Integration of the deep learning prediction tool Prosit into Skyline for high-accuracy, on-demand fragment intensity and iRT prediction

Tobias Rohde[1], Tobias Schmidt[2], Bernhard Kuster[2, 3], Michael J. MacCoss[1], Mathias Wilhelm[2], Brendan MacLean[1]

[1]Department of Genome Sciences, University of Washington, Seattle, WA 98195, [2]Chair of Proteomics and Bioanalytics, Technical University of Munich, Freising, Germany,
[3]Bavarian Center for Biomolecular Mass Spectrometry, Freising, Germany

Introduction:

Acquisition methods in Mass spectrometry-based proteomics heavily benefit from libraries for matching MS/MS spectra and choosing transitions. **Skyline** is a popular open-source tool for building and analyzing such methods, but like most other tools, requires empirically measured spectral libraries. These libraries are usually acquired by time-consuming and potentially expensive DDA experiments. While publicly available spectral libraries can be used as well, they are often incomplete and may have been acquired using different LC/MS settings. Recently, a deep neural network named **Prosit** has been developed to predict MS/MS fragment ion intensities and retention time indices (iRT) with high accuracy. **Skyline** is the first tool into which **Prosit** has been integrated. Usage of Prosit in Skyline can save users time and money in many workflows, such as targeted assays (**Figure 1.**)

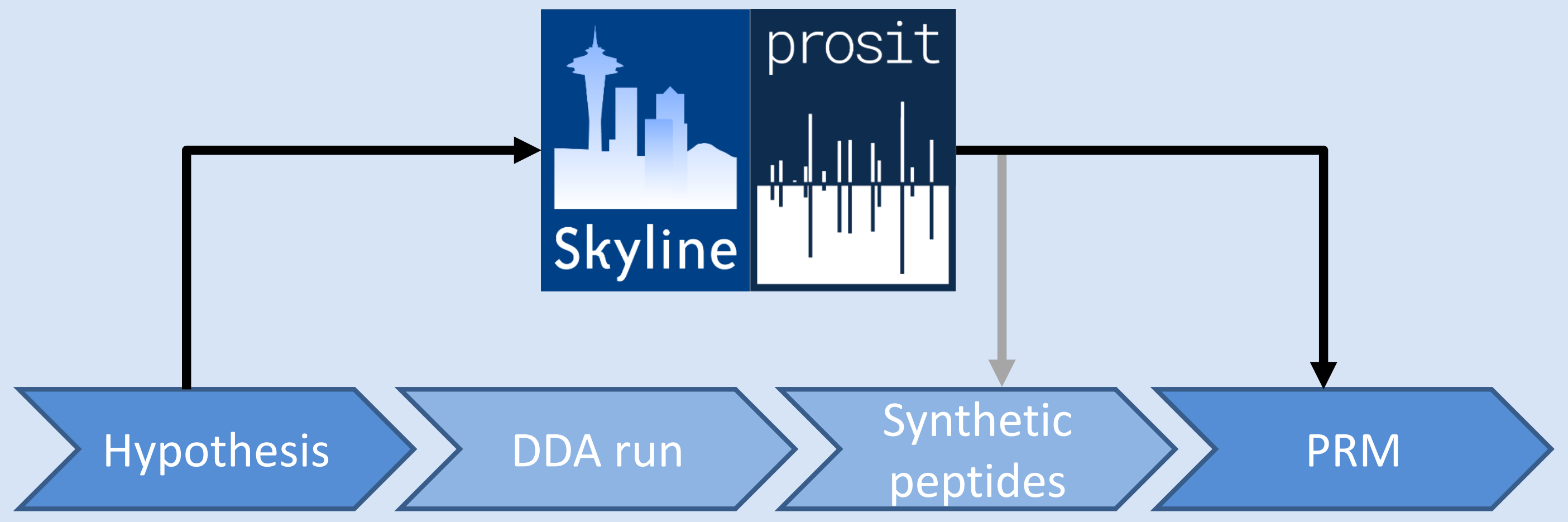


Figure 1. When developing a targeted assay for identifying or quantifying a set of peptides, Skyline with Prosit can eliminate the need for DDA runs and potentially the use of synthetic peptides for verification.

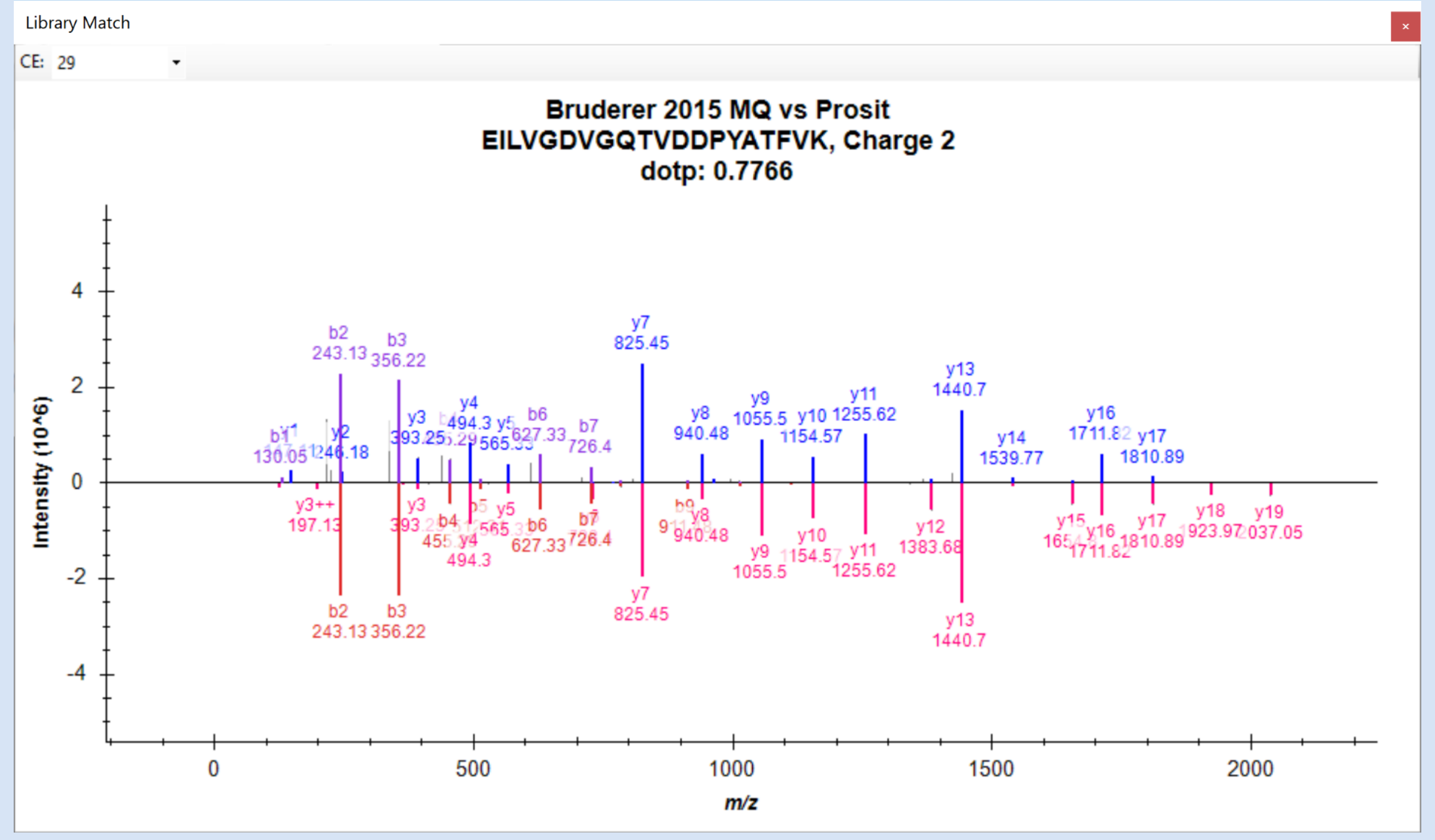


Figure 4. Mirror plot in Skyline comparing the experimental spectrum (blue) of *EILVGDVGQTVDDPYATFVK* (z=2) with the Prosit prediction (red). In the title the normalized contrast angle (dotp) is displayed (A value of 1 indicates a perfect match).

Overview:

Skyline is a windows client for building SRM/MRM, PRM, DIA/SWATH and DDA methods and analyzing the resulting mass spectrometer data.

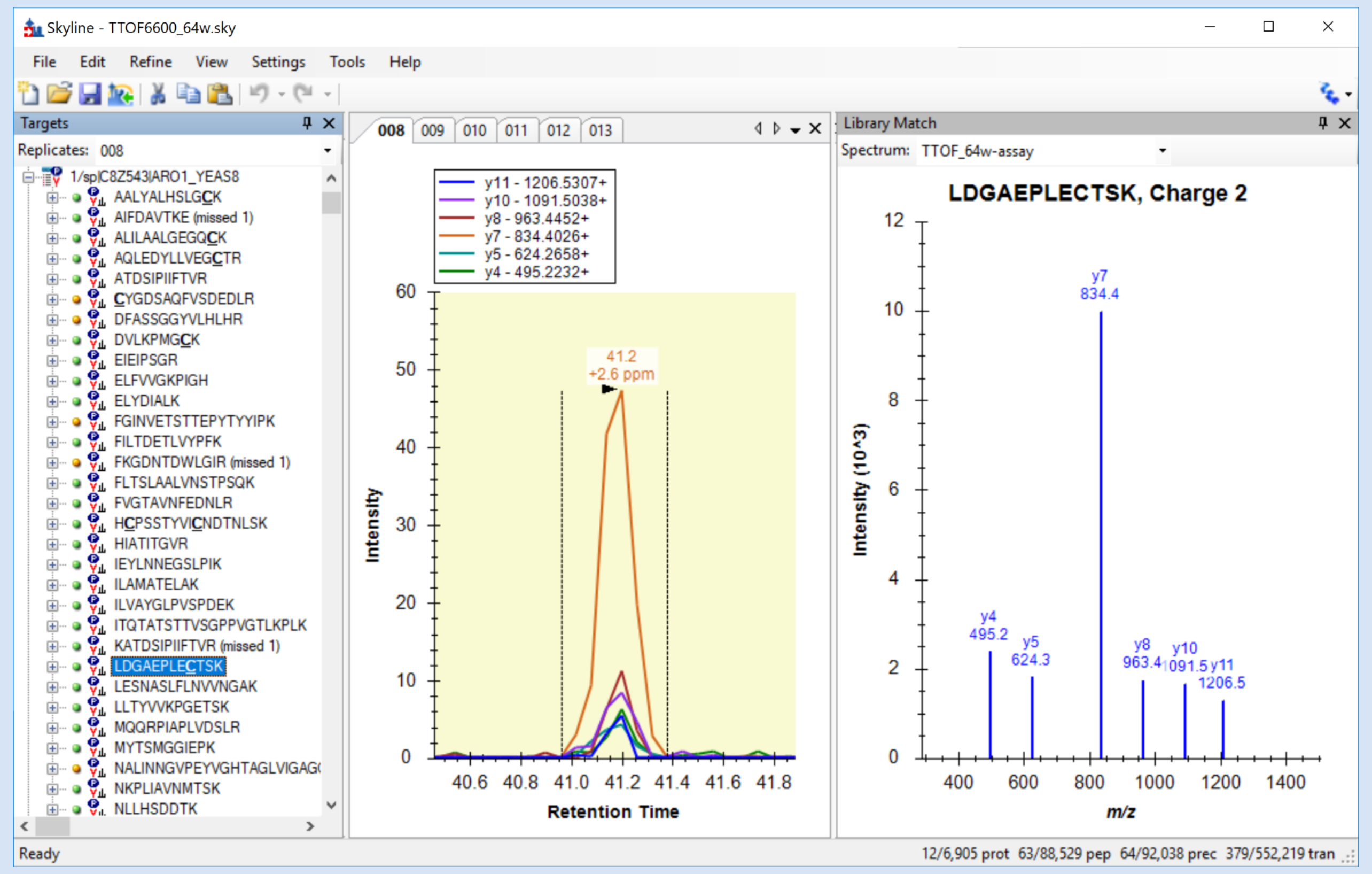


Figure 2. Left: The targets window, where a charge 2 precursor is selected. Middle: The chromatogram of the selected precursor, zoomed to the best peak. Right: The mass spectrum from the assay library matching the selected precursor.

Prosit is a deep neural network for predicting MS/MS fragment ion intensities and indexed Retention Time (iRT) values.

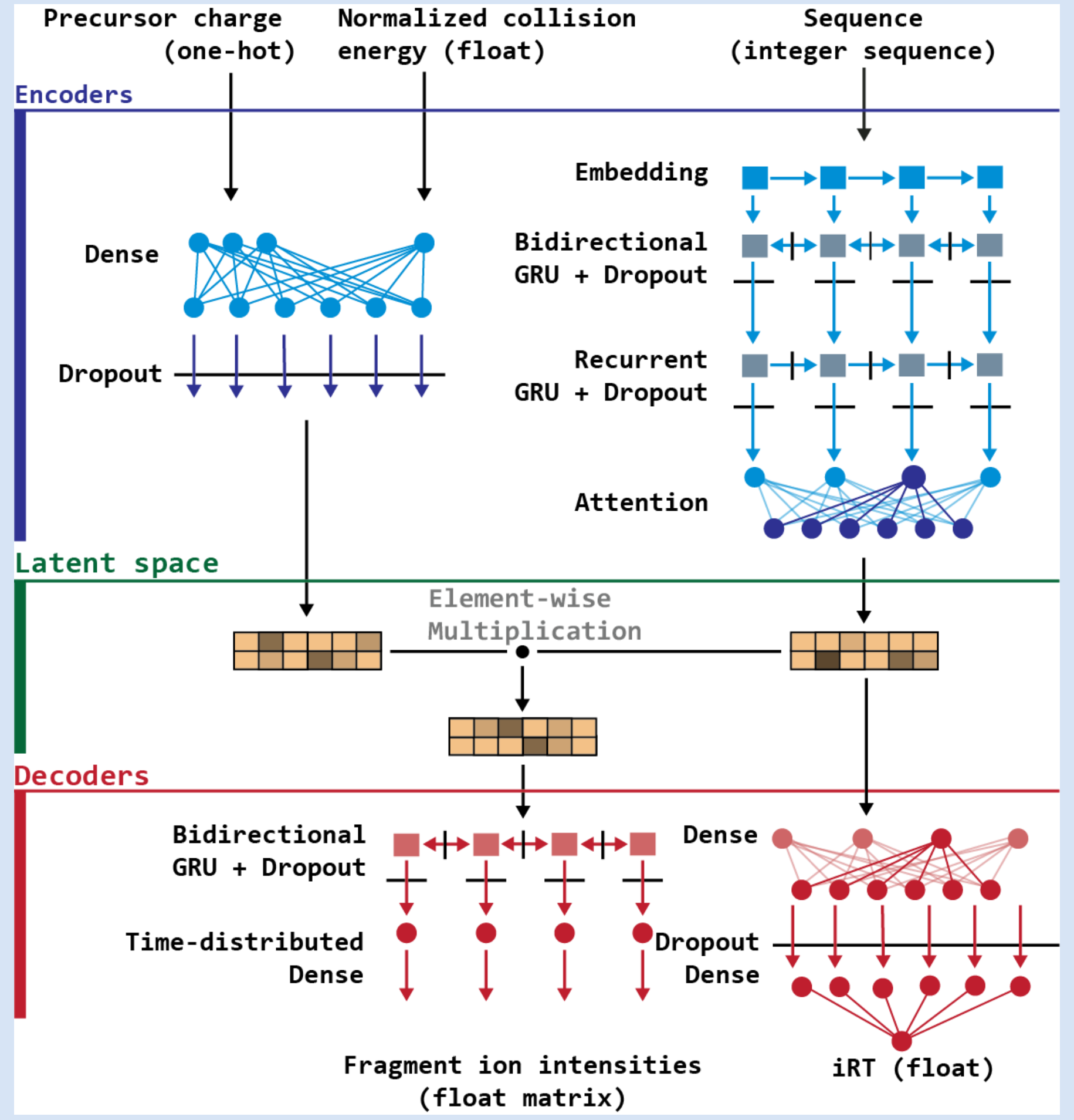
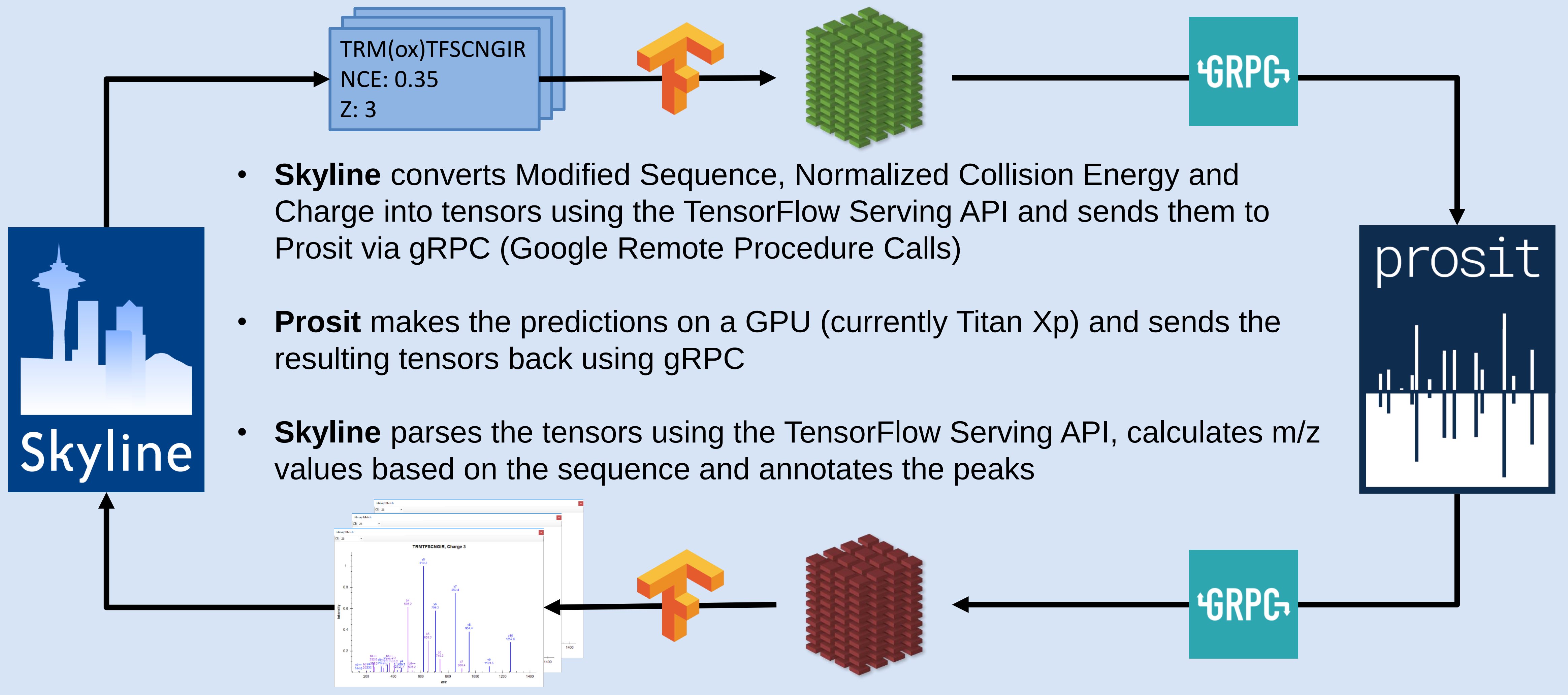

Precursor charge (one-hot)
Normalized collision energy (float)
Sequence (integer sequence)
Encoders
Embedding
Dense
Bidirectional GRU + Dropout
Dropout
Recurrent GRU + Dropout
Attention
Latent space
Element-wise Multiplication
Decoders
Bidirectional GRU + Dropout
Dense
Time-distributed Dense
Dropout Dense
Fragment ion intensities (float matrix)
iRT (float)

Figure 3. Architecture of the MS/MS and iRT models. The architecture is based on Natural Language Processing (NLP) models. The models were trained on ~460.000 tryptic human peptides. (ProteomeTools)



- **Skyline** converts Modified Sequence, Normalized Collision Energy and Charge into tensors using the TensorFlow Serving API and sends them to Prosit via gRPC (Google Remote Procedure Calls)
- **Prosit** makes the predictions on a GPU (currently Titan Xp) and sends the resulting tensors back using gRPC
- **Skyline** parses the tensors using the TensorFlow Serving API, calculates m/z values based on the sequence and annotates the peaks

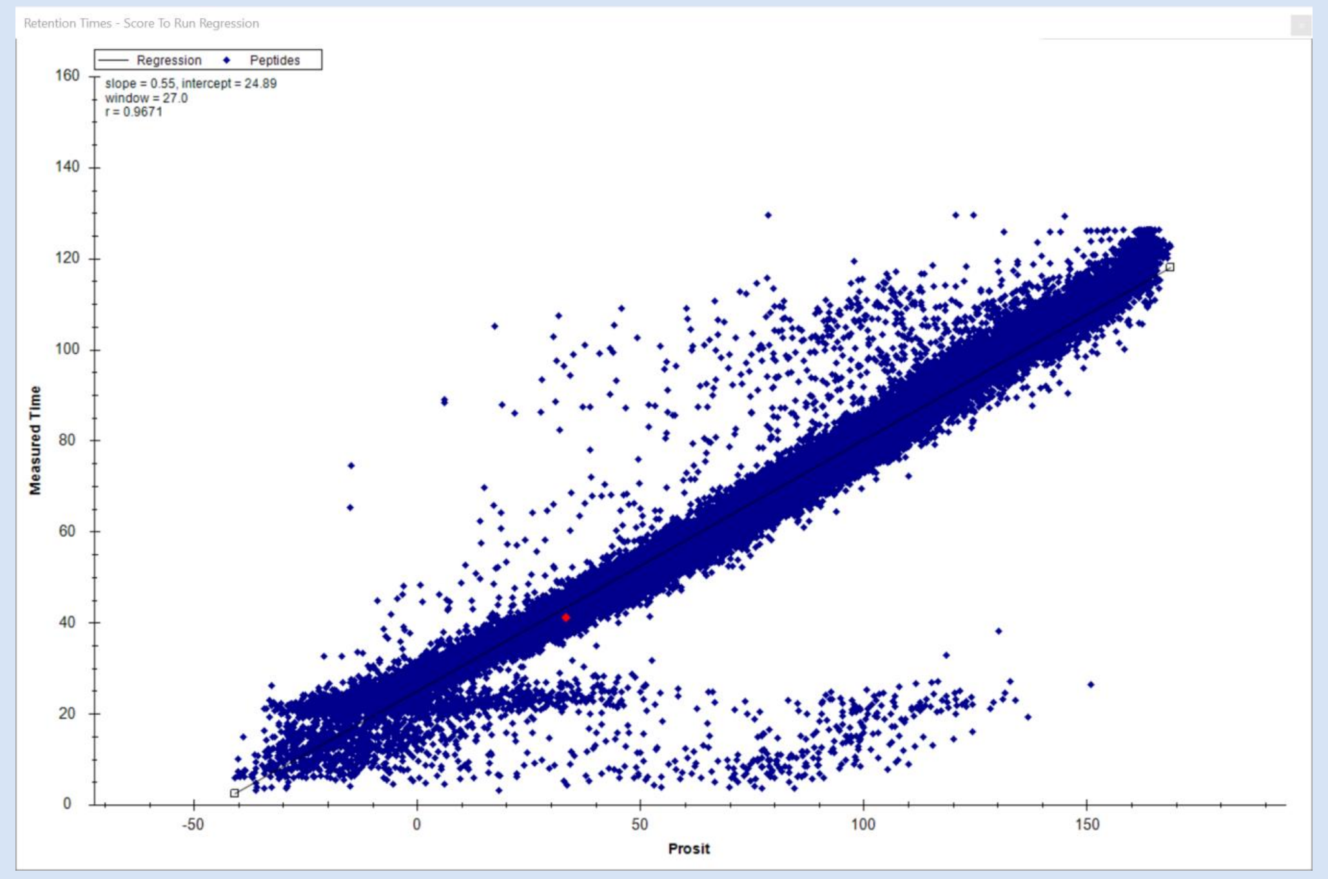


Figure 5. Retention time regression using iRT predictions from Prosit

Results:

We have found that **Prosit** spectral libraries are larger than experimental libraries (assuming the peptides of interest are supported by **Prosit**) and of similar quality. A benchmark DIA experiment has shown the same number of peptide identifications compared to when using an experimental library. Furthermore we have seen a significant decrease in retention time regression residuals when using **Prosit's** iRT predictions instead of algorithms such as SSRCalc3. Lastly, we expect that the integration of **Prosit** into **Skyline** will serve as a reference for other developers to benefit from **Prosit** predictions in the future.

Current work:

Recently a new model has been under development for predicting the predominant charge state of a peptide and potentially the relative intensities of all occurring charges. The model is based on previous **Prosit** architectures. The model is trained on DDA data, which in the past has been shown to not generalize well. However we still expect the current model to perform reasonably well on DIA data but are also looking for a large enough DIA dataset to train the model.

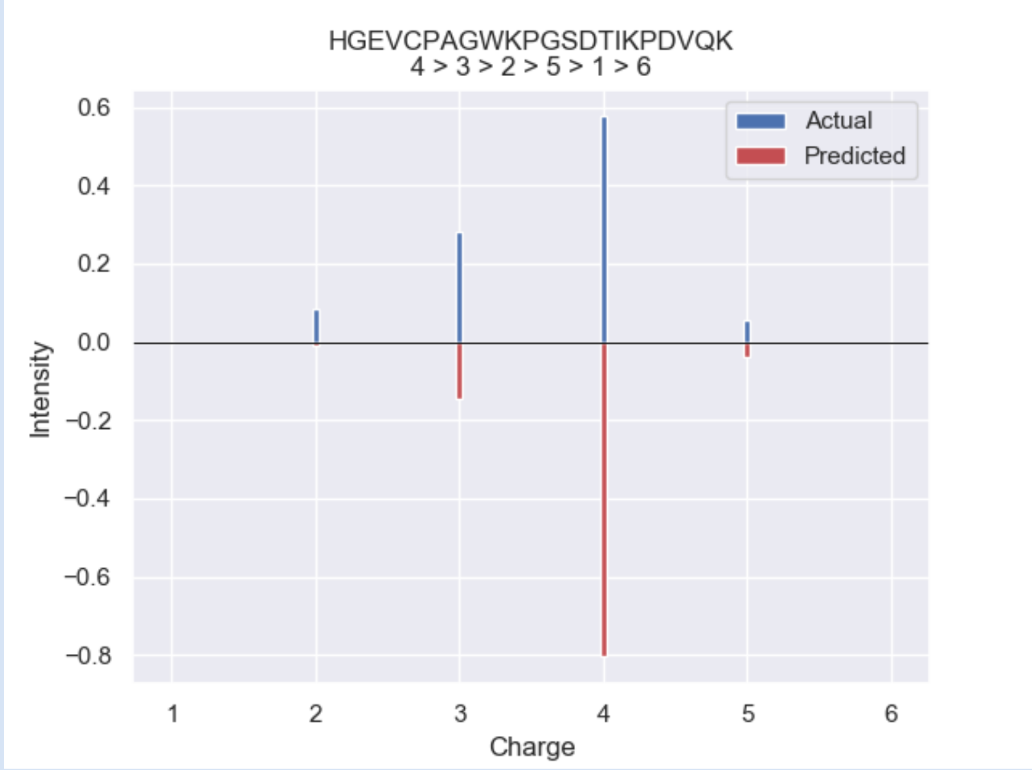


Figure 6: A precursor charge state distribution mirror plot. The graph displays predictions from a regular regression, while the ordering in the title was predicted through a ranking model trained as a Siamese neural network. Preliminary results show accurate predictions for DDA data; however the model still needs to be evaluated on DIA data.