Electronics, Computer and Artificial Intelligence

2 papers

A Multi-Task Transformer for Protein Secondary Structure and Solvent Accessibility Prediction

Image Acharya[1*], Suresh Manandhar[1], Jhashanath Adhikari Subin[1]

[1]Institute of Applied Sciences, Madan Bhandari University of Science and Technology, Nepal
[*]Corresponding author: image.acharya@mbust.edu.np

Abstract

Predicting protein secondary structure (SS) and relative solvent accessibility (RSA) from sequence is valuable because experimental structure determination is time-consuming and costly; however, the two properties are usually modelled separately, and convolutional neural network (CNN) and recurrent neural network (RNN) models have difficulty capturing long-range residue interactions. This paper proposes a Transformer encoder with two task-specific heads that jointly predicts three-class SS (helix, sheet, coil) and three-class RSA (buried, intermediate, exposed) from amino acid sequence alone. Training proceeds in two stages: joint multi-task learning, followed by task-specific fine-tuning of each head. Using Protein Data Bank (PDB) data, with DSSP labels mapped from eight to three classes and RSA computed with FreeSASA, joint training achieved macro F1-scores of 0.905 (SS) and 0.774 (RSA). Fine-tuning increased these to 0.929 and 0.783, compared with 0.838 and 0.681 for a CNN, 0.865 and 0.712 for a bidirectional long short-term memory (BiLSTM) network, and 0.837 and 0.664 for a frozen ProtBERT baseline. On the external CB513 benchmark, the fine-tuned model achieved an SS macro F1-score of 0.896. All labels were derived computationally from PDB structures, and no experimental validation was performed.

Keywords: protein secondary structure prediction, relative solvent accessibility, Transformer, multi-task learning, fine-tuning

Introduction

Protein function depends on three-dimensional structure, which is encoded by the amino acid sequence (Alberts et al., 2002; Dill et al., 2008). Secondary structure (SS), comprising helices, sheets and coils, constitutes the first level of local folding, and relative solvent accessibility (RSA) quantifies the extent to which a residue is exposed to solvent (Tien et al., 2013); both properties support the analysis of protein folding, binding sites and drug targets. UniProt contains over 190 million sequences, whereas the Protein Data Bank (PDB) contains only about 200,000 experimentally determined structures (Berman et al., 2000); sequence-based prediction is therefore required.

Convolutional neural networks (CNNs) capture only local patterns within a fixed receptive field, whereas bidirectional long short-term memory (BiLSTM) networks process positions sequentially and degrade on very long sequences. This limitation is consequential for β-sheets, in which interacting residue pairs may be separated by hundreds of positions (Heffernan et al., 2017; Wang et al., 2016). Most existing models treat SS and RSA as separate single-task problems (Cheng et al., 2008; Yang et al., 2016), and the few multi-task models are based on CNN or recurrent neural network (RNN) backbones (Al-Shameri et al., 2025; Görmez & Aydin, 2022). Transformers (Vaswani et al., 2017) model all residue pairs through self-attention; however, their application to joint SS and RSA prediction has received little attention, and single-stage training does not distinguish shared from task-specific representations.

Accordingly, this study addresses two research questions: (i) whether a self-attention model can surpass CNN and BiLSTM baselines in predicting SS and RSA, and (ii) whether joint training followed by task-specific fine-tuning outperforms single-stage training. The study contributes a multi-task Transformer for joint SS and RSA prediction from raw amino acid sequences, a two-stage training strategy, and a systematic comparison with CNN, BiLSTM and frozen ProtBERT baselines on a PDB-derived dataset and the external CB513 benchmark.

Literature Review / Related Work

Early methods relied on residue propensities (Chou & Fasman, 1979) and, subsequently, on sequence profiles combined with neural networks; PSIPRED achieved 76–78% Q3 accuracy (Jones, 1999), against an estimated accuracy ceiling of 88–90% (Rost & Sander, 1993). Deep learning removed the need for hand-crafted features: DeepCNF combined convolutional networks with conditional random fields (Wang et al., 2016), BiLSTMs exploited context from both directions (Heffernan et al., 2017), and SPIDER2 jointly predicted SS, solvent accessibility and torsion angles (Yang et al., 2016). Transformer-based protein language models such as ProtTrans (Elnaggar et al., 2022) learn representations from very large sequence collections, and TAPE provided a benchmark for such pretrained models (Rao et al., 2019). Multi-task learning (Caruana, 1997) has been applied to SS, solvent accessibility and torsion angles using LSTM backbones (Görmez & Aydin, 2022); however, Transformer-based multi-task frameworks for SS and RSA remain largely unexplored.

Secondary structure and solvent accessibility are interrelated. A residue in an α-helix may be highly exposed when the helix lies on the protein surface, or relatively buried when it is packed within the interior; β-sheet residues are frequently buried because many strands are hydrophobic, although exposed edges can be highly accessible; and coil residues, which lack regular hydrogen bonding, tend to be more exposed on average. This correlation motivates the prediction of both properties within a single model. RSA is defined as the solvent-accessible surface area (SASA) of a residue, measured with a 1.4 Å probe sphere approximating a water molecule (Richards, 1977), divided by the maximum area of the same residue in a reference Gly-X-Gly tripeptide (Tien et al., 2013); values near 0 therefore indicate buried core residues, and values near 1 indicate surface residues.

The Transformer encoder (Vaswani et al., 2017) replaces recurrence with multi-head self-attention. Each position forms query, key and value vectors, and the attention weights are given by $\mathrm{softmax}(QK^{T}/\sqrt{d_k})$, such that every residue can attend directly to every other residue. Each block is completed by residual connections, layer normalisation and positional encodings. For sequence labelling, only the encoder is required, and its per-residue outputs are passed to task-specific heads. Recent Transformer studies include MFTrans, which combines evolutionary profiles with learned embeddings (Chen et al., 2024), and a Transformer trained directly on sequences that achieved competitive Q3 accuracy with substantially fewer computational resources than large language models (Maxime, 2025). AlphaFold2 (Jumper et al., 2021) predicts several structural properties within a single attention-based pipeline, demonstrating the benefit of learning structural features jointly.

Materials and Methods

Data

Sequences and eight-class (Q8) DSSP labels (Kabsch & Sander, 1983) for experimentally determined PDB structures were obtained from the public dataset protein_secondary_structure_from_PDB (LAMM, 2024). The Q8 labels were mapped to Q3 (G, H and I to Helix; E and B to Sheet; T, S and L/C to Coil). RSA was computed per residue with FreeSASA (Mitternacht, 2016), normalised by the Gly-X-Gly maximum areas of Tien et al. (2013), and categorised as Buried (RSA ≤ 0.09), Intermediate (0.09 < RSA ≤ 0.36) or Exposed (RSA > 0.36). The data were split 80/10/10 into training, validation and test sets, identically for all models. Most sequences contained 50–300 residues. SS labels comprised approximately 40% coil, 36% helix and 22% sheet residues, and RSA labels approximately 36% buried and 31% each intermediate and exposed.

Model

Each residue is represented as one of 21 tokens (20 amino acids and a padding token) and mapped by a learnable embedding to 512 dimensions; no evolutionary profiles are used. Proteins are divided into overlapping segments (length 512, stride 256), padded and masked. The shared encoder comprises six Transformer blocks with eight attention heads, a 1024-unit feed-forward layer, ReLU activation, residual connections, layer normalisation and a dropout rate of 0.2. Two linear heads produce per-residue logits for SS (H, E, C) and RSA (B, I, E), with padding excluded from the loss (Figure 1). Hyperparameters were selected on validation data. A dropout rate of 0.1 led to overfitting and 0.3 slowed training, so 0.2 was adopted; the loss weight λ was tested at 0.5, 1.0, 1.5 and 2.0, and 2.0 gave the best trade-off between the tasks; and among the learning rates 1e-3, 5e-4, 1e-4 and 5e-5, 5e-4 performed best in Stage 1. The model dimension, number of layers and learning-rate schedule had the largest effect on performance, dropout and λ a moderate effect, and the number of heads the least.

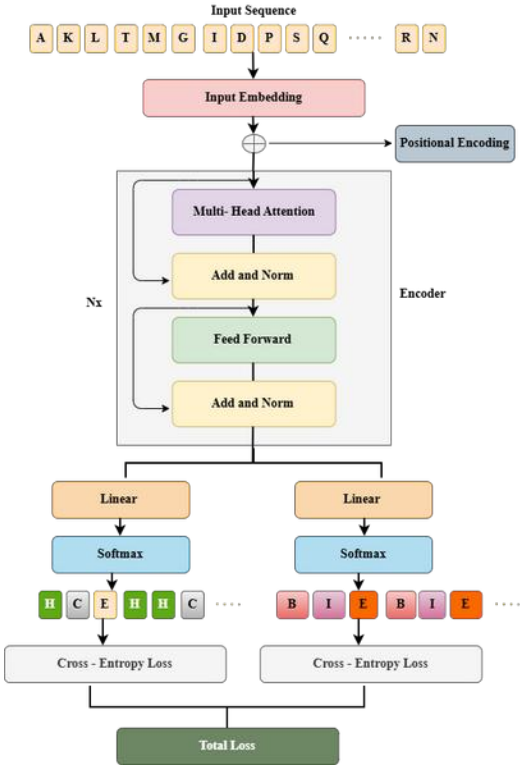
Training and Baselines

In Stage 1, the whole model is trained end-to-end on $L = L_{SS} + \lambda L_{RSA}$ with $\lambda = 2.0$, using class-weighted cross-entropy with label smoothing of 0.1, the Adam optimiser (learning rate 5e-4, cosine annealing), gradient clipping (1.0) and early stopping on the combined validation loss (patience 15). Stages 2A and 2B each start from the best Stage 1 checkpoint: one head is frozen, and the encoder (learning rate 1e-4) and the other head (5e-4) are trained on the loss of that task alone (patience 10). Stage 2B is not initialised from Stage 2A.

Baselines comprise a multi-task CNN (shared multi-scale convolutional encoder with residual connections), a multi-task BiLSTM (stacked bidirectional LSTM layers) and a frozen ProtBERT encoder with a multilayer perceptron (MLP) head. All models were implemented in PyTorch with identical preprocessing, data split, optimiser and early stopping, and were run on a single NVIDIA GeForce RTX 5090 GPU (32 GB) with Python 3.10, PyTorch 2.2.0 and CUDA 13.2. Because the classes are imbalanced, the primary metric is macro F1, the unweighted mean of the per-class F1-scores, where F1 is the harmonic mean of precision (TP / (TP + FP)) and recall (TP / (TP + FN)); TP, FP and FN denote true positives, false positives and false negatives, respectively. Per-class F1, residue-level accuracy and normalised confusion matrices are also reported, with padding excluded.

Figure 1

Joint Multi-Task Architecture With a Shared Transformer Encoder and Two Prediction Heads



Results

Joint training reduced the training and validation losses together without divergence, and test results were consistent with validation results; no evidence of overfitting was observed. Table 1 and Figure 2 report the test results. The jointly trained Transformer achieved SS and RSA macro F1-scores of 0.905 and 0.774, and fine-tuning increased these to 0.929 (Stage 2A) and 0.783 (Stage 2B). The CNN, BiLSTM and frozen ProtBERT baselines scored lower on both tasks.

Figure 2

Macro F1-Score Across Training Stages and Baseline Models

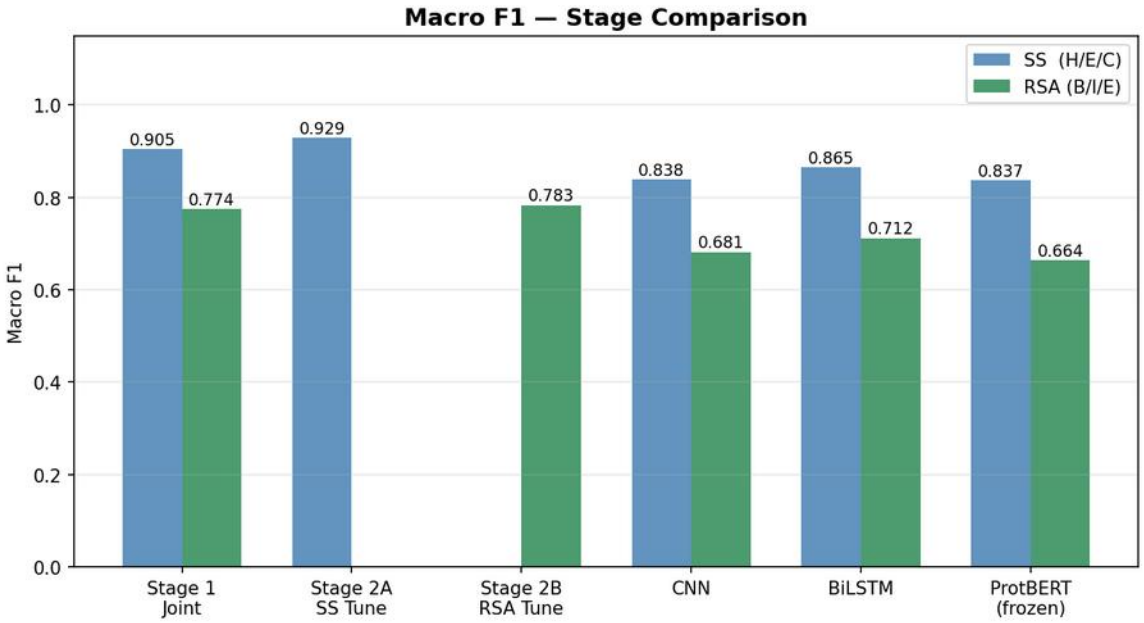


Table 1

Test Results for All Models

Model	SS macro-F1	RSA macro-F1	SS acc.	RSA acc.	CB513 SS macro-F1
CNN	0.838	0.681	0.8402	0.6884	0.792
BiLSTM	0.865	0.712	0.8660	0.7163	0.824
ProtBERT (frozen) + MLP	0.837	0.664	–	–	–
Transformer, Stage 1 (joint)	0.905	0.774	0.9049	0.7781	0.860
Transformer, fine-tuned (2A / 2B)	**0.929**	**0.783**	**–**	**–**	**0.896**

Note. SS = secondary structure; RSA = relative solvent accessibility; acc. = accuracy; CB513 = external benchmark (SS only). A dash indicates a value not reported.

Among the SS classes, Helix was predicted best and Sheet worst. Sheet was most frequently confused with Coil, and Intermediate with Exposed, and the Transformer exhibited the lowest Sheet–Coil confusion of the three architectures (Figure 3). Joint RSA training and self-attention together appear to reduce this structural ambiguity.

Compared with the frozen ProtBERT encoder, the fine-tuned Transformer achieved higher per-class F1-scores for every class: 0.942 (H), 0.925 (E) and 0.912 (C) for SS, against 0.881, 0.819 and 0.810, and 0.838 (B), 0.702 (I) and 0.787 (E) for RSA, against 0.748, 0.543 and 0.701. The difference was largest for the difficult Intermediate RSA class.

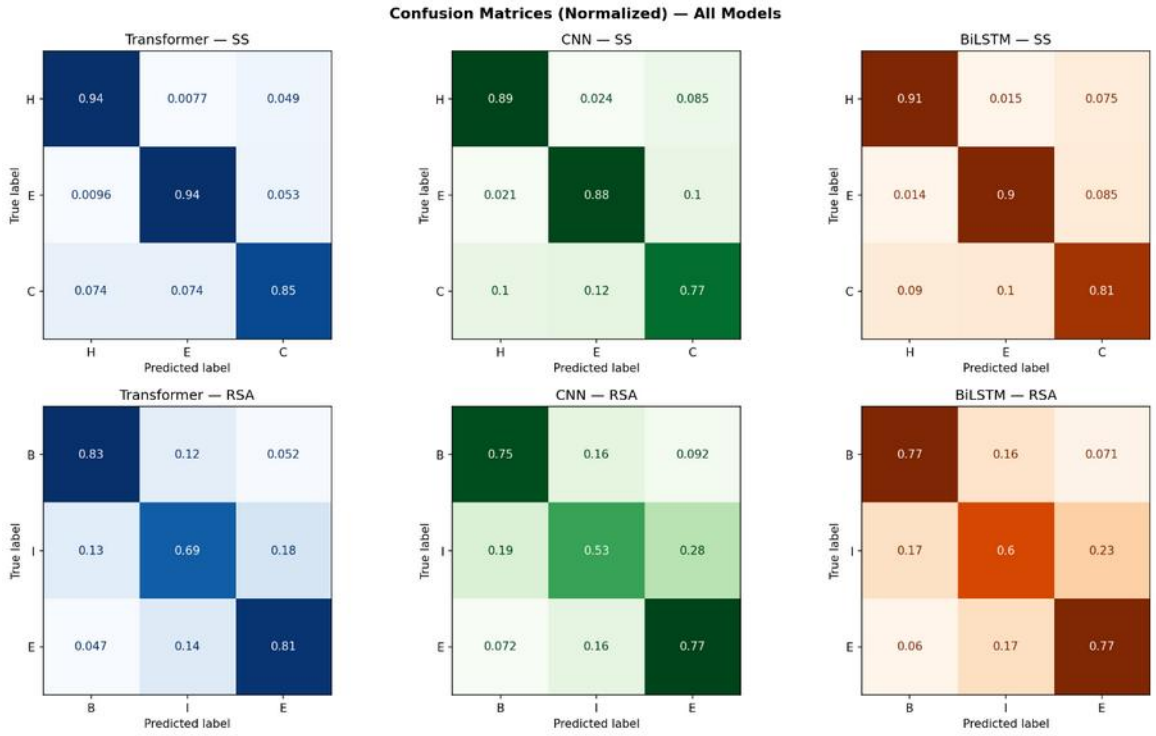
Performance increased with sequence length. Proteins longer than 512 residues achieved macro F1-scores of 0.908 (SS) and 0.775 (RSA), against 0.865 and 0.740 for proteins of 512 residues or fewer, and the 0–100 residue range scored lowest (0.775 and 0.613), with the largest increase occurring between the two shortest length groups (SS 0.775 to 0.861; RSA 0.613 to 0.741). The 512-residue threshold corresponds to the chunk length of the sliding-window procedure: longer proteins require several overlapping chunks, with predictions averaged where chunks overlap.

For proteins longer than 512 residues, the SS F1-scores were 0.931 (H), 0.904 (E) and 0.889 (C), and the RSA F1-scores were 0.839 (B), 0.705 (I) and 0.780 (E).

On the external CB513 benchmark, the fine-tuned model achieved an SS macro F1-score of 0.896, exceeding the joint model (0.860), the BiLSTM (0.824) and the CNN (0.792), with Helix, Sheet and Coil F1-scores of 0.917, 0.888 and 0.883 and diagonal confusion-matrix values of 0.94, 0.92 and 0.85. The CNN exhibited the greatest Coil confusion, with Coil accuracy decreasing to 0.73. RSA could not be evaluated externally because no external benchmark was available.

Figure 3

Normalised Confusion Matrices for All Models (Top: Secondary Structure; Bottom: RSA)



Discussion

Transformer architecture. The Transformer outperformed the CNN and BiLSTM on both tasks, most clearly for Sheet, in which interacting residues are separated by long distances in sequence. CNNs are constrained by local receptive fields and BiLSTMs by sequential processing, whereas self-attention accesses every position directly.

Multi-task learning. Joint training appeared to assist the encoder in separating Sheet from Coil, consistent with the biological relationship between burial and regular secondary structure, as indicated by the lower Sheet–Coil confusion in Figure 3.

Two-stage training. Fine-tuning yielded a larger gain for SS (+0.024) than for RSA (+0.009). Dedicated optimisation of a single loss, once shared representations have been established, refines task-specific decision boundaries without interference from the gradient of the other task.

The smaller RSA gain suggests that the shared representation is already well suited to RSA, which is a smoother problem than three-class secondary-structure prediction.

Pretrained models. ProtBERT was pretrained in a self-supervised manner on approximately 393 million sequences to learn general-purpose representations, and was not designed specifically for SS or RSA prediction (Elnaggar et al., 2022). The proposed model contributes a task-specific two-stage strategy, a lightweight encoder that requires no large-scale pretraining, and empirical evidence that this strategy outperforms a frozen ProtBERT encoder with an MLP head under the reported setup; the comparison does not extend to fine-tuned pretrained protein language models. The principal distinction lies between general pretraining and task-specific adaptation: frozen representations cannot adapt to the joint objective.

Error analysis. Performance increased with sequence length; this is an observed association rather than evidence that longer sequences are inherently easier to predict, and it may reflect the greater context available to self-attention, although differences in composition between length groups cannot be excluded. Sliding-window inference performed well above 512 residues. Sheet remains the most difficult SS class because it depends on long-range interactions, and Intermediate remains the most difficult RSA class because RSA is a continuous value discretised with fixed thresholds, so that residues near the class boundaries are inherently ambiguous.

Limitations

The model uses simple trainable embeddings rather than pretrained embeddings or evolutionary profiles. All labels were computed from PDB structures, and the predictions were not verified experimentally. RSA was evaluated only on the study dataset, and discretisation into three classes introduces boundary ambiguity. The model predicts only SS and RSA; disordered regions, binding sites and higher-order interactions are not addressed. In addition, no single-task ablation isolating the contribution of multi-task learning is reported, and variation across random seeds is not reported.

Conclusion

A multi-task Transformer trained in two stages predicted secondary structure and solvent accessibility from sequence with macro F1-scores of 0.929 and 0.783, exceeding the CNN, BiLSTM and frozen ProtBERT baselines, and achieved 0.896 on CB513. The results are consistent with contributions from self-attention, multi-task learning and task-specific fine-tuning, and support joint learning followed by task-specific fine-tuning as an effective strategy for per-residue prediction from sequence alone.

In practice, the proposed approach provides a computationally efficient way to predict both properties directly from amino acid sequences without requiring large-scale pretraining or multiple sequence alignments. By combining both prediction tasks within a single model, it offers an alternative to training separate models for each property. Future work will explore the integration of three-dimensional or AlphaFold2-derived features, evaluation on more diverse datasets, experimental validation, and extension to binding-site and disordered-region prediction.

Acknowledgments

The authors thank the faculty and staff of the Institute of Applied Sciences, Madan Bhandari University of Science and Technology, for supervision, feedback, computational resources and

institutional support. They are also grateful to Dr. Shree Krishna Acharya for constructive feedback on this work, and to the Hugging Face and Protein Data Bank communities for making the datasets publicly available.

Author Contributions

Image Acharya: Conceptualization, Methodology, Software, Validation, Formal analysis, Investigation, Visualization, Writing – original draft. Suresh Manandhar and Jhashanath Adhikari Subin: Supervision, Writing – review and editing.

Funding

The authors received no financial support for the research, authorship, or publication of this paper.

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

The sequence and secondary-structure data are publicly available (LAMM, 2024); RSA labels were computed from PDB structures with FreeSASA. Code is available from the corresponding author on reasonable request.

Ethical Approval / Consent

Not applicable. The study used only publicly available protein sequence and structure data.

References

Alberts, B., Johnson, A., Lewis, J., Raff, M., Roberts, K., & Walter, P. (2002). *Molecular biology of the cell* (4th ed.). Garland Science. https://www.ncbi.nlm.nih.gov/books/NBK21054/

Al-Shameri, Y. N. H., Kösesoy, İ., Gündüz, H., & Yılmaz, Ö. F. (2025). Scalable and efficient protein secondary structure prediction using autoencoder-reduced ProtBERT embeddings. *Applied Sciences, 15*(13), 7112.

Berman, H. M., Westbrook, J., Feng, Z., Gilliland, G., Bhat, T. N., Weissig, H., Shindyalov, I. N., & Bourne, P. E. (2000). The Protein Data Bank. *Nucleic Acids Research, 28*(1), 235–242.

Caruana, R. (1997). Multitask learning. *Machine Learning, 28*(1), 41–75.

Cheng, J., Tegge, A. N., & Baldi, P. (2008). Machine learning methods for protein structure prediction. *IEEE Reviews in Biomedical Engineering, 1*, 41–49.

Chen, Y., Chen, G., & Chen, C. Y. C. (2024). MFTrans: A multi-feature transformer network for protein secondary structure prediction. *International Journal of Biological Macromolecules, 267*, 131311.

Chou, P. Y., & Fasman, G. D. (1979). Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology and Related Areas of Molecular Biology, 47*, 45–148.

Dill, K. A., Ozkan, S. B., Shell, M. S., & Weikl, T. R. (2008). The protein folding problem. *Annual Review of Biophysics, 37*(1), 289–316.

Elnaggar, A., Heinzinger, M., Dallago, C., Rehawi, G., Wang, Y., Jones, L., Gibbs, T., Feher, T., Angerer, C., Steinegger, M., Bhowmik, D., & Rost, B. (2022). ProtTrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 44*(10), 7112–7127.

Görmez, Y., & Aydin, Z. (2022). IGPRED-MultiTask: A deep learning model to predict protein secondary structure, torsion angles and solvent accessibility. *IEEE/ACM Transactions on Computational Biology and Bioinformatics, 20*(2), 1104–1113.

Heffernan, R., Yang, Y., Paliwal, K., & Zhou, Y. (2017). Capturing non-local interactions by long short-term memory bidirectional recurrent neural networks for improving prediction of protein secondary structure, backbone angles, contact numbers and solvent accessibility. *Bioinformatics, 33*(18), 2842–2849.

Jones, D. T. (1999). Protein secondary structure prediction based on position-specific scoring matrices. *Journal of Molecular Biology, 292*(2), 195–202.

Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., Bridgland, A., Meyer, C., Kohl, S. A. A., Ballard, A. J., Cowie, A., Romera-Paredes, B., Nikolov, S., Jain, R., Adler, J., … Hassabis, D. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature, 596*(7873), 583–589.

Kabsch, W., & Sander, C. (1983). Dictionary of protein secondary structure: Pattern recognition of hydrogen-bonded and geometrical features. *Biopolymers, 22*(12), 2577–2637.

LAMM: MIT Laboratory for Atomistic and Molecular Mechanics. (2024). *protein_secondary_structure_from_PDB* [Data set]. Hugging Face. https://huggingface.co/datasets/lamm-mit/protein_secondary_structure_from_PDB

Maxime. (2025). *Protein secondary structure prediction using transformer model* [Preprint]. arXiv.

Mitternacht, S. (2016). FreeSASA: An open source C library for solvent accessible surface area calculations. *F1000Research, 5*, 189.

Rao, R., Bhattacharya, N., Thomas, N., Duan, Y., Chen, P., Canny, J., Abbeel, P., & Song, Y. (2019). Evaluating protein transfer learning with TAPE. *Advances in Neural Information Processing Systems, 32*.

Richards, F. M. (1977). Areas, volumes, packing, and protein structure. *Annual Review of Biophysics and Bioengineering, 6*(1), 151–176.

Rost, B., & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology, 232*(2), 584–599.

Tien, M. Z., Meyer, A. G., Sydykova, D. K., Spielman, S. J., & Wilke, C. O. (2013). Maximum allowed solvent accessibilities of residues in proteins. *PLoS ONE, 8*(11), e80635.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., & Polosukhin, I. (2017). Attention is all you need. *Advances in Neural Information Processing Systems, 30*.

Wang, S., Peng, J., Ma, J., & Xu, J. (2016). Protein secondary structure prediction using deep convolutional neural fields. *Scientific Reports, 6*(1), 18962.

Yang, Y., Heffernan, R., Paliwal, K., Lyons, J., Dehzangi, A., Sharma, A., Wang, J., Sattar, A., & Zhou, Y. (2016). SPIDER2: A package to predict secondary structure, accessible surface area, and main-chain torsional angles by deep neural networks. In *Prediction of protein secondary structure* (pp. 55–63). Springer.

Differentiating Gram-positive and Gram-negative Bacteria using Machine Learning assisted Raman Spectroscopy

Iswori Prasad Paudel[1,2,*], Suresh Manandhar[1] and Rijan Maharjan[1,2]

[1]Madan Bhandari University of Science and Technology, Thaha-9, Makwanpur
[2]Phutung Research Institute, Tarakeshwor-7, Kathmandu
[*]iswori.prasad.paudel@mbust.edu.np

Abstract

Bacterial infections kill millions of people every year, and a major reason is the time for diagnosis. Delay in diagnosis leads to the unnecessary use of antibiotics, which enhances antimicrobial resistance (AMR). By classifying bacteria into Gram-positive and Gram-negative, clinicians can narrow down the broad-spectrum use of antibiotics. Conventional methods such as Gram-staining require culturing of bacteria, which takes 24–72 hr. Here, we present a machine learning-assisted Raman spectroscopy pipeline for culture-independent, rapid Gram classification. We use a Partial Least Squares Discriminant Analysis (PLS-DA) model, which was trained on 56,000 Raman spectra spanning 28 bacterial species from the publicly available Ho et al. dataset (Ho et al., 2019) and validated using Leave-One-Species-Out Cross-Validation (LOSO-CV). The decision threshold was optimised on an independent tune set. On the held-out reference test set, the model achieved an ROC-AUC of 0.954 with 88.7% sensitivity and 90.6% specificity. To assess cross-instrument generalisability, the model was further evaluated on 147 spectra acquired on an in-house Raman system from *Escherichia coli* (*E. coli*) ATCC 25922 (Gram-negative) and *Enterococcus faecalis* (*E. faecalis*) ATCC 19433 together with *Enterococcus* spp. isolates from Nepali fermented foods (Gram-positive), which yielded 84.5% sensitivity and 92.1% specificity. These results demonstrate that a computationally lightweight, interpretable linear classifier can generalise Gram-type discrimination across instruments and bacterial origins without retraining.

Keywords: Raman spectroscopy, Gram classification, partial least squares discriminant analysis, machine learning, antimicrobial resistance

Introduction

Bacterial infections remain one of the most serious threats to global health. In 2019 only, approximately 7.7 million worldwide deaths were caused by 33 bacterial pathogens, roughly one in every eight deaths that year (Ikuta et al., 2022). A critical but often overlooked bottleneck for the management of these infections is the time required to determine a bacterium's Gram type, a significant actionable piece of information a clinician needs before prescribing antibiotics. Gram-positive and Gram-negative have differences in their cell wall architecture, and these differences govern antibiotic effectiveness. This is crucial for antibiotic management because every hour of delay in appropriate antibiotic initiation increases sepsis mortality by approximately 7.6% (Kumar et al., 2006).

The widely used Gram staining process involves culture growth, colony isolation, Gram staining, and microscopic examination, which requires a minimum of 24–48 hours and even longer for some slow-growing species (Reller et al., 2009). Other molecule-based alternatives such as polymerase chain reaction (PCR) help to reduce this delay but require expensive reagents and infrastructure that are rarely available in resource-limited settings (Van Belkum & Dunne, 2013).

Raman spectroscopy as an analytical technique offers an innovative and powerful alternative. It provides the molecular-level information of a bacterial sample in seconds, without chemical labels, stains, or sample destruction, producing a spectrum that encodes rich information about cell constituents including cell wall composition, peptidoglycan and lipopolysaccharide (LPS), the molecular determinants of Gram type. However, Raman spectra are high-dimensional, noisy, and strongly multicollinear, and subtle differences in these spectra of chemically similar bacteria make visual interpretation difficult. Machine Learning (ML) bridges this gap. Among different ML methods, Partial Least Squares Discriminant Analysis (PLS-DA) stands out for its ability to handle multicollinearity and high-dimensional spectral data while remaining computationally efficient and interpretable (Barker & Rayens, 2003; Wold et al., 2001). Several studies have shown Raman spectroscopy for Gram-type classification (Jarvis & Goodacre, 2004; Lorenz et al., 2017), but rigorous cross-instrument validation remains uncommon in the literature. In this study, we trained the model on publicly available spectra obtained from one instrument, and demonstrated performance of the model on spectra from a completely independent laboratory setup. For the classification task, we (i) train and cross-validate a PLS-DA Gram classifier using species-level holdout validation on the large publicly available reference dataset (Ho et al., 2019); (ii) optimise its decision threshold on an independent tune set; (iii) evaluate it on a held-out reference test set; and (iv) validate it on spectra acquired in our own laboratory on a custom-built, in-house Raman spectroscopy setup, including environmental *Enterococcus* isolates from Nepali fermented foods (a bacterial population not represented in the training data).

Materials and Methods

2.1. Dataset

2.1.1. Reference Dataset

Raman spectra from reference dataset consist of 56,000 spectra from 28 bacterial species (17 Gram-positive and 11 Gram-negative) (Ho et al., 2019). Each spectrum spanned 381.98–1792.4 cm^{-1} at 1,000 wavenumber points. There are three different datasets: a training set (2,000 spectra per species; 56,000 total; used for training and selection of LVs), a tune set (100 spectra per species; 2,800 total; used solely for threshold optimisation), and a test set (100 spectra per species; 2,800 total; evaluated exactly once). Training and test cultures were entirely independent, preventing culture-level data leakage.

2.1.2. Laboratory Dataset

A total of 147 spectra were collected in our laboratory using a Raman spectroscopy setup equipped with a 784.65 ± 0.02 nm laser diode (Thorlabs LD-785-SEV-300), a water dipping objective (Olympus XLUMPLFLN N20X-PFH) with high Numerical Aperture (NA) = 1, and a cooled spectrometer (Avantes AVASPEC-HSC1021X58TEC-EVO). Samples comprised *E. coli* ATCC 25922 (Gram-negative; n = 76) and *E. faecalis* ATCC 19433 together with *Enterococcus* spp. isolates from Nepali fermented foods (Gram-positive; n = 71). Bacterial suspensions were prepared in phosphate-buffered saline (PBS) and deposited on calcium fluoride (CaF_2) substrates for spectrum acquisition.

2.2. Preprocessing steps

For the reference dataset, a fifth-degree polynomial was subtracted to correct the background, and spectra were min-max normalised. For laboratory spectra, a four-stage pipeline was applied: (i) PBS background subtraction; (ii) autofluorescence baseline correction via quantile-regression

polynomial fitting (degree 11) using the pybaselines library (Erb, 2022); (iii) cropping to the fingerprint region and interpolation to align wavenumbers with the reference dataset; and (iv) min-max normalisation to [0, 1].

2.3. PLS-DA and Cross Validation

PLS-DA was implemented using PLSRegression from scikit-learn (Pedregosa et al., 2011). The number of latent variables (LVs) was selected over the range $A \in \{2, \ldots, 15\}$ by maximising the cross-validated ROC-AUC under Leave-One-Species-Out Cross-Validation (LOSO-CV), implemented with LeaveOneGroupOut from scikit-learn with species identity as the grouping variable (28 folds). PLS-DA produces a continuous discriminant score $\hat{y} \in \mathrm{R}$. Rather than using the conventional threshold of 0.5, the decision threshold was optimised on the tune set using the ROC curve: the threshold was set to the smallest value at which sensitivity equalled specificity. This threshold was then applied, unchanged, to both the reference test set and the laboratory dataset. Model performance was assessed using ROC-AUC, sensitivity, specificity, and Matthews Correlation Coefficient (MCC). All analyses were performed in Python 3.10 on a standard laptop (Intel Core i5-1135G7, 8 GB RAM) without GPU resources.

Results

3.1. Model Training: Selecting the Optimal Number of LVs

The first step was to find the optimal number of latent variables for the PLS-DA model. Too few may underfit the model, and too many start picking up noise. We evaluated LVs $A = 2$ to 15 using LOSO-CV, a validation strategy that withholds each bacterial species entirely in one fold to ensure the model is tested on species it has never seen. This is more appropriate than standard *k*-fold cross-validation, where spectra from the same species often appear in both training and validation folds, and produces a more realistic estimate of how the model will behave in practice. The optimal number of latent variables was $A = 6$, at which the cross-validated ROC-AUC, sensitivity, and specificity all reached their peak as shown in Figure 1. Beyond six components, performance slightly declined and plateaued, suggesting that additional latent variables began to encode species-specific noise rather than Gram-type biochemistry.

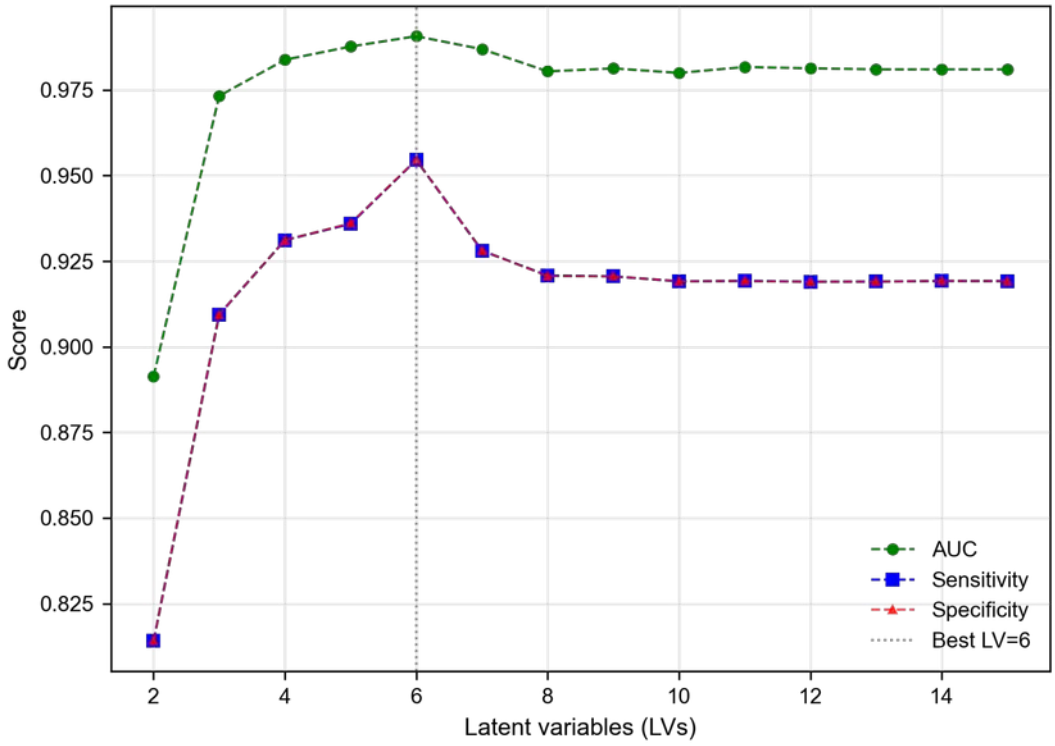


Figure 1
Selection of optimal number of latent variables using LOSO-CV. The x-axis shows the number of latent variables (A = 2–15); the y-axis shows the cross-validated score (AUC, sensitivity, specificity). The dotted vertical line marks A = 6.

3.2. Decision Threshold Optimisation on the Tune Set

PLS-DA outputs a continuous score $\hat{y}$. To convert this into a binary Gram-type prediction, a decision threshold must be chosen. Using the conventional default of 0.5 is problematic here because of class imbalance. The training set has more Gram-positive spectra than Gram-negative (17:11), which pushes the model's score distribution toward the dominant class. Applying a threshold of 0.5 to an imbalanced classifier systematically misclassifies one class at the expense of the other. Therefore, we derived the threshold (t) from the ROC curve of an independent tune set with 2,800 spectra. The threshold value is set to t = 0.802 at which sensitivity equals specificity (Figure 2). This threshold was then fixed and applied without modification to the reference test set and the laboratory dataset.

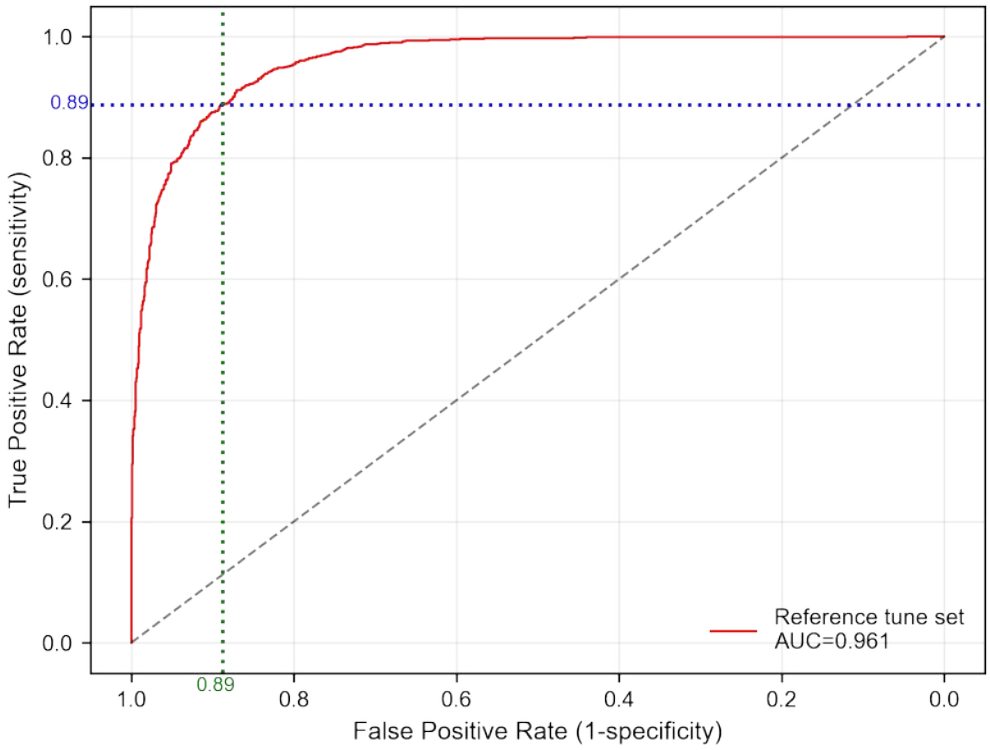


Figure 2
ROC curve on the tune set used for threshold selection. The operating point (marked by a dashed cross-hair) corresponds to t = 0.802, where sensitivity = specificity = 89%.

3.3. Performance on the Reference Test Set

With *A* and *t* fixed, we evaluated the model on the held-out reference test set. The test set also has 2,800 spectra from 28 species, from cultures entirely independent of training. This set was used exactly once, after all modelling decisions had been finalised. The model achieved an ROC-AUC of 0.954, sensitivity of 88.7%, specificity of 90.6%, and an MCC of 0.784 (Figure 3c). The score plot (Figure 3b) shows separation between Gram-classes along LV1 and LV2. The confusion matrix (Figure 3d) shows 1,508 Gram-positive spectra correctly classified against 192 misclassified, and 997 Gram-negative correctly classified against 103 misclassified.

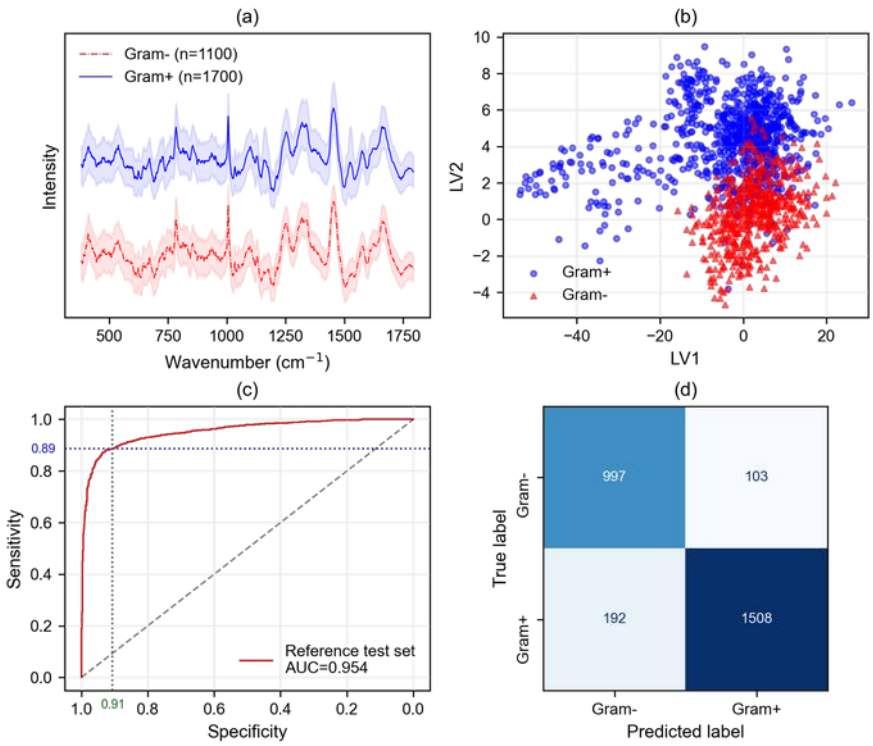


Figure 3
Performance on the reference test set. (a) Mean-variance plot of Gram-positive and Gram-negative spectra; (b) PLS-DA score plot (LV1 vs LV2) showing class separation; (c) ROC curve with AUC = 0.954 and operating point marked; (d) Confusion matrix.

3.4. Cross-Instrument Validation on the Laboratory Dataset

The most stringent test of any trained spectral model is its performance on data from a completely different instrument, acquired by a different team, in a different laboratory, on bacterial cultures that were never part of the training set. We carried out exactly this test. The 147 laboratory spectra required a more involved preprocessing pipeline than the reference data because the custom-built spectrometer introduces different background sources (PBS medium scattering, variable autofluorescence from environmental isolates). After the four-stage pipeline described in Section 2.2, the model trained on Ho et al. (2019) reference spectra was applied directly to the laboratory spectra with no retraining. The results are shown in Figure 4. The model achieved an ROC-AUC of 0.939, sensitivity of 84.5%, and specificity of 92.1%. The MCC of 0.770 confirms solid balanced performance despite the small dataset. The confusion matrix (Figure 4d) shows 70 of 76 *E. coli* spectra correctly classified as Gram-negative, and 60 of 71 *Enterococcus* spectra correctly classified as Gram-positive.

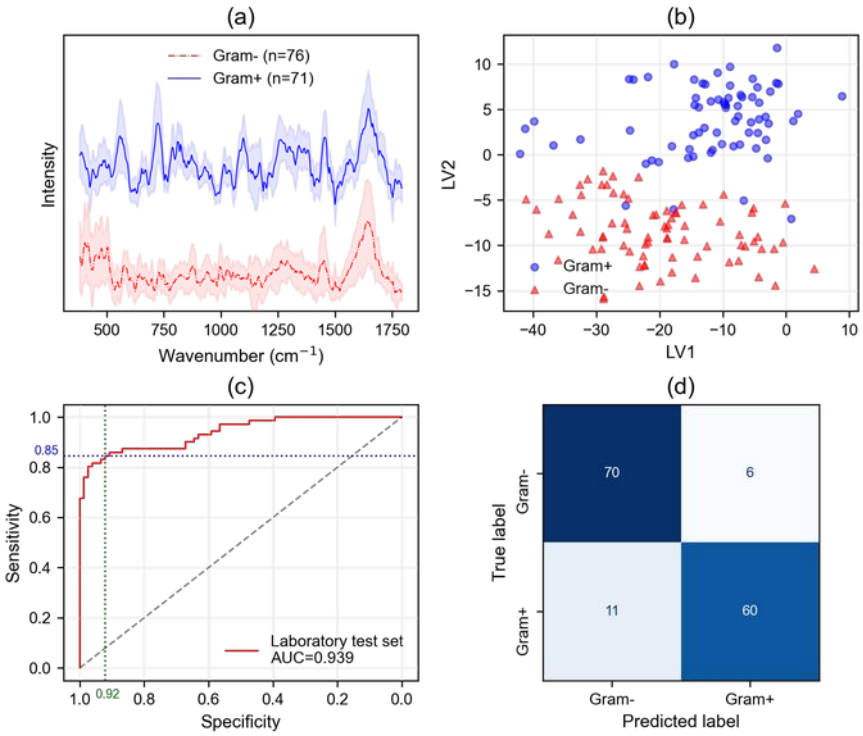


Figure 4
Performance on the independent laboratory test set. (a) Mean-variance plot of E. coli (n = 76) and Enterococcus (n = 71) spectra from the custom-built setup; (b) Score plot; (c) ROC curve with AUC = 0.939; (d) Confusion matrix.

Discussion

This study pursued a deliberate trade-off: simplicity and generalisability over peak performance. PLS-DA with six latent variables is not the most powerful classifier available for Raman spectral data. Convolutional neural networks, for example, have achieved species-level identification accuracies above 97% on the same reference dataset (Ho et al., 2019). But neural networks of that capacity require GPU resources, thousands of spectra per class for stable training, and produce opaque decision boundaries that are difficult to audit in a clinical context. By contrast, the PLS-DA model trained here runs in milliseconds on a standard laptop. The most important finding of this study is not the reference test set performance, which is consistent with published results from Jarvis and Goodacre (2004) and Lorenz et al. (2017), but rather that this performance is largely preserved when the model is applied to data from a completely different instrument and from bacterial isolates it has never encountered. The 4.2% drop in sensitivity (88.7% to 84.5%) and 1.5% increase in specificity (90.6% to 92.1%) across this inter-laboratory gap, achieved with only a generic preprocessing pipeline and no model retraining, supports the feasibility of deploying this approach in resource-limited settings where instruments will inevitably differ from the reference system used for training. The LOSO-CV strategy was central to achieving this generalisability. Standard *k*-fold cross-validation on spectral datasets, where each species with 2,000 spectra, allows spectra from the same species to appear in both training and validation folds. The model can then exploit species-specific spectral signatures rather than

the Gram-type biochemistry it needs to learn. LOSO-CV however eliminates this shortcut and makes the model to learn features that work for new unseen species.

Limitations

Several limitations warrant acknowledgement. First, the laboratory validation dataset is small (147 spectra, two bacterial types). The inter-laboratory claim would be substantially strengthened by expanding to at least six to eight species. Second, the model does not include a rejection mechanism for out-of-class spectra; in a clinical setting, samples from species outside the 28-species training distribution, or samples contaminated with biological fluids, would be forced into a Gram-positive or Gram-negative decision with false confidence. Third, the environmental *Enterococcus* isolates from Nepali fermented foods are treated as a single group together with the ATCC reference strain; their spectral diversity relative to laboratory strains deserves dedicated characterisation. Each of these limitations defines a clear priority for future work.

Conclusion

We have demonstrated that a computationally lightweight PLS-DA model trained on a large public reference dataset of bacterial Raman spectra can classify bacteria as Gram-positive or Gram-negative with clinically meaningful accuracy (89% sensitivity, 91% specificity on the reference test set; 85% sensitivity, 92% specificity on an independent laboratory dataset) without any retraining, using only a standard preprocessing pipeline. The key methodological contribution is the use of species-level LOSO-CV for model selection, which produces performance estimates that generalize more reliably to new bacterial species and new instruments than the spectrum-level cross-validation commonly reported in the literature. These results support the feasibility of developing portable, culture-independent, reagent-free Gram-classification devices for resource-limited settings, where faster and more accessible bacterial diagnostics could meaningfully reduce the inappropriate antibiotic use that drives antimicrobial resistance.

Acknowledgments

The authors thank Madan Bhandari University of Science and Technology and Phutung Research Institute for providing the facilities and support, and Ho et al. for making their Raman spectra dataset publicly available.

Funding

This work was funded by The World Academy of Sciences (TWAS).

Conflict of Interest

The authors declare no conflict of interest.

Data Availability

The data are available from the corresponding author on request.

References

Barker, M., & Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, *17*(3), 166–173. https://doi.org/10.1002/cem.785

Erb, D. (2022). pybaselines: A Python library of algorithms for the baseline correction of experimental data. Zenodo. https://doi.org/10.5281/zenodo.5608581

Ho, C.-S., Jean, N., Hogan, C. A., Blackmon, L., Jeffrey, S. S., Holodniy, M., Banaei, N., Saleh, A. A. E., Ermon, S., & Dionne, J. (2019). Rapid identification of pathogenic bacteria using Raman spectroscopy and deep learning. *Nature Communications*, *10*(1), Article 4927. https://doi.org/10.1038/s41467-019-12898-9

Ikuta, K. S., Swetschinski, L. R., Aguilar, G. R., et al. (2022). Global mortality associated with 33 bacterial pathogens in 2019: A systematic analysis for the Global Burden of Disease Study 2019. *The Lancet*, *400*(10369), 2221–2248. https://doi.org/10.1016/S0140-6736(22)02185-7

Jarvis, R. M., & Goodacre, R. (2004). Discrimination of bacteria using surface-enhanced Raman spectroscopy. *Analytical Chemistry*, *76*(1), 40–47. https://doi.org/10.1021/ac034689c

Kumar, A., Roberts, D., Wood, K. E., Light, B., Parrillo, J. E., Sharma, S., Suppes, R., Feinstein, D., Zanotti, S., Taiberg, L., Gurka, D., Kumar, A., & Cheang, M. (2006). Duration of hypotension before initiation of effective antimicrobial therapy is the critical determinant of survival in human septic shock. *Critical Care Medicine*, *34*(6), 1589–1596. https://doi.org/10.1097/01.CCM.0000217961.75225.E9

Lorenz, B., Wichmann, C., Stöckel, S., Rösch, P., & Popp, J. (2017). Cultivation-free Raman spectroscopic investigations of bacteria. *Trends in Microbiology*, *25*(5), 413–424. https://doi.org/10.1016/j.tim.2017.01.002

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, *12*, 2825–2830.

Reller, L. B., Weinstein, M., Jorgensen, J. H., & Ferraro, M. J. (2009). Antimicrobial susceptibility testing: A review of general principles and contemporary practices. *Clinical Infectious Diseases*, *49*(11), 1749–1755. https://doi.org/10.1086/647952

Van Belkum, A., & Dunne, W. M., Jr. (2013). Next-generation antimicrobial susceptibility testing. *Journal of Clinical Microbiology*, *51*(7), 2018–2024. https://doi.org/10.1128/JCM.00313-13

Wold, S., Sjöström, M., & Eriksson, L. (2001). PLS-regression: A basic tool of chemometrics. *Chemometrics and Intelligent Laboratory Systems*, *58*(2), 109–130. https://doi.org/10.1016/S0169-7439(01)00155-1