Title

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

Authors

H. Schweke[a], MH. Mucchielli[ab], S. Sacquin-Mora[c], W. Bei[a], A. Lopes[a]

Major Subject Area(s)

Computational and Systems Biology

Authors affiliations

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

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

[c] 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

Abstract

In the crowded cell, functional and non-functional interactions undergo severe competition. Understanding how a protein binds the right piece in the right way in this complex jigsaw puzzle is crucial and difficult to address experimentally. To interrogate how this competition constrains the behavior of proteins with respect to their partners or random encounters, we (i) performed thousands of docking simulations to systematically characterize the interaction energy landscapes of functional and non-functional protein pairs and (ii) developed an original theoretical framework based on two-dimensional energy maps that reflect the propensity of a protein surface to interact. Strikingly, we show that the interaction propensity of not only binding sites but also of the rest of protein surfaces is conserved for homologous partners, and this feature holds for both functional and non-functional partners. We highlight a new role for non-interacting regions in preventing non-functional interactions and guiding the interaction process toward functional interactions.

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 (Garzón et al., 2016; Janin et al., 2008; Nobeli et al., 2009; Nooren and Thornton, 2003; Robinson et al., 2007). The concentration of proteins in a cell has been estimated to be approximately 2-4 million proteins per cubic micron (Milo, 2013). In such a highly crowded environment, proteins constantly encounter each other and numerous non-specific interactions are likely to occur (McGuffee and Elcock, 2010; Yu et al., 2016). For example, in the cytosol of *S. cerevisiae* a protein can encounter no less than 2000 different proteins (Levy et al., 2014). In this complex jigsaw puzzle, each protein has evolved to bind the right piece in the right way (positive design) and to prevent misassembly and non-functional interactions (negative design) (Deeds et al., 2007; Garcia-Seisdedos et al., 2017; Karanicolas et al., 2011; Pechmann et al., 2009; Richardson and Richardson, 2002).

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. For example, homologs sharing at least 30% sequence identity almost invariably interact in the same way (Aloy et al., 2003). Conversely, negative design prevents proteins to be trapped in the numerous competing non-functional interactions inherent to the crowded environment of the cell. Particularly, the misinteraction avoidance shapes the evolution and physico-chemical properties of abundant proteins, resulting in slower evolution and less sticky surfaces than what is observed for less abundant ones (Drummond and Wilke, 2008; Heo et al., 2011; Levy et al., 2012; Pál et al., 2001; Yang et al., 2012; Zhang et al., 2008). 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 (Levy et al., 2012). 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 (Schavemaker et al., 2017). 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 (Schavemaker et al., 2017).

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 and 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. Concretely, 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 ability to interact 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-related partners or arbitrary partners. Docking simulations enable to energetically characterize 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 (Fiorucci and Zacharias, 2010; Pierce et al., 2011; Ritchie and Venkatraman, 2010), opening the way for extensive cross-docking experiments (Lopes et al., 2013; Ohue et al., 2013; Vamparys et al., 2016; Wass et al., 2011). Protein docking enables to explore the interaction propensity of the whole protein surface by simulating alternative binding modes. Here, we performed a cross-docking experiment involving 72 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 landscape resulting from each docking calculation with a two dimensional (2-D) 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 homologous partners, thus addressing the evolution of the propensity of the whole protein surface to interact with homology-related partners either native or arbitrary.

Results

The interaction propensity of a protein to interact either with native or arbitrary partners is conserved during evolution

We ask whether the propensity of a protein surface to interact with native-related homologous partners is conserved among homologous ligands, and whether this remains true for arbitrary homologous partners. Arbitrary pairs refer to pairs of proteins for which no interaction between them or their respective homologs has been experimentally characterized in the Protein Data Bank (Berman et al., 2000). To test the aforementioned hypothesis, we built a database comprising 72 protein structures divided into 12 structural homolog families (Table S1). 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 a native-related partner family (Fig. S1). We performed a cross-docking experiment using the ATTRACT software (Fiorucci and Zacharias, 2010) on all possible protein pairs in the database (Fig. 1*A* and *Materials and Methods*). ATTRACT enables a homogenous and exhaustive conformational sampling and is well suited to investigate the propensity of the whole surface of a protein to interact with a given ligand. For each docking calculation, we produced a 2-D energy map, which provides the distribution of interaction energies for all docking solutions over the whole receptor surface (Fig. 1*B* and *Materials and Methods*, Fig. 2*A-C*). Consequently, energy maps reflect the propensity of the whole surface of a protein (namely, the receptor) to interact with the docked partner (namely, the ligand). Because energy maps computed for two unrelated receptors are not comparable, the procedure is ligand-centered and focuses on the comparison of energy maps produced with different ligands docked with the same receptor. The comparison of two energy maps enables to evaluate the similarity of the interaction

141 propensities of a receptor with the two corresponding ligands. To prevent any bias, each
142 protein plays alternatively the role of receptor and ligand leading to a total of 72x72 docking
143 calculations.
144 Fig. 3*A* represents the energy maps computed for the receptor 2AYN_A, the human ubiquitin
145 carboxyl-terminal hydrolase 14 (family UCH) docked with (i) its native partner (1XD3_B,
146 ubiquitin-related family), a homolog of its partner (1NDD_B) and (ii) two arbitrary
147 homologous ligands (1YVB_A and 1NQD_B from the papain-like family). For all four
148 ligands, either native-related or arbitrary partners, docking calculations lead to an
149 accumulation of low-energy solutions (hot regions in red) around the two experimentally
150 known binding sites of the receptor. The first one corresponds to the interaction site with the
151 native partner, ubiquitin (pdb id 2ayo). The second one corresponds to its homodimerisation
152 site (pdb id 2ayn). This indicates that native-related but also arbitrary partners tend to bind
153 onto the native binding sites of native partners as observed in earlier studies (Fernandez-Recio
154 et al., 2004; Vamparys et al., 2016). The same tendency is observed for all 72 ligands in the
155 database (Fig. 3*B*). Their 20 best docking poses systematically tend to accumulate in the
156 vicinity of the two native interaction sites. Whereas the low-energy solutions for most ligands
157 accumulate around the same interaction sites (i.e. the native binding sites), we observe that,
158 globally, 2-D energy maps (i) seem to be more similar between ligands of a same family than
159 between ligands belonging to different families (Fig. 3*A*). The two energy maps obtained with
160 the ligands of the native-related partners family both reveal two sharp hot regions around the
161 native sites and a subset of well-defined cold regions (i.e. blue regions corresponding to high
162 energy solutions) placed in the same area in the map's upper-right quadrant. In contrast, the
163 energy maps obtained for the two ligands of the papain-like family display a large hot region
164 around the two native binding sites of the receptor, extending to the upper-left and bottom-
165 right regions of the map, suggesting a large promiscuous binding region for these ligands.

166 We ask whether the observation made for the receptor 2AYN_A that energy maps produced
167 by homologous ligands are more similar than those produced by unrelated ones can be
168 generalized to all proteins of the dataset. Therefore, we systematically compared the energy
169 maps produced for a single receptor docked successively with the 72 ligands of the dataset by
170 calculating the Manhattan distance between each pair of maps (see *Materials and Methods*).
171 To prevent any bias from the choice of the receptor, the procedure was repeated for each
172 receptor of the dataset. Fig. 4 represents the boxplots of energy map pairwise distances
173 calculated between random ligand pairs or between homologous ligand pairs docked with
174 their native-related partners or with arbitrary partners. Homologous ligands docked either
175 with their native-related partners or arbitrary partners display significantly lower energy map
176 distances than random ligand pairs (Wilcoxon test $p = 0$). This indicates that energy maps
177 produced by homologous ligands docked with a given receptor are more similar than those
178 produced with non-homologous ligands. Interestingly, this observation holds whether the
179 receptor-ligand pair is a native pair or not.

180

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

182 The results presented above prompt us to assess the extent to which energy maps are specific
183 to ligand families. Therefore, we tested our ability to predict the homologs of a given ligand
184 based only on the comparison of its energy maps with those of the other ligands. For each
185 receptor, we built an energy map distance (EMD) matrix, where each entry (i,j) corresponds
186 to the similarity between the energy maps of ligands i and j respectively docked with that
187 receptor (Fig. 1*C* and *Materials and Methods*). The 72 resulting EMD matrices (Fig. 1*D* and
188 *Materials and Methods*), are then averaged in an averaged distances matrix (ADM). Each
189 entry of the ADM corresponds to the similarity between two sets of 72 energy maps produced
190 by two different ligands and reflects the behavior similarity for these two ligands docked on

191 each receptor in the dataset. We computed a receiver operating characteristic (ROC) curve
192 from the ADM (see Fig. 1*E* and *Materials and Methods*) which evaluates our capacity to
193 discriminate the homologs of a given ligand from non-homologous ligands by comparing
194 their respective energy maps computed with all 72 receptors of the dataset. The true positive
195 set consists in the homologous protein pairs while the true negative set consists in any
196 homology-unrelated protein pair. The resulting Area Under the Curve (AUC) is equal to 0.8
197 (Fig. 5). We evaluated the robustness of the ligand's homologs prediction depending on the
198 size of the receptor subset with a bootstrap procedure by randomly removing receptor subsets
199 of different sizes (from 1 to 71 receptors). The resulting AUCs range from 0.777 to 0.801, and
200 show that from a subset size of five receptors, the resulting prediction accuracy no longer
201 significantly varies (risk of wrongly rejecting the equality of two variances (F-test) >5%), and
202 is thus robust to the nature of the receptor subset (Fig. S2). All together, these results indicate
203 that energy maps are specific to ligand families and that homology relationships between
204 protein ligands can be detected solely on the basis of the comparison of their energy maps.
205 Interestingly, this observation holds for families displaying important sequence variations
206 (Table S1). For example, the AUC computed for the UCH and ubiquitin-related families are
207 0.95 and 0.88 respectively despite the fact that the average sequence identity of these families
208 does not exceed 45% (Fig. S3 and Table S1). This indicates that energy maps are similar even
209 for homologous pairs displaying large sequence variations.
210 We then specifically investigate the similarity of the energy maps produced by ligands
211 belonging to a same family in order to see whether some ligands behave energetically
212 differently from the rest of their family members. On the 72 ligands, only five (2L7R_A,
213 4BNR_A, 1BZX_A, 1QA9_A, 1YAL_B) display energy maps that are significantly different
214 from those of their related homologs (Z-tests *p-values* for the comparison of the averaged
215 distance of each ligand with their homologs versus the averaged distance of all ligands with

216 their homologous ligands ≤ 5%). In order to identify the factors leading to differences
217 between energy maps involving homologous ligands, we computed the pairwise sequence
218 identity and the root mean square deviation (RMSD) between the members of each family.
219 Interestingly, none of these criteria can explain the energy map differences observed within
220 families (Fisher test *p* of the linear model estimated on all protein families >0.1) (see Fig. 6*B*-
221 *C* for the ubiquitin-related family, Fig. S4-S14 *B-C* for the other families, and Table S2 for
222 details). Fig. 6*A* represents a subsection of the ADM for the ubiquitin-related family (i.e. the
223 energy map distances computed between all the members of the ubiquitin-related family and
224 averaged over the 72 receptors). Low distances reflect pairs of ligands with similar energy
225 behaviors (i.e. producing similar energy maps when interacting with the same receptor) while
226 high distances reveal pairs of ligands with distant energy behaviors. 2L7R_A distinguishes
227 itself from the rest of the family, displaying high-energy map distances with all of its
228 homologs. RMSD and sequence identity contribute modestly to the energy map distances
229 observed in Fig. 6*A* (Spearman correlation test $p^{RMSD} = 0.01$ and $p^{seq} = 0.04$ (Table S2 , Fig.
230 6*B-C*)). Fig. 6*D* shows a projection of the contribution from the electrostatic term in the
231 energy function of ATTRACT on the surface of the seven ubiquitin-related family members
232 (for more details, see Fig. S15 and *Material and Methods*). Fig. 6*E* represents the electrostatic
233 maps distances computed between all members of the family. 2L7R_A stands clearly out,
234 displaying a negative electrostatic potential over the whole surface while its structural
235 homologs harbor a remarkable fifty-fifty electrostatic distribution (Fig. 6*D*). The negatively
236 charged surface of 2L7R_A is explained by the absence of the numerous lysines that are
237 present in the others members of the family (referred by black stars, Fig. 6*D*). Lysines are
238 known to be essential for ubiquitin function by enabling the formation of polyubiquitin chains
239 on target proteins. Among the seven lysines of the ubiquitin, K63 polyubiquitin chains are
240 known to act in non-proteolytic events while K48, K11, and the four other lysines

241 polyubiquitin chains are presumed to be involved into addressing proteins to the proteasome
242 (Xu et al., 2009). 2L7R_A is a soluble UBL domain resulting from the cleavage of the fusion
243 protein FAU (Welchman et al., 2005). Its function is unrelated to proteasomal degradation,
244 which might explain the lack of lysines on its surface and the differences observed in its
245 energy maps. Interestingly, the differences observed for the energy maps of 1YAL_B
246 (Papain-like family) (Fig. S4) and 4BNR_A (eukaryotic proteases family) (Fig. S5) regarding
247 their related homologs can be explained by the fact that they both display a highly charged
248 surface. These two proteins are thermo-stable (Molnár et al., 2013; Sumner et al., 1993),
249 which is not the case for their related homologs, and probably explains the differences
250 observed in their relative energy maps. The V-set domain family is split into two major
251 subgroups according to their averaged energy map distances (Fig. S6*A*). The first group
252 corresponds to CD2 proteins (1QA9_A and its unbound form 1HNF_A) and differs
253 significantly from the second group (Z-test $p = 0.03$ and $p = 0.05$ respectively). The second
254 group corresponds to CD58 (1QA9_B and its unbound form 1CCZ_A) and CD48 proteins
255 (2PTT_A). Interestingly, CD2 is known to interact with its structural homologs (namely
256 CD58 and CD48) through an interface with a striking electrostatic complementarity (Wang et
257 al., 1999). The two subgroups have thus evolved distinct and specific binding sites to interact
258 together. They should have different interaction propensities which probably impact their
259 energy maps. These five cases illustrate the capacity of our theoretical framework to reveal
260 functional or biophysical specificities of homologous proteins that could not be revealed by
261 classical descriptors such as RMSD or sequence identity.
262 The AUC of 0.8 calculated previously with energy maps produced by the docking of either
263 native-related or arbitrary pairs indicates that energy maps are specific to ligand families. To
264 see whether this observation is not mainly due to the native-related pairs, we repeated the
265 previous test while removing that time all energy maps computed with native-related pairs

266 and calculated the resulting ADM. We then measured our ability to retrieve the homologs of
267 each ligand by calculating the ROC curve as previously. The resulting AUC is still equal to
268 0.8, revealing that our ability to identify a ligand's homologs is independent from the fact that
269 the corresponding energy maps were computed with native-related or arbitrary pairs (Fig. 5).
270 This shows that the energy maps are specific to protein families whether the docked pairs are
271 native-related or not. Consequently, the propensity of the whole protein surface to interact
272 with a given ligand is conserved and specific to the ligand family whether the ligand is native-
273 related or not. This striking result may reflect both positive and negative design operating on
274 protein surfaces to maintain functional interactions and to limit random interactions that are
275 inherent to a crowded environment.
276
277 **The interaction propensity of all surface regions of a receptor is evolutionary conserved**
278 **for homologous ligands**
279 To see whether some regions contribute more to the specificity of the maps produced by
280 homologous ligands, we next dissected the effective contribution of the surface regions of the
281 receptor defined according to their docking energy value, in the identification of ligand's
282 homologs. We discretized the energy values of each energy map into five categories, leading
283 to a palette of five energy classes (or colors) (see Fig. 2*D* and *Materials and Methods*). These
284 five-color maps highlight low-energy regions (i.e. hot regions in red), intermediate-energy
285 regions (i.e. warm, lukewarm and cool regions in orange, light-green and dark-green
286 respectively) and high-energy regions (i.e. cold regions in blue). We first checked that the
287 discretization of the energy maps does not affect our ability to identify the homologs of each
288 of the 72 ligands from the comparison of their five-colors maps. The resulting AUC is 0.78
289 (Table 1)**,** showing that the discretization step does not lead to an important loss of
290 information.

291 Then, we evaluated the contribution of each of the five energy classes separately in the
292 ligand's homologs identification by testing our ability to retrieve the homologs of the 72
293 ligands from their one-color energy maps (either red, orange, yellow, green or blue) (see Fig.
294 1*E* and *Materials and Methods*). Table 1 shows the resulting AUCs. Interestingly, the
295 information provided by each energy class taken separately is sufficient for discriminating the
296 homologs of a given ligand from the rest of the dataset (Table 1). The resulting AUCs range
297 from 0.77 to 0.79 for the orange, light green, dark green, and blue classes and are comparable
298 to those obtained with all classes taken together (0.78). This shows that (i) warm, lukewarm,
299 cool, and cold regions alone are sufficient to retrieve homology relationships between ligands
300 and (ii) the localization of a given energy class is specific to protein families. Hot regions are
301 less discriminative and lead to an AUC of 0.74. In order to see how regions corresponding to
302 a specific energy class are distributed over a receptor surface, we summed the 72
303 corresponding one-color maps into a stacked map (Fig. 8-supplement 1 – see Materials and
304 Methods for more details). For each color, the resulting stacked map reflects the tendency of a
305 map cell to belong to the corresponding energy class. Fig. 8 shows an example of the five
306 stacked maps (i.e. for cold, cool, lukewarm, warm and hot regions) computed for the receptor
307 1P9D_U. Intermediates regions (i.e. warm, lukewarm and cool regions) are widespread on the
308 stacked map while cold and hot regions are localized on few small spots (three and one
309 respectively) no matter the nature of the ligand. Fig. 8-supplement 2, shows for the receptor
310 1P9D_U, the 12 blue and red stacked maps computed for each ligand family separately. We
311 can see that some cold spots are specific to ligand families and that their area distribution is
312 specific to families while all 12 ligand families display the same hot spot in the map's upper-
313 right quadrant. These observations can be generalized to each receptor. On average,
314 intermediate regions are widespread on the stacked maps and cover respectively 742, 1142
315 and 609 cells for cool, lukewarm and warm regions, while cold and hot regions cover no more

316 than respectively 105 and 106 cells respectively (Fig. S16). Interestingly, hot regions are more
317 colocalized than cold ones and are restricted to 2 distinct spots on average per stacked map,
318 while cold regions are spread on 3.6 spots on average (t-Test p = 9.61e-12). These results
319 show that ligands belonging to different families tend to dock preferentially on the same
320 regions and thus lead to similar hot region distributions on the receptor surface. This
321 observation recalls those made by *Fernandez-Recio et al*. (Fernandez-Recio et al., 2004), who
322 showed that docking random proteins against a single receptor leads to an accumulation of
323 low-energy solutions around the native interaction site and suggested that different ligands
324 will bind preferentially on the same localization.
325 We can hypothesize that hot regions present universal structural and biochemical features that
326 make them more prone to interact with other proteins. To test this hypothesis, we computed
327 for each protein of the dataset, the 2-D projection of three protein surface descriptors (see
328 *Materials and Methods* and Fig. S15): the Kyte-Doolittle (KD) hydrophobicity (Kyte and
329 Doolittle, 1982), the circular variance (CV) (Mezei, 2003) and the stickiness (Levy et al.,
330 2012). The CV measures the density of protein around an atom and is a useful descriptor to
331 reflect the local geometry of a surface region. CV values are comprised between 0 and 1. Low
332 values reflect protruding residues and high values indicate residues located in cavities.
333 Stickiness reflects the propensity of amino acids to be involved in protein-protein interfaces
334 (Levy et al., 2012). It has been calculated as the log ratio of the residues frequencies on
335 protein surfaces versus their frequencies in protein-protein interfaces. For each receptor, we
336 calculated the correlation between the docking energy and the stickiness, hydrophobicity or
337 CV over all cells of the corresponding 2D-maps. We found a significant anti-correlation
338 between the docking energy and these three descriptors (correlation test p between docking
339 energies and respectively stickiness, hydrophobicity and CV < 2.2e-16, see Table S3)). Fig. 7
340 represents the boxplots of the stickiness, hydrophobicity and CV of each energy class (see

341 Fig. S15 and *Materials and Methods* section for more details). We observe a clear effect of
342 these factors on the docking energy: cold regions (i.e. blue class) are the less sticky, the less
343 hydrophobic and the most protruding while hot ones (i.e. red class) are the most sticky, the
344 most hydrophobic and the most planar (Tukey HSD test (Tukey, 1949), p of the differences
345 observed between each energy classes < 2.2e-16). One should notice that stickiness has been
346 defined from a statistical analysis performed on experimentally characterized protein
347 interfaces and therefore between presumed native partners. The fact that docking energies
348 (physics-based) calculated either between native-related or arbitrary partners is anti-correlated
349 with stickiness (statistics-based) defined on native interfaces, strengthens strongly the concept
350 of stickiness as the propensity of interacting promiscuously and provides physics-based pieces
351 of evidence for sticky regions as a proxy for promiscuous interactions.
352 We show that not only the area distribution on a receptor surface of hot regions but also those
353 of intermediate and cold regions are similar for homologous ligands and are specific to ligand
354 families (AUC ranging from 0.74 to 0.79) whether the ligands are native-related or not. This
355 tendency is even stronger for intermediate and cold regions. Interestingly, the information
356 contained in the cold regions that cover on average no more than 4.9% of the energy maps is
357 sufficient to identify homology relationships between ligands.

358 **Discussion**

359 In this study, we address the impact of both positive and negative design on thousands of
360 interaction energy landscapes by the mean of a synthetic and efficient representation of the
361 docking energy landscapes: two-dimensional energy maps that reflect the interaction
362 propensity of the whole surface of a protein (namely the receptor) with a given partner
363 (namely the ligand). We show that all regions of the energy maps, including cold,
364 intermediate and hot regions are similar for homologous ligands and are specific to ligand
365 families whether the ligands are native-related or arbitrary. This reveals that the interaction
366 propensity of the whole surface of proteins is constrained by functional and non-functional
367 interactions, reflecting both positive and negative design operating on whole protein surfaces,
368 thus shaping the interaction energy landscapes of functional partners and random encounters.
369 While most studies that aim at depicting protein-protein interactions focus on native binding
370 sites of proteins (Andreani et al., 2012; Chakrabarti and Janin, 2002; Keskin et al., 2005; Li et
371 al., 2004; Lo Conte et al., 1999; Pechmann et al., 2009), we bring a new regard on protein-
372 protein interactions by providing a systematic and physical characterization of all regions of
373 the surface of a protein in interaction with a given ligand (i.e. cold, intermediate and hot
374 regions). Here, we address the energy behavior of not only known binding sites, but also of
375 the rest of the protein surface, which plays an important role in protein interactions by
376 constantly competing with the native binding site. We show that the interaction propensity of
377 the rest of the surface is not homogeneous and displays regions with different binding
378 energies that are specific to ligand families. This probably reflects the negative design
379 operating on these regions to limit non-functional interactions. We can hypothesize that these
380 regions participate to favor functional assemblies (i.e. functional assembly modes with
381 functional partners) over non-functional ones and are thus evolutionary constrained by non-
382 functional assemblies. On the other side, we show that hot regions are very localized (4.9% of

383 the cells of an energy map) and tend to be similar no matter the ligand. Hot regions are likely
384 to display universal properties of binding, i.e. they are more hydrophobic and more planar,
385 and thus more "sticky" than the other regions. They provide a non-specific binding patch that
386 is suitable for many ligands. Despite the fact that hot regions are similar for functional and
387 arbitrary ligands, we can hypothesize that functional partners have evolved to optimize their
388 binding site (positive design) so that functional interactions prevail over non-functional
389 competing ones. Indeed, we have previously shown that the docking of native partners lead to
390 more favorable binding energies than the docking of non-native partners when the ligand is
391 constrained to dock around the receptor's native binding site (Lopes et al., 2013; Sacquin-
392 Mora et al., 2008). All these results suggest a new physical model of protein surfaces where
393 protein surface regions, in the crowded cellular environment, serve as a proxy for regulating
394 the competition between functional and non-functional interactions. In this model,
395 intermediate and cold regions play an important role by preventing non-functional assemblies
396 and by guiding the interaction process towards functional ones and hot regions may select the
397 functional assembly among the competing ones through optimized interfaces with the native
398 partner.
399 In this work, we used and extended the application of the 2-D energy map representation
400 developed in (Fernandez-Recio et al., 2004) to develop an original theoretical framework that
401 enables the efficient, automated and integrative analysis of different protein surface features.
402 2-D maps provide the area distribution of a given feature on the whole protein surface and
403 their discretization enables to specifically study a given property (e.g. protruding, planar,
404 sticky, positively charged, cold or hot regions for example). They are easy to manipulate and
405 their straightforward comparison enables (i) to study relationships between different
406 properties with the comparison of their area distributions on a protein surface and (ii) to shed
407 light on the evolutionary constraints exerted on a given feature by comparing its area

408 distribution on homologous protein surfaces. Particularly, this enables to highlight and
409 characterize hot regions on a protein surface which can be either specific or conserved for all
410 ligands and opens up new possibilities for the development of novel methods for protein
411 binding sites prediction and classification as functional or promiscuous in the continuity of
412 previous developments based on arbitrary docking (Fernandez-Recio et al., 2004; Lopes et al.,
413 2013; Sacquin-Mora et al., 2008; Vamparys et al., 2016).
414 Our framework provides a proxy for further protein functional characterization as shown with
415 the five proteins discussed in the *Results* section *Energy maps are specific to protein families*.
416 The comparison of their respective energy maps enables to reveal biophysical and functional
417 properties that could not be revealed with classical monomeric descriptors such as RMSD or
418 sequence identity. Indeed, our framework highlights the energy behavior of a protein
419 interacting with a subset of selected partners either functional or arbitrary, thus embedding
420 functional and systemic properties of proteins. This work goes beyond the classical use of
421 binary docking to provide a systemic point of view of protein interactions**,** for example by
422 exploring the propensity of a protein to interact with hundreds of selected ligands, and thus
423 addressing the behavior of a protein in a specific cellular environment. Particularly, exploring
424 the dark interactome (i.e. non-functional assemblies and interactions with non-functional
425 partners) can provide a wealth of valuable information to understand mechanisms driving and
426 regulating protein-protein interactions. Precisely, our 2D energy maps based strategy enables
427 to explore it in an efficient and automated way.

Materials and Methods

Protein dataset

The dataset comprises 72 protein structures divided into 12 families of structural homologs (Table S1). Each family is related to at least one other family (its native-related partners family) through a pair of interacting proteins (except the V set domain family: the two native partners are homologs and belong to the same family) (Fig. S1). Each family is composed of a monomer selected from the protein-protein docking benchmark 5.0 (Vreven et al., 2015) in its bound and unbound forms, which is called the master protein. Each master protein has a native partner (for which the 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 (Faure et al., 2012) according to the following criteria: (i) experimental structure resolution lower 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. As we are limited by the number of available structures of interologs, we completed families with unbound monomers homologous to the master following the same criteria. We consider that all members of a family correspond to native-related partners of all members of their native-related partner family. To address the impact of conformational changes of a protein on its interaction energy maps, we added different NMR conformers. We show that energy map involving pairs of conformers are significantly more similar than those obtained for other pairs of homologous ligands (unilateral Wilcoxon test, $p < 2.2e-16$) showing that the conformational changes in a protein (lower than 3Å) have a low impact on the resulting energy maps (Fig. S17).

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

455 A complete cross-docking experiment was realized with the ATTRACT software (Fiorucci
456 and Zacharias, 2010) on the 72 proteins of the dataset, leading to 5184 (72 x 72) docking
457 calculations (Fig 1*A*). ATTRACT uses a coarse-grain reduced protein representation and a
458 simplified energy function comprising a pseudo Lennard-Jones term and an electrostatic term.
459 The calculations took approximately 20000 hours on a 2.7GHz processor. Prior to docking
460 calculations, all PDB structures were prepared with the DOCKPREP software (Lang et al.,
461 2009).

462 During a docking calculation, the ligand L_i explores exhaustively the surface of the receptor
463 R_k (whose position is fixed during the procedure), sampling and scoring thousands of different
464 ligand docking poses (between 10000 and 50000 depending of the sizes of the proteins
465 involved) (Fig. 2*A*). For each protein couple R_k-L_i, a 2-D energy map is computed which
466 shows the distribution of the energies of all docking solutions over the receptor surface. To
467 compute these maps, for all docking poses, the spherical coordinates (ϕ, θ) (with respect to
468 the receptor center of mass (CM)) of the ligand CM are represented onto a 2-D map in an
469 equal-area 2-D sinusoidal projection (Fig. 2*B*) (see (Fernandez-Recio et al., 2004) for more
470 details). Each couple of coordinates (ϕ, θ) is associated with the energy of the corresponding
471 docking conformation (Fig. 2*B*). A continuous energy map is then derived from the discrete
472 one, where the map is divided into a grid of 36 x 72 cells. Each cell represents the same
473 surface and, depending on the size of the receptor, can span from 2.5 $Å^2$ to 13$Å^2$. For each
474 cell, all solutions with an energy score below 2.7 kcal/mol^{-1} from the lowest solution of the
475 cell are retained, according to the conformations filtering protocol implemented in (Lopes et
476 al., 2013). The average of the retained energy scores is then assigned to the cell. If there is no
477 docking solution in a cell, a score of 0 is assigned to it. Finally, the energies of the cells are

478 smoothed, by averaging the energy values of each cell and of the eight surrounding neighbors
479 (Fig. 2*C*).
480 For each map, the energy values are discretized into five energy classes of same range leading
481 to a discrete five-colors energy map (Fig. 2*D*). The range is calculated for each energy map
482 and spans from the minimum to the maximum scores of the map cells. The range of the
483 energy classes of the map R_k-L_i is equal to (maxE – minE)/5, where maxE and minE
484 correspond to the maximal and minimal energy values in the R_k-L_i map. Each five-colors
485 energy map is then split into five one-color maps, each one representing an energy class of the
486 map (Fig. 2*E*). The continuous, five-colors and one-color energy maps are calculated for the
487 5184 energy maps (Fig. 1*B*).
488
489 **Comparison of energy maps and identification of ligand's homologs**
490 For each receptor R_k, we computed a 72x72 energy map distance (EMD) matrix where each
491 entry (*i*,*j*) corresponds to the pairwise distance between the energy maps R_k-L_i and R_k-L_j
492 resulting from the docking of the ligands L_i and L_j on the receptor R_k (Fig. 1*C*). The pairwise
493 distance $d_{Man}(R_k\text{-}L_i, R_k\text{-}L_j)$ between the energy maps is calculated with a Manhattan distance
494 according to

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

495
496 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
497 and R_k-L_j respectively. Low distances reflect pairs of ligands that induce similar energy maps
498 when they are docked on the same receptor. The procedure is repeated for each receptor of the
499 database leading to 72 EMD matrices. The 72 EMD matrices are averaged into an averaged
500 distances matrix (ADM). Each entry (*i*,*j*) of the ADM reflects the similarity of the R_k-L_i and
501 R_k-L_j energy maps averaged over all the receptors R_k in the dataset (Fig. 1*D*). In order to test

502 to what extent family members display similar energy maps when they are docked with the
503 same receptor, we tested our ability to correctly identify the homologs of the 72 ligands from
504 the only comparison of its energy maps with those of the other ligands. Because, energy maps
505 are receptor-centered, we cannot compare the energy maps computed for two unrelated
506 receptors. The procedure is thus ligand-centered and consists in the comparison of energy
507 maps produced with different ligands docked with a same receptor. Two ligands (i,j) are
508 predicted as homologs according to their corresponding distance (i,j) in the ADM. Values
509 close to zero should reflect homologous ligand pairs, while values close to one should reflect
510 unrelated ligand pairs. A Receiver Operating Characteristic (ROC) curve and its Area Under
511 the Curve (AUC) are computed on the ADM. True positives (TP) are all the homologous
512 ligand pairs and predicted as such, true negatives (TN) are all the unrelated ligand pairs and
513 predicted as such. False positives (FP) are unrelated ligand pairs but incorrectly predicted as
514 homologous pairs. False negatives (FN) are homologous ligand pairs but incorrectly predicted
515 as unrelated pairs. ROC curves and AUC values were calculated with the R package pROC
516 (Robin et al., 2011). The ligand's homologs identification was also realized using the five-
517 color energy maps or the one-color energy maps taken separately. The five energy class
518 regions display very different sizes, with median ranging from 65 cells for the red and blue
519 regions to 630 cells for the yellow one. To prevent any bias due to the size of the different
520 classes, we normalized the Manhattan distance by the size of the regions compared in the
521 map. The rest of the procedure is the same than those used for continuous energy maps (Fig.
522 1).
523 To visualize the area distribution of the regions of a given energy class for all ligands on the
524 receptor surface, the 72 corresponding one-color maps are summed into a stacked map where
525 each cell's intensity varies from 0 to 72 (Fig. 8-supplement 1). To remove background-image
526 from these maps, i.e. cells with low intensity (intensity < 17) and the areas of small size (< 4

527 cells), we used a Dirichlet process mixture model simulation for image segmentation (R
528 package *dpmixsim*) (Ferreira da Silva, 2007).
529
530 **2-D projection of monomeric descriptors of protein surfaces**
531 We computed KD hydrophobicity (Kyte and Doolittle, 1982), stickiness (Levy et al., 2012),
532 CV maps of each protein of the dataset, in order to compare their topology with the energy
533 maps. Prior to all, proteins belonging to the same families were structurally aligned with TM-
534 align (Zhang and Skolnick, 2005) in order to place them in the same reference frame, making
535 their maps comparable. Particles were generated around the protein surface with a slightly
536 modified Shrake-Rupley algorithm (Saladin et al., 2009). The density of spheres is fixed at
537 1Å^2, representing several thousands particles per protein. Each particle is located at 5Å from
538 the surface of the protein. The CV, stickiness and KD hydrophobicity values of the closest
539 atom of the protein are attributed to each particle. We also generated electrostatic maps
540 reflecting the distribution of the contribution of the coulombic term as encoded in the
541 ATTRACT force field on a protein surface. The procedure is slightly different: each particle *i*
542 has a +1 positive charge, and receives the coulombic value:

$$Q_i = \sum_{j=1}^{n} q_i q_j / \mathcal{E} r_{ij}$$

543
544 with *n* the number of pseudo-atom in the protein, q_i the charge of the particle, q_j the charge of
545 the pseudo-atom *j*, r_{ij} the distance between the particle *i* and the pseudo-atom *j*, and ε a
546 distant-dependent dielectric constant (ε = 15r_{ij}). CV was calculated following the protocol
547 described in (Mezei, 2003) on the all-atom structures. Stickiness, electrostatics and
548 hydrophobicity were calculated on ATTRACT coarse-grain models. Pseudo-atom charges are
549 defined according to the ATTRACT force field (Fiorucci and Zacharias, 2010). After
550 attributing a value to each particle, the position of their spherical coordinates is represented in

551 a 2-D sinusoidal projection, following the same protocol as described in Fig. 2 and *Materials*
552 *and Methods* section *Docking experiment and construction of energy maps*. The map is then
553 smoothed following the protocol in Fig. 2.

Acknowledgment

We thank F. Fraternali, R. Guerois, E. Laine, M. Montes for their constructive comments on the manuscript. SSM work is supported by the “Initiative d’Excellence” program from the French State (Grant “DYNAMO”, ANR-11-LABX-0011-01). HS work is supported by a French government fellowship. This work was supported by GENCI-CINES (grant No. x2016077460).

Aloy P, Ceulemans H, Stark A, Russell RB. The relationship between sequence and interaction divergence in proteins. J Mol Biol 2003;332:989–98.

Andreani J, Faure G, Guerois R. Versatility and invariance in the evolution of homologous heteromeric interfaces. PLoS Comput Biol 2012;8:e1002677. doi:10.1371/journal.pcbi.1002677.

Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, et al. The Protein Data Bank. Nucleic Acids Res 2000;28:235–42.

Chakrabarti P, Janin J. Dissecting protein–protein recognition sites. Proteins: Structure, Function, and Bioinformatics 2002;47:334–343.

Deeds EJ, Ashenberg O, Gerardin J, Shakhnovich EI. Robust protein protein interactions in crowded cellular environments. Proc Natl Acad Sci USA 2007;104:14952–7. doi:10.1073/pnas.0702766104.

Drummond DA, Wilke CO. Mistranslation-induced protein misfolding as a dominant constraint on coding-sequence evolution. Cell 2008;134:341–52. doi:10.1016/j.cell.2008.05.042.

Faure G, Andreani J, Guerois R. InterEvol database: exploring the structure and evolution of protein complex interfaces. Nucleic Acids Res 2012;40:D847-856. doi:10.1093/nar/gkr845.

Fernandez-Recio J, Totrov M, Abagyan R. Identification of protein–protein interaction sites from docking energy landscapes. Journal of Molecular Biology 2004;335:843–865.

Ferreira da Silva AR. A Dirichlet process mixture model for brain MRI tissue classification. Med Image Anal 2007;11:169–82. doi:10.1016/j.media.2006.12.002.

Fiorucci S, Zacharias M. Binding site prediction and improved scoring during flexible protein–protein docking with ATTRACT. Proteins: Structure, Function, and Bioinformatics 2010;78:3131–3139. doi:10.1002/prot.22808

Fox NK, Brenner SE, Chandonia J-M. SCOPe: Structural Classification of Proteins--extended, integrating SCOP and ASTRAL data and classification of new structures. Nucleic Acids Res 2014;42:D304-309. doi:10.1093/nar/gkt1240.

Garcia-Seisdedos H, Empereur-Mot C, Elad N, Levy ED. Proteins evolve on the edge of supramolecular self-assembly. Nature 2017;548:244–7. doi:10.1038/nature23320.

Garzón JI, Deng L, Murray D, Shapira S, Petrey D, Honig B. A computational interactome and functional annotation for the human proteome. ELife Sciences 2016;5:e18715. doi:10.7554/eLife.18715.

Heo M, Maslov S, Shakhnovich E. Topology of protein interaction network shapes protein abundances and strengths of their functional and nonspecific interactions. Proc Natl Acad Sci USA 2011;108:4258–63. doi:10.1073/pnas.1009392108.

Janin J, Bahadur RP, Chakrabarti P. Protein-protein interaction and quaternary structure. Q Rev Biophys 2008;41:133–80. doi:10.1017/S0033583508004708.

Karanicolas J, Corn JE, Chen I, Joachimiak LA, Dym O, Peck SH, et al. A de novo protein binding pair by computational design and directed evolution. Mol Cell 2011;42:250–60. doi:10.1016/j.molcel.2011.03.010.

Keskin O, Ma B, Nussinov R. Hot regions in protein--protein interactions: the organization and contribution of structurally conserved hot spot residues. J Mol Biol 2005;345:1281–94. doi:10.1016/j.jmb.2004.10.077.

Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol 1982;157:105–32.

Lang PT, Brozell SR, Mukherjee S, Pettersen EF, Meng EC, Thomas V, et al. DOCK 6: combining techniques to model RNA-small molecule complexes. RNA 2009;15:1219–30. doi:10.1261/rna.1563609.

Levy ED, De S, Teichmann SA. Cellular crowding imposes global constraints on the chemistry and evolution of proteomes. Proc Natl Acad Sci USA 2012;109:20461–6. doi:10.1073/pnas.1209312109.

Levy ED, Kowarzyk J, Michnick SW. High-resolution mapping of protein concentration reveals principles of proteome architecture and adaptation. Cell Rep 2014;7:1333–40. doi:10.1016/j.celrep.2014.04.009.

Li X, Keskin O, Ma B, Nussinov R, Liang J. 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 2004;344:781–795. doi:10.1016/j.jmb.2004.09.051

Lo Conte L, Chothia C, Janin J. The atomic structure of protein-protein recognition sites. J Mol Biol 1999;285:2177–98.

Lopes A, Sacquin-Mora S, Dimitrova V, Laine E, Ponty Y, Carbone A. Protein-protein interactions in a crowded environment: an analysis via cross-docking simulations and evolutionary information. PLoS Comput Biol 2013;9:e1003369. doi:10.1371/journal.pcbi.1003369.

McGuffee SR, Elcock AH. Diffusion, crowding & protein stability in a dynamic molecular model of the bacterial cytoplasm. PLoS Comput Biol 2010;6:e1000694. doi:10.1371/journal.pcbi.1000694.

Mezei M. A new method for mapping macromolecular topography. Journal of Molecular Graphics and Modelling 2003;21:463–72. doi:10.1016/S1093-3263(02)00203-6.

Milo R. What is the total number of protein molecules per cell volume? A call to rethink some published values. Bioessays 2013;35:1050–5. doi:10.1002/bies.201300066.

Molnár T, Vörös J, Szeder B, Takáts K, Kardos J, Katona G, et al. Comparison of complexes formed by a crustacean and a vertebrate trypsin with bovine pancreatic trypsin inhibitor - the key to achieving extreme stability? FEBS J 2013;280:5750–63. doi:10.1111/febs.12491.

Nobeli I, Favia AD, Thornton JM. Protein promiscuity and its implications for biotechnology. Nature Biotechnology 2009;27:157–67. doi:10.1038/nbt1519.

Nooren IMA, Thornton JM. Diversity of protein–protein interactions. The EMBO Journal 2003;22:3486–92. doi:10.1093/emboj/cdg359. doi:10.1093/emboj/cdg359

Ohue M, Matsuzaki Y, Shimoda T, Ishida T, Akiyama Y. Highly precise protein-protein interaction prediction based on consensus between template-based and de novo docking methods. BMC proceedings, vol. 7, BioMed Central; 2013, p. S6. doi :10.1186/1753-6561-7-S7-S6

Pál C, Papp B, Hurst LD. Highly expressed genes in yeast evolve slowly. Genetics 2001;158:927–31.

Pechmann S, Levy ED, Tartaglia GG, Vendruscolo M. Physicochemical principles that regulate the competition between functional and dysfunctional association of proteins. Proc Natl Acad Sci USA 2009;106:10159–64. doi:10.1073/pnas.0812414106.

Pierce BG, Hourai Y, Weng Z. Accelerating protein docking in ZDOCK using an advanced 3D convolution library. PLoS ONE 2011;6:e24657. doi:10.1371/journal.pone.0024657.

Richardson JS, Richardson DC. Natural beta-sheet proteins use negative design to avoid edge-to-edge aggregation. Proc Natl Acad Sci USA 2002;99:2754–9. doi:10.1073/pnas.052706099.

Ritchie DW, Venkatraman V. Ultra-fast FFT protein docking on graphics processors. Bioinformatics 2010;26:2398–405. doi:10.1093/bioinformatics/btq444.

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

Robinson CV, Sali A, Baumeister W. The molecular sociology of the cell. Nature 2007;450:973–82. doi:10.1038/nature06523.

Sacquin-Mora S, Carbone A, Lavery R. Identification of protein interaction partners and protein–protein interaction sites. Journal of Molecular Biology 2008;382:1276–1289. doi :10.1016/j.jmb.2008.08.002

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

Schavemaker PE, Śmigiel WM, Poolman B. Ribosome surface properties may impose limits on the nature of the cytoplasmic proteome. Elife 2017;6. doi:10.7554/eLife.30084.

Snedecor GW, Cochran WC. Statistical methods. lowa state university press, Ames. Iowa N Vadivukkarasi et Al 1989.

Sumner IG, Harris GW, Taylor MA, Pickersgill RW, Owen AJ, Goodenough PW. Factors effecting the thermostability of cysteine proteinases from Carica papaya. Eur J Biochem 1993;214:129–34.

Tukey JW. Comparing Individual Means in the Analysis of Variance. Biometrics 1949;5:99–114. doi:10.2307/3001913.

Vamparys L, Laurent B, Carbone A, Sacquin-Mora S. Great interactions: How binding incorrect partners can teach us about protein recognition and function. Proteins 2016;84:1408–21. doi:10.1002/prot.25086.

Vreven T, Moal IH, Vangone A, Pierce BG, Kastritis PL, Torchala M, et al. Updates to the integrated protein–protein interaction benchmarks: docking benchmark version 5 and affinity benchmark version 2. Journal of Molecular Biology 2015;427:3031–3041. doi:10.1016/j.jmb.2015.07.016

Wang JH, Smolyar A, Tan K, Liu JH, Kim M, Sun ZY, et al. Structure of a heterophilic adhesion complex between the human CD2 and CD58 (LFA-3) counterreceptors. Cell 1999;97:791–803.

Wass MN, Fuentes G, Pons C, Pazos F, Valencia A. Towards the prediction of protein interaction partners using physical docking. Mol Syst Biol 2011;7:469. doi:10.1038/msb.2011.3.

Welchman RL, Gordon C, Mayer RJ. Ubiquitin and ubiquitin-like proteins as multifunctional signals. Nat Rev Mol Cell Biol 2005;6:599–609. doi:10.1038/nrm1700.

Xu P, Duong DM, Seyfried NT, Cheng D, Xie Y, Robert J, et al. Quantitative Proteomics Reveals the Function of Unconventional Ubiquitin Chains in Proteasomal Degradation. Cell 2009;137:133–45. doi:10.1016/j.cell.2009.01.041.

Yang J-R, Liao B-Y, Zhuang S-M, Zhang J. Protein misinteraction avoidance causes highly expressed proteins to evolve slowly. Proc Natl Acad Sci USA 2012;109:E831-840. doi:10.1073/pnas.1117408109.

Yu I, Mori T, Ando T, Harada R, Jung J, Sugita Y, et al. Biomolecular interactions modulate macromolecular structure and dynamics in atomistic model of a bacterial cytoplasm. Elife 2016;5. doi:10.7554/eLife.19274.

Zhang J, Maslov S, Shakhnovich EI. Constraints imposed by non-functional protein-protein interactions on gene expression and proteome size. Mol Syst Biol 2008;4:210. doi:10.1038/msb.2008.48.

Zhang Y, Skolnick J. TM-align: a protein structure alignment algorithm based on the TM-score. Nucleic Acids Res 2005;33:2302–9. doi:10.1093/nar/gki524.

560 **Figure Legends**

561

562 **Table 1.** The AUC are calculated from the ADM with the continuous energy maps (Fig. 2*C*),
563 the five-color energy maps (Fig. 2*D*) and the one-color energy maps (Fig. 2*E*) following the
564 protocol of Fig. 1*D-E*.

565

566 **Fig. 1.** Experimental Protocol. (*A*) A complete cross-docking calculation is realized on the
567 dataset of 72 protein structures. This represents 72 x 72 = 5184 docking calculations. (*B*) For
568 each docking calculation, an energy map is computed as well as its corresponding five-color
569 and one-color energy maps, with the procedure described in Fig. 2 and *Materials and*
570 *Methods*. (*C*) For each receptor, an energy map distance (EMD) matrix is computed,
571 representing the pairwise distances between the 72 energy maps resulting from the docking of
572 all ligands with this receptor. Each value (*i*,*j*) of the matrix represents the Manhattan distance
573 between the two energy maps, resulting of the docking of ligands *i* and *j* with the receptor. A
574 small distance indicates that the ligands *i* and *j* produce similar energy maps when docked
575 with this receptor. To prevent any bias from the choice of the receptor, the procedure is
576 repeated for each receptor of the database, leading to 72 EMD matrices. (*D*) An averaged
577 distance matrix (ADM) is computed by calculating the average of the 72 EMD matrices. Each
578 entry (*i*,*j*) of the resulting matrix is the mean of the (*i*,*j*) values of the 72 receptor distance
579 matrices. A low distance between two ligands indicates that, on average, these ligands induce
580 similar energy maps when docked with a same receptor. Inversely a large distance indicates
581 that, on average, these ligands induce dissimilar energy maps when docked with a same
582 receptor. (*E*) To assess the extent to which energy maps are specific to ligand families, a
583 receiver operating characteristic (ROC) curve and the Area Under this Curve (AUC) are
584 computed from the ADM.

585

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

595

596 **Fig. 3.** Subset of energy maps and of ligand docking poses for receptor 2AYN_A. (*A*)
597 Examples of maps for the receptor 2AYN_A (ubiquitin carboxyl-terminal hydrolase (UCH)
598 family) docked with the ligands 1XD3_B (native partner), 1NDD_B (homolog of the native
599 partner), 1YVB_A and 2NQD_B (false partners). The star indicates the localization of the
600 experimentally determined interaction site of the ubiquitin, the circle-cross indicates the
601 homodimerization site of 2AYN_A. (*B*) Centers of mass (CM) of the 20 best docking poses
602 obtained for each of the 72 ligands of the database docked with the receptors 2AYN_A.
603 Receptor protein is represented in cartoon (black), its native ligand and its homodimere are
604 represented in cartoon with transparency (red and black respectively). CMs of the ligands
605 belonging to the ubiquitin-related family are colored in red, CMs of the proteins belonging to
606 the papain-like family are colored in blue.

607

608 **Fig. 4.** Boxplots of energy map pairwise distances between homologous ligand pairs from
609 native-related partner families, homologous ligand pairs from arbitrary partner families and

610 random ligand pairs. For each receptor, we computed (i) the average of energy map distances
611 of pair of homologous ligands belonging to its native-related partner family(ies), (ii) the
612 average of energy map distances of pair of homologous ligands belonging to its non-native-
613 related partner families, and (iii) the average of energy map distances of random pairs. P-
614 values are calculated with an unilateral Wilcoxon test.
615
616 **Fig. 5.** Receiver operating characteristic (ROC) curve and its Area Under the Curve (AUC).
617 ROC are calculated on the averaged distance matrix (ADM) including either all pairs (blue) or
618 only arbitrary pairs (red) (see Fig. 1 for more details).
619
620 **Fig. 6.** Ubiquitin-related family. (*A*) Energy map distances matrix. It corresponds to the
621 subsection of the ADM for the ubiquitin-related family (for the construction of the ADM, see
622 Fig. 1*D*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair (*i*,*j*)
623 averaged over the 72 receptors of the dataset (for more details, see Fig. 1*D*). (*B*) Pairwise
624 sequence identity matrix between all members of the family. (*C*) Pairwise root mean square
625 deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps and
626 cartoon representations of the seven members of the family. An electrostatic map represents
627 the distribution of the electrostatic potential on the surface of a protein (for more details, see
628 Fig. S15 and *Materials and Methods*). On the electrostatic maps, lysines positions are
629 indicated by stars. Cartoon structures are colored according to the distribution of their
630 electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix
631 represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.
632
633 **Fig. 7.** Boxplots of three descriptors of the protein surface: (*A*) the stickiness values, (*B*) the
634 Kyte-Doolittle hydrophobicity and (*C*) the CV values, depending on the energy class. The

635 stickiness, hydrophobicity and CV values are calculated for each protein following the
636 protocol described in *Materials and Methods*. For each of these criteria, *p-values* between the
637 median values of two "successive" energy classes were computed using the Tukey HSD
638 statistical test (Tukey, 1949).
639
640 **Fig. 8.** Stacked maps of 1P9D_U after the filtering of cells with too low intensity and areas of
641 too small size. The protocol to generate stacked maps is presented in Fig. 8-supplement 1. (*A*)
642 Blue stacked map (i.e. stacked cold regions). (*B*) Dark green stacked map (i.e. stacked cool
643 regions). (*C*) Light green stacked map (i.e. stacked lukewarm regions). (*D*) Orange stacked
644 map (i.e. stacked warm regions). (*E*) Red stacked map (i.e. stacked hot regions). One should
645 notice that stacked maps of two different colors can overlap because a cell can be associated
646 to different energy classes depending on the docked ligands. Fig. 8-supplement 2 presents
647 blue and red stacked maps of 1P9D_U computed for each ligand family.
648
649
650 **Fig. 8-supplement 1.** Generation of stacked maps of a receptor. (*A*) Calculation of the 72
651 one-color maps (red ones in the example) of receptor #1. A value of one is associated to
652 colored cells while zero is assigned to white cells. (*B*) Sum of the 72 one-color maps into a
653 stacked map. Cell's intensity varies from 0 to 72 and corresponds to the number of time the
654 cell is colored over the 72 ligands. (*C*) Filtering of cells of low cell intensity (intensity < 17)
655 and areas of too small size (< 4 cells) with a Dirichlet process mixture model simulation for
656 image segmentation (Ferreira da Silva, 2007). The procedure is repeated for each color
657 stacked map.
658

659 **Fig. 8-supplement 2.** Blue and red stacked maps of 1P9D_U computed for each ligand
660 family. (*A-L*) We compute the one-color stacked map of each family as the sum of the one-
661 color maps resulting from the docking of each ligand of a same family with 1P9D_U.
662
663
664 **Supplementary Files**

665 Table S1. List of proteins in the dataset and their structural families
666
667 Table S2. Estimation of the effective contribution of the three descriptors (sequence identity, RMSD and
668 electrostatic distance) in the pairwise ADM distances observed for each ligand pair belonging to a same family
669 Table S3. Spearman's rank correlation between docking energy values and three descriptors of protein surfaces
670 Fig. S1. Interactions between structural families of the dataset. Interactions are symbolized by links between
671 families
672 Fig. S2. AUC values calculated on random subsets of receptor of different sizes
673 Fig. S3. Receiver operating characteristic (ROC) curve and Area Under this Curve (AUC) calculated for each
674 family
675 Fig. S4. Papain-like family
676 Fig. S5. Eukaryotic-proteases family
677 Fig. S6. V set domains family
678 Fig. S7. UCH-L family
679 Fig. S8. UCH family
680 Fig. S9. Ubiquitin activating enzymes family
681 Fig. S10. UBC-related family
682 Fig. S11. Kunitz (STI) inhibitors family
683 Fig. S12. Retrovirus capsid proteins family
684 Fig. S13. Cystatins family
685 Fig. S14. Cyclophilins family
686 Fig. S15. Generation of electrostatics, stickiness, hydrophobicity and circular variance (CV) maps
687 Fig. S16. Boxplots of the size (in number of cells) of each energy class for all stacked maps

Fig. S17. Boxplots of pairwise distances between ligand pairs of conformers or pairs of homologous ligands (i.e. non-conformers pairs)

Table 1. AUC obtained with different types of energy maps

type of map	continous energy maps	five-colors energy maps	red energy maps	orange energy maps	light green energy maps	dark green energy maps	blue energy maps
AUC	0.8	0.78	0.74	0.77	0.77	0.78	0.79

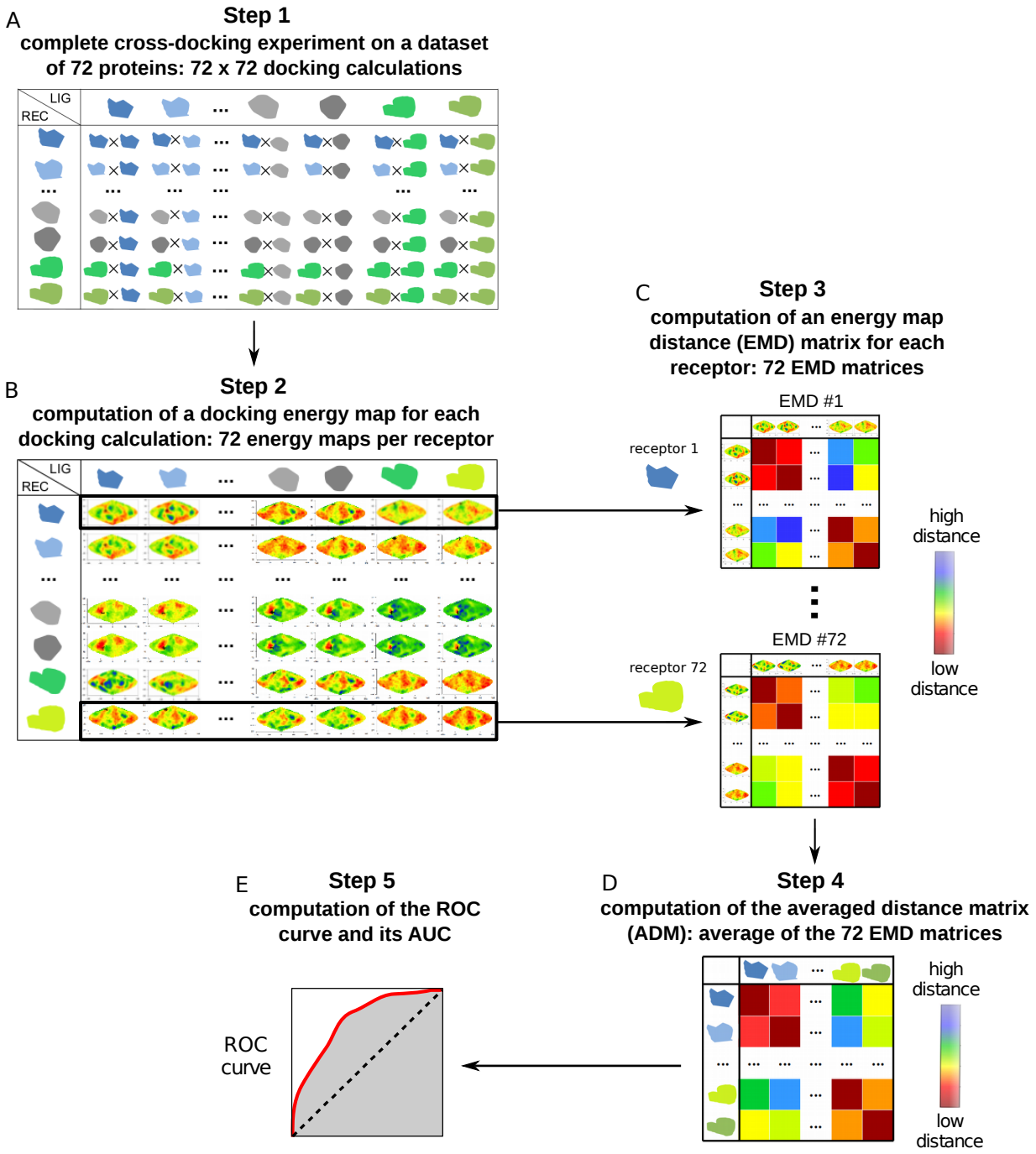
A
Step 1
complete cross-docking experiment on a dataset of 72 proteins: 72 x 72 docking calculations
LIG
REC
B
Step 2
computation of a docking energy map for each docking calculation: 72 energy maps per receptor
LIG
REC
C
Step 3
computation of an energy map distance (EMD) matrix for each receptor: 72 EMD matrices
EMD #1
receptor 1
EMD #72
receptor 72
high distance
low distance
D
Step 4
computation of the averaged distance matrix (ADM): average of the 72 EMD matrices
high distance
low distance
E
Step 5
computation of the ROC curve and its AUC
ROC curve

Fig. 1.

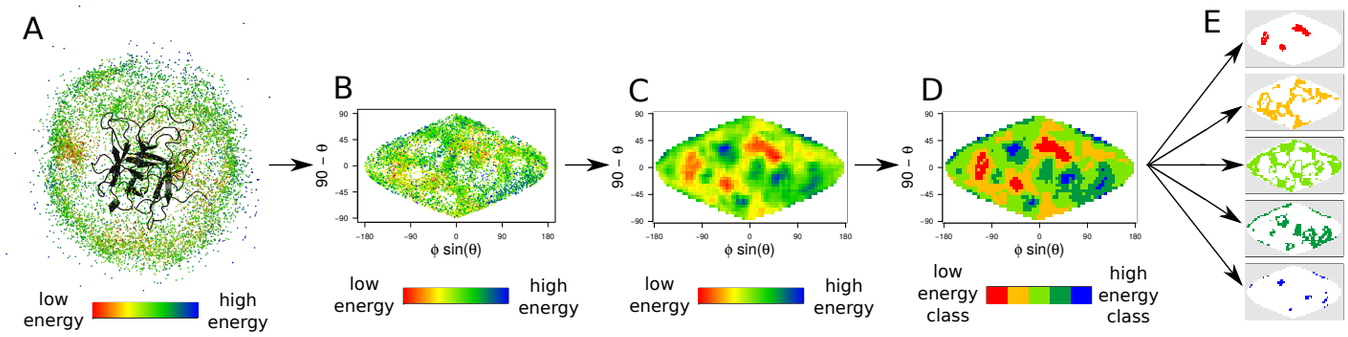
A
low energy
high energy
B
90 − θ
ϕ sin(θ)
low energy
high energy
C
90 − θ
ϕ sin(θ)
low energy
high energy
D
90 − θ
ϕ sin(θ)
low energy class
high energy class
E

Fig. 2.

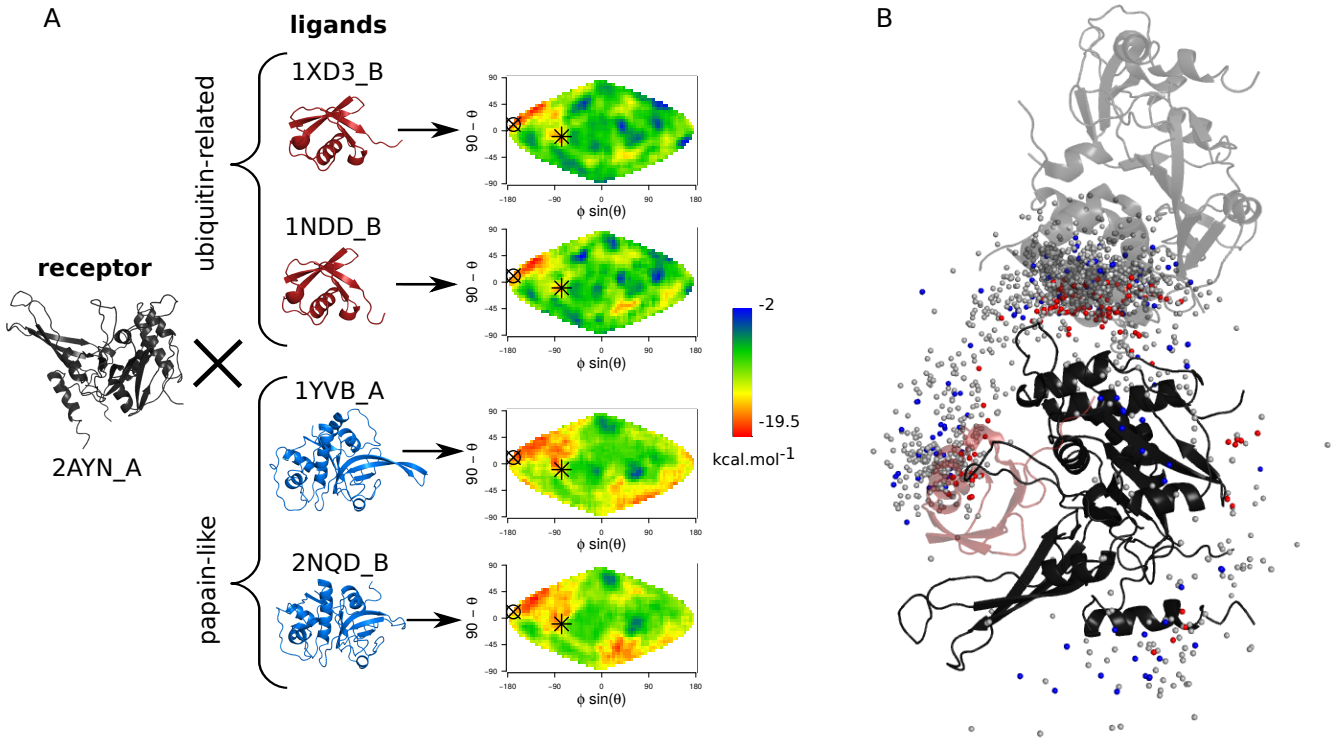
A
ligands
ubiquitin-related
1XD3_B
1NDD_B
receptor
2AYN_A
1YVB_A
2NQD_B
papain-like
90 − θ
ϕ sin(θ)
-2
-19.5
kcal.mol-1
B

Fig. 3.

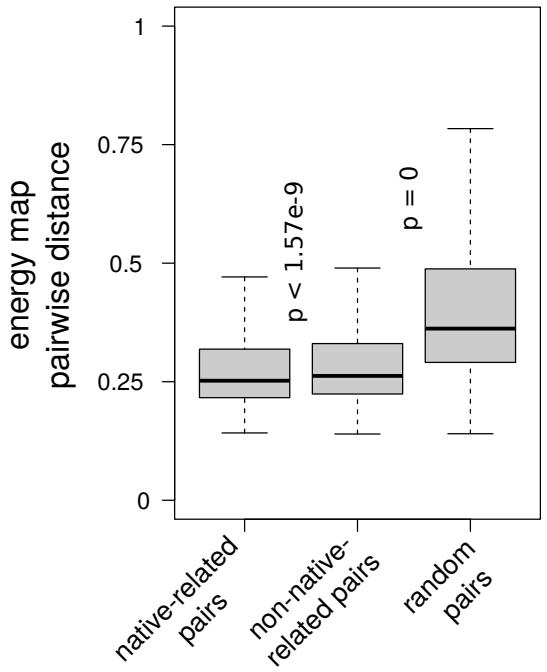
energy map
pairwise distance
1
0.75
0.5
0.25
0
p < 1.57e-9
p = 0
native-related pairs
non-native-related pairs
random pairs

Fig. 4.

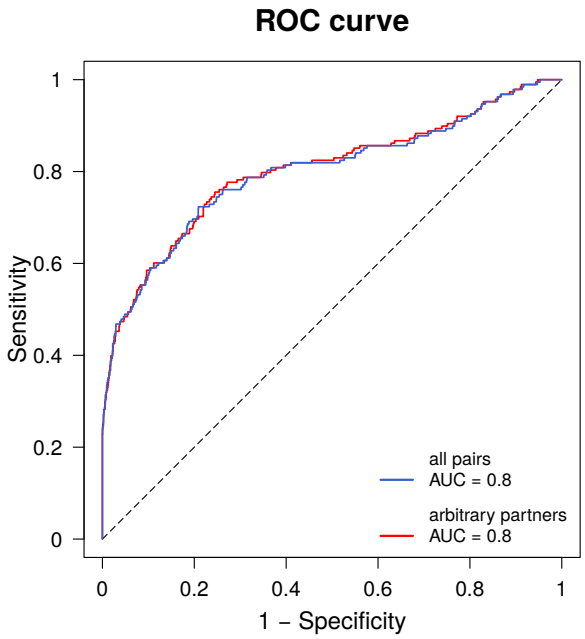
ROC curve
Sensitivity
1 − Specificity
0
0.2
0.4
0.6
0.8
1
all pairs
AUC = 0.8
arbitrary partners
AUC = 0.8

Fig. 5.

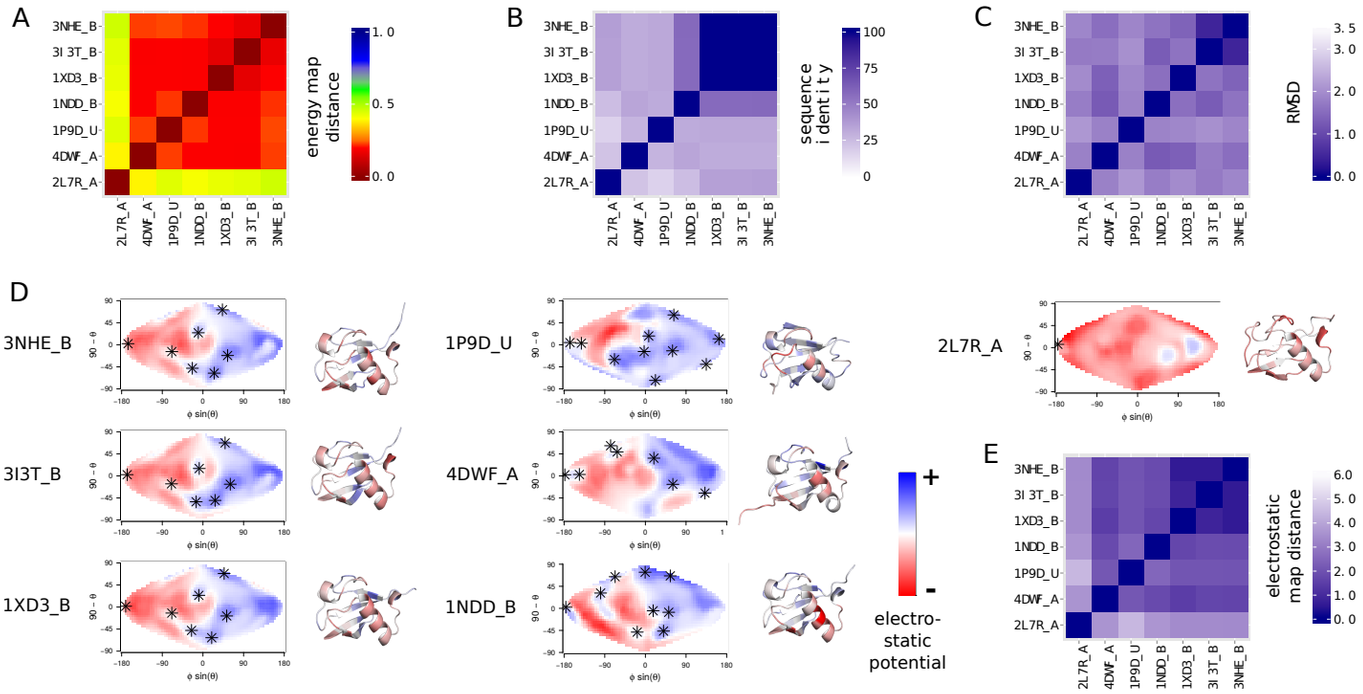
A
3NHE_B
3I 3T_B
1XD3_B
1NDD_B
1P9D_U
4DWF_A
2L7R_A
energy map distance
1. 0
0. 5
0. 0
B
sequence identity
100
75
50
25
0
C
RMSD
3. 5
3. 0
2. 0
1. 0
0. 0
D
3NHE_B
3I3T_B
1XD3_B
1P9D_U
4DWF_A
1NDD_B
2L7R_A
ϕ sin(θ)
90 − θ
+
-
electro-
static
potential
E
electrostatic map distance
6.0
5. 0
4. 0
3. 0
2. 0
1. 0
0. 0

Fig. 6.

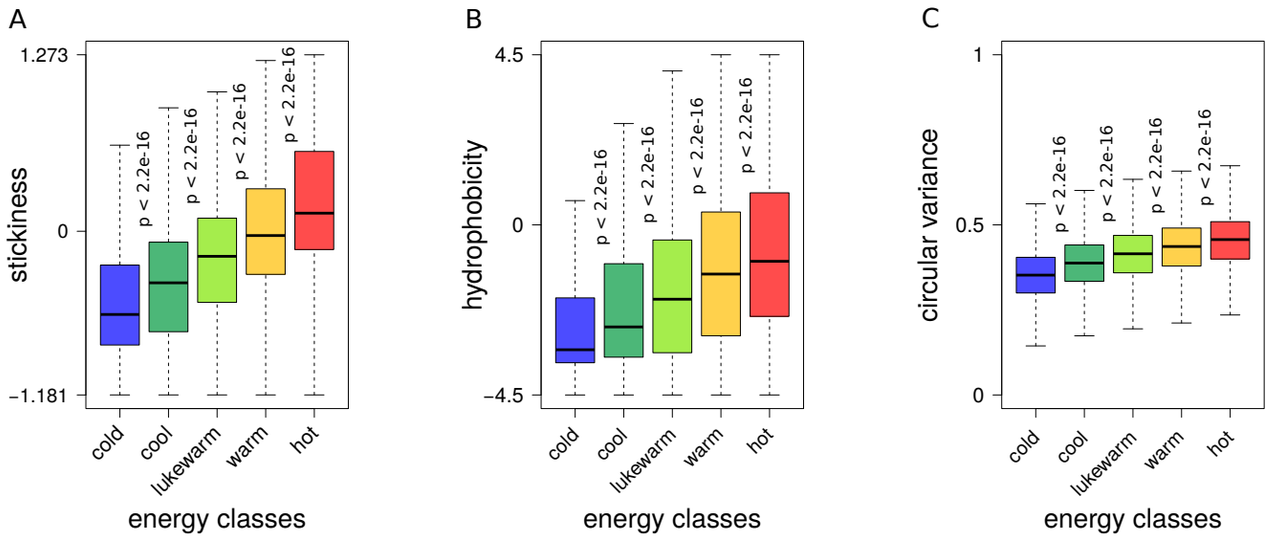
A
1.273
0
−1.181
stickiness
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16
cold
cool
lukewarm
warm
hot
energy classes
B
4.5
0
−4.5
hydrophobicity
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16
cold
cool
lukewarm
warm
hot
energy classes
C
1
0.5
0
circular variance
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16
cold
cool
lukewarm
warm
hot
energy classes

Fig. 7.

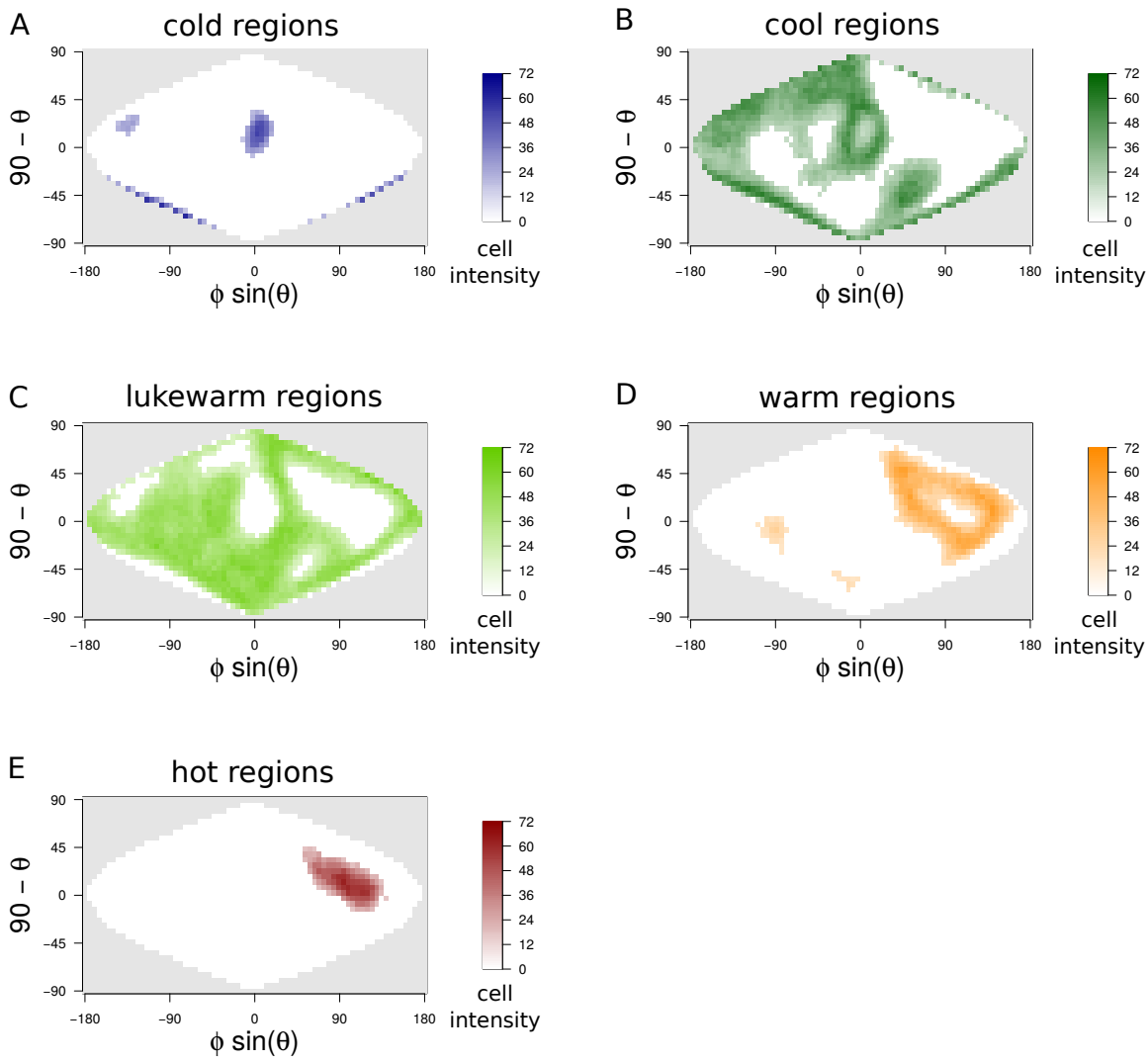
A
cold regions
B
cool regions
C
lukewarm regions
D
warm regions
E
hot regions
90 − θ
ϕ sin(θ)
cell intensity
90
45
0
−45
−90
−180
−90
0
90
180
72
60
48
36
24
12
0

Fig. 8.

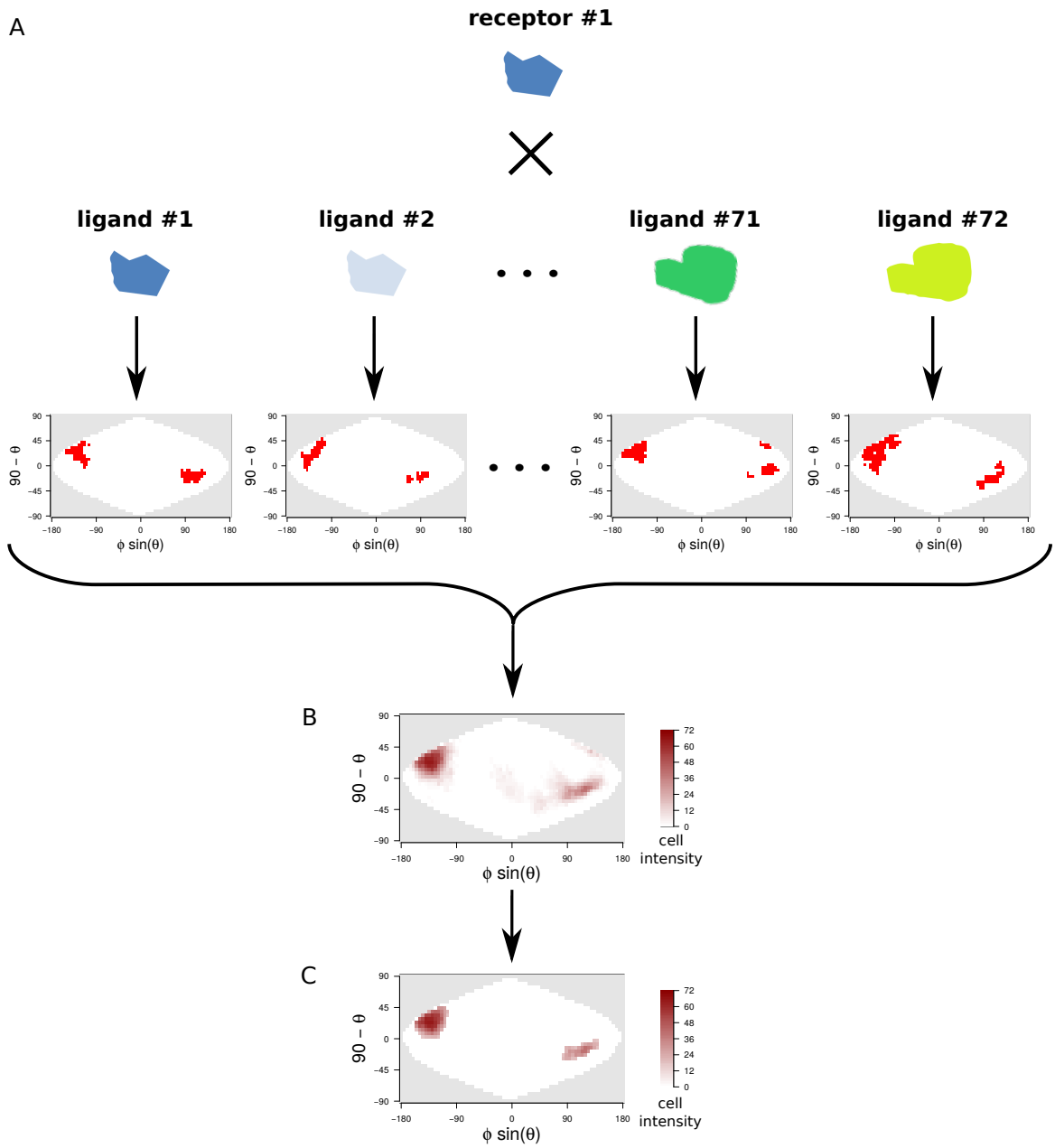

Fig. 8-supplement 1.

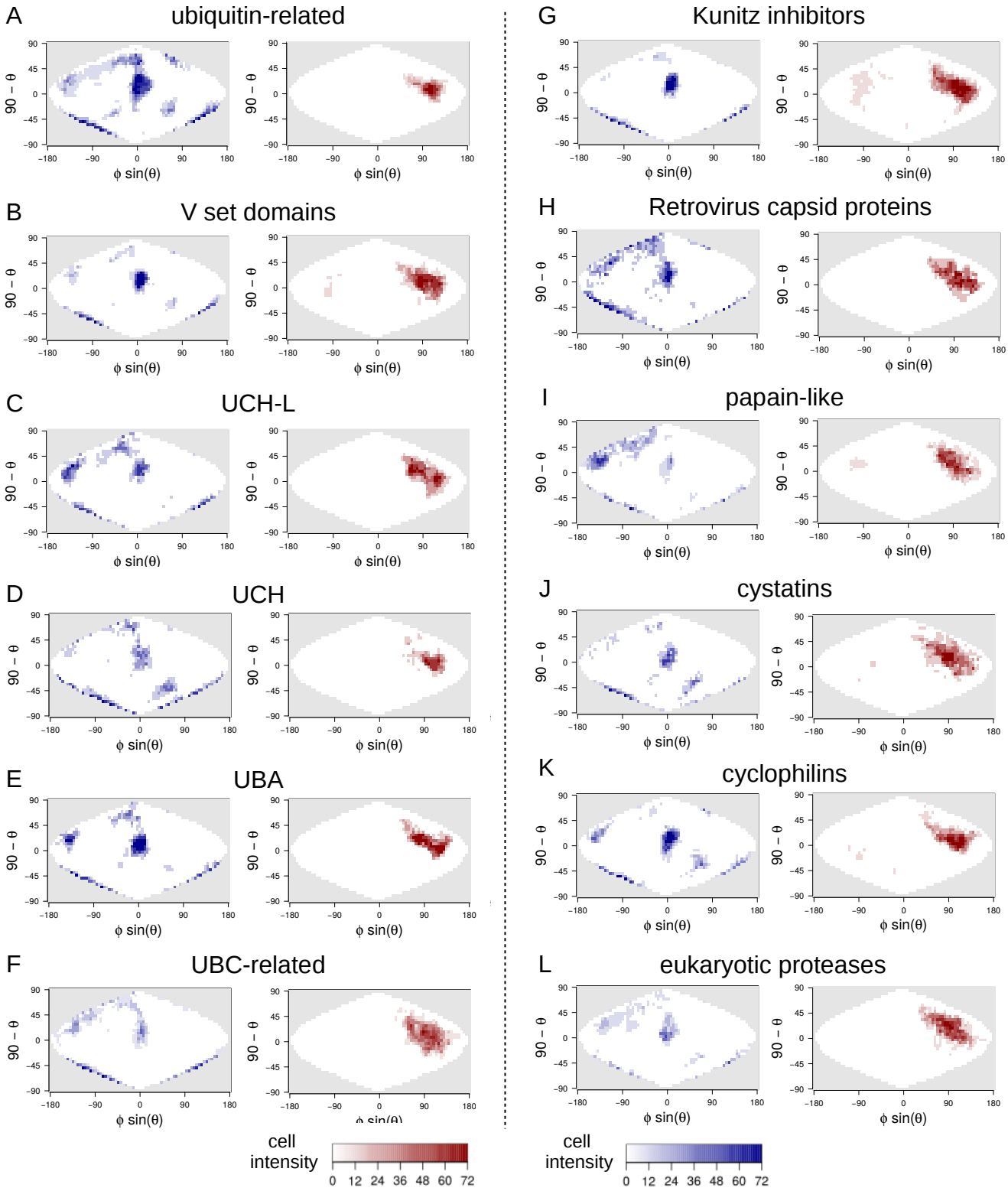

Fig. 8-supplement 2.