nature portfolio

Peer Review File

A foundation model of transcription across human cell types

Corresponding Author: Professor Raul Rabadan

Parts of this Peer Review File have been redacted as indicated to maintain the confidentiality of unpublished data.
This file contains all reviewer reports in order by version, followed by all author rebuttals in order by version.

Version 0:

Reviewer comments:

Referee #1

(Remarks to the Author)
Here, Fu and Mo et al introduce the General Expression Transformer (GET) – a novel model to enhance the understanding of transcription regulation across 213 diverse human cell types. The
GET model is extensively trained on chromatin accessibility data, enabling it to learn the syntax of transcriptional regulation. The authors demonstrate that GET can accurately predict gene expression in both previously encountered and novel cell types, demonstrating compatibility with different sequencing platforms. A notable feature of GET is its capacity for zero-shot prediction of reporter assay readout in new cell types, distinguishing it from preceding models, delivering impressive performance in identifying crucial regulatory elements. Through its application, GET study unveils rich regulatory insights for many genes within the explored cell types, identifying both novel and known upstream regulators of fetal hemoglobin, and assembling a comprehensive interaction catalog of human transcription factors and coactivators. Furthermore, GET reveals a significant interaction between the transcription factors PAX5 and the retinoic acid receptor family, shedding light on a potential disease-driving mechanism associated with lymphoma. In summary, the versatility and precise understanding of transcription regulation demonstrated by GET significantly contributes to advancing the understanding of noncoding genetic variants. Moreover, it lays a robust foundation for the de novo design of cell-type specific regulatory frameworks, showcasing its potential as a transformative tool in the domain of transcriptional regulation.

Major Strengths:
• The GET model architecture is well-designed and justified to capture regulatory grammar. Using self-supervised pretraining on 213 scATAC-seq datasets allows the model to learn generalizable regulatory principles.
• GET demonstrates remarkable accuracy in predicting gene expression for unseen cell types, achieving near experimental-level performance (Pearson r ~0.94). This significantly outperforms all baselines such as accessibility, mean expression, etc.
• The model demonstrates impressive adaptability across diverse assays through effective transfer learning. Predictions on ATAC-seq and zero-shot lentiMPRA experiments are compelling validations of its versatility.
• The comprehensive regulatory analyses enabled by GET are a major highlight. It effectively identifies distal enhancers, regulators of fetal hemoglobin, and downstream targets of key regulators like GATA1. The motif aggregation strategy is powerful.
• The GET Catalog database of predicted regulatory networks and TF interactions is an invaluable resource for the community, where the interactive website is well-designed to disseminate this knowledge.
• The study is technically strong, reflecting rigorous model development, training, evaluation and interpretation. The multi-pronged validation and biological insights demonstrate the usefulness of this foundation model.

Limitations and suggestions:

• My major concern centers on the validation of the approach to uncover novel biological insights. GET uncovers mechanism of germline mutation in disorder region of transcription factors – at this level I would expect a much deeper demonstration to uncover novel biology, benchmarking against other approaches to show that these insights are only possible with GET.
• The authors have demonstrated their ability to accurately predict gene expression in left-out cell types, using astrocytes as an example. However, it's worth noting that astrocytes are well-characterized and highly specialized cells where transcription is rigorously regulated. It would be greatly appreciated if the authors could extend their findings to a broader range of cell types. I anticipate particular interest in the accuracy of prediction when dealing with cell types that exhibit more dynamic chromatin landscapes, such as stem cells. If there are variations in prediction accuracy among different cell types, it could be intriguing to investigate further what could contribute to the variation.
• Related to the previous point, I am curious to know whether the GET model can also predict gene expression in non-

physiological cell types, such as iPSC or cancer cells.
• The authors make a claim regarding the transferability of the GET model to new sequencing platforms. While I see the demonstration of this transferability in the context of 10x multiome data in Figure 2, I would appreciate additional clarity regarding its applicability to other sequencing methods, such as plate-based methods or Share-seq.
• In Figure 2d, the authors illustrate the readout distribution for different types of elements, including Promoters, Peaks, Heterochromatin, and a control group. It would be helpful if the definitions of each element were clarified, and it would be even more informative to include enhancers in this comparison.
• The authors have effectively demonstrated that GET can characterize TF-TF interactions. However, in Figure 4g, the network is displayed among TF families, which can be quite broad. It would be beneficial if the researchers could provide insights into how to precisely identify the specific TFs involved within this network or offer explanations for their selection.
• In Figure 6, the authors have proposed that the interaction between Pax5 and nuclear receptors, as well as its potential disruption, could play a role in contributing to B-All. It might enhance the strength of this assertion if a Western blot analysis were employed to demonstrate how this interaction weakens in the presence of the somatic mutation G183S, and if the subsequent effect can be rescued by expressing wild-type Pax5.
• More details should be provided on model training - optimization hyperparameters, compute infrastructure used, convergence criteria etc. This will aid reproducibility.
• Quantitative metrics such as Pearson r, AUC scores, etc. must be included in some figure legends to allow better assessment of method performance.
• Adding nucleotide-level perturbations or randomizations in training data could help improve variant effect predictions.
• Future work could explore incorporating TF binding, 3D contacts, kinetics and cellular perturbations to make GET more dynamic.
• The discussion could be expanded on limitations and future directions like generalizing to more assays, non-coding variants, gene regulation kinetics, etc.

Referee #2

(Remarks to the Author)
Fu et al. report GET, a foundation model based on chromatin accessibility data that predicts transcription, regulatory grammar, and protein interactions. They demonstrate benchmarking for tasks such as predicting the results of reporter assays and apply the model to identify new regulatory regions in erythroblasts as well as transcription factor cooperative interactions. A benefit of the approach is using chromatin accessibility as the input, which yields cell context specificity compared to using genomic sequence alone. However, the authors should address the following points to bolster confidence in their approach:

Major points:

Benchmarking:

- The authors should also benchmark against HyenaDNA as that model allows for greater genomic context given the subquadratic time complexity of the convolutional method and was also shown to outperform DeepSEA, which is one of the other models benchmarked in this manuscript.

- Generally for all benchmarking, authors should include simple approaches using the same input data of chromatin accessibility (e.g. standard neural network, CatBoost, SVM, Random Forest, Logistic Regression).

- For the fine-tuning, the authors should report results for testing leave-one-out with every individual cell type (not just the one example of astrocytes). They should also test leaving out entire general classes of cell types to test the ability of the model to generalize even when it hasn't seen a very similar cell type (e.g. test leaving out all vascular cells).

- Similarly, for the leave-one-out chromosome test, authors should repeat this with each chromosome and report summary of results rather than only testing chr 11. Additionally, because as the authors note the good performance can be explained by learning motifs from the various other chromosomes since transcription factors have a total global effect across all chromosomes, the authors should also test leaving out all regions that have a particular motif and testing whether the model can accurately predict for those held-out motif regions (and should repeat for multiple different motifs to get overall summary performance).

- The authors should better characterize the computational cost compared to alternatives including Enformer and HyenaDNA. They mention the number of elements they were able to test as a rough estimate but they should expand on quantitative comparison of compute required. Additionally, they should detail why the computational cost is improved with their architecture so that it is clear to the readers.

Input data:

- The authors should include more information on the pretraining dataset. Are they only using chromatin accessibility data from references 12-14? If so, how many total cells does this encompass? Are they presenting the data to the model only as pseudobulk? Was the data from primary tissues that were normal, or were there cell lines or disease states included?

- The authors should elaborate on how they annotate the 213 cell types. Is this based on a specific ontology? Because any

number of fine or coarse cell types can be selected, the number is less informative than what they were. Are there particular tissues that were represented? Also, what was the relative proportion of different tissues / cell types? Was the data curated or balanced?

Input data encoding:

- Similarly, the authors should more clearly state how the 2M bps are being included as an effective sequence length. Are these non-overlapping as suggested by the Input data section of the Methods, or is there scanning that occurs base pair by base pair in sections of 2M? If they are non-overlapping, is there only 2M bps of context present for the central gene within the section? Are the borders assigned randomly based on where the 2M ends or are they being selected based on CTCF sites or an other informed method?

- Are the features being defined by only previously known motifs based on Jeff et al.? Can the model make predictions for transcription factors without defined motifs?

- Given the usage of features defined by motifs, the authors should elaborate on how the model encodes overlapping motifs and test whether cooperative binding can be predicted based on ground truth of ChIPseq datasets targeting co-binding factors both when the two factors bind at overlapping motifs as well as when one factor directly binds DNA but the other acts indirectly through the binding of the transcription factor complex.

- The authors should also elaborate on how the effects of regulatory variants would be encoded - would variants be encoded differently only if they had a large effect on a known motif?

Multiomic data:

- For the ATACseq-RNAseq pairing in cases with separate but paired experiments, the authors are compelled to only match data by similarity, which we know from multiome in the same single cells is not always accurate, as chromatin accessibility can change asynchronously from transcription. The authors should discuss this limitation in the main body of the paper but should also test whether the model is able to predict expression in particular regions where the chromatin accessibility and transcription are known not to synchronize as defined by multiome. There are more multiome datasets available that do not seem to be accessed by the authors that can be used as a held-out test.

Fine-tuning:

- Fig. 2a seems to imply that the platform 2 data was used during the fine-tuning. If so, the authors should repeat the experiment using only the platform 1 for fine-tuning to evaluate on a held-out platform. Otherwise, the test does not truly demonstrate ability to transfer to a different sequencing platform.

- It seems unsurprising that promoters had greater readout compared to heterochromatin, etc. and this categorization is very coarse. The authors should expand their analysis to subcategorize different types of elements (for example, different types of enhancers based on distance from gene, presence of overlapping motifs indicating potential co-binding, location with respect to TAD boundaries, presence of pioneering vs. follower transcription factor motifs, etc.).

- The authors should test whether the model can predict variation in gene expression between individuals with variable genetic sequence or whether it is only predicting the average gene expression.

Identification of CREs:

- The authors should address whether the type of base-editing data they used is based on a technique that already has different efficiency/effect in open chromatin and therefore may already be more closely correlated with the chromatin accessibility data used to train GET to begin with.

- Given that GET has seen data from erythroblasts, the authors should indicate whether the other methods (e.g. Enformer, etc.) have seen data from this cell type. The authors should also show the efficacy in a cell type not seen by GET and generally a wider set of different cell types to get a better measure of its generalizability.

- The authors should comment further on the particular cis-regulatory region prediction benchmarks where other models outperformed GET (e.g. HBG2 distal pairs) and if there is any known differences between those benchmarks compared to the others that would explain the difference in performance.

Embedding characterization:

- The authors should show the data underlying their statement that the embedding from final layers correlates with gene expression while earlier layers are more indicative of differences in regulatory grammar.

- In the UMAP visualization of embeddings, the embeddings are compressed to 2 dimensions so necessarily, the primary category distinguishing subsets of the data will dominate the visualization. It shows that the motif yields a greater distinction than the cell type but does not indicate that there is no cel type differentiation. The authors should show clustering with the full number of embedding dimensions (e.g. hierarchical clustering with heatmap) to further evaluate if the cell type does not

play as much of a role in distinguishing the embeddings. However, in the data the authors show in Fig. 4b-c, the astrocytes do appear separate from erythroblasts, which are more similar as 1 and 3, so their data does seem to indicate that cell types are distinguished.

Protein interaction networks:

- The use of the 1st layer embedding in Fig. 4 should be compared to the use of the input data directly given there is only 1 layer of attention at play at that point, as well as comparing to a later layer.

- It is unclear what the causal direction in the neighborhood graph in Fig. 4g means. Is the predicted causal direction validated in any way compared to ground truth? What does a negative interaction mean with regard to the protein binding interactions? How is this directionality incorporated into the comparison with the STRING database?

- Macro F1 should be reported rather than true positive rate alone since this can be artificially inflated by repeated positive predictions.

- The authors should perform basic experimental validation of their prediction that the G183S germline mutation disrupts interactions between PAX5 and other binding partners.

Minor Points:

- In the introduction, the authors cite models that use sequence to make predictions, but sequence is not context-specific (the same genome is present in every cell). Given that one benefit of using chromatin accessibility is that it encodes the context of each cell's state, it would be important in the introduction to provide the context to readers of other recent foundation models that use transcriptomic data as the input since that also encodes the context-specific networks in each cell state.

- In the section "GET accurately predicts gene expression in unseen cell types at experimental accuracy", there appear to be missing figure references in the second to last 2 paragraphs. For example, "Our findings showed an average R^2 of 0.53 across diverse adult cell types..." does not have a related figure reference. Similarly, data should be shown for the result that performance dropped comparing fine-tuning with random initialization to fine-tuning with using the pretrained weights.

- Also in the above section, there are three cell types referenced where the model cannot beat baseline but only one is explicitly mentioned (pancreatic acinar cell). The authors should list all three cell types here.

- The "Input data" section of the methods appears to have 2 repeated paragraphs written in different ways.

- It is unclear whether Supp Fig 3a only refers to chr 11 or not. Authors should clarify this.

- Authors should add x and y axis labels to Fig. 2e.

- In Fig. 3c, it is unclear why the particular areas were chosen to be highlighted with gray shaded regions given that there are others with high correlation with ground truth.

- In Supp Fig. 4, the authors should indicate why ABC did not achieve prediction (it's labeled "No prediction" without an explanation for this in the figure legend).

- In Supp Fig. 1, the authors should more clearly delineate which part of the bottom row is GET vs. Enformer (separate two halves more clearly with labels). They should also define abbreviations in the figure legend.

Referee #3

(Remarks to the Author)
Fu et al. report a computational model using genomic DNA sequence and cell type-specific epigenomic data to predict gene expression and regulatory element activity. Notably, this model makes predictions for cell types outside of the training set (or left out), which the authors report as a major advance over the previous state of the art. The model predictions are presented for 213 human adult and fetal cell types, as well as a collection of cell types in the lymph node. The model is interpreted to characterize regulatory element syntax at the level of TF motifs and identify TF pairs with likely physical protein interactions. The authors present some interesting results, but there is a critical lack of well-considered comparisons to previously reported models allowing the assessment of the benefit of this approach. Additionally, the manuscript suffers from consistent structural issues – e.g., a lack of the text referencing the correct figures, the figure legends matching the data in the figures, missing references to data, numbers quoted in the text do not match those in figures, misleading statistics that do not capture the data, etc. Together, this manuscript requires significant improvement to demonstrate clearly that the model provides improved and biologically meaningful insights compared to previous efforts.

Major points
1. The manuscript is missing a clear, quantitative assessment of model performance on gene expression and its distribution across single cells. The authors need to clearly show prediction performance within and outside of the training cell types, as

well a comparison to Enformer performance on the same cell types. The authors should show the performance for all left-out cell types rather than just astrocytes as an anecdote in Figure 1. Additionally, a claim that prediction matches experimental accuracy is made and must be quantitatively supported and explained (is this a comparison to technical or biological replicate measurements?). There should also be deeper discussion of where and how the model succeeds and fails since much of the treatment is superficial and, in some places, misleading. This is also true for the data presented in Figures 4 - 6.
2. Prediction of quantitative chromatin accessibility seems impressive (correlation 0.98) but is hard to evaluate without context. How does this metric compare to Enformer or other models? Also, why is performance so much better for ATAC data than other data modalities?
3. Prediction of lentiMPRA activity with random insertion averages is a creative trick and the performance compared to Enformer for promoters is promising. Why do the authors believe the performance is worse in peak vs non-peak regions? Presumably non-peak regions are not enhancers, so what type of regulatory elements are these? Also, is Pearson correlation the best metric to capture performance?
4. Prediction of enhancer-promoter pairs looks promising, but the comparison to other methods raises questions. Why do the other methods perform so poorly (basically at random level) while ATAC by itself stands out? In the Enformer and ABC papers, for example, they report better than random performance. If the authors are using a different metric that they believe is more relevant, this should be explained and the metric used by others should also be reported. The Enformer paper (Nature Methods) Fig. 2 compares to Gasperini et al. and Fulco et al, and achieves above random prediction of functional regulatory elements.
5. Enrichment of physically interacting TF pairs is interesting. To understand the result, it will be important to know the contribution of motif co-localization as a baseline. For example, are two TFs predicted to interact because their motifs tend to fall in the same regions? Or, is the model adding some other information to improve upon this baseline?

Other points
1. Please reference all figure panels
2. Please describe all work thoroughly; the logic connecting points and sections is often tenuous.
3. Missing description of statistical test in Figure 2d.
4. Incorrect figure references (“Figure 2” noted on page 4 when referring to Figure 1c)
5. Fig 4h is not referenced in the text (there were several other instances of a similar lack of reference).
6. Figure 5 legends do not match data (e refers to c, g does not refer to co-IP, which has no legend).
7. Methods section typos
o Methods section “Input data” is repeated
o “Multimer structure prediction” end of text is cut off
8. Text needs an additional proofread both for grammar and scientific accuracy.
9. How does source of training data (multi ome or individual ome impact results)?

Version 1:

Reviewer comments:

Referee #1

(Remarks to the Author)
The authors have addressed my previous concerns in a very comprehensive manner.

Referee #2

(Remarks to the Author)
The authors have performed additional experiments to address the concerns raised by reviewers. They have broadened their analyses to raise confidence in the generalizability of their approach and performed experimental validation of predicted interactions. The authors have sufficiently addressed the prior concerns.

Referee #3

(Remarks to the Author)
Our major concerns related to benchmarking GET against previous models and various analysis choices were largely addressed.
• The expansion of model performance to many “left out” cell types is comprehensive and addresses issues from the first submission.
• The comparison to Enformer is now clearer in the ATAC and lentiMPRA prediction tasks.
• The updated comparison to Enformer and ABC models for enhancer-promoter pair prediction seems to have addressed concerns about improper treatment of alternative models leading to artificially poor performance.
• Inclusion of motif co-localization provides context for TF functional pair prediction.

The revision, however, raises several minor points and issues that will be important to address:
• The TFAP2A co-IP in Figure 5 should be labeled so it is clear that TFAP2A is the IP target and negative controls (such as beta actin and another TF that is not predicted to interact with TFAP2A) should be included.
• The suggestion that SNAI1 and RELA interact due to a mutual interaction with EP300 is interesting but raises the question of specificity (as many TFs are thought to interact with EP300). The authors should provide an analysis asking whether the

TFs that they predict to interact are enriched for EP300 interaction (this could be done using ChIP-seq for the TFs and EP300 and potentially AlphaFold).
• The BioID experiment is described in very inconsistent and confusing ways in the main text and methods. The authors repeatedly refer to it as a co-immunoprecipitation experiment using anti-PAX5 and anti-HA antibodies, which is not correct. The lysates were enriched with streptavidin-linked beads (not antibody-linked beads as in an immunoprecipitation) and the enriched sample should not be labeled “IP.” Labeling those samples as “enrichment” or “streptavidin enrichment” would accurately reflect the experiment and avoid confusing readers. Also, the authors refer to the experiment as “proximity ligation assay” which is incorrect (this name refers to a very specific alternative assay using DNA-conjugated antibodies commonly implemented using the Sigma Duolink kit). There is a methods typo referring to it as “PLS” as well. The general type of assay the author’s preformed is called “proximity labeling.” Also, it appears the data using streptavidin-HRP, anti-PAX5, or anti-HA are not shown although those reagents are listed in the methods. Antibody information for anti-NR3C1 is not shared.
• We would also suggest another edit pass to improve readability
• We also find the lack of code availability during review problematic

Version 2:

Reviewer comments:

Referee #3

(Remarks to the Author)
The current revision addresses our outstanding concerns sufficently.

Referee expertise:

Referee #1: scRNA-seq/dev bio

Referee #2: transfer learning

Referee #3: systems biology

Author comments

Thank you for the opportunity to revise our manuscript entitled "GET: a foundation model of transcription across human cell types.” We appreciate the thorough review by the referees and the valuable feedback provided. We have carefully considered each of the reviewers' comments and suggestions and have made substantial revisions to improve the quality and clarity of our manuscript. Below, we provide a point-by-point response to each of the three referees’ comments.

Referees' comments

Referee #1 (Remarks to the Author):

R1[0]: Here, Fu and Mo et al introduce the General Expression Transformer (GET) – a novel model to enhance the understanding of transcription regulation across 213 diverse human cell types. The GET model is extensively trained on chromatin accessibility data, enabling it to learn the syntax of transcriptional regulation. The authors demonstrate that GET can accurately predict gene expression in both previously encountered and novel cell types, demonstrating compatibility with different sequencing platforms. A notable feature of GET is its capacity for zero-shot prediction of reporter assay readout in new cell types, distinguishing it from preceding models, delivering impressive performance in identifying crucial regulatory elements. Through its application, GET study unveils rich regulatory insights for many genes within the explored cell types, identifying both novel and known upstream regulators of fetal hemoglobin, and assembling a comprehensive interaction catalog of human transcription factors and coactivators. Furthermore, GET reveals a significant interaction between the transcription factors PAX5 and the retinoic acid receptor family, shedding light on a potential disease-driving mechanism associated with lymphoma. In summary, the versatility and precise understanding of transcription regulation demonstrated by GET significantly contributes to advancing the understanding of noncoding genetic variants. Moreover, it lays a robust foundation for the de novo design of cell-type specific regulatory frameworks, showcasing its potential as a transformative tool in the domain of transcriptional regulation.

<u>Major Strengths</u>

• The GET model architecture is well-designed and justified to capture regulatory grammar. Using self-supervised pretraining on 213 scATAC-seq datasets allows the model to learn generalizable regulatory principles.

• GET demonstrates remarkable accuracy in predicting gene expression for unseen cell types, achieving near experimental-level performance (Pearson r ~0.94). This significantly outperforms all baselines such as accessibility, mean expression, etc.

• The model demonstrates impressive adaptability across diverse assays through effective transfer learning. Predictions on ATAC-seq and zero-shot lentiMPRA experiments are compelling validations of its versatility.

• The comprehensive regulatory analyses enabled by GET are a major highlight. It effectively identifies distal enhancers, -alin, and downstream targets of key regulators like GATA1. The motif aggregation strategy is powerful.

• The GET Catalog database of predicted regulatory networks and TF interactions is an invaluable resource for the community, where the interactive website is well-designed to disseminate this knowledge.

• The study is technically strong, reflecting rigorous model development, training, evaluation and interpretation. The multi-pronged validation and biological insights demonstrate the usefulness of this foundation model.

Limitations and suggestions

R1[1]: My major concern centers on the validation of the approach to uncover novel biological insights. GET uncovers mechanism of germline mutation in disorder region of transcription factors – at this level I would expect a much deeper demonstration to uncover novel biology, benchmarking against other approaches to show that these insights are only possible with GET.

We thank the reviewer for their insightful comments. To demonstrate the capability of GET to uncover novel biological findings, we have included the following analyses and examples in the main manuscript:

1. The identification of new transcription factor-transcription factor (TF-TF) functional interactions. Given GET's proficiency in elucidating intricate regulatory mechanisms across diverse cellular contexts, it can be applied to discover unknown TF-TF functional interactions. In the main manuscript, we exemplified this application identifying different factors (motifs) and their interactions, including four subnetworks: MBD2, NR/17, NFKB/1 and ZFX. (Please see section "**GET-based causal discovery identifies potential transcription factor-transcription factor interaction"** and **Figure 4** of the main manuscript for further details.) To demonstrate the method's capability to uncover new TF-TF functional interactions, we investigated two cases in more detail:

1.1. TFAP2A and ZFX: Taking TFAP2A and ZFX as an example, we segmented both proteins into four distinct structured or disordered domains based on predicted local distance difference test (pLDDT) (**Figure 5b**) and predicted the multimer structure of all pairwise combinations between these segments. Remarkably, the originally unstructured ZFX intrinsically disordered region (IDR) (**Figure 5c**) folded into a well-defined multimeric structure when paired with TFAP2A structured domains, mainly driven by electrostatic interactions (**Figure 5d**).

To provide another line of evidence, we employed molecular dynamics simulations (**Methods:Molecular dynamics simulation**), discovering that the monomer IDR exhibited a more collapsed structure after 100 ns (**Figure 5e**) and fewer inter-chain hydrogen bonds. Moreover, the per-residue pLDDT of ZFX IDR and TFAP2A in the multimer structure correlated strongly with residue instability, as measured by root mean squared distance (RMSD), aligning with previous studies indicating AlphaFold's implicit

learning of protein folding energy functions. To validate the predicted interactions between these two proteins, we performed co-immunoprecipitation experiments. As shown in **Figure 5f**, we were able to pull down ZFX using a TFAP2A antibody.

1.2. PAX5-NR/3 interaction: In our initial investigation, we analyzed the functional interactions of the transcription factor PAX5 in B-cell precursor acute lymphoblastic leukemia (B-ALL). We uncovered new connections between PAX5 and NR/2 transcription factors such as E2F3, MZF1, and NR4A2, indicating their potential involvement in gene regulation. This interaction was further corroborated by positive affinity purification-mass spectrometry (AP-MS) data of their paralog PAX2-NR2C2.

In the revised version of the manuscript, we have included further experimental validation to characterize the PAX5 interaction with NR/3 transcription factors. Since NR4A1 and NR2C2 yielded the highest values, we selected them for further experimental validation with both wild-type PAX5 and the G183S mutant. The results of our experiment demonstrated that PAX5 is indeed part of the same regulatory complex with NR2C2 in B-ALL cell lines. This conclusion was further supported by colocalization of ChIP-seq peaks of analyses for PAX5, NR2C1 (heterodimer of NR2C2) in B-ALL cell lines, as well as previous BioID-MS analysis of PAX5 interacting proteins[1]. Our experimental results agree with the GET prediction, showing that a NR/3 transcription factor NR2C2 is interacting with PAX5 (wt/mut). NR4A1 did not show interaction with PAX5.

Please see response to R1[7] to find all details of the characterization of the PAX5-NR/3 interaction.

2. Elucidating the effect of a TF germline mutation in B-cell precursor acute lymphoblastic leukemia (B-ALL). To assess how PAX5 G183S affects the interaction with NR2C2, we quantified the interaction by normalizing to NR2C2 expression level in 3 independent replicates. Our findings suggest that both PAX5 wild-type and G183S mutant interact with NR2C2 and their interaction strength differs, potentially affecting co-regulated transcriptional programs.

To investigate the effect of PAX5 G183S mutation on transcriptional programs on NR/3 and PAX5 regulated genes, we examined 141 sporadic childhood B-ALL samples including 4 familial B-ALL samples with PAX5 G183S germline variants[2]. We first selected only CDKN2A loss cases (n=20) to match the condition with familial cases (n=4, biallelic mutations, loss of WT PAX5 and CDKN2A). We further stratified the patients into two groups to perform two different comparisons: 1) PAX5 WT (n=10) vs PAX5 Loss (n=10), as a negative control; and 2) PAX5 G183S (n=4) vs PAX5 loss of function cases (e.g.

PAX5 P80R, PAX5 Loss, n=12), for defining the G183S-specific transcriptome signature. The analyses suggest that there is a specific transcriptional program associated with the G183S mutation. To evaluate if this specific transcriptional program was relevant to PAX5-NR interaction, we performed further gene set enrichment analysis, finding an enrichment in the PAX5-loss associated differential expression genes (fold enrichment 1.22, p-value: 1.5e-3) with predicted PAX5 targets and no significant enrichment in NR/3 regulated and PAX5+NR/3 regulated genes, as expected. However, the PAX5 G183S-specific differential expression genes show significant enrichment in NR/3 and PAX5+NR/3 regulated genes (fold enrichment: 1.41, 1.35; p-value: 1.05e-30, 1.57e-18). Moreover, we found that this mutant specific differential expression gene set is also enriched in the genes activated in ProB cells (fold enrichment: 1.57, p-value: 2.2e-3), including well-known functionally relevant genes like CD19, EBF1, BACH2, and PML.

Please see response to R1[7] to find all the details of the effect of the PAX5 G183S germline mutation on the PAX5-NR/3 interaction.

[REDACTED]

[REDACTED]

Currently, GET is pretrained from data spanning 213 “healthy” cell types. However, it can also be finetuned with data specific to cancer or other disease-related cell types. As detailed in R1[3], we also evaluated the performance of GET on IDH1-wild type glioblastoma (GBM) patient samples. We finetuned GET on a single patient sample

(referred to as the “one-shot” setting) and evaluated performance against the pretrained model (“zero-shot”) on held-out patient samples in expression prediction. Comparing the zero-shot and one-shot settings, we identified genomic regions that become transcriptionally active by inspecting the network’s Jacobian with respect to its input during expression prediction. In our preliminary analysis, we found that TERT is activated when finetuning GET on one patient sample when compared to zero-shot prediction, recapitulating known findings about the TERT promoter’s role in glioma[4].

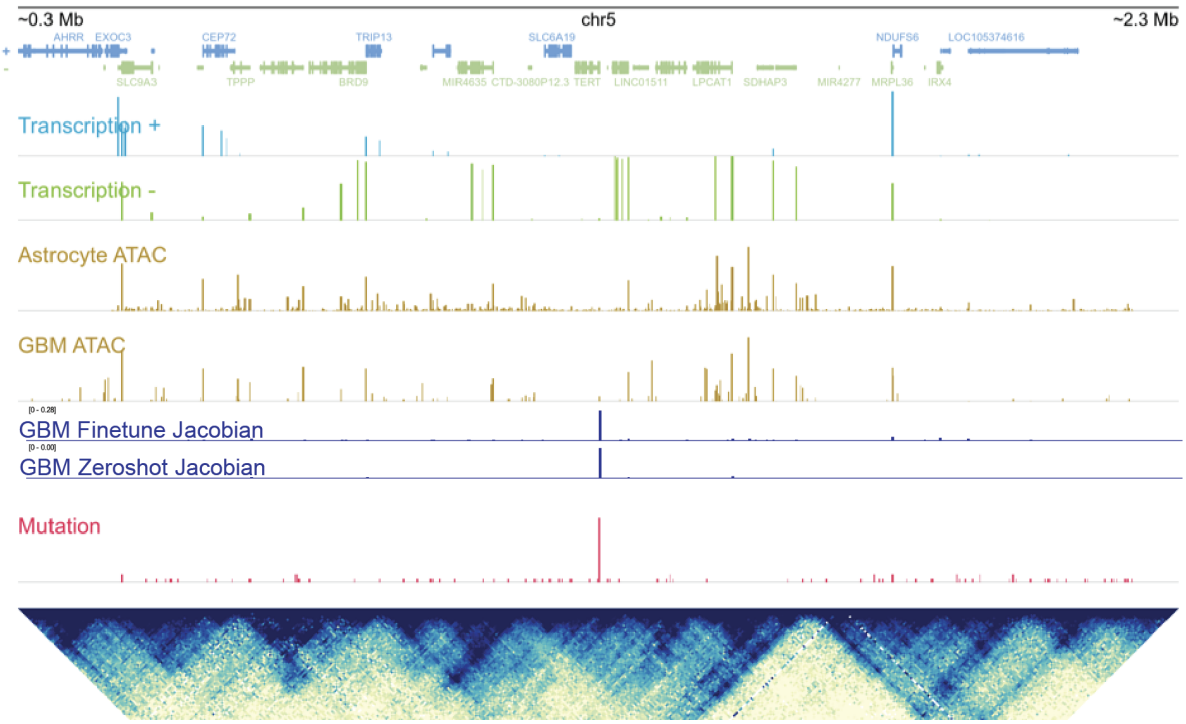


Figure. *Comparison of genomic tracks for astrocyte expression, astrocyte ATAC-seq, glioblastoma (GBM) ATAC-seq, GET’s Jacobian (finetuned), GET’s Jacobian (zero-shot), and mutation frequency at the TERT locus (chr5:1,253,167-1,295,068). Although pretrained on ATAC-seq from astrocyte cells, the model is better able to capture the glioblastoma transcriptional profile at the TERT locus after fine-tuning on a single glioblastoma patient sample. Genomic regions labeled as influential to the prediction are identified via the L2-norm of the network’s Jacobian expression matrix with respect to its input, with improved granularity of the finetuned Jacobian when compared to the zero-shot.*

We consider these examples as relevant case controls to demonstrate GET’s potential to uncover novel biological findings in glioblastoma samples and other non-physiological cell types, but leave further analysis to future work. We have included this potential application as future work in the **Discussion** section of the revised manuscript: “With our efficient

finetune framework, comparative interpretation analysis using pretrained and finetuned GET can be performed to highlight important regulatory regions or motifs driving cell state changes."

R1[2]: The authors have demonstrated their ability to accurately predict gene expression in left-out cell types, using astrocytes as an example. However, it's worth noting that astrocytes are well-characterized and highly specialized cells where transcription is rigorously regulated. It would be greatly appreciated if the authors could extend their findings to a broader range of cell types.

We have conducted additional experiments using the GET model enhanced by quantitative ATAC-seq data in a broader range of cell types. This model, referred to as the "Quantitative ATAC" model, incorporates quantitative peak counts from ATAC-seq as input features. Additionally, we employed a "Binary ATAC" model for comparison, where all non-zero ATAC peak counts are set to 1.

To test the model's generalizability, we executed three independent runs, each one with a different random seed, across various leave-out cell types. Our findings, as documented in the revised manuscript (**Supplementary Figure 2c**), demonstrate that the GET model consistently surpasses the mean expression baseline across all evaluated fetal cell types ("Training cell type mean"). Please refer to the figure below.

We also benchmark expression prediction for the same set of cell types with Enformer. Because Enformer was not trained on RNA-seq output tracks, we follow the linear probing procedure described in the original publication's benchmark against ExPecto (Extended Data Figure 4 from Enformer[5]). We fit a linear model on top of Enformer's CAGE output track predictions using RNA-seq from the particular cell type as targets.

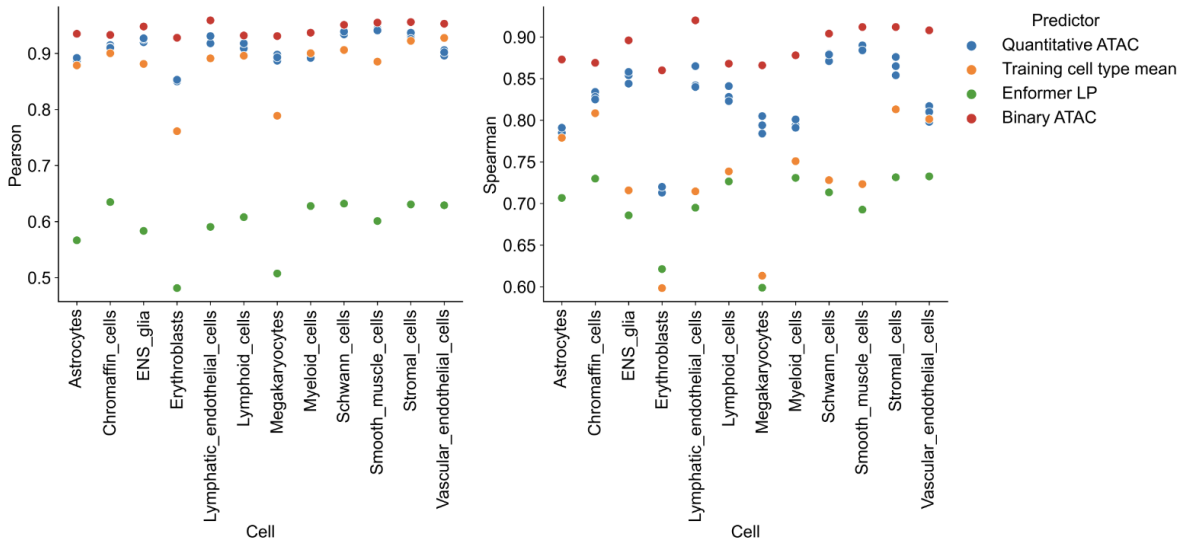


Figure. *Expression prediction for two settings of GET in which quantitative ATAC ("Quantitative ATAC") and binarized ATAC ("Binary ATAC") are included as input features. GET was finetuned on expression data from all other cell types and evaluated on the reported, held-out cell type. We report two baselines for comparison: the mean expression from the pretraining set ("Training cell type mean") and predictions from linear probing of Enformer on RNA-seq ("Enformer LP").*

We have incorporated this analysis in the **Methods** section (section: **Model evaluation: Cross-cell-type prediction**) and updated **Supplementary Figure 2c**. Additionally, in **Figure 1e** of our manuscript (see below) we show the zero-shot prediction performance in predicting gene expression across all adult cell types, which were entirely left out during the finetuning procedure on fetal data.

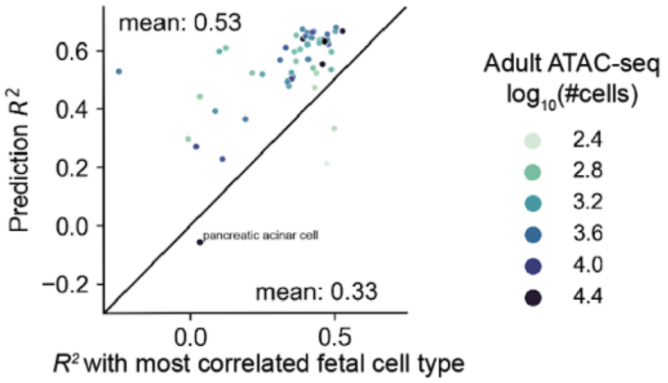


Figure. *GET trained on fetal cell types generalizes to adult cell types without retraining, outperforming the most correlated cell type baseline. x-axis showing R^2 score between GET prediction in adult cell types and observed expression in most similar fetal cell types. y-axis showing R^2 score between GET prediction and observed expression in the adult cell type.*

The reviewer can also find performance evaluation in cross-platform (SHARE-seq) perturbed embryonic stem cells and 10x Multiome glioblastoma tumor cells in R1[3] as a demonstration of transferability to non-physiological cell types.

R1[3]: I anticipate particular interest in the accuracy of prediction when dealing with cell types that exhibit more dynamic chromatin landscapes, such as stem cells. If there are variations in prediction accuracy among different cell types, it could be intriguing to investigate further what could contribute to the variation. Related to the previous point, I am curious to know whether the GET model can also predict gene expression in non-physiological cell types, such as iPSC or cancer cells.

Following the suggestion of Reviewer 1, we have conducted analyses on a variety of cell types and cell states using different datasets to probe the factors influencing prediction accuracy. Beyond K562 (a chronic myeloid leukemia cell line) mentioned in the original manuscript, we applied GET to both H1 human embryonic stem cells (hESC), and tumor cells from IDH1-wild type glioblastoma (GBM) patients. The hESC data is from the TF Atlas[6], a SHARE-seq based multiome profiling of H1 hESC after multiplexed overexpression of different transcription factor isoforms. This dataset is thus appropriate for evaluating GET's performance on non-physiological cell states with dynamic chromatin landscapes. The GBM data is profiled with 10x Multiome for 18 patients from the NCI Human Tumor Atlas Network (GEO accession number GSE240822)[7], providing an adequate test for across-patient prediction. Both datasets involve not only cell type shift but also platform shift (sci-seq to SHARE-seq or 10x Multiome).

Zero-shot evaluation of the base state hESC (subcluster #0 in the below figure) using a model trained on expression data of fetal cell types yields a Pearson correlation of 0.56. Further finetuning on the hESC data improved the Pearson correlation to 0.73 on held-out chromosomes 10 and 11. The Pearson correlation between hESC and the training cell type mean expression is 0.85.

To evaluate the leave-out cell-state performance of GET on different transcriptional states of hESC, we finetuned the model on pseudobulk data from all but one subcluster defined in the hESC 10x Multiome data. In this scenario, GET is able to reach a Pearson

correlation of 0.96, while the maximum expression correlation between unseen and seen clusters is 0.98. Using this model, when trying to predict the differential expression between unseen subcluster #6_0 and the most similar seen subcluster (#3_3, having a Pearson correlation of 0.98 with #6_0), the correlation between predicted fold change and observed fold change is only 0.18. A more dissimilar pair (#6_0 vs. #0, Pearson correlation 0.95) yields a 0.50 Pearson correlation of fold changes (see figure below). This indicates that GET is able to capture cell-state specificity after TF perturbation.

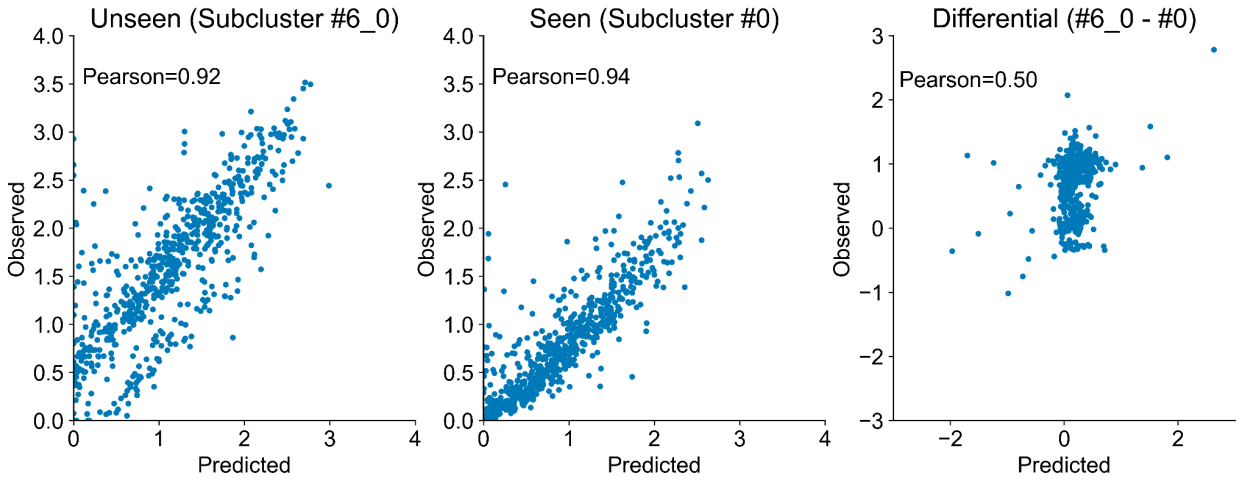


Figure. *Evaluating GET transferability to perturbed hESC SHARE-seq multiome data. Pearson correlations and scatter plots are shown between predicted and observed values for expression of one unseen cell state (#6_0) and one training cell state (#0) as well as their differential expression (log10 fold change).*

We next evaluated the performance of GET on tumor cells from IDH1-wild type glioblastoma (GBM) patients. The zero-shot performance on GBM tumor cells across 18 patients achieves a mean Pearson correlation of 0.68. We hypothesize that this higher performance in GBM cells compared to zero-shot hESC is due to their underlying similarity to astrocytes whose ATAC-seq data appeared in the pretraining set. This similarity likely makes it easier for the model to recall from pretrained embeddings, though the specific RNA-seq data for astrocytes were also not seen by the model.

We next finetuned GET on a single patient sample (referred to as the “one-shot” setting) and evaluated performance against the pretrained model (“zero-shot”) on 16 held-out patient samples. (Here we exclude two patients from evaluation to be used in the training set when evaluating robustness of the finetuning procedure; we finetune on only one patient sample at a time.) Finetuning on a single tumor patient sample shows that GET achieves a Pearson correlation of above 0.9 in the one-shot setting when predicting expression for held-out patients.

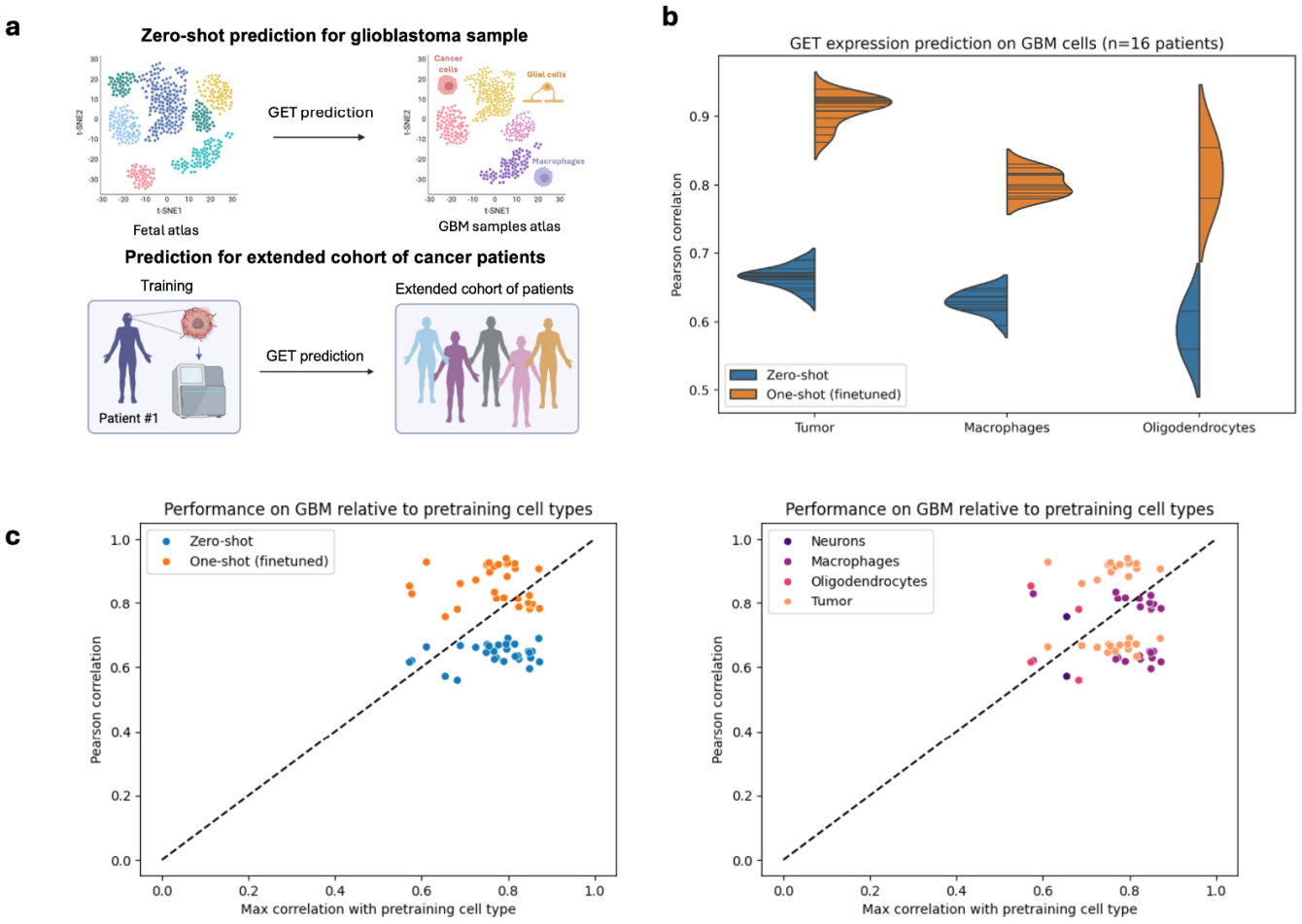
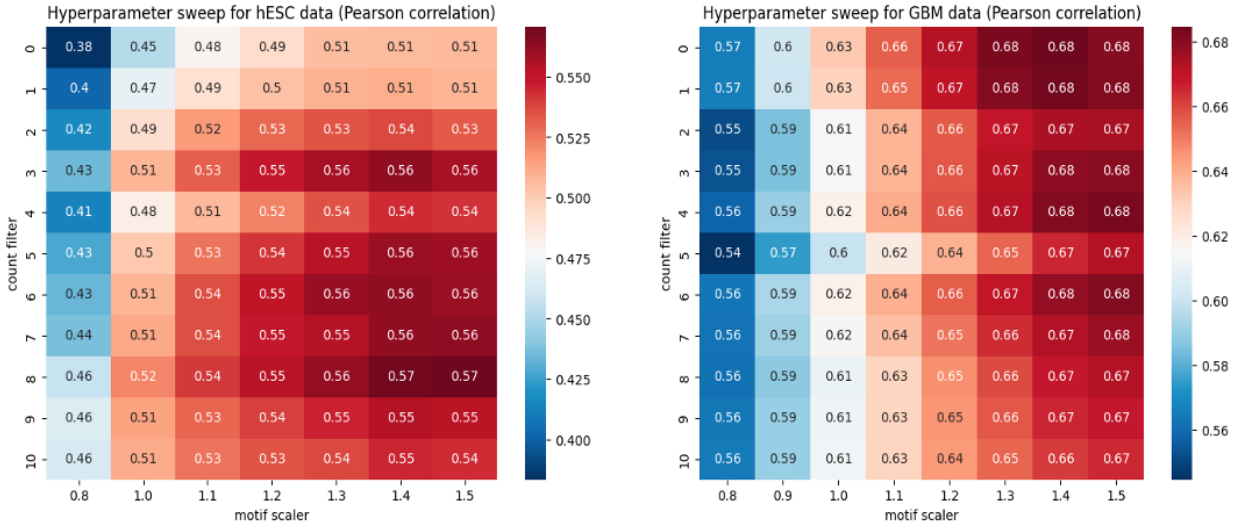


Figure. a) *GET was applied in the pretrained setting ("zero-shot") to predict gene expression from GBM patient samples. This prediction task represents both cell type and sequencing platform shift when compared to GET's original pretraining set. GET was also finetuned on a single patient sample ("one-shot") and used to predict gene expression for an extended cohort of glioblastoma patients.* ***b)*** *GET shows zero-shot and finetuned capability on a task to predict gene expression from held-out GBM patient samples.* ***c)*** *The one-shot model outperforms a baseline of predicting expression using the maximally correlated cell type in GET's pretraining set. Compared to the zero-shot setting, the one-shot finetuning improves performance across various cell types from GBM samples.*



Supplementary Figure. *To evaluate robustness of zero-shot performance in the hESC and GBM prediction settings, GET was evaluated with a sweep over two internal hyperparameters. We retain the optimal hyperparameter choices in subsequent finetuning experiments.*

In our revised manuscript, we have expanded the discussion elaborating on how different biological and technical factors could affect the predictive performance of our model across various cell types, and describe the potential of GET to be applied to non-physiological cell types, including cancer cells. Figures a and b above are now included in the main manuscript as **Figure 2d** and **Figure 2e**.

R1[4]: The authors make a claim regarding the transferability of the GET model to new sequencing platforms. While I see the demonstration of this transferability in the context of 10x multiome data in Figure 2, I would appreciate additional clarity regarding its applicability to other sequencing methods, such as plate-based methods or SHARE-seq.

In our manuscript, we showcased the applicability of GET to 10x Multiome data, as well as NEAT-seq (10x Multiome and protein quantification) and bulk ATAC-seq data on the K562 cell line (**Figure 2**). Following the reviewer's suggestion, we have incorporated additional analyses involving other sequencing methods, including zero-shot evaluation in both H1 hESC SHARE-seq data and GBM 10x Multiome data (please see R1[3] for more details).

The main obstacle when transferring our model to new data lies in making the new input space compatible with the training input space. Different cell types, sequencing technologies, and preprocessing pipelines may result in substantially different ATAC peak sets from what the model encounters during training, leading to incompatibility between the input space and embedding space under dataset shift. We ensure a compatible peak set by forming a combined peak set of new peaks and training (i.e. preexisting) peaks. Specifically, when there is an overlap between training peaks and new peaks, we use the training peak set coordinates. If a peak is unique to the new peak set, we include it as is. A uniform peak calling pipeline was used to ensure similar or equal (e.g. 400 bp in the case of the fetal-adult atlas) peak length of the training and new peak sets. Due to the comprehensive coverage of the 1.3M fetal-only and fetal-adult peak sets, new unseen peaks in most cases will only contribute to less than 10% of the total peaks. This approach leads to promising transferability to SHARE-seq data, as shown in R1[3]. As more ATAC-seq and Multiome data becomes available, more comprehensive reference peak sets like the ENCODE DHS index[8] and cPeaks[9] will facilitate adaptation of GET to new cell types.

We have updated the **Methods** section under **"Transferring to new datasets"** to clarify the applicability of GET to other sequencing methods.

R1[5]: In Figure 2d, the authors illustrate the readout distribution for different types of elements, including Promoters, Peaks, Heterochromatin, and a control group. It would be helpful if the definitions of each element were clarified, and it would be even more informative to include enhancers in this comparison.

In our manuscript, we followed the categorization of the different types of elements originally provided by Agarwal et al.[10]. "Promoters" are defined as protein-coding gene promoters, "Peaks" are defined as accessible peaks called from ENCODE K562 ATAC-seq data (which contain most of the enhancers), "Heterochromatin" are defined as genomic tiles of heterochromatin regions around selected genes (e.g. GATA1), and "Control Group" are defined as synthetic elements.

Following the suggestion in R1[5] and Reviewer 2's comments in R2[14], we have further stratified the categorization to include more nuanced information. We have now further clarified the definition of the elements by overlapping the annotated 15 ENCODE ChromHMM states computed from histone marks and other ChIP-seq data for K562. We selected the elements overlapping with states "12_EnhBiv," "6_EnhG," and "7_Enh" as enhancers; and "13_ReprPC," "14_ReprPCWk," and "15_Quies" as repressive and quiescent regions. The updated figure is shown below and is now updated in the revised

manuscript as **Figure 2g**. We have also included the corresponding element definitions in the **Methods** section under **"LentiMPRA zeroshot prediction."**

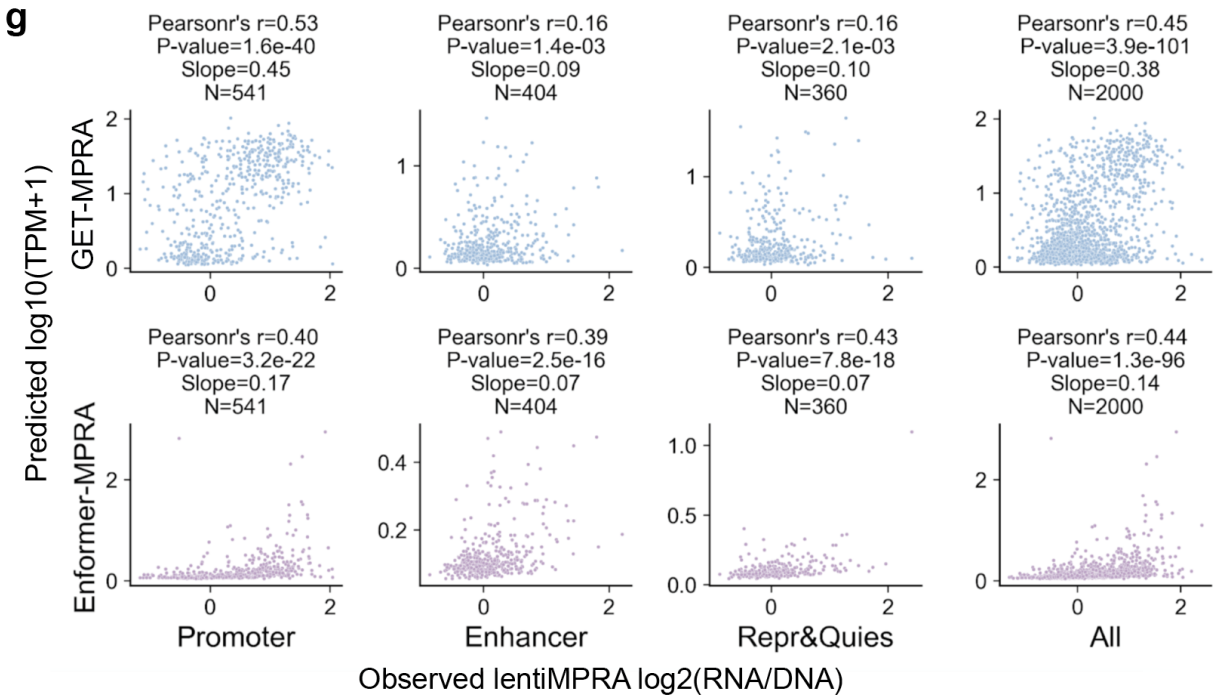


Figure. *Benchmarking GET lentiMPRA prediction against Enformer on a random subset of elements (n=2000).*

R1[6]: The authors have effectively demonstrated that GET can characterize TF-TF interactions. However, in Figure 4g, the network is displayed among TF families, which can be quite broad. It would be beneficial if the researchers could provide insights into how to precisely identify the specific TFs involved within this network or offer explanations for their selection.

As the reviewer has indicated, our approach to modeling TF-TF interactions (**Figure 4g,h**) is indeed based on motifs, which presents challenges in precisely selecting individual TFs from broader families. This led us to include additional information like TF expression and structure prediction to identify specific TFs for downstream analysis.

In our case study of PAX5, we started from GET predicted cell type specific motif-motif interaction networks (n=5 for PAX5, **Figure 6b**), and performed exhaustive AlphaFold multimer prediction between pairs of TF domains to narrow down potential structure interactions (n=1, NR/3 motif with 8 potential TF candidates, **Figure 6f**). To select the most specific TFs, a logical approach is to select the set of TFs with higher expression in

the cell type under investigation. This method is underpinned by the rationale that the expressed TF in a particular cellular context is likely to be functionally significant and thus a plausible candidate for mediating the observed regulatory interactions. This approach leads to only four candidate NR/3 motif transcription factors (NR4A1, NR2C2, RXRB, RARA; **Figure 6f**). This indicates that cell type specific TF expression helps to narrow down potential interactors among all TFs with the same motif; however, expression itself does not fully resolve this issue. Additional data like BioID-MS (BioID proximity labeling followed by mass spectrometry), although in a different cell type (mouse Pre-B cell[1]), completely narrowed down candidates to the validated TF NR2C2.

We have added a paragraph in the **Discussion** section with additional recommendations to pinpoint the functional transcription factor interaction by combining GET prediction, a structure prediction model, expression data, and other types of information like high throughput protein interactome characterization: “GET helps to identify functionally relevant motif-motif interaction networks. As in our PAX5-NR2C2 example, other types of information like TF expression in specific cell types or protein-protein interaction data could help to further narrow down specific transcription factors.”

R1[7]: In Figure 6, the authors have proposed that the interaction between Pax5 and nuclear receptors, as well as its potential disruption, could play a role in contributing to B-All. It might enhance the strength of this assertion if a Western blot analysis were employed to demonstrate how this interaction weakens in the presence of the somatic mutation G183S, and if the subsequent effect can be rescued by expressing wild-type Pax5.

We appreciate the reviewer's suggestion to perform an experimental validation of PAX5’s interaction with NR/3 transcription factors. We have tried two different approaches to characterize the interactions: Co-immunoprecipitation (Co-IP) and BioID-based proximity labeling. Our first try with Co-IP failed to detect the interaction between PAX5 and NR2C2, among other NR TFs. However, as a previous study on TF-TF interaction has recently shown, direct interaction shown by cryoEM can be missed by IP-MS, an *in vitro* pull-down assay, due to weaker interaction strength, which often happens when disordered regions are involved in the interaction[11].

To overcome the limitations of co-IP in capturing transient interactions, we performed BioID proximity labeling experiments to characterize PAX5-NR/3 interaction in B-ALL cell lines REH. To select candidate NR transcription factors, we prioritized all TFs in the NR/3 motif family based on their expression in 15 B-ALL PAX wildtype patient samples in a longitudinal study[12] (see Figure below; Diagnosis: green; Relapsed: orange).

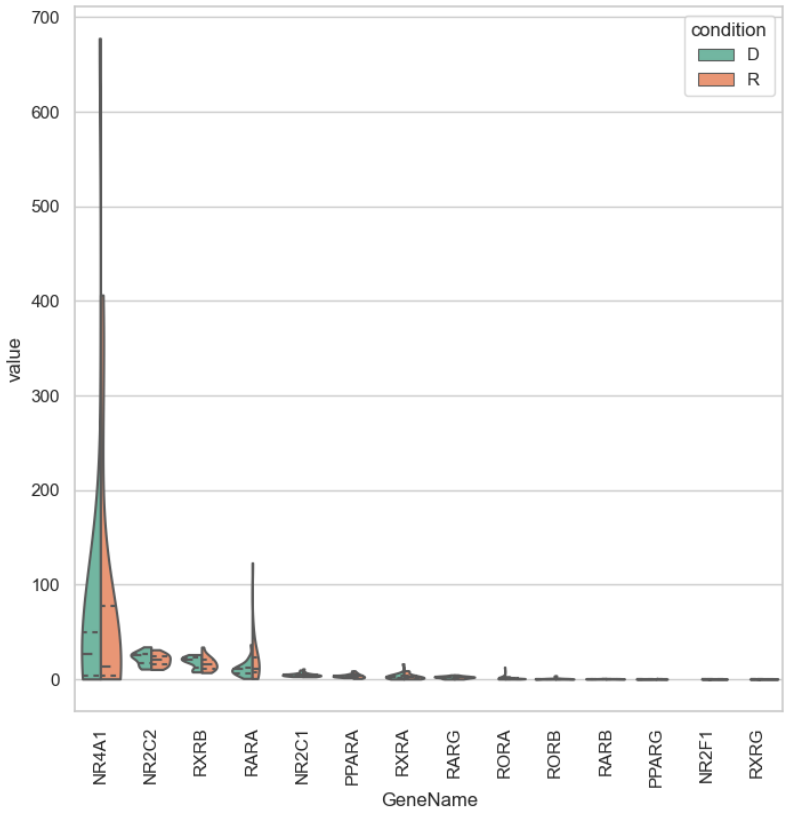


Figure. *Nuclear receptor transcription factor candidates based on expression in 15 B-ALL PAX wildtype patient samples in a longitudinal study.*

Since NR4A1 and NR2C2 yielded the highest expression values, we selected them for further experimental validation with both wild-type PAX5 and the G183S mutant. As positive controls, we also included previously reported nuclear receptor corepressor/cofactors NCOR1 and NRIP1[13], and a known interacting NR/20 TF, NR3C1, whose motif was predicted by GET to be non-interacting with PAX5. RHOA was used as a negative control.

The results of our experiment demonstrated that PAX5 is indeed part of the same regulatory complex with NR2C2 in B-ALL cell lines. This conclusion was further supported by colocalization of ChIP-seq peaks in analyses for PAX5, NR2C1 (heterodimer of NR2C2) in B-ALL cell lines, as well as a previous BioID-MS analysis of PAX5 interacting proteins[1].

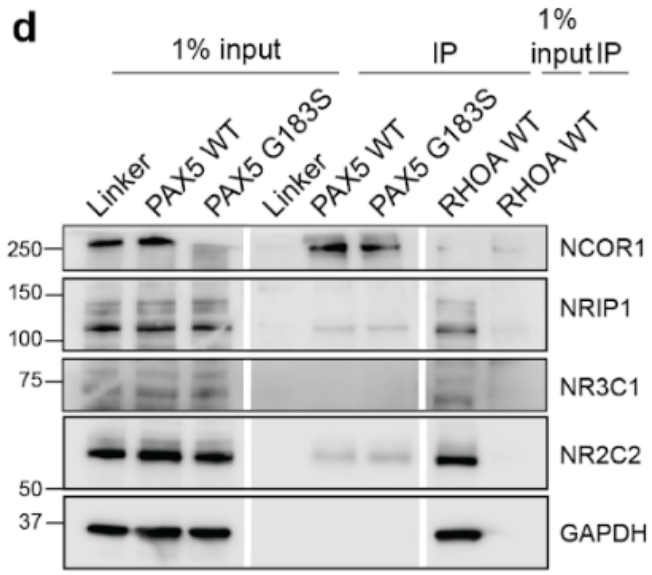


Figure. *Detection of PAX5 and RHOA BioID-fusion proteins in REH cells by western blot analysis using anti-PAX5 and anti-HA antibodies. 1% of total protein lysate was used as input.*

The experimental results agree with the GET prediction, showing that a NR/3 transcription factor NR2C2 is interacting with PAX5 (wt/mut). As a negative control, the NR/20 transcription factor (predicted to be interacting with NR/3 but not PAX5 in B cells) NR3C1 showed no interaction with PAX5 (wt/mut), although they are known to be both in the NR3C1-NR2C2-NRIP1 complex[13]. This suggests a direct PAX5-NR2C2 interaction, as predicted. NR4A1 did not show interaction with PAX5. The western blot analysis is included in the revised manuscript as **Figure 6d**.

To assess how PAX5 G183S affects the interaction with NR2C2, we quantified the interaction by normalizing to NR2C2 expression level (band intensity) in 3 independent replicates (normalizing to GAPDH leads to similar results). Our findings suggest that both PAX5 wild-type and G183S mutant interact with NR2C2 and their interaction strength

differs, potentially affecting co-regulated transcriptional programs. The following figure has been included in the main manuscript as **Figure 6e**.

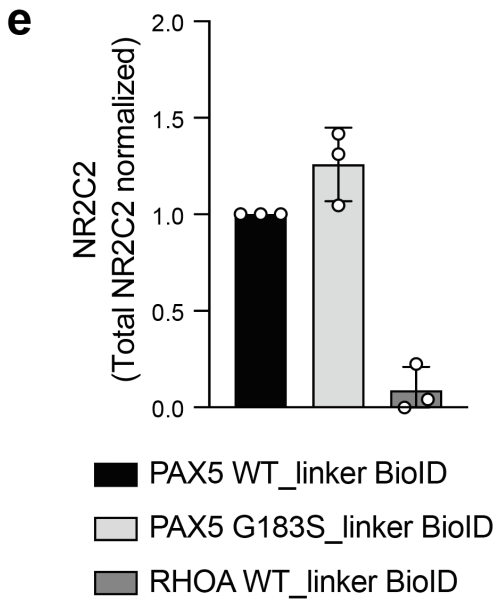


Figure. *Quantification of PAX5-NR2C2 interaction in the streptavidin immunoprecipitation in 3 replicates.*

To investigate the effect of the PAX5 G183S mutation on transcriptional programs on NR/3 and PAX5 regulated genes, we examined 141 sporadic childhood B-ALL samples including 4 familial B-ALL samples with PAX5 G183S germline variants[2]. We first selected

only CDKN2A loss cases (n=20) to match the condition with familial cases (n=4, biallelic mutations, loss of WT PAX5 and CDKN2A).

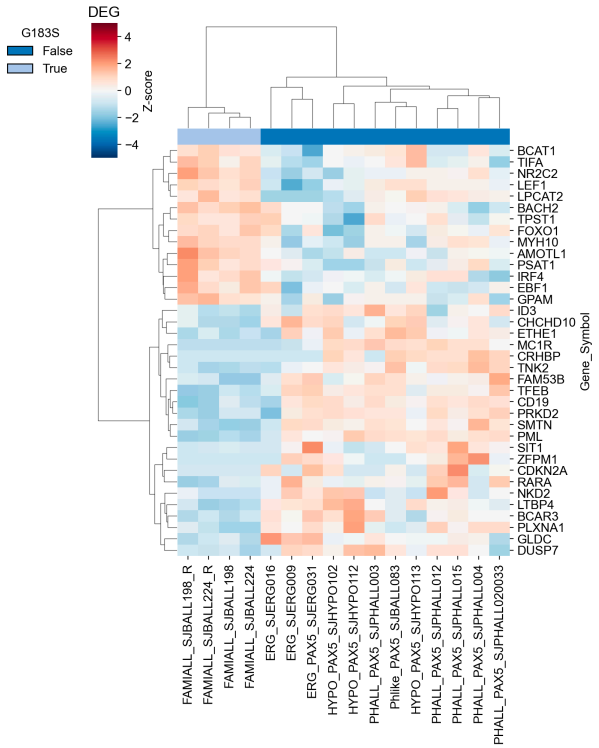


Figure. *Heatmap with the specific transcriptional program for 141 cases of sporadic childhood B-ALL with the germline G183S mutation.*

We further stratified the patients into two groups to perform two different comparisons: 1) PAX5 WT (n=10) vs PAX5 Loss (n=10), as a negative control; and 2) PAX5 G183S (n=4) vs PAX5 loss of function cases (e.g. PAX5 P80R, PAX5 Loss, n=12), for defining the G183S-specific transcriptome signature. The analyses suggest that there is a specific transcriptional program associated with the G183S mutation. This figure has been included in the revised manuscript as **Supplementary Figure 9e**.

To evaluate if this specific transcriptional program is relevant to the PAX5-NR interaction, we performed further gene set enrichment analysis. We defined the PAX5-specific, NR/3-specific, and PAX5+NR/3 commonly regulated genes with GET (**Methods**). We found an enrichment in the PAX5-loss associated differentially expressed genes (fold enrichment 1.22, p-value: 1.5e-3, hypergeometric test with Benjamini-Hochberg correction) with predicted PAX5 targets and no significant enrichment in NR/3 regulated and PAX5+NR/3 regulated genes, as expected. However, the PAX5 G183S-specific differentially expressed genes showed significant enrichment in NR/3 and PAX5+NR/3 regulated genes (fold enrichment: 1.41, 1.35; p-value: 1.05e-30, 1.57e-18, respectively). Moreover,

we found that this mutant specific differentially expressed gene set is also enriched in the genes activated in ProB cells (fold enrichment: 1.57, p-value: 2.2e-3), including well-known functionally relevant genes like CD19, EBF1, BACH2, and PML. This is now included as **Figure 6h** in the revised manuscript.

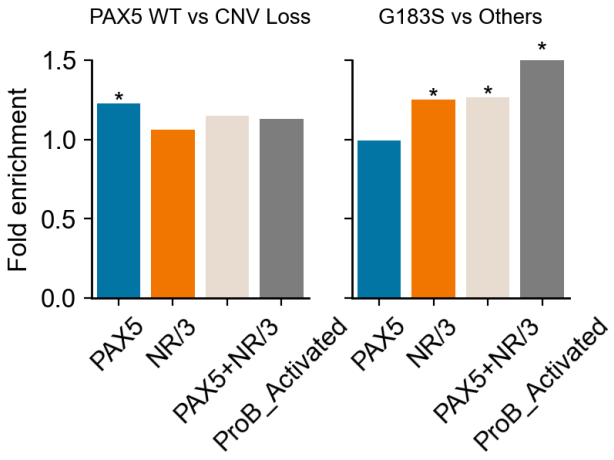


Figure. *Enrichment analysis for differentially expressed genes between PAX5 wild type v.s. PAX5 loss (left) and PAX5 G183S v.s. other PAX5 alterations (CNV loss, P80R). * indicates statistical significance from a hypergeometric test (Benjamini-Hochberg adjusted p-values are reported).*

R1[8]: More details should be provided on model training - optimization hyperparameters, compute infrastructure used, convergence criteria etc. This will aid reproducibility.

Following the reviewer's suggestion, we have now included a more detailed description of the optimization hyperparameters, compute infrastructure, and convergence criteria used in developing the model in the **Methods** section. We also include these details below.

Pretraining phase:

1. Compute Infrastructure: GET was pretrained on 16 NVIDIA V100 GPUs or 8 NVIDIA A100 GPUs for 800 epochs of training, which spanned approximately one week.
2. Learning Rate: A base learning rate of 1e-3 together with the cosine scheduler and linear annealing warmup in the first epoch.
3. Mask Ratio: 0.5.
4. Optimizer: AdamW with weight decay of 0.05.

Finetuning phase:

1. Compute Infrastructure: Similar to the pretraining phase, GET was finetuned on 8 NVIDIA A100 GPUs for 100 epochs, and completed in around one day.
2. Learning Rate: A base learning rate of 1e-3 together with the cosine scheduler and linear annealing warmup in the first epoch.
3. Optimizer: AdamW with weight decay of 0.05.
4. Early Stopping: To optimize performance and prevent overfitting, we employed early stopping based on validation loss, allowing us to select the best model checkpoints for subsequent evaluation.

This information has been added to the **Methods** section under **"Training details."**

R1[9]: Quantitative metrics such as Pearson r, AUC scores, etc. must be included in some figure legends to allow better assessment of method performance.

We have updated the figure legends to include these quantitative measures. More specifically, we have updated the legends for **Figure 3d** and **Figure 4h** to include the suggested metrics.

R1[10]: Adding nucleotide-level perturbations or randomizations in training data could help improve variant effect predictions.

We appreciate the reviewer's suggestion. This is a very interesting area of future work, as we mentioned in the **Discussion** section: "Future enhancements to GET can be envisioned through the incorporation of multiple layers of biological information, including but not limited to nucleotide-level ATAC footprints, three-dimensional chromatin architecture, and regulator expression profiles. Multiplexed nucleotide-level perturbations or randomizations will be instrumental in calibrating GET for precise predictions of the functional impact of noncoding genetic variants."

[REDACTED]

R1[11]: Future work could explore incorporating TF binding, 3D contacts, kinetics and cellular perturbations to make GET more dynamic.

We agree with the reviewer's suggestions to broaden the scope of GET to encompass transcription factor (TF) binding, 3D genomic contacts, kinetics, and cellular perturbations. The integration of these elements into GET represents a promising direction for making the method more comprehensive and reflective of dynamic cell states.

We are laying the groundwork for future work that aims to integrate these aspects into the GET model:

1. TF Binding: We plan to incorporate more direct measures of TF binding activity, leveraging ChIP-seq data and potentially integrating TF binding predictions to refine our understanding of TF-gene regulatory relationships.
2. 3D Genomic Contacts: Acknowledging the significant role of chromatin architecture in gene regulation, we intend to incorporate data on 3D genomic contacts, such as those obtained from Hi-C experiments. This will allow GET to account for the regulatory influences of enhancer-promoter interactions that occur across spatial genomic distances.
3. Kinetics: We intend to incorporate temporal changes and response to cellular signals in order to model the dynamic aspects of gene expression and TF activity. We aim to model these kinetics by incorporating time-series expression data and dynamic TF activity profiles.
4. Cellular Perturbations: Understanding how gene expression responds to various perturbations is essential to dissecting regulatory networks. We plan to integrate data from perturbation experiments, such as CRISPR knockout and

overexpression studies, to model the impact of specific regulatory elements being activated or silenced.

R1[12]: The discussion could be expanded on limitations and future directions like generalizing to more assays, non-coding variants, gene regulation kinetics, etc.

Following the reviewer's suggestion, we have extended our discussion to elaborate more about current limitations of GET. One limitation is its reliance primarily on chromatin accessibility data, which, while informative, does not encompass the full spectrum of molecular interactions and events that dictate gene expression dynamics. Future iterations of GET will aim to incorporate a broader range of assays, including those that directly measure transcription factor binding, histone modifications, and RNA polymerase activity, to provide a more holistic view of the regulatory landscape.

Elucidating the effect of non-coding variants in modulating gene expression and disease susceptibility remains an important area of exploration. Integrating genomic variants into the GET framework will enable us to predict their impact on gene regulation more accurately, offering insights into the genetic basis of complex traits and diseases. Additionally, the kinetics of gene regulation, reflecting the temporal changes in transcriptional activity in response to developmental cues or environmental stimuli, is another dimension of complexity that can be integrated into the model.

Finally, extending GET's pretraining set to include data from cellular perturbations, such as drug treatments or genetic modifications, or from cancer cells, will enhance our ability to predict the outcomes of experimental interventions. This may facilitate the design of therapeutic strategies and the identification of novel drug targets.

We have updated the **Discussion** section of the revised manuscript to include these limitations and future directions.

Referee #2 (Remarks to the Author):

R2[0]: Fu et al. report GET, a foundation model based on chromatin accessibility data that predicts transcription, regulatory grammar, and protein interactions. They demonstrate benchmarking for tasks such as predicting the results of reporter assays and apply the model to identify new regulatory regions in erythroblasts as well as transcription factor cooperative interactions. A benefit of the approach is using chromatin accessibility as the input, which yields cell context specificity compared to using genomic sequence alone. However, the authors should address the following points to bolster confidence in their approach:

Major points:

Benchmarking:

R2[1]: The authors should also benchmark against HyenaDNA as that model allows for greater genomic context given the subquadratic time complexity of the convolutional method and was also shown to outperform DeepSEA, which is one of the other models benchmarked in this manuscript.

We thank the reviewer for their insightful comments. HyenaDNA is a DNA language model trained to predict the next nucleotide from unidirectional input context[16]. In the original study, HyenaDNA was finetuned on the same task used to train DeepSEA, i.e. a 919-way multi-task prediction of chromatin profiles. However, at the moment of writing, this finetuned model has not been released by the authors.

Because HyenaDNA is not benchmarked for expression prediction in the original study, we focused on a downstream task of predicting enhancer-gene pairs in Erythroblasts[17] and K562[18]. We overlapped pair coordinates with our defined ATAC peaks for the following benchmarks. For HyenaDNA, we used the largest pretrained model available through Hugging Face (https://huggingface.co/LongSafari/hyenadna-large-1m-seqlen-hf, context length 1 million base pairs). To score enhancer-gene pairs, we performed in-silico mutagenesis by knocking down the enhancer element (i.e. setting each base pair in the enhancer region to the unknown nucleotide “N” in the vocabulary set) and comparing against wild-type likelihood of observing the promoter sequence. We also finetuned HyenaDNA on fetal erythroblast chromatin accessibility data using a multiclass classification approach (i.e. by binning targets and assigning token labels), but found that performance degraded compared to inference of the publicly released model with in-silico mutagenesis. We benchmark Enformer for the enhancer-gene prediction task, in which we used Enformer’s contribution score (gradient × input) with background normalization,

following the normalization procedure described in Gschwind et al.[18] More details about Enformer's performance are included in R3[6].

Although HyenaDNA is not trained on cell-type specific genomic data, it shows above-random performance in both the Erythroblast and K562 enhancer-gene benchmark tasks, likely due to its 1 million base pair receptive field with positional embeddings. However, GET outperforms HyenaDNA in both short- and long-range predictions, demonstrating the importance of cell-type specific pretraining.

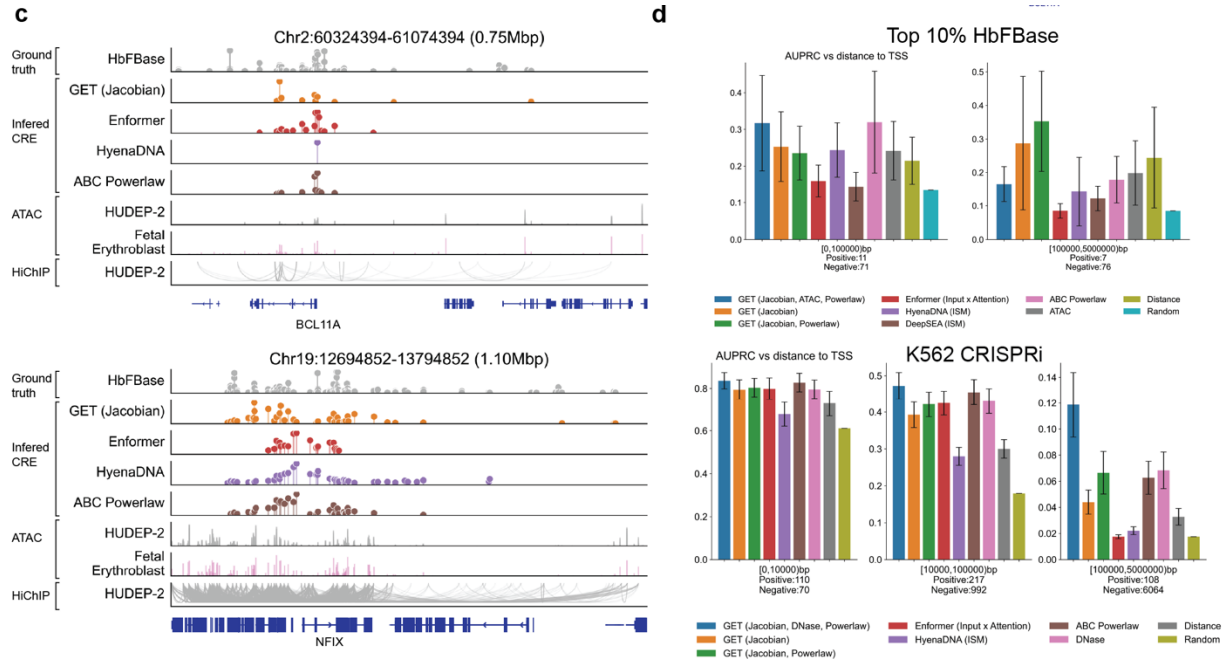


Figure. c. *Genome tracks for HbFBase enhancer-gene pair prediction, showing HbFBase ground truth, GET (Jacobian), Enformer, HyenaDNA, ABC Powerlaw, HUDEP-2 ATAC, Fetal Erythroblast ATAC, and HUDEP-2 Hi-ChIP.* ***d****. (top) We score enhancer-gene pairs using data from an ABEmax editor fetal erythroblast experiment*[17]*. A top 10% HbFBase threshold was used as the cutoff for the max HbFBase score per region. (bottom) We score enhancer-gene pairs in the K562 cell line using a recent benchmark from ENCODE*[19]*. GET shows significantly improved performance in the distal setting (i.e. distance greater than 100,000 base pairs), owing to its capacity for longer input sequences.*

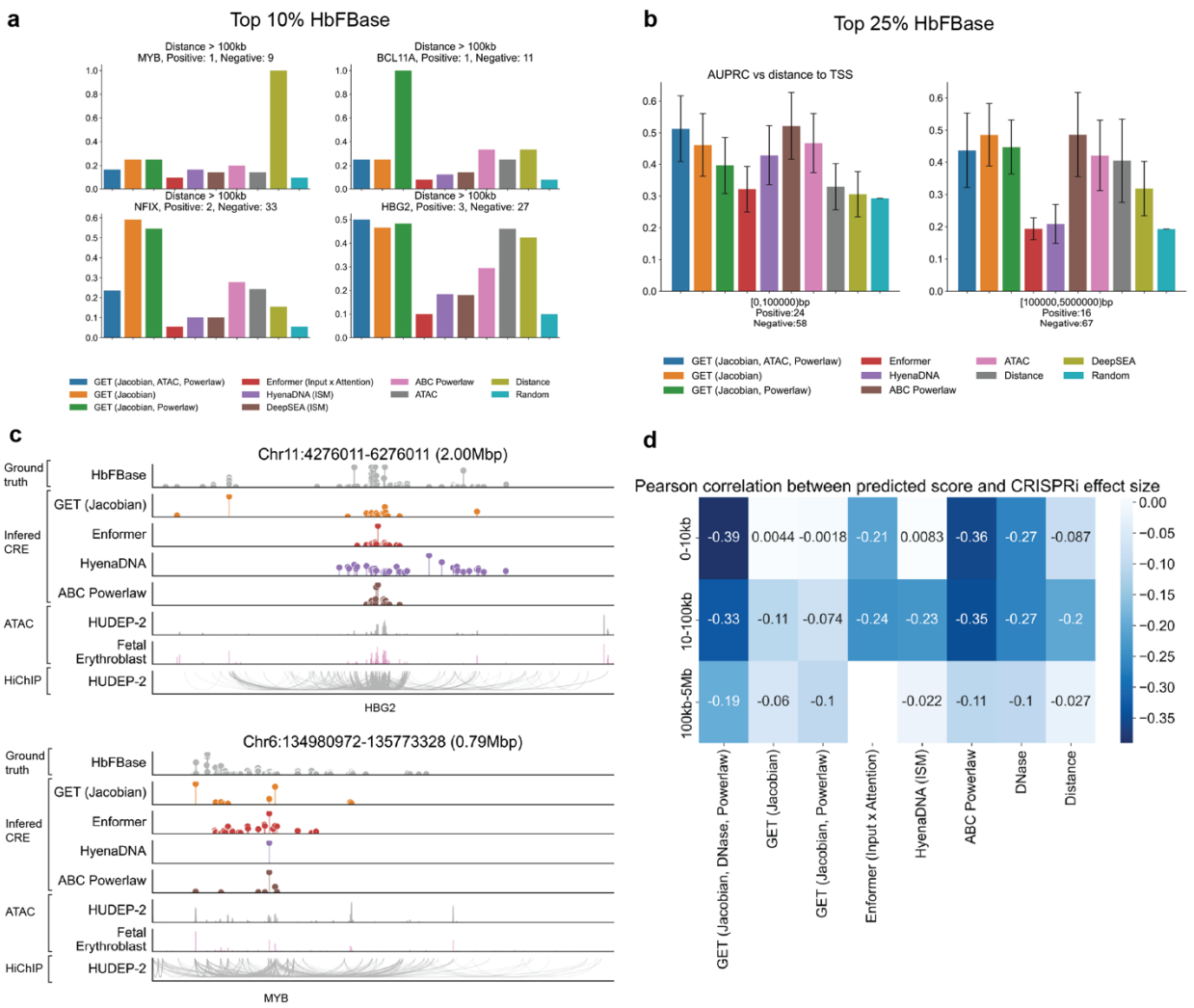


Figure a. Per-gene AUPRC benchmark of cis-regulatory region prediction with top 10% HbFbase as the label cut off. b. Distance stratified AUPRC benchmark using top 25% HbFbase as the label cut off. c. Cis-regulatory region prediction of HBG2 and MYB loci. d. Distance stratified Pearson correlation between predicted scores and K562 CRISPRi effect size.

The updated panels appear as **Figure 3d** and **Supplementary Figure 4a-d** in the revised manuscript.

R2[2]: Generally for all benchmarking, authors should include simple approaches using the same input data of chromatin accessibility (e.g. standard neural network (MLP, CNN), CatBoost, SVM, Random Forest, Logistic Regression).

In response to the reviewer's suggestion, we have expanded our benchmarking to include several conventional supervised machine learning models. We have implemented comparisons with the following methods on the task of expression prediction when leaving out chromosome 11 and leaving out astrocytes, using the same input data as GET. We provide parameters used in our implementation:

1. MLP: 3 linear layers separated by ReLU (layer dimensions: 283 input, 512, 256, 2 output). SoftPlus is used for output activation.
2. CNN: 3 Conv1d layers (layer dimensions: 283 input, 128, 64, 32, 3 kernel size) followed by FC(32, 512) → ReLU → FC(512, 2). SoftPlus is used for output activation. We used the same optimizer and parameters as used in GET (base learning rate: 1e-3, cosine scheduler, linear annealing warmup, AdamW optimizer with weight decay of 0.05).
3. CatBoost: We used CatBoostRegressor with loss function "MultiRMSE" for 1000 iterations (learning rate 1e-3).
4. SVM: We used scikit-learn Support Vector Regression (SVR) with epsilon 0.2, linear kernel, and max iterations 1000. MultiOutputRegressor was used to handle 2-dimensional output.
5. Random Forest: We used scikit-learn Random Forest Regressor with 10 estimators and max depth 10. MultiOutputRegressor was used to handle 2-dimensional output.
6. Linear Regression: We opted to use linear regression instead of logistic regression because our setting aligns better with regression rather than classification. We used scikit-learn LinearRegression and MultiOutputRegressor with default parameters.

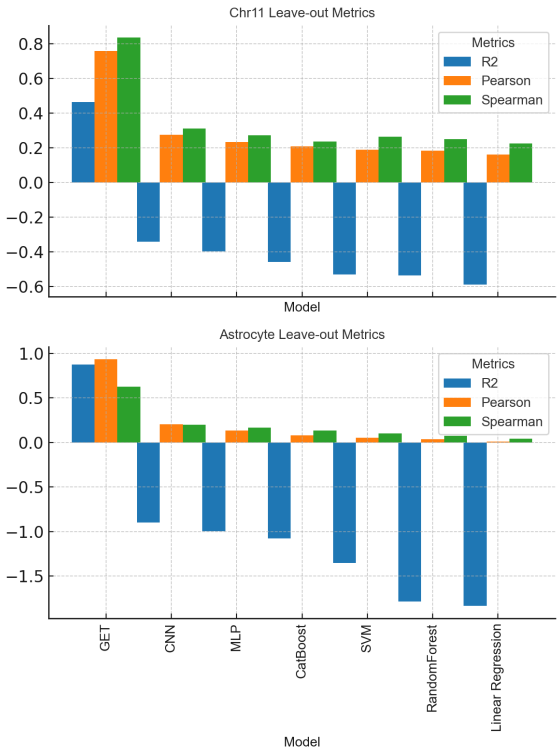


Figure. *GET benchmarked against supervised approaches for expression prediction in held-out chromosome 11 and held-out astrocytes. GET with quantitative ATAC input consistently outperforms other approaches, showing the utility of pretraining.*

For most downstream tasks, our approach diverges from the supervised methods used in these benchmarks, as we primarily employ a zero-shot interpretation approach. This strategy leverages the model's ability to generalize from the pretraining data to unseen data without the need for retraining, thus demonstrating its robustness and utility in practical applications. We have included this benchmarking figure as **Supplementary Figure 2d** in the revised manuscript.

R2[3]: For the fine-tuning, the authors should report results for testing leave-one-out with every individual cell type (not just the one example of astrocytes). They should also test leaving out entire general classes of cell types to test the ability of the model to generalize even when it hasn't seen a very similar cell type (e.g. test leaving out all vascular cells).

We appreciate the reviewer’s suggestion and note that this point was also mentioned in R1[2]. We further validated our model beyond astrocytes to include a broader range of cell types, including cell types transcriptionally different from those that appear in the

training set. For example, we left out all vascular endothelial cells for training and evaluated performance on vascular endothelial cells.

We have conducted additional experiments using the GET model enhanced by quantitative ATAC-seq data. This model ("Quantitative ATAC") incorporates quantitative peak counts from ATAC-seq as input features. Additionally, we employed a "Binary ATAC" model for comparison, where all non-zero peak counts are set to 1.

To test the model's generalizability, we executed three independent runs, each one with a different random seed, across various leave-out cell types. Our findings, as documented in the revised manuscript (**Supplementary Figure 2c**), demonstrate that the GET model consistently surpasses the mean expression baseline across all evaluated fetal cell types ("Training cell type mean"). Please refer to the figure below.

We also benchmark expression prediction for the same set of cell types with Enformer. Because Enformer was not trained on RNA-seq output tracks, we follow the linear probing procedure described in the original publication's benchmark against ExPecto (Extended Data Figure 4 from Enformer[5]). We fit a linear model on top of Enformer's CAGE output track predictions using RNA-seq from the cell type as targets.

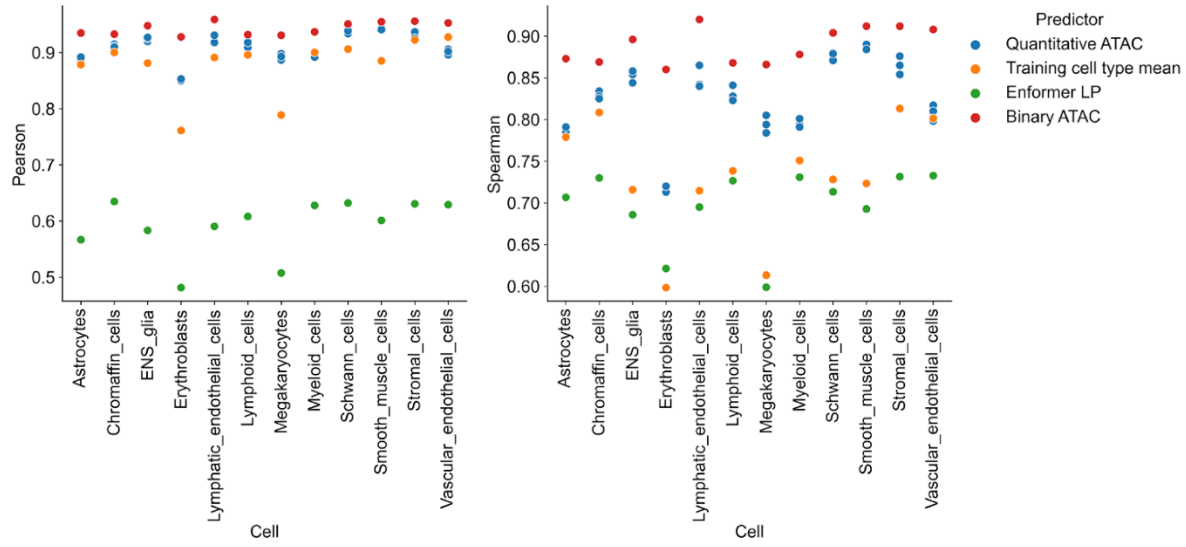


Figure. *Expression prediction for two settings of GET in which quantitative ATAC ("Quantitative ATAC") and binarized ATAC ("Binary ATAC") are included as input features. GET was finetuned on expression data from all other cell types and evaluated on the reported, held-out cell type. We report two baselines for comparison: the mean expression from the pretraining set ("Training cell type mean") and predictions from linear probing of Enformer on RNA-seq ("Enformer LP").*

We have incorporated this analysis in the **Methods** section (section: **Model evaluation: Cross-cell-type prediction**) and updated **Supplementary Figure 2c**. Additionally, in **Figure 1e** of our manuscript (see below) we show the zero-shot prediction performance in predicting gene expression across all adult cell types, which were entirely left out during the finetuning procedure on fetal data.

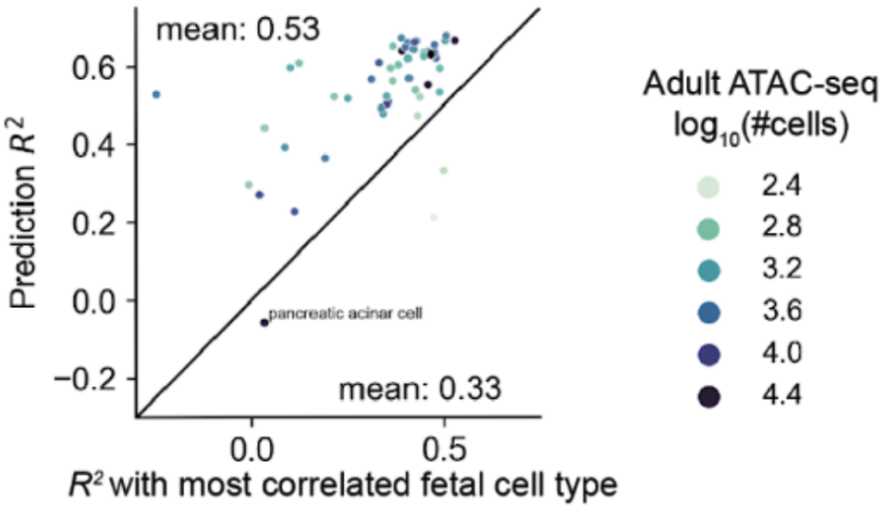


Figure. *GET trained on fetal cell types generalizes to adult cell types without retraining, outperforming the most correlated cell type baseline. x-axis showing R^2 score between GET prediction in adult cell types and observed expression in most similar fetal cell types. y-axis showing R^2 score between GET prediction and observed expression in the adult cell type.*

The reviewer can also find performance evaluation in cross-platform (SHARE-seq) perturbed embryonic stem cells and 10x Multiome glioblastoma tumor cells in R1[3] as a demonstration of transferability to non-physiological cell types.

R2[4]: Similarly, for the leave-one-out chromosome test, authors should repeat this with each chromosome and report summary of results rather than only testing chr 11.

Following the suggestion of the reviewer, we have expanded the benchmarking to include all leave-out chromosomes. We found that the performance remains consistent across chromosomes, conditioned on the same sequencing platform and data sources. We find an average Pearson correlation of 0.78 (min: 0.73, max: 0.84) on fetal astrocytes.

We also extended our evaluation of leave-out chromosomes to tumor cells from IDH1-wild type glioblastoma patients from the Human Tumor Atlas Network, as detailed in R1[3]. We performed finetuning of the base GET model on tumor cells from a single patient and evaluated performance on each leave-out chromosome. This evaluation shows an average Pearson correlation of 0.73 (min: 0.66, max: 0.80) on leave-out chromosomes.

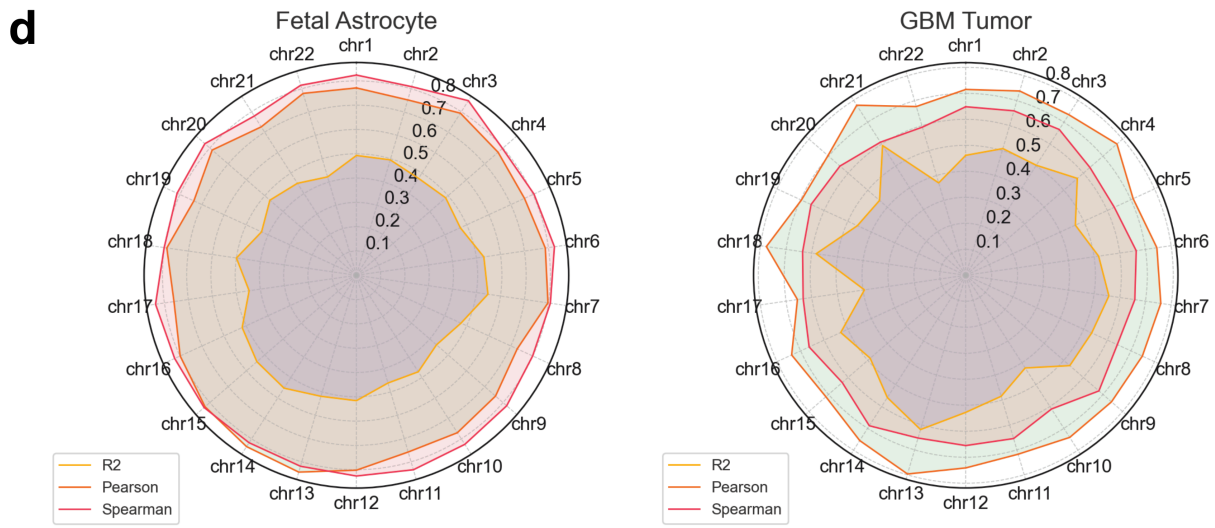


Figure. *(left) Finetuned GET benchmarked on the leave-one-out chromosome task, showing consistent performance across chromosomes for fetal astrocytes (average Pearson correlation: 0.78, min: 0.73, max: 0.84). (right) Finetuned GET benchmarked on the leave-one-out chromosome text, showing consistent performance across chromosomes for glioblastoma (GBM) samples (average Pearson correlation: 0.78, min: 0.73, max: 0.84).*

We have added these results as **Supplementary Figure 2f**.

R2[5]: Additionally, because as the authors note the good performance can be explained by learning motifs from the various other chromosomes since transcription factors have a total global effect across all chromosomes, the authors should also test leaving out all regions that have a particular motif and testing whether the model can accurately predict for those held-out motif regions (and should repeat for multiple different motifs to get overall summary performance).

We first note that we discovered an error in the original K562 ATAC analysis pipeline, in which a bug resulted in data leakage for evaluation on leave-out chromosome 11, leading to an unrealistic reported performance of 0.98. We have corrected this in the original

manuscript. After redoing this analysis, we find a Pearson correlation of 0.75 for the left-out chromosome 14.

Following the reviewer's suggestion, we have performed the held-out motif analysis. The motif encoding in our input data is not a binary token but a continuous value obtained by summing non-overlapping motif instance binding scores in a peak, which should reflect the total capacity of a peak to host a particular family of transcription factors. This motif encoding is gathered across several neighboring regions to encode a locus as an input sample. This has three implications. First, longer peaks may have larger motif scores in general, and they also tend to be more open. This aligns well with our intention to encode cell-type specific regulatory information. Second, a continuous score means that it is difficult to determine regions that have or do not have certain motifs. Third, directly omitting peaks from an input sample results in a similar effect as image corruption in deep vision models. While our pretraining target learns to "repair" the corruption, such pretext pretraining was not performed on the K562 cell line, which has a different peak distribution than cell types which appeared in the pretraining dataset, e.g. fetal astrocyte. It is thus expected that when dropping a larger number of motifs, the performance will degrade. To proceed with this task, we used K562 scATAC-seq data from ENCODE (accession: ENCFF998SLH) and called peaks with MACS2 with a threshold of q=0.05. We merged this peak set with the union peak set from the fetal pretraining data, keeping the peaks with at least 10 counts in K562. For finetuning computational efficiency, we used Low-Rank Adaptation (LoRA)[20] parameter-efficient finetuning of the binary ATAC checkpoint (pretrained checkpoint used for motif analysis in **Figure 4** and onward), pretrained on fetal and adult ATAC data with a 200-region receptive field.

We explored holding out randomly selected 1, 2, 3, 4, 10, and 20 motifs. For each motif, we checked whether a peak's binding score is larger than the top 20% scores in its score distribution across the genome. During the training stage, if a peak has any of the leave-out motifs passing this threshold, we set all input motif features of that peak as well as the observation accessibility TPM (aTPM) to zero. In this approach, these "knock-out" peaks do not contribute to the loss. During the evaluation stage, we calculated Pearson and Spearman correlation of aTPM only on these "knock-out" peaks with the original observed aTPM. For example, when there is only one leave-out motif "CTCF," we will in effect be training with about 80% of peaks that have low CTCF binding score on the training chromosomes, assuming evenly distributed binding sites across chromosomes. Similarly, when evaluating, we will be evaluating with 20% of peaks with higher CTCF binding in the test chromosomes. In these experiments, we evaluated on held-out chromosome 14.

In general, GET shows robust performance when leaving out 1 to 10 motifs. The performance degrades heavily when using 20 motifs with top 20% cutoff for each motif independently, due to removal of most of the training data.

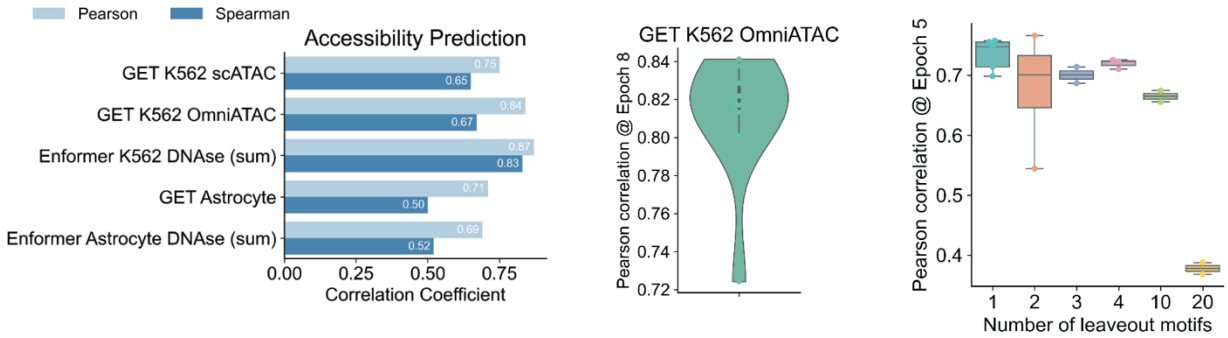


***Figure**. Performance of GET on ATAC prediction. (left) GET performance for K562 and Astrocyte ATAC compared against Enformer's predictions for DNAse on K562 and Astrocyte output tracks. (middle) GET performance for K562 bulk OmniATAC, showing leave-one-chromosome-out results for all autosomes. (right) GET performance for K562 scATAC, showing leave-out-motif analysis for randomly selected 1, 2, 3, 4, 10, or 20 motif features from the training set and evaluating on peaks with the left-out motifs (**Method: Leave-out motif evaluation**).*

We have added this analysis to a new **Supplementary Figure 2h** and updated **Methods** with a **"Leave-out-motif evaluation"** section.

R2[6]: The authors should better characterize the computational cost compared to alternatives including Enformer and HyenaDNA. They mention the number of elements they were able to test as a rough estimate, but they should expand on quantitative comparison of compute required. Additionally, they should detail why the computational cost is improved with their architecture so that it is clear to the readers.

We provide inference time, number of model parameters, maximum sequence length, and memory requirements for GET, Enformer, and HyenaDNA as a measure of computational cost. We trained or finetuned GET, Enformer, and HyenaDNA for different tasks with various dataset sizes and approaches (i.e. masked pretraining on chromatin accessibility and expression head finetuning for GET; linear probing on expression from CAGE predictions for Enformer; multiclass finetuning on chromatin profile prediction for HyenaDNA). We do not report a side-by-side comparison of training times for these disparate tasks.

	GET	Enformer	HyenaDNA
Training batch size	64 examples per batch	12 examples per batch	1 example per batch
RAM per batch (GPU)	3 GB on 1 A6000 (200 region setting) 10 GB on 1 A6000 (900 region setting)	48 GB on 1 A6000	35 GB on 1 A6000
Inference time per 1000 samples	2.6 seconds on 1 RTX 3090 (200 region setting) 13.8 seconds on 1 RTX 3090 (900 region setting)	107 seconds on 1 RTX 3090	12,000 seconds on 1 A6000†
Number of model parameters	86M	251M	6.6M
Maximum sequence length	Dependent on density of peak calling (typical sequence length: 2.7M)‡	393,216	1M

† 1 RTX 3090 with 24 GB GPU memory was sufficient for running inference for GET and Enformer. Running inference for HyenaDNA required more RAM, in which we opted to run inference on 1 A6000 with 48 GB GPU memory.

‡ GET's effective sequence length depends on the density of peak calling for a particular region, i.e. a very sparsely sampled peak set achieves a theoretical upper bound of capturing the entire genome with one input sequence. We estimate a typical ATAC-seq peak set to contain 100,000 peaks. Assuming the peaks are evenly distributed across the genome, a 200-peak input sample into the model spans 6M base pairs, while a 900-peak input sample spans 27M base pairs. In our study, we primarily use 900-peak input samples in tandem with an atlas-level joint peak set (1M peaks), such that one input sequence encodes 2.7M base pairs on average. Capturing a receptive field at this scale aligns well with the biological prior that further genomic distances are rarely found to interact in three-dimensions, as measured by Hi-C data.

When compared to Enformer and HyenaDNA, GET has a better compression ratio of regulatory sequence information due to its region-based formulation and focus on only accessible peak regions. For example, a typical ATAC-seq peak with a 1000-bp sequence is represented as a one-hot encoded 1000-by-4 matrix in Enformer and a 1000-length token sequence in HyenaDNA. In GET, the same peak is represented as a 283-length vector of motifs. The accessible genome in a typical cell type accounts for only around 10% of the total genome size, allowing for greater compression efficiency and significantly reduced input size into GET. Because the transformer-based architectures used in GET and Enformer scale as $O(n^2)$ to input sequence length, the region-based encoding used by GET significantly reduces the computation overhead when compared to an equivalent genomic sequence.

We have updated the **Methods** section with the elaborated analysis of computational cost compared to Enformer and HyenaDNA under **"Computational cost comparison**."

Input data:

R2[7]: The authors should include more information on the pretraining dataset. Are they only using chromatin accessibility data from references 12-14? If so, how many total cells does this encompass? Are they presenting the data to the model only as pseudobulk? Was the data from primary tissues that were normal, or were there cell lines or disease states included?

The authors should elaborate on how they annotate the 213 cell types. Is this based on a specific ontology? Because any number of fine or coarse cell types can be selected, the number is less informative than what they were. Are there particular tissues that were represented? Also, what was the relative proportion of different tissues / cell types? Was the data curated or balanced?

We thank the reviewer for the requested clarification and address both questions together in this response. We used ATAC-seq and expression data from references[21–23] in training. In total, the dataset encompasses 1.3 million single nuclei and is only presented to the model in pseudo-bulked format. All cell types are primary cell types from normal tissue. No disease states were included in the pretraining dataset. We incorporated additional datasets in downstream finetuning tasks, including K562, hESC, and glioblastoma tumor cells (see R1[3]).

The annotation of the 213 cell types used in pretraining followed the original cell type classification provided in the fetal and adult accessibility atlases[52,53][21–23]. This classification

was achieved through clustering of ATAC-seq count profiles, with subsequent labeling based on the expression of known marker genes. The comprehensive list of tissues and cell types, along with their annotations, can be found at the Human Cell Atlas website (http://catlas.org/humanenhancer/#!/cellType). We refer the reviewer to Zhang et al.[23] for more details on the pretraining dataset.

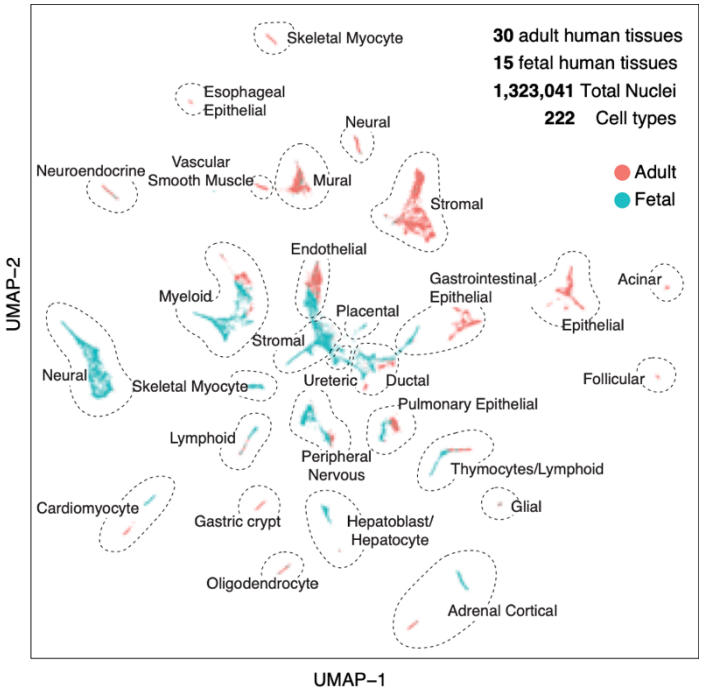


Figure. *UMAP plot of the 222 fetal and adult cell types included in the original pretraining dataset (figure from http://catlas.org/humanenhancer/#!/). GET was pretrained on 213 cell types after filtering for sequencing coverage.*

The original atlas included 222 fetal and adult cell types but was filtered to 213 cell types to remove cell types with low sequencing coverage (i.e. cell counts less than 600). This approach ensured that our model was trained across diverse cellular contexts while retaining enough coverage in the chromatin accessibility pseudo-bulk track. The data is neither further balanced nor curated.

We have updated the description of data in the **"Input data"** section of the **Methods** section with the information above.

Input data encoding.

R2[8]: Similarly, the authors should more clearly state how the 2M bps are being included as an effective sequence length. Are these non-overlapping as suggested by the Input

data section of the Methods, or is there scanning that occurs base pair by base pair in sections of 2M? If they are non-overlapping, is there only 2M bps of context present for the central gene within the section? Are the borders assigned randomly based on where the 2M ends or are they being selected based on CTCF sites or another informed method?

Our approach for encoding genomic data is peak-centric, focusing on chromatin accessibility peaks as the primary feature of interest. This method allows us to capture the most relevant regulatory regions for our analysis of transcriptional regulation.

For the analyses in our study, we employed a fixed number of 900 peaks per sample as our standard input configuration. This number was selected based on a balance between computational efficiency and the need to encompass a representative sample of the regulatory landscape for each cell type. It's important to note that the actual genomic span covered by these 900 peaks can vary depending on several factors, including cell-type-specific variations in chromatin accessibility, the threshold applied during peak calling, and the chromosomal distribution of the peaks.

In the context of the uniformly processed datasets covering fetal and adult cell types, one input sequence encodes approximately 2.7M base pairs on average. Using a joint peak set of approximately 1M over the entire genome (approximately 3 billion base pairs), the 900-peak input encodes input sequences of 2.7M base pairs on average. We used non-overlapping sampling during the pretraining and a sliding window approach during finetuning. The stride of the sliding window is set to half of the number of regions in one sample (i.e. 450 peaks per step for 900-peak samples). We note similar results when using non-overlapping windows for finetuning (i.e. stride 900).

When pretraining GET, the borders were entirely dependent on the boundary of sampled peaks; no other priors were used. Sampling was performed independently in each chromosome starting from the beginning of the chromosomes.

We have clarified these points in the updated **Methods** section.

R2[9]: Are the features being defined by only previously known motifs based on Jeff et al.? Can the model make predictions for transcription factors without defined motifs?

The reviewer is correct. We used the 282 nonredundant motif clusters defined in version 1 of Vierstra et al.'s non redundant motifs[24] (https://www.vierstra.org/resources/motif_clustering#v10-2020-4-14). GET is based on the transcription factor binding motifs that have been previously characterized; the

upstream regulator prediction and motif-interaction network are all based on the predefined motifs. If a transcription factor does not have a known motif, GET cannot be applied to directly study its regulatory relationship and motif interaction.

To extend GET to transcription factors with uncharacterized motifs, we suggest using experimental methods like ChIP-seq or SELEX to define its motif and retrain the model if the defined motifs do not align with one of the 282 motif clusters used during pretraining. Alternatively, structure prediction models like AlphaFold 3[25] can be used to screen the potential binding motifs of all human transcription factors, as both protein sequence and potential binding sequence can be supplied to AlphaFold 3 for structure prediction.

R2[10]: Given the usage of features defined by motifs, the authors should elaborate on how the model encodes overlapping motifs and test whether cooperative binding can be predicted based on ground truth of ChIPseq datasets targeting co-binding factors both when the two factors bind at overlapping motifs as well as when one factor directly binds DNA but the other acts indirectly through the binding of the transcription factor complex.

GET encodes overlapping motifs in two ways:

1. Some motifs used in our model are already composite (e.g. dimer) motifs, such as NR/13 (AGGTCA/AGGTCA) (see right image). These composite motifs inherently capture the overlapping binding sites of the constituent transcription factors.
2. Different motifs are matched independently, and all scores are summarized in the motif vector. When a sequence has two overlapping motifs, both will have a matching score. We use continuous matching scores (the log likelihood of the PWM) to retain more information for the model.

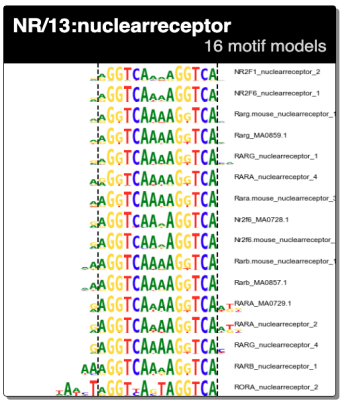


Figure. *NR/13, an example of a composite motif used in the model.*

Our current model relies on only accessibility and sequence information, without incorporating ChIP-seq data for training. As a result, our focus is on predicting functional interactions between transcription factor motifs rather than specific binding events, including cooperative binding at specific sites. We acknowledge that the co-localization of ChIP-seq signals between different factors is a valuable ground truth that reflects motif interactions and functional interactions between transcription factors. Following the suggestions of Reviewer 2 and Reviewer 3 (see R3[7]), we have incorporated ChIP-seq co-localization (using results from ChIP-Atlas[26] with data from ENCODE[27–29] and GEO[30]) as a ground truth comparison and accessible motif co-localization as a baseline.

Additionally, we have switched from precision to macro F1 score as the evaluation metric, as suggested in R2[22].

The updated results are presented in **Figure 4h** of the revised manuscript, which compares various transcription factor interaction prediction methods across different recall thresholds. The plot shows the macro F1 score of different approaches, including ground truth from affinity purification mass spectrometry in Göös et al.[31], ChIP-seq co-localization from the HepG2 cell line (ENCODE, ChIP-Atlas), motif-motif interactions derived from the GET model using the causal discovery method LiNGAM (https://github.com/cdt15/lingam), correlation analysis of the GET Jacobian matrix, and accessible motif co-localization baselines performed either across all cell types or specifically using hepatocytes corresponding to HepG2.

ChIP-seq co-localization provides a 5% recall at a 0.14 macro F1 score, slightly lower than the GET Causal approach at the same recall. The GET causal discovery method consistently outperforms the motif co-localization baselines across all specificity levels.

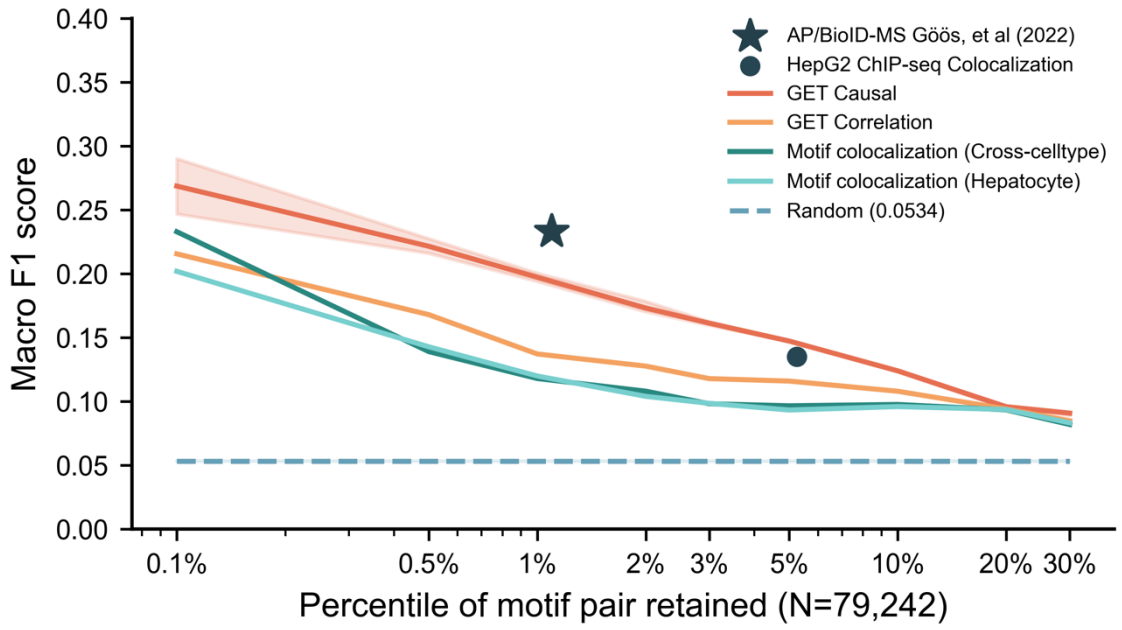


Figure. *Comparison of transcription factor interaction prediction methods across different recall thresholds. The plot shows the precision (macro F1 score) of various approaches, including ground truth from AP-MS (Göös et al., star), ChIP-seq co-localization from the HepG2 cell line (ENCODE, ChIP-Atlas, dot), motif-motif interactions from the GET model derived using five independent runs of LiNGAM ("GET Causal"), correlation analysis of GET's Jacobian matrix ("GET Correlation"), and accessible motif co-localization baselines performed either across all cell types ("Motif colocalization (Cross-celltype)") or specifically using hepatocytes ("Motif*

colocalization (Hepatocyte)"). The x-axis represents the percentile of motif pairs retained based on their interaction scores (recall), with lower percentiles indicating more stringent thresholds. The random baseline (dotted line) represents the expected performance of random guessing. ChIP-seq co-localization provides a 5% recall at 0.14 macro F1 score, slightly lower than the GET Causal approach at the same recall. GET Causal consistently outperforms the motif co-localization baselines across all specificity levels.

Supplementary Notes. The authors of ChIP-Atlas first stratify the binding peak for each TF ChIP-seq data into 3 tiers (high, mid, low; see https://github.com/inutano/chip-atlas/wiki#colocalization_doc). Then for a pair of transcription factors P1 and P2, they check colocalization between every tier pair and assign scores with a preference for high-high colocalization. If the strong binding peaks of P1 overlap with strong binding peaks of P2, the P1-P2 interaction is more robust than the case in which the P1 strong binding sites only overlap with the P2 weak binding sites. We presented the colocalization stronger than mid-mid interactions (score>=4) in **Figure 4h** as those present the more reliable interactions. A stronger cutoff (score>=9, keeping only high-high interactions) actually reduced its performance to 0.097 Macro F1 score at 2% recall.

To study the case mentioned by the reviewer in which one factor (P1) directly binds the DNA but the other (P2) acts indirectly through the binding of the transcription factor complex, one can examine the ChIP-seq for colocalized peaks between P1 and P2 and check whether any known P2 motif is present in those peaks. GET can also partially detect these when P1 and P2 are known to have different motifs: when both P1 and P2 motifs are present, the peak motif vector contains scores for both motifs; when only the P1 motif is present, the peak motif vector will only have a high match score for P1. The advantage of ChIP-seq is that it can directly pinpoint specific proteins with the same motif, assuming the availability of specific antibodies. Of course, when studying transcription regulatory complexes, ChIP-seq can also miss certain interactions. We also share an interesting observation in which we found the AP1/2-NR/11 (RXRA-NFE2L2) motif predicted by GET to be interacting across many cell types, while the colocalization signal between these two TF families in ChIP-seq is not obvious (see figure below). We note that the pair is not classified as colocalized in the ChIP-Atlas. The AP1/2 and NR/11 motifs show partial and imperfect overlapping for some TFs (e.g. RARA in NR/11; see figure below). However, the physical and functional interaction between RXRA-NFE2L2 has been shown in previous studies[32], as well as presented in the STRING database[33]. In the future, a model that integrates TF footprints and TF-DNA complex structures, and even Hi-C information, would provide better guidance to identify new transcriptional regulatory complexes.

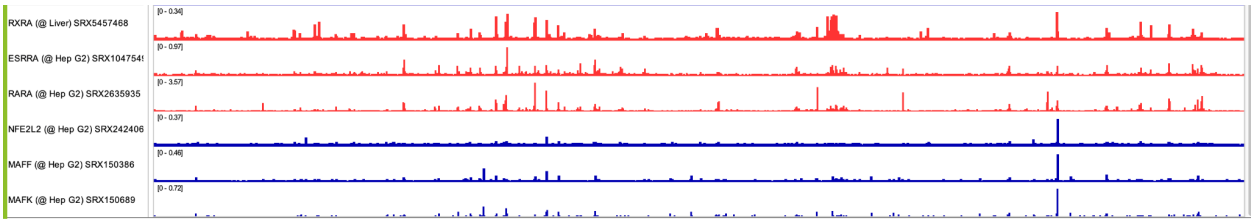


Figure. *Example ChIP-seq tracks of AP1/2 (RXRA, ESRRA, RARA) and NR/11 (NFE2L2, MAFF, MAFK) transcription factors in HepG2.*

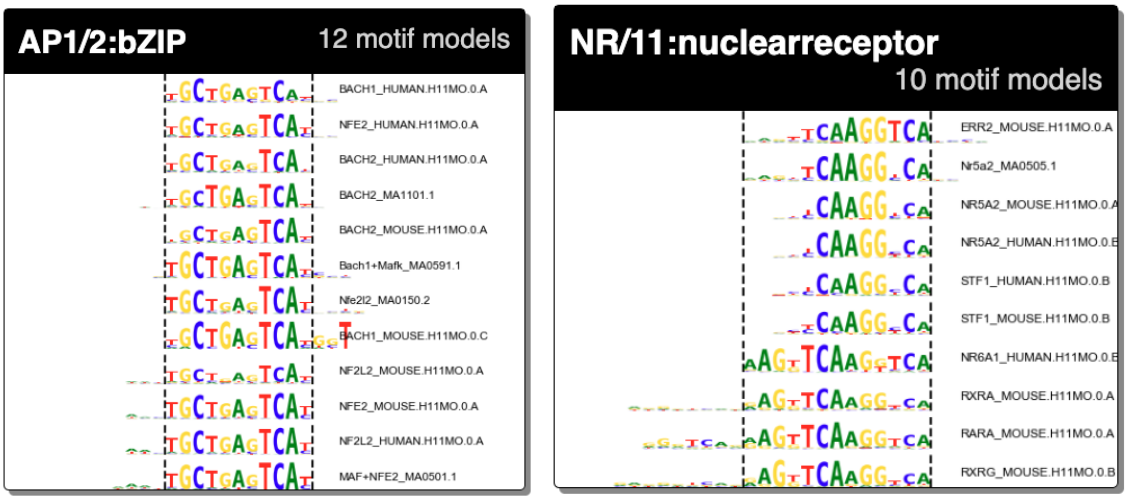


Figure. *Motif visualization of AP1/2 and NR/11.*

R2[11]: The authors should also elaborate on how the effects of regulatory variants would be encoded - would variants be encoded differently only if they had a large effect on a known motif?

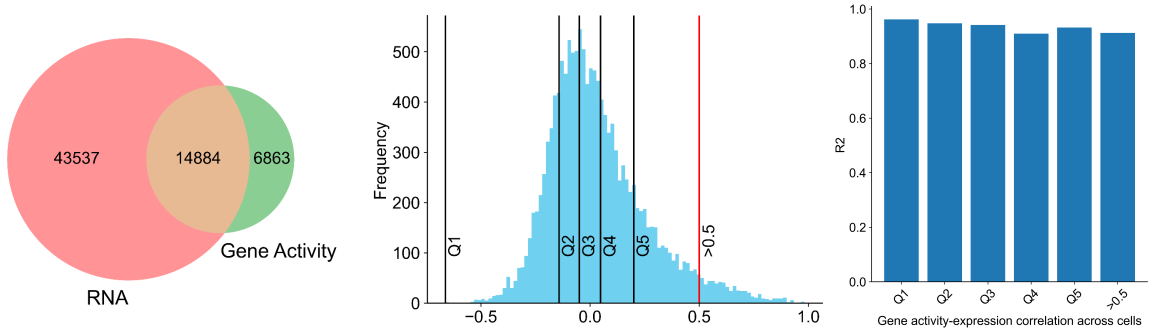
In our initial pipeline, we utilized the Vierstra et al.[24] motif cluster version 1, derived from approximately 2,000 PWMs of transcription factor binding profiles, which were categorized into 282 motif clusters. However, no consensus motif is available for each cluster. We implemented a preprocessing approach in which we used pre-scanned tracks of all motifs and consolidated the scanning results at each position. Specifically, if two motif instances in the same cluster overlapped, we retained only the maximum score within the same motif cluster. The motif scores across multiple non-overlapping motif instances are summed if they do not overlap with each other. Due to this sum-max operation, the resulting motif score vector is indeed not very sensitive to slight changes in the binding score.

As an example, we use the altered sequence to rescan for the 282 motif clusters and compute an updated motif vector for prediction. When applying this to a known MYC distal enhancer element that is approximately 2M base pairs away from its promoter, we successfully detect a complete loss of OCT motif and show that the loss of OCT motif leads to a change in MYC expression (see R1[1] Figure b).

Multiomic data:

R2[12]: For the ATACseq-RNAseq pairing in cases with separate but paired experiments, the authors are compelled to only match data by similarity, which we know from multiome in the same single cells is not always accurate, as chromatin accessibility can change asynchronously from transcription. The authors should discuss this limitation in the main body of the paper but should also test whether the model is able to predict expression in particular regions where the chromatin accessibility and transcription are known not to synchronize as defined by multiome. There are more multiome datasets available that do not seem to be accessed by the authors that can be used as a held-out test.

Following the reviewer's suggestions, we selected genes with deviating gene expression and gene activity and found the prediction performance to be similar across genes with or without synchronization. We performed this analysis using both the main GET checkpoint applied to left-out fetal astrocytes and the TF-perturbed embryonic stem cell SHARE-seq data (see R1[3] for dataset details). For fetal astrocytes, a separate brain multiome dataset from Mannens et al.[34] was used to define the gene sets with different ATAC-RNA synchronous levels. Please refer to the figures below.



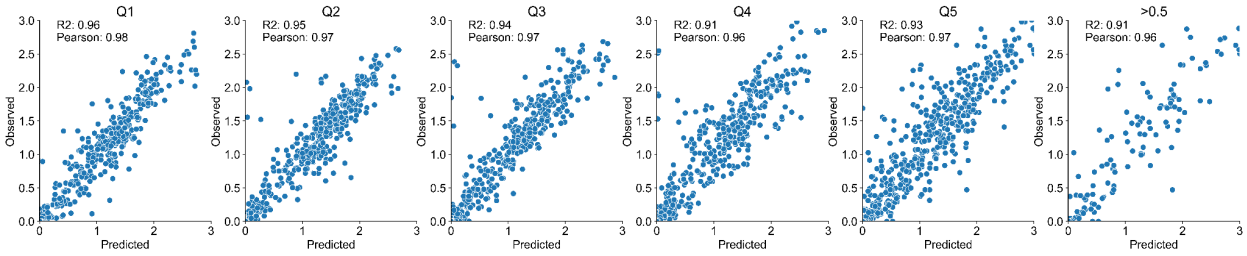


Figure. *GET prediction performance in fetal astrocytes across different ATAC-RNA synchronization gene sets. (top left) Overlap between GET-predicted genes and genes with ATAC-seq based "gene activity" calculated in the brain multiome dataset. (top right) Correlation between RNA and gene activity across pseudo-bulked subclusters, split into 6 different groups. (bottom) GET performance (predicted vs. observed) across the 6 gene groups.*

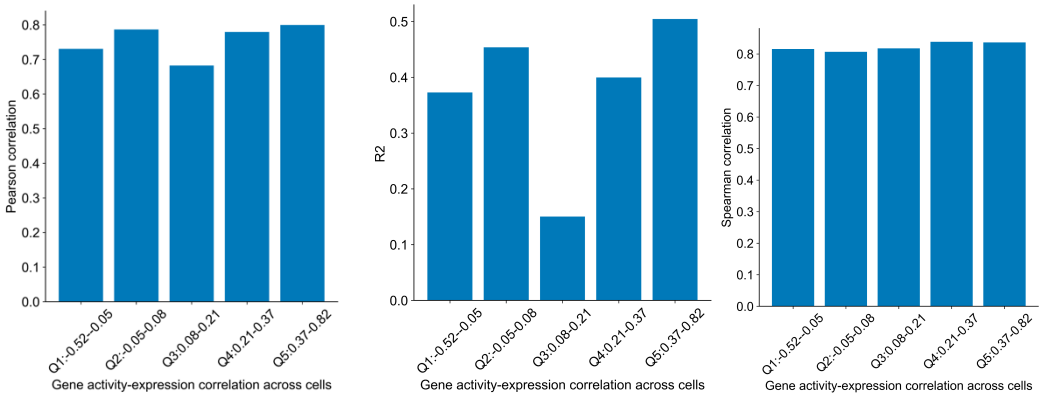


Figure: *Computed gene activity (peak counts in gene body ± 5,000 bp) across cell subclusters in the entire TF Atlas SHARE-seq multiome data. A "meta cell" with fine-grained Louvain clustering was used to alleviate sparsity and increase power of correlation detection. Correlation is stratified into five quantiles (Q1-Q5) and evaluated for genes in chromosomes 10 and 11.*

We now mention that the model is trained only on coarse-grained static cell states in the **Discussion** section of the revised manuscript.

Fine-tuning

R2[13]: Fig. 2a seems to imply that the platform 2 data was used during the fine-tuning. If so, the authors should repeat the experiment using only the platform 1 for fine-tuning to

evaluate on a held-out platform. Otherwise, the test does not truly demonstrate ability to transfer to a different sequencing platform.

To evaluate performance under sequencing platform shift, we have conducted analyses on a variety of cell types and cell states sequenced with different platforms (details for these experiments are also contained in R[3]). We applied GET to both H1 human embryonic stem cells (hESC) and tumor cells from IDH1-wild type glioblastoma (GBM) patients. The hESC data is from the TF Atlas[6], a SHARE-seq based multiome profiling of H1 hESC after multiplexed overexpression of different transcription factor isoforms. This dataset is thus appropriate for evaluating GET's performance on non-physiological cell states with dynamic chromatin landscapes. The GBM data is profiled with 10x Multiome for 18 patients from the NCI Human Tumor Atlas Network (GEO accession number GSE240822)[7], providing an adequate test for across-patient prediction. Both datasets involve not only cell type shift but also platform shift (sci-seq to SHARE-seq or 10x Multiome).

Zero-shot evaluation of the base state hESC (subcluster #0 in the below figure) using a model trained on expression data of fetal cell types yielded a Pearson correlation of 0.56, showing transferability to a new sequencing platform. Further finetuning on the hESC data improved the Pearson correlation to 0.73 on held-out chromosomes 10 and 11. The Pearson correlation between hESC and the training cell type mean expression is 0.85.

To evaluate the leave-out cell-state performance of GET on different transcriptional states of hESC, we finetuned the model on pseudobulk data from all but one subcluster defined in the hESC 10x Multiome data. In this scenario, GET can reach a Pearson correlation of 0.96, while the maximum expression correlation between unseen and seen clusters is 0.98. Using this model, when trying to predict the differential expression between unseen subcluster #6_0 and the most similar seen subcluster (#3_3, having a Pearson correlation of 0.98 with #6_0), the correlation between predicted fold change and observed fold change is only 0.18. A more dissimilar pair (#6_0 vs. #0, Pearson correlation 0.95) yields a 0.50 Pearson correlation of fold changes (see figure below). This indicates that GET is able to capture cell-state specificity after TF perturbation.

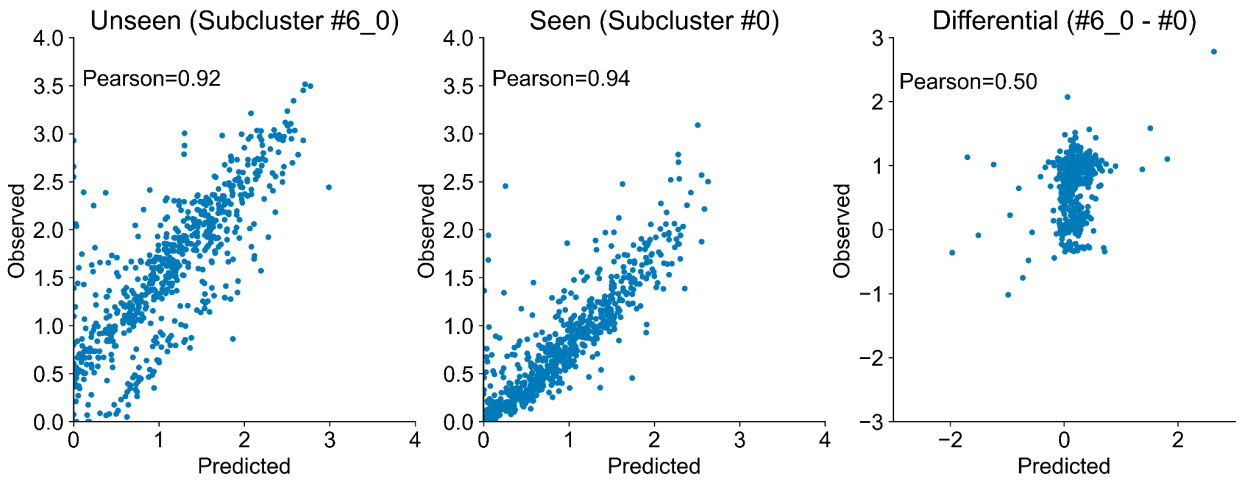


Figure. *Evaluating GET transferability to perturbed hESC SHARE-seq multiome data. Pearson correlations and scatter plots are shown between predicted and observed values for expression of one unseen cell state (#6_0) and one training cell state (#0) as well as their differential expression (log10 fold change).*

We next evaluated the performance of GET on tumor cells from IDH1-wild type glioblastoma (GBM) patients. The zero-shot performance on GBM tumor cells across 18 patients achieves a mean Pearson correlation of 0.68. We hypothesize that this higher performance in GBM cells compared to zero-shot hESC is due to their underlying similarity to astrocytes whose ATAC-seq data appeared in the pretraining set. This similarity likely makes it easier for the model to recall from pretrained embeddings, though the specific RNA-seq data for astrocytes were also not seen by the model. We note here that the model performs well under sequencing platform shift in the zero-shot setting.

We next finetuned GET on a single patient sample (referred to as the “one-shot” setting) and evaluated performance against the pretrained model on 16 held-out patient samples. (Here we exclude two patients from evaluation to be used in the training set when evaluating robustness of the finetuning procedure; we finetune on only one patient sample at a time.) Finetuning on a single tumor patient sample shows that GET achieves a Pearson correlation of above 0.9 in the one-shot setting when predicting expression for held-out patients.

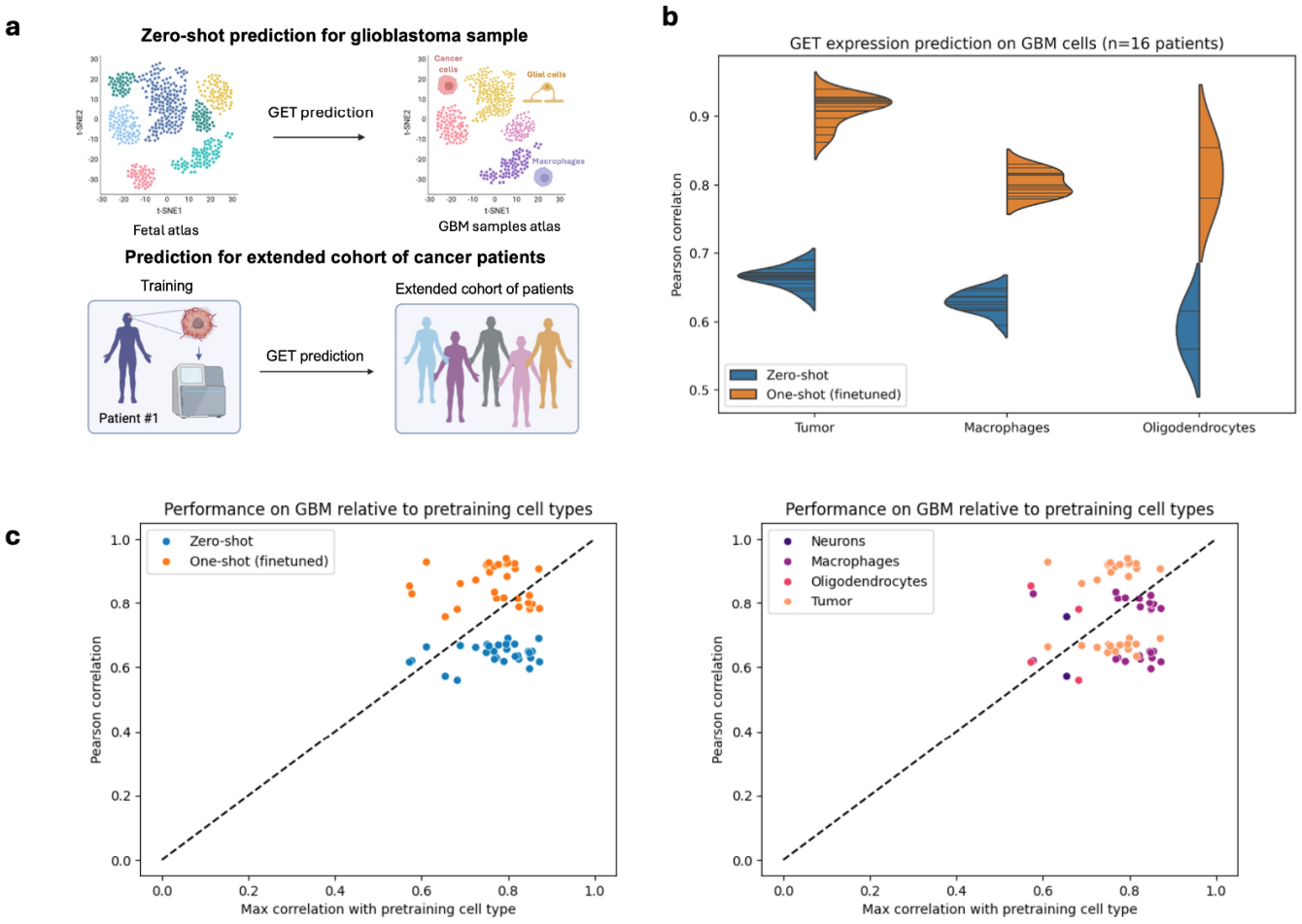
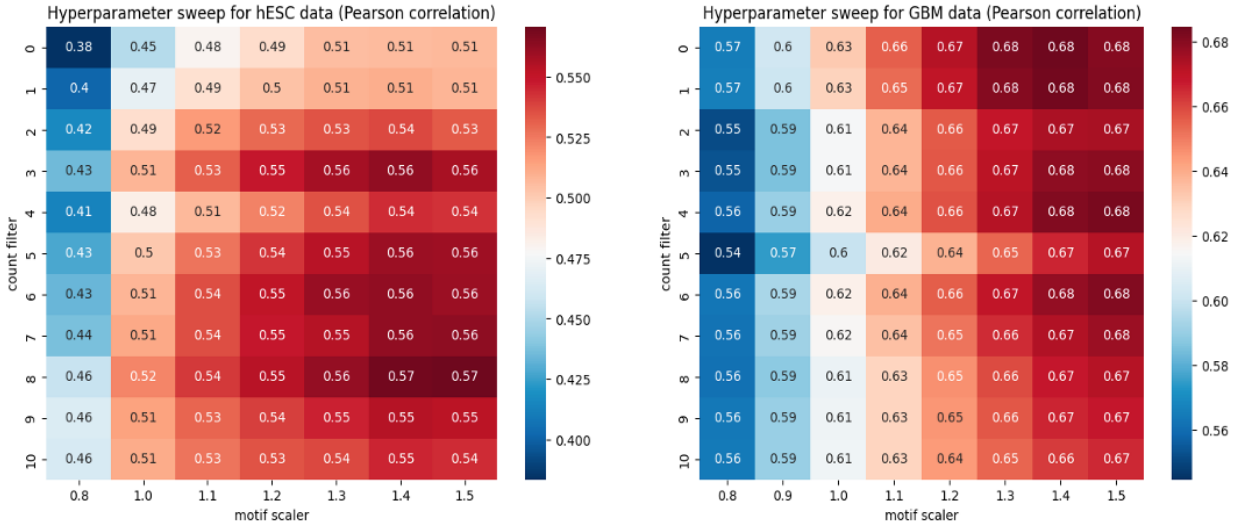


Figure. a) *GET was applied in the pretrained setting ("zero-shot") to predict gene expression from GBM patient samples. This prediction task represents both cell type and sequencing platform shift when compared to GET's original pretraining set. GET was also finetuned on a single patient sample ("one-shot") and used to predict gene expression for an extended cohort of glioblastoma patients.* ***b)*** *GET shows zero-shot and finetuned capability on a task to predict gene expression from held-out GBM patient samples.* ***c)*** *The one-shot model outperforms a baseline of predicting expression using the maximally correlated cell type in GET's pretraining set. Compared to the zero-shot setting, the one-shot finetuning improves performance across various cell types from GBM samples.*



Supplementary Figure. *To evaluate robustness of zero-shot performance in the hESC and GBM prediction settings, GET was evaluated with a sweep over two internal hyperparameters. We retain the optimal hyperparameter choices in subsequent finetuning experiments.*

In our revised manuscript, we have expanded the discussion elaborating on how different biological and technical factors could affect the predictive performance of GET across various cell types and sequencing platforms, and describe the potential of GET to be applied to non-physiological cell types, including cancer cells. Figures a and b above are now included in the main manuscript as **Figure 2d** and **Figure 2e**.

R2[14]: It seems unsurprising that promoters had greater readout compared to heterochromatin, etc. and this categorization is very coarse. The authors should expand their analysis to subcategorize different types of elements (for example, different types of enhancers based on distance from gene, presence of overlapping motifs indicating potential co-binding, location with respect to TAD boundaries, presence of pioneering vs. follower transcription factor motifs, etc.).

In our manuscript, we previously followed the categorization in the data provided by the LentiMPRA paper[35] in the ENCODE data portal, where “Promoters” are protein coding gene promoters, “Peaks” are accessible peaks called from ENCODE K562 ATAC-seq data which contains most of the enhancers, “Heterochromatin” are genomic tiles of

heterochromatin regions around selected genes (e.g. GATA1), and "Control" are synthetic elements defined in the original publication[35].

Following reviewer 2's comments, we have further stratified the categorization to include more nuanced information. Accordingly, we have updated **Figure 2f** and **Figure 2g** as shown below. We have also included the corresponding element definitions in the **Methods** section **"LentiMPRA zeroshot prediction."**

Although lentiMPRA measures in-chromatin readout, the readout is averaged from multiplexed random insertion by the lentivirus across the entire genome, and the exact location of these insertions is not measured in the data. Although it's not possible to annotate these elements based on distance from gene or TAD boundaries, it is possible to annotate them based on their original chromatin states and transcription factor binding features. Using all K562 (untreated) ChIP-seq peaks from the ChIP-Atlas, including 32 targeting histone marks and 562 targeting TFs (more than 500) and other targets (e.g. G-quadruplex, AGO1/2, Cas9), we performed enrichment analysis against the lentiMPRA promoter (approximately 57,000) and ATAC peak (approximately 169,000) library, which was further subdivided into 4 groups respectively based on GET prediction (P) (P+: >1; P-: <1) and observed readout (O) (O+: > 0.5; O-: <0.5, visualized in the figure below).

The enrichment Fisher's exact test was performed using the bedtools fisher command. The overlapping ChIP-seq peaks for the same target were merged before testing. Log2 fold enrichment for significant results ($p < 0.05$ in any group) were visualized and compared using a heatmap for all histone marks and the 100 most variable TFs (in terms of enrichment across the four expression-based groups). We noticed that O+ promoters showed larger H3K4me3 enrichment, and P+ promoters showed larger H3K4me3 and H3K27ac enrichment. Interestingly, P+O- promoters have a distinct signature of H3K9me3, H2AK119Ub, and H3K79me2 compared to the P+O+ group, potentially relevant to promoters that were originally regulated by the Polycomb complex or in heterochromatin. Since lentiMPRA elements are relatively newly inserted to the genome, these repressive mechanisms may have not yet been set up, leading to a discrepancy between RNA-seq based prediction and lentiMPRA readout. Similarly, accessible peaks are enriched in H3K27ac and H3K4me1/2/3 in general, while P+O+ peaks are particularly enriched in H3K122ac, a signature of active enhancers and transcription[36]. Comparing P-O+ peaks to P+O+ peaks, H3K27me2, H3K14cr, and H4K20me1 are more enriched. Pioneer factors like GABPA, NFYA, and NEUROD1 are enriched in P+ and particularly P+O+ promoter and enhancers. Other signature transcription factors are CEBPZ (P+ vs. P- promoters and P-O+ vs. P-O- peaks) and WDR5 (P+O+ vs. P-O+ peaks).

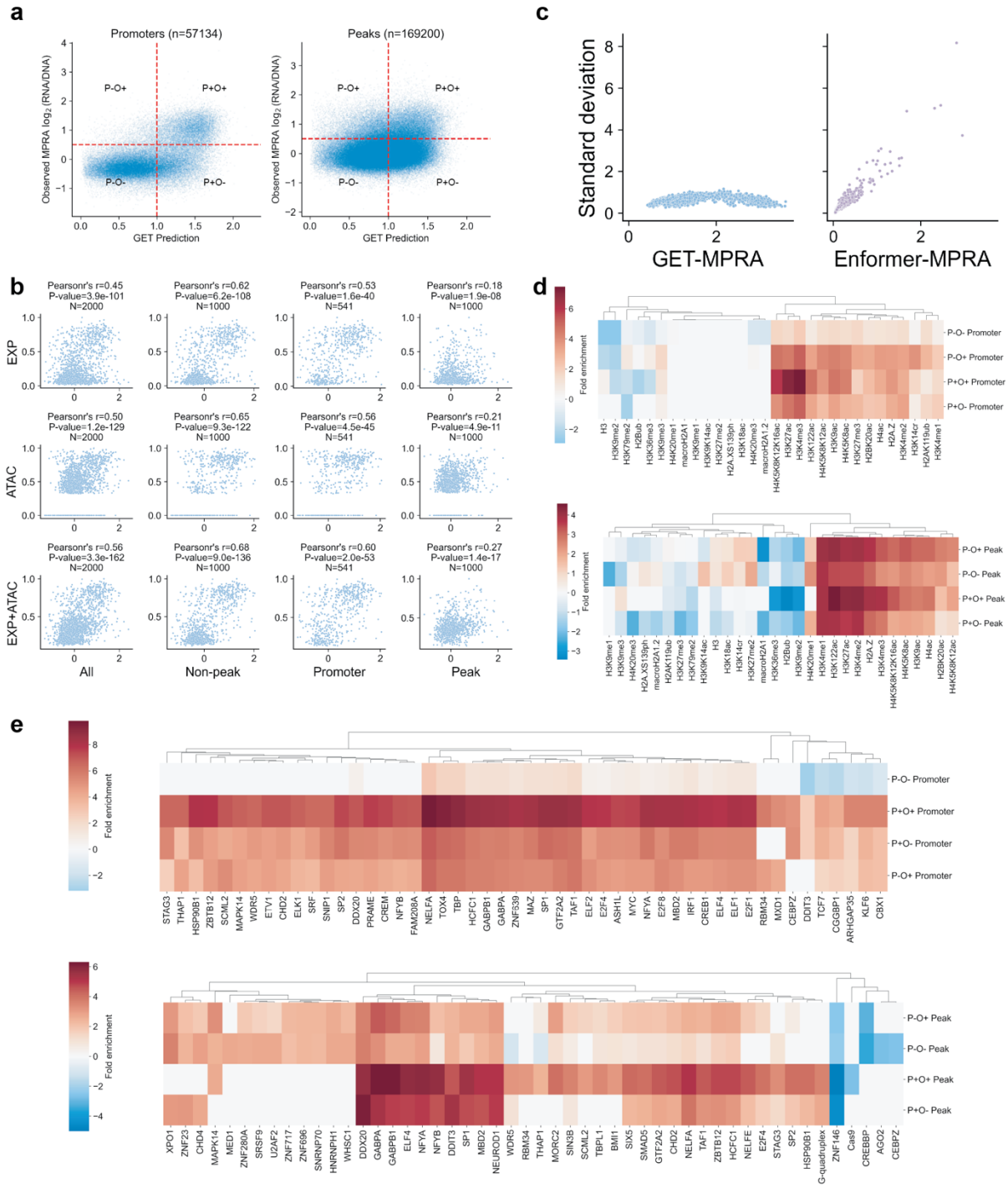


Figure. a) *Scatterplot of lentiMPRA readout and GET-MPRA prediction. Promoter (left) and ATAC peak (right) elements are classified into four sub-categories, respectively, based on high (+) or low (-) in Prediction (cutoff = 1) and Observation (cutoff = 0.5).* ***b)*** *Detailed ablation of expression and ATAC components of the GET prediction.* ***c)*** *Benchmarking standard deviation of MPRA zero-shot prediction across random*

insertions. ***d)*** *Histone mark enrichment analysis of promoter (top) and peak (bottom) elements, respectively, using ENCODE K562 ChIP-seq data.* ***e)*** *Transcription factor binding site enrichment analysis of promoter (top) and peak (bottom) elements respectively using ENCODE K562 ChIP-seq data. A Fisher exact test was performed, where tests with p < 0.05 are shown. Color shows log10 fold enrichment. For transcription factors, the variance of fold enrichment across four groups was calculated and the top 50 TFs are visualized.*

We have now further clarified the definition of the elements by overlapping the annotated 15 ENCODE ChromHMM states computed from histone marks and other ChIP-seq data for K562. We selected the elements overlapping with states "12_EnhBiv," "6_EnhG," and "7_Enh" as enhancers; and "13_ReprPC," "14_ReprPCWk," and "15_Quies" as repressive and quiescent regions. The updated figure is shown below and is now updated in the revised manuscript as **Figure 2g**. We have also included the corresponding element definitions in the **Methods** section under **"LentiMPRA zeroshot prediction."**

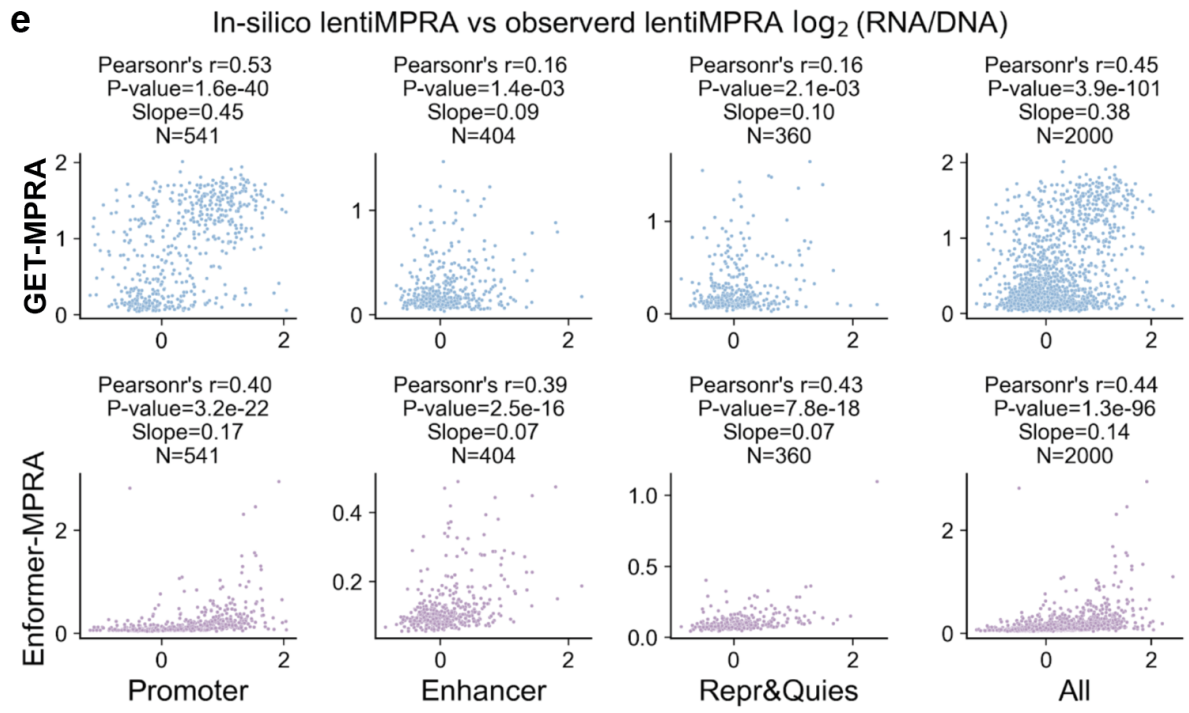


Figure. *Benchmarking GET lentiMPRA prediction against Enformer on a random subset of elements (n=2000).*

R2[15]: The authors should test whether the model can predict variation in gene expression between individuals with variable genetic sequence or whether it is only predicting the average gene expression.

Since our model uses a motif-based tokenization approach (see R2[11]) the model is not very sensitive to nucleotide-level variation and often predicts minimal changes in expression if only one nucleotide is changed. However, the chromatin accessibility landscape across different individuals indeed enables the model to predict variable expression profiles, as we show by finetuning the base model on glioblastoma patient samples (see R1[3]).

This current limitation of our model can be addressed by replacing the motif preprocessing step with an end-to-end architecture where motifs are used as convolutional kernels.

[REDACTED]

Identification of CREs

R2[16]: The authors should address whether the type of base-editing data they used is based on a technique that already has different efficiency/effect in open chromatin and therefore may already be more closely correlated with the chromatin accessibility data used to train GET to begin with.

To address this concern, we refer to a recent publication by Song et al.[37] The authors evaluated the efficiencies of an adenine base editor (ABE) and a cytosine base editor (CBE) at a large number of target sequences in human cells: “Here, we found that ABE and CBE activities at positions 3–10 in DNase I hypersensitive (DHS) target sites were, respectively, barely (1.1-fold, P = 0.55) or significantly (1.9-fold, P = 0.0077) higher than at positions in non-DHS sites (Supplementary Fig. 3), indicating that CBE activities are higher at sites with high chromatin accessibility than at those with low chromatin accessibility and that ABE activities are hardly affected by chromatin accessibility. Here, we found that ABE and CBE activities at positions 3–10 in DNase I hypersensitive (DHS) target sites were, respectively, barely (1.1-fold, P = 0.55) or significantly (1.9-fold, P = 0.0077) higher than at positions in non-DHS sites (Supplementary Fig. 3), indicating that CBE activities are higher at sites with high chromatin accessibility than at those with low chromatin accessibility and that ABE activities are hardly affected by chromatin accessibility.”

Importantly, their results suggest that the efficiency of the ABE was not significantly influenced by chromatin accessibility: "We found that there were only modest asymmetric correlations between the activities of the base editors and Cas9 at the same targets." This finding indicates that the ABE-mediated base-editing data used in our study is unlikely to be heavily biased by chromatin accessibility.

We have updated the section **“Enhancer-gene pair prediction”** in **Methods** to include a discussion of this relevant study and its implications for our work.

R2[17]: Given that GET has seen data from erythroblasts, the authors should indicate whether the other methods (e.g. Enformer, etc.) have seen data from this cell type. The authors should also show the efficacy in a cell type not seen by GET and generally a wider set of different cell types to get a better measure of its generalizability.

We used erythroid cell types consistently in all benchmarking in our study. For the Enformer model, we utilized the “CAGE:CD34 cells differentiated to erythrocyte lineage”

dataset. For the ABC score, we used the file "AllPredictions.AvgHiC.ABC0.015.minus150.ForABCPaperV3.txt.gz" downloaded from https://www.engreitzlab.org/resources and filtered for erythroblast predictions. For DeepSEA, we used the original scores produced in Cheng et al.[17], which are cell-type agnostic.

We acknowledge the importance of assessing the generalizability of GET across a wider range of cell types, including those not seen during training. To address this, we have updated the manuscript in section "**Zeroshot GET prediction of expression-driving regulatory elements in new cell types"** to include additional analyses on the performance of GET in patient glioblastoma (GBM) tumor cells. We refer the reviewer to R2[13] for a detailed overview of these results, as well as reported performance on a perturbed H1 hESC cell line.

We have conducted additional experiments using GET enhanced by quantitative ATAC-seq data in a broader range of cell types. This model, referred to as the "Quantitative ATAC" model, incorporates quantitative peak counts from ATAC-seq as input features. Additionally, we employed a "Binary ATAC" model for comparison, where all non-zero ATAC peak counts are set to 1.

To test the model's generalizability, we executed three independent runs, each one with a different random seed, across various leave-out cell types. Our findings demonstrate that GET consistently surpasses the mean expression baseline across all evaluated fetal cell types ("Training cell type mean"). We also benchmark expression prediction for the same set of cell types with Enformer. Because Enformer was not trained on RNA-seq output tracks, we follow the linear probing procedure described in the original publication's benchmark against ExPecto (Extended Data Figure 4 from Enformer[5]). We fit a linear model on top of Enformer's CAGE output track predictions using RNA-seq from the particular cell type as targets.

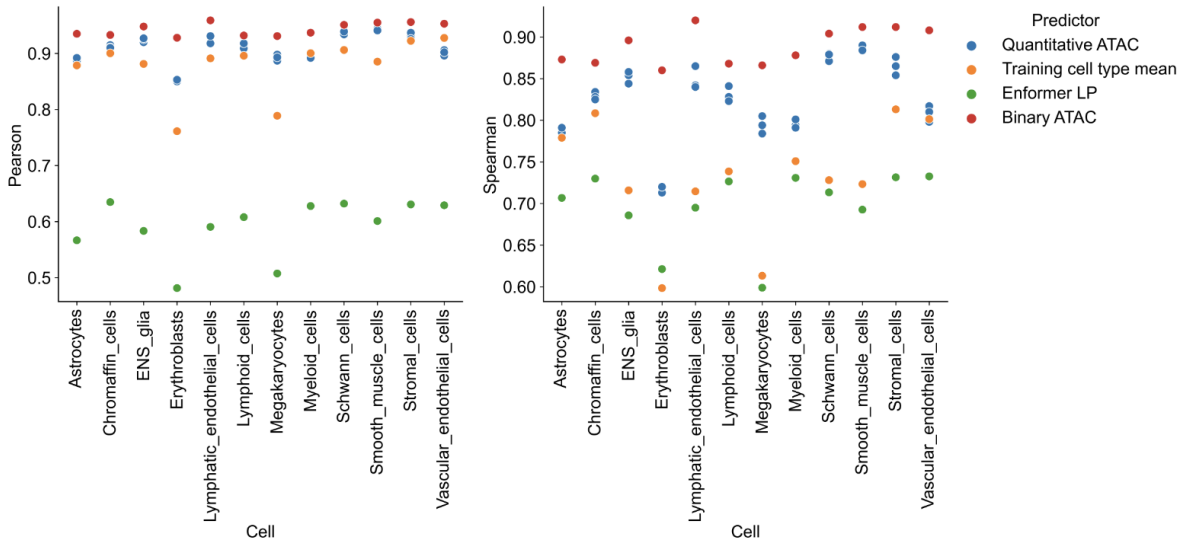


Figure. *Expression prediction for two settings of GET in which quantitative ATAC ("Quantitative ATAC") and binarized ATAC ("Binary ATAC") are included as input features. GET was finetuned on expression data from all other cell types and evaluated on the reported, held-out cell type. We report two baselines for comparison: the mean expression from the pretraining set ("Training cell type mean") and predictions from linear probing of Enformer on RNA-seq ("Enformer LP").*

We have incorporated this analysis in the **Methods** section (section "**Model evaluation: Cross-cell-type prediction"**) and updated **Supplementary Figure 2c**.

R2[18]: The authors should comment further on the particular cis-regulatory region prediction benchmarks where other models outperformed GET (e.g. HBG2 distal pairs) and if there are any known differences between those benchmarks compared to the others that would explain the difference in performance.

We refer the reviewer to **Figure 3** and **Supplementary Figure 4** of the revised manuscript, in which we show improved performance on the CRE prediction task for Erythroblasts and include an additional benchmark on K562 (please also see R2[1] and in particular R3[6]). The promoter annotation used in the original experiments for HBG2 originally was not directly overlapped with our ATAC-seq peak, potentially leading to degraded performance.

We have updated our CRE prediction analysis and developed a new way of calculating the score. Recent studies have highlighted the dominant importance of 1D genomic

distance in governing CRISPRi enhancer knockout effects (e.g. Gschwind et al.[18]). In this benchmark, most methods include a component of genomic distance. For example, Enformer incorporates exponential decay in its positional encodings. HyenaDNA incorporates a sinusoidal positional encoding over the DNA sequence, and our benchmarking results follow an exponential decay from the TSS (see **Figure 3c** NFIX). We have also extended GET to incorporate distance information. In particular, we designed a simple DistanceContactMap module for GET to convert the pairwise 1D distance map between peaks to a pseudo-Hi-C contact map. DistanceContactMap is a simple 3-layer 2D convolutional neural network (kernel size 3) with log10 (Pairwise Distance+1) as input and SCALE-normalized observed contact frequency as output. A Poisson negative log-likelihood loss was used to train the model. We trained DistanceContactMap with the same K562 Hi-C data (ENCFF621AIY) used for training ABC Powerlaw, resulting in a 0.855 Pearson correlation, which mostly captured the exponential decay in contact frequency. We termed the prediction of this model "GET Powerlaw." The other two scores shown in **Figure 3d** are defined as:

1. GET (Jacobian, DNAse/ATAC, Powerlaw) = GET Jacobian + aTPM × GET Powerlaw
2. GET (Jacobian, Powerlaw) = GET Jacobian × GET Powerlaw

This model can be further improved as future work by taking GET region embeddings as additional input and learning to predict cell type specific 3D contacts.

Embedding characterization

R2[19]: The authors should show the data underlying their statement that the embedding from final layers correlates with gene expression while earlier layers are more indicative of differences in regulatory grammar.

We thank the reviewer for the requested clarification. Below we provide a visualization using UMAP of two final layer embeddings in GET. Each point represents a (TSS, cell type) pair, transformed from the 768-dimensional embedding vector corresponding to a specific TSS from a specific cell type. The plot on the left is colored by TSS accessibility, while the plot on the right is colored by the expression level. The two arms in the figure correspond to directional transcription on two strands.

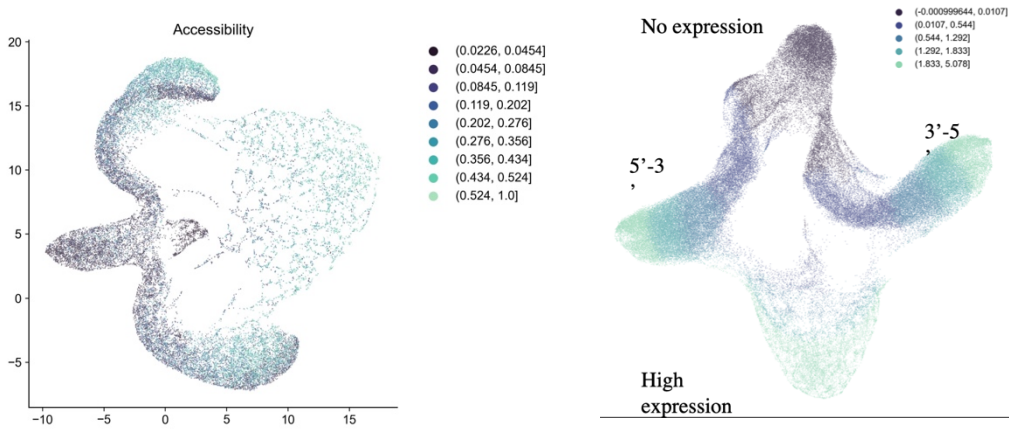


Figure. UMAP plots for two final layer embeddings from pretrained GET. (left) Final layer embedding colored by TSS accessibility. (right) Final layer embedding colored by expression level, with arms showing separation by directional transcription on two strands.

We have updated the section **"Regulatory embedding"** in **Methods** and a discussion of our findings, providing evidence for layer-wise interpretation of the model's hierarchical representation learning of regulatory elements and their effects on gene expression.

R2[20]: In the UMAP visualization of embeddings, the embeddings are compressed to 2 dimensions so necessarily, the primary category distinguishing subsets of the data will dominate the visualization. It shows that the motif yields a greater distinction than the cell type but does not indicate that there is no cell type differentiation. The authors should show clustering with the full number of embedding dimensions (e.g. hierarchical clustering with heatmap) to further evaluate if the cell type does not play as much of a role in distinguishing the embeddings. However, in the data the authors show in Fig. 4b-c, the astrocytes do appear separate from erythroblasts, which are more similar as 1 and 3, so their data does seem to indicate that cell types are distinguished.

We thank the reviewer for raising this point. We agree with the reviewer that this embedding would be able to preserve cell type specificity, as shown in **Figure 4b,c** and explained in the main text: "Nonetheless, when we subset the embeddings to only three specific cell types (fetal astrocyte and two fetal erythroblast subclusters), the UMAP exhibited distinct clusters for astrocytes and erythroblast genes (**Figure 4b**). This result further corroborates that GET is proficient at discerning cell-type-specific regulatory information."

One of the most unique characteristics of GET is its cell type specific input with cell type agnostic training. This enables the model to construct a unified latent space across all

regulatory elements and all cell types, while also preserving cell type specificity when jointly considering all genes in a cell type. Following the reviewer's suggestion, we have investigated hierarchical clustering using the full embedding matrix. Two examples are shown below. On the left, we show a subsample of 1,000 promoters, with red row color annotating fetal cell types. On the right, we show a subsample of 10,000 promoters from the entire space, with red row color annotating genes from astrocytes. No clear patterns of cell types can be identified within clusters.

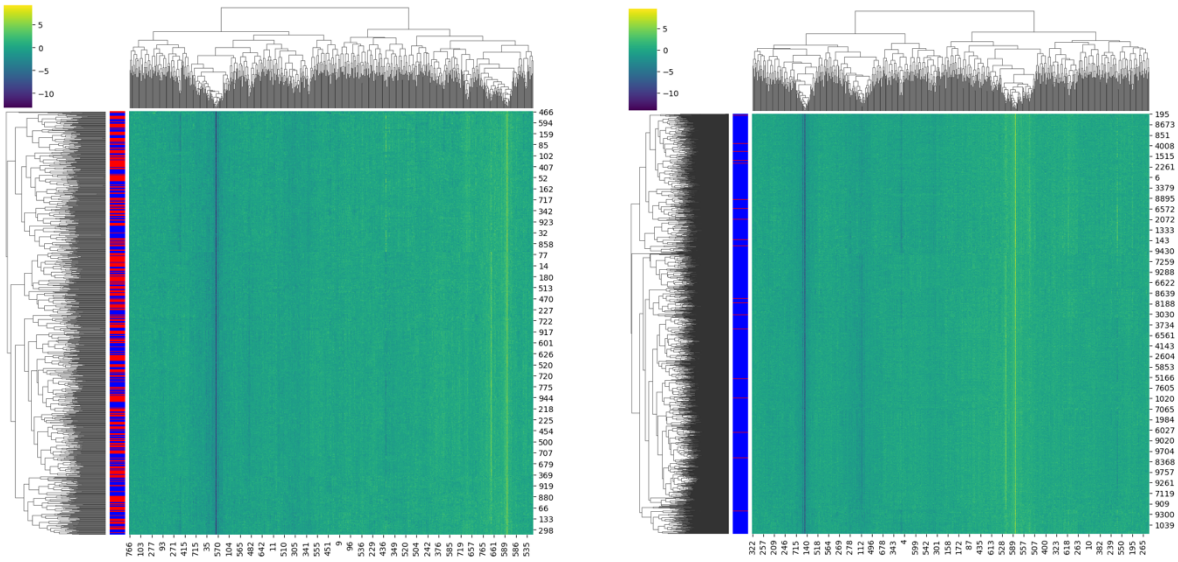


Figure. Heatmaps visualizing hierarchical clustering of the embedding space, indicating that cell types are distributed across clusters. (left) Subsample of 1,000 promoters, in which red rows annotate fetal cell types. (right) Subsample of 10,000 promoters, in which red rows annotate genes from astrocytes.

This unified regulatory embedding space is shared across all cell types without partitioning of the space by cell type or data source. In **Figure 4**, a cell type is represented as a collective set of points in this space, in which each point is a promoter. As a set, these points show cell type specificity by residing in different clusters. This is confirmed by the heatmap in **Figure 4e**, with Astrocytes enriched for a NFIA-related space, while Erythroblasts are enriched for an AP/1-related space.

R2[21]: The use of the 1st layer embedding in Fig. 4 should be compared to the use of the input data directly given there is only 1 layer of attention at play at that point, as well as comparing to a later layer.

We also conducted an analysis in which we visualized the same set of (gene, cell type) points using UMAP embeddings derived from promoter input motif vectors. These results are displayed in the below figure, with color coding consistent with that used in **Figure 4**. This analysis revealed that while the input motif vectors do not adequately distinguish between Astrocytes and Erythroblasts, they slightly preserve the clustering patterns evident from the first layer embedding. This observation suggests that even the initial transformation of the input data through the model's first layer adds significant value in data separation and interpretation.

Figure. UMAP visualization of promoter input motif vectors. Some clustering patterns visible in the first layer embedding (**Figure 4b**) is shared by the embedding space of the promoter input motif vectors.

Moreover, as we progress deeper into the model, the layers are increasingly compressed towards the expression space, which is characterized by parameters such as expression level (low to high) and strand orientation (positive or negative). This compression makes deeper layers less suitable for motif-related interpretations as they are more abstracted from the input features and more aligned with the output gene expression characteristics.

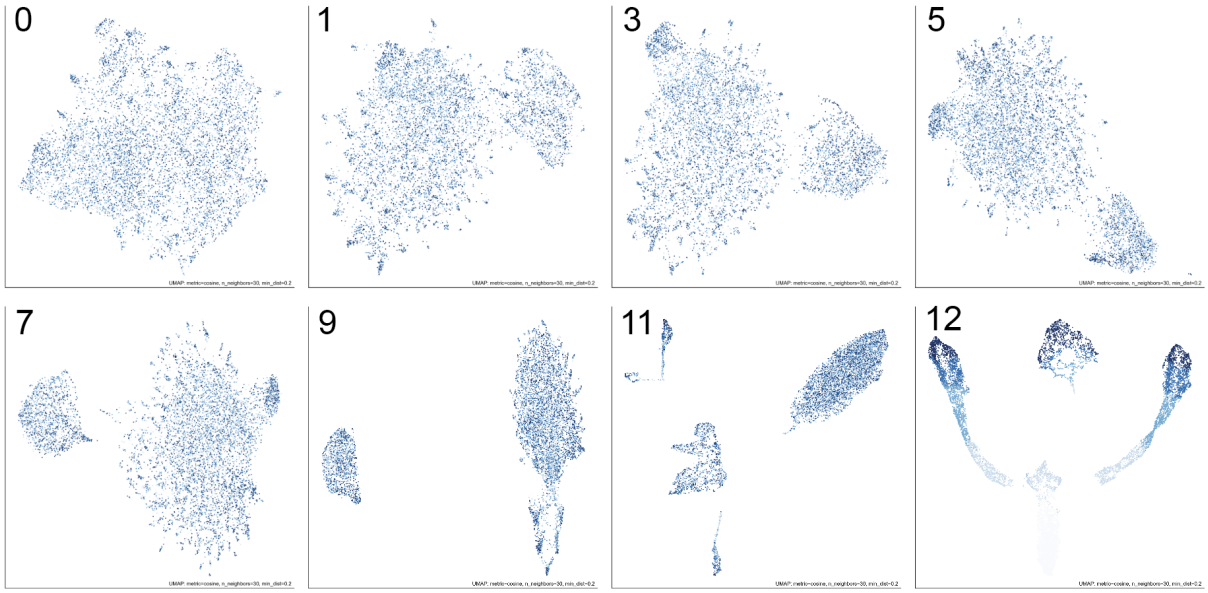


Figure. Latent embeddings after each transformer layer. Each dot is a gene colored by expression value.
This figure has been now incorporated in **Supplementary Figure 5b**.

Protein interaction networks

R2[22]: It is unclear what the causal direction in the neighborhood graph in Fig. 4g means. Is the predicted causal direction validated in any way compared to ground truth? What does a negative interaction mean with regard to the protein binding interactions? How is this directionality incorporated into the comparison with the STRING database?

Causal inference of direction and effect size of motif-motif interactions is determined using the established causal discovery method LiNGAM (https://github.com/cdt15/lingam). Positive and negative scores represent an estimated effect size of the causal effect from one motif to another. However, we are not aware of high throughput experiments that can serve as a ground truth for this causal direction.

In the main text, we provided some known examples where potential negative regulation occurs, e.g. methylation related motifs which are usually anti-correlated with active transcription like MBD2, MECP2, and NRF1. When benchmarking against STRING, we ignore the causal direction and use only absolute effect size to rank the motif pairs.

R2[23]: Macro F1 should be reported rather than the true positive rate alone since this can be artificially inflated by repeated positive predictions.

We refer the reviewer to R2[10] for the updated results. We include a figure with macro F1 score, as well as added ChIP-seq colocalization ground truth. We have also updated **Figure 4h** in the manuscript accordingly.

R2[24]: The authors should perform basic experimental validation of their prediction that the G183S germline mutation disrupts interactions between PAX5 and other binding partners.

We appreciate the reviewer's suggestion to perform an experimental validation of PAX5's interaction with NR/3 transcription factors. We have tried two different approaches to characterize the interactions: Co-immunoprecipitation (Co-IP) and BioID-based proximity labeling. Our first try with Co-IP failed to detect the interaction between PAX5 and NR2C2, among other NR TFs. However, as a previous study on TF-TF interaction has recently shown, direct interaction shown by cryoEM can be missed by IP-MS, an *in vitro* pull-down assay, due to weaker interaction strength, which often happens when disordered regions are involved in the interaction[11].

To overcome the limitations of co-IP in capturing transient interactions, we performed BioID proximity labeling experiments to characterize PAX5-NR/3 interaction in B-ALL cell lines REH. To select candidate NR transcription factors, we prioritized all TFs in the NR/3 motif family based on their expression in 15 B-ALL PAX wildtype patient samples in a longitudinal study[12] (see Figure below; Diagnosis: green; Relapsed: orange).

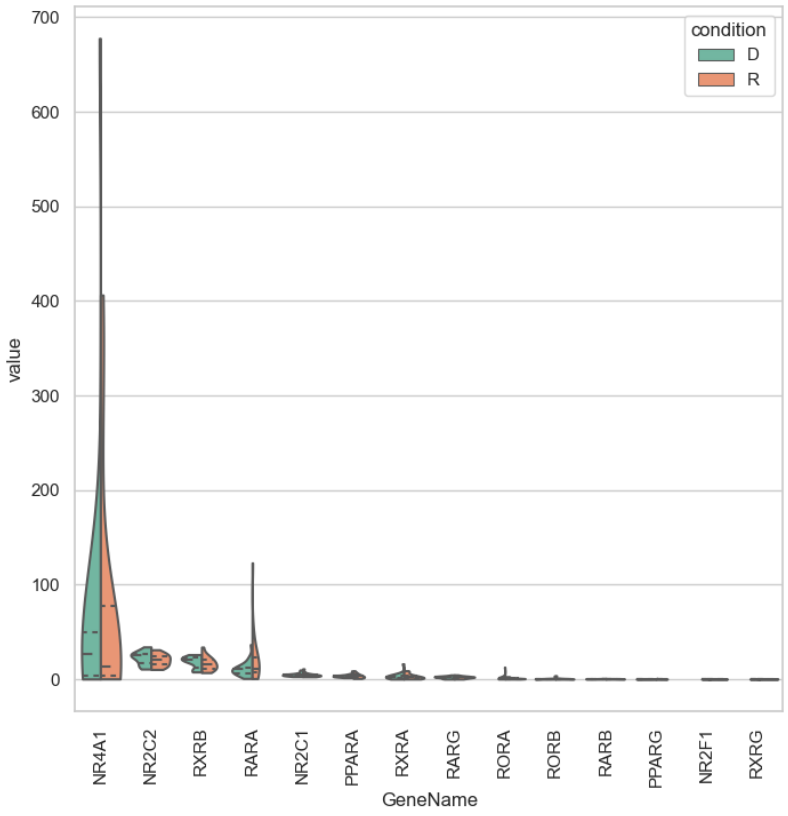


Figure. *Nuclear receptor transcription factor candidates based on expression in 15 B-ALL PAX wildtype patient samples in a longitudinal study.*

Since NR4A1 and NR2C2 yielded the highest expression values, we selected them for further experimental validation with both wild-type PAX5 and the G183S mutant. As positive controls, we also included previously reported nuclear receptor corepressor/cofactors NCOR1 and NRIP1[13], and a known interacting NR/20 TF, NR3C1, whose motif was predicted by GET to be non-interacting with PAX5. RHOA was used as a negative control.

The results of our experiment demonstrated that PAX5 is indeed part of the same regulatory complex with NR2C2 in B-ALL cell lines. This conclusion was further supported by colocalization of ChIP-seq peaks in analyses for PAX5, NR2C1 (heterodimer of NR2C2) in B-ALL cell lines, as well as a previous BioID-MS analysis of PAX5 interacting proteins[1].

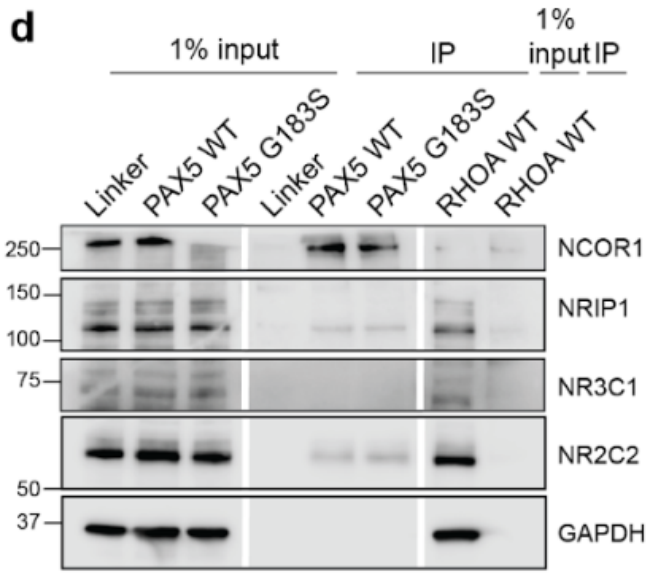


Figure. *Detection of PAX5 and RHOA BioID-fusion proteins in REH cells by western blot analysis using anti-PAX5 and anti-HA antibodies. 1% of total protein lysate was used as input.*

The experimental results agree with the GET prediction, showing that a NR/3 transcription factor NR2C2 is interacting with PAX5 (wt/mut). As a negative control, the NR/20 transcription factor (predicted to be interacting with NR/3 but not PAX5 in B cells) NR3C1 showed no interaction with PAX5 (wt/mut), although they are known to be both in the NR3C1-NR2C2-NRIP1 complex[13]. This suggests a direct PAX5-NR2C2 interaction, as predicted. NR4A1 did not show interaction with PAX5. The western blot analysis is included in the revised manuscript as **Figure 6d**.

To assess how PAX5 G183S affects the interaction with NR2C2, we quantified the interaction by normalizing to NR2C2 expression level (band intensity) in 3 independent replicates (normalizing to GAPDH leads to similar results). Our findings suggest that both PAX5 wild-type and G183S mutant interact with NR2C2 and their interaction strength differs, potentially affecting co-regulated transcriptional programs. The following figure has been included in the main manuscript as **Figure 6e**.

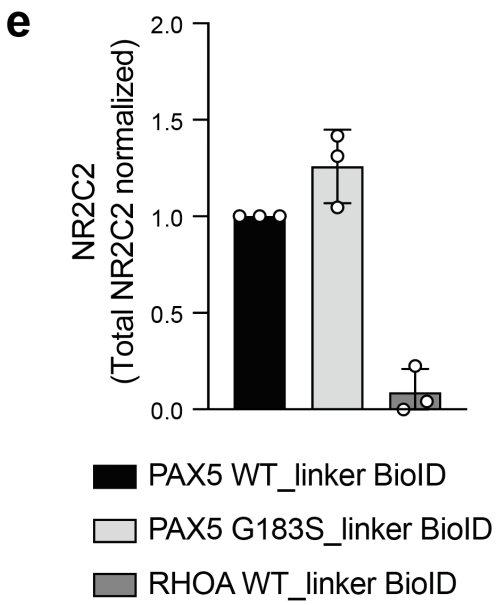
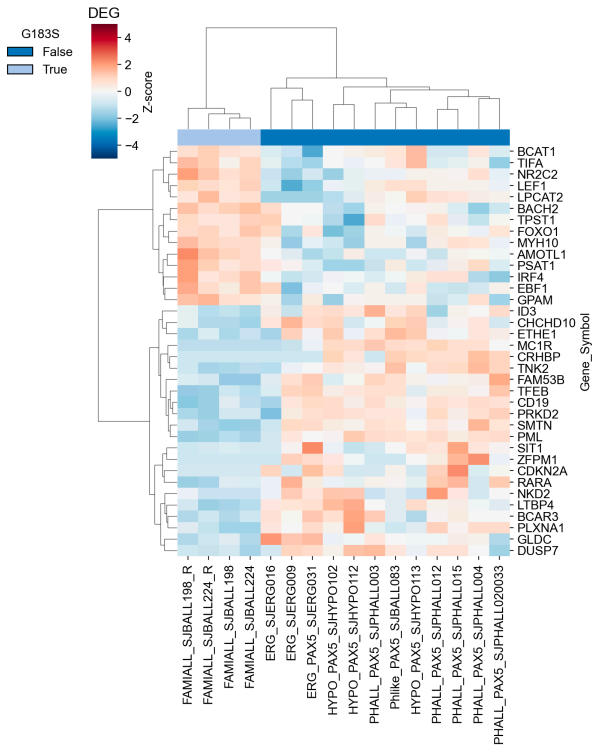


Figure. *Quantification of PAX5-NR2C2 interaction in the streptavidin immunoprecipitation in 3 replicates.*

To investigate the effect of the PAX5 G183S mutation on transcriptional programs on NR/3 and PAX5 regulated genes, we examined 141 sporadic childhood B-ALL samples including 4 familial B-ALL samples with PAX5 G183S germline variants[2]. We first selected

only CDKN2A loss cases (n=20) to match the condition with familial cases (n=4, biallelic mutations, loss of WT PAX5 and CDKN2A).

Figure. *Heatmap with the specific transcriptional program for 141 cases of sporadic childhood B-ALL with the germline G183S mutation.*



We further stratified the patients into two groups to perform two different comparisons: 1) PAX5 WT (n=10) vs PAX5 Loss (n=10), as a negative control; and 2) PAX5 G183S (n=4) vs PAX5 loss of function cases (e.g. PAX5 P80R, PAX5 Loss, n=12), for defining the G183S-specific transcriptome signature. The analyses suggest that there is a specific transcriptional program associated with the G183S mutation. This figure has been included in the revised manuscript as **Supplementary Figure 9e**.

To evaluate if this specific transcriptional program is relevant to the PAX5-NR interaction, we performed further gene set enrichment analysis. We defined the PAX5-specific, NR/3-specific, and PAX5+NR/3 commonly regulated genes with GET (**Methods**). We found an enrichment in the PAX5-loss associated differentially expressed genes (fold enrichment 1.22, p-value: 1.5e-3, hypergeometric test with Benjamini-Hochberg correction) with predicted PAX5 targets and no significant enrichment in NR/3 regulated and PAX5+NR/3 regulated genes, as expected. However, the PAX5 G183S-specific differentially expressed genes showed significant enrichment in NR/3 and PAX5+NR/3 regulated genes (fold enrichment: 1.41, 1.35; p-value: 1.05e-30, 1.57e-18, respectively). Moreover, we found that this mutant specific differentially expressed gene set is also enriched in the

genes activated in ProB cells (fold enrichment: 1.57, p-value: 2.2e-3), including well-known functionally relevant genes like CD19, EBF1, BACH2, and PML. This is now included as **Figure 6h** in the revised manuscript.

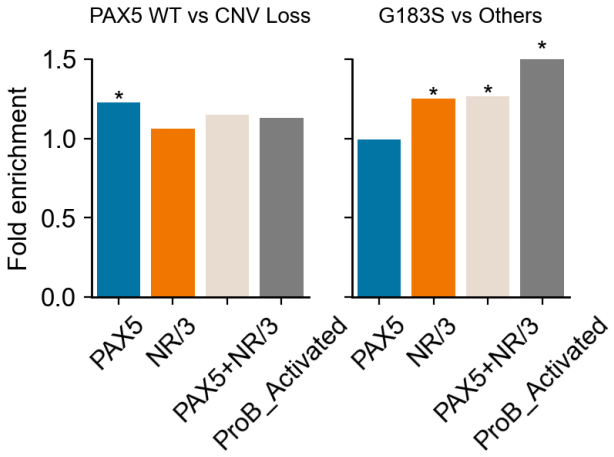


Figure. *Enrichment analysis for differentially expressed genes between PAX5 wild type v.s. PAX5 loss (left) and PAX5 G183S vs. other PAX5 alterations (CNV loss, P80R). * indicates statistical significance from a hypergeometric test (Benjamini-Hochberg adjusted p-values are reported).*

Minor Points

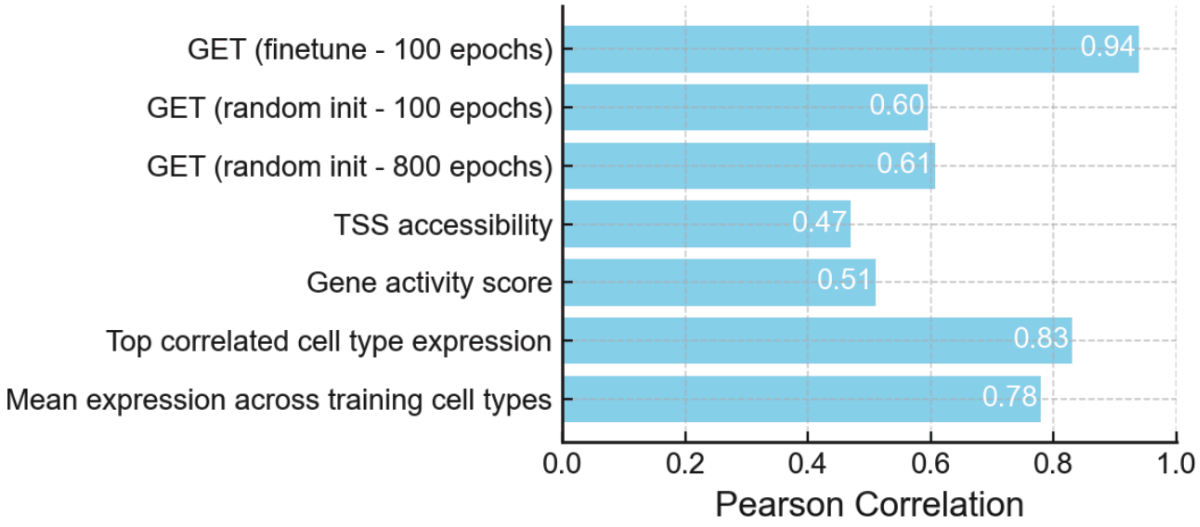
R2[25]: In the introduction, the authors cite models that use sequence to make predictions, but sequence is not context-specific (the same genome is present in every cell). Given that one benefit of using chromatin accessibility is that it encodes the context of each cell's state, it would be important in the introduction to provide the context to readers of other recent foundation models that use transcriptomic data as the input since that also encodes the context-specific networks in each cell state.

We have now included the following text in the introduction to contextualize our work against existing transcriptomic foundation models: “In the field of transcriptional regulation, a foundation model has the potential to synthesize the vast complexities of regulatory mechanisms across various cell types, offering a versatile framework that can be finetuned to target specific applications, cell types, or conditions. For example, recent developments in single cell foundation models like Geneformer, scGPT, and scFoundation have demonstrated how diverse transcriptome profiles can be encoded in one model, enabling various contextualized downstream tasks including cell type annotation and perturbation prediction. However, a foundation model of how transcription emerges from the chromatin landscape has not yet been explored.”

R2[26]: In the section "GET accurately predicts gene expression in unseen cell types at experimental accuracy", there appear to be missing figure references in the second to last 2 paragraphs. For example, "Our findings showed an average R^2 of 0.53 across diverse adult cell types..." does not have a related figure reference. Similarly, data should be shown for the result that performance dropped comparing fine-tuning with random initialization to fine-tuning with using the pretrained weights.

We thank the reviewer for spotting this. We have now included the figure reference in the text: “Our findings showed an average R^2 of 0.53 across diverse adult cell types, once again surpassing the baseline (R^2 = 0.33) obtained using corresponding fetal cell types for prediction (e.g. utilizing fetal astrocyte to predict adult astrocyte) (**Figure 1e**).”

For the finetune vs. random initialization comparison, we refer the reviewer to the newly added **Supplementary Figure 2d** (as shown below), which shows improved performance when finetuning from pretrained weights.



R2[27]: Also in the above section, there are three cell types referenced where the model cannot beat baseline but only one is explicitly mentioned (pancreatic acinar cell). The authors should list all three cell types here.

The additional two cell types in which the model cannot beat baseline performance are:

1. Blood Brain Barrier Endothelial Cell, endothelial cell
2. Tuft Cell, intestinal tuft cell

As shown in the coloring in **Figure 1e**, the poor performance of these samples can be attributed to low cell count. We have updated the manuscript to incorporate this

information: “The only 3 cell types where we cannot beat the baseline are cell types with low cell counts in either ATAC-seq (blood brain barrier endothelial cell and tuft cell) or RNA-seq label (pancreatic acinar cell).”

R2[28]: The "Input data" section of the methods appears to have 2 repeated paragraphs written in different ways.

We thank the reviewer for spotting this. We have removed the redundant **“Input data”** paragraph from the **Methods** section.

R2[29]: It is unclear whether Supp Fig 3a only refers to chr 11 or not. Authors should clarify this.

As mentioned in R2[5], the original K562 ATAC experiment had a data leakage problem for evaluation on leave-out chromosome 11, leading to an unrealistic reported performance of 0.98. We have corrected this in the original manuscript. After redoing this analysis, we find a Pearson correlation of 0.75 for the left-out chromosome 14 (chosen for proper comparison with Enformer, which entirely leaves out chromosome 14 from its training set). Details for this evaluation are included in R3[4]. We have updated **Supplementary Figure 2g** and **Supplementary Figure 2h** with the K562 ATAC analysis.

R2[30]: Authors should add x and y axis labels to Fig. 2e.

We thank the reviewer for pointing this out. In this figure, each scatterplot has an x-axis of experimental readout (i.e. lentiMPRA log2(RNA/DNA)) and a y-axis of predicted expression (i.e. in-silico MPRA, as predicted by GET-MPRA or Enformer-MPRA). We have now further clarified the figure legend and axes labels in **Figure 2g**.

R2[31]: In Fig. 3c, it is unclear why the particular areas were chosen to be highlighted with gray shaded regions given that there are others with high correlation with ground truth.

In the original presentation of **Figure 3c**, we highlighted distal regions that are correctly predicted by our method but did not want to crowd the figure with too many highlights, choosing instead to select a few regions. To avoid this confusion, we have removed the highlights for a more unbiased presentation of the results in the figure.

R2[32]: In Supp Fig. 4, the authors should indicate why ABC did not achieve prediction (it's labeled "No prediction" without an explanation for this in the figure legend).

We downloaded the ABC prediction for Erythroblast from https://mitra.stanford.edu/engreitz/oak/public/Nasser2021/AllPredictions.AvgHiC.ABC0.015.minus150.ForABCPaperV3.txt.gz and overlapped the predicted enhancer-gene pairs with the benchmark dataset. The “No prediction” label indicates that the provided ABC score does not contain a prediction for the gene. To avoid this issue, we have taken a new approach to computing the ABC score by multiplying the Fetal Erythroblast ATAC signal with a reference distance power law trained on K562 Hi-C data (see R2[18] for details). We also refer the reviewer to R2[1] for a comparison with other methods on this benchmark.

R2[33]: In Supp Fig. 1, the authors should more clearly delineate which part of the bottom row is GET vs. Enformer (separate two halves more clearly with labels). They should also define abbreviations in the figure legend.

Following the reviewer's suggestion, we have updated **Supplementary Figure 1** to more clearly represent the GET architecture. We include the figure below.

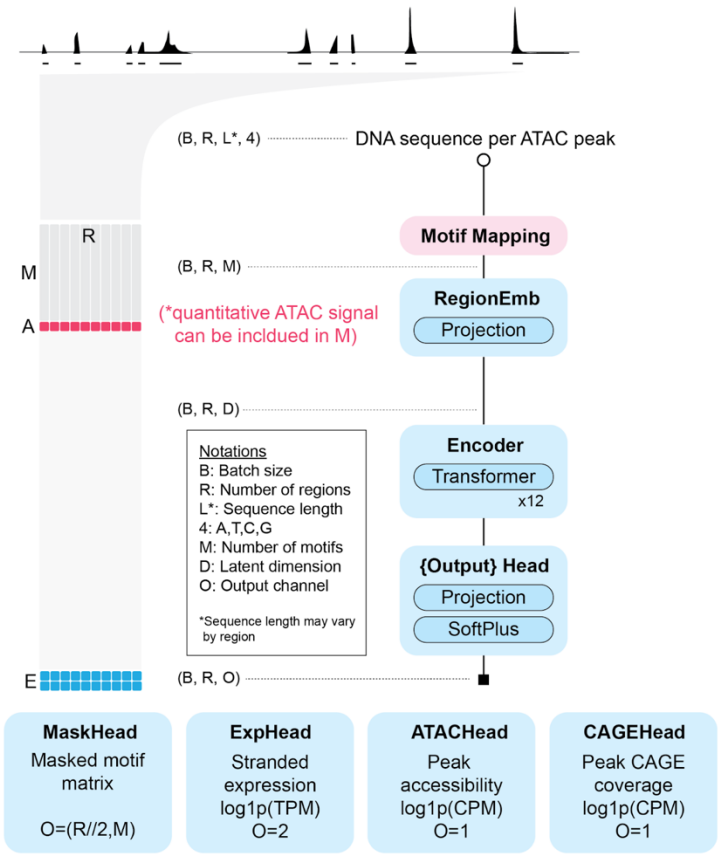


Figure. *Architecture of GET.*

Referee #3 (Remarks to the Author):

R3[0]: Fu et al. report a computational model using genomic DNA sequence and cell type-specific epigenomic data to predict gene expression and regulatory element activity. Notably, this model makes predictions for cell types outside of the training set (or left out), which the authors report as a major advance over the previous state of the art. The model predictions are presented for 213 human adult and fetal cell types, as well as a collection of cell types in the lymph node. The model is interpreted to characterize regulatory element syntax at the level of TF motifs and identify TF pairs with likely physical protein interactions. The authors present some interesting results, but there is a critical lack of well-considered comparisons to previously reported models allowing the assessment of the benefit of this approach. Additionally, the manuscript suffers from consistent structural issues – e.g., a lack of the text referencing the correct figures, the figure legends matching the data in the figures, missing references to data, numbers quoted in the text do not match those in figures, misleading statistics that do not capture the data, etc. Together, this manuscript requires significant improvement to demonstrate clearly that the model provides improved and biologically meaningful insights compared to previous efforts.

Major points

R3[1]: The manuscript is missing a clear, quantitative assessment of model performance on gene expression and its distribution across single cells. The authors need to clearly show prediction performance within and outside of the training cell types, as well a comparison to Enformer performance on the same cell types. The authors should show the performance for all left-out cell types rather than just astrocytes as an anecdote in Figure 1.

We thank the reviewer for their insightful comments. To address the comments of the reviewer and the previous reviewers about generalizability (see R1[2] and R2[3]), we validated our model in a broader range of cell types. In addition to the zero-shot prediction performance across all adult cell types using a model trained on only fetal data (**Figure 1e**), we have extended our evaluation of the model to 12 coarse classes of fetal cell type groups, 18 cell states in a SHARE-seq data of a perturbed H1 hESC cell line, and 10x Multiome data from 18 IDH1-wild type glioblastoma patients (see R1[3]).

To test the model's generalizability, we validated our model beyond astrocytes to include a broader range of cell types, including cell types transcriptionally different from those that appear in the training set. We executed three independent runs, each with a different random seed, across various leave-out cell types. For example, we left out all vascular endothelial cells for training and evaluated performance on vascular endothelial cells. We

have conducted these experiments enhanced by quantitative ATAC-seq data. This model ("Quantitative ATAC") incorporates quantitative peak counts from ATAC-seq as input features. Additionally, we trained a "Binary ATAC" model for comparison, where all non-zero peak counts are set to 1.

Our findings demonstrate that GET consistently surpasses the mean expression baseline across all evaluated fetal cell types ("Training cell type mean"). We benchmark expression prediction for the same set of cell types with Enformer. Because Enformer was not trained on RNA-seq output tracks, we follow the linear probing procedure described in the original publication's benchmark against ExPecto (Extended Data Figure 4 from Enformer[5]). We fit a linear model on top of Enformer's CAGE output track predictions using RNA-seq from the cell type as targets.

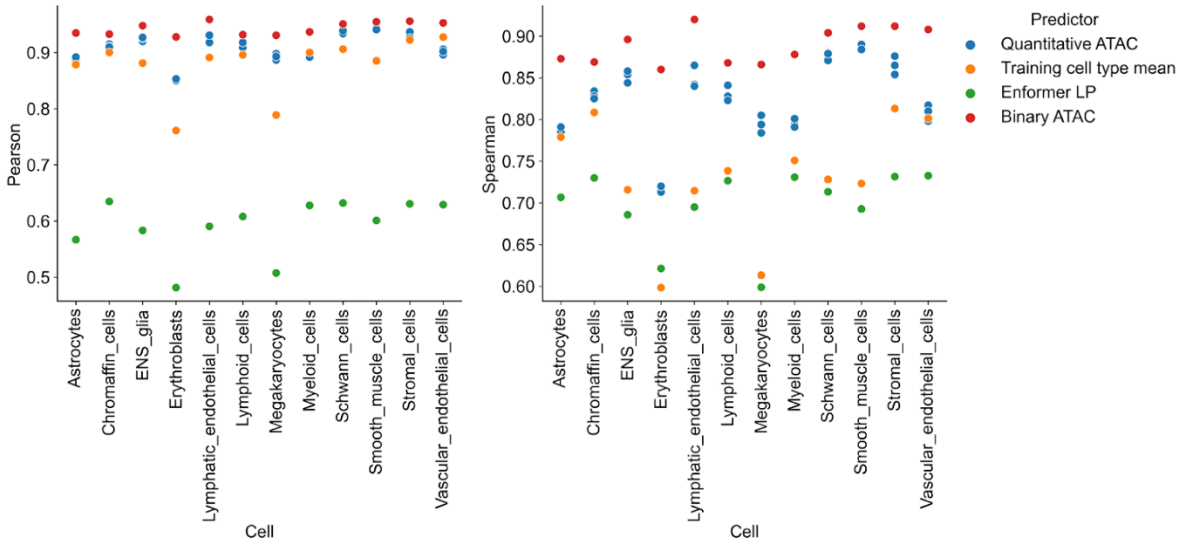


Figure. *Expression prediction for two settings of GET in which quantitative ATAC ("Quantitative ATAC") and binarized ATAC ("Binary ATAC") are included as input features. GET was finetuned on expression data from all other cell types and evaluated on the reported, held-out cell type. We report two baselines for comparison: the mean expression from the pretraining set ("Training cell type mean") and predictions from linear probing of Enformer on RNA-seq ("Enformer LP").*

We provide an additional expression evaluation against Enformer, in which we apply GET to predict K562 CAGE (FANTOM5 sample ID: CNhs12336). We first note that this comparison privileges Enformer, which was trained extensively on CAGE tracks, including K562 (track ID: 4828 and 5111), while GET must be transferred to the new assay. Here we evaluated finetuned GET against Enformer predictions summed across the two CAGE output tracks for a leave-out peak set across chromosome 14. We select

chromosome 14 because it did not appear in the public Enformer checkpoint's training or validation set. Pretrained GET was finetuned in three ways:

1. BATAC LoRA from BATAC pretrain: In this setting, the base model was trained on the fetal and adult atlases with binarized ATAC signal. In the finetuning, the ATAC data was binarized.
2. QATAC LoRA from BATAC pretrain: In this setting, the base model was trained on the fetal and adult atlases with binarized ATAC signal. In the finetuning, we used the original accessibility TPM (aTPM) for ATAC signal.
3. QATAC LoRA from QATAC finetuned: In this setting, the base model was the leave-out astrocyte RNA-seq prediction model trained on the fetal accessibility and expression atlas. In the finetuning, we used the original accessibility TPM (aTPM) for ATAC signal.

We note that these experiments leverage Low-Rank Adaptation (LoRA)[20] parameter efficient finetuning to achieve significant gains in time and storage complexity. On a single RTX 3090 GPU, all finetuning converged within 30 minutes, resulting in a 3 MB K562-CAGE-specific adaptor that can be merged into the base model.

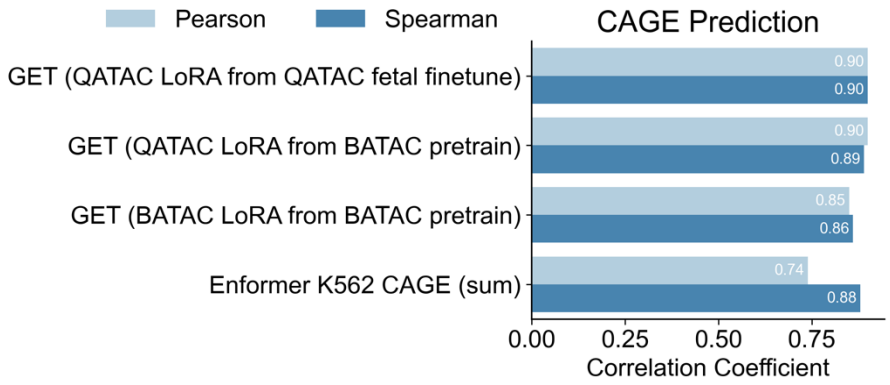


Figure. Three settings of GET finetuned on CAGE K562 dataset and evaluated on leave-out chromosome 14 as compared to Enformer predictions. GET finetuned with quantitative ATAC ("QATAC LoRA from QATAC fetal finetune" and "QATAC LoRA from BATAC pretrain") outperforms GET finetuned with binarized ATAC ("BATAC LoRA from BATAC pretrain") and Enformer predictions for the peak evaluation regions. Evaluation was performed for all TSS in chromosome 14.

We have incorporated this analysis in the **Methods** section ("**Transferring to new assays"**) and updated **Supplementary Figure 2f**.

Beyond K562 (a chronic myeloid leukemia cell line) mentioned in the original manuscript, we applied GET to both H1 human embryonic stem cells (hESC), and tumor cells from IDH1-wild type glioblastoma (GBM) patients. The hESC data is from the TF Atlas[6], a SHARE-seq based Multiome profiling of H1 hESC after multiplexed overexpression of different transcription factor isoforms. This dataset is thus appropriate for evaluating GET's performance on non-physiological cell states with dynamic chromatin landscapes. The GBM data is profiled with 10x Multiome for 18 patients from the NCI Human Tumor Atlas Network (GEO accession number GSE240822)[7], providing an adequate test for across-patient prediction. Both datasets involve not only cell type shift but also platform shift (sci-seq to SHARE-seq or 10x Multiome).

Zero-shot evaluation of the base state hESC (subcluster #0 in the below figure) using a model trained on expression data of fetal cell types yielded a Pearson correlation of 0.56. Further finetuning on the hESC data improved the Pearson correlation to 0.73 on held-out chromosomes 10 and 11. The Pearson correlation between hESC and the training cell type mean expression is 0.85.

To evaluate the leave-out cell-state performance of GET on different transcriptional states of hESC, we finetuned the model on pseudobulk data from all but one subcluster defined in the hESC 10x Multiome data. In this scenario, GET can reach a Pearson correlation of 0.96, while the maximum expression correlation between unseen and seen clusters is 0.98. Using this model, when trying to predict the differential expression between unseen subcluster #6_0 and the most similar seen subcluster (#3_3, having a Pearson correlation of 0.98 with #6_0), the correlation between predicted fold change and observed fold change is only 0.18. A more dissimilar pair (#6_0 vs. #0, Pearson correlation 0.95) yields a 0.50 Pearson correlation of fold changes (see figure below). This indicates that GET is able to capture cell-state specificity after TF perturbation.

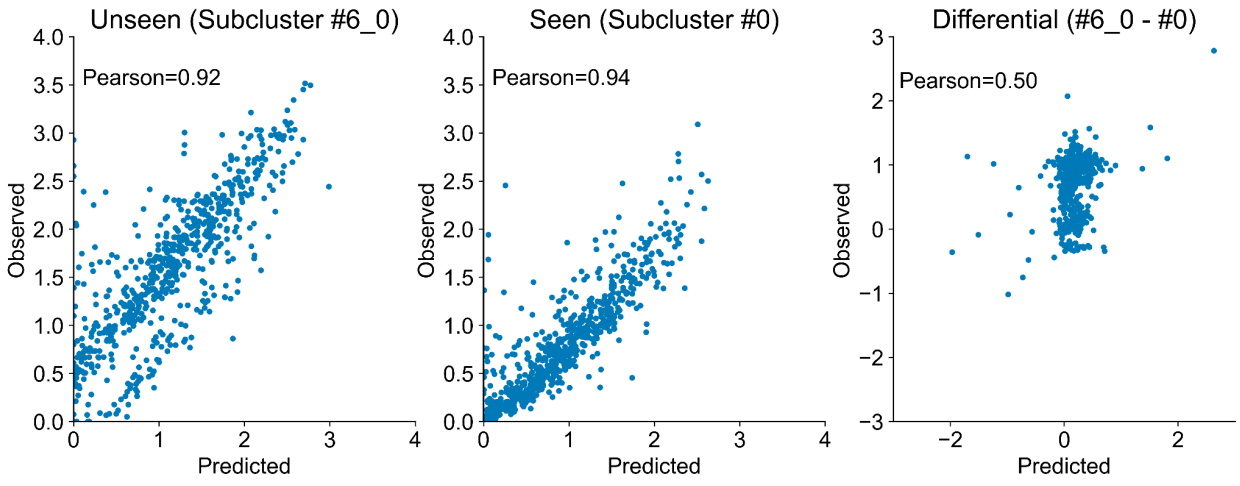


Figure. *Evaluating GET transferability to perturbed hESC SHARE-seq multiome data. Pearson correlations and scatter plots are shown between predicted and observed values for expression of one unseen cell state (#6_0) and one training cell state (#0) as well as their differential expression (log10 fold change).*

We next evaluated the performance of GET on tumor cells from IDH1-wild type glioblastoma (GBM) patients. The zero-shot performance on GBM tumor cells across 18 patients achieves a mean Pearson correlation of 0.68. We hypothesize that this higher performance in GBM cells compared to zero-shot hESC is due to their underlying similarity to astrocytes whose ATAC-seq data appeared in the pretraining set. This similarity likely makes it easier for the model to recall from pretrained embeddings, though the specific RNA-seq data for astrocytes were also not seen by the model.

We next finetuned GET on a single patient sample (referred to as the “one-shot” setting) and evaluated performance against the pretrained model (“zero-shot”) on 16 held-out patient samples. (Here we exclude two patients from evaluation to be used in the training set when evaluating robustness of the finetuning procedure; we finetune on only one patient sample at a time.) Finetuning on a single tumor patient sample shows that GET achieves a Pearson correlation of above 0.9 in the one-shot setting when predicting expression for held-out patients.

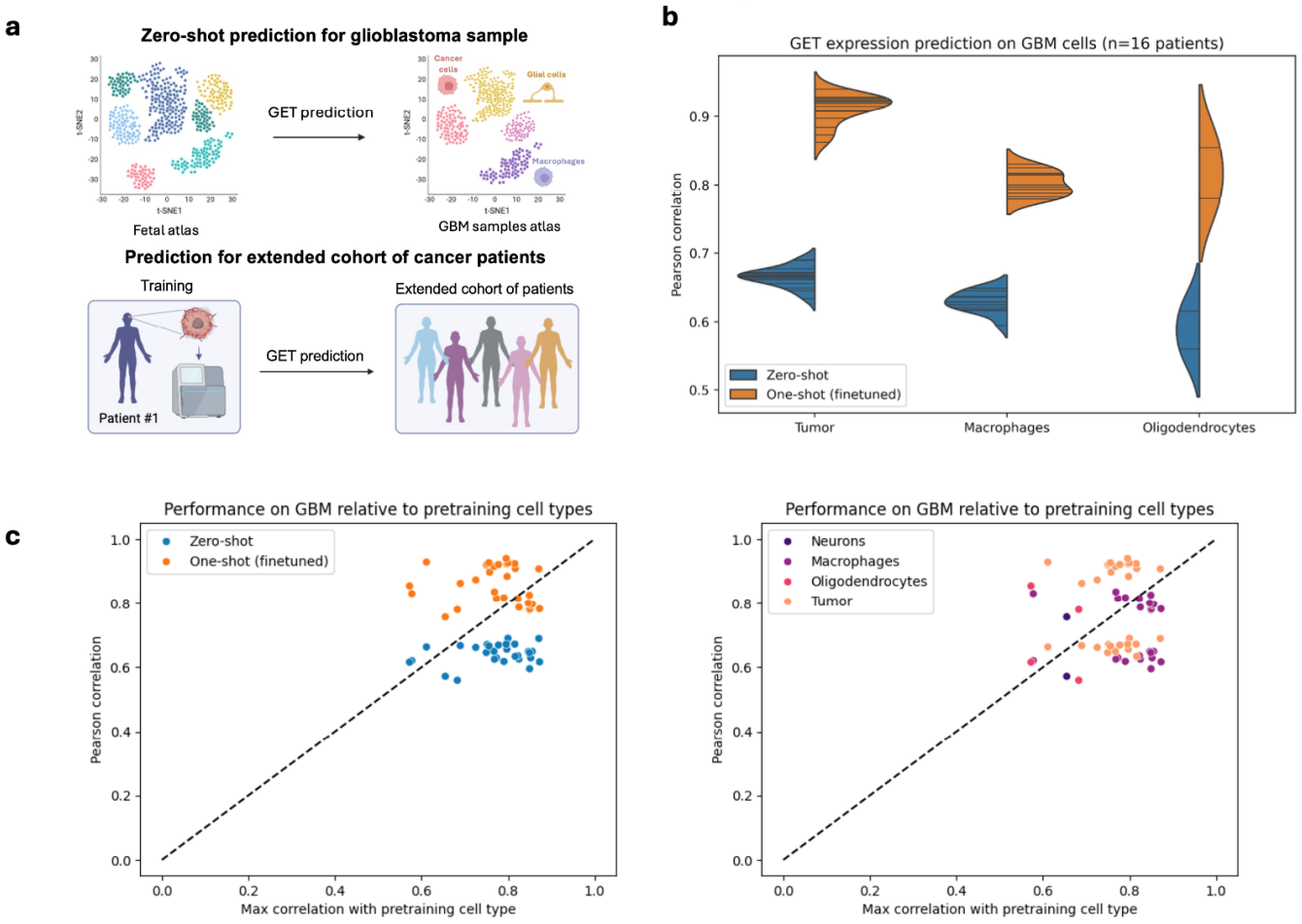
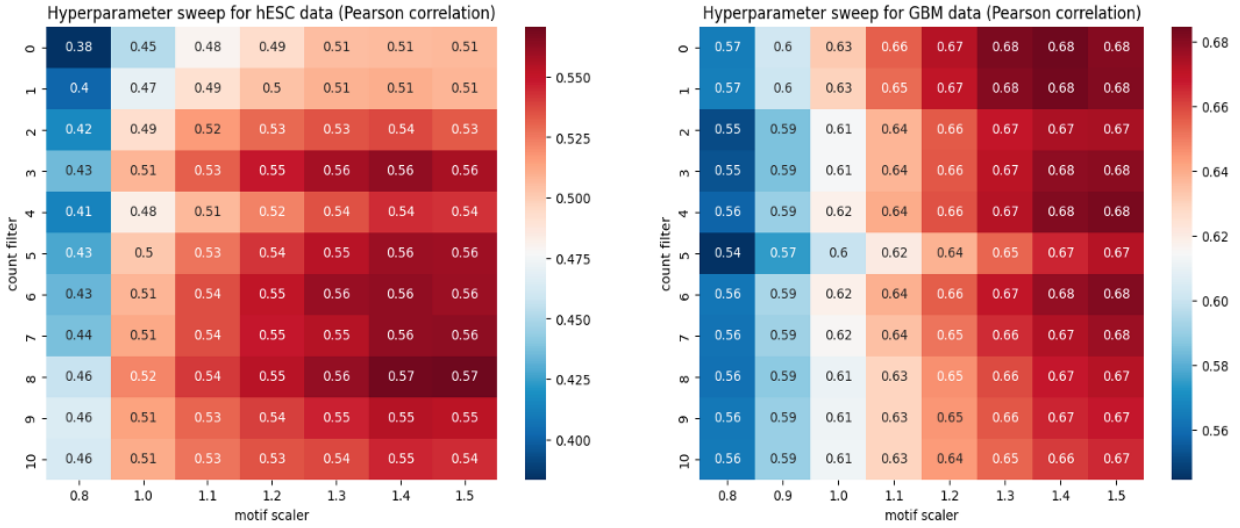


Figure. a) *GET was applied in the pretrained setting ("zero-shot") to predict gene expression from GBM patient samples. This prediction task represents both cell type and sequencing platform shift when compared to GET's original pretraining set. GET was also finetuned on a single patient sample ("one-shot") and used to predict gene expression for an extended cohort of glioblastoma patients.* ***b)*** *GET shows zero-shot and finetuned capability on a task to predict gene expression from held-out GBM patient samples.* ***c)*** *The one-shot model outperforms a baseline of predicting expression using the maximally correlated cell type in GET's pretraining set. Compared to the zero-shot setting, the one-shot finetuning improves performance across various cell types from GBM samples.*



Supplementary Figure. *To evaluate robustness of zero-shot performance in the hESC and GBM prediction settings, GET was evaluated with a sweep over two internal hyperparameters. We retain the optimal hyperparameter choices in subsequent finetuning experiments.*

In our revised manuscript, we have expanded the discussion elaborating on how different biological and technical factors could affect the predictive performance of our model across various cell types and describe the potential of GET to be applied to non-physiological cell types, including cancer cells. Figures a and b above are now included in the main manuscript as **Figure 2d** and **Figure 2e**.

R3[2]: Additionally, a claim that prediction matches experimental accuracy is made and must be quantitatively supported and explained (is this a comparison to technical or biological replicate measurements?)

In our study, we leveraged human RNA-seq data from GSE147870[38], which comprises data on Astrocytes cultured under various conditions and biological replicates. Please find attached the violin plot of correlation between different culture conditions (acutely purified, serum selected, serum-free cultures) across biological replicates.

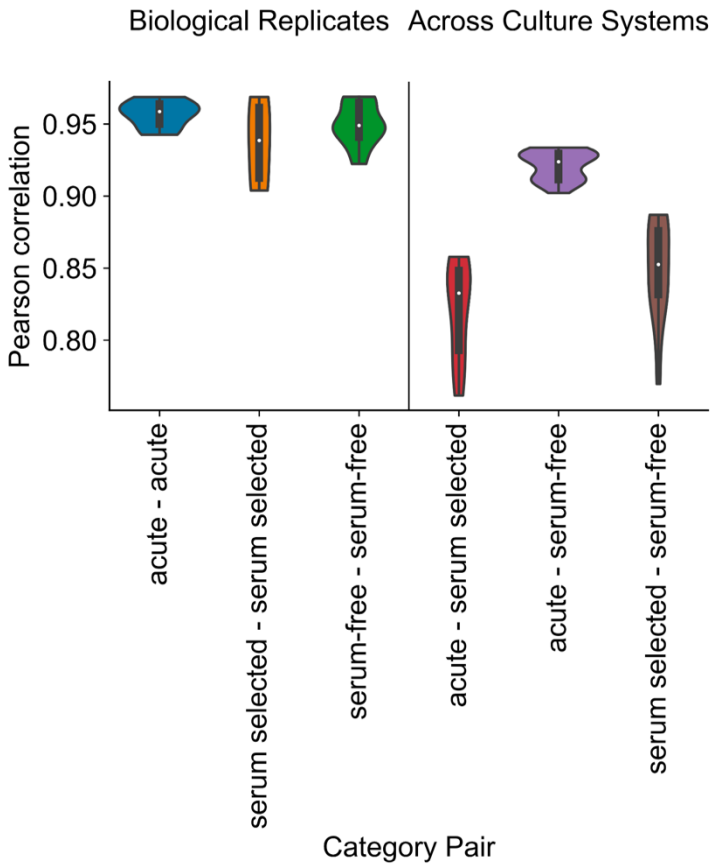


Figure. *Pairwise Pearson correlation between different culture conditions (acutely purified, serum selected, and serum-free) across biological replicates for human astrocytes.*

To clarify this point, we have updated the section **"GET accurately predicts gene expression in unseen cell types at experimental accuracy"** in the main text: "An example can be taken from left-out astrocytes, where the Pearson correlation between GET's predicted expression values and the observed expression reached 0.94 (R^2 = 0.88) (**Figure 1c,d**), a result that is in line with experimental accuracy across different culture systems and biological replicates of human astrocytes[17] (Pearson r = 0.92-0.99, **Supplementary Figure 2a**)."

R3[3]: There should also be deeper discussion of where and how the model succeeds and fails since much of the treatment is superficial and, in some places, misleading. This is also true for the data presented in Figures 4 - 6.

We observe that GET is successful at making accurate and generalizable predictions of gene expression when trained with only single cell ATAC-seq and RNA-seq data (see R3[1] and R3[2]). Prediction performance will not be as ideal when the target dataset has

a large biological or technical distribution shift, as we have demonstrated in the perturbed SHARE-seq hESC prediction results (see R3[1]). We have developed uniform data preprocessing and normalization pipelines to alleviate these discrepancies as much as possible. In the future, comprehensive peak sets like cPeaks[9] or ENCODE DHS index[24] can be constructed and used to make transfer to new cell types less prone to technological and processing biases. Trained on separately constructed scATAC-seq and scRNA-seq datasets, GET can efficiently learn important motifs that regulate genes, as well as functional interactions between motifs, which can lead to novel discovery of transcription factor-transcription factor physical interactions. However, GET does not directly pinpoint a specific TF among the same motif family. Additional information from multiome data and analysis of correlation between individual TF expression and potential target gene expression is required to further narrow down its predictions.

For motif-motif interaction analysis, due to the highly conserved nature of transcription factor DNA binding domains, most TFs in the same family have very similar motifs. Since our method pools the motif matching event at the peak level, the model will lose sensitivity to slight changes in motifs which may be essential to further delineate TFs in a family. We note that recent studies have shown redundant binding activities for TFs in the same family. For example, the homeodomain family TFs ALX1 and DLX1 both bind with E-box family TFs to form the “Coordinator” regulatory complex, yet major actors in the complex depend on the protein expression level in different cell types[11]. It is also worth noting that all TFs have transactivation domains that bind to cofactors and coregulators, which can also shift their binding profiles. Future study of GET should consider protein features such as sequence or protein-protein interactions to alleviate these issues. Even still, many of these uncertainties may only be solved using complementary experimental approaches. As suggested by the reviewer, we have now updated the **Discussion** section to analyze the advantages and caveats of our current model.

We also review here some additional analysis we have conducted to supplement **Figure 4** and **Figure 6** in response to other reviewers’ questions. For **Figure 4a,b,c**, we refer the reviewer to R2[19-21] for a discussion and additional analyses of the regulatory embedding space used in GET. We also refer the reviewer to R3[7] below, where we provide an improved version of the motif-motif interaction benchmark appearing in **Figure 4h**, which now includes a comparison with accessibility-aware motif-colocalization and TF ChIP-seq colocalization.

In **Figure 6**, we provide additional analyses and experimental validation of PAX5’s interaction with NR/3 transcription factors. We have updated the structure prediction in **Figure 6c** with AlphaFold 3[25], showing a more detailed view of G183, which appears to

be mediated by a hydrophobic interaction between isoleucine residues and hydrogen bonds between serine residues.

In **Figure 6d**, we tried two experimental approaches to characterize the PAX5 case study described in section **"GET uncovers the mechanism of germline mutations in disordered regions of transcription factors"** of the main text: Co-immunoprecipitation (Co-IP) and BioID-based proximity labeling. Our first try with Co-IP failed to detect the interaction between PAX5 and NR2C2, among other NR TFs. However, as a previous study on TF-TF interaction has recently shown, direct interaction shown by cryoEM can be missed by IP-MS, an *in vitro* pull-down assay, due to weaker interaction strength, which often happens when disordered regions are involved in the interaction[11].

To overcome the limitations of co-IP in capturing transient interactions, we performed BioID proximity labeling experiments to characterize the PAX5-NR/3 interaction in B-ALL cell line REH. To select candidate NR transcription factors, we prioritized all TFs in the NR/3 motif family based on their expression in 15 B-ALL PAX wildtype patient samples in a longitudinal study[12] (see Figure below; Diagnosis: green; Relapsed: orange).

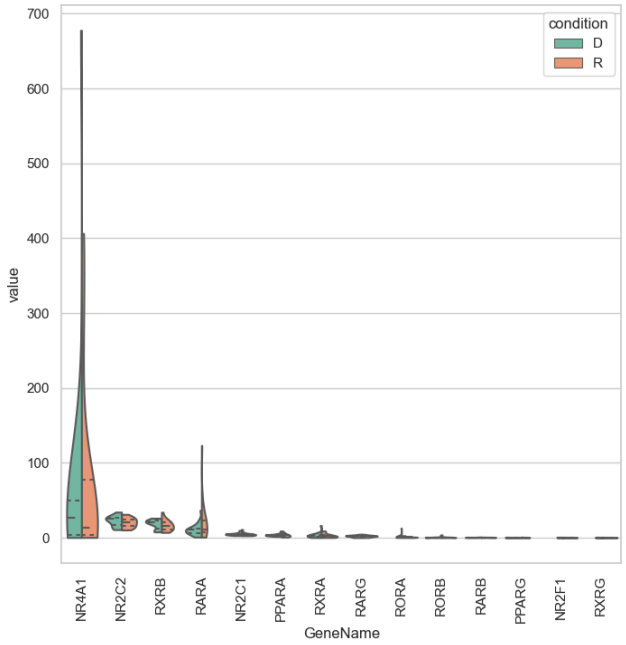


Figure. *Nuclear receptor transcription factor candidates based on expression in 15 B-ALL PAX wildtype patient samples in a longitudinal study.*

Since NR4A1 and NR2C2 yielded the highest expression values, we selected them for further experimental validation with both wild-type PAX5 and the G183S mutant. As positive controls, we also included previously reported nuclear receptor

corepressor/cofactors NCOR1 and NRIP1[13], and a known interacting NR/20 TF, NR3C1, whose motif was predicted by GET to be non-interacting with PAX5. RHOA was used as a negative control.

The results of our experiment demonstrated that PAX5 is indeed part of the same regulatory complex with NR2C2 in B-ALL cell lines. This conclusion was further supported by colocalization of ChIP-seq peaks in analyses for PAX5, NR2C1 (heterodimer of NR2C2) in B-ALL cell lines, as well as a previous BioID-MS analysis of PAX5 interacting proteins[1].

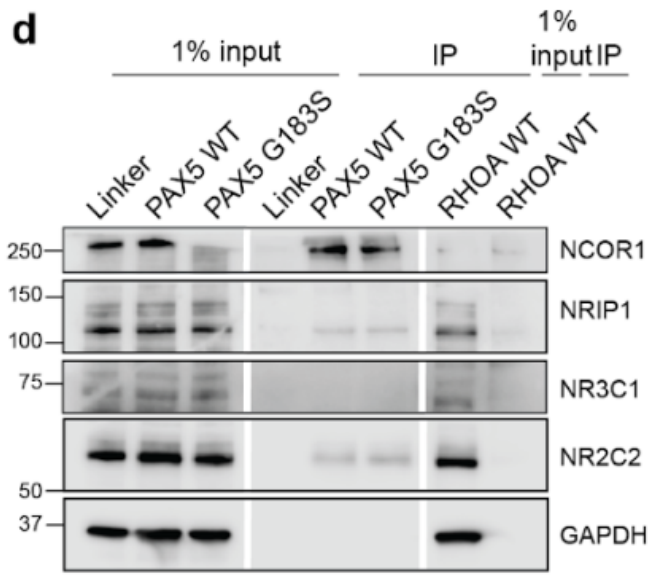


Figure. *Detection of PAX5 and RHOA BioID-fusion proteins in REH cells by western blot analysis using anti-PAX5 and anti-HA antibodies. 1% of total protein lysate was used as input.*

The experimental results agree with the GET prediction, showing that a NR/3 transcription factor NR2C2 is interacting with PAX5 (wt/mut). As a negative control, the NR/20 transcription factor (predicted to be interacting with NR/3 but not PAX5 in B cells) NR3C1 showed no interaction with PAX5 (wt/mut), although they are known to be both in the NR3C1-NR2C2-NRIP1 complex[13]. This suggests a direct PAX5-NR2C2 interaction, as predicted. NR4A1 did not show interaction with PAX5. The western blot analysis is included in the revised manuscript as **Figure 6d**.

To assess how PAX5 G183S affects the interaction with NR2C2, we quantified the interaction by normalizing to NR2C2 expression level (band intensity) in 3 independent replicates (normalizing to GAPDH leads to similar results). Our findings suggest that both

PAX5 wild-type and G183S mutant interact with NR2C2 and their interaction strength differs, potentially affecting co-regulated transcriptional programs. The following figure has been included in the main manuscript as **Figure 6e**.

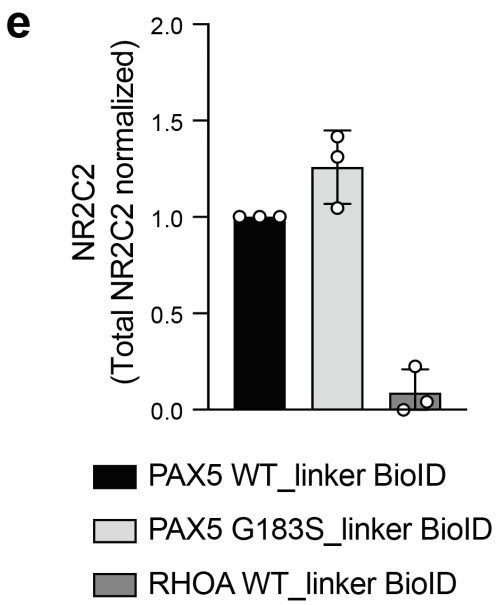


Figure. *Quantification of PAX5-NR2C2 interaction in the streptavidin immunoprecipitation in 3 replicates.*

To investigate the effect of the PAX5 G183S mutation on transcriptional programs on NR/3 and PAX5 regulated genes, we examined 141 sporadic childhood B-ALL samples including 4 familial B-ALL samples with PAX5 G183S germline variants[2]. We first selected

only CDKN2A loss cases (n=20) to match the condition with familial cases (n=4, biallelic mutations, loss of WT PAX5 and CDKN2A).

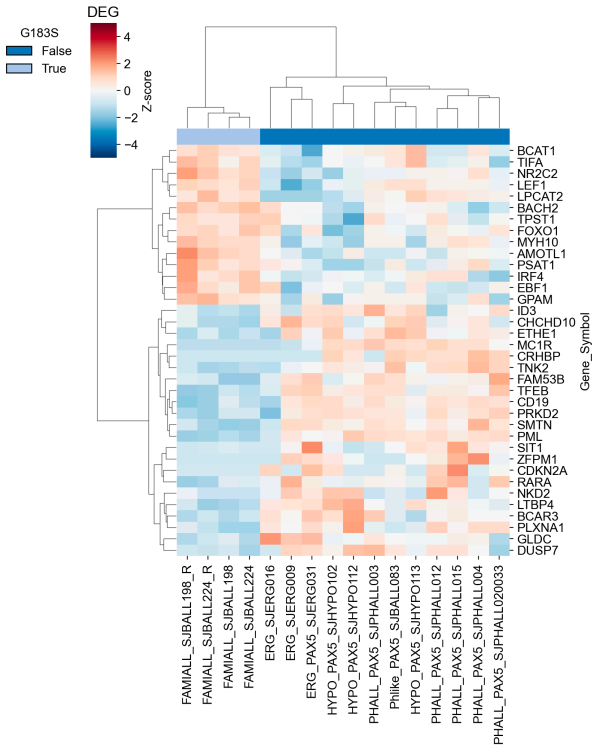


Figure. *Heatmap with the specific transcriptional program for 141 cases of sporadic childhood B-ALL with the germline G183S mutation.*

We further stratified the patients into two groups to perform two different comparisons: 1) PAX5 WT (n=10) vs PAX5 Loss (n=10), as a negative control; and 2) PAX5 G183S (n=4) vs PAX5 loss of function cases (e.g. PAX5 P80R, PAX5 Loss, n=12), for defining the G183S-specific transcriptome signature. The analyses suggest that there is a specific transcriptional program associated with the G183S mutation. This figure has been included in the revised manuscript as **Supplementary Figure 9e**.

To evaluate if this specific transcriptional program is relevant to the PAX5-NR interaction, we performed further gene set enrichment analysis. We defined the PAX5-specific, NR/3-specific, and PAX5+NR/3 commonly regulated genes with GET (**Methods**). We found an enrichment in the PAX5-loss associated differentially expressed genes (fold enrichment 1.22, p-value: 1.5e-3, hypergeometric test with Benjamini-Hochberg correction) with predicted PAX5 targets and no significant enrichment in NR/3 regulated and PAX5+NR/3 regulated genes, as expected. However, the PAX5 G183S-specific differentially expressed genes showed significant enrichment in NR/3 and PAX5+NR/3 regulated genes (fold enrichment: 1.41, 1.35; p-value: 1.05e-30, 1.57e-18, respectively). Moreover,

we found that this mutant specific differentially expressed gene set is also enriched in the genes activated in ProB cells (fold enrichment: 1.57, p-value: 2.2e-3), including well-known functionally relevant genes like CD19, EBF1, BACH2, and PML. This is now included as **Figure 6h** in the revised manuscript.

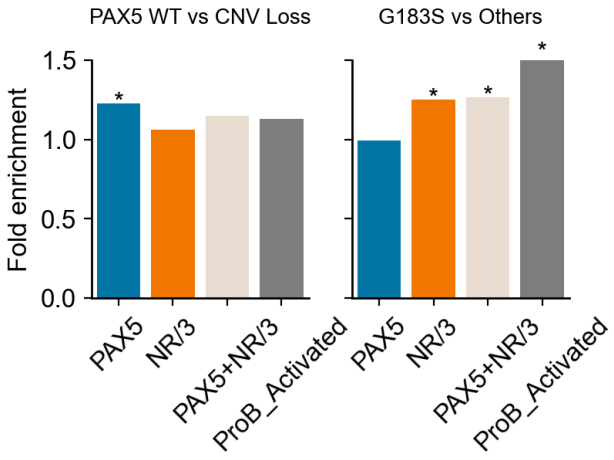


Figure. *Enrichment analysis for differentially expressed genes between PAX5 wild type v.s. PAX5 loss (left) and PAX5 G183S vs. other PAX5 alterations (CNV loss, P80R). * indicates statistical significance from a hypergeometric test (Benjamini-Hochberg adjusted p-values are reported).*

R3[4]: Prediction of quantitative chromatin accessibility seems impressive (correlation 0.98) but is hard to evaluate without context. How does this metric compare to Enformer or other models? Also, why is performance so much better for ATAC data than other data modalities?

We first note that we discovered an error in the original K562 ATAC analysis pipeline, in which a bug resulted in data leakage for evaluation on leave-out chromosome 11, leading to an unrealistic reported performance of 0.98. We have corrected this in the original manuscript. After redoing this analysis, we find finetuned GET achieves a 0.81 Pearson correlation across different leave-one-chromosome-out experiments and 0.84 on chr14, the Enformer test chromosome (Supplementary Figure 3h, left, middle, N=22).

We further evaluated performance of GET finetuned on both single cell and bulk chromatin accessibility data in comparison to Enformer's predictions for both K562 and astrocytes. We selected chromosome 14 for evaluation because it did not appear in the public Enformer checkpoint's training or validation sets. For evaluation on K562, we sum Enformer's predictions over the peak set using its K562 DNase output tracks (track IDs: 121, 122, 123, 625). For evaluation of astrocytes, we sum Enformer's predictions over

the peak set using its astrocyte DNase output tracks (track IDs: 76, 77, 78, 133). GET finetuned on astrocytes achieves a Pearson correlation of 0.71 compared to Enformer's performance of 0.69. However, GET finetuned on K562 ATAC achieves a Pearson correlation of 0.75 compared to Enformer's performance of 0.84. We note that Enformer was trained extensively on K562 functional genomics tracks, especially 461 transcription factor ChIP-seq tracks which provides more fine-grained supervision to the model, while this setting represents dataset shift for GET which must be learned during finetuning. In these experiments, GET was finetuned using LoRA parameter-efficient finetuning as described in R3[1].

For comparison, we also note that ChromBPNet[39] (a sequence-based ATAC-seq model) is reported by the authors to have 0.7 Spearman correlation on predicting ATAC-seq in a leave-out chromosome setting.

We also explored holding out randomly selected 1, 2, 3, 4, 10, and 20 motifs. For each motif, we checked whether a peak's binding score is larger than the top 20% scores in its score distribution across the genome. During the training stage, if a peak has any of the leave-out motifs passing this threshold, we set all input motif features of that peak as well as the observation accessibility TPM (aTPM) to zero. In this approach, these "knock-out" peaks do not contribute to the loss. During the evaluation stage, we calculated Pearson and Spearman correlation between predicted and observed aTPM only on these "knock-out" peaks with the original observed aTPM. For example, when there is only one leave-out motif "CTCF," we will in effect be training with about top 20\% of peaks that have low CTCF binding score on the training chromosomes, assuming evenly distributed binding sites across chromosomes. Similarly, when evaluating, we will be evaluating with 20\% of peaks with higher CTCF binding in the test chromosomes. In these experiments, we evaluated on held-out chromosome 14. In general, GET shows robust performance when leaving out 1-10 motif. The performance degrades heavily when using 20 motifs with top 20\% cutoff for each motif independently, due to removal of most of the training data (Figure below, right).

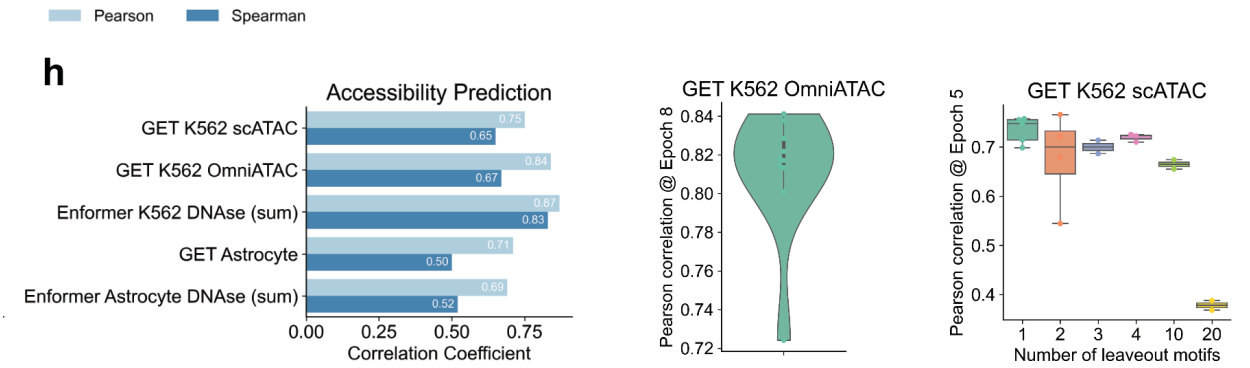


***Figure**. Performance of GET on ATAC prediction. (left) GET performance for K562 and Astrocyte ATAC compared against Enformer's predictions for DNAse on K562 and Astrocyte output tracks. (middle) GET performance for K562 bulk OmniATAC, showing leave-one-chromosome-out results for all autosomes. (right) GET performance for K562 scATAC, showing leave-out-motif analysis for randomly selected 1, 2, 3, 4, 10, or 20 motif features from the training set and evaluating on peaks with the left-out motifs (**Method: Leave-out motif evaluation**).*

R3[5]: Prediction of lentiMPRA activity with random insertion averages is a creative trick and the performance compared to Enformer for promoters is promising. Why do the authors believe the performance is worse in peak vs non-peak regions? Presumably non-peak regions are not enhancers, so what type of regulatory elements are these? Also, is Pearson correlation the best metric to capture performance?

In our manuscript, we followed the categorization of the different types of elements originally provided by Agarwal et al.[10] "Promoters" are defined as protein-coding gene promoters, "Peaks" are defined as accessible peaks called from ENCODE K562 ATAC-seq data (which contain most of the enhancers), "Heterochromatin" are defined as genomic tiles of heterochromatin regions around selected genes (e.g. GATA1), and "Control Group" are defined as synthetic elements.

Following the suggestions in R1[5] and R2[14], we have further stratified the categorization to include more nuanced information. We have now further clarified the definition of the elements by overlapping the annotated 15 ENCODE ChromHMM states computed from histone marks and other ChIP-seq data for K562. We selected the elements overlapping with states "12_EnhBiv," "6_EnhG," and "7_Enh" as enhancers; and "13_ReprPC," "14_ReprPCWk," and "15_Quies" as repressive and quiescent regions. The updated figure is shown below and is now updated in the revised manuscript as **Figure 2g**. We have also included the corresponding element definitions in the **Methods** section under **"LentiMPRA zeroshot prediction."**

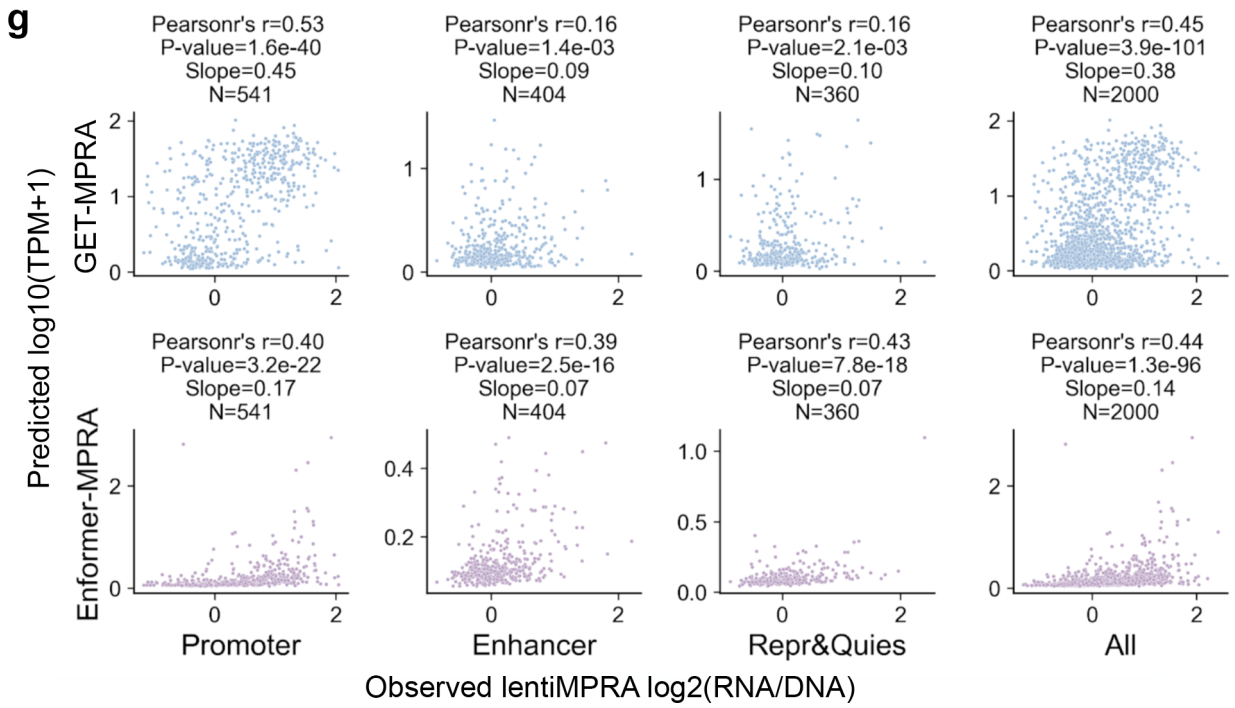
Non-peak elements contain the "Promoters," "Heterochromatin," and "Control Group" elements. We believe the Pearson correlation of GET is worse than Enformer in peak regions for the following reasons:

1. GET was finetuned using 3' scRNA-seq data, while Enformer was trained on a broad range of K562 functional genomics data, including TF binding, histone marks, and CAGE. CAGE captures not only mRNA but also other transcripts like eRNA and PROMPTs. This leads to much broader expression signal coverage over the

genome. With 3' scRNA-seq data, GET's ability to capture enhance expression is reduced. Consequently, we use ATAC-seq signals in this task to improve performance.

2. GET was trained at the peak level, which may be less sensitive to detect more nuanced and less conserved sequence patterns in generally accessible regions when compared to Enformer.

Regarding the performance metric, we have also added a slope metric to reflect the scaling of prediction versus observation, in which GET is consistently better than Enformer. We also note that this task can potentially be improved with our method finetuned on CAGE data (see R3[1]); however, since CAGE is not widely available at single cell resolution, we choose not to include this analysis.



We have updated our description of the lentiMPRA experiment in the main text and the lentiMPRA panel in **Figure 2g** of the revised manuscript.

R3[6]: Prediction of enhancer-promoter pairs looks promising, but the comparison to other methods raises questions. Why do the other methods perform so poorly (basically at random level) while ATAC by itself stands out? In the Enformer and ABC papers, for example, they report better than random performance. If the authors are using a different metric that they believe is more relevant, this should be explained and the metric used by others should also be reported. The Enformer paper (Nature Methods) Fig. 2 compares

to Gasperini et al. and Fulco et al, and achieves above random prediction of functional regulatory elements.

We thank the reviewer for the requested clarification. We have made multiple updates to this benchmark and added the referenced Gasperini et al. and Fulco et al. benchmark to ensure a fair comparison between methods.

We first note that in the Erythroblast enhancer benchmark, we used erythroid cell types consistently across methods. For the Enformer model, we utilized the "CAGE:CD34 cells differentiated to erythrocyte lineage" dataset. For the ABC score, we used the file "AllPredictions.AvgHiC.ABC0.015.minus150.ForABCPaperV3.txt.gz" downloaded from https://www.engreitzlab.org/resources and filtered for erythroblast predictions. For DeepSEA, we used the original scores produced in Cheng et al.[17], which are cell-type agnostic.

We have recomputed Enformer and ABC scores in the benchmark using the following methodology. We have also added the DNA language model HyenaDNA to this benchmark, noting the request by reviewer 2 in R2[1].

1. Enformer: We used Enformer's contribution score (gradient × input) with background normalization, following the normalization procedure described in Gschwind et al.[18] (i.e. normalization of the enhancer's score divided by mean contribution score over the entire context sequence). We found that this normalization procedure significantly improved Enformer's performance compared to the results we originally reported.
2. ABC: We computed ABC's power law score, using ATAC-seq from the same fetal erythroblast data as GET and the reference power law provided in the ABC GitHub repository (https://github.com/broadinstitute/ABC-Enhancer-Gene-Prediction/tree/main). This improves performance of the ABC model to avoid cases in which it cannot give a prediction due to mismatch between potential differences in ATAC-seq profiles, when comparing the authors' adult erythroblast dataset[40] and our fetal erythroblast dataset.
3. HyenaDNA: For HyenaDNA, we used the largest pretrained model available through Hugging Face (https://huggingface.co/LongSafari/hyenadna-large-1m-seqlen-hf, context length 1 million base pairs). To score enhancer-gene pairs, we performed in-silico mutagenesis by knocking down the enhancer element (i.e. setting each base pair in the enhancer region to the unknown nucleotide "N" in the vocabulary set) and comparing against wild-type likelihood of observing the promoter sequence.

We note that Enformer's degraded performance in the Erythroblast setting compared to the Gasperini et al. and Fulco el al. benchmarks cited in the Enformer paper is likely due

to this benchmark's use of K562, a cell type extensively trained on by Enformer, with more than 10% its output tracks dedicated to the K562 cell type. The performance discrepancy of ABC can be attributed to the lack of a cell type specific Hi-C contact map and the relative sparsity of scATAC-seq data compared to DNase-seq used in the ABC study.

In the erythroblast base-editing experiment, the overall level of fetal hemoglobin (HbF high vs. low) is used as the functional readout, which is affected by the expression of four genes (MYB, NFIX, HBG, BCL11A) on four different chromosomes whose neighboring enhancers are perturbed. The expression change of other genes in the same loci is not measured. As a result, when the gene is fixed, the strength of the enhancer-gene pair is likely more related to the strength of the enhancer itself, making ATAC a good predictor. However, in an all-against-all setting where expression change of all genes in a locus is measured, enhancer ATAC alone cannot distinguish pairs between the same enhancer and different genes.

To extend the evaluation on the enhancer-target prediction task in a more comprehensively measured setting, we also benchmarked on a recent CRISPRi dataset provided in Gschwind et al.[18] (which includes the Gasperini et al. and Fulco et al. benchmarks used in Enformer's paper). We compared GET with both pretrained Enformer and HyenaDNA. To ensure a fair comparison, GET, ABC, and Enformer benchmarked here all use the ENCODE K562 DNAse-seq data ENCFF257HEE. We called peaks using the ENCFF257HEE bam file and MACS3 with recommended parameters for ATAC-seq and the default threshold ($q < 0.05$). We took the union of the K562 peaks with the fetal atlas peak set used in GET's pretraining. For peaks that overlap between the two, we kept the fetal peak boundary. For peaks that are specific to K562, we retained the K562 peak boundary. Peaks are filtered with a minimum count of 10 in K562 to select the final K562 peak set, resulting in approximately 250,000 peaks.

Although HyenaDNA is not trained on cell-type specific genomic data, it shows above-random performance in both the Erythroblast and K562 enhancer-gene benchmark tasks, likely due to its 1 million base pair receptive field with positional embeddings. However, GET outperforms HyenaDNA in both short- and long-range predictions, demonstrating the importance of cell-type specific pretraining. GET also outperforms Enformer in every setting, even when Enformer is heavily trained on K562 functional genomics profiles. Compared to ABC, GET achieves comparable performance in the short-range setting, yet has a two-fold increase in AUPR in the K562 benchmark using the same functional genomics input.

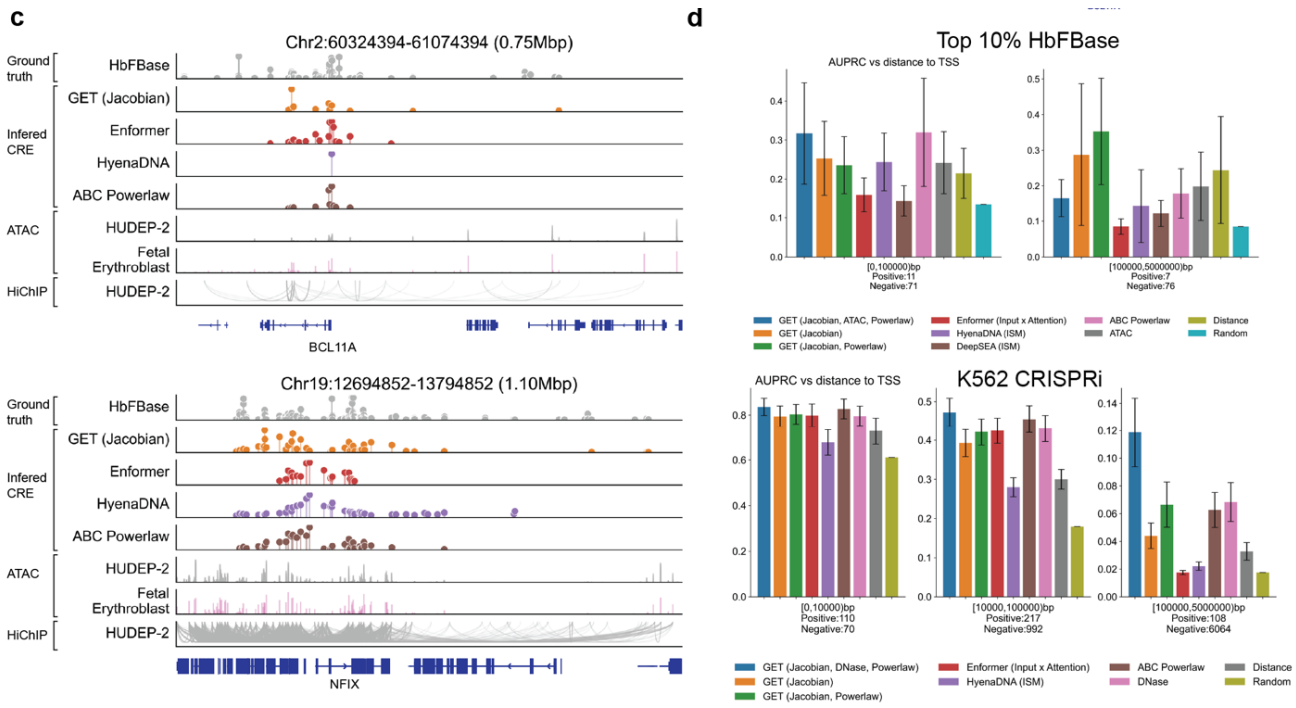


***Figure. c**) Genome tracks displaying inferred cis-regulatory elements (CREs) for BCL11A, MYB, NFIX, and HBG2 across different loci. The loci shown are Chr2:60324394-61074394 (0.75 Mbp) and Chr19:12694852-13794852 (1.10 Mbp). From top to bottom, the tracks represent: HbFBase, showing the gRNA enrichment score from base-editing experiments; GET, showing the inferred region importance score; Enformer, showing the inferred region importance score; HyenaDNA, showing in silico mutagenesis (ISM) result using pretrained HyenaDNA language model; ABC Powerlaw, showing the Activity-by-Contact prediction using fetal erythroblast ATAC and K562 Hi-C power law; ATAC-seq data from HUDEP-2, an erythroblast cell line; ATAC-seq data from fetal erythroblast cells, used in the training of GET; and HiChIP-seq data from HUDEP-2, demonstrating chromatin interactions. **d)** Benchmarking results comparing GET to other methods for predicting enhancer-promoter pairs, including analysis of distal (>100 kb) interactions. (top) Erythroblast fetal hemoglobin regulating enhancer prediction. (bottom) K562 CRISPRi enhancer target prediction. Area under precision-recall curve (AUPRC) is shown. Ablation of different GET prediction components (Jacobian, DNase, Powerlaw; see **Methods: Enhancer-gene pair prediction**) is also shown in the plot.*

The updated panels appear as **Figure 3c,d** in the revised manuscript.

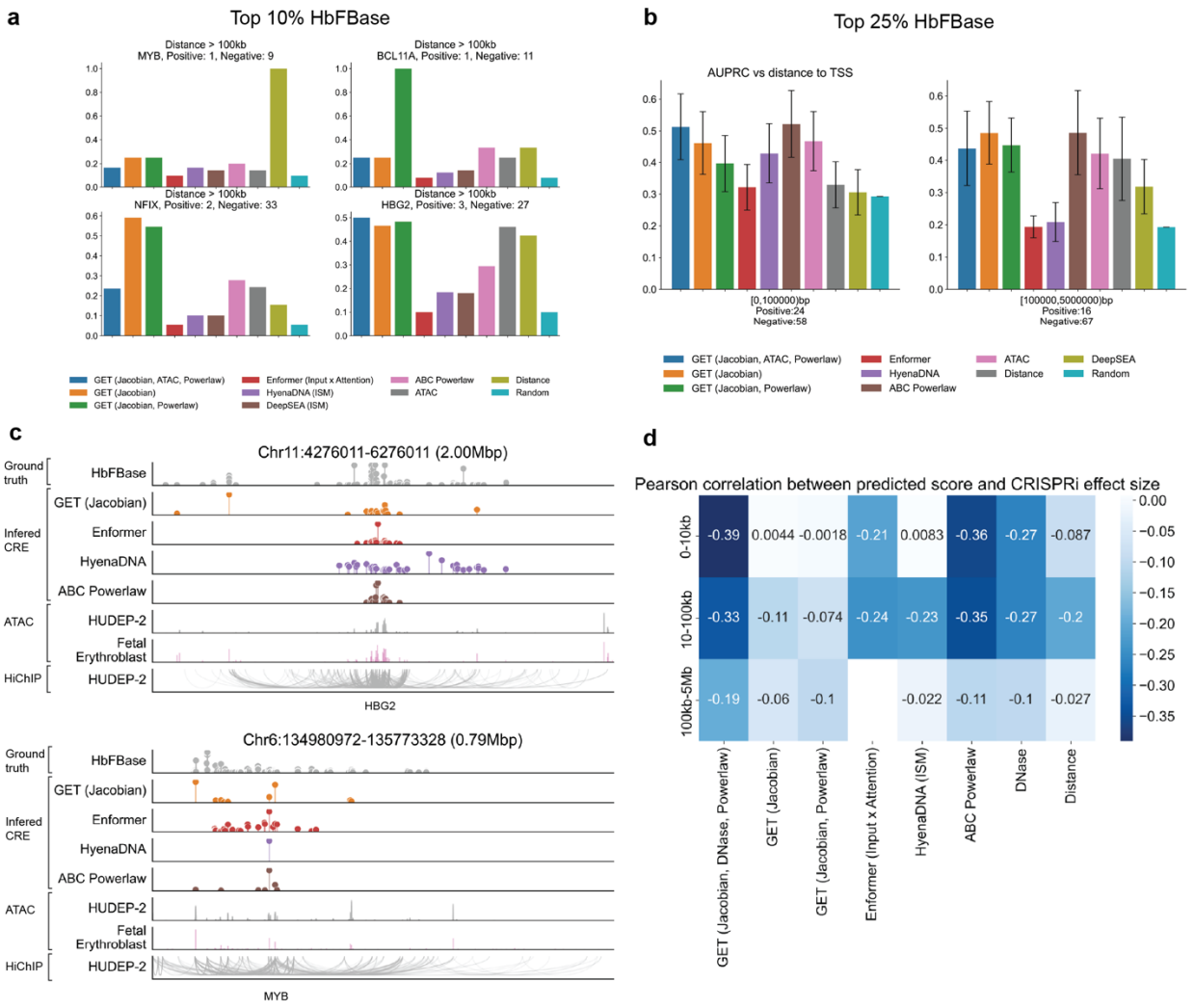


Figure. ***a)*** *Per-gene AUPRC benchmark of cis-regulatory region prediction with top 10% HbFbase as the label cutoff.* ***b)*** *Distance stratified AUPRC benchmark using top 25% HbFbase as the label cutoff.* ***c)*** *Cis-regulatory region prediction of HBG2 and MYB loci.* ***d)*** *Distance stratified Pearson correlation between predicted scores and K562 CRISPRi effect size.*

The updated panels appear as **Supplementary Figure 4a-d** in the revised manuscript.

R3[7]: Enrichment of physically interacting TF pairs is interesting. To understand the result, it will be important to know the contribution of motif co-localization as a baseline. For example, are two TFs predicted to interact because their motifs tend to fall in the same regions? Or, is the model adding some other information to improve upon this baseline?

Following the suggestions of Reviewer 2 (R2[10]) and Reviewer 3[7], we have incorporated ChIP-seq co-localization (using 677 TF ChIP-seq in HepG2 cell line from ChIP-Atlas) as a ground truth comparison and accessible motif co-localization as a baseline, computed either across all cell types or in Hepatocytes for a comparison with HepG2 ChIP-seq colocalization. Additionally, we have switched from Precision to macro F1 as the evaluation metric to avoid artificial inflation of the metric by repeated positive predictions, suggested in R2[22].

The updated results are presented in **Figure 4h** of the revised manuscript, which compares various transcription factor interaction prediction methods across different recall thresholds. The plot shows the macro F1 score of different approaches, including ground truth from affinity purification mass spectrometry in Göös et al.[31], ChIP-seq co-localization from the HepG2 cell line (ENCODE, ChIP-Atlas), motif-motif interactions derived from the GET model using the causal discovery method LiNGAM (https://github.com/cdt15/lingam), correlation analysis of the GET Jacobian matrix, and accessible motif co-localization baselines performed either across all cell types or specifically using hepatocytes corresponding to HepG2.

ChIP-seq co-localization provides a 5% recall at a 0.14 macro F1 score, slightly lower than the GET Causal approach at the same recall. The GET causal discovery method consistently outperforms the motif co-localization baselines across all specificity levels, indicating that the model has learned useful information from the interplay between ATAC-seq and RNA-seq across various cell types.

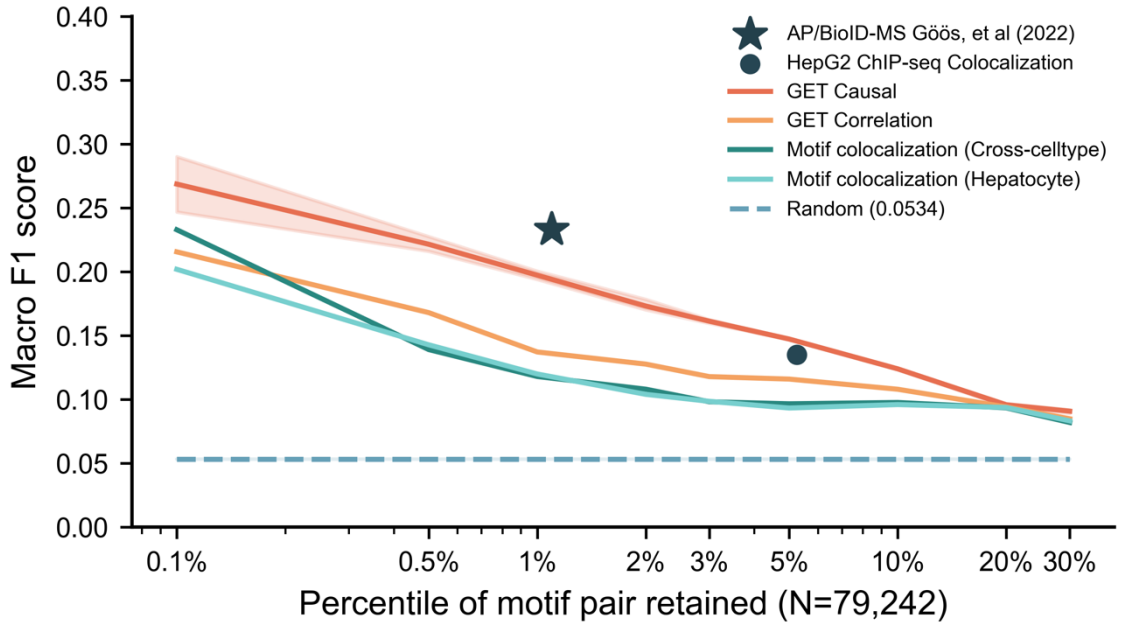


Figure. *Comparison of transcription factor interaction prediction methods across different recall thresholds. The plot shows the precision (macro F1 score) of various approaches, including ground truth from AP-MS (Göös et al., star), ChIP-seq co-localization from the HepG2 cell line (ENCODE, ChIP-Atlas, dot), motif-motif interactions from the GET model derived using five independent runs of LiNGAM ("GET Causal"), correlation analysis of GET's Jacobian matrix ("GET Correlation"), and accessible motif co-localization baselines performed either across all cell types ("Motif colocalization (Cross-celltype)") or specifically using hepatocytes ("Motif colocalization (Hepatocyte)"). The x-axis represents the percentile of motif pairs retained based on their interaction scores (recall), with lower percentiles indicating more stringent thresholds. The random baseline (dotted line) represents the expected performance of random guessing. ChIP-seq co-localization provides a 5% recall at 0.14 macro F1 score, slightly lower than the GET Causal approach at the same recall. GET Causal consistently outperforms the motif co-localization baselines across all specificity levels.*

Supplementary Notes. The authors of ChIP-Atlas first stratify the binding peak for each TF ChIP-seq data into 3 tiers (high, mid, low; see https://github.com/inutano/chip-atlas/wiki#colocalization_doc). Then for a pair of transcription factors P1 and P2, they check colocalization between every tier pair and assign scores with a preference for high-high colocalization. If the strong binding peaks of P1 overlap with strong binding peaks of P2, the P1-P2 interaction is more robust than the case in which the P1 strong binding sites only overlap with the P2 weak binding sites. We presented the colocalization stronger than mid-mid interactions (score>=4) in **Figure 4h** as those present the more reliable interactions. A stronger cutoff (score>=9, keeping only high-high interactions) reduced its performance to 0.097 Macro F1 score at 2% recall.

Other points

R3[8]: Please reference all figure panels

We have ensured that all the figure panels are now referenced in the text.

R3[9]: Please describe all work thoroughly; the logic connecting points and sections is often tenuous.

We thank the reviewer for raising this concern. We have rewritten the **Methods** section to include more details and improved the main manuscript.

R3[10]: Missing description of statistical test in Figure 2d.

We performed a two-sided Mann-Whitney U-test and updated the figure legend (**Figure 2f**) to include the following information:

Two sided Mann-Whitney U-test p-value annotation legend: ***: 1.00e-04 < p <= 1.00e-03; ****: p <= 1.00e-04. The detailed p-values are Promoter vs. Peak: p < 1e-237; Peak vs. Heterochromatin: p < 1e-237; Heterochromatin vs. Control: p = 4.049e-04.

R3[11]: Incorrect figure references (“Figure 2” noted on page 4 when referring to Figure 1c)

We thank the reviewer for catching this. The typo has been corrected.

R3[12]: Fig 4h is not referenced in the text (there were several other instances of a similar lack of reference).

We have ensured that all figures and panels are now referenced in the text.

R3[13]: Figure 5 legends do not match data (e refers to c, g does not refer to co-IP, which has no legend).

We have corrected **Figure 5** legends to ensure coherence.

R3[14]: Methods section typos
o Methods section “Input data” is repeated
o “Multimer structure prediction” end of text is cut off

We have updated the **Methods** section **“Input data”** and **“Multimer structure prediction”** to correct these typos.

R3[15]: Text needs an additional proofread both for grammar and scientific accuracy.

We thank the reviewer for this suggestion. We have proofread the manuscript to improve the writing and scientific accuracy.

R3[16]: How does source of training data (multi ome or individual ome) impact results?

We thank the reviewer for this question. The impact of multiome vs. individual-ome varies with the sample type and research focus. As shown in the TF perturbed hESC example (see R3[1]), the expression correlation between perturbed hESC clusters can be as high as 0.98, which is higher than the prediction accuracy of GET. In such cases, using two very similar pseudobulk datasets for pretraining is probably not very meaningful as most of the genomic loci would look the same. However, during the finetuning stage, for a study focused on perturbation or cell states, multiome is required to ensure the model can learn from the data as accurately as possible. For the more general training of transcriptional rules across major cell types, we believe individual-omes are crucial to reduce data scarcity, as we recommend ensuring at least 10 million reads per cell type or 500 to 1,000 cells per cell type for robust modeling. In these cases, higher quantities greatly improve model performance, which presents a cost challenge for the qualified multiome data. We have incorporated this information into the **Method: Dataset considerations** section as a general guide to the user.

References

1. Okuyama, K. *et al.* PAX5 is part of a functional transcription factor network targeted in lymphoid leukemia. *PLoS Genet* **15**, e1008280 (2019).
2. Shah, S. *et al.* A recurrent germline PAX5 mutation confers susceptibility to pre-B cell acute lymphoblastic leukemia. *Nat Genet* **45**, 1226–1231 (2013).
3. Yanchus, C. *et al.* A noncoding single-nucleotide polymorphism at 8q24 drives *IDH1* -mutant glioma formation. *Science* **378**, 68–78 (2022).
4. Killela, P. J. *et al. TERT* promoter mutations occur frequently in gliomas and a subset of tumors derived from cells with low rates of self-renewal. *Proc. Natl. Acad. Sci. U.S.A.* **110**, 6021–6026 (2013).
5. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203 (2021).
6. Joung, J. *et al.* A transcription factor atlas of directed differentiation. *Cell* **186**, 209-229.e26 (2023).
7. Terekhanova, N. V. *et al.* Epigenetic regulation during cancer transitions across 11 tumour types. *Nature* **623**, 432–441 (2023).
8. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
9. Meng, Q. *et al.* The full set of potential open regions (PORs) in the human genome defined by consensus peaks of ATAC-seq data. 2023.05.30.542889 Preprint at https://doi.org/10.1101/2023.05.30.542889 (2023).

10. Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. Preprint at https://doi.org/10.1101/2023.03.05.531189 (2023).
11. Kim, S. *et al.* DNA-guided transcription factor cooperativity shapes face and limb mesenchyme. *Cell* **187**, 692-711.e26 (2024).
12. Oshima, K. *et al.* Mutational and functional genetics mapping of chemotherapy resistance mechanisms in relapsed acute lymphoblastic leukemia. *Nat Cancer* **1**, 1113–1127 (2020).
13. Windahl, S. H. *et al.* The nuclear-receptor interacting protein (RIP) 140 binds to the human glucocorticoid receptor and modulates hormone-dependent transactivation. *The Journal of Steroid Biochemistry and Molecular Biology* **71**, 93–102 (1999).
14. 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).
15. Dudnyk, K., Cai, D., Shi, C., Xu, J. & Zhou, J. Sequence basis of transcription initiation in the human genome. *Science* **384**, eadj0116 (2024).
16. Nguyen, E. *et al.* HyenaDNA: Long-Range Genomic Sequence Modeling at Single Nucleotide Resolution. Preprint at https://doi.org/10.48550/arXiv.2306.15794 (2023).
17. Cheng, L. *et al.* Single-nucleotide-level mapping of DNA regulatory elements that control fetal hemoglobin expression. *Nat Genet* **53**, 869–880 (2021).
18. Gschwind, A. R. *et al.* An encyclopedia of enhancer-gene regulatory interactions in the human genome. 2023.11.09.563812 Preprint at https://doi.org/10.1101/2023.11.09.563812 (2023).

19. Martyn, G. E. *et al.* Rewriting regulatory DNA to dissect and reprogram gene expression. 2023.12.20.572268 Preprint at https://doi.org/10.1101/2023.12.20.572268 (2023).
20. Hu, E. J. *et al.* LoRA: Low-Rank Adaptation of Large Language Models. Preprint at http://arxiv.org/abs/2106.09685 (2021).
21. Domcke, S. *et al.* A human cell atlas of fetal chromatin accessibility. *Science* **370**, (2020).
22. Cao, J. *et al.* A human cell atlas of fetal gene expression. *Science* **370**, (2020).
23. Zhang, K. *et al.* A single-cell atlas of chromatin accessibility in the human genome. *Cell* **184**, 5985-6001.e19 (2021).
24. Vierstra, J. *et al.* Global reference mapping of human transcription factor footprints. *Nature* **583**, 729–736 (2020).
25. Abramson, J. *et al.* Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature* **630**, 493–500 (2024).
26. Zou, Z., Ohta, T. & Oki, S. ChIP-Atlas 3.0: a data-mining suite to explore chromosome architecture together with large-scale regulome data. *Nucleic Acids Research* gkae358 (2024) doi:10.1093/nar/gkae358.
27. Partridge, E. C. *et al.* Occupancy maps of 208 chromatin-associated proteins in one human cell type. *Nature* **583**, 720–728 (2020).
28. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
29. Hitz, B. C. *et al.* The ENCODE Uniform Analysis Pipelines.

30. Barrett, T. *et al.* NCBI GEO: archive for functional genomics data sets--update. *Nucleic Acids Res* **41**, D991-995 (2013).
31. Göös, H. *et al.* Human transcription factor protein interaction networks. *Nat Commun* **13**, 766 (2022).
32. Wang, H. *et al.* RXRα Inhibits the NRF2-ARE Signaling Pathway through a Direct Interaction with the Neh7 Domain of NRF2. *Cancer Research* **73**, 3097–3108 (2013).
33. Szklarczyk, D. *et al.* The STRING database in 2021: customizable protein–protein networks, and functional characterization of user-uploaded gene/measurement sets. *Nucleic Acids Research* **49**, D605–D612 (2021).
34. Mannens, C. C. A. *et al.* Chromatin accessibility during human first-trimester neurodevelopment. *Nature* 1–8 (2024) doi:10.1038/s41586-024-07234-1.
35. Agarwal, V. *et al.* Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. 2023.03.05.531189 Preprint at https://doi.org/10.1101/2023.03.05.531189 (2023).
36. Tropberger, P. *et al.* Regulation of Transcription through Acetylation of H3K122 on the Lateral Surface of the Histone Octamer. *Cell* **152**, 859–872 (2013).
37. Song, M. *et al.* Sequence-specific prediction of the efficiencies of adenine and cytosine base editors. *Nat Biotechnol* **38**, 1037–1043 (2020).
38. Li, J. *et al.* Conservation and divergence of vulnerability and responses to stressors between human and mouse astrocytes. *Nat Commun* **12**, 3958 (2021).

39. Brennan, K. J. *et al.* Chromatin accessibility is a two-tier process regulated by transcription factor pioneering and enhancer activation. Preprint at https://doi.org/10.1101/2022.12.20.520743 (2022).
40. Corces, M. R. *et al.* Lineage-specific and single-cell chromatin accessibility charts human hematopoiesis and leukemia evolution. *Nature Genetics* **48**, 1193–1203 (2016).

COLUMBIA UNIVERSITY MEDICAL CENTER

RAUL RABADAN
PROFESSOR
PROGRAM FOR MATHEMATICAL GENOMICS
DEPARTMENT OF SYSTEMS BIOLOGY

622 West 168th Street, PH18-200
New York, NY 10032

(212) 305-3896
rr2579@cumc.columbia.edu

Referee expertise:

Referee #1: scRNA-seq/developmental biology

Referee #2: network biology/transfer learning

Referee #3: systems/computation bio

Referees' comments:

Referee #1 (Remarks to the Author):

The authors have addressed my previous concerns in a very comprehensive manner.

Referee #2 (Remarks to the Author):

The authors have performed additional experiments to address the concerns raised by reviewers. They have broadened their analyses to raise confidence in the generalizability of their approach and performed experimental validation of predicted interactions. The authors have sufficiently addressed the prior concerns.

Referee #3 (Remarks to the Author):

Our major concerns related to benchmarking GET against previous models and various analysis choices were largely addressed.
• The expansion of model performance to many “left out” cell types is comprehensive and addresses issues from the first submission.
• The comparison to Enformer is now clearer in the ATAC and lentiMPRA prediction tasks.
• The updated comparison to Enformer and ABC models for enhancer-promoter pair prediction seems to have addressed concerns about improper treatment of alternative

COLUMBIA UNIVERSITY MEDICAL CENTER

RAUL RABADAN
PROFESSOR
PROGRAM FOR MATHEMATICAL GENOMICS
DEPARTMENT OF SYSTEMS BIOLOGY

622 West 168th Street, PH18-200
New York, NY 10032

(212) 305-3896
rr2579@cumc.columbia.edu

models leading to artificially poor performance.
• Inclusion of motif co-localization provides context for TF functional pair prediction.

The revision, however, raises several minor points and issues that will be important to address:
• The TFAP2A co-IP in Figure 5 should be labeled so it is clear that TFAP2A is the IP target and negative controls (such as beta actin and another TF that is not predicted to interact with TFAP2A) should be included.

We thank the reviewer for this suggestion. We have now included beta actin and another TF that is not predicted to interact with TFAP2A, SRF, as negative controls. SRF was chosen because it has near-zero effect size from LiNGAM, a good antibody available, and a relatively small motif cluster size, in contrast to e.g. homeodomain TFs which have a large amount of TFs in one motif cluster.

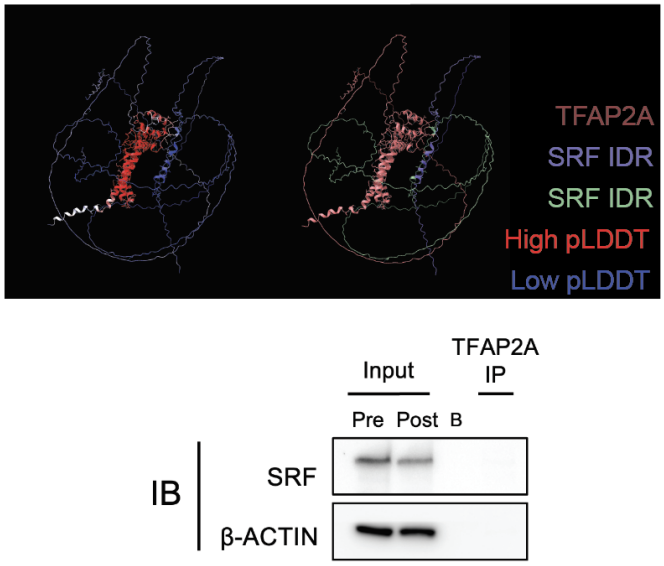


Figure legend:

Upper panel show AF3-based structure prediction between SRF IDR with its DNA-binding domain (133-230aa) removed and full length TFAP2A, demonstrating no clear interactions between these two protein and low prediction confidence for SRF IDR, indicating its highly disordered structure. Color represent pLDDT (red-high/blue-low, left) or chains (pink/purple/green, right). Lower micrographs show immunoblot detection of SRF, and β-ACTIN from HeLa cell lysates subjected to co-immunoprecipitation using a TFAP2A antibody, β-ACTIN serves as a loading control. SRF and β-ACTIN both show no interactions with TFAP2A as determined by co-immunoprecipitation. Abbreviation:

COLUMBIA UNIVERSITY MEDICAL CENTER

RAUL RABADAN
PROFESSOR
PROGRAM FOR MATHEMATICAL GENOMICS
DEPARTMENT OF SYSTEMS BIOLOGY

622 West 168th Street, PH18-200
New York, NY 10032

(212) 305-3896
rr2579@cumc.columbia.edu

Pre, HeLa cell lysate prior to bead incubation; Post, HeLa cell lysate after bead incubation; IP, immunoprecipitated proteins. B, empty lane.

We have updated **Figure 5g** to include these changes and updated the corresponding main text as below.

Main:

*"To validate the predicted interactions between these two proteins, we performed co-immunoprecipitation experiments. As shown in **Figure 5g**, we were able to pull down ZFX using a TFAP2A antibody, while a negative control TF, SRF, predicted to be not interacting with TFAP2A, show no interaction."*

Method:

Cell Culture

HeLa cells were purchased from ATCC (CCL-2). HeLa cells were cultured in DMEM (Gibco, 11965) supplemented with 10% defined FBS (HyClone, SH30070), at 37°C/ 5% CO_2.

TFAP2A co-immunoprecipitation

HeLa cell protein lysates were generated with 0.5% NP-40 lysis buffer containing 50mM Tris-HCl, 150mM NaCl and a phosphatase-protease inhibitor cocktail (Sigma-Aldrich, PPC1010). 500 µg of protein lysate was pre-cleared with Protein A agarose beads (Cell Signaling Technology, 9863) and incubated with 5 µg of agarose-conjugated TFAP2A antibody (Santa Cruz Biotechnology, sc-12726 AC) overnight at 4°C. Beads were washed 3 times in lysis buffer and proteins were eluted in Laemmli sample buffer (BioRad, 1610737). Proteins were detected using primary antibodies against TFAP2A (ABclonal, A2294), ZFX (ThermoFisher, PA5-34376), SRF (Abclonal, A16718) and β-ACTIN (Santa Cruz Biotechnology, sc-47778 or Cell Signaling Technology, 4967) followed by chemiluminescence imaging.

• The suggestion that SNAI1 and RELA interact due to a mutual interaction with EP300 is interesting but raises the question of specificity (as many TFs are thought to interact with EP300). The authors should provide an analysis asking whether the TFs that they predict to interact are enriched for EP300 interaction (this could be done using ChIP-seq for the TFs and EP300 and potentially AlphaFold).

We thank the reviewer for this suggestion. We have performed further characterization of the GET Causal prediction and identified enrichment in EP300 interaction. We started

COLUMBIA UNIVERSITY MEDICAL CENTER

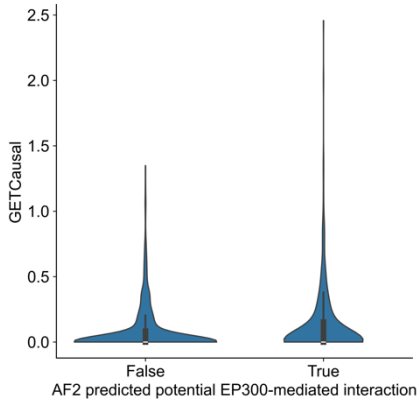
RAUL RABADAN
PROFESSOR
PROGRAM FOR MATHEMATICAL GENOMICS
DEPARTMENT OF SYSTEMS BIOLOGY

622 West 168th Street, PH18-200
New York, NY 10032

(212) 305-3896
rr2579@cumc.columbia.edu

with the across cell type interaction prediction used in the ChIP-seq benchmark. We defined a motif-motif interaction to be potentially EP300-mediated if any TF in each motif cluster has a strong ChIP-seq defined colocalization (ChIP-Atlas HepG2 colocalization score of 9, as used in **Figure 4g**). Focusing on only motif-motif interactions that have a EP300-based annotation (n=15,252), we found that those with a top 5% interaction (among all motif-motif pairs, n=763) have an enrichment in potentially EP300-mediated interaction (n=61 among 930 potentially EP300-mediated pairs), with a Fisher exact test P-value of 0.03. This enrichment is even stronger if we further consider STRING-based physical interaction between TFs and EP300 to complement the ChIP-seq defined pairs, leading to a P-value of 6.53e-07 (n=117 among 1490 pairs).

Due to the computational complexity of AlphaFold multimer prediction, we focused on a group of predicted interacting TFs. More specifically, for each GET Causal interaction identified across all cell types (with 95 percentile absolute effect size in each cell type), we map the motif to the corresponding highest expressed TF and convert the interaction to TF-TF interaction. Then we selected all TF-TF interactions that appear in more than 5 cell types, resulting in interactions spanning 188 TFs. For each of these TFs we used AlphaFold2 to model their interactions with EP300 in an all-against-all manner across all pLDDT defined protein domains. Then for each TF, we defined the AlphaFold-based EP300 interaction score as the maximum interface-pTM (ipTM) across all domain-domain multimer predictions with at least 50 prediction confidence (pLDDT). Accordingly, we defined the TF-TF AlphaFold-based EP300 interaction score by multiplying their corresponding scores, and these scores were mapped back to motif-motif interaction for comparison with GET Causal effect size. When considering all pairs of motif with a defined AlphaFold-based score, we found that the pairs with higher ipTM tends to have higher GET Causal effect size (p-value 9.73e-3, Mann-Whitney U test, see figure to the right; Fisher exact test p-value 0.011 with the same GET Causal cut off as above). This analysis has now been incorporated in **Method: Characterization of EP300-mediated Interactions in GET Causal Predictions** due to space constraint.



• The BioID experiment is described in very inconsistent and confusing ways in the main text and methods. The authors repeatedly refer to it as a co-immunoprecipitation experiment using anti-PAX5 and anti-HA antibodies, which is not correct. The lysates were enriched with streptavidin-linked beads (not antibody-linked beads as in an immunoprecipitation) and the enriched sample should not be labeled “IP.” Labeling those samples as “enrichment” or “streptavidin enrichment” would accurately reflect the

COLUMBIA UNIVERSITY MEDICAL CENTER

RAUL RABADAN
PROFESSOR
PROGRAM FOR MATHEMATICAL GENOMICS
DEPARTMENT OF SYSTEMS BIOLOGY

622 West 168th Street, PH18-200
New York, NY 10032

(212) 305-3896
rr2579@cumc.columbia.edu

experiment and avoid confusing readers. Also, the authors refer to the experiment as "proximity ligation assay" which is incorrect (this name refers to a very specific alternative assay using DNA-conjugated antibodies commonly implemented using the Sigma Duolink kit). There is a methods typo referring to it as "PLS" as well. The general type of assay the author's preformed is called "proximity labeling." Also, it appears the data using streptavidin-HRP, anti-PAX5, or anti-HA are not shown although those reagents are listed in the methods. Antibody information for anti-NR3C1 is not shared.

We thank the reviewer for raising these points. We have now streamlined the description in the revised manuscript. Altered parts are listed below:

Methods

Proximity labeling assay to detect PAX5-NR2C2 interactions
We initially cloned PAX5 wild type (WT) and PAX5 G183S mutant into the pCDNA3.1-MCS-13Xlinker-BioID2-HA (Addgene #80899)[28]. After verification, we subcloned PAX5-WT-13Xlinker-BioID2-HA and PAX5-G183S-13Xlinker-BioID2-HA into the pCDH-GFP-puro vector (System bioscience #CD513B-1). We transduced the REH B-ALL cell line (ATCC #CRL-8286) with pCDH- PAX5-WT-13Xlinker-BioID2-HA-GFP and with pCDH-PAX5-G183S-13Xlinker-BioID2-HA-GFP and selected transduced cells with puromycin (1ug mL-1) to generate stable cell lines. Proximity labeling assay was performed following previously published methods[28–30]. …Biotinylated proteins in total protein extracts or streptavidin enriched extracts were detected by western blot using standard protocols and the following antibodies: streptavidin-HRP antibody (Life Technologies #S911), anti-PAX5 (Cell Signaling #8970), anti-HA (Cell Signaling #3724), anti-NR2C2 (Cell Signaling #31646), anti-NCOR1 (Cell Signaling #5948), NRIP1-HRP (Santa Cruz Biotechnology #sc-518071) and NR3C1 (Cell Signaling #12041). Proteins were detected using a Li-Cor Odyssey OFC instrument and quantified using the GelAnalyzer 23.1software.

Main text

We performed proximity labeling assay using PAX5 wild type and G183S mutant linked to a biotin ligase in the B-ALL REH cell line. Analysis of proteins biotinylated by their proximity to the PAX5-BioID fusions identified the nuclear corepressor NCOR1, a previously described PAX5 interactor **68**,**69**, NR2C2 NR/3 motif-containing nuclear corepressor, and the NRIP1 nuclear receptor interacting protein. The experiment shows a clear interaction between PAX5 and NR2C2, aligned with the previous report, while NR4A1 shows no interaction with PAX5. NR3C1, a NR/20 motif TF predicted to be interacting with NR/3 but not PAX/1 by GET, shows no interaction with PAX5. Interestingly, the presence of the G183S mutation results in an increased interaction between PAX5 and NR2C2 (**Figure 6d-e, Supplementary Figure 10)**.

COLUMBIA UNIVERSITY MEDICAL CENTER

RAUL RABADAN
PROFESSOR
PROGRAM FOR MATHEMATICAL GENOMICS
DEPARTMENT OF SYSTEMS BIOLOGY

622 West 168th Street, PH18-200
New York, NY 10032

(212) 305-3896
rr2579@cumc.columbia.edu

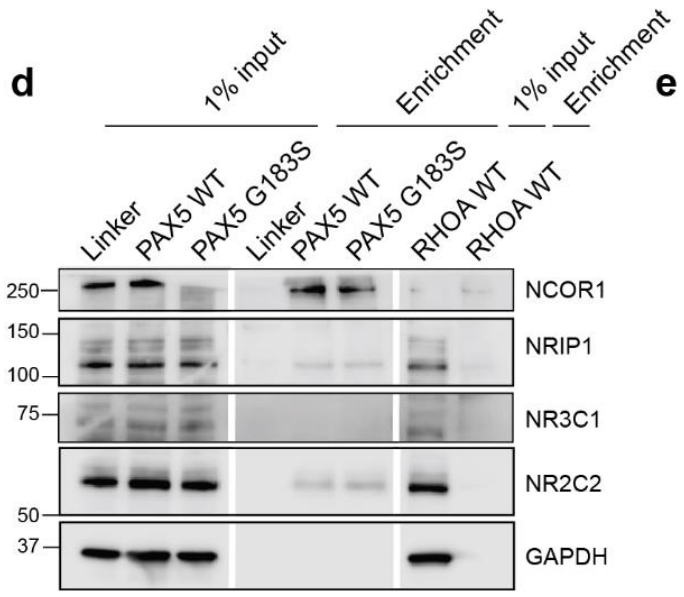


Figure 6d. Detection of NCOR1, NRIP1, NR3C1, and NR2C2 PAX5 interacting proteins in Streptavidin-enriched eluates from PAX5 WT, PAX5 G183S and RHOA WT-BioID expressing B-ALL REH cell line and in total protein lysates (1% input) using proximity labeling Assays.

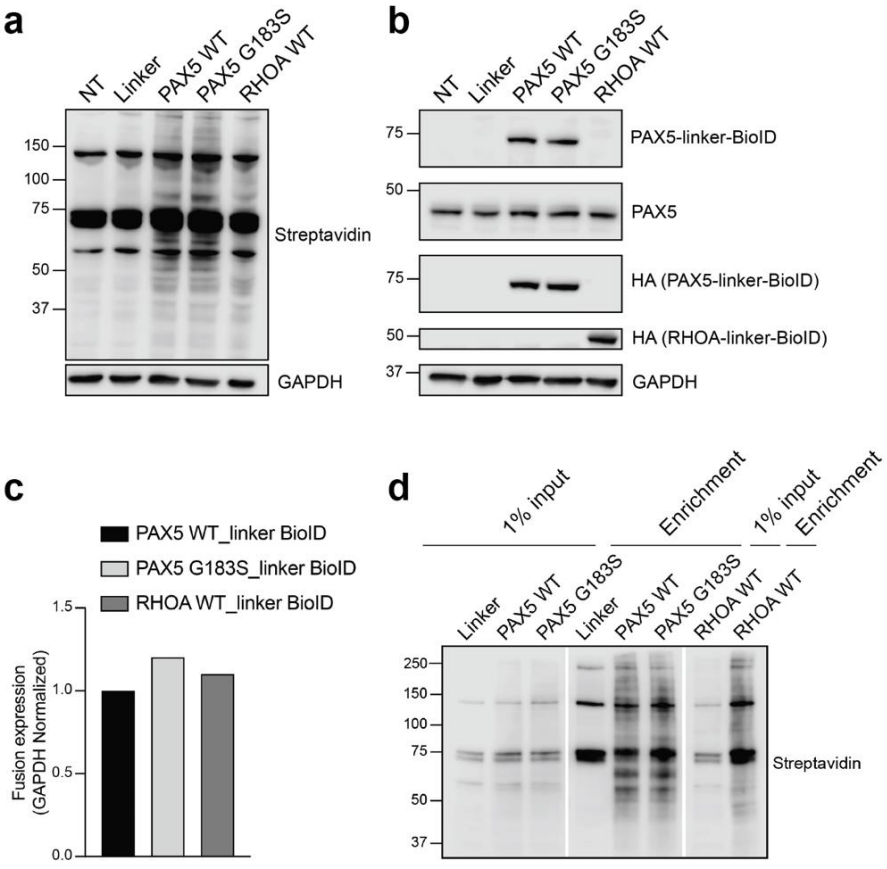
Supplementary Figure 10. Identification of PAX5 interacting proteins by Protein Ligation Assays in B-ALL REH cell line. a. Detection of biotinylated proteins in REH cell line transduced with PAX5 WT-linker-BioID (PAX5 WT) and PAX5 G183S-linker-BioID (PAX5 G183S) fusion proteins by western blot analysis using streptavidin-HRP. Expression of RHOA WT-linker-BioID is used as specificity control (RHOA WT). b. Detection of PAX5 and RHOA BioID-fusion proteins in REH cells by western blot analysis using anti-PAX5 and anti-HA antibodies. c. Fusion protein quantification from (b) d. Detection of biotinylated proteins in whole cell lysates (input; 1% of total protein lysate) and Streptavidin-enriched eluates (Enrichment) by streptavidin-HRP staining.

COLUMBIA UNIVERSITY MEDICAL CENTER

RAUL RABADAN
PROFESSOR
PROGRAM FOR MATHEMATICAL GENOMICS
DEPARTMENT OF SYSTEMS BIOLOGY

622 West 168th Street, PH18-200
New York, NY 10032

(212) 305-3896
rr2579@cumc.columbia.edu



• We would also suggest another edit pass to improve readability

Thanks. We have performed additional edit pass to improve the readability and fixed inconsistent parts.

• We also find the lack of code availability during review problematic

Referee #3 (Remarks on code availability):

Code is not available (authors say it will be made available upon publication) so we cannot review.
We also get an error when we use the link above https://github.com/fuxialexander/get_model

COLUMBIA UNIVERSITY
MEDICAL CENTER

RAUL RABADAN
PROFESSOR
PROGRAM FOR MATHEMATICAL GENOMICS
DEPARTMENT OF SYSTEMS BIOLOGY

622 West 168th Street, PH18-200
New York, NY 10032

(212) 305-3896
rr2579@cumc.columbia.edu

Thank the reviewer for raising this up. The code has now been made available.