An orphan protein drove the ecological expansion of nitrogen fixation

Bruno Cuevas-Zuviría[1,2,*], Amanda K. Garcia[1,*], Brooke M. Carruthers[1], Betül Kaçar[1,**]

[1]Department of Bacteriology, University of Wisconsin-Madison, Madison, WI, USA

[2]Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid, Madrid, Spain

*These authors contributed equally to this work

**To whom correspondence should be addressed: bkacar@wisc.edu

ABSTRACT

Nitrogenases have catalyzed biological nitrogen fixation for billions of years and revolutionized planet Earth by supplying essential nitrogen to the biosphere. How these molecular machines were built and distributed by microbial processes and a shifting geochemical landscape remains an open question. Here, we probe the birth and functional evolution of the G-subunit, an integral, Precambrian component of certain nitrogenases that makes its appearance midway through the history of nitrogen fixation. Our results establish the G-subunit as an unprecedented molecular novelty that was integrated into preexisting interaction pathways for nitrogenase metallocluster incorporation. In the wake of Earth oxygenation, the G-subunit primed nitrogenases and their microbial hosts to exploit newly diversified geochemical environments. The history of the nitrogenase G-subunit showcases the elaborate tango between contingent evolutionary events and environmental conditions that together open an adaptive frontier for molecular innovations with global consequences.

INTRODUCTION

Over billions of years, life has generated an enormous wealth of biomolecular diversity, producing an estimated 10^5 extant protein families[1] and a much larger but inaccessible multitude of extinct proteins. In parallel, biology has become a defining component of the Earth system, which has experienced tremendous revolutions in climate and biogeochemistry that are largely attributable to the molecular innovations of life itself[2]. A unified understanding of the Earth system requires investigating how the proteins and metabolic processes that power planetary phenomena have been built and distributed.

In many cases, nature has evidently been a tinkerer[3], generating novel functions through the progressive duplication, accretion, and recombination of pre-existing protein elements[4-7]. In rarer cases, complete novelty can arise through the *de novo* birth of genes from ancestral, non-genic DNA[4,8]. This scenario represents a more difficult evolutionary hurdle due to the improbability that formerly non-genic DNA would produce a structurally or functionally coherent peptide[9]. Retracing these different evolutionary events in the histories of biogeochemically critical enzymes is itself a considerable challenge given their antiquity, with many of these enzymes, including those required for essential elemental cycling, likely having evolved more than three billion years ago[2,10,11]. Nevertheless, from an evolutionary perspective, it is of considerable interest to uncover the unique conditions under which the benefits of molecular innovations in these systems have outweighed their inherent costs[4] and expanded the planet's biogeochemical repertoire.

Here, we demonstrate an unusual case of molecular novelty in the evolutionary history of biological nitrogen fixation. For >3 billion years[12,13], nitrogenase metalloenzymes have provided the sole biochemical gateway for essential nitrogen into the Earth biosphere, consequently transforming the planet[10,14-16]. This family is today represented by three homologous isozymes—molybdenum- (Nif), vanadium- (Vnf), and iron- (Anf) nitrogenases—that differ in their metal dependencies but otherwise share a core, two-component oxidoreductase architecture[17-19]. The antiquity and biological significance of this architecture is highlighted by its conservation in distantly related enzymes within the nitrogenase superfamily[20], the origin of which likely dates back to the last universal common ancestor[20,21]. Nevertheless, nitrogenase evolution appears to have since uniquely elaborated upon this structural scheme. A striking distinction in all experimentally characterized Vnf and Anf nitrogenases is an additional, functionally enigmatic protein subunit that is not present in Nif, referred to here as the "G-subunit"[22-25]. However, the

origin and role of the G-subunit in the diversification of Vnf and Anf nitrogenases—and thus in the ability of biological nitrogen fixation to exploit a different suite of geochemical ingredients—remains unknown. Given the ancient transitions in redox-sensitive metal availabilities triggered by Earth surface oxygenation[26,27], elucidating these evolutionary features would elucidate the roles of environmental conditions versus molecular evolutionary events in the ecological diversification of biological nitrogen fixation[26,28,29].

In the present study, we replay the sequence of events that gave rise to the nitrogenase G-subunit. We interrogate the genetic and environmental context of its emergence, recruitment, and functional significance. Our findings reveal the unprecedented nature of G-subunit sequence and structure within broader protein diversity, plausibly derived from the serendipitous, *de novo* birth of an orphan gene. By reconstructing its ancestral interactions, we find that G-subunit recruitment >1.5 billion years ago likely exploited and improved the specificity of existing pathways that shape nitrogenase metal usage. The G-subunit thus represents an unusual molecular novelty that, in the wake of Earth oxygenation and its impact on surface geochemistry, fueled the proliferation of nitrogen fixation strategies to diverse microorganisms and ecological niches.

RESULTS

The nitrogenase G-subunit is an ancient orphan protein

We began our evolutionary study of the nitrogenase G-subunit by constraining its place within the broader diversity of protein sequence and structure. The G-subunit protein is a small, structural element of the larger, two-component nitrogenase complex. The first component is a H-subunit homodimer that delivers electrons to the second, catalytic component, which, in Nif, is a D- and K-subunit heterotetramer but, in Vnf/Anf (collectively referred to as “alternative nitrogenases”), is a D-, G-, and K-subunit heterohexamer[22] (**Figure 1A**). Thus, the G-subunit is a primary distinction in the otherwise structurally comparable Nif, Vnf, and Anf complexes. To date, the sole crystallographic structure of the G-subunit is from *A. vinelandii* Vnf[22], a 113-amino-acid, four-helix protein (**Figure 1B**). In Vnf (and presumably Anf), the G-subunit neighbors the D-subunit that houses the active-site metallocluster where N_2 is bound and reduced[30]. The metal content of the metallocluster differs between Nif, Vnf, and Anf nitrogenase isozymes, incorporating Fe and Mo (“FeMo-co”), Fe and V (“FeV-co”), or only Fe (“FeFe-co”), respectively. Finally, though sharing a

common ancestry[28,31,32] and core N_2-reduction mechanism[33], Nif, Vnf, and Anf nitrogenases vary in their kinetic behavior and reactivities to different substrates[19,33-35].

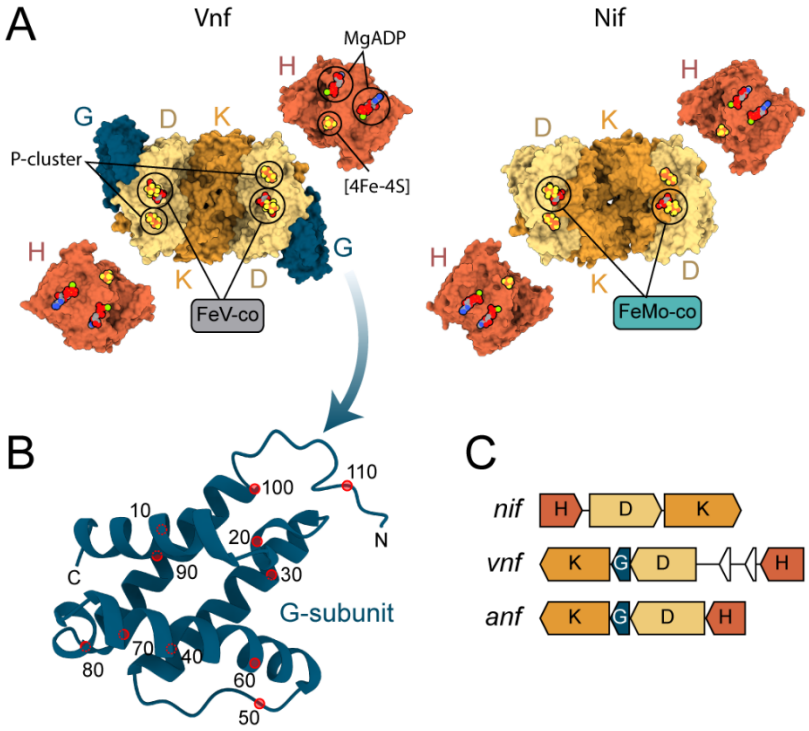


Figure 1. Structure and genetics of the nitrogenase G-subunit. A) *Left, crystallographic structure of the extant A. vinelandii Vnf nitrogenase complex (VnfH, PDB 6Q93[36]; VnfDGK, PDB 5N6Y[22]) containing the G-subunit (dark blue). The Anf nitrogenase is inferred to be comparable in its multimeric structure, binding FeFe-co instead of FeV-co. Right, crystallographic structure of the extant Nif nitrogenase complex (NifH, PDB 1M34[37]; NifDK, PDB 3U7Q[38]) lacking the G-subunit.* ***B)*** *Secondary structure of the Vnf G-subunit.* ***C)*** *Arrangement of the Nif, Vnf, and Anf structural genes in A. vinelandii. Vnf/AnfG (dark blue) is located between Vnf/AnfD and Vnf/AnfK genes.*

In addition to surveying the diversity of members of the G-subunit protein family, we sought to identify distant homologs that might provide insights into G-subunit ancestry and function. Though genetic studies have established the G-subunit as essential for an active, N_2-reducing Vnf/Anf nitrogenase complex[24,39], its precise functional contribution remains enigmatic. We implemented a homology search workflow that included both sequence-based (BLASTP, TBLASTN[40], PSI-BLAST[41], HMMER[42]) and structure-based (Foldseek[43]) methods appropriate for detection of distantly related proteins. Our search returned >400 canonical G-subunit proteins, but, notably,

114 we did not find convincing similarity to any other protein. PSI-BLAST and HMMER searches
115 together returned <10 hits that we could exclude from the canonical G-subunit protein family (i.e.,
116 were not encoded by genes within *vnf*/*anf* structural operons) (**Table 1, Figure S1**). In both cases,
117 only short sequence regions of scattered taxonomic affinities aligned to our query with poor
118 statistical significance. We were thus unable to rule out spurious matches. Additionally, we used
119 the structure-based method, FoldSeek, leveraging the greater conservation of protein structure
120 compared to sequence[44] in homology detection. We identified two uncharacterized, AlphaFold-
121 predicted structures that are not canonical G-subunit proteins but nevertheless share its four-helix
122 structural arrangement (**Table 1, Figure S2, S3**). However, sequence-level similarity to G-subunit
123 proteins is negligible and, given the lack of functional annotations for these hits, their relevance
124 to early G-subunit evolution is unclear. Finally, we note that we did not detect sequence or
125 structural similarity to NafY and CowN proteins, both of which have previously been proposed to
126 resemble the G-subunit in their interactions with the nitrogenase complex[22,45,46].
127
128 If the G-subunit shares an ancient ancestry with another extant protein family, no strong vestige
129 of this evolutionary relationship appears presently detectable. The absence of distant homologs
130 to the G-subunit family is comparable to the case of "orphan genes," which are individual genes
131 that either have no resemblance to other sequences or have few, taxonomically restricted
132 relatives[8,47]. Proposed mechanisms for the emergence of orphan genes include duplication
133 followed by rapid divergence and *de novo* birth from previously non-genic DNA[4,8]. The latter has
134 more recently been recognized as a significant source of orphan genes[48-50] and thus would
135 plausibly explain G-subunit origins. Regardless of the specific originating mechanism, our results
136 indicate that the G-subunit family is a molecular innovation as unique to nitrogenases as it is to
137 broader protein diversity.
138

139 **Table 1. Summary of G-subunit homolog search results. *Template modeling score for structural alignments.**

Method	Database	Total hits (including canonical G-subunit proteins)	Distant hits	E-value range of distant hits	TM-score* of distant hits	Annotation of distant hits	Taxonomy of distant hits
BLASTP	NCBI non-redundant proteins (nr)	433	0	n/a	n/a	n/a	n/a
TBLASTN	NCBI nucleotide collection (nr/nt)	205	0	n/a	n/a	n/a	n/a
PSI-BLAST	NCBI non-redundant proteins (nr)	446	7	0.026-0.064	n/a	TIGR04141 family sporadically distributed protein, Uncharacterized protein	*Pseudomonas* spp.
HMMER	UniProt Reference Proteomes	94	1	0.061	n/a	Uncharacterized protein	*Coregonus* sp.
FoldSeek	Protein Data Bank	3	0	n/a	n/a	n/a	n/a
FoldSeek	AlphaFold-Uniprot/Swiss-Prot	39	2	5.2-8.7	0.52-0.54	Uncharacterized protein	*Vibrio vulnificus, Deinococcus grandis*

140

141 **Invention of the G-subunit coincided with expansion of nitrogenase metal dependence and**
142 **host ecology**
143
144 We reconstructed phylogenies of nitrogenase G-subunit proteins and examined their genetic
145 distribution across extant genomes and microbial host environments. We identified G-subunit
146 homologs exclusively from *vnf* and *anf* gene clusters in bacterial and archaeal genomes (similarly,
147 all identified *vnf*/*anf* clusters include the G-subunit), confirming that the G-subunit is a significant
148 genetic distinction between Vnf/Anf and Nif nitrogenases[51]. In all analyzed clusters, *vnf/anfG*
149 genes are located between *vnf/anfD and vnf/anfK* loci (**Figure 1C**). By contrast, *nifD* and *nifK* are
150 universally contiguous with small intergenic distances (median ≈ 25 bp) (**Figure S4**). Finally, all
151 complete genomes harboring *vnf* and/or *anf* nitrogenase genes in our dataset also possess *nif*
152 genes. Therefore, the presence of the G-subunit correlates with the capacity of a host diazotroph
153 to express at least two different nitrogenase isozymes with distinct metal dependencies. We built
154 two maximum-likelihood G-subunit phylogenies: the first from individual G-subunit protein
155 sequences and the second from G-subunit sequences concatenated with other nitrogenase
156 subunit sequences. Both trees show separate clustering of Vnf and Anf sequences, though deep
157 branch support in the concatenated tree is significantly improved (**Figure 2, S5, S6**).
158
159 The overall topology of our concatenated nitrogenase tree mirrors previous phylogenetic analyses
160 of nitrogenase proteins reconstructed without G-subunit sequences. Nesting of Vnf and Anf
161 clades within Nif sequences indicates that Vnf/Anf are more recently evolved[28,32], which, by
162 extension, establishes the G-subunit as the most recently evolved structural component of the
163 nitrogenase complex. Minimum age estimates of the Vnf/Anf clade range from ~1.5 to 2.5 Ga,
164 based on the timing of horizontal gene transfer events that pervade nitrogenase evolutionary
165 history[13]. These estimates indicate that the Precambrian origin of the G-subunit roughly coincided
166 with the initial accumulation of oxygen in the Earth surface environment[27], the earliest whiffs of
167 which date back to ~2.5 Ga[52].
168
169 Our analyses demonstrate that the origin of the G-subunit was associated both with the gain of
170 novel metal dependence and the ecological expansion of nitrogen fixation following the rise of
171 atmospheric oxygen. Vnf and Anf sequences specifically branch from within the "Group III"
172 nitrogenase lineage[31], which is additionally populated by Nif homologs of either experimentally
173 characterized or computationally predicted Mo-dependence[28,53]. Group III Nif are exclusively
174 hosted by anaerobic taxa (primarily methanogenic Euryarchaeota and Firmicutes), many of which

are probable thermophiles or hyperthermophiles (optimal growth temperatures predicted from genome content, see **Materials and Methods**) (**Figure 2, Supplementary Tables**). By contrast, Vnf/Anf are hosted by comparatively more taxonomically diverse and mesophilic archaea and bacteria that also harbor "Group I and II" Nif nitrogenases[31], including the well-studied diazotrophic models like *A. vinelandii*, *Clostridium pasteurianum,* and *Rhodopseudomonas palustris*. Vnf/Anf are not hosted by organisms that harbor Group III Nif, despite Vnf/Anf emerging from within the Group III lineage. Within Group III, only Vnf and Anf clades contain sequences from aerobic taxa, including cyanobacteria which appear to only host Nif and Vnf nitrogenases. Together, these data indicate that G-subunit containing Vnf and Anf nitrogenases were able to proliferate into ecological niches inaccessible to earlier diverged Group III Nif.

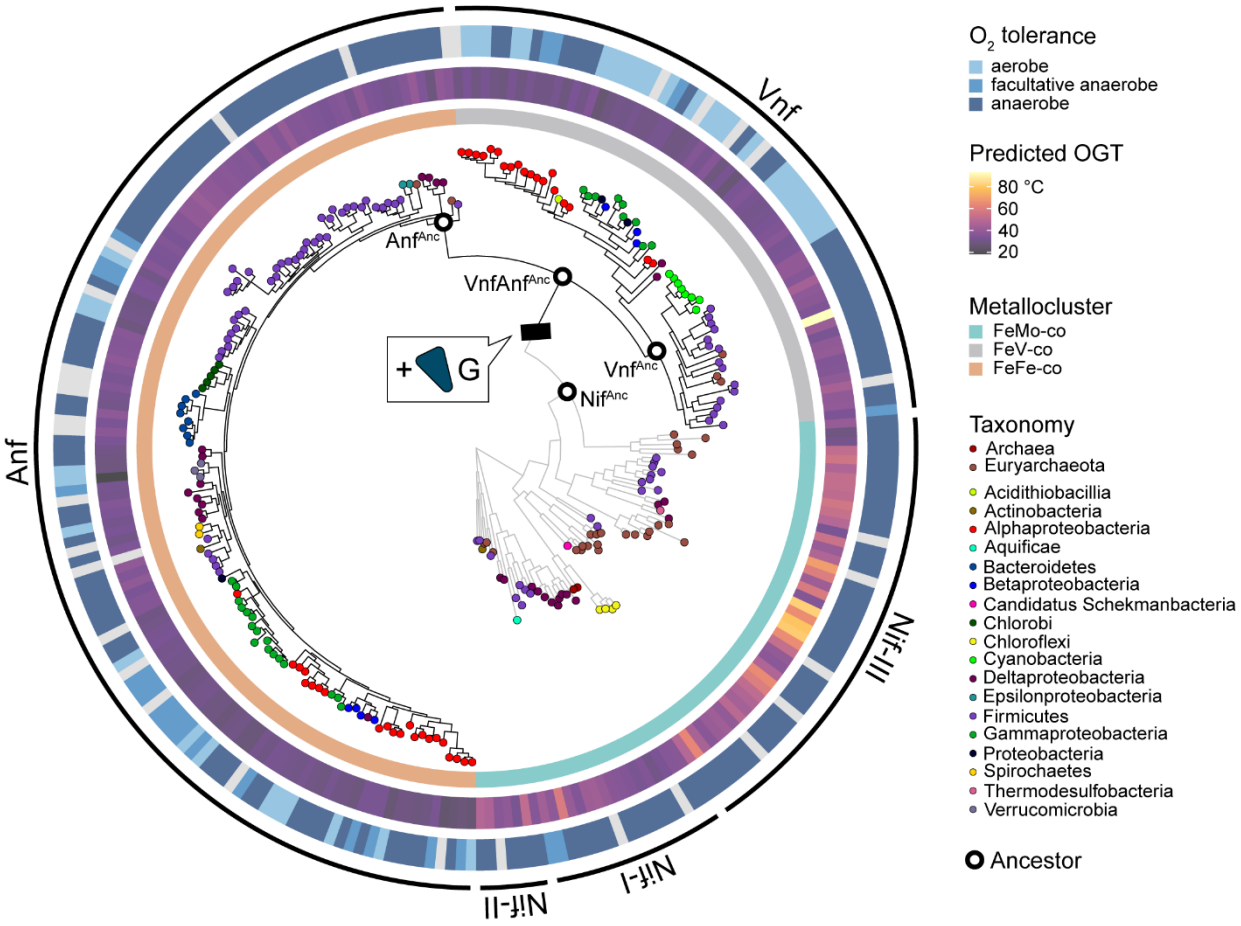


Figure 2. Evolutionary and environmental history of the nitrogenases and the G-subunit. *Maximum-likelihood phylogeny built from concatenated nitrogenase subunit proteins. Oxygen tolerance, predicted optimal growth temperatures (OGT), and taxonomy of represented host microbes, as well as the metal content of the active-site metallocluster, are mapped to the*

phylogeny (light grey bars for O_2 tolerance and predicted OGT indicate missing data). Nif clades are labeled according to the nomenclature used by Raymond et al.[31] (e.g., Nif-I = Group I Nif, etc.). The G-subunit origin event is indicated by a black hash mark. Branches that track the emergence of the G-subunit and its diversification into Vnf and Anf groups are shown in black, whereas branches lacking the G-subunit are shown in grey. Ancestral nodes (Nif^{Anc}, $VnfAnf^{Anc}$, Vnf^{Anc}, Anf^{Anc}) targeted for sequence reconstruction and analysis are highlighted by bold outlined, white circles.

An ancient assembly pathway preconditioned early nitrogenases for G-subunit interaction

To replay the first evolutionary steps in the birth of the G-subunit, we reconstructed and analyzed ancestral nitrogenase protein sequences. We examined the degree to which ancestral protein surface features were remodeled to accommodate a new subunit, with the expectation that G-subunit interactions would have driven the evolution of complementary residues within the interface. Because the G-subunit is exclusive to Vnf and Anf nitrogenases, the most parsimonious evolutionary scenario is that it emerged along the branch leading to the Vnf/Anf last common ancestor, $VnfAnf^{Anc}$ (**Figure 2**). We therefore compared sequences and structures of nitrogenase ancestors reconstructed both immediately before (Nif^{Anc}) and after ($VnfAnf^{Anc}$) gain of the G-subunit.

We mapped the interface between ancestral G- and D-subunit proteins ("G-D interface") by building its residue interaction network from the predicted $VnfAnf^{Anc}$ structure (**Figure 3A-3D**). Nearly all well-conserved, G-subunit residues lie within the G-D interface and contribute significantly to its stability (**Figure S7**). The few exceptions are conserved residues distal from the interface with side chains oriented towards the interior of the helical bundle, which likely stabilize its globular arrangement.

We tracked the evolution of D-subunit residues through the emergence of the ancestral G-D interface. We identified 42 D-subunit sites interacting with the G-subunit following its initial recruitment in ancestral $VnfAnf^{Anc}$ (**Supplementary Tables**). Of the 42 sites, 26 are well-conserved among extant Vnf/Anf nitrogenases and form a core interaction patch that was likely integral to the early G-D interface (**Figure 3E, 3F**). We inspected these 26 sites and found that nine sites evolved (i.e., were substituted) alongside G-subunit emergence, imposing subtle changes to the electrostatic surface potential of the $VnfAnf^{Anc}$ interaction patch (**Figure S8, S9**).

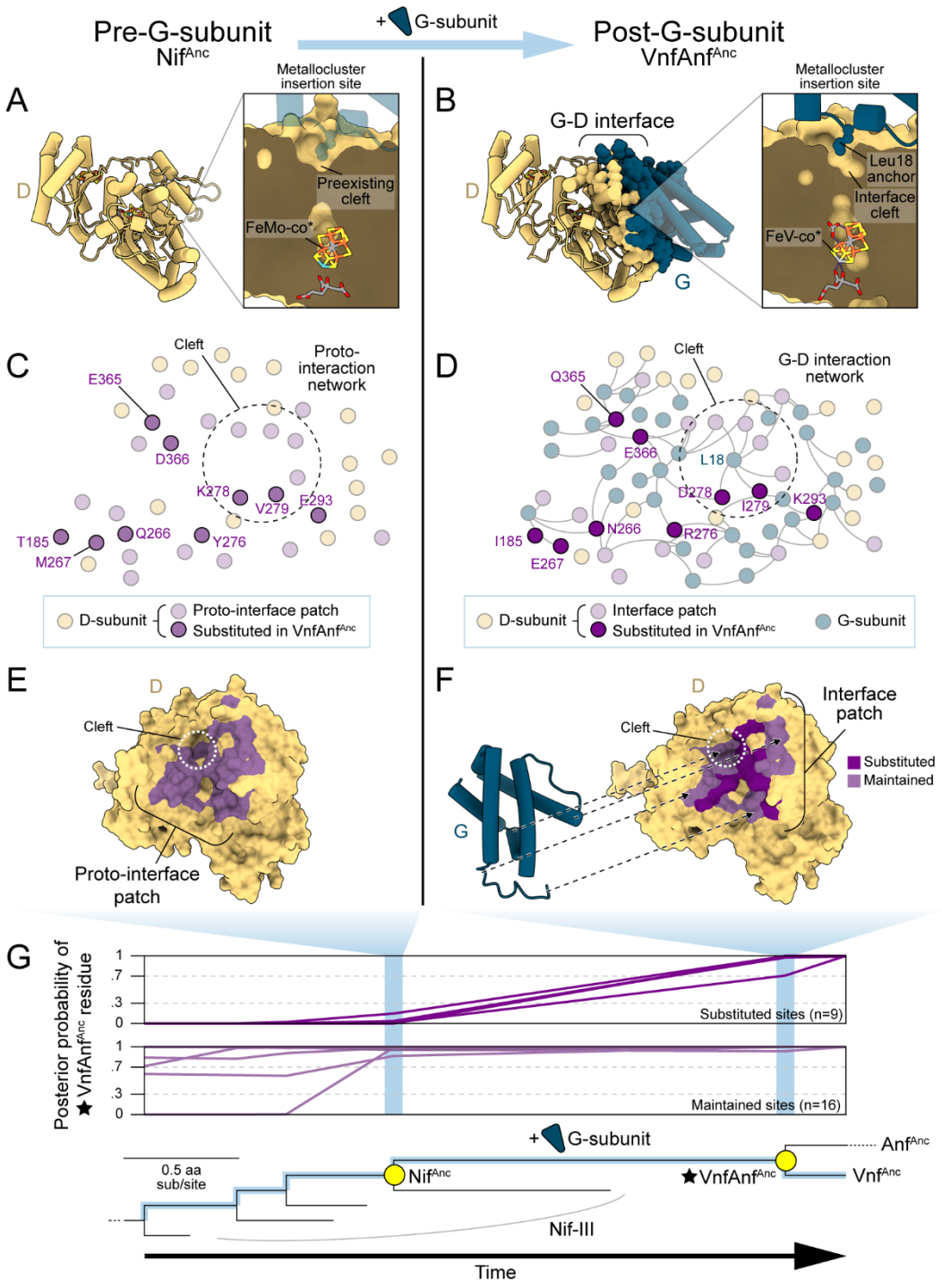


Figure 3. Birth of the nitrogenase G-subunit interface. A) *Predicted Nif^{Anc} D-subunit structure, prior to G-subunit emergence. Inset shows cross-section at the putative metallocluster insertion site, the entrance of which forms a "cleft" in the D-subunit surface. Superimposed G-subunit*

structure (transparent) highlights relative position of the cleft. ***B)*** *Predicted VnfAnf^{Anc} D- and G-subunit structures, following G-subunit emergence, shown in the same view as A). Side chain of the G-subunit Leu18 residue is anchored within the D-subunit cleft.* ***A-B)*** *Heterometal content for the ancestor metalloclusters is inferred.* ***C)*** *Nif^{Anc} G-D proto-interaction network. Purple residues are located within an interface patch conserved in extant Vnf/Anf. Residues that are later substituted in VnfAnf^{Anc} are shown bold and labeled to permit comparison with D). Residues within the cleft are circled.* ***D)*** *VnfAnf^{Anc} G-D interaction network, shown in the same view as C). Residues newly substituted in VnfAnf^{Anc} are shown in dark purple and labeled.* ***E)*** *Surface view of the Nif^{Anc} D-subunit proto-interface patch (purple).* ***F)*** *Surface view of the VnfAnf^{Anc} D-subunit interface patch, shown in the same view as E). Residues that are maintained between NifAnc and VnfAnfAnc are shown in purple, and those that are substituted are shown in dark purple.* ***G)*** *Evolution of residues within the interaction patch through the emergence of the G-subunit. For substituted sites, the posterior probability of the most likely VnfAnf^{Anc} residue is <.3 in Nif^{Anc} and >.7 in VnfAnf^{Anc} and, for maintained sites, >.7 in Nif^{Anc} and >.7 in VnfAnf^{Anc}.*

Intriguingly, we find that multiple ancestral D-subunit surface features appear to have been preconditioned for G-subunit binding, collectively forming a "proto-interface" upon which interactions with the newly evolved G-subunit is built. Out of 26 residues within the core, 16 (~62%) show that the D-subunit interaction patch had already become VnfAnf^{Anc}-like prior to G-subunit emergence (**Figure 3F,G**). This residue-level conservation translates to maintained hydrophobicity (median apolar fraction ≈ .2) of the (proto-)interaction patch both before and after G-subunit recruitment (**Figure S10**). Further, shape complementarity as an integral feature of the G-D interface appears to have preceded the G-subunit. In *A. vinelandii* Vnf[22], a cleft within the D-subunit interaction patch is occupied by a strictly conserved G-subunit residue, Leu18. This residue is among the most well-connected within the G-D interaction network (**Figure 3A-D**), is critical for interface stability (**Figure S7**), and is among the most allosterically active, capable of generating long-distance effects that reach across opposite ends of the nitrogenase complex (**Figure S11**). These effects reflect the importance of the G-D interface in controlling the global motions of the nitrogenase complex and evoke similar, long-distance structural perturbations that are induced through the nitrogenase catalytic cycle[54,55]. The cleft itself provides an important anchor point for G-subunit binding and is present in both ancestral and representative extant nitrogenase structures (**Figure 3A,B, Figure 4**).

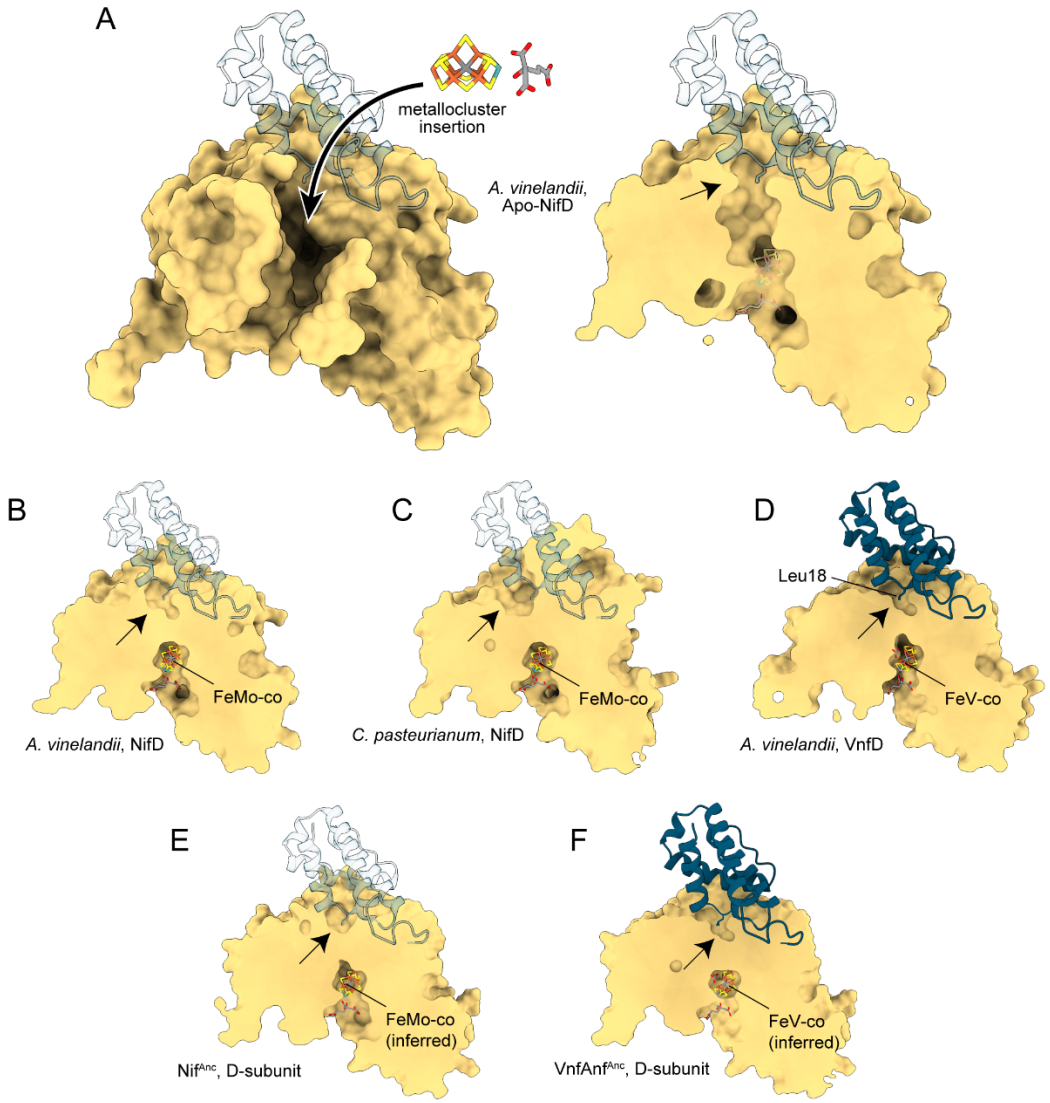


Figure 4. Conservation of the D-subunit cofactor insertion cleft across extant and ancestral nitrogenases. A) *Crystallographic structure of the A. vinelandii apo-NifD protein (PDB 1L5H). Cross-section view (right) reveals the proposed metallocluster insertion site[56], the entrance of which structurally aligns with the G-D interface of VnfD. Aligned, cross-section views are displayed for extant* ***B)*** *A. vinelandii NifD (PDB 3U7Q),* ***C)*** *Clostridium pasteurianum NifD (PDB 4WES), and* ***D)*** *A. vinelandii VnfD (PDB 5N6Y) crystallographic structures as well as predicted, ancestral* ***E)*** *Nif^{Anc} D-subunit and* ***F)*** *VnfAnf^{Anc} D-subunit structures.* ***(A-C, E)*** *A. vinelandii VnfG (PDB 5N6Y; transparent) is superimposed to enable comparison across nitrogenases with and without native G-subunit proteins.* ***(A-F)*** *Small black arrow indicates position of the proposed cofactor insertion cleft.*

276

277 The functional significance of the D-subunit cleft provides an explanation for why the interface
278 patch resisted significant remodeling with G-subunit addition, as well as implicates an adaptive
279 role for the G-subunit. In extant Nif, the cleft aligns with the entrance to a proposed nitrogenase
280 metallocluster insertion site, which, in the apo-NifD protein, adopts an open conformation to permit
281 cluster loading[56]. Given the consistent presence of the cleft across Nif, Vnf, and Anf nitrogenases,
282 a similar mechanism can be predicted for Vnf and Anf and is likely an ancient, universal feature
283 of nitrogenase maturation that predates the G-subunit and was maintained due to its essentiality.
284 In fact, we find that 12 interface residues surrounding the insertion site are well-conserved in
285 extant Group III Nif clades that diverge prior to the origin of the G-subunit, and five of these are
286 either universally or nearly universally conserved among all nitrogenases (**Figure S12**). Structural
287 alignments indicate that the G-subunit caps the opening of the insertion site, with the side chain
288 of the conserved Leu18 residue oriented into the cleft and toward the loaded active site. An
289 unusual feature of the G-subunit, a carboxyl-terminal tail that terminates in a strictly conserved
290 tyrosine residue (Tyr113), bridges the presumed opening of the apo-D-subunit protein and
291 neighbors universally conserved D-subunit residues. Together, these features implicate a role for
292 the G-subunit in nitrogenase assembly, which is supported by experimental evidence that VnfG
293 and VFe-co associate *in vitro*, as well as of the requirement of both VnfG and VFe-co to stabilize
294 an active complex[24]. Conservation of G-subunit residues surrounding the cleft (and thus the
295 metallocluster insertion site) suggests that this putative role was early-evolved and maintained
296 through its evolution.
297

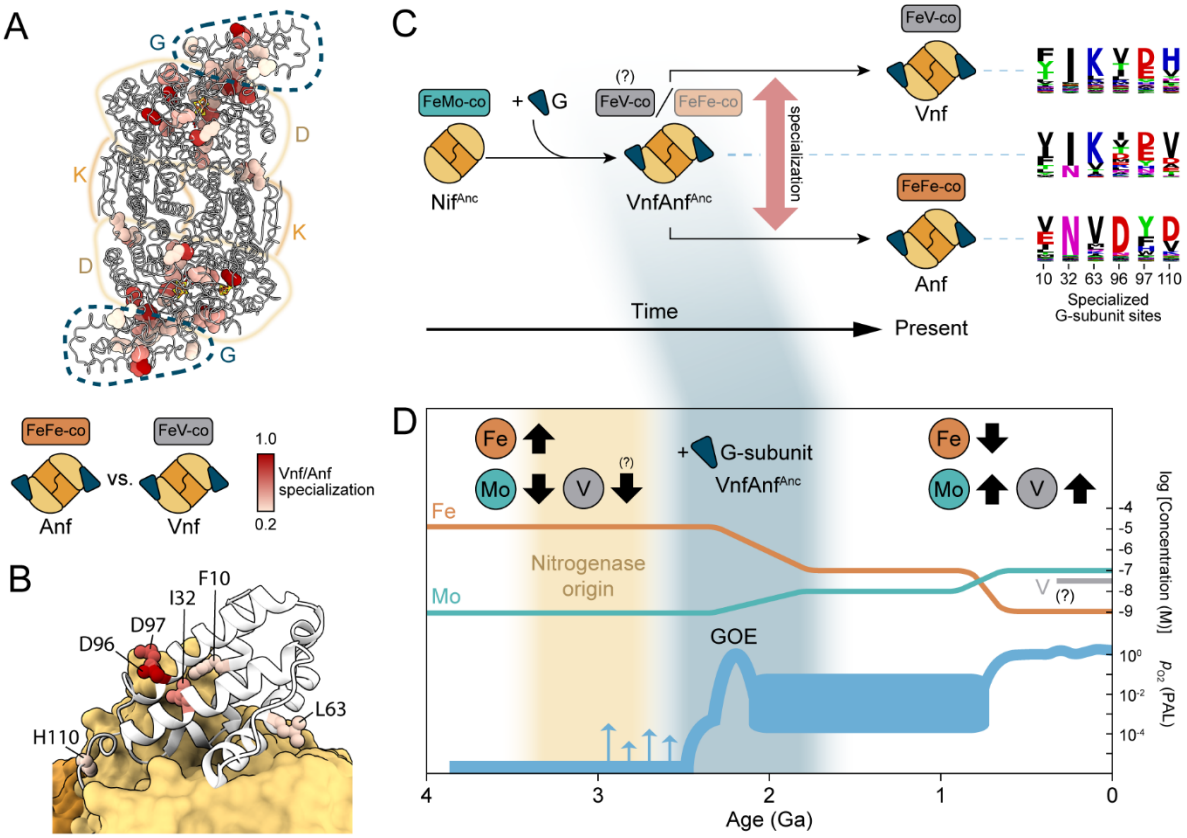
298 **The G-subunit confers specialization between Vnf and Anf nitrogenases**

299

300 We assessed to what degree sequence variation within the G-subunit is capable of driving the
301 functional distinctions between Vnf and Anf isozymes, which are each dependent on distinct
302 metalloclusters and exhibit varied N_2-reduction efficiencies and reactivities to different chemical
303 substrates[19,33-35]. We identified protein sites that are "specialized" between Vnf and Anf (i.e.,
304 diverge between Vnf and Anf but are otherwise conserved in each of the two groups), and
305 therefore most likely drive functional variation. We find the degree of specialization (relative to
306 protein size) within the G-subunit to be on par with that of the D-subunit but observe comparatively
307 little specialization within the K-subunit (**Figure 5A, 5B, Supplementary Tables**). Therefore, the
308 G-subunit contributes an outsized fraction of the total sequence specialization between Vnf and
309 Anf, despite its remoteness from any nitrogenase metalloclusters. Specialized sites within the G-

subunit are distal from the G-D interface. These include one residue close to the presumed H-subunit binding site, which might contribute to a previously proposed role in mediating H-subunit binding specificity[25].

We next examined how these signatures of sequence-level specialization manifested early in G-subunit evolution, tracking their amino acid content through the divergence of Vnf and Anf clades. We find that the specialized sequence motifs of Vnf^{Anc} (Vnf ancestor) and Anf^{Anc} (Anf ancestor) most closely resemble those of their respective descendants (**Figure 5C**). The most likely Vnf^{Anc} motif across divergent sites is "FIKIDV" (extant Vnf consensus "FIKVDH") whereas the most likely Anf^{Anc} motif is "VNVDYD" (extant Anf consensus "VNVDYD"). However, we find that the most likely motif of the common $VnfAnf^{Anc}$ ancestor, "VIKIDV," is more like that of Vnf than Anf, even accounting for the statistical uncertainty associated with ancestral reconstruction. This result suggests that the earliest G-subunit proteins were Vnf-like (in agreement with previous inferences of ancestral nitrogenases that indicate Vnf evolved prior to Anf[28]), suggesting that the G-subunit contributed to the specificity required for vanadium dependency early in the birth of a new nitrogenase.



328 ***Figure 5. Coevolution of G-subunit specialization and Vnf/Anf metal dependence**. **A)***
329 *Specialized sites mapped to the nitrogenase VnfDGK structure (PDB 5N6Y[22]). Scale in A) applies*
330 *to B).* ***B)*** *Detailed view of specialized sites across the G-subunit.* ***C)*** *Evolution of specialized G-*
331 *subunit sequence motifs through the divergence of Vnf and Anf. Sequence logos scale either with*
332 *amino acid frequencies (for extant Vnf or Anf nitrogenases) or posterior probabilities (for ancestral*
333 *VnfAnf^{Anc}) at each site. VnfAnf^{Anc} dependence on FeV-co is inferred based on comparisons of*
334 *specialized sequence motifs.* ***D)*** *The bulk marine concentrations of redox-sensitive metals*
335 *involved in nitrogen fixation (Mo, V, Fe) have changed with progressive oxygenation of the Earth*
336 *surface environment, with Fe more abundant in the Archean and Mo/V more abundant today.*
337 *Though quantitative estimates of historical V concentrations are not presently available,*
338 *oxygenation likely impacted the speciation and solubility of V and Mo similarly over geologic*
339 *time[57,58]. Metal concentrations from Zerkle et al.[59] and references therein and p_{O2} from Lyons et*
340 *al.[27]. Age constraints of nitrogenase origins and VnfAnf^{Anc} (i.e., G-subunit emergence) from*
341 *Stueken et al.[12] and Parsons et al.[13], respectively. GOE: Great Oxidation Event, PAL: present*
342 *atmospheric level.*
343
344 **DISCUSSION**
345
346 In the present study, we interrogated the origin, early evolution, and functional significance of a
347 major novelty in the history of nitrogen fixation: the emergence of the G-subunit in Vnf and Anf
348 nitrogenases >1.5 billion years ago. With the birth of the nitrogenase G-subunit, not only did a
349 rare evolutionary event occur with associated fitness costs overcome[4], but a novel structural
350 domain became fully integrated into the nitrogenase complex and was retained in one of the most
351 central metabolic processes to Earth's biosphere.
352
353 We demonstrate that there currently exists no candidate, evolutionary predecessor of the G-
354 subunit family, whether another protein, protein fragment, or segment of non-genic DNA. This
355 finding sets the origin of the G-subunit apart from other major evolutionary events in the history
356 of nitrogen fixation that are explainable by gene duplication, including the divergence between
357 other catalytic subunits, between nitrogenases and their maturation protein scaffolds, and
358 between nitrogenase isozymes[32,60]. The addition of such a small and evolutionarily enigmatic
359 protein domain appears to be an anomaly across biogeochemical enzymes.
360

361 We propose that the G-subunit was formed *de novo* from previously non-genic DNA, more
362 recently acknowledged as a viable process for generating genetic novelty[48-50]. Such an event may
363 have accompanied the genetic duplication and divergence of Nif that produced alternative
364 nitrogenases[32]. Regardless of its originating mechanism, retainment of the G-subunit gene within
365 the nitrogenase operon appears to have been improbable. Insertions between nitrogenase D- and
366 K-subunit genes must either be rare or frequently deleterious, as evidenced by their small
367 intergenic distance and absence of any other disrupting open reading frames across extant Nif
368 clusters in our dataset. Thus, the evolution of the G-subunit constitutes an unprecedented gain in
369 structural and genetic complexity in the evolutionary history of nitrogen fixation.
370
371 By analyzing predicted, nitrogenase ancestors, we reveal an evolutionary steppingstone in the
372 initial recruitment of the G-subunit. Our reconstruction of ancestral, D-subunit surface features
373 indicates that the G-subunit binding site was built from an essential, preexisting metallocluster
374 insertion site, which additionally appears preconditioned for G-subunit interactions. How did this
375 essential surface feature accommodate burial by a new interaction partner? Metallocluster
376 delivery in extant nitrogenases is thought to be mediated by transient protein-protein
377 interactions[56,61,62] that would presumably have shaped nitrogenase surface properties, perhaps
378 explaining the early evolved hydrophobicity of the ancestral, proto-interface. Our evolutionary
379 analyses, coupled with experimental support of a role in nitrogenase assembly[24], suggest that the
380 G-subunit was integrated within the assembly pathway, replacing and/or mediating earlier evolved
381 protein or metallocluster interactions. Thus, the birth of the permanent G-D interface represents
382 a gradual subsummation of formerly transient interactions that provided an evolutionary "cradle"
383 for a permanent, novel structural domain.
384
385 In its capacity to mediate newly evolved incorporation of V- and Fe-based clusters, the G-subunit
386 would have enabled the evolutionary burst of both metal dependence and ecological diversity
387 within Group III nitrogenase lineage from which alternative nitrogenases originated. We find that
388 G-subunit residues, despite being relatively distal from the active-site and not retaining bound,
389 reactive metalloclusters like other nitrogenase subunits, have an outsized contribution to the
390 sequence-level specialization between alternative nitrogenases. Possible functional outcomes
391 include improved specificity toward metalloclusters or associated protein partners during
392 nitrogenase assembly (preventing costly mismetallation[63])[24] and mediating crosstalk between
393 electron transfer and catalytic nitrogenase components[25]. Each of these functions would have
394 equipped diazotrophs to coordinate multiple nitrogenases systems with unique metal

395 dependencies within the same cellular environment. The G-subunit therefore primed these
396 organisms for fluctuations in redox-sensitive metal availabilities following the rise of oxygenic
397 phototrophs[27,64], one of the most significant ecological revolutions in Earth history.
398
399 An ongoing debate in evolutionary biology centers on the degree to which environmental
400 conditions drive biological innovations. In the evolution of biological nitrogen fixation, an intuitive
401 narrative is that diversification of nitrogenase metal dependence, intimately tied to the productivity
402 and ecological distribution of nitrogen fixation, was triggered largely by transitions in bulk,
403 environmental metal availabilities[26], the historical records of which are primarily built from
404 geochemical analyses of ancient marine sediments (**Figure 5D**). Our study instead highlights the
405 importance of chance events shaping evolutionary innovations with global consequences. The
406 birth of the nitrogenase G-subunit, by providing a crucial component in novel metal incorporation,
407 expanded the adaptive landscape for biological nitrogen fixation, forever altering the evolutionary
408 trajectory of the biosphere. This ancient, orphan protein was birthed by seemingly improbable
409 events but was nevertheless fostered by molecular features that were predisposed for its arrival.
410 Whether this evolutionary transition in the history of biological nitrogen fixation could have been
411 crossed in the absence of the G-subunit awaits further studies. Similar, future synergies between
412 historical studies of molecular novelties and the environmental circumstances that shaped them
413 can refine our understanding of how the Earth-life planetary system has evolved over billions of
414 years.
415
416 **METHODS**
417
418 **Nitrogenase protein sequence curation**
419
420 A preliminary 412-sequence dataset of nitrogenase Vnf/AnfG proteins was compiled by BLASTP[40]
421 search (Expect value threshold = 0.1) against the National Center for Biotechnology Information
422 (NCBI) non-redundant protein sequence database (nr)[65] (accessed January 2022), using the *A.*
423 *vinelandii* VnfG query sequence (WP_012698949.1). BLASTP hits were curated by filtering
424 against partial sequences and homologs whose encoding genes are distantly located from *vnf* or
425 *anf* operons (and thus unlikely to be canonical *vnf/anfG* genes). Certain VnfG sequences were
426 manually extracted from fused VnfDG proteins, using MAFFT v7.490[66] alignments to other D- or
427 G-subunit sequences as a guide. The preliminary Vnf/AnfG sequence dataset was pruned by CD-
428 HIT[67] (sequence identity threshold = 90%), yielding a final 188-sequence G-subunit dataset.

429

430 **G-subunit homolog search**

431

432 In addition to the BLASTP search described above, three sequence-based search tools–
433 TBLASTN[40], PSI-BLAST[41], and HMMER[42]–were used to identify distantly related homologs of the
434 nitrogenase G-subunit (conducted January 2022). TBLASTN and PSI-BLAST searches against
435 the NCBI nucleotide collection (nr/nt) and nr protein database, respectively, were executed with
436 the *A. vinelandii* VnfG query (WP_012698949.1), an Expect value threshold = 0.1, and word size
437 = 2 (all other parameters default). HMMER search (hmmsearch) against the UniProt Reference
438 Proteome database[68] was performed using an HMM profile generated from our MAFFT-aligned,
439 188-sequence G-subunit dataset (see above) with an Expect value threshold = 0.1 for both
440 sequences and hits (all other parameters default).

441

442 Structure-based searches of distant G-subunit homologs were performed by the FoldSeek
443 server[43] using the *A. vinelandii* VnfG crystallographic structure (PDB 5N6Y[22]) query against the
444 Protein Data Bank and AlphaFold-UniProt/Swiss-Prot databases. We employed both 3Di-AA and
445 TM-Align modes with an 3Di-AA Expect value threshold = 0.01 or a TM-score = 0.5 (template
446 modeling score).

447

448 **Phylogenetic and ancestral sequence reconstruction**

449

450 The curated, 188-sequence Vnf/AnfG dataset was aligned by MAFFT. No outgroup to G-subunit
451 sequences was included because the distant homology search described above did not yield
452 convincing hits (see **Results**). IQ-TREE v1.6.12[69] was used to reconstruct a maximum-likelihood
453 G-subunit protein phylogeny, using the best-fit evolutionary model, WAG+R6 (selected by the
454 Bayesian Information Criterion, BIC), tested by ModelFinder[70], with branch support calculated by
455 ultra-fast bootstrap approximation (UFBoot)[71] and the Shimodaira-Hasegawa-like approximate
456 likelihood-ratio test (SH-like aLRT)[72].

457

458 A phylogeny of concatenated Vnf/AnfHDGK proteins was generated by retrieving syntenous
459 subunit sequences to the curated VnfG/AnfG dataset. Certain VnfH sequences that recently
460 diverged from NifH clades were omitted due to having different evolutionary histories than those
461 of other Vnf/Anf proteins. NifHDK sequences from a prior phylogenetic analysis[73] were included
462 as an outgroup (n = 67, closely branching to Anf/Vnf) and were used to calculate intergenic *nifD*-

463 to-*nifK* distances (n = 306). Each subunit sequence dataset was aligned by MAFFT prior to
464 concatenation. Maximum-likelihood phylogenetic reconstruction was performed by IQ-TREE as
465 described above (best-fit evolutionary model, BIC: LG+F+R9).
466
467 Marginal maximum-likelihood ancestral sequence reconstruction of concatenated Vnf/AnfHDGK
468 sequences was performed by PAML v4.9j[74] (LG+F+G4 model). Mean sitewise posterior
469 probabilities D- and G-subunit ancestors analyzed in this study range from .75 to .96 (**Table S1**).
470 Sequence gaps were reconstructed as described by Aadland et al.[75]. Briefly, the input protein
471 sequence alignment was recoded as a binary amino-acid presence/absence matrix. Ancestral
472 presence/absence states were reconstructed by PAML using a binary character model and state
473 frequencies inferred by maximum likelihood.
474
475 Logo representations of extant multiple sequence alignments were generated with the Python
476 library LogoMaker[76] using the probability transformation of the count matrix. Representations of
477 ancestral sequences were generated from the posterior probability matrices produced by PAML.
478
479 Sequence and phylogenetic tree datasets are available at
480 https://github.com/kacarlab/nitrogenase-g-subunit.
481
482 **Optimal growth temperature prediction**
483
484 For taxa represented in the nitrogenase G-subunit sequence dataset, optimal growth
485 temperatures (OGT) were predicted from genomic content using the method described by Sauer
486 and Wang[77]. Due to genome incompleteness and the absence of 16S rRNA sequences in
487 fragmented genomes for certain representative taxa, the selected "Superkingdom Bacteria" and
488 "Superkingdom Archaea" regression models for OGT prediction excluded genome size and 16S
489 rRNA.
490
491 **Protein conservation and divergence analysis**
492
493 Sequence conservation for D- and G-subunit proteins was determined by the ConSurf server[78].
494 Sites were considered well-conserved if their ConSurf grade were >7. Sequence specialization
495 was determined by the TwinCons package[79]. TwinCons was run separately on individual
496 alignments of extant H, D, G, and K subunit protein sequences with user-defined Vnf and Anf

497 groups, using default parameters and the LG substitution matrix. TwinCons specialization scores
498 were normalized to a 0-to-1 scale per analyzed subunit.
499
500 **Structural Modeling**
501
502 Ancestral and extant NifHDGK protein structures were predicted using ColabFold[80], which
503 combines the AlphaFold2 method for structure prediction from multiple-sequence alignments[18]
504 with the MMSeq2 method to obtain evolutionary information[81]. Specifically, the colabfold_batch
505 v1.4.0 implementation (https://github.com/YoshitakaMo/localcolabfold) was used with the
506 AlphaFold-v2.0-multimer prediction model and standard options (3 recycles and AMBER all-atom
507 optimization in GPU).Protein structures were visualized by ChimeraX[82].
508
509 **Structure Analysis**
510
511 Residue interaction networks[83] were generated using ProLif[84], which finds the different possible
512 interactions (hydrophobic, electrostatic, hydrogen bonds, etc) between two groups of atoms (here
513 defined as G- and D-subunit groups) within a structure snapshot (e.g. single structure, trajectory
514 frame). Interactions were compiled into networks by NetworkX[85], which were later visualized in
515 Gephi[86]. Scripts used to generate the networks are available at
516 https://github.com/kacarlab/nitrogenase-g-subunit.
517
518 Surface electrostatic maps were computed using the Poisson Boltzmann Solvent-Accessible
519 model implemented in APBS[87] through the PyMOL plugin (Schrödinger, Inc), using standard
520 PDB2PQR and APBS protocols at pH 7.
521
522 Sitewise contribution to D-G interface stability was computed using *in-silico* alanine scanning,
523 which substitutes residues with a minimal methyl moiety (alanine)[88]. The changes on the binding
524 free energy values upon substitution were obtained using the machine-learning enabled mCSM2
525 method[89] with default parameters.
526
527 Allosteric effects across the nitrogenase complex was modeled using Perturbation Response
528 Scanning[90], implemented in ProDy[90] with default parameters and provided with PDB 5N6Y[22] alpha
529 carbon coordinates. Scripts used to model allosteric effects are available at
530 https://github.com/kacarlab/nitrogenase-g-subunit.

ACKNOWLEDGEMENTS

This research was supported by the National Aeronautics and Space Administration (NASA) Interdisciplinary Consortium for Astrobiology Research: Metal Utilization and Selection Across Eons, MUSE (19-ICAR19_2–0007), the NASA Arizona Space Grant (B.M.C), the John Templeton Foundation (B.K.; 61926), the National Science Foundation (B.K.; 2228495), the NASA Early Career Faculty Award (B.K.), the Hypothesis Fund Award (B.K.), and the Universidad Politécnica de Madrid (UPM) Margarita Salas Fellowship, founded by the Unión Europea - Next Generation EU (B.C.-Z.; UP2021-035). We thank Morgan Sobol for assistance with optimal growth temperature prediction, as well as Alex Rivier, Holly Rucker, Lance Seefeldt, Derek Harris, Joanna Masel and Steven Russell for valuable discussions and feedback.

REFERENCES

1 Choi, I.-G. & Kim, S.-H. Evolution of protein structural classes and protein sequence families. *Proc Natl Acad Sci USA* **103**, 14056-14061, doi:10.1073/pnas.0606239103 (2006).

2 Knoll, A. H. The geological consequences of evolution. *Geobiology* **1**, 3-14, doi:10.1046/j.1472-4669.2003.00002.x (2003).

3 Jacob, F. Evolution and Tinkering. *Science* **196**, 1161-1166, doi:10.1126/science.860134 (1977).

4 Andersson, D. I., Jerlstrom-Hultqvist, J. & Nasvall, J. Evolution of new functions de novo and from preexisting genes. *Cold Spring Harb Perspect Biol* **7**, doi:10.1101/cshperspect.a017996 (2015).

5 Eck, R. V. & Dayhoff, M. O. Evolution of the Structure of Ferredoxin Based on Living Relics of Primitive Amino Acid Sequences. *Science* **152**, 363-366, doi:10.1126/science.152.3720.363 (1966).

6 Copley, S. D. Evolution of new enzymes by gene duplication and divergence. *FEBS J* **287**, 1262-1283, doi:10.1111/febs.15299 (2020).

7 Alva, V., Soding, J. & Lupas, A. N. A vocabulary of ancient peptides at the origin of folded proteins. *Elife* **4**, e09410, doi:10.7554/eLife.09410 (2015).

8 Van Oss, S. B. & Carvunis, A. R. De novo gene birth. *PLoS Genet* **15**, e1008160, doi:10.1371/journal.pgen.1008160 (2019).

565 9 Beasley, J. R. & Hecht, M. H. Protein design: the choice of de novo sequences. *J Biol*
566 *Chem* **272**, 2031-2034, doi:10.1074/jbc.272.4.2031 (1997).
567 10 Falkowski, P. G., Fenchel, T. & Delong, E. F. The microbial engines that drive Earth's
568 biogeochemical cycles. *Science* **320**, 1034-1039, doi:10.1126/science.1153213 (2008).
569 11 Moore, E. K., Jelen, B. I., Giovannelli, D., Raanan, H. & Falkowski, P. G. Metal availability
570 and the expanding network of microbial metabolisms in the Archaean eon. *Nat Geosci* **10**,
571 629-636, doi:10.1038/ngeo3006 (2017).
572 12 Stueken, E. E., Buick, R., Guy, B. M. & Koehler, M. C. Isotopic evidence for biological
573 nitrogen fixation by molybdenum-nitrogenase from 3.2 Gyr. *Nature* **520**, 666-669,
574 doi:10.1038/nature14180 (2015).
575 13 Parsons, C., Stüeken, E. E., Rosen, C. J., Mateos, K. & Anderson, R. E. Radiation of
576 nitrogen-metabolizing enzymes across the tree of life tracks environmental transitions in
577 Earth history. *Geobiology* **19**, 18-34, doi:10.1111/gbi.12419 (2020).
578 14 Falkowski, P. G. Evolution of the nitrogen cycle and its influence on the biological
579 sequestration of CO_2 in the ocean. *Nature* **387**, 272-275, doi:10.1038/387272a0 (1997).
580 15 Allen, J. F., Thake, B. & Martin, W. F. Nitrogenase Inhibition Limited Oxygenation of
581 Earth's Proterozoic Atmosphere. *Trends Plant Sci* **24**, 1022-1031,
582 doi:10.1016/j.tplants.2019.07.007 (2019).
583 16 Sanchez-Baracaldo, P., Ridgwell, A. & Raven, J. A. A neoproterozoic transition in the
584 marine nitrogen cycle. *Curr Biol* **24**, 652-657, doi:10.1016/j.cub.2014.01.041 (2014).
585 17 Einsle, O. & Rees, D. C. Structural Enzymology of Nitrogenase Enzymes. *Chem Rev* **120**,
586 4969-5004, doi:10.1021/acs.chemrev.0c00067 (2020).
587 18 Mus, F., Alleman, A. B., Pence, N., Seefeldt, L. C. & Peters, J. W. Exploring the
588 alternatives of biological nitrogen fixation. *Metallomics* **10**, 523-538,
589 doi:10.1039/c8mt00038g (2018).
590 19 Eady, R. R. Structure-function relationships of alternative nitrogenases. *Chemical*
591 *Reviews* **96**, 3013-3030, doi:10.1021/cr950057h (1996).
592 20 Ghebreamlak, S. M. & Mansoorabadi, S. O. Divergent Members of the Nitrogenase
593 Superfamily: Tetrapyrrole Biosynthesis and Beyond. *ChemBioChem* **21**, 1723-1728,
594 doi:10.1002/cbic.201900782 (2020).
595 21 Weiss, M. C. *et al.* The physiology and habitat of the last universal common ancestor. *Nat*
596 *Microbiol* **1** (2016). <https://www.ncbi.nlm.nih.gov/pubmed/27562259>.
597 22 Sippel, D. & Einsle, O. The structure of vanadium nitrogenase reveals an unusual bridging
598 ligand. *Nat Chem Biol* **13**, 956-960, doi:10.1038/nchembio.2428 (2017).

599 23 Eady, R. R., Robson, R. L., Richardson, T. H., Miller, R. W. & Hawkins, M. The vanadium
600 nitrogenase of Azotobacter chroococcum. Purification and properties of the VFe protein.
601 *Biochemical Journal* **244**, 197-207, doi:10.1042/bj2440197 (1987).
602 24 Chatterjee, R., Ludden, P. W. & Shah, V. K. Characterization of VNFG, the delta subunit
603 of the vnf-encoded apodinitrogenase from Azotobacter vinelandii. Implications for its role
604 in the formation of functional dinitrogenase 2. *J Biol Chem* **272**, 3758-3765,
605 doi:10.1074/jbc.272.6.3758 (1997).
606 25 Pence, N., Lewis, N., Alleman, A. B., Seefeldt, L. C. & Peters, J. W. Revealing a role for
607 the G subunit in mediating interactions between the nitrogenase component proteins. *J*
608 *Inorg Biochem* **214**, 111273, doi:10.1016/j.jinorgbio.2020.111273 (2021).
609 26 Anbar, A. D. & Knoll, A. H. Proterozoic ocean chemistry and evolution: a bioinorganic
610 bridge? *Science* **297**, 1137-1142, doi:10.1126/science.1069651 (2002).
611 27 Lyons, T. W., Reinhard, C. T. & Planavsky, N. J. The rise of oxygen in Earth's early ocean
612 and atmosphere. *Nature* **506**, 307-315, doi:10.1038/nature13068 (2014).
613 28 Garcia, A. K., McShea, H., Kolaczkowski, B. & Kacar, B. Reconstructing the evolutionary
614 history of nitrogenases: Evidence for ancestral molybdenum-cofactor utilization.
615 *Geobiology* **18**, 394-411, doi:10.1111/gbi.12381 (2020).
616 29 Glass, J. B., Wolfe-Simon, F. & Anbar, A. D. Coevolution of metal availability and nitrogen
617 assimilation in cyanobacteria and algae. *Geobiology* **7**, 100-123, doi:10.1111/j.1472-
618 4669.2009.00190.x (2009).
619 30 Hoffman, B. M., Lukoyanov, D., Yang, Z. Y., Dean, D. R. & Seefeldt, L. C. Mechanism of
620 nitrogen fixation by nitrogenase: the next stage. *Chem Rev* **114**, 4041-4062,
621 doi:10.1021/cr400641x (2014).
622 31 Raymond, J., Siefert, J. L., Staples, C. R. & Blankenship, R. E. The natural history of
623 nitrogen fixation. *Mol Biol Evol* **21**, 541-554, doi:10.1093/molbev/msh047 (2004).
624 32 Boyd, E. S. & Peters, J. W. New insights into the evolutionary history of biological nitrogen
625 fixation. *Front Microbiol* **4**, 201, doi:10.3389/fmicb.2013.00201 (2013).
626 33 Harris, D. F. *et al.* Mo-, V-, and Fe-Nitrogenases Use a Universal Eight-Electron
627 Reductive-Elimination Mechanism To Achieve N_2 Reduction. *Biochemistry* **58**, 3293-3301,
628 doi:10.1021/acs.biochem.9b00468 (2019).
629 34 Zheng, Y. *et al.* A pathway for biological methane production using bacterial iron-only
630 nitrogenase. *Nat Microbiol* **3**, 281-286, doi:10.1038/s41564-017-0091-5 (2018).

631 35 Hu, Y., Lee, C. C. & Ribbe, M. W. Extending the Carbon Chain: Hydrocarbon Formation
632 Catalyzed by Vanadium/Molybdenum Nitrogenases. *Science* **333**, 753-755,
633 doi:10.1126/science.1206883 (2011).
634 36 Rohde, M., Trncik, C., Sippel, D., Gerhardt, S. & Einsle, O. Crystal structure of VnfH, the
635 iron protein component of vanadium nitrogenase. *J Biol Inorg Chem* **23**, 1049-1056,
636 doi:10.1007/s00775-018-1602-4 (2018).
637 37 Schmid, B. *et al.* Biochemical and structural characterization of the cross-linked complex
638 of nitrogenase: comparison to the ADP-AlF4(-)-stabilized structure. *Biochemistry* **41**,
639 15557-15565 (2002).
640 38 Spatzal, T. *et al.* Evidence for interstitial carbon in nitrogenase FeMo cofactor. *Science*
641 **334**, 940, doi:10.1126/science.1214025 (2011).
642 39 Waugh, S. I. *et al.* The genes encoding the delta subunits of dinitrogenases 2 and 3 are
643 required for mo-independent diazotrophic growth by Azotobacter vinelandii. *Journal of*
644 *Bacteriology* **177**, 1505-1510, doi:10.1128/jb.177.6.1505-1510.1995 (1995).
645 40 Altschul, S. F., Gish, W., Miller, W., Myers, E. W. & Lipman, D. J. Basic local alignment
646 search tool. *Journal of Molecular Biology* **215**, 403-410, doi:10.1016/s0022-
647 2836(05)80360-2 (1990).
648 41 Altschul, S. F. *et al.* Gapped BLAST and PSI-BLAST: a new generation of protein
649 database search programs. *Nucleic Acids Research* **25**, 3389-3402,
650 doi:10.1093/nar/25.17.3389 (1997).
651 42 Finn, R. D., Clements, J. & Eddy, S. R. HMMER web server: interactive sequence
652 similarity searching. *Nucleic Acids Res* **39**, W29-37, doi:10.1093/nar/gkr367 (2011).
653 43 van Kempen, M. *et al.* Foldseek: fast and accurate protein structure search.
654 doi:10.1101/2022.02.07.479398 (2022).
655 44 Illergard, K., Ardell, D. H. & Elofsson, A. Structure is three to ten times more conserved
656 than sequence--a study of structural response in protein cores. *Proteins* **77**, 499-508,
657 doi:10.1002/prot.22458 (2009).
658 45 Medina, M. S. *et al.* CowN sustains nitrogenase turnover in the presence of the inhibitor
659 carbon monoxide. *J Biol Chem* **296**, 100501, doi:10.1016/j.jbc.2021.100501 (2021).
660 46 Rubio, L. M. & Ludden, P. W. Maturation of nitrogenase: a biochemical puzzle. *J Bacteriol*
661 **187**, 405-414, doi:10.1128/JB.187.2.405-414.2005 (2005).
662 47 Tautz, D. & Domazet-Loso, T. The evolutionary origin of orphan genes. *Nat Rev Genet*
663 **12**, 692-702, doi:10.1038/nrg3053 (2011).

664 48 Heames, B., Schmitz, J. & Bornberg-Bauer, E. A Continuum of Evolving De Novo Genes
665 Drives Protein-Coding Novelty in Drosophila. *J Mol Evol* **88**, 382-398, doi:10.1007/s00239-
666 020-09939-z (2020).
667 49 Vakirlis, N. *et al.* De novo emergence of adaptive membrane proteins from thymine-rich
668 genomic sequences. *Nat Commun* **11**, 781, doi:10.1038/s41467-020-14500-z (2020).
669 50 Vakirlis, N., Carvunis, A. R. & McLysaght, A. Synteny-based analyses indicate that
670 sequence divergence is not the main source of orphan genes. *Elife* **9**,
671 doi:10.7554/eLife.53500 (2020).
672 51 Dos Santos, P. C., Fang, Z., Mason, S. W., Setubal, J. C. & Dixon, R. Distribution of
673 nitrogen fixation and nitrogenase-like sequences amongst microbial genomes. *BMC*
674 *Genomics* **13**, 162, doi:10.1186/1471-2164-13-162 (2012).
675 52 Anbar, A. D. *et al.* A whiff of oxygen before the great oxidation event? *Science* **317**, 1903-
676 1906, doi:10.1126/science.1140325 (2007).
677 53 Kessler, P. S., McLarnan, J. & Leigh, J. A. Nitrogenase phylogeny and the molybdenum
678 dependence of nitrogen fixation in Methanococcus maripaludis. *J Bacteriol* **179**, 541-543
679 (1997).
680 54 Rutledge, H. L., Cook, B. D., Nguyen, H. P. M., Herzik, M. A. & Tezcan, F. A. Structures
681 of the nitrogenase complex prepared under catalytic turnover conditions. *Science* **377**,
682 865-869, doi:10.1126/science.abq7641 (2022).
683 55 Huang, Q. *et al.* Mechanical coupling in the nitrogenase complex. *PLoS Comput Biol* **17**,
684 e1008719, doi:10.1371/journal.pcbi.1008719 (2021).
685 56 Schmid, B. *et al.* Structure of a cofactor-deficient nitrogenase MoFe protein. *Science* **296**,
686 352-356, doi:10.1126/science.1070010 (2002).
687 57 Moore, E. K., Hao, J., Spielman, S. J. & Yee, N. The evolving redox chemistry and
688 bioavailability of vanadium in deep time. *Geobiology* **18**, 127-138, doi:10.1111/gbi.12375
689 (2020).
690 58 Emerson, S. R. & Huested, S. S. Ocean anoxia and the concentrations of molybdenum
691 and vanadium in seawater. *Marine Chemistry* **34**, 177-196, doi:10.1016/0304-
692 4203(91)90002-e (1991).
693 59 Zerkle, A. L., House, C. H. & Brantley, S. L. Biogeochemical signatures through time as
694 inferred from whole microbial genomes. *Am J Sci* **305**, 467-502, doi:DOI
695 10.2475/ajs.305.6-8.467 (2005).

696 60 Garcia, A. K., Kolaczkowski, B. & Kacar, B. Reconstruction of Nitrogenase Predecessors
697 Suggests Origin from Maturase-Like Proteins. *Genome Biol Evol* **14**,
698 doi:10.1093/gbe/evac031 (2022).
699 61 Ruttimann-Johnson, C., Rubio, L. M., Dean, D. R. & Ludden, P. W. VnfY is required for
700 full activity of the vanadium-containing dinitrogenase in Azotobacter vinelandii. *J Bacteriol*
701 **185**, 2383-2386, doi:10.1128/JB.185.7.2383-2386.2003 (2003).
702 62 Perez-Gonzalez, A. *et al.* Specificity of NifEN and VnfEN for the Assembly of Nitrogenase
703 Active Site Cofactors in Azotobacter vinelandii. *mBio* **12**, e0156821,
704 doi:10.1128/mBio.01568-21 (2021).
705 63 Rebelein, J. G., Lee, C. C., Newcomb, M., Hu, Y. & Ribbe, M. W. Characterization of an
706 M-Cluster-Substituted Nitrogenase VFe Protein. *MBio* **9**, doi:10.1128/mBio.00310-18
707 (2018).
708 64 Hazen, R. M. *et al.* Mineral evolution. *American Mineralogist* **93**, 1693-1720,
709 doi:10.2138/am.2008.2955 (2008).
710 65 Sayers, E. W. *et al.* Database resources of the national center for biotechnology
711 information. *Nucleic Acids Res* **50**, D20-D26, doi:10.1093/nar/gkab1112 (2022).
712 66 Katoh, K. & Standley, D. M. MAFFT multiple sequence alignment software version 7:
713 improvements in performance and usability. *Mol Biol Evol* **30**, 772-780,
714 doi:10.1093/molbev/mst010 (2013).
715 67 Fu, L., Niu, B., Zhu, Z., Wu, S. & Li, W. CD-HIT: accelerated for clustering the next-
716 generation sequencing data. *Bioinformatics* **28**, 3150-3152,
717 doi:10.1093/bioinformatics/bts565 (2012).
718 68 The Uniprot Consortium. UniProt: the Universal Protein Knowledgebase in 2023. *Nucleic*
719 *Acids Res* **51**, D523-D531, doi:10.1093/nar/gkac1052 (2023).
720 69 Nguyen, L. T., Schmidt, H. A., von Haeseler, A. & Minh, B. Q. IQ-TREE: a fast and effective
721 stochastic algorithm for estimating maximum-likelihood phylogenies. *Mol Biol Evol* **32**,
722 268-274, doi:10.1093/molbev/msu300 (2015).
723 70 Kalyaanamoorthy, S., Minh, B. Q., Wong, T. K. F., von Haeseler, A. & Jermiin, L. S.
724 ModelFinder: fast model selection for accurate phylogenetic estimates. *Nature Methods*
725 **14**, 587-589, doi:10.1038/nmeth.4285 (2017).
726 71 Hoang, D. T., Chernomor, O., von Haeseler, A., Minh, B. Q. & Vinh, L. S. UFBoot2:
727 Improving the Ultrafast Bootstrap Approximation. *Mol Biol Evol* **35**, 518-522,
728 doi:10.1093/molbev/msx281 (2018).

729 72 Guindon, S. *et al.* New algorithms and methods to estimate maximum-likelihood
730 phylogenies: assessing the performance of PhyML 3.0. *Syst Biol* **59**, 307-321,
731 doi:10.1093/sysbio/syq010 (2010).
732 73 Garcia, A. K. *et al.* Nitrogenase resurrection and the evolution of a singular enzymatic
733 mechanism. *Elife* **12**, doi:10.7554/eLife.85003 (2023).
734 74 Yang, Z. PAML 4: phylogenetic analysis by maximum likelihood. *Mol Biol Evol* **24**, 1586-
735 1591, doi:10.1093/molbev/msm088 (2007).
736 75 Aadland, K., Pugh, C. & Kolaczkowski, B. High-Throughput Reconstruction of Ancestral
737 Protein Sequence, Structure, and Molecular Function. *Methods Mol Biol* **1851**, 135-170,
738 doi:10.1007/978-1-4939-8736-8_8 (2019).
739 76 Tareen, A. & Kinney, J. B. Logomaker: beautiful sequence logos in Python. *Bioinformatics*
740 **36**, 2272-2274, doi:10.1093/bioinformatics/btz921 (2020).
741 77 Sauer, D. B. & Wang, D. N. Predicting the optimal growth temperatures of prokaryotes
742 using only genome derived features. *Bioinformatics* **35**, 3224-3231,
743 doi:10.1093/bioinformatics/btz059 (2019).
744 78 Ashkenazy, H. *et al.* ConSurf 2016: an improved methodology to estimate and visualize
745 evolutionary conservation in macromolecules. *Nucleic Acids Res* **44**, W344-350,
746 doi:10.1093/nar/gkw408 (2016).
747 79 Penev, P. I., Alvarez-Carreno, C., Smith, E., Petrov, A. S. & Williams, L. D. TwinCons:
748 Conservation score for uncovering deep sequence similarity and divergence. *PLoS*
749 *Comput Biol* **17**, e1009541, doi:10.1371/journal.pcbi.1009541 (2021).
750 80 Mirdita, M. *et al.* ColabFold: making protein folding accessible to all. *Nature Methods* **19**,
751 679-682, doi:10.1038/s41592-022-01488-1 (2022).
752 81 Steinegger, M. & Söding, J. MMseqs2 enables sensitive protein sequence searching for
753 the analysis of massive data sets. *Nature Biotechnology* **35**, 1026-1028,
754 doi:10.1038/nbt.3988 (2017).
755 82 Pettersen, E. F. *et al.* UCSF ChimeraX: Structure visualization for researchers, educators,
756 and developers. *Protein Sci* **30**, 70-82, doi:10.1002/pro.3943 (2021).
757 83 Csermely, P. Creative elements: network-based predictions of active centres in proteins
758 and cellular and social networks. *Trends Biochem Sci* **33**, 569-576,
759 doi:10.1016/j.tibs.2008.09.006 (2008).
760 84 Bouysset, C. & Fiorucci, S. ProLIF: a library to encode molecular interactions as
761 fingerprints. *Journal of Cheminformatics* **13**, doi:10.1186/s13321-021-00548-6 (2021).

85 Hagberg, A. A., Schult, D. A. & Swart, P. J. in *Proceedings of the 7th Python in Science Conference.* (eds Gäel Varoquaux, Travis Vaught, & Jarrod Millman) 11-15.

86 Bastian, M., Heymann, S. & Jacomy, M. Gephi: An Open Source Software for Exploring and Manipulating Networks. *Proceedings of the International AAAI Conference on Web and Social Media* **3**, 361-362, doi:10.1609/icwsm.v3i1.13937 (2009).

87 Jurrus, E. *et al.* Improvements to the APBS biomolecular solvation software suite. *Protein Science* **27**, 112-128, doi:10.1002/pro.3280 (2017).

88 Kortemme, T., Kim, D. E. & Baker, D. Computational Alanine Scanning of Protein-Protein Interfaces. *Science's STKE* **2004**, doi:10.1126/stke.2192004pl2 (2004).

89 Rodrigues, C. H. M., Myung, Y., Pires, D. E. V. & Ascher, D. B. mCSM-PPI2: predicting the effects of mutations on protein–protein interactions. *Nucleic Acids Research* **47**, W338-W344, doi:10.1093/nar/gkz383 (2019).

90 Gerek, Z. N. & Ozkan, S. B. Change in allosteric network affects binding affinities of PDZ domains: analysis through perturbation response scanning. *PLoS Comput Biol* **7**, e1002154, doi:10.1371/journal.pcbi.1002154 (2011).