An automated pipeline to extract the Drosophila modular transcription regulators and targets from massive literature articles

Tzu-Hsien Yang*, Sheng-Hang Wu, Fang-Yuan Zhang, Hsiu-Chun Tsai, Ya-Chiao Yang, Yan-Yuan Tseng, Wei-Sheng Wu*

Supplementary Materials

1 Learning curves for the deep learning models in DMLS Step 1 and Step 3

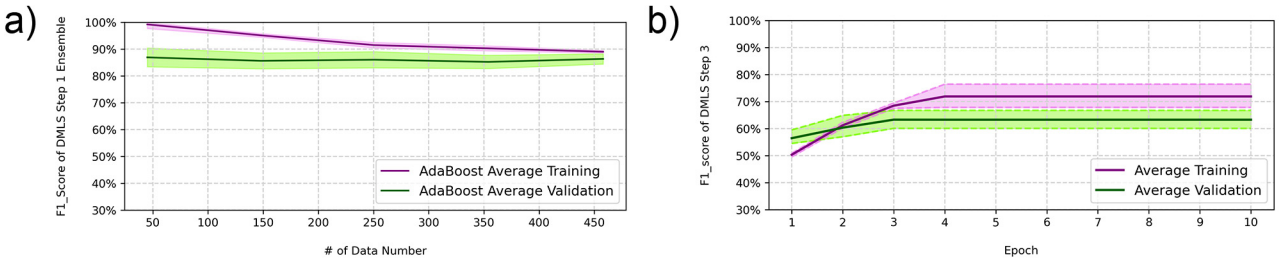

Figure S1: (a) The learning curves of the DMLS Step 1 ensemble model. (b) The learning curves of the DMLS Step 3 deep network.

2 Learning curves for the retrained CNN and LSTM models for identifying CRM literature

We retrained the CNN and LSTM modeling networks using the BioBERT word embeddings. The hyperparameters of the CNN modeling chosen by five-fold cross-validation were as the follows: (1) learning rate schedule: linear warm-up to 1e-4 in 10 epochs followed by exponential decay (decay rate = 0.85); (2) Optimization method: Adam; (3) number of epochs: 50; (4) neuron initialization: Xavier initialization; (5) Batch training size: 64. (6) dropout = 0.4. The hyperparameters of LSTM modeling were the following: (1) learning rate schedule: cosine warm-up to 1e-5 in 5 epochs followed by exponential decay (decay rate = 0.9 before epoch 20; 0.85 after that epoch); (2) Optimization method: Adam; (3) number of epochs: 50; (4) neuron initialization: Xavier initialization; (5) Batch training size: 64. (6) dropout = 0.5. The CNN and LSTM models were trained using NVIDIA RTX 3090 GPUs.

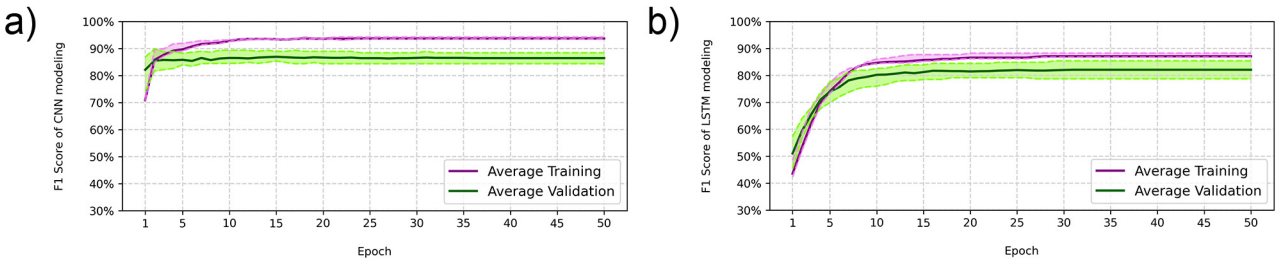

Figure S2: The learning curves for obtaining (a) the CNN modeling and (b) LSTM modeling proposed by Burns *et al.* [**?**].

3 Learning curves for the compared machine learning methods for obtaining the ensemble model in DMLS Step 1

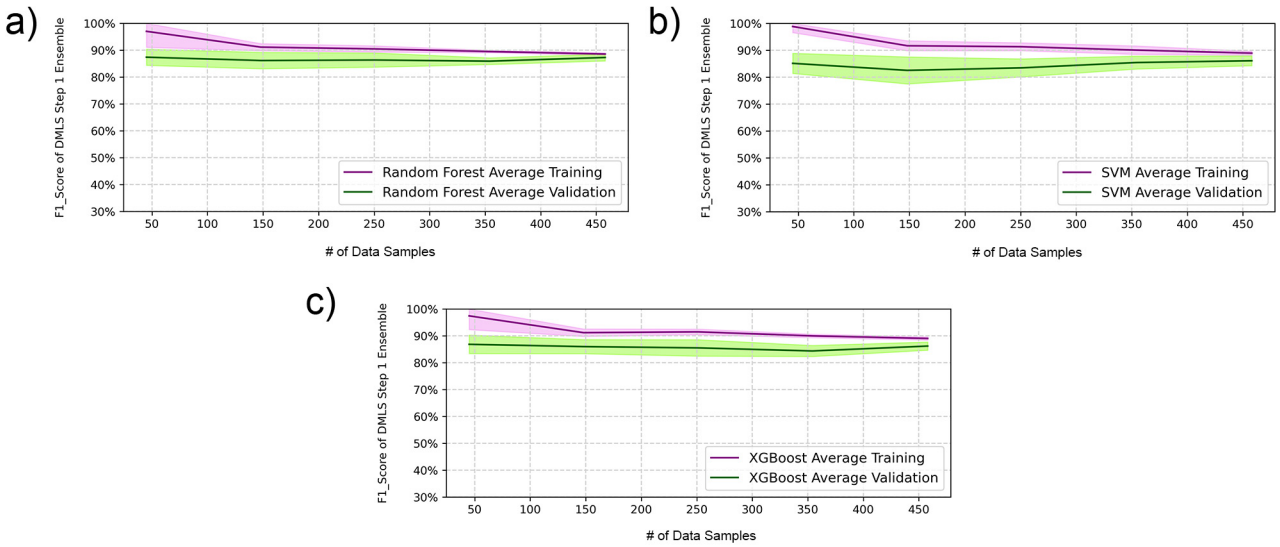

Figure S3: (a) The learning curves of the DMLS Step 1 ensemble model based the random forest algorithm. (b) The learning curves of the DMLS Step 1 ensemble model based the SVM algorithm. (c) The learning curves of the DMLS Step 1 ensemble model based the XGBoost algorithm.