Article- Discoveries Section

Quantifying Evolutionary Importance of Protein Sites: A Tale of Two Measures

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

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

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

Abstract

A key challenge in evolutionary biology is the accurate quantification of selective pressure on proteins and other biological macromolecules at single-site resolution. The evolutionary importance of a protein site under purifying selection is typically measured by the degree of conservation of the protein site itself. A possible alternative measure is the strength of the site-induced conservation gradient in the rest of the protein structure. However, the quantitative relationship between these two measures remains unknown. Here, we show that despite major differences, there is a strong linear relationship between the two measures such that more conserved protein sites also induce stronger conservation gradient in the rest of the protein. This linear relationship is universal as it holds for different types of proteins and functional sites in proteins. Our results show that the strong selective pressure acting on the functional site in general percolates through the rest of the protein via residue-residue contacts. Surprisingly however, catalytic sites in enzymes are the principal exception to this rule. Catalytic sites induce significantly stronger conservation gradients in the rest of the protein than expected from the degree of conservation of the site alone. The unique requirement for the active site to selectively stabilize the transition state of the catalyzed chemical reaction imposes additional selective constraints on the rest of the enzyme.

Introduction

The evolutionary importance of protein sites under purifying selection can be quantified in two very different ways. The classical, “intrinsic” measure for the evolutionary importance of a protein site is the degree of conservation or evolutionary rate of the protein site itself. Protein residues experience different degrees of selective pressure as a result of the different roles they play in protein stability and function(Worth et al. 2009; Echave et al. 2016). For example, residues in a protein core are generally under stronger selective pressure than surface residues due to their importance in stabilizing the protein. Indeed, structural determinants such as solvent exposure(Overington et al. 1992; Conant and Stadler 2009; Franzosa and Xia 2009; Franzosa et al. 2013; Sharir-Ivry and Xia 2017) and degree of packing (Hamelryck 2005; Yeh et al. 2014; Marcos and Echave 2015) were shown to explain a large portion of the variability in the observed site-specific evolutionary rates. In addition, residues in functional sites such as catalytic sites(Bartlett et al. 2002) and ligand-binding sites are also under stronger selective pressure than non-functional residues.

An alternative, “extrinsic” measure for the evolutionary importance of a protein site is the conservation gradient the site exerts on the rest of the protein. Rather than quantifying evolutionary conservation of the protein site itself, this measure captures how the evolutionary conservation of other residues surrounding the site gradually decreases with distance from the site in the tertiary structure. Several studies have indicated the possibility for selective pressure to propagate from the functional site to the rest of the protein via physical interactions between neighboring residues in the three-dimensional structure (Tóth-Petróczy and Tawfik 2011; Jack et al. 2016; Nelson and Grishin 2016; Sharir-Ivry and Xia 2019a). While it is clear that the two

69 measures of the evolutionary importance of protein sites are substantially different, it remains
70 unknown how the two measures relate to each other.

71 In this paper, we addressed the fundamental question whether there is a direct relationship
72 between the intrinsic and extrinsic measures of the evolutionary importance of protein sites. We
73 have based our study on a dataset of homology-based structural models of the yeast
74 proteome(Franzosa and Xia 2009; Sharir-Ivry and Xia 2017). Despite their major differences, we
75 show here that there is a strong linear relationship between the degree of conservation of a
76 protein site and the conservation gradient induced from it. In other words, more conserved
77 protein sites tend to induce stronger conservation gradient in the rest of the protein, as selective
78 pressure acting on the protein site percolates via residue-residue interactions. This linear
79 evolutionary conservation-percolation relationship is universal in that it holds for different types
80 of proteins as well as for different types of functional sites in proteins. Remarkably however,
81 catalytic sites in enzymes are the only exception to this universal rule, as catalytic sites induce
82 significantly stronger long-range conservation gradient than other types of functional sites with
83 similar degrees of conservation. We conclude that for many different types of functional sites,
84 site-induced conservation gradient can be explained by the percolation of site-specific selective
85 pressure through the rest of the protein via residue-residue contacts. However, catalytic sites in
86 enzymes induce significantly stronger long-range conservation gradient in the rest of the protein
87 than expected from the percolation theory. This is likely due to the unique requirement for the
88 enzyme active site to selectively bind to and stabilize the transition state of the catalyzed
89 chemical reaction(Warshel et al. 2006).

90 Overall, we show that a more complete understanding of the selective pressure on protein sites
91 can be achieved by integrating the intrinsic measure of site-specific evolutionary conservation
92 with the extrinsic measure of site-induced conservation gradient, with potential implications in
93 protein design, study of disease mutations and better phylogenetic inference(Yang 1996).

94

95

96 # Results

97 ***Evolutionary conservation gradient induced from a protein residue is linearly correlated with***
98 ***the conservation rank of the residue***

99 We based our study on a dataset of homology-based structural models of the *S. cerevisiae*
100 proteome. This basic dataset contains structural templates from the Protein Data Bank
101 (PDB)(Berman et al. 2000) mapped to ORFs of *S. cerevisiae* via sequence alignment (see Methods).
102 We had 1274 yeast ORFs with structural models from the PDB for which residue conservation
103 scores are available in ConSurf-DB (Goldenberg et al. 2009; Celniker et al. 2013). Each residue in
104 the dataset was ranked according to its relative conservation within the protein (the residue's
105 rank of conservation divided by the total number of residues). This normalized conservation rank
106 of a residue ranges from 0 to 1, with higher rank corresponding to lower conservation within the
107 protein. We have also calculated for each residue the Pearson correlation between the
108 conservation scores of all other residues in the protein and their distance from that reference
109 residue. This Pearson correlation between conservation and distance describes the degree of
110 percolation of the evolutionary conservation from a residue throughout the protein tertiary
111 structure, i.e., the 'long-range conservation gradient'. All residues were then binned into 20
112 equally spaced bins of conservation rank and the average conservation gradient for each bin was
113 calculated.

114 A clear negative linear trend is observed between the conservation rank of a residue and the
115 strength of the evolutionary conservation gradient induced from it (Fig 1a). The more a residue
116 is conserved within the protein (lower conservation rank), the stronger the evolutionary

117 conservation gradient it induces. The distribution of overall per-protein Pearson correlations
118 between conservation ranks and conservation gradients is mainly between -0.3 and -0.7, and the
119 average correlation in the dataset is -0.5 (Fig 1b), indicating that the linear relationship between
120 site-specific conservation and site-induced conservation gradient is high for all types of proteins.

121 We have also examined the relationship between site-specific conservation and site-induced
122 conservation gradient specifically for functional sites. We identified different functional sites in
123 the dataset: catalytic sites, non-catalytic ligand-binding sites (on enzymes and on nonenzymatic
124 proteins), protein-protein interaction sites, and allosteric sites. We identified catalytic sites using
125 the Catalytic Site Atlas(Furnham et al. 2014), ligand-binding sites using BioLip(Yang et al. 2013),
126 allosteric sites using the Allosteric Database (ASD)(Huang et al. 2011; Shen et al. 2016) and
127 protein-protein interaction sites using our previous protocol of identifying protein-protein
128 interfaces in the yeast proteome(Franzosa and Xia 2009). Relative solvent accessibility (RSA) of
129 residues was calculated and residues were classified as buried if RSA=0.0, exposed if RSA>0.8 and
130 middle for 0.0<RSA≤0.8. The linear relationship between site-specific conservation rank and site-
131 induced conservation gradient holds regardless of the residue's location within the protein or its
132 functional role (Fig 2). Overall, our results reveal the existence of a conservation–percolation
133 relationship in which higher residue conservation leads to stronger long-range percolation of
134 selective pressure to adjacent sites in the tertiary structure. Furthermore, this relationship holds
135 for residues in different types of functional sites.

136

137

138 ***Conservation gradient induced from a protein site is linearly correlated with the evolutionary***

139 ***rate of the site***

140 The linear trend between conservation and percolation shown above is based on conservation
141 ranks of residues, which are relative within a protein. In these calculations, we have lumped
142 together residues from different proteins having similar conservation rank, however these
143 residues, being from different proteins, can be under different selective pressure. To address this
144 caveat, we examined whether the conservation-percolation linear trend still holds when
145 conservation is measured in terms of absolute evolutionary rate (dN/dS). The evolutionary rate
146 is calculated for the yeast proteins in *S. cerevisiae* compared with its orthologs in four closely
147 related yeast species (see Methods). Conservation score annotations are transferred from
148 structural homologs onto yeast proteins using sequence alignment. Then, we binned all residues
149 into 100 equally spaced bins of conservation rank and calculated their average evolutionary rate
150 (dN/dS). The average evolutionary rate increases for residues with increasing conservation rank
151 (decreasing conservation) (S1 Fig), showing high correlation between the evolutionary rate of
152 yeast protein sites and the conservation scores of their structural homologs. The conservation-
153 percolation linear trend is shown to still hold here, where conservation is measured in terms of
154 absolute evolutionary rate (Fig 3). The lower the average evolutionary rate of a protein site is,
155 the stronger percolation of evolutionary conservation is induced from it.

156

157

158

159 ***Catalytic sites induce stronger conservation gradients than predicted by the conservation-***
160 ***percolation trend***

161 The annotation of functional and structural site residues was transferred to yeast proteins from
162 their structural models using sequence alignment. The average conservation gradient induced
163 from them is shown in Fig 3. As expected, functional residues evolve more slowly (low dN/dS)
164 than other residues. The average conservation gradients induced from ligand binding sites and
165 protein-protein interaction sites can all be predicted reasonably well from the linear
166 conservation-percolation trend. Protein-protein interaction sites induce slightly weaker
167 conservation gradient than expected, possibly due to their tendency to rewire during evolution.
168 Exposed non-functional residues induce slightly stronger conservation gradient than expected,
169 possibly due to hidden functional sites on protein surfaces. Remarkably, catalytic sites are the
170 only functional sites with the most significant deviation from the linear conservation-percolation
171 trend. The average conservation gradient induced from catalytic site residues is significantly
172 stronger than expected by the linear conservation-percolation trend. This result suggests that
173 the strong long-range conservation gradient induced from catalytic residues cannot be solely
174 attributed to the percolation of their low evolutionary rate.

175 ***Catalytic site residues induce stronger conservation gradients than non-catalytic functional site***
176 ***residues with similar evolutionary rates***

177 We have shown that on average, catalytic site residues induce stronger conservation gradients
178 than expected by their average evolutionary rate. Next, we further reinforce this result by
179 comparing the conservation gradient induced from subsets of catalytic and non-catalytic site

180 residues with similar evolutionary rates. We have sampled 1,000 random subsets of residues
181 from each type of functional site. For each such subset (represented by a circle in Fig 4), we
182 calculated the average evolutionary rate and average conservation gradient. As expected,
183 catalytic site residue subsets span the lowest dN/dS values (x-axis), followed by non-catalytic
184 ligand binding sites and allosteric sites, protein-protein interaction sites, ligand binding sites in
185 non-enzymes and finally buried residues. The linear trend between evolutionary rate and
186 conservation gradient holds for each of these functional site types (Fig 4). This result indicates
187 the robustness of the conservation-percolation linear trend regardless of the functional or
188 structural role of the residues. Interestingly, residue subsets from different functional sites with
189 similar evolutionary rates induce conservation gradients with different magnitudes. In particular,
190 catalytic sites induce significantly stronger long-range conservation gradients than all other non-
191 catalytic sites with similar evolutionary rates, including non-catalytic ligand binding sites (Fig 4).
192 Finally, protein-protein interaction site residues induce lower conservation gradients than most
193 other functional site residues with similar evolutionary rates, possibly due to the tendency for
194 protein-protein interactions to rewire during evolution.

195 ***Catalytic site residues often induce stronger conservation gradients than more conserved non-***
196 ***catalytic functional site residues within the same protein***

197 We have shown that conservation gradients induced by functional site residues in general
198 correlate linearly with the evolutionary rates of these functional site residues. We have also
199 shown that catalytic site residues are special in that they induce the strongest conservation
200 gradients among subsets of different functional site residues with similar average evolutionary

201 rates. However, our analyses have so far been carried out by grouping together functional site
202 residues from different proteins. In this section, we perform a stringent, per-protein analysis by
203 focusing on multi-functional proteins with at least two distinct functional sites and comparing
204 the evolutionary properties of different functional site residues within the same protein. Some
205 proteins in our dataset contain both catalytic sites as well as non-catalytic functional sites,
206 whereas other proteins in our dataset contain two distinct non-catalytic functional sites. We
207 found that for the majority of the functional site residue pairs within the same multi-functional
208 protein, the more conserved functional site residue indeed induces stronger conservation
209 gradient (86% if the more conserved residue is catalytic, Fig 5A; 67% if both residues are non-
210 catalytic, Fig 5B). These results agree with the hypothesis that induced conservation gradients
211 are largely driven by the percolation of selective pressure acting on functional sites.

212 Remarkably, in cases where the catalytic site residue is less conserved than the non-catalytic
213 functional site residue within the same multi-functional protein, the catalytic site residue still
214 induces stronger conservation gradient than the non-catalytic functional site for many of these
215 cases (discordance of 49%, Fig 5C). This large discordance shown in cases where the lower
216 conservation residue is a catalytic site residue, is significantly higher compared to the cases
217 where the lower conservation residue is a non-catalytic site residue (binomial test $P<<0.01$).
218 These results clearly show that within the same protein, less conserved catalytic site residues
219 often induce stronger conservation gradient than the more conserved non-catalytic site.
220 Therefore, the strong conservation gradients from catalytic sites cannot be entirely explained by
221 the percolation of the strong selective pressure acting on the catalytic sites.

Discussion

In this paper we have shown a linear relationship between two measures of evolutionary importance of protein sites under purifying selection. These are the degree of evolutionary conservation of the site itself, as well as the percolation of evolutionary conservation induced from the site via neighboring residues in the protein tertiary structure. Despite major differences between these two measures, we have shown that the linear relationship between the two measures holds for different types of proteins as well as for different types of functional sites in proteins. However, catalytic sites in enzymes are the principal exception to this rule. We have shown here that catalytic sites in enzymes induce significantly stronger conservation gradients in the rest of the protein than expected from the degree of conservation of the site alone. Catalytic sites have a unique and complex functionality as they both bind a substrate as well as reduce the free energy barrier required for a chemical reaction to occur. These catalytic sites were shown to be under stronger selective pressure compared with other functional sites such as protein-protein interaction sites and ligand binding sites(Jack et al. 2016; Sharir-Ivry and Xia 2019a). It was also shown that they induce a significantly stronger long-range evolutionary rate gradient than other functional sites. One hypothesis regarding the origin of the strong conservation gradient from catalytic sites is that it is simply due to the strong selective pressure acting on these sites percolating through the rest of the protein via residue-residue contacts. However, we have shown here that the strong selective pressure acting on catalytic sites cannot entirely explain the strong conservation gradients induced from catalytic sites.

The main determinant of the long-range conservation gradient from catalytic sites is still not completely understood. Local structural constraints (such as residue burial and packing) are

usually potential contributors as they are known to generally have a significant effect on residue evolutionary rate(Franzosa and Xia 2009; Worth et al. 2009; Marcos and Echave 2015; Echave et al. 2016). However, it was shown that local structural constraints are not the main determinants of long-range conservation gradients in enzymes(Jack et al. 2016; Sharir-Ivry and Xia 2018; Sharir-Ivry and Xia 2019b). Moreover, the fact that non-catalytic ligand binding sites and allosteric binding sites induce significantly weaker evolutionary rate gradients implies that the ligand binding and allosteric function are not the main determinants of the long-range conservation gradients in enzymes either(Sharir-Ivry and Xia 2019a). We are therefore left with the hypothesis that the uniquely strong conservation gradient in enzymes is imposed by the special requirement for catalytic sites to differentially bind the transition state of a chemical reaction rather than the reactants or products with very similar properties (Warshel et al. 2006), a function which is unique to catalytic sites compared with other functional binding sites. A recent physical model of residue evolutionary rates in enzymes introduced an activation term in addition to a stability term and showed improved predictive ability(Echave 2019). The model attributed the activity to the free energy required to transform a distorted catalytic site upon mutation to its native conformation. The improved ability of the model supports the hypothesis that the main determinant of the observed long-range conservation gradient in enzymes is a functional rather than structural constraint. Overall, our results suggest that the stringent requirement for the catalytic site to differentially bind to and stabilize the transition state of the catalyzed chemical reaction imposes extensive evolutionary constraints on a large portion of the enzyme beyond just the catalytic site, all of which play key roles in maintaining the catalytic function.

The results in this paper provide a complete, quantitative picture of evolutionary conservation patterns within proteins induced by functional sites. The linear relationship between the degree of conservation of a protein functional site and the induced conservation gradient in the rest of the protein suggests that the strong selective pressure acting on the functional site percolates through the rest of the protein via residue-residue contacts. Our results also clearly show the unique evolutionary behavior of enzymes in which the catalytic site induces significantly stronger evolutionary constraints on their surroundings than can be explained by the percolation theory alone. Finally, our proposed measure of the evolutionary importance of protein sites can assist in better phylogenetic inference, functional site prediction and protein design.

277 # Methods

278 ***Protein dataset collection and functional site annotations***

279 The current dataset is based on a dataset of structural homologs of yeast proteins(Sharir-Ivry and
280 Xia 2019a). The dataset was created first by using gapped BLAST(Altschul et al. 1997) searches
281 between protein subunit sequences with solved structure from the Protein Data Bank(Berman et
282 al. 2000) and 5,861 translated open reading frames (ORFs) of the yeast *Saccharomyces*
283 *cerevisiae*(Cherry et al. 1998). The ORF–subunit pairs were chosen such that both the subunit
284 sequence and the ORF sequence had coverage of ≥50% in the alignment and E-value $<10^{-5}$ and
285 could be paired with their orthologs in four other closely-related yeast species *S. paradoxus*, *S.*
286 *mikatae*, *S. bayanus and S. pombe*. This way, 1,555 yeast ORFs were mapped to homologs in the
287 PDB. The procedure included the following steps:

288 First, if one of the homology-based structural models of a yeast ORF had an annotated allosteric
289 site, this model was chosen. For 115 yeast proteins the structural model was identified with a
290 known allosteric site as well as pre-calculated conservation scores in ConSurf-DB(Goldenberg et
291 al. 2009). For all other yeast ORFs, if they had structural models with known ligand binding sites
292 that do not overlap with catalytic sites and the closest catalytic site residue is at least 6Å away
293 (non-catalytic ligand binding site), the model with the lowest E-value out of them was chosen.
294 Overall, 54 non-enzymes with 57 ligand binding sites and 41 enzymes with 47 non-catalytic
295 ligand-binding sites were part of the dataset. For all other yeast ORFs, the structural model with
296 the lowest E-value was chosen. In this manner, 1064 more ORFs for which the best structural
297 model had pre-calculated conservation scores in ConSurf-DB were added to the dataset. Overall,
298 1,274 ORF-subunit pairs were included in the study. Out of them, 148 yeast proteins were

299 identified with 280 protein-protein interactions sites and 348 proteins were identified with
300 catalytic sites. Full list of yeast proteins, their structural models along with identified functional
301 sites can be found in Supplementary Table S1.

302 ***Functional site annotations***

303 Allosteric sites within the structural subunits were found and their residues annotated using the
304 Allosteric Database (ASD)(Huang et al. 2011; Shen et al. 2016). Biologically-relevant ligand-
305 binding sites were found using Binding MOAD (Mother of all Databases)(Hu et al. 2005; Ahmed
306 et al. 2015), a database of biologically significant protein-ligand binding in the PDB. Using MOAD,
307 sites with bound crystallographic additives, buffers, salts, metals and sites with covalently linked
308 ligands are excluded. Ligand-binding residues were identified using the BioLip(Yang et al. 2013)
309 database. Catalytic sites within the protein structural subunits were found using the Catalytic Site
310 Atlas(Furnham et al. 2014). In order to find proteins that participate in protein-protein
311 interactions in our dataset, we identified structural subunits where each subunit is both in
312 physical contact with another subunit and the corresponding modelled ORFs are reported as
313 interacting by at least one physical experiment in the BioGRID (Stark et al. 2006; Chatr-
314 Aryamontri et al. 2017). Our dataset contains 148 proteins with 280 protein-protein interaction
315 interfaces. Interfacial residues were identified as residues with different solvent accessibility
316 values when in complex compared to when the interacting partner is manually deleted from the
317 tertiary structure. Distances between residues were calculated as distances between their
318 respective Cα atoms. All functional site annotations of the chosen structural subunits were
319 transferred to the yeast ORF sequence according to the sequence alignment.

320

321 ***Evolutionary conservation and rate calculations***

322 In this study we calculated both evolutionary conservation scores for the residues of the
323 structural subunits as well as average absolute evolutionary rates (dN/dS) for the yeast ORFs.
324 Evolutionary conservation scores were taken from ConSurf-DB(Goldenberg et al. 2009), which is
325 a database of pre-calculated conservation scores of residues in proteins with known structures
326 in the Protein Data Bank (PDB). ConSurf-DB conservation scores are based on collected sequence
327 homologs of the PDB structure and using the Rate4Site algorithm(Pupko et al. 2002).

328 To calculate the average evolutionary rates (dN/dS) for residues of *S. cerevisiae*, we first used the
329 orthology assignment of the protein-coding genes of *S. cerevisiae* with four other closely-related
330 yeast species (*S. paradoxus*, *S. mikatae*, *S. bayanus, and S. pombe*), according to the Fungal
331 Orthogroup Repository(Wapinski et al. 2007). We then aligned the ORFs using MAFFT(Katoh and
332 Standley 2013). Then, evolutionary rates were calculated using the program codeml within the
333 PAML software package(Yang 2007). The tree was specified as ((((*S. cerevisiae*, *S. paradoxus*), *S.*
334 *mikatae*), *S. bayanus*), *S. pombe*). Codon frequencies were assumed equal (CodonFreq=0) and
335 other parameters in codeml were left to their default values. The codon alignments can be found
336 in Supplementary S2 File.

337 ***Statistical analysis***

338 1000 random sample of 250 residues were collected from each type of functional site residues
339 for Fig 4.

340 Estimated standard errors in our measurements of conservation gradients (Pearson correlations)
341 and of dN/dS values were done using 50 rounds of bootstrap resampling.

Acknowledgments

This work was supported by Natural Sciences and Engineering Research Council of Canada (grant numbers RGPIN-2019-05952, RGPAS-2019-00012), Canada Foundation for Innovation (grant number JELF-33732), and Canada Research Chairs program.

Figure Legends

Figure 1. **Long-range conservation gradient induced from a protein residue is linearly correlated with its conservation rank within the protein**. (a) Average long-range conservation gradient from a residue as a function of conservation rank for all residues in the dataset binned into 20 equally spaced bins of conservation rank. (b) Distribution of per-protein Pearson correlation between residues' conservation ranks and conservation gradients.

Figure 2. Residues in functional and structural sites exhibit linear correlation between conservation and conservation gradient. Conservation gradient of residues as a function of their conservation rank within a protein along with the corresponding Pearson correlation for residues in different types of structural and functional sites.

Figure 3. **Conservation gradient induced from residues is linearly (negatively) correlated with their evolutionary rate (dN/dS), with catalytic site residues inducing stronger conservation gradient than expected by the linear trend**. Average conservation gradient (calculated as the average Pearson correlation between conservation of residues and their distance from a site) as a function of average evolutionary rate (dN/dS) for all yeast protein residues binned according to their annotated conservation rank into 100 equally spaced bins.

Figure 4. **Catalytic site residues induce stronger conservation gradients than non-catalytic functional site residues with similar evolutionary rates.** Average conservation gradients of subsets of functional residues. Each circle represents a subset of residues, colored by different types of functional sites.

Figure 5. Catalytic site residues often induce stronger conservation gradients than more conserved non-catalytic functional site residues within the same protein. Within the same protein, (A) more conserved catalytic site residues tend to induce stronger conservation gradient than less conserved non-catalytic site residues (binomial test, $P<<0.001$); (B) more conserved non-catalytic site residues tend to induce stronger conservation gradient than less conserved non-catalytic site residues (binomial test, $P<<0.001$); (C) less conserved catalytic site residues often induce stronger conservation gradient than more conserved non-catalytic site residues (discordance for half of the cases, binomial test, $P=0.13$). Functional site residue pairs for which the ordering of residue conservation agrees with the ordering of induced conservation gradient (concordance) are marked in blue. Functional site residue pairs for which the ordering of residue conservation disagrees with the ordering of induced conservation gradient (discordance) are marked in orange.

393 # References

394 Ahmed A, Smith RD, Clark JJ, Dunbar JB, Carlson HA. 2015. Recent improvements to Binding
395 MOAD: a resource for protein–ligand binding affinities and structures. *Nucleic Acids Res.*
396 43:D465–D469.

397 Altschul SF, Madden TL, Schäffer AA, Zhang J, Zhang Z, Miller W, Lipman DJ. 1997. Gapped
398 BLAST and PSI-BLAST: a new generation of protein database search programs. *Nucleic*
399 *Acids Res.* 25:3389–3402.

400 Bartlett GJ, Porter CT, Borkakoti N, Thornton JM. 2002. Analysis of catalytic residues in enzyme
401 active sites. *J. Mol. Biol.* 324:105–121.

402 Berman HM, Westbrook J, Feng Z, Gilliland G, Bhat TN, Weissig H, Shindyalov IN, Bourne PE.
403 2000. The Protein Data Bank. *Nucleic Acids Res.* 28:235–242.

404 Celniker G, Nimrod G, Ashkenazy H, Glaser F, Martz E, Mayrose I, Pupko T, Ben-Tal N. 2013.
405 ConSurf: Using Evolutionary Data to Raise Testable Hypotheses about Protein Function. *Isr.*
406 *J. Chem.* 53:199–206.

407 Chatr-Aryamontri A, Oughtred R, Boucher L, Rust J, Chang C, Kolas NK, O'Donnell L, Oster S,
408 Theesfeld C, Sellam A, et al. 2017. The BioGRID interaction database: 2017 update. *Nucleic*
409 *Acids Res.* 45:D369–D379.

410 Cherry JM, Adler C, Ball C, Chervitz SA, Dwight SS, Hester ET, Jia Y, Juvik G, Roe T, Schroeder M,
411 et al. 1998. SGD: Saccharomyces Genome Database. *Nucleic Acids Res.* 26:73–79.

412 Conant GC, Stadler PF. 2009. Solvent exposure imparts similar selective pressures across a
413 range of yeast proteins. *Mol. Biol. Evol.* 26:1155–1161.

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

416 Echave J, Spielman SJ, Wilke CO. 2016. Causes of evolutionary rate variation among protein
417 sites. *Nat. Rev. Genet.* 17:109–121.

418 Franzosa EA, Xia Y. 2009. Structural determinants of protein evolution are context-sensitive at
419 the residue level. *Mol. Biol. Evol.* 26:2387–2395.

420 Franzosa EA, Xue R, Xia Y. 2013. Quantitative residue-level structure-evolution relationships in
421 the yeast membrane proteome. *Genome Biol. Evol.* 5:734–744.

422 Furnham N, Holliday GL, De Beer TAP, Jacobsen JOB, Pearson WR, Thornton JM. 2014. The
423 Catalytic Site Atlas 2.0: Cataloging catalytic sites and residues identified in enzymes.
424 *Nucleic Acids Res.* 42:D485–D489.

425 Goldenberg O, Erez E, Nimrod G, Ben-Tal N. 2009. The ConSurf-DB: pre-calculated evolutionary
426 conservation profiles of protein structures. *Nucleic Acids Res.* 37:D323-7.

427 Hamelryck T. 2005. An amino acid has two sides: A new 2D measure provides a different view
428 of solvent exposure. *Proteins* 59:38–48.

429 Hu L, Benson ML, Smith RD, Lerner MG, Carlson HA. 2005. Binding MOAD (Mother Of All
430 Databases). *Proteins Struct. Funct. Bioinforma.* 60:333–340.

431 Huang Z, Zhu L, Cao Y, Wu G, Liu X, Chen Y, Wang Q, Shi T, Zhao Y, Wang Y, et al. 2011. ASD: a
432 comprehensive database of allosteric proteins and modulators. *Nucleic Acids Res.*
433 39:D663–D669.

434 Jack BR, Meyer AG, Echave J, Wilke CO. 2016. Functional sites induce long-range evolutionary
435 constraints in enzymes. *PLOS Biol.* 14:e1002452.

436 Katoh K, Standley DM. 2013. MAFFT multiple sequence alignment software version 7:
437 Improvements in performance and usability. *Mol. Biol. Evol.* 30:772–780.

438 Marcos ML, Echave J. 2015. Too packed to change: side-chain packing and site-specific
439 substitution rates in protein evolution. *PeerJ* 3:e911.

440 Nelson ED, Grishin N V. 2016. Evolution of off-lattice model proteins under ligand binding
441 constraints. *Phys. Rev. E* 94:022410.

442 Overington J, Donnelly D, Johnson MS, Sali A, Blundell TL. 1992. Environment-specific amino
443 acid substitution tables: tertiary templates and prediction of protein folds. *Protein Sci.*
444 1:216–226.

445 Pupko T, Bell RE, Mayrose I, Glaser F, Ben-Tal N. 2002. Rate4Site: an algorithmic tool for the
446 identification of functional regions in proteins by surface mapping of evolutionary
447 determinants within their homologues. *Bioinformatics* 18:S71–S77.

448 Sharir-Ivry A, Xia Y. 2017. The impact of native state switching on protein sequence evolution.
449 *Mol. Biol. Evol.* 34:1378–1390.

450 Sharir-Ivry A, Xia Y. 2018. Nature of Long-Range Evolutionary Constraint in Enzymes: Insights
451 from Comparison to Pseudoenzymes with Similar Structures. *Mol. Biol. Evol.* 35:2597–
452 2606.

453 Sharir-Ivry A, Xia Y. 2019a. Non-catalytic Binding Sites Induce Weaker Long-Range Evolutionary
454 Rate Gradients than Catalytic Sites in Enzymes. *J. Mol. Biol.* 431:3860–3870.

455 Sharir-Ivry A, Xia Y. 2019b. Using Pseudoenzymes to Probe Evolutionary Design Principles of
456 Enzymes. *Evol. Bioinforma.* 15:117693431985593.

457 Shen Q, Wang G, Li S, Liu X, Lu S, Chen Z, Song K, Yan J, Geng L, Huang Z, et al. 2016. ASD v3.0:
458 unraveling allosteric regulation with structural mechanisms and biological networks.
459 *Nucleic Acids Res.* 44:D527–D535.

460 Stark C, Breitkreutz B-J, Reguly T, Boucher L, Breitkreutz A, Tyers M. 2006. BioGRID: a general
461 repository for interaction datasets. *Nucleic Acids Res.* 34:D535-9.

462 Tóth-Petróczy A, Tawfik DS. 2011. Slow protein evolutionary rates are dictated by surface-core

463 association. *Proc. Natl. Acad. Sci. U. S. A.* 108:11151–11156.

464 Wapinski I, Pfeffer A, Friedman N, Regev A. 2007. Natural history and evolutionary principles of
465 gene duplication in fungi. *Nature* 449:54–61.

466 Warshel A, Sharma PK, Kato M, Xiang Y, Liu H, Olsson MHM. 2006. Electrostatic basis for
467 enzyme catalysis. *Chem. Rev.* 106:3210–3235.

468 Worth CL, Gong S, Blundell TL. 2009. Structural and functional constraints in the evolution of
469 protein families. *Nat. Rev. Mol. Cell Biol.* 10:709–720.

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

472 Yang Z. 1996. Among-site rate variation and its impact on phylogenetic analyses. *Trends Ecol.*
473 *Evol.* 11:367–372.

474 Yang Z. 2007. PAML 4: phylogenetic analysis by maximum likelihood. *Mol. Biol. Evol.* 24:1586–
475 1591.

476 Yeh SW, Huang TT, Liu JW, Yu SH, Shih CH, Hwang JK, Echave J. 2014. Local packing density is
477 the main structural determinant of the rate of protein sequence evolution at site level.
478 *Biomed Res. Int.*:572409.

479

480

481

482

483

484

485

486

Figures

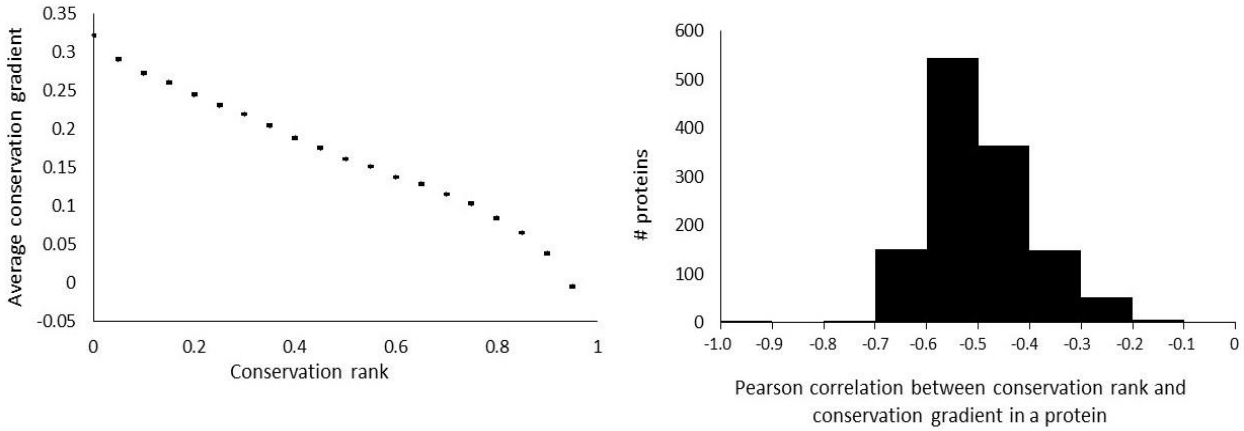


Figure 1. **Long-range conservation gradient induced from a protein residue is linearly correlated with its conservation rank within the protein**. (a) Average long-range conservation gradient from a residue as a function of conservation rank for all residues in the dataset binned into 20 equally spaced bins of conservation rank. (b) Distribution of per-protein Pearson correlation between residues' conservation ranks and conservation gradients.

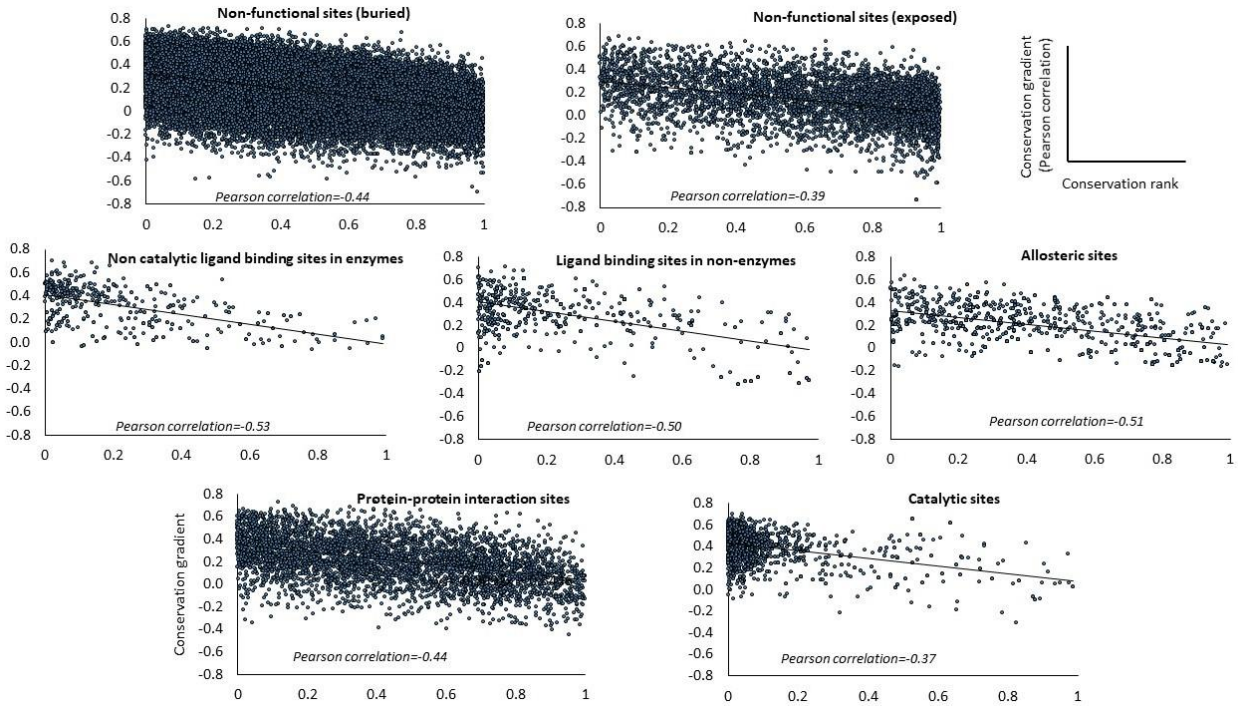


Fig 2. Residues in functional and structural sites exhibit linear correlation between conservation and conservation gradient. Conservation gradient of residues as a function of their conservation rank within a protein along with the corresponding Pearson correlation for residues in different types of structural and functional sites.

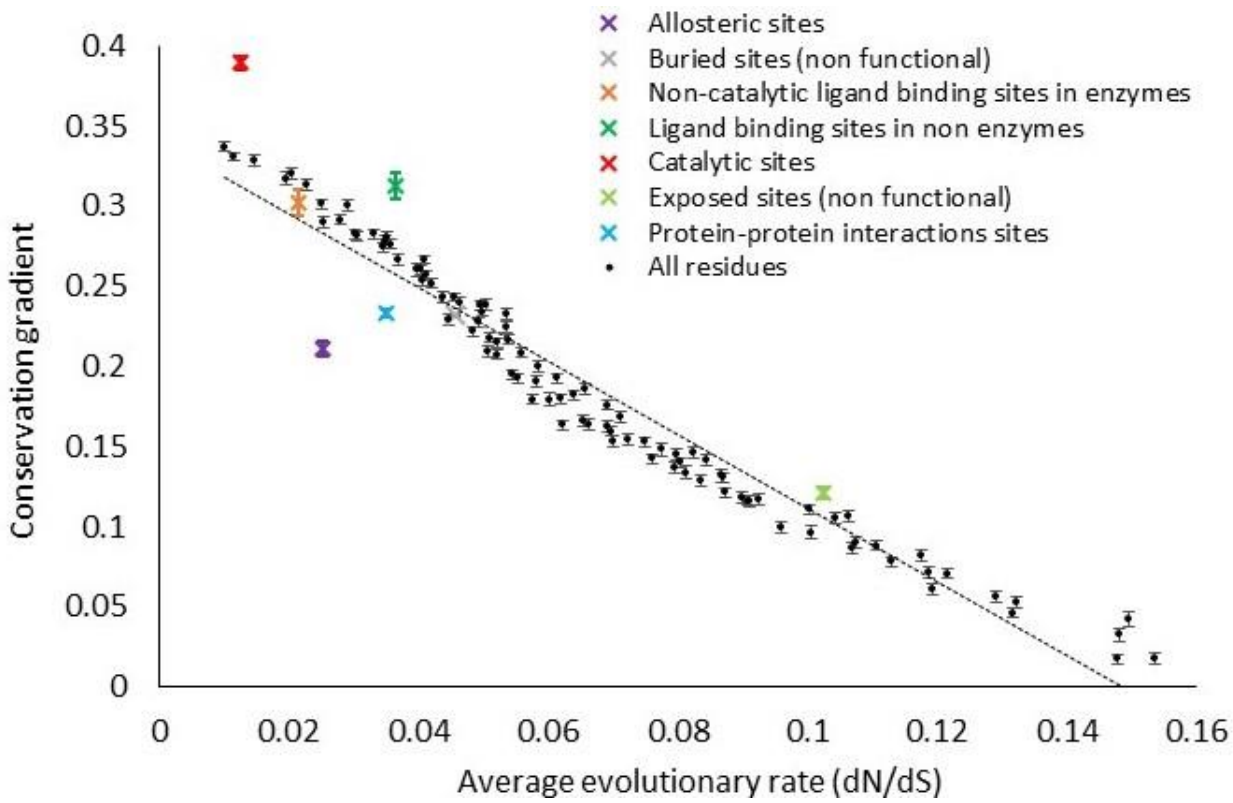
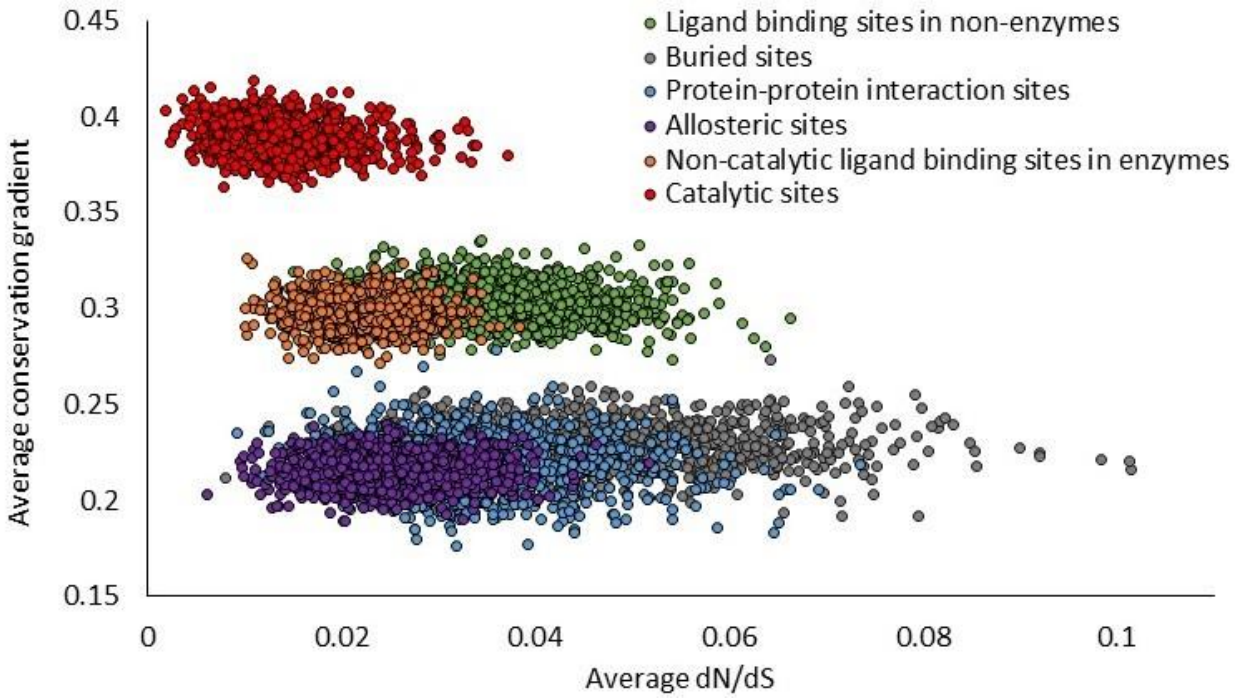


Fig 3. **Conservation gradient induced from residues is linearly (negatively) correlated with their evolutionary rate (dN/dS), with catalytic site residues inducing stronger conservation gradient than expected by the linear trend**. Average conservation gradient (calculated as the average Pearson correlation between conservation of residues and their distance from a site) as a function of average evolutionary rate (dN/dS) for all yeast protein residues binned according to their annotated conservation rank into 100 equally spaced bins.



518

519

520 **Figure 4**. **Catalytic site residues induce stronger conservation gradients than non-catalytic functional**
521 **site residues with similar evolutionary rates.** Average conservation gradients of subsets of functional
522 residues. Each circle represents a subset of residues, colored by different types of functional sites.

523

524

525

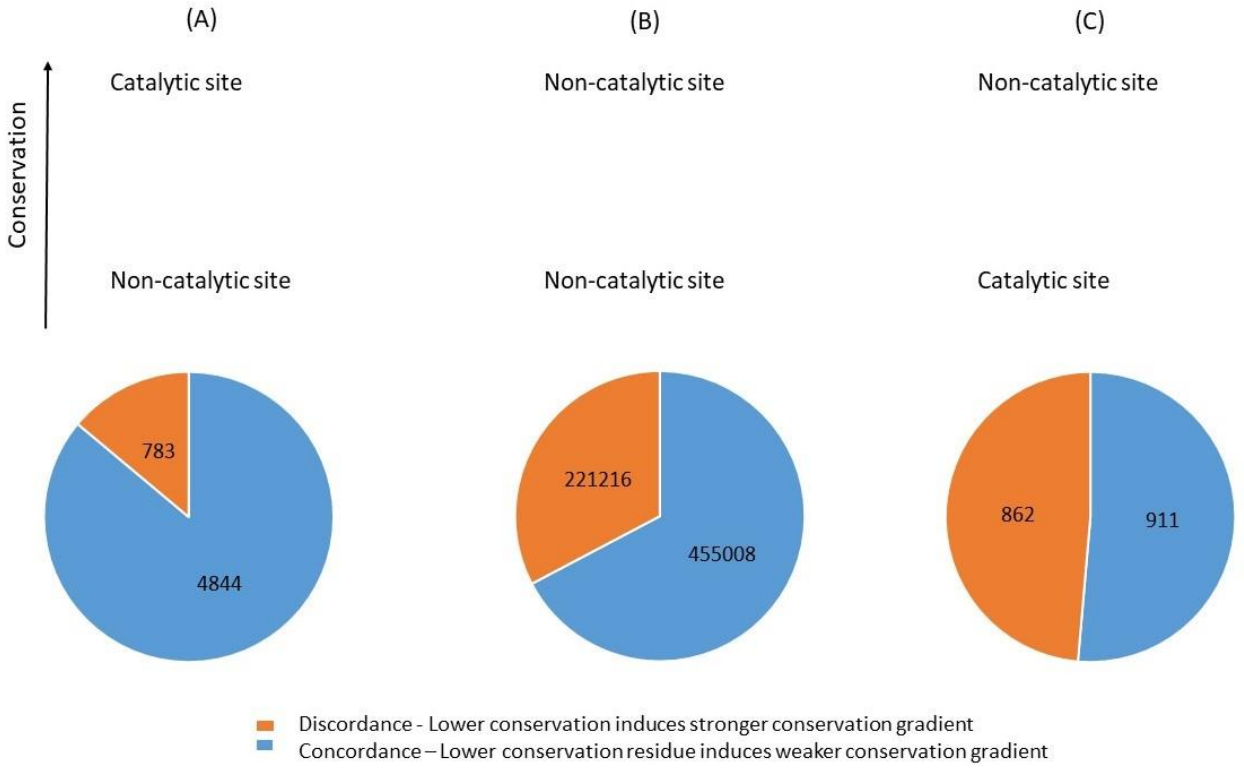
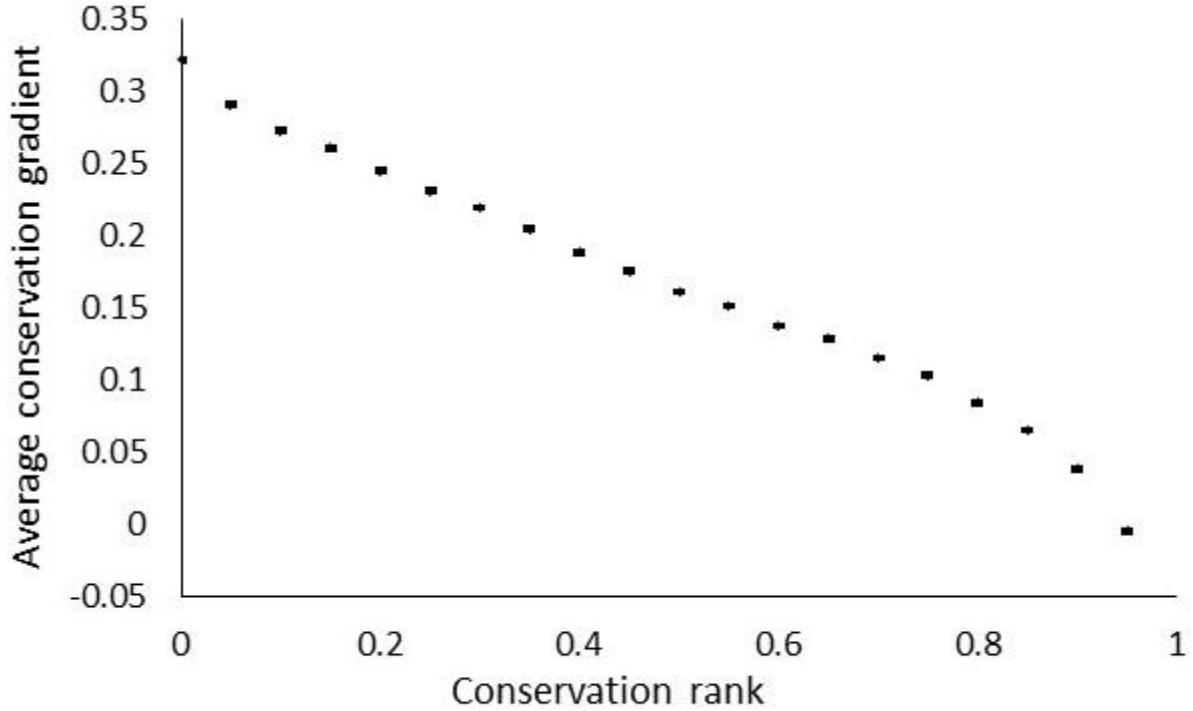
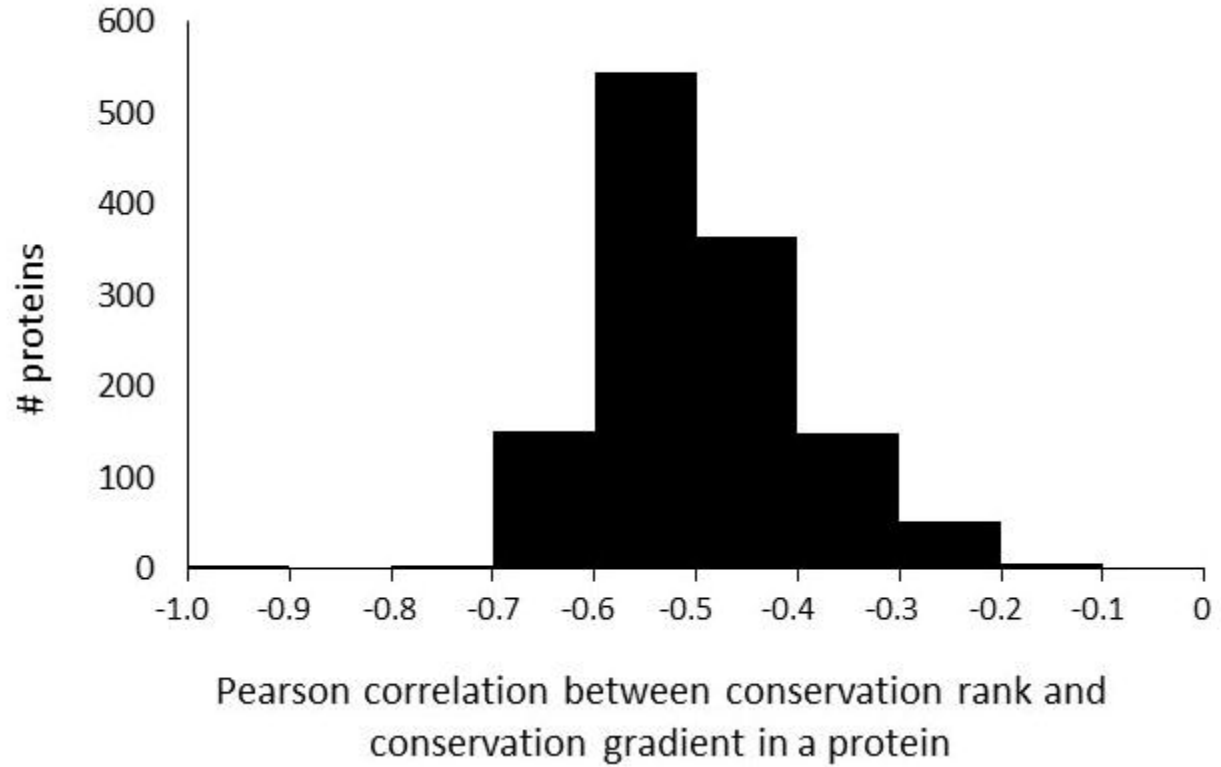
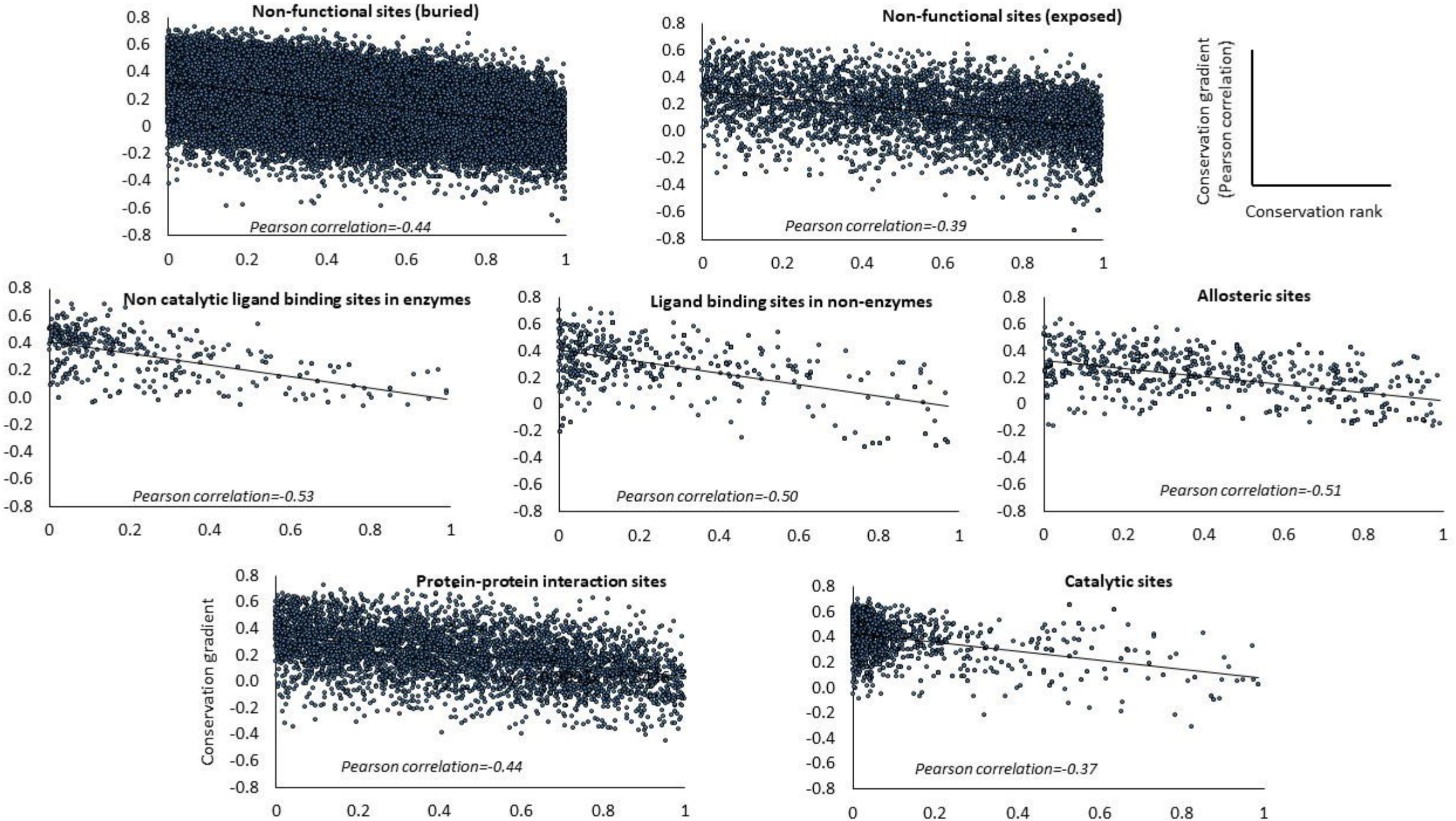
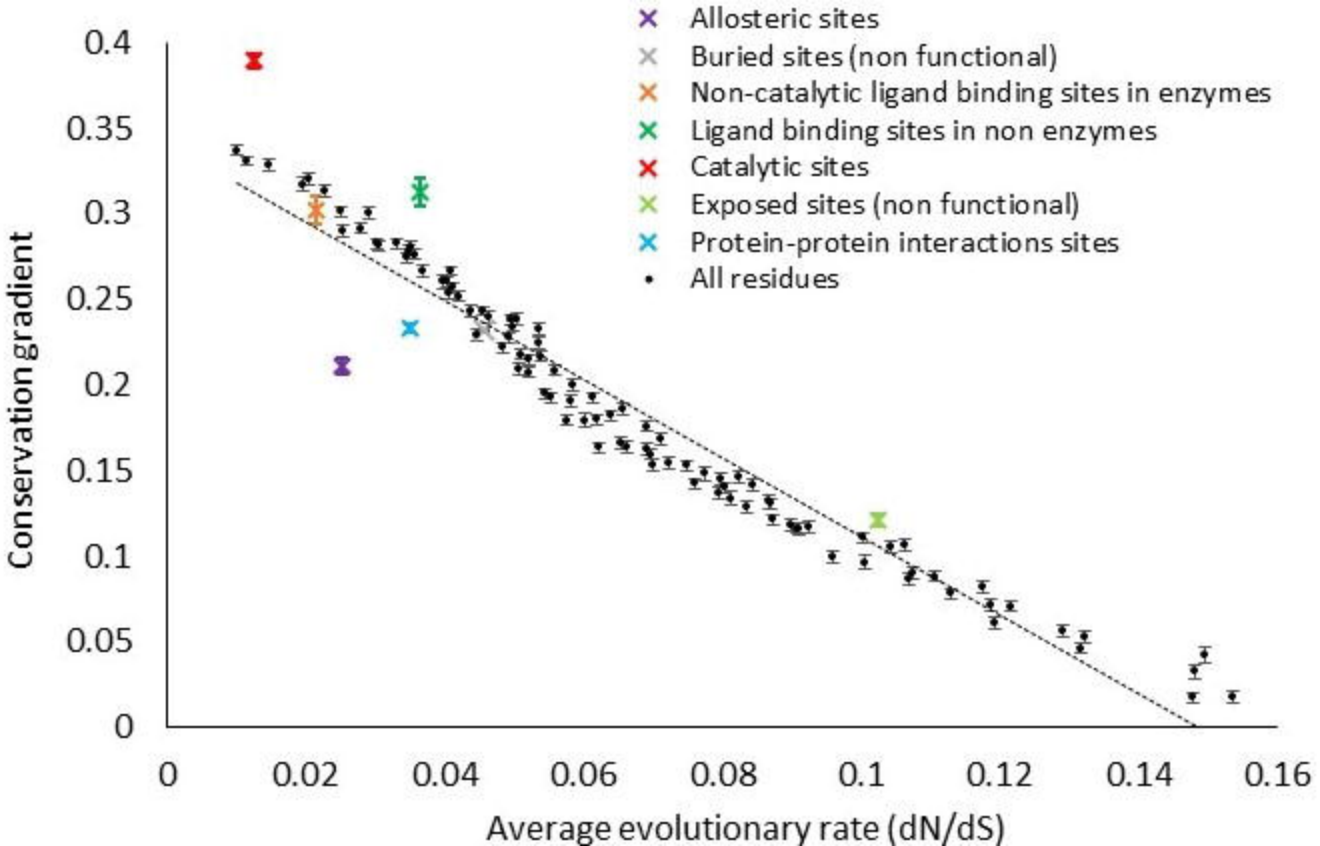


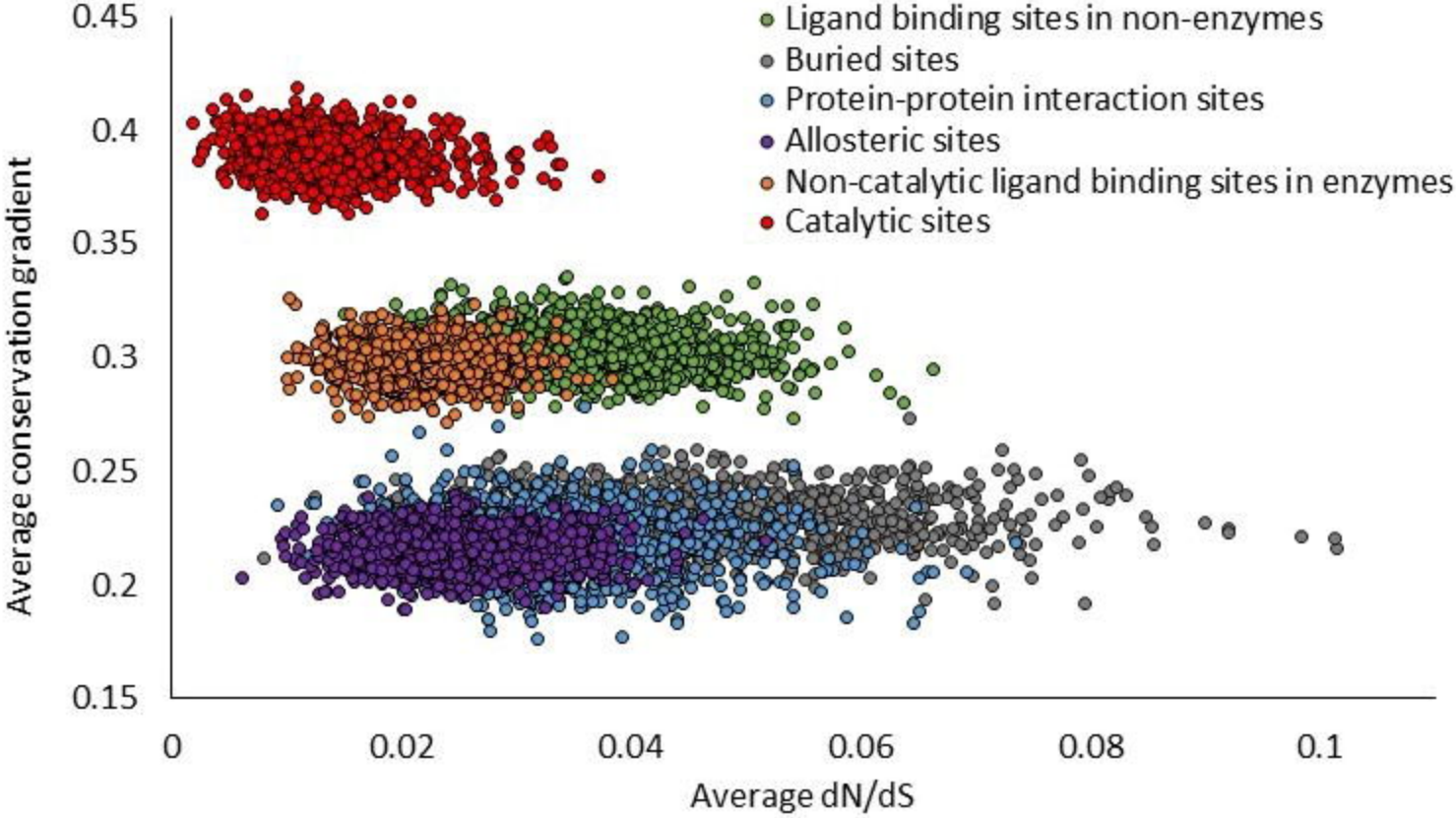
Fig 5. Catalytic site residues often induce stronger conservation gradients than more conserved non-catalytic functional site residues within the same protein. Within the same protein, (A) more conserved catalytic site residues tend to induce stronger conservation gradient than less conserved non-catalytic site residues (binomial test, P<<0.001); (B) more conserved non-catalytic site residues tend to induce stronger conservation gradient than less conserved non-catalytic site residues (binomial test, P <<0.001); (C) less conserved catalytic site residues often induce stronger conservation gradient than more conserved non-catalytic site residues (discordance for half of the cases, binomial test, P=0.13). Functional site residue pairs for which the ordering of residue conservation agrees with the ordering of induced conservation gradient (concordance) are marked in blue. Functional site residue pairs for which the ordering of residue conservation disagrees with the ordering of induced conservation gradient (discordance) are marked in orange.

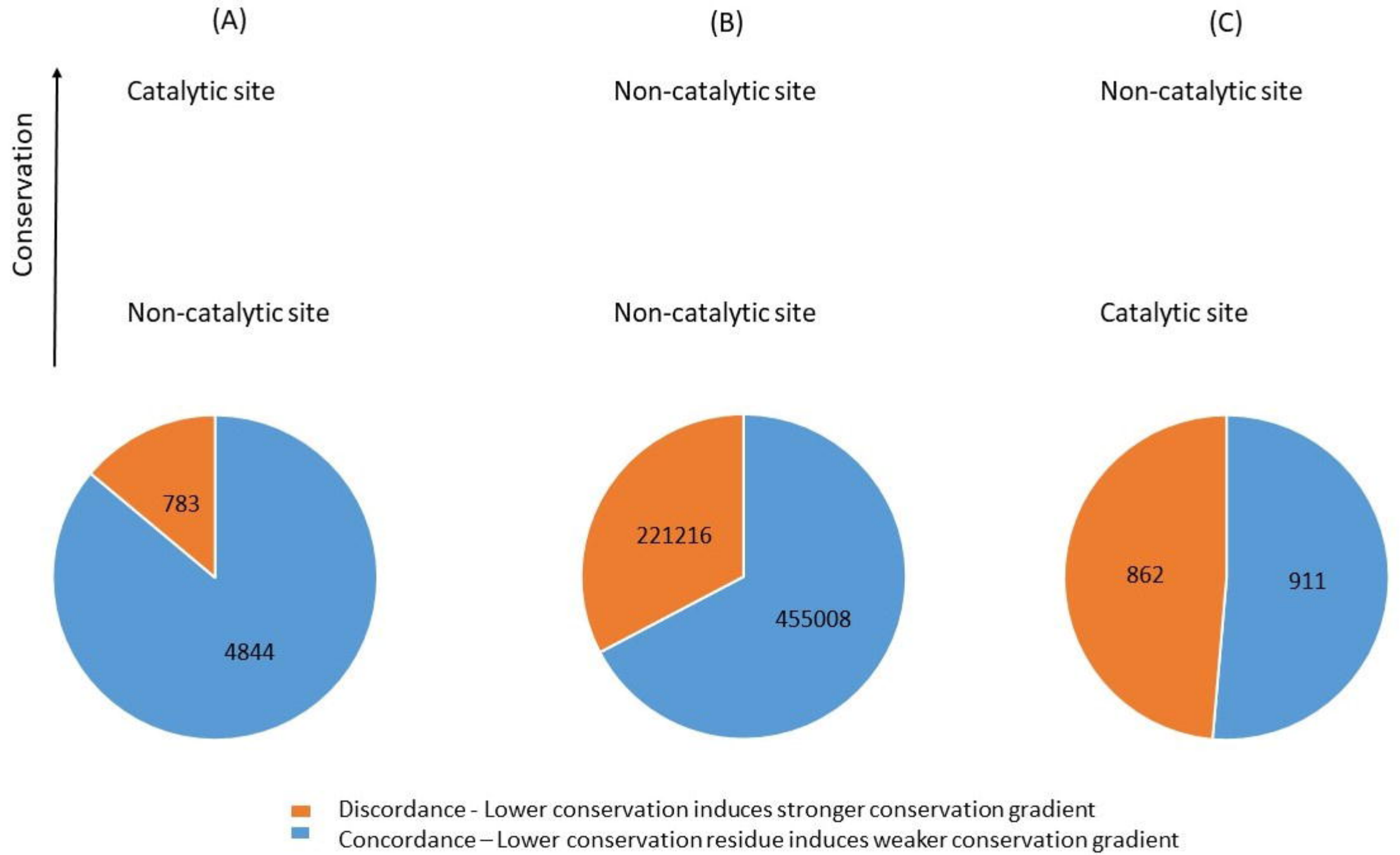

Average conservation gradient
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
-0.05
0
0.2
0.4
0.6
0.8
1
Conservation rank


proteins
600
500
400
300
200
100
0
-1.0
-0.9
-0.8
-0.7
-0.6
-0.5
-0.4
-0.3
-0.2
-0.1
0
Pearson correlation between conservation rank and conservation gradient in a protein


Non-functional sites (buried)
Pearson correlation=-0.44
Non-functional sites (exposed)
Pearson correlation=-0.39
Conservation gradient (Pearson correlation)
Conservation rank
Non catalytic ligand binding sites in enzymes
Pearson correlation=-0.53
Ligand binding sites in non-enzymes
Pearson correlation=-0.50
Allosteric sites
Pearson correlation=-0.51
Protein-protein interaction sites
Pearson correlation=-0.44
Catalytic sites
Pearson correlation=-0.37
Conservation gradient


Allosteric sites
Buried sites (non functional)
Non-catalytic ligand binding sites in enzymes
Ligand binding sites in non enzymes
Catalytic sites
Exposed sites (non functional)
Protein-protein interactions sites
All residues
Conservation gradient
0.4
0.35
0.3
0.25
0.2
0.15
0.1
0.05
0
0
0.02
0.04
0.06
0.08
0.1
0.12
0.14
0.16
Average evolutionary rate (dN/dS)


Ligand binding sites in non-enzymes
Buried sites
Protein-protein interaction sites
Allosteric sites
Non-catalytic ligand binding sites in enzymes
Catalytic sites
Average conservation gradient
0.45
0.4
0.35
0.3
0.25
0.2
0.15
0
0.02
0.04
0.06
0.08
0.1
Average dN/dS


(A)
(B)
(C)
Conservation
Catalytic site
Non-catalytic site
Non-catalytic site
Non-catalytic site
Non-catalytic site
Catalytic site
783
4844
221216
455008
862
911
Discordance - Lower conservation induces stronger conservation gradient
Concordance – Lower conservation residue induces weaker conservation gradient