Mechanisms of protein evolution

Running Title: On the evolution of proteins

Vijay Jayaraman[#,1], Saacnicteh Toledo-Patiño[#,2], Lianet Noda-García[*,3] and Paola Laurino[*,2]

[1]Department of Molecular Cell Biology, Weizmann Institute of Science, Rehovot, Israel, 7610001

[2]Protein Engineering and Evolution Unit, Okinawa Institute of Science and Technology Graduate University, 1919-1 Tancha, Onna 904-0495 Okinawa, Japan

[3]Department of Plant Pathology and Microbiology, Institute of Environmental Sciences, Robert H. Smith Faculty of Agriculture, Food and Environment, Hebrew University of Jerusalem, Rehovot, Israel, 7610001

[#] These authors contributed equally

* Corresponding authors: paola.laurino@oist.jp, lianet.noda@mail.huji.ac.il

Abstract

How do proteins evolve? How do changes in sequence mediate changes in protein structure, and in turn in function? This question has multiple angles, ranging from biochemistry and biophysics to evolutionary biology. This review provides a brief integrated view of some key mechanistic aspects of protein evolution. First, we explain how protein evolution is primarily driven by randomly acquired genetic mutations and selection for function, and how these mutations can even give rise to completely new folds. Then, we also comment on how phenotypic protein variability, including promiscuity, transcriptional and translational errors, may also accelerate this process, possibly *via* 'plasticity-first' mechanisms. Finally, we highlight open questions in the field of protein evolution, with respect to the emergence of more sophisticated protein systems such as protein complexes, pathways and the emergence of pre-LUCA enzymes.

Introduction

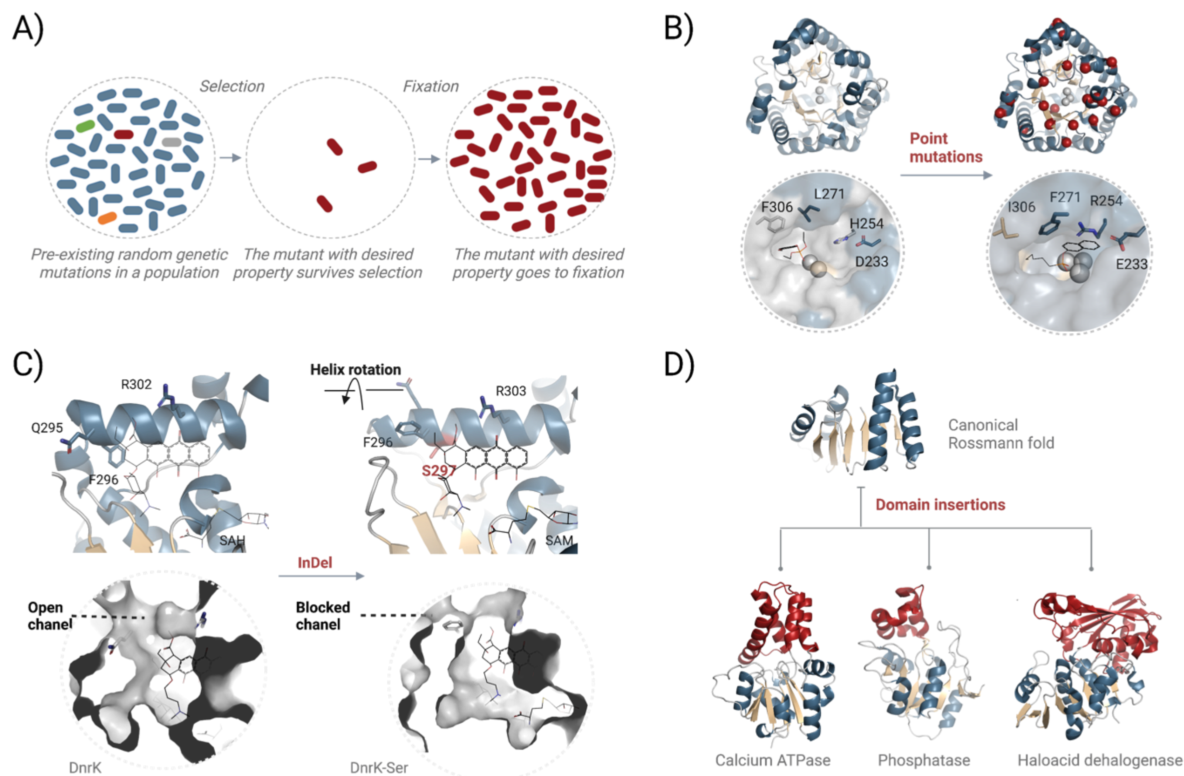
The first version of this manuscript was written by Paola Laurino, Lianet Noda-García and Prof. Dan S. Tawfik, whom the authors deeply miss.

Protein evolution encompasses a large variety of phenomena addressed by multiple disciplines including biophysics, biochemistry and evolutionary biology. The mechanistic aspects of protein evolution may be broadly phrased as: how do changes in protein sequence occur, and how do they mediate changes in protein structure, and in turn in function? Each discipline has its own angle with respect to these questions. Here we present an integrated view, through the eyes of protein scientists. We attempted to portray how multi-faceted the research of protein evolution is and discuss relatively unexplored aspects and fundamental questions that remain unanswered. However, breadth inevitably trades off with depth. Thus, we apologise if significant achievements of specific fields are not thoroughly cited.

A fundamental paradox in protein evolution is that: 'nothing evolves unless it already exists', or in other words as stated by DeVries: 'Natural selection may explain the survival of the fittest, but it cannot explain the arrival of the fittest' [1]. Mutations, insertions/deletions, and recombination mostly induce minor changes in protein structure (*micro-transitions*) that are sufficient for the rise of new functions, although in rare cases, these can generate completely new protein folds (*macro-transitions*) (**Box 1**, *1*). Our review revolves around this classic, 'Darwinian model', and covers cases where the pre-existing sequence diversity in a population give rise to new functions.

Further, we describe various mechanisms that may expedite this process. For instance, it is possible that the genomic mutations needed conferring a novel function might not be present in a population, they can however, rise by non-genetic mechanisms mediated by errors in replication, transcription and translation (phenotypic mutations)[2–4]. Thus, the upcoming new function is already present, as fortuitous, latent variation at the phenotypic level within identical genotypes (phenotypic variability) [5,6]. These changes are observed at all biological levels of organization, from single proteins to entire organisms [7,8]. Indeed, the pre-existence of protein activities as latent promiscuous functions, is by now, a well-established hypothesis understood in atomic detail [9–11]. We also highlight additional aspects of phenotypic variability that underlie the arrival of the fitter. A seemingly attractive, yet controversial hypothesis, is that phenotypic variability (and possibly also genetic changes) is directly induced by environmental challenges.

66 These so-called ‘Baldwin-effects’[12,13] may apply to protein evolution, and are presented here
67 under a general model, coined ‘plasticity-first’.
68 Much of the current work revolves around the evolution of individual biochemical
69 activities such as ligand binding (DNA, RNA, small molecules or proteins) or enzymatic
70 functions (for recent examples see [14–23]). However, beyond biochemical activity *per se*, other
71 protein features are also shaped by evolution, such as the regulation of the protein expression,
72 folding, stability and oligomerization [24–27] or avoiding undesired interactions with other
73 metabolites or proteins [28]. Further, proteins also coevolve with other proteins and biomolecules
74 with whom they interact, and with the cellular components responsible for protein synthesis,
75 maintenance, and clearance [29,30]. Here, we discuss some open questions related to these aspects.
76 As proteins have been evolving for ca. 3.7 billion years, the mechanisms underlying the
77 divergence of recently evolved enzymes [11,31] may appear largely inapplicable to the emergence
78 of the very first protein(s) [32]. There are, however, some unifying themes that we describe here
79 alongside differences and unknowns. We conclude by discussing how short and functional
80 protein fragments may have been recruited prior to the appearance of LUCA’s proteome to
81 give rise to primitive metabolic systems.
82
83 ***Box 1***
84 **Microtransitions in protein evolution**
85 Protein evolution is driven by mutations that can occur biasedly [33,34] or at random and with no
86 relation to selection [35]. Deleterious mutations are purged, whereas new challenges drive the
87 fixation of mutations that give rise to proteins with modified or new functions (**Box 1**, *2*, and
88 **Figure 1**).



89

90 **Figure 1. Darwinian evolution driven by pre-existing genetic changes, ranging from single amino acid** 91 **mutations to gene rearrangements**. **(A)** Schematic representation of Darwinian selection: selection purges most 92 of the variations in the population, leading to survival of the fittest mutant, eventually undergoing fixation. **(B)** 93 The outcome of a laboratory evolutionary trajectory of 18 consecutive point mutations (PDB codes: 1DPM, 2R1N, 94 4E3T) [36]. The original and evolved active sites are depicted with their corresponding reaction intermediates (a 95 phosphotriesterase (left) and aryl-esterase (right). The mutated positions are denoted in red. The overall structure 96 (cartoon) and the key catalytic residues remained unchanged (the catalytic metals are presented as grey spheres). 97 **(C)** A switch between two fundamentally different activities, methyltransferase (left) and monooxygenase (right) 98 may be triggered by an insertion of a single amino acid. An inserted serine at position 297 (red) induces a flip of 99 the adjacent side-chain of Phe296 (blue sticks) that reshapes the active-site (surface) and triggers the activity 100 change (PDB codes: 4WXH and 5EEG) [37]. **(D)** Domain insertions into an existing enzyme drive the divergence 101 of new functions [38,39], as exemplified here for three different enzymes that share a Rossmman-fold core domain a 102 Haloacid dehalogenase (PDB 1ZRN), a phosphatase (1N9K) and a calcium pump-driving ATPase (1SU4). The 103 canonical Rossmann fold is represented by a dehydrogenase (5KKA). For other examples of microtransitions see 104 [19–22].

105

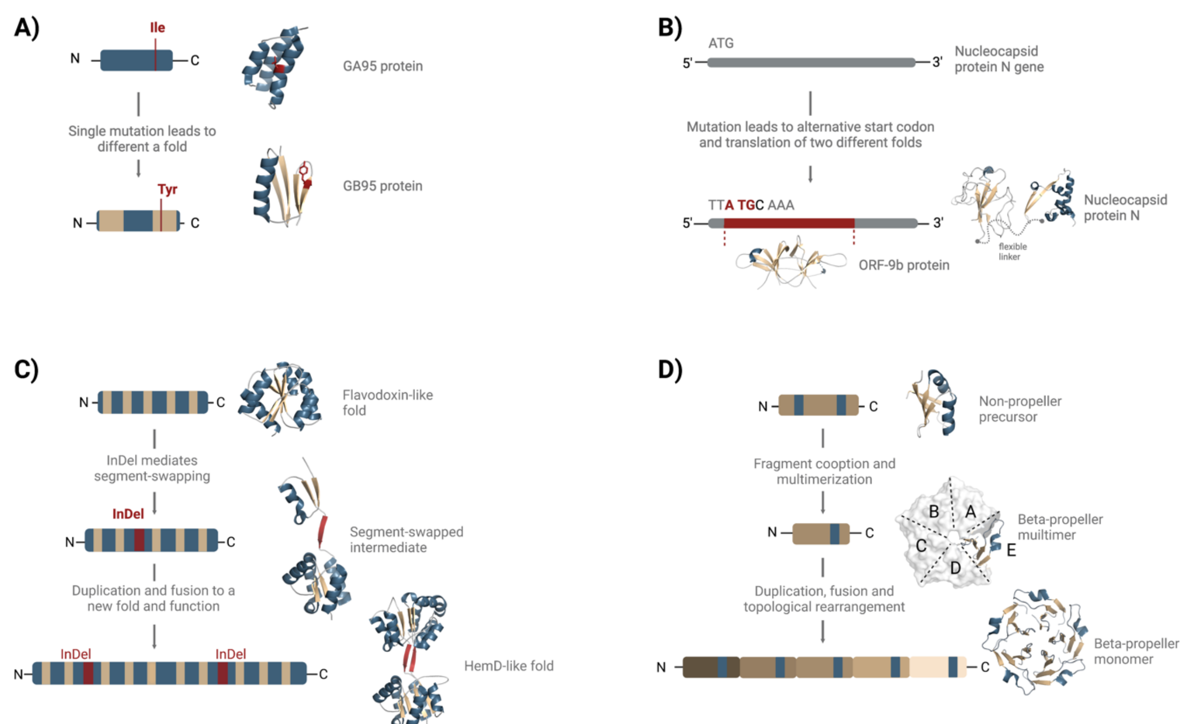
106 As exemplified in **Figure 1**, most, if not all, of the extant protein repertoire emerged by small 107 structural modifications while maintaining their basic fold. Such changes, dubbed 108 *microtransitions* (**Box 1**, *1*), have been demonstrated in the laboratory, largely *via* point 109 mutations, insertions/deletions (InDels), homologous or non-homologous recombination [23] and 110 domain fusions [40]. While the effects of point mutations have been widely explored (e.g., **Figure** 111 **1b**), we know less about how other types of genetic changes lead to new proteins. InDels, for 112 example, have high adaptive potential. For instance, a single InDel can induce functional 113 transitions [37,41] (e.g., **Figure 1c**). Additionally, domains frequently mix and match (gene fusion 114 or fission) to yield new proteins [39,42–44]. The addition of a single relatively small domain allows

Rossmann fold enzymes to catalyse different reactions e.g., calcium ATPase, phosphatase and haloacid dehalogenase (**Figure 1d**). InDels or larger genetic rearrangements are on average, even more deleterious than point mutations and therefore intensely purged [45,46]. Acceptance of mutations in general, and InDels or larger genetic rearrangements especially, typically demands compensation by other mutations (**Box 1,** *6*) [45,47].

In contrast to the divergence of functions in existing domains, the birth of new protein topology and architecture is driven primarily by duplication and fusion of short segments, as discussed in the following section.

Macrotransitions in protein evolution

Protein domains whose secondary structural elements adopt similar orientation in space are classified under the same architecture. If in addition, these elements display identical topological connections, they are further sorted under the same fold. Substantiated cases of homology between different folds are rare. Only recently, the development of sensitive homology prediction tools has allowed drawing evolutionary bridges between folds that were previously thought unrelated [48–55]. Despite these efforts, most of the evolutionary relationships between distant homologs remain a mystery. How did the first protein folds emerge? Did transitions between these architectures occur at any stage; and if so, how?



135 **Figure 2. Macrotransitions: genetic mutations induce changes in protein structure. (A)** A single amino acid
136 mutation (I45Y, red) leads to a fold change as exemplified by protein GA95 (PDB 2KDL): the all-alpha structure
137 protein acquires alpha and beta secondary elements in GB95 (PDB 2KDM). **(B)** Mutations at the DNA level can
138 lead to alternative reading frames. Such is the case for the Nucleocapsid protein N gene that gives rise to the
139 nucleocapsid N ORF-9b protein (PDB 2CME) [56]. The new protein adopts an all-beta fold, in contrast to the alpha
140 and beta elements of the original protein [57]. (**C**) An insertion (red) within the flavodoxin-like fold (PDB 1REQ),
141 results in an additional beta element that segment-swaps the original fold in two. This structural rearrangement
142 creates a protein interface that is now able to associate with another monomer, inducing the topological changes,
143 resulting in the hemD-like fold (PDB 1jr2) [53]. **(D)** Short fragments within proteins can act as building blocks to
144 create novel architectures. A fragment from a non-propeller precursor (PDB 3WHI) upon oligomerisation,
145 duplication and fusion rearrange in a monomeric propeller fold (PDB 5C2N) [58].
146

147 Studies of metamorphic proteins have provided some hints [59–62], which demonstrate that the
148 topology and architecture of protein domains can be altered, herein called *macrotransitions,* by
149 introducing a few or even one single amino acid substitution [56] (**Box 1**, *1*). Such is the case of
150 the G_A protein, a serum binding domain that is converted into G_B, an IgG-binding domain upon
151 a L45Y substitution (**Figure 2a**). This type of structural transition suggests the existence of
152 critical residues that stabilize certain tertiary interactions while abolishing others. Likewise, a
153 single protein sequence can fold into more than one structure. These sequences have more than
154 one energetically favoured minimum (scaffold plasticity) that allows the interconversion
155 between different structures upon changes in the environment such as pH; lipid or buffer
156 composition [63]. *De novo* emergence of proteins by overprinting is another example of a
157 *macrotransition*, where alternative frames of coding sequences from short segments of existing
158 proteins are translated. This phenomenon can give rise to new amino acid sequences, and
159 ultimately to new protein architectures [64–66]. For instance, by incorporating an alternative start
160 codon within the nucleocapsid protein N, an additional reading frame is created, giving rise to
161 ORF-9b protein, which adopts a new fold (**Figure 2b**) [57]. This process is not to be confused
162 with the *de novo* emergence of proteins from non-coding DNA (see open questions), where
163 arbitrary transcripts occasionally overlap with randomly attained open-reading-frames and
164 become translated[67–70]. Further architectural rearrangements can emerge through trading of
165 structurally similar regions (segment-swaps) between two or more domains, which can be
166 found in around 13% of the PDB structures [71]. This type of *macrotransition* can also be induced
167 by InDels within a protein sequence [53] as exemplified by the flavodoxin-like fold, which upon
168 insertion, duplication and fusion gave rise to a new functionality, adopting the bi-lobular
169 hemD-like architecture (**Figure 2c**). Duplication and fusion of short segments can also lead to
170 open-ended (solenoid) structures as indicated by the internal symmetry that underlines many
171 protein folds, [72,73], e.g., αα-hairpin repeats generate TRP, HEAT, Armadillo, and Ankyrin
172 structures, whereas βαβ units generate leucine rich repeats. In other instances, repeating units

173 create globular structures, such is the case for the TIM barrels [74,75] and beta-propellers (**Figure**
174 **2d**) [58,76,77]. Overall, the above-mentioned examples highlight how novel protein architectures
175 can emerge from structurally unrelated scaffolds through relatively small changes, illustrating
176 their plasticity and resilience potential.
177
178 While it is well known that mutations, gene rearrangements and InDels can cause functional
179 and structural changes in proteins, not all these mutations go to fixation. In the next section,
180 we discuss how selection and fixation occur, based on results of various directed evolution
181 experiments on individual proteins.
182
183 **Selection and fixation of mutations**
184 Following their appearance, most mutations are purged while some are fixed, by selection, but
185 also by chance (**Box 1**, *2*). This leads to the critical question: out of all possible mutations in a
186 protein, which fraction of these is neutral versus what fraction is deleterious, and to what degree.
187 Equally crucial is the frequency of potentially beneficial mutations and their effects on the
188 protein's original function and stability as this dictate whether they might be fixed or rapidly
189 purged.
190
191 The answer to these is embedded in the distribution of fitness effects of mutations
192 (DFE) – a subject of extensive research. Systematic mappings of the effects of all possible
193 single amino acid mutations in a given protein have become routine [78–81]. These mutational
194 scans yield distributions of the effects of mutations in individual proteins, and also insights
195 regarding the structural and biochemical parameters that dictate them [82,83]. The cumulative
196 knowledge of protein DFEs indicates that the vast majority of mutations, probably ≥80%, are
197 deleterious [84], with the primary reason being impaired folding and/or decreased stability [84].
198 Mutations that alter biochemical function are rare and also purged more intensely [82,83]. The
199 effects of mutations on folding and stability are complex, as they also relate to how the cellular
200 machinery deals with impaired mutants (see below). Indeed, in the short term, mildly-
201 deleterious mutations may be tolerated owing to various cellular buffering mechanisms, thus
202 facilitating protein evolution [85–87].
203
204 The evolutionary interpretation of deep mutational scans is problematic, not the least
205 because the measured 'fitness' values rarely relate to organismal fitness. Accordingly, most
206 experiments indicate higher mutational tolerance than what observed in nature amongst

207 homologous proteins, suggesting that most mutations, in laboratory conditions, do not affect
208 structure and/or function [84]. It appears that the deleterious effects of mutations are masked in
209 most laboratory experiments [83], rendering the results more relevant to the understanding of
210 short-term genetic diversity (*e.g.* population polymorphism), as opposed to long-term
211 evolutionary processes [84,88]. Similarly, when it comes to adaptation (acquiring new or modified
212 protein properties), laboratory selections may typically be too stringent, thus funnelling
213 adaptations towards one trajectory in a limited and defined environment (a single growth
214 medium, temperature, etc.). The gradual selection pressures and diverse environments that
215 underlie natural evolution may shape protein adaptation in ways that differ from what has been
216 observed in most laboratory experiments [89,90].
217
218 **Evolutionary rates of proteins**
219 When it comes to long-term evolution, the rates by which proteins evolve vary dramatically.
220 Even when comparing proteins of the same species, or orthologues only (i.e. assuming minor
221 changes in protein function), evolutionary rates (substitutions per site, per generation) typically
222 span over 2 orders of magnitude among the proteins in the same genomes. The factors that
223 dictate the rate of protein evolution is of major interest [91]. One key determinant is epistasis,
224 namely interdependency between different positions of the same gene/protein (intragenic
225 epistasis; **Box 1**, *5*). Globular proteins in general exhibit negative epistasis (deleterious effects
226 of two different mutations is greater than the sum of individual ones) [92]. As proteins evolve,
227 deleterious mutations can still be fixed. However, their acceptance depends on the pre-
228 existence of other mutations (permissive, enabling mutations) or on the subsequent
229 accumulation of compensatory mutations (**Box 1**, *6*). This context dependency of mutations
230 dictates a slower rate of evolution. Biophysical and functional constraints also affect rates of
231 protein evolution. These include high expression levels that make proteins more prone to
232 aggregation and promiscuous associations; and multi-functionality, thereby engaging a large
233 fraction of the protein's surface [91,93].The latter two constraints act primarily on the protein
234 surface -namely surface residues mutate 4-fold faster than the core residues. Interestingly, the
235 surface constraints slow down the divergence of other residues, in particular core residues,
236 resulting in an overall very slow evolutionary rate [94].
237
238 Finally, the acquisition of new functions is the strongest driving force to protein sequence
239 changes. Accordingly, mutational trajectories that lead to new protein functions have been
240 extensively studied, revealing in atomic detail the effects of mutations on protein structure and

241 function (**Box 1**, *3-9*) [95]. We note that nearly every long adaptive trajectory beyond few
242 mutations, includes multiple mutations at positions distal to the active site. Despite the
243 importance of these so-called 3rd shell mutations their contribution to the emergence of new
244 protein function remains poorly understood [96,97].
245

246 **Mutagenic hotspots**

247 Mutations that confer modified or new protein functions (adaptive mutations), may pre-exist
248 in the population when a new challenge appears, or may arise within subsequent generations –
249 for example, both pre-existing and arising mutations have been identified in insect esterases
250 that evolved towards insecticide resistance [98–100]. Mutations that are neutral or nearly neutral,
251 with respect to the protein's existing function, and are therefore not purged, may become
252 beneficial upon the emergence of a new challenge (**Box 1**, 7). Still, the occurrence of point
253 mutations is rare (10^{-9} per site, per generation, on average). Thus, the genetic diversity available
254 at any given moment is limited, especially in organisms with small population size. In cases
255 where mutation(s) with adaptive potential do not pre-exist in a population, the initial response
256 to a new challenge is critical. In this context, we review and discuss several mechanisms that
257 may hence expedite adaptation in the absence of pre-existing genetic diversity.
258

259 Cellular stresses correlate with higher mutation rates [101]. Also, the rate and type of mutations
260 vary dramatically, depending on local DNA context *e.g.* short sequence repeats [102] and in a
261 global one *e.g.* highly transcribed regions [33,103]. These so-called adaptive mutations arise due
262 to high mutability of single-stranded DNA in active transcription bubbles and from replication-
263 transcription collisions [104,105]. Similarly, highly transcribed genes may be duplicated *via* cDNA
264 intermediates (retro-genes) [106]. Duplications can vary from gene segments to whole genomes
265 and may also be considered as 'adaptive mutations', whenever they are stress induced and auto-
266 amplified [107]. Under strong selection, multiple copies of a gene mediating survival may emerge
267 within a strikingly small number of generations and disappear immediately after selection is
268 removed [108,109].
269

270 Given high replication fidelity, the above discussed mechanisms may be crucial in shortening
271 the time gap between new challenges and the arrival of mutations that promote survival [110]. It
272 is not trivial to establish direct causality between stress, the induction of genetic changes, and
273 adaptation [101]. Nonetheless, their relevance is highlighted by the existence of explicitly evolved

‘hotspots’ regions in specific genes which encode rapid heritable genetic switches, such as in surface antigen proteins of pathogenic bacteria [102,111].

Protein noise and phenotypic mutations

In most cases, mutations conferring new function are pre-existing in a population. Alternatively, the yet-to-become new function could be already present, as latent, coincidental phenotypic variation whereby a single genotype (a given gene sequence) may give rise to a range of protein sequences, structures and functions, and thereby to multiple phenotypes. If phenotypic protein variability is neutral in the environment(s) under which a protein evolved – this variability comprises ‘molecular noise’. Nonetheless, upon appearance of a new challenge phenotypic variability may provide an immediate survival advantage and increase the adaptive potential. In proteins, phenotypic variation can be displayed in multiple ways, including: (*i*) variable protein levels in a population of cells due to expression noise; (*ii*) latent, promiscuous protein conformations and activities due to drift; and (*iii*) alternative protein sequences due to transcriptional, splicing and/or translational errors.

Here, we focus on transcriptional and translational errors. For the adaptive potential of (*i*) see Refs. [112,113], regarding (*ii*), see **Box 1**, *4*. Translational and transcriptional errors are ~10^5 times more frequent than genetic mutations [114,115]. As a representative example, it was shown that in yeast ADH1 gene, transcriptional errors alone can affect almost every aspect of enzymes function including oligomerization, substrate binding, cofactor binding, metal binding and post-translational modification site (**Figure 3b**) [116]. Like genetic mutations, phenotypic mutations are not limited to single amino acid exchanges – frameshifts, alternative starts and/or stops codons, and larger rearrangements (*e.g., via* alternative splicing) are also common (**Figure 3b-d**). Overall, given the wealth of noise associated with transcription, mRNA processing and protein synthesis, protein copies that deviate from the expected translated gene sequence are abundant [114,117,118]. These so-called phenotypic mutations have a role in the evolutionary shaping of proteins [2,119], and may also provide starting points for emergence of new proteins or functions [115,120]. Short peptide segments that result from ‘illegitimate’ translation (smORFs) are also prevalent due to alternative start/stop codons, off-frame translation of coding sequences, or translation of complementary strands or even of noncoding regions [65,121–123]. Such segments might also comprise the starting material for novel proteins (**Figure 3c, d**) [124,125].

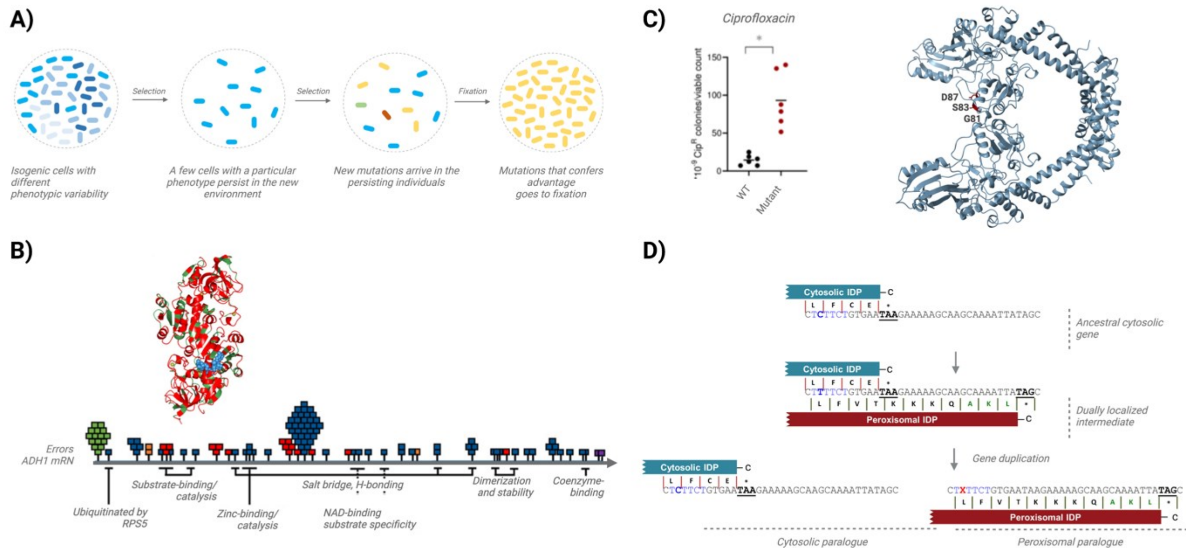


Figure 3. Plasticity-first mechanisms driving protein evolution (A) A schematic representation of selection that follows the plasticity-first mechanism. A new environmental challenge selects a subset of phenotypically variable isogenic cells. The phenotype permits the survival of cells, providing time for the occurrence of a mutation which confers an adaptive advantage. The mutant cells take over the entire population (fixation). **(B)** Transcriptional errors in yeast ADH1 mRNA mapped on to the structure. The residues with errors are highlighted in red. The scheme at the bottom shows that the mutations can affect several aspects of the enzyme: oligomerization, substrate binding, cofactor binding, metal binding and post-translational modification sites [116]. The figure panel is reprinted from the reference (116). (**C**) An *E. coli* mutant exhibiting higher mistranslation rates (phenotypic plasticity) displays higher frequency of genetic mutations that confer antibiotic resistance (adaptation). This panel is reprinted from the reference (135). The right panel shows the structure of DNA gyrase with the mutations conferring ciprofloxacin resistance highlighted in red. (**D**) Translational errors may provide the raw material to new proteins [120]. In the depicted example, a coincidental translational slippage at a TCTTTT site produces an alternative protein form with a C-terminal peroxisomal signal. In the second step, a mutation of C-to-T, that is silent with respect to the original frame, increases slippage rate, thus generating two alternative protein forms from one gene: the original cytosolic form, and a minor peroxisomal form (the AKL peroxisomal signal peptide, denoted in green). Finally, following gene duplication, a single base deletion gives rise to a new, legitimate peroxisomal paralogue, whereas the original, cytosolic gene loses the cryptic peroxisomal signal.

It is important to highlight the fact that, although phenotypic mutations are not heritable as such, the potential to make them can be [118,126–129]. For example, the frequency of transcriptional/translational errors is highly variable and sequence dependent. Codon usage strongly affects the rate of mistranslation [130]. The frequency of slippage to yield phenotypic frameshifts is directly proportional to repeat length, 8 consecutive A's being an example of programmed slippage [131–133]. Therefore, selection may favour gene sequences that increase the frequency of alternative protein variants while retaining the original wild-type protein sequence. In this manner, errors that occur largely at random can be amplified at specific sites and can also be heritable.

Although phenotypic mutations occur at higher frequency, and are shown to be important for the adaptation of organisms [134–137], the experimental evidence for their adaptive role in protein evolution is only recently emerging. Direct evidence for the evolutionary role

337 of phenotypic mutations came from the emergence of a new yeast enzyme paralogue (**Figure**
338 **3d**) [120]. The ancestral gene of isocitrate dehydrogenase (IDP) encodes two enzyme forms
339 (isozymes) – a major cytosolic form by intact translation, and a minor form that possess a C-
340 terminal peroxisomal signal peptide due to a translational frameshift. Following duplication, a
341 single nucleotide genetic deletion gave rise to a new, legitimate peroxisomal paralogue,
342 whereas the cytosolic paralogue lost the translational frameshift that leads to a peroxisomal
343 signal.
344
345 **Genetic accommodation of phenotypic mutations**
346 Phenotypic mutations may bridge the time gap between the appearance of a new challenge and
347 the emergence of a mutation that resolves it (a gap that can be much longer than intuitively
348 assumed). If a challenge persists, what initially comprises coincidental noise often becomes a
349 'legitimate' function *via* the fixation of mutations at the genomic level that refine this function.
350 For example, typically following gene duplication a weak promiscuous enzymatic activity may
351 increase in both rate and selectivity to become the primary function. This was demonstrated
352 recently in a study employing *E. coli* strains with varying levels of translation error rates. The
353 authors show that the *E. coli* mutants with higher error rates show higher frequency of
354 ciprofloxacin resistant colonies compared to WT strains (**Figure 3c**) [138]. Accordingly, lowering
355 the mistranslation rates, reduced the frequency of resistant colonies as well. It is worth noting
356 that the genotypic mutation is often different from the phenotypic mutation.
357 Promiscuous protein activities seem to have a unique evolutionary advantage –
358 mutations that increase them usually have either weak or no deleterious effects on the protein's
359 primary activity (**Box 1**, *4*, *7*, *8*). Phenotypic mutations may also have a unique advantage in
360 how they are genetically accommodated. In the yeast IDP case described above, single base
361 deletions that accommodate the new trait at the DNA level (*i.e.*, in-frame translation of the
362 peroxisomal signal to direct all protein molecules to the peroxisome) occur at the very same
363 mRNA site at which translational slippage occurs [120]. Overlaps between sites of genetic and
364 phenotypic mutations have also been observed in an *in vitro* study [131]. Thus, selection of
365 genotypes exhibiting a higher rate of a specific phenotypic mutation also gives rise to a hotspot
366 for genetic mutations that accommodate the very same trait [120,131]. Similarly, ambiguous
367 decoding (translation of a given codon to two different amino acids) was genetically
368 accommodated in certain organisms via divergence of a dedicated tRNA [117]. More recently, it
369 was also shown that phenotypic mutation can reduce the mutational load in a population by

370 efficiently purging deleterious mutation. Accordingly, phenotypic mutation exhibits negative
371 epistasis with DNA or genotypic mutation [139].
372
373 **Plasticity-first: an emerging model for protein evolution**
374 The so-called Baldwin-effect [12], or in its more modern form, the 'plasticity-first' model [13] refers
375 to the phenomena when non-hereditary molecular variability induced by an environmental
376 change enables initial survival. This buys time for the emergence and accommodation of
377 genetic mutations, ensuring long-term survival of the population in the new environment. Both
378 phenotypic plasticity and the ensuing genetic accommodation of mutations have been
379 extensively examined and debated in the context of developmental plasticity and evolutionary
380 adaptations [8]. Here, we adapted the Baldwin effect [12] and following a recent and insightful
381 review [13], present the key criteria for such a mechanism to be applied to protein evolution
382 (**Figure 3a**).
383
384 The most critical criterion for proving the 'plasticity-first' model for protein evolution is that
385 the yet-to-evolve trait becomes more variable in response to the physiological stress that
386 accompanies the new challenge. For example, the magnitude of certain promiscuous activities
387 or the frequency of translational errors may increase due to changes in metabolite
388 concentrations or pH. Similarly, if some pre-existing, cryptic genetic variation happens to
389 increase the magnitude of trait variability, this would of course promote the 'plasticity-first'
390 mechanism. This criterion is not trivial to establish, and to the best of our knowledge, has not
391 been directly examined in relation to a proven case of protein evolution.
392
393 Indeed, in many cases where the history of acquisition of new protein functions had been
394 tracked down; pre-existing promiscuous functions [140,141], or phenotypic mutations [115,120], were
395 found to have been starting points and even provide initial survival of the population [142,143].
396 However, such trajectories may be perfectly accounted by a Darwinian mutation-selection
397 model, since the pre-existence of mutants with an optimal activity in the population was not
398 examined. Therefore, a key challenge remains: to show that the latent activity was present at a
399 sufficient level to provide a selective advantage before genetic accommodation of mutations.
400
401 Increased molecular noise is inevitably associated with reduced fitness. The cost of increased
402 rate of translational errors, may be tolerable in short-term [114], but in the long-term, high error
403 rates rarely persist [101]. Overall, whilst the 'plasticity-first' model presents an elegant shortcut

404 to the ‘arrival of the fitter’, direct evidence for its role in protein functional evolution is yet to
405 be provided.

406 **Protein evolution - beyond biochemical activity**

407 Biochemical activity —be it ligand binding or catalysis— is the primary driving force of
408 evolutionary innovation. However, within their natural context, proteins are shaped by
409 additional needs and forces that are complex (see **Box 1**, *8*). Following their translation,
410 proteins fold into their native state, and must be sufficiently stable to avoid misfolding,
411 aggregation and/or proteolysis. The interactions of proteins with the cellular machineries that
412 control protein quality are therefore crucial. Chaperones and also proteases, therefore impacts
413 the type and number of mutations tolerated and thus impacts protein evolution [85,86,144].
414
415 Regulation of protein expression is another key property shaping evolutionary trajectory. As
416 indicated by their faster sequence divergence, non-coding elements are more evolvable than
417 the proteins they regulate [145]. Often, the initial steps, and even the driving force for divergence
418 may involve a new mode of transcriptional regulation. Further, the divergence of a new
419 biochemical function is often initiated by increase in expression of an existing protein with a
420 latent, promiscuous function [146]. This divergence may occur via mutations in the gene’s own
421 promoter, in genes encoding other regulatory elements, or via gene duplication (**Box 1**, *3*). By
422 the current view, most new genes, and paralogues especially, diverged in their transcriptional
423 regulation and not in their biochemical function [147,148]. A classic example is the divergence by
424 duplication of yeast Gal1/3. The ancestral, pre-duplicated gene, Gal1, encoded an enzyme, b-
425 galactosidase that also acts as transcriptional co-inducer. Upon duplication, the new paralogue,
426 Gal3, specialized as co-inducer, primarily *via* changes in the promoter that enabled faster
427 triggering of Gal1’s transcription upon appearance of lactose [146].
428
429 Changes in the regulation of protein expression can also affect the evolvability of proteins. In
430 fact, expression levels and protein concentration correlate with evolutionary rates – the higher
431 the protein amount in the cell, the slower the rate [91], although to our knowledge, direct causality
432 has not been established. In the case of Gal3 although the key adaptive step was due to the
433 changes in the promoter, [146] protein activity was also changed. Specialized as a co-inducer,
434 Gal3 lost its enzymatic activity, but gain the ability to bind to Gal80 (the transcriptional
435 repressor) with >10-fold higher affinity compared to Gal1, thus providing a distinct advantage
436 upon switching to lactose as carbon source [149]. The divergence of new genes therefore involves

437 changes in gene expression, that in turn enable changes in protein activity, and *vice versa* – in
438 other words, noncoding and coding regions coevolve [149,150]. Beyond transcription, levels of
439 translation are regulated, as are cellular protein levels (*via* changes in protein turnover rates).
440 The mechanisms and dynamics behind the coevolution of protein expression, turnover and
441 function remain to be elucidated.
442
443 Proteins seldom work as independent subunits, and often self-assemble (homomers) or
444 associated with other proteins (heteromers). About 60% of proteins are known to form
445 complexes [151]. How these multimeric assemblies emerge and if there is adaptive value for these
446 complexes is not clear. Recent experimental [152] and theoretical work [151] suggest that these
447 complexes can emerge by neutral drift just like in the case of catalytic promiscuity. Often, as
448 little as one or two mutations are enough to form new homomeric complexes [153]. Though it is
449 tempting to associate an adaptive value for these assemblies, this remains to be investigated.
450 Finally, protein evolution is also constrained by its cellular location. A new localization
451 imposes new challenges. ~30% of the yeast paralogues and ~15% of *Arabidopsis* paralogues
452 diverged in localization [154,155]. Beyond retargeting, typically by the acquisition of a signal
453 peptide[120], a change in localization enforces adaptation towards export (that may involve
454 unfolding and refolding), different pH and/or redox state, and new protein partners. Overall,
455 protein adaptation is a comprehensive process involving multiple parameters in addition to
456 biochemical activity. Foremost, it is a process of coevolution involving the protein itself, its
457 transcriptional and translational regulatory elements, the cellular protein-handling machineries
458 and other proteins and biomolecules that interact with the evolving protein.
459
460 **Open questions**
461 Beyond the series of questions mentioned above, there are, in our view, three key aspects in
462 protein evolution that remain largely unanswered.
463
464 *Multiple, interlocked protein components.* Proteins rarely confer physiological advantage on
465 their own. Typically, they are part of a system – a pathway, or whole network involving several
466 proteins, -whereby loss of any one of these proteins results in loss of function of the entire
467 system. For example, biosynthetic pathways comprise several enzymes, and loss of any of
468 which of these enzymes typically results in no product. How did these multiple, interlocked
469 protein systems (*MIPSs*) emerge in the first instance?
470

471 Many MIPSs can be unlocked – suffice to say that free-living natural bacteria with <1,400
472 genes are known, and even these genomes can probably be reduced [156]. Thus, the current state
473 of a MIPS does not reflect its initial, emergent state. Relatively simple scenarios for the
474 emergence of MIPSs have been hypothesized [32]. With respect to metabolic pathways,
475 bifunctional enzymes are commonly found, suggesting that certain pathways may have *a priori*
476 evolved to catabolize more than one nutrient, or produce more than one product, and at later
477 stages diverged and specialized (**Box 1**) [157–159]. Nonetheless, the emergence of the first MIPSs,
478 and specifically of the core biosynthetic pathways, remains enigmatic. Spontaneous occurrence
479 of reactions, alongside a few multi-functional enzymes, may have enabled the formation of key
480 metabolites, thus seeding the future pathways [160–163].
481
482 *Pre-LUCA recruitment of the first enzymes.* In the pre-LUCA world, modern enzymes did not
483 exist. Rather, ribozymes, metals, and H^+ and OH^- ions [164] may have been the principal catalysts.
484 In this scenario, it has been postulated that the first peptides could have emerged to assist these
485 early catalysts [165,166]. In fact, the exceptional abilities of peptides to chelate metals, catalyse
486 reaction by themself and concentrate in condensate to enhance their activity, make them ideal
487 seeds for the emergence of complex enzymes.[167,168] An alternative scenario includes amyloids
488 as plausible catalytic unit at the origin of life [170,171]. Not only they show an extraordinary
489 stability against UV radiation, different pH and high salt concentrations, but they also catalyse
490 diverse reactions, including their own formation and correction while being replicated. For
491 these reasons the catalytic role of prototype peptides and/or amyloids prior to the putative pre-
492 LUCA world cannot be excluded [169]. An early form of metabolism could have started *via* the
493 recruitment of small peptides with catalytic properties. These units can be seen as minimalistic
494 representations of enzymes [172–174]. Sequence and structural studies on protein domains suggest
495 that the first proteins may have emerged by repetition, fusion, recombination, and augmentation
496 of primordial peptides [175]. These peptide units can be found in modern protein domains with
497 distinct global architecture [55,176,177] and were probably catalytically active as stand-alone, even
498 if less efficient than their contemporary descendants, as well as stable enough to survive. Many
499 questions remain unanswered on how these minimal and functional structures were recruited
500 to replace pre-biotic catalysts and eventually lead to modern protein world.
501
502
503
504

505 *De novo emergence of proteins*. So far, we have addressed a large body of evidence related to
506 transitions (micro- and macro-) in proteins that have a pre-existing globular 3D-structure (and
507 function), but how does structure and function evolve in *de novo* proteins? *De novo* proteins
508 are encoded in genes that emerge from non-coding segments of the DNA sequence [178–180].
509 These new proteins are highly disordered and represent an excellent model system to study
510 how globular proteins evolved from a disordered precursor. The foldability of a *de novo* protein
511 was examined in detail, showing that it adopts a rudimentary fold, exhibits amyloid-like
512 properties and could act as a precursor for the emergence of fully folded proteins [179]. The study
513 of *de novo* proteins might provide in the future new general principles for the evolution of
514 folded proteins.
515
516 Overall, the evolution of MIPSs, the recruitment of first enzymes, and de novo emergence of
517 proteins are aspects where our knowledge is still at infancy. As our understanding of how
518 proteins evolve advances, new insights will emerge that address these and other key questions.

519 **Box 1: Concepts and mechanisms in protein evolution - a very brief guide**

520 The text focuses on a few less explored aspects of protein evolution, while more established
521 aspects are covered in this box that lists key concepts and guiding references (reviews and
522 recent papers describing specific case studies). Scientific concepts and mechanisms are
523 inevitably schematic (if not dogmatic). Alternative scenarios or mechanisms are denoted here
524 side-by-side in blue (noted as 'versus', 'alternatively', etc.). In reality, these are not mutually
525 exclusive and may be even complementary. Many concepts are also interrelated as indicated
526 in our cross-referencing.

527 (*1*) Transitions in protein evolution can be categorized to:

528 **Microtransitions** – divergence of new functions while maintaining the original architecture
529 (fold) and key active-site features (divergence within protein families and superfamilies).

530 **Macrotransitions** – transitions between different folds including the emergence of the earliest
531 protein folds.

532 (*2*) Protein sequences diverge with time (this is what evolution means). Schematically, these
533 changes may relate to drift or adaptation:

534 **Drift** – sequence changes occurring due to random sampling while preserving the protein's
535 structure and function (see *purifying selection*).

536 **Adaptation** – changes in protein properties including the acquisition of new biochemical
537 activities (see *positive selection*).

538 **Selection** may drive a reduction in the frequency of certain mutations (alleles) within a given
539 population (purging, *purifying selection*) and/or the enrichment of other mutations (*positive*
540 *selection*). Selection shapes protein traits including their biochemical activity (binding,
541 catalysis, etc.) and biophysical properties (folding, stability, etc.). Traits such as enzyme
542 selectivity relate to *positive selection*, i.e., enrichment of mutations that increase binding, or
543 catalytic efficiency with the target ligand/substrate, but also by mutations that reduce activity
544 with undesirable, non-cognate substrates [19,181] (see also *trade-offs*). The latter is often addressed
545 as '*negative selection'* (although in population genetics this term is used in relation to purifying
546 selection).

547 (*3*) ***Gene duplication*** provides the raw material for new proteins. Several different mechanisms
548 may underline the emergence of new genes via duplication [4,31,182]. Briefly, duplicated genes

may evolve towards a novel function that had not been present in the ancestral, pre-duplicated gene (*neo-functionalization*). Alternatively, a bifunctional ancestor (*generalist*) may split to two *specialist* genes (*sub-functionalization*, or *divergence before duplication*). Duplication may also provide an adaptive advantage *per se*, by increasing protein dose and thereby augmenting a weak, pre-existing promiscuous function [110].

(*4*) ***Promiscuity*** relates to the coincidental pre-existence of functions that may serve as the starting point for new functions [9–11]. If such *latent*, *promiscuous* functions come under selection, they give rise to bi-functional, *generalist* intermediates. Upon gene duplication, generalist intermediates split, giving rise to two *specialists*, each performing one function (*sub-functionalization*) [14,18,21]. Although duplication and going from *generalists to specialists* is a general trend, the opposing process of gene loss and/or *specialist to generalist* also occurs [183].

(*5*) ***Epistasis*** – the effects of mutations in different genes, but also within the same gene/protein can be non-additive, *i.e.*, *epistatic*. Epistasis has a profound impact on evolution in general, and protein evolution in particular [184–186].

(*6*) **Enabling/compensatory mutations.** The dominance of epistasis also means that many (probably most) mutations that eventually get fixed in evolving proteins are deleterious on their own (during *drift*, and certainly during *adaptation*). Their acceptance may therefore occur via two alternative mechanisms: A deleterious mutation transiently accumulates and is later followed by a *compensatory mutation* [47]. Alternatively, mutations that accumulate initially as neutral enable deleterious mutations to fix at a later stage (*enabling, permissive* mutations) [45,187,188].

Enabling and compensation (and hence epistasis) can be *local* or *specific* [184] – *i.e.*, the deleterious and enabling/compensatory mutations occur in a specific pair of residues (typically, in two contacting residues, *e.g.* within active-sites); or *global*, *nonspecific* – a given mutation may enable/compensate a range of different deleterious mutations (*e.g.* stabilizing mutations that may compensate many different destabilizing mutations).

(*7*) ***Neutrality, robustness*** relates to the ability of proteins to accumulate mutations with no change of structure, stability or function. ***Evolvability***, or ***innovability***, relate to the ability of one, or a few mutations to introduce a new structure and/or function.

While seemingly contradictory, these properties are actually complementary [189,190] - this is primarily because mutations may be neutral in one context (function, environment) yet

580 beneficial in another (*e.g.*, neutral mutations with respect to a protein's native, physiological
581 function may augment a latent, promiscuous activity; see also original-new function *tradeoff).*

582 (*8*) ***Trade-offs in protein evolution*** – mutations almost always affect more than one protein
583 trait (*pleiotropy*) and often in contradictory ways. Epistasis and trade-offs are the key elements
584 shaping the trajectories of protein evolution [36]. Several types of evolutionary trade-offs are
585 known with respect to proteins:

586 *Original vs New-function trade-off* – a mutation improving a new, evolving function is likely
587 to decrease the original one. A strong trade-off enforces *neo-functionalization*, *i.e.*, duplication
588 must occur to complete divergence and specialization (*escape from adaptive conflict*) [4]. In
589 many cases this trade-off is, initially weak, thus enabling divergence towards a bifunctional,
590 *generalist* intermediate (see *sub-functionalization* above). The magnitude of original-new
591 trade-offs tends to vary along adaptive trajectories, starting from weak trade-offs that give rise
592 to *generalist* intermediates, and shifting to strong trade-offs as selection progresses, thus
593 yielding a new *specialist* (typically after *duplication*) [36,95].

594 *Stability-activity trade-off* – most mutations decrease protein stability and thereby lead to
595 misfolding, aggregation and/or proteolysis. New-function mutations are even more so, thus
596 making their accumulation dependent on *enabling/compensatory mutations* [191,192].

597 *Folding-stability trade-off* – beyond the thermodynamic and kinetic stability of the native,
598 folded state, the folding process itself imposes severe constrains. Trade-offs between monomer
599 folding and assembly of oligomers, or between the ability of a protein to fold and the stability
600 of its final, folded state, may underline the birth of new proteins [58].

601 *Rate-accuracy trade-off* – a mutation that improves the catalytic efficiency of an enzyme may
602 reduce its selectivity. Similarly, improvement in the affinity towards the cognate ligand may
603 also increase cross-reactivity with noncognate ligands (see also *positive* versus *negative*
604 *selection*) [181].

605 (*9*) ***Diminishing returns*** – evolutionary optimizations, including protein optimizations, are
606 subject to strong diminishing returns – early mutations confer large advantages per mutation
607 but as the new, evolving trait improves, the improvement per mutations decreases.[36,95]. *Trade-*
608 *offs*, *diminishing returns* and other factors result in many proteins being suboptimal with
609 respect to individual traits such as catalytic efficiency, selectivity and stability [181,193].

610

611 (*10*) ***Phenotypic variation***
612 Variation that exists in a genetically identical population due to the noise associated with
613 various biological processes like transcription, translation, splicing etc.
614
615
616 **Acknowledgments**
617 We thank all members of Dan S Tawfik's lab for insightful discussions on the first version of
618 this manuscript. We thank Vikram Alva, Yitzhak Pilpel, Amy Stanton Gooch and Raul Mireles
619 for the critical reading of the manuscript. Financial support from the Okinawa Institute of
620 Science and Technology to P.L. and from the Hebrew University of Jerusalem to L.N.-G. is
621 gratefully acknowledged. V.J is a senior post-doctoral fellow from the Feinberg Graduate
622 School, Weizmann Institute of Science.
623
624 **References**
625 1. Vries H de. *Species and Varieties: Their Origin by Mutation : Lectures Delivered at the*
626 *University of California*. Open Court Publishing Company; 1904.
627 2. Goldsmith M, Tawfik DS. Potential role of phenotypic mutations in the evolution of
628 protein expression and stability. *Proc Natl Acad Sci U S A*. 2009;106(15):6197-6202.
629 doi:10.1073/pnas.0809506106
630 3. Whitehead DJ, Wilke CO, Vernazobres D, Bornberg-Bauer E. The look-ahead effect
631 of phenotypic mutations. *Biol Direct*. 2008;3:1-15. doi:10.1186/1745-6150-3-18
632 4. Sikosek T, Chan HS, Bornberg-Bauer E. Escape from adaptive conflict follows from
633 weak functional trade-offs and mutational robustness. *Proc Natl Acad Sci U S A*.
634 2012;109(37):14888-14893. doi:10.1073/pnas.1115620109
635 5. Ackermann M. A functional perspective on phenotypic heterogeneity in
636 microorganisms. *Nat Rev Microbiol*. 2015;13(8):497-508. doi:10.1038/nrmicro3491
637 6. Wagner A. *The Origins of Evolutionary Innovations: A Theory of Transformative*
638 *Change in Living Systems*.; 2011. doi:10.1093/acprof:oso/9780199692590.001.0001
639 7. Tawfik DS. Messy biology and the origins of evolutionary innovations. *Nat Chem Biol*.
640 2010;6(10):692-696. doi:10.1038/nchembio.441
641 8. Mary Jane West-Eberhard. *Developmental Plasticity and Evolution*.; 2003.
642 doi:10.1093/oso/9780195122343.001.0001
643 9. Pandya C, Farelli JD, Dunaway-Mariano D, Allen KN. Enzyme promiscuity: Engine of
644 evolutionary innovation. *J Biol Chem*. 2014;289(44):30229-30236.
645 doi:10.1074/jbc.R114.572990

646 10. Copley SD. An evolutionary biochemist's perspective on promiscuity. *Trends Biochem*
647 *Sci*. 2015;40(2):72-78. doi:10.1016/j.tibs.2014.12.004
648 11. Khersonsky O, Tawfik DS. Enzyme promiscuity: A mechanistic and evolutionary
649 perspective. *Annu Rev Biochem*. 2010;79:471-505. doi:10.1146/annurev-biochem-
650 030409-143718
651 12. Crispo E. The Baldwin effect and genetic assimilation: Revisiting two mechanisms of
652 evolutionary change mediated by phenotypic plasticity. *Evolution (N Y)*.
653 2007;61(11):2469-2479. doi:10.1111/j.1558-5646.2007.00203.x
654 13. Levis NA, Pfennig DW. Evaluating "Plasticity-First" Evolution in Nature: Key Criteria
655 and Empirical Approaches. *Trends Ecol Evol*. 2016;31(7):563-574.
656 doi:10.1016/j.tree.2016.03.012
657 14. Hudson WH, Kossmann BR, De Vera IMS, et al. Distal substitutions drive divergent
658 DNA specificity among paralogous transcription factors through subdivision of
659 conformational space. *Proc Natl Acad Sci U S A*. 2016;113(2):326-331.
660 doi:10.1073/pnas.1518960113
661 15. Escudero JA, Loot C, Parissi V, Nivina A, Bouchier C, Mazel D. Unmasking the
662 ancestral activity of integron integrases reveals a smooth evolutionary transition
663 during functional innovation. *Nat Commun*. 2016;7. doi:10.1038/ncomms10937
664 16. Salmon M, Laurendon C, Vardakou M, et al. Emergence of terpene cyclization in
665 Artemisia annua. *Nat Commun*. 2015;6:4-13. doi:10.1038/ncomms7143
666 17. Elliott AG, Delay C, Liu H, et al. Evolutionary origins of a bioactive peptide buried
667 within Preproalbumin. *Plant Cell*. 2014;26(3):981-995. doi:10.1105/tpc.114.123620
668 18. Pougach K, Voet A, Kondrashov FA, et al. Duplication of a promiscuous transcription
669 factor drives the emergence of a new regulatory network. *Nat Commun*. 2014;5.
670 doi:10.1038/ncomms5868
671 19. McKeown AN, Bridgham JT, Anderson DW, Murphy MN, Ortlund EA, Thornton JW.
672 Evolution of DNA specificity in a transcription factor family produced a new gene
673 regulatory module. *Cell*. 2014;159(1):58-68. doi:10.1016/j.cell.2014.09.003
674 20. Coyle SM, Flores J, Lim WA. XExploitation of latent allostery enables the evolution of
675 new modes of MAP kinase regulation. *Cell*. 2013;154(4):875-887.
676 doi:10.1016/j.cell.2013.07.019
677 21. Bar-Rogovsky H, Hugenmatter A, Tawfik DS. The evolutionary origins of detoxifying
678 enzymes: The mammalian serum paraoxonases (PONs) relate to bacterial
679 homoserine lactonases. *J Biol Chem*. 2013;288(33):23914-23927.
680 doi:10.1074/jbc.M112.427922

681 22. Coelho PS, Wang ZJ, Ener ME, et al. A serine-substituted P450 catalyzes highly
682 efficient carbene transfer to olefins in vivo. *Nat Chem Biol*. 2013;9(8):485-487.
683 doi:10.1038/nchembio.1278

684 23. Trudeau DL, Smith MA, Arnold FH. Innovation by homologous recombination. *Curr*
685 *Opin Chem Biol*. 2013;17(6):902-909. doi:10.1016/j.cbpa.2013.10.007

686 24. Agozzino L, Dill KA. Protein evolution speed depends on its stability and abundance
687 and on chaperone concentrations. *Proc Natl Acad Sci U S A*. 2018;115(37):9092-
688 9097. doi:10.1073/pnas.1810194115

689 25. Manhart M, Morozov A V. Protein folding and binding can emerge as evolutionary
690 spandrels through structural coupling. *Proc Natl Acad Sci U S A*. 2015;112(6):1797-
691 1802. doi:10.1073/pnas.1415895112

692 26. Modi T, Campitelli P, Kazan IC, Ozkan SB. Protein folding stability and binding
693 interactions through the lens of evolution: a dynamical perspective. *Curr Opin Struct*
694 *Biol*. 2021;66:207-215. doi:10.1016/j.sbi.2020.11.007

695 27. Levy E, Teichmann S. *Structural, Evolutionary, and Assembly Principles of Protein*
696 *Oligomerization*. Vol 117.; 2013. doi:10.1016/B978-0-12-386931-9.00002-7

697 28. Castro-Fernandez V, Herrera-Morande A, Zamora R, et al. Reconstructed ancestral
698 enzymes reveal that negative selection drove the evolution of substrate specificity in
699 ADP-dependent kinases. *J Biol Chem*. 2017;292(38):15598-15610.
700 doi:10.1074/jbc.M117.790865

701 29. Hartl FU, Bracher A, Hayer-Hartl M. Molecular chaperones in protein folding and
702 proteostasis. *Nature*. 2011;475(7356):324-332. doi:10.1038/nature10317

703 30. Chen B, Retzlaff M, Roos T, Frydman J. Cellular strategies of protein quality control.
704 *Cold Spring Harb Perspect Biol*. 2011;3(8):1-14. doi:10.1101/cshperspect.a004374

705 31. Soskine M, Tawfik DS. Mutational effects and the evolution of new protein functions.
706 *Nat Rev Genet*. 2010;11(8):572-582. doi:10.1038/nrg2808

707 32. Noda-Garcia L, Liebermeister W, Tawfik DS. Metabolite-Enzyme Coevolution: From
708 Single Enzymes to Metabolic Pathways and Networks. *Annu Rev Biochem*.
709 2018;87:187-216. doi:10.1146/annurev-biochem-062917-012023

710 33. Monroe JG, Srikant T, Carbonell-Bejerano P, et al. Mutation bias reflects natural
711 selection in Arabidopsis thaliana. *Nature*. 2022;602(February). doi:10.1038/s41586-
712 021-04269-6

713 34. McDonald J, Kreitman M. Adaptive protein evolution at Adh in Drosophila. *Nature*.
714 1991;351(June):652-654.
715 http://ib.berkeley.edu/labs/slatkin/popgenjclub/pdf/mcdonald-kreitman1991.pdf.

716 35. Luria, S. E. & Delbruck M. Mutations of Bacteria from Virus Sensitivity to Virus
717 Resistance. *Genetics*. 1943;28:491-511.
718 36. Kaltenbach M, Tokuriki N. Dynamics and constraints of enzyme evolution. *J Exp Zool*
719 *Part B Mol Dev Evol*. 2014;322(7):468-487. doi:10.1002/jez.b.22562
720 37. Grocholski T, Dinis P, Niiranen L, Niemi J, Metsä-Ketelä M. Divergent evolution of an
721 atypical S-adenosyl-L-methionine-dependent monooxygenase involved in
722 anthracycline biosynthesis. *Proc Natl Acad Sci U S A*. 2015;112(32):9866-9871.
723 doi:10.1073/pnas.1501765112
724 38. Meng EC, Babbitt PC. Topological variation in the evolution of new reactions in
725 functionally diverse enzyme superfamilies. *Curr Opin Struct Biol*. 2011;21(3):391-397.
726 doi:10.1016/j.sbi.2011.03.007
727 39. Allen KN, Dunaway-Mariano D. Markers of fitness in a successful enzyme
728 superfamily. *Curr Opin Struct Biol*. 2009;19(6):658-665. doi:10.1016/j.sbi.2009.09.008
729 40. Pandya C, Brown S, Pieper U, et al. Consequences of domain insertion on sequence-
730 structure divergence in a superfold. *Proc Natl Acad Sci U S A*. 2013;110(36):3381-
731 3387. doi:10.1073/pnas.1305519110
732 41. Tóth-Petróczy Á, Tawfik DS. Hopeful (protein InDel) monsters? *Structure*.
733 2014;22(6):803-804. doi:10.1016/j.str.2014.05.013
734 42. Moore AD, Björklund ÅK, Ekman D, Bornberg-Bauer E, Elofsson A. Arrangements in
735 the modular evolution of proteins. *Trends Biochem Sci*. 2008;33(9):444-451.
736 doi:10.1016/j.tibs.2008.05.008
737 43. Dohmen E, Klasberg S, Bornberg-Bauer E, Perrey S, Kemena C. The modular nature
738 of protein evolution: Domain rearrangement rates across eukaryotic life. *BMC Evol*
739 *Biol*. 2020;20(1):1-13. doi:10.1186/s12862-020-1591-0
740 44. Weiner J, Beaussart F, Bornberg-Bauer E. Domain deletions and substitutions in the
741 modular protein evolution. *FEBS J*. 2006;273(9):2037-2047. doi:10.1111/j.1742-
742 4658.2006.05220.x
743 45. Tóth-Petróczy Á, Tawfik DS. Protein insertions and deletions enabled by neutral
744 roaming in sequence space. *Mol Biol Evol*. 2013;30(4):761-771.
745 doi:10.1093/molbev/mst003
746 46. Emond S, Petek M, Kay EJ, et al. Accessing unexplored regions of sequence space
747 in directed enzyme evolution via insertion/deletion mutagenesis. *Nat Commun*.
748 2020;11(1):1-14. doi:10.1038/s41467-020-17061-3
749 47. Leushkin E V., Bazykin GA, Kondrashov AS. Insertions and deletions trigger adaptive
750 walks in Drosophila proteins. *Proc R Soc B Biol Sci*. 2012;279(1740):3075-3082.

751 doi:10.1098/rspb.2011.2571

752 48. Grishin N V. Fold change in evolution of protein structures. *J Struct Biol*. 2001;134(2-
753 3):167-185. doi:10.1006/jsbi.2001.4335

754 49. Andreeva A, Murzin AG. Evolution of protein fold in the presence of functional
755 constraints. *Curr Opin Struct Biol*. 2006;16(3):399-408. doi:10.1016/j.sbi.2006.04.003

756 50. Andreeva A, Prlić A, Hubbard TJP, Murzin AG. SISYPHUS - Structural alignments for
757 proteins with non-trivial relationships. *Nucleic Acids Res*. 2007;35(SUPPL. 1):253-
758 259. doi:10.1093/nar/gkl746

759 51. Alva V, Koretke KK, Coles M, Lupas AN. Cradle-loop barrels and the concept of
760 metafolds in protein classification by natural descent. *Curr Opin Struct Biol*.
761 2008;18(3):358-365. doi:10.1016/j.sbi.2008.02.006

762 52. Alva V, Remmert M, Biegert A, Lupas AN, Söding J. A galaxy of folds. *Protein Sci*.
763 2010;19(1):124-130. doi:10.1002/pro.297

764 53. Toledo-Patiño S, Chaubey M, Coles M, Höcker B. Reconstructing the Remote Origins
765 of a Fold Singleton from a Flavodoxin-Like Ancestor. *Biochemistry*. 2019;58(48):4790-
766 4793. doi:10.1021/acs.biochem.9b00900

767 54. josé arcadio Farías-rico steffen schmidt & BH. Evolutionary relationship of two
768 ancient protein superfolds. 2014;(july). doi:10.1038/nchembio.1579

769 55. Kolodny R, Nepomnyachiy S, Tawfik DS, Ben-Tal N. Bridging Themes: Short Protein
770 Segments Found in Different Architectures. *Mol Biol Evol*. 2021;38(6):2191-2208.
771 doi:10.1093/molbev/msab017

772 56. Shortle D. One sequence plus one mutation equals two folds. *Proc Natl Acad Sci U S*
773 *A*. 2009;106(50):21011-21012. doi:10.1073/pnas.0912370107

774 57. Meier C, Aricescu AR, Assenberg R, et al. The Crystal Structure of ORF-9b, a Lipid
775 Binding Protein from the SARS Coronavirus. *Structure*. 2006;14(7):1157-1165.
776 doi:10.1016/j.str.2006.05.012

777 58. Smock RG, Yadid I, Dym O, Clarke J, Tawfik DS. De Novo Evolutionary Emergence
778 of a Symmetrical Protein Is Shaped by Folding Constraints. *Cell*. 2016;164(3):476-
779 486. doi:10.1016/j.cell.2015.12.024

780 59. Vila JA. Metamorphic Proteins in Light of Anfinsen's Dogma. *J Phys Chem Lett*.
781 2020;11(13):4998-4999. doi:10.1021/acs.jpclett.0c01414

782 60. Das M, Chen N, LiWang A, Wang LP. Identification and characterization of
783 metamorphic proteins: Current and future perspectives. *Biopolymers*. 2021;112(10).
784 doi:10.1002/bip.23473

785 61. Madhurima K, Nandi B, Sekhar A. Metamorphic proteins: The Janus proteins of
786 structural biology. *Open Biol*. 2021;11(4). doi:10.1098/rsob.210012
787 62. Dishman AF, Tyler RC, Fox JC, et al. Evolution of fold switching in a metamorphic
788 protein. *Science (80-)*. 2021;371(6524):86-90. doi:10.1126/science.abd8700
789 63. Lella M, Mahalakshmi R. Metamorphic Proteins: Emergence of Dual Protein Folds
790 from One Primary Sequence. *Biochemistry*. 2017;56(24):2971-2984.
791 doi:10.1021/acs.biochem.7b00375
792 64. Carter CW. Simultaneous codon usage, the origin of the proteome, and the
793 emergence of de-novo proteins. *Curr Opin Struct Biol*. 2021;68:142-148.
794 doi:10.1016/j.sbi.2021.01.004
795 65. Sabath N, Wagner A, Karlin D. Evolution of viral proteins originated de novo by
796 overprinting. *Mol Biol Evol*. 2012;29(12):3767-3780. doi:10.1093/molbev/mss179
797 66. Pavesi A, Magiorkinis G, Karlin DG. Viral Proteins Originated De Novo by Overprinting
798 Can Be Identified by Codon Usage: Application to the "Gene Nursery" of
799 Deltaretroviruses. *PLoS Comput Biol*. 2013;9(8). doi:10.1371/journal.pcbi.1003162
800 67. Bornberg-Bauer E, Schmitz JF. Fact or fiction: Updates on how protein-coding genes
801 might emerge de novo from previously non-coding DNA. *F1000Research*. 2017;6(0).
802 doi:10.12688/f1000research.10079.1
803 68. Baalsrud HT, Tørresen OK, Solbakken MH, et al. De Novo Gene Evolution of
804 Antifreeze Glycoproteins in Codfishes Revealed by Whole Genome Sequence Data.
805 *Mol Biol Evol*. 2018;35(3):593-606. doi:10.1093/molbev/msx311
806 69. Cai J, Zhao R, Jiang H, Wang W. De novo origination of a new protein-coding gene in
807 Saccharomyces cerevisiae. *Genetics*. 2008;179(1):487-496.
808 doi:10.1534/genetics.107.084491
809 70. Lange A, Patel PH, Heames B, et al. Structural and functional characterization of a
810 putative de novo gene in Drosophila. *Nat Commun*. 2021;12(1):1-13.
811 doi:10.1038/s41467-021-21667-6
812 71. Szilágyi A, Zhang Y, Závodszky P. Intra-chain 3D segment swapping spawns the
813 evolution of new multidomain protein architectures. *J Mol Biol*. 2012;415(1):221-235.
814 doi:10.1016/j.jmb.2011.10.045
815 72. Kobe B, Kajava A V. When protein folding is simplified by protein coiling solenoid
816 structures. *Trend Biochem Sci*. 2000;25(10):509-515.
817 73. Andrade MA, Perez-Iratxeta C, Ponting CP. Protein repeats: Structures, functions,
818 and evolution. *J Struct Biol*. 2001;134(2-3):117-131. doi:10.1006/jsbi.2001.4392
819 74. Romero-Romero S, Kordes S, Michel F, Höcker B. Evolution, folding, and design of

820 TIM barrels and related proteins. *Curr Opin Struct Biol*. 2021;68:94-104.
821 doi:10.1016/j.sbi.2020.12.007
822 75. Söding J, Remmert M, Biegert A. HHrep: De novo protein repeat detection and the
823 origin of TIM barrels. *Nucleic Acids Res*. 2006;34(WEB. SERV. ISS.):137-142.
824 doi:10.1093/nar/gkl130
825 76. Kopec KO, Lupas AN. β-Propeller Blades as Ancestral Peptides in Protein Evolution.
826 *PLoS One*. 2013;8(10). doi:10.1371/journal.pone.0077074
827 77. Lee J, Blaber M. Experimental support for the evolution of symmetric protein
828 architecture from a simple peptide motif. *Proc Natl Acad Sci U S A*. 2011;108(1):126-
829 130. doi:10.1073/pnas.1015032108
830 78. Fowler DM, Stephany JJ, Fields S. Measuring the activity of protein variants on a
831 large scale using deep mutational scanning. *Nat Protoc*. 2014;9(9):2267-2284.
832 doi:10.1038/nprot.2014.153
833 79. Heyne M, Shirian J, Cohen I, et al. Climbing up and down Binding Landscapes
834 through Deep Mutational Scanning of Three Homologous Protein-Protein Complexes.
835 *J Am Chem Soc*. 2021;143(41):17261-17275. doi:10.1021/jacs.1c08707
836 80. Newberry RW, Leong JT, Chow ED, Kampmann M, DeGrado WF. Deep mutational
837 scanning reveals the structural basis for α-synuclein activity. *Nat Chem Biol*.
838 2020;16(6):653-659. doi:10.1038/s41589-020-0480-6
839 81. Chen JZ, Fowler DM, Tokuriki N. Comprehensive exploration of the translocation,
840 stability and substrate recognition requirements in vim-2 lactamase. *Elife*. 2020;9:1-
841 31. doi:10.7554/eLife.56707
842 82. Firnberg E, Labonte JW, Gray JJ, Ostermeier M. A comprehensive, high-resolution
843 map of a Gene's fitness landscape. *Mol Biol Evol*. 2014;31(6):1581-1592.
844 doi:10.1093/molbev/msu081
845 83. Rockah-Shmuel L, Tóth-Petróczy Á, Tawfik DS. Systematic Mapping of Protein
846 Mutational Space by Prolonged Drift Reveals the Deleterious Effects of Seemingly
847 Neutral Mutations. *PLoS Comput Biol*. 2015;11(8):1-28.
848 doi:10.1371/journal.pcbi.1004421
849 84. Boucher JI, Bolon DNA, Tawfik DS. Quantifying and understanding the fitness effects
850 of protein mutations: Laboratory versus nature. *Protein Sci*. 2016;25:1219-1226.
851 doi:10.1002/pro.2928
852 85. Wyganowski KT, Kaltenbach M, Tokuriki N. GroEL/ES buffering and compensatory
853 mutations promote protein evolution by stabilizing folding intermediates. *J Mol Biol*.
854 2013;425(18):3403-3414. doi:10.1016/j.jmb.2013.06.028

855 86. Bershtein S, Mu W, Serohijos AWR, Zhou J, Shakhnovich EI. Protein Quality Control
856 Acts on Folding Intermediates to Shape the Effects of Mutations on Organismal
857 Fitness. *Mol Cell*. 2013;49(1):133-144. doi:10.1016/j.molcel.2012.11.004
858 87. Kadibalban AS, Bogumil D, Landan G, Dagan T. DnaK-dependent accelerated
859 evolutionary rate in prokaryotes. *Genome Biol Evol*. 2016;8(5):1590-1599.
860 doi:10.1093/gbe/evw102
861 88. Young DL, Fields S. The role of functional data in interpreting the effects of genetic
862 variation. *Mol Biol Cell*. 2015;26(22):3904-3908. doi:10.1091/mbc.E15-03-0153
863 89. De Vos MGJ, Dawid A, Sunderlikova V, Tans SJ. Breaking evolutionary constraint
864 with a tradeoff ratchet. *Proc Natl Acad Sci U S A*. 2015;112(48):14906-14911.
865 doi:10.1073/pnas.1510282112
866 90. Noda-García L, Davidi D, Korenblum E, et al. Chance and pleiotropy dominate
867 genetic diversity in complex bacterial environments. *Nat Microbiol*. 2019;4(7):1221-
868 1230. doi:10.1038/s41564-019-0412-y
869 91. Echave J, Spielman SJ, Wilke CO. Causes of evolutionary rate variation among
870 protein sites. *Nat Rev Genet*. 2016;17(2):109-121. doi:10.1038/nrg.2015.18
871 92. Bershtein S, Segal M, Bekerman R, Tokuriki N, Tawfik DS. Robustness-epistasis link
872 shapes the fitness landscape of a randomly drifting protein. *Nature*.
873 2006;444(7121):929-932. doi:10.1038/nature05385
874 93. Yang JR, Liao BY, Zhuang SM, Zhang J. Protein misinteraction avoidance causes
875 highly expressed proteins to evolve slowly. *Proc Natl Acad Sci U S A*.
876 2012;109(14):5158-5159. doi:10.1073/pnas.1117408109
877 94. Tóth-Petróczy Á, Tawfik DS. Slow protein evolutionary rates are dictated by surface -
878 core association. *Proc Natl Acad Sci U S A*. 2011;108(27):11151-11156.
879 doi:10.1073/pnas.1015994108
880 95. Tokuriki N, Jackson CJ, Afriat-Jurnou L, Wyganowski KT, Tang R, Tawfik DS.
881 Diminishing returns and tradeoffs constrain the laboratory optimization of an enzyme.
882 *Nat Commun*. 2012;3. doi:10.1038/ncomms2246
883 96. Gade M, Tan LL, Damry AM, et al. Substrate Dynamics Contribute to Enzymatic
884 Specificity in Human and Bacterial Methionine Adenosyltransferases. *JACS Au*. 2021.
885 doi:10.1021/jacsau.1c00464
886 97. Ben-David M, Soskine M, Dubovetskyi A, et al. Enzyme evolution: An epistatic ratchet
887 versus a smooth reversible transition. *Mol Biol Evol*. 2020;37(4):1133-1147.
888 doi:10.1093/molbev/msz298
889 98. Clarkson CS, Temple HJ, Miles A. The genomics of insecticide resistance: insights

890 from recent studies in African malaria vectors. *Curr Opin Insect Sci*. 2018;27:111-115.
891 doi:10.1016/j.cois.2018.05.017

892 99. Menozzi P, Shi MA, Lougarre A, Tang ZH, Fournier D. Mutations of
893 acetylcholinesterase which confer insecticide resistance in Drosophila melanogaster
894 populations. *BMC Evol Biol*. 2004;4:1-7. doi:10.1186/1471-2148-4-4

895 100. Hartley CJ, Newcomb RD, Russell RJ, et al. Amplification of DNA from preserved
896 specimens shows blowflies were preadapted for the rapid evolution of insecticide
897 resistance. *Proc Natl Acad Sci U S A*. 2006;103(23):8757-8762.
898 doi:10.1073/pnas.0509590103

899 101. MacLean RC, Torres-Barceló C, Moxon R. Evaluating evolutionary models of stress-
900 induced mutagenesis in bacteria. *Nat Rev Genet*. 2013;14(3):221-227.
901 doi:10.1038/nrg3415

902 102. Zhou K, Aertsen A, Michiels CW. The role of variable DNA tandem repeats in bacterial
903 adaptation. *FEMS Microbiol Rev*. 2014;38(1):119-141. doi:10.1111/1574-6976.12036

904 103. Jee J, Rasouly A, Shamovsky I, et al. Rates and mechanisms of bacterial
905 mutagenesis from maximum-depth sequencing. *Nature*. 2016;534(7609):693-696.
906 doi:10.1038/nature18313

907 104. Morreall J, Kim A, Liu Y, Degtyareva N, Weiss B, Doetsch PW. Evidence for
908 Retromutagenesis as a Mechanism for Adaptive Mutation in Escherichia coli. *PLoS*
909 *Genet*. 2015;11(8):1-12. doi:10.1371/journal.pgen.1005477

910 105. Sankar TS, Wastuwidyaningtyas BD, Dong Y, Lewis SA, Wang JD. The nature of
911 mutations induced by replication-transcription collisions. *Nature*. 2016;535(7610):178-
912 181. doi:10.1038/nature18316

913 106. Carelli FN, Hayakawa T, Go Y, Imai H, Warnefors M, Kaessmann H. The life history of
914 retrocopies illuminates the evolution of new mammalian genes. *Genome Res*.
915 2016;26(3):301-314. doi:10.1101/gr.198473.115

916 107. Slack A, Thornton PC, Magner DB, Rosenberg SM, Hastings PJ. On the mechanism
917 of gene amplification induced under stress in Escherichia coli. *PLoS Genet*.
918 2006;2(4):385-398. doi:10.1371/journal.pgen.0020048

919 108. Wannarat W, Wannarat W, Motoyama S, Masuda K, Kawamura F, Inaoka T.
920 Tetracycline tolerance mediated by gene amplification in Bacillus subtilis. *Microbiol*
921 *(United Kingdom)*. 2014;160:2474-2480. doi:10.1099/mic.0.081505-0

922 109. Adler M, Anjum M, Berg OG, Andersson DI, Sandegren L. High fitness costs and
923 instability of gene duplications reduce rates of evolution of new genes by duplication-
924 divergence mechanisms. *Mol Biol Evol*. 2014;31(6):1526-1535.

925 doi:10.1093/molbev/msu111

926 110. Näsvall J, Sun L, Roth JR, Andersson DI. Real-time evolution of new genes by
927 innovation, amplification, and divergence. *Science (80-)*. 2012;338(6105):384-387.
928 doi:10.1126/science.1226521

929 111. Graves CJ, Ros VID, Stevenson B, Sniegowski PD, Brisson D. Natural Selection
930 Promotes Antigenic Evolvability. *PLoS Pathog*. 2013;9(11).
931 doi:10.1371/journal.ppat.1003766

932 112. Rotem E, Loinger A, Ronin I, et al. Regulation of phenotypic variability by a threshold-
933 based mechanism underlies bacterial persistence. *Proc Natl Acad Sci U S A*.
934 2010;107(28):12541-12546. doi:10.1073/pnas.1004333107

935 113. Garcia-Bernardo J, Dunlop MJ. Phenotypic Diversity Using Bimodal and Unimodal
936 Expression of Stress Response Proteins. *Biophys J*. 2016;110(10):2278-2287.
937 doi:10.1016/j.bpj.2016.04.012

938 114. Ribas de Pouplana L, Santos MAS, Zhu JH, Farabaugh PJ, Javid B. Protein
939 mistranslation: Friend or foe? *Trends Biochem Sci*. 2014;39(8):355-362.
940 doi:10.1016/j.tibs.2014.06.002

941 115. Gordon AJE, Satory D, Halliday JA, Herman C. Lost in transcription: Transient errors
942 in information transfer. *Curr Opin Microbiol*. 2015;24:80-87.
943 doi:10.1016/j.mib.2015.01.010

944 116. Verheijen BM, van Leeuwen FW. Commentary: The landscape of transcription errors
945 in eukaryotic cells. *Front Genet*. 2017;8(DEC). doi:10.3389/fgene.2017.00219

946 117. Ling J, O'Donoghue P, Söll D. Genetic code flexibility in microorganisms: Novel
947 mechanisms and impact on physiology. *Nat Rev Microbiol*. 2015;13(11):707-721.
948 doi:10.1038/nrmicro3568

949 118. Mordret E, Dahan O, Asraf O, et al. Systematic Detection of Amino Acid Substitutions
950 in Proteomes Reveals Mechanistic Basis of Ribosome Errors and Selection for
951 Translation Fidelity. *Mol Cell*. 2019;75(3):427-441.e5.
952 doi:10.1016/j.molcel.2019.06.041

953 119. Bratulic S, Gerber F, Wagner A. Mistranslation drives the evolution of robustness in
954 TEM-1 β-lactamase. *Proc Natl Acad Sci U S A*. 2015;112(41):12758-12763.
955 doi:10.1073/pnas.1510071112

956 120. Yanagida H, Gispan A, Kadouri N, et al. The Evolutionary Potential of Phenotypic
957 Mutations. *PLoS Genet*. 2015;11(8):1-20. doi:10.1371/journal.pgen.1005445

958 121. Ruiz-Orera J, Messeguer X, Subirana JA, Alba MM. Long non-coding RNAs as a
959 source of new peptides. *Elife*. 2014;3:1-24. doi:10.7554/eLife.03523

960 122. Pauli A, Valen E, Schier AF. Identifying (non-)coding RNAs and small peptides:
961 Challenges and opportunities. *BioEssays*. 2015;37(1):103-112.
962 doi:10.1002/bies.201400103
963 123. Saghatelian A, Couso JP. Discovery and characterization of smORF-encoded
964 bioactive polypeptides. *Nat Chem Biol*. 2015;11(12):909-916.
965 doi:10.1038/nchembio.1964
966 124. Chen S, Krinsky BH, Long M. New genes as drivers of phenotypic evolution. *Nat Rev*
967 *Genet*. 2013;14(9):645-660. doi:10.1038/nrg3521
968 125. McLysaght A, Guerzoni D. New genes from non-coding sequence: The role of de
969 novo protein-coding genes in eukaryotic evolutionary innovation. *Philos Trans R Soc*
970 *B Biol Sci*. 2015;370(1678). doi:10.1098/rstb.2014.0332
971 126. Halfmann R, Jarosz DF, Jones SK, Chang A, Lancaster AK, Lindquist S. Prions are a
972 common mechanism for phenotypic inheritance in wild yeasts. *Nature*.
973 2012;482(7385):363-368. doi:10.1038/nature10875
974 127. Hornung G, Bar-Ziv R, Rosin D, et al. Noise-mean relationship in mutated promoters.
975 *Genome Res*. 2012;22(12):2409-2417. doi:10.1101/gr.139378.112
976 128. Raser JM, O'Shea EK. Control of stochasticity in eukaryotic gene expression. *Science*
977 *(80-)*. 2004;304(5678):1811-1814. doi:10.1126/science.1098641
978 129. Levin BR, Rozen DE. Non-inherited antibiotic resistance. 2006;4(July).
979 130. Allan Drummond D, Wilke CO. The evolutionary consequences of erroneous protein
980 synthesis. *Nat Rev Genet*. 2009;10(10):715-724. doi:10.1038/nrg2662
981 131. Rockah-Shmuel L, Tóth-Petróczy Á, Sela A, Wurtzel O, Sorek R, Tawfik DS.
982 Correlated Occurrence and Bypass of Frame-Shifting Insertion-Deletions (InDels) to
983 Give Functional Proteins. *PLoS Genet*. 2013;9(10). doi:10.1371/journal.pgen.1003882
984 132. van der Woude MW. Phase variation: How to create and coordinate population
985 diversity. *Curr Opin Microbiol*. 2011;14(2):205-211. doi:10.1016/j.mib.2011.01.002
986 133. Atkins JF, Loughran G, Bhatt PR, Firth AE, Baranov P V. Ribosomal frameshifting and
987 transcriptional slippage: From genetic steganography and cryptography to
988 adventitious use. *Nucleic Acids Res*. 2016;44(15):7007-7078. doi:10.1093/nar/gkw530
989 134. Yi X, Dean AM. Phenotypic plasticity as an adaptation to a functional trade-off. *Elife*.
990 2016;5:1-12. doi:10.7554/elife.19307
991 135. Miller SR, Longley R, Hutchins PR, Bauersachs T. Cellular Innovation of the
992 Cyanobacterial Heterocyst by the Adaptive Loss of Plasticity. *Curr Biol*.
993 2020;30(2):344-350.e4. doi:10.1016/j.cub.2019.11.056

994 136. Corl A, Bi K, Luke C, et al. The Genetic Basis of Adaptation following Plastic Changes
995 in Coloration in a Novel Environment. *Curr Biol*. 2018;28(18):2970-2977.e7.
996 doi:10.1016/j.cub.2018.06.075
997 137. Li A, Li L, Zhang Z, et al. Noncoding Variation and Transcriptional Plasticity Promote
998 Thermal Adaptation in Oysters by Altering Energy Metabolism. *Mol Biol Evol*.
999 2021;38(11):5144-5155. doi:10.1093/molbev/msab241
1000 138. Samhita L, Raval PK, Agashe D. Global mistranslation increases cell survival under
1001 stress in Escherichia coli. *PLoS Genet*. 2020;16(3):1-21.
1002 doi:10.1371/journal.pgen.1008654
1003 139. Zheng J, Guo N, Wagner A. Mistranslation Reduces Mutation Load in Evolving
1004 Proteins through Negative Epistasis with DNA Mutations. *Mol Biol Evol*.
1005 2021;38(11):4792-4804. doi:10.1093/molbev/msab206
1006 140. Copley SD. Moonlighting is mainstream: Paradigm adjustment required. *BioEssays*.
1007 2012;34(7):578-588. doi:10.1002/bies.201100191
1008 141. Rauwerdink A, Lunzer M, Devamani T, et al. Evolution of a catalytic mechanism. *Mol*
1009 *Biol Evol*. 2016;33(4):971-979. doi:10.1093/molbev/msv338
1010 142. Amitai G, Gupta RD, Tawfik DS. Latent evolutionary potentials under the neutral
1011 mutational drift of an enzyme. *HFSP J*. 2007;1(1):67. doi:10.2976/1.2739115
1012 143. Bershtein S, Tawfik DS. Ohno's model revisited: Measuring the frequency of
1013 potentially adaptive mutations under various mutational drifts. *Mol Biol Evol*.
1014 2008;25(11):2311-2318. doi:10.1093/molbev/msn174
1015 144. Sabater-Muñoz B, Prats-Escriche M, Montagud-Martínez R, et al. Fitness trade-offs
1016 determine the role of the molecular chaperonin GroEL in buffering mutations. *Mol Biol*
1017 *Evol*. 2015;32(10):2681-2693. doi:10.1093/molbev/msv144
1018 145. Liang H, Lin YS, Li WH. Fast evolution of core promoters in primate genomes. *Mol*
1019 *Biol Evol*. 2008;25(6):1239-1244. doi:10.1093/molbev/msn072
1020 146. Hittinger CT, Carroll SB. Gene duplication and the adaptive evolution of a classic
1021 genetic switch. *Nature*. 2007;449(7163):677-681. doi:10.1038/nature06151
1022 147. Gu Z, Nicolae D, Lu HHS, Li WH. Rapid divergence in expression between duplicate
1023 genes inferred from microarray data. *Trends Genet*. 2002;18(12):609-613.
1024 doi:10.1016/S0168-9525(02)02837-8
1025 148. Makova KD, Li WH. Divergence in the spatial pattern of gene expression between
1026 human duplicate genes. *Genome Res*. 2003;13(7):1638-1645.
1027 doi:10.1101/gr.1133803
1028 149. Lavy T, Yanagida H, Tawfik DS. Gal3 binds Gal80 tighter than Gal1 indicating

1029 adaptive protein changes following duplication. *Mol Biol Evol*. 2016;33(2):472-477.
1030 doi:10.1093/molbev/msv240

1031 150. Noda-Garcia L, Romero Romero ML, Longo LM, Kolodkin-Gal I, Tawfik DS. Bacilli
1032 glutamate dehydrogenases diverged via coevolution of transcription and enzyme
1033 regulation . *EMBO Rep*. 2017;18(7):1139-1149. doi:10.15252/embr.201743990

1034 151. Lynch M. The evolution of multimeric protein assemblages. *Mol Biol Evol*.
1035 2012;29(5):1353-1366. doi:10.1093/molbev/msr300

1036 152. Pillai AS, Chandler SA, Liu Y, et al. Origin of complexity in haemoglobin evolution.
1037 *Nature*. 2020;581(7809):480-485. doi:10.1038/s41586-020-2292-y

1038 153. Garcia-Seisdedos H, Empereur-Mot C, Elad N, Levy ED. Proteins evolve on the edge
1039 of supramolecular self-assembly. *Nature*. 2017;548(7666):244-247.
1040 doi:10.1038/nature23320

1041 154. Marques AC, Vinckenbosch N, Brawand D, Kaessmann H. Functional diversification
1042 of duplicate genes through subcellular adaptation of encoded proteins. *Genome Biol*.
1043 2008;9(3):1-12. doi:10.1186/gb-2008-9-3-r54

1044 155. Liu SL, Pan AQ, Adams KL. Protein subcellular relocalization of duplicated genes in
1045 Arabidopsis. *Genome Biol Evol*. 2014;6(9):2501-2515. doi:10.1093/gbe/evu191

1046 156. Nilsson AI, Koskiniemi S, Eriksson S, Kugelberg E, Hinton JCD, Andersson DI.
1047 Bacterial genome size reduction by experimental evolution. *Proc Natl Acad Sci U S A*.
1048 2005;102(34):12112-12116. doi:10.1073/pnas.0503654102

1049 157. Jensen RA. Enzyme recruitment in evolution of new function. *Annu Rev Microbiol*.
1050 1976;30:409-425. doi:10.1146/annurev.mi.30.100176.002205

1051 158. Weng JK. The evolutionary paths towards complexity: A metabolic perspective. *New*
1052 *Phytol*. 2014;201(4):1141-1149. doi:10.1111/nph.12416

1053 159. Notebaart RA, Szappanos B, Kintses B, et al. Network-level architecture and the
1054 evolutionary potential of underground metabolism. *Proc Natl Acad Sci U S A*.
1055 2014;111(32):11762-11767. doi:10.1073/pnas.1406102111

1056 160. Caetano-Anollés K, Caetano-Anollés G. Structural Phylogenomics Reveals Gradual
1057 Evolutionary Replacement of Abiotic Chemistries by Protein Enzymes in Purine
1058 Metabolism. *PLoS One*. 2013;8(3). doi:10.1371/journal.pone.0059300

1059 161. Keller MA, Turchyn A V., Ralser M. Non-enzymatic glycolysis and pentose phosphate
1060 pathway-like reactions in a plausible Archean ocean. *Mol Syst Biol*. 2014;10(4):1-12.
1061 doi:10.1002/msb.20145228

1062 162. Laurino P, Tawfik DS. Spontaneous Emergence of S -Adenosylmethionine and the
1063 Evolution of Methylation . *Angew Chemie*. 2017;129(1):349-351.

1064 doi:10.1002/ange.201609615

1065 163. Lazcano A, Miller SL. On the origin of metabolic pathways. *J Mol Evol*.
1066 1999;49(4):424-431. doi:10.1007/PL00006565

1067 164. Cornish-Bowden A, Cárdenas ML. Life before LUCA. *J Theor Biol*. 2017;434:68-74.
1068 doi:10.1016/j.jtbi.2017.05.023

1069 165. Piette BMAG, Heddle JG. A Peptide–Nucleic Acid Replicator Origin for Life. *Trends*
1070 *Ecol Evol*. 2020;35(5):397-406. doi:10.1016/j.tree.2020.01.001

1071 166. Chatterjee S, Yadav S. The origin of prebiotic information system in the peptide/RNA
1072 world: A simulation model of the evolution of translation and the genetic code. *Life*.
1073 2019;9(1). doi:10.3390/life9010025

1074 167. Rode BM. Peptides and the origin of life. *Peptides*. 1999;20(6):773-786.
1075 doi:10.1016/S0196-9781(99)00062-5

1076 168. Ikehara K. Possible steps to the emergence of life: The [GADV]-protein world
1077 hypothesis. *Chem Rec*. 2005;5(2):107-118. doi:10.1002/tcr.20037

1078 169. Fried SD, Fujishima K, Makarov M, Cherepashuk I, Hlouchova K. Peptides before and
1079 during the nucleotide world: An origins story emphasizing cooperation between
1080 proteins and nucleic acids. *J R Soc Interface*. 2022;19(187).
1081 doi:10.1098/rsif.2021.0641

1082 170. Maury CPJ. Amyloid and the origin of life: self-replicating catalytic amyloids as
1083 prebiotic informational and protometabolic entities. *Cell Mol Life Sci*. 2018;75(9):1499-
1084 1507. doi:10.1007/s00018-018-2797-9

1085 171. Friedmann MP, Torbeev V, Zelenay V, Sobol A, Greenwald J, Riek R. Towards
1086 prebiotic catalytic amyloids using high throughput screening. *PLoS One*.
1087 2015;10(12):1-16. doi:10.1371/journal.pone.0143948

1088 172. Romero Romero ML, Yang F, Lin YR, et al. Simple yet functional phosphate-loop
1089 proteins. *Proc Natl Acad Sci U S A*. 2018;115(51):E11943-E11950.
1090 doi:10.1073/pnas.1812400115

1091 173. Longo LM, Petrovic D, Kamerlin SCL, Tawfik DS. Short and simple sequences
1092 favored the emergence of N-helix phospho-ligand binding sites in the first enzymes.
1093 *Proc Natl Acad Sci U S A*. 2020;117(10):5310-5318. doi:10.1073/pnas.1911742117

1094 174. Medvedev KE, Kinch LN, Schaeffer RD, Grishin N V. *Functional Analysis of*
1095 *Rossmann-like Domains Reveals Convergent Evolution of Topology and Reaction*
1096 *Pathways*. Vol 15.; 2019. doi:10.1371/journal.pcbi.1007569

1097 175. Söding J, Lupas AN. More than the sum of their parts: On the evolution of proteins
1098 from peptides. *BioEssays*. 2003;25(9):837-846. doi:10.1002/bies.10321

1099 176. Ferruz N, Lobos F, Lemm D, et al. Identification and Analysis of Natural Building
1100 Blocks for Evolution-Guided Fragment-Based Protein Design. *J Mol Biol*.
1101 2020;432(13):3898-3914. doi:10.1016/j.jmb.2020.04.013

1102 177. Alva V, Söding J, Lupas AN. A vocabulary of ancient peptides at the origin of folded
1103 proteins. *Elife*. 2015;4:1-19. doi:10.7554/elife.09410

1104 178. Van Oss SB, Carvunis AR. De novo gene birth. *PLoS Genet*. 2019;15(5):1-23.
1105 doi:10.1371/journal.pgen.1008160

1106 179. Wilson BA, Foy SG, Neme R, Masel J. Young genes are highly disordered as
1107 predicted by the preadaptation hypothesis of de novo gene birth. *Nat Ecol Evol*.
1108 2017;1(6):1-19. doi:10.1038/s41559-017-0146

1109 180. Bungard D, Copple JS, Yan J, et al. Foldability of a Natural De Novo Evolved Protein.
1110 *Structure*. 2017;25(11):1687-1696.e4. doi:10.1016/j.str.2017.09.006

1111 181. Tawfik DS. Accuracy-rate tradeoffs: How do enzymes meet demands of selectivity
1112 and catalytic efficiency? *Curr Opin Chem Biol*. 2014;21:73-80.
1113 doi:10.1016/j.cbpa.2014.05.008

1114 182. Innan H, Kondrashov F. The evolution of gene duplications: Classifying and
1115 distinguishing between models. *Nat Rev Genet*. 2010;11(2):97-108.
1116 doi:10.1038/nrg2689

1117 183. Noda-García L, Camacho-Zarco AR, Medina-Ruíz S, et al. Evolution of substrate
1118 specificity in a recipient's enzyme following horizontal gene transfer. *Mol Biol Evol*.
1119 2013;30(9):2024-2034. doi:10.1093/molbev/mst115

1120 184. Dellus-Gur E, Elias M, Caselli E, et al. Negative epistasis and evolvability in TEM-1 β-
1121 lactamase - The thin line between an enzyme's conformational freedom and disorder.
1122 *J Mol Biol*. 2015;427(14):2396-2409. doi:10.1016/j.jmb.2015.05.011

1123 185. Starr TN, Thornton JW. Epistasis in protein evolution. *Protein Sci*. 2016;25:1204-
1124 1218. doi:10.1002/pro.2897

1125 186. Miton CM, Tokuriki N. How mutational epistasis impairs predictability in protein
1126 evolution and design. *Protein Sci*. 2016;25:1260-1272. doi:10.1002/pro.2876

1127 187. Harms MJ, Thornton JW. Historical contingency and its biophysical basis in
1128 glucocorticoid receptor evolution. *Nature*. 2014;512(7513):203-207.
1129 doi:10.1038/nature13410

1130 188. Anderson DW, McKeown AN, Thornton JW. Intermolecular epistasis shaped the
1131 function and evolution of an ancient transcription factor and its DNA binding sites.
1132 *Elife*. 2015;4(JUNE2015):1-26. doi:10.7554/eLife.07864

1133 189. Dellus-Gur E, Toth-Petroczy A, Elias M, Tawfik DS. What makes a protein fold

1134 amenable to functional innovation? fold polarity and stability trade-offs. *J Mol Biol*.
1135 2013;425(14):2609-2621. doi:10.1016/j.jmb.2013.03.033

1136 190. Payne JL, Wagner A. The robustness and evolvability of transcription factor binding
1137 sites. *Science (80-)*. 2014;343(6173):875-877. doi:10.1126/science.1249046

1138 191. Sikosek T, Chan HS. Biophysics of protein evolution and evolutionary protein
1139 biophysics. *J R Soc Interface*. 2014;11(100). doi:10.1098/rsif.2014.0419

1140 192. Liskova V, Bednar D, Prudnikova T, et al. Balancing the stability-activity trade-off by
1141 fine-tuning dehalogenase access tunnels. *ChemCatChem*. 2015;7(4):648-659.
1142 doi:10.1002/cctc.201402792

1143 193. Newton MS, Arcus VL, Patrick WM. Rapid bursts and slow declines: On the possible
1144 evolutionary trajectories of enzymes. *J R Soc Interface*. 2015;12(107).
1145 doi:10.1098/rsif.2015.0036

1146

1147