the same position on the *SEM* describing the second

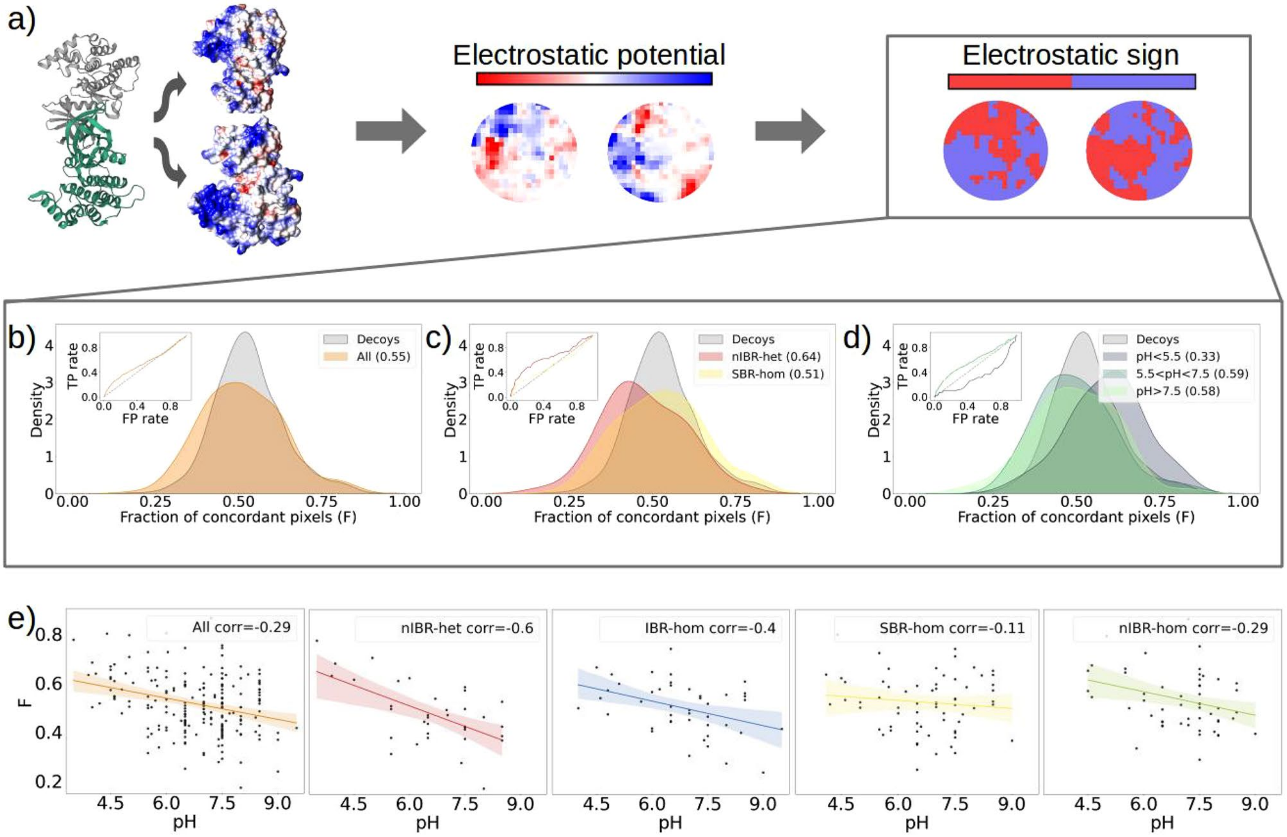

Figure 2. Electrostatic complementarity contribution in protein-protein complexes **(a)** On the left, the 3D representation of two proteins forming a complex and their electrostatic potential surfaces. In the center, the *EM*s of two interacting patches. Each pixel of the matrices is colored according to the electrostatic potential value of the surface points projected in that region. On the right, the *SEM*s of the same two patches: red and blue pixels correspond respectively to positive and negative values of the electrostatic potential. **(b)** Distributions of the *F* values computed for interacting (orange) and random (grey) patches taken from the 'Human' dataset. In the insert the corresponding ROC curve. **(c)** Distributions of the *F* values of the interacting patches in complexes from the nIBR-het (red) and SBR-hom (yellow) classes. In the insert the corresponding ROC curves. **(d)** The distributions of the *F* values computed for interacting patches in a) are classified in pH ranges: low in light purple, physiological in green, and high in light green. In the insert the corresponding ROC curves. **(e)** Fraction of concordant regions as a function of the pH and computed correlation (in the legend). From left to right the considered complexes are the whole 'Human' dataset, the nIBR-het, the IBR-hom, the SBR-hom, and the nIBR-hom.

one. In this way, we can check if surface points that face each other on the binding partners have electrostatic potentials with opposite signs. Operatively, we defined *F* as the fraction of pixel pairs in which the two pixels have the same sign: patches with a high electrostatic complementarity, i.e. a high number of opposite neighbor points with discordant electrostatic potentials, should have a low value of *F*. The measured complementarity is compared with that one would obtain by chance, randomly selecting surface points and building around each one a patch on the surface of the partner protein.

Figure 2b shows the distribution of *F* scores for the whole 'Human' dataset. In particular, one can see that random patches (i.e. decoys) have a gaussian-like distribution with a mode of $\tilde{0}.49$, as we expect that for two random patches the probability that a spatial corresponding region has an opposite sign is 0.5. The distribution of all complexes instead is shifted toward values lower than 0.5. In particular, it has a mode of 0.38, indicating that protein binding regions have a degree of electrostatic complementarity higher than what one would expect by chance. Indeed, this can be quantified by computing the ROC curve (see inset Fig. 2b) and evaluating the Area Under the Curve (AUC), which is 0.55 in this case.

Looking at the shape of the distribution, one can see that it appears to be composed of different populations of proteins; in fact, it presents bi/tri modalities. We thus proceeded to separate the dataset according to the complex classes.

Doing so, we found different behaviors for the various classes. To quantify the differences between the distributions, we evaluated (see Fig. 2c and Figure S3 in the Supplementary) the ROC curves of each distribution with respect to the decoy's one and computed the corresponding AUC. The classes with the lowest AUC of the ROC curve are SBR and nIBR homodimers (at 0.51 and 0.53 respectively), for which the classification performance can not be distinguished from that of random decoys. A slightly better classification is obtained for the IBR homodimers, with an AUC of the ROC curve of 0.56. On the other hand, heterodimers present an AUC of the ROC curve of 0.64.

Interestingly, a trend is observed also stratifying the dataset according to the pH of each complex. Figure 2d shows the distributions and relative ROC curves for three ranges of pH values, i.e. high (pH> 7.5), low (pH< 5.5), and physiological (5.5 <pH< 7.5). Complexes in the latter range have a high degree of electrostatic complementarity, having an AUC of the ROC curve of 0.58, whereas the *F* score of interacting patches in the low pH range is shifted to higher values (resulting in an AUC of the ROC curve of 0.33), meaning that in this case, the binding regions have a higher fraction of concordant points facing each other.

Moreover, in Fig. 2e, we show the fraction of concordant regions as a function of the experimental pH value for every complex of the 'Human' dataset and each subclass. It is interesting to note that while the whole Human dataset does not show a strong correlation with the pH (-0.29), the correlation values vastly differ among the subclasses. The *F* values computed for SBR and nIBR homodimers have a correlation of -0.11 and -0.29 respectively. On the other hand, IBR homodimers and heterodimers anti-correlate with the pH value, correlating -0.4 and -0.6. Nevertheless, Table 1 shows that independently from the class, the interacting patches of complexes in the low range have an AUC lower than 0.5, meaning that for low pH the electrostatic potential in points facing each other on interacting patches has the same sign. Table 1 reports as well how the performances of the *F* score for increasing radius R of the patch: after $R = 12$Å its characterization of interacting patches does not improve or even worsen. Even if larger regions include more charged residues by extending out of the hydrophobic binding sites, the complementarity of the charges is lost when the surfaces of the complex are not interacting. This analysis confirms our choice of a 9 Å radius to define the interacting patches.

Low-affinity interactions use electrostatic complementarity to achieve specificity. Since both the stratification by classes and pH did not fully account for the observed shape of the distribution, we look for binding affinity data. To do so, we collected a dataset of complexes with known structure and experimental dissociation constant, K_d. In particular, we took the dataset proposed by Desantis *et al.*, which we refer to as the 'Affinity' dataset (see Methods for details), which is exclusively composed of 123 heterodimers.

Figure 3a shows the *F* score distribution of interacting and random patches for the whole 'Affinity' dataset. The latter have a gaussian-like distribution with a mode of 0.49, as for the random patches of the 'Human' dataset. The former instead is shifted to lower values, having a mode of 0.31 and an AUC of the ROC curve of 0.69 (as shown in the insert of Fig. 3a).

Stratifying the dataset in three groups according to the complex binding affinities [$B_a = \log_{10}(K_d)$], we obtain the results shown in Fig. 3b,c. The three distributions, corresponding to high ($B_a < -9.0$), medium ($-9.0 < B_a < -6.0$) and low ($B_a > -6.0$) binding affinity are well separated and shifted on different ranges of *F* scores. Low-affinity complexes are moved to lower values of *F*, resulting in an AUC of the ROC curve of 0.81, while the ones with high binding affinity can not be distinguished from random decoys, having an AUC of the

F	$R = 6$	$R = 9$	$R = 12$	$R = 15$
$pH < 5.5$	0.32	0.33	0.33	0.35
$5.5 < pH < 7.5$	0.54	0.59	0.61	0.62
$pH > 7.5$	0.52	0.58	0.62	0.60

Table 1. AUC of the ROC curves of the *F* score for varying patch radius and pH. The AUC of the ROC curves are computed using the random distribution and the distributions of the 'Human' dataset interacting patches, divided according to the pH. Increasing values of the radius R defining the patches are tested.

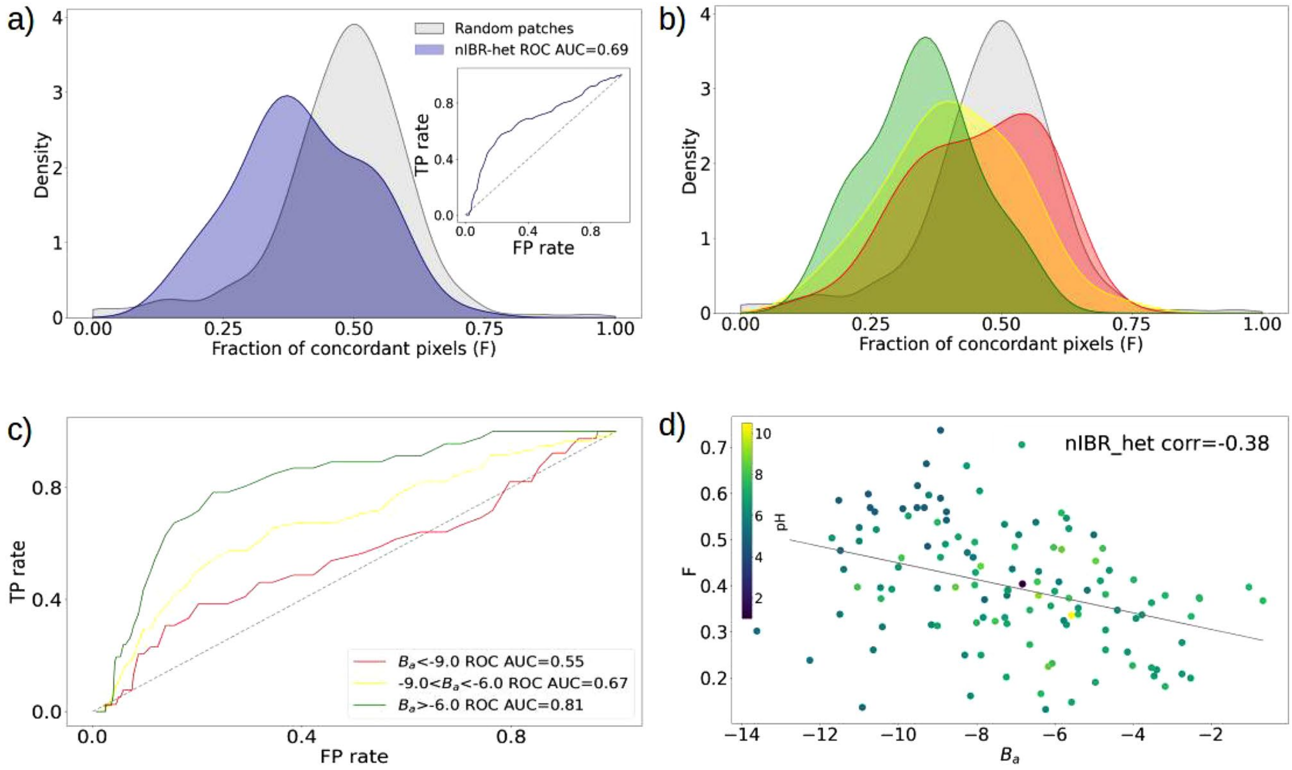

Figure 3. Electrostatic complementarity contribution in the binding stability of complexes. **(a)** Distributions of the *F* values computed for interacting (violet) and random (grey) patches taken from the 'Affinity' dataset. In the insert the corresponding ROC curve. **(b)** The distribution of the *F* values of the interacting patches in a) is divided according to the binding affinity of the complex: high affinity in red, medium in yellow, and low in green. In grey is the distribution of the random decoys. **(c)** ROC curves and corresponding AUC (in the legend) of the distributions of the interacting patches in b), computed against the random distribution. **(d)** Fraction of concordant regions as a function of the binding affinity and computed correlation (in the legend). Each point is colored according to the pH value, as indicated by the color bar.

ROC curve of 0.55. The medium binding affinity complexes cover an intermediate range of *F* values and have an AUC of 0.67. Interestingly, if we look at the *F* value of each complex as a function of its binding affinity, we get a negative correlation (-0.38) with the binding affinity, as shown in Fig. 3d. Table 2 shows how the performance of the F score changes for increasing radius R of the patch: as already discussed for Table I, after R = 12 Å its characterization of interacting patches does not improve.

Projecting the molecular surface on an orthogonal basis allows to compactly describe the electrostatic contribution to the interface of complexes.

Leveraging on the results of the previous sections, we looked for a compact method to simultaneously measure both electrostatic and shape complementarity between protein patches.

To describe and compare the electrostatic surfaces more efficiently, we apply the 2D Zernike polynomials, which constitute a complete basis in which any function of two variables defined in a unitary disk can be decomposed. The Zernike expansion associates each portion of a surface to an ordered set of numerical descriptors, invariant under rotation, allowing an easy and fast metric comparison between different protein regions for complementarity evaluation (see Methods for details). The rotational invariance is a fundamental property in the blind search for interacting patches. The complementarity of the binding regions can then be evaluated in

F	$R = 6$	$R = 9$	$R = 12$	$R = 15$
$B_a < -9.0$	0.54	0.55	0.51	0.54
$-6.0 < B_a < -9.0$	0.67	0.67	0.65	0.65
$B_a > -6.0$	0.77	0.81	0.86	0.86

Table 2. AUC of the ROC curves of *F* for varying patch radius and binding affinity. The AUC of the ROC curves are computed using the binding and random regions *F*s distributions. The binding region distributions are divided into three groups, according to the binding affinity B_a of their complex. Increasing values of the radius R defining the patches are tested. The complexes are part of the 'Affinity' dataset.

terms of the euclidean distance between their corresponding Zernike vectors. In particular, we measure how much the distance between the Zernike descriptors of a pair of interacting sites is smaller than the distances between random patches.

Figure 4 (panels from a to d) shows a schematic representation of the computational protocol for comparing, in terms of shape and electrostatic, interacting proteins. For each protein, the molecular surface and the electrostatic potential surface are built. The former corresponds to the solvent-accessible surface, the latter is obtained by assigning to each point of the molecular surface the value of the electrostatic potential computed in that region as obtained by solving the Poisson-Boltzmann equation[44]. On each surface, a patch is iteratively selected, and the corresponding regions of both the molecular and electrostatic surfaces are separately projected onto a plane. An example of both projections for two interacting patches is shown in Fig. 4e. More details can be found in the Methods. We assess the shape and electrostatic complementarity between the patches by expanding in terms of Zernike polynomials the 2D projections of the molecular and electrostatic potential surfaces respectively. The distance between the Zernike vectors of interacting patches is smaller than the one between the vectors of two random patches, as shown in Fig. 4f. As shown in Fig. 4g,h and in Figure S4a,b in the Supplementary our results are in line with what has been observed in[16]: interacting patches are efficiently distinguished from random decoys, with an AUC of the ROC curve close to ~0.8. The class whose interacting sites can be better identified includes IBR homodimers (with a success rate of 0.96), whereas the lowest efficiency is obtained for nIBR homodimers (AUC at 0.72). Next, we extended the Zernike method for the study of electrostatic complementarity. Since Zernike coefficients can represent only real-valued functions over the unit disk we define the Confined Electrostatic Matrix (*CEM*). *CEM*s are obtained by capping the *EM*s pixels above +30 and below −30[45]. This allows us to obtain Zernike-expandable functions. We then define electrostatic complementarity as the distance between the Zernike vectors associated with the *CEM*s. Figure 4i-j and Figure S4c,d in the Supplementary show that this definition of complementarity reaches an efficiency in distinguishing interacting and random patches comparable with the one obtained with *F* values, with an AUC of the ROC close to 0.60. IBR and nIBR homodimers correspond to the best (0.68) and worst (0.55) performances. As previously assessed, complexes with low pH have a low electrostatic complementarity and can not be easily distinguished from random decoys, reaching an AUC of 0.55, as shown in Table 3. Interestingly, the opposite behavior can be observed for the shape complementarity, which is higher when the pH is low.

Finally, to evaluate the impact of the structural characteristics on electrostatic complementarity, we stratified our results -both concerning the F value and the Zernike distance- according to the prevailing secondary structures of the complex.

As shown in Fig. 5a, the F value distribution does not significantly vary between the three structural classes. Moreover, when the complexes are divided in SS, HH and SH the classification performance of the F value can not be distinguished from that of random decoys, reaching a value of the ROC AUC of 0.54, 0.56 and 0.5 respectively.

When looking instead at the fraction of concordant regions as a function of the experimental pH value for each of the three structural classes of the 'Human' dataset, more interesting observations can be done. Figure 5b confirms that complexes with a higher pH value tend to have a higher degree of electrostatic complementarity. This is particularly true for SH complexes, which have a correlation of -0.88 (p-value at 0.002). The F values of SS and SH complexes are more randomly distributed (correlation at -0.25 and -0.27 respectively), nevertheless this division results in an overall better correlation of the single classes with the pH, compared to what is obtained when considering the dimer class.

Next, we stratified the Zernike distances between the molecular surface patches according to this division. As shown in Figs. 5c and d, when considering shape complementarity SS complexes are the most easy to distinguish from random decoys (ROC AUC of 0.79). Note that face-to-face interactions between β-strands are usually characterized by a high shape complementarity[46]. SH complexes reach a ROC AUC of 0.76, instead.

The opposite trend can be observed for the electrostatic complementarity studied with the Zernike method: in this case, as depicted in Fig. 5e and f, SH complexes are the most distinguishable from random decoys (ROC AUC at 0.68), whereas SS complexes are the most difficult to classify (ROC AUC at 0.6).

Transient from permanent interactions can be distinguished solely based on the electrostatic complementarity.

At last, we apply our method to the 'Affinity' dataset to test the ability of our descriptor to distinguish between permanent and transient interactions, as this property has important effects on biological functions[47]. In particular, defining permanent (respectively transient) interactions based on the binding affinity[48] being lower (resp. higher) than $B_a = -6$, we obtained the distributions shown in Fig. 6a. Interestingly, transient interactions display higher than random electrostatic complementarity values (green distribution), while permanent interactions (red distribution) have Zernike distances slightly higher than that one would expect by chance. This can be quantified again by looking at the ROC curves in Fig. 6b and evaluating the AUC values. Indeed, transient interactions display an AUC of 0.69 with respect to the decoy distribution, while permanent interactions have an AUC of 0.43. Permanent interactions can be distinguished from transient ones with an AUC of the ROC of 0.78. Notably, knowing the pH of the considered complexes allows for an even better classification as one can see from Fig. 6c, d.

Finally, we tested the capability of the Zernike method to quantify the effect of point mutations at the interface on the binding affinity of the complex. Results, discussed in the SI, indicate that the electrostatic complementarity evaluated on five mutants with known mutations at the binding sites and experimentally determined dissociation constant is higher the lower is the stability of the complex.

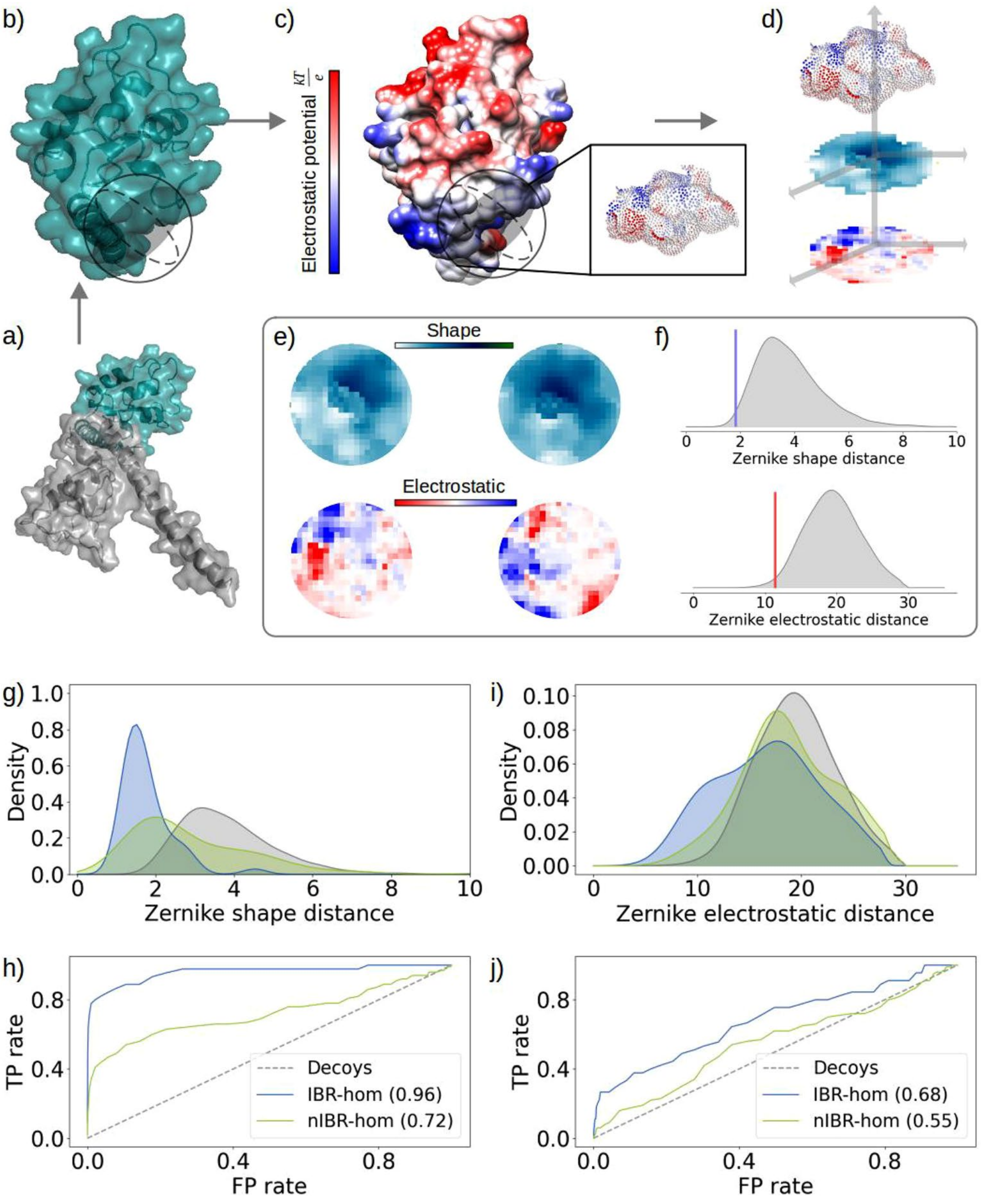

Figure 4. Schematic representation of the computational protocol and Zernike evaluation of complementarity. **(a)** Molecular representations of the surfaces of two proteins forming a complex. **(b)** Molecular representation of the surface of one of the proteins depicted in a). A sphere is used to select a possible patch on the surface: the dark shadow highlights the selected points. **(c)** Electrostatic potential surface, where each point is colored according to its electrostatic potential value. In the zoom, the region of the electrostatic potential surface corresponding to the patch selected in b).**(d)** 2D projections of the patch. In the blue scale the shape projection, for which the colors in the plane are determined by the distance of the surface points from a predefined origin (see Methods for details). In the blue-red scale the electrostatic projection, where the colors are determined by the electrostatic potential values of the above points. **(e)** Comparison between the shape and electrostatic projections of two binding regions. **(f)** In grey the distributions of the Zernike shape (top) and electrostatic (bottom) distances between random patches. The blue and red lines correspond to the distances between the Zernike vectors describing the two patches on top and bottom respectively in e). **(g)** Distributions of the distances between the Zernike vectors describing the molecular surface of IBR-hom and nIBR-hom interacting (blue and green respectively) and random (grey) patches in the 'Human' dataset. **(h)** ROC curves of the distributions in g) and corresponding AUC (in the legend) computed against the random distribution. **(i)** For each patch the distance between the Zernike vectors describing the electrostatic potential surface in that region is computed. Then the same analysis and classification as in g) is performed. **(j)** ROC curves of the distributions in i) and corresponding AUC (in the legend) computed against the random distribution.

	$Zernike_{Shape}$	$Zernike_{Electrostatic}$
$pH < 5.5$	0.84	0.55
$5.5 < pH < 7.5$	0.76	0.63
$pH > 7.5$	0.77	0.61

Table 3. Discriminating power of the Zernike-based expansion of molecular and electrostatic potential surface for different pH ranges. In the $Zernike_{Shape}$ column, the AUC of the ROC curves is computed from the distribution of the Zernike distances between the molecular surface of interacting patches and the random decoys. In the $Zernike_{Electrostatic}$ column the AUC of the ROC curves is computed considering the electrostatic potential description of the patches. The results are divided according to the pH of the complexes.

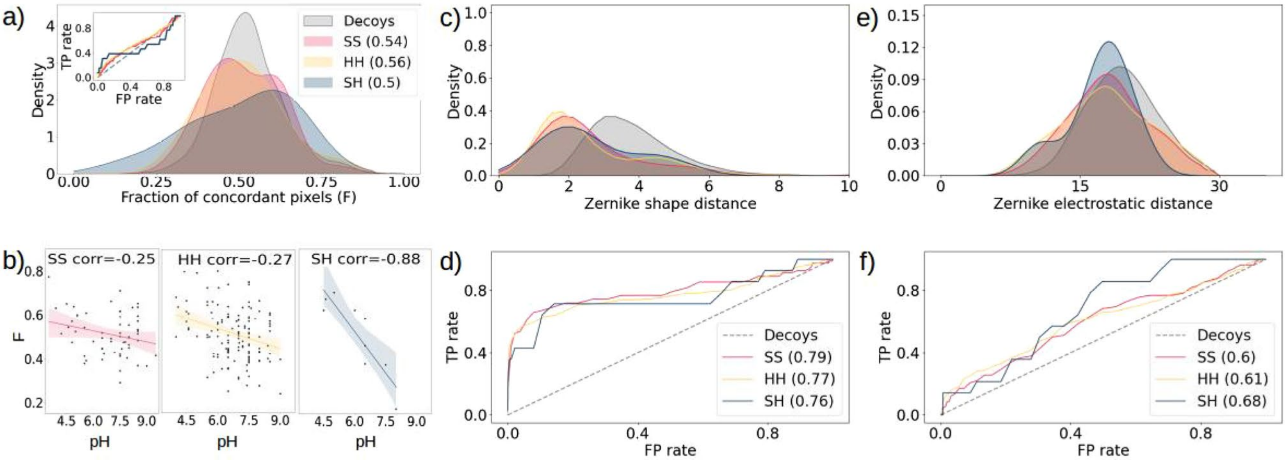

Figure 5. Electrostatic complementarity contribution in protein-protein complexes divided according to their secondary structure **(a)** Distributions of the F values of the interacting patches in complexes from the SS (red), HH (yellow) and SH (blue) classes. In the insert the corresponding ROC curves. **(b)** Fraction of concordant regions as a function of the pH and computed correlation (in the legend). From left to right the considered complexes are the SS, HH and SH complexes. **(c)** Distributions of the distances between the Zernike vectors describing the molecular surface of SS, HH and SH interacting (red, yellow and blue respectively) and random (grey) patches in the 'Human' dataset. **(d)** ROC curves of the distributions in c) and corresponding AUC (in the legend) computed against the random distribution. **(e)** For each patch the distance between the Zernike vectors describing the electrostatic potential surface in that region is computed. Then the same analysis and classification as in c) is performed. **(f)** ROC curves of the distributions in e) and corresponding AUC (in the legend) computed against the random distribution.

Discussion

The full mapping of the organisms' interactomes is fundamental for understanding molecular interactions and their many physiological and pathological implications. The well-tested toolbox of experimental techniques we dispose of, such as X-ray crystallography[49], NMR[50,51] and cryo-EM[52,53], is allowing for the detection of protein-protein binding and the determination of the complexes atomistic structure. However, all these techniques are expensive and time-consuming[54] so that up to now only small fractions of the organisms' interactomes have been experimentally determined at the structural level[55–57]. In this respect, computational methods represent a powerful tool to unveil the uncharted landscape of protein complexes[58] by predicting protein-protein associations in normal conditions and under mutations/modifications[59–62], which further complicate the compilation of the interactomes by increasing the number of matches to probe.

To predict the protein complex, the identification of putative binding interfaces plays a key role and most of the proposed strategies identify the interfaces as those showing some geometrical/chemical complementary between the molecular partners. In particular, the side chain rearrangements minimize the van der Waals interaction thus determining shape complementarity at the interfaces, which is typically evaluated by geometrical approaches requiring structural alignment between the two interacting molecules.

Notably, the geometric complementarity of the final complexes does not depend on the dynamical specifics of the binding process. In fact, partners can undergo very few changes upon binding (i.e they follow a "lock-and-key" model) or the interactions between two approaching structures can induce conformational changes ("induced fit") or the protein conformation suitable for binding (bound state) can be explored by the protein even in the absence of the molecular partner (the "conformational selection" model); these three views suggest different key contributors to the conformational changes between the unbound and bound structures, but for all of them shape complementarity is a necessary condition for the complex formation.

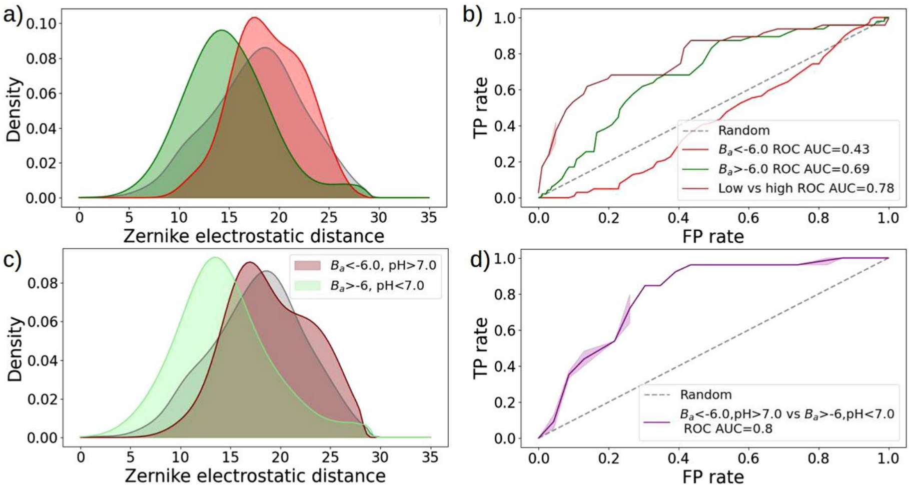

Figure 6. Superposition-free classification of transient and permanent interactions. **(a)** Probability density functions of the Zernike electrostatic distances of the 'Affinity' dataset complexes. Green (respectively red) distribution corresponds to complexes having B_a values higher (respectively lower) than -6, corresponding to transient and permanent interactions, respectively. Grey curves correspond to the distances of random decoy patches on the protein surfaces. **(b)** ROC curves of the transient (green) and permanent (red) interactions with respect to the decoy distribution, together with the ROC curve of the transient distribution with respect to the permanent one (brown). **(c)** Same as in a) but considering transient complexes with pH higher than 7 and permanent interactions with pH lower than 7, in maroon and light green respectively. **(d)** ROC curves of the two distributions displayed in panel c).

Usually, by including the electrostatic contribution to the binding process investigation, in addition to the van der Waals forces, one aims at more precise discrimination of the biological interfaces.

In this respect, computational methods can be divided into two categories: model-based and feature-based approaches. The former exploits the residue-conservation found between similar proteins, the latter is based on local features of protein sequences and/or structures. Feature-based approaches are more general and can work on any type of protein. Even if the availability of protein structures is less abundant than sequences, structural features are fundamental for understanding binding between proteins. Moreover, the recent advances in the field of protein structure prediction starting only from its amino acid sequence[63], vouch for an even more important role of the structural-based method than in the past years. Nonetheless, even using structural information, the identification of interfaces remains a challenge in structural biology. Machine learning-based approaches give promising results, but they require the definition and training of several parameters and lack a clear physical-chemical interpretation. Here, we proposed an approach for the rapid and quantitative evaluation of electrostatic complementarity and we probed its role in the identification of binding regions and complexes' stability. Analysis of the electrostatic potential of protein-protein complexes has led to the general assertion that electrostatic complementarity must be of importance at the interfaces of biological complexes[22]; nevertheless, a well-settled definition of how electrostatic complementarity should be quantified and what is its role was still missing.

With this aim, we collected two large datasets of protein dimers with known structural information stratified by dimer type, prevailing secondary structure and stability (quantified by means of experimental binding affinity). At first, we analyzed the amino acid composition of the binding region with respect to those of the proteins' cores and solvent-exposed regions. Next, we looked at the presence and disposition of the charged residues on the binding regions finding that different classes of dimers and structures have slightly different disposition/abundances of charged-charged interactions. Finally, we further increase the complexity of the electrostatic description, considering the full electrostatic potential generated by the protein partial charges on the solvent-exposed molecular surface. This representation allows for a high-level measurement of the electrostatic complementary at the interface of the interacting molecules. Indeed, comparing the spatial correspondence of the potential sign (see F descriptor), we found that the binding regions exhibit a complementary higher than the one we could expect by chance. Notably, the signal is influenced both by the complex class and the experimental pH and binding affinity. In particular, we observed that the maximum complementarity is shown by low-affinity complexes ($B_a > -6.0$), while SH complexes and homodimers sharing some residues on the binding regions (SBR-hom) exhibit a nearly random match. Finally, we propose a novel method to assess electrostatic complementarity without the need of having complex structures. Indeed, we already developed a novel computational protocol based on the Zernike polynomials to describe the shape of portions of the molecular surface in the form of a vector of numbers[16,37–40]. Here, the method is extended to molecular surfaces for which the electrostatic potential has been calculated through the Poisson-Boltzmann equation[44]. Indeed, after a proper projection of the electrostatic potential surfaces on the 2D plane, electrostatic complementarity can then be defined again as the Euclidean distance between these new Zernike invariant descriptors (see Methods).

Comparing the electrostatic complementarities at the complex interface via the Zernike method, we found that we are able to discriminate between transient and permanent interactions with an AUC of the ROC of 0.8. Interestingly, the electrostatic complementarity evaluated with the Zernike method on five mutants with known mutations at the binding sites and experimentally determined dissociation constant seems to indicate that our method is able to capture to a certain extent also the effect of point mutations on the complex binding affinity.

In conclusion, we found that electrostatic complementarity in the binding region is efficiently measured simply requiring a spatial match between the signs of the electrostatic potentials. Moreover, such complementarity strongly depends on both the kind of the considered complex, the pH of the environment, and the transient/permanent nature of the binding. In particular, we observe an evident inversely proportional relationship between electrostatic complementarity and the experimental binding affinity. Our results thus help shed light on the often contrasting conclusions of previous works that measured electrostatic complementarity using large datasets. Leveraging on our findings, we adapted the Zernike formalism to measure both shape and electrostatic complementarity in a fast and superposition-free manner. Finally, we note that our findings could be used to reinforce the docking algorithm, or/and to perform pose selection. Moreover, our method could be adapted to other properties that can be described with numerical values assigned to each surface point, since the Zernike expansion can be applied to any function.

Methods

Protein complex datasets.

To probe the degree of electrostatic complementarity in protein-protein binding regions, we collect a dataset of protein-protein dimers for which structure information was available from the 3D complex database[64]. Selecting only non-redundant human dimers, with an x-ray crystal resolution better than 3.0 Å and no missing residues in the binding region, we ended up with 199 human protein complexes in PDB format[65]. We opted to restrict to only one organism to avoid spurious effects on the charges distribution in the protein structure, due for instance to thermal adaptation[10,13,66].

Looking at the dimer composition and spatial orientation, we classify the dataset, that we call ‘Human‘ dataset, into four groups:

- 44 homodimers with Identical Binding Regions (IBR-hom), i.e. binding regions that have at least 70% of common residues.
- 66 homodimers with Shifted Binding Regions (SBR-hom), i.e. interacting patches that have between 30% and 70% of common residues.
- 54 homodimers with non-Identical Binding Regions (nIBR-hom), i.e. binding regions that share less than 30% of the residues.
- 35 heterodimers (nIBR-het), where two different proteins are interacting.

To gain more insights into the structural dependence of electrostatic complementarity, we considered a second independent classification of the same dataset. For this classification we looked at the secondary structure of each protein. Per residue secondary structure assignment was done using the DSSP[67] module implemented in Python. Proteins with a prevalence of residues associated to helices are classified as H, otherwise as S. Looking at the structural composition of the binding partners, the ’Human’ dataset is classified in three classes:

- 133 HH complexes, i.e. complexes where both partners have a prevalence of helices over strands.
- 57 SS complexes, i.e. complexes where both partners have a prevalence of strands over helices.
- 9 SH complexes, i.e. complexes where one of the partners has more helices residues and the other more strands residues.

Table SI shows the list of the PDB id of the complexes in the ’Human’ dataset, together with their dimer and structural classification. Figure S2 shows the amino acid composition and charge properties of the three classes.

To analyze the correlation between electrostatic complementarity and binding stability we consider a second dataset, composed of 123 complexes extracted from the dataset used in[10]. To our knowledge, that dataset is the largest available collection of complexes with experimental data of binding affinity B_a, defined as the log_{10} of the equilibrium dissociation constant K_d[68]. We then select the complexes with known pH and no missing atoms or residues and call the resulting collection ’Affinity’ dataset. Two of these complexes are also part of the ‘Human‘ dataset as heterodimers: 2HTH and 3MZG. The list of the complexes together with their B_a is reported in Table SII.

Computation of the surfaces and surface residues definition.

The solvent-accessible surface for each structure of the dataset was computed using DMS[69], with a density of 5 points per Å^2 and a water probe radius of 1.4 Å. For each surface point, the unit normal vector was calculated with the flag −n. Starting from these surfaces, the electrostatic potential of each protein was calculated independently from the partner using the APBS code[43], considering the experimental pH. The electrostatic potential surface was then defined by building a grid and selecting the values of the electrostatic potential in the grid cells corresponding to each surface point.

To select among the residues included in the surface the mainly superficial ones, we computed the Relative Solvent Accessibility as the ratio between Solvent Accessibility and the maximum Solvent Accessible Surface Area of the considered amino acid. The Solvent Accessibility is calculated with DMS by computing the portion exposed to the solvent of each residue involved in the interaction, while the maximum Solvent Accessible Surface

Area of the twenty natural amino acids was taken from[70]. A residue is considered superficial if it has a Relative Solvent Accessibility higher than 0.25. The interacting regions were defined as the points on a protein surface closer than 6Å to its partner surface.

Patch definition and projection. To define a surface patch, we use a spherical region with radius R centered at one point of the surface. This point is randomly extracted for the decoy random patches, while to study the binding regions the geometrical center of the experimental interacting regions is considered. For this study, we chose $R = 9$Å to be able to study simultaneously the shape and electrostatic complementarity with the 2D Zernike-based method. Indeed, in a previous work, we discussed the range of R values resulting in the best identification of binding regions when considering shape complementarity[16].

Once the patch has been selected, we re-orient the coordinates. When two random patches are compared, for each patch we build a plane passing through it and we orient the coordinates so that the z-axis is perpendicular to the plane. It must be remembered that when comparing the shape of patches, their relative orientation must be evaluated: to assess their shape complementarity, we have to orient the patches contrariwise, i.e. one patch with the solvent-exposed part toward the positive z-axis ('up') and one toward the negative z-axis ('down').

On the other hand, to compare the *EM*s and *SEM*s of interacting patches we compute the mean of the normal vector of the first partner and the inverse of the normal vector of the second one. The binding patches are then rotated so that this averaged vector is along the z-axis. This step results again in two patches contrariwise oriented, but in addition to this, we can preserve the spatial correspondence of the surface points after the rotation. We want to remark here how this correspondence is not necessary when the projections are decomposed in the Zernike basis and compared with the Zernike protocol, giving the rotation invariance of the Zernike polynomials. Therefore, to study the patches in terms of the Zernike polynomials expansion we reorient each binding site along the z-axis independently from its partner.

Once the patches have been rotated, two protocols can be implemented. The first one is used to obtain the projections of the corresponding regions of the electrostatic potential surface, whereas the second one provides the projections of the molecular surfaces.

Electrostatic projection. Each point of the re-oriented electrostatic surface is projected on the x-y plane. Next, we build a square grid (25×25 pixels) and associate each pixel with the mean value of the electrostatic potential of the points projected inside of it, and call it the Electrostatic Matrix (*EM*).

Shape projection. Once the patch has been rotated, given a point C on the z-axis we define the angle θ as the largest angle between the z-axis and a secant connecting C to any point of the patch. C is then set so that $\theta = 45°$.

To study the shape of the patch, each surface point is labeled with its distance r to C. We then build a square grid (25×25 pixels), associating each pixel with the mean r value calculated on the points inside it.

Zernike 2D protocol. Each function of two variables $f(r, \psi)$ defined in polar coordinates inside the region of the unitary circle ($r < 1$) can be decomposed in the Zernike basis as

$$f(r,\psi) = \sum_{n'=0}^{\infty} \sum_{m=0}^{n'} c_{n'm} Z_{n'm}(r,\psi), \tag{1}$$

where

$$c_{n'm} = \frac{n'+1}{\pi} \int_0^1 dr\, r \int_0^{2\pi} d\psi\, Z^*_{n'm}(r,\psi) f(r,\psi) \tag{2}$$

and

$$Z_{n'm} = R_{n'm}(r) e^{im\psi}. \tag{3}$$

$c_{n'm}$ are the expansion coefficients, while the complex functions $Z_{n'm}(r, \psi)$ are the Zernike polynomials. The radial part $R_{n'm}$ is given by

$$R_{n'm}(r) = \sum_{k=0}^{\frac{n'-m}{2}} \frac{(-1)^k (n'-k)!}{k!\left(\frac{n'+m}{2}-k\right)!\left(\frac{n'-m}{2}-k\right)!}. \tag{4}$$

Since for each couple of polynomials, it is true that

$$\langle Z_{n'm} | Z_{n''m'} \rangle = \frac{\pi}{n'+1} \delta_{n'n''} \delta_{mm'}, \tag{5}$$

the complete sets of polynomials form a basis, and knowing the set of complex coefficients $c_{n'm}$ allows for a univocal reconstruction of the original patch. The resolution of this reconstruction depends on the order of expansion $N = max(n^{'})$.

The norm of the coefficients $z_{n'm} = |c_{n'm}|$ defines the Zernike invariant descriptor, which is invariant for rotations around the origin of the unitary circle.

The complementarity between two given patches defined with a sphere of radius R can then be measured as the Euclidean distance between the two corresponding invariant vectors: the more the complementary the smaller the distance between their corresponding Zernike vectors. This evaluation can be applied to any properties of the patches that can be described by assigning a numerical value to each surface point.

The efficiency of this method depends on two key parameters: the radius R and the Zernike maximum expansion order N. When R is too low, the patches lack sufficient surface to distinguish the compatibility between interacting regions, whereas too-large patches would include non-interacting regions that have a low complementarity per se. N, on the other hand, determines the level of details captured: too low orders could confuse interacting and random patches because the surfaces are excessively "smoothed", while an excessively accurate level of description would model unnecessary (and time-consuming) details.

In this study, we performed the Zernike protocol using $R = 9$Å and $N = 20$, in accordance with the most efficient parameters identified in the previously mentioned work[16].

Code availability

All codes and relevant data are within the Main Text, and at: https://github.com/matmi8/Zernike2D.

Received: 14 February 2023; Accepted: 16 June 2023
Published online: 23 June 2023

References

1. Ryan, D. P. & Matthews, J. M. Protein-protein interactions in human disease. *Curr. Opin. Struct. Biol.* **15**(4), 441–446 (2005).
2. Rabbani, G., Baig, M. H., Ahmad, K. & Choi, I. Protein-protein interactions and their role in various diseases and their prediction techniques. *Curr. Protein Pept. Sci.* **19**(10), 948–957 (2018).
3. Gracia, P., Polanco, D., Tarancón-Díez, J., Serra, I., Bracci, M., Oroz, J., Laurents, D.V., García, I., & Cremades, N. Molecular mechanism for the synchronized electrostatic coacervation and co-aggregation of alpha-synuclein and tau. *Nat. Commun.*, **13**(1), (2022).
4. Berggård, T., Linse, S. & James, P. Methods for the detection and analysis of protein-protein interactions. *Proteomics* **7**(16), 2833–2842 (2007).
5. Sheinerman, F. B. & Honig, B. On the role of electrostatic interactions in the design of protein–protein interfaces. *J. Mol. Biol.* **318**(1), 161–177 (2002).
6. Koshland Jr, D. E. The key-lock theory and the induced fit theory. *Angew. Chem. Int. Ed. Engl.* **33**(23–24), 2375–2378 (1995).
7. Csermely, P., Palotai, R., & Nussinov, R. Induced fit, conformational selection and independent dynamic segments: an extended view of binding events. *Nat. Precedings*, pp. 1–1 (2010).
8. Paul, F. & Weikl, T. R. How to distinguish conformational selection and induced fit based on chemical relaxation rates. *PLoS Comput. Biol.* **12**(9), e1005067 (2016).
9. Gabb, H. A., Jackson, R. M. & Sternberg, M. J. E. Modelling protein docking using shape complementarity, electrostatics and biochemical information 1 1edited by j. thornton. *J. Mol. Biol.* **272**(1), 106–120 (1997).
10. Desantis, F., Miotto, M., Di Rienzo, L., Milanetti, E., & Ruocco, G. Spatial organization of hydrophobic and charged residues affects protein thermal stability and binding affinity. *Sci. Rep.*, **12**(1), (2022).
11. Skrabanek, L., Saini, H. K., Bader, G. D. & Enright, A. J. Computational prediction of protein-protein interactions. *Mol. Biotechnol.* **38**(1), 1–17 (2008).
12. Van Dan, B. *et al.* Protein stabilization by hydrophobic interactions at the surface. *Eur. J. Biochem.* **220**(3), 981–985 (1994).
13. Miotto, M. *et al.* Insights on protein thermal stability: a graph representation of molecular interactions. *Bioinformatics* **35**(15), 2569–2577 (2018).
14. Miotto, M., Di Rienzo, L., Gosti, G., Bo' Leonardo, P., Giacomo, P., Roberta, B., Alberto, R., Giancarlo & Milanetti, E. Inferring the stabilization effects of SARS-CoV-2 variants on the binding with ACE2 receptor. *Commun. Biol.*, **5**(1), (2022).
15. Erijman, A., Rosenthal, E. & Shifman, J. M. How structure defines affinity in protein-protein interactions. *PLoS ONE* **9**(10), e110085 (2014).
16. Milanetti, E. *et al.* 2d zernike polynomial expansion: Finding the protein-protein binding regions. *Comput. Struct. Biotechnol. J.* **19**, 29–36 (2021).
17. Kihara, D., Sael, L., Chikhi, R. & Esquivel-Rodriguez, J. Molecular surface representation using 3d zernike descriptors for protein shape comparison and docking. *Curr. Protein Pept. Sci.* **12**(6), 520–530 (2011).
18. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**(2), 184–192 (2020).
19. Daberdaku, S. & Ferrari, C. Antibody interface prediction with 3d zernike descriptors and svm. *Bioinformatics* **35**(11), 1870–1876 (2019).
20. Zhu, X., Xiong, Y. & Kihara, D. Large-scale binding ligand prediction by improved patch-based method patch-surfer2. 0. *Bioinformatics* **31**(5), 707–713 (2015).
21. Venkatraman, V., Yang, Y. D., Sael, L. & Kihara, D. Protein-protein docking using region-based 3d zernike descriptors. *BMC Bioinformatics* **10**(1), 1–21 (2009).
22. Bauer, M. R. & Mackey, M. D. Electrostatic complementarity as a fast and effective tool to optimize binding and selectivity of protein–ligand complexes. *J. Med. Chem.* **62**(6), 3036–3050 (2019).
23. Di Rienzo, L., Miotto, M., Bò, L., Ruocco, G., Raimondo, D., & Milanetti, E. Characterizing hydropathy of amino acid side chain in a protein environment by investigating the structural changes of water molecules network. *Front. Mol. Biosci.*, **8**, (2021).
24. Milanetti, E. *et al.* Correlation analysis based on the hydropathy properties of non-steroidal anti-inflammatory drugs in solid-phase extraction (spe) and reversed-phase high performance liquid chromatography (hplc) with photodiode array detection and their applications to biological samples. *J. Chromatogr. A* **1605**, 360351 (2019).
25. Shashikala, H.M., Chakravorty, A., & Alexov, E. Modeling electrostatic force in protein-protein recognition. *Front. Mol. Biosci.*, **6**, (2019).
26. Zhang, Z., Witham, S. & Alexov, E. On the role of electrostatics in protein–protein interactions. *Phys. Biol.* **8**(3), 035001 (2011).
27. Vascon, F. *et al.* Protein electrostatics: From computational and structural analysis to discovery of functional fingerprints and biotechnological design. *Comput. Struct. Biotechnol. J.* **18**, 1774–1789 (2020).
28. Kundrotas, P. J. & Alexov, E. Electrostatic properties of protein-protein complexes. *Biophys. J* . **91**(5), 1724–1736 (2006).
29. Tsuchiya, Y. Analyses of homo-oligomer interfaces of proteins from the complementarity of molecular surface, electrostatic potential and hydrophobicity. *Protein Eng. Des. Sel.* **19**(9), 421–429 (2006).
30. Zhou, H.-X. & Pang, X. Electrostatic interactions in protein structure, folding, binding, and condensation. *Chem. Rev.* **118**(4), 1691–1741 (2018).

31. Yoshida, K., Kuroda, D., Kiyoshi, M., Nakakido, M., Nagatoishi, S., Soga, S., Shirai, H., & Tsumoto, K. Exploring designability of electrostatic complementarity at an antigen-antibody interface directed by mutagenesis, biophysical analysis, and molecular dynamics simulations. *Sci. Rep.*, **9**(1), (2019).
32. McCoy, A. J., Chandana Epa, V. & Colman, P. M. Electrostatic complementarity at protein/protein interfaces 1 1edited by b. honig. *J. Mol. Biol.* **268**(2), 570–584 (1997).
33. Ghaemi, Z., Guzman, I., Gnutt, D., Luthey-Schulten, Z. & Gruebele, M. Role of electrostatics in protein–RNA binding: The global vs the local energy landscape. *J. Phys. Chem. B* **121**(36), 8437–8446 (2017).
34. McCoy, A. J., Chandana Epa, V. & Colman, P. M. Electrostatic complementarity at protein/protein interfaces. *J. Mol. Biol.* **268**(2), 570–584 (1997).
35. Miotto, M. *et al.* Thermometer: a webserver to predict protein thermal stability. *Bioinformatics* **38**(7), 2060–2061 (2022).
36. Maleki, M., Vasudev, G. & Rueda, L. The role of electrostatic energy in prediction of obligate protein-protein interactions. *Proteome Sci.* **11**(1), 1–12 (2013).
37. Milanetti, E., Miotto, M., Di Rienzo, L., Nagaraj, M., Monti, M., Golbek, T.W., Gosti, G., Roeters, S.J., Weidner, T., Otzen, D.E. & Ruocco, G. In-silico evidence for a two receptor based strategy of SARS-CoV-2. *Front. Mol. Biosci.*, **8**, (2021).
38. Miotto, M., Di Rienzo, D., Bò, L., Boffi, A., Ruocco, G., & Milanetti, E. Molecular mechanisms behind anti SARS-CoV-2 action of lactoferrin. *Front. Mol. Biosci.*, **8**, (2021).
39. Bò, L., Miotto, M., Di Rienzo, L., Milanetti, E. & Ruocco, G. Exploring the association between sialic acid and sars-cov-2 spike protein through a molecular dynamics-based approach. *Front. Med. Technol.* **2**, 24 (2020).
40. Grassmann, G., Miotto, M. , Di Rienzo, L., Salaris, F., Silvestri, B., Zacco, E., Rosa, A., Tartaglia, G.G., Ruocco, G., & Milanetti, E. A computational approach to investigate tdp-43 rna-recognition motif 2 c-terminal fragments aggregation in amyotrophic lateral sclerosis. *Biomolecules*, **11**(12), (2021).
41. Grassmann, G. *et al.* A novel computational strategy for defining the minimal protein molecular surface representation. *PLoS ONE* **17**(4), e0266004 (2022).
42. Yan, C., Feihong, W., Jernigan, R. L., Dobbs, D. & Honavar, V. Characterization of protein-protein interfaces. *Protein J.* **27**(1), 59–70 (2008).
43. Jurrus, E. *et al.* Improvements to the apbs biomolecular solvation software suite. *Protein Sci.* **27**(1), 112–128 (2017).
44. Chakavorty, A., Li, L. & Alexov, E. Electrostatic component of binding energy: Interpreting predictions from poisson-boltzmann equation and modeling protocols. *J. Comput. Chem.* **37**(28), 2495–2507 (2016).
45. Gainza, P. *et al.* Deciphering interaction fingerprints from protein molecular surfaces using geometric deep learning. *Nat. Methods* **17**(2), 184–192 (2019).
46. Cheng, P.-N., Pham, J. D. & Nowick, J. S. The supramolecular chemistry of β-sheets. *J. Am. Chem. Soc.* **135**(15), 5477–5492 (2013).
47. Nooren, I. M. A. & Thornton, J. M. Diversity of protein-protein interactions. *EMBO J.* **22**(14), 3486–3492 (2003).
48. La, D., Kong, M., Hoffman, W., Choi, Y. I. & Kihara, D. Predicting permanent and transient protein-protein interfaces. *Proteins* **81**(5), 805–818 (2013).
49. Laurent, M. & Lionel, M. Protein x-ray crystallography and drug discovery. *Molecules* **25**(5), 1030 (2020).
50. Takahashi, H., Nakanishi, T., Kami, K., Arata, Y. & Shimada, I. A novel nmr method for determining the interfaces of large protein-protein complexes. *Nat. Struct. Biol.* **7**(3), 220–223 (2000).
51. Foster, M. P. *et al.* Chemical shift as a probe of molecular interfaces: Nmr studies of dna binding by the three amino-terminal zinc finger domains from transcription factor iiia. *J. Biomol. NMR* **12**(1), 51–71 (1998).
52. Bai, X.-C., McMullan, G. & Scheres, S. H. W. How cryo-em is revolutionizing structural biology. *Trends Biochem. Sci.* **40**(1), 49–57 (2015).
53. Cheng, Y. Single-particle cryo-em at crystallographic resolution. *Cell* **161**(3), 450–457 (2015).
54. Berggård, T., Linse, S. & James, P. Methods for the detection and analysis of protein-protein interactions. *Proteomics* **7**(16), 2833–2842 (2007).
55. Haibin, G., Zhu, P., Jiao, Y., Meng, Y. & Chen, M. Prin: a predicted rice interactome network. *BMC Bioinformatics* **12**(1), 1–13 (2011).
56. Plewczyński, D. & Ginalski, K. The interactome: predicting the protein-protein interactions in cells. *Cell. Mol. Biol. Lett.* **14**(1), 1–22 (2009).
57. Li, S. *et al.* A map of the interactome network of the metazoan c. elegans. *Science* **303**(5657), 540–543 (2004).
58. Zhang, Q. C. *et al.* Structure-based prediction of protein-protein interactions on a genome-wide scale. *Nature* **490**(7421), 556–560 (2012).
59. Ezkurdia, I. *et al.* Progress and challenges in predicting protein-protein interaction sites. *Brief. Bioinform.* **10**(3), 233–246 (2009).
60. Lichtarge, O., Bourne, H. R. & Cohen, F. E. An evolutionary trace method defines binding surfaces common to protein families. *J. Mol. Biol.* **257**(2), 342–358 (1996).
61. Wang, B. *et al.* Predicting protein interaction sites from residue spatial sequence profile and evolution rate. *FEBS Lett.* **580**(2), 380–384 (2006).
62. Brender, J. R. & Zhang, Y. Predicting the effect of mutations on protein-protein binding interactions through structure-based interface profiles. *PLoS Comput. Biol.* **11**(10), e1004494 (2015).
63. Varadi, M. *et al.* AlphaFold protein structure database: massively expanding the structural coverage of protein-sequence space with high-accuracy models. *Nucleic Acids Res.* **50**(D1), D439–D444 (2021).
64. Levy, E. D., Pereira-Leal, J. B., Chothia, C. & Teichmann, S. A. 3d complex: A structural classification of protein complexes. *PLoS Comput. Biol.* **2**(11), e155 (2006).
65. Berman, H.M., Bhat, T. N., Bourne, P.E., Feng, Z., Gilliland, G., Weissig, H., & Westbrook, J. *Nature Structural Biology*, **7**, 957–959, (2000).
66. Miotto, M. *et al.* Simulated epidemics in 3d protein structures to detect functional properties. *J. Chem. Inf. Model.* **60**(3), 1884–1891 (2020).
67. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolym. Original Res. Biomol.* **22**(12), 2577–2637 (1983).
68. Vangone, A., & Bonvin, A.M.J.J. Contacts-based prediction of binding affinity in protein–protein complexes. *elife*, **4**, (2015).
69. Richards, F. M. Areas, volumes, packing and protein structure. *Annu. Rev. Biophys. Bioeng.* **6**, 151–76 (1977).
70. Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J. & Wilke, C. O. Maximum allowed solvent accessibilites of residues in proteins. *PLoS ONE* **8**(11), e80635 (2013).

Acknowledgements

The research leading to these results was also supported by European Research Council through its Synergy grant programme, project ASTRA (grant agreement No 855923) and by European Innovation Council through its Pathfinder Open Programme, project ivBM-4PAP (grant agreement No 101098989).

Author contributions

E.M and M.M. conceived and supervisioned the work. G.Grassmann performed computational analysis. L.D.R. and G.Gosti collected the dataset and performed additional analysis. M.L. and G.R. contributed with additional ideas. G.Grassmann, M.M and E.M. wrote the manuscript. All authors revised the work.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1038/s41598-023-37130-z.

Correspondence and requests for materials should be addressed to M.M. or E.M.

Reprints and permissions information is available at www.nature.com/reprints.