Supplemental Materials for "Class-balanced negative training sets for enhancer-promoter pairs to enhance predictive models"

Osamu Maruyama and Tsukasa Koga

1 Modification of TransEPI code for more equitable evaluation of negative datasets

To ensure a more equitable evaluation of the negative datasets, we utilized a slightly modified version of the TransEPI code. In its original implementation, the loss function combined a standard binary cross-entropy loss with an additional term penalizing the gap between the actual and predicted EP distances. We removed this distance-based term because it hindered the model's ability to learn essential relationships between enhancer and promoter features. This term introduced significant bias when the distance distributions of positive and negative EP pairs differed substantially, artificially affecting the model's predictions. In cases where the distance distributions were similar, the term had negligible impact. By excluding this component, we aimed to eliminate confounding effects and focus on learning the fundamental features driving EP interactions.

2 Supplemental figures

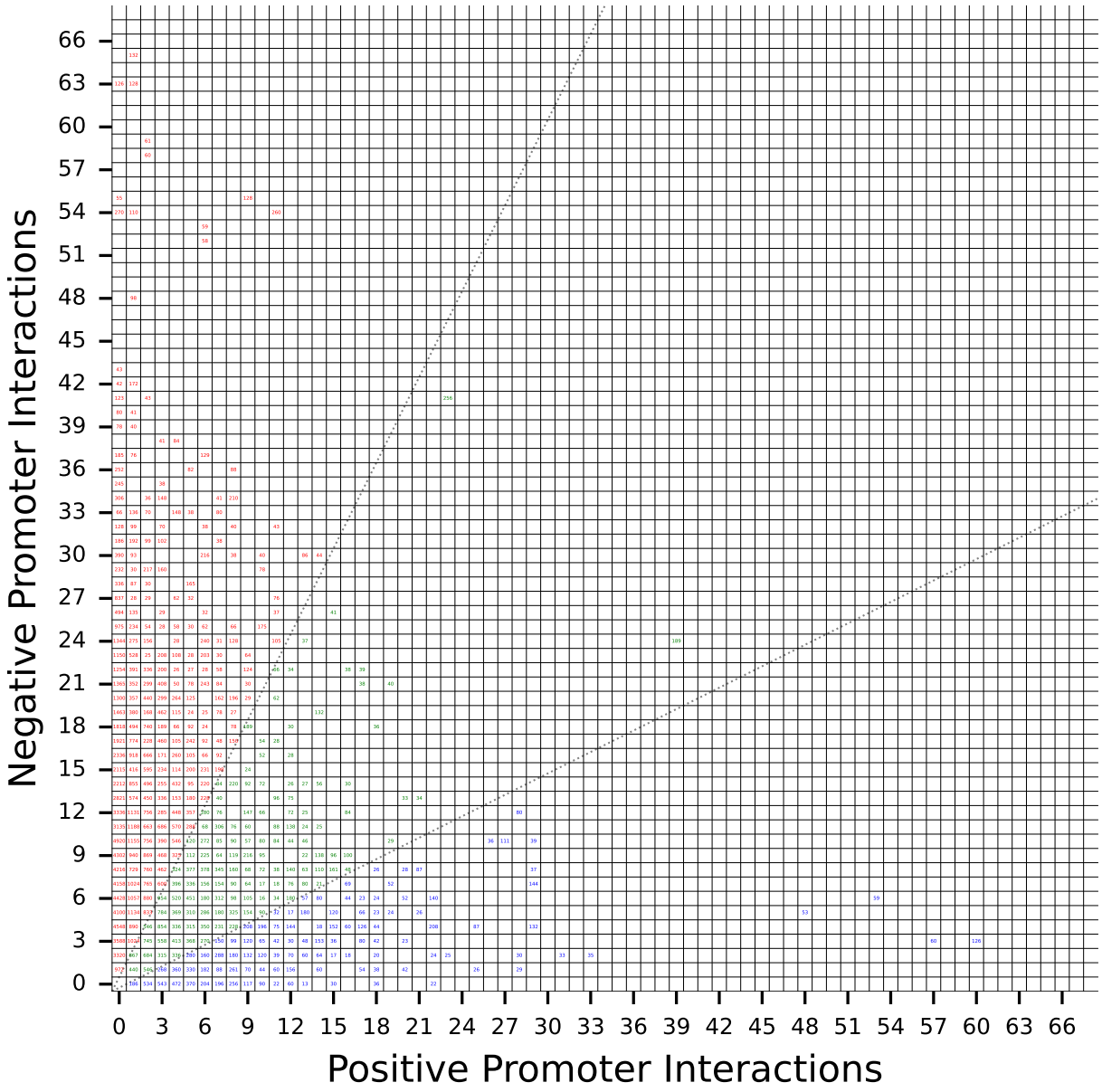


Figure S1: Full size of the promoter-frequency matrix for the BENGI-positive and original BENGI-negative sets in the GM12878 cell line. Number of positive and negative EP pairs sharing a promoter. If there are N different promoters such that each of them occurs in n_1 positive and n_2 negative EP pairs, the corresponding (n_1, n_2)-element shows the total EP pair, $N(n_1 + n_2)$. The positive dataset is the BENGI EP pair of GM12878.

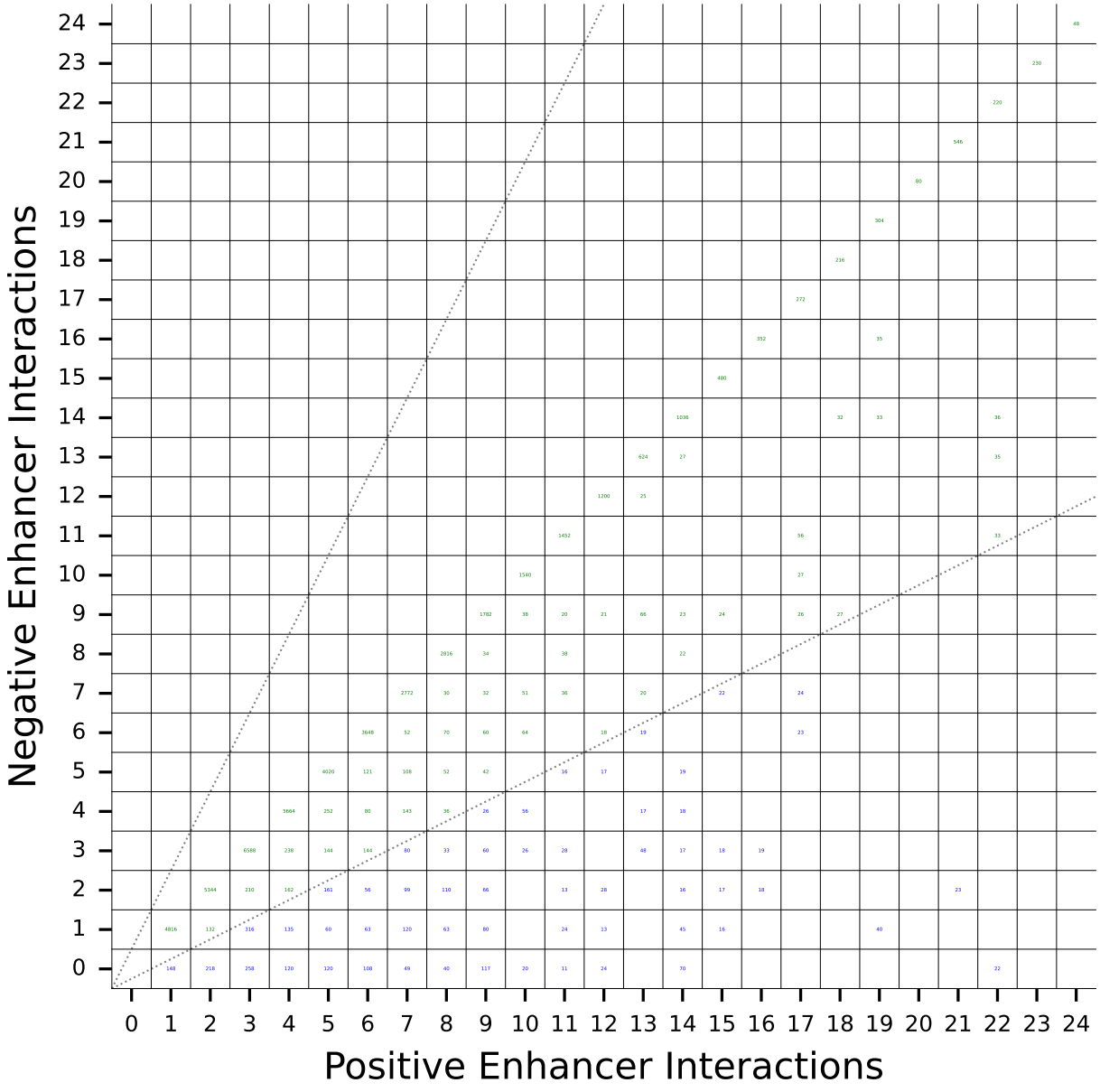


Figure S2: Enhancer-frequency matrix for the pair of the BENGI-positive and CBMF-negative sets in the GM12878 cell line.

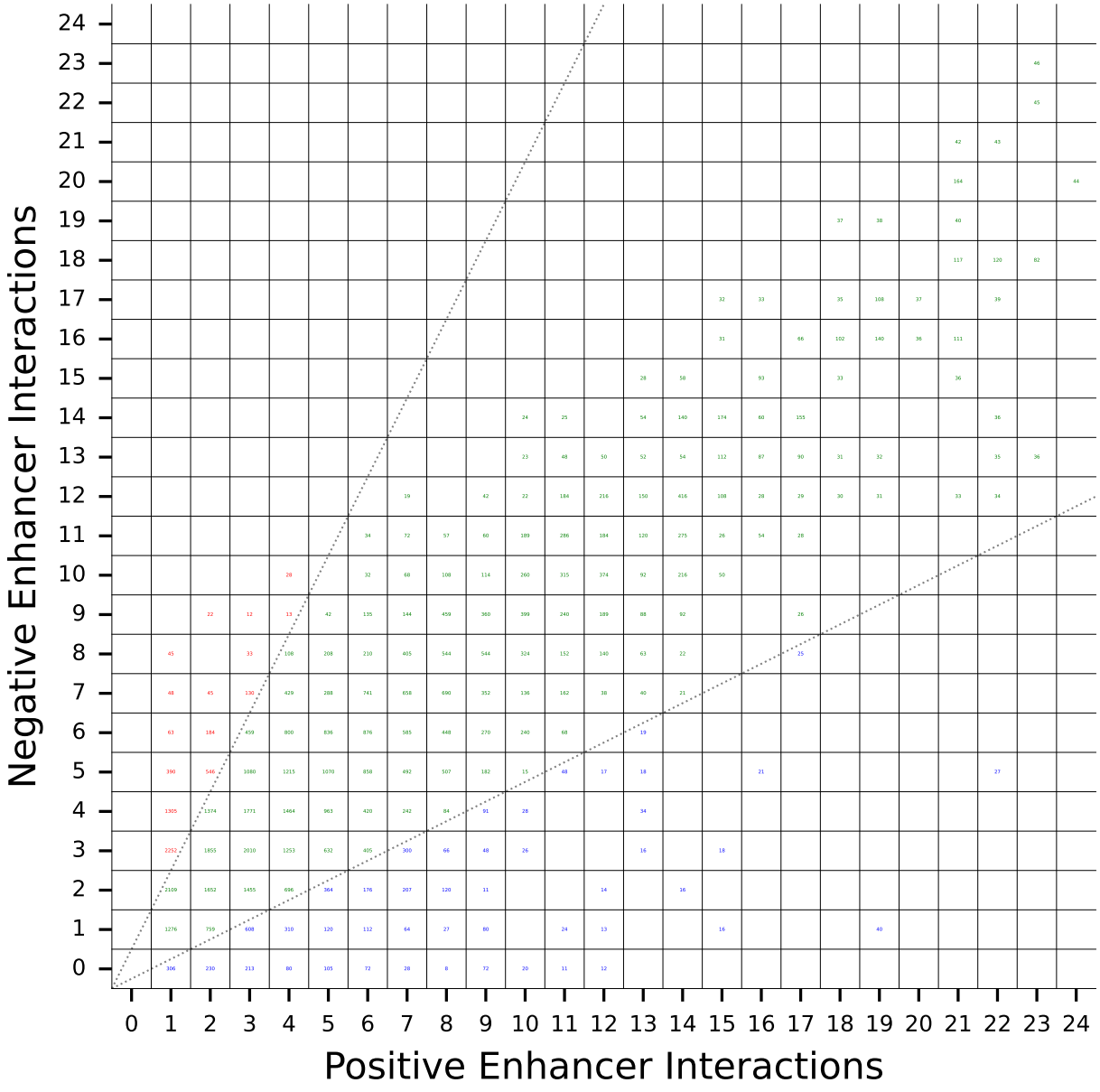
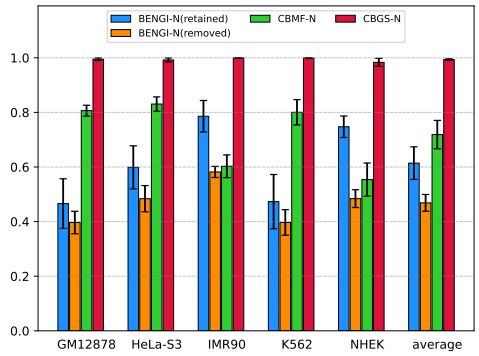
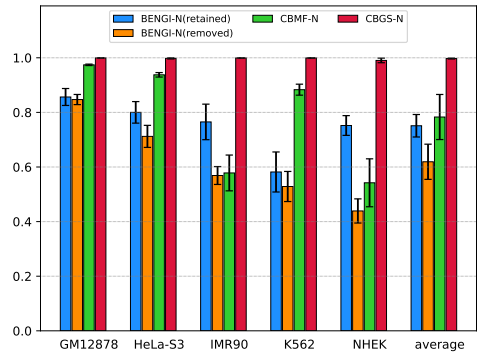


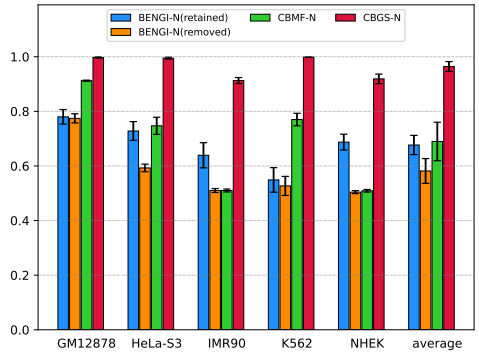
Figure S3: Enhancer-frequency matrix for the pair of the BENGI-positive and CBGS-negative sets in the GM12878 cell line.



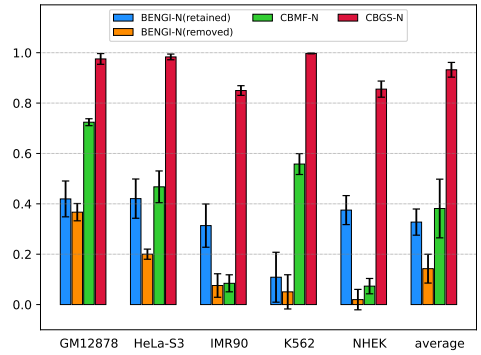
(a) AUPR



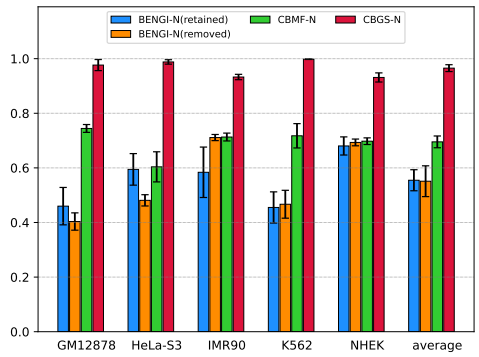
(b) AUC



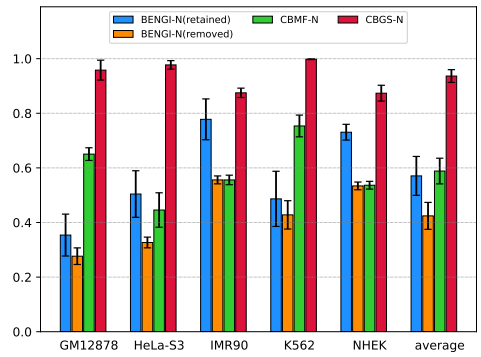
(c) Balanced accuracy



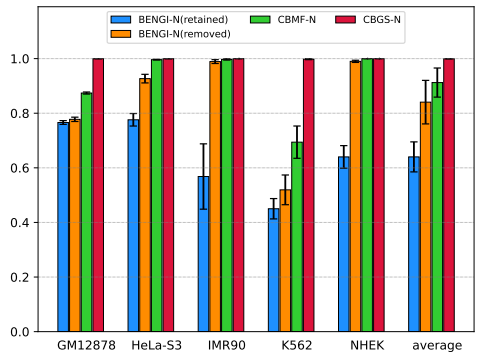
(d) MCC



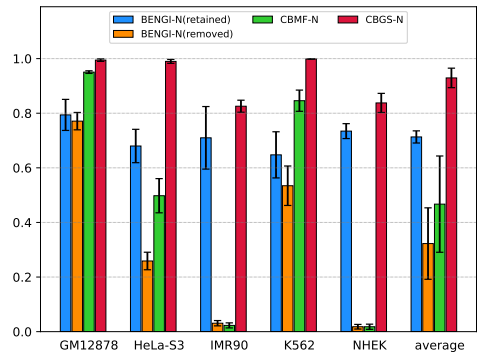
(e) F1



(f) Precision

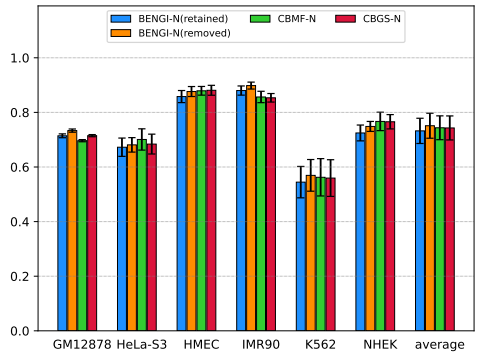


(g) Recall

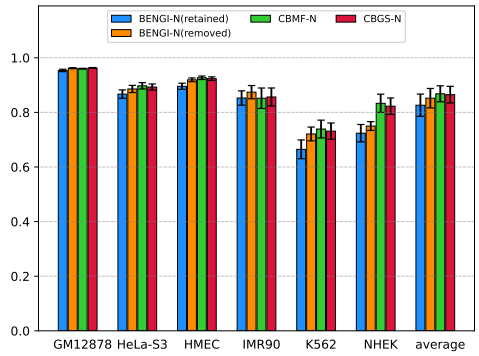


(h) Specificity

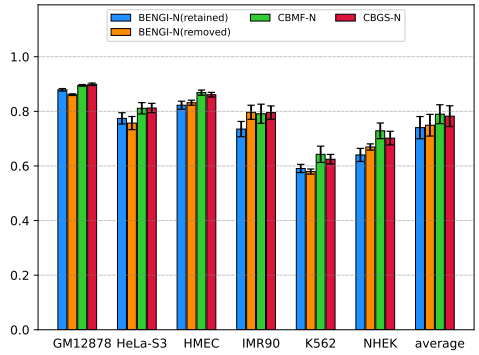
Figure S4: Prediction performance of TargetFinder using the common negative test set with unique elements removed. Panels (a)-(h) show AUPR, AUC, balanced accuracy, MCC, F1, precision, recall, and specificity, respectively, for the negative training sets: the BENGI-negative set with unique elements retained (blue), the BENGI-negative set with unique elements removed (orange), CBMF-negative set (green), and CBGS-negative set (red).



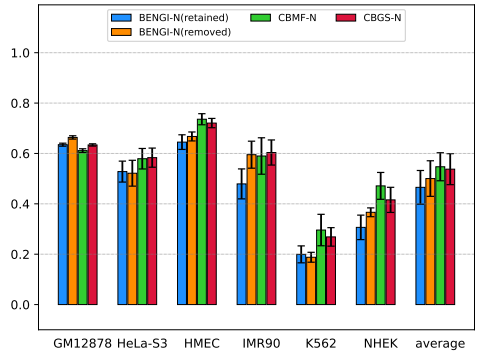
(a) AUPR



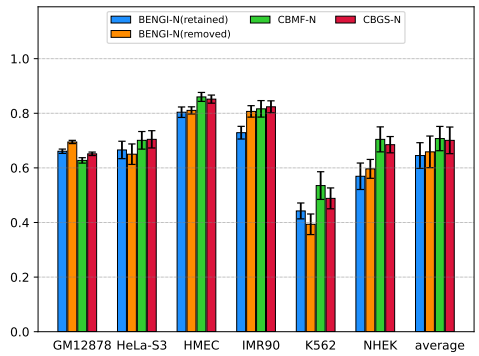
(b) AUC



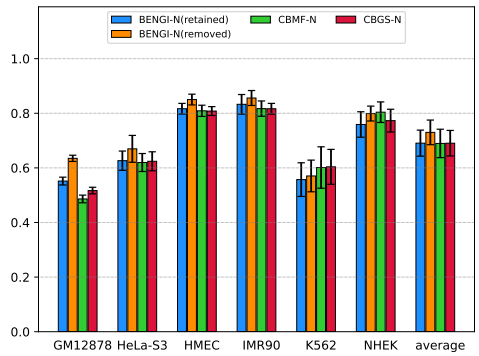
(c) Balanced accuracy



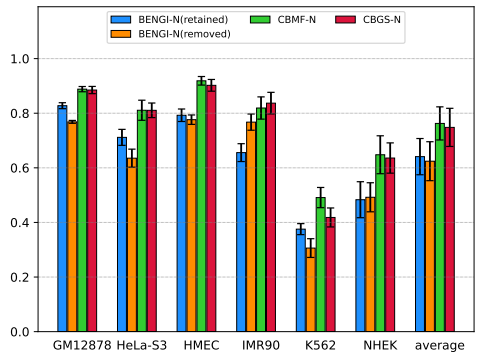
(d) MCC



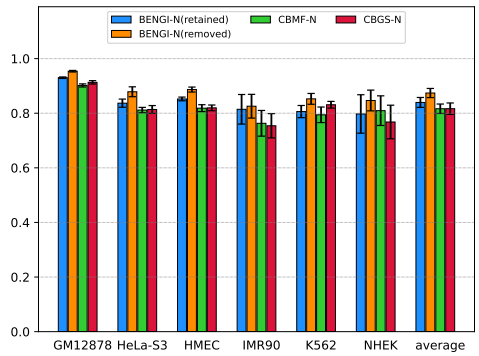
(e) F1



(f) Precision

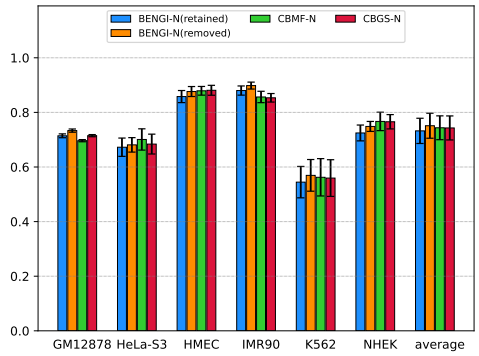


(g) Recall

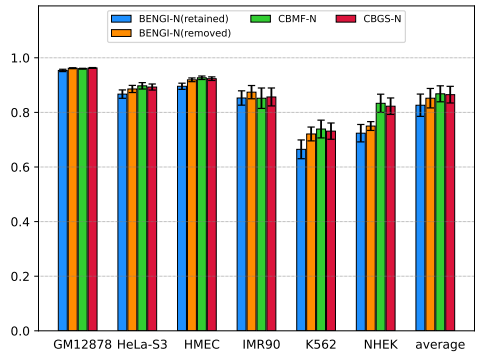


(h) Specificity

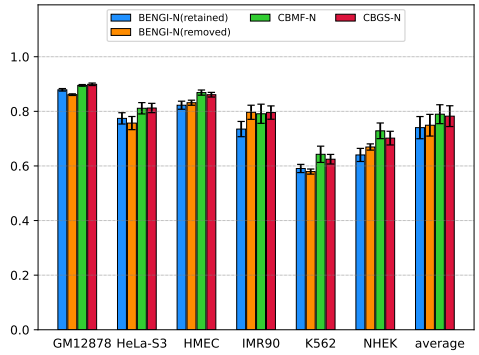
Figure S5: Prediction performance of TransEPI using the common negative test set with unique elements retained. (a)-(h) show AUPR, AUC, balanced accuracy, MCC, F1, precision, recall, and specificity, respectively, for the negative training sets: the BENGI-negative set with unique elements retained (blue), the BENGI-negative set with unique elements removed (orange), CBMF-negative set (green), and CBGS-negative set (red).



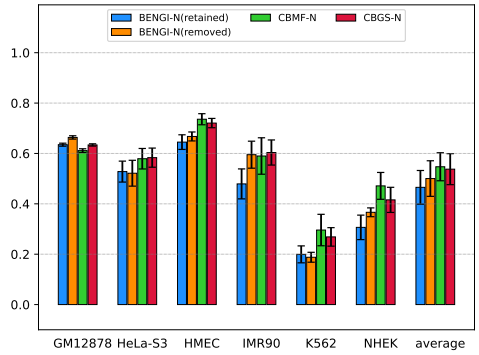
(a) AUPR



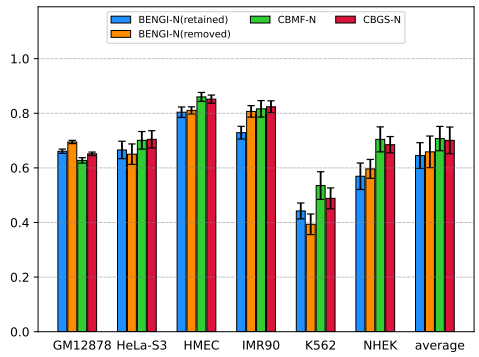
(b) AUC



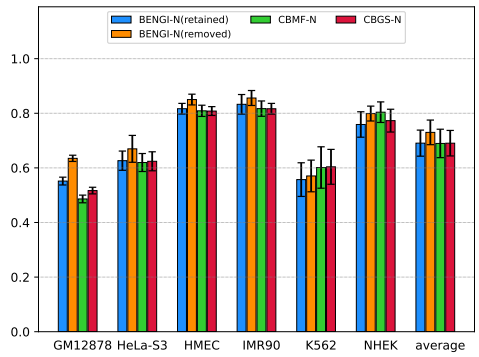
(c) Balanced accuracy



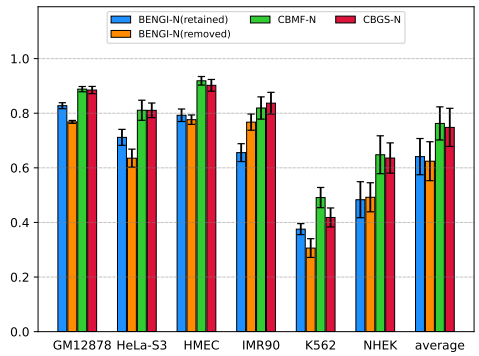
(d) MCC



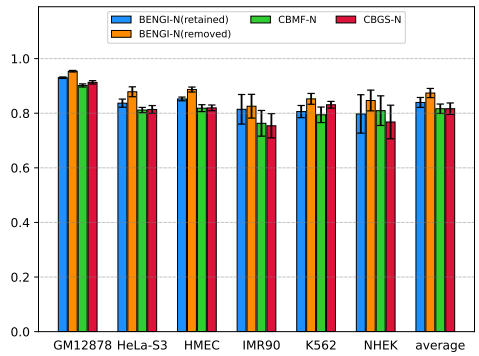
(e) F1



(f) Precision



(g) Recall



(h) Specificity

Figure S6: Prediction performance of TransEPI using the common negative test set with unique elements removed. (a)-(h) show AUPR, AUC, balanced accuracy, MCC, F1, precision, recall, and specificity, respectively, for the negative training sets: the BENGI-negative set with unique elements retained (blue), the BENGI-negative set with unique elements removed (orange), CBMF-negative set (green), and CBGS-negative set (red).