Evaluating Protein Language Models Against Biochemical Property Features for SARS-CoV-2 Variant Classification

Florian Ranbir Vaid Daefler(3180408) Emilija Milanovic(3191332)
Changchen Yu(3161666)

Bocconi University Deep Learning and Reinforcement Learning

Abstract

Recent advances in protein language models offer a promising alternative to traditional feature engineering for viral classification tasks. However, the comparative effectiveness of these models against both simple biochemical descriptors and more carefully designed biochemical descriptors remains unclear. Here we evaluate three feature extraction approaches for SARS-CoV-2 spike protein variant classification using 261,042 sequences spanning six WHO lineages: (i) a six-dimensional aggregate hydrophobicity baseline, (ii) ProtBERT CLS embeddings differenced against the Wuhan reference, and (iii) biochemical feature vectors encoding site-specific mutations explicitly.

Using hyperparameter-optimized XGBoost classifiers, we find that ProtBERT embeddings substantially outperform the simple baseline (macro-F1: 0.958 vs 0.928, accuracy: 98.6% vs 97.9%), with particularly strong gains on minority lineages. However, biochemical features that explicitly encode mutation locations achieve superior performance to both approaches (macro-F1: 0.967, accuracy: 99.0%), with statistically significant improvements confirmed by paired t-tests and McNemar's test.

Our results suggest that while off-the-shelf protein language models provide an effective "quick go-to" solution that reliably outperforms simple baselines without domain expertise, carefully engineered biochemical feature vectors can achieve higher accuracy when computational resources and domain knowledge permit. This positions protein language models as valuable intermediate solutions for rapid deployment in viral surveillance, while highlighting the continued importance of specialized feature engineering for optimal performance. Future work investigating targeted fine-tuning of protein language models on viral classification tasks may potentially bridge this performance gap, combining the accessibility of pre-trained models with the accuracy of domain-informed approaches.

1 Introduction

Effective public health response to emerging infectious diseases, such as COVID-19, critically depends on timely identification and classification of viral variants. The spike (S) protein of SARS-CoV-2, crucial for virus-host interactions and immune evasion, has become a key focus in genomic surveillance. Understanding and tracking mutations within this protein enables researchers and health authorities to anticipate epidemiological trends, vaccine effectiveness, and the emergence of new Variants of Concern.

Traditional approaches to viral variant classification have relied on handcrafted biochemical descriptors that encode functional differences between strains [Nanni et al., 2020, Solis and Rackovsky, 2010]. These range from simple aggregate statistics derived from physicochemical properties to sophisticated biochemical feature encoding schemes that explicitly capture mutation locations and their functional context [Solis and Rackovsky, 2010, Lin et al., 2011]. While effective, such approaches require domain expertise and manual feature engineering [Lin et al., 2011], limiting their accessibility and rapid deployment potential.

Recent breakthroughs in deep learning have introduced protein language models like ProtBERT that automatically derive dense, high-dimensional embeddings directly from raw amino acid sequences. These models promise to capture subtle biological relationships without explicit feature design, potentially democratizing high-performance viral classification. However, their practical advantages over both simple baselines and more carefully engineered per-site biochemical feature vectors remain unclear.

This study addresses this knowledge gap through a systematic three-way comparison of feature extraction approaches: (1) simple aggregate hydrophobicity statistics between sequences, (2) ProtBERT-derived embeddings, and (3) per-site hydrophobicity feature encoding. Using a comprehensive dataset of 261,042 SARS-CoV-2 spike sequences, we evaluate each approach using hyperparameter-optimized XGBoost classifiers to isolate the effect of feature representation.

Our results reveal a clear performance hierarchy, positioning protein language models as effective intermediate solutions that outperform simple baselines while being outperformed by per-site biochemical feature encoding.

2 Dataset

All spike protein sequences used in this study were obtained from the open-access Kaggle repository *"SARS-CoV-2 Spike Protein Sequences"* maintained by M. Jamhuri (`edumath`). This repository, distributed under the Apache 2.0 licence, mirrors GISAID records and provides variant-specific FASTA files that facilitate systematic analysis of major SARS-CoV-2 lineages.

The complete dataset comprises 261,042 full-length spike protein sequences distributed across six variant categories: Alpha (182,869 sequences), Beta (3,937), Gamma (16,385), Delta (35,561), Omicron (1,847), and Others (20,443). The substantial size imbalance reflects the natural epidemiological prevalence of these variants during the pandemic, with Alpha dominating the early 2021 period when extensive sequencing efforts were underway.

The **Others** category serves as a catch-all class for sequences that could not be assigned to the five major variants of concern. This heterogeneous group includes sequences whose FASTA headers either lacked WHO Variant of Concern tags, carried PANGO lineages not mapping to the named variants (such as early Wuhan/B.1 lineages, Lambda/C.37, or Mu/B.1.621), or were flagged by GISAID as *Unassigned*. Rather than discarding these sequences, we retain them to provide useful background diversity while preventing rare or ambiguous lineages from contaminating the well-defined variant classes used for supervised learning.

The dataset includes a small fraction of records containing IUPAC ambiguity symbols that represent sequencing uncertainties. These symbols include `X` (unknown amino acid), `B` (aspartic acid or asparagine), `Z` (glutamic acid or glutamine), and `J` (isoleucine or leucine). To maintain analytical rigor, we treat these ambiguous symbols as non-informative throughout our analysis. Specifically, they are excluded when computing per-site amino acid frequencies, biochemical entropy, AAindex-based feature vectors, and Hamming distances, thereby preventing sequencing ambiguities from introducing systematic bias into our statistical analyses or downstream classification models.

For reference-based comparisons and evolutionary analysis, we obtained the original Wuhan-Hu-1 spike protein sequence from NCBI (NCBI accession YP_009724390.1) , which serves as the ancestral baseline for measuring mutational divergence across all variant lineages. All sequences were pre-aligned to this reference, ensuring consistent positional indexing across the 1,273-residue spike protein and enabling direct comparison of amino acid substitution patterns between variants.

3 Methods

3.1 Feature Extraction Approaches

We compare three distinct approaches to represent SARS-CoV-2 spike protein sequences, designed to span the spectrum from simple aggregate statistics to sophisticated biochemical feature encoding and modern BERT-Style embeddings

3.1.1 AAindex Aggregate Features (Simple Baseline)

The simplest approach encodes each spike sequence as a **six-dimensional vector** derived from aggregate statistics of Kyte–Doolittle hydrophobicity changes (AAindex entry KYTJ820101). This baseline represents the type of rapid, domain-agnostic feature extraction commonly used in resource-constrained surveillance settings.

Kyte–Doolittle hydrophobicity scale. The Kyte–Doolittle index quantifies the hydrophobicity (water-repelling tendency) of each amino acid on a scale from highly hydrophilic (−4.5 for arginine) to highly hydrophobic (+4.5 for isoleucine) [Kyte and Doolittle, 1982]. This scale, derived from experimental measurements of amino acid transfer energies between water and organic solvents, captures fundamental physicochemical properties that govern protein folding, membrane interactions, and structural stability. For viral surveillance, hydrophobicity changes are particularly relevant because: (i) the SARS-CoV-2 spike protein is a membrane-associated glycoprotein where hydrophobic regions facilitate viral-host membrane interactions; (ii) mutations altering local hydrophobicity can destabilize protein domains or modify binding affinity to the ACE2 receptor; and (iii) hydrophobicity patterns distinguish functional regions such as the transmembrane domain from the receptor-binding domain [Lan et al., 2020]. As a single physicochemical dimension, hydrophobicity provides a biologically meaningful yet computationally efficient baseline that requires little domain expertise to implement.

Let $h(a)$ denote the hydrophobicity assigned to amino-acid $a \in \{\mathrm{A}, \ldots, \mathrm{V}\}$. Given the Wuhan–Hu-1 reference sequence $R_{1:L}$ ($L = 1\,273$) and a variant sequence $S_{1:L}$, we iterate over aligned positions and, for every substitution $R_i \neq S_i$, compute the scalar

$$\Delta h_i = h\big(S_i\big) \;-\; h\big(R_i\big). \tag{1}$$

Sequence-level aggregation. The set $\{\Delta h_i\}$ is summarised into a fixed-width feature vector

$$\boldsymbol{x}_{\mathrm{agg}} = \big[\, \mu,\ \sigma,\ \Sigma,\ \max,\ \min,\ m/L \,\big], \tag{2}$$

where μ is the mean, σ the sample standard deviation, Σ the arithmetic sum, and m the number of mutated sites. The mutation rate m/L captures sequence-level divergence from the reference. This aggregate feature vector serves as a very simple low-dimensional baseline

3.1.2 AAindex Biochemical Feature Vectors (Enhanced Baseline)

To evaluate whether explicit biochemical information improves upon aggregate statistics, we construct a **1273-dimensional vector** that preserves the location of each hydrophobicity change across the spike protein sequence.

Position-specific encoding. For each alignment position $i \in \{1, \ldots, 1273\}$, we compute:

$$\boldsymbol{x}_{\mathrm{biochem}}[i] = \begin{cases} \Delta h_i = h(S_i) - h(R_i) & \text{if } R_i \neq S_i \\ 0 & \text{if } R_i = S_i \end{cases} \tag{3}$$

This representation explicitly encodes both the magnitude and location of hydrophobicity changes, enabling the classifier to learn position-specific mutation patterns that may be diagnostic of particular variants. Unlike aggregate features, this approach preserves the spatial organization of mutations across functional domains of the spike protein.

3.1.3 ProtBERT Embeddings

ProtBERT is a transformer-based protein language model pre-trained on over 200 million protein sequences using masked language modeling objectives. By learning to predict masked amino acids from surrounding context, ProtBERT captures complex sequence patterns, evolutionary relationships, and functional constraints that are difficult to encode through traditional physicochemical descriptors. This makes it particularly suitable for viral classification tasks where subtle sequence variations may carry functional significance.

Model architecture and setup. We employed the public checkpoint `Rostlab/prot_bert`, keeping all weights frozen during classification to evaluate the quality of off-the-shelf representations. Notably, the ProtBERT authors acknowledge that "in some tasks you could gain more accuracy by fine-tuning the model rather than using it as a feature extractor," suggesting potential for performance improvements beyond our frozen encoder evaluation.

Embedding extraction and delta calculation. Sequences were processed in batches of 64 on an NVIDIA RTX 4090. The 1,024-dimensional hidden state of the `[CLS]` token serves as the holistic sequence representation $\mathbf{e} \in \mathbb{R}^{1024}$, capturing global sequence properties learned during pre-training.

To focus on variant-specific features while removing common background signal, we computed delta embeddings by subtracting the Wuhan-Hu-1 reference representation:

$$\Delta\mathbf{e} = \mathbf{e}_{\text{variant}} - \mathbf{e}_{\text{Wuhan}}.$$

This delta calculation mirrors the approach used in AAindex features and ensures that the learned representation emphasizes deviations from the original strain—precisely the signal most relevant for variant classification.

3.2 Classification Framework

To rigorously compare feature extraction approaches while ensuring optimal performance for each method, we implemented a two-stage experimental design combining hyperparameter optimization with robust cross-validation evaluation.

Experimental design rationale. Rather than using identical hyperparameters across all approaches—which could unfairly disadvantage methods with different optimal configurations—we conducted method-specific hyperparameter tuning followed by fair comparison under optimized conditions. This design isolates the effect of feature representation while maximizing each method's potential, providing a more definitive assessment of their relative capabilities.

Hyperparameter optimization. We performed minimal but systematic hyperparameter tuning for each feature extraction approach using macro-F1 score as the selection criterion. Macro-F1 was chosen over accuracy because it provides equal weight to all variant classes, ensuring that optimization focuses on minority lineage performance—critical for surveillance applications where rare variants are often most important.

For computational efficiency, while preserving statistical validity, we used stratified subsampling (maximum 15,000 sequences) with 3-fold cross-validation to evaluate parameter combinations. The search space focused on XGBoost's most impactful parameters: `max_depth` $\in \{4, 5, 6, 8\}$, `n_estimators` $\in \{200, 300, 500\}$, and `learning_rate` $\in \{0.05, 0.1, 0.15\}$. Base parameters were fixed across methods (`subsample = 0.9`, `colsample_bytree = 0.9`, `objective = 'multi:softprob'`) to ensure consistent optimization conditions.

Cross-validation evaluation. Final performance assessment employed 5-fold stratified cross-validation on the complete dataset ($N = 261{,}042$), using the method-specific optimal hyperparameters identified during tuning. Stratification preserves the original class distribution across folds: Alpha (70.1%), Beta (1.5%), Gamma (6.3%), Delta (13.6%), Omicron (0.7%), and Others (7.8%). This approach yields five independent performance measurements per method, enabling rigorous statistical comparison while ensuring that each sequence appears in exactly one test set, eliminating data leakage concerns.

Evaluation metrics. We report both overall accuracy and macro-averaged F1 score, with primary emphasis on macro-F1 due to the substantial class imbalance in our dataset. Macro-F1 treats all variant classes equally regardless of their frequency, making it particularly suitable for viral surveillance where detecting minority variants (e.g., Beta, Omicron) is often more critical than optimizing performance on dominant lineages. Per-class precision, recall, and F1 scores provide additional insight into method-specific strengths and limitations across individual variant types.

4 Exploratory Analysis: Understanding Variant Differences

Before presenting our main classification results, we first establish the theoretical foundation for why variant classification should be feasible. By analyzing the distributional and sequence-level differences between WHO lineages, we can set expectations for classifier performance and understand the underlying biological signal that both traditional and modern feature extraction methods aim to capture.

4.1 Information-theoretic Variant Distances

To quantify how different variants differ in their spike protein sequences, we employ an information-theoretic approach that measures the divergence between amino acid distributions across variants. Rather than examining individual sequence differences, this method captures the overall compositional patterns that characterize each lineage, providing insight into how variants have evolved distinct amino acid usage signatures at each position along the spike protein.

Our analysis begins by constructing position-specific amino acid distributions for each variant. We define the standard 20 canonical amino acids as $\mathcal{A} = \{\text{A,C,D,E,F,G,H,I,K,L,M,N,P,Q,R,S,T,V,W,Y}\}$ and represent the sequences belonging to variant v as $S_v = \{\, s_v^{(i)} \}_{i=1}^{N_v}$, where each sequence has the canonical spike protein length of $L = 1\,273$ amino acids.[1] For each alignment position $j \in \{1, \ldots, L\}$, we count how many sequences within variant v contain each amino acid a:

$$n_{v,j}(a) \;=\; \big|\{\, i \mid s_v^{(i)}[j] = a \,\}\big|, \qquad a \in \mathcal{A}.$$

These counts are then normalized to obtain empirical probabilities representing the likelihood of observing each amino acid at position j within variant v:

$$p_{v,j}(a) \;=\; \frac{n_{v,j}(a)}{\displaystyle\sum_{a' \in \mathcal{A}} n_{v,j}(a')}. \tag{1}$$

This procedure yields a $L \times 20$ probability matrix $P_v = \big[p_{v,j}(a)\big]$ for each variant, where each row represents the amino acid distribution at a specific position and sums to unity.

To measure the divergence between any two variants A and B, we employ the Jensen–Shannon divergence (JSD), a symmetric and bounded information-theoretic distance measure. For each alignment position j, we compute the JSD between the corresponding amino acid distributions:

$$\mathrm{JSD}_j(A,B) \;=\; \frac{1}{2}\, D_{\mathrm{KL}}\big(p_{A,j} \,\|\, M_j\big) \;+\; \frac{1}{2}\, D_{\mathrm{KL}}\big(p_{B,j} \,\|\, M_j\big), \quad M_j = \frac{p_{A,j} + p_{B,j}}{2}, \tag{2}$$

where M_j represents the average distribution and $D_{\mathrm{KL}}(P\|Q) = \sum_{a \in \mathcal{A}} P(a)\, \log_2 \frac{P(a)}{Q(a)}$ is the Kullback–Leibler divergence. The JSD effectively measures how much information is needed to distinguish between the two amino acid distributions at each position, with values bounded between 0 and 1 bit and satisfying the symmetry property essential for a proper distance metric.

Since we require a single scalar distance to characterize the overall divergence between lineages, we average the position-specific JSD values across all L positions:

$$\boxed{D(A,B) \;=\; \frac{1}{L} \sum_{j=1}^{L} \mathrm{JSD}_j(A,B)} \tag{3}$$

This final distance measure provides an intuitive interpretation: $D(A,B) = 0$ bits indicates that variants A and B exhibit identical amino acid usage patterns at every position, while larger values signal increasing compositional divergence. The theoretical maximum of 1 bit would occur if the two variants never shared the same amino acid at any position, representing complete compositional orthogonality.

Computing equation (3) for all pairs of variants produces the symmetric distance matrix visualized in Figure 1. The resulting 6×6 matrix reveals several compelling evolutionary patterns. Most notably, Beta emerges as the most divergent variant, showing the greatest distances from both the early Alpha reference (0.134 bits) and from Omicron (0.146 bits). In contrast, Alpha and Gamma form the closest pair with only 0.035 bits of divergence. The catch-all *Others* is closest to Omicron, and additionally is farthest from Beta. Interestingly Beta is also farthest from Omicron suggesting similarity amongst these two variants.

The hierarchical clustering pattern displayed in the dendrogram confirms that these lineages form well-separated clusters at the amino acid distribution level, providing strong evidence for distinct evolutionary

[1] Sequences containing ambiguous residues (`X`) are retained but those `X` symbols are ignored during counting; indels were removed during the alignment step.

signatures that differentiate the major SARS-CoV-2 variants. This information-theoretic foundation establishes that variants can be distinguished not only by individual mutations but by their overall compositional fingerprints across the entire spike protein sequence.

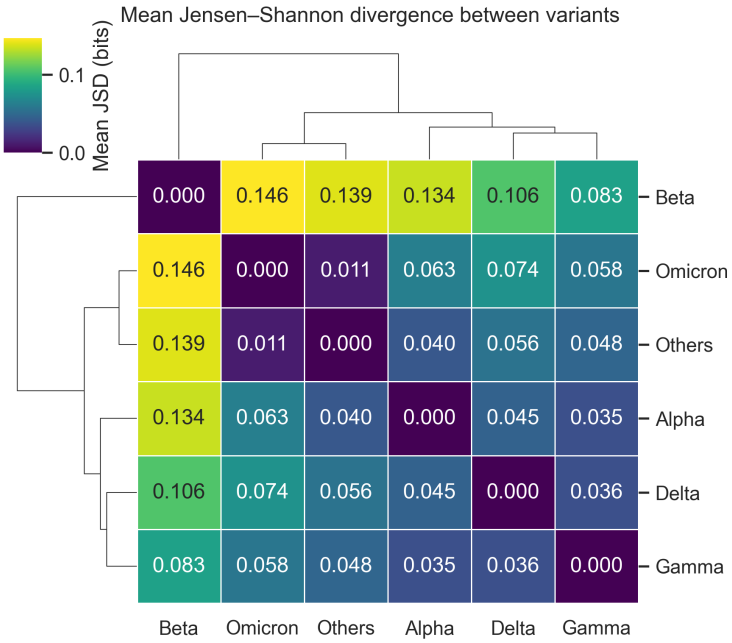

Figure 1: Mean Jensen–Shannon divergence between variants. Darker colours indicate similar amino-acid usage; lighter colours highlight lineages with strongly diverging residue statistics. The dendrogram reflects hierarchical clustering of the symmetric distance matrix.

4.2 Spike-sequence divergence via Hamming distance

While the previous subsection quantified *distributional* differences between variants, we now examine variation at the *sequence* level using the familiar normalised Hamming distance. This approach allows us to assess sequence-level separability both between different lineages (across variants) and within lineages (within variants).

To quantify sequence-level differences, we compute the normalised Hamming distance between pairs of aligned spike sequences. For two sequences $s^{(1)}, s^{(2)} \in \mathcal{A}^L$ where $L = 1\,273$ represents the spike protein length, the Hamming distance is defined as

$$d_{\mathrm{H}}\big(s^{(1)}, s^{(2)}\big) \;=\; \frac{1}{L}\sum_{j=1}^{L} \mathbf{1}\big[s_j^{(1)} \neq s_j^{(2)}\big], \tag{4}$$

which simply measures the proportion of amino acid sites that differ between the two sequences. Since computing distances for all $O(N^2)$ possible sequence pairs would be computationally impractical, we randomly sampled 1,000 sequences per lineage (or all available sequences if fewer than 1,000 existed) and calculated pairwise distances for two distinct sets: **within-variant** pairs $\big\{(s_v^{(i)}, s_v^{(k)})\big\}_{i<k}$ comparing sequences from the same lineage, and **between-variant** pairs $\big\{(s_A^{(i)}, s_B^{(k)})\big\}_{A\neq B}$ comparing sequences from different lineages.

The resulting distance distributions, shown as kernel-smoothed histograms in Figure 2, reveal a clear bimodal pattern that demonstrates strong sequence-level separation between variants. The blue distribution represents within-variant distances and exhibits a tight mode at $d_{\mathrm{H}} \approx 0.02$, indicating that sequences within the same lineage differ at only approximately 2% of amino acid sites on average. In stark contrast, the orange distribution for between-variant pairs shows a much broader spread, peaking around 0.10 and extending

beyond 0.30, reflecting the substantial sequence divergence that has accumulated between different SARS-CoV-2 lineages.

To gain deeper insight into the relationships between specific variant pairs, we aggregated the pairwise distances from equation (4) to construct a symmetric distance matrix $D_{\mathrm{H}}(A, B) = \mathbb{E}[d_{\mathrm{H}}(s_A, s_B)]$, shown in Figure 3. This matrix reveals patterns that closely mirror those observed in the Jensen-Shannon divergence analysis (Section 4.1). The Beta variant emerges as the most divergent, showing the greatest distances from both Alpha and Delta variants (0.121–0.134), while Gamma exhibits the closest relationship to Alpha with a mean distance of only 0.044. Notably, the miscellaneous *Others* class maintains the greatest distance from every named lineage (≥ 0.199), consistent with its role as a catch-all category for diverse, less-common variants.

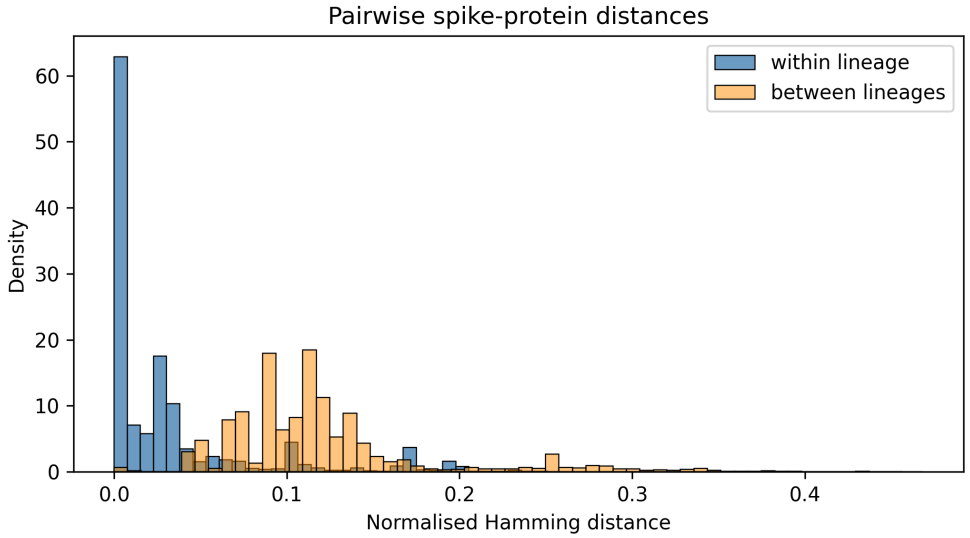

Figure 2: Density of normalised Hamming distances d_{H} for 6,000 sampled sequences. Within-variant pairs (blue) cluster tightly around 0.02, whereas between-variant pairs (orange) occupy a broader range centred near 0.10.

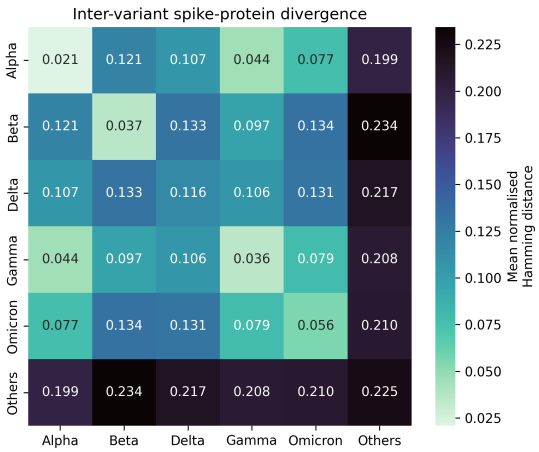

Figure 3: Mean normalised Hamming distance $D_{\mathrm{H}}(A, B)$ between variants. Darker cells denote greater sequence-level divergence.

These exploratory analyses reveal that the six WHO lineages already form well-separated clusters in raw

sequence space, with within-variant Hamming distances centring around 2 percent while between-variant distances consistently exceed 10 percent. The Jensen-Shannon divergence analysis confirms this same hierarchical ordering.

5 Results

5.1 Classification Performance Hierarchy

Our systematic comparison with method-specific hyperparameter optimization reveals a clear performance hierarchy that positions domain-informed biochemical feature vectors as the superior approach for SARS-CoV-2 variant classification.

Table 1: XGBoost performance comparison with method-specific optimized hyperparameters. Results show mean ± standard deviation over 5-fold cross-validation (N = 261,042).

Feature set	**Optimal Parameters**	**Accuracy**	**Macro-F1**
AAindex per-site hydrophobicity (1273D)	*depth=6, trees=200, lr=0.05*	**0.9903 ± 0.0003**	**0.9671 ± 0.0018**
ProtBERT Δ-embeds (1024D)	*depth=6, trees=500, lr=0.15*	0.9868 ± 0.0003	0.9585 ± 0.0017
AAindex hydrophobicity aggregate (6D)	*depth=5, trees=300, lr=0.1*	0.9795 ± 0.0006	0.9277 ± 0.0026
Biochemical vs ProtBERT		**+0.35%**	**+0.86%**
ProtBERT vs Aggregate		**+0.73%**	**+3.08%**

Hyperparameters optimized using macro-F1 criterion with 3-fold CV on stratified subsets

Hyperparameter optimization impact. Method-specific hyperparameter tuning using macro-F1 as the selection criterion reveals that different approaches require distinct optimal configurations. AAindex biochemical feature vectors achieve peak performance with conservative settings (depth=6, 200 trees, learning rate=0.05), suggesting efficient parameter utilization. In contrast, ProtBERT Δ-embeddings require more model capacity (depth=6, 500 trees, learning rate=0.15) to reach their optimum, indicating that the high-dimensional learned representations benefit from increased ensemble complexity.

Performance hierarchy with optimization. Table 1 establishes AAindex per-site hydrophobicity vectors as the superior approach:

- **AAindex per-site hydrophobicity dominance**: Achieves the highest performance across both metrics (99.03% accuracy, 96.71% macro-F1), establishing domain-informed feature engineering as the superior approach for this task.
- **ProtBERT competitive but secondary**: Despite sophisticated pre-training on 200M protein sequences, ProtBERT Δ-embeddings achieve strong but secondary results (98.68% accuracy, 95.85% macro-F1), falling short of biological biochemical encoding.
- **Meaningful performance gaps**: AAindex biochemical feature vectors provide +0.35% accuracy and +0.86% macro-F1 improvements over ProtBERT—differences that are statistically significant ($p = 0.002$) and represent meaningful advances for surveillance applications where minority variant detection is critical.

Statistical significance with rigorous cross-validation. We conducted paired statistical tests on 5-fold cross-validation results to rigorously assess performance differences while avoiding data bias. Each method comparison uses five paired measurements from identical data splits to ensure fair evaluation.

ProtBERT vs AAindex aggregate: ProtBERT significantly outperforms the simple baseline across both metrics. Paired t-tests show substantial improvements in accuracy (+0.66% ± 0.06%, t-statistic = 51.8, p-value = 8.3×10^{-7}) and macro-F1 (+2.98% ± 0.26%). McNemar's test on 262,042 pooled predictions

yields $\chi^2 = 587.4$ (p-value = 9.0×10^{-130}), revealing that ProtBERT corrects twice as many aggregate baseline errors as it introduces, confirming overwhelming superiority of learned representations over simple physicochemical summaries.

AAindex biochemical vs ProtBERT: Domain-informed biochemical feature vectors significantly outperform ProtBERT embeddings. Paired t-tests demonstrate consistent advantages in accuracy (+0.35% ± 0.05%, t-statistic = 14.5, p-value = 1.3×10^{-4}) and macro-F1 (+0.86% ± 0.25%, t-statistic = 6.92, p-value = 0.002). McNemar's test ($\chi^2 = 248.13$, p-value = 6.6×10^{-56}) confirms systematic prediction differences, establishing that explicit biochemical encoding provides genuine methodological advantages over general-purpose protein language models.

The cross-validation framework ensures unbiased statistical inference by using non-overlapping test sets, with each sequence appearing in exactly one fold's evaluation.

5.2 Per-class Analysis

Table 2: Per-class precision, recall, and F_1 scores with optimized hyperparameters. Bold figures mark the best score for each variant.

Variant	Support	Biochemical (1273D)			ProtBERT (1024D)			Aggregate (6D)		
		Prec.	Rec.	F_1	Prec.	Rec.	F_1	Prec.	Rec.	F_1
Alpha	182,869	**0.997**	**0.999**	**0.998**	0.993	0.996	0.994	0.979	0.990	0.984
Beta	3,937	**0.995**	**0.989**	**0.992**	0.988	0.944	0.965	0.948	0.887	0.916
Delta	35,561	**0.982**	**0.979**	**0.980**	0.975	0.976	0.975	0.948	0.955	0.952
Gamma	16,385	**0.996**	**0.994**	**0.995**	0.990	0.973	0.982	0.982	0.965	0.973
Omicron	1,847	**0.972**	**0.939**	**0.955**	0.966	0.886	0.924	0.879	0.754	0.812
Others	20,443	0.886	0.879	0.882	**0.917**	**0.903**	**0.910**	0.822	0.730	0.773

Strong performance across most variant classes. The optimized AAindex biochemical approach achieves superior performance for 5 out of 6 variant classes, with particularly pronounced advantages for challenging minority lineages that are most critical for surveillance applications:

- **Beta (N=3,937)**: F1 advantage of +2.7 pp over ProtBERT (0.992 vs 0.965), demonstrating superior detection of this historically important variant of concern.
- **Omicron (N=1,847)**: F1 advantage of +3.1 pp over ProtBERT (0.955 vs 0.924), crucial for early detection of emerging variants with high surveillance priority.
- **Gamma (N=16,385)**: F1 advantage of +1.3 pp over ProtBERT (0.995 vs 0.982), showing consistent superiority across medium-frequency variants.
- **Others/Unassigned (N=20,443)**: ProtBERT achieves +2.8 pp F1 advantage (0.910 vs 0.882), representing the only class where learned representations outperform biochemical feature vectors.

6 Conclusion

This study provides a systematic comparison of feature extraction approaches for SARS-CoV-2 variant classification using over 260,000 sequences. We demonstrate a clear performance hierarchy: biochemical feature vectors achieve the highest accuracy (macro-F1: 0.967), followed by ProtBERT embeddings (0.958), and simple aggregate features (0.928). Statistical tests confirm these differences are significant, with biochemical features showing particular advantages for challenging minority variants like Beta and Omicron.

These results position protein language models as valuable intermediate solutions for viral surveillance. ProtBERT provides substantial improvements over simple baselines without requiring domain expertise, making it attractive for rapid deployment scenarios. However, carefully engineered biochemical feature vectors achieve superior performance, particularly for minority variant detection.

Rather than competing alternatives, these approaches serve complementary roles. Protein language models offer immediately deployable solutions with strong performance, while specialized biochemical feature engineering achieves optimal accuracy when resources permit. The choice between approaches should depend on specific operational constraints, performance requirements, and available expertise.

Future work investigating fine-tuning of these protein language models would provide valuable insights into their potential when greater computational resources are available. The ProtBERT authors note that "in some tasks you could gain more accuracy by fine-tuning the model rather than using it as a feature extractor" [Elnaggar et al., 2020], suggesting that task-specific adaptation might enable protein language models to match or exceed the performance of our best biochemical feature approaches.

References

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rihawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Protbert. https://huggingface.co/Rostlab/prot_bert, 2020. Accessed: 2024.

Jack Kyte and Russell F Doolittle. A simple method for displaying the hydropathic character of a protein. *Journal of Molecular Biology*, 157(1):105–132, 1982.

Jun Lan, Jiwan Ge, Jinfang Yu, Sisi Shan, Huan Zhou, Shilong Fan, Qi Zhang, Xuanling Shi, Qisheng Wang, Linqi Zhang, and Xinquan Wang. Structure of the sars-cov-2 spike receptor-binding domain bound to the ace2 receptor. *Nature*, 581(7807):215–220, 2020.

Wei-Zhong Lin, Jian-An Fang, Xuan Xiao, and Kuo-Chen Chou. Predicting and analyzing dna-binding domains using a systematic approach to identifying a set of informative physicochemical and biochemical properties. *BMC Systems Biology*, 5(1):47, 2011.

Loris Nanni, Sheryl Brahnam, Michelangelo Paci, and Stefano Ghidoni. Deep learning and handcrafted features for virus image classification. *Journal of Imaging*, 6(12):143, 2020.

Andres D Solis and Shalom Rackovsky. Physicochemical property distributions for accurate and rapid pairwise protein homology detection. *BMC Bioinformatics*, 11(1):145, 2010.