Supplemental Figures:

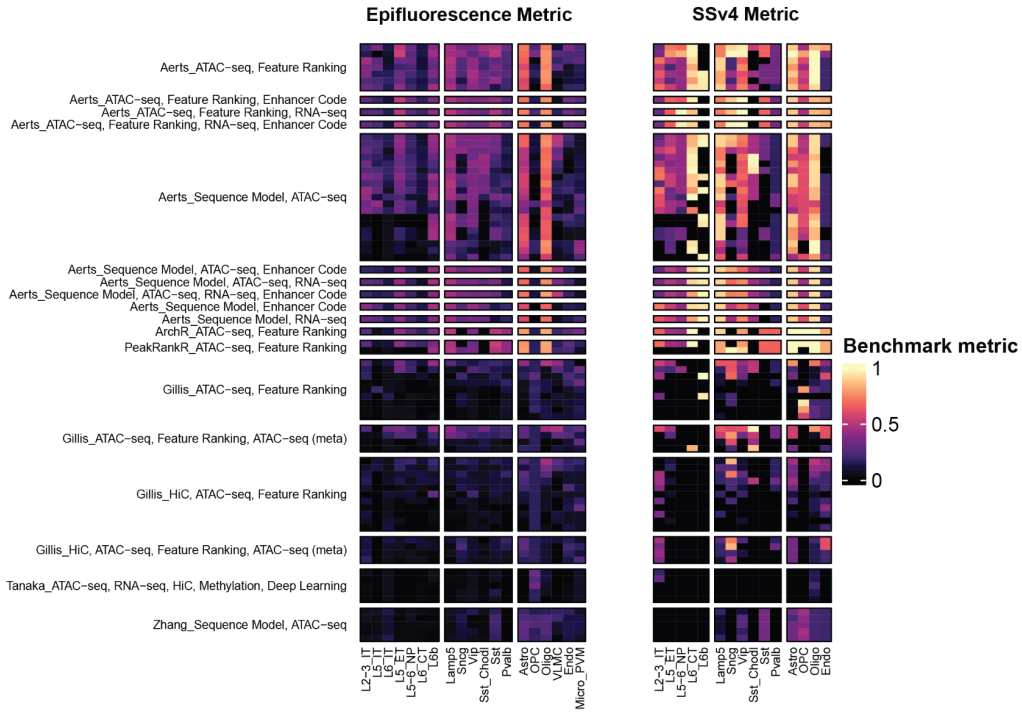

Figure S1. Scoring of all team submissions based on two measures of *in vivo* enhancer activity. Submissions are grouped by team and approach. The heatmaps visualize the epifluorescence and SSv4 normalized benchmark metrics for each submission per subclass. Higher benchmark metric values represent agreement with the validated enhancer collection.

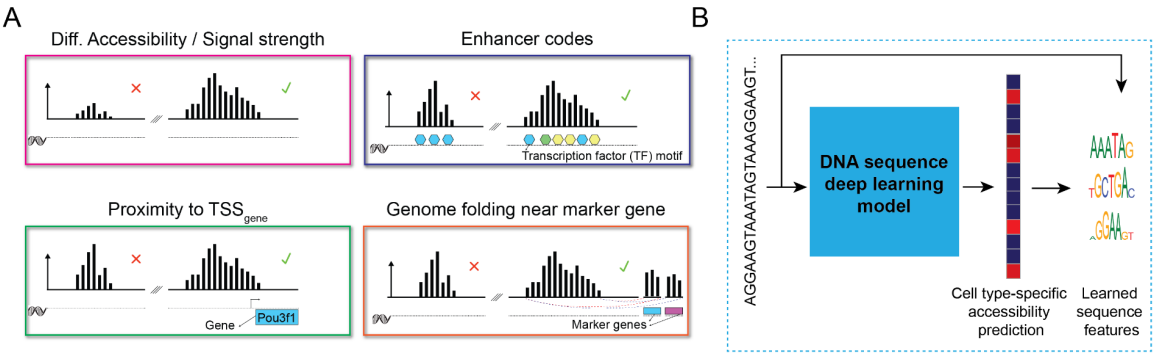

Figure S2. Illustration of biological priors, methods and DNA sequence models
Illustration of several genomic features and computational methods used by top-performing teams.

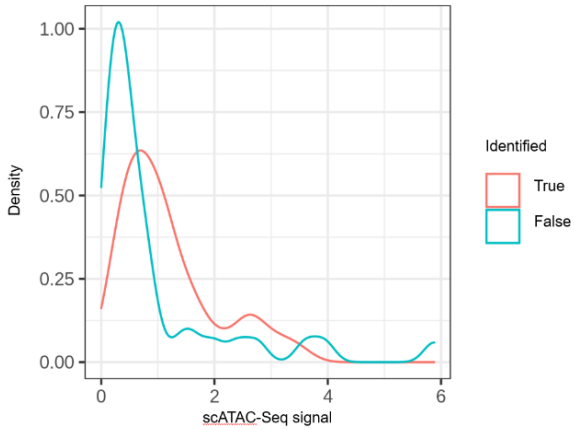

Figure S3. ATAC-seq signal for peaks called or missed by teams.
Density plots of ATAC-seq signal for enhancers identified by at least one team or missed by all teams.

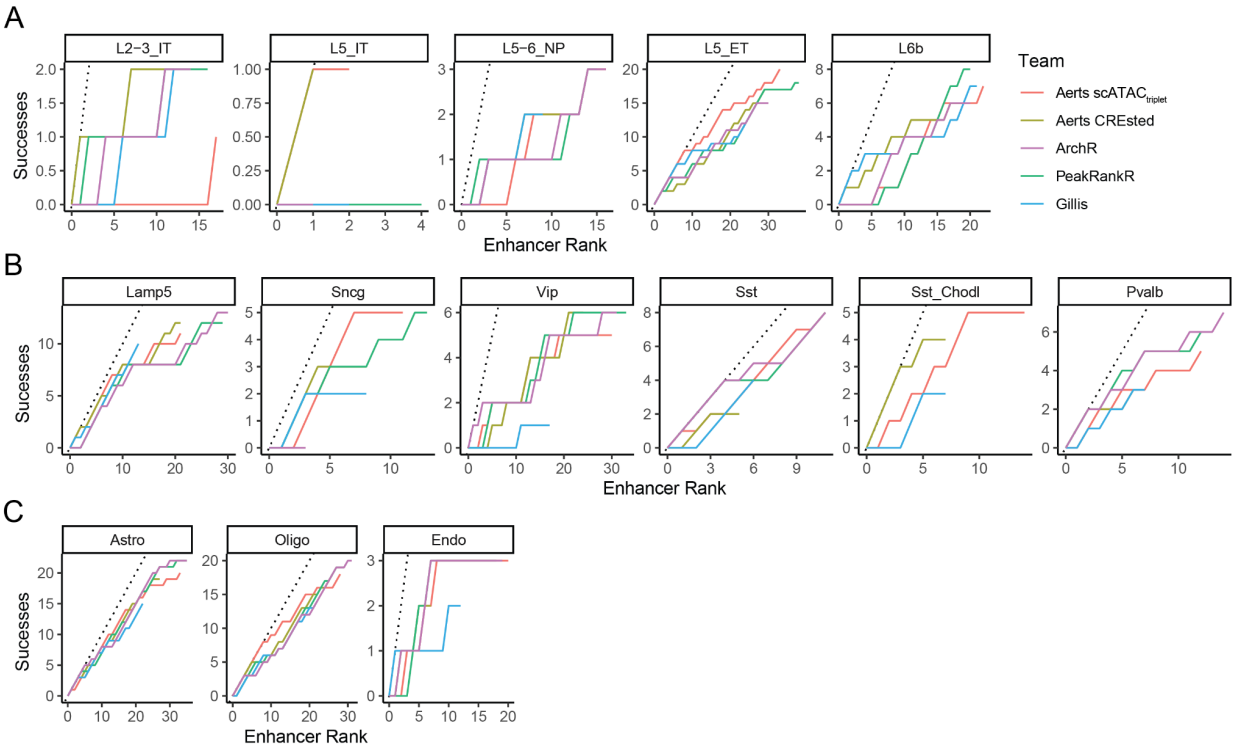

Figure S4. Rates of functional enhancer identification vary by cell type and submission.
Counts of On-Target enhancers identified for (A) excitatory and (B) inhibitory neuronal types and (C) non-neuronal types as a function of the number of validated enhancers for each type. Note that for some cell types (e.g. L5 IT), only some teams submitted any On-Target enhancer.

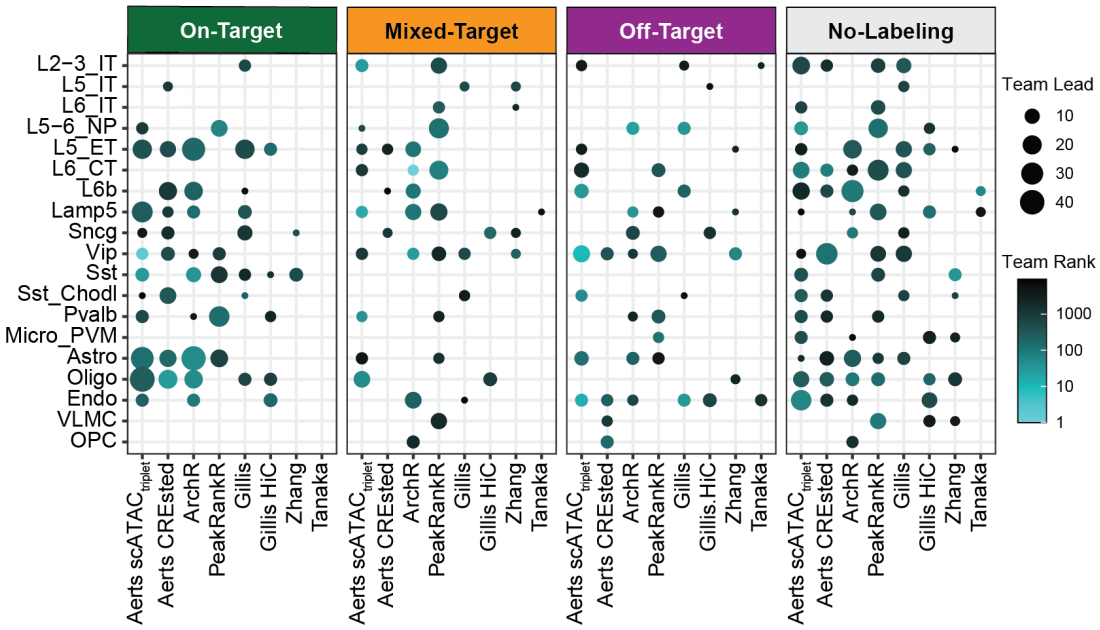

Figure S5. Summary of the highest ranked enhancers across submissions.
Enhancers were grouped by *in vivo* validated activity (e.g. On-Target), and point size indicates the number of enhancers that each submission scored the highest (i.e. lowest rank) compared to the other submissions. The color corresponds to the median rank of the enhancers included in each point. Note that better performing teams had larger, lighter points in the On-Target category (i.e., more enhancers that were scored higher) and smaller, darker points in the Mixed-Target, Off-Target, and No-Labeling categories (i.e., fewer enhancers that were scored lower).

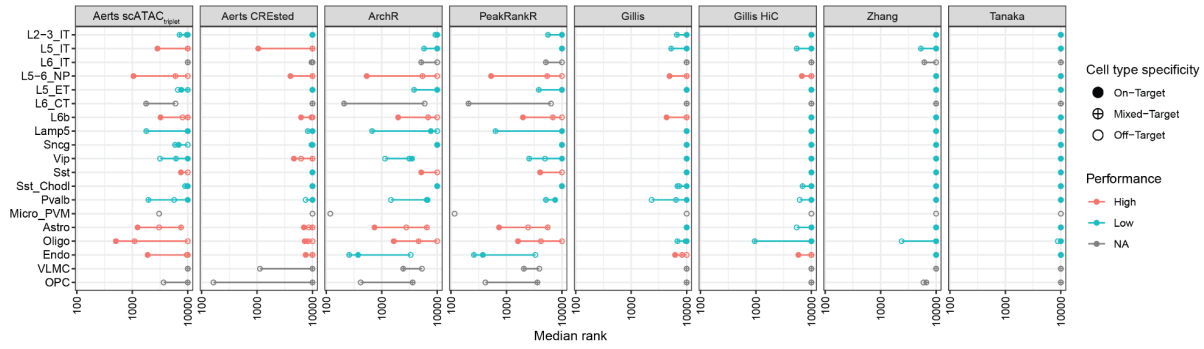

Figure S6. Comparison of enhancer rankings across submissions, cell types, and *in vivo* labeling results.
Median ranks of enhancers grouped by submission, cell type, and specificity. Rankings were labeled "high performance" if the median rank of On-Target enhancers was lower than the rankings of Mixed-Target and Off-Target enhancers. If no On-Target enhancer rankings were submitted, then the performance was labeled "NA".

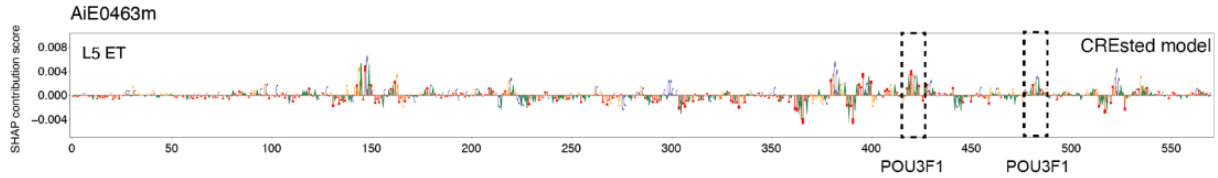

Figure S7. *Pou3f1* enrichment in CREsted model for AiE0463m.
The nucleotide contribution score track for enhancer AiE0463m for the L5 ET class of the Aerts CREsted model. Two *Pou3f1* motifs were identified around position 420 and 485.

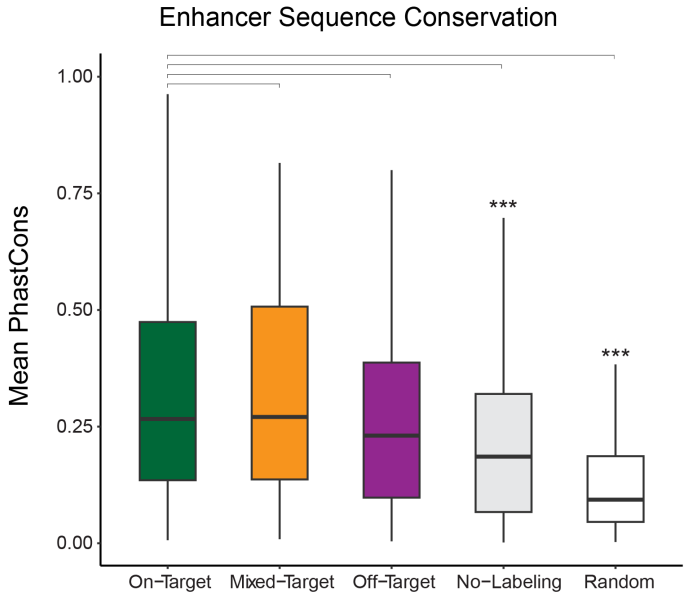

Figure S8. DNA sequence conservation of screened enhancers grouped by cell type labeling results.
Comparison between On-Target and other enhancer categories. *** P < 0.0001, Wilcoxon rank-sum test two-sided, unpaired, Bonferroni-corrected P-values.

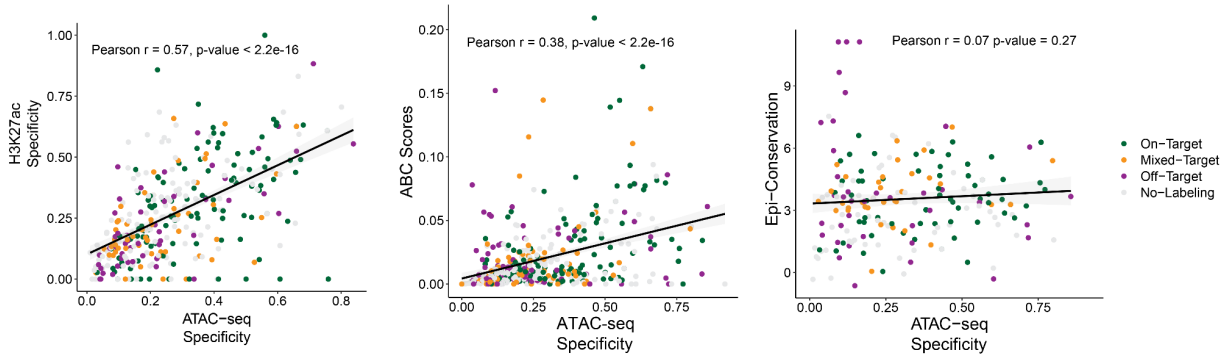

Figure S9. Variable correlation of genomic features with ATAC-seq specificity.
H3K27ac specificity, ABC score, and open chromatin conservation between mouse and human compared to ATAC-seq specificity for all screened enhancers. Note that the candidate enhancers with the highest conservation of open chromatin (upper left quadrant of the third plot) were located at promoters and lacked cell type-specific activity.

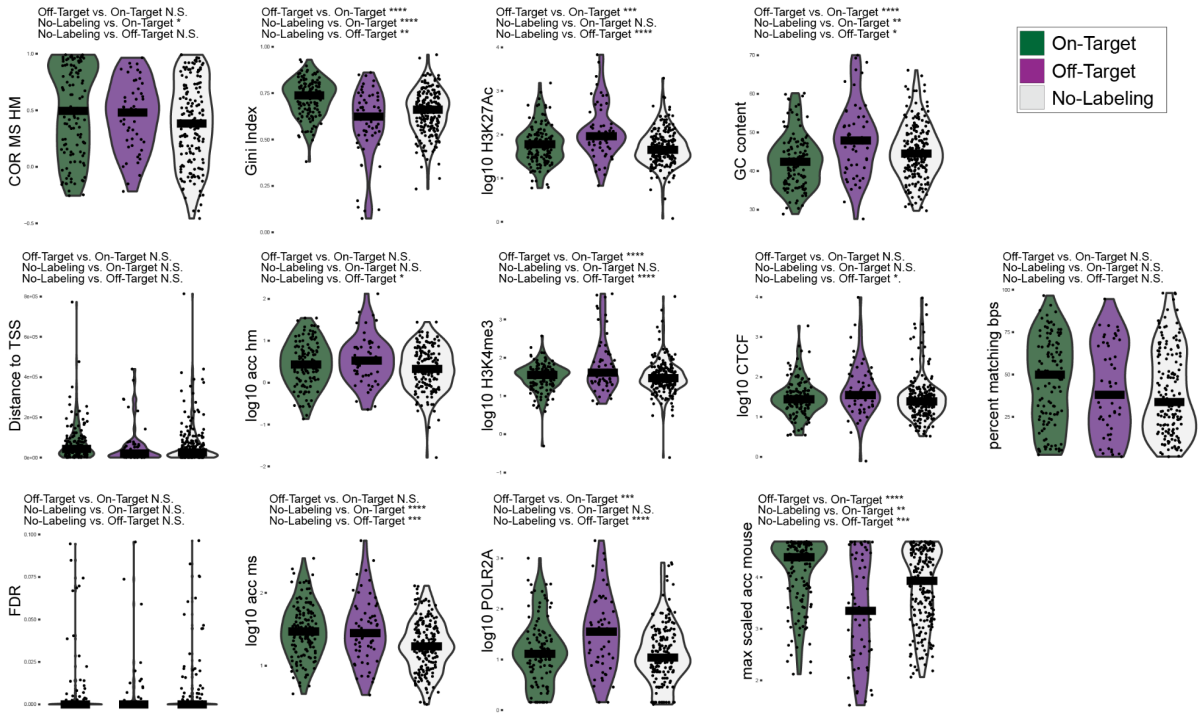

Figure S10. Genomic features compared across high- and low-performance enhancers.
Enhancer sequence features, species conservation and published bulk ChIP-seq data (ENCODE dataset ENCFF203KID) were used to train supervised random forest models to predict *in vivo* enhancer activity based on primary scoring data. Models were developed and optimized using 10-fold cross-validation in scikit-learn. Models were then tested using a held-out test set (70% training, 30% testing). Violin plots were generated for all features grouped by enhancer activity. * P < 0.05, ** P < 0.01, *** P < 0.001, **** P < 0.0001, ANOVA and Tukey post hoc tests, Bonferroni-corrected P-values.

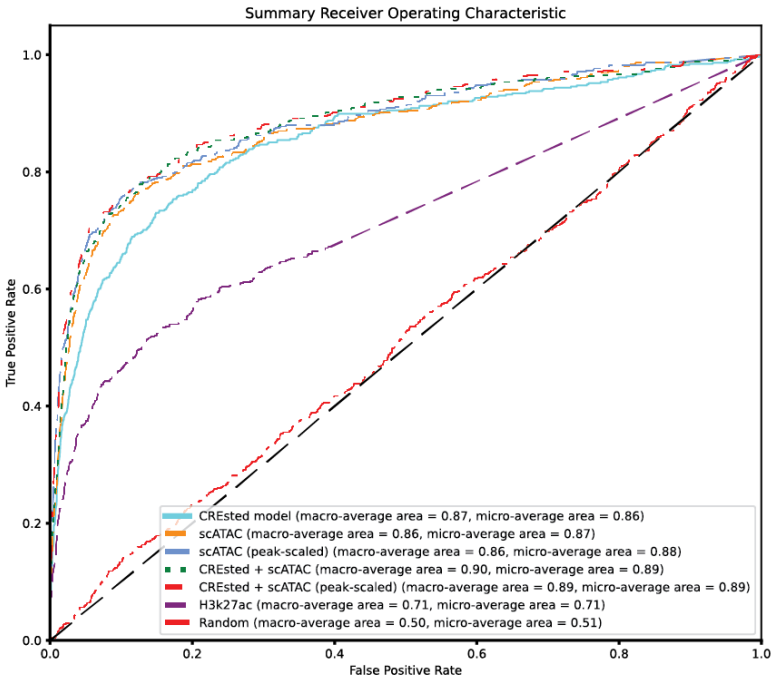

Figure S11. Performance comparison of several models predicting *in vivo* enhancer activity.
Area under the receiver operator curve to assess performance of enhancer ranking methods using the rescored enhancer activities.

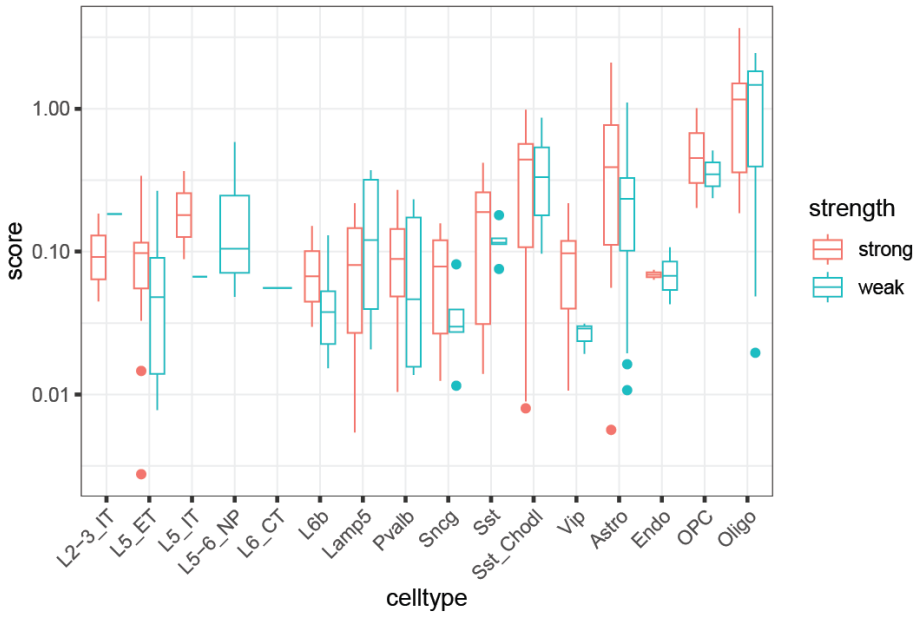

Figure S12. Comparisons of DNA sequence model scores for strong and weak enhancers.
Boxplots of CREsted scores for candidate enhancers grouped by *in vivo* validated cell type target and strength of activity. Median +/- inter-quartile interval, outliers are points.

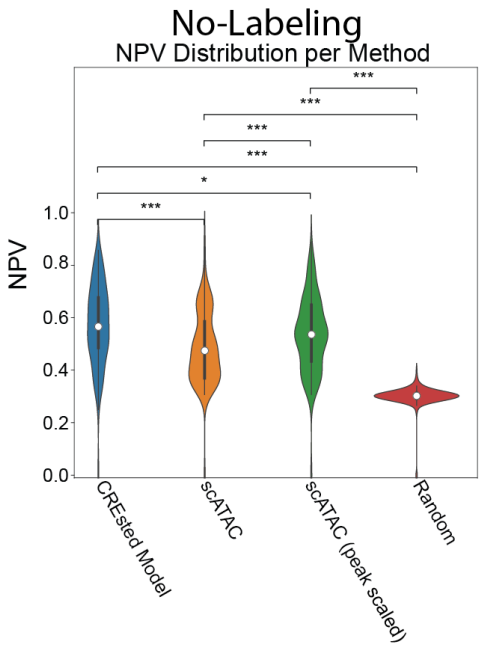

Figure S13. CREsted model is better at identifying enhancers lacking *in vivo* activity.
The negative predictive value (NPV) of the CREsted model is significantly higher than the NPVs of the initial and peak-scaled ATAC-seq models. * P < 0.05, *** P < 0.001, Wilcoxon rank-sum test two-sided, unpaired, Bonferroni-corrected P-values.

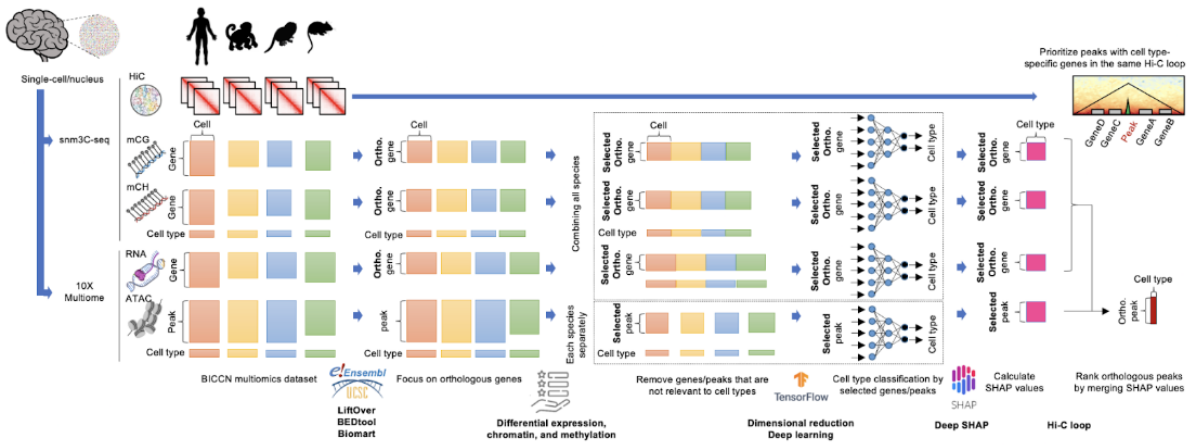

Figure S14. Yoshiaki Tanaka lab enhancer prediction overview. cisMultiDeep: Identifying Cell Type-Specific Cis-Regulatory Regions by Automatically-Tuned Deep Neural Network and SHAP.

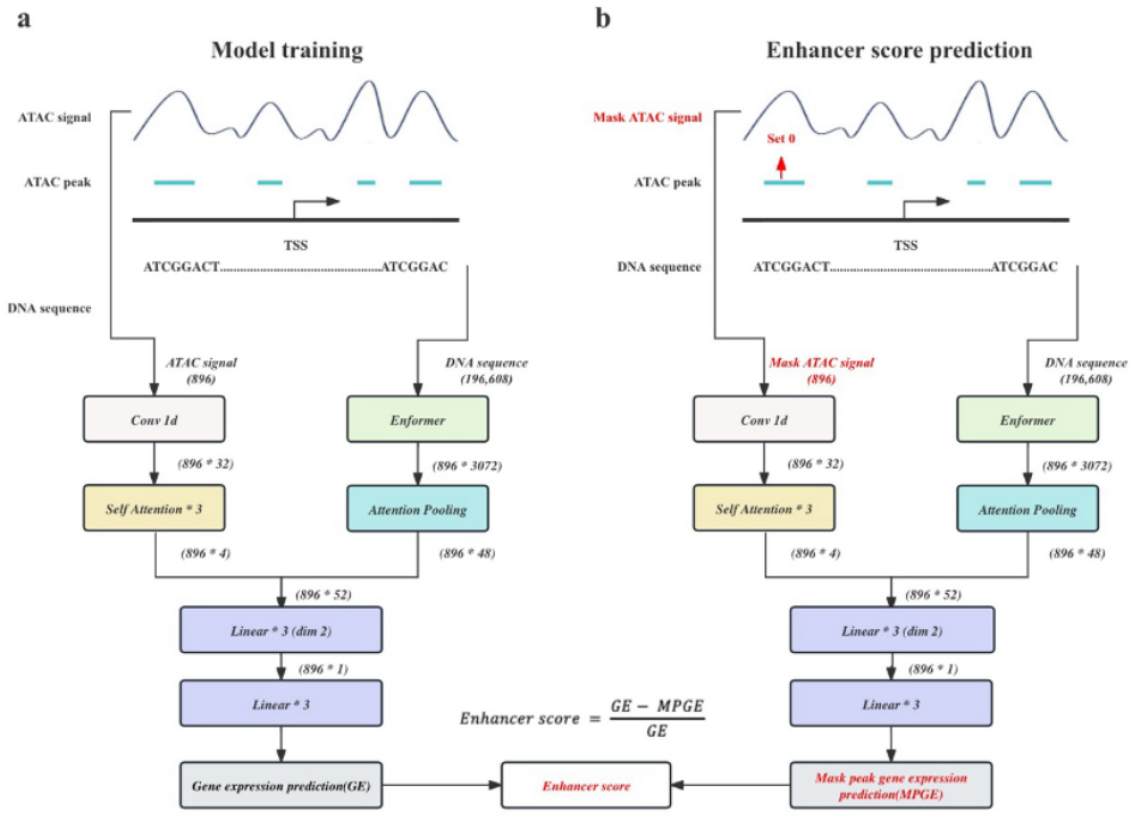

Figure S15. Kai Zhang lab enhancer prediction overview.

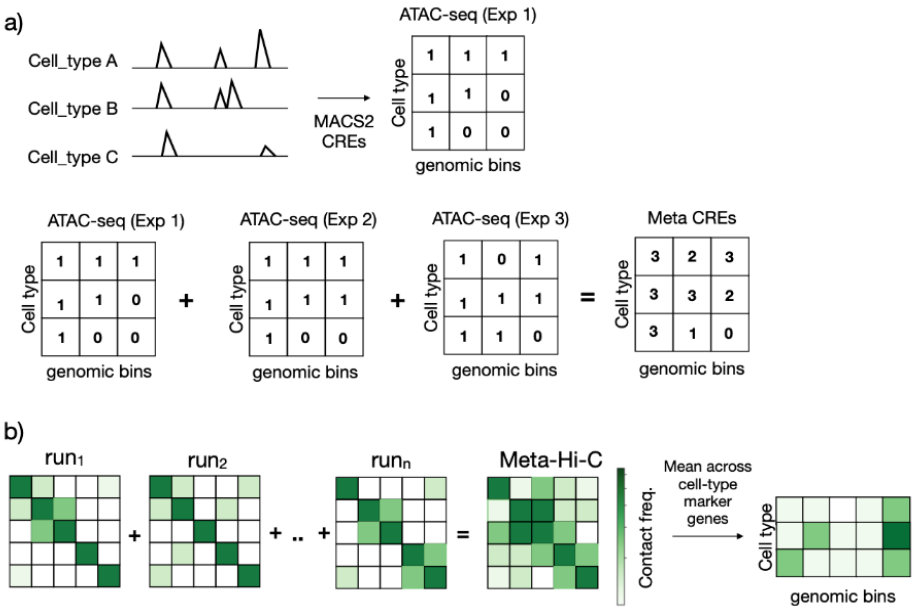

Figure S16. Gillis lab enhancer prediction overview.

Supplemental Tables:

Supplement Table 1. Summary of validated enhancers.
Table of validated enhancer data showing all counts.

Supplement Table 2. Summary of team submissions.
Table of benchmark metrics for all submissions during the BICCN challenge.

Supplement Table 3. Summary of team submission groups.
Table of submission names used in figures.