

Using large-scale mutagenesis to guide single amino acid scanning experiments

Vanessa E. Gray¹, Ronald J. Hause¹, Douglas M. Fowler^{1,2}

¹Department of Genome Sciences, University of Washington, Seattle, WA, USA

²Department of Bioengineering, University of Washington, Seattle, WA, USA

Correspondence to dfowler@uw.edu

1 **Abstract**

2 Alanine scanning mutagenesis is a widely-used method for identifying protein positions
3 that are important for function or ligand binding. Alanine was chosen because it is
4 physicochemically innocuous and constitutes a deletion of the side chain at the β -
5 carbon. Alanine is also thought to best represent the effects of other mutations;
6 however, this assumption has not been formally tested. To determine whether alanine
7 substitutions are always the best choice, we analyzed 34,373 mutations in fourteen
8 proteins whose effects were measured using large-scale mutagenesis approaches. We
9 found that several substitutions, including histidine and asparagine, are better at
10 recapitulating the effects of other substitutions. Histidine and asparagine also correlated
11 best with the effects of other substitutions in different structural contexts. Furthermore,
12 we found that alanine is among the worst substitutions for detecting ligand interface
13 positions, despite its frequent use for this purpose. Our work highlights the utility of
14 large-scale mutagenesis data and can help to guide future single substitution mutational
15 scans.

16 **Main text**

17

18 **Introduction**

19 Making and studying mutants is a fundamental way to learn about proteins, revealing
20 functionally important positions, validating specific hypotheses about catalytic
21 mechanism and yielding insights into protein folding and stability. Single amino acid
22 scanning mutagenesis, in which every position in a protein is sequentially mutated to
23 one particular amino acid, was a key advance. By searching sequence space
24 systematically, scanning mutagenesis enabled the unbiased identification of positions
25 important for protein function. The first application of scanning mutagenesis used
26 alanine substitutions to identify positions in human growth hormone important for
27 receptor binding¹. Alanine was chosen because it represents a deletion of the side
28 chain at the β -carbon, and because, being uncharged and of modest size, it is
29 physicochemically innocuous. Furthermore, alanine is the second most abundant amino
30 acid in natural sequences and is found in a variety of structural contexts²⁻⁴. In addition to
31 alanine, many other amino acids including arginine⁵, cysteine⁶, glycine⁷, methionine⁸,
32 phenylalanine⁹, proline¹⁰ and tryptophan¹¹ have been used for scanning mutagenesis,
33 often with a specific hypothesis in mind (e.g. that bulky amino acids are important).
34 Nevertheless, the vast majority of scanning mutagenesis experiments are conducted
35 using alanine under the assumption that alanine substitutions are especially useful for
36 identifying functionally important positions.
37 Does alanine best represent the effect of other substitutions? Are alanine substitutions
38 ideal for finding functionally important positions, such as those that participate in binding

39 interfaces? Answering these questions is important because alanine scans continue to
40 be used to understand and engineer proteins. Despite the large investment in alanine
41 scanning mutagenesis, little work has been done to determine which substitutions are
42 ideal. Some scanning mutagenesis studies compare two different types of scans (*e.g.*
43 alanine and cysteine), but generally find that the information revealed by each
44 substitution is distinct^{9,12}. Computational predictions for all substitutions at 1,073
45 positions across 48 proteins in the Alanine Scanning Energetics Database suggested
46 that alanine substitutions correlated best with the mean effect of every mutation at each
47 position¹³. However, concrete answers to these questions require comparing the
48 empirical effects of different substitutions in many proteins. Thus, we analyzed large-
49 scale experimental mutagenesis data sets comprising 34,373 mutations in fourteen
50 proteins. We found that proline is the most disruptive substitution and methionine is the
51 most tolerated. Global and position-centric analyses revealed that histidine and
52 asparagine substitutions best represent the effects of other substitutions. We evaluated
53 the utility of each amino acid substitution for determining whether a position is in a
54 ligand-binding interface, and found that aspartic acid, glutamic acid, asparagine and
55 glutamine performed best. Thus, our results suggest that, compared to other
56 substitutions, alanine substitutions are not especially representative, nor are they the
57 best choice for finding ligand-binding interfaces.

58

59 **Results**

60 Deep mutational scanning is a method that enables measurement of the effects of
61 hundreds of thousands of mutations in a protein simultaneously^{14,15}. Deep mutational

62 scanning can be used to quantify the effects of all mutations at each position in a
63 protein, and is therefore a conceptual extension of single amino acid scanning
64 mutagenesis. Broad application of deep mutational scanning has resulted in an
65 explosion of protein mutagenesis data¹⁴. These large-scale mutagenesis data sets
66 create the opportunity to assess relationships between the effects of different amino
67 acid substitutions comprehensively.

68
69 We curated sixteen large-scale mutagenesis data sets from published deep mutational
70 scans of fourteen proteins (**Fig. 1A, Table 1**). Here, we included two distinct data sets
71 for the BRCA1 RING domain and for UBI4 because mutations in these proteins have
72 been independently assayed for different protein functions (e.g. BRCA1 BARD1 binding
73 and E3 ligase activity). Our collection of data sets is ideal for an unbiased analysis of
74 the general effects of mutations because the mutagenized proteins are highly diverse,
75 encompassing enzymes, structural proteins and chaperones from organisms ranging
76 from bacteria to humans. The frequency of amino acids in the wild type sequences of
77 the fourteen proteins was similar to amino acid frequencies in all known proteins² (**Fig.**
78 **1B**). For example, leucine (frequency = 11%) and alanine (8%) were the most frequently
79 occurring wild type amino acids in the fourteen proteins, while tryptophan (<1%) was the
80 rarest. However, the unbiased and massively parallel nature of deep mutational
81 scanning experiments yielded a relatively uniform distribution of amino acid
82 substitutions (**Fig. 1C**). Furthermore, the data sets were generated by different labs at
83 different times using different types of assays, reducing the chances of bias arising from
84 specific experimental or analytical practices. Importantly, the assay formats used for the

85 deep mutational scans included many commonly employed in alanine scanning like
86 phage display and yeast two-hybrid. Collectively, these large-scale mutagenesis data
87 sets comprised 34,373 nonsynonymous mutations at 2,236 positions in the fourteen
88 proteins. The data sets contained effect scores for most mutations at each position. To
89 facilitate comparisons between each data set, we rescaled mutational effect scores for
90 each protein, using synonymous mutations to define wild type-like activity and the
91 bottom 1% of mutations to define lack of activity (**Fig. S1A**). Thus, each mutational
92 effect score reflects the impact of the mutation, relative to wild type, with a score of zero
93 meaning no activity and a score of one meaning wild type-like activity.

94
95 To validate the large-scale mutagenesis data, we examined expected patterns of
96 mutational effect. For example, mutations to proline should generally disrupt protein
97 function, as proline restricts the conformation of the polypeptide backbone and
98 eliminates the amide hydrogen necessary for hydrogen bonding. Indeed, proline
99 substitutions were overwhelmingly more damaging than other substitutions to protein
100 function (**Fig. 1D; Fig.S1B**). In fact, proline was the most damaging amino acid in
101 eleven of fourteen proteins and second most damaging in the remaining three proteins
102 (**Fig. 1E**). Additionally, as expected from the Dayhoff¹⁶, Blosum¹⁷ and Grantham¹⁸
103 substitution matrices, tryptophan tended to be deleterious. Methionine was the best-
104 tolerated substitution. Many other substitutions were also well-tolerated, with seven
105 different amino acids appearing as the most tolerated across the fourteen proteins (**Fig.**
106 **1D, E**). Tolerance to substitutions depends on structural context, so the variability in the
107 best tolerated substitution might be due to diversity in the structural composition of each

108 protein in our data set. Thus, the large-scale mutagenesis data sets we collected
109 generally recapitulated our expectations about the effects of mutations, despite coming
110 from fourteen distinct proteins that were each assayed independently.

111
112 Next, we determined which amino acid substitution best represented the effects of all
113 other substitutions. To avoid bias arising from incomplete data, we restricted this
114 analysis to the 882 positions in the fourteen proteins with measured effects for all
115 nineteen possible substitutions. We calculated the median mutational effect at each of
116 these 882 positions. Overall, the median effects across these positions were mildly
117 damaging, with a mean of 0.82 (stop ~ 0, wild-type ~ 1). We found that the effects of
118 phenylalanine, glycine, histidine, isoleucine, leucine, asparagine, glutamine and tyrosine
119 substitutions were all indistinguishable from the median effects (**Fig. 2A, Table S1**).
120 However, proline, aspartic acid and tryptophan substitutions are much more damaging
121 than the median substitution. Alanine, cysteine, methionine, serine, threonine and valine
122 are considerably less damaging than the median substitution. These well-tolerated
123 amino acid substitutions might be useful for detecting the most mutationally sensitive
124 positions in a protein. However, these substitutions, alanine included, are not especially
125 representative of the effects of other substitutions.

126
127 We also examined the dispersion of each amino acid's mutational effect about the
128 median at all 882 positions, reasoning that representative substitutions would have
129 minimal dispersion. Of substitutions whose effects were indistinguishable from the
130 median effect, histidine and asparagine have the smallest dispersion (standard

131 deviation = 0.15 and 0.14, respectively; **Fig. 2B**), while tyrosine (0.18), glutamine (0.16),
132 phenylalanine (0.19), glycine (0.17), leucine (0.17) and isoleucine (0.19) all had larger
133 dispersions. Thus, of all possible substitutions, histidine and asparagine tended to have
134 effects closest to the median effect at the 882 positions we examined.

135

136 Because of the comprehensive nature of the large-scale mutagenesis data sets, we
137 could ask how well the mutational effect scores of each substitution correlated with the
138 scores of every other substitution at each position. Thus, we calculated Pearson
139 correlation coefficients for the mutational effect scores of each substitution pair across
140 all positions (**Fig. 2C, Fig. S2**). The effects of histidine and asparagine substitutions
141 correlated best with the effects of all other substitutions, while the effect of proline
142 substitutions correlated worst. To visualize the relationships between each pair of
143 substitutions, we constructed a force-directed graph (**Fig. 2D**). As expected,
144 substitutions cluster by physicochemical type in the graph, meaning that similar
145 substitutions have similar effects. Proline is not represented because its effects are
146 poorly correlated with other substitutions. Histidine and asparagine are connected to
147 many other amino acids, owing to the high correlation of the effects of these
148 substitutions with many other substitutions.

149

150 We next asked whether the secondary structural context of a position altered the effect
151 of each substitution. We excluded DBR1 and GB1 from this analysis because they did
152 not have structures of a sufficiently close homologs. We used DSSP to identify 1,007
153 positions in the remaining proteins that were in an α -helix, a β -sheet or a turn¹⁹. Overall,

154 substitutions in turns are less damaging than substitutions in α -helices or β -sheets (**Fig.**
155 **3A**). However, the relative effects of each substitution in the three structural contexts
156 were mostly consistent, especially between α -helices and β -sheets (**Fig. 3B, S3A**).
157 Surprisingly, the tolerance for each amino acid substitution in the different secondary
158 structural contexts was not strongly correlated with the frequency of that amino acid's
159 occurrence in known structures²⁰. For example, alanine occurs more frequently in α -
160 helices, relative to β -sheets. However, in our large-scale mutagenesis data sets,
161 alanine substitutions were mildly damaging in both structural contexts. These
162 observations suggest that secondary structure does not dominate mutational tolerance,
163 at least for the proteins we examined.

164
165 We next investigated which substitutions were the most representative regardless of
166 structural context. We found that histidine substitutions have close to the median effect
167 in α -helices and turns, but were more damaging than the median effect in β -sheets (**Fig.**
168 **3B**). Asparagine and glutamine substitutions had near median effects in all three
169 contexts. As above, we examined how well the effects of each substitution correlated
170 with every other substitution at each position in each context. We found that the effects
171 of histidine, asparagine and glutamine substitutions correlated best with the effects of
172 other substitutions (**Fig. S3B, C**). Thus, the effects of histidine, asparagine and
173 glutamine are relatively consistent in the different structural contexts we examined,
174 highlighting the representativeness of these substitutions.

175

176 An important use of alanine scanning is to identify positions in protein-ligand interfaces.
177 In order to determine whether alanine is ideal for that purpose, we analyzed the effects
178 of substitutions in four proteins with ligand-bound structures: the hYAP65 WW domain,
179 the PSD95 pdz3 domain, the BRCA1 RING domain and GAL4. Amongst these four
180 proteins there were 4,884 mutations at 282 positions. We used relative solvent
181 exposure to classify each position as either buried or on the surface. We also
182 determined interface positions based on published structures and functional studies
183 (see **Methods**). We found that substitutions at interface positions are substantially more
184 damaging than substitutions at buried, non-interface or surface, non-interface positions
185 (**Fig. 4A**). This result is expected, given that all four deep mutational scans were
186 conducted using selections that depended on ligand binding. Alanine, along with
187 phenylalanine, isoleucine and methionine, are the least disruptive amino acid
188 substitutions at interface positions, suggesting that they may not be ideal for interface
189 detection.

190
191 We reasoned that the ideal substitution for detecting protein-ligand interfaces would
192 exhibit a large difference in mutational effect between interface and non-interface
193 positions. To formalize this idea, we used a mutational effect threshold. If a substitution
194 at a particular position had a mutational effect below the threshold, we classified that
195 position as “interface.” Conversely, if the mutational effect was above the threshold that
196 position was classified as “non-interface.” For each substitution, we varied the
197 mutational effect threshold from the maximum mutational effect score to the minimum
198 effect in 200 steps. At each step, we compared the true interface positions to those

199 determined using the mutational effect threshold procedure. We then constructed
200 receiver operating characteristic (ROC) curves. The area under each ROC curve
201 revealed the ability of that substitution to discriminate between true interface and non-
202 interface positions. Surprisingly, we found that alanine had among the worst
203 discriminatory power (**Fig. 4B, Fig. S4**). Substitutions that were highly disruptive at
204 interfaces, like asparagine, glutamine, aspartic acid and glutamic acid, had the best
205 discriminatory power. Next, we calculated the fraction of true interface positions
206 detected by each amino acid substitution at a 5% false positive rate. Here, we found
207 that asparagine and glutamine substitutions revealed over 60% of the true interface
208 positions; aspartic acid and glutamic acid substitutions also performed well (**Fig. S5**).
209 However, alanine substitutions detected fewer than 20% of the true interface positions
210 at a 5% false positive rate. Thus, asparagine, glutamine, aspartic acid or glutamic acid
211 substitutions are all better choices than alanine for detecting protein-ligand interfaces.

212

213 **Discussion**

214 Alanine scanning mutagenesis is a widely-used method for identifying protein positions
215 that are important for function or ligand binding. Alanine was selected on rational
216 grounds: it is physicochemically innocuous and constitutes a deletion of the side chain
217 at the β -carbon. By analyzing tens of thousands of mutations in fourteen proteins, we
218 have determined that alanine is not the most revealing substitution. In fact, many
219 superior choices exist. For example, histidine and asparagine substitutions have an
220 effect close to the median, and these substitutions correlate best with the effects of all
221 other substitutions. Thus, they better represent the effects of mutations generally.

222 Asparagine, glutamine, aspartic acid and glutamic acid are the most useful substitutions
223 for detecting ligand interface positions. Thus, our work highlights the utility of large-scale
224 mutagenesis data and suggests that alanine is not necessarily the best choice for future
225 single substitution mutational scans whose goal is to identify functionally important
226 positions or map protein-ligand interfaces.

227
228 However, our conclusions are based on only fourteen proteins. While these proteins are
229 diverse in structure and function, they may not fully reflect the mutational propensities of
230 other proteins. For example, tryptophan scanning mutagenesis is often applied to
231 transmembrane domains²¹⁻²³, which were absent from the proteins we analyzed. Thus,
232 our conclusions are most applicable to soluble proteins. Furthermore, we do not
233 address specialized applications of single amino acid scanning mutagenesis. For
234 example, cysteine scanning mutagenesis has been used to introduce disulfide bridges⁶
235 and glycine scanning mutagenesis has been used to increase conformational
236 flexibility²⁴. Our conclusions do not apply to these situations. Finally, the deep
237 mutational scanning data we analyzed arises from genetic selections for protein
238 function. Biochemical assays might reveal different patterns. However, we note that a
239 few of the large-scale mutagenesis data sets we used were benchmarked against and
240 found to be consistent with biochemical assay results^{25,26}.

241
242 Deep mutational scanning can reveal the functional consequences of all possible single
243 amino acid substitutions in a protein. However, these experiments can be expensive or
244 unwieldy. Therefore, scanning mutagenesis with one or a few amino acids will remain

245 useful for determining functionally important positions, probing protein-ligand
246 interactions and answering other specific questions. Our results could be used to guide
247 future single amino acid scanning mutagenesis experiments, enabling selection of the
248 amino acid best suited for the goals of the experiment.

249

250 **Acknowledgements**

251 This work was supported by the National Institute of General Medical Sciences
252 (1R01GM109110 to D.M.F.). V.E.G. is a National Science Foundation Graduate
253 Research Fellow (DGE-1256082) and R.J.H. is a Damon Runyon Cancer Research
254 Foundation Fellow (DRG-2224-15). We thank Lea Starita for helpful discussions and
255 advice.

256

257 **Author Contributions**

258 D.M.F. conceived of the project. V.E.G. and R.J.H. curated and rescaled the data sets.
259 V.E.G. and D.M.F. analyzed the data and wrote the paper.

260

261 **Materials and Methods**

262

263 **Data curation and rescaling**

264 We curated a subset of the published deep mutational scanning data sets. We excluded
265 deep mutational scans of non-natural proteins, because the mutational properties of
266 natural and non-natural proteins could differ. The result was a set of sixteen deep
267 mutational scans of fourteen proteins (**Table 1**). BRCA1 and UBI4 each have two large-

268 scale mutagenesis data sets corresponding independent experiments in which different
269 functions were assayed (e.g. ligand binding or catalytic activity). We treated these data
270 sets separately, and did not perform any averaging of mutational effects between the
271 data sets. Additionally, we removed any variants with more than one amino acid
272 substitution from all the data sets.

273
274 Most of the data sets reported mutational effect scores as the log-transformed ratio of
275 mutant frequency before and after selection, divided by wild type frequency before and
276 after selection. For data sets that used a different scoring scheme, we recalculated
277 mutational effect scores as the log-transformed ratio of mutant frequency before and
278 after selection, divided by wild type frequency before and after selection. Given that the
279 assays used to detect mutational effect differ, we rescaled the reported mutational effect
280 scores for each data set. First, we subtracted the median effect of synonymous
281 mutations from each reported effect score and then divided by the negative of the
282 bottom 1% of reported effect scores. Finally, we added 1. In cases where synonymous
283 mutational effect scores were unavailable, we omitted the synonymous score median
284 subtraction step. Our rescaling scheme is expressed as

$$S_{i,scaled} = \frac{S_{i,reported} - S_{median\ synonymous}}{-S_{median\ bottom\ 1\%}} + 1$$

285 where S is the mutational effect score. Our normalization scheme resulted in scaled
286 mutational effect scores where the most damaging mutations have effect scores ≈ 0 and
287 wild-type-like mutations have scores ≈ 1 .

288

289 Unless otherwise stated, we used all of the rescaled mutational effect data for each
290 analysis. In each analysis, we used median as a summary statistic rather than mean
291 because the frequency distributions of mutational effect are bimodal rather than
292 Gaussian (**Fig. S1**).

293

294 **Variant annotation**

295 DSSP was used to annotate the secondary structure and absolute solvent accessibility
296 of each wild type amino acid in our data set (http://swift.cmbi.ru.nl/gv/dssp/DSSP_3.html).

297 To estimate the relative solvent accessibility of amino acids, we divided absolute solvent
298 accessibility as determined using DSSP by the total surface area of each amino acid.

299 Amino acids with relative solvent accessibilities greater than 0.2 were labeled as

300 “surface”, whereas amino acids with relative solvent accessibilities less than 0.2 were
301 labeled as “buried”²⁷.

302

303 **Identification of interface positions**

304 Four proteins in our data set had high-resolution PDB structures with peptide or

305 nucleotide ligands, Gal4 (3COQ), BRCA1 RING domain (1JM7), PSD95 pdz3 domain

306 (1BE9) and hYAP65 WW domain (1JMQ). We determined interface positions from the

307 literature^{15,28-30}. The interface positions in hYAP65 WW domain were 188, 190, 197 and

308 199. The interface positions in BRCA1 RING domain were 11, 14, 18, 93 and 96. PSD95

309 pdz3 domain positions were 318, 322-327, 329, 339, 372 and 379. Gal4 interface

310 positions were 9, 15, 17, 18, 20, 21, 43, 46 and 51.

311

312 **Construction of ROC curves**

313 We constructed empirical ROC curves to illustrate the power of each substitution to
314 discriminate between interface and non-interface positions, determined as described
315 above. First, we defined a discrimination threshold, such that positions with a mutational
316 effect score below the threshold were classified “interface” and positions with a
317 mutational effect score above the threshold were classified as “non-interface.” For each
318 substitution, we varied this discrimination threshold from the maximum mutational effect
319 score to the minimum mutational effect score in 200 steps, calculating the true positive
320 interface detection rate (TPR) and false positive interface detection rate (FPR) at each
321 step. The TPR was calculated by dividing the number of interface positions with scores
322 below the mutational effect threshold by the total number of interface positions. The
323 FPR was calculated by dividing the number of non-interface positions with scores below
324 the mutational effect threshold by the total number of non-interface positions. ROC
325 curves were constructed by plotting the TPR and FPR for each of the 200 mutational
326 effect thresholds. The area under each ROC curve was determined in R using the auc()
327 function in the pROC package (<https://cran.r-project.org/web/packages/pROC/pROC.pdf>).

328

329 **Data and code availability**

330 The data sets used in this study came from a variety of published works (see **Table 1**).

331 The curated data sets and code for generating figures can be found at:

332 <https://github.com/FowlerLab/>

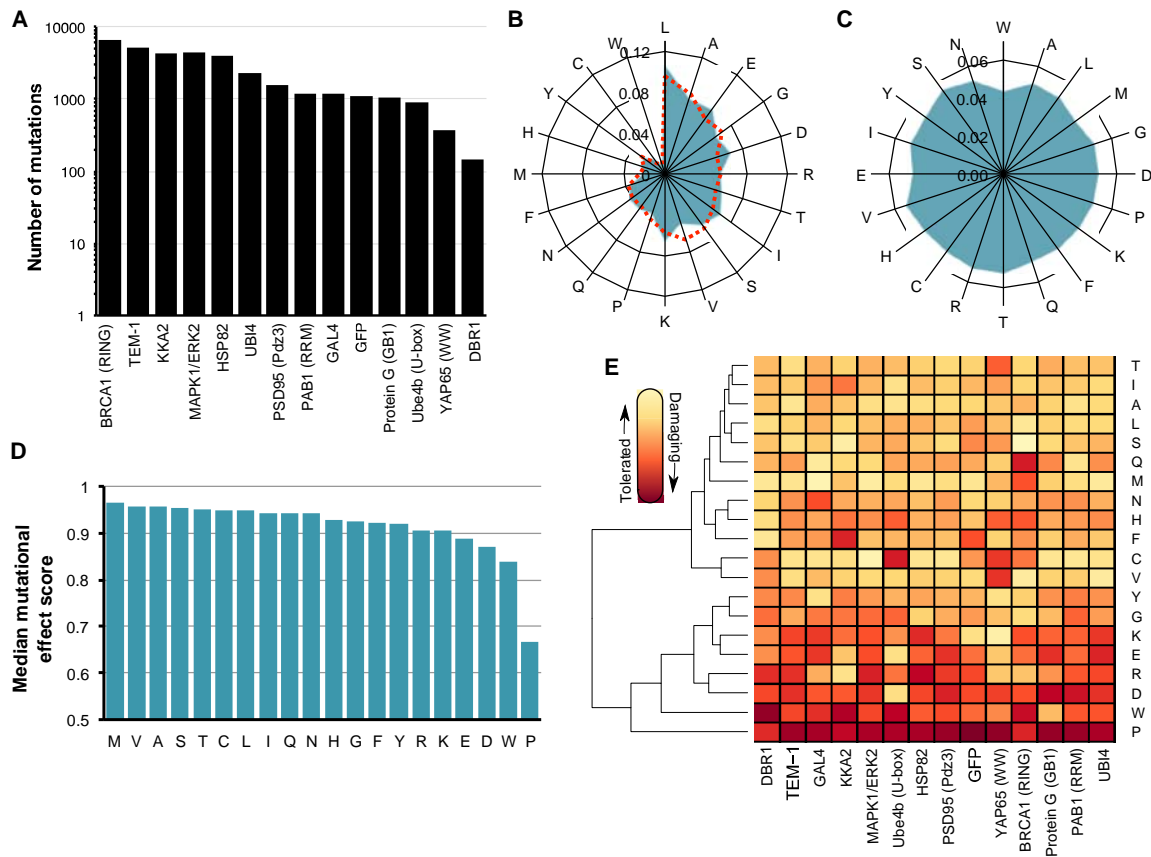
| Data set | Number of mutations | Mutagenized positions | Mutational completeness* | Organism | Selected Phenotype | Citation |
|--|---------------------|-----------------------|--------------------------|----------------------------------|-------------------------------|----------|
| Aminoglycoside kinase | 4,234 | 264 | 84% | <i>K. pneumoniae</i> | Antibiotic resistance | 31 |
| BRCA1 RING domain - BARD1 binding | 1,748 | 102 | 90% | <i>H. sapiens</i> | Binding activity (Y2H) | 28 |
| BRCA1 RING domain - E3 ligase activity | 4,872 | 303 | 85% | <i>H. sapiens</i> | Ubiquitin ligase activity | 28 |
| DBR1 | 144 | 25 | 30% | <i>H. sapiens</i> | RNA enzyme activity | 32 |
| Gal4 | 1,196 | 64 | 98% | <i>H. sapiens</i> | Transcription factor activity | 33 |
| GFP | 1,084 | 235 | 24% | <i>A. victoria</i> | Fluorescence | 34 |
| Hsp82 | 4,021 | 219 | 97% | <i>S. cerevisiae</i> | Chaperone activity | 35 |
| hYAP65 WW domain | 363 | 33 | 58% | <i>H. sapiens</i> | Ligand binding | 15 |
| MAPK1/ERK2 | 4,470 | 359 | 66% | <i>H. sapiens</i> | Kinase activity | 36 |
| Pab1 | 1,188 | 75 | 83% | <i>S. cerevisiae</i> | mRNA binding | 37 |
| Protein G GB1 domain | 1,045 | 55 | 100% | <i>Streptococcus sp. group G</i> | IgG-Fc binding | 25 |
| PSD95 pdz3 domain | 1,577 | 83 | 100% | <i>H. sapiens</i> | Ligand binding | 26 |
| TEM1 β -lactamase | 5,198 | 287 | 95% | <i>E. coli</i> | Antibiotic resistance | 38 |
| Ube4b U-box | 899 | 102 | 46% | <i>S. cerevisiae</i> | Ubiquitin ligase activity | 39 |
| Ubi4 - activity | 1,249 | 75 | 88% | <i>S. cerevisiae</i> | Ubiquitin ligase activity | 40 |
| Ubi4 - activation by E1 | 1,085 | 60 | 95% | <i>S. cerevisiae</i> | Ubiquitin ligase activity | 41 |

*proportion of all possible single amino acid mutations in mutagenized region observed

333

334 **Table 1. Large-scale mutagenesis data sets used in this study**

335 **Figure Legends**



336

337 **Figure 1. Large-scale mutagenesis data from fourteen proteins. (A)** The number of

338 single amino acid mutations with effect scores in each of the fourteen proteins is shown.

339 **(B)** A radar plot shows the relative frequency of occurrence of each amino acid in the

340 wild type sequences of the fourteen proteins (blue) or in 554,515 proteins in the UniProt

341 Knowledgebase² (dashed red). **(C)** A radar plot shows the relative frequency of each of

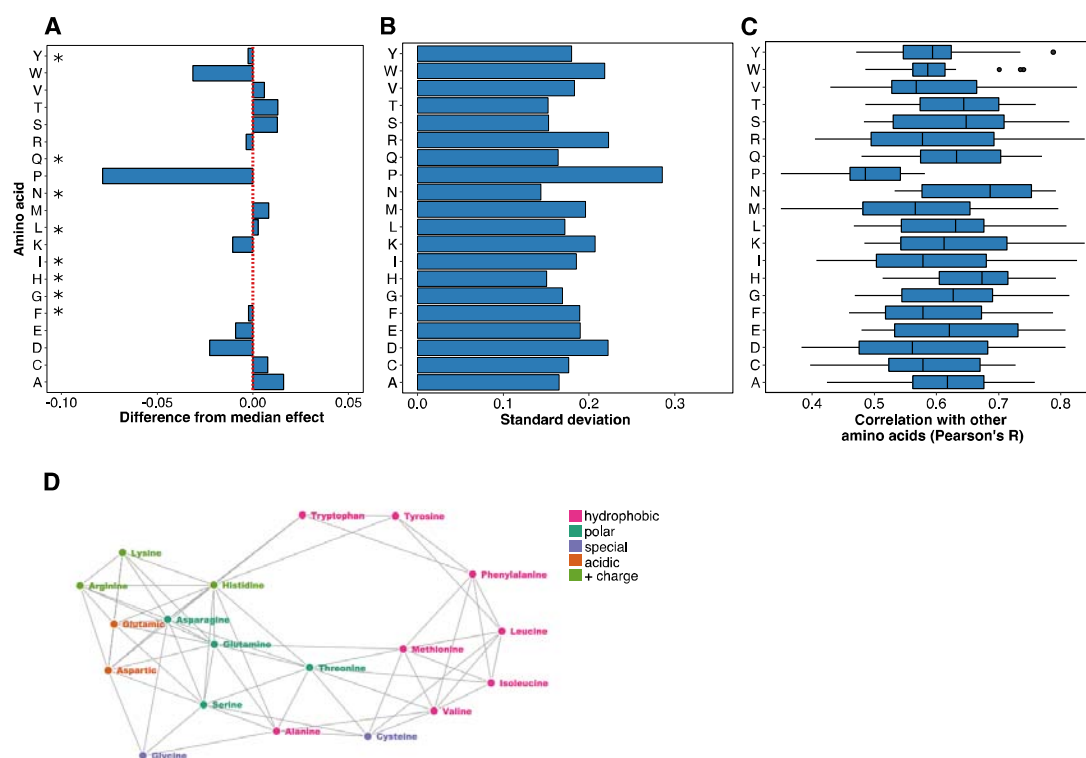
342 the twenty amino acid substitutions in the large-scale mutagenesis data sets for all

343 fourteen proteins. **(D)** The median mutational effect score of each amino acid

344 substitution is shown for 34,373 mutations at 2,236 positions in all fourteen proteins. **(E)**

345 A heat map shows the median mutational effect score of each amino acid substitution

346 for each protein separately. Yellow indicates tolerated substitutions while orange
347 indicated damaging substitutions. Amino acids and proteins were ordered according to
348 similarity using hierarchical clustering with the hclust function from the heatmap2
349 package in R. The dendrogram is shown only for amino acid clustering.
350



351

352 **Figure 2. Histidine and asparagine substitutions best represent the effect of other**

353 **substitutions. (A)** For each of the 882 positions where the mutational effects of all

354 nineteen substitutions were measured, the difference from the median effect was

355 calculated for each substitution at each position. The median of these differences

356 across all positions for each substitution is shown, with the red line indicating a median

357 difference of zero. A paired, two-sided Wilcoxon rank sum test was used to determine

358 whether each substitution's difference from the median effect across all positions was

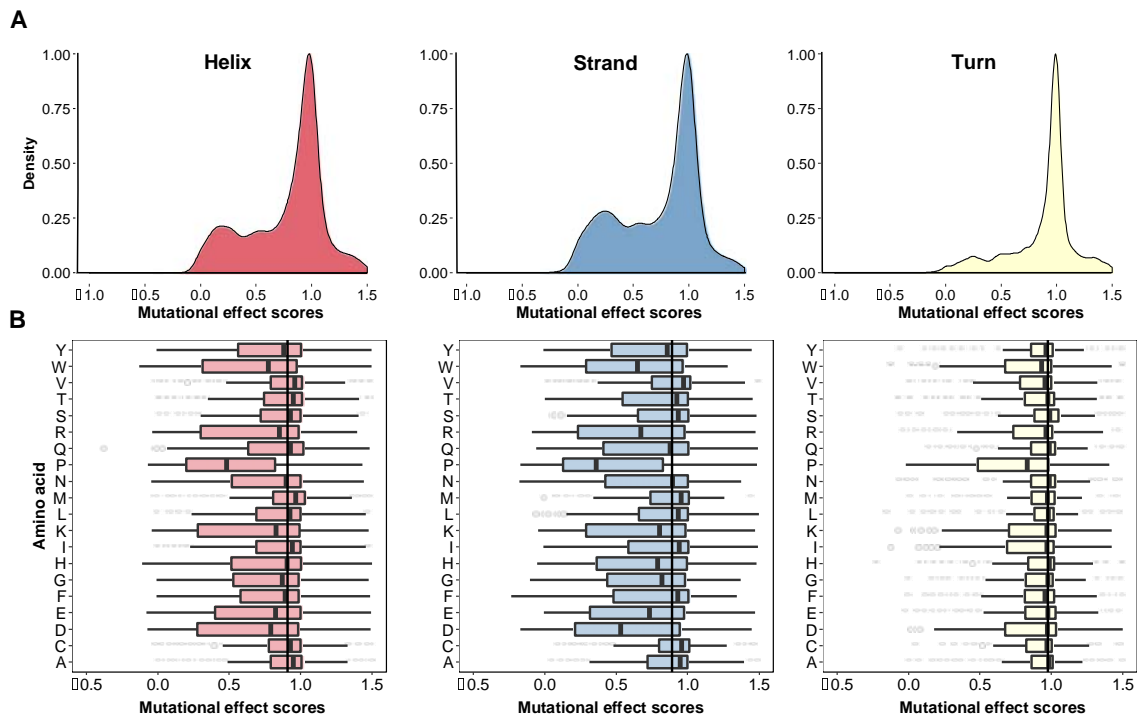
359 equal to zero (* indicates substitutions with a Bonferroni-corrected p -value > 0.01 ; Table

360 S1). **(B)** The standard deviation of each substitution's differences from the median effect

361 at the 882 positions where the mutational effects of all nineteen substitutions were

362 measured is shown. **(C)** For each substitution, Pearson correlation coefficients were

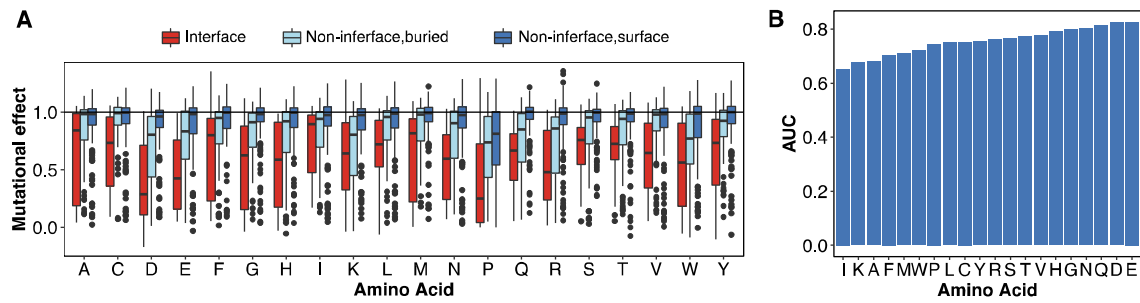
363 calculated for the mutational effects of that substitution with every other substitution at
364 each position. The distribution of correlation coefficients for each substitution is shown.
365 **(D)** These pairwise mutational effect score correlations are also illustrated using a force
366 directed graph. Each node represents an amino acid and each edge force value is the
367 Pearson correlation coefficient for the mutational effect scores of the two amino acid
368 substitutions connected by the edge. To reduce the density of edges, only the top 40%
369 of Pearson correlation coefficients were included. This cutoff removed proline from the
370 graph. Amino acids are colored by physicochemical type. The graph was constructed
371 using the networkD3 package in R.
372
373



374

375 **Figure 3. Secondary structural context of mutational effects.** (A) Density functions
376 describing the distribution of mutational effect scores for each substitution are shown for
377 three different structural contexts as determined using DSSP: β -sheets (left panel, N =
378 4,796), α -helices (middle panel, N = 8,669) and turns (right panel, N = 3,329). (B) The
379 mutational effect score distributions for each substitution in β -sheets (left panel), α -
380 helices (middle panel), and turns (right panel) are shown. The vertical line in each panel
381 represents the median effect score for all substitutions in that secondary structure type.

382



383

384 **Figure 4. Alanine is not especially useful for identifying positions in protein-**
385 **ligand interfaces. (A)** The distribution of mutational effect scores for every substitution
386 in four proteins with ligand-bound structures (hYAP65 WW domain, PSD95 pdz3
387 domain, BRCA1 RING domain (BARD1 binding) and Gal4) is shown at ligand interface
388 positions as reported in the literature, and for non-interface buried positions or non-
389 interface surface positions. **(B)** A mutational effect threshold was defined such that
390 positions with a mutational effect below the threshold were classified as “interface,”
391 whereas positions with a mutational effect above the threshold were classified as “non-
392 interface.” ROC curves for each amino acid were generated by varying this threshold.
393 The area under each ROC curve is shown, illustrating the power of each substitution to
394 discriminate between interface and non-interface positions.

395

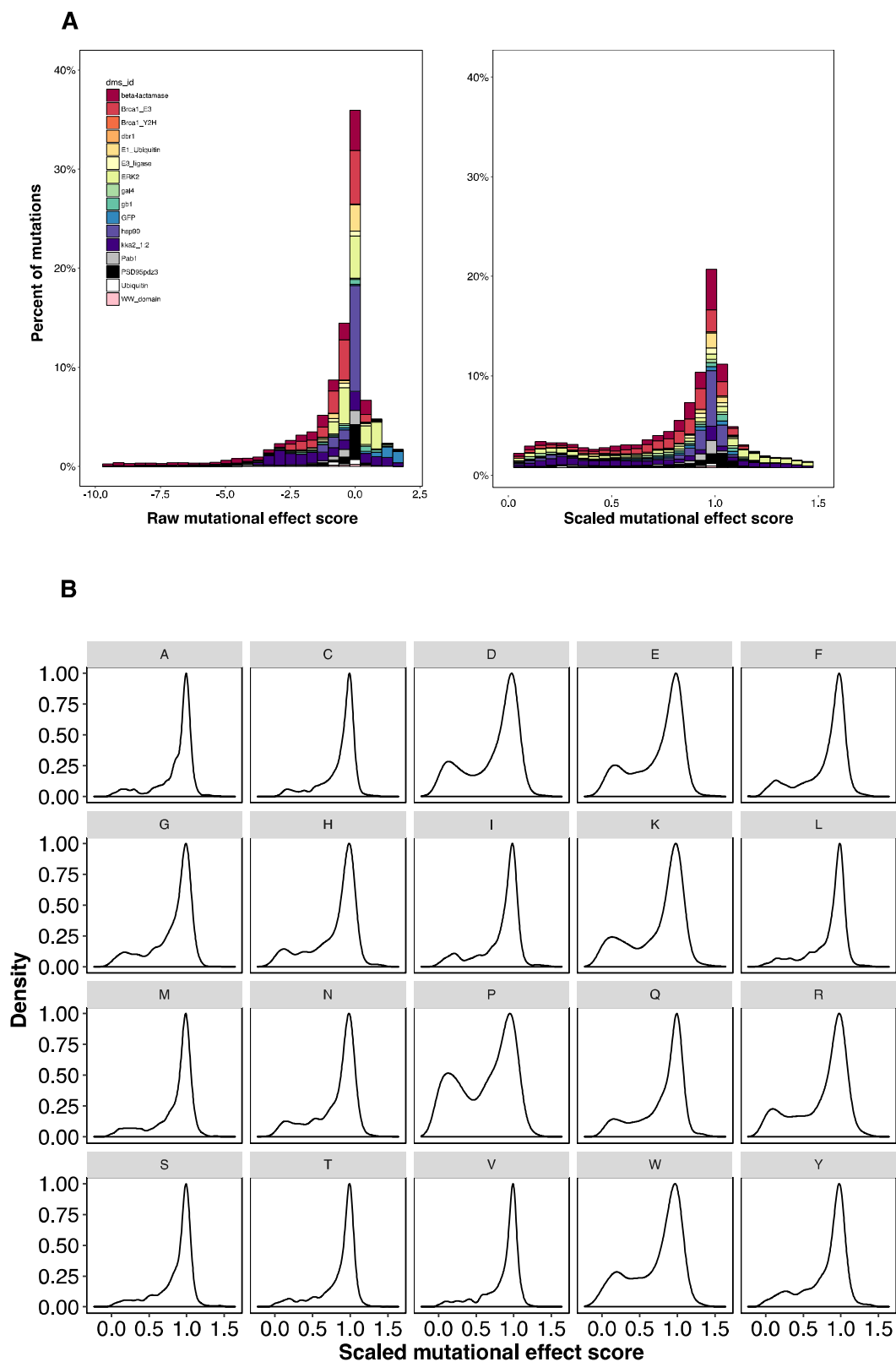
396 References Cited

- 397 1. Cunningham, B. C. & Wells, J. A. High-resolution epitope mapping of hGH-
398 receptor interactions by alanine-scanning mutagenesis. *Science* **244**, 1081–1085
399 (1989).
- 400 2. Magrane, M. & Consortium, U. UniProt Knowledgebase: a hub of integrated
401 protein data. *Database* **2011**, 1–13 (2011).
- 402 3. Otzen, D. E. & Fersht, A. R. Side-chain determinants of beta-sheet stability.
403 *Biochemistry* **34**, 5718–5724 (1995).
- 404 4. Myers, J. K., Pace, C. N. & Scholtz, J. M. Helix propensities are identical in
405 proteins and peptides. *Biochemistry* **36**, 10923–10929 (1997).
- 406 5. Nanevicz, T. *et al.* Mechanisms of thrombin receptor agonist specificity. Chimeric
407 receptors and complementary mutations identify an agonist recognition site. *J.*
408 *Biol. Chem.* **270**, 21619–21625 (1995).
- 409 6. Kanaya, E., Kanaya, S. & Kikuchi, M. Introduction of a non-native disulfide bridge
410 to human lysozyme by cysteine scanning mutagenesis. *Biochem. Biophys. Res.*
411 *Commun.* **173**, 1194–1199 (1990).
- 412 7. Valbuena, J. J. *et al.* Plasmodium falciparum normocyte binding protein (PfNBP-
413 1) peptides bind specifically to human erythrocytes. *Peptides* **24**, 1007–1014
414 (2003).
- 415 8. Woods, A. C., Guillemette, J. G., Parrish, J. C., Smith, M. & Wallace, C. J.
416 Synergy in protein engineering. Mutagenic manipulation of protein structure to
417 simplify semisynthesis. *J. Biol. Chem.* **271**, 32008–32015 (1996).
- 418 9. Borngräber, S. *et al.* Shape and specificity in mammalian 15-lipoxygenase active
419 site. The functional interplay of sequence determinants for the reaction specificity.
420 *J. Biol. Chem.* **274**, 37345–37350 (1999).
- 421 10. Vandemeulebroucke, A., De Vos, S., Van Holsbeke, E., Steyaert, J. & Versées,
422 W. A flexible loop as a functional element in the catalytic mechanism of
423 nucleoside hydrolase from *Trypanosoma vivax*. *J. Biol. Chem.* **283**, 22272–22282
424 (2008).
- 425 11. Zhang, L. *et al.* Mapping hydration dynamics around a protein surface. *Proc. Natl.*
426 *Acad. Sci.* **104**, 18461–18466 (2007).
- 427 12. Xiao, Y., Wigneshweraraj, S. R., Weinzierl, R., Wang, Y.-P. & Buck, M.
428 Construction and functional analyses of a comprehensive sigma54 site-directed
429 mutant library using alanine-cysteine mutagenesis. *Nucl. Acids Res.* **37**, 4482–
430 4497 (2009).
- 431 13. Bromberg, Y. & Rost, B. Comprehensive in silico mutagenesis highlights
432 functionally important residues in proteins. *Bioinformatics* **24**, i207–i212 (2008).
- 433 14. Fowler, D. M. & Fields, S. Deep mutational scanning: a new style of protein
434 science. *Nat. Methods* **11**, 801–807 (2014).
- 435 15. Fowler, D. M. *et al.* High-resolution mapping of protein sequence-function
436 relationships. *Nat. Methods* **7**, 741–746 (2010).
- 437 16. Dayhoff, M. O. *Atlas of Protein Sequence and Structure.* (1978).
- 438 17. Henikoff, S. & Henikoff, J. G. Amino acid substitution matrices from protein
439 blocks. *Proc. Natl. Acad. Sci.* **89**, 10915–10919 (1992).
- 440 18. Grantham, R. Amino acid difference formula to help explain protein evolution.

- 441 *Science* **185**, 862–864 (1974).
- 442 19. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern
443 recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–
444 2637 (1983).
- 445 20. Costantini, S., Colonna, G. & Facchiano, A. M. Amino acid propensities for
446 secondary structures are influenced by the protein structural class. *Biochem.*
447 *Biophys. Res. Commun.* **342**, 441–451 (2006).
- 448 21. Sharp, L. L., Zhou, J. & Blair, D. F. Tryptophan-scanning mutagenesis of MotB, an
449 integral membrane protein essential for flagellar rotation in *Escherichia coli*.
450 *Biochemistry* **34**, 9166–9171 (1995).
- 451 22. Rasmussen, T. *et al.* Properties of the Mechanosensitive Channel MscS Pore
452 Revealed by Tryptophan Scanning Mutagenesis. *Biochemistry* **54**, 4519–4530
453 (2015).
- 454 23. Depriest, A., Phelan, P. & Martha Skerrett, I. Tryptophan scanning mutagenesis of
455 the first transmembrane domain of the innexin Shaking-B(Lethal). *Biophys. J.* **101**,
456 2408–2416 (2011).
- 457 24. Weinglass, A. B., Smirnova, I. N. & Kaback, H. R. Engineering conformational
458 flexibility in the lactose permease of *Escherichia coli*: use of glycine-scanning
459 mutagenesis to rescue mutant Glu325-->Asp. *Biochemistry* **40**, 769–776 (2001).
- 460 25. Olson, C. A., Wu, N. C. & Sun, R. A Comprehensive Biophysical Description of
461 Pairwise Epistasis throughout an Entire Protein Domain. *Current Biology* **24**,
462 2643–2651 (2014).
- 463 26. McLaughlin, R. N., Jr, Poelwijk, F. J., Raman, A., Gosal, W. S. & Ranganathan, R.
464 The spatial architecture of protein function and adaptation. *Nature* **491**, 138–142
465 (2012).
- 466 27. Chen, H. & Zhou, H.-X. Prediction of solvent accessibility and sites of deleterious
467 mutations from protein sequence. *Nucl. Acids Res.* **33**, 3193–3199 (2005).
- 468 28. Starita, L. M. *et al.* Massively Parallel Functional Analysis of BRCA1 RING
469 Domain Variants. *Genetics* **200**, 413–422 (2015).
- 470 29. Doyle, D. A. *et al.* Crystal Structures of a Complexed and Peptide-Free
471 Membrane Protein–Binding Domain: Molecular Basis of Peptide Recognition by
472 PDZ. *Cell* **85**, 1067–1076 (1996).
- 473 30. Marmorstein, R. & Carey, M. DNA recognition by GAL4: structure of a protein-
474 DNA complex. *Nature* (1992).
- 475 31. Melnikov, A., Rogov, P., Wang, L., Gnirke, A. & Mikkelsen, T. S. Comprehensive
476 mutational scanning of a kinase in vivo reveals substrate-dependent fitness
477 landscapes. *Nucl. Acids Res.* **42**, gku511–e112 (2014).
- 478 32. Findlay, G. M., Boyle, E. A., Hause, R. J., Klein, J. C. & Shendure, J. Saturation
479 editing of genomic regions by multiplex homology-directed repair. *Nature* **513**,
480 120–123 (2014).
- 481 33. Kitzman, J. O., Starita, L. M., Lo, R. S., Fields, S. & Shendure, J. Massively
482 parallel single-amino-acid mutagenesis. *Nat. Methods* **12**, 203–6– 4 p following
483 206 (2015).
- 484 34. Sarkisyan, K. S. *et al.* Local fitness landscape of the green fluorescent protein.
485 *Nature* **533**, 397–401 (2016).
- 486 35. Mishra, P., Flynn, J. M., Starr, T. N. & Bolon, D. N. A. Systematic Mutant

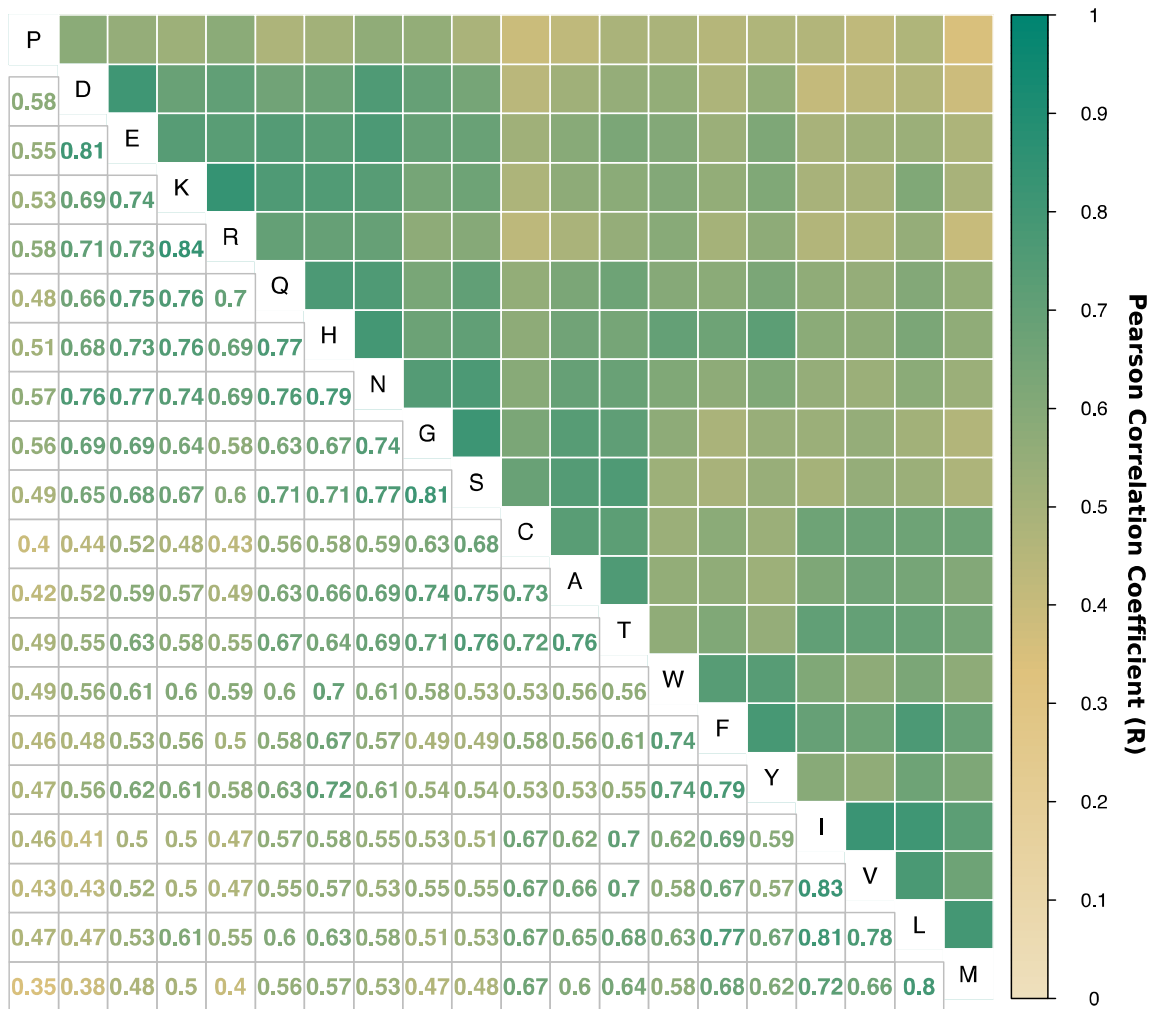
- 487 Analyses Elucidate General and Client-Specific Aspects of Hsp90 Function. *Cell*
488 *Reports* **15**, 588–598 (2016).
- 489 36. Brenan, L. *et al.* Phenotypic Characterization of a Comprehensive Set of
490 MAPK1/ERK2 Missense Mutants. *Cell Reports* **17**, 1171–1183 (2016).
- 491 37. Melamed, D., Young, D. L., Gamble, C. E., Miller, C. R. & Fields, S. Deep
492 mutational scanning of an RRM domain of the *Saccharomyces cerevisiae* poly(A)-
493 binding protein. *RNA* **19**, 1537–1551 (2013).
- 494 38. Firnberg, E., Labonte, J. W., Gray, J. J. & Ostermeier, M. A comprehensive, high-
495 resolution map of a gene's fitness landscape. *Mol. Biol. Evol.* **31**, 1581–1592
496 (2014).
- 497 39. Starita, L. M. *et al.* Activity-enhancing mutations in an E3 ubiquitin ligase identified
498 by high-throughput mutagenesis. *Proc. Natl. Acad. Sci.* **110**, E1263–72 (2013).
- 499 40. Roscoe, B. P., Thayer, K. M., Zeldovich, K. B., Fushman, D. & Bolon, D. N. A.
500 Analyses of the effects of all ubiquitin point mutants on yeast growth rate. *J. Mol.*
501 *Biol.* **425**, 1363–1377 (2013).
- 502 41. Roscoe, B. P. & Bolon, D. N. A. Systematic Exploration of Ubiquitin Sequence, E1
503 Activation Efficiency, and Experimental Fitness in Yeast. *J. Mol. Biol.* **426**, 2854–
504 2870 (2014).
- 505

506 Supplemental Figure Legends



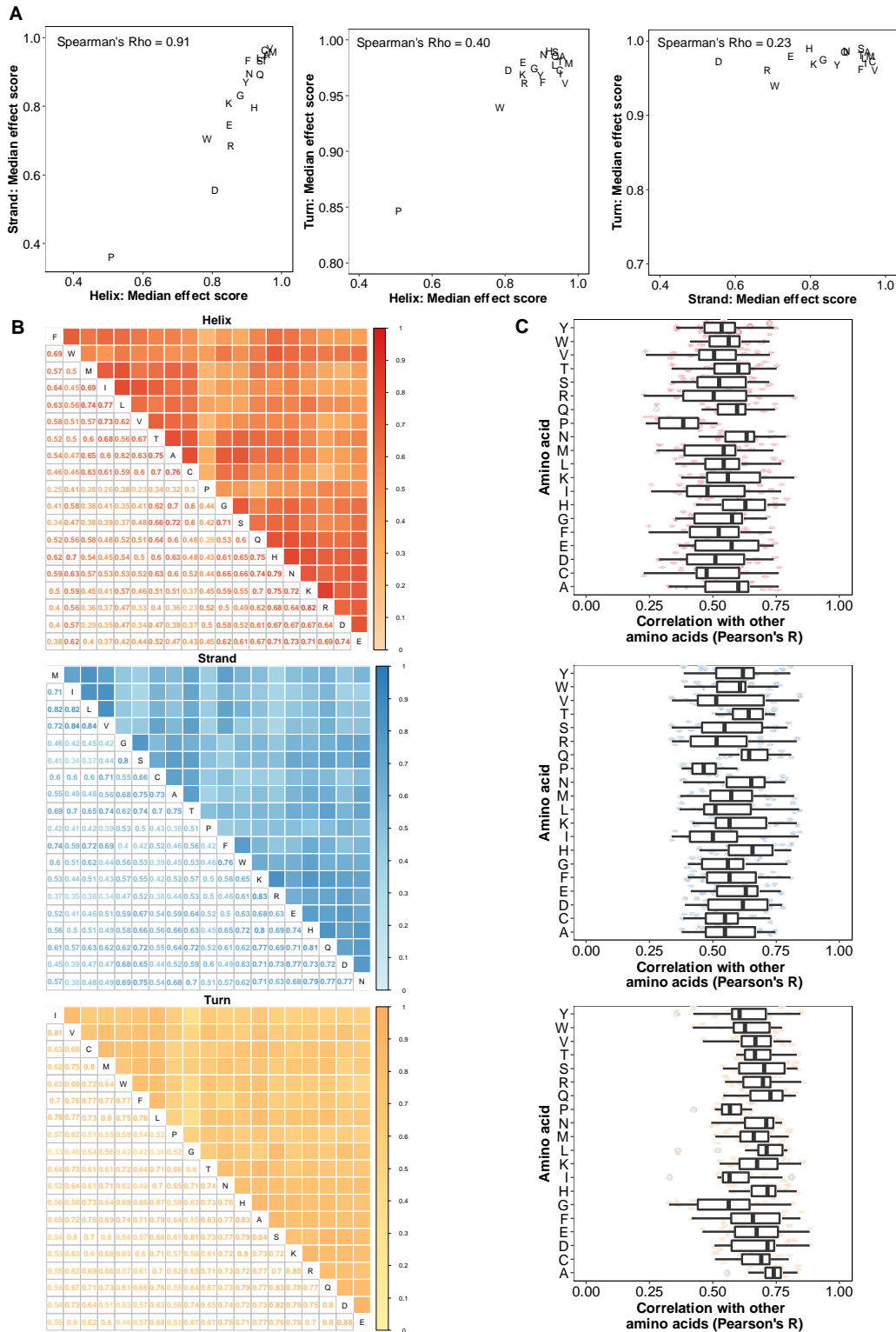
507

508 **Figure S1.** We curated large-scale mutagenesis data sets describing the effects of
509 34,373 mutations at 2,236 positions in fourteen proteins. To facilitate comparisons
510 between each data set, we rescaled mutational effect scores for each protein by
511 subtracting the median mutational effect score of all synonymous mutations in that
512 protein from each nonsynonymous mutational effect score and then dividing that
513 difference by the median of the bottom 1% of mutational effect scores (see **Methods**).
514 **(A)** Stacked histograms of the original scores (**left panel**) and rescaled scores (**right**
515 **panel**) are shown. **(B)** Density plots of the scaled mutational effect scores for each
516 amino acid substitution are shown.
517

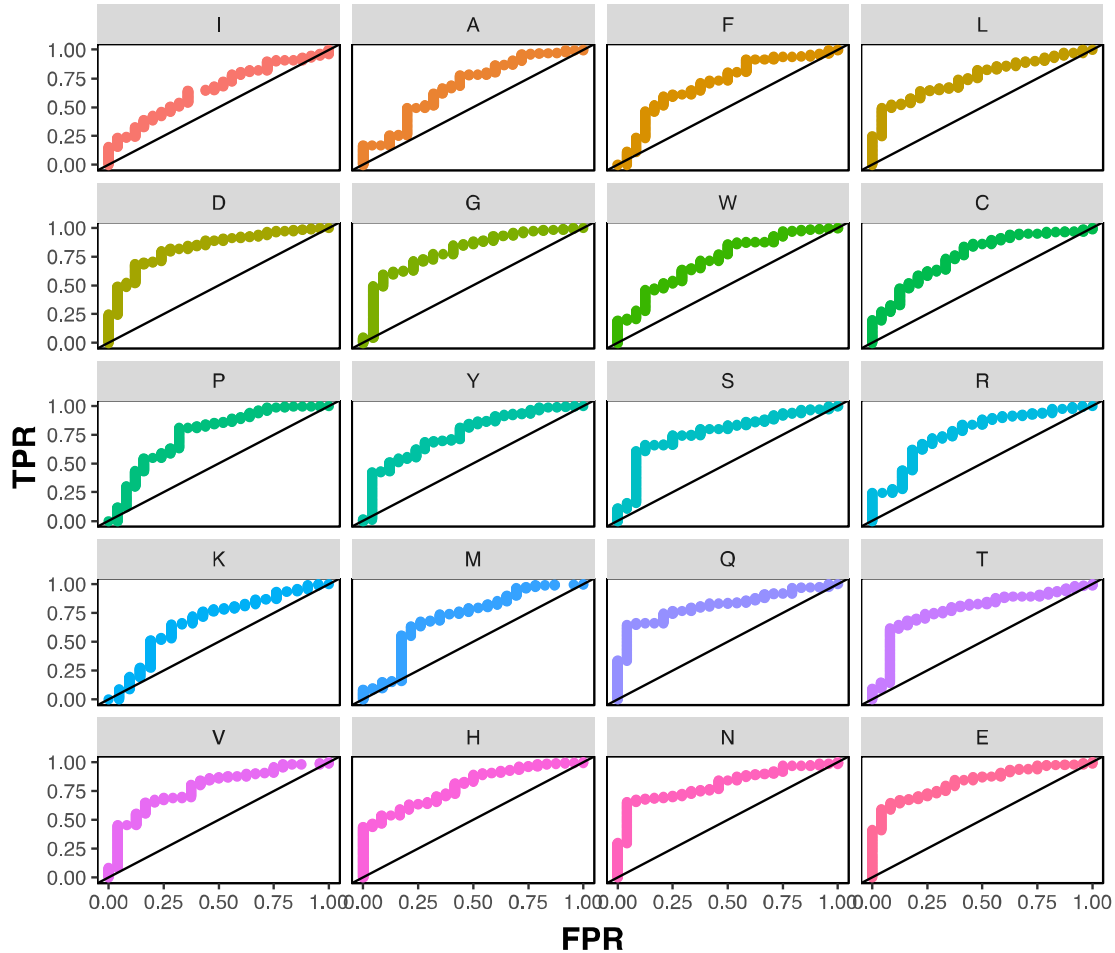


518

519 **Figure S2.** For each substitution, Pearson correlation coefficients were calculated for
 520 the mutational effect scores of that substitution with every other substitution at each
 521 position. A correlation plot of these Pearson coefficients is shown. Color indicates the
 522 Pearson correlation coefficient ranging from 0 (light brown) to 1 (green).



524 **Figure S3. (A)** For each amino acid substitution, the median mutational effect score
525 was calculated. The correlation between the median mutational effects for each
526 substitution in helices, strand and turns are shown in scatterplots, and Spearman's Rho
527 indicates the degree of rank correlation within each scatterplot. **(B)** Pearson correlation
528 coefficients were calculated for the mutational effects of each substitution with every
529 other substitution at every position. The Pearson correlation coefficient plots are shown
530 separately for α -helices (top), β -sheets (middle), and turns (bottom). **(C)** Boxplots show
531 the distribution of Pearson correlation coefficients for each amino acid type in three
532 structural contexts.

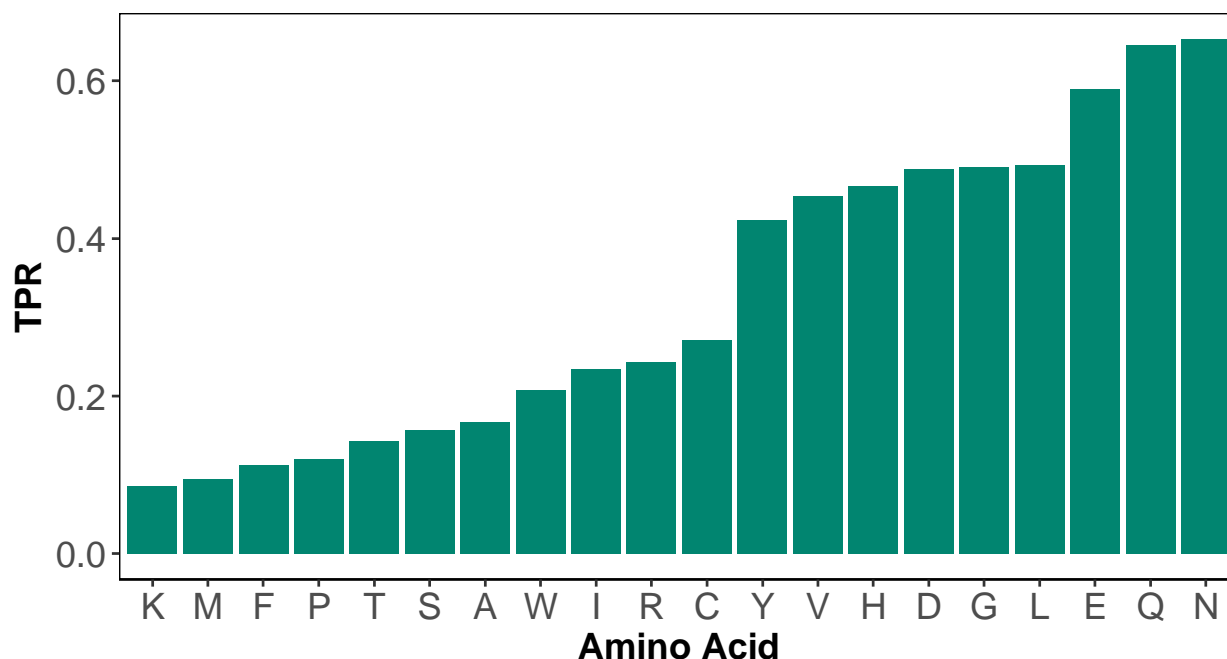


533

534 **Figure S4.** A mutational effect threshold was defined such that positions with a
535 mutational effect score below the threshold were classified as “interface,” whereas
536 positions with a mutational effect score above the threshold were classified as “non-
537 interface.” ROC curves were generated by varying this threshold for each amino acid
538 type in the four proteins with protein or DNA ligand-bound structures (hYAP65 WW
539 domain, PSD95 pdz3 domain, Gal4 and BRCA1 RING domain (BARD1 binding)).

540

541



542

543 **Figure S5.** A mutational effect threshold was defined such that positions with a
544 mutational effect score below the threshold were classified as “interface,” whereas
545 positions with a mutational effect score above the threshold were classified as “non-
546 interface.” A barplot shows each amino acid substitution’s true positive rate (TPR) for
547 detecting interface positions at a fixed, 5% non-interface position false positive rate.

548

549 **Supplemental Table**

550 **Table S1.** A table showing sample size, p -value and Bonferroni corrected p -value for
551 paired, two-sided Wilcoxon rank sum tests of the position median effect scores versus
552 each amino acid substitution’s effect scores. This analysis was restricted to the 882
553 positions where the effects of all 19 possible substitutions were scored.