

Supplementary Note 1: Supplementary Figures

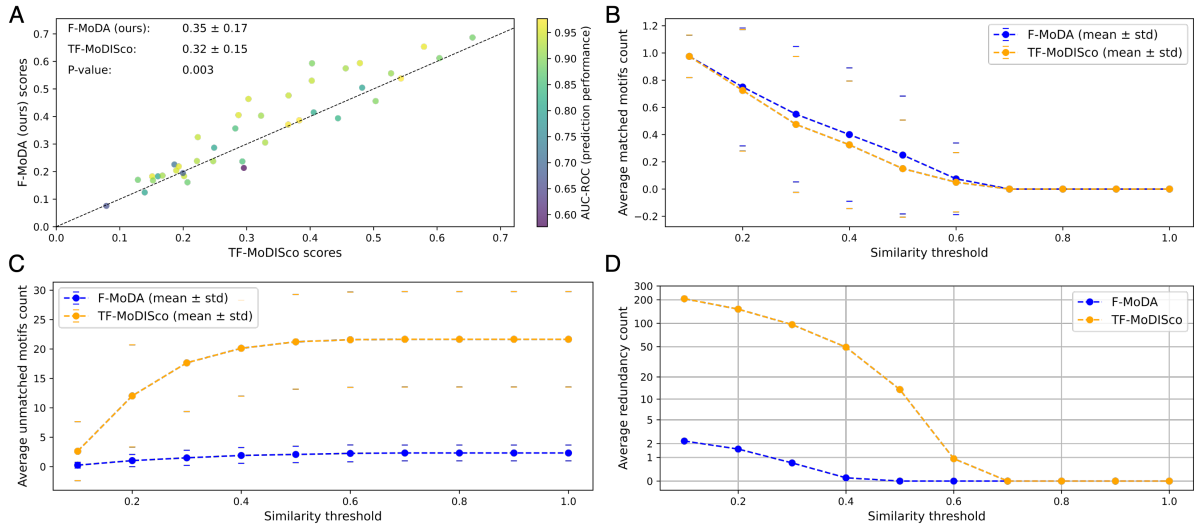


Fig. S1. Performance comparison on benchmark ChIP-seq datasets based on the continuous Jaccard similarity measure. **(A) Motif identification quality across 40 ChIP-seq experiments in K562 cells.** Each point represents the maximum similarity score between the discovered motifs and a SELEX-derived reference motif (which is considered the ground truth) for one experiment (max-similarity measure). Point colors indicate the model's prediction performance on the validation subset. Average similarity scores for F-MoDA and TF-MoDISco are also shown. Statistical significance was assessed using the Wilcoxon signed-rank test. **(B-D) Sensitivity, specificity, and conciseness measures.** Average counts per dataset versus similarity threshold for: (B) matched motifs (relative to ground truth; sensitivity measure); (C) unmatched motifs (relative to ground truth; specificity measure); and (D) internal redundancy of reported motifs (conciseness measure).

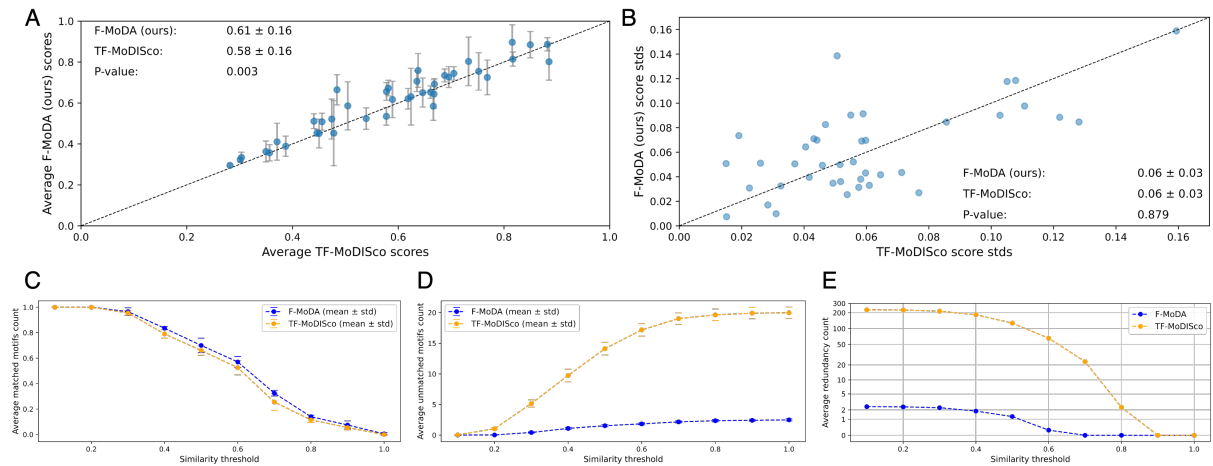


Fig. S2. Robustness across multiple training runs, evaluated using the cross-correlation similarity measure. Performance measures were averaged over five independent training runs of the CNN-Classifier model. **(A) Motif identification quality.** Each point represents the average max-similarity score for one of the 40 experiments, calculated over the five runs. Error bars denote the standard deviation across runs for F-MoDA. Overall average similarity scores for F-MoDA and TF-MoDISco are shown. Statistical significance was assessed using the Wilcoxon signed-rank test. **(B) Stability of motif similarity scores.** Each point compares the standard deviation (std) of the max-similarity scores across runs for a single experiment between F-MoDA and TF-MoDISco. The average standard deviation for each method is reported, with statistical significance assessed by the Wilcoxon signed-rank test. **(C-E) Average sensitivity, specificity, and conciseness measures.** Average counts per dataset versus similarity threshold. Each line represents the mean result across the five runs, and the corresponding error bands show the standard deviation. The plots show: (C) matched motifs (relative to ground truth; sensitivity measure); (D) unmatched motifs (relative to ground truth; specificity measure); and (E) internal redundancy of reported motifs (conciseness measure).

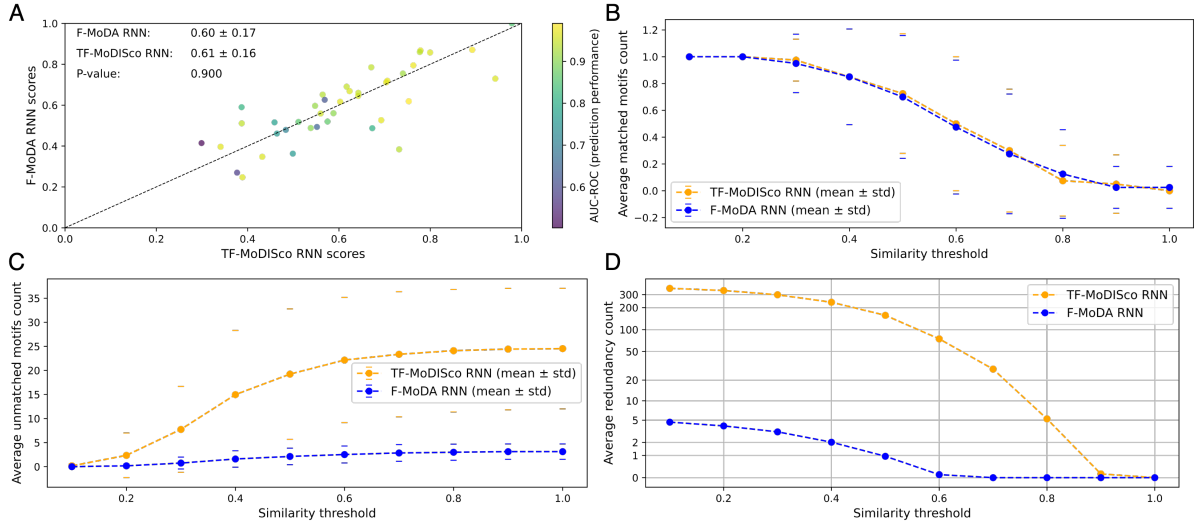


Fig. S3. Performance comparison on benchmark ChIP-seq datasets using attribution maps generated from the RNN-Classifier model. All comparisons are based on the cross-correlation similarity measure. **(A) Motif identification quality across 40 ChIP-seq experiments in K562 cells.** Each point represents the maximum similarity score between the discovered motifs and a SELEX-derived reference motif (which is considered the ground truth) for one experiment (max-similarity measure). Point colors indicate the model's prediction performance on the validation subset. Average similarity scores for F-MoDA and TF-MoDISco are also shown. No statistical significance was found between the methods (Wilcoxon signed-rank test). **(B-D) Sensitivity, specificity, and conciseness measures.** Average counts per dataset versus similarity threshold for: (B) matched motifs (relative to ground truth; sensitivity measure); (C) unmatched motifs (relative to ground truth; specificity measure); and (D) internal redundancy of reported motifs (conciseness measure).

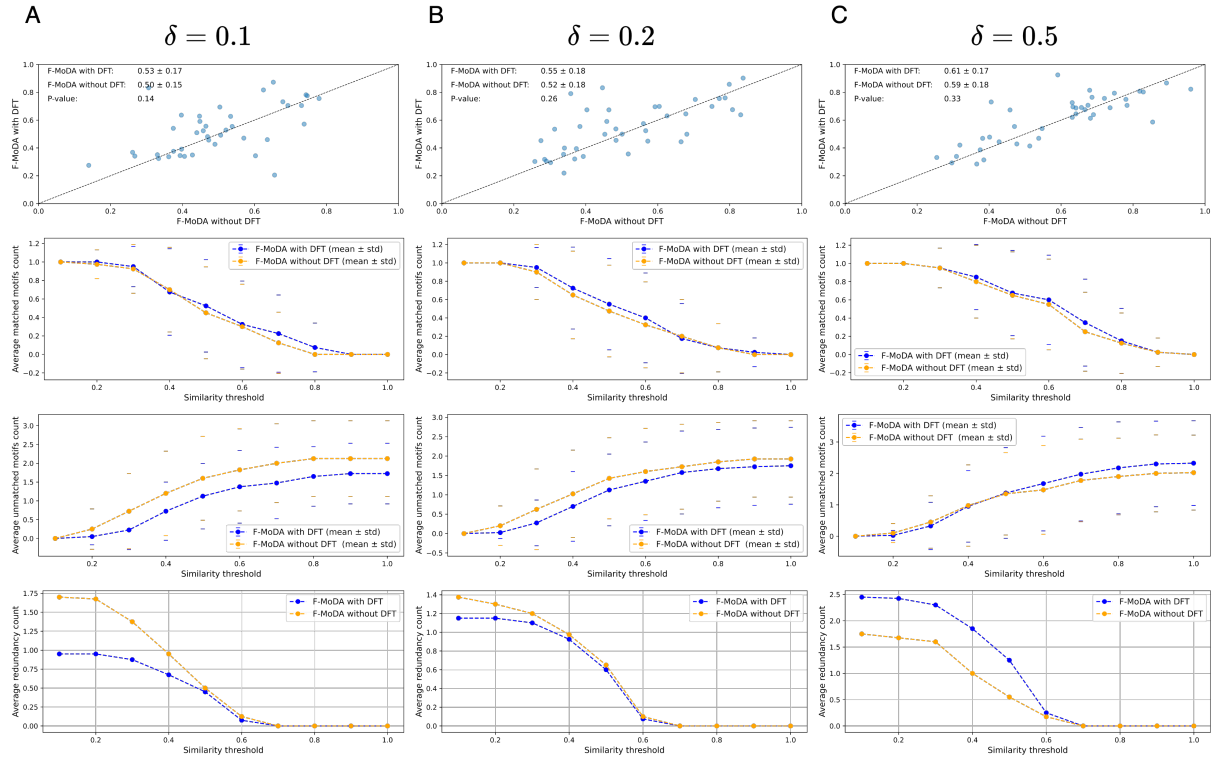


Fig. S4. Impact of the Fourier transform step on F-MoDA performance across the experimental benchmark. **(A-C) F-MoDA with the 2D-discrete Fourier transform (DFT) is compared to a version without it across all four evaluation measures.** The comparison is shown at three different drop threshold (δ) values: $\delta = 0.1$ (A), $\delta = 0.2$ (B), and $\delta = 0.5$ (C). Within each panel, the four plots compare performance on the max-similarity, sensitivity, specificity, and conciseness measures.