

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

Yie-Hwa Chang

yiehwa.chang@health.slu.edu

Saint Louis University School of Medicine <https://orcid.org/0000-0002-1583-0627>

Research Article

Keywords: methionine aminopeptidase, initiator methionine, N-terminal processing, protein evolution, orthologs, AlphaFold, Protein Data Bank, protein fold

Posted Date: October 7th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-11272653/v1>

License:  This work is licensed under a Creative Commons Attribution 4.0 International License.

[Read Full License](#)

Additional Declarations: The authors declare no competing interests.

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

Yie-Hwa Chang¹

¹ Edward A. Doisy Department of Biochemistry and Molecular Biology, Saint Louis University School of Medicine, 1100 South Grand Boulevard, St. Louis, MO 63104, USA

Corresponding author: Yie-Hwa Chang, Edward A. Doisy Department of Biochemistry and Molecular Biology, Saint Louis University School of Medicine, 1100 South Grand Boulevard, St. Louis, MO 63104, USA. Email: yiehwa.chang@health.slu.edu. ORCID 0000-0002-1583-0627.

Running title: Conservation of the MetAP class at residue two

Supplementary material (posted with this preprint): Supplementary_Methods_and_Results.pdf; Supplementary_Tables_S1-S27_workbook.xlsx; Supplementary_Figure_S1_design_schematic.pdf; code archive at <https://doi.org/10.5281/zenodo.23112899>.

This manuscript is a preprint and has not been peer reviewed.

ABSTRACT

Removal of the initiator methionine by methionine aminopeptidase is decided by the second residue, which therefore predicts whether a mature protein begins with methionine. This study asks whether this predicted processing fate is itself conserved by evolution, and why. Comparing orthologous proteins across 40 proteomes anchored in human, *Escherichia coli* and *Haloferax volcanii*, and independently across 34 proteomes without an anchor, it finds that when the second residue changes it changes to a residue of the same processing class far more often than a composition-matched expectation, in each of 53 species and in all three domains of life, even where residue identity is not conserved. The two classes are conserved for different reasons. Proteins conserving a removal-type second residue are enriched in the translation apparatus. Proteins conserving a retention-type second residue show increased initiator-methionine packing in predicted models, with experimental-coordinate support strongest in human proteins, and class-switching lineages are associated with less packing. A core of 41 human and 21 *Escherichia coli* proteins has fold-packed initiator methionine supported by predicted and deposited structures; 37 of the 41 human core proteins begin with a beta-strand. The ranked sets and stated predictions provide a framework for testing methionine processing as a conserved trait.

Keywords: methionine aminopeptidase; initiator methionine; N-terminal processing; protein evolution; orthologs; AlphaFold; Protein Data Bank; protein fold

STATEMENT FOR A BROADER AUDIENCE

Most proteins are synthesized with methionine as their first amino acid, and the second amino acid helps determine whether that methionine can be removed. Across bacteria, archaea and eukaryotes, this study finds that the sequence class predicting removal or retention is conserved beyond a composition-matched expectation. Conserved retention is associated with packing of the first methionine into the protein fold, suggesting a structural constraint on N-terminal processing that can be tested experimentally.

ABBREVIATIONS

MetAP, methionine aminopeptidase; Met1, initiator methionine; OR, odds ratio; PDB, Protein Data Bank; PDBe, Protein Data Bank in Europe; SIFTS, Structure Integration with Function, Taxonomy and Sequences; NatB and NatC, N-terminal acetyltransferases B and C; DC, deeply conserved; AF, AlphaFold model; exp, experimental structure; iMet, initiator methionine (in UniProt annotation); CC BY, Creative Commons Attribution licence.

INTRODUCTION

Almost every protein is synthesized with an initiator methionine, and in all three domains of life a methionine aminopeptidase (MetAP) decides whether that methionine survives in the mature chain. MetAP removes the initiator methionine when the second residue has a small side chain (Gly, Ala, Ser, Cys, Thr, Pro or Val) and leaves it in place when the side chain is larger (Sherman et al. 1985; Hirel et al. 1989; Giglione et al. 2004; Xiao et al. 2010); the rule holds across bacteria, archaea and eukaryotes and is routinely used to predict the N-terminus of a protein from its sequence (Wingfield 2017). Two points frame this study. First, the rule is a prediction: a sequence with a small second residue is MetAP-permissive and one with a larger second residue is MetAP-nonpermissive, but whether the methionine is actually removed is a matter of measurement, and exceptions are known (Giglione et al. 2004; Frottin et al. 2006). Second, the rule places the processing fate of a protein at the mercy of a single codon: a point mutation at position two can switch a protein between a methionine-lacking and a methionine-bearing N-terminus without changing the fold, and the mature N-terminus in turn feeds the N-terminal acetyltransferases (Aksnes et al. 2016), the N-degron pathways (Kim et al. 2014; Varshavsky 2024) and other modifications (Varland et al. 2015).

How strongly evolution guards this position is largely unexamined. The one comparative survey (Bonissone et al. 2013), using proteomics data from *Shewanella* strains, *Saccharomyces* species and mammals, found the second position conserved beyond expectation, mainly through Ala and Ser, and no remaining excess for the pooled MetAP-permissive class once Ala and Ser were removed. It did not include archaea, did not compare class retention with a composition-matched expectation for each residue separately, and did not ask whether the fold has any bearing on which second residues are tolerated.

The question asked here is whether the predicted processing fate is itself conserved across orthologs, over and above conservation of the second residue's identity, and, if it is, what constrains it on each side. The two sides need not share a reason: a protein may keep a permissive second residue because the small residue exposed after processing is read by a downstream enzyme, and a nonpermissive one because the retained methionine has a role of its own. If so, the initiator methionine of such proteins should be buried in the fold more often than in other nonpermissive proteins, and lineages that switch class should be those in which the methionine has no such role.

These ideas were tested by building ranked sets of proteins whose MetAP class is deeply conserved, anchored in human, *Escherichia coli* and *Haloferax volcanii* across 40 proteomes, and independently in an anchor-free analysis of 34 proteomes using orthogroups from OrthoFinder (Emms and Kelly 2019) and DIAMOND (Buchfink et al. 2021), and by scoring whether the initiator methionine is packed into the fold on AlphaFold models (Jumper et al. 2021; Varadi et al. 2024) and on deposited coordinates mapped through SIFTS (Dana et al. 2019). The analysis identifies class, not residue identity, as the conserved quantity in every one of 53 species; assigns

different correlates to the two classes; recovers a core of proteins in which a retained initiator methionine begins a buried β -strand; and leaves a residual set whose conserved N-termini point to questions a fold-based explanation does not answer.

RESULTS

The MetAP class of the second residue is conserved above a composition-matched expectation in all three domains

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), from 86.2 against 57.7 percent for 1,698 substitutions between human and other eukaryotes ($z = 24.3$) to 76.0 against 57.9 percent between *H. volcanii* and other archaea ($z = 9.5$), with z of at least 4.0 in every comparison. 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 is the class, not 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. All 53 target species individually showed a positive excess (Figure 1C; sign-test P from 0.016 to 3×10^{-5} where five or more species were available), and with the anchor protein as the unit the Wilcoxon signed-rank P ranged from 6×10^{-39} to 4×10^{-3} , with only the 45-protein human–bacteria comparison falling short ($P = 0.10$) (Supplementary Table S5). These tests establish conservation above a compositional expectation; selection is the inference drawn from them, not a quantity measured here.

Because the earlier survey had read conservation at this position as conservation of Ala and Ser (Bonissone et al. 2013), the test was repeated for each starting residue separately. 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) and for Ala, Gly, Pro and Ser in *E. coli* and Ala, Gly, Pro, Ser and Val in *H. volcanii*, the rarer starts Thr and Val falling short in the prokaryotes (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, $z = 18.3$; 0.657 against 0.479 in *E. coli*; 0.648 against 0.430 in *H. volcanii*) (Figure 1D). One qualification applies: the test removes Ala and Ser only as starting residues, and among class-keeping permissive substitutions from other starts 73 percent moved to Ala or Ser against 66 percent in the control, so Ala and Ser are the preferred destinations as well as the most common starting residues.

Ranked, deeply conserved sets in each domain

For each anchor, a protein was called deeply conserved when its class is kept in at least half of its orthologs in each of a minimum number of lineage groups with at least 80 percent agreement overall (Materials and Methods). Of 605 scored human proteins, 359 met the criterion (263 permissive, 96 nonpermissive); of 204 *E. coli* proteins, 85 (45 and 40); and of 226 *H. volcanii* proteins, 99 (46 and 53) (Table I; Supplementary Tables S2 to S4). The top entries are shown in Figure 2A, and fourteen human proteins keep their class in bacteria and archaea as well as across eukaryotes. The sets are robust to the lineage-count and agreement thresholds (set changes of 7 to 28 percent, every downstream test still significant), to the N-terminal matching window

(Jaccard 0.91 to 0.95) and to exclusion of proteins with an internal methionine within residues 2 to 10, a proxy for start-codon misannotation (Supplementary Table S5). 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 are observed states reported alongside, not as validation of, the sequence class.

To remove the dependence on an anchor species, the analysis was repeated on 4,845 orthogroups inferred de novo across 34 proteomes, of which 435 of 1,254 scored under a cross-domain criterion were deeply conserved (336 permissive, 99 nonpermissive) (Supplementary Tables S7 to S10). Of the anchored human deeply conserved proteins 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), and the orthogroup design added 168 deeply conserved orthogroups, containing 187 human proteins, that the anchored design could not score, mostly paralog-rich families, listed by human representative as candidates (Supplementary Table S11). Thirty-nine orthogroups were deeply conserved across all three domains, led by EF-Tu/eEF1A, RPS15A/rpsH/rps8 and glutamine–fructose-6-phosphate amidotransferase (Cys2 in 29 of 29 species) on the permissive side and GDP-mannose pyrophosphorylase and riboflavin synthase (Phe2 in 18 of 18) on the nonpermissive side (Figure 2C).

Permissive and nonpermissive conservation have different correlates

On the permissive side, the conserved sets are dominated by the translation apparatus (ribosomal proteins, translation factors and aminoacyl-tRNA synthetases; 87 human, 30 *E. coli* and 30 *H. volcanii* proteins): membership was 34 against 15 percent among human permissive proteins (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), replicated in the anchor-free sets (OR 3.6, 5.1 and 13) and 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. Two smaller permissive groups are locked at the residue level for chemical reasons: the glutamine amidotransferases (GFPT1, GFPT2, ASNS, glmS), whose Cys2 becomes the N-terminal nucleophile after methionine removal, and the N-myristoylated Gly2 proteins (ARF1, ARL1 and others) (Supplementary Table S2). Conserved permissive proteins did not differ from other permissive proteins in Met1 packing (13 against 14 percent).

On the nonpermissive side there is no functional enrichment; the conserved sets differ structurally. In AlphaFold models, Met1 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, $P = 1.5 \times 10^{-4}$), 80 against 51 percent in *E. coli* (OR 3.9) and 87 against 57 percent in *H. volcanii* (OR 5.0) (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), 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 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). The direction is the same in all six comparisons, and these 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 (conserved and other permissive proteins have indistinguishable N-terminal order), and a more ordered N-terminus is itself expected if Met1 is part of the fold. The anchor-free sets reproduced the association (human OR 2.3, $P = 0.013$; *E. coli* OR 2.4, $P = 0.056$; *H. volcanii* 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). Most packed 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; counting both, 72 of 96 (75 percent) have a buried Met1 against 46 percent of the rest (OR 3.5, $P = 4 \times 10^{-5}$).

The packing correlate holds on experimental coordinates

Because the packing scores above come from monomer models with a confidence term, the same prediction was tested on deposited structures whose chain begins 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, $P = 0.27$; 96 against 74 percent when interface burial is counted, OR 9.7, $P = 0.017$) (Figure 3D), with median long-range contacts of 5 against 2 ($P = 0.038$) and 6 against 3.5 ($P = 0.039$) (Figure 3E). Permissive proteins showed no significant difference (37 against 25 percent in human, $P = 0.34$). The effect held in entries without a tag remnant (human OR 3.0, $P = 0.044$), was weaker with one best-resolution entry per protein (OR 2.1, $P = 0.087$) and pointed the same way in X-ray, cryo-EM and high-resolution subsets, none individually significant. The AlphaFold and experimental calls 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, 16 and 9 of 62) (Figure 3F). The strongest cases were DTD1 and *E. coli* dtd (12 long-range contacts each), dnaN and RAD9A (11) and SMC3 (10).

The same entries carry indicators of processing state that follow the class prediction (Table II): Met1 was modelled for 64 and 88 percent of human and *E. coli* nonpermissive proteins but for only 26 and 16 percent of permissive ones (OR 5.0 and 36.5). These indicators are indirect, since an unmodelled Met1 can reflect terminal disorder or processing by a heterologous host; disorder, however, should affect both classes alike and did not. Deposition from residue 2 and the within-class comparisons are detailed in Supplementary Results.

Class switches avoid packed termini

A test within proteins asks what happens when one lineage changes class. 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 (OR 1.5, $P = 0.42$), as expected if the constraint on that side is processing fate rather than the Met1 environment (Supplementary Tables S19 to S21). Switch status overlaps almost completely with the deeply conserved label, so this result is not independent of the association above; its added content lies within the 65 proteins that are not deeply conserved, where the association still held (OR 0.23, $P = 0.030$). In the seven proteins with a packed human Met1 and AlphaFold models of both a switched and an unswitched ortholog, the switched ortholog had fewer Met1 contacts in all seven (means 3.5 against 5.9, Wilcoxon $P = 0.016$) (Figure 4C); none of the seven is deeply conserved, but their switched orthologs are mostly from distant lineages, so lineage and switch status are confounded, and seven cases are consistent with the prediction but too few for a claim. Orthologs that gained a nonpermissive residue showed no structural signature (Figure 4D).

A β 1-strand core and an unexplained residual

The 96 deeply conserved human nonpermissive proteins were partitioned by the structural evidence for each (Figure 5A): 43 are fold-packed on experimental coordinates, of which 41 also pass the AlphaFold 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 and form the residual. In *E. coli* the partition of 40 proteins gave 22 fold-packed (21 core), 5 interface, 9 predicted-only and 4 residual, so the residual is largely a eukaryotic set in these data (Supplementary Tables S22 and S23).

The core is a β 1-strand class: Met1 begins a β -strand in 37 of the 41 core proteins, a helix in one (TRMT112) and coil or mixed structure in three (EIF2D, MCTS1, CFAP20), and 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 dtd; PCNA with dnaN), the ubiquitin β -grasp, the longin folds of adaptor subunits, RNA polymerase small subunits, the SMC ATPase N-lobe and a large group of enzymes in which Met1 opens the first strand (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: identity at Met1-contacting positions averaged 0.71 against 0.68 at other buried positions (Wilcoxon $P = 0.25$) (Figure 5D).

The 24 residual proteins separate into five groups (Figure 6A; Supplementary Table S24): three predicted packed but experimentally unresolved or just below the confidence gate (SRPRA, GDI1, BBS2), three near a threshold (EIF3I, SHQ1, SEPSECS), one uninformative (SOD2, whose Met1 lies in the cleaved presequence), three individual leads (ARFRP1, RAD21, EFTUD2) and 14 unexplained, among them PWP2, KDELR1, KDELR2, RBBP5, RABGGTA and LSM2, the one case in which Met1 is resolved and solvent-exposed. The unexplained proteins with EFTUD2 (15) share features the core lacks: disordered N-termini (median pLDDT at residue 2 of 64 against 93, $P = 3 \times 10^{-6}$), NatB-type residues at position two (9 of 15 against 6 of 41, OR 8.8, $P = 0.0015$) and annotated Met1 acetylation in five. However, the residual of the 98 scored human nonpermissive proteins that are not deeply conserved (57 proteins) had the same profile: NatB-type 46 percent in the deeply conserved residual against 56 percent ($P = 0.47$), disordered 58 against 63 percent, acetylated 25 against 40 percent ($P = 0.22$) (Figure 6B; Supplementary Table S25). These features are therefore properties of unpacked nonpermissive N-termini as a population and do not define a class of signal-conserved termini; among 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. Its second residue was conserved as an identity at a median of 0.86 across all orthologs, exceeding the identity of exposed residues in the same proteins by 0.23 (Wilcoxon $P = 5 \times 10^{-5}$; 21 of 24 above baseline), the same excess as in the fold-packed core (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 core is the informative comparison: 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) and RAD21 (Phe2 in 15 of 15), while residues 2 to 10 downstream were no more conserved than in the core (Supplementary Table S27).

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); 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 has a non-structural, acetylation-dependent explanation. RAD21 holds Phe2 in all 15 orthologs at the start of a conserved 30-residue segment that folds onto the SMC3 coiled coil in cohesin (Gligoris et al. 2014; Cheng et al. 2020); deposited entries cover the region without resolving Met1, so whether Met1 itself contacts SMC3 is open. EFTUD2 holds Asp2 in 8 of 11 deep orthologs within a conserved motif, Md..LYDEFGNYIGPeld in consensus notation (11 of residues 2 to 30 identical in at least 80 percent of 14 orthologs), at the start of a 115-residue acidic domain that loses conservation almost completely thereafter; the domain was not placed in the tri-snRNP density (Agafonov et al. 2016), and no function has been described for the motif. Eukaryotic elongation factor 2, deeply conserved in the permissive class, supplies the contrast (Figure 6E): its N-terminal segment is at least as conserved but is ordered and buried as the first strand and P-loop of the G domain, so it is explained by the structure, whereas a disordered, exposed segment held as strictly, as in EFTUD2, is not, and is the pattern that points to a role acted out in trans.

DISCUSSION

The second residue has long been read as the position that decides whether the initiator methionine is removed (Sherman et al. 1985; Hirel et al. 1989; Giglione et al. 2004; Xiao et al. 2010). The present analysis asks whether the decision itself, rather than the residue that encodes it, is what evolution keeps, and finds that across 53 species and all three domains of life the second residue, when it changes, changes to a residue of the same MetAP class more often than a composition-matched expectation predicts, for every permissive starting residue separately and after Ala and Ser are removed. The conserved quantity is therefore the predicted processing fate, a statement about sequence class that does not by itself show that methionine is removed or retained in any given protein. This agrees with the observation of Bonissone and colleagues (Bonissone et al. 2013) that the position is conserved beyond expectation, mainly at Ala and Ser, but differs from their reading of it as conservation of residue identity, which rested on the pooled permissive class losing its excess once Ala and Ser were removed: with a composition-matched expectation for each residue separately, class is retained where identity is not, for Gly, Pro, Thr and Val as well as Ala and Ser; archaea show the same pattern; and the retention side carries a structural correlate that survives control for model confidence and is reproduced on experimental coordinates. The rebuttal is partial, because Ala and Ser are also the preferred destinations of class-keeping substitutions (73 against 66 percent), so a model in which

exposure of Ala or Ser specifically is the functional outcome is narrowed rather than excluded, and the shift in second-residue usage from Lys in bacteria to Ser and Ala in eukaryotes (Shemesh et al. 2010) is the reason a composition-matched control was required.

Two constraints on one position

The two sides of the class behave differently, and the difference is the main biological content of the study. On the removal side, the deeply conserved proteins are dominated by the translation apparatus; a minority is locked to a single residue by chemistry (the Cys1 nucleophile or the Gly2 myristoyl acceptor) and the majority only to the class, as expected if these proteins need a small exposed residue at the mature N-terminus and are indifferent to which one. Why the translation apparatus carries this requirement is not resolved here; the companion ribosome census (Y.-H. Chang, manuscript submitted; <https://doi.org/10.5281/zenodo.22981236>) shows that termini embedded in rRNA preferentially lose the initiator methionine, and the present enrichment is consistent with that observation without adding a mechanism.

On the retention side there is no functional category. The deeply conserved nonpermissive proteins are distinguished structurally: Met1 is more often packed into the protein's own fold, and the fold-packed human core is a β 1-strand core in which the methionine is the first residue of a buried strand. The second residue is then conserved as the residue that keeps Met1 there. The class-switch analysis supports this reading: orthologs that lose the nonpermissive class are ten times more often proteins whose Met1 is not packed, and in the seven paired cases where a packed Met1 did switch, the ortholog's Met1 environment is predicted to have loosened. Gains of Met retention carry no packing signature, so the structural constraint explains the deeply conserved core rather than the nonpermissive class as a whole.

The size of the packing correlate and where the evidence rests

The correlate should not be overstated. The binary odds ratios from AlphaFold models are inflated by a confidence confound, and the confidence-controlled estimates of 2.1, 2.1 and 4.4 are the effect sizes reported here. A twofold enrichment in odds is a property of the set, not of every conserved protein: a little over half of deeply conserved human nonpermissive proteins are predicted packed, a further fifth are buried at a subunit interface and a quarter are neither. Because the models carry a confidence term, the load-bearing evidence is the test on deposited coordinates (OR 2.7, $P = 0.029$ in human, with 92 percent agreement between model and coordinates), which is free of the confound; the *E. coli* test points the same way but is underpowered, and no archaeal test is possible with the present structural record. The processing-state indicators from the same entries are consistent with the class prediction but come from heterologous expression hosts and are not a substitute for measurement in native cells.

What the residual set does and does not show

About 15 human proteins remain in which the class is conserved and structure offers no reason. Their termini are typically disordered, NatB-type and often acetylated, but the same features are at least as common among nonpermissive termini that are not deeply conserved, so the residual does not define a mechanism and no class-level claim is made for it. What it does show is that its second residue is held as a strict identity, indistinguishable from the fold-packed core, while the residues downstream are no more conserved than elsewhere; a residue with no function would not survive in this way. ARFRP1, RAD21 and EFTUD2 carry individual hypotheses drawn from published work (Behnia et al. 2004; Setty et al. 2004; Grunwald et al. 2020; Varland et al. 2023;

Gligoris et al. 2014; Cheng et al. 2020; Agafonov et al. 2016); a partner-binding role is the natural guess for the others, but it is a guess. The strict identity of the residual set and the EFTUD2 motif are recorded because they are reproducible, not because they are understood, and the eEF2 contrast shows what the residual proteins are not: there the conserved segment serves the GTPase fold, and the second residue is simply part of it.

Limitations and predictions

The analysis is associative: it shows that predicted class is conserved and that conserved retention travels with a packed Met1, not that packing causes retention to be selected, and the class-switch test is a finer view of the same association rather than independent evidence. No processing state is measured. Anchor-based orthology misses paralog-rich families and divergent N-termini, and the anchor-free rerun recovers only the former. Start-codon annotation in prokaryotes can be wrong, and the internal-methionine sensitivity analysis reduces this concern without removing it. The nonpermissive substitution signal is weaker than the permissive one and, in eukaryotes, partly reflects conservation of the N-acetylation sub-class; the prokaryotic comparisons are the cleaner evidence that methionine retention itself is selected.

The model makes testable predictions. N-terminomics in native cells should find the initiator methionine retained in the fold-packed core and in the residual set and removed in the deeply conserved permissive set. In the core, substitution of the second residue by a permissive one should produce a folding or stability defect while substitution by another nonpermissive residue is tolerated; the residual set, where identity rather than class is held, predicts the opposite tolerance; and for the three leads the predictions are loss of Golgi targeting (ARFRP1), loss of SMC3 binding (RAD21) and a tri-snRNP assembly or splicing defect (EFTUD2). The ranked sets, packing scores and structure annotations are provided so that each test can be chosen from a defined list.

MATERIALS AND METHODS

This is a computational study; no reagents, cell lines, strains or animals were used. Full procedural detail, every threshold and all sensitivity analyses are given in Supplementary Methods; the essentials follow.

Data, class definition and orthologs

Protein sequences came from 40 UniProt reference proteomes (release 2026_03; 21 eukaryotes in seven lineage groups, 11 bacteria, 8 archaea; Supplementary Table S1), keeping canonical, full-length, methionine-initiated entries (Swiss-Prot subset for human). Predicted structures came from the AlphaFold Protein Structure Database and experimental structures from the PDB through SIFTS, with contacts read from the RCSB ModelServer. Each protein was classed from its second residue as MetAP-permissive (Gly, Ala, Ser, Cys, Thr, Pro or Val) or MetAP-nonpermissive, with NatB-type and NatC-type sub-classes recorded; UniProt processing annotations are reported separately as annotated processing state. Anchored orthology was defined by DIAMOND reciprocal best hits from human, *E. coli* K-12 and *H. volcanii*, retaining a pair only if both alignment starts fall within the first 50 residues and differ by no more than about 10 percent, so that position two is collinear. Anchor-free orthogroups were inferred with OrthoFinder on 34 of the proteomes, keeping one N-terminally matched representative per species.

Class-conservation test and deeply conserved sets

The fraction of substitutions at position two keeping the anchor's class was compared with the fraction at positions 3 to 12 of the same alignments, reweighted to the mix of starting residues at position two; the excess was expressed as a z score from the binomial variance and retested with the species (sign test) and the anchor protein (Wilcoxon signed-rank test) as the unit, for each starting residue separately and with Ala2 and Ser2 removed. A protein was called deeply conserved when its class is kept in at least half of its orthologs in at least a minimum number of lineage groups (human 4 of 8; *E. coli* 6 of 10 bacteria; *H. volcanii* 4 of 7 archaea) with at least 80 percent agreement over all retained orthologs; a Poisson-binomial score was used for ranking only. Orthogroups were scored under cross-domain and domain-level criteria mirroring these rules.

Structural scoring

Long-range contacts of Met1 were counted as residues at least six positions away with any heavy atom within 4.5 Å of any Met1 heavy atom. On AlphaFold models the packing call required pLDDT at residue 2 of at least 70 and at least three such contacts, and every comparison was repeated within pLDDT terciles, restricted to proteins passing the confidence gate, and on contact counts; the confidence-controlled estimates are the reported effect sizes. Interface burial of human N-termini was taken from the companion census of human experimental assemblies (Chang 2026a; <https://doi.org/10.5281/zenodo.22966146>). On deposited coordinates, entries qualified if the chain begins at UniProt residue 1 or exactly at residue 2; the packing call was at least three long-range intrachain contacts with no confidence term, other chains and crystallographic symmetry copies were counted as interface contacts, per-protein values are the maximum over qualifying entries, and all counts are from the corrected, symmetry-aware recount described in Supplementary Methods. Class switches were scored on AlphaFold models of eukaryotic orthologs of each human protein; the deeply conserved nonpermissive proteins were partitioned as fold-packed on experimental coordinates (core when also passing the AlphaFold gates), interface-buried, predicted-only or residual, with secondary structure, Met1-pocket conservation and second-residue identity baselines computed across the orthologs as described in Supplementary Methods.

Statistics, software and data

Proportions were compared with Fisher's exact test and reported as odds ratios with two-sided P values (with a conditional maximum-likelihood 95 percent confidence interval for the primary experimental comparison), counts with the Mann–Whitney U test and paired comparisons with the Wilcoxon signed-rank test; all P values are two-sided, no correction for multiple testing was applied to the pre-specified primary comparisons, and secondary comparisons are descriptive. Analyses were run in Python 3.14 with pandas 3.0 and SciPy 1.18; the 36-script pipeline is deposited on Zenodo (Chang 2026b) under CC BY 4.0, and database releases and software versions (September 2026) are listed in Supplementary Methods. Scripts were developed with generative-AI assistance and independently audited; see the AI Disclosure Statement.

SUPPLEMENTARY MATERIAL

Supplementary Material available with the article: Supplementary Methods and Supplementary Results (Supplementary_Methods_and_Results.docx; the full procedural description and the extended results text condensed from the main text); Supplementary Tables S1 to S27 in one

workbook (Supplementary_Tables_S1-S27_workbook.xlsx; ranked and deeply conserved sets for each anchor and for the anchor-free design, robustness and confound tests, per-protein structural scores on predicted and deposited structures, class-switch tables, the core and residual analyses and the identity baselines; the README sheet lists the contents); Supplementary Figure S1, the study design (Supplementary_Figure_S1_design_schematic.tif); and the analysis pipeline (Supplementary_Data_pipeline_scripts.zip, 36 scripts with a step-by-step description), which is also deposited on Zenodo (<https://doi.org/10.5281/zenodo.23112899>).

ACKNOWLEDGMENTS

This research received no specific grant from any funding agency in the public, commercial or not-for-profit sectors. Generative-AI assistance with code, figures and text is described in the AI Disclosure Statement.

AUTHOR CONTRIBUTIONS

Yie-Hwa Chang: Conceptualization; Data curation; Formal analysis; Investigation; Methodology; Software; Validation; Visualization; Writing – original draft; Writing – review and editing.

CONFLICT OF INTEREST STATEMENT

The author declares no competing interests.

DATA AVAILABILITY STATEMENT

All input data are public: UniProt reference proteomes, the AlphaFold Protein Structure Database and the Protein Data Bank (mapped through SIFTS). The complete set of ranked sets, per-protein structural scores, experimental-structure extractions, class-switch tables and per-protein analyses is provided as Supplementary Tables S1 to S27 (one workbook). The analysis code (36 scripts with a step-by-step pipeline description), the supplementary workbook and the figure-build script are deposited on Zenodo (Chang 2026b; <https://doi.org/10.5281/zenodo.23112899>, CC BY 4.0). Interface-burial values for human proteins were taken from the per-protein census table of the companion human assembly census (Chang 2026a; <https://doi.org/10.5281/zenodo.22966146>).

AI DISCLOSURE STATEMENT

Generative artificial-intelligence assistance (Perplexity Computer, Perplexity AI; used August to October 2026) was employed under the author's direction in three parts of this work. (i) Code and analysis: the pipeline, statistical and figure scripts deposited at Zenodo (10.5281/zenodo.23112899) were written and run with AI assistance to the author's specification; the author reviewed the code, and all contact counts and tests were independently recomputed in a second pass (symmetry-aware recount, two-sided tests) and reconciled with the published census archive (version 1.4.0). (ii) Figures and tables: Figures 1 to 6, Figure S1 and Tables I and II were drawn by those scripts directly from the output tables; no image was generated or altered by an AI image tool, and the author inspected every panel against the underlying values. (iii) Text: the manuscript was drafted and edited with AI assistance from the author's outline, results and interpretation, and the author checked every statement, number and reference against the primary sources. The AI tool did not originate the hypotheses, study design or conclusions, is not an author, and the author takes full responsibility for the content.

REFERENCES

- Agafonov DE, Kastner B, Dybkov O, Hofele RV, Liu W-T, Urlaub H, Lührmann R, Stark H (2016) Molecular architecture of the human U4/U6.U5 tri-snRNP. *Science* 351:1416-1420. <https://pubmed.ncbi.nlm.nih.gov/26912367/>
- Aksnes H, Drazic A, Marie M, Arnesen T (2016) First things first: vital protein marks by N-terminal acetyltransferases. *Trends Biochem Sci* 41:746-760. <https://doi.org/10.1016/j.tibs.2016.07.005>
- Bonissone S, Gupta N, Romine M, Bradshaw RA, Pevzner PA (2013) N-terminal protein processing: a comparative proteogenomic analysis. *Mol Cell Proteomics* 12:14-28. <https://pmc.ncbi.nlm.nih.gov/articles/PMC3536895/>
- Behnia R, Panic B, Whyte JRC, Munro S (2004) Targeting of the Arf-like GTPase Arl3p to the Golgi requires N-terminal acetylation and the membrane protein Sys1p. *Nat Cell Biol* 6:405-413. <https://pubmed.ncbi.nlm.nih.gov/15077113/>
- Buchfink B, Reuter K, Drost H-G (2021) Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* 18:366-368. <https://pubmed.ncbi.nlm.nih.gov/33828273/>
- Chang Y-H (2026a) Human protein complex interfaces reveal candidate assembly roles for initiator methionine excision. *bioRxiv*, version 2, 28 September 2026. <https://doi.org/10.64898/2026.08.31.748389> (data archive version 1.4.0: <https://doi.org/10.5281/zenodo.22966146>)
- Chang Y-H (2026b) Conservation of the MetAP processing class of the second residue across three domains of life: analysis pipeline, supplementary workbook and figure scripts. Version 1.1.0. Zenodo. <https://doi.org/10.5281/zenodo.23112899>
- Cheng H, Zhang N, Pati D (2020) Cohesin subunit RAD21: from biology to disease. *Gene* 758:144966. <https://pmc.ncbi.nlm.nih.gov/articles/PMC7949736/>
- Dana JM, Gutmanas A, Tyagi N, Qi G, O'Donovan C, Martin M, Velankar S (2019) SIFTS: updated Structure Integration with Function, Taxonomy and Sequences resource. *Nucleic Acids Res* 47:D482-D489. <https://academic.oup.com/nar/article/47/D1/D482/5184711>
- Emms DM, Kelly S (2019) OrthoFinder: phylogenetic orthology inference for comparative genomics. *Genome Biol* 20:238. <https://genomebiology.biomedcentral.com/articles/10.1186/s13059-019-1832-y>
- Frottin F, Martinez A, Peynot P, Mitra S, Holz RC, Giglione C, Meinnel T (2006) The proteomics of N-terminal methionine cleavage. *Mol Cell Proteomics* 5:2336-2349. <https://www.sciencedirect.com/science/article/pii/S1535947620306940>
- Giglione C, Boularot A, Meinnel T (2004) Protein N-terminal methionine excision. *Cell Mol Life Sci* 61:1455-1474. <https://pubmed.ncbi.nlm.nih.gov/15197470/>
- Gligoris TG, Scheinost JC, Bürmann F, Petela N, Chan K-L, Uluocak P, Beckouët F, Gruber S, Nasmyth K, Löwe J (2014) Closing the cohesin ring: structure and function of its Smc3-kleisin interface. *Science* 346:963-967. <https://www.science.org/doi/abs/10.1126/science.1256917>

Grunwald S, Hopf LVM, Bock-Bierbaum T, Lally CCM, Spahn CMT, Daumke O (2020) Divergent architecture of the heterotrimeric NatC complex explains N-terminal acetylation of cognate substrates. *Nat Commun* 11:5506. <https://www.nature.com/articles/s41467-020-19321-8>

Hirel PH, Schmitter JM, Dessen P, Fayat G, Blanquet S (1989) Extent of N-terminal methionine excision from *Escherichia coli* proteins is governed by the side-chain length of the penultimate amino acid. *Proc Natl Acad Sci USA* 86:8247-8251. <https://www.pnas.org/doi/pdf/10.1073/pnas.86.21.8247>

Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al. (2021) Highly accurate protein structure prediction with AlphaFold. *Nature* 596:583-589. <https://pmc.ncbi.nlm.nih.gov/articles/PMC8371605/>

Kim H-K, Kim R-R, Oh J-H, Cho H, Varshavsky A, Hwang C-S (2014) The N-terminal methionine of cellular proteins as a degradation signal. *Cell* 156:158-169. <https://pubmed.ncbi.nlm.nih.gov/24361105/>

Setty SRG, Strohlic TI, Tong AHY, Boone C, Burd CG (2004) Golgi targeting of ARF-like GTPase Arl3p requires its N α -acetylation and the integral membrane protein Sys1p. *Nat Cell Biol* 6:414-419. <https://pubmed.ncbi.nlm.nih.gov/15077114/>

Shemesh R, Novik A, Cohen Y (2010) Follow the leader: preference for specific amino acids directly following the initial methionine in proteins of different organisms. *Genomics Proteomics Bioinformatics* 8:180-189. <https://pubmed.ncbi.nlm.nih.gov/20970746/>

Sherman F, Stewart JW, Tsunasawa S (1985) Methionine or not methionine at the beginning of a protein. *BioEssays* 3:27-31. <https://pubmed.ncbi.nlm.nih.gov/3024631/>

Varadi M, Bertoni D, Magana P, Paramval U, Pidruchna I, Radhakrishnan M, Tsenkov M, Nair S, Mirdita M, Yeo J, et al. (2024) AlphaFold Protein Structure Database in 2024: providing structure coverage for over 214 million protein sequences. *Nucleic Acids Res* 52:D368-D375. <https://pubmed.ncbi.nlm.nih.gov/37933859/>

Varland S, Osberg C, Arnesen T (2015) N-terminal modifications of cellular proteins: the enzymes involved, their substrate specificities and biological effects. *Proteomics* 15:2385-2401. <https://pubmed.ncbi.nlm.nih.gov/25914051/>

Varland S, Silva RD, Kjosås I, Faustino A, Bogaert A, Billmann M, Boukhatmi H, Kellen B, Costanzo M, Drazic A, et al. (2023) N-terminal acetylation shields proteins from degradation and promotes age-dependent motility and longevity. *Nat Commun* 14:6774. <https://www.nature.com/articles/s41467-023-42342-y>

Varshavsky A (2024) N-degron pathways. *Proc Natl Acad Sci USA* 121:e2408697121. <https://www.pnas.org/doi/abs/10.1073/pnas.2408697121>

Wingfield PT (2017) N-terminal methionine processing. *Curr Protoc Protein Sci* 88:6.14.1-6.14.3. <https://pmc.ncbi.nlm.nih.gov/articles/PMC5663234/>

Xiao Q, Zhang F, Nacev BA, Liu JO, Pei D (2010) Protein N-terminal processing: substrate specificity of *Escherichia coli* and human methionine aminopeptidases. *Biochemistry* 49:5588-5599. <https://pubs.acs.org/doi/abs/10.1021/bi1005464>

TABLES

Table I. Conservation of the MetAP class at position two against a composition-matched expectation, and the deeply conserved sets, for the six anchored comparisons. Columns: anchor; target group; species in group; ortholog pairs with N-termini matched from residue 1; substitutions at position two; fraction keeping the anchor's class; composition-matched expectation from positions 3 to 12; z score; species with a positive excess (of total); Wilcoxon signed-rank P with the anchor protein as the unit; identity conservation at position two and at positions 3 to 12. A second block gives, for each anchor, the number of scored proteins and of deeply conserved proteins by class (human 605 scored, 359 deeply conserved, 263 permissive and 96 nonpermissive from 411 and 194 scored; E. coli 204, 85, 45 and 40 from 103 and 101; H. volcanii 226, 99, 46 and 53 from 122 and 104). Class is a sequence prediction; UniProt-annotated initiator-methionine removal, where present, is an observed state listed for information (134 deeply conserved human permissive proteins, 41 E. coli and 3 H. volcanii; none nonpermissive, after correction of one human entry, UBE2M, whose removal annotation belongs to a non-canonical isoform). Per-residue and Ala2/Ser2-removed values are in Supplementary Table S6.

Block A. Class conservation at position two.

Anchor	Target group	Species	Pairs	Subs.	Kept	Expected	z	Species +	P (sign)	P (Wilcoxon)	Id. 2	Id. 3 -12
Human	Other eukaryotes (non-vertebrate)	16	5272	1698	0.862	0.577	24.3	16/16	3.1×10^{-5}	5.9×10^{-39}	0.68	0.58
Human	Vertebrates	4	31023	4886	0.714	0.654	9.1	4/4	n.t.	9.7×10^{-7}	0.84	0.77
Human	Bacteria	11	115	70	0.771	0.542	4.0	8/8	0.008	0.10	0.39	0.44
Human	Archaea	8	162	83	0.807	0.514	5.5	8/8	0.008	0.004	0.49	0.44
E. coli	Bacteria	10	1943	1010	0.741	0.599	9.4	10/10	0.002	7.5×10^{-4}	0.48	0.51
H. volcanii	Archaea	7	1260	621	0.76	0.579	9.5	7/7	0.016	2.2×10^{-12}	0.51	0.57

Block B. Scored and deeply conserved proteins by anchor and class.

Anchor	Scored	DC	Perm. scored	Perm. DC	Nonperm. scored	Nonperm. DC	Perm. DC, iMet removal annotated	Nonperm. DC, iMet removal annotated
Human	605	359	411	263	194	96	134	0
E. coli	204	85	103	45	101	40	41	0
H. volcanii	226	99	122	46	104	53	3	0

Column key: Species, target species with ortholog pairs aligned from residue 1; Pairs, ortholog pairs aligned from residue 1 in both proteins; Subs., substitutions at position two; Kept, fraction of substitutions keeping the anchor class; Expected, composition-matched expectation from positions 3 to 12; Species +, species with a positive excess of species with at least five substitutions; P (sign), sign test over species (n.t., fewer than five species); P (Wilcoxon), signed-rank test with the anchor protein as the unit; Id. 2 and Id. 3–12, identity between orthologs at position two and at positions 3 to 12. DC, deeply conserved; iMet, initiator methionine.

Table II. Met1 packing on experimental coordinates and processing-state indicators from deposited structures, by class and conservation, for human and E. coli. Upper block, nonpermissive proteins with a modelled Met1: number of proteins, number fold-packed (at least three long-range contacts), percent, odds ratio and Fisher P for deeply conserved against other (human 43 of 61 against 16 of 34, OR 2.7, P = 0.029; E. coli 22 of 28 against 22 of 34, OR 2.0, P

= 0.27), median long-range contacts (5 against 2; 6 against 3.5), the same test with interface burial counted (human 89 against 74 percent, $P = 0.086$; *E. coli* 96 against 74 percent, $P = 0.017$), the same test restricted to entries without a tag remnant (human OR 3.0, *E. coli* OR 2.4), and agreement between the AlphaFold and experimental calls (92 and 85 percent). Lower block, processing-state indicators by class among all proteins with a qualifying structure (475 human and 165 *E. coli* anchor-scored proteins): proteins with Met1 modelled in at least one entry (nonpermissive 64 and 88 percent, permissive 26 and 16 percent; OR 5.0 and 36.5), proteins with a chain deposited from residue 2 (permissive 34 and 67 percent, nonpermissive 13 and 12 percent; OR 3.4 and 14.7), and, in constructs with a native N-terminus, proteins in which Met1 is absent while residue 2 is modelled (permissive 40 and 57 percent, nonpermissive 7 and 7 percent; OR 8.2 and 18.3). Footnote: these indicators are consistent with the class prediction but are not direct evidence of MetAP action, because an unmodelled Met1 can reflect terminal disorder or processing by a heterologous expression host; the 19 human nonpermissive proteins with a chain deposited from residue 2 include β - and γ -actin, whose acetylated Met1 is removed by a dedicated pathway, and are listed in Supplementary Table S16. Structures qualified when the deposited chain began at UniProt residue 1 or exactly at residue 2; only one *H. volcanii* protein qualified. The tag-remnant-restricted and one-entry-per-protein odds ratios quoted in the title are from Supplementary Table S14; the main tests are tabulated here. All P values are two-sided; contact counts are from the corrected recount described in Materials and Methods.

Block A. Met1 packing on experimental coordinates (nonpermissive proteins with Met1 modelled).

Species	Measure	Group A	Value A	Group B	Value B	Odds ratio	P
Human	Nonpermissive, Met1 modelled: Met1 fold-packed (≥ 3 long-range intrachain residues within 4.5 Å)	deeply conserved	0.70 (43/61)	other	0.47 (16/34)	2.7 (95% CI 1.0–7.0)	0.029
Human	Nonpermissive, Met1 modelled: Met1 fold-packed or at interface	deeply conserved	0.89 (54/61)	other	0.74 (25/34)	2.8	0.086
Human	Nonpermissive, Met1 modelled: long-range contact count (median)	deeply conserved	5 (n=61)	other	2 (n=34)		0.038
Human	AlphaFold fold-packed vs experimental fold-packed (nonpermissive, Met1 modelled)	agreement	0.92 (n=95)	AF+/exp+, AF+/exp-, AF-/exp+, AF-/exp-	56, 5, 3, 31	116	1.8×10^{-16}
<i>E. coli</i>	Nonpermissive, Met1 modelled: Met1 fold-packed (≥ 3 long-range intrachain residues within 4.5 Å)	deeply conserved	0.79 (22/28)	other	0.65 (22/34)	2.0	0.27
<i>E. coli</i>	Nonpermissive, Met1 modelled: Met1 fold-packed or at interface	deeply conserved	0.96 (27/28)	other	0.74 (25/34)	9.7	0.017
<i>E. coli</i>	Nonpermissive, Met1 modelled: long-range contact count (median)	deeply conserved	6 (n=28)	other	3.5 (n=34)		0.039

Species	Measure	Group A	Value A	Group B	Value B	Odds ratio	P
E. coli	AlphaFold fold-packed vs experimental fold-packed (nonpermissive, Met1 modelled)	agreement	0.85 (n=62)	AF+/exp+, AF+/exp-, AF-/exp+, AF-/exp-	37, 2, 7, 16	42	1.0×10^{-7}

Block B. Processing-state indicators among all proteins with a qualifying structure.

Species	Measure	Group A	Value A	Group B	Value B	Odds ratio	P
Human	Met1 modelled in at least one entry	nonpermissive	0.64 (95/148)	permissive	0.26 (86/327)	5.0	6.4×10^{-15}
Human	Entry deposited with chain starting at residue 2	permissive	0.34 (110/327)	nonpermissive	0.13 (19/148)	3.4	1.3×10^{-6}
Human	Native construct: Met1 absent while residue 2 modelled	permissive	0.40 (103/259)	nonpermissive	0.07 (9/121)	8.2	8.8×10^{-12}
E. coli	Met1 modelled in at least one entry	nonpermissive	0.88 (64/73)	permissive	0.16 (15/92)	36.5	4.2×10^{-21}
E. coli	Entry deposited with chain starting at residue 2	permissive	0.67 (62/92)	nonpermissive	0.12 (9/73)	14.7	3.0×10^{-13}
E. coli	Native construct: Met1 absent while residue 2 modelled	permissive	0.57 (40/70)	nonpermissive	0.07 (4/59)	18.3	5.7×10^{-10}

FIGURE LEGENDS

All figures were drawn by the deposited scripts directly from the output tables, with generative-AI assistance in writing those scripts; no image was generated or altered by an AI image tool (see the AI Disclosure Statement).

Figure 1. The MetAP class of the second residue is conserved above a composition-matched expectation in all three domains of life. (A) Fraction of substitutions at position two that kept the anchor's MetAP class (filled bars) against the composition-matched expectation derived from positions 3 to 12 of the same ortholog pairs (open bars), for the six anchored comparisons: human against other eukaryotes (n = 1,698 substitutions), human against vertebrates (4,886), human against bacteria (70), human against archaea (83), E. coli against other bacteria (1,010) and H. volcanii against other archaea (621). The z score of the excess is given above each pair; all six are positive (z from 4.0 to 24.3). (B) Identity conservation against class conservation at position two and at positions 3 to 12 in the two prokaryotic comparisons. In bacteria the second residue is identical between orthologs no more often than its neighbours (0.48 against 0.51) and in archaea less often (0.51 against 0.57), yet keeps its class far more often than they do, separating conservation of class from conservation of the side chain. (C) Robustness of the

excess of class retention over expectation when the unit of analysis is the species or the protein rather than the substitution. Filled bars give the median excess across the target species of each comparison (16 non-vertebrate eukaryotes, 4 vertebrates, 8 bacteria and 8 archaea against human; 10 bacteria against *E. coli*; 7 archaea against *H. volcanii*); open bars give the mean excess across anchor proteins. The number of species with a positive excess is given above each pair (53 of 53 overall) with the sign-test P value for groups of five or more species; the per-protein Wilcoxon P values (from 1×10^{-1} for human against bacteria to 6×10^{-39} for human against other eukaryotes) are in Supplementary Table S5b. (D) Class retention against the matched control for each permissive starting residue separately (Ala, Gly, Pro, Ser, Thr, Val) and for all permissive starts with Ala2 and Ser2 removed, in the human–eukaryote, *E. coli*–bacteria and *H. volcanii*–archaea comparisons; substitution counts are given beneath each bar. Exact values are in Table I and Supplementary Table S6. Class is a sequence prediction (MetAP-permissive or MetAP-nonpermissive) throughout and is not a measured processing state.

Figure 2. Ranked, deeply conserved sets in each domain and their anchor-free replication.

(A) Ranking of anchor-scored proteins by conservation score for human (605 scored, 359 deeply conserved: 263 permissive, 96 nonpermissive), *E. coli* (204 scored, 85: 45 and 40) and *H. volcanii* (226 scored, 99: 46 and 53). Deeply conserved proteins are coloured by class (permissive, blue; nonpermissive, orange); the top entries on each side are labelled, and the fourteen human proteins that keep their class in bacteria and archaea as well as across eukaryotes are marked. The score is used for ranking only, because its ceiling depends on the class background (Materials and Methods). (B) Agreement between the anchored sets and the sets defined on orthogroups inferred de novo across 34 proteomes. For anchored deeply conserved proteins that fall in a scored orthogroup, the fraction also deeply conserved at the orthogroup level is 236 of 277 for human (85 percent; OR 31, $P = 1 \times 10^{-53}$), 74 of 76 for *E. coli* (OR 167) and 66 of 69 for *H. volcanii* (OR 93). The 168 orthogroups (187 human proteins) scored only by the orthogroup design, mostly paralog-rich families, are listed by human representative in Supplementary Table S11. (C) The 39 orthogroups deeply conserved across all three domains, ordered by number of species and coloured by class, with the residue at position two in each species shown as a strip; EF-Tu/eEF1A (32 species), RPS15A/rpsH/rps8, glutamine–fructose-6-phosphate amidotransferase (Cys2 in 29 of 29 species), Hsp70/DnaK and asparagine synthetase B lead the permissive side, and GDP-mannose pyrophosphorylase (31 species), riboflavin synthase (Phe2 in 18 of 18), UDP-glucose dehydrogenase and exonuclease III/APEX2 the nonpermissive side. Set membership and scores are in Supplementary Tables S2 to S4 and S7 to S10.

Figure 3. Deeply conserved permissive and nonpermissive proteins have different correlates.

(A) Translation-apparatus membership among permissive proteins, deeply conserved against not deeply conserved, in the anchored sets (human 34 against 15 percent, OR 2.9, $P = 3 \times 10^{-5}$; *E. coli* 67 against 21 percent, OR 7.7, $P = 3 \times 10^{-6}$; *H. volcanii* 65 against 14 percent, OR 11, $P = 1 \times 10^{-8}$) and in the anchor-free sets (eukaryotes OR 3.6, bacteria OR 5.1, archaea OR 13). No enrichment is present among nonpermissive proteins. (B) Fold-packing of the initiator methionine on AlphaFold models among nonpermissive proteins, deeply conserved against not deeply conserved, restricted to proteins whose N-terminus passes the confidence gate (pLDDT at residue 2 of at least 70) so that model confidence is held fixed. Bars show the fraction with at least three long-range contacts (human 71 against 55 percent, OR 2.1, $P = 0.063$; *E. coli* 80 against 66 percent, OR 2.1, $P = 0.16$; *H. volcanii* 92 against 73 percent, OR 4.4, $P = 0.021$); box plots show the number of long-range contacts of Met1 (medians 5 against 3, 6 against 4 and 5.0

against 4.5; Mann–Whitney $P = 0.006, 0.019$ and 0.053). These confidence-controlled values are the effect sizes reported in the text; the uncontrolled binary odds ratios and the confidence difference that inflates them are given in Supplementary Tables S12 and S13. (C) The same packing comparison in the anchor-free orthogroup sets (human OR 2.3, $P = 0.013$; *E. coli* OR 2.4, $P = 0.056$; *H. volcanii* OR 6.8, $P = 1.6 \times 10^{-3}$). (D) Fold-packing of Met1 on deposited experimental coordinates, with the same contact definition and no confidence term, among nonpermissive proteins whose Met1 is modelled: human 43 of 61 deeply conserved against 16 of 34 other (70 against 47 percent, OR 2.7, $P = 0.029$); *E. coli* 22 of 28 against 22 of 34 (79 against 65 percent, OR 2.0, $P = 0.27$, not significant). Permissive proteins show no significant difference (human 37 against 25 percent, $P = 0.34$). Only one *H. volcanii* protein has a qualifying structure, so the archaeal test is not possible. (E) Number of long-range contacts of Met1 on experimental coordinates for the same proteins (medians 5 against 2 in human, $P = 0.038$; 6 against 3.5 in *E. coli*, $P = 0.039$). (F) Agreement between the AlphaFold packing call and the experimental call for nonpermissive proteins with a modelled Met1: 92 percent in human (56 both packed, 31 both unpacked, 8 discordant of 95) and 85 percent in *E. coli* (37 both packed, 16 both unpacked, 9 discordant of 62). Long-range contacts are residues at least six positions away in the same chain instance with any heavy atom within 4.5 Å of a Met1 heavy atom; symmetry copies and other chains count as interface contacts (Materials and Methods). All P values are two-sided (Fisher exact test in A, C and D; Mann–Whitney U test in B and E). Panels D to F are built from the experimental-structure validation data in Supplementary Tables S14 to S18.

Figure 4. Class switches between orthologs avoid packed initiator methionines. (A) Among 160 scored human nonpermissive proteins with at least five eukaryotic orthologs, the fraction whose human Met1 is fold-packed on the AlphaFold model, for proteins with at least one ortholog that switched to the permissive class ($n = 28$) against proteins with no switch ($n = 132$): 14 against 55 percent (OR 0.13, $P = 6 \times 10^{-5}$). The median human Met1 long-range contact count of each group is given above the bars (0 against 3, Mann–Whitney $P = 2 \times 10^{-4}$). Permissive proteins show no relationship (22 against 16 percent, OR 1.5, $P = 0.42$). (B) Switch rate counted per ortholog pair: 7 of 936 (0.7 percent) when the human Met1 is packed against 63 of 902 (7.0 percent) when it is not. (C) Paired test in the seven human nonpermissive proteins whose Met1 has at least three long-range contacts and which have AlphaFold models of both a switched and an unswitched ortholog (SEC31A, RPL18A, ATP6V1A, APEX2, ASRGL1, UBE2J1, PRELID3B). Lines join the mean Met1 contact count in unswitched orthologs to that in switched orthologs of the same protein; the switched ortholog has fewer contacts in all seven (means 5.9 against 3.5, Wilcoxon $P = 0.016$), with SEC31A the exception in which the switched orthologs keep Met1 packed. (D) Orthologs that gained a nonpermissive residue at position two from a permissive human protein, showing no difference in predicted Met1 packing between gained and ungained orthologs. Switch status overlaps almost completely with the deeply conserved label (92 deeply conserved without a switch, 40 neither, 25 switched and not deeply conserved, 3 both); the association in A holds within the 65 proteins that are not deeply conserved (OR 0.23, $P = 0.030$). Ortholog-level packing data are in Supplementary Tables S19 to S21.

Figure 5. The deeply conserved nonpermissive core is a β 1-strand class. (A) Partition of the 96 deeply conserved human nonpermissive proteins by the structural evidence for Met1: fold-packed on experimental coordinates (43, of which 41 also pass the AlphaFold packing and confidence gates and form the core), interface-buried but not fold-packed (18), no structure covering Met1 but a confident packing prediction (11), and neither fold-packed nor interface-

buried (24, the residual). The same partition of the 40 E. coli proteins (22 fold-packed, 21 core; 5 interface; 9 predicted only; 4 residual) is shown alongside. (B) Secondary structure at Met1 in the 41 core proteins, assigned from model dihedrals of residues 2 to 6: β -strand in 37, helix in 1 (TRMT112), coil or mixed in 3 (EIF2D, MCTS1, CFAP20). The adjacent panel gives, for each core protein, the number of residues at least six positions away contacted by the Met1 side chain (median 7, dashed line) and their composition (hydrophobic against other residues). (C) Representative Met1 environments from experimental coordinates, one per recurring architecture: DTD1 with its E. coli counterpart dtd (12 long-range contacts in both), the ubiquitin β -grasp of RPS27A, the SMC ATPase N-lobe of SMC3, the longin fold of AP2M1, and the zinc-ribbon RNA polymerase subunit POLR2L. Met1 is drawn with all heavy atoms (orange) and the residues at least six positions away with any heavy atom within 4.5 Å of it (teal, labelled), over the C α trace within 16 Å (grey), projected onto the principal plane of the local environment; the contact counts in the panel titles are those measured on the deposited coordinates of the single chain shown (PDB 2OKV chain A, 1JKE chain A, 3NHE chain B, 8ROK chain A, 6URI chain M, 9EHZ chain J: 12, 12, 9, 9, 9 and 6). (D) Conservation of the Met1-contacting positions against other buried positions across the orthologs of each core protein: per-position identity 0.71 against 0.68 (Wilcoxon $P = 0.25$), with 23 proteins higher and 18 lower at the contact positions; hydrophobic-class conservation (0.92 against 0.91, $P = 0.28$) is given in Supplementary Table S23. Each point is one protein; the diagonal marks equality. Partition assignments and structural values are in Supplementary Tables S22 and S23.

Figure 6. The residual set holds position two strictly without a structural reason, and three proteins are carried forward as individual hypotheses. (A) The 24 residual human proteins separated into five groups: predicted packed but experimentally unresolved or just below the confidence gate (SRPRA, GDI1, BBS2), near a threshold (EIF3I, SHQ1, SEPSECS), uninformative (SOD2, Met1 in the cleaved presequence), individual leads (ARFRP1, RAD21, EFTUD2) and unexplained (14 proteins, listed in Supplementary Table S24). (B) Control for the features shared by the unexplained residual proteins. Frequencies of NatB-type residues at position two, predicted disorder at the N-terminus, annotated Met1 acetylation and membrane keywords are compared between the residual of the deeply conserved set and the residual of the 98 scored human nonpermissive proteins that are not deeply conserved (57 residual): NatB-type 46 percent in the deeply conserved residual against 56 percent in the other residual ($P = 0.47$), disordered 58 against 63 percent, acetylated 25 against 40 percent ($P = 0.22$), membrane 33 against 30 percent, with no difference within a direct stratum match of 35 disordered NatB-type termini. Among nonpermissive proteins with ordered termini, deep conservation is depleted for NatB-type residues (22 against 42 percent, $P = 0.02$). (C) Identity conservation of position two across all orthologs, expressed as the excess over the exposed-residue baseline of the same protein, for the fold-packed core ($n = 41$; median excess +0.23, Wilcoxon $P = 2 \times 10^{-8}$), the deeply conserved residual ($n = 24$; median identity 0.86, excess +0.23, $P = 5 \times 10^{-5}$; 21 of 24 above baseline) and the non-deeply-conserved residual, which is shown for completeness but is circular because deep conservation is defined at this position. Core against residual: $P = 0.56$. The strictest residual proteins are labelled (RABGGTA His2 10 of 10, RAD21 Phe2 15 of 15, RBBP5 Asn2 13 of 13, LSM2 Leu2 14 of 14). (D) N-terminal consensus strings and per-position identity over residues 1 to 30 for the three leads, with predicted order (pLDDT) and burial beneath each: ARFRP1 (Tyr2 in 5 and Phe2 in 6 of 11 deep orthologs, an aromatic NatC-type terminus recognised after acetylation by SYS1), RAD21 (Phe2 in 15 of 15 at the start of a conserved 30-residue segment that binds the SMC3 coiled coil) and EFTUD2 (Asp2 in 8 of 11

within the conserved motif Md..LYDEFGNYIGPeld, 11 positions of residues 2 to 30 at 80 percent identity or more across 14 orthologs, after which identity falls from 0.85 for residues 2 to 10 to 0.25 for residues 21 to 30). In the consensus strings, upper case marks identity of at least 80 percent and lower case a residue class kept in at least 80 percent of orthologs. (E) The same display for eukaryotic elongation factor 2 as a contrast: permissive at position two (Val2 in 15, Gly2 in 6, Pro2 in 2 of 23 deep species), 19 of residues 2 to 30 held at 80 percent identity or more across 20 orthologs (mean identity 0.98 over residues 21 to 30), but with residues 3 onward confidently ordered (pLDDT 70 to 97), residue 4 and most positions from 17 on buried, and residues 26 to 34 forming the P-loop of the G domain. The conserved segment in eEF2 is the first strand of a nucleotide-binding fold; the conserved segment in EFTUD2 is disordered and exposed. Values are in Supplementary Tables S24 to S27.

Figure 1

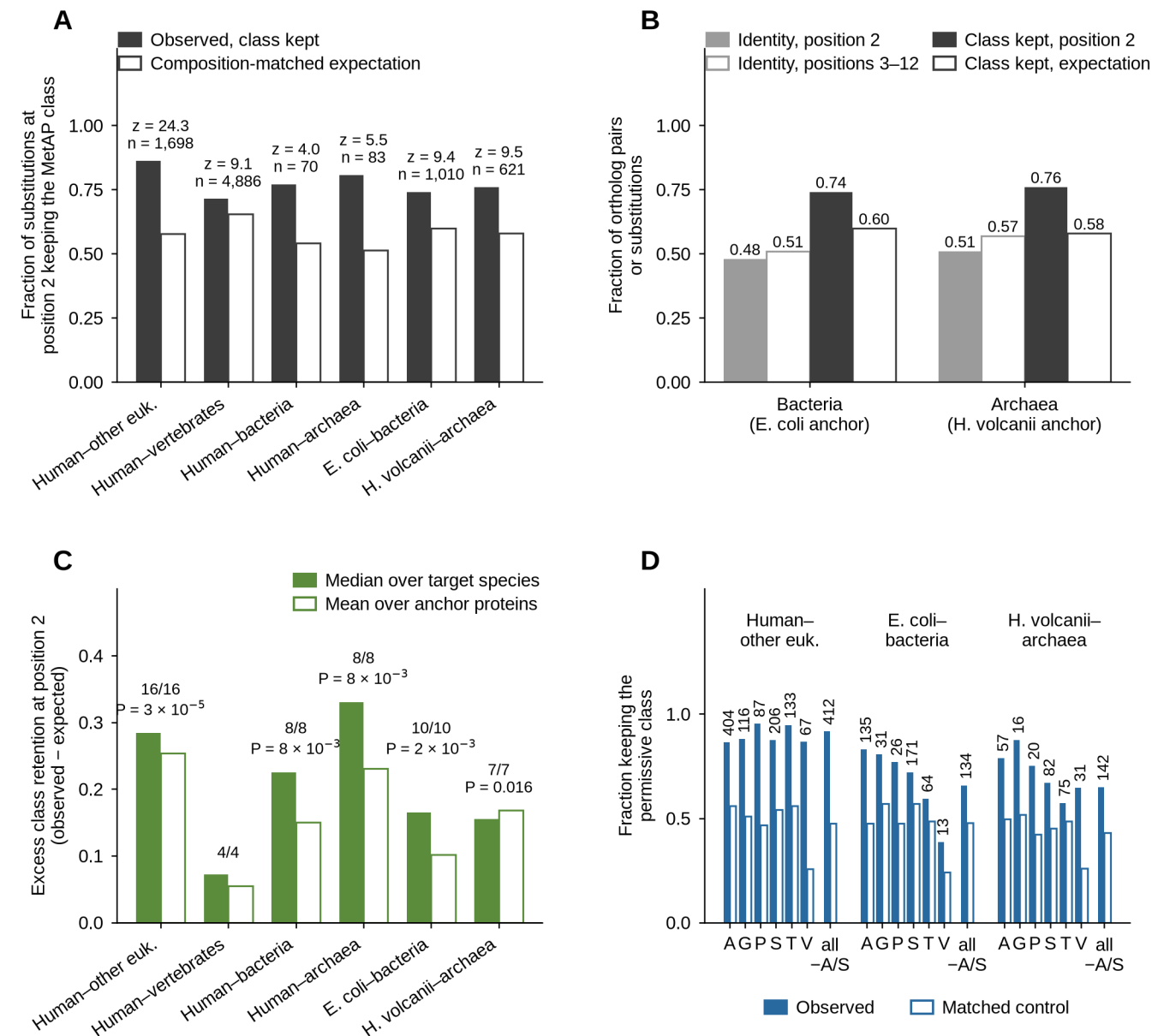


Figure 2

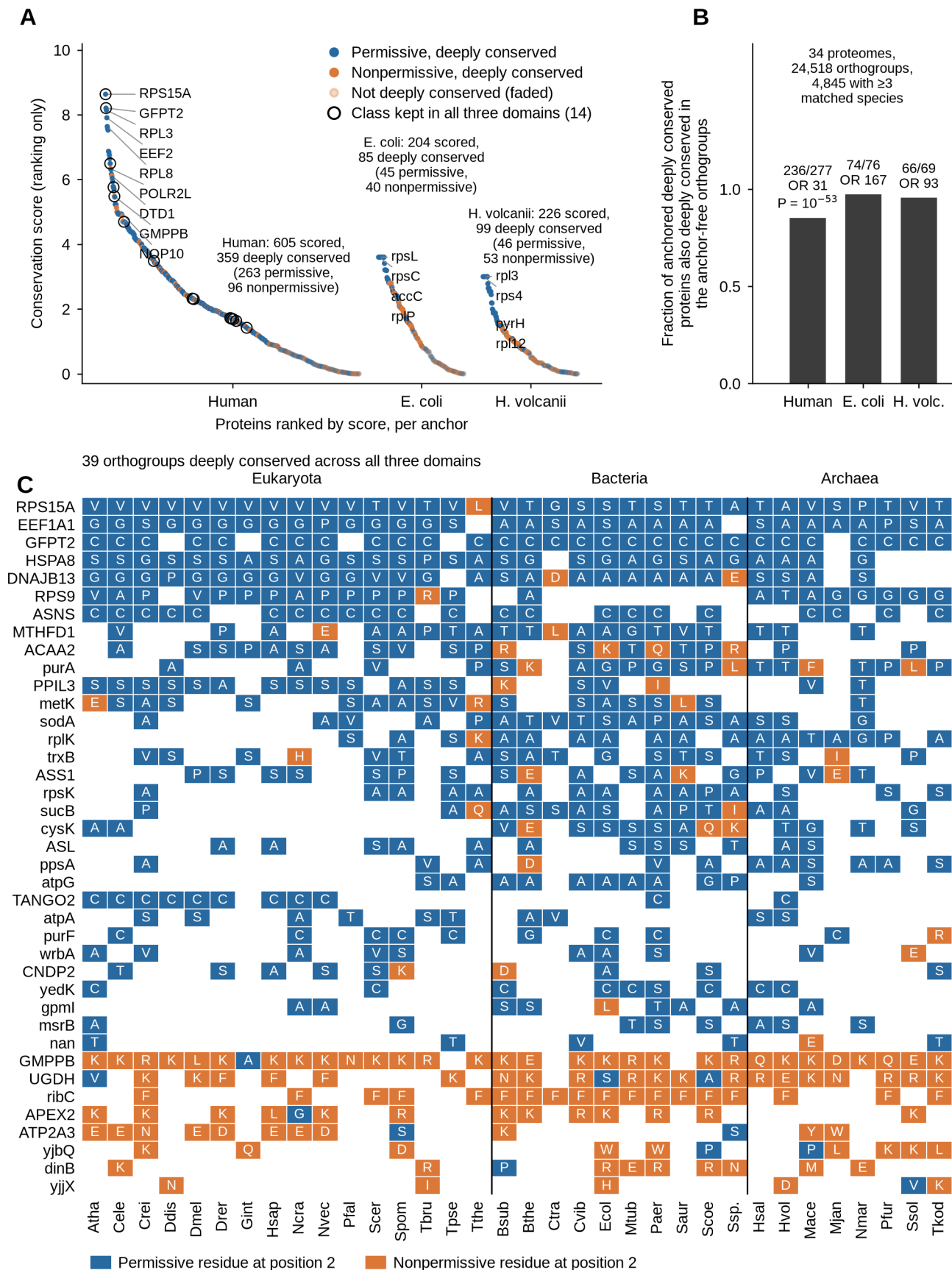


Figure 3

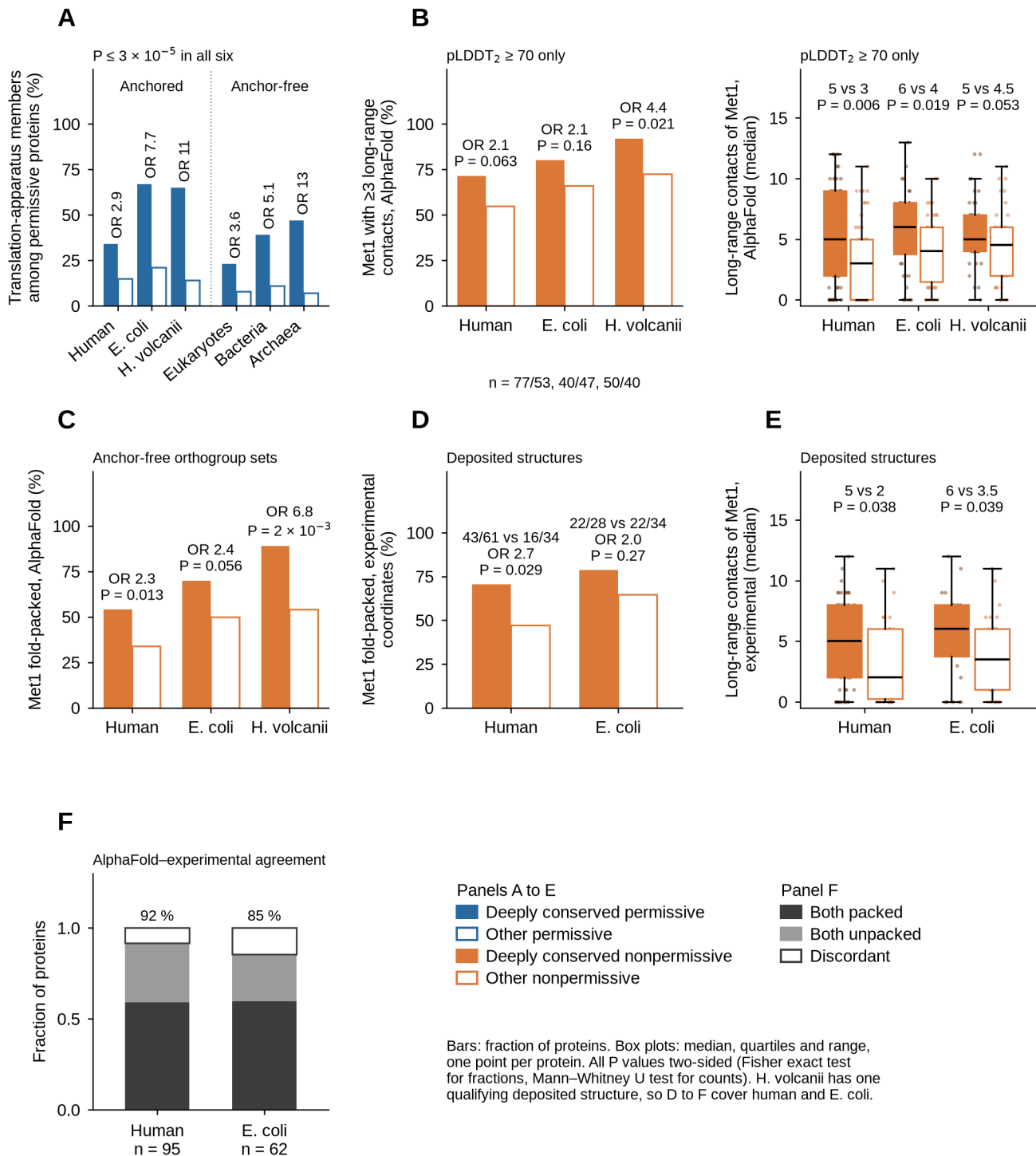


Figure 4

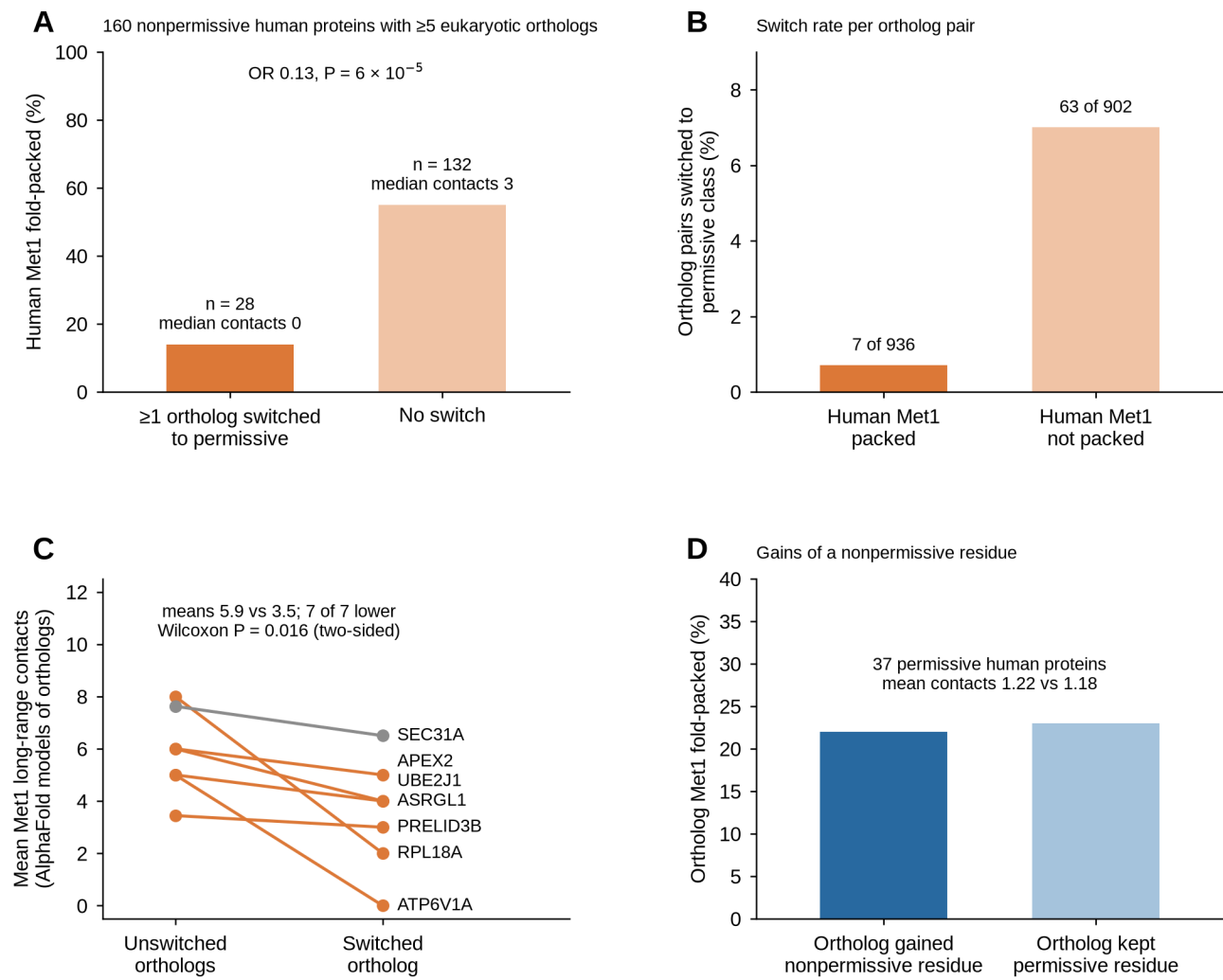


Figure 5

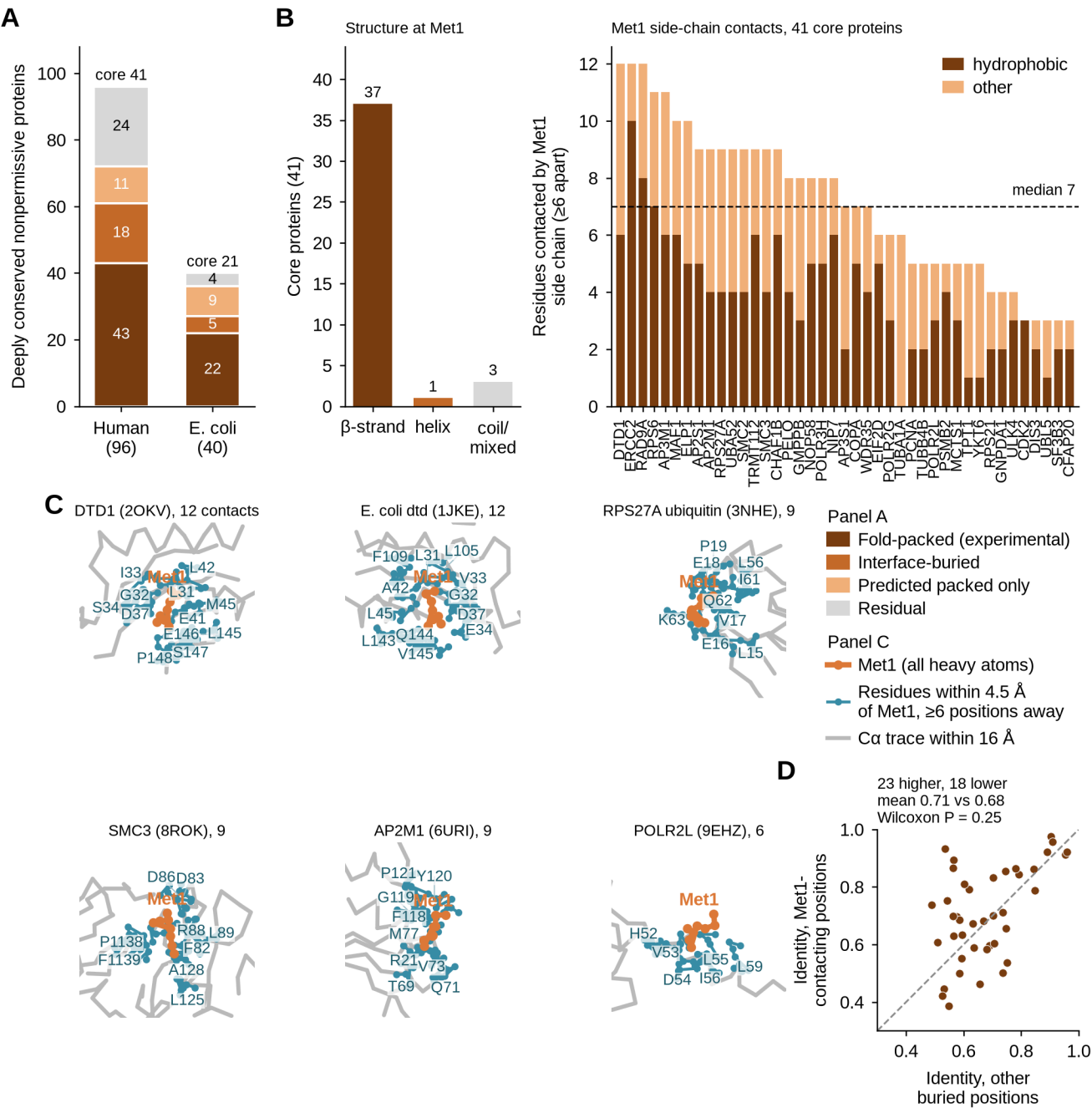
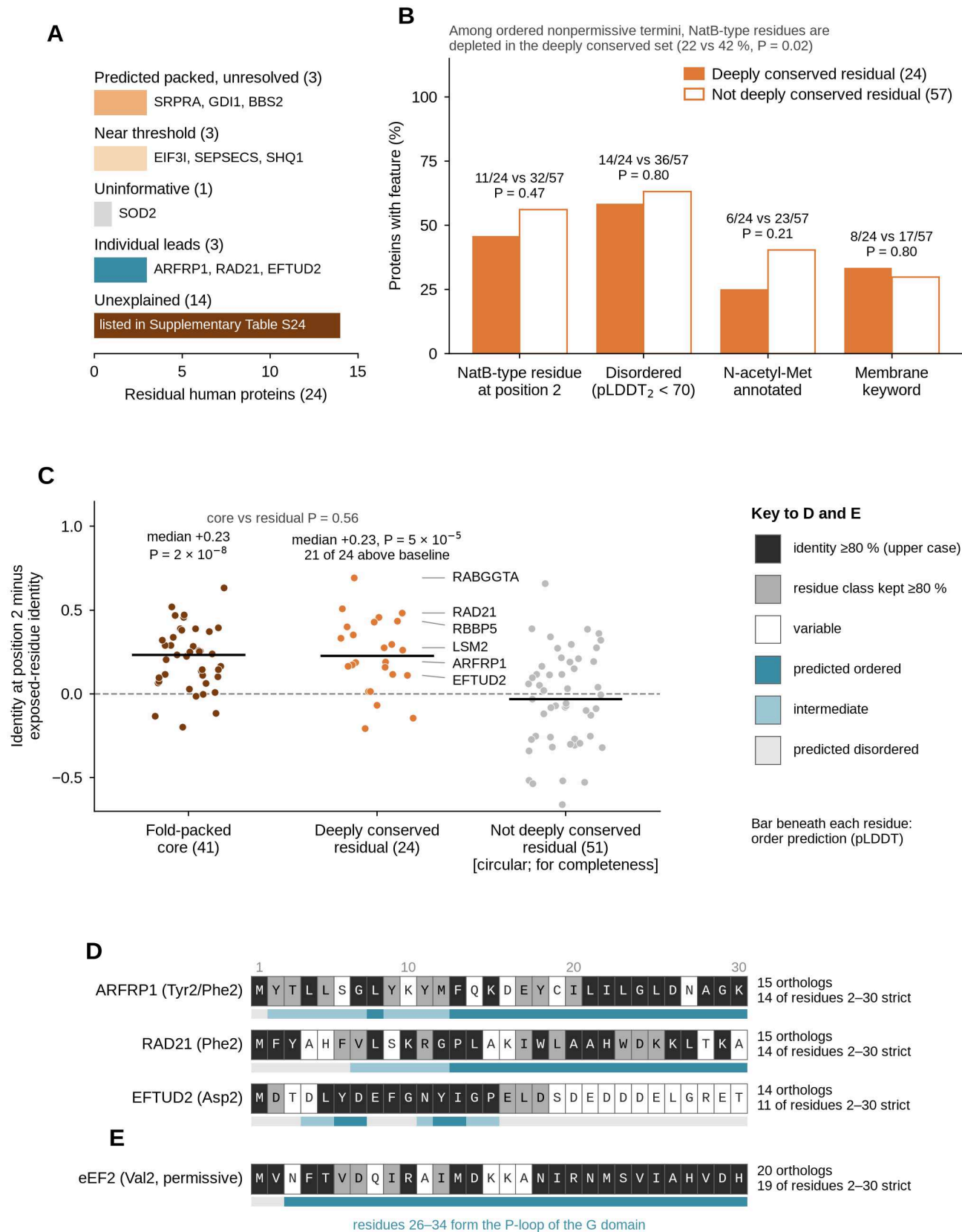


Figure 6



Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- [SupplementaryFigureS1designschematic.pdf](#)
- [SupplementaryMethodsandResults.pdf](#)
- [SupplementaryTablesS1S27workbook.xlsx](#)