

NAISS · KÄLL LAB



# dion

A self-supervised transformer for tandem mass spectra

Alfred Nilsson

KTH

Joel Lapin

TUM

Samuel H. Payne

BYU

Mathias Wilhelm

TUM

Lukas Käll

KTH

Preprint: [doi.org/10.48550/arXiv.2610.06282](https://doi.org/10.48550/arXiv.2610.06282)



● ● ● claude · lukas@arrhenius:~/dion

lukas@arrhenius:~/dion\$ claude

> I am presenting our dIon paper at NAISS. They know computations better than peptides. How should I structure the talk?

- Answer five questions, in this order:

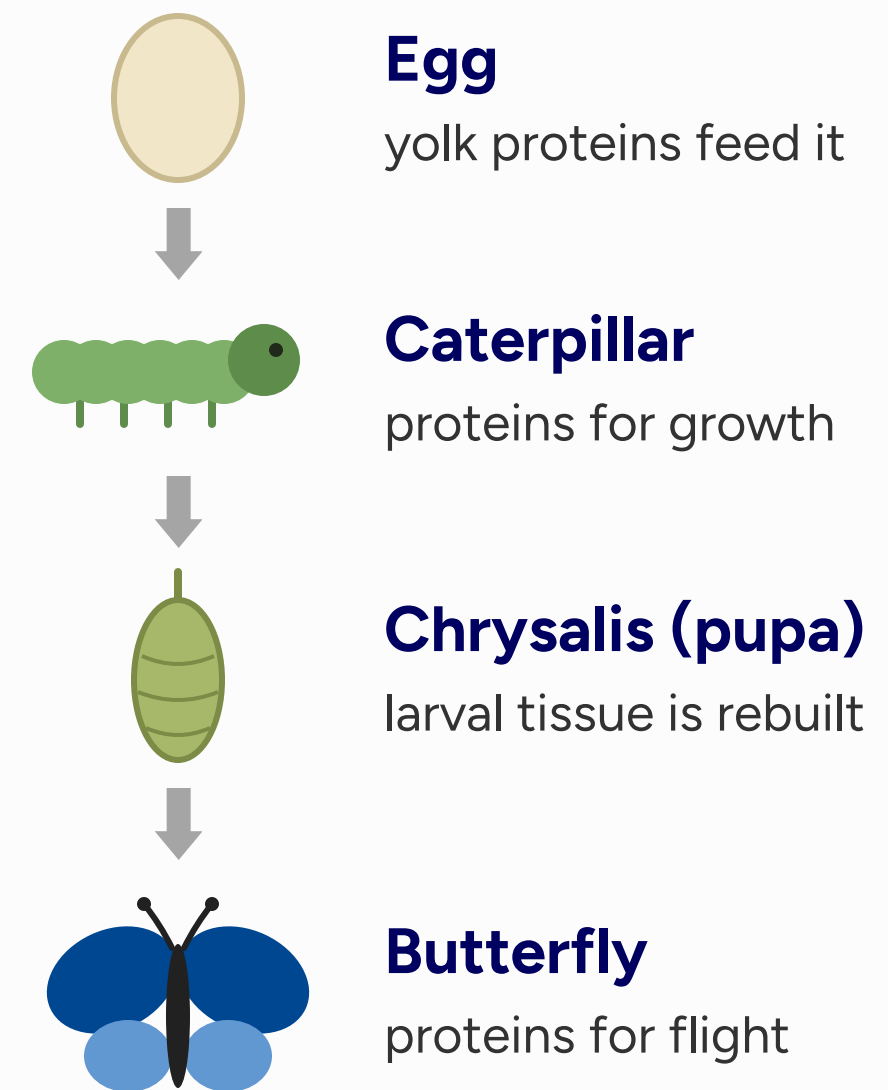
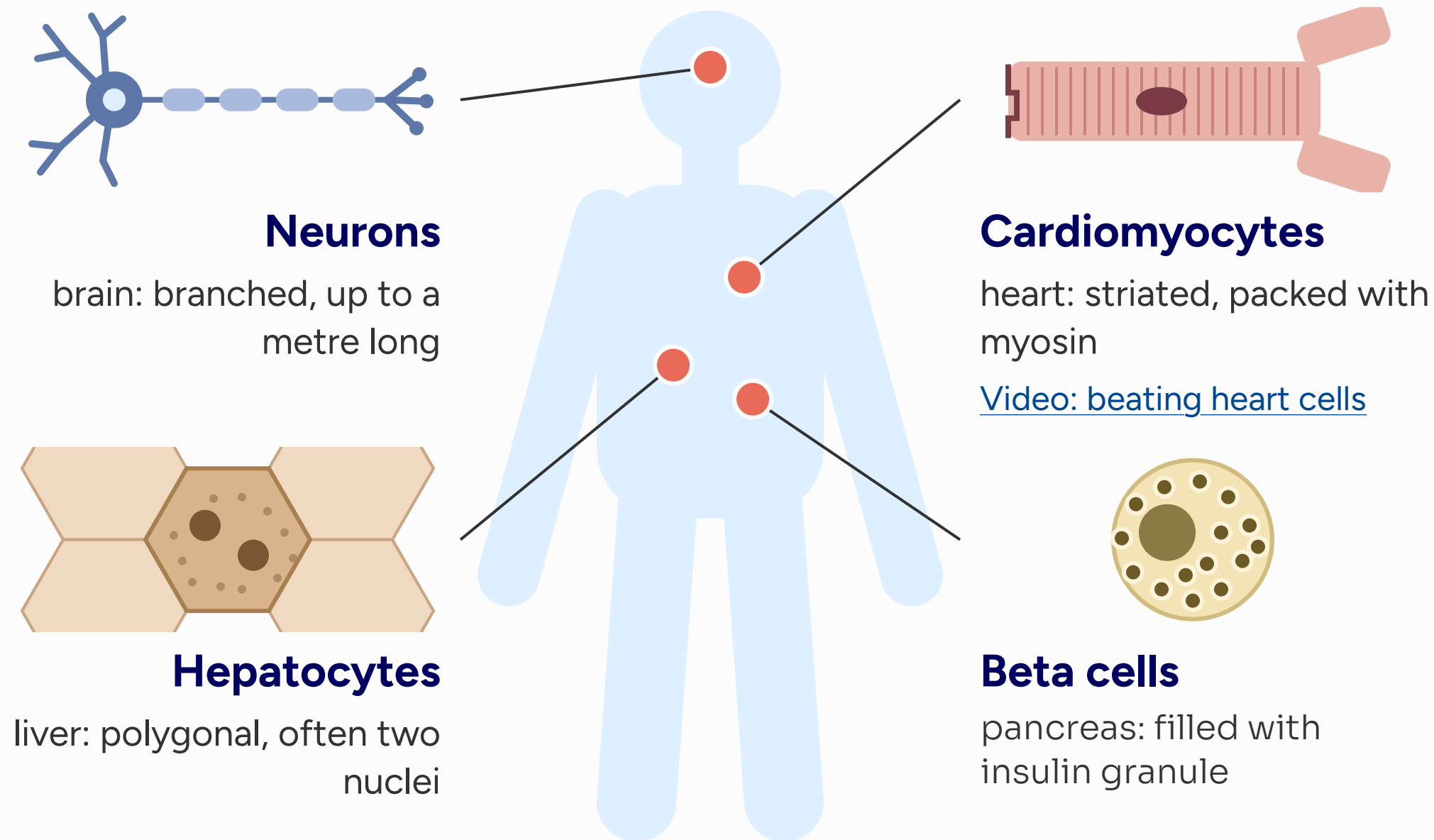
1. Why proteomics?
2. Why learn without labels?
3. How does DINO learn without labels?
4. What does dIon change compared with DINO?
5. What can the pretrained model be used for?

└ starting with question 1

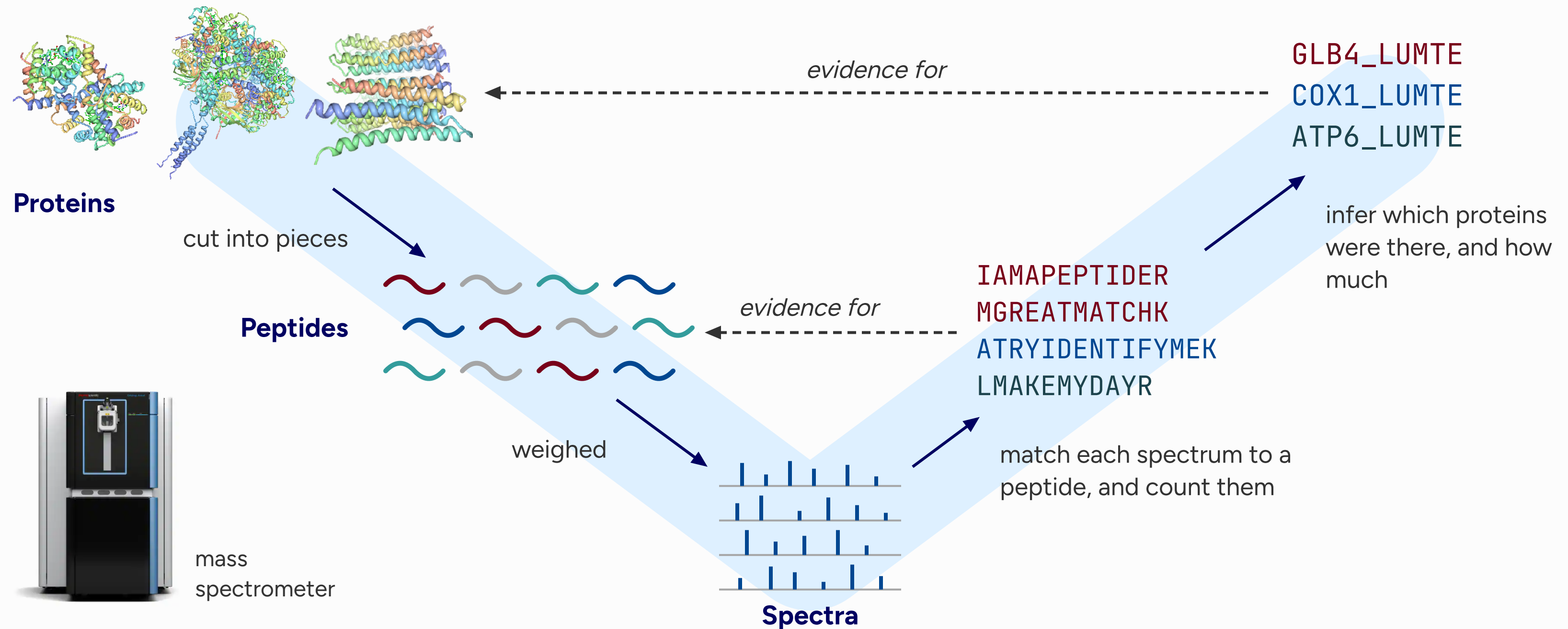


# Same DNA, different proteins

Same genome in every cell type and every life stage. What differs is which proteins are made, and how they are configured.



# From spectra back to proteins



# More famous people than me do mass spectrometry

Anthropic's CEO Dario Amodei was a postdoc within computational proteomics before he moved to other tasks.

Correspondence | Published: 10 October 2012

## A cross-platform toolkit for mass spectrometry and proteomics

[Matthew C Chambers](#), [Brendan Maclean](#), [Robert Burke](#), [Dario Amodei](#), [Daniel L Ruderman](#), [Steffen Neumann](#),

Nature Biotechnology 2012: ProteoWizard, the msconvert toolkit

## Building high-quality assay libraries for targeted analysis of SWATH MS data

Olga T Schubert<sup>1,2,9</sup>, Ludovic C Gillet<sup>1,9</sup>, [Ben C Collins](#)<sup>1,9</sup>, Pedro Navarro<sup>3</sup>, George Rosenberger<sup>1,2</sup>, Witold E Wolski<sup>1,4</sup>, Henry Lam<sup>5</sup>, [Dario Amodei](#)<sup>6</sup>, Parag Mallick<sup>6</sup>, Brendan MacLean<sup>7</sup> & Ruedi Aebersold<sup>1,8</sup>

Nature Protocols 2015: SWATH assay libraries

## Identification of a Set of Conserved Eukaryotic Internal Retention Time Standards for Data-independent Acquisition Mass Spectrometry\*<sup>§</sup>

Sarah J. Parker<sup>††</sup>, Hannes Rost<sup>§¶</sup>, George Rosenberger<sup>§¶</sup>, Ben C. Collins<sup>§</sup>, Lars Malmström<sup>||</sup>, [Dario Amodei](#)<sup>\*\*</sup>, Vidya Venkatraman<sup>††</sup>, Koen Raedschelders<sup>††</sup>, Jennifer E. Van Eyk<sup>†††</sup>, and Ruedi Aebersold<sup>§§§¶¶</sup>

Mol. Cell. Proteomics 2015: retention time standards

## Improving Precursor Selectivity in Data-Independent Acquisition Using Overlapping Windows

[Dario Amodei](#)<sup>1</sup>, Jarrett Egertson<sup>2</sup>, Brendan X. MacLean<sup>2</sup>, Richard Johnson<sup>2</sup>, Gennifer E. Merrihew<sup>2</sup>, Austin Keller<sup>2</sup>, Don Marsh<sup>2</sup>, Olga Vitek<sup>3</sup>, Parag Mallick<sup>1</sup>, Michael J. MacCoss<sup>2</sup>

<sup>1</sup>Department of Radiology, Stanford University, 3155 Porter Drive, Palo Alto, CA, USA

J. Am. Soc. Mass Spectrom. 2019: overlapping DIA windows

> Why is self-supervised learning important?

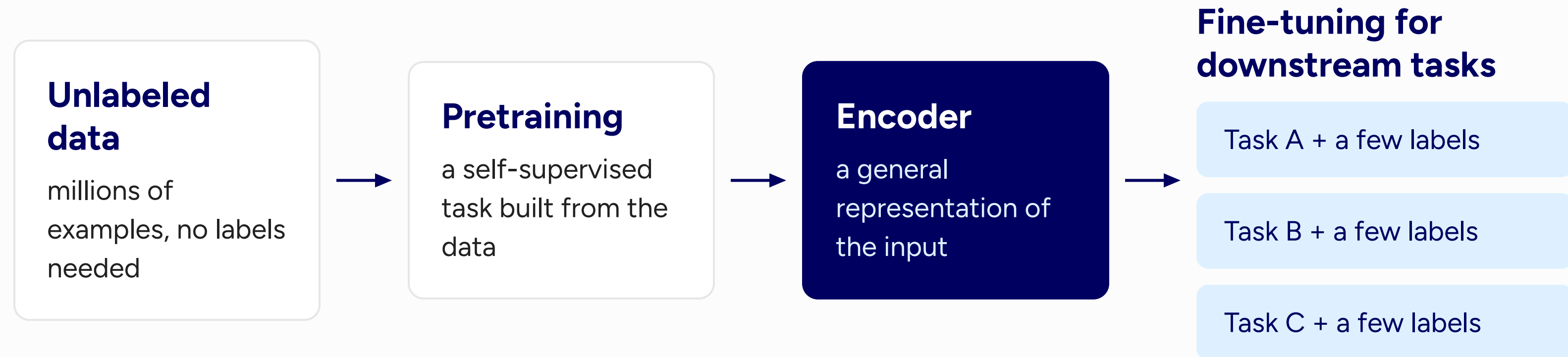
- Because most data comes without labels.

A label carries a few bits per example. The data itself carries millions.

**Eat some cake. The revolution will not be supervised.**

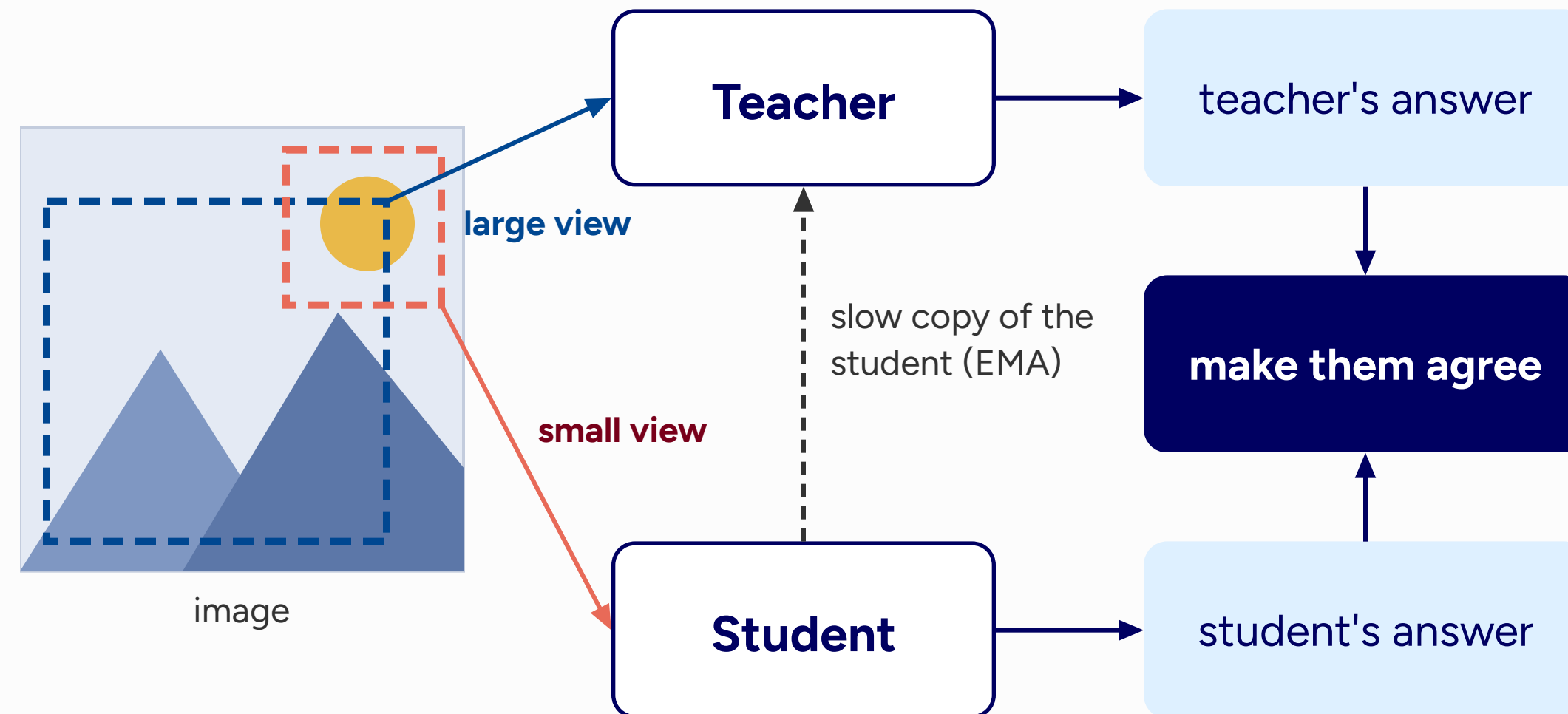
↳ showing slide: Pretrain once, fine-tune for many tasks

# Pretrain once, fine-tune for many tasks



Throughout the results we compare the same downstream task fine-tuned from the pretrained encoder and trained from random weights, which is called training from scratch.

# DINO: learning by agreeing with yourself



**The student** sees a small crop, **the teacher** a large one.

The student learns to give the teacher's answer; the teacher is a moving average of the student.

**No labels** anywhere.

DINO features, Meta AI



# Image views do not fit spectra

## Images

- Pixels on a regular grid
- A crop still shows the same object
- Colour and brightness can change freely

## Tandem mass spectra

- An unordered set of peaks
- Peak masses are exact to a few ppm: shifting one changes its meaning
- Hiding peaks and predicting them back is a weak task: spectra already miss peaks at random

**We need views that come from the physics of the measurement.**

● ● ● claude · lukas@arrhenius:~/dion

> Can we not just run DINO on mass spectra?

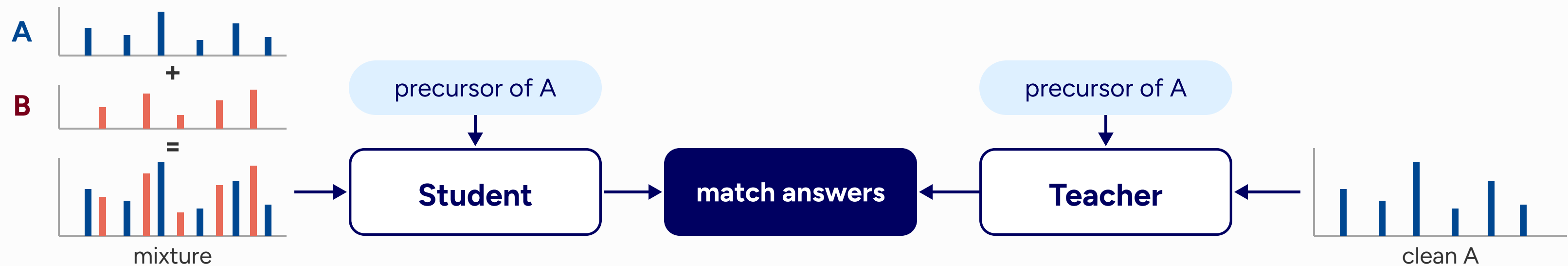
- The machinery, yes. The views carry over poorly: you can crop a spectrum, but neighbouring peaks are not related the way neighbouring pixels are.

dIon builds its views from fragmentation physics instead: every fragment comes from the precursor. It mixes two spectra, or hides part of one.

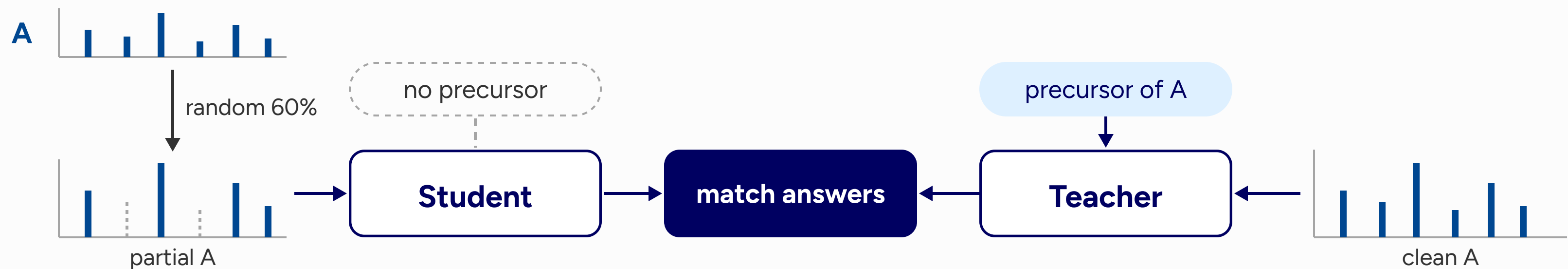
↳ showing slide: Two training tasks

# Two training tasks

## Task 1: find your peptide in a mixture



## Task 2: recognise a partial spectrum



## Pretraining with no peptide labels

**9.2 M**

unlabeled bacterial spectra (PXD010000)

**127 M**

parameters in a 15-block Transformer encoder

**4,800**

GPU hours: 64 NVIDIA GH200 for about 75 hours

**95%**

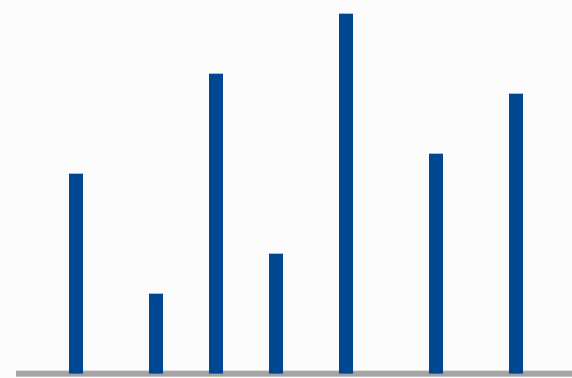
average GPU utilisation, median 100% (sampled on one of the 16 nodes)

> What can I use the pretrained model for?

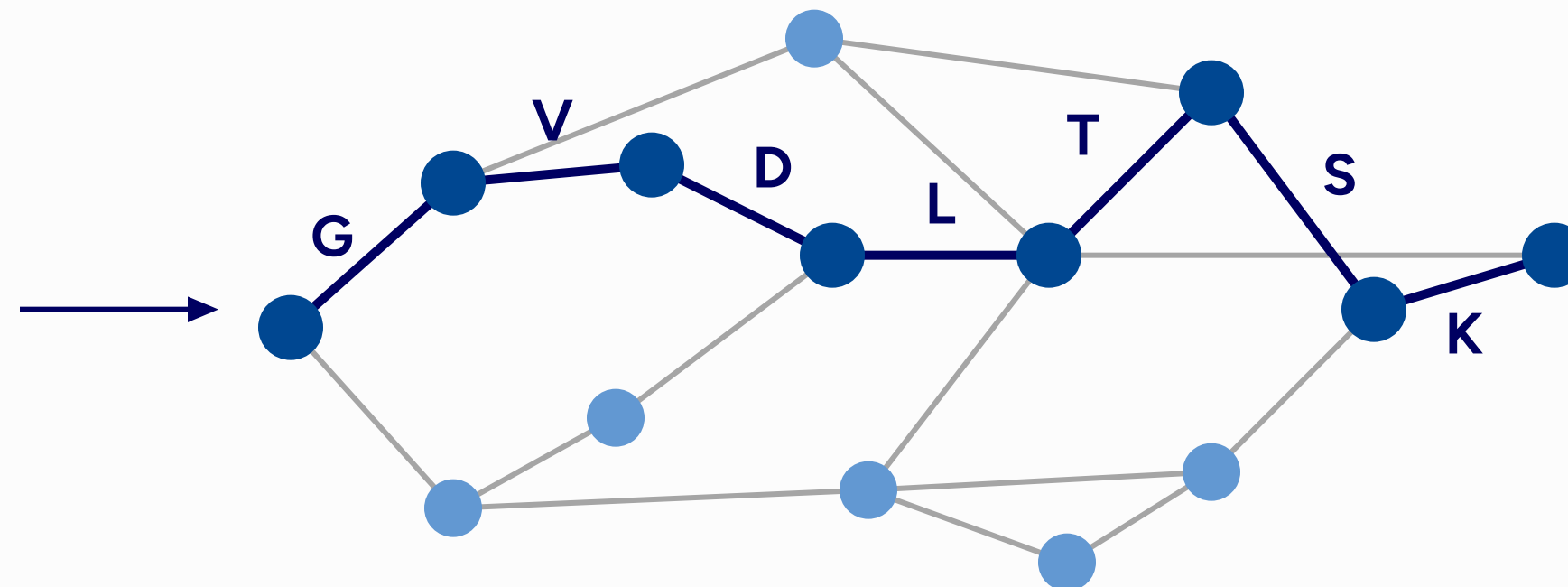
- Three downstream applications so far, all from the same pretrained encoder:
    1. *De novo* sequencing, fine-tuned: ahead of Casanovo and InstaNovo
    2. Chimeric spectra, frozen: the precursor mass picks the peptide
    3. Retrieval among near-identical spectra, frozen and fine-tuned
- └ showing slide: Classical approach to *de novo* sequencing

# Classical approach to *de novo* sequencing

Spectrum



Spectrum graph

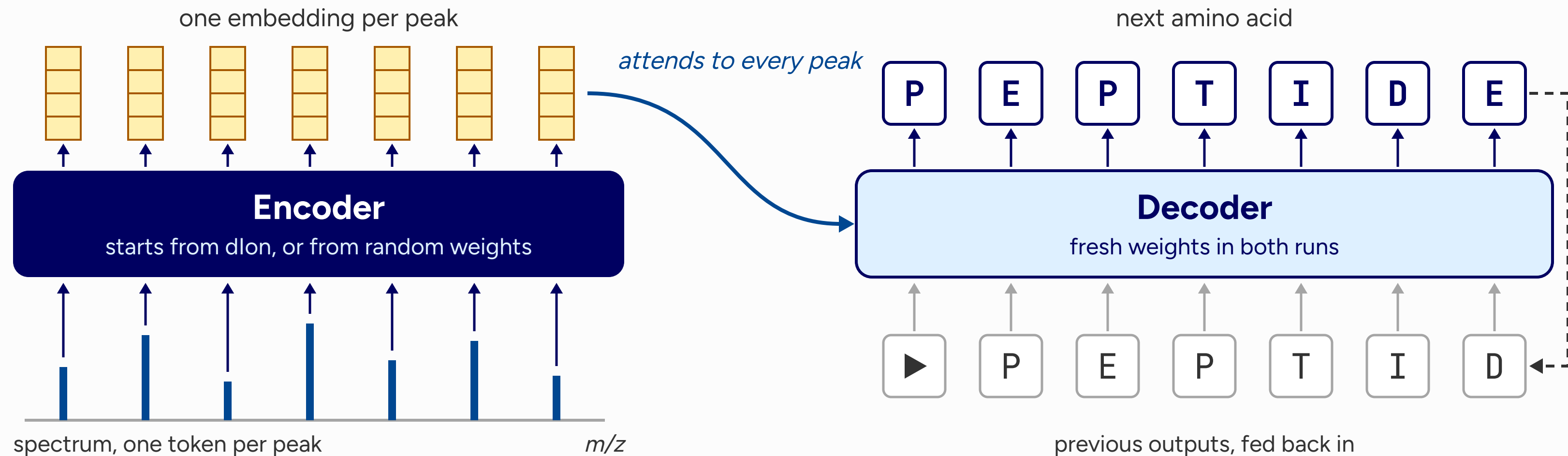


De novo reconstruction

**GVDLTSK**

Nodes are fragment masses. An edge joins two masses that differ by one amino acid.  
The best path through the graph spells the peptide.

# Fine-tuning task: *de novo* sequencing



**Fair test:** only the encoder's start differs.  
Same decoder, data and decoding.

**Tested** on MassIVE-KB and on Kingdoms,  
70 species.

**Score:** fraction of spectra whose whole  
peptide is correct.

# Ahead of supervised models on 70 species

All 70 species, 4.9 M spectra

dlon, trained on DNL-v1



InstaNovo 1.2.2



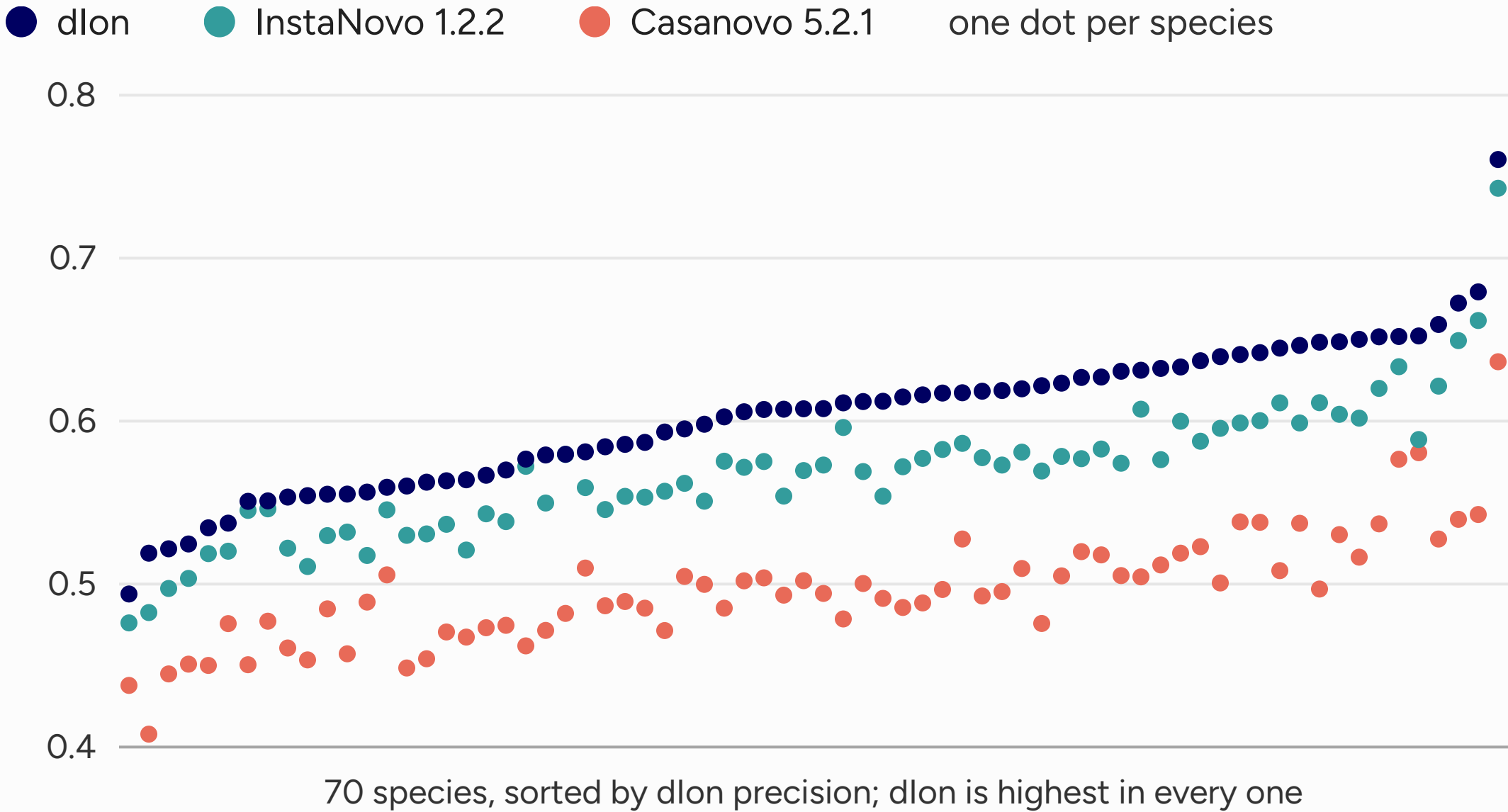
From scratch, trained on DNL-v1



Casanovo 5.2.1

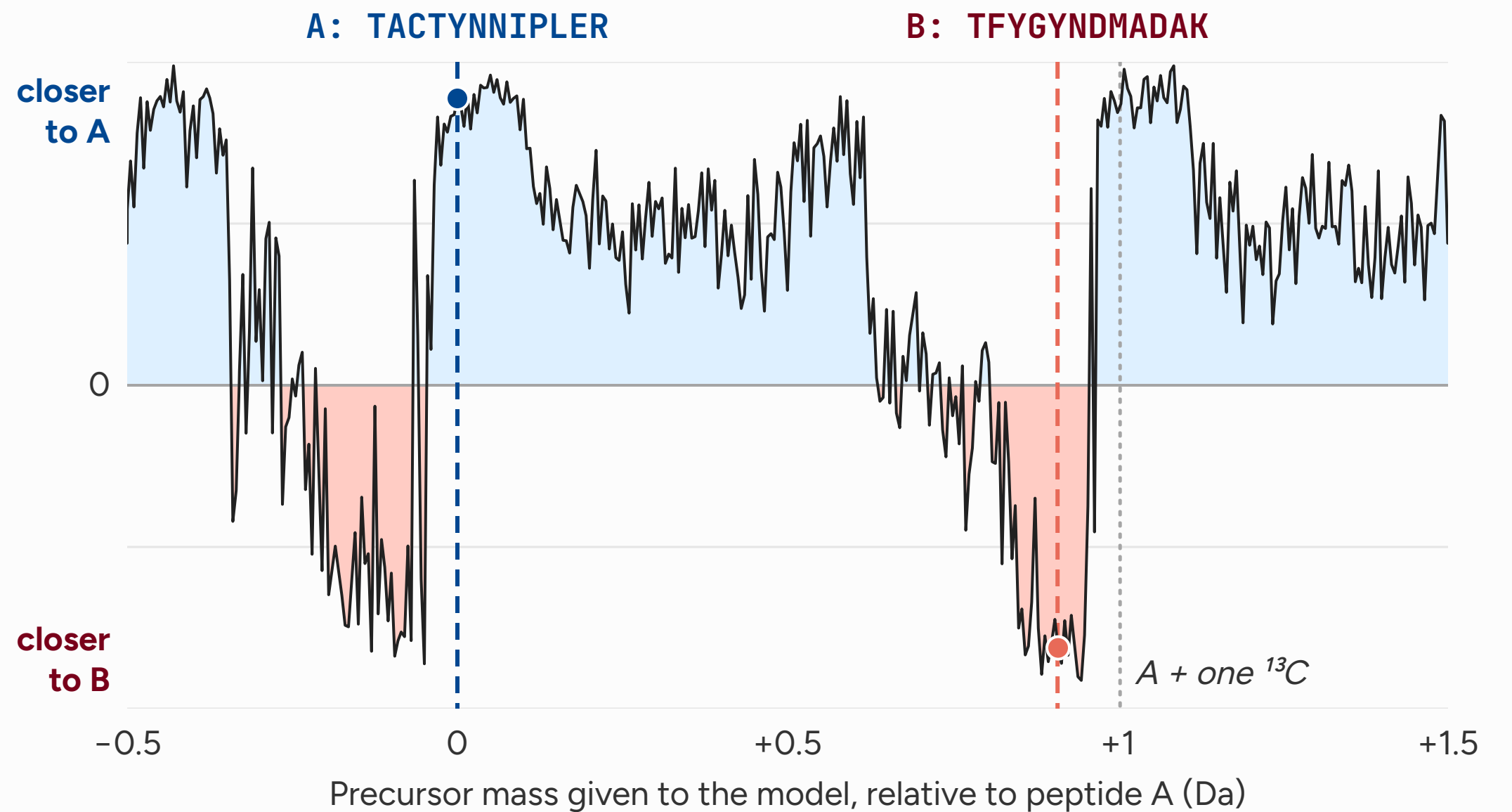
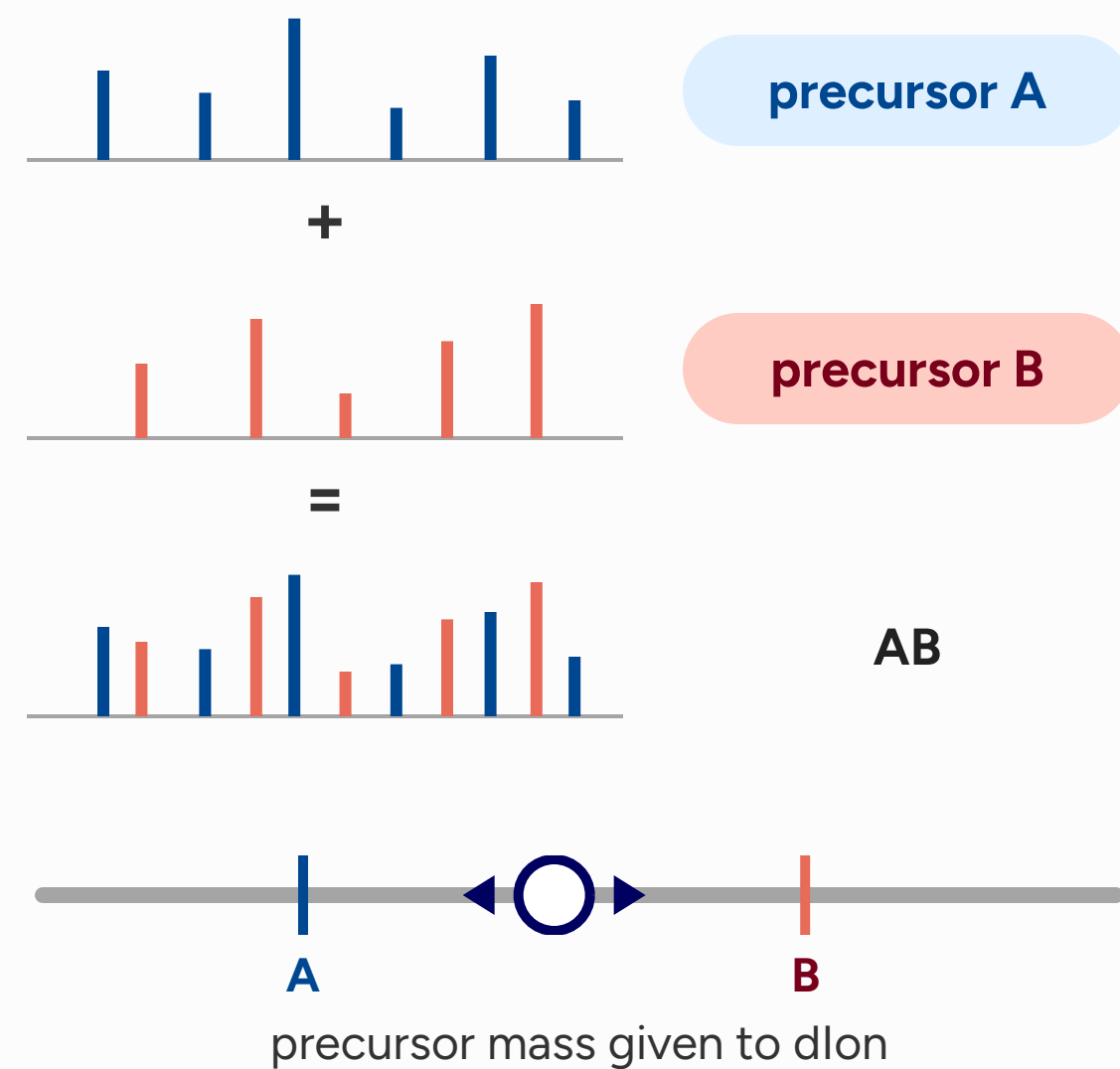


Peptide precision at full coverage, bars from 0.  
All models decode greedily with one beam.

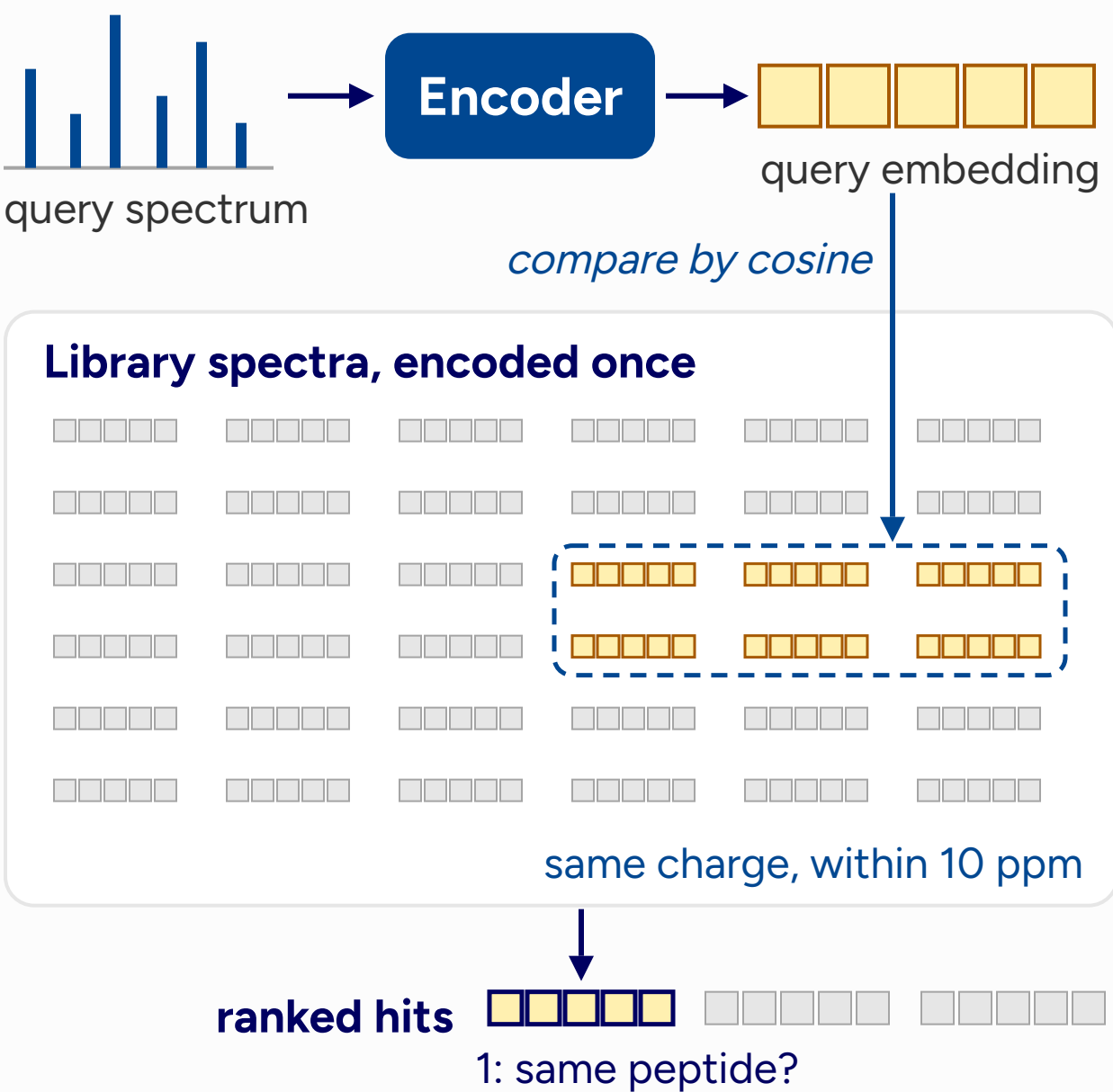


## 5 · RESULTS

# The precursor mass picks the peptide



# Retrieval among near-identical spectra



## Retrieval precision (mAP), Kingdoms test set



Same charge, precursor within 10 ppm. 70 species. Labels: peptide-supervised contrastive fine-tuning.

# Take-home messages

- 1 Labels from database search cannot be fully trusted. Self-supervision needs none.
- 2 A new invariance, every fragment belongs to its precursor, opens spectra to DINO training. Two tasks impose it: mixtures and partial spectra.
- 3 Fine-tuning from dlon raises *de novo* precision by 2.3 to 8.4 points, ahead of Casanovo and InstaNovo on all 70 species.
- 4 If you or your users are new to bioinformatic format conversions, SLURM syntax or git merges, run a coding agent on the HPC system itself.



Swedish  
Research  
Council



SciLifeLab

Thank you!



Preprint

*Knut and Alice  
Wallenberg  
Foundation*

