# *De novo* sequencing of chimeric spectra using Casanovo

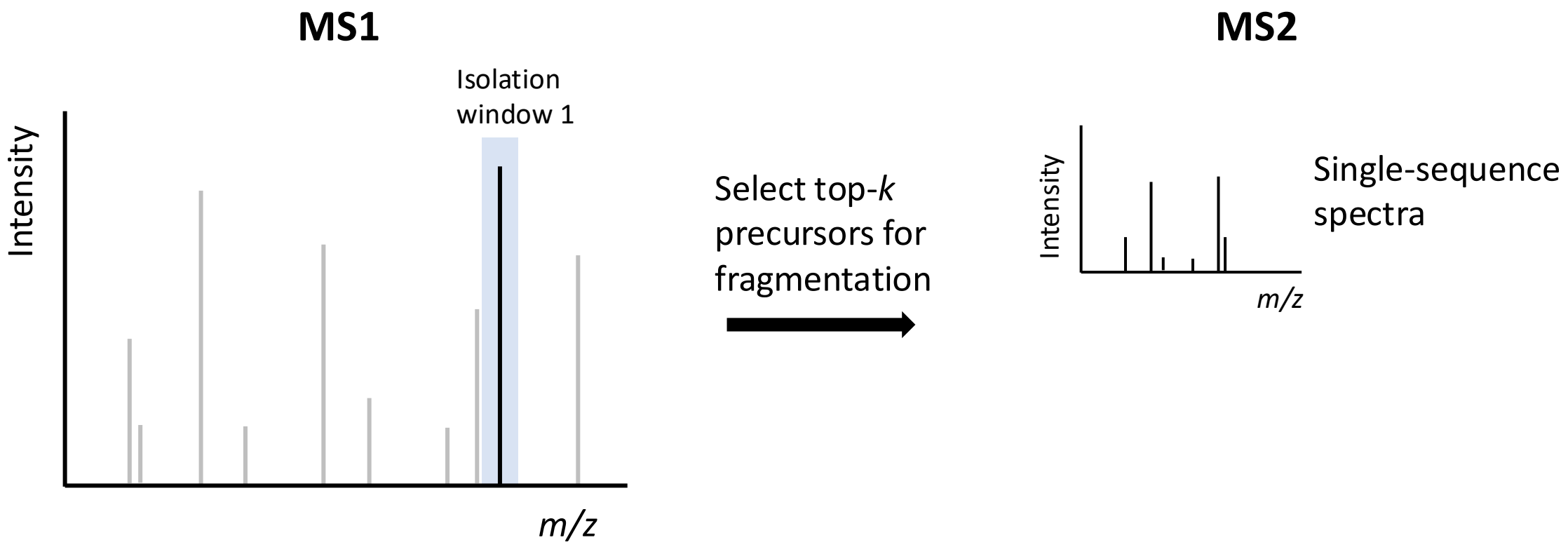
Yuhui Hong[1], Isha Gokhale[1], Justin Sanders[2], Melih Yilmaz[1], Gwenneth Straub[1], Wout Bittremieux[3], and William Stafford Noble[12]

The authors declare no competing financial interest.

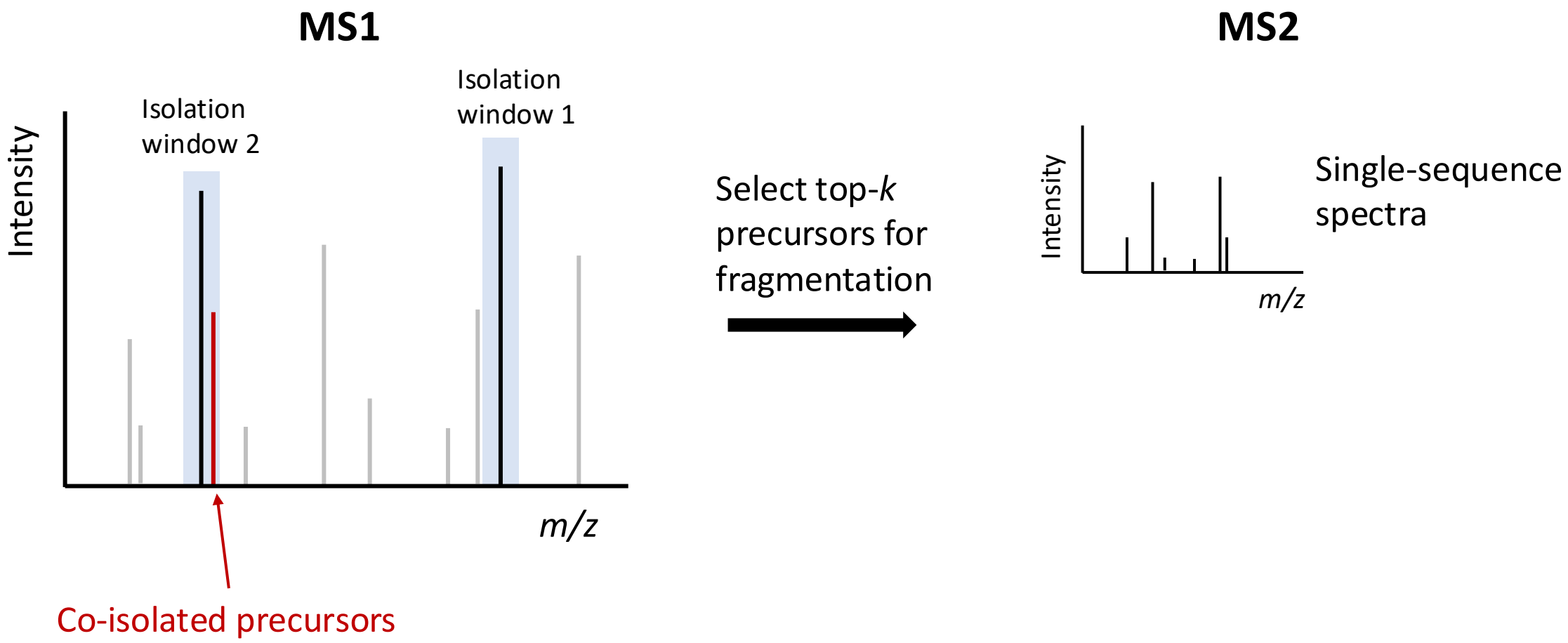
[1]*Department of Genome Sciences, University of Washington*
[2]*Paul G. Allen School of Computer Science and Engineering, University of Washington*
[3]*Department of Computer Science, University of Antwerp*

2026/05/20

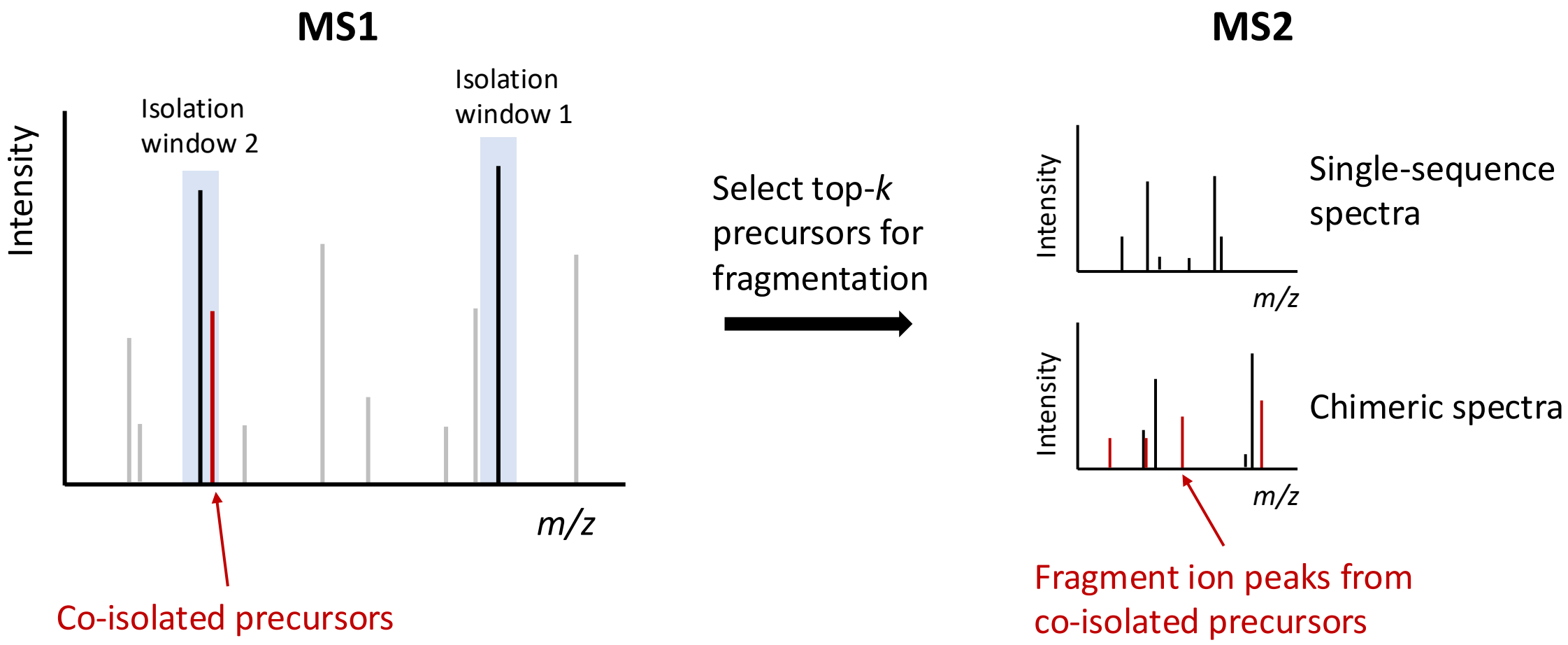
# DDA Selects and Isolates the Top-*k* Precursors for Fragmentation



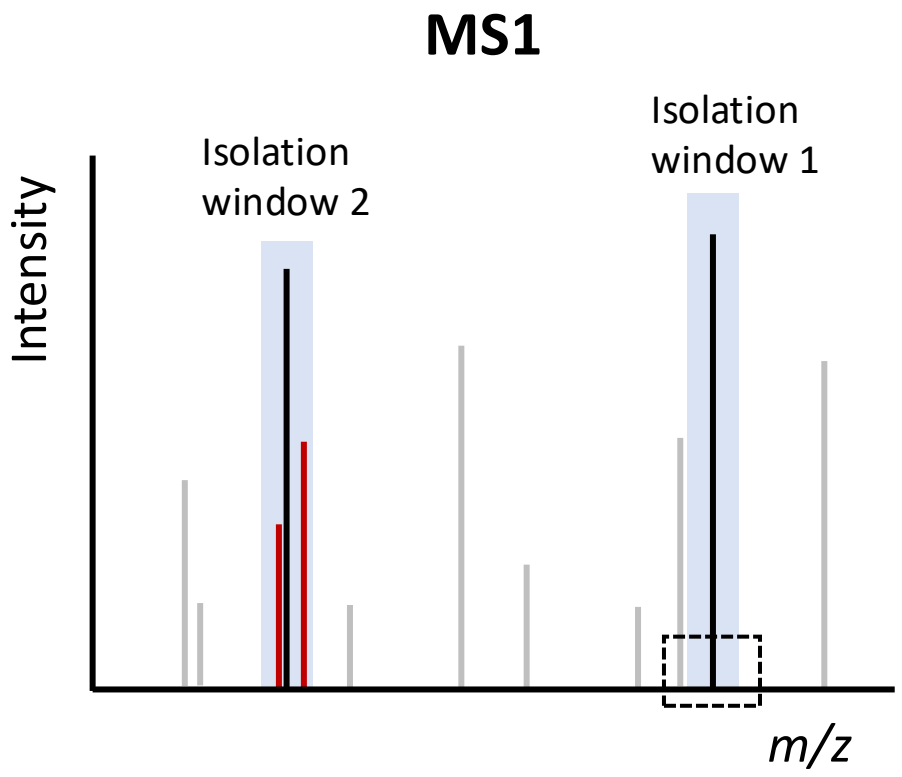
# Co-isolated Precursors in DDA Give Rise to Chimeric Spectra



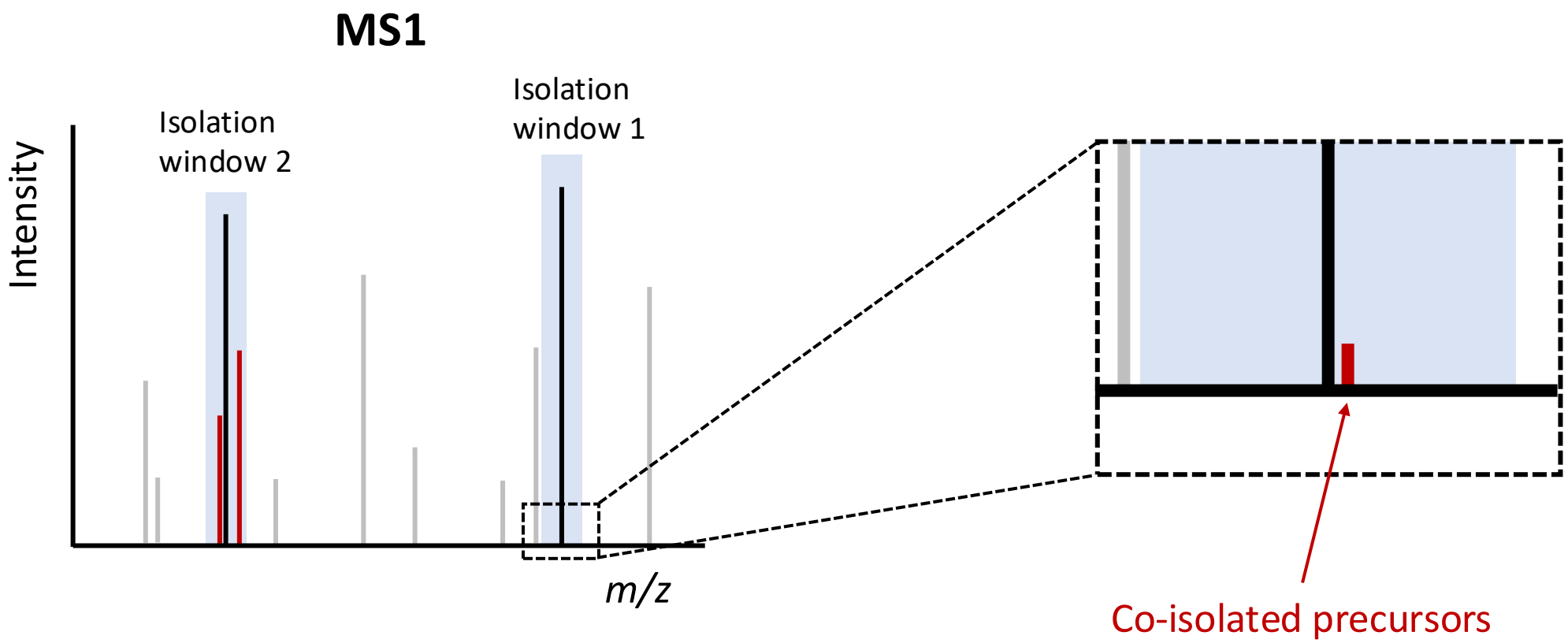
# Chimeric Spectra Contain Fragment Ions from Multiple Co-isolated Precursors



# Every Spectrum Is Potentially Chimeric



# Every Spectrum Is Potentially Chimeric — Sensitive Methods Can Reveal Hidden Co-isolated Peptides



# Existing *de novo* Methods on DDA Are Not Designed for Chimeric Spectra

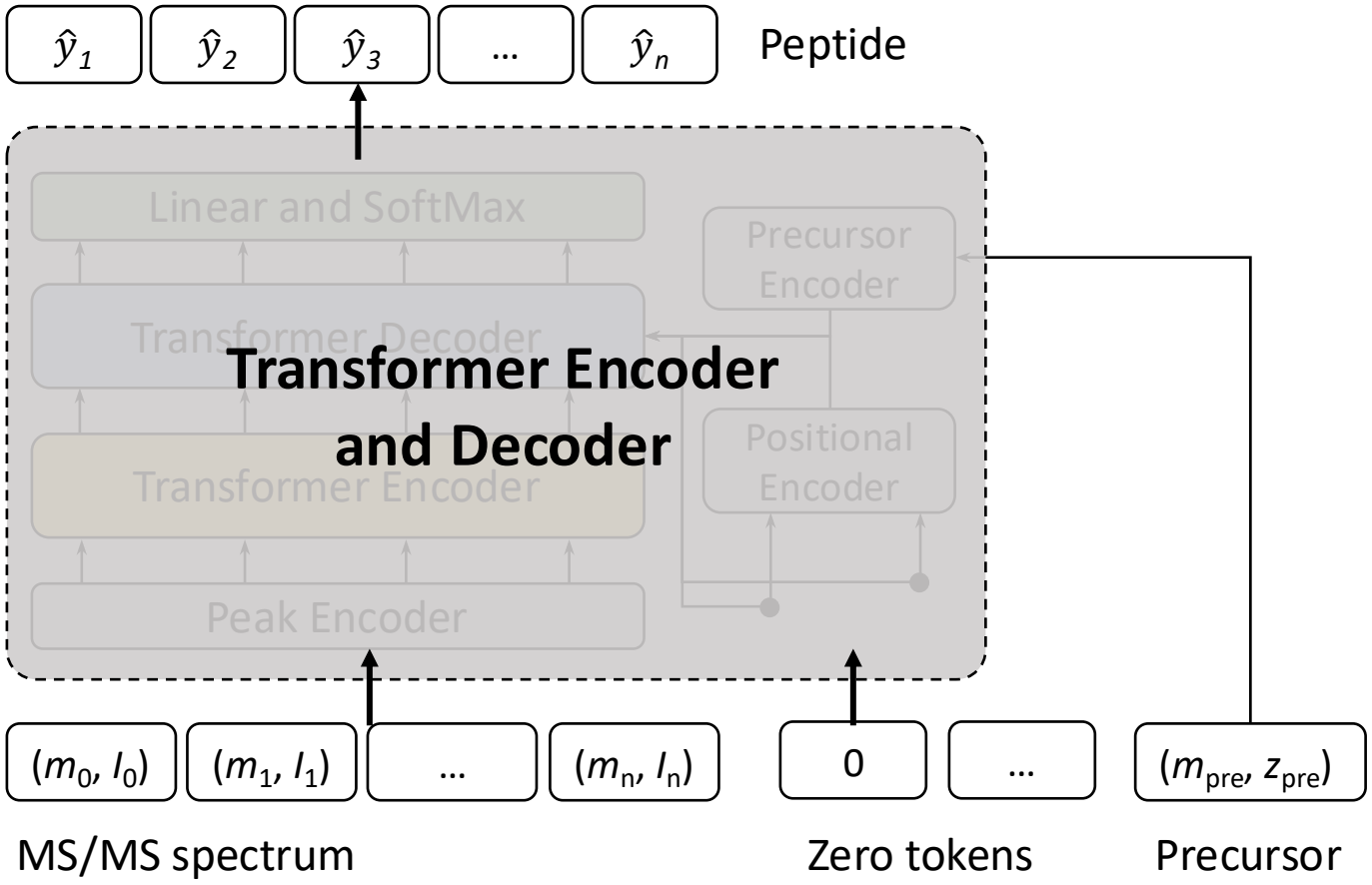
| Method | Year | Model architecture | | | Chimeric spectra | Citations |
|---|---|---|---|---|---|---|
| | | CNN | Transformer | Others | | |
| DeepNovo (Tran et al. 2017) | 2017 | ✓ | | | | 564 |
| SMSNet (Karunratanakul et al. 2019) | 2019 | ✓ | | | | 97 |
| PepNet (Liu et al. 2023) | 2023 | ✓ | | | | 86 |
| Casanovo (Yilmaz et al. 2022) | 2022 | | ✓ | | | 117 |
| GraphNovo (Mao et al. 2023) | 2023 | | ✓ | | | 59 |
| $\pi$-HelixNovo (Yang et al. 2024) | 2024 | | ✓ | | | 46 |
| PointNovo (Qiao et al. 2021) | 2021 | | | ✓ | | 128 |
| pNovo 3 (Yang et al. 2019) | 2019 | | | ✓ | | 112 |

.....

Original table is from [1], where the citations are updated by counting from Google Scholar on May 26, 2026.

[1] Bittremieux, Wout, et al. "Deep learning methods for de novo peptide sequencing." Mass Spectrometry Reviews 45.3 (2026): 507-526.

# Casanovo Is a State-of-the-Art Transformer-Based *de novo* Sequencer

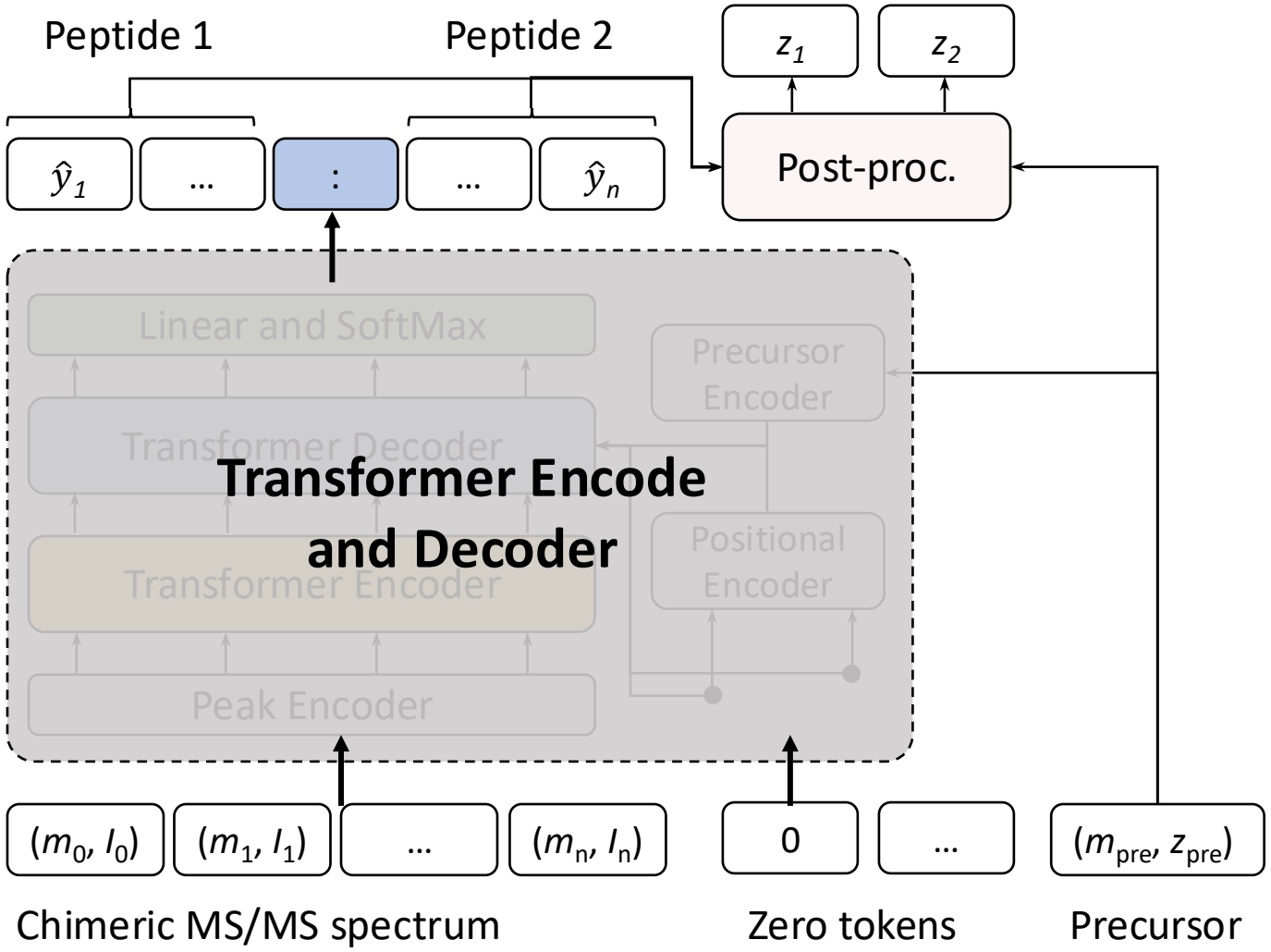



[2] Yilmaz, Melih, et al. "Sequence-to-sequence translation from mass spectra to peptides with a transformer model." Nature communications 15.1 (2024): 6427.

# Three Key Modifications Enable Casanovo to Directly Predict Chimera

**Casanovo-chim**

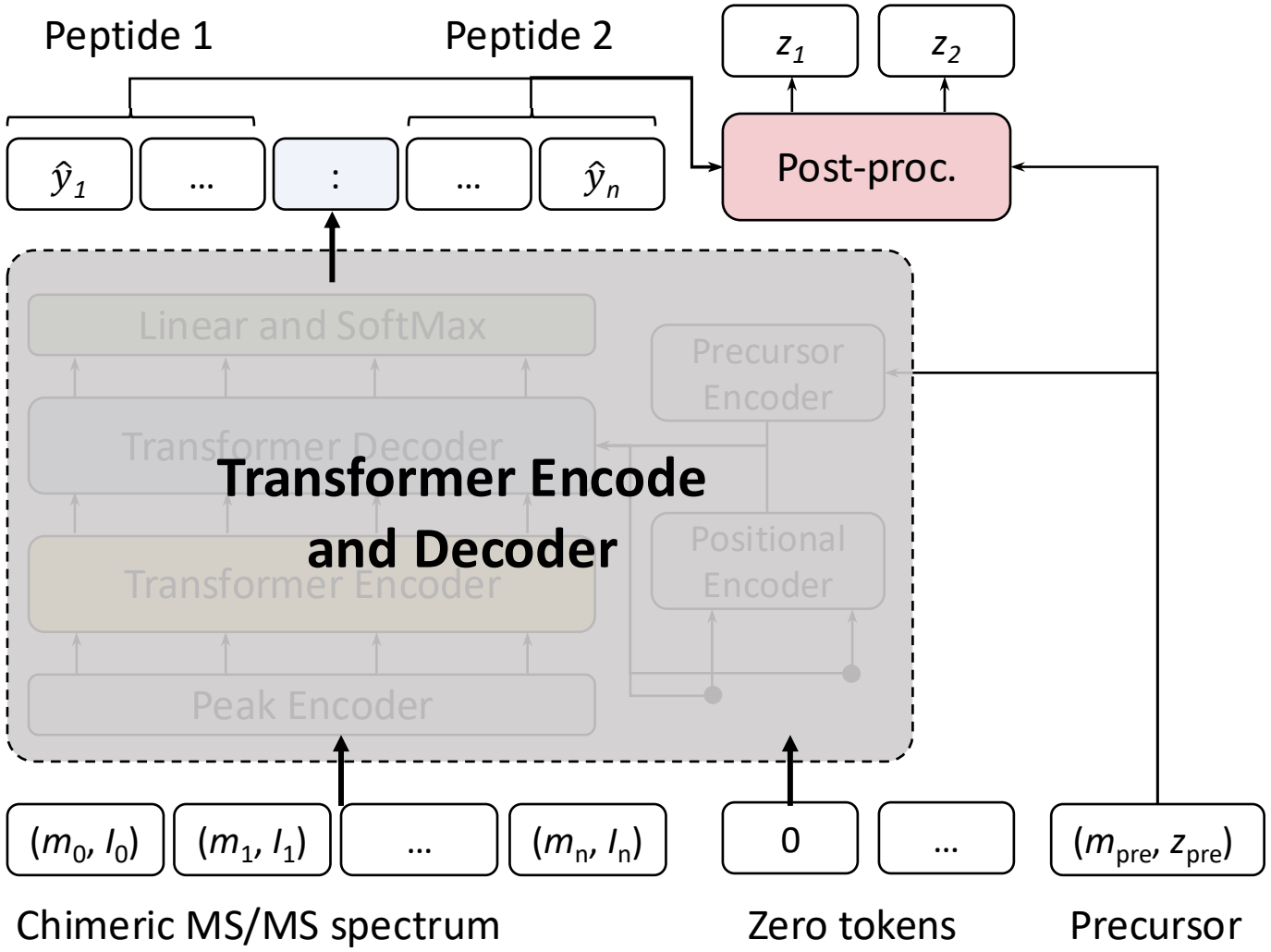
1. Special separator token ":" for dual-peptide output



# Three Key Modifications Enable Casanovo to Directly Predict Chimera

**Casanovo-chim**

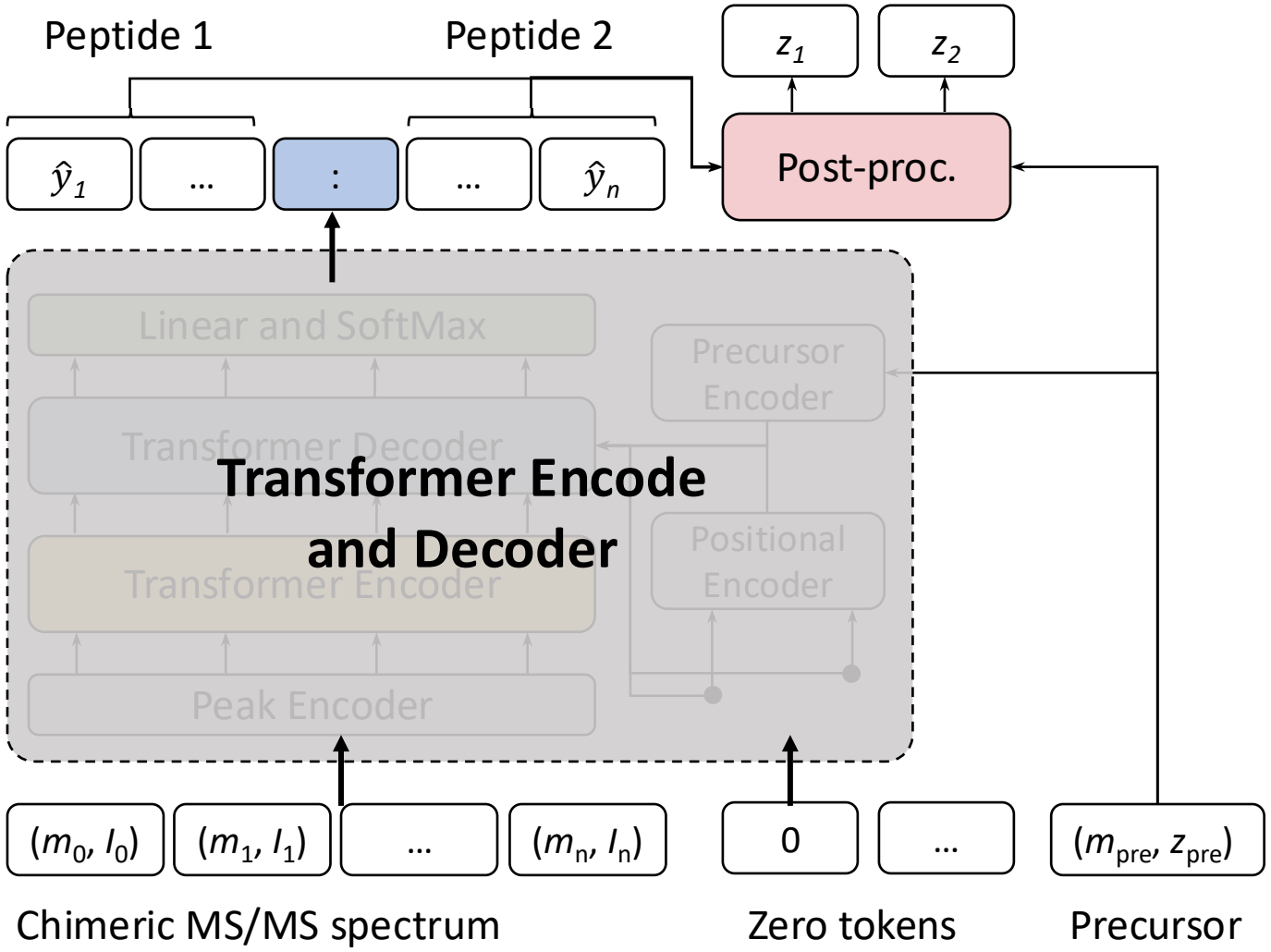
1. Special separator token ":" for dual-peptide output
2. Post-processing to correct chimeric charge states



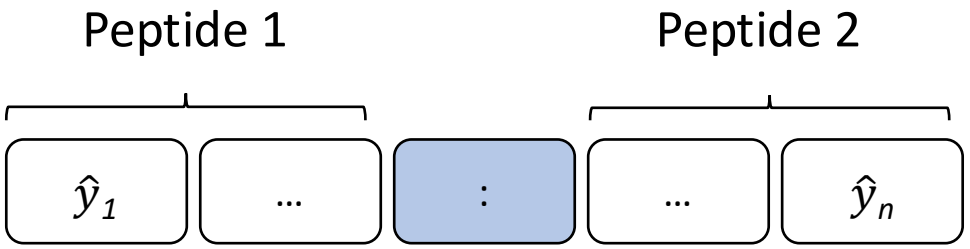
# Three Key Modifications Enable Casanovo to Directly Predict Chimera

**Casanovo-chim**

1. Special separator token ":" for dual-peptide output
2. Post-processing to correct chimeric charge states
3. Permutation-invariant loss function for chimeric output



# Permutation-Invariant Loss Function for Chimeric Output



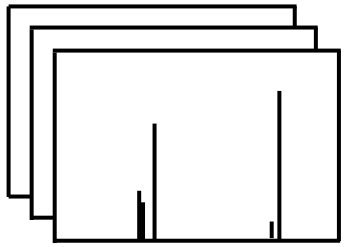

predicted peptide $\hat{y}$

$$\mathcal{L} = \min(\mathcal{H}(p_1 : p_2, \hat{y}), \mathcal{H}(p_2 : p_1, \hat{y}))$$

mean per-token
cross-entropy loss

# We Constructed a Ten-Species Chimeric and Single-Peptide Spectra Dataset Using MSFragger DDA+

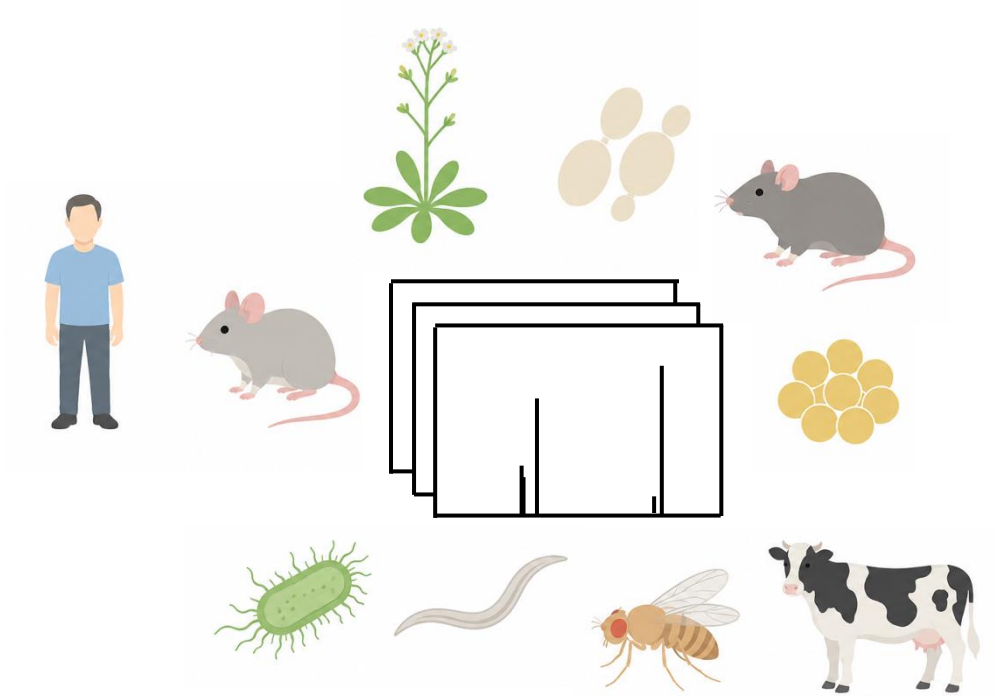
PRIDE
PRoteomics IDEntifications Database



Experimental spectra

FRAGPIPE

MSFragger DDA+



Spectra with peptide annotations across ten species

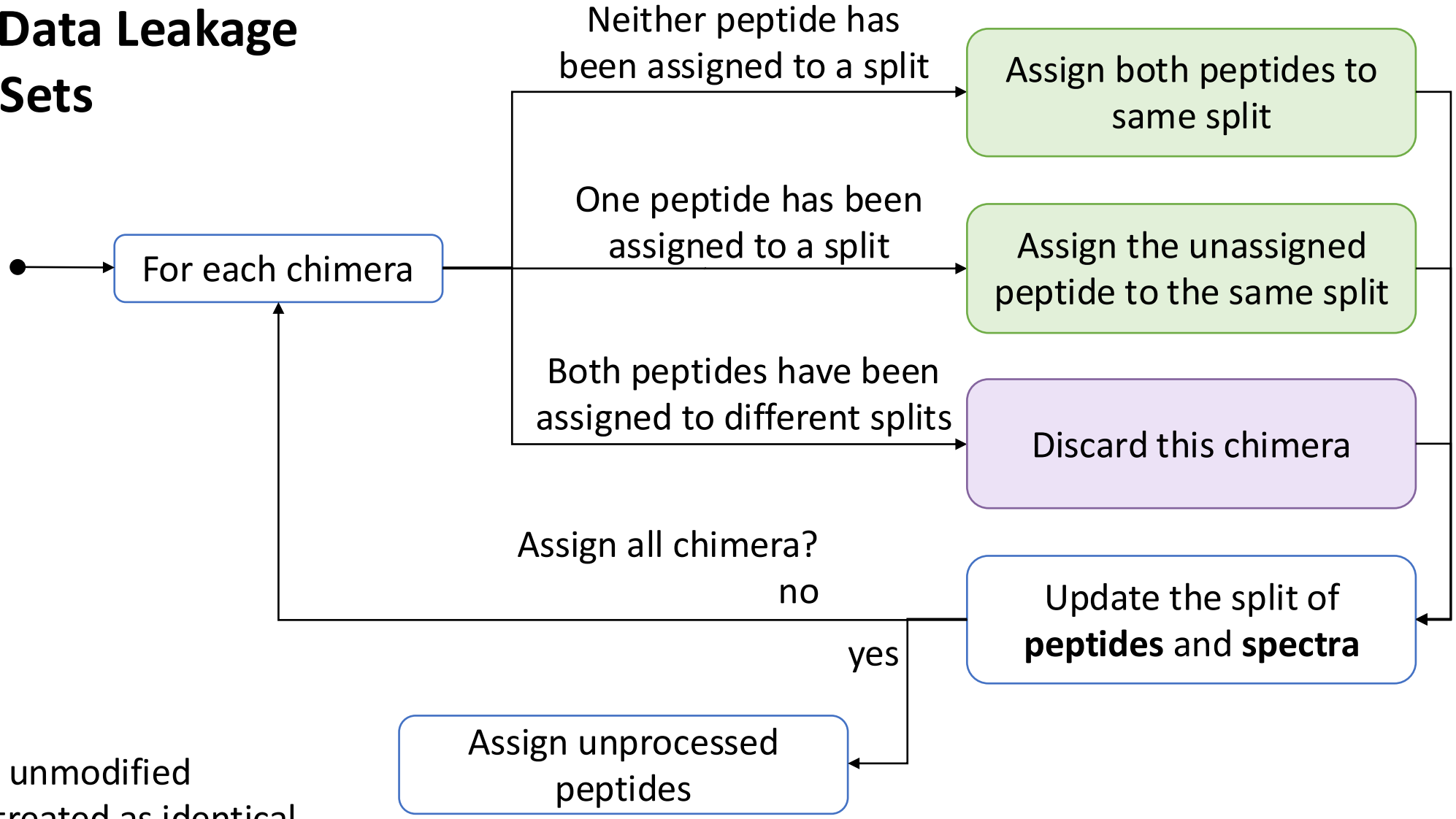
[3] Yu, Fengchao, Yamei Deng, and Alexey I. Nesvizhskii. "MSFragger-DDA+ enhances peptide identification sensitivity with full isolation window search." Nature Communications 16.1 (2025): 3329.

# Preprocessing Removes Unreliable Chimeras and Balances Peptide Representation

1. Spectra with 3+ co-isolated precursors excluded.
2. Top 2 spectra retained per unique peptide or chimeric pair (by q-value).
3. Chimeras downsampled to ≤10 per unique peptide for balanced representation.

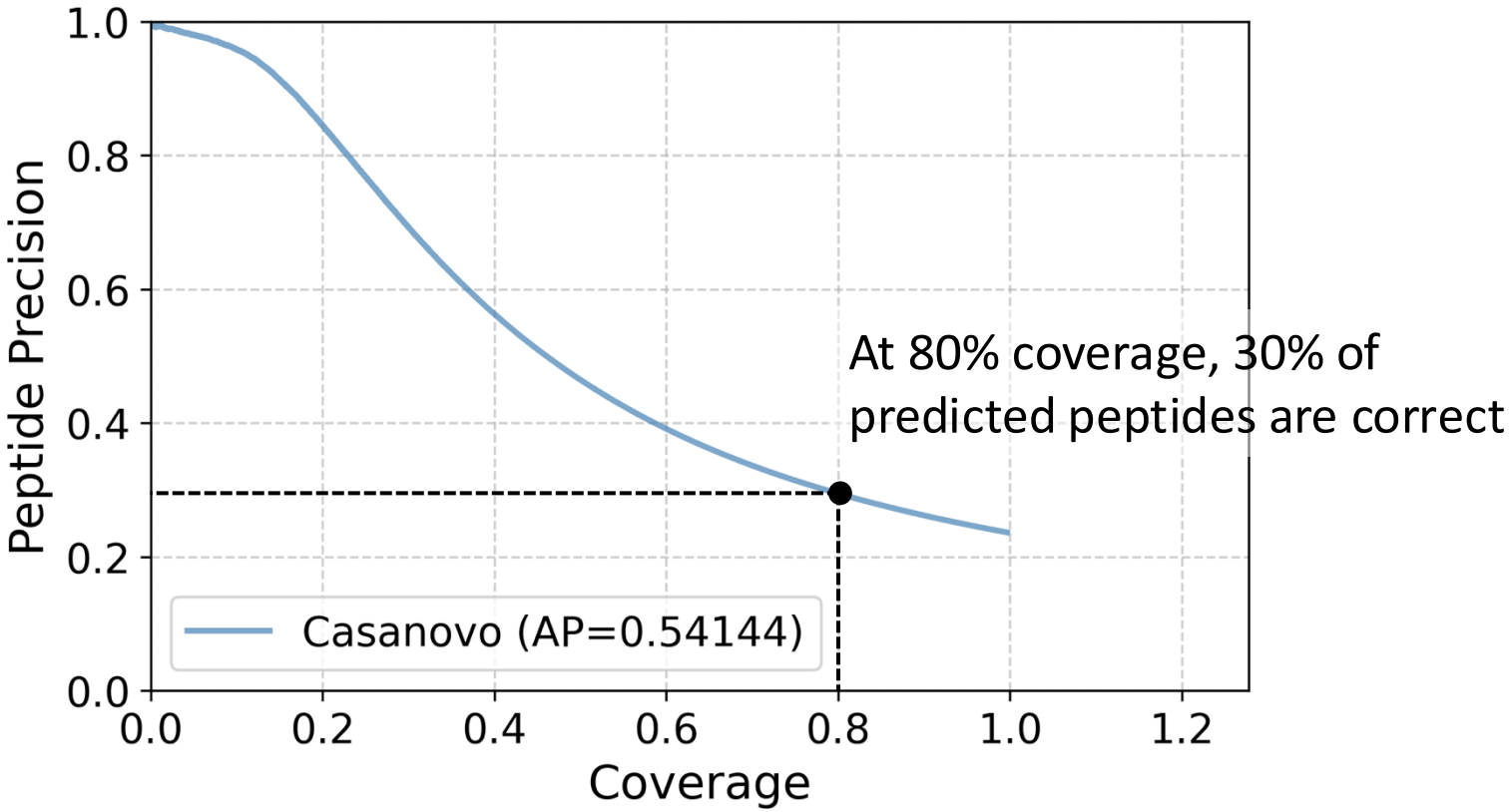
**~5M spectra from ~ 1M peptides — 53.3% chimeric**

# Peptide-Level Partitioning Prevents Data Leakage between Sets

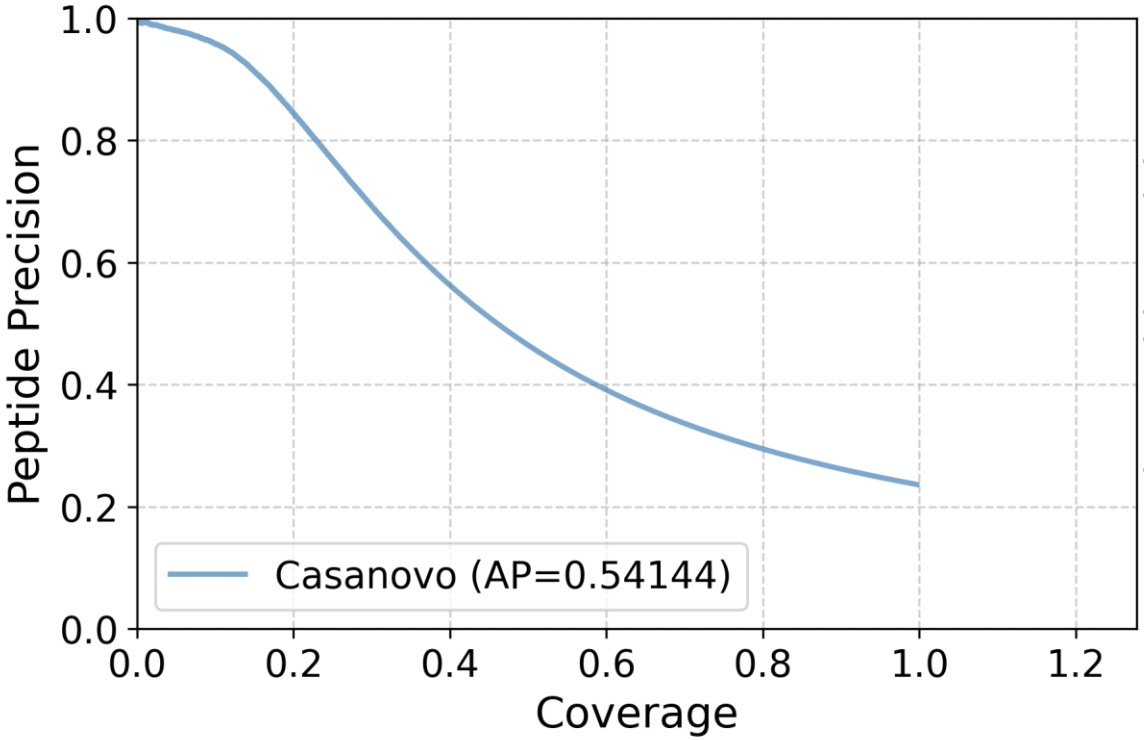


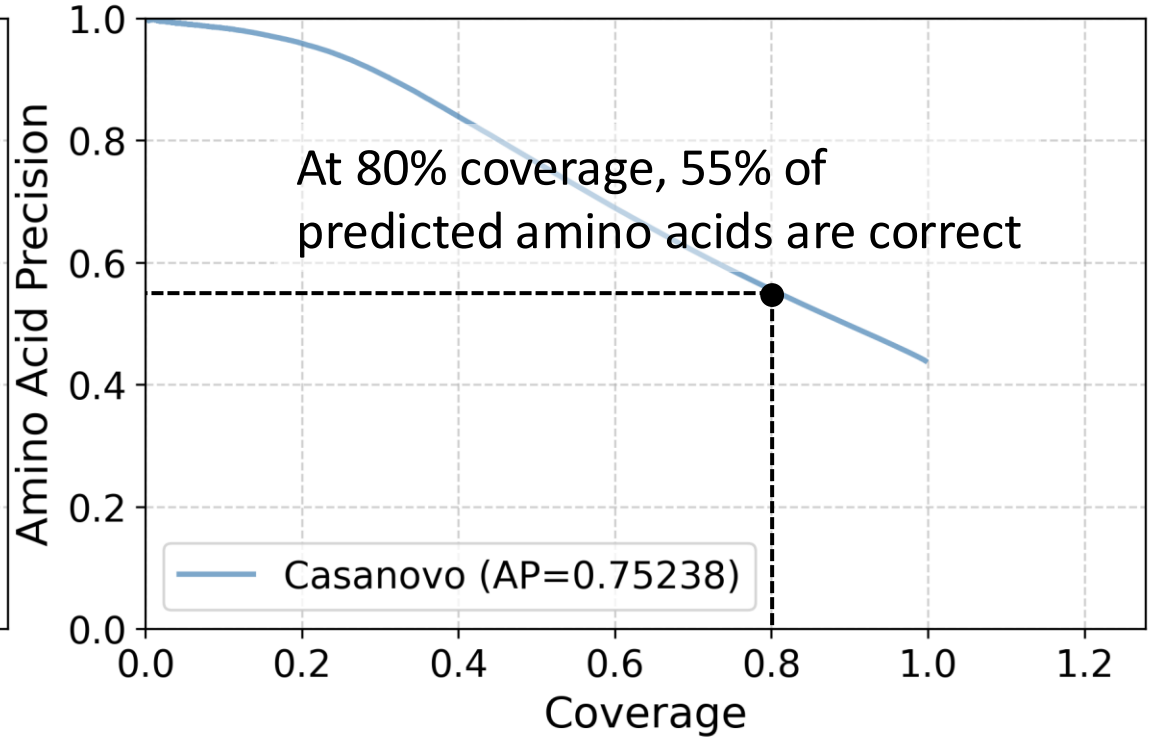

Modified and unmodified peptides are treated as identical.

# Precision-Coverage Curves Evaluate Both Accuracy and Completeness of Predictions



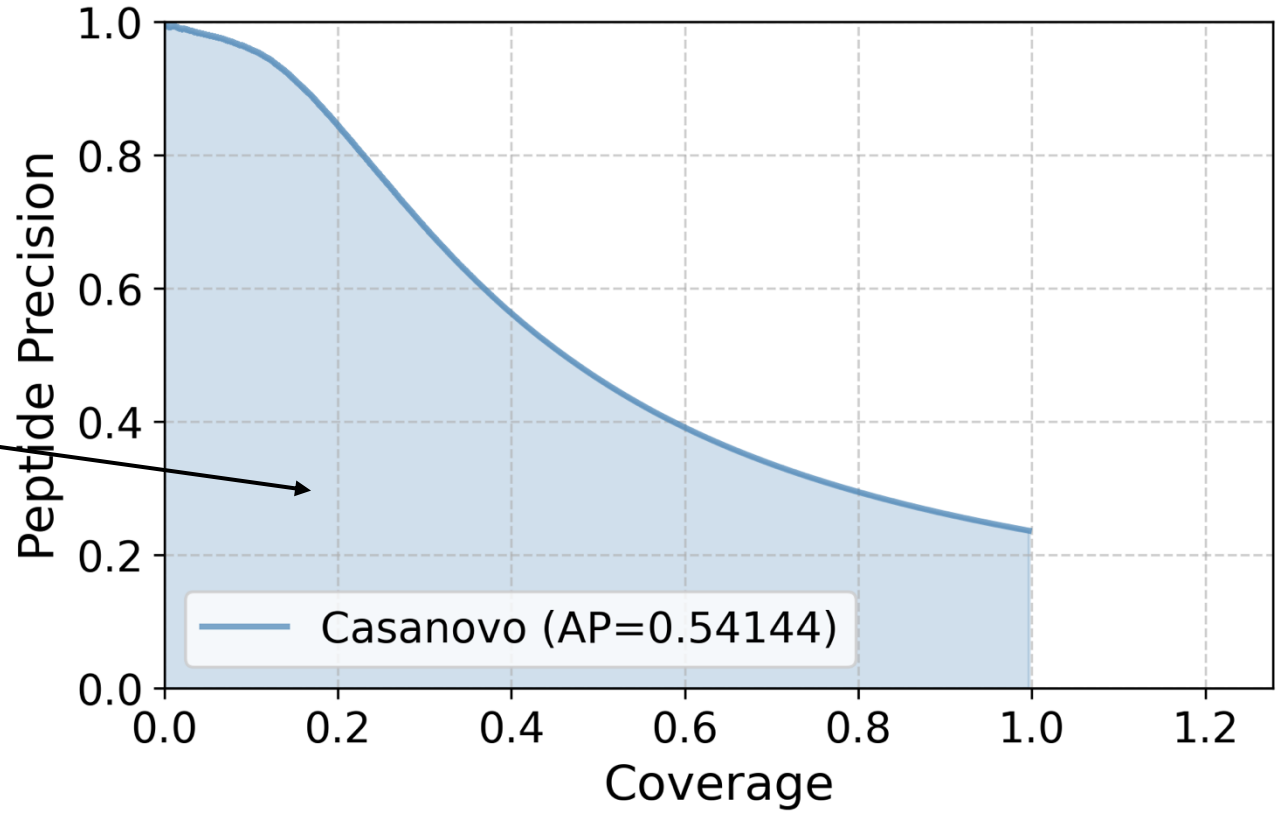
# Precision-Coverage Curve Can Also be Calculated at Amino Acid Level



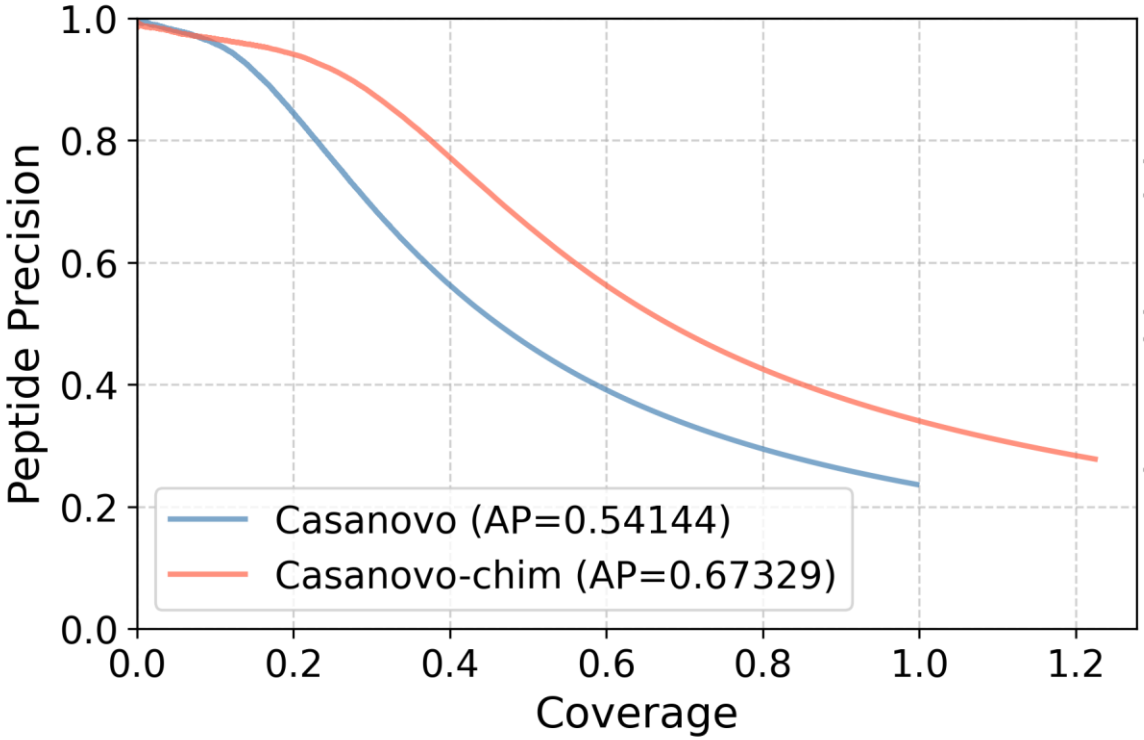



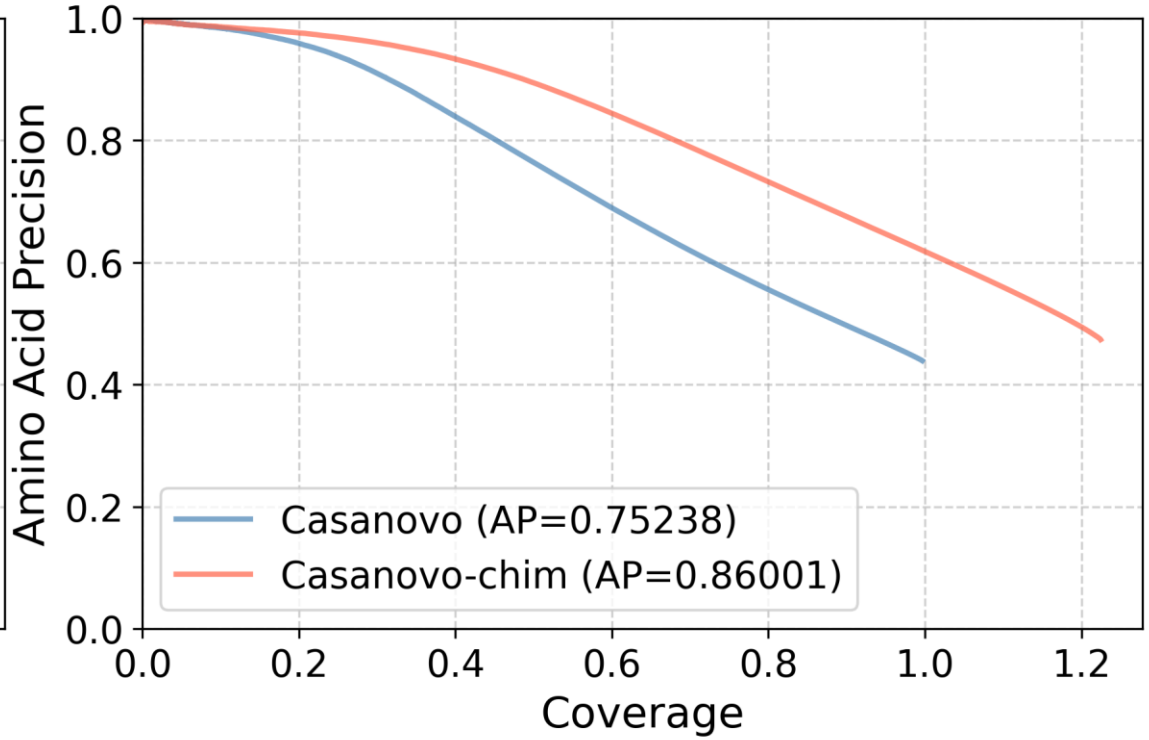
# Chimeric Predictions Can Exceed Full Coverage — Average precision (AP) Is Evaluated at Coverage ≤ 1

Area under the curve
= average precision

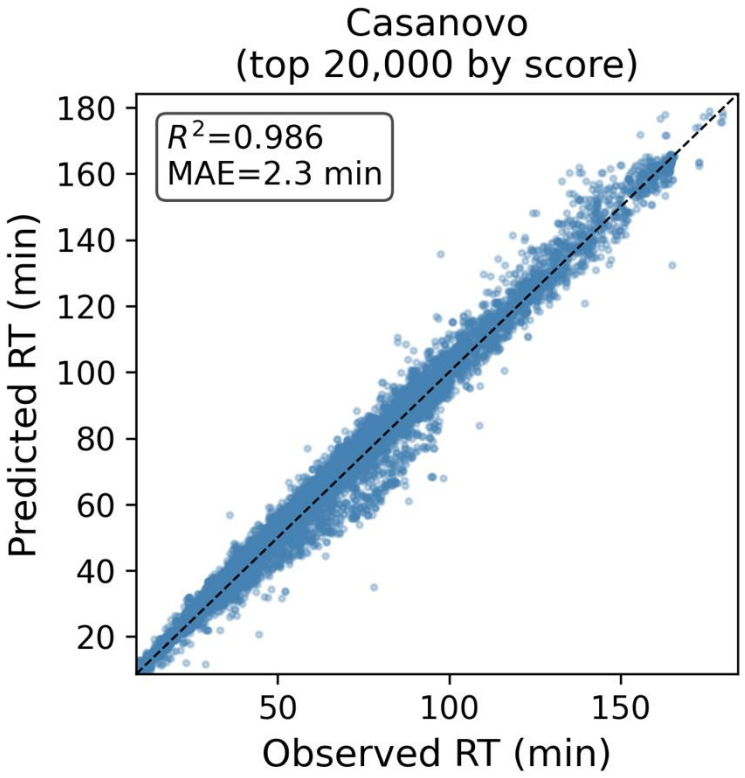


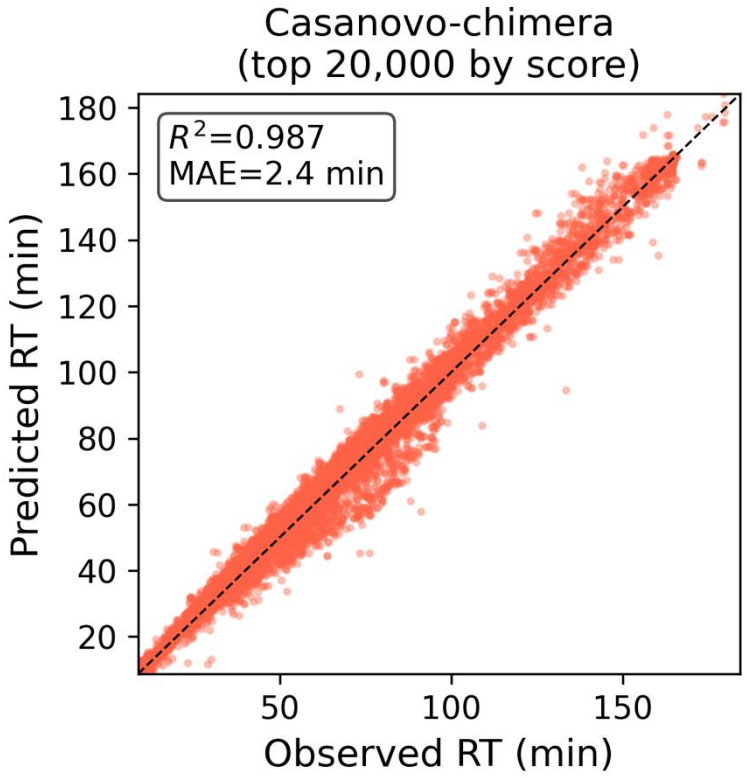
# Casanovo-chim Outperforms non-chimeric model at Both Peptide and Amino Acid Levels



# Top-Ranked Predictions Match Casanovo's Retention Time Reliability

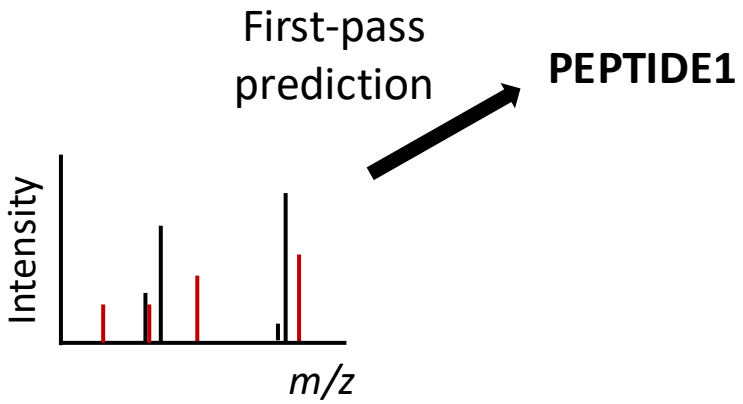



Bo Wen
Noble Lab

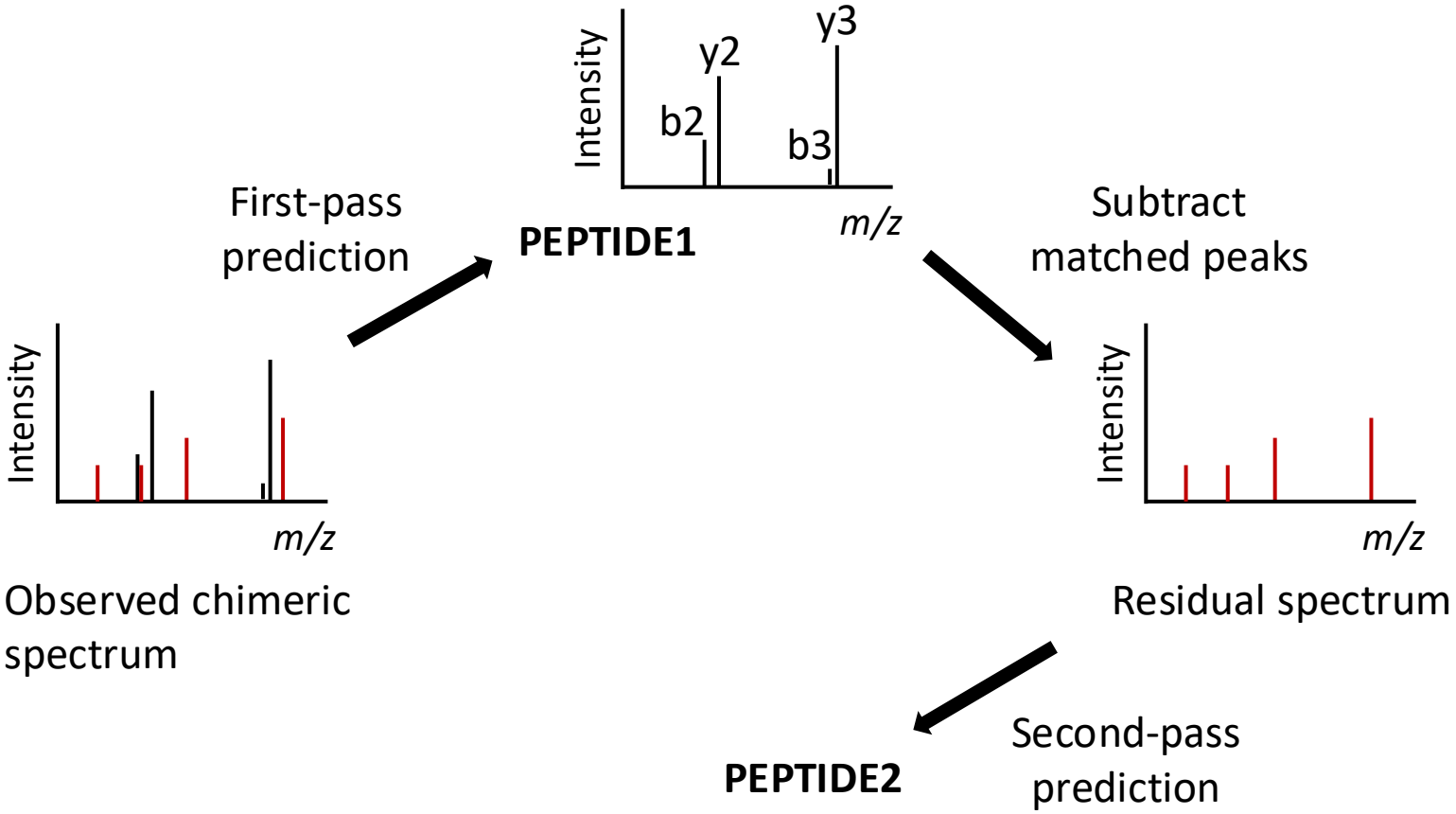
[4] Wen, Bo, et al. "Carafe enables high quality in silico spectral library generation for data-independent acquisition proteomics." Nature Communications 16.1 (2025): 9815.
[5] Wen, Bo, et al. "Carafe2 enables high quality in silico spectral library generation for timsTOF data-independent acquisition proteomics." bioRxiv (2026): 2026-03.

# An Iterative Baseline: Casanovo Performs First-Pass Prediction on Observed Spectra

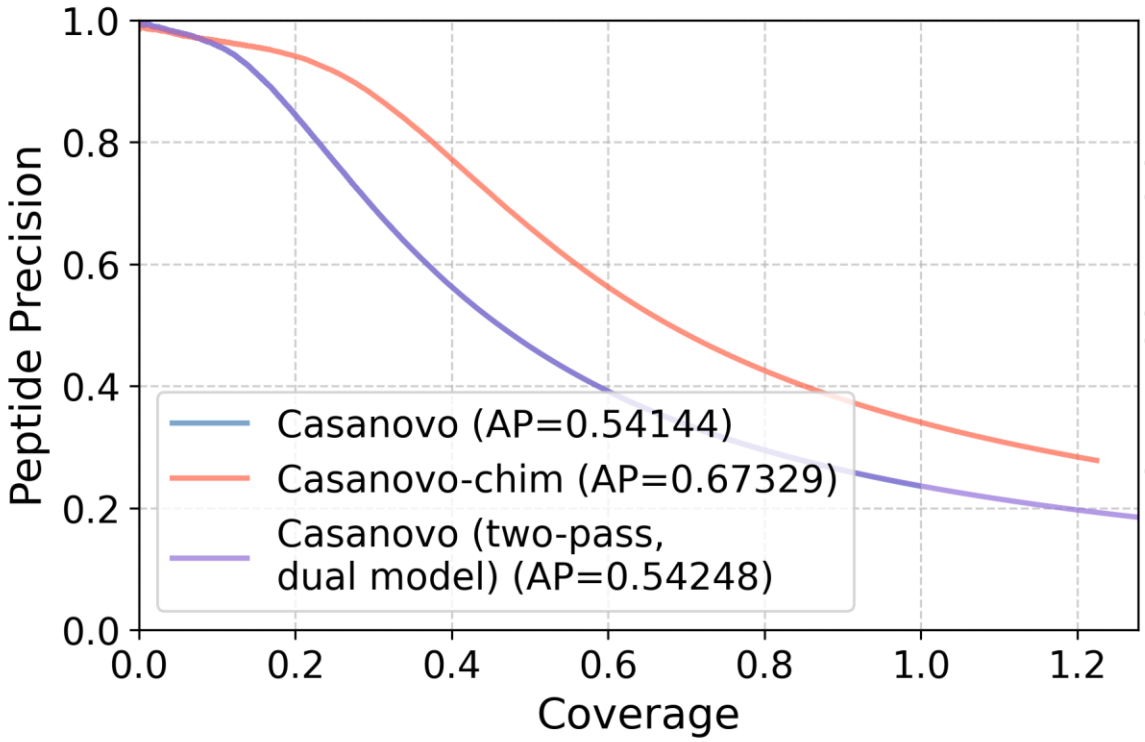


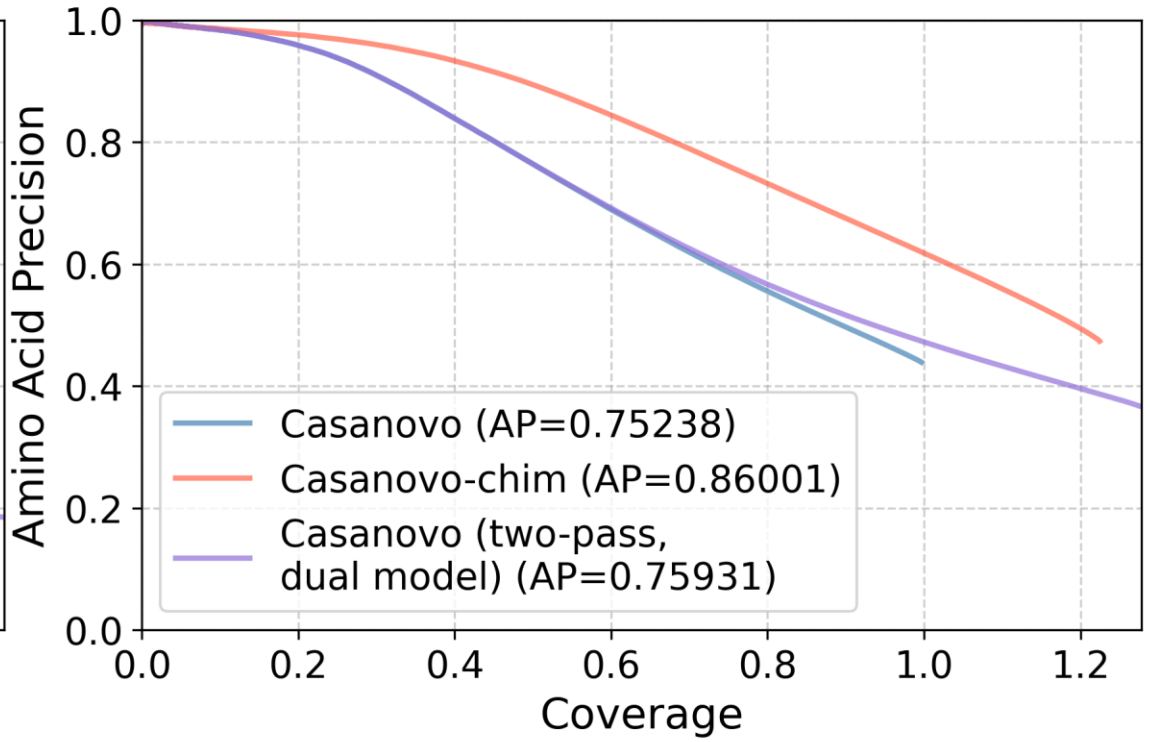

Observed chimeric spectrum

# Second-Pass Prediction Is Performed on Residual Peaks After Subtracting Annotated Peaks



# Casanovo-Chim Outperforms Two-Pass Casanovo



# Takeaway

- We extend Casanovo to directly predict peptides from chimeric spectra.
- At precision 0.8, Casanovo-Chim identifies 17,281 more unique peptides than the non-chimeric baseline — a ~37% lift.
- Casanovo-Chim predictions are validated by consistent retention time.
- Casanovo-Chim outperforms the two-pass Casanovo baseline.

# Acknowledgement


**University of Washington**
**Noble Lab**

William Stafford Noble
Ghulam Murtaza
Ananthan Nambiar
Doudou Yu
Shengqi Hang
Lincoln Harris
Carlo Melendez
Justin Sanders
Bo Wen
Shuhua Xu
Patrick Yu
Xumeng Zhang
Kylie Trousil
Charles Grant
Gwenneth Straub
Priank Dasgupta
Devanshi Desai
Bhavya Krishnan

**University of Antwerp**
Wout Bittremieux