DIP/Dpr interactions and the evolutionary design of specificity in protein families

Alina P. Sergeeva, Phinikoula S. Katsamba, Filip Cosmanescu, Joshua J. Brewer, Goran Ahlsen, Seetha Mannepalli, Lawrence Shapiro, Barry Honig

Affiliations

Department of Biochemistry and Molecular Biophysics, Columbia University, New York, NY, USA

Alina P. Sergeeva, Phinikoula S. Katsamba, Filip Cosmanescu, Joshua J. Brewer, Goran Ahlsen, Seetha Mannepalli, Lawrence Shapiro, Barry Honig

Zuckerman Mind, Brain and Behavior Institute, Columbia University, New York, NY, USA

Alina P. Sergeeva, Phinikoula S. Katsamba, Filip Cosmanescu, Joshua J. Brewer, Goran Ahlsen, Seetha Mannepalli, Lawrence Shapiro, Barry Honig

Department of Systems Biology, Columbia University, New York, NY, USA

Alina P. Sergeeva, Lawrence Shapiro, Barry Honig

Department of Medicine, Columbia University, New York, NY, USA

Barry Honig

Abstract

Differential binding affinities among closely related protein family members underlie many biological phenomena, including cell-cell recognition. *Drosophila* DIP and Dpr proteins mediate neuronal targeting in the fly through highly specific protein-protein interactions. DIPs/Dprs segregate into seven specificity subgroups defined by binding preferences between their DIP and Dpr members. Here we describe a novel sequence-, structure- and energy-based computational approach, combined with experimental binding affinity measurements, to reveal how specificity is coded on the canonical DIP/Dpr interface. We show that binding specificity of DIP/Dpr subgroups is controlled by

27 “negative constraints”, which interfere with binding. To achieve specificity, each subgroup utilizes a
28 different combination of negative constraints, which are broadly distributed and cover the majority of
29 the protein-protein interface. We discuss the structural origins of negative constraints, and potential
30 general implications for the evolutionary origins of binding specificity in multi-protein families.
31

32 **Introduction**

33 Over the course of evolution, gene duplications followed by sequence divergence have
34 generated numerous protein families whose homologous members have distinct binding
35 specificities. In many cases there are only subtle differences in the sequence and structure of closely
36 related family members, yet these can have profound functional consequences. Families of cell-cell
37 adhesion proteins offer many such examples where seemingly small changes in sequence generate
38 new specificities for protein-protein interaction specificities. For example, invertebrate Dscam[1,2] and
39 vertebrate clustered protocadherin[3,4] neuronal “barcoding” proteins each display strict intra-family
40 homophilic specificity. While one Dscam (or Pcdh) isoform may have >90% sequence identity to
41 another family member, the few differences between isoforms selected by nature ensure that only
42 homophilic recognition occurs; this is critical to their mechanism of function in diversification of
43 neuronal identities and self/non-self-discrimination. In other cases, for example the type I and type II
44 classical cadherins[5-7], which pattern epithelia and other tissue structures, specificity is less strict.
45 While classical cadherins also exhibit homophilic binding, in addition they show strong heterophilic
46 binding to other select family members. In another example, the nectin[8] adhesion protein family
47 encodes interactions between members that are mainly heterophilic, in some cases to form a
48 checkerboard pattern between two cell types, each expressing a cognate nectin. The specificity
49 patterns of these protein families underlie their biological functions, and are conserved in evolution.
50 Understanding how specificity is coded on multi-protein adhesion protein families is thus critical.
51 Here we carry out a comprehensive computational and experimental study of specificity
52 determinants in two interacting families of neuronal recognition proteins, the 21-member Dpr

53 (Defective proboscis extension response) and the 11-member DIP (Dpr Interacting Proteins). These
54 proteins have been extensively characterized structurally[9-11], and their interactions were
55 characterized quantitatively with biophysical measurements[11], thus offering an ideal system to study
56 the evolutionary design of specificity on protein-protein interfaces. DIPs and Dprs are expressed in
57 cell-specific patterns throughout the developing nervous system[12]. DIPs preferentially bind Dprs,
58 and a network of specific heterophilic interactions is formed between members of the two families.
59 This molecular binding network is correlated with synaptic specificity in the fly retina, suggesting that
60 DIP/Dpr interactions play an important role in neuronal patterning[9,13].

61 The extracellular regions of DIP and Dpr family members consist of three and two tandem Ig-
62 like domains, respectively[13]. Homodimerization is observed for some DIPs and Dprs, and homo- and
63 hetero-dimerization is mediated by an interface formed between the membrane-distal Ig1 domains
64 (Figure 1A). Surface plasmon resonance (SPR) showed that members of both families have distinct
65 binding profiles, with DIPs and Dprs initially classified as forming four specificity subgroups.[11] In the
66 current work we extended the number of subgroups to seven based primarily on the strongest
67 heterophilic binding preferences but also on DIP/Dpr sequence similarity (see color-coded subgroup
68 assignment in Figure 1B). Our DIP/Dpr grouping is somewhat different than that published by Cheng
69 et al.[10] due in part to the fact that these authors did not include DIP-κ and DIP-λ, whose binding
70 preferences had been previously mapped[11]. Additional differences could be due to the biophysical
71 approaches used to measure DIP/Dpr binding affinities in Cosmanescu et al.[11] and Cheng et al.[10]
72 (see methods).

73 The Ig1 domains of the *Drosophila Melanogaster* DIPs and Dprs have intra-family pairwise
74 sequence identities greater than about 50% and 40%, respectively, while the average identity
75 between individual DIPs and Dprs is about 30%. Binding interfaces for crystallographically
76 determined hetero-dimer structures are essentially identical – superimposing to within 1Å (Figure
77 1C)[11]. The central question we address here is how DIPs and Dprs that are so closely related in
78 sequence and structure can exhibit such highly specific pairwise interactions. Previous studies have

79 identified specificity residues for select DIP/Dpr interactions[9-11]. Here, we analyze specificity for the
80 family as a whole. Our results reveal the central role of “negative constraints” - destabilizing
81 interfacial positions, which preclude complex formation between members of some specificity
82 groups, while allowing it between others. We use the term negative constraints here to denote an
83 amino acid in a cognate interface that interferes with binding to a non-cognate partner. The term
84 negative constraint has been used in the field of protein design[14-17] to denote a domain that must be
85 designed against, in effect an “anti-target”. In contrast, our use of the term here focuses on
86 individual amino acids rather than entire domains.
87

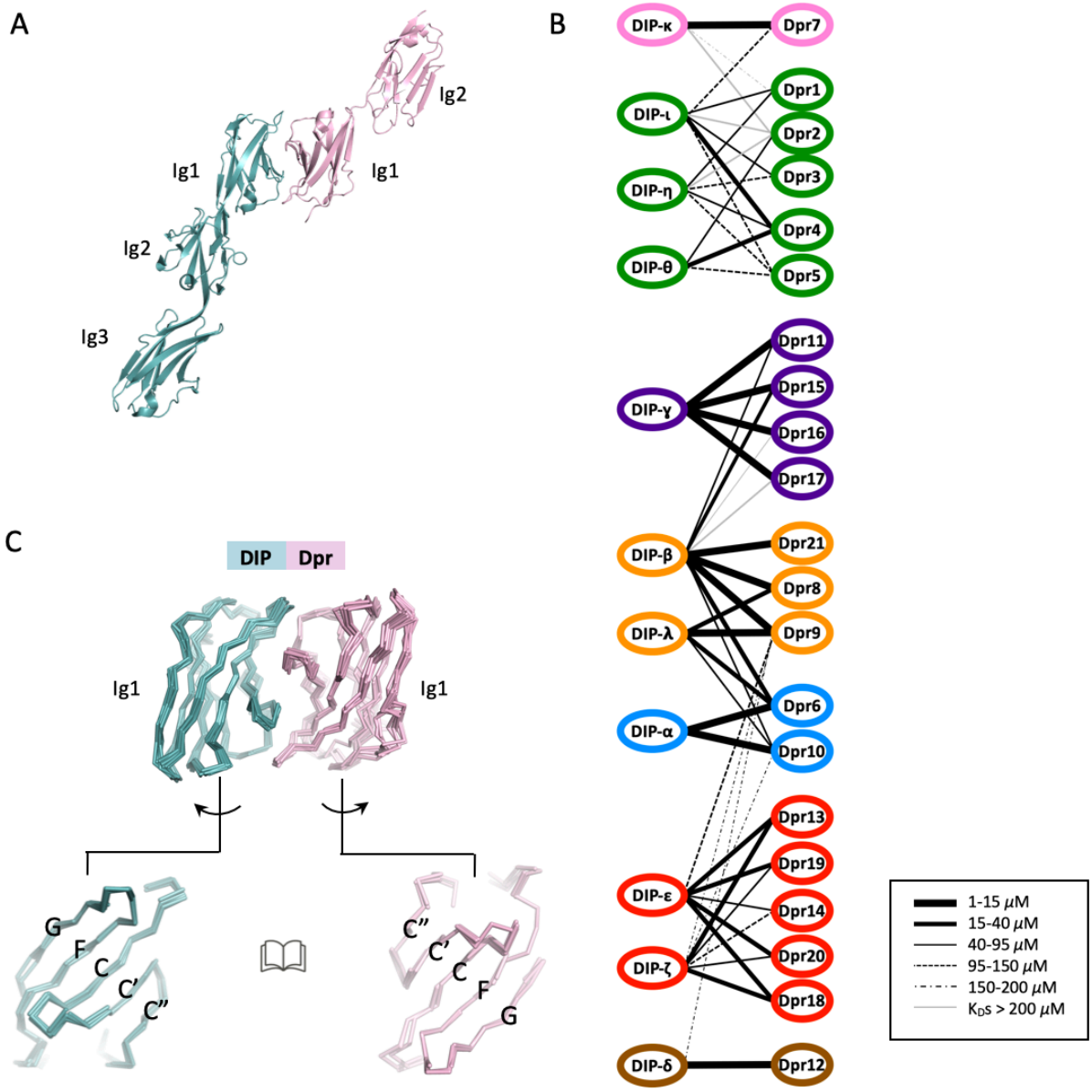

Figure 1. Structure and interaction properties of DIPs and Dprs. (A) Ribbon representation of the DIP/Dpr heterodimer (PDBID: 6EG0)[11] – DIP shown in cyan, Dpr in pink. **(B)** Affinity-based binding interactome of DIPs and Dprs of *Drosophila Melanogaster*. Line thickness indicates the K_D for a given pairwise interaction, as specified in the boxed inset. Groupings are color-coded based on their heterophilic binding preference. The interactome is based on previously published SPR data[11]. **(C)** Structural alignment of Ig1 domains in DIP/Dpr complexes (see Methods for PDBIDs). DIPs and Dprs interact via their CC'C"FG surfaces.

Since there are a total of forty-nine possible DIP/Dpr subfamily pairs and only seven bind strongly, there must be forty-two sets of negative constraints that preclude incorrect pairing. These are coded on a pseudo-symmetric Ig1'-Ig1" interface of about 1900 Å^2 buried surface area and comprising 33 interfacial residues on the DIP side and 33 interfacial residues on the Dpr side. We

99 recognize that non-interfacial residues may also contribute to specificity but the primary
100 determinants will, in most cases, be part of the interface and these are the focus of the current work.
101 We begin by asking what can be learned from sequence alone and find that the information
102 available is useful but incomplete. Our structure-based approach involves building homology models
103 of hypothetical complexes formed between all DIPs and Dprs and, also using solved complexes
104 where available, to calculate the energetic consequences of mutating interfacial residues in one
105 family member to those of every other. We tested various programs that calculate binding free
106 energy changes ($\Delta\Delta$Gs) resulting from mutations and found that FoldX[18] offered the best
107 combination of accuracy and computational efficiency. Using FoldX, we identified negative
108 constraints for every DIP/Dpr pair and then confirmed a subset of our predictions using quantitative
109 biophysical assays (AUC and SPR).
110 Our results provide a detailed account of the design of complex specificity on a canonical
111 protein-protein interface. Most of the interface is used to create negative constraints with virtually all
112 expected energetic terms (steric hindrance, Coulombic repulsion, unsatisfied buried charges and/or
113 polar groups) used to weaken undesired complex formation between different subgroup pairs. Some
114 locations on the interface are used by multiple pairs while others are used in only a small number of
115 cases. In general, more than one negative constraint is required to significantly weaken binding. We
116 also discuss the trade-off between specificity and affinity, a phenomenon that appears essential for
117 the generation of multiple specificity groups[14,15,19]. Overall, our analysis reveals general principles
118 about the design of protein families whose members are very similar in sequence and structure and
119 yet exhibit exquisitely controlled binding affinities and specificities.
120
121 **Results**
122 Sequence analysis of DIPs and Dprs
123 Figure 2 plots an alignment of sequence logos corresponding to interfacial residues of DIP
124 and Dpr *Insecta* orthologs (see Methods for details). Interfacial positions of DIPs and Dprs were

125 assigned based on non-zero change in accessible surface area per residue upon complex formation
126 in experimentally solved crystal structures. Interfacial positions of DIP/Dpr complexes of unknown
127 structure were inferred from multiple sequence alignment (residue correspondence with interfacial
128 positions of solved complexes). The family-wide numbering of interfacial positions (*i* = 1-33,
129 designated above the logos in Figure 2) is used throughout the paper (see correspondence between
130 the family-wide numbering and Uniprot numbering (given throughout this paper in curly brackets) in
131 Table S1).

132 Four positions shown in yellow are occupied almost exclusively by hydrophobic residues
133 which form the hydrophobic core of the DIP/Dpr interface[10] (Figure 2A). Additionally, shown in purple
134 is a conserved Gln in both subfamilies and a conserved Asn in DIPs, which interlock the Ig1'-Ig1"
135 interface by forming buried sidechain-to-backbone hydrogen bonds (Figure 2B). Residues that are
136 conserved across subfamilies cannot be responsible for subgroup specificity but changes in the
137 identity of individual conserved hydrophobic residues can play a role (see below). Rather, potential
138 specificity determinants will generally correspond to positions that are conserved in at least one
139 subfamily but differ in between subfamilies.

140 We tested the ability of a number of sequence-based methods to identify specificity
141 determinants in the seven specificity subgroups. These include GroupSim[20], SDPpred[21], SPEER[22]
142 and Multi-Harmony[23].
143

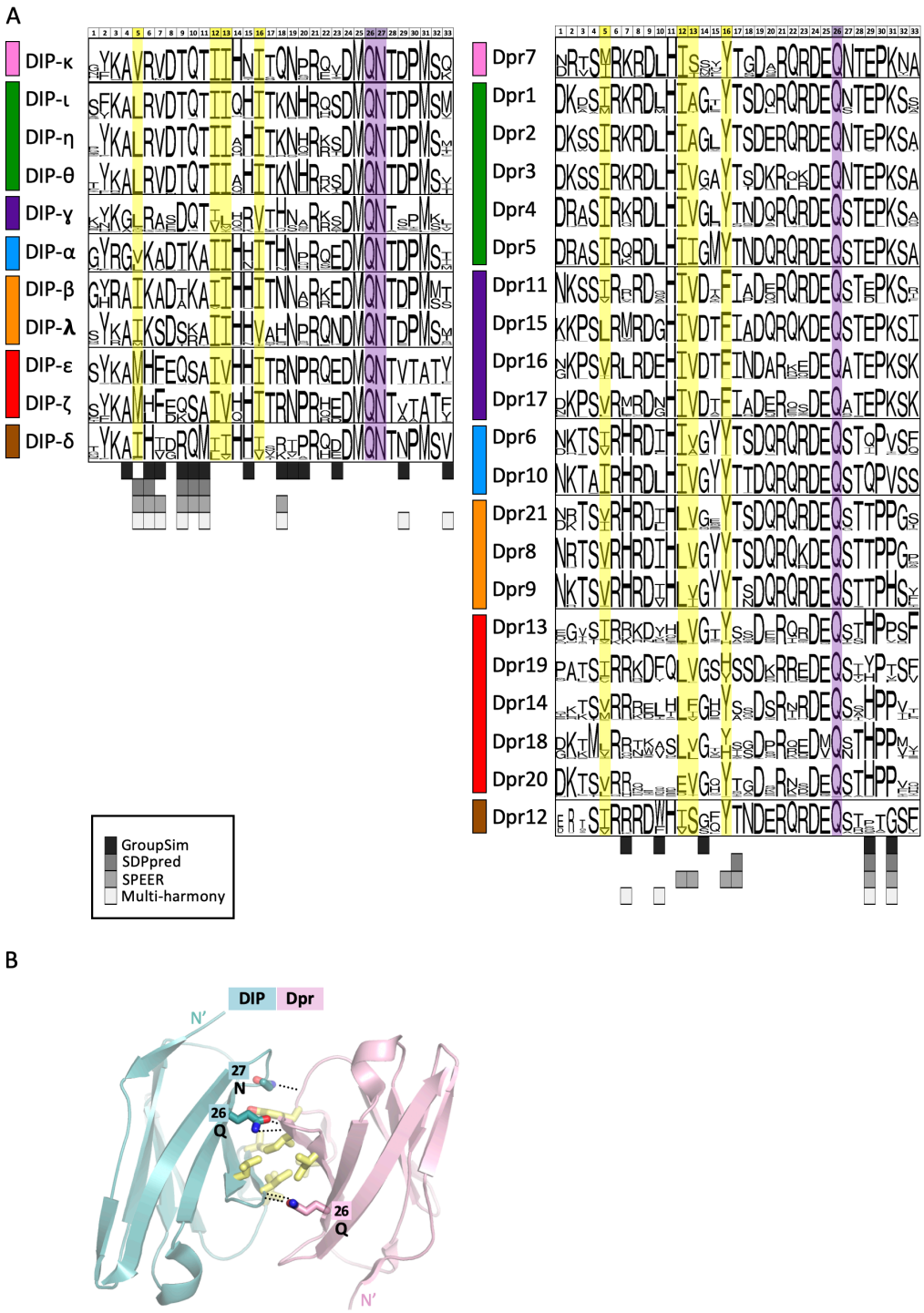

Figure 2. Sequence-based predictions of DIP/Dpr specificity residues and common binding determinants. (A) Alignment of sequence logos corresponding to interfacial residues of DIP and Dpr Insecta orthologs. Predictions of specificity determining positions using sequence-based methods (GroupSim, SDPpred, SPEER, and Multi-Harmony) are given below logos in different shades of grey as specified in boxed inset. Bar on the left of logos indicates color of the DIP/Dpr subgroup (Figure 1B). Family-wide numbering of interfacial positions is given above the alignment. The correspondence between protein specific (UniProt) and family-wide numbering is given in Table S1. See Methods for details on sequence logo preparation. **(B)** Common binding determinants of DIP/Dpr dimers depicted on ribbon Ig1'-Ig1" interfaces in sticks. Hydrophobic interfacial residues are in yellow. Conserved buried hydrogen bonds are depicted by dotted lines. Interfacial positions enclosed in boxes follow family-wide numbering and are annotated with identities of conserved amino-acids.

155
156 As input, we used a multiple sequence alignment of just interfacial residues with protein sequences
157 of 1732 DIP or 2570 Dpr orthologs segregated into seven specificity subgroups (see Figure 2 and
158 methods for details). Nine specificity determining positions were identified based on a consensus of
159 at least 3 out of 4 of these methods, and 14 additional positions, were predicted by at least one
160 method (see predictions below logos in grey, Figure 2). These expand the number of predicted
161 specificity determining sites compared to our previous study[11] where we relied on visual inspection
162 of a multiple sequence alignment of DIP/Dpr sequences in *Drosophila* Melanogaster alone. Below
163 we compare the predictions of these sequence-based methods to an energy-based analysis of DIP
164 and Dpr structures.
165
166 Identifying negative constraints
167 We hypothesized that negative constraints are the essential factor that defines DIP/Dpr
168 subgroups and developed a computational strategy to identify and characterize these putative
169 constraints. Since each subgroup (e.g. *i* and *j*) has both DIP and Dpr members, for purposes of
170 discussion we can define four sets of proteins, DIP-*i*, DIP-*j*, Dpr*i* and Dpr*j*. We need to explain why
171 members of set DIP-*i* don't bind to members of set Dpr*j* and, in turn, why members of set DIP-*j* don't
172 bind to members of set Dpr*i*. Since Dpr*i* binds to DIP-*i* while Dpr*j* does not, we first use FoldX
173 calculations to identify residues of Dpr*i* that weaken binding to DIP-*i* when mutated into a residue of
174 Dpr*j*. Thus, we identify negative constraints of Dpr*j* by studying their predicted effects on the binding
175 of mutated Dpr*i* to DIP-*i*. The experimental tests follow the same logic: the K_Ds of mutant Dpr*i*
176 binding to DIP-*i* are used to validate predicted negative constraints on Dpr*j*.
177 The overall procedure was to start with one or more structures of DIP/Dpr complexes for
178 each subgroup (the template complex) and; 1) to mutate every interfacial residue in the DIP
179 protomer to all residues appearing at the same position in DIPs of other specificity subgroups and
180 then 2) calculate ΔΔGs for the binding of each DIP to the Dpr in the complex. 3) Repeat the

181 procedure for the Dpr in the complex and then calculate ΔΔGs for its binding to each DIP.
182 Templates that were used include: *blue subgroup* - crystal structure of Dpr6/DIP-α and two
183 conformations of Dpr10/DIP-α[9,10]; *purple subgroup* – crystal structure of Dpr11/DIP-γ[10]; *green*
184 *subgroup* – crystal structures of Dpr4/DIP-η and for two conformations each for Dpr1/DIP-η and
185 Dpr2/DIP-θ)[10,11]; *pink subgroup* - homology model of Dpr7/DIP-κ; *orange subgroup* - homology
186 model of Dpr8/DIP-β; *brown subgroup* - homology model of Dpr12/DIP-δ; *red subgroup* -
187 homology model of Dpr13/DIP-ε. Overall, we predicted the effects of 3044 mutations of 33 interfacial
188 residues of 13 template complexes on the DIP and Dpr side into residues occurring in DIPs and
189 Dprs of other subfamilies (see a supplementary excel file for the FoldX data). Notably, the results
190 depend heavily on the use of homology models (see Methods for details of how they were
191 constructed).
192 Negative constraints were defined as positions predicted to cause destabilizing effects for
193 *every* DIP or Dpr fly subgroup member in the context of binding to a member of another Dpr or DIP
194 subgroup, respectively. Our energy-based filter also requires that the effect is destabilizing for every
195 subfamily template complex used in the FoldX calculations (see energy filter in methods for details).
196 In addition, we introduced a second filtering step, whereby we only defined a negative constraint as
197 a position that would also be predicted for other *Insecta* species so as to focus on negative
198 constraints that are evolutionarily conserved. The evolutionary filter is based on the assumption that
199 orthologs in the same subfamily conserve their interaction specificity (see methods for data
200 supporting the assumption). To clarify, the energy-based filter is based on FoldX calculations of
201 structures of fly proteins whereas the evolutionary filter is based on comparison of ortholog
202 sequences of different species of insects. The evolutionary filter involved a qualitative analysis of
203 interfacial residues at corresponding positions of DIP-*i* and DIP-*j* (or Dpr*i* and Dpr*j*) in *Insecta*
204 multiple sequence alignments to see if the identities, size or biophysical properties of amino acids in
205 all members of subfamily "*i*" were different from those of subfamily "*j*" at the predicted position (see
206 evolutionary filter in methods for details).

207 A subset of our predictions (see below) was tested experimentally. Overall, we identified
208 negative constraints for 42 combinations of non-interacting DIP and Dpr subgroups (see
209 supplemental excel file for all FoldX data).
210
211 Evaluation of methods to calculate ΔΔG
212 We tested six algorithms that calculate binding affinity changes upon mutation of individual
213 residues (ΔΔGs) based on reports of their performance in the literature[24-30]. These include FoldX,[18,31]
214 mCSM,[30] BeAtMusic,[27] MutaBind,[29] Rosetta flex ddG,[24] and BindProfX[32]. FoldX, mCSM, and
215 BeAtMusic require from seconds to few minutes of CPU time per single point mutation, while
216 MutaBind, Rosetta flex ddG, and BindProfX take more than an hour. Our evaluation tests were
217 based on 25 experimental ΔΔG values obtained via SPR or AUC measurements of DIP/Dpr and
218 DIP/DIP dimers. Of these, two were published[11] and twenty-three are new (see Figure S1 for
219 supporting AUC and SPR data). Data on the performance of these methods is summarized in Table
220 S2.
221 As can be seen in Table S2, FoldX, MutaBind, and Rosetta flex ΔΔG perform best on our
222 dataset as they feature the highest Pearson correlation coefficients (PCC). MutaBind is the best
223 performer but it fails to identify stabilizing mutations (shown in green, Table S2) - a failure of all
224 machine learning methods we tested, which is probably reflective of an overrepresentation of
225 destabilizing mutations in training datasets. In addition, we preferred to use methods based on
226 physics-based force fields since these were easiest to interpret in structural terms. Given its
227 performance in terms of speed and accuracy, we settled on FoldX as a computational guide for the
228 identification of negative constraints.
229
230 Negative constraints between blue and purple DIP/Dpr subgroups
231 We discuss Dpr10/DIP-α(blue subgroup) and Dpr11/DIP-γ (purple subgroup) in detail since
232 these pairwise DIP/Dpr interactions have revealed clear associated phenotypes in the central and

233 peripheral nervous systems of *Drosophila Melanogaster*[9,33-35]. We consider four sets of negative
234 constraints that weaken blue/purple inter-subgroup DIP/Dpr binding: a) On purple subgroup Dprs; b)
235 On blue subgroup DIP-α; c) On blue subgroup Dprs; and d) On purple subgroup DIP-γ.
236
237 a) *Negative constraints on purple subgroup Dprs that weaken binding to blue subgroup DIP-*
238 *α.* As described in the protocol summarized above, we identify negative constraints on purple Dprs
239 by computationally mutating every interfacial amino acid on blue Dprs to those of purple Dprs and
240 calculating $\Delta\Delta$Gs for binding to DIP-α. This procedure is applied for every template complex listed
241 above. Figure 3A summarizes results using one of the DIP-α/Dpr10 crystal conformations (chains C
242 and D of the asymmetric unit). The figure shows an alignment of the four purple Dprs with blue
243 Dpr10. Each column in the alignment is color-coded based on the $\Delta\Delta$G for mutating a residue in
244 Dpr10 to the aligned residue in a purple Dpr. Every position where the resulting $\Delta\Delta$G is positive for
245 all purple Dprs is a candidate for a negative constraint. Predictions are then filtered based on the
246 requirement that $\Delta\Delta$G >= 0.05 kcal/mol for all available blue subgroup crystal structures (see Figure
247 S2). In addition to the energy filter, only positions that pass evolutionary filter (see above) are
248 accepted.
249

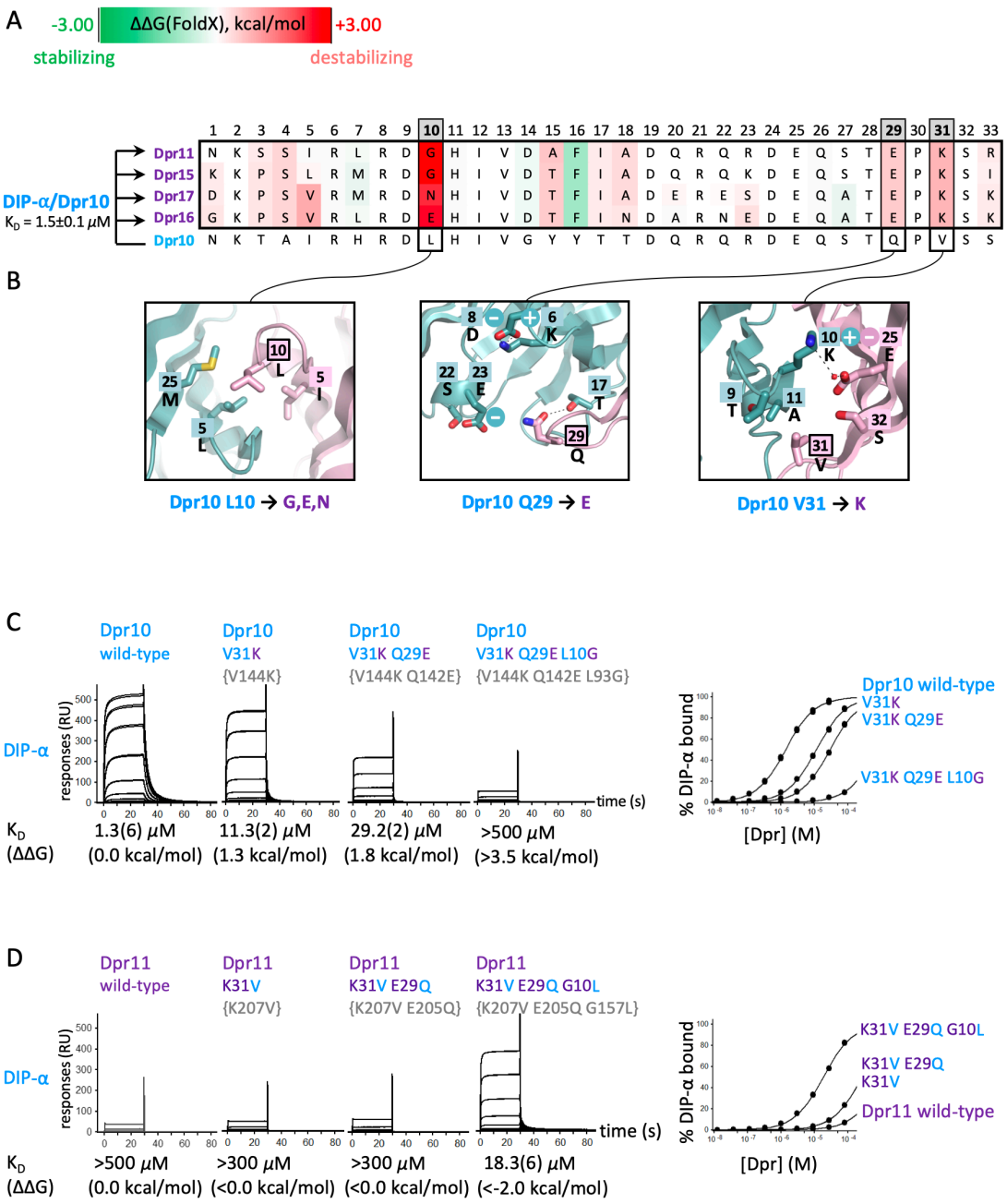

Figure 3. Negative constraints on purple Dprs 11/15/16/17 which prevent binding to the blue DIP-α. (A) Alignment of interfacial residues of the four purple group Dprs and of Dpr10. Residues are color-coded based on the calculated ΔΔG, as specified in the color bar at the top of the figure. Predicted negative constraint positions are in grey. The average K_D for wild-type DIP-α/Dpr10 interaction based on six independent experimental measurements is given on the left of the alignment. All calculations are carried out on the DIP-α/Dpr10 complex indicated to the left of the figure. The arrows indicate that residues in Dpr10 were computationally mutated to those of the four purple Dprs, as indicated. **(B)** Structural details of wild-type environment for each position where a mutation is predicted to cause a negative constraint (enclosed in a bold box). **(C) and (D).** SPR binding analysis of Dpr10 and Dpr11 mutants, representing negative constraints, to DIP-α. Each row shows SPR sensorgrams of Dpr analytes binding over a DIP-α immobilized surface. An overlay of the binding isotherms for each surface is shown to the right. K_D (ΔΔG) values for each DIP/Dpr interaction are listed below the SPR sensorgrams. For each K_D, the number in parenthesis represents the fitting error in the last significant figure, in μM. Family-wide numbering of interfacial positions is used for all the mutants. Uniprot numbering is given in curly brackets.

265 This procedure identified three interfacial positions, 10, 29, and 31 (Figure 3A), as negative
266 constraints. Notably, stabilizing mutations (shown in green) suggest that DIP/Dpr complexes are not
267 fully optimized for binding (see below). The destabilizing effect of these four negative constraints can
268 be understood in biophysical terms (Figure 3B). Dpr10 has a conserved hydrophobic residue at
269 position 10 whose interaction with buried hydrophobic residues of DIP-α contributes favorably to
270 binding (first panel in Figure 3B). When this interaction is lost via mutation to glycine (Dpr11 and
271 Dpr15), a negatively charged glutamate (Dpr16), or a polar asparagine (Dpr17), the DIP/Dpr
272 complexes are destabilized either through the creation of a cavity in the hydrophobic interface or
273 from the desolvation penalty associated with placing charged or polar atoms in the hydrophobic
274 environment. Two other negative constraints correspond to conserved charged Glu and Lys at
275 interfacial positions 29 and 31. These residues form salt bridges in the purple DIP/Dpr subfamily, but
276 would create unsatisfied charges in the context of binding to DIP-α.
277 We tested these predictions with SPR measurements of mutant proteins. Figure 3C displays
278 SPR sensorgrams of Dpr10 wild-type and its mutants passed over a chip immobilized with DIP-α. As
279 can be seen in the figure, the V31K single mutant increases the K_D relative to wild-type by about an
280 order of magnitude; adding a second mutation, Q29E, further increases the K_D by about a factor of 2
281 while adding a third mutation L10G results in a triple mutant with a $K_D > 500$ μM. Thus, introducing
282 Dpr11-like negative constraints to Dpr10 at these three positions is sufficient to kill binding between
283 Dpr10 and DIP-α. We then carried out a reverse set of experiments aimed at removing negative
284 constraints on Dpr11 to enhance its binding to DIP-α. As can be seen in Figure 3D, the K31V single
285 mutant and the K31V,E29Q double mutant have very little effect, while the triple Dpr11 mutant binds
286 to DIP-α with a K_D of 18.3 μM – weaker than wild-type Dpr10 but still quite strong. Overall, our
287 results indicate that we have successfully identified the negative constraints on the Dpr side of the
288 interface that preclude binding of purple group Dprs to blue group DIPs.
289

290 *b) Negative constraints on the blue subgroup DIP-α that weaken binding to purple subgroup*
291 *Dprs.* FoldX calculations identify positions 6,9,11,15,22 as negative constraints (Figure 4A). We
292 validated a subset of these negative constraints with SPR, positions 9, 15 and 22, and found that all
293 weakened the binding of DIP-γ to Dpr11, although only position 9 had a substantial effect, increasing
294 the K_D of DIP-γ/Dpr11 binding from 7.9 to 37.2 μM (Figure S3A). As can be seen in Figure 4A, there
295 is a charged residue at each of these positions in DIP-γ that forms an ion pair with a complementary
296 charge on purple group Dprs. Each of these positions contains a neutral polar residue in DIP-α
297 which would then disrupt the ion pair and leave an unsatisfied partially buried charge in a
298 hypothetical complex with purple group Dprs.
299
300 c) *Negative constraints on blue subgroup Dprs that weaken binding to purple subgroup DIP-γ.*
301 Positions 7,14,16,29,31 on Dpr6 and Dpr10 are predicted to be negative constraints (Figure 4B).
302 Positions 29 and 31 were tested and, as can be seen in the figure, the single K31V mutant increased
303 the K_D of wild-type Dpr11 binding to DIP-γ from 7.9 μM to 22.7 μM while the K31V,E29Q double
304 mutant with two negative constraints had a K_D of 131 μM, almost a factor of 20 greater than wild-
305 type (Figure S3B). The source of the destabilization is evident from the structure figures shows in
306 Figure 4B.
307 The K to V mutation at position 31 would create a steric clash with Thr at position 11 of DIP-γ
308 and leave two unsatisfied negative charges in the DIP-γ/Dpr11 interface (Asp9 of DIP-γ and Glu25 of
309 Dprs), one of which will also be in a hypothetical complex with blue Dprs (Glu25). The E29Q
310 mutation breaks a salt bridge with Arg6 of DIP-γ leaving an unsatisfied positive charge.

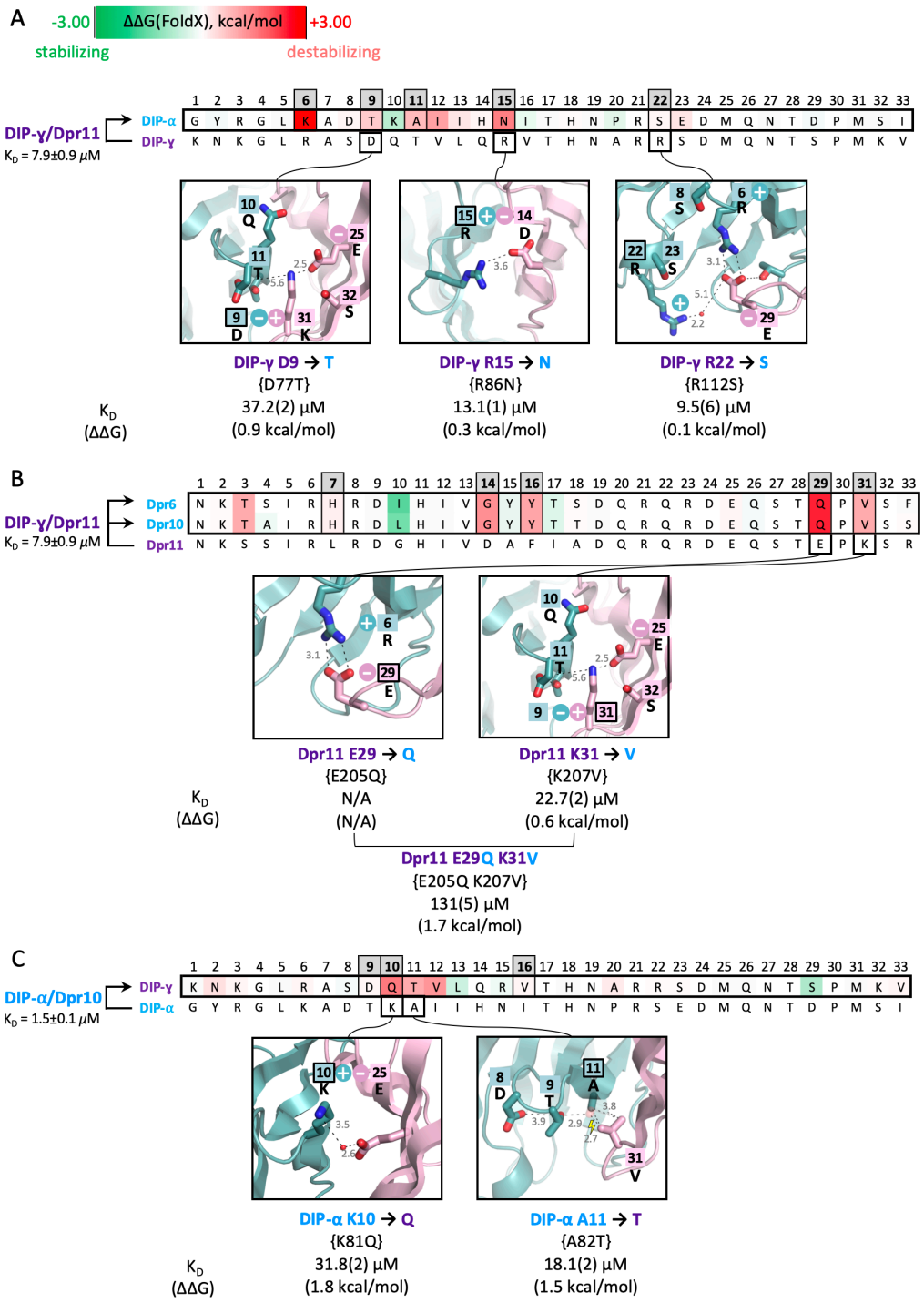

Figure 4. Negative constraints on A) DIP-α, (B), blue Dprs 6/10, (C), DIP-γ that inhibit purple/blue inter-subgroup binding. A) Alignment of interfacial residues of DIP-α and DIP-γ followed by structural details used to explain negative constraint. The average K_D for wild-type DIP-γ/Dpr11 interaction based on five independent experimental measurements is given on the left of the alignment. Experimental K_Ds and ΔΔGs relative to wild type DIP-α/Dpr11 are shown below each insert. All symbols as in Figure 3. B) Alignment of interfacial residues of Dpr6 and Dpr10 with Dpr11. All other details as in 4A. C) Alignment of interfacial residues of DIP-α and DIP-γ. All other details as in 4A. The supporting SPR data for the mutants presented in A, B, and C can be found in Figure S3A, S3B and S3C. For each K_D, the number in parenthesis represents the fitting error in the last significant figure, in μM.

320 *d) Negative constraints on the purple subgroup DIP-γ that weaken binding to blue subgroup*
321 *Dprs.* Three positions (9, 10, 16) in DIP-γ serve as negative constraints based on energy and
322 evolutionary filters (Figure 4C). Position 11 would also be predicted if two of the three template
323 complexes were used in energy filter (Figure S3D). We mutated two of these positions, K10Q and
324 A11T, and tested the extent to which they weaken binding of DIP-α to Dpr10. As seen in Figure
325 S3C, mutations at both of these positions increased the K_D of binding by about an order of
326 magnitude, thus confirming the prediction. Again, the physical basis of the negative constraint is
327 evident from the structure: K10 forms a water-mediated salt-bridge with sequence invariant Glu at
328 position 25 of light blue Dprs that would be disrupted with a neutral Gln; The A to T mutation at
329 position 11 of DIP-γ would introduce a steric clash with Val31 of light blue Dprs.
330
331 <u>DIP/Dpr family overview</u>
332 Figure 5 plots negative constraints identified by FoldX on the surface of hypothetical interfaces
333 that would be formed by a number of non-interacting subgroups. Most of the interface is used to create
334 negative constraints but different regions are exploited by different subgroups (Figure 5). Overall, our
335 predictions implicate 21 positions on the Dpr side of an interface and 17 positions on the DIP side as
336 a source of negative constraints (see positions highlighted in grey, Figure 6). These positions are
337 buried to different degrees in the interface, are located both in loops and secondary structure elements
338 and have different physical-chemical properties (hydrophobic, polar or charged). The majority of the
339 negative constraints reside in loop regions and code for polar or charged residues (21 positions) with
340 10 positions burying more than 50Å^2 of accessible surface area upon complex formation. Each non-
341 cognate DIP/Dpr subfamily pair was found to be destabilized by a set of about 7 interfacial residues
342 involving both sides of the interface. The physical source of the negative constraint can be shape non-
343 complementarity in the hydrophobic core (Figure 1) (through steric clashes or cavity creation),
344 electrostatic effects (Coulombic repulsion or a desolvation penalty associated with burying unsatisfied
345 charges or polar residues), or both (see examples in Figure 5). Notably, our SPR results suggest that

one negative constraint is not sufficient to ablate binding but that at least three are required to assure that members of different subgroups do not form a stable complex.

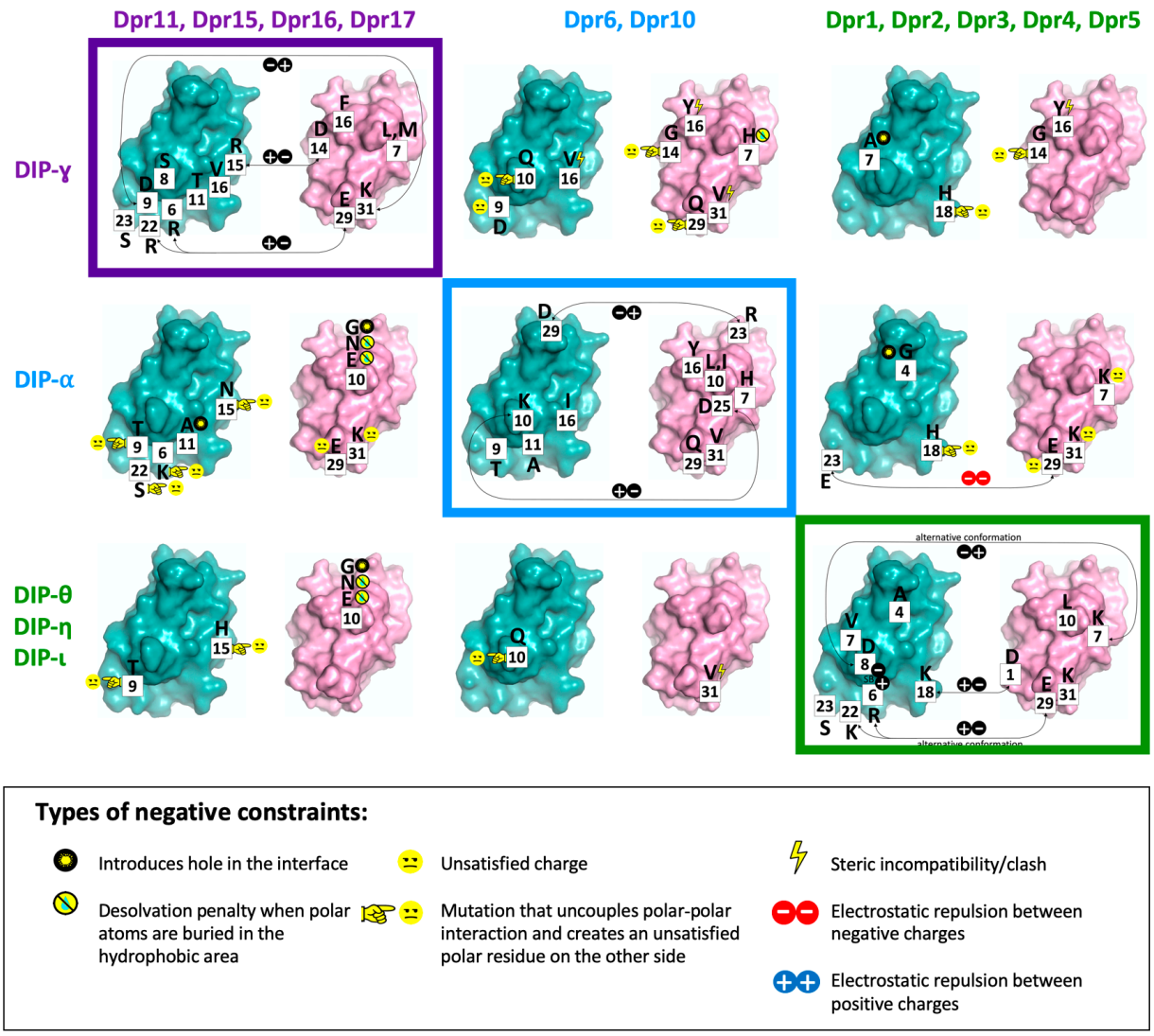

Figure 5. Negative constraints in DIP/Dpr non-cognate pairs in *Insecta*. Predicted negative constraints in three DIP/Dpr subfamilies are shown on the DIP (in cyan) and Dpr (in pink) interacting surfaces of purple, blue, and green subfamilies using an open book representation, with cognate interacting pairs shown enclosed in boxes coded with subfamily color. Family-wide interfacial position number are in filled white boxes. Amino acid residue identity at a negative constraint position is given in a one letter code. The physical origin of each negative constraint is shown with a pictogram as detailed in the enclosed rectangle at bottom.

Our structure-based predictions of negative constraints include all positions predicted by sequence-based methods, but identify about twice as many positions. Sequence-based methods often

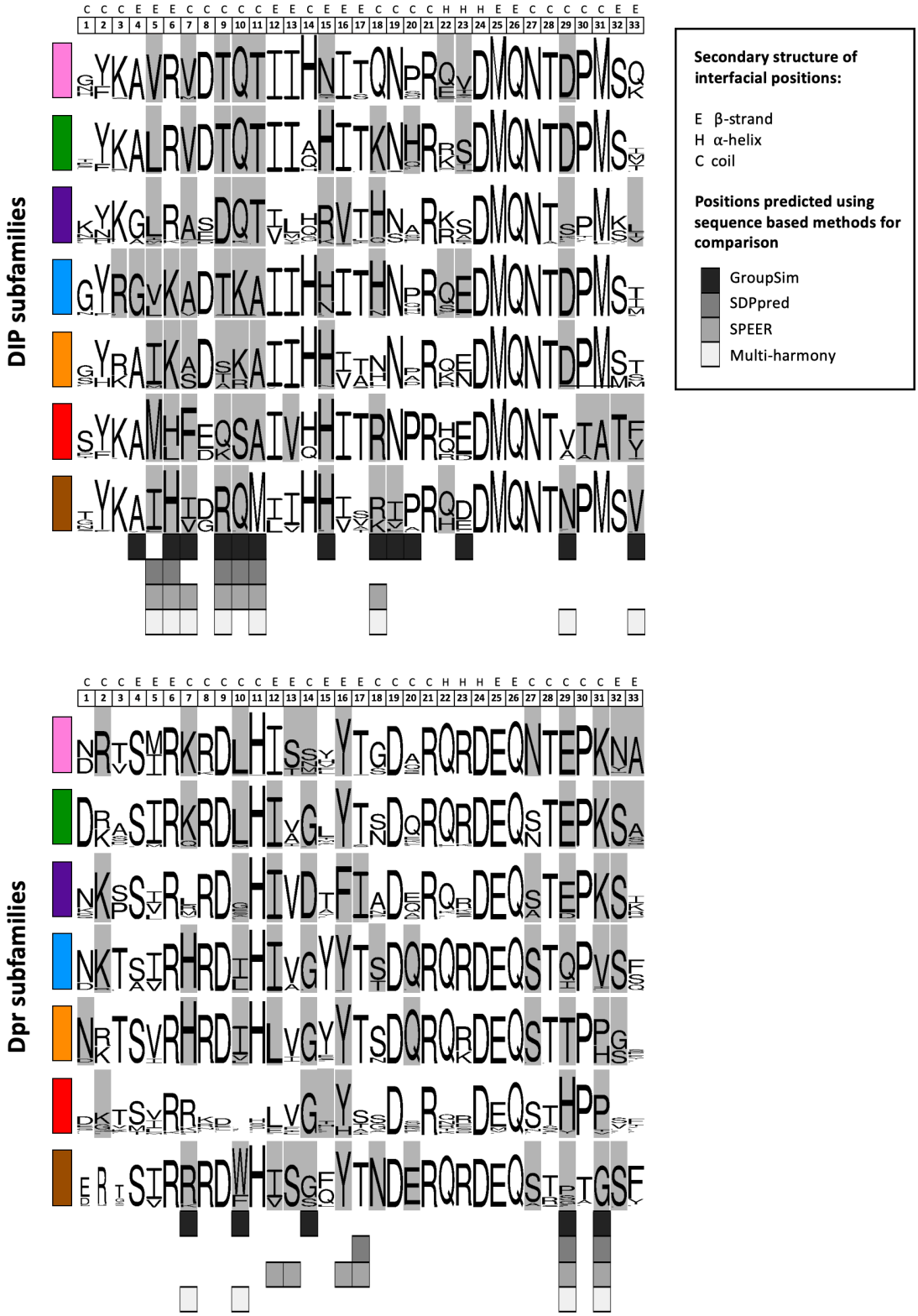

Figure 6. Predicted positions of negative constraints DIP/Dpr subgroups. Logos created for Insecta with residues that function as negative constraints highlighted in grey. Subfamily color-code as in Figure 1B. Each interfacial position (1-33) is annotated above logos with associated secondary structure, as detailed in the boxed legend. Sequence-based predictions of specificity determining positions using sequence-based methods (GroupSim, SDPpred, SPEER, and Multi-Harmony, same as in Figure 2A) are given below logos as specified in boxed inset.

364
365 fail to identity positions not fully conserved within a subfamily[22,36]. We also note that the majority of
366 sequence-based methods encounter difficulties in identifying sites that are specific to only one
367 subfamily. For example, positions 14 and 16 correspond to conserved Asp and Phe residues in the
368 purple Dpr subfamily, while other Dprs have Gly and Tyr in these positions, respectively; see Figure
369 2). Similarly, we predict several residues in the hydrophobic core as potential negative constraints
370 (12, 13, 16 in Dprs and 5, 13, 16 in DIPs), whereas sequence-based methods predict only one of these
371 positions (Figure 2). Of note, these positions were not among those we had experimentally
372 confirmed in this study.
373
374 Balancing affinity and specificity
375 In some interfacial positions, FoldX predicts that mutations to residues found in other
376 subgroups would actually increase affinity (Figures 3 and 4 and Supplemental FoldX data). Some of
377 these suboptimal positions in cognate DIP/Dpr pairs appear to act as negative constraints. As an
378 example, position 10 in purple Dprs, which is either a Gly or a polar residue, is predicted by FoldX to
379 disrupt nonpolar interactions with DIPs in other subgroups. SPR experiments confirmed this
380 prediction (Figure S4); L10G in Dpr10 decreased binding to DIP-α and DIP-β ($\Delta\Delta G$ >1.6 kcal/mol),
381 Notably, G10L increased binding between Dpr11 and DIP-γ by about the same amount, indicating
382 that some cognate binding affinity has been "sacrificed" so as to weaken non-cognate binding.
383 In a second example, position 7 is predicted to be used by the green Dpr subfamily to
384 weaken non-cognate binding to DIP-α by placing an unsatisfied charge, K7, in non-cognate
385 interfaces (Figure 5). Confirming this prediction, the H7K {H110K} mutation in Dpr6 weakens
386 cognate Dpr6/DIP-α binding by >1.9 kcal/mol[11]. In contrast, the K7H {K82H} mutation in Dpr4
387 strengthens cognate binding to DIP-η by 0.3 kcal/mol[11], a small effect but within the error bars of our
388 measurements. This is another case where cognate binding is designed to be suboptimal.
389

390 **Discussion**

391 We have described the structural and energetic origins of the partition of DIPs and Dprs into
392 orthogonal specificity groups defined by SPR-derived binding affinity measurements. We previously
393 analyzed specificity determinants in type II cadherins[5], nectins[8] and DIPs and Dprs[11], primarily
394 through visual inspections of sequences guided by structural data. Here we have adopted a far more
395 extensive and quantitative approach as required by the complexity of the problem we set out to
396 address. Specifically, how specificity is coded on seven subgroups of DIPs and Dprs so that out of
397 the 49 possible subgroup pairings of closely related subfamilies of DIPs and Dprs, only seven form
398 strong pairwise interactions. The problem is complicated by the fact that each subgroup contains
399 between 1 and 5 members on the Dpr side and 1 and 3 members on the DIP side. Asking how 42
400 combinations of DIP and Dpr subfamilies are designed not to interact, or in some cases to interact
401 weakly, poses a unique set of challenges. Notably, sequence-based methods were able to provide
402 only partial answers to this question.

403 The approach we have adopted is essentially to build a large set of homology models of
404 complexes that form and do not form and to ask, using FoldX calculations, SPR measurements and
405 visual inspection what is wrong with those that do not form. Since the calculations for each mutation
406 have to be repeated many times so as to obtain proper averages, the findings reported in this paper
407 are based on over 30440 separate FoldX calculations. These led to the identification of 38 negative
408 constraints and to 25 validation tests carried out with SPR. While it is possible of course that some of
409 our predictions are mistaken, the fact that each can be rationalized following visual inspection, that
410 each predicted negative constraint has to be common to all members of a subfamily and be present
411 in orthologous species adds confidence to the validity of our results.

412 The negative constraints we identified for the most part exploit strong localized unfavorable
413 energetic contributions rather, for example, than more subtle effects that are distributed over many
414 residues (as is the case for type I cadherins where entropic effects involving the movement of entire
415 domains play a role in determining dimerization affinities)[7]. The physical origins of negative

416 constraints include: replacing a charge in an ion pair with one of opposite sign leading to Coulombic
417 repulsion; replacing one member of an ion pair with a neutral group, which has the effect of burying
418 an unsatisfied charge; mutations in the hydrophobic core to larger amino acids that create steric
419 clashes or to smaller ones that create cavities which lead to packing defects and to weaker
420 hydrophobic contributions to binding. Of course, the same "trick" cannot be reused at the same
421 location for 42 different subgroup pairs. In fact, every set involving negative constraints on DIP and
422 Dpr side of the 42 non-cognate pairs is unique (see supplementary excel file). This is why as
423 discussed above, 38 of 66 interfacial residues, spread over the entire interface needs to be
424 exploited. Notably, in the cases we tested,[11] at least three negative constraints had to be added or
425 removed to essentially switch the specificity of a particular protein from one subgroup to another. As
426 can be seen in Figure 1B, intra-subgroup K_Ds are generally in the range of 10 μM (although some
427 are as weak as 40 μM) while inter-subgroup K_Ds (with a few exceptions) are undetectable, implying
428 a K_D > 500 μM. This suggests that the generation of a new DIP/Dpr specificity requires a change in
429 binding free energy of ~2.5 kcal/mol, a value that is difficult (but not impossible) to reach with one or
430 two mutations.

431 How is the DIP/Dpr interface designed to achieve affinity and specificity? First, we note that,
432 in common with most cell-cell recognition proteins, DIP/Dpr affinities are in the μM range, a
433 thousand-fold weaker, for example, than many antibody-antigen complexes with substantially
434 smaller binding interfaces. That lower affinities are a general feature of adhesion receptors, even
435 those which are not members of large families, suggest that there are other factors involved that
436 have little to do with negative constraints, for example cellular motility. Moreover, cell-cell avidity
437 involves the interactions of multiple receptors so that there may not be a need to evolve proteins with
438 nanomolar affinities. This may account for the observation above that even cognate interfaces are
439 not optimized for affinity. Moreover, a trade-off between affinity and specificity is to be expected
440 in large protein families with only limited sequence divergence. If an interaction between two
441 family members, say 'a' and 'b', is very strong, then it is more of a challenge, for example, to

442 generate an isoform ‘c’ that is similar in sequence to ‘b’, but interacts very weakly with ‘a’. Thus,
443 the higher the specificity requirements, the greater the constraints on absolute affinities.
444 We find that more than a half of the DIP/Dpr interface (38 of 66 interfacial residues) is utilized
445 for negative constraints across different insect species. However, these only play a role in the
446 context of non-canonical subgroup-subgroup interactions, whereas in any given cognate DIP/Dpr
447 complex, a majority of the interfacial residues likely play a stabilizing role. One can imagine that an
448 ancestral DIP/Dpr complex relied entirely on stabilizing interactions in the hydrophobic core along
449 with additional stabilizing or energy neutral interactions in the periphery. As multiple family members
450 evolved the need for specificity was satisfied through the introduction of negative constraints
451 throughout the interface. This evolutionary strategy is likely used by other large protein families.
452 In addition to manifesting themselves as destabilizing interfacial positions via non-synonymous
453 mutations (as in the case of DIP/Dpr family), negative constraints could potentially be introduced via
454 loop insertions or deletions whose purpose is to avoid undesired interaction, as it was shown in
455 some proteins[37].
456 Detailed binding affinity measurements have now been reported for a number of families of
457 adhesion receptors. In some cases, for example type II cadherins,[5] the family can be divided into
458 subgroups whose members exhibit considerable promiscuity in their intra-subgroup binding
459 properties while not binding to members of other subgroups. This behavior is reminiscent of the
460 subgroup characteristics of DIPs and Dprs although since there are fewer cadherin specificity groups
461 their evolutionary design is less of a challenge. In both cases binding promiscuity may well be a
462 source of functional redundancy but this has not been established. In contrast, Dscams[1,2] and
463 clustered protocadherins[3,4] are strictly homophilic which, as mentioned above, is an absolute
464 requirement for their role of establishing a unique identity for every neuron. In contrast to DIPs and
465 Dprs, the exquisite specificity of these protein families is established through multi-domain interfaces
466 that allow for a greater degree of fine-tuning. Negative constraints are obviously a central component
467 of their evolutionary design but these have not been characterized in detail yet.

468 The range of DIP/Dpr binding affinities indicated in Figure 1B raises an additional set of
469 interesting questions regarding the biological role of weak interactions. In a number of cases there is
470 a clearly established connection between molecular affinities and cellular phenotype[8,38] but in general
471 very little is known about this topic. The situation becomes more complex for cells that express more
472 than one family member and/or more than one subgroup member. The ability to design mutants of
473 altered affinities and specificities combined with modern gene editing techniques opens the door to a
474 new type of experiment which probes the molecular basis of cell-cell recognition. We have reported
475 preliminary studies on this problem involving DIP and Dpr mutants, which abolished homophilic or
476 both homophilic and heterophilic interactions in the *Drosophila* retina[11,35]. The next logical step will be
477 to test how subtle changes in affinity and specificity would affect biological phenotype.
478

479 ## Methods

480 **Protein structure alignment**

481 The DIP/Dpr structures used in alignment presented in Figure 1C were, *blue subgroup* - crystal
482 structures of Dpr6/DIP-α (PDBID: 5EO9) and two conformations of Dpr10/DIP-α (PDBID: 6NRQ)[9,10];
483 *purple subgroup* – crystal structure of Dpr11/DIP-ɣ (PDBID: 6NRR)[10]; *green subgroup* – crystal
484 structures of Dpr4/DIP-η (PDBID: 6EG0) and for two conformations each for Dpr1/DIP-η (PDBID:
485 6NRW) and Dpr2/DIP-θ (PDBID: 6EG1)[10,11].
486

487 **Construction of homology models for all superfamily members**

488 We built homology models for each DIP and Dpr Ig1 domain (using the crystal structure of the DIP-
489 η/Dpr4 complex as a template, PDBID: 6EG0). The models are expected in principle to be high
490 quality since there is considerable sequence and structural similarity between intra-family DIP and
491 Dpr Ig1 domains. A few modeled proteins have longer loops than those of the template but these
492 regions do not appear critical for Ig1'-Ig1" interface. Homology models of DIP and Dpr monomers

493 built using MODELLER[39] were superimposed onto Dpr4/DIP-η crystal structures yielding 231
494 DIP/Dpr heterodimers. The side chains of the dimer models were further minimized using the Scwrl4
495 algorithm[40].
496
497 **FoldX protocol to evaluate ΔΔG(binding)**
498 Although FoldX can be used as a "black box" we summarize here the procedures used to
499 calculate ΔΔG(binding). We first refine the structure of a protein complex using the FoldX utility
500 called "RepairPDB" (http://foldxsuite.crg.eu/command/RepairPDB). This procedure "repairs"
501 improbable dihedral angles and van der Waals clashes in a protein. We then run five rounds of
502 "RepairPDB" and confirm that total energy of a protein complex reaches a plateau. The complex
503 whose structure has now been optimized with respect to the FoldX energy function is then used as a
504 starting point for a computational mutagenesis procedure with "BuildModel"
505 (http://foldxsuite.crg.eu/command/BuildModel). A desired mutation is made to the repaired structure
506 and then energy-minimized by sampling rotamers for the mutated residue and its neighboring amino
507 acids. The wild type protein is then energy-minimized with respect to rotamers of the same set of
508 residues. The difference in interaction energy between the mutant and wild type is then calculated
509 between the two protomers in the complex using the 'AnalyseComplex' FoldX utility
510 (http://foldxsuite.crg.eu/command/AnalyseComplex).
511 To test if FoldX achieves convergence for a given mutation we run "BuildModel" ten times by
512 setting the "numberOfRuns" parameter to 10 (http://foldxsuite.crg.eu/parameter/numberOfRuns),
513 which produces ten pairs of wild type and mutant structures, as different neighbors of the mutated
514 residue could be moved during minimization. The average value of the interaction energy obtained
515 via energy decomposition of the total energy using 'AnalyseComplex' defines ΔΔG(binding).
516 Since we carried out 10 separate runs for each FoldX prediction, this yields 30,440 FoldX
517 single point mutation calculations (1 week of computing time on 1 CPU).
518

519 **Calculating the effects of mutations on protein-protein binding**

520 The following is a summary of the methods evaluated for the calculation of the effects of mutations
521 on binding free energies. ***FoldX*** evaluates the effects of mutations using an empirical force-field that
522 allows side chains to move but keeps the backbone rigid[18,31]; ***mCSM*** is a machine learning method
523 based on the assumption that the impact of a mutation is correlated with atomic-distance patterns in
524 surrounding amino acids[30]; ***BeAtMusic*** relies on a set of coarse-grained statistical potentials derived
525 from known protein structures[27]. BeAtMusic is the fastest of all the methods we tested – it takes
526 seconds to assess all possible mutations in a protein chain or at the interface; ***MutaBind*** utilizes a
527 combination of molecular mechanics force fields, statistical potentials and a fast side-chain
528 optimization algorithm; ***Rosetta flex ddG***[24] samples conformational diversity using "backrub" to
529 generate an ensemble of models and then applies torsional minimization, side-chain repacking, and
530 averaging across this ensemble to calculate $\Delta\Delta G$; ***BindProfX***[32] combines a sequence profile
531 conservation score of structural homologues with the FoldX potential. Data on the performance of
532 these methods is summarized in Table S2.

533

534

535 **Sequence-logo preparation**

536 To get orthologs of DIPs and Dprs in Insecta species we performed the search using reciprocal best
537 hit BLAST[41] with an e-value cutoff of 10^{-35} using in-house developed scripts on non-redundant
538 BLAST database. The resulting 2950 Dpr sequences and 2005 DIP sequences (121 species) were
539 screened using CD-HIT[42] with 100% threshold to get rid of redundancy arising from multiple GI
540 (GenInfo Identifier) numbers pointing to the same protein product. The obtained non-redundant set
541 of full-length sequences corresponding to DIP(*i*) or Dpr(*j*) was combined with a reference sequence
542 (DIP-α or Dpr6 respectively), and aligned using Clustal-Omega[43]. The multiple sequence alignments
543 (MSA) of each DIP(i) and Dpr(j) were further processed to remove sequences that had non-complete
544 Ig1 domains and to extract interfacial positions based on the reference sequences. Sequence

545 logos[44] based on the MSAs presented in Figure 2 were created separately for each *Drosophila* DIP
546 and Dpr based on 2570 Dpr and 1732 DIP sequences. Sequence logos were also prepared on a
547 subfamily level by combining sequences of all its members, see Figure 6.
548
549 **Sequence-based methods used to predict specificity residues**
550 GroupSim[20] analyzes evolutionary conservation patterns using physico-chemical properties of amino
551 acids by means of similarity matrices; SDPpred[21] uses information theory schemes based on
552 Shannon entropy and frequency scores; SPEER[22] uses amino acid properties, entropy and
553 evolutionary rates by analyzing quantitative measures of the conservation patterns of protein sites
554 based on their physico-chemical properties and the heterogeneity of evolutionary changes between
555 and within the protein subfamilies. Multi-Harmony[23] uses a combination of Shannon entropy and
556 machine learning.
557 All the sequence-based methods we use for determination of specificity positions require
558 predefined subgrouping of sequences based on their binding preferences. We grouped the above
559 2570 Dpr and 1732 DIP sequences (truncated to 33 interfacial residues only) into seven sets (DIP-
560 θ/η/ι/, DIP-κ, DIP-α, DIP-β/λ, DIP-γ, DIP-δ, DIP-ε,ζ for DIPs, and Dpr1/2/3/4/5, Dpr7, Dpr6/10,
561 Dpr8/9/21, Dpr11/15/16/17, Dpr12, Dpr13,14,18,19,20 for Dprs).
562 We rely on the assumption that orthologs in the same subfamily conserve their interaction
563 specificity. This assumption would not work for comparison of proteins in species that are too far
564 away from each other on the tree of life, but since we consider orthologs within one class, *Insecta*,
565 the assumption is likely to be correct. To justify this, we calculated the pairwise sequence identities
566 within *Insecta* orthologs for each DIP and Dpr, and the intra-subgroup pairwise sequence identity for
567 each DIP and Dpr *Insecta* subgroup (~78%) using multiple sequence alignments of interfacial
568 residues. These values are comparable to intra-subgroup pairwise sequence identities of interfacial
569 residues in *Drosophila Melanogaster* (~80%) and are consistently higher that inter-subgroup
570 pairwise sequence identities in *Drosophila Melanogaster* or *Insecta* (~50%). Only one subfamily (red

571 Dpr) is an exception to the rule - it has relatively low conservation in *Insecta* (48%), but it is as
572 diverse within the group in *Drosophila Melanogaster* (40%).

573 **Protocol for identifying negative constraints.**

574 *Energy filter*

575 Positions were filtered based on FoldX calculations of *Drosophila Melanogaster* DIP/Dpr complexes.
576 If positions were predicted to cause destabilizing effects ($\Delta\Delta G$ >=0.05 kcal/mol) for every DIP or Dpr
577 fly subgroup member in the context of binding to a member of another Dpr or DIP subgroup, the
578 position were accepted provided $\Delta\Delta G$ >=0.05 kcal/mol for every subfamily template complex used in
579 the FoldX calculations. The cutoff value for $\Delta\Delta G$ was chosen based on estimated standard deviation
580 value for mutating an interfacial amino acid into self ($\Delta\Delta G = 0.00 \pm 0.05$ kcal/mol). Positions that
581 passed energy filter were further subject to the evolutionary filter.

582 *Evolutionary filter*

583 Positions were filtered using multiple sequence alignments (MSAs) for DIP and Dpr subfamilies
584 based on sequences of insect orthologs (see above how these were obtained). For every non-
585 cognate DIP(i)/Dpr(j) family, to filter positions on DIP side we compared corresponding positions
586 from DIP(i) and DIP(j) subfamilies in MSAs. To filter positions on Dpr side we compared positions
587 from Dpr(i) and Dpr(j) subfamilies in MSAs. The position was accepted if at least one of the three
588 conditions outlined below was true.

589 1) If positions in "i" and "j" subgroup were both conserved (S<0.23) and different in identity – the
590 position was accepted (if purifying selection imposed evolutionary pressure to preserve a
591 certain type of an amino acid, even differences in shape/size of similar amino acids (L and I,
592 or R and K, might be of significance). Here S stands for Shannon entropy as the measure of
593 conservation. It's calculated using the following formula $S = -\sum_{i=1}^{N} f_i log_{21} f_i$ where f_i is
594 the fraction (frequency of occurrence) of residues of amino acid type present at a position in

595 MSA and N is the number of amino acid types (base of 21 was picked to account for 20
596 amino acids types and a gap).
597 2) If positions in "i" and "j" subgroup had difference in biophysical property (to account for
598 negative constraints associated with either desolvation penalty or electrostatic repulsion). We
599 binned 20 amino acids into four property groups: hydrophobic (I, V, L, M, F, W, A, C, G, P),
600 polar (S,T,Q,N,H,Y), positively charged (K,R), and negatively charged (E,D). Then we
601 summed up the instances where amino acids at a position of subfamily "i" (aa_i) were different
602 in biophysical property from aa_j weighted by the occurrences of aa_i and aa_j in MSAs
603 ($\sum_{i=1}^{N} f_{aa_i} \sum_{j=1}^{M} f_{aa_j} \times 100\%$). The positions were accepted if there was >90% difference in
604 biophysical property between two positions.
605 3) If positions in "i" and "j" subgroup had differences in size (amino acids of one subgroup
606 always bigger, or always smaller in size than those appearing in the other subgroup to justify
607 negative constraint involving steric clashes or introducing holes in the interfaces,
608 respectively). We binned 20 amino acids into five size groups, which were based on the
609 volume of amino acids: tiny (G, A, S), small (C, D, P, N, T), medium (E, V, Q, H), large (M, I,
610 L, K, R), bulky (F, Y, W). The percent difference was weighted by amino acid occurrences (in
611 the same manner as above) and the positions were accepted at >90% cutoff.

612 **Plasmid construction and protein expression**

613 All proteins were produced by expression in human embryonic kidney cells (HEK293F). DNA
614 sequences encoding DIP and Dpr extracellular regions were amplified by PCR and sub-cloned into
615 the mammalian expression vector VRC-8400[45] between the NotI and BamHI sites. Sequence
616 boundaries for each DIP and Dpr were the same as those defined in Cosmanescu et al.[11] All
617 sequences were preceded by the signal sequence of human binding immunoglobulin protein BiP
618 (MKLSLVAAMLLLLSAARA), and a kozak sequence (GCCACC). Constructs were followed by a C-
619 terminal hexa-histidine tag. Point mutations were introduced using the QuickChange method
620 (Agilent). HEK293F cells were transiently transfected with each expression construct using the

621 polyethylenimine method[46], and cells were expanded and grown in shake flasks in a CO2 incubator
622 for 3-6 days.
623
624 **Protein purification**
625 Conditioned media was equilibrated to 10mM Tris-HCl pH 8.0, 500mM NaCl, 3mM $CaCl_2$ and 5mM
626 Imidazole pH 8.0 and incubated with Ni^{2+} charged IMAC Sepharose 6 Fast Flow resin (GE
627 Healthcare) for 1 hr at 25°C. Resin was washed with at least 20 column volumes of buffer containing
628 10mM Imidazole pH 8.0. SDS gel electrophoresis was used to detect which elution fractions
629 contained the desired protein.
630 Proteins were further purified by size-exclusion chromatography (Superdex 200 HiLoad 26/60 or
631 Superdex S200 Increase 10/300 GL; GE Healthcare) on an AKTA pure fast protein liquid
632 chromatography system (GE Healthcare). Most proteins were stored in a buffer of 10mM Bis-Tris pH
633 6.6 and 150mM NaCl. UV absorbance at 280nm was used to determine protein concentration and
634 verification of purity was determined by gel electrophoresis. Accurate molecular weights were
635 determined through MALDI-TOF mass spectrometry at the Proteomics Shared Resource facility at
636 Columbia University.
637
638 **Sedimentation equilibrium by analytical ultracentrifugation**
639 Experiments were performed in a Beckman XL-A/I analytical ultracentrifuge (Beckman-Coulter, Palo
640 Alto CA, USA), utilizing six-cell centerpieces with straight walls, 12 mm path length and sapphire
641 windows. Protein samples were dialyzed to 10mM Bis-Tris pH 6.6, 150mM NaCl. The samples were
642 diluted to an absorbance of 0.65, 0.43 and 0.23 at a 10 mm path length and 280 nm wavelength in
643 channels A, B and C, respectively. Dilution buffer were used as blank. The samples were run at four
644 speeds. Most proteins were run at 15000, 19000, 23000, and 27000 RPM. For all runs the lowest
645 speed was held for 20hr and then four scans were taken with a 1hr interval, the second lowest held
646 for 10hr then four scans with a 1hr interval, and the third lowest and highest speed measured as the

647 second lowest speed. Measurements were done at 25°C, and detection was by UV at 280 nm.
648 Solvent density and protein v-bar were determined using the program SednTerp. (Alliance Protein
649 Laboratories) To calculate the K_D and apparent molecular weight, data were fit to a global fit model,
650 using HeteroAnalysis software package, obtained from University of Connecticut[47]
651 (http://www.biotech.uconn.edu/auf).
652
653 **Surface Plasmon Resonance (SPR) binding experiments**
654 SPR binding assays were performed using a Biacore T100 biosensor equipped with a Series S CM4
655 sensor chip. To minimize artificial binding resulting from enhanced- avidity effects of oligomers
656 interacting with an immobilized ligand, DIPs and their respective mutants, were consistently used as
657 ligands rather than analytes, and immobilized over independent flow cells using amine-coupling
658 chemistry in HBS pH 7.4 (10mM HEPES, 150mM NaCl) buffer at 25°C using a flow rate of 20
659 µL/min. Dextran surfaces were activated for 7 minutes using equal volumes of 0.1M NHS(N-
660 Hydroxysuccinimide) and 0.4M EDC(1-Ethyl-3-(3-dimethylaminopropyl)carbodiimide). Each protein
661 of interest was immobilized at ~30µg/mL in 10 mM sodium acetate, pH 5.5 until the desired
662 immobilization level was achieved. The immobilized surface was blocked using a 4-minute injection
663 of 1.0 M ethanolamine, pH 8.5. Typical immobilization levels ranged between 760-980 RU. To
664 minimize nonspecific binding the reference flow cell was blocked by immobilizing BSA in 10 mM
665 sodium acetate, pH 4.25 for 3 minutes using a similar amine-coupling protocol as described above.
666 For each experiment where DIP mutants were immobilized to the chip surface, the wild type
667 molecule was immobilized on an adjacent flow cell as a positive control.
668 Binding analysis was performed at 25°C in a running buffer of 10 mM Tris-HCl, pH 7.2, 150mM NaCl,
669 1mM EDTA, 1 mg/mL BSA and 0.01% (v/v) Tween-20. Analytes were prepared in running buffer and
670 tested at nine concentrations using a three-fold dilution series ranging from 81 to 0.012 µM for data
671 shown in Figure 3C and D, Figure S3A-C, Figure S4, with the of exception Dpr10 binding to DIP-α
672 A11T in Figure S3C, which was tested at nine concentration ranging from 27 to 0.004 µM, prepared

673 in a three-fold dilution series. Dpr6 was tested at eight concentrations ranging from 9 to 0.004 μM,
674 prepared in a running buffer using a three- fold dilution series for data presented in Figure S1A, with
675 the exception of experiments over the DIP-α I16A and DIP-α K10Q G4S surfaces where Dpr6 was
676 used at ten concentrations ranging from 81-0.004 μM, using a three-fold dilution series. Dpr10 binding
677 experiments in Figure S1B, was tested at nine concentrations ranging from 27 to 0.004 μM, prepared
678 in a running buffer using a three- fold dilution series, with the exception of experiments over surfaces
679 immobilized with DIP-α I16A, K10Q G4S and K10Q D29S where the analyte was used at ten
680 concentrations ranging from 81-0.004 μM, using a three-fold dilution series. In each experiment, every
681 concentration was tested in duplicate. Within each experiment, there two technical replicates starting
682 with a single concentration series, where samples are tested in order of increasing concentration,
683 followed by a repeat of the same concentration series, performed again from low to high concentration.
684 During a binding cycle, the association phase between each analyte and the immobilized molecule
685 was monitored for either 30 or 40 seconds as indicated by the plotted sensorgrams, followed by 120-
686 second dissociation phase, each at 50 μL/min. At the end of the dissociation phase the signal returned
687 to baseline thus eliminating the need for a regeneration step. The last step was buffer wash injection
688 at 100 μL/min for 60 seconds. The analyte was replaced by buffer every two or three binding cycles
689 to double-reference the binding signals by removing systematic noise and instrument drift. The
690 duplicate binding responses were fit globally, using an 1:1 interaction model and a single K_D was
691 calculated as the analyte concentration that would yield 0.5 R_{max}[48] and a fitting error. In cases where
692 the highest analyte concentration is only around the K_D, we use the R_{max} value for a saturating binding
693 isotherm shown on the same plot, as a global R_{max}, to fit the binding isotherms for weaker interactions
694 and calculate a K_D. The data was processed using Scrubber 2.0 (BioLogic Software).
695 For several wild type interactions discussed in this manuscript, Dpr10/DIP-α, Dpr6/DIP-α, and
696 Dpr11/DIP-γ, we have also determined K_Ds from independent experiments. For the Dpr10/DIP-α
697 binding pair, the K_D from six independent experiments is 1.5±0.1 μM, and similarly for the Dpr6/DIP-
698 α, the K_D from six independent experiments is 2.0±0.2 μM. The K_D from five independent

699 experiments for Dpr11/DIP-γ is 7.9±0.9μM. The standard deviation in ΔΔG values computed based
700 on K_D measurements of the above three cases do not exceed 0.1 kcal/mol. Therefore, mutations
701 resulting in ΔΔG (SPR) > | 0.2 kcal/mol | are likely to be significant.
702 In a note added in proof, Cheng et al.[10] recently criticized K_Ds we previously determined via SPR
703 experiments[11], specifically for the interacting pairs Dpr6/DIP-α, Dpr11/DIP-γ and Dpr1/DIP-η, for
704 being consistently weaker than K_Ds they reported. The following discussion explains how differences
705 in experimental methodology likely explains these differences and raises issues about the approach
706 used by Cheng et al.[10]. The underlying challenge in both sets of measurements is that some DIPs
707 homodimerize, a problem that is discussed by Rich and Myszka in their 2006 review of SPR data[49].
708 We addressed this problem by immobilizing DIP proteins to the chip surface, using Dprs as analytes,
709 to minimize artificially lower K_Ds resulting from homodimers used as analytes. It has been suggested
710 that an alternate "fix" is to use the free monomer analyte concentration in K_D calculations[10].
711 However, the monomer/dimer equilibrium in the analyte can shift the moment the analyte is injected
712 over a surface immobilized with a heterophilic binding partner, which changes the monomeric
713 analyte concentration available for a heterophilic binding reaction over the course of the experiment.
714 Therefore, this correction does not adequately solve the problem.
715 Moreover, to calculate the free monomer analyte concentration of DIP-η, Cheng et al. used size
716 exclusion chromatography and SPR to calculate a homophilic K_D of 23 μM and 14 μM,
717 respectively,[10] compared to our analytical ultracentrifugation-determined K_D of 56.2 μM[11]. Given the
718 reliability of AUC measurements, this highlights the problems associated with Chang et al.'s
719 approach. Most importantly these numbers suggest that the differences in K_D reported in both
720 studies are due in part to the use of different biophysical techniques used to study these molecules.
721 In addition, for many of the results that differ between our studies and Cheng et al.[10], these authors
722 relied on ECIA experiments with artificially multimerized molecules. As previously discussed in detail
723 in Cosmanescu et al.[11] although ECIA can be used to detect interactions in a high-throughput
724 fashion and without the need for purified proteins, it involves the use of multimerized forms of both

725 the prey and the bait molecules, and thus introduces artificial avidity to amplify binding signals,
726 enhancing the likelihood of positive detection, but masking the real binding affinities between
727 interactants. In addition, the use of unpurified protein supernatants[10] could introduce experimental
728 bias toward identifying interactions between molecules that are more easily expressed at higher
729 levels compared to proteins that can be difficult to express.
730

731 **Acknowledgments**

732 This work was supported by US National Science Foundation grant MCB-1412472 (to B.H.) and NIH
733 grant R01 MH114817 (to L.S.). B.H. is Investigator of the Howard Hughes Medical Institute. We
734 thank Engin Ozkan for providing DIP-α/Dpr10 and DIP-γ/Dpr11 structures and density maps prior to
735 their appearance in the PDB database.
736

737 **Author Contributions**

738 A.P.S., L.S. and B.H. conceived the project, designed experiments, and analyzed data. A.P.S.
739 carried out all the theoretical and computational work. P.S.K. performed and analyzed SPR
740 experiments. G.A. performed AUC experiments. J.J.B, F.C., S.M. produced expression vectors and
741 proteins. A.P.S. and F.C. designed mutant constructs. A.P.S., B.H., and L.S. wrote the manuscript.
742

743 **Data and code availability**

744 The raw data that support the findings of this study, and the code used to generate it, are available
745 from the corresponding authors upon request.
746

747 **References**

748

749 1. Zipursky, S.L., Wojtowicz, W.M. & Hattori, D. Got diversity? Wiring the fly brain with
750 Dscam. *Trends Biochem Sci* **31**, 581-8 (2006).

751 2. Wojtowicz, W.M. et al. A vast repertoire of Dscam binding specificities arises from
752 modular interactions of variable Ig domains. *Cell* **130**, 1134-45 (2007).
753 3. Thu, C.A. et al. Single-cell identity generated by combinatorial homophilic interactions
754 between alpha, beta, and gamma protocadherins. *Cell* **158**, 1045-1059 (2014).
755 4. Schreiner, D. & Weiner, J.A. Combinatorial homophilic interaction between gamma-
756 protocadherin multimers greatly expands the molecular diversity of cell adhesion. *Proc*
757 *Natl Acad Sci U S A* **107**, 14893-8 (2010).
758 5. Brasch, J. et al. Homophilic and Heterophilic Interactions of Type II Cadherins Identify
759 Specificity Groups Underlying Cell-Adhesive Behavior. *Cell Rep* **23**, 1840-1852 (2018).
760 6. Katsamba, P. et al. Linking molecular affinity and cellular specificity in cadherin-
761 mediated adhesion. *Proc Natl Acad Sci U S A* **106**, 11594-9 (2009).
762 7. Vendome, J. et al. Structural and energetic determinants of adhesive binding specificity
763 in type I cadherins. *Proc Natl Acad Sci U S A* **111**, E4175-84 (2014).
764 8. Harrison, O.J. et al. Nectin ectodomain structures reveal a canonical adhesive interface.
765 *Nat Struct Mol Biol* **19**, 906-15 (2012).
766 9. Carrillo, R.A. et al. Control of Synaptic Connectivity by a Network of Drosophila IgSF Cell
767 Surface Proteins. *Cell* **163**, 1770-1782 (2015).
768 10. Cheng, S. et al. Molecular basis of synaptic specificity by immunoglobulin superfamily
769 receptors in Drosophila. *Elife* **8**(2019).
770 11. Cosmanescu, F. et al. Neuron-Subtype-Specific Expression, Interaction Affinities, and
771 Specificity Determinants of DIP/Dpr Cell Recognition Proteins. *Neuron* (2018).
772 12. Tan, L. et al. Ig Superfamily Ligand and Receptor Pairs Expressed in Synaptic Partners in
773 Drosophila. *Cell* **163**, 1756-69 (2015).
774 13. Ozkan, E. et al. An extracellular interactome of immunoglobulin and LRR proteins reveals
775 receptor-ligand networks. *Cell* **154**, 228-39 (2013).
776 14. Havranek, J.J. & Harbury, P.B. Automated design of specificity in molecular recognition.
777 *Nat Struct Biol* **10**, 45-52 (2003).
778 15. Humphris, E.L. & Kortemme, T. Design of multi-specificity in protein interfaces. *PLoS*
779 *Comput Biol* **3**, e164 (2007).
780 16. Leaver-Fay, A. et al. Computationally Designed Bispecific Antibodies using Negative
781 State Repertoires. *Structure* **24**, 641-651 (2016).
782 17. Mandell, D.J. & Kortemme, T. Computer-aided design of functional protein interactions.
783 *Nat Chem Biol* **5**, 797-807 (2009).
784 18. Schymkowitz, J. et al. The FoldX web server: an online force field. *Nucleic Acids Res* **33**,
785 W382-8 (2005).
786 19. Grigoryan, G., Reinke, A.W. & Keating, A.E. Design of protein-interaction specificity gives
787 selective bZIP-binding peptides. *Nature* **458**, 859-64 (2009).
788 20. Capra, J.A. & Singh, M. Characterization and prediction of residues determining protein
789 functional specificity. *Bioinformatics* **24**, 1473-80 (2008).
790 21. Kalinina, O.V., Mironov, A.A., Gelfand, M.S. & Rakhmaninova, A.B. Automated selection
791 of positions determining functional specificity of proteins by comparative analysis of
792 orthologous groups in protein families. *Protein Sci* **13**, 443-56 (2004).
793 22. Chakrabarti, S., Bryant, S.H. & Panchenko, A.R. Functional specificity lies within the
794 properties and evolutionary changes of amino acids. *J Mol Biol* **373**, 801-10 (2007).

795 23. Brandt, B.W., Feenstra, K.A. & Heringa, J. Multi-Harmony: detecting functional
796 specificity from sequence alignment. *Nucleic Acids Res* **38**, W35-40 (2010).
797 24. Barlow, K.A. et al. Flex ddG: Rosetta Ensemble-Based Estimation of Changes in Protein-
798 Protein Binding Affinity upon Mutation. *J Phys Chem B* **122**, 5389-5399 (2018).
799 25. Buss, O., Rudat, J. & Ochsenreither, K. FoldX as Protein Engineering Tool: Better Than
800 Random Based Approaches? *Comput Struct Biotechnol J* **16**, 25-33 (2018).
801 26. Choi, Y., Furlon, J.M., Amos, R.B., Griswold, K.E. & Bailey-Kellogg, C. DisruPPI: structure-
802 based computational redesign algorithm for protein binding disruption. *Bioinformatics*
803 **34**, i245-i253 (2018).
804 27. Dehouck, Y., Kwasigroch, J.M., Rooman, M. & Gilis, D. BeAtMuSiC: Prediction of changes
805 in protein-protein binding affinity on mutations. *Nucleic Acids Res* **41**, W333-9 (2013).
806 28. Geng, C., Vangone, A., Folkers, G.E., Xue, L.C. & Bonvin, A. iSEE: Interface structure,
807 evolution, and energy-based machine learning predictor of binding affinity changes
808 upon mutations. *Proteins* (2018).
809 29. Li, M., Simonetti, F.L., Goncearenco, A. & Panchenko, A.R. MutaBind estimates and
810 interprets the effects of sequence variants on protein-protein interactions. *Nucleic Acids*
811 *Res* **44**, W494-501 (2016).
812 30. Pires, D.E., Ascher, D.B. & Blundell, T.L. mCSM: predicting the effects of mutations in
813 proteins using graph-based signatures. *Bioinformatics* **30**, 335-42 (2014).
814 31. Guerois, R., Nielsen, J.E. & Serrano, L. Predicting changes in the stability of proteins and
815 protein complexes: a study of more than 1000 mutations. *J Mol Biol* **320**, 369-87 (2002).
816 32. Xiong, P., Zhang, C., Zheng, W. & Zhang, Y. BindProfX: Assessing Mutation-Induced
817 Binding Affinity Change by Protein Interface Profiles with Pseudo-Counts. *J Mol Biol* **429**,
818 426-434 (2017).
819 33. Ashley, J. et al. Transsynaptic interactions between IgSF proteins DIP-α and Dpr10 are
820 required for motor neuron targeting specificity in Drosophila. *bioRxiv* (2018).
821 34. Venkatasubramanian, L. et al. Stereotyped Terminal Axon Branching of Leg Motor
822 Neurons Mediated by IgSF Proteins DIP-α and Dpr10. *bioRxiv* (2018).
823 35. Xu, S. et al. Interactions between the Ig-Superfamily Proteins DIP-alpha and Dpr6/10
824 Regulate Assembly of Neural Circuits. *Neuron* (2018).
825 36. Chakraborty, A. & Chakrabarti, S. A survey on prediction of specificity-determining sites
826 in proteins. *Brief Bioinform* **16**, 71-88 (2015).
827 37. Akiva, E., Itzhaki, Z. & Margalit, H. Built-in loops allow versatility in domain-domain
828 interactions: lessons from self-interacting domains. *Proc Natl Acad Sci U S A* **105**, 13292-
829 7 (2008).
830 38. Togashi, H. et al. Nectins establish a checkerboard-like cellular pattern in the auditory
831 epithelium. *Science* **333**, 1144-7 (2011).
832 39. Eswar, N. et al. Comparative protein structure modeling using Modeller. *Curr Protoc*
833 *Bioinformatics* **Chapter 5**, Unit-5 6 (2006).
834 40. Krivov, G.G., Shapovalov, M.V. & Dunbrack, R.L., Jr. Improved prediction of protein side-
835 chain conformations with SCWRL4. *Proteins* **77**, 778-95 (2009).
836 41. Altschul, S.F., Gish, W., Miller, W., Myers, E.W. & Lipman, D.J. Basic local alignment
837 search tool. *J Mol Biol* **215**, 403-10 (1990).

838 42. Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-
839 generation sequencing data. *Bioinformatics* **28**, 3150-2 (2012).
840 43. Sievers, F. et al. Fast, scalable generation of high-quality protein multiple sequence
841 alignments using Clustal Omega. *Mol Syst Biol* **7**, 539 (2011).
842 44. Crooks, G.E., Hon, G., Chandonia, J.M. & Brenner, S.E. WebLogo: a sequence logo
843 generator. *Genome Res* **14**, 1188-90 (2004).
844 45. Barouch, D.H. et al. A human T-cell leukemia virus type 1 regulatory element enhances
845 the immunogenicity of human immunodeficiency virus type 1 DNA vaccines in mice and
846 nonhuman primates. *J Virol* **79**, 8828-34 (2005).
847 46. Baldi, L., Hacker, D.L., Meerschman, C. & Wurm, F.M. Large-scale transfection of
848 mammalian cells. *Methods Mol Biol* **801**, 13-26 (2012).
849 47. Cole, J.L., Lary, J.W., T, P.M. & Laue, T.M. Analytical ultracentrifugation: sedimentation
850 velocity and sedimentation equilibrium. *Methods Cell Biol* **84**, 143-79 (2008).
851 48. Rich, R.L., Myszka, D.G. Extracting affinity constants from biosensor binding responses.
852 in *Label-free biosensors : techniques and applications* (ed. Cooper, M.A.) pp. 48-84
853 (Cambridge; New York: Cambridge University Press, 2009).
854 49. Rich, R.L. & Myszka, D.G. Survey of the year 2006 commercial optical biosensor
855 literature. *J Mol Recognit* **20**, 300-66 (2007).
856