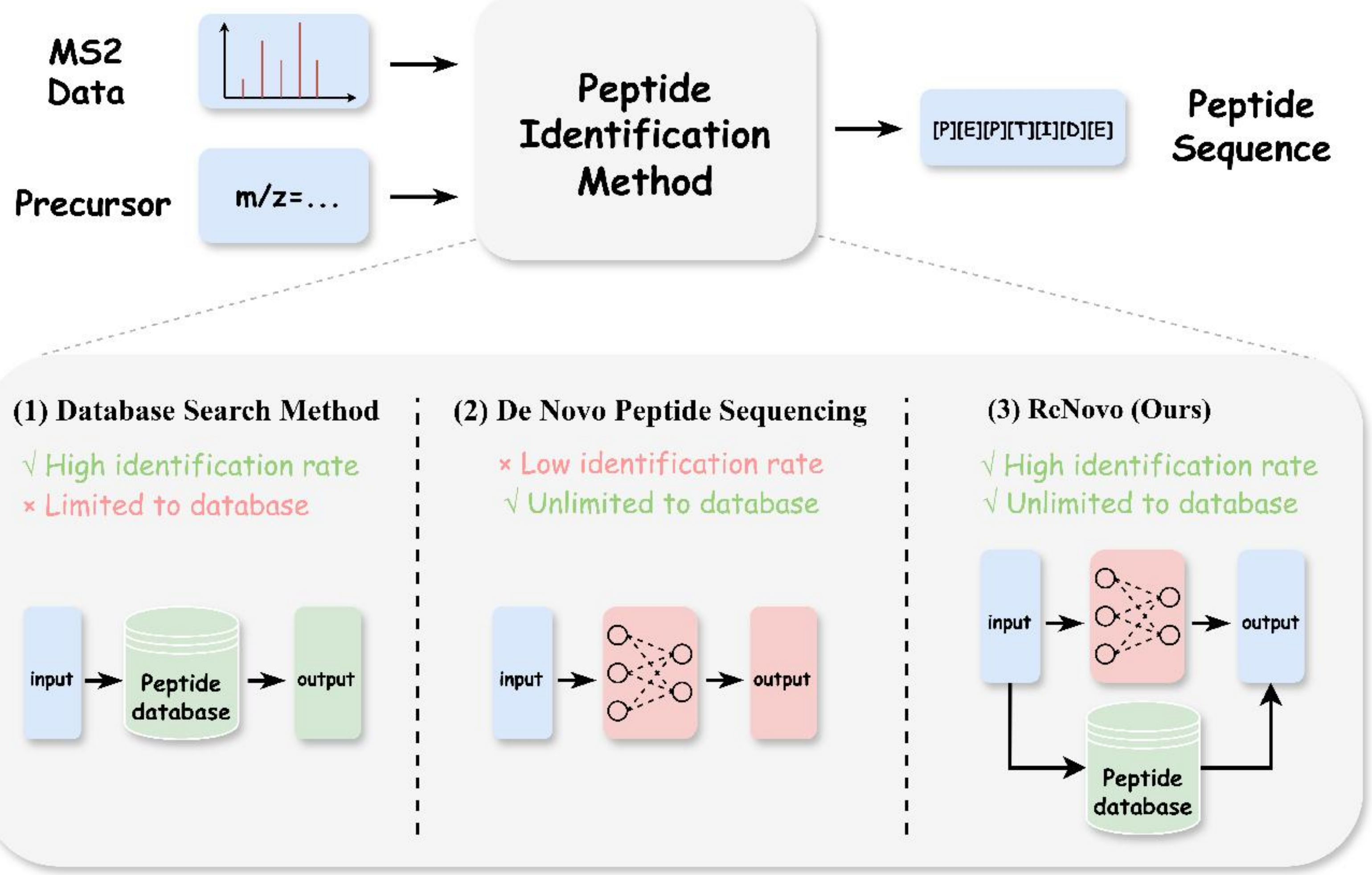
ReNovo: Retrieval-Based De Novo Mass Spectrometry Peptide Sequencing

Authors: Shaorong Chen, Jun Xia, Jingbo Zhou, Lecheng Zhang, Zhangyang Gao, Bozhen Hu, Cheng Tan, Wenjie Du, Stan Z. Li

Keywords: AI for Science, Peptide Sequencing, Proteomics, Computational Biology

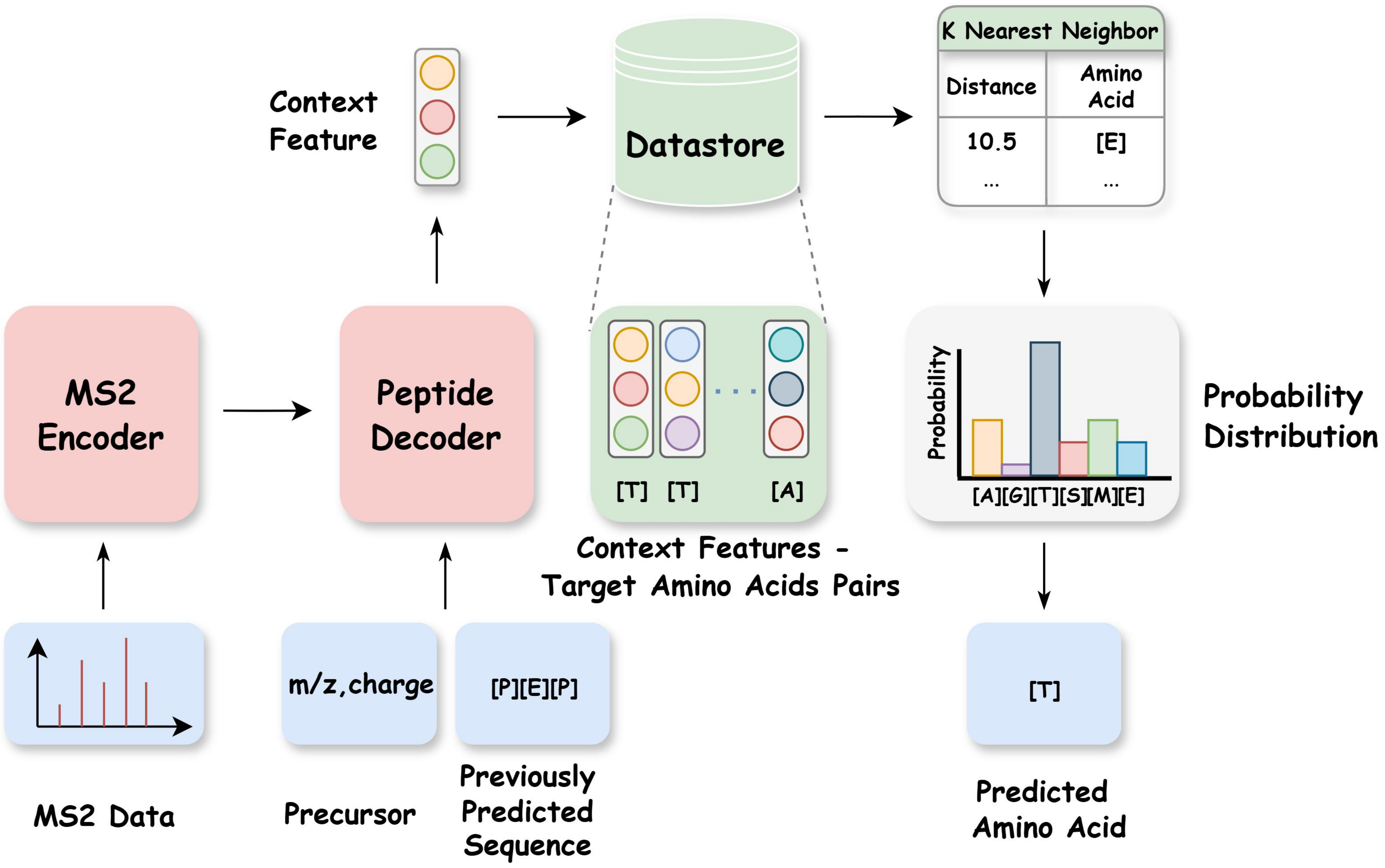
Contact (WeChat):

Motivation: ReNovo Combines the Strengths of Both Database Search and De Novo Sequencing Methods



ReNovo: A Novel Retrieval-based De Novo Peptide Sequencing Methodology

- **Model Training**: ReNovo undergoes supervised training using training data
- **Datastore Building Stage**: ReNovo generates context feature - target amino acid pairs using the training dataset, which are then stored in the datastore
- **Retrieval-Based Inference Stage**: ReNovo model will retrieve the datastore and incorporate the retrieved pairs information to make the final prediction



Experimental Results

- ReNovo Achieves Significant Improvement in Peptide-level Metrics

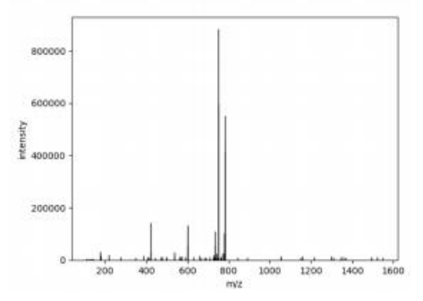
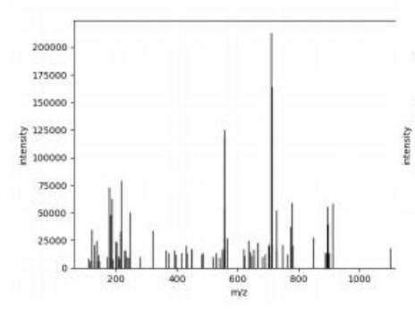
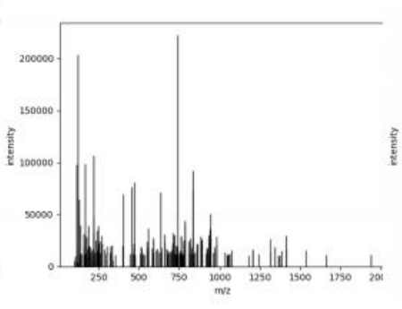
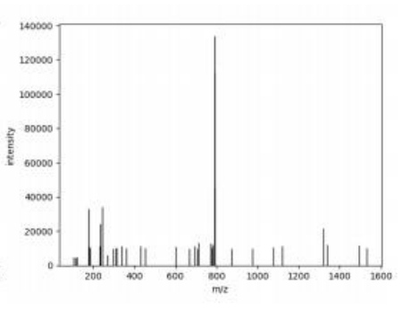
Method	Peptide-level performance					
	Seven-species		Nine-species		HC-PT	
	Prec.	AUC	Prec.	AUC	Prec.	AUC
DeepNovo	0.204	0.136	0.428	0.376	0.313	0.255
PointNovo	0.022	0.007	0.480	0.436	<u>0.419</u>	<u>0.373</u>
CasaNovo	0.119	0.084	0.481	0.439	0.211	0.177
HelixNovo	<u>0.234</u>	<u>0.173</u>	<u>0.517</u>	0.453	0.356	0.318
AdaNovo	0.174	0.135	0.505	<u>0.469</u>	0.212	0.178
ReNovo	**0.278**	**0.228**	**0.568**	**0.528**	**0.467**	**0.436**

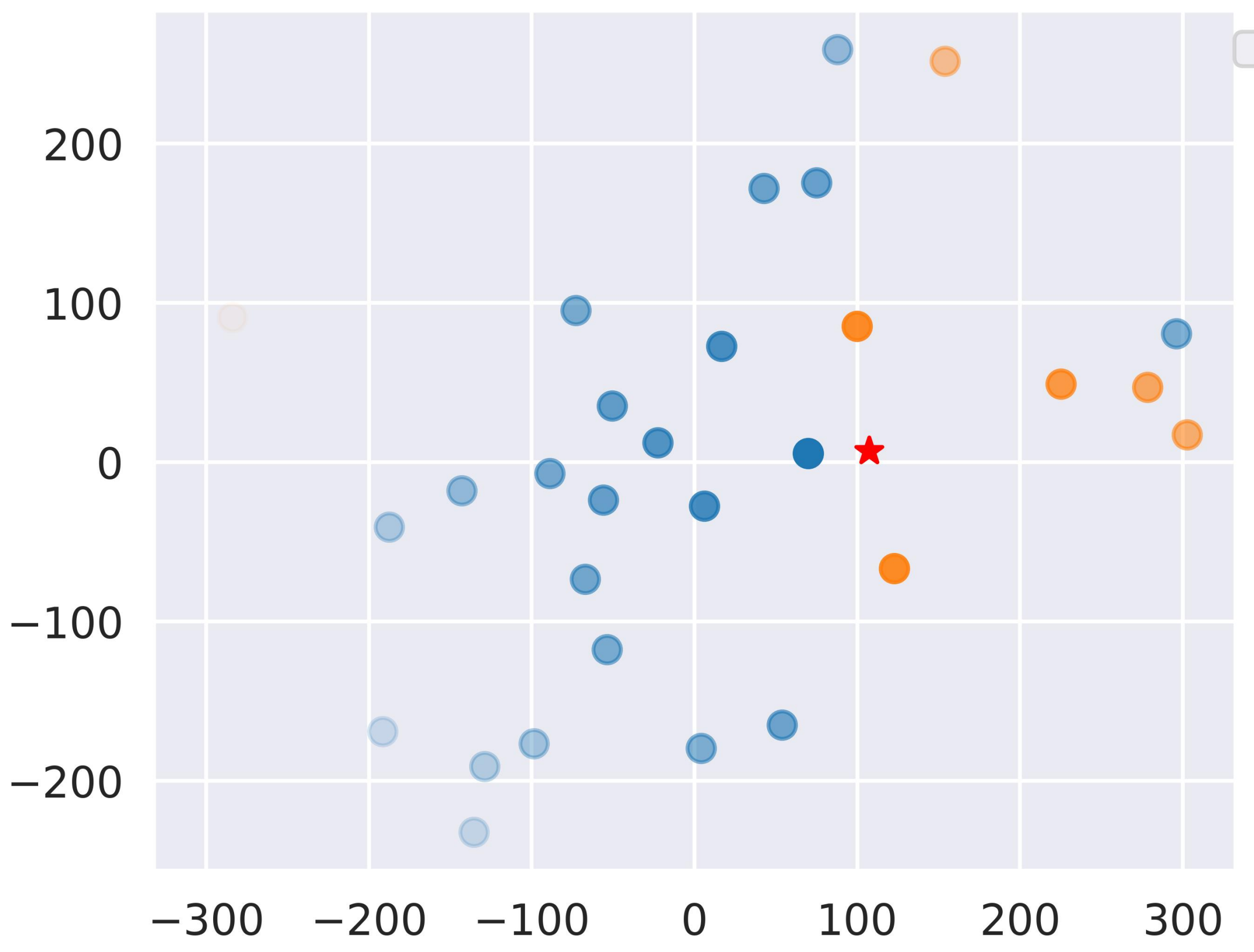
- The Time and Storage Consumption of the ReNovo Model Is Minor

	Model Training	**Datastore Building**	**Retrieval-Based Inference**
Time(s)	84,703	2,207	8,278
Percentage	88.98%	2.32%	8.70%

	Seven-species Dataset	**Nine-species Dataset**	**HC-PT Dataset**
Pairs Number	5,626,944	8,456,240	3,232,616
Storage (GB)	11.16	16.77	6.42

- ReNovo Can Be More Accurate with The Assistance of The Datastore

Input MS2	**Previously Predicted**			
	RVNLARIDNE			
	Sample 1	**Sample 2**	**Sample 3**	**...**
Retrieved MS2				...
Retrieved Amino Acid	D	E	D	...
Retrieved Distance	22.11	26.21	26.23	...
Predicted Amino Acid($K = 0$) **Predicted Amino Acid($K = 32$)** **Ground Truth Amino Acid**	$p(\mathrm{D}) = 0.45, p(\mathrm{E}) = 0.50, \ldots$ $p(\mathrm{D}) = 0.48, p(\mathrm{E}) = 0.47, \ldots$ D			
Predicted Peptide($K = 0$) **Predicted Peptide($K = 32$)** **Ground Truth Peptide**	RVNLARIDNE<u>E</u>VM(+15.99) RVNLARIDNE<u>D</u>VM(+15.99) RVNLARIDNE<u>D</u>VM(+15.99)			


200
100
0
−100
−200
−300
−200
−100
0
100
200
300