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[1-3], Aniketh Janardhan Reddy[4], Sameer Gabbita[2,5], Zach Nussbaum[6], César Miguel Valdez Córdova[7], Aaron Wenteler[8], Noah Weber[9], Tin M. Tunjic[9], Martino Mansoldo[10], Talha Ahmad Khan[10], Zelun Li[11,12], Cameron Smith[1-3], Matei Bejan[13], Lithin Karmel Louis[11,12], Paola Cornejo[11,12], Will Connell[10], Emily S. Wong[11,12], Wouter Meuleman[14,15], Luca Pinello[1-3†]

[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]Department of Electrical Engineering and Computer Sciences, University of California, Berkeley, CA, USA
[5]Johns Hopkins University, Baltimore, MD, USA
[6]Nomic AI
[7]Johannes Kepler University, Linz, Austria
[8]Queen Mary University of London, London, UK
[9]TU Vienna, Austria
[10]Independent Researcher
[11]Victor Chang Cardiac Institute, Darlinghurst, New South Wales, Australia
[12]School of Biotechnology and Biomolecular Sciences, Faculty of Science, UNSW Sydney, Sydney, Australia
[13]University of Bucharest, Bucharest, Romania
[14]Altius Institute for Biomedical Sciences, Seattle, WA, USA
[15]Paul G. Allen School of Computer Science & Engineering, University of Washington, Seattle, WA, USA
[*] 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 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.

Introduction

The systematic modification and optimization of regulatory elements to control gene expression is one of the key challenges in modern genomics and synthetic biology. This process offers the potential to correct disease-related misregulation and to direct cells to specific functional states. Large consortia such as ENCODE[1–3], Roadmap Epigenomics[4], Blueprint[5], FANTOM[6], and others have uncovered the complexity of gene regulation and provide rich data sources for learning about regulatory element features. The field of generative Artificial Intelligence (AI) has advanced tremendously over the last few years, yielding approaches that enable researchers to discover, represent and generate patterns in biological data unlike ever before. Such approaches have great potential for designing synthetic sequences and identifying genomic locations to integrate them with the goal of safely and precisely modulating gene expression.

Results

In this study, we propose to use diffusion probabilistic models to design context-specific DNA regulatory sequences that have the potential to modify gene expression and be employed in new therapeutic applications that require precise perturbation of gene regulation. Diffusion models have shown remarkable performance in generating audio, pictures (Stable Diffusion[7]), 3D objects (DreamFusion[8]), and proteins (RFdiffusion[9]) (Fig. 1a,b). Our framework utilizes the DHS index dataset curated by Meuleman et al.[10], 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 carcinoma cell line) for liver biology and disease studies. A stratified chromosome sampling strategy proposed by Meuleman et al.[11] was utilized to partition the dataset into mutually exclusive subsets for testing (chr1), validation (chr2), and training (remaining chromosomes). Training of the model consisted of transforming endogenous DHS sequences into a hot-encoded format and introducing a fixed amount of standard normal noise. By learning to predict the introduced noise, the model, whose backbone is based on the U-Net[12] architecture, can then create new cell type-specific sequences from randomly initialized

noise. Utilizing the trained model a total of 100,000 DNA-Diffusion sequences per cell type were generated for downstream validation.

In this work, we used a spectrum of metrics to ensure our sequences are diverse while retaining key properties of endogenous sequences with respect to potential binding specificity, composition, accessibility, and regulatory potential.

First, we assessed transcription factor (TFs) binding site composition by comparing cell type-specific DNA-Diffusion sequences with endogenous sequences using the Jensen-Shannon divergence between their TFs binding probability vectors in three cell types (Fig. 2a). Comparison of TF motif composition reveals significant differences between DNA-Diffusion and endogenous sequences, yielding an average JS divergence of 0.101 across the three cell types when compared to the average mean distance between endogenous train and test of 0.048. This shift, driven by modulated motif density and inclusion of known cell type-specific motifs, demonstrates the model's ability to enhance cell type specificity during sequence generation.

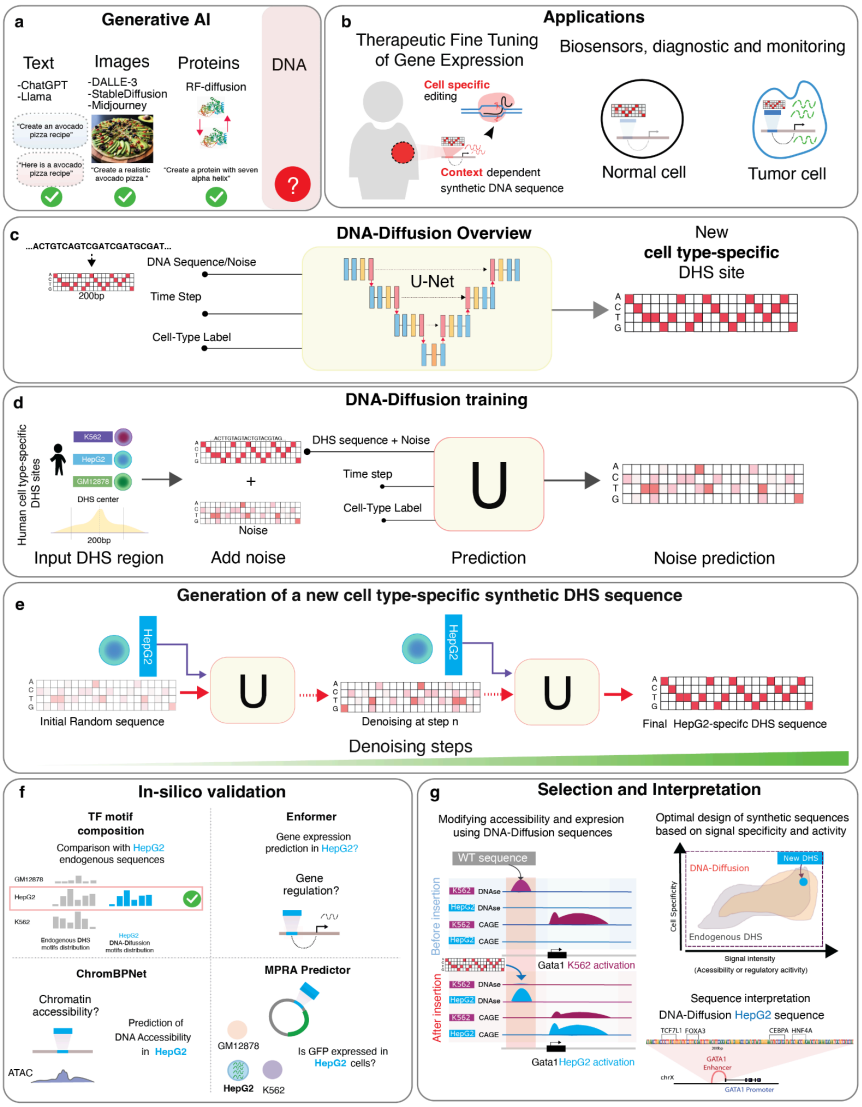

Figure 1. a,b,c,d,e) Proposed framework for training and sampling from DNA-Diffusion along with potential therapeutic applications f,g) In-silico and selection scheme for sequences

Second, we evaluated the cell type-specific chromatin accessibility potential of these sequences by using ChromBPnet[13], a state-of-the-art model to predict chromatin accessibility from DNA sequences. As before, we considered three classes of cell type-specific sequences generated by the DNA-Diffusion model and assessed their chromatin accessibility patterns using three different ChromBPnet models. Similar to the motif composition evaluation, sequences generated within a specific cell type context scored higher than those generated across cell type contexts, according to the three ChromBPnet models. Focusing on the GATA1 locus, known for its cell type-specific chromatin accessibility in K562, the endogenous DHS sites from the training set had a mean log-normalized predicted ATAC value of 1.88 in the K562 ChromBPNet model, compared to 1.65 for the HepG2 model and 1.66 for the GM12878 model. Notably, the predicted accessibility values of the K562-specific DNA-Diffusion sequences showed a value of 1.91 for the K562 model, suggesting a 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 cell type models, both showing predicted values of 1.64. Similar to the observations in K562 cells, the DNA-Diffusion sequences for HepG2 and GM12878 showed stronger signals in their respective ChromBPNet models compared to predictions in other cell types. These results indicate that our generated sequences are detected as accessible regions in a cell context-specific manner and present the same range of in-silico accessibility as endogenous sequences.

Finally, to evaluate the capacity of our DNA-Diffusion sequences to activate gene expression in different cell contexts, we use the Enformer[14] model, a transformer-based deep learning model for predicting properties of DNA sequences including chromatin histone modification and CAGE information in a cell type-specific manner using genomic sequence as the sole input. 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. In the GATA1 locus, Enformer DNase prediction analysis showed that DNA-Diffusion sequences specific for K562 maintained or enhanced the predicted accessibility in the GATA1 enhancer compared to the native sequence, while showing no activity for GM12878 and HepG2, consistent with the ChromBPNet analysis (Fig. 2c).

Similarly when introducing cell type-specific HepG2 and GM12878 DNA-Diffusion sequences (Fig. 2c), we observe a statistically significant (p-value < 0.001, single-sided t-test) increase in accessibility within previously non-activated GATA1 DHS regulatory regions in those cell lines. To assess the impact of these sequences on cell type-specific expression of GATA1, we analyzed Enformer predictions of CAGE within the 1kb region proximal to the TSS. GM12878 DNA-Diffusion sequences replacing the GATA1 regulatory region showed a mean CAGE signal of 8.83 in GM12878, a statistically significant increase in activity compared to the best endogenous GM12878 DHS training sequence (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). Fig. 2b shows the impact of the insertion of a HepG2 DNA-Diffusion sequence in a GATA1 enhancer region. The Enformer prediction demonstrates that the nearby gene, in this case GATA1, increases expression as shown by the CAGE HepG2 track.

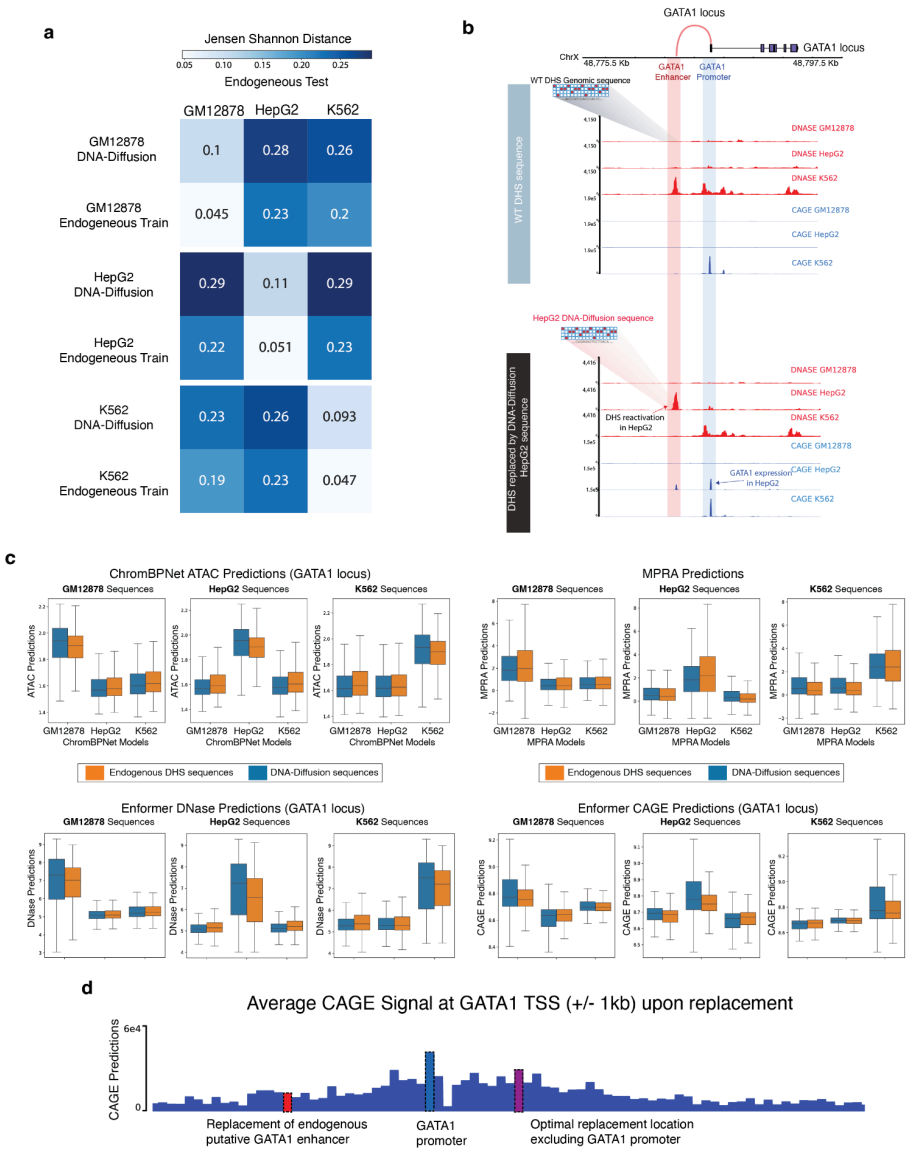

Figure 2 a) Cell-type specific TF binding site motif composition. b) Reactivation of GATA1 in HepG2 by a DNA-Diffusion sequence. c) In silico predictions of ATAC, MPRA, DNase, and CAGE signal in the GATA1 locus d) Optimal placement of DNA-Diffusion sequence for maximal GATA1 expression in HepG2.

As a way to define genomic regions susceptible to synthetic element insertion, we mapped the changes in GATA1 expression mediated by a HepG2 DNA-Diffusion sequence for all the possible insertion locations within the GATA1 locus (Fig. 2d). This tiling approach was able to detect a DNase-accessible region ~1kb apart from the GATA1 promoter that demonstrates the potential to regulate GATA1 expression. In addition to this DHS region, some GATA1 intronic regions demonstrated differential impact on GATA1 expression. The tiling approach 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.

Discussion

DNA-Diffusion presents the first end-to-end solution to sequence design that requires no external model guidance nor orchestration of different models to generate cell type-specific sequences. While it presents promising in silico results it will be paramount to have robust experimental techniques that can replace endogenous sequences with generated ones to elucidate the true functional potential of synthetically designed sequences.

Further exploration into scalable DNA diffusion models presents the potential to revolutionize synthetic biology, enabling precise perturbation of cell type-specific gene regulation and advancing precision gene therapies. This could facilitate highly efficient, targeted treatments for disorders while improving our understanding of the relationship between sequence content, context, and their impact on gene expression.

References

1. ENCODE Project Consortium. An integrated encyclopedia of DNA elements in the human genome. *Nature* **489**, 57–74 (2012).
2. Luo, Y. *et al.* New developments on the Encyclopedia of DNA Elements (ENCODE) data portal. *Nucleic Acids Res* **48**, D882–D889 (2020).
3. Hitz, B. C. *et al.* The ENCODE Uniform Analysis Pipelines.
4. Kundaje, A. *et al.* Integrative analysis of 111 reference human epigenomes. *Nature* **518**, 317–330 (2015).
5. Martens, J. H. A. & Stunnenberg, H. G. BLUEPRINT: mapping human blood cell epigenomes. *Haematologica* **98**, 1487–1489 (2013).
6. Noguchi, S. *et al.* FANTOM5 CAGE profiles of human and mouse samples. *Sci Data* **4**, 170112 (2017).
7. Rombach, R., Blattmann, A., Lorenz, D., Esser, P. & Ommer, B. High-Resolution Image Synthesis with Latent Diffusion Models. in *2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)* 10674–10685 (IEEE, New Orleans, LA, USA, 2022). doi:10.1109/CVPR52688.2022.01042.
8. Poole, B., Jain, A., Barron, J. T. & Mildenhall, B. DreamFusion: Text-to-3D using 2D Diffusion. Preprint at https://doi.org/10.48550/arXiv.2209.14988 (2022).
9. Watson, J. L. *et al.* De novo design of protein structure and function with RFdiffusion. *Nature* **620**, 1089–1100 (2023).
10. Meuleman, W. *et al.* Index and biological spectrum of human DNase I hypersensitive sites. *Nature* **584**, 244–251 (2020).
11. Meuleman, W. Synthetic DNA sequences. *meuleman.org* https://www.meuleman.org/research/synthseqs/ (2018).
12. Ronneberger, O., Fischer, P. & Brox, T. U-Net: Convolutional Networks for Biomedical Image Segmentation. Preprint at https://doi.org/10.48550/arXiv.1505.04597 (2015).
13. Pampari, A. *et al.* Bias factorized, base-resolution deep learning models of chromatin accessibility reveal cis-regulatory sequence syntax, transcription factor footprints and regulatory variants. Zenodo https://doi.org/10.5281/ZENODO.7567627 (2023).
14. Avsec, Ž. *et al.* Effective gene expression prediction from sequence by integrating long-range interactions. *Nat Methods* **18**, 1196–1203 (2021).