PRACTICE PROJECT · COMPUTATIONAL BIOCHEMISTRY

Modeling Structural and Functional Patterns of Proteins

Do parameters of protein portray any predictive patterns?

Abir Hossain

March 30, 2026

Keywords: computational biochemistry · computational modeling · protein · data science

Contents

1. Executive Summary

This study investigates the relationship between a protein's functional classification, its structural characteristics, and its amino acid sequence. By analyzing proteins as polymers with quantifiable physicochemical properties — such as **molecular weight**, **hydrophobicity**, **amino acid composition**, and **entropy** — we aim to uncover predictive patterns within the dataset. Protein data from three different organisms (30,000 proteins) were obtained from UniProt to ensure both reliability and diversity in the sample.

Proteins are chains of 20 distinct amino acids, where the `R` group is the **variable side chain**. Although the chemical properties of individual amino acids are comprehensively established, forecasting the behavior of the complete chain inside an actual biological system presents a significant difficulty. Optimal chemical conditions seldom occur in reality, resulting in noise and unpredictability. This intricacy offers a significant opportunity for computational methods.

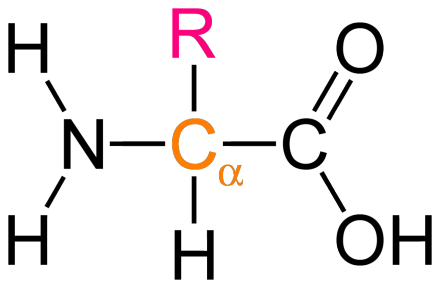

Figure 1: General structure of an amino acid.

The study concludes: despite hundreds of millions of years of evolutionary divergence, physicochemical characteristics of over 30,000 proteins from three mammalian proteomes remained remarkably intact. Instead of being arranged into distinct clusters, protein sequence space is a continuous physicochemical manifold in which location indicates biophysical function rather than species identity.

2. Purpose of This Project

This project is the fourth in a set of hands-on data science activities. The main objective was to become more proficient with the **computational ecosystem of Python**, with a focus on using **machine learning modeling tools** on biological data. A significant focus was placed on understanding and implementing an **ETL (Extract, Transform, Load) pipeline**.

The insights gained here directly informed subsequent analyses, evidently the latter study on R&D Spendings of Countries and Results where the understanding of the purpose and process of ETL implementation became much clearer.

Coming from a physics background rather than biology, this study has been an exercise in stepping outside my comfort zone. It demonstrates the power of applying numerical and computational tools to hyper-specialized research fields.

3. Project Structure & Workflow

```
├── clean_data
├── raw_data
├── results
└── src
```

3.1. Data Extraction

Protein sequence data was retrieved from UniProt using its REST API. The scripts — `download_data_test_failed.py` and `download_data.py` — reflect an iterative approach where the first attempt exposed pagination and formatting issues corrected in the latter implementation. The downloaded data was stored in TSV format under the `raw_data` directory.

The data files: `data.tsv` — a small test pull to confirm response format and pagination — and `proteins_raw.tsv`, the full dataset. The data covers three organisms: **Homo sapiens** (15,000 proteins), **Mus musculus** (7,000 proteins), and **Bos taurus** (8,000 proteins).

3.2. Data Cleaning

Raw data quality turned out to be high. The `validate.py` script performed upfront checks on sequence length, amino acid characters and sequence validity.

> Of the 75 sequences initially flagged as invalid, most were later found to be biologically legitimate. The amino acid codes **X**, **O**, and **U** — which triggered the original rejection — are in fact valid for the organisms in this dataset: X represents an unknown or ambiguous residue, O denotes pyrrolysine, and U denotes selenocysteine, both of which occur in mammalian proteomes.

`clean.py` then handled the actual cleaning. The cleaned dataset was saved in `clean_data` as `proteins_clean.tsv` sequences that failed validation were written to `removed_sequences.tsv`.

3.3. Feature Extraction

`feature.py` computed four per-sequence descriptors stored in `features.tsv` under the `results` directory:

- **Molecular Weight** — estimated from amino acid composition
- **Kyte–Doolittle Hydrophobicity** — a residue-averaged score reflecting the overall hydrophobic character of the sequence
- **Shannon Entropy** — a measure of compositional complexity based on residue frequency distributions
- **Sequence Length** — recalculated as an independent verification step

3.4. Analysis

`stats.py` was scripted for a lightweight quality check, grouping sequences by organism and computing feature means as a sanity check before deeper analysis.

Each computational parameter was selected to address a different structural question:

- **Correlation Parameter (Pearson's r)** — determines the linear relationship between two continuous parameters, ranging from -1 (inverse) to $+1$ (proportional).
- **Mutual Information** — measures non-linear statistical dependence between variables (units: bits). Finds intricate, non-monotonic relationships unlike correlation.

- **Allometric Scaling** — tests whether a protein feature scales as a power law with size: $Y = a \cdot X^b$. The exponent b reveals scaling regime: $b \approx 1$ (isometric), $b < 1$ (diminishing returns), $b > 1$ (accelerating).
- **PCA and PC Loadings** — correlated features are rotated into orthogonal axes arranged by variance explained. High absolute loadings show which features drive each axis.
- **Statistical Significance (ANOVA/Kruskal-Wallis + Bonferroni)** — tests whether feature distributions differ across organisms. Bonferroni correction adjusts for multiple testing ($p_{\text{adj}} = p \times n_{\text{tests}}$).
- **t-SNE** — non-linear dimensionality reduction preserving local neighborhood structure. Output coordinates cannot be interpreted separately; only relative distances and cluster patterns provide meaning.
- **Silhouette Score** — clustering quality measured as $s = \frac{b-a}{\max(a,b)}$ where a = mean intra-cluster distance and b = mean nearest-cluster distance. Scores >0.5 denote acceptable structure;

$$< 0.25$$

  indicates absence of natural clusters.
- **Outlier Proteins** — biologically extreme cases (intrinsically disordered regulators, transmembrane structural proteins) highlighted to identify functional specializations.

> - All python scripts can be found in the `src` directory.
> - The `results` directory houses all datasets and plots.
> - Organism name reference: Human = **Homo sapiens**, Cow = **Bos taurus**, Mouse = **Mus musculus**.

4. Results

4.1. Physicochemical Features by Organism

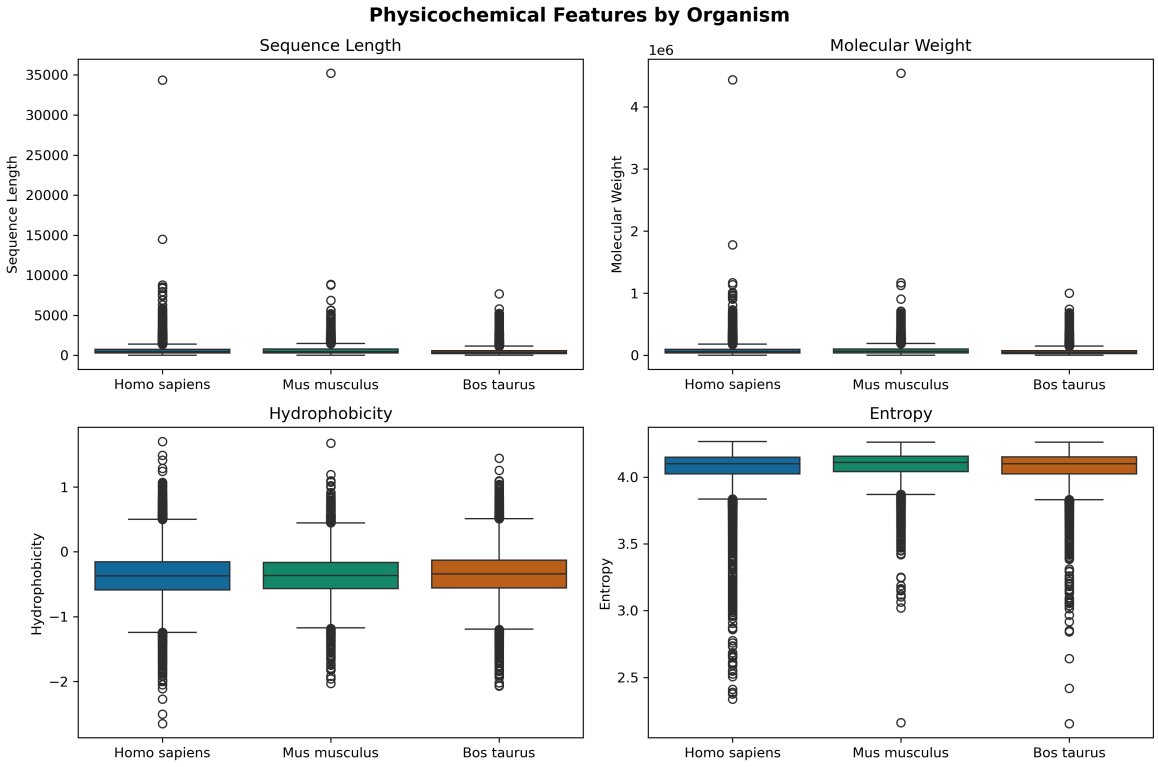

Figure 2: Physicochemical features by organism.

- **Sequence Length & Molecular Weight**: All three organisms show similar right-skewed distributions with medians around 400–500 amino acids (50 kDa), with extreme outliers reaching 15,000–35,000 residues. Typical proteome architecture is dominated by moderate-sized proteins.
- **Hydrophobicity**: Median values cluster near −0.4, indicating proteins are slightly hydrophilic overall — optimal for solubility in aqueous cellular environments.
- **Sequence Entropy**: High median values (4.1–4.2 bits) demonstrate that proteins across all species maintain high sequence complexity with diverse amino acid usage.

The striking similarity across Human, Mouse, and Cow indicates that **fundamental biophysical constraints on protein properties are conserved across 100 million years of mammalian evolutionary divergence**.

4.2. Feature Correlation Heatmap

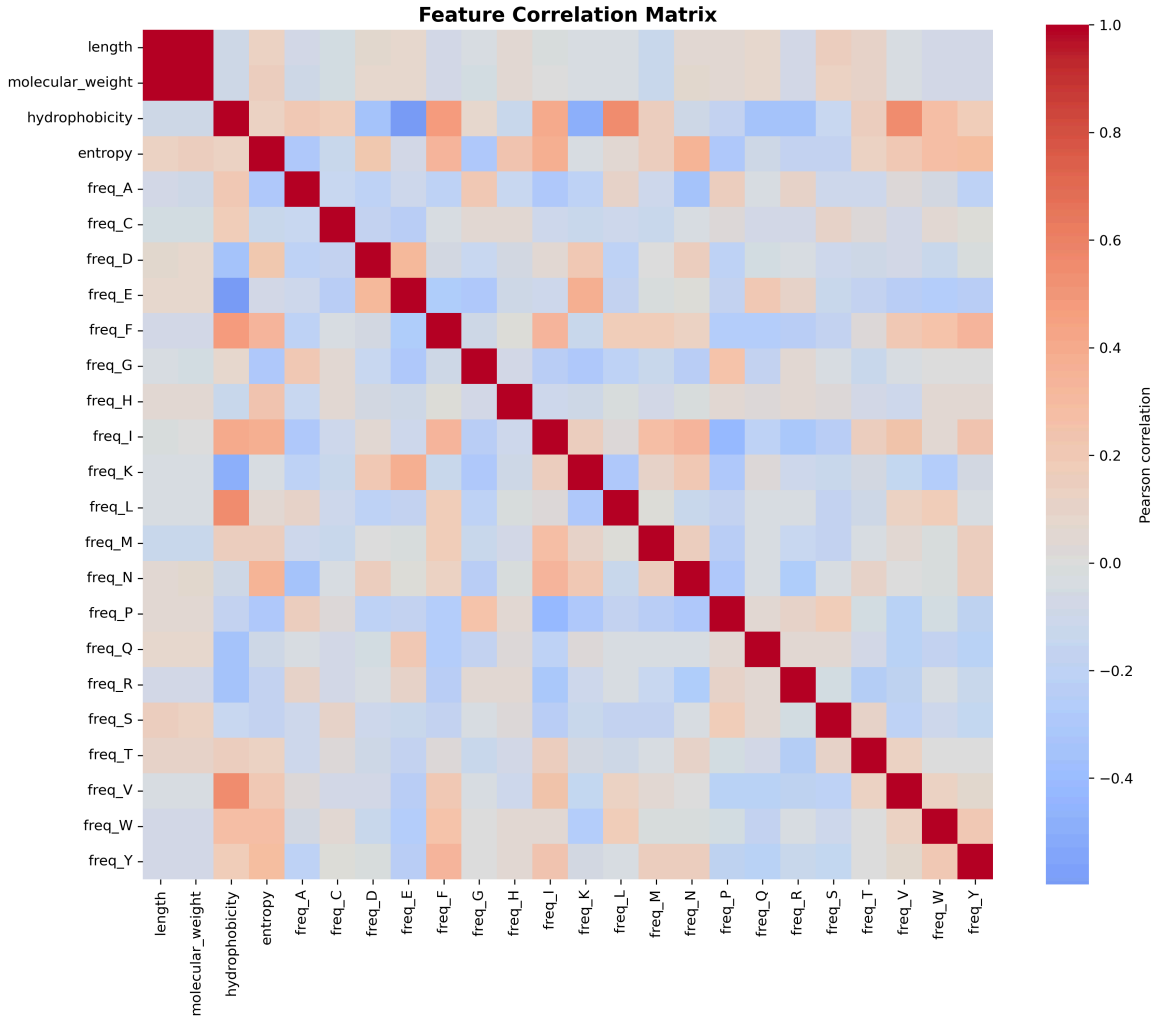

Figure 3: Feature correlation heatmap.

- **Length and molecular weight** show near-perfect correlation ($r > 0.9$), confirming linear scaling.
- **Hydrophobicity** correlates positively with I, L, M, V and negatively with E, K, Q — charged residues of the protein sequence.
- **Entropy and hydrophobicity** show predominantly weak correlations ($|r| < 0.4$), justifying their use as independent features throughout this study.

> The diagonal appears to have perfect correlation as a result of cross-matching the same features on both axes. Consider ignoring the perfectly matched diagonal when exploring the heatmap.

4.3. Top Amino Acid Pairs by Mutual Information

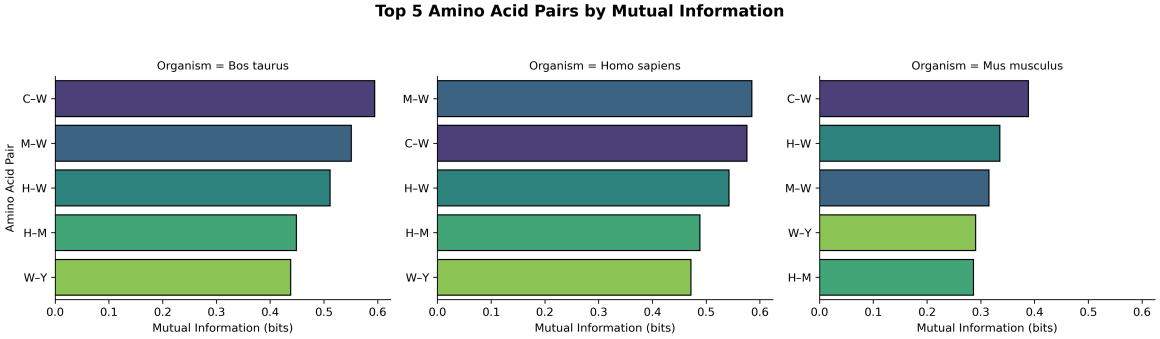

Figure 4: Top amino acid pairs by mutual information across the three organisms.

> Values of 0.5–0.6 bits represent strong statistical dependency — knowing one amino acid is present substantially increases the probability of finding its partner, far exceeding random expectation.

With 20 amino acids, there are 190 possible unique pairs. Displaying all across 3 organisms (190×3) would obscure the strongest signals, so the top 5 pairs per organism are discussed.

- **Tryptophan (W)** appears in 4 of 5 top pairs across all organisms. It is the **rarest amino acid** yet shows the strongest co-occurrence patterns, suggesting it occupies specialized structural or functional niches (active sites, protein–protein interfaces).
- **Cysteine** pairs like C–W and C–M rank highest in cow and human, reflecting their shared roles in metal binding, redox chemistry, and disulfide bridge formation.

The same amino acid pairs dominate across all three species with similar MI magnitudes (0.25–0.6 bits), indicating **universal constraints on protein sequence architecture rather than species-specific adaptations**.

4.4. Allometric Scaling with Protein Length

4.4.1. Hydrophobicity

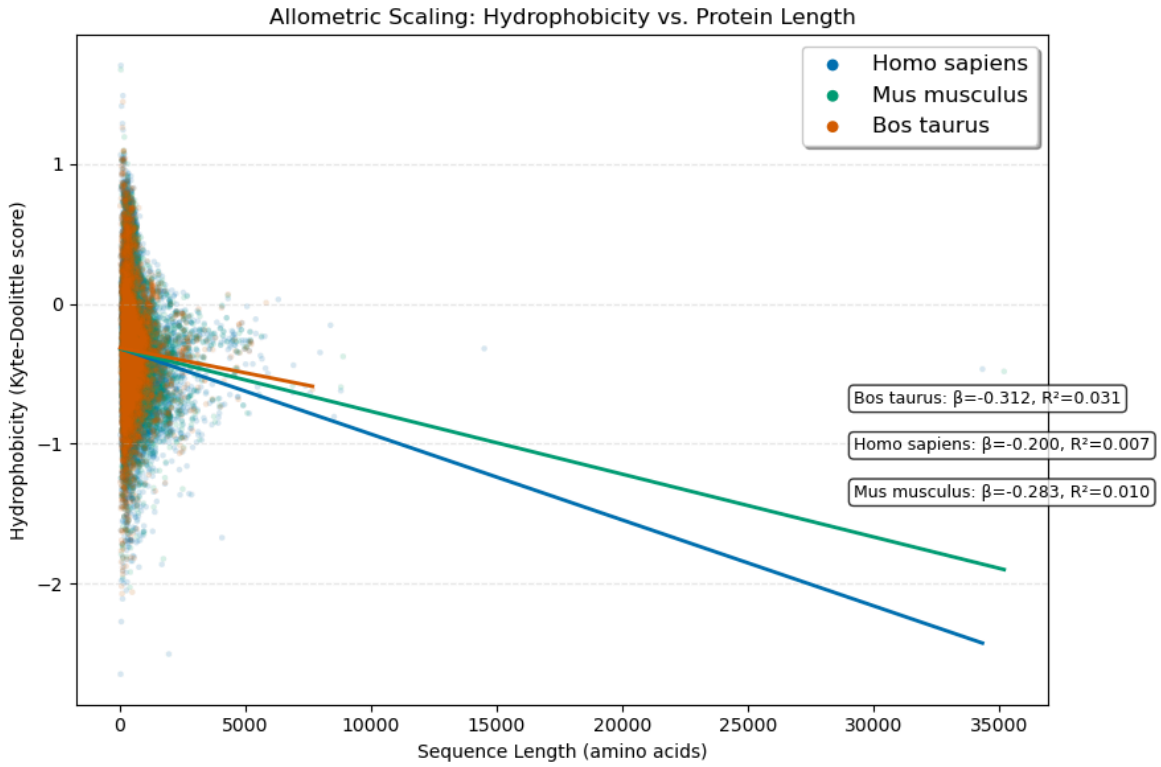

Figure 5: Allometric scaling of hydrophobicity with protein length.

All three organisms exhibit consistent negative slopes ($\beta = -0.20$ to -0.31), indicating that longer proteins are systematically less hydrophobic — suggestive of **larger proteins maintaining solubility by reducing hydrophobic residue content**. Low R^2 values (0.007–0.031) confirm length contributes little to hydrophobicity variance, yet the consistent direction and statistical significance ($p < 10^{-1}$) confirm a real biological trend.

4.4.2. Entropy

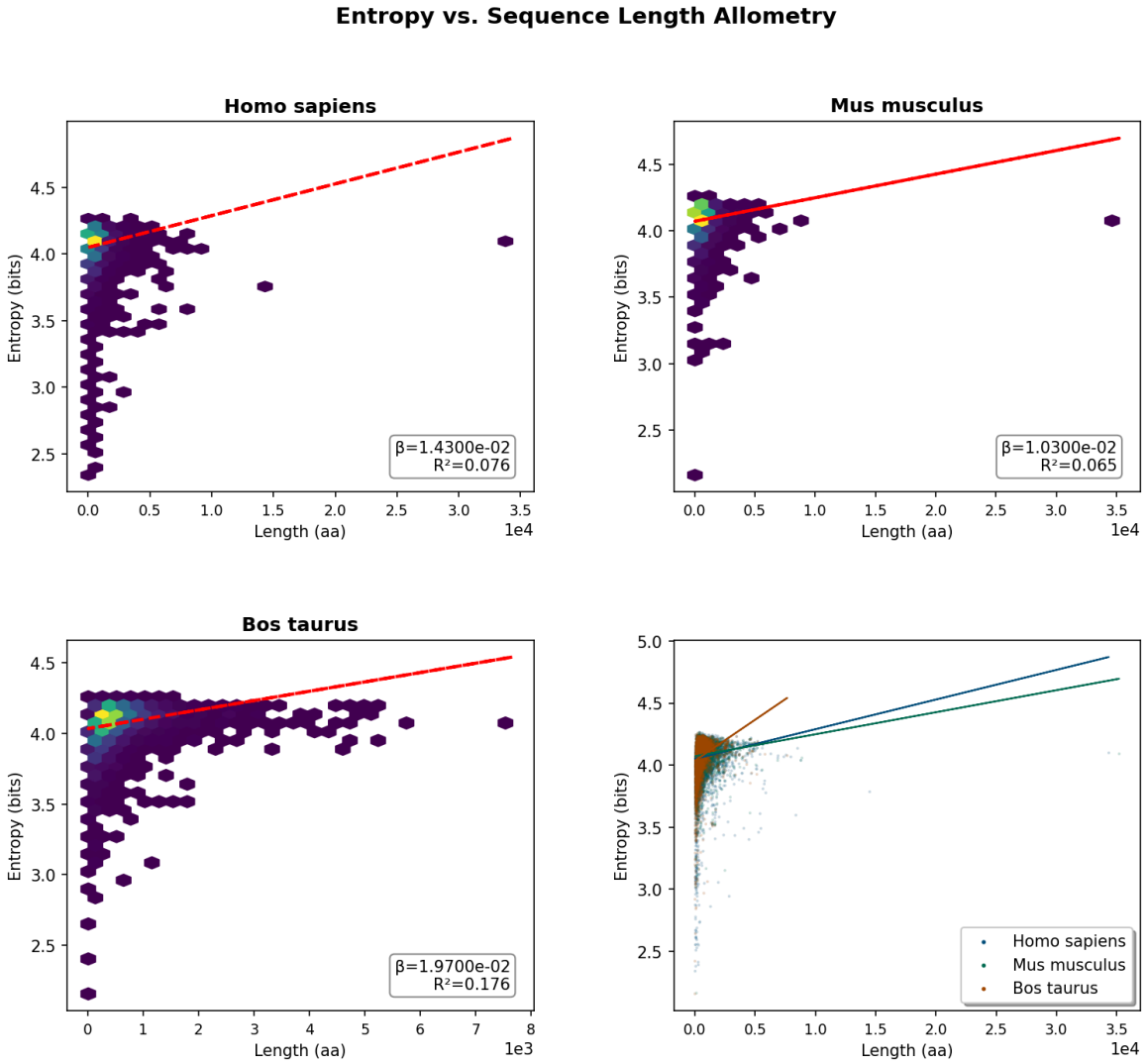

Figure 6: Allometric scaling of entropy with protein length.

Sequence entropy shows minimal dependence on protein length, with very weak positive slopes ($\beta = 0.01$–0.02). R^2 values of 0.065–0.176 mean length explains only 6.5–17.6% of entropy variance. **Compositional complexity is governed by factors other than protein size** — this contrasts sharply with hydrophobicity. Evolutionary pressures on sequence diversity operate independently of size constraints.

4.5. PCA Variance

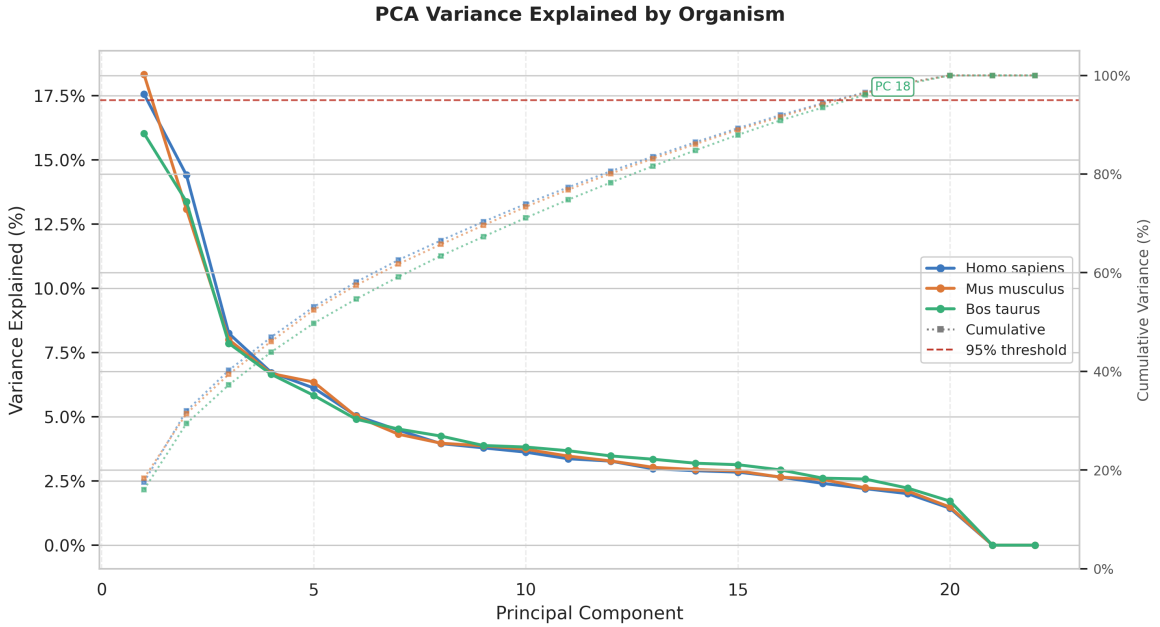

Figure 7: PCA scree plot showing variance explained per component for all three organisms.

All three species share a strikingly similar pattern: PC1 alone captures 16–18% of variance, followed by a steep elbow through PC5. The near-perfect overlap of all three curves implies **the dimensionality structure of proteome composition is highly conserved across mammals**. 95% of variance is not reached until PC18, meaning no single dominant axis exists and **the signal is genuinely multi-dimensional**. Retaining only the first few PCs would discard the majority of information.

4.6. PC1 Loadings Interpretation

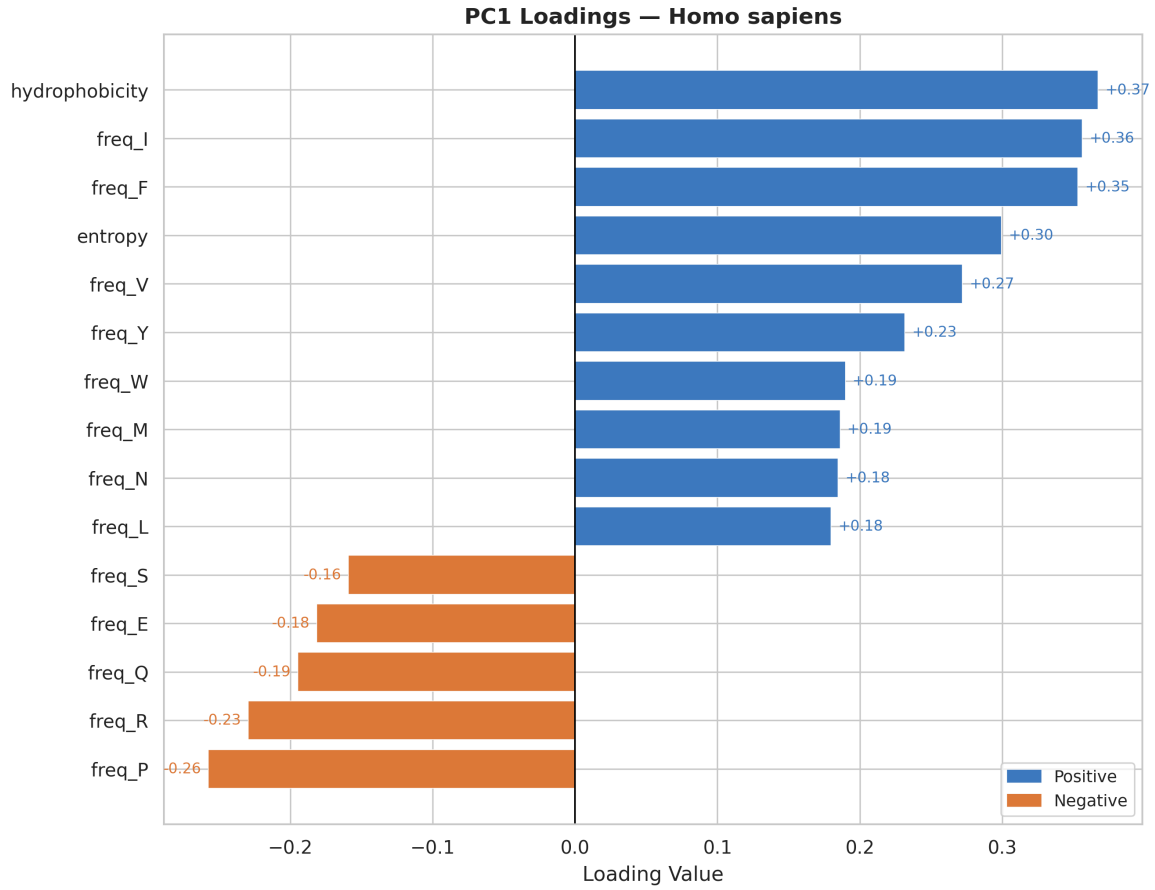

Figure 8: PC1 loadings — Human (Homo sapiens).

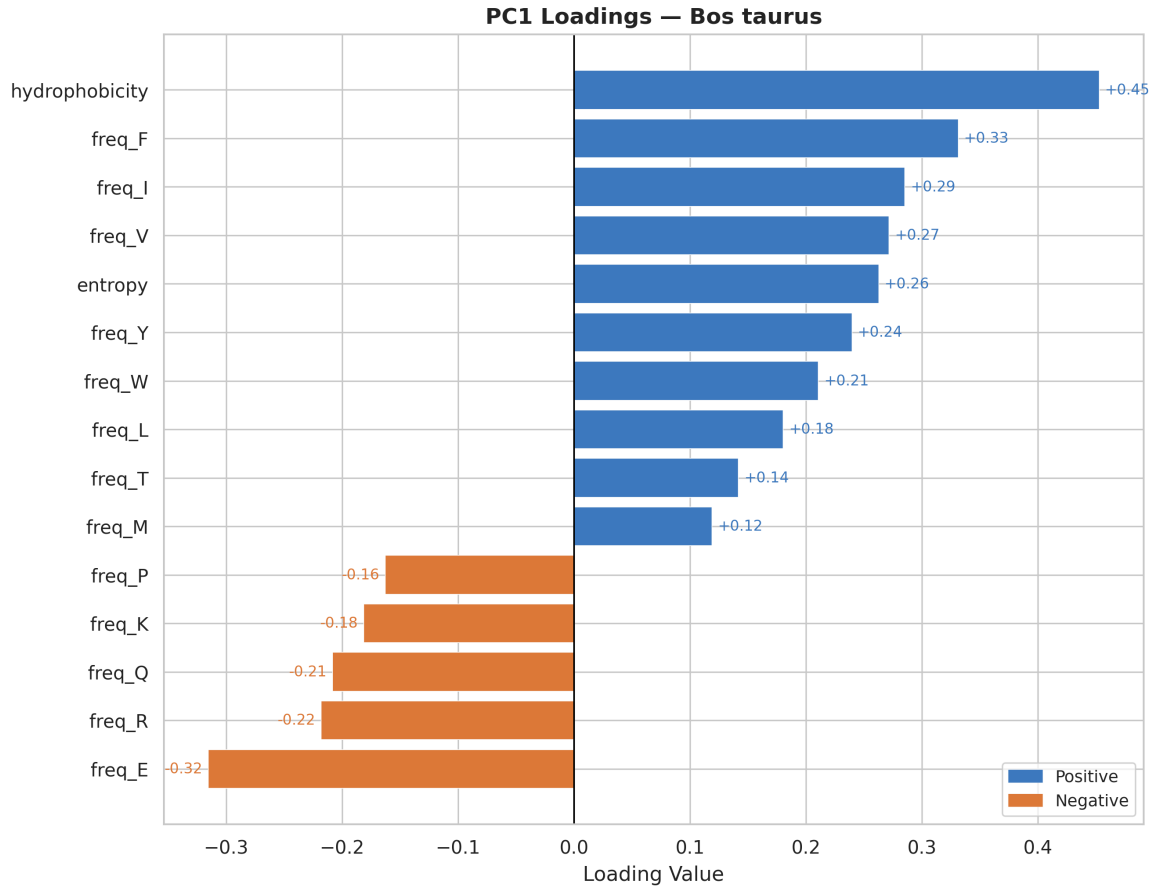

Figure 9: PC1 loadings — Cow (Bos taurus).

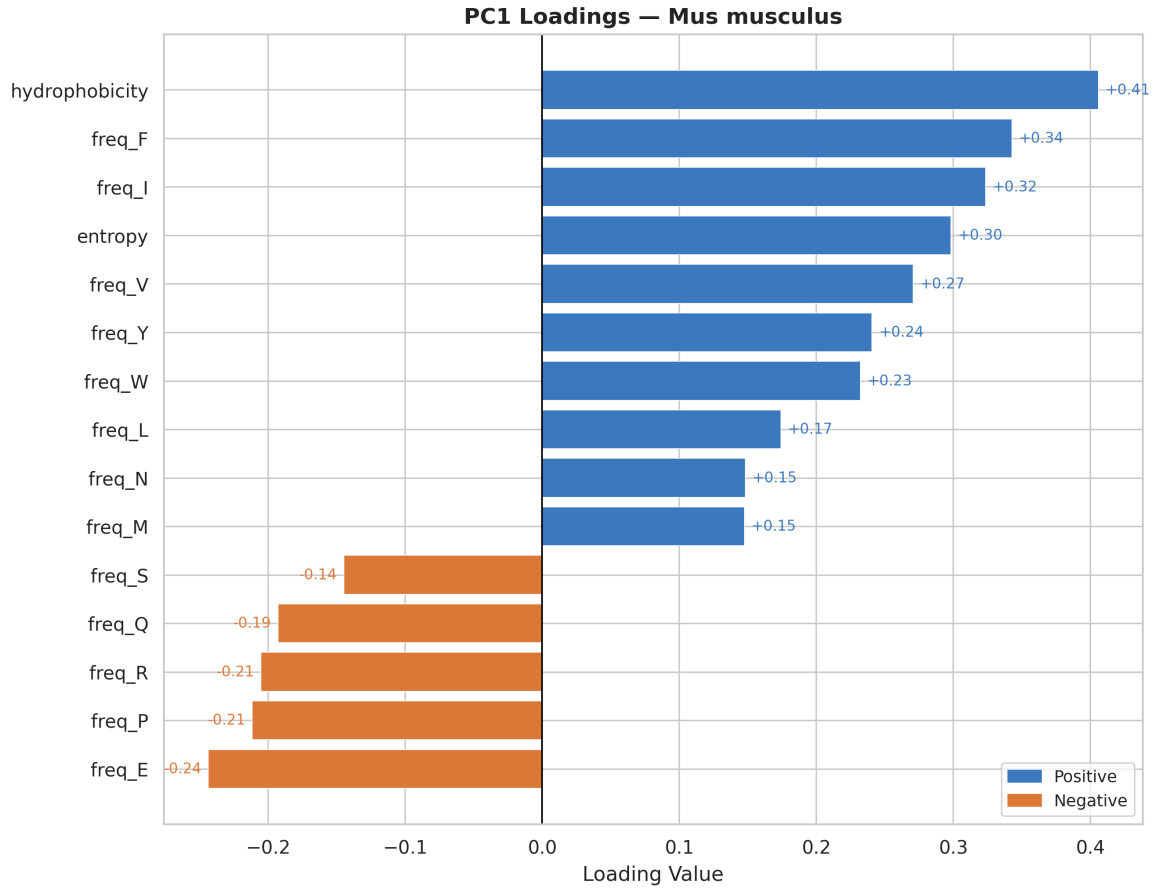

Figure 10: PC1 loadings — Mouse (Mus musculus).

- **Hydrophobicity** is the strongest positive contributor to PC1 across all organisms.
- **Hydrophobic amino acids** consistently show positive loadings — PC1 captures proteins enriched in non-polar residues.
- **Charged/polar residues** display negative loadings, representing surface-exposed, soluble protein characteristics.

PC1 effectively separates proteins by their **structural topology** — hydrophobic core dominated versus surface-exposed charged residue dominated architectures. The loading patterns are remarkably consistent across all three organisms.

4.7. PC1 vs PC2

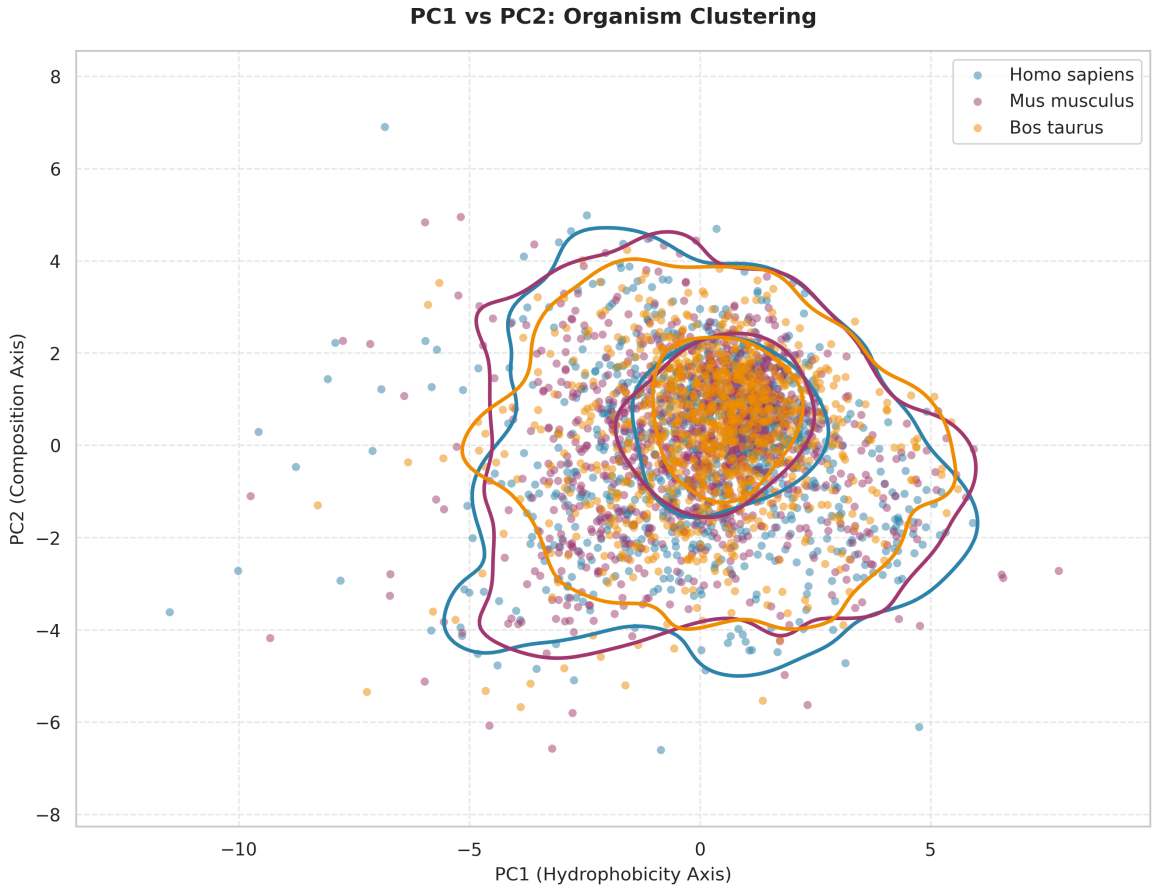

Figure 11: PC1 vs PC2 scatter plot with kernel density contours coloured by organism.

All three organisms occupy largely the **same region** of PC1–PC2 space with no distinct clustering by species. The horizontal axis (PC1) represents the "hydrophobicity axis"; the vertical axis (PC2) represents the "composition axis". The substantial overlap indicates that despite 100 million years of evolutionary divergence, mammalian proteins are constrained to similar physicochemical property spaces.

4.8. Feature Distribution Comparison

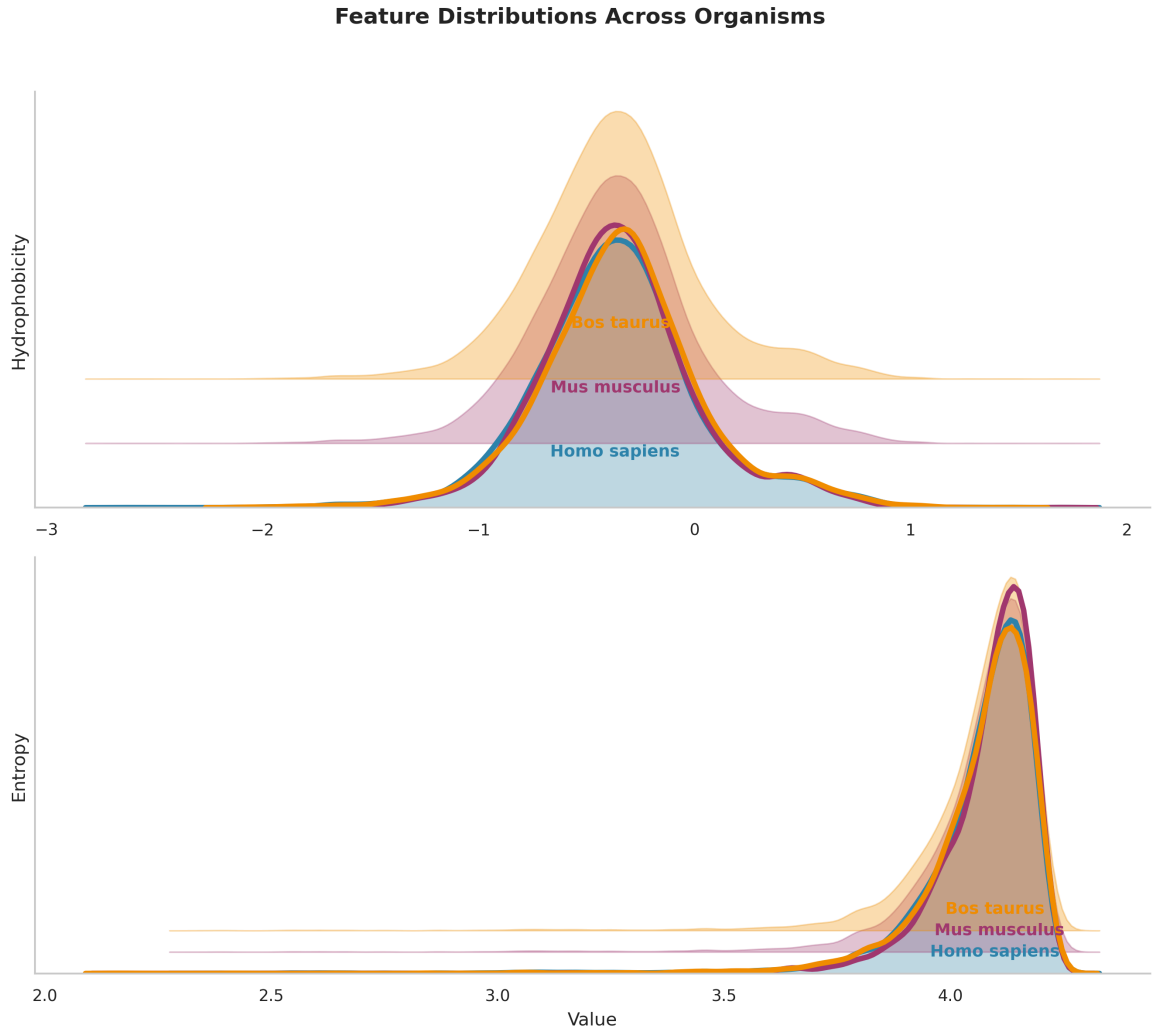

Figure 12: Probability density distributions of hydrophobicity (Panel A) and entropy (Panel B) across organisms.

Panel A (Hydrophobicity): All three organisms show right-skewed distributions with a primary peak around −0.4 to −0.3. All distributions show a long right tail extending to +2.0, representing the minority of highly hydrophobic proteins (likely membrane proteins).

Panel B (Entropy): Extremely narrow, left-skewed distributions with sharp peaks around 4.1–4.2 bits, indicating most proteins have very similar sequence complexity. The near-perfect alignment of entropy peaks across all three organisms reflects strong evolutionary constraints.

4.9. Statistical Significance

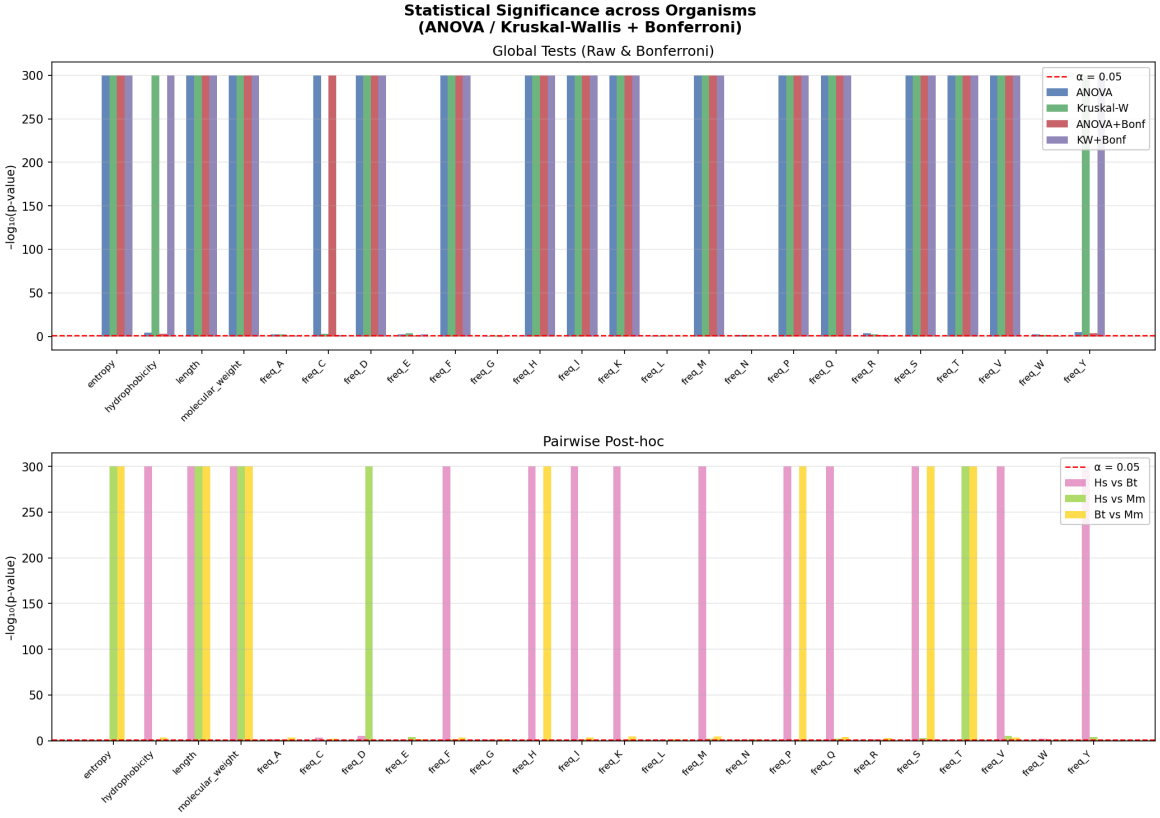

Figure 13: Statistical significance of feature differences across organisms (ANOVA/Kruskal-Wallis + Bonferroni correction). Bars above the red line ($\alpha = 0.05$) are significant; ceiling bars (y=300) indicate $p \approx 0$.

- **Global Tests**: Most features — length, weight, most amino acids — differ universally across all three species. Hydrophobicity is only significant in Kruskal-Wallis, suggesting outliers or non-normal distributions drive the difference.
- **Pairwise Drivers**: Length/Weight differ across all organism pairs. Entropy distinguishes Mouse from the Human/Cow cluster, while Hydrophobicity distinguishes Human from Cow specifically.
- Usage varies by residue: `freq_T` (Threonine) differs universally, while `freq_A` (Alanine) and `freq_E` (Glutamic Acid) are conserved across all three species.

4.10. Physicochemical Property Gradients (t-SNE)

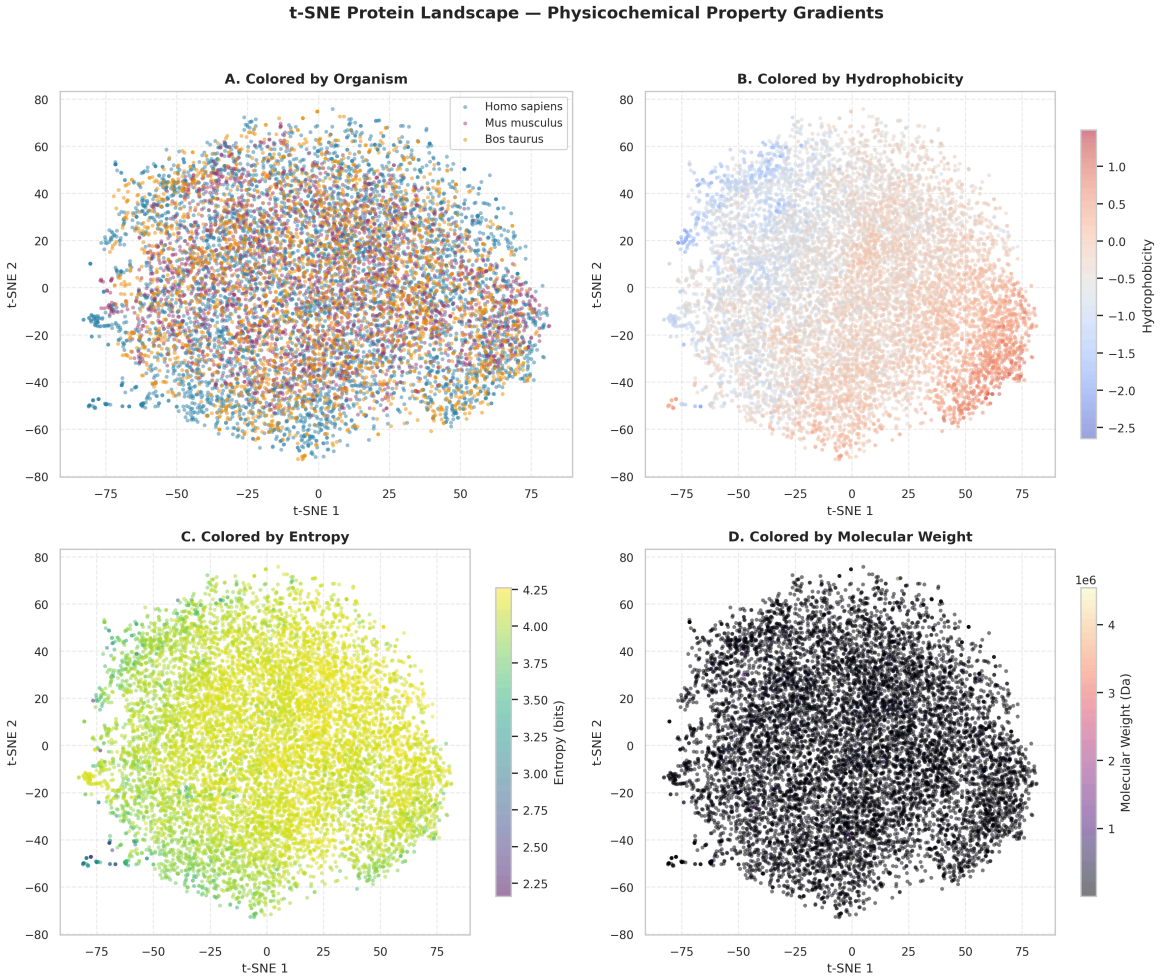

Figure 14: 4-panel t-SNE projection of proteins coloured by organism, hydrophobicity, entropy, and molecular weight.

- **Panel A (Organism)**: All three species intermix completely with no organism-specific clustering — evolutionary divergence has left no detectable species-specific signature at the global physicochemical level.
- **Panel B (Hydrophobicity)**: A smooth diagonal gradient from hydrophilic to hydrophobic, with no sharp transitions. Proteins continuously balance solubility requirements against structural stability.
- **Panel C (Entropy)**: High-entropy and low-entropy proteins show moderate spatial separation, suggesting sequence complexity correlates with specific physicochemical regions but is not the primary organizing axis.
- **Panel D (Molecular Weight)**: Largely independent of the t-SNE axes, suggesting protein size evolves orthogonally to hydrophobicity and sequence complexity.

4.11. Cluster Quality — Silhouette Analysis

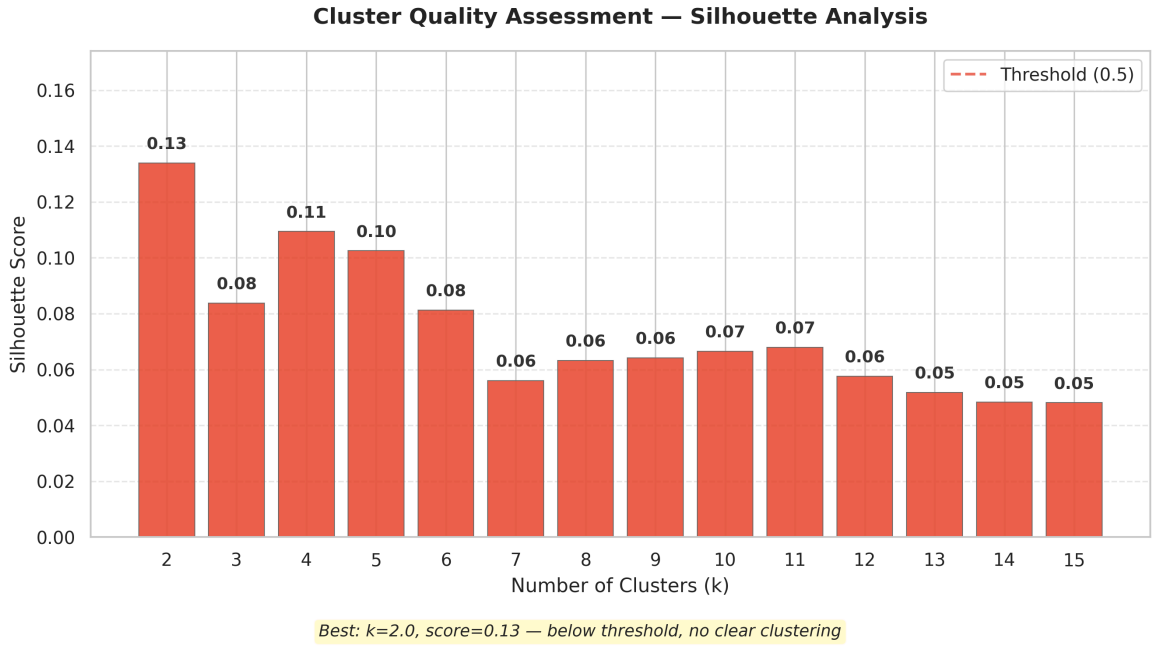

Figure 15: Silhouette scores for K-means clustering at k = 2–15. Dashed line at 0.5 marks the conventional threshold for meaningful structure.

All scores fall far below 0.5, ranging from 0.05 (k = 14–15) to 0.13 (k = 2). **Proteins do not naturally partition into discrete physicochemical groups.** Categorical labels like "hydrophobic" vs. "hydrophilic" impose artificial structure on a fundamentally continuous distribution. This negative result rules out clustering as an appropriate framework and points toward gradient-based or manifold-learning approaches.

4.12. Extreme Proteins — Outlier Analysis

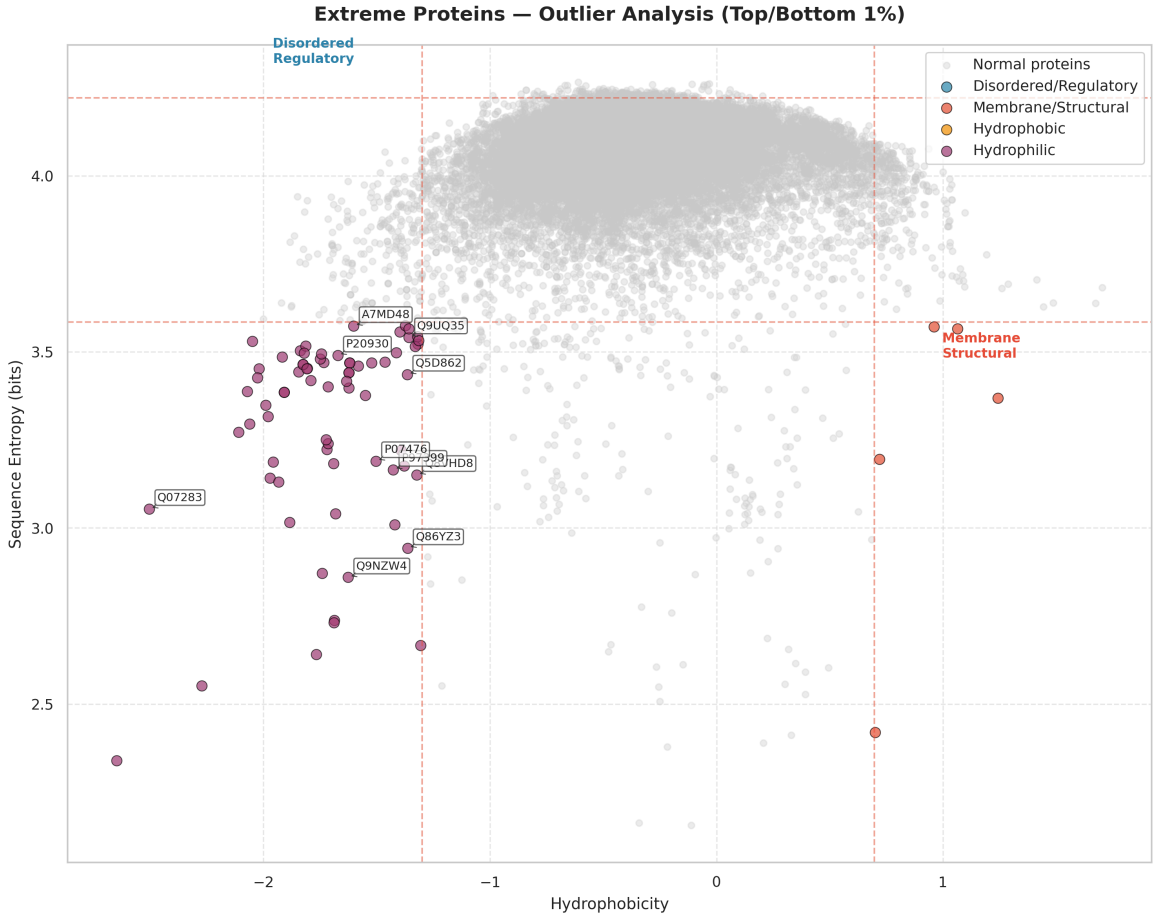

Figure 16: Outlier proteins mapped in hydrophobicity–entropy space. Dense central cloud (gray) = typical proteins; colored points = outliers in the top/bottom 1% of both features.

Most proteins cluster around hydrophobicity ≈ −0.5 to 0.0 and entropy ≈ 3.8–4.2 bits.

- **Low hydrophobicity + high entropy**: intrinsically disordered and regulatory proteins (transcription factors, signaling proteins).
- **High hydrophobicity + low entropy**: likely membrane proteins and structural components with repetitive motifs (transmembrane helices, collagen repeats).
- Very few proteins combine **high hydrophobicity with high entropy** — this combination promotes aggregation and is selected against.

5. Insights

The core finding is unusual from a common-sense perspective: no feature examined cleanly separates proteins by species. **The mammalian proteome, viewed through the lens of physicochemical properties, behaves as a single conserved landscape rather than separate species-specific ones.** This points to selection acting on biophysical viability rather than on sequence identity by itself.

- Categorical labels such as “hydrophobic” or “disordered” are useful shorthand but overall poor analytical categories. Any downstream modelling should treat these properties as coordinates on a manifold rather than discrete class memberships.
- Rare amino acids constrain local-sequence-context strongly. This points toward active sites and interface residues as disproportionate sources of co-occurrence signal.
- There is a genuine biological constraint (larger proteins reduce hydrophobic exposure to prevent aggregation) that explains only a small fraction of observed variance.
- 95% variance is not recovered until PC18, arguing for caution in any dimensionality-reduction strategy retaining only the first few components.

The findings reinforce a view of proteome evolution as operating under strong universal biophysical constraints — solubility, folding stability, resistance to aggregation — largely conserved across mammalian lineages while leaving room for species-level divergence in properties such as size distribution and specific residue frequencies.

6. The Horizon to be Explored

6.1. Possible Expansions

Enhancing this study by increasing taxonomic breadth — adding non-mammalian vertebrates, invertebrates, prokaryotes, plants, fungi, and archaea — would test whether the conserved physicochemical manifold observed here is a mammalian phenomenon or a universal feature of life. Extremophile proteomes from thermophiles, halophiles, and psychrophiles are of particular interest, as their proteins must function under conditions that should impose measurable shifts in hydrophobicity and compositional bias.

Structurally informed data from AlphaFold-predicted structures would enable the same physicochemical analyses to be performed more thoroughly. Applying supervised and deep learning models opens a new front: transformer-based protein language models (ESM-2, ProtTrans) encode richer representations than amino acid frequencies, and graph neural networks can incorporate structural context directly.

Stricter mathematical frameworks — topological data analysis to characterise the manifold geometry, information-theoretic measures beyond pairwise mutual information — would allow more rigorous tests of the continuity and dimensionality conclusions reached here.

6.2. Current State of Research

Computational biochemistry has undergone a step-change in the past decade, driven primarily by the application of deep learning to protein structure prediction. **AlphaFold2** (2021) and its successors have largely solved the single-chain structure prediction problem. The field has since moved toward **structure-based function prediction, protein–protein interaction modelling, and the inverse problem of de novo protein design**, where tools such as **RFdiffusion** and **ProteinMPNN** generate novel sequences with target folds.

Protein language models — trained on hundreds of millions of sequences — now produce dense vector representations capturing evolutionary, structural, and functional information simultaneously. At the systems level, proteome-scale mass spectrometry and single-cell proteomics are generating datasets of previously impossible scale and resolution.

7. Author's Note

Stepping outside my comfort zone into biochemistry — and the seemingly infinite world of proteins — has meaningfully boosted my confidence as a researcher. What fascinated me most were the patterns revealed by the results: counterintuitive, often unpredictable, and all the more rewarding for it.

My background in physics, with grounding in organic chemistry fundamentals, along with prior experience in data analysis and computational techniques, gave me the footing needed to navigate concepts that might otherwise have been overwhelming. Through this work, I deepened my practical knowledge of scikit-learn and Python-based machine learning, while independently exploring Julia libraries and GPU-accelerated computation.

PS: This has been a practice project in an ongoing endeavour to get a strong footing on data science and computational techniques. It is not to be viewed as a reference or a peer-reviewed scientific paper — only a humble effort of a curious mind fascinated by the possibilities.