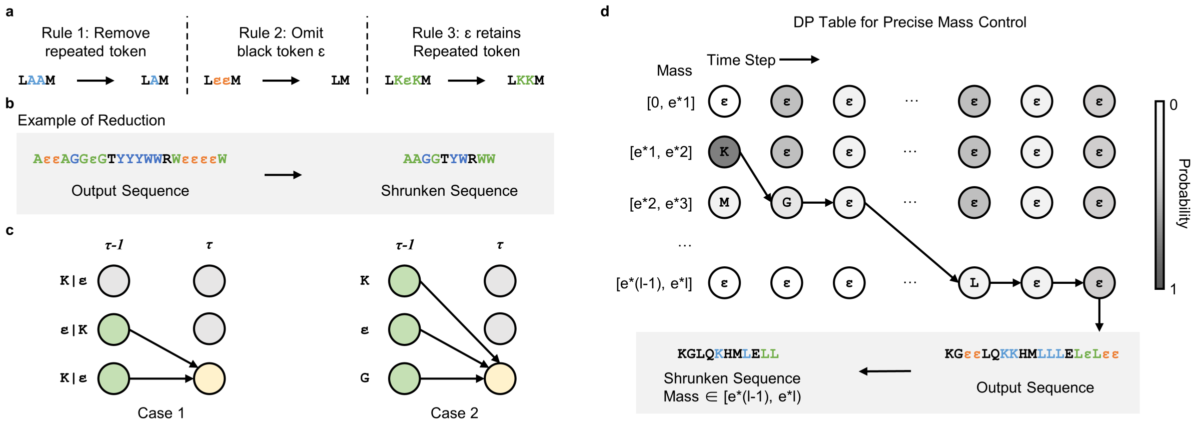
Supplementary Information for: π-PrimeNovo

Xiang Zhang♣[1,2], Tianze Ling♣[3,4], Zhi Jin♣[1], Sheng Xu♣[1,5], Zhiqiang Gao[1], Boyan Sun[4], Zijie Qiu[1,5], Jiaqi Wei[1,6], Nanqing Dong[1], Guangshuai Wang[1,5], Guibin Wang[4], Leyuan Li[4], Muhammad Abdul-Mageed[2,7], Laks V.S. Lakshmanan[2], Fuchu He[4,8], Wanli Ouyang†[1], Cheng Chang†[4], and Siqi Sun†[5,1]

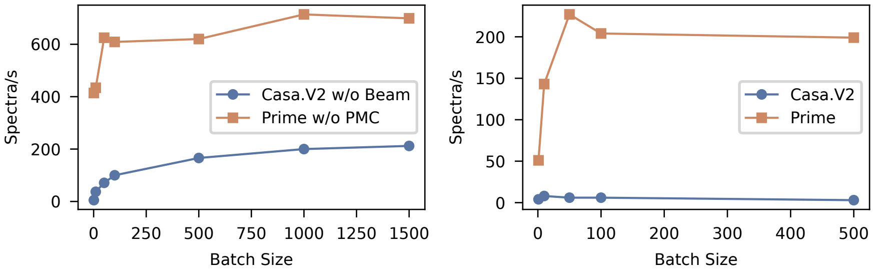
[1] Shanghai Artificial Intelligence Laboratory
[2] University of British Columbia
[3] Tsinghua University
[4] State Key Laboratory of Medical Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics
[5] Research Institute of Intelligent Complex Systems, Fudan University
[6] Zhejiang University
[7] MBZUAI
[8] International Academy of Phronesis Medicine (Guangdong)

Supplementary Figures

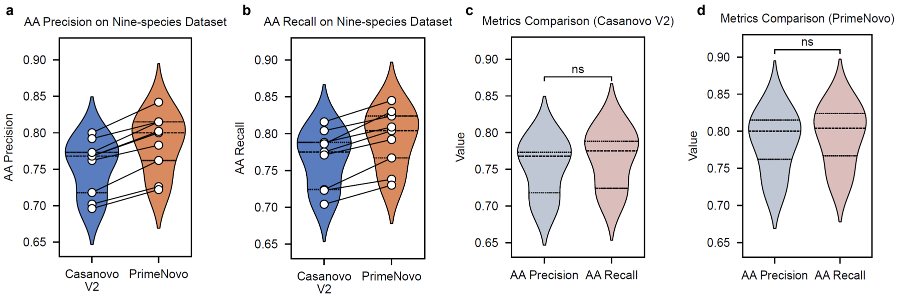


Supplementary Fig. 1: **a.** Visualization of the CTC Token Processing Reduction Rule. **b.** A case study of how the reduction rule can be applied to a sequence to obtain the final sequence. **c.** Two Cases in CTC Loss Calculation: On the left, the scenario where the upcoming token is a repeat of the previous one. On the right, the scenario where the next token differs from the preceding one **d.** Overview of the PMC Decoding Process: Sequential decoding from the first to last time step with mass constraint adherence in Each Cell, culminating in the application of the CTC Reduction Rule for the final output.

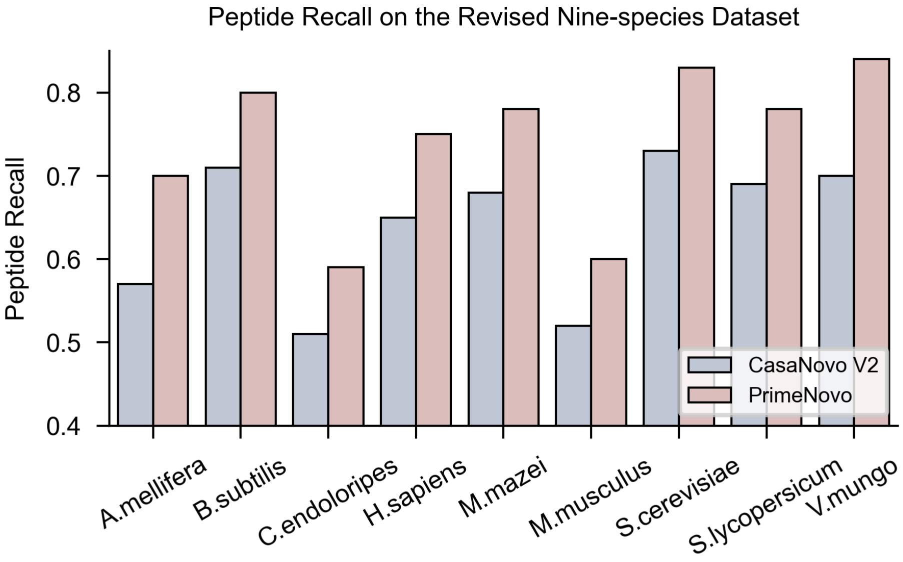
♣ These authors contributed equally to this work.
† Corresponding authors: siqisun@fudan.edu.cn, changcheng@ncpsb.org.cn, ouyangwanli@pjlab.org.cn



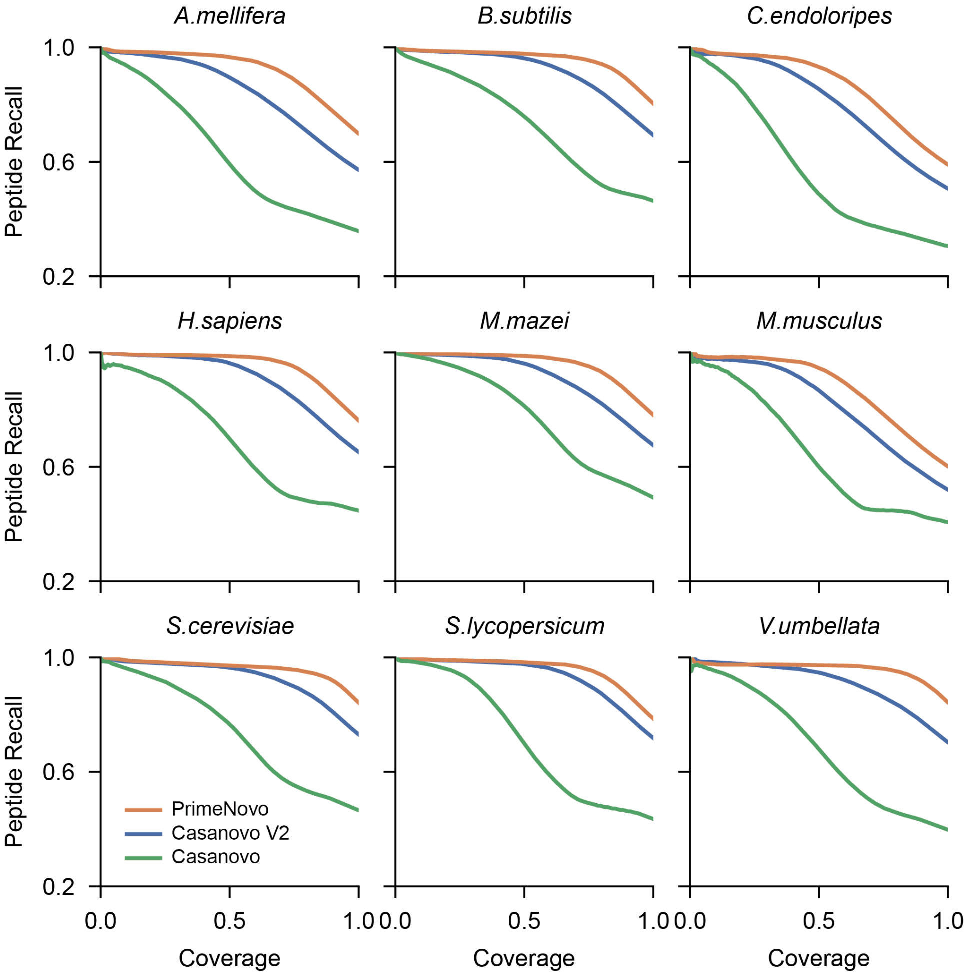
Supplementary Fig. 2: The left-hand side shows results when no post-decoding algorithms (PMC and beam search) are used for either model, while the right-hand side shows results with PMC and beam search applied.



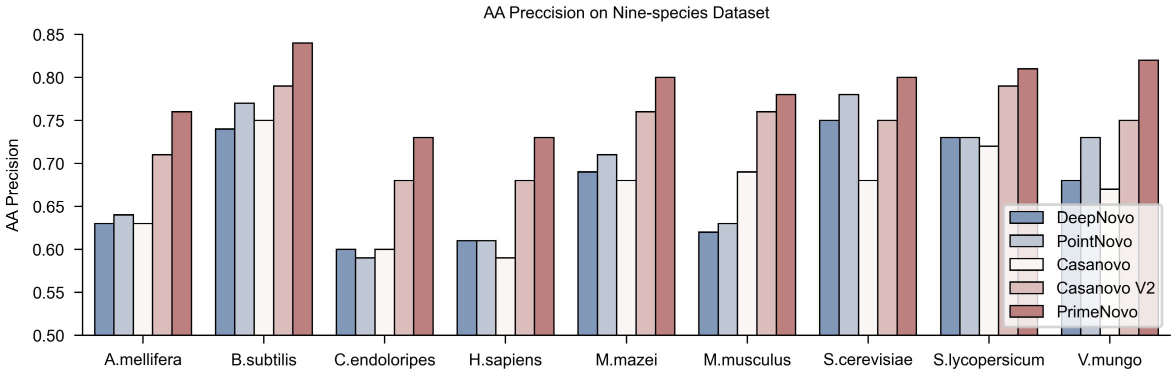
Supplementary Fig. 3: **a**. Comparison of AA precision across 9 different test species between PrimeNovo and Casanovo V2. **b**. Comparison of AA recall across the same 9 test species between PrimeNovo and Casanovo V2. **c**. The difference in distribution between AA recall and AA precision for Casanovo V2 across the nine-species dataset (n=9, ns indicates `not significant` (p-value set to 0.05)) according to the Mann-Whitney U test). **d**. The difference in distribution between AA recall and AA precision for PrimeNovo across the nine-species dataset (n=9, ns indicates `not significant` according to the Mann-Whitney U test).



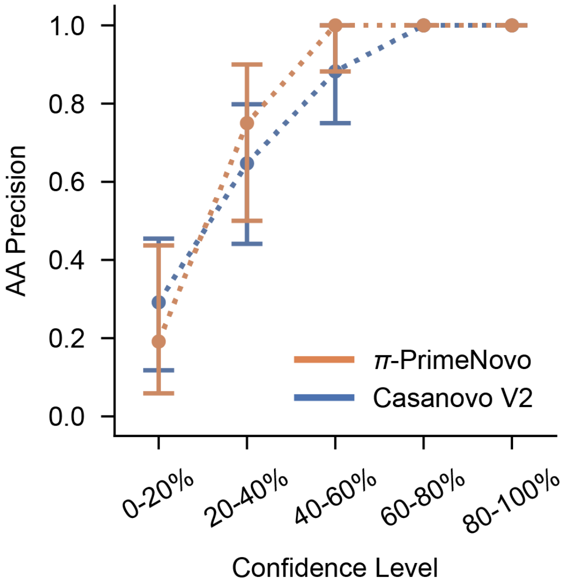
Supplementary Fig. 4: Peptide recall on the revised nine-species dataset.



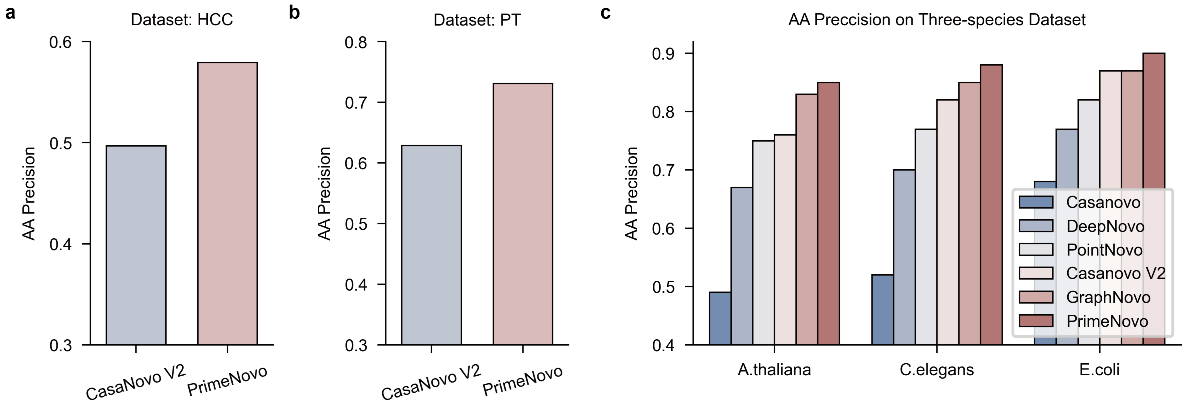
Supplementary Fig. 5: Peptide recall-coverage curves on the revised nine-species dataset.



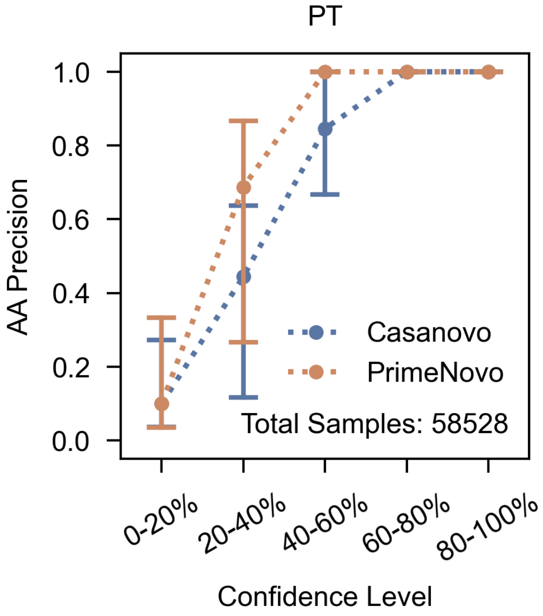
Supplementary Fig. 6: The prediction precision at AA level on the nine-species benchmark dataset. The AA precision of PrimeNovo is significantly higher than that of other de novo methods.



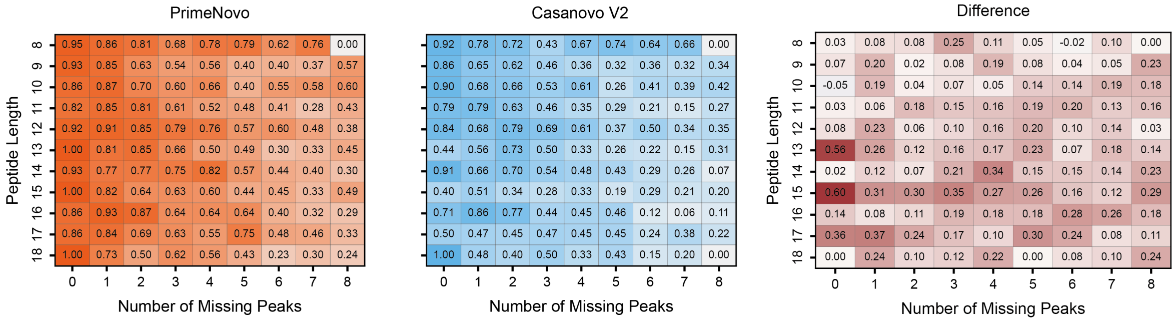
Supplementary Fig. 7: The prediction precision at AA level with the increase of confidence score on nine-species benchmark dataset. PrimeNovo exhibits an overall higher AA precision compared to Casanova V2. Data are presented as median values of each confidence level with interquartile range (50% percentile interval, n=89,928).



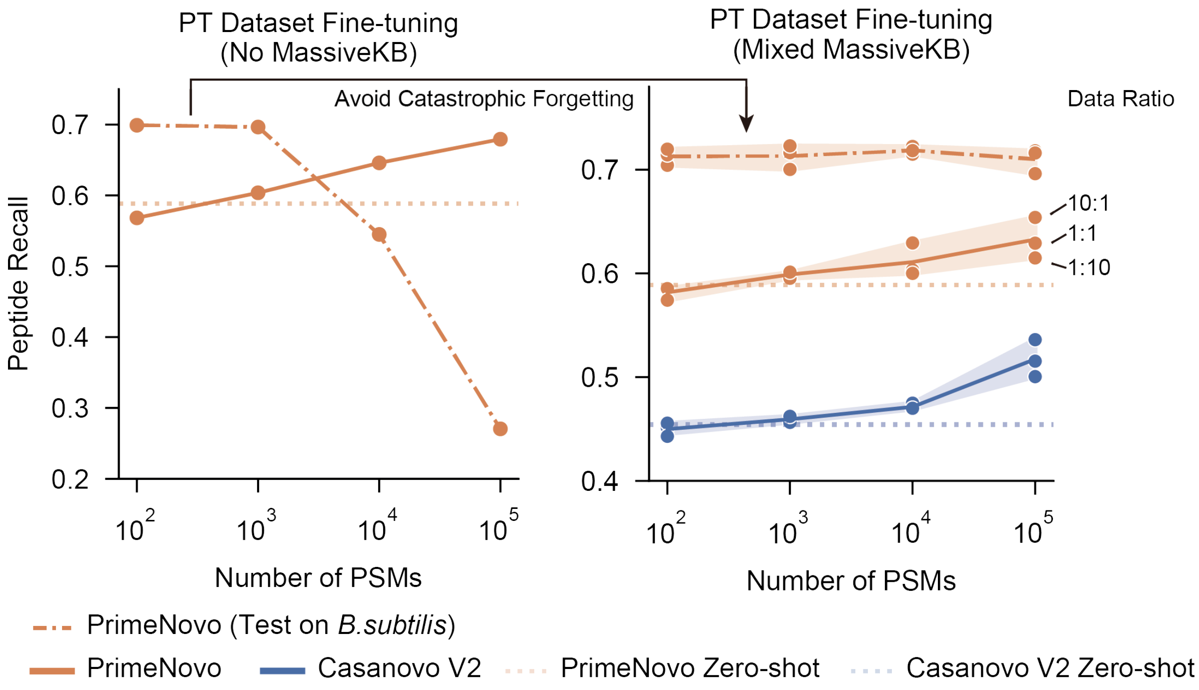
Supplementary Fig. 8: The prediction precision at Amino Acid level on the test datasets HCC (a), PT (b), and Three-species (c), respectively.



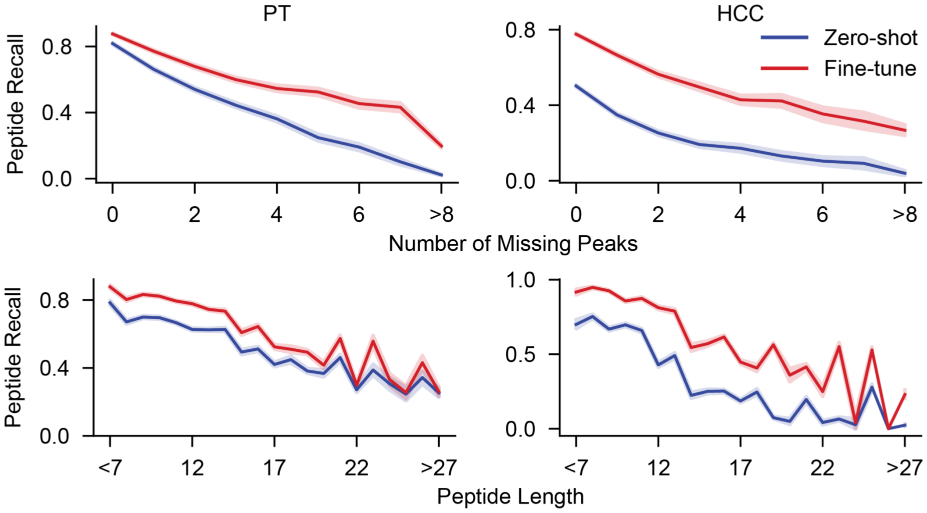
Supplementary Fig. 9: The prediction precision at AA level with the increase of confidence score on the PT dataset. PrimeNovo exhibits an overall higher AA precision compared to Casanova V2. Data are presented as median values of each confidence level with interquartile range (50% percentile interval, n=58,528).



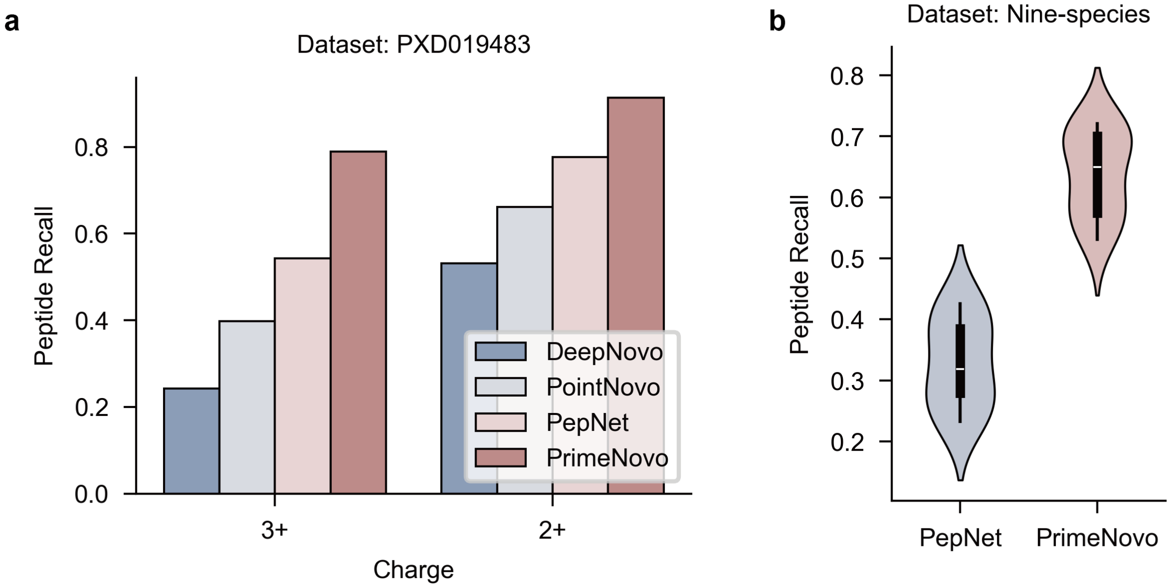
Supplementary Fig. 10: The peptide recalls are shown by heatmap for different combinations of the number of missing peaks and peptide lengths of PrimeNovo (left) and Casanovo V2 (middle). The right heatmap shows the differences in peptide recalls between PrimeNovo and Casanovo V2 for different combinations of the number of missing peaks and peptide lengths.



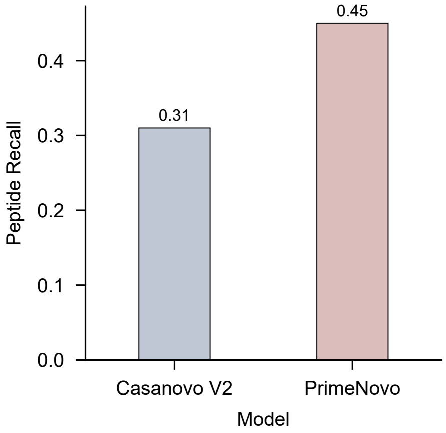
Supplementary Fig. 11: The performance of fine-tuning on the PT test dataset as more PT training data is added during fine-tuning. The left panel shows fine-tuning using only the PT dataset, which leads to a catastrophic forgetting of the original data distribution (on part of the nine-species benchmark dataset *B.subtilis*). The right side illustrates fine-tuning with a mixture of PT and MassIVE-KB training data. The data points in the right figure shows the performance of three different data ratio during the fine-tuning stage. We plot a central curve that connects the mean values of the data points, with light background representing the s.d.



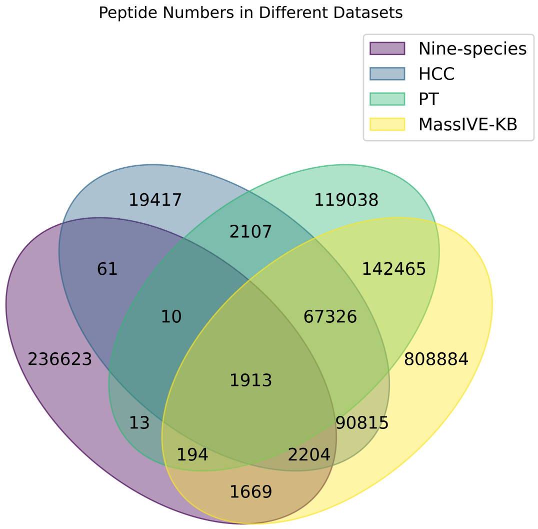
Supplementary Fig. 12: The performance of PrimeNovo for varying peptide lengths and missing peaks (n=117,056 for PT dataset and n=26,000 for HCC dataset; the curve represent mean values of the data points, with light background representing the s.d.).



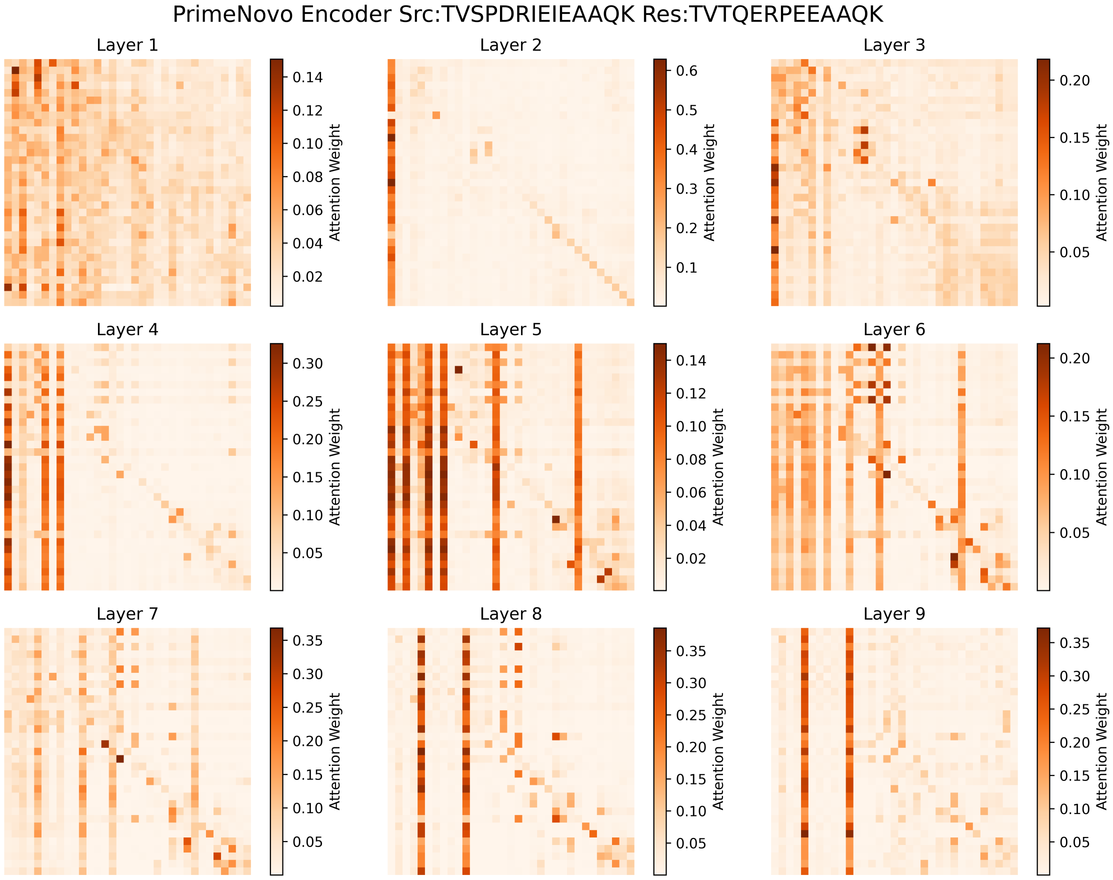
Supplementary Fig. 13: Performance comparison with PepNet. **a.** Comparing PrimeNovo with DeepNovo, PointNovo and PepNet on the PepNet(Human) dataset at $2+$ and $3+$ charges, respectively. **b.** Comparison of PrimeNovo on the nine-species benchmark (n=9) dataset with PepNet.



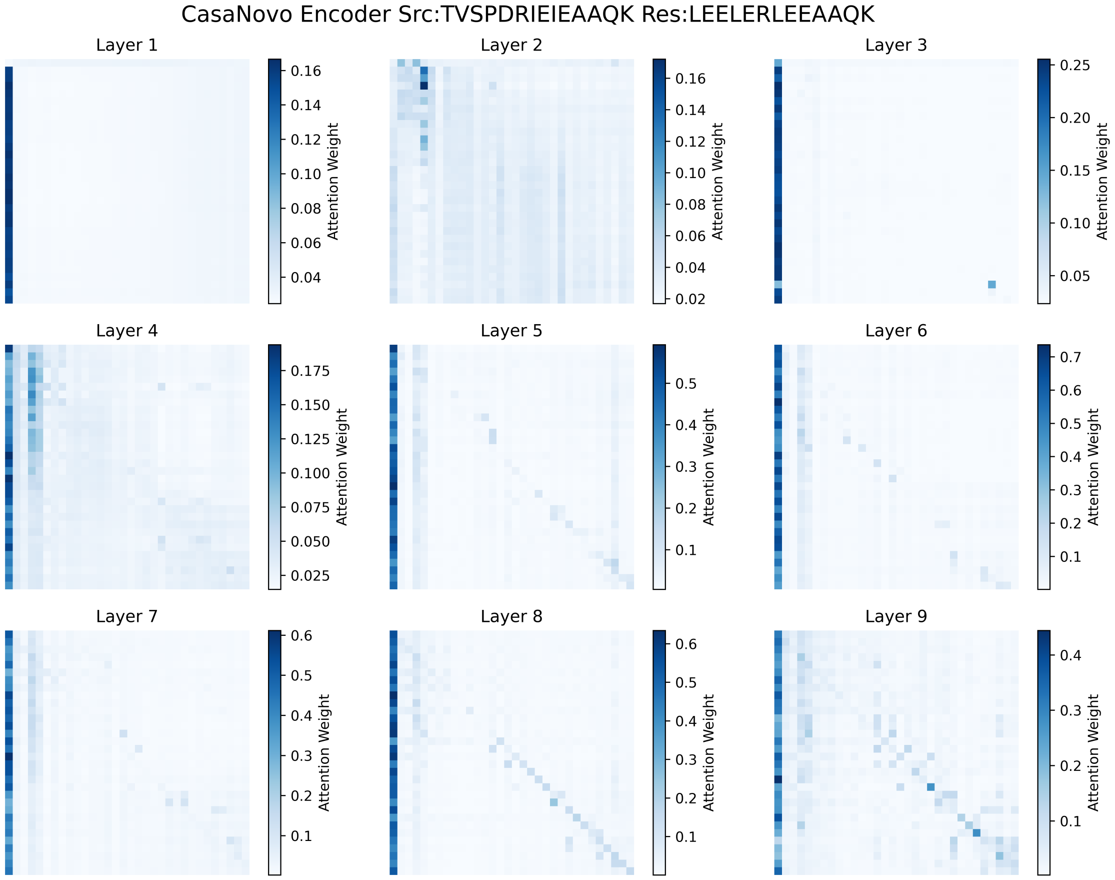
Supplementary Fig. 14: Performance of PrimeNovo and Casanovo V2 on a TimsTOF based dataset.



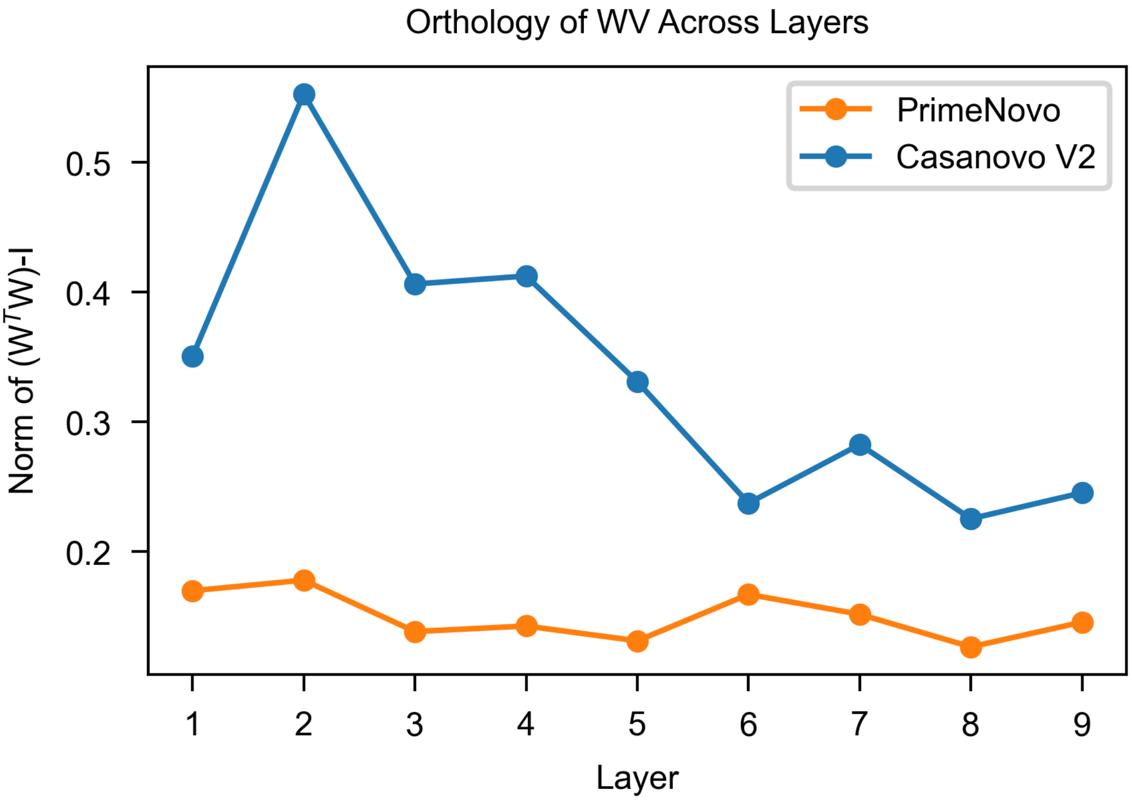
Supplementary Fig. 15: Venn diagram of peptide numbers in different datasets (no modification).



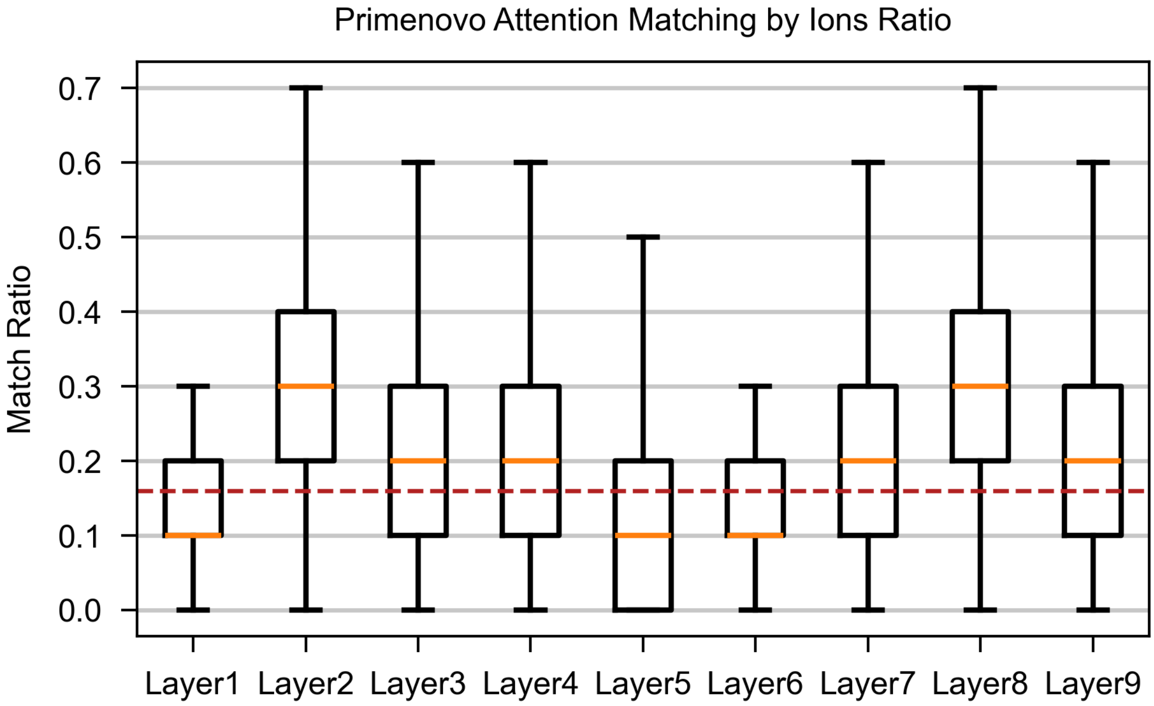
Supplementary Fig. 16: Visualization of the attention score matrix of PrimeNovo's encoder. A striped distribution in the highlighted sections is revealed. The columns that are highlighted represent the ions selectively preferred by the attention mechanism.



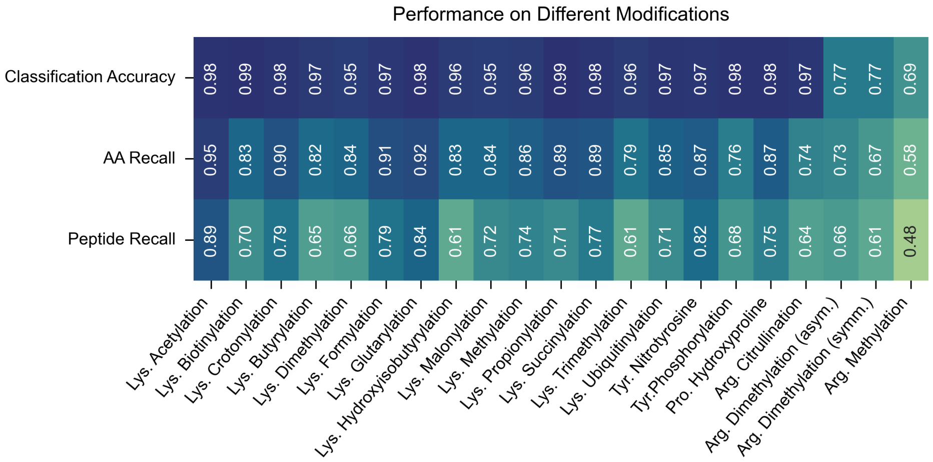
Supplementary Fig. 17: The visualization of the attention score matrix of Casanovo's encoder. There is no obvious striped distribution in the highlighted sections is revealed. The attention mechanism's bias for selecting specific ions is not pronounced.



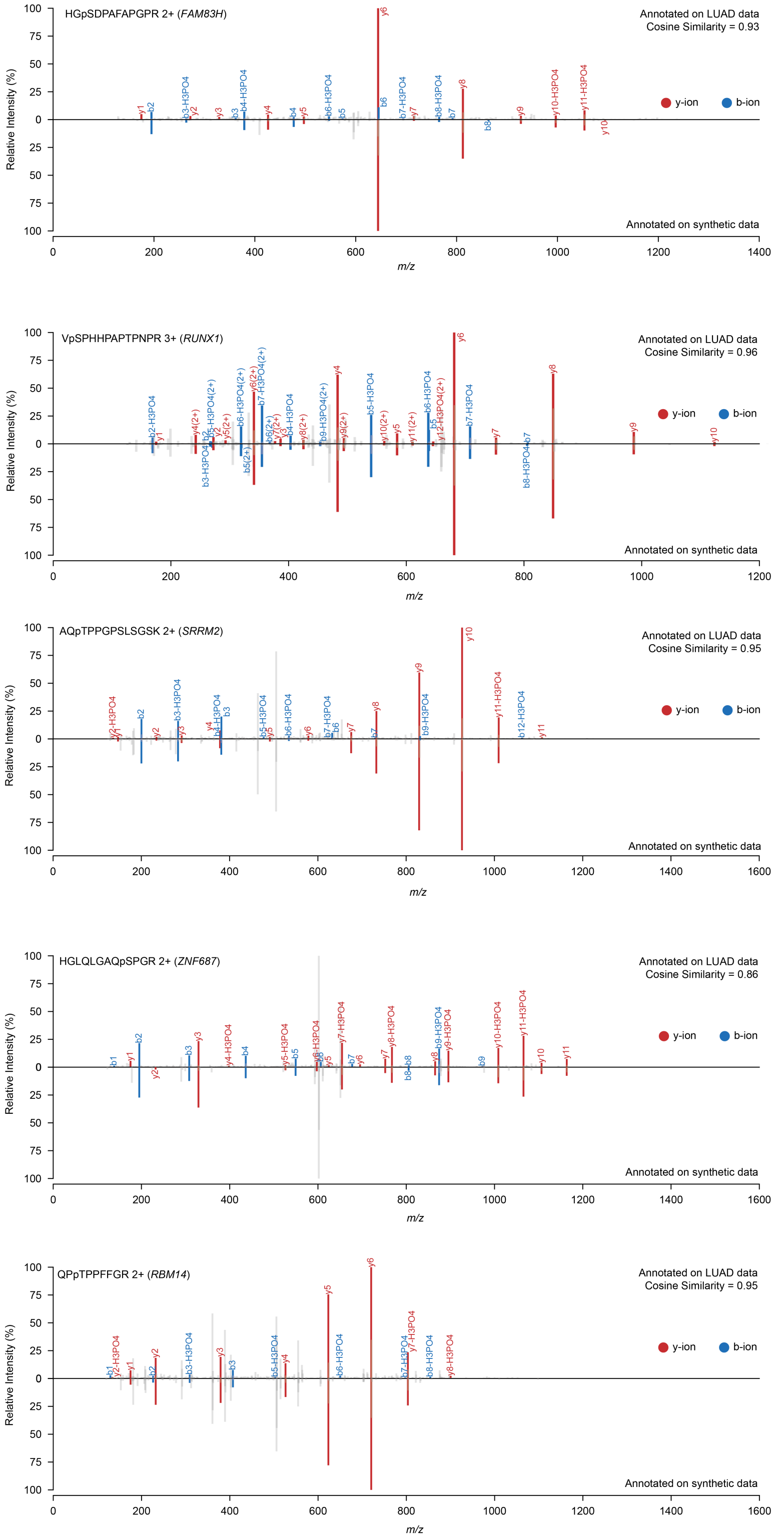
Supplementary Fig. 18: Orthogonality analysis of value matrix projections by the Gram matrix. The y-axis denotes the Frobenius norm of the Gram matrix of the normalized value matrix after subtracting an identity matrix. A smaller value of this norm indicates a better orthogonality of the value matrix projections.



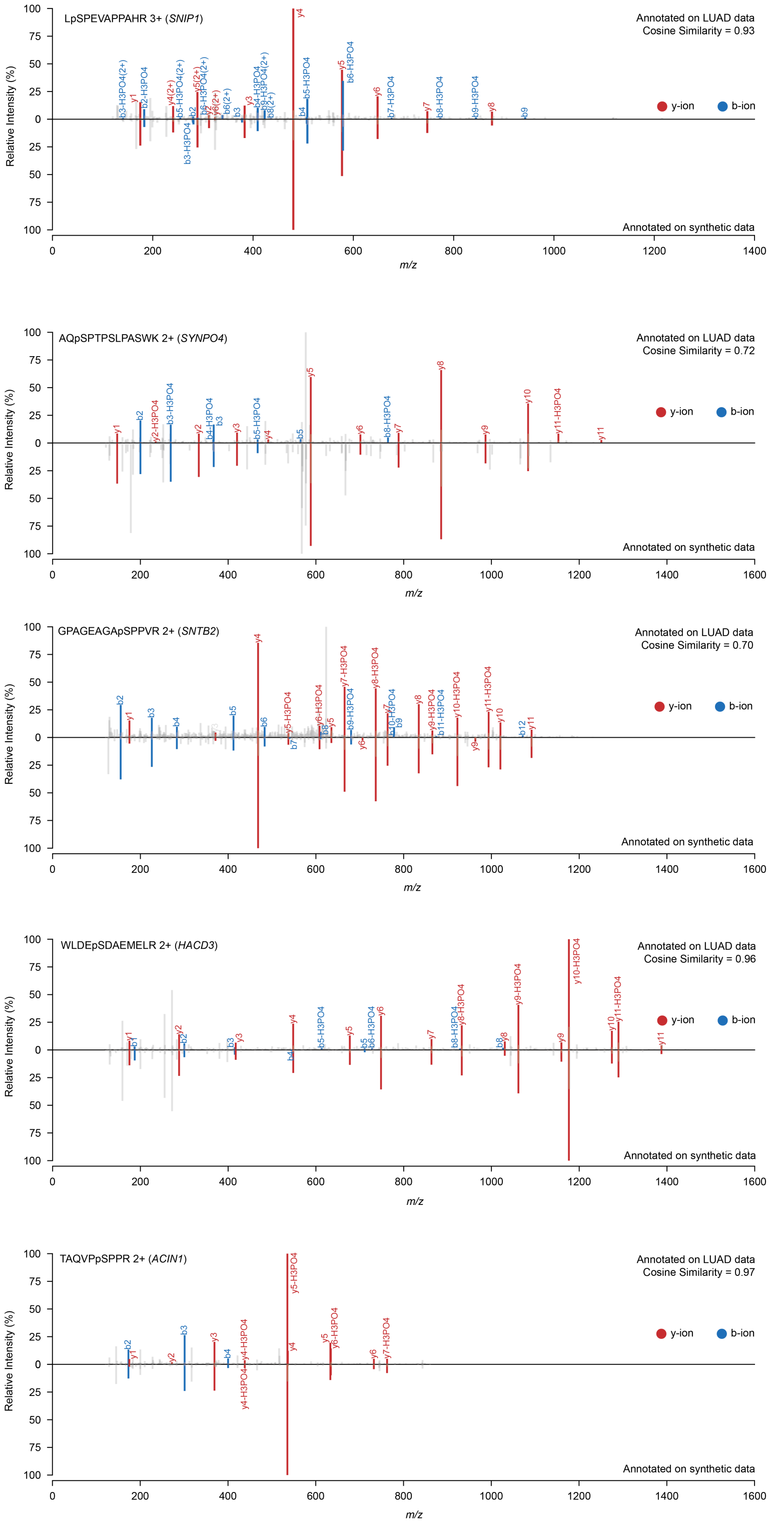
Supplementary Fig. 19: Boxplots to visualize the distribution of the match ratio in the attention score matrices of PrimeNovo's encoder (n=90,594). The yellow line denotes the median, and the upper and lower edges of the box indicate the first and third quartiles, respectively. The short lines at the top and bottom represent the maximum and minimum values. The red dashed line is the proportion of by ions among all peaks.



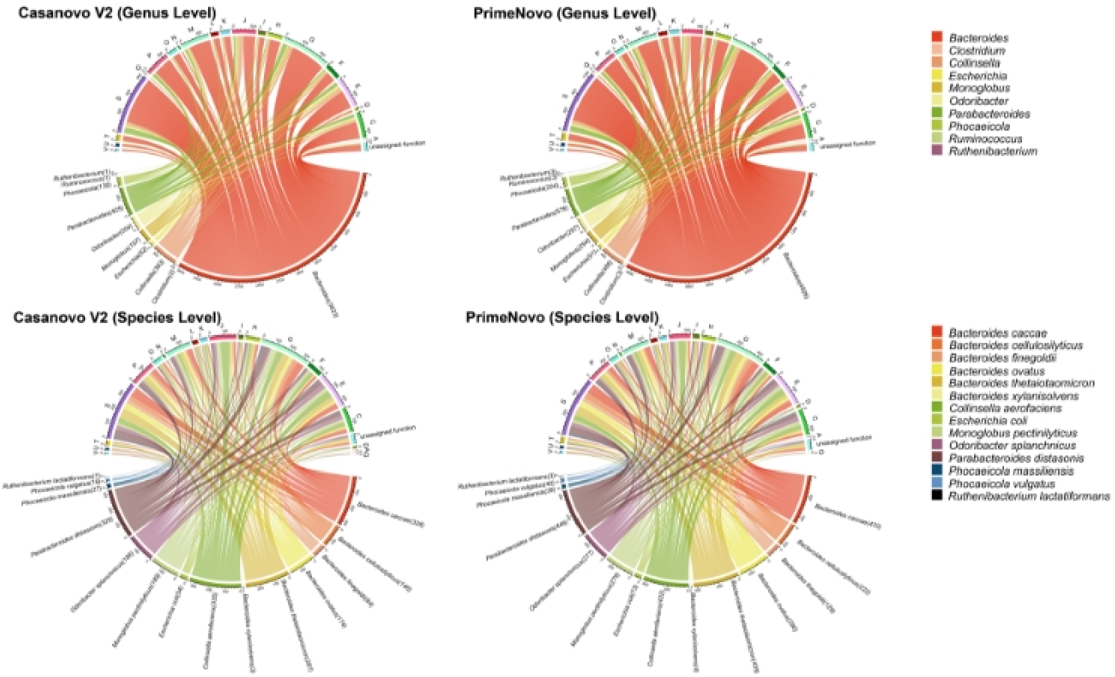
Supplementary Fig. 20: The performance of PrimeNovo for each specific PTM on the 21PTMs dataset.



Supplementary Fig. 21: Comparison of the actual input spectrum and spectrum of 10 synthesized peptides predicted by PrimeNovo. The upper section of each diagram displays the original input spectrum from non-enriched 2020-Cell-LUAD dataset, while the lower section showcases the spectrum from the synthetic data. All overlapping peaks are marked in red and blue for b-y ions.



Supplementary Fig. 22: (Cont.) Comparison of the actual input spectrum and spectrum of 10 synthesized peptides predicted by PrimeNovo. The upper section of each diagram displays the original input spectrum from non-enriched 2020-Cell-LUAD dataset, while the lower section showcases the spectrum from the synthetic data. All overlapping peaks are marked in red and blue for b-y ions.



Supplementary Fig. 23: This figure presents the taxonomic and functional annotation results at the protein level. Function abbreviation can be found in Supplementary Table 11. PrimeNovo notably enhances both taxonomic and functional resolution at the genus and species levels compared to Casanovo V2. Readers are advised to zoom into this figure to observe the detailed differences, as the resolution is too high to display clearly in the standard page view.

Supplementary Tables

Name	Species	Instrument	Accession	No. of PSMs	Description
nine-species benchmark	*V. mungo*	Q-Exactive	MSV000081382	37,775	A standard benchmark previously employed. Analysis conducted through leave-one-species-out cross-validation or a zero-shot approach
	M. musculus			37,021	
	M. mazei			164,421	
	B. subtilis			291,783	
	C. endoloripes			150,611	
	S. lycopersicum			290,050	
	S. cerevisiae			111,312	
	A. mellifera			314,571	
	H. sapiens			130,583	
MassIVE-KB	*H. sapiens*	Q-Exactive	-	30,633,841	Foundational dataset for the development of Casanovo V2 and PrimeNovo algorithms
Proteometools (PT)	*H. sapiens*	Orbitrap Fusion Lumos	PXD004732	28,065,572	Employed as test datasets for PrimeNovo and baseline approaches, analyzed under zero-shot or few-shot fine-tuning scenarios
HCC	*H. sapiens*	Orbitrap Fusion Lumos	IPX0000937000	20,217,331	
IgG1-Human-HC	*H. sapiens*	LTQ Orbitrap	MSV000079801	14,087	
three-species	*A. thaliana*	Orbitrap Q Exactive HF-X	zenodo.8000316	12,222	
	C. elegans	QExactive HF		12,103	
	E. coli	Q-Exactive		12,330	
PXD019483	*H. sapiens*	Q Exactive HFX Orbitrap	PXD019483	1,915,890	
revised nine-species benchmark	nine-species	Q-Exactive	MSV000090982	2,812,194	
PepNet training dataset	multiple	multiple	zenodo.13352403	3,041,555	

Supplementary Table 1: Summary of training and testing datasets in our study.

Name	Species	Instrument	Accession	No. of PSMs	Description
Cell-metaproteome	Microbes	LTQ Orbitrap	MSV000082287	26,457,107	metaproteomics analysis purpose
21PTMs	*H. sapiens*	Orbitrap Fusion Lumos	PXD009449	703,606	PTM analysis purpose
2020-Cell-LUAD	*H. sapiens*	Orbitrap Fusion	IPX0001804000	26,162,410	

Supplementary Table 2: Summary of datasets used in downstream applications.

Species	**Peaks.**	**PepNet**	**Deep.**	**Point.**	**Casa.**	**Casa.V2**	**Prime.**
A. mellifera	0.29	0.28	0.33	0.40	0.41	0.49	**0.60**
B. subtilis	0.39	0.42	0.45	0.52	0.54	0.62	**0.72**
C. endoloripes	0.20	0.24	0.25	0.30	0.33	0.45	**0.53**
H. sapiens	0.28	0.23	0.29	0.35	0.34	0.45	**0.57**
M.mazei	0.36	0.38	0.42	0.48	0.48	0.56	**0.65**
M. musculus	0.20	0.28	0.29	0.36	0.43	0.48	**0.57**
S. cerevisiae	0.43	0.37	0.46	0.53	0.49	0.60	**0.70**
S. lycopersicum	0.40	0.43	0.45	0.51	0.52	0.62	**0.70**
V. mungo	0.36	0.32	0.44	0.51	0.51	0.59	**0.70**
Average	0.32	0.33	0.38	0.44	0.45	0.54	**0.64**

Supplementary Table 3: Peptide recall on the nine-species benchmark dataset. Note: the bold text indicates the highest performance in each row.

Species	**Peaks.**	**Deep.**	**Point.**	**Casa.**	**Casa.V2**	**Prime.**
A. mellifera	0.63	0.63	0.64	0.63	0.71	**0.76**
B. subtilis	0.72	0.74	0.77	0.75	0.79	**0.84**
C. endoloripes	0.59	0.60	0.59	0.60	0.68	**0.72**
H. sapiens	0.64	0.61	0.61	0.59	0.68	**0.72**
M.mazei	0.67	0.69	0.71	0.68	0.76	**0.80**
M. musculus	0.60	0.62	0.63	0.69	0.76	**0.78**
S. cerevisiae	0.75	0.75	0.78	0.68	0.75	**0.80**
S. lycopersicum	0.73	0.73	0.73	0.72	0.79	**0.82**
V. mungo	0.64	0.68	0.73	0.67	0.75	**0.82**
Average	0.66	0.67	0.69	0.67	0.74	**0.79**

Supplementary Table 4: Amino acid precision on the nine-species benchmark dataset. Note: the bold text indicates the highest performance in each row.

Species	**Casa.**	**Casa.V2**	**Prime.**
A. mellifera	0.3578	0.57	**0.70**
B. subtilis	0.4637	0.71	**0.80**
C. endoloripes	0.3055	0.51	**0.59**
H. sapiens	0.4468	0.65	**0.75**
M.mazei	0.4927	0.68	**0.78**
M. musculus	0.4063	0.52	**0.60**
S. cerevisiae	0.4657	0.73	**0.83**
S. lycopersicum	0.4354	0.69	**0.78**
V. mungo	0.3977	0.7	**0.84**
Average	0.4191	0.64	**0.74**

Supplementary Table 5: Peptide recall on the revised nine-species benchmark dataset. Note: the bold text indicates the highest performance in each row.

	Amino acid precision			Peptide recall		
Test dataset	**Casanovo**	**Casanovo V2**	**PrimeNovo**	**Casanovo**	**Casanovo V2**	**PrimeNovo**
HCC	0.0908	0.4968	0.5793	0.0001	0.161	0.3817
IgG1-Human-HC	0.34465	0.605916667	0.693133333	0.1154	0.4065	0.5416
PT	0.4443	0.6284	0.7307	0.2873	0.4543	0.5857
Three-species	0.5633	0.8167	0.8767	0.5037	0.6873	0.7867

Supplementary Table 6: The average performance of PrimeNovo compared to Casanovo and Casanovo V2 across four distinct large-scale MS/MS datasets.

Enzyme	**Amino acid precision**			**Peptide recall**		
	Casanovo	**Casanovo V2**	**PrimeNovo**	**Casanovo**	**Casanovo V2**	**PrimeNovo**
AspN	0.3089	0.5236	0.6059	0.0726	0.2611	0.3593
Chymotrysin	0.2114	0.4794	0.6227	0.0298	0.2493	0.446
GluC	0.2878	0.5214	0.6427	0.0686	0.3133	0.4625
LysC	0.4359	0.7002	0.749	0.2108	0.5224	0.6204
ProteinaseK	0.3844	0.6735	0.7567	0.089	0.4568	0.5972
Trypsin	0.4395	0.7374	0.7818	0.1703	0.5483	0.6649

Supplementary Table 7: The performance of PrimeNovo compared to Casanovo and Casanovo V2 on six different proteolytic enzymes in the IgG1-Human-HC dataset.

Training dataset	Test dataset	Peptide recall
MassIVE-KB	MassIVE-KB	0.8975
MassIVE-KB	HCC	0.3817
MassIVE-KB	PT	0.5887
MassIVE-KB	nine-species (bacillus)	0.7210
nine-species(exclude bacillus)	MassIVE-KB	0.5596
nine-species (exclude bacillus)	HCC	0.0002
nine-species (exclude bacillus)	PT	0.3448
nine-species (exclude bacillus)	nine-species (bacillus)	0.6168
PT	MassIVE-KB	0.7524
PT	HCC	0.5283
PT	PT	0.6998
PT	nine-species (bacillus)	0.2585
HCC	MassIVE-KB	0.4918
HCC	HCC	0.6929
HCC	PT	0.4824
HCC	nine-species (bacillus)	0.2480

Supplementary Table 8: The average peptide recalls when assessing the model's generalization ability.

Data type	Classification accuracy	AA recall	Peptide recall
Tumor tissue	0.98	0.80	0.66
Non-cancerous adjacent tissue	0.98	0.85	0.66

Supplementary Table 9: The performance of PrimeNovo on the phosphorylation dataset 2020-Cell-LUAD.

Charge	**DeepNovo**	**PointNovo**	**Pepnet.**	**PrimeNovo.**
2+	0.513	0.661	0.776	**0.914**
3+	0.242	0.398	0.543	**0.789**

Supplementary Table 10: Peptide recall on the Pepnet test dataset under zero-shot setting. The performance for the different precursor charges is reported separately.

Abbreviations	Functions
-	Function Unassigned
A	Rna Processing And Modification
B	Chromatin Structure And Dynamics
C	Energy Production And Conversion
D	Cell Cycle Control, Cell Division, Chromosome Partitioning
E	Amino Acid Transport And Metabolism
F	Nucleotide Transport And Metabolism
G	Carbohydrate Transport And Metabolism
H	Coenzyme Transport And Metabolism
I	Lipid Transport And Metabolism
J	Translation, Ribosomal Structure And Biogenesis
K	Transcription
L	Replication, Recombination And Repair
M	Cell Wall/Membrane/Envelope Biogenesis
N	Cell Motility
O	Posttranslational Modification, Protein Turnover, Chaperones
P	Inorganic Ion Transport And Metabolism
Q	Secondary Metabolites Biosynthesis, Transport And Catabolism
R	General Function Prediction Only
S	Function Unknown
T	Signal Transduction Mechanisms
U	Intracellular Trafficking, Secretion, And Vesicular Transport
V	Defense Mechanisms
W	Extracellular Structures
X	Mobilome: Prophages, Transposons
Y	Nuclear Structure
Z	Cytoskeleton

Supplementary Table 11: The corresponding abbreviations for COG function entries.

peptide sequence	Paper	Cancer type	Validation method
AQpSPTPSLPASWK	Zheng J, et al. [1]	Lung cancer	Experiment
AQpTPPGPSLSGSK	Sharifnia T, et al. [2]	Osteosarcoma	Experiment
	Li Y, et al. [3]	Prostate cancer	Bioinformatics analysis
	Gao L, et al. [4]	Colorectal cancer	Experiment & Bioinformatics analysis
GPAGEAGApSPPVR	Hungermann D, et al. [5]	Breast cancer	Bioinformatics analysis
	Panea RI, et al. [6]	Burkitt lymphoma	Bioinformatics analysis
	Li Z, et al. [7]	Melanoma	Bioinformatics analysis
HGpSDPAFAPGPR	Gan J, et al. [8]	Non-small cell lung cancer	Bioinformatics analysis
HGLQLGAQpSPGR	Sayeeram D, et al. [9]	Lung adenocarcinoma	Bioinformatics analysis
	Li M, et al. [10]	Lung adenocarcinoma	Experiment & Bioinformatics analysis
LpSPEVAPPAHR	Fujii M, et al. [11]	Non-small cell lung cancer	Experiment
	Jeon HS, et al. [12]	Non-small cell lung cancer	Experiment & Bioinformatics analysis
	Zhong Y, et al. [13]	Non-small cell lung cancer	Experiment & Bioinformatics analysis
	Wu J, et al. [14]	Non-small cell lung cancer	Experiment
LGpSFGSLTR	Windhorst S, et al. [15]	Lung adenocarcinoma	Experiment
	Pal J, et al. [16]	Non-small cell lung cancer	Bioinformatics analysis
	Wu J, et al. [17]	Lung squamous cell carcinoma	Bioinformatics analysis
	Wu ZX, et al. [18]	Lung adenocarcinoma	Bioinformatics analysis
	Ding H, et al. [19]	Non-small cell lung cancer	Bioinformatics analysis
	Xuan Y, et al. [20]	Lung cancer	Bioinformatics analysis
	Qin C, et al. [21]	Lung adenocarcinoma	Bioinformatics analysis
	Yang Z, et al. [22]	Lung adenocarcinoma	Bioinformatics analysis
	Hui G, et al. [23]	Non-small cell lung cancer	Bioinformatics analysis
QApSLELPSMAVASTK	Hu Z, et al. [24]	Non-small cell lung cancer	Experiment & Bioinformatics analysis
	Park SL, et al. [25]	Lung cancer	Bioinformatics analysis
QPpTPPFFGR	Sui Y, et al. [26]	Non-small cell lung cancer	Experiment
	Hu, Y.et al. [27]	Lung adenocarcinoma	Experiment
TAQVPpSPPR	Shu Y, et al. [28]	Lung adenocarcinoma	Experiment
	Ujifuku K, et al. [29]	Non-small cell lung cancer	Experiment
	Bacolod MD, et al. [30]	Lung cancer	Bioinformatics analysis
VpSPHHPAPTPNPR	Ramsey J, et al. [31]	Lung adenocarcinoma	Experiment & Bioinformatics analysis
	Kim Y, et al. [32]	Lung adenocarcinoma	Experiment
	Ma H, et al. [33]	Non-small cell lung cancer	Experiment & Bioinformatics analysis
WLDEpSDAEMELR	Barupal DK, et al. [34]	Breast cancer	Bioinformatics analysis
	Xu Y, et al. [35]	Breast cancer	Bioinformatics analysis

Supplementary Table 12: Detailed information about cancer research for the proteins of the 12 synthesized phosphopeptides

Peptide information			Matched protein information			Phosphorylation site
Modified peptide	Matched peptide	Similarity	Accession	Gene	Protein	HTP** / LTP***
AQpSPTPSLPASWK	AQSPTPSLPASWK	0.72	Q9UMS6	SYNPO2	Synaptopodin-2	43/0
AQpTPPGPSLSGSK	AQTPPGPSLSGSK	0.95	Q9UQ35	SRRM2	Serine/arginine repetitive matrix 2	137/0
GPAGEAGApSPPVR	GPAGEAGASPPVR	0.70	Q13425	SNTB2	Beta-2-syntrophin	86/1
HGpSDPAFAPGPR	HGSDPAFAPGPR	0.93	Q6ZRV2	FAM83H	Protein FAM83H	72/0
HGLQLGAQpSPGR	HGLQLGAQSPGR	0.86	Q8N1G0	ZNF687	Zinc finger protein 687	45/0
LpSPEVAPPAHR	LSPEVAPPAHR	0.93	Q8TAD8	SNIP1	Smad nuclear-interacting protein 1	70/1
LGpSFGSLTR	LGSFGSITR	0.96	Q14315	FLNC	Filamin-C	210/3
QApSLELPSMAVASTK	QASIELPSMAVASTK	0.97	P33241	LSP1	Lymphocyte-specific protein 1	216/3
QPpTPPFFGR	QPTPPFFGR	0.95	Q96PK6	RBM14	RNA-binding protein 14	251/0
TAQVPpSPPR	TAQVPSPPR	0.97	Q9UKV3	ACIN1	Apoptotic chromatin condensation inducer in the nucleus	118/0
VpSPHHPAPTPNPR	VSPHHPAPTPNPR	0.96	Q01196	RUNX1	Runt-related transcription factor 1	24/0
WLDEpSDAEMELR	WLDESDAEMELR	0.96	Q9P035	HACD3	Very-long-chain (3R)-3-hydroxyacyl-CoA dehydratase 3	257/1

Supplementary Table 13: Detailed annotations for all the 12 synthesized phosphopeptides

Supplementary Notes

SI Note 1: Dataset Details

Supplementary Table 1 provides a comprehensive overview of the training and testing datasets utilized in our study. This includes the Massive-KB dataset, which was used for training PrimeNovo, and the nine-species benchmark dataset, employed for training PrimeNovo CV and for conducting comparative evaluations against baseline models. Following the training phase, PrimeNovo was subsequently evaluated on a diverse set of datasets, encompassing PT, HCC, IgG1-human-HC, three-species, PXD019483, and the revised nine-species benchmark datasets. Except for the nine-species benchmark dataset and PepNet training dataset, all other datasets were downloaded as raw files along with identification results. For these datasets, we first converted the raw files into mgf files by MSConvert (version 3.0), then incorporated the peptide sequences, charge states, and precursor ion masses from the identification results into these mgf files. For datasets where the distinction between training and testing sets was not made during download, we devised a new strategy for partitioning into training and testing sets. Specifically, we randomly selected a certain number of PSMs from the dataset for testing, while reserving the remaining annotated data for training and fine-tuning. To ensure that none of spectra with the same peptides appear in both the testing set and the training set, we implemented a selection method where the inclusion of any PSM in the test set automatically results in the selection of all other PSMs with the same true peptide sequence to the test set. This prevents any leakage of answers during fine-tuning/training and guarantees a fair testing environment. The detailed generation processes of the training and testing dataset are as follows:

1. **nine-species benchmark** dataset [36]: It was originally curated by Tran *et al.* during the development of DeepNovo [36], comprises a diverse assortment of data from nine distinct species. It was meticulously compiled by aggregating contributions from multiple research teams, with the aim of minimizing biases associated with species and laboratory sources. This dataset has since gained popularity as a preferred choice for evaluating various de novo sequencing algorithms, including Casanovo [37] and PointNovo [38], utilizing the leave-one-species-out cross-validation method established by DeepNovo. In our study, we accessed this dataset from the MassIVE repository (MSV000090982) and adopted the same cross-validation strategy to ensure a fair and consistent comparison. The mgf files were directly downloaded from the MassIVE repository (identifier: MSV000081382), shared by the authors of the DeepNovo paper. We did not undertaken any re-partitioning for this dataset and used downloaded test set directly for testing our model.

2. **MassIVE-KB** [39]: This dataset played a pivotal role in training our model. It encompasses an extensive collection of over 2.1 million precursors derived from 19,610 proteins. This dataset was meticulously assembled, drawing from more than 31 TB of human data originating from 227 public proteomics datasets. Notably, it upholds rigorous false discovery rate controls, ensuring a comprehensive and robust training environment for our model. The raw files (in mzML format) and the filtered identification results from the "All Candidate library spectra" section of the MassIVE Knowledge Base spectral library v1 (https://massive.ucsd.edu/ProteoSAFe/static/massive-kb-libraries.jsp) were obtained by downloading from the MassIVE repository. We did not split this dataset and used all psms as training set.

3. **ProteomeTools (PT)** [40]: The PT dataset, chosen for PrimeNovo's performance evaluation and fine-tuning, constitutes a carefully curated assembly of synthetic human peptides. As a prominent component of the Proteometools project, it comprises in excess of 330,000 synthetic tryptic peptides, encompassing a diverse spectrum of human gene products. The presence of well-established peptide sequences within this dataset serves as a reliable foundation for evaluating the algorithm's precision and accuracy. The raw files and MaxQuant identification results were obtained by downloading from the PRIDE [41] repository by PXD004732. We applied the training and testing splitting strategy as described above, selecting 58,000 PSMs as the test set and using the remaining data for training and fine-tuning.

4. **HCC** [42]: From a real-world application standpoint, the HCC dataset, centered on proteomics data from early-stage human hepatocellular carcinoma (HCC) patients, was incorporated. This dataset provides a distinctive insight into the proteomics of both tumor and non-tumor tissues, facilitating an evaluation of our model's generalizability in a clinical context.

5. **IgG1-Human-HC** [43]: The IgG1-Human-HC dataset, employed to assess the generalizability of de novo sequencing models, comprises human antibody sequences that have undergone processing with various enzymes, including trypsin, chymotrypsin, and others. This dataset holds particular importance due to its applicability in identifying novel or unfamiliar protein sequences, especially in the realm of immunotherapy antibodies where variable sequences are often unavailable, rendering traditional database search algorithms ineffective. Hence, we utilized this dataset to evaluate our model's proficiency in unraveling the amino acid sequences of unknown antibodies.

6. **three-species** [44]: We employed the three-species dataset, made available by GraphNovo [44], which comprises samples from *Arabidopsis thaliana*, *Caenorhabditis elegans*, and *Escherichia coli*, for our comparative analysis. This dataset, subjected to trypsin digestion and analyzed using state-of-the-art mass spectrometry techniques, allowed us to conduct a thorough performance comparison between PrimeNovo and GraphNovo.

7. **PXD019483** [45]: In order to facilitate a comparison with the recently published PepNet [46], we acquired the publicly available test dataset PXD019483, which is the testing set used by PepNet. This dataset, consisting of extensive human proteomics data, is accessible through the PRIDE repository under the identifier PXD019483. In line with PepNet's approach, we filtered the spectra to include only those with charge states 2+ and 3+, as identified by MaxQuant with a false discovery rate (FDR) threshold of less than or equal to 1%, and with precursor mass differences of no more than 10 ppm. The raw files and MaxQuant identification results were obtained by downloading from the PRIDE [41] repository by PXD019483. This dataset was used as the test dataset for PepNet [46]. We did not undertaken any re-partitioning for this dataset and used downloaded test set directly for testing our model.

8. **Revised nine-species benchmark** [37] : The dataset was acquired by the authors of Casanova V2 [37] through a process that involved downloading the raw files of the nine-species benchmark dataset and subsequently reanalyzing them using Crux version 4.1 [47]. Following this analysis, peptides shared between species were removed. The resulting dataset, known as the new nine-species benchmark, encompasses approximately 2.8 million PSMs and is derived from a total of 343 raw files. For our study, we directly obtained the identification results that were shared by the authors.

9. **PepNet training dataset** [46]: The mgf files were provided by the authors [46] after we contacted them. We downloaded their training and testing data and made no further changes or splitting before using them for training/fine-tuning and testing.

Supplementary Table 2 presents a comprehensive overview of the datasets utilized for downstream applications in our investigation.

1. **Cell-metaproteome** [48]: This dataset was utilized to assess the performance of our model on complex samples containing multiple coexisting species, such as microbial communities that often lack reference sequences. We accessed a publicly available human-gut-derived bacterial metaproteomics dataset, which was sampled from the human gastric organ and processed using an LTQ Orbitrap mass spectrometer, resulting in the generation of MS/MS spectra. These MS/MS spectra were subjected to analysis using MyriMatch v.2.2. We obtained approximately 26 million PSMs by downloading the identification results generously shared by the authors of the analysis [48]

2. **21PTMs** [49]: The 21PTMs dataset represents a publicly available benchmark dataset notable for containing the most extensive variety of post-translational modifications (PTMs) to date. This dataset comprises approximately 5,000 peptides, which in turn represent 21 different naturally occurring human PTMs, encompassing modifications of lysine, arginine, proline, and tyrosine side chains, along with their corresponding unmodified counterparts. the raw files and MaxQuant identification results were obtained by downloading from the PRIDE [41] repository by PXD009449. Following our splitting strategy, we split all 24 portions using a 95% training to 5% testing ratio for training and fine-tuning. Each PTM portion was then combined with its corresponding non-PTM portion. For example, the Trimethyl of K amino acid training data was combined with Non-PTM of K amino acid training data to form the training set for Trimethyl of K amino acid. Similarly, Trimethyl of K amino acid testing data was combined with Non-PTM of K amino acid testing data to evaluate the accuracy of PTM identification.

3. **2020-Cell-LUAD** [50]: To assess our model's ability to infer phosphorylated peptides, we obtained a publicly available phosphoproteomics dataset, known as 2020-Cell-LUAD [50]. This dataset was specifically designed for the study of lung adenocarcinoma (LUAD) and was compiled from the tumors of 103 LUAD patients and their corresponding non-cancerous adjacent tissues. It offers both phospho-enriched data (the phosphoproteomic data) and non-enriched data (the proteomic data), facilitating a comprehensive analysis of phosphorylation events.

SI Note 2: Additional model details

CTC loss calculation with dynamic programming.

As we mentioned in “Section Methods. Definition of CTC loss” of our main manuscript, CTC loss hinges on calculating the total probability of all paths that lead to a desired target sequence. However, enumerating all paths in the model output to calculate the $P(A|\mathcal{S})$ is too time-consuming as there are $(\mathrm{m}+1)^{\mathbf{t}}$ different decoding paths, with m being the total number of amino acids adding one ϵ token. To enable efficient calculation, dynamic programming is used during the training using CTC. Let denote $\alpha(\tau, r)$ as the probability of generating $A_{1:r}$, the first r amino acids in the ground truth sequence A, using only the output positions up to τ, formally:

$$\alpha(\tau, r) = P(A_{1:r}|\mathcal{S}) = \sum\nolimits_{\mathbf{y}_{1:\tau}:\Gamma(\mathbf{y}_{1:\tau})=A_{1:r}} \sum_{\mathrm{y}_i \in \mathbf{y}}^{\tau} P(\mathrm{y}_i|\mathcal{S}) \tag{1}$$

Consequently, $\alpha(\tau, 0)$ is the probability of generating ϵ tokens at all first τ positions in the model output, as $|\Gamma(\mathbf{y}_{1:\tau})| = 0 \Rightarrow \mathbf{y}_{1:\tau} = (\epsilon, \epsilon, \cdots, \epsilon)$. And $\alpha(1, 1)$ is the probability of generating the exact first true amino acid at the first position. We initialize our DP as follows:

$$\alpha(\tau, 0) = P(\mathrm{y}_1 = \epsilon|\mathcal{S}) * P(\mathrm{y}_2 = \epsilon|\mathcal{S}) * \cdots * P(\mathrm{y}_\tau = \epsilon|\mathcal{S}), \;\; \forall 1 \leq \tau \leq \mathbf{t} \tag{2}$$

$$\alpha(1, 1) = P(\mathrm{y}_1 = \mathrm{a}_1|\mathcal{S}) \tag{3}$$

$$\alpha(1, r) = 0, \;\; \forall r > 1 \tag{4}$$

For the convenience of explanation, we further decompose the $\alpha(\tau, r)$ into $\alpha(\tau, r|\mathrm{y}_\tau = \epsilon) + \alpha(\tau, r|\mathrm{y}_\tau \neq \epsilon)$ by the law of total probability. We then recursively calculate $\alpha(\tau, r)$ using the following rule:

$$\alpha(\tau, r) = \begin{cases} (\alpha(\tau-1, r|\mathrm{y}_{\tau-1} \neq \epsilon) + \alpha(\tau-1, r-1)|\mathrm{y}_{\tau-1} = \epsilon) * P(\mathrm{y}_\tau = \mathrm{a_r}|\mathcal{S}) & \\ \qquad +(\alpha(\tau-1, r)) * P(\mathrm{y}_{\tau-1} = \epsilon|\mathcal{S}), & \text{if } \mathrm{a}_r = \mathrm{a}_{r-1} \\ (\alpha(\tau-1, r|\mathrm{y}_\tau \neq \epsilon) + \alpha(\tau-1, r-1)) * P(\mathrm{y}_{\tau-1} = \mathrm{a_r}|\mathcal{S}) & \\ \qquad +(\alpha(\tau-1, r)) * P(\mathrm{y}_\tau = \epsilon|\mathcal{S}), & \text{if } \mathrm{a}_r \neq \mathrm{a}_{r-1}. \end{cases} \tag{5}$$

When $\mathrm{a}_r = \mathrm{a}_{r-1}$, either by adding the a_r to decoded sequence up to$\tau - 1$ time step $\mathbf{y}_{\tau-1}$ which can be reduced to $A_{1:r-1}$, with $\mathrm{y}_{\tau-1} = \epsilon$, or repeating the last token of $\mathbf{y}_{\tau-1}$ which can be reduced to $A_{1:r}$ can both yield $A_{1:r}$. Otherwise, when $\mathrm{a}_r \neq \mathrm{a}_{r-1}$, we wouldn’t need to consider $\mathrm{y}_{\tau-1} = \epsilon$ when adding a_r, with everything else being the same.

Then, our loss function $\mathcal{L}_{\text{ctc}}$ is:

$$\mathcal{L}_{\text{ctc}} = -P(A|\mathcal{S}) = -\alpha(\mathbf{t}, |A|), \tag{6}$$

as we are trying to maximize the total probability of all paths that can be reduced to A. This is precisely the value of $\alpha(\mathbf{t}, |A|)$ as we defined above. We make it negative to make it a minimization problem.

It’s worth noting that dynamic programming is employed to efficiently compute the CTC loss by identifying all valid paths without affecting the backpropagation algorithm. Instead of naively enumerating all possible paths that can be reduced to a target label (which grows exponentially), dynamic programming simplifies the process by applying CTC recursion rules, reducing complexity. This method finds all valid paths more efficiently but does not alter the loss calculation, which remains the sum of probabilities for the correct paths.

CUDA algorithm for PMC unit

Algorithm 1 CUDA PMC Algorithm.

```
procedure $PMC\_inference(lock, aa\_mass, grid\_size)$
    $w \leftarrow blockIdx.x$
    $dim \leftarrow blockDim.x$
    $h \leftarrow threadIdx.x$                    ▷ The [w,h] grid is calculated by this thread.
    if $w = 0 \vee h = 0$ then
        return
    end if                    ▷ The zero position and the zero mass do not need calculation.
    if $w > maxW$ then
        $lock[w \times dim + h] \leftarrow 1$
        return
    end if
    if $h = 1$ then
        $Calculate_{h}1(w, h)$    ▷ Find the amino acid which could be placed in [w,h] grid, with the maximum log probability.
        $lock[w \times dim + h] \leftarrow 1$
        return
    end if
    while $lock[w \times dim + h - 1] \neq 1$ do
    end while                    ▷ Waiting the dependency lock to release.
    $PutIn\epsilon OrSim()$
    for $i \leftarrow 1$ to $AA\_num - 1$ do
        $minw \leftarrow \lfloor w \times grid\_size - aa\_mass[i]/grid\_size \rfloor$
        for each $minw$ and $minw + 1$ do
            while $lock[minw \times dim + h - 1] \neq 1$ do
            end while                    ▷ Waiting for the dependency lock to release.
            $PutInNumiAA()$
        end for
    end for
    $lock[w \times dim + h] \leftarrow 1$
    return
end procedure
```

Our CUDA algorithm begins by constructing a two-dimensional dynamic table. In this table, we store the most probable partial sequence at time step t, ensuring that it satisfies the mass constraint for each cell. Each cell independently carries out its calculations. Since the computation of each cell relies on all the cells in its upper left column, we employ a lock mechanism to prevent computational race conditions. Starting from the first column, we select the amino acid whose mass aligns with the mass constraint for each row, filling in the respective cell. After completing its computation, the cell releases its dependency lock, allowing the subsequent cell to utilize its results for further computation. Throughout the column, we perform recursive updates in accordance with the PMC rule. Once all cells in the table have completed their calculations, we can retrieve the value in the last cell, which represents the final sequence we seek.

Model configurations and training process

PrimeNovo utilizes an encoder-decoder transformer network configuration. More precisely, both the encoder and decoder components consist of 9 multi-head attention layers. Each attention layer within the network is composed of 8 heads. The hidden embedding dimension for our model is set to 400.

For the training process, we employed a Cosine learning rate scheduler in conjunction with the AdamW optimizer. To mitigate the risk of excessive overfitting, we set the dropout rate at 0.18.

During inference, we use the PMC with precision e of 0.1Da. This precision constraint ensures that the decoded peptide always falls within a range of 0.1 Da from the precursor mass indicated in the spectrum.

Our model utilizes a vocabulary of size 28, including 27 amino acid-related tokens: ”G”, ”A”, ”S”, ”P”, ”V”, ”T”, ”C+57.021”, ”L”, ”I”, ”N”, ”D”, ”Q”, ”K”, ”E”, ”M”, ”H”, ”F”, ”R”, ”Y”, ”W”, ”M+15.995”, ”N+0.984”,

”Q+0.984”, ”+42.011”, ”+43.006”, ”-17.027”, and ”+43.006-17.027”, along with one CTC placeholder token ϵ.

The feedforward layer in the Transformer has a dimension of 1024. We have fixed the maximum generational length to 40 tokens, which will be reduced using CTC post-reduction. Training is conducted over a maximum of 150 epochs, with each epoch covering the sampled 1 million data in the training dataset. We use the nine-species Bacillus validation set for all training validation, as we believe it can reflect the model's generalization on out-of-distribution data. Training is stopped when the validation loss converges, which is indicated by minimal changes (+- 0.05) or an increase (overfitting). The last checkpoint before convergence is used for testing. Our PMC unit functions as the post-generational decoding module and does not require additional training or tuning.

During Training, we use a base learning rate of 0.0004, with a linear warmup over the first 5 epochs. From the 6th epoch onward, a cosine learning rate scheduler with a 0.5 decay factor is applied. The total batch size is 3200, distributed across 8 GPUs. We use the Adam optimizer with $\beta_1 = 0.9$ and $\beta_2 = 0.999$, and no additional weight decay (set to 0).

Non-autoregressive versus autoregressive

As our model is fundamentally different from the traditional autoregressive architect, we further clarify the differences in architect design to facilitate future research. The differences between autoregressive (AR) models and non-autoregressive (NAR) models, specifically in the context of Transformer architecture (AT vs NAT), are detailed as follows, with each counterpart outlined:

1. **Decoder Input and Decoder Masking:** During the training of AT models, all tokens are fed into the Transformer decoder to parallelize the training. However, because the modeling objective for autoregression is the conditional probability $P(a_{i+1}|a_{1:i})$, exposing all tokens to the AR model would lead to information leakage and fail to model the conditional probability correctly. Therefore, AR model training employs ”causal masking.” This mask forces each attention mechanism to focus only on the preceding tokens by masking the values after each token with $-\infty$, effectively zeroing out those values after applying the softmax function, as $\text{softmax}(-\infty) = 0$. In contrast, during NAT model training, none of the true tokens are fed into the decoder, and the model learns to generate all positions from ”scratch.” Since the model's objective is not a conditional probability like in AT models, the decoder does not use ”masking” to prevent information leakage. This approach also allows the model to be bidirectional, as each position can attend to both preceding and succeeding positions.

2. **Token-Level Loss vs. Sequence-Level Loss:** AT model training uses cross-entropy loss on each token, meaning that the ”reward” signal is applied to each correctly predicted token. On the other hand, NAT models apply the connectionist temporal classification (CTC) loss, as detailed in the manuscript. CTC's reward signal is applied to entire ”paths,” meaning that partially predicted tokens do not receive a reward. Only entirely correct paths receive a positive reward, while any incorrect path receives a negative reward. This loss mechanism better enhances global sequence coherence.

3. **Decoding Dynamics:** Since NAT models generate probabilities independently at each position, the resulting probability distribution is static. However, in AT models, each token's probability is conditional, making it dynamic. At each position of an AT model, the selection of a particular token affects the distribution of all subsequent positions. For instance, selecting ”good” at the first position might increase the probability of ”bye” in the second position, while selecting ”see” might increase the probability of ”you” in the second position. This dependency prevents the model from having accurate global control over the generational output, as any modification to a previous token requires re-decoding all subsequent tokens due to the shifted distribution. In contrast, because the NAR model's probability is static, any position can be modified without such concerns, which is the motivation behind the Precise Mass Control (PMC) unit.

SI Note 3: Peptide annotation for metaproteomic research

The quality assessment of decoded peptides from both PrimeNovo and Casanovo V2, when applied to unlabeled spectra, encompassed various facets, including peptide counts, taxonomic annotations, and functional annotations. Taxonomic annotation was conducted utilizing Unipept (version 5.1.2, employing UniProt 2023.04), while functional annotation was carried out using eggNOG-mapper v2 (utilizing the eggNOG 5 database). To construct the taxon-function network, we employed a protein-peptide bridge approach with the assistance of a pre-existing Python script [51].

SI Note 4: Additional results on speed of PrimeNovo

We observed that a larger batch size does not always result in faster inference. We discuss the cases where post-modification processes—PMC for PrimeNovo and beam search for Casanovo v2—are enabled, compared to when these processes are disabled.

- When both post-modification processes were included, we found that the optimal inference speed was achieved with a batch size between 10 and 50 for both models (Supplementary Figure 2 right-hand side). Increasing or decreasing the batch size beyond this range resulted in reduced speed. Specifically, Casanovo with beam search reached its peak performance at a batch size of 10, processing 8 spectra per second. PrimeNovo with PMC achieved its best performance at a batch size of 50, processing 227 spectra per second—28 times faster than Casanovo V2. The small optimal batch size is due to the heavy reliance of both post-decoding algorithms on CPU processing after our investigation, which involves many non-parallelizable components and conditional branches (e.g., if statements). Larger batch sizes lead to CPU offloading inefficiencies beyond the optimal threshold.
- When both post-decoding algorithms were disabled, we observed a continuous increase in speed with larger batch sizes (Supplementary Figure 2 left hand side). Specifically, both models reached their speed plateau at a batch size more than 1000. In this case, PrimeNovo without PMC was able to decode up to 714 spectra per second, while Casanovo without beam search reached 212 spectra per second. This is because, without the post-decoding algorithms, both models primarily rely on GPU processing, which efficiently handles large amounts of parallel work. Larger batch sizes improve GPU utilization.

Additionally, we tested other factors such as the number of spectra peaks and GPU card numbers and found minimal impact on speed, indicating that batch size is likely the primary factor affecting performance.

SI Note 5: Additional results on nine-species benchmark dataset

PrimeNovo extends its strong performance to the revised nine-species benchmark dataset. The widely accepted nine-species benchmark dataset has faced criticism due to its limited amount of testing data. In an effort to address this limitation and provide a more comprehensive evaluation of model performance, Yilmaz et al. [37] introduced a new nine-species test dataset in conjunction with the Casanovo V2 model. This updated test set offers a significantly larger volume of data points compared to its predecessor and features a broader distribution of data across each species. In our study, we conducted tests on this new nine-species dataset using PrimeNovo and compared its performance with that of Casanovo V2. As illustrated in Supplementary Fig. 4, Fig. 5 and Table 5, PrimeNovo consistently outperforms Casanovo V2 across all test species, achieving an average peptide recall rate that surpasses Casanovo V2 by 10%. This outcome reinforces the superior accuracy and efficacy of our model.

PrimeNovo Achieves State-of-the-Art Amino Acid Level Accuracy. While sequence-level accuracy, measured by peptide recall, remains a paramount metric for evaluating model performance, amino acid (AA) level accuracy holds significance for de novo models. Partially correct amino acids can often provide valuable biological insights and aid in identifying essential components within a peptide. In our comparison of AA level accuracy between PrimeNovo and all other baseline models using the nine-species benchmark dataset, PrimeNovo consistently outperforms its counterparts in AA precision, as illustrated in Supplementary Fig. 6.

Furthermore, we delve into the AA level accuracy at various confidence levels of the model. We rank predictions from both PrimeNovo and Casanovo V2 based on their confidence scores and assess amino acid accuracy within each confidence level range. As demonstrated in Supplementary Fig. 7, amino acid accuracy increases with higher confidence scores for both PrimeNovo and Casanovo. This indicates that model confidence scores serve as meaningful indicators of token-level accuracy. PrimeNovo exhibits superior AA level accuracy in the high-confidence region compared to Casanovo V2, highlighting its ability to make higher-quality token-level predictions when confident in its own predictions. For detailed metrics regarding peptide level accuracy and amino acid level accuracy on the nine-species benchmark dataset, please refer to Supplementary Tables 3 and 4.

We also compare the AA recall performance by PrimeNovo and Casanovo V2. As shown in Supplementary Figure 3.a-b, PrimeNovo consistently outperforms Casanovo V2 in both AA precision and AA recall across all species. For both models, we observe that AA recall and AA precision are very similar, with AA recall being, on average, 0.5%–1% higher than AA precision. We further investigated the difference between AA precision and AA recall using statistical testing. As shown in Supplementary Figure 3.c-d, the Mann-Whitney U test indicates no significant difference (p-value < 0.05) between recall and precision across all species, with the mean recall value slightly higher (less than 1%) than precision. This lack of significance is consistent for both PrimeNovo and Casanovo V2, with PrimeNovo showing even smaller differences between recall and precision (Supplementary Figure 3.d) and higher confidence level (lower p-value) in rejecting the non-difference using Mann-Whitney U test. In conclusion, these two metrics do not show substantial differences in representing Amino Acid level prediction performance and can be used interchangeably. Thus, our reported AA precision across all datasets serves as a robust evaluation of AA-level accuracy.

SI Note 6: Additional results on other test datasets

PrimeNovo Demonstrates Superior Amino Acid Prediction Across Diverse Unseen Datasets. We also conducted a comparison of PrimeNovo with other de novo algorithms based on amino acid (AA) level precision across three additional datasets. Notably, all models were evaluated under the zero-shot setting, without any fine-tuning on the target data distribution. As depicted in Supplementary Fig. 8a and 8b, along with the results provided in Supplementary Table 6, PrimeNovo excels in AA level accuracy, surpassing the previous best model, in both the HCC and PT datasets. Furthermore, PrimeNovo outperforms all baseline models on the three-species dataset across all tested species, as indicated in Supplementary Table 6. To provide a more detailed analysis, we compared AA level accuracy within each confidence range, having ranked the peptides based on their output confidence scores. Our observations reveal that PrimeNovo consistently achieves higher accuracy levels across all confidence ranges compared to Casanovo V2. These results collectively highlight PrimeNovo's superior capability in accurately predicting amino acids across a range of unseen datasets, even when operating in a zero-shot setting.

PrimeNovo Demonstrates Robust Performance in the Face of Missing Peaks and Varied Peptide Lengths. We conducted an analysis of PrimeNovo's performance, as measured by peptide recall, when dealing with peptides of different lengths and input spectra containing missing data. To accomplish this, we utilized the PT dataset and categorized the MS/MS data based on the length of the target peptide. Within each length category, we quantified the number of missing peaks in each spectrum, following the methodology outlined in previous research [52]. Subsequently, we reported the peptide-level recall under a zero-shot setting for each combination of target peptide length and missing peak count. As observed in Supplementary Fig. 12 and Fig. 10, both models experience a drop in accuracy (indicated by lighter colors) when more peaks are missing from the spectrum, regardless of the peptide length. However, the heat map illustrating the performance gap between PrimeNovo and Casanovo V2 highlights that PrimeNovo consistently maintains a significantly higher level of peptide-level accuracy across all combinations of missing peaks and target lengths. Furthermore, PrimeNovo consistently outperforms Casanovo V2 across all target lengths, underscoring its robustness in handling variations in peptide length as well as the presence of missing peaks in the input data.

PrimeNovo Demonstrates Enhanced Adaptability on the PT Dataset. In our investigation, we conducted fine-tuning experiments on both Casanovo V2 and PrimeNovo using the PT dataset. The objective was to evaluate the adaptability of both models to unseen data distributions. Similar to the observations in the "Results" section of our main manuscript, when the model is exclusively fine-tuned using the PT training data, we observe a phenomenon known as catastrophic forgetting, where the model's performance on the original data distribution significantly declines. This is evident in the left portion of Supplementary Fig. 11, where the addition of more PT data results in a drop in performance on the nine-species benchmark dataset. To address this issue, we implemented a strategy of mixing the training data from PT with the original MassiveKB training data in a proportional manner. As illustrated in the right-hand side of Supplementary Fig. 11, fine-tuning with a mixture of the original training data and PT data led to stable performance on the original nine-species dataset. Moreover, with the gradual addition of more PT training data, performance on the PT test set consistently improved. Crucially, throughout these fine-tuning experiments with varying dataset sizes and mixing ratios (10:1, 1:1, and 1:10), PrimeNovo consistently outperformed Casanovo V2 across all configurations. This robust performance underlines PrimeNovo's superior adaptability to different data distributions and highlights its strong generalization capabilities.

PrimeNovo Excels in Predicting Peptides with Various Charges, Outperforming Recently Published PepNet. We conducted a comparison of PrimeNovo's prediction accuracy with the most advanced CNN-based deep learning model, PepNet [46]. For this analysis, we utilized the PXD019483 dataset, which was employed as the test set by the authors, and organized predictions based on the charges of each peptide. As demonstrated in Supplementary Fig. 13a, PrimeNovo attains the highest peptide-level accuracy across peptides with different charges, outperforming PepNet. Additionally, we compared the performance of the PepNet model with PrimeNovo on the nine-species benchmark dataset, as depicted in Supplementary Fig. 13b. In this evaluation as well, PrimeNovo exhibits a significant advantage in terms of peptide-level recall over PepNet, underscoring its superior predictive capabilities.

We conducted a comparison of PrimeNovo's prediction accuracy with the most advanced CNN-based deep learning model, PepNet [46]. For this analysis, we utilized the PXD019483 dataset, which was employed as the test set by the authors, and organized predictions based on the charges of each peptide. As demonstrated in Supplementary Fig. 13a, PrimeNovo attains the highest peptide-level accuracy across peptides with different charges, outperforming PepNet. Additionally, we compared the performance of the PepNet model with PrimeNovo on the nine-species benchmark dataset, as depicted in Supplementary Fig. 13b. In this evaluation as well, PrimeNovo exhibits a significant advantage in terms of peptide-level recall over PepNet, underscoring its superior predictive capabilities.

PrimeNovo perform accurate sequencing on data generated from the instrument other than Thermo orbitrap. The mass spectrometry data generated by different types of experimental instruments can exhibit significant differences. Models trained on mass spectrometry data sourced from Thermo orbitrap instrument may not necessarily perform effectively on other types of instruments. To test our method's robustness against spectrum data from different instruments, we have chosen a public DDA-PASEF dataset [53] generated from timsTOF Pro instrument to evaluate our model's performance. The raw files (in `.d` format) and MSFragger identification results were obtained by downloading from the PRIDE repository by PXD041421. Mzml files generated by MSFragger were converted to mgf files and added identified information. This dataset contains 3667330 PSMs with 83272 peptides.

We randomly sampled 5% of the entire dataset for evaluation, with the results shown in Supplementary Figure 14. As observed, PrimeNovo demonstrates the ability to effectively decode spectra generated by the timsTOF Pro instrument. Specifically, it achieved a 45% peptide recall in a zero-shot setting, where no fine-tuning was performed on this data distribution. In comparison, Casanovo V2 achieved a peptide recall of 31%, significantly lower than our approach. This 14% improvement over Casanovo V2 is consistent with the performance advantage observed in other datasets (ranging from 10% to 30%).

SI Note 7: Additional results on model explainability

Analyzing the orthogonality of the Value matrix projection by the norm of the Gram matrix. In order to maintain consistency with previous studies, we utilized the norm of the Gram matrix to measure the orthogonality of the Value matrix projection. In this methodology, the Value matrices from all heads at each layer are concatenated to form a complete Value matrix. Subsequently, each column of the Value matrix is normalized according to the L2 norm to transform it into a vector with a magnitude of 1. This normalized matrix is then transposed and multiplied by itself to generate a Gram matrix. The measure of orthogonality is quantified by taking the Frobenius norm of the resulting matrix after the subtraction of an identity matrix. A smaller value of this norm indicates better orthogonality of the Value matrix projection, implying reduced redundancy in the extracted features. As shown in Supplementary Fig. 18, we employ a line graph to visualize and compare the orthogonality of the Value matrix projections in the Casanovo and PrimeNovo encoders. Notably, the curve for PrimeNovo consistently lies below that of Casanovo, demonstrating superior orthogonality in the Value matrix projection of the PrimeNovo encoder, which suggests more diverse feature extraction.

Analyzing the bias of the PrimeNovo encoder's attention mechanism towards $b - y$ ions. On the nine-species benchmark dataset, we random select a spectrum as an example and examine the distribution of match ratio for different layers in the PrimeNovo encoder (Supplementary Fig. 16 and 17). Here, match ratio refers to the proportion of by ions in the top 10 peaks (excluding the first token) that receive the most attention in each sample. This attention level is measured by the L1 norm of the corresponding column in the attention score matrix, with a larger norm indicating a higher attention level.

Visualization of the distribution of match ratio. As shown in Supplementary Fig. 19, we conducted the analysis on the test dataset of H.sapiens, in which the proportion of b ions and y ions among all peaks is 0.16. This threshold signifies the expected match ratio if the attention mechanism's bias towards b ions and y ions were random. A value exceeding this threshold indicates a preference for ions. The majority of layers exhibit a preference for focusing on b ions and y ions, particularly the second and eighth layers, aligning with the prior knowledge that ions contain crucial information for de novo sequencing. Conversely, layers 1, 5, and 6 tend to overlook b ions and y ions, possibly due to their function in extracting useful information from other peaks. A common feature across all layers is the deviation of match ratio from that of random selection, suggesting a definitive selective bias of the attention mechanism towards b ions and y ions.

Generation of all possible fragment ions. In this analysis, we considered only the fragment ions generated by the cleavage of chemical bonds along the peptide backbone. It is well-known that in mass spectrometry analysis of peptides, the focus is primarily on six ions generated by the cleavage of three types of chemical bonds. These include a-ions and x-ions formed by the cleavage between the α-carbon atom and the carboxyl group, b ions and y ions formed by the cleavage between the carbonyl group and the amino group of the peptide bond, and c ions and z ions formed by the cleavage between the α-carbon atom and the amino group. For this study, we also focused on these three types of chemical bond cleavages. However, we not only considered these six ions but also took into account all possible intermediate fragment ions of variable lengths. This process was implemented by the programming language Python.

SI Note 8: Additional Result on Synthesizing phospho-peptide from PrimeNovo

Peptide Selection Process

We applied the fine-tuned PrimeNovo model to the non-enriched 2020-Cell-LUAD dataset, notably a dataset without pre-identified peptide labels. The model's confidence score served as the primary criterion for selection due to the absence of these labels. We initially filtered for the top 300 predicted peptides featuring phospho modifications, all of which boasted confidence scores exceeding 0.99. Subsequent manual scrutiny allowed us to refine this selection based on quality indicators, specifically: 1) Preference for peptides terminating with the amino acids K or R; 2) Selection of peptides of an optimal length, avoiding those excessively long or short; 3) Requirement that more than half of the 20 most intense peaks could be aligned with the theoretical ions of the peptide in question. 4) The presence of a matched ion both preceding and following the phospho modification site.

Following these criteria, we identified 12 peptides as prime candidates for laboratory synthesis and subsequent functional analyses.

Mass Spectrometry Analysis of Synthetic Phosphopeptides

Liquid chromatography tandem mass spectrometry (LC-MS/MS) analyses were conducted using an Orbitrap Fusion Tribrid Lumos mass spectrometer (Thermo Fisher Scientific), directly interfaced with a nanoflow LC system (EASY-nLC 1200, Thermo Fisher Scientific). A quantity of 50 ng of synthetic phosphopeptides was resuspended in solvent A (0.1% formic acid (FA) in HPLC-grade water) and introduced onto a 2 cm self-packed trapping column (100-μm inner diameter, 1.9 μm resin, ReproSil-Pur C18-AQ, Dr. Maisch GmbH) in solvent A. Following loading and washing steps, peptides were eluted to a 30 cm analytical column (150-μm inner diameter, 1.9 μm resin, ReproSil-Pur C18-AQ, Dr. Maisch GmbH) and separated using a non-linear gradient of solvent B (0.1% FA in acetonitrile, ACN) over 30 minutes at a flow rate of 600 nL/min. The gradient program was as follows: from 0 to 2 minutes, 7–12% solvent B; 2 to 13 minutes, 12–32% solvent B; 13 to 19 minutes, 32–45% solvent B; 19 to 21 minutes, 45–95% solvent B; and maintaining 95% solvent B from 21 to 30 minutes. Ionization was achieved with a 2.0 kV spray voltage and a capillary temperature set at 320 °C. Data acquisition was carried out in OT-OT mode with full MS scans (350 to 1,500 m/z) at a resolution of 120,000, a maximum injection time of 50 ms, and an Automatic Gain Control (AGC) target of 1e6. MS2 fragmentation was performed using higher-energy collision dissociation (HCD) with a normalized collision energy of 32%, on 2+ to 7+ precursor ions within a 3 s duty cycle, in top-speed mode. MS2 spectra were acquired in the ion trap in rapid mode, with an AGC target of 50,000 and a maximum injection time of 22 ms, at a resolution of 15,000. Dynamic exclusion was applied for 25 s to prevent the redundant selection of peptides.

Analysis of MS Data for Synthetic Phosphopeptides

The acquired Thermo raw files were analyzed using MaxQuant (version 2.1.4), querying against the human UniProt database (SwissProt, 20,266 entries, release date 2021-01-22). The analysis incorporated variable modifications such as oxidation on methionine, phosphorylation on serine, threonine, and tyrosine, and acetylation at the N-termini of proteins. Fixed modifications included carbamidomethylation of cysteine. The search parameters permitted up to two missed cleavages by trypsin (with full specificity). To ensure data quality, the Percolator algorithm was employed to maintain the false discovery rate (FDR) below 1% at both the peptide spectrum match (PSM) and protein levels.

Details of 12 synthesized Phosphopeptides

We analyzed each one of 12 phosphopeptides in detail from various aspects, with summarized information in Supplementary Table 11 and Table 12.

More about Three LUAD-Unrelated Peptides from Twelve Selected Predictions

SRRM2 is indispensable for pre-mRNA splicing, serving as a vital component of the spliceosome. Within the minor spliceosome, it potentially contributes to the splicing of U12-type introns found in pre-mRNAs. Research highlights a notable correlation between SRRM2 expression and prostate cancer prognosis [51]. Additionally, SRRM2 possesses phosphorylation sites uniquely regulated by GSK3α, significantly impacting the survival outcomes of colorectal cancer patients [4].

SNTB2 is crucial for the subcellular localization and organization of various membrane proteins. It may connect different receptors to the actin cytoskeleton and the dystrophin-glycoprotein complex, influencing the regulation of secretory granules through interactions with PTPRN. Recent studies have identified SNTB2 as a critical immune marker in macrophages, linked to melanoma metastasis [7]. Mutations within the SNTB2 gene are prevalent across Burkitt lymphoma (BL) subtypes, underscoring its potential significance in the disease's pathology [6].

HACD3 plays a crucial role in the third of four reactions in the long-chain fatty acid elongation cycle. It is involved in producing very long-chain fatty acids (VLCFAs) of various lengths, serving as precursors for membrane lipids and lipid mediators, essential for numerous biological functions. Studies have linked HACD3 to the aggressiveness and prognosis of breast cancer, including its association with patient survival rates [34, 35].

References

[1] Zheng, J. *et al.* Molecular changes of lung malignancy in hiv infection. *Scientific reports* **8**, 13128 (2018).

[2] Sharifnia, T. *et al.* Mapping the landscape of genetic dependencies in chordoma. *Nature Communications* **14**, 1933 (2023).

[3] Li, Y. *et al.* Identification of bicalutamide resistance-related genes and prognosis prediction in patients with prostate cancer. *Frontiers in Endocrinology* **14**, 1125299 (2023).

[4] Gao, L. *et al.* Deciphering the clinical significance and kinase functions of gsk3α in colon cancer by proteomics and phosphoproteomics. *Molecular & Cellular Proteomics* **22** (2023).

[5] Hungermann, D. *et al.* Influence of whole arm loss of chromosome 16q on gene expression patterns in oestrogen receptor-positive, invasive breast cancer. *The Journal of pathology* **224**, 517–528 (2011).

[6] Panea, R. I. *et al.* The whole-genome landscape of burkitt lymphoma subtypes. *Blood, The Journal of the American Society of Hematology* **134**, 1598–1607 (2019).

[7] Li, Z. *et al.* Development of a macrophage-related risk model for metastatic melanoma. *International Journal of Molecular Sciences* **24**, 13752 (2023).

[8] Gan, J., Meng, Q. & Li, Y. Systematic analysis of expression profiles and prognostic significance for fam83 family in non-small-cell lung cancer. *Frontiers in Molecular Biosciences* **7**, 572406 (2020).

[9] Sayeeram, D. *et al.* Identification of potential biomarkers for lung adenocarcinoma. *Heliyon* **6** (2020).

[10] Li, M. *et al.* Oncogenic zinc finger protein znf687 accelerates lung adenocarcinoma cell proliferation and tumor progression by activating the pi3k/akt signaling pathway. *Thoracic Cancer* **14**, 1223–1238 (2023).

[11] Fujii, M. *et al.* Snip1 is a candidate modifier of the transcriptional activity of c-myc on e box-dependent target genes. *Molecular cell* **24**, 771–783 (2006).

[12] Jeon, H.-S. *et al.* High expression of snip1 correlates with poor prognosis in non-small cell lung cancer and snip1 interferes with the recruitment of hdac1 to rb in vitro. *Lung Cancer* **82**, 24–30 (2013).

[13] Zhong, Y. *et al.* Long non-coding rna afap1-as1 accelerates lung cancer cells migration and invasion by interacting with snip1 to upregulate c-myc. *Signal Transduction and Targeted Therapy* **6**, 240 (2021).

[14] Wu, J. *et al.* Mir-138-5p suppresses the progression of lung cancer by targeting snip1. *Thoracic Cancer* **14**, 612–623 (2023).

[15] Windhorst, S. *et al.* Functional role of inositol-1, 4, 5-trisphosphate-3-kinase-a for motility of malignant transformed cells. *International journal of cancer* **129**, 1300–1309 (2011).

[16] Pal, J. *et al.* Systematic analysis of migration factors by migexpress identifies essential cell migration control genes in non-small cell lung cancer. *Molecular oncology* **15**, 1797–1817 (2021).

[17] Wu, J. *et al.* Comprehensive analysis of tumor microenvironment and identification of an immune signature to predict the prognosis and immunotherapeutic response in lung squamous cell carcinoma. *Annals of translational medicine* **9** (2021).

[18] Wu, Z.-X., Huang, X., Cai, M.-J., Huang, P.-D. & Guan, Z. Development and validation of a prognostic index based on genes participating in autophagy in patients with lung adenocarcinoma. *Frontiers in Oncology* **11**, 799759 (2022).

[19] Ding, H. *et al.* Construction and evaluation of a prognostic risk model of tumor metastasis-related genes in patients with non-small cell lung cancer. *BMC Medical Genomics* **15**, 187 (2022).

[20] Xuan, Y., Jin, X., Wang, M., Wang, Z. & Sun, Y. Necroptosis-related prognostic signature and nomogram model for predicting the overall survival of patients with lung cancer. *Genetics Research* **2022**, e36 (2022).

[21] Qin, C. *et al.* Development and validation of a dna damage repair-related gene-based prediction model for the prognosis of lung adenocarcinoma. *Journal of Thoracic Disease* **15**, 6928 (2023).

[22] Yang, Z. *et al.* Comprehensive analysis of resistance mechanisms to egfr–tkis and establishment and validation of prognostic model. *Journal of Cancer Research and Clinical Oncology* **149**, 13773–13792 (2023).

[23] Hui, G., Xie, Y., Niu, L. & Liu, J. A novel gene signature related to focal adhesions for distinguishing and predicting the prognosis of lung squamous cell carcinoma. *Frontiers in Medicine* **10**, 1284490 (2024).

[24] Hu, Z. *et al.* Genetic variants of mirna sequences and non–small cell lung cancer survival. *The Journal of clinical investigation* **118**, 2600–2608 (2008).

[25] Park, S. L. *et al.* Pleiotropic associations of risk variants identified for other cancers with lung cancer risk: the page and tricl consortia. *Journal of the National Cancer Institute* **106**, dju061 (2014).

[26] Sui, Y. *et al.* Gene amplification and associated loss of 5 regulatory sequences of coaa in human cancers. *Oncogene* **26**, 822–835 (2007).

[27] Hu, Y., Mu, H. & Deng, Z. Rbm14 as a novel epigenetic-activated tumor oncogene is implicated in the reprogramming of glycolysis in lung cancer. *World Journal of Surgical Oncology* **21**, 132 (2023).

[28] Shu, Y. *et al.* The acin1 gene is hypermethylated in early stage lung adenocarcinoma. *Journal of Thoracic Oncology* **1**, 160–167 (2006).

[29] Ujifuku, K. *et al.* Exploration of pericyte-derived factors implicated in lung cancer brain metastasis protection: a pilot messenger rna sequencing using the blood–brain barrier in vitro model. *Cellular and Molecular Neurobiology* **42**, 997–1004 (2022).

[30] Bacolod, M. D., Fisher, P. B. & Barany, F. Multi-cpg linear regression models to accurately predict paclitaxel and docetaxel activity in cancer cell lines. *Advances in Cancer Research* **158**, 233–292 (2023).

[31] Ramsey, J. *et al.* Loss of runx1 is associated with aggressive lung adenocarcinomas. *Journal of cellular physiology* **233**, 3487–3497 (2018).

[32] Kim, Y. *et al.* Clinicopathological significance of runx1 in non-small cell lung cancer. *Journal of Clinical Medicine* **9**, 1694 (2020).

[33] Ma, H. *et al.* Runx1 promotes proliferation and migration in non-small cell lung cancer cell lines via the mtor pathway. *The FASEB Journal* **37**, e23195 (2023).

[34] Barupal, D. K. *et al.* Prioritization of metabolic genes as novel therapeutic targets in estrogen-receptor negative breast tumors using multi-omics data and text mining. *Oncotarget* **10**, 3894 (2019).

[35] Xu, Y. *et al.* Prognostic signature and therapeutic value based on membrane lipid biosynthesis-related genes in breast cancer. *Journal of Oncology* **2022** (2022).

[36] Tran, N. H., Zhang, X., Xin, L., Shan, B. & Li, M. De novo peptide sequencing by deep learning. *Proceedings of the National Academy of Sciences* **114**, 8247–8252 (2017).

[37] Yilmaz, M. *et al.* Sequence-to-sequence translation from mass spectra to peptides with a transformer model. *Nature communications* **15**, 6427 (2024).

[38] Qiao, R. *et al.* Computationally instrument-resolution-independent de novo peptide sequencing for high-resolution devices. *Nature Machine Intelligence* **3**, 420–425 (2021).

[39] Wang, M. *et al.* Assembling the Community-Scale Discoverable Human Proteome. *Cell Systems* **7**, 412–421.e5 (2018).

[40] Zolg, D. P. *et al.* Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods* **14** (2017).

[41] Vizcaíno, J. A. *et al.* 2016 update of the PRIDE database and its related tools. *Nucleic acids research* **44**, D447–56 (2016).

[42] Jiang, Y. *et al.* Proteomics identifies new therapeutic targets of early-stage hepatocellular carcinoma. *Nature* **567**, 257–261 (2019).

[43] Tran, N. H. *et al.* Complete de Novo Assembly of Monoclonal Antibody Sequences. *Scientific Reports* **6**, 1–10 (2016).

[44] Mao, Z., Zhang, R., Xin, L. & Li, M. Mitigating the missing-fragmentation problem in de novo peptide sequencing with a two-stage graph-based deep learning model. *Nature Machine Intelligence* **5**, 1250–1260 (2023).

[45] Müller, J. B. *et al.* The proteome landscape of the kingdoms of life. *Nature* **582**, 592–596 (2020).

[46] Liu, K., Ye, Y., Li, S. & Tang, H. Accurate de novo peptide sequencing using fully convolutional neural networks. *Nature Communications* **14** (2023).

[47] McIlwain, S. *et al.* Crux: Rapid Open Source Protein Tandem Mass Spectrometry Analysis. *Journal of Proteome Research* **13**, 4488–4491 (2014).

[48] Patnode, M. L. *et al.* Interspecies Competition Impacts Targeted Manipulation of Human Gut Bacteria by Fiber-Derived Glycans. *Cell* **179**, 59–73.e13 (2019).

[49] Paul Zolg, D. *et al.* Proteometools: Systematic characterization of 21 post-translational protein modifications by liquid chromatography tandem mass spectrometry (lc-ms/ms) using synthetic peptides. *Molecular and Cellular Proteomics* **17**, 1850–1863 (2018).

[50] Xu, J. Y. *et al.* Integrative Proteomic Characterization of Human Lung Adenocarcinoma. *Cell* **182**, 245–261.e17 (2020).

[51] Li, L. *et al.* Revealing proteome-level functional redundancy in the human gut microbiome using ultra-deep metaproteomics. *Nature Communications* **14**, 3428 (2023).

[52] Beslic, D., Tscheuschner, G., Renard, B. Y., Weller, M. G. & Muth, T. Comprehensive evaluation of peptide de novo sequencing tools for monoclonal antibody assembly. *Briefings in Bioinformatics* **5**, 1–12 (2022).

[53] Wang, H. *et al.* MultiPro: DDA-PASEF and diaPASEF acquired cell line proteomic datasets with deliberate batch effects. *Scientific Data* **10**, 858 (2023).