3D Chromosome Structure Reconstruction of *E. coli* Using Machine Learning Methods

1st Xiaofeng Dai
Department of Chemistry
University of Michigan
Ann Arbor, US
xiaofend@umich.edu

***Abstract*—This project applies two machine learning models: REACH-3D (unsupervised) and TECH-3D (transfer learning) to reconstruct the 3D genome structure of *E. coli* from Hi-C data, which captures frequency of pairwise DNA fragments interactions. These models are evaluated for their performance and adaptability in a prokaryotic system. The results will offer insight into the generalizability of the ML model and its practical application in genomic data analysis.**

I. Introduction

The 3D folding of DNA inside cells plays a key role in gene regulation. Hi-C technology captures how often different regions of the genome interact with each other, providing data to infer this structure. Although traditional simulation methods are resource intensive [6], recent machine learning (ML) models offer faster and scalable alternatives.

This project reconstructs the *E. coli* genome in 3D using REACH-3D and TECH-3D and compares their accuracy and adaptability, aiming to assess how well the ML methods generalize across biological systems. The results aim to support broader use of pre-trained ML models in genomic analysis, helping make 3D genome structure data more accessible for research in various fields.

II. Related Workd

Machine learning approaches have shown their potential to reconstruct 3D genome structures from Hi-C data. REACH-3D is an unsupervised recurrent autoencoder that learns 3D representations directly from contact matrices without requiring explicit distance conversion. It preserves genomic order through LSTM-based embeddings and has been successfully applied to eukaryotic cells [2]. TECH-3D takes a supervised approach, using synthetic Hi-C/structure pairs to train a deep encoder. This method allows structure inference even in the absence of ground-truth 3D coordinates, addressing a major limitation in experimental data availability [3].

Despite these advances, most of the prior work has focused on eukaryotic genomes, which differ significantly from bacterial systems in size, organization, and structural constraints. The adaptability of these models to prokaryotic genomes like *E. coli* has not been systematically evaluated. Additionally, while REACH-3D and TECH-3D represent two fundamentally different strategies (unsupervised learning vs. transfer learning), direct comparisons between them on the same dataset are lacking.

This project addresses these gaps by applying both models to the same bacterial Hi-C dataset, allowing a side-by-side evaluation of their performance and generalizability.

III. Method

A. Problem Formulation, Dataset and Model Design

The learning problem is formulated as a mapping from an N × N Hi-C contact matrix to an N × 3 matrix of 3D coordinates, where N is the number of genomic bins (each bin representing 5 kbp of DNA). The contact matrix encodes how frequently pairs of genomic regions are found in proximity, while the predicted 3D coordinates represent the inferred spatial positions of those regions within the cell.

I used a Hi-C dataset from Lioy et al. (2018) [1], which provides high-resolution contact maps for *E. coli* under different growth phases. The conditions choose have the most drastic difference in genome structure, which is shown by biological researches [5] thus providing a robust benchmark for evaluating model performance.

Two deep learning models are evaluated:

- The architecture of REACH-3D consists of an encoder and decoder, both implemented with single-layer Long Short-Term Memory (LSTM) networks to capture the sequential nature of genomic data. The encoder processes each N × N Hi-C contact matrix—reshaped as a sequence of N vectors—to generate low-dimensional embeddings that preserve genomic order and proximity. The decoder maps transform these embeddings back to the original contact space. This design implicitly learning both local and long-range interactions from Hi-C data [2].
- TECH-3D is trained on synthetic datasets where 3D coordinates are paired with their corresponding computed Hi-C contact matrices. The model architecture includes a bidirectional LSTM encoder that processes each N × N contact matrix as a sequence of N vectors. The output is passed through a fully connected linear layer to produce 3D embeddings of dimension. During training, small Gaussian noise is optionally added to break symmetry. This architecture allows the model to learn spatial representations from simulated data and generalize them to experimental Hi-C inputs [3].

B. Problems and Solutions

Adapting REACH-3D to the compact *E. coli* genome required adjusting the model's constraints on distances between consecutive genomic bins. Originally designed for larger eukaryotic genomes, REACH-3D needed tighter spacing to reflect the high-density organization of bacterial chromosomes. In addition, the loss function was modified to incorporate bacterial-specific features, combining Hi-C reconstruction accuracy, bond length, and overall genome compactness to ensure biologically plausible 3D structures.

For TECH-3D, the main challenge was generating realistic training data. The original synthetic structures were unsuitable for bacterial systems, so a new generator was developed based on a toroidal backbone with random loops, noise, and compaction—parameterized from distributions to capture natural variability. The model was retrained on this dataset to improve alignment with bacterial genome architecture.

Another key issue was the instability caused by the Kabsch alignment loss. This was replaced with a correlation-based loss that compares pairwise distance matrices, providing rotation- and translation-invariant structure comparison. The revised loss combines biological smoothness, distance correlation, and bond length consistency, resulting in more stable and interpretable predictions. To improve training efficiency, a learning rate scheduler and early stopping were introduced. Finally, both models underwent hyperparameter tuning to optimize performance for the bacterial dataset.

IV. Results

With the original models adapted to better suit the structural characteristics of the *E. coli* genome, the data pipeline and performance of REACH-3D and TECH-3D in reconstructing bacterial genome are shown in the subsections below.

A. Data Pipeline

Distance normalization, log scaling, and min–max rescaling are provided as options for each 5kb Hi-C matrix (928×928), which is then reshaped from an $N \times N$ grid into a length-N sequence of contact-frequency vectors—the format required by the REACH-3D LSTM encoder. These sequences are streamed to the GPU for processing. Training is unsupervised: Adam (learning rate $= 10^{-3}$) minimizes a composite loss comprising Hi-C reconstruction error, a bond-length regularizer, and a compactness term, with early stopping. The final encoder embeddings are saved as $N \times 3$ coordinates, and reconstructed contact maps are saved for quantitative comparison.

TECH-3D's pipeline has two stages. **(1) Synthetic-data generation.** A custom generator creates thousands of *E. coli*–like 3-D coordinate sets and converts them to Hi-C contact matrices. Each coordinate–matrix pair is stored as a compressed `.npy` file, then split 90 / 10 % into training and test subsets. **(2) Supervised training.** Each batch supplies an input matrix $[B, N, N]$ together with its ground-truth coordinates $[B, N, 3]$ to a bidirectional LSTM encoder. Outputs are centered and perturbed with small Gaussian noise during training, then evaluated with three-component loss (biological smoothness, distance correlation, distance MSE). Adam optimizer with a step-based learning rate decay schedule is applied, and early stopping terminates optimization when validation loss plateaus. After convergence, predicted coordinates for both synthetic test data and real Hi-C maps are exported for downstream visualization and metric comparison.

B. Training and Reconstruction evaluation

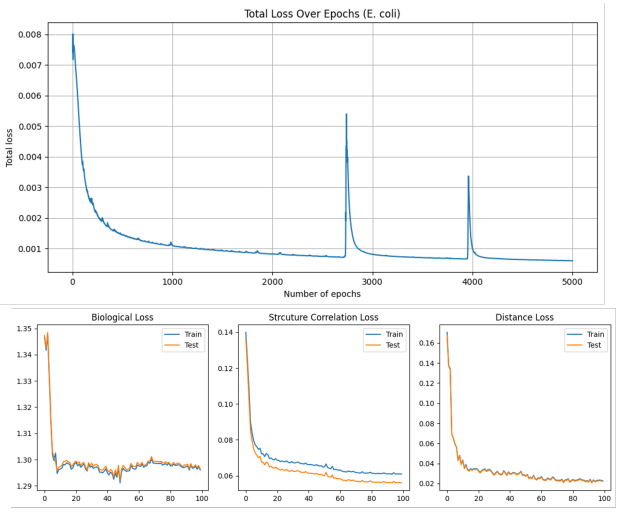

Fig. 1. Training-loss trajectories. **Top**: total loss for REACH-3D on the *E. coli* dataset. **Bottom**: evolution of TECH-3D's three component losses shown for both training and test sets across 100 epochs.

Figure 1 shows the loss trajectories over the course of training. REACH-3D's total loss decreases sharply during the first 300 epochs, then gradually converges to a stable minimum (6×10^{-4}). The two transient spikes indicates jumping out from the local geometry; each is followed by another round of improvement, demonstrating the optimizer escaped shallow minima but required several hundred iterations to stabilize. TECH-3D seems to demonstrates efficient convergence within 100 epochs: biological smoothness, structure-correlation, and distance-reconstruction losses all fall steeply in the first 20 epochs and then level off, with train and test curves tracking closely, indicating good generalization. After hyperparameter tuning, the best-performing versions of both models were applied to predict the 3D genome structure of *E. coli* under two distinct growth conditions for downstream evaluation.

Figure 2 demonstrates the reconstruction performance of both models on experimental Hi-C data. For REACH-3D, the reconstructed contact map closely matches the input, indicating effective training and a high-fidelity reconstruction. In contrast, the contact map reconstructed by TECH-3D coarsely captures the general interaction pattern but the overall similarity is not ideal, suggesting less precise reconstruction. The 3D structure predicted by REACH-3D displays characteristic short-range curvature consistent with local chromosomal folding, but fails to recover long-range interactions and loop structures previously reported by orthogonal methods.

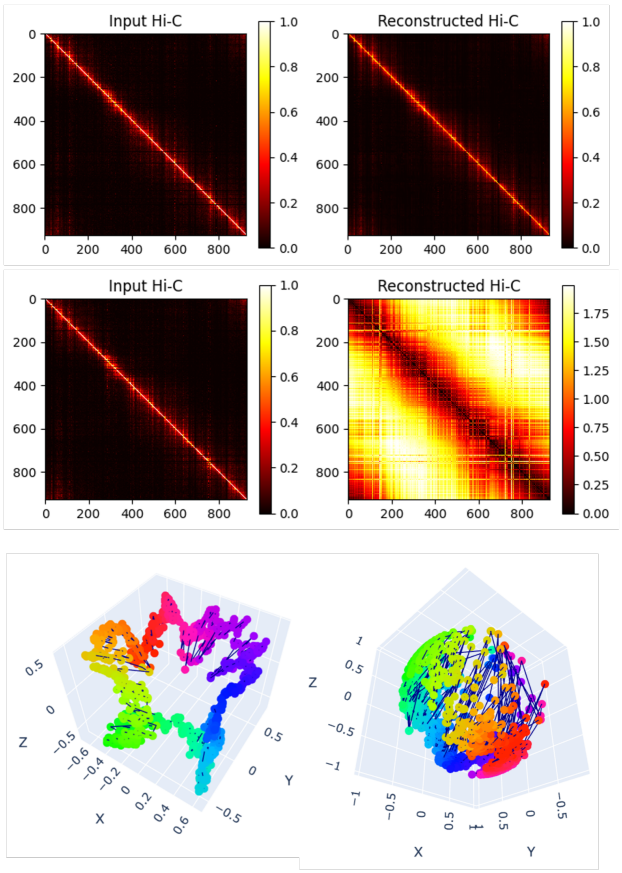

Fig. 2. Prediction visualization. **Top-left and middle-left**: experimental Hi-C maps for two *E. coli* conditions. **Top-right**: corresponding map reconstructed by REACH-3D. **Middle-right**: map reconstructed by TECH-3D. **Bottom**: 3-D genome structure predicted by REACH-3D (left) and TECH-3D (right); color traces the genomic coordinate from origin (red) to terminus (violet).

This limitation arises from the nature of manifold learning, where the latent space does not necessarily correspond to physical 3D coordinates. Nevertheless, the model's ability to recover biologically meaningful local features highlights its potential, though further adaptation is required for robust application to bacterial genomes. The structure predicted by TECH-3D tends to form a compact, spherical configuration—morphologically similar to structures typically observed in eukaryotic chromatin. This suggests that the model's generalization to prokaryotic systems is limited. To determine whether this limitation stems from the model's intrinsic design or from training failure, we tested its predictions on synthetic data with ground-truth structures (Figure 3).

The three loops present in the synthetic structure are clearly captured by the TECH-3D model, demonstrating its ability to learn meaningful spatial patterns under the transfer learning framework. However, the predicted structure lacks the twisting observed in the ground truth, suggesting that the mismatch may stem from suboptimal training—even though convergence was observed in the loss trajectory shown in Figure 1. Additionally, the limited performance on experimental Hi-C data likely reflects a disconnect between the synthetic training data and the true structural complexity of bacterial chromosomes.

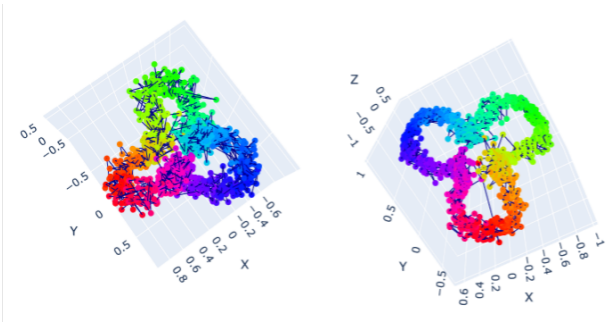

Fig. 3. TECH-3D evaluation on a synthetic test sample. **Left**: ground-truth 3-D chromosome conformation generated by the toroidal–loop model. **Right**: corresponding structure predicted by TECH-3D.

Conclusion

This project evaluated REACH-3D and TECH-3D for reconstructing the 3D genome structure of *E. coli*. Beyond hyperparameter tuning, the results emphasize the importance of tailoring loss functions to biological context and ensuring the training set reflects the system being studied.

REACH-3D was easier to train and produced biologically reasonable predictions, particularly for local structures. However, due to the nature of manifold learning, it struggled to capture long-range interactions and loop formations. TECH-3D, while requiring more careful adjustment and synthetic data design, showed strong potential in reconstructing known structures, though its generalization to real bacterial data was limited—likely due to the mismatch between synthetic and experimental distributions. A planned quantitative evaluation was omitted because visual inspection proved effective for comparing outputs and identifying artifacts. Overall, both models show promise, with TECH-3D requiring further adaptation for prokaryotic systems.

References

[1] Lioy, V.S., Cournac, A., Marbouty, M., Duigou, S., Mozziconacci, J., Espéli, O., Boccard, F. and Koszul, R., 2018. Multiscale structuring of the E. coli chromosome by nucleoid-associated and condensin proteins. Cell, 172(4), pp.771-783.

[2] Cristescu, B.C., Borsos, Z., Lygeros, J., Martínez, M.R. and Rapsomaniki, M.A., 2018. Inference of the three-dimensional chromatin structure and its temporal behavior. arXiv preprint arXiv:1811.09619.

[3] Georges, T.M. and Rapsomaniki, M.A., 2021. Modeling the three-dimensional chromatin structure from hi-c data with transfer learning. bioRxiv, pp.2021-12.

[4] Verma, S.C., Qian, Z. and Adhya, S.L., 2019. Architecture of the Escherichia coli nucleoid. PLoS genetics, 15(12), p.e1008456.

[5] Dame, R.T., Rashid, F.Z.M. and Grainger, D.C., 2020. Chromosome organization in bacteria: mechanistic insights into genome structure and function. Nature Reviews Genetics, 21(4), pp.227-242.

[6] Liu, T., Qiu, Q.T., Hua, K.J. and Ma, B.G., 2024. Chromosome structure modeling tools and their evaluation in bacteria. Briefings in Bioinformatics, 25(2), p.bbae044.