IceQream: Quantitative chromosome accessibility analysis using physical TF models

Akhiad Bercovich [1,2], Aviezer Lifshitz [1,2], Michal Eldar[1,2], Saifeng Cheng [2], Roni Stok Ranen[1,2], Yonatan Stelzer [2] & Amos Tanay [1,2] ✉

Single-cell mapping of chromosomal accessibility patterns has recently led to improved predictive modelling of epigenomic activity from sequence. However, quantitative models explaining the epigenome using directly interpretable components are still lacking. Here we develop IceQream (IQ), a modelling strategy and inference algorithm for regressing accessibility from sequences using physical models of transcription factor (TF) binding. IQ uses spatial integration of sequences over a range of TF-DNA affinities and localization relative to the target locus. It infers TF effective concentrations as latent variables that activate or repress regulatory elements in a non-linear fashion. These are supplemented with synergistic and antagonistic pairwise interactions between TFs. Analysis of both human and mouse data shows that IQ derives similar, and in some cases, better performance compared to state-of-the-art deep neural network models. IQ provides an essential mechanistic and explicable baseline for further developments toward understanding gene and genome regulation from sequence.

The diversity of cellular functions encoded by animal genomes relies on flexible, plastic, and specific genome regulation. Such regulation helps cells to determine, stabilize, and then differentiate their transcriptional states. Recent breakthroughs in single-cell genomics and epigenomics facilitate the reconstruction of detailed cellular manifold models that describe phenomenologically these states at high resolution[1–4]. Most current manifold models (nicknamed atlases) specify transcriptional states and possibly coupled profiles of chromosomal accessibility[5–7]. The latter lists genomic loci in which nucleosome depletion or instability is recorded either at transcription start sites (TSSs) or, more abundantly, in distal genomic loci, which can be broadly classified as cis-regulatory elements (CREs). Explaining TSSs and CREs accessibility mechanistically and based on their underlying sequences is a major open challenge in the field.

Predicting CREs from genomic sequence has a long history, and it is recently being approached using increasingly complex models leveraging deep neural networks and the extensive genome-wide data provided by single-cell data organized over cellular manifolds[8–13]. Complex black-box sequence specificity models can be equipped with exploratory tools to enhance interpretability[14]. But whether their successful deployment indicates simply rediscovery of mechanisms that are already known, or alternatively, augments these non-linearly and takes advantage of a gap in our mechanistic understanding of CRE specificity, remains unknown. In simple words: it is still unclear if we learned something new[15–18].

DNA targeting by sequence-specific transcription factors (TFs) is the major, and at present almost the only known mechanism for genomic specificity of CREs[19–22]. Naively, CREs are defined by short DNA sequences (or motifs) and their combinatorics. But as a rule, mammalian (and other eukaryotic) TFs recognize their targets loosely, such that their DNA binding preferences alone mark millions of false positive targets across the genome[23]. Moreover, eukaryotic TFs also engage with suboptimal binding sites given the appropriate chromosomal context and co-factor presence[24]. Predicting CRE activity from

the presence or absence of optimal TF binding motifs is therefore showing both low specificity and low sensitivity. Reconstructing the precise and robust accessibility landscapes that can be observed in cells in practice must therefore consider quantitative TF-DNA biophysics and complex multivariate interactions among dozens of TFs that are typically active concurrently in any cell state.

The chromosomal state of a cell depends not only on the current activity of TFs, but also on interactions occurring during the differentiation leading to the current cell state-representing the cell's epigenetic memory[25,26]. Not enough is known about the kinetics and impact of such memory, but it is clear that modeling dynamics at TSSs and CREs must consider at least the recent cellular history leading to the state under investigation. Single-cell genomics data cannot directly specify these dynamics, as it recovers snapshots of single-cell states. But cellular manifold models can be interpreted as defining differentiation trajectories, suggesting possible epigenetic histories for any states based on the putative differentiation trajectory leading to it.

Here, we introduce IceQream (IQ), a model and inference algorithm for quantitative fitting of chromosomal accessibility levels using fully explicable physical TF model components. IQ sets the stage for fully mechanistic models explaining gene and genome regulation, and opens up multiple hypotheses for further investigation and development, including integration of low-affinity TF-DNA interactions, condition-specific non-linear interaction of CREs with TFs, and pairwise TF-TF links. We demonstrate this using the analysis of differentiation manifolds describing mouse gastrulation and human hematopoiesis. We suggest that bottom-up genome regulation analysis, as implemented by IQ, can be developed in parallel to top-down deep learning models. IQ will serve both as a platform for representing and updating the continuously increasing mechanistic understanding of the specificity of genome and cellular regulation and as a device for discovering novel mechanisms if and when bottom-up models fail to reproduce the performance of their deep learning analogs.

Results

Overview of IceQream quantitative accessibility modeling
IceQream (IQ) transforms single-cell ATAC raw counts (Fig. 1A, (i)) to estimated *access probability* (AP), representing probabilities of loci accessibility in an absolute range of 0 to 1. AP values of 1 represent a locus that is accessible in all cells at all times, and lower values indicate either transient accessibility in all cells or more stable accessibility in a subset of the cells. For an idealized single cell and single locus, accessibility may be represented as a binary variable of either open or closed state, yet AP aims to quantify the frequency of observed loci being accessible when sampling a large number of cells from the same cellular state (and across time for dynamic processes). This motivates the focus on quantification in a continuous variable rather than a binary one. The IQ model (Fig. 1A, (ii)) is built upon components representing transcription factors that are coupled together with epigenomic context variables and pairwise interactions. Each TF model integrates contributions from strong and weak affinity sequences as predicted by a PSSM model, and these are weighted by a spatial curve specifying binding preferences within and around the center of the accessible hotspot (i.e., the nucleosome-depleted region). TF models are linked together via a family of pre-defined non-linear dose-response-like curves. For model initialization, the IceQream algorithm (Fig. 1A (iii)) scans through candidate initial TF models from PSSM databases[12,27–30] and de-novo motif regression. Filtered clusters of these candidates are used to initialize *consolidative* regression and generalized linear modeling, so that one integrated model predicts (Fig. 1A, (iv)) the differential AP over a selected manifold differentiation trajectory. To allow systematic analysis of scATAC dynamics, IceQream can then fuse IQ models inferred from different manifold trajectories, and recompute them using a common universal set of TF models (Fig. 1A, (v)). IceQream is implemented modularly to provide users

with control over the main steps of the process (https://github.com/tanaylab/iceqream) and to allow incorporation into scATAC and Multiome analysis pipelines.

Transforming scATAC read counts to normalized access probabilities
We tested conversion of raw ATAC read counts to APs using data on two manifolds capturing complex differentiation programs in mouse gastrulation and human hematopoiesis, respectively (Fig. S1). To derive estimations of APs, we identify a set of genomic loci that can be assumed to be accessible with probability 1 (i.e., AP = 1) across all cells regardless of their manifold state (denoted constitutively open loci). Importantly, a natural set of loci with high and extremely well-conserved raw ATAC UMI count is indeed observed in both manifolds (and any other dataset we studied). This set includes mostly promoters of housekeeping genes (Fig. S2A) for which accessibility is indeed constantly high. The raw ATAC UMI counts on such loci represent significant technical variation that must be normalized. Variation in raw ATAC read counts for constitutively open loci is strongly correlated with local and regional GC content or, more generally, with the background ATAC signal around it (defined as mean ATAC count in 20 kbp windows around the element but excluding the element itself) (Fig. S2B). We thereby normalize the ATAC signal of any local element through subtraction by the mean ATAC in 20 kbp windows around it (excluding the center 1 kbp window). This corrects the initial asymmetry in the intensity of constitutive and differential ATAC signal (Fig. 1B), recovering some differential peaks as similarly accessible to constitutive peaks. Assuming that constitutively open loci are representing AP = 1, we further transform the region-normalized data to ensure that normalized ATAC in all other loci is scaled appropriately (Fig. 1B, middle). We note that this is resulting in different cell types showing up with higher or lower total accessibility (totalAP = $\sum_i \alpha_i$), which may be correlated with an overall more plastic or more restricted epigenetic landscape per cell state. Indeed, we observe that the more differentiated states in both the gastrulation (i.e., primitive erythrocyte) and hematopoiesis (i.e., pro-B cells, platelets) datasets show accessibility that is more concentrated on constitutive loci (and therefore involves less activity on regulatory, non-constitutive loci), when compared with the pluri- or multi-potent states (e.g., Epiblast, HSC/MPP, Figure S2C). We note that the IQ normalization approach is designed to maximize quantitative readout when regressing sequence models, defining a different goal compared to information-theoretic or statistical peak finding approaches (see Fig. S2D for comparison to TF-IDF[31,32]).

Following normalization, IQ can process differential AP values on endpoints of differentiation trajectories (Fig. 1C, D), focusing on loci with high absolute and differential AP.

IceQream generates concise models without compromising accuracy
We selected for analysis differentiation trajectories that provide unambiguous endpoints, focusing on the transition of mouse epiblast to mesoderm or endoderm and the differentiation of human hematopoietic stem cells toward myeloid and erythroid lineages. Figure 2A shows key statistics on the IQ inference process as applied to mouse mesoderm differentiation. The algorithm consolidates thousands of database and de-novo derived TF components (level 0) into refined, concise models (level 1–4), and adds pairwise interactions between the consolidated models (level 5). This leads to models with a remarkably lower number of TF components while improving prediction accuracy. Overall, the mesoderm model reduced 21867 initial models into 13 final models. We note that model accuracy can be measured conservatively using the fit to the differential AP signal (Fig. 2B, C) or based on the inherently better fit to absolute AP values in the differentiation endpoint (Fig. 2D). We note that an even better fit would be reported if
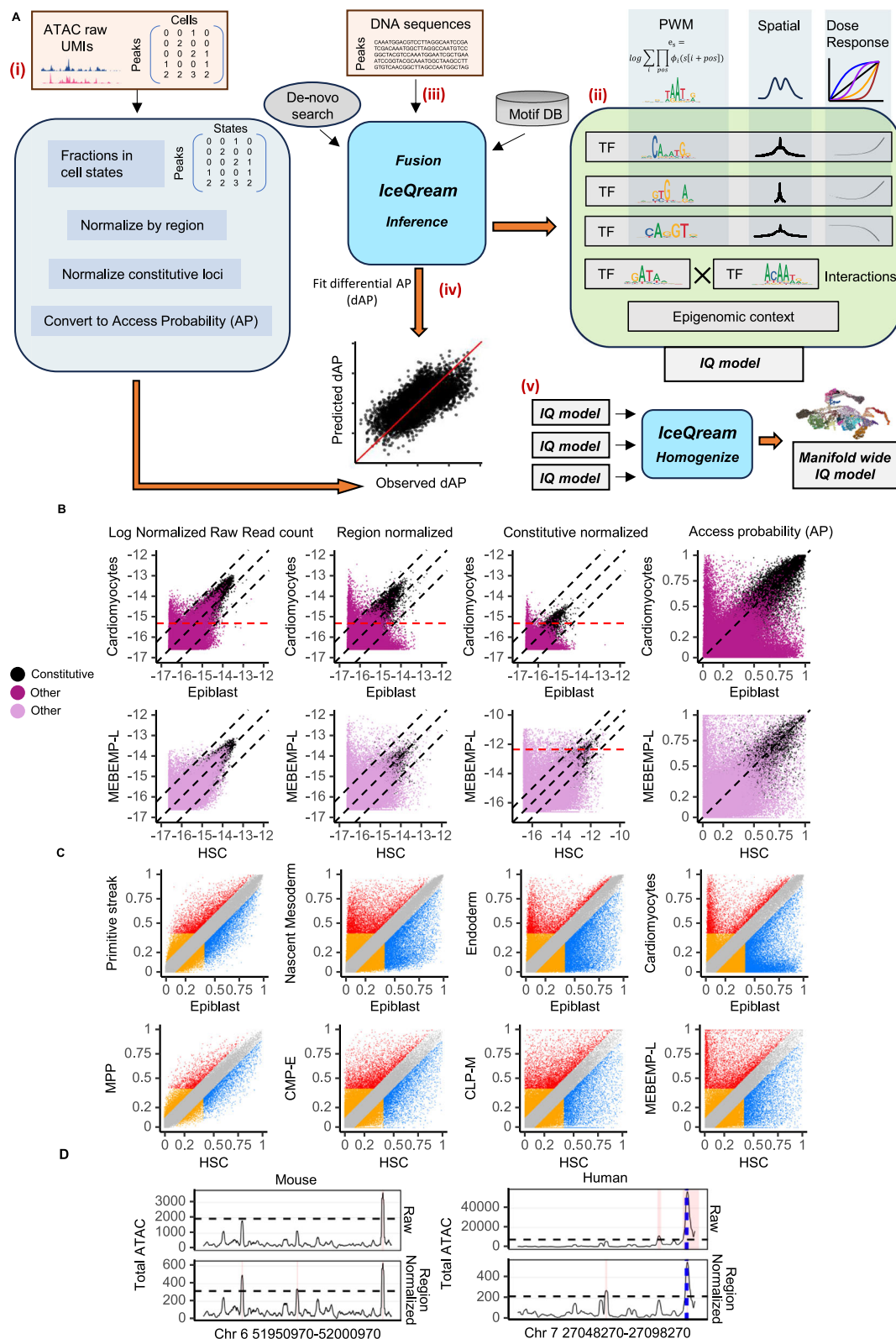
using unnormalized ATAC values and including constitutive peaks or TSS, as these are all highly predictable by GC content and other low-information content sequence features. But predicting constitutive or unnormalized ATAC will, of course, not be informative for understanding true CRE dynamics.

We compared the IQ complete scheme to alternative regression approaches in Fig. 2E. We observed poor performance for a simple linear model on the database motif, and gradual improvement when using IQ consolidation of motif components with linear models, generalized additive models (GAMs), or IQ nonlinear functions without pairwise interactions. Gradient boosting with IQ features performs better but is still inferior to the complete IQ scheme. IQ is also performing better when compared to a previous motif clustering and consolidation strategy by *SCENIC*[12] (Fig. S3A).

More crucially than improved technical performance metrics, the IQ model allows direct interpretation globally and over individual loci.

Fig. 1 | IceQream overview. A Schematic of the IceQream (IQ) workflow: (i) Single-cell ATAC raw counts are transformed into estimated access probabilities (AP). (ii) The IQ model incorporates transcription factor (TF) models, epigenomic context variables, and pairwise interactions of TF models. Each TF model integrates contributions from strong and weak affinity sequences, weighted by spatial preferences around the accessible hotspot, which are transformed into dose-response-like spatial binding preference curves using pre-defined non-linear functions. (iii) Model initialization involves scanning candidate TF models from PSSM (position-specific scoring matrix) databases and de novo motif regression. (iv) An integrated IQ model predicts differential AP (dAP) across a selected manifold trajectory. (v) IQ models from multiple trajectories are fused to create a manifold-wide set of common TF motif models. **B** Normalization steps from raw ATAC-seq data on peaks to access probabilities (AP, left to right): Raw counts, region-normalized counts, constitutive-loci-normalized counts, and final APs for mouse gastrulation (top) and human hematopoiesis (bottom) datasets. Black points represent the constitutive loci. The red dashed line indicates the threshold for loci with AP = 1 (-15.3 for mouse gastrulation, −12.4 for human hematopoiesis). **C** AP for various cell types compared to epiblast in mouse gastrulation (top) and compared to HSC in human hematopoiesis (bottom) manifolds. Red and blue dots represent loci that opened or closed during the trajectory; gray dots show loci that did not change; and orange dots represent loci with a small change (dAP ≤ 0.4). **D** Examples of scATAC-seq signal at specific genomic loci before and after region normalization in mouse (left) and human (right) genomes. Top panels show the raw scATAC-seq signal (total number of reads). Bottom panels show the signal after region normalization. This is calculated as the raw signal divided by the mean ATAC signal in a 20 kbp window around it, excluding the center 1 kbp window. Dashed horizontal lines indicate the threshold for peak calling. Red shaded areas denote called peaks. Blue dashed vertical lines represent transcription start sites (TSS). Source data are provided as a Source Data file.

The TF components of the model in the mesoderm trajectory of the mouse model (see Fig. S3B, C for similar analysis in the human MEP trajectory), can be divided into a group of components with positive response (including Mesp, Eomes, and Gata models, Fig. 2F, left) and components with negative response (including Sox/Nanog, Pou3, and Snai, Fig. 2F, right). TF components vary in their non-linearity, but only two of the 14 models (annotated as Dmrta2 and Tbx1) showed a non-monotonic response curve, suggesting that for these sequences, additional interactions between factors are leading to preferences for medium sequence affinity over both low and high affinity. The low dimension of the final IQ model also facilitates a simple combinatorial depiction of the relationship between TF models. For example, in the mouse mesoderm model (Fig. 2G, Fig. S3E, F), Eomes (positive) and Tcf (negative) showed generally consistent correlation with AP differential, while Pou3, Nanog-Oct, and their combination were observed only in defined subsets of the loci. We note the inferred PBMs represent TFs that have established roles in these developmental trajectories, and we validate both their expected expression (Fig. 2H) and enrichment in selected TF-ChIP-seq tracks (Fig. 2I). In summary, IQ can infer a directly interpretable model without compromising accuracy, thus supporting downstream quantitative and biophysically-inspired analysis.

IceQream fits quantitative affinities to continuous accessibility probabilities

DNA accessibility per locus is commonly conceptualized and modeled as a binary state. However, when observing ATAC signals over cell populations, the average accessibility state is represented by IQ quantitatively through AP levels, defining a regression task, rather than a classification problem. As shown in Fig. 3A, when grouping loci into seven bins based on their IQ predicted ATAC difference on the epiblast to mesoderm mouse trajectory, we demonstrate a gradual shift in the mean observed ATAC difference rather than a change in the mixture of differential vs conserved sites. This implies that using sequence and epigenomic features alone, the model separates strongly induced mesoderm CRE from weakly (but still significant) induced ones. Similar separation is observed for endoderm-repressed CREs and for other trajectories (Fig. S4A, B, $p \ll 10^{-6}$).

Quantitative regulation of CRE activity can be modeled by IQ since it fits the integrated contributions of short sequences across the respective regulatory element, considering both optimal consensus motifs and sequences that are only partially compatible with the TF's preferences. This approach is crucial for model accuracy, as shown by tests in which we replace IQ integrated motif energy with the affinity of only the top binding site for each CRE, leading to a significant reduction in model R^2 (Fig. 3B). Combinations of lower-affinity sites are therefore likely to contribute to CRE specificity and can compensate for the lack of an optimal motif. This integrative capability may, however, be restricted to only some of the TFs (as represented by individual components in the integrated model). To test this directly for any given motif, we classified CRE sequences into those containing a single high-affinity locus or those containing multiple medium-affinity loci. We then compared the differential ATAC signal on CREs from these groups. Interestingly, we identified consistent behaviors for specific factors over multiple trajectories (Fig. 3C). For example, Mesp2 (promoting mouse mesoderm CREs) and Nanog/Sox (promoting mouse epiblast CREs) showed a preference for a single near-optimal binding site. In contrast, the Snai and Tcf mouse models were susceptible to occurrences of two or more medium-affinity sites. Affinity preferences for all inferred motifs are shown in Data S1, Fig. S4C, D.

Spatial and epigenomic features provide IQ models with predictive gains

Our analysis of AP differentials over trajectories is focused on localized peaks of ATAC coverage that generally represent nucleosome-depleted regions (NDRs). These elements are assumed to be observed over a length scale of two spaced nucleosomes, or approximately 400–500 bp. IQ modeling can account for preferential positioning of TF binding relative to the NDR through spatial curves defining probabilities per distance from the NDR center that are inferred for each motif. For mouse gastrulation, we observed remarkably homogeneous spatial curves representing general preference within −80: + 80 bp to −120: + 120 bp around the center (Fig. 3D, compared to nucleosome positioning data in Fig. S5A–D). Human models showed more varied spatial preference, with a highly localized effect for key TFs, including Gata and PU.1. IQ can take into account epigenomic features, if such features are measured for the states at the base of the modeled trajectory (data on the epigenomics of the trajectory itself or its endpoint is not used to avoid information leakage). We indeed observed some improvement in model performance (Fig. 3E) when considering the initial local epigenetic activity of a locus as indicated by the histone modifications H3K27ac or H3K4me3 (Fig. 3F). The model can also consider the regional activity around of locus, by integrating total accessibility within 20 kbp around it at the base of the trajectory, gaining some additional predictive value in some trajectories.

Inferred IQ Interactions highlight cooperative CRE specificity

Cooperative or antagonistic interactions between pairs of PBMs or between PBMs and epigenetic features are modeled by IQ explicitly and improve model predictions significantly (Fig. 2E, compare IQ to IQ-no interactions). Of note, depending on the number of PBMs and data size (number of considered CREs), IQ filters potential pairwise interactions during the learning process (METHODS). Many interaction terms showed non-negligible coefficients (Fig. 3G), and distributions of ATAC difference for loci with strong affinity for such pairs were suggestive of a true synergistic effect (Fig. 3H). Support for synergistic
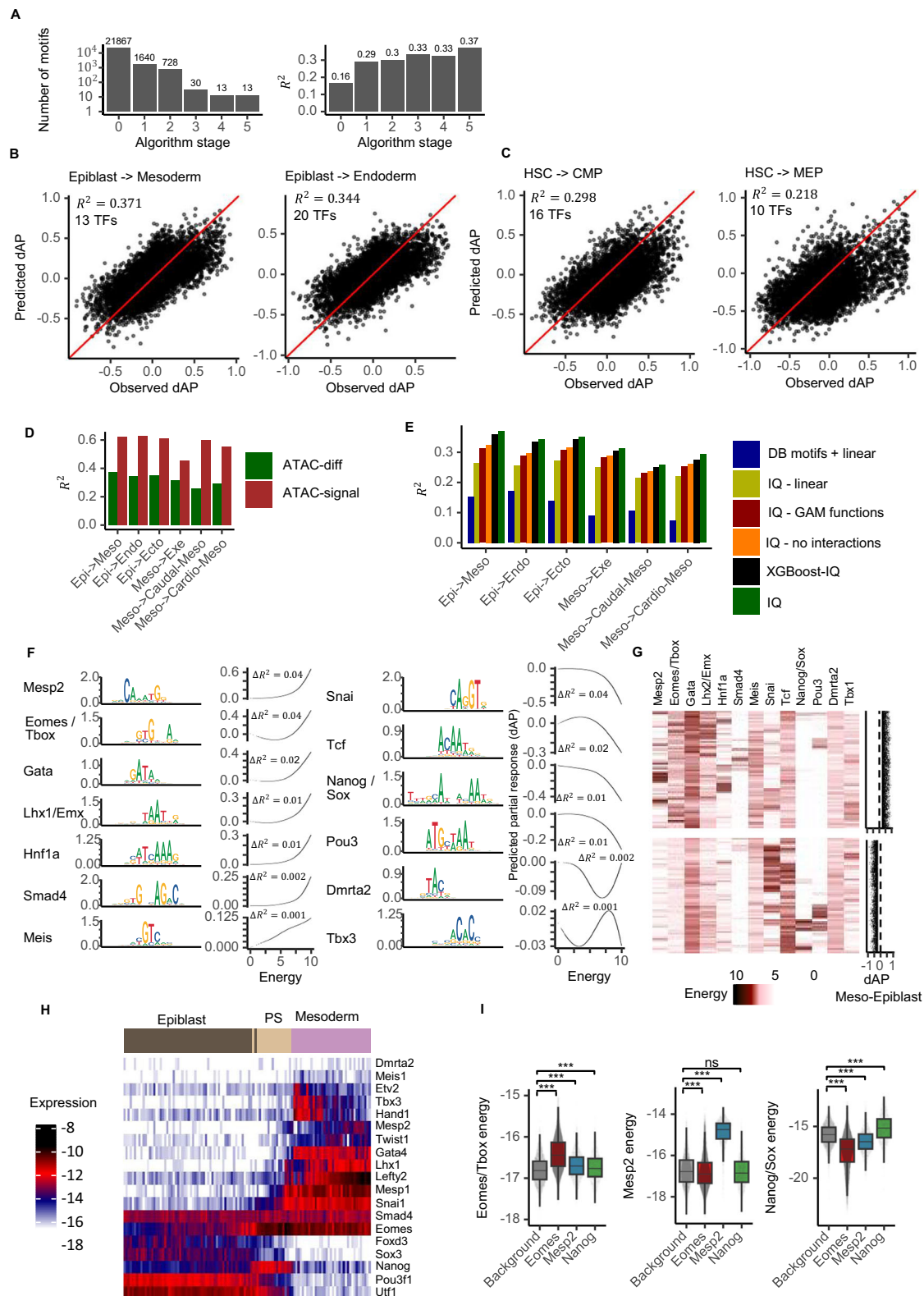
interactions between TFs can be provided when comparing differential ATAC in loci that are grouped according to the relative spacing of their putative interacting binding sites. For the potential interaction between Eomes and Mesp in the mesoderm trajectory, such analysis clearly shows that localization of binding sites for the two models within 20 bp is correlated with much stronger responses when compared to pairs observed within 100–200 bp ($p < 0.001$, One-sided Kolmogorov–Smirnov test) and even when comparing pairs within

50 bp ($p < 0.05$). Significant pairwise interactions may represent the formation of heterodimers and specific stable structures over a CRE[33], but such structures are yet to be fully defined. We hypothesize that the quantitative impact of genomic spacing (exemplified also for Atf4-Atf3, Fig. 3J–L) may also implicate some synergistic effect with a more dynamical model leading to cooperativity through maintenance of an open state over time. Such synergism may even involve changing TF activities over the developmental trajectories under study, where a

Fig. 2 | IQ generates precise and interpretable accessibility models. A Number of PWM motifs at different stages of the IQ algorithm (methods, left) and the R^2 values at each stage (right). **B**, **C** - Scatter plots comparing observed vs. predicted dAP on the sets of 20% test CREs for the mouse (B; $n = 19858$), and human (C; $n = 11057$) trajectories. **D** Model performance (R^2) when predicting the change in accessibility (dAP, the ATAC differential) vs predicting the absolute access probability (AP, the ATAC signal). **E** Model performance (R^2) across different regression approaches for mouse gastrulation trajectories (methods). In brief, "DB motifs" uses top database motifs without the IQ fusion and spatial weighting. "IQ-linear", "IQ-GAM functions", and "XGBoost-IQ" respectively use a logistic regression model, Generalized Additive Models (GAM), or an XGBoost approach to combine the IQ models instead of the non-linear response functions used by "IQ-no interactions". These model variants are compared to the full IQ model, including interactions. **F** Response curves for inferred transcription factors (TFs) in the epiblast to mesoderm trajectory, showing positive (left) and negative (right) components that are annotated by comparison to databases and manual curation. Left - sequence logo representing the position weight matrix (PSSM) of the fused motif, where y-axis shows information content in bits. Right - response curve where X-axis represents binding energy, and Y-axis shows the predicted response (dAP). ΔR^2 denotes the predictive value added when the TF is included in the model. **G** - Heatmap illustrating the relationship between TF binding and accessibility changes in the epiblast to mesoderm trajectory. Columns represent transcription factors (TFs) and rows represent loci with absolute dAP ≥ 0.2. Color intensity indicates motif energy. The heatmap is split into two parts: the top shows loci that became more accessible ($n = 8972$), and the bottom shows loci that became less accessible ($n = 8815$) during the trajectory. Scatter plots at the right display the differential AP between epiblast and mesoderm states for each locus. Positive values (top) indicate regions that became more accessible during the trajectory, while negative values (bottom) show regions that became less accessible. **H** Heatmap showing gene expression of TFs inferred by IQ in the epiblast to mesoderm trajectory over metacells in the mouse gastrulation manifold. **I** Box plots comparing motif binding energies between background regions ($n = 5000$) and ChIP-seq peaks for Eomes/Tbox ($n = 1618$), Mesp2 ($n = 2156$), and Nanog/Sox motifs (Nanog $n = 1614$, Sox2 $n = 1827$). Higher energies indicate stronger binding affinity. Boxes show median, quartiles, and 1.5× IQR whiskers. Two-sided Kolmogorov-Smirnov test: *** indicates $p < 0.001$; ns indicates not significant. Eomes/Tbox peaks vs background: Eomes D = 0.4064, $p < 1 \times 10^{-300}$; Mesp2 D = 0.1613, $p = 2.35 \times 10^{-34}$; Nanog D = 0.1052, $p = 3.73 \times 10^{-39}$. Mesp2 peaks vs background: Eomes D = 0.0662, $p = 2.08 \times 10^{-15}$; Mesp2 D = 0.8004, $p < 1 \times 10^{-300}$; Nanog D = 0.0202, $p = 7.45 \times 10^{-2}$ [ns]. Nanog/Sox peaks vs background: Eomes D = 0.4391, $p < 1 \times 10^{-300}$; Mesp2 D = 0.2640, $p = 3.30 \times 10^{-91}$; Nanog D = 0.1308, $p = 3.87 \times 10^{-60}$. Source data are provided as a Source Data file.

reduction in the activity of an early factor (e.g., Eomes) is compensated by an increase in the activity of a later factor (e.g., Mesp1), and communication is facilitated by the epigenomic state of the locus.

Homogenization of CRE models across a single cell manifold
Each IQ regression task is solved given differential AP on a select differentiation trajectory. In gastrulation, hematopoiesis, and in most other cases, multiple trajectories on a manifold may be of interest. Understanding regulatory activity through comparison of PBMs across trajectories requires homogenization, where one set of PSSM and spatial curves is inferred universally and used with different parametrization of response curves, weighting, and pairwise interaction to regress all trajectories. To facilitate such homogenization, given a set of individual trajectory models, IQ is using clustering of motif models across trajectories (exemplified in Fig. 4A), followed by resubstituting each cluster's motifs with one common model that is inferred de novo to fit partial responses over all trajectories. We applied homogenization in the mouse gastrulation manifold (Fig. 4B) to derive one common set of 61 motifs. Analyzing the human HSPC manifold led to a model with 20 motifs. We demonstrated that homogenization is highly effective, since when rerunning modeling on individual trajectories using the homogenized motifs, we did not record any loss in performance, and in fact quantified improved performance on cross validations (Fig. 4C, Fig. S5A). Taken together, we can now study CRE dynamics over an entire manifold (Fig. 4D) using A) clusters of loci (columns in Fig. 4D) that show B) common AP levels across the manifold states (Fig. 4D, top rows), which are compared to C) a controlled set of motifs (Fig. 4D, bottom rows). Motifs that are enriched in specific clusters of loci, or in combinations of such clusters, are powerful building blocks for developing integrated regulatory models.

IQ predictive power matches deep learning counterparts
Language models and, more generally, deep learning (DL) models have recently seen increased use in genomic predictive tasks. The flexibility of these models to receive inputs on multiple scales, together with the capacity to learn complex functions from large datasets across entire genomes, holds great promise for modeling the regulatory code. However, the black box nature of current DL models raises questions about what new biological phenomena we uncover with this increased predictive capacity. Comparing the predictions of IQ to current genomic models can therefore direct research into unmapped or underappreciated mechanisms learned by DL models, as well as assist in the development of future models. To address these goals, we first generated IQ models using motifs from all the gastrulation or hematopoiesis manifolds combined (Fig. 5A, METHODS). These yield the most accurate IQ model variants, though with higher interpretation costs (80-180 motifs instead of 17−31 motifs per model). We then trained several DL models using the CREsted[34] or Enformer[35] frameworks to predict all trajectories from the gastrulation and HSPC manifolds (METHODS). These DL models can be subdivided into convolution (CNN) or transformer (LM) based models, of which DeepTopic[34] (CNN) and Borzoi[36] (LM) showed the best-in-class predictive power. Comparing the R^2 values across differential accessibility scores (Fig. 5B, C, and METHODS), we observed IQ accuracy on par with these state-of-the-art DL models. The best predictive power was achieved by model ensembles, with IQ + Borzoi (LM) slightly outperforming an ensemble of all DL models (Fig. 5D). CREs with significant errors in Borzoi predictions but accurate predictions in IQ and ensemble models show higher energies for the IQ PBMs, such as Eomes and Mesp2 (Figs. 5E, S6C). LM models are structured to model long-range patterns (sometimes defined as regulatory "grammar") that are inaccessible to the IQ localized modeling approach. But powerful DL models can also recast known local TF-DNA interactions within their extensive parameter space. Our results show that if non-local, grammar-like structures were discovered by current DL models, their impact is somewhat limited to the performance gap quantified between IQ and the ensemble IQ+Borozoi model. We believe that further comparison of current and upcoming DL models to IQ models can highlight gaps in our current mechanistic understanding of genome and chromosomal regulation, as well as guide the development of DL models toward the more difficult challenges in genome regulation.

Exploratory analysis of CREs given an IQ model
IQ models can be interrogated globally (in-trans) given putative TFs and their interactions over a trajectory, but also provide a valuable toolkit (denoted IQ-plots or IQ-P) for understanding CREs in cis, given the directly interpretable and localized semantics of the model's motifs. For each CRE, we can compare using IQ-P the spatial distribution of normalized ATAC intensities to the localization of sequences providing high contributions to model predictions in each motif component of the model (Fig. 6A). To allow exploration of these sequences, IQ-P implements a new visualization scheme focusing on and color-coding "important" sequences over a CRE. This strategy identifies cases of partial overlap between putative binding sites
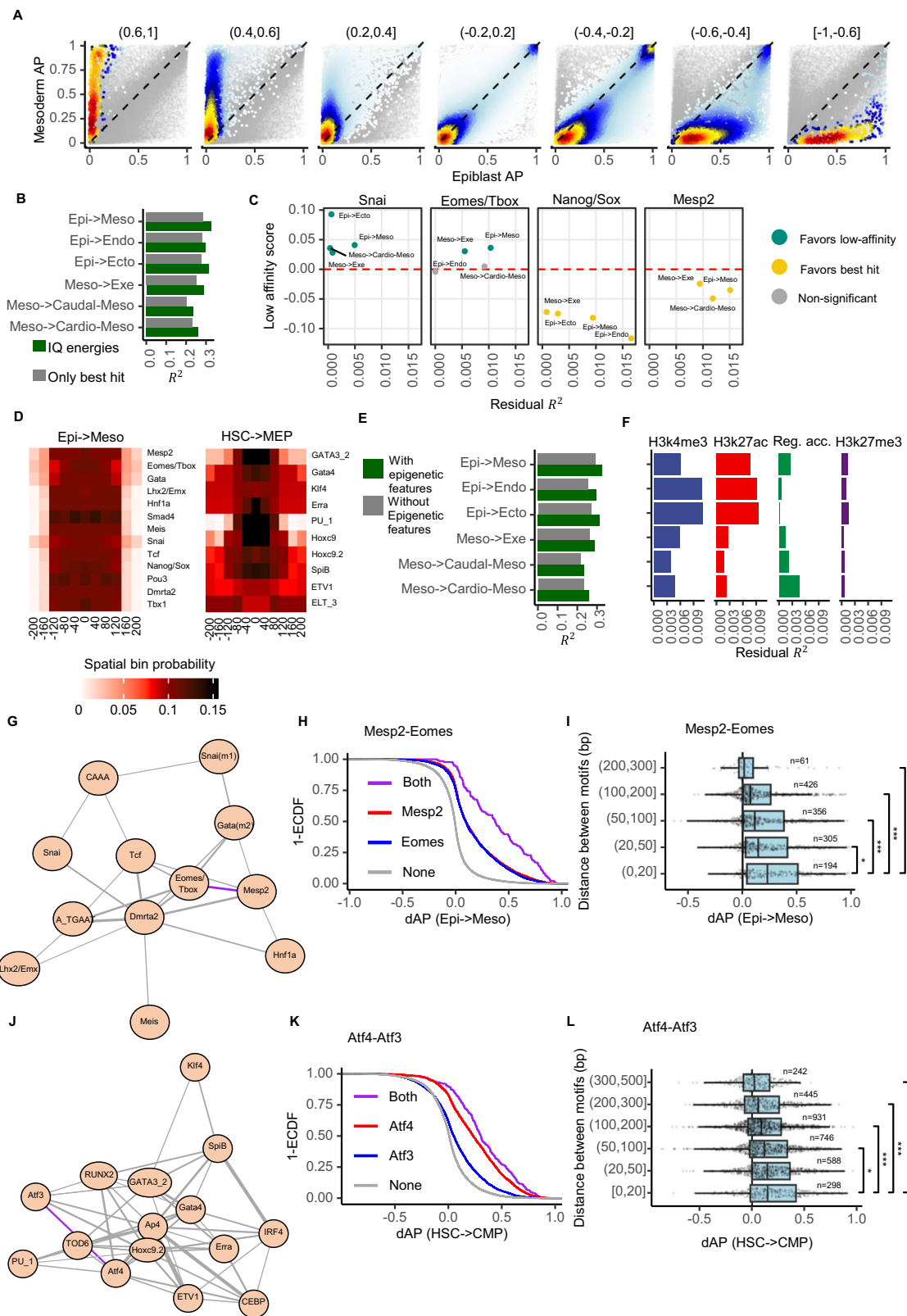
(Eomes and Snai, Fig. 6A), cases of flanked localization (Mesp and Hnf, Fig. 6A), and the overall sequence architecture of a locus. All of these sequence features are difficult to discern within the CRE 300 bp without IQ's highly specific modeling. IQ-P also detects and allows examination of CREs with multiple sub-optimal binding sites (Fig. 6B). Detailed architectural analysis of a CRE, as rendered by IQ-P, immediately opens questions regarding the role of higher-level architecture,

or "grammar" between binding sites. Importantly, the IQ physical model is completely blind to such higher-order effects and only integrates independent spatial preferences for each motif and global contribution for their pairwise interactions. As discussed above, comparisons to DL models with the potential to infer more complex grammar are inconclusive concerning the functional impact that such higher-order signatures can have.

Fig. 3 | IQ use of low-affinity binding, spatial preferences, and transcription factor interactions. A Density plots showing the distribution of observed AP for loci grouped into seven bins based on their IQ-predicted dAP on the epiblast to mesoderm mouse trajectory. Gray dots represent all loci in the dataset, while colored dots represent the density of loci in the specific bin of each plot. The gradual shift in distributions demonstrates IQ's ability to distinguish between strongly and weakly induced/repressed cis-regulatory elements (CREs). **B** Comparison of model performance (R^2) when using IQ's integrated motif energy approach (methods) versus considering only the top binding site for each CRE, across different mouse gastrulation trajectories. **C** Comparison of the impact of low-affinity versus high-affinity binding sites. Y-axis: mean dAP for loci with multiple low-affinity sites minus the mean dAP for loci with a single high-affinity site (methods). X-axis: residual R^2 of each motif in the trajectory model. Each point represents one mouse gastrulation trajectory. **D** Heatmaps showing the inferred spatial binding preferences for transcription factor models in a mouse (left) and human (right) trajectories. Columns represent different distances from the CRE center. **E** Comparison of model performance (R^2) with and without epigenomic features for different trajectories in mouse gastrulation. **F** Residual R^2 values showing the predictive power of different epigenomic features across the mouse gastrulation trajectories in 3E. **G** Network diagram illustrating inferred pairwise interactions between transcription factor models in the mouse epiblast to mesoderm trajectory. Edge thickness represents the magnitude of the beta coefficient

for each pairwise interaction. The purple edge highlights the interaction between Mesp2 and Eomes, further examined in panels H and I. **H** Empirical cumulative distribution function (ECDF) of dAP for loci with strong affinity for Mesp2 (red, $n = 2715$), Eomes (blue; $n = 2844$), both factors (purple; $n = 136$), and negative controls (gray; $n = 93596$), demonstrating synergistic effects in the mouse epiblast to mesoderm trajectory. Strong affinity was defined as motif energy ≥ 8. **I** Boxplots showing the distribution of dAP for loci grouped by the distance between Mesp2 and Eomes binding sites within the same CRE. Boxes: median (center), IQR (box), whiskers = 1.5×IQR; points outside whiskers = outliers. One-sided Kolmogorov–Smirnov tests compared the (0,20) bp group against each other spacing group. * $p < 0.05$, *** $p < 0.001$. (0,20) bp vs. others: (20,50] D = 0.1126, $p = 4.94 \times 10^{-2}$; (50,100] D = 0.1719, $p = 5.99 \times 10^{-4}$; (100,200] D = 0.2362, $p = 3.49 \times 10^{-7}$; (200,300] D = 0.4205, $p = 7.47 \times 10^{-8}$. **J–L** As in (G–I), showing data from the human HSC to CMP trajectory and the Atf4-Atf3 interaction. Sample sizes: Atf4 $n = 3514$; Atf3 $n = 2532$; both $n = 251$; negative controls $n = 48,990$. Boxplots: median (center), IQR (box), whiskers = 1.5×IQR; points outside whiskers = outliers. One-sided Kolmogorov–Smirnov tests compared the (0,20) bp group against each other spacing group. * $p < 0.05$, *** $p < 0.001$. (0,20) bp vs. others: (20,50] D = 0.0694, $p = 1.49 \times 10^{-1}$ [ns]; (50,100] D = 0.0869, $p = 4.0 \times 10^{-2}$; (100,200] D = 0.1480, $p = 5.1 \times 10^{-5}$; (200,300] D = 0.1625, $p = 8.1 \times 10^{-5}$; (300,500] D = 0.2728, $p = 2.3 \times 10^{-9}$. Source data are provided as a Source Data file.

Discussion

We introduce IceQream (IQ) to derive physical models fitting regulatory activity in cis-regulatory elements (CREs). The model is composed of a limited number of components ("motifs"), each defining sequence-specific binding probabilities integrated over the CRE sequence with spatial preference and non-linear response. Interaction between motifs and epigenetic memory indicators is added to the model. This model is fitted to estimated CRE access probabilities (APs) ranging between 0 and 1, normalized from raw ATAC-seq reads. APs are assumed to specify the expected fraction of cells in which a certain locus is accessible at any point in time, for a given cell state. The IQ inference algorithm is running on endpoints of specific manifold trajectories (e.g., epiblast to early mesoderm, or hematopoietic stem cells to myeloid progenitors). It can then fuse models from all manifold trajectories using a common set of motifs. We showed that using this approach, IQ maximizes model interpretability without compromising (and in fact while improving) precision and robustness.

The strength of the IQ algorithm relies on concrete assumptions about the interaction between TFs and DNA within nucleosome-depleted regions. Our data is demonstrating that model performance is enhanced by integrating the contribution of multiple potential binding sites, rather than considering only a single best sequence motif hit. We also show that inferring spatial preference curves and channeling the total predicted sequence affinity through a non-linear dose-response curve are also important for model accuracy. Another key feature of IQ is the modeling of pairwise TF interactions. These can represent physical heterodimers or other direct effects (e.g., Eomes-Gata4 predicted by IQ as a significant interaction and shown to have physical interaction potential in[33]). But cooperativity of two TFs over a trajectory can also be observed when the activity of one TF early in a trajectory transitions into the activity of another one later in the trajectory. IQ Interactions in these cases will represent forms of epigenetic memory. More work is needed to allow distinction between the different modes of action underlying the pairwise TF interactions captured by IQ.

Recent modeling trends are promoting the use of deep networks, and in particular, large language models (LLMs), for predicting chromosomal accessibility from sequence. IQ analysis defines a biophysics-inspired baseline for such models, and we suggest that it is the method of choice for understanding and further exploring CRE dynamics, given both its expressive power and direct functional

relevance. Whenever LLMs (or any black-box architecture) can outperform significantly the principled physical model outlined here, it will suggest the existence of mechanisms that are currently unappreciated and lead to new discoveries. Importantly, we already observe improved performance of ensembles of IQ and DL models. Understanding the residual benefits gained from such ensembles can lead to possible new insight into the biological mechanisms ensuring CRE specificity.

The modeling strategy of IQ avoids linkage between motifs and the expression of genes coding for the TFs they represent. Motif-TF association remains ambiguous, given the complexity of the TF repertoire regulating any cell state. In most studied conditions, a large number of TFs are expressed simultaneously, and their effective regulatory role may be driven by post-translational modification and protein-protein interactions. Furthermore, rich TF repertoires are organized in families that show massive overlap in sequence binding preferences, and in many instances, members of one family are regulating distinct differentiation lineages (e.g., GATA factors in gastrulation). IQ results should therefore *not* be interpreted at this stage as outlining a gene regulatory network with complex feedback and implied dynamics.

To move toward more complete modeling of genome regulation over a manifold, IQ (or IQ-like) models would have to be extended to include more explicit models of TF identities and activities. IQ can be enhanced to go beyond regression of ATAC differentials over trajectory endpoints and toward modeling of accessibility *kinetics* across developmental and fine-grained trajectories or an entire manifold. Aiming at models that involve an explicit temporal dimension[37,38] would be essential to derive predicted regulatory kinetics. Only once this can be demonstrated, the road toward unified analysis of predicted TF binding and downstream transcriptional effects will be open. Even before this grand challenge can be tackled, the precision and simplicity of the IQ model can lead to a mechanistic understanding and follow-up experiments interrogating the regulatory effect of specific factors in trans and specific sequences in cis.

Methods

The IceQream Model Overview

IceQream (IQ) is a quantitative framework designed to infer chromosomal accessibility differentials from sequence and epigenetic features. It is a hierarchical model based on a small number of motif models $PB_1..PB_n$ that compute *energies* from typically short (500 bp)
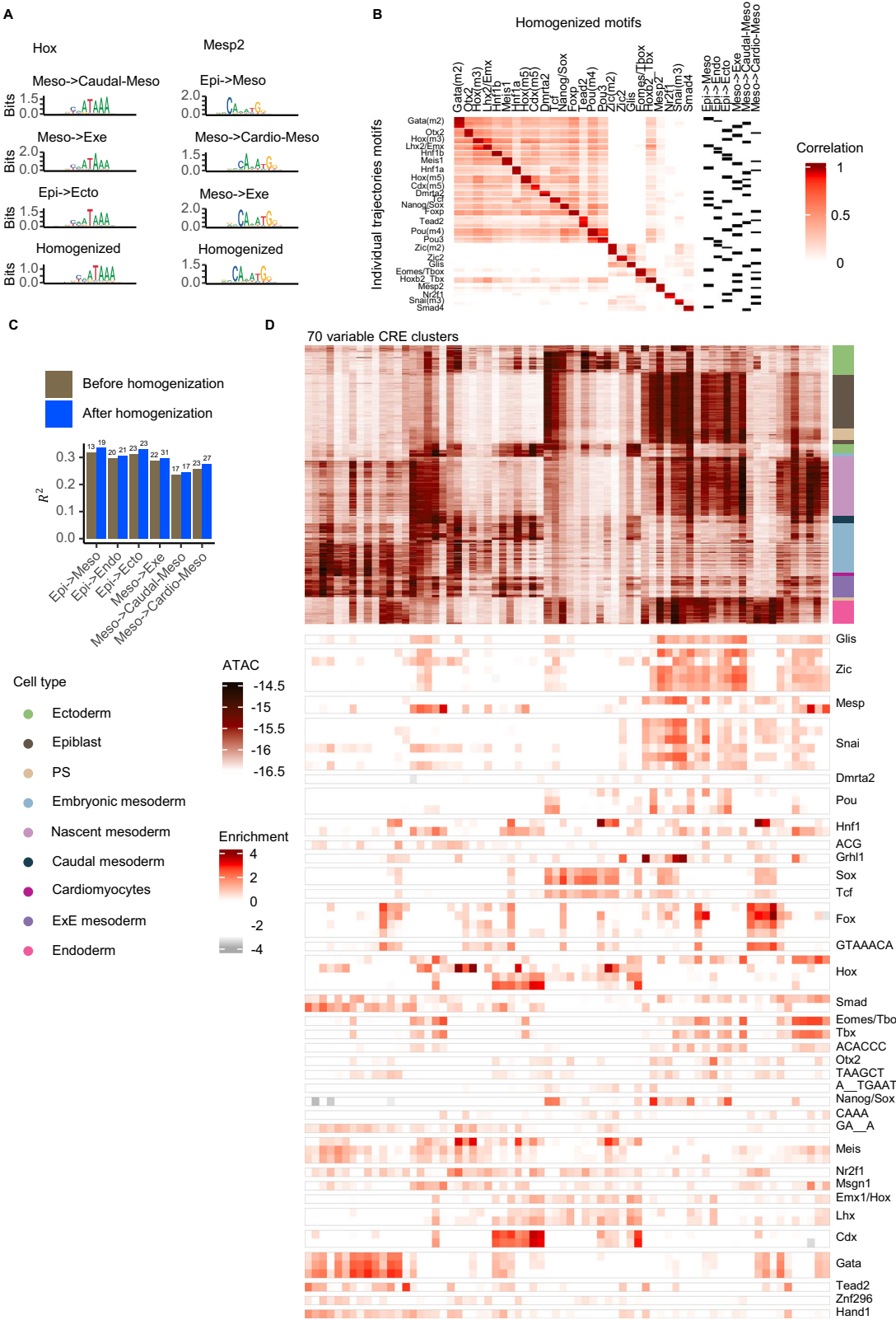
Fig. 4 | IQ model homogenization over multiple trajectories. A Comparison of sequence motifs for Hox and Mesp2 transcription factors before and after (bottom row) homogenization across three trajectories in the mouse gastrulation manifold. **B** Matrix showing the PSSM (Position-Specific Scoring Matrix) correlation between 63 motif models for individual trajectories (rows) and 25 homogenized models (columns). The rightmost columns indicate the trajectory of each model. **C** Model performance (R^2) before (brown) and after (blue)

homogenization for mouse gastrulation trajectories. The number above the bar indicates the number of motifs in each model. **D** CRE dynamics across the full mouse gastrulation manifold. Top heatmap shows the normalized ATAC signal for 70 variable CRE clusters (columns) across 369 metacells (rows), colored by their cell type (right annotation). The bottom heatmap shows the log2 enrichment of 61 full-manifold motifs in the CRE clusters. TF motifs are manually grouped into families of transcription factors (TFs).
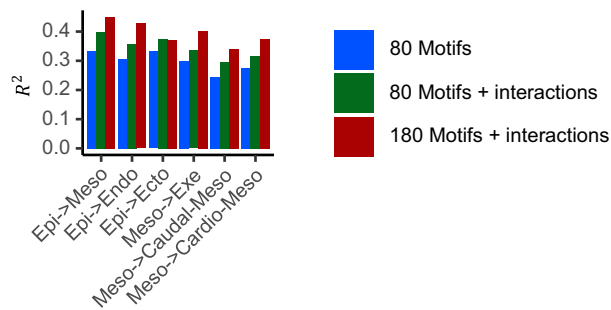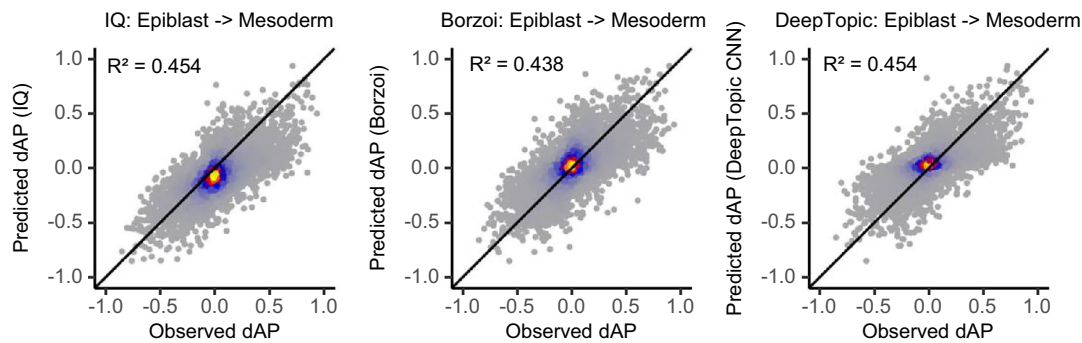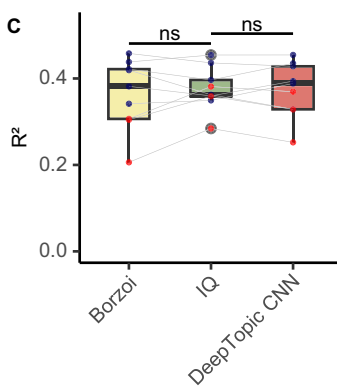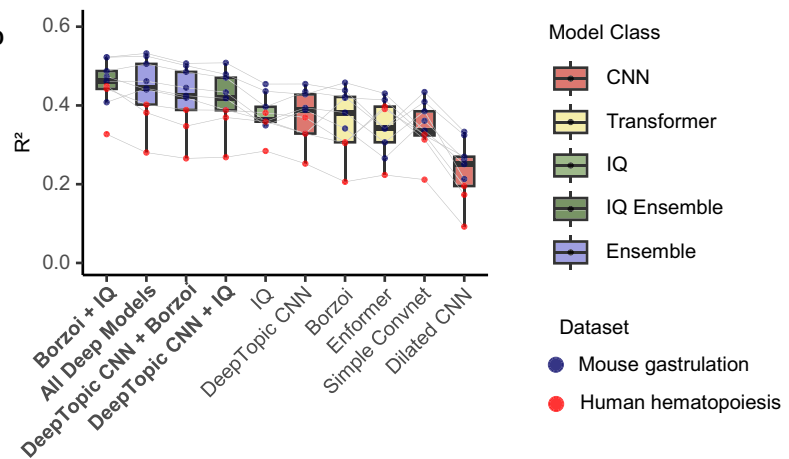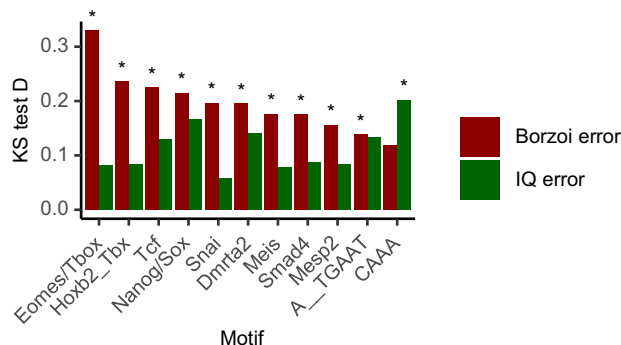
DNA sequences around hotspots of accessibility (and not from the entire genomic sequence). The energies are integrated using a generalized non-linear logistic model employing a basis of *response functions* $f_i \in F$ that channel the motif energies through either positive or negative non-linear effects.

An IQ model is thus regressing the differential Access Probability *dAP* from the binding energies predicted from sequence by a set of

motif models PB_i, and the epigenetic features E_k:

$$dAP = \sigma \left(\sum_{i=1}^{n} \sum_{j=1}^{l} \beta_{i,j} f_j(PB_i) + \sum_{(i_1,i_2) \in I} \sum_{j=1}^{l} \gamma_{k,j} f_j\left(PB_{i_1} PB_{i_2}\right) + \sum_{k=1}^{p} \sum_{j} \delta_{k,j} f_j(E_k) \right)$$

(1)

Fig. 5 | Benchmarking of IQ and leading deep learning models. A Model performance (R²) for IQ trajectory models using the full set of the gastrulation manifold motifs, without interactions (blue) and with interactions (green), compared to an expanded 180-motif IQ model with interactions (red) trained on our gastrulation data. **B** Scatter plots comparing observed versus predicted differential accessibility (dAP) scores using IQ (left), Borzoi (middle), and DeepTopic (right) models. Points are colored by local point density from low (gray) to high (blue to red). **C** R² values for three models and mouse (blue) or human (red) trajectories ($n = 6$ mouse trajectories, $n = 3$ human trajectories). Gray lines connect trajectories across models. Box plots show median (center line), first and third quartiles (box bounds), and whiskers extending to 1.5× the interquartile range. Paired two-sided Wilcoxon signed-rank tests with Benjamini-Hochberg correction: Borzoi vs IQ V(8) = 19, $p = 0.813$, r = 0.118, 95% CI = [−0.054, 0.025]; Borzoi vs DeepTopic V(8) = 14, $p = 0.813$, r = 0.316, 95% CI = [−0.046, 0.033]; IQ vs DeepTopic V(8) = 20, $p = 0.813$, r = −0.079, 95% CI = [−0.034, 0.031]. No significant differences between models

were observed. **D** R² values for individual models and ensembles, ranked by mean performance from best (left) to worst (right) ($n = 6$ mouse trajectories, $n = 3$ human trajectories). Points and colors as in panel (**C**). Gray lines connect trajectories across models. Box colors indicate model class. Box plots show median (center line), first and third quartiles (box bounds), and whiskers extending to 1.5× the interquartile range. **E** Kolmogorov-Smirnov test D statistics comparing motif energies in model-specific errors versus all correct predictions ($n = 141$, RMS < 0.05 in all models). Red bars: Borzoi errors ($n = 680$, RMS > 0.1 in Borzoi, <0.05 in IQ and ensemble). Green bars: IQ errors ($n = 304$, RMS > 0.1 in IQ, <0.05 in Borzoi and ensemble). Asterisks indicate FDR < 0.05. Significant FDR-adjusted q-values for Borzoi errors: Eomes/Tbox (3.5×10^{-10}), Tcf (9.8×10^{-5}), Hoxb2_Tbx (4.4×10^{-5}), Nanog/Sox (2.0×10^{-4}), Dmrta2 (8.1×10^{-4}), Snai (8.1×10^{-4}), Meis (3.6×10^{-3}), Smad4 (3.6×10^{-3}), Mesp2 (1.4×10^{-2}), A_TGAAT (4.3×10^{-2}); IQ errors: CAAA (1.5×10^{-2}). Source data are provided as a Source Data file.

Where:
- σ is the logistic function
- f_j are non-linear transformations of binding energies
- $\beta_{i,j}$ are the coefficients for each motif model and non-linear transformation
- I are a set of interaction pairs between motifs
- $\gamma_{k,j}$ are the coefficients for TF-TF interactions
- $\delta_{k,j}$ are the coefficients for epigenetic features

The complete IQ algorithm typically requires approximately 17 min of training time for a single differentiation trajectory on a 32-core CPU system, with memory usage peaking at approximately 30 GB.

Physical Binding Models

A motif model (also denoted as a physical binding model or PBM) in IQ consists of two components, a Position Weight Matrix (PWM) and a spatial integration module. PWM_k models the pointwise sequence preference in a window of L bps using position-specific nucleotide probabilities $P_j^k \cdot PWM_k(s_1..s_{L_k}) = \sum_j \log_2(P_j^k[s_j])$. To consider the contributions of all possible binding sites in a sequence of length len(S), rather than simply taking the maximum binding site, the model calculates the weighted average of the PWM in a running window across the sequence, with weights taken by the inferred spatial factors w_i^k. The absolute binding energy for a sequence $PB^{abs}(S)$ is then defined as:

$$PB_k^{abs}(S) = \log_2 \left(\sum_{i=1}^{len(S)-L_k+1} 2^{(PWM_k(S[i..(i+L_k)]) \cdot w_i)} \right) \quad (2)$$

We normalize absolute binding energies to facilitate homogeneous modeling, setting:

$$PB_k(S) = \min \left(\max \left(10 + PB_k^{abs}(S) - Q_{S' \in NORM}^{0.99} \left(PB_k^{abs}(S') \right), 0 \right), 10 \right) \quad (3)$$

Where $Q^{0.99}$ defines the 99th quantile and *NORM* is a set of sequences drawn from some background set. Following normalization, PB values of 10 imply maximum binding affinity, and values of 0 are assumed to have no affinity, allowing a range spanning three orders of magnitude but not more.

Normalized energies are then transformed through a set of logistic functions with chosen parameters reflecting potential levels of TF concentrations:

$$f_j(B) = \frac{\mu_j}{1 + \exp\left(-\alpha_j \left(B - \beta_j\right)\right)} \quad (4)$$

Where α_j, β_j and μ_j are logistic functions shape parameters. In this work, we used $k = 4$ parametrizations in F.

Estimation of access probabilities (AP)

We use pooled ATAC reads over a metacell or group of metacells to define a genome-wide raw coverage track u_i. We used 20 bp resolution in current applications. To take into account regional and nucleotide composition bias, we normalize raw coverage by a (punctured) regional mean:

$$un_i = u_i - \text{mean}_{1k \leq |i'-i| < 20k} u_{i'} \quad (5)$$

We define peaks as local maxima of un_i which are larger than the 98th quantile of the normalized coverage $Q^{0.98}(un_i)$. Peaks that are less than 500 bp apart are merged. The peak raw intensity up_i is then defined by the mean of the normalized coverage in the 500 bp around the peak center.

To scale the peak raw intensity and take into consideration variation in overall chromosomal promiscuity, we identify a set of constitutive peaks i_{const} that are empirically highly accessible and show low variability between conditions. Such loci are, for example, TSSs of housekeeping genes. We then compute a normalizer for each state using a high percentile ν of raw coverage for constitutive peaks:

$$\phi_{const} = Q_{i \in i_{const}}^{\nu}(up_i) \quad (6)$$

In applications, we used $\nu = 0.85$. We assume that ϕ_{const} coverage implies a fully accessible element and that all other values are scaling accordingly:

$$AP_i = \phi_{const}^{-1} up_i \quad (7)$$

To calculate the differential Access Probability (dAP) for regression analysis, IceQream (IQ) uses two manifold states as input, which may be individual metacells or groups of metacells enhancing coverage. The dAP, initially ranging from −1 to 1, is transformed to a 0 to 1 range for logistic regression.

Motif regression using PREGO

We adapted PREGO[39] to allow regression of PWMs and spatial parameters given either dAP values directly, or model residual values $dAP - mod(S)$ at different stages of the IQ algorithm as discussed below. Briefly, PREGO detects PWM seeds by screening through all DNA words (k-mers) of a predefined structure (K positions with a potential gap). Next, a seed k-mer is selected (or several seeds) with maximal correlation between the number of appearances in a sequence and the dAP signal. PREGO then maximizes the R² values of one motif model *PB* by brute force local optimization of PWM weights and spatial curve parameters. The algorithm terminates when it converges on a local maximum. As a direct extension, we allow PREGO regression to consider simultaneously several objectives dAP^s defining differential AP in multiple trajectories s, where the optimizer goal is
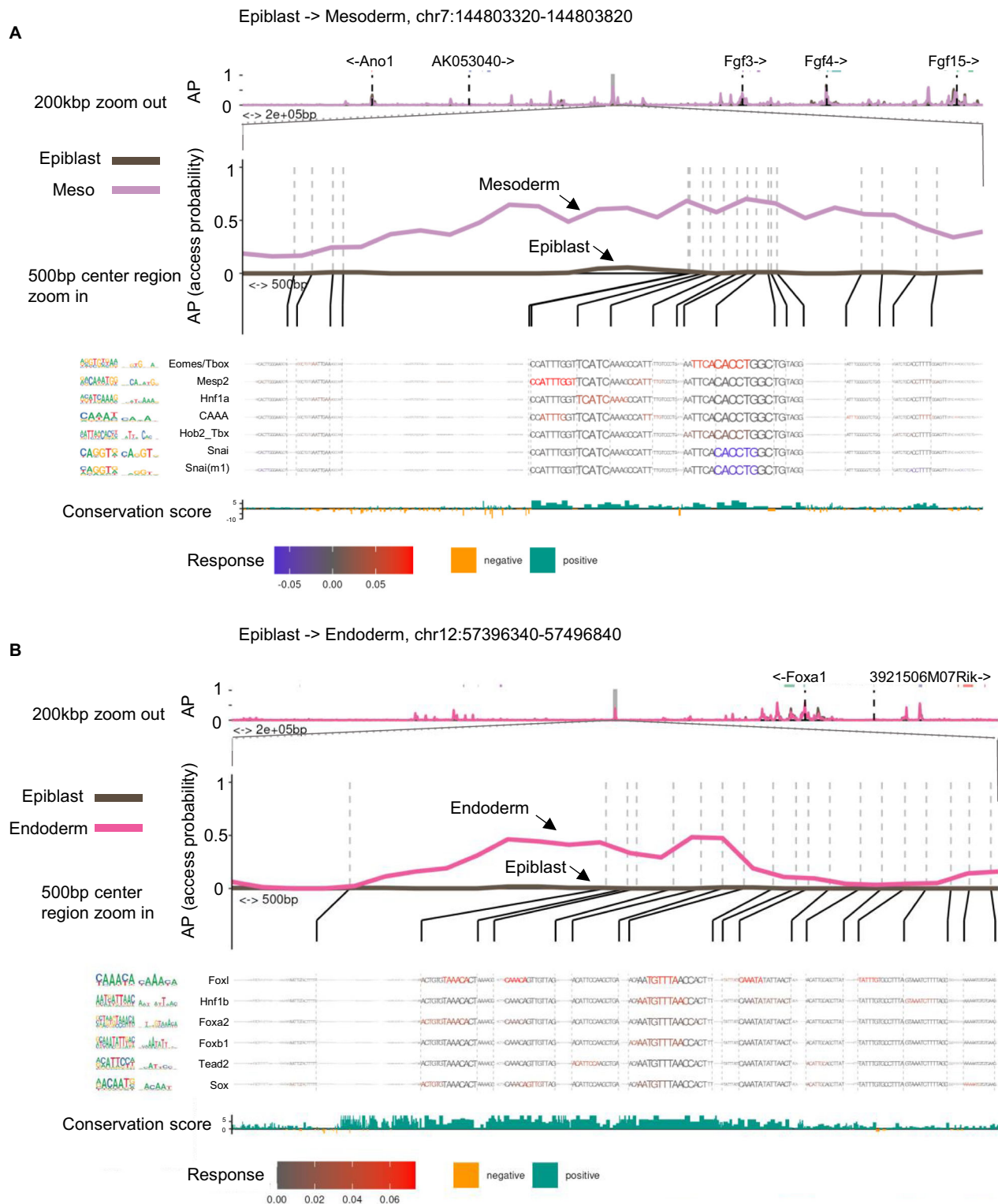
Fig. 6 | IQ-plots enable detailed architectural analysis of individual cis-regulatory elements. A, B - IQ-plots (IQP) for the Fgf3 locus in the epiblast to mesoderm trajectory (**A**) and the Foxa1 locus in the epiblast to endoderm trajectory (**B**). Each IQ-plot consists of three main components: Top panels: Normalized ATAC signal for the respective trajectories at 200 kbp (upper) and 500 bp (lower) resolutions. Black dashed lines indicate transcription start sites (TSS) with gene names above. Colored lines above plots represent exons. Middle panel: 500 bp DNA sequence, focusing on central gray loci from the top panel, with letter size indicating binding affinity to any model. Letter colors represent predicted contributions of each TF motif (rows) to the dAP. Red letters indicate predicted positive dAP (chromatin opening) while blue letters indicate negative dAP (chromatin closing). Bottom panel: Conservation score for each base pair. Green indicates positive scores, and orange indicates negative scores. Vertical gray and black guide lines connect corresponding positions between the top panels and the DNA sequence panel. Source data are provided as a Source Data file.

defined as the mean R² over all objectives. This added multiple objectives approach is used for IQ model homogenization.

IQ step 1 – motif library screen

Given a dAP signal, IQ is initiated using a library of PWMs and a trivial (uniform) spatial component. We compute $PB_m(S_i)$ for all models and sequences and retain for further analysis only models with $|cor(PB_m, dAP)| > T_c$, where T_c is set to a very low value (e.g., 0.05) and can be further downgraded if no motifs are identified. IQ can add to the filtered library PWMs inferred using de-novo PREGO analysis of the dAP signal. We, however, use de novo screening carefully as it may reduce generalization by overfitting the data.

IQ step 2 – Further motif filtering using a simple model with L1 regularization

In IQ step 2, we fit an initial logistic regression with L1 regularization model, with an elastic-net based logistic regression model using *glmnet*[40], with the *binomial* family, a *logit* link function, with alpha = 1 and lambda = 1e-5. We then filter PWMs with near-zero coefficients based on a defined threshold.

At this stage, we also add epigenetic features to the model based on user input. These can include histone markers of each peak, but also regional averages of accessibility within topologically associated domains (TAD, see below). Sequence features such as GC content and di-nucleotide signatures can also be incorporated.

IQ step 3 – PREGO-based motif cluster consolidation

We cluster the BP_i vectors of the motifs retained in step 2 into B (user-defined, default 50) clusters (or blocks) I_b using hierarchical clustering. We then break down the model predictions in level 2 into B components: $Mod_b = \sum_{i \in I_b} \beta_i(PB_i)$. For each block, we can now apply PREGO regression on Mod_b, with an initial condition using the PWM $i \in I_b$ showing maximal model contribution. This derives one PBM, or motif model (PWM and spatial curve), that approximates the contributions of $|I_b|$ motifs in level 2, leading to a concise (and empirically more generalizable) model. Each inferred PB_b is then transformed using the *response functions* in F. Together with the epigenetic features, these constitute the inputs for the logistic regression model of this stage, again using the *binomial* family, a *logit* link function, with alpha = 1 and lambda = 1e-5.

IQ step 4 – second round of filtering and consolidation

We optionally apply a second round of consolidation, similarly to step 3, except with the changed definition of $Mod = \sum_{i \in I_b} \sum_{j=1}^{l} \beta_{i,j} f_j(PB_i)$ to include the non-linear transformations. We also filter motif models based on a residual R² threshold. Residual R² is defined by the loss of R² in the model inferred in level 3 and a similar model from which one motif model was eliminated.

IQ step 5 – adding interactions

To add the interaction component $\sum_{(i_1, i_2) \in I} \sum_{j=1}^{l} \gamma_{k,j} f_j(PB_{i_1} PB_{i_2})$ to the model, we iterate over pairs of step 4 models and consider the utility of their product in improving regression accuracy. We only consider pairs in which at least one of the motif models has a residual R² above a defined threshold. These interactions are added to the input features of step 4 and are used in the final logistic regression model with alpha = 1 and lambda = 1e-5.

IQ model homogenization

The set of inferred PBMs predictive for a single dAP score for a differentiation trajectory in a manifold of interest may overlap with PBMs with other trajectories in the same manifold. To ease interpretation and obtain a full manifold model, given a set of IQ models per trajectory from step 4, we use a homogenization approach to reduce the number of PBMs. First, we cluster the PB_i from all IQ models using

hierarchical clustering with the "complete" method and a user-defined number of clusters. For each cluster, we then fit a new PBM with multiple dAPs as the learning objective for PREGO. The objective for a cluster includes any dAP that a PBM in the cluster was used in its IQ regression for the single trajectory, using uniform weights.

R² comparisons (Fig. 2E)

In order to assess the contribution of the different components of IQ and other possible models of regression, we conducted comparisons with the full IQ and other models trained on randomly chosen 80% of the peaks filtered to a distance >5 kbp from any TSS. Each model was used to predict the dAP on the 20% held-out test set of peaks, and the R² of the models' predictions versus observed dAP was used as the accuracy metric. The *GLM-IQ* model does not incorporate the non-linear transformations L. The *GAM-IQ* model uses IQ PBMs, but incorporates the default non-linearities as in the GAM[41] framework. *XGBoost-IQ* uses IQ PBMs but conducts logistic regression using *xgboost*[42]. With the same R² approach, we also compare the accuracy of models with an increasing number of PWMs from the literature to IQ, where literature PWMs were chosen based on their Pearson correlation of columns of $B_{peak,PWM}$ to the dAP. In Fig. S3A, we added a comparison of choosing the PWMs as the most correlated PWM for each *SCENIC* cluster. In addition, we tested an approach of fusing motifs across their *SCENIC* clusters instead of the de novo *hclust* approach. We also conducted experiments of training and testing IQ on distal (> 5 kbp), proximal (500–5 kbp), and promoter (< 500 bp) CREs shown in Fig. S3D.

Low and Medium affinity binding site model contribution (Fig. 3B,C)

To generate the data in Fig. 3B, we created a model with altering energy computation to max based integration:

$$PB_k^{abs}(S) = \max_i PWM_k\left(S[i..(i+L_k)]\right) \cdot w_i \qquad (8)$$

and retraining an alternative logistic regression model using these energies as the input features across the same train/test splits.

As a more direct test (Fig. 3C), we created for each motif model two sets of peaks based on the number of sites above a quantile of binding energy scores (one-hit versus multiple hits) and a threshold on the maximum binding energy score across any position across the peaks. Peaks with a single best motif hit above the threshold were classified as single-high ($dAP_{single-high}$), and peaks with multiple hits but all below the threshold were classified as multiple low-affinity peaks ($dAP_{multi-low}$). We then calculated a low-affinity score for each PBM using:

$$LAS_{PBM} = mean\left(dAP_{multi-low}\right) - mean\left(dAP_{single-high}\right) \qquad (9)$$

A positive LAS_{PBM} indicates that the PBM has a stronger effect when multiple low-affinity sites are present, while a negative score suggests a stronger effect for a single high-affinity site.

AP normalization comparison to TF-IDF normalization

To compare AP normalization with the common TF-IDF normalization, we implemented TF-IDF transformation on the raw metacell ATAC count matrix. Term frequency (TF) was calculated as the fraction of peaks in each metacell by dividing each metacell's counts by its total count sum. The TF values were then log-transformed using $\log(1 + TF \times 10^4)$. Inverse document frequency (IDF) was calculated as $\log(1 + N/df)$, where N is the total number of metacells and df is the number of metacells containing each peak. The final TF-IDF matrix was computed as the element-wise product of TF and IDF values. Due to the dominance of the constitutively accessible peaks, we also tested a

probability-normalized TF-IDF variant where values were first divided by the 60th percentile of constitutive peaks, then capped at 1, and finally normalized to the [0,1] range per metacell to match the AP range. For comparison, Pearson correlation matrices were computed between cell type accessibility profiles using each normalization method (TF-IDF, probability-normalized TF-IDF, and AP). Cell type-specific accessibility profiles were generated by averaging the normalized values across all metacells belonging to each cell type.

ChIP-Seq validation of selected PBMs

To validate the inferred motif models, we compared motif binding energies between ChIP-seq peaks and background regions. ChIP-seq datasets were obtained for key transcription factors: Eomes (embryoid body differentiation from mESCs[43]), Mesp1 (embryoid body differentiation from PSCs[44]), Nanog and Sox2 (embryonic stem cells[45]), and Oct4 (embryonic stem cells[46]).ChIP-seq peaks were defined using a threshold of $-\log2(1\text{-signal}) \geq 13$. For each transcription factor, we extracted motif binding energies from 500 bp regions centered on ChIP-seq peaks using the corresponding IceCream motif models. Background regions were generated by randomly sampling an equal number of 500 bp genomic intervals. Statistical significance was assessed using a two-sample Kolmogorov-Smirnov test comparing the distributions of motif binding energies between ChIP-seq peaks and background regions, testing the alternative hypothesis that ChIP-seq peaks have higher binding energies than background.

Nucleosome positioning data analysis

To validate IceQream's inferred spatial binding preferences, we compared them to experimental nucleosome positioning data using MNase-seq profiles from mouse embryonic stem cells[43]The MNase-seq data includes nucleosome occupancy profiles under different digestion conditions (1U, 4U, 16U, and 64U enzyme concentrations). For each transcription factor (Nanog/Sox, Hnf1a, Eomes/Tbox, and Gata), we identified peaks with motif energy scores above 9 and extracted MNase-seq signals in 2 kbp windows centered on these peaks. Missing values were replaced with zero, and signals were averaged across all peaks for each factor, with each position's signal divided by a 16-bp rolling mean.

Optional filtering of candidate pairwise interactions

The number of possible pairwise interactions scales with the square of the number of motif models, specifically $\left(\frac{n^2}{2}\right) - n$. For scenarios where n is chosen to be big enough, considering all pairwise interactions might not be optimal/feasible. We therefore developed strategies to rank and possibly filter candidate interactions. Before including interactions, we utilize either the max magnitude of the coefficients for each feature or the residual R^2 values for ranking the different motif model features. We then set a threshold or a rank to select which motif models will be considered as significant candidates for pairwise interactions. Pairwise interactions are then constructed between significant motif models, and possibly between pairs of significant and low-significant motif models.

IQ full manifold models for benchmarking comparisons

To enable a full comparison of the IQ approach to alternative models, we generated manifold models encompassing all the trajectories of the gastrulation or hematopoiesis manifolds. Specifically, we trained IQ models with 60 PBMs per trajectory, combined all the PBMs, and homogenized the full set from all trajectories into 180 full manifold PBMs. We divided the CREs into train, validation, and test sets based on chromosomes, with chr8/chr10 for validation and chr9/chr18 for testing (mouse gastrulation: $n = 80046$ train, 10286 validation, 8959 test; human hematopoiesis: $n = 45747$ train, 5705 validation, 3835 test), as was done for all the alternative models. Due to the larger number of PBMs in these models, we filtered the number of possible interactions

per model as described above. For the additional features, we also added predictions of the IQ model for the state ATAC values (trained only on the training chromosomes). We also incorporated energy features calculated across multiple lengths from the peak center (100, 300, 500, 1000, 2000 bp), resulting in a small performance improvement. For these most expressive variants of IQ, the number of total parameters is on the order of 10^5 per trajectory, whereas the alternative models we compare to have on the order of 10^8 total parameters.

Training and benchmarking the DL models for ATAC regression

We utilized the CREsted framework[34] (v1.3.0) to train the deep learning models using our gastrulation and hematopoiesis ATAC data with the same normalization scheme as for IQ. Peak values were normalized using crested.pp.normalize_peaks() with default top_k_percent=0.03. CREs were divided into train, validation, and test sets based on chromosomes, with chr8/chr10 for validation and chr9/chr18 for testing.

The analysis included six mouse gastrulation trajectories (Epiblast to Mesoderm, Epiblast to Endoderm, Epiblast to Ectoderm, Mesoderm to Extraembryonic mesoderm, Mesoderm to Cardiomyocytes, and Mesoderm to Caudal Mesoderm) and three human hematopoiesis trajectories (HSC to CLP, HSC to CMP, and HSC to MEP).

We tested multiple model architectures with varying input sequence lengths: ChromBPNet[47] (dilated CNN, 2114 bp), DeepTopic CNN[34] (500 bp), Simple Convnet (2114 bp), and Borzoi (2048 bp). CNN models were trained for 60 epochs with early stopping, batch sizes of 128–256, Adam optimizer (learning rate 1e-3), and CosineMSELogLoss (max_weight = 100). We tested between training directly on the differential ATAC values, versus training on the ATAC states and predicting differential values by the subtraction of states, and took the best result for each model. For ChromBPNet, the version using cell state ATAC values was chosen, and for the other models, training directly on differential ATAC signals was chosen. For Borzoi finetuning, we modified pretrained models (Borzoi_mouse_rep0 for mouse, Borzoi_human_rep0 for human) by replacing the head with a flattened output layer for scalar prediction per region. We employed two-phase training: initial finetuning on all regions (learning rate 1e-5) followed by 5 epochs of focused training on high-specificity regions (learning rate 5e-5, Gini threshold = 1.0) filtered using crested.pp.filter_regions_on_specificity, as recommended in CREsted. Borzoi used 2048 bp sequences with target_length = 64 bins to maintain compatibility with the pretrained model's binning structure.

For Enformer finetuning, we extracted embeddings from the pretrained Enformer model and trained a custom ParallelVectorMLP architecture on top of these frozen features. The ParallelVectorMLP processes each of the 896 Enformer output vectors (3072-dimensional each) through separate MLPs with hidden dimensions of [256, 128], focusing on the central 14 vectors to reduce computational complexity while maintaining performance. We used distributed training across multiple L40 GPUs with batch size 8, Adam optimizer (learning rate 1e-4), MSE loss, and trained for 4 epochs. The model aggregates the processed vectors through a final linear layer to produce scalar predictions for each trajectory's differential ATAC signal.

Ensemble methods and evaluation

Ensemble models combined individual predictions using linear regression trained on the chromosome-based training splits. For the R^2 comparisons in Fig. 5B–D, we chose model variants that provided the best performance on differential accessibility prediction. Model performance was evaluated using R^2 between predicted and observed differential accessibility on held-out test chromosomes. Statistical comparisons used Wilcoxon tests across trajectory pairs, and model-specific error distributions for transcription factor binding sites were analyzed using Kolmogorov-Smirnov tests (Fig. 5E).

IQ Plots

IQ plots (IQ-Ps) are a visualization tool developed to aid in the interpretation of IceQream models and the exploration of cis-regulatory elements (CREs) architectures. IQ-Ps combine several layers of information:

1. ATAC signal: APs at 200 kbp and 500 bp resolutions, for the two cell states that constitute the inputs to calculate a dAP for a trajectory of interest. Gene names of annotated TSSs in the 200 kbp are also marked.
2. DNA sequence: 500 bp DNA sequence with letter size scaled by binding energy to any PBM model.
3. Motif models' contributions: Color-coded letters representing predicted contributions of each PBM to the ATAC difference.
4. Representative logo plots for each motif model.
5. Conservation score: Conservation scores were taken from Sullivan et al.[29] for each base pair.

The resulting IQP provides a comprehensive view of the sequence features, binding preferences, and predicted regulatory effects for a given CRE.

Multiome RNA-seq analysis and metacell manifolds derivation

The gastrulation metacell[48] model includes 10X Multiome experiments obtained from published data[49], merged with new 10X Multiome data collected by us from E6.5-E7.5 stage mouse embryos. After low UMI counts, doublets, and outlier filtering, a total of 55,783 single-cell RNA profiles were used to generate a metacell model based on *metacell-2*[50] with an equivalent of 25 cells as the target number of cells per metacell. The resulting metacells *EGC* matrix was used in a projection[51] schema on our comprehensive mouse gastrulation atlas[37]. Cell state annotations were taken from the projected atlas metacells, and ordering of metacells across differentiation trajectories was also conducted using the projected metacells RNA signatures. The hematopoiesis metacell model used public BM CD34 + 10X Multiome data[52]. A total of 12947 single-cell RNA profiles were used to generate a metacell model with an equivalent of 35 cells as the target number of cells per metacell.

Nuclei isolation from mouse embryos for Multiome sequencing

All animal procedures were approved by the Institutional Animal Care and Use Committee and were performed in strict adherence to Weizmann Institute guidelines. Mice were monitored for health and activity and were given ad libitum access to water and standard mouse chow with 12 h light/dark cycles.

Mouse embryos from timed pregnant immune competent Hsd: ICR (CD−1) females at E6.5-E7.5 were harvested and dissected in ice-cold 1×PBS. Nuclei for the Multiome assay were then prepared from these embryos. In general, 45-70 K cells were used for nuclei isolation for each Multiome assay. Single nuclei suspension was obtained following the demonstrated low cell input nuclei isolation protocol from 10x Genomics, with minor changes. Briefly, cells were washed once with 500 µl cold PBS/0.04% BSA and resuspended in 50 µl PBS/0.04% BSA in a 200 µl DNase/RNase-free tube. Following a centrifuge at 300 rcf for 5 min at 4 °C, cell pellets were incubated in 45 µl chilled lysis buffer (10 mM Tris-HCl, pH7.4, 10 mM NaCl, 3 mM $MgCl_2$, 0.1% Tween-20, 0.1% Nonident P40 substitute, 0.01% digitonin, 1% BSA, 1 mM DTT) with freshly added RNase inhibitor (final concentration 1U/µl, Sigma, 3335399001) for 3 min on ice after gentle pipetting for 5 times. Afterwards, 100 µl chilled wash buffer (10 mM Tris-HCl, pH7.4, 10 mM NaCl, 3 mM $MgCl_2$, 1% BSA, 0.1% Tween-20, 1 mM DTT, 1U/µl RNase inhibitor) was added directly to the tube, followed by another 5 times gentle mixing. The suspension was spun at 500 rcf for 5 min at 4 °C. After removing 145 µl supernatant, 45 µl chilled diluted nuclei buffer (10x Genomics) was added to the pellet without dislodging the pellet. Followed by a centrifuge at 500 rcf for 5 min at 4 °C, the final nuclei were resuspended in 7 µl diluted nuclei buffer. 2 µl nuclei suspension was taken to quantify the concentration with a Countess II (Thermo) and to check the quality under an inverted microscope. >95% of nuclei stained positive for trypan blue, and the nuclei were found to have the expected morphology. Nuclei were further diluted to the suggested concentration range suitable for the preparation of a 4000-nuclei library if needed.

Multiome library preparation

Single nuclei suspensions were loaded on Chromium Next GEM Chip J (10x Genomics) to generate single-cell GEMs. Single-cell ATAC libraries and gene expression libraries were constructed separately following the manufacturer's instructions. Library size distribution and abundance were assessed with a D1000 or D5000 ScreenTape (Agilent), and their concentration was quantified with Qubit 4 Fluorometer (Thermo). Libraries were sequenced on an Illumina NovaSeq 6000 instrument in paired-end mode. Sequencing settings are: ATAC library: read1: 50 cycles, index1: 8 cycles, index2: 24 cycles, read2: 49 cycles. RNA library: read1:28 cycles, index1:10 cycles, index2: 10 cycles, read2: 90 cycles.

Reporting summary

Further information on research design is available in the Nature Portfolio Reporting Summary linked to this article.

Data availability

Data that support the findings of this work have been deposited in the Gene Expression Omnibus under accession code GSE305339. Previously published mouse gastrulation Multiome data that were reanalyzed here are available under GSE205117. Previously published Human BM CD34+ Multiome data is available under GSE200046. All other materials and data supporting the findings of this study are available from the corresponding authors upon request. Source data are provided with this paper.

Code availability

IQ code and workflows used in this work are available in the GitHub repository at https://github.com/tanaylab/iceqream. All scripts and auxiliary scripts used to analyze data and generate figures have been deposited in the GitHub repository at https://github.com/tanaylab/IQ-Paper[53].

References

1. Buenrostro, J. D. et al. Single-cell chromatin accessibility reveals principles of regulatory variation. *Nature* **523**, 486–490 (2015).
2. Jaitin, D. A. et al. Massively parallel single-cell RNA-Seq for marker-free decomposition of tissues into cell types. *Science* **343**, 776–779 (2014).
3. Macosko, E. Z. et al. Highly parallel genome-wide expression profiling of individual cells using nanoliter droplets. *Cell* **161**, 1202–1214 (2015).
4. Picelli, S. et al. Full-length RNA-seq from single cells using Smart-seq2. *Nat. Protoc.* **9**, 171–181 (2014).
5. Granja, J. M. et al. Single-cell multiomic analysis identifies regulatory programs in mixed-phenotype acute leukemia. *Nat. Biotechnol.* **37**, 1458–1465 (2019).
6. Ma, S. et al. Chromatin potential identified by shared single-cell profiling of RNA and chromatin. *Cell* **183**, 1103–1116.e20 (2020).
7. Trevino, A. E. et al. Chromatin and gene-regulatory dynamics of the developing human cerebral cortex at single-cell resolution. *Cell* **184**, 5053–5069.e23 (2021).
8. de Boer, C. G. & Regev, A. BROCKMAN: deciphering variance in epigenomic regulators by k-mer factorization. *BMC Bioinforma.* **19**, 253 (2018).
9. Avsec, Ž. et al. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nat. Genet.* **53**, 354–366 (2021).

10. Schep, A. N., Wu, B., Buenrostro, J. D. & Greenleaf, W. J. chromVAR: inferring transcription-factor-associated accessibility from single-cell epigenomic data. *Nat. Methods* **14**, 975–978 (2017).

11. Yuan, H. & Kelley, D. R. scBasset: sequence-based modeling of single-cell ATAC-seq using convolutional neural networks. *Nat. Methods* **19**, 1088–1096 (2022).

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

13. Ashuach, T., Reidenbach, D. A., Gayoso, A. & Yosef, N. PeakVI: A deep generative model for single-cell chromatin accessibility analysis. *Cell Rep. Methods* **2**, 100182 (2022).

14. Shrikumar, A., Greenside, P. & Kundaje, A. Learning important features through propagating activation differences. Preprint at *arXiv* https://doi.org/10.48550/arXiv.1704.02685 (2019).

15. Huang, C. et al. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nat. Genet.* **55**, 2056–2059 (2023).

16. Karollus, A., Mauermeier, T. & Gagneur, J. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biol.* **24**, 56 (2023).

17. Sasse, A. et al. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nat. Genet.* **55**, 2060–2064 (2023).

18. Kathail, P. et al. Current genomic deep learning models display decreased performance in cell type-specific accessible regions. *Genome Biol.* **25**, 202 (2024).

19. Levine, M. & Tjian, R. Transcription regulation and animal diversity. *Nature* **424**, 147–151 (2003).

20. Francois, S. & Eileen, E. M. F. Transcription factors: from enhancer binding to developmental control. Nat. Rev. Genet. **13**, 613–626 (2012).

21. Lee, B.-K. et al. Cell-type specific and combinatorial usage of diverse transcription factors revealed by genome-wide binding studies in multiple human cells. *Genome Res* **22**, 9 (2012).

22. Gifford, C. A. et al. Transcriptional and epigenetic dynamics during specification of human embryonic stem cells. *Cell* **153**, 1149–1163 (2013).

23. Stormo, G. D. & Zhao, Y. Determining the specificity of protein–DNA interactions. *Nat. Rev. Genet.* **11**, 751–760 (2010).

24. Badis, G. et al. Diversity and complexity in DNA recognition by transcription factors. *Science* **324**, 1720–1723 (2009).

25. Kim, K. et al. Epigenetic memory in induced pluripotent stem cells. *Nature* **467**, 285–290 (2010).

26. Long, H. K., Prescott, S. L. & Wysocka, J. Ever-changing landscapes: transcriptional enhancers in development and evolution. *Cell* **167**, 1170 (2016).

27. Heinz, S. et al. Simple combinations of lineage-determining transcription factors prime cis-regulatory elements required for macrophage and B cell identities. *Mol. Cell* **38**, 576–589 (2010).

28. Castro-Mondragon, J. A. et al. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Res.* **50**, D165–D173 (2022).

29. Jolma, A. et al. DNA-dependent formation of transcription factor pairs alters their binding specificity. *Nature* **534**, S15–S16 (2016).

30. Kulakovskiy, I. V. et al. HOCOMOCO: towards a complete collection of transcription factor binding models for human and mouse via large-scale ChIP-Seq analysis. *Nucleic Acids Res.* **46**, D252–D259 (2018).

31. Cusanovich, D. A. et al. A single-cell atlas of in vivo mammalian chromatin accessibility. *Cell* **174**, 1309–1324.e18 (2018).

32. Salton, G. & Buckley, C. Term-weighting approaches in automatic text retrieval. *Inf. Process. Manag.* **24**, 513–523 (1988).

33. Xie, Z. et al. DNA-guided transcription factor interactions extend human gene regulatory code. *Nature* **641**, 1329–1338 (2025).

34. Kempynck, N. et al. CREsted: modeling genomic and synthetic cell type-specific enhancers across tissues and species. Preprint at *bioRxiv* https://doi.org/10.1101/2025.04.02.646812 (2025).

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

36. Linder, J., Srivastava, D., Yuan, H., Agarwal, V. & Kelley, D. R. Predicting RNA-seq coverage from DNA sequence as a unifying model of gene regulation. *Nat. Genet.* **57**, 949–961 (2025).

37. Mittnenzweig, M. et al. A single-embryo, single-cell time-resolved model for mouse gastrulation. *Cell* **184**, 2825–2842.e22 (2021).

38. Qiu, C. et al. Systematic reconstruction of cellular trajectories across mouse embryogenesis. *Nat. Genet.* **54**, 328–341 (2022).

39. Tanay, A. Extensive low-affinity transcriptional interactions in the yeast genome. *Genome Res* **16**, 962–972 (2006).

40. Friedman, J. H., Hastie, T. & Tibshirani, R. Regularization paths for generalized linear models via coordinate descent. *J. Stat. Softw.* **33**, 1–22 (2010).

41. Wood, S. N. Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **73**, 3–36 (2011).

42. Chen, T. & Guestrin, C. XGBoost: a scalable tree boosting system. in *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* 785–794 (Association for Computing Machinery, New York, NY, USA, 2016).

43. Schüle, K. M. et al. Eomes restricts Brachyury functions at the onset of mouse gastrulation. *Dev. Cell* **58**, 1627–1642.e7 (2023).

44. Lin, X. et al. Mesp1 controls the chromatin and enhancer landscapes essential for spatiotemporal patterning of early cardiovascular progenitors. *Nat. Cell Biol.* **24**, 1114–1128 (2022).

45. Chronis, C. et al. Cooperative binding of transcription factors orchestrates reprogramming. *Cell* **168**, 442–459.e20 (2017).

46. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).

47. 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 *bioRxiv* https://doi.org/10.1101/2024.12.25.630221 (2025).

48. Baran, Y. et al. MetaCell: analysis of single-cell RNA-seq data using K-nn graph partitions. *Genome Biol.* **20**, 206 (2019).

49. Argelaguet, R. et al. Decoding gene regulation in the mouse embryo using single-cell multi-omics. 2022.06.15.496239 Preprint at *bioRxiv* https://doi.org/10.1101/2022.06.15.496239 (2022).

50. Ben-Kiki, O., Bercovich, A., Lifshitz, A. & Tanay, A. Metacell-2: a divide-and-conquer metacell algorithm for scalable scRNA-seq analysis. *Genome Biol.* **23**, 100 (2022).

51. Ben-Kiki, O. et al. MCProj: metacell projection for interpretable and quantitative use of transcriptional atlases. *Genome Biol.* **24**, 220 (2023).

52. Persad, S. et al. SEACells infers transcriptional and epigenomic cellular states from single-cell genomics data. *Nat. Biotechnol.* **41**, 1746–1757 (2023).

53. Bercovich, A. et al IceQream: quantitative chromosome accessibility analysis using physical TF models https://github.com/tanaylab/IQ-paper, https://doi.org/10.5281/zenodo.16936494 (2025).

Acknowledgments

Author contributions

A.B., A.L., and A.T. conceived the project. S.C and R.S. performed the experiments. A.B., A.L., and A.T. analyzed the data. A.B., A.L., and A.T. developed the computational algorithms and analysis pipeline. A.B., A.L., and M.E. performed the benchmarking versus deep models. A.B., A.L., and A.T. wrote the manuscript. A.T. and Y.S. supervised the project. All authors reviewed the results and approved the final version of the manuscript.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information The online version contains supplementary material available at https://doi.org/10.1038/s41467-025-63925-x.

Correspondence and requests for materials should be addressed to Amos Tanay.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.