DIA-NN: enabling peptidoform confidence in DIA proteomics

Katharina Faisst, Kate Lau, Justus Grossmann, Lukasz Szyrwiel, Ludwig R. Sinn, Vadim Demichev
Laboratory for Quantitative Proteomics, Charité – Universitätsmedizin Berlin

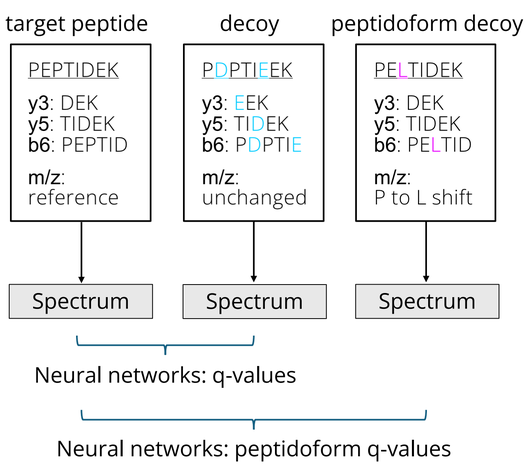
The challenge of DIA

Data-independent acquisition (DIA) proteomics has a range of advantages, from high proteomic depth and data completeness to the capability for precise and accurate quantification[1] of peptides and proteins. Further, DIA scales well to high-throughput workflows and experiments comprising thousands of samples. In the past years, DIA has gained capabilities for reliable identification and localisation of post-translational modifications (PTMs)[2] as well as for multiplexing[3], further broadening the range of its applications. However, so far DIA has had a key limitation: **lack of peptidoform confidence**. This matters in numerous applications which require distinguishing amino acid substitutions, such as:

- **Metaproteomics**: distinguishing between orthologues;
- **Population-scale plasma proteomics**: matching sequence variants to correct spectra, including when heterozygous;
- **General proteomics**: can mismatched proteoform assignment affect protein quantification?

Neural network-based peptidoform scoring

DIA-NN[4] is based on the application of neural networks to distinguish between true and false signals. The networks are trained based on **target** peptide-spectrum matches (PSMs) – originating from the peptides of interest – and so-called **decoy** PSMs, obtained by matching in silico-generated faux peptides that are not present in the sample. Typically, the decoys used have very different in silico-generated spectra, i.e. all fragment masses are different. We now show that making decoys similar to targets, combined with neural network-based scoring, enables peptidoform confidence.



Peptidoform confidence in DIA: how?

- Quality MS1 signal, correlating with MS2 XICs
- Sequence coverage by fragment ions

The above seems to suffice to achieve higher confidence than typical DDA, with still higher proteomic depth

Old DIA-NN PTM module[5]

- Generation of decoys via amino acid mutation at PTM sites
- Need to know specific PTMs
- Low confidence in any other PTMs – relevant when those are present in the sample

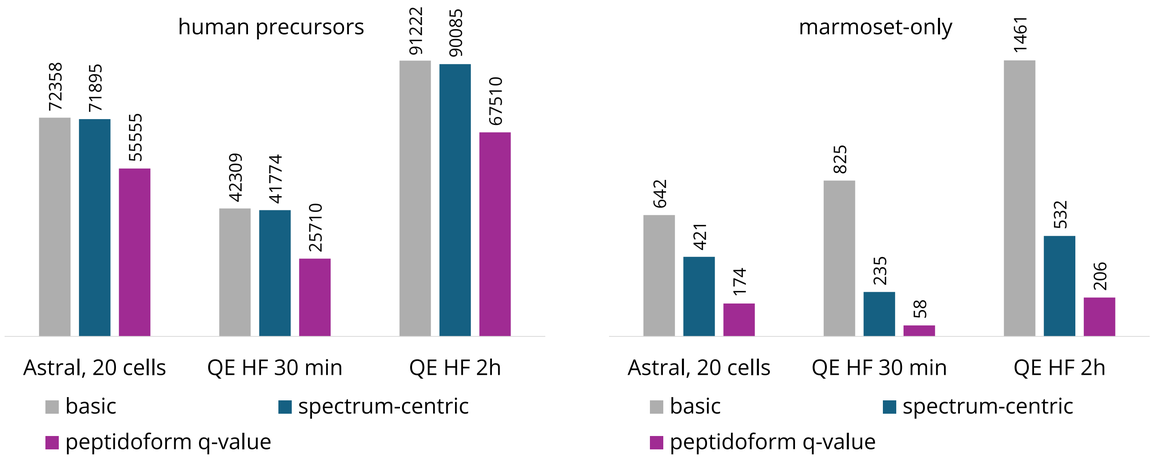
New peptidoform module

- Generation of decoys via amino acid mutation at a random position
- Works against any PTMs with mass shifts beyond the isotopic envelope
- Separate PTM-specific module responsible for PTM localisation

Benchmarks

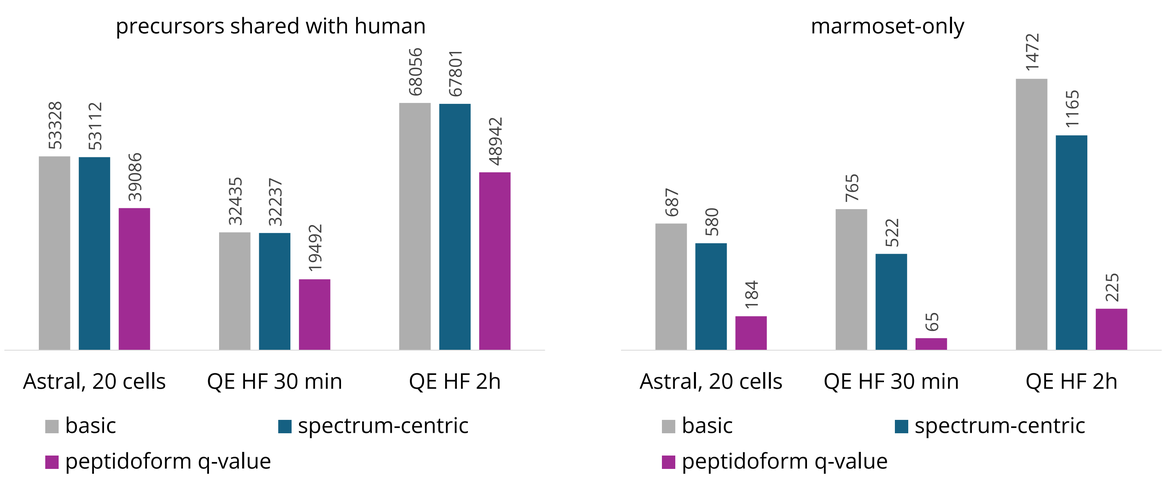
1. Searching human data against human + monkey (marmoset) database.

Marmoset-only peptides are counted as known false positives. This benchmark tests the ability of the software to choose the best peptide match for a spectrum out of known options. Basic search, spectrum centric module[4] (default in DIA-NN, v1.9.2) and the peptidoform scoring module are compared, q-value < 0.01 filter applied in each case. Runs: QE HF (30-min and 2h gradients, 2µg)[6], Orbitrap Astral (40 SPD, 4m/z isolation windows, 20 HeLa cells)[7].



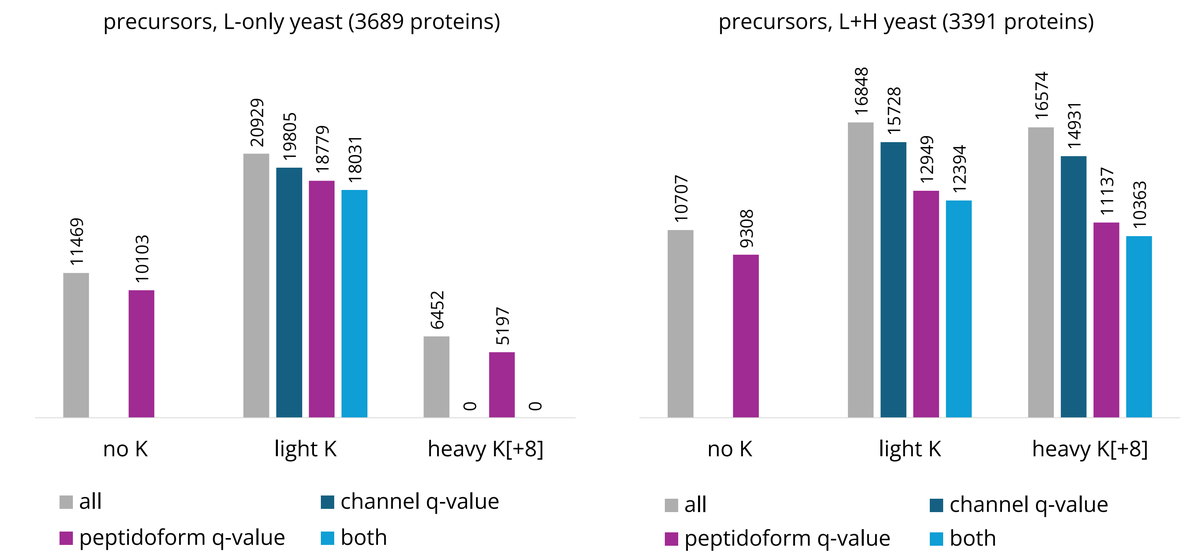
2. Searching human data against monkey database.

Marmoset-only peptides (not shared with human) are counted. This benchmark tests the ability of the software to determine if the peptide-spectrum match is correct when alternative peptidoforms are not known.



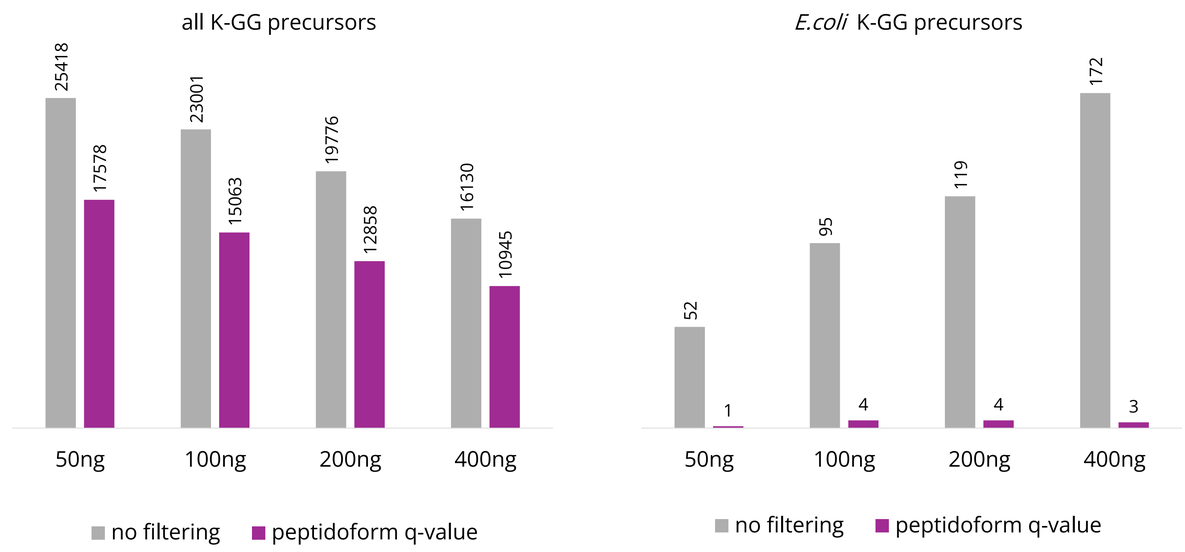
3. Slice-PASEF and multiplexing.

SILAC (light and heavy K) yeast lysC digests, 2-frame Slice-PASEF[8], Evosep 60 SPD coupled to timsTOF Ultra. This benchmark tests (i) the performance with very wide isolation windows and (ii) compatibility with multiplexing.



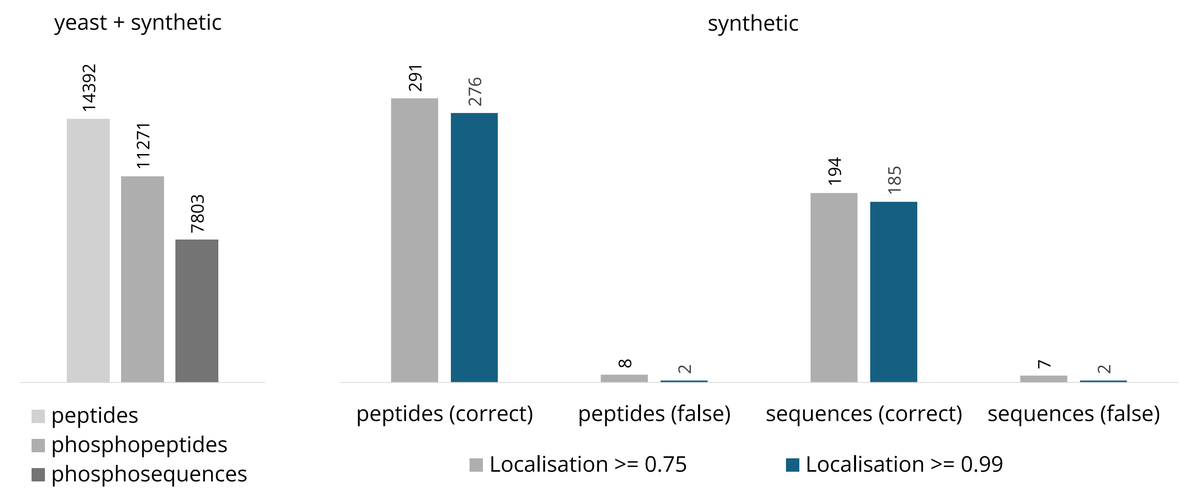
4. Ubiquitinomics: direct FDR validation.

E.coli digest spiked in different amounts (50ng – 400ng) in human K-GG-enriched sample[5], all *E.coli* K-GG peptides called by the software are false. MBR disabled.



5. Phosphoproteomics.

Yeast + synthetic phosphopeptides data set[9], reanalysed in ultra-fast mode with MBR.



Conclusions

- The spectrum-centric module in DIA-NN 1.9.2 ('No shared spectra' option) already ensures inherently low peptidoform FDR, if DIA-NN is searching all peptidoforms present in the sample (e.g. both unmodified and modified).
- The new peptidoform scoring module further reduces peptidoform FDR. Whether or not all possible peptidoforms are searched is irrelevant.
- Peptidoforms scoring is compatible even with very wide (100 m/z or greater) isolation windows as in Slice-PASEF.
- Peptidoform scoring complements channel-specific scoring in multiplexed DIA.

[1] Kistner et al., biorxiv, 2023
[2] Rosenberger et al. Nature Biotechnology, 2017
[3] Derks et al. Nature Biotechnology, 2023
[4] Demichev et al. Nature Methods, 2020
[5] Steger et al. Nature Communications, 2021
[6] Bruderer et al. MCP, 2017
[7] PXD049211 (Olsen lab)
[8] Szyrwiel et al, biorxiv, 2022
[9] Bekker-Jensen et al. Nature Communications, 2020

MSTARS

CHARITÉ
UNIVERSITÄTSMEDIZIN BERLIN