PepTrust: A Probabilistic Framework for High-Confidence Peptide Identification

Zhao Juntao [1], Bian Yuqian [1], Hao Jingyu [1], Yu Weichuan [1]

[1]Department of Electronic and Computer Engineering, Hong Kong University of Science and Technology

Abstract

Mass spectrometry-based peptide identification is fundamental to proteomics, yet evaluating identification accuracy without ground truth remains a critical challenge. We present **PepTrust**, a probabilistic framework that estimates confidence scores for peptide identifications by modeling search algorithms as information sources in a truth discovery framework.

Innovation: PepTrust evaluates multiple combination methods (MSblender, Voting, Geometric Mean) by measuring their ability to explain observed data distributions, identifying the optimal method without requiring ground truth.

Results: On PRIDE dataset PDX004732, PepTrust achieved **70.29% precision**, surpassing individual search methods (avg. 64.91%). MSblender emerged as the optimal combination method with significantly higher log-likelihood scores across all datasets.

Key Words: Mass Spectrometry, Peptide identification, ground-truth-absent evaluation, probabilistic framework, truth discovery

Method

1. Problem Formulation

Given MS/MS spectra $\mathcal{X} = \{x_1, x_2, \ldots, x_N\}$ and peptide database $\mathcal{P}$, each search algorithm S_i produces:

$$S_i : \mathcal{X} \to \mathcal{P} \times \mathbb{R}^+$$

mapping spectrum x_j to peptide p_{ij}.

2. Bayesian Source Reliability Estimation

The posterior reliability of search algorithm S_i given combination method M_j:

$$\begin{aligned} \tau_{S_i|M_j} &= \frac{P\left(M_j \mid \tau_{S_i}\right) P\left(\tau_{S_i}\right)}{P\left(M_j\right)} \\ &= \frac{\prod_{k=1}^{N} P\left(p_{M_j,k} \mid p_{S_i,k}, \tau_{S_i}\right) P\left(\tau_{S_i}\right)}{\sum_{\tau'} \prod_{k=1}^{N} P\left(p_{M_j,k} \mid p_{S_i,k}, \tau'\right) P\left(\tau'\right)} \end{aligned} \tag{1}$$

3. Information-Theoretic Method Selection

The optimal combination method minimizes KL divergence:

$$M^* = \arg\max_{M_j} \sum_{S_i \in \mathcal{S}} \sum_{k=1}^{N} P(\phi_{S_i,k}) \log P(\phi_{S_i,k} \mid M_j) \tag{2}$$

where $\phi_{S_i,k}$ represents the observation of source S_i on spectrum k.

4. Truth Discovery Framework

Algorithm 1 PepTrust Algorithm

```
for each combination method M_j do
    for each search engine S_i do
        Calculate trustworthiness τ_{S_i|M_j}
    end for
    Compute log-likelihood L(M_j)
end for
return arg max_{M_j} L(M_j)
```

Implementation

Search Engines

Comet, *X!Tandem*, *MS-GF+*

Combination Methods

- *MSblender:* Weighted probabilistic model
- *Voting:* Majority consensus
- *Geometric Mean:* $p_{combined} = \sqrt[n]{\prod_{i=1}^{n} p_i}$

Results on Synthetic Data

- Five sub-pools from PRIDE PDX004732.
- Each 10K spectra.
- Ground truth available for validation (not used in method).

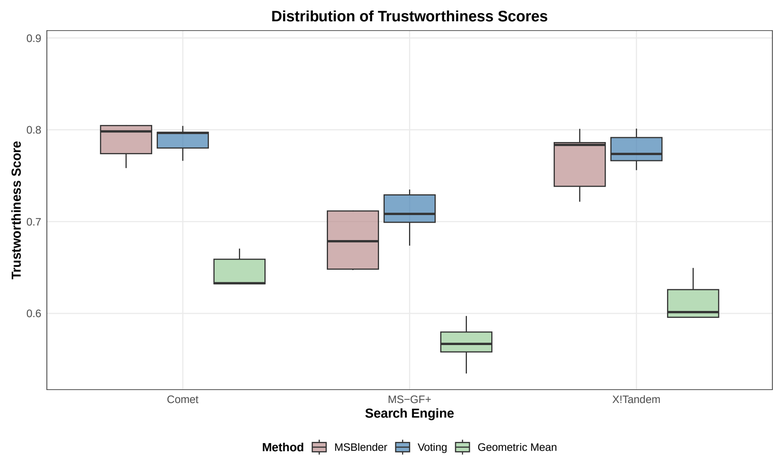

Figure 1. Search engine trustworthiness scores.

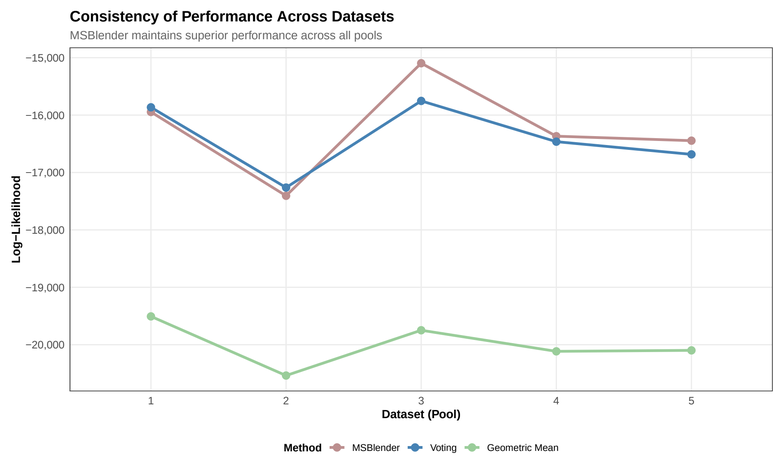

Figure 2. Log-likelihood across 5 dataset pools. MSblender maintains superior performance.

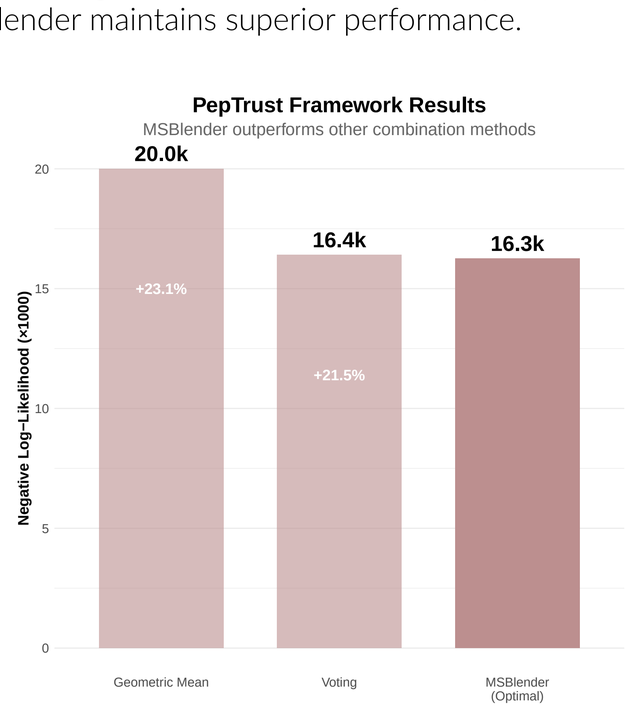

Figure 3. MSblender achieves significantly better performance (lower is better).

Validations on Synthetic Data

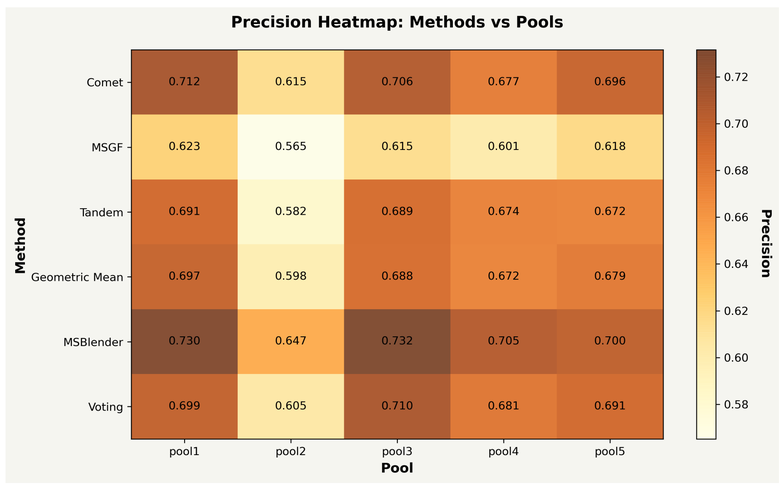

Figure 4. Precision heatmap of the 5 sub-pools.

Method	Log-Likelihood	Precision
MSblender	-16252.1	**70.29%**
Voting	-16405.4	67.73%
Geometric Mean	-20001.3	66.68%
Individual Avg.	-	64.91%

Table 1. Performance comparison on the 5 sub-pools.

Discussion & Future Work

Discussion

- **No ground truth dependency:** Enables evaluation on any proteomics dataset.
- **Principled framework:** Based on established truth discovery theory.
- **Extensible:** Can incorporate new search engines and combination methods.

Future Work

- Extension to post-translational modification (PTM) identification.
- Development of confidence intervals for peptide identifications.

References

1. Fang et al. (2020) From Appearance to Essence: Comparing Truth Discovery Methods without Using Ground Truth. *ACM TIST* 11(6): 1-24.
2. Zolg et al. (2017) Building ProteomeTools based on a complete synthetic human proteome. *Nature Methods* 14: 259-262.
3. T. Kwon*, H. Choi*, C. Vogel, A.I. Nesvizhskii, and E.M. Marcotte, MSblender: a probabilistic approach for integrating peptide identifications from multiple database search engines. J. Proteome Research, 10(7): 2949–2958 (2011)