Structural evolution of nitrogenase enzymes over geologic time

Bruno Cuevas-Zuviría[1,2†], Franka Detemple[3], Kaustubh Amritkar[1], Amanda K. Garcia[1], Lance C. Seefeldt[4], Oliver Einsle[3] and Betül Kaçar[1*]

1. Department of Bacteriology, University of Wisconsin-Madison, Madison, Wisconsin, United States
2. Centro de Biotecnología y Genómica de Plantas, Universidad Politécnica de Madrid (UPM)—Instituto Nacional de Investigación y Tecnología Agraria y Alimentaria-CSIC (INIA/CSIC), Campus de Montegancedo, Madrid, Spain
3. Institute of Biochemistry, University of Freiburg, Freiburg, Germany
4. Department of Chemistry and Biochemistry, Utah State University, Logan, Utah, United States

[†]Affiliation where research was conducted: Department of Bacteriology, University of Wisconsin-Madison, Madison, United States

[*]To whom correspondence should be addressed: betul.kacar@wisc.edu

Short title: Three billion years of nitrogenase diversity

ABSTRACT

Life on Earth is more than 3.5 billion years old—nearly as old as the age of the planet. Over this vast expanse of time, life and its biomolecules adapted to and triggered profound changes to the Earth's environment. Certain critical enzymes evolved early in the history of life and have persisted through planetary extremes. While sequence data is widely used to trace evolutionary trajectories, enzyme structure remains an underexplored resource for understanding how proteins evolve over long timescales. Here, we implement an integrated approach to study nitrogenase, an ancient, globally critical enzyme essential for nitrogen fixation. Despite the ecological diversity of its host microbes, nitrogenase has strict functional limitations, including extreme oxygen sensitivity, energy requirements and substrate availability. We combined phylogenetics, ancestral sequence reconstruction, protein crystallography and deep-learning based structural prediction, and resurrected three billion years of nitrogenase structural history. We present the first effort to predict all extant and ancestral structures along the evolutionary tree of an enzyme and present a total of over 5000 structures. This approach lays the foundation for reconstructing key structural constraints that influence protein evolution and examining ancient enzymes in the context of phylogenetic and environmental change across geological timescales.

INTRODUCTION

Over more than 3.5 billion years of Earth-life history, certain globally significant biogeochemical processes have shaped planetary surface conditions. To understand the history of life on our planet as environments changed, one needs to understand how enzyme activities evolved over time. Traditional approaches for reconstructing the history of life depend on the study of geological remains and the fossils that can be recovered from them. Today, many important evolutionary questions remain unanswered due to the rarity and degradation of fossils from the earliest stages of Earth's history (Kaçar, 2024). Genetic sequences, particularly those of proteins and enzymes, serve as an independent repository to access life's evolution over long time scales. A next frontier in the study of protein evolution is to wholly uncover the evolutionary history of key enzymes and metabolisms in the context of shifts in ecosystems and global biogeochemistry (Garcia and Kaçar, 2019). Such a goal necessitates a new interdisciplinary synthesis, leveraging state-of-the-art evolutionary models and computational tools to reconstruct extinct enzymes and study their behavior under changing environmental conditions. In addition, the ability to generate structural predictions by high-throughput, *ab initio* methods (e.g., AlphaFold, (Jumper et al., 2021)) now makes it possible to investigate a significantly broader swath of structural variation than has been achieved to date for ecologically critical protein families.

A prime example of an ancient, globally influential enzyme is nitrogenase. Nitrogenase provides access to bioessential nitrogen via the reduction of highly inert atmospheric dinitrogen (N_2) to ammonia (NH_3) (Rucker and Kaçar, 2024). In select microbes called diazotrophs, this pathway is exclusively performed by members of the nitrogenase metalloenzyme family that catalyze the cleavage of the strong N≡N bond. Age-calibrated, phylogenomic analyses of nitrogenase genes (Parsons et al., 2021) and nitrogen isotopic signatures preserved in the geologic record (Stüeken et al., 2015) both provide evidence for the existence of biological nitrogen fixation more than three billion years ago. Thus, as the sole biological means for fixing essential nitrogen, nitrogenases have underlaid the

productivity of the biosphere for most of Earth's history (Falkowski, 1997; Navarro-González et al., 2001; Sánchez-Baracaldo et al., 2014).

Despite the antiquity and ecological importance of nitrogenases, nearly nothing is known about how their enzymatic properties have varied in the past (Rucker and Kaçar, 2024), particularly in response to ancient, global environmental transitions and the distribution of nitrogenase genes across ecologically diverse diazotrophs (Garcia et al., 2023, 2020). Notably, nitrogenases are extremely sensitive to oxygen, which degrades the bound metalloclusters required for their enzymatic activity and multimeric structure (Robson and Postgate, 1980). Nevertheless, these enzymes, which originated in anaerobic organisms inhabiting an anoxic planet >3.2 billion years ago (Stüeken et al., 2015), evolved through the progressive oxygenation of Earth's surface environment (beginning ~2.4 to 3 billion years ago (Lyons et al., 2024, 2014) and genetic acquisition by aerobic organisms (Boyd et al., 2015). Though several regulatory mechanisms and accessory proteins that today shield nitrogenase from oxygen have been identified (Dixon and Kahn, 2004; Robson and Postgate, 1980; Schlesier et al., 2016; Takimoto et al., 2022), molecular features associated with oxygen protection adaptation have not been investigated within the nitrogenase enzyme itself. Another outcome of Earth surface oxygenation is the spaciotemporal shift in the environmental bioavailability of redox-sensitive trace metals used by nitrogenases. Specifically, molybdenum (Mo), vanadium (V), and iron (Fe) (Johnson et al., 2021; Moore et al., 2017) are proposed to have shaped the diversification of the variably metal-dependent Mo-, V-, and Fe-nitrogenase isozymes (Anbar and Knoll, 2002). The emergence of novel features within the nitrogenase enzyme has been suggested as necessary for maintaining biological nitrogen fixation amid global marine geochemical shifts (Cuevas-Zuviría et al., 2024), yet little is known about the molecular changes and constraints that accompanied their diversification and evolution across time. This knowledge gap limits our ability to link molecular innovations to environmental transitions in nitrogenase evolution, obscuring the principles behind an enzyme capable of catalyzing one of nature's most challenging reactions.

A window into the evolution of nitrogenase function can be generated by probing the historical protein structure space representative of this ancient enzyme family. In this study, we present a combinatorial approach for reconstructing ancient nitrogenase structures and tracking their evolution across Earth history. We draw from a phylogenetic dataset of extant and inferred ancient nitrogenase proteins to reconstruct more than three billion years of nitrogenase structural evolution. Our study combines ancestral sequence reconstruction with massive machine-learning-enabled structure prediction of present-day and ancestral nitrogenases, as well as protein crystallographic investigations of select, functionally characterized ancestors. We thus present the first effort to predict all extant (770) and ancestral (4,608) enzyme structures across the nitrogenase evolutionary tree, and interpret key observations across an enzyme's structural history within the context of planetary environmental change.

RESULTS

Structural prediction of ancient nitrogenase enzymes across time

We began our structural investigation of nitrogenase evolutionary history by conducting a large-scale structure prediction analysis, encompassing 5378 protein structures, a more than threefold increase compared to available nitrogenase structures in the PDB. We then analyzed our phylogenetic dataset to identify notable structural changes. Our dataset includes representatives of all three known nitrogenase isozymes, which are differentiated by the metal composition of their active-site metallocluster: Nif (encoded by *nifHDK* genes; incorporates an iron-molybdenum cofactor, "FeMoco"), Vnf (encoded by *vnfHDGK* genes; incorporates an iron-vanadium cofactor, "FeVco"), and Anf (encoded by *anfHDGK* genes; incorporates an iron-only cofactor, "FeFeco"). H-, D-, and K-subunit sequences were analyzed for each nitrogenase homolog (G-subunit sequences unique to Vnf and Anf were not included in the analysis; see **Materials and Methods**). In all nitrogenase isozymes, H-subunits form a homodimeric reductase component ("HH") that transiently associates with and transfers electrons to a catalytic core (Einsle and Rees, 2020). This catalytic component

is a heterotetramer composed of two D- and two K-subunits ("DDKK" or $\alpha_2\beta_2$) and contains the active site metallocluster where N_2 is bound and reduced to NH_3.

The topology of the nitrogenase phylogenetic dataset (Garcia et al., 2023) used in the present study reveals five major clades: Nif-I, Nif-II, Nif-III, Vnf, and Anf (**Figure 1A**). Both Vnf and Anf homologs are nested within the Nif-III clade. Aside from sequence-level homology of their constituent proteins, clades are also distinguished by associated features including metal dependence of the nitrogenase enzyme, taxonomy, host ecology, and complexity of the nitrogenase-associated gene network (Cuevas-Zuviría et al., 2024). In extant diazotrophs, nitrogenases require a suite of associated genes for their regulation, assembly, and maintenance (Campo et al., 2022). The content and complexity of this cellular network for nitrogen fixation tends to be characteristic of the major clades and host ecologies of nitrogenases. For example, Nif-I nitrogenases are primarily hosted by aerobic or facultatively anaerobic diazotrophs that have among the largest number of nitrogenase-associated genes (Boyd et al., 2015). By comparison, Nif-II and Nif-III nitrogenases are hosted nearly exclusively by anaerobic diazotrophs and have comparatively smaller nitrogen fixation gene networks (Cuevas-Zuviría et al., 2024; Garcia et al., 2020). Vnf and Anf isozymes, though sharing certain associated genes with Nif, are also supported by distinct, dedicated gene clusters (Garcia et al., 2022, 2020). We hypothesized that these differences would manifest in correlated structural differences across the enzyme family.

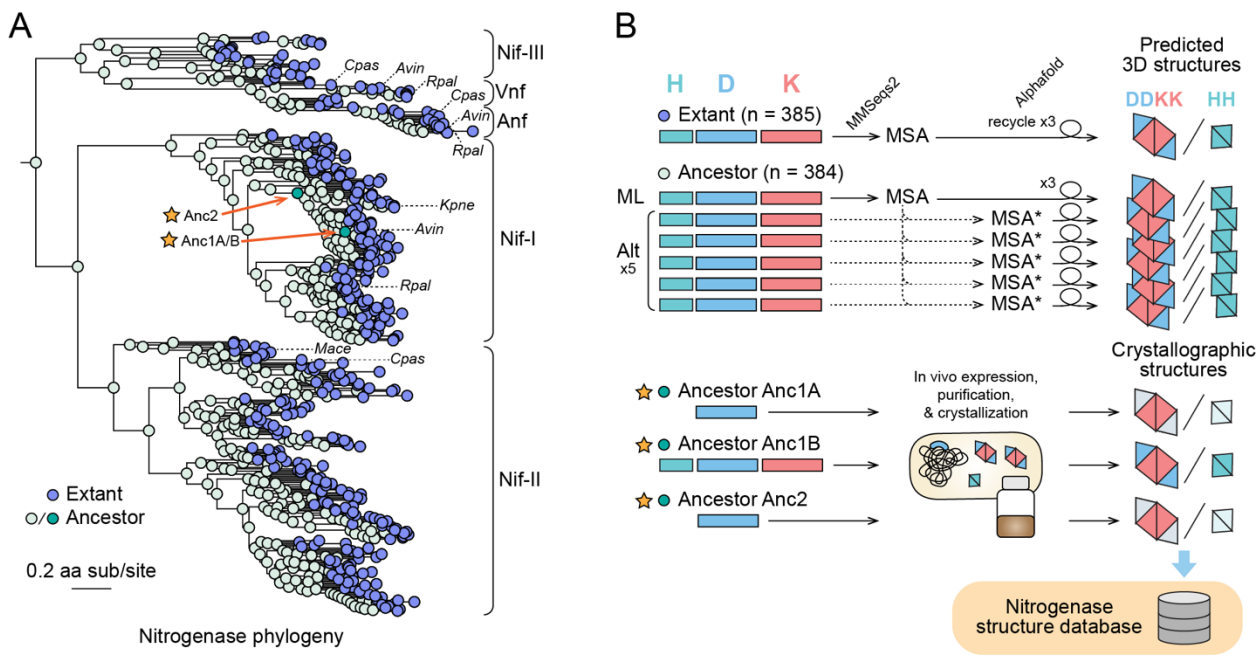

Figure 1. Overview of the methodological pipeline for massive protein structure prediction and experimental crystallographic analysis of extant and ancestral nitrogenase enzymes. A) Nitrogenase phylogeny built from concatenated Nif/Vnf/AnfHDK protein sequences. Clades are labeled according to the nomenclature used by Raymond et al. 2004. (Nif-I: Group I Nif, Nif-II: Group II Nif, Nif-III: Group III Nif). Nif, Vnf, or Anf homologs from select model organisms are labeled with dashed lines (*Avin*: *Azotobacter vinelandii*, *Cpas*: *Clostridium pasteurianum*, *Kpne*: *Klebsiella pneumoniae*, *Mace*: *Methanosarcina acetivorans*, *Rpal*: *Rhodopseudomonas palustris*). Anc1A/B and Anc2 ancestors targeted for crystallographic analysis are labeled with stars. **B)** Graphical overview of the pipeline for nitrogenase protein structure prediction and crystallization (see Materials and Methods for further details). Colored rectangles correspond to protein sequences for H, D, and K subunits. For each ancestral node, protein structures were predicted for the most likely ancestral sequence ("ML") and five alternative sequences ("Alt") reconstructed based on the site-wise posterior probability distributions in the ancestral sequence. Ancestors Anc1A and Anc2 hybrid enzymes were crystallized containing an ancestral NifD subunit and WT NifH and NifK subunits (WT subunits indicated by lighter color). All predicted structures are publicly available at https://nsdb.bact.wisc.edu.

We employed a high-throughput, AlphaFold-based computational pipeline to generate multimeric HH and DDKK structural predictions for each nitrogenase homolog in our dataset (**Figure 1B**). For each of the 385 extant targets, a single structure was predicted. For each of the 384 ancestral targets, six structures were predicted, sourced from the single most likely ancestral protein sequence and five alternative ancestral sequences generated by substitutions at ambiguously reconstructed sites (see **Materials and Methods**). In total, 2,689 unique extant and ancestral nitrogenase variants were targeted. All structures were generated in approximately 805 hours, including GPU computations and MMseqs2 alignments performed using two different protocols: one for extant or most likely ancestral sequences, and another for ancestral variants. HH and DDKK structures were predicted independently, resulting in a total of 5378 individual nitrogenase protein structures generated from our pipeline. For comparison, the Protein Data Bank currently contains structures of nitrogenase homologs (from seven different organisms, spanning 15 HH and 24 DDKK 104 structures) (**Table S1**).

We found that all nitrogenase structures were predicted with high confidence, assessed by the predicted Local Distance Difference Test (pLDDT) (Jumper et al., 2021; Mariani et al., 2013). pLDDT scores ranged between ~82 and 98 (**Figure S1**). Within this range, confidence was lowest for H-subunit structures (median pLDDT ≈ 88) and highest for K-subunit structures (median pLDDT ≈ 98). Furthermore, we built a public web database, "Nitrogenase Structural DB," (accessible at https://nsdb.bact.wisc.edu), to host all nitrogenase structures predicted in this study. Structures are visualizable, downloadable (PDB format files), and can be searched by sequence or node ID (annotated on the nitrogenase phylogeny in **Figure S2**), NCBI taxon ID, and species name.

Ancient structural insertion events distinguish major nitrogenase clades

We observe that predicted nitrogenase structures have sufficient variations to resolve phylogenetic relationships. To assess structural variation, we calculated the Root Mean Squared Deviation (RMSD) using US-Align (Zhang et al., 2022). Although in this case, this

203 metric indicates overall small differences (< 3 Å) (**Figure 2A**), it still strongly correlates with
204 sequence identity (**Figure 2B**), indicating that even minor structural variations can
205 recapitulate sequence-based phylogenetic distinctions. Hierarchical clustering of
206 nitrogenase structures by pairwise RMSD largely reproduces the major clades (Nif-I, Nif-II,
207 Nif-III, Vnf/Anf) (**Figure 2C)**. To confirm that this correlation between sequence identity and
208 structure is not simply an artifact resulting from alignment length bias, we applied two
209 complementary metrics: $RMSD_{100}$ (Carugo and Pogor 2001) and TM-score (Zhang and
210 Skolnick, 2004). $RMSD_{100}$ is a normalized variant of RMSD designed specifically to correct
211 structural alignment length bias. The results obtained with these alternative metrics (**Figure**
212 **S3**) were similar to the results of **Figure 2**, differing primarily by a scaling factor of
213 approximately one-third.
214
215 TM-score enables a global comparison of structural similarity ranging between 1 (identical)
216 and 0 (no similarity). All DDKK structures have relatively high pairwise TM-scores (> .84); TM-
217 score > 0.5 indicates comparable folding (Xu and Zhang, 2010) and hierarchical clustering of
218 nitrogenase structures by pairwise TM-scores largely reproduces the major clades (Nif-I, Nif-
219 II, Nif-III, Vnf/Anf) observed in the protein sequence phylogeny (**Figure S4**).
220
221 Upon finding a correlation between structure similarity and phylogenetic relationships, we
222 developed a method to sample the inherent uncertainty of ASR outputs by generating
223 sequences with alternative residues at ambiguous positions (see **Materials and Methods**).
224 We predicted the resulting structures and compared the variability of these outputs against
225 the maximum-likelihood sequence. **Figure S5A** shows the overall distribution of RMSDs for
226 randomized variants against the maximum likelihood variant. RMSD for most divergent
227 variants fall within two angstroms, indicating very high structural similarity. As shown, the
228 RMSD does not correlate with the distance to the tree's root, indicating no systemic
229 relationship between structural similarity and tree depth (**Figure S5B**). Inferred structures,
230 even those derived from deep nodes within the phylogenetic tree, are robust against the
231 statistical uncertainty of ancestral sequence inference. Finally, we computed the all-vs-all

RMSD at the DDKK active sites (FeMoco and 8Fe7S; **Figures S6A** and **S6B**). The results indicate that the active sites have high structural conservation, especially around the 8Fe7S cluster (RMSD < 0.5Å, **Figure S6C**), which could be expected given the tight coordination of the metal cofactor with its surrounding cysteines. It is worth noting that AlphaFold2 does not use information about metal cofactors when building the model.

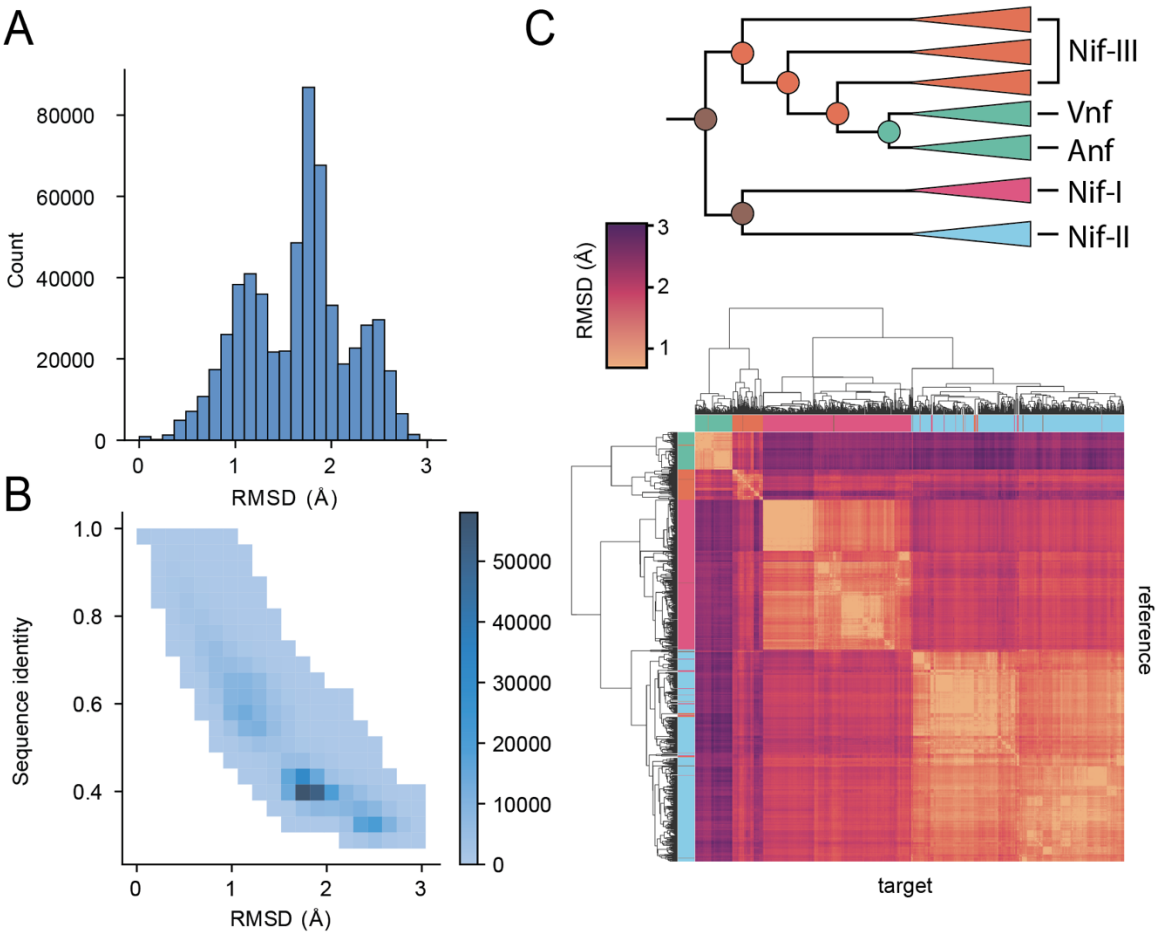

Figure 2. Global analyses of nitrogenase DDKK sequence and structural diversity. A) Root Mean Squared Deviation (RMSD) distribution on paired nitrogenase alignments across extant and ML-ancestor pairs. **B)** Sequence identity and structural similarity (quantified by root mean square deviation (RMSD) of aligned predicted structures) distribution of paired nitrogenase alignments. **C)** Hierarchical clustering of predicted nitrogenase structures based on structural similarity (RMSD). Each tile in the heatmap corresponds to the RMSD between two nitrogenase structures.

Our global comparisons of predicted nitrogenase structures suggest that, despite relatively high conservation, certain structural distinctions between major clades evolved early and persisted in the extant descendants. We examined our dataset to identify these distinctions and map them to the evolutionary trajectory on the nitrogenase family history.

The structural predictions identified three insertion events that distinguish nitrogenase variants. These insertions likely contribute significantly to the phylogenetic signal encoded within structural features (**Figure 3A**): a D-subunit C-terminus extension (~60 residues) unique to Anf (**Figure 3B**), a K-subunit N-terminus extension (~8 to 40 residues) unique to Nif-I (**Figure 3C**), and a D-subunit insertion (~55 residues) unique to Nif-II (**Figure 3D**). The relationships between these insertion events and the major nitrogenase clades are also associated with the significant differences in host ecology and nitrogen fixation genetics across clades.

We found that the K-subunit N-terminus is predominantly found in oxygen-tolerant, aerobic or facultatively anaerobic diazotrophs, whereas the D-subunit insertion of Nif-II is mostly found in anaerobic diazotrophs. Our ancestral reconstructions constrain the initial appearance of each insertion to the branch leading to the last common ancestor of the clade. Both D-subunit insertions appear within a single branch and thereafter remain relatively conserved in length. By contrast, we observed that the K-subunit N-terminus insertion evolved progressively in length (**Figure 3B, 3D**) through several deep-time ancestors of the Nif-I clade (**Figure 3C).** The elongation of this insertion recapitulates trends in the K-subunit N-terminus lengths of extant Nif-I homologs, where early-diverged lineages have short (~8 residue) extensions and later-diverged lineages have longer (~40 residue) extensions. The extent to which the N-terminus of the K-subunit affects functional importance in relation to oxygen demand warrants further empirical investigation. As a side observation, our analysis detected no major deletion events in the evolutionary history of nitrogenase. Oldest ancestors (including the last common ancestor of the entire nitrogenase

family) and Group III nitrogenases are among the shortest in sequence length and do not contain any of the insertions described above.

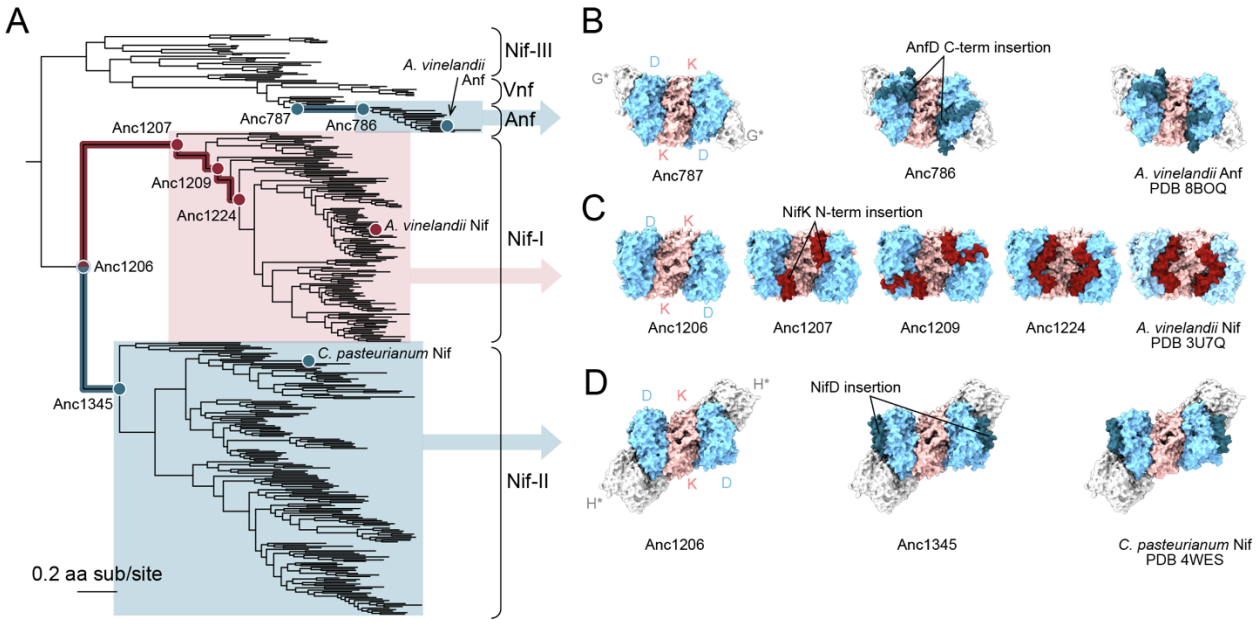

Figure 3. Nitrogenase structure variation in a phylogenetic context. A) Nitrogenase protein phylogeny. Branches and ancestral nodes corresponding to structural insertion events, as well as representative extant variants conserving those insertions, are highlighted and/or labeled. Clade and node colors correspond to the subunit for which an insertion is observed (i.e., blue for the D subunit, red for the K subunit). **B)** Elongation of the NifD C-terminus coincident with the origin of the Anf clade. **C)** Progressive elongation of the NifK N-terminus through the early evolution of the Nif-I clade. **D)** Insertion within NifD coincident with the origin of the Nif-II clade. **B-D)** All visualized structures are predicted unless otherwise specified with the corresponding Protein Data Bank identifier. Bound G- and H-subunit structures were not predicted together with the NifDK structures and are thus indicated with an asterisk. The binding positions of the G- and H-subunit structures are inferred based on alignment with PDB 8BOQ (Trncik et al., 2023) and PDB 1M34 (Schmid et al., 2002), respectively.

294 The locations of these insertions within the nitrogenase structure and their encoding genes
295 provide insight into their possible functional contributions. First, all insertions are exposed
296 at the surface of the DDKK protein complex (**Figures 3B-D**), and all comprise varying
297 proportions of helix or loop regions in both extant and ancestral structures. The position of
298 the D-subunit C-terminus extension (**Figure 3B**) and the K-subunit N-terminus extension
299 (**Figure 3C**) is similar. Both line the edges of the interface between D- and K- subunit proteins,
300 possibly stabilizing the complex. It is possible that the N-terminus of NifK wraps around the
301 entry point of small-molecules (Barney et al., 2009; Morrison et al., 2015), and help in
302 reducing access for oxygen in Nif-I hosted by aerobes. The D-subunit insertion in Nif-II
303 instead protrudes near the inferred HH binding site. Though no crystallographic structures
304 of a bound HH-DDKK nitrogenase complex are available for Nif-II homologs (and we did not
305 predict bound structures here), the Nif-II D-subunit insertion is likely capable of H-subunit
306 interactions based on alignment with the binding site in Nif-I nitrogenases (**Figure 3D**).
307
308 **Phylogenetic trends of nitrogenase DDKK complex structures**
309
310 Following the identification of structural innovations in the history of nitrogenases, we
311 investigated evolutionary trends in calculated structural attributes for DDKK structures,
312 including surface properties and intermolecular contacts (**Figure S7A**). Broadly, we found
313 that many structural features carry clade-specific trends and/or are associated with the
314 presence of the insertions described above. For example, total surface area (summed over
315 each protein subunit in the DDKK complex) is distinct between Nif-I, Nif-II, and Nif-
316 III/Vnf/Anf, and greatest in the Nif-II clade (**Figure S7B**). Trends in surface area are likely
317 driven in part by the presence and conformation of the surface-exposed insertions. We also
318 examined attributes related to protein-protein interactions within DDKK. In all cases, we
319 found that trends in these attributes were influenced most strongly by the presence of the K-
320 subunit N-terminus extension in the Nif-I clade. Nif-I homologs (and particularly those with
321 longer K-subunit extensions) have the largest number of intersubunit molecular contacts,
322 and intersubunit binding affinities, as calculated with Prodigy (Vangone and Bonvin 2015)

323 (**Figure S7C, Figure 4B**). These patterns support our hypothesis that the K-subunit N-
324 terminal extension (**Figure 3C** and **4C**) contributes positively to DDKK complex interaction
325 strength and stability. By contrast, homologs with the least number of intersubunit contacts
326 and weakest binding affinities are Nif-III nitrogenases, which notably contain none of the
327 described insertions. We also observed an increase in the total surface area of the
328 nitrogenase complex within the Nif-II group (**Figure S7B**). The only structural attribute that
329 we found correlated with ancestor age was the proportion of charged residues across DDKK
330 proteins (**Figure S7E**). Accordingly, we observed a weak negative correlation between
331 phylogenetic distance from the root and proportion of charged residues (Spearman R ≈ -
332 0.261). We did not detect similar trends for other attributes. Rather, the attributes of
333 nitrogenase ancestors largely resemble those of their descendants in a clade-specific
334 manner. These results reinforce the idea that major structural distinctions were set early in
335 nitrogenase history and conserved in respective clades.
336
337
338

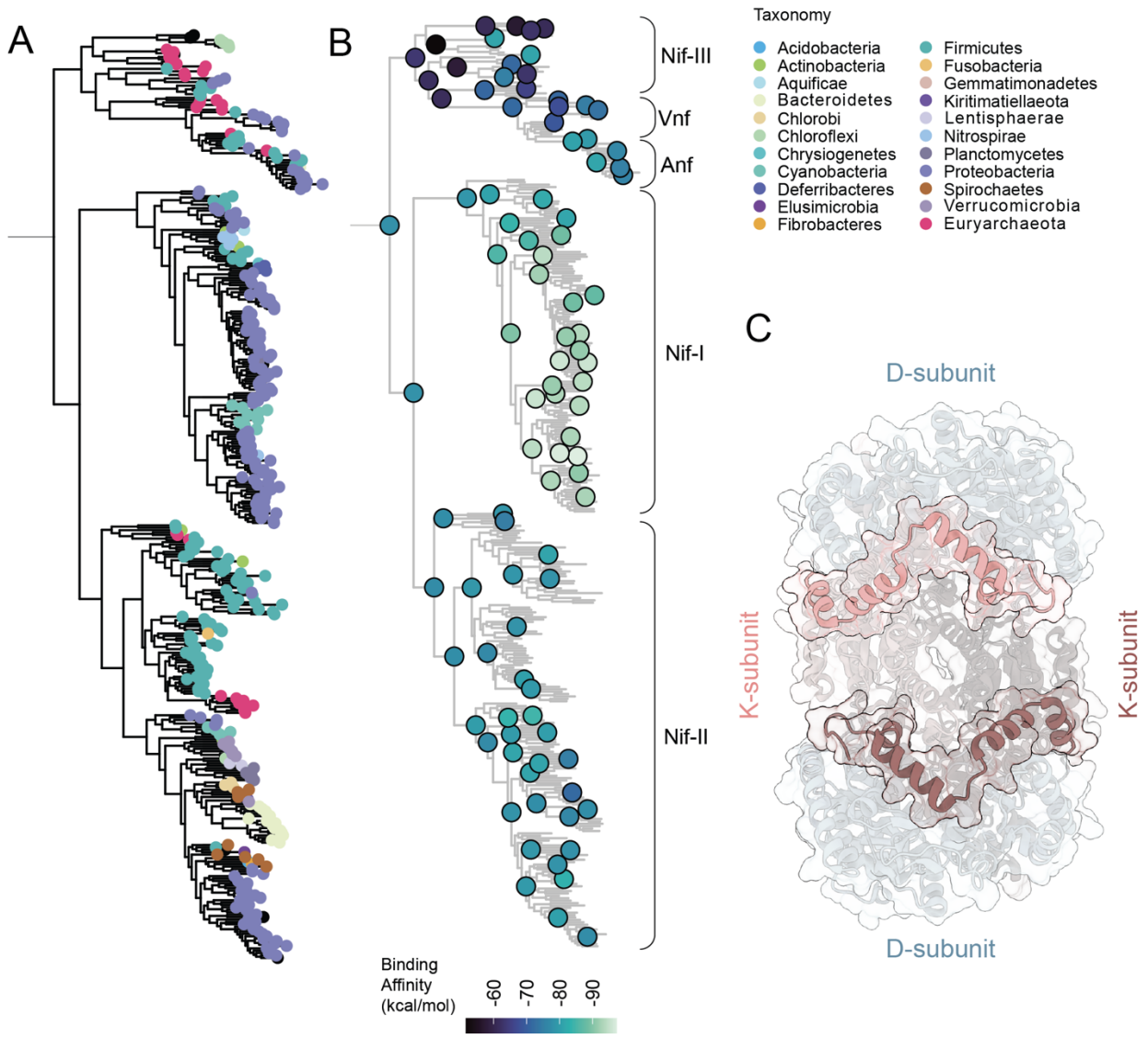

Figure 4. Phylogenetic patterns of nitrogenase structural attributes calculated across DDKK proteins. A) Node taxonomy overview. **B)** Binding affinity prediction for extant and ancestral nitrogenase nodes mapped to the nitrogenase phylogeny. **C)** Structure overview of the D and K subunit interactions around the N-terminal insertion in Nif-I nitrogenases. Structure corresponding to *Azotobacter vinelandii* NifDK (PDB code: 3U7Q). Note: We reduced the number of displayed phylogenetic nodes to mitigate visual overcrowding, refer to **Supplementary Figure S7** for a complete visualization of all nodes.

349 **Crystalized nitrogenase ancestors confirm evolutionary substitutions at the HH-DDKK**
350 **interface**
351 In this work, we created a large dataset of Nitrogenase structures, expanding upon the
352 limited structural reference data that the prediction methods rely on. Therefore, we
353 continued our structural investigation of nitrogenase evolutionary history by crystallographic
354 characterization of select nitrogenase ancestors which helped us assess the accuracy of the
355 structure prediction methods. Three previously characterized ancestral variants were
356 targeted from phylogenetic nodes within the evolutionary lineage of *A. vinelandii* Nif (wild-
357 type, "WT") (**Figure 5A**) (Garcia et al., 2023): Anc1A, Anc1B and Anc2. Anc1A and Anc1B are
358 the same age, whereas Anc2 is reconstructed from an older phylogenetic node. WT, Anc1,
359 and Anc2 belong to a Mo-nitrogenase clade, formerly designated as 'Group I' by (Raymond
360 et al., 2004). This clade includes homologs from various aerobic and facultatively anaerobic
361 organisms, such as proteobacteria and cyanobacteria (Garcia et al., 2023). An estimated
362 maximum age of approximately 2.5 billion years for Group I nitrogenases —and
363 consequently for Anc1 and Anc2— is inferred from the timing of the GOE (Garcia et al., 2023;
364 Lyons et al., 2014). As an additional distinction, Anc1A and Anc2 contain an ancestral D-
365 subunit protein complexed with WT H- and K-subunit proteins. Anc1B, by contrast, is
366 composed of ancestral H-, D-, and K-subunit proteins. Ancestral variants preferentially bind
367 to the ATP cofactor (Harris et al., 2024) and exhibit the same general mechanism for N_2
368 binding and reduction conserved across all known nitrogenases to date, but with decreased
369 activity for reduction of N_2 or other substrates (Garcia et al., 2023). Thus, the sequence
370 differences between ancestors and WT are sufficient to generate phenotypic changes.
371
372 Ancestral nitrogenase DDKK proteins were isolated from genomically modified strains of *A.*
373 *vinelandii* (Garcia et al., 2023), purified by chromatographic methods, and crystallized under
374 anoxic conditions (**Figure S8**) (see **Materials and Methods**). We found that ancestral
375 nitrogenases required different conditions for crystallization relative to that required for
376 modern proteins. Specifically, all ancestors crystallized at pH 6.5, which is unusually acidic
377 for nitrogenase (Peters et al., 1997; Spatzal et al., 2011). In addition, all ancestral proteins

378 crystallized in space group *P* $2_12_12_1$, which differs from the space groups for the same
379 proteins in *A. vinelandii* NifDK (i.e., WT; space group *P* 2_1), *Clostridium pasteurianum* Nif
380 (space group *P* 2_1), and *Klebsiella pneumoniae* NifDK (space group *C* 2). We observed one
381 ancestral DDKK protein per asymmetric unit. All three ancestors showed anisotropic
382 diffraction: 2.65 to 2.93 Å for Anc1A, 1.82 to 2.8 Å for Anc1B, and 1.82 to 2.42 Å for Anc2. All
383 structures were solved by molecular replacement using WT *A. vinelandii* (PDB 3U7Q).
384
385 As in extant nitrogenase proteins, each ancestral D- or K-subunit structure exhibits three
386 Rossmann-fold domains (**Figure 5B**). The P-cluster, an 8Fe7S metallocluster that mediates
387 electron flow to the active site, is located on the pseudo-twofold axis relating the structurally
388 similar D- and K-subunits and was modeled in the all-ferrous P^N-state. Both the P-cluster and
389 the active-site cofactor, FeMoco, were modeled with full occupancy. Global structural
390 differences between the ancestors and WT *A. vinelandii* DDKK are relatively minor, and
391 within the range of variation for available extant structures. Among all ancestors, root mean
392 square deviations (RMSD) for all atom positions are largest between Anc1B and WT DDKK
393 (0.36 Å; compared to RMSD of 0.50 Å between *A. vinelandii* and *K. pneumoniae* DDKK). This
394 observation may be explained by the larger number of amino acid substitutions between the
395 fully ancestral Anc1B complex and WT, whereas only the D-subunit contains substitutions
396 in the Anc1A and Anc2 hybrid complexes (**Figure 5C**). Among the two hybrid ancestors, the
397 RMSD between Anc2 and WT (0.24 Å) is smaller than that between Anc1A and WT (0.31 Å).
398 Furthermore, we observe that the position of the twofold symmetry axis relating the two DK
399 heterodimers of a single MoFe protein shifts slightly between Anc1B and WT (**Figure S9**). The
400 same was previously observed in VFe protein (Sippel and Einsle, 2017) and FeFe protein
401 (Trncik et al., 2023) and may reflect slight differences in the interface between the two copies
402 of NifK.
403
404

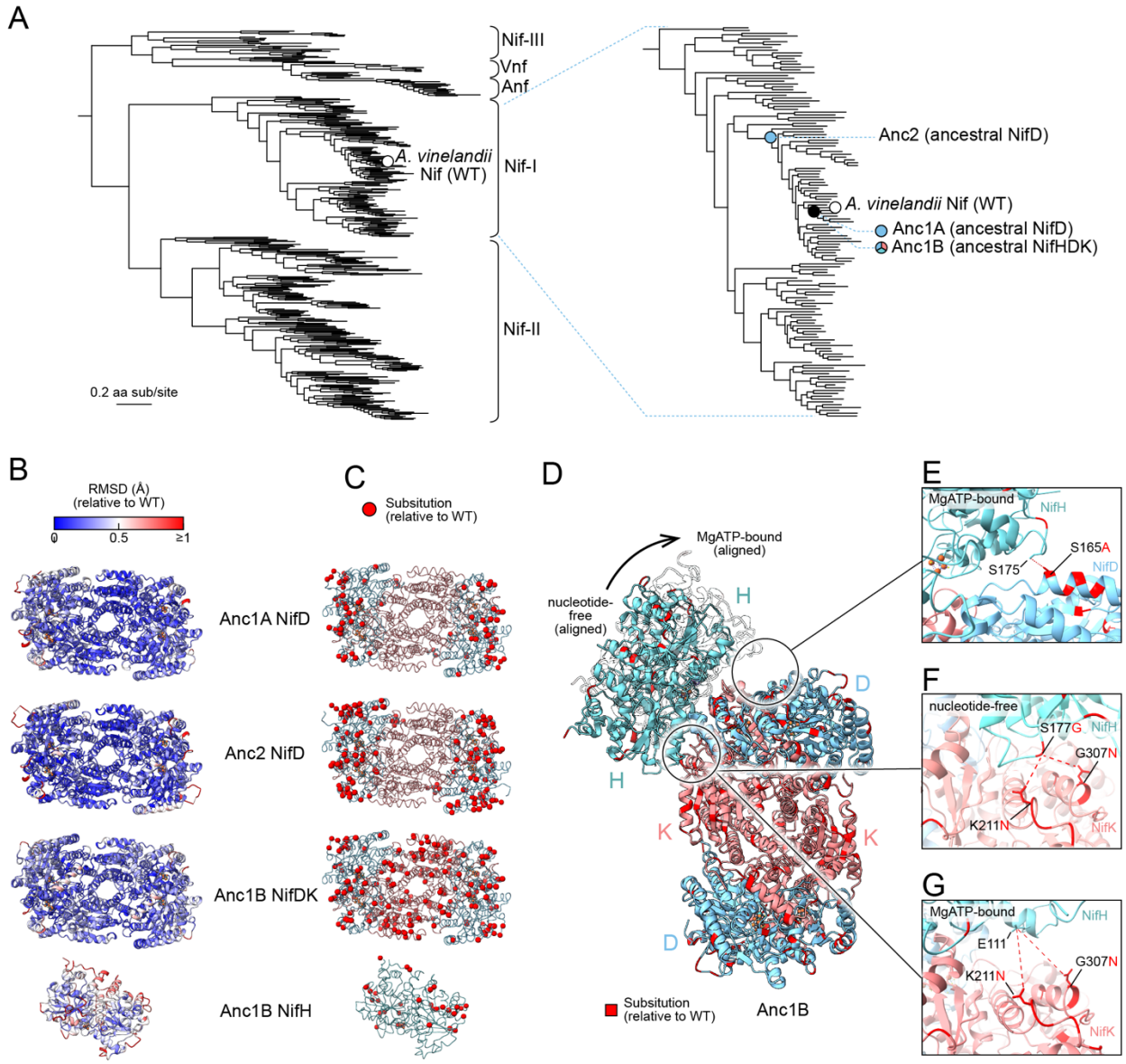

Figure 5. Crystal structures of targeted ancestral nitrogenases. A) Nitrogenase protein phylogeny from which nitrogenase ancestral proteins were reconstructed and crystallized for structural characterization. Major clades are labeled following (Garcia et al., 2023). **B)** Residue-level RMSD between the crystallized ancestral and wild-type (*A. vinelandii*) NifDK/NifH structures. **C)** Spatial distribution of ancestral amino acid substitutions relative to WT. Crystallized protein complexes for Anc1A and Anc2 contain ancestral NifD and WT NifH and NifK. Therefore, ancestral substitutions are only in the NifD subunit for these structures. **D)** Ancestral amino acid substitutions within the NifH-NifDK interface of Anc1B.

Bound positions of NifH are inferred by alignment with either the nucleotide-free (PDB 2AFH (Tezcan et al., 2005)) or MgATP-bound (PDB 7UT8 (Rutledge et al., 2022)) structures. **E-G)** Close views of specific, ancestral amino acid substitutions that are inferred to impact NifH-NifDK interactions.

In our previous study, we found that ancestral nitrogenases had decreased N_2-reduction activity *in vitro* relative to WT (Garcia et al., 2023). The generation of crystallographic structures for these ancestors provide an opportunity to examine any structural features that might be responsible for the decrease in activity. Given that most ancestral DDKK substitutions were not close to metalloclusters, we hypothesized that those situated at the DDKK-HH interface might have modified the activity of the ancestors by impeding protein-protein interactions that are necessary for electron transfer to the nitrogenase active site, as well as cofactor loading and complex maturation (Burén et al., 2020). Snapshots of crystal structure interface, highlighting mutations and residue interactions are shown in **Figure 5E-G**.

Notably, a non-conserved Nif-D α-helix (positions 165-174 in Anc1B) lies within interacting distance of HH. This helix contains several substitutions across all ancestors, albeit the positioning of polar amino acids differs among Anc1A, Anc1B, and Anc2. Herein, the D-subunit S165A substitution is consistently present across all ancestors and may influence interactions with HH S175 in the ATP-bound complex, potentially reducing stability (**Figure 5E)**. Additionally, two unique single-point substitutions in the ancestral K-subunit—K211N and G307N, specific to Anc1B—are likely to impact interactions with HH E111 in the ATP-bound state (**Figure 5G**). Positioned centrally within the DDKK-HH interface, these substitutions alter the electrostatic surface of HH. In Anc1B, the substitution HH S177G is located near the interface with NifDK, with G307N in Anc1B NifK identified as a potential interaction partner (**Figure 5F**).

443 Predictions for nitrogenase variants had high similarity (average RMSD < 1.4 Å) to
444 experimentally determined structures (**Figure S10B**). This was also true for the ancestral
445 nodes selected for crystallographic analysis in this study (**Figure S10A, Figure S10C**),
446 notably, none of which were included in the structural database used for AlphaFold training.
447 Beyond the limited number of substitutions within the interfaces of Anc1A/B and Anc2, the
448 changes in the ancestral enzymes were distributed quite evenly across the protein subunits.
449 Importantly, we did not observe substitutions within regions considered important for
450 nitrogenase catalysis (e.g., close to metalloclusters), nor were any sites overlapping with
451 those previously targeted for mutagenesis experiments.
452
453 **DISCUSSION**
454
455 In this study, we combined targeted crystallization experiments of ancestral nitrogenase
456 enzymes with large-scale 3D structure predictions spanning both present-day and ancestral
457 nitrogenase diversity. Building on the recent expansion of predicted structures for extant
458 proteins, this work marks a pioneering effort to apply these methods to explore the ancestry
459 of an entire enzyme family. The three crystallographic and 5,378 predicted structures
460 generated in this study represent approximately a 50-fold increase in available nitrogenase
461 complex structures, and importantly includes an extensive sampling of historical structural
462 variation.
463
464 Both crystallographic and predicted structure of nitrogenase ancestors highlight the long-
465 term evolutionary conservation of core structural features, consistent with previously
466 characterized preservation of mechanistic aspects in nitrogenase catalysis (Garcia et al.,
467 2023, Harris et al. 2024). Nevertheless, historical structural variation sufficiently
468 distinguishes major nitrogenase clades independent of protein sequence information and
469 this phylogenetic variation is driven primarily by different insertion events that occurred early
470 in nitrogenase evolutionary history. For example, we show that the extension of the K-subunit
471 N-terminus emerged during the early diversification of Group I nitrogenases coinciding with

472 the proliferation of nitrogenases into oxygen-tolerant bacterial lineages and the planetary
473 rise in atmospheric oxygen.
474
475 Nitrogenase structural features that might be associated with increased compatibility
476 between the oxygen-sensitive nitrogenase enzyme and oxygen-tolerant host organisms are
477 of significant interest given efforts to transfer N-fixation genes to crops (Vicente and Dean,
478 2017). Furthermore, if such structural features are in fact required for compatibility with
479 oxygen tolerant hosts, their emergence would represent a pivotal transition in Earth history
480 given the importance of aerobic N-fixers for present-day global nitrogen cycling. For
481 example, *Trichodesmium*, a single genus of oxygenic phototrophic cyanobacteria, today
482 contributes 60-80 Tg of the total 100-200 Tg fixed N produced annually in modern oceans
483 (Bergman et al., 2013). It was also proposed that the O_2 sensitivity of nitrogenase limited the
484 extent of oxygen production by primary producers (Mrnjavac et al., 2024).
485
486 The occurrence of the NifK N-terminus extension in oxygen tolerant diazotrophs, combined
487 with increased inter-subunit interactions and affinity predicted by our *in-silico* analyses,
488 points to a possible functional role in response to the planetary rise of oxygen. Aerobic
489 diazotrophs today utilize a variety of oxygen protection strategies for N-fixation. These
490 include physical separation of nitrogenases into anaerobic, differentiated cells, temporal
491 decoupling of N-fixation from oxygen production in cyanobacteria, and increased rates of
492 aerobic respiration (Mus et al., 2019; Robson and Postgate, 1980). Further, the Shethna
493 protein II was found to shield nitrogenase from temporary oxygen exposure by forming a
494 complex with nitrogenase proteins and rendering the enzyme inactive yet oxygen-tolerant
495 (Schlesier et al., 2016). Nevertheless, evolutionary modifications to the nitrogenase enzyme
496 itself that improve its compatibility in aerobic diazotrophs have not yet been identified. The
497 NifK N-terminus extension discussed here may play a direct role, by stabilizing oxygen-
498 sensitive, intersubunit protein interactions within the nitrogenase enzyme. Alternatively, it
499 might play an indirect role by shaping potential interaction sites with accessory proteins (like
500 the Shethna protein II) that improve oxygen protection. Targeted experimental studies aimed

at the NifK N-terminus, as well as other insertions that we identified in our work, in both extant and ancestral nitrogenases could illuminate potential functional roles. Given the dramatic environmental shifts brought by the Great Oxidation Event (GOE), it remains uncertain to what extent these emergent changes were beneficial at the time of their appearance. Within the framework of constructively neutral evolution, these changes may have initially been neutral, only later acquiring adaptive value as environmental conditions or biological contexts shifted (Brunet and Doolittle, 2018).

We identify major structural transitions in the history of nitrogenase enzymes that can serve as a foundation for future experimental studies examining the interplay between molecular and ecological factors that shaped the evolution of this globally important enzyme. Nonetheless, the insights gained by these structure prediction methods carry inherent uncertainty. For example, accurate active-site side chain prediction would require incorporation of metal-cofactors into the active site, which is currently missing in our AlphaFold2-based pipeline. Future versions of AlphaFold, with enhanced capacity to capture finer-scale features (e.g., side chain structures), may enable more detailed estimates and reveal subtle structural and historical variations.

Our study shows that despite a conservation of a core multimeric structure, nitrogenase evolved novel structural features, coinciding with the key environmental transitions in Earth's history. Subtle, modular structural changes away from the active site likely played a crucial role in the evolution and long-term persistence of nitrogenases over geologic time. These findings may also indicate a prominent role for transient regulatory adaptations in response to planetary-scale environmental transitions shaping the course of protein evolution. Reconstructing lost evolutionary histories of globally essential enzymes, including the structural changes during key planetary transitions provides critical insights into how these proteins persist over long timescales and opens avenues for future research into the climatological and ecological factors driving protein evolution.

530 **MATERIALS AND METHODS**

531

532 **Nitrogenase ancestral sequence reconstruction, selection, and resurrection in *A.***
533 ***vinelandii***

534

535 All extant and ancestral nitrogenase protein sequences were drawn from a previous
536 nitrogenase sequence dataset (Garcia et al., 2023). Briefly, 385 nitrogenase sequences and
537 385 outgroup sequences were curated from the NCBI nr database by BLASTp for
538 phylogenetic tree building and ancestral sequence reconstruction. For ancestors Anc1A and
539 Anc2, aligned sequences (MAFFT v7.450 (Katoh and Standley, 2013) were trimmed (trimAl
540 v1.2 (Capella-Gutiérrez et al., 2009)) and both tree reconstruction and ancestral sequence
541 inference were performed by RAxML v8.2.10 (Stamatakis, 2014) with the LG + G + F
542 evolutionary model. Model selection was performed by ModelFinder (Kalyaanamoorthy et
543 al., 2017) in IQ-TREE v.1.6.12 (Nguyen et al., 2015). For ancestor Anc1B, the untrimmed
544 sequence alignment was used for tree reconstruction by RAxML and ancestral sequence
545 inference by PAML v4.9j (Yang, 2007) using the LG + G + F model. This second analysis was
546 conducted due to concerns that RAxML v8.2 does not perform full, marginal ancestral
547 sequence reconstructed as described by (Yang et al., 1995). Anc1A and Anc1B are
548 equivalent in their set of descendants and, despite the difference in reconstruction
549 methods, have NifD proteins with 95% identity (**Figure S11**).

550

551 In addition to the ancestral sequences described above, we calculated the statistical
552 uncertainty of ASR by generating five alternative variants for each ancestor in the nitrogenase
553 phylogeny. These include: one variant that contains the second most probable residue at
554 each ambiguous position (probability of most probable ancestral state < 0.7) ("*altall*") and
555 another four variants selected by randomly sampling the ASR posterior probability
556 distribution at each ambiguous site ("*alt2*," "*alt3*," "*alt4*," "*alt5*"). The *altall* and *alt2-5*
557 sequences correspond to the alternative sequences considered in **Figure 1**.

558

559 Engineering of *A. vinelandii* strains was performed as previously described, replacing the
560 native WT *nifD* gene with an ancestral variant (Garcia et al., 2023). A. *vinelandi*i WT ('DJ'),
561 DJ2278 (ΔnifD::KanR), and DJ2102 (Strep-II-tagged WT NifD were generously provided by
562 Dennis Dean (Virginia Tech) for strain engineering (Santos, 2018). Following genomic
563 integration of ancestral nitrogenase genes into *A. vinelandii*, transformants were passaged
564 at least three times to ensure phenotypic stability prior to storage at –80 °C in phosphate
565 buffer containing 7% DMSO. Verification of the engineered strains was performed by Sanger
566 sequencing of the *nifD* region as well as whole genome sequencing.
567
568 **Expression and purification of nitrogenase proteins *A. vinelandii***
569
570 Cells from DMSO stock were directly grown in a pre-culture using liquid Burk-medium
571 containing ammonium sulfate at 37 °C and 180 rpm until an optical density at 600 nm of at
572 least 1.5 was reached. Further a second pre-culture without ammonium-sulfate was grown
573 under the same conditions and used to inoculated 500 mL main cultures at 30 °C and 180
574 rpm. Cells were harvested when an optical density at 600 nm of 1.7-2.5 was reached.
575
576 **Anc1A:** Cells were grown in a 10 L Eppendorf BF120 bioreactor at 30 °C in Burk-medium
577 containing urea as the fixed N source. Dissolved oxygen (DO) was held at 20% using lab air
578 at a constant 10 SLPM and cascade control of agitation (100 – 500 RPM) and supplemental
579 O_2 (0 – 30%). At an optical density at 600 nm of ~8 the cells were pelleted by centrifugation
580 and resuspended in Burk-medium without a fixed N source to induce expression of
581 nitrogenase. Cells were allowed to express for 4 hours with DO held at 10% with the same air
582 flow and cascade control before harvesting.
583
584 **Isolation of MoFe proteins:** MoFe protein is oxygen sensitive therefore all consecutive steps
585 were conducted either in an anoxic chamber (Coy Laboratories) under 95% N_2 and 5% H_2
586 atmosphere or using modified Schlenk techniques under N_2 flow. Anc1 contains two
587 additional crystallographic forms, Anc1A and Anc1B; the two are the same age, and Anc1A

588 contains an ancestral D-subunit protein complexed with WT H- and K-subunit proteins,
589 Anc1B is by contrast is composed of ancestral H-, D-, and K-subunit proteins. They were
590 isolated as follows:
591
592 **Anc1B (MoFe and Fe protein):** Cell pellets were resuspended in lysis buffer (50 mM Tris/HCl
593 at pH 7.4, 2.5 mM $Na_2S_2O_4$) and opened at 1500 bar in a SUP Maximator HPL-6 at 2 °C under
594 anoxic conditions. The cell free extract was centrifuged at 87,000 x g at 4 °C for 60 minutes,
595 filtered through a 0.45 µM syringe filter (Filtropur) and were loaded onto two equilibrated 5
596 mL HiTrap Q HP columns. After washing with 12.5 % elution buffer (50 mM Tris/HCl at pH
597 7.4, 1000 mM NaCl, 2.5 mM Na2S2O4), the amount of elution buffer was increased in 2.5 %
598 increments up to 35 %, followed by steps at 40 and 50 %. MoFe and Fe protein eluted after
599 300 mM NaCl and 325 mM NaCl respectively. Step elution with 40 % elution buffer was used
600 to concentrate. The proteins were loaded separately onto a size exclusion chromatography
601 (Superdex S200, GE Healthcare) equilibrated with 20 mM Tris/HCl at pH 7.4, 200 mM NaCl,
602 2.5 mM $Na_2S_2O_4$, concentrated (Vivaspin 20, 50 or 30 kDa MWCO, Sartorius) and flash frozen
603 in liquid nitrogen. A second size exclusion chromatography step (Superdex 200 Increase
604 10/300 GL) was carried out with MoFe protein, which was subsequently concentrated, and
605 flash frozen in liquid nitrogen.
606
607 **Anc1A:** Cell pellets were resuspended in lysis buffer (50 mM Tris/HCl pH 7.9, 20% glycerol
608 (v/v), 500 mM NaCl, and 2 mM $Na_2S_2O_4$). Cells were lysed with a Nano DeBee 45-2 High
609 Pressure Homogenizer and then ultra-centrifugated to produce cell free lysate. Cell free
610 lysate was loaded onto an equilibrated (50 mM Tris/HCl pH 7.9, 500 mM NaCl, and 2 mM
611 $Na_2S_2O_4$) 20 mL StrepTactinXT 4Flow column. The column was washed with 2 column
612 volumes of equilibration buffer, followed by elution of the protein in a single fraction with
613 elution buffer (50 mM Tris/HCl pH 8, 150 mM NaCl, 50 mM biotin, and 2 mM $Na_2S_2O_4$). Protein
614 was then loaded on an equilibrated (20 mM Tris/HCl at pH 7.4, 200 mM NaCl, 2.5 mM
615 $Na_2S_2O_4$) size exclusion chromatography (Superdex 200 Increase 10/300 GL). Subsequently
616 it was concentrated, and flash frozen in liquid nitrogen.

617 **Anc2:** Cell pellets were resuspended in lysis buffer (20 mM Tris/HCl at pH 7.4, 200 mM NaCl,
618 2.5 mM $Na_2S_2O_4$) and opened, centrifuged and filtered following the Anc1B protocol. The cell
619 free extract was loaded onto a with lysis buffer equilibrated Streptavidin XT 4 Flow column.
620 Through addition of 50 mM Biotin to the lysis buffer Anc2 eluted. Size exclusion
621 chromatography (Superdex S200, GE Healthcare) was equilibrated using lysis buffer and the
622 protein loaded. Protein was concentrated (Vivaspin 20, 50 kDa MWCO, Sartorius), and flash
623 frozen in liquid nitrogen until further use.
624
625 **Crystallization and data collection**
626
627 Protein crystallization was carried out in an anoxic chamber with less than 5 ppm of O_2.
628 Solutions used for crystallization were degassed using modified Schlenck techniques and
629 crystallization was carried out using the sitting-drop vapor diffusion technique in 96 well
630 plates (Anc1A and Anc1B) (Swissci 96-well two-drop plate, Hampton Research) or 24 well
631 plates (Anc2) (24-well crystallization Cryschem M plates, Hampton Research). Crystals
632 which were used for microseeding were added to 50 µl reservoir solution and a PTFE seed
633 bead (Hampton Research). The crystals were crushed by 6 times vortexing for 30 s and then
634 diluted with reservoir solution. 7.5 mM of $Na_2S_2O_4$ were added to the protein prior to
635 crystallization. 2 µl of 10 mg/ml protein solution (Anc2) were mixed with the same volume of
636 reservoir solution containing 4.5% (w/v) of polyethylene glycol 2000, 3350, 4000 and
637 polyethylene glycol mono methylether 5000, 0.1 M MES/NaOH at pH 6.5, 10% (v/v) of
638 ethylene glycol and 0.15 M Mg acetate. 0.2 µl of 1:100 diluted seed were added.
639
640 For Anc1A, 0.3 µl of 9 mg/ml protein solution were mixed with 0.4 µl of reservoir solution
641 containing 0.2 M $LiSO_4$ x 1 H_2O, 0.1 M bis-Tris/HCl at pH 6.5, 25%(w/v) of polyethylene glycol
642 3350. 0.1 µl of 1:400 diluted seed was added. 0.4 µl of 9 mg/ml protein solution (Anc1B) was
643 mixed with 0.4 µl of reservoir solution containing 0.1 M MES/imidazole at pH 6.5, 12.5% of
644 (w/v) of polyethylene glycol 1000, 3350 and (v/v) 2-methyl-2,4-pentandiol each and 0.02 M
645 of 1,6-hexanediol, 1-butanol, 1,4-butanediol, (RS)-1,2-propanediol, 2-propanol, and 1,3-

646 propanediol, each with addition of 0.1 µl of 1:400 diluted seed. 0.8 µl of 8 mg/ml Anc1B-H
647 was mixed with 0.8 µl of reservoir containing 0.1 M MES/NaOH at pH 6.5, 7.5% of (w/v)
648 polyethylene glycol 6000, 8000, 10000 each, 10%(vol/vol) isopropanol. 0.2 µl of 1:400
649 diluted seed stock was added.
650
651 Crystals were harvested into nylon loops and flash-frozen in liquid nitrogen. Diffraction data
652 for Anc2 were collected at the Swiss Light Source (Paul Scherrer Institute) on beamline
653 X06SA using an EIGER 16M X detector at an X-ray wavelength of 1.0000 Å. Diffraction data
654 for Anc1A and B were collected at the ESRF on beamline ID30B using an EIGER2 X 9M at an
655 X-ray wavelength of 0.8856 A. Diffraction date for Anc1B-H was collected at the ESRF on
656 beamline ID23-1 using an EIGER2 16M CdTe detector at an X-ray wavelength of 0.8856 A.
657
658 **Structure solution and refinement**
659
660 The crystallographic phase problem was solved using the WT structure of MoFe or Fe
661 protein, respectively, from *A. vinelandii* (PDB:3U7Q, 1G1M). A single solution was obtained
662 for the mutants using MOLREP (Anc1B, Anc1B-NifH) (Vagin and Teplyakov, 2010) or
663 PhaserMR (Anc1A, Anc2) (McCoy et al., 2007). The obtained model was refined using
664 iterative rounds of BUSTER (Anc1A, Anc2) (Blanc et al., 2004) or REFMAC5 (all) (Murshudov
665 et al., 2011) and model building in COOT (Emsley et al., 2010). **Tables S2 to S5** show the
666 statistics of the refinement process.
667
668 **Ancient and extant protein structure prediction**
669
670 Structure predictions based on Alphafold-derived methods require two steps: generating
671 deep sequence alignments and predicting structures based on those alignments (Jumper et
672 al., 2021; Senior et al., 2020; Tunyasuvunakool et al., 2021). Additionally, structural
673 templates can be added to the process, but they are not mandatory. The first step depends
674 on searching large sequence databases. The optimization of this step requires significant

675 memory resources which are limiting in many cases. Colabfold provides a public MMSeqs2
676 server (Steinegger and Söding, 2017) that generates multiple-sequence alignments tailored
677 to Alphafold requirements to solve this issue (Mirdita et al., 2022). However, the public
678 nature of this server implies that users can only perform a limited number of queries per unit
679 of time. The second step consists of running forward predictions on a deep neural network.
680 This step is usually performed in high-performance graphical processing units (GPU). To
681 predict nitrogenase structures, we have considered a *bonafide* protocol and a recycling
682 protocol.
683
684 ***Bonafide protocol –*** In this protocol, we rely on the Colabfold methodology, which consists
685 of using MMSeqs2 to generate a deep sequence alignment covering the whole sequence of
686 the query and running forward predictions with 3 recycles using the Alphafold neural
687 network. In this step, we avoided using structural templates and optimizing side chains using
688 the AMBER force field.
689
690 ***Recycling protocol –*** To predict variant structures with a high level of sequence identity to
691 another reference sequence, we implemented a different pipeline based on recycling the
692 MSA generated by MMSeqs2. The main reason for this additional step is to avoid blocking a
693 public resource with thousands of similar requests. We re-aligned the corresponding
694 concatenation of sequences to the MSA generated on the structure prediction of the
695 maximum-likelihood ancestral sequence using MAFFT with its options "--keepLength" and
696 "--addResidue" (Katoh and Standley, 2013). Previous conversions from A3M to FASTA format
697 were carried out using HH-suite convert.pl script (Steinegger et al., 2019). The result of this
698 re-alignment was converted to A3M again, and the files were modified to include the headers
699 specifying the number of chains for each sequence and the length of each sequence in the
700 concatenated dataset. The rest of the protocol proceeded as the *bonafide* protocol, using 3
701 recycles without templates nor side-chain optimization.
702
703

Protein structure analysis pipeline

The final dataset includes a total of 5378 structures corresponding to extant and ancestral structures. We utilized semi-automatic methods to analyze the high number of structures. Our pipeline for the analysis of structures includes the following steps. (1) Aligning every structure against a template using US-Align (Zhang et al., 2022). For the component-1 complexes (chains DDKK), Vnf nitrogenase (PDB: 5N6Y) was used as a template. For the component-2, the NifH dimer was used as a template (PDB: 1G1M), both templates from A. *vinelandii*. (2) Modifying the name of chains to match the reference structure, and (3) Extracting the pLDDT, predicted local distance difference test, which is the standard uncertainty predictor of Alphafold (Jumper et al., 2021) from the B-factor column of PDB files and trimming regions of the structure with low prediction confidence.

The similarity of two protein structures is computed using US-Align (Zhang et al., 2022), which extends the structural alignment capabilities of the TM-align (Zhang and Skolnick, 2005) method to proteins containing multiple chains. US-Align provides two metrics that provide an assessment of the similarity between proteins: first, the Root Mean Square Deviation (RMSD), which measures the mean distance in which a set of atoms deviates from another; and the TM-score, which is a metric bound between 0 (no similarity) and 1 (absolute similarity).

[FORMULA 1: RMSD]

$$RMSD = \sqrt{\frac{1}{N}\sum_{i}^{N}\left(x_i^{ref} - x_i\right)^2}$$

Where N is the number of atoms, x_i are the 3D coordinates of the aligned atoms, and x_i^{ref} are the 3D coordinates of the reference atoms. In this work, we compute this function solely based on the coordinates of α-carbons.

[FORMULA 2: TM-SCORE]

$$TMScore = \max\left(\frac{1}{L^{ref}}\sum_{i}^{L}\frac{1}{1+\left(\frac{d_i}{d_0(L^{ref})}\right)^2}\right)$$

Where L^{ref} is the length of the reference protein, L is the number of aligned residues, d_i is the distance between a par of aligned residues, and $d_0(L^{ref})$ is a normalization term to account for the size of the proteins under alignment. We additionally attempted to correct possible bias connected to the structural alignment size using the RMSD100 metric (Carugo and Pongor 2001), defined as:

[FORMULA 3: RMSD100]

$$RMSD_{100} = RMSD \cdot \frac{1}{1+\ln\sqrt{\frac{L}{100}}}$$

Where L is the length of the structural alignment.

We have combined different programs' outputs to study the structural features among the different nitrogenases. The list of studied features and the respective programs are summarized in **Table S6**. For normal-mode derived properties and geometric properties (e.g., gyration radius), we used ProDy (Bakan et al. 2011). For affinity predictions, we used PRODIGY (Vangone and Bonvin 2015). For surface solvent-accessible areas, we used FreeSASA (Mitternacht 2016). Finally, for residue-interaction network properties, we used RING (Clementel et al. 2022).

Nitrogenase structure repository

We implemented a lightweight server using Streamlit framework in Python providing access and the ability to download to all reconstructed structures using taxon IDs, gene names, or species names available at https://nsdb.bact.wisc.edu.

757 **Visualization**

758

759 To visualize quantitative data, we employed the Seaborn Python library. To visualize protein
760 structures, we used ChimeraX-v1.5 (Goddard et al., 2018).

761

762 **DATA AVAILABILITY**

763 All codes are available at Github: http://github.com/kacarlab/nif-structures

764 All crystal structures will be available at the Protein Database upon publication: Anc1A
765 DDKK: 9HB9, Anc 1B HH: 9HAZ, Anc1B DDKK: 9HBN and Anc2 DDKK: 9HBC.

766 All structures are available at the nitrogenase structure repository at the UW-Madison
767 Department of Bacteriology hosted https://nsdb.bact.wisc.edu and at Zenodo
768 10.5281/zenodo.13351063.

769

770 **AUTHOR CONTRIBUTIONS**

771 **Conceptualization and study design:** BCZ and BK

772 **Data collection**: BCZ, FD, KA, AKG

773 **Writing (first draft)**: BCZ and BK

774 **Writing (reviewing/editing):** All authors

775

776 **ACKNOWLEDGEMENTS**

777 B.C.Z acknowledges the Margarita Salas Postdoctoral Fellowship, founded by the Unión
778 Europea - Next Generation EU (B.C.Z.; UP2021-035). This work was supported by the
779 National Aeronautics and Space Administration (NASA) Interdisciplinary Consortium for
780 Astrobiology Research: Metal Utilization and Selection Across Eons, MUSE
781 [*80NSSC17K0296*] with additional support from the Human Frontier Science Program
782 (HFSP) [*RGY0072/2021*] (KA), W. M. Keck Foundation, the Hypothesis Fund and the NASA
783 Exobiology Program [*NNH23ZDA001N*]. O.E. acknowledges support from the European

Research Council (Horizon Europe, AdG 101141673). We thank the Center for High Throughput Computing (CHTC), Dr. Derek Harris for assistance in nitrogenase protein purification, Dr. Janet Newlands for the IT support, the Department of Bacteriology at the University of Wisconsin-Madison for providing computing resources and the members of the Kaçar laboratory for the valuable feedback.

REFERENCES

Anbar AD, Knoll AH. 2002. Proterozoic Ocean Chemistry and Evolution: A Bioinorganic Bridge? *Science* **297**:1137–1142. doi:10.1126/science.1069651

Bakan A, Meireles LM, Bahar I. 2011. ProDy: Protein Dynamics Inferred from Theory and Experiments. *Bioinformatics* **27**:1575–1577. doi:10.1093/bioinformatics/btr168

Barney BM, Yurth MG, Santos PCD, Dean DR, Seefeldt LC. 2009. A substrate channel in the nitrogenase MoFe protein. *JBIC J Biol Inorg Chem* **14**:1015–1022. doi:10.1007/s00775-009-0544-2

Bergman B, Sandh G, Lin S, Larsson J, Carpenter EJ. 2013. Trichodesmium – a widespread marine cyanobacterium with unusual nitrogen fixation properties. *FEMS Microbiol Rev* **37**:286–302. doi:10.1111/j.1574-6976.2012.00352.x

Blanc E, Roversi P, Vonrhein C, Flensburg C, Lea SM, Bricogne G. 2004. Refinement of severely incomplete structures with maximum likelihood in BUSTER–TNT. *Acta Crystallogr Sect D: Biol Crystallogr* **60**:2210–2221. doi:10.1107/s0907444904016427

Boyd ES, Costas AMG, Hamilton TL, Mus F, Peters JW. 2015. Evolution of Molybdenum Nitrogenase during the Transition from Anaerobic to Aerobic Metabolism. *J Bacteriol* **197**:1690–1699. doi:10.1128/jb.02611-14

Brunet TDP, Doolittle WF. 2018. The generality of Constructive Neutral Evolution. *Biol Philos* **33**:2. doi:10.1007/s10539-018-9614-6

Burén S, Jiménez-Vicente E, Echavarri-Erasun C, Rubio LM. 2020. Biosynthesis of Nitrogenase Cofactors. *Chem Rev* **120**:4921–4968. doi:10.1021/acs.chemrev.9b00489

Campo JSM del, Rigsbee J, Batista MB, Mus F, Rubio LM, Einsle O, Peters JW, Dixon R, Dean DR, Santos PCD. 2022. Overview of physiological, biochemical, and regulatory aspects of nitrogen fixation in Azotobacter vinelandii. *Crit Rev Biochem Mol Biol* **57**:492–538. doi:10.1080/10409238.2023.2181309

Capella-Gutiérrez S, Silla-Martínez JM, Gabaldón T. 2009. trimAl: a tool for automated alignment trimming in large-scale phylogenetic analyses. *Bioinformatics* **25**:1972–1973. doi:10.1093/bioinformatics/btp348

Carugo O, Pongor S. 2001. A normalized root-mean-spuare distance for comparing protein three-dimensional structures. *Protein Sci* **10**:1470–1473. doi:10.1110/ps.690101

Clementel D, Conte AD, Monzon AM, Camagni GF, Minervini G, Piovesan D, Tosatto SCE. 2022. RING 3.0: fast generation of probabilistic residue interaction networks from structural ensembles. *Nucleic Acids Res* **50**:W651–W656. doi:10.1093/nar/gkac365

Cuevas-Zuviría B, Garcia AK, Rivier AJ, Rucker HR, Carruthers BM, Kaçar B. 2024. Emergence of an orphan nitrogenase protein following atmospheric oxygenation. *Mol Biol Evol* msae067. doi:10.1093/molbev/msae067

Dixon R, Kahn D. 2004. Genetic regulation of biological nitrogen fixation. *Nat Rev Microbiol* **2**:621–631. doi:10.1038/nrmicro954

Einsle O, Rees DC. 2020. Structural Enzymology of Nitrogenase Enzymes. *Chem Rev* **120**:4969–5004. doi:10.1021/acs.chemrev.0c00067

Emsley P, Lohkamp B, Scott WG, Cowtan K. 2010. Features and development of Coot. *Acta Crystallogr Sect D: Biol Crystallogr* **66**:486–501. doi:10.1107/s0907444910007493

Falkowski PG. 1997. Evolution of the nitrogen cycle and its influence on the biological sequestration of CO2 in the ocean. *Nature* **387**:272–275. doi:10.1038/387272a0

Garcia AK, Harris DF, Rivier AJ, Carruthers BM, Pinochet-Barros A, Seefeldt LC, Kaçar B. 2023. Nitrogenase resurrection and the evolution of a singular enzymatic mechanism. *eLife* **12**:e85003. doi:10.7554/elife.85003

Garcia AK, Kaçar B. 2019. How to resurrect ancestral proteins as proxies for ancient biogeochemistry. *Free Radic Biol Med* **140**:260–269. doi:10.1016/j.freeradbiomed.2019.03.033

Garcia AK, Kolaczkowski B, Kaçar B. 2022. Reconstruction of Nitrogenase Predecessors Suggests Origin from Maturase-Like Proteins. *Genome Biol Evol* **14**:evac031. doi:10.1093/gbe/evac031

Garcia AK, McShea H, Kolaczkowski B, Kaçar B. 2020. Reconstructing the evolutionary history of nitrogenases: Evidence for ancestral molybdenum-cofactor utilization. *Geobiology* **18**:394–411. doi:10.1111/gbi.12381

Goddard TD, Huang CC, Meng EC, Pettersen EF, Couch GS, Morris JH, Ferrin TE. 2018. UCSF ChimeraX: Meeting modern challenges in visualization and analysis. *Protein Sci* **27**:14–25. doi:10.1002/pro.3235

Harris DF, Rucker HR, Garcia AK, Yang Z-Y, Chang SD, Feinsilber H, Kaçar B, Seefeldt LC. 2024. Ancient nitrogenases are ATP dependent. *mBio* **15**:e01271-24. doi:10.1128/mbio.01271-24

Johnson AC, Ostrander CM, Romaniello SJ, Reinhard CT, Greaney AT, Lyons TW, Anbar AD. 2021. Reconciling evidence of oxidative weathering and atmospheric anoxia on Archean Earth. *Sci Adv* **7**:eabj0108. doi:10.1126/sciadv.abj0108

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, Bridgland A, Meyer C, Kohl SAA, Ballard AJ, Cowie A, Romera-Paredes B, Nikolov S, Jain R, Adler J, Back T, Petersen S, Reiman D, Clancy E, Zielinski M, Steinegger M, Pacholska M, Berghammer T, Bodenstein S, Silver D, Vinyals O, Senior AW, Kavukcuoglu K, Kohli P, Hassabis D. 2021. Highly accurate protein structure prediction with AlphaFold. *Nature* **596**:583–589. doi:10.1038/s41586-021-03819-2

Kaçar B. 2024. Reconstructing Early Microbial Life. *Annu Rev Microbiol*. doi:10.1146/annurev-micro-041522-103400

Kalyaanamoorthy S, Minh BQ, Wong TKF, Haeseler A von, Jermiin LS. 2017. ModelFinder: fast model selection for accurate phylogenetic estimates. *Nat Methods* **14**:587–589. doi:10.1038/nmeth.4285

Katoh K, Standley DM. 2013. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Mol Biol Evol* **30**:772–780. doi:10.1093/molbev/mst010

Lyons TW, Reinhard CT, Planavsky NJ. 2014. The rise of oxygen in Earth's early ocean and atmosphere. *Nature* **506**:307–315. doi:10.1038/nature13068

Lyons TW, Tino CJ, Fournier GP, Anderson RE, Leavitt WD, Konhauser KO, Stüeken EE. 2024. Co-evolution of early Earth environments and microbial life. *Nat Rev Microbiol* **22**:572–586. doi:10.1038/s41579-024-01044-y

Mariani V, Biasini M, Barbato A, Schwede T. 2013. lDDT: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics* **29**:2722–2728. doi:10.1093/bioinformatics/btt473

McCoy AJ, Grosse-Kunstleve RW, Adams PD, Winn MD, Storoni LC, Read RJ. 2007. Phaser crystallographic software. *J Appl Crystallogr* **40**:658–674. doi:10.1107/s0021889807021206

Mirdita M, Schütze K, Moriwaki Y, Heo L, Ovchinnikov S, Steinegger M. 2022. ColabFold: making protein folding accessible to all. *Nat Methods* **19**:679–682. doi:10.1038/s41592-022-01488-1

Mitternacht S. 2016. FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research* **5**:189. doi:10.12688/f1000research.7931.1

Moore EK, Jelen BI, Giovannelli D, Raanan H, Falkowski PG. 2017. Metal availability and the expanding network of microbial metabolisms in the Archaean eon. *Nat Geosci* **10**:629–636. doi:10.1038/ngeo3006

Morrison CN, Hoy JA, Zhang L, Einsle O, Rees DC. 2015. Substrate Pathways in the Nitrogenase MoFe Protein by Experimental Identification of Small Molecule Binding Sites. *Biochemistry* **54**:2052–2060. doi:10.1021/bi501313k

Mrnjavac N, Esposti MD, Mizrahi I, Martin WF, Allen JF. 2024. Three enzymes governed the rise of O2 on Earth. *Biochim Biophys Acta (BBA) - Bioenerg* **1865**:149495. doi:10.1016/j.bbabio.2024.149495

Murshudov GN, Skubák P, Lebedev AA, Pannu NS, Steiner RA, Nicholls RA, Winn MD, Long F, Vagin AA. 2011. REFMAC5 for the refinement of macromolecular crystal structures. *Acta Crystallogr Sect D* **67**:355–367. doi:10.1107/s0907444911001314

Mus F, Colman DR, Peters JW, Boyd ES. 2019. Geobiological feedbacks, oxygen, and the evolution of nitrogenase. *Free Radic Biol Med* **140**:250–259. doi:10.1016/j.freeradbiomed.2019.01.050

Navarro-González R, McKay CP, Mvondo DN. 2001. A possible nitrogen crisis for Archaean life due to reduced nitrogen fixation by lightning. *Nature* **412**:61–64. doi:10.1038/35083537

Nguyen L-T, Schmidt HA, Haeseler A von, Minh BQ. 2015. IQ-TREE: A Fast and Effective Stochastic Algorithm for Estimating Maximum-Likelihood Phylogenies. *Mol Biol Evol* **32**:268–274. doi:10.1093/molbev/msu300

Parsons C, Stüeken EE, Rosen CJ, Mateos K, Anderson RE. 2021. Radiation of nitrogen-metabolizing enzymes across the tree of life tracks environmental transitions in Earth history. *Geobiology* **19**:18–34. doi:10.1111/gbi.12419

Peters JW, Stowell MHB, Soltis SM, Finnegan MG, Johnson MK, Rees DC. 1997. Redox-Dependent Structural Changes in the Nitrogenase P-Cluster † , ‡. *Biochemistry* **36**:1181–1187. doi:10.1021/bi9626665

Raymond J, Siefert JL, Staples CR, Blankenship RE. 2004. The Natural History of Nitrogen Fixation. *Mol Biol Evol* **21**:541–554. doi:10.1093/molbev/msh047

Robson RL, Postgate JR. 1980. Oxygen and Hydrogen in Biological Nitrogen Fixation. *Annu Rev Microbiol* **34**:183–207. doi:10.1146/annurev.mi.34.100180.001151

Rucker HR, Kaçar B. 2024. Enigmatic evolution of microbial nitrogen fixation: insights from Earth's past. *Trends Microbiol* **32**:554–564. doi:10.1016/j.tim.2023.03.011

Rutledge HL, Cook BD, Nguyen HPM, HerzikJr. MA, Tezcan FA. 2022. Structures of the nitrogenase complex prepared under catalytic turnover conditions. *Science* **377**:865–869. doi:10.1126/science.abq7641

Sánchez-Baracaldo P, Ridgwell A, Raven JA. 2014. A Neoproterozoic Transition in the Marine Nitrogen Cycle. *Curr Biol* **24**:652–657. doi:10.1016/j.cub.2014.01.041

Santos PCD. 2018. Metalloproteins, Methods and Protocols. *Methods Mol Biol* **1876**:91–109. doi:10.1007/978-1-4939-8864-8_6

Schlesier J, Rohde M, Gerhardt S, Einsle O. 2016. A Conformational Switch Triggers Nitrogenase Protection from Oxygen Damage by Shethna Protein II (FeSII). *J Am Chem Soc* **138**:239–247. doi:10.1021/jacs.5b10341

Schmid B, Einsle O, Chiu H-J, Willing A, Yoshida M, Howard JB, Rees DC. 2002. Biochemical and Structural Characterization of the Cross-Linked Complex of Nitrogenase: Comparison to the ADP-AlF4 --Stabilized Structure † , ‡. *Biochemistry* **41**:15557–15565. doi:10.1021/bi026642b

Senior AW, Evans R, Jumper J, Kirkpatrick J, Sifre L, Green T, Qin C, Žídek A, Nelson AWR, Bridgland A, Penedones H, Petersen S, Simonyan K, Crossan S, Kohli P, Jones DT, Silver D, Kavukcuoglu K, Hassabis D. 2020. Improved protein structure prediction using potentials from deep learning. *Nature* **577**:706–710. doi:10.1038/s41586-019-1923-7

Sippel D, Einsle O. 2017. The structure of vanadium nitrogenase reveals an unusual bridging ligand. *Nat Chem Biol* **13**:956–960. doi:10.1038/nchembio.2428

Spatzal T, Aksoyoglu M, Zhang L, Andrade SLA, Schleicher E, Weber S, Rees DC, Einsle O. 2011. Evidence for Interstitial Carbon in Nitrogenase FeMo Cofactor. *Science* **334**:940–940. doi:10.1126/science.1214025

Stamatakis A. 2014. RAxML version 8: a tool for phylogenetic analysis and post-analysis of large phylogenies. *Bioinformatics* **30**:1312–1313. doi:10.1093/bioinformatics/btu033

Steinegger M, Meier M, Mirdita M, Vöhringer H, Haunsberger SJ, Söding J. 2019. HH-suite3 for fast remote homology detection and deep protein annotation. *BMC Bioinform* **20**:473. doi:10.1186/s12859-019-3019-7

Steinegger M, Söding J. 2017. MMseqs2 enables sensitive protein sequence searching for the analysis of massive data sets. *Nat Biotechnol* **35**:1026–1028. doi:10.1038/nbt.3988

Stüeken EE, Buick R, Guy BM, Koehler MC. 2015. Isotopic evidence for biological nitrogen fixation by molybdenum-nitrogenase from 3.2 Gyr. *Nature* **520**:666–669. doi:10.1038/nature14180

Takimoto R, Tatemichi Y, Aoki W, Kosaka Y, Minakuchi H, Ueda M, Kuroda K. 2022. A critical role of an oxygen-responsive gene for aerobic nitrogenase activity in Azotobacter vinelandii and its application to Escherichia coli. *Sci Rep-uk* **12**:4182. doi:10.1038/s41598-022-08007-4

Tezcan FA, Kaiser JT, Mustafi D, Walton MY, Howard JB, Rees DC. 2005. Nitrogenase Complexes: Multiple Docking Sites for a Nucleotide Switch Protein. *Science* **309**:1377–1380. doi:10.1126/science.1115653

Trncik C, Detemple F, Einsle O. 2023. Iron-only Fe-nitrogenase underscores common catalytic principles in biological nitrogen fixation. *Nat Catal* 1–10. doi:10.1038/s41929-023-00952-1

Tunyasuvunakool K, Adler J, Wu Z, Green T, Zielinski M, Žídek A, Bridgland A, Cowie A, Meyer C, Laydon A, Velankar S, Kleywegt GJ, Bateman A, Evans R, Pritzel A, Figurnov M, Ronneberger O, Bates R, Kohl SAA, Potapenko A, Ballard AJ, Romera-Paredes B, Nikolov S, Jain R, Clancy E, Reiman D, Petersen S, Senior AW, Kavukcuoglu K, Birney E, Kohli P, Jumper J, Hassabis D. 2021. Highly accurate protein structure prediction for the human proteome. *Nature* **596**:590–596. doi:10.1038/s41586-021-03828-1

Vagin A, Teplyakov A. 2010. Molecular replacement with MOLREP. *Acta Crystallogr Sect D: Biol Crystallogr* **66**:22–25. doi:10.1107/s0907444909042589

Vangone A, Bonvin AM. 2015. Contacts-based prediction of binding affinity in protein–protein complexes. *Elife* **4**:e07454. doi:10.7554/elife.07454

Vicente EJ, Dean DR. 2017. Keeping the nitrogen-fixation dream alive. *Proc Natl Acad Sci* **114**:3009–3011. doi:10.1073/pnas.1701560114

Xu J, Zhang Y. 2010. How significant is a protein structure similarity with TM-score = 0.5? *Bioinformatics* **26**:889–895. doi:10.1093/bioinformatics/btq066

Yang Z. 2007. PAML 4: Phylogenetic Analysis by Maximum Likelihood. *Mol Biol Evol* **24**:1586–1591. doi:10.1093/molbev/msm088

Yang Z, Kumar S, Nei M. 1995. A new method of inference of ancestral nucleotide and amino acid sequences. *Genetics* **141**:1641–1650. doi:10.1093/genetics/141.4.1641

Zhang C, Shine M, Pyle AM, Zhang Y. 2022. US-align: universal structure alignments of proteins, nucleic acids, and macromolecular complexes. *Nat Methods* **19**:1109–1115. doi:10.1038/s41592-022-01585-1

Zhang S, Krieger JM, Zhang Y, Kaya C, Kaynak B, Mikulska-Ruminska K, Doruker P, Li H, Bahar I. 2021. ProDy 2.0: increased scale and scope after 10 years of protein dynamics modelling with Python. *Bioinformatics* **37**:3657–3659. doi:10.1093/bioinformatics/btab187

Zhang Y, Skolnick J. 2005. TM-align: a protein structure alignment algorithm based on the TM-score. *Nucleic Acids Res* **33**:2302–2309. doi:10.1093/nar/gki524

Zhang Y, Skolnick J. 2004. Scoring function for automated assessment of protein structure template quality. *Proteins: Struct, Funct, Bioinform* **57**:702–710. doi:10.1002/prot.20264