Title

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

Short Title

The interaction propensity of the whole surface of proteins is conserved during evolution

Authors

Hugo Schweke[a], Marie-Hélène Mucchielli[ab], Sophie Sacquin-Mora[c], Wanying Bei[a], Anne Lopes[a]

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

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. Understanding how this competition constrains the behavior of proteins with respect to their partners or random encounters is very difficult to address experimentally. Here, we developed an original theoretical framework in order to investigate the propensity of protein surfaces to interact with functional and arbitrary partners and ask whether their interaction propensity is conserved during evolution. Therefore, we performed 5476 cross-docking simulations to systematically characterize the energy landscapes of 74 proteins interacting with different sets of homologs, corresponding to their functional partner's family or arbitrary protein families. Our framework relies on an original representation of interaction energy landscapes with two-dimensional energy maps that reflect the propensity of a protein surface to interact. To address the evolution of interaction energy landscapes, we systematically compared the energy maps resulting from the docking of a protein with several homologous partners. 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 arbitrary partners. While most studies aiming at depicting protein-protein interactions focus on native binding sites of proteins, our analysis framework enables in an efficient and automated way, the physical characterization of not only known binding sites, but also of the rest of the protein surfaces, and provides a wealth of valuable information to understand mechanisms driving and regulating protein-protein interactions. It enables to address the energy behavior of a protein in interaction with hundreds of selected partners, providing a functional and systemic point of view of protein interactions, and opening the way for further developments to study the behavior of proteins in a specific environment.

Author Summary

In the crowded cell, the competition between functional and non-functional interactions is severe. Understanding how a protein binds the right piece in the right way in this complex jigsaw puzzle is crucial and very 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 cross-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 be they functional or not. We show that exploring non-functional interactions (i.e. non-functional assemblies and interactions with non-functional partners) is a viable route to investigate the mechanisms underlying protein-protein interactions. Precisely, our 2D energy maps based strategy enables it in an efficient and automated way. Moreover, our theoretical framework opens the way for the developments of a variety of applications covering functional characterization, binding site prediction, or characterization of protein behaviors in a specific environment.

77 ## Introduction

78 Biomolecular interactions are central for many physiological processes and are of utmost
79 importance for the functioning of the cell. Particularly protein-protein interactions have
80 attracted a wealth of studies these last decades [1–5]. The concentration of proteins in a cell
81 has been estimated to be approximately 2-4 million proteins per cubic micron [6]. In such a
82 highly crowded environment, proteins constantly encounter each other and numerous non-
83 specific interactions are likely to occur [7,8]. For example, in the cytosol of *S. cerevisiae* a
84 protein can encounter no less than 2000 different proteins [9]. In this complex jigsaw puzzle,
85 each protein has evolved to bind the right piece in the right way (positive design) and to
86 prevent misassembly and non-functional interactions (negative design) [10–14]).

87 Consequently, positive design constrains the physico-chemical properties and the evolution of
88 protein-protein interfaces. Indeed, a strong selection pressure operates on binding sites to
89 maintain the functional assembly. For example, homologs sharing at least 30% sequence
90 identity almost invariably interact in the same way [15]. Conversely, negative design prevents
91 proteins to be trapped in the numerous competing non-functional interactions inherent to the
92 crowded environment of the cell. Particularly, the misinteraction avoidance shapes the
93 evolution and physico-chemical properties of abundant proteins, resulting in a slower
94 evolution and less sticky surfaces than what is observed for less abundant ones [16–21]. The
95 whole surface of abundant proteins is thus constrained, preventing them to engage deleterious
96 non-specific interactions that could be of dramatic impact for the cell at high concentration
97 [20]. Recently, it has been shown in *E. coli* that the net charge as well as the charge
98 distribution on protein surfaces affect the diffusion coefficients of proteins in the cytoplasm
99 [22]. Positively charged proteins move up to 100 times more slowly as they get caught in non-

100 specific interactions with ribosomes which are negatively charged and therefore, shape the
101 composition of the cytoplasmic proteome [22].

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

115 In this work, we focus on protein surfaces as a proxy for functional and non-functional
116 protein-protein interactions. We investigate their interaction energy landscapes with native
117 and non-native partners and ask whether their interaction propensity is conserved during
118 evolution. With this aim in mind, we performed large-scale docking simulations to
119 characterize interactions involving either native and/or native-related (i.e. partners of their
120 homologs) partners or arbitrary partners. Docking simulations enable the characterization of
121 all possible interactions involving either functional or arbitrary partners, and thus to simulate
122 the interaction of arbitrary partners which is very difficult to address with experimental
123 approaches. Docking algorithms are now fast enough for large-scale applications and allow
124 for the characterization of interaction energy landscapes for thousand of protein couples.

125 Typically, a docking simulation takes from a few minutes to a couple of hours on modern
126 processors [23–25], opening the way for extensive cross-docking experiments [26–29].
127 Protein docking enables the exploration of the interaction propensity of the whole protein
128 surface by simulating alternative binding modes. Here, we performed a cross-docking
129 experiment involving 74 selected proteins docked with their native-related partners and their
130 corresponding homologs, as well as arbitrary partners and their corresponding homologs. We
131 represented the interaction energy landscape resulting from each docking calculation with a
132 two dimensional (2D) energy map in order to (i) characterize the propensity of all surface
133 regions of a protein to interact with a given partner (either native-related or not) and (ii) easily
134 compare the energy maps resulting from the docking of a same protein with different
135 homologous partners, thus addressing the evolution of the propensity of the whole protein
136 surface to interact with homology-related partners either native or arbitrary.

137 **Results**

138

139 **The interaction propensity of a protein to interact either with native-related or arbitrary**
140 **partners is conserved during evolution**

141 We ask whether the interaction propensity of a protein surface is conserved for homologous
142 native-related partners**,** and whether this remains true for homologous arbitrary partners. For a
143 protein A, we refer as native-related partners its native partner (when its three dimensional
144 (3D) structure is available) and native partners of proteins that are homologous to the protein
145 A. Arbitrary pairs refer to pairs of proteins for which no interaction between them or their
146 respective homologs has been experimentally characterized in the Protein Data Bank [30]. To
147 test the aforementioned hypothesis, we built a database comprising 74 protein structures
148 divided into 12 families of homologs (S1 Table and *Materials and Methods*). Each family
149 displays different degrees of structural variability and sequence divergence in order to see the
150 impact of these properties on the conservation of the interaction propensity inside a protein
151 family. Each family has at least a native-related partner family (S1 Fig). Docking calculations
152 were performed with the ATTRACT software [25]. ATTRACT enables a homogeneous and
153 exhaustive conformational sampling and is well suited to investigate the propensity of the
154 whole surface of a protein to interact with a given ligand. Our procedure is asymmetrical
155 since we aim at characterizing the interaction propensity of a protein (namely the receptor)
156 with a subset of proteins (namely the ligands). Therefore, a given receptor is docked with a
157 subset of ligands (here the 74 proteins of the dataset) (Fig 1A and *Materials and Methods*).
158 For each docking calculation, we produced a 2D energy map, which provides the distribution
159 of interaction energies of all docking solutions over the whole receptor surface (Fig 1B and
160 *Materials and Methods*, Fig 2A-C). The resulting energy map reflects the propensity of the
161 whole surface of the receptor to interact with the docked ligand. One should notice that

162 energy maps computed for two unrelated receptors are not comparable since their surfaces are
163 not comparable. Therefore, the procedure is ligand-centered and allows only the comparison
164 of energy maps produced by different ligands docked with the same receptor. The comparison
165 of two energy maps enables the evaluation of the similarity of the interaction propensity of the
166 receptor with the two corresponding ligands. In order to investigate the interaction propensity
167 of all proteins of the dataset, each protein plays alternately the role of receptor and ligand.
168 Consequently, the procedure presented in Fig 1 is repeated for the whole dataset where each
169 protein plays the role of the receptor and is docked with the 74 proteins that play the role of
170 ligands.

171

172 **Fig 1. Experimental Protocol.** (*A*) A receptor protein is docked with all proteins of the
173 dataset (namely the ligands) resulting in 74 docking calculations. (*B*) For each docking
174 calculation, an energy map is computed as well as its corresponding five-color and one-color
175 energy maps, with the procedure described in Fig 2 and *Materials and Methods*. (*C*) An
176 energy map distance (EMD) matrix is computed, representing the pairwise distances between
177 the 74 energy maps resulting from the docking of all ligands with this receptor. Each cell (i,j)
178 of the matrix represents the Manhattan distance between the two energy maps resulting from
179 the docking of ligands i and j with the receptor. A small distance indicates that the ligands i
180 and j produce similar energy maps when docked with this receptor. In other words, it reflects
181 that the interaction propensity of this receptor is similar for these two ligands. To prevent any
182 bias from the choice of the receptor, the whole procedure is repeated for each receptor of the
183 database, leading to 74 EMD matrices.

184

185 **Fig 2. 2D asymmetrical representation of docking energy landscapes and resulting**
186 **energy maps.** (*A*) Three-dimensional (3D) representation of the ligand docking poses around

187 the receptor. Each dot corresponds to the center of mass (CM) of a ligand docking pose. It is
188 colored according to its docking energy score. (*B*) Representation of the CM of the ligand
189 docking poses after an equal-area 2D sinusoidal projection. CMs are colored according to the
190 same scale as in A. (*C*) Continuous energy map (see *Materials and Methods* for more details).
191 (*D*) Five-color map. The energy map is discretized into five energy classes (*E*) One-color
192 maps. Top to bottom: red, orange, green, dark green and blue maps highlight respectively hot,
193 warm, lukewarm, cool and cold regions.

194

195 Fig 3A represents the energy maps computed for the receptor 2AYN_A, the human ubiquitin
196 carboxyl-terminal hydrolase 14 (family UCH) docked with (i) its native partner (1XD3_B,
197 ubiquitin-related family), a homolog of its partner (1NDD_B) and (ii) two arbitrary
198 homologous ligands (1YVB_A and 1NQD_B from the papain-like family). For all four
199 ligands, either native-related or arbitrary partners, docking calculations lead to an
200 accumulation of low-energy solutions (hot regions in red) around the two experimentally
201 known binding sites of the receptor. The first one corresponds to the interaction site with the
202 native partner, ubiquitin (pdb id 2ayo). The second one corresponds to its homodimerisation
203 site (pdb id 2ayn). This indicates that native-related but also arbitrary partners tend to bind
204 onto the native binding sites of native partners as observed in earlier studies [29,31]. The
205 same tendency is observed for all 74 ligands in the database (Fig 3B). Their 20 best docking
206 poses systematically tend to accumulate in the vicinity of the two native interaction sites.
207 Whereas the low-energy solutions for most ligands accumulate around the same interaction
208 sites (i.e. the native binding sites), we observe that, globally, 2-D energy maps (i) seem to be
209 more similar between ligands of a same family than between ligands belonging to different
210 families (Fig 3A). The two energy maps obtained with the ligands of the native-related

211 partners family both reveal two sharp hot regions around the native sites and a subset of well-
212 defined cold regions (i.e. blue regions corresponding to high energy solutions) placed in the
213 same area in the map's upper-right quadrant. In contrast, the energy maps obtained for the two
214 ligands of the papain-like family display a large hot region around the two native binding sites
215 of the receptor, extending to the upper-left and bottom-right regions of the map, suggesting a
216 large promiscuous binding region for these ligands.

217

218 **Fig 3. Subset of energy maps and of ligand docking poses for receptor 2AYN_A.** (*A*)
219 Examples of maps for the receptor 2AYN_A (ubiquitin carboxyl-terminal hydrolase (UCH)
220 family) docked with the ligands 1XD3_B (native partner), 1NDD_B (homolog of the native
221 partner), 1YVB_A and 2NQD_B (false partners). The star indicates the localization of the
222 experimentally determined interaction site of the ubiquitin, the circle-cross indicates the
223 homodimerization site of 2AYN_A. (*B*) Centers of mass (CM) of the 20 best docking poses
224 obtained for each of the 74 ligands of the database docked with the receptors 2AYN_A.
225 Receptor protein is represented in cartoon (black), its native ligand and its homodimere are
226 represented in cartoon with transparency (red and black respectively). CMs of the ligands
227 belonging to the ubiquitin-related family are colored in red, CMs of the proteins belonging to
228 the papain-like family are colored in blue.

229

230 We ask whether the observation made for the receptor 2AYN_A, that energy maps produced
231 with homologous ligands are more similar than those produced with unrelated ligands could
232 be generalized to all proteins of the dataset. Therefore, we systematically compared the
233 energy maps computed for a single receptor docked successively with the 74 ligands of the
234 dataset by calculating of the Manhattan distance between each pair of maps (Fig 1C and
235 *Materials and Methods*). The resulting distances are stored in an energy map distance (EMD)

236 matrix, where each entry (i,j) corresponds to the distance $d_{i,j}$ between the energy maps of
237 ligands i and j docked with the receptor of interest (Fig 1C and *Materials and Methods*).
238 Consequently, a small distance $d_{i,j}$ between ligands i and j docked with the receptor k, reflects
239 that their energy maps are similar. In other words, the interaction propensity of the surface of
240 the receptor k is similar for both ligands i and j. The procedure is repeated for each receptor of
241 the dataset resulting in 74 EMD matrices. In order to quantify the extent to which the
242 interaction propensity of the receptor is conserved for homologous ligands, we investigate
243 whether distances calculated between homologous ligand pairs (be they native-related to the
244 receptor or not) are smaller than distances calculated between random pairs. Fig 4 represents
245 the boxplots of energy map distances calculated between random ligand pairs or between
246 homologous ligand pairs docked with their native-related receptors or with the other receptors
247 of the dataset. Homologous ligands docked either with their native-related or arbitrary
248 receptors display significantly lower energy map distances than random ligand pairs
249 (Wilcoxon test $p = 0$). This indicates that energy maps produced by homologous ligands
250 docked with a given receptor are more similar than those produced with non-homologous
251 ligands. Interestingly, this observation holds whether the receptor-ligand pair is a native pair
252 or not. This suggests that the interaction propensity of a receptor is conserved for homologous
253 ligands be they native-related or not.

254

255 **Fig 4. Boxplots of energy map pairwise distances between homologous ligand pairs from**
256 **native-related partner families, homologous ligand pairs from arbitrary partner families**
257 **and random ligand pairs.** For each receptor, we computed (i) the average of energy map
258 distances of pair of homologous ligands belonging to its native-related partner family(ies), (ii)
259 the average of energy map distances of pair of homologous ligands belonging to its non-

260 native-related partner families, and (iii) the average of energy map distances of random pairs.
261 P-values are calculated with an unilateral Wilcoxon test.
262
263 **Energy maps are specific to protein families**
264
265 The results presented above prompt us to assess the extent to which the interaction propensity
266 of a receptor is specific to the ligand families. In other words, we quantify the extent to which
267 energy maps are specific to ligand families. If so, we should be able to retrieve ligand
268 homology relationships solely with the comparison of their corresponding 2D energy maps.
269 Therefore, we tested our ability to predict the homologs of a given ligand based only on the
270 comparison of its energy maps with those of the other ligands. In order to prevent any bias
271 from the choice of the receptor, the 74 EMD matrices are averaged in an averaged distances
272 matrix (ADM) (see *Materials and Methods*). Each entry *(i,j)* of the ADM corresponds to the
273 averaged distance between two sets of 74 energy maps produced by two ligands *i* and *j*. A
274 low distance indicates that the two ligands display similar energy maps whatever the receptor
275 is. We computed a receiver operating characteristic (ROC) curve from the ADM (see
276 *Materials and Methods*) which evaluates our capacity to discriminate the homologs of a given
277 ligand from non-homologous ligands by comparing their respective energy maps computed
278 with all 74 receptors of the dataset. The true positive set consists in the homologous protein
279 pairs while the true negative set consists in any homology-unrelated protein pair. The
280 resulting Area Under the Curve (AUC) is equal to 0.79 (Fig 5). We evaluated the robustness
281 of the ligand's homologs prediction depending on the size of the receptor subset with a
282 bootstrap procedure by randomly removing receptor subsets of different sizes (from 1 to 73
283 receptors). The resulting AUCs range from 0.769 to 0.79, and show that from a subset size of
284 five receptors, the resulting prediction accuracy no longer significantly varies (risk of wrongly

285 rejecting the equality of two variances (F-test) >5%), and is thus robust to the nature of the
286 receptor subset (S2 Fig). Finally, we evaluated the robustness of the predictions according to
287 the number of grid cells composing the energy maps. Therefore, we repeated the procedure
288 using energy maps with resolutions ranging from 144x72 to 48x24 cells. S2 Table presents
289 the AUCs calculated with different grid resolutions. The resulting AUCs range from 0.78 to
290 0.8 showing that the grid resolution has a weak influence on the map comparison. All
291 together, these results indicate that homology relationships between protein ligands can be
292 detected solely on the basis of the comparison of their energy maps. In other words, the
293 energy maps calculated for a given receptor docked with a set of ligands belonging to a same
294 family are specific to these families. Interestingly, this observation holds for families
295 displaying important sequence variations (S1 Table). For example, the AUC computed for the
296 UCH and ubiquitin-related families are 0.98 and 0.88 respectively despite the fact that the
297 average sequence identity of these families does not exceed 45% (S3 Fig and S1 Table). This
298 indicates that energy maps are similar even for homologous ligands displaying large sequence
299 variations.
300
301 **Fig 5. Receiver operating characteristic (ROC) curve and its Area Under the Curve**
302 **(AUC).** ROC are calculated on the averaged distance matrix (ADM) including either all pairs
303 (blue) or only arbitrary pairs (red) (see *Materials and Methods* for more details).
304
305 We then specifically investigate the similarity of the energy maps produced by ligands
306 belonging to a same family in order to see whether some ligands behave energetically
307 differently from their family members. On the 74 ligands, only five (2L7R_A, 4BNR_A,
308 1BZX_A, 1QA9_A, 1YAL_B) display energy maps that are significantly different from those
309 of their related homologs (Z-tests *p-values* for the comparison of the averaged distance of

310 each ligand with their homologs versus the averaged distance of all ligands with their
311 homologous ligands ≤ 5%). In order to identify the factors leading to differences between
312 energy maps involving homologous ligands, we computed the pairwise sequence identity and
313 the root mean square deviation (RMSD) between the members of each family. Interestingly,
314 none of these criteria can explain the energy map differences observed within families (Fisher
315 test p of the linear model estimated on all protein families >0.1) (see Fig 6B-C for the
316 ubiquitin-related family, S4-S14B-C Fig for the other families, and S3 Table for details). Fig
317 6A represents a subsection of the ADM for the ubiquitin-related family (i.e. the energy map
318 distances computed between all the members of the ubiquitin-related family and averaged
319 over the 74 receptors). Low distances reflect pairs of ligands with similar energy behaviors
320 (i.e. producing similar energy maps when interacting with a same receptor) while high
321 distances reveal pairs of ligands with distant energy behaviors. 2L7R_A distinguishes itself
322 from the rest of the family, displaying high-energy map distances with all of its homologs.
323 RMSD and sequence identity contribute modestly to the energy map distances observed in Fig
324 6A (Spearman correlation test p^{RMSD} = 0.01 and p^{seq} = 0.02 (S3 Table, Fig 6B-C)). Fig 6D
325 shows a projection of the contribution from the electrostatic term in the energy function of
326 ATTRACT on the surface of the seven ubiquitin-related family members (for more details,
327 see S15 Fig and *Materials and Methods*). Fig 6E represents the electrostatic maps distances
328 computed between all members of the family. 2L7R_A stands clearly out, displaying a
329 negative electrostatic potential over the whole surface while its homologs harbor a remarkable
330 fifty-fifty electrostatic distribution (Fig 6D). The negatively charged surface of 2L7R_A is
331 explained by the absence of the numerous lysines that are present in the others members of
332 the family (referred by black stars, Fig 6D). Lysines are known to be essential for ubiquitin
333 function by enabling the formation of polyubiquitin chains on target proteins. Among the
334 seven lysines of the ubiquitin, K63 polyubiquitin chains are known to act in non-proteolytic

335 events while K48, K11, and the four other lysines polyubiquitin chains are presumed to be
336 involved into addressing proteins to the proteasome [32]. 2L7R_A is a soluble UBL domain
337 resulting from the cleavage of the fusion protein FAU [33]. Its function is unrelated to
338 proteasomal degradation, which might explain the lack of lysines on its surface and the
339 differences observed in its energy maps. Interestingly, the differences observed for the energy
340 maps of 1YAL_B (Papain-like family) (S4 Fig) and 4BNR_A (eukaryotic proteases family)
341 (S5 Fig) regarding their related homologs can be explained by the fact that they both display a
342 highly charged surface. These two proteins are thermo-stable [34,35], which is not the case
343 for their related homologs, and probably explains the differences observed in their relative
344 energy maps. The V-set domain family is split into two major subgroups according to their
345 averaged energy map distances (S6A Fig). The first group corresponds to CD2 proteins
346 (1QA9_A and its unbound form 1HNF_A) and differs significantly from the second group (Z-
347 test $p = 0.03$ and $p = 0.05$ respectively). The second group corresponds to CD58 (1QA9_B
348 and its unbound form 1CCZ_A) and CD48 proteins (2PTT_A). Interestingly, CD2 is known
349 to interact with its homologs (namely CD58 and CD48) through an interface with a striking
350 electrostatic complementarity [36]. The two subgroups have thus evolved distinct and specific
351 binding sites to interact together. We can hypothesize that they have different interaction
352 propensities resulting in the differences observed between their corresponding energy maps.
353 These five cases illustrate the capacity of our theoretical framework to reveal functional or
354 biophysical specificities of homologous proteins that could not be revealed by classical
355 descriptors such as RMSD or sequence identity.
356
357 **Fig 6. Ubiquitin-related family.** (*A*) Energy map distances matrix. It corresponds to the
358 subsection of the ADM for the ubiquitin-related family (for the construction of the ADM, see
359 *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the

360 ligand pair (*i*,*j*) averaged over the 74 receptors of the dataset. (*B*) Pairwise sequence identity
361 matrix between all members of the family. (*C*) Pairwise root mean square deviation (RMSD)
362 matrix between all members of the family. (*D*) Electrostatic maps and cartoon representations
363 of the seven members of the family. An electrostatic map represents the distribution of the
364 electrostatic potential on the surface of a protein (for more details, see S15 Fig and *Materials*
365 *and Methods*). On the electrostatic maps, lysines positions are indicated by stars. Cartoon
366 structures are colored according to the distribution of their electrostatic potential. (*E*)
367 Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan
368 distance between the electrostatic maps of the proteins *(i,j)*.
369
370 The AUC of 0.79 calculated previously with energy maps produced by the docking of either
371 native-related or arbitrary pairs indicates that energy maps are specific to ligand families. To
372 see whether this observation is not mainly due to the native-related pairs, we repeated the
373 previous test while removing that time all energy maps computed with native-related pairs
374 and calculated the resulting ADM. We then measured our ability to retrieve the homologs of
375 each ligand by calculating the ROC curve as previously. The resulting AUC is still equal to
376 0.79, revealing that our ability to identify a ligand's homologs is independent from the fact
377 that the corresponding energy maps were computed with native-related or arbitrary pairs (Fig
378 5). This shows that the energy maps are specific to protein families whether the docked pairs
379 are native-related or not. Consequently, the propensity of the whole protein surface to interact
380 with a given ligand is conserved and specific to the ligand family whether the ligand is native-
381 related or not. This striking result may reflect both positive and negative design operating on
382 protein surfaces to maintain functional interactions and to limit random interactions that are
383 inherent to a crowded environment.
384

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

387 To see whether some regions contribute more to the specificity of the maps produced by
388 homologous ligands, we next dissected the effective contribution of the surface regions of the
389 receptor defined according to their docking energy value, in the identification of ligand's
390 homologs. We discretized the energy values of each energy map into five categories, leading
391 to a palette of five energy classes (or colors) (see Fig 2D and *Materials and Methods*). These
392 five-color maps highlight low-energy regions (i.e. hot regions in red), intermediate-energy
393 regions (i.e. warm, lukewarm and cool regions in orange, light-green and dark-green
394 respectively) and high-energy regions (i.e. cold regions in blue). We first checked that the
395 discretization of the energy maps does not affect our ability to identify the homologs of each
396 of the 74 ligands from the comparison of their five-colors maps. The resulting AUC is 0.77
397 (Table 1)**,** showing that the discretization step does not lead to an important loss of
398 information.

399

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

type of map	continuous energy maps	five-colors energy maps	red energy maps	orange energy maps	light green energy maps	dark green energy maps	blue energy maps
AUC	0.79	0.77	0.73	0.76	0.76	0.76	0.79

401 The AUC are calculated from the ADM with the continuous energy maps (Fig 2C), the five-
402 color energy maps (Fig 2D) and the one-color energy maps (Fig 2E) (see *Materials and*
403 *Methods* for more details).

404

405

406

407 Then, we evaluated the contribution of each of the five energy classes separately in the
408 ligand's homologs identification by testing our ability to retrieve the homologs of the 74
409 ligands from their one-color energy maps (either red, orange, yellow, green or blue) (see
410 *Materials and Methods*). Table 1 shows the resulting AUCs. Interestingly, the information
411 provided by each energy class taken separately is sufficient for discriminating the homologs
412 of a given ligand from the rest of the dataset (Table 1). The resulting AUCs range from 0.76
413 to 0.79 for the orange, light green, dark green, and blue classes and are comparable to those
414 obtained with all classes taken together (0.77). This shows that (i) warm, lukewarm, cool, and
415 cold regions alone are sufficient to retrieve homology relationships between ligands and (ii)
416 the localization on the receptor surface of a given energy class is specific to the ligand
417 families. Hot regions are less discriminative and lead to an AUC of 0.73. In order to see how
418 regions corresponding to a specific energy class are distributed over a receptor surface, we
419 summed its 74 corresponding one-color maps into a stacked map (S16 Fig – see *Materials*
420 *and Methods* for more details). For each color, the resulting stacked map reflects the tendency
421 of a map cell to belong to the corresponding energy class. Fig 7 shows an example of the five
422 stacked maps (i.e. for cold, cool, lukewarm, warm and hot regions) computed for the receptor
423 1P9D_U. Intermediates regions (i.e. warm, lukewarm and cool regions) are widespread on the
424 stacked map while cold and hot regions are localized on few small spots (three and one
425 respectively) no matter the nature of the ligand. S17 Fig shows for the receptor 1P9D_U the
426 12 blue and red stacked maps computed for each ligand family separately. We can see that
427 some cold spots are specific to ligand families and that their area distribution is specific to
428 families while all 12 ligand families display the same hot spot in the map's upper-right
429 quadrant. These observations can be generalized to each receptor. On average, intermediate
430 regions are widespread on the stacked maps and cover respectively 744, 1164 and 631 cells
431 for cool, lukewarm and warm regions, while cold and hot regions cover no more than

432 respectively 104 and 110 cells respectively (S18 Fig). Interestingly, hot regions are more
433 colocalized than cold ones and are restricted to 2 distinct spots on average per stacked map,
434 while cold regions are spread on 3.7 spots on average (t-Test p = 7.42e-13). These results
435 show that ligands belonging to different families tend to dock preferentially on the same
436 regions and thus lead to similar hot region distributions on the receptor surface. This
437 observation recalls those made by *Fernandez-Recio et al*. [31], who showed that docking
438 random proteins against a single receptor leads to an accumulation of low-energy solutions
439 around the native interaction site and who suggested that different ligands will bind
440 preferentially on the same localization.
441
442 **Fig 7. Stacked maps of 1P9D_U after the filtering of cells with too low intensity and**
443 **areas of too small size.** The protocol to generate stacked maps is presented in S16 Fig. (*A*)
444 Blue stacked map (i.e. stacked cold regions). (*B*) Dark green stacked map (i.e. stacked cool
445 regions). (*C*) Light green stacked map (i.e. stacked lukewarm regions). (*D*) Orange stacked
446 map (i.e. stacked warm regions). (*E*) Red stacked map (i.e. stacked hot regions). One should
447 notice that stacked maps of two different colors can overlap because a cell can be associated
448 to different energy classes depending on the docked ligands. S17 Fig presents blue and red
449 stacked maps of 1P9D_U computed for each ligand family.
450
451 We can hypothesize that hot regions present universal structural and biochemical features that
452 make them more prone to interact with other proteins. To test this hypothesis, we computed
453 for each protein of the dataset, the 2D projection of three protein surface descriptors (see
454 *Materials and Methods* and S15 Fig): the Kyte-Doolittle (KD) hydrophobicity [37], the
455 circular variance (CV) [38] and the stickiness [20]. The CV measures the density of protein
456 around an atom and is a useful descriptor to reflect the local geometry of a surface region. CV

457 values are comprised between 0 and 1. Low values reflect protruding residues and high values
458 indicate residues located in cavities. Stickiness reflects the propensity of amino acids to be
459 involved in protein-protein interfaces [20]. It has been calculated as the log ratio of the
460 residues frequencies on protein surfaces versus their frequencies in protein-protein interfaces.
461 For each receptor, we calculated the correlation between the docking energy and the
462 stickiness, hydrophobicity or CV over all cells of the corresponding 2D maps. We found a
463 significant anti-correlation between the docking energy and these three descriptors
464 (correlation test *p* between docking energies and respectively stickiness, hydrophobicity and
465 CV < 2.2e-16, see S4 Table)). Fig 8 represents the boxplots of the stickiness, hydrophobicity
466 and CV of each energy class (see S15 Fig and *Materials and Methods* section for more
467 details). We observe a clear effect of these factors on the docking energy: cold regions (i.e.
468 blue class) are the less sticky, the less hydrophobic and the most protruding while hot ones
469 (i.e. red class) are the most sticky, the most hydrophobic and the most planar (Tukey HSD test
470 [39], *p* of the differences observed between each energy classes < 2.2e-16). One should notice
471 that stickiness has been defined from a statistical analysis performed on experimentally
472 characterized protein interfaces and therefore between presumed native partners. The fact that
473 docking energies (physics-based) calculated either between native-related or arbitrary partners
474 is anti-correlated with stickiness (statistics-based) defined from native interfaces, strengthens
475 strongly the concept of stickiness as the propensity of interacting promiscuously and provides
476 physics-based pieces of evidence for sticky regions as a proxy for promiscuous interactions.
477 We show that not only the area distribution on a receptor surface of hot regions but also those
478 of intermediate and cold regions are similar for homologous ligands and are specific to ligand
479 families (AUC ranging from 0.73 to 0.79) whether the ligands are native-related or not. This
480 tendency is even stronger for intermediate and cold regions. Interestingly, the information

481 contained in the cold regions that cover on average no more than 5.0% of the energy maps is
482 sufficient to identify homology relationships between ligands.
483
484 **Fig 8. Boxplots of three descriptors of the protein surface.** (*A*) the stickiness values, (*B*) the
485 Kyte-Doolittle hydrophobicity and (*C*) the CV values, depending on the energy class. The
486 stickiness, hydrophobicity and CV values are calculated for each protein following the
487 protocol described in *Materials and Methods*. For each of these criteria, *p-values* between the
488 median values of two "successive" energy classes were computed using the Tukey HSD
489 statistical test [39].

490 **Discussion**

491 In this study, we address the impact of both positive and negative design on thousands of
492 interaction energy landscapes by the mean of a synthetic and efficient representation of the
493 docking energy landscapes: two-dimensional energy maps that reflect the interaction
494 propensity of the whole surface of a protein (namely the receptor) with a given partner
495 (namely the ligand). We show that all regions of the energy maps, including cold,
496 intermediate and hot regions are similar for homologous ligands and are specific to ligand
497 families whether the ligands are native-related or arbitrary. This reveals that the interaction
498 propensity of the whole surface of proteins is constrained by functional and non-functional
499 interactions, reflecting both positive and negative design operating on the whole surface of
500 proteins, thus shaping the interaction energy landscapes of functional partners and random
501 encounters. These observations were made on a dataset of 74 protein structures belonging to
502 12 families of structural homologs. 54 out of the 74 proteins of the dataset have at least one
503 known partner in the dataset. For the 20 remaining proteins, we were not able to find
504 evidences that they indeed interact with a protein of the dataset. However, we showed that the
505 interaction propensity of a receptor is conserved for homologous ligands independently from
506 the fact that these ligands correspond to native partners or not. Indeed, we showed that ligand
507 homology relationships could be retrieved from their energy maps whether the maps were
508 computed with native-related pairs or not (the corresponding AUCs calculated with and
509 without native pairs both equal to 0.79).

510 While most studies that aim at depicting protein-protein interactions focus on native binding
511 sites of proteins [12,40–44], we bring a new perspective on protein-protein interactions by
512 providing a systematic and physical characterization of all regions of the surface of a protein
513 in interaction with a given ligand (i.e. cold, intermediate and hot regions). Here, we address
514 the energy behavior of not only known binding sites, but also of the rest of the protein surface,

515 which plays an important role in protein interactions by constantly competing with the native
516 binding site. We show that the interaction propensity of the rest of the surface is not
517 homogeneous and displays regions with different binding energies that are specific to ligand
518 families. This may reflect the negative design operating on these regions to limit non-
519 functional interactions [12,14,45]. We can hypothesize that non-interacting regions participate
520 to favor functional assemblies (i.e. functional assembly modes with functional partners) over
521 non-functional ones and are thus evolutionary constrained by non-functional assemblies. The
522 fact that cold regions seem to be more specific to ligand families than hot ones may be
523 explained by the fact that they are on average more protuberant and more charged. They thus
524 display more variability than hot ones. Indeed, there is more variability in being positively or
525 negatively charged and protuberant (with an important range of protuberant shapes) than in
526 being neutral and flat. S19 Fig presents the electrostatic potential distribution of all energy
527 classes. Cold regions display a larger variability of electrostatic potential (F-test, $p < 2.2e-16$)
528 than hot regions that are mainly hydrophobic thus displaying neutral charge distributions in
529 average. Consequently, a same hot region may be attractive for a large set of ligands while a
530 cold region may be unfavorable to specific set of ligands, depending on their charges, shapes
531 and other biophysical properties.
532 On the other hand, we show that hot regions are very localized (4.9% of the cells of an energy
533 map) and tend to be similar no matter the ligand. Similarly to protein interfaces that have been
534 extensively characterized in previous studies [2,40-43], hot regions are likely to display
535 universal properties of binding, i.e. they are more hydrophobic and more planar, and thus
536 more "sticky" than the other regions. They may provide a non-specific binding patch that is
537 suitable for many ligands. However, we can hypothesize that native partners have evolved to
538 optimize their interfaces (positive design) so that native interactions prevail over non-native
539 competing ones. Indeed, we have previously shown that the docking of native partners lead to

540 more favorable binding energies than the docking of non-native partners when the ligand is
541 constrained to dock around the receptor's native binding site [28,46]. All these results suggest
542 a new physical model of protein surfaces where protein surface regions, in the crowded
543 cellular environment, serve as a proxy for regulating the competition between functional and
544 non-functional interactions. In this model, intermediate and cold regions play an important
545 role by preventing non-functional assemblies and by guiding the interaction process towards
546 functional ones and hot regions may select the functional assembly among the competing ones
547 through optimized interfaces with the native partner.
548
549 In this work, we used and extended the application of the 2D energy map representation
550 developed in [31] to develop an original theoretical framework that enables the efficient,
551 automated and integrative analysis of different protein surface features. 2D maps provide the
552 area distribution of a given feature on the whole protein surface and their discretization
553 enables the study of a given surface property (e.g. protuberance, planarity, stickiness,
554 positively charged regions, or cold and hot regions for example). They are easy to manipulate
555 and their straightforward comparison enables (i) the study of relationships between different
556 surface properties through the comparison of their area distributions on a protein surface and
557 (ii) the highlight of the evolutionary constraints exerted on a given feature by comparing its
558 area distribution on the surfaces of homologous proteins. Particularly, this enables the
559 identification and characterization of hot regions on a protein surface which can be either
560 specific or conserved for all ligands and opens up new possibilities for the development of
561 novel methods for protein binding sites prediction and their classification as functional or
562 promiscuous in the continuity of previous developments based on arbitrary docking
563 [28,29,31,46].
564

565 Our framework provides a proxy for further protein functional characterization as shown with
566 the five proteins discussed in the *Results* section *Energy maps are specific to protein families*.
567 The comparison of their respective energy maps enables us to reveal biophysical and
568 functional properties that could not be revealed with classical monomeric descriptors such as
569 RMSD or sequence identity. Indeed, our framework can reflect the energy behavior of a
570 protein interacting with a subset of selected partners either functional or arbitrary, thus
571 revealing functional and systemic properties of proteins. This work goes beyond the classical
572 use of binary docking to provide a systemic point of view of protein interactions**,** for example
573 by exploring the propensity of a protein to interact with hundreds of selected ligands, and thus
574 addressing the behavior of a protein in a specific cellular environment. Particularly, exploring
575 the dark interactome (i.e. non-functional assemblies and interactions with non-functional
576 partners) can provide a wealth of valuable information to understand mechanisms driving and
577 regulating protein-protein interactions. Precisely, our 2D energy maps based strategy enables
578 its exploration in an efficient and automated way.

579 ## Materials and Methods

580

581 **Protein dataset**

582 The dataset comprises 74 protein structures divided into 12 families of structural homologs
583 (see S1 Table for a detailed list of each family). Each family is related to at least one other
584 family (its native-related partners family) through a pair of interacting proteins for which the
585 3D structure of the complex is characterized experimentally (except the V set domain family:
586 the two native partners are homologous and belong to the same family) (S1 Fig). Each family
587 is composed of a monomer selected from the protein-protein docking benchmark 5.0 [47] in
588 its bound and unbound forms, which is called the master protein. Each master protein has a
589 native partner (for which the 3D structure of the corresponding complex has been
590 characterized experimentally) in the database, which is the master protein for another family,
591 except the V set domain family, which is a self-interacting family. When available, we
592 completed families with interologs (i.e. pairs of proteins which have interacting homologs in
593 an other organism) selected in the INTEREVOL database [48] according to the following
594 criteria: (i) experimental structure resolution better than 3.25 Å, (ii) minimum alignment
595 coverage of 75% with the rest of the family members and (iii) minimum sequence identity of
596 30% with at least one member of the family. Since we were limited by the number of
597 available interologs, we completed families with unbound monomers homologous to the
598 master following the same criteria and by searching for their partners in the following protein-
599 protein interactions databases [49–54]. We consider that all members of a family correspond
600 to native-related partners of all members of their native-related partner family. To address the
601 impact of conformational changes of a protein on its interaction energy maps, we added
602 different NMR conformers. We show that energy maps involving pairs of conformers are
603 significantly more similar than those obtained for other pairs of homologous ligands

604 (unilateral Wilcoxon test, $p < 2.2e-16$) showing that the conformational changes in a protein
605 (lower than 3Å) have a low impact on the resulting energy maps (S20 Fig).
606
607 **Docking experiment and construction of energy maps**

608 A complete cross-docking experiment was realized with the ATTRACT software [25] on the
609 74 proteins of the dataset, leading to 5476 (74 x 74) docking calculations (Fig 1A).
610 ATTRACT uses a coarse-grain reduced protein representation and a simplified energy
611 function comprising a pseudo Lennard-Jones term and an electrostatic term. The calculations
612 took approximately 20000 hours on a 2.7GHz processor. Prior to docking calculations, all
613 PDB structures were prepared with the DOCKPREP software [55].

614 During a docking calculation, the ligand L_i explores exhaustively the surface of the receptor
615 R_k (whose position is fixed during the procedure), sampling and scoring thousands of different
616 ligand docking poses (between 10000 and 50000 depending on the sizes of the proteins) (Fig
617 2A). For each protein couple R_k-L_i, a 2D energy map is computed which shows the
618 distribution of the energies of all docking solutions over the receptor surface. To compute
619 these maps, for all docking poses, the spherical coordinates (ϕ, θ) (with respect to the
620 receptor center of mass (CM)) of the ligand CM are represented onto a 2D map in an equal-
621 area 2D sinusoidal projection (Fig 2B) (see [31] for more details). Each couple of coordinates
622 (ϕ, θ) is associated with the energy of the corresponding docking conformation (Fig 2B). A
623 continuous energy map is then derived from the discrete one, where the map is divided into a
624 grid of 36 x 72 cells. Each cell represents the same surface and, depending on the size of the
625 receptor, can span from 2.5 $Å^2$ to 13$Å^2$. For each cell, all solutions with an energy score below
626 2.7 kcal/mol^{-1} from the lowest solution of the cell are retained, according to the conformations
627 filtering protocol implemented in [28]. The average of the retained energy scores is then

assigned to the cell. If there is no docking solution in a cell, a score of 0 is assigned to it. Finally, the energies of the cells are smoothed, by averaging the energy values of each cell and of the eight surrounding neighbors (Fig 2C).

For each map, the energy values are discretized into five energy classes of same range leading to a discrete five-colors energy map (Fig 2D). The range is calculated for each energy map and spans from the minimum to the maximum scores of the map cells. The range of the energy classes of the map R_k-L_i is equal to (maxE – minE)/5, where maxE and minE correspond to the maximal and minimal energy values in the R_k-L_i map. Each five-colors energy map is then split into five one-color maps, each one representing an energy class of the map (Fig 2E). The continuous, five-colors and one-color energy maps are calculated for the 5476 energy maps.

Comparison of energy maps and identification of ligand's homologs

Since, we cannot compare energy maps computed for two unrelated receptors, the procedure is ligand-centered and only compares energy maps produced with different ligands docked with the same receptor. The referential (i.e. the receptor) is thus the same (in other words all grid cells are comparable) for all the energy maps that are compared. For each receptor R_k, we computed a 74x74 energy map distance (EMD) matrix where each entry (i,j) corresponds to the pairwise distance between the energy maps R_k-L_i and R_k-L_j resulting from the docking of the ligands L_i and L_j on the receptor R_k (Fig 1). The pairwise distance $d_{Man}(R_k\text{-}L_i, R_k\text{-}L_j)$ between the energy maps is calculated with a Manhattan distance according to equation (1)

$$d_{Man}\left(R_k L_i, R_k L_j\right) = \sum_{n=1}^{36} \sum_{m=1}^{72} |a_{nm} - b_{nm}| \qquad (1)$$

652 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
653 and R_k-L_j respectively. Low distances reflect pairs of ligands that induce similar energy maps
654 when they are docked on the same receptor. The procedure presented in Fig 1 is repeated for
655 each receptor of the database resulting in 74 EMD matrices. The 74 EMD matrices are
656 averaged into an averaged distances matrix (ADM). Each entry (*i*,*j*) of the ADM reflects the
657 similarity of the R_k-L_i and R_k-L_j energy maps averaged over all the receptors R_k in the dataset.
658 In order to estimate the extent to which family members display similar energy maps when
659 they are docked with the same receptor, we tested our ability to correctly identify the
660 homologs of the 74 ligands from the only comparison of its energy maps with those of the
661 other ligands. Because, energy maps are receptor-centered, we cannot compare the energy
662 maps computed for two unrelated receptors. The procedure consists in the comparison of
663 energy maps produced with different ligands docked with a same receptor. Two ligands (*i*,*j*)
664 are predicted as homologs according to their corresponding distance (*i*,*j*) in the ADM. Values
665 close to zero should reflect homologous ligand pairs, while values close to one should reflect
666 unrelated ligand pairs. A Receiver Operating Characteristic (ROC) curve and its Area Under
667 the Curve (AUC) are computed from the ADM. True positives (TP) are all the homologous
668 ligand pairs and predicted as such, true negatives (TN) are all the unrelated ligand pairs and
669 predicted as such. False positives (FP) are unrelated ligand pairs but incorrectly predicted as
670 homologous pairs. False negatives (FN) are homologous ligand pairs but incorrectly predicted
671 as unrelated pairs. ROC curves and AUC values were calculated with the R package pROC
672 [56]. The ligand's homologs identification was also realized using the five-color energy maps
673 or the one-color energy maps taken separately. The five energy class regions display very
674 different sizes, with median ranging from 63 and 66 cells for the blue and red regions to 633
675 cells for the yellow one. To prevent any bias due to the size of the different classes, we

676 normalized the Manhattan distance by the size of the regions compared in the map. The rest of
677 the procedure is the same than those used for continuous energy maps (Fig 1).

678 To visualize the area distribution of the regions of a given energy class for all ligands on the
679 receptor surface, the 74 corresponding one-color maps are summed into a stacked map where
680 each cell's intensity varies from 0 to 74 (S16 Fig). To remove background-image from these
681 maps, i.e. cells with low intensity (intensity < 17) and the areas of small size (< 4 cells), we
682 used a Dirichlet process mixture model simulation for image segmentation (R package
683 *dpmixsim*) [57].

684

685 **2D projection of monomeric descriptors of protein surfaces**

686 We computed KD hydrophobicity [37], stickiness [20], CV [38] maps of each protein of the
687 dataset, in order to compare their topology with the energy maps. Prior to all, proteins
688 belonging to the same families were structurally aligned with TM-align [58] in order to place
689 them in the same reference frame, making their maps comparable. Particles were generated
690 around the protein surface with a slightly modified Shrake-Rupley algorithm [59]. The density
691 of spheres is fixed at 1Å^2, representing several thousands particles per protein. Each particle is
692 located at 5Å from the surface of the protein. The CV, stickiness and KD hydrophobicity
693 values of the closest atom of the protein are attributed to each particle. We also generated
694 electrostatic maps reflecting the distribution of the contribution of the coulombic term as
695 encoded in the ATTRACT force field on a protein surface. The procedure is slightly different:
696 each particle *i* has a +1 positive charge, and receives the coulombic value Q_i (see equation
697 (2)).

698 $$Q_i = \sum_{j=1}^{n} q_i q_j / \mathcal{E} r_{ij} \quad (2)$$

with n the number of pseudo-atom in the protein, q_i the charge of the particle, q_j the charge of the pseudo-atom j, r_{ij} the distance between the particle i and the pseudo-atom j, and ε a distant-dependent dielectric constant ($\varepsilon = 15r_{ij}$). CV was calculated following the protocol described in [38] on the all-atom structures. Stickiness, electrostatics and hydrophobicity were calculated on ATTRACT coarse-grain models. Pseudo-atom charges are defined according to the ATTRACT force field [25]. 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 2 and *Materials and Methods* section *Docking experiment and construction of energy maps*. The map is then smoothed following the protocol in Fig 2.

Acknowledgments

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

References

[1].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.

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

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

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

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

[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.

[7] 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.

[8] 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.

[9] 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.

[10] 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.

[11] 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.

[12] 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.

[13] 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.

[14] 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.

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

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

[17] 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.

[18] Zhang J, Maslov S, Shakhnovich EI. Constraints imposed by non-functional protein–protein interactions on gene expression and proteome size. Molecular Systems Biology 2008;4:210.

[19] 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.

[20] 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.

[21] 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.

[22] 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.

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

[24] 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.

[25] de Vries S, Zacharias M. Flexible docking and refinement with a coarse-grained protein model using ATTRACT. Proteins 2013;81:2167–74. doi:10.1002/prot.24400.

[26] 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.

[27] 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.

[28] 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.

[29] 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.

[30] 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.

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

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

[38] 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.

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

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

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

[42] 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.

[43] 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.

[44] 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.

[45] Kastritis PL, Rodrigues JPGLM, Folkers GE, Boelens R, Bonvin AMJJ. Proteins feel more than they see: fine-tuning of binding affinity by properties of the non-interacting surface. J Mol Biol 2014;426:2632–52. doi:10.1016/j.jmb.2014.04.017.

[46] 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.

[47] 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.

[48] 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.

[49] Salwinski L, Miller CS, Smith AJ, Pettit FK, Bowie JU, Eisenberg D. The Database of Interacting Proteins: 2004 update. Nucleic Acids Res 2004;32:D449–51. doi:10.1093/nar/gkh086.

[50] Alfarano C, Andrade CE, Anthony K, Bahroos N, Bajec M, Bantoft K, et al. The Biomolecular Interaction Network Database and related tools 2005 update. Nucleic Acids Res 2005;33:D418–24. doi:10.1093/nar/gki051.

[51] Güldener U, Münsterkötter M, Oesterheld M, Pagel P, Ruepp A, Mewes H-W, et al. MPact: the MIPS protein interaction resource on yeast. Nucleic Acids Res 2006;34:D436–41. doi:10.1093/nar/gkj003.

[52] Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. BioGRID: a general repository for interaction datasets. Nucleic Acids Res 2006;34:D535–9. doi:10.1093/nar/gkj109.

[53] Kerrien S, Alam-Faruque Y, Aranda B, Bancarz I, Bridge A, Derow C, et al. IntAct—open source resource for molecular interaction data. Nucleic Acids Res 2007;35:D561–5. doi:10.1093/nar/gkl958.

[54] Alonso-López D, Gutiérrez MA, Lopes KP, Prieto C, Santamaría R, De Las Rivas J. APID interactomes: providing proteome-based interactomes with controlled quality for multiple species and derived networks. Nucleic Acids Res 2016;44:W529–35. doi:10.1093/nar/gkw363.

[55] 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.

[56] 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.

[57] 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.

[58] 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.

[59] 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.

[60] 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.

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

712 ## Supporting information

713

714 **S1 Table. List of proteins of the dataset and their structural families.** Proteins are referred
715 by their PDB identifiers, followed by their chain identifier. The NMR conformers are referred
716 with their conformation identifier. The conformational state of the structures are indicated in
717 brackets ((b) for bound conformation, (u) for unbound conformation). Structural families are
718 named according to the SCOPe database [60] at the family level. Averaged sequence identity
719 and RMSD are given for each family.

720

721 **S2 Table. AUC according to the grid resolution used for the energy maps.** A linear model
722 was constructed from the dataset constituted of all the intra-family ligand pairs (202 protein
723 pairs). This model allows the estimation of the linear correlation between the three descriptors
724 and the pairwise ADM distance. The model takes into account the individual contribution of
725 each descriptor as well as their crossed contributions with each other. The p-value of each
726 individual contribution calculated over the 202 pairs is estimated with a Fisher test and are
727 given in the table line "all proteins". We then individually looked each family to see whether
728 the contribution of the descriptors is dependent from the family. Inside each family, the
729 number of protein pairs is too small to estimate a linear model. Consequently, we used a
730 Spearman correlation coefficient test to estimate the p-value of each contribution.

731

732 **S3 Table. Estimation of the effective contribution of sequence identity, RMSD and**
733 **electrostatic distance in the pairwise ADM distances for each ligand pair belonging to a**
734 **same family.** The correlation is computed between each cell of the 74 energy maps of each of
735 the 74 receptors and the corresponding cell in receptor's maps of stickiness, hydrophobicity
736 and CV.

S4 Table. Correlation between energy scores and stickiness, hydrophobicity and circular variance (CV). The grid resolution corresponds to the number of cells composing the energy maps. The AUC is calculated following the same protocol used in the main text (see *Materials and Methods*)

S1 Fig. Interactions between structural families of the dataset. Interactions are symbolized by links between families. An interaction is established between two families when, there is at least one PDB reporting a structure of complex involving members of the two families [30]. Consequently, all members of a family do not necessarily have its native partner in its native-related partner family. The V set domains family is a special case of self-interacting family, where members form dimers of structural homologs.

S2 Fig. AUC values calculated on random subsets of receptor of different sizes. The AUC is computed following the protocol described in Fig. 1 with random subsets composed from 1 to 73 receptors. Receptors of each subset are randomly chosen among the 74 receptors of the dataset. For each subset size, the procedure is repeated 100 times. Red vertical lines indicate the standard deviation of the AUC for each subset size. Above a subset size of five receptors, the AUC does not significantly fluctuate (risk of wrongly rejecting the equality of two variances (F-test) >5% [61]).

S3 Fig. Receiver operating characteristic (ROC) curve and Area Under this Curve (AUC) calculated for each family.

S4 Fig. Papain-like family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the papain-like 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 (for more details, see *Materials and Methods*). (*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 (see Fig. S15 and *Materials and Methods*). 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)*.

S5 Fig. Eukaryotic-proteases family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the Eukaryotic proteases 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 (for more details, see *Materials and Methods*). (*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 Fig. S15 and *Materials and Methods*). 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)*.

S6 Fig. V set domains family. (*A*) Energy map distances matrix. It corresponds to the subsection of the ADM for the V set domain family (for the construction of the ADM, see

789 *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the
790 ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials*
791 *and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*)
792 Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*)
793 Electrostatic maps and cartoon representations of the six members of the family. An
794 electrostatic map represents the distribution of the electrostatic potential on the surface of a
795 protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are
796 colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map
797 distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the
798 electrostatic maps of the proteins *(i,j)*.
799
800 **S7 Fig. UCH-L family.** (*A*) Energy map distances matrix. It corresponds to the subsection of
801 the ADM for the UCH-L family (for the construction of the ADM, see *Materials and*
802 *Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)*
803 averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*).
804 (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root
805 mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic
806 maps and cartoon representations of the seven members of the family. An electrostatic map
807 represents the distribution of the electrostatic potential on the surface of a protein (for more
808 details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to
809 the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each
810 entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of
811 the proteins *(i,j)*.
812
813 **S8 Fig. UCH family.** (*A*) Energy map distances matrix. It corresponds to the subsection of the
814 ADM for the UCH family (for the construction of the ADM, see *Materials and Methods*).

815 Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)* averaged
816 over the 74 receptors of the dataset (for more details, see *Materials and Methods*). (*B*)
817 Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root mean
818 square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic maps
819 and cartoon representations of the seven members of the family. An electrostatic map
820 represents the distribution of the electrostatic potential on the surface of a protein (for more
821 details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to
822 the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each
823 entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of
824 the proteins *(i,j)*.
825
826 **S9 Fig. Ubiquitin activating enzymes family.** (*A*) Energy map distances matrix. It
827 corresponds to the subsection of the ADM for the Ubiquitin activating enzymes family (for
828 the construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the
829 pairwise energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the
830 dataset (for more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix
831 between all members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix
832 between all members of the family. (*D*) Electrostatic maps and cartoon representations of the
833 seven members of the family. An electrostatic map represents the distribution of the
834 electrostatic potential on the surface of a protein (for more details, see Fig. S15 and *Materials*
835 *and Methods*). Cartoon structures are colored according to the distribution of their
836 electrostatic potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix
837 represents the Manhattan distance between the electrostatic maps of the proteins *(i,j)*.
838
839 **S10 Fig. UBC-related family.** (*A*) Energy map distances matrix. It corresponds to the
840 subsection of the ADM for the UBC-related family (for the construction of the ADM, see

841 *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the
842 ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials*
843 *and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*)
844 Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*)
845 Electrostatic maps and cartoon representations of the seven members of the family. An
846 electrostatic map represents the distribution of the electrostatic potential on the surface of a
847 protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are
848 colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map
849 distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the
850 electrostatic maps of the proteins *(i,j)*.
851
852 **S11 Fig. Kunitz (STI) inhibitors family.** (*A*) Energy map distances matrix. It corresponds to
853 the subsection of the ADM for the Kunitz (STI) inhibitors family (for the construction of the
854 ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map
855 distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details,
856 see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all members of
857 the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all members of
858 the family. (*D*) Electrostatic maps and cartoon representations of the seven members of the
859 family. An electrostatic map represents the distribution of the electrostatic potential on the
860 surface of a protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon
861 structures are colored according to the distribution of their electrostatic potential. (*E*)
862 Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan
863 distance between the electrostatic maps of the proteins *(i,j)*.
864
865 **S12 Fig. Retrovirus capsid proteins family.** (*A*) Energy map distances matrix. It
866 corresponds to the subsection of the ADM for the retrovirus capsid proteins family (for the

867 construction of the ADM, see *Materials and Methods*). Each entry *(i,j)* represents the pairwise
868 energy map distance of the ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for
869 more details, see *Materials and Methods*). (*B*) Pairwise sequence identity matrix between all
870 members of the family. (*C*) Pairwise root mean square deviation (RMSD) matrix between all
871 members of the family. (*D*) Electrostatic maps and cartoon representations of the seven
872 members of the family. An electrostatic map represents the distribution of the electrostatic
873 potential on the surface of a protein (for more details, see Fig. S15 and *Materials and*
874 *Methods*). Cartoon structures are colored according to the distribution of their electrostatic
875 potential. (*E*) Electrostatic map distances matrix. Each entry *(i,j)* of the matrix represents the
876 Manhattan distance between the electrostatic maps of the proteins *(i,j)*.
877
878 **S13 Fig. Cystatins family.** (*A*) Energy map distances matrix. It corresponds to the subsection
879 of the ADM for the cystatins family (for the construction of the ADM, see *Materials and*
880 *Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the ligand pair *(i,j)*
881 averaged over the 74 receptors of the dataset (for more details, see *Materials and Methods*).
882 (*B*) Pairwise sequence identity matrix between all members of the family. (*C*) Pairwise root
883 mean square deviation (RMSD) matrix between all members of the family. (*D*) Electrostatic
884 maps and cartoon representations of the seven members of the family. An electrostatic map
885 represents the distribution of the electrostatic potential on the surface of a protein (for more
886 details, see Fig. S15 and *Materials and Methods*). Cartoon structures are colored according to
887 the distribution of their electrostatic potential. (*E*) Electrostatic map distances matrix. Each
888 entry *(i,j)* of the matrix represents the Manhattan distance between the electrostatic maps of
889 the proteins *(i,j)*.
890
891 **S14 Fig. Cyclophilins family.** (*A*) Energy map distances matrix. It corresponds to the
892 subsection of the ADM for the cyclophilins family (for the construction of the ADM, see

893 *Materials and Methods*). Each entry *(i,j)* represents the pairwise energy map distance of the
894 ligand pair *(i,j)* averaged over the 74 receptors of the dataset (for more details, see *Materials*
895 *and Methods*). (*B*) Pairwise sequence identity matrix between all members of the family. (*C*)
896 Pairwise root mean square deviation (RMSD) matrix between all members of the family. (*D*)
897 Electrostatic maps and cartoon representations of the seven members of the family. An
898 electrostatic map represents the distribution of the electrostatic potential on the surface of a
899 protein (for more details, see Fig. S15 and *Materials and Methods*). Cartoon structures are
900 colored according to the distribution of their electrostatic potential. (*E*) Electrostatic map
901 distances matrix. Each entry *(i,j)* of the matrix represents the Manhattan distance between the
902 electrostatic maps of the proteins *(i,j)*.
903
904 **S15 Fig. Generation of electrostatics, stickiness, hydrophobicity and circular variance**
905 **(CV) maps.** Here is presented an example of generation of the stickiness map for the structure
906 1AVW_A. (*A*) Generation of particles with a slightly modified Shrake-Rupley algorithm [59]
907 around the protein surface, leads to a homogenous shell of particles with a 1Å^2 density. Each
908 sphere is located at 5Å from the surface of the protein. The stickiness value of the closest
909 atom of the protein is attributed to each particle. In this example, spheres are colored
910 according to the stickiness of the protein surface. The procedure is similar for hydrophobicity
911 and CV. (*B*) The spherical coordinates of each sphere is represented on a 2-D map with an
912 equal-area sinusoidal projection, following the same protocol as described in Fig. 2 and
913 *Materials and Methods*. Each resulting dot is colored according to the same scale of (*A*). (*C*)
914 The map is smoothed following the protocol in Fig. 2 and *Materials and Methods*. The scale
915 is the same as in (*A*).
916

S16 Fig. Generation of stacked maps of a receptor. (*A*) Calculation of the 74 one-color maps (red ones in the example) of receptor #1. A value of one is associated to colored cells while zero is assigned to white cells. (*B*) Sum of the 74 one-color maps into a stacked map. Cell's intensity varies from 0 to 74 and corresponds to the number of time the cell is colored over the 74 ligands. (*C*) Filtering of the cells of low cell intensity (intensity < 17) and areas of too small size (< 4 cells) with a Dirichlet process mixture model simulation for image segmentation [57]. The procedure is repeated for each color stacked map.

S17 Fig. Blue and red stacked maps of 1P9D_U computed for each ligand family. (*A-L*) We compute the one-color stacked map of each family as the sum of the one-color maps resulting from the docking of each ligand of a same family with 1P9D_U.

S18 Fig. Boxplots of the size (in number of cells) of each energy class for all stacked maps. One should notice that the sum of the sizes of the 5 energy classes is superior to 1548 cells, which is the total size of a map, because a same cell of a stacked map can be assigned to several energy classes (Fig 8).

S19 Fig. Boxplots of the electrostatic potential of the protein surfaces depending on the energy class. The electrostatic potential is calculated for each protein following the protocol described in *Materials and Methods*. *p-values* between the variances of two "successive" energy classes were computed using the F-test.

S20 Fig. Boxplots of energy map pairwise distances between ligand pairs of conformers and pairs of homologous ligands (i.e. non-conformers pairs). For each receptor, we computed (i) the average of energy map distances of pairs of conformers, (ii) the average of

942 energy map distances of pairs of homologous ligands. P-values are calculated with an
943 unilateral Wilcoxon test.

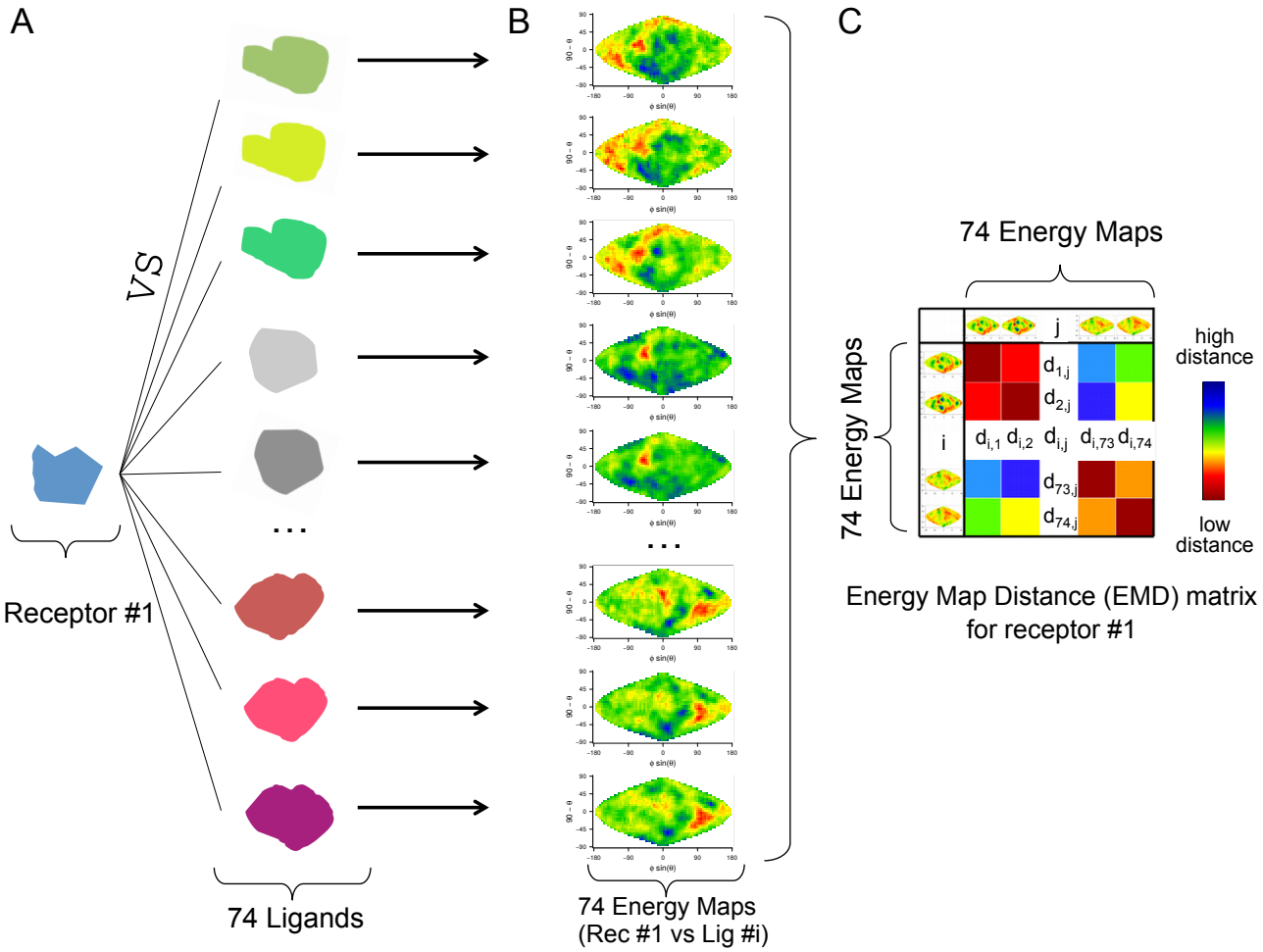
A
B
C
VS
Receptor #1
74 Ligands
74 Energy Maps
(Rec #1 vs Lig #i)
74 Energy Maps
74 Energy Maps
j
i
$d_{1,j}$
$d_{2,j}$
$d_{i,1}$ $d_{i,2}$ $d_{i,j}$ $d_{i,73}$ $d_{i,74}$
$d_{73,j}$
$d_{74,j}$
high distance
low distance
Energy Map Distance (EMD) matrix
for receptor #1

Fig 1.

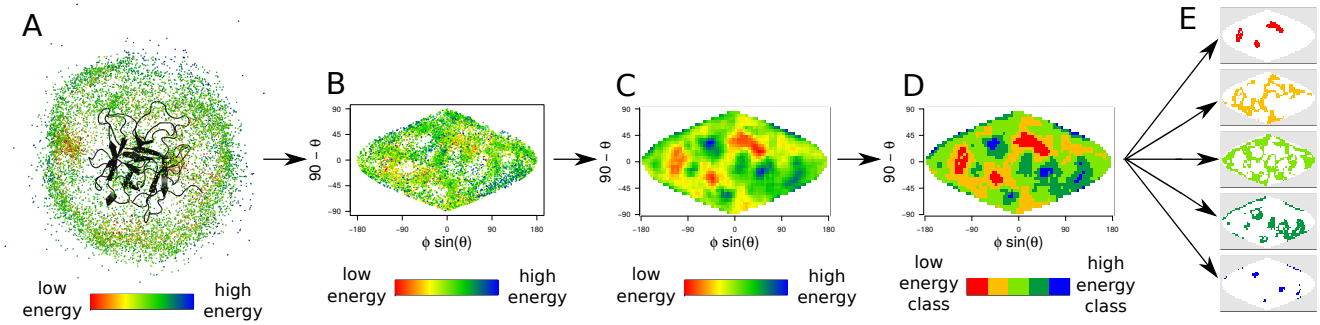
A
low energy
high energy
B
90 − θ
ϕ sin(θ)
low energy
high energy
C
90 − θ
ϕ sin(θ)
low energy
high energy
D
90 − θ
ϕ sin(θ)
low energy class
high energy class
E

Fig 2.

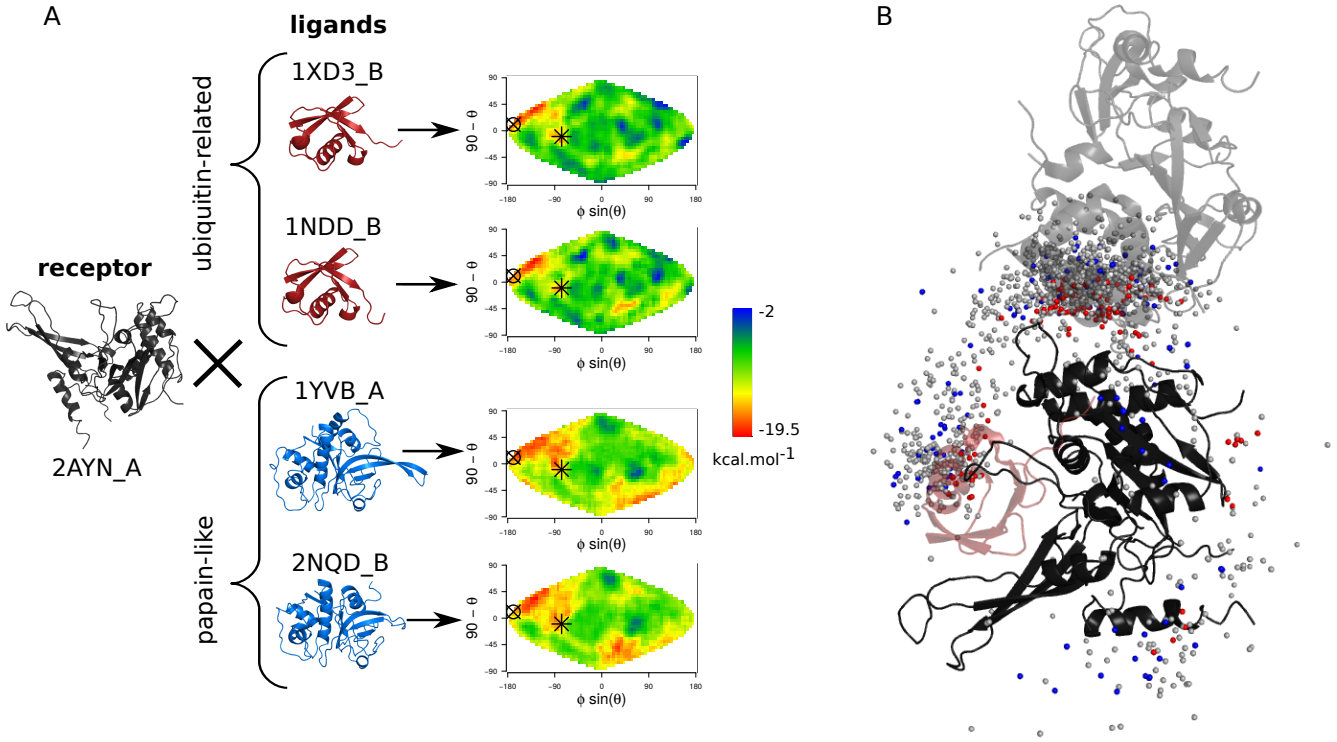
A
ligands
1XD3_B
1NDD_B
1YVB_A
2NQD_B
ubiquitin-related
papain-like
receptor
2AYN_A
90 − θ
ϕ sin(θ)
-2
-19.5
kcal.mol-1
B

Fig 3.

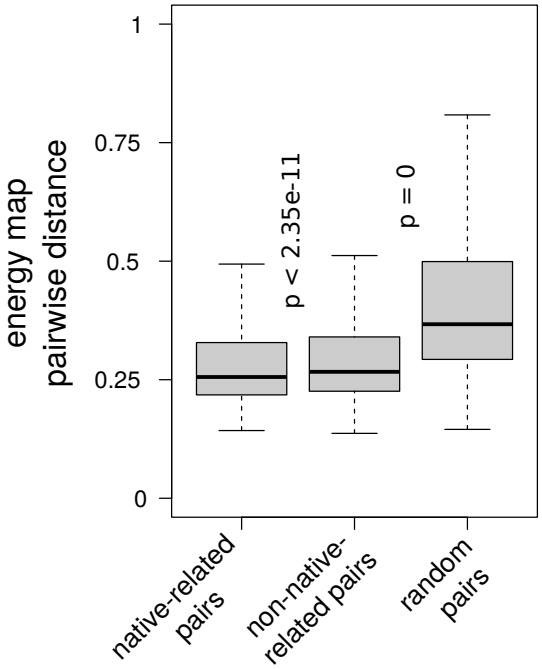
energy map
pairwise distance
1
0.75
0.5
0.25
0
p < 2.35e-11
p = 0
native-related pairs
non-native-related pairs
random pairs

Fig 4.

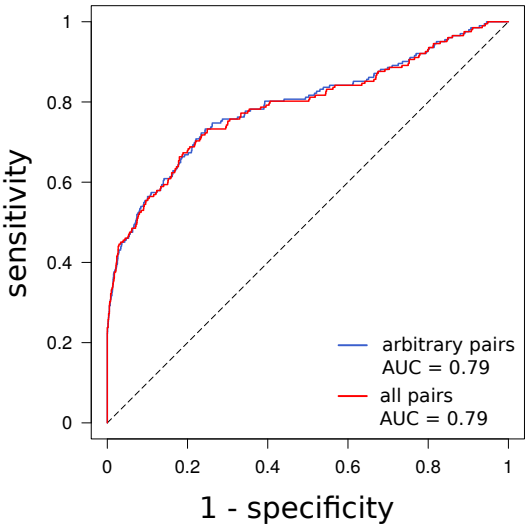
sensitivity
1 - specificity
0
0.2
0.4
0.6
0.8
1
arbitrary pairs
AUC = 0.79
all pairs
AUC = 0.79

Fig 5.

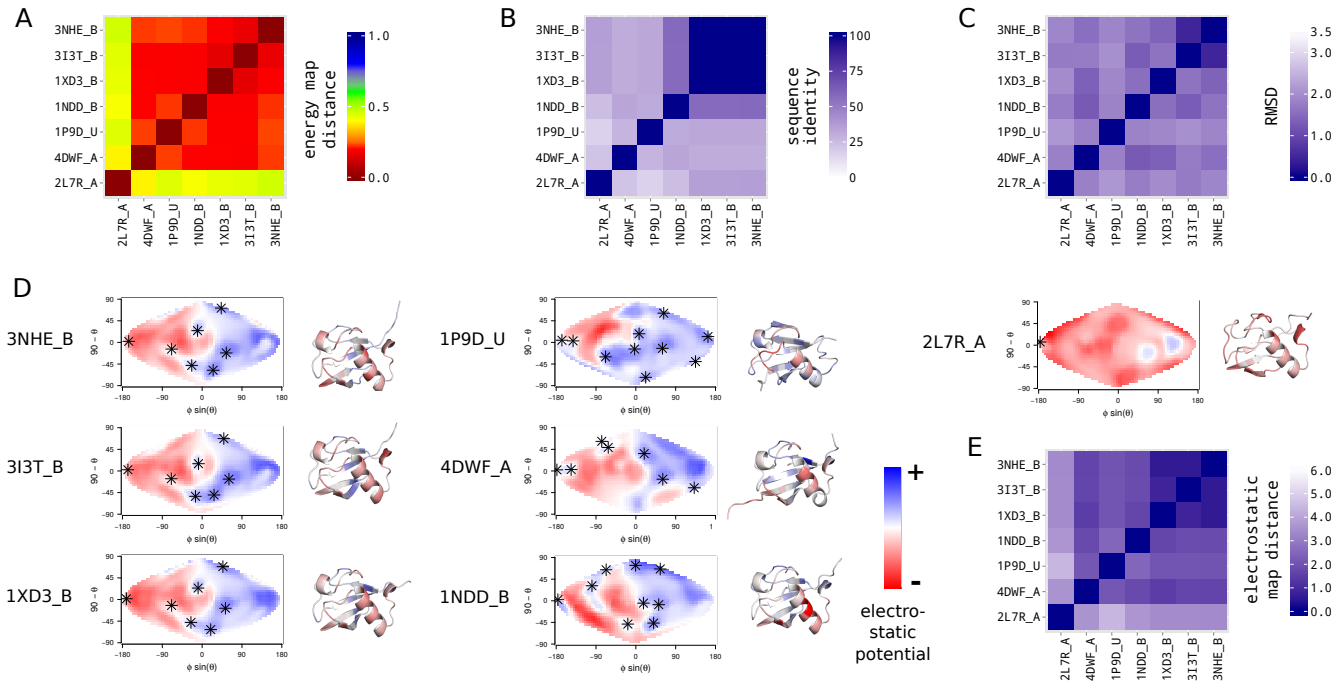
A
3NHE_B
3I3T_B
1XD3_B
1NDD_B
1P9D_U
4DWF_A
2L7R_A
energy map distance
1.0
0.5
0.0
B
sequence identity
100
75
50
25
0
C
RMSD
3.5
3.0
2.0
1.0
0.0
D
3NHE_B
1P9D_U
2L7R_A
3I3T_B
4DWF_A
1XD3_B
1NDD_B
ϕ sin(θ)
90 − θ
+
-
electro-
static
potential
E
electrostatic map distance
6.0
5.0
4.0
3.0
2.0
1.0
0.0

Fig 6.

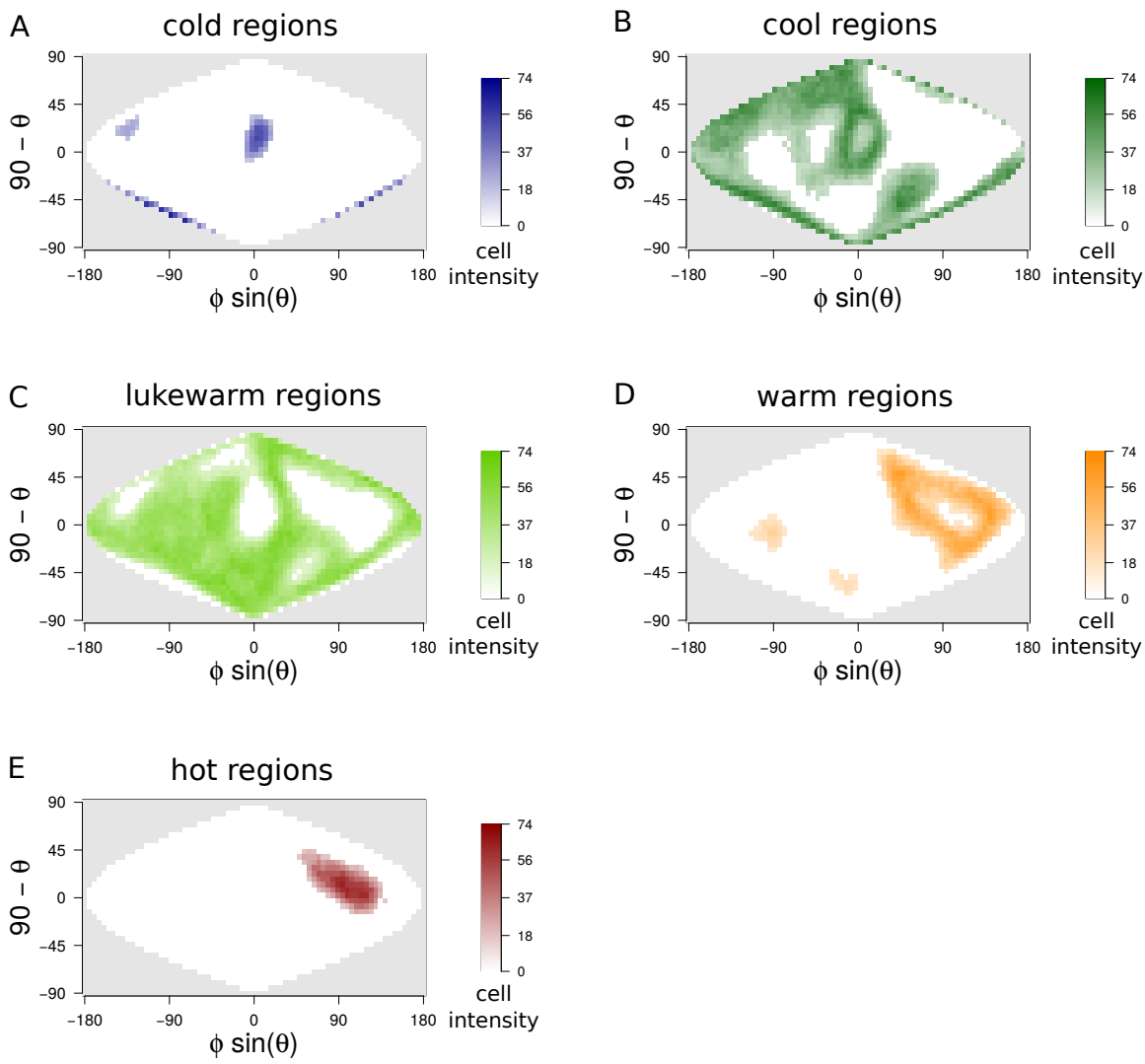
A
cold regions
B
cool regions
C
lukewarm regions
D
warm regions
E
hot regions
90 − θ
ϕ sin(θ)
90
45
0
−45
−90
−180
−90
0
90
180
74
56
37
18
0
cell
intensity

Fig 7.

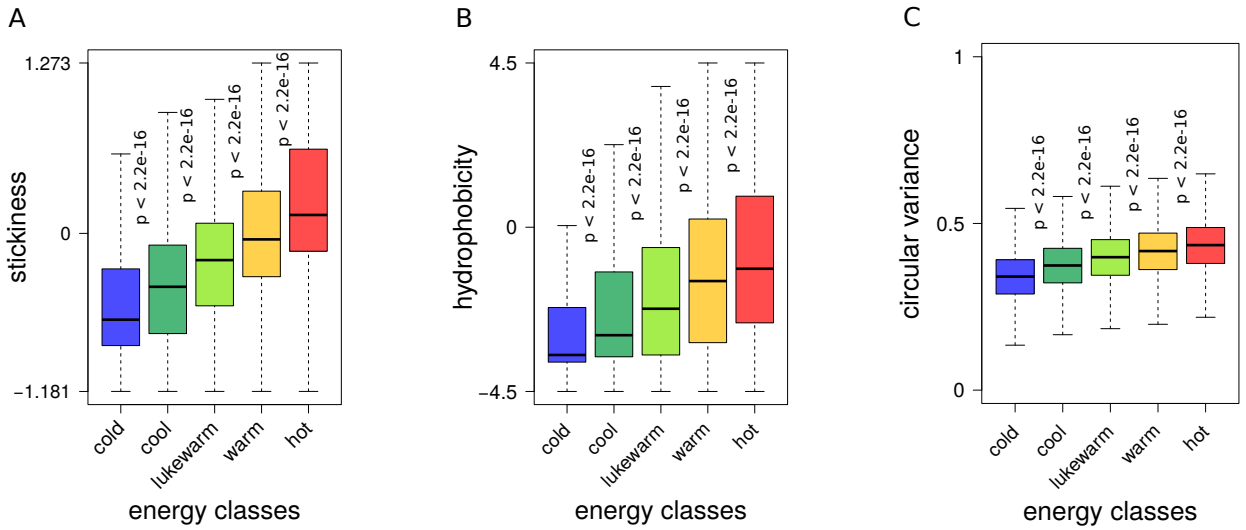
A
1.273
0
−1.181
stickiness
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16
cold
cool
lukewarm
warm
hot
energy classes
B
4.5
0
−4.5
hydrophobicity
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16
cold
cool
lukewarm
warm
hot
energy classes
C
1
0.5
0
circular variance
p < 2.2e-16
p < 2.2e-16
p < 2.2e-16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

Fig 8.