

Supplementary Methods and Supplementary Results

Chang Y-H. The processing class of the second residue is conserved across three domains of life, and retention of the initiator methionine is coupled to packing into the protein fold. *Protein Science* (submitted).

This document holds the full procedural description of the analysis (Supplementary Methods) and the extended results text (Supplementary Results) that were condensed in the main text to meet the length guideline for Full-Length Articles. All values are those of the audited analysis of 2 October 2026 and match Supplementary Tables S1 to S27 and Figures 1 to 6. Citations refer to the reference list of the main text.

SUPPLEMENTARY METHODS

SM. Sequence and structure data

This is a computational study; no reagents, cell lines, strains or animals were used. All sequences and structures were obtained from public databases. Protein sequences came from 40 UniProt reference proteomes (release 2026_03): 21 eukaryotes spanning seven lineage groups (vertebrates, invertebrate metazoa, fungi, Archaeplastida, Amoebozoa, SAR and Discoba/Metamonada), 11 bacteria and 8 archaea; the species and proteome identifiers are listed in Supplementary Table S1. For each proteome only canonical, full-length, methionine-initiated entries were kept; for human the Swiss-Prot subset was used (20,213 proteins). Predicted structures were taken from the AlphaFold Protein Structure Database (model version 6, released 1 August 2025). Experimental structures were mapped from UniProt accessions to Protein Data Bank entries through the SIFTS best-structures service of PDBe, and residue-level contacts were read from the RCSB ModelServer. Annotation of initiator-methionine removal and N-terminal acetylation was taken from the UniProt “Modified residue” and “Initiator methionine” features, and is referred to throughout as annotated processing state, distinct from the sequence-based MetAP class defined below.

SM. Definition of MetAP class at the second residue

Every protein was assigned a sequence class from its second residue: MetAP-permissive when the residue is Gly, Ala, Ser, Cys, Thr, Pro or Val, and MetAP-nonpermissive otherwise. The class is a prediction of processing from the specificity of methionine aminopeptidase and is not treated as evidence that the initiator methionine is removed or retained in any given protein. Within eukaryotic nonpermissive proteins, sub-classes were recorded for the acetyltransferase-type N-termini (NatB-type: Met followed by Asp, Glu, Asn or Gln; NatC-type: Met followed by Leu, Ile, Phe, Trp or Tyr) for use in the confound analysis.

SM. Orthologs and N-terminal matching

Anchored orthology was defined by reciprocal best hits from DIAMOND 2.1.10 in sensitive mode, run from three anchors: human against each of the other 39 proteomes, *Escherichia coli* K-12 against the 10 other bacteria, and *Haloferax volcanii* against the 7 other archaea. Because the second residue can be compared only when the two N-termini are collinear, an ortholog pair was retained only if the alignment starts in both sequences fall within the first 50 residues and differ by no more than about 10 percent of the shorter start position; pairs failing this window were not scored. This length-matching step is deliberately strict and discards most reciprocal best hits (for example 524 of 2,237 human–yeast pairs pass). Its effect on the results was examined by

varying the window from strict (both starts within 10 residues, offset at most 2) to relaxed (within 100 residues, 20 percent) and by dropping it altogether, as reported in Results.

SM. Class-conservation test

For each anchored comparison, substitutions at the second residue were counted over all retained ortholog pairs, and the analysis asked how often the substituted residue kept the anchor's MetAP class. The expectation was taken from substitutions at positions 3 to 12 of the same alignments, reweighted so that the mix of starting residues matched the mix at position two; this composition-matched expectation removes the effect of the abundance of small residues at the N-terminus and, because it is built from the same starting residues, carries the same codon-neighbourhood structure. The excess of class retention at position two over the expectation was expressed as a z score from the binomial variance. Because ortholog pairs are not independent, the test was repeated with the target species as the unit (sign test over species, each species contributing one signed difference) and with the anchor protein as the unit (Wilcoxon signed-rank test over proteins, each contributing one averaged difference). The same test was run separately for each starting residue and with Ala2 and Ser2 removed.

SM. Deeply conserved sets and ranking

A protein was called deeply conserved when its MetAP class is kept in at least half of its retained orthologs in at least a minimum number of lineage groups (human: 4 of the 8 groups formed by the six non-vertebrate eukaryotic lineages plus Bacteria and Archaea; *E. coli*: 6 of 10 bacteria; *H. volcanii*: 4 of 7 archaea) and with at least 80 percent class agreement over all retained orthologs; a protein scored in more groups than the minimum was not required to keep the class in every one of them, and the agreement threshold limited how many groups could disagree. Within each class, proteins were ranked by a Poisson-binomial P-value for the number of orthologs sharing the anchor's class, with each species' expected probability set to its own proteome-wide class frequency; vertebrate orthologs were excluded from the human score as uninformative. The score was used for ranking only and not as a gate. Sensitivity of the sets to the lineage-count and agreement thresholds, to the matching window, and to exclusion of proteins with an internal methionine within residues 2 to 10 (a proxy for start-codon misannotation in prokaryotes) is reported in Results and in Supplementary Tables S2 to S5.

SM. Anchor-free orthogroup analysis

To remove the dependence on an anchor species, orthogroups were inferred with OrthoFinder 2.5.5 on 34 of the 40 proteomes (six large, lineage-redundant proteomes were omitted to keep the run tractable; every lineage group remained represented). The all-versus-all searches were run with DIAMOND in sensitive mode (E-value at most 0.001, up to 2,000 hits per query) and clustered by MCL within OrthoFinder. Within each orthogroup the member whose N-terminus is length-matched, under the window above, to the largest number of other species served as the local reference, and one matched representative per species was kept. Class conservation was assessed with a cross-domain criterion (at least 6 species; at least 4 of 9 lineage groups each keeping the class in at least half of their members; agreement at least 0.8) and with domain-level criteria mirroring the anchored rules (eukaryotes: at least 4 supergroups; bacteria: at least 6 species; archaea: at least 5 species; agreement at least 0.8). Concordance with the anchored sets was measured by Fisher's exact test on the proteins present in both analyses.

SM. Structural scoring on predicted models

For every anchor-scored protein with an AlphaFold model, the initiator methionine was scored for packing into the fold. Using gemmi to read coordinates and FreeSASA for accessibility, the analysis counted the residues at least six positions away in sequence with any heavy atom within 4.5 Å of any Met1 heavy atom (long-range contacts) and recorded the predicted local confidence (pLDDT) of residues 1 to 3 and the relative solvent accessibility of Met1. The binary packing call required pLDDT at residue 2 of at least 70 and at least three long-range contacts. Because this call contains a confidence term, and because deeply conserved nonpermissive proteins have more ordered predicted N-termini than other nonpermissive proteins, every packing comparison was repeated in three confidence-controlled forms: within terciles of median pLDDT over residues 1 to 3; restricted to proteins passing the confidence gate in both groups, comparing the proportion with at least three contacts; and as a comparison of contact counts by Mann–Whitney test. The confidence-controlled estimates are the ones reported as effect sizes. For human proteins, interface burial of the N-terminus was taken from the per-protein table of the companion census of human experimental assemblies (Chang 2026a; data archive <https://doi.org/10.5281/zenodo.22966146>), using its descriptive threshold of at least 25 percent of the N-terminal segment's accessible area buried at an interface.

SM. Experimental-structure validation

All anchor-scored proteins (human 605, *E. coli* 204, *H. volcanii* 226) and the anchor-free human additions were mapped to the PDB through SIFTS. An entry qualified if the deposited chain begins at UniProt residue 1 (so that Met1 was present in the construct) or exactly at residue 2 (chain deposited as the mature protein). For each protein the best-resolution X-ray, cryo-EM and NMR entry was analysed together with the best entry deposited from residue 2. For each entry the pipeline recorded any residues preceding Met1 in the construct (tag remnants, from the PDBe molecules service), the first modelled residue (PDBe polymer coverage) and, when Met1 was modelled, the number of residues of the same chain instance at least six positions away and the number of residues of other chain instances with heavy atoms within 4.5 Å of a Met1 heavy atom (RCSB ModelServer residue-interaction query at 4.5 Å, model 1, hydrogen atoms and zero-occupancy atoms excluded, distances recomputed from the returned coordinates). Crystallographic symmetry copies of the same chain were treated as other chain instances and counted as interface contacts, not as intrachain packing; an earlier version of the script had merged symmetry copies into the intrachain count, and all experimental counts reported here are from the corrected recount, in which three entries whose Met1 lacked usable coordinates or was not methionine were excluded. The packing call on experimental coordinates was at least three long-range intrachain contacts, with no confidence term; interface burial was at least three contacts to another chain instance or to nucleic acid. Per-protein values are the maximum over qualifying entries; one-best-entry, tag-free, X-ray, cryo-EM and high-resolution subsets are reported as sensitivity analyses. Analyses were repeated on entries without tag remnants and within X-ray, cryo-EM and high-resolution subsets. Concordance between the AlphaFold call and the experimental call was tabulated per protein. Where a chain deposited from residue 2 lacks Met1 and the construct carried no tag, this was recorded as a processing-state indicator with the caveat that the expression host is often heterologous.

SM. Class-switch test

For every human protein with an AlphaFold model, eukaryotic orthologs aligned from residue 1 in both sequences were split into those keeping the human MetAP class and those switching it. AlphaFold models were retrieved for the ortholog accessions and Met1 packing measured with the definitions above. Human proteins with at least five eukaryotic orthologs were compared by whether any ortholog had switched class (Fisher's exact test on the human packing call; Mann–Whitney on contact counts), switch rates were counted per ortholog according to whether the human Met1 is packed, and, for proteins with models of both a switched and an unswitched ortholog, the paired difference in Met1 contact count was tested by two-sided Wilcoxon signed-rank test. The overlap between switch status and the deeply conserved label was tabulated and the test repeated within the proteins that are not deeply conserved.

SM. Per-protein analysis of conserved nonpermissive proteins

The deeply conserved nonpermissive proteins of human ($n = 96$) and *E. coli* ($n = 40$) were partitioned by the structural evidence available for each: fold-packed on experimental coordinates (the core when the protein also passes the AlphaFold packing and confidence gates); interface-buried but not fold-packed; predicted-only (no experimental coverage of Met1 but a confident packing prediction); and residual (neither fold-packed nor interface-buried). For core proteins, secondary structure at residues 2 to 6 was assigned from AlphaFold backbone dihedrals, Met1 contact residues were listed, and identity at Met1-contacting positions was compared with identity at other buried positions (relative accessibility below 0.2, excluding the first six residues) across the anchored orthologs by Wilcoxon test. For the residual set, the same partition was mirrored on the scored nonpermissive human proteins that are not deeply conserved, and the two residual sets were compared for N-terminal disorder (pLDDT), NatB-type sub-class, annotated N-terminal acetylation and membrane annotation by Fisher's exact test, including a stratum-matched comparison. Conservation of the identity of the second residue was compared with identity at exposed positions and at positions 3 to 12 in the same orthologs for the core, the deeply conserved residual and the non-deeply-conserved residual. A per-position identity profile of residues 1 to 30 across orthologs was computed for individual proteins of interest and for the disordered residual group as a whole.

SM. Statistical analysis

Proportions were compared with Fisher's exact test and reported as odds ratios with two-sided P-values (and, for the primary experimental comparison, a conditional maximum-likelihood 95 percent confidence interval); contact counts and confidence values were compared with the two-sided Mann–Whitney U test; paired comparisons used the two-sided Wilcoxon signed-rank test; all P values in the text, tables and figures are two-sided; the class-conservation excess was tested by a z score with binomial variance, by a sign test over species and by a Wilcoxon signed-rank test over proteins; enrichment of translation-apparatus proteins (UniProt keyword “Ribosomal protein” or a description matching ribosomal protein, elongation factor, initiation factor, tRNA ligase or synthetase, or translation) was tested by Fisher's exact test against all scored proteins of the same class. The unit of analysis is stated for every test; where ortholog pairs share an anchor protein or species, the species-level and protein-level tests are the ones relied on. No correction for multiple testing was applied to the pre-specified primary comparisons; secondary and exploratory comparisons are labelled as such and their P-values are descriptive. Exact n for every

group appears in the text, tables and figure legends, and full outputs are in Supplementary Tables S1 to S27. Analyses were run in Python 3.14 with pandas 3.0 and SciPy 1.18.

SM. Software and code

The complete analysis pipeline, from proteome download to workbook and figures, consists of 36 scripts documented in a step-by-step pipeline file (steps 1 to 14) and is deposited with a permanent DOI on Zenodo (Chang 2026b) under a CC BY 4.0 licence and require only Python 3, DIAMOND, OrthoFinder, gemmi and FreeSASA. Two elements are new to this work: the composition-matched substitution test for class retention at the second residue, and the per-protein Met1 packing and processing-state extraction from SIFTS-mapped PDB entries.

SM. Data sources and software versions

UniProt reference proteomes (<https://www.uniprot.org/proteomes>); AlphaFold Protein Structure Database (<https://alphafold.ebi.ac.uk>); Protein Data Bank in Europe, SIFTS best-structures, molecules and polymer-coverage services (<https://www.ebi.ac.uk/pdbe/api>); RCSB ModelServer (<https://models.rcsb.org>); DIAMOND 2.1.10 (<https://github.com/bbuchfink/diamond>); OrthoFinder 2.5.5 (<https://github.com/davideemms/OrthoFinder>); gemmi 0.7.5 (<https://gemmi.readthedocs.io>); FreeSASA 2.2.1 (<https://freesasa.github.io>). Versions and database releases are those used at the time of analysis (September 2026).

SUPPLEMENTARY RESULTS (extended text)

SR1. Class conservation at position two: all six anchored comparisons and the species- and protein-level tests

The first question was whether, when the second residue changes between orthologs, the new residue tends to stay in the same MetAP class. For ortholog pairs whose N-termini are length-matched from residue 1, substitutions at position two were counted and the fraction that kept the anchor's class was compared with the same fraction at positions 3 to 12, reweighted to the same mix of starting residues (Materials and Methods). In every one of the six anchored comparisons the class was kept far more often than expected (Figure 1A; Table I). Between human and other eukaryotes, 86.2 percent of 1,698 substitutions at position two kept the class against an expectation of 57.7 percent ($z = 24.3$); between human and vertebrates, 71.4 against 65.4 percent (4,886 substitutions, $z = 9.1$); between human and bacteria, 77.1 against 54.2 percent (70 substitutions, $z = 4.0$); between human and archaea, 80.7 against 51.4 percent (83 substitutions, $z = 5.5$); between *E. coli* and other bacteria, 74.1 against 59.9 percent (1,010 substitutions, $z = 9.4$); and between *H. volcanii* and other archaea, 76.0 against 57.9 percent (621 substitutions, $z = 9.5$).

The prokaryotic comparisons separate class from identity. In bacteria the second residue is no more often identical between orthologs than its neighbours (0.48 against 0.51), and in archaea it is less so (0.51 against 0.57), yet in both it keeps its class far more often than they do (Figure 1B). What is conserved at this position is therefore the processing class rather than a particular side chain. Because ortholog pairs share anchor proteins and species share history, the test was repeated with species and with proteins as the unit. Every target species individually showed a positive excess: 16 of 16 non-vertebrate eukaryotes, 4 of 4 vertebrates, 8 of 8 bacteria and 8 of 8 archaea against human, 10 of 10 bacteria against *E. coli* and 7 of 7 archaea against *H. volcanii* (Figure 1C; sign-test P from 0.016 to 3×10^{-5} where five or more species were available). With the anchor protein as the unit, the Wilcoxon signed-rank P was 6×10^{-39} for human against other

eukaryotes, 1×10^{-6} for human against vertebrates, 4×10^{-3} for human against archaea, 8×10^{-4} for *E. coli* and 2×10^{-12} for *H. volcanii*; only the human–bacteria comparison, with 45 proteins, fell short ($P = 0.10$). These tests establish conservation above a compositional expectation; selection is the inference drawn from them, not a quantity measured here.

SR2. Per-residue substitution tests and the Ala2/Ser2-removed test

The earlier proteogenomic survey (Bonissone et al. 2013) found that the pooled permissive class lost its excess conservation once Ala and Ser were removed, and read conservation at this position as conservation of Ala and Ser. To test whether the class signal is carried by one or two residues, the substitution test was repeated for each starting residue separately, keeping the composition-matched control, which by construction also carries the same codon-neighbourhood structure. Class retention exceeded the control for Ala, Gly, Pro, Ser, Thr and Val individually between human and other eukaryotes (z 8.0 to 12.3, each $P < 10^{-16}$), for Ala, Gly, Pro and Ser in *E. coli* against bacteria (z 2.6 to 8.3) and for Ala, Gly, Pro, Ser and Val in *H. volcanii* against archaea (z 2.9 to 4.9) (Supplementary Table S6). With Ala2 and Ser2 removed as starting residues, permissive starts still kept their class well above control: 0.917 against 0.476 between human and other eukaryotes (412 substitutions, $z = 18.3$), 0.703 against 0.588 against vertebrates (1,072, $z = 7.9$), 0.657 against 0.479 in *E. coli* (134, $z = 4.2$) and 0.648 against 0.430 in *H. volcanii* (142, $z = 5.3$) (Figure 1D). One qualification limits how far this revises the earlier reading: the test removes Ala and Ser only as starting residues, and among permissive substitutions starting from other residues that kept their class, 73 percent moved to Ala or Ser, against 66 percent in the matched control, so Ala and Ser are the preferred destinations as well as the most common starting residues. What the per-residue test establishes is that class membership is conserved for permissive residues other than Ala and Ser, which a positional comparison without composition matching could not have detected.

SR3. Deeply conserved sets: ranking score, top entries, three-domain proteins and robustness

The next step defined, for each anchor, the set of proteins whose class is kept in at least half of their orthologs in each of a minimum number of lineage groups and with at least 80 percent class agreement over all retained orthologs (Materials and Methods). Of 605 scored human proteins, 359 met the criterion (263 MetAP-permissive, 96 MetAP-nonpermissive, from a scored mix of 411 and 194); of 204 *E. coli* proteins, 85 (45 and 40, from 103 and 101); and of 226 *H. volcanii* proteins, 99 (46 and 53, from 122 and 104) (Table I; Supplementary Tables S2 to S4). The archaeal set is balanced between classes; an earlier version of the analysis that gated on the ranking score had produced a 43-to-7 permissive excess, which proved to be an artefact of the score's dependence on the class background (the score of a protein conserved in all seven other archaea cannot exceed 1.5 for the majority nonpermissive class but reaches 3.0 for the minority permissive class), and the score is therefore used for ranking only.

The top human entries by score are RPS15A, GFPT2, RPL3, EEF2, RPL8, RPL15, RPS17, FARSB, RPL39, RPL21, RPS13, ASNS, EEF1A1, RPL17, RPS19 and FEN1 on the permissive side and POLR2L, DTD1, GMPPB, NOP10, POLR3H, NIP7, PELO, RPS6, PWP2, SMC2, IMP4 and TRMT112 on the nonpermissive side (Figure 2A). Fourteen human proteins keep their class in bacteria and archaea as well as across eukaryotes: RPS15A, GFPT2, ASNS, RPS9, GMPPB, SMC2, ENO2, ASS1, ACO1, OPLAH, DTYMK, ATIC, GFPT1 and PGAM2.

The sets are robust to the choices made in defining them (Supplementary Table S5). Requiring 3 or 5 lineage groups instead of 4, or 90 or 100 percent agreement instead of 80, changed the human set by 7 to 28 percent and left every downstream test significant. Widening or narrowing the N-terminal matching window changed the set by at most 9 percent (Jaccard 0.91 to 0.95), whereas dropping N-terminal matching altogether doubled the scored set and weakened the structural test described below, because non-collinear N-termini entered the comparison. Excluding every protein with an internal methionine within residues 2 to 10, the positions where a misannotated start codon would be most plausible, removed about 20 percent of the sets and left the human tests significant. UniProt annotates initiator-methionine removal for 134 of the deeply conserved human permissive proteins and 41 of the *E. coli* ones, and for none of the deeply conserved nonpermissive proteins; these annotations are observed states and are reported here alongside, not as validation of, the sequence class.

SR4. Anchor-free orthogroup analysis

To remove the dependence on an anchor species, the analysis was repeated on orthogroups inferred de novo across 34 proteomes (Materials and Methods). OrthoFinder assigned 81 percent of 263,844 proteins to 24,518 orthogroups, of which 4,845 had at least three N-terminally matched species. Under a cross-domain criterion, 435 of 1,254 scored orthogroups were deeply conserved (336 permissive, 99 nonpermissive); under domain-level criteria, 491 of 1,074 in eukaryotes (376 and 115), 194 of 456 in bacteria (89 and 105) and 177 of 327 in archaea (63 and 114) (Supplementary Tables S7 to S10). The anchor-free sets agreed closely with the anchored ones: of the anchored human deeply conserved proteins that fell in a scored orthogroup, 236 of 277 (85 percent) were also deeply conserved at the orthogroup level (OR 31, $P = 1 \times 10^{-53}$), as were 74 of 76 for *E. coli* (OR 167) and 66 of 69 for *H. volcanii* (OR 93) (Figure 2B). The orthogroup analysis also added 168 deeply conserved orthogroups, containing 187 human proteins, that the reciprocal-best-hit design could not score, mostly paralog-rich families, listed by human representative in Supplementary Table S11; among them are RRM1, CTPS1, CALR, P4HB, GDI2, RPN1, SEC11A, PRIM2, GINS1 and NMD3 on the nonpermissive side and HSPA8, SRP54, METAP1, POLR3A, AHCY, GLUL, APRT, RPS2, RPS5 and RPS25 on the permissive side; these are listed as candidates (Supplementary Table S11). Thirty-nine orthogroups were deeply conserved across all three domains, led by EF-Tu/eEF1A (permissive, 32 species, agreement 1.0), RPS15A/rpsH/rps8, glutamine-fructose-6-phosphate amidotransferase (Cys at position two in 29 of 29 species), Hsp70/DnaK and asparagine synthetase B (Cys, 22 of 22), with GDP-mannose pyrophosphorylase (nonpermissive, 31 species), riboflavin synthase (Phe at position two in all 18 species), UDP-glucose dehydrogenase and exonuclease III/APEX2 among the nonpermissive entries (Figure 2C).

SR5. Correlates of permissive conservation: functional groups

The two classes of deeply conserved proteins differ in what distinguishes them from other proteins of the same class (Figure 3). On the permissive side, the conserved sets are dominated by the translation apparatus: ribosomal proteins, EF-Tu/eEF1A, EF-G/eEF2, EF-Ts, IF1/eIF1A and aminoacyl-tRNA synthetases account for 87 human, 30 *E. coli* and 30 *H. volcanii* proteins. Among permissive proteins, translation-apparatus membership was 34 percent in the conserved human set against 15 percent otherwise (OR 2.9, $P = 3 \times 10^{-5}$), 67 against 21 percent in *E. coli* (OR 7.7, $P = 3 \times 10^{-6}$) and 65 against 14 percent in *H. volcanii* (OR 11, $P = 1 \times 10^{-8}$) (Figure 3A). The enrichment replicated in the anchor-free sets (23 against 8 percent in eukaryotes, OR 3.6, $P = 1 \times 10^{-7}$; 39 against 11 percent in bacteria, OR 5.1, $P = 2 \times 10^{-5}$; 47 against 7 percent in

archaea, OR 13, $P = 3 \times 10^{-8}$) and was absent among nonpermissive proteins. Within the translation apparatus the residue itself varies (63 of 87 human members keep the class but not the residue), so what is held is removability rather than a specific residue. Two smaller permissive groups are locked at the residue level for known chemical reasons: the glutamine amidotransferases GFPT1, GFPT2, ASNS and the *E. coli* and *H. volcanii* glmS, all beginning Met-Cys-Gly, whose Cys becomes the N-terminal nucleophile after methionine removal, and the N-myristoylated Gly2 proteins ARF1, ARL1, ARL2, ARL3, PIK3R4, PSMC1, RPL15 and RPS8. Chaperones and folding factors form a fourth recognisable group (CCT6A, CCT8, PFDN6, HSPE1, DNAJA1, GroEL, DnaK and others), and 151 conserved human permissive proteins remain in an undivided remainder that includes proteasome α subunits, calmodulins, histones H3 and H4 and glycolytic enzymes (Supplementary Table S2). Conserved permissive proteins did not differ from other permissive proteins in intrachain Met1 packing (13 against 14 percent), were slightly more often engaged at interfaces in the census (19 against 12 percent, OR 1.9, $P = 0.04$) and were less often annotated as N-terminally acetylated (32 against 43 percent, OR 0.62, $P = 0.025$).

SR6. Correlates of nonpermissive conservation on AlphaFold models, confidence control and recurring architectures

On the nonpermissive side, the conserved sets are not enriched for any functional category; instead they differ structurally. In AlphaFold models, the initiator methionine of deeply conserved nonpermissive proteins was packed into the fold (pLDDT at residue 2 of at least 70 and at least three long-range contacts) in 57 percent of 96 human proteins against 30 percent of 98 other nonpermissive proteins (OR 3.2, $P = 1.5 \times 10^{-4}$), in 80 percent of 40 against 51 percent of 61 in *E. coli* (OR 3.9, $P = 3.4 \times 10^{-3}$) and in 87 percent of 53 against 57 percent of 51 in *H. volcanii* (OR 5.0, $P = 9 \times 10^{-4}$) (Supplementary Table S12). These binary odds ratios overstate the effect, because deeply conserved nonpermissive proteins also have more confidently predicted N-termini (median pLDDT over residues 1 to 3 of 87.4 against 71.1 in human, $P = 2 \times 10^{-6}$; 93.7 against 88.6 in *E. coli*; 94.5 against 86.6 in *H. volcanii*), and within terciles of local confidence the binary difference largely disappeared. Holding confidence fixed by restricting both groups to proteins passing the gate, the enrichment remained but was smaller: 71 against 55 percent with at least three contacts in human (OR 2.1, $P = 0.063$), 80 against 66 percent in *E. coli* (OR 2.1, $P = 0.16$) and 92 against 73 percent in *H. volcanii* (OR 4.4, $P = 0.021$), with median contact counts of 5 against 3 ($P = 0.006$), 6 against 4 ($P = 0.019$) and 5.0 against 4.5 ($P = 0.053$) (Figure 3B; Supplementary Table S13). Of these six confidence-controlled comparisons, the contact-count comparisons in human and *E. coli* and the binary comparison in *H. volcanii* reach P below 0.05 individually; the direction is the same in all six. The confidence-controlled values are the effect sizes reported here. Two observations argue that the residual effect is not merely a difference in predicted order: the confidence difference is specific to the nonpermissive class (deeply conserved and other permissive proteins have indistinguishable N-terminal order and identical median contact counts), and a more ordered N-terminus is itself expected if Met1 is part of the fold. The anchor-free orthogroup sets reproduced the packing association with the same direction and size (human 54 against 34 percent, OR 2.3, $P = 0.013$; *E. coli* 70 against 50 percent, OR 2.4, $P = 0.056$; *H. volcanii* 89 against 54 percent, OR 6.8, $P = 1.6 \times 10^{-3}$) (Figure 3C), and among permissive proteins there was no packing effect in either design (OR 0.4 to 0.9).

The packed nonpermissive proteins include recurring architectures: the zinc-ribbon RNA polymerase subunits POLR2L and rpoN; the Rossmann or P-loop β 1 strands of GMPPB, DTD1,

adk, purD, pyrH and dnaN; the SMC ATPase N-lobes of SMC2 and SMC3; PCNA; the tubulin autoregulation motif (TUBB4B Met-Arg-Glu-Ile, TUBA1A Met-Arg-Glu-Cys); and the ubiquitin Met1 of UBA52 and RPS27A (Met-Gln-Ile-Phe-Val). In all three anchors most of these proteins keep the class but not the residue (34 of 55 in human, 29 of 32 in *E. coli*, 42 of 46 in *H. volcanii*), so the methionine is retained while the residue behind it varies. A further 12 conserved human nonpermissive proteins have Met1 buried at a subunit interface in the companion census of human experimental assemblies (Chang 2026a) rather than in their own fold (PSMA1, PSMA5, UBE2M, ARPC2, SPCS3, CRCP, ERCC8, ANAPC11, POLR2J, POLR3K, SEH1L, XPR1); counting intrachain packing and interface burial together, 72 of 96 (75 percent) conserved human nonpermissive proteins have a buried Met1 against 46 percent of the rest (OR 3.5, $P = 4 \times 10^{-5}$).

SR7. Experimental-structure validation: sensitivity analyses, strongest cases and processing-state indicators

Because the packing scores above come from monomer models with a confidence term, the same prediction was tested on deposited structures. All 1,035 anchor-scored proteins and 162 anchor-free human additions were mapped to the PDB, and entries were retained when the deposited chain began at UniProt residue 1 or exactly at residue 2 (Materials and Methods). Structures were found for 746 proteins: 475 human and 165 *E. coli* anchor-scored proteins, 105 anchor-free human additions and one *H. volcanii* protein, so the archaeal test could not be repeated (Table II; Supplementary Tables S14 to S18).

Among nonpermissive proteins whose Met1 is modelled, the deeply conserved ones were more often fold-packed, with the same contact definition and no confidence term: 70 percent (43 of 61) against 47 percent (16 of 34) in human (OR 2.7, 95 percent CI 1.0 to 7.0, $P = 0.029$) and 79 percent (22 of 28) against 65 percent (22 of 34) in *E. coli* (OR 2.0, $P = 0.27$, not significant on its own; 96 against 74 percent when interface burial is counted, OR 9.7, $P = 0.017$) (Figure 3D). The median number of long-range contacts of Met1 was 5 against 2 in human ($P = 0.038$) and 6 against 3.5 in *E. coli* ($P = 0.039$) (Figure 3E). Among permissive proteins with a modelled Met1 the difference was smaller and not significant (37 against 25 percent in human, $P = 0.34$). The effect held in entries without any tag remnant (human OR 3.0, $P = 0.044$; *E. coli* OR 2.4, $P = 0.22$), was weaker when one best-resolution entry per protein was scored instead of the maximum over entries (human OR 2.1, $P = 0.087$), and pointed the same way in X-ray-only, cryo-EM-only and high-resolution subsets, none of which was individually significant. The AlphaFold call and the experimental call agreed for 92 percent of human proteins (56 both packed, 31 both unpacked, 8 discordant of 95) and 85 percent of *E. coli* proteins (37 both packed, 16 both unpacked, 9 discordant of 62) (Figure 3F). The strongest experimental cases were DTD1 and *E. coli* dtd (12 long-range contacts in both organisms), dnaN and RAD9A (11), SMC3 (10) and, with 9 each, the ubiquitin Met1 of RPS27A, GMPPB, pheT, murF, ERCC2, RPS6, MAF1, AP2M1 and TRMT112.

The same entries carry indicators of processing state, with the class comparison as the control (Table II). Met1 was modelled in at least one entry for 64 percent of human and 88 percent of *E. coli* nonpermissive proteins but for only 26 and 16 percent of permissive ones (OR 5.0 and 36.5; $P = 6 \times 10^{-15}$ and 4×10^{-21}). A chain deposited from residue 2 existed for 34 percent of human and 67 percent of *E. coli* permissive proteins against 13 and 12 percent of nonpermissive ones (OR 3.4 and 14.7). In constructs with a native N-terminus, Met1 was absent while residue 2 was modelled for 40 percent of human and 57 percent of *E. coli* permissive proteins against 7 percent

of nonpermissive proteins in both organisms (OR 8.2 and 18.3; $P = 9 \times 10^{-12}$ and 6×10^{-10}). None of these indicators is direct evidence of MetAP action: an unmodelled Met1 can reflect terminal disorder or processing by a heterologous expression host, which follows the same rule. Terminal disorder, however, should affect both classes alike and did not, so the indicators are consistent with the class prediction at the level of deposited coordinates while remaining indirect. Within the permissive class, deeply conserved proteins were more often deposited from residue 2 than the rest (human 38 against 26 percent, $P = 0.036$; E. coli 81 against 55 percent, $P = 0.008$). Among the 14 three-domain human proteins, the structures of ASNS, GFPT1, ENO2, ACO1, RPS9 and RPS15A begin at residue 2, whereas the two nonpermissive members with structures, GMPPB and SMC2, have Met1 modelled and fold-packed (9 and 5 long-range contacts). Many ribosomal proteins share the same entries, so those observations are not independent, and the 19 human nonpermissive proteins with a chain deposited from residue 2 include known special cases such as β - and γ -actin, whose acetylated Met1 is removed by a dedicated pathway; they are listed rather than excluded (Supplementary Table S16).

SR8. Class-switch test: ortholog counts, overlap with the deeply conserved label and the paired comparisons

The association tests compare proteins. A test within proteins asks what happens when one lineage changes class. For the 82 human proteins (41 nonpermissive, 41 permissive) with at least one eukaryotic ortholog that switched class, the 126 switched ortholog pairs were compared with the 437 unswitched pairs, and Met1 packing was measured on AlphaFold models of 487 of the 555 ortholog accessions (Supplementary Tables S19 to S21). Among the 160 scored human nonpermissive proteins with at least five eukaryotic orthologs, 28 had at least one ortholog that switched to permissive; their human Met1 was fold-packed in 14 percent of cases against 55 percent for the 132 proteins with no switch (OR 0.13, $P = 6 \times 10^{-5}$; median long-range contacts 0 against 3, $P = 2 \times 10^{-4}$) (Figure 4A). Counted per ortholog, the switch rate was 0.7 percent (7 of 936) when the human Met1 is packed and 7.0 percent (63 of 902) when it is not (Figure 4B). Permissive proteins showed no such relationship (22 against 16 percent packed, OR 1.5, $P = 0.42$), as expected if the constraint on that side is processing fate rather than the Met1 environment.

This result is not independent of the association in the preceding section: switch status and the deeply conserved label overlap almost completely (92 of the 160 proteins are deeply conserved with no switch, 40 are neither, 25 switched and are not deeply conserved, and 3 are deeply conserved with a switch). Its added content lies within the 65 proteins that are not deeply conserved, where the association still held (OR 0.23, $P = 0.030$); within the deeply conserved set there were too few switches to test ($P = 0.07$). In a paired comparison across 39 human nonpermissive proteins with models of both a switched and an unswitched ortholog, the mean Met1 contact count was 1.02 in switched against 1.32 in unswitched orthologs (median paired difference 0, Wilcoxon $P = 0.33$), an expected null because 32 of the 39 have an unpacked human Met1 and so nothing to lose. In the seven proteins whose human Met1 has at least three long-range contacts (SEC31A, RPL18A, ATP6V1A, APEX2, ASRGL1, UBE2J1, PRELID3B), the switched ortholog had fewer Met1 contacts than the unswitched orthologs in all seven (means 3.5 against 5.9; $P = 0.016$) (Figure 4C), most clearly in RPL18A (a Thr2 ortholog with 2 contacts against 8) and ATP6V1A (a Ser2 ortholog with 0 against 5), with SEC31A the exception in which the switched orthologs keep Met1 packed. None of the seven is in the deeply conserved set, and the switched orthologs are mostly from distant lineages, so lineage and switch status are

confounded; seven cases are consistent with the prediction but are too few for a claim. Orthologs that gained a nonpermissive residue showed no structural signature (Figure 4D).

SR9. The core and the residual: partition, residual groups, control comparison and identity baselines

The final question, asked protein by protein, was what accounts for conservation of the nonpermissive class. The 96 deeply conserved human nonpermissive proteins were partitioned by the structural evidence available for each (Figure 5A). Forty-three are fold-packed on experimental coordinates, of which 41 also pass the AlphaFold packing and confidence gates and form the core; 18 are interface-buried but not fold-packed; 11 have no structure covering Met1 but a confident packing prediction; and 24 are neither fold-packed nor interface-buried and form the residual. In *E. coli* the same partition of 40 proteins gave 22 fold-packed (21 core), 5 interface, 9 predicted-only and 4 residual (serS, clpB, fluC, rsfS), so the residual is largely a eukaryotic set in these data (Supplementary Tables S22 and S23).

The core is a β 1-strand class. Assigning secondary structure to residues 2 to 6 from the model dihedrals, Met1 begins a β -strand in 37 of the 41 core proteins, starts a helix in one (TRMT112) and lies in coil or mixed structure in three (EIF2D, MCTS1, CFAP20). Its side chain contacts a median of seven residues at least six positions away, about half of them hydrophobic (Figure 5B). The recurring architectures are cross-domain pairs with the same fold and the same conserved class (DTD1 with *E. coli* dtd; PCNA with dnaN), the ubiquitin β -grasp (RPS27A, UBA52, UBL5), the longin folds of adaptor subunits (AP2M1, AP2S1, AP3M1, AP3S1), RNA polymerase small subunits (POLR2G, POLR2L, POLR3H), the SMC ATPase N-lobe (SMC2, SMC3), the tubulin N-terminus and a large group of enzymes and multi-domain proteins in which Met1 opens the first strand of the fold (Figure 5C). Whether a chain begins with a strand is a property of the fold rather than of function, which is why the core cuts across functional categories. The Met1 pocket is not specially co-conserved: across the orthologs of the 41 core proteins, identity at Met1-contacting positions averaged 0.71 against 0.68 at other buried positions (Wilcoxon $P = 0.25$), and hydrophobic-class conservation 0.92 against 0.91 ($P = 0.28$), with 23 proteins higher and 18 lower at the contact positions (Figure 5D). The pocket is an ordinary part of the hydrophobic core.

The 24 residual human proteins separate into five groups (Figure 6A; Supplementary Table S24). Three are predicted packed but experimentally unresolved or just below the confidence gate (SRPRA, GDI1, BBS2); three are near a threshold (EIF3I and SHQ1 with two long-range contacts, SEPSECS with weak interface burial); one is uninformative (SOD2, whose Met1 lies in the cleaved mitochondrial presequence); three are individual leads with a non-structural explanation in hand or in prospect (ARFRP1, RAD21, EFTUD2, below); and 14 are unexplained (PWP2, CHMP5, KDELR1, KDELR2, UTP23, SEC61G, PSMD14, RBBP5, SNF8, RABGGTA, DNAJB12, UFD1, TNPO3 and LSM2, the last being the one case in which Met1 is resolved and solvent-exposed). The unexplained proteins, taken together with EFTUD2 (15 proteins), share three features that the core lacks: predicted disordered N-termini (median pLDDT at residue 2 of 64 against 93, Mann–Whitney $P = 3 \times 10^{-6}$), enrichment of NatB-type residues Asp, Glu, Asn and Gln at position two (9 of 15 against 6 of 41, OR 8.8, $P = 0.0015$) and annotated N-terminal acetylation of Met1 in five. However, when the same partition was applied to the 98 scored human nonpermissive proteins that are not deeply conserved (16 fold-packed, 19 interface, 6 predicted-only, 57 residual), their residual had the same profile: NatB-type residues in 46 percent of the deeply conserved residual against 56 percent of the other residual ($P = 0.47$),

disordered termini in 58 against 63 percent, annotated acetylation in 25 against 40 percent ($P = 0.22$), membrane keywords in 33 against 30 percent, and no difference within a direct stratum match of 35 disordered NatB-type termini (Figure 6B; Supplementary Table S25). Disorder, NatB-type residues and acetylation are therefore properties of unpacked nonpermissive N-termini as a population, and they do not define a class of signal-conserved termini among the deeply conserved proteins. Consistent with this, among nonpermissive proteins with ordered termini, deep conservation was depleted for NatB-type residues (22 against 42 percent, $P = 0.02$), as expected if packing favours hydrophobic and basic residues at position two.

What does distinguish the residual is the strictness with which position two is held. Across all orthologs, the second residue of the deeply conserved residual proteins was conserved as an identity at a median of 0.86, exceeding the identity of exposed residues in the same proteins by 0.23 (Wilcoxon $P = 5 \times 10^{-5}$) and of residues 3 to 12 by 0.16 ($P = 0.025$), with 21 of 24 above their exposed-residue baseline; the fold-packed core showed the same excess (0.23, $P = 2 \times 10^{-8}$; core against residual $P = 0.56$) (Figure 6C; Supplementary Table S26). The comparison with the non-deeply-conserved residual is circular, since deep conservation is defined at this position, so the informative comparison is with the core: the residual proteins hold position two as strictly as proteins whose Met1 is packed, with no packing to account for it. Strictest are RABGGTA (His2 in 10 of 10 orthologs), RAD21 (Phe2, 15 of 15), RBBP5 (Asn2, 13 of 13), LSM2 (Leu2, 14 of 14) and PWP2, KDELR1 and KDELR2 (11 of 12 each). The segment downstream of Met1 was not conserved as a group: per-position identity over residues 2 to 10 was 0.69 in the 14 disordered residual proteins, 0.71 in the 31 disordered non-deeply-conserved residual proteins and 0.67 in the core, with no difference in the excess over baseline ($P = 0.11$ and 0.71) (Supplementary Table S27).

SR10. Individual leads (ARFRP1, RAD21, EFTUD2) and the eEF2 contrast

Two individual proteins do carry a conserved motif in a disordered segment, and a third supplies a positive control, so three proteins are carried forward as individual hypotheses (Figure 6D). ARFRP1 conserves an aromatic, NatC-type residue at position two (Tyr in 5 and Phe in 6 of 11 deep orthologs, identity +0.19 over its exposed baseline); its acetylated Met-Tyr/Met-Phe terminus is recognised by SYS1 for Golgi targeting, as first shown for yeast Arl3p (Behnia et al. 2004; Setty et al. 2004) and consistent with the NatC substrate architecture (Grunwald et al. 2020; Varland et al. 2023), so its conservation is explained by a non-structural, acetylation-dependent signal. RAD21 holds Phe2 in all 15 orthologs (+0.48) at the start of a conserved 30-residue segment, disordered in isolation, that folds onto the SMC3 coiled coil in cohesin (Gligoris et al. 2014; Cheng et al. 2020); two deposited entries (8ROK, 9HVV) cover the region without resolving Met1, so whether Met1 itself contacts SMC3 is open. EFTUD2 holds Asp2 in 8 of 11 deep orthologs (+0.11) within a conserved N-terminal motif, Met-Asp-x-x-Leu-Tyr-Asp-Glu-Phe-Gly-Asn-Tyr-Ile-Gly-Pro-(Glu-Leu-Asp), in which 11 of residues 2 to 30 are identical in at least 80 percent of 14 orthologs and three further positions keep their residue class, at the start of a 115-residue acidic N-terminal domain that loses conservation almost completely thereafter (identity 0.85 for residues 2 to 10 against 0.25 for residues 21 to 30 and 0.63 for exposed residues); the domain is not homologous to EF-2, was not placed in the tri-snRNP density (Agafonov et al. 2016), and one entry (9S2E) covers it without resolving Met1. Met1 acetylation is annotated. No function has been described for this motif, and a conserved short motif inside an otherwise unconserved disordered domain is the signature of a partner-binding element.

A contrast fixes what these three are not. Eukaryotic elongation factor 2 is deeply conserved in the permissive class (Val2 in 15, Gly in 6 and Pro in 2 of 23 deep species) and holds 19 of residues 2 to 30 at 80 percent identity or more across 20 orthologs, with mean identity 0.98 over residues 21 to 30 (Figure 6E). But residues 3 onward are confidently ordered (pLDDT 70 to 97), residue 4 and most positions from 17 on are buried, and residues 26 to 34 form the P-loop of the G domain; eEF1A shows the same pattern. In these proteins the conserved N-terminal segment is the first strand of a nucleotide-binding fold, and the downstream conservation is the fold. In EFTUD2 the conserved residues 1 to 18 lie in a predicted disordered segment with no buried position, and conservation collapses immediately after them. The two patterns look alike in an alignment and mean different things: an ordered, buried conserved segment is explained by the structure; a disordered, exposed one held as strictly is not, and is the pattern that points to a role acted out in trans. The eEF2 observation also raises a question these data do not answer, namely why a protein whose first strand is buried from residue 4 keeps a permissive rather than a nonpermissive residue at position two so consistently across eukaryotes; it is recorded as a reproducible observation that marks where the next question lies.

SR11. Observations kept without explanation

Three observations are recorded because they are reproducible, not because they are understood. The first is the strict P2 identity of the residual set just described. The second is the EFTUD2 motif, Md..LYDEFGNYIGPeld in consensus notation (upper case, identity of at least 80 percent in 14 orthologs; lower case, class kept), within a domain whose other residues are not conserved. The third is eukaryotic elongation factor 2, which is deeply conserved in the permissive class yet holds 19 of residues 2 to 30 at 80 percent identity or more across 20 orthologs, with mean identity 0.98 over residues 21 to 30 and the P-loop at 26 to 34; here the conservation of the N-terminal segment clearly serves the GTPase fold and the second residue is simply part of it, so the permissive class at P2 is not the object of selection in the way it appears to be in the residual set. The contrast is useful precisely because it shows what the residual proteins are not. Interesting observations of this kind can direct attention to a position even before their mechanism is known, and they are presented in that spirit.