Cell Reports, Volume *41*

Supplemental information

Interpretable deep learning translation of GWAS and multi-omics findings to identify pathobiology and drug repurposing in Alzheimer's disease

Jielin Xu, Chengsheng Mao, Yuan Hou, Yuan Luo, Jessica L. Binder, Yadi Zhou, Lynn M. Bekris, Jiyoung Shin, Ming Hu, Fei Wang, Charis Eng, Tudor I. Oprea, Margaret E. Flanagan, Andrew A. Pieper, Jeffrey Cummings, James B. Leverenz, and Feixiong Cheng

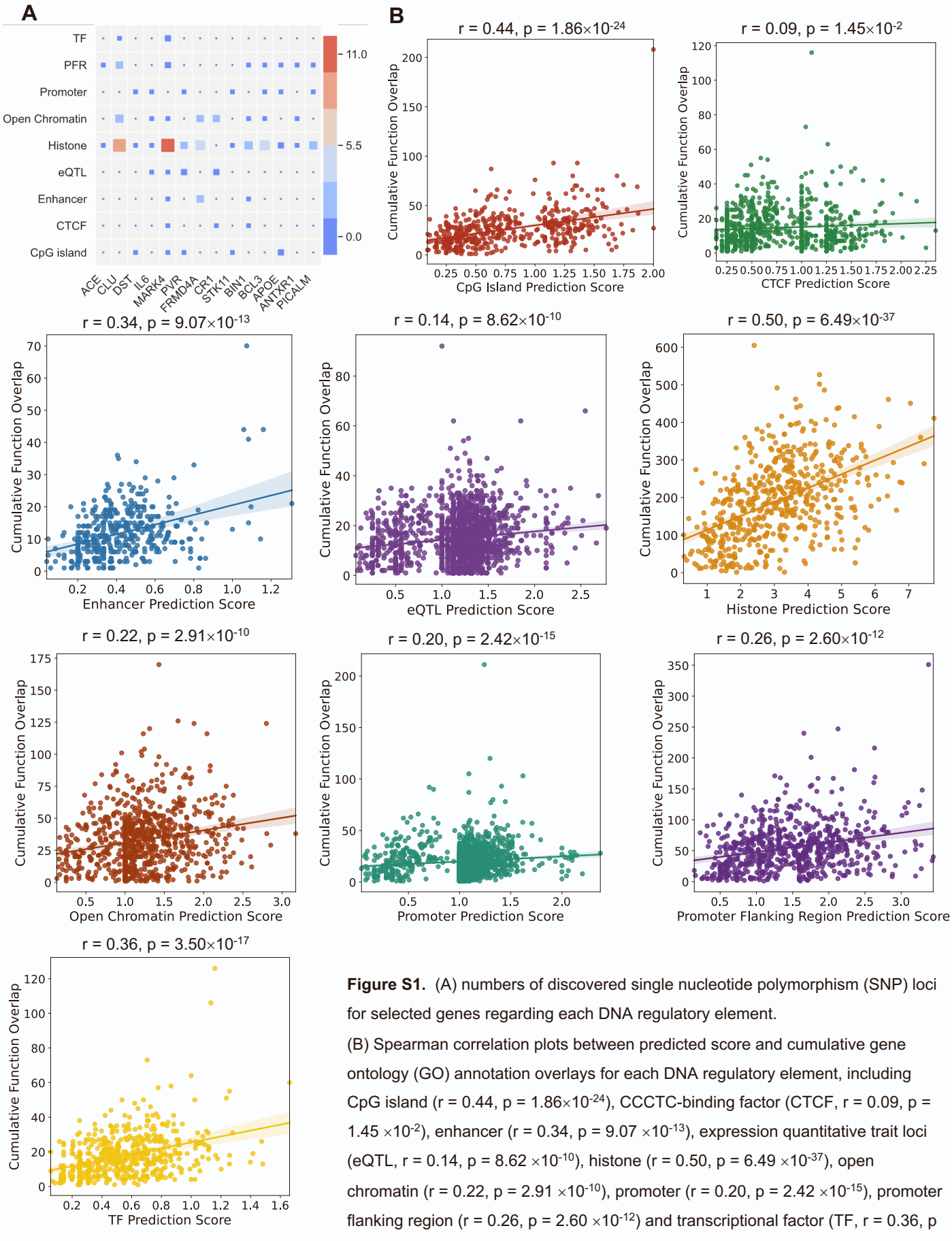


Figure S1. (A) numbers of discovered single nucleotide polymorphism (SNP) loci for selected genes regarding each DNA regulatory element.

(B) Spearman correlation plots between predicted score and cumulative gene ontology (GO) annotation overlays for each DNA regulatory element, including CpG island (r = 0.44, p = 1.86×10^{-24}), CCCTC-binding factor (CTCF, r = 0.09, p = 1.45×10^{-2}), enhancer (r = 0.34, p = 9.07×10^{-13}), expression quantitative trait loci (eQTL, r = 0.14, p = 8.62×10^{-10}), histone (r = 0.50, p = 6.49×10^{-37}), open chromatin (r = 0.22, p = 2.91×10^{-10}), promoter (r = 0.20, p = 2.42×10^{-15}), promoter flanking region (r = 0.26, p = 2.60×10^{-12}) and transcriptional factor (TF, r = 0.36, p = 3.50×10^{-17}).

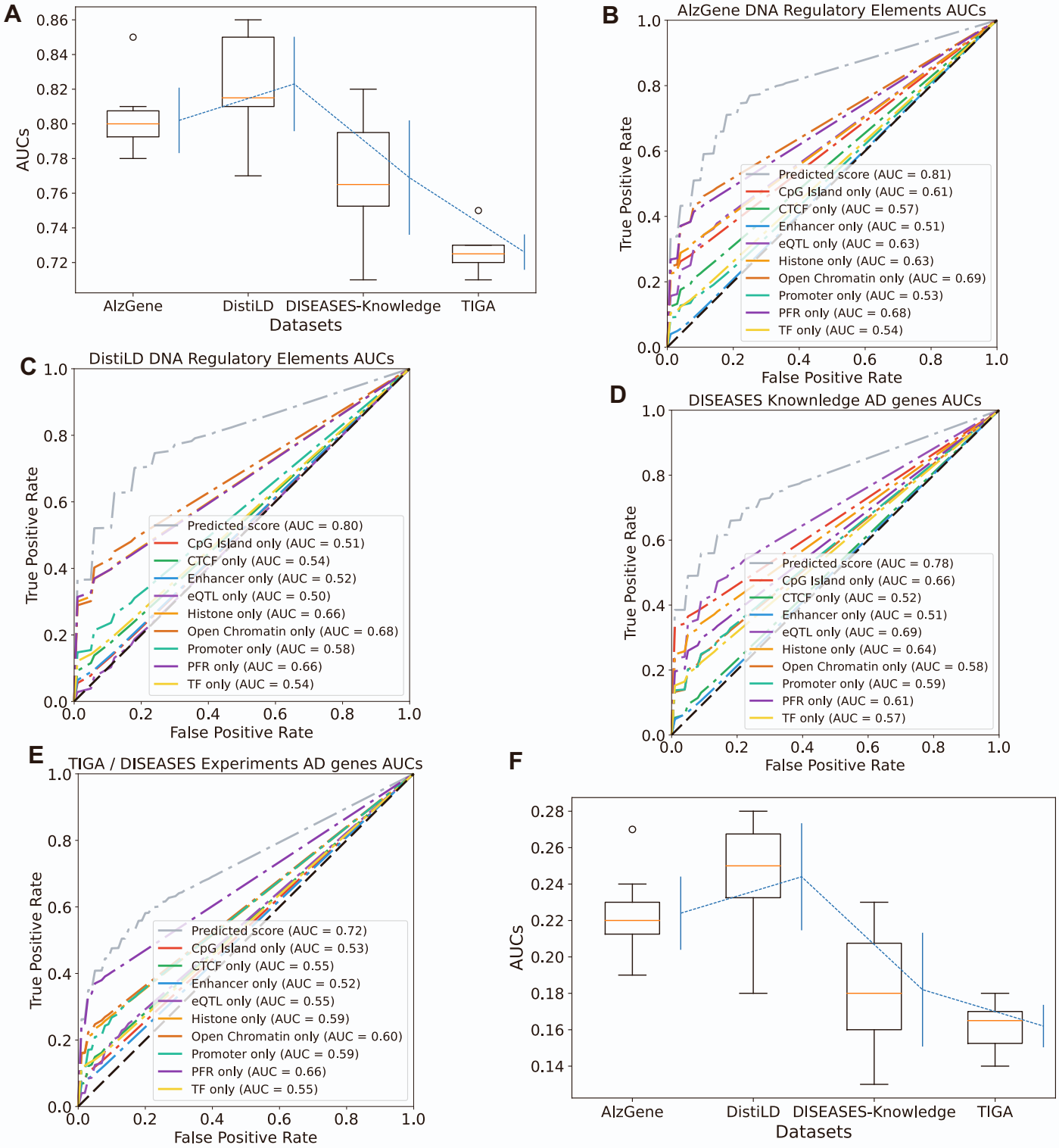


Figure S2. (A) ROC analyses based on AD seed gene sets collected (#s of replicates = 10).
(B) Receiver operating characteristic (ROC) analyses between predicted (integrated) score and scores by considering each DNA regulatory element alone based on AD seed gene set from AlzGene.
(C) ROC analyses between predicted (integrated) score and scores by considering each DNA regulatory element alone based on AD seed gene set from DistiLD.
(D) ROC analyses between predicted (integrated) score and scores by considering each DNA regulatory element alone based on AD seed gene set from DISEASES (knowledge).
(E) ROC analyses between predicted (integrated) score and scores by considering each DNA regulatory element alone based on AD seed gene set from TIGA.
(F) The differences of area under the curves (AUCs) between integration and single DNA regulatory element with respect to four AD knowledgebases, i.e., AlzGene, DistiLD, DISEASES (knowledge) and TIGA with ten repeats of NETTAG (#s of replicates = 10).

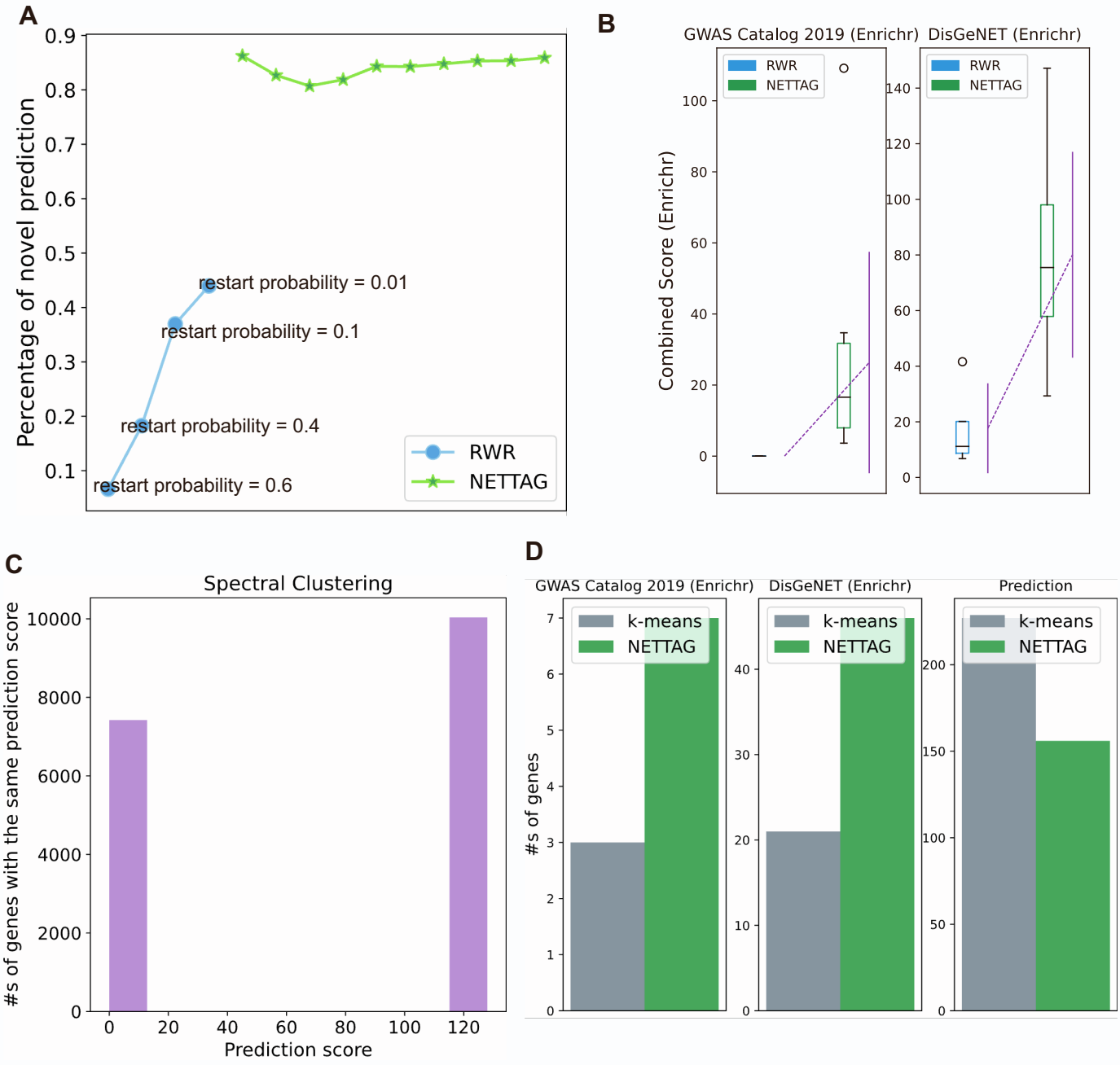


Figure S3. (A) Percentage of novel genes predicted by random walk with restart (RWR) with different restarting probabilities (blue) and NETTAG including 10 runs with different seeds (green) (**Table S3**) .
(B) Disease enrichment analyses with GWAS Catalog and DisGeNET considering novel predicted genes only (predicted genes that are not overlapped with input genes). The y-axis represents the combined score, which equaled product of –log(p-value) and z-score as defined by Enrichr (see **STAR Methods** and **Table S3**). The higher the combined score is, the more likely the predicted genes are specific disease-related. For GWAS Catalog, the plotted combined score is the maximum combined scores with disease names containing "Alzheimer". Novel genes predicted by RWR were not enriched with any disease item that having "Alzheimer". For DisGeNET, the combined score is extracted with "Alzheimer's Disease" as the disease name. In this plot, we consider 4 (#s of replicates) different restarting probabilities (0.01, 0.1, 0.4 and 0.6) for RWR, and 10 (#s of replicates) repeats of NETTAG.
(C) Spectral clustering cannot differentiate genes by prediction score (see **STAR Methods** and **Table S3**), i.e., multiple genes shared the same gene scores.
(D) Compared to k-means, NETTAG has less total predicted genes (NETTAG 156, k-means 227), but more AD related genes according to GWAS Catalog (NETTAG:7, k-means: 3, disease name = "Alzheimer's Disease) and DisGeNET (NETTAG: 46, k-means: 21, disease name = 'Alzheimer's Disease") (see **STAR Methods** and **Table S3**).

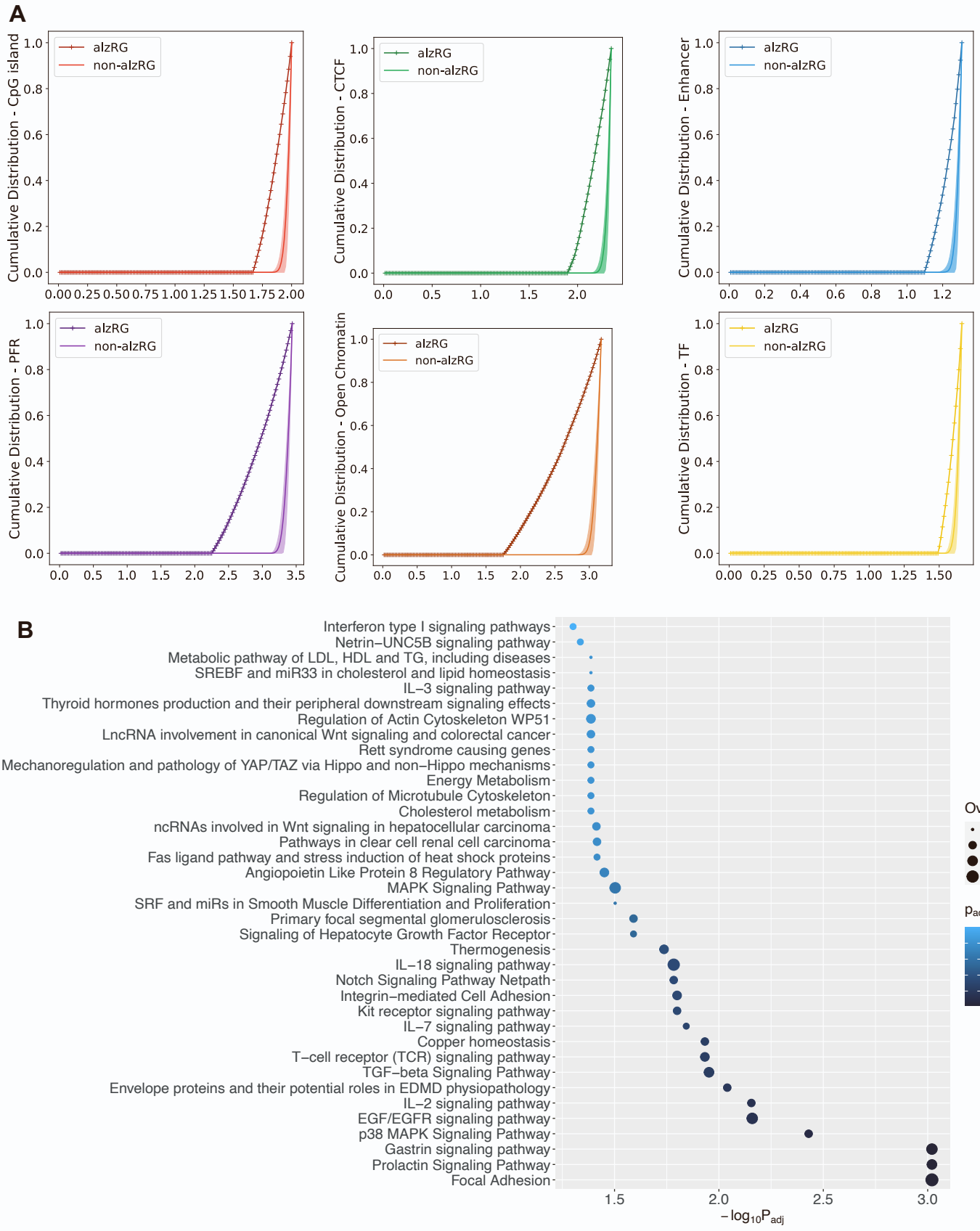
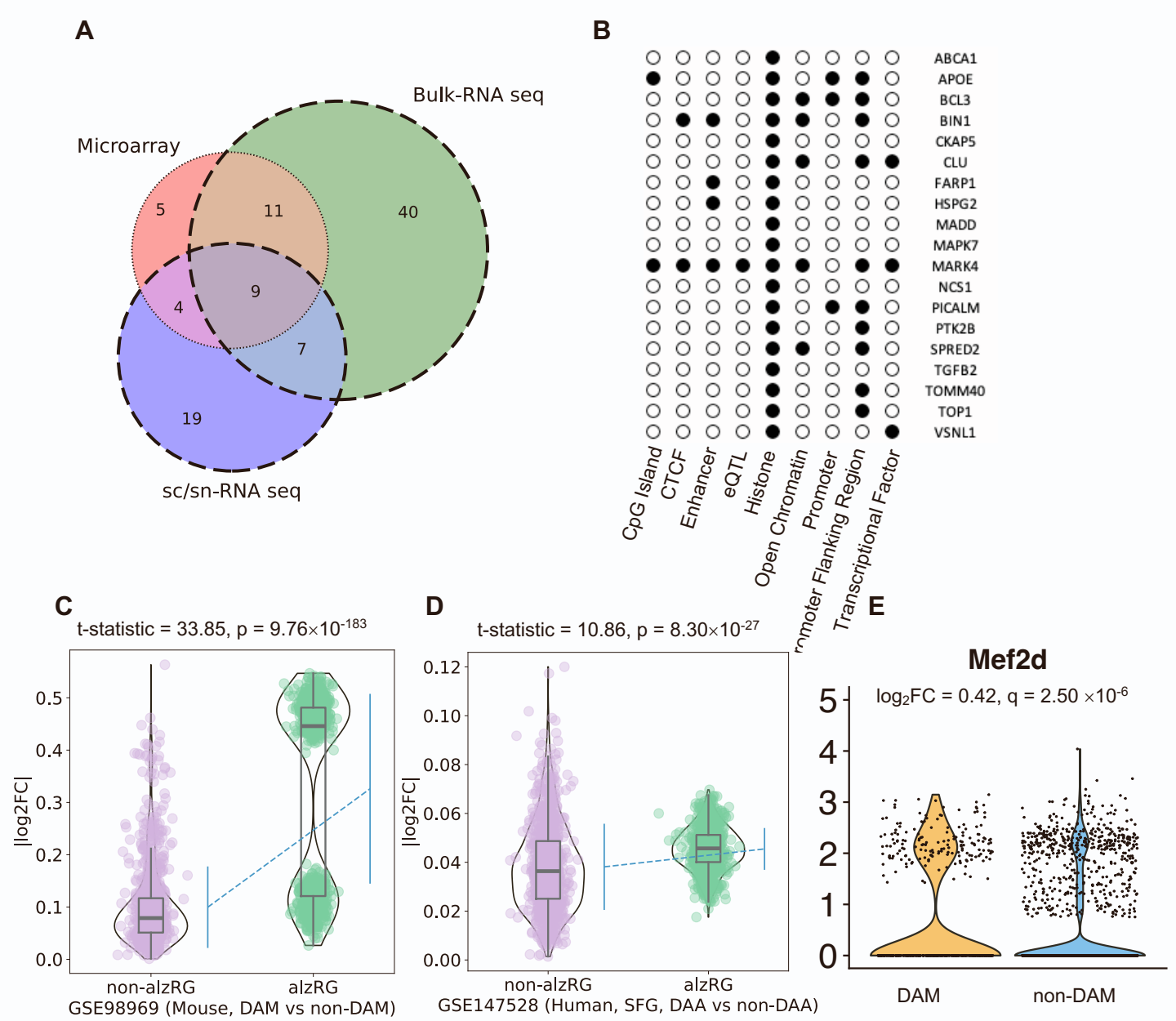


Figure S4 (A) Cumulative distributions of predicted scores with alzRGs and same amount of random non-alzRGs with similar degree distribution for CpG island, CCCTC-binding factor (CTCF), enhancer, open chromatin, promoter flanking region (PFR) and transcriptional factor (TF) regulatory elements, respectively.
(B) Pathway enrichment analysis (Wikipathway, see **STAR Methods**).



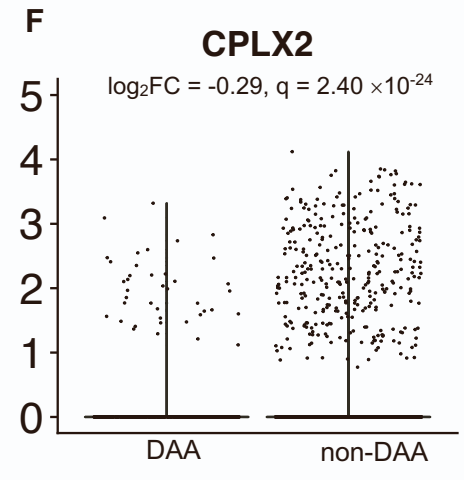


Figure S5. (A) Venn diagram shows the numbers of common differentially expressed genes (DEGs) according to microarray, bulk RNA-seq and single cell / nucleus RNA-seq studies.

(B) Nineteen DEGs are associated with various types of DNA regulatory elements (black dots).

(C) and (D) Violin plots show alzRGs are more likely differentially expressed in disease-associated microglia (DAM) and astrocyte (DAA) according to one mouse single-cell RNA seq dataset GSE98969 (C, 16 5xFAD and 16 wild type mice) and one human single-nucleus RNA seq dataset GSE147528 (D, 3,4 and 3 human postmortem brain samples with Braak stage 0, 2, and 6, respectively) [unpaired t-test, t-test statistic = 33.85 and 10.86, p = 9.76 $\times 10^{-183}$ and 8.30 $\times 10^{-27}$, separately, #s of replicates = 1000].

(E) One of our top predicted likely AD-associated gene MEF2D is found to upregulated [log2FC = 0.42, q = 2.50 $\times 10^{-6}$] in disease-associated microglia (DAM) according to one mouse single-nucleus RNA-seq dataset GSE140511 [3 mice for each group: 5xFAD, wild type, Trem2 knock out 5xFAD, and Trem2 knock out wild type].

(F) Another top predicted likely AD-associated gene CPLX2 is found to down-regulated [log2FC = -0.29, q = 2.40 $\times 10^{-24}$] in DAA according to one human single-nucleus RNA-seq dataset GSE157827 [9 normal control and 12 AD human postmortem brain samples].

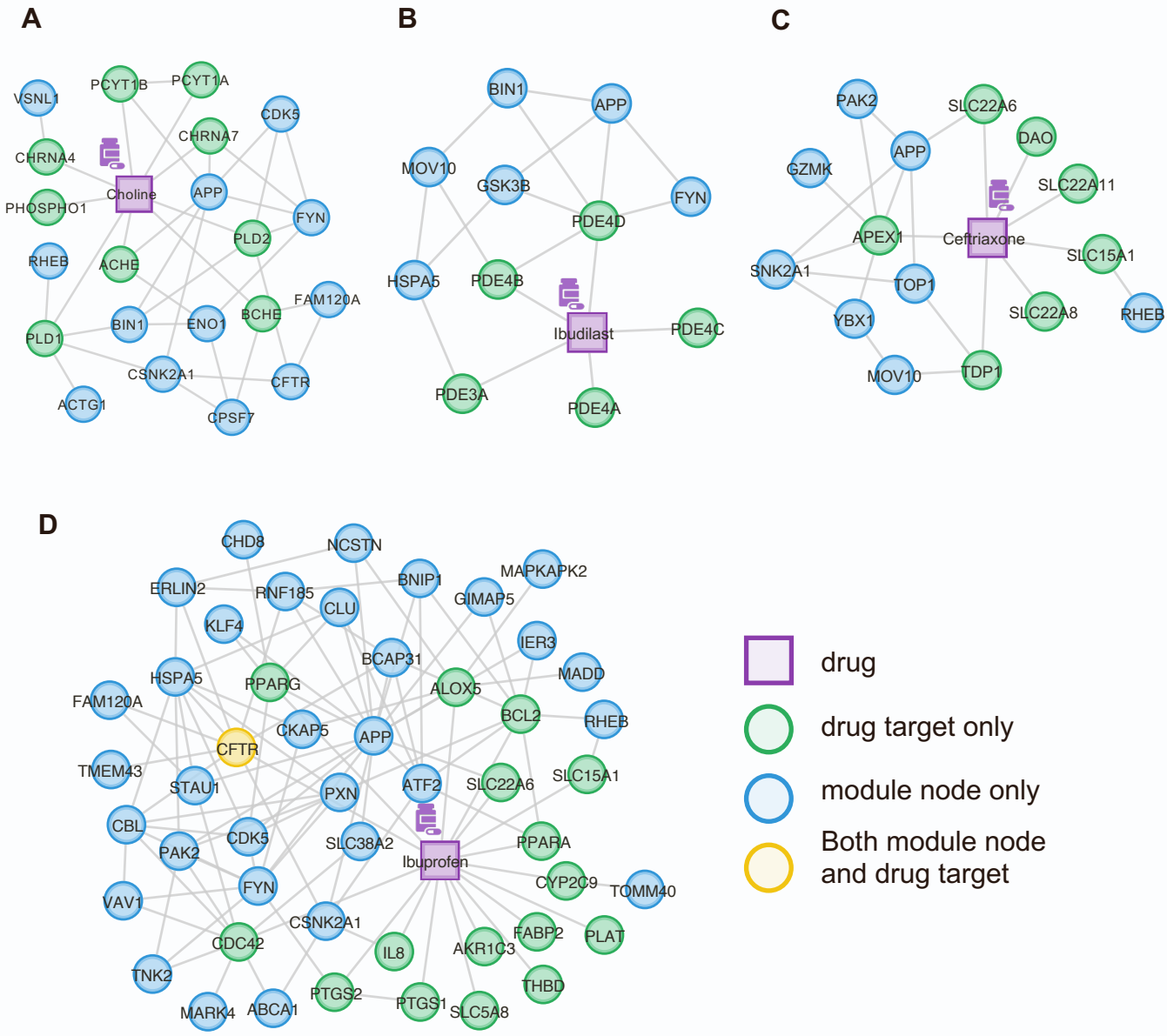


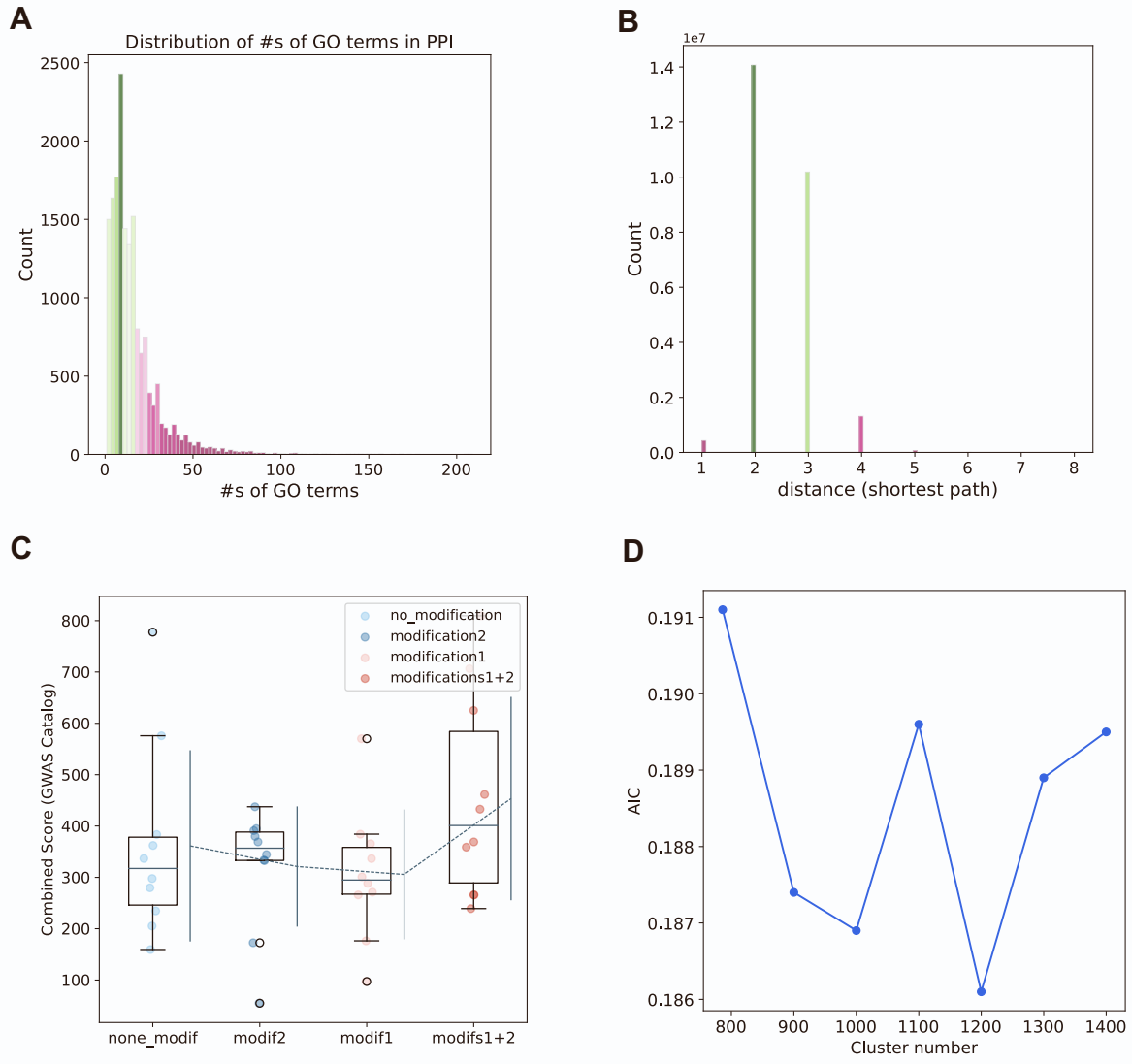
Figure S6. Drug target network analyses for 4 prioritized drugs.

(**A**) Choline (FDA approved).

(**B**) Ibudilast (Investigational, anti-inflammatory and neuroprotective oral agent).

(**C**) Ceftriaxone (FDA approved, antibiotic).

(**D**) Ibuprofen (FDA approved, NSAID and non-selective COX inhibitor).



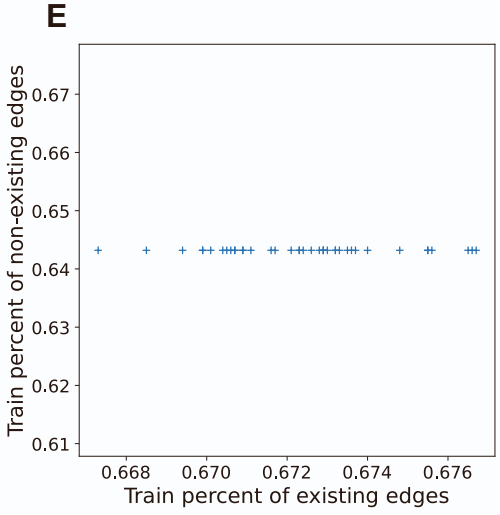


Figure S7. (**A**) Histogram of proteins' numbers of gene ontology (GO) terms.
(**B**) Histogram of shorted distanced regarding our protein-protein interaction (PPI) with respect to all protein pairs that share at least one common GO annotation.
(**C**) Disease enrichment analysis: boxplot of GWAS Catalog combined scores (from left to right): without any modification, modification 2 (subsampled topological similar subgraphs, see **STAR Methods**), modification 1 (coupling 2nd order term, see **STAR Methods**), and both modifications. The combined score equals product of –log(p-value) and z-score as defined by Enrichr (see **STAR Methods**). The higher the combined score is, the more likely the predicted genes are specific disease-related. Here we consider 'Alzheimer's Disease' from GWAS Catalog (#s of replicates = 10).
(**D**) Akaike information criterions (AICs) used for cluster number determination (see **STAR Methods**).
(**E**) Percentages of trained existing edges and those of trained non-existing edges in all training iterations are quite similar.

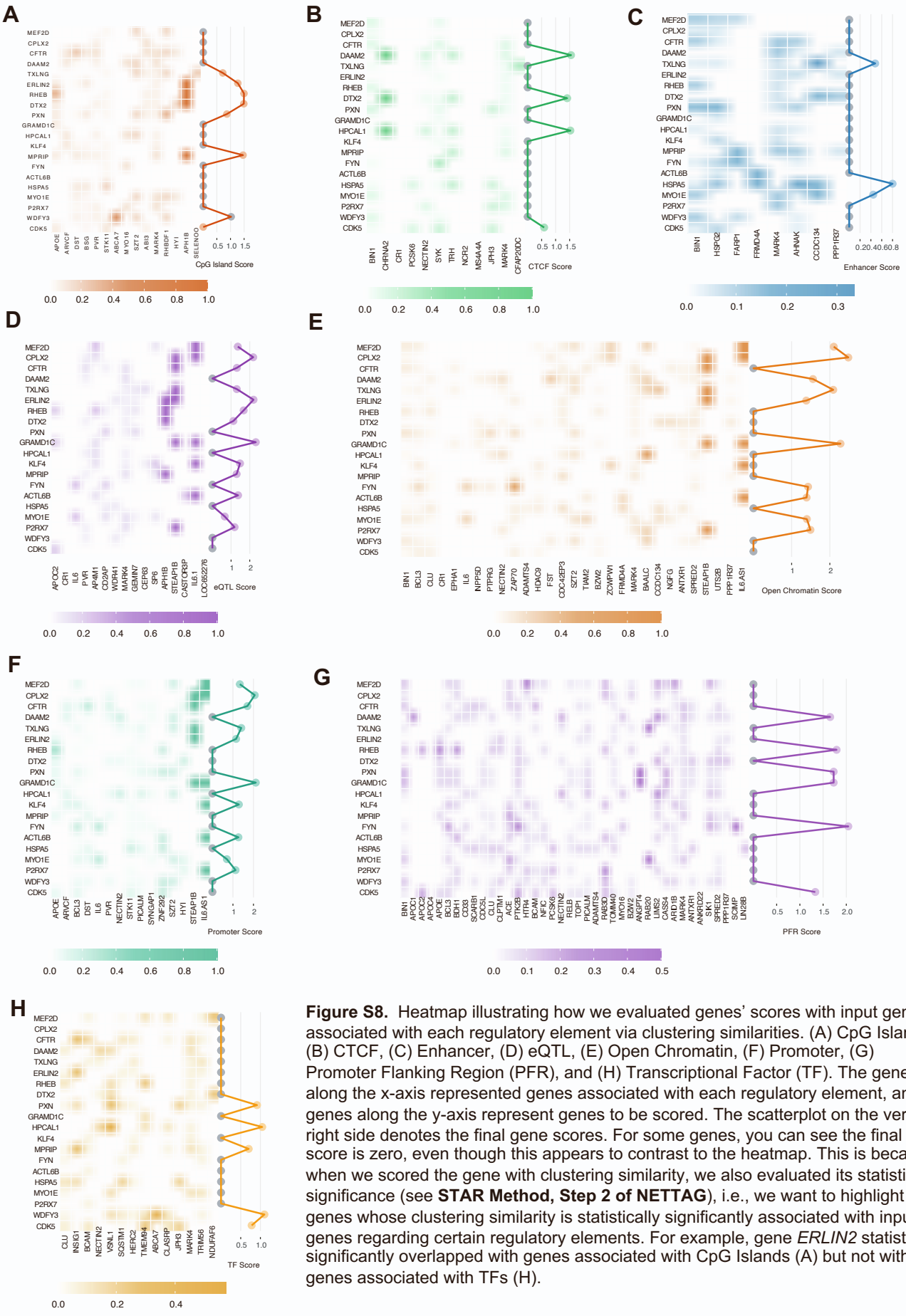


Figure S8. Heatmap illustrating how we evaluated genes' scores with input genes associated with each regulatory element via clustering similarities. (A) CpG Island, (B) CTCF, (C) Enhancer, (D) eQTL, (E) Open Chromatin, (F) Promoter, (G) Promoter Flanking Region (PFR), and (H) Transcriptional Factor (TF). The genes along the x-axis represented genes associated with each regulatory element, and genes along the y-axis represent genes to be scored. The scatterplot on the very right side denotes the final gene scores. For some genes, you can see the final score is zero, even though this appears to contrast to the heatmap. This is because when we scored the gene with clustering similarity, we also evaluated its statistical significance (see **STAR Method, Step 2 of NETTAG**), i.e., we want to highlight only genes whose clustering similarity is statistically significantly associated with input genes regarding certain regulatory elements. For example, gene *ERLIN2* statistically significantly overlapped with genes associated with CpG Islands (A) but not with genes associated with TFs (H).