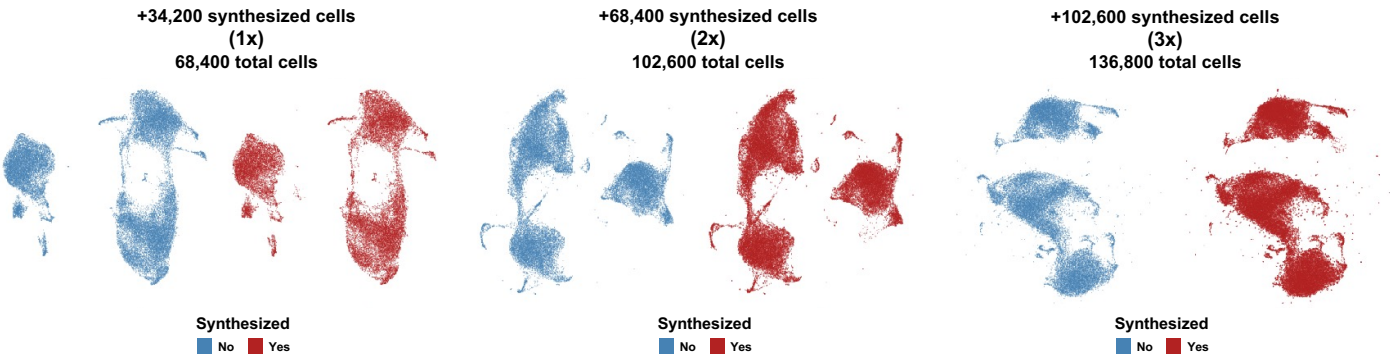


Single-cell RNA-seq data augmentation using generative Fourier transformer

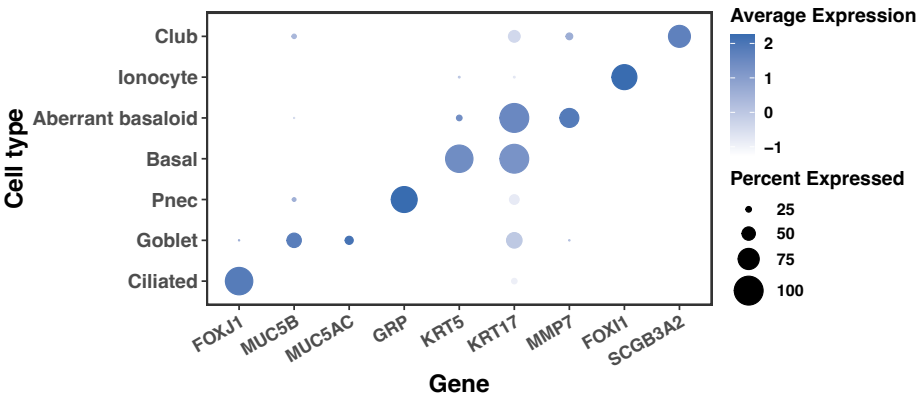
Nima Nouri

Precision Medicine and Computational Biology, Sanofi, 350 Water Street, Cambridge, MA 02141, USA
nima.nouri@sanofi.com



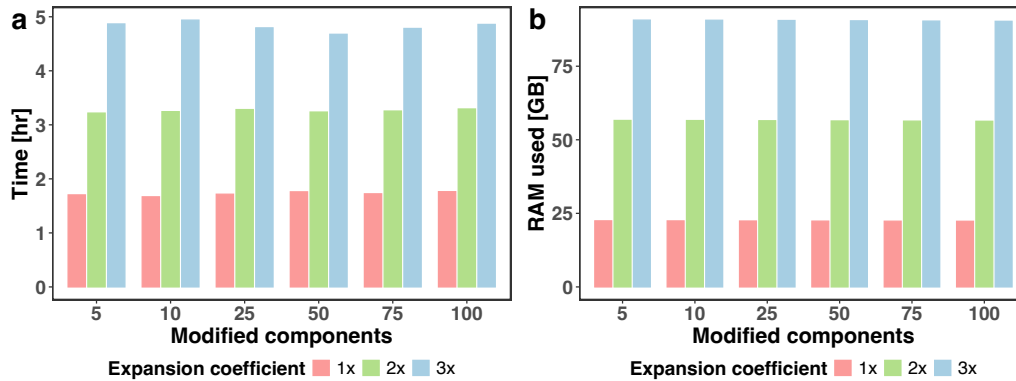
Supplementary Figure 1

Evaluation of synthesized cells at different expansion levels. UMAP representation of the PRJEB44878 dataset, consisting of 34,200 processed single cells from primary small airway epithelial cells. Each column illustrates original cells (blue) and cells synthesized by scGFT (red), shown separately for 100 complex components modified under varying expansion coefficients: 1x synthesis, resulting in 68,400 total cells (left panel); 2x synthesis, resulting in 102,600 total cells (middle panel); and 3x synthesis, resulting in 136,800 total cells (right panel). Each point in the UMAP plot represents an individual cell.



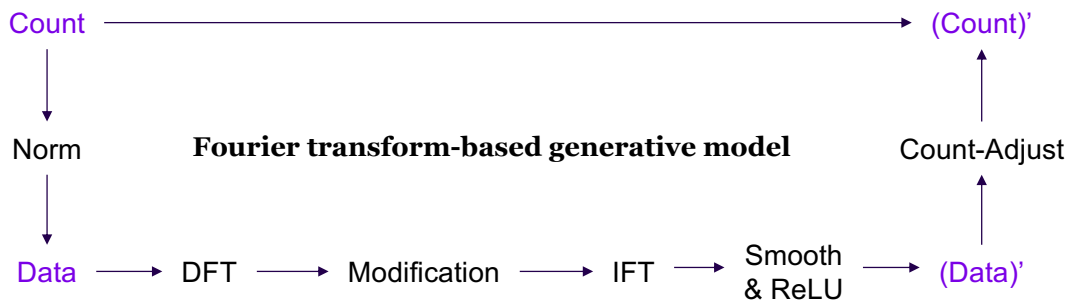
Supplementary Figure 2

Expression levels of epithelial subtype markers. Dot-plot displaying selected marker genes for each of the different subtypes of epithelial cells in the PRJEB44878 dataset. The color of each circle represents the average expression level of the gene, while its size indicates the percentage of cells expressing that particular gene.



Supplementary Figure 3

Scalability and resource efficiency of scGFT on large datasets from the Tabula Sapiens Consortium (~500K single cells). **a**, Runtime (y-axis) of the synthesis process on a uniprocessor system, plotted against the number of complex components modified (x-axis). **b**, Memory usage (y-axis) of the synthesis process, plotted against the number of complex components modified (x-axis). The color legend denotes the expansion coefficients of the synthesis process.



Supplementary Figure 4

Overview of the scGFT Fourier transform-based generative model. The process commences with the normalization of raw counts, followed by the application of the Discrete Fourier Transform (DFT) to extract complex components. These components are then modified and subjected to an Inverse Fourier Transform (IFT) to map back to the original data space. To mitigate potential noise introduced during the IFT, a smoothing procedure is implemented, followed by a Rectified Linear Unit (ReLU) activation function to clip any residual negative values generated during the IFT process to zero. Finally, the synthesized raw counts are derived from their original raw counterparts by taking into account the relative change rate between the synthesized data and their original normalized counterparts.

	scGAN	scDiffusion	scVI	scGFT
Core Methodology	Generative Adversarial Net (GAN)	Latent Diffusion Model (LDM)	Variational Autoencoder (VAE)	Fourier Transform (FT)
Approach Type	Deep learning	Deep learning	Deep learning	Analytical solution
Training Requirement	Required	Required	Required	Train-free
UMAP Alignment	Good overlap	Good overlap	Good overlap	Good overlap
Fidelity to Original Clusters	11% lower compared to scGFT	9% lower compared to scGFT	9% lower compared to scGFT	Baseline
Fidelity to Profile Variability	30% lower compared to scGFT	36% lower compared to scGFT	20% lower compared to scGFT	Baseline
Data Sparsity Impact	5% increase in sparsity	14% imputation	1% increase in sparsity	2% imputation

Supplementary Table 1

Key methodological, computational, and performance metrics of scGAN, scDiffusion, scVI, and scGFT.