DNA-DIFFUSION: LEVERAGING GENERATIVE MODELS FOR CONTROLLING CHROMATIN ACCESSIBILITY AND GENE EXPRESSION VIA SYNTHETIC REGULATORY ELEMENTS

Lucas Ferreira DaSilva[1,2,†], Simon Senan[2,3,†], Zain Munir Patel[1,2,3,††], Judith F Kribelbauer[4,5,††], Aniketh Janardhan Reddy[6], Sameer Gabbita[7], Jonathan D. Rosen[8], Zach Nussbaum[9], César Miguel Valdez Córdova[10], Aaron Wenteler[11], Noah Weber[12], Tin M. Tunjic[12], Martino Mansoldo[13,**], Talha Ahmad Khan[13], Zelun Li[14,15], Cameron Ray Smith[1,2,3], Matei Bejan[16], Lithin Karmel Louis[14,15], Paola Cornejo[14,15], Will Connell[13], Bart Deplanke[4,5], Michael I. Love[8], Emily S. Wong[14,15], Wouter Meuleman[17,18], Luca Pinello[1,2,3,‡]

[1]Department of Pathology, Harvard Medical School, Boston, MA, USA
[2]Molecular Pathology Unit, Center for Cancer Research, Massachusetts General Hospital, Boston, MA, USA
[3]Broad Institute of Harvard and MIT, Cambridge, MA, USA
[4]Laboratory of Systems Biology and Genetics, Institute of Bioengineering, School of Life Sciences, École Polytechnique Fédérale de Lausanne (EPFL), Lausanne, Switzerland
[5]Swiss Institute of Bioinformatics (SIB), Lausanne, Switzerland
[6]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA
[7]Johns Hopkins University, Baltimore, MD, USA
[8]Department of Genetics, Department of Biostatistics, University of North Carolina at Chapel Hill, Chapel Hill, NC, USA
[9]Nomic AI
[10]Johannes Kepler University, Linz, Austria
[11]Queen Mary University of London, London, UK
[12]TU Vienna, Austria
[13]Independent Researcher
[14]Victor Chang Cardiac Institute, Darlinghurst, New South Wales, Australia
[15]School of Biotechnology and Biomolecular Sciences, Faculty of Science, UNSW Sydney, Sydney, Australia
[16]University of Bucharest
[17]Altius Institute for Biomedical Sciences, Seattle, WA, USA
[18]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
[†]These authors contributed equally
[††]These authors contributed equally
[**]Employed by GSK at time of publication. All work was completed prior to starting at GSK
[‡]Corresponding author: lpinello@mgh.harvard.edu (L.P.)

ABSTRACT

The challenge of systematically modifying and optimizing regulatory elements for precise gene expression control is central to modern genomics and synthetic biology. Advancements in generative AI have paved the way for designing synthetic sequences with the aim of safely and accurately modulating gene expression. We leverage diffusion models to design context-specific DNA regulatory sequences, which hold significant potential toward enabling novel therapeutic applications requiring precise modulation of gene expression. Our framework uses a cell type-specific diffusion model to generate synthetic 200 bp DNA regulatory elements based on chromatin accessibility across different cell types. We evaluate the generated sequences based on key metrics to ensure they retain properties of

endogenous sequences: transcription factor binding site composition, potential for cell type-specific chromatin accessibility, and capacity for sequences generated by DNA diffusion to activate gene expression in different cell contexts using state-of-the-art prediction models. Our results demonstrate the ability to robustly generate DNA sequences with cell type-specific regulatory potential. DNA-Diffusion paves the way for revolutionizing a regulatory modulation approach to mammalian synthetic biology and precision gene therapy.

1 INTRODUCTION

Gene regulation is a complex process orchestrated at different levels: genetic, epigenetic, and post-transcriptional. The genomic DNA encodes the blueprint for proteins and the regulatory elements that control when, where, and how much of each protein is made. These regulatory elements, such as promoters, enhancers, silencers, and insulators, interact with various proteins and RNA molecules to modulate the transcriptional activity of genes. Despite the availability of technologies to annotate regulatory elements, thanks to the efforts of large consortia such as ENCODE (ENCODE Project Consortium, 2012; Hitz et al.; Kagda et al., 2023), Roadmap Epigenomics (Inoue et al., 2017), Blueprint (Kundaje et al., 2015), FANTOM (Martens and Stunnenberg, 2013; Noguchi et al., 2017), and others, and the ability to understand their critical nucleotides through techniques like MPRA (Massively Parallel Reporter Assays) and CRISPR-based perturbations, there remains a significant challenge in fully comprehending these regulatory elements.

Recent advances in synthetic DNA sequence design have utilized machine learning models such as Generative Adversarial Networks (GANs) and Convolutional Neural Networks (CNNs), which have demonstrated potential in sequence generation (Gosai et al., 2023; Taskiran et al., 2024; Zrimec et al., 2022). Initial efforts have been starting to explore diffusion models (Li et al., 2023; Avdeyev et al., 2023; Penzar et al., 2023) without employing conditional diffusion, therefore limiting the generation of cell type-specific regulatory elements.

All existing methods require complex workflows involving multiple models trained independently on individual cell types, with subsequent steps to ensure the cell type specificity of the generated sequences across different cell types. This impedes end-to-end, multi-cell-type model training, which is a significant limitation for tailored sequence design and hampers the ability to develop foundational models that include a wide range of cell types.

Our DNA-Diffusion model bridges the gap between AI-driven generative models and practical applications of synthetic DNA sequence generation for cell type-specific gene regulation. These sequences hold the potential to modify gene expression and be employed in new therapeutic applications requiring precise perturbation of gene regulation (Fig. 1b).

2 RESULTS

2.1 DNA-DIFFUSION: CONDITIONAL DIFFUSION MODEL TO GENERATE CELL TYPE-SPECIFIC REGULATORY ELEMENTS

DNA-Diffusion is a conditional diffusion model (Nichol and Dhariwal, 2021) that operates in the space of DNA sequences. Sequences are encoded using a strategy akin to one-hot encoding, but each nucleotide has a support range of [-1,1] to facilitate the injection of Gaussian noise centered around zero. The backbone of the model is a denoising U-Net with two embedding layers for cell label and timestep, respectively (Fig. 1c). During training, it receives three inputs: DNA sequences, a timestep, and cell type labels. After training, the model takes in input a cell type label and can generate novel cell type-specific sequences.

The primary goal in training this model was to generate cell type-specific regulatory elements. We utilized the DHS index dataset curated by Meuleman et al. (2020), which includes 733 biosamples from 438 cell and tissue types, to derive cell type-specific sequences for GM12878, K562, and HepG2. These cell types were chosen for their distinct biological contexts, diverse tissue origins and encompassing different germ layer lineages: GM12878 (a B lymphocyte cell line) for the immune system, K562 (a leukemia cell line) for blood cancer research, and HepG2 (a hepatocellular carci-

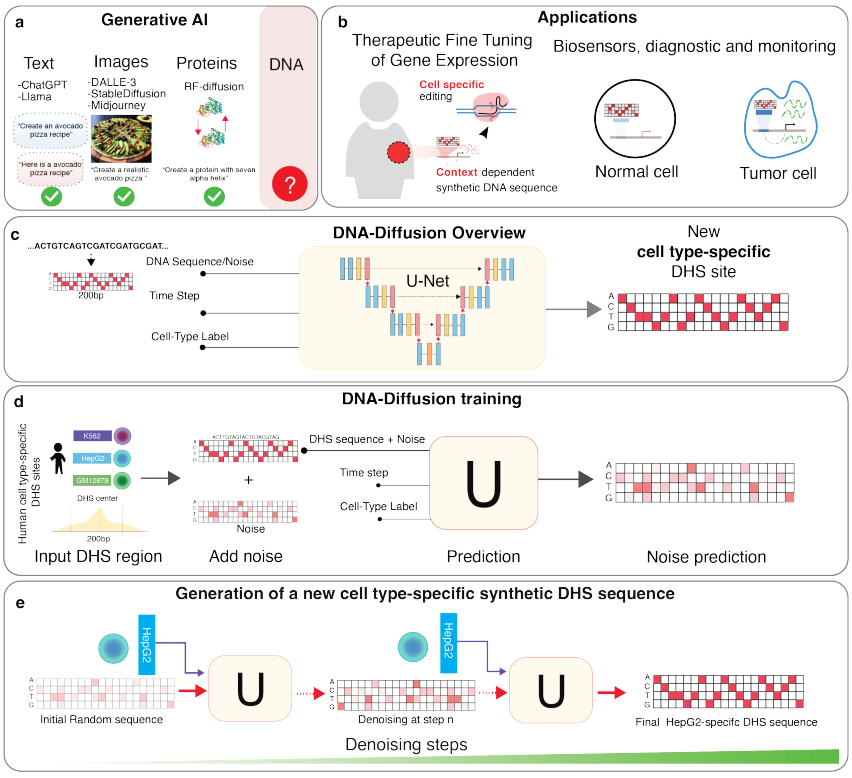


Figure 1: a) Examples of generative AI models across domains (language, images, biology). b) Potential applications of DNA-Diffusion generated synthetic cell type-specific sequences. c) DNA-Diffusion model architecture. U-Net iteratively generates DNA sequences from cell type data. d) DNA-Diffusion trained simultaneously on DHS sequences from K562, HepG2, and GM12878 cells. The model transforms input sequence by adding Gaussian noise and iteratively denoises them during training. e) New sequences are generated by iteratively denoising a randomized sequence with the trained U-Net toward a specific target cell type.

noma cell line) for liver biology and disease studies. For each cell type, we assigned a categorical class label and selected DNA sequences spanning 200 bp around the summit of cell type-specific peaks (Methods). Prior to model training, we adopted a strategy proposed by Meuleman (2018). Briefly, a chromosome-based stratified sampling strategy was employed to partition the dataset into mutually exclusive subsets for training, validation, and testing. The training involved a forward process where varying levels of noise were introduced into the encoded sequence representations, aiming to learn a function that can effectively denoise a sequence at each step by predicting the patterns of introduced noise (Fig. 1d). Upon completing training, a reverse process synthesizes novel DNA sequences, given a specified number of steps and the desired cell type-specific label (Fig. 1e).

Following model training (Supplementary Fig. 1), we generated 100,000 200-bp DNA sequences per cell type. We evaluated these sequences for uniqueness, transcription factor composition, chromatin accessibility, and their ability to drive cell-type-specific gene expression.

2.2 Diffusion-generated sequences exhibit motifs present on endogenous sequences without memorizing

We wanted to assess if the model was memorizing the training sequences and recapitulating TF motif composition of each cell type (Supplementary Note 1). These analyses confirm that DNA-Diffusion sequences are not mere replicas of the training set, but are instead novel and original sequences. The model enhances cell type-specificity by modulating motif density and incorporating known cell type-specific transcription factors, thereby closely mirroring the motif vocabulary of endogenous sequences for a given cell type (Supplementary Fig.2). This highlights the model's capability to

generate realistic and diverse genomic sequences for precise cellular contexts without resorting to memorization.

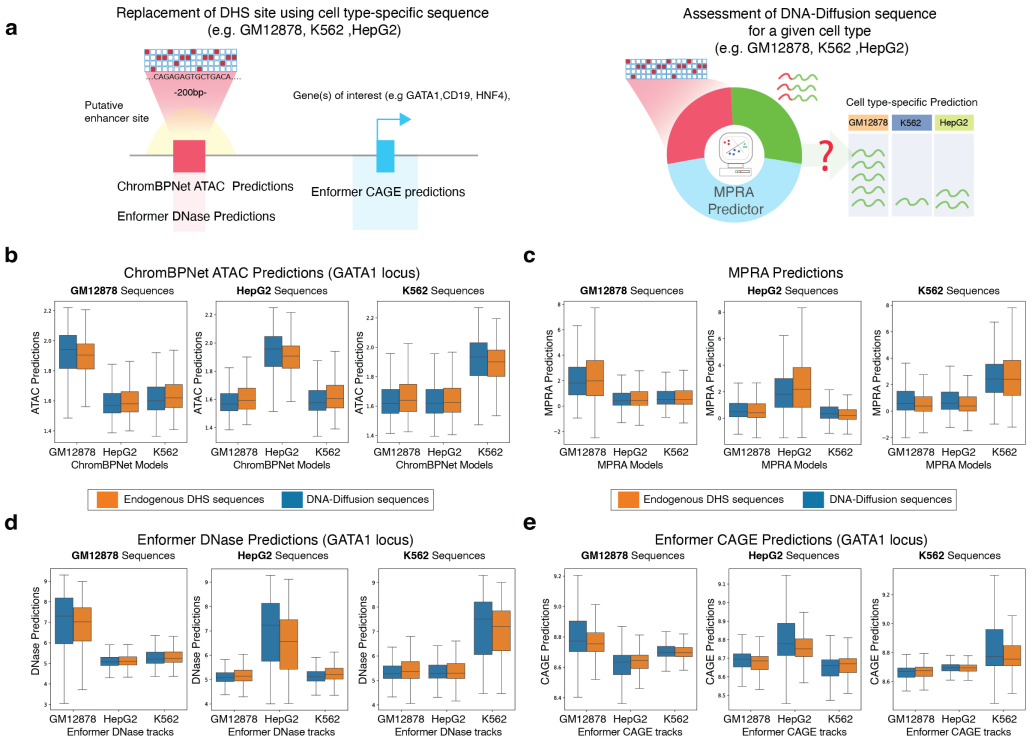


Figure 2: a) Strategy for inserting cell type-specific sequences into enhancer regions for in silico assessment of accessibility (ChromBPNet), gene expression (Enformer), or context-independent expression potential (MPRA predictor). b) ChromBPNet predicts ATAC accessibility for DNA-Diffusion-generated and endogenous sequences in the GATA1 locus across cell lines. c) MPRA assay confirms cell type-specific expression patterns of DNA-Diffusion and endogenous sequences. d) Enformer predicts chromatin accessibility (DNase) for DNA-Diffusion and endogenous sequences in the GATA1 locus. e) DNA-Diffusion sequences drive cell type-specific GATA1 expression.

2.3 DNA-DIFFUSION GENERATED SEQUENCES DEMONSTRATE CELL SPECIFICITY, ACCESSIBILITY AND CAN ACTIVATE CIS GENES IN A CELL TYPE-SPECIFIC MANNER IN SILICO.

To investigate the effects of cell type-specific DNA-Diffusion sequences on chromatin accessibility and gene expression, we replaced endogenous sequences at accessible DHS sites specific to each cell type with our generated DNA-Diffusion sequences. The objective was to determine if these sequences could maintain or change the existing accessibility in their respective cell types. Concurrently, we investigated whether the DNA-Diffusion sequences could induce accessibility in areas that were previously inaccessible, depending on the sequence's initial cellular context. Additionally, we aimed to evaluate the influence of these sequences on the expression of genes potentially regulated by these elements. To this end we considered different "oracles", state of the art prediction methods that have demonstrated the ability to recapitulate gene expression and chromatin accessibility patterns across cell types in human cells (Fig. 2a).

This approach was applied to the DHS regulatory regions of three distinct and well-known cell type-specific genes for a comprehensive analysis (Supplementary Table 1). Specifically, this included GATA1, a transcription factor that plays an essential role in red blood development and is implicated in various blood- related disorders (Wu et al., 2019; Yu et al., 2019). Importantly, the activity of its proximal enhancers has been validated through CRISPR interference (CRISPRi) in K562 cells

(Reilly et al., 2021; Gasperini et al., 2019). Based on these findings, we selected the region of 200bp that showed the strongest activity. HNF4A, another selected gene, plays a crucial role in triggering the transcriptional response in liver cells and is expressed in HepG2 (Meuleman et al., 2020; Argemi et al., 2019). Finally, CD19 is an important gene in B-cell development and serves as a key cellular marker for chimeric antigen receptor (CAR) T-cell therapies (Miller and Maus, 2015). Lacking experimentally validated enhancers for HNF4A and CD19, we leveraged endogenous chromatin accessibility patterns to identify promising candidate regions for evaluating the impact of our engineering sequences. To delineate a panel of candidate enhancers for each gene, we selected a 200bp region within 100KB of its transcription start site (TSS) and characterized by a pronounced cell type-specific pattern of chromatin accessibility (Supplementary Fig. 3).

Having selected the genomic location of these distal regulatory sequence for each locus in different cellular contexts, we next assessed the impact of replacing the endogenous sequences with different cell type-specific sequences for each cell type from two groups: either from DNA-Diffusion sequences or endogenous DHS from other locations. Our objective was to evaluate the extent to which DNA-Diffusion sequences or endogenous sequences, presumed to be active in a particular cellular context, exhibit cell type-specific patterns upon integration into the regulatory regions of these loci. To this end, we utilized ChromBPNet (Pampari et al., 2023), a recent state-of-the-art method capable of predicting chromatin accessibility (ATAC signal) at nucleotide level resolution based exclusively on DNA sequence information. We concurrently utilized three distinct ChromBPNet models, each specifically trained on K562, GM12878, and HepG2 data to analyze the predicted chromatin accessibility patterns of each sequence in the various groups (Methods).

Initially, we focused on the GATA1 locus, replacing the endogenous sequence of the validated enhancer (Supplementary Fig. 3a), known for its cell type-specific chromatin accessibility in K562, with sequences from our two groups. Our observations using the three ChromBPNet models as oracles indicated that only DNA-Diffusion and endogenous sequences specific to K562 successfully maintained or enhanced accessibility in K562. As a baseline, predictions from endogenous DHS sites in the training set and specific for K562 cells had a mean log-normalized predicted ATAC value of 1.88 based on the K562 model, compared to 1.65 for the HepG2 model and 1.66 for the GM12878 model. Notably, the predictions for the DNA-Diffusion sequences showed a value of 1.91 for the K562 model, suggesting slightly higher but significant activity ($p < 0.01$, one-sided t-test) when compared to the endogenous baseline values and slightly lower but significant activity ($p < 0.01$, one- sided t-test) for the other two cells type models, both showing a predicted value of 1.64 (Fig 2b). These results indicate that synthetic K562-specific sequences demonstrated the ability to maintain or enhance chromatin accessibility in K562 cells compared to the endogenous sequence. Conversely, these sequences exhibit reduced accessibility in non-target cell types, suggesting increased context-specificity. Contrasting with the K562 cells, where an established DHS signature demarcates the chosen GATA1 regulatory region (Fig. 2b), HepG2 and GM12878 cells typically display a lack of accessibility at this element (Supplementary Fig. 3 b,c).

This distinction provided an opportunity to evaluate whether DNA-Diffusion sequences, specifically crafted for these cell types, could effectively reactivate chromatin accessibility in regions usually inactive. Despite the inherent inactivity of the selected region in these cell types, both DNA-Diffusion and endogenous sequences specific to GM12878 and HepG2 managed to successfully induce chromatin accessibility (Fig 3b). Similar to the observations in K562 cells, the DNA-Diffusion sequences for HepG2 and GM12878 showed stronger signal in their respective ChromBPNet models compared to predictions in other cell types. Specifically, DNA-Diffusion sequences for the GATA1 regulatory region in HepG2 had a predicted mean ATAC signal of 1.92, with lower values in GM12878 and K562 models (1.60 and 1.63, respectively). Conversely, GM12878 DNA-Diffusion sequences in the GATA1 regulatory region exhibited a mean predicted value of 1.91 in the GM12878 ChromBPNet model, compared to 1.60 in HepG2 and 1.63 in K562 (Fig 2b).

Collectively, these results underscore a context-dependent functionality of DNA-Diffusion sequences, with an enhanced activity within the appropriate cell type and diminished activity in others. Analysis using cell type- specific signals consistently showed that DNA-Diffusion sequences designed for K562, HepG2, and GM12878 exhibited higher predicted accessibility values within their respective models, as opposed to predictions from models trained on different cell types. This evidence supports the cell type-specific utility of the DNA-Diffusion strategy, demonstrating its tailored activation capacity in designated cellular environments.

Having established that our synthetic DNA-Diffusion sequences can modulate chromatin accessibility, we aimed to investigate their impact on gene expression, a task that our DNA-Diffusion model was not specifically trained on. Massively Parallel Reporter Assays (MPRA) have been used to measure the cell type-specific ability of a given sequence to drive a reporter gene's expression in a given cell type and in a genomic context-independent manner. In addition, prior work has shown that the MPRA activity of a sequence can be accurately predicted across a variety of human cell lines (Agarwal et al., 2023) by fine-tuning the Enformer (Avsec et al., 2021a) model. We utilized a similar approach and fine-tuned Enformer to predict MPRA activity in 5 cell lines, including K562, GM12878, and HepG2 cells (Methods). We aimed to validate, *in silico*, cell type-specific expression mediated by our synthetic sequences. Our analysis revealed that sequences designed for a particular target cell showed higher predicted MPRA activity within the corresponding cell compared to predictions for other cell types (Fig. 3c). For example, K562 DNA-Diffusion sequence predictions exhibited a mean normalized value of 2.60 according to the K562 MPRA model, markedly higher than the values of 0.68 and 0.67 for the GM12878 and HepG2 models respectively (Fig.3c). As a baseline, the predicted activity distribution for endogenous train K562 sequences (2.53) closely aligned with the DNA-Diffusion predictions. Similar patterns emerged when analyzing DNA- Diffusion sequences for HepG2 and GM12878, with sequences specifically tailored for each cell type demonstrating higher predicted MPRA values within their respective models (Fig.3c). While the DNA-Diffusion model was not primarily trained to optimize gene expression regulation, it successfully learned to capture inherent regulatory patterns from the endogenous sequences, thereby enabling cell type-specific modulation of predicted MPRA expression.

Building on the evidence that DNA-Diffusion sequences can modulate in silico chromatin accessibility and drive MPRA expression in a cell type-specific manner as demonstrated by ChromBPNet and MPRA predictive models, we sought to delve deeper into their potential to regulate the expression of specific genes. We focused on the replacement of synthetic sequences at the GATA1, HNF4A, and CD19 loci and assessed the impact on gene expression by modulating chromatin accessibility within the cellular contexts of K562, HepG2, and GM12878. A key goal was to ascertain whether these sequences could activate genes not natively expressed in certain cell types.

To this end, we used Enformer, a cutting-edge deep-learning model that has demonstrated the ability to recapitulate gene expression, chromatin histone modification, and accessibility patterns across cell types using only DNA sequence input. In the case of the GATA1 locus, analysis of Enformer DNase predictions after replacing endogenous sequences with generated sequences showed that DNA-Diffusion sequences specific for K562 maintained or enhanced the predicted accessibility in the GATA1 enhancer for K562 compared to the native sequence, while showing no activity for GM12878 and HepG2, consistent with the ChromBPNet analysis (Fig. 2d). Similarly, when introducing cell type-specific HepG2 and GM12878 DNA-Diffusion sequences (Fig. 2d), we observed an increase in accessibility within previously non-activated GATA1 DHS regulatory regions in those cell lines. These sequences induced a statistically significant increase in accessibility in previously quiescent GATA1 DHS regulatory regions for GM12878 and HepG2, exceeding even the levels observed for the endogenous training sequences used during model training (t-test p-value < 0.001).

To assess the impact of these sequences on cell type-specific expression of GATA1, we analyzed the Enformer prediction of the CAGE (Cleavage and Polyadenylation Specificity Factor) assay within the 1KB region proximal to its TSS, noting that GATA1 is already expressed in K562 but not GM12878 and HepG2. The Enformer CAGE analysis revealed a significant increase in predicted promoter activity following the insertion of DNA-Diffusion sequences for each cell type (Fig. 3e). For example, GM12878 DNA-Diffusion sequences replacing the GATA1 regulatory region show a mean CAGE signal of 8.83 in GM12878, surpassing the activity of the best endogenous GM12878 DHS training sequences (8.77, t-test $p < 0.001$) in the same cell type. A similar reactivation was observed for HepG2 DNA-Diffusion sequences (8.33) in HepG2 while K562 DNA-Diffusion sequences maintained or slightly increased expression of GATA1 in K562 (8.71). Similar results were found for the other two loci (Supplementary Note 2 and Supplementary Figs. 4,5). A tiling experiment was also performed that presented alternative enhancer insertion locations (Supplementary Note 3, Supplementary Fig. 6).

Taken together, these analyses highlight the model's versatility in creating sequences with cell-specific potential in modulating chromatin accessibility and gene expression, showcasing its broader applicability and effectiveness in varied genomic and cellular contexts.

3 DISCUSSION

In this study, we introduced DNA-Diffusion, a novel generative approach leveraging conditional diffusion probabilistic models for the design of cell type-specific DNA regulatory sequences. Our findings demonstrate the model's remarkable capacity to generate sequences that not only retain the essential properties of endogenous sequences, such as transcription factor binding motif composition, but also exhibit enhanced regulatory activity and accessibility specific to the targeted cell types.

Importantly, our model presents the first end-to-end solution to sequence design that requires no external model guidance nor orchestration of different models to achieve cell type-specificity. While DNA-Diffusion marks a substantial advance, it is not without limitations. The current model relies on binarized labels for DHS peak identification, which might oversimplify the complexity of chromatin accessibility landscapes. Future iterations could benefit from incorporating continuous measures of chromatin accessibility to capture more nuanced regulatory dynamics. Additionally, expanding the model to encompass a broader range of cell types from the DHS index dataset would be an invaluable extension, allowing for more comprehensive applications in diverse biological contexts.

Additionally, exploring the model's potential in disease contexts, such as cancer or genetic disorders, could open new therapeutic avenues. Specifically, generating regulatory elements that can modulate gene expression in these contexts could lead to novel, patient-specific treatments. To translate *in silico* predictions of DNA-Diffusion sequences into tangible biological outcomes, robust experimental techniques that are able to replace endogenous sequences with generated sequences are paramount. These approaches would validate and elucidate the true functional potential and applicability of these designed sequences.

In conclusion, DNA-Diffusion exemplifies the potential of generative AI in advancing our understanding of genomic regulation and synthetic biology. By generating functional, cell-type-specific regulatory elements, this model holds immense promise for the future of precision medicine and synthetic biology, marking a significant milestone in the journey toward effective therapeutic strategies aimed at fine-tuning gene expression.

REFERENCES

S. J. Gosai, R. I. Castro, N. Fuentes, J. C. Butts, S. Kales, R. R. Noche, K. Mouri, P. C. Sabeti, S. K. Reilly, and R. Tewhey. Machine-guided design of synthetic cell type-specific cis-regulatory elements, August 2023. URL https://www.biorxiv.org/content/10.1101/2023.08.08.552077v1. Pages: 2023.08.08.552077 Section: New Results.

Ibrahim I. Taskiran, Katina I. Spanier, Hannah Dickmänken, Niklas Kempynck, Alexandra Pančíková, Eren Can Ekşi, Gert Hulselmans, Joy N. Ismail, Koen Theunis, Roel Vandepoel, Valerie Christiaens, David Mauduit, and Stein Aerts. Cell-type-directed design of synthetic enhancers. *Nature*, 626(7997):212–220, February 2024. ISSN 1476-4687. doi: 10.1038/s41586-023-06936-2. URL https://www.nature.com/articles/s41586-023-06936-2. Number: 7997 Publisher: Nature Publishing Group.

Jan Zrimec, Xiaozhi Fu, Azam Sheikh Muhammad, Christos Skrekas, Vykintas Jauniskis, Nora K. Speicher, Christoph S. Börlin, Vilhelm Verendel, Morteza Haghir Chehreghani, Devdatt Dubhashi, Verena Siewers, Florian David, Jens Nielsen, and Aleksej Zelezniak. Controlling gene expression with deep generative design of regulatory DNA. *Nature Communications*, 13(1): 5099, August 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-32818-8. URL https://www.nature.com/articles/s41467-022-32818-8. Number: 1 Publisher: Nature Publishing Group.

Zehui Li, Yuhao Ni, Tim August B. Huygelen, Akashaditya Das, Guoxuan Xia, Guy-Bart Stan, and Yiren Zhao. Latent Diffusion Model for DNA Sequence Generation, December 2023. URL http://arxiv.org/abs/2310.06150. arXiv:2310.06150 [cs].

Pavel Avdeyev, Chenlai Shi, Yuhao Tan, Kseniia Dudnyk, and Jian Zhou. Dirichlet Diffusion Score Model for Biological Sequence Generation, June 2023. URL http://arxiv.org/abs/2305.10699. arXiv:2305.10699 [cs, q-bio].

Dmitry Penzar, Daria Nogina, Elizaveta Noskova, Arsenii Zinkevich, Georgy Meshcheryakov, Andrey Lando, Abdul Muntakim Rafi, Carl De Boer, and Ivan V Kulakovskiy. LegNet: a best-in-class deep learning model for short DNA regulatory regions. *Bioinformatics*, 39(8):btad457, August 2023. ISSN 1367-4811. doi: 10.1093/bioinformatics/btad457. URL https://academic.oup.com/bioinformatics/article/doi/10.1093/bioinformatics/btad457/7230784.

Alex Nichol and Prafulla Dhariwal. Improved Denoising Diffusion Probabilistic Models, February 2021. URL http://arxiv.org/abs/2102.09672. arXiv:2102.09672 [cs, stat].

Wouter Meuleman, Alexander Muratov, Eric Rynes, Jessica Halow, Kristen Lee, Daniel Bates, Morgan Diegel, Douglas Dunn, Fidencio Neri, Athanasios Teodosiadis, Alex Reynolds, Eric Haugen, Jemma Nelson, Audra Johnson, Mark Frerker, Michael Buckley, Richard Sandstrom, Jeff Vierstra, Rajinder Kaul, and John Stamatoyannopoulos. Index and biological spectrum of human DNase I hypersensitive sites. *Nature*, 584(7820):244–251, August 2020. ISSN 1476-4687. doi: 10.1038/s41586-020-2559-3. URL https://www.nature.com/articles/s41586-020-2559-3. Number: 7820 Publisher: Nature Publishing Group.

Wouter Meuleman. Synthetic DNA sequences, 2018. URL https://www.meuleman.org/research/synthseqs/.

Yuxuan Wu, Jing Zeng, Benjamin P. Roscoe, Pengpeng Liu, Qiuming Yao, Cicera R. Lazzarotto, Kendell Clement, Mitchel A. Cole, Kevin Luk, Cristina Baricordi, Anne H. Shen, Chunyan Ren, Erica B. Esrick, John P. Manis, David M. Dorfman, David A. Williams, Alessandra Biffi, Carlo Brugnara, Luca Biasco, Christian Brendel, Luca Pinello, Shengdar Q. Tsai, Scot A. Wolfe, and Daniel E. Bauer. Highly efficient therapeutic gene editing of human hematopoietic stem cells. *Nature Medicine*, 25(5):776–783, May 2019. ISSN 1546-170X. doi: 10.1038/s41591-019-0401-y. URL https://www.nature.com/articles/s41591-019-0401-y. Number: 5 Publisher: Nature Publishing Group.

Shan Yu, Xuepeng Jiang, Juan Li, Chao Li, Mian Guo, Fei Ye, Maomao Zhang, Yufei Jiao, and Baoliang Guo. Comprehensive analysis of the GATA transcription factor gene family in breast carcinoma using gene microarrays, online databases and integrated bioinformatics. *Scientific Reports*, 9(1):4467, March 2019. ISSN 2045-2322. doi: 10.1038/s41598-019-40811-3. URL https://www.nature.com/articles/s41598-019-40811-3. Number: 1 Publisher: Nature Publishing Group.

Steven K. Reilly, Sager J. Gosai, Alan Gutierrez, Ava Mackay-Smith, Jacob C. Ulirsch, Masahiro Kanai, Kousuke Mouri, Daniel Berenzy, Susan Kales, Gina M. Butler, Adrianne Gladden-Young, Redwan M. Bhuiyan, Michael L. Stitzel, Hilary K. Finucane, Pardis C. Sabeti, and Ryan Tewhey. Direct characterization of cis-regulatory elements and functional dissection of complex genetic associations using HCR-FlowFISH. *Nature Genetics*, 53(8):1166–1176, August 2021. ISSN 1546-1718. doi: 10.1038/s41588-021-00900-4.

Molly Gasperini, Andrew J. Hill, José L. McFaline-Figueroa, Beth Martin, Seungsoo Kim, Melissa D. Zhang, Dana Jackson, Anh Leith, Jacob Schreiber, William S. Noble, Cole Trapnell, Nadav Ahituv, and Jay Shendure. A Genome-wide Framework for Mapping Gene Regulation via Cellular Genetic Screens. *Cell*, 176(1-2):377–390.e19, January 2019. ISSN 1097-4172. doi: 10.1016/j.cell.2018.11.029.

Josepmaria Argemi, Maria U. Latasa, Stephen R. Atkinson, Ilya O. Blokhin, Veronica Massey, Joel P. Gue, Joaquin Cabezas, Juan J. Lozano, Derek Van Booven, Aaron Bell, Sheng Cao, Lawrence A. Vernetti, Juan P. Arab, Meritxell Ventura-Cots, Lia R. Edmunds, Constantino Fondevila, Peter Stärkel, Laurent Dubuquoy, Alexandre Louvet, Gemma Odena, Juan L. Gomez, Tomas Aragon, Jose Altamirano, Juan Caballeria, Michael J. Jurczak, D. Lansing Taylor, Carmen Berasain, Claes Wahlestedt, Satdarshan P. Monga, Marsha Y. Morgan, Pau Sancho-Bru, Philippe Mathurin, Shinji Furuya, Carolin Lackner, Ivan Rusyn, Vijay H. Shah, Mark R. Thursz, Jelena Mann, Matias A. Avila, and Ramon Bataller. Defective HNF4alpha-dependent gene expression as a driver of hepatocellular failure in alcoholic hepatitis. *Nature Communications*, 10(1):3126, July 2019. ISSN 2041-1723. doi: 10.1038/s41467-019-11004-3. URL https://www.nature.com/articles/s41467-019-11004-3. Number: 1 Publisher: Nature Publishing Group.

Brian C. Miller and Marcela V. Maus. CD19-Targeted CAR T Cells: A New Tool in the Fight against B Cell Malignancies. *Oncology Research and Treatment*, 38(12):683–690, 2015. ISSN 2296-5262. doi: 10.1159/000442170.

Anusri Pampari, Anna Shcherbina, Surag Nair, Jacob Schreiber, Aman Patel, Austin Wang, Soumya Kundu, Avanti Shrikumar, and Anshul Kundaje. Bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants., December 2023. URL https://zenodo.org/doi/10.5281/zenodo.7567627.

Vikram Agarwal, Fumitaka Inoue, Max Schubach, Beth K. Martin, Pyaree Mohan Dash, Zicong Zhang, Ajuni Sohota, William Stafford Noble, Galip Gürkan Yardimci, Martin Kircher, Jay Shendure, and Nadav Ahituv. Massively parallel characterization of transcriptional regulatory elements in three diverse human cell types. preprint, Genomics, March 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.03.05.531189.

Žiga Avsec, Vikram Agarwal, Daniel Visentin, Joseph R. Ledsam, Agnieszka Grabska-Barwinska, Kyle R. Taylor, Yannis Assael, John Jumper, Pushmeet Kohli, and David R. Kelley. Effective gene expression prediction from sequence by integrating long-range interactions. *Nature Methods*, 18(10):1196–1203, October 2021a. ISSN 1548-7105. doi: 10.1038/s41592-021-01252-x. URL https://www.nature.com/articles/s41592-021-01252-x. Number: 10 Publisher: Nature Publishing Group.

Niels Rogge and Kashif Rasul. The Annotated Diffusion Model. URL https://huggingface.co/blog/annotated-diffusion.

Olaf Ronneberger, Philipp Fischer, and Thomas Brox. U-Net: Convolutional Networks for Biomedical Image Segmentation, May 2015. URL http://arxiv.org/abs/1505.04597. arXiv:1505.04597 [cs].

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep Residual Learning for Image Recognition, December 2015. URL http://arxiv.org/abs/1512.03385. arXiv:1512.03385 [cs].

Zhuoran Shen, Mingyuan Zhang, Haiyu Zhao, Shuai Yi, and Hongsheng Li. Efficient Attention: Attention with Linear Complexities, November 2020. URL http://arxiv.org/abs/1812.01243. arXiv:1812.01243 [cs].

Jonathan Ho and Tim Salimans. Classifier-Free Diffusion Guidance, July 2022. URL http://arxiv.org/abs/2207.12598. arXiv:2207.12598 [cs].

Diederik P. Kingma and Jimmy Ba. Adam: A Method for Stochastic Optimization, January 2017. URL http://arxiv.org/abs/1412.6980. arXiv:1412.6980 [cs].

Žiga Avsec, Melanie Weilert, Avanti Shrikumar, Sabrina Krueger, Amr Alexandari, Khyati Dalal, Robin Fropf, Charles McAnany, Julien Gagneur, Anshul Kundaje, and Julia Zeitlinger. Base-resolution models of transcription-factor binding reveal soft motif syntax. *Nature Genetics*, 53 (3):354–366, March 2021b. ISSN 1546-1718. doi: 10.1038/s41588-021-00782-6. URL https://www.nature.com/articles/s41588-021-00782-6. Number: 3 Publisher: Nature Publishing Group.

Aaron R. Quinlan and Ira M. Hall. BEDTools: a flexible suite of utilities for comparing genomic features. *Bioinformatics (Oxford, England)*, 26(6):841–842, March 2010. ISSN 1367-4811. doi: 10.1093/bioinformatics/btq033.

W. James Kent. BLAT–the BLAST-like alignment tool. *Genome Research*, 12(4):656–664, April 2002. ISSN 1088-9051. doi: 10.1101/gr.229202.

Janne Korhonen, Petri Martinmäki, Cinzia Pizzi, Pasi Rastas, and Esko Ukkonen. MOODS: fast search for position weight matrix matches in DNA sequences. *Bioinformatics (Oxford, England)*, 25(23):3181–3182, December 2009. ISSN 1367-4811. doi: 10.1093/bioinformatics/btp554.

Jaime A. Castro-Mondragon, Rafael Riudavets-Puig, Ieva Rauluseviciute, Roza Berhanu Lemma, Laura Turchi, Romain Blanc-Mathieu, Jeremy Lucas, Paul Boddie, Aziz Khan, Nicolás Manosalva Pérez, Oriol Fornes, Tiffany Y. Leung, Alejandro Aguirre, Fayrouz Hammal, Daniel Schmelter, Damir Baranasic, Benoit Ballester, Albin Sandelin, Boris Lenhard, Klaas Vandepoele, Wyeth W. Wasserman, François Parcy, and Anthony Mathelier. JASPAR 2022: the 9th release of the open-access database of transcription factor binding profiles. *Nucleic Acids Research*, 50 (D1):D165–D173, January 2022. ISSN 1362-4962. doi: 10.1093/nar/gkab1113.

Alexander Karollus, Thomas Mauermeier, and Julien Gagneur. Current sequence-based models capture gene expression determinants in promoters but mostly ignore distal enhancers. *Genome Biology*, 24(1):56, March 2023. ISSN 1474-760X. doi: 10.1186/s13059-023-02899-9. URL https://doi.org/10.1186/s13059-023-02899-9.

Jeffrey L. Hansen and Barak A. Cohen. A quantitative metric of pioneer activity reveals that HNF4A has stronger in vivo pioneer activity than FOXA1. *Genome Biology*, 23(1):221, October 2022. ISSN 1474-760X. doi: 10.1186/s13059-022-02792-x.

Catherine S. Lee, Joshua R. Friedman, James T. Fulmer, and Klaus H. Kaestner. The initiation of liver development is dependent on Foxa transcription factors. *Nature*, 435(7044):944–947, June 2005. ISSN 1476-4687. doi: 10.1038/nature03649.

G. Celine Han, Vinesh Vinayachandran, Alain R. Bataille, Bongsoo Park, Ka Yim Chan-Salis, Cheryl A. Keller, Maria Long, Shaun Mahony, Ross C. Hardison, and B. Franklin Pugh. Genome-Wide Organization of GATA1 and TAL1 Determined at High Resolution. *Molecular and Cellular Biology*, 36(1):157–172, December 2015. ISSN 0270-7306. doi: 10.1128/MCB.00806-15. URL https://www.ncbi.nlm.nih.gov/pmc/articles/PMC4702602/.

4 METHODS

4.1 DATA PREPROCESSING FOR TRAINING

To facilitate data filtering and analysis, the binary peaks annotation from the DHS Index was augmented with the following metadata columns: chromosome (chr), DHS start position, DHS end position, DHS width, DHS summit (peak maximum), total signal strength, component, proportion, and sequence (Meuleman, 2018). Given the variable length of ENCODE DHS sites, we standardized sequence analysis by extracting a 200bp region centered on the peak summit (+/- 100bp). We use this threshold by assuming that the peak center should represent the most accessible regions and take into consideration the expected number of nucleotides protected by a single chromosome 18. Our analysis of 733 biosamples from 438 cell types aim to identify replicates exhibiting both strong component association in the non-negative factorization data and ease of laboratory culture; ultimately yielding four replicates from K562, GM12878, HepG2, and hESCT0 cell lines, respectively.

Using the selected replicates, the complete dataset was filtered only to include the DHS sites that were cell- type specific. This cell type-specificity refinement strategy consisted of prioritizing peaks that had limited occurrences in cell lines outside our replicates, while also putting emphasis on samples that had peaks in other replicates of the same cell. From there we filtered out samples that occurred in more than one of our cells of interest, so that we were left with a dataset that was sorted in descending order of peak presence in other replicates and ascending order for peak presence in other cell types. GM12878 was the cell line with the smallest number of cell type-specific peaks (11968 sites). To have a balanced number of regions across all cells every replicate was restricted to the same number. From there, regions present in chr1 and chr2 were held out as test and validation shuffle sets, respectively.

4.2 MODEL ARCHITECTURE

The DNA-Diffusion model takes inspiration from the Annotated Diffusion Model (Rogge and Rasul), with some key modifications. At its core, the diffusion model consists of a U-Net (Ronneberger et al., 2015) that first projects the batch of encoded sequences to the desired channel dimension of 200. Each downsampling layer consists of 2 ResNet (He et al., 2015) blocks, followed by a linear attention (Shen et al., 2020) layer, and then downsampling convolutional layers until the output

channel dimension is reached. The middle layers of the U-Net consist of a ResNet block and an attention layer before an additional ResNet block. The structure of the upsampling stage is the same as the downsampling stage with the exception of the downsampling convolutional layers being replaced with a upsampling convolutional layers. The output of the U-Net is then projected through one final ResNet block before one convolutional layer is applied to return the output with the same dimensionality as the input.

During the forward process of this diffusion model, a batch of DNA sequences and their corresponding cell type-specific labels are fed through the forward process where the cell type labels are randomly masked (setting to null), allowing for classifier-free guidance (Ho and Salimans, 2022) during training and downstream sampling. Using this sampling process 100,000 sequences per cell type were generated, resulting in a generated set of 300,000 sequences.

4.3 TRAINING PROCESS

The main diffusion model was implemented in PyTorch 2.0.0 and trained on 4x 40gb Nvidia A100s using a batch size of 960 for 2000 epochs. We used the Adam optimizer (Kingma and Ba, 2017) with a learning rate of $1e^{-4}$ along with PyTorch defaults values for other hyperparameters: $\beta_1 = 0.9$ and $\beta_2 = 0.999$. Linear noising schedule was used with $\beta_{start} = 0.0001$ and $\beta_{end} = 0.005$.

4.4 *In silico* VALIDATION

4.4.1 ASSESSING DNA CHROMATIN ACCESSIBILITY WITH CHROMBPNET

Accessible chromatin is a notable marker of regulatory activity for a DNA sequence. We use the ChromBPNet model as a biological oracle to assess the cell-type specificity of generated diffusion sequences' chromatin accessibility. ChromBPNet consists of two convolutional neural networks that are similar in structure to BPNet (Avsec et al., 2021b). One convolutional neural network learns to predict base-pair resolution chromatin accessibility, while the second network learns to predict the noise of an experimental assay (e.g. DNase-Seq, ATAC-seq). The outputs of the two networks are summed to predict the log of total counts and base-resolution probability distribution of counts of a 1000 bp sequence. Because the ChromBPNet architecture consists of two models to predict the ground truth data of an assay, the first convolutional neural network's output can serve as a bias-corrected chromatin accessibility at base-pair resolution without the noise of an assay affecting the prediction.

We used ChromBPNet models trained on ENCSR000EPC, ENCSR000EMT, and ENCSR000ENP to predict base-pair chromatin accessibility of DNA-Diffusion sequences in K562, GM12878, and HepG2 contexts respectively. Each ChromBPNet model's input was a one-hot encoded 2114 bp DNA sequence (A = [1,0,0,0], C = [0,1,0,0], G = [0,0,1,0], T = [0,0,0,1]).

To evaluate how DNA sequences generated by the diffusion model affect chromatin accessibility, we inserted them within known or putative regulatory regions for the following genes: GATA1, CD19, and HNF4A and calculated the mean of the predicted base-resolution chromatin accessibility score.

Given a generated sequence designed for a particular cell type, we ideally expect that the ChromBPNet model trained to predict accessibility in that cell type will yield greater chromatin accessibility predictions relative to the other two ChromBPNet models, indicating greater regulatory potential. However, the three ChromBPNet models used were trained on experimental data with different read depths: ENCSR000EPC had a read depth of 245 million, ENCSR000EMT had a read depth of 68 million, and ENCSR000EMT had a read depth of 323 million. Thus, we observed that models trained on deeper data typically predicted higher chromatin accessibility compared to the other two models for a generated sequence, irrespective of the cell type specificity of a sequence - leading to inaccurate comparison of model outputs. To combat this potential confounder, we applied quantile normalization (see Normalization and scaling of the data).

4.4.2 MPRA PREDICTOR

We train our MPRA activity predictor using MPRA data from GM12878, K562, HepG2, SK-N-SH, and A549 cells (accession ids are available in Supplementary Table 3), collected by one lab using a uniform protocol (Gosai et al., 2023). The assayed sequences are 200bp long and most of them are

derived from human genomic segments containing the reference and alternate alleles for variants from UK Biobank and GTEx. These sequences are cloned upstream of a reporter gene and this construct is delivered to host cells using transient transfection. Each sequence's MPRA activity is measured as the log2 of the ratio of the number of mRNA molecules produced using the construct to the number of constructs delivered to the host cells. From ENCODE, we obtain 318734, 636185, 750298, 750084, and 318734 MPRA activity measurements from GM12878, K562, HepG2, SK-N-SH, and A549 cells respectively.

To predict a given sequence's activity in each of the cells, we input the 200bp sequence to Enformer and obtain its sequence embeddings from the last transformer layer. Then, we perform attention pooling (over the sequence's length) to get the final sequence embedding. This sequence embedding is supplied to a fully connected output layer that simultaneously predicts MPRA activity in all five cell lines using multi-task learning. We initialize Enformer using pretrained weights provided by the authors. Following recent work from the lab that collected and used a portion of this data11, we trained our model using all sequences except those from chromosomes 7, 13, 19, 21, and X (79% of all sequences). Sequences from chromosomes 7 and 13 are used in the validation set (13% of all sequences), and sequences from chromosomes 19, 21, and X (7% of all sequences) are used in the test set. The model is trained for a maximum of 50 epochs using the AdamW optimizer45 with a 1e-4 learning rate and 1e-4 weight decay. Training is stopped if model performance on the validation set does not improve for 5 consecutive epochs.

Our trained model is quite accurate and achieves Spearman's rank correlation coefficients of 0.4527, 0.7285, 0.7682, 0.7745 and 0.6682 on the test sets for GM12878, K562, HepG2, SK-N-SH, and A549 cells respectively. It achieves Pearson's correlation coefficients of 0.4479, 0.8115, 0.8257, 0.8199 and 0.7218 on the test sets for GM12878, K562, HepG2, SK-N-SH, and A549 cells respectively.

We get model predictions for the generated sequences, control sequences, and sequences used in the training, validation, and test sets of the diffusion models. Since we do not have MPRA data from hESCT0 cells, we only analyze generated sequences that are designed to have regulatory activity in GM1278, K562, or HepG2 cells.

4.4.3 Assessing DNA chromatin accessibility with Enformer

We replaced the endogenous sequence of putative regulatory regions with DNA-Diffusion or other endogenous sequences and considered Enformer predictions using a window of 393,216 base pairs around the region used for replacement. The resulting Enformer predictions offered an approximate 128 bp resolution window for DNase and CAGE tracks across all three cell lines. The extracted predictions across enhancer and promoter regions were averaged to generate a single value that described the prediction signal at the given coordinate. A shortcoming of Enformer is that it is difficult to decrease or remove the CAGE signal in gene regions where a cell is already expected to display promoter activity. As a result, in-silico validation results were only utilized from regions where the cell line previously had no CAGE activity. This resulted in the following process: K562 sequence activity was examined in HNF4A and CD19, HepG2 in GATA1 and CD19, and GM12878 in HNF4A and GATA1 (Supplementary Table 2).

4.5 Normalization and scaling of the data

Since Enformer (DNase and CAGE) and ChromBPNet (ATAC-seq) were trained using data with different sequencing depths, their baseline predictions exhibited a wide dynamic range, making it difficult to compare the difference in raw predictions across cell lines. Utilizing quantile normalization, we adjusted these predictions and rescaled the data to a common range. This process helped to minimize the differences and biases, allowing for more accurate comparisons and interpretations of the data across the loci and cell types.

For ChromBPNet predictions in all loci, we normalized the three cell lines' accessibility predictions, generating a similarly distributed set across K562, HepG2, and GM12878. Instead of different models, Enformer outputs a separate prediction signal track for each cell type and data modality; we used the by-cell procedure mentioned from ChromBPNet to normalize the Enformer DNase and CAGE track predictions.

During normalization, we artificially introduced prediction values from 20k annotated GENCODE v43 protein- coding transcript promoters and 5k random endogenous genomic regions to ensure a final dynamic range representing natural endogenous sequence occurrence distribution.

During the ChromBPNet and Enformer DNase normalization, we introduced 5k prediction values captured from sequences selected directly from the random genome coordinates. The sequence coordinates were defined using the Bedtools Random function Quinlan and Hall (2010) in the hg38 genome version. While for the Enformer CAGE prediction normalization, we added prediction values from 20k promoters, utilizing predictions captured from their original genomic occurrence.

After normalizing all the signal predictions, we used the rank percentile to transform all the normalized values from ChromBPNet, DNase, and CAGE along each cell type to accommodate the values between a 0-1 scale.

4.6 Cell type-specificity and ranking of sequences based on the predictions of different oracles

We call a predicted signal cell-specific when a prediction presents a bias towards a single cell type. For example, a GM12878 endogenous positive sequence is expected to have a stronger accessibility signal or targeted gene expression (e.g. GATA1) for GM12878 cells while presenting almost no signal in HepG2 and K562 cells.

We calculated the cell specificity for a given metric by selecting a specific cell type as the target and the other two remaining cells as off-targets. The specificity was calculated by dividing the signal in the target cell by the sum of the signal in all different cells (target + off-target1 + off-target2). The signal specificity is 1.0 when only the target cell shows activity. The signal specificity is 0.33 when all the cell types have exactly the same signal intensity.

We calculated specificity metrics for all the model predictions from ChromBPNet, Enformer DNase, and Enformer CAGE resulting in a total of twenty-four metrics per sequence and considering the average rank across these metrics as described below.

4.7 DNA-Diffusion specificity focused sequences

To evaluate the role of cell specificity in our model, we first filtered the set of sequences to keep a moderate baseline signal for all metrics (DNase, ChromBPNet, and CAGE $>$ predicted signal percentile 0.5) and a baseline specificity of 0.5. We removed sequences that performed well only in a single metric to ensure a moderate baseline signal and specificity. After filtering out sequences with prediction signal and specificity under the baseline thresholds, we selected sequences presenting high specificity (> 0.8) in at least one of the metrics (DNase, ChromBPNet, and CAGE).

Sequences with high specificity in at least one of the metrics were ranked by creating first an averaged accessibility rank (DNase + ChromBPNet) and selecting the final sequences by averaging the previously generated accessibility rank and the CAGE rank.

DNase and ChromBPNet were ranked in their respective groups before being averaged to create an average accessibility rank. From there, the CAGE values were ranked and then averaged with the accessibility rank to obtain an overall rank across the baseline signal metrics. This two-step procedure allowed us to avoid the specificity being biased through accessibility since Enformer DNase predictions and ChromBPNet are correlated. This new rank was then utilized in conjunction with the designated high-specificity sequences to select a set of 400 highest-ranked DNA-Diffusion specificity-focused sequences.

4.8 DNA-Diffusion signal focused sequences

A different DNA-Diffusion sequence group was selected to investigate the role of signal strength without considering cell specificity. We selected 400 DNA-diffusion High Signal sequences for this set by ranking the strongest signal and averaging the rank for all metrics (DNase, ChromBPNet, and CAGE) in both cell-specific loci. The final top 400 sequences present the highest possible signal average among all generated sequences. No specificity metric was used in this group.

4.9 MOTIF SCANNING

To scan motifs the MOODS package was used using the vertebrate motifs from the JASPAR database (n = 949 PWMs). The parameters used in MOODS were a p-value of 0.0001 and a pseudocount (–ps 0.0001).

5 ACKNOWLEDGMENTS

We would like to acknowledge Bo Wang, Vallijah Subasri, Micaela Consens, Rex Ma, Judith Franziska Kribelbauer, Bart Deplancke, Jiecong Lin and other people in the Pinello Lab, Niccolò Zanichelli and the OpenBioML.org community for helpful feedback or discussions. Anshul Kundaje, Anusri Pampari for their support in using the ChromBPnet models. We thank StabilityAI for the use of its HPC cluster to train and test our models. L.P. is partially supported by 1R35HG010717-01. W.M. is partially supported by 1R35HG011317-01.

6 SUPPLEMENTARY NOTES

6.1 SUPPLEMENTARY NOTE 1. DIFFUSION-GENERATED SEQUENCES EXHIBIT MOTIFS PRESENT IN ENDOGENOUS SEQUENCES WITHOUT MEMORIZING EXISTING SEQUENCES.

To verify the uniqueness of the DNA-Diffusion sequences and confirm that they were not mere copies of the training sequences, we conducted a comparative sequence alignment between the generated DNA-Diffusion sequences and the sets of train, test and randomly sampled endogenous sequences, employing BLAT (Kent, 2002) analysis for this purpose.

The average BLAT alignment length distribution between the DNA-Diffusion sequences and endogenous training sequences was found to be significantly shorter (23.4bp, t-test p-value < 0.001) than between endogenous test and training sequences (48.8bp) or between randomly sampled genomic sequences and endogenous training sequences (49.3bp) (Fig. 2c). To assess sequence diversity of the generated sequences, we also perform comparative sequence alignment of each set with itself. Comparing all the possible pairwise DNA-Diffusion sequences we observed a significantly lower average match (23bp, t-test p-value < 0.001) as compared to train (29bp) or random (39bp) endogenous sequences (Fig. 2d).

Among the analyzed DNA-Diffusion sequences, only 111 out of 300k (0.037%) displayed greater than 40bp overlap with the DHS training dataset (52 for GM12878, 30 for HepG2, and 29 for K562). Additionally, assessing the diversity of DNA-Diffusion sequences, we found that of the 300k sequences, only 223 matched with others, with the matches per cell-type being 103 in GM12878 (0.104% of 100,000), 78 in HepG2 (0.079%), and 42 in K562 (0.043%). These findings support that DNA-Diffusion sequences are not copying large sequence segments from the endogenous training set, while also presenting a diverse set of novel sequences.

We next wanted to demonstrate that the diffusion model learned cell type-specific TF composition. Using MOODS (Korhonen et al., 2009), we scanned the sequences for recognizable TF binding sites using TF position frequency matrices (PFMs) from the JASPAR database (Castro-Mondragon et al., 2022). To assess the fidelity of our model's output to biological reality, we employed the Jensen-Shannon (JS) divergence for comparing the distribution of TF motif hits in cell type-specific DNA-Diffusion sequences against the endogenous test set. This analysis yielded a mean JS divergence of 0.101 across the three cell-types (Fig. 2c), revealing a high degree of similarity between the DNA-Diffusion sequences and the endogenous test set motif distributions. Further reinforcing this finding, we established a baseline by comparing endogenous training sequences to endogenous test sequences within each cell type, which showed a closer similarity with a mean JS divergence of 0.048 across the three cell-types (Fig. 2b). Importantly, we observed a clear separation of motif composition across cell types in both comparisons, with a slightly more distinct separation in generated sequences (GM12878 DNA-Diffusion: 0.27, GM12878 endogenous train: 0.22, HepG2 DNA-Diffusion: 0.29, HepG2 endogenous train: 0.23, K562 DNA- Diffusion: 0.25, K562 endogenous train: 0.21), suggesting some divergence of TF motif usage from the endogenous sequences.

6.2 SUPPLEMENTARY NOTE 2. ASSESSMENT OF DNA-DIFFUSION SEQUENCES FOR HNF4A AND CD19 LOCI

Extending our investigation to assess the generalizability of our approach, we explored its applicability to the regions selected for the HNF4A and CD19 loci, with the former being accessible in HepG2 but not in K562 and GM12878, and the latter accessible in GM12878 but not in K562 and HepG2 (Supplementary Fig 4b,c). Using the same approach, we observed that DNA-Diffusion sequences were able to maintain, augment, or initiate chromatin accessibility across various cell types at these loci, which illustrates these effects in different genomic contexts (Supplementary Figs. 4a, 5a). These observations suggest that DNA-Diffusion sequences can achieve a spectrum of chromatin accessibility that is comparable or greater than that of endogenous sequences in the right cellular type and in different genomic contexts. The DNA-Diffusion model not only preserved and modified accessibility in regions with established DHS activity specific to a cell type but also successfully induced accessibility in genomic regions previously inactive in the respective cell types.

6.3 SUPPLEMENTARY NOTE 3: SELECTION OF OPTIMAL REGION FOR INSERTION OF A SYNTHETIC SEQUENCE FOR REACTIVATION OF A TARGET GENE

To explore the effects of replacing DNA-Diffusion sequences beyond putative regulatory regions annotated based on known cell type-specific DHS sites, we aimed to pinpoint novel regions suitable for the insertion of synthetic elements that could enhance cell type-specific reactivation of a target gene, considering the potential interactions with other regulatory elements within the same locus.

We implemented a tiling strategy, analyzing every 200bp segment within the GATA1 locus (Fig 4a). By substituting a DNA-Diffusion sequence tailored for HepG2 into consecutive windows, we assessed changes in GATA1 expression, as reflected by the predicted CAGE signal around its TSS (Fig 4b).

As already discussed in the previous section, substituting this sequence at the endogenous K562 DNase- accessible region substantially increased chromatin accessibility in HepG2 cells but also reactivated GATA1 expression (Fig 3c), here we investigated whether other locations might exhibit similar or improved reactivation effects. This analysis indicated that the level of predicted reactivation varied with the sequence replacement location. Interestingly, certain regions enhanced GATA1 reactivation in HepG2 cells beyond what was observed at the initial replacement location determined by the existing regulatory element in K562, specifically in an intronic region of GATA1, with the promoter region of GATA1 also exhibiting significant reactivation (Fig 4c; Supplementary Figure 8). Importantly, replacing the sequence outside an existing regulatory element for a given cell line offers a crucial advantage: it avoids disrupting the function of preexisting regulatory elements crucial for the cell type where the gene is naturally active.

This method effectively identified both optimal sites for element insertion and previously unrecognized regulatory regions, thereby providing a versatile approach for precise genomic modifications, considering the broader genomic context for any gene of interest. Nonetheless, it's essential to recognize that our strategy, which relies on Enformer, may not fully account for the impact of regulatory elements distant from the TSS of the target gene, as suggested by prior studies analyzing the predicted effects of these models on the perturbation or deletion of distal regulatory elements (Karollus et al., 2023).

To elucidate the underlying factors contributing to the observed discrepancy in motif composition between DNA-Diffusion sequences and endogenous training sequences, we compared individual motif abundances across these two sets (Fig. 2c). Our analysis revealed that cell type-specific TFs are present in a higher number of DNA-Diffusion sequences compared to endogenous train sequences. For example, HNF4A—a liver-specific pioneer factor (Hansen and Cohen, 2022)—was detected in approximately 29% of endogenous training sequences, yet it was present in 48% of DNA-Diffusion sequences. Similar trends were observed with other liver-specific factors, such as FOXA1 and FOXA3 (Hansen and Cohen, 2022; Lee et al., 2005). Furthermore, this pattern of enriched cell type-specific TFs in DNA- Diffusion sequences extended to other cell-types. For instance, in the GM12878 cell-type, the immune-related factor IRF1 showed a presence in 63% of DNA-Diffusion sequences, contrasting with 47% in endogenous training sequences. Similarly, in the K562 cell line, the combined GATA1-TAL1 motif (Lee et al., 2005; Han et al., 2015) was present in 42% of DNA-Diffusion sequences, compared to 31% in endogenous training sequences. Interestingly, the DNA-Diffusion sequences exhibited a lower abundance of cell type-specific motifs for cell types other than the targeted one. For instance, IRF1 was found less frequently in HepG2 and K562 specific sequences as compared to the sequences generated for GM12878. This suggests that the model is effectively using cell type-specific motifs to create sequences with greater specificity. On the other hand, endogenous sequences may exhibit a "leaky" or non-specific expression when placed in an alternate cellular context. For example, endogenous sequences selected for HepG2, which more frequently contain the IRF1 motif, might still demonstrate some degree of weak expression when introduced into GM12878 cells, albeit in a less targeted manner.

In addition to the changes in motif abundance across different sets of sequences, we also noted an increase in the average number of motif hits within individual DNA-Diffusion sequences, particularly for cell type-specific transcription factors (Fig. 2d). This indicates not just a shift in the overall presence of motifs in the set of DNA-Diffusion sequences, but also an enhanced density of binding sites per sequence. For instance, DNA-Diffusion sequences for HepG2 contained on average about 0.62 HNF4A motifs, in contrast to approximately 0.32 in the endogenous training sequences. In a

similar vein, K562 sequences demonstrated an increased number of GATA1-TAL1 and SMAD1 motifs. Likewise, GM12878 sequences were characterized by more frequent occurrences of IRF1 and STAT1 motifs. Overall, this underscores a pattern of increased motif representation per sequence in the DNA-Diffusion sequences across diverse cell types.

6.4 SELECTION AND CHARACTERIZATION OF SYNTHETIC SEQUENCES WITH DIFFERENT REGULATORY POTENTIAL AND CELL TYPE-SPECIFICITY

Having established that our DNA-Diffusion sequences match or exceed the accessibility and gene expression levels of the DHS sites used for training across different loci and cell types, we explored strategies for selecting sequences with particular regulatory characteristics along two dimensions: "signal intensity" and "signal specificity" (Fig. 5a). Generating cell type-specific sequences that fine-tune chromatin accessibility and gene expression can greatly enhance the development of synthetic gene circuits and precision gene therapies, potentially leading to innovative therapeutics for targeted drug release, diagnostic applications, or the dynamic modulation of cellular functions. Importantly, the generative aspect of our model enables us to sample a broad set of sequences to fulfill specific design requirements.

To this end, we developed a framework to select DNA-Diffusion sequences that can display different levels of accessibility and regulatory activity across different cell types for a given locus of interest. Specifically, we combined in-silico oracles used to evaluate the regulatory potential of these sequences, averaging the Enformer CAGE, Enformer DNase, and ChromBPNet ATAC predictive scores to define overall signal specificity and specificity (Methods). Briefly, we characterized signal specificity for a particular target cell and sequence by calculating the ratio of the signal intensity in the target cell to the total signal intensities across all cells.

Comparing the training endogenous sequences with DNA-Diffusion sequences across these two metrics for the GATA1 locus, particularly focusing on HepG2-specific sequences, we noted significant variations. Despite the use of specificity filtration criteria during training, aimed at promoting cell type-specific sequences, the specificity of these sequences showed substantial variability, a trend effectively mirrored by the DNA-Diffusion sequences (Fig 5 b,c). Remarkably, a larger number of DNA-Diffusion sequences exhibited superior performance, outperforming the endogenous sequences in both signal strength and specificity. This highlights the model's refined capability to discern sequence features unique to specific cell types. Following the observation of variations among DNA-Diffusion sequences, particularly with a subset showing outstanding performance in signal strength and specificity, we decided to investigate the defining features of these top-performing sequences further. We carried out a motif enrichment analysis to compare the sequences identified as high-signal with those deemed high-specificity (Fig. 6a,b). This analysis aimed to identify overrepresented motifs within each subgroup to understand the sequence characteristics that contribute to their remarkable performance and cell type-specific effectiveness (Fig. 6c).

For GM12878 and HepG2, multiple TF motifs were predominantly observed in sequences classified as high- signal or high-specificity. Notably, in HepG2 sequences, a significant enrichment of the HNF4A motif was found in high-signal DNA-Diffusion sequences, with 82% showcasing this motif. In contrast, high-specificity sequences displayed the HNF4A motif less frequently, at 41%. Conversely, the FOXA1 motif was present in 58% of high-specificity sequences, compared to 48% in high-signal ones, suggesting HNF4A's role in enhancing accessibility in HepG2, while FOXA1 might contribute to achieving greater specificity in this cell type. In GM12878 sequences, immune-related TFs, such as MEF2A (50% in high-specificity vs. 12% in high-signal) and IRF7 (77% vs. 53%), were predominantly found in high-specificity sequences. Contrastingly, SPIB, another immune-associated factor, was more prevalent in high-signal sequences (64% vs. 87%). Additionally, IRF1 showed significant enrichment in both high-signal and high-specificity sequences (96% vs. 97%). Interestingly, for K562 sequences, our analysis suggested fewer TFs (including GATA1-TAL1) are required for generating sequences with high specificity and signal intensity. However, our framework couldn't distinguish between the two categories solely based on motif composition. The average motif hits for the aforementioned TFs also increased in their respective categories compared to others and endogenous training sequences. For instance, HNF4A had an average of 1.195 hits in high-signal DNA-Diffusion sequences, 0.49 hits in high-specificity sequences, and 0.32 hits in endogenous training sequences (Fig. 6c).

Further investigation of TF expression across cell types revealed that many TFs enriched in high-specific and high-signal sequences are predominantly expressed in a cell type-specific manner (Supplementary Fig. 9a). These findings suggest a sequence-based regulatory logic distinguishing these two categories of sequences, underscoring the nuanced relationship between TF motifs and their regulatory impact in specific cellular contexts.

7 SUPPLEMENTARY TABLES

Table 1: Cell type-specific genes and regulatory regions selected for each cell type

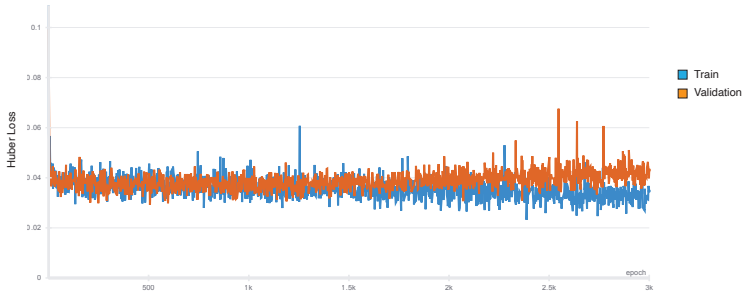
Gene	**Enhancer Region**	**Gene Region**	**Cell Type**
GATA1	chrX:48782929-48783129	chrX:48785536-48787536	K562
HNF4A	chr20:44370692-44370892	chr20:44355699-44432845	HepG2
CD19	chr16:28930777-28930977	chr16:28931971-28939342	GM12878

Table 2: Genes tested for reactivation for each cell type

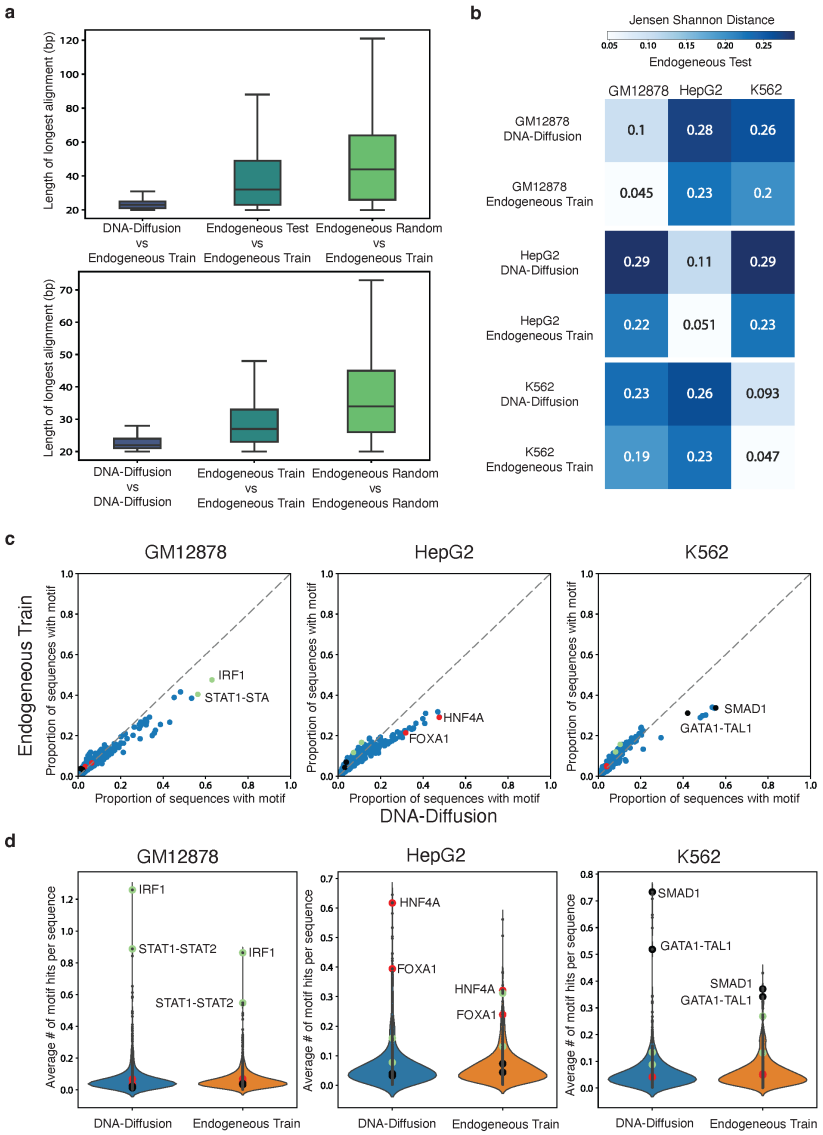
Cell Type	Gene
GM12878	GATA1 HNF4A
K562	HNF4A CD19
HepG2	CD19 GATA1

Table 3: ENCODE IDs of samples use for training the MPRA predictor model

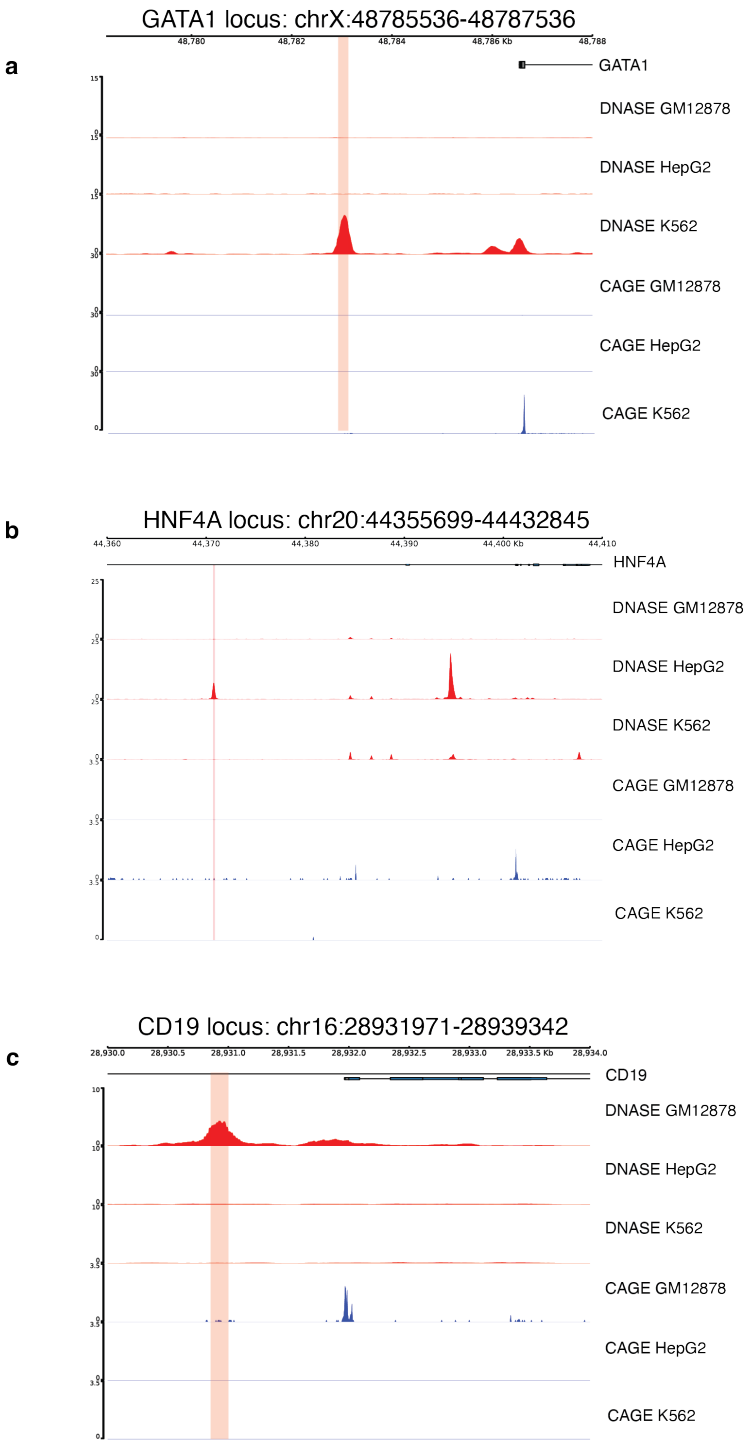
ENCFF996ECA
ENCFF018AMJ
ENCFF345ASG
ENCFF970OLE
ENCFF318XMJ
ENCFF821XQZ
ENCFF358MBK
ENCFF379XWL
ENCFF774CHX
ENCFF138DJM
ENCFF277DDE
ENCFF334EKU
ENCFF857FQR
ENCFF259NMG
ENCFF477LDL
ENCFF484JFE
ENCFF227KRF
ENCFF102ZVT
ENCFF418GRL
ENCFF333BAD
ENCFF307HBZ
ENCFF771HPB
ENCFF359KJL
ENCFF035HKU
ENCFF759PPO
ENCFF705AES
ENCFF256WKS
ENCFF352JAC
ENCFF147SMK
ENCFF311DJW
ENCFF350IJA
ENCFF815ORW
ENCFF402GOL
ENCFF865LNO
ENCFF755GRH
ENCFF440YVF
ENCFF703OIL
ENCFF927USI
ENCFF476FXK
ENCFF742ENC
ENCFF112HAT
ENCFF792IHA
ENCFF267VJ



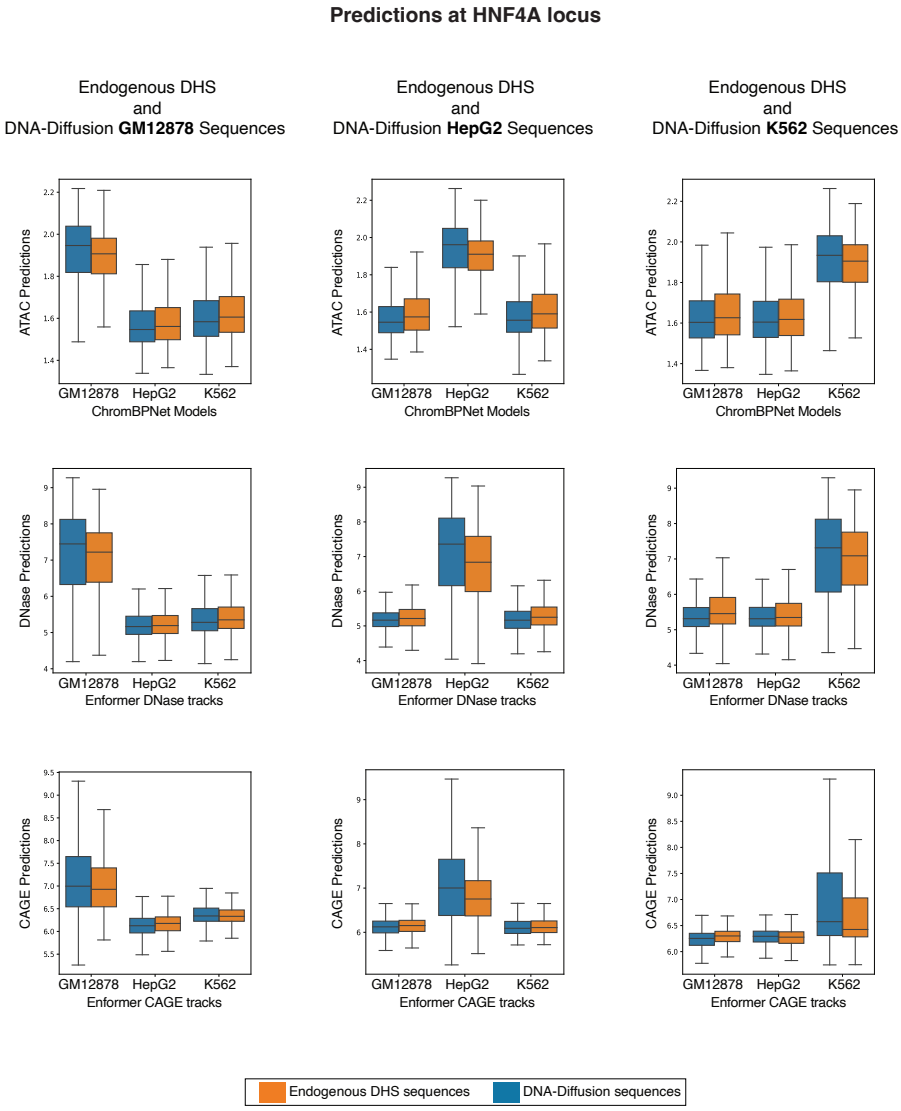
Supplementary Figure 1: DNA-Diffusion training loss plot



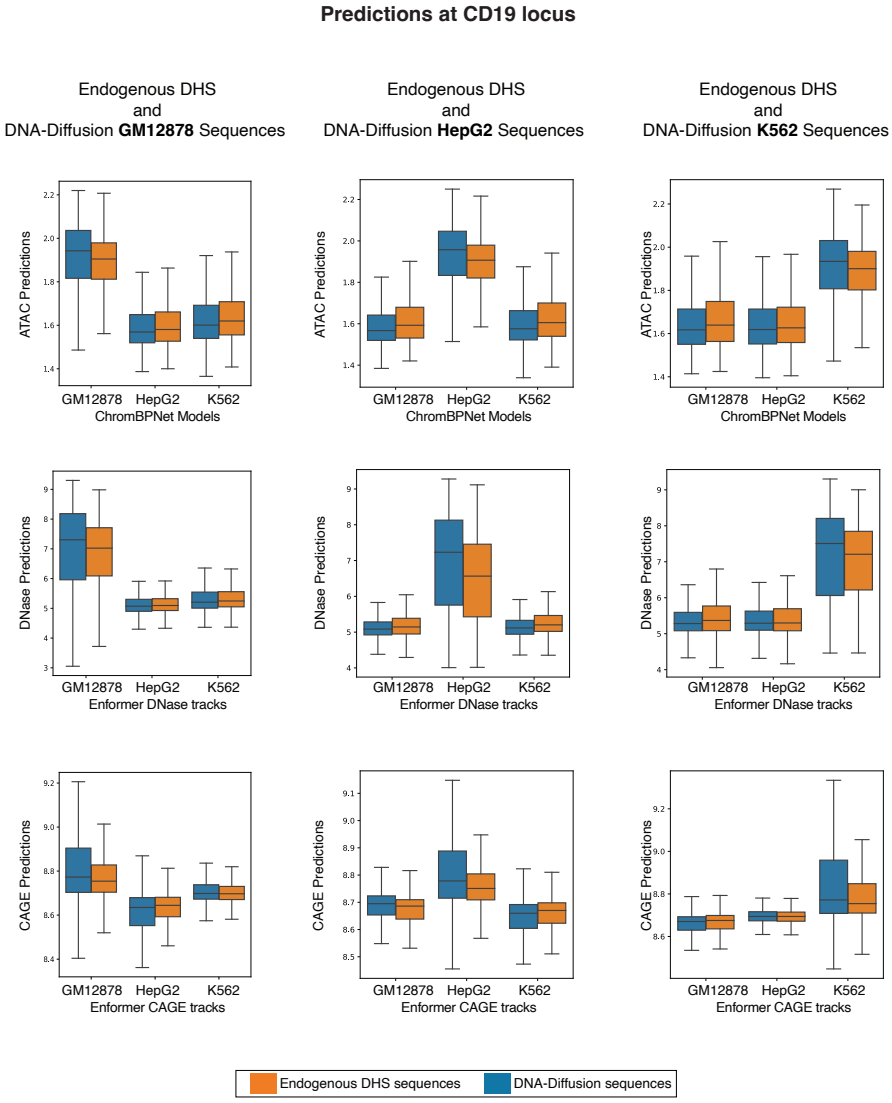
Supplementary Figure 2: a) Top, bar plots displaying sequence alignment length comparing the amount of overlap between DNA- Diffusion, endogenous test, and endogenous random sequences against endogenous train sequences. Bottom, distribution of alignment length within each group. b) Jensen Shannon distance comparing TF motif hit distributions between DNA-Diffusion sequences and endogenous train sequences against the endogenous test sets across GM12878, HepG2, and K562 cell-types. c) Scatterplot showing the proportion of sequences with a given TF motif in endogenous training sequences vs DNA-Diffusion sequences across GM12878, HepG2, and K562 cell-types. Important cell type- specific TFs are annotated based on existing literature (green: GM12878, red: HepG2, black: K562). d) Violin plot showing the average number of TF motif hits per sequence in endogenous training sequences vs DNA-diffusion sequences across GM12878, HepG2, and K562 cell-types. Important cell type-specific TFs are annotated based on existing literature (green: GM12878, red: HepG2, black: K562).



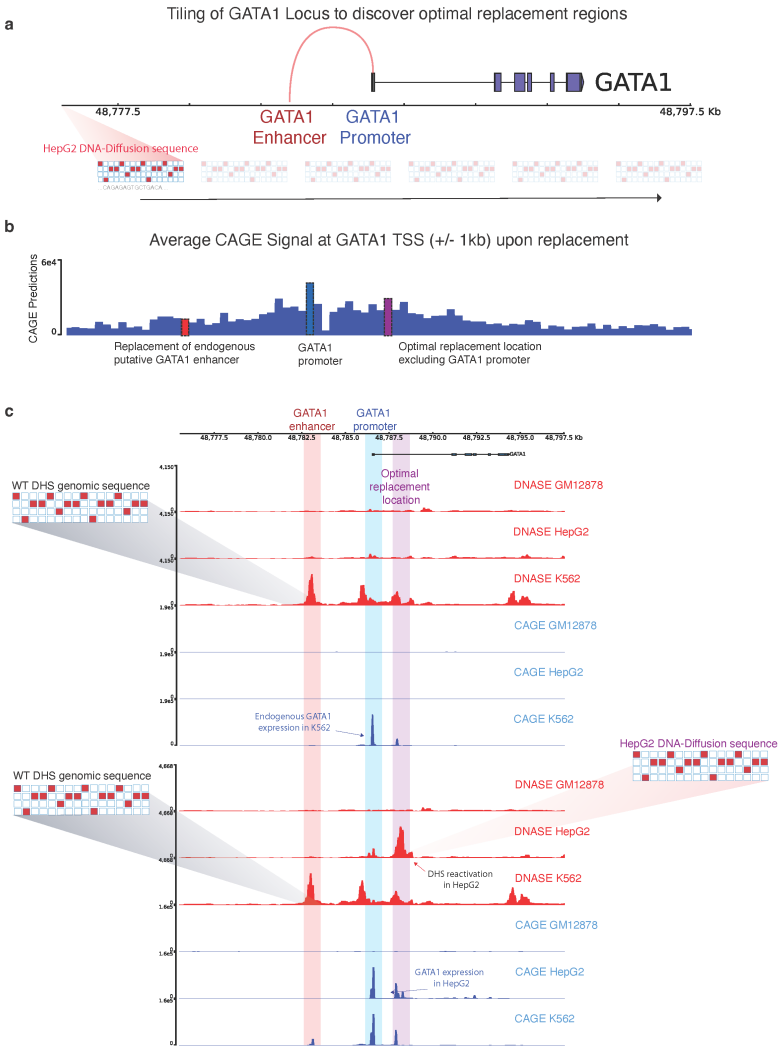
Supplementary Figure 3: Endogenous DNase and CAGE Tracks. Chromatin accessibility (ENCODE DNase, red track) and gene expression (FANTOM CAGE, blue track) profiles are shown for three key gene loci. Each panel displays a window in the genomic region potentially controlling a specific gene (GATA1, HNF4A, CD19).



Supplementary Figure 4: Endogenous train and DNA-Diffusion predictions across all three downstream oracles within the HNF4A gene locus. a) Boxplots showing the predicted chromatin accessibility/gene expression activity (ChromBPNet ATAC, Enformer DNase, Enformer CAGE) upon replacement within the HNF4A locus with endogenous DHS train and DNA-Diffusion sequences specific for each cell line (GM12878, HepG2, K562).



Supplementary Figure 5: Endogenous train and DNA-Diffusion predictions across all three downstream oracles within the CD19 gene locus. a) Boxplots showing the predicted chromatin accessibility/gene expression activity (ChromBPNet ATAC, Enformer DNase, Enformer CAGE) upon replacement within the CD19 locus with endogenous DHS train and DNA-Diffusion sequences specific for each cell line (GM12878, HepG2, K562).



Supplementary Figure 6: a) Tiling approach for exploring the optimal location of replacement of a synthetic sequence to reactivate a gene of interest. b) The average CAGE predictions for the GATA1 TSS+/- 1KB are shown for each replacement of the same HepG2-specific sequence throughout the entire locus. Key regions with strong predicted functional effects on GATA1 reactivation are highlighted; the original GATA1 enhancer in K562 (red), the GATA1 promoter (blue), and the optimal replacement region discovered by the tiling procedure (purple). c) Top: DNase and CAGE activity predictions at the GATA1 locus based on the wild-type sequence. Bottom: post-insertion of a HepG2-optimized DNA-Diffusion sequence into the optimal replacement location, resulting in heightened chromatin accessibility and GATA1 expression in HepG2, maintenance of accessibility and GATA expression in K562 and no significant changes in GM12878.