Multimodal learning decodes the global binding landscape of chromatin-associated proteins

Jimin Tan[1,2,3#*], Xi Fu[4,5*], Xinyu Ling[1*], Shentong Mo[6], Jiangshan Bai[1], Raúl Rabadán[4,5], David Fenyö[2,7], Jef D. Boeke[2,7,8], Aristotelis Tsirigos[3,9#], Bo Xia[1,2,10#]

1 Gene Regulation Observatory, Broad Institute of MIT and Harvard, Cambridge, MA, USA
2 Institute for Systems Genetics, NYU Grossman School of Medicine, New York, NY, USA
3 Division of Precision Medicine, NYU Grossman School of Medicine, New York, NY, USA
4 Program of Mathematical Genomics, Department of System Biology, Columbia University, New York, NY, USA
5 Department of Biomedical Informatics, Columbia University, New York, NY, USA
6 School of Computer Science, Carnegie Mellon University, Pittsburgh, PA, USA
7 Department of Biochemistry and Molecular Pharmacology, NYU Grossman School of Medicine, New York, NY, USA
8 Department of Biomedical Engineering, NYU Tandon School of Engineering, Brooklyn, NY, USA
9 Department of Pathology, NYU Grossman School of Medicine, New York, NY, USA
10 Society of Fellows, Harvard University, Cambridge, MA, USA
[*]Co-first Authors
[#]Corresponding Authors: Jimin.Tan@nyulangone.org; Aristotelis.Tsirigos@nyulangone.org; xiabo@broadinstitute.org

Abstract

Chromatin-associated proteins (CAPs), including over 1,600 transcription factors, bind directly or indirectly to the genomic DNA to regulate gene expression and determine a myriad of cell types. Mapping their genome-wide binding and co-binding landscape is essential towards a mechanistic understanding of their functions in gene regulation and resulting cellular phenotypes. However, due to the lack of techniques that effectively scale across proteins and biological samples, their genome-wide binding profiles remain challenging to obtain, particularly in primary cells. Here we present Chromnitron, a multimodal foundation model that accurately predicts CAP binding landscapes across hundreds of proteins in unseen cell types. Via *in silico* perturbation experiments, we show that the model learned principles of CAP binding from multimodal features including DNA sequence motifs, chromatin accessibility levels, and protein functional domains. Applying Chromnitron to study cell fate transitions, we discovered novel CAPs regulating the T cell exhaustion process. Furthermore, Chromnitron can predict the dynamic CAP binding landscapes during development, revealing the global orchestration of protein and regulatory element activities in neurogenesis. We expect Chromnitron to accelerate discovery and engineering in regulatory genomics, particularly in human primary cells, and empower future therapeutic opportunities.

INTRODUCTION

How does the genome encode gene activities to determine the full spectrum of cell types and their functions in a multicellular organism? While DNA sequence carries the shared basal genetic information across all cell types, chromatin-associated proteins (CAPs) establish cell-type-specific gene regulatory programs in both normal and disease states[1–5]. In the human genome, there are an estimated 2,500 CAPs, including approximately 1,600 transcription factors (TF) and around 900 other transcriptional co-factors and chromatin regulators[6–8]. These proteins interact with over one million putative *cis*-regulatory elements (CREs), mediated by chromatin states, to confer gene regulatory programs[9,10]. Such CAP-DNA interactions can be achieved both directly and indirectly. Through DNA binding domains (DBDs), TFs directly recognize and bind to short consensus DNA sequence motifs within appropriate chromatin states[11–16]. Transcriptional co-factors and chromatin regulators typically interact with DNA indirectly via histones, other CAPs, or chromatin-associated RNAs[17–20]. Together, the specificity of CAP binding across the genome and cell types is collectively determined by DNA sequence, cell-type-specific chromatin states, and protein structural features encoded in the amino acid sequence.

Mapping the genome-wide binding landscape of all CAPs across cell types is crucial for decoding cell-type-specific gene regulatory programs and predicting cell fate transitions[21–23]. However, traditional experimental approaches for mapping CAP binding profiles, such as ChIP-seq, have limited scalability across proteins and biological samples, especially primary cells due to their heterogeneity and rarity[24,25]. As a result, only around 1,200 human CAPs have been mapped, primarily in cultured cell lines[26–29]. Compared to experimental approaches, *in silico* predictions are highly scalable and hold the promise to generalize to primary cells that are otherwise not measurable. Recent advances in machine learning have shown early promises in predicting CAP binding profiles from DNA sequence or chromatin features[30–33]. However, existing models are mostly unimodal and limited in their ability to generalize across proteins and cell types. We hypothesize that a multimodal foundation model is required to learn a comprehensive set of rules governing CAP-chromatin interactions and enable transferable prediction across unseen cell types and chromatin states.

Here we present Chromnitron (Chromatin omni-modal transformer), a biologically informed multimodal deep neural network that integrates information from DNA sequence, chromatin accessibility signals, and CAP residues to predict cell type-specific CAP binding landscapes (Fig. 1). Chromnitron was pre-trained on 1,105 CAP ChIP-seq datasets across five cell types to learn the general principles of CAP-chromatin interaction and individually fine-tuned for each CAP to further improve prediction performance. The trained model can predict the global CAP binding landscape across 767 distinct CAPs in cell types that were unseen during training. We show that Chromnitron can be used to discover critical regulators at key regulatory elements and categorize co-regulation structures across CAPs and CREs. Applying Chromnitron to study cell fate transitions, we discovered ZNF865 and ZNF766 as novel regulators of the T cell exhaustion process. Furthermore, Chromnitron can predict global CAP binding landscape and reveal the orchestration of CAPs and CRE activities during continuous cell fate transitions in developmental systems. Together, these results demonstrate the effectiveness of a multimodal foundation model in understanding the principles of CAP-chromatin interactions in gene regulatory programs and accelerating discoveries in cell fate determination.

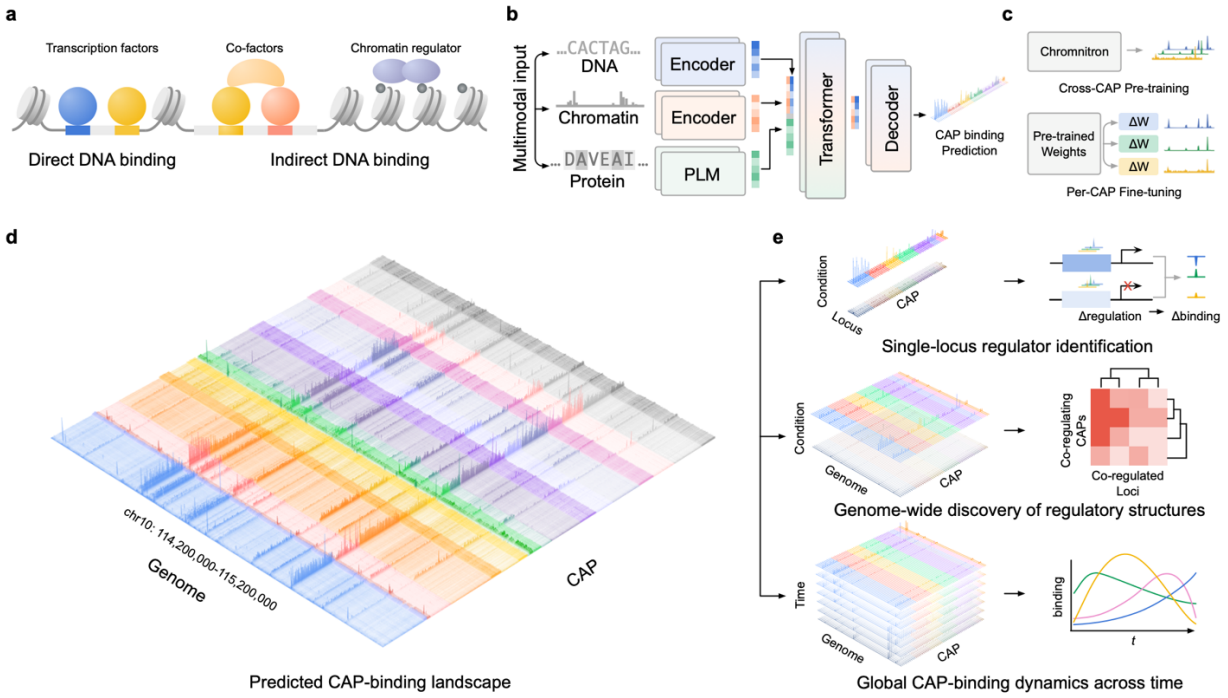


Fig. 1: Chromnitron predicts the binding landscape of chromatin-associated proteins.

a. Schematic of varied CAP types and their interaction modes with chromatin. **b.** Chromnitron model architecture. Chromnitron uses a cascaded architecture across multimodal encoding, modality fusion, transformer, and decoding modules to predict CAP binding profiles. **c.** Two-phase training strategy showing cross-CAP pre-training using shared weights (grey) followed by individual fine-tuning for each CAP with LoRA, visualized as weight deltas (ΔW). **d.** Predicted CAP-binding landscape, presented as two-dimensional heatmap of binding intensity across genomic regions for 767 CAPs (CAP axis). CAPs are colored by clusters based on binding similarity. **e.** Downstream applications of Chromnitron to accelerate discoveries. Top: discovering CRE-specific regulators at single locus. Middle: genome-wide discovery of regulatory structures across CAPs and CREs. Bottom: mapping global CAP-binding dynamics across continuous cell fate transition trajectories.

RESULTS

Multimodal learning of CAP binding principles

To create a model that learns the principles of CAP binding, we designed a multimodal architecture that integrates information from DNA sequence features, cell-type-specific chromatin states, and CAP structural features encoded in the amino acid sequences (Fig.1b and Supplementary Fig. 1). This design allows Chromnitron to learn features from all modalities, enabling accurate predictions of CAP binding that generalize across new genomic sequences and cell types. To efficiently learn and integrate input information across different modalities, Chromnitron employs a cascaded architecture including encoding, modality fusion, transformer, and decoding modules (Fig. 1b and Supplementary Fig. 1). Each input modality was first processed by its own encoder module. DNA sequence and chromatin accessibility signals were separately encoded by two convolutional modules, while CAP residues were encoded with a pre-trained protein language model[34] (Methods). These encoders yield latent embeddings that were subsequently merged in two stages. First, DNA sequence and chromatin accessibility embeddings were concatenated to form an integrated chromatin embedding aligned on the genomic axis. Second, we fused the protein embedding to the integrated chromatin embedding on the length dimension (Methods). This fusion enables the following transformer to attend every protein residue to every location in the chromatin embedding and allow implicit learning of both intra- and cross-modality attentions. In the final decoder module, the chromatin embedding was processed by a transposed convolutional module to generate predicted CAP binding intensities at single nucleotide resolution (Fig. 1b and Supplementary Fig. 1).

Training a high-performance foundation model requires large-scale and standardized data[34–37]. To minimize experimental noise across proteins and data sources, we first curated a collection of 1,105 qualified ChIP-seq from five different cell types, covering 767 unique CAPs, as well as the corresponding cell-type-specific ATAC-seq datasets (Supplementary Fig. 2a-d, Methods). Next, we developed a robust bioinformatic infrastructure, Chrom2Vec, to convert raw sequencing data (i.e. FASTQ files) to training-ready data vectors (Supplementary Fig. 2e, Methods). Chrom2Vec generates CAP binding profiles that are comparable across proteins and cell types, effectively minimizes noise-induced errors to ensure model performance and robustness[35,38,39].

Next, we trained Chromnitron in two-phases: cross-CAP pre-training and individual fine-tuning for each CAP. Compared with individualized training across each CAP from scratch, this two-phase training strategy substantially reduced the amount of total training time and allows the model to learn integrated binding principles across CAPs (Methods). However, the pre-training step can be computationally heavy due to the compounded sampling space across genomic region, cell type, and CAP dimensions (Supplementary Fig. 3a-c). To improve pre-training efficiency, we developed an algorithm, OuterMix, to reuse embeddings generated by each encoding step (Supplementary Fig. 3d-f , Methods). This approach dramatically reduced sampling complexities and accelerated data loading and encoding by five folds during pre-training. We then used low-rank adaptation (LoRA) to fine-tune individual sub-models for each CAP[40]. During pre-training we observed distinct learning dynamics across CAP groups, possibly driven by their differential binding mechanisms. For instance, performance of cohesin-related CAPs quickly improved in early stages, whereas MAF-related CAPs did not increase in performance until the middle stage of pre-training (Supplementary Fig. 4a). After pre-training, individual CAP fine-turning substantially increased prediction performance overall (Supplementary Fig. 4a-b).

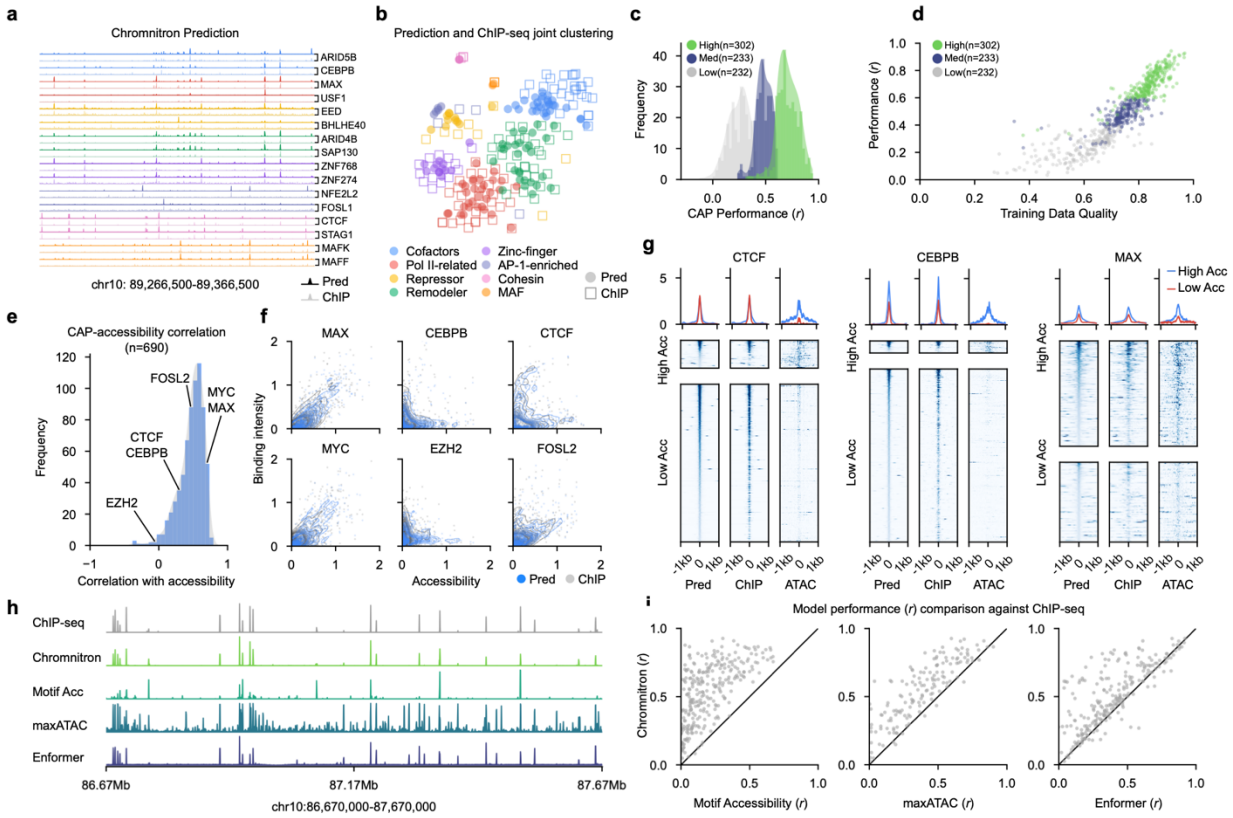


Fig. 2: High-quality prediction across chromatin contexts.

a. Paired Chromnitron prediction (dark) and ChIP-seq (light) tracks for 16 CAPs in leave-out chromosome region. **b.** Dimensionality reduction (t-SNE) visualization of predicted (circle) and experimental (squares) binding profiles from a subset of CAPs. Colors define CAPs clustered with similar binding profiles. **c-d.** CAP-binding prediction performance distribution (**c**), measured by Pearson's *r*, and their correlation with training data quality (**d**, Methods). Green, blue, and grey colors mark high- (302), medium- (233) and low- (232) performance CAP categories, respectively (Methods, and Supplementary Fig. 5-6). **e.** Distribution of correlation coefficients (Pearson's *r*) between experimental ChIP-seq signal and chromatin accessibility in HepG2 cells. **f.** Scatter plots of CAP binding versus chromatin accessibility signals. Each CAP presents both experimental ChIP-seq signal (grey) and predicted binding signal (blue). **g.** Predicted and experimental CAP binding signals in both high- and low-accessibility regions. Presented heatmaps include only 2kb genomic regions in leave-out chromosome (chr10). Acc: accessibility. **h.** Example CAP binding tracks comparing Chromnitron, motif accessibility, maxATAC and Enformer predictions, all aligned together with the experimental ChIP-seq data in leave-out chromosome 10. **i.** Performance comparison (Pearson's *r*) between Chromnitron and other methods across CAPs.

High-quality and cross-modality CAP binding predictions

Accurate prediction is the foundation for downstream biological discovery. Chromnitron was trained to predict high-quality CAP binding profiles across 767 proteins at single-nucleotide resolution and can generalize to unseen chromatin and cell types. In leave-out genomic regions

170 (chromosome 10), Chromnitron accurately predicts CAP-binding profiles comparable to
171 experimental measurements (Fig. 2a). Projecting both Chromnitron predictions and experimental
172 ChIP-seq profiles into t-distributed stochastic neighbor embedding (t-SNE) space, the predicted
173 data form clusters that align closely with their experimental counterparts, demonstrating high-
174 fidelity prediction performance (Fig. 2b). We quantified prediction performance by calculating
175 Pearson's correlation (r) between Chromnitron-predicted and experimental signals in the leave-
176 out chromosome. Across the 767-CAP catalogue, more than half of the CAPs have Pearson's
177 correlation coefficients larger than 0.5 (Fig. 2c, Methods). We found that Chromnitron's
178 predictions match or outperform the quality of experimental replicates (Supplementary Fig. 4c-e).
179 Chromnitron also achieves high-quality prediction when generalizing to new cell types and CAPs
180 via ablation studies (Supplementary Fig. 4f-g). We notice that Pearson's correlation is highly
181 sensitive to noise at background regions of the genome. Correlating with data quality, we found
182 that prediction performance of individual CAP improves consistently as training-data quality
183 increases (i.e. reduced background noise level) (Fig. 2d and Methods). Despite the modest
184 correlation coefficients in some CAPs ($r < 0.5$), Chromnitron-predicted CAP-binding peaks have
185 high correspondence to ChIP-seq peaks (Supplementary Fig. 5-6).
186
187 Most existing tools for inferring candidate CAP binding from ATAC-seq signals are more sensitive
188 to highly accessible region and less so to lowly accessible regions[41–43]. We found that
189 experimentally measured CAP binding profiles includes substantial binding events at low-
190 accessibility or nominally inaccessible regions (Fig. 2e-g). Comparing experimental CAP binding
191 profiles with chromatin accessibility, we observed a wide range of relationships, including positive
192 correlation (e.g. MAX and MYC), negative correlation (e.g. CEBPB and EZH2), and non-linear
193 patterns (e.g. CTCF and FOSL2, Fig. 2f). Since Chromnitron is not restricted by accessibility
194 signals but rather by learning the binding principles across different chromatin states, we observed
195 that Chromnitron faithfully captures different CAP binding modes as observed in experimental
196 data (Fig. 2f-g). Together, these results demonstrate that Chromnitron, with its multimodal design,
197 successfully learned the intricate relationships between CAP binding and local chromatin
198 accessibility states.
199

Chromnitron is a multimodal foundation model for predicting cell-type-specific CAP-binding profiles in unseen cell types during training. Although no other model employs the same multimodal framework, we nevertheless benchmarked Chromnitron against several approaches that aimed to infer or predict CAP binding profiles. These approaches include "motif accessibility" (MA) score, which multiplies TF motif strength by chromatin accessibility as a first-order approximation of CAP binding. Comparing with ground truth ChIP-seq data, the MA score tends to be noisy and shows weak correlation due to its frequent false positive and false negative signals (Fig. 2h, Supplementary Fig. 7). Chromnitron prediction outperforms MA score by a large margin (Fig. 2i). Next, we compared Chromnitron with two machine learning approaches, maxATAC and Enformer[31,44]. MaxATAC was trained on sparse CAP-binding peak files and predicts the probability of CAP binding, while Enformer is a DNA sequence-based model trained with multi-task learning and unable to generalize to new cell types or CAPs[31,44]. Using experimental ChIP-seq data as the reference, we found that Chromnitron substantially outperforms both methods across various metrics (Fig. 2h-i, and Supplementary Fig. 7-9).

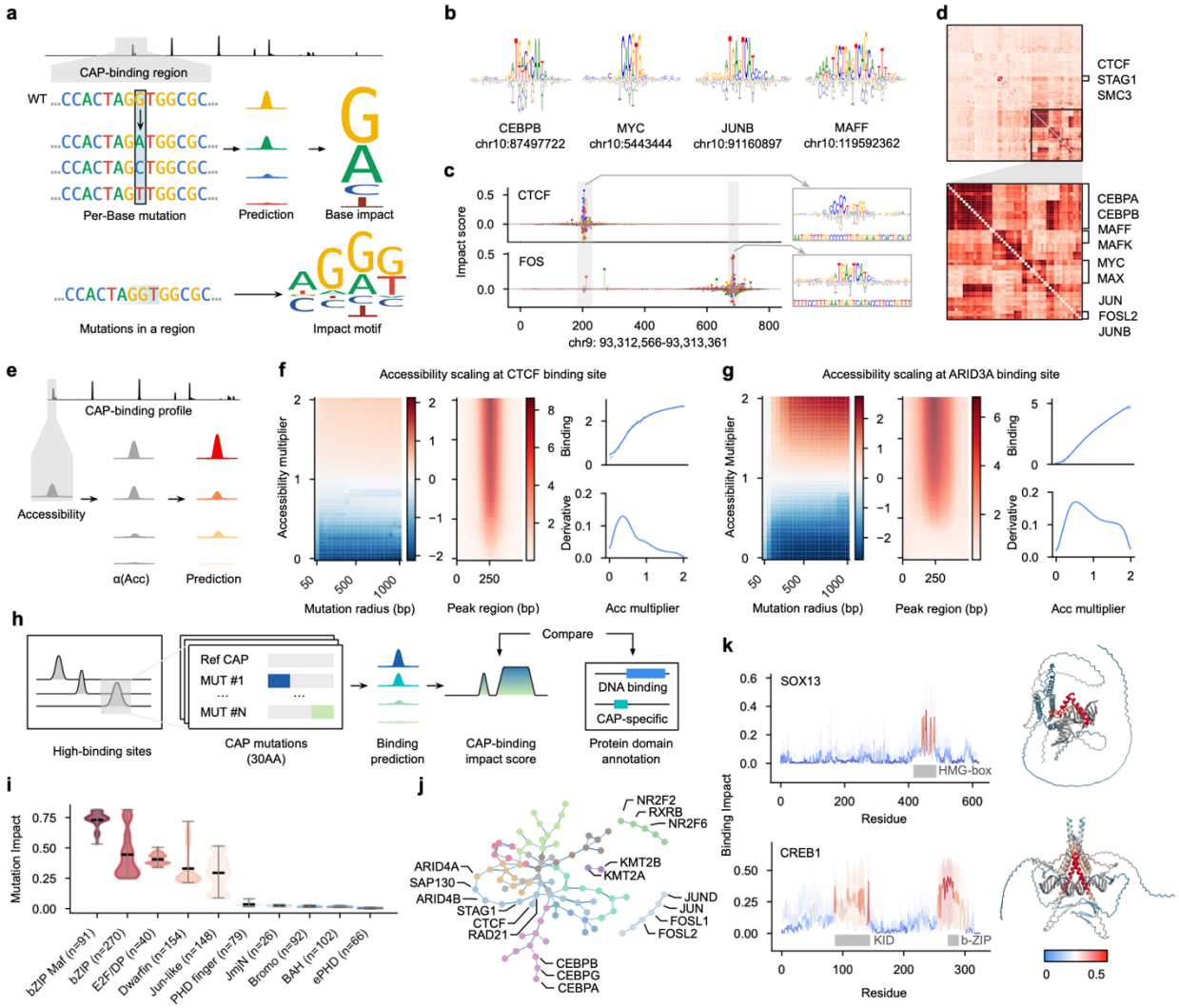


Fig. 3: *In silico* perturbation reveals principles governing CAP binding.

a. Quantification of CAP-specific sequence impact scores through base-resolution *in silico* mutations and Chromnitron prediction (Methods). **b.** Representative impact motifs for CEBPB, MYC, JUNB, and MAFF. **c.** Impact motifs generated from the same locus for two different CAPs (CTCF and FOS). **d.** Hierarchical clustering using impact motif cross-correlations of CAPs. **e.** Chromatin accessibility perturbation using *in silico* scaling multipliers (Methods). **f-g.** Accessibility scaling results at CTCF (**f**) and ARID3A (**g**) binding sites. Left: predicted CAP binding changes responding to perturbation radius and accessibility multiplier (Methods). Middle: heatmap of predicted binding intensity upon perturbing accessibility at fixed 500bp window, centered at binding peaks. Right: CAP binding changes (top) and the rate of change (bottom) as a function of accessibility multiplier. **h.** Schematic of *in silico* mutation for CAP protein sequences. Upon mutation, predicted CAP binding intensities of mutant CAPs were compared to reference CAP to calculate protein residue impact scores (Methods). **i.** Violin plot distribution of impact scores across CAP residues, stratified by domain annotations. **j.** Protein–protein network inferred from predicted CAP co-localization patterns, colored by groups. **k.** Per-residue binding impact

231 scores for SOX13 and CREB1, with AlphaFold-predicted structures highlighting high-impact
232 residues (HMG-box, KID, b-ZIP domains) which bind proximately to DNA.
233

234 **Multimodal *in silico* perturbations reveal principles of CAP binding**

235 The multimodal design allows Chromnitron to learn CAP binding principles from different
236 modalities and improve prediction performance. To investigate whether the learned principles
237 conform to known biological mechanisms, we performed *in silico* perturbations to interrogate the
238 model across all three input modalities, starting with *in silico* DNA sequence mutagenesis. For a
239 locus with CAP binding, we sequentially mutated every nucleotide to each of the three other
240 nucleotides, followed by quantification of the mutation impact by computing the resulting changes
241 in predicted binding (Fig. 3a). We found that these single-locus, per-nucleotide impact scores
242 recapitulate TF consensus motifs that are typically derived from hundreds to thousands of
243 experimental binding sequences[45,46] (Fig. 3b). Applying *in silico* mutagenesis across all CAPs at a
244 given locus, Chromnitron can identify multiple, spatially distinct impact motifs within the locus,
245 reflecting the precise sequence-dependent CAP binding syntax (Fig. 3c).
246

247 Chromnitron's sensitivity to individual nucleotide alternations enables systematic discovery of
248 uncharacterized CAP-DNA interactions. By collecting impact motifs across multiple loci for each
249 CAP and clustering them by similarity, we found several modules corresponding to distinct CAP
250 families, such as the C/EBP, bHLH, and AP-1 families, and a large set of unique impact motifs,
251 many uncharacterized (Fig. 3d, and Supplementary Fig. 10, Methods). Moreover, *in silico*
252 mutagenesis with Chromnitron identified impact motifs for proteins that lack canonical DNA
253 binding domains, such as cohesin subunits resembling CTCF motif (Fig. 3d and Supplementary
254 Fig. 10). We found these motifs often resemble the sequence features of their frequent co-binding
255 partners, suggesting a promising strategy for discovering co-binding partners by comparing impact
256 motif matrices (Fig. 3d and Supplementary Fig. 10). Last, Chromnitron is sensitive to non-coding
257 variants and captures their effect on CAP binding in both gain-of-binding and loss-of-function
258 mutation events (Supplementary Fig. 10).
259

260 Chromatin accessibility differences represent different epigenetic states across genomic regions
261 and cell types. Therefore, we evaluated the effect of perturbing chromatin accessibility levels (Fig.
262 3e). By scaling accessibility intensity *in silico* across different window sizes, we found that binding

263 prediction is indeed sensitive to accessibility changes, and in a CAP-dependent manner. For
264 example, the relationship between CTCF binding and accessibility level follows a logarithmic
265 curve, saturating at relatively low accessibility levels (Fig. 3f), whereas ARID3A binding
266 intensifies linearly with increasing accessibility (Fig. 3g). These results confirmed that
267 Chromnitron learned both the non-linearity and specificity in CAP binding, echoing our previous
268 results correlating ChIP-seq to chromatin accessibility (Fig. 2e-g).
269
270 Last, we carried out *in silico* mutagenesis on CAP amino acid sequences to assess whether
271 Chromnitron learned informative features from protein residues. We randomly mutated amino
272 acids in segments with a sliding window approach and quantified the impact of mutant residues
273 by calculating the resulting changes in predicted CAP binding (Fig. 3h, Methods). Scanning
274 through the entire protein sequence, we observed strong enrichment in evolutionary conserved
275 DNA-binding domains such as bZIP and E2F (Fig. 3i-j, Supplementary Fig. 11). For examples,
276 we observed high impact scores within the direct DNA binding regions for both SOX13 and
277 CREB1, which were confirmed by both domain annotation and AlphaFold-predicted protein-DNA
278 binding structures (Fig. 3k). In contrast, functional domains that do not contribute to sequence
279 specificity of CAP binding, such as bromodomain and PHD finger domain, tends to have much
280 lower impact scores (Fig. 3k, Supplementary Fig. 11). Collectively, these *in silico* perturbation
281 results demonstrate that Chromnitron effectively integrates multimodal information to uncover
282 both canonical and hidden principles of CAP binding.
283

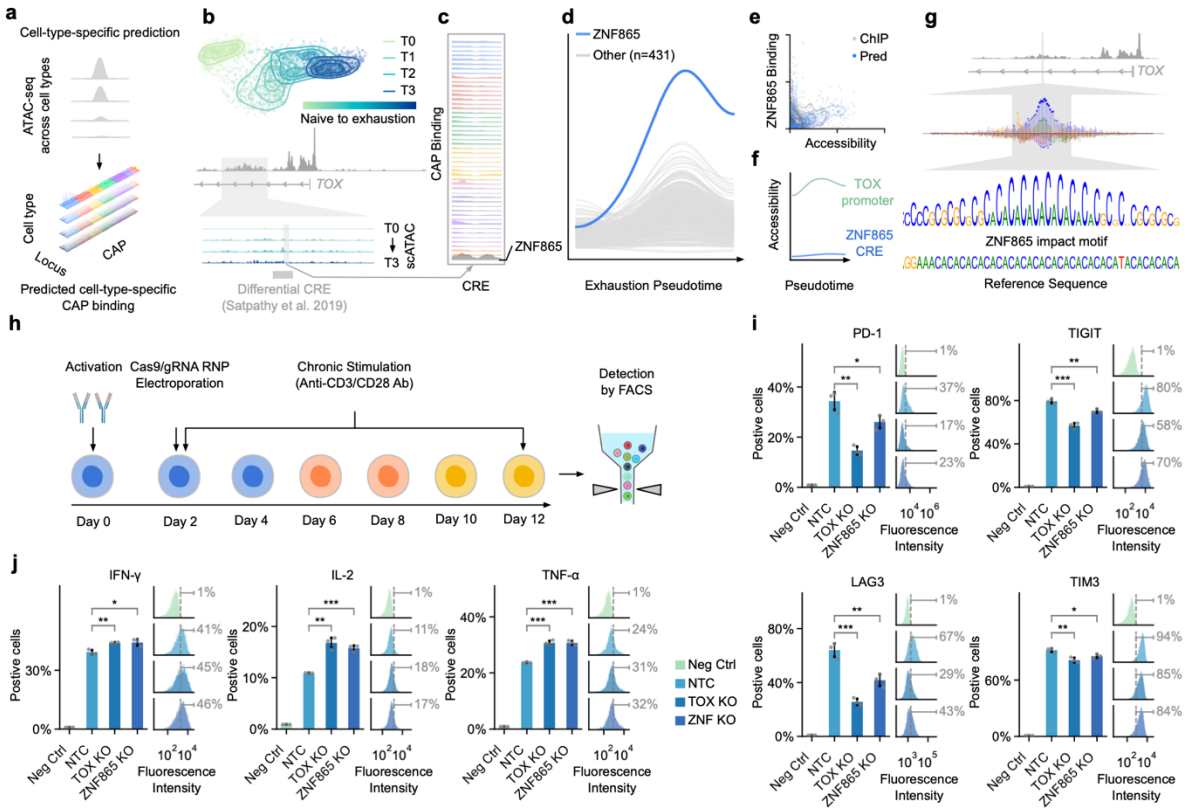


Fig. 4: Predicting novel regulatory CAPs at individual regulatory element.

a. Predicting cell-type-specific CAP binding using ATAC-seq datasets across conditions. **b.** Dimensionality reduction (UMAP) showing T cell fate trajectory from naïve (T0) to exhausted (T3) states (top), together with chromatin accessibility changes at a *TOX* promoter-proximal regulatory element. **c.** Predicted CAP binding signals at this *TOX* promoter-proximal CRE, highlighting ZNF865 binding. **d.** ZNF865 binding stood out among other CAPs at the *TOX* CRE during late pseudo-time stage. **e.** Scatter plot of ZNF865 experimental ChIP-seq binding (grey) and predicted binding (blue) versus chromatin accessibility, indicating that ZNF865 preferentially binds at low-accessibility regions. **f.** Chromatin accessibility signals at *TOX* promoter and the ZNF865-bound CRE across pseudo-time. **g.** Chromatin accessibility at the ZNF865-bound CRE (top) and the corresponding ZNF865 impact motif (bottom). **h.** Schematic of *in vitro* model of T cell exhaustion through persistent anti-CD3/CD28 stimulation, coupled with CRISPR/Cas9-based gene knockout in primary T cells. **i-j.** Flow cytometry quantification of T cell exhaustion markers (**i**, PD1, TIGIT, LAG3 and TIM3) and cytokines (**j,** IFN-γ, IL-2, and TNF-α) across negative control, non-targeting control (NTC), *TOX* knockout, or *ZNF865* knockout conditions. Bar plots show mean ± s.d. for n = 3 independent replicates. Statistical comparisons performed using two-sided unpaired Student's t-tests with significance denoted as *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

Discovering CRE-specific regulators at single locus

305 The collection of CAPs binding at a cis-regulatory element (CRE) determines its regulatory
306 function. Given the chromatin accessibility signals at a genomic locus, Chromnitron predicts high-
307 quality CAP-binding profiles that enables the discovery of key regulators (Fig. 4a). We applied
308 Chromnitron to study T cell fate transition during chronic stimulation. Persistent antigen exposure
309 pushes tumor-infiltrated T cells into a dysfunctional or exhausted state characterized by enhanced
310 expression of inhibitory cell surface receptors, reduced cytokine production, and activation of
311 *TOX*[47–54]. Single-cell ATAC-seq (scATAC-seq) of tumor-infiltrated T cells has captured the
312 transition from naïve T cell to exhausted T cells[55]. To pinpoint the locus-specific regulators that
313 control *TOX* expression, we applied Chromnitron to track CAP binding dynamics at the *TOX* locus
314 across the T cell exhaustion trajectory derived from scATAC-seq data (Fig. 4b, Methods).
315 Surprisingly, ZNF865, an uncharacterized C2H2 zinc-finger protein, stands out as a unique
316 regulator at a *TOX* promoter-proximal regulatory element with a steady increase of binding along
317 the exhaustion pseudo-time (Fig. 4c-d). We noticed that ZNF865 preferentially binds at relatively
318 inaccessible regions, including this *TOX* promoter-proximal element, whose accessibility is
319 substantially lower than that of the promoter region (Fig. 4e-f). Further *in silico* mutagenesis
320 revealed that ZNF865 recognizes a previously uncharacterized CA-repeat binding pattern, which
321 explains in part of its unique binding to the *TOX* locus (Fig. 4g). These results underscore
322 Chromnitron's ability to leverage fine-grained DNA sequence features and nuanced changes in
323 accessibility for discovering novel regulators of gene expression.
324
325 To validate this finding, we employed an established *in vitro* T cell exhaustion model using
326 persistent anti-CD3 antibody or anti-CD3/CD28 cocktail stimulation[56–58] (Methods). After
327 persistent stimulation, human T cells were analyzed by flow cytometry for cell surface exhaustion
328 markers and cytokine production (Methods). As expected, exhaustion markers including PD1,
329 TIGIT, LAG3, and TIM3 were significantly upregulated in the persistently stimulated group, while
330 IFN-γ, IL-2, and TNF-α expression level were significantly reduced, confirming successful
331 induction of T cell exhaustion phenotype *in vitro* (Supplementary Fig. 12). Consistently, both
332 *ZNF865* and *TOX* expression increase along the progression of persistent T cells stimulation
333 (Supplementary Fig. 13a-b). To evaluate the function of ZNF865, we electroporated Cas9
334 ribonucleoprotein with sgRNAs targeting *ZNF865*, *TOX* (positive control), and a non-targeting
335 control (NTC) before persistent anti-CD3/CD28 stimulation of the T cells (Methods). Consistent

with the hypothesis, we found that knockout *ZNF865* reduced *TOX* expression level (Supplementary Fig. 13c-d). Notably, *ZNF865* knockout led to a significant reduction in surface exhaustion markers compared to NTC, closely resembling the effects observed in the *TOX* knockout group (Fig. 4i). Moreover, ZNF865-deficient T cells produced elevated cytokine production, suggesting restoration of effector function (Fig. 4j).

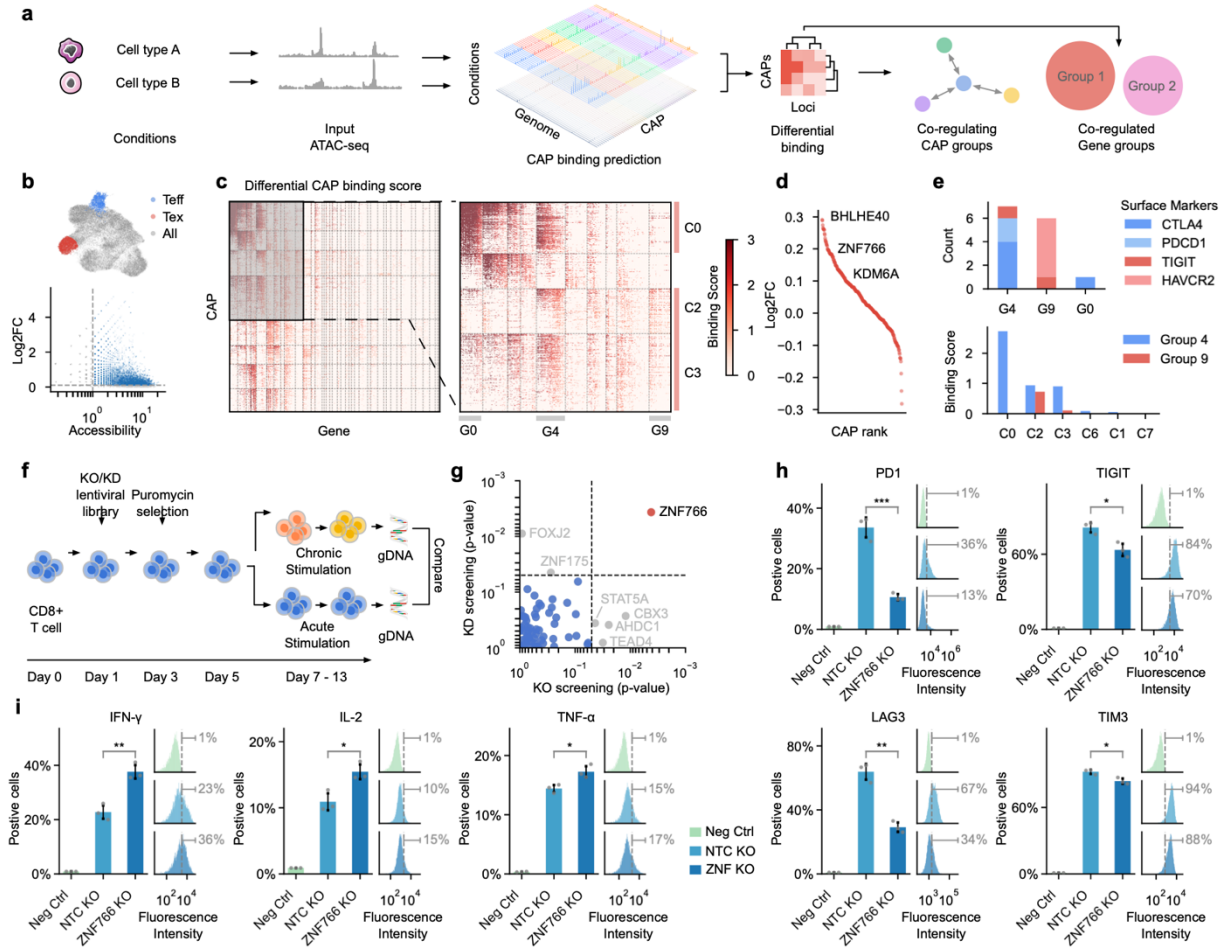


Fig. 5: Genome-wide characterization and discovery of novel regulatory CAPs driving T cell exhaustion.

a. Workflow of genome-wide characterization of co-regulation structures across CAPs and CREs between paired cell conditions. **b.** The distribution of effector T cells (Teff, blue) and exhausted T cells (Tex, red) in UMAP (top) and the chromatin accessibility differences between the two cell population (bottom, Methods). **c.** Predicted CAP binding difference matrix between Teff and Tex cells, clustered by both co-regulating CAPs (rows) and co-regulated gene (promoter) groups (columns). A zoom-in plot highlights a subset of CAP modules (C0, C2, C3) and gene groups (G0, G4, G9) enriching reported Tex marker genes (Methods). **d.** Elbow plot of enriched CAPs ranked

352 by log2 fold change of Tex over Teff. **e.** Distribution of reported Tex marker genes (CTLA4,
353 PDCD1, TIGIT, HAVCR2) in co-regulated CRE groups (top) and their corresponding co-
354 regulating CAP clusters (bottom). **f.** CRISPR dropout screens in human $CD8^+$ T cells using *in vitro*
355 chronic stimulation model (Methods). **g.** Scatter plot of targeting-gene sgRNA enrichment p-
356 values in the knockout (x-axis) and knockdown (y-axis) CRISPR screens, highlighting *ZNF766* as
357 the top hit in both conditions. **h-i.** Flow cytometry quantification of T cell exhaustion markers (**i**,
358 PD1, TIGIT, LAG3 and TIM3) and cytokine secretion levels (**j,** IFN-γ, IL-2, and TNF-α) across
359 negative control, non-targeting control (NTC), or *ZNF766* knockout conditions. Bar plots show
360 mean ± s.d. for n = 3 independent replicates. Statistical comparisons performed using two-sided
361 unpaired Student's t-tests with significance denoted as *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.
362

363

364 **Genome-wide discovery of CAP-dependent regulatory structures**

365 Having demonstrated Chromnitron's ability to discover candidate regulators at a single regulatory
366 element of interest, we next asked whether it could generalize these insights to uncover genome-
367 wide regulatory structures. Cell fate transitions are regulated by groups of CAPs interacting with
368 CREs across the genome. Given chromatin accessibility datasets of differential cell states,
369 Chromnitron can predict their global CAP binding changes to identify co-regulating CAPs and co-
370 regulated CREs (Fig. 5a). We applied this *in silico* framework to dissect the genome-wide
371 regulatory structures driving T cell exhaustion, focusing on the group of co-regulated genes
372 corresponding to differentially accessible gene promoters between effector T cells and exhausted
373 T cells (Fig. 5b). Using Chromnitron, we predicted CAP binding profiles across these promoters
374 in both groups of T cell population (Methods). Comparing the CAP binding signals between the
375 two populations, we calculated the differential CAP binding matrix (CRE-by-CAP) and clustered
376 on both axes to uncover modules of co-regulating CAPs and co-regulated CREs (Fig. 5c). The top
377 co-regulating CAP module (C0) includes both previously identified CAPs such as BHLHE40 and
378 KDM6A[58,59], and uncharacterized CAPs such as ZNF766 (Fig. 5d). Mapping known exhaustion
379 surface markers onto our co-regulated CRE modules revealed strong enrichment in group G4, G9
380 and G0. Corresponding to these CRE groups, we found that co-regulating CAP clusters C0, C2,
381 and C3 exhibit the strongest differential binding, suggesting potential regulatory relationship (Fig.
382 5e).
383

384 We hypothesized that these co-regulating CAPs in the above modules include uncharacterized
385 upstream regulators of T cell exhaustion progress. To assess the functions of these candidate CAPs,

386 we conducted a small-scale CRISPR-based pooled screen using the *in vitro* T cell exhaustion
387 model. We designed two separates sgRNA libraries comprising 370 sgRNAs targeting 65 genes,
388 respectively optimized for gene knock-out and transcriptional knock-down via lentiviral
389 transduction (Methods). Following sgRNA delivery, human primary $CD8^+$ T cells were subjected
390 to persistent anti-CD3 treatment, with T cell proliferation serving as the phenotypic readout (Fig.
391 5f, Methods). Across both knock-out and knock-down screens, ZNF766 consistently stood out as
392 the most prominent regulator, exhibiting the strongest impact on T cell persistence during chronic
393 stimulation (Fig. 5g). To independently validate this result, we performed targeted gene knockout
394 of ZNF766 via RNP electroporation (Supplementary Fig. 14, Methods). Subsequent flow
395 cytometry analyses revealed significant decreases in surface exhaustion markers, along with
396 increases in effector cytokine production (Fig. 5h-i). RNA-seq following *ZNF766* knockout
397 showed gene expression changes that reflect a shift toward reduced exhaustion (Supplementary
398 Fig. 14). Together, these findings suggest that ZNF766 may play an important role in regulating T
399 cell exhaustion.
400

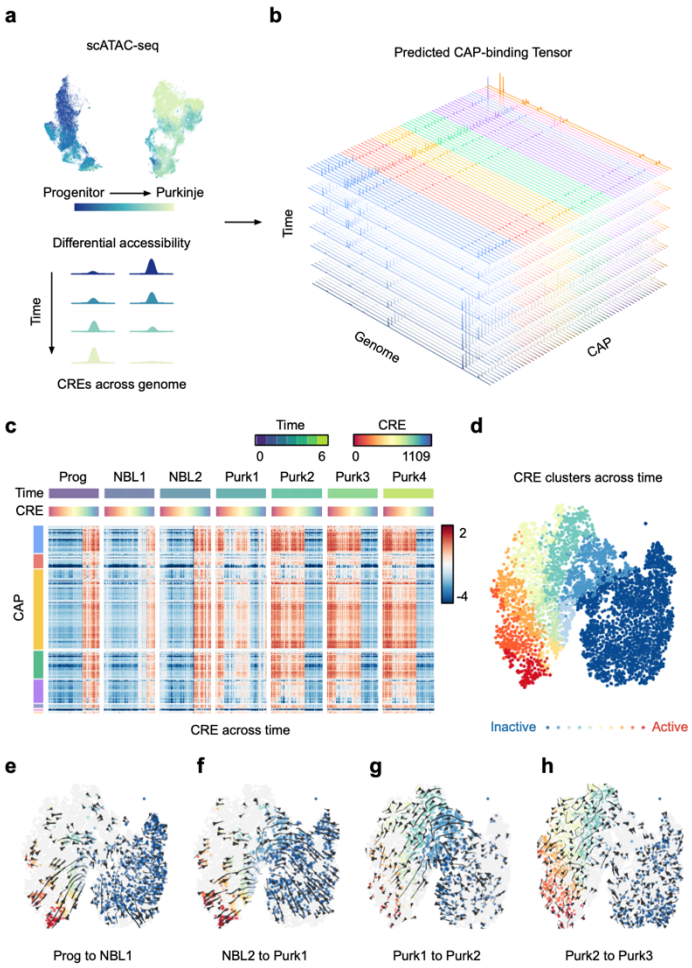


Fig. 6: Global dynamics of CAP binding and CRE activities during cell differentiation.

a. Single cell ATAC-seq is used to compute pseudo-time and identify differentially accessible CREs across the genome for Chromnitron inference. **b.** Three-dimensional tensor representing predicted CAP binding landscapes across developmental time. **c.** Heatmap of predicted CAP binding signals across CREs grouped by developmental stages. Rows and columns represent CAPs and CREs activities, respectively. **d.** T-SNE projection of CREs colored by activity clusters across pseudo-time. Clusters were defined by CAP binding profiles and sorted from inactive (blue) to active (red) based on the average binding activity. **e-h**. Vector fields of directional CRE activity transition to their following developmental stage, projected on the t-SNE map in **d**.

Global orchestration of CAP binding and CRE activities

Cell fate transition typically goes through drastic state changes to achieve functional maturity in developmental process. To characterize the shifting gene regulatory programs, we applied Chromnitron to predict genome-wide CAP-binding landscapes along cell fate transition trajectory using single cell chromatin accessibility data (Fig. 6a-b). The predicted three-dimensional CAP-

binding tensor across genomic position, CAP, and time offers unprecedented analytical scales to characterize how CAP binding and CRE activities are globally orchestrated (Fig. 6b). We applied this analytical framework to Purkinje neuron maturation during first trimester neurodevelopment[60]. After generating the CAP binding landscapes across Purkinje neuron maturation, we first transformed the three-dimensional data tensor into a CAP-by-CRE matrix with sub-grouping by time (Fig. 6c, Methods). We observed a clear transition of CRE activities, measured by global CAP occupancy, from neural progenitor cells to mature Purkinje cells, with a sharp transition between neuroblast to early Purkinje state (Fig. 6c).

To dissect stage-specific transitions in CRE activities along the developmental time trajectory, we embedded all CREs in a t-SNE map using their CAP-binding features. Subsequent clustering of all CREs revealed the global distribution of CRE identities ranging from inactive to active states across genomic space and temporal stages (Fig. 6d, Supplementary Fig. 15a-d, Methods). Next, we tracked each CRE from its current developmental stage to the following stage, computing a vector field that visualizes its trajectory (Methods). The resulting vector field uncovers the migration snapshots of CRE identities (Fig. 6e-h). From neural progenitor (Prog) to neuroblast state 1 (NBL1), stem-cell-specific active CREs start to drift toward the inactive region (Fig. 6e). From neuroblast state 2 (NBL2) to early Purkinje cells (Purk1), most inactive CREs start to shift towards activation, representing the strongest wave of CAP-driven CRE activity changes, while those already active CREs continue to migrate to the inactive state (Fig. 6f). Purk1 to Purk2 interval marks the final major wave of CRE activity shift, where the last set of inactive CREs complete their transition (Fig. 6g). Starting from Purk2, all CREs converge to a relatively stable configuration to form the terminal regulatory landscape (Fig. 6h).

The three-dimensional tensor can be interrogated along any axis to offer new insights about CAP binding and CRE activities. Using developmental time as the organizing axis, we can track the binding trend for each CAP on every CRE. This analysis allows us to uncover cohorts of early- and late-acting regulators during Purkinje cell maturation trajectory, most of which follow monotonic gain- or loss-of-binding trends (Supplementary Fig. 15e). In parallel, the tensor can be transformed using the CRE dimension as the feature, followed by re-clustering CAPs. We found distinct CAP binding signatures across developmental time. For example, unlike transcription

activators, cohesin-related protein, chromatin remodelers, and repressors assume relatively stable, factor-specific binding patterns (Supplementary Fig. 15f-g). These CAP bindings tend to have minor changes within their own group and rarely shift between groups, consistent with the regulatory constraints imposed by their associated functions on chromatin (Supplementary Fig. 15g). Together, these multidimensional analyses illustrate how the CAP-binding tensor offers an unprecedented lens on the dynamic orchestration of CAP binding and CRE activities, setting the stage for systematic discovery of gene regulation programs that controls cell fate transition, and ultimately, guiding precision rewiring of cell fate.

DISCUSSION

A major challenge in understanding the molecular mechanisms of human health and disease is to systematically characterize how chromatin-associated proteins configure gene regulatory programs. Here we show that Chromnitron, a biologically informed multimodal foundation model, can learn the core principles of CAP-chromatin interactions and achieve accurate prediction of global CAP-binding landscapes across hundreds of proteins. The multimodal and cell-type-specific design allows the model to generalize to virtually any cell type, enabling systematic discovery of core transcriptional regulators and their downstream impacts on cellular phenotypes. Chromnitron acts as a prism to transform single chromatin accessibility signal into a spectrum of functional genomics data with high-fidelity. By harnessing the power of such high-throughput prediction, we discovered previously uncharacterized regulators of T-cell exhaustion and mapped the global dynamics of CAP binding and CRE activities during neurodevelopment with unprecedented scale. Chromnitron exemplifies how biologically grounded genomics foundation models can accelerate hypothesis generation and enable targeted experimental validation, revealing new biological insights.

Beyond its practical applications, Chromnitron also uncovers both known CAP-chromatin binding rules and non-canonical interactions at low-accessibility or absence-of-accessibility regions with previously unrecognized sequence motifs. Compared with sequence-based models that excel at pinpointing nucleotide-level attribution of regulatory activity, Chromnitron learns the fundamental rules of CAP binding through a multimodal architecture, enabling generalizable high-performance prediction across cell types and chromatin states. Moreover, Chromnitron's multimodal design

allows the integration of pre-trained DNA language models as alternative or additional encoders to further increase training efficiency and performance[32–34,36,43,44,61].

Despite its generalizability and high prediction performance, Chromnitron nonetheless has limitations. As a machine learning model, Chromnitron's prediction accuracy depends on the training data quality. Current training datasets are derived primarily from conventional ChIP-seq experiments, which are subject to variable antibody quality, lack of dosage considerations, and batch effects. Although substantial computational efforts were made to denoise and standardize the data, not all processed ChIP-seq reach high quality. These challenges underscore the need for generating standardized, high-quality CAP-binding datasets. Importantly, the pre-trained foundation model can be readily updated by fine-tuning with improved CAP-binding datasets. Generating standardized CAP-binding datasets covering the remaining uncharacterized human CAPs will further improve Chromnitron's prediction performance and scalability across diverse biological systems.

Biologically informed multimodal foundation models like Chromnitron offer a new paradigm in human biology research. By accurately predicting large-scale functional data from information-rich measurement, these models can enable high-throughput *in silico* screening to aid hypotheses generation and validation, therefore accelerating biological discovery and new therapeutic development.

500 **Methods:**
501
502 **Model architecture**
503 We designed Chromnitron (Chromatin Omni-modal Transformer) to integrate three information
504 streams that jointly determine CAP-chromatin binding: DNA sequence, local chromatin
505 accessibility, and the amino-acid sequence of the assayed CAP. Chromnitron comprises three
506 successive components: modality-specific encoders, a multimodal self-attention module and a
507 decoder (Supplementary Fig. 1).
508
509 The model has three modality-specific encoders. The encoders for DNA sequence and chromatin
510 accessibility consists of similar convolutional blocks (ConvBlocks). Each block begins with a
511 strided Conv1D layer that covers large local windows while halving the sequence length. The
512 convolution is followed by group normalization (groups = 1) and a Gaussian Error Linear Unit
513 (GELU) activation function[62]. Feature maps then traverse 3 residual ConvBlocks (kernel = 9;
514 stride = 1) that maintain length while expanding the hidden dimensionality to 384. Outputs from
515 the DNA and accessibility encoders were projected to the same embedding of 512 × 384. The two
516 embeddings were fused on the feature dimension to form a joint embedding of size 512 × 768. The
517 joint embedding was scaled down to a 512 × 384 chromatin embedding with a header for
518 subsequent fusion with the protein embedding.
519
520 For the protein encoder, we leveraged a pre-trained protein language model ESM2 as a protein
521 feature extractor[34]. Running ESM2 inference during training can add to the computational
522 overhead. To increase model training efficiency, we extracted and cached the protein features
523 before the training step. A non-redundant list of human transcription factors was assembled by
524 merging all proteins for which we had ChIP-seq data and the available proteins in UniProt database.
525 To encode the embeddings, we used the publicly available ESM2 transformer with 36 layers and
526 3 billion parameters (esm2_t36_3B_UR50D, hidden size = 2560). For each protein, we requested
527 the hidden representations of the final, 36th transformer block. Since ESM2 accepts up to 1,024
528 tokens, we split any long proteins into non-overlapping 1,024-residue segments, each embedded
529 independently. The resulting matrices were concatenated along the sequence dimension to
530 reconstitute the full-length protein representation. The number of residues was capped at 4,096 to
531 accommodate most of the full-length CAP sequences while keeping a low compute cost. The per-
532 residue embeddings were converted to NumPy arrays and stored in a compressed npz file. These
533 cached embeddings are loaded at run-time, down-sampled by the Chromnitron protein header, and
534 fused with genomic features in the multimodal attention module.
535
536 We implemented the attention mechanism in transformer to fuse contextual information both
537 within and across modalities[63]. Chromatin feature embeddings and protein embeddings were
538 concatenated along the length dimension, separated by a single zero vector that marks the modality
539 boundary. Positional information was added via fixed sinusoidal encodings. The resulting fused
540 feature embedding has a length of 769 (512 from chromatin embedding, 256 from protein
541 embedding, and a separator with a length of 1). The fused embedding was processed by 16 layers

of pre-layer-normalized transformer. Each transformer layer has 8 attention heads, 384 hidden state dimension, 2,048 feed-forward layer dimension, and a dropout rate of 0.1. Self-attention allows every genomic position to attend not only to distant nucleotides but also cross-modal to any residue of the corresponding CAP.

After transformer module, we only kept features corresponding to genomic positions and forwarded them to the decoder. The decoder has four ConvTranspose blocks successively up-sample the latent representation to the original 8,192-bp resolution while preserving contextual information through residual connections. A final convolution header produces a one-dimensional per-base binding profile for each target CAP.

Pre-trained Chromnitron model has 72 million parameters. Each LoRA fine-tuned sub-model weight 9 million parameters. The fully fine-tuned Chromnitron model with pre-trained and LoRA weights has over 6.9 billion parameters in total.

Genomic data collection

To obtain training data with more standardized quality, we primarily focused on consortium-generated datasets for consistency. We developed a systematic approach to collect and curate high-quality CAPs ChIP-seq datasets from the ENCODE database[9]. Our data collection strategy focused on obtaining non-revoked and non-archived datasets with sufficient sequencing depth to ensure reliable modeling. We required a minimum of 40 million reads to achieve adequate genome coverage for CAPs ChIP-seq and ATAC-seq, while immunoprecipitation control experiments required 100,000 reads. Chromatin accessibility information is essential for understanding the regulatory context of transcription factor binding. To ensure experimental coherence and enable integrative analysis, we retained only those CAPs ChIP-seq datasets that had good-quality ATAC-seq data in the corresponding cell type, resulting to 1,105 qualified ChIP-seq from five different cell types, covering 767 unique CAPs.

Genomic data preprocessing with Chrom2Vec

To prepare precise and highly consistent training data, we designed Chrom2Vec, a modular bioinformatic platform that streamlines training data processing for Chromnitron. Chrom2Vec platform starts from raw FASTQ files and delivers training-ready data tensors. Chrom2Vec is based on Apache Airflow and orchestrates next-generation-sequencing processing for both ChIP-seq and ATAC-seq. It implements configurable pipelines that handle distributed workers, tracking, and reproducible containers, performs robust assay-specific normalization, and signal compression to optimize data storage. Pipelines are described by lightweight Directed Acyclic Graphs (DAGs) that call a library of preprocessing operators, which are Airflow wrappers around containerized shell or Python scripts to ensure maximum reproducibility. While tasks are being scheduled and executed by Airflow, the metadata are streamed to Neo4j for full data-lineage tracking. To reduce storage and I/O overheads without sacrificing data quality on key genomic regions, we implemented an adaptive dynamic-bin compression algorithm. We observed as high as 50% reduction in storage for some of the high-quality tracks. The final compressed training

584 dataset is around 952 GB in size. This architecture was optimized for machine-learning workflow
585 as it guarantees reproducibility, supports scaling, and writes outputs in chunked Zarr format which
586 is directly ingestible by the model and supports fast random-access during training.
587
588 To enable quantitative comparisons across cell types and experimental conditions, we performed
589 peak-based quantile normalization on bulk ATAC-seq, ChIP-seq and pseudo-bulk single cell
590 ATAC-seq data. We followed a two-step approach that first removed noise and then rescaled data.
591 The noise level can be represented by variations in the background, non-active genomic regions.
592 To first define active regions, we used cPeaks from Meng et al., a union of accessible peaks across
593 human cell types[64]. The background region is defined as its complement regions on the genome,
594 excluding ENCODE blacklist[65].
595
596 For ChIP-seq, we performed a two-stage correction that explicitly models both biochemical
597 background and sequencing bias. First, chromatin signal was extracted from immunoprecipitated
598 (IP) and matched input libraries. Using the pre-defined background regions, we calculated the
599 mean IP and input signal in the background bins. The input was rescaled to reach similar
600 background signal intensity as IP. The rescaled input was then subtracted from the IP track,
601 effectively removing experimental bias across the genome. To harmonize dynamic range across
602 ChIP-seq experiments, the corrected IP signal was divided by the average of the 99 to 99.9
603 percentile values with 1 percent step sizes, computed over the pre-defined active peak regions from
604 cPeaks. This normalization puts the strongest peaks near a value of one while retaining linearity at
605 regions with lower signal intensities. The result is a bias-corrected, unit-less collection of ChIP-
606 seq datasets that are comparable across CAPs and experiments, dramatically facilitating training
607 stability. After scaling the training datasets, we further measured a background-noise level for each
608 dataset by calculating standard deviation in the background chromatin regions defined above. The
609 training data quality metric was calculated by (1 – noise) for downstream model performance
610 evaluation.
611
612 For ATAC-seq, where a control library was not available, we removed noise from background
613 directly followed by percentile scaling. The algorithm first estimates systematic noise by averaging
614 the coverage signal over the same background regions. This background mean was subtracted
615 from every genomic position, eliminating basal accessibility level. We collected mean cPeak
616 signals, and computed the scaling constant based on top 99 to 99.9 percentiles similar to ChIP-seq.
617 Dividing the background-corrected track by this constant compresses the dynamic range so that
618 the signals of typical accessible loci fall between 0 and 1 (though exceptionally strong regulatory
619 regions can exceed 1), producing normalized ATAC-seq datasets that are consistent and
620 comparable across batches and cell types.
621
622 For single cell ATAC-seq data, we started with fragment files and processed them through a
623 pipeline with shared preprocessing steps as Chrom2Vec and generates quantitative chromatin
624 accessibility profiles. Fragment coordinates were standardized to the hg38 reference genome. Cells
625 with 1,000-50,000 fragments per cell were retained to exclude low-quality cells with insufficient

chromatin accessibility signal and potential doublets with artificially inflated fragment counts. Individual cells were subsequently aggregated into pseudo-bulk profiles based on cell cluster annotations (i.e. cell types or states) in original publication, ensuring each pseudo-bulk has over 10 million unique sequencing reads (or 20 million Tn5 transposase insertions). These accessibility reads from all cells within each cluster were summed to generate high-resolution chromatin accessibility tracks with improved signal-to-noise ratios compared to individual cells. Each pseudo-bulk data represents a unique cell type or state, and was normalized by the same pipeline described in the data normalization section.

Model pre-training and fine-tuning

Training data were assembled from the three input feature modalities and the target CAP binding profiles. Each training sample unit was set to an 8,192-bp genomic region. For the two chromatin input features, we extracted a five-channel (A, C, G, T, N) one-hot representation of the hg38 DNA sequence and a single-channel chromatin-accessibility track (ATAC-seq Tn5-transposase insertion frequency signal). To stabilize training, the ATAC-seq signal was transformed with a monotonic log1p function. CAP binding profiles were retrieved from the normalized ChIP-seq coverage for each CAP in every cell type. In a separate preprocessing step, we embedded the primary amino-acid sequence of every CAP with the 36-layer ESM2 protein language model (esm2_t36_3B_UR50D). This embedding was cached and loaded on to the protein header during training.

Random samples from different modalities are recombined to form the final training batch. Batched tensors were shaped as (batch, length, channels) for DNA and ATAC, and (batch, residues, channels) for ESM2 embeddings. The model takes in these inputs and predicts the target CAP binding profile of shape (batch, length). Validation loss was calculated as the mean-squared error (MSE) between predicted and observed log1p-signal. We used Adam[66] as the optimizer with a learning rate of 0.00002 and weight decay of 0. The scheduler used ReduceLROnPlateau strategy with a patience of 5 epochs. After validation loss stabilizes at 390 epochs, we fine-tuned the model for another 10 epochs for each CAP with low-rank adaptation (LoRA), updating the CAP-specific sub-model parameters[67]. LoRA layers were directly applied to feedforward layers in the attention. For convolution and convolution transpose layers, we custom-designed the LoRA layers, with details described in the following sections.

The model was pre-trained on a node with 4 NVIDIA A100 80GB for 300 Hours with FP32 full precision. The pre-training stage took around 1,200 GPU hours and has a total compute of 8.7e19 FLOPs (floating-point operations). The fine-tuned model was trained in parallel on a cluster with 32 NVIDIA A100 GPUs using TF32 precision, which has much higher throughput. Each fine-tuning was set to run 8 hours for 10 epochs. The fine-tuning step was performed on all 767 CAPS, leading to a total compute of 3.4e21 FLOPs, and took around 6,000 GPU hours using the lower precision TF32 tensor format. The total amount of compute for training and fine-tuning is around 3.5e21 FLOPs.

Compared with previous methods that trained separate models for individual CAPs, Chromnitron training employed a cross-CAP pretraining with per-CAP fine-tuning strategy which dramatically increased both training efficiency and model performance. To measure the training efficiency difference between individualized training approach versus LoRA finetuning, we used CTCF as an example to compare the two approaches. We found that training the model from scratch took more than 30 epochs to reach good performance, whereas fine-tuning from a pre-trained model only took 1 epoch to reach comparable performance. As the pre-training compute is only a small fraction of the fine-tuning compute (< 3% by estimation), assuming all CAPs requires a similar level of compute training from scratch, we estimated that the pre-training with fine-tuning strategy is at least one magnitude more efficient compared with individually training models for each CAP from scratch. The stored LoRA weight matrix is also 8 times smaller than full weight matrix. Meanwhile, pre-training substantially increases overall model performance through its generalized learning of CAP-binding principles across different CAPs.

Efficient multimodal sampling with OuterMix during pre-training

To improve model training efficiency at the cross-CAP pre-training phase, we developed OuterMix, an algorithm to generate Cartesian product of encoded features from randomly sampled data points on each dimension as the training units. OuterMix maximizes throughput by re-using modality-specific embeddings generated by each encoding step, thus dramatically reduces training complexities, and increases data loading and encoding efficiency exponentially.

First of all, a naïve sampling approach for multimodal training through randomly selecting a fixed number of examples for each batch would not be practical due to exponentially increased complexities. Each sample consists of random examples from different modalities. Given a fixed U number of samples from each modality with V modalities in total, the compute complexity would be O(U^V). Formally, if we are sampling $N \cdot M \cdot K$ samples in a batch representing the total amount of genomic loci, chromatin feature, and CAPs during pre-training, the three encoders need to encode all the samples. Using l, t, c to represent locus, cell type and CAP, respectively, the total random sample space for a batch $\mathcal{B}_{rand}$ can be defined as:

$$\mathcal{B}_{rand} = \bigcup_{i=1}^{NMK} \{(l_i, t_i, c_i)\},\ (l_i, t_i, c_i) \sim \mathcal{U}(\mathcal{L} \times \mathcal{T} \times \mathcal{C})$$

where each encoder needs to run across all samples.

OuterMix algorithm was designed to reuse embeddings generated by each encoder and reduce an exponential complexity O(U^V) down to a linear O(UV) complexity, increasing data loading and encoding efficiency exponentially, especially with large encoders and a high number of modalities. With OuterMix, we started by randomly selecting fixed number of samples from each dimension and produce a Cartesian product of loci (l_i), cell types (t_j) and proteins (c_k):

$$\mathcal{B}_{OuterMix} = \bigcup_{j=1}^{N} \bigcup_{i=1}^{M} \bigcup_{k=1}^{K} \{(l_i, t_j, c_k)\}$$

Here, each encoder only runs through samples from their predefined subset (Supplementary Fig. 3). In our implementation, the genome was first divided into non-overlapping 100kb chunks. In each iteration, the data loader selects one genomic chunk per cell type, followed by sampling 8,192bp windows and applying lightweight augmentations. For each window, we randomly selected up to 4 CAPs with their corresponding DNA sequence, ChIP-seq, and ESM2 data. Modality-specific encoders then generate feature-specific embeddings, which are tiled and concatenated to form the Cartesian product samples ready to be fed into the Transformer. This design allows the same sequence-ATAC tensors to be paired with multiple proteins, increasing Transformer trunk throughput while avoiding additional disk I/O and encoder overheads.

LoRA implementation during fine-tuning

Since Low-Rank Adaption (LoRA) was originally developed for fine-tuning natural language applications such as Large Language Models (LLMs), its implementation for one-dimensional convolution and convolution-transpose layers used in Chromnitron is not trivial[67]. Here we implemented LoRA for both Conv1D and ConvTranspose1D. On a high level, LoRA for convolution layers were implemented by freezing the original kernel and learning a low-rank update that can be added onto the kernel at evaluation time. The 1-dimensional convolutional layer weights are 3D tensors representing input dimension, kernel size and output dimension. In this implementation, we reshaped input and kernel size into one dimension and used it to construct two low-rank matrices.

For Conv1D, the convolutional kernel weights $W \in \mathbb{R}^{C_{\text{in}} \times C_k \times C_{\text{out}}}$ is kept frozen during fine-tuning, with C_{in}, C_{k}, and C_{out} representing input size, number of kernels and output size, respectively. The weight is separated into two low-rank matrices A and B with rank r, defined as:

$$A \in \mathbb{R}^{(r \cdot k) \times (C_{\text{in}} \cdot k)}, B \in \mathbb{R}^{C_{\text{out}} \times (r \cdot k)}$$

and is updated through $\Delta W = B\,A$. In the forward pass, the weight is added to the frozen weight to apply fine-tuning changes:

$$y = \text{conv1d}\left(x, W + \frac{\alpha}{r} \cdot \Delta W\right)$$

where y is the convolution activation and α defines the strength for LoRA weights in which the default is set to 1.

For ConvTranspose1D, the same decomposition is applied with a custom conv_transpose1d forward pass applying the ΔW, similar to conv1d:

$$y = \text{conv1d_transpose}\left(x, W + \frac{\alpha}{r} \cdot \Delta W\right)$$

We tested a range of ranks from 4 to 16 and did not observe significant performance difference. To reduce weight size and fine-tuning efficiency, we used r = 4 for the final large-scale fine-tuning. The custom LoRA implementation is available on https://github.com/tanjimin/LoRA.

Predicting CAP-binding profiles

During prediction, the LoRA checkpoint corresponding to each CAP was loaded together with the backbone weights. Since Chromnitron prediction operates on individual 8,192bp windows, to

predict CAP-binding signals across the entire chromosome, we used a sliding window approach with 5,120bp step size to create overlap between adjacent predictions. Chromnitron's predictions are robust to window shifts, but the predicted signal at the window edge can have lower quality due to the absence of flanking chromatin context. To mitigate this effect, we retained only the central 6,144bp predicted profiles out of the 8,192bp as valid predictions. With a 5,120bp step size and a 6,144 effective window size, adjacent predictions will have a 1,024bp effective overlap. To further ensure a high-quality transition between the overlapped regions between adjacent predictions, we implemented a weighted average in the overlapping regions to give more weights to bases that are closer to the center of their prediction windows. The weighted signal $\bar{S}$ is an average of left signal S_l and right signal S_r, weighted by their relative base location i in the region: $\bar{S} = \frac{i}{1024} \cdot S_l + \frac{1024-i}{1024} S_r$.

Prediction performance evaluation

All prediction performance evaluations were conducted on the leave-out chromosome 10. To compare with experimentally measured ChIP-seq profiles, we first predicted CAP-binding profiles across all five cell types on chromosome 10. For each CAP, we extracted profiles from both prediction and ChIP-seq profiles on the chromosome 10p arm between 6 Mb and 30 Mb. We evaluated model performance by calculating a set of metrics (detailed in the following section) between prediction and experimental ChIP-seq profiles in the leave-out chromosome across all cell types. For cell-type-specific prediction performance evaluation, we first calculated the experimental data difference and the predicted CAP-binding difference between HepG2 and K562 cells, the two cell types with the highest number of overlapping CAP profiles. We then applied the same set of evaluation metrics to compare the difference signal between the two groups.

For a comprehensive comparison, we computed in total eight statistics, including Pearson's correlation (r), Spearman's correlation (ρ), the coefficient of determination (R squared), F1 score, area under the receiver-operating characteristic curve (AUROC), and area under the precision-recall curve (AUPR) with both target-centric and prediction-centric labelling. We repeated the statistics over a grid of different bin sizes and top-percentile subset of the bins. Predicted or ChIP-seq signals were first converted to numerical vectors, with missing values set to zero. Subsequently, the vectors were mean-reduced into non-overlapping bins of 50, 250, 500 or 1000bp. This multi-scale representation allows assessment of both fine-grained and coarse regulatory patterns. The top-percentile bins measure how well the model can differentiate between peaks and their relative enrichment level. For this, we set an exponential increment of percentile using a list of 100% (1), 50% (1/2), 25% (1/4), 12.5% (1/8), …, 0.05% (1/2048). For AUROC and AUPR calculation, we used the median value of the selected bins as threshold to separate positive and negative labels. We also tested an alternative strategy to select most characteristic bins by selecting both top and bottom percentile bins (Supplementary Fig. 6).

To categorize CAP-binding prediction performance, we focused on the Pearson's correlation (r) as a simple metric to help interpreting the prediction results. We categorized CAP-binding prediction performance with both validation performance and cell-type-specific performance on

789 the leave-out chromosome (chr10) for all CAPs in cell types where training data are available. The
790 validation performance was defined as the highest performance across all cell types, while the cell-
791 type-specific performance was calculated as described above. To establish a relationship between
792 validation performance and cell-type-specific performance, we correlated the validation
793 performance against cell-type-specific performance and found a general positive correlation. The
794 cell-type-specific performance is a relatively more challenging metric due to additional
795 experimental noises across cell types. We selected 187 CAPs that are shared between HepG2 and
796 K562 and fitted a line between validation performance and cell-type-specific performance. After
797 custom evaluation of prediction performance, we assigned CAPs to different performance groups
798 based on both validation and cell-type-specific prediction performance metrics. The specific
799 performance cutoff was chosen based on the fitted curve between the two metrics (Supplementary
800 Fig. 4f). We classified a CAP into high-performance group if its prediction performance has either
801 a validation performance > 0.6 or the corresponding cell-type-specific performance. CAPs with
802 either a validation performance > 0.4 or the corresponding cell-type-specific performance were
803 assigned in medium-performance group. The rest of CAPs were categorized into low-performing
804 group. Notably, although some CAPs were classified into low-performance group by Pearson's
805 correlation, their predicted binding profiles recapitulate ChIP-seq peaks (Supplementary Fig. 5-6).
806 Thus, the proposed CAP-performance categorization should be regarded as a practical metric for
807 interpreting prediction results, rather than a strict measure of biological validity.
808
809 **Global comparison of prediction to experimental data at CAP-binding regions**
810 To evaluate prediction performance across different chromatin backgrounds, we compared CAP-
811 binding signals by aligning Chromnitron predictions with matched ChIP-seq and ATAC-seq
812 profiles centered in peak regions. For each CAP of a particular cell type, three one-dimensional
813 signal tracks were collected: (i) Chromnitron-predicted CAP-binding signal, (ii) experimental
814 ChIP-seq data (reference), and (iii) ATAC-seq data to represent chromatin accessibility.
815 Chromnitron-predicted CAP-binding data on chromosome 10 were generated and documented, as
816 described in the previous performance evaluation section. Peaks were independently identified in
817 prediction, ChIP-seq and ATAC-seq by thresholding the signal at 0.5 (half of very high peaks),
818 merging contiguous bins above this level, and discarding events narrower than 100 bp. The mean
819 signal within each retained interval was recorded to generate a peak set annotated by genomic
820 coordinates and intensity.
821
822 To visualize and directly compare prediction to the reference CAP-binding data, all peaks detected
823 in Chromnitron prediction were stratified into two groups according to chromatin accessibility
824 signal. Each predicted peak was labelled according to its overlap with ATAC-seq peaks, generating
825 an accessible set and an inaccessible set. For every peak in each group, we extracted a 2kb genomic
826 window centered at the highest score in each peak. Finally, we visualized the predicted CAP-
827 binding signal, ChIP-seq signal, and ATAC-seq signals as heatmaps to evaluate global prediction
828 performance across chromatin background.
829
830 **Ablation study**

We performed ablation studies to measure how the model performs under different constraints. This analysis includes performance evaluation across models trained without fine-tuning, and models trained with missing cell types or CAPs to test their generalizability across modalities in unseen contexts. We performed ablation studies on a random subset of chromosome 10 to reduce computation and storage consumption. Our preliminary calculation found that for one set of experiments, inferring 767 CAPs on chromosome 10 for five cell types could take two weeks and 6.5 TB of storage. Instead, we used Pearson's correlation coefficient calculated during the validation step embedded in the evaluation process to save computation and storage. This evaluation was performed on 500 random loci on chromosome 10 across all cell types and works as a proxy of the genome-wide prediction performance statistics. The performance statistics used for both the full Chromnitron model and ablated models were all calculated from validation step using the same metrics.

To evaluate the performance improvements of fine-tuning, we extracted performance metrics from the pre-trained model at its last epoch (390) and metrics from fine-tuned models at their last fine-tuning epoch (≤ 10). The last fine-tuned epoch can change due to our early stopping strategy. Then, we compared the performance individually across cell types and CAPs.

For cell-type-specific prediction evaluation through ablation study, we kept data from HepG2 and K562 cells as the training set, and left out data from IMR-90, GM12878 and HCT116 cells as testing. Chromosome 10 region was again used as leave-out chromosome. This model was then pre-trained (i.e. without fine-tuning) and compared with the full Chromnitron pre-training model in leave-out cell types. For CAP-specific evaluation, we randomly left out half of the CAPs (n=382) and kept the other half (n=385) in the training set. Then we extracted the performance metrics from the ablated model and compared with the full Chromnitron pre-training model on leave-out CAPs (again in chromosome 10 regions).

Benchmarking prediction performance

Although there was no directly comparable model that used the same multimodality framework, we benchmarked performance of Chromnitron against previous machine learning approaches that aimed for similar goals, including Enformer and maxATAC, and the motif accessibility score as a baseline approximation[31,44]. Detailed inference of CAP binding signal for each approach was described below. For each benchmark, we compared the inferred CAP-binding tracks against the experimental data reference (Chrom2Vec processed ENCODE ChIP-seq profiles). We calculated eight different metrics across bins and top binding regions, similar to the procedure described in the model performance evaluation section.

Enformer: We used GRCh38 reference genome sequence as input to Enformer. Chromosome 10 was divided into non-overlapping 100 kb windows; for each window, 196,608 bp centered on the window mid-point were retrieved (zero-padded at chromosome ends) and one-hot encoded. Each sequence was forwarded through the PyTorch Enformer implementation (EleutherAI/enformer-

official-rough) which has similar performance as the TensorFlow implementation. The model returned an 896-bin output tensor per window which was cached on disk. CHIP channels relevant to HepG2 or K562 cells were identified from human ChIP-seq targets. For every requested TF channel, the central 50,176 bp (±50,176 bp) of each window were extracted to avoid model padding artefacts. The selections were written into an array representing chromosome 10 at 1bp resolution. Overlaps occurred only at window boundaries and were overwritten by later windows. The final arrays were exported to BigWig files for visualization and comparison with other approaches.

MaxATAC: Raw ATAC-seq data were obtained from ENCODE as BAM files (HepG2: ENCFF939JOT, K562: ENCFF804MXV). These files were down-sampled to 40M reads and deduplicated to generate the BAM files required for maxATAC pipeline. MaxATAC pre-processing ("maxatac prepare") was performed with "-c chr10" to build the ATAC coverage tracks necessary its prediction. For every CAP that is included in both Chromnitron and maxATAC model, we used the corresponding sub-model to predict the CAP binding profile as a BigWig file using "maxatac predict".

Motif-Accessibility: Non-redundant motif from Jeff Viestra (https://resources.altius.org/~jvierstra/projects/motif-clustering/releases/v1.0/) were filtered to only retain human CAPs. Motif scanning results on chr10 were converted into a dense 1-bp signal for each motif with the corresponding matching scores. A 20-bp moving average was applied to the dense track. The final results were stored as BigWig files. To weight motif by chromatin accessibility, we loaded ATAC-seq for all five cell types (HepG2, K562, GM12878, IMR-90, HCT116). For each cell type, the smoothed motif track was multiplied elementwise by the corresponding ATAC-seq signal and stored as BigWig files.

***In silico* DNA sequence mutagenesis**

To identify the impact of DNA sequence on the binding of a CAP, we focused on regions where Chromnitron predicted intense CAP-binding signals. For each CAP-binding peak, we extended the locus by ±3 kb and records its midpoint base location. We defined a 300-bp "mutation window" (±150 bp) flanking the midpoint and systematically substituted every base within the window. Subsequently, we performed *in silico* saturated DNA base mutagenesis of the 300bp window, one base at a time. For every position, we generated four sequences, each carrying a single A, C, G or T substitution, and predicted the corresponding CAP-binding signal for each sequence. We recorded the original sequence and its predicted CAP-binding signal as wild-type (WT) reference. We then calculated the effect of every mutation as the difference between mutant and WT predictions. The resulting inferred CAP-binding data at this window was stored as a 300 × 4 × 8,192 tensor, with the three dimensions representing the number of mutation positions, number of substitution nucleotides, and the window size of the CAP-binding prediction, respectively. The signals represent the predicted CAP-binding intensities. To represent only the CAP-binding region, the tensor was cropped from 8,192bp to the central 1kb and averaged, generating a 300 × 4 matrix whose values represent CAP-binding intensities across the 300bp mutated positions and 4

nucleotides. This matrix served as the nucleotide-resolution perturbation signals for downstream motif analyses and interpretation. For visualization purposes, the impact motif was normalized by subtracting the median CAP-binding intensities across the 4 nucleotides (i.e. the mean binding intensities of the middle two nucleotides) at each position. In essence, the impact motif is computed from a single locus, differing from the conventional motif calling which computes consensus sequence from a large number of loci.

Cross-CAP motif discovery and clustering

To analyze motifs across all CAPs, we generated weight matrices (represented by median-subtracted CAP-binding signals as described above) for all CAPs at their top binding loci. These weight matrices were used as input for TF-MoDISco to generate motifs for each individual CAP[68]. For every CAP, the tool scans up to 20,000 high-scoring 200-bp windows ("seqlets", or high-importance windows of the sequences, as referred in TF-MoDISco), clustering them by k-mer similarity and reports positive and negative patterns (i.e. sequels). After identifying relevant seqlets for a CAP, we mapped their corresponding seqlets back to the genomic coordinates and applies a strict quality control procedure: a pattern must contain at least one seqlet whose high (90th-percentile) absolute weight difference is nontrivial (>0.1) after median subtraction. This procedure excluded noisy results from the analyses. We then stored the remaining patterns for every CAP. To pick the most representative seqlets for each CAP, we ranked them by (1) the number of retained seqlets, (2) the overall motif intensity represented by the median difference value and (3) the highest motif intensity represented by median 90th percentile difference. Patterns whose seqlets overlapped more than ten distinct loci and repeated in different CAPs were discarded to avoid trivial repeats.

Then, we built normalized weight matrices for cross-CAP comparison from the top ranked seqlets. For each seqlet, we subtracted its per-position median and scaling by its mean absolute difference. Each motif was flattened into a single vector (length by 4). Pairwise distances were computed as the maximal cross-correlation between motif A and motif B or B's reverse complement. We performed hierarchical clustering with Ward linkage based on the distance metrics, generating the clustering heatmap for different motifs.

***In silico* chromatin accessibility perturbations**

We measured the changes of predicted CAP binding upon *in silico* scaling of chromatin accessibility signal to examine Chromnitron's sensitivity to local chromatin states. We employed a scaling approach through multiplying the ATAC-seq profiles by a defined factor (multiplier) to alter accessibility intensities while preserving peak shape. Meanwhile, the size of the perturbation window (i.e. number of bases) also matters in this *in silico* perturbation. Thus, we created an *in silico* perturbation grid by combining multiplicative factor between 0 to 2, and with a perturbation radius that ranged between 50 and 1,000bps. For a given radius, the slice of the ATAC vector centered on the locus was multiplied by the selected factor; the remainder of the signal was kept the same. For each combination of multiplier and perturbation radius size values, we predicted the corresponding CAP-binding signal for perturbation and calculated the binding differential by

957 subtracting the unperturbed condition, generating a data tensor of 40 × 20 × 8,192 (multipliers ×
958 window-sizes × predicted profile). Similar to *in silico* DNA sequence mutagenesis, we averaged
959 the center 400bp region of difference between perturbed prediction and wild-type prediction to
960 generate a 40 × 20 matrix that measures the accessibility perturbation effects.
961
962 ***In silico* protein residue mutagenesis**
963 To investigate how the model leverages protein embedding features in prediction, we performed
964 systematic mutagenesis directly on the input CAP residues without referring to particular protein
965 domain or functional data. Each perturbation operates on a 30-amino acid sliding window with
966 single-residue steps, replacing the window with random amino acids to generate the corresponding
967 mutant protein sequence. Mutated CAP sequence was feed into the same protein encoder (ESM2)
968 to generate the corresponding mutant embeddings, which were used in downstream prediction of
969 mutant binding profile and comparison with wild-type protein.
970
971 To ensure more robust comparison of protein mutation effects between the mutant and wild-type
972 CAP, we selected the top 200 Chromnitron-predicted binding peaks for each CAP in HepG2 cells,
973 filtered for ATAC-seq overlap, and equally extended to the flanking sequences up to 8,192bp to
974 facilitate prediction. Chromnitron predictions of the corresponding CAP were generated for wild-
975 type and mutant embeddings across all selected loci. We then extracted maximum scores from the
976 center 400bp region of each window (recorded as "wildtype_max" and "mutant_max",
977 respectively). CAP mutation impact scores were calculated as ($\frac{|\text{mutant_max} - wildtype_max|}{wildtype_max}$). For
978 protein domain interpretation, we retrieved InterPro domain annotations and computed precision-
979 recall curves using domain boundaries as labels and mutation scores as predictions[69].
980 Domain enrichment was quantified using AUPR compared to random baselines.
981
982 **Inferring CAP interactions through colocalization analysis**
983 To investigate whether model prediction serves as a potential source to study protein-protein
984 interaction between CAPs, we followed our previous approach and applied linear non-Gaussian
985 acyclic model (LiNGAM) causal-discovery algorithm to a selection of high-quality CAPs
986 prediction across HepG2 ATAC-seq peaks[37,70]. We found that, as expected, the strongest
987 colocalizations were often between transcription factors that belong to the same family as they
988 usually share similar conserved DNA-binding domains. To systematically evaluate whether
989 colocalization calculated from experimental ChIP-seq or predicted CAP-binding data reconciles
990 with known physical protein-protein interactions, we benchmarked our casual discovery-based
991 interaction graph with Pearson's correlation-based interaction graph. This analysis was performed
992 on both experimental and predicted ChIP-seq data using confident interaction annotation from
993 STRING database (using scores over 700 as positive, and 400 to 700 as negative)[71]. The casual
994 discovery on predicted data yields the highest AUPR (0.426), followed by causal discovery on
995 observed data (0.372), while the correlation-based approaches achieved AUPR scores of 0.351 and
996 0.321 for predicted and experimental ChIP-seq data, respectively (Supplementary Fig. 11). The
997 lower performance when using ChIP-seq data could be a result from heterogeneous noise

distribution in the experimental data compared to the predicted results, leading to prediction of only strongest interactions.

Single cell differential ATAC-seq peak calling

To identify differentially accessible chromatin regions between pairs of cell types from single-cell ATAC-seq data, we employed a computational workflow utilizing SCANPY[72]. For each pairwise comparison, we first extracted the cells from the two specified populations the from complete peak-by-cell matrix. Peaks with low signal, defined as those with fewer than 500 total fragment counts across the selected cells, were removed from consideration. To account for technical variations in sequencing depth, we normalized the counts for each cell to a total of 100,000, followed by a log1p transformation to stabilize the variance. Subsequently, we identified highly variable peaks between the two cell populations by using a dispersion-based method implemented in SCANPY (scanpy.pp.highly_variable_genes), retaining peaks that exhibited a normalized dispersion of at least 0.1. The genomic coordinates of these candidate regions were used as candidate differentially accessible regions and exported for downstream analysis.

Predicting CAP-binding at single locus across T cell states

We used a public single-cell ATAC-seq data of T cell exhaustion collected from human patient PBMC samples[55]. Four representative T cell stages (from effector to exhausted states) were prepared to relate CAP binding changes to T-cell states. Pseudo-bulk ATAC-seq data for each stage was generated by sampling 10M reads from cells in the vicinity of selected center. The pseudo-bulk ATAC-seq was then processed and normalized as described in the single cell data processing method section. Differential accessible peaks of the corresponding cell populations were called as described in the above section. These peak regions served as putative *cis*-regulatory elements that might play important functions in T cell states. Next, for a given regulatory element of interest, we predicted CAP-binding profiles at an element-centered 8,192bp window across the cell states. As single locus is highly susceptible to noise, we selected a set of inference CAPs using a more stringent cutoff of 0.5 (Pearson's r) validation performance for this analysis. Predictions for every T cell state were cropped to a 300bp window around the CRE center. The binding signal was normalized by median on the genome dimension to remove global scaling differences. Trajectories were plotted as ridge plots that overlay the complete track profile per CAP. The mean CAP-binding signals across trajectory were smoothed with Gaussian kernel and visualized with line charts tracing binding dynamics across cell state transition trajectory.

Genome-wide identification of regulatory CAPs in T cell exhaustion

The same T cell exhaustion scATAC-seq and annotation in single-locus-analysis were used for this analysis. We first isolated reference pseudo-bulks for both effector T cells and exhausted T cells according to the original cell type annotations[55]. To obtain a cleaner causal relationship between CAPs and target genes, we focused on promoter regions as the putative regulatory elements. We used GENCODE v45 transcript annotation (GFF3) and the hg38.chrom.sizes file hosted by IGV to define promoter coordinates and chromosome boundaries[73]. Promoter regions (±500 bp around the transcription-start site, TSS) were extracted from the GENCODE annotation.

1040 Genes on sex chromosomes or mitochondrial DNA were discarded. We also removed transcripts
1041 not matching a whitelist of supported biotypes (protein-coding, lincRNA, miRNA, snRNA,
1042 snoRNA, rRNA). The maximal and mean ATAC-seq accessibility signal were computed in both
1043 T-cell states for each promoter. Promoters with a cross-condition maximum accessibility ≤ 1 were
1044 removed to exclude inactive loci. We then calculated log2 fold change (log2FC) of accessibility
1045 for all the loci, and partitioned them into three categories: (1) up-regulated promoters, defined as
1046 top 2,000 positive log2FC values, (2) down-regulated promoters defined as top 2,000 negative
1047 log2FC values, and (3) stable promoters, defined as 100 promoters with the smallest absolute
1048 log2FC. We also considered alternative TSS as separate promoters and implemented additional
1049 constraints to merge overlapping TSS loci.
1050
1051 Due to the high compute requirement for this task, Chromnitron prediction was sharded by cell
1052 type and all CAPs, and executed in parallel on a GPU cluster. For this task, we used the full set of
1053 CAPs since genome-wide analysis can better tolerate noisy predictions. After all shards completed,
1054 we merged the predictions by iterating over chromosomes, stacking the individual CAP arrays to
1055 generate genome_length × CAPs matrices. This matrix was rechunked to optimize downstream
1056 biological analyses for characterizing putative regulator CAPs.
1057
1058 Promoter-level average CAP-binding signals were extracted by recording the maximum predicted
1059 signal in the 1kb window centered at TSS, generating a three-dimensional tensor with dimension
1060 (cell types × promoters × CAPs). Downstream analyses used the log-transformed values. A
1061 promoter-specific differential CAP-binding score was calculated by subtracting the predicted
1062 CAP-binding signals of the exhaustion group by that of the effector T cell group. To account for
1063 basal CAP-specific variability, the difference matrix was z-standardized by subtracting the mean
1064 and dividing by the standard deviation of the difference values measured at stable promoters. To
1065 increase signal-to-noise ratio, mid-range values between the 10th and 90th percentiles were set to
1066 zeros separately across loci and CAPs, thereby highlighting the most differential – potentially
1067 functional – signals for informing downstream analysis. Next, we performed k-means clustering
1068 on both the loci and CAP dimensions. Silhouette analysis over k-means clustering using k ranging
1069 from 3 to 40 was used to derive optimal number of clusters. To identify important CAPs and loci
1070 within each cluster, we ranked loci and CAPs by revaluating the average binding against the
1071 opposing cluster assignments iteratively on both dimension until convergence.
1072
1073 **Global CAP and CRE dynamics characterization during Purkinje cell differentiation**
1074 Single-cell ATAC-seq datasets for Purkinje cell development were obtained from Mannens et al[60].
1075 We adopted the original assignment of Purkinje cell differentiation trajectory encompassing seven
1076 stages: progenitor cells (Prog), two neuroblast stages (NBL1, NBL2), and four Purkinje cell
1077 maturation stages (Purk1 to Purk4). All datasets were processed using the scATAC-seq
1078 preprocessing steps described earlier. Only autosomes were included for this analysis.
1079
1080 To identity genomic regions of interest, we used multiple filtering approaches based on chromatin
1081 accessibility dynamics across developmental stages. Initial candidate regions were generated from

cPeaks representing the repertoire of accessible regions across human cell types[64]. We extended each peak center by 500 bp flanking region on each side to create 1 kb inference windows to include the complete peak region. Then, we filtered these loci by classifying them into two categories based on temporal accessibility patterns: differential and stable regions. Differential loci were defined as regions exhibiting significant accessibility changes across developmental stages (log2 fold-change larger than 0.1 compared with average accessibility), representing dynamic regulatory elements. Stable loci have relatively minimal accessibility change, defined as having absolute log2FC smaller than 0.1. Similar to the genome-wide regulator screening in T cell exhaustion, these stable loci served as normalization controls for downstream analysis. We selected the top 1,000 differential loci and top 100 most stable loci. To ensure non-redundant genomic coverage, overlapping regions were merged. Regions with insufficient chromatin accessibility across all conditions were excluded to prevent analysis of inactive or noisy genomic elements. The final dataset comprised 1,073 high-confidence genomic loci distributed across autosomes.

This task was computationally intensive and was thus parallelized on a GPU cluster. To reduce compute, we only keep a set of high-performance CAPs similar to the single locus T cell exhaustion inference experiment. Individual CAP predictions were concatenated to construct a 3D data tensor of developmental time × genomic loci × CAPs. For each genomic locus, we calculated the maximum peak heights within 1 kb windows centered at peak summits. The resulting 3D tensor was filtered for missing values, infinite numbers, and outliers. Data normalization was performed using stable loci as internal controls to account for technical variation and global accessibility changes across developmental stages. We performed z-score normalization using the mean and standard deviation of differential patterns calculated from stable loci. Outlier detection and removal were implemented using percentile-based clipping (1st and 99th percentiles) to prevent extreme values from dominating downstream clustering analyses.

Regulatory pattern discovery was performed using multiple complementary approaches to capture different aspects of the temporal regulatory landscape. The 3D tensor was reshaped into matrices emphasizing different analytical perspectives. There are three ways to transform the 3D tensor to 2D matrix for specific analyses, each enabling discovery of distinct regulatory relationships. First, we can use CAP as features, reshaping the tensor to a matrix with dimension (time × loci) by CAPs to identify loci whose activities change across time. Second, we can use loci as features to generate the (time × CAPs) by loci matrix. This approach can represent how CAP binding profiles change across time. Finally, using time as the feature, the reshaped matrix can tell how a binding event specific to a CAP and a locus varies across time.

We focused on the first analysis using CAP as features. To observe CRE activity shifts across time, we performed hierarchical clustering of CREs on the first pseudo-time point. This approach grouped CREs with similar regulatory dynamics. To further elucidate the relationship between CRE and time, we applied k-means clustering using a k between 5 and 20 to identify regulatory locus modules across all time points and CREs. The optimal cluster number was determined using

silhouette score. Clustering was performed on unit-normalized data using Euclidean distance metrics, with multiple random initializations to ensure stable cluster assignments. For visualization and interpretation, t-distributed Stochastic Neighbor Embedding (t-SNE) was used to generate 2D embeddings of high-dimensional regulatory patterns. Each dot on the t-SNE embedding represents a CRE at a specific time point and was colored by clustering results representing its identity.

To visualize transition of CREs between adjacent pseudo-time points in Purkinje cell development, we calculated a vector field where the start and end were defined by the t-SNE location of each locus at first and second time points. This velocity field represents how CREs change their regulatory identity between consecutive developmental stages. We applied Gaussian kernel smoothing to the vector field to reduce noise while preserving temporal trends.

Cell culture and primary cell isolation

HEK293T was obtained from the American Type Culture Collection (ATCC). HEK293T cells were maintained in Dulbecco's Modified Eagle Medium (DMEM; Gibco, 11965092) supplemented with 10% fetal bovine serum (FBS; Gibco, A5670701) and 1% penicillin–streptomycin (Pen/Strep; Gibco, 15140122). Human Peripheral blood mononuclear cells (PBMCs) were isolated from healthy donor leukopaks (Charles River Laboratories Cell Solutions, 103051-034) using Ficoll-Paque PLUS density gradient centrifugation (Cytiva, 17144002) according to the manufacturer's protocol. Isolated PBMCs were cryopreserved in 90% human AB serum (MP Biomedicals, 092938249) plus 10% dimethyl sulfoxide (Sigma-Aldrich, D4540) at a density of 30M cells per ml. For in vitro T cell assays, thawed PBMCs or purified $CD8^+$ T cells (isolated using EasySep™ Human CD8+ T Cell Isolation Kit, STEMCELL Technologies, 17953) were cultured in X-VIVO 20 medium (Lonza, BW04-448Q) supplemented with 5% heat-inactivated human AB serum (MP Biomedicals, 092938249) and 10 ng/ml recombinant human IL-2 (PeproTech, 200-02-50UG). T cell cultures were maintained at 37 °C in a humidified incubator with 5% CO_2.

sgRNA library preparation

For each selected gene, five sgRNAs were designed to either knock out or repress gene expression using the CRISPick online tool (Broad Institute). The complete list of sgRNA sequences is provided in Supplementary File 1. A total of 370 DNA oligonucleotides corresponding to sgRNAs were synthesized as pooled libraries (500 ng per library) by Twist Bioscience. For gene knockout experiments, sgRNAs were cloned into the lentiviral vector pXPR_023 vector (Broad Institute, also available through Addgene #52961), which encodes Cas9, a puromycin resistance gene, and the sgRNA scaffold. For gene repression (CRISPRi), sgRNAs were cloned into pXPR_066 vector (obtained via Broad Institute Genetic Perturbation Platform), which expresses KRAB-dCas9 under the EFS promoter along with a puromycin resistance cassette and sgRNA scaffold. Oligonucleotides were cloned into the respective backbone vectors using Golden Gate assembly with BsmBI restriction enzyme (New England Biolabs, R0739S), following the protocol described by the Feng Zhang lab (Broad Institute). Ligation products were transformed into NEB Stable Competent *E. coli* (New England Biolabs, C3040H) and plated on LB-agar plates supplemented

1166 with carbenicillin (100 μg/ml). Colonies were scraped from the plates and pooled, and plasmid
1167 DNA was extracted using the ZymoPURE II Midiprep Kit (Zymo Research, D4201) and
1168 quantified by NanoDrop for downstream lentivirus production.
1169
1170 **Lentivirus production**
1171 Lentivirus was produced in HEK293T cells (ATCC; passage number <20) seeded at 4×10^7 cells
1172 per 150-mm tissue culture dish. Twenty-four hours after seeding, the culture medium was replaced
1173 with fresh pre-warmed DMEM containing 10% FBS. Transfection was performed using
1174 polyethyleneimine (PEI 40K; Polysciences, 24765) as follows: 13.3 μg of the sgRNA vector, 10 μg
1175 of packaging plasmid psPAX2, and 6.7 μg of envelope plasmid pMD2.G were mixed in 400 μl of
1176 Opti-MEM (Thermo Fisher Scientific, 31985070) (Tube A). Separately, 90 μl of PEI (1 mg/ml)
1177 was diluted in 310 μl of Opti-MEM (Tube B). The two solutions were incubated separately for
1178 5 min at room temperature, combined, and incubated for an additional 15 min before being added
1179 dropwise to the HEK293T cells. Seventy-two hours post-transfection, the culture supernatant was
1180 harvested and centrifuged at 3,000g for 15 min to remove cellular debris. Lentiviral particles were
1181 precipitated by adding polyethylene glycol 8000 (PEG8000; Sigma-Aldrich, P2139-500G) to a
1182 final concentration of 8% (w/v) and incubating the mixture overnight at 4 °C. The precipitated
1183 lentiviral particles were pelleted by centrifugation at 1,500g for 30 min and resuspended in 500 μl
1184 of X-VIVO 20 medium (Lonza, BW04-448Q). Aliquots were stored at −80 °C until use.
1185
1186 **T cell stimulation assays**
1187 Cryopreserved human CD8+ T cells were thawed and resuspended in T cell culture medium (X-
1188 VIVO 20 supplemented with 5% human AB serum and 10 ng ml^{-1} recombinant human IL-2) at a
1189 density of 1×10^6 cells per ml in 12-well plates. Cells were stimulated by the addition of Human
1190 CD3/CD28 T Cell Activation Beads (BioLegend, 422603) at a 1:1 bead-to-cell ratio and incubated
1191 for 48 h at 37 °C with 5% CO_2. Following activation, beads were removed using a magnetic
1192 separation rack according to the manufacturer's instructions. Cells were then washed twice with
1193 fresh T cell medium and split into two different stimulation conditions. For the acute stimulation
1194 group, cells were maintained in T cell medium alone. For the chronic stimulation group with
1195 antibody cocktail, cells were cultured in T cell medium supplemented with ImmunoCult Human
1196 CD3/CD28 T Cell Activator (Stem Cell Technologies, 10971), following the manufacturer's
1197 protocol. For the chronic stimulation group with anti-CD3 antibody (eBioscience, 16-0037-81),
1198 the antibody was coated at 12 well plate with 5 ug/ml concentration. Cells in the chronic condition
1199 were passaged every 48 h, with medium replacement and re-activate for three consecutive
1200 stimulation rounds. At the end point, cells were harvested and divided into two groups for either
1201 surface staining of exhaustion markers or intracellular cytokine analysis by flow cytometry.
1202
1203 **Lentiviral transduction and titration in $CD8^+$ T cells**
1204 To determine optimal lentiviral transduction conditions for primary human $CD8^+$ T cells, cells
1205 were activated for 48 h using CD3/CD28 beads as described above. After bead removal, cells were
1206 plated in 96-well plates at 2×10^5 cells per well in 200 μl of T cell medium (X-VIVO 20
1207 supplemented with 5% human AB serum and 10 ng/ml IL-2). Serial dilutions of lentivirus

encoding puromycin resistance gene were added to the wells in the presence of 8 μg/ml polybrene (Millipore, TR-1003-G). Plates were centrifuged at 1,500g for 90 min at 37 °C, the medium was changed to T cell medium after additional 4h incubation, the cell was continued culture at 37 °C, 5% CO2. At 48 h post-transduction, cells were diluted to 1×10^5 cells per ml and cultured in the presence of 1 μg/ml puromycin for 48 h. cell viability was assessed by counting viable cells using trypan blue. Viral titer and multiplicity of infection (MOI) were calculated based on the fraction of puromycin-resistant cells relative to total input.

CRISPR screening in primary $CD8^+$ T cells

Human $CD8^+$ T cells were isolated from PBMCs by negative selection and activated with CD3/CD28 beads for 24 h. After bead removal, cells were transduced with pooled lentiviral sgRNA libraries at an MOI of 0.3 in the presence of 8 μg/ml polybrene. Cells were spinoculated at 1000g and 32 °C for 90 min. 6h after spin infection, the cell were resuspended to T cell medium for overnight culture. 24 hours after spin infection, the cell were treated with puromycin (1.5 μg/ml) to select the transduced cells. Following selection, cells were washed and resuspended into two treatment conditions corresponding to acute and chronic stimulation, as described previously. Cells were maintained under these conditions for 10 days, with medium refreshed and IL-2 re-supplemented every 2 days. On the final day of the screen, genomic DNA was harvested from each condition using the QIAamp DNA Blood Maxi Kit (QIAGEN, 51192) following the manufacturer's instructions. Genomic DNA was quantified using Qubit dsDNA BR Assay Kit (Thermo Fisher Scientific, Q32850), and sgRNA sequences were PCR-amplified from extracted DNA using a one-step amplification protocol compatible with next-generation sequencing, as previously described by Broad Institute Genetic Perturbation Platform ("sgRNA/shRNA/ORF PCR for Illumina Sequencing" linked through https://portals.broadinstitute.org/gpp/public/resources/protocols), PCRprimers were provided in Supplementary File 2. PCR products were purified using AMPure XP beads (Beckman Coulter, A63880). Libraries were quantified, normalized, and sequenced on an Illumina NovaSeq X platform using a 25B flow cell, generating paired-end 150 bp reads. Raw sequencing data were demultiplexed using bcl2fastq (Illumina) and further processed as described below.

CRISPR screening analysis

Sequencing reads from CRISPR screens were first demultiplexed and trimmed to extract sgRNA sequences. Reads were aligned to the sgRNA reference library using the MAGeCK count function in the MAGeCK software package with default parameters[74]. Only sgRNAs with perfect matches to the reference library were retained for downstream analysis. Read count matrices were generated for all samples and normalized using median normalization. Differential sgRNA abundance between experimental conditions (Chronic vs. Acute stimulation) was analyzed using the MAGeCK test module, which applies a negative binomial model to assess gene-level enrichment or depletion. For each gene, MAGeCK aggregates sgRNA-level statistics to compute a robust gene-level score and associated p value using a robust ranking aggregation (RRA) algorithm. Genes were ranked by their adjusted p values (Benjamini–Hochberg FDR correction),

1249 and significance thresholds were defined as indicated in the figure legends. All analyses were
1250 performed using MAGeCK with default settings unless otherwise noted.
1251
1252 **RNP electroporation in primary human CD8$^+$ T cells**
1253 To perform Cas9 RNP-mediated gene knockout, two synthetic sgRNAs targeting the same gene
1254 were used per condition (sequence available in Supplementary File 2). Each sgRNA was
1255 resuspended to a final concentration of 100 μM in RNase-free water and stored at −80 °C until use.
1256 Cas9 protein (EnGen® SpyCas9 NLS, 20 μM stock, New England Biolabs, M0646T) was used
1257 for complex formation. For each reaction, 60 pmol of Cas9 protein was complexed with 180 pmol
1258 total sgRNA by mixing 3 μl of Cas9 protein, 0.9 μl of each sgRNA, and 0.4 μl of 10× NEBuffer
1259 3.1 in a total volume of 5.2 μl. The mixture was incubated at room temperature for 15 min to allow
1260 RNP formation. Primary CD8$^+$ T cells were activated for 48 h with CD3/CD28 beads as described
1261 above, then washed once with 1× PBS and resuspended in Lonza P3 electroporation buffer at a
1262 density of 1×10^6 cells per 20 μl. The entire 5.2 μl RNP mixture was added to 20 μl of cell
1263 suspension, mixed gently, and transferred into a well of a 16-well Nucleocuvette™ strip (Lonza).
1264 Air bubbles were removed by tapping the strip on the benchtop to prevent disruption of the electric
1265 pulse. Electroporation was performed using the Lonza 4D-Nucleofector system with the program
1266 EH115, optimized for primary human T cells. Immediately after electroporation, 80 μl of pre-
1267 warmed T cell medium (X-VIVO 20 supplemented with 5% human AB serum and 10 ng ml^{-1} IL-
1268 2) was added to each well. The cells were incubated in the cuvette strip at 37 °C for 10 min to
1269 enhance recovery before being gently transferred to pre-warmed 24-well plates.
1270
1271 **Flow cytometry staining and analysis**
1272 For intracellular cytokine staining, T cells were stimulated for 3 h at 37 °C with Cell Activation
1273 Cocktail containing Brefeldin A (BioLegend, 423303). Cells were washed with Cell Staining
1274 Buffer (BioLegend, 420201), fixed and permeabilized with Fixation/Permeabilization solution
1275 (BD Biosciences, AB_2869008) for 20 min at 4 °C, and washed twice with 1× Perm/Wash buffer.
1276 After Fc receptor blocking (BioLegend, 422302), cells were stained for 2 h at 4 °C in the dark with
1277 fluorophore-conjugated antibodies in staining buffer (BioLegend, 420201) along with NIR
1278 Live/Dead dye (1:10000 dilution, ThermoFisher, L34975). Antibodies included APC anti-human
1279 CD8 Antibody (BioLegend, 344722), Brilliant Violet 785™ anti-human IFN-γ Antibody
1280 (BioLegend, 502541), Brilliant Violet 605™ anti-human TNF-α Antibody (BioLegend, 502935),
1281 and Brilliant Violet 421™ anti-human IL-2 Antibody (BioLegend, 500327).
1282
1283 For surface marker staining, cells were blocked with FC receptor blocking reagent (BioLegend,
1284 422302), then stained with antibodies and NIR Live/Dead dye (1:10000 dilution, ThermoFisher,
1285 L34975) at 4 °C for 1 h. Antibodies included CD8 Antibody (BioLegend, 344722), Brilliant Violet
1286 785™ anti-human TIGIT (VSTM3) Antibody (BioLegend, 372735), Brilliant Violet 605™ anti-
1287 human CD366 (Tim-3) Antibody (BioLegend, 345017), Brilliant Violet 510™ anti-human CD223
1288 (LAG-3) Antibody (BioLegend, 369317), and Brilliant Violet 421™ anti-human CD279 (PD-1)
1289 Antibody (BioLegend, 329920). Flow cytometry was performed on a CytoFLEX LE (Beckman
1290 Coulter, six-laser configuration). Gating strategy included sequential gating on FSC-A/SSC-A

1291 (lymphocytes), FSC-A/FSC-H (single cells), and exclusion of dead cells using viability dye.
1292 Fluorescence compensation was calculated using single-stained compensation beads (BioLegend).
1293 Data were initially analyzed with FlowJo v10.8, with gates being defined using isotype controls,
1294 followed by custom reproducing of the analysis for visualization purpose only.
1295
1296 **Surveyor assay for genome editing efficiency**
1297 To evaluate genome editing efficiency following RNP electroporation, genomic DNA was
1298 extracted from T cells 72 h post-nucleofection using the Monarch® Genomic DNA Purification
1299 Kit (New England Biolabs, T3010S), following the manufacturer's instructions. Genomic regions
1300 flanking the Cas9 cleavage site were PCR-amplified using 2× Phanta Max Master Mix (Vazyme,
1301 P525-02) in a 50 µl reaction volume (PCR primers provided in Supplementary File 2). PCR
1302 products were verified by agarose gel electrophoresis and purified using the Zymo DNA Clean &
1303 Concentrator Kit (Zymo Research, D4014), with final elution in 30 µl of nuclease-free water. To
1304 generate heteroduplex DNA for mismatch detection, 200 ng of purified PCR product was
1305 combined with 2 µl of 10× NEBuffer 2 (NEB) and nuclease-free water to a total volume of 19 µl.
1306 The mixture was denatured and reannealed in a thermocycler using the following program: 95 °C
1307 for 5 min; ramp at −2 °C/s to 85 °C; then ramp at −0.1 °C/s to 25 °C; and held at 4 °C. After
1308 annealing, 1 µl of T7 Endonuclease I (NEB, M0302) was added to each reaction and incubated at
1309 37 °C for 60 min. Reactions were terminated by adding 1.5 µl of 0.25 M EDTA and analyzed by
1310 electrophoresis on a 1% agarose gel to visualize cleavage products.
1311
1312 **mRNA knockdown analysis by quantitative PCR**
1313 To assess transcript-level gene knockdown following RNP electroporation, edited T cells were
1314 collected at the end point for RNA extraction. Total RNA was isolated using a RNeasy Mini Kit
1315 (Qiagen, 74104), and RNA concentration was quantified by nanodrop. Reverse transcription was
1316 performed using the HiScript® III 1st Strand cDNA Synthesis Kit (+gDNA wiper) (Vazyme,
1317 R312-01) according to the manufacturer's instructions. Gene-specific primers targeting the
1318 transcript of interest were used during amplification (Supplementary File 2). Quantitative PCR
1319 was performed using PowerUp™ SYBR® Green Master Mix (Thermo Fisher Scientific, A25742)
1320 on a CFX96 Real-Time PCR Detection System (Bio-Rad). ACTB (β-actin) was used as the internal
1321 control gene for normalization. Each sample was run in triplicates, and specificity of amplification
1322 was confirmed by melt-curve analysis. Relative gene expression was quantified using the ΔΔCt
1323 method, and results were reported as fold change compared to the control condition.
1324
1325 **RNA-seq analyses**
1326 Total RNA was extracted on the final day from ZNF766 knockout and NTC control groups using
1327 the RNeasy Mini Kit (Qiagen, 74004). Two biological replicates were collected per condition.
1328 RNA-seq libraries were generated with the NEBNext Ultra II Directional RNA Library Prep Kit
1329 (NEB, E7765L) following the manufacturer's poly(A) mRNA workflow, including enrichment
1330 with the NEBNext Poly(A) mRNA Magnetic Isolation Module (NEB, E7490L). Raw RNA-seq
1331 reads were mapped to the human reference genome (GRCh38/hg38) using STAR (v2.7.2a) with
1332 default settings. Strand-specific read counts were aggregated into a gene-level count matrix for

subsequent analyses. Transcriptomic variation across samples was evaluated by principal component analysis (PCA) based on the 500 most variable genes. Differential expression analysis was conducted using DESeq2 (v1.40.2), and genes with an absolute log2 fold change greater than 0.5 and an adjusted P-value below 0.05 were defined as significantly differentially expressed. Functional enrichment analyses were performed using clusterProfiler (v4.10.0), together with org.Hs.eg.db (v3.18.0), enrichplot (v1.22.0), and DOSE (v3.28.2). Significant genes were converted to Entrez Gene identifiers and assessed for enrichment in Gene Ontology biological processes and KEGG pathways. Enrichment results were visualized as dot plots (GO) and bar plots (KEGG).

Data Availability

All ATAC-seq, ChIP-seq used in this study are public data collected from the ENCODE data portal (https://www.encodeproject.org). T cell scATAC-seq data is obtained from Gene Expression Omnibus (GEO) under accession GSE129785. Purkinje neuron developmental scATAC-seq data is obtained from https://github.com/linnarsson-lab/fetal_brain_multiomics. CRISPR screen data is available on GEO under accession number GSE305611. RNA-seq data on ZNF766 KO is available on GEO under accession number GSE305610.

Code Availability

The code for Chromnitron is available at https://github.com/tanjimin/Chromnitron.

Competing Interests

J.T., X.F., X.L., B.X., and A.T. are inventors on provisional patents covering the models and tools reported herein. A.T. is a scientific advisor to Intelligencia AI. J.T. is a consultant for Flagship Pioneering. J.D.B is a Founder and Director of CDI Labs, Inc., a Founder of and consultant to Opentrons LabWorks/Neochromosome, Inc, and serves or served on the Scientific Advisory Board of CZ Biohub New York, LLC, Logomix, Inc., Rome Therapeutics, Inc., SeaHub, Seattle, WA, Tessera Therapeutics, Inc., and the Wyss Institute. All these services are unrelated to this work.

Acknowledgement

This work is supported by NIH grants DP5OD033430 and R03DE033371 (to B.X.), R01CA252239 and P01CA229086 (to A.T.). X.F. and R.R. are supported by R35CA253126 and P30CA013696. J.T. was supported by the Judith & Stewart Colton Center for Autoimmunity at NYU Langone and Arthritis National Research Foundation Grant Award (No. 1471427). X.L. is a Fellow of The Jane Coffin Childs Fund for Medical Research. B.X. is a Junior Fellow of the Harvard Society of Fellows. We thank the High Performance Computing Core Facility at NYU School of Medicine and the Broad Institute for the support of computation. We thank the Broad Genetic Perturbation Platform and Broad Clinical Labs for the support of genetic and genomic experiments. We also acknowledge Milk Bar, where Xi and Jimin enjoyed coffee and met weekly for over two years to work on Chromnitron. We thank Huiyuan Zhang, Chenzhen Zhang, Hanqing Liu, Lloyd Bod, Jason D. Buenrostro, Romanos S. Pistofidis, Liz Gaskell, Wenke Liu, Jose U. Scher, and the members of the Xia lab and Tsirigos lab for their valuable suggestions and discussion. Lastly, we are grateful to the ENCODE consortium for generating, organizing, and sharing standardized genomics data that enabled this study.

Author Contribution

B.X. and J.T. conceived the project. J.T., B.X. and A.T. designed the computational and experimental components and interpreted the results with inputs from X.F. and X.L. J.T., F.X. and S.M. designed, implemented and optimized the neural network. J.T. built the preprocessing platform with input from X.F. X.F. performed single cell data preprocessing and protein feature analysis. J.T. trained the model, performed inference and downstream computational analysis with input from X.F, B.X. and A.T. J.B. performed RNA-seq analysis. X.L. performed the experimental

components with help from J.B. R.R., D.F., and J.D.B. provided feedback and contributed to discussion. J.T. and B.X. led the writing of the manuscript with inputs from X.F., X.L. and A.T. All authors commented on the manuscript.

1388 **References**

1389 1. Arendt, D. *et al.* The origin and evolution of cell types. *Nat. Rev. Genet.* **17**, 744–757 (2016).

1390 2. Xia, B. & Yanai, I. A periodic table of cell types. *Dev. Camb. Engl.* **146**, dev169854 (2019).

1391 3. Gurdon, J. B. Cell Fate Determination by Transcription Factors. *Curr. Top. Dev. Biol.* **116**,
1392 445–454 (2016).

1393 4. Chen, X. *et al.* Integration of external signaling pathways with the core transcriptional
1394 network in embryonic stem cells. *Cell* **133**, 1106–1117 (2008).

1395 5. Boyer, L. A. *et al.* Core transcriptional regulatory circuitry in human embryonic stem cells.
1396 *Cell* **122**, 947–956 (2005).

1397 6. Lander, E. S. *et al.* Initial sequencing and analysis of the human genome. *Nature* **409**, 860–
1398 921 (2001).

1399 7. Lambert, S. A. *et al.* The Human Transcription Factors. *Cell* **172**, 650–665 (2018).

1400 8. Vaquerizas, J. M., Kummerfeld, S. K., Teichmann, S. A. & Luscombe, N. M. A census of
1401 human transcription factors: function, expression and evolution. *Nat. Rev. Genet.* **10**, 252–263
1402 (2009).

1403 9. Gerstein, M. B. *et al.* Architecture of the human regulatory network derived from ENCODE
1404 data. *Nature* **489**, 91–100 (2012).

1405 10. Li, Y. E. *et al.* A comparative atlas of single-cell chromatin accessibility in the human
1406 brain. *Science* **382**, eadf7044 (2023).

1407 11. Weirauch, M. T. *et al.* Determination and inference of eukaryotic transcription factor
1408 sequence specificity. *Cell* **158**, 1431–1443 (2014).

1409 12. Jolma, A. *et al.* DNA-binding specificities of human transcription factors. *Cell* **152**, 327–
1410 339 (2013).

1411 13. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints.
1412 *Nature* **583**, 729–736 (2020).
1413 14. Murre, C. Helix-loop-helix proteins and the advent of cellular diversity: 30 years of
1414 discovery. *Genes Dev.* **33**, 6–25 (2019).
1415 15. Doughty, B. R. *et al.* Single-molecule states link transcription factor binding to gene
1416 expression. *Nature* **636**, 745–754 (2024).
1417 16. Durdu, S. *et al.* Chromatin-dependent motif syntax defines differentiation trajectories.
1418 *Mol. Cell* **85**, 2900-2918.e16 (2025).
1419 17. Kim, S. & Wysocka, J. Deciphering the multi-scale, quantitative cis-regulatory code.
1420 *Mol. Cell* **83**, 373–392 (2023).
1421 18. Kim, S. *et al.* DNA-guided transcription factor cooperativity shapes face and limb
1422 mesenchyme. *Cell* **187**, 692-711.e26 (2024).
1423 19. Statello, L., Guo, C.-J., Chen, L.-L. & Huarte, M. Gene regulation by long non-coding
1424 RNAs and its biological functions. *Nat. Rev. Mol. Cell Biol.* **22**, 96–118 (2021).
1425 20. Smith, E. & Shilatifard, A. The chromatin signaling pathway: diverse mechanisms of
1426 recruitment of histone-modifying enzymes and varied biological outcomes. *Mol. Cell* **40**, 689–
1427 701 (2010).
1428 21. Yosef, N. *et al.* Dynamic regulatory network controlling TH17 cell differentiation.
1429 *Nature* **496**, 461–468 (2013).
1430 22. Johnson, D. S., Mortazavi, A., Myers, R. M. & Wold, B. Genome-wide mapping of in
1431 vivo protein-DNA interactions. *Science* **316**, 1497–1502 (2007).
1432 23. Tsankov, A. M. *et al.* Transcription factor binding dynamics during human ES cell
1433 differentiation. *Nature* **518**, 344–349 (2015).

1434 24. Park, P. J. ChIP-seq: advantages and challenges of a maturing technology. *Nat. Rev.*
1435 *Genet.* **10**, 669–680 (2009).
1436 25. Skene, P. J. & Henikoff, S. An efficient targeted nuclease strategy for high-resolution
1437 mapping of DNA binding sites. *eLife* **6**, e21856 (2017).
1438 26. Partridge, E. C. *et al.* Occupancy maps of 208 chromatin-associated proteins in one
1439 human cell type. *Nature* **583**, 720–728 (2020).
1440 27. Hammal, F., de Langen, P., Bergon, A., Lopez, F. & Ballester, B. ReMap 2022: a
1441 database of Human, Mouse, Drosophila and Arabidopsis regulatory regions from an
1442 integrative analysis of DNA-binding sequencing experiments. *Nucleic Acids Res.* **50**, D316–
1443 D325 (2022).
1444 28. Moyers, B. A. *et al.* Characterization of human transcription factor function and patterns
1445 of gene regulation in HepG2 cells. *Genome Res.* **33**, 1879–1892 (2023).
1446 29. Wang, J. *et al.* Sequence features and chromatin structure around the genomic regions
1447 bound by 119 human transcription factors. *Genome Res.* **22**, 1798–1812 (2012).
1448 30. Avsec, Ž. *et al.* Base-resolution models of transcription-factor binding reveal soft motif
1449 syntax. *Nat. Genet.* **53**, 354–366 (2021).
1450 31. Cazares, T. A. *et al.* maxATAC: Genome-scale transcription-factor binding prediction
1451 from ATAC-seq with deep neural networks. *PLOS Comput. Biol.* **19**, e1010863 (2023).
1452 32. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting RNA-seq
1453 coverage from DNA sequence as a unifying model of gene regulation. *Nat. Genet.* **57**, 949–
1454 961 (2025).
1455 33. Avsec, Ž. *et al.* AlphaGenome: advancing regulatory variant effect prediction with a
1456 unified DNA sequence model.

34. Lin, Z. *et al.* Evolutionary-scale prediction of atomic-level protein structure with a language model.

35. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).

36. Nguyen, E. *et al.* Sequence modeling and design from molecular to genome scale with Evo. *Science* **386**, eado9336 (2024).

37. Fu, X. *et al.* A foundation model of transcription across human cell types. *Nature* (2025) doi:10.1038/s41586-024-08391-z.

38. Hayes, T. *et al.* Simulating 500 million years of evolution with a language model. Preprint at https://doi.org/10.1101/2024.07.01.600583 (2024).

39. Theodoris, C. V. *et al.* Transfer learning enables predictions in network biology. *Nature* **618**, 616–624 (2023).

40. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at https://doi.org/10.48550/arXiv.2106.09685 (2021).

41. Bravo González-Blas, C. *et al.* SCENIC+: single-cell multiomic inference of enhancers and gene regulatory networks. *Nat. Methods* **20**, 1355–1367 (2023).

42. Hu, Y. *et al.* Multiscale footprints reveal the organization of cis-regulatory elements. *Nature* **638**, 779–786 (2025).

43. Pampari, A. *et al.* ChromBPNet: bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. Preprint at https://doi.org/10.1101/2024.12.25.630221 (2024).

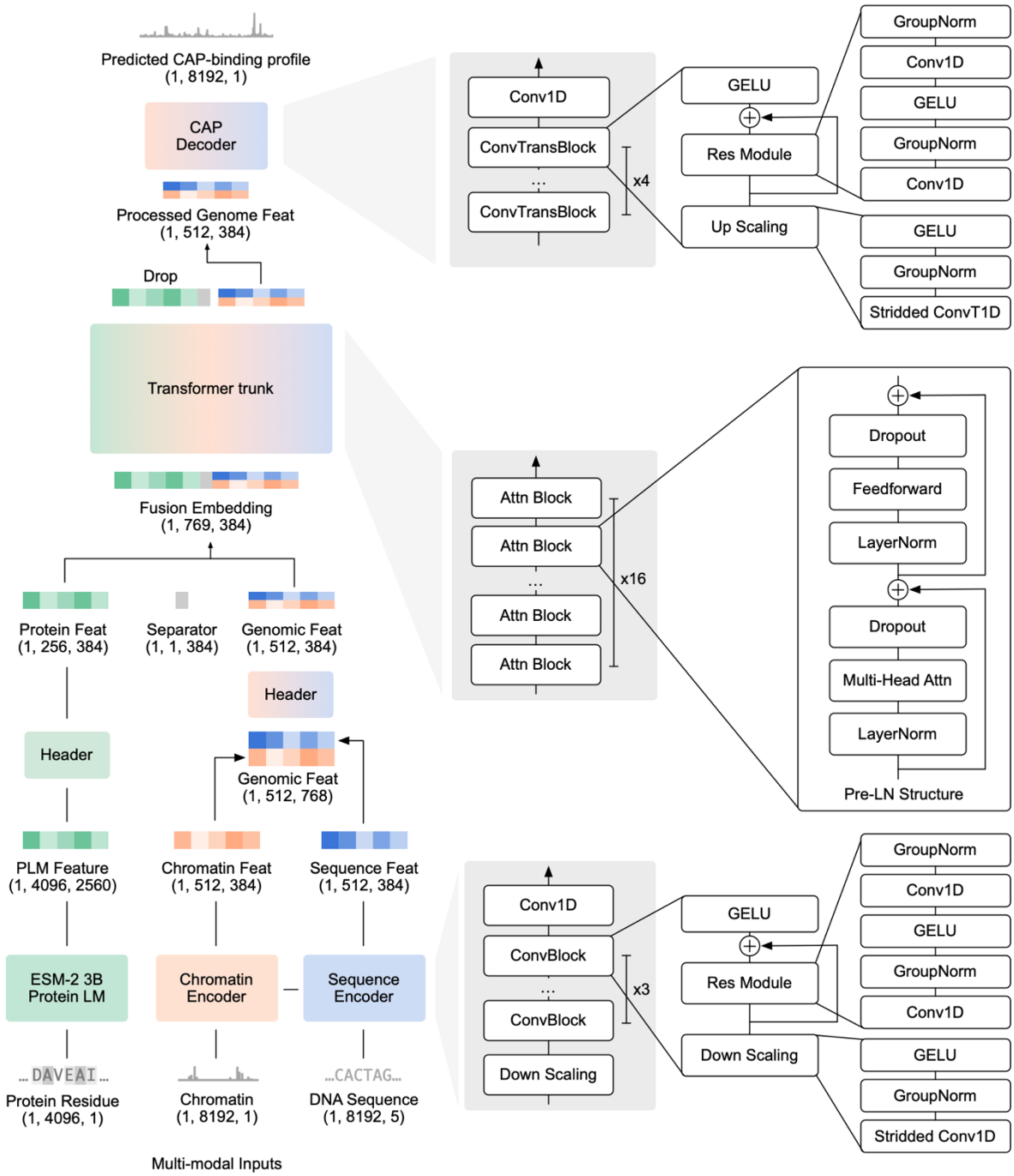
44. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat. Methods* **18**, 1196–1203 (2021).

1480 45. Stoltenburg, R., Reinemann, C. & Strehlitz, B. SELEX—A (r)evolutionary method to
1481 generate high-affinity nucleic acid ligands. *Biomol. Eng.* **24**, 381–403 (2007).
1482 46. Ren, B. *et al.* Genome-Wide Location and Function of DNA Binding Proteins. *Science*
1483 **290**, 2306–2309 (2000).
1484 47. Khan, O. *et al.* TOX transcriptionally and epigenetically programs CD8+ T cell
1485 exhaustion. *Nature* **571**, 211–218 (2019).
1486 48. Seo, H. *et al.* TOX and TOX2 transcription factors cooperate with NR4A transcription
1487 factors to impose CD8 + T cell exhaustion. *Proc. Natl. Acad. Sci.* **116**, 12410–
1488 12415 (2019).
1489 49. Yao, C. *et al.* Single-cell RNA-seq reveals TOX as a key regulator of CD8+ T cell
1490 persistence in chronic infection. *Nat. Immunol.* **20**, 890–901 (2019).
1491 50. Mann, T. H. & Kaech, S. M. Tick-TOX, it's time for T cell exhaustion. *Nat. Immunol.*
1492 **20**, 1092–1094 (2019).
1493 51. Scott, A. C. *et al.* TOX is a critical regulator of tumour-specific T cell differentiation.
1494 *Nature* **571**, 270–274 (2019).
1495 52. Alfei, F. *et al.* TOX reinforces the phenotype and longevity of exhausted T cells in
1496 chronic viral infection. *Nature* **571**, 265–269 (2019).
1497 53. Hashimoto, M. *et al.* CD8 T Cell Exhaustion in Chronic Infection and Cancer:
1498 Opportunities for Interventions. *Annu. Rev. Med.* **69**, 301–318 (2018).
1499 54. Baessler, A. & Vignali, D. A. A. T Cell Exhaustion. *Annu. Rev. Immunol.* **42**, 179–206
1500 (2024).

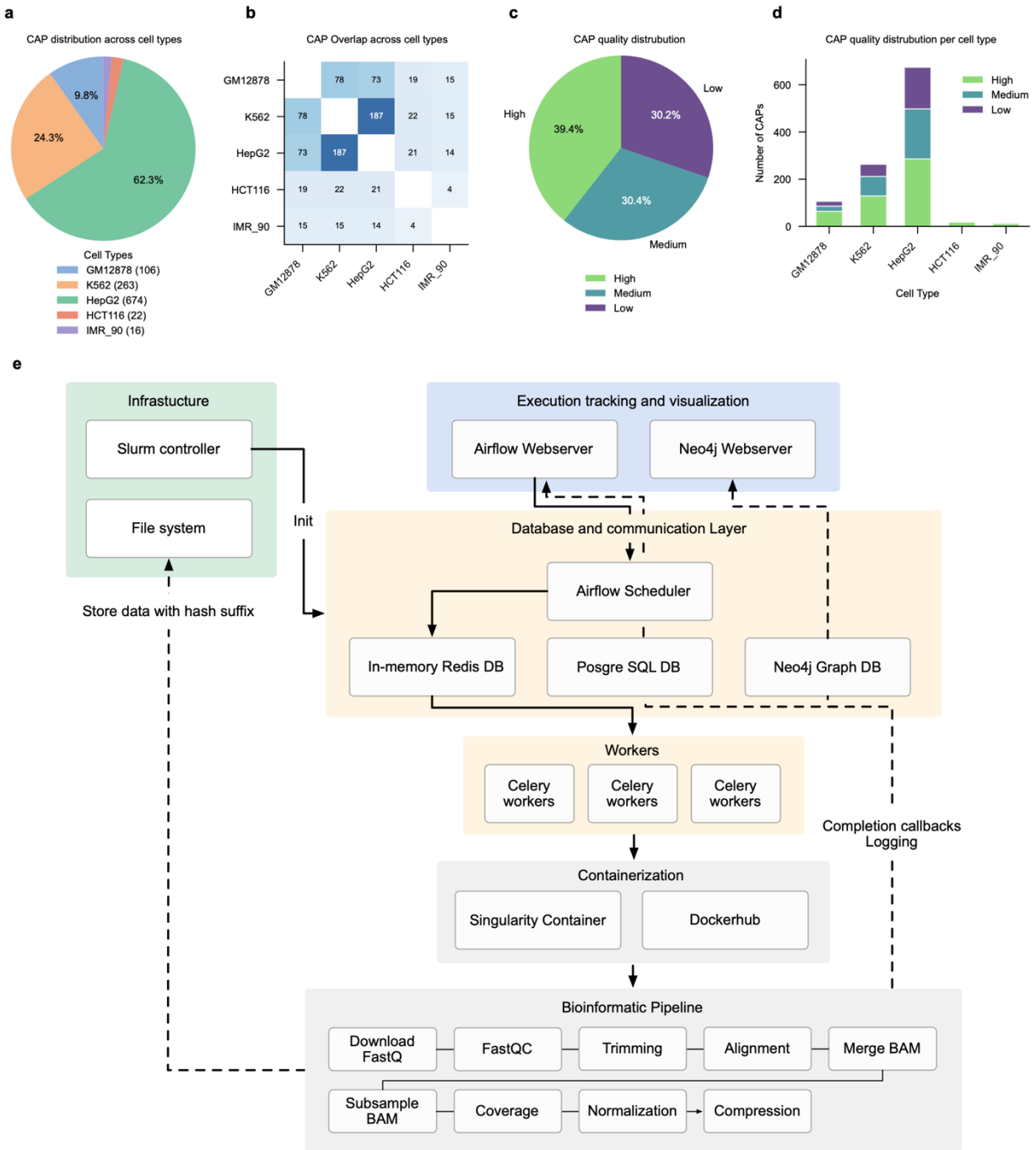
1501 55. Satpathy, A. T. *et al.* Massively parallel single-cell chromatin landscapes of human
1502 immune cell development and intratumoral T cell exhaustion. *Nat. Biotechnol.* **37**, 925–936
1503 (2019).
1504 56. Jenkins, E., Whitehead, T., Fellermeyer, M., Davis, S. J. & Sharma, S. The current state
1505 and future of T-cell exhaustion research. *Oxf. Open Immunol.* **4**, iqad006 (2023).
1506 57. Tay, T. *et al.* Degradation of IKZF1 prevents epigenetic progression of T cell exhaustion
1507 in an antigen-specific assay. *Cell Rep. Med.* **5**, 101804 (2024).
1508 58. Belk, J. A. *et al.* Genome-wide CRISPR screens of T cell exhaustion identify chromatin
1509 remodeling factors that limit T cell persistence. *Cancer Cell* **40**, 768-786.e7 (2022).
1510 59. Wu, J. E. *et al.* In vitro modeling of CD8+ T cell exhaustion enables CRISPR screening
1511 to reveal a role for BHLHE40. *Sci. Immunol.* **8**, eade3369 (2023).
1512 60. Mannens, C. C. A. *et al.* Chromatin accessibility during human first-trimester
1513 neurodevelopment. *Nature* (2024) doi:10.1038/s41586-024-07234-1.
1514 61. Chen, K. M., Wong, A. K., Troyanskaya, O. G. & Zhou, J. A sequence-based global map
1515 of regulatory activity for deciphering human genetics. *Nat. Genet.* **54**, 940–949 (2022).
1516 62. Hendrycks, D. & Gimpel, K. Gaussian Error Linear Units (GELUs). Preprint at
1517 https://doi.org/10.48550/arXiv.1606.08415 (2023).
1518 63. Vaswani, A. *et al.* Attention Is All You Need. *ArXiv170603762 Cs* (2017).
1519 64. Meng, Q. *et al.* A generic reference defined by consensus peaks for scATAC-seq data
1520 analysis. Preprint at https://doi.org/10.1101/2023.05.30.542889 (2023).
1521 65. Amemiya, H. M., Kundaje, A. & Boyle, A. P. The ENCODE Blacklist: Identification of
1522 Problematic Regions of the Genome. *Sci. Rep.* **9**, 9354 (2019).

1523 66. Kingma, D. P. & Ba, J. Adam: A Method for Stochastic Optimization. Preprint at
1524 https://doi.org/10.48550/arXiv.1412.6980 (2017).
1525 67. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at
1526 https://doi.org/10.48550/arXiv.2106.09685 (2021).
1527 68. Shrikumar, A. *et al.* Technical Note on Transcription Factor Motif Discovery from
1528 Importance Scores (TF-MoDISco) version 0.5.6.5. Preprint at
1529 https://doi.org/10.48550/arXiv.1811.00416 (2020).
1530 69. Blum, M. *et al.* InterPro: the protein sequence classification resource in 2025. *Nucleic*
1531 *Acids Res.* **53**, D444–D456 (2025).
1532 70. Ikeuchi, T., Ide, M., Zeng, Y., Maeda, T. N. & Shimizu, S. Python package for causal
1533 discovery based on LiNGAM.
1534 71. Szklarczyk, D. *et al.* The STRING database in 2023: protein–protein association
1535 networks and functional enrichment analyses for any sequenced genome of interest. *Nucleic*
1536 *Acids Res.* **51**, D638–D646 (2023).
1537 72. Wolf, F. A., Angerer, P. & Theis, F. J. SCANPY: large-scale single-cell gene expression
1538 data analysis. *Genome Biol.* **19**, 15 (2018).
1539 73. Robinson, J. T. *et al.* Integrative genomics viewer. *Nat. Biotechnol.* **29**, 24–26 (2011).
1540 74. Li, W. *et al.* MAGeCK enables robust identification of essential genes from genome-
1541 scale CRISPR/Cas9 knockout screens. *Genome Biol.* **15**, 554 (2014).
1542

Supplementary Figures

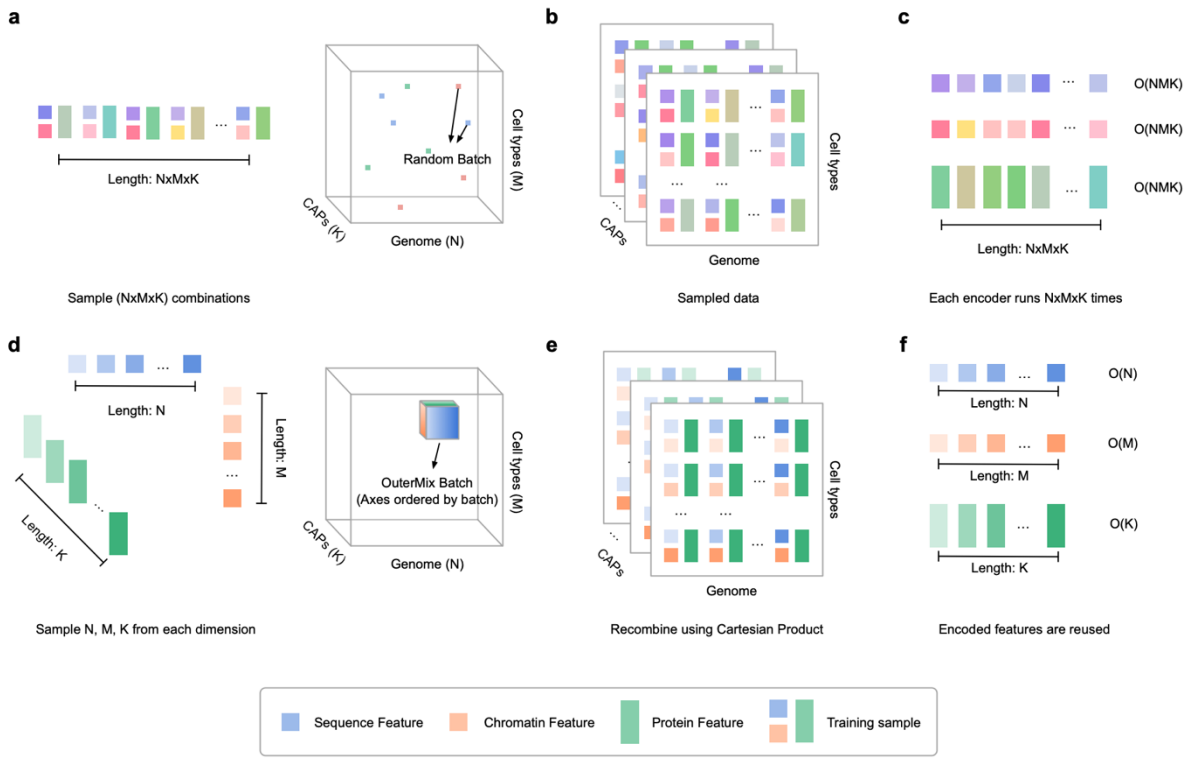


Supplementary Fig. 1: Chromnitron model architecture.
Left: overall model structure, convolution and transformer modules, and layer configurations. The feature size of embeddings (including batch as the first dimension) are denoted in brackets.



Supplementary Fig. 2: Training data statistics and the Chrom2Vec platform.
a. CAP-binding ChIP-seq datasets and the corresponding percentages across five human cell lines.
b. Heatmap of pairwise overlaps of CAP-binding datasets across cell lines. Diagonal entries are removed.
c. CAP categorization by their Chromnitron-prediction performance.
d. Distribution of CAPs across prediction-performance categories in each cell type.
e. Workflow of Chrom2Vec platform for processing training datasets. Detailed workflow was also described in Methods sections. Slurm workload controller was used to initialize schedulers

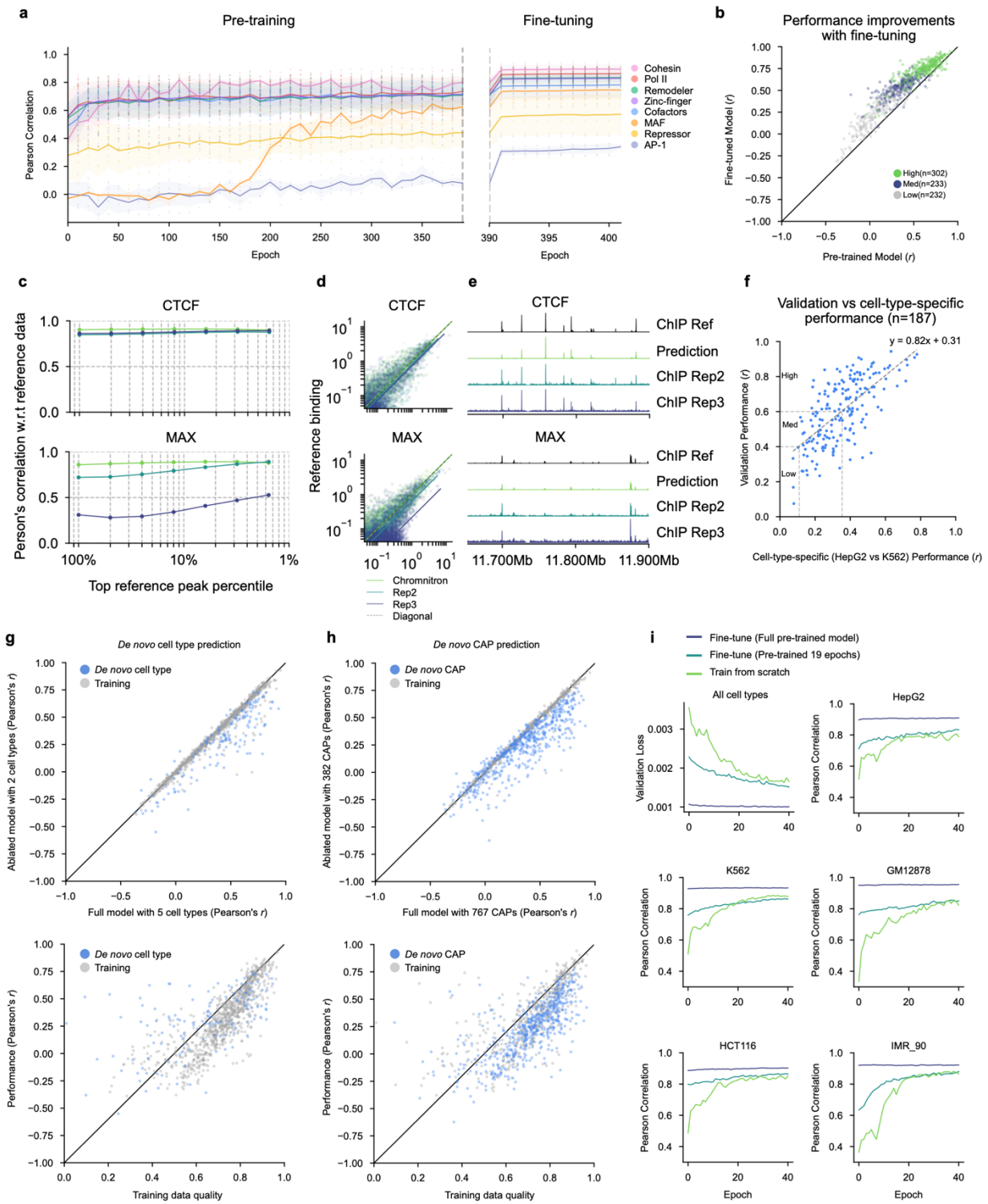
tracking tools including Airflow and Neo4j webservers. Databases used for information relay (Redis in-memory DB, PostgreSQL DB, Neo4j graph DB) were initialized separately. Celery workers perform the containerized (Singularity, DockerHub) bioinformatic pipeline, including Download FastQ, FastQC quality control, Trimming, Alignment, Merge BAM, Subsample, Coverage, Normalization, and Compression steps. Each action includes a completion callback that logs through Airflow to the Neo4j database.



Supplementary Fig. 3: OuterMix algorithm increases multimodal training efficiency via factorized batch sampling.

a-c. Sampling and computation with random sampling strategy. A randomly sampled batch has $N \times M \times K$ combinations of genome sequence (N), chromatin feature (M), and CAP (K), selected from the full three-dimensional data manifold (**a**). This random process generates large data matrices for encoding (**b**). Thus, each encoder for the corresponding modality would process every combination independently, yielding a computational complexity of $O(N \times M \times K)$ per encoder (**c**).

d-f. Sampling and computation with OuterMix strategy. OuterMix first samples *N, M, K* training samples from each dimension, effectively sampling a dense data cube from the full three-dimensional data manifold (**d**). Each sampling was encoded, with the encoded embeddings being recombined via a Cartesian product to form $N \times M \times K$ training samples (**e**). These encoded features are reused in the final Cartesian product and avoids redundant computation, reducing encoder computation complexity down to $O(N) + O(M) + O(K)$.



Supplementary Fig. 4: Chromnitron model performance evaluation and ablation studies. **a.** Model performance across pre-training and fine-tuning epochs, grouped by CAP categories. Line and the shade represent scores of mean ± s.d. for each CAP category. **b.** Scatter plot of model performance (Pearson's correlation between prediction and experimental data of individual CAPs) between fine-tuned model and pre-trained model.

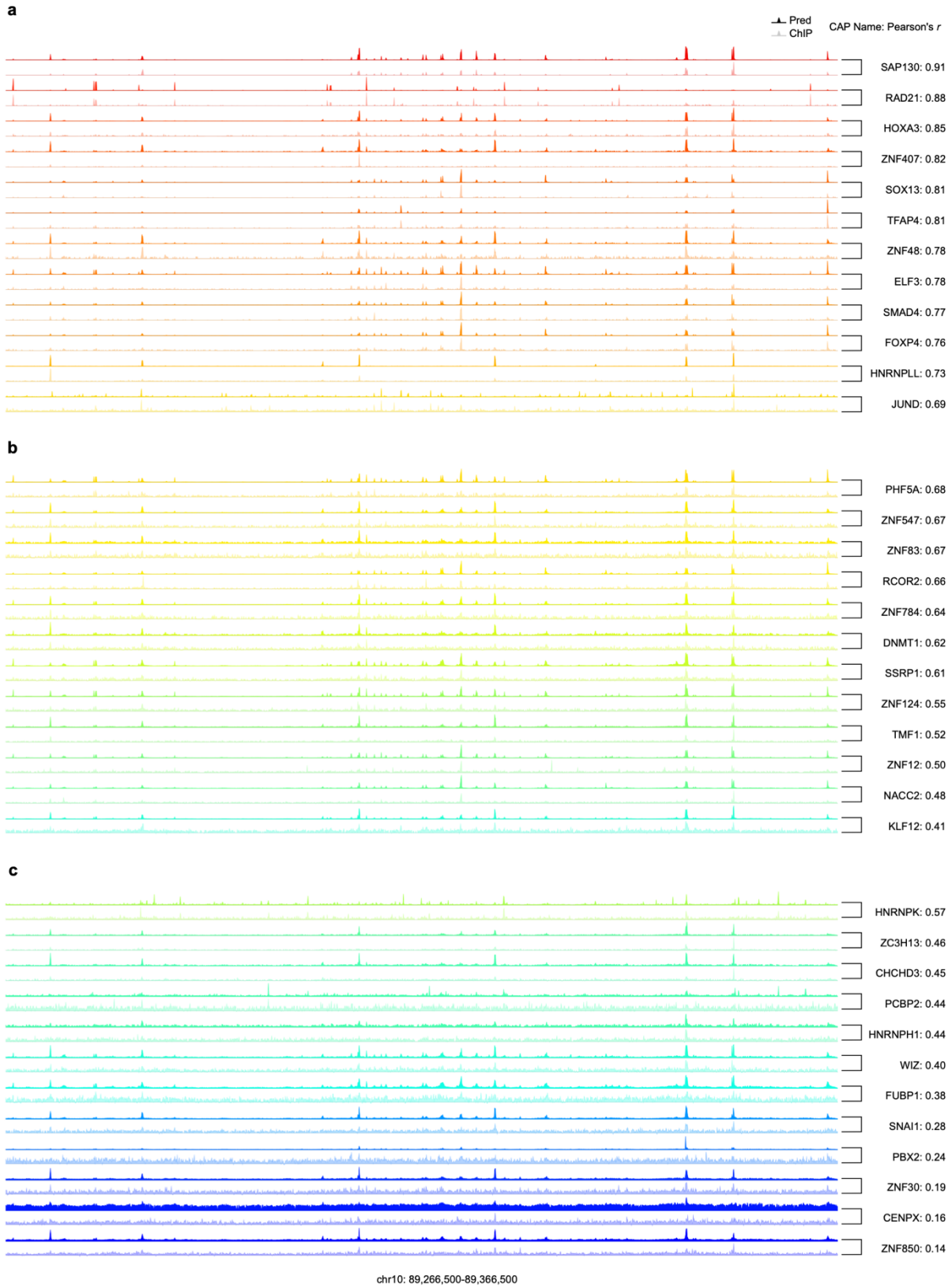
c-e. Evaluation of Chromnitron predictions against experimental ChIP-seq replicates for CTCF (top) and MAX (bottom). Each evaluation was performed by comparing prediction performance with that of replicating experimental measurements. These evaluations include: (1) calculating Pearson's correlation to the reference experimental data across different percentiles of genomic bins (**c**), (2) scatter plot visualization of predicted or experimental replicate binding signals (x-axis) to the reference experimental data (y-axis, **d**), and (3) direct visualization of CAP-binding tracks of prediction alongside two other experimental replicates (Rep2, Rep3), all aligning to the reference experimental data (Ref, **e**).

f. Establishment of CAP performance groups. Two cutoffs (0.4 and 0.6) based on validation performance were selected on the regression line between validation performance and cell-type-specific performance. Regression coefficient and offset is shown. Areas split by the horizontal and vertical line segments represent regions for high, medium, and low performance CAP groups (annotated on the right of y axis). The grouping rules were later generalized to all CAPs.

g. Ablation study to test *de novo* prediction of CAP-binding in unseen cell types. Ablation-study model was trained with ChIP-seq data from only HepG2 and K562 cells. Top: scatter plot of prediction performance (Pearson's *r*) comparing ablation-study model trained using only two cell types (HepG2 and K562) versus the full Chromnitron model trained on all five cell types (HepG2, K562, GM12878, IMR-90, HCT116). Bottom: performance of ablation-study model versus training data quality. Blue and grey points represent CAPs in *de novo* cell types and in training cell types, respectively.

h. Ablation study to test *de novo* CAP prediction. Ablation-study model was trained with ChIP-seq corresponding to 385 CAPs. Top: scatter plot of ablation-study model performance for model trained with 385 CAPs versus the full Chromnitron model trained on all 767 CAPs. Bottom: performance of ablation-study model versus data quality. Blue and grey points represent *de novo* CAPs and training CAPs, respectively.

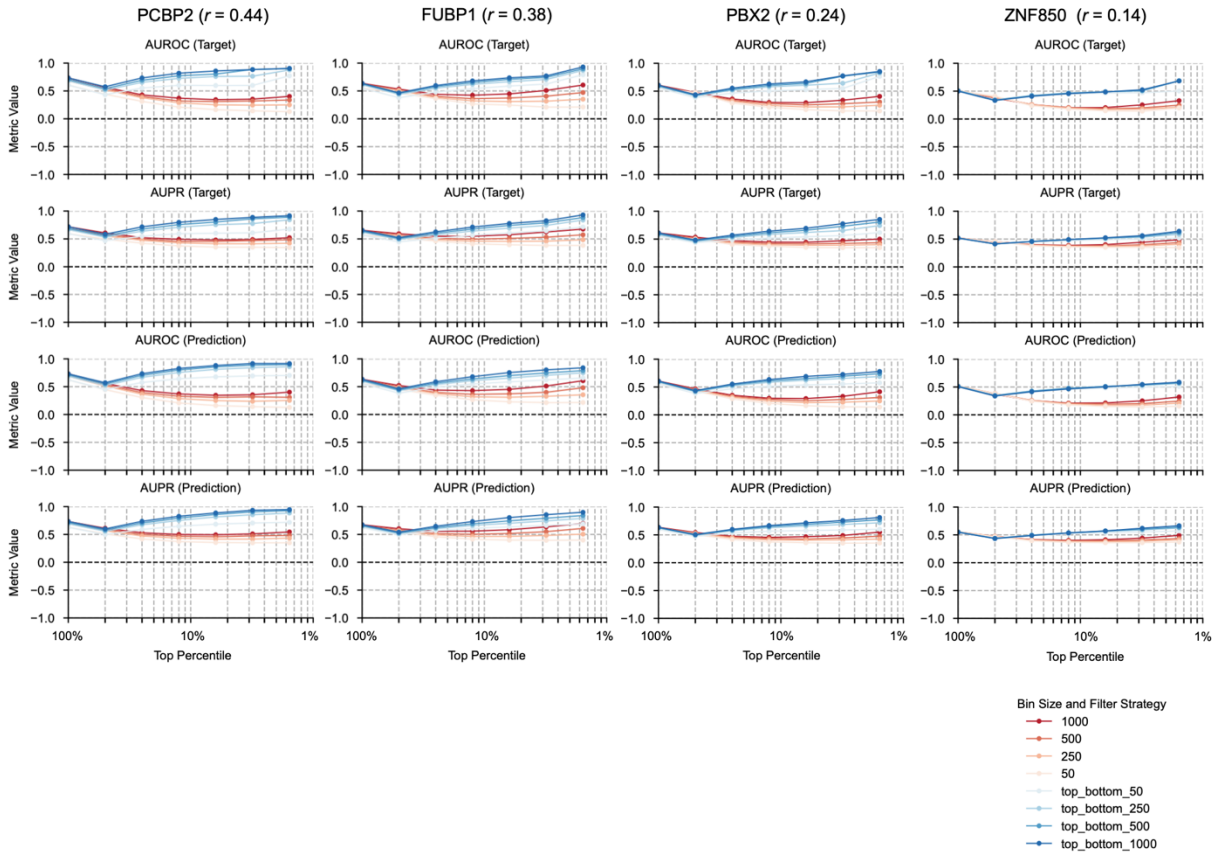
i. Training curves comparisons across three model-training strategies: fine-tuning a fully pre-trained model (dark blue), fine-tuning after 19 epochs pre-training (light blue), and training from scratch (green). The comparison includes overall validation loss (across all cell types) and Pearson's *r* relating to reference data in each cell type across training epochs.


a
Pred
ChIP
CAP Name: Pearson's r
SAP130: 0.91
RAD21: 0.88
HOXA3: 0.85
ZNF407: 0.82
SOX13: 0.81
TFAP4: 0.81
ZNF48: 0.78
ELF3: 0.78
SMAD4: 0.77
FOXP4: 0.76
HNRNPLL: 0.73
JUND: 0.69
b
PHF5A: 0.68
ZNF547: 0.67
ZNF83: 0.67
RCOR2: 0.66
ZNF784: 0.64
DNMT1: 0.62
SSRP1: 0.61
ZNF124: 0.55
TMF1: 0.52
ZNF12: 0.50
NACC2: 0.48
KLF12: 0.41
c
HNRNPK: 0.57
ZC3H13: 0.46
CHCHD3: 0.45
PCBP2: 0.44
HNRNPH1: 0.44
WIZ: 0.40
FUBP1: 0.38
SNAI1: 0.28
PBX2: 0.24
ZNF30: 0.19
CENPX: 0.16
ZNF850: 0.14
chr10: 89,266,500-89,366,500

Supplementary Fig. 5: Paired visualization of Chromnitron-predicted and ChIP-seq-measured CAP binding profiles.

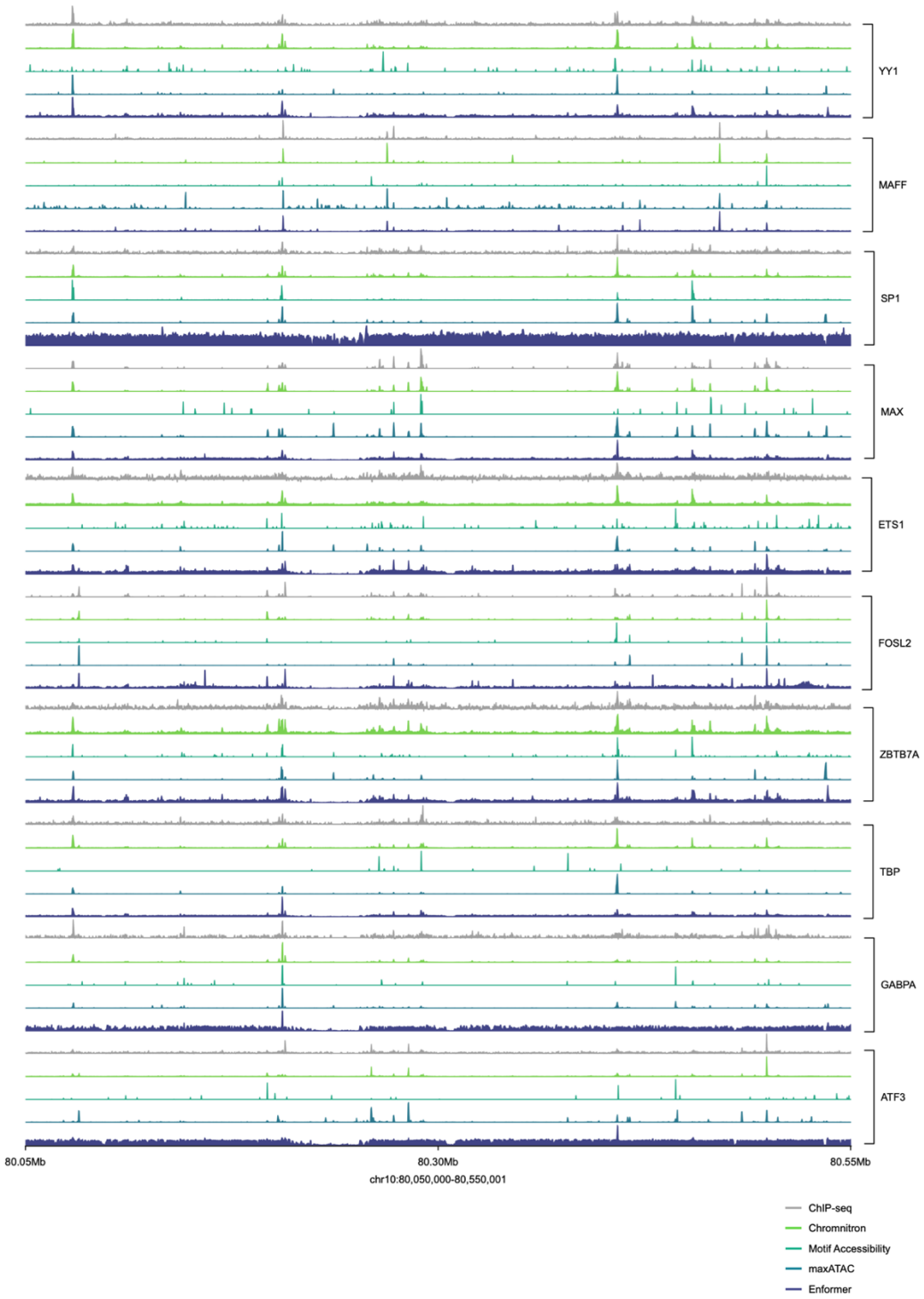
a. Paired CAP-binding tracks of prediction (top) and experimental data (bottom) for 12 randomly selected high-performance CAPs across the leave-out genomic region (chr 10). CAP-binding tracks were colored by their descending performance scores (Pearson's *r*).

b-c. Paired CAP-binding tracks for randomly selected medium-performance CAPs (**b**) and low-performance CAPs (**c**). Visualizations followed the same pattern as presented in **a**.

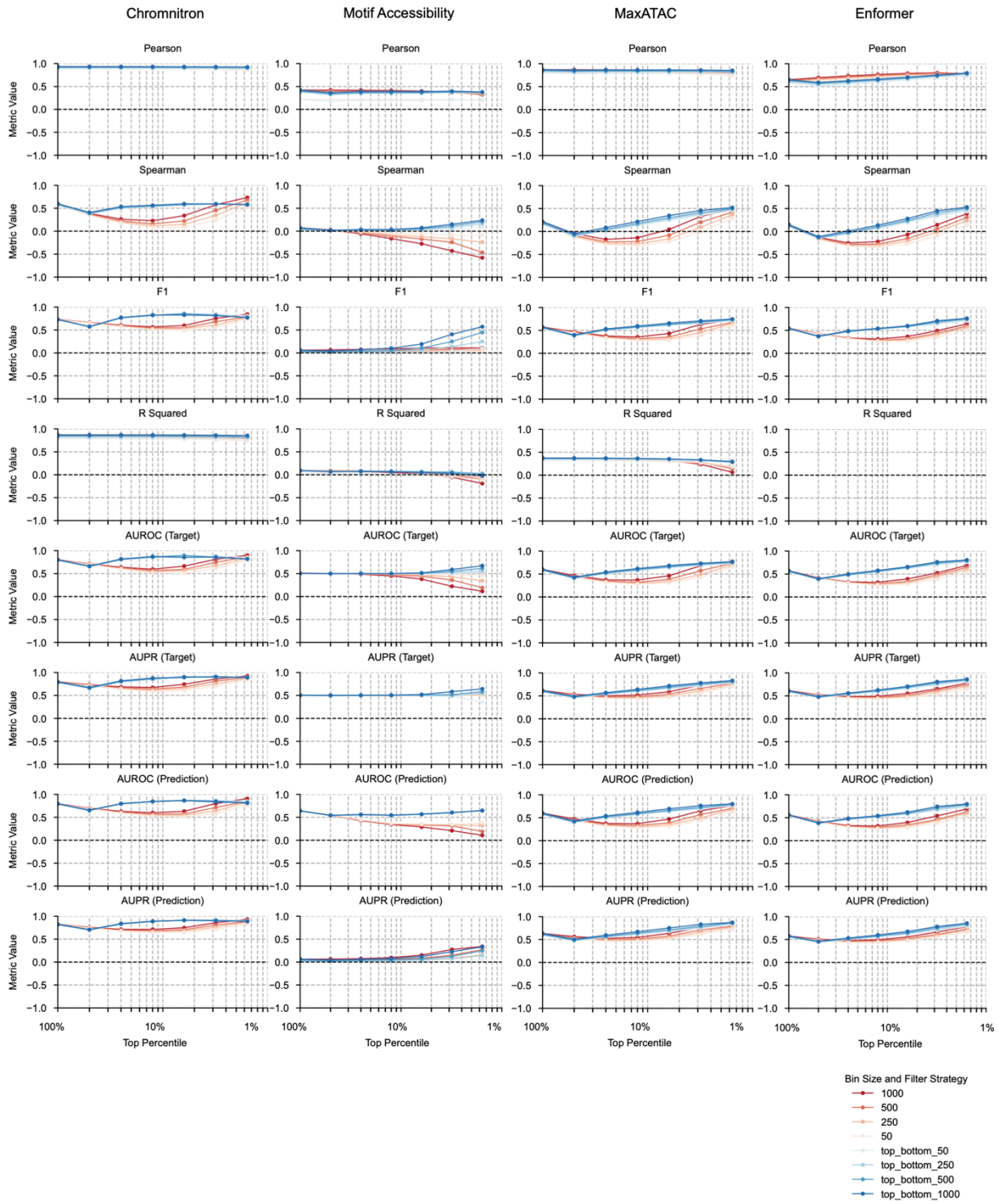


Supplementary Fig. 6: Peak detection-based performance evaluation for low-correlation CAPs.

Performance metrics for four randomly selected low-correlation (Pearson's $r < 0.5$) CAPs, PCBP2, FUBP1, PBX2, ZNF850. Pearson's correlation was calculated by comparing to reference experimental ChIP-seq data. For each CAP, four statistics were calculated, including area under the receiver operating characteristic curve (AUROC) and area under the precision-recall curve (AUPR) using both target-centric and prediction-centric labelling. Metrics were calculated using bin sizes across 50, 250, 500, and 1,000 bp (color intensity denotes bin size). Red curves indicate selection of bins from the top percentile only, while blue curves indicate selection of bins from both the top and bottom percentiles of bins.



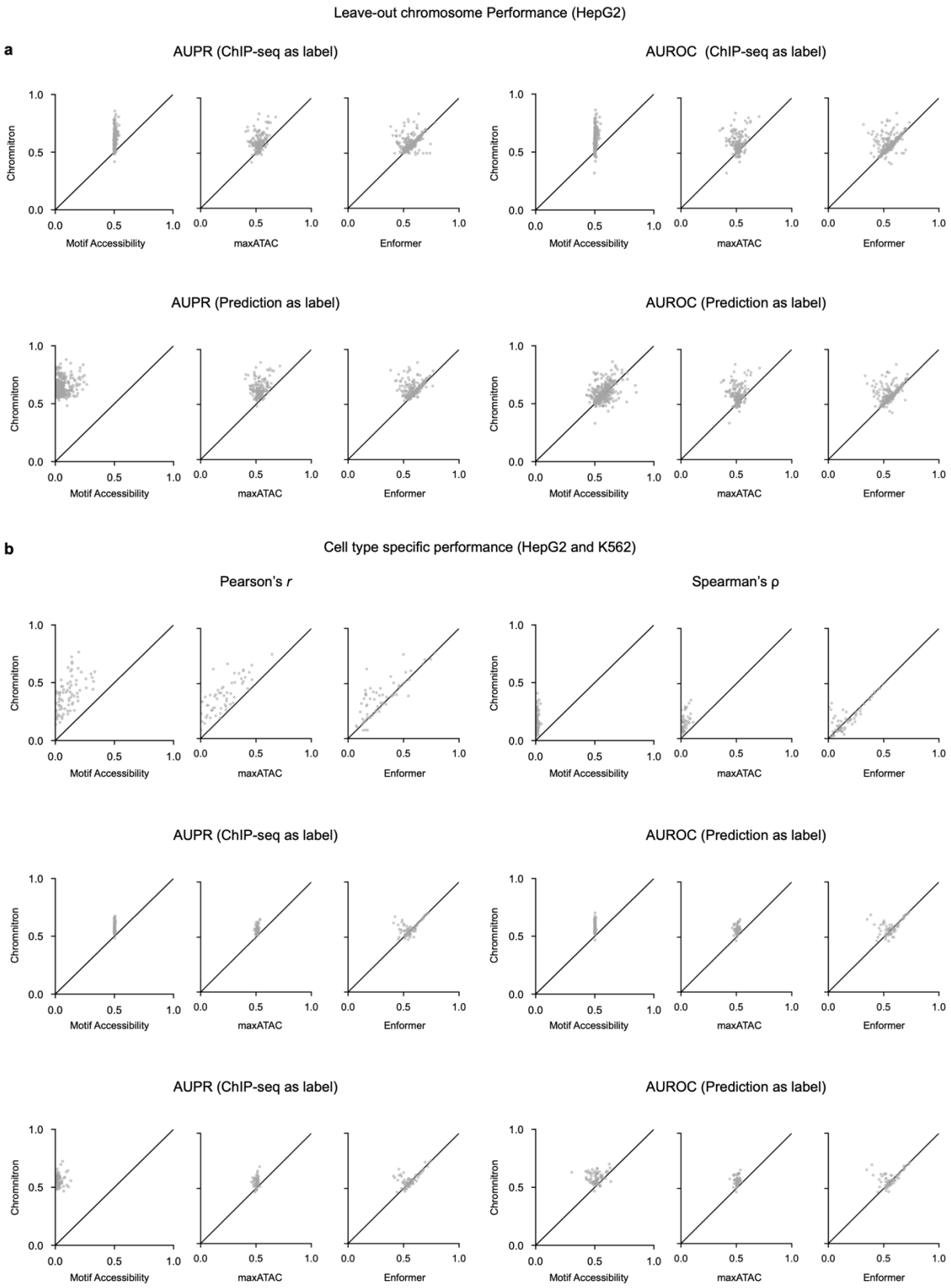
Supplementary Fig. 7: Performance comparison of CAP binding predictions. Relating to Fig. 2h, 10 more randomly selected CAPs were presented at the same 500kb genomic window in the leave-out chromosome 10. Each CAP includes predicted binding tracks from Chromnitron, motif accessibility, maxATAC and Enformer models, all referring to the experimental ChIP-seq data in leave-out chromosome 10. Min-max scaling of the binding signal was applied to each track separately.



Supplementary Fig. 8: Performance metrics comparison between Chromnitron and three other models, using MAX as an example.

All performance metrics were performed against reference experimental ChIP-seq data of MAX. In total, eight statistics were calculated, including Pearson's correlation (r), Spearman's correlation (ρ), the coefficient of determination (R squared), F1 score, AUROC and AUPR

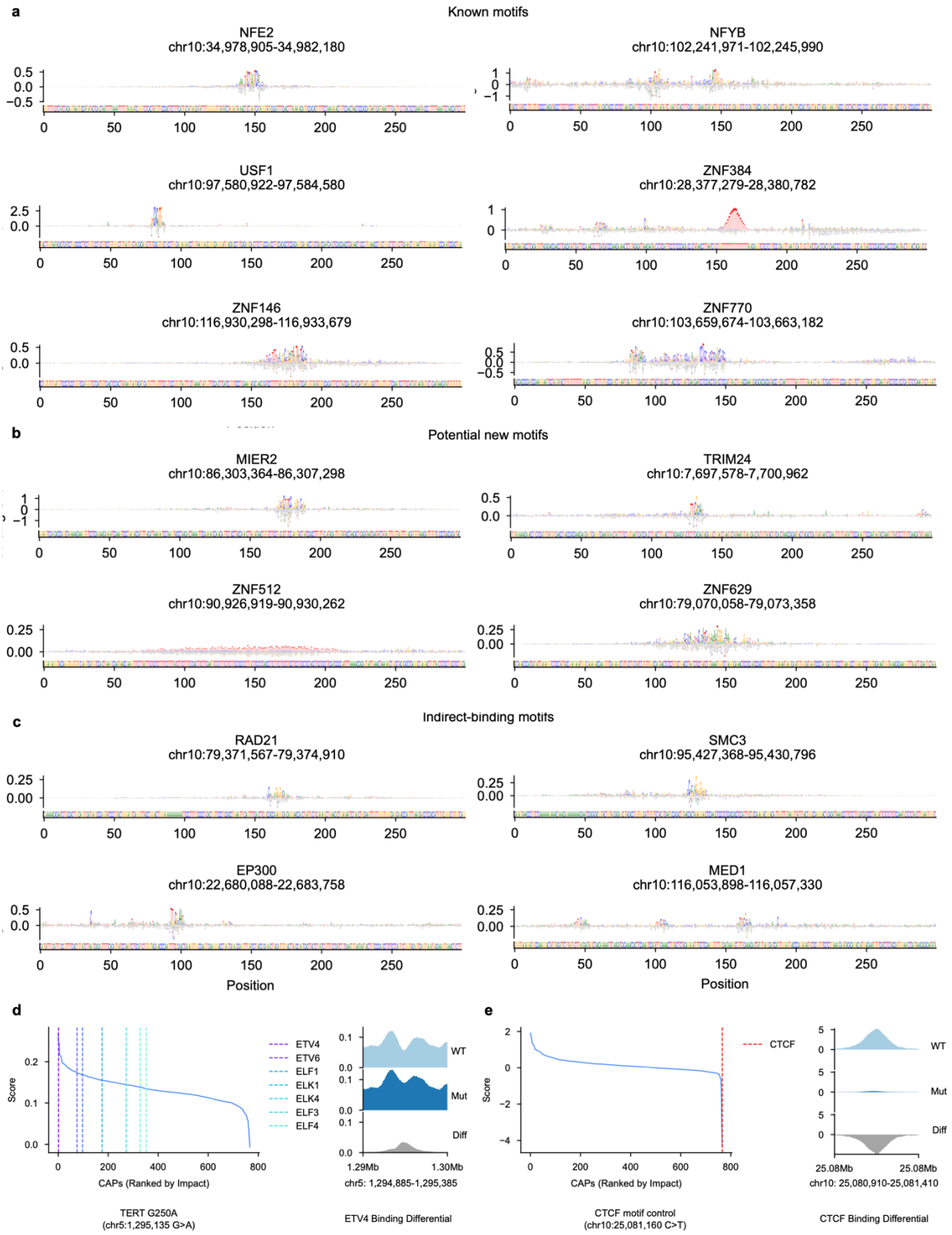
curves using both target-centric and prediction-centric labelling. Each of these methods was calculated using bins sizes of 50, 250, 500, and 1,000 bp (color intensity denotes bin size). Each plot shows a performance metric (y-axis) dynamic as the different percentages of genomic bins included in the analysis (x-axis, log-scaled). Red lines indicate selection of bins from the top percentiles only, while blue lines indicate selection of bins from both the top and bottom percentiles of bins.



Supplementary Fig. 9: Global statistics of Chromnitron benchmark.

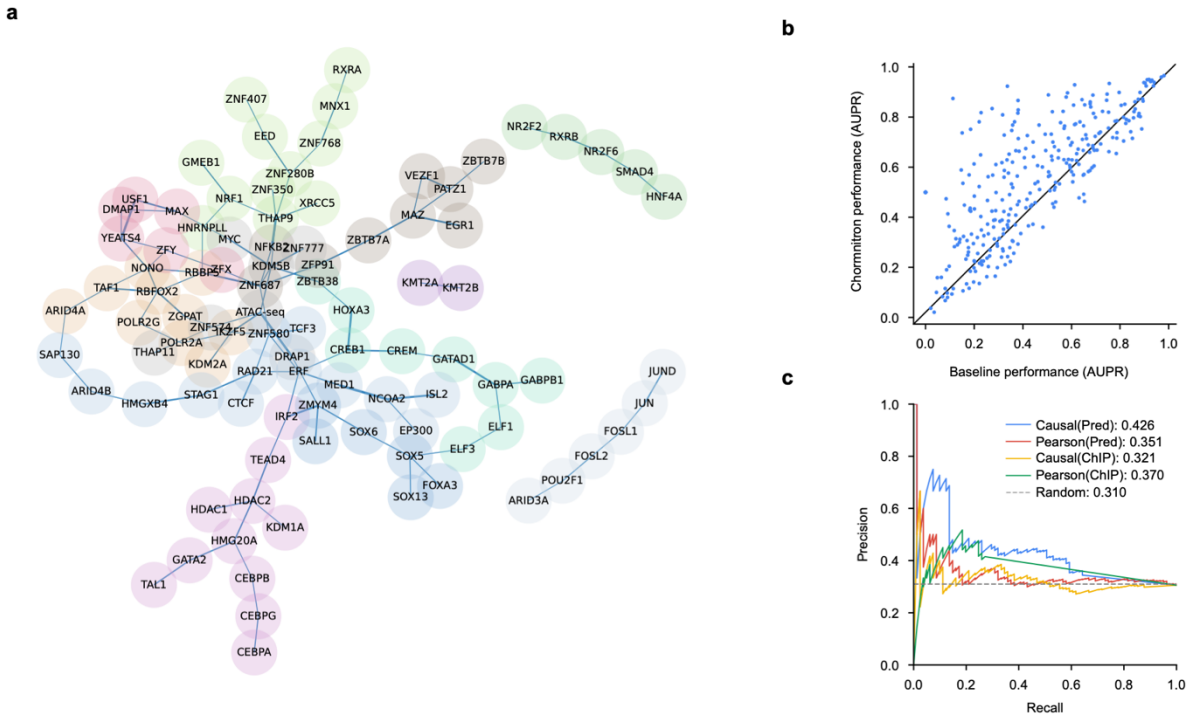
a. Benchmarking prediction performance in the leave-out genomic region (chromosome 10) of Chromnitron. The benchmarking was performed in HepG2 cells. Each dot represents a CAP prediction performance by Chromnitron (y-axis) versus the comparison model (x-axis). Area under the precision–recall curve (AUPR) and area under the receiver operating characteristic curve (AUROC) were calculated by using ChIP-seq (top) or prediction (bottom) as the reference label of binarized binding status.

b. Performance benchmarking by predicting cell type-specific CAP-binding. The benchmarking was performed by comparing model-predicted binding difference between HepG2 and K562 cells (HepG2 – K562) to the experimental ChIP-seq difference (HepG2 – K562). Pearson's r and Spearman's ρ were included as evaluation metrics in addition to the AUPR and AUROC as in (**a**).



Supplementary Fig. 10: *In silico* DNA sequence mutagenesis identifies CAP impact motifs and enables variant effect prediction.

a-c. Chromnitron prediction-based *in silico* DNA sequence mutagenesis identifies CAP-specific impactful DNA bases (i.e. impact motifs) that resembles both known (**a**) and novel (**b**) motifs, as well as indirect binding motifs that resemble their interaction partner CAPs which binds DNA(**c**). Impact scores were presented in the form of sequence logos which were normalized per base to median value. Each impact motif was predicted at a specific locus as denoted in panel title.
d. Chromnitron predicts gain-of-binding effect at *TERT* promoter G250A variant (hg38 chr5:1295135 G>A). Left: CAPs ranked by mutation impact score with vertical dashed colored lines marking ETS-family TFs. Right: top ETS-family candidate ETV4 binding profiles on wild type (WT) and mutant (Mut) sequences and their differential binding (Diff: Mut – WT).
e. Chromnitron predicts loss-of-binding effect at a CTCF-binding site (hg38 chr10:25081160 C>T). Left: CAPs ranked by mutation impact score with red dashed line indicating CTCF. Right: CTCF binding profiles across WT and Mut sequences and their differential binding.

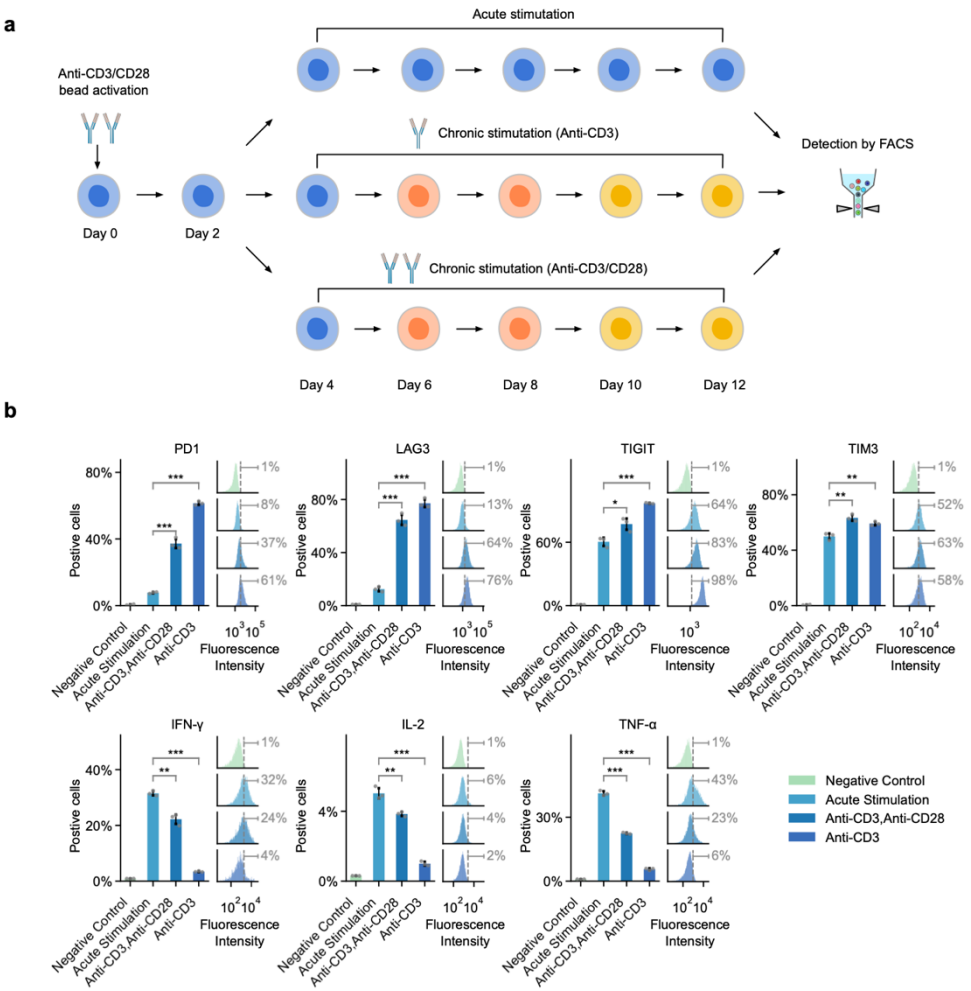


Supplementary Fig. 11: Chromnitron-based *in silico* CAP residue mutagenesis captures protein functional domains and protein–protein interactions.

a. Visualization of LiNGAM causal-discovery algorithm using predicted CAP binding profiles (Methods). Edges represent high interaction potential between adjacent CAPs. Nodes were colored by CAP community clusters.

b. *In silico* CAP residue mutagenesis by Chromnitron captures protein residues from annotated functional domains. Each dot in the scatter plot represents a CAP's AUPR scores against protein domain annotations using *in silico* mutagenesis-identified residues (y-axis) and compared with that using randomly selected residues (x-axis).

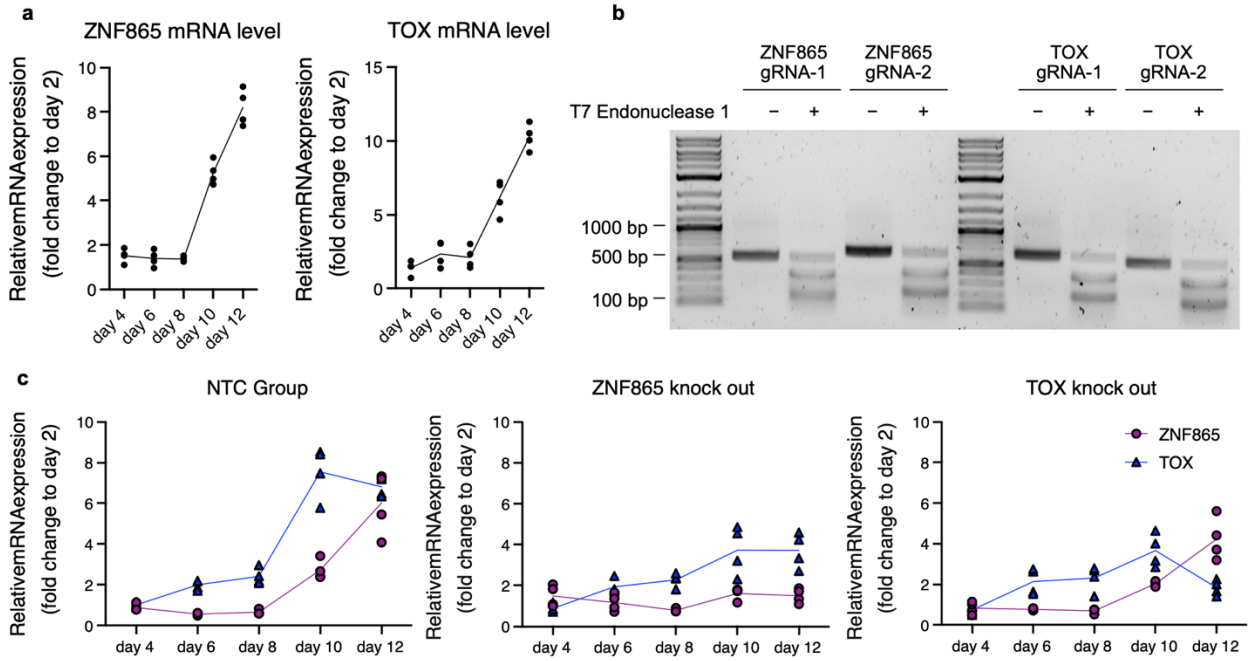
c. Performance comparison between different methods for discovering protein-protein interactions. We calculated both causal-inference-based interaction graph and Pearson's correlation-based interaction graph for both Chromnitron prediction and ChIP-seq and compared them against experimental STRING database-annotated interactions to calculate AUPR (Methods).



Supplementary Fig. 12: *In vitro* chronic stimulation of human primary CD8$^+$ T cells to model T cell exhaustion trajectory.

a. Schematic of the *in vitro* treatment of human CD8$^+$ T cells across acute stimulation (top), chronic stimulation by anti-CD3 antibodies (middle), and chronic stimulation by anti-CD3/CD28 antibodies (bottom).

b. Flow cytometry quantification of cell surface markers of T cell exhaustion (PD1, LAG3, TIM3, and TIGIT) and intracellular cytokine production (IFN-γ, TNF-α, IL-2) in different stimulation groups at day 12**.** Bar plots show mean ± s.d. for n = 3 independent replicates. Statistical comparisons performed using two-sided unpaired Student's t-tests with significance denoted as *, $P < 0.05$; **, $P < 0.01$; ***, $P < 0.001$.

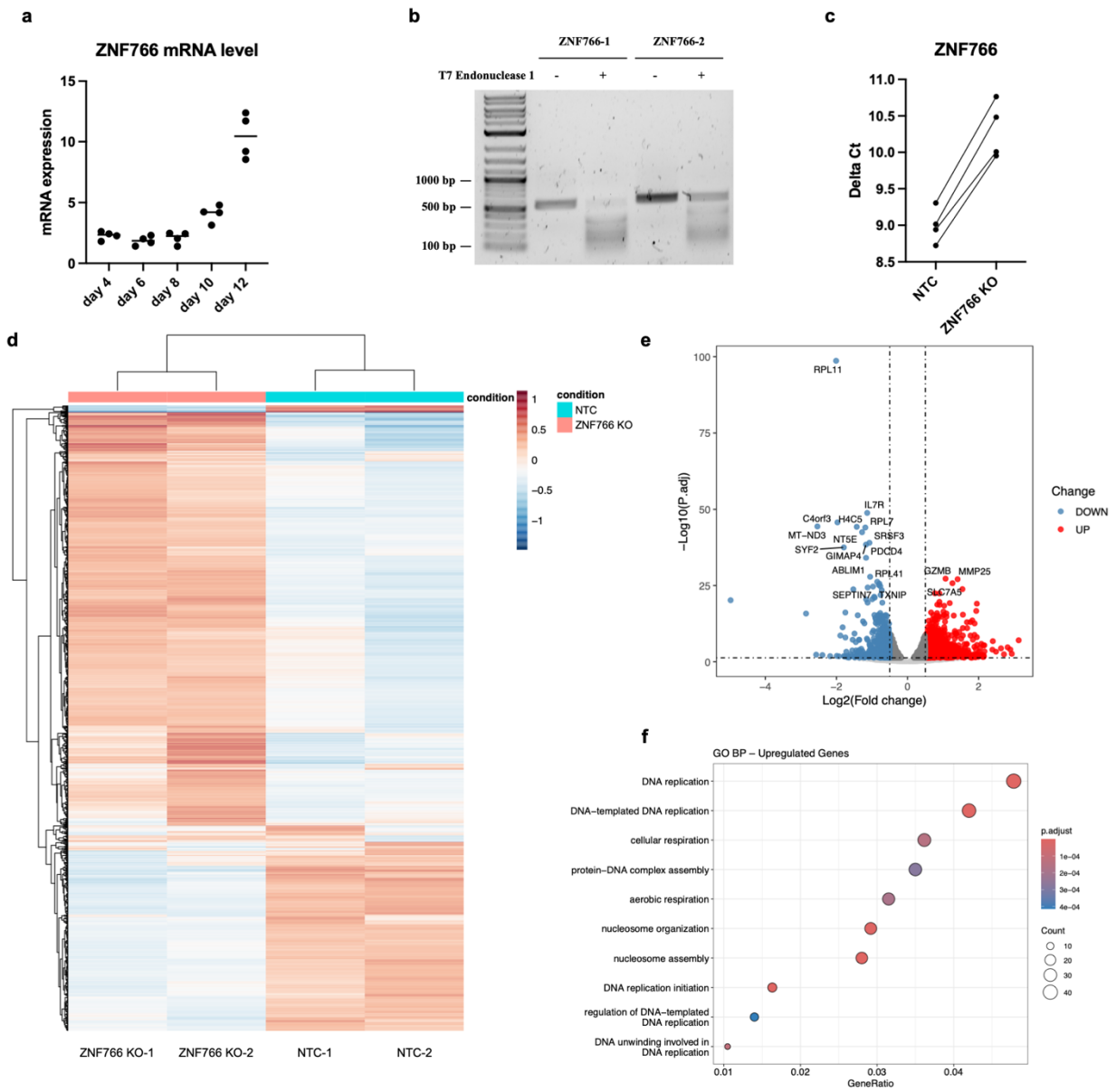


Supplementary Fig. 13: Validation of CRISPR-mediated perturbation and transcript-level perturbation of *TOX* and *ZNF865*.

a. RT–qPCR quantification of *TOX* and *ZNF865* mRNA expression level changes across chronic stimulation of human $CD8^+$ T cells from day 4 to day 12, indicating increased expression along the trajectory. Relative expression levels were normalized to day 2 using the ΔΔCt method. Each dot represents an independent biological replicate (n = 4); error bars indicate mean ± SEM.

b. Surveyor assay confirmed efficient CRISPR knockout gRNA targeting *TOX* and *ZNF865*. Each gene includes two gRNAs for knock-out perturbation. Genomic DNA was amplified around the gRNA target sites and the fragments were digested with T7 Endonuclease I.

c. RT–qPCR quantification of *ZNF865* and *TOX* mRNA level changes across chronic stimulation days with individual gene knockouts (*ZNF865* KO and *TOX* KO). Relative expression levels were normalized to day 2. Each dot represents an independent biological replicate (n = 4); error bars indicate mean ± SEM.



Supplementary Fig. 14: Validation of CRISPR-mediated *ZNF766* knock-out and transcriptomic profiling of perturbation effects on chronic-stimulated T cells.

a. RT–qPCR quantification of *ZNF766* mRNA level changes across chronic stimulation of $CD8^+$ T cells from day 4 to day 12, indicating increased expression along the trajectory. Relative expression levels were normalized to day 2 using the ΔΔCt method. Each dot represents an independent biological replicate (n = 4); error bars indicate mean ± SEM.

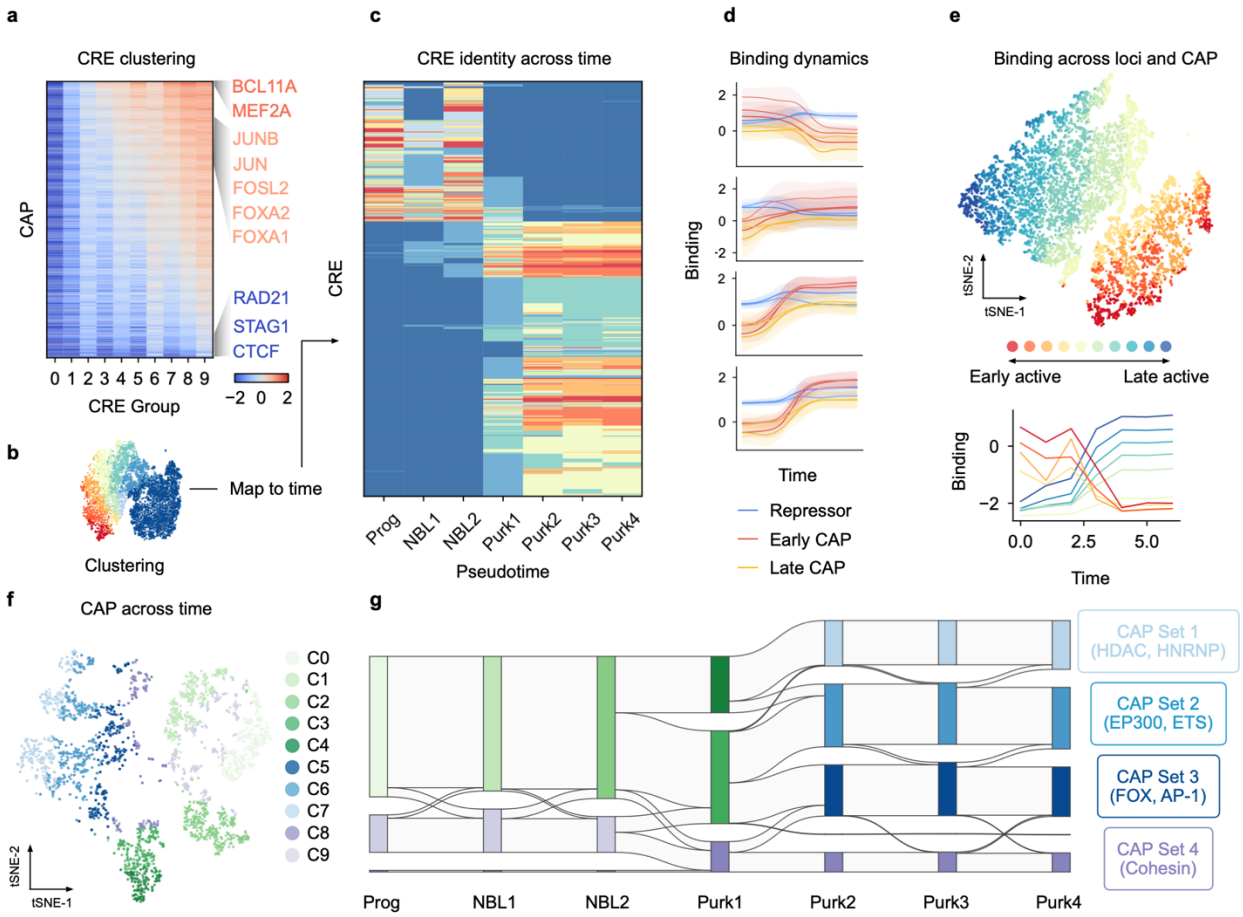
b. Surveyor assay confirmed efficient CRISPR knockout gRNA targeting *ZNF766*.

c. RT–qPCR quantification of *ZNF766* mRNA expression levels after gene knockout. Primers were designed flanking the CRISPR cutting site to confirm transcript disruption. Increased ΔCt number after perturbation indicates efficient knock down of unperturbed *ZNF766* mRNA.

d. Heatmap of differentially expressed genes between *ZNF766* knockout and non-targeting control (NTC) T cells.

e. Volcano plot of differentially expressed genes between *ZNF766* knockout and NTC groups. Red and blue dots indicate significantly upregulated and downregulated genes, respectively, in the knock-out condition ($|\log_2 FC| > 1$, $p.adj < 0.05$). Representative genes involved in T cell activation (e.g., *GZMB*, *MMP25*, *SLC7A5*) and exhaustion-associated regulators (e.g., *PDCD4*, *NT5E*, *TXNIP*) were individually labeled.

f. Gene Ontology (Biological Process) enrichment analysis of significantly upregulated genes in *ZNF766* knockout group. Dot size corresponds to gene count per term; color indicates adjusted p-value. Enriched pathways highlight enhanced DNA replication, chromatin organization, and mitochondrial respiration, suggesting a functional restoration of exhausted T cells.



Supplementary Fig. 15: Dynamics of CAP binding profile and CRE activities during Purkinje cell differentiation.

a. CAP binding profiles across CRE groups. Representative CAPs were labeled. Color scale indicates normalized binding.

b. T-SNE embedding of CREs colored by activity clusters (same as Fig. 6d).

c. Heatmap of CRE identity transitions across Purkinje cell differentiation stages.

d. Average binding dynamics for four different CRE sets across pseudo-time. Each CRE set includes three CAP classes (Repressor: e.g. TCF7, RCOR1; Early CAP: e.g. BCL11A, MEF2A, ATF2; Late CAP: e.g. FOXA1, FOXA2). Shading denotes 1 standard deviation.

e. T-SNE embedding of individualized CAP binding across loci using time trajectory as feature, colored by trajectory trend cluster. Each dot represents a CAP binding at a single locus across all pseudo-time stages. Average binding trajectories for each cluster across time were presented at bottom.

f. T-SNE embedding of CAP across time using CRE (binding profile) as feature, colored by CAP cluster identities. Each dot represents a CAP's binding profile across all loci at a single time point.

g. Alluvial diagram showing CAP cluster identity changes across time. Representative CAPs in each cluster were annotated on the right.