Large-Scale Statistical Dissection of Sequence-Derived Biochemical Features Distinguishing Soluble and Insoluble Proteins

Nguyen Huy Hoang Vu [1], Bao Long Nguyen [2,*]

1 Faculty of Mathematics, Informatics and Mechanics, University of Science, Vietnam National University, Hanoi (VNU-HUS), Hanoi, Vietnam
2 School of Economics and Management, Hanoi University of Science and Technology (HUST), Hanoi, Vietnam

* baolongnguyen243@gmail.com

Abstract

Protein solubility critically influences recombinant expression efficiency and downstream biotechnological applications. While deep learning models have improved predictive accuracy, the intrinsic magnitude, redundancy, and interpretability of classical sequence-derived determinants remain insufficiently characterized. We performed a statistically rigorous large-scale univariate analysis on a curated dataset of 78,031 proteins (46,450 soluble; 31,581 insoluble). Thirty-six biochemical descriptors were evaluated using Mann–Whitney U tests with Benjamini–Hochberg false discovery rate correction. Effect sizes were quantified using Cliff's δ, and discriminative performance was assessed by ROC–AUC. Although 34 features remained significant after correction, most exhibited small effect sizes and substantial class overlap, consistent with a weak-signal regime. The strongest effects were associated with size-related features (sequence length and molecular weight; $\delta \approx -0.21$), whereas charge-related descriptors, particularly the proportion of negatively charged residues ($\delta = 0.150$; AUC = 0.575), showed consistent but modest shifts. Spearman correlation analysis revealed near-complete redundancy among major size-related variables (ρ up to 0.998). Applying a redundancy threshold ($|\rho| \geq 0.85$), we derived a parsimonious composite integrating sequence length and negative charge proportion, achieving AUC = 0.624 (MCC = 0.1746). These findings demonstrate that sequence-level solubility information is intrinsically low-dimensional and governed by coordinated weak effects, establishing a transparent statistical baseline for large-scale solubility characterization.

Introduction

Protein solubility is a fundamental physicochemical constraint governing protein folding, intracellular stability, recombinant expression efficiency, and downstream biotechnological deployment [1, 6]. Insolubility and aggregation frequently result in inclusion body formation, impaired functional yield, and substantial experimental cost during heterologous expression [2, 5]. From a molecular perspective, solubility reflects the balance between intermolecular attractive forces that promote aggregation and intramolecular stabilizing interactions that favor properly folded, soluble conformations. Consequently, understanding how primary sequence encodes this balance remains a central problem in computational and structural biology. A substantial body of experimental evidence indicates that solubility is modulated by multiple intrinsic

sequence-level properties. Chain length influences translational burden, folding kinetics, and the probability of partially folded intermediates. Amino acid composition shapes the global physicochemical landscape, including hydrophobic core formation, surface polarity, and electrostatic repulsion. Charge distribution affects long-range electrostatic interactions and colloidal stability, whereas hydrophobic clustering and aggregation-prone segments can nucleate intermolecular association [3, 4]. Importantly, these factors are not independent: size-related properties, mass, and hydrophobic content are often structurally coupled, reflecting shared underlying constraints of polypeptide architecture.

Early computational approaches sought to capture these determinants through interpretable sequence-derived descriptors, including amino acid composition profiles, Chou Fasman secondary structure propensities, and hydropathy scales such as the Kyte Doolittle index [7, 8]. These features are computationally efficient and mechanistically transparent; however, their relative contribution and practical magnitude at scale remain insufficiently quantified. In particular, while many descriptors exhibit statistical significance in large datasets, it remains unclear whether their effect sizes are biologically substantial or merely reflect subtle distributional shifts amplified by sample size. Recent machine learning and protein language model (PLM) architectures have achieved state of the art predictive performance in solubility benchmarks [9–12]. Nevertheless, these high capacity models often obscure the marginal contribution of individual physicochemical dimensions and may implicitly encode correlated features without explicit redundancy assessment. As predictive accuracy increases, mechanistic interpretability and quantitative effect characterization become correspondingly more important.

This contrast exposes a critical gap: the intrinsic dimensionality and practical magnitude of classical sequence-derived solubility signals have not been systematically evaluated at large scale under rigorous statistical control. In high-powered datasets, extremely small p-values can arise from negligible shifts [17], necessitating effect size measures such as Cliff's δ to quantify stochastic dominance [15]. Furthermore, the mathematical equivalence between rank-based statistics and ROC AUC [19] enables a direct link between distributional separation and univariate discriminative capacity.

In the present study, we conduct a comprehensive statistical dissection* of 36 sequence-derived biochemical descriptors in a curated benchmark of soluble and insoluble proteins. Rather than prioritizing high-capacity prediction, our objective is to quantify effect magnitude, redundancy structure, and practical discrimination under strict multiple-testing control. By integrating non-parametric testing, effect size estimation, correlation-based redundancy analysis, and ROC-based evaluation, we aim to clarify whether solubility is driven by dominant single determinants or by coordinated, low-dimensional weak signals. This framework provides a biologically grounded and statistically transparent foundation for parsimonious composite modeling and for interpreting the limits of sequence-based solubility prediction.

Materials and methods

Dataset

The dataset was obtained from the curated solubility benchmark introduced by Zhang *et al.* (2024) [12], comprising more than 78,000 protein sequences annotated as soluble or insoluble. The original release provides independent training, validation, and test splits constructed for predictive benchmarking.

For the present analysis, all splits were merged into a single dataset because the objective was large-scale distributional characterization rather than predictive

generalization. All sequences and labels are publicly available via Zenodo [12].

Feature Extraction

Each protein sequence was normalized by retaining only the 20 canonical amino acids.

A total of 36 sequence-derived biochemical features were computed for each protein, including:

- 20 amino acid frequency variables,
- functional residue group ratios (e.g., positively charged, negatively charged, polar, hydrophobic, small, sulfur-containing),
- global physicochemical descriptors (molecular weight, isoelectric point, net charge at physiological pH, mean hydropathy),
- secondary structure propensity proxies based on Chou Fasman parameters [7],
- an intrinsic disorder proxy [21, 22],
- an aggregation-related proxy defined as the longest contiguous hydrophobic segment.

These descriptors form a structured feature matrix for subsequent statistical evaluation.

Analytical workflow

The analytical workflow was designed to quantify practical effect magnitude, control false discoveries, eliminate redundancy, and subsequently construct an interpretable composite index.

For each of the 36 sequence-derived descriptors, distributional differences between soluble and insoluble proteins were first evaluated using the Mann Whitney U test [13]. Resulting p-values were adjusted via the Benjamini Hochberg procedure to control the false discovery rate (FDR) [16]. Only features satisfying $q < 0.05$ were retained for subsequent effect size analysis, thereby ensuring that magnitude estimation was restricted to statistically reliable signals.

For FDR-significant features, effect magnitude was quantified using Cliff's δ [15], which measures stochastic dominance without distributional assumptions. Median shifts were expressed using the Hodges Lehmann estimator with 95% confidence intervals [14]. Stability of δ was assessed via percentile bootstrap resampling ($B = 2{,}000$) [18]. Univariate discriminative capacity was evaluated using ROC AUC [19], with optimal thresholds determined by Youden's J statistic [20].

Features were subsequently ranked according to absolute δ to identify descriptors exhibiting the strongest practical separation. Because high-effect features may capture overlapping physicochemical dimensions, pairwise redundancy was assessed prior to composite modeling using Spearman's rank correlation coefficient [24, 28]. Spearman correlation was selected due to its robustness to non-normality and its ability to capture monotonic relationships.

Feature pairs exhibiting strong monotonic association ($|\rho| \geq 0.85$) were considered redundant. Within correlated clusters, only a single representative descriptor was retained to avoid double-counting latent physicochemical axes. This redundancy-aware filtering ensured that subsequent integration combined orthogonal dimensions rather than correlated proxies of the same underlying property.

The final composite-δ index was constructed as a linear combination of retained features, robustly scaled using median and interquartile range (IQR) [25, 26] and weighted by their corresponding δ values. Proteins were classified according to the sign of the composite score. Because no parameter fitting or hyperparameter optimization was performed, the resulting index reflects intrinsic distributional separability under a statistically controlled and redundancy-filtered feature set.

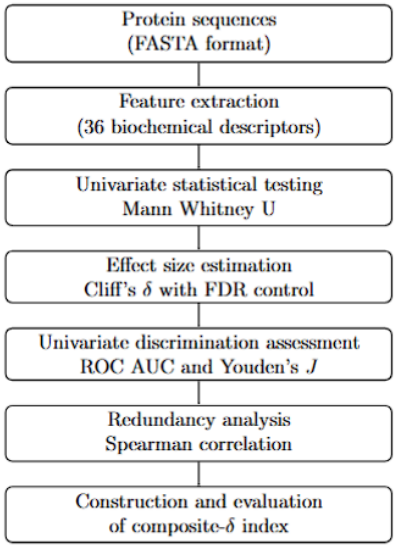

Fig 1. Analytical workflow.

Results

Dataset overview and multiple-testing-adjusted significance

Following preprocessing and feature extraction, the final analysis dataset comprised $N = 78{,}031$ protein sequences, including 46,450 soluble and 31,581 insoluble entries. For each sequence, 36 sequence-derived biochemical descriptors were computed, spanning composition, charge-related variables, hydrophobicity measures, structural propensities, and aggregation proxies.

Group-wise distributional differences were evaluated using the Mann Whitney U test [13], with Benjamini Hochberg correction to control the false discovery rate (FDR) [16]. After adjustment, 34 of 36 features remained statistically significant ($q < 0.05$), indicating that most classical sequence-level descriptors exhibit detectable distributional shifts between soluble and insoluble proteins.

However, given the large sample size, statistical significance alone is insufficient to infer biological relevance. In high-powered datasets, even negligible shifts can produce extremely small p-values [17]. We therefore focus on two complementary quantities: (i) effect magnitude quantified by Cliff's δ [15], which measures stochastic dominance independent of distributional assumptions, and (ii) univariate discriminative capacity assessed via ROC AUC [19], which directly quantifies separability between classes.

To express group differences on the original measurement scale, we report the Hodges Lehmann (HL) location shift with 95% confidence intervals [14]. Uncertainty of δ estimates was evaluated using percentile bootstrap intervals ($B = 2{,}000$) [18]. For completeness, Youden's J statistic and its corresponding optimal threshold T are also reported [20]; in the presence of substantial distributional overlap, J remains close to zero even when statistical significance is achieved.

Collectively, this framework distinguishes statistical detectability from practical magnitude, enabling a biologically grounded interpretation of whether sequence derived physicochemical features exert strong or merely subtle influence on protein solubility.

Global physicochemical features

Table 1. Results for 16 global sequence-derived features.

Feature	Soluble (median [Q1,Q3])	Insoluble (median [Q1,Q3])	HL [95% CI]	δ [95% CI]	FDR q	AUC	J	T^*
length	209 [129, 338]	277 [172, 426]	-70 [-77, -62]	-0.215 [-0.237, -0.196]	0.000	0.392	0.158	227
molecular_weight	19 743 [12 201, 31 715]	25 966 [16 148, 39 869]	-6322.485 [-7147.547, -5784.646]	-0.214 [-0.244, -0.204]	0.000	0.393	0.157	21520
aggregation_ratio	0.02262 [0.01502, 0.03371]	0.01869 [0.01252, 0.02857]	0.004 [0.003, 0.004]	0.166 [0.138, 0.178]	0.000	0.583	0.125	0.020
neg_ratio	0.1298 [0.1104, 0.1510]	0.1219 [0.1034, 0.1414]	0.008 [0.007, 0.009]	0.150 [0.134, 0.173]	0.000	0.575	0.114	0.126
isoelectric_point	6.401 [4.971, 8.357]	6.931 [5.384, 8.941]	-0.588 [-0.702, -0.468]	-0.132 [-0.167, -0.125]	0.000	0.434	0.102	6.523
net_charge_pH7	-1.882 [-7.066, 3.770]	-0.207 [-5.000, 6.476]	-1.618 [-1.931, -1.217]	-0.120 [-0.143, -0.099]	0.000	0.440	0.107	2.413
mean_hydropathy	-0.3511 [-0.4695, -0.2277]	-0.3085 [-0.4388, -0.1834]	-0.046 [-0.060, -0.035]	-0.080 [-0.106, -0.064]	0.000	0.460	0.067	-0.305
sulfur_ratio	0.03448 [0.01961, 0.04712]	0.03681 [0.02193, 0.05013]	-0.002 [-0.003, -0.002]	-0.076 [-0.096, -0.056]	0.000	0.462	0.066	0.029
helix_prop_mean	1.029 [1.019, 1.038]	1.025 [1.015, 1.034]	0.004 [0.003, 0.005]	0.060 [0.049, 0.089]	0.000	0.530	0.054	1.041
hydrophobic_ratio	0.4019 [0.3786, 0.4253]	0.4063 [0.3829, 0.4286]	-0.003 [-0.005, -0.001]	-0.051 [-0.053, -0.012]	0.000	0.475	0.046	0.419
sheet_prop_mean	0.9937 [0.9857, 1.0024]	0.9963 [0.9883, 1.0048]	-0.002 [-0.003, -0.000]	-0.050 [-0.056, -0.015]	0.000	0.475	0.046	1.008
disorder_ratio	0.3976 [0.3690, 0.4310]	0.3937 [0.3662, 0.4274]	0.003 [0.001, 0.005]	0.048 [0.021, 0.063]	0.000	0.524	0.039	0.412
polar_ratio	0.2006 [0.1765, 0.2247]	0.2031 [0.1798, 0.2278]	-0.002 [-0.005, -0.001]	-0.024 [-0.055, -0.013]	0.000	0.488	0.023	0.195
tiny_ratio	0.1389 [0.1111, 0.1667]	0.1411 [0.1127, 0.1695]	-0.000 [-0.003, 0.002]	-0.021 [-0.035, 0.009]	0.000	0.489	0.037	0.171
pos_ratio	0.1412 [0.1212, 0.1636]	0.1419 [0.1230, 0.1636]	-0.000 [-0.002, 0.001]	-0.015 [-0.026, 0.015]	0.000	0.492	0.022	0.116
turn_prop_mean	0.9667 [0.9584, 0.9759]	0.9672 [0.9589, 0.9763]	-0.001 [-0.003, 0.000]	-0.011 [-0.036, 0.004]	0.011	0.495	0.021	0.928

Table 1 summarizes 16 global (non-`freq_*`) descriptors. The largest absolute effects were observed for `length` and `molecular_weight`, both exhibiting $\delta < 0$, indicating that insoluble proteins are, on average, longer and heavier. On the original measurement scale, the Hodges Lehmann estimates correspond to a median shift of -70 amino acids (95% CI $[-77, -62]$) for `length` and $-6{,}322$ Da (95% CI $[-7{,}148, -5{,}785]$) for `molecular_weight`.

From a mechanistic standpoint, increased chain length elevates folding complexity, prolongs exposure of partially folded intermediates, and increases the probability of intermolecular encounters that nucleate aggregation [1, 2]. Larger proteins also present greater solvent-accessible surface area and a higher combinatorial risk of hydrophobic clustering. Despite these biologically plausible trends, univariate discrimination remained limited (AUC $= 0.392$ and 0.393, respectively). Values below 0.5 reflect reversed directionality rather than meaningful separability; indeed, inverting the decision rule yields AUC ≈ 0.607, yet substantial distributional overlap persists, as confirmed by small Youden's J statistics.

Charge-related descriptors exhibited coherent and biologically interpretable shifts. The proportion of negatively charged residues (`neg_ratio`) was enriched in soluble proteins ($\delta = 0.150$, 95% CI $[0.134, 0.173]$, AUC $= 0.575$; HL $= 0.00810$ with 95% CI $[0.00674, 0.00937]$). Conversely, `isoelectric_point` and `net_charge_pH7` showed $\delta < 0$, indicating higher pI and reduced net negativity among insoluble proteins. These findings align with electrostatic stabilization theory: increased net negative charge enhances intermolecular repulsion, reduces colloidal attraction, and suppresses aggregation propensity [2, 3, 5]. The moderate AUC values (≈ 0.57) indicate measurable but limited standalone discriminatory power.

Hydrophobicity-related descriptors exhibited smaller effects. Both `mean_hydropathy` and `hydrophobic_ratio` showed $\delta < 0$, consistent with the role of exposed hydrophobic patches in promoting solvent exclusion and intermolecular association [4, 8]. However, effect sizes were modest and AUC values remained close to 0.5, underscoring extensive overlap between soluble and insoluble distributions. These observations suggest that global hydrophobic averages are insufficient to capture context-dependent aggregation mechanisms.

Disorder and secondary structure proxies

Intrinsic disorder tendency, approximated by `disorder_ratio`, exhibited a small positive effect ($\delta = 0.048$, AUC $= 0.524$), suggesting slight enrichment of disorder-associated residues among soluble proteins. Disordered or flexible regions may increase solvent accessibility and modulate intermolecular interaction landscapes, although their impact appears context-dependent and modest in magnitude [21, 22].

Secondary structure propensity proxies (`helix_prop_mean`, `sheet_prop_mean`, and `turn_prop_mean`) demonstrated very small effect sizes. This indicates that simple global aggregation of Chou Fasman propensities across full-length sequences does not yield strong univariate separation in a large and heterogeneous dataset [7].

Amino acid composition

Table 2. Results for amino acid frequency features.

Feature	Soluble (median [Q1,Q3])	Insoluble (median [Q1,Q3])	HL [95% CI]	δ [95% CI]	FDR q	AUC	J	T^*
freq_C	0.01017 [0.00312, 0.01882]	0.01282 [0.00615, 0.02143]	-0.003 [-0.003, -0.002]	-0.131 [-0.151, -0.111]	0.000	0.434	0.100	0.000
freq_E	0.07064 [0.05488, 0.08889]	0.06529 [0.05072, 0.08187]	0.006 [0.005, 0.007]	0.121 [0.119, 0.160]	0.000	0.560	0.092	0.071
freq_R	0.05350 [0.03663, 0.07273]	0.05894 [0.04061, 0.07944]	-0.005 [-0.006, -0.005]	-0.105 [-0.101, -0.061]	0.000	0.447	0.083	0.060
freq_K	0.06019 [0.04255, 0.07843]	0.05350 [0.03636, 0.07339]	0.007 [0.006, 0.008]	0.091 [0.067, 0.109]	0.000	0.546	0.080	0.049
freq_D	0.05721 [0.04098, 0.07391]	0.05473 [0.04000, 0.07074]	0.003 [0.002, 0.003]	0.080 [0.055, 0.097]	0.000	0.540	0.067	0.061
freq_S	0.06349 [0.04724, 0.08065]	0.06630 [0.05000, 0.08333]	-0.003 [-0.003, -0.002]	-0.068 [-0.092, -0.044]	0.000	0.466	0.043	0.067
freq_P	0.04244 [0.02899, 0.05755]	0.04494 [0.03125, 0.06061]	-0.003 [-0.003, -0.002]	-0.064 [-0.088, -0.040]	0.000	0.468	0.050	0.048
freq_A	0.07273 [0.05556, 0.09091]	0.07650 [0.05825, 0.09524]	-0.004 [-0.005, -0.003]	-0.051 [-0.075, -0.026]	0.000	0.474	0.052	0.092
freq_Q	0.03768 [0.02500, 0.05172]	0.03643 [0.02424, 0.05000]	0.001 [0.001, 0.002]	0.042 [0.017, 0.066]	0.000	0.521	0.042	0.030
freq_Y	0.03061 [0.01923, 0.04211]	0.02913 [0.01818, 0.04110]	0.001 [0.001, 0.002]	0.041 [0.016, 0.066]	0.000	0.521	0.041	0.009
freq_W	0.01000 [0.00000, 0.01754]	0.01031 [0.00000, 0.01852]	-0.000 [-0.001, 0.000]	-0.039 [-0.063, -0.014]	0.000	0.481	0.047	0.000
freq_I	0.05128 [0.03604, 0.06818]	0.05333 [0.03788, 0.07059]	-0.002 [-0.003, -0.001]	-0.039 [-0.063, -0.014]	0.000	0.481	0.047	0.045
freq_V	0.06140 [0.04545, 0.07813]	0.06349 [0.04688, 0.08000]	-0.002 [-0.003, -0.001]	-0.035 [-0.060, -0.010]	0.000	0.482	0.043	0.077
freq_N	0.03922 [0.02632, 0.05263]	0.04000 [0.02740, 0.05405]	-0.001 [-0.002, -0.000]	-0.033 [-0.058, -0.008]	0.000	0.484	0.040	0.056
freq_G	0.06195 [0.04598, 0.07937]	0.06383 [0.04688, 0.08108]	-0.002 [-0.003, -0.001]	-0.028 [-0.053, -0.003]	0.000	0.487	0.039	0.031
freq_L	0.09091 [0.07317, 0.10976]	0.09195 [0.07407, 0.11111]	-0.001 [-0.002, -0.000]	-0.026 [-0.051, -0.001]	0.000	0.487	0.038	0.079
freq_F	0.03636 [0.02532, 0.04878]	0.03704 [0.02632, 0.05000]	-0.001 [-0.001, -0.000]	-0.018 [-0.043, 0.007]	0.000	0.491	0.029	0.018
freq_H	0.02381 [0.01389, 0.03448]	0.02439 [0.01471, 0.03571]	-0.001 [-0.001, 0.000]	-0.016 [-0.041, 0.009]	0.000	0.492	0.029	0.012
freq_M	0.02174 [0.01220, 0.03125]	0.02174 [0.01282, 0.03125]	0 [-0.001, 0.001]	-0.003 [-0.028, 0.023]	0.514	0.499	0.023	0.011
freq_T	0.05263 [0.03704, 0.06897]	0.05263 [0.03788, 0.06897]	0 [-0.001, 0.001]	0.000 [-0.025, 0.025]	0.989	0.500	0.022	0.072

Table 2 reports results for the 20 amino acid frequency variables. Many compositional descriptors achieved extremely small q-values, reflecting high statistical power; however, most displayed small $|\delta|$ and AUC values only marginally deviating from 0.5. Notably, `freq_E` and `freq_D` were elevated in soluble proteins, consistent with the `neg_ratio` signal, whereas `freq_R` and `freq_C` were relatively enriched in insoluble proteins. These trends are compatible with electrostatic contributions and residue-specific side chain chemistry influencing folding stability and aggregation kinetics [2–4]. Only two features, `freq_M` and `freq_T`, did not remain significant after FDR correction ($q \geq 0.05$), with $\delta \approx 0$ and AUC ≈ 0.5, indicating negligible univariate discriminatory information within this dataset.

Emergent weak-signal regime and biological interpretation

Taken together, these findings indicate that protein solubility is not governed by a dominant sequence-level determinant but instead reflects coordinated contributions of multiple weak physicochemical signals [5, 6]. Size-related variables capture a structural burden axis, while charge-related descriptors reflect electrostatic stabilization; hydrophobic and compositional features contribute smaller contextual effects. The generally small Youden's J values across individual descriptors confirm that no single feature provides practical threshold-based separation. Rather, the data support a low-signal, potentially low-dimensional structure in which overlapping physicochemical

axes jointly influence solubility. This observation motivates subsequent redundancy-aware integration of selected descriptors into a parsimonious composite index, as detailed below.

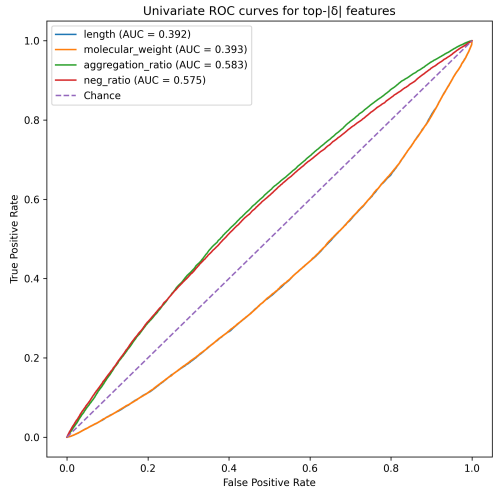

Fig 2. Univariate ROC curves for the four features with the largest absolute Cliff's δ.

Redundancy structure and composite-δ refinement

Spearman redundancy analysis

Prior to finalizing the composite formulation, pairwise redundancy among selected high-effect features was evaluated using Spearman's rank correlation coefficient [24]. The resulting correlation matrix is shown in Figure 3.

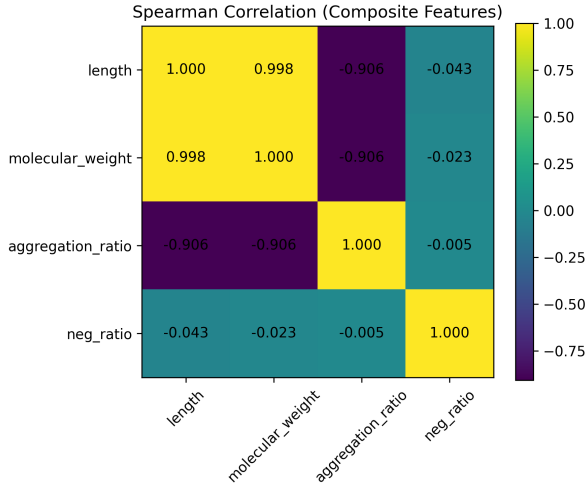

Fig 3. Spearman correlation matrix among top-ranked global features.

Strong monotonic associations were observed among size-related descriptors. In particular, sequence length and molecular weight exhibited near-complete collinearity ($\rho \approx 0.998$), reflecting their deterministic structural coupling. Aggregation-related metrics also displayed strong monotonic association with size-related variables

($|\rho| \approx 0.9$), indicating that these descriptors largely capture a shared latent structural axis.

In contrast, the proportion of negatively charged residues (`neg_ratio`) showed minimal correlation with size-related variables ($|\rho| < 0.05$), suggesting relative independence of electrostatic and size dimensions. Applying a predefined redundancy criterion ($|\rho| \geq 0.85$), correlated size-related descriptors were considered redundant. To avoid double-counting a single latent physicochemical axis, only one representative size descriptor was retained. We selected `length` due to its direct structural interpretability and deterministic relation to molecular mass.

The initial composite-δ index integrated the top-ranked descriptors by absolute effect size, robustly scaled using median and interquartile range (IQR):

$$x_j^{(\mathrm{rs})} = \frac{x_j - \mathrm{median}(x_j)}{\mathrm{IQR}(x_j)}, \qquad \mathrm{IQR}(x_j) = Q_{0.75}(x_j) - Q_{0.25}(x_j) \tag{1}$$

$$S = \sum_{j=1}^{k} \delta_j \, x_j^{(\mathrm{rs})} \tag{2}$$

$$\hat{y} = \begin{cases} 1, & \text{if } S > 0 \quad \text{(soluble)} \\ 0, & \text{if } S \leq 0 \quad \text{(insoluble)} \end{cases} \tag{3}$$

Reduced composite formulation

Table 3. Robust summaries of the two orthogonal physicochemical dimensions defining the reduced composite index

Feature	**Median [IQR]**
length	236 [231]
neg_ratio	0.1265 [0.0399]

Using global medians and IQR values (Table 3), the reduced composite-δ score becomes:

$$S = -0.215 \, \frac{L - 236}{231} + 0.150 \, \frac{N_- - 0.126531}{0.039934} \tag{4}$$

This two-dimensional formulation preserves the dominant size and electrostatic axes while eliminating redundant contributions. As shown below, discriminative performance remains comparable to the full multi-feature composite, supporting the interpretation that sequence-level solubility information is intrinsically low-dimensional.

Table 4. Performance comparison across existing solubility prediction methods and the proposed composite-δ baseline.

Model	T	**AUC**	**Acc.**	**F1**	**MCC**	**Prec.**	**Sens.**	**Spec.**
Protein_sol	0.5	0.5985	0.5649	0.6350	0.1405	0.5470	0.7569	0.3729
SKADE	0.5	0.6882	0.6443	0.5304	0.3305	0.7811	0.4015	0.8874
SWI	0.5	0.5597	0.5423	0.6334	0.0971	0.5284	0.7905	0.2938
Solupro	0.5	0.7126	0.6308	0.6639	0.2667	0.6095	0.7290	0.5325
EPSOL	0.5	0.6664	0.6663	0.5915	0.3578	0.7630	0.4830	0.8499
NetSolP	0.5	0.6183	0.5502	0.6557	0.1268	0.5312	0.8564	0.2437
DeepSol-E	0.4	0.6660	0.5990	0.6407	0.2036	0.5804	0.7150	0.4830
PLM_Sol	0.5	0.8342	0.7299	0.7542	0.4690	0.6919	0.8289	0.6308
Composite-δ	0	**0.6240**	**0.5734**	**0.5875**	**0.1746**	**0.6922**	**0.5102**	**0.6663**

Table 5. Computational complexity comparison across solubility prediction models

Model	Architecture Type	Inference Complexity	Training Requirement
Protein_sol	Feature-based ML	$O(d)$-$O(t)$	Supervised training
SKADE	Ensemble model	$O(t)$	Supervised training
SWI	Linear sequence index	$O(L)$	None
Solupro	Feature-based ML	$O(d)$-$O(t)$	Supervised training
EPSOL	Ensemble / ML	$O(t)$	Supervised training
NetSolP	Deep neural network	$O(L \cdot k)$	Supervised training
DeepSol-E	CNN-based model	$O(L \cdot k)$	Supervised training
PLM_Sol	Transformer (PLM-based)	$O(H \cdot L^2 \cdot h)$	Pretrained + fine-tuned
Composite-δ	Interpretable linear score	$\mathbf{O(1)}$	None

The redundancy-aware composite-δ baseline achieved AUC $= 0.6240$ and MCC $= 0.1746$. Although this performance remains below that of high-capacity protein language model architectures such as PLM_Sol [12], it is comparable to, and in several cases exceeds, traditional physicochemical feature-based predictors reported in the literature. Importantly, the composite-δ formulation involves no parameter fitting, no embedding extraction, and no hyperparameter optimization. The decision rule is fully determined by robust scaling [25, 26] and statistically estimated effect sizes derived directly from empirical distributions. Consequently, the observed discrimination reflects intrinsic separability of sequence-level biochemical features rather than model-driven representation learning.

From a computational standpoint, the proposed model operates in constant time per sequence ($O(1)$), requiring only a small number of arithmetic operations. In contrast, classical machine learning predictors based on handcrafted descriptors scale at least linearly with the number of features or ensemble components ($O(d)$–$O(t)$) [29]. Convolutional neural networks introduce sequence-length-dependent cost ($O(L \cdot k)$) [30], while transformer-based protein language models incur quadratic complexity with respect to sequence length due to the self-attention mechanism ($O(H \cdot L^2 \cdot h)$) [31]. This progression represents multiple orders of magnitude increase in computational demand.

The moderate performance gap between composite-δ and PLM-based architectures therefore reflects a clear trade-off between representational capacity and computational efficiency. While transformer models capture higher-order contextual interactions at substantial resource cost, the present results demonstrate that a low-dimensional, redundancy-controlled linear formulation retains a non-trivial portion of discriminative signal with negligible computational overhead.

From a mechanistic perspective, the achieved performance indicates that global physicochemical descriptors encode a measurable but limited solubility signal. The modest AUC and MCC values are consistent with the substantial distributional overlap observed in univariate analyses and reinforce the interpretation of solubility as a weak-signal, low-dimensional, and context-dependent phenotype [5, 6]. Rather than suggesting inadequacy of classical descriptors, these findings clarify their role: global sequence-derived features provide a statistically grounded lower bound on achievable discrimination and define the interpretable physicochemical axes upon which more complex models may implicitly build. In this sense, the composite-δ baseline serves as both a transparent reference model and a mechanistic anchor for evaluating the added value of high-capacity predictive frameworks under explicit computational constraints.

Conclusion

We conducted a large-scale, statistically rigorous, and fully interpretable dissection of sequence-derived biochemical determinants of protein solubility using a curated benchmark of $N = 78{,}031$ labeled proteins. Across 36 descriptors, distributional shifts were quantified using non-parametric testing with FDR control, complemented by dominance-based effect size estimation, uncertainty quantification, and ROC-based discrimination analysis.

Although 34 features remained statistically significant after multiple-testing correction, most exhibited small effect magnitudes and extensive distributional overlap between soluble and insoluble proteins. No single descriptor demonstrated strong standalone separability. Nonetheless, the directionality of key effects—greater chain length and molecular weight in insoluble proteins, enrichment of negatively charged residues in soluble proteins, higher isoelectric point in insoluble proteins, and modest hydrophobicity-associated shifts—aligns closely with established biophysical principles and prior experimental observations. Our effect-size-centered framework clarifies that these influences, while coherent and mechanistically plausible, operate within a weak-signal regime whose magnitude is substantially more moderate than often implied by mean-based or p-value-driven interpretations.

Redundancy analysis further revealed strong collinearity among size-related descriptors, indicating that multiple high-ranking features largely reflect a shared latent structural axis. After applying a correlation-based filtering criterion, a parsimonious two-dimensional composite integrating sequence length and negative charge proportion preserved comparable discrimination (AUC = 0.624), supporting the view that solubility-associated sequence information is intrinsically low-dimensional.

Collectively, these findings reinforce the interpretation of protein solubility as an emergent, multifactorial phenotype governed by coordinated weak physicochemical signals rather than dominant single determinants. The primary contribution of this work lies not in proposing novel predictors, but in establishing a statistically principled, effect-size-centered reference characterization of a modern large-scale solubility benchmark. By combining dominance statistics, uncertainty intervals, redundancy control, and interpretable composite modeling within a reproducible and computationally efficient framework, this study refines the quantitative understanding of classical sequence-level descriptors and provides a transparent baseline that complements—rather than competes with—high-capacity machine learning approaches.

Supporting information

Supplementary Tables

- **Table S1.** Comprehensive statistical results for the 16 global (non compositional) descriptors, including group wise medians and interquartile ranges, Hodges Lehmann location shifts with 95% confidence intervals, raw p-values, FDR adjusted q-values, Cliff's δ with 95% bootstrap confidence intervals, ROC AUC, Youden's J, and optimal thresholds T^*.

- **Table S2.** Complete statistical results for the 20 amino acid frequency features, reported with the same metrics as in Table S1 to enable direct comparison across descriptor classes.

- **Table S3.** These robust summaries define the centering and scaling constants for the redundancy-aware composite, capturing a size axis (length) and an electrostatic axis (negative charge proportion).

- **Table S4.** Full comparative performance metrics for all evaluated predictors, including the redundancy-aware composite-δ baseline and reference models. Reported metrics comprise ROC AUC, accuracy, F1-score, Matthews correlation coefficient (MCC), precision, sensitivity, specificity, and decision thresholds.

- **Table S5.** Presents a qualitative comparison of computational complexity across solubility prediction models, following the same ordering as the performance comparison. The models span a broad spectrum of architectural paradigms, ranging from simple interpretable linear scoring to deep neural networks and transformer-based protein language models. Inference complexity increases substantially from feature-based or linear approaches ($O(1)$ to $O(d)$) to convolutional architectures ($O(L \cdot k)$) and ultimately to transformer-based models with quadratic sequence length scaling ($O(H \cdot L^2 \cdot h)$). Notably, the proposed composite-δ baseline operates in constant time and requires no parameter training, highlighting a substantial trade-off between computational cost and predictive performance across model families.

Supplementary Figures

- **Figure S1.** Detailed workflow diagram of the analytical pipeline, from dataset acquisition and preprocessing to non-parametric statistical testing, effect size estimation, redundancy assessment, and composite-δ construction.

- **Figure S2.** Comparative performance visualization of existing solubility predictors and the redundancy-aware composite-δ baseline, including ROC AUC and MCC.

- **Figure S3.** Spearman correlation matrix among top-ranked global physicochemical descriptors, highlighting strong monotonic associations among size-related variables and the relative independence of electrostatic features. Correlation coefficients are displayed within each cell; feature pairs with $|\rho| \geq 0.85$ were considered redundant in composite construction.

Data and Code Availability

The protein solubility dataset analyzed in this study was obtained from the benchmark dataset published by Zhang *et al.* [12] and is publicly accessible via Zenodo at `https://zenodo.org/records/12881509`.

All scripts used for sequence preprocessing, feature extraction, statistical testing, composite score construction, and figure generation were developed by the authors. The complete analysis pipeline, together with processed feature matrices, statistical summaries, composite score outputs, and manuscript figures, is publicly available at:

`https://github.com/huyhoang6723/protein-solubility-effectsize.git`

Within the repository, the file `data.zip` contains the dataset splits used in this study, while `results.zip` includes all derived features and analytical outputs reported in the manuscript. Raw sequence datasets are referenced to their original public source in order to respect licensing constraints. This structure ensures full computational reproducibility of all reported results. Long-term archival deposition with DOI assignment (e.g., via Zenodo) will be finalized upon publication.

Repository Structure

```
protein-solubility-effectsize/
├── README.md
├── LICENSE
└── .gitignore
```

```
│
├── code.ipynb
├── data.zip
│   ├── Train_dataset.fasta
│   ├── validation_dataset.fasta
│   └── test_dataset.fasta
└── results.zip
    ├── stats_features.csv
    ├── summary_global.csv
    ├── summary_by_class.csv
    ├── composite_delta_metrics.csv
    ├── composite_delta_predictions.csv
    ├── features_per_protein.csv
    ├── corr_4features_spearman.png
    └── roc_top4.png
```


Acknowledgments

The authors thank the contributors of the publicly available protein solubility benchmark dataset used in this study and acknowledge the open-source software community for providing essential computational tools that enabled large-scale statistical analysis. We are grateful to colleagues and reviewers for constructive discussions and critical feedback that improved the clarity and rigor of this work.

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors. The authors declare no competing interests.

References

1. Baneyx F. Recombinant protein expression in *Escherichia coli*. Curr Opin Biotechnol. 1999 Oct;10(5):411-421. doi: 10.1016/S0958-1669(99)00003-8.
2. Wilkinson DL, Harrison RG. Predicting the solubility of recombinant proteins in *Escherichia coli*. Biotechnology (N Y). 1991 May;9(5):443-448. doi: 10.1038/nbt0591-443.
3. Idicula-Thomas S, Balaji PV. Understanding the relationship between the primary structure of proteins and its propensity to be soluble on overexpression in *Escherichia coli*. Protein Sci. 2005 Feb;14(2):582-592. doi: 10.1110/ps.041129005.
4. Ventura S. Sequence determinants of protein aggregation: Tools to increase protein solubility. Microb Cell Fact. 2005;4:11. doi: 10.1186/1475-2859-4-11.
5. Chiti F, Dobson CM. Protein aggregation in disease and biotechnology. Annu Rev Biochem. 2006;75:333-366. doi: 10.1146/annurev.biochem.75.101304.123901.
6. Chiti F, Dobson CM. Protein misfolding, amyloid formation, and human disease. Annu Rev Biochem. 2017;86:27-68. doi: 10.1146/annurev-biochem-061516-045115.
7. Chou PY, Fasman GD. Prediction of protein conformation. Biochemistry. 1974 Jan;13(2):222-245. doi: 10.1021/bi00699a002.
8. Kyte J, Doolittle RF. A simple method for displaying the hydropathic character of a protein. J Mol Biol. 1982 May;157(1):105-132. doi: 10.1016/0022-2836(82)90515-0.

9. Khurana S, et al. DeepSol: A deep learning framework for sequence-based protein solubility prediction. Bioinformatics. 2018 Aug;34(15):2605-2613. doi: 10.1093/bioinformatics/bty166.

10. Hon J, et al. SoluProt: Prediction of soluble protein expression in *Escherichia coli*. Bioinformatics. 2021 Jan;37(1):23-28. doi: 10.1093/bioinformatics/btaa1102.

11. Thumuluri V, et al. NetSolP: Predicting protein solubility in *Escherichia coli* using language models. Bioinformatics. 2022 Feb;38(4):941-946. doi: 10.1093/bioinformatics/btab782.

12. Zhang X, et al. PLM_Sol: Predicting protein solubility by benchmarking multiple protein language models with the updated *Escherichia coli* protein solubility dataset. Brief Bioinform. 2024;25(5):bbae404. doi: 10.1093/bib/bbae404.

13. Mann HB, Whitney DR. On a test of whether one of two random variables is stochastically larger than the other. Ann Math Stat. 1947 Mar;18(1):50-60. doi: 10.1214/aoms/1177730491.

14. Hodges JL, Lehmann EL. Estimates of location based on rank tests. Ann Math Stat. 1963 Jun;34(2):598-611. doi: 10.1214/aoms/1177704172.

15. Cliff N. Dominance statistics: Ordinal analyses to answer ordinal questions. Psychol Bull. 1993 Nov;114(3):494-509. doi: 10.1037/0033-2909.114.3.494.

16. Benjamini Y, Hochberg Y. Controlling the false discovery rate: A practical and powerful approach to multiple testing. J R Stat Soc B. 1995;57(1):289-300. doi: 10.1111/j.2517-6161.1995.tb02031.x.

17. Wasserstein RL, Schirm AL, Lazar NA. Moving to a world beyond "p¡0.05". Am Stat. 2019;73(Suppl 1):1-19. doi: 10.1080/00031305.2019.1583913.

18. Efron B, Tibshirani RJ. *An Introduction to the Bootstrap*. Chapman & Hall; 1993.

19. Hanley JA, McNeil BJ. The meaning and use of the area under a receiver operating characteristic (ROC) curve. Radiology. 1982 Apr;143(1):29-36. doi: 10.1148/radiology.143.1.7063747.

20. Youden WJ. Index for rating diagnostic tests. Cancer. 1950 Jan;3(1):32-35. doi: 10.1002/1097-0142(1950)3:1<32::AID-CNCR2820030106>3.0.CO;2-3.

21. Dunker AK, et al. Intrinsically disordered protein. J Mol Graph Model. 2001;19(1):26-59. doi: 10.1016/S1093-3263(00)00138-8.

22. Uversky VN. Intrinsic disorder-based protein interactions and their modulators. Chem Rev. 2013 Sep;113(9):6844-6879. doi: 10.1021/cr300363r.

23. Matthews BW. Comparison of the predicted and observed secondary structure of T4 phage lysozyme. Biochim Biophys Acta. 1975 Oct;405(2):442-451. doi: 10.1016/0005-2795(75)90109-9.

24. Spearman C. The proof and measurement of association between two things. Am J Psychol. 1904;15(1):72-101. doi: 10.2307/1412159.

25. Huber PJ. *Robust Statistics*. New York: John Wiley & Sons; 1981.

26. Hampel FR, Ronchetti EM, Rousseeuw PJ, Stahel WA. *Robust Statistics: The Approach Based on Influence Functions*. New York: John Wiley & Sons; 1986.

27. van Rijsbergen CJ. *Information Retrieval*. London: Butterworth-Heinemann; 1979.

28. Conover WJ. *Practical Nonparametric Statistics*. 3rd ed. New York: Wiley; 1999.

29. Bishop CM. *Pattern Recognition and Machine Learning*. New York: Springer; 2006.

30. LeCun Y, Bottou L, Bengio Y, Haffner P. Gradient-based learning applied to document recognition. Proc IEEE. 1998 Nov;86(11):2278-2324. doi: `10.1109/5.726791`.

31. Vaswani A, et al. Attention Is All You Need. Adv Neural Inf Process Syst. 2017;30:5998-6008. doi: `10.48550/arXiv.1706.03762`.