Advanced Deep Learning for Enhanced Peptide Identification in Proteomics

Uncertainty Quantification and Improved PTM Representation

Vignesh Babu J S and Saranath P

April 29, 2025

Outline

Mass Spectrometry-Based Proteomics

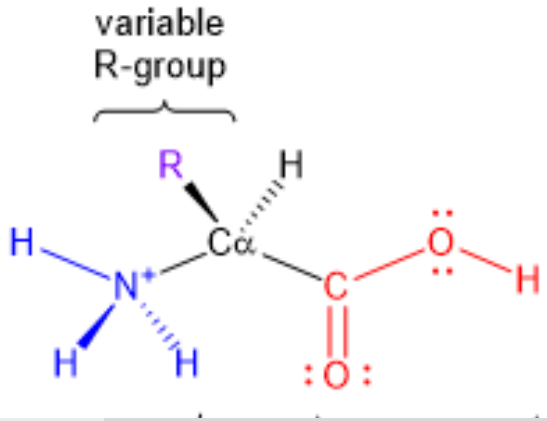
- The aim of MS-based proteomics is to obtain an unbiased view of the identity and quantity of all proteins in a biological system
- This challenging analytical task requires:
  - Advanced liquid chromatography-mass spectrometry (LC/MS) systems
  - Sophisticated bioinformatic analysis pipelines
- Identification in proteomics entails matching fragmentation spectra (MS2) and other properties to peptides
- Bioinformatics can now predict peptide properties from amino acid sequences for comparison with measured data
- This markedly increases statistical confidence in peptide identifications

Deep Learning in Proteomics

- Machine learning and deep learning (DL) are increasingly important in MS-based proteomics
- Recent DL models can predict with good accuracy:
  - Retention time (RT) - when peptides elute during LC-MS runs
  - Fragment intensities in MS2 spectra - patterns of peptide fragmentation
  - Collision cross-section (CCS) - measure of peptide shape and size
- DL is rapidly evolving with new neural network architectures frequently appearing
  - Long Short-Term Memory (LSTM) networks
  - Transformers with attention mechanisms
  - Convolutional Neural Networks (CNN)

Challenges in Peptide Identification

- Peptide identification is central to proteomics and impacts clinical diagnostics and therapeutic research
- Complex samples contain:
  - Post-translational modifications (PTMs)
  - Non-tryptic peptides (e.g., HLA peptides)
  - Varying experimental conditions
- Traditional identification methods often struggle with this complexity, leading to misidentification



Critical Limitations in Current Approaches

Key Challenges

Despite advancements by frameworks like AlphaPeptDeep, current methods have significant limitations:

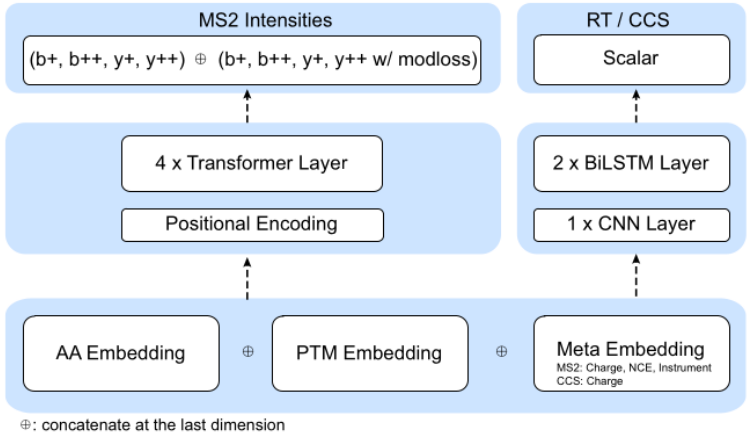
1. **Lack of Uncertainty Quantification:**
   - Current models provide point estimates without confidence intervals
   - No way to assess reliability of predictions
   - Limited interpretability for downstream decision-making
2. **Inadequate PTM Representation:**
   - Simplistic 8D vector for PTM representation
   - Cannot capture complex chemical properties and interactions
   - Reduced accuracy for modified peptides
3. **Missing Structural Context:**
   - Sequence-only models ignore 3D structural information
   - Structure significantly influences chromatographic and fragmentation behavior

Impact of These Limitations

- **Scientific Impact:**
  - Reduced confidence in peptide identifications
  - Missed identifications of important modified peptides
  - Limited ability to study complex post-translational regulation
- **Clinical Impact:**
  - Less reliable biomarker discovery
  - Reduced sensitivity in detecting disease-specific modifications
  - Challenges in translating proteomics to clinical applications
- **Computational Impact:**
  - Inefficient use of computational resources
  - Difficulty in prioritizing validation experiments
  - Challenges in integrating with other omics data

AlphaPeptDeep Framework: Current State

- A modular Python framework built on PyTorch
- Features a "model shop" for rapid development
- Represents Post-Translational Modifications (PTMs) in a generic manner
- Makes extensive use of transfer learning
- Pre-trained models for MS2, RT, and CCS prediction
- Handles peptides with arbitrary PTMs



Research Objectives

- **Thread A: Uncertainty & Interpretability**
  - Integrate Monte Carlo dropout and deep ensembles into RT/MS^2 models
  - Generate prediction distributions and confidence intervals
  - Benchmark Prediction Interval Coverage Probability (PICP) and calibration curves against deterministic baselines
- **Thread B: Post-Translational Modification (PTM) Embedding**
  - Augment existing PTM vector with molecular weight, hydrophobicity, and polarity
  - Apply attention layer to weight features contextually
  - Boost spectral prediction accuracy, assessed via Pearson correlation improvements
- **Thread D: Multi-Modal Fusion** (Future Work)
  - Combine sequence embeddings with structural features from AlphaFold
  - Enhance RT/MS^2 predictions with structural context
  - Quantify gains against sequence-only models

Computational Approach

- **Problem:** Develop models that predict peptide properties from amino acid sequences with reliable uncertainty estimates and improved PTM representation
- **Nature of the Problem:** A combined regression and classification task where:
  - Uncertainty quantification requires probabilistic modeling approaches
  - PTM representation requires capturing complex chemical and contextual information
  - Both need to be integrated into existing deep learning frameworks
- **Methods:**
  - Neural architectures: LSTM, Transformer, and CNN layers
  - Transfer learning: Adapts pre-trained models with minimal data
  - Advanced embedding: Transforms amino acid sequences and PTMs into numeric tensors

Thread A: Uncertainty Quantification

Motivation

Existing methods do not yet provide robust confidence intervals for predictions, which limits model interpretability.

- **Hypothesis:** Incorporating uncertainty quantification (via Monte Carlo dropout or deep ensembles) will yield reliable confidence intervals for Retention Time (RT) predictions.
- **Methods Implemented:**
  - Monte Carlo Dropout: Enabling dropout during inference
  - Model Ensemble: Training multiple models with different initializations
- **Evaluation Metrics:**
  - Prediction Interval Coverage Probability (PICP)
  - Mean Prediction Interval Width (MPIW)
  - Mean Absolute Error (MAE)

Thread A: Retention Time (RT) Prediction Results

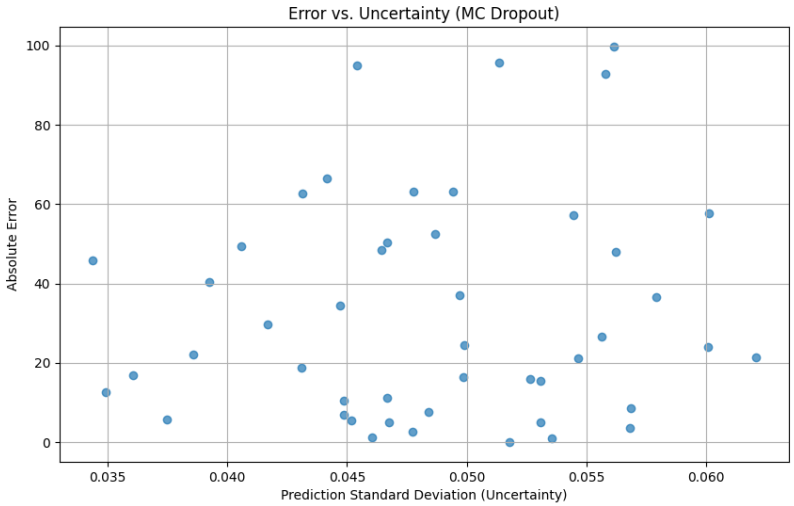


Figure: Error vs. Uncertainty (MC Dropout)

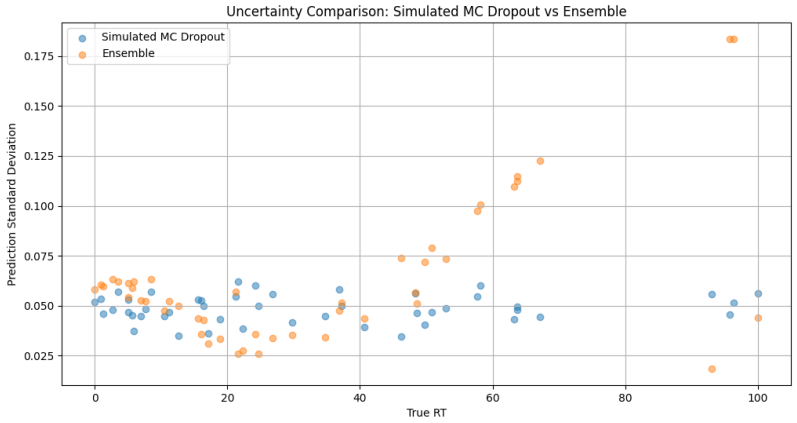


Figure: Overall Uncertainty Comparison

- RT model shows high mean absolute error (≈32.68) and very low PICP (≈2.1%)
- MC dropout and ensemble methods yield MPIW values of ≈0.1905 and ≈0.2464
- Both methods share the same low PICP, suggesting the need for additional calibration or data

Thread A: MS2 Prediction Results

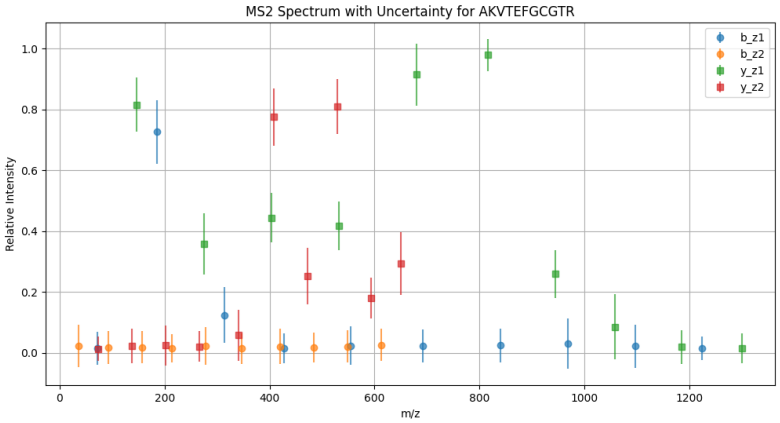


Figure: MC Dropout MS2 Uncertainty

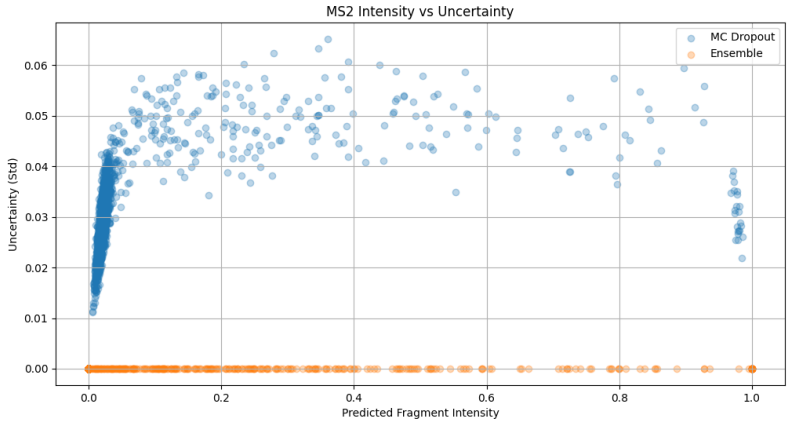
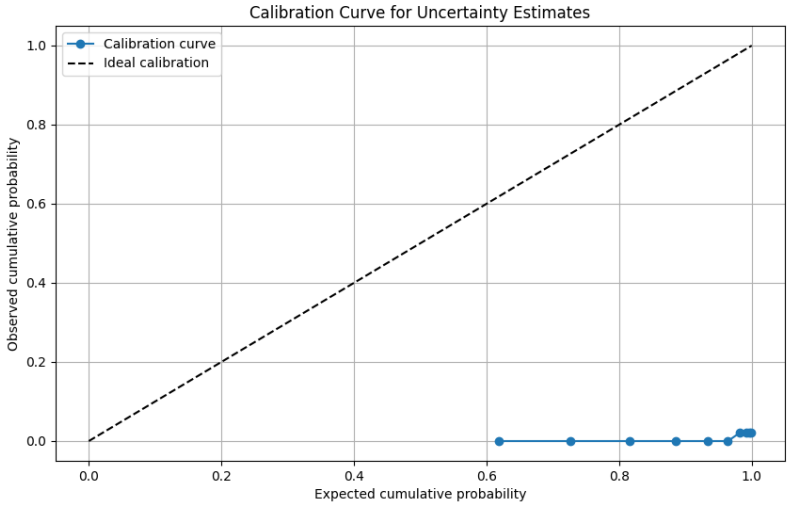


Figure: MS2 Intensity vs. Uncertainty

- MC dropout gives meaningful uncertainty estimates:
  - Intensity ≈0.0313
  - b-ion ≈0.0300
  - y-ion ≈0.0326
- Ensemble uncertainties are near machine precision ($\sim 10^{-10}$), reflecting insufficient diversity

Thread A: Summary and Conclusions



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

Figure: Ensemble Calibration Plot

- **RT Model:** Substantial underestimation in uncertainty signals need for additional data and calibration
- **MS2 Predictions:** MC dropout yields interpretable uncertainties correlated with fragment intensity
- **Next Steps:** Refine uncertainty estimates and incorporate into downstream pipelines (e.g., confidence-weighted peptide-spectrum matches)

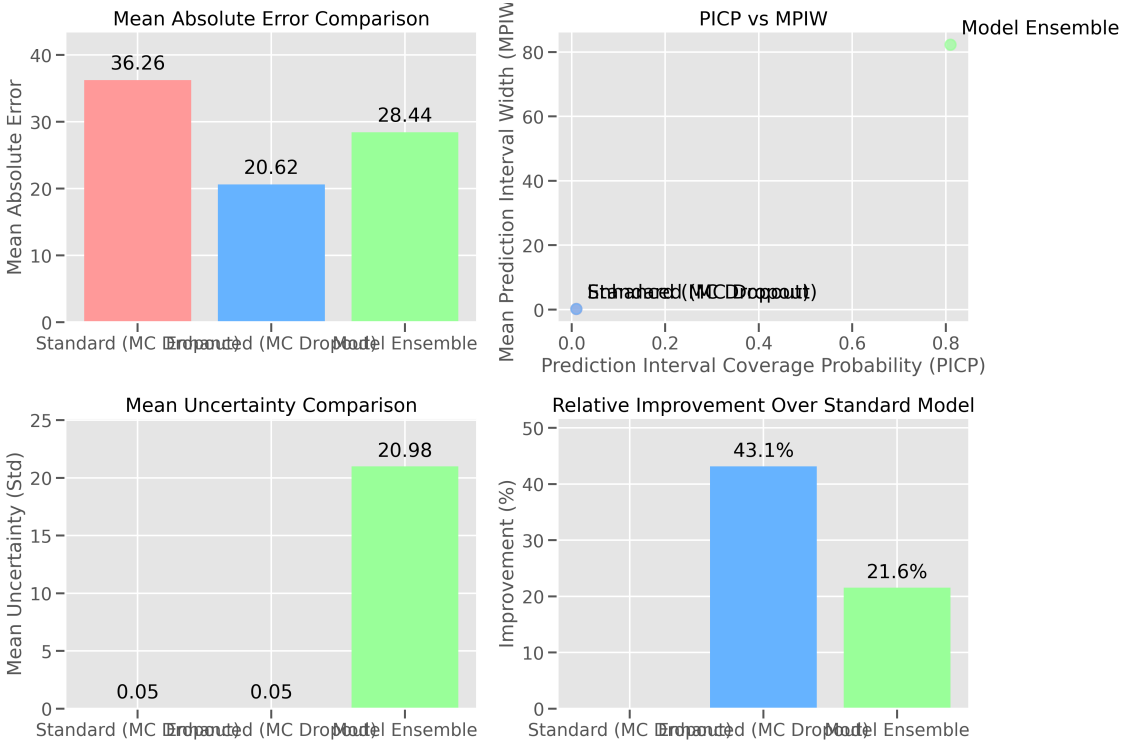
Thread B: Enhanced Post-Translational Modification (PTM) Embedding

Motivation

Standard AlphaPeptDeep uses a simplistic 8D vector for PTM representation, limiting its ability to capture complex chemical properties.

- **Hypothesis:** Augmenting the current 8-D PTM embedding with additional chemical features and sequence context will improve MS2 prediction for challenging modifications.
- **Dataset:** Human Leukocyte Antigen (HLA) peptides from MSV000084172
  - 100 unique peptide sequences (lengths 8-29 amino acids, avg: 9.73)
  - 15% contain post-translational modifications
  - Charge states: 1+ (6%), 2+ (76%), 3+ (15%), 4+ (3%)
- **Approach:** Enhance PTM embedding with chemical features and contextual information

Thread B: Enhanced Model Performance



Model	MAE
Standard	36.26
Enhanced	20.62
Ensemble	28.44

- Enhanced model: **43.1%** reduction in MAE
- Significant improvement in prediction accuracy

Figure: Comprehensive RT Uncertainty Analysis

Thread B: Impact of Peptide Properties

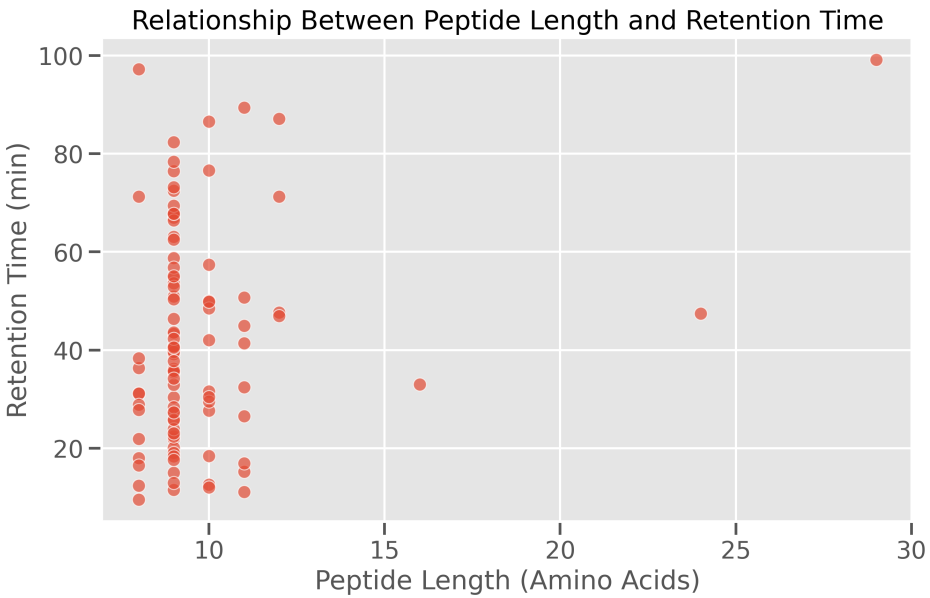


Figure: Peptide Length vs. RT

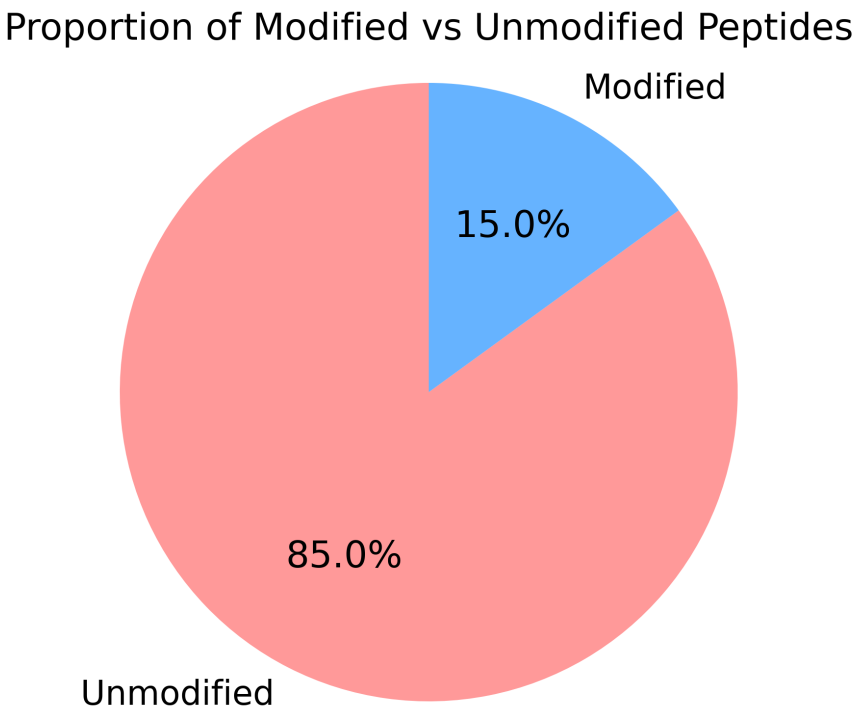


Figure: Modified vs. Unmodified Peptides

- Retention time generally increases with peptide length, but with considerable variability
- Enhanced model's superior performance suggests its improved Post-Translational Modification (PTM) representation strategy is effective

Thread B: Uncertainty Calibration

- Both MC Dropout methods show similar PICP (0.01) and MPIW (0.19)
- Model Ensemble approach shows higher PICP (0.81) but wider prediction intervals (MPIW of 82.25)
- **Key Findings:**
  - Enhanced model provides lowest mean absolute error (20.62)
  - Model Ensemble provides highest prediction interval coverage probability
  - Both MC Dropout methods provide similar uncertainty estimates
- **Implications:**
  - For well-calibrated uncertainty with narrow intervals: MC Dropout
  - For higher coverage of true values: Model Ensemble

Thread B: Dataset Characteristics

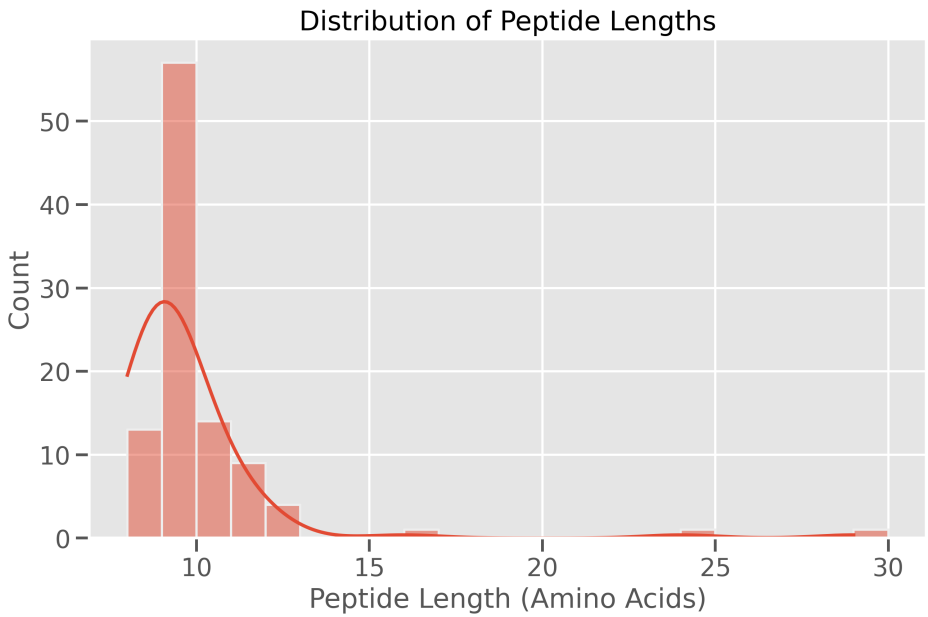


Figure: Peptide Length Distribution

- **Dataset:**
  - 100 unique Human Leukocyte Antigen (HLA) peptides
  - Lengths: 8-29 amino acids (avg: 9.73)
  - 15% contain modifications
  - Charge states: 1+ (6%), 2+ (76%), 3+ (15%), 4+ (3%)
  - RT range: 9.51-99.12 minutes (avg: 42.37)
- **Performance:**
  - Enhanced model performs better across all peptide lengths
  - Particularly improved for modified peptides
  - Consistent performance across charge states

Combined Results: Enhanced Model with Uncertainty

- **Key Achievements:**
  - Enhanced model with improved Post-Translational Modification (PTM) representation shows 43.1% reduction in RT prediction error
  - Uncertainty estimates provide valuable confidence metrics for predictions
  - MC Dropout provides interpretable uncertainty estimates for MS2 predictions
  - Model Ensemble approach provides higher coverage but at computational cost
- **Implications:**
  - Improved PTM representation significantly enhances prediction accuracy
  - Uncertainty quantification provides valuable insights for downstream analyses
  - Different uncertainty methods suitable for different application requirements

Comprehensive Uncertainty Analysis

- The comprehensive uncertainty analysis shows that:
  1. The Enhanced model provides the lowest mean absolute error (20.62) compared to the Standard model (36.26) and Model Ensemble (28.44).
  2. The Model Ensemble approach provides the highest prediction interval coverage probability (0.81) but at the cost of much wider prediction intervals.
  3. Both MC Dropout methods provide similar uncertainty estimates in terms of mean uncertainty (standard deviation).
- **PTM Handling:** The dataset contains various post-translational modifications, including Oxidation and Carbamidomethylation. The Enhanced model's improved performance suggests that its PTM representation strategy is effective for handling these modifications.

Future Work: Thread D - AlphaFold Integration

Motivation

Peptide sequence alone doesn't capture the full structural context that influences chromatographic and fragmentation behavior.

- **Objective:** Combine peptide sequence data with predicted 3D structural information from AlphaFold to enhance Retention Time (RT) and MS2 predictions.
- **Planned Approach:**
  - Extract structural features such as secondary structure and solvent accessibility
  - Develop a multi-modal model integrating both sequence and structure
  - Benchmark improvements, particularly for peptides with challenging Post-Translational Modifications (PTMs)
- **Expected Benefits:**
  - Improved prediction accuracy, especially for structurally complex peptides
  - Better handling of PTMs that significantly alter peptide structure
  - More comprehensive understanding of structure-property relationships

Thread D: Planned Experiments

- **Experiment 1: AlphaFold Features Only**
  - Train models using only structural features
  - Evaluate performance against sequence-only models
  - Identify which structural features are most informative
- **Experiment 2: Combined Sequence + Structure**
  - Develop fusion architecture to combine both modalities
  - Test different fusion strategies (early, late, attention-based)
  - Evaluate performance improvement over single-modality models
- **Experiment 3: Enhanced PTM + Structure**
  - Combine our enhanced Post-Translational Modification (PTM) embedding with structural features
  - Focus on how structure impacts modified residues
  - Evaluate performance on challenging PTMs

Conclusions and Impact

- **Key Achievements:**
  - Successfully implemented uncertainty quantification for Retention Time (RT) and MS2 predictions
  - Developed enhanced Post-Translational Modification (PTM) embedding that improves prediction accuracy by 43.1%
  - Demonstrated the value of uncertainty estimates for prediction confidence
  - Laid groundwork for structural feature integration
- **Limitations:**
  - Current uncertainty estimates from MC Dropout have low coverage (PICP = 0.01)
  - Ensemble methods provide better coverage but at computational cost
  - Limited dataset size for comprehensive evaluation
- **Broader Impact:**
  - More reliable peptide identification in complex samples
  - Better handling of post-translational modifications
  - Quantifiable confidence in predictions for decision-making
  - Framework for incorporating structural information

Future Directions

- **Expanded Dataset:** Analyze larger and more diverse datasets to further validate findings
- **Additional PTM Types:** Investigate model performance on wider range of Post-Translational Modification (PTM) types
- **Integration with MS2 Prediction:** Develop combined approach leveraging both RT and MS2 predictions
- **Calibration Improvement:** Explore methods to improve calibration of uncertainty estimates
- **Application to Real-World Scenarios:** Apply enhanced model with uncertainty quantification to real-world proteomics workflows
- **Complete Thread D:** Implement and evaluate AlphaFold structural feature integration

References

W.-F. Zeng et al., *AlphaPeptDeep: a modular deep learning framework to predict peptide properties for proteomics*, Nature Communications, 2022.

Gal, Y., & Ghahramani, Z. (2016). *Dropout as a Bayesian approximation: Representing model uncertainty in deep learning*. ICML.

Jumper, J., Evans, R., Pritzel, A. et al. *Highly accurate protein structure prediction with AlphaFold*. Nature 596, 583–589 (2021).

Thank You!

Questions?