Article

Not peer-reviewed version

Reading Between the ABCs: Intrinsic Disorder and Evolutionary Dynamics of Non-Canonical Regions in ABC Transporters

Ichda Arini Dinana , Yukihiko Kubota , Masahiro Ito *

Posted Date: 24 April 2026

doi: 10.20944/preprints202604.1710.v1

Keywords: ABC transporter; intrinsic disorder; post-translational modification; non-canonical regions; evolutionary dynamics; linker regions; phylogenetic signal; site-specific selection; architectural class

 doi:10.20944/preprints202604.1710.v1

Article

Reading Between the ABCs: Intrinsic Disorder and Evolutionary Dynamics of Non-Canonical Regions in ABC Transporters

Ichda Arini Dinana [1], Yukihiko Kubota [2] and Masahiro Ito [1,2,*]

[1] Advanced Life Sciences Program, Graduate School of Life Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577, Japan

[2] Department of Bioinformatics, College of Life Sciences, Ritsumeikan University, 1-1-1 Nojihigashi, Kusatsu, Shiga, 525-8577, Japan

* Correspondence: maito@sk.ritsumei.ac.jp;

Abstract

ATP-binding cassette (ABC) transporters constitute one of the largest membrane protein superfamilies, yet the structural and evolutionary properties of their non-domain regions remain poorly characterized. To elucidate the diversity of these non-canonical regions across evolutionary lineages, we analyzed intrinsic disorder, site-specific selection, and predicted post-translational modification (PTM) sites across five architectural classes comprising 1,581 prokaryotic and eukaryotic sequences. Linker and flanking regions were consistently more disordered than transmembrane and nucleotide-binding domains across all architectures. Disorder fraction differed significantly among region types after phylogenetic correction (Pagel's $\lambda \approx 0.97$). Predicted PTM sites are enriched in disordered non-domain segments, with N-linked glycosylation and phosphoserine showing the strongest positive enrichment; 140 sites satisfied a tiered conservation criterion (MusiteDeep score ≥ 0.5; cross-species conservancy ≥ 30%), including 40 high-confidence or moderate-confidence sites (conservancy ≥ 50%) as well as novel phosphotyrosine candidates in half transporters and NBD-only proteins. Site-specific selection analyses indicated that episodic positive selection was concentrated at inter-domain boundaries, whereas NBD cores were subject to pervasive purifying selection. Together, these findings establish that non-canonical regions of ABC transporters are evolutionarily dynamic and harbor conserved predicted modification sites, supporting their roles as potential regulatory interfaces rather than passive structural linkers.

Keywords: ABC transporter; intrinsic disorder; post-translational modification; non-canonical regions; evolutionary dynamics; linker regions; phylogenetic signal; site-specific selection; architectural class

1. Introduction

ATP-binding cassette (ABC) transporters are among the largest and most phylogenetically widespread membrane protein superfamilies, with members identified in organisms ranging from bacteria to humans [1,2]. These proteins function as primary active transporters, using ATP hydrolysis to translocate diverse substrates, including lipids, ions, peptides, and xenobiotics, across the cellular membrane [3,4]. In humans, 48 members are classified into seven subfamilies (ABCA–ABCG) based on sequence similarity and domain architecture [1,5]. The core structural unit of ABC transporters consists of transmembrane domains (TMDs) and nucleotide-binding domains (NBDs), arranged either as full transporters within a single polypeptide (TMD–NBD–TMD–NBD), as half transporters requiring dimerization (TMD–NBD or NBD–TMD), or as NBD-only soluble proteins lacking transmembrane segments[2,3,6,7]. Domain topology is largely conserved, with most subfamilies following a forward TMD–NBD arrangement, though ABCG members adopt a reversed

 doi:10.20944/preprints202604.1710.v1

NBD–TMD configuration [8,9]. Within this overall architecture, NBDs are the most conserved components, characterized by Walker A, Walker B, and LSGGQ signature motifs, whereas TMDs are more variable and largely determine substrate specificity [2,10].

Beyond these canonical domains, ABC transporters contain inter-domain linkers and N- and C-terminal flanking regions whose structural and evolutionary properties remain comparatively underexplored. Ford et al. [11] reviewed the disordered linker regions in eukaryotic ABC transporters and proposed that these segments may be regulated by post-translational modifications (PTMs), particularly phosphorylation. The best-characterized example is the regulatory R-domain of CFTR (ABCC7), in which phosphorylation of multiple serine residues within a disordered ~200-residue linker between NBD1 and TMD2 controls channel gating through disorder-dependent conformational switching [12,13]. Similar phosphorylation-dependent regulation of cytosolic non-domain regions has been reported in other ABC transporters. In ABCC1, PTM sites within the intrinsically disordered L1 linker modulate protein–protein interactions and transport activity [14,15]. In ABCA1, CK2-mediated phosphorylation of residues within the R1 inter-domain linker reduces cholesterol efflux and apoA-I binding [16], whereas JAK2-mediated tyrosine phosphorylation enhances cholesterol efflux in response to apoA-I stimulation [17]. Whether comparable non-domain PTM landscapes are conserved across all ABC transporter architectures, and how they are shaped by evolutionary constraints, have not been examined at the superfamily level.

The broader context for this question comes from studies of intrinsically disordered regions (IDRs) in other multidomain protein families. Holehouse & Kragelund [18] demonstrated that linker IDRs connecting folded domains influence the effective concentration of adjacent domains and modulate inter-domain interactions in a length- and sequence-dependent manner, with PTMs providing a mechanism for dynamic regulation. IDRs are also enriched in PTM sites relative to structured domains across multiple species, suggesting that this enrichment reflects a general feature of disordered protein segments rather than a family-specific feature[19,20]. From an evolutionary perspective, IDRs typically evolve more rapidly than ordered domains due to relaxed structural constraints. Nevertheless, their biophysical properties—including disorder propensity and compositional bias — can remain phylogenetically conserved even when the primary sequence diverges [21,22]. Whether non-domain regions of ABC transporters exhibit similar patterns, and how their structural properties vary across the five architectural classes, have not been systematically investigated.

Here, we characterize the structural and evolutionary features of non-canonical regions across the ABC transporter superfamily by analyzing intrinsic disorder, site-specific selection, and predicted PTM sites across five architectural classes comprising 1,581 prokaryotic and eukaryotic sequences. We show that non-domain regions are consistently more disordered than domain cores across all architectures, are enriched in predicted PTM sites that are partially conserved across species, and experience distinct evolutionary pressures. In particular, episodic positive selection is concentrated at inter-domain boundaries, whereas NBD cores are subject to pervasive purifying selection. These findings provide a systematic view of the non-canonical sequence space of ABC transporters and identify candidate regulatory regions for experimental investigation.

2. Results

2.1. Domain Architecture and Transmembrane Profiles Across ABC Subfamilies

To characterize the structural organization of ABC transporters across evolutionary lineages, 1,581 sequences from prokaryotic and eukaryotic taxa were annotated using HMM-based domain prediction and classified into architectural classes based on the arrangement of transmembrane domains (TMDs) and nucleotide-binding domains (NBDs). Consensus transmembrane topology was independently assessed using DeepTMHMM and TOPCONS to define membrane-embedded regions (Figure 1; Table S2).

 doi:10.20944/preprints202604.1710.v1

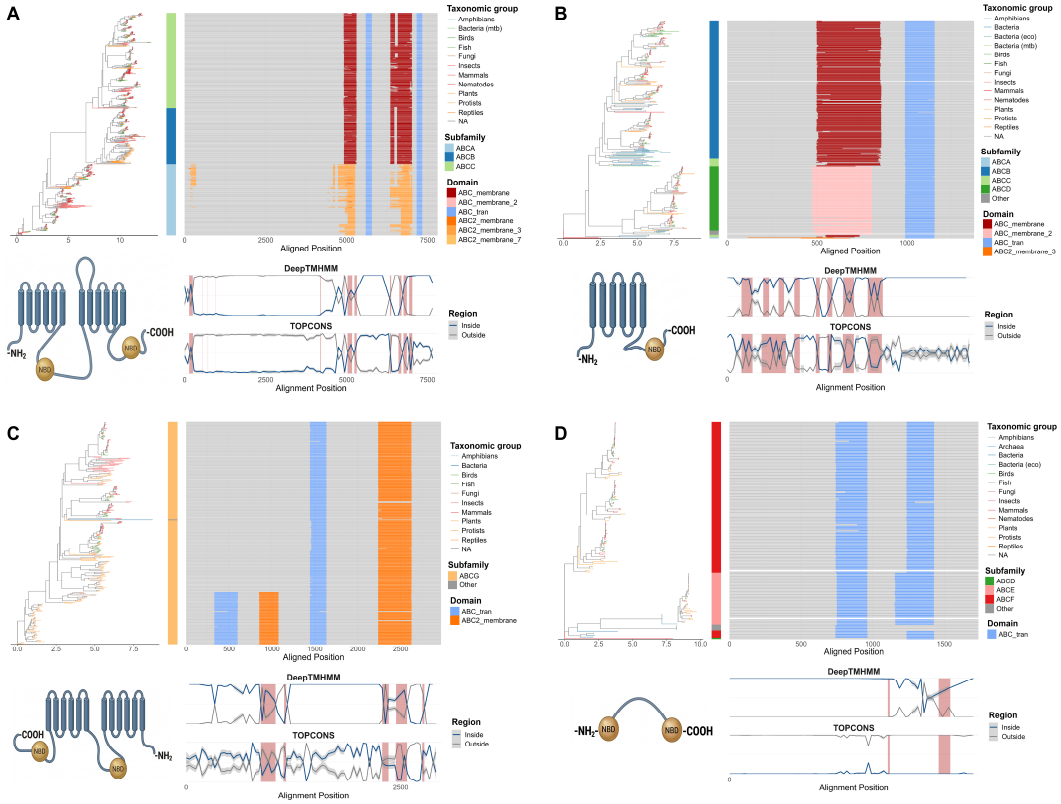

Figure 1. Domain architecture and transmembrane topology across ABC transporter architectural classes. For each panel, sequences are ordered by phylogenetic topology (left), with branch colors indicating taxonomic group and vertical color strips indicating subfamily membership (*see legends*). Main Panel: per-sequence HMMER-annotated domains mapped onto trimmed multiple sequence alignment coordinates; shown as colored horizontal bars (ABC_membrane variants, orange/red tones; ABC_tran, blue; *see Domain legend*). Schematic cartoon (bottom left): topology diagram illustrating membrane orientation, with TMD (blue), NBDs (yellow), and N-/C-termini indicated. Transmembrane topology profiles (bottom right): consensus transmembrane helix occupancy per alignment position from DeepTMHMM (upper trace) and TOPCONS (lower trace); blue line = inside (cytoplasmic) loop frequency; grey line = outside (extracellular/luminal) loop frequency; red shading = positions with >50% transmembrane helix occupancy. (A) Full forward transporters. (B) Half forward transporters. (C) Full and half reverse transporters. (D) NBD-only soluble proteins.

Full forward transporters (TMD–NBD–TMD–NBD; n = 751) were distributed across the ABCA, ABCB, and ABCC subfamilies (Figure 1A). TMD2 showed annotation heterogeneity across subfamily members, with ABC2_membrane, ABC2_membrane_3, and ABC2_membrane_7 variants detected, reflecting the structural diversity in C-terminal TMD configurations in ABCA and ABCC members. A subset of ABCC sequences contained an additional N-terminal TMD0 domain, corresponding to the regulatory transmembrane segment characteristic of long MRP-type transporters, concentrated in mammalian and vertebrate lineages. Topology predictions from DeepTMHMM and TOPCONS showed two concordant transmembrane helix clusters corresponding to TMD1 and TMD2 in the N-terminal and C-terminal thirds of the alignment, respectively. Intervening NBD regions and inter-domain linkers showed minimal transmembrane signal, consistent with cytoplasmic localization. Inside/outside topology profiles further indicated an extracellular N-terminus and cytoplasmic NBDs.

Half forward transporters (TMD–NBD; n = 372) comprising ABCB and ABCD members, were represented as single polypeptides requiring homo- or heterodimerization (Figure 1B). Both predictors identified a single transmembrane helix cluster in the N-terminal half of the alignment, with no membrane association in the C-terminal NBD region, consistent with the canonical six-helix

 doi:10.20944/preprints202604.1710.v1

TMD bundle. Prokaryotic and eukaryotic sequences were phylogenetically intermixed, supporting an early evolutionary origin of the half-transporter configuration.

Full and half reverse transporters (NBD–TMD; n = 69 full, n = 228 half) were restricted to the ABCG subfamily (Figure 1C). In these proteins, ABC_tran domains precede ABC2_membrane domains. DeepTMHMM and TOPCONS profiles confirmed transmembrane helix predictions were shifted toward the C-terminus, with the N-terminal NBD region lacked membrane association. Inside/outside topology profiles indicated a cytoplasmic N-terminus — the inverse of forward transporter topology — consistent with the established architecture of ABCG2 (BCRP) and the obligatory heterodimer ABCG5/G8.

"NBD-only proteins (n = 154), corresponding to ABCE and ABCF, lacked all membrane-associated domain annotations (Figure 1D). Neither predictor identified transmembrane regions at any position, confirming their soluble cytoplasmic nature. Both predictors showed flat, near-zero probability profiles across the full alignment length.

Across all four architectural classes, positions flanking annotated domain boundaries contained extended segments not covered by domain annotation. These non-domain regions, including inter-domain linkers (L1, L2, L3) and N- and C-terminal flanking sequences (Nflank, Cflank), varied substantially in length across sequences and subfamilies. Their structural and evolutionary properties were characterized in subsequent analyses

2.2. Disorder Propensity Across ABC Transporter Classes

To characterize the structural features of ABC transporter sequences beyond annotated domain cores, per-residue intrinsic disorder was predicted across all five architectural classes using AIUPred, and secondary structure propensity was independently predicted using NetSurfP-3.0. Disorder scores were visualized as phylogenetically ordered heatmaps with corresponding mean disorder profiles (Figure 2), and region-specific disorder fractions were quantified at the gene level (Figures 3 and 4). Secondary structure profiles confirmed that annotated NBD and TMD cores show consistently high helix and sheet propensity, while inter-domain linker and flanking regions are coil-dominant across all architectural classes, corroborating the AIUPred predictions (see Supplementary Figure S4).

 doi:10.20944/preprints202604.1710.v1

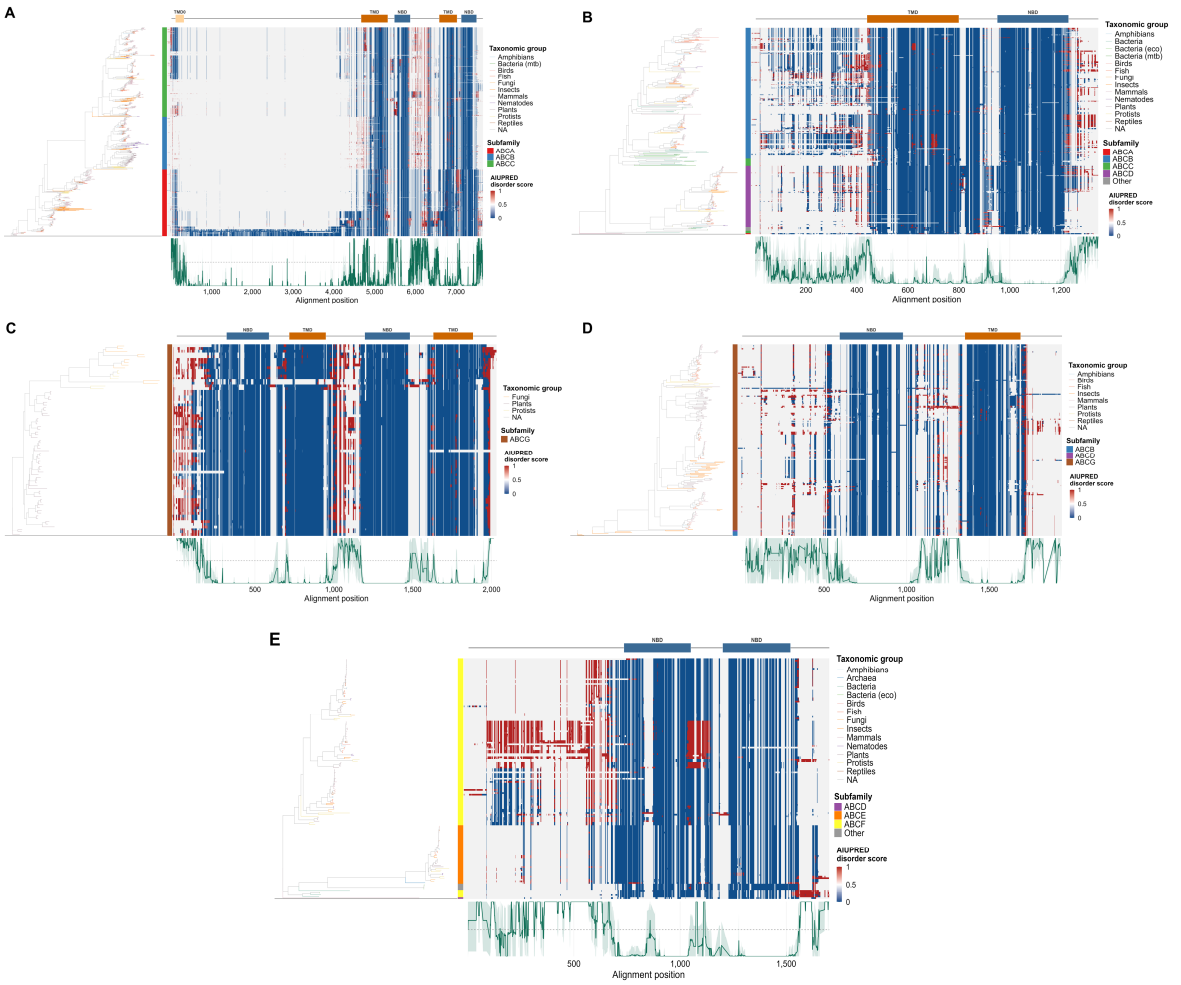

Figure 2. Intrinsic disorder heatmap across ABC transporter architectural classes. For each panel, sequences are ordered by phylogenetic topology. Main panel: per-position AIUPred disorder score mapped onto multiple sequence alignment coordinates; blue = ordered (score 0); red = disordered (score 1); grey = alignment gap; color scale shown at right with 0.5 threshold indicated. Domain annotation track (top): TMD regions (orange) and NBD regions (blue) with boundaries derived from HMMER coordinates. Mean disorder track (bottom): mean AIUPred disorder ± 1 SD per alignment position (dark green line; shaded band); dashed horizontal line = 0.5 disorder threshold. **(A)** Full forward transporters. **(B)** Half forward transporters. **(C)** Full reverse transporters. **(D)** Half reverse transporters. **(E)** NBD-only soluble proteins.

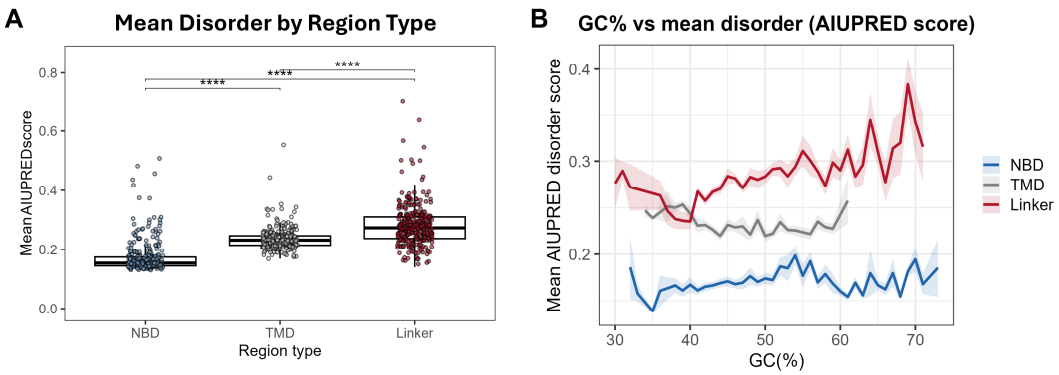

Figure 3. Intrinsic disorder hierarchy across region types and its relationship with genomic GC content. **(A)** Mean AIUPred disorder score per gene-level region grouped by region type; individual data points shown; box plots show median and interquartile range; significance brackets indicate BH-adjusted pairwise Wilcoxon comparisons (all $p < 0.0001$). **(B)** Binned means AIUPred disorder score as a function of GC content (30-75%)

 doi:10.20944/preprints202604.1710.v1

stratified by region type (NBD, blue; TMD, grey; linker, red); shaded bands = 95% CI; This positive GC–disorder relationship in linker regions was preserved under binary thresholding of disorder scores (Supplementary Figure S2), confirming that the trend is not an artefact of the continuous scoring metric.

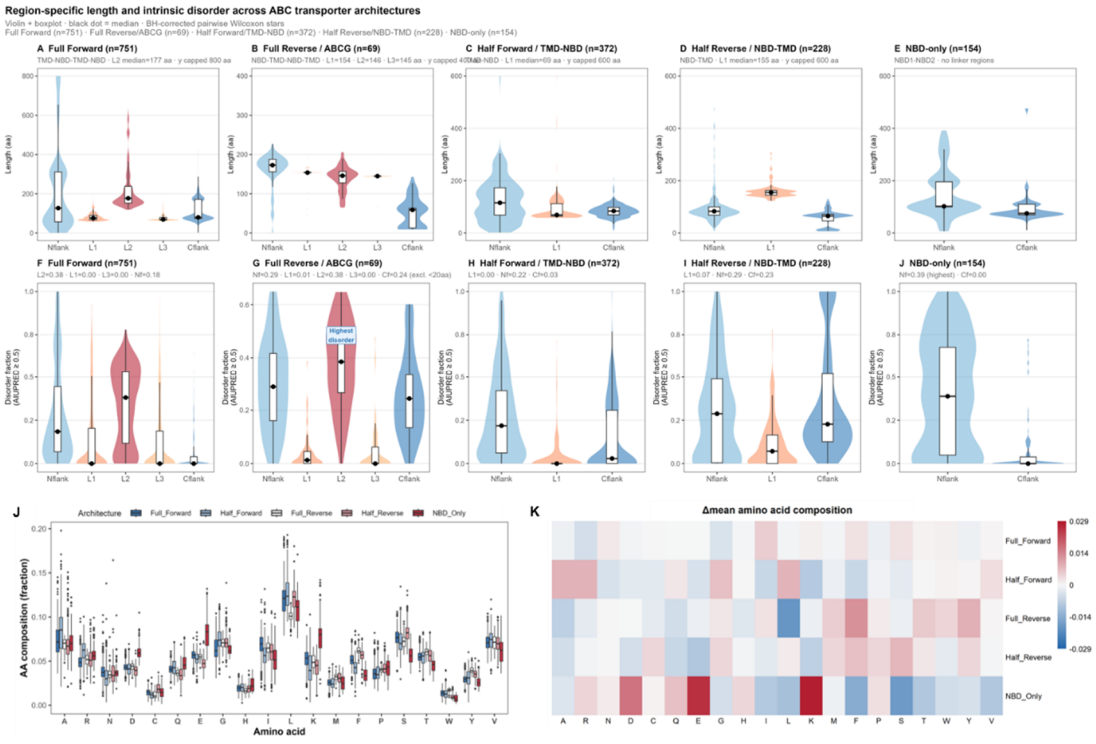

Figure 4. Region-specific length, intrinsic disorder, and amino acid composition of ABC transporter non-canonical regions across architectural classes. Violin plots with embedded box plots; black dot = median; significance brackets = BH-corrected pairwise Wilcoxon comparisons. Pink violins = inter-domain linker regions; blue violins = flanking regions. Top row **(A–E):** region length in amino acids (y-axis capped as indicated). Middle row **(F–J)**: disorder fraction, defined as the proportion of residues with AIUPred score ≥ 0.5. **(A, F)** Full forward transporters (TMD–NBD–TMD–NBD; n = 751); L2 median length 176 aa, median disorder fraction 0.38; L1, L3, and Cflank largely ordered (median ≈ 0.00); Nflank median disorder fraction 0.19. **(B, G)** Full reverse transporters (NBD–TMD–NBD–TMD; ABCG; n = 69); five non-domain regions shown (Nflank, L1, L2, L3, Cflank); L1/L2/L3 median lengths 145–154 aa; disorder distributed across Nflank (0.288), L2 (0.385), and Cflank (0.355); L1 and L3 largely ordered (median ≤ 0.013). **(C, H)** Half forward transporters (TMD–NBD; n = 372); single linker **(L)** median length 155 aa, median disorder fraction 0.07; Nflank median 0.29; Cflank median 0.23. **(D, I)** Half reverse transporters (NBD–TMD; n =228); single linker **(L)** median length 69 aa, median disorder fraction 0.00; Nflank median 0.22; Cflank median 0.03. **(E, J)** NBD-only soluble proteins (n = 154); Nflank carries the highest median disorder fraction across all regions and architectures (median 0.38); Cflank largely ordered (median 0.00). Bottom left: amino acid composition (fraction) per residue across all 20 standard amino acids, stratified by architectural class. Bottom right: Δmean amino acid composition heatmap showing deviation of each architectural class from the global mean per residue (red = enriched; blue = depleted; scale ±0.029); full statistical annotation in Supplementary Figure S1.

In full forward transporters (Figure 2A), disorder scores were elevated at three positional intervals: the N-terminal pre-TMD0 segment in long ABCC members, the L2 linker between NBD1 and TMD2, and C-terminal extensions beyond NBD2. Within ABCC sequences, the TMD0 helix bundle corresponded to a locally ordered interval, whereas the L0 linker and the NBD1-TMD2 regulatory region were highly disordered across most sequences, consistent with crystallographic and cryo-EM data showing that these segments are poorly resolved in structural studies [12,23]. In ABCA and ABCB sequences, disorder was more restricted to the inter-domain linker positions and

 doi:10.20944/preprints202604.1710.v1

terminal flanking regions. Quantification confirmed that L2 had the highest median disorder fraction among all linker segments in full forward transporters (median 0.38), while L1, L3, and Cflank were largely ordered (median approximately 0.00), and Nflank showed moderate disorder (median 0.19; Figure 4A, F). Within the full forward class, notable subfamily variation was observed in L2: ABCC members showed the highest L2 disorder (median 0.526) with median length 158 aa, ABCB showed intermediate disorder (median 0.395, length 158 aa), while ABCA members carried the longest L2 in the dataset (median 303 aa) yet the lowest disorder fraction (median 0.083), suggesting a structured rather than regulatory role for this segment in the ABCA subfamily. This structured character may reflect the conformational constraints imposed by the large extracellular domains (~600 aa each) that are characteristic of ABCA members and are resolved as ordered folds in available cryo-EM structures [24].

Half forward transporters showed a structurally more compact disorder pattern, with elevated disorder in the N-terminal and C-terminal flanking regions (Nflank median 0.29; Cflank median 0.23) and in the NBD–TMD linker (Figure 2B, Figure 4C, H). The single inter-domain linker had a median disorder fraction of 0.07 and median length 155 aa. Within ABCB half transporters, this region showed elevated disorder, consistent with the intracellular loop architecture of P-glycoprotein-related proteins [25] ABCD half transporters displayed a similar profile.

Full reverse transporters showed a distributed disorder pattern across multiple non-domain segments, in contrast to the single dominant disordered linker seen in full forward transporters (Figure 2C, Figure 4B, G). Disorder was elevated in the N-terminal flanking region preceding NBD1 (Nflank median disorder fraction 0.288), in the L2 linker connecting TMD1 to NBD2 (median 0.385, length 146 aa), and in the C-terminal flanking region beyond TMD2 (Cflank median 0.355; note that 36% of full reverse sequences have Cflank lengths below 20 aa, and the disorder fraction for sequences with Cflank ≥ 20 aa is 0.244). The intervening linkers L1 and L3, which connect NBD1 to TMD1 and NBD2 to TMD2 respectively, were largely ordered (median disorder fraction ≤ 0.013; Figure 4G). This multi-segment disorder distribution reflects the inverted NBD–TMD–NBD–TMD topology of ABCG members, in which no single inter-domain gap is long enough to serve as a dedicated regulatory linker. Half reverse transporters showed a simpler profile, with moderate disorder in the N-terminal flank (median 0.218) and near-zero disorder in the single TMD–NBD linker (median 0.000; Figure 2D, Figure 4D, I).

NBD-only proteins lacked transmembrane-flanking disorder by definition, as these classes carry no TM segments. Disorder was localized to the inter-NBD linker and terminal extensions, with the Nflank exhibiting the highest median disorder fraction recorded across all regions and architectures (median 0.388) while the Cflank remained largely ordered (median 0.000; Figure 2E, Figure 4E, J).

Across all architectural classes, linker and flanking regions were significantly more disordered than TMD and NBD cores (BH-adjusted pairwise Wilcoxon, all comparisons $p < 0.0001$; Figure 3B). PGLS models incorporating branch-length-scaled covariance yielded Pagel's λ of 0.955–0.972 across model specifications (Table 1), confirming strong phylogenetic signal in linker disorder and indicating that the observed hierarchy reflects genuine evolutionary constraint rather than phylogenetic non-independence alone.

Table 1. Phylogenetic generalized least squares (PGLS) models of GC content and amino acid composition as predictors of linker intrinsic disorder in ABC transporter genes.

GC predictor	n	λ [a]	R^2	adj. R^2	AIC	β (GC)	SE	t	p-value
M1: GC% only — direct effect of GC content on linker disorder (n = 1581)									
Total GC%	1581	0.955	0.073	0.072	−5800.5	0.0024	0.0002	11.13	$<2\times10^{-16}$ ***
GC1% (1st position)	1581	0.958	0.082	0.081	−5816.6	0.0038	0.0003	11.84	$<2\times10^{-16}$ ***

 doi:10.20944/preprints202604.1710.v1

GC2% (2nd position)		**1581**	**0.963**	**0.164**	**0.164**	**−5966.3**	**0.0078**	**0.0004**	**17.62**	**<2×10⁻¹⁶ *****
GC3% (3rd position)		1581	0.955	0.037	0.036	−5740.2	0.0008	0.0001	7.75	1.6×10⁻¹⁴ ***
M2: GC% + AA composition — residual direct GC effect after controlling for AA composition (n = 1581)										
Total GC%		1581	0.968	0.413	0.412	−6519.2	0.0015	—	—	1.8×10⁻¹⁵ ***
GC1% (1st position)		1581	0.969	0.407	0.406	−6503.8	0.0019	—	—	5.2×10⁻¹² ***
GC2% (2nd position)		**1581**	**0.971**	**0.427**	**0.426**	**−6554.1**	**0.0040**	**—**	**—**	**<2×10⁻¹⁶ *****
GC3% (3rd position)		1581	0.968	0.405	0.404	−6499.1	0.0005	—	—	4.7×10⁻¹¹ ***
M3: AA composition only — baseline model without GC predictor (n = 1581)										
AA composition (IDP index) [b]		1581	0.972	0.391	0.390	−6458.5	—	—	—	—

Outcome variable: mean AIUPred disorder score of ABC transporter linker regions; n = 1581 genes. [a] λ: Pagel's lambda estimated by maximum likelihood. $\lambda = 1$ indicates complete phylogenetic signal (Brownian motion); $\lambda = 0$ indicates phylogenetic independence. All models returned $\lambda = 0.955–0.972$. [b] M3 contains only amino acid composition (IDP index) as a predictor; β, SE, t, and p are not applicable (—). M3 provides the baseline R^2 against which the residual direct effect of GC% in M2 can be assessed (ΔR^2(M2 vs M3) = +0.016–0.037). SE and t-statistic available for M1 only; — indicates not extracted for M2. R^2 values reflect phylogenetically corrected residual variance and are not directly comparable to OLS R^2. *** $p < 0.001$.

The best-fitting PGLS model incorporated amino acid composition principal components with GC2% as predictors (M2: AIC = −6554.1, adj. $R^2 = 0.426$; Table 1), with amino acid composition alone accounting for most of the explained variance (M3: adj. $R^2 = 0.390$; AIC = −6458.5; ΔAIC M2 vs M3 = −95.6). Among the four GC predictors tested individually, GC2% was the strongest single predictor of linker disorder (M1: adj. $R^2 = 0.164$), followed by GC1% (0.081), total GC% (0.072), and GC3% (0.036), This pattern indicates that the GC–disorder association is driven primarily by non-synonymous codon positions via amino acid composition.

Region-specific quantification further resolved these patterns across architectural classes (Figure 4A–J; Supplementary Table S3). In half forward transporters (TMD–NBD; n = 232), the single inter-domain linker showed a median disorder fraction of 0.07, whereas in half reverse transporters (NBD–TMD; n = 228) the linker was largely ordered (median 0.00; Supplementary Table S3). Nflank disorder did not differ significantly across architectural classes in any pairwise comparison (all BH-adjusted $p > 0.05$; Cliff's d −0.148 to +0.128; Supplementary Table S3), indicating that elevated N-terminal flank disorder is largely independent of domain topology.

The amino acid composition of linker regions differed significantly across architectural classes (Figure 4, bottom panels). Full reverse transporters were enriched in lysine and depleted in isoleucine relative to the global mean (permutation FDR $p < 0.001$ for both), while NBD-only proteins showed the most divergent composition overall, with strong enrichment of glutamate and lysine. Both glutamate and lysine are established disorder-promoting residues [26–28], consistent with the

 doi:10.20944/preprints202604.1710.v1

observed architectural differences in disorder. Full statistical annotations per amino acid and architecture are shown in Supplementary Figure S1.

To assess whether the GC content–disorder association reflected a direct effect independent of amino acid composition, formal mediation analysis was performed with the IDP amino acid index as the mediator (Figure 5). Total GC% showed a positive association with mean linker disorder across GC% quartiles (Kruskal-Wallis H = 67.93, p = 1.18 × 10^{-14}; Figure 5B). Spearman correlations between GC content and mean disorder were positive and significant in linker regions across all four codon positions (total GC%: ϱ = 0.206; GC1%: ϱ = 0.222; GC2%: ϱ = 0.141; GC3%: ϱ = 0.177; all p < 0.001; Figure 5C ; full codon-position breakdowns with OLS R^2 and standardized direct effects are shown in Supplementary Figure S3), consistently negative in TMDs (total GC%: ϱ = −0.332, p < 0.001), and weak in NBDs (total GC%: ϱ = 0.057, p < 0.05). Mediation analysis confirmed that the direct effect of total GC% on linker disorder was independent of amino acid composition (c′ = 0.200, 95% CI [0.160, 0.241], p < 0.001; Figure 5A), while the indirect effect through the IDP amino acid index was small (ACME β = 0.031; proportion mediated 13.4%). The consistency of the direct GC effect across both non-synonymous positions (GC1%, GC2%) and synonymous wobble position (GC3%) suggest that the association reflects broader genomic GC environment rather than a codon-position-specific mechanism. Taken together with the PGLS results, these analyses indicate that linker disorder is primarily determined by amino acid composition, with genomic GC content contributing a secondary but statistically significant effect that is region-specific—positive in linkers, negative in TMDs, and minimal in NBDs.

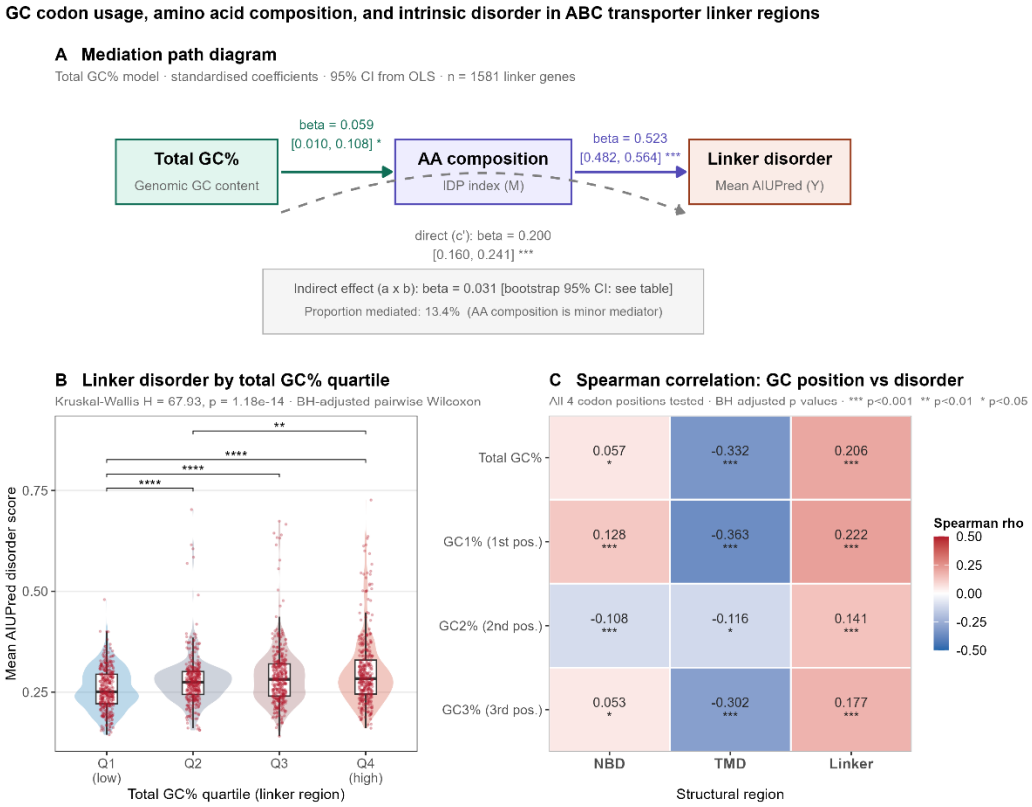

Figure 5. GC codon usage, amino acid composition, and intrinsic disorder in ABC transporter linker regions. **(A)** Mediation path diagram for the total-GC% model (*n* = 1,581 linker genes): total genomic GC content (exposure), IDP index (mediator *M*), and mean AIUPred linker disorder (outcome *Y*); standardised β coefficients with 95% CI from OLS regression and bootstrap mediation (5,000 resamples). **(B)** Mean AIUPred linker disorder score stratified by total GC% quartile; Kruskal–Wallis H = 67.93, p = 1.18 × 10^{-14}; BH-adjusted pairwise Wilcoxon significance shown. **(C)** Spearman rank correlation (ϱ) between GC content at each codon position (total GC%, GC1%, GC2%, GC3%; rows) and mean AIUPred disorder per structural region type (NBD, TMD, linker; columns); BH-adjusted *p*-values annotated.

 doi:10.20944/preprints202604.1710.v1

2.3. PTM Site Distribution and Co-Localization with Disordered Regions

To assess the regulatory potential of disordered regions across ABC transporter subfamilies, predicted PTM sites from MusiteDeep were mapped onto aligned sequence positions and overlaid with disorder scores (Figure 6).

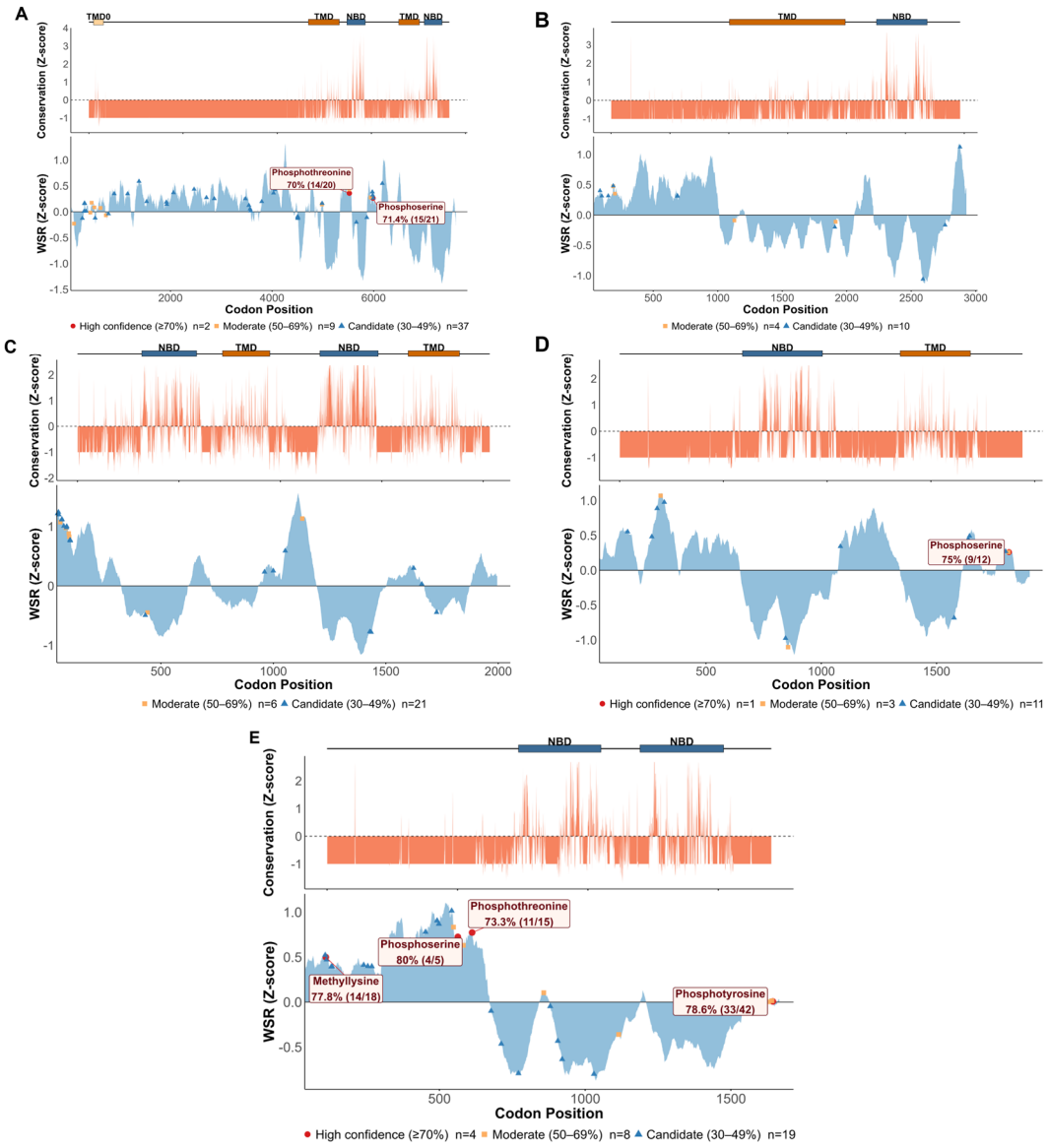

Figure 6. Sequence conservation and site-wise evolutionary rate with conserved PTM site annotation across ABC transporter architectural classes. For each panel, two tracks are shown along the codon alignment position (x-axis). Upper track: AL2CO sequence conservation Z-score; positive values indicate conservation above the alignment mean; dashed line = 0; grey shading = annotated NBD and TMD regions. Lower track: IQ-TREE2 site-wise relative evolutionary rate (WSR) Z-score overlaid with conserved predicted PTM sites from MusiteDeep (green circles; score ≥ 0.5 and cross-species conservancy ≥ 50%); PTM type labelled per site. **(A)** Full forward transporters. **(B)** Full reverse transporters (ABCG). **(C)** Half forward transporters. **(D)** Half reverse transporters. **(E)** NBD-only soluble proteins.

AL2CO conservation Z-scores revealed a strong asymmetric pattern across sequence alignment in all five architectural classes. Positions within annotated NBD regions showed the highest positive conservation (Z-scores up to approximately +4 in full forward transporters (Figure 6A), consistent with the well-documented conserved Walker A, Walker B, and LSGGQ signature motifs [6,7]. TMD regions also exhibited elevated conservation relative to non-domain regions, though at lower

 doi:10.20944/preprints202604.1710.v1

magnitude than NBDs, reflecting greater structural diversity across subfamilies [6]. In contrast, non-domain regions, including the TMD0 region, the N-terminal pre-TMD0 segment, and the NBD1-TMD2 inter-domain region were dominated by negative conservation Z-scores approaching −1, indicating high sequence variability relative to the alignment average. This pattern was consistent across all five architectural classes and is in line with prior observations that inter-domain linkers in multidomain proteins evolve more rapidly than structured domain cores [29].

WSR Z-scores showed a partially inverted pattern relative to conservation. NBD regions exhibited negative WSR Z-scores across all five architectural classes, indicating slow evolutionary rates under strong purifying selection acting to maintain the NBD fold and catalytic residues. In contrast, non-domain and inter-domain regions showed positive WSR Z-scores, with the highest peaks concentrated outside annotated domains. Several non-domain sites exceeded the +2 Z-score threshold, marking them as fast-evolving outliers.

Mapping predicted PTM sites onto the WSR landscape revealed a spatial pattern across architecture. In full forward transporters, predicted phosphoserine, phosphothreonine, methyllysine, and ubiquitination sites co-localized predominantly with regions showing negative WSR Z-scores within the NBD1 and NBD2 blocks, rather than with the most rapidly evolving non-domain positions (Figure 6A). This pattern was also observed in half forward transporters (Figure 6C) and in NBD-only proteins, where phosphothreonine, phosphoserine, methylarginine, and O-linked glycosylation sites mapped to slowly evolving positions near or within the NBD cores (Figure 6D, E). In full reverse transporters, a single phosphoserine site was identified at a conserved slowly evolving position within the NBD-TMD interface (Figure 6B). Half reverse transporters showed a richer set of conserved PTM sites, including phosphoserine, phosphothreonine, O-linked glycosylation, methylarginine, N6-acetyllysine, and phosphotyrosine, with several co-localizing at slowly evolving positions spanning the NBD-TMD boundary (Figure 6D). The preferential localization of predicted PTM sites to slowly evolving positions, rather than at rapidly evolving non-domain regions, is not expected under a neutral model and suggests that a subset of these sites may be maintained by stabilizing selection for their modifiable state across evolutionary timescales.

2.4. PTM Site Distribution, Conservancy, and Enrichment

To characterize the PTM landscape more broadly and identify sites showing cross-species conservation, predicted PTM sites were mapped across all architectural classes as a function of alignment position and scored for modification type enrichment and cross-species conservancy (Figures 7 and 8).

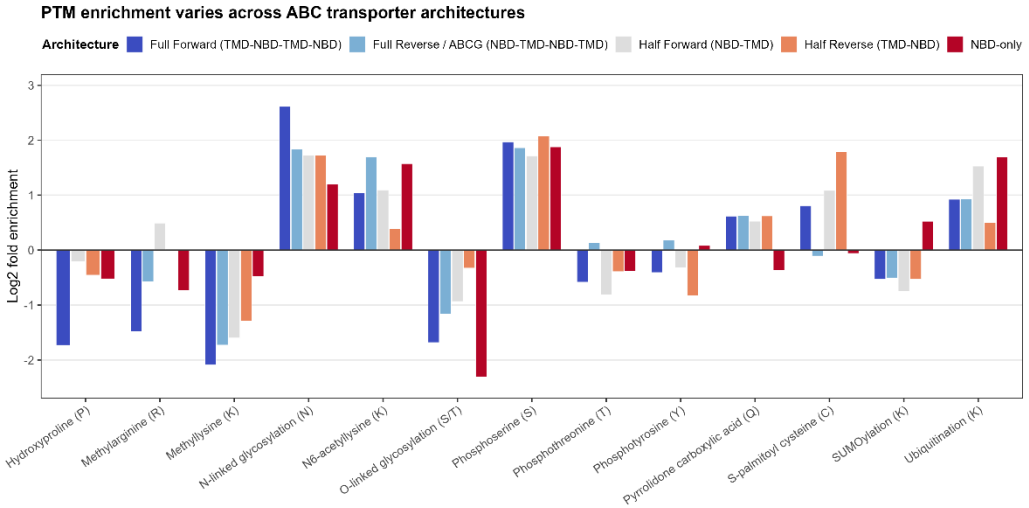

Figure 7. Per PTM type enrichment across ABC transporter architectural classes. Log_2 fold enrichment (Log_2FE) per modification type per architectural class, calculated relative to the expected rate given residue composition. Positive values indicate enrichment; negative values indicate depletion. Modification types shown:

 doi:10.20944/preprints202604.1710.v1

phosphoserine (pSer), phosphothreonine (pThr), phosphotyrosine (pTyr), N-linked glycosylation (N-glyc), O-linked glycosylation (O-glyc), ubiquitination (Ub), N6-acetyllysine (AcK), methylarginine (MeR), methyllysine (MeK). Bars represent mean Log_2FE; error bars = 95% CI.

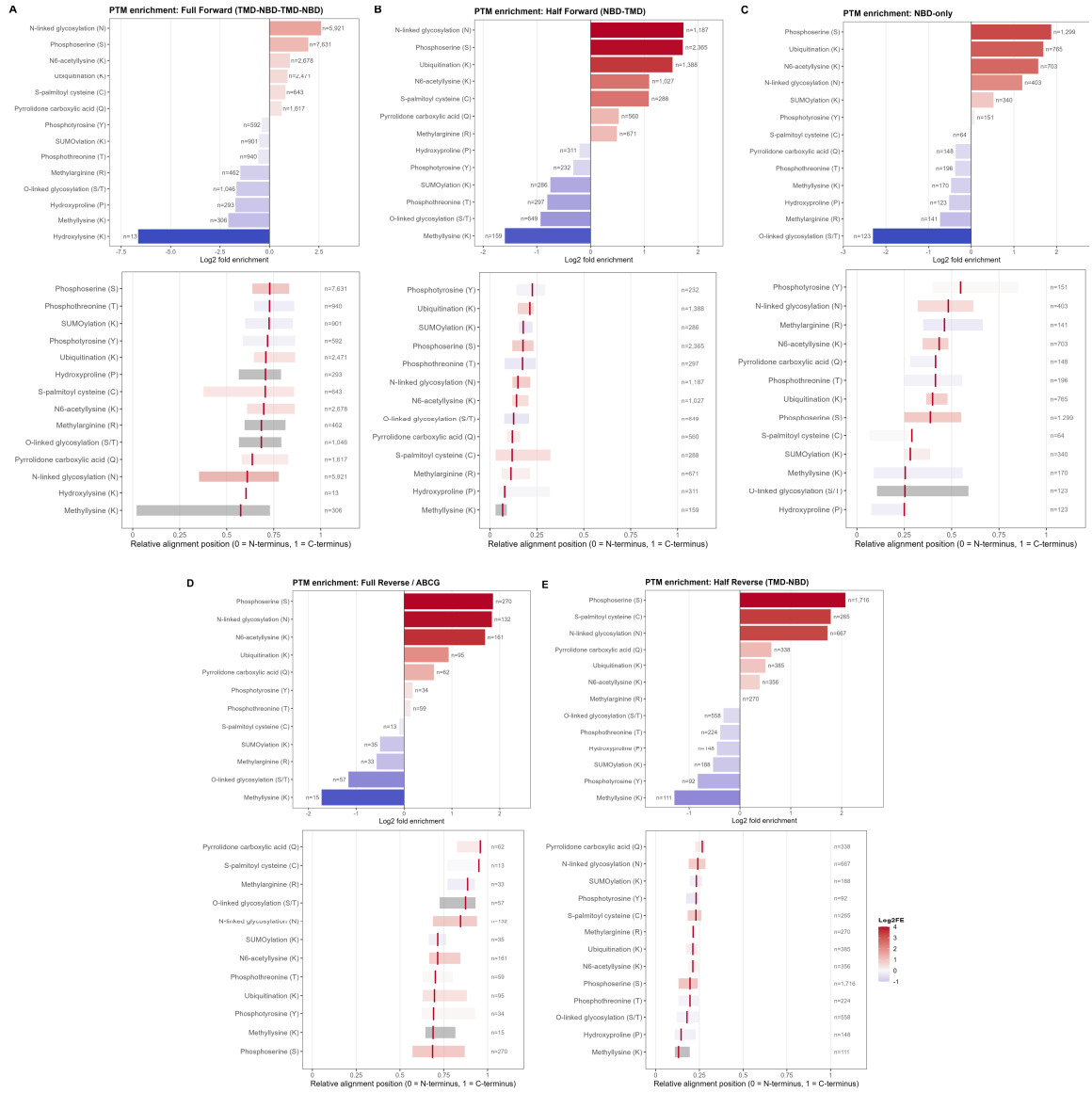

Figure 8. Log_2FE PTM enrichment and distribution across architectural classes and subfamilies. Additional breakdown of PTM distribution by subfamily and region type, with conserved sites (MusiteDeep score ≥ 0.5 and cross-species conservancy ≥ 50%) highlighted.

Predicted PTM sites were enriched in disordered non-domain regions across all architectural classes, with phosphoserine and phosphothreonine showing the highest predicted frequencies in full forward, half forward, and ABCG configurations. Enrichment analysis (Figure 8) further indicated that N-linked glycosylation and phosphoserine exhibited the strongest positive Log_2 fold-enrichment (Log_2FE), consistent with the established over-representation of phosphorylation and N-glycosylation sites in disordered protein regions. In full forward transporters, high predicted PTM scores were disproportionately concentrated in the N-terminal pre-TMD0 region of long ABCC members, the NBD1-TMD2 linker, and the C-terminal tail beyond NBD2. In ABCA and ABCB sequences, predicted PTM sites were less dense but showed a similar positional bias toward non-domain regions.

In full reverse (ABCG) transporters, predicted PTM density was more broadly distributed along the aligned length compared to full forward transporters. This likely reflects the proportionally

 doi:10.20944/preprints202604.1710.v1

smaller domain core in ABCG members, which exposes a larger fraction of the total sequence to modification-permissive disordered regions[8,9]. Taxonomic variation within the ABCG PTM landscape was apparent: fungal sequences showed PTM-dense patches at positions distinct from those in plant sequences, suggesting that the positional distribution of predicted regulatory sites has diverged across kingdoms within the same subfamily.

A total of 140 predicted PTM sites satisfied a tiered conservation criterion (MusiteDeep score ≥ 0.5; cross-species conservancy ≥ 30%), stratified as 7 high-confidence sites (conservancy ≥ 70%), 33 moderate-confidence sites (conservancy 50–70%), and 100 lower-confidence candidate sites (conservancy 30–50%; Supplementary Table S3). High-confidence sites were identified in full forward (n = 2), half reverse (n = 1), and NBD-only (n = 4) architectures. Among the most conserved predictions were a phosphotyrosine site at NBD-only alignment position 1644, corresponding to ABCE1 Y594 (conservancy 78.6%, score 0.906), and a methyllysine site at position 110 (conservancy 77.8%, score 0.743). These represent previously unreported regulatory modifications in NBD-only proteins not annotated in UniProt or PhosphoSitePlus.

2.5. Site-Specific Selection Pressure Across Architectural Classes

FEL analysis revealed pervasive purifying selection across both domain and non-domain positions in all five architectural classes, with the strongest purifying signal concentrated within annotated NBD regions, where dN/dS values were most negative (Figure 9). This pattern reflects the functional constraint on ATP-binding residues and was consistent across architectures. A smaller subset of sites showed positive FEL estimates, indicative of diversifying selection at specific codon positions. These sites were concentrated at inter-domain boundaries rather than uniformly distributed throughout the sequence. In full forward transporters, FEL-positive sites were identified at positions 30, 50, 165, 278, 953, 1169, and 1999, with β values ranging from 0.403 to 2.045 (Figure 9A). In full reverse transporters, FEL-positive sites clustered near the N-terminal pre-NBD region and the NBD-TMD interface (Figure 9C).

 doi:10.20944/preprints202604.1710.v1

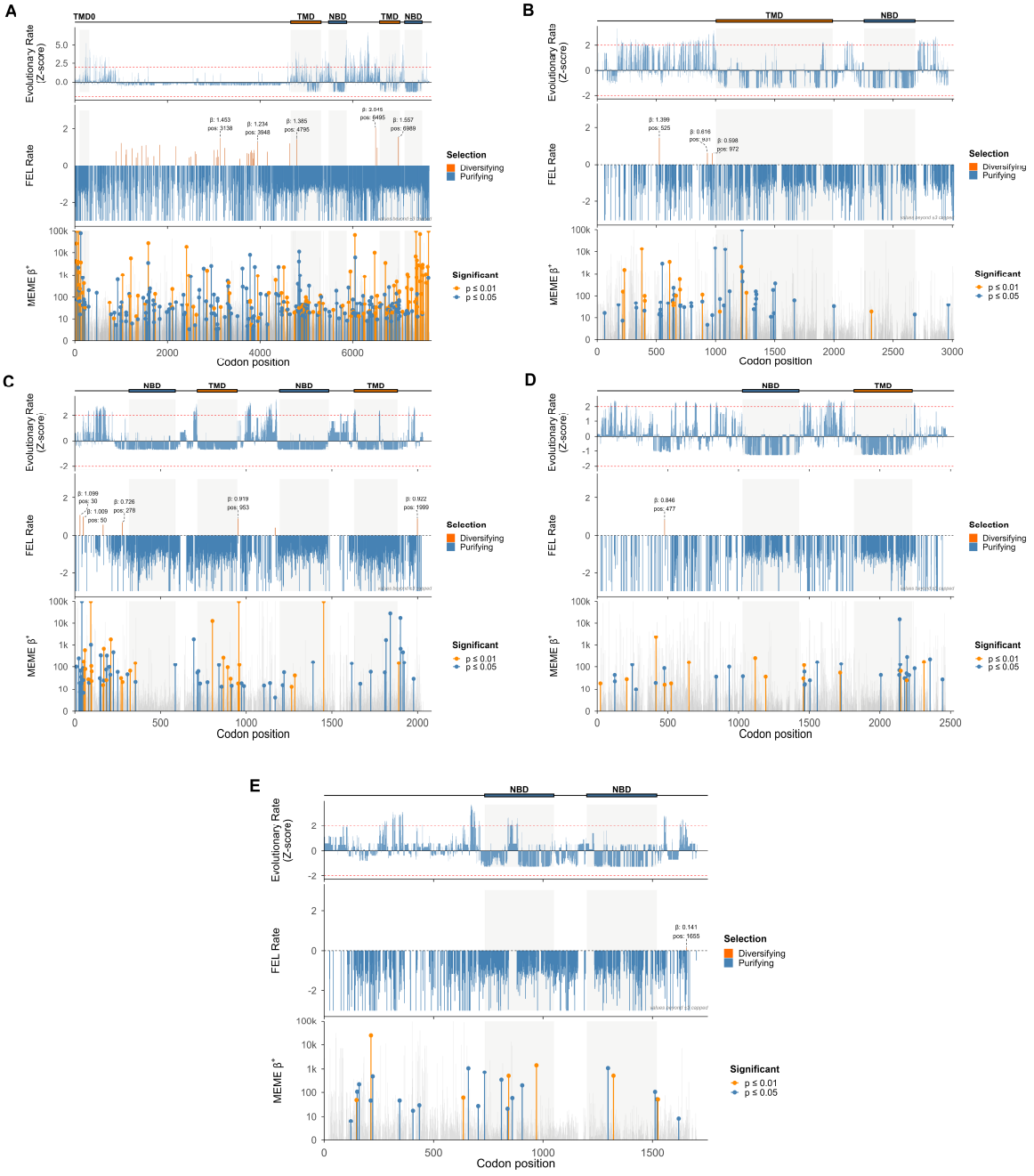

Figure 9. Site-specific evolutionary rates and selection pressure across ABC transporter architectural classes. For each panel, three tracks are shown along the codon alignment position (x-axis); grey shading = annotated domain regions (NBD and TMD). Top track: IQ-TREE2 site-wise relative evolutionary rate Z-score; red dashed lines = ±2 Z-score threshold. Middle track: Fixed Effects Likelihood (FEL) rate per codon position; orange bars = diversifying selection (dN > dS); blue bars = purifying selection (dN < dS); labelled sites indicate highest-βFEL-positive positions. Bottom track: MEME episodic positive selection β+ values (log scale); orange circles = p ≤ 0.01; blue circles = p ≤ 0.05; grey stems = non-significant sites. **(A)** Full forward transporters. **(B)** Half forward transporters. **(C)** Full reverse transporters (ABCG). **(D)** Half reverse transporters. **(E)** NBD-only soluble proteins.

In half forward transporters, three FEL-positive sites were identified (positions 525, 931, 972; Figure 9B), all in non-domain regions flanking the TMD. Half reverse transporters showed a single FEL-positive site (position 477, β = 0.846; Figure 9D), while NBD-only proteins exhibited minimal diversifying signal (one site, position 1655, β = 0.141; Figure 9E), consistent with the strong functional constraint in soluble NBD-only proteins.

 doi:10.20944/preprints202604.1710.v1

MEME analysis identified episodic positive selection at a larger number of sites than FEL in all architectural classes, consistent with MEME's greater sensitivity for detecting selection acting in only a subset of lineages rather than across the full tree[30]. The highest β+ values were observed at non-domain positions, with several sites exceeding β+ = 100 and a subset approaching 10^4-10^5 in full forward and full reverse transporters, indicating rare but highly accelerated substitution in specific lineages. In full forward transporters, significant MEME sites ($p \leq 0.01$) were concentrated in the N-terminal pre-TMD0 region and at positions flanking NBD1 and NBD2 (Figure 9A). In full reverse transporters, significant MEME sites were distributed across both domain-flanking and non-domain positions, with the highest β+ values in the N-terminal pre-NBD region (Figure 9C). In NBD-only proteins, MEME-positive sites were mainly located in the N-terminal flank and inter-NBD linker, flanking rather than within the conserved NBD cores (Figure 9E).

Evolutionary rate profiles inferred using IQ-TREE2 corroborated the FEL and MEME results across all five classes: NBD cores showed compressed, near-zero or negative WSR Z-scores, while non-domain regions exhibited broader variance, with multiple positions exceeding the +2 threshold (Figure 9, top tracks). No positions outside NBD cores showed Z-scores below the −2 threshold, suggesting a lack of unusually constrained non-domain sites at the resolution of these datasets.

Together, these analyses show that the evolutionary landscape of ABC transporters is dominated by strong purifying selection within domain cores across all architectural classes, with episodic diversification concentrated at inter-domain boundaries. The enrichment of high β+ MEME sites and positive FEL outliers in non-domain regions supports the view that these regions are not neutral spacer sequence but are subject to lineage-specific adaptive pressures that vary across architectures.

3. Discussion

3.1. Non-Domain Regions Are Structurally and Evolutionarily Distinct from Domain Cores

The central finding of this study is that non-domain regions of ABC transporters, including inter-domain linkers, N-terminal flanks, and C-terminal extensions, are not passive structural connectors but exhibit distinct structural, compositional, and evolutionary signatures that vary systematically across architectural classes. Linker disorder was significantly higher than that of NBDs and TMDs across all five architectures ($p < 0.0001$, BH-adjusted Wilcoxon), and this hierarchy remained robust after phylogenetic correction (Pagel's λ = 0.955–0.972 across PGLS models), indicating that the observed pattern reflects genuine evolutionary constraint rather than an artifact of phylogenetic non-independence. These results are consistent with broader observations that IDRs are functionally active elements whose properties are maintained by selection rather than arising solely from relaxed structural constraints [18,22].

The strong phylogenetic signal in linker disorder is itself notable. It implies that closely related sequences tend to exhibit similar linker disorder profiles, a pattern inconsistent with neutral evolution of linker composition. The high λ values across all three PGLS models suggest that the amino acid identity of linker regions has been inherited and conserved along lineages in a manner consistent with Brownian motion under stabilizing constraint[31,32]. This parallels observations in other IDR-containing proteins, where functional properties are conserved at the level of biophysical features rather than primary sequence[21,22].

It should be noted that the disorder predictions reported here rely on a single algorithm (AIUPred); although AIUPred integrates energy estimation with deep learning and was cross-validated by NetSurfP-3.0 secondary structure predictions (Supplementary Figure S4), consensus approaches incorporating multiple disorder predictors may yield quantitatively different disorder fraction estimates for individual sequences.

3.2. L2 Emerges as the Primary Regulatory Linker in Full Forward Transporters

Among all inter-domain segments, the L2 linker connecting NBD1 to TMD2 showed the highest median length (176 aa) and the highest median disorder fraction (0.38) in full forward transporters,

 doi:10.20944/preprints202604.1710.v1

while L1, L3, and Cflank were largely ordered. This positional specificity is consistent with experimental evidence: L2 corresponds to the regulatory (R) domain of CFTR (ABCC7) and related ABCC members, where extensive phosphorylation of disordered serine residues modulates NBD dimerization and channel gating[12,13]. Recent cryo-EM structures of phosphorylated Ycf1p, a yeast ABCC homolog, directly resolved the L2/R-domain bridging NBD1 and NBD2 in a transport-competent conformation, demonstrating that its disordered character is functionally coupled to the conformational transport cycle[23]. Our computational results extend this observation from individual well-studied transporters to the full forward transporter class, suggesting that elevated L2 disorder and PTM enrichment represent a general organizational feature of ABCA, ABCB, and ABCC transporters rather than a property unique to CFTR. The co-localization of predicted phosphoserine, phosphothreonine, ubiquitination, and methyllysine sites with slowly evolving positions within the NBD1 and NBD2 further suggests that a subset of these regulatory sites is maintained by stabilizing selection, consistent with the constitutive phosphorylation of ABCA1 at Ser2054 required for cholesterol efflux activity[17,24,33].

3.3. Genomic GC Content Influences Linker Disorder Primarily Through Amino Acid Composition

PGLS modelling across all four codon positions indicates that the GC–disorder association in ABC transporter linker regions is driven primarily by non-synonymous sites. GC2% (second codon position) was the strongest single GC predictor of linker disorder (adj. $R^2 = 0.164$; Table 1), followed by GC1% (adj. $R^2 = 0.081$), total GC% (adj. $R^2 = 0.072$), and GC3% at the synonymous wobble position (adj. $R^2 = 0.036$). This ranking is mechanistically interpretable: the second and first codon positions directly determine amino acid identity, and high GC at these positions biases the encoded amino acid composition toward disorder-promoting residues, as codons for alanine, glycine, and proline are GC-rich [27,28]. That GC2% is the strongest predictor is consistent with the established role of the second codon position in determining amino acid hydrophobicity: U at the second position preferentially encodes hydrophobic residues (Leu, Ile, Val, Phe, Met), whereas C and G at this position encode disorder-promoting residues including Ala, Pro, Arg, and Gly [34]. Amino acid composition alone explained 39.1% of variance in linker disorder (M3; Table 1), substantially exceeding any single GC predictor. This confirms that the GC–disorder association operates largely through the amino acid composition pathway, as proposed previously by Peng et al. [27] and Basile et al. [28]

Nevertheless, all four GC predictors retained statistically significant partial effects after controlling for amino acid composition (all $p < 0.001$ in M2 models; Table 1), and the combined model (GC2% + AA composition PCs; adj. $R^2 = 0.426$, AIC = −6554.1) explained substantially more variance than amino acid composition alone (ΔAIC = −95.6 relative to M3). The residual GC effect after composition control ($\Delta R^2 = 0.016$–0.037 across predictor variants) is consistent with the mediation analysis, which indicated partial mediation (total GC% proportion mediated 13.4%; direct effect $c' = 0.200$, $p < 0.001$). The GC–disorder relationship was strongly region-specific across all four codon positions: positive in linker regions, negative in TMDs, and minimal in NBDs, suggesting that genome-wide analyses pooling region types would substantially obscure this association.

3.4. Domain Boundaries Are the Sites of Structural State Lability

GLOOME-inferred structural transition rates, which quantify disorder-to-order gain and loss events across lineages, were highest at domain boundaries rather than within linker cores. This finding challenges the view that linker regions are the primary sites of structural evolutionary change. Instead, the data suggest that linker cores tend to maintain a relatively stable disordered state, while regions at the interface between structured domains and flanking disordered segments undergo more frequent structural remodeling. This pattern is consistent with the concept of molecular recognition features (MoRFs), which are short segments within IDRs that can undergo coupled folding upon binding to interaction partners [18,35]. Domain boundaries in ABC transporters may therefore function as evolutionarily labile MoRF-like elements whose structural state is subject to lineage-specific selection, while the bulk of linker regions maintains a stable

 doi:10.20944/preprints202604.1710.v1

disordered character. This interpretation is consistent with structural evidence that the intracellular coupling helices at the TMD-NBD interface of ABCB and ABCC transporters undergo substantial conformational reorganization during the transport cycle [2]. A caveat is that the alignment-dependent analyses underlying these observations (GLOOME, FEL, MEME, AL2CO) are sensitive to alignment quality, and highly divergent non-domain regions may be poorly aligned despite ClipKIT trimming, potentially inflating evolutionary rate estimates at boundary positions. Independent validation using structure-aware alignment methods or pairwise dN/dS approaches would help to confirm these patterns.

3.5. PTM Distribution Reflects Both Disorder Enrichment and Positional Conservation

Predicted PTM sites were enriched in disordered non-domain segments across all architectural classes, with phosphoserine and phosphothreonine predominating in full forward, half forward, and ABCG configurations. This enrichment is consistent with the well-established over-representation of phosphorylation sites in IDRs [36,37] and with experimental phosphoproteomic data from CFTR, ABCA1, MRP1, and Ycf1p demonstrating that functionally critical phosphorylation events occur in disordered linker regions [12,13,23]. The unexpected finding was that several predicted PTM sites co-localized with slowly evolving positions within NBD cores, regions under strong purifying selection. This suggests a subset of regulatory modification sites are deeply conserved and maintained by stabilizing selection, distinct from the broader population of PTM sites located in rapidly evolving non-domain regions. A similar two-tier organization of PTM sites, with constitutively conserved regulatory residues embedded within more rapidly evolving disordered backgrounds, has been proposed for other multidomain signaling proteins [22,38]. The presence of O-linked glycosylation, methylarginine, and N6-acetyllysine sites in the half forward and NBD-only classes expands the predicted regulatory PTM landscape beyond phosphorylation in these less-studied architectural classes and warrants experimental validation. All PTM sites reported here are computational predictions from MusiteDeep; cross-referencing against experimentally annotated sites in UniProt Swiss-Prot for human ABC transporters confirmed recovery of the majority of available reference sites, though experimental PTM coverage in UniProt is sparse for most subfamilies and concentrated in well-characterised members such as CFTR and ABCG1 (Supplementari Table S6). Novel candidate sites — particularly those in non-human sequences or understudied architectural classes — require experimental validation by phosphoproteomics or site-directed mutagenesis before functional claims can be made.

3.6. Implications for Linker-Targeted Drug Discovery

Non-domain regions of ABC transporters represent a largely underexplored space for functional modulation. Current pharmacological strategies predominantly target the structurally well-characterised NBD and TMD cores through ATP-competitive or substrate-competitive inhibition, which raises selectivity concerns given the high conservation of these sites across family members. The disordered, PTM-enriched linker regions identified here offer a mechanistically distinct alternative: these segments regulate transporter activity through phosphorylation-dependent conformational changes and protein–protein interactions rather than through direct catalysis and are therefore amenable to modulation that does not compete with substrates or nucleotides.

That such regulation is functionally consequential is well established. Linker mutations in the CFTR R-domain and the MRP1 L0 linker cause disease by disrupting phosphorylation-dependent gating [12–14], demonstrating that these segments are essential regulatory elements. The clinical success of CFTR modulators further illustrates the therapeutic potential of inter-domain communication: the corrector VX-809 (lumacaftor) acts on transmembrane domain 1 and propagates allosteric stabilisation to the NBD1 interface, rescuing folding defects caused by the F508del mutation [39]. This allosteric principle extends beyond CFTR — extracellular non-domain mutations in MRP1 couple to transmembrane conformational changes [40], establishing that perturbations in non-domain segments can propagate functional effects across the full transporter architecture. More

 doi:10.20944/preprints202604.1710.v1

recently, Heinkel et al. [41]demonstrated that the disordered, phosphothreonine-rich linker of the *M. tuberculosis* transporter Rv1747 undergoes phosphorylation-dependent liquid–liquid phase separation, regulated by multiple Ser/Thr kinases, suggesting that condensate formation in non-domain regions may represent a regulatory mechanism in bacterial ABC transporters. A broader perspective on disorder-targeted drug design is emerging: computational strategies for targeting IDPs/IDRs include fragment-based mapping to identify transient pockets in disordered ensembles [42], approaches to modulate biomolecular condensate formation as a therapeutic strategy [43], and the design of conformationally adaptive therapeutic peptides that match the dynamic ensembles of disordered target regions [44]. More broadly, the taxonomic composition of the dataset is weighted toward mammals (~40%), and some architectural classes have limited representation (full reverse, n = 69), which constrains the statistical power of class-specific conclusions and may not fully capture the diversity of linker architectures in undersampled lineages. The disorder and PTM maps presented here therefore constitute a prioritisation resource for guiding future experimental and structural studies rather than definitive evidence of regulatory function or therapeutic druggability.

4. Materials and Methods

4.1. Sequence Retrieval, Quality Filtering, and Architecture Classification

A total of 1,581 ABC transporter protein sequences were retrieved from the KEGG database [45]using a custom Python script interfacing with the KEGG REST API, and querying by KEGG Orthology (KO) codes. The database included prokaryotes (*E. coli, M. tuberculosis, P. aeruginosa, V. cholerae, S. aureus*), fungi (*S. cerevisiae, C. albicans*), plants (*A. thaliana, Z. mays, O. sativa*), invertebrates (*D. melanogaster, C. elegans*), and vertebrates across five classes. Exact duplicate sequences were removed using SeqKit v2.8.0 rmdup --by-seq [46], and sequences containing non-standard residues or truncated open reading frames were filtered using HyPhy v2.5 [47]. Domain boundaries were annotated using HMMER v3.4 against Pfam v36.0 [48,49], searching for ABC_tran (NBD), ABC_membrane, ABC_membrane_2, ABC2_membrane, ABC2_membrane_3, and ABC2_membrane_7 profiles at E-value ≤ 1×10^{-3}. A custom Python script extracted domain boundary coordinates and classified sequences into five architectural classes based on domain count and order, following established structural taxonomy of Wilkens [25] and Dean & Annilo[1]: Full_Forward (TMD–NBD–TMD–NBD; n = 751), Full_Reverse (NBD–TMD–NBD–TMD, ABCG; n = 69), Half_Forward (TMD–NBD; n = 372), Half_Reverse (NBD–TMD; n = 228), and NBD_only (NBD–NBD, ABCE/ABCF; n = 154). The full reverse class comprises plant (n = 53), fungal (n = 9), and protist (n = 6) sequences corresponding to full-size ABCG transporters of the pleiotropic drug resistance (PDR) type, which are expressed as single-chain NBD–TMD–NBD–TMD polypeptides [50–53]. Structural regions were defined as N-terminal flank (Nflank), inter-domain linkers (L1–L3, minimum gap ≥ 10 residues), and C-terminal flank (Cflank). Organism information list and KO codes are provided in Table S1, alignment statistics per class are provided in Table S2.

4.2. Sequence Alignment, Phylogenetic Inference, and Transmembrane Topology

Protein sequences per architectural class were aligned using MAFFT v7.520 L-INS-i [54] and trimmed with ClipKIT v1.3.0 kpic-smart-gap mode [55]. Codon-aware nucleotide alignments were generated using Pal2Nal [56]. Phylogenetic trees were inferred with IQ-TREE2 using ModelFinder Plus (MFP) for automatic model selection and ultrafast bootstrap approximation (1,000 replicates)[57,58]. Trees served as input for downstream analyses, and for ordering sequences in heatmap visualizations. Transmembrane topology was predicted per sequence using DeepTMHMM v1.0.24[59] and validated with TOPCONS [60]. Consensus TM frequency profiles were constructed by mapping predicted TM positions onto alignment coordinates and computing positional occupancy frequencies.

 doi:10.20944/preprints202604.1710.v1

4.3. Structural Characterisation: Disorder, Secondary Structure, and Region-Specific Analysis

Per-residue intrinsic disorder scores were predicted using AIUPred [61], which integrates energy estimation with deep learning (score range 0–1; threshold 0.5 for disorder classification). Secondary structure propensities (helix, sheet, coil) were predicted using NetSurfP-3.0 [62]. All predictions were performed on unaligned sequences and subsequently mapped onto trimmed alignment coordinates per architectural class for heatmap visualization. Heatmaps were constructed in R v4.4.2 using a custom pipeline with ggtree, ggplot2, and patchwork [63,64]. Disorder fraction and region length distributions were compared using Kruskal-Wallis tests with Benjamini-Hochberg-corrected pairwise Wilcoxon tests (rstatix package). Cross-architecture comparisons were performed using Mann-Whitney U tests with Cliff's delta as a non-parametric effect size measure.

4.4. Post-Translational Modification Prediction and Enrichment

PTM sites were predicted across all sequences using MusiteDeep [65], a deep learning framework trained on experimentally verified data from UniProt and PhosphoSitePlus. Predictions were generated for 14 modification types, including phosphoserine, phosphothreonine, phosphotyrosine, N6-acetyllysine, methylarginine, methyllysine, N-linked and O-linked glycosylation, S-palmitoyl cysteine, SUMOylation, ubiquitination, hydroxyproline, hydroxylysine, and pyrrolidone carboxylic acid. Sites were retained at score ≥ 0.5 and verified for chemical validity (residue–modification pairing).

Conservancy (%) = (n_{passing} / $n_{\text{with-residue}}$) × 100

where n_{passing} is the number of sequences with a chemically valid prediction scoring ≥ 0.5 at that column and $n_{\text{with-residue}}$ is the number of sequences carrying a non-gap residue, following the gap-correction principle of Valdar [66] and Capra & Singh [67]. Three filters were applied before tier assignment: (i) $n_{\text{with-residue}} \geq \max(5, \lceil 0.02 \times n_{\text{matched}} \rceil)$, where n_{matched} is the number of sequences present in both the alignment and the MusiteDeep output; (ii) $n_{\text{passing}} \geq 3$; and (iii) the score and chemical-validity filters above. Sites passing all filters were stratified into three tiers: high-confidence (≥ 70%), moderate-confidence (50–69%), and candidate (30–49%), following the stratified PTM conservation approach of Minguez et al. [68].

Log_2 fold enrichment (Log_2FE) per modification type per architectural class was calculated as:

$$\textbf{\textit{Log}}_2\textbf{\textit{FE(t)}} = \log_2([n\_sites(t) / n\_target(t)] / [n\_total\_sites / n\_total\_residues])$$

where *n_target(t)* represents the number of chemically eligible residues across all sequences in the architecture. Conserved PTM sites were identified using dual criteria: MusiteDeep score ≥0.5 and cross-species conservancy ≥ 50%. Computational predictions were benchmarked against 27 experimentally verified PTM annotations from UniProt's Swiss-Prot for 47 human ABC transporters, with a positional tolerance of ±2 residues (Supplementary Table S5).

4.5. Amino Acid Composition, IDP Index, GC Content, and Mediation Analysis

Per-gene amino acid composition was computed as the fractional residue frequency for each structural region. A net intrinsic disorder propensity (IDP) index was derived as:

$$\textbf{\textit{IDP index}} = \Sigma f(\text{disorder-promoting}) - \Sigma f(\text{disorder-inhibiting})$$

where disorder-promoting residues were {E, K, R, S, Q, D, P, G, T, N} and disorder-inhibiting residues were {I, L, V, F, W, Y, C, M}, following Dunker et al. [69] and Uversky et al. [70]. Architecture-specific compositional deviations (Δmean) were assessed using per-cell permutation testing (10,000 iterations, Benjamini-Hochberg FDR across 100 cells). GC content was calculated per gene per region at each codon position (total GC%, GC1%, GC2%, GC3%) and correlated with mean AIUPred disorder using Spearman rank correlation per region type. Causal mediation analysis was conducted on both canonical and non-canonical regions (n = 1581 genes) with GC% as the predictor, IDP index as the mediator, and mean AIUPred disorder as the outcome. The analysis followed the Baron and Kenny stepwise regression (mediation package v4.5.0, 5,000 bootstrap resamples) [71] and confirmed

 doi:10.20944/preprints202604.1710.v1

by structural equation modelling in lavaan v0.6-21 [72] (1,000 bootstrap resamples; model fit assessed by CFI, RMSEA, SRMR).

4.6. Phylogenetic Generalized Least Squares

To account for phylogenetic non-independence, three PGLS models were fitted using the caper R package [73] with Pagel's λ optimized by maximum likelihood [31]. Three model families were evaluated for each of four GC predictors (total GC%, GC1%, GC2%, GC3%): M1 (GC alone), M2 (GC + amino acid composition principal components), and M3 (composition PCs only, providing a baseline for ΔAIC and ΔR^2 assessment). The outcome was mean AIUPred linker disorder (n = 1,581). Pagel's λ ranged from 0.955 to 0.972 across all models, confirming strong phylogenetic structuring of linker disorder. GC2% was the strongest single predictor (M1: adj. $R^2 = 0.164$) and yielded the best combined model (M2: adj. $R^2 = 0.426$, AIC = −6554.1). Full results for all twelve model fits are reported in Table 1.

4.7. GLOOME Structural State Transition Analysis

Per-site binary disorder-state assignments (0 = ordered, 1 = disordered; AIUPred threshold 0.5; gaps encoded as?) were formatted as phyletic-pattern FASTA files per architectural class. GLOOME (gainLoss executable) [74] was applied with IQ-TREE2 phylogenies supplied in Newick format. Analyses used a mixture model for gain and loss rates (_gainLossDist 1), GENERAL_GAMMA_PLUS_INV distributions for gain and loss (_gainDistributionType, _lossDistributionType), three categories each, joint maximum-likelihood optimization of parameters and branch lengths (_performOptimizations 1, _performOptimizationsBBL 1), at the mid-optimization level. Site-specific gain/loss expectations from PosteriorExpectationOfChange.txt were standardized to Z-scores; sites with $Z > +2$ were classified as structurally labile.

4.8. Site-Specific Selection and Sequence Conservation

Codon-level selection analyses were conducted using HyPhy v2.5 [47] on Pal2Nal codon alignments with IQ-TREE2 phylogenies. Site-specific dN/dS ratios were estimated using Fixed Effects Likelihood (FEL) under the MG94 × GTR substitution model with synonymous rate variation[75]. Sites classified as diversifying ($dN > dS$) or purifying ($dN < dS$) at $p \leq 0.05$. Episodic positive selection was detected using MEME[30] ($p \leq 0.05$; highly significant sites at $p \leq 0.01$). Sequence conservation was quantified per site using AL2CO v1.0 [76], standardized to Z-scores.

4.9. Software and Statistical Environment

All analyses were performed in R v4.4.2. Key packages: tidyverse, ggplot2 v3.5, ggtree/treeio, patchwork, rstatix, lavaan v0.6-21, mediation v4.5.0, caper, RColorBrewer. Python analyses used pandas. Sequence processing used SeqKit v2.8.0 and ClipKIT v1.3.0. Figures were exported at 300 dpi (supplementary) and 600 dpi (main figure).

5. Conclusions

This study provides the first systematic characterization of the non-canonical sequence space of ABC transporters across all five architectural classes. Non-domain regions, including inter-domain linkers and terminal flanking segments, are consistently more disordered than NBD and TMD cores, with this hierarchy phylogenetically conserved across lineages (Pagel's $\lambda \approx 0.97$) and driven primarily by amino acid composition with an additional region-specific contribution from GC codon usage. The concentration of predicted PTM sites at both disordered non-domain positions and slowly evolving domain-proximal sites, together with episodic positive selection concentrated at inter-domain boundaries across all architectural classes, support the view that non-canonical regions are not passive spacers but are structurally and evolutionarily active components of the transporter architecture. Together, these findings provide a positional map of predicted regulatory sites,

 doi:10.20944/preprints202604.1710.v1

compositional signatures, and evolutionary pressures across the ABC transporter superfamily, guiding the prioritization of non-domain positions for experimental and therapeutic investigation.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org, Figure S1: Architecture-specific amino acid composition of ABC transporter linker regions with permutation FDR significance; Figure S2: GC content versus fraction of disordered residues by region type; Figure S3: Comparison of GC content at all codon positions versus mean disorder by region type and direct effect on linker disorder; Figure S4: NetSurfP-3.0 secondary structure propensity profiles across ABC transporter architectural classes; Table S1: KEGG organism code abbreviations for all species represented in the dataset; Table S2: Dataset composition per architectural class; Table S3: Cross-architecture pairwise comparisons of regional intrinsic disorder fraction and linker length; Table S4: Tiered predicted PTM conservation sites across ABC transporter architectural classes; Table S5: Within-architecture pairwise comparisons of region length and disorder fraction; Table S6: Validation of MusiteDeep predictions against experimentally verified UniProt PTM annotations for human ABC transporters.

Author Contributions: Conceptualization, I.A.D.; methodology, I.A.D. and M.I.; software, I.A.D.; validation, I.A.D. and Y.K; formal analysis, I.A.D.; investigation, I.A.D. and Y.K; resources, I.A.D. and M.I.; data curation, I.A.D.; writing—original draft preparation, I.A.D.; writing—review and editing, I.A.D. and M.I.; visualization, I.A.D.; supervision, M.I. and Y.K; project administration, M.I and I.A.D.; funding acquisition, M.I. All authors have read and agreed to the published version of the manuscript.

Funding: This research was funded by the Ministry of Education, Culture, Sports, Science and Technology (MEXT), Japan.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: The data presented in this study are available in the article and Supplementary Materials.

Acknowledgments: The authors would like to thank the reviewers for their helpful comments and the members of information biology laboratory of Ritsumeikan University for their support and helpful comments.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

ABC	ATP-Binding Cassette
IDR	Intrinsically Disordered Region
NBD	Nucleotide Binding Domain
PDR	Pleiotropic Drug Resistance
PGLS	Phylogenetic Generalized Least Squares
PTM	Post-translational Modification
TMD	Transmembrane Domain

References

1. Dean, M.; Annilo, T. EVOLUTION OF THE ATP-BINDING CASSETTE (ABC) TRANSPORTER SUPERFAMILY IN VERTEBRATES. *Annu. Rev. Genom. Hum. Genet.* **2005**, *6*, 123–142, doi:10.1146/annurev.genom.6.080604.162122.
2. Thomas, C.; Tampé, R. Structural and Mechanistic Principles of ABC Transporters. *Annu. Rev. Biochem.* **2020**, *89*, 605–636, doi:10.1146/annurev-biochem-011520-105201.
3. Alam, A.; Locher, K.P. Structure and Mechanism of Human ABC Transporters. *Annu. Rev. Biophys.* **2023**, *52*, 275–300, doi:10.1146/annurev-biophys-111622-091232.

 doi:10.20944/preprints202604.1710.v1

4. Rees, D.C.; Johnson, E.; Lewinson, O. ABC Transporters: The Power to Change. *Nat Rev Mol Cell Biol* **2009**, *10*, 218–227, doi:10.1038/nrm2646.
5. Dean, M.; Rzhetsky, A.; Allikmets, R. The Human ATP-Binding Cassette (ABC) Transporter Superfamily. *Genome Res.* **2001**, *11*, 1156–1166, doi:10.1101/gr.184901.
6. Dean, M. The Genetics of ATP-Binding Cassette Transporters. In *Methods in Enzymology*; Elsevier, 2005; Vol. 400, pp. 409–429 ISBN 978-0-12-182805-9.
7. Murina, V.; Kasari, M.; Takada, H.; Hinnu, M.; Saha, C.K.; Grimshaw, J.W.; Seki, T.; Reith, M.; Putrinš, M.; Tenson, T.; et al. ABCF ATPases Involved in Protein Synthesis, Ribosome Assembly and Antibiotic Resistance: Structural and Functional Diversification across the Tree of Life. *Journal of Molecular Biology* **2019**, *431*, 3568–3590, doi:10.1016/j.jmb.2018.12.013.
8. Woodward, O.M.; Köttgen, A.; Köttgen, M. ABCG Transporters and Disease. *The FEBS Journal* **2011**, *278*, 3215–3225, doi:10.1111/j.1742-4658.2011.08171.x.
9. Ferreira, R.J.; Bonito, C.A.; Cordeiro, M.N.D.S.; Ferreira, M.-J.U.; Dos Santos, D.J.V.A. Structure-Function Relationships in ABCG2: Insights from Molecular Dynamics Simulations and Molecular Docking Studies. *Sci Rep* **2017**, *7*, 15534, doi:10.1038/s41598-017-15452-z.
10. Yu, J.; Ge, J.; Heuveling, J.; Schneider, E.; Yang, M. Structural Basis for Substrate Specificity of an Amino Acid ABC Transporter. *Proc. Natl. Acad. Sci. U.S.A.* **2015**, *112*, 5243–5248, doi:10.1073/pnas.1415037112.
11. Ford, R.C.; Marshall-Sabey, D.; Schuetz, J. Linker Domains: Why ABC Transporters 'Live in Fragments No Longer.' *Trends in Biochemical Sciences* **2020**, *45*, 137–148, doi:10.1016/j.tibs.2019.11.004.
12. Bickers, S.C.; Benlekbir, S.; Rubinstein, J.L.; Kanelis, V. Structure of Ycf1p Reveals the Transmembrane Domain TMD0 and the Regulatory Region of ABCC Transporters. *Proc. Natl. Acad. Sci. U.S.A.* **2021**, *118*, e2025853118, doi:10.1073/pnas.2025853118.
13. Baker, J.M.R.; Hudson, R.P.; Kanelis, V.; Choy, W.-Y.; Thibodeau, P.H.; Thomas, P.J.; Forman-Kay, J.D. CFTR Regulatory Region Interacts with NBD1 Predominantly via Multiple Transient Helices. *Nat Struct Mol Biol* **2007**, *14*, 738–745, doi:10.1038/nsmb1278.
14. Ambadipudi, R.; Georges, E. Sequences in Linker-1 Domain of the Multidrug Resistance Associated Protein (MRP1 or ABCC1) Bind to Tubulin and Their Binding Is Modulated by Phosphorylation. *Biochemical and Biophysical Research Communications* **2017**, *482*, 1001–1006, doi:10.1016/j.bbrc.2016.11.147.
15. I. Stolarczyk, E.; J. Reiling, C.; M. Paumi, C. Regulation of ABC Transporter Function Via Phosphorylation by Protein Kinases. *CPB* **2011**, *12*, 621–635, doi:10.2174/138920111795164075.
16. Roosbeek, S.; Peelman, F.; Verhee, A.; Labeur, C.; Caster, H.; Lensink, M.F.; Cirulli, C.; Grooten, J.; Cochet, C.; Vandekerckhove, J.; et al. Phosphorylation by Protein Kinase CK2 Modulates the Activity of the ATP Binding Cassette A1 Transporter. *Journal of Biological Chemistry* **2004**, *279*, 37779–37788, doi:10.1074/jbc.M401821200.
17. Tang, C.; Liu, Y.; Kessler, P.S.; Vaughan, A.M.; Oram, J.F. The Macrophage Cholesterol Exporter ABCA1 Functions as an Anti-Inflammatory Receptor. *J Biol Chem* **2009**, *284*, 32336–32343, doi:10.1074/jbc.M109.047472.
18. Holehouse, A.S.; Kragelund, B.B. The Molecular Basis for Cellular Function of Intrinsically Disordered Protein Regions. *Nat Rev Mol Cell Biol* **2024**, *25*, 187–211, doi:10.1038/s41580-023-00673-0.
19. Gao, C.; Ma, C.; Wang, H.; Zhong, H.; Zang, J.; Zhong, R.; He, F.; Yang, D. Intrinsic Disorder in Protein Domains Contributes to Both Organism Complexity and Clade-Specific Functions. *Sci Rep* **2021**, *11*, 2985, doi:10.1038/s41598-021-82656-9.
20. Fahmi, M.; Ito, M. Evolutionary Approach of Intrinsically Disordered CIP/KIP Proteins. *Sci Rep* **2019**, *9*, 1575, doi:10.1038/s41598-018-37917-5.
21. Singleton, M.D.; Eisen, M.B. Evolutionary Analyses of Intrinsically Disordered Regions Reveal Widespread Signals of Conservation. *PLoS Comput Biol* **2024**, *20*, e1012028, doi:10.1371/journal.pcbi.1012028.
22. Zarin, T.; Strome, B.; Peng, G.; Pritišanac, I.; Forman-Kay, J.D.; Moses, A.M. Identifying Molecular Features That Are Associated with Biological Function of Intrinsically Disordered Protein Regions. *Elife* **2021**, *10*, e60220, doi:10.7554/eLife.60220.

 doi:10.20944/preprints202604.1710.v1

23. Souza Amado De Carvalho, R.; Rasel, M.S.I.; Khandelwal, N.K.; Tomasiak, T.M. Cryo-EM Reveals a Phosphorylated R-Domain Envelops the NBD1 Catalytic Domain in an ABC Transporter. *Life Sci. Alliance* **2024**, *7*, e202402779, doi:10.26508/lsa.202402779.
24. Qian, H.; Zhao, X.; Cao, P.; Lei, J.; Yan, N.; Gong, X. Structure of the Human Lipid Exporter ABCA1. *Cell* **2017**, *169*, 1228-1239.e10, doi:10.1016/j.cell.2017.05.020.
25. Wilkens, S. Structure and Mechanism of ABC Transporters. *F1000Prime Rep* **2015**, *7*, doi:10.12703/P7-14.
26. Pechmann, S.; Frydman, J. Evolutionary Conservation of Codon Optimality Reveals Hidden Signatures of Cotranslational Folding. *Nat Struct Mol Biol* **2013**, *20*, 237–243, doi:10.1038/nsmb.2466.
27. Peng, Z.; Uversky, V.N.; Kurgan, L. Genes Encoding Intrinsic Disorder in Eukaryota Have High GC Content. *Intrinsically Disordered Proteins* **2016**, *4*, e1262225, doi:10.1080/21690707.2016.1262225.
28. Basile, W.; Sachenkova, O.; Light, S.; Elofsson, A. High GC Content Causes Orphan Proteins to Be Intrinsically Disordered. *PLoS Comput Biol* **2017**, *13*, e1005375, doi:10.1371/journal.pcbi.1005375.
29. Homma, K.; Anbo, H.; Noguchi, T.; Fukuchi, S. Both Intrinsically Disordered Regions and Structural Domains Evolve Rapidly in Immune-Related Mammalian Proteins. *IJMS* **2018**, *19*, 3860, doi:10.3390/ijms19123860.
30. Murrell, B.; Wertheim, J.O.; Moola, S.; Weighill, T.; Scheffler, K.; Kosakovsky Pond, S.L. Detecting Individual Sites Subject to Episodic Diversifying Selection. *PLoS Genet* **2012**, *8*, e1002764, doi:10.1371/journal.pgen.1002764.
31. Freckleton, R.P.; Harvey, P.H.; Pagel, M. Phylogenetic Analysis and Comparative Data: A Test and Review of Evidence. *The American Naturalist* **2002**, *160*, 712–726, doi:10.1086/343873.
32. Pagel, M. Inferring the Historical Patterns of Biological Evolution. *Nature* **1999**, *401*, 877–884, doi:10.1038/44766.
33. See, R.H.; Caday-Malcolm, R.A.; Singaraja, R.R.; Zhou, S.; Silverston, A.; Huber, M.T.; Moran, J.; James, E.R.; Janoo, R.; Savill, J.M.; et al. Protein Kinase A Site-Specific Phosphorylation Regulates ATP-Binding Cassette A1 (ABCA1)-Mediated Phospholipid Efflux. *Journal of Biological Chemistry* **2002**, *277*, 41835–41842, doi:10.1074/jbc.M204923200.
34. D'Onofrio, G.; Jabbari, K.; Musto, H.; Alvarez-Valin, F.; Cruveiller, S.; Bernardi, G. Evolutionary Genomics of Vertebrates and Its Implications. *Annals of the New York Academy of Sciences* **1999**, *870*, 81–94, doi:10.1111/j.1749-6632.1999.tb08867.x.
35. Moesa, H.A.; Wakabayashi, S.; Nakai, K.; Patil, A. Chemical Composition Is Maintained in Poorly Conserved Intrinsically Disordered Regions and Suggests a Means for Their Classification. *Molecular BioSystems* **2012**, *8*, 3262–3273, doi:10.1039/C2MB25202C.
36. Iakoucheva, L.M. The Importance of Intrinsic Disorder for Protein Phosphorylation. *Nucleic Acids Research* **2004**, *32*, 1037–1049, doi:10.1093/nar/gkh253.
37. Ahmed, S.S.; Rifat, Z.T.; Lohia, R.; Campbell, A.J.; Dunker, A.K.; Rahman, M.S.; Iqbal, S. Characterization of Intrinsically Disordered Regions in Proteins Informed by Human Genetic Diversity. *PLoS Comput Biol* **2022**, *18*, e1009911, doi:10.1371/journal.pcbi.1009911.
38. Newcombe, E.A.; Delaforge, E.; Hartmann-Petersen, R.; Skriver, K.; Kragelund, B.B. How Phosphorylation Impacts Intrinsically Disordered Proteins and Their Function. *Essays in Biochemistry* **2022**, *66*, 901–913, doi:10.1042/EBC20220060.
39. Ren, H.Y.; Grove, D.E.; De La Rosa, O.; Houck, S.A.; Sopha, P.; Van Goor, F.; Hoffman, B.J.; Cyr, D.M. VX-809 Corrects Folding Defects in Cystic Fibrosis Transmembrane Conductance Regulator Protein through Action on Membrane-Spanning Domain 1. *MBoC* **2013**, *24*, 3016–3024, doi:10.1091/mbc.e13-05-0240.
40. Bin Kanner, Y.; Ganoth, A.; Tsfadia, Y. Extracellular Mutation Induces an Allosteric Effect across the Membrane and Hampers the Activity of MRP1 (ABCC1). *Sci Rep* **2021**, *11*, 12024, doi:10.1038/s41598-021-91461-3.
41. Heinkel, F.; Abraham, L.; Ko, M.; Chao, J.; Bach, H.; Hui, L.T.; Li, H.; Zhu, M.; Ling, Y.M.; Rogalski, J.C.; et al. Phase Separation and Clustering of an ABC Transporter in *Mycobacterium Tuberculosis*. *Proc. Natl. Acad. Sci. U.S.A.* **2019**, *116*, 16326–16331, doi:10.1073/pnas.1820683116.
42. Ruan, H.; Sun, Q.; Zhang, W.; Liu, Y.; Lai, L. Targeting Intrinsically Disordered Proteins at the Edge of Chaos. *Drug Discovery Today* **2019**, *24*, 217–227, doi:10.1016/j.drudis.2018.09.017.

 doi:10.20944/preprints202604.1710.v1

43. Biesaga, M.; Frigolé-Vivas, M.; Salvatella, X. Intrinsically Disordered Proteins and Biomolecular Condensates as Drug Targets. *Current Opinion in Chemical Biology* **2021**, *62*, 90–100, doi:10.1016/j.cbpa.2021.02.009.
44. Fantini, J.; Azzaz, F.; Di Scala, C.; Aulas, A.; Chahinian, H.; Yahi, N. Conformationally Adaptive Therapeutic Peptides for Diseases Caused by Intrinsically Disordered Proteins (IDPs). New Paradigm for Drug Discovery: Target the Target, Not the Arrow. *Pharmacology & Therapeutics* **2025**, *267*, 108797, doi:10.1016/j.pharmthera.2025.108797.
45. Kanehisa, M.; Furumichi, M.; Sato, Y.; Kawashima, M.; Ishiguro-Watanabe, M. KEGG for Taxonomy-Based Analysis of Pathways and Genomes. *Nucleic Acids Research* **2023**, *51*, D587–D592, doi:10.1093/nar/gkac963.
46. Shen, W.; Le, S.; Li, Y.; Hu, F. SeqKit: A Cross-Platform and Ultrafast Toolkit for FASTA/Q File Manipulation. *PLoS ONE* **2016**, *11*, e0163962, doi:10.1371/journal.pone.0163962.
47. Kosakovsky Pond, S.L.; Poon, A.F.Y.; Velazquez, R.; Weaver, S.; Hepler, N.L.; Murrell, B.; Shank, S.D.; Magalis, B.R.; Bouvier, D.; Nekrutenko, A.; et al. HyPhy 2.5—A Customizable Platform for Evolutionary Hypothesis Testing Using Phylogenies. *Molecular Biology and Evolution* **2020**, *37*, 295–299, doi:10.1093/molbev/msz197.
48. Eddy, S.R. Accelerated Profile HMM Searches. *PLoS Comput Biol* **2011**, *7*, e1002195, doi:10.1371/journal.pcbi.1002195.
49. Mistry, J.; Chuguransky, S.; Williams, L.; Qureshi, M.; Salazar, G.A.; Sonnhammer, E.L.L.; Tosatto, S.C.E.; Paladin, L.; Raj, S.; Richardson, L.J.; et al. Pfam: The Protein Families Database in 2021. *Nucleic Acids Research* **2021**, *49*, D412–D419, doi:10.1093/nar/gkaa913.
50. Shibata, Y.; Ojika, M.; Sugiyama, A.; Yazaki, K.; Jones, D.A.; Kawakita, K.; Takemoto, D. The Full-Size ABCG Transporters Nb-ABCG1 and Nb-ABCG2 Function in Pre- and Postinvasion Defense against *Phytophthora Infestans* in *Nicotiana Benthamiana*. *Plant Cell* **2016**, *28*, 1163–1181, doi:10.1105/tpc.15.00721.
51. Kang, J.; Park, J.; Choi, H.; Burla, B.; Kretzschmar, T.; Lee, Y.; Martinoia, E. Plant ABC Transporters. *The Arabidopsis Book* **2011**, *9*, e0153, doi:10.1199/tab.0153.
52. Lamping, E.; Baret, P.V.; Holmes, A.R.; Monk, B.C.; Goffeau, A.; Cannon, R.D. Fungal PDR Transporters: Phylogeny, Topology, Motifs and Function. *Fungal Genetics and Biology* **2010**, *47*, 127–142, doi:10.1016/j.fgb.2009.10.007.
53. Crouzet, J.; Trombik, T.; Fraysse, Å.S.; Boutry, M. Organization and Function of the Plant Pleiotropic Drug Resistance ABC Transporter Family. *FEBS Letters* **2006**, *580*, 1123–1130, doi:10.1016/j.febslet.2005.12.043.
54. Katoh, K.; Standley, D.M. MAFFT Multiple Sequence Alignment Software Version 7: Improvements in Performance and Usability. *Molecular Biology and Evolution* **2013**, *30*, 772–780, doi:10.1093/molbev/mst010.
55. Steenwyk, J.L.; Buida, T.J.; Li, Y.; Shen, X.-X.; Rokas, A. ClipKIT: A Multiple Sequence Alignment Trimming Software for Accurate Phylogenomic Inference. *PLoS Biol* **2020**, *18*, e3001007, doi:10.1371/journal.pbio.3001007.
56. Suyama, M.; Torrents, D.; Bork, P. PAL2NAL: Robust Conversion of Protein Sequence Alignments into the Corresponding Codon Alignments. *Nucleic Acids Research* **2006**, *34*, W609–W612, doi:10.1093/nar/gkl315.
57. Minh, B.Q.; Schmidt, H.A.; Chernomor, O.; Schrempf, D.; Woodhams, M.D.; Von Haeseler, A.; Lanfear, R. IQ-TREE 2: New Models and Efficient Methods for Phylogenetic Inference in the Genomic Era. *Molecular Biology and Evolution* **2020**, *37*, 1530–1534, doi:10.1093/molbev/msaa015.
58. Kalyaanamoorthy, S.; Minh, B.Q.; Wong, T.K.F.; Von Haeseler, A.; Jermiin, L.S. ModelFinder: Fast Model Selection for Accurate Phylogenetic Estimates. *Nat Methods* **2017**, *14*, 587–589, doi:10.1038/nmeth.4285.
59. Hallgren, J.; Tsirigos, K.D.; Pedersen, M.D.; Almagro Armenteros, J.J.; Marcatili, P.; Nielsen, H.; Krogh, A.; Winther, O. DeepTMHMM Predicts Alpha and Beta Transmembrane Proteins Using Deep Neural Networks 2022.
60. Tsirigos, K.D.; Peters, C.; Shu, N.; Käll, L.; Elofsson, A. The TOPCONS Web Server for Consensus Prediction of Membrane Protein Topology and Signal Peptides. *Nucleic Acids Res* **2015**, *43*, W401–W407, doi:10.1093/nar/gkv485.
61. Erdős, G.; Dosztányi, Z. AIUPred: Combining Energy Estimation with Deep Learning for the Enhanced Prediction of Protein Disorder. *Nucleic Acids Research* **2024**, *52*, W176–W181, doi:10.1093/nar/gkae385.

 doi:10.20944/preprints202604.1710.v1

62. Høie, M.H.; Kiehl, E.N.; Petersen, B.; Nielsen, M.; Winther, O.; Nielsen, H.; Hallgren, J.; Marcatili, P. NetSurfP-3.0: Accurate and Fast Prediction of Protein Structural Features by Protein Language Models and Deep Learning. *Nucleic Acids Research* **2022**, *50*, W510–W515, doi:10.1093/nar/gkac439.
63. Xu, S.; Li, L.; Luo, X.; Chen, M.; Tang, W.; Zhan, L.; Dai, Z.; Lam, T.T.; Guan, Y.; Yu, G. *Ggtree* : A Serialized Data Object for Visualization of a Phylogenetic Tree and Annotation Data. *iMeta* **2022**, *1*, e56, doi:10.1002/imt2.56.
64. Wickham, H. *Ggplot2*; Use R!; Springer International Publishing: Cham, 2016; ISBN 978-3-319-24275-0.
65. Wang, D.; Liu, D.; Yuchi, J.; He, F.; Jiang, Y.; Cai, S.; Li, J.; Xu, D. MusiteDeep: A Deep-Learning Based Webserver for Protein Post-Translational Modification Site Prediction and Visualization. *Nucleic Acids Research* **2020**, *48*, W140–W146, doi:10.1093/nar/gkaa275.
66. Valdar, W.S.J. Scoring Residue Conservation. *Proteins* **2002**, *48*, 227–241, doi:10.1002/prot.10146.
67. Capra, J.A.; Singh, M. Predicting Functionally Important Residues from Sequence Conservation. *Bioinformatics* **2007**, *23*, 1875–1882, doi:10.1093/bioinformatics/btm270.
68. Minguez, P.; Parca, L.; Diella, F.; Mende, D.R.; Kumar, R.; Helmer-Citterich, M.; Gavin, A.; Van Noort, V.; Bork, P. Deciphering a Global Network of Functionally Associated Post-translational Modifications. *Molecular Systems Biology* **2012**, *8*, 599, doi:10.1038/msb.2012.31.
69. Dunker, A.K.; Lawson, J.D.; Brown, C.J.; Williams, R.M.; Romero, P.; Oh, J.S.; Oldfield, C.J.; Campen, A.M.; Ratliff, C.M.; Hipps, K.W.; et al. Intrinsically Disordered Protein. *Journal of Molecular Graphics and Modelling* **2001**, *19*, 26–59, doi:10.1016/S1093-3263(00)00138-8.
70. Uversky, V.N.; Gillespie, J.R.; Fink, A.L. Why Are ?Natively Unfolded? Proteins Unstructured under Physiologic Conditions? *Proteins* **2000**, *41*, 415–427, doi:10.1002/1097-0134(20001115)41:3<415::AID-PROT130>3.0.CO;2-7.
71. Tingley, D.; Yamamoto, T.; Hirose, K.; Keele, L.; Imai, K. **Mediation** : *R* Package for Causal Mediation Analysis. *J. Stat. Soft.* **2014**, *59*, doi:10.18637/jss.v059.i05.
72. Rosseel, Y. **Lavaan** : An *R* Package for Structural Equation Modeling. *J. Stat. Soft.* **2012**, *48*, doi:10.18637/jss.v048.i02.
73. Orme, D. Caper: Comparative Analyses of Phylogenetics and Evolution in R 2018.
74. Cohen, O.; Ashkenazy, H.; Belinky, F.; Huchon, D.; Pupko, T. GLOOME: Gain Loss Mapping Engine. *Bioinformatics* **2010**, *26*, 2914–2915, doi:10.1093/bioinformatics/btq549.
75. Kosakovsky Pond, S.L.; Frost, S.D.W. Not So Different After All: A Comparison of Methods for Detecting Amino Acid Sites Under Selection. *Molecular Biology and Evolution* **2005**, *22*, 1208–1222, doi:10.1093/molbev/msi105.
76. Pei, J.; Grishin, N.V. AL2CO: Calculation of Positional Conservation in a Protein Sequence Alignment. *Bioinformatics* **2001**, *17*, 700–712, doi:10.1093/bioinformatics/17.8.700.