De *novo* mass spectrometry peptide sequencing with a transformer model

Melih Yilmaz, Will Fondrie, Wout Bittremieux, Sewoong Oh, William Noble

Mass spectrometry provides a high-throughput framework for identifying proteins

- Proteins are digested into ~15-20 amino acid long peptides
- Peptides are analyzed in the mass spectrometer

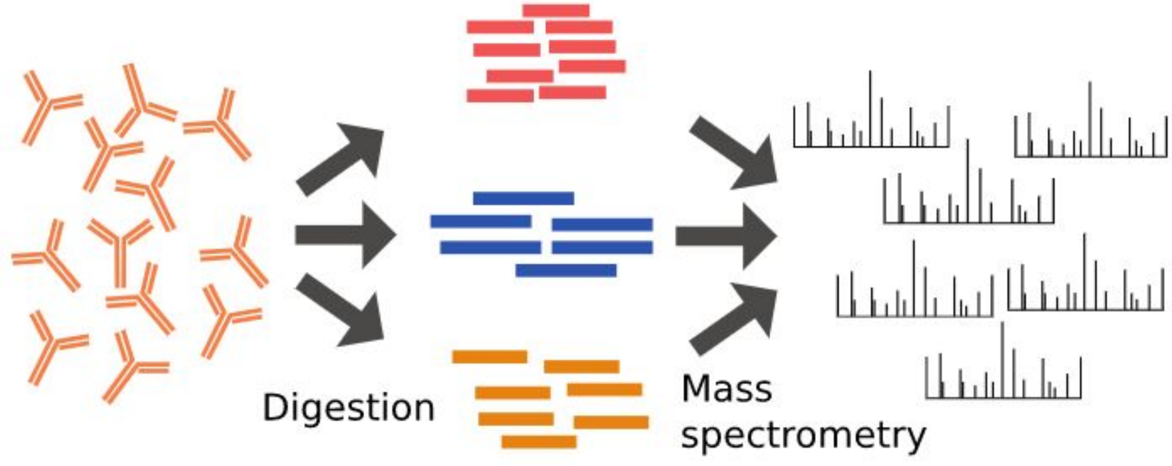
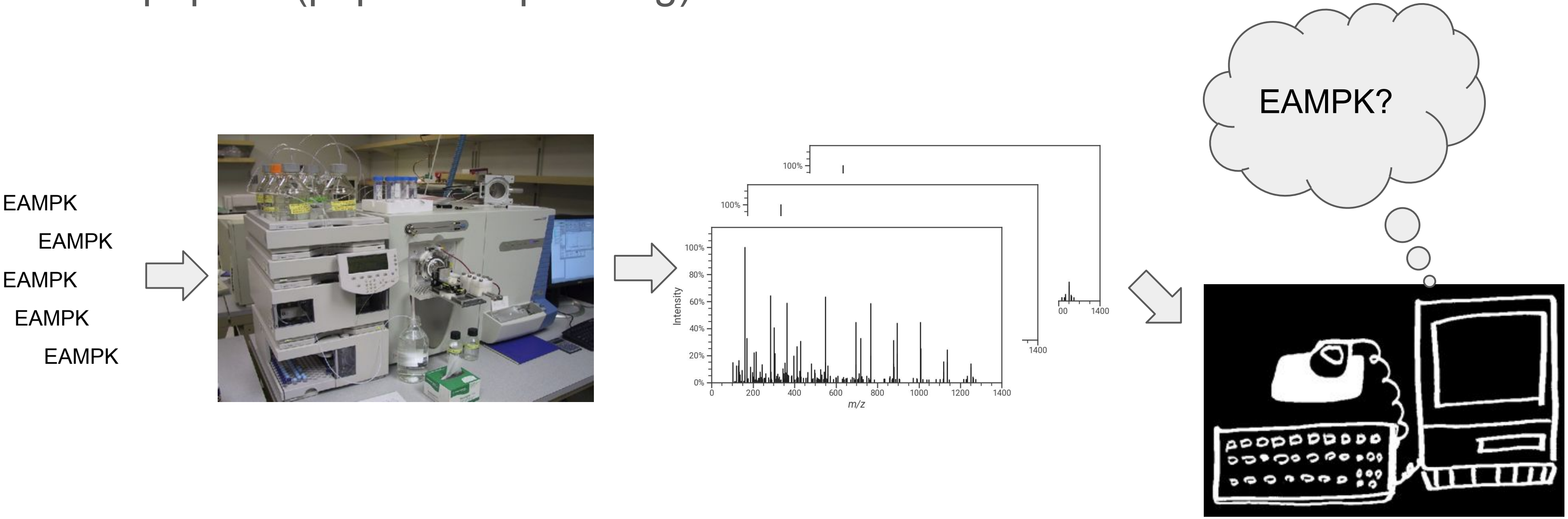


Image https://abterrabio.com/2018/08/06/de-novo-antibody-sequencing-an-overview/

The goal: assign a generating peptide to each spectrum

- Given a spectrum, computationally identify the amino acid sequence of its peptide (peptide sequencing)



EAMPK

EAMPK

EAMPK

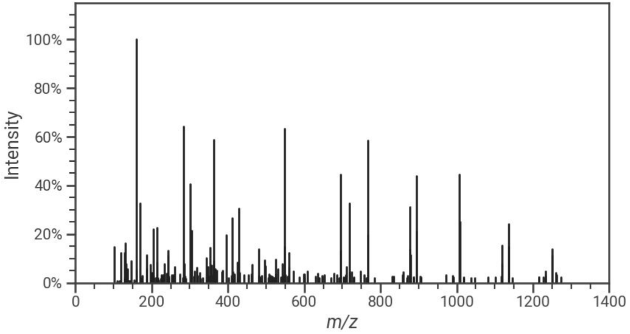
EAMPK

EAMPK

EAMPK?

De novo sequencing infers peptide directly from spectrum

Observed mass spectrum



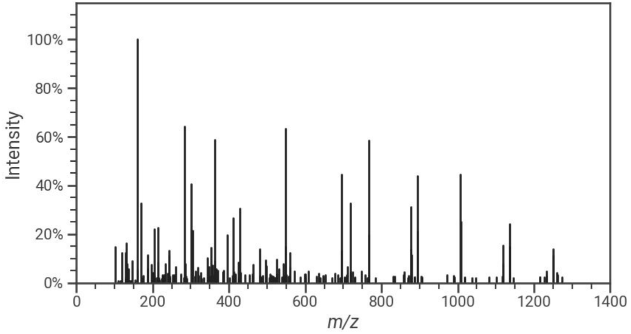
Generating peptide

DNTIEINVEPK

De novo sequencing infers peptide directly from spectrum

- In addition, we also observe precursor mass, i.e. full mass of the peptide, and charge

Observed mass spectrum

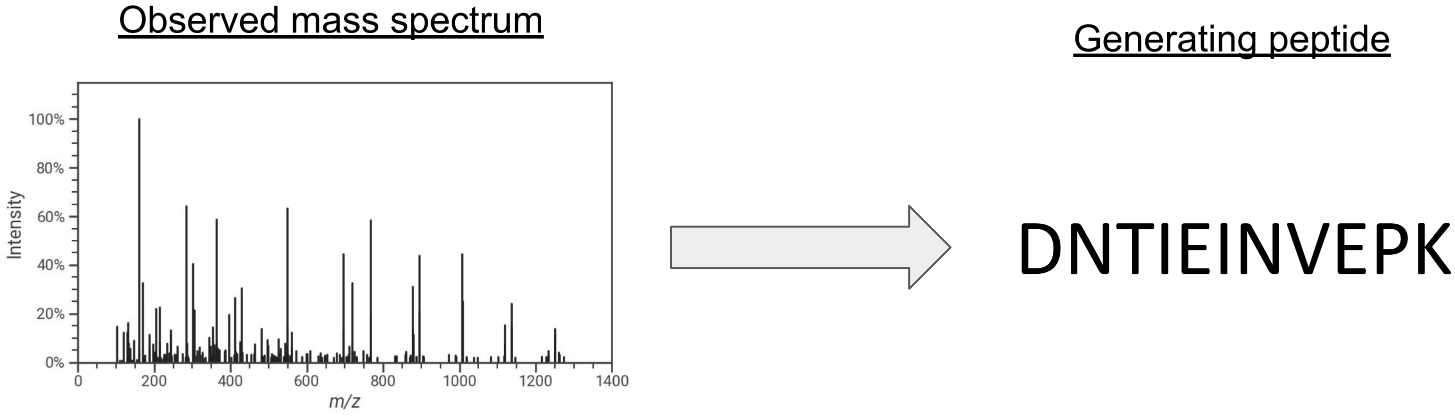


Generating peptide

DNTIEINVEPK

De novo sequencing infers peptide directly from spectrum

- In addition, we also observe precursor mass, i.e. full mass of the peptide, and charge
- Hard to *de novo* sequence accurately

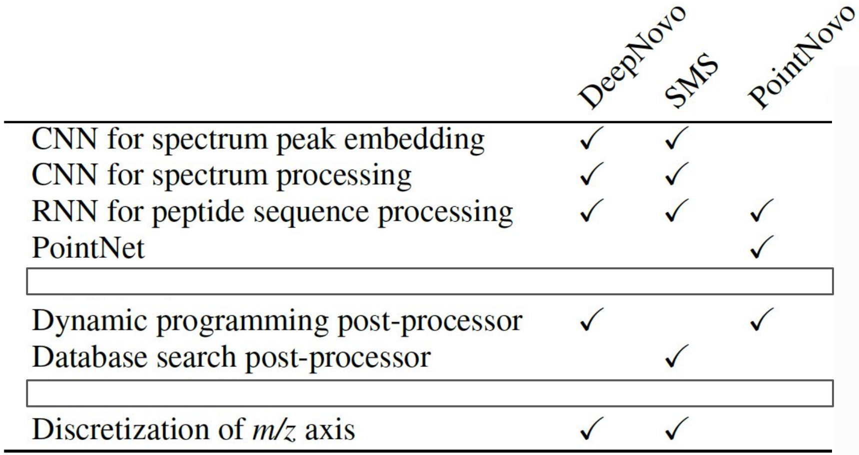


Shortcomings of existing methods

- **Accuracy is still low:** correctly assigns peptides to 40-60% of spectra

Shortcomings of existing methods

- **Accuracy is still low:** correctly assigns peptides to 40-60% of spectra
- **Complex models:** combines multiple neural nets and post-processing steps
  - → higher # of parameters and slow inference

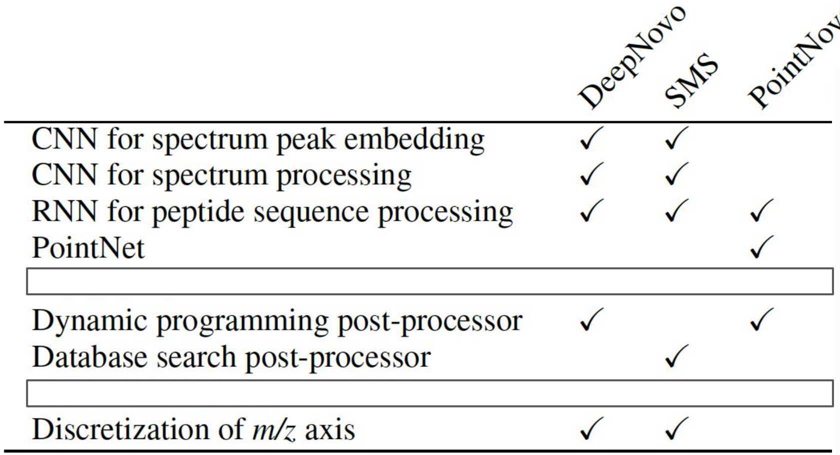


	DeepNovo	SMS	PointNovo
CNN for spectrum peak embedding	✓	✓	
CNN for spectrum processing	✓	✓	
RNN for peptide sequence processing	✓	✓	✓
PointNet			✓
Dynamic programming post-processor	✓		✓
Database search post-processor		✓	
Discretization of *m/z* axis	✓	✓	

Table: Comparison of existing deep learning methods for de novo peptide sequencing.

Shortcomings of existing methods

- **Accuracy is still low:** correctly assigns peptides to 40-60% of spectra
- **Complex models:** combines multiple neural nets and post-processing steps
  - → higher # of parameters and slow inference
- ***m/z* axis discretization:** presents a tradeoff between low binning resolution and higher model complexity



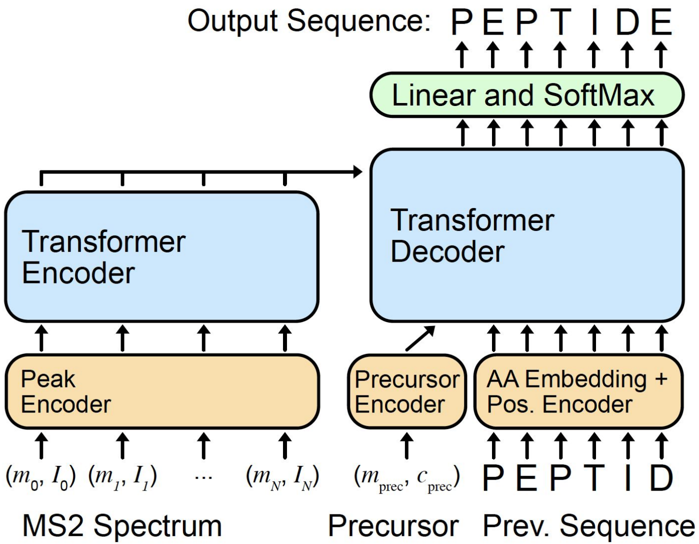
	DeepNovo	SMS	PointNovo
CNN for spectrum peak embedding	✓	✓	
CNN for spectrum processing	✓	✓	
RNN for peptide sequence processing	✓	✓	✓
PointNet			✓
Dynamic programming post-processor	✓		✓
Database search post-processor		✓	
Discretization of *m/z* axis	✓	✓	

Table: Comparison of existing deep learning methods for de novo peptide sequencing.

Peptide sequencing can be conceived as translation between two sequences (spectrum ⇢ peptide)

Peptide sequencing can be conceived as translation between two sequences (spectrum ⇢ peptide)

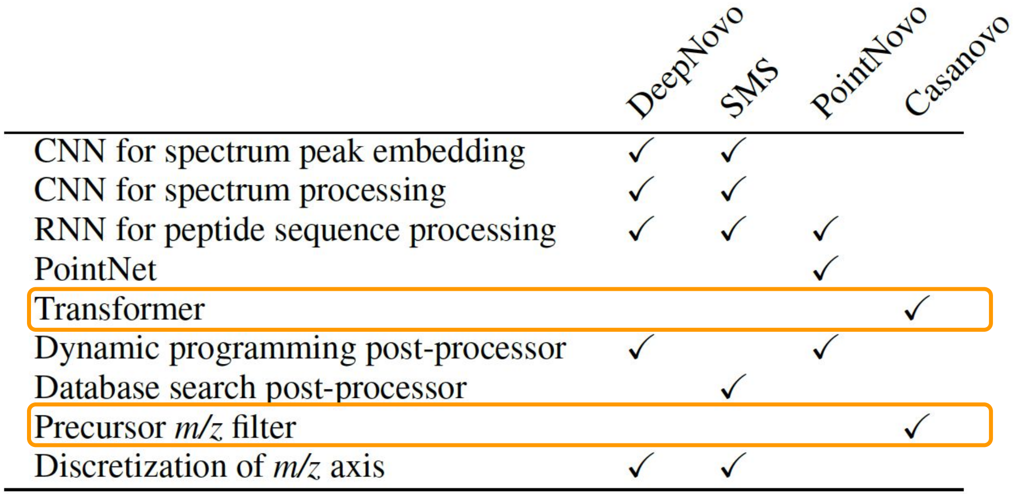
And learned with a transformer model



Output Sequence: P E P T I D E
Linear and SoftMax
Transformer Decoder
Transformer Encoder
Peak Encoder
Precursor Encoder
AA Embedding + Pos. Encoder
(m_0, I_0) (m_1, I_1) ··· (m_N, I_N)
$(m_{\text{prec}}, c_{\text{prec}})$
P E P T I D
MS2 Spectrum
Precursor
Prev. Sequence

Casanovo: a *de novo* peptide sequencing transformer

- We propose a **unified** solution to sequencing sub-tasks

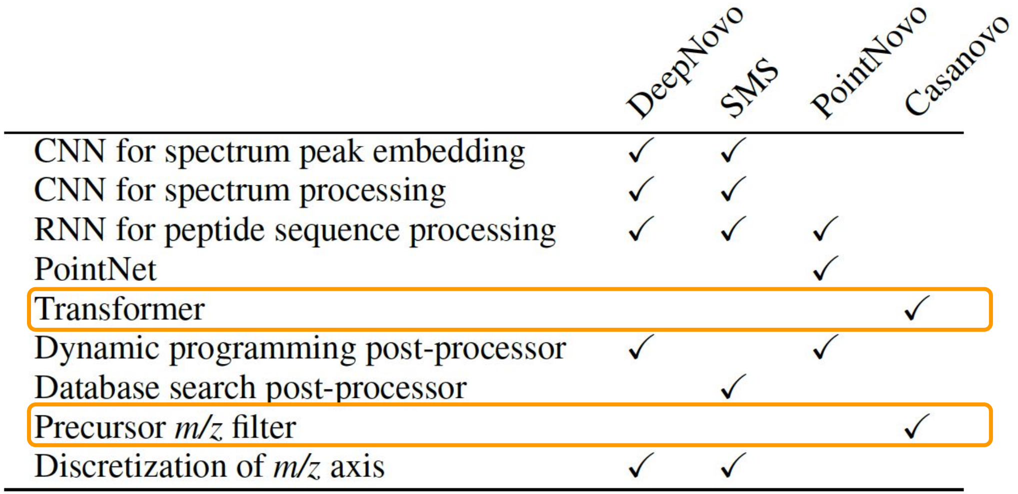


	DeepNovo	SMS	PointNovo	Casanovo
CNN for spectrum peak embedding	✓	✓		
CNN for spectrum processing	✓	✓		
RNN for peptide sequence processing	✓	✓	✓	
PointNet			✓	
Transformer				✓
Dynamic programming post-processor	✓		✓	
Database search post-processor		✓		
Precursor *m/z* filter				✓
Discretization of *m/z* axis	✓	✓		

Table: Comparison of deep learning methods for de novo peptide sequencing.

Casanovo: a *de novo* peptide sequencing transformer

- We propose a **unified** solution to sequencing sub-tasks
- Casanovo directly models spectrum peaks
  - No need for m/z discretization!

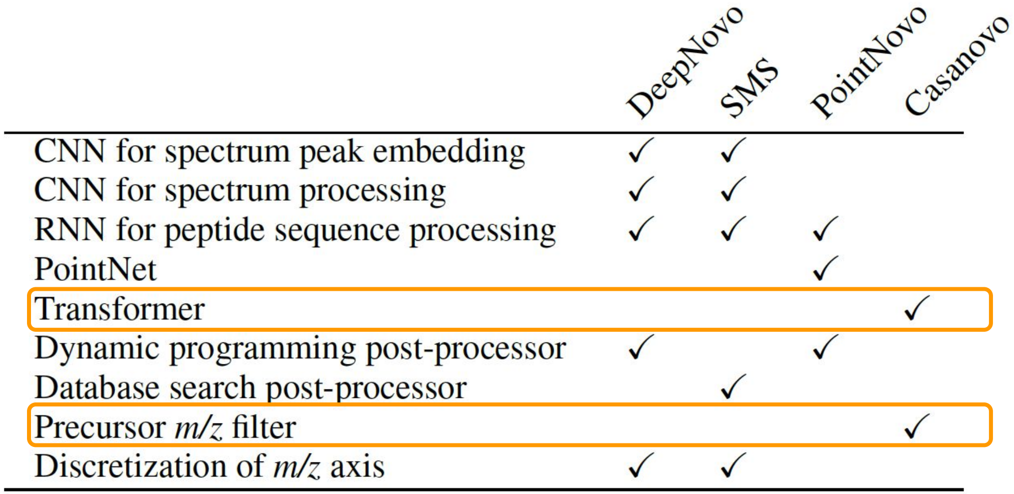


	DeepNovo	SMS	PointNovo	Casanovo
CNN for spectrum peak embedding	✓	✓		
CNN for spectrum processing	✓	✓		
RNN for peptide sequence processing	✓	✓	✓	
PointNet			✓	
Transformer				✓
Dynamic programming post-processor	✓		✓	
Database search post-processor		✓		
Precursor *m/z* filter				✓
Discretization of *m/z* axis	✓	✓		

Table: Comparison of deep learning methods for de novo peptide sequencing.

Casanovo: a *de novo* peptide sequencing transformer

- We propose a **unified** solution to sequencing sub-tasks
- Casanovo directly models spectrum peaks
  - No need for m/z discretization!
- Filters out implausible de novo sequences based on precursor m/z



	DeepNovo	SMS	PointNovo	Casanovo
CNN for spectrum peak embedding	✓	✓		
CNN for spectrum processing	✓	✓		
RNN for peptide sequence processing	✓	✓	✓	
PointNet			✓	
Transformer				✓
Dynamic programming post-processor	✓		✓	
Database search post-processor		✓		
Precursor *m/z* filter				✓
Discretization of *m/z* axis	✓	✓		

Table: Comparison of deep learning methods for de novo peptide sequencing.

Cross-species evaluation framework

- Benchmark dataset with **~1.5M peptide-spectra matches** from **9 species** was used

8 species

- Train/Validation
  - 90/10
  - ~1.4M spectra

1 species (e.g. yeast)

- Test
  - ~100k spectra

Cross-species evaluation framework

- Benchmark dataset with **~1.5M peptide-spectra matches** from **9 species** was used
- Test set peptides are mostly unique, i.e. not seen in the training set

8 species

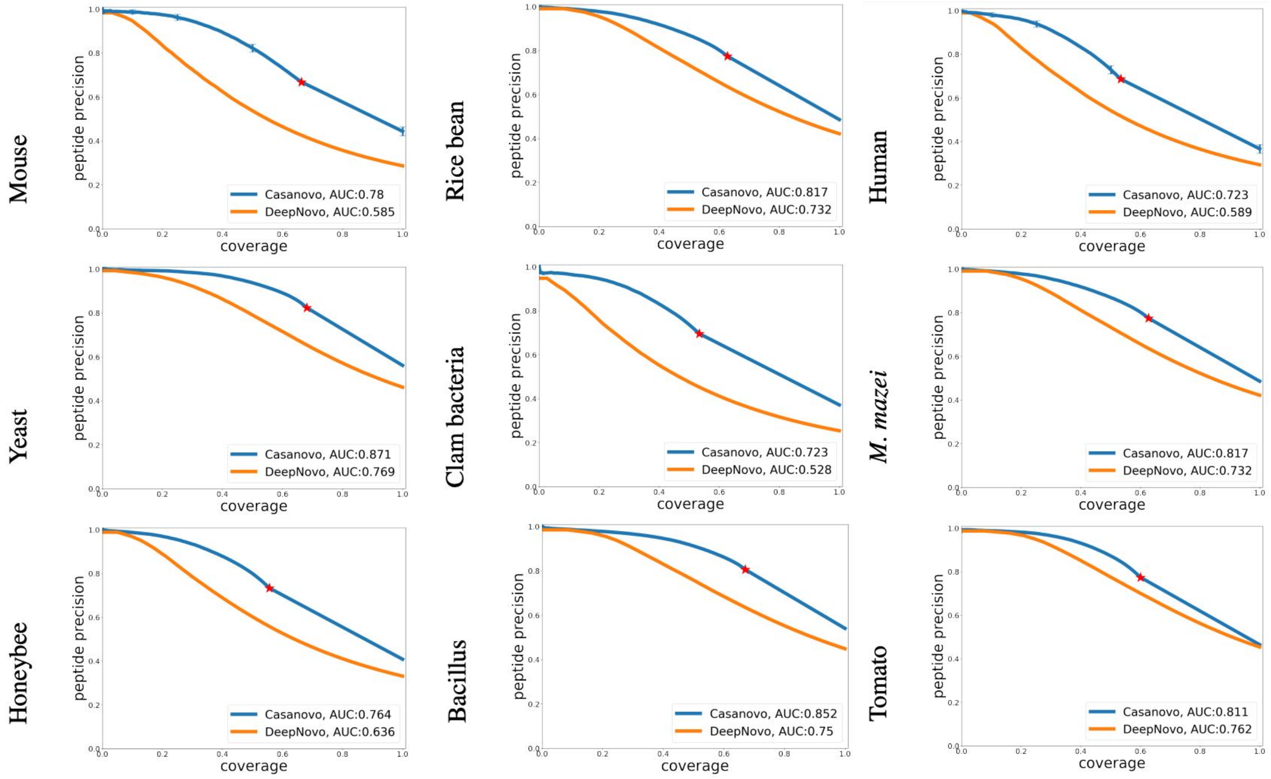
- Train/Validation
  - 90/10
  - ~1.4M spectra

1 species (e.g. yeast)

- Test
  - ~100k spectra

Casanovo achieves higher peptide precision in all species

- Consistently better precision at the same coverage
- Higher overall precision in all
- Mean AUC improvement of **0.13**

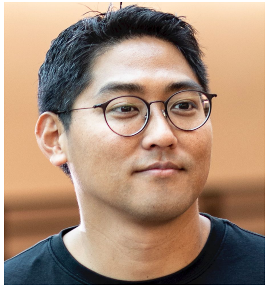


Thanks!

Code @ github.com/Noble-Lab/casanovo

Bill Noble

Will Fondrie


Sewoong Oh

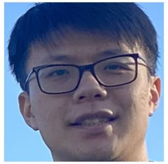
Wout Bittremieux

Noble Lab

Dejun

Kris

Bobby


Gang

Ran

Gesine

Ayse

Anu

Kianna

Lincoln

Alan

Mu

Robin

Yang

Melih

W PAUL G. ALLEN SCHOOL
OF COMPUTER SCIENCE & ENGINEERING

Genome Sciences
UNIVERSITY OF WASHINGTON