P270

https://doi.org/10.17952/38EPS.2026.P270

universität wien
Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences

Vienna Doctoral Schools

Can we Confidently Sequence the „Peptide Dark Matter“? Reaching a Consensus for De Novo Peptide Sequencing

Sebastian Johannes Eder[1,2,3], Manuela Schmidt[1,2], David Gomez-Varela[1,2]

1. University of Vienna, Division of Pharmacology and Toxicology,
2. University of Vienna, Bruker Daltonics Center of Excellence for Metaproteomics

Overview

Database search controls error by target-decoy competition. De novo sequencing has none, so it returns model-specific confidence scores rather than an error rate, and those scores are not calibrated. Any threshold placed on them is arbitrary, which leaves the unidentified majority of a run unquantifiable. Here I present QUORUM, a model to effectively increase accuracy and FDR control for de novo sequencing

How it works

CasaNovo[1], InstaNovo[2] and PrimeNovo[3] sequence the same spectrum independently. A gradient-boosted model scores every aligned residue from what all three proposed there together with the peak evidence supporting it and assembles fourth candidate. All four are described by one identical feature set, dominated by the ion evidence class at each boundary and by agreement between observed and predicted retention time. A selector, based on a logistic regression, selects the best one and assigns a probability score.

The consensus beats every model that built it

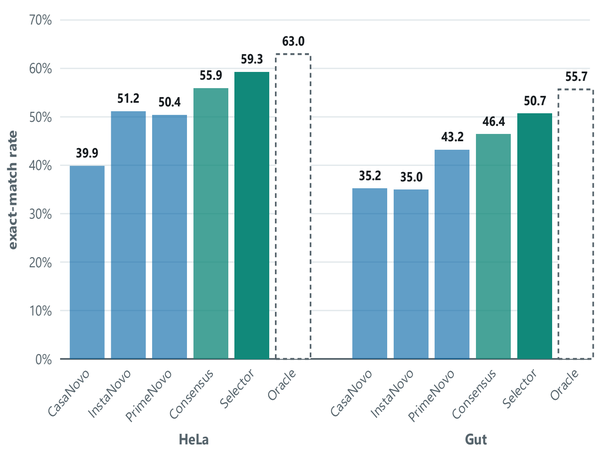


Fig. 1: Model accuracy against database-search-derived peptides. Measured on a HeLa lysis sample and a Gut fecal metaproteome fraction. Consensus refers to the assembled candidate. Oracle refers to the maximum possible percent, where at least 1 of the models predicted the correct peptide.

Calibration is conservative off training

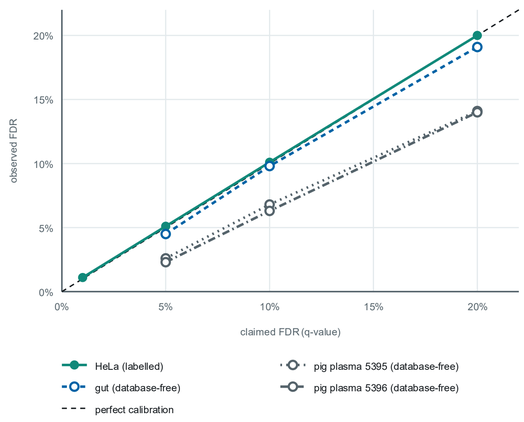
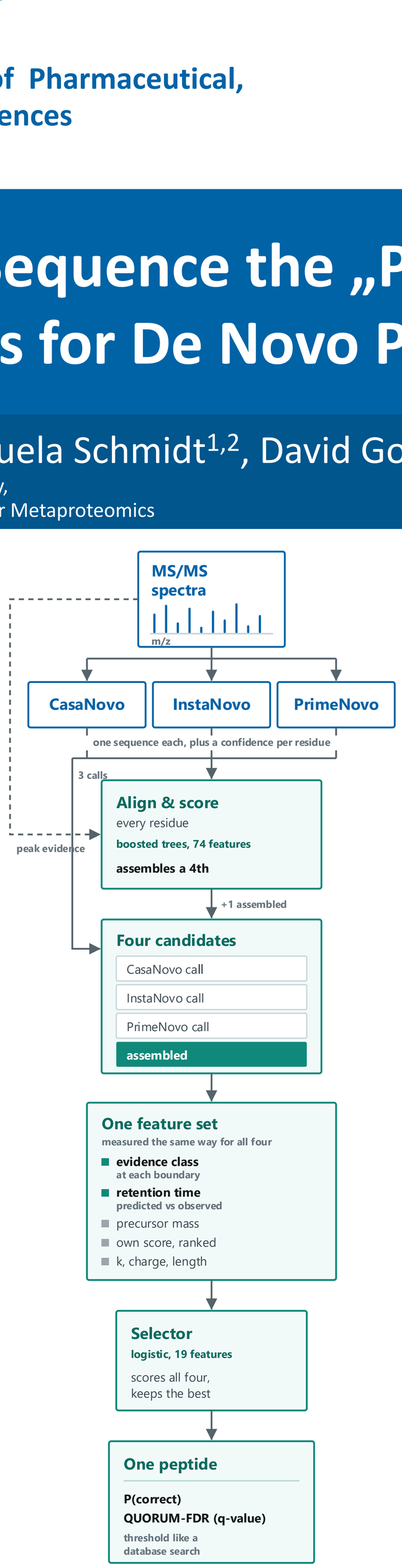


Fig. 2: Claimed against delivered false discovery rate. Within a dataset a nominal 5% delivers 5.1%. Transferred truth-free to a dataset the selector had never seen (pig plasma and gut metaproteomics) it delivers 3.7% - 4.9%, conservative at every cut, because nothing is refitted there except the base rate, by an estimator known to under-shoot.



More peptides at a controlled error rate

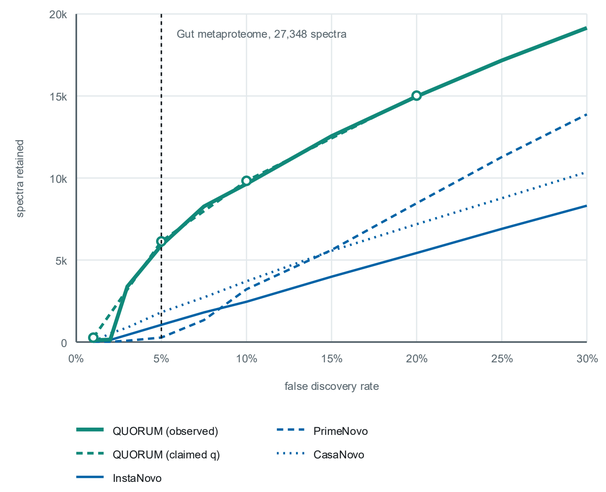


Fig. 3: Spectra retained against delivered false discovery rate. At a 5% delivered error rate QUORUM returns 5,893 gut spectra where the best model returns 1,814. The model curves use a threshold derived from ground truth, which is unavailable in practice without a database, so they flatter the models rather than quorum.

Benchmark FDR against Winnow

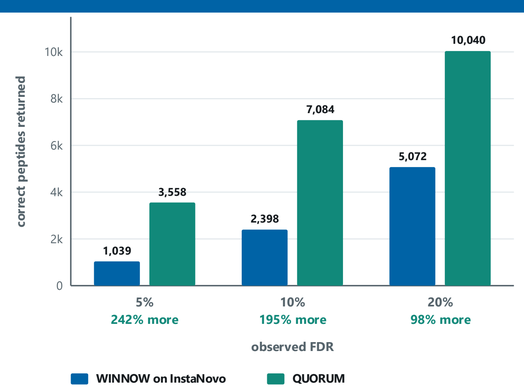


Fig. 4: Correct peptides at a matched error rate. Gut sample. Both tools run without a database, then graded on the 18,756 spectra both scored. Winnow[4], another FDR-rescorer, rescores InstaNovo, the only one of our three models whose output it can, and both score the same peptides. Both models deliver a conservative FDR, but QUORUM delivers over 240% more peptides.

Confirmation of QUORUM predicted peptides not found by FragPipe database search

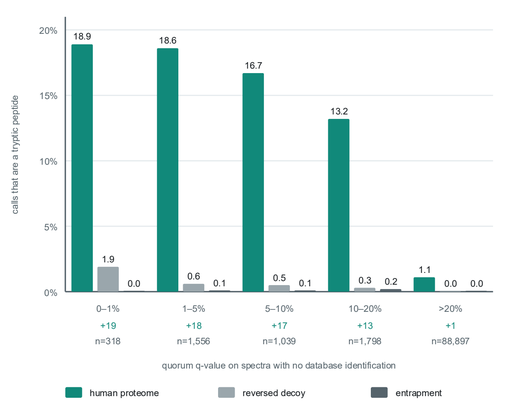


Fig. 5: Tryptic entrapment validation of calls on spectra no database search identified. HeLa, 38,039 spectra the database search never identified. Each call is looked up as a tryptic peptide in the human database and in two nulls: the reversed decoy half, and Arabidopsis plus yeast with every human-shared peptide removed. All three matched at 2,369,611 peptides.

Conclusion

QUORUM follows the approach PostNovo[5] introduced in 2018, combining several de novo models under FDR control. What differs is the evidence it runs on: the ion evidence class at each boundary and retention time, neither of which PostNovo used, and a full-length candidate assembled residue by residue rather than a consensus subsequence.

With no database and no PSM table at any stage, the pipeline reaches 50.7% exact matches on unseen gut metaproteomics, and 52.6% on a pig plasma sample from a species absent from training. Validation stands at two sample types. A synthetic community of known composition is the next test and would make the currently unmeasurable fraction gradable.

University of Vienna
Vienna Doctoral School of Pharmaceutical, Nutritional and Sport Sciences
Department of Pharmaceutical Sciences

Sebastian Johannes Eder:
sebastian.johannes.eder@univie.ac.at

References:

1. **Yilmaz M et al. Casanovo. Nat Commun 2024 (orig. ICML 2022)**
2. **Eloff K et al. InstaNovo. Nat Mach Intell 2025**
3. **Zhang et al. N-PrimeNovo. Nat Commun 2025**
4. **Mabona A et al. Winnow, arXiv 2025**
5. **Miller SE et al. PostNovo. J Proteome Res 2018**