MULTI-MODAL SINGLE-CELL FOUNDATION MODELS VIA DYNAMIC TOKEN ADAPTATION

Wenmin Zhao, Ana Solaguren-Beascoa, Grant Neilson, Louwai Muhammed, Liisi Laaniste, and Sera Aylin Cakiroglu
Cosyne Therapeutics
London, UK
{millie, ana, grant, louwai, liisi, aylin}@cosyne.com

ABSTRACT

Recent advances in applying deep learning in genomics include DNA-language and single-cell foundation models. However, these models take only one data type as input. We introduce dynamic token adaptation and demonstrate how it allows combining these models to predict gene regulation at single-cell level in different genetic contexts. Although the method is generalisable, we focus on an illustrative example by training an adapter from DNA-sequence embeddings to a single-cell foundation model's token embedding space. As qualitative evaluation, we assess the impact of DNA sequence changes on the model's learned gene regulatory networks by mutating the transcriptional start site of the transcription factor *GATA4 in silico*, observing predicted expression changes in its target genes in fetal cardiomyocytes.

1 INTRODUCTION

There have been rapid advances in training single-cell foundation models on large single-cell RNA sequencing (scRNA-seq) datasets (Theodoris et al., 2023; Yang et al., 2022; Cui et al., 2023). These models represent each gene in a single-cell transcriptome as an input text token and do not integrate genetic information, making it hard to interpret gene expression predictions under genetic changes. On the other hand, DNA language models have been trained to predict epigenetic signals from the DNA sequence (Kelley et al., 2018; Kelley, 2020; Avsec et al., 2021b;a), and can be fine-tuned to predict the expression values of individual genes across cells in a scRNA-seq dataset (Schwessinger et al., 2023). However, these models do not take cell-level co-regulation into account when predicting a gene's expression, instead, they focus on predicting epigenetic signals or the expression for each gene separately.

Our method of combining the modelling of both DNA sequences and single-cell transcriptomics data is inspired by unified embedding architectures for multi-modal large language models (LLMs), which convert an image into embedding vectors as a set of special tokens that are prepended to the input text tokens (Cho et al., 2021). Similar approaches using an adapter to provide additional information to the transcriptome have recently been applied to single-cell models (Maleki et al., 2025; Levine et al., 2023), however, these methods have been restricted to a few additional tokens encoding a single entity (e.g. cell-type, disease state, molecule of drug treatment).

In this paper, we propose extending the approach to all tokens in the input to allow their embeddings to flexibly encode additional information from a different modality that may change between data samples, which we call dynamic token adaptation (DTA). As an application, we introduce Bio-DTA, a novel multi-modal model that learns from single-cell transcriptomes and DNA sequences jointly. Finally, we demonstrate that the model has learned dynamic co-regulation by assessing the impact of genetic changes to the DNA sequence of the transcription factor *GATA4 in silico* on the model's predictions for its targets.

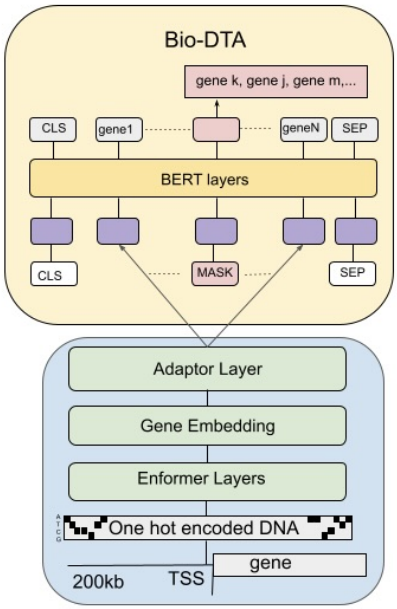

Figure 1: Schematic of Bio-DTA. Fixed token embeddings are replaced with a projection of aggregated Enformer embeddings of the gene's DNA sequence.

2 METHODS

Training data We downloaded scRNA-seq data from $33,364,242$ unique cells across 265 datasets in the census dataset (version 2023-07-25) from the CellXGene data portal (CZI Single-Cell Biology Program et al., 2023). Data processing followed Theodoris et al. (2023), representing each single-cell transcriptome as a sequence of gene names of maximum length $2,048$ ordered by their median-normalised expression. We excluded cancer cells and cells with < 500 expressed genes. Transcriptional start sites (TSSs) of all protein coding genes were obtained from Ensembl (GRCh38.108). For each gene, $196,608$bp around the TSS of the reference genome was inputted to Enformer to compute the mean embedding of dimension $3,072$ across the positions of the pooling convolutional filters. The Enformer checkpoint was obtained from https://github.com/lucidrains/enformer-pytorch.

Model architecture and training Bio-DTA combines a DNA language model with a single-cell foundation model via token adapters. Although our method is flexible and can accommodate other architectures, in the experiments presented here, the single-cell foundation model is based on a bidirectional transformer encoder-only architecture (BERT). The model receives a single-cell transcriptome of length $2,048$ as an ordered sequence of gene names as input, which are mapped to integer identifiers called token IDs (Theodoris et al., 2023). In a usual BERT model, each token ID is mapped to a unique and trainable embedding vector that forms the input to the transformer encoder (Devlin et al., 2019). In contrast, Bio-DTA (Figure 1) projects Enformer's aggregated embeddings for each input gene to the token embedding size using an adapter layer (e.g. a multilayer perceptron followed by a softplus activation). This forms the input to the transformer encoder capturing genetic information. Here, a gene's token embedding is not unique and may change if the gene's input DNA sequence changes, and the same adapter is used for all the genes. Bio-DTA outputs the token IDs of the input gene names according to the single-cell foundation model.

Bio-DTA was trained end-to-end with a masked language modelling task (masking 15% of input tokens) for three epochs. For more implementation details as well as hyperparameters used for the final model and training procedure see Appendix A.1.

3 IN SILICO MUTAGENESIS REVEALS LEARNED CONNECTION **DYNAMICS** IN BIO-DTA

To assess how changes in the input DNA sequence of one gene affects learned co-regulation networks, we followed Theodoris et al. (2023) and focussed on *GATA4* and *TBX5*, two known congenital heart disease genes, which are co-expressed during cardiac morphogenesis, physically interact and have co-bound targets (Misra et al., 2014). We obtained 103 transcriptomes from fetal cardiomyocytes expressing *GATA4* (Knight-Schrijver et al., 2022). We then introduced random mutations *in silico* in 100bp around the *GATA4* TSS, which reduced Enformer's predicted expression (Figure 2). For each single-cell transcriptome in the dataset, we performed a forward pass and extracted the contextualised embedding for each input gene from the penultimate layer of Bio-DTA, once when using the unchanged *GATA4* embedding, and once when using the mutated embedding as input for *GATA4*. As the retained gene embeddings are from deeper layers of Bio-DTA's BERT encoder, they may capture information about other co-regulated genes in the input sequence. To test if these depend on genetic changes, we calculated the cosine similarity between the gene embeddings before and after *GATA4*'s *in silico* mutagenesis for each transcriptome. Indeed, the embeddings of experimentally identified GATA4 and TBX5 targets (defined based on ChIP-seq data in Theodoris et al. (2023)) had a significant drop in cosine similarity compared with the remainder of the genome ($p < 0.05$, Wilcoxon test), while embeddings of housekeeping genes remained stable (Table 1).

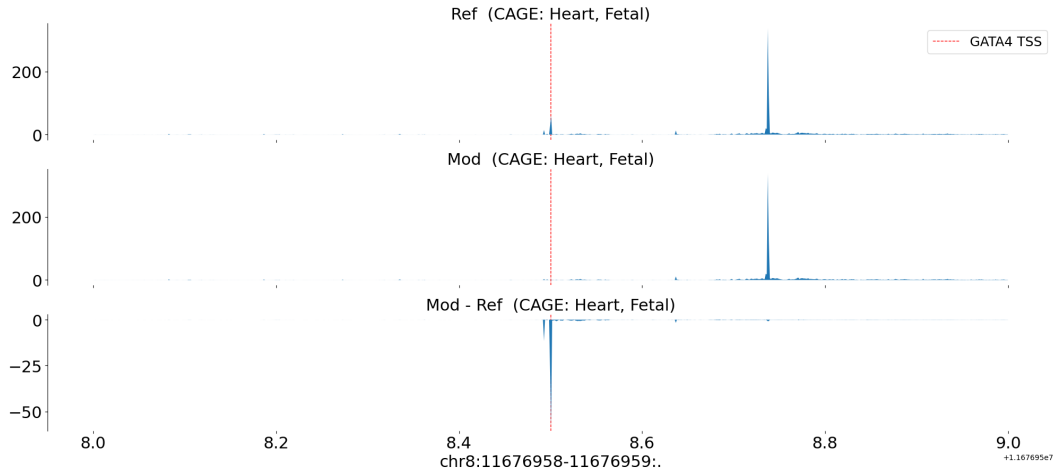

Figure 2: Enformer predicted Cap analysis of gene expression (CAGE) tracks of gene expression centered around the *GATA4* TSS for the reference genome (top), *in silico* mutated sequence (random sequence replacing 100bp around TSS) (middle), and the difference in predictions between the *in silico* mutated and reference sequence showing a large decrease around the TSS in the middle of the track (bottom).

Next, we selected the 50 genes with the largest changes in their embeddings and assessed the model's capability to identify experimentally derived targets compared with a random gene set of the same size. To compare the performance of Bio-DTA with Geneformer, we perform *in silico* deletion of GATA4 by removing it from the input sequence as described in Theodoris et al. (2023). We also train a BERT model similar to Geneformer with the same parameters as Bio-DTA that does not use adapters and instead uses the gene names as input sequence on the same training data. We perform *in silico* deletion of GATA4 for this model as described for Geneformer.

Figure 3 shows the precision and recall of true targets for each of the models in each group. Bio-DTA performs best on both metrics for direct GATA4 and TBX5 and their co-bound targets, and is comparable to the adapter-free model on indirect GATA4 targets. However, it is outperformed by the adapter-free model on indirect TBX5 targets. Across all groups and metrics, Geneformer is outperformed by both models and sometimes the random gene set baseline. In particular, it can not recover any experimentally verified GATA4 targets. A similar trend can be seen using the top 100 genes (Figure A.1). Taken together, our results demonstrate how the dynamic token embeddings

	P-value	FDR adjusted P-value
GATA4 direct	6.42e-10	**3.85e-09**
GATA4 indirect	3.41e-07	**5.11e-07**
GATA4 & TBX5 combination	1.51e-09	**4.52e-09**
TBX5 direct	2.64e-09	**5.27e-09**
TBX5 indirect	5.75e-07	**6.90e-07**
housekeeping genes	7.24e-01	7.24e-01

Table 1: P-values of a Wilcoxon test comparing the cosine similarities of the gene embeddings before and after *in silico* mutagenesis of *GATA4* in a target group with those of the remainder of the genome not in any of the target groups. Significant FDR-adjusted p-values at a significance threshold of 0.05 are highlighted in bold. Target groups as indicated were obtained from Theodoris et al. (2023).

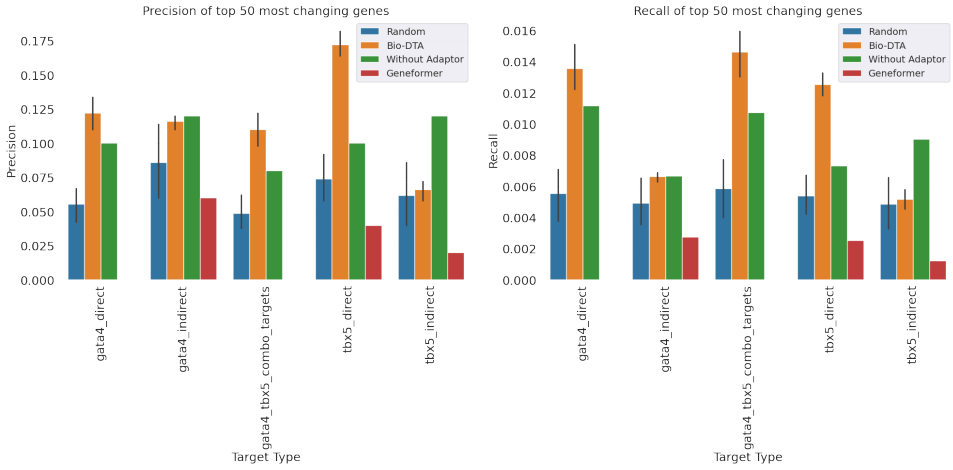

Figure 3: Barplots of precision and recall of putative GATA4 and TBX5 targets (orange) based on contextualised embedding changes under *in silico* mutagenesis of *GATA4* in fetal cardiomyocytes, comparing with the in-silico deletion of *GATA4* in the BERT model without an adapter (green), Geneformer (red) and size-matched random samples (blue). Target group annotations based on ChIP-seq data were obtained from (Theodoris et al., 2023).

allow the model to be sensitive to small genetic changes and their impact on its learned co-regulation networks.

4 Conclusion

We introduced dynamic token adaption to project embeddings of a different data modality into the text token embeddings of a foundation model. As an application, we presented Bio-DTA, a multi-modal model that combines a DNA language model with a single-cell transcriptome model.

We showed that Bio-DTA learns the impact of small genetic changes on co-regulation networks. We evaluated the model's ability to capture gene co-regulation by performing *in silico* mutagenesis of the GATA4 transcription factor and observing changes in the contextualized embeddings of target genes in fetal cardiomyocytes. Next, we showed that Bio-DTA successfully returns experimentally verified targets amongst the 50 genes with the largest changes in their embeddings. We compared these results with an adapter-free model that was trained on the same data with the same hyperparameters. We showed that Bio-DTA outperforms the adapter-free model on direct GATA4 and TBX5 targets. This may be due to the method of *in silico* deletion completely removing GATA4 from the

input sequence, which changes the entire input transcriptome and removes learned connections between GATA4 input embeddings and its targets. In contrast, Bio-DTA uses small changes to the embedding of GATA4 to signal reduced expression without changing the overall input and allows the assessment of these changes on learned connections. On the other hand, the adapter-free model outperforms Bio-DTA on indirect targets. This may be due to the complete removal of GATA4's input during the *in silico* deletion, resulting in a much larger change to the input, which may influence the contextualised embeddings more in the adapter-free model.

In future work, we plan to extend our evaluation beyond the GATA4 case study by exploring additional transcription factors and cell types to validate the generalisability of Bio-DTA further. However, identifying suitable benchmark cases for such evaluation remains a significant challenge. While there are many examples of the impact of genetic changes on one gene's expression (for example, eQTLs), high-quality, well-characterized instances where both the genetic perturbation and its downstream co-regulation impact are experimentally validated — especially in a cell-type-specific context — are rare. This limits the availability of ground truth data against which to benchmark model predictions. Furthermore, we will expand our evaluations to more complex genetic variations such as SNPs, indels, and structural variants which introduce additional biological layers that will provide a more comprehensive evaluation of the model.

In the work presented here, the reference genome was used as input to Enformer and *in silico* mutagenesis was performed as a zero-shot approach. DNA language models trained on the reference genome such as Enformer struggle to reliably predict the direction of eQTLs and the expression variation for different individuals (Huang et al., 2023; Sasse et al., 2023). For a personalised medicine approach, future work will also include fine-tuning Enformer on different genomes with allelic information as described in Drusinsky et al. (2024) to provide more nuanced DNA sequence embeddings.

Other applications of DTA will include projecting embeddings of the gene's RNA isoforms or amino acid sequence into the token embedding space. While we mapped the DNA sequence to one token in the presented experiments, we will also evaluate using several tokens per input gene representing the DNA sequence to allow the encoding of larger genetic contexts.

REFERENCES

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 1548-7091, 1548-7105. doi: 10.1038/s41592-021-01252-x. URL https://www.nature.com/articles/s41592-021-01252-x.

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 1061-4036, 1546-1718. doi: 10.1038/s41588-021-00782-6. URL https://www.nature.com/articles/s41588-021-00782-6.

Jaemin Cho, Jie Lei, Hao Tan, and Mohit Bansal. Unifying Vision-and-Language Tasks via Text Generation, 2021. URL https://arxiv.org/abs/2102.02779. Version Number: 2.

Haotian Cui, Chloe Wang, Hassaan Maan, Kuan Pang, Fengning Luo, and Bo Wang. scGPT: Towards Building a Foundation Model for Single-Cell Multi-omics Using Generative AI. preprint, Bioinformatics, May 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.04.30.538439.

CZI Single-Cell Biology Program, Shibla Abdulla, Brian Aevermann, Pedro Assis, Seve Badajoz, Sidney M. Bell, Emanuele Bezzi, Batuhan Cakir, Jim Chaffer, Signe Chambers, J. Michael Cherry, Tiffany Chi, Jennifer Chien, Leah Dorman, Pablo Garcia-Nieto, Nayib Gloria, Mim Hastie, Daniel Hegeman, Jason Hilton, Timmy Huang, Amanda Infeld, Ana-Maria Istrate, Ivana Jelic, Kuni Katsuya, Yang Joon Kim, Karen Liang, Mike Lin, Maximilian Lombardo, Bailey Marshall, Bruce Martin, Fran McDade, Colin Megill, Nikhil Patel, Alexander Predeus, Brian Raymor, Behnam Robatmili, Dave Rogers, Erica Rutherford, Dana Sadgat, Andrew Shin, Corinn

Small, Trent Smith, Prathap Sridharan, Alexander Tarashansky, Norbert Tavares, Harley Thomas, Andrew Tolopko, Meghan Urisko, Joyce Yan, Garabet Yeretssian, Jennifer Zamanian, Arathi Mani, Jonah Cool, and Ambrose Carr. CZ CELL×GENE Discover: A single-cell data platform for scalable exploration, analysis and modeling of aggregated data. preprint, Cell Biology, November 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.10.30.563174.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding, May 2019. URL http://arxiv.org/abs/1810.04805. arXiv:1810.04805 [cs].

Shiron Drusinsky, Sean Whalen, and Katherine S. Pollard. Deep-learning prediction of gene expression from personal genomes, July 2024. URL http://biorxiv.org/lookup/doi/10.1101/2024.07.27.605449.

Connie Huang, Richard W. Shuai, Parth Baokar, Ryan Chung, Ruchir Rastogi, Pooja Kathail, and Nilah M. Ioannidis. Personal transcriptome variation is poorly explained by current genomic deep learning models. *Nature Genetics*, 55(12):2056–2059, December 2023. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-023-01574-w. URL https://www.nature.com/articles/s41588-023-01574-w.

David R. Kelley. Cross-species regulatory sequence activity prediction. *PLOS Computational Biology*, 16(7):e1008050, July 2020. ISSN 1553-7358. doi: 10.1371/journal.pcbi.1008050. URL https://dx.plos.org/10.1371/journal.pcbi.1008050.

David R. Kelley, Yakir A. Reshef, Maxwell Bileschi, David Belanger, Cory Y. McLean, and Jasper Snoek. Sequential regulatory activity prediction across chromosomes with convolutional neural networks. *Genome Research*, 28(5):739–750, May 2018. ISSN 1088-9051, 1549-5469. doi: 10.1101/gr.227819.117. URL http://genome.cshlp.org/lookup/doi/10.1101/gr.227819.117.

Vincent R. Knight-Schrijver, Hongorzul Davaapil, Semih Bayraktar, Alexander D. B. Ross, Kazumasa Kanemaru, James Cranley, Monika Dabrowska, Minal Patel, Krzysztof Polanski, Xiaoling He, Ludovic Vallier, Sarah Teichmann, Laure Gambardella, and Sanjay Sinha. A single-cell comparison of adult and fetal human epicardium defines the age-associated changes in epicardial activity. *Nature Cardiovascular Research*, 1(12):1215–1229, December 2022. ISSN 2731-0590. doi: 10.1038/s44161-022-00183-w. URL https://www.nature.com/articles/s44161-022-00183-w.

Daniel Levine, Syed Asad Rizvi, Sacha Lévy, Nazreen Pallikkavaliyaveetil MohammedSheriff, Ruiming Wu, Insu Han, Zihe Zhang, Antonio Fonseca, Xingyu Chen, Sina Ghadermarzi, Amin Karbasi, Rahul M. Dhodapkar, and David Van Dijk. Cell2Sentence: Teaching Large Language Models the Language of Biology. preprint, Bioinformatics, September 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.09.11.557287.

Sepideh Maleki, Jan-Christian Huetter, Kangway V. Chuang, David Richmond, Gabriele Scalia, and Tommaso Biancalani. Efficient Fine-Tuning of Single-Cell Foundation Models Enables Zero-Shot Molecular Perturbation Prediction, April 2025. URL http://arxiv.org/abs/2412.13478. arXiv:2412.13478 [cs].

Chaitali Misra, Sheng-Wei Chang, Madhumita Basu, Nianyuan Huang, and Vidu Garg. Disruption of myocardial Gata4 and Tbx5 results in defects in cardiomyocyte proliferation and atrioventricular septation. *Human Molecular Genetics*, 23(19):5025–5035, October 2014. ISSN 0964-6906, 1460-2083. doi: 10.1093/hmg/ddu215. URL https://academic.oup.com/hmg/article-lookup/doi/10.1093/hmg/ddu215.

Jeff Rasley, Samyam Rajbhandari, Olatunji Ruwase, and Yuxiong He. DeepSpeed: System Optimizations Enable Training Deep Learning Models with Over 100 Billion Parameters. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pp. 3505–3506, Virtual Event CA USA, August 2020. ACM. ISBN 978-1-4503-7998-4. doi: 10.1145/3394486.3406703. URL https://dl.acm.org/doi/10.1145/3394486.3406703.

Alexander Sasse, Bernard Ng, Anna E. Spiro, Shinya Tasaki, David A. Bennett, Christopher Gaiteri, Philip L. De Jager, Maria Chikina, and Sara Mostafavi. Benchmarking of deep neural networks for predicting personal gene expression from DNA sequence highlights shortcomings. *Nature Genetics*, 55(12):2060–2064, December 2023. ISSN 1061-4036, 1546-1718. doi: 10.1038/s41588-023-01524-6. URL https://www.nature.com/articles/s41588-023-01524-6.

Ron Schwessinger, Jacob Deasy, Rob T. Woodruff, Stephen Young, and Kim M. Branson. Single-cell gene expression prediction from DNA sequence at large contexts. preprint, Genomics, July 2023. URL http://biorxiv.org/lookup/doi/10.1101/2023.07.26.550634.

Christina V. Theodoris, Ling Xiao, Anant Chopra, Mark D. Chaffin, Zeina R. Al Sayed, Matthew C. Hill, Helene Mantineo, Elizabeth M. Brydon, Zexian Zeng, X. Shirley Liu, and Patrick T. Ellinor. Transfer learning enables predictions in network biology. *Nature*, 618(7965):616–624, June 2023. ISSN 0028-0836, 1476-4687. doi: 10.1038/s41586-023-06139-9. URL https://www.nature.com/articles/s41586-023-06139-9.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. HuggingFace's Transformers: State-of-the-art Natural Language Processing, July 2020. URL http://arxiv.org/abs/1910.03771. arXiv:1910.03771 [cs].

Fan Yang, Wenchuan Wang, Fang Wang, Yuan Fang, Duyu Tang, Junzhou Huang, Hui Lu, and Jianhua Yao. scBERT as a large-scale pretrained deep language model for cell type annotation of single-cell RNA-seq data. *Nature Machine Intelligence*, 4(10):852–866, September 2022. ISSN 2522-5839. doi: 10.1038/s42256-022-00534-z. URL https://www.nature.com/articles/s42256-022-00534-z.

A Appendix

A.1 Implementation

We used the same BERT-based architecture as in Theodoris et al. (2023) with 6 transformer layers with input size $2,048$, embedding dimensions 256, 4 attention heads per layer and feed-forward size of 512. The adapter is a single MLP layer with a Softplus activation. Hyperparameters were chosen to allow for distributed learning: max learning rate, 1×10^{-3} scaled by the number of GPUs; a learning scheduler, linear with warm-up (10k steps) and linear decay; Adam optimizer with weight decay parameter 0.001. Training was distributed over 4 GPUs in one node with minibatch size 11 and 2 gradient accumulation steps.

To speed up pretraining we used dynamic padding combined with a length-grouped sampler to minimise computation on padding. This sampler takes a randomly sampled megabatch and then orders minibatches by their length in descending order. Mini-batches are then dynamically padded, minimising the computation wasted on padding as sequences of similar lengths are grouped. The authors of Geneformer extended an existing version of this sampler from Huggingface transformers for the distributed case (Theodoris et al., 2023; Wolf et al., 2020). However, neither of these samplers shuffle the mini-batches within the megabatch before passing them to the model, which resulted in a 60x-performance-drop of the trained model in our tests (in terms of training and test perplexity on smaller sample datasets) compared to model runs not employing the grouped-length batching. We implemented a shuffling of the mini batches which slightly diminishes the speed up during training.

For efficient data parallelisation across the GPUS, we used Deepspeed (Rasley et al., 2020). Overall, pre-training was achieved in just over 7 days distributed across one node with four Nvidia A10G 24GB GPUs.

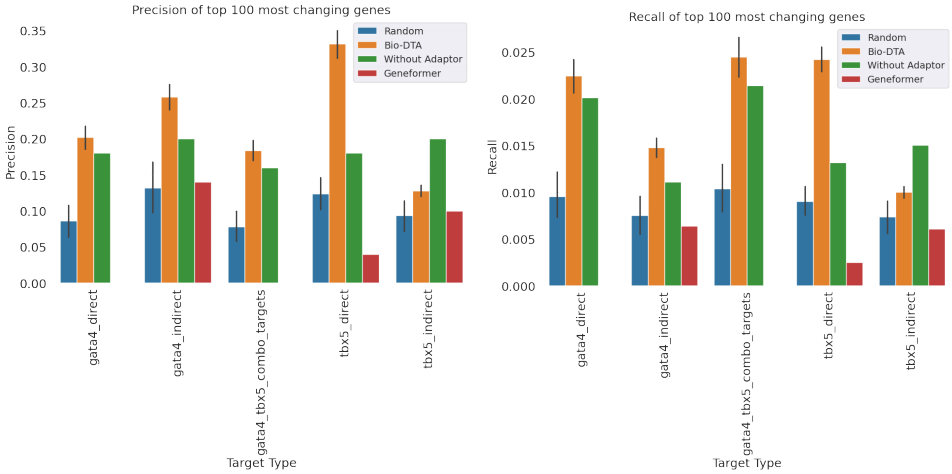

Figure A.1: Barplots of precision and recall of putative GATA4 and TBX5 targets (orange) based on contextualised embedding changes under *in silico* mutagenesis of *GATA4* in fetal cardiomyocytes, comparing with the in-silico deletion of *GATA4* in the BERT model without an adapter (green), Geneformer (red) and size-matched random samples (blue). Target group annotations based on ChIP-seq data were obtained from (Theodoris et al., 2023).