# Supporting Information

to

# AP3: An Advanced Proteotypic Peptide Predictor for Targeted Proteomics by Incorporating Peptide Digestibility

Zhiqiang Gao[1,3‖], Cheng Chang[2‖], Jinghan Yang[1,3], Yunping Zhu[2,4*], Yan Fu[1,3*]

1. National Center for Mathematics and Interdisciplinary Sciences, Key Laboratory of Random Complex Structures and Data Science, Academy of Mathematics and Systems Science, Chinese Academy of Sciences, Beijing 100190, China.

2. State Key Laboratory of Proteomics, Beijing Proteome Research Center, National Center for Protein Sciences (Beijing), Beijing Institute of Lifeomics, Beijing 102206, China.

3. School of Mathematical Sciences, University of Chinese Academy of Sciences, Beijing 100049, China.

4. Anhui Medical University, Hefei 230032, China.

‖Contributed equally to this work

*Corresponding Authors

Email: yfu@amss.ac.cn

Email: zhuyunping@gmail.com

# Supplementary Notes

## 1. Feature selection

Among the 588 features used for peptide detectability prediction, some may be redundant or highly correlated with others. For example, there is a strong correlation between the peptide length and the molecular weight. Therefore, we performed feature selection using the mRMR method. mRMR can sort features by simultaneously considering their relevancies to the dependent variable and their redundancies with higher-ranked features. The top 50 features were added into the random forest classifier one by one in order. Ten-fold cross-validation was adopted to evaluate the performance of the trained model as the number of added features increased. The minimum set of features with sufficient predictive performance was selected. Finally, 29 features were ultimately selected, as the further added features just had little effect on the 10-fold AUC of the trained model (Figure S1). As shown in Table S2, we grouped the 29 selected features into six categories: digestion, hydrophobicity, structure, charge, energy and others. The peptide digestibility was the top one feature selected, and the number of missed cleavage sites was the third. Both features are related with protein proteolytic digestion, which is the first step of LC-MS/MS based shotgun proteomics. There is broad agreement that the hydrophobicity, structure, charge and energy have large impacts on peptide detectability[1–4]. In reversed-phase high-performance liquid chromatography (RP-HPLC), hydrophobic residues (2-th and 11-th features) make peptides hard to dissolve, while hydrophilic residues (16-th feature) make peptides hard to retain. Eleven of the 29 selected features were related to secondary or tertiary structures, suggesting that peptide structure also influenced peptide detectability. Some peptides

might never be generated because they exist in a region of the protein's structure that is very stable and thus resistant to proteolytic digestion by trypsin. There is no surprising that charge has a significant influence on the peptide detectability, especially the ionization efficiency. The selected feature "Activation Gibbs energy of unfolding" is consistent with a previous study[5], which claimed that Gibbs free-energy transfer between amino acids leads to an increased response in peptides with nonpolar regions. To validate the effectiveness of feature selection, we compared the prediction performances between the full model using all 588 features and the simplified model using the 29 selected features, respectively. The results indicated that the simplified model showed similar accuracy on the training set and even better generalization ability on the test sets in comparison with the full model (Figure S2). We also analyzed with examples why one feature was selected while another one was not. For example, the Pearson Correlation Coefficient (PCC) of the features "number of nonpolar hydrophobic residues" and "number of neutral residues" is 0.8922, which shows that they are correlated features. The Kullback-Leibler (KL) distances between the distributions of these two features on the observed peptides and the unobserved peptides are 0.584 ("number of nonpolar hydrophobic residues") and 0.529 ("number of neutral residues"), respectively, which suggested that the feature "number of nonpolar hydrophobic residues" can better distinguish the positive and negative peptides. The PCC of the features "peptide length" and "peptide molecular weight" is 0.9923, which shows that they are correlated features. The KL distances between the distributions of these two features on the observed peptides and the unobserved peptides are 0.581 ("peptide length") and 0.529 ("peptide molecular weight"). Both examples above are consistent with our feature selection result (i.e., "number of nonpolar hydrophobic residues" and "peptide length" are kept).

## 2. Relationship between peptide digestibility and the number of missed cleavage sites

Integrating peptide digestibility into the feature set of the peptide detectability model was inspired by the close relationship between protein proteolytic digestion and peptide LC-MS/MS detection. However, the number of missed cleavage sites in the peptide sequence is also a feature related to protein digestion. To demonstrate that peptide digestibility is a better representative of protein digestion than the number of missed cleavage sites in terms of peptide detectability prediction, we performed a comparative analysis of these two features. First, we analyzed the correlation between them. The peptide digestibility distributions of peptides with different numbers of missed cleavage sites showed that peptides having more missed cleavage sites tended to have smaller peptide digestibilities (**Figure S7A**). Second, each of the two features was combined with the remaining 586 features to form two sets of 587 features. Training the model on the two feature sets resulted in 10-fold cross-validation AUCs of 0.9183 (the set including peptide digestibility) and 0.8826 (the set including the number of missed cleavage sites), respectively. We then generated three models using one or both of the two features. As shown in **Figure S7B**, the 10-fold cross-validation AUCs of the models using both features, the peptide digestibility only and the number of missed cleavage sites only were 0.8427, 0.8266 and 0.8066, respectively. Third, the KL distance was calculated to measure the distinguishing ability of peptide digestibility and the number of missed cleavage sites for positive peptides and negative peptides. The results showed that the KL distance of peptide digestibility (3.06) was much larger than that of the number of missed cleavage sites (1.42). All the results demonstrated that peptide digestibility is a more powerful feature for peptide

detectability prediction than the number of missed cleavage sites although the two features are correlated.

## Supplementary Figures

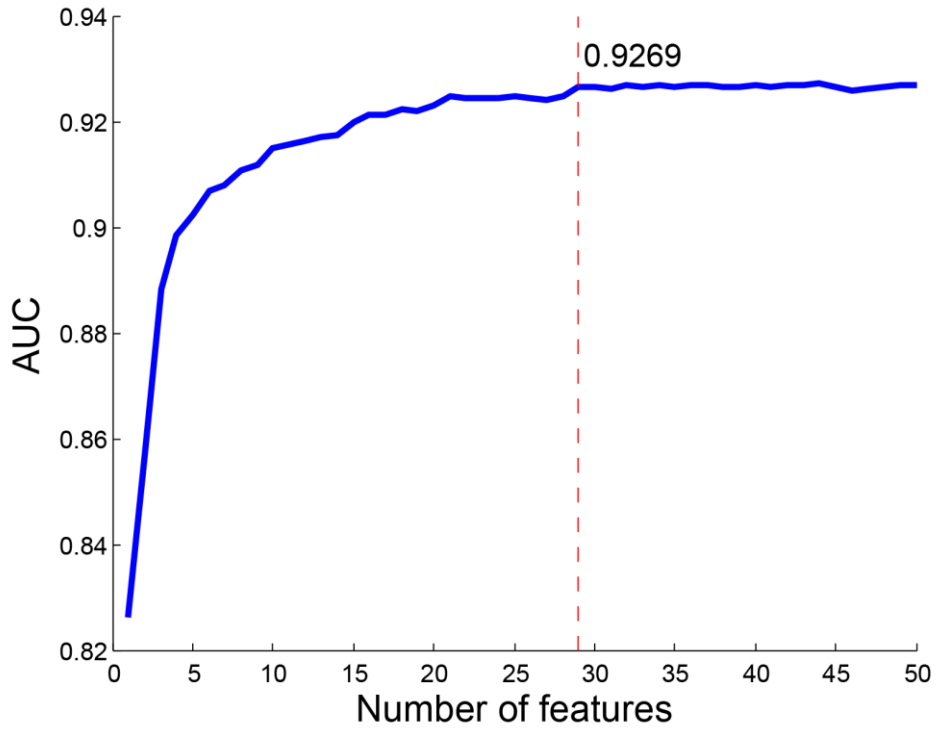


**Figure S1.** The incremental feature selection curve was plotted by 10-fold cross-validation, as the top 50 features selected by mRMR were added successively to the random forest classifier. Ultimately, we selected 29 features with AUC of 0.9269.

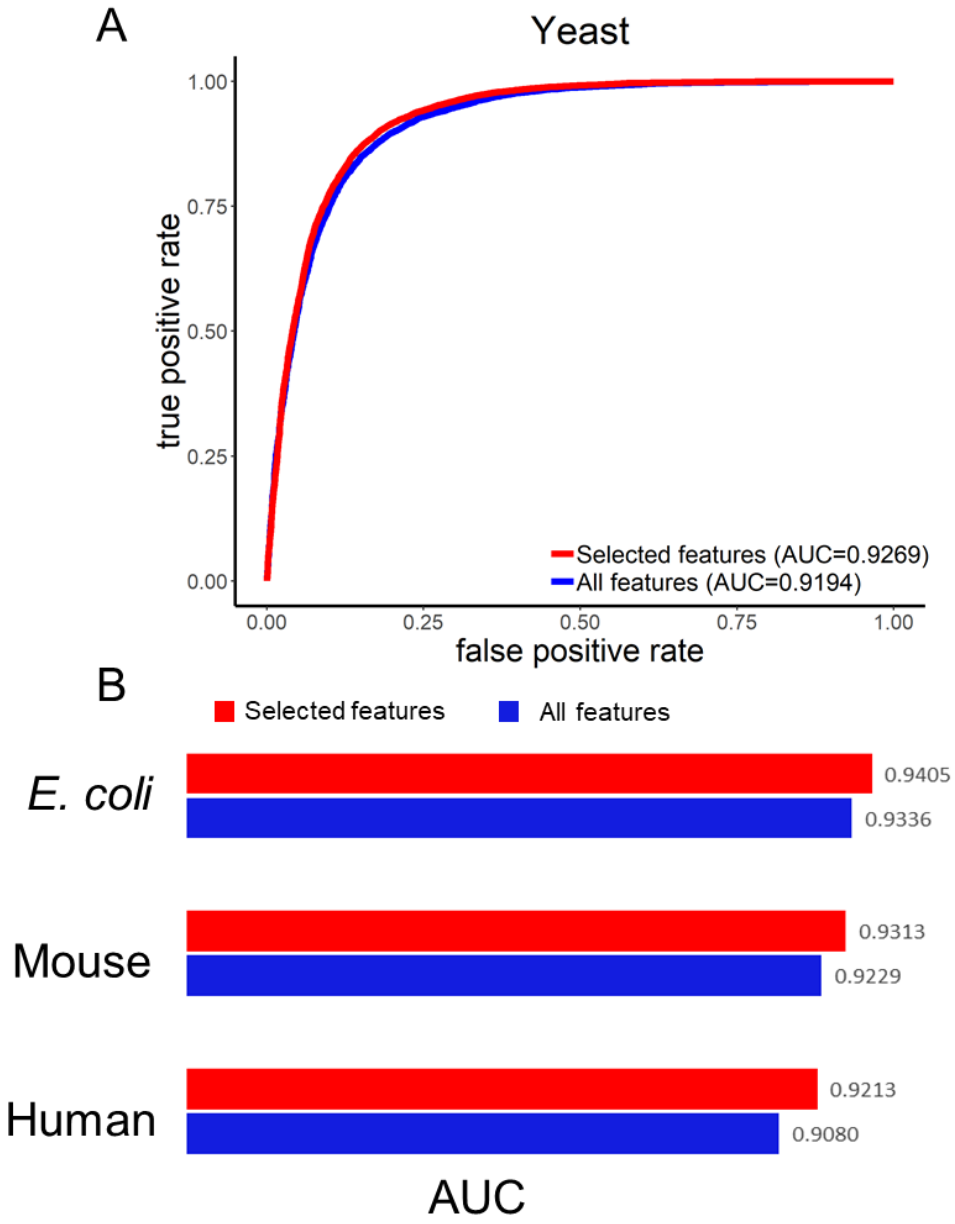


**Figure S2.** Validation of feature selection. (A) Comparison of 10-fold cross-validation ROC curves obtained using the selected features and all features on the training data set. (B) AUCs on three independent test datasets.

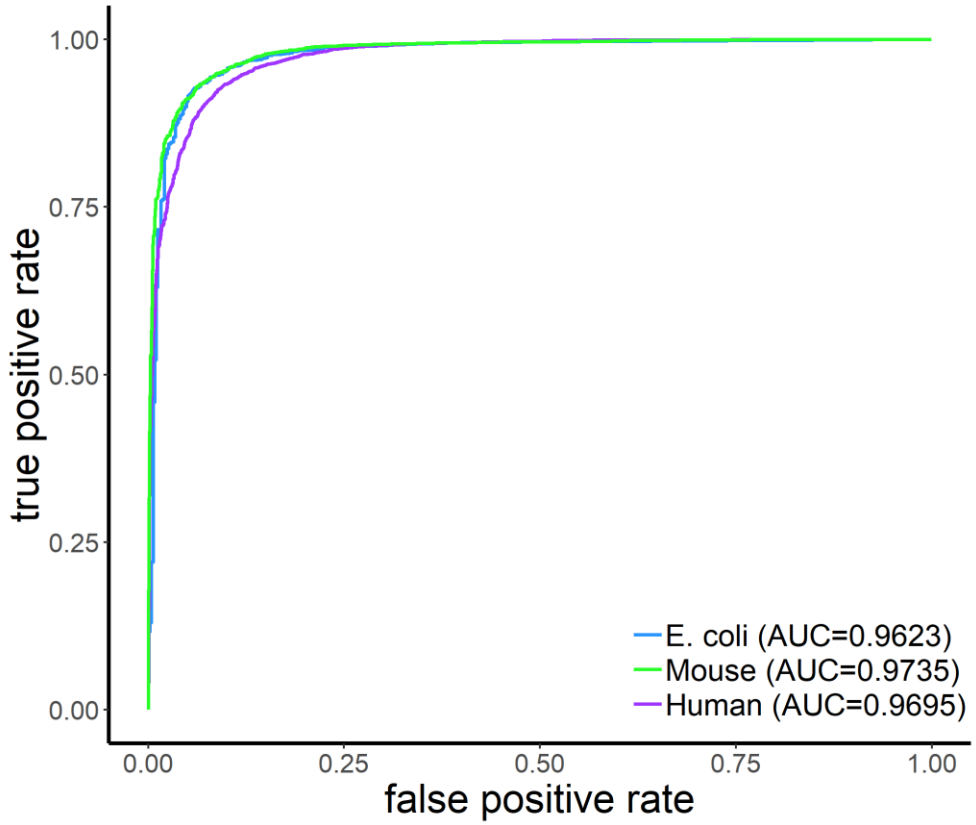


**Figure S3**. ROC curves demonstrating the good performance of cleavage probability model applied to the three test data sets. The cleavage probability model was trained on the Yeast data set.

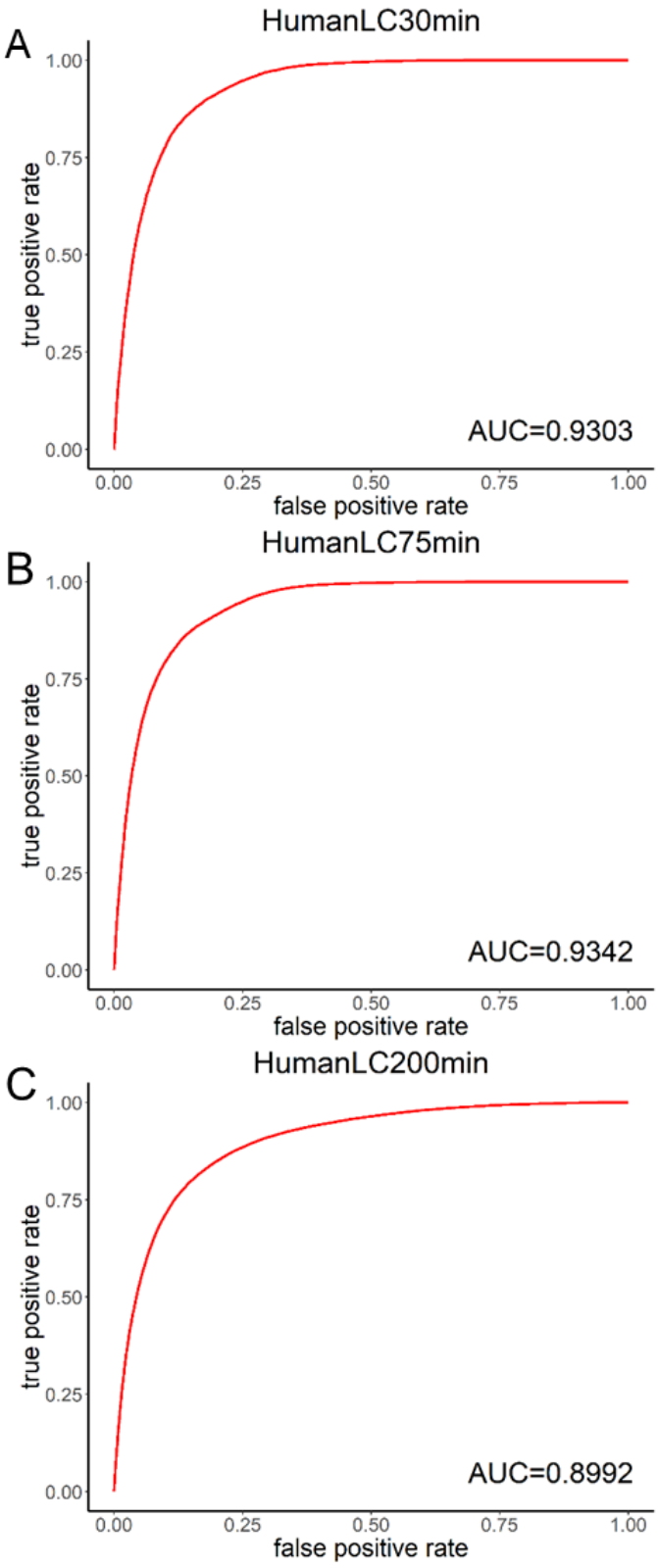


**Figure S4.** The performance of AP3 on three published data sets with different LC columns and elution conditions: (A) 30 min LC elution time, (B) 75 min LC elution time, and (C) 200 min LC elution time.

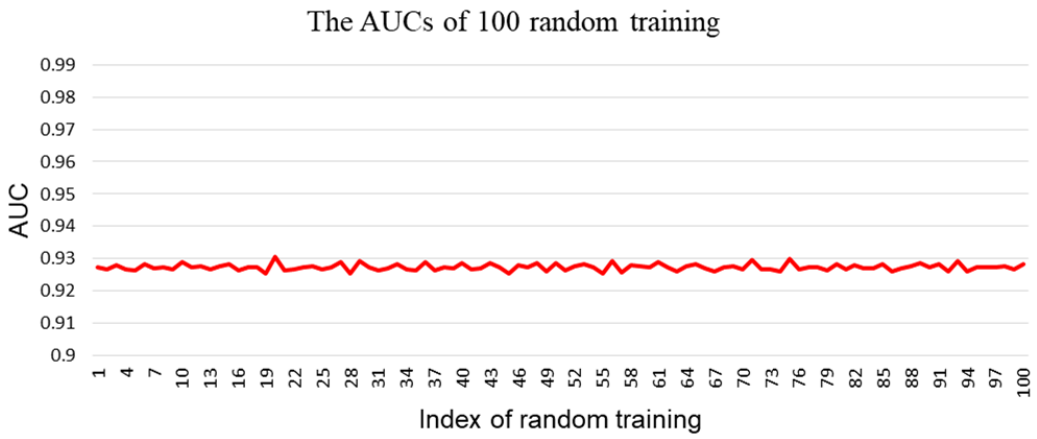


**Figure S5.** Demonstration of the stability of peptide detectability model. The red curve represents the 10-fold cross-validation AUC against the index of random sampling of negative peptides on the Yeast training data set.

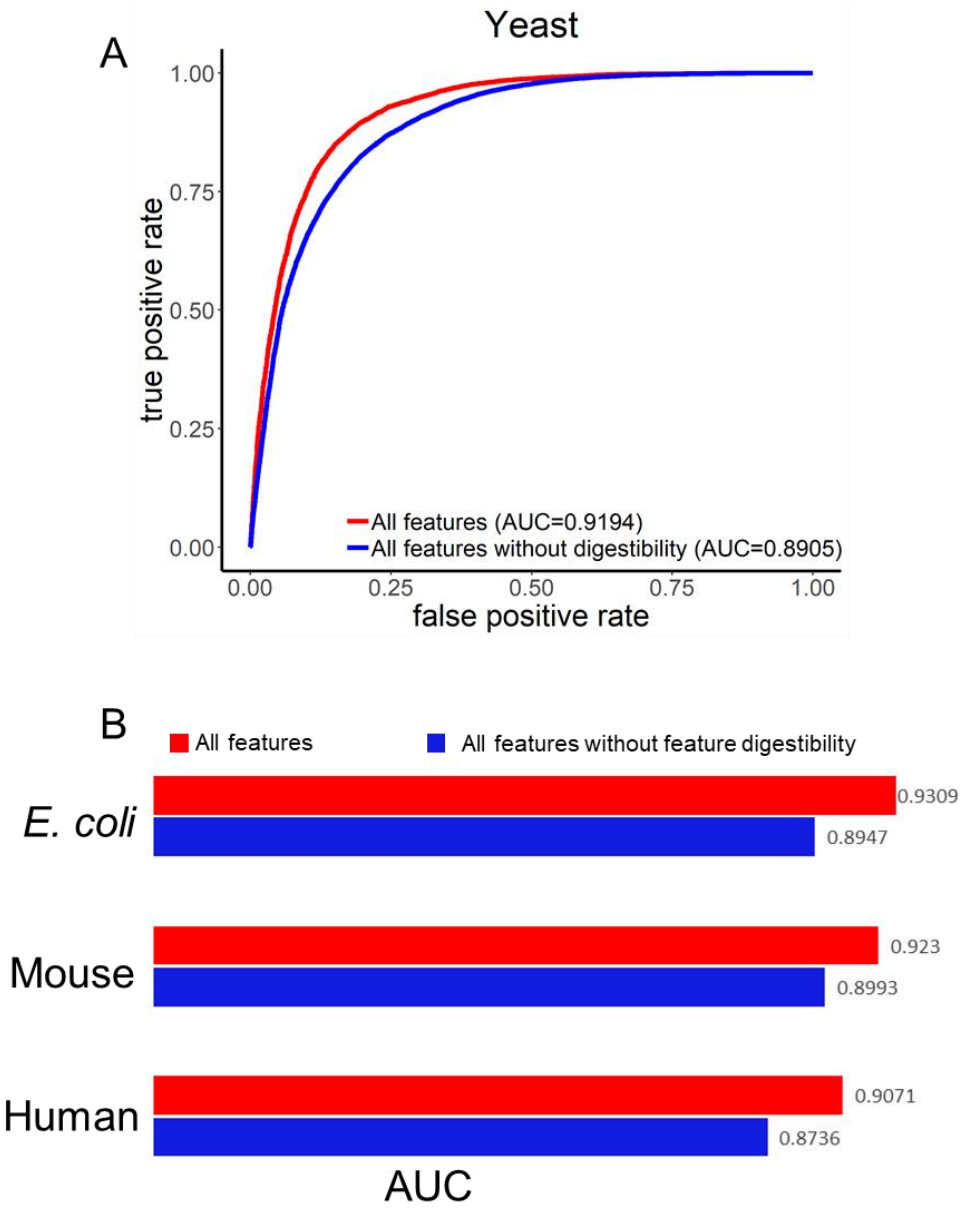


**Figure S6.** Performance comparisons of the AP3 algorithm with/without the peptide digestibility feature included in all the 588 features. (A) The 10-fold cross-validation ROC curves on the Yeast training data set. (B) The AUCs on the three test data sets.

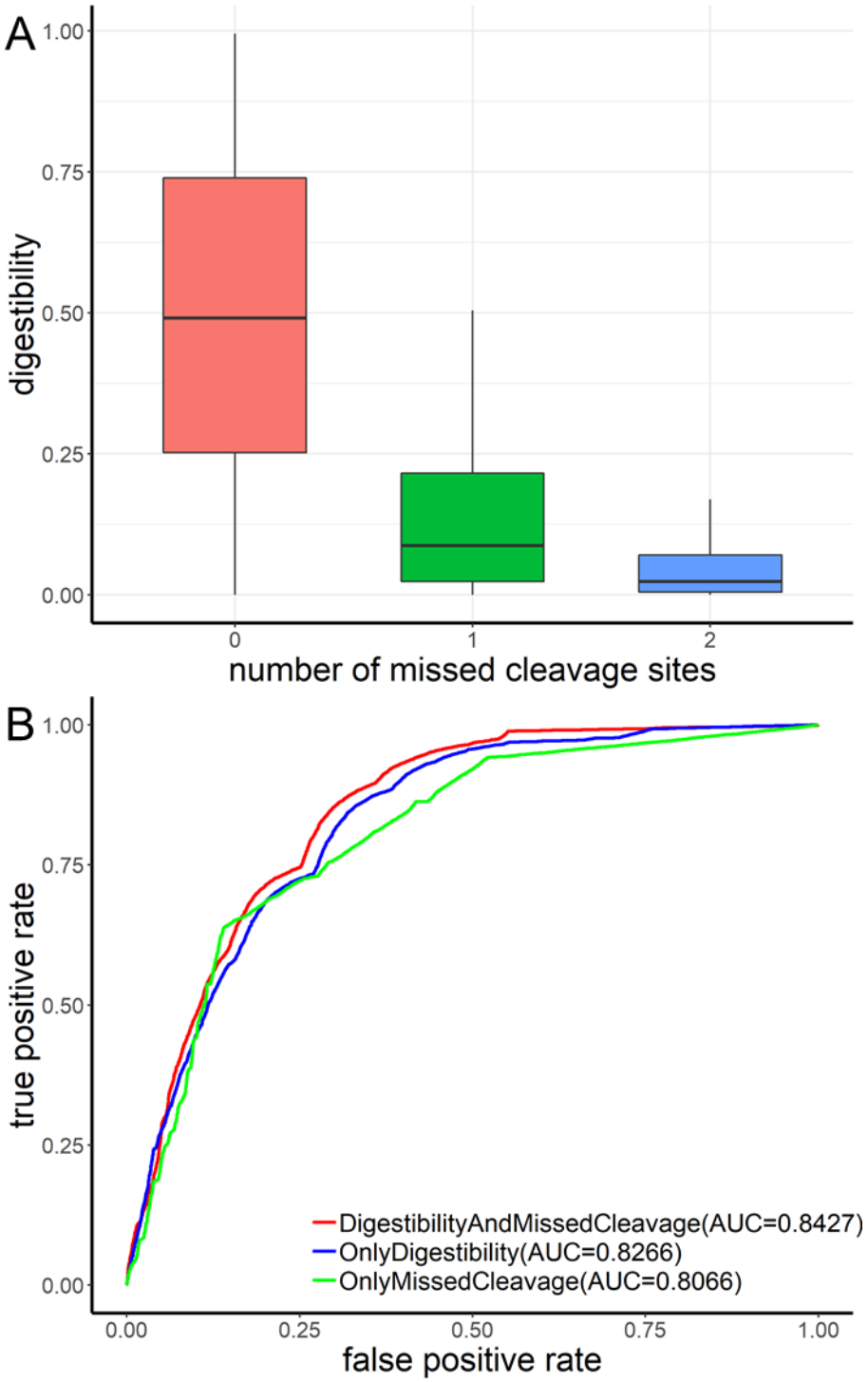


**Figure S7.** Comparative analysis of peptide digestibility and the number of missed cleavage sites. (A) Boxplots of the predicted digestibilities for the peptides with different numbers of missed cleavage sites. (B) The 10-fold cross-validation ROC curves of the models using the number of missed cleavage sites alone (green), the digestibility alone (blue) and both features (red), respectively.

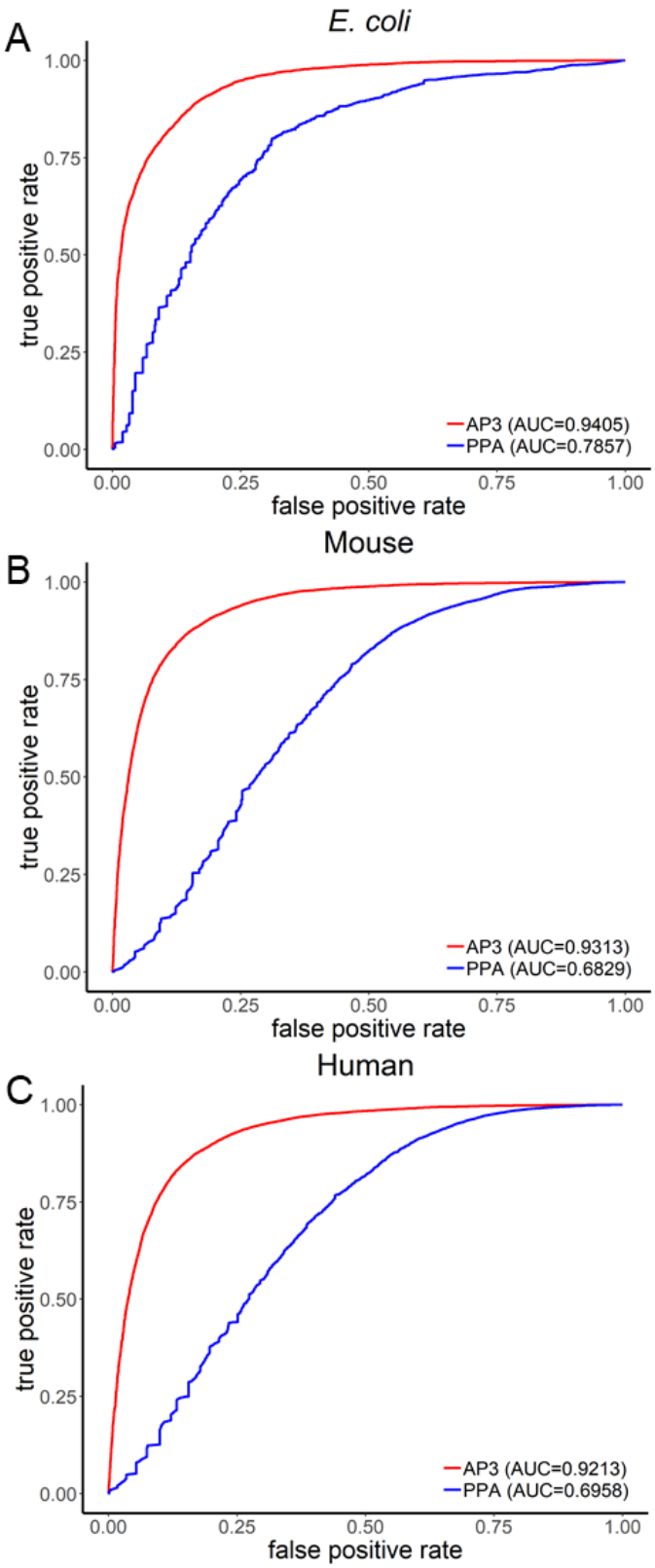


**Figure S8.** Performance comparison between AP3 and PPA on the three independent published data sets, (A) *E. coli*, (B) Mouse, and (C) Human, respectively. Both models were trained on the Yeast data set.

# Supplementary Tables

**Table S1**. The 588 features used to characterize the peptides in AP3.

**Table S2**. The lists of selected features in four data sets by mRMR.

**Table S3**. Summary of the three additive published data sets with different LC columns and elution conditions.

**Table S4**. Validation result on MRM assay data set.

(See excel files for Tables S1, S2 and S4)

**Table S3**. Summary of the three additive published data sets with different LC columns and elution conditions

| Data set | Instrument | LC column | Elution time | Reference |
|---|---|---|---|---|
| HumanLC30min | Q Exactive | $75\mu m \times 10cm$ | 30 min | (Ding, et al.)[6] |
| HumanLC75min | Orbitrap Fusion Lumos | a home-made $150\mu m \times 12cm$ silica microcolumn | 75 min | (Ge, et al.)[7] |
| HumanLC200min | Q Exactive | $75\mu m \times 30cm$ | 200 min | (Tyanova, et al.)[8] |

# References

(1) Fusaro, V. A.; Mani, D. R.; Mesirov, J. P.; Carr, S. A. *Nat. Biotechnol.* **2009**, *27* (2), 190–198.

(2) Mallick, P.; Schirle, M.; Chen, S. S.; Flory, M. R.; Lee, H.; Martin, D.; Ranish, J.; Raught, B.; Schmitt, R.; Werner, T.; et al. *Nat. Biotechnol.* **2007**, *25* (1), 125–131.

(3) Sanders, W. S.; Bridges, S. M.; McCarthy, F. M.; Nanduri, B.; Burgess, S. C. *BMC Bioinf.* **2007**, *8 (Suppl 7)*, S23.

(4) Eyers, C. E.; Lawless, C.; Wedge, D. C.; Lau, K. W.; Gaskell, S. J.; Hubbard, S. J. *Mol. Cell. Proteomics* **2011**, *10* (11), M110.003384-M110.003384.

(5) Cech, N. B.; Enke, C. G. *Anal. Chem.* **2000**, *72* (13), 2717–2723.

(6) Ding, C.; Jiang, J.; Wei, J.; Liu, W.; Zhang, W.; Liu, M.; Fu, T.; Lu, T.; Song, L.; Ying, W.; et al. *Mol. Cell. Proteomics* **2013**, *12* (8), 2370–2380.

(7) Ge, S.; Xia, X.; Ding, C.; Zhen, B.; Zhou, Q.; Feng, J.; Yuan, J.; Chen, R.; Li, Y.; Ge, Z.; et al. *Nat. Commun.* **2018**, *9* (1), 1–16.

(8) Tyanova, S.; Albrechtsen, R.; Kronqvist, P.; Cox, J.; Mann, M.; Geiger, T. *Nat. Commun.* **2016**, *7*, 10259.