Supplementary Materials
mbDiffusion: Deep Conditional Diffusion Generative Modeling for Missing Microbiome Data Recovery

I. Algorithm of mbDiffusion

We provide mbDiffusion pseudocode to clarify the conditional mechanism in training and inference. The training procedure is detailed in Algorithm 1, while the inference process is described in Algorithm 2.

Algorithm 1 Training of mbDiffusion

Input: Latent space $X_{latent} \in \mathbb{R}^{n \times p}$, Metadata $X_{meta} \in \mathbb{R}^{n \times k}$, observed data $X_{observed} \in \mathbb{R}^{n \times p}$
Output: Trained denoising function ϵ_θ

Add noise to the X_{latent} for T steps to obtain Y_{latent}
Use the pretrained large model to encode Metadata X_{meta} into embedding vectors $Y_{meta} = \tau(X_{meta})$
for $i = 1$ to N_{iter} **do**
  $x_0 = \phi(Y_{latnet})$
  $c_{meta} = \psi_1(Y_{meta})$, $c_{observed} = \psi_2(X_{observed})$
  $x_c = c_{meta} + c_{observed}$
  $t \sim \text{Uniform}(\{1, ..., T\})$
  Updata to minimize $\|\epsilon - \epsilon_\theta(\mathbf{x}_t^*, t \mid x_c) \odot M\|_2^2$
end for
until converged

Algorithm 2 Inference of mbDiffusion

Input: Gaussian Noise $\mathcal{N}(0, I)$, Metadata $X_{meta} \in \mathbb{R}^{n \times k}$, observed data $X_{observed} \in \mathbb{R}^{n \times p}$
Output: Predicted latent space expression x_0

Sample initial noise $x_t \sim \mathcal{N}(0, I)$
Use the pretrained large model to encode Metadata X_{meta} into embedding vectors $Y_{meta} = \tau(X_{meta})$
for $t = T$ to 1 **do**
  Sample $\epsilon_t \sim \mathcal{N}(0, I)$ if $t > 1$, else $\epsilon_t = 0$
  $c_{meta} = \psi_1(Y_{meta})$, $c_{observed} = \psi_2(X_{observed})$
  $x_c = c_{meta} + c_{observed}$
  Update using reverse diffusion step $x_{t-1} \leftarrow \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{1-\alpha_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t, t|x_c)) + \sqrt{\beta_t}\epsilon_t$
  $t \leftarrow t - 1$
end for
return the denoised sample x_0

II. Evaluation Metrics

To evaluate the performance of mbDiffusion and baseline methods, we used four evaluation metrics on three datasets: Pearson correlation coefficient (PCC), Cosine distance (Cosine), Root Mean Square Error (RMSE), and Mean Absolute Error (MAE). The calculation method for four evaluation metrics are as follows:

$$\text{PCC} = \frac{\text{Cov}(x, y)}{\sqrt{\text{Var}(x)} \cdot \sqrt{\text{Var}(y)}} \tag{1}$$

$$\text{Cosine}(x, y) = \frac{x \cdot y}{\|x\| \cdot \|y\|} \tag{2}$$

$$\mathrm{RMSE}(x, y) = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - y_i)^2} \tag{3}$$

$$\mathrm{MAE}(x, y) = \frac{1}{n}\sum_{i=1}^{n}|x_i - y_i| \tag{4}$$

where x denotes the original data with the first-stage masking applied, and y represents the imputed values corresponding to the masked regions after completion. Cov() is the covariance, and Var() is the variance.

III. Ablation Studies

To validate the functional components of mbDiffusion, we perform systematic ablation studies examining the following aspects: (1) Different backbone networks; (2) Whether to include metadata as condition; (3) Whether to pre-training VAE.

Different backbone networks. In the backbone network of mbDiffusion, we employ DiT blocks based on a cross-attention mechanism. The purpose of using cross-attention is to effectively fuse different types of data, enhance feature interaction, improve interpretability and performance of the model, and enable the generation of higher-quality synthetic data. To validate the effectiveness of the cross-attention mechanism, we compared it with a diffusion model that uses a traditional Unet network as its backbone.

The experimental results in Table I demonstrate that, compared to the traditional U-Net architecture, employing DiT blocks with a cross-attention mechanism as the backbone is more suitable for our specific task. This superior performance may be attributed to the cross-attention mechanism's ability to more effectively integrate input data that carry heterogeneous types of information. By explicitly modeling the interactions between different modalities, the mechanism can capture the underlying dependencies and complementary features across inputs.

TABLE I: Ablation experiments on different backbone networks. The results present the mean and standard deviation across five independent experiments.

PCC	S/HNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
Backbone w/ Unet	0.544±0.025	0.582±0.077	0.535±0.029	0.511±0.083	0.553±0.117	0.472±0.099
Backbone w/ DiT	**0.624±0.048**	**0.701±0.055**	**0.626±0.051**	**0.619±0.093**	**0.627±0.106**	**0.535±0.111**
Cosine	S/HNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
Backbone w/ Unet	0.722±0.028	0.718±0.034	0.751±0.011	0.638±0.079	0.798±0.084	0.832±0.079
Backbone w/ DiT	**0.803±0.044**	**0.773±0.062**	**0.792±0.036**	**0.694±0.048**	**0.873±0.068**	**0.871±0.058**
RMSE	S/HNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
Backbone w/ Unet	1.231±0.064	1.232±0.031	1.422±0.037	1.426±0.078	1.233±0.056	1.124±0.081
Backbone w/ DiT	**1.162±0.046**	**0.964±0.049**	**1.328±0.071**	**1.314±0.074**	**1.139±0.042**	**0.962±0.072**
MAE	S/HNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
Backbone w/ Unet	0.918±0.033	0.732±0.082	1.247±0.041	1.127±0.032	0.957±0.027	0.833±0.031
Backbone w/ DiT	**0.827±0.018**	**0.522±0.043**	**0.977±0.073**	**1.064±0.042**	**0.836±0.069**	**0.749±0.048**

Whether to include metadata as condition. In the field of bioinformatics, each dataset contains a wealth of valuable information. For instance, in a specific cancer patient microbiome dataset, not only is there microbial data, but also patient metadata, which encompasses a significant amount of information. In our approach, we propose a method to integrate these diverse types of metadata. To verify whether this metadata contributes to performance of the model, we conducted ablation experiments.

The experimental results in Table II indicate that enriching the diffusion model's conditions with metadata is beneficial for our specific task, leading to a noticeable improvement in model performance. This indicates that the information extracted from metadata plays a beneficial role in the imputation task. This enhancement also provides new insights for subsequent experiments, such as exploring better embedding methods for text-based metadata and more effective ways to integrate diverse types of metadata. At the same time, this also provides inspiration for our future research. It raises an important question: if, in addition to incorporating metadata, we also integrate corresponding single-cell sequencing data or other types of multi-omics data to enrich the conditioning information, will this lead to further improvements in imputation performance? This is a question worth investigating.

Whether to pre-train VAE. Due to the limited sample size of individual datasets, training a model on a single dataset alone does not achieve the desired results. Therefore, we adopted a transfer learning approach. Specifically, we

TABLE II: Ablation experiments on whether the proposed mbDiffusion with the condition (metadata). The results present the mean and standard deviation across five independent experiments.

PCC	S/HNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
w/o Metadata	0.599±0.008	0.682±0.051	0.602±0.037	0.595±0.088	0.602±0.075	0.511±0.079
w/ Metadata	**0.624±0.048**	**0.701±0.055**	**0.626±0.051**	**0.619±0.093**	**0.627±0.106**	**0.535±0.111**
Cosine	S/HNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
w/o Metadata	0.786±0.023	0.755±0.026	0.773±0.017	0.654±0.074	0.847±0.024	0.856±0.077
w/ Metadata	**0.803±0.044**	**0.773±0.062**	**0.792±0.036**	**0.694±0.048**	**0.873±0.068**	**0.871±0.058**
RMSE	S/HNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
w/o Metadata	1.175±0.038	0.975±0.024	1.343±0.077	1.339±0.036	1.169±0.033	0.981±0.065
w/ Metadata	**1.162±0.046**	**0.964±0.049**	**1.328±0.071**	**1.314±0.074**	**1.139±0.042**	**0.962±0.072**
MAE	S/HNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
w/o Metadata	0.837±0.031	0.558±0.069	0.966±0.033	1.078±0.025	0.852±0.037	0.768±0.022
w/ Metadata	**0.827±0.018**	**0.522±0.043**	**0.977±0.073**	**1.064±0.042**	**0.836±0.069**	**0.749±0.048**

TABLE III: Ablation experiments on whether pre-training VAE module. The results present the mean and standard deviation across five independent experiments.

PCC	S/HNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
w/o Pre_train	0.583±0.036	0.629±0.050	0.533±0.021	0.521±0.094	0.533±0.146	0.472±0.161
w/ Pre_train	**0.624±0.048**	**0.701±0.055**	**0.626±0.051**	**0.619±0.093**	**0.627±0.106**	**0.535±0.111**
Cosine	S/SHNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
w/o Pre_train	0.738±0.048	0.725±0.032	0.662±0.029	0.634±0.092	0.797±0.057	0.832±0.074
w/ Pre_train	**0.803±0.044**	**0.773±0.062**	**0.792±0.036**	**0.694±0.048**	**0.873±0.068**	**0.871±0.058**
RMSE	S/HNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
w/o Pre_train	1.261±0.044	0.992±0.018	1.436±0.030	1.387±0.088	1.176±0.140	1.117±0.062
w/ Pre_train	**1.162±0.046**	**0.964±0.049**	**1.328±0.071**	**1.314±0.074**	**1.139±0.042**	**0.962±0.072**
MAE	S/HNSC	S/COAD	S/STAD	B/HNSC	B/COAD	B/STAD
w/o Pre_train	0.879±0.025	0.552±0.031	1.037±0.046	1.103±0.066	0.871±0.093	0.789±0.022
w/ Pre_train	**0.827±0.018**	**0.522±0.043**	**0.977±0.073**	**1.064±0.042**	**0.836±0.069**	**0.749±0.048**

first used other datasets to learn an initial set of weights, which were then transferred to the backbone network as its starting weights. To validate the effectiveness of this pre-training strategy for our specific task, we conducted ablation experiments comparing scenarios with and without the pre-training strategy.

The experimental results shown in Table III demonstrate that adopting a pre-training and fine-tuning strategy leads to a substantial improvement in model performance. This clearly indicates that pre-training plays a crucial role in mitigating the challenges associated with training models on small-sample datasets by providing a strong initialization that helps the model converge more effectively during fine-tuning. Moreover, these findings offer valuable insights for future research, such as exploring whether using larger-scale pre-training datasets could further improve model performance.