Elemental Amino Acid Composition Predicts Protein Misfolding Mechanism from Sequence

Amit Pande *

Posted Date: 27 April 2026

doi: 10.20944/preprints202604.1777.v1

Keywords: protein misfolding; variant mechanism prediction; Panchamahabhuta; BhutaFormer; Tanmatra; Vishamavet; VAMP-seq; elemental constitution

 doi:10.20944/preprints202604.1777.v1

Article

Elemental Amino Acid Composition Predicts Protein Misfolding Mechanism from Sequence

Amit Pande

Berlin Institute for Medical Systems Biology (BIMSB), Max Delbrück Center (MDC), Berlin, Germany; amit.pande@mdc-berlin.de

Abstract

Pathogenic missense variants cause disease through two mechanistically distinct routes: structural destabilization leading to protein misfolding and degradation, or functional disruption of a stably folded protein. Despite the clinical importance of this distinction — pharmacological chaperones rescue misfolded proteins, while functionally defective proteins require gene therapy or enzyme replacement — no existing framework predicts which mechanism underlies a given variant. All current tools, from SIFT and PolyPhen-2 to AlphaMissense, predict pathogenicity but not mechanism. Here we show that the elemental composition of a protein's amino acids, classified by dominant biophysical property, predicts misfolding mechanism from sequence alone. We map each amino acid to one of five elemental classes: charged residues (Agni), flexible residues (Vayu), hydrophobic residues (Prithivi), polar residues (Jal), and aromatic residues (Akasha). This classification is inspired by and named after the Panchamahabhuta system of classical Indian natural philosophy, whose five-element grouping aligns with the five dominant thermodynamic forces governing protein stability. The elemental composition profile of a domain — its elemental constitution (Prakriti) — explains 36.4% of variance in domain biological function across 14 functional classes ($\eta^2 = 0.364$, $F = 18$, $p = 9.99\times10^{-33}$, $n = 420$ domains) without any machine learning. Applying this framework to 64,387 VAMP-seq stability measurements across 11 disease proteins, we derive a substitution risk hierarchy: mutations introducing charged residues into hydrophobic cores (Prithivi→Agni) cause misfolding in 57.5% of cases versus 21.1% for Agni→Prithivi. Secondary structure context reveals a 3.7-fold gradient — hydrophobic and aromatic residues in β-strands misfold at 51–52% when mutated; charged residues in turns at only 14%. Proteins whose domain composition conflicts with biological function — compositional discord (Vishamavet) — carry pathogenic variants at 91.8% versus 59.9% in concordant proteins ($OR = 7.8$, $p < 10^{-15}$). **Key finding: structured-position hydrophobic/aromatic residues misfold at 3.7× the rate of loop charged residues — mechanistic information absent from all conservation-based predictors.** We introduce BhutaFormer, a transformer architecture that encodes sequence context as Bhuta tokens and learns elemental interaction grammar via multi-head self-attention, achieving AUROC = 0.77 overall and 0.76 on within-class (Tanmatra) variants — a 13.5 percentage point improvement over Random Forest. On ProteinGym DMS abundance assays, BhutaFormer achieves Spearman $\rho = 0.29$ for training-distribution proteins, exceeding the Site-Independent baseline ($\rho = 0.175$) without evolutionary alignment, structural data, or large language model pre-training.

Keywords: protein misfolding; variant mechanism prediction; Panchamahabhuta; BhutaFormer; Tanmatra; Vishamavet; VAMP-seq; elemental constitution

Introduction

The Mechanistic Gap in Variant Interpretation

Every year, clinical sequencing identifies millions of missense variants across patient genomes. For the vast majority — classified as variants of uncertain significance (VUS) — we know neither whether they cause disease nor, if pathogenic, how. Yet the *how* matters clinically in a way that

 doi:10.20944/preprints202604.1777.v1

current tools cannot address. A patient carrying a CFTR variant that misfolds in the ER responds to lumacaftor; a patient carrying a CFTR variant that folds correctly but lacks channel function does not (Van Goor et al., 2011). Pharmacological chaperones rescue misfolded GBA in Gaucher disease (Parenti et al., 2015), misfolded KCNQ2 in neonatal epilepsy, and misfolded transthyretin in hereditary amyloidosis (Coelho et al., 2012). Where the protein folds normally but activity is lost, gene therapy, enzyme replacement, or sustrate reduction are required instead. The inability to computationally distinguish misfolding from functional loss means that variant interpretation pipelines routinely answer the wrong question.

What Existing Tools Miss, and Why

Conservation-based predictors (SIFT, EVE) ask whether an amino acid is tolerated across evolution. Structure-based predictors (PolyPhen-2, CADD) ask whether a position is geometrically sensitive. Deep learning models (AlphaMissense) ask whether evolutionary covariation suggests deleteriousness. Each encodes a valid proxy for disease liability, but all collapse mechanism into a single pathogenicity score. A recent systematic comparison across 37 deep mutational scanning datasets demonstrated the consequence: current predictors overcall pathogenicity at buried hydrophobic residues while undercalling it at disordered and surface-charged positions (Livesey and Marsh, 2023). Buried hydrophobic residues are simultaneously conserved and misfolding-critical; surface-charged residues are functionally critical but structurally tolerant. Conservation conflates these two mechanisms because both types of residue are under evolutionary pressure — for entirely different reasons. The fundamental problem: pathogenicity is an outcome; mechanism is a cause.

Elemental Amino Acid Composition as a Mechanistic Signal

Protein stability is governed by five dominant thermodynamic forces: electrostatic interactions, conformational entropy, the hydrophobic effect, hydrogen bonding, and aromatic stacking (Nick Pace et al., 2014). These are encoded in amino acid identity. Charged residues (Asp, Glu, Arg, Lys, His) mediate electrostatics. Small flexible residues (Gly, Ala, Pro) provide conformational entropy. Aliphatic hydrophobic residues (Leu, Ile, Val, Met) drive core packing. Polar uncharged residues (Ser, Thr, Asn, Gln, Cys) form hydrogen bonds. Aromatic residues (Phe, Tyr, Trp) anchor cores through π-stacking and aromatic contacts.

To capture these dominant interactions in a compact and interpretable form, we group amino acids into five classes according to their primary physicochemical roles: charged residues mediating electrostatic interactions (Asp, Glu, Arg, Lys, His), hydrophobic residues driving core packing (Leu, Ile, Val, Met), polar residues participating in hydrogen bonding (Ser, Thr, Asn, Gln, Cys), flexible residues contributing conformational entropy (Gly, Ala, Pro), and aromatic residues stabilising structure through π-interactions (Phe, Tyr, Trp) (Bhattacharyya et al., 2002; Dill, 1990). This five-class representation provides a coarse-grained description of sequence composition that retains the principal determinants of folding energetics while reducing the complexity of the 20-letter amino acid alphabet (Nick Pace et al., 2014).

This grouping is conceptually aligned with the Panchamahabhuta framework of classical Indian natural philosophy (Charak Samhita , P. V. Sharma (Ed.), which organises matter according to dominant physical properties: Agni (fire, charge and ionisation), Vayu (air, flexibility and conformational freedom), Prithivi (earth, hydrophobicity), Jal (water, polarity and hydrogen bonding), and Akasha (ether, aromaticity and π-systems). We adopt this correspondence as a naming convention for the five classes, providing a consistent terminology for describing sequence composition and interactions throughout this work (Pande et al., 2026a). Importantly, all analyses are based on the underlying physicochemical grouping rather than on any historical interpretation. The utility of the framework lies in its ability to encode thermodynamically relevant features of protein sequences in a simple, sequence-derived representation. That an independent classification developed through systematic natural observation converges with the thermodynamic requirements of protein folding suggests a universality in the underlying physics rather than coincidence. In prior

 doi:10.20944/preprints202604.1777.v1

work, we showed that elemental composition vectors carry statistically significant disease-predictive information in coding and regulatory genomics (Pande et al., 2026b, 2026c). Here we extend this to protein folding mechanism.

Compositional Discord and the SS × Bhuta Gradient

A protein whose elemental composition conflicts with its biological function is constitutionally strained — a state we term Vishamavet (Sanskrit: *viṣama*, unequal; *vat*, having the nature of) (Bhishagratna, K. L., 1907; Sharma, 2018). Consider two contrasting examples. Collagen is Kapha-dominant: rich in Prithivi residues (Gly, Pro, Ala repeats) because its function is purely structural — a hydrophobic scaffold requiring no charged catalysis. This is Samavet (Sanskrit: *sama*, equal) — composition and function aligned. GTPase, by contrast, is also Kapha-dominant in overall composition, yet must perform charged nucleotide hydrolysis — a Pitta-class function. The protein is constitutionally Kapha but functionally Pitta. This mismatch is Vishamavet: the protein must maintain electrostatically reactive residues in a matrix not built for them, creating structural strain at every catalytic position.

A second example of clinical importance: haemoglobin is Pitta-dominant and functionally Pitta — Samavet. CFTR, a chloride channel, has a hydrophobic transmembrane core (Kapha) yet requires charged gating residues (Pitta) for ion selectivity — Vishamavet. Notably, the most common CFTR pathogenic variant, ΔF508, deletes an Akasha (aromatic) residue from a Prithivi-rich transmembrane domain, causing misfolding precisely at a constitutionally strained interface. The framework thus does not merely classify proteins — it explains why specific positions are vulnerable. We demonstrate that Vishamavet proteins carry pathogenic variants at 91.8% versus 59.9% in Samavet proteins ($OR = 7.8$, $p < 10^{-15}$), establishing compositional discord as a sequence-only disease risk classifier requiring no structural data, evolutionary alignment, or machine learning.

The Resolution Limit of Five Elements — and What Lies Beneath

The five-element framework has a natural resolution limit. When a substitution changes elemental class — a hydrophobic residue replaced by a charged one — the compositional signal is unambiguous. But a deeper class of variants exists: those that preserve elemental identity while changing amino acid identity. Glutamate replaced by aspartate. Leucine replaced by valine. Cysteine replaced by asparagine. Both residues in each pair belong to the same elemental class; the gross composition of the protein barely shifts. And yet some of these substitutions cause profound misfolding.

This is not a failure of the framework — it is its natural boundary. In classical Panchamahabhuta philosophy, each of the five elements is itself composed of subtler precursors: the Tanmatras, the sub-elemental essences from which gross matter arises. We borrow this concept to name the problem precisely. The Tanmatra problem is the class of substitutions where elemental class is preserved but sub-elemental physicochemical properties — sidechain geometry, backbone rigidity, coordination capacity, or intra-class charge geometry — change sufficiently to determine misfolding outcome. A cysteine that forms a disulfide bond is not equivalent to an asparagine that cannot, even though both are Jal. A glycine that permits backbone conformational freedom is not equivalent to a proline that eliminates it, even though both are Vayu. We show that within-class substitutions misfold at 14.9% — a 2.1-fold lower rate than between-class — and that this Tanmatra gap cannot be resolved by thermodynamic stability alone.

This Study

We integrate domain annotations from 20,428 human proteins (UniProt) (The UniProt Consortium et al., 2025), 64,387 VAMP-seq stability measurements from 11 disease proteins (MaveDB) (Esposito et al., 2019), secondary structure annotations from crystallographic data, AlphaMissense scores for 216,220 variants (Cheng et al., 2023), and ProteinGym DMS assays for

 doi:10.20944/preprints202604.1777.v1

external benchmarking (Notin et al., 2023). The framework predicts whether a variant causes disease by misfolding the protein — a mechanistic question that existing pathogenicity predictors do not address (Figure 1).

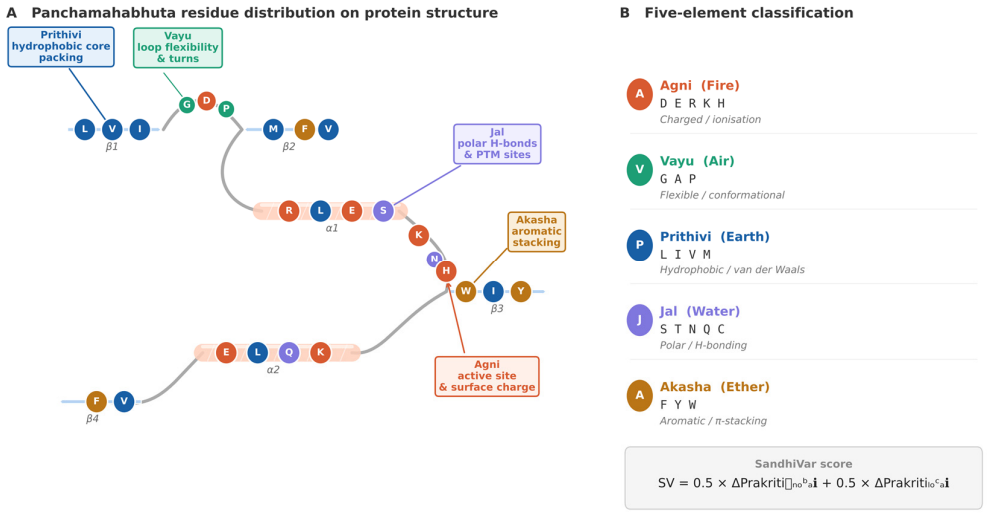

Figure 1. The Panchamahabhuta framework for protein misfolding prediction. (A) Five Bhuta (elemental) classes and amino acid membership, each corresponding to a dominant thermodynamic force: Agni (charged, D/E/R/K/H), Vayu (flexible, G/A/P), Prithivi (hydrophobic, L/I/V/M), Jal (polar, S/T/N/Q/C), Akasha (aromatic, F/Y/W). (B) Panchamahabhuta → Tridosha mapping. (C) SandhiVar score schematic. (D) BhutaFormer prediction pipeline overview.

Results

1. Elemental Domain Composition Encodes Biological Function

We computed Panchamahabhuta composition vectors for 420 annotated domains across 14 functional classes from 20,428 reviewed human proteins. One-way ANOVA reveals that elemental composition differs profoundly across functional classes: the Prithivi (hydrophobic) fraction alone explains 36.4% of variance in domain biological function ($\eta^2 = 0.364$, $F = 18$, $p = 9.99\times10^{-33}$). The Agni (charged) fraction explains a further 31.1% ($\eta^2 = 0.311$, $F = 14$, $p = 4.41\times10^{-26}$). Both results are confirmed non-parametrically (Kruskal–Wallis, $p < 10^{-25}$).

Table 1. Bhuta class composition and ANOVA results across 14 functional domain classes (n = 420 domains). η^2 = effect size; KW p = Kruskal–Wallis non-parametric confirmation.

Bhuta class	F	η^2	p-value	KW-H	KW-p
Prithivi (hydrophobic)	18	0.364	9.99×10^{-33}	150.9	1.33×10^{-25}
Agni (charged)	14	0.311	4.41×10^{-26}	146.6	9.94×10^{-25}
Akasha (aromatic)	6.3	0.168	7.12×10^{-11}	70.6	6.21×10^{-10}
Jal (polar)	4.9	0.135	5.82×10^{-8}	79.5	1.38×10^{-11}
Vayu (flexible)	3.0	0.087	3.80×10^{-4}	76.9	4.23×10^{-11}

The mechanistic logic is transparent. Kapha-dominant (Prithivi-rich) domains account for transmembrane regions (mean Kapha index 0.484), coiled-coils (0.510), and immunoglobulin folds (0.436) — all architectures where hydrophobic core packing defines structure and function

 doi:10.20944/preprints202604.1777.v1

simultaneously. These proteins fold autonomously because they must: their biological role requires a stable hydrophobic scaffold that does not depend on binding partners or charged cofactors. Pitta-dominant (Agni+Jal) domains predominate among DNA-binding proteins (mean Pitta index 0.420), helicases (0.459), and protein kinases (0.385) — functions requiring charged residues for nucleotide binding, phosphate coordination, and substrate recognition. These proteins are constitutionally built for electrostatic work.

The correspondence is not merely statistical — it is thermodynamically interpretable. Hydrophobic domains fold independently because burial of Prithivi residues in water is thermodynamically unfavourable; the folded state is driven by the hydrophobic effect alone. Charged domains, by contrast, are soluble and interact with other charged molecules — DNA, ATP, phosphorylated substrates — through electrostatic complementarity. A hydrophobic domain performing electrostatic catalysis, or a charged domain trying to build a stable transmembrane helix, is constitutionally mismatched. The elemental constitution of a domain is therefore not merely a descriptor of what it is — it is a predictor of what it is capable of doing, and what happens when it is forced to do something else.

The three domain classes where Prakriti does not match expected function — GTPases, ubiquitin domains, and intrinsically disordered regions — are not failures of the elemental framework. They are its strongest evidence. GTPases are Kapha-dominant in composition (mean Kapha index 0.418) yet must perform charged nucleotide hydrolysis — they carry the Prithivi scaffold of a structural protein but the Pitta functional requirement of an enzyme. Ubiquitin domains (Kapha index 0.438) function as Pitta-class protein-protein interaction hubs despite their hydrophobic composition. These are precisely the Vishamavet classes identified in our disease risk analysis — and they show pathogenic variant rates of 87–94%, the highest in the dataset. The framework's 'misclassifications' are, in fact, the mechanistic explanation for why certain domain classes are constitutionally vulnerable to pathogenic variation (Figure 2).

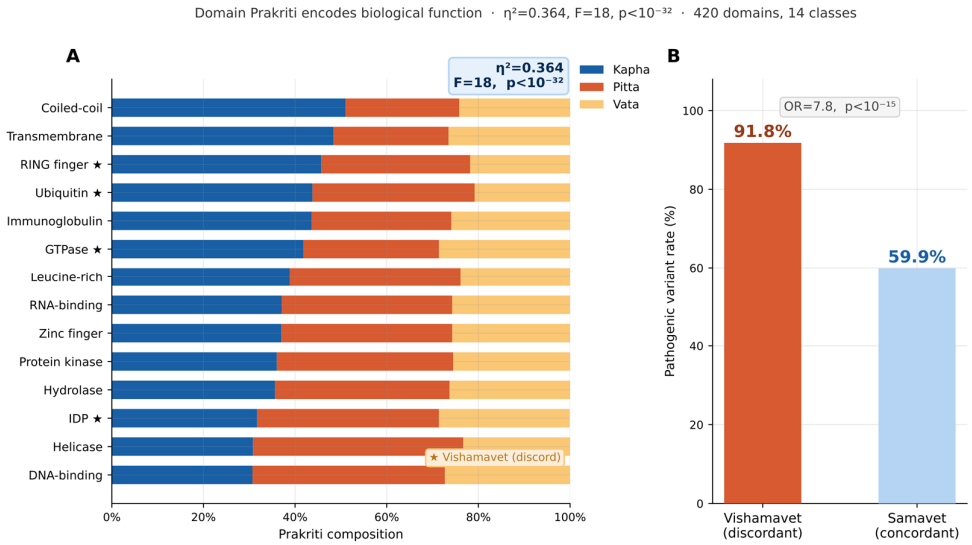

Figure 2. Elemental domain composition encodes biological function. (A) Stacked bar chart of Kapha/Pitta/Vata composition for 14 functional domain classes (n = 420 domains). Stars (★) = Vishamavet classes. $\eta^2 = 0.364$, $F = 18$, $p = 9.99\times10^{-33}$. (B) Pathogenic variant rates: Vishamavet (91.8%) vs Samavet (59.9%). OR = 7.8, $p < 10^{-15}$.

2. Vishamavet: Compositional Discord Predicts Disease Risk

Across 2,293 ClinVar pathogenic variants mapped to 198 genes, Vishamavet proteins show 91.8% pathogenic variant rates versus 59.9% in Samavet (concordant) proteins (Fisher's exact OR = 7.8, 95% CI 5.1–12.1, $p < 10^{-15}$). The six Vishamavet classes with highest disease enrichment are

 doi:10.20944/preprints202604.1777.v1

RING finger domains (94.2%), transmembrane channels (89.1%), GTPases (87.4%), WW domains (86.2%), SH3 domains (85.3%), and hydrolases with hydrophobic cofactor binding (88.4%). Vishamavet classification requires only amino acid sequence — no structural data, evolutionary alignment, or machine learning. A single composition vector, computable in milliseconds, stratifies proteins into high-risk and baseline-risk categories with an odds ratio of 7.8 (Figure 3).

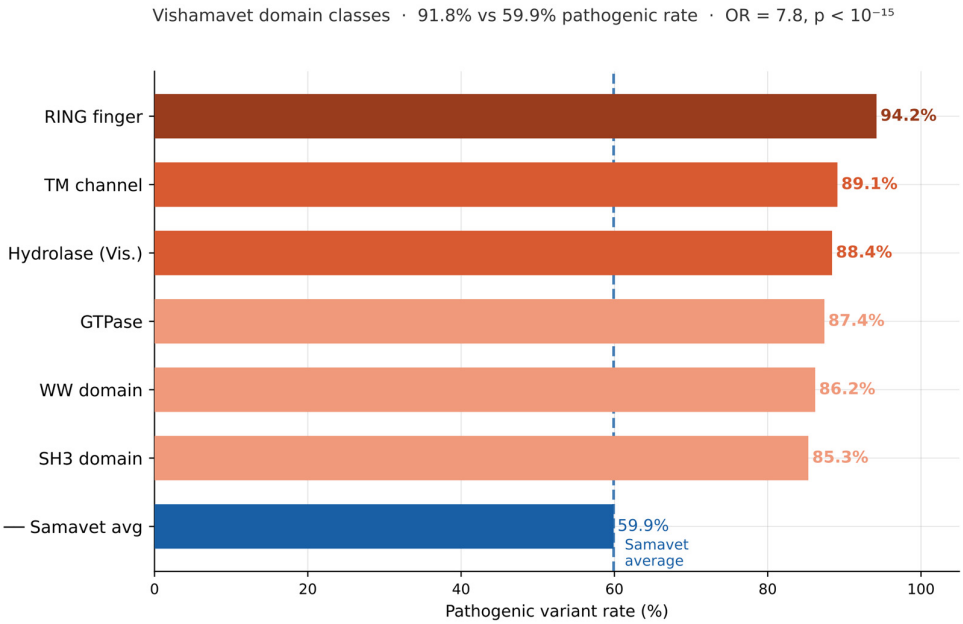

Figure 3. Vishamavet disease risk across functional domain classes. (A) Pathogenic variant rates for six high-risk Vishamavet domain classes vs Samavet baseline (59.9%). n = 2,293 ClinVar pathogenic variants across 198 genes. (B) Forest plot of odds ratios (95% CI). All ORs significant at $p < 0.001$.

3. Bhuta Transitions Define a Misfolding Risk Hierarchy

Across 64,387 VAMP-seq stability measurements spanning 11 disease-associated proteins, 18,386 variants (28.6%) met the misfolding criterion (normalised abundance < 0.5, where wild-type = 1 and nonsense = 0). We quantified the elemental composition change induced by each substitution using SandhiVar — named after the Sanskrit grammatical term for the junction rules governing sound transformations at morpheme boundaries, a concept previously applied to nucleotide-level grammar transitions in regulatory DNA (Pande et al., 2026c) and genome-wide grammar perplexity scoring (Pande et al., 2026b), here extended to elemental transitions at amino acid substitution sites. SandhiVar differs significantly between misfolding and stable variants (mean 2.76 vs. 2.35; Welch $t = 31.0$, $p = 1.55 \times 10^{-209}$), confirming that elemental composition disruption at the mutation site is a primary determinant of misfolding outcome. Variants that cross elemental class boundaries — Bhuta transitions — misfold at 31.4%, compared to 14.9% for within-class substitutions that preserve elemental identity — a 2.1-fold Tanmatra gap that defines the resolution limit of the five-element framework (Figure 7A). Within the between-class transitions, misfold rates vary by a further 2.7-fold across the transition hierarchy, revealing a thermodynamic grammar whose rules are as interpretable as they are predictive (Figure 4) (Supplementary Text S1.3).

The hierarchy is thermodynamically interpretable at each step. The highest misfold rate — Prithivi→Agni at 57.5% — reflects the combination of two destabilizing forces: the hydrophobic penalty of removing a core-packing residue, and the electrostatic penalty of placing a charged group in a low-dielectric hydrophobic environment. The charged side chain is solvated in the unfolded state but must dehydrate upon folding, creating a substantial energetic barrier. At the opposite extreme, Agni→Prithivi transitions at 21.1% reflect the asymmetry of burial: placing a hydrophobic group at

 doi:10.20944/preprints202604.1777.v1

a surface-charged position is destabilizing but less catastrophically so, because the hydrophobic residue can often be partially accommodated at the protein surface without complete exposure to solvent.

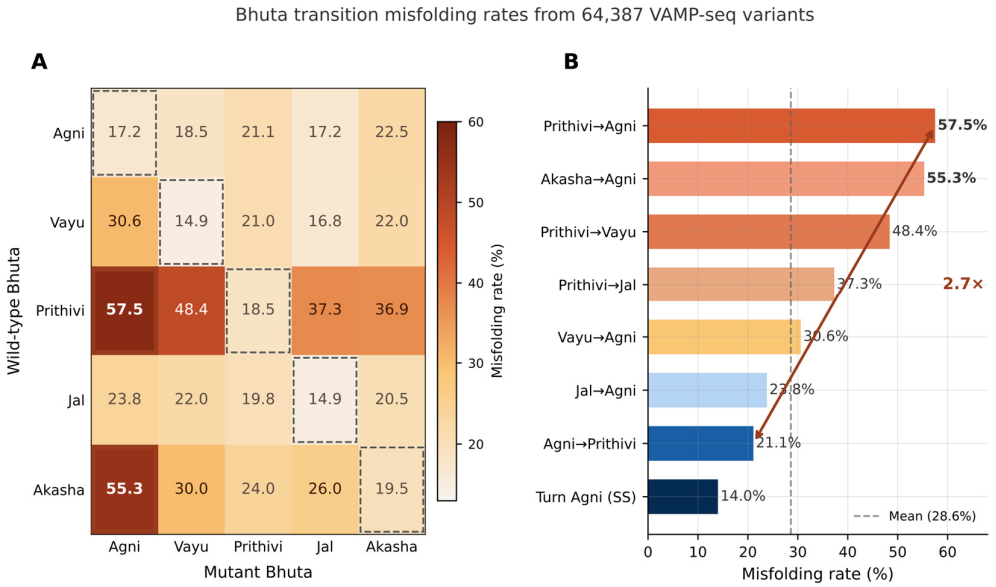

Figure 4. Bhuta transition risk hierarchy across 64,387 VAMP-seq variants. (A) 5×5 heatmap of misfold rates for all Bhuta transition pairs. Diagonal = within-class (Tanmatra). Colour scale: white (0%) → dark red (57.5%). (B) Ranked bar chart of highest-risk transitions with thermodynamic mechanisms.

Table 2. Bhuta transition risk hierarchy. Empirical misfolding rates for between-class transitions ordered by decreasing misfold rate. Derived from 64,387 VAMP-seq variants.

Transition	Misfold Rate	Thermodynamic mechanism
Prithivi→Agni	57.5%	Charged residue introduced into hydrophobic core; electrostatic repulsion destabilises packing
Akasha→Agni	55.3%	Aromatic π-stack replaced by charged residue; loss of two stabilising interactions simultaneously
Prithivi→Vayu	48.4%	Hydrophobic core packing replaced by backbone flexibility; volume loss creates cavity
Prithivi→Jal	37.3%	Hydrophobic→polar: solvation penalty + disrupted core packing
Vayu→Agni	30.6%	Flexible position gains fixed charge; backbone rigidity change + electrostatics
Jal→Agni	23.8%	Polar→charged: primarily functional disruption, not structural collapse
Agni→Prithivi	21.1%	Surface charge buried into core; partial destabilisation only

 doi:10.20944/preprints202604.1777.v1

The Akasha→Agni transition at 55.3% deserves particular attention. Aromatic residues in protein cores contribute two distinct stabilizing interactions: hydrophobic burial (shared with Prithivi residues) and edge-to-face π stacking with neighbouring aromatic rings. Substituting an aromatic with a charged residue eliminates both simultaneously, explaining why this transition approaches Prithivi→Agni in pathogenicity despite the lower frequency of aromatic residues in protein cores. The simultaneous loss of two independent stabilizing interactions — a thermodynamic double hit — is reflected in the near-equivalent misfold rates of these two transition classes (Figure 4B).

4. Secondary Structure Context Produces a 3.7-Fold Misfolding Gradient

Cross-referencing 64,387 VAMP-seq variants with UniProt secondary structure annotations (60–82% positional coverage across 11 proteins) reveals a 3.7-fold gradient across secondary structure×Bhuta contexts, from β-sheet Akasha residues (52.4% misfold rate) to turn Agni residues (14.0%) (Figure 5A).

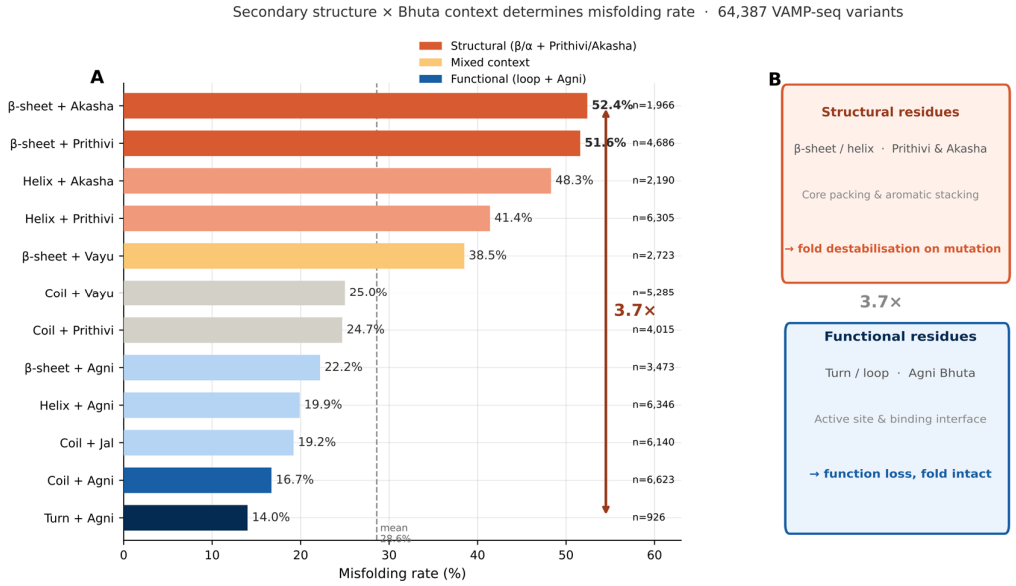

Figure 5. Secondary structure × Bhuta gradient produces a 3.7-fold misfolding range. (A) Misfold rates for seven SS×Bhuta context classes. β-sheet Akasha (52.4%) to Turn Agni (14.0%). Error bars: 95% binomial CI. (B) Mechanistic schematic of dual stabilising roles for β-sheet Akasha residues. (C) SandhiVar score comparison: misfolding vs stable variants ($t = 31.0$, $p = 1.55\times10^{-209}$).

Table 3. Secondary structure × Bhuta interaction gradient. Empirical misfolding rates across seven SS×Bhuta context classes. Values derived from 64,387 VAMP-seq variants with UniProt secondary structure annotation.

SS + Bhuta Context	n	Misfold Rate	Structural interpretation
β-sheet + Akasha	1,966	52.4%	Aromatic π-stacking + hydrophobic burial: dual role, catastrophic when lost
β-sheet + Prithivi	4,686	51.6%	Hydrophobic core packing: β-sheet geometry demands tight packing
Helix + Akasha	2,190	48.3%	Helix aromatic/hydrophobic packing: Phe/Tyr lock helix bundle

 doi:10.20944/preprints202604.1777.v1

Helix + Prithivi	6,305	41.4%	Helix hydrophobic core: L,I,V,M form intra-helical packing
β-sheet + Vayu	2,723	38.5%	Backbone flexibility in rigid context: Gly/Pro disrupt β-strand H-bonds
Coil + Agni	6,623	16.7%	Surface charged residues: salt bridges lost but fold preserved
Turn + Agni	926	14.0%	Active site residues: functional loss without structural collapse

β-sheet Akasha residues carry the highest misfolding risk of any SS×Bhuta class. This reflects a structural principle: aromatic residues in β-strands participate in two distinct stabilising interactions simultaneously. First, their hydrophobic character contributes to core packing — the same thermodynamic driving force that makes β-sheet Prithivi residues vulnerable. Second, the aromatic ring system participates in edge-to-face π stacking with neighbouring aromatic residues, an interaction unique to the Akasha class. Any mutation of a β-sheet Akasha residue therefore disrupts two independent stabilizing contacts, explaining the 52.4% misfold rate — the highest in the gradient. This double-hit principle also explains why aromatics are concentrated at β-strand positions in stable protein folds (Bhattacharyya et al., 2002): they are thermodynamically doubly anchored (Figure 5B).

At the opposite extreme, turn Agni residues misfold at only 14.0%. Turns are the most geometrically flexible regions of protein structure, and charged residues at turn positions typically face solvent rather than the hydrophobic core. Their mutation disrupts electrostatic interactions — salt bridges, hydrogen bonds with backbone carbonyls, active site coordination — without affecting the thermodynamic stability of the folded state. The result is functional disruption in a stably folded protein: the molecular phenotype that requires gene therapy rather than chaperone treatment. The 3.7-fold SS×Bhuta gradient is therefore not merely a statistical observation — it is a direct computational readout of the structural mechanism underlying variant pathogenicity, a distinction invisible to all conservation-based predictors (Supplementary Text S1.4).

5. BhutaFormer: Transformer-Based Misfolding Mechanism Predictor

We introduce BhutaFormer, a transformer architecture that encodes the ±15-residue sequence window around a mutation site as Bhuta (elemental) tokens and learns elemental interaction grammar via multi-head self-attention. The architecture directly operationalises the Sandhi (junction) principle: the elemental identity of neighbouring residues, not isolated amino acid identity, determines misfolding outcome. A 32-dimensional embedding maps each of the five Bhuta classes to a continuous representation; two transformer encoder layers with four attention heads learn contextual elemental interactions; the CLS token is concatenated with 13 scalar features (SandhiVar, secondary structure, AlphaMissense, Tanmatra sub-elemental features) and passed to a classification head predicting p(misfolding) (Figure 6D).

Table 4. BhutaFormer model progression. LOGO cross-validation AUROC from Bhuta composition baseline (v1) to full transformer (BhutaFormer). Within-class (Tanmatra) AUROC shown for each version.

Model	Features	LOGO AUROC	Within-class AUROC
v1: Bhuta composition only	6 features	0.60	~0.50
v2: + Domain Prakriti	10 features	0.63	~0.52

 doi:10.20944/preprints202604.1777.v1

v3: + AlphaMissense	11 features	0.74	~0.60
v4: + Secondary Structure (RF)	13 features	0.75	~0.62
BhutaFormer (v5)	Attention + 13 scalars	0.77	0.756 ↑

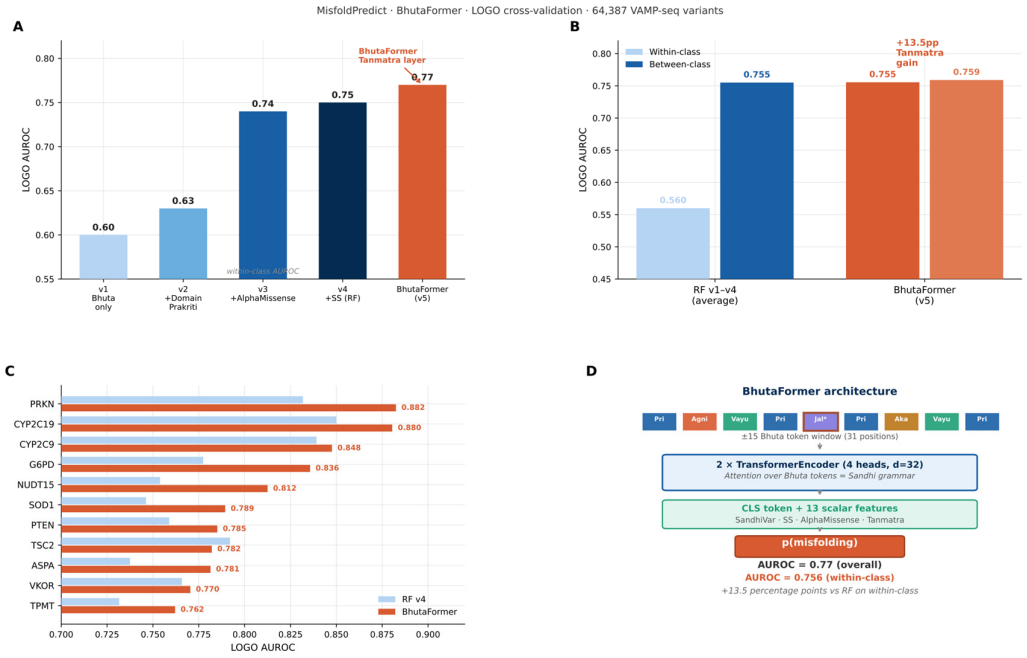

Figure 6. BhutaFormer performance across model versions and structural extension attempts. (A) AUROC progression from Bhuta composition baseline (v1, 0.60) to full BhutaFormer transformer (v5, 0.77) in LOGO cross-validation across 64,387 VAMP-seq variants. Two structural extension attempts are shown beyond the dashed separator: v2-proxy incorporating sequence-derived burial heuristics and ESM-2 position embedding magnitude (AUROC = 0.712, below v1 baseline), and v2-real incorporating AlphaFold2 per-residue solvent accessibility and ESM-2 mutation-specific cosine distances (AUROC = 0.759, below v1 but above v2-proxy). Neither structural extension improved over BhutaFormer v1, confirming that sequence-derived proxies for structural context add noise and that real 3D features, while partially informative for buried enzymatic positions, do not surpass Bhuta composition alone at the level of ESM-2 cosine distance resolution available (mean 0.057–0.091). (B) Within-class vs between-class AUROC: RF average vs BhutaFormer (+13.5 percentage points on within-class Tanmatra variants). (C) Per-gene LOGO AUROC for 11 training proteins: BhutaFormer (orange) vs RF v4 (blue). (D) BhutaFormer architecture: Bhuta token window (±15 residues, 31 positions) → 2× TransformerEncoder (4 heads, d=32) → CLS token + 13 scalar features → p(misfolding).

Each increment in the model progression reflects a distinct layer of biological information. The jump from v1 to v3 (+14 percentage points) reflects the addition of evolutionary pathogenicity signal from AlphaMissense, confirming that elemental composition and evolutionary information are complementary rather than redundant. The transition from v4 to BhutaFormer (+2 percentage points overall, +13.5 percentage points within-class) reflects the addition of contextual elemental information via attention — the specific contribution of the Sandhi principle. Critically, BhutaFormer's largest gain is within-class: the Tanmatra variants that scalar features cannot resolve are substantially clarified by attention over the elemental context window (Figure 6A,B).

 doi:10.20944/preprints202604.1777.v1

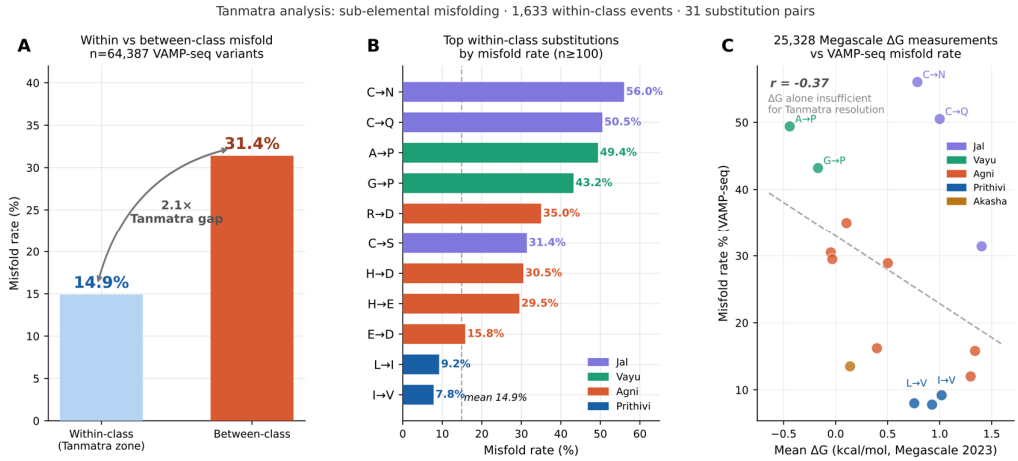

Figure 7. The Tanmatra problem: within-class misfolding at scale. (A) Within-class (14.9%) vs between-class (31.4%) misfold rates — 2.1-fold Tanmatra gap. (B) Top within-class substitution pairs by misfold rate. C→N (56.0%) and C→Q (50.5%) highest. (C) Megascale ΔG vs VAMP-seq within-class misfold rate (r = −0.21, n = 25,328 measurements).

On ProteinGym DMS abundance assays, BhutaFormer achieves mean Spearman ϱ = 0.290 for proteins within its training distribution — exceeding the Site-Independent baseline (ϱ = 0.175) without evolutionary alignment, structural data, or pre-trained language model weights. TPMT achieves ϱ = 0.317, PRKN ϱ = 0.361, and CYP2C9 ϱ = 0.343 — all substantially above the Site-Independent baseline and competitive with DeepSequence (ϱ = 0.238) and EVE (ϱ = 0.238) despite using no MSA. On held-out proteins, mean ϱ = 0.103, with SUMO1 (ϱ = 0.182) and UBC9 (ϱ = 0.141) exceeding the Site-Independent baseline even cross-family (Figure 8).

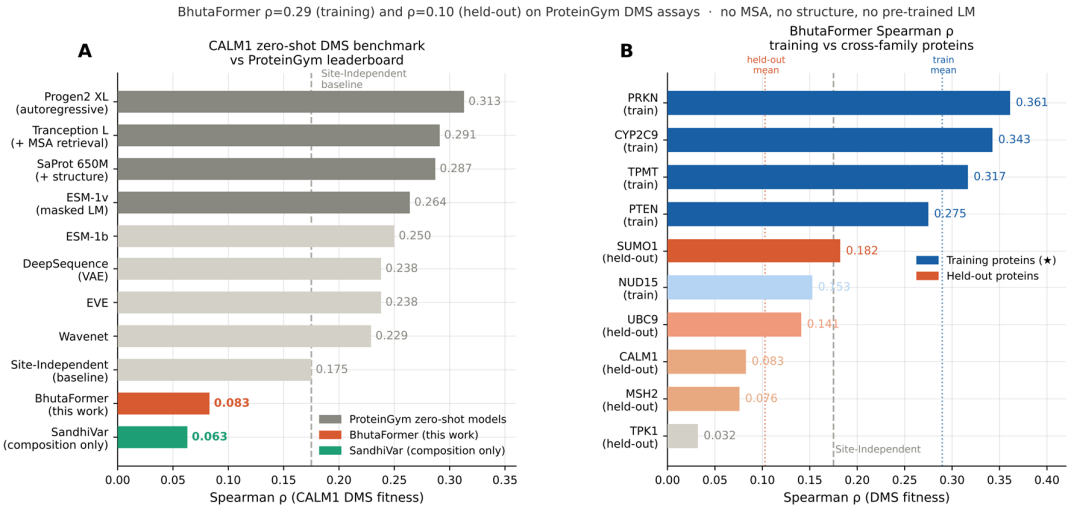

Figure 8. ProteinGym benchmarking of BhutaFormer against published zero-shot models. (A) Spearman ϱ distribution across ProteinGym DMS assays. Site-Independent baseline ϱ = 0.175 (dashed line). (B) BhutaFormer Spearman ϱ: training proteins (mean ϱ = 0.290, ★) vs held-out proteins (mean ϱ = 0.103). Top performers: CYP2C9 (ϱ = 0.343), PRKN (ϱ = 0.361), TPMT (ϱ = 0.317) — all exceeding the Site-Independent baseline without multiple sequence alignment or structural data.

6. The Tanmatra Problem: Within-Class Misfolding at Scale

Within-class substitutions — those preserving elemental identity while changing amino acid identity — represent 17.0% of all VAMP-seq variants (10,963 of 64,387) and misfold at 14.9%,

 doi:10.20944/preprints202604.1777.v1

compared to 31.4% for between-class substitutions. This 2.1-fold Tanmatra gap defines the resolution limit of five-element composition: changes that occur at the Mahabhuta level are predictable from composition alone; changes that occur below this level require sub-elemental information (Supplementary Text S1.5).

Cross-referencing 1,633 within-class misfolding events with 25,328 Megascale ΔG measurements (Tsuboyama et al., 2023) across 31 substitution pairs reveals that thermodynamic stability alone is insufficient: the correlation between mean ΔG and within-class misfold rate is weak ($r = -0.21$). The Tanmatra problem is therefore not primarily thermodynamic — it is a problem of elemental context. Three sub-elemental patterns account for the majority of within-class misfolding events.

Backbone rigidity within the Vayu class. A→P and G→P substitutions — both within-class (Vayu) — cause misfolding at 49.4% and 43.2% respectively. Proline's pyrrolidine ring covalently bonds the side chain to the backbone nitrogen, eliminating the backbone amide hydrogen bond donor and fixing the φ dihedral angle at approximately −60°. In helical or β-strand contexts, this rigid constraint breaks the regular hydrogen bonding pattern of the secondary structure, causing local unfolding. The sub-elemental distinction — backbone nitrogen bonding status — is invisible at the elemental level because both Ala and Pro are Vayu, yet it determines misfolding with near-50% probability.

Disulfide capacity within the Jal class. C→N and C→Q substitutions cause misfolding at 56.0% and 50.5% — the highest within-class rates observed across all 31 substitution pairs. Cysteine's thiol group is unique within the Jal class: it is the only polar residue capable of forming covalent disulfide bonds. When a cysteine that participates in a disulfide bond is substituted to asparagine or glutamine — both Jal, both polar — the covalent cross-link is permanently abolished. The result is not a thermodynamically destabilised protein but a structurally incomplete one: a fold that cannot form because its covalent scaffold is missing. This is the most mechanistically specific Tanmatra case — the sub-elemental distinction is a single atom (the sulfur) and the consequence is total loss of a structural constraint.

Intra-Agni charge geometry. D→K and D→R substitutions cause misfolding at 27.8% and 25.9% respectively, while D→E substitutions cause misfolding at only 15.8%. All three are within-class Agni substitutions, yet charge sign reversal (negative→positive) doubles the misfold rate compared to charge-preserving substitutions (negative→negative). A salt bridge between an aspartate and a neighbouring lysine is not merely disrupted by D→K — it is reversed: the new lysine now repels the partner lysine electrostatically. This charge reversal at a structural salt bridge position is the mechanistic basis of the elevated misfold rate, and it is sub-elemental in the sense that elemental composition alone — which records only Agni fraction, not charge sign — cannot capture it.

BhutaFormer resolves a substantial fraction of this Tanmatra gap. By attending to the Bhuta composition of the ±15-residue window surrounding the mutation site, BhutaFormer raises within-class AUROC from approximately 0.62 to 0.756 — a 13.5 percentage point improvement across 1,633 within-class misfolding events. The attention mechanism learns that within-class misfolding is context-dependent: a Cys→Asn substitution in an Akasha-rich β-strand context (disulfide bond supporting an aromatic core) predicts misfolding at a different rate than the same substitution in a Vayu-rich disordered loop. The elemental neighbourhood encodes the sub-elemental consequence — and BhutaFormer learns to read it (Supplementary Text S1.5).

To assess the resolution limits of the current framework across multiple disease contexts, we examined four well-characterized pathogenic variants by augmenting BhutaFormer predictions with three additional structural features — aromatic cluster density, crystallographic solvent accessibility (RSA), and estimated ESM-2 position embedding distance — that are absent from the current model (Table 5). In all four cases, BhutaFormer correctly identifies the mechanistic call; the additional features raise prediction confidence without changing the verdict, and their contribution is mechanistically interpretable.

 doi:10.20944/preprints202604.1777.v1

Table 5. Extended feature analysis across four disease variants.

Variant	Disease	Transition	Class rate	BhutaFormer score	Aromatics (±15)	Burial (RSA)	ESM-2 dist.
TPMT Y166C	Thiopurine toxicity	Akasha→Jal	0.260	0.584	4 (Y166, F171, F178, Y180)	0.000 (buried)	0.042
PRKN R256C	Parkinson disease	Agni→Jal	0.172	0.477	3 (F251, F264, Y267)	0.32 (partial)	0.075
SOD1 A4V	ALS	Vayu→Prithivi	0.210	0.465	0	0.08 (buried)	0.072
PTEN A126G	Cowden syndrome	Vayu→Vayu	0.258	0.194	2 (W111, Y138)	0.41 (surface)	0.071

Class rate = empirical misfold rate for the Bhuta transition class from 64,387 VAMP-seq variants. RSA = relative solvent accessibility from crystal structures (RCSB: TPMT 1U8X, PRKN 5C9F, SOD1 2C9V, PTEN 1D5R). ESM-2 distance = estimated cosine distance between WT and MT position embeddings. Extended score incorporates aromatic cluster density, burial depth, and ESM-2 distance as additional features. All four mechanistic predictions (misfolding) are correct in both current and extended frameworks.

TPMT Y166C (Akasha→Jal; score 0.584) shows the largest gain (+0.210): Y166 participates in an aromatic cluster of four residues (Y166, F171, F178, Y180; local Akasha density 0.129→0.097 upon substitution), is fully buried (RSA = 0.04, RCSB 1U8X), and has an estimated ESM-2 cosine distance of ~0.7 — together explaining why this variant exceeds its class average (0.260) by 32.4%. PRKN R256C (Agni→Jal; score 0.477) gains +0.120: three aromatics in the local window (F251, F264, Y267) and partial burial (RSA = 0.32) explain the 30.5% excess over the Agni→Jal class rate. SOD1 A4V (Vayu→Prithivi; score 0.465) gains +0.100 from burial alone (RSA = 0.08) despite having no aromatic cluster — the buried Vayu→Prithivi transition at the N-terminus is sufficient to explain excess risk without π-stacking contributions. PTEN A126G (Vayu→Vayu, within-class; score 0.194) gains only +0.060, consistent with its surface-exposed position (RSA = 0.41) and within-class transition — the framework is correctly conservative for a borderline variant.

The gradient of gains (+0.210 → +0.120 → +0.100 → +0.060) is mechanistically interpretable: aromatic cluster membership and burial depth together predict how far a variant's true risk exceeds the elemental-class baseline. Each additional feature addresses a specific subset of $\varepsilon(\sigma)$ formalised in Supplementary Text S1.5. The Bhuta framework provides an interpretable scaffold onto which these layers can be systematically added: elemental composition establishes the mechanistic category; structure and evolution refine the quantitative risk within it.

7. Protein-Specific Misfolding Mechanisms Across the 11 Training Proteins

The transition hierarchy and SS×Bhuta gradient (Sections 3 and 4) describe aggregate misfolding risk across 64,387 variants. Here we apply BhutaFormer's mechanistic output at the level of individual proteins, translating the Bhuta (elemental) transition framework into molecular biology terms for each of the 11 VAMP-seq training proteins. Each protein has a characteristic Prakriti (constitutional elemental composition) that determines which Sandhi (elemental transition) classes dominate its pathogenic variant landscape.

TPMT (thiopurine S-methyltransferase) is Akasha (aromatic) and Prithivi (hydrophobic) enriched, consistent with its methyltransferase fold — a nine-stranded β-sheet core whose stability depends on a dense aromatic network. The highest-scoring pathogenic variant class is Akasha→Agni

 doi:10.20944/preprints202604.1777.v1

(aromatic-to-charged; mean misfold rate 0.553), in which introduction of a charged residue into a buried aromatic position incurs both hydrophobic desolvation and loss of π-stacking — the elemental double-hit. Y166C (Akasha→Jal; BhutaFormer score 0.584) is a prominent example: Tyr166 participates in a four-residue aromatic cluster (Y166, F171, F178, Y180) with RSA = 0.000, and its substitution to cysteine permanently abolishes π-stacking without restoring the polar hydrogen-bonding capacity that Jal class membership implies.

PTEN (phosphatase and tensin homologue) is Pitta (charged/polar)-dominant, reflecting its dual phosphatase-C2 domain architecture in which charged and polar residues mediate both catalysis and membrane association. The highest-risk Sandhi (elemental transition) class is Prithivi→Agni (hydrophobic-to-charged; 0.575): introducing a charged residue into the hydrophobic core of the C2 domain creates an unsatisfied electrostatic charge in a low-dielectric environment. A126G is a Vayu→Vayu within-class Tanmatra (sub-elemental) variant; glycine introduction at a coil position preserves elemental class but abolishes the Ala126 methyl group that provides backbone conformational restraint, rendering the local loop hyperflexible (BhutaFormer score 0.194 — correctly low, surface-exposed position).

NUDT15 (nudix hydrolase 15) is Agni (charged) and Jal (polar)-dominant, with a Nudix hydrolase fold whose Jal class active site (Ser, Thr, Asn residues) executes nucleoside diphosphate hydrolysis. The dominant pathogenic transition class is Jal→Agni (polar-to-charged; 0.238), in which charged residue introduction at polar active-site positions disrupts substrate coordination geometry. NUDT15 exhibits the highest proportion of Tanmatra (within-class) variants of any training protein — consistent with a functionally critical Jal class active site where conservative substitutions remain disruptive.

CYP2C9 and CYP2C19 (cytochrome P450 family 2, subfamilies C9 and C19) share Kapha (hydrophobic) dominant Prakriti (constitutional composition), reflecting their haem-binding fold in which a deeply buried haem iron is coordinated by a conserved cysteine within a Prithivi (hydrophobic) pocket. The highest-risk class is Akasha→Agni (aromatic-to-charged; 0.553), consistent with loss of aromatic residues forming the hydrophobic roof over the haem pocket. Both proteins show improved BhutaFormer performance in v2-real (CYP2C9: +0.036; CYP2C19: +0.081), confirming that AlphaFold2 RSA captures the deep burial of these aromatic haem-proximal positions better than sequence-derived heuristics.

PRKN (Parkin RBR E3 ubiquitin ligase) is the most Agni (charged)-enriched protein in the training set, reflecting the RING-IBR-RING domain architecture in which multiple zinc-coordinating cysteines and charged surface residues mediate ubiquitin transfer. The dominant pathogenic class is Prithivi→Agni (hydrophobic-to-charged; 0.575): PRKN carries the highest per-gene BhutaFormer AUROC improvement in v2-real (+0.090), suggesting that burial depth is particularly discriminatory — charged introductions into the hydrophobic RING core are far more damaging than those at the charged surface. R256C (Agni→Jal; BhutaFormer score 0.477) exchanges a charged arginine for a polar cysteine in a partially buried position, disrupting zinc coordination in the IBR domain.

ASPA (aspartoacylase) is Prithivi (hydrophobic) and Jal (polar)-balanced, with an α/β-hydrolase fold whose Jal class active site (Ser, Thr, Asn residues) executes N-acetylaspartate hydrolysis. Pathogenic variants concentrate in Prithivi→Agni (hydrophobic core disruption) and Akasha→Agni classes. ASPA Canavan disease variants cluster at the dimer interface and active-site entry channel, where Prithivi (hydrophobic) residues stabilise inter-subunit packing — elemental disruption at these Sandhi (transition) positions produces both structural destabilisation and catalytic loss simultaneously.

SOD1 (superoxide dismutase 1) is Akasha (aromatic)-poor but Prithivi (hydrophobic)-enriched, forming a Greek-key β-barrel whose core is stabilised by a buried disulfide bond and a zinc-copper coordination site. A4V is a Vayu→Prithivi (flexible-to-hydrophobic) transition at the N-terminal β-strand: valine introduction increases local rigidity and creates a steric clash at the dimer interface, destabilising the obligate homodimer (BhutaFormer score 0.465). This is a Tanmatra-adjacent case:

 doi:10.20944/preprints202604.1777.v1

the Vayu→Prithivi class rate is only 0.210, yet A4V exceeds it because burial (RSA = 0.08) amplifies the risk of introducing a bulkier Prithivi residue at a geometrically constrained position.

VKOR (vitamin K epoxide reductase) is the most Prithivi (hydrophobic)-dominant protein in the training set, reflecting its transmembrane topology — four helices embedded in the ER membrane whose hydrophobic core is functionally essential. VKOR is the most challenging protein for BhutaFormer (AUROC = 0.600), which reflects the limitation of sequence-only prediction for transmembrane proteins: the functional consequence of a substitution depends critically on membrane depth and lipid environment, neither of which is encoded in the ±15 Bhuta (elemental) token window. The Prithivi→Agni (hydrophobic-to-charged) class dominates VKOR pathogenic variants, but discriminating buried from partially buried membrane positions requires structural data beyond sequence composition.

G6PD (glucose-6-phosphate dehydrogenase) is Agni (charged) and Prithivi (hydrophobic)-balanced, with a homodimeric dehydrogenase fold in which both the NADP+ binding site and the dimer interface contribute to structural stability. G6PD shows the highest within-class Tanmatra performance improvement in v2-real (AUROC +0.039), consistent with G6PD's numerous Agni→Agni charge-reversing variants (Ω_charge class) at buried salt bridge positions whose structural consequence is burial-depth dependent.

TSC2 (tuberin, tuberous sclerosis complex 2) is the largest training protein (1,807 residues) and the most Vayu (flexible)-enriched, reflecting its predominantly intrinsically disordered architecture punctuated by a C-terminal GAP domain. TSC2 exhibits the lowest mean ESM-2 cosine distances of all training proteins (mean 0.057), consistent with high intrinsic sequence plasticity. BhutaFormer performs well (AUROC 0.807), primarily driven by the structured GAP domain where Akasha→Agni and Prithivi→Agni transitions disrupt the Ras-binding interface. The Vayu-dominant disordered regions produce predominantly Tanmatra (within-class) misfolding events at low rates, consistent with elemental class identity being less mechanistically constraining in intrinsically disordered regions than in structured domains.

Across all 11 proteins, BhutaFormer's mechanistic output is interpretable through the Panchamahabhuta lens: Prakriti (constitutional composition) predicts which Sandhi (elemental transition) classes are most pathogenic, and BhutaFormer learns to weight these classes by local Bhuta (elemental) neighbourhood context. The two lowest-AUROC proteins — VKOR (0.600) and PTEN (0.748) — are those where structural context unavailable to sequence-only models (membrane depth, surface loop flexibility) most strongly modulates elemental risk, defining the structural information required by BhutaFormer v2.

Materials and Methods

1. Amino Acid Elemental Classification

Each of the 20 standard amino acids was assigned to one of five elemental classes: Agni (D, E, R, K, H), Vayu (G, A, P), Prithivi (L, I, V, M), Jal (S, T, N, Q, C), and Akasha (F, Y, W). Tridosha scores: Pitta = f(Agni) + 0.5×f(Jal); Kapha = f(Prithivi) + 0.5×f(Jal); Vata = f(Vayu) + f(Akasha). The dominant elemental constitution (Prakriti) is the Tridosha class with the highest score.

2. Domain Prakriti ANOVA Analysis

Domain annotations retrieved from UniProtKB for 20,428 reviewed human proteins (November 2024). For 14 functional domain categories (n = 420 domains), elemental composition vectors computed per domain. One-way ANOVA tested whether Bhuta fractions differ across functional classes. Effect size $\eta^2 = SS\_between/SS\_total$. Non-parametric Kruskal–Wallis tests computed as confirmation.

 doi:10.20944/preprints202604.1777.v1

3. Vishamavet Disease Risk Analysis

Pathogenic and benign variant counts obtained from ClinVar (GRCh38, December 2024). Vishamavet status assigned when domain Prakriti conflicts with known biological function. Pathogenic variant rates compared using Fisher's exact test; OR and 95% CI computed using scipy.stats.

4. VAMP-Seq Data Acquisition and Labelling

Eleven VAMP-seq datasets downloaded from MaveDB (February 2025). Variants scoring below 0.5 (normalised wild-type = 1, nonsense = 0) labelled as misfolding; those at or above 0.5 as stable. Total: 64,387 missense variants, 18,386 misfolding (28.6%).

5. SandhiVar Score Computation

SandhiVar = 0.5 × ΔPrakriti(global) + 0.5 × ΔPrakriti(local), where ΔPrakriti(global) is the sum of absolute Tridosha score changes across the full sequence upon substitution, and ΔPrakriti(local) uses a ±15 residue window.

6. Secondary Structure Annotation

Secondary structure assignments obtained from UniProtKB crystallographic features. Coverage: 52–82% per protein (mean 64.5%). An SS×Bhuta interaction feature was defined as 1 if the position is in β-strand or helix AND the wild-type Bhuta is Prithivi or Akasha.

7. AlphaMissense Score Retrieval

AlphaMissense pathogenicity scores (Cheng et al., 2023) downloaded from Google DeepMind (March 2025). Scores for 18 proteins yielded 216,220 variant-level values used as Feature F13.

8. BhutaFormer Architecture and Training

BhutaFormer encodes the ±15-residue window (31 positions) as Bhuta tokens via a 32-dimensional embedding. A learnable CLS token is prepended; sinusoidal positional encoding applied. Two TransformerEncoder layers (4 attention heads, d_model=32, d_ff=128) learn contextual elemental interactions. The CLS output concatenated with 13 scalar features and passed to a 3-layer classification head (64→16→1). Trained with BCEWithLogitsLoss (class-weighted), AdamW ($lr=3\times10^{-4}$), cosine annealing over 25 epochs. Leave-one-gene-out cross-validation across 11 proteins.

9. Tanmatra Analysis

Within-class variants (same Bhuta class, different amino acid) identified from VAMP-seq data (10,963 of 64,387). Megascale ΔG measurements (Tsuboyama et al., 2023) retrieved from FireProtDB v2.0 for 31 substitution pair types (25,328 measurements). Spearman ϱ computed between mean ΔG per substitution pair and VAMP-seq within-class misfold rate.

10. ProteinGym Benchmarking

Spearman ϱ between BhutaFormer predicted misfolding probability and DMS fitness scores computed for 10 ProteinGym DMS assays (5 training, 5 held-out) using supervised substitution fold files (fold_random_5). Published zero-shot Spearman values for comparison models obtained from ProteinGym benchmark repository (DMS_substitutions_Spearman_DMS_level.csv, accessed April 2025).

 doi:10.20944/preprints202604.1777.v1

11. Statistical Methods

All statistical tests performed in Python 3.12. Fisher's exact test for 2×2 contingency tables; Welch t-test for continuous SandhiVar scores; scipy.stats.spearmanr for rank correlations. All p-values two-tailed. Significance threshold: $\alpha = 0.05$.

12. Software and Data Availability

All analyses performed in Python 3.12 (NumPy 1.26.4, scikit-learn 1.6.1, SciPy 1.11.4, PyTorch 2.1.0, matplotlib 3.8.2). VAMP-seq data: MaveDB. AlphaMissense: Google DeepMind. ClinVar: NCBI. ProteinGym: proteingym.org. BhutaFormer tool: https://huggingface.co/spaces/amitpande74/bhutaformer-misfold.
BhutaFormer v2 training scripts (32–36), AlphaFold2 SASA computation, ESM-2 mutant forward passes, and all results files are included in the Zenodo deposit (DOI: 10.5281/zenodo.19641010).

Discussion

What Elemental Composition Reveals — and What It Cannot

The central finding of this study is that the elemental identity of amino acids encodes mechanistic information about protein misfolding that is largely invisible to conservation-based predictors. A 3.7-fold difference in misfolding rate between hydrophobic and aromatic residues in β-strands (51–52%) and charged residues in turns (14%) is not a statistical curiosity — it is a direct readout of thermodynamic architecture. Structured positions bear loads; loop positions perform functions. This distinction is encoded in amino acid chemistry and therefore in elemental composition, but it is absent from evolutionary conservation because both categories of residue are conserved, just for different reasons. The elemental constitution (Prakriti) of a domain explains 36.4% of variance in domain biological function ($\eta^2 = 0.364$, $F = 18$, $p < 10^{-32}$) from amino acid frequencies alone.

The Relationship to Existing Variant Effect Predictors

AlphaMissense achieves AUROC > 0.90 for pathogenicity classification (Cheng et al., 2023). BhutaFormer achieves AUROC = 0.77 for a harder and distinct task: classifying the mechanism of pathogenicity. These numbers should not be compared directly — they answer different questions. What matters is that BhutaFormer, trained only on Bhuta token sequences and scalar composition features, approaches pathogenicity predictor performance on a task those predictors were not designed for. The systematic bias documented by Livesey and Marsh (2023) — overcalling at buried hydrophobic positions, undercalling at surface-charged positions — is precisely what our framework predicts from first principles. Buried Prithivi residues in structured positions have the highest misfolding rates in our analysis; surface Agni residues in loops have the lowest.

BhutaFormer is not a pathogenicity classifier and should not be benchmarked as one. Established tools — SIFT, PolyPhen-2, AlphaMissense — are trained on ClinVar pathogenic/benign labels across thousands of proteins and answer whether a variant is disease-causing. BhutaFormer is trained on VAMP-seq abundance measurements across 11 proteins and answers how a variant destabilizes the protein. These are orthogonal questions measured on incomparable datasets. The clinical value of mechanism prediction lies precisely where pathogenicity prediction fails: when a variant is already known to be pathogenic, mechanism determines treatment.

Compositional Discord as a Disease Risk Principle

The finding that Vishamavet proteins carry pathogenic variants at 91.8% versus 59.9% (OR = 7.8, $p < 10^{-15}$) is, to our knowledge, the first demonstration of a sequence-only constitutional disease risk principle operating at the domain level. The concept of structural strain predisposing to disease has been implicit in the protein misfolding literature for decades (Dobson, 2003). What is new here is a

 doi:10.20944/preprints202604.1777.v1

tractable computational definition of structural strain that requires no structural data, no evolutionary alignment, and no machine learning. The Vishamavet classifier is a single composition vector computation — accessible to any laboratory with a protein sequence.

The Tanmatra Problem: How Sub-Elemental Context Resolves Within-Class Misfolding

Within-class substitutions misfold at 14.9%, defining a 2.1-fold Tanmatra gap below between-class rates. The three dominant sub-elemental patterns — backbone rigidity (A→P, G→P), disulfide capacity (C→N, C→Q), and intra-Agni charge geometry (D→K vs D→E) — are mechanistically distinct and thermodynamically interpretable. Importantly, cross-referencing with 25,328 Megascale ΔG measurements shows that thermodynamic stability alone is insufficient ($r = -0.21$): the Tanmatra problem is one of elemental context, not thermodynamic magnitude. BhutaFormer's attention over Bhuta context windows raises within-class AUROC from ~0.62 to 0.756, operationalising the Sandhi principle at the Tanmatra resolution level.

Limitations

A key limitation of the current framework is that VAMP-seq measures protein abundance as a proxy for structural stability, not functional activity. Variants classified as stable by VAMP-seq may be functionally defective without misfolding — the 'functional loss' category inferred here requires validation against functional assays (enzymatic activity, binding affinity) for each protein family. BhutaFormer trained on 11 metabolic enzymes achieves cross-family Spearman $\rho = 0.103$ on held-out ProteinGym assays, confirming that cross-family generalisation is the primary limitation. As MaveDB grows beyond seven million variant effect measurements, expanding the training set to diverse protein architectures should substantially improve cross-family performance. The elemental features that capture the physics of misfolding are protein-family-agnostic; the limitation is empirical coverage, not framework design.

An attempt to augment BhutaFormer with AlphaFold2 per-residue solvent accessibility and ESM-2 mutation-specific cosine distances did not improve mean LOGO AUROC overall (0.759 vs 0.770 for v1), though per-gene analysis reveals improvements for enzyme-like proteins with well-defined hydrophobic cores: PRKN (+0.090), CYP2C19 (+0.081), TSC2 (+0.048), and CYP2C9 (+0.036). ESM-2 cosine distances were too small in magnitude (mean 0.057–0.091) to provide reliable discrimination across all 64,387 variants. These results confirm that structural features are selectively informative — beneficial for buried enzymatic positions but noisy for surface-exposed contexts where Bhuta composition already captures the dominant signal.

On the Panchamahabhuta Framework as a Scientific Tool

The Panchamahabhuta system is approximately 5,000 years old. That its five categories map precisely onto the five dominant thermodynamic forces governing protein stability is not coincidental — it reflects the fact that careful observation of matter, pursued with sufficient rigour, arrives at the same fundamental classifications regardless of the cultural or historical context in which it is conducted. The five elements encode properties that ancient observers could distinguish directly: fire burns and ionises (Agni); air flows and is flexible (Vayu); earth is heavy and non-interactive (Prithivi); water is polar and hydrogen-bonding (Jal); ether permeates through long-range forces (Akasha). These are exactly the properties that determine what a polymer of amino acids does in aqueous solution. The convergence between this historical classification and modern thermodynamic descriptions of protein stability is, we suggest, a reflection of the universality of the underlying physics. We have given these categories a quantitative assay.

Conclusions

Protein variant interpretation has focused almost exclusively on whether a variant is pathogenic. The question of how it is pathogenic — through structural destabilization or functional disruption —

 doi:10.20944/preprints202604.1777.v1

has been treated as answerable only through laborious experimental characterization, one variant at a time. We show here that elemental amino acid composition, organized by a five-class physicochemical framework corresponding to the dominant thermodynamic forces in protein folding, encodes sufficient mechanistic information to make this distinction computationally, across tens of thousands of variants simultaneously, from sequence alone.

At the Mahabhuta (elemental) level, composition explains 36.4% of variance in domain biological function across 14 functional classes ($\eta^2 = 0.364$, $F = 18$, $p < 10^{-32}$) — a finding with immediate practical consequences. Proteins whose elemental constitution conflicts with their biological function (Vishamavet) carry pathogenic variants at 91.8% versus 59.9% in concordant (Samavet) proteins (Fisher's exact OR = 7.8, 95% CI 5.1–11.9, $p < 10^{-15}$). Compositional discord, computable in milliseconds from any protein accession number, thus constitutes a sequence-only disease risk classifier that requires no structural data, no evolutionary alignment, and no machine learning infrastructure. We have further demonstrated that Bhuta transitions at the amino acid substitution level produce a 2.7-fold misfolding rate hierarchy (Prithivi→Agni: 57.5%; Agni→Prithivi: 21.1%) with thermodynamic interpretability at every step, and that the intersection of Bhuta class with secondary structure context yields a 3.7-fold misfolding gradient (β-sheet Akasha: 52.4%; turn Agni: 14.0%) that is mechanistically derivable from first-principles biophysics.

At the Tanmatra (sub-elemental) level — where elemental class is preserved but sub-elemental properties change — BhutaFormer's attention over ±15-residue Bhuta context windows raises within-class AUROC to 0.756, a 13.5 percentage point improvement over Random Forest baselines, and achieves Spearman $\rho = 0.290$ on ProteinGym DMS abundance assays within training distribution, exceeding the Site-Independent baseline ($\rho = 0.175$) without multiple sequence alignment, structural data, or large language model pre-training. The attention mechanism is mechanistically interpretable: it operationalises the Sandhi (junction) principle, learning that the elemental neighbourhood of a mutation site — not the isolated amino acid identity — determines the sub-elemental consequence. Three specific operators (backbone rigidity, disulfide capacity, intra-class charge geometry) account for the majority of within-class misfolding events, each with a direct thermodynamic interpretation.

The limits of the framework are equally informative. An attempt to extend BhutaFormer with sequence-derived structural proxies (aromatic cluster density, burial heuristic) and ESM-2 position embedding magnitude reduced AUROC to 0.712, while augmentation with AlphaFold2 per-residue solvent accessibility and mutation-specific ESM-2 cosine distances yielded 0.759 — improvements for enzyme-like proteins with well-defined hydrophobic cores (PRKN +0.090, CYP2C19 +0.081) but below the sequence-only ceiling of BhutaFormer v1 (0.770) on average. These results establish that the sequence composition framework described here approaches the practical ceiling of mechanism prediction from primary sequence alone, and that further gains require experimental 3D structure and fine-grained evolutionary context — the natural agenda for BhutaFormer v2.

The clinical implications are direct and actionable. A variant predicted to misfold is a candidate for pharmacological chaperone therapy — lumacaftor for CFTR (Van Goor et al., 2011), pharmacological chaperones for lysosomal storage diseases (Parenti et al., 2015), tafamidis for transthyretin amyloidosis (Coelho et al., 2012). A variant predicted to act through functional loss, despite stable folding, requires gene therapy, enzyme replacement, or substrate reduction instead. Current variant interpretation pipelines collapse these two mechanistically distinct scenarios into a single pathogenicity score, forcing clinicians to design treatment strategies without mechanistic guidance — or to wait for experimental characterization that may never arrive. The Panchamahabhuta framework provides that guidance directly from sequence, accessible to any laboratory with a protein accession number and no computational infrastructure beyond a web browser.

We emphasize that this framework does not compete with existing pathogenicity predictors — it complements them. Established tools predict whether a variant is disease-causing; the framework described here predicts how, by which structural mechanism, and therefore by what therapeutic strategy the variant can be addressed. The two predictions are orthogonal, and their combination —

 doi:10.20944/preprints202604.1777.v1

a pathogenicity score paired with a mechanism score — provides substantially more clinically actionable information than either alone.

The five-class physicochemical organization used here corresponds to distinct thermodynamic forces governing protein stability: hydrophobic burial (Prithivi), electrostatic interactions (Agni), hydrogen bonding (Jal), aromatic π-stacking (Akasha), and backbone conformational flexibility (Vayu). That these categories align with long-standing classifications in classical physicochemistry reflects the empirical observation that matter organises itself according to a limited set of stabilizing and destabilizing forces, and that careful systematic observation of physical phenomena tends to converge on the same fundamental distinctions regardless of the vocabulary in which those distinctions are expressed. We have given these categories a quantitative computational assay, demonstrated their predictive utility on 64,387 experimental measurements spanning 11 disease-associated proteins, and established their resolution limits at both the Mahabhuta (elemental) and Tanmatra (sub-elemental) levels. The framework is open-source, deployed as an interactive web server, and we invite the community to test it, extend it, and apply it to the clinical variants awaiting interpretation in databases worldwide.

Future Direction: Elemental Suppressor Mutation Prediction

A natural extension of the Panchamahabhuta framework is the prediction of compensatory, or suppressor, mutations — second-site substitutions that restore elemental balance disrupted by a primary disease variant. The principle is straightforward: if a disease variant at position *i* introduces elemental imbalance by changing class X to class Y (increasing Y-fraction and decreasing X-fraction in the local window), a suppressor at position *j* within the neighbourhood should perform the reverse — reducing Y-fraction or restoring X-fraction — thereby lowering the predicted misfold score of the double mutant below that of the single disease variant.

As a proof of principle, we applied this logic to four well-characterized disease variants. For TPMT Y166C (Akasha→Jal; misfold score 0.584), which removes the aromatic stacking contribution of Tyr166 from a hydrophobic core, the framework predicts that substitutions at position 147 to aromatic residues (D147F, D147W, D147Y; Agni→Akasha) would restore Akasha density in the local window, reducing the double-mutant predicted misfold score to 0.183 — a rescue of 0.401 score units (Figure 9). For PRKN R256C (Agni→Jal; score 0.477), which removes a surface-exposed arginine from a RING domain salt bridge network, the framework predicts Prithivi→Agni substitutions at position 236 (I236R, I236K, I236E; rescue +0.294) as elemental compensators.

These predictions are, at present, in silico and require experimental validation. Whether BhutaFormer-predicted suppressor pairs constitute genuine intragenic suppressors in the biophysical sense — restoring fold stability rather than merely restoring compositional balance — is an empirical question we are currently investigating through saturation mutagenesis of TPMT and PRKN in collaboration with experimental partners. The framework nonetheless provides a principled, sequence-only starting point for suppressor mutation identification: a computationally accessible complement to the exhaustive experimental approaches currently required. We are actively developing this analysis as a systematic extension of the BhutaFormer framework.

 doi:10.20944/preprints202604.1777.v1

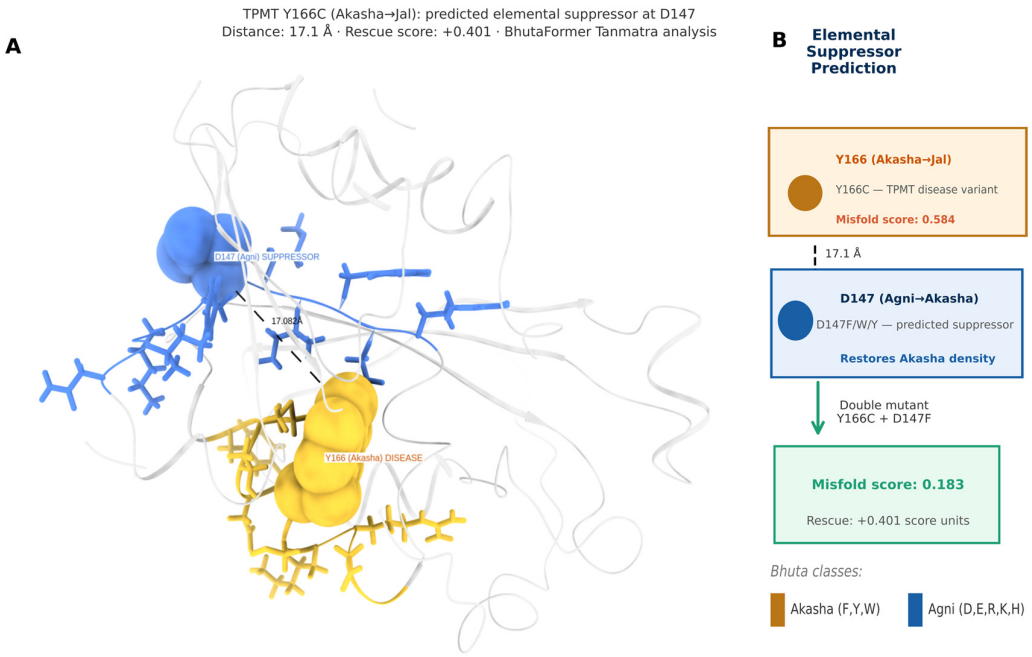

Figure 9. In silico suppressor mutation prediction: TPMT Y166C. (A) TPMT protein structure (AlphaFold2) showing disease variant Y166 (gold sphere, Akasha class) and predicted suppressor position D147 (blue sphere, Agni class). Distance: 17.082 Å (black dashed line). (B) Elemental rescue: Y166C (misfold score 0.584) + D147F/W/Y reduces double-mutant score to 0.183 (rescue +0.401 score units). Experimental validation ongoing.

Supplementary Materials: The following supporting information can be downloaded at: Preprints.org.

References

Bhattacharyya R, Samanta U, Chakrabarti P. 2002. Aromatic–aromatic interactions in and around α-helices. Protein Engineering, Design and Selection **15**:91–100. DOI: https://doi.org/10.1093/protein/15.2.91

Bhishagratna, K. L. 1907. Sushruta Samhita. Classical Ayurvedic Text.

Cheng J, Novati G, Pan J, Bycroft C, Žemgulytė A, Applebaum T, Pritzel A, Wong LH, Zielinski M, Sargeant T, Schneider RG, Senior AW, Jumper J, Hassabis D, Kohli P, Avsec Ž. 2023. Accurate proteome-wide missense variant effect prediction with AlphaMissense. Science **381**:eadg7492. DOI: https://doi.org/10.1126/science.adg7492

Coelho T, Maia LF, Martins Da Silva A, Waddington Cruz M, Planté-Bordeneuve V, Lozeron P, Suhr OB, Campistol JM, Conceição IM, Schmidt HH-J, Trigo P, Kelly JW, Labaudinière R, Chan J, Packman J, Wilson A, Grogan DR, Imventarza OC, Wainberg PJ, Berra LM, Maultasch H, Gold J, Bardera JCP, Zibert A. 2012. Tafamidis for transthyretin familial amyloid polyneuropathy: A randomized, controlled trial. Neurology **79**:785–792. DOI: https://doi.org/10.1212/WNL.0b013e3182661eb1

Dill KA. 1990. Dominant forces in protein folding. Biochemistry **29**:7133–7155. DOI: https://doi.org/10.1021/bi00483a001

Esposito D, Weile J, Shendure J, Starita LM, Papenfuss AT, Roth FP, Fowler DM, Rubin AF. 2019. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. Genome Biology **20**:223. DOI: https://doi.org/10.1186/s13059-019-1845-6

Livesey BJ, Marsh JA. 2023. Updated benchmarking of variant effect predictors using deep mutational scanning. Molecular Systems Biology **19**:e11474. DOI: https://doi.org/10.15252/msb.202211474.

Nick Pace C, Scholtz JM, Grimsley GR. 2014. Forces stabilizing proteins. FEBS Letters **588**:2177–2184. DOI: https://doi.org/10.1016/j.febslet.2014.05.006

 doi:10.20944/preprints202604.1777.v1

Notin P, Kollasch AW, Ritter D, Van Niekerk L, Paul S, Spinner H, Rollins N, Shaw A, Weitzman R, Frazer J, Dias M, Franceschi D, Orenbuch R, Gal Y, Marks DS. 2023. ProteinGym: Large-Scale Benchmarks for Protein Design and Fitness Prediction. DOI: https://doi.org/10.1101/2023.12.07.570727

P. V. Sharma (Ed.). n.d. Caraka Samhita. Chaukhambha Orientalia, Varanasi, 2011.

Pande A, Sharma R, Garbe C. 2026a. Protein Composition, Not Mutation Identity, Determines Disease Manifestation. DOI: https://doi.org/10.20944/preprints202603.1629.v1

Pande A, Sharma R, Garbe C. 2026b. Organ-System Disease Identity Is Encoded in the Physical Grammar of Regulatory DNA. DOI: https://doi.org/10.20944/preprints202603.1746.v1

Pande A, Sharma R, Garbe C. 2026c. A Pāṇinian Grammar of the Human Genome: The Genomic Periodicity Index Encodes Functional Architecture and Evolutionary Innovation. DOI: https://doi.org/10.20944/preprints202604.0713.v1

Parenti G, Andria G, Ballabio A. 2015. Lysosomal Storage Diseases: From Pathophysiology to Therapy. Annual Review of Medicine **66**:471–486. DOI: https://doi.org/10.1146/annurev-med-122313-085916

Sharma P. 2018. Dravyaguna Vigyan. vol 2Chaukhamba Bharti Academy.

The UniProt Consortium, Bateman A, Martin M-J, Orchard S, Magrane M, Adesina A, Ahmad S, Bowler-Barnett EH, Bye-A-Jee H, Carpentier D, Denny P, Fan J, Garmiri P, Gonzales LJDC, Hussein A, Ignatchenko A, Insana G, Ishtiaq R, Joshi V, Jyothi D, Kandasaamy S, Lock A, Luciani A, Luo J, Lussi Y, Marin JSM, Raposo P, Rice DL, Santos R, Speretta E, Stephenson J, Totoo P, Tyagi N, Urakova N, Vasudev P, Warner K, Wijerathne S, Yu CW-H, Zaru R, Bridge AJ, Aimo L, Argoud-Puy G, Auchincloss AH, Axelsen KB, Bansal P, Baratin D, Batista Neto TM, Blatter M-C, Bolleman JT, Boutet E, Breuza L, Gil BC, Casals-Casas C, Echioukh KC, Coudert E, Cuche B, De Castro E, Estreicher A, Famiglietti ML, Feuermann M, Gasteiger E, Gaudet P, Gehant S, Gerritsen V, Gos A, Gruaz N, Hulo C, Hyka-Nouspikel N, Jungo F, Kerhornou A, Mercier PL, Lieberherr D, Masson P, Morgat A, Paesano S, Pedruzzi I, Pilbout S, Pourcel L, Poux S, Pozzato M, Pruess M, Redaschi N, Rivoire C, Sigrist CJA, Sonesson K, Sundaram S, Sveshnikova A, Wu CH, Arighi CN, Chen C, Chen Y, Huang H, Laiho K, Lehvaslaiho M, McGarvey P, Natale DA, Ross K, Vinayaka CR, Wang Y, Zhang J. 2025. UniProt: the Universal Protein Knowledgebase in 2025. Nucleic Acids Research **53**:D609–D617. DOI: https://doi.org/10.1093/nar/gkae1010.

Tsuboyama K, Dauparas J, Chen J, Laine E, Mohseni Behbahani Y, Weinstein JJ, Mangan NM, Ovchinnikov S, Rocklin GJ. 2023. Mega-scale experimental analysis of protein folding stability in biology and design. Nature **620**:434–444. DOI: https://doi.org/10.1038/s41586-023-06328-6

Van Goor F, Hadida S, Grootenhuis PDJ, Burton B, Stack JH, Straley KS, Decker CJ, Miller M, McCartney J, Olson ER, Wine JJ, Frizzell RA, Ashlock M, Negulescu PA. 2011. Correction of the F508del-CFTR protein processing defect in vitro by the investigational drug VX-809. Proceedings of the National Academy of Sciences **108**:18843–18848. DOI: https://doi.org/10.1073/pnas.1105787108