Title

Protein interaction energy landscapes are shaped by functional and also non-functional partners

Authors

Hugo Schweke[1], Marie-Hélène Mucchielli[1,2], Sophie Sacquin-Mora[3], Wanying Bei[1], Anne Lopes[1]

Authors affiliations

[1] Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 91198, Gif-sur-Yvette cedex, France

[2] Sorbonne Universités, UPMC Univ Paris 06, UFR927, F-75005 Paris, France.

[3] Laboratoire de Biochimie Théorique, UPR 9080 CNRS Institut de Biologie Physico-Chimique, Paris, France

Author contributions

HS performed research

HS, MHM, SSM, WB, AL analyzed data

HS, MHM, AL designed research

HS, MHM, AL wrote the paper

AL conceived the project

The authors declare no conflict of interest.

Corresponding author

Anne Lopes, Institute for Integrative Biology of the Cell (I2BC), CEA, CNRS, Univ. Paris-Sud, Université Paris-Saclay, 1 avenue de la Terrasse, 91198 Gif-sur-Yvette, France.

Tel: +33 (0)1 69 15 35 60

email: anne.lopes@u-psud.fr

The authors have declared that no competing interests exist.

Abstract

In the crowded cell, a strong selective pressure operates on the proteome to limit the competition between functional and non-functional protein-protein interactions. We developed an original theoretical framework in order to interrogate how this competition constrains the behavior of proteins with respect to their partners or random encounters. Our theoretical framework relies on a two-dimensional (2D) representation of interaction energy landscapes with 2D energy maps that reflect in a synthetic way the propensity of a protein to interact with another protein. We investigated the propensity of protein surfaces to interact with functional and arbitrary partners and asked whether their interaction propensity is conserved during the evolution. Therefore, we performed several thousands of cross-docking simulations to systematically characterize the whole energy landscapes of 74 proteins interacting with different sets of homologs, corresponding to their functional partner's family or arbitrary protein families. Then, we systematically compared the energy maps resulting from the docking of a given protein with the different protein families of the dataset. Strikingly, we show that the interaction propensity not only of the binding site but also of the rest of the protein surface is conserved for docking partners belonging to the same protein family. Interestingly, this observation holds for docked proteins corresponding to true but also to arbitrary partners. Our theoretical framework enables the characterization of the energy behavior of a protein in interaction with hundreds of selected partners and opens the way for further developments to study the behavior of proteins in a specific environment.

Introduction

Biomolecular interactions are central for many physiological processes and are of utmost importance for the functioning of the cell. Particularly protein-protein interactions have attracted a wealth of studies these last decades [1–5]. The concentration of proteins in a cell has been estimated to be approximately 2-4 million proteins per cubic micron [6]. In such a highly crowded environment, proteins constantly encounter each other and numerous non-specific interactions are likely to occur [7–10]. For example, in the cytosol of *S. cerevisiae* a protein can encounter up to 2000 different proteins [11]. In this complex jigsaw puzzle, each protein has evolved to bind the right piece(s) in the right way (positive design) and to prevent misassembly and non-functional interactions (negative design) [12–16]).

Consequently, positive design constrains the physico-chemical properties and the evolution of protein-protein interfaces. Indeed, a strong selection pressure operates on binding sites to maintain the functional assembly including the functional partner and the functional binding mode. For example, homologs sharing at least 30% sequence identity almost invariably interact in the same way [17]. Conversely, negative design prevents proteins to be trapped in the numerous competing non-functional interactions inherent to the crowded environment of the cell. Many studies were reported on the relationship between the propensity of proteins for promiscuous interactions and their abundances or surface properties [18–21]. Particularly, it has been shown that the misinteraction avoidance shapes the evolution and physico-chemical properties of abundant proteins, resulting in a slower evolution and less sticky surfaces than what is observed for less abundant ones [18,22–26]. The whole surface of abundant proteins is thus constrained, preventing them to engage deleterious non-specific interactions that could be of dramatic impact for the cell at high concentration [25]. Recently, it has been shown in *E. coli* that the net charge as well as the charge distribution on protein surfaces affect the

diffusion coefficients of proteins in the cytoplasm [19,27]. Positively charged proteins move up to 100 times more slowly as they get caught in non-specific interactions with ribosomes which are negatively charged and therefore, shape the composition of the cytoplasmic proteome [27].

All these studies show that both positive and negative design effectively operate on the whole protein surface. Binding sites are constrained to maintain functional assemblies (i.e. functional binding modes and functional partners) while the rest of the surface is constrained to avoid non-functional assemblies. Consequently, these constraints should shape the energy landscapes of functional but also non-functional interactions so that non-functional interactions do not prevail over functional ones. This should have consequences (i) on the evolution of the propensity of a protein to interact with its environment (including functional and non-functional partners) and (ii) on the evolution of the interaction propensity of the whole surface of proteins, non-interacting surfaces being in constant competition with functional binding sites. We can hypothesize that the interaction propensity of the whole surface of proteins is constrained during evolution in order to (i) ensure that proteins correctly bind functional partners**,** and (ii) limit non-functional assemblies as well as interactions with non-functional partners.

In this work, we focus on protein surfaces as a proxy for functional and non-functional protein-protein interactions. We investigate their interaction energy landscapes with native and non-native partners and ask whether their interaction propensity is conserved during evolution. With this aim in mind, we performed large-scale docking simulations to characterize interactions involving either native or native-related (i.e. partners of their homologs) partners or arbitrary partners. Docking simulations enable the characterization of all possible interactions involving either functional or arbitrary partners, and thus to simulate

the interaction of arbitrary partners which is very difficult to address with experimental approaches. Docking algorithms are now fast enough for large-scale applications and allow for the characterization of interaction energy landscapes for thousand of protein couples. Typically, a docking simulation takes from a few minutes to a couple of hours on modern processors [28–30], opening the way for extensive cross-docking experiments [31–35]. Protein docking enables the exploration of the interaction propensity of the whole protein surface by simulating alternative binding modes. Here, we performed a cross-docking experiment involving 74 selected proteins docked with their native-related partners and their corresponding homologs, as well as arbitrary partners and their corresponding homologs. We represented the interaction energy landscapes resulting from each docking calculation with a two dimensional (2D) energy map in order to (i) characterize the propensity of all surface regions of a protein to interact with a given partner (either native-related or not) and (ii) easily compare the energy maps resulting from the docking of a same protein with different sets of homologous partners, thus addressing the evolution of the propensity of a protein to interact with homologous partners either native or arbitrary.

Results

The interaction propensity of the whole surface of the human ubiquitin carboxyl-terminal hydrolase 14 is conserved for homologous protein ligands, be they functional partners or random encounters

If positive and negative design constraint the propensity of the whole surface of proteins to interact with their functional partners or random encounters, this should shape the evolution of interaction energy landscapes of functional protein pairs but also of random encounter pairs. Consequently, we expect that the interaction energy landscape involving a protein pair (functional or arbitrary) is conserved for a homologous pair. Testing this hypothesis involves being able to characterize the interaction propensity of the whole surface of a protein. Therefore we designed a procedure based on a two-dimensional (2D) representation of docking energy landscapes with 2D energy maps which reflect the propensity of a protein (i.e. the receptor) to interact with the docked partner (i.e. the ligand) (*Materials and Methods*, Fig 1A-C). The procedure is asymmetrical and the resulting energy map provides the distribution of all docking energies over the whole receptor surface thus reflecting the propensity of the receptor to interact with the docked ligand. Fig 2 represents the energy maps computed for the receptor 2AYN_A, the human ubiquitin carboxyl-terminal hydrolase 14 (family UCH) docked with (i) its native partner (1XD3_B, ubiquitin-related family), a homolog of its partner (defined as a native-related partner) (1NDD_B) and (ii) two arbitrary homologous ligands (1YVB_A and 1NQD_B from the papain-like family). For all four ligands, either native-related or arbitrary partners, docking calculations lead to an accumulation of low-energy solutions (hot regions in red) around the two experimentally known binding sites of the receptor. The first one corresponds to the interaction site with the native partner, ubiquitin

(pdb id 2ayo). The second one corresponds to its homodimerisation site (pdb id 2ayn). This indicates that native-related but also arbitrary partners tend to bind onto the native binding sites of native partners as observed in earlier studies [34,36]. Indeed, the low energy solutions tend to accumulate systematically in the vicinity of the two native interaction sites. Whereas the low energy solutions obtained for both ligand families accumulate around the native binding sites of 2AYN_A, the two ligand families display clear differences in the rest of the map. Indeed, the energy maps obtained with the ligands of the ubiquitin-like family both reveal two sharp hot regions around the native sites and a subset of well-defined cold regions (i.e. blue regions corresponding to high energy solutions) placed in the same area in the map's upper-right quadrant. In contrast, the energy maps obtained for the ligands of the papain-like family display a large hot region around the two native binding sites of the receptor, extending to the upper-left and bottom-right regions of the map and suggesting a large promiscuous binding region for these ligands. The interaction propensity of the two binding sites of 2AYN_A but also of the other regions of its surface seems to be conserved for homologous ligands and specific to each ligand family whether the ligands correspond to native-related partners or not (Fig 2).

Generalization to a large set of proteins

We asked whether this observation could be generalized to a large set of proteins. Therefore we built a database comprising 74 protein structures divided into 12 families of homologs (S1 Table and *Materials and Methods*). Each family displays different degrees of structural variability and sequence divergence in order to see the impact of these properties on the conservation of the interaction propensity inside a protein family. Each family has at least one native-related partner family (S1 Fig). For a protein A, we refer as native-related partners its native partner (when its three dimensional (3D) structure is available) and native partners of

proteins that are homologous to the protein A. Arbitrary pairs refer to pairs of proteins for which no interaction has been experimentally characterized in the Protein Data Bank neither for their respective homologs [37]. Docking calculations are performed with the ATTRACT software [30]. Each protein (namely the receptor) is docked with the 74 proteins (namely the ligands) of the dataset (Fig 3A and *Materials and Methods*) and the 74 corresponding energy maps are calculated (Fig 3B and *Materials and Methods*). The 74 resulting energy maps are compared together with a Manhattan distance and all the energy map distances are stored in an energy map distance (EMD) matrix (Fig 3C and *Materials and Methods*). Each matrix entry (*i*,*j*) corresponds to the distance $d_{i,j}$ between the energy maps of ligands *i* and *j* docked with a receptor *k* (Fig 3C and *Materials and Methods*). Consequently, a small distance $d_{i,j}$ between ligands *i* and *j* docked with a receptor *k*, reflects a high similarity of their energy maps. In other words, the interaction propensity of the surface of the receptor *k* is similar for both ligands *i* and *j*. One should notice that energy maps computed for two unrelated receptors are not comparable since their surfaces are not comparable as well. Therefore, the procedure is asymmetrical and receptor-centered. It only compares energy maps calculated for different ligands docked with the same receptor. In order to prevent any bias from the choice of the receptor, each of the 74 proteins plays alternately the role of receptor and ligand. Consequently, the protocol presented in Fig 3 is repeated for the entire dataset where each protein plays the role of the receptor and is docked with the 74 proteins that play the role of ligands, thus resulting in 74 EMD matrices. In order to quantify the extent to which the interaction propensity of a receptor is conserved for homologous ligands, we evaluated to what extent distances calculated between homologous ligand pairs were smaller than distances calculated between random pairs. Fig 4 represents the boxplots of energy map distances calculated between random ligand pairs or between homologous ligand pairs docked with their native-related receptor or with the other receptors of the dataset. Homologous

194 ligands display significantly lower energy map distances than random ligand pairs (Wilcoxon
195 test $p = 0$) indicating that energy maps produced by homologous ligands are more similar than
196 those produced by non-homologous ligands. Interestingly, this observation holds whether the
197 receptor-ligand pair is a native pair or not. This suggests that the interaction propensity of a
198 receptor is conserved for homologous partners be they native-related or not.

199

200 **Energy maps are specific to protein families**

201 The results presented above prompted us to assess the extent to which the interaction
202 propensity of a receptor is specific to the ligand families it interacts with. If so, a receptor
203 should lead to energy maps that are specific to the different ligand families and we should be
204 able to retrieve homology relationships of ligands solely from the comparison of their energy
205 maps. Therefore, we tested our ability to predict the homologs of a given ligand based only on
206 the comparison of its energy maps with those of the other ligands. In order to prevent any bias
207 from the choice of the receptor, the 74 EMD matrices are averaged in an averaged distances
208 matrix (ADM) (see *Materials and Methods*). Each entry *(i,j)* of the ADM corresponds to the
209 averaged distance between two sets of 74 energy maps produced by two ligands *i* and *j*. A low
210 distance indicates that the two ligands display similar energy maps whatever the receptor is.
211 We computed a receiver operating characteristic (ROC) curve from the ADM (see *Materials*
212 *and Methods*) which evaluates our capacity to discriminate the homologs of a given ligand
213 from non-homologous ligands by comparing their respective energy maps computed with all
214 74 receptors of the dataset. The true positive set consists in the homologous protein pairs
215 while the true negative set consists in any homology-unrelated protein pair. The resulting
216 Area Under the Curve (AUC) is equal to 0.79 (Fig 5). We evaluated the robustness of the
217 ligand's homologs prediction depending on the size of the receptor subset with a bootstrap

procedure by randomly removing receptor subsets of different sizes (from 1 to 73 receptors). The resulting AUCs range from 0.77 to 0.79, and show that from a subset size of five receptors, the resulting prediction accuracy no longer significantly varies (risk of wrongly rejecting the equality of two variances (F-test) >5%), and is robust to the nature of the receptor subset (S2 Fig). Finally, we evaluated the robustness of the predictions according to the number of grid cells composing the energy maps. Therefore, we repeated the procedure using energy maps with resolutions ranging from 144x72 to 48x24 cells. S2 Table presents the AUCs calculated with different grid resolutions. The resulting AUCs range from 0.78 to 0.8 showing that the grid resolution has a weak influence on the map comparison. All together, these results indicate that homology relationships between protein ligands can be detected solely on the basis of the comparison of their energy maps. In other words, the energy maps calculated for a receptor docked with a set of ligands belonging to a same family are specific to this family. Interestingly, this observation holds for families displaying important sequence variations (S1 Table). For example, the AUC computed for the UCH and ubiquitin-related families are 0.98 and 0.88 respectively despite the fact that the average sequence identity of these families does not exceed 45% (S3 Fig and S1 Table). This indicates that energy maps are similar even for homologous ligands displaying large sequence variations.

We then specifically investigated the energy maps of each family in order to see whether some ligands behave energetically differently from their family members. On the 74 ligands, only five (2L7R_A, 4BNR_A, 1BZX_A, 1QA9_A, 1YAL_B) display energy maps that are significantly different from those of their related homologs (Z-tests *p-values* for the comparison of the averaged distance of each ligand with their homologs versus the averaged

distance of all ligands with their homologous ligands ≤ 5%). In order to identify the factors leading to differences between energy maps involving homologous ligands, we computed the pairwise sequence identity and the root mean square deviation (RMSD) between the members of each family. Interestingly, none of these criteria can explain the energy map differences observed within families (Fisher test p of the linear model estimated on all protein families >0.1) (see Fig 6B-C for the ubiquitin-related family, S4-S14B-C Fig for the other families, and S3 Table for details). Fig 6A represents a subsection of the ADM for the ubiquitin-related family (i.e. the energy map distances computed between all the members of the ubiquitin-like family and averaged over the 74 receptors). Low distances reflect pairs of ligands with similar energy behaviors (i.e. producing similar energy maps when interacting with a same receptor) while high distances reveal pairs of ligands with different energy behaviors. 2L7R_A distinguishes itself from the rest of the family, displaying high-energy map distances with all of its homologs. RMSD and sequence identity contribute modestly to the energy map distances observed in Fig 6A (Spearman correlation test $p^{RMSD} = 0.01$ and $p^{seq} = 0.02$ (S3 Table, Fig 6B-C)). Fig 6D shows a projection of the electrostatic potential calculated with APBS [38] on the surface of the seven ubiquitin-related family members (for more details, see S15 Fig and *Materials and Methods*). Fig 6E represents the electrostatic maps distances computed between all members of the family. 2L7R_A clearly stands out, displaying a negative electrostatic potential over the whole surface while its homologs harbor a remarkable fifty-fifty electrostatic distribution (Fig 6D). The negatively charged surface of 2L7R_A is explained by the absence of the numerous lysines that are present in the others members of the family (referred by black stars, Fig 6D). Lysines are known to be essential for ubiquitin function, enabling the formation of polyubiquitin chains on target proteins. Among the seven lysines of the ubiquitin, K63 polyubiquitin chains are known to act in non-proteolytic events while K48, K11, and the four other lysines polyubiquitin chains are presumed to be involved

into addressing proteins to the proteasome [39]. 2L7R_A is a soluble UBL domain resulting from the cleavage of the fusion protein FAU [40]. Its function is unrelated to proteasomal degradation, which might explain the lack of lysines on its surface and the differences observed in its energy maps. Interestingly, the differences observed for the energy maps of 1YAL_B (Papain-like family) (S4 Fig) and 4BNR_A (eukaryotic proteases family) (S5 Fig) regarding their related homologs can be explained by the fact that they both display a highly charged surface. These two proteins are thermostable [41,42], which is not the case for their related homologs, and probably explains the differences observed in their relative energy maps. The V-set domain family is split into two major subgroups according to their averaged energy map distances (S6A Fig). The first group corresponds to CD2 proteins (1QA9_A and its unbound form 1HNF_A) and differs significantly from the second group (Z-test $p = 0.03$ and $p = 0.05$ respectively). The second group corresponds to CD58 (1QA9_B and its unbound form 1CCZ_A) and CD48 proteins (2PTT_A). Interestingly, CD2 is known to interact with its homologs (namely CD58 and CD48) through an interface with a striking electrostatic complementarity [43]. The two subgroups have thus evolved distinct and specific binding sites to interact together. We can hypothesize that they have different interaction propensities resulting in the differences observed between their corresponding energy maps. These five cases illustrate the capacity of our theoretical framework to reveal functional or biophysical specificities of homologous proteins that could not be revealed by classical descriptors such as RMSD or sequence identity.

The AUC of 0.79 calculated previously with energy maps produced with the docking of either native-related or arbitrary pairs indicates that energy maps are specific to ligand families. To see whether this observation is not mainly due to the native-related pairs, we repeated the previous test while removing that time all energy maps computed with native-related pairs and calculated the resulting ADM. We then measured our ability to retrieve the homologs of

each ligand by calculating the ROC curve as previously. The resulting AUC is still equal to 0.79, revealing that our ability to identify a ligand's homologs is independent from the fact that the corresponding energy maps were computed with native-related or arbitrary pairs (Fig 5). This shows that the energy maps are specific to protein families whether the docked pairs are native-related or not. Consequently, the propensity of the whole protein surface to interact with a given ligand is conserved and specific to the ligand family whether the ligand is native-related or not. This striking result may reflect both positive and negative design operating on protein surfaces to maintain functional interactions and to limit random interactions that are inherent to a crowded environment.

The interaction propensity of all surface regions of a receptor is evolutionary conserved for homologous ligands

To see whether some regions contribute more to the specificity of the maps produced by homologous ligands, we next dissected the effective contribution of the surface regions of the receptor defined according to their docking energy value, in the identification of ligand's homologs. We discretized the energy values of each energy map into five categories, leading to a palette of five energy classes (see Fig 1D and *Materials and Methods*). These five-classes maps highlight low-energy regions (i.e. hot regions in red), intermediate-energy regions (i.e. warm, lukewarm and cool regions in orange, light-green and dark-green respectively) and high-energy regions (i.e. cold regions in blue). We first checked that the discretization of the energy maps does not affect our ability to identify the homologs of each of the 74 ligands from the comparison of their five-classes maps. The resulting AUC is 0.77 (Table 1)**,** showing that the discretization step does not lead to an important loss of information.

315

316 Then, we evaluated the contribution of each of the five energy classes separately in the
317 ligand's homologs identification by testing our ability to retrieve the homologs of the 74
318 ligands from their one-class energy maps (either hot, warm, lukewarm, cool or cold) (see
319 *Materials and Methods*). Table 1 shows the resulting AUCs. Interestingly, the information
320 provided by each energy class taken separately is sufficient for discriminating the homologs
321 of a given ligand from the rest of the dataset (Table 1). The resulting AUCs range from 0.76
322 to 0.79 for the warm, lukewarm, cool and cold classes and are comparable to those obtained
323 with all classes taken together (0.77). This shows (i) that warm, lukewarm, cool, and cold
324 regions alone are sufficient to retrieve homology relationships between ligands and (ii) that
325 the localization on the receptor surface of a given energy class is specific to the ligand
326 families. Hot regions are less discriminative and lead to an AUC of 0.73. In order to see how
327 regions of an energy class are distributed over a receptor surface, we summed the one-class
328 maps of the corresponding energy class calculated for this receptor into a stacked map (S16
329 Fig – see *Materials and Methods* for more details). A stacked map reflects the tendency of a
330 surface region (i.e. map cells) to belong to the corresponding energy class. Fig 7 shows an
331 example of the five stacked maps (i.e. for cold, cool, lukewarm, warm and hot regions)
332 computed for the receptor 1P9D_U. Intermediates regions (i.e. warm, lukewarm and cool
333 regions) are widespread on the stacked map while cold and hot regions are localized on few
334 small spots (three and one respectively) no matter the nature of the ligand. S17 Fig shows for
335 the receptor 1P9D_U the 12 cold and hot stacked maps computed for each ligand family
336 separately. We can see that some cold spots are specific to ligand families and that their area
337 distribution is specific to families while all 12 ligand families display the same hot spot in the
338 map's upper-right quadrant. These observations can be generalized to each receptor. On
339 average, intermediate regions are widespread on the stacked maps and cover respectively 744,

1164 and 631 cells for cool, lukewarm and warm regions, while cold and hot regions cover no more than respectively 104 and 110 cells respectively (S18 Fig). Interestingly, hot regions are more colocalized than cold ones and are restricted to 2 distinct spots on average per stacked map**,** while cold regions are spread on 3.7 spots on average (t-Test p = 7.42e-13). These results show that ligands belonging to different families tend to dock preferentially on the same regions and thus lead to similar hot region distributions on the receptor surface. This observation recalls those made by *Fernandez-Recio et al*. [36], who showed that docking random proteins against a single receptor leads to an accumulation of low-energy solutions around the native interaction site and who suggested that different ligands will bind preferentially on the same localization.

We can hypothesize that hot regions present universal structural and biochemical features that make them more prone to interact with other proteins. To test this hypothesis, we computed for each protein of the dataset, the 2D projection of three protein surface descriptors (see *Materials and Methods* and S15 Fig): the Kyte-Doolittle (KD) hydrophobicity [44], the circular variance (CV) [45] and the stickiness [25]. The CV measures the density of protein around an atom and is a useful descriptor to reflect the local geometry of a surface region. CV values are comprised between 0 and 1. Low values reflect protruding residues and high values indicate residues located in cavities. Stickiness reflects the propensity of amino acids to be involved in protein-protein interfaces [25]. It is calculated as the log ratio of the residues frequencies on protein surfaces versus their frequencies in protein-protein interfaces. For each receptor, we calculated the correlation between the docking energy and the stickiness, hydrophobicity or CV over all cells of the corresponding 2D maps. We found a significant anti-correlation between the docking energy and these three descriptors (correlation test p

364 between docking energies and respectively stickiness, hydrophobicity and CV < 2.2e-16, see
365 S4 Table)). Fig 8 represents the boxplots of the stickiness, hydrophobicity and CV of each
366 energy class (see S15 Fig and *Materials and Methods* section for more details). We observe a
367 clear effect of these factors on the docking energy: cold regions are the less sticky, the less
368 hydrophobic and the most protruding while hot ones are the most sticky, the most
369 hydrophobic and the most planar (Tukey HSD test [46], *p* of the differences observed
370 between each energy classes < 2.2e-16). One should notice that stickiness has been defined
371 from a statistical analysis performed on experimentally characterized protein interfaces and
372 therefore between presumed native partners. The fact that docking energies (physics-based)
373 calculated either between native-related or arbitrary partners is anti-correlated with stickiness
374 (statistics-based) defined from native interfaces, strengthens strongly the concept of stickiness
375 as the propensity of interacting promiscuously and provides physics-based pieces of evidence
376 for sticky regions as a proxy for promiscuous interactions.

377 We show that not only the area distribution on a receptor surface of hot regions but also those
378 of intermediate and cold regions are similar for homologous ligands and are specific to ligand
379 families (AUC ranging from 0.73 to 0.79) whether the ligands are native-related or not. This
380 tendency is even stronger for intermediate and cold regions. Interestingly, the information
381 contained in the cold regions that cover on average no more than 5.0% of the energy maps is
382 sufficient to identify homology relationships between ligands.

Discussion

In this study, we address the impact of both positive and negative design on thousands of interaction energy landscapes by the mean of a synthetic and efficient representation of the docking energy landscapes: two-dimensional energy maps that reflect the interaction propensity of the whole surface of a protein (namely the receptor) with a given partner (namely the ligand). We show that the distribution on the protein surface of all regions, including cold, intermediate and hot regions are similar for homologous ligands and are specific to ligand families whether the ligands are native-related or arbitrary. This reveals that the interaction propensity of the whole surface of proteins is constrained by functional and non-functional interactions, reflecting both positive and negative design operating on the whole surface of proteins, thus shaping the interaction energy landscapes of functional partners and random encounters. These observations were made on a dataset of 74 protein structures belonging to 12 families of structural homologs. 54 out of the 74 proteins of the dataset have at least one known partner in the dataset. For the 20 remaining proteins, we were not able to find evidences that they indeed interact with a protein of the dataset. However, we showed that the interaction propensity of a receptor is conserved for homologous ligands independently from the fact that these ligands correspond to native partners or not. Indeed, we showed that ligand homology relationships could be retrieved from their energy maps whether the maps were computed with native-related pairs or not (the corresponding AUCs calculated with and without native pairs both equal to 0.79).

Most studies that aim at depicting protein interactions focus on the functional ones and on the characterization of the native assembly modes [14,47–51]. Nevertheless, the importance of non-specific interactions and non-native assembly modes in protein interactions is no longer in doubt [7,19,21,27,52–55]. Experimental and *in-silico* studies showed the impact of non-

specific interactions on the in-cell mobility of proteins [7,19,21,27]. In addition, an important literature describes the relationship between the physico-chemical properties of proteins and their ability for non-specific interactions [7,19,21,25,53]. In particular, Wang *et al* showed that the propensity for non-specific interactions is determined by multiple factors such as the protein charge, the conformational flexibility and the distribution of hydrophobic residues on the protein surface [19]. Finally, recent studies have demonstrated the importance of non-native assembly modes and non-interacting regions in the protein association process [54] and showed that it is relevant to consider them for predicting protein partners and binding affinities [56,57]. Particularly, Marin-Lopez *et al* developed a method based on the sampling of the conformational space of the encounter complexes formed during the binding process and showed that ΔG can be predicted accurately from the scoring of all encounter complexes sampled during a docking simulation, suggesting that the knowledge of the native pose is not necessary [57]. All these works highlight the importance of taking into account the whole surface of proteins as well as all the binding modes of a protein pair. This calls for the development of new methods that enable the systematic and physical characterization of the whole surface of a protein in interaction with a given partner. Here, we address the energy behavior of not only known binding sites, but also of the rest of the protein surface, which plays an important role in protein interactions by constantly competing with the native binding site. We show that the interaction propensity of the rest of the surface is not homogeneous and displays regions with different binding energies that are specific to ligand families. This may reflect the negative design operating on these regions to limit non-functional interactions [14,16,58]. We can hypothesize that non-interacting regions participate to favor functional assemblies (i.e. functional assembly modes with functional partners) over non-functional ones and are thus evolutionary constrained by non-functional assemblies. The fact that cold regions seem to be more specific to ligand families than hot ones may be

explained by the fact that they are on average more protuberant and more charged. They thus display more variability than hot ones. Indeed, there is more variability in being positively or negatively charged and protuberant (with an important range of protuberant shapes) than in being neutral and flat. S19 Fig presents the electrostatic potential distribution of all energy classes. Cold regions display a larger variability of electrostatic potential (F-test, $p < 2.2e\text{-}16$) than hot regions that are mainly hydrophobic thus displaying neutral charge distributions in average. Consequently, a same hot region may be attractive for a large set of ligands while a cold region may be unfavorable to specific set of ligands, depending on their charges, shapes and other biophysical properties.

Moreover, we show that hot regions are very localized (4.9% of the cells of an energy map) and tend to be similar no matter the ligand. Similarly to protein interfaces that have been extensively characterized in previous studies [47,48,48–50], hot regions are likely to display universal properties of binding, i.e. they are more hydrophobic and more planar, and thus more "sticky" than the other regions. They may provide a non-specific binding patch that is suitable for many ligands. However, we can hypothesize that native partners have evolved to optimize their interfaces (positive design) so that native interactions prevail over non-native competing ones. Then positive design results in conserved binding sites and coevolved interfaces in order to maintain the charge and shape complementarity between functional partners. Indeed, we have previously shown that the docking of native partners lead to more favorable binding energies than the docking of non-native partners when the ligand is constrained to dock around the receptor's native binding site [33,59]. All these results suggest a new physical model of protein surfaces where protein surface regions, in the crowded cellular environment, serve as a proxy for regulating the competition between functional and non-functional interactions. In this model, intermediate and cold regions play an important role by preventing non-functional assemblies and by guiding the interaction process towards

functional ones and hot regions may select the functional assembly among the competing ones through optimized interfaces with the native partner. This model recalls the transitive model proposed by Marin-Lopez *et al* where a path connecting what they call "productive" (near-native) and "non-productive" (non-native) assemblies can be defined [57]. This path consists in distinct conformational states where each one is a macro-state of the binding process involving either the native binding site of each partner, a single native binding site or no native ones. The initial steps consist in macro-states which do not involve native binding sites. Macro-states appearing later during the assembly process would play a mechanistic role by drawing near the binding sites of the two partners. The latest stage would correspond to near-native conformations where van der Waals and de-solvation energies play a major role in the energy of interaction of the corresponding complexes while the electrostatic forces contribute mostly in the energy of non-native assemblies [60,61]. Figure S21 shows the effective electrostatic and van der Waals contributions in the total docking energy for the different surface regions (i.e. cold, intermediate and hot regions). Interestingly, our results concur with the observations made in [60,61] since we show that the contribution of electrostatic in the total docking energy is more important in cold regions while van der Waals energies predominate in hot ones. Characterizing the relationship between the macro-states defined by Marin-Lopez *et al* and the surface regions of different energy levels could provide at the same time a structural, physical and readable characterization of the binding process of two interacting proteins. In particular, it would be interesting to compare the properties of the different macro-states (involving or not the native binding sites of the two proteins) identified for functional and arbitrary pairs to see whether functional pairs displays specific features that would have resulted from an optimization of the binding process.

In this work, we used and extended the application of the 2D energy map representation developed in [36] to develop an original theoretical framework that enables the efficient, automated and integrative analysis of different protein surface features. Many other surface representations have been developed to characterize protein surface properties [62–67]. These representations include 2D projections or more sophisticated methods such as for example using 3D Zernike descriptors as a representation of the protein surface shape [68,69] which is a powerful tool to compare surface properties of either homologous or unrelated proteins since it does not require any prior alignment. 2D maps provide the area distribution of a given feature on the whole protein surface and their discretization enables the study of a given surface property (e.g. protuberance, planarity, stickiness, positively charged regions, or cold and hot regions for example). The advantage with the 2D energy maps is that they are easy to build and manipulate and their straightforward comparison enables (i) the study of relationships between different surface properties through the comparison of their area distributions on a protein surface and (ii) the highlight of the evolutionary constraints exerted on a given feature by comparing its area distribution on the surfaces of homologous proteins. Particularly, this enables the identification and characterization of hot regions on a protein surface which can be either specific or conserved for all ligands and opens up new possibilities for the development of novel methods for protein binding sites prediction and their classification as functional or promiscuous in the continuity of previous developments based on arbitrary docking [33,34,36,59].

Finally, our framework provides a proxy for further protein functional characterization as shown with the five proteins discussed in the *Results* section *Energy maps are specific to protein families*. The comparison of their respective energy maps enables us to reveal

505 biophysical and functional properties that could not be revealed with classical monomeric
506 descriptors such as RMSD or sequence identity. Indeed, our framework can reflect the energy
507 behavior of a protein interacting with a subset of selected partners either functional or
508 arbitrary, thus revealing functional and systemic properties of proteins. This work goes
509 beyond the classical use of binary docking to provide a systemic point of view of protein
510 interactions**,** for example by exploring the propensity of a protein to interact with hundreds of
511 selected ligands, and thus addressing the behavior of a protein in a specific cellular
512 environment. Particularly, exploring the dark interactome (i.e. non-functional assemblies and
513 interactions with non-functional partners) can provide a wealth of valuable information to
514 understand mechanisms driving and regulating protein-protein interactions. Precisely, our 2D
515 energy maps based strategy enables its exploration in an efficient and automated way.

Materials and Methods

Protein dataset

The dataset comprises 74 protein structures divided into 12 families of structural homologs which were selected from the protein docking benchmark 5.0. (see S1 Table for a detailed list of each family). We decided to systematically remove all Antibody/Antigens complexes since they display specific evolutionary properties. Indeed, they did not co-evolve to interact and we can hypothesize that the evolutionary constraints operating on their interaction energy landscapes are different from those of other complexes. Each family is related to at least one other family (its native-related partners family) through a pair of interacting proteins for which the 3D structure of the complex is characterized experimentally (except the V set domain family: the two native partners are homologous and belong to the same family) (S1 Fig). Each family is composed of a monomer selected from the protein-protein docking benchmark 5.0 [70] in its bound and unbound forms, which is called the master protein. Each master protein has a native partner (for which the 3D structure of the corresponding complex has been characterized experimentally) in the database, which is the master protein for another family, except the V set domain family, which is a self-interacting family. When available, we completed families with interologs (i.e. pairs of proteins which have interacting homologs in an other organism) selected in the INTEREVOL database [71] according to the following criteria: (i) experimental structure resolution better than 3.25 Å, (ii) minimum alignment coverage of 75% with the rest of the family members and (iii) minimum sequence identity of 30% with at least one member of the family. Since we were limited by the number of available interologs, we completed families with unbound monomers homologous to the

539 master following the same criteria and by searching for their partners in the following protein-
540 protein interactions databases [72–77]. We consider that all members of a family correspond
541 to native-related partners of all members of their native-related partner family. To address the
542 impact of conformational changes of a protein on its interaction energy maps, we added
543 different NMR conformers. We show that energy maps involving pairs of conformers are
544 significantly more similar than those obtained for other pairs of homologous ligands
545 (unilateral Wilcoxon test, p < 2.2e-16) showing that the conformational changes in a protein
546 (lower than 3Å) have a low impact on the resulting energy maps (S20 Fig).

547

548 **Docking experiment and construction of energy maps**

549 A complete cross-docking experiment was realized with the ATTRACT software [30] on the
550 74 proteins of the dataset, leading to 5476 (74 x 74) docking calculations (Fig 1A).
551 ATTRACT uses a coarse-grain reduced protein representation and a simplified energy
552 function comprising a pseudo Lennard-Jones term and an electrostatic term. The calculations
553 took approximately 20000 hours on a 2.7GHz processor. Prior to docking calculations, all
554 PDB structures were prepared with the DOCKPREP software [78].

555 During a docking calculation, the ligand L_i explores exhaustively the surface of the receptor
556 R_k (whose position is fixed during the procedure), sampling and scoring thousands of
557 different ligand docking poses (between 10000 and 50000 depending on the sizes of the
558 proteins) (Fig 1A). For each protein couple R_k-L_i, a 2D energy map is computed which shows
559 the distribution of the energies of all docking solutions over the receptor surface. To compute
560 these maps, for all docking poses, the spherical coordinates (ϕ, θ) (with respect to the
561 receptor center of mass (CM)) of the ligand CM are represented onto a 2D map in an equal-

562 area 2D sinusoidal projection (Fig 1B) (see [36] for more details). Each couple of coordinates
563 (ϕ, θ) is associated with the energy of the corresponding docking conformation (Fig 1B). A
564 continuous energy map is then derived from the discrete one, where the map is divided into a
565 grid of 36 x 72 cells. Each cell represents the same surface and, depending on the size of the
566 receptor, can span from 2.5 $Å^2$ to 13$Å^2$. For each cell, all solutions with an energy score
567 below 2.7 $kcal/mol^{-1}$ from the lowest solution of the cell are retained, according to the
568 conformations filtering protocol implemented in [33]. The average of the retained energy
569 scores is then assigned to the cell. If there is no docking solution in a cell, a score of 0 is
570 assigned to it. Finally, the energies of the cells are smoothed, by averaging the energy values
571 of each cell and of the eight surrounding neighbors (Fig 1C).

572 For each map, the energy values are discretized into five energy classes of same range leading
573 to a discrete five-colors energy map (Fig 1D). The range is calculated for each energy map
574 and spans from the minimum to the maximum scores of the map cells. The range of the
575 energy classes of the map R_k-L_i is equal to (maxE – minE)/5, where maxE and minE
576 correspond to the maximal and minimal energy values in the R_k-L_i map. Each five-classes
577 energy map is then split into five one-class maps, each one representing an energy class of the
578 map (Fig 1E). The continuous, five-classes and one-class energy maps are calculated for the
579 5476 energy maps.

580

581 **Comparison of energy maps and identification of ligand's homologs**

582 Since, we cannot compare energy maps computed for two unrelated receptors, the procedure
583 is receptor-centered and only compares energy maps produced with different ligands docked
584 with the same receptor. The referential (i.e. the receptor) is thus the same (in other words all
585 grid cells are comparable) for all the energy maps that are compared. For each receptor R_k, we

586 computed a 74x74 energy map distance (EMD) matrix where each entry (*i*,*j*) corresponds to
587 the pairwise distance between the energy maps R_k-L_i and R_k-L_j resulting from the docking of
588 the ligands L_i and L_j on the receptor R_k (Fig 3). The pairwise distance $d_{Man}(R_k\text{-}L_i, R_k\text{-}L_j)$
589 between the energy maps is calculated with a Manhattan distance according to equation (1)

590

591 $$d_{Man}(R_kL_i, R_kL_j) = \sum_{n=1}^{36}\sum_{m=1}^{72}|a_{nm} - b_{nm}| \quad (1)$$
592

593 where a_{nm} and b_{nm} are the cells of row index *n* and column index *m* of the energy maps R_k-L_i
594 and R_k-L_j respectively. Low distances reflect pairs of ligands that induce similar energy maps
595 when they are docked on the same receptor. The procedure presented in Fig 3 is repeated for
596 each receptor of the database resulting in 74 EMD matrices. The 74 EMD matrices are
597 averaged into an averaged distances matrix (ADM). Each entry (*i*,*j*) of the ADM reflects the
598 similarity of the R_k-L_i and R_k-L_j energy maps averaged over all the receptors R_k in the dataset.
599 In order to estimate the extent to which family members display similar energy maps when
600 they are docked with the same receptor, we tested our ability to correctly identify the
601 homologs of the 74 ligands from the only comparison of its energy maps with those of the
602 other ligands. Because, energy maps are receptor-centered, we cannot compare the energy
603 maps computed for two unrelated receptors. The procedure consists in the comparison of
604 energy maps produced with different ligands docked with a same receptor. Two ligands (*i*,*j*)
605 are predicted as homologs according to their corresponding distance (*i*,*j*) in the ADM. Values
606 close to zero should reflect homologous ligand pairs, while values close to one should reflect
607 unrelated ligand pairs. A Receiver Operating Characteristic (ROC) curve and its Area Under
608 the Curve (AUC) are computed from the ADM. True positives (TP) are all the homologous
609 ligand pairs and predicted as such, true negatives (TN) are all the unrelated ligand pairs and
610 predicted as such. False positives (FP) are unrelated ligand pairs but incorrectly predicted as

611 homologous pairs. False negatives (FN) are homologous ligand pairs but incorrectly predicted
612 as unrelated pairs. ROC curves and AUC values were calculated with the R package pROC
613 [79]. The ligand's homologs identification was also realized using the five-classes energy
614 maps or the one-class energy maps taken separately. The five energy class regions display
615 very different sizes, with median ranging from 63 and 66 cells for the cold and hot regions to
616 633 cells for the yellow one. To prevent any bias due to the size of the different classes, we
617 normalized the Manhattan distance by the size of the regions compared in the map. The rest
618 of the procedure is the same than those used for continuous energy maps (Fig 3).

619 To visualize the area distribution of the regions of a given energy class for all ligands on the
620 receptor surface, the 74 corresponding one-class maps are summed into a stacked map where
621 each cell's intensity varies from 0 to 74 (S16 Fig). To remove background-image from these
622 maps, i.e. cells with low intensity (intensity < 17) and the areas of small size (< 4 cells), we
623 used a Dirichlet process mixture model simulation for image segmentation (R package
624 *dpmixsim*) [80].

625

626 **2D projection of monomeric descriptors of protein surfaces**

627 We computed KD hydrophobicity [44], stickiness [25], CV [45] maps of each protein of the
628 dataset, in order to compare their topology with the energy maps. Prior to all, proteins
629 belonging to the same families were structurally aligned with TM-align [81] in order to place
630 them in the same reference frame, making their maps comparable. Particles were generated
631 around the protein surface with a slightly modified Shrake-Rupley algorithm [82]. The
632 density of spheres is fixed at 1Å^2, representing several thousands particles per protein. Each
633 particle is located at 5Å from the surface of the protein. The CV, stickiness and KD

hydrophobicity values of the closest atom of the protein are attributed to each particle. We also generated electrostatic maps reflecting the distribution of the contribution of the electrostatic potential on a protein surface. The electrostatics potential was computed with the APBS software suite [38] using the CHARMM force field [83]. In this case the procedure is different as the electrostatic potential is calculated at each particle position, using the multivalue executable from the APBS software suite.

The CV was calculated following the protocol described in [45] on the all-atom structures. Stickiness and hydrophobicity were calculated on ATTRACT coarse-grain models. After attributing a value to each particle, the position of their spherical coordinates is represented in a 2-D sinusoidal projection, following the same protocol as described in Fig 1 and *Materials and Methods* section *Docking experiment and construction of energy maps*. The map is then smoothed following the protocol in Fig 1.

646 **Acknowledgments**

647 We thank F. Fraternali, R. Guerois, E. Laine, and M. Montes for their constructive comments

648 on the manuscript.

References

[1] J.I. Garzón, L. Deng, D. Murray, S. Shapira, D. Petrey, B. Honig, A computational interactome and functional annotation for the human proteome, ELife. 5 (2016) e18715. doi:10.7554/eLife.18715.

[2] J. Janin, R.P. Bahadur, P. Chakrabarti, Protein–protein interaction and quaternary structure, Quart. Rev. Biophys. 41 (2008). doi:10.1017/s0033583508004708.

[3] I. Nobeli, A.D. Favia, J.M. Thornton, Protein promiscuity and its implications for biotechnology, Nat Biotechnol. 27 (2009) 157–167. doi:10.1038/nbt1519.

[4] I.M.A. Nooren, J.M. Thornton, NEW EMBO MEMBER'S REVIEW: Diversity of protein-protein interactions, The EMBO Journal. 22 (2003) 3486–3492. doi:10.1093/emboj/cdg359.

[5] C.V. Robinson, A. Sali, W. Baumeister, The molecular sociology of the cell, Nature. 450 (2007) 973–982. doi:10.1038/nature06523.

[6] R. Milo, What is the total number of protein molecules per cell volume? A call to rethink some published values: Insights & Perspectives, BioEssays. 35 (2013) 1050–1055. doi:10.1002/bies.201300066.

[7] S.R. McGuffee, A.H. Elcock, Diffusion, Crowding & Protein Stability in a Dynamic Molecular Model of the Bacterial Cytoplasm, PLoS Computational Biology. 6 (2010) e1000694. doi:10.1371/journal.pcbi.1000694.

[8] I. Yu, T. Mori, T. Ando, R. Harada, J. Jung, Y. Sugita, M. Feig, Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm, ELife. 5 (2016). doi:10.7554/eLife.19274.

[9] J.T. Mika, B. Poolman, Macromolecule diffusion and confinement in prokaryotic cells, Curr. Opin. Biotechnol. 22 (2011) 117–126. doi:10.1016/j.copbio.2010.09.009.

[10] R.J. Ellis, Macromolecular crowding: an important but neglected aspect of the intracellular environment, Curr. Opin. Struct. Biol. 11 (2001) 114–119. doi:10.1016/S0959-440X(00)00172-X.

[11] E.D. Levy, J. Kowarzyk, S.W. Michnick, High-Resolution Mapping of Protein Concentration Reveals Principles of Proteome Architecture and Adaptation, Cell Reports. 7 (2014) 1333–1340. doi:10.1016/j.celrep.2014.04.009.

[12] J.S. Richardson, D.C. Richardson, Natural -sheet proteins use negative design to avoid edge-to-edge aggregation, Proceedings of the National Academy of Sciences. 99 (2002) 2754–2759. doi:10.1073/pnas.052706099.

[13] E.J. Deeds, O. Ashenberg, J. Gerardin, E.I. Shakhnovich, Robust protein protein interactions in crowded cellular environments, Proceedings of the National Academy of Sciences. 104 (2007) 14952–14957. doi:10.1073/pnas.0702766104.

[14] S. Pechmann, E.D. Levy, G.G. Tartaglia, M. Vendruscolo, Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins, Proceedings of the National Academy of Sciences. 106 (2009) 10159–10164. doi:10.1073/pnas.0812414106.

[15] J. Karanicolas, J.E. Corn, I. Chen, L.A. Joachimiak, O. Dym, S.H. Peck, S. Albeck, T. Unger, W. Hu, G. Liu, S. Delbecq, G. T. Montelione, C. P. Spiegel, D.R. Liu, D. Baker, A De Novo Protein Binding Pair By Computational Design and Directed Evolution, Molecular Cell. 42 (2011) 250–260. doi:10.1016/j.molcel.2011.03.010.

[16] H. Garcia-Seisdedos, C. Empereur-Mot, N. Elad, E.D. Levy, Proteins evolve on the edge of supramolecular self-assembly, Nature. 548 (2017) 244–247. doi:10.1038/nature23320.

[17] P. Aloy, H. Ceulemans, A. Stark, R.B. Russell, The Relationship Between Sequence and Interaction Divergence in Proteins, Journal of Molecular Biology. 332 (2003) 989–998. doi:10.1016/j.jmb.2003.07.006.

[18] M. Heo, S. Maslov, E. Shakhnovich, Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions, Proceedings of the National Academy of Sciences. 108 (2011) 4258–4263. doi:10.1073/pnas.1009392108.

[19] Q. Wang, A. Zhuravleva, L.M. Gierasch, Exploring Weak, Transient Protein–Protein Interactions in Crowded In Vivo Environments by In-Cell Nuclear Magnetic Resonance Spectroscopy, Biochemistry. 50 (2011) 9225–9236. doi:10.1021/bi201287e.

[20] N. Zhang, L. An, J. Li, Z. Liu, L. Yao, Quinary Interactions Weaken the Electric Field Generated by Protein Side-Chain Charges in the Cell-like Environment, J. Am. Chem. Soc. 139 (2017) 647–654. doi:10.1021/jacs.6b11058.

[21] X. Mu, S. Choi, L. Lang, D. Mowray, N.V. Dokholyan, J. Danielsson, M. Oliveberg, Physicochemical code for quinary protein interactions in *Escherichia coli*, Proc Natl Acad Sci USA. 114 (2017) E4556–E4563. doi:10.1073/pnas.1621227114.

[22] C. Pal, B. Papp, L.D. Hurst, Highly Expressed Genes in Yeast Evolve Slowly, Genetics. 158 (2001) 927–931.

[23] D.A. Drummond, C.O. Wilke, Mistranslation-Induced Protein Misfolding as a Dominant Constraint on Coding-Sequence Evolution, Cell. 134 (2008) 341–352. doi:10.1016/j.cell.2008.05.042.

[24] J. Zhang, S. Maslov, E.I. Shakhnovich, Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size, Molecular Systems Biology. 4 (2008). doi:10.1038/msb.2008.48.

[25] E.D. Levy, S. De, S.A. Teichmann, Cellular crowding imposes global constraints on the chemistry and evolution of proteomes, Proceedings of the National Academy of Sciences. 109 (2012) 20461–20466. doi:10.1073/pnas.1209312109.

[26] J.-R. Yang, B.-Y. Liao, S.-M. Zhuang, J. Zhang, Protein misinteraction avoidance causes highly expressed proteins to evolve slowly, Proceedings of the National Academy of Sciences. 109 (2012) E831–E840. doi:10.1073/pnas.1117408109.

[27] P.E. Schavemaker, W.M. Śmigiel, B. Poolman, Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome, ELife. 6 (2017) e30084. doi:10.7554/eLife.30084.

[28] D.W. Ritchie, V. Venkatraman, Ultra-fast FFT protein docking on graphics processors, Bioinformatics. 26 (2010) 2398–2405. doi:10.1093/bioinformatics/btq444.

[29] B.G. Pierce, Y. Hourai, Z. Weng, Accelerating Protein Docking in ZDOCK Using an Advanced 3D Convolution Library, PLoS ONE. 6 (2011) e24657. doi:10.1371/journal.pone.0024657.

[30] S. de Vries, M. Zacharias, Flexible docking and refinement with a coarse-grained protein model using ATTRACT: Flexible Protein-Protein Docking and Refinement, Proteins: Structure, Function, and Bioinformatics. 81 (2013) 2167–2174. doi:10.1002/prot.24400.

[31] M.N. Wass, G. Fuentes, C. Pons, F. Pazos, A. Valencia, Towards the prediction of protein interaction partners using physical docking, Molecular Systems Biology. 7 (2011) 469–469. doi:10.1038/msb.2011.3.

[32] M. Ohue, Y. Matsuzaki, T. Shimoda, T. Ishida, Y. Akiyama, Highly precise protein-protein interaction prediction based on consensus between template-based and de novo docking methods, BMC Proc. 7 (2013) S6. doi:10.1186/1753-6561-7-S7-S6.

[33] A. Lopes, S. Sacquin-Mora, V. Dimitrova, E. Laine, Y. Ponty, A. Carbone, Protein-Protein Interactions in a Crowded Environment: An Analysis via Cross-Docking Simulations and Evolutionary Information, PLoS Comput Biol. 9 (2013) e1003369. doi:10.1371/journal.pcbi.1003369.

[34] L. Vamparys, B. Laurent, A. Carbone, S. Sacquin-Mora, Great interactions: How binding incorrect partners can teach us about protein recognition and function: Predicting Binding Sites From Cross-Docking, Proteins: Structure, Function, and Bioinformatics. 84 (2016) 1408–1421. doi:10.1002/prot.25086.

[35] E. Laine, A. Carbone, Protein social behavior makes a stronger signal for partner identification than surface geometry: Protein Social Behavior, Proteins: Structure, Function, and Bioinformatics. 85 (2017) 137–154. doi:10.1002/prot.25206.

[36] J. Fernández-Recio, M. Totrov, R. Abagyan, Identification of Protein–Protein Interaction Sites from Docking Energy Landscapes, Journal of Molecular Biology. 335 (2004) 843–865. doi:10.1016/j.jmb.2003.10.069.

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

[38] E. Jurrus, D. Engel, K. Star, K. Monson, J. Brandi, L.E. Felberg, D.H. Brookes, L. Wilson, J. Chen, K. Liles, M. Chun, P. Li, D.W. Gohara, T. Dolinsky, R. Konecny, D.R. Koes, J.E. Nielsen, T. Head-Gordon, W. Geng, R. Krasny, G.-W. Wei, M.J. Holst, J.A. McCammon, N.A. Baker, Improvements to the APBS biomolecular solvation software suite, Protein Science. 27 (2018) 112–128. doi:10.1002/pro.3280.

[39] P. Xu, D.M. Duong, N.T. Seyfried, D. Cheng, Y. Xie, J. Robert, J. Rush, M. Hochstrasser, D. Finley, J. Peng, Quantitative proteomics reveals the function of unconventional ubiquitin chains in proteasomal degradation, Cell. 137 (2009) 133–145. doi:10.1016/j.cell.2009.01.041.

[40] R.L. Welchman, C. Gordon, R.J. Mayer, Ubiquitin and ubiquitin-like proteins as multifunctional signals, Nat Rev Mol Cell Biol. 6 (2005) 599–609. doi:10.1038/nrm1700.

[41] T. Molnár, J. Vörös, B. Szeder, K. Takáts, J. Kardos, G. Katona, L. Gráf, Comparison of complexes formed by a crustacean and a vertebrate trypsin with bovine pancreatic trypsin inhibitor – the key to achieving extreme stability?, The FEBS Journal. 280 (2013) 5750–5763. doi:10.1111/febs.12491.

[42] I.G. Sumner, G.W. Harris, M.A.J. Taylor, R.W. Pickersgill, A.J. Owen, P.W. Goodenough, Factors effecting the thermostability of cysteine proteinases from Carica papaya, European Journal of Biochemistry. 214 (1993) 129–134. doi:10.1111/j.1432-1033.1993.tb17904.x.

[43] J. Wang, A. Smolyar, K. Tan, J. Liu, M. Kim, Z.J. Sun, G. Wagner, E.L. Reinherz, Structure of a Heterophilic Adhesion Complex between the Human CD2 and CD58 (LFA-3) Counterreceptors, Cell. 97 (1999) 791–803. doi:10.1016/S0092-8674(00)80790-4.

[44] J. Kyte, R.F. Doolittle, A simple method for displaying the hydropathic character of a protein, Journal of Molecular Biology. 157 (1982) 105–132. doi:10.1016/0022-2836(82)90515-0.

[45] M. Mezei, A new method for mapping macromolecular topography, Journal of Molecular Graphics and Modelling. 21 (2003) 463–472. doi:10.1016/S1093-3263(02)00203-6.

[46] J.W. Tukey, Comparing Individual Means in the Analysis of Variance, Biometrics. 5 (1949) 99–114. doi:10.2307/3001913.

[47] Lo Conte, Loredana, C. Chothia, J. Janin, The atomic structure of protein-protein recognition sites, Journal of Molecular Biology. 285 (1999) 2177–2198. doi:10.1006/jmbi.1998.2439.

[48] P. Chakrabarti, J. Janin, Dissecting protein-protein recognition sites, Proteins: Structure, Function, and Genetics. 47 (2002) 334–343. doi:10.1002/prot.10085.

[49] X. Li, O. Keskin, B. Ma, R. Nussinov, J. Liang, Protein–Protein Interactions: Hot Spots and Structurally Conserved Residues often Locate in Complemented Pockets that Pre-organized in the Unbound States: Implications for Docking, Journal of Molecular Biology. 344 (2004) 781–795. doi:10.1016/j.jmb.2004.09.051.

[50] O. Keskin, B. Ma, R. Nussinov, Hot Regions in Protein–Protein Interactions: The Organization and Contribution of Structurally Conserved Hot Spot Residues, Journal of Molecular Biology. 345 (2005) 1281–1294. doi:10.1016/j.jmb.2004.10.077.

[51] J. Andreani, G. Faure, R. Guerois, Versatility and Invariance in the Evolution of Homologous Heteromeric Interfaces, PLoS Computational Biology. 8 (2012) e1002677. doi:10.1371/journal.pcbi.1002677.

[52] C. Tang, J. Iwahara, G.M. Clore, Visualization of transient encounter complexes in protein–protein association, Nature. 444 (2006) 383. doi:10.1038/nature05201.

[53] W.B. Monteith, R.D. Cohen, A.E. Smith, E. Guzman-Cisneros, G.J. Pielak, Quinary structure modulates protein stability in cells, Proceedings of the National Academy of Sciences. 112 (2015) 1739–1742. doi:10.1073/pnas.1417415112.

[54] D. Kozakov, K. Li, D.R. Hall, D. Beglov, J. Zheng, P. Vakili, O. Schueler-Furman, I.C. Paschalidis, G.M. Clore, S. Vajda, Encounter complexes and dimensionality reduction in protein–protein association, ELife. 3 (2014) e01370. doi:10.7554/eLife.01370.

[55] G. Schreiber, A.E. Keating, Protein binding specificity versus promiscuity, Current Opinion in Structural Biology. 21 (2011) 50–61. doi:10.1016/j.sbi.2010.10.002.

[56] J. Planas-Iglesias, J. Bonet, J. García-García, M.A. Marín-López, E. Feliu, B. Oliva, Understanding Protein–Protein Interactions Using Local Structural Features, Journal of Molecular Biology. 425 (2013) 1210–1224. doi:10.1016/j.jmb.2013.01.014.

[57] M.A. Marín-López, J. Planas-Iglesias, J. Aguirre-Plans, J. Bonet, J. Garcia-Garcia, N. Fernandez-Fuentes, B. Oliva, On the mechanisms of protein interactions: predicting their affinity from unbound tertiary structures, Bioinformatics. 34 (2018) 592–598. doi:10.1093/bioinformatics/btx616.

[58] P.L. Kastritis, J.P.G.L.M. Rodrigues, G.E. Folkers, R. Boelens, A.M.J.J. Bonvin, Proteins Feel More Than They See: Fine-Tuning of Binding Affinity by Properties of the Non-Interacting Surface, Journal of Molecular Biology. 426 (2014) 2632–2652. doi:10.1016/j.jmb.2014.04.017.

[59] S. Sacquin-Mora, A. Carbone, R. Lavery, Identification of Protein Interaction Partners and Protein–Protein Interaction Sites, Journal of Molecular Biology. 382 (2008) 1276–1289. doi:10.1016/j.jmb.2008.08.002.

[60] R. Alsallaq, H.-X. Zhou, Electrostatic rate enhancement and transient complex of protein–protein association, Proteins: Structure, Function, and Bioinformatics. 71 (2008) 320–335. doi:10.1002/prot.21679.

[61] H.-X. Zhou, P.A. Bates, Modeling protein association mechanisms and kinetics, Current Opinion in Structural Biology. 23 (2013) 887–893. doi:10.1016/j.sbi.2013.06.014.

[62] K. Pawłowski, A. Godzik, Surface Map Comparison: Studying Function Diversity of Homologous Proteins, Journal of Molecular Biology. 309 (2001) 793–806. doi:10.1006/jmbi.2001.4630.

[63] T.V. Pyrkov, A.O. Chugunov, N.A. Krylov, D.E. Nolde, R.G. Efremov, PLATINUM: a web tool for analysis of hydrophobic/hydrophilic organization of biomolecular complexes, Bioinformatics. 25 (2009) 1201–1202. doi:10.1093/bioinformatics/btp111.

[64] A.D. Koromyslova, A.O. Chugunov, R.G. Efremov, Deciphering Fine Molecular Details of Proteins' Structure and Function with a *Protein Surface Topography (PST)* Method, Journal of Chemical Information and Modeling. 54 (2014) 1189–1199. doi:10.1021/ci500158y.

[65] H. Yang, R. Qureshi, A. Sacan, Protein surface representation and analysis by dimension reduction, Proteome Science. 10 (2012) S1. doi:10.1186/1477-5956-10-S1-S1.

[66] G. Levieux, M. Montes, Towards real-time interactive visualization modes of molecular surfaces: examples with udock, in: 2015 IEEE 1st International Workshop on Virtual and Augmented Reality for Molecular Science (VARMS@IEEEVR), 2015: pp. 19–23. doi:10.1109/VARMS.2015.7151723.

[67] D.G. Kontopoulos, D. Vlachakis, G. Tsiliki, S. Kossida, Structuprint: a scalable and extensible tool for two-dimensional representation of protein surfaces, BMC Structural Biology. 16 (2016). doi:10.1186/s12900-016-0055-7.

[68] L. Sael, B. Li, D. La, Y. Fang, K. Ramani, R. Rustamov, D. Kihara, Fast protein tertiary structure retrieval based on global surface shape similarity, Proteins: Structure, Function, and Bioinformatics. 72 (2008) 1259–1273. doi:10.1002/prot.22030.

[69] L. Sael, D. La, B. Li, R. Rustamov, D. Kihara, Rapid comparison of properties on protein surface, Proteins: Structure, Function, and Bioinformatics. 73 (2008) 1–10. doi:10.1002/prot.22141.

[70] T. Vreven, I.H. Moal, A. Vangone, B.G. Pierce, P.L. Kastritis, M. Torchala, R. Chaleil, B. Jiménez-García, P.A. Bates, J. Fernandez-Recio, A.M.J.J. Bonvin, Z. Weng, Updates to the Integrated Protein–Protein Interaction Benchmarks: Docking Benchmark Version 5 and Affinity Benchmark Version 2, Journal of Molecular Biology. 427 (2015) 3031–3041. doi:10.1016/j.jmb.2015.07.016.

[71] G. Faure, J. Andreani, R. Guerois, InterEvol database: exploring the structure and evolution of protein complex interfaces, Nucleic Acids Research. 40 (2012) D847–D856. doi:10.1093/nar/gkr845.

[72] L. Salwinski, The Database of Interacting Proteins: 2004 update, Nucleic Acids Research. 32 (2004) 449D – 451. doi:10.1093/nar/gkh086.

[73] C. Alfarano, The Biomolecular Interaction Network Database and related tools 2005 update, Nucleic Acids Research. 33 (2004) D418–D424. doi:10.1093/nar/gki051.

[74] U. Guldener, MPact: the MIPS protein interaction resource on yeast, Nucleic Acids Research. 34 (2006) D436–D441. doi:10.1093/nar/gkj003.

[75] C. Stark, BioGRID: a general repository for interaction datasets, Nucleic Acids Research. 34 (2006) D535–D539. doi:10.1093/nar/gkj109.

[76] S. Kerrien, Y. Alam-Faruque, B. Aranda, I. Bancarz, A. Bridge, C. Derow, E. Dimmer, M. Feuermann, A. Friedrichsen, R. Huntley, C. Kohler, J. Khadake, C. Leroy, A. Liban, C. Lieftink, L. Montecchi-Palazzi, S. Orchard, J. Risse, K. Robbe, B. Roechert, D. Thorneycroft, Y. Zhang, R. Apweiler, H. Hermjakob, IntAct--open source resource for molecular interaction data, Nucleic Acids Research. 35 (2007) D561–D565. doi:10.1093/nar/gkl958.

[77] D. Alonso-López, M.A. Gutiérrez, K.P. Lopes, C. Prieto, R. Santamaría, J. De Las Rivas, APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks, Nucleic Acids Res. 44 (2016) W529–W535. doi:10.1093/nar/gkw363.

[78] P.T. Lang, S.R. Brozell, S. Mukherjee, E.F. Pettersen, E.C. Meng, V. Thomas, R.C. Rizzo, D.A. Case, T.L. James, I.D. Kuntz, DOCK 6: Combining techniques to model RNA–small molecule complexes, RNA. 15 (2009) 1219–1230. doi:10.1261/rna.1563609.

[79] X. Robin, N. Turck, A. Hainard, N. Tiberti, F. Lisacek, J.-C. Sanchez, M. Müller, pROC: an open-source package for R and S+ to analyze and compare ROC curves, BMC Bioinformatics. 12 (2011) 77. doi:10.1186/1471-2105-12-77.

[80] A.R. Ferreira da Silva, A Dirichlet process mixture model for brain MRI tissue classification, Medical Image Analysis. 11 (2007) 169–182. doi:10.1016/j.media.2006.12.002.

[81] Y. Zhang, J. Skolnick, TM-align: a protein structure alignment algorithm based on the TM-score, Nucleic Acids Research. 33 (2005) 2302–2309. doi:10.1093/nar/gki524.

[82] A. Saladin, S. Fiorucci, P. Poulain, C. Prévost, M. Zacharias, PTools: an opensource molecular docking library, BMC Struct Biol. 9 (2009) 27. doi:10.1186/1472-6807-9-27.

[83] A.D. Mackerell Jr., M. Feig, C.L. Brooks III, Extending the treatment of backbone energetics in protein force fields: Limitations of gas-phase quantum mechanics in reproducing protein conformational distributions in molecular dynamics simulations, Journal of Computational Chemistry. 25 (2004) 1400–1415. doi:10.1002/jcc.20065.

Fig. 1. 2D asymmetrical representation of docking energy landscapes and resulting energy maps. (*A*) Three-dimensional (3D) representation of the ligand docking poses around the receptor. Each dot corresponds to the center of mass (CM) of a ligand docking pose and is colored according to its docking energy score. (*B*) Representation of the CM of the ligand docking poses after an equal-area 2D sinusoidal projection. CMs are colored according to the same scale as in A. (*C*) Continuous energy map (see *Materials and Methods* for more details). (*D*) Five-class map. The energy map is discretized into five energy classes (*E*) One-class maps. Top to bottom: one-class maps that highlight respectively hot, warm, lukewarm, cool and cold regions.

Fig. 2. Interaction propensity for the receptor 2AYN_A and four different ligands. 2D energy maps for the receptor 2AYN_A (ubiquitin carboxyl-terminal hydrolase (UCH) family) docked with the ligands 1XD3_B (native partner), 1NDD_B (homolog of the native partner), 1YVB_A and 2NQD_B (arbitrary partners). The star indicates the localization of the experimentally determined interaction site for the ubiquitin, the circle-cross indicates the homodimerization site of 2AYN_A.

Fig. 3. Experimental Protocol. (*A*) A receptor protein is docked with all proteins of the dataset (namely the ligands) resulting in 74 docking calculations. (*B*) For each docking calculation, an energy map is computed as well as its corresponding five-classes and one-class energy maps, with the procedure described in Fig 1 and *Materials and Methods*. (*C*) An energy map distance (EMD) matrix is computed, representing the pairwise distances between the 74 energy maps resulting from the docking of all ligands with this receptor. Each cell (*i*,*j*)

of the matrix represents the Manhattan distance between the two energy maps resulting from the docking of ligands *i* and *j* with the receptor. A small distance indicates that the ligands *i* and *j* produce similar energy maps when docked with this receptor. In other words, it reflects that the interaction propensity of this receptor is similar for these two ligands. To prevent any bias from the choice of the receptor, the whole procedure is repeated for each receptor of the database, leading to 74 EMD matrices.

Fig. 4. Boxplots of energy map pairwise distances between homologous ligand pairs from native-related partner families, homologous ligand pairs from arbitrary partner families and random ligand pairs. For each receptor, we computed (i) the average of energy map distances of pair of homologous ligands belonging to its native-related partner family(ies), (ii) the average of energy map distances of pair of homologous ligands belonging to its non-native-related partner families, and (iii) the average of energy map distances of random pairs. P-values are calculated with a unilateral Wilcoxon test.

Fig. 5. Receiver operating characteristic (ROC) curve and its Area Under the Curve (AUC). ROC are calculated on the averaged distance matrix (ADM) including either all pairs (blue) or only arbitrary pairs (red) (see *Materials and Methods* for more details).

Fig. 6. Ubiquitin-related family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the ubiquitin-related family (for the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair (*i,j*) averaged over the 74 receptors of the dataset. (*B*) Pairwise sequence identity

matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the family. An electrostatic map represents the distribution of the electrostatic potential on the surface of a protein (for more details, see S15 Fig and *Materials and Methods*). On the electrostatic maps, lysines positions are indicated by stars. Cartoon structures are colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.

Fig. 7. Stacked maps of 1P9D_U after the filtering of cells with too low intensity and areas of too small size. The protocol to generate stacked maps is presented in S16 Fig. (*A-E*) Stacked map for cold, cool, lukewarm, warm and hot regions respectively. The cell intensity in a stacked map of a given energy class indicates the number of times the cell has been associated to this energy class in all the corresponding one-class maps. One should notice that stacked maps of two different energy classes can overlap because a map cell can be associated to different energy classes depending on the docked ligands. S17 Fig presents cold and hot stacked maps of 1P9D_U computed for each ligand family.

Fig. 8. Boxplots of three descriptors of the protein surface. (*A*) the stickiness values, (*B*) the Kyte-Doolittle hydrophobicity and (*C*) the CV values, depending on the energy class. The stickiness, hydrophobicity and CV values are calculated for each protein following the protocol described in *Materials and Methods*. For each of these criteria, *p-values* between the median values of two "successive" energy classes were computed using the Tukey HSD statistical test [46].

71

72

Table 1. AUC obtained with different types of energy maps.

type of map	continuous energy maps	five-classes energy maps	hot energy maps	warm energy maps	lukewarm energy maps	cool energy maps	cold energy maps
AUC	0.79	0.77	0.73	0.76	0.76	0.76	0.79

The AUC are calculated from the ADM with the continuous energy maps (Fig 1C), the five-classes energy maps (Fig 1D) and the one-class energy maps (Fig 1E) (see Materials and Methods for more details).

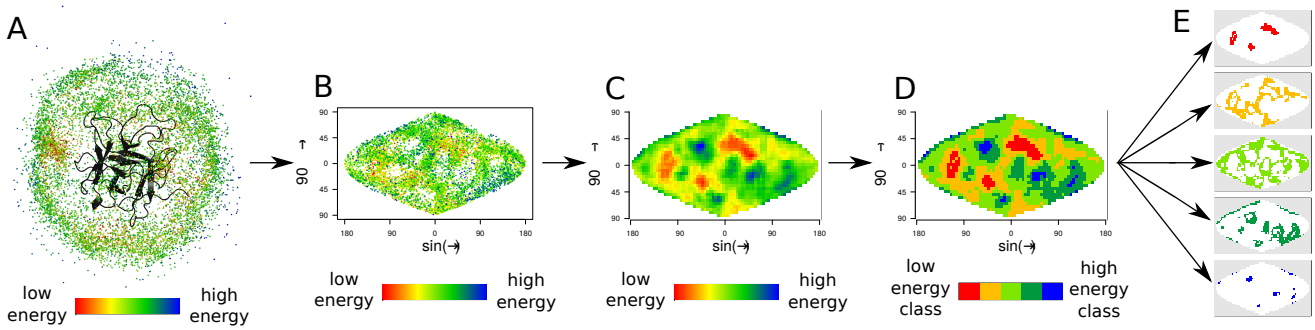

Fig. 1

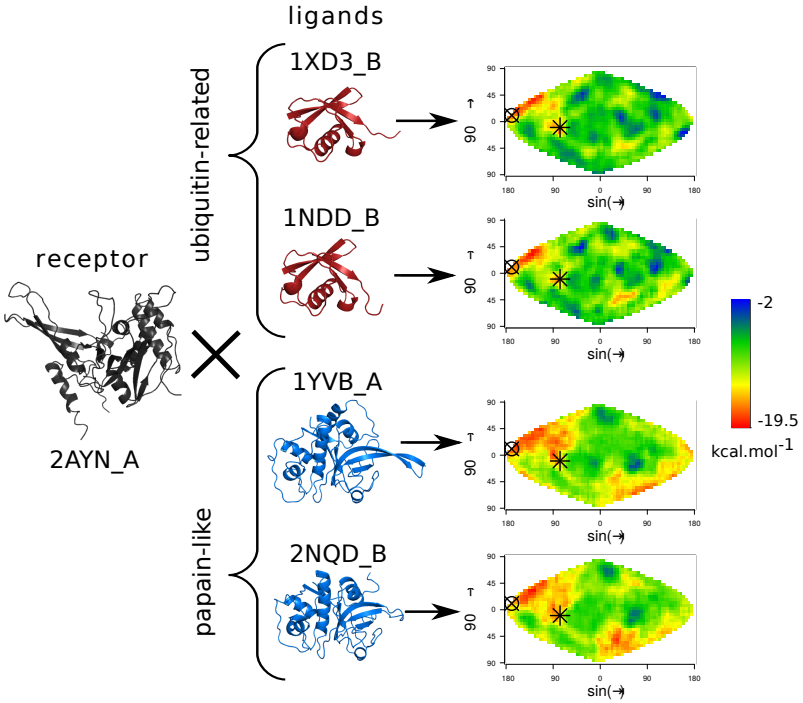

Fig 2.

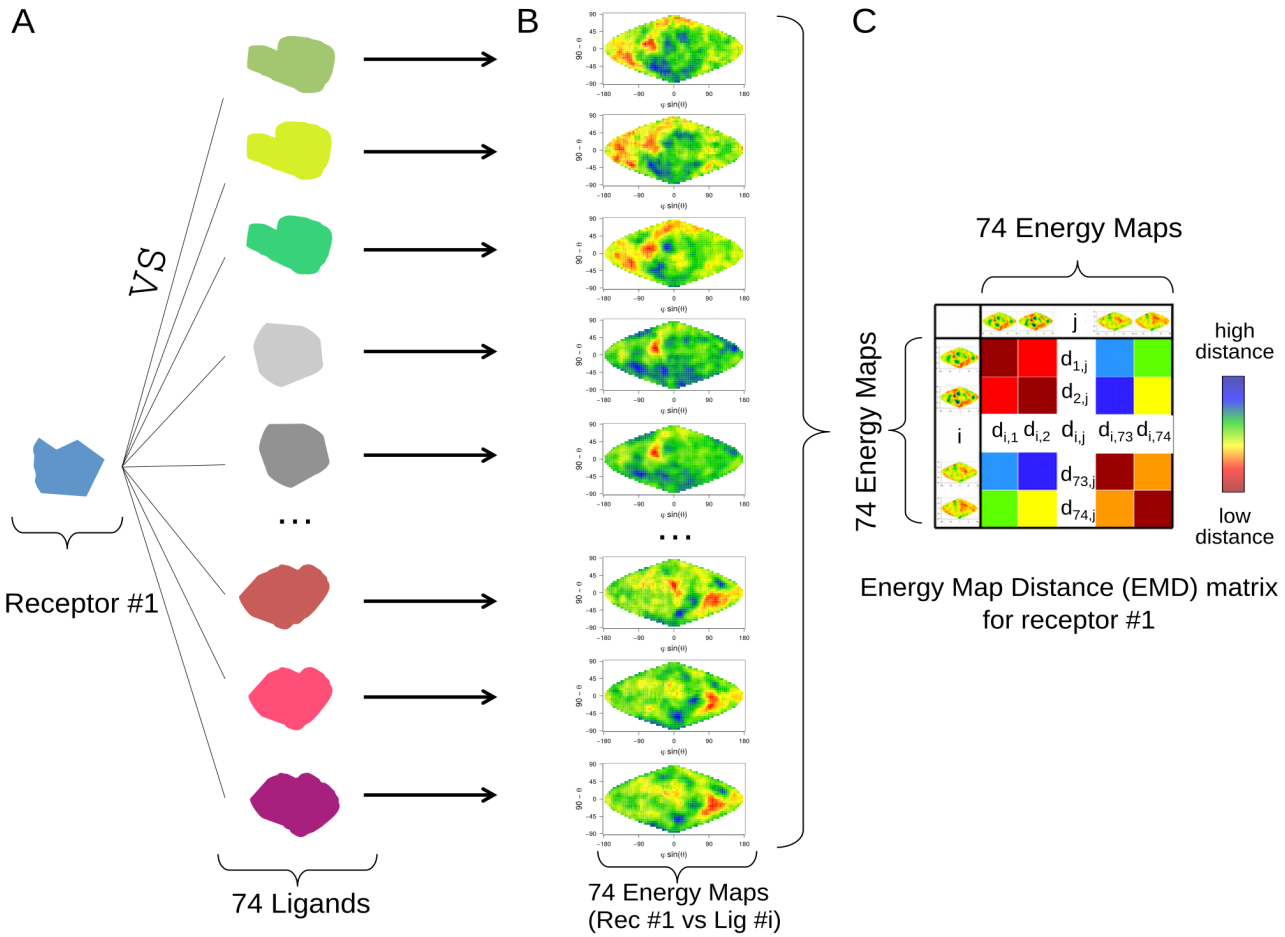

Fig. 3

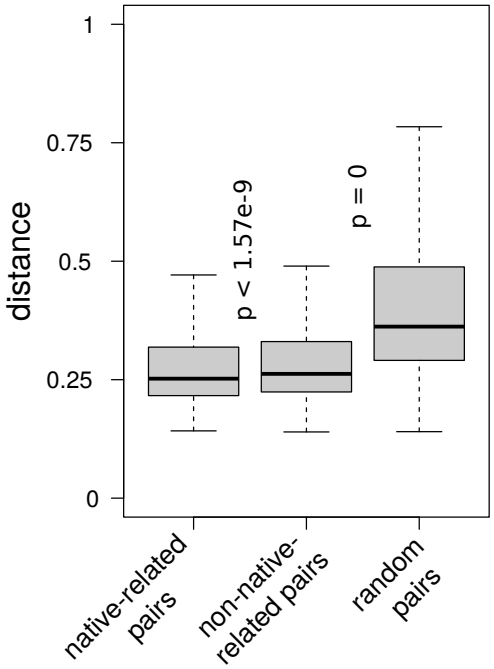

Fig 4.

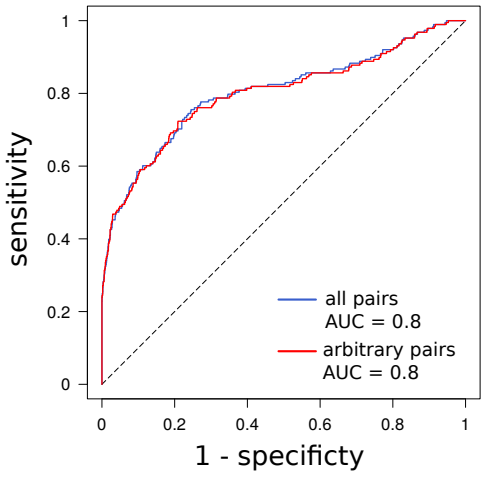

Fig. 5

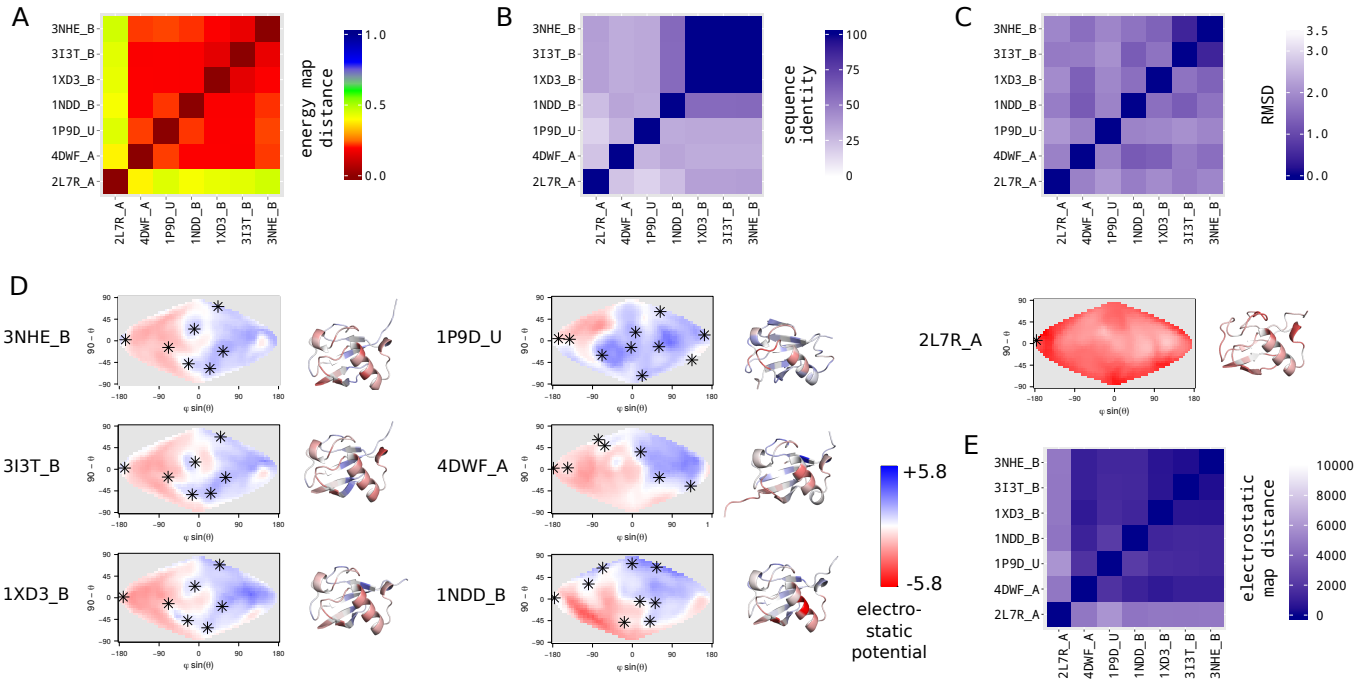

Fig 6.

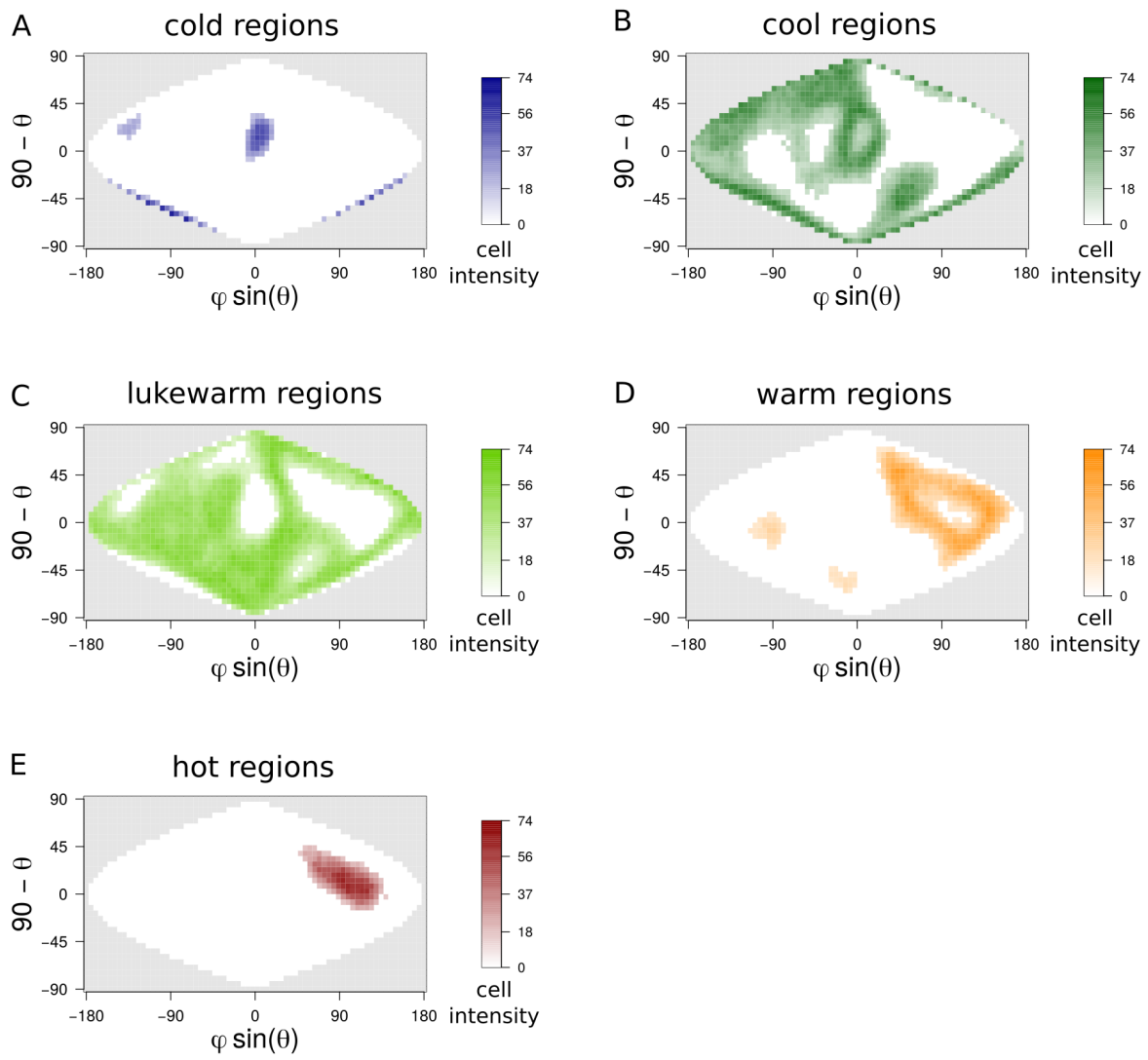

Fig. 7

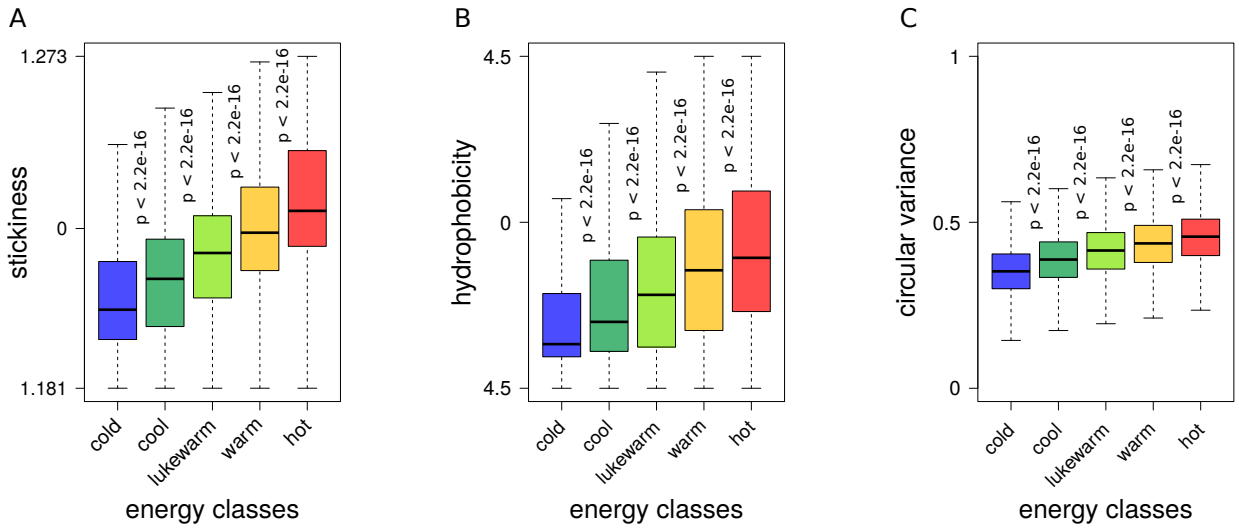

Fig 8.