ICML
International Conference
On Machine Learning

THE HONG KONG
POLYTECHNIC UNIVERSITY
香港理工大學

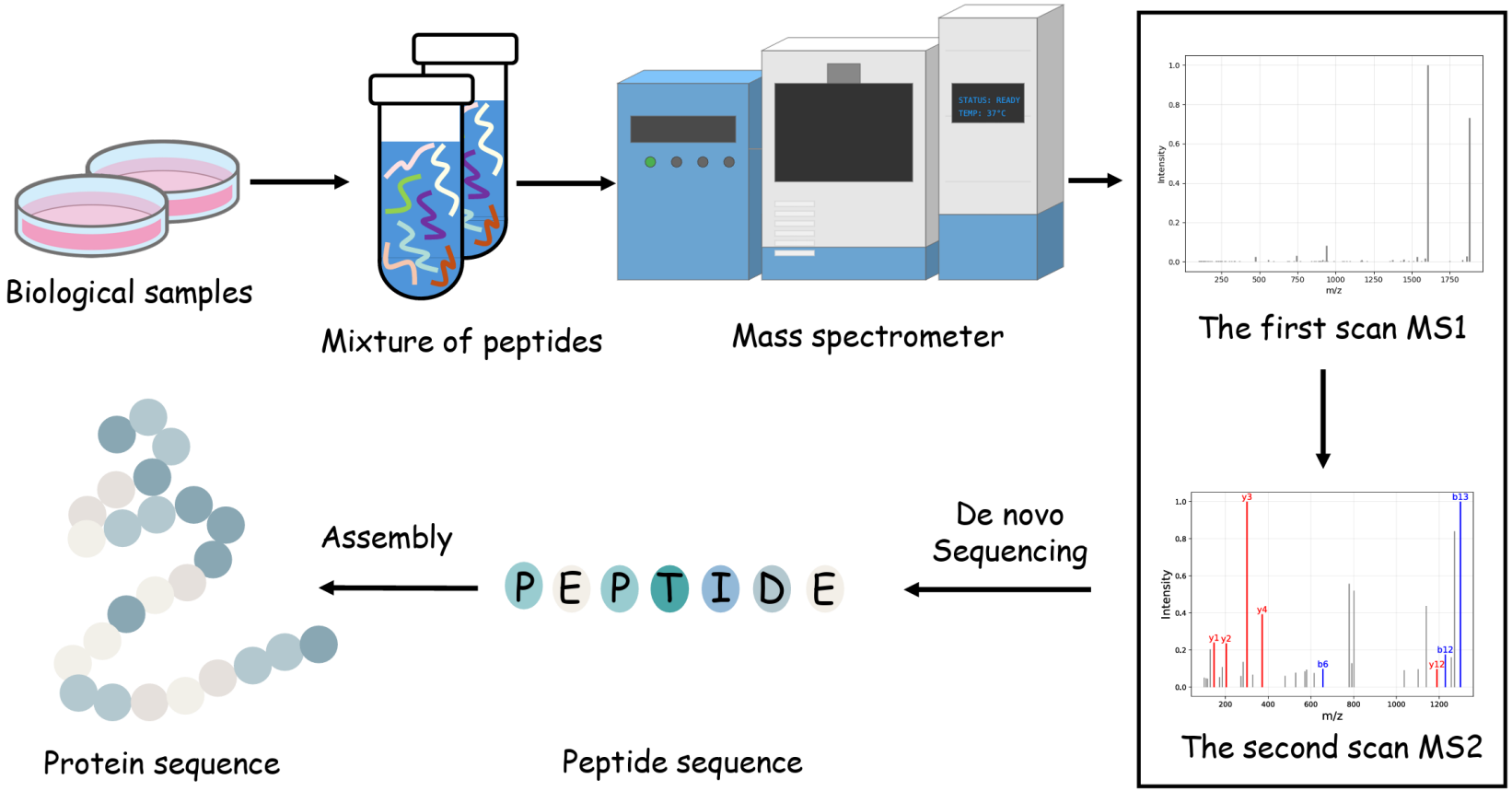
Latent Imputation before Prediction: A New Computational Paradigm for *De Novo* Peptide Sequencing

Ye Du, Chen Yang, Nanxi Yu, Wanyu Lin, Qian Zhao, ShujunWang*

The Hong Kong Polytechnic University

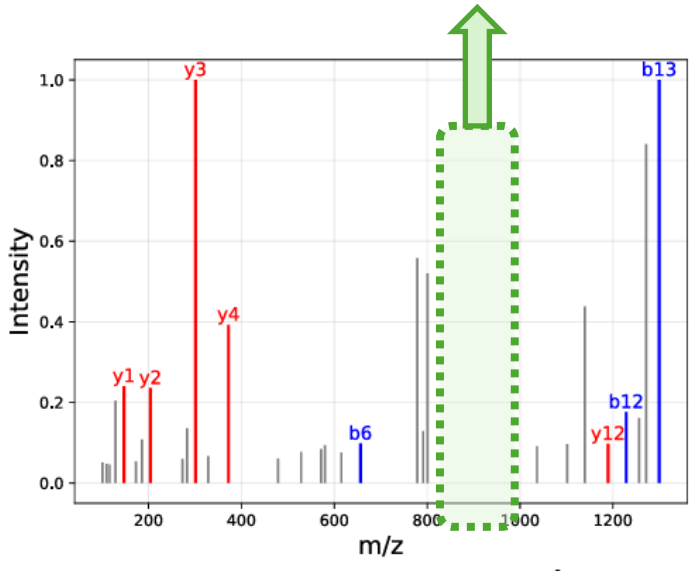
*Correspondence to: shu-jun.wang@polyu.edu.hk

The Identification Workflow of Shotgun Proteomics

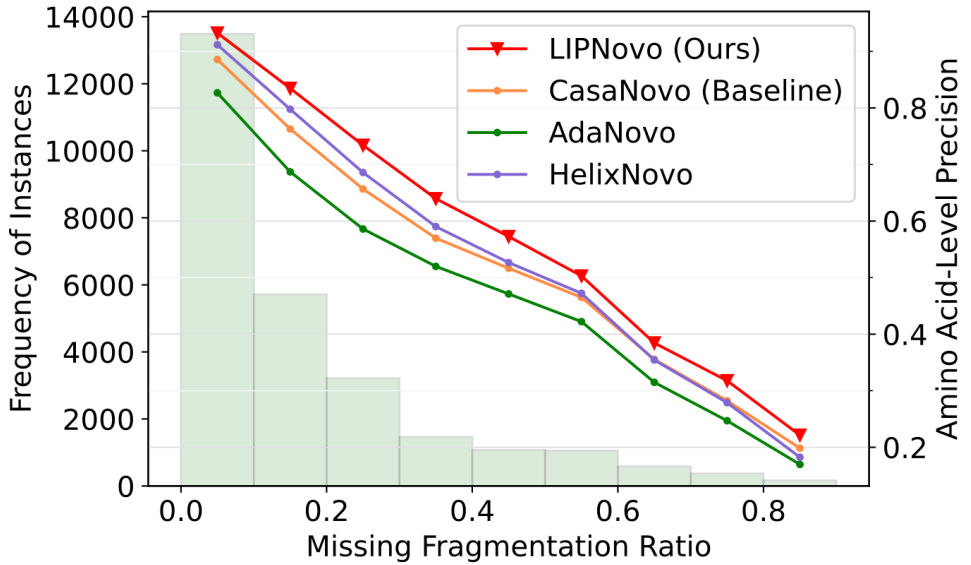


Missing Fragmentation Issue

Missing signal peaks, such as b7,y8 …

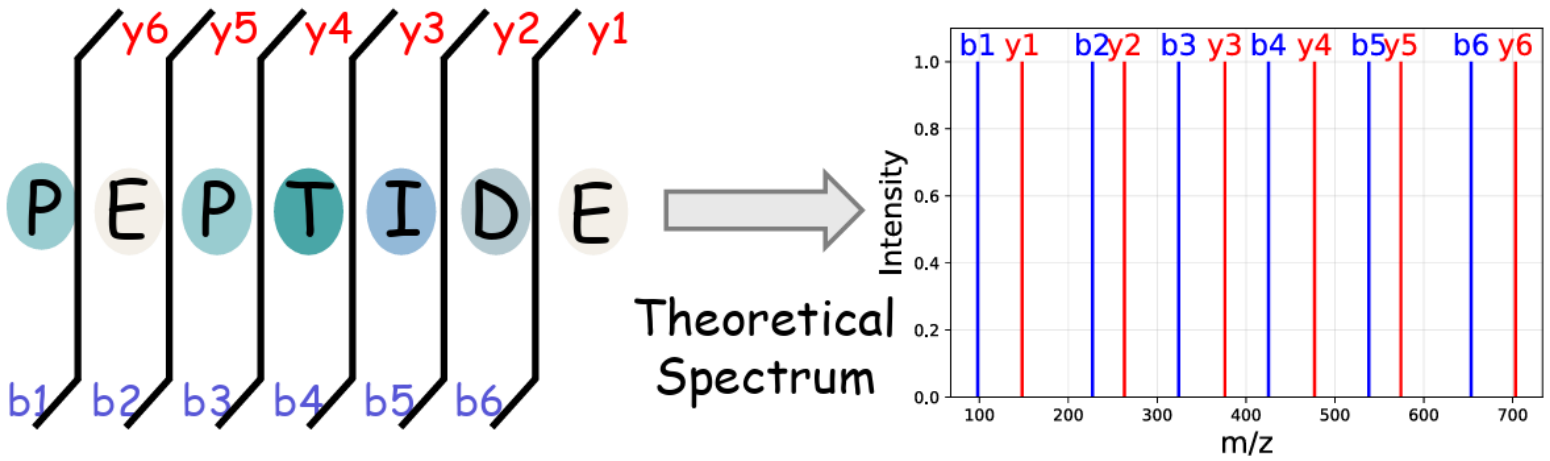


As the missing ratio increases, performance deteriorates dramatically.

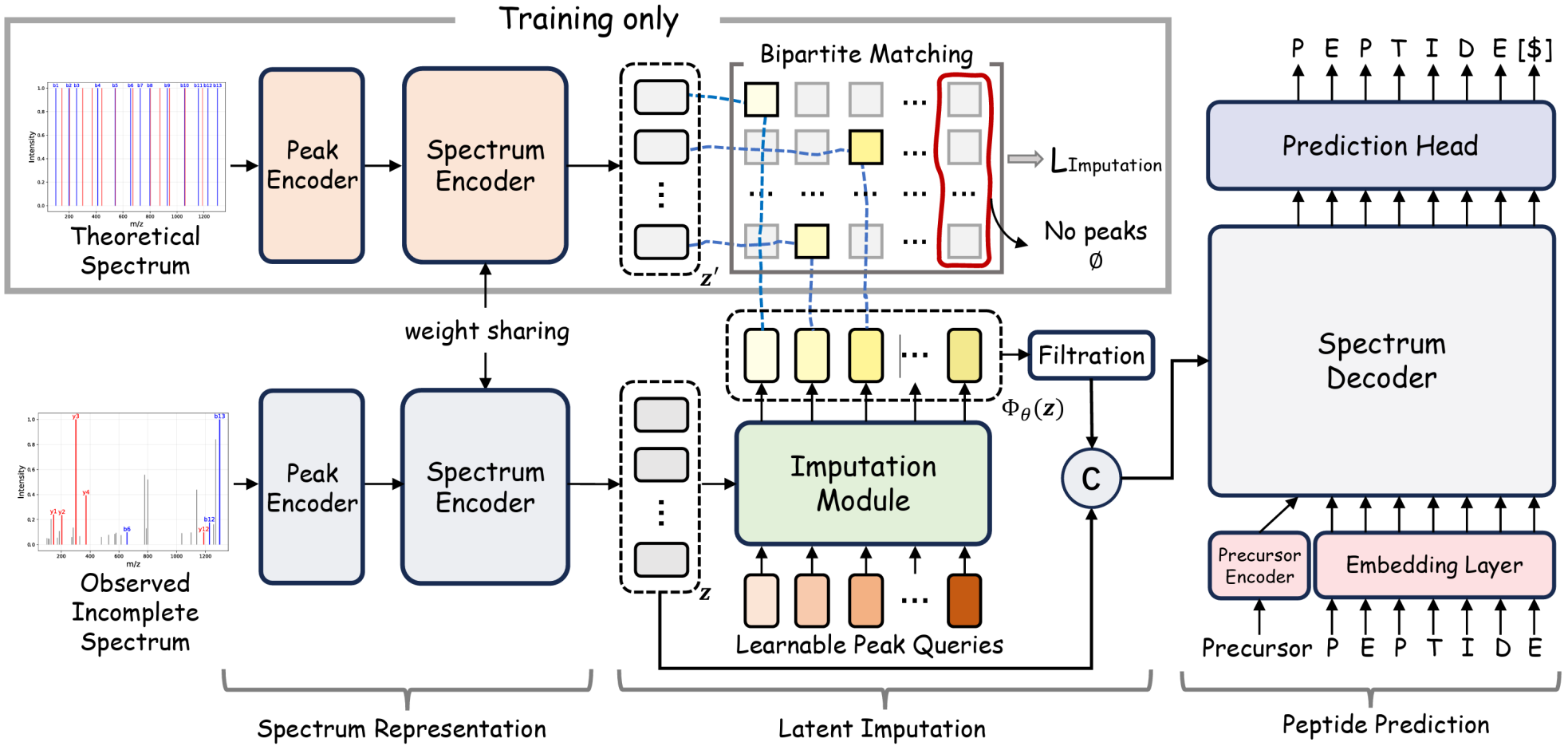


Learn to Impute the Missing Peaks

We can calculate the theoretical spectrum during training.



Latent Imputation before Prediction



Comparison with State-of-the-arts

LIPNovo achieves SOTA performance on three public datasets, e.g., achieve +20% amino acid-level improvements on the Seven-species dataset.

Table 1. Empirical comparison with state-of-the-art methods on Nine-species, Seven-species, and HC-PT datasets in amino acid-level and peptide-level performance. † denotes our retrained results, and other results are provided by NovoBench. The best is marked in bold.

Method	Amino Acid-Level Performance						Peptide-Level Performance					
	Nine-species		Seven-species		HC-PT		Nine-species		Seven-species		HC-PT	
	Prec.	Recall	Prec.	Recall	Prec.	Recall	Prec.	AUC	Prec.	AUC	Prec.	AUC
PEAKS (Ma et al., 2003)	0.748	-	-	-	-	-	0.428	-	-	-	-	-
DeepNovo (Tran et al., 2017)	0.696	0.638	0.492	0.433	0.531	0.534	0.428	0.376	0.204	0.136	0.313	0.255
PointNovo (Qiao et al., 2021)	0.740	0.671	0.196	0.169	0.623	0.622	0.480	0.436	0.022	0.007	0.419	0.373
InstaNovo (Eloff et al., 2023)	0.420	0.395	0.192	0.176	0.289	0.285	0.164	0.123	0.031	0.009	0.057	0.034
CasaNovo (Yilmaz et al., 2024)	0.697	0.696	0.322	0.327	0.442	0.453	0.481	0.439	0.119	0.084	0.211	0.177
AdaNovo (Xia et al., 2024)	0.698	0.709	0.379	0.385	0.442	0.451	0.505	0.469	0.174	0.135	0.212	0.178
AdaNovo† (Xia et al., 2024)	0.681	0.681	0.403	0.405	0.492	0.496	0.473	0.439	0.189	0.149	0.289	0.254
π-HelixNovo (Yang et al., 2024)	0.765	0.758	0.481	0.472	0.588	0.582	0.517	0.453	0.234	0.173	0.356	0.318
π-HelixNovo† (Yang et al., 2024)	0.765	0.752	0.465	0.462	0.532	0.537	0.509	0.431	0.218	0.156	0.301	0.261
Baseline† (Yilmaz et al., 2024)	0.741	0.740	0.357	0.366	0.525	0.530	0.529	0.493	0.159	0.119	0.324	0.290
LIPNovo (Ours)	**0.797**	**0.797**	**0.557**	**0.560**	**0.637**	**0.643**	**0.582**	**0.547**	**0.327**	**0.281**	**0.458**	**0.427**

Model Analysis

Ablation experiments show the effectiveness of each component.

Table 5. Component ablation. "Impu." denotes the imputation module, and $\mathcal{L}_{\text{CE}}(z')$ means the CE loss supervised on the theoretical spectrum. "Comp." means the complementary spectrum.

	Baseline	Impu.	$\mathcal{L}_{\text{CE}}(z')$	Comp.	Amino Acid Level		Peptide Level	
					Prec.	Recal	Prec.	AUC
1	✓	✗	✗	✗	0.741	0.740	0.529	0.493
2	✓	✗	✗	✓	0.755	0.755	0.537	0.500
3	✓	✓	✗	✓	0.766	0.764	0.546	0.513
4	✓	✓	✓	✗	0.782	0.782	0.569	0.536
5	✓	✓	✓	✓	**0.797**	**0.797**	**0.582**	**0.547**

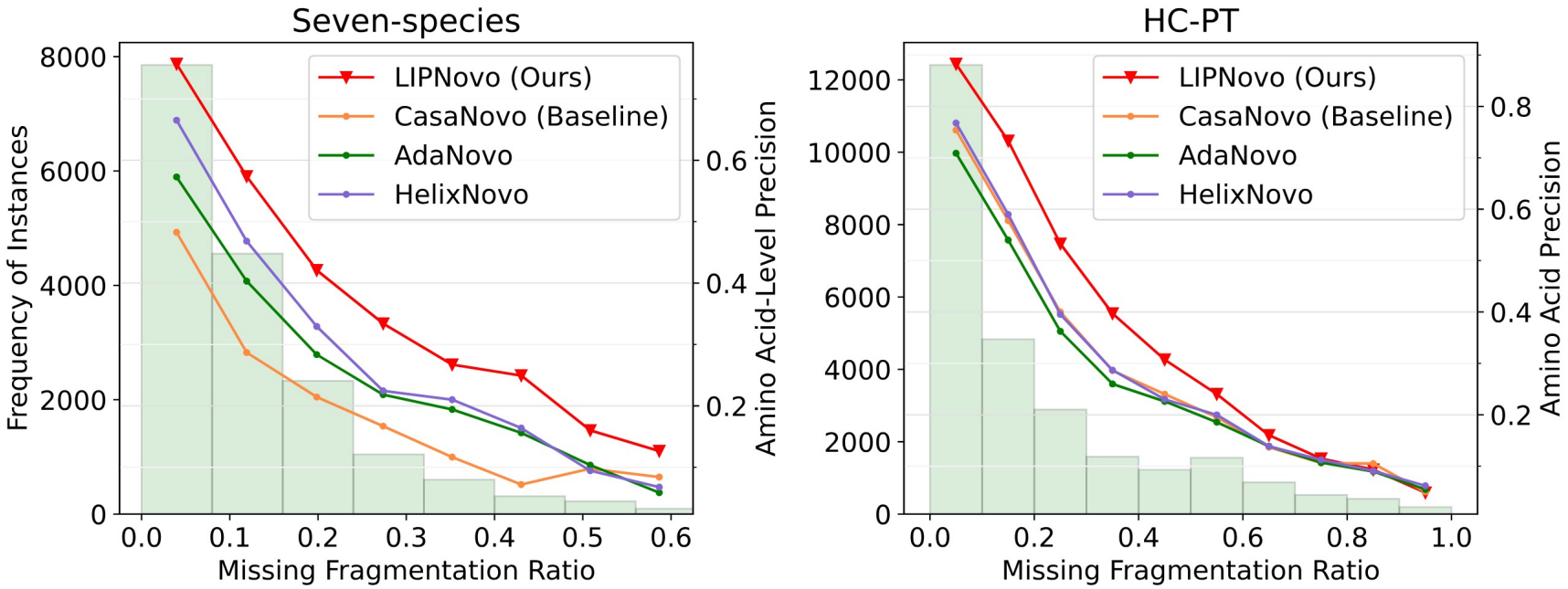
Performance enhancement originates from imputation mechanism rather than the additional parameters.

Table 7. Parameters *vs.* model performance. ¶ is the extension.

Method	# Params	Amino Acid Level		Peptide Level	
		Prec.	Recall	Prec.	AUC
Baseline	47.4M	0.741	0.740	0.529	0.493
Baseline¶	69.4M	0.750	0.751	0.539	0.494
LIPNovo	68.4M	**0.797**	**0.797**	**0.582**	**0.547**

Comparison across Missing Fragmentation Ratios (MFR)

LIPNovo achieves consistent performance improvements across MFRs.



Thanks

Paper

Code

For more details, refer to our paper and code