# Non-catalytic binding sites induce weaker long-range evolutionary rate gradients than catalytic sites in enzymes

Avital Sharir-Ivry[1], Yu Xia[1*]

[1] Department of Bioengineering, McGill University, Montreal, QC H3A 0E9, Canada

Avital Sharir-Ivry: avital.ivry@mail.mcgill.ca

Corresponding author:

Yu Xia
Department of Bioengineering
McGill University
Montreal, Quebec H3A 0E9, Canada

**Email:** brandon.xia@mcgill.ca
**Office:** McConnell Engineering Building, Room 360
**Phone:** 514-398-5026 (O), 514-246-8751 (C)

# Abstract

Enzymes exhibit a strong long-range evolutionary constraint that extends from their catalytic site and affects even distant sites, where site-specific evolutionary rate increases monotonically with distance. While protein-protein sites in enzymes was previously shown to induce only a weak conservation gradient, a comprehensive relationship between different types of functional sites in proteins and the magnitude of evolutionary rate gradients they induce has yet to be established. Here, we systematically calculate the evolutionary rate (dN/dS) of sites as a function of distance from different types of binding sites on enzymes and other proteins: catalytic sites, non-catalytic ligand binding sites, allosteric binding sites, and protein-protein interaction sites. We show that catalytic binding sites indeed induce significantly stronger evolutionary rate gradient than all other types of non-catalytic binding sites. In addition, catalytic sites in enzymes with no known allosteric function still induce strong long-range conservation gradients. Notably, the weak long-range conservation gradients induced by non-catalytic binding sites on enzymes is nearly identical in magnitude to those induced by ligand binding sites on non-enzymes. Finally, we show that structural determinants such as local solvent exposure of sites cannot explain the observed difference between catalytic and non-catalytic functional sites. Our results suggest that enzymes and non-enzymes share similar evolutionary constraints only when examined from the perspective of non-catalytic functional sites. Hence, the unique evolutionary rate gradient from catalytic sites in enzymes is likely driven by the optimization of catalysis rather than ligand binding and allosteric functions.

45 # Introduction

46 Enzymes are key players in metabolic pathways and are crucial for cell function. They bind their chemical
47 reactants and enhance the production rate of chemical products by several orders of magnitude. The
48 catalytic function occurs in the catalytic site, which is usually composed of relatively buried residues in a
49 large cleft [1]. Due to their function, catalytic site residues are highly conserved and even their close vicinity
50 residues are under strong selective pressure where their conservation decreases with distance from the
51 catalytic site [1]. Recently it was shown that enzymes exhibit a long-range, nearly-linear conservation
52 gradient from their catalytic site that extends even up to ~30Å in distance [2].

53 Possible confounding factors of the observed long-range conservation gradient in enzymes include various
54 local structural properties known to drive evolutionary rate variation of protein sites [3–12]. The main local
55 structural determinants known are residue solvent accessibility and residue packing [6,8,9] such that residues
56 are generally less conserved with increasing solvent exposure or decreasing packing. Indeed, catalytic sites
57 are generally relatively buried and therefore could induce a conservation gradient based solely on solvent
58 exposure gradients. However, for enzymes where the catalytic site is on the surface, a strong conservation
59 gradient was observed as well [2], demonstrating that distance and solvent exposure contribute
60 independently to the conservation gradient. Still, given that protein fold usage is very different between
61 enzymes and non-enzymes [13–15], it is possible that other local structural determinants (other than solvent
62 exposure and residue packing) could potentially cause the strong conservation gradient from catalytic
63 sites in enzymes. However, we have previously shown that catalytically-inactive pseudoenzymes with
64 nearly identical tertiary structure as enzymes exhibit a significantly reduced conservation gradient,
65 thereby demonstrating that the observed conservation gradient in enzymes cannot be dominated by any
66 backbone-based local structural determinants [16].

This study addresses the following outstanding question: Are catalytic sites unique in their abilities to induce such strong long-range conservation gradients? What about other non-catalytic functional sites which share similar functional capacities of binding another molecule as catalytic binding sites? Using off-lattice protein model, it was shown that the requirement to maintain a specific ligand-binding site gives rise to a conservation gradient from the ligand-binding site [17,18]. For protein-protein interaction sites however, it was shown that interfacial sites induce significantly weaker conservation gradients than catalytic sites in enzymes [2]. Here, we present a data-driven proteome-level study of the evolutionary rate (dN/dS) of residues as a function of their distance from different types of functional sites in proteins. We consider four types of functional sites: catalytic sites, non-catalytic ligand binding sites, allosteric binding sites, and protein-protein interaction sites. We show that all types of non-catalytic binding sites induce significantly weaker long-range evolutionary rate gradients than catalytic binding sites in enzymes. Surprisingly, non-catalytic ligand-binding sites on enzymes do not induce significant long-range evolutionary rate gradients. Instead, the weak evolutionary rate gradient induced from non-catalytic ligand-binding sites on enzymes resembles that from ligand-binding sites on non-enzymes. Moreover, we show that catalytic sites in enzymes with no known allosteric function still induce a strong long-range evolutionary rate gradient, indicating that allosteric function is not the main driving force of the observed evolutionary rate gradients from catalytic sites. Lastly, we show that solvent exposure gradients cannot explain the differences in magnitude between evolutionary rate gradients induced from catalytic and non-catalytic ligand-binding sites. Taken together, our results suggest that the observed evolutionary rate gradient from catalytic sites in enzymes is primarily driven by the optimization and maintenance of catalytic function rather than ligand-binding or allosteric function.

# Results

## Non-catalytic ligand-binding sites induce weaker evolutionary rate gradients than catalytic binding sites

To examine whether evolutionary rate gradients are also induced from non-catalytic ligand-binding sites, we first identified such functional binding sites within enzymes and non-enzymes in the yeast proteome. As a starting point, we used a dataset of the yeast proteome that contains 1744 structurally annotated yeast proteins (see Methods). We screened them to find yeast proteins for which the structural model contains a biologically significant ligand bound to it based on the BioLip database [19], excluding ions. We found 486 ligand-binding sites on 240 enzymes and 462 ligand-binding sites on 276 proteins which are not known to be enzymes (non-enzymes). Ligand-binding residues were identified from BioLip, and catalytic site residues in enzymes were identified using the Catalytic Site Atlas [20]. The ligand-binding sites on enzymes were then subdivided into 250 ligand-binding sites that include catalytic site residues, and 236 ligand-binding sites that are completely separated from the catalytic site. On average, the evolutionary rates of the structurally modeled non-enzymes and enzymes in our datasets are not significantly different (0.0676±0.0009 and 0.0656 ±0.0007 respectively). For each residue, we calculated the distance to the closest ligand-binding residue. Average evolutionary rate (dN/dS) was then calculated for residues over distance bins to obtain the evolutionary rate gradients.

The slope of the linear fit for the evolutionary rate gradient from catalytic binding sites is significantly larger than the slope for the evolutionary rate gradient from non-catalytic ligand-binding sites (t-test, $P<0.001$, Table 1 and Figure 1). Surprisingly, the evolutionary rate gradient induced from non-catalytic ligand-binding sites on enzymes is weak and actually resembles the evolutionary rate gradient induced from ligand-binding sites on non-enzymes (t-test, $P=0.17$). The weak evolutionary rate gradient from non-

112 catalytic ligand-binding sites suggests that ligand-binding functionality of catalytic sites is probably not the
113 main determinant of the strong evolutionary rate gradients induced from them.

## 114 Allosteric binding sites induce weaker long-range evolutionary rate gradients than catalytic 115 sites

116 Next, we focused on the special case of allosteric binding sites. We collected those yeast proteins that
117 have a structural model which is known to have an allosteric function with its allosteric site residues
118 annotated in the allosteric database (ASD) [21]. We found 190 allosteric proteins, of which 108 are known
119 enzymes with known catalytic site residues. All residues were binned according to their distance from the
120 closest allosteric binding residue as well as from the closest catalytic residue in the case of enzymes.
121 Average evolutionary rate (dN/dS) was calculated for the residues in each distance bin.

122 The slope of the evolutionary rate gradient induced from allosteric binding sites is significantly smaller
123 than that induced from catalytic sites (t-test, $P<0.001$, Figure 2 and Table 2). Allosteric binding sites in
124 enzymes modulate the activity of catalytic sites in that the catalytic site is shifting into its active
125 conformation upon a binding event in the distant allosteric binding site [22,23]. Despite this long-range
126 interaction between catalytic sites and allosteric binding sites, our results suggest that allosteric function
127 is not the main determinant of the long-range evolutionary rate gradient induced from catalytic sites in
128 enzymes.

## 129 Yeast-annotated non-catalytic ligand-binding sites induce weaker evolutionary rate gradients 130 than catalytic binding sites

131 The annotations of functional binding sites were made according to the homology-based
132 structural models of the *Saccharomyces cerevisiae* proteins. These models are usually based on
133 solved structures of homologous proteins in other species. In order to rule out the confounding

134 factor that some of our homology-based annotations of functional sites are false positives in
135 yeast, we examined a subset of yeast proteins which are also known to have the relevant binding
136 functionality in yeast. The results for these yeast proteins with high-confidence annotations in
137 Figure 3 and Table 3 clearly show a significant difference between the strong evolutionary rate
138 gradients induced from catalytic binding sites on enzymes and all other non-catalytic binding sites.
139 Due to the significantly smaller size of the yeast-annotated allosteric site dataset (only 11 sites),
140 we reduced the number of distance bins to five and the slope is compared to the slope of the
141 gradient induced from catalytic binding sites based on five bins as well. Overall, our results are
142 robust to possible false positives of our homology-based functional binding site annotations for
143 yeast proteins.

144 **Catalytic sites in non-allosteric enzymes induce strong evolutionary rate gradients**

145 For comparison, we also identified 'non-allosteric enzymes' in our dataset as those proteins with an
146 enzyme structural model with known catalytic residues but with no known allosteric function (219
147 proteins). Notably, similar to allosteric enzymes, catalytic sites in non-allosteric enzymes also exhibit a
148 strong evolutionary rate gradient that extends to distant sites (Figure 4 and Table 2). The existence of
149 evolutionary rate gradient from catalytic sites appears to be independent of the existence of allosteric
150 function in the enzyme. This result further supports the hypothesis that the allosteric function is not the
151 main determinant of evolutionary rate gradients induced from catalytic site in enzymes.

152

153

154 **Protein-protein interaction sites induce weaker long-range evolutionary rate gradients than**
155 **catalytic binding sites**

156 Using relative conservation scores, it was previously shown that protein-protein interactions sites on
157 enzymes induce only a minor conservation gradient which is significantly weaker than that from catalytic
158 sites [2]. Here, we studied residue evolutionary rate (dN/dS) as a function of distance from protein-protein
159 interaction sites in the yeast proteome. We screened the structurally-annotated yeast proteome dataset
160 to find proteins that are known to interact with other proteins (see Methods). We found 459 interfacial
161 sites for 250 proteins. Interfacial residues were identified as those residues that that have different
162 solvent accessibility when in complex compared to when the interacting partner is deleted. Residues were
163 binned according to their distance from interfacial residues, and average dN/dS was calculated for each
164 bin. The slope of the evolutionary rate gradient induced from interfacial sites is indeed significantly smaller
165 than that induced from catalytic binding sites (t-test, $P<0.001$, Figure 5 and Table 4), either on enzymes
166 or on non-enzymes.

167 **Evolutionary rate gradient variations among different functional sites cannot be explained by**
168 **local structural determinants**

169 We examined whether local structural determinants such as solvent exposure gradients are responsible
170 for the observed difference in magnitude of evolutionary rate gradients induced from catalytic and non-
171 catalytic binding sites. We calculated the relative solvent accessibility (RSA) of each residue and then the
172 expected linear trend of dN/dS versus RSA for each protein dataset. We then used the calculated slope
173 and intercept of this dN/dS versus RSA relationship (fig. S1 and table S1 in the Supplementary Material)
174 to plot the expected evolutionary rate at each distance bin according to the average RSA of the residues
175 at that bin. As shown in Figure 6, there is no significant expected increase of evolutionary rate with
176 distance based on solvent exposure alone for any of the datasets of functional binding sites. We therefore

177 conclude that solvent exposure patterns cannot explain the differences in evolutionary rate gradients
178 induced from catalytic and non-catalytic binding sites.

179

## 180 Discussion

181 Strong long-range rate gradient is observed from catalytic sites in enzymes where there is a monotonic,
182 nearly-linear increase in selective pressure on residues with their proximity to the catalytic site. We
183 studied here functional non-catalytic ligand-binding sites on enzymes and non-enzymes to examine
184 whether they induce similar evolutionary rate gradients as enzymatic catalytic sites. For this purpose, we
185 used a structurally-annotated yeast proteome as a model to systematically study site-specific evolutionary
186 rates as a function of distance from functional binding sites. We show that non-catalytic binding sites,
187 even on enzymes, induce significantly weaker evolutionary rate gradients than catalytic sites.

188 Previous studies show that evolutionary rate gradients in enzymes cannot be explained by local structural
189 properties of the protein [2,16], suggesting that the gradient is driven by functional determinants. A recent
190 biophysical model of enzyme evolution that incorporates enzymatic activity constraint along with stability
191 constraints better explains the evolutionary rate gradient from catalytic sites [24]. Enzymatic function is
192 very complex, involving both substrate binding and accelerating the production rate of chemical products
193 by decreasing the free energy barrier of the chemical reaction. By showing that non-catalytic ligand-
194 binding sites induce significantly weaker evolutionary rate gradients than catalytic sites, we can rule out
195 several hypotheses on the origin of the long-range evolutionary rate gradient. The first hypothesis that
196 can be ruled out is that the ligand-binding function of the catalytic site is the main determinant of the
197 long-range rate gradient. While ligand binding is part of the function of the catalytic site, it appears
198 improbable that the specific binding of the substrate is the main determinant of the long-range selection
199 pressure in enzymes.

200 The second hypothesis that can be ruled out is that the long-range evolutionary rate gradient from
201 catalytic sites is mainly driven by the allosteric function of the enzyme. Allosteric function exhibit a long-
202 range effect in which binding of an effector molecule at a distant allosteric site shifts the catalytic site into
203 an active conformation [22,23]. It is reasonable to expect that the need to maintain proper long-range
204 allosteric function can explain the observed long-range evolutionary rate gradient from the catalytic site.
205 On the contrary, we have shown that allosteric binding sites induce a significantly weaker evolutionary
206 rate gradient than catalytic sites. Moreover, a strong evolutionary rate gradient exists even in enzymes
207 which are not known to be allosteric. These results suggest that optimizing the allosteric function of the
208 enzyme is not the main determinant of the observed evolutionary rate gradient.

209 Finally, we show that for all the functional sites we have studied here, there is no significant solvent
210 exposure gradients extending from them. Hence, the observed evolutionary rate gradient variations
211 among different functional sites cannot be explained by known local structural determinants such as
212 solvent exposure gradients.

213 We are therefore left with two plausible hypotheses on the origin of the unique evolutionary rate gradient
214 from catalytic sites. The first hypothesis is that strongly conserved sites in general induce stronger long-
215 range evolutionary rate gradient. Here, the evolutionary rate gradient is a result of a percolating effect of
216 penetrating conservation [18,25,26]. Thus, the higher the selective pressure on a site is, the stronger the
217 percolation of evolutionary rates will be from it. The second hypothesis is that evolutionary rate gradients
218 are driven by the chemistry of the enzyme where the main determinant of the increasing selective
219 pressure is to optimize the catalytic power of the enzyme by specifically stabilizing the reaction transition
220 state [27]. Indeed, a chemistry-centered view on the evolution of new catalytic functions in enzymes
221 suggests that new function in an enzyme can evolve from the ability of the enzyme to stabilize the
222 transition state of the new chemical reaction (rather than just binding of the new ligand) [28–30]. These two
223 hypotheses are both consistent with the observations in this study, and they are not mutually exclusive.

224 Further work is needed to critically examine the validity and applicability of these hypotheses in terms of
225 explaining the origin of the long-range evolutionary rate gradient observed in enzymes.

226

# 227 Methods

228 **Structural annotation of yeast proteins**

229 We based our study on a dataset of structural homologs for the yeast proteome. This dataset was created
230 using gapped BLAST [31] searches between protein subunit sequences with solved structure from the
231 Protein Data Bank [32] and 5,861 translated open reading frames (ORFs) of the yeast *Saccharomyces*
232 *cerevisiae* [33]. We kept those ORF–subunit pairs in which both the subunit sequence and the ORF sequence
233 had coverage of ≥50% in the alignment and E-value $<10^{-5}$ and could be paired with their orthologs in three
234 other closely related yeast species *S. paradoxus*, *S. mikatae*, and *S. Bayanus*. The subunit that produced
235 the lowest alignment E-value with the ORF was chosen as the structural homolog. This way, 1798 yeast
236 ORFs were mapped to a homolog in the PDB. The site-specific structural features of the PDB homolog we

237 *Annotation of ligand-binding sites and catalytic sites*

238 The structural subunits were screened to find those with an identified biologically-relevant ligand-binding
239 site according to the BioLip protein function database [19]. We excluded binding sites where the ligand is an
240 ion. Yeast ORF was classified as an enzyme if its paired structural subunit has a known EC number. Catalytic
241 sites of protein models were identified using the Catalytic Site Atlas [20]. 240 enzymes with 486 ligand-
242 binding sites were identified. We then subdivided the ligand-binding sites on yeast enzymatic proteins
243 into those with an overlap between binding site and catalytic site residues (250 catalytic ligand-binding
244 sites) and those where none of the binding site residues is a catalytic residue (236 non-catalytic ligand-
245 binding sites). 276 proteins with 462 ligand-binding sites are not assigned with an EC number and were
246 therefore termed as 'non-enzymes'. A subset of the dataset which contains only yeast proteins which are

247 known to contain the relevant functional binding sites was created as well and included 59 enzymes with
248 60 catalytic ligand binding sites and 75 non-catalytic ligand binding sites and 76 non-enzymes with 125
249 ligand binding sites. The distance of the Cα atom of each residue in these proteins to the closest Cα of the
250 binding site residues and catalytic site residues was calculated. Residues were binned according these
251 distances up until 30Å into equally spaced bins of 3Å in size. The first bin essentially contained all the
252 ligand-binding residues and only them.

253 *Annotation of allosteric proteins*

254 Using the dataset of proteins with known allosteric function from the allosteric database (ASD) [21], we
255 screened the yeast ORFs by means of their structural subunit to those that are known to have an allosteric
256 function and their allosteric sites residues are known. 196 yeast ORFs matched this criterion, and out of
257 them we kept 190 proteins for which the structural alignment contained the allosteric binding sites.
258 Within the dataset, 108 proteins are identified as enzymes with an EC number and with known catalytic
259 site from the Catalytic Site Atlas (Furnham et al., 2014). Other proteins in our dataset which have an EC
260 number assigned and their catalytic site residues are identified, but whose structural models are not
261 known to have an allosteric function (219 proteins), were classified as 'non-allosteric' enzymes. In addition
262 we created which are known to have allosteric function in ASD. The distance of the Cα atom of each
263 residue to the closest Cα atom of allosteric site residue and closest Cα atom of catalytic site residue was
264 calculated. Residues were binned according these distances up until 30Å into equally spaced bins of 3Å in
265 size and for the subset of 11 proteins into equally spaced bins of 6 Å to have sufficient number of residues
266 in each bin.

267 *Protein interfaces*

268 We screened the protein complexes from which the best ORF-subunit were derived to those that are in
269 physical contact with another ORF-subunit pair (not necessarily the pair with highest similarity) and were
270 reported as interacting with the other ORF-subunit pair by at least one physical experiment in the BioGRID

271 [34,35]. Our dataset contained 250 yeast proteins with 459 interfaces. Interfacial residues were identified as
272 the residues which have different solvent accessibility values when in complex compared to when the
273 interaction partner is manually deleted.

274 All of the optimal ORF-subunit pairs and the corresponding subunit binding site residues are listed in
275 supplementary table S2 in the Supplementary Material and their alignments in supplementary file S1.

276 **dN/dS calculations**

277 We used previous alignments made between the translated ORFs from *S. cerevisiae* and their orthologs
278 in three other closely related species *S. paradoxus*, *S. mikatae*, and *S. Bayanus* [9]. The aligned codons were
279 binned according to their distance from different binding sites as explained above and then concatenated.
280 dN/dS values were calculated over the multiple sequence alignments using the program codeml within
281 the PAML software package [36]. As our four yeast species are closely related, we opted to calculate a single
282 dN/dS value for the entire tree (using model=0), which we specified as ((*S. cerevisiae*, *S. paradoxus*), *S.*
283 *mikatae*, *S. bayanus*). Codon frequencies were assumed equal (CodonFreq=0) and other parameters in
284 codeml were left to their default values. The codon alignments can be found in supplementary files S2 in
285 the Supplementary Material.

286 **Statistical analysis**

287 For each bin, we estimated the standard error in our measurements of dN/dS using 50 rounds of
288 bootstrap resampling. We used a weighted least squares regression to fit dN/dS versus distance where
289 the standard errors of dN/dS in each bin was considered such that distance bins with small dN/dS
290 estimation errors receive greater weight in the line fitting process. We used two-tailed t-tests to compute
291 the significance of the difference between slopes.

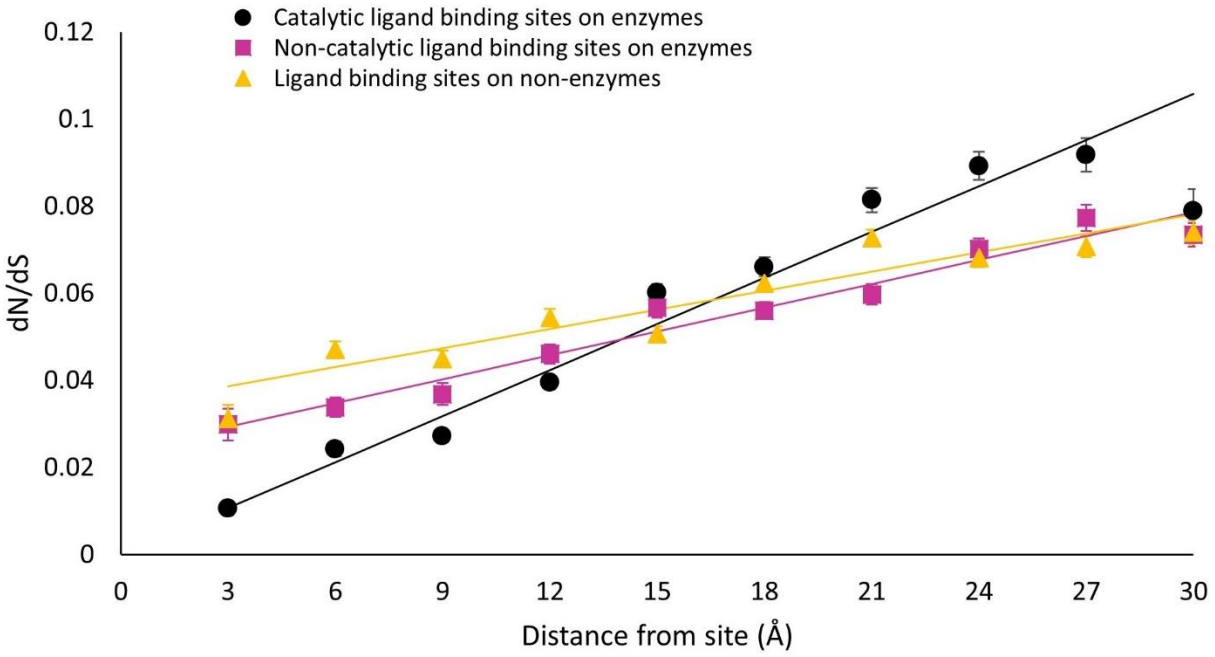
292

## Acknowledgments

This work was supported by Natural Sciences and Engineering Research Council of Canada (grant number RGPIN-2014-03892 to Y.X.), Canada Research Chairs program (to Y.X.), Canada Foundation for Innovation (grant numbers JELF-33732, IF-33122 to Y.X.) and Dalia and Dan Maydan Post-doctoral Fellowship (to A.S.-I.).

314 # Figures

315




316

317 **Figure 1. Non-catalytic ligand binding sites induce weaker evolutionary rate gradients than catalytic**
318 **binding sites.** Evolutionary rate (dN/dS) as a function of distance from catalytic and non-catalytic ligand-
319 binding sites in enzymes, as well as from ligand-binding sites in non-enzymatic proteins.

320

321

322

323

324

325

326

327

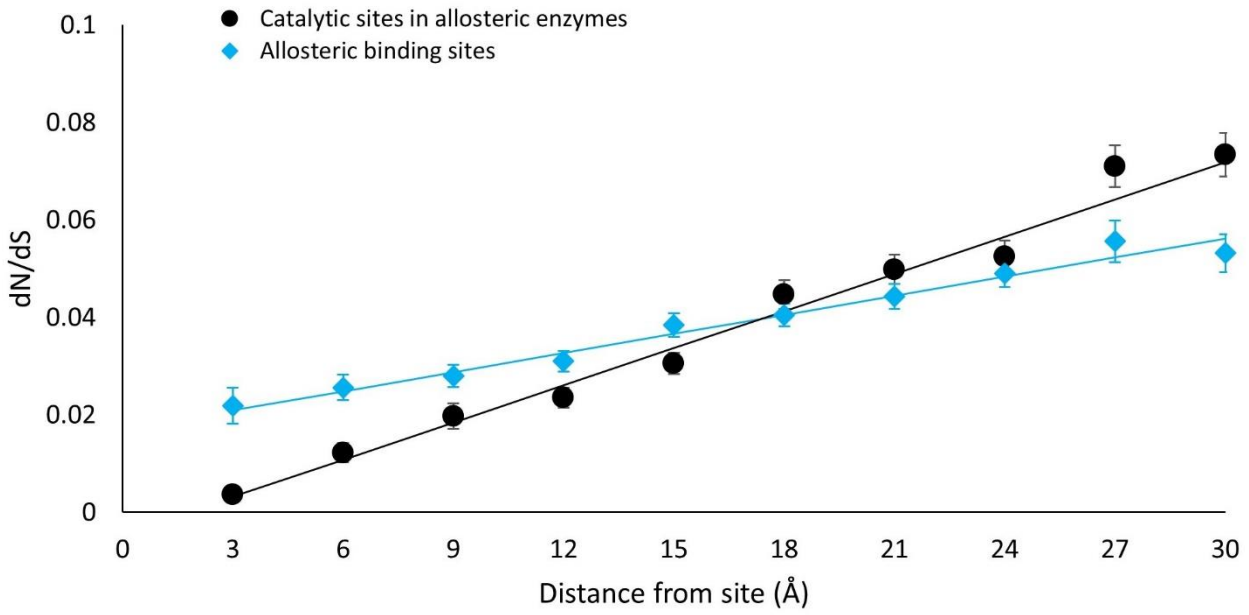
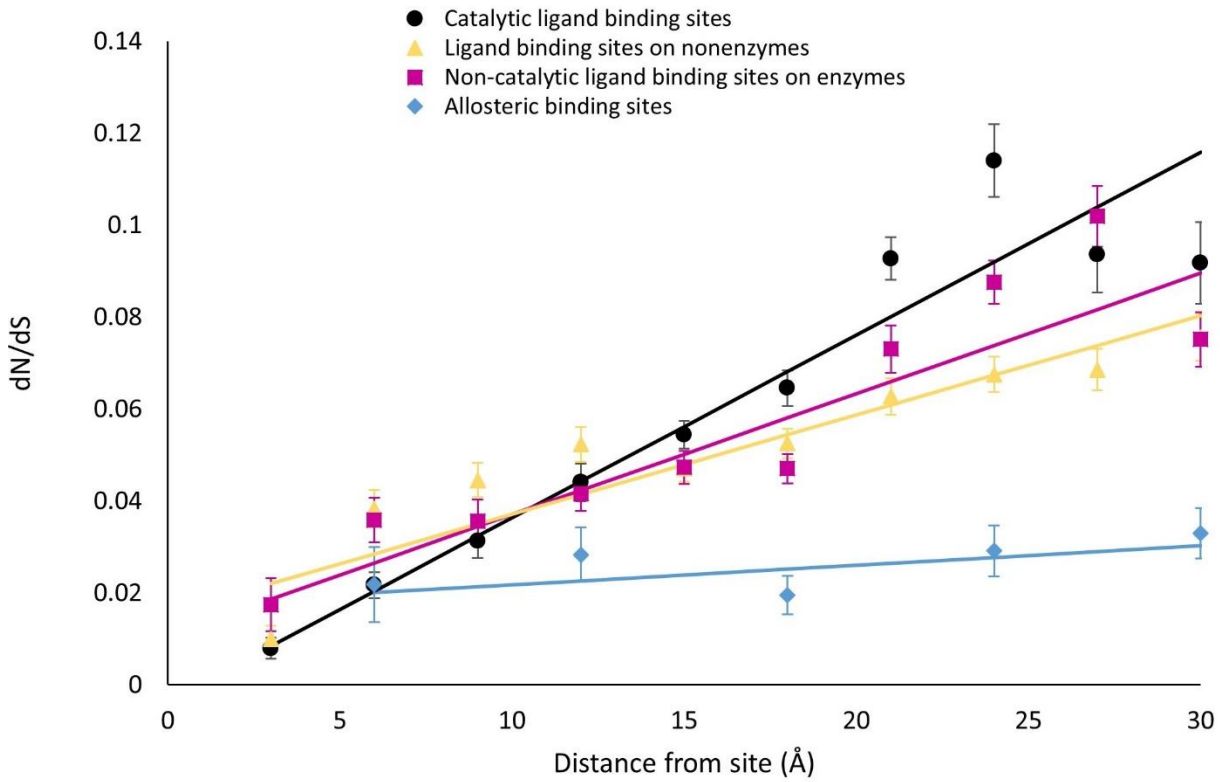



**Figure 2. Allosteric binding sites induce weaker evolutionary rate gradients than catalytic sites.** Evolutionary rate (dN/dS) as a function of distance from allosteric binding sites and from catalytic sites in allosteric enzymes.




338

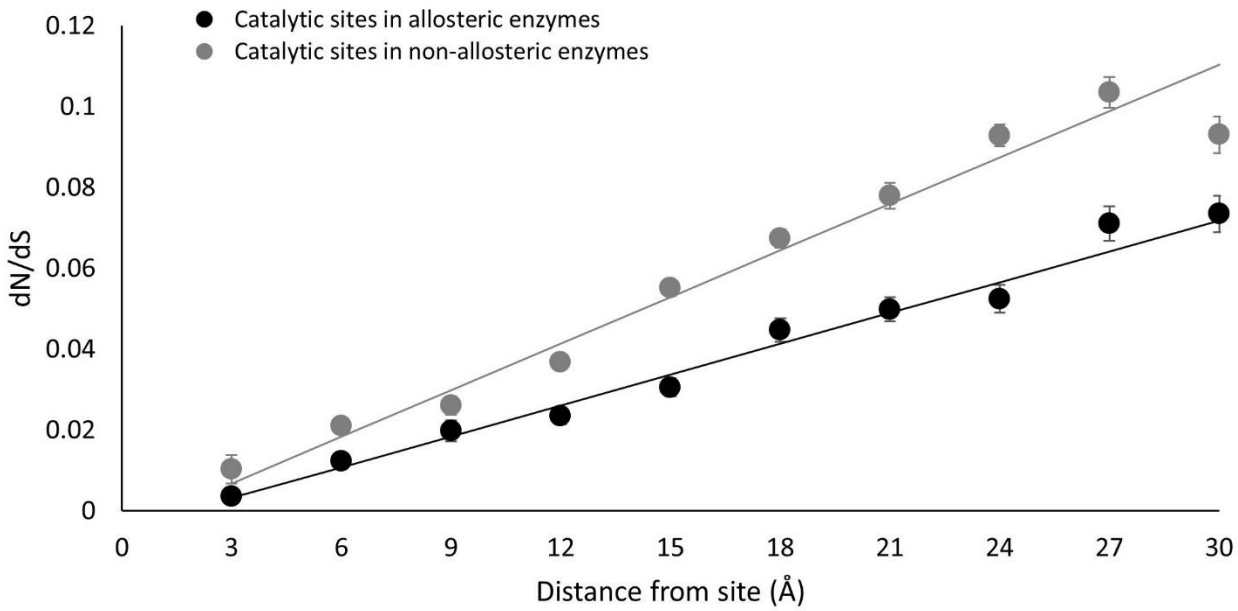
339 **Figure 3**. **Yeast-annotated non-catalytic ligand binding sites induce weaker evolutionary rate gradients**
340 **than yeast-annotated catalytic binding sites.** Evolutionary rate (dN/dS) as a function of distance from
341 catalytic and non-catalytic ligand-binding sites in enzymes as well as from ligand-binding sites in non-
342 enzymatic proteins and allosteric binding sites for the subset of *Saccharomyces cerevisiae* proteins known
343 to contain the respective functional binding sites.

344

345

346

347




348

349 **Figure 4. Catalytic sites in non-allosteric enzymes induce strong evolutionary rate gradients.**
350 Evolutionary rate (dN/dS) as a function of distance from catalytic sites in allosteric and non-allosteric
351 enzymes.

352

353

354

355

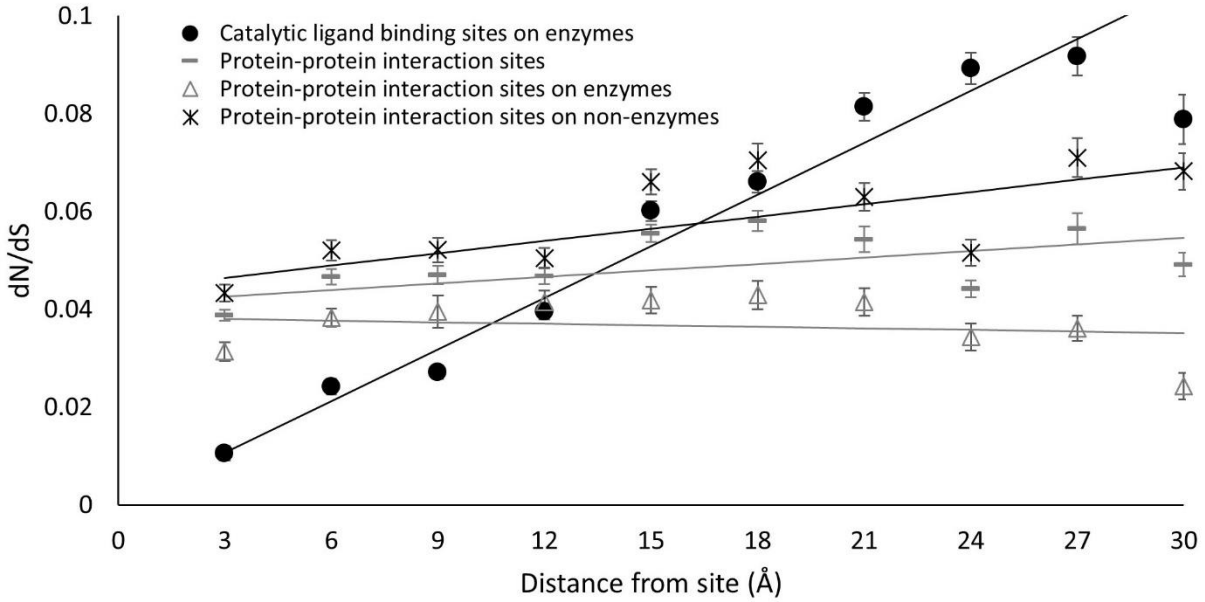



**Figure 5. Protein-protein interaction sites on enzymes or non-enzymes induce weaker long-range evolutionary rate gradients than catalytic binding sites.** Evolutionary rate (dN/dS) as a function of distance from catalytic binding sites and from protein-protein interaction sites on enzymes and non-enzymes.

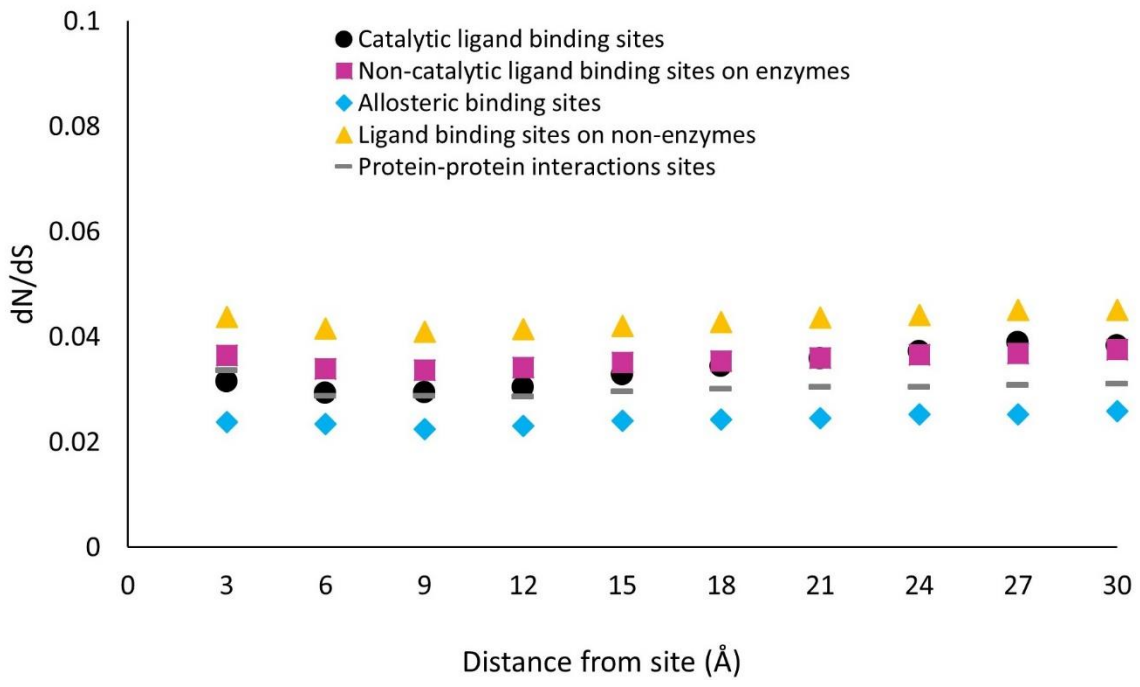



**Figure 6. Evolutionary rate based on solvent exposure alone does not show an increase with distance from catalytic and non-catalytic binding sites.** Expected dN/dS according to average relative solvent accessibility (RSA) of residues as a function of distance from binding site.

380 **Table 1**. **Non-catalytic ligand binding sites induce weaker evolutionary rate gradients than catalytic**
381 **binding sites**. Slope of the linear fit of the average evolutionary rate (dN/dS) versus distance from catalytic
382 sites and non-catalytic ligand-binding sites, as well as the average evolutionary rate of binding site
383 residues.

| | **Type of binding site** | **Slope** | **Evolutionary rate (dN/dS) of binding site residues** |
|---|---|---|---|
| Enzymes | Catalytic ligand-binding sites | 0.0035 (±0.0003) | 0.011 (±0.002) |
| Enzymes | Non-catalytic ligand-binding site | 0.0018 (±0.0001) | 0.030 (±0.004) |
| Non-enzymes | Non-catalytic ligand-binding site | 0.0015 (±0.0002) | 0.0312 (±0.003) |

384 • In brackets – standard errors

385
386
387
388
389
390
391
392
393
394
395
396
397
398
399
400
401
402
403
404
405

406 **Table 2**. **Allosteric binding sites induce weaker evolutionary rate gradients than catalytic sites and**
407 **catalytic sites in non-allosteric enzymes induce strong evolutionary rate gradients**. Slope of the linear fit
408 of the average evolutionary rate (dN/dS) versus distance from allosteric binding sites and catalytic sites in
409 allosteric enzymes, as well as the average evolutionary rate of binding site residues.

| | **Type of functional site** | **Slope** | **Evolutionary rate (dN/dS) of functional site residues** |
|---|---|---|---|
| Allosteric enzymes | Catalytic site | 0.0025 (±0.0001) | 0.004 (±0.002) |
| Allosteric enzymes | Allosteric binding sites | 0.0013 (<0.0001) | 0.022 (±0.004) |
| Non-allosteric enzymes | Catalytic site | 0.0038 (±0.0002) | 0.010 (±0.004) |

410 • In brackets – standard errors

411

412

413

414

415

416

417

418

419

420

421

422

423

424

425

426

427

428

429

430

431 **Table 3**. **Yeast-annotated non-catalytic ligand binding sites induce weaker evolutionary rate gradients**
432 **than yeast-annotated catalytic binding sites.** Slope of the linear fit of the average evolutionary rate
433 (dN/dS) versus distance from catalytic sites and non-catalytic ligand-binding sites, as well as the average
434 evolutionary rate of binding site residues for yeast proteins known to contain functional binding sites.

| Type of binding site | Slope | Evolutionary rate (dN/dS) of binding site residues |
|---|---|---|
| Catalytic binding site | 0.0040 (±0.0003)<br>0.0037(±0.0005)[#] | 0.008 (±0.002) |
| Non-catalytic ligand binding sites on enzymes | 0.0026 (±0.0005) | 0.017 (±0.006) |
| Ligand binding sites on non-enzymes | 0.0022 (±0.0003) | 0.010 (±0.003) |
| Allosteric binding sites | 0.0004 (±0.0003)[#] | 0.022 (±0.008) |

435 • In brackets – standard errors
436 [#] Based on slope calculated over 5 distance bins of 6Å
437
438
439
440
441
442
443
444
445
446
447
448
449
450
451
452
453
454
455

456 **Table 4**. **Protein-protein interaction sites on enzymes or non-enzymes induce weaker long-range**
457 **evolutionary rate gradients than catalytic binding sites.** Slope of the linear fit of the average evolutionary
458 rate (dN/dS) versus distance from protein-protein interaction sites, as well as the average evolutionary
459 rate of interfacial residues.

| Protein-protein interaction sites on: | Slope | Evolutionary rate (dN/dS) of binding site residues |
|---|---|---|
| All proteins | 0.0004 (±<0.0003) | 0.039 (±0.001) |
| Non-enzymes | 0.0008 (±<0.0002) | 0.043 (±0.002) |
| Enzymes | -0.0001 (±<0.0002) | 0.031 (±0.002) |

460 • In brackets – standard errors

461

462

463

464

465

466

467

468

469

470

471

472

473

474

475

476

477

478

479

480

481

482

## References


1. Bartlett, G. J., Porter, C. T., Borkakoti, N. & Thornton, J. M. (2002). Analysis of catalytic residues in enzyme active sites. *J. Mol. Biol.* **324**, 105–121

2. Jack, B. R., Meyer, A. G., Echave, J. & Wilke, C. O. (2016). Functional sites induce long-range evolutionary constraints in enzymes. *PLOS Biol.* **14**, e1002452

3. Franzosa, E. A., Xue, R. & Xia, Y. (2013). Quantitative residue-level structure-evolution relationships in the yeast membrane proteome. *Genome Biol. Evol.* **5**, 734–744

4. Bloom, J. D., Drummond, D. A., Arnold, F. H. & Wilke, C. O. (2006). Structural determinants of the rate of protein evolution in yeast. *Mol. Biol. Evol.* **23**, 1751–1761

5. Echave, J., Spielman, S. J. & Wilke, C. O. (2016). Causes of evolutionary rate variation among protein sites. *Nat. Rev. Genet.* **17**, 109–121

6. Marcos, M. L. & Echave, J. (2015). Too packed to change: side-chain packing and site-specific substitution rates in protein evolution. *PeerJ* **3**, e911

7. Oberai, A., Joh, N. H., Pettit, F. K. & Bowie, J. U. (2009). Structural imperatives impose diverse evolutionary constraints on helical membrane proteins. *Proc. Natl. Acad. Sci. U. S. A.* **106**, 17747–17750

8. Sharir-Ivry, A. & Xia, Y. (2017). The impact of native state switching on protein sequence evolution. *Mol. Biol. Evol.* **34**, 1378–1390

9. Franzosa, E. A. & Xia, Y. (2009). Structural determinants of protein evolution are context-sensitive at the residue level. *Mol. Biol. Evol.* **26**, 2387–2395

10. Overington, J., Donnelly, D., Johnson, M. S., Sali, A. & Blundell, T. L. (1992). Environment-specific amino acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.* **1**, 216–226

11. Conant, G. C. & Stadler, P. F. (2009). Solvent exposure imparts similar selective pressures across a range of yeast proteins. *Mol. Biol. Evol.* **26**, 1155–1161

12. Ramsey, D. C., Scherrer, M. P., Zhou, T. & Wilke, C. O. (2011). The relationship between relative solvent accessibility and evolutionary rate in protein evolution. *Genetics* **188**, 479–488

13. Hegyi, H. & Gerstein, M. (1999). The relationship between protein structure and function: a

comprehensive survey with application to the yeast genome. *J. Mol. Biol.* **288**, 147–164

14. Day, T. J. F., Soudackov, A. V., Čuma, M., Schmitt, U. W. & Voth, G. a. (2002). A second generation multistate empirical valence bond model for proton transport in aqueous systems. *J. Chem. Phys.* **117**, 5839–5849

15. Martin, A. C. *et al.* (1998). Protein folds and functions. *Structure* **6**, 875–884

16. Sharir-Ivry, A. & Xia, Y. (2018). Nature of Long-Range Evolutionary Constraint in Enzymes: Insights from Comparison to Pseudoenzymes with Similar Structures. *Mol. Biol. Evol.* **35**, 2597–2606

17. Nelson, E. D. & Grishin, N. V. (2016). Evolution of off-lattice model proteins under ligand binding constraints. *Phys. Rev. E* **94**, 022410

18. Nelson, E. D. & Grishin, N. V. (2016). Long-Range Epistasis Mediated by Structural Change in a Model of Ligand Binding Proteins. *PLoS One* **11**, e0166739

19. Yang, J., Roy, A. & Zhang, Y. (2013). BioLiP: A semi-manually curated database for biologically relevant ligand-protein interactions. *Nucleic Acids Res.* **41**, D1096-1103

20. Furnham, N. *et al.* (2014). The Catalytic Site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes. *Nucleic Acids Res.* **42**, D485–D489

21. Shen, Q. *et al.* (2016). ASD v3.0: unraveling allosteric regulation with structural mechanisms and biological networks. *Nucleic Acids Res.* **44**, D527–D535

22. Stadtman, E. R. (1966). Allosteric regulation of enzyme activity. *Adv. Enzymol. Relat. Areas Mol. Biol.* **28**, 41–154

23. Liu, J. & Nussinov, R. (2016). Allostery: An overview of its history, concepts, methods, and applications. *PLoS Computational Biology* **12**, e1004966

24. Echave, J. (2019). Beyond Stability Constraints: A Biophysical Model of Enzyme Evolution with Selection on Stability and Activity. *Mol. Biol. Evol.* **36**, 613–620

25. Tóth-Petróczy, A. & Tawfik, D. S. (2011). Slow protein evolutionary rates are dictated by surface-core association. *Proc. Natl. Acad. Sci. U. S. A.* **108**, 11151–6

26. Rajasekaran, N., Suresh, S., Gopi, S., Raman, K. & Naganathan, A. N. (2017). A General Mechanism for the Propagation of Mutational Effects in Proteins. *Biochemistry* **56**, 294–305

27. Warshel, A. *et al.* (2006). Electrostatic basis for enzyme catalysis. *Chem. Rev.* **106**, 3210–3235

28. Gerlt, J. A. & Babbitt, P. C. (2001). Divergent Evolution of Enzymatic Function: Mechanistically Diverse Superfamilies and Functionally Distinct Suprafamilies. *Annu. Rev. Biochem.* **70**, 209–246

29. Khersonsky, O. & Tawfik, D. S. (2010). Enzyme promiscuity: a mechanistic and evolutionary perspective. *Annu. Rev. Biochem.* **79**, 471–505

30. Luo, J., van Loo, B. & Kamerlin, S. C. L. (2012). Catalytic promiscuity in *Pseudomonas aeruginosa* arylsulfatase as an example of chemistry-driven protein evolution. *FEBS Lett.* **586**, 1622–1630

31. Altschul, S. F. *et al.* (1997). Gapped BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic Acids Res.* **25**, 3389–3402

32. Berman, H. M. *et al.* (2000). The Protein Data Bank. *Nucleic Acids Res.* **28**, 235–242

33. Cherry, J. M. *et al.* (1998). SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* **26**, 73–79

34. Stark, C. *et al.* (2006). BioGRID: a general repository for interaction datasets. *Nucleic Acids Res.* **34**, D535-9

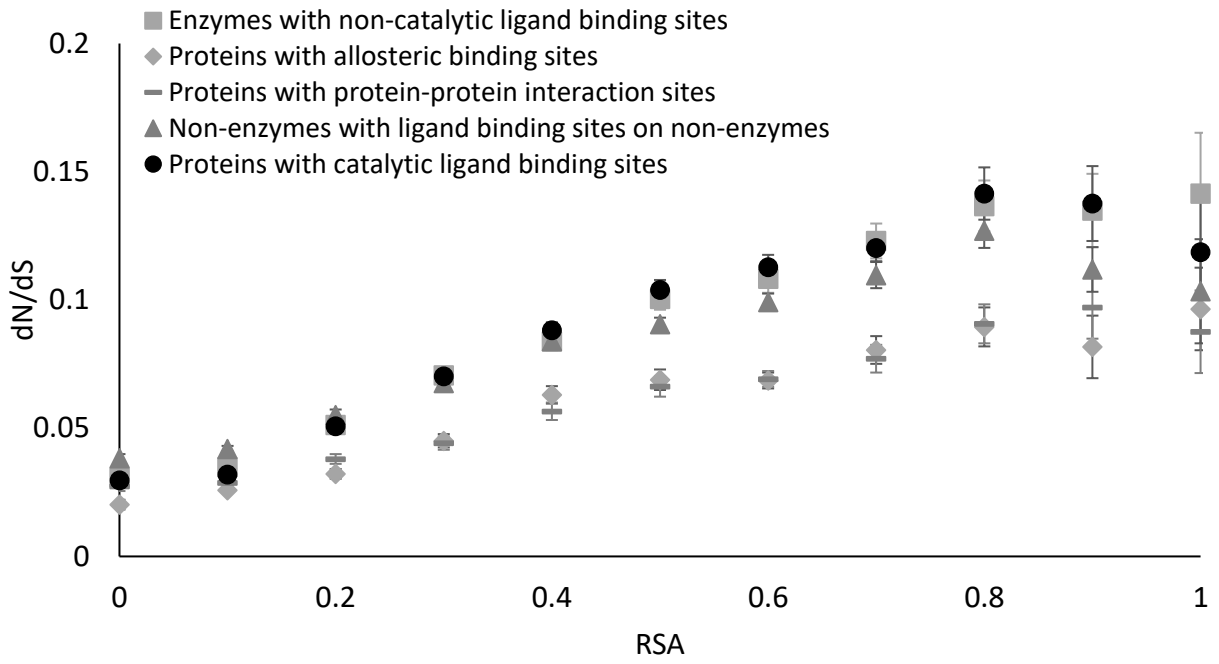
35. Chatr-Aryamontri, A. *et al.* (2017). The BioGRID interaction database: 2017 update. *Nucleic Acids Res.* **45**, D369–D379

36. Yang, Z. (2007). PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* **24**, 1586–1591

564 # Supplementary material

565

566 .




567

568 **Figure S1. Average evolutionary rate (dN/dS) as a function of relative solvent accessibility (RSA).**

569

570 **Table S1. Slope and intercept of the linear fit of the average evolutionary rate (dN/dS) versus**
571 **RSA.**

572

| | Slope | Intercept |
|---|---|---|
| Proteins with catalytic binding sites | 0.145 (±0.011) | 0.023 (±0.003) |
| Enzymes with non-catalytic ligand binding sites | 0.136 (±0.008) | 0.026 (±0.002) |
| Non-enzymes with ligand binding sites | 0.109 (±0.006) | 0.034 (±0.002) |
| Proteins with allosteric binding sites | 0.089 (±0.006) | 0.018 (±0.002) |
| Proteins with protein-protein interaction sites | 0.078 (±0.004) | 0.023 (±0.001) |

573

574

575