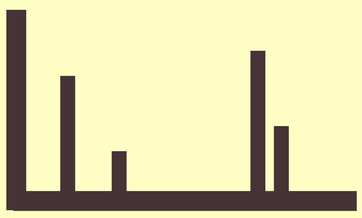
A LEARNED EMBEDDING FOR EFFICIENT JOINT ANALYSIS OF MILLIONS OF MASS SPECTRA

INTRODUCTION

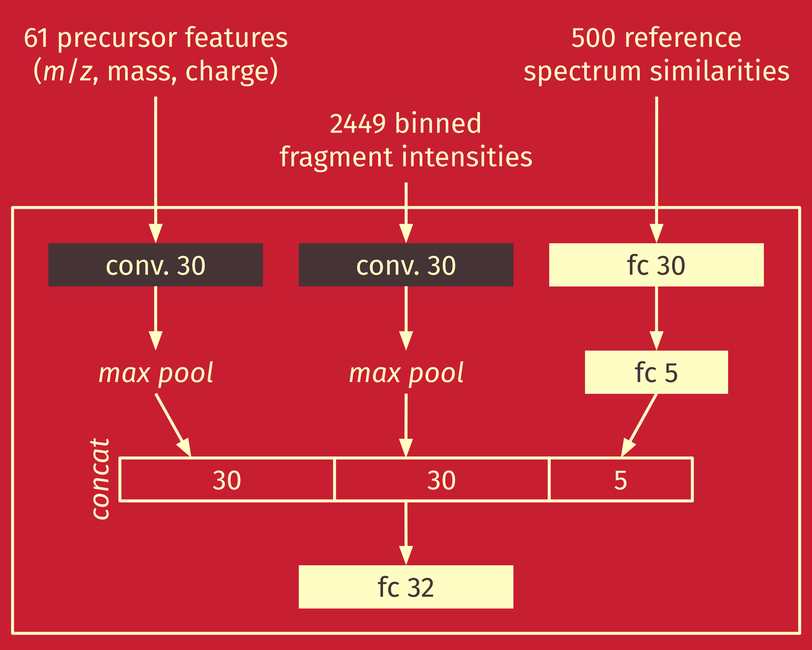


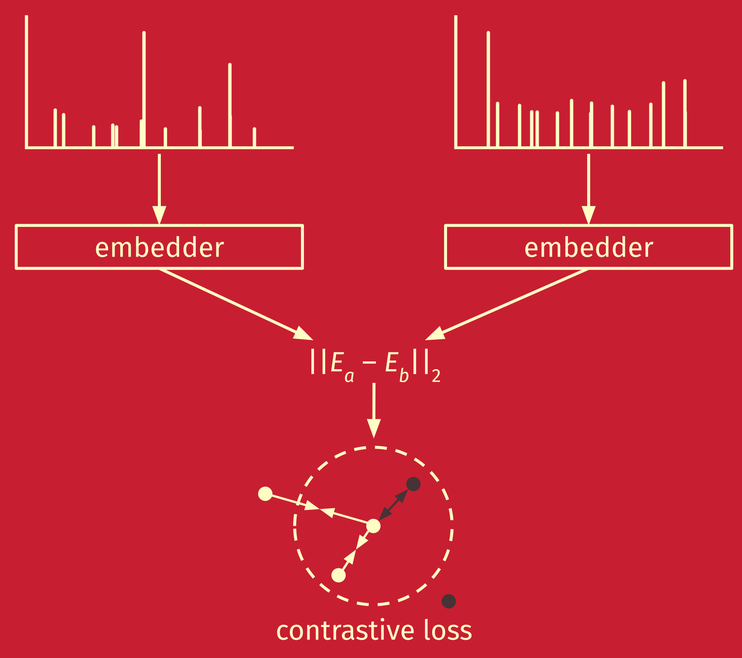
We propose to train a **Siamese neural network** using peptide–spectrum assignments to **embed spectra in a low-dimensional space** such that spectra generated by the same peptide are close to one another. We demonstrate that this learned embedding captures latent properties of the mass spectra, clusters related spectra in the low-dimensional space, and identifies the "dark matter" of the human proteome.

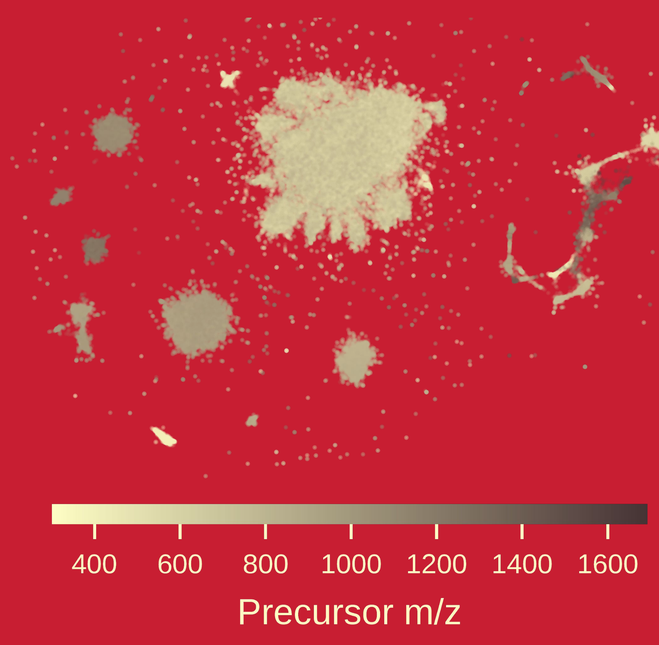
GLEAMS

The network, called "GLEAMS" (GLEAMS is a Learned Embedding for Annotating Mass Spectra), enables the efficient joint analysis of millions of mass spectra.

- Input to the embedder network consists of three feature types: **precursor attributes**, binned **fragment intensities**, and similarities to a set of **reference spectra**.
- A **Siamese network** containing two instances of the embedder with tied weights is trained using the **contrastive loss** function (Hadsell *et al. CVPR 2006*).
- The **learned embedding captures latent properties** of the mass spectra, such as precursor mass and charge, and protein modifications correspond to translations in the latent space.





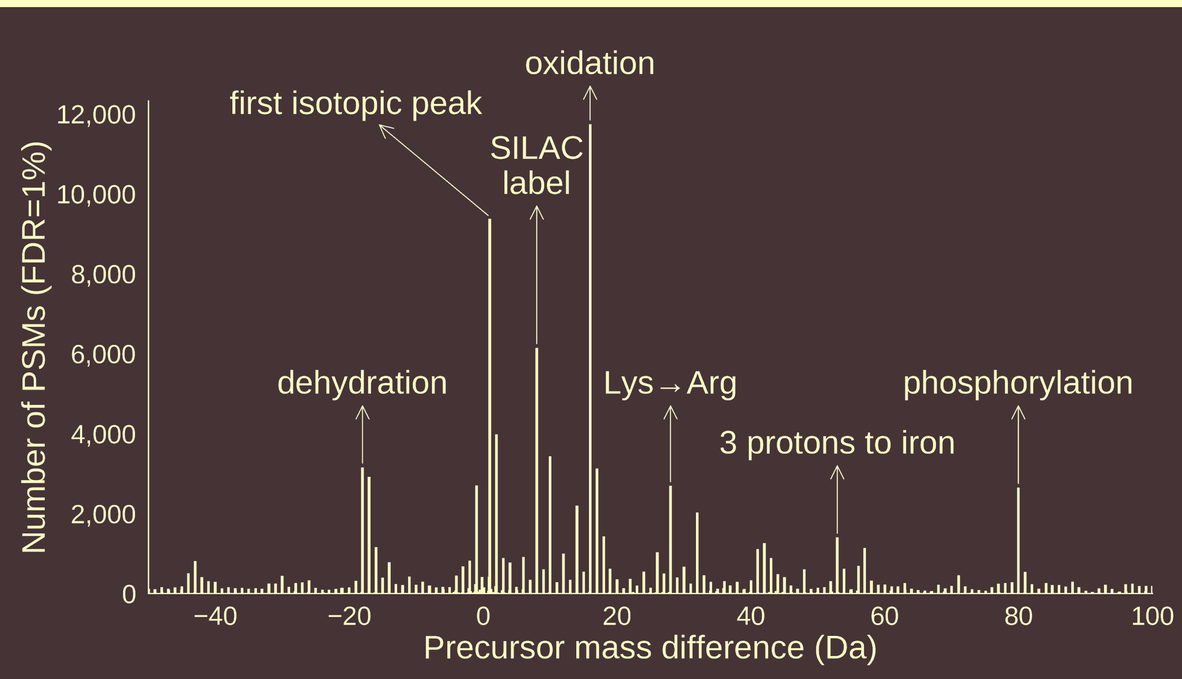


ELUCIDATING THE "DARK PROTEOME"

The "**dark proteome**" consists of spectra that are observed repeatedly across many experiments but consistently remain unidentified. We process hundreds of millions of spectra from the MassIVE-KB human dataset using:

- **DBSCAN clustering** in the low-dimensional space to propagate identifications within high-confidence clusters.
- **Open modification searching** using the ANN-SoLo spectral library search engine to identify modified peptides.

For a subset of the MassIVE-KB dataset, consisting of 4,808,492 spectra, we identify 759,505 previously unknown spectra, corresponding to a 132.6% increase in identifications relative to standard database search.



Wout Bittremieux[1*], Damon H. May[2], Jeffrey Bilmes[2], William Stafford Noble[2]

[1]University of California San Diego, La Jolla, CA, USA; [2]University of Washington, Seattle, WA, USA
*wbittremieux@health.ucsd.edu