Advanced Deep Learning for Enhanced Peptide Identification in Proteomics

Vignesh Babu J S & Saranath P

Introduction and Motivation

Peptide identification via mass spectrometry is central to proteomics and impacts clinical diagnostics and therapeutic research. Despite advancements by frameworks like AlphaPeptDeep [**?**] and Prosit [**?**], current methods lack:

- Reliable uncertainty estimates that enhance model interpretability.
- Rich PTM embeddings that capture complex chemical properties.

Addressing these limitations is crucial for improving identification accuracy and enabling faster, more reliable proteomic analyses.

Literature Survey

Deep learning has revolutionized the prediction of peptide properties in proteomics, surpassing traditional methods like the iRT-calculator and ELUDE. **AlphaPeptDeep** [**?**] is a prime example of this progress due to its modular design and ease of use.

Key Features

- **Rich Embedding Strategies:** The framework uses detailed chemical descriptors for both amino acid sequences and post-translational modifications (PTMs), allowing generalization to complex or poorly characterized modifications.
- **Robust Transfer Learning:** Pre-trained models can be fine-tuned on limited data from new experiments, significantly enhancing predictive performance (e.g., improvements in PCC90).
- **User-Friendly Model Shop:** Ready-to-use templates based on LSTM, CNN, and transformer architectures simplify the development and customization of deep learning models.

Research Gap

Although significant progress has been made with methods such as AlphaPeptDeep and Prosit, several key challenges remain unaddressed:

1. Existing methods do not yet provide robust confidence intervals for predictions, which limits model interpretability.
2. Detailed chemical descriptors for PTM embedding are not fully leveraged in current approaches.
3. There is a lack of real-time inference capabilities on resource-constrained devices.
4. Integration of sequence data with structural insights is still not comprehensively implemented.

Our approach aims to tackle these gaps by further enhancing uncertainty quantification, improving PTM embedding quality, and exploring multi-modal integration strategies.

Project Objectives & Methodology

Thread A: Uncertainty & Interpretability. We will integrate Monte Carlo dropout and deep ensembles into our RT/MS^2 models to generate prediction distributions and confidence intervals, then benchmark PICP and calibration curves against deterministic baselines.

Thread B: PTM Embedding. By augmenting the existing PTM vector with molecular weight, hydrophobicity and polarity—and applying an attention layer to weight these features contextually—we aim to boost spectral-prediction accuracy, assessed via Pearson correlation improvements.

Thread C: Multi-Modal Fusion. Sequence embeddings will be fused with structural features (secondary structure, solvent accessibility from AlphaFold) to enhance RT/MS^2 predictions, with gains quantified against sequence-only models.

Expected Outcomes and Evaluation Strategy

Outcomes:

- Models with robust uncertainty quantification offering reliable confidence intervals.
- Enhanced spectral predictions through enriched PTM embeddings.

Evaluation:

- Use statistical metrics (Pearson correlation, R^2, and PICP) for benchmarking.
- Compare against current state-of-the-art methods to assess improvements.

Timeline (1.5 Months)

Week	Milestones
1	Literature review, dataset acquisition, and baseline experiments
2	Implement and test Thread A (Uncertainty Quantification)
3	Develop and integrate Thread B (Enhanced PTM Embedding)
4	Preliminary benchmarking and refinement of Threads A & B
5	Explore structural feature extraction (Thread C)
6	Final evaluations, documentation, and preparation of deliverables

Preliminary Results and Preparatory Work

Initial experiments using the AlphaPeptDeep framework [**?**] have achieved MS2 predictions with Pearson correlations exceeding 90% and RT model fine-tuning that improves R^2 from 0.93 to above 0.98. These promising results validate our planned improvements. We have access to GPU clusters, and our implementation will use Python with PyTorch. A GitHub repository will host the code along with thorough documentation.

Results Summary: Thread A

Overview: Experiments on RT and MS2 uncertainty were conducted using the dataset available here.

RT Predictions: The RT model shows a high mean absolute error ($\approx$32.68) and very low PICP ($\approx$2.1%), indicating overconfident, narrow intervals. MC dropout and ensemble methods yield MPIW values of $\approx$0.1905 and $\approx$0.2464, respectively, but both share the same low PICP, suggesting the need for additional calibration or data.

MS2 Predictions: While the model predicts plausible fragment spectra, MC dropout gives meaningful uncertainty estimates (Intensity $\approx$0.0313, b-ion $\approx$0.0300, y-ion $\approx$0.0326). In contrast, ensemble uncertainties are near machine precision ($\sim 10^{-10}$), reflecting insufficient diversity.

Summary Table:

Method	Metric	RT Value	MS2 (Intensity)
MC Dropout	PICP	0.0213	–
MC Dropout	MPIW	0.1905	0.0313
Ensemble	PICP	0.0213	–
Ensemble	MPIW	0.2464	~ 0

Visual Summaries

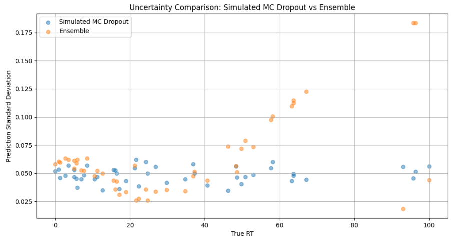

Figure 1: Overall Uncertainty Comparison between Methods.

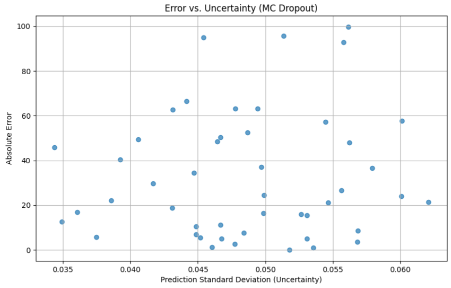

Figure 2: Error vs. Uncertainty (MC Dropout).

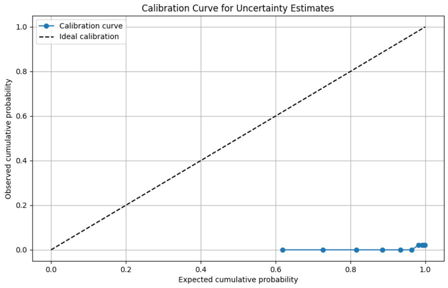

Figure 3: Ensemble Calibration Plot.

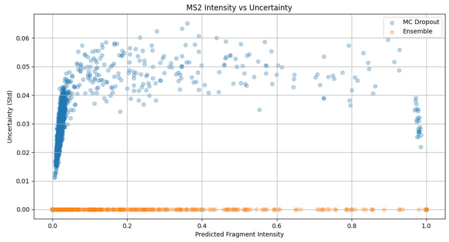

Figure 4: MS2 Intensity vs. Uncertainty.

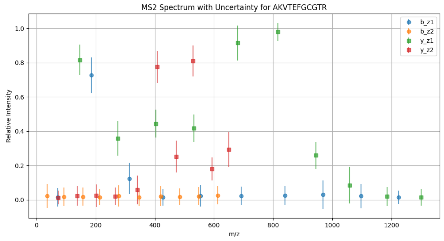

Figure 5: MC Dropout MS2 Uncertainty for Peptide AKVTEFGCGTR.

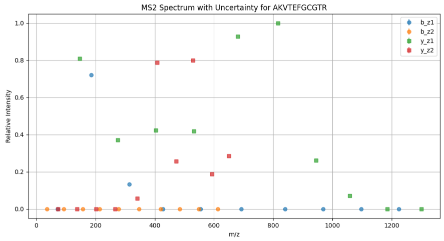

Figure 6: Ensemble MS2 Uncertainty for Peptide AKVTEFGCGTR.

Conclusions and Next Steps

- **RT Model:** The substantial underestimation in uncertainty (low PICP, high absolute error) signals that additional data, retraining, and calibration adjustments are required.
- **MS2 Predictions:** The MC dropout method yields interpretable uncertainties correlated with fragment intensity predictions. The ensemble method, however, needs revision to improve its diversity.
- **General Integration:** Refining uncertainty estimates and incorporating them into downstream pipelines (e.g., confidence-weighted peptide-spectrum matches) is a priority for improving overall proteomic reliability.

Conclusion

This work enhances proteomics by integrating robust uncertainty estimates and enriched PTM embeddings for more accurate, interpretable peptide identification. Our modular framework promises broad impact in research and clinical settings.