Pathogenic Missense Mutations in Intrinsically Disordered Regions Reveal Functional and Clinical Signals

Norbert Deutsch, Gábor Erdős, *Zsuzsanna Dosztányi

Department of Biochemistry, Eötvös Loránd University, Pázmány Péter stny 1/c, Budapest H-1117, Hungary
*zsuzsanna.dosztanyi@ttk.elte.hu

Abstract

Intrinsically disordered regions (IDRs) play essential roles in signaling and regulation but remain under-characterized in the context of human disease. We analyzed nearly one million ClinVar missense variants, focusing on the subset within IDRs defined by curated and predicted annotations. Pathogenic variants were significantly enriched in short linear motifs (SLiMs) and disordered binding regions, consistent with their functional importance. To extend these insights beyond existing annotations, we applied AlphaMissense, a deep learning model for variant effect prediction, revealing localized “island-like” pathogenicity patterns within IDRs. Based on these signals, we developed a classifier to prioritize predicted ELM motifs (PEMs), identifying motifs linked to pathogenicity and major disease classes, including neurological, cardiovascular, and cancer-associated genes. Case studies in POLK, FOXP2, and LMOD3 demonstrate how this framework yields mechanistic insight into disease-relevant mutations within disordered protein regions, advancing clinical interpretation in the disordered proteome.

Introduction

Missense variants represent the most common form of genetic variation in humans, yet their clinical interpretation remains a major challenge. Despite advances in cataloging human variation, the majority of missense mutations remain classified as variants of uncertain significance (VUS), limiting their utility in diagnosis and precision medicine [1–3]. Historically, efforts to understand the pathogenicity of missense variants have focused predominantly on globular protein domains, due to their stable structures and rich functional annotations. However, over one-third of the human proteome consists of intrinsically disordered regions (IDRs)—segments that lack stable three-dimensional structure yet play essential roles in signaling, regulation, and molecular interactions [4–7].

IDPs typically form weak, transient interactions, often mediated by short linear motifs (SLiMs) that regulate binding, localization, and post-translational modification [8,9]. Recent proteome-scale studies have shown that pathogenic mutations can disrupt SLiM-mediated interactions, rewiring protein networks or creating novel motifs that alter protein function and localization [10–12]. IDRs also play pivotal roles in organizing biomolecular condensates—phase-separated compartments essential for diverse cellular processes [13,14]. Misregulation of condensate formation can also lead to disease, as seen with tau aggregation in Alzheimer's disease [15]. Even small sequence changes can disrupt condensate dynamics; for example, mis-splicing of a microexon in the IDR of the RNA-binding protein CPEB4 converts reversible condensates into solid aggregates, impairing neuronal function in autism [16]. In cancer, IDRs are similarly critical: approximately 20% of known driver genes harbor mutations within disordered regions, often affecting long, conserved, and ancient domains [5,17]. Despite their broad biological and clinical relevance, IDRs remain systematically under-annotated, and most missense variants affecting disordered regions still lack mechanistic interpretation.

To better characterize this landscape, we first systematically analyzed pathogenic and uncertain missense variants from ClinVar [1–3], focusing on their distribution within disordered regions. Despite comprising over one-third of the proteome, IDRs account for only ~8% of pathogenic ClinVar mutations. Nevertheless, pathogenic variants in IDRs are significantly enriched in functional elements such as experimentally validated disordered regions, post-translational modification sites, and SLiMs. These observations underscore the functional relevance of IDRs and highlight the urgent need for improved approaches to interpret variation in these regions.

While advances in machine learning have improved variant effect prediction, including the development of AlphaMissense[18], the interpretation of IDR mutations remains a major gap[19]. General-purpose predictors often miss context-dependent features such as transient binding or phase separation propensity. Recent models incorporating phase separation properties have improved predictions in IDRs[20] , underscoring the need for specialized approaches.

Here, we integrate pathogenicity predictions from AlphaMissense with motif-centric analysis to systematically map missense variation within IDRs. We reveal that AlphaMissense scores highlight localized "island-like" pathogenicity patterns within SLiMs, enabling the identification of constrained and disease-relevant motifs. Building on this signal, we develop a classifier to prioritize predicted ELM motifs (PEMs) and link them to clinically significant

mutations, offering a scalable framework to advance variant interpretation across disordered regions.

Results

Clinvar mutations in disordered regions

To investigate the landscape of clinically relevant missense variants within intrinsically disordered regions (IDRs), we first systematically analyzed pathogenic and uncertain variants from ClinVar. We examined their distribution relative to structural states across the human proteome, using 19,692 canonical isoforms from GENCODE (Frankish et al., 2021). Approximately one million single nucleotide polymorphisms (SNPs) were mapped and classified by clinical significance as Pathogenic, Benign, or Uncertain. Overall, these ClinVar variants covered less than 8% of all amino acid positions in the human proteome, with ~70,000 classified as pathogenic (Figure 1a).

By classifying positions using the “Combined Disorder” model, we found that disordered regions comprise 36.8% of the proteome (Figure 1b). Uncertain mutations (35%) aligned with this distribution, while Benign mutations were enriched in disordered regions (47.3%, Figure 1c). Pathogenic mutations were less common in disordered regions (8%), which may reflect functional constraints in ordered regions or biases toward studying mutations in ordered regions.

In total, 2,984 genes contained pathogenic mutations, of which 798 had mutations in IDRs, including 244 genes exclusively within disordered regions (Figure 1d). A large majority of genes had mutations with uncertain significance, in most cases (13268, 73%) both in ordered and disordered regions. Benign mutations were found in 11,732 genes, with 3,250 in disordered regions and 4,176 in ordered regions (Figure 1d). In total, 5,467 pathogenic mutations (2,298 unique positions) were identified in IDRs, with 2,957 distinct position-disease pairs (Figure 1c, Supplementary Table 1). The list of collected pathogenic mutations in IDRs is provided in Supplementary Data 1.

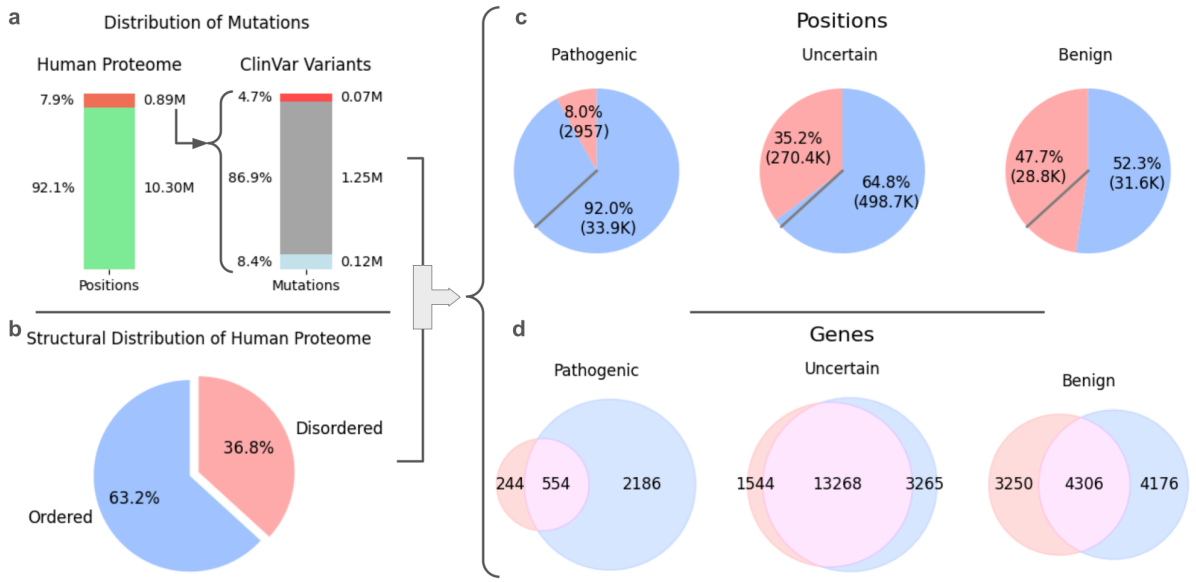


Figure 1) Structural and Functional Analysis of ClinVar Variants. a) Distribution of ClinVar Mutations Across the Human Proteome. Left panel illustrates the percentage of proteome positions with mutations reported in ClinVar (red). The right panel presents mutation classifications according to clinical significance: benign (blue), uncertain (grey), and pathogenic (red). b) Structural Composition of the Human Proteome. Based on our combined disorder model, positions classified as disordered are represented in light pink, and ordered regions in light blue. c) Structural Classification of ClinVar Mutations by Position. Positional distribution of ClinVar-reported mutations classified based on their structural context within the proteome—ordered (light blue) or disordered (light pink)—highlighting the distribution of mutation-containing positions according to their structural state. d) Venn Diagram of Genes Containing ClinVar Mutations. Illustrates the overlap and specificity of mutated genes categorized by the structural regions (ordered, disordered, or both) that harbor mutations.

Specificity of pathogenic mutations within IDRs

We further examined the relationship between disease complexity and pathogenic mutations by categorizing diseases into Monogenic (single gene), Multigenic (2–4 genes), and Complex (>4 genes). Over 90% of diseases found with mutations in ClinVar involve fewer than five genes, however, nearly half of the mutations are linked to complex diseases and around 20% to monogenic diseases. Mutations within disordered regions showed a markedly higher association with complex diseases compared to those in ordered regions, suggesting that IDRs may play a particularly important role in signaling pathways and regulatory networks disrupted in these conditions (Figure 2a).

Pathogenic mutations within IDRs were classified based on their distribution across proteins: those occurring predominantly (>60%) in IDRs versus those found equally or predominantly in ordered regions. Approximately 43.9% of mutations were specific to disordered regions, with genes predominantly or exclusively mutated within IDRs. This increased to 50.5% for monogenic diseases (Figure 2b).

Diseases most impacted by mutations include neurodegenerative, neurodevelopmental, cardiovascular disorders, and cancer (Supplementary Figure 1a). Notably, in cancer, cardiovascular, and neurodevelopmental diseases, these mutations predominantly were more specific to IDR regions (Supplementary Figure 1b). In line with this, genes affected by IDR-localized mutations are most frequently associated with neurodegenerative, cardiovascular, and cancer-related disease categories (Figure 2c), highlighting the contribution of disordered regions at the gene level as well.

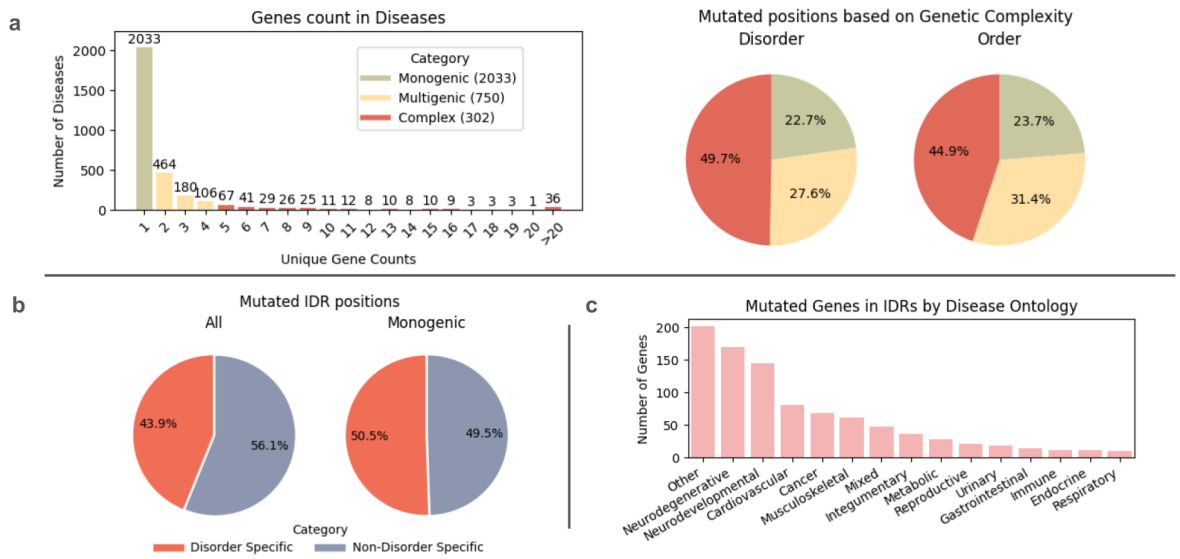


Figure 2) Analysis of genetic complexity and disease ontology of ClinVar variants. (a) Genetic complexity distributions. The left panel illustrates the distribution of genetic complexity based on genes and diseases, highlighting that most diseases are monogenic. The right panel shows the proportion of mutated positions categorized by structural classification. Notably, a higher proportion of mutations in disordered regions is observed in complex diseases. (b) Distribution of mutated positions within IDRs by genetic complexity. For monogenic diseases, the majority of mutations predominantly or exclusively occur within IDRs. (c) Distribution of mutated genes across disease classes with only IDR mutations.

Functional Annotations of pathogenic mutations within IDRs.

We further examined the structural and functional characterization of pathogenic mutations within IDRs. Structural data was gathered from the PDB and experimentally verified disordered regions from the MobiDB database. Functional annotations were compiled from UniProt and disorder-specific resources, such as ELM, DIBS, MFIB, and PhasePro, along with post-translational modification (PTM) data sourced from dbPTM and PhosphoSitePlus.

Pathogenic mutations exhibited greater experimental structural coverage compared to uncertain variants; however, more than 50% of ordered and 65% of disordered sites remained uncharacterized (Figure 3a). Among the 2,957 pathogenic positions analyzed, 15.5% were located in experimentally verified disordered regions, 20.7% overlapped with functionally annotated sites, and 8.5% were annotated for both disorder and function. While there is a clear enrichment in annotations for pathogenic mutations compared to uncertain

ones, 54.5% of pathogenic mutations within IDRs still lack any functional annotation (Figure 3b), underscoring the need for further characterization.

In terms of specific functional annotations, the highest number of pathogenic positions occurred in experimentally disordered regions (778) and UniProt functional sites (530). Additionally, binding sites (ELM: 176, DIBS: 62) and phase separation-associated regions (PhasePro: 47) were also enriched, highlighting their role in disease mechanisms. Among post-translational modifications, phosphorylation (206) showed the highest association with pathogenicity (Figure 3c). Pathogenic mutations that occur in annotated disordered functional regions are typically found in proteins where mutations primarily or exclusively affect disordered regions. This trend is especially strong for binding sites (e.g., ELM, DIBS) and phase separation-associated regions (PhasePro), indicating that mutations in these functionally important IDR elements may play a central role in driving disease. Supporting this notion, most UniProt-annotated pathogenic positions within IDRs (Supplementary Figure 1c) are linked to interaction and binding functions, further emphasizing that such regions act as critical molecular interfaces whose disruption is more likely to have pathogenic consequences.

This functional relevance is further reflected in the distribution of benign mutations across disordered regions. While unannotated IDRs showed a benign mutation rate of 1.48%, this rate was lower in annotated functional regions, including DIBS (1.21%), PhasePro (1.06%), and MFIB (0.97%). Notably, ELM motifs displayed the lowest benign mutation rate (0.94%), suggesting they are under the strongest purifying selection. These findings further reinforce that functionally annotated disordered elements—particularly short linear motifs—represent critical sites of molecular interactions, where variation is more likely to be deleterious (Supplementary Figure 1d).

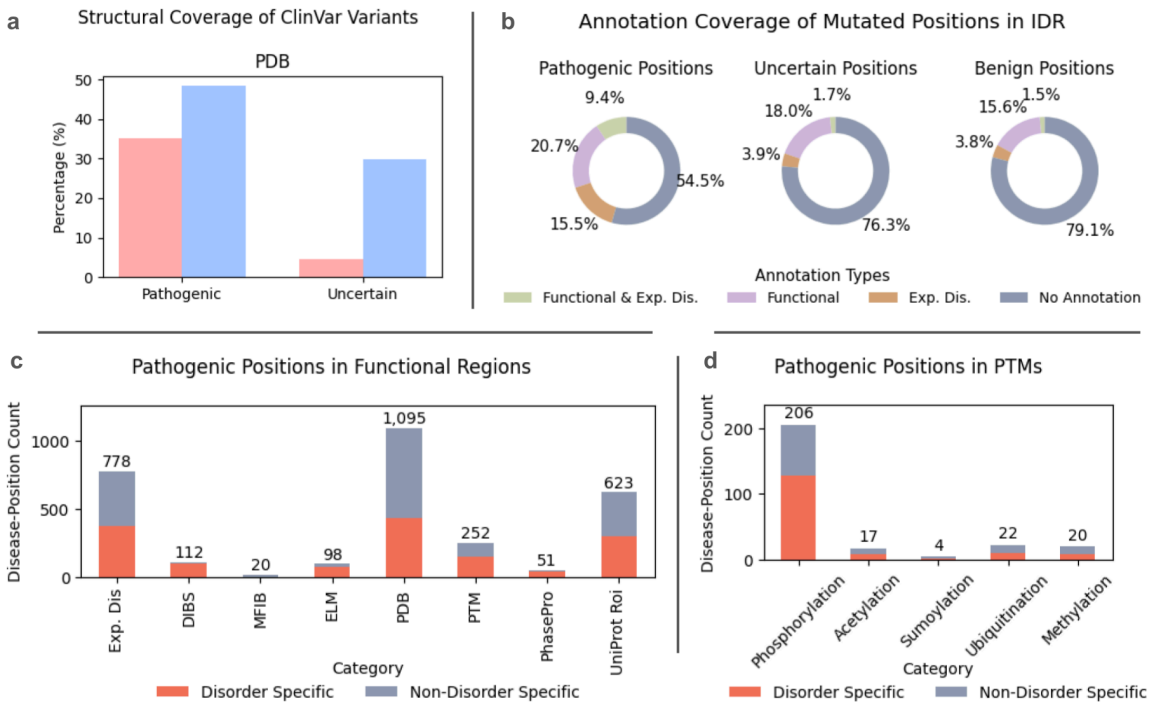


Figure 3) Analysis of Functional Annotation of Variants. (a) Structural Distribution of PDB for Uncertain and Pathogenic Mutated Positions. (b) Functional Distribution of Mutated Disordered Positions. Pathogenic positions exhibit a higher proportion of functional annotations compared to both uncertain and benign positions; however, only 42.8% of these positions possess any functional annotation. (c-d) Functional Distribution of Pathogenic Positions and same for PTM.

AlphaMissense as a Predictive Framework for Identifying Pathogenic and Functional Regions in IDRs

AlphaMissense Captures Functionally Relevant Mutations in Disordered Regions

While pathogenic mutations in ClinVar currently cover less than 0.25% of all possible amino acid positions in the human proteome, pathogenicity predictors enable the assessment of a much broader mutational landscape—including variants that are embryonically lethal or have not yet been observed in humans. Here we systematically evaluated the utility of AlphaMissense for predicting functionally relevant mutations in IDRs. AlphaMissense[18] (AM) is a recent method that leverages a modified AlphaFold2[21] architecture to estimate whether a genetic variant is likely to cause disease. A key advantage of the method is that it was not directly trained on ClinVar variants. AM was shown to outperform other predictors in classifying ClinVar variants and showed the highest correlation with multiplexed assays of variant effects (MAVEs), as cataloged in ProteinGym[22].

AM provides pathogenicity scores for 11 million positions across nearly the entire human proteome and all possible missense variants. However, we noted that AM tended to classify all possible mutations in IDRs as either fully pathogenic or benign, whereas in ordered regions, the distribution was more gradual (Supplementary Figure 2). Based on this, we characterized pathogenicity at the positional level by averaging scores across all 19 possible missense variants and classified positions with scores above 0.5 as pathogenic.

AM scores were generally lower for disordered regions than for globular proteins, with average values of 0.6 and 0.36, respectively (Figure 4a, right panel). We evaluated AM's performance on a non-redundant set of ClinVar variants, comparing its accuracy across ordered and disordered regions (1000 bootstrap, 75% random sampling). At the variant level, AM performed better in structured regions, correctly classifying 86.6% of benign and 81.5% of pathogenic variants, compared to 93.7% and 57.2%, respectively, in IDRs (Figure 4a, left panel). However, this difference diminished at the position level, with balanced accuracy scores of 76–77% for both structured and disordered regions (Figure 4a, center panel).

Across the human proteome, AM predicted 4.7 million positions as pathogenic. However, their distribution differed markedly between ordered and disordered regions (Figure 4b). While 55% of positions in globular regions were predicted to be pathogenic, only 19% of IDR positions met this threshold. Despite this, IDRs contained a disproportionately high fraction (17.8%) of predicted pathogenic positions relative to their representation in ClinVar (8%).

This suggests that a substantially higher proportion of IDR positions are likely to have functional consequences when mutated, exceeding what is currently annotated in ClinVar.

We further assessed AM's accuracy across different functional and structural categories of ClinVar pathogenic positions. Within the IDRs the disorder specific mutations were predicted more accurately (71%) compared to non-disorder specific mutations (62%). In terms of disease complexity, monogenic mutations had the highest prediction accuracy (73%), followed by complex (68%) and multigenic (62%) diseases (Supplementary Figure 3). Moreover, pathogenic mutations located within IDRs with annotations were classified correctly in 75% of the cases, compared to the 62% for non-annotated positions. Experimentally validated disordered regions, particularly those involved in binding interactions (e.g., annotated in DIBS, MFIB, or ELM), were consistently well classified (Figure 4c).

We also examined how AM predictions relate to secondary structure elements, based on AlphaFold2 models. IDR positions predicted to be pathogenic by AM exhibited a higher tendency to adopt regular secondary structures, although coil regions remained dominant (Supplementary Figure 4). However, known pathogenic mutations in coil regions showed the highest misclassification rate, whereas mutations in helices and strands were more accurately predicted (Figure 4 d). Finally, we analyzed evolutionary conservation and found that pathogenic variants in IDRs tend to occur at more conserved positions than benign variants, both within ClinVar-annotated sites and across the proteome (Figure 4e). Altogether, these results support the functional relevance of pathogenic sites predicted by AM within IDRs, particularly those that are evolutionarily conserved and likely to adopt secondary structure elements.

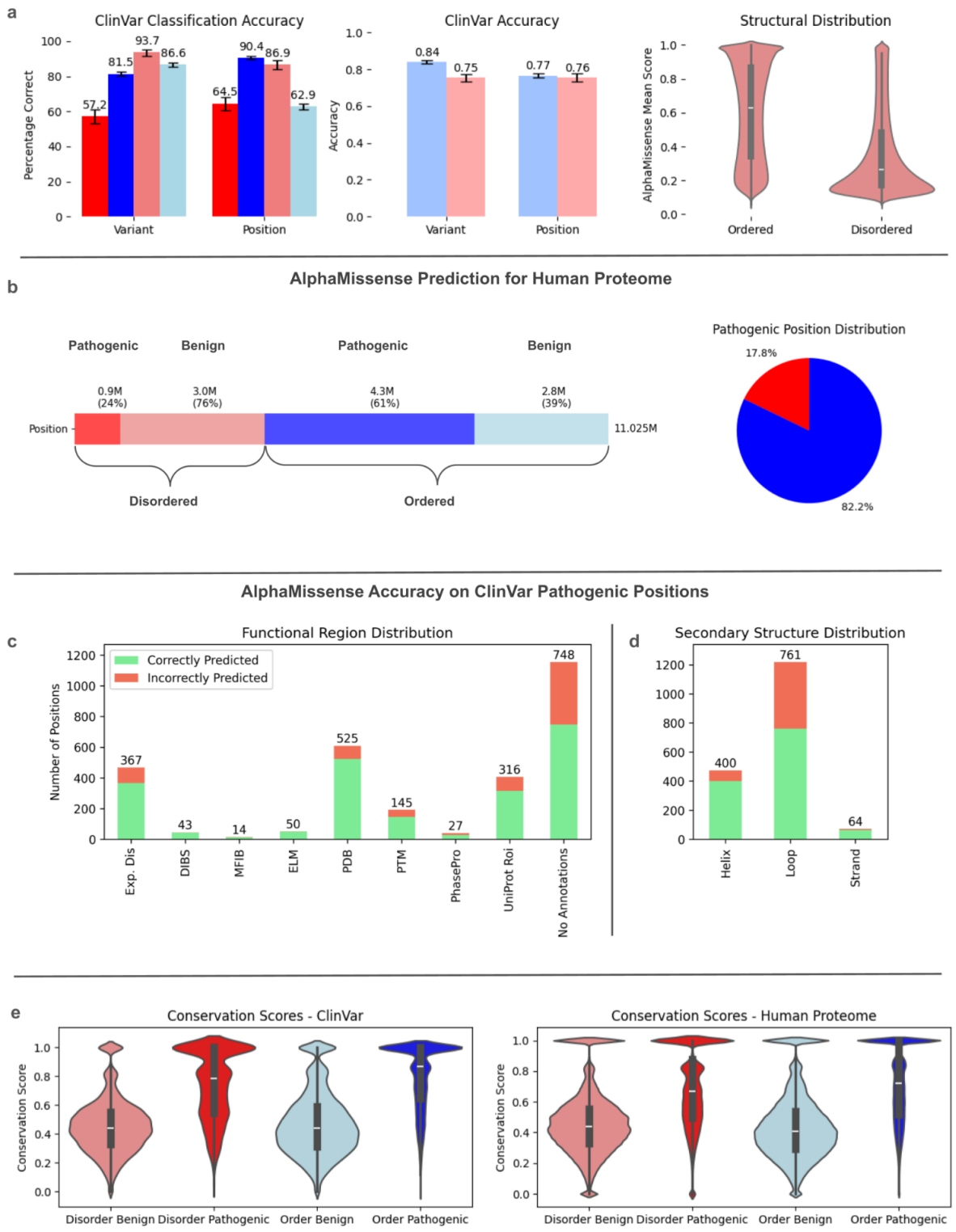


Figure 4) AlphaMissense score Distribution. a) Left: Accuracy of AlphaMissense in classifying ClinVar variants in ordered (blue) and disordered (red) regions, based on variant-level and position-level scores. Light and dark shades indicate correct classification of benign and pathogenic variants, respectively. Middle: Accuracy of non-redundant ClinVar dataset across structural contexts, highlighting reduced predictive performance in IDRs (1000 bootstrap, 75% random sampling). Right: Mean AlphaMissense positional

pathogenicity scores across the human proteome, showing generally lower scores in disordered compared to ordered regions. b) The positional pathogenicity prediction for the human proteome (>0.5 mean pathogenicity is pathogen). AlphaMissense accuracy of pathogenic positions based on function (c) and secondary structure (d), the bar labels show the number of correctly predicted positions. e) Evolutionary Conservation score distribution of pathogenic and benign variants (ClinVar in left and AlphaMissense predicted of Human Proteome in right).

AlphaMissense Captures Functional Specificity and Enables Motif Discovery in IDRs

We further investigated how AM scores vary across functional annotations within IDRs. Compared to experimentally verified disordered regions, we observed higher scores in functionally annotated regions — particularly those that undergo disorder-to-order transitions, such as DIBS and MFIB. In contrast, phase-separated regions generally showed lower scores, likely due to their broader annotations (Figure 5a).

Among functional elements, short linear motifs (SLiMs) stood out by exhibiting a distinctive and localized pathogenicity signal. These motifs displayed consistently higher AM scores than their flanking sequences, forming a characteristic “island-like” pattern not observed in globular domains (Figure 5a). AM also captured internal motif structure, with key residues showing elevated pathogenicity scores relative to non-key residues, underscoring their likely importance in mediating interactions (Supplementary Figure 5a). Many motif classes—especially degradation, ligand-binding, docking, and targeting—showed strong class-specific signal enrichment (Supplementary Figure 6a). To quantify this effect, we computed the score difference between each functional module and its flanking sequence (±10 amino acids). The contrast was strongest for annotated entries in ELM, DIBS, and MFIB databases, suggesting that AM captures localized functional constraints within IDRs. Specifically, the mean score difference was +0.17 across SLiMs (Figure 5b).

Building on this, we assessed whether AM could support motif discovery beyond variant classification. Due to their short and variable sequences, SLiMs are notoriously difficult to detect, and regex-based methods often yield numerous false positives. We refer to these raw, regex-identified motifs that fall within IDRs as RawPEM (Raw Predicted ELM Motifs). To improve these predictions, we trained a decision tree model integrating AM-derived metrics. Although multiple features were considered during training, the algorithm automatically selected two as most informative: (1) the island-like effect—defined as the difference between the motif’s mean AM score and that of its flanking ±10 residue region—and (2) the mean AM score of key residues. A motif was classified as likely functional if either the island-like effect exceeded 0.104, or if the mean score of its key residues surpassed 0.808 in the absence of an island pattern (Figure 5c). Motif matches retained by this decision tree filter are termed AMPEM (AlphaMissense-filtered Predicted ELM Motifs).

Of the 1,526 known motifs in the ELM database, 1,067 (69.9%) were correctly recovered by our classifier, with a mean classification accuracy of 0.747 ± 0.040 (Supplementary Figure 6b). This filtering reduced the initial 1.1 million RawPEM to approximately 192,000 AMPEM (~17%). Among 51 pathogenic positions within annotated ELM motifs, 50 were retained as AMPEMs, corresponding to 31 of 32 disease-associated motifs. This translates to true

positive rates of 98% at the motif level and 97% at the positional level (Figure 5d, Supplementary Figure 6c).

Overall, 46.7% of IDRs contain at least one predicted AMPEM. However, these motifs cover only 15.8% of all disordered positions (Figure 5d, right), suggesting that AMPEM clusters in specific subregions—likely reflecting functionally important elements.

To further reduce false positives, we removed broadly matching ELM classes with more than 1,000 hits across the proteome. This refinement resulted in a high-confidence subset of 8,654 motifs, which we designate CorePEM (High-Confidence Predicted ELM Motifs). The CorePEM set retained 62% of the original ELM classes, with LIG-type motifs being the most prevalent (87 classes), followed by DOC (12), DEG (10), and TRG (9) (Supplementary Figure 5b–c).

We next investigated whether CorePEMs are subject to stronger evolutionary constraints by analyzing the distribution of benign variants. Key residues within CorePEMs showed a significantly lower benign mutation rate (0.64%) compared to all regex matched motif key residues (1.58%); Fisher's exact test $p = 3.518\times10^{-43}$, odds ratio = 0.38) (Supplementary Figure 6d).

To further validate our predictions, we compared them to SLiMPrint results. Of the 92 high-confidence motifs identified by SLiMPrint, 87% (80) were recovered as AMPEMs. However, only five of these overlapped disease-associated mutations, suggesting strong agreement in motif prediction but limited intersection with known pathogenic variants. We also benchmarked our results against a recent peptide-phage display study assessing how IDR mutations affect protein interactions [12]. Among 64 motifs where mutations disrupted binding, 38 (59%) were correctly identified as AMPEM—further supporting the functional relevance of our predictions. Collectively, these results demonstrate that our classifier substantially reduces false positives while retaining a large fraction of biologically meaningful motifs.

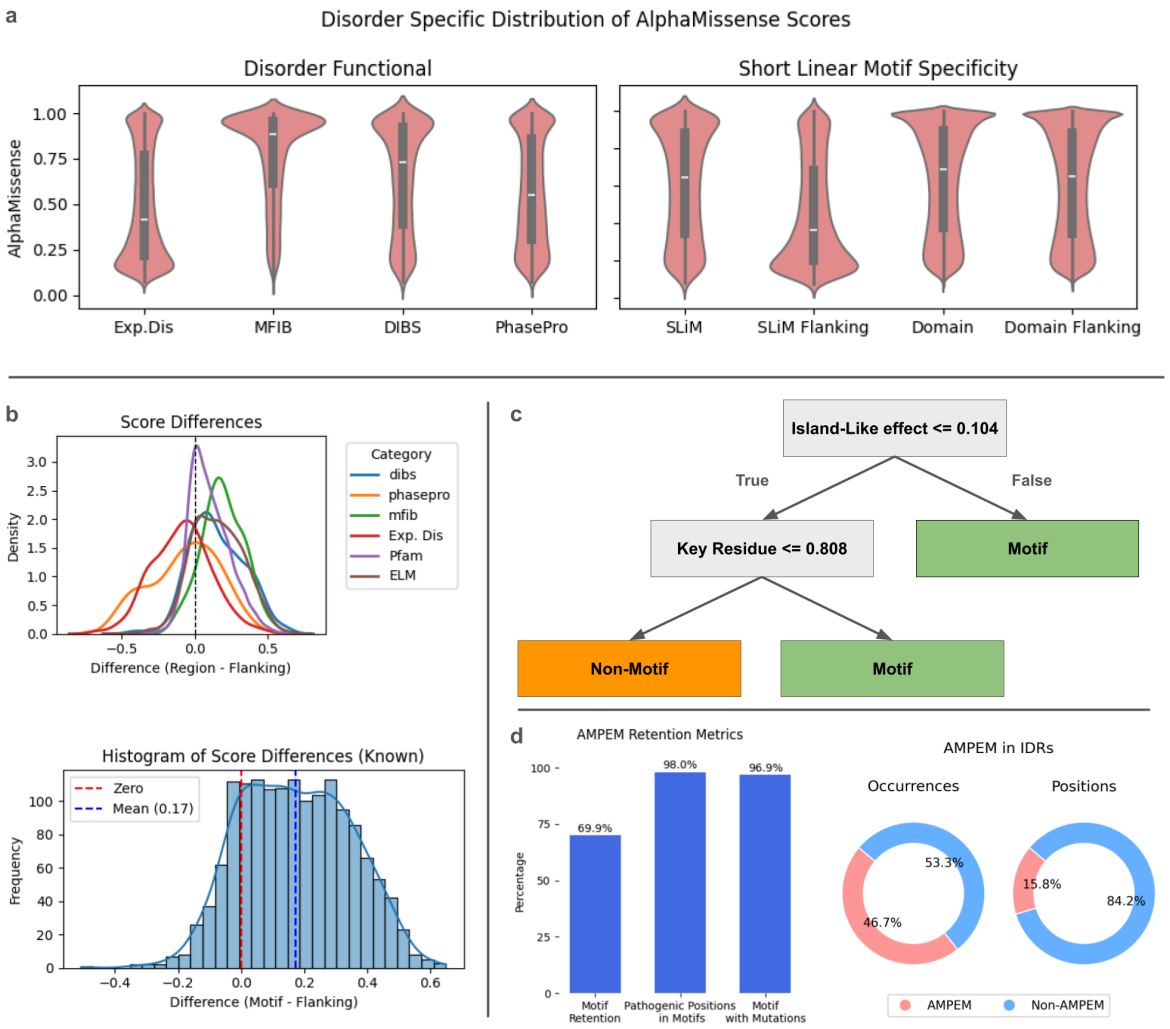


Figure 5) Predicting SLiM Regions with AlphaMissense. (a) AlphaMissense score distribution for functional regions in IDRs. The right panel shows the sequential environment scores for ELM and Pfam domains (sequential length is maximised 10 amino acid long segment) (b) Special Property of AlphaMissense for SLiMs. The upper panel shows the sequential differences between various annotations. The lower panel presents a histogram illustrating the difference in AlphaMissense scores between motifs and their surrounding sequences. c) AMPEM (AlphaMissense-filtered Predicted ELM Motifs) - Decision Tree model. The tree considers the motif as a real hit if the Island-like effect is =>0.104 or the key residue AlphaMissense positional score is =>0.808. d) Metrics for AMPEM. Left: Retention of known motifs, pathogenic mutations, and motifs containing pathogenic variants. Right: Proportion of the human proteome's IDRs covered by predicted motifs.

Prioritizing Predicted ELM Motifs for Clinical Interpretation of Disease-Associated Mutations

Having established the biological relevance of the CorePEM dataset, we next prioritized motifs with potential clinical impact. We focused on ClinVar variants classified as pathogenic or of uncertain significance, selecting motifs that either (i) showed a ≥3-fold enrichment of uncertain mutations relative to the protein sequence, or (ii) contained mutations exclusively

within the motif and not elsewhere in the protein. Variants that were not associated with specific diseases, such as “Inborn Genetic Disease” or “not provided” were excluded.

This filtering yielded 228 HotspotPEM (clinically prioritized, mutation-enriched CorePEM): 67 containing pathogenic mutations (24 known, 43 novel) and 161 with uncertain mutations (29 known, 132 novel). Consistently across categories, the majority of mutations were disorder-specific, supporting the hypothesis that disruption of functional motifs within IDRs is a key mechanism of disease association (Figure 6b, top panels). HotspotPEM with enriched pathogenic mutations were predominantly associated with neurodevelopmental and neurodegenerative conditions. Notably, several entries grouped under the “Other” category in the disease ontology plot also correspond to these functional classes—specifically, rare monogenic disorders with developmental, immunological, or neuromuscular phenotypes—highlighting the broader clinical relevance of motif disruption in IDRs across a diverse spectrum of predominantly rare but medically significant diseases (Figure 6b, bottom panel).

We also examined which genes and ELM classes were most frequently affected by pathogenic HotspotPEM (Figure 6c). Several genes—including HIVEP2, KAT6B and AFF3—exhibited strong enrichment of pathogenic mutations within HotspotPEM, but did not overlap previously annotated motifs, highlighting their potential clinical relevance. Among ELM classes, LIG-type motifs were the most frequently observed, but other types of motifs also occurred. These results underscore the functional diversity of disease-associated HotspotPEM and point to specific classes of short linear motifs as recurrent targets of pathogenic variation.

To facilitate further exploration of these prioritized motifs, we provide two curated resources: the “HotspotPEM mutation analysis” dataset (Supplementary Data 2), highlighting known and predicted motif-disrupting mutations, and the full CorePEM list (Supplementary Data 3), cataloging all predicted motifs within IDRs. These resources offer a valuable framework for interpreting ClinVar variants in disordered contexts and for linking sequence-level disruptions to molecular dysfunction.

Finally, to illustrate how HotspotPEM can mechanistically inform disease biology, we present selected examples in the following section, highlighting both known and novel motifs whose disruption is linked to pathogenic or uncertain mutations.

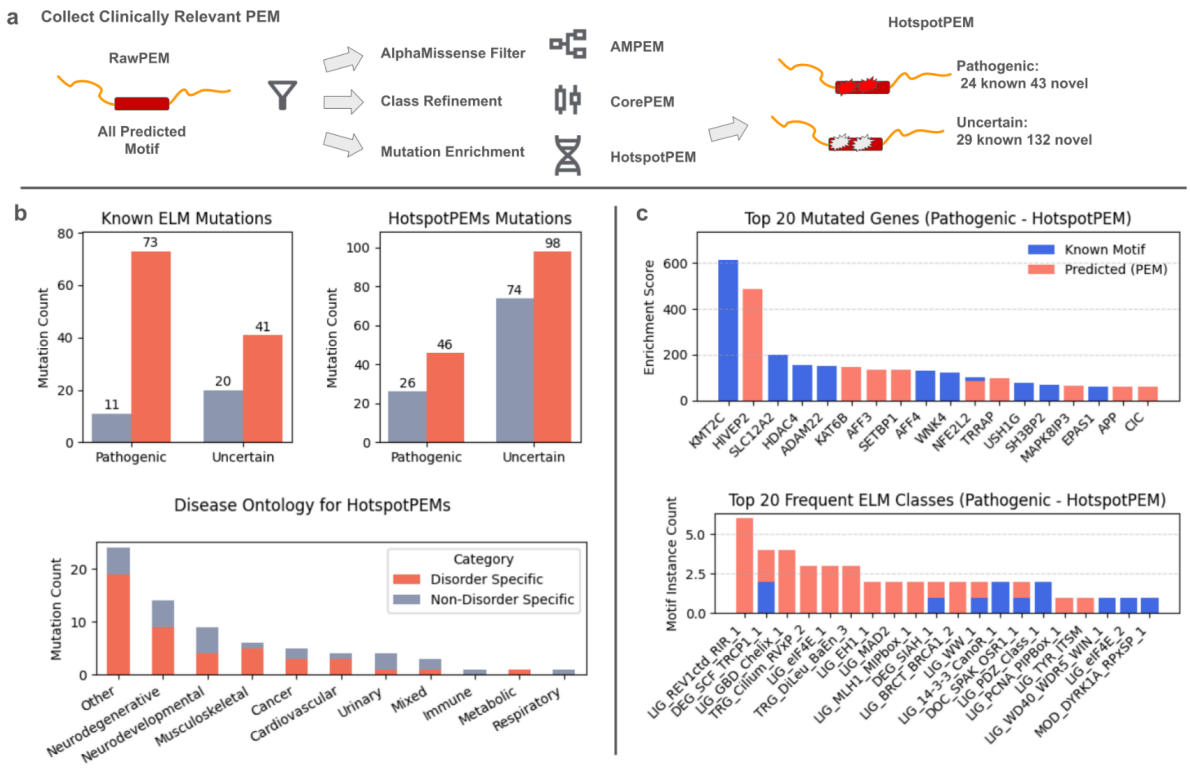


Figure 6) Identifying Pathogenic mutations in Motifs a) Overview of collecting HotspotPEM (clinically prioritized, mutation-enriched CorePEM) b) In the top site of the figure we show Pathogenic and Uncertain mutations in ELM and HotspotPEM. In the bottom site of the figure we show Disease Ontology of Pathogenic Variants within Predicted ELM Motifs. c) Top mutated genes of enriched pathogenic HotspotPEM and the most frequent ELM classes labeled the already known and predicted motifs.

Mechanistic Interpretation of Disease Mutations in IDRs through Predicted Motif Integration

Actin-Associated Motif Disruption in Nemaline Myopathy

In LMOD3, a protein critical for sarcomeric actin filament organization, pathogenic variants associated with nemaline myopathy were found within a previously unannotated but predicted actin-binding motif (LIG_Actin_WH2_1) (Figure 7a). This motif spans residues 535–553 in the C-terminal disordered region and aligns with LMOD3's known role in actin polymerization and filament stabilization. The existance of two pathogenic mutations (R543L, L550F) with this motif suggests that disrupted actin binding may impair filament dynamics and contribute to disease pathology. We propose that pathogenic mutations at this site impair actin binding and filament stabilization, offering a mechanistic explanation for the disease phenotypes observed in LMOD3-related nemaline myopathy. This example illustrates how HotspotPEMs can reveal previously unannotated yet functionally important elements within clinically relevant disordered regions.

Predicted SIAH Degron Motifs Are Enriched in Disease-Associated Variants

Predicted degron motifs recognized by the SIAH1 E3 ubiquitin ligase emerged as a notable subclass of HotspotPEMs enriched in disease-associated mutations. These short linear motifs mediate targeted protein degradation and were identified in four disordered proteins using our HotspotPEMs pipeline: AFF2, AFF3, AFF4, and FOXP2.

AFF3 and AFF4, two paralogous transcriptional regulators implicated in KINSSHIP and CHOPS syndromes [23], respectively, showed strong enrichment of pathogenic mutations at predicted degron sites. AFF2, another paralog, harbored uncertain variants at the predicted degron, suggesting a related degradation mechanism may be disrupted in non-syndromic X-linked intellectual disability.

In FOXP2, uncertain missense variants clustered at a predicted SIAH degron motif located within its C-terminal IDR. FOXP2 is a nuclear transcriptional repressor involved in chromatin binding and RNA polymerase II regulation, particularly during cortical development. Disruption of this motif may impair FOXP2 degradation, leading to protein accumulation and persistent repression of neuronal genes — a potential mechanism underlying language disorder phenotypes linked to uncertain variants.

To assess structural compatibility, we used AlphaFold Server[24] to model interactions between the predicted degrons and SIAH1. In all four proteins, the degron is aligned with a high-confidence binding interface, supporting its functional plausibility. In the FOXP2–SIAH1 complex, introducing the T451M variant reduced prediction confidence at the canonical degron and suggested a potential shift toward an alternate interface (residues 426–432).

Together, these results suggest that dysregulation of degron motifs may contribute to neurodevelopmental diseases through impaired protein turnover. Our predictions extend known SIAH targets to new paralogs and functionally important regulators such as FOXP2, offering mechanistic hypotheses for both pathogenic and uncertain variants.

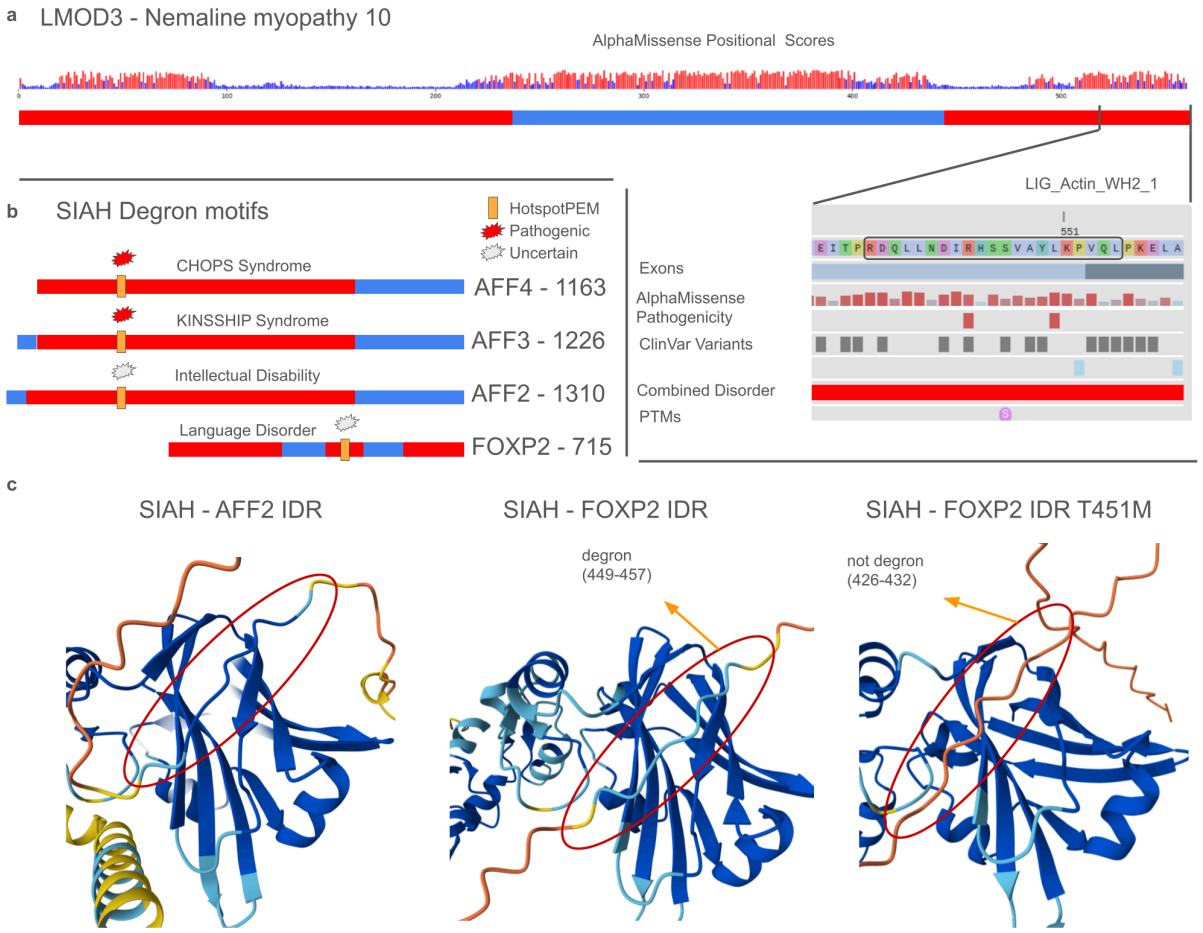


Figure 7) Functional Motif Disruption in IDRs Highlights Pathogenic Variants in Neuromuscular and Neurodevelopmental Diseases. This figure illustrates two examples of how predicted functional motifs within IDRs can highlight disease-relevant mutations. a) In LMOD3, a predicted actin-binding WH2 motif overlaps with two pathogenic mutations (R543L, L550F) linked to nemaline myopathy. DisCanVis visualization shows the motif, ClinVar mutations, and AlphaMissense pathogenicity scores, supporting a role for disrupted actin binding in disease. b-c) Predicted SIAH degron motifs in AFF2, AFF3, AFF4, and FOXP2 are associated with pathogenic and uncertain ClinVar variants and supported by AlphaFold3 multimeric structure predictions. The identified PEMs are located in disordered regions of these regulators. b) Pathogenic mutations in AFF3 and AFF4 are linked to KINSSHIP and CHOPS syndromes, respectively, while AFF2 harbors uncertain variants associated with syndromic X-linked intellectual disability. In FOXP2, uncertain variants coincide with the predicted degron and are implicated in specific language disorder. c) AlphaFold3 multimer predictions confirm high-confidence interactions between the degron motifs and SIAH in wild-type sequences. Notably, in FOXP2, the T451M variant—a variant of uncertain significance—disrupts this interaction by shifting the binding away from the degron site and reducing confidence in the predicted complex. This suggests a potential disease mechanism involving impaired degradation and altered protein turnover.

Predicted PCNA and REV1 Interaction Motifs Link Disordered POLK Regions to Cancer Mutations

POLK (UniProt: Q9UBT6, DisCanVis: POLK-201) is a Y-family DNA polymerase involved in translesion synthesis (TLS), a DNA damage tolerance pathway essential for maintaining genome integrity under replicative stress (Yadav et al., 2015). While POLK's role in cancer is increasingly recognized, the functional impact of many cancer-associated missense mutations—particularly those outside the catalytic domain—remains unclear.

Previous work identified a REV1-binding RIR motif in POLK (residues 565–576; PDB: 4GK5) and proposed a noncanonical C-terminal PIP motif, which required flanking residues for stable interaction with PCNA (Hishiki et al., 2009). Using our HotspotPEMs pipeline, we detected a previously unannotated PIP-like motif (residues 526–535) immediately downstream of the catalytic domain. AlphaFold-based docking supports high-confidence binding of this motif to PCNA's canonical site, and this interaction is favored even when a longer region containing both motifs is modeled (Figure 8b) . We also identified a C-terminal RIR-like motif (residues 866–870) that overlaps the previously proposed PIP site. This motif can bind REV1 at an alternative site from the known RIR motif and continues to prefer REV1 (Figure 8c) even in multimer docking with both PCNA and REV1 (Figure 8d). These findings suggest that the newly identified PIP-like motif is the primary PCNA-binding site, whereas the C-terminal motif may act as a less specific or context-dependent interface. The overlapping sequence characteristics and binding flexibility of PIP- and RIR-like motifs underscore the limitations of current motif definitions in disordered regions.

Both newly predicted interaction motifs coincide with pathogenic prostate cancer mutations, providing a structural basis to interpret previously uncharacterized variants. We propose that mutations in these regions disrupt TLS complex assembly—either by impairing PCNA docking (526–535) or interfering with REV1-mediated recruitment (866–870)—ultimately compromising DNA damage bypass and promoting genomic instability. However, the low information content and overlapping nature of PIP-like motifs make it challenging to assign precise interaction roles.

These findings extend previous work by offering structural and functional hypotheses for the contribution of disordered interaction motifs to TLS dysfunction. While further experimental validation and refinement of motif definitions are needed—particularly for the C-terminal region—this case illustrates how PEM predictions, when filtered through AlphaMissense scores, can help uncover mechanistic explanations for cancer-associated mutations in intrinsically disordered regions.

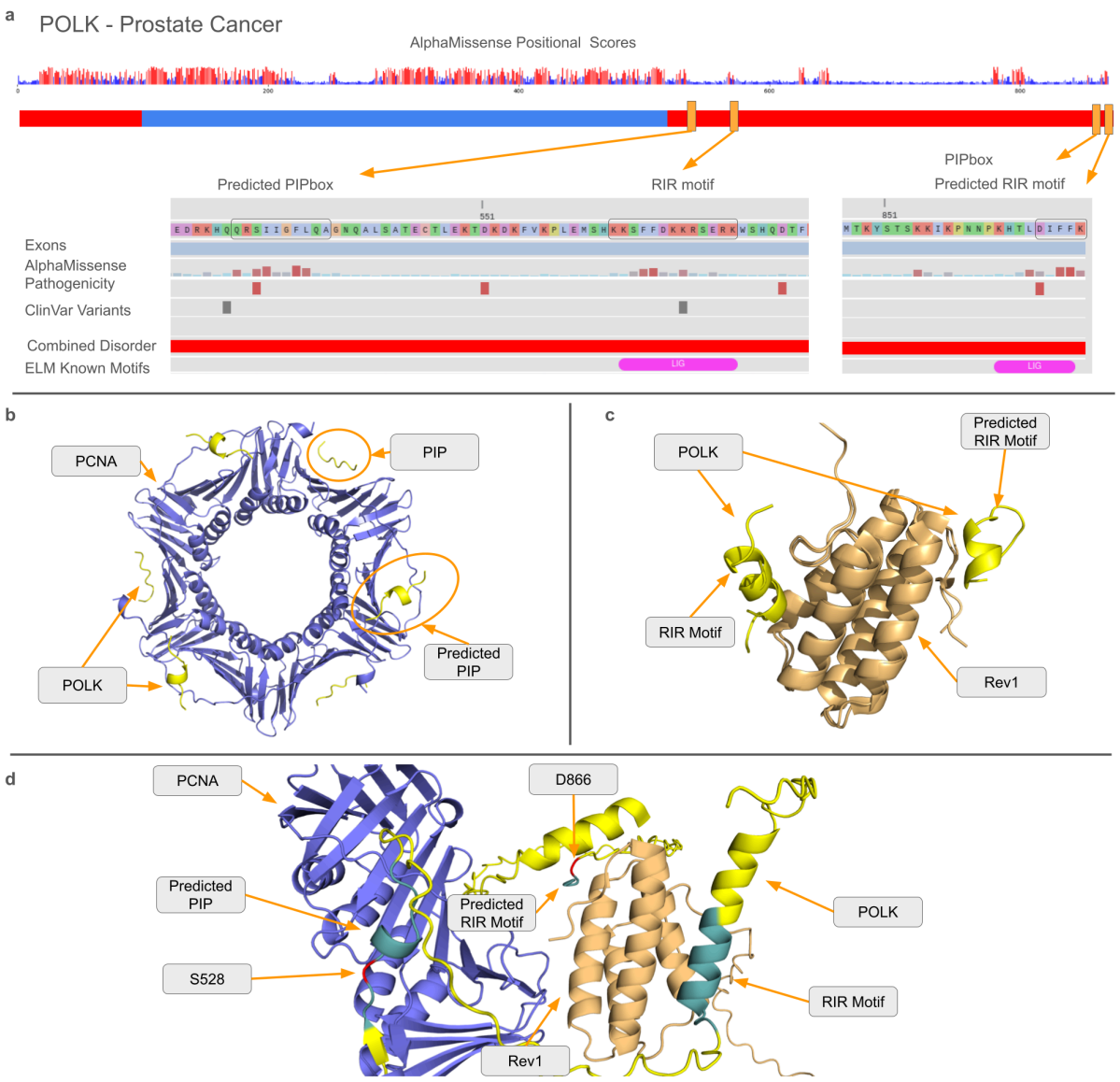


Figure 8) Cancer-Linked Motif Disruption in Disordered Regions of POLK Reveals Potential DNA Repair Defects. This figure illustrates a HotspotPEM within the IDR of *POLK*, a polymerase implicated in prostate cancer. a) Visualization from the DisCanVis webserver, showing the predicted PCNA-interacting protein (PIP) box and Rev1-interacting region (RIR) motifs within IDRs, alongside mapped ClinVar missense mutations and AlphaMissense pathogenicity scores. b) AlphaFold3 multimer model depicting POLK bound to PCNA, showing both the predicted and annotated PIP-box motifs. c) Structural superposition of the experimentally determined Rev1–POLK complex (PDB:4GK5) onto the AlphaFold3 multimer prediction, showing Rev1's C-terminal domain interacting with the known motif (residues 565–576) and the predicted RIR motif (residues 866–870). d) Full multimeric model including full-length POLK, full-length PCNA, and the Rev1 C-terminal domain. Predicted motifs are color-coded, and pathogenic missense mutations at residues 528 and 866 are highlighted in red.

Discussion

Intrinsically disordered regions (IDRs) constitute over one-third of the human proteome, yet remain significantly underrepresented in clinical variant annotations, accounting for only ~8% of pathogenic mutations in ClinVar. Traditionally, pathogenic variant studies had largely focused on globular domains, whose well-defined structures have made them more amenable to experimental and computational analyses. As a result, a critical gap persists in our understanding of how mutations in disordered regions contribute to diseases. Our study addresses this gap by examining missense mutations within IDRs through integrating structural disorder predictions with experimental and functional annotations. We demonstrate that missense mutations within IDRs — particularly those linked to monogenic diseases — often coincide with experimentally validated disordered segments or overlap with functional elements such as short linear motifs (SLiMs). Nonetheless, more than half of these mutations lack any functional annotation, underscoring the limitations of current resources and the strong need for improved strategies to characterize functional sites within disordered regions.

A central strength of our study is the application of AlphaMissense (AM), a deep learning–based pathogenicity predictor that was not trained on ClinVar data and therefore provides an unbiased assessment. We found that AM performs well not only for structured regions, but —perhaps unexpectedly— also in disordered context, effectively highlighting regions of functional importance. Furthermore, we also reveal a previously unexploited layer of information: its ability to identify “island-like” pathogenicity patterns in SLiMs. These patterns, where pathogenicity scores sharply peak within the motif relative to their surrounding regions, mirror conservation-based signals observed in evolutionary studies and suggest localized functional relevance[25]. We have shown that his feature can be used to enhance the prediction of binding regions in IDRs [26].

Building on this insight, we developed a decision-tree classifier that leverages these AM-derived patterns to identify predicted ELM motifs (PEMs) across the human proteome. Remarkably, many PEMs overlap with known pathogenic mutations or variants of uncertain significance (VUS), revealing their potential disease relevance. Moreover, residues within PEMs—particularly those annotated as key functional positions—are significantly depleted in benign mutations, indicating purifying selection and functional constraint.

The clinical utility of PEMs is further highlighted by several illustrative case studies. In POLK, pathogenic mutations were found at predicted PCNA and REV1 interaction motifs within its disordered regions, offering mechanistic explanations for their contribution to prostate cancer. Similarly, in FOXP2 and LMOD3, mutations within predicted degron and actin-binding motifs, respectively, align with neurodevelopmental and muscular disease phenotypes. These examples underscore how predicted motifs can connect sequence-level variants to broader biological dysfunction, particularly in diseases lacking structural context.

Overall, our findings underscore the need for a distinct interpretative framework when analyzing missense mutations in intrinsically disordered regions, which behave fundamentally differently from structured domains. Tools like AlphaMissense, when combined with motif prediction and visualization platforms such as DisCanVis, provide an integrated framework for identifying and prioritizing pathogenic mutations in disordered

regions. The functionally constrained elements identified by PEMs within capture a distinct mechanistic layer compared to regions involved in LLPS [20]. This complementarity reinforces the need to consider multiple functional dimensions when interpreting pathogenic variants in disordered regions. Altogether, this strategy not only enhances our understanding of protein function and disease mechanisms, but also opens new avenues for the clinical interpretation of variants of uncertain significance (VUS), which remain a major bottleneck in genetic diagnostics.

Methods

Transcript Mapping and Annotation

We mapped GENCODE v44[27] transcripts to the corresponding canonical protein sequences from UniProt[28] (release 2024_01), following the procedure previously implemented in DisCanVis[29]. For all downstream analyses, we included only the main isoforms to ensure consistency across the dataset. All structural, functional, and pathogenicity annotations were assigned to these canonical sequences.

Combined Disorder Model

To define the disordered state of protein positions, we combined experimental and computational data. Experimentally verified IDRs were obtained from MobiDB[30] (categories: "homology-disorder-merge" and "curated-disorder-merge"). For proteins without experimental data, disorder was predicted using AlphaFold2-derived[21] RSA values and IUPred3[31] scores.

Pfam[32] domains were mapped to the proteome, and each domain was labeled as disordered or ordered based on whether ≥50% of its residues were disordered. In disordered domains, all positions were considered disordered. In ordered domains, only terminal residues verified as disordered were retained. The final disorder state at each position was assigned hierarchically: experimental > AlphaFold2-RSA (>0.582) > IUPred3 (>0.4). Positions located in coiled-coil regions (DeepCoil[33], threshold: 0.5) and collagen genes were excluded.

Functional and Structural Annotations

Functional annotations were sourced from the second version of DisCanVis and mapped to Gencode v44. We focused on disordered-specific annotations, post-translational modifications (PTMs), and curated regions from UniProt. UniProt's "Region of Interest" entries were filtered to exclude general structural descriptors.

Disordered binding regions were extracted from MFIB [34] and DIBS [35] databases; SLiMs were obtained from the ELM database [36]. Phase-separating regions were derived from PhasePro [37]. PTM annotations were compiled from dbPTM [38] and PhosphoSitePlus [39]. PDB entries were mapped to Gencode transcripts, and missing residues were excluded. Secondary structure assignments from AlphaFold2 were computed using the 4th version of DSSP [40] .

ClinVar Variant Processing

Missense variant data were obtained from ClinVar (VCFv4.1, release date: 2023-08-13, GRCh38) and mapped to Gencode v44 transcripts. Only non-synonymous single-nucleotide variants (SNVs) were retained, excluding those at the first codon position to avoid potential mapping ambiguities. Variants were filtered to include only those labeled as "Pathogenic," "Likely pathogenic," or "Benign," with at least one ClinVar star rating.

Variants annotated as "Uncertain significance," "Conflicting interpretations of pathogenicity," or "Not provided" were grouped together as variants of uncertain significance (VUS) for downstream analyses.

Disease Ontology Classification

Disease ontology assignments were extracted from the MONDO Disease Ontology database[41]. Diseases were classified into 16 major organ/system-specific categories, such as Cancer, Cardiovascular, Neurodevelopmental, and Musculoskeletal. If a disease matched multiple categories, it was assigned to an organ-specific group when possible; otherwise, it was classified as "Mixed." Any disease with "Cancer" in its ontology path was prioritized into the Cancer group. Diseases without mapped terms were assigned to the "Unknown" category.

These groups were used to assess genetic complexity by classifying diseases as Monogenic (1 gene), Multigenic (2–4 genes), or Complex (>4 genes).

AlphaMissense

AlphaMissense scores were downloaded from the original publication and mapped to the transcripts. Scores from the first codon position were excluded. A score threshold of 0.564 was used to classify positions as pathogenic based on the original paper.

To evaluate predictive accuracy, ClinVar variants were filtered to reduce redundancy using DIAMOND[42] (40% sequence identity). Proteins with equal numbers of pathogenic and benign variants were selected. Variants were classified as pathogenic or benign using the 0.567 threshold. For positional scoring, AlphaMissense scores were averaged across residues, and a threshold of 0.5 was used for binary classification.

Short Linear Motif Prediction (PEM)

A custom motif prediction pipeline was developed using ELM regex patterns to scan the human proteome. For each motif instance, we extracted the following features:

- The mean AlphaMissense (AM) score for the motif
- The mean AM score for the sequential motif region (up to 10 amino acids)
- The mean AM scores for key residues versus non-key residues
- The maximum AM residue mean score within the motif region
- The difference between sequential and motif AM scores

- The difference between key and non-key AM scores

To capture potential “island-like” pathogenic zones, we computed the difference between the motif’s mean AM score and that of its sequential region, as well as the difference between key-residue and non-key-residue AM scores. We focused on ELM classes predominantly located within intrinsically disordered regions (IDRs) and retained only those motif instances for which all key metrics were available. Additionally, we required that at least 80% of the motif sequence fall within disordered regions, as defined by our Combined Disorder model. This filtering yielded 1.1 million predicted motifs and 1,523 known motif sets.

For classification, we treated known motifs as the positive set and predicted motifs as the negative set. We trained a decision tree using scikit-learn’s DecisionTreeClassifier (maximum depth of 3, entropy-based splitting, and balanced class weights). We then applied 10-fold cross-validation with 1,000 bootstrap samples, each time selecting a random subset of predicted motifs as the negative set. An 80/20 train-test split was employed during each iteration.

For motif prioritization, we calculated local variant density within each predicted motif and compared it to the background density across the entire protein. Motifs were retained if their density was at least threefold higher than the background or if variants were exclusive to the motif region.

Motif Validation Datasets

Peptide Interaction Dataset Validation

To evaluate the functional relevance of PEMs, we utilized a dataset from a recent study[12], which included experimentally tested domain–peptide interactions. The dataset contains peptide pairs (wild-type and mutant) mapped to interaction domains, with annotations on whether the mutation affected binding. We selected peptides containing regex-defined SLiMs and mapped them to our reference proteome. Motifs were filtered based on disorder using our Combined Disorder model, and each instance was scored using our decision-tree–based PEM predictor to assess whether it would be classified as a motif.

SLiMPrint validation

We also validated our predictions using SLiMPrint, a conservation-based method for identifying candidate SLiMs. Human motif instances were downloaded and mapped to our proteome and filtered based on the Combined Disorder model. We evaluated all mapped SLiMs as well as a high-confidence subset annotated with “Good,” “Strong,” “Motif,” or “Ok” scores. Each motif was then passed through our PEM prediction pipeline to determine classification.

Benign Mutation Rate Estimation

To assess selective constraint, we calculated the benign mutation rate in annotated and unannotated IDRs. This was restricted to proteins containing benign variants. The mutation rate was defined as the proportion of disordered positions harboring at least one benign

variant in each functional category (e.g., ELM, MFIB, PhasePro, experimental IDRs, and unannotated IDRs). For PEMs, we applied the same approach focusing on refined PEM and all predicted key residues.

Data Availability

The datasets used in this study are publicly available from the following sources:

- **GENCODE v44** transcript annotations were obtained from the GENCODE database: https://ftp.ebi.ac.uk/pub/databases/gencode/Gencode_human/release_44/
- **ClinVar missense variant data** (release 2023-08-13, GRCh38) was downloaded from the NCBI FTP archive: https://ftp.ncbi.nlm.nih.gov/pub/clinvar/vcf_GRCh38/archive_2.0/2023/clinvar_20230813.vcf.gz
- **AlphaMissense variant effect scores** were retrieved from the DeepMind repository: https://storage.googleapis.com/dm_alphamissense/AlphaMissense_aa_substitutions.tsv.gz
- **Short linear motif annotations (ELM)** were accessed via the ELM database: http://elm.eu.org/downloads.html#instances
- **MFIB** (Mutual Folding Induced by Binding) and **DIBS** (Disordered Binding Sites) datasets were downloaded from their respective databases: https://mfib.pbrg.hu/downloads.php https://dibs.enzim.ttk.mta.hu/downloads.php
- **Experimental Disorder regions** were collected from **MobiDB** (accessed on 2024-04-11): https://mobidb.org/statistics?proteome=UP000005640
- **Disease ontology information** was obtained from the **MONDO Disease Ontology** project: https://mondo.monarchinitiative.org/pages/download/
- Shown examples was extracted from the second version of **DisCanVis**: https://v2.discanvis.elte.hu

Supplementary Data

Supplementary Data 1: Pathogenic Variants in IDRs

Supplementary Data 2: Variants in HotspotPEM

Supplementary Data 3: CorePEM dataset

Funding

N.D. acknowledges “Supported by the EKÖP-KDP-24 University Excellence Scholarship Program, the Cooperative Doctoral Program of the Ministry for Culture and Innovation, and the National Research, Development and Innovation Fund.”

E.G. acknowledges “Supported by the EKÖP-24 University Excellence Scholarship Program of the Ministry for Culture and Innovation, funded by the National Research, Development, and Innovation Fund.”

Z.D. acknowledges “HORIZON-MSCA-2023-SE - Grant Agreement 101182949 “IDPfun2”.

This work was funded by the European Union. “HORIZON WIDERA 2023 IDP2Biomed - Grant Agreement 101160233”.

References

1. Landrum, M. J. *et al.* ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic Acids Res* **46**, D1062–D1067 (2018).
2. Van Hout, C. V. *et al.* Exome sequencing and characterization of 49,960 individuals in the UK Biobank. *Nature* **586**, 749–756 (2020).
3. Karczewski, K. J. *et al.* The mutational constraint spectrum quantified from variation in 141,456 humans. *Nature* **581**, 434–443 (2020).
4. van der Lee, R. *et al.* Classification of intrinsically disordered regions and proteins. *Chem. Rev.* **114**, 6589–6631 (2014).
5. Pajkos, M., Zeke, A. & Dosztányi, Z. Ancient Evolutionary Origin of Intrinsically Disordered Cancer Risk Regions. *Biomolecules* **10**, (2020).
6. Tompa, P., Davey, N. E., Gibson, T. J. & Babu, M. M. A million peptide motifs for the molecular biologist. *Mol Cell* **55**, 161–169 (2014).
7. Holehouse, A. S. & Kragelund, B. B. The molecular basis for cellular function of intrinsically disordered protein regions. *Nat Rev Mol Cell Biol* **25**, 187–211 (2024).
8. Van Roey, K. *et al.* Short linear motifs: ubiquitous and functionally diverse protein interaction modules directing cell regulation. *Chem Rev* **114**, 6733–6778 (2014).
9. Wright, P. E. & Dyson, H. J. Intrinsically disordered proteins in cellular signalling and regulation. *Nat. Rev. Mol. Cell Biol.* **16**, 18–29 (2015).
10. Meyer, K. *et al.* Mutations in Disordered Regions Can Cause Disease by Creating

Dileucine Motifs. *Cell* **175**, 239–253.e17 (2018).

11. Rrustemi, T. *et al.* Pathogenic mutations of human phosphorylation sites affect protein-protein interactions. *Nat Commun* **15**, 3146 (2024).
12. Kliche, J. *et al.* Proteome-scale characterisation of motif-based interactome rewiring by disease mutations. *Mol Syst Biol* **20**, 1025–1048 (2024).
13. Chen, J.-C. *et al.* WNK4 kinase is a physiological intracellular chloride sensor. *Proc Natl Acad Sci U S A* **116**, 4502–4507 (2019).
14. Banani, S. F., Lee, H. O., Hyman, A. A. & Rosen, M. K. Biomolecular condensates: organizers of cellular biochemistry. *Nat Rev Mol Cell Biol* **18**, 285–298 (2017).
15. Ambadipudi, S., Biernat, J., Riedel, D., Mandelkow, E. & Zweckstetter, M. Liquid-liquid phase separation of the microtubule-binding repeats of the Alzheimer-related protein Tau. *Nat Commun* **8**, 275 (2017).
16. Garcia-Cabau, C. *et al.* Mis-splicing of a neuronal microexon promotes CPEB4 aggregation in ASD. *Nature* **637**, 496–503 (2025).
17. Mészáros, B., Hajdu-Soltész, B., Zeke, A. & Dosztányi, Z. Mutations of Intrinsically Disordered Protein Regions Can Drive Cancer but Lack Therapeutic Strategies. *Biomolecules* **11**, (2021).
18. Cheng, J. *et al.* Accurate proteome-wide missense variant effect prediction with AlphaMissense. *Science* **381**, eadg7492 (2023).
19. Luppino, F., Lenz, S., Chow, C. F. W. & Toth-Petroczy, A. Deep learning tools predict variants in disordered regions with lower sensitivity. *BMC Genomics* **26**, 367 (2025).
20. Feng, M. *et al.* Decoding Missense Variants by Incorporating Phase Separation via Machine Learning. *Nat Commun* **15**, 8279 (2024).
21. Jumper, J. *et al.* Highly accurate protein structure prediction with AlphaFold. *Nature* **596**, 583–589 (2021).
22. Notin, P. *et al.* ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. *bioRxiv* (2023) doi:10.1101/2023.12.07.570727.
23. Inoue, Y. *et al.* Three KINSSHIP syndrome patients with mosaic and germline AFF3

variants. *Clin Genet* **103**, 590–595 (2023).

24. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
25. Davey, N. E. *et al.* SLiMPrints: conservation-based discovery of functional motif fingerprints in intrinsically disordered protein regions. *Nucleic Acids Res* **40**, 10628–10641 (2012).
26. Erdős, G., Deutsch, N. & Dosztányi, Z. AIUPred - Binding: Energy Embedding to Identify Disordered Binding Regions. *J Mol Biol* 169071 (2025).
27. Frankish, A. *et al.* GENCODE 2021. *Nucleic Acids Res.* **49**, D916–D923 (2021).
28. UniProt Consortium. UniProt: the Universal Protein Knowledgebase in 2025. *Nucleic Acids Res* **53**, D609–D617 (2025).
29. Deutsch, N., Pajkos, M., Erdős, G. & Dosztányi, Z. DisCanVis: Visualizing integrated structural and functional annotations to better understand the effect of cancer mutations located within disordered proteins. *Protein Sci.* **32**, e4522 (2023).
30. Piovesan, D. *et al.* MOBIDB in 2025: integrating ensemble properties and function annotations for intrinsically disordered proteins. *Nucleic Acids Res* **53**, D495–D503 (2025).
31. Mészáros, B., Erdos, G. & Dosztányi, Z. IUPred2A: context-dependent prediction of protein disorder as a function of redox state and protein binding. *Nucleic Acids Res.* **46**, W329–W337 (2018).
32. Mistry, J. *et al.* Pfam: The protein families database in 2021. *Nucleic Acids Res.* **49**, D412–D419 (2021).
33. Ludwiczak, J., Winski, A., Szczepaniak, K., Alva, V. & Dunin-Horkawicz, S. DeepCoil-a fast and accurate prediction of coiled-coil domains in protein sequences. *Bioinformatics* **35**, 2790–2795 (2019).
34. Fichó, E., Reményi, I., Simon, I. & Mészáros, B. MFIB: a repository of protein complexes with mutual folding induced by binding. *Bioinformatics* **33**, 3682–3684 (2017).
35. Schad, E. *et al.* DIBS: a repository of disordered binding sites mediating interactions

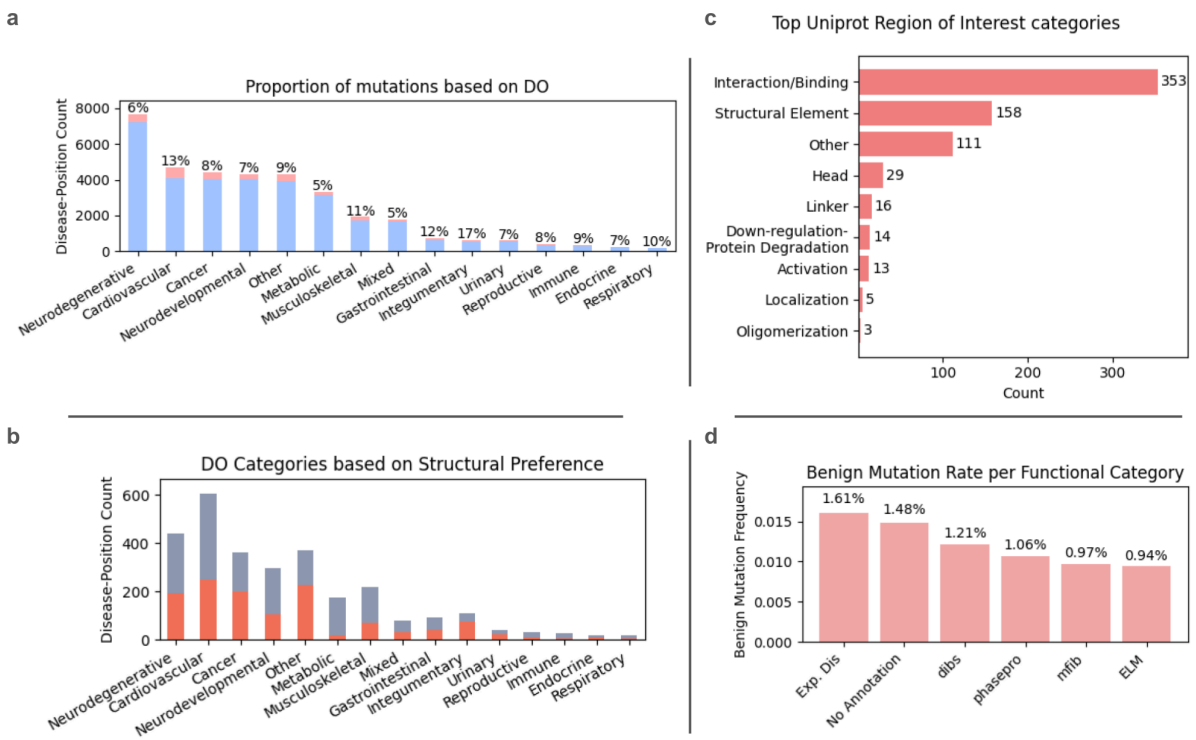
with ordered proteins. *Bioinformatics* **34**, 535–537 (2018).

36. Dinkel, H. *et al.* ELM--the database of eukaryotic linear motifs. *Nucleic Acids Res* **40**, D242–51 (2012).
37. Mészáros, B. *et al.* PhaSePro: the database of proteins driving liquid-liquid phase separation. *Nucleic Acids Res.* **48**, D360–D367 (2020).
38. Huang, K.-Y. *et al.* dbPTM in 2019: exploring disease association and cross-talk of post-translational modifications. *Nucleic Acids Res* **47**, D298–D308 (2019).
39. Hornbeck, P. V. *et al.* PhosphoSitePlus, 2014: mutations, PTMs and recalibrations. *Nucleic Acids Res* **43**, D512–20 (2015).
40. Kabsch, W. & Sander, C. Dictionary of protein secondary structure: pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers* **22**, 2577–2637 (1983).
41. Vasilevsky, N. A. *et al.* Mondo: Unifying diseases for the world, by the world. *medRxiv* 2022.04.13.22273750 (2022) doi:10.1101/2022.04.13.22273750.
42. Buchfink, B., Reuter, K. & Drost, H.-G. Sensitive protein alignments at tree-of-life scale using DIAMOND. *Nat Methods* **18**, 366–368 (2021).

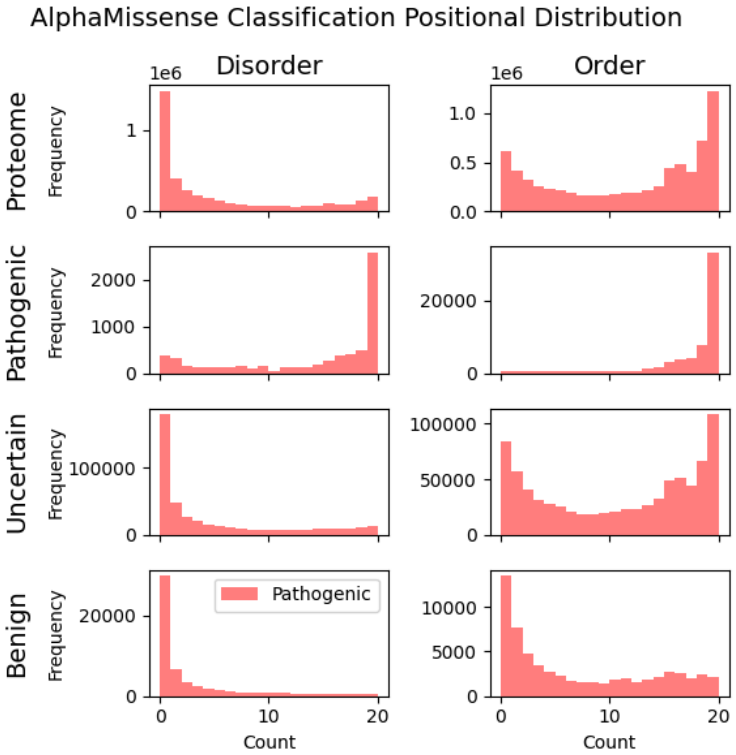
Supplementary Materials

Category	Structure	Unique Protein Position	ClinVar Residue Percentage	Human Proteome Residue Percentage	Mutation Count	Gene Count	Disease Count
Pathogenic	disorder	2 298	0.26 %	0.02 %	5 467	798	813
Pathogenic	order	26 256	2.96 %	0.23 %	61 264	2 740	2 921
Uncertain	disorder	281 501	31.79 %	2.52 %	435 736	14 812	2 585
Uncertain	order	521 001	58.83 %	4.66 %	810 538	16 533	3 842
Benign	disorder	36 539	4.13 %	0.33 %	56 926	7 556	1 569
Benign	order	40 739	4.6 %	0.36 %	64 260	8 482	1 899

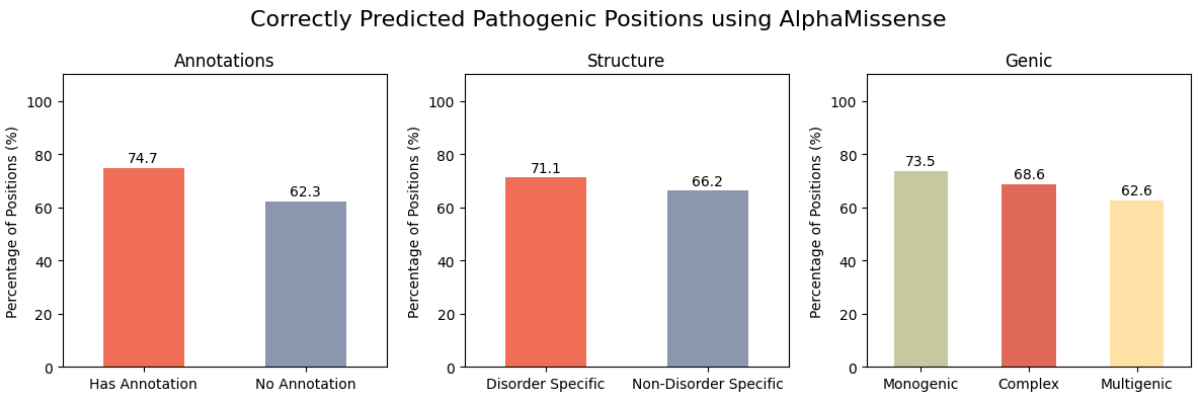
Supplementary Table 1) Structural and Functional Summary Table of ClinVar Variants.



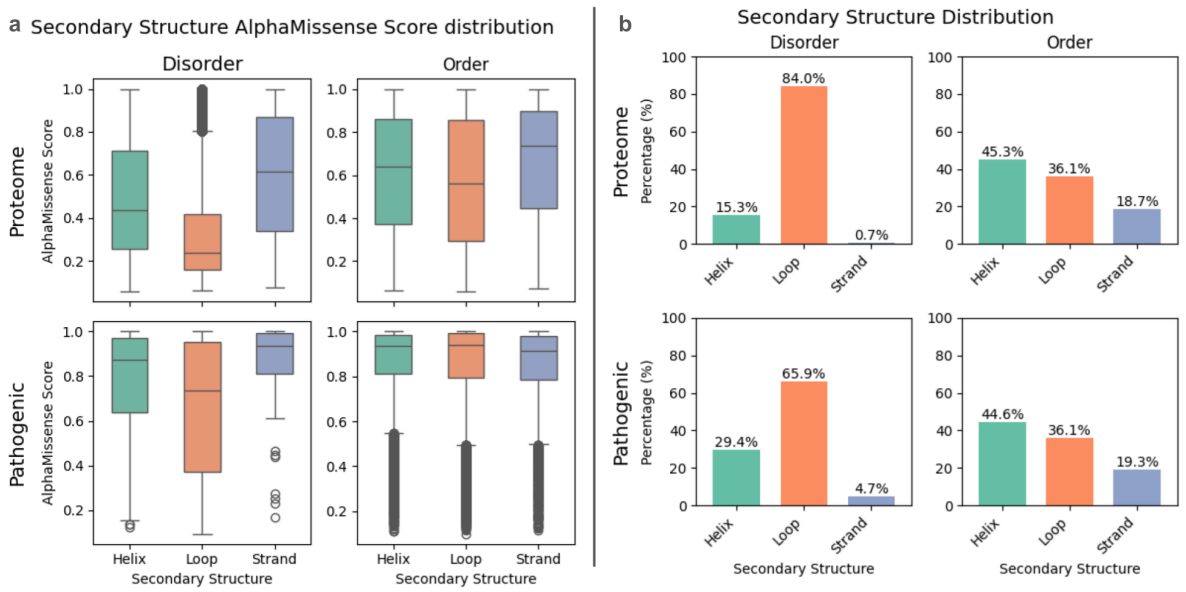
Supplementary Figure 1) a) Structural classification of pathogenic positions by disease ontology. Blue indicates ordered regions, while light pink denotes disordered regions. The percentages represent the proportion of pathogenic positions located in disordered regions within each disease class. d) Distribution of pathogenic mutations within disordered regions across different disease ontology categories. c) Top UniProt region-of-interest (ROI) sites ranked by the number of pathogenic mutations. d) Benign mutation rates across different functional categories compared to unannotated disordered regions.



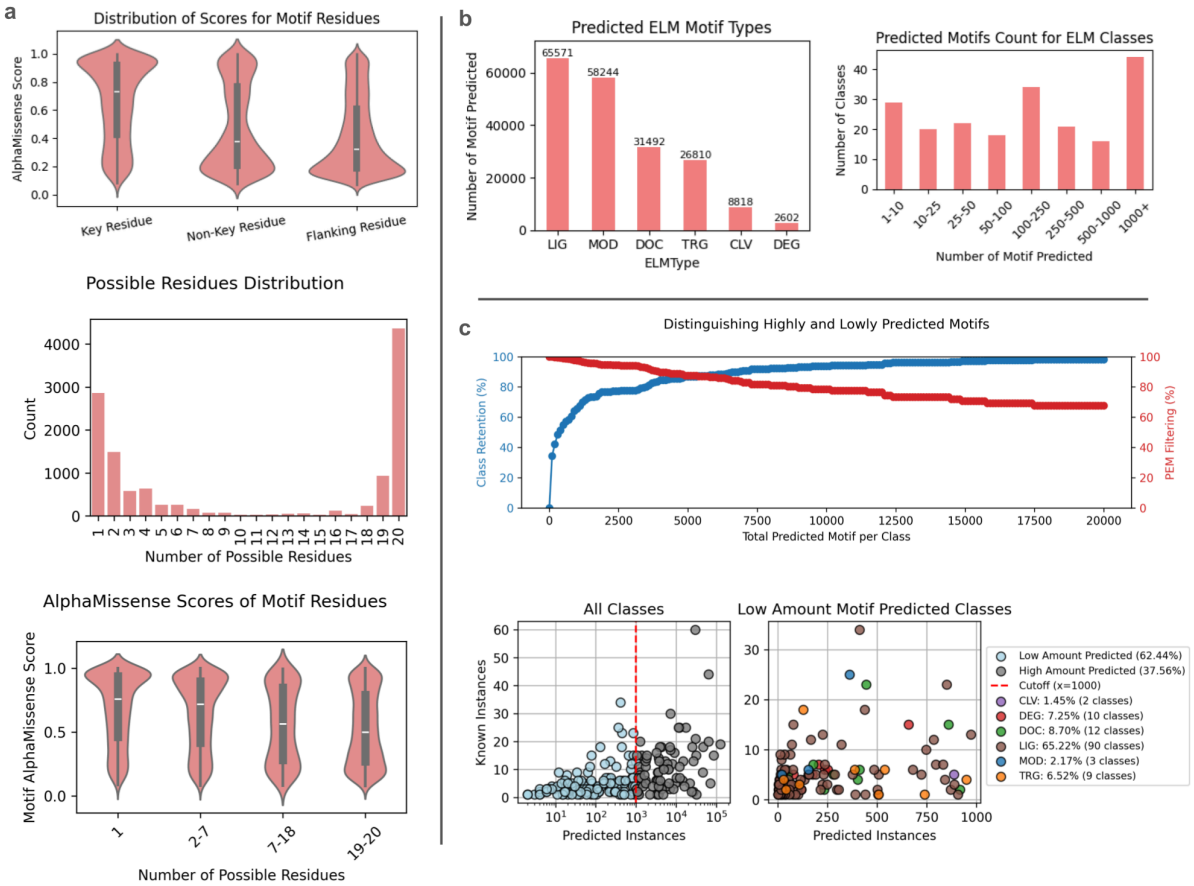
Supplementary Figure 2) Distribution of AlphaMissense pathogenicity classifications for positions across the human proteome and for ClinVar-reported variants. Bars indicate the number of positions classified as pathogenic based on the mean AlphaMissense scores.



Supplementary Figure 3) AlphaMissense Correctly Predicted Pathogenic Positions. In the top we present the accuracy based on annotations. In the bottom segment we show based on structural distribution and genic category.

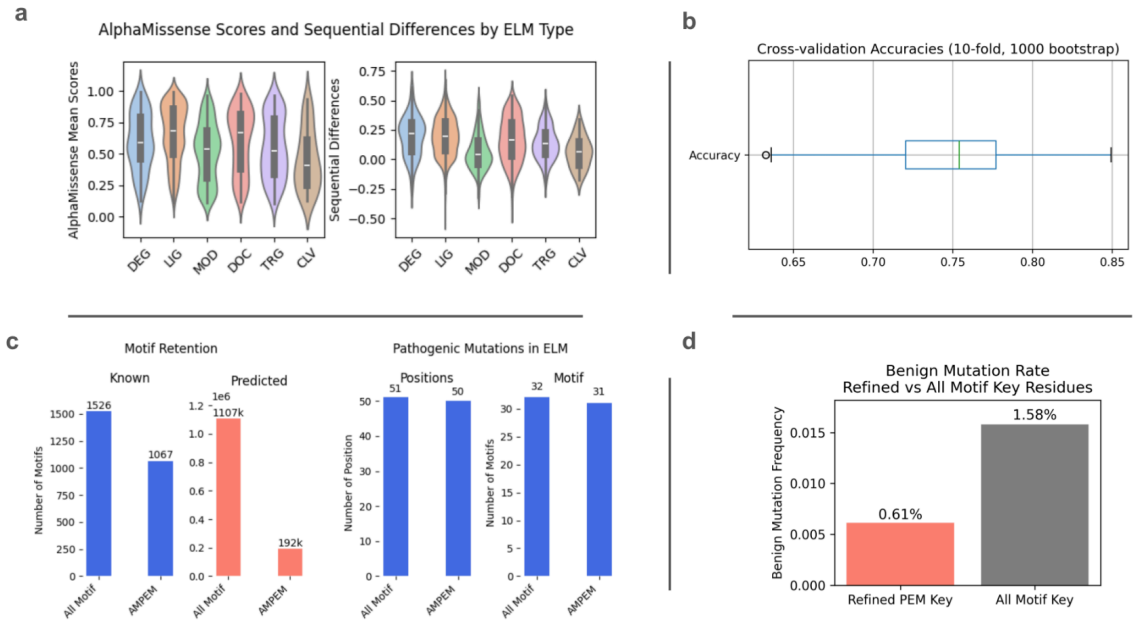


Supplementary Figure 4) a) Distribution of Global AlphaMissense Scores for ClinVar and Human Proteome Positions b) Distribution for Secondary Structure for Proteome and ClinVar Pathogenic Positions.



Supplementary Figure 5) a) Distribution of AlphaMissense Scores for Motif Residues. Key residues (spanning 1–5 amino acids) within each motif display higher pathogenicity scores than both non-key residues (5–20 amino acids) and the surrounding flanking regions. Below we show a detailed distribution of the number of possible residues within known ELM motifs

and the AlphaMissense score distribution of these residues. b) PEM prediction on human proteome. In the left figure we show the number of regions predicted for each protein. The right figure shows the motif length distribution for the predicted motif regions. c) Removal of ELM Classes with high Instance Prediction and the High Confidence Classes ELM Type Distribution. The bottom part contains predicted motif statistics. The left figure shows the ELM Type distribution of the predicted instances. The right figure shows the ELM classes' predicted instances counts.



Supplementary Figure 6) (a) AlphaMissense Mean Score Distribution for Various ELM Types and Differences Between Sequential and Motif Scores. Distinct SLiM types exhibit noticeable variation in AlphaMissense scores, as well as in their differences from the local sequence environment. Among these, LIG, DOC, and DEG motifs typically show the highest pathogenicity, with most scoring above 0.5. Across all SLiM types, the majority of motifs display higher pathogenicity scores than their flanking regions, suggesting that AlphaMissense may be a valuable tool for SLiM detection. b) Cross-validation accuracy distribution of classification of the Decision Tree for Known and PEMs. c) Metrics with the identified PEM motifs. Left panel shows the retention metrics while the right panel shows pathogenic mutations and motifs with pathogenic mutations. d) Benign Mutation Rate for refined PEMs compared to all regex matched motifs.