

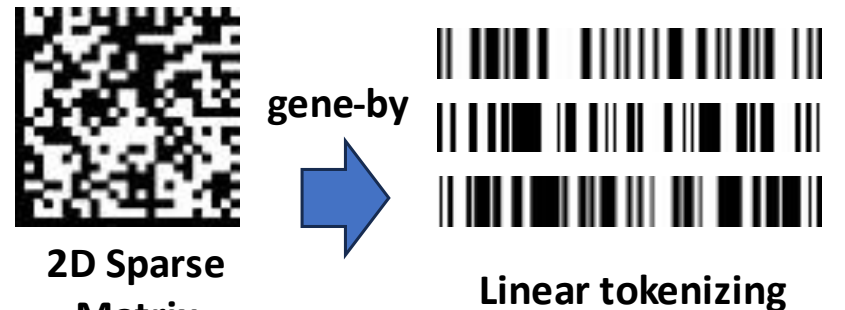
Abstract

In the field of biology, numerous emerging methodologies and analytical approaches are uncovering implicit associations among multiple domains. These interconnected branches of biological study, collectively known as multi-omics, include genomics, transcriptomics, phenomics, proteomics, and more. To achieve the goal of multi-omics integration, we propose a novel deep learning approach called Transformer Architecture with Bayesian Inference for Cascading Omics (TABICO). This method aims to provide comprehensive interpretations by integrating observations measured through various methodologies. The core concept of TABICO is that the attention mechanism can be probabilistically interpreted, and the distribution of observations with multiple (random) variables can be modeled by chaining the attention with variational inference. Thus, TABICO leverages the attention mechanism to understand the intricate relationships and dependencies among the various biological domains.

Theoretical Implementation

1. Overview

Data Serialization



Gene-centric collections

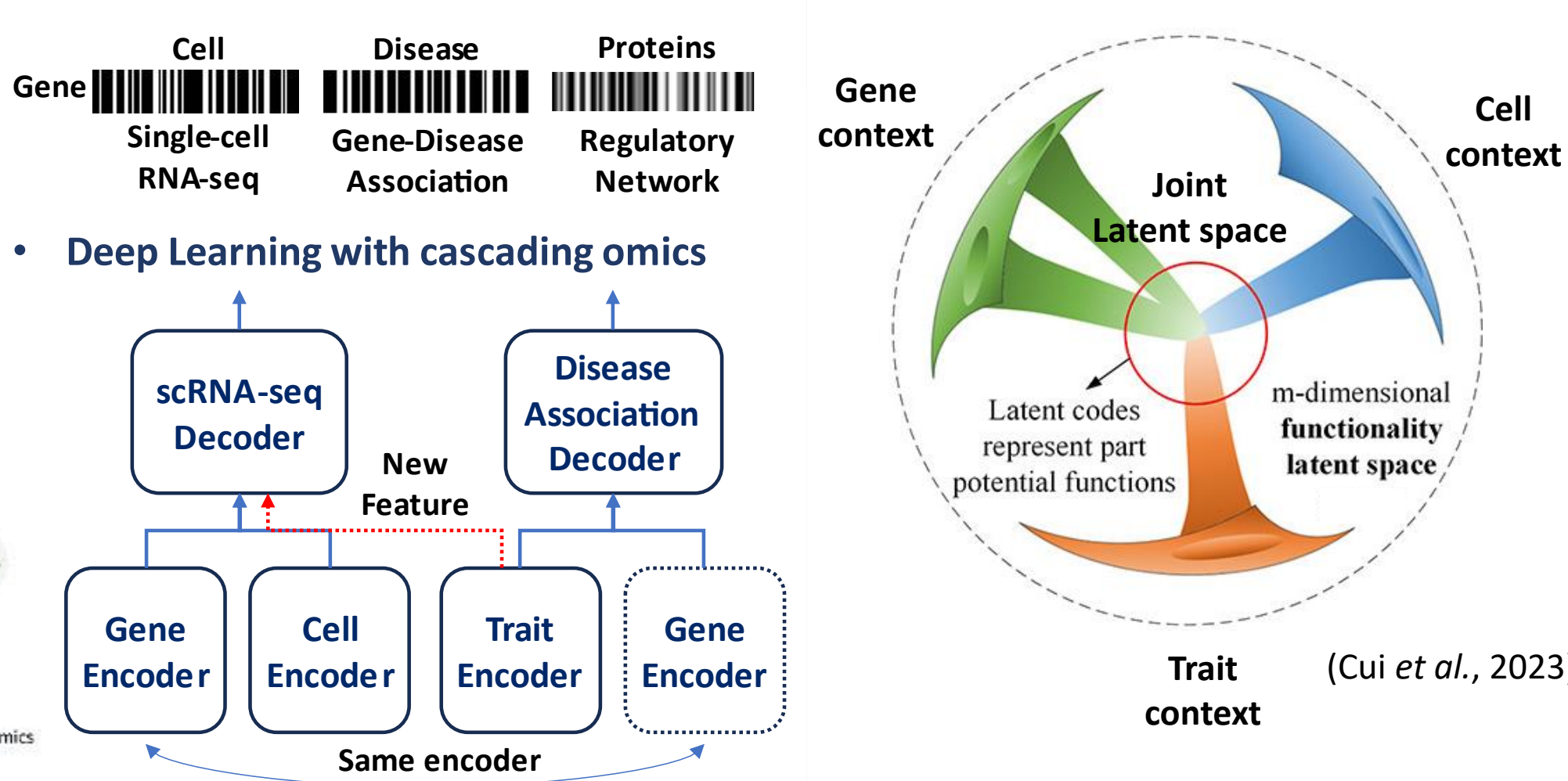


Figure 1. Overview of TABICO. We designed deep learning-based strategy to learn domain information with multi-omics data. Graphical representation showing how we perform analytical approximation of posterior probabilities for unobserved expressions.

2. Mathematical derivation

2-a. Variational inference for marginal likelihood

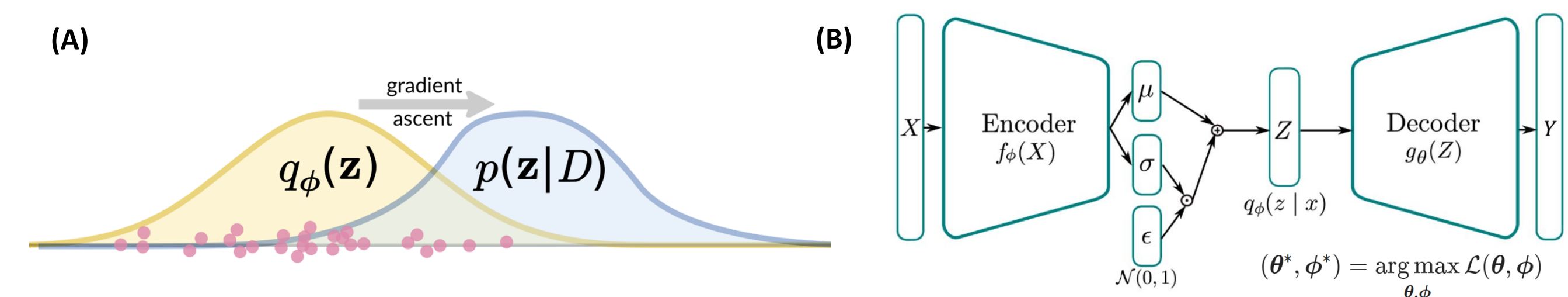


Figure 2. Variational Bayesian model framework. (A) Graphical representation of posterior probability approximation for unobserved variables. (B) Model optimization example illustrating parameter adjustment to maximize observed data log-likelihood.

Under the assumption that conditional independence between latent variables is guaranteed,

$$\begin{aligned} \log p_{\theta}(y|g, c) &= \log \int_{z_g} \int_{z_c} p_{\theta}(y, z_g, z_c | g, c) \frac{q_{\phi_g}(z_g | g)}{q_{\phi_g}(z_g | g)} \frac{q_{\phi_c}(z_c | c)}{q_{\phi_c}(z_c | c)} dz_g dz_c \\ &= \log \mathbb{E}_{q_{\phi_g}(z_g | g), q_{\phi_c}(z_c | c)} \left[\frac{p_{\theta}(y, z_g, z_c | g, c)}{q_{\phi_g}(z_g | g) q_{\phi_c}(z_c | c)} \right] \\ &= \log \mathbb{E}_{q_{\phi_g}(z_g | g), q_{\phi_c}(z_c | c)} \left[\frac{p_{\theta}(y | z_g, z_c) \cdot p(z_g, z_c | g, c)}{q_{\phi_g}(z_g | g) q_{\phi_c}(z_c | c)} \right] \\ &\geq \mathbb{E}_{q_{\phi_g}(z_g | g), q_{\phi_c}(z_c | c)} [\log p_{\theta}(y | z_g, z_c)] \\ &\quad + \mathbb{E}_{q_{\phi_g}(z_g | g)} \left[\log \frac{p(z_g | g)}{q_{\phi_g}(z_g | g)} \right] + \mathbb{E}_{q_{\phi_c}(z_c | c)} \left[\log \frac{p(z_c | c)}{q_{\phi_c}(z_c | c)} \right] \\ &= \mathbb{E}_{q_{\phi_g}(z_g | g), q_{\phi_c}(z_c | c)} [\log p_{\theta}(y | z_g, z_c)] \\ &\quad - D_{\text{KL}}(q_{\phi_g}(z_g | g) \| p(z_g | g)) - D_{\text{KL}}(q_{\phi_c}(z_c | c) \| p(z_c | c)) \\ &= \text{ELBO} \end{aligned}$$

$$\begin{aligned} \log p_{\theta}(y | z_g, z_c) &= \eta(y, z_g, z_c)^{\top} \mathcal{T}_{\theta}(y) - \log Z_{\theta}(\eta) \\ &\approx \eta(y, z_g, z_c)^{\top} \mathcal{T}_{\theta}(y) - \eta(y, z_g, z_c)^{\top} \nabla_{\eta} \log Z_{\theta}(\eta) \\ &= \eta(y, z_g, z_c)^{\top} \mathcal{T}_{\theta}(y) - \eta(y, z_g, z_c)^{\top} \mathbb{E}_{y \sim p_{\theta}(y | \eta)} [\mathcal{T}_{\theta}(y)] \\ &= \eta(y, z_g, z_c)^{\top} (\mathcal{T}_{\theta}(y) - \mathbb{E}_{y \sim p_{\theta}(y | \eta)} [\mathcal{T}_{\theta}(y)]) \end{aligned}$$

Observed energy Expected energy

$$\mathcal{D} = \{y_i, g_i, c_i\}_{i=1}^N$$

Our goals:
Maximize ELBO!

$$\text{KL}(q(z_i) \| \mathbb{P}(z_i | x_i, \theta))$$

$$\begin{aligned} \log \text{likelihood} &= \log(\mathbb{P}(x_i | \theta)) \\ \text{ELBO} &= \log(q(z_i | \theta)) \\ \mathcal{L}(q, \theta) &= \log \text{likelihood} - \text{ELBO} \end{aligned}$$

(Bai et al., 2020)

Layer-wise modulation

$$\log p_{\theta}(y | z_g, z_c) = \sum_{l=1}^L \log p_l(y | \eta_l, \{\eta_r\}_{r < l})$$

$$\begin{aligned} \eta_l &= \eta_{l-1} + \Delta \eta_l(z_g, z_c, \{\eta_r\}_{r < l}) \\ \mathcal{T}_l(y) &= \mathcal{T}_{l-1}(y) + \Delta \mathcal{T}_l(y) \end{aligned}$$

$$\begin{aligned} \log p_l(y | \eta_l, \{\eta_r\}_{r < l}) &= \sum_{l=1}^L \log p_l(y | \eta_l) \\ &= \sum_{l=1}^L \eta_l^{\top} (\mathcal{T}_l(y) - \mathbb{E}_{y \sim p_{\theta}} [\mathcal{T}_l(y)]) \end{aligned}$$

2-b. Probabilistic interpretation of Attention

Exponential Family

$$p(y|\eta) = \frac{1}{Z(\eta)} h(y) \exp[\eta^{\top} \mathcal{T}(y)]$$

$$Z(\eta) = \int h(y) \exp[\eta^{\top} \mathcal{T}(y)] dy$$

Attention

$$W_{i,j}^{l,h} = \frac{\exp(\Phi_{i,j}^{l,h})}{\sum_{j'=1}^n \exp(\Phi_{i,j'}^{l,h})}$$

$$\Phi_{i,j}^{l,h} = f(Q_{i,\cdot}^{l,h} \cdot K_{\cdot,j}^{l,h}) = \frac{Q_{i,\cdot}^{l,h} \cdot (K_{\cdot,j}^{l,h})^{\top}}{\sqrt{d_k}}$$

Alternative expression of Attention:

$$\mathbb{E}_{y \sim p_{\theta}(y|\eta)} [\mathcal{T}_{\theta}(y)] = \nabla_{\theta} \log Z(\theta)$$

3. Architectures

Bayesian Attention

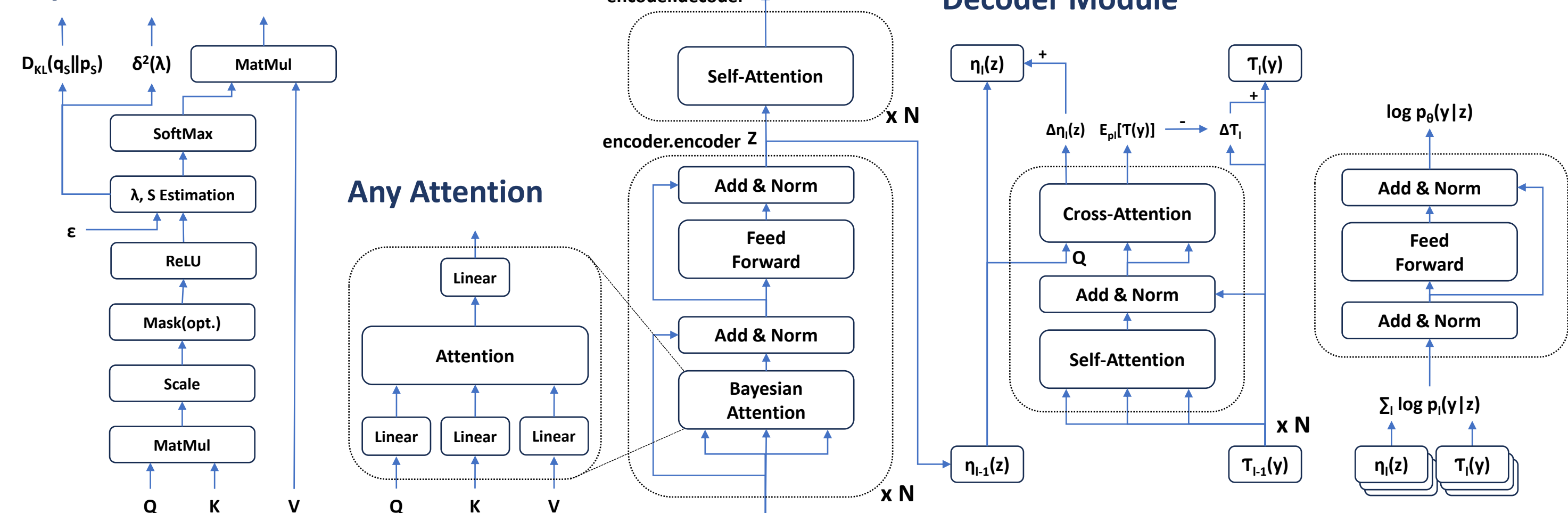


Figure 3. TABICO architecture consisting of two primary modules: 1) A Bayesian attention-based context-encoder that processes ontology-derived ID tokens with added stochastic noise, and 2) An omics-decoder designed to predict omics observations from the encoded data.

4. Objectives

Gene expression prediction: CDW-CE loss

$$\mathcal{L} = \frac{1}{N} \sum_{i=1}^N \underbrace{CE(y_i, \hat{y}_i)}_{\text{Cross Entropy}} \cdot \underbrace{(1 + d_i)^{\alpha}}_{\text{Distance Weight}} \cdot \underbrace{(1 + \max(d_i - m, 0))}_{\text{Margin Term}}$$

$$d_i = |\arg \max_c(\hat{y}_i) - y_i|$$

Token recovery: CE loss + KL Div. loss

$$\mathcal{L}_{INV} = w_{inv} \sum_{k \in K} (\mathcal{L}_{CE}(\hat{y}_k, y_k) + \lambda \mathcal{L}_{aux,k})$$

Semi-analytic KL divergence under non-informative prior

$$\mathcal{L}_{aux,k} = D_{\text{KL}}(\text{Weibull}(k, \lambda) \| \text{Gamma}(\alpha, \beta)) = \frac{\gamma \alpha}{k} - \alpha \log \lambda + \log k + \beta \lambda \Gamma(1 + \frac{1}{k})$$

$$\gamma = \gamma - 1 - \alpha \log \beta + \log \Gamma(\alpha)$$

*α=1e-10, β=1e-15, k=1

Bayesian Attention

$$W_i^{l,h} = \tilde{g}(S_i^{l,h}) := S_i^{l,h} / \sum_j S_{i,j}^{l,h}$$

$$p(S|k, \lambda) = \frac{k}{\lambda^k} S^{k-1} e^{-(S/\lambda)^k}, S \sim \text{Weibull}(k, \lambda)$$

$$S = \tilde{g}(\epsilon) := \lambda(-\log(1 - \epsilon))^{1/k}, \epsilon \sim \text{Uniform}(0, 1)$$

$$S_{i,j}^{l,h} \sim \text{Weibull}(k, \lambda_{i,j}^{l,h}), \lambda_{i,j}^{l,h} = \exp\left(\frac{\Phi_{i,j}^{l,h}}{\Gamma(1 + 1/k)}\right)$$

Preprocessing & Training strategy

❖ Tokenizing with CELLxGENE and OpenTargets database

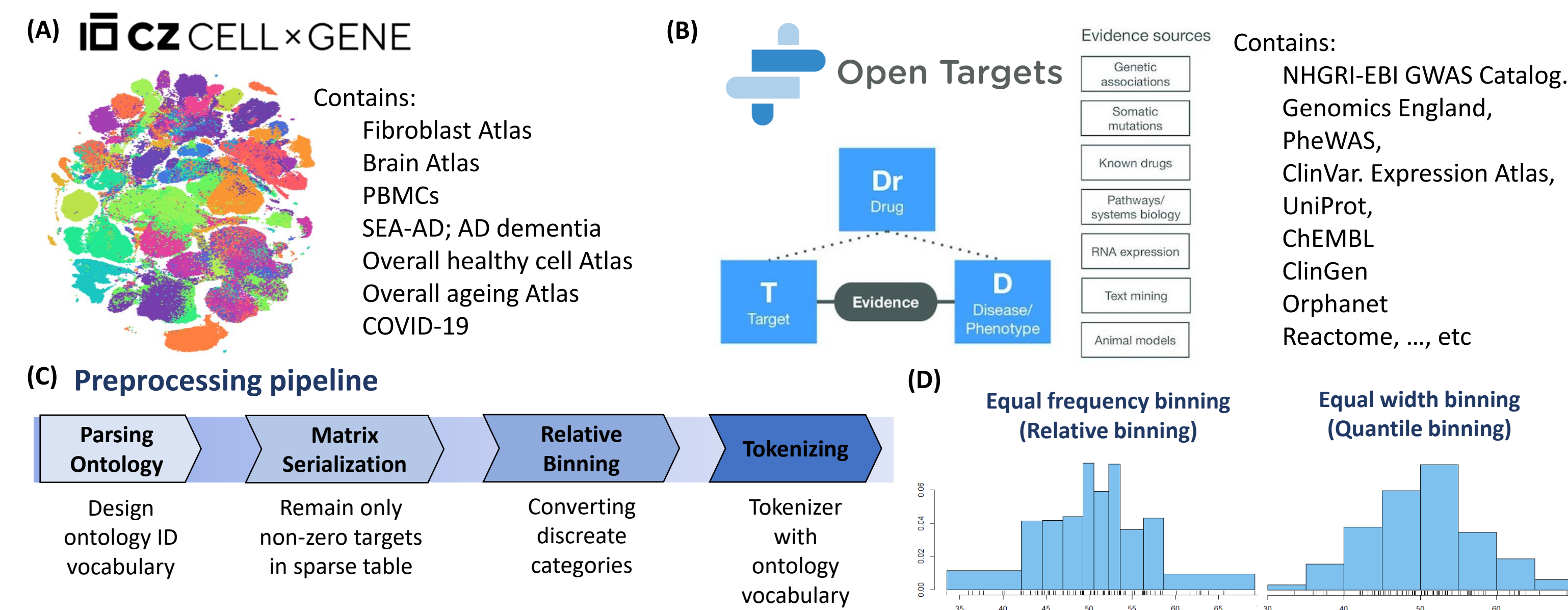


Figure 4. (A) Overall single-cell transcriptomics on CELLxGENE. (B) Integrated gene-disease associations through the evidences from multiple sources on OpenTargets. Furthermore, OpenTargets provides drug target validity including genetic associations, somatic mutations, known drugs, and pathways. (C) Preprocessing pipelines. We mapped the context ID tokens from ontologies including cell ontology(CO), gene ontology(GO) with ENSEMBL, and experimental factor ontology(EFO). We designed tokenizer that supports sparse data format, AnnData, and tabular data format, Spark-induced table. (D) The simple comparison between relative binning and general quantile binning.

Model	Baseline	Dataset	Architecture & Loss	Other options
TABICO	VAE Transformer	1. CellxGene 2. OpenTargets = gene-associated 2D matrix 3. Ontologies(GO, CO, EFO)	TABICO + MLP(option) or + Bayesian attention - Loss Binning with CDW-CE + Random replaced with BCE *CDW-CE; class distance weighted cross-entropy	- Optimizer SGD, Adam - Metrics Accuracies of 1) Expression prediction, 2) Token recovery, 3) Random replace recognition

Table 1. (A) Overall single-cell transcriptomics on CELLxGENE. (B) Integrated gene-disease associations through the evidences from multiple sources on OpenTargets

❖ Example workflow

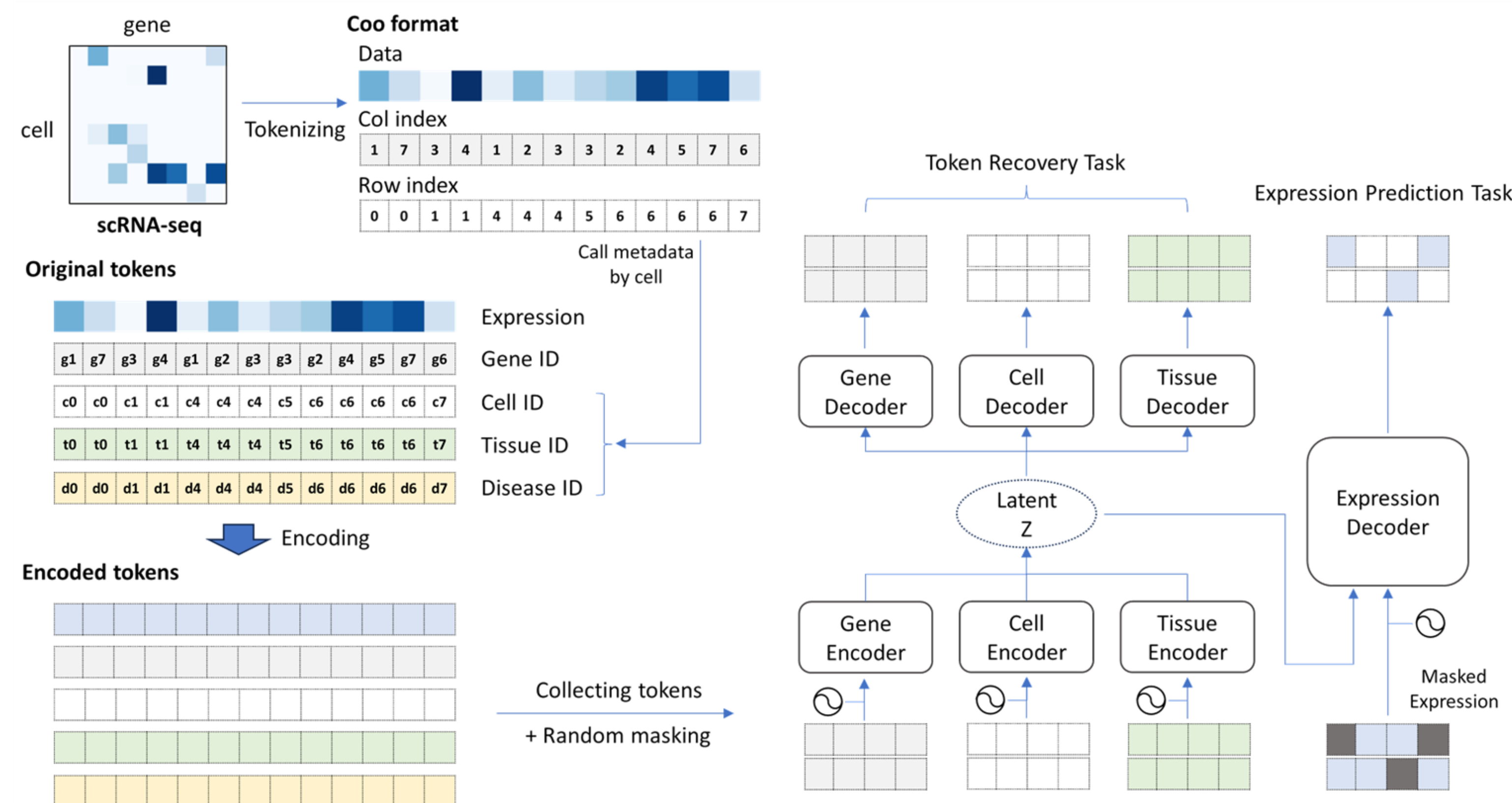


Figure 5. Overview of TABICO workflow. We developed a framework for directly learning from multiple sparse matrices centered on gene domains. Using single-cell RNA-seq data as an example, expression levels, gene IDs, and cell IDs (or metadata) were tokenized and pre-encoded from Coordinate(COO) sparse matrices. An encoder-decoder architecture was then implemented: the encoder learns domain conditions, while the decoder learns observed expression data. Bayesian inference is incorporated, treating the encoder output as a latent variable and enabling the decoder to estimate likelihoods for expression tokens.

Results

❖ TABICO

Learning curve for training

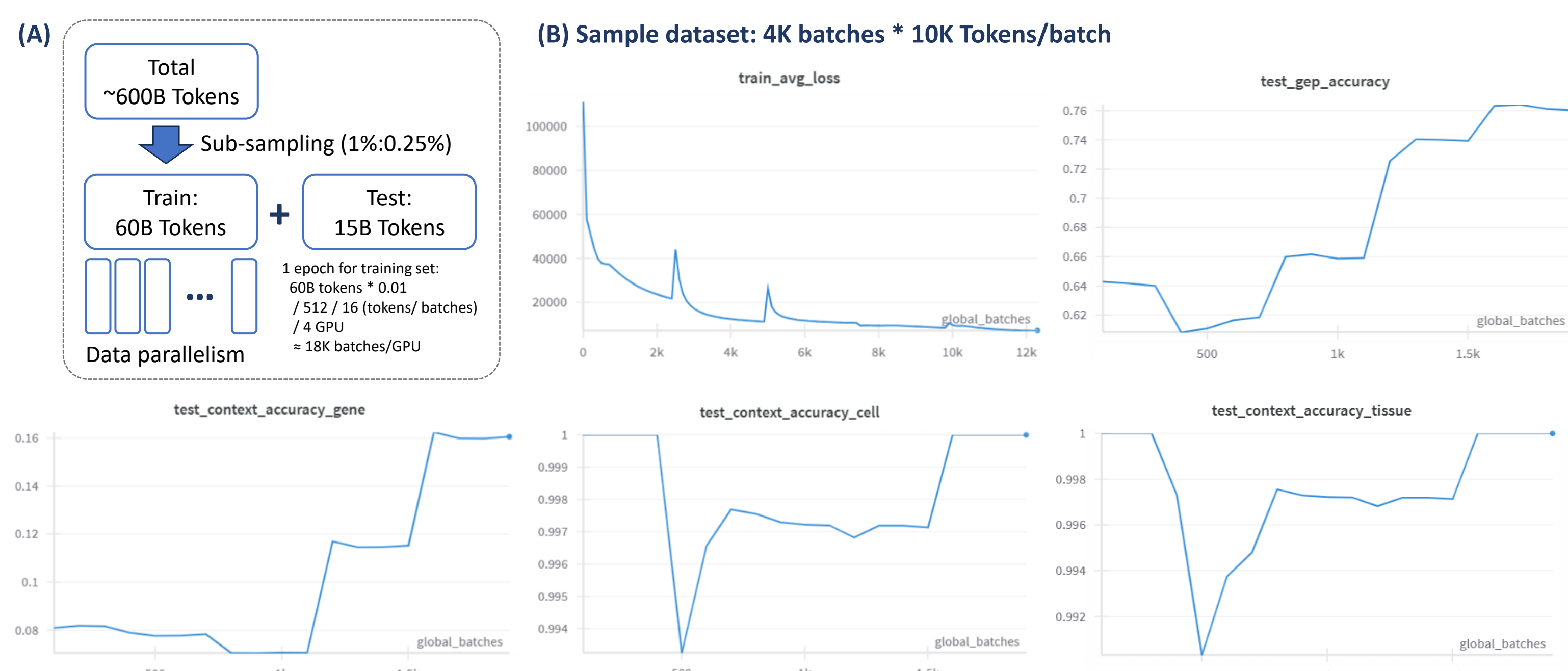


Figure 6. Learning performance of TABICO on single-cell RNA-seq data from the CELLxGENE database. To manage the extensive dataset (600B expression tokens), evaluations were conducted with progressively larger token sub-samples. (A) The efficient sub-sampling strategy in generating manageable training sets. (B) Training process evaluation using sample datasets (e.g., 40M tokens). Total loss comprises metrics for two primary tasks: Gene Expression Prediction (GEP) and context-specific ID token recovery from latent variables (INV). Furthermore, the performance for these tasks, including ID token recovery specific to each context encoder, is presented.

Conclusions and future works

We have presented a Bayesian network architecture integrating variational inference and attention mechanisms, optimized via MLE to maximize the ELBO of conditional likelihood. The modularity enables independent learning of domain-specific latent variables. Our future work focuses on several key areas:

- Integration of diverse multi-omics:** we will perform pre-training and fine-tuning strategies to transfer knowledge among diverse multi-omics datasets.
- Causal Inference:** we will incorporate causal inference to identify driver genes and regulatory circuits in biological systems, moving beyond associations to establish causal relationships for enhanced biological discovery.
- Disease Modeling:** We will predict disease subtypes and identify therapeutic targets, using attention mechanisms to reveal key genes and pathways associated with different disease conditions.

These future directions aim to broaden TABICO's applicability, improve its theoretical foundations, and ultimately drive new biological insights.