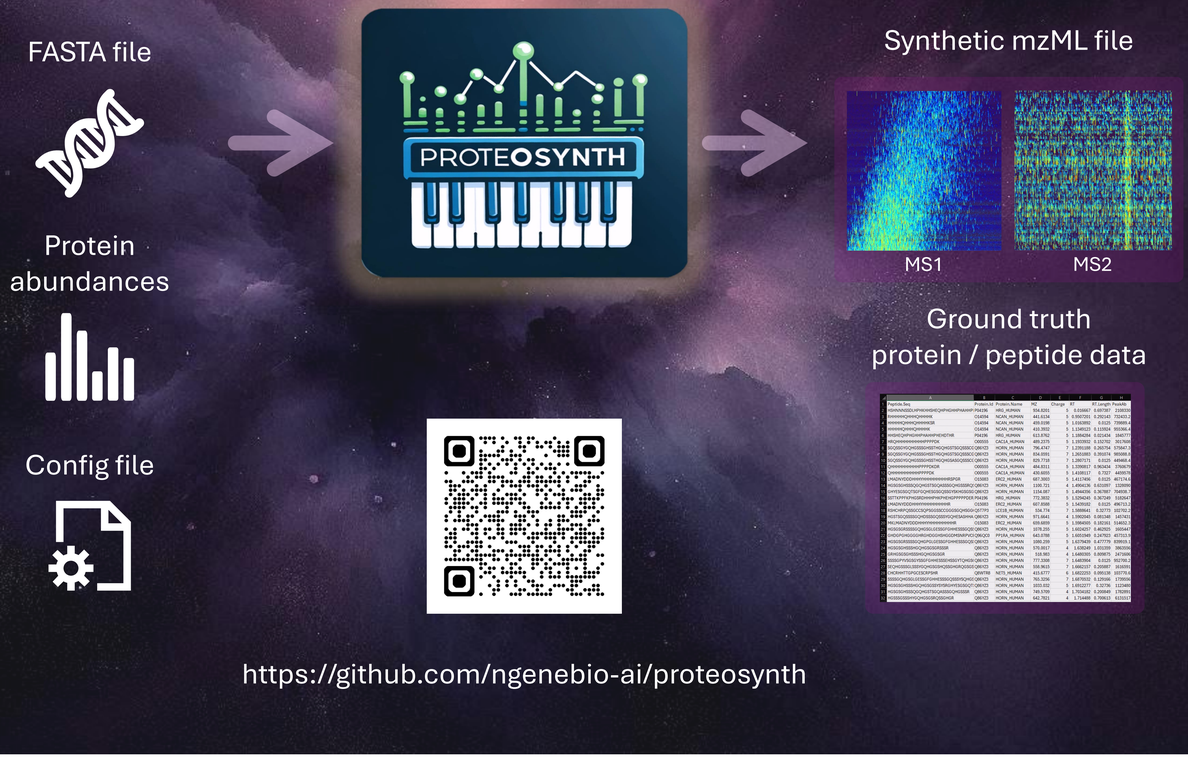
NGeneBioAI

Matthew N. Dailey, Dmitry Avtonomov, Eric Kim, Chris Wilkins, Stefanie Kairs, Sangtae Kim
NGeneBioAI, San Diego, CA

ProteoSynth

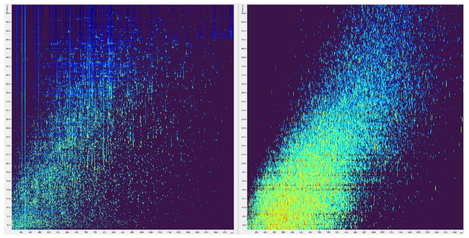
*Helps You **Test** Your Proteomics Data Processing **Pipelines** Against **Ground Truth***



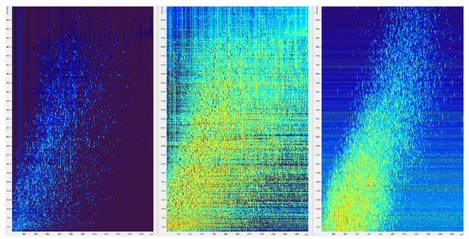
The Problem

- Evaluating tools that analyze proteomics mass spectrometry data is hard.
- Without ground truth, we might say, “the more IDs the better!”
- Simulators and emulators exist, but none are both public and currently supported.

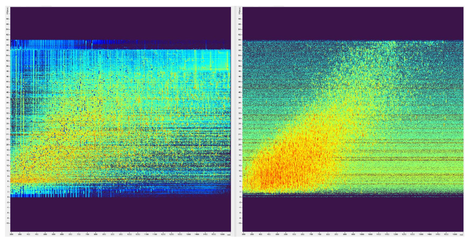
The Solution



Left: Survey MS1 scans from a real Mag-Net[1] enriched human plasma sample run on Exploris 480. **Right:** ProteoSynth synthesized mzML file from 2000 randomly sampled Human proteins with no noise.



Left: Real Survey MS1 scans as above. **Middle:** HDR MS1 scans, similar to BoxCar. **Right:** ProteoSynth synthesized mzML file from 2000 randomly sampled Human proteins with noise.

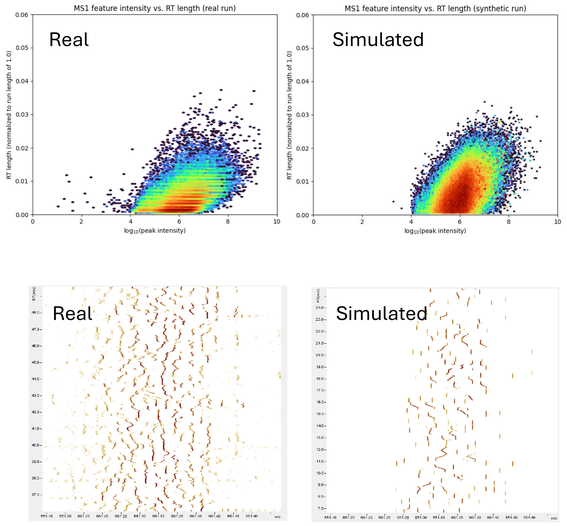


Within-scan dynamic range limit of the instrument is simulated in synthetic data, manifesting as the darker horizontal swaths – regions where fewer ions are observed. **Left:** Real. **Right:** Synthesized.

[1] We et al. (2024), Mag-Net: Rapid Enrichment of membrane-bound particles enables high coverage quantitative analysis of the plasma proteome. bioRxiv, PMC11014469.

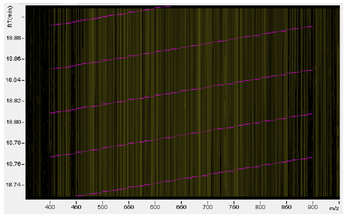
Features

- Customizable DIA, DDA schedules; output to mzML
- Customizable models for predicting peptide retention time, intensity, charge state, noise density / intensity
- Synthesis with or without noise peaks, intensity noise, and m/z jitter
- Simulating within-scan dynamic range cutoff
- Mass resolution – generate spectra with detailed isotopic fine structure or merge peaks according to simulated “Resolution” setting
- Transfer of MS1 intensity distribution and LC elution duration distribution to profiles extracted from real runs



Synthetic DIA Run Against DIA-NN

- 60-minute DIA runs, 10-Da MS2 isolation windows
- 2000 random proteins from UniProt Human
- Trypsin digestion in silico yields 176K peptides
- Random log normal protein abundances
- Predicted peptide charge state, retention time peaks
- Peptide intensities predicted, modulated by protein intensities, transformed to match real run intensity distribution
- Peptide retention time length distributions matched to real run distribution conditional on intensity
- Differential expression of one protein, TAU_HUMAN

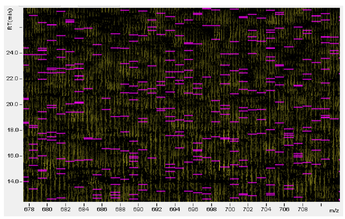


DIA-NN @ FDR=0.01	Precision	Recall	Actual FDR
Peptides, no noise	0.9924	0.3863	0.008
Proteins, no noise	0.8737	0.9935	0.126
Peptides, with noise	0.9962	0.0931	0.004
Proteins, with noise	0.9687	0.7274	0.031

- Relative protein abundance quantification for the target protein is perfect both with and without noise!

Synthetic DDA Run Against Sage

- All parameters identical to the DIA runs except for simulation of MS2 DDA isolations with 1 Da window.



Sage @ FDR=0.01	Precision	Recall	Actual FDR
Peptides, no noise	0.9928	0.1894	0.007
Proteins, no noise	0.9414	0.9560	0.059
Peptides, with noise	0.9867	0.0801	0.013
Proteins, with noise	0.9571	0.9030	0.043

Conclusion

ProteoSynth does not aim to replace real data in testing, but measurable success on synthetic data can and should be a bare minimum requirement for tool testing.

Conflict of interest disclosure: All authors are employees of NGeneBioAI, a company developing proteomics-based diagnostics solutions.