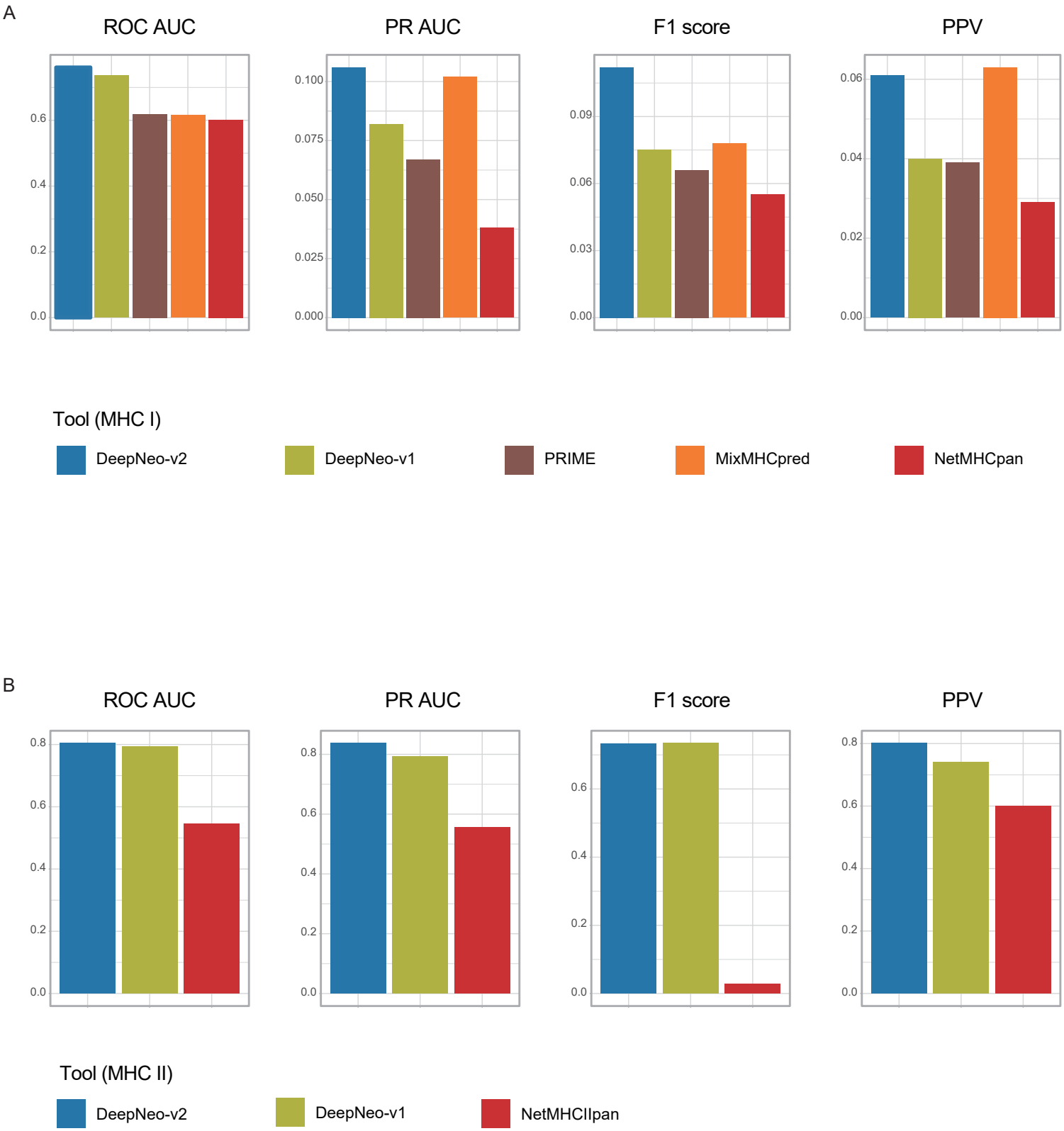
Supplementary Figure 1



Supplementary Figure 1. Accuracy metrics for external validation sets. In MHC I (A) and II (B), external validation sets were applied to examine prediction performances. The prediction power is measured with ROC AUC, PR AUC, F1 score, and PPV. In both classes, DeepNeo-v2 outperforms other tools.

Supplementary Table 1

Input size	DeepNeo-v1	DeepNeo-v2.0
10	4.56	0.47
100	6.16	0.63
1000	13.69	1.72

Supplementary Table 1. Comparison of runtime. After code optimization and platform upgrade, the mean runtime(sec) per input size showed significant decrease in DeepNeo-v2, appropriate to be implemented as webserver.

Supplementary Table 2

Training set		Test set performance			External validation performance		
Positive %	Data	ROC AUC	PR AUC	F1 score	ROC AUC	PR AUC	F1 score
42.6%	**19,149**	**0.805**	**0.713**	**0.687**	**0.761**	**0.106**	**0.112**
25.7%	31,681	0.796	0.548	0.481	0.757	0.081	0.134
30%	27,160	0.801	0.603	0.521	0.723	0.078	0.108
35%	23,280	0.803	0.660	0.599	0.761	0.087	0.122
45%	18,107	0.795	0.722	0.700	0.763	0.092	0.091
50%	16,296	0.796	0.767	0.737	0.758	0.090	0.093

Supplementary Table 2. Comparison of model performance according to label composition of training set. To find optimal label composition, multiple ratios were employed and tested. The label composition with the most balanced accuracy metrics across the three datasets are selected as the optimal data composition.