# Squidiff: Predicting cellular development and responses to perturbations using a diffusion model

Siyu He[1,2,3†], Yuefei Zhu[1†], Daniel Naveed Tavakol[1†], Haotian Ye[4], Yeh-Hsing Lao[1,5], Zixian Zhu[1], Cong Xu[1], Sharadha Chauhan[6], Guy Garty[7], Raju Tomer[1,6], Gordana Vunjak-Novakovic[1,13], James Zou[3,4,8*], Elham Azizi[1,2,9,10,11*], Kam W. Leong[1,13*]

[1]Department of Biomedical Engineering, Columbia University, NY.
[2]Irving Institute for Cancer Dynamics, Columbia University, NY.
[3]Department of Biomedical Data Science, Stanford University, CA.
[4]Department of Computer Sciences, Stanford University, CA.
[5]Department of Pharmaceutical Sciences, University at Buffalo, NY.
[6]Department of Biological Sciences, Columbia University, NY.
[7]Center for Radiological Research, Columbia University, NY.
[8]Department of Electrical Engineering, Stanford University, CA.
[9]Department of Computer Sciences, Columbia University, NY.
[10]Data Science Institute, Columbia University, NY.
[11]Herbert Irving Comprehensive Cancer Center, Columbia University, NY.
[12]Department of Medicine, Columbia University,NY.
[13]Department of Systems Biology, Columbia University Irving Medical Center, NY.

*Corresponding author(s). E-mail(s): jamesz@stanford.edu; ea2690@columbia.edu; kam.leong@columbia.edu;
Contributing authors: siyuhe@stanford.edu; yz3713@columbia.edu; dnt2114@columbia.edu; haotianye@stanford.edu; sima@buffalo.edu; zz2894@columbia.edu; cx2230@columbia.edu; sc4638@columbia.edu¿; gyg2101@cumc.columbia.edu; raju.tomer@columbia.edu; gv2131@columbia.edu;
†These authors contributed equally to this work.

**Abstract**

Single-cell sequencing has revolutionized our understanding of cellular heterogeneity and responses to environmental stimuli. However, mapping transcriptomic changes across diverse cell types in response to various stimuli and elucidating underlying disease mechanisms remains challenging. Studies involving physical stimuli, such as radiotherapy, or chemical stimuli, like drug testing, demand labor-intensive experimentation, often hindering the rapid advancement of mechanistic insight and drug discovery. To address this, we present Squidiff, a diffusion model-based generative framework designed to predict transcriptomic changes across diverse cell types in response to a wide range of environmental changes. We demonstrate Squidiff's robustness across various scenarios, including cell differentiation, gene perturbation, and drug response prediction. Through continuous denoising and semantic feature integration, Squidiff effectively learns transient cell states and predicts high-resolution transcriptomic landscapes over time and conditions. Furthermore, we applied Squidiff to model the development of blood vessel organoids and cellular responses to neutron irradiation and growth factors. Our results demonstrate that Squidiff enables *in silico* screening of cell molecular landscapes, facilitating rapid hypothesis generation and providing valuable insights for precision medicine.

**Keywords:** Transcriptomic prediction, diffusion model, gene perturbation, cell differentiation, drug testing, blood vessel organoids, radiotherapy

# Introduction

Cells in biological systems coordinate and respond to environmental changes collectively, with responses varying by location and cell type due to both intrinsic tissue heterogeneity and external signaling cues[1, 2]. These living cells are complex, dissipative systems that operate far from chemical equilibrium and often exhibit highly non-linear responses[3]. While single-cell sequencing has enabled the unbiased characterization of cellular heterogeneity, predicting transcriptomic changes in responses to stimuli remains challenging. Exploring disease mechanisms and optimizing drug components require sequencing large-scale screens, which is both laborious and expensive[4, 5].

Several machine-learning based models, such as scGEN[6], CellOT[7], GEARS[8], and GenePert[9] have been developed to predict single-cell perturbation using variational autoencoders (VAE)[10], optimal transport methods[7], graph neural networks[8] and large language models. Despite the advances they offer, these models face limitations in predicting high-resolution dynamic transcriptional responses, especially in highly transient states such as in organ development. Reconstructing smooth transitions of learned features remains challenging. Multiple cell types or cross-cell type transfer of prediction is also not considered in these methods. Additionally, these models require both the unperturbed transcriptomic and perturbed data as inputs but do not fully leverage underlying biological knowledge to guide their predictions, often leading to failure in interpolating cell states.

Diffusion models have demonstrated remarkable potential in data generation, capable of producing high-quality and realistic samples through iterative refinement processes and learning richer and more complex data distributions than variational autoencoders[11–13]. Their integration with variational autoencoders further enhances their ability to model smooth feature transitions and capture complex latent spaces[14]. Recent perspectives highlight their potential in creating virtual cells establishing them as a promising approach in artificial intelligence[15]. However, their application in predicting single-cell perturbations or cell differentiation remains largely untapped. Here we introduce Squidiff (Single-cell QUantitative Inference of stimuli responses by DIFFusion models), a computational framework designed to predict transcriptomic responses of diverse cell types to a spectrum of environmental changes, including cell differentiation, gene perturbation, and drug treatment. Squidiff is a conditional denoising diffusion implicit model[16] which generates new transcriptomes that represent distinct cellular states. Squidiff excels in predicting the differentiation of induced pluripotent stem cells (iPSCs) into mesendoderm and endoderm, guided by stimuli vectors. Notably, Squidiff captures transient cellular states that other methods often miss. Moreover, it effectively predicts non-additive gene perturbation and cell type-specific drug responses, as demonstrated in its application to analysis of temporal cell states in glioblastoma dataset in response to novel drug combinations.

We applied Squidiff to investigate neovascularization and vasculopathy induction in blood vessel organoids exposed to high linear energy transfer (LET) radiation, uncovering potential mechanisms related to radiation-induced changes in vascular differentiation trajectories. Organoid technologies, in this case blood vessel organoids derived from human iPSCs, have shown promise in disease modeling and as drug testing platforms, especially in the context of self-organizing models that can start from a single source of stem cells[17]. This platform is thus particularly useful for understanding injury and identifying potential drug candidates with targeting extreme injury states on human cells, such as high LET radiation exposures. In particular, these models are critical for scenarios that cannot be feasibly tested with real patients, including the risks anticipated in deep space missions to the Moon and Mars. Despite the benefits of 3D systems for modeling cell types *in vitro*, the complexity of organoids and the labor-intensive nature of drug testing have limited the widespread application of these models.

To evaluate Squidiff's capabilities in the characterization of cellular dynamics and facilitating candidate drug discovery, we constructed blood vessel organoids and probed the potential mechanisms of neutron radiation damage, associated with secondary radiation sources present in deep space. Using Squidiff, we predicted the molecular profiles of various cell compositions within blood vessel organoids throughout their differentiation and development, including endothelial cells, fibroblasts, and mural cells – the principal components of blood vessels. We validated these predictions with real experimental single-cell sequencing. Interestingly, Squidiff suggested a potential developmental pathway from mural cells to endothelial cells, consistent with recent observations in a time-series study of blood vessel organoids[18]. It also predicted the effects of irradiation on various cell types by generating the single-cell transcriptomic

data upon irradiation, even when only limited data on irradiated cell types were available, and identified the affected signaling pathways at each time point of injury. By incorporating an FDA-approved radioprotective drug, G-CSF, Squidiff predicted that G-CSF might have a protective effect by promoting vascular specification compared to irradiated groups. These findings provide a detailed assessment of cell heterogeneity and differentiation trajectories in vessel organoids exposed to radiation and G-CSF proteins. Squidiff thus shows the power of characterizing high-resolution cell dynamics under environmental changes by leveraging the strengths of diffusion models.

# Results

## Squidiff predicts transcriptomic changes through denoising

Squidiff is a diffusion model designed to predict single-cell transcriptomic changes in response to cell state changes such as cell differentiation, development, and exposure to various physical or chemical stimuli across different cell types (Fig. 1A). Leveraging a neural network architecture and continuous denoising processes, Squidiff can predict both future and past cell states when stimuli are specified, and can further predict the effects of multiple drug combinations as well as multiple gene perturbations.

Squidiff integrates a diffusion model for data generation with a variational autoencoder (VAE) for encoding latent representations[10, 12, 18]. This integration enables the generation of new transcriptomic data while modulating cellular states and conditions using latent variables (Fig. 1B). Specifically, the VAE serves as an encoder that maps cell data as well as treatments into a unified latent representation space comprising semantic variables ($z_{sem}$), while the diffusion model can generate any target cell conditioned on the semantic variable via a standard denoising process (see Methods for details). Overall, Squidiff is capable of generating transcriptomic data reflecting cell-type variations, cell-state transitions, and cell-type-specific responses to multiple stimuli such as drug and gene perturbations.

To explain the Squidiff framework, we first show an application to synthetic data generated using Splatter[19]. We simulated single-cell transcriptomics data for three cell types, introducing noise at each step to visualize the diffusion processes (Supplementary Fig. 1). The semantic latent variable ($z_{sem}$) captured meaningful variations in gene expression associated with specific cell states or responses to stimuli, resulting in a clear separation between different conditions in the latent space(Fig. 1C). Additionally, Squidiff can encode input data into a stochastic subcode $x_T$ by running its deterministic generative process backwards. However, $x_T$ primarily captures stochastic details rather than retaining the semantic information of the input (Fig. 1C), as the diffusion process introduces noise to the input data. The reverse diffusion process then denoises and reconstructs the data given the condition, successfully recovering the original transcriptomic profiles (Fig. 1D), demonstrating the model's ability to accurately capture and reproduce cellular states.

To generate new single-cell gene expression data over time and in response to stimuli, Squidiff employs two methods of latent manipulation: interpolation and addition (Fig. 1E). Addition involves combining the original latent representation with

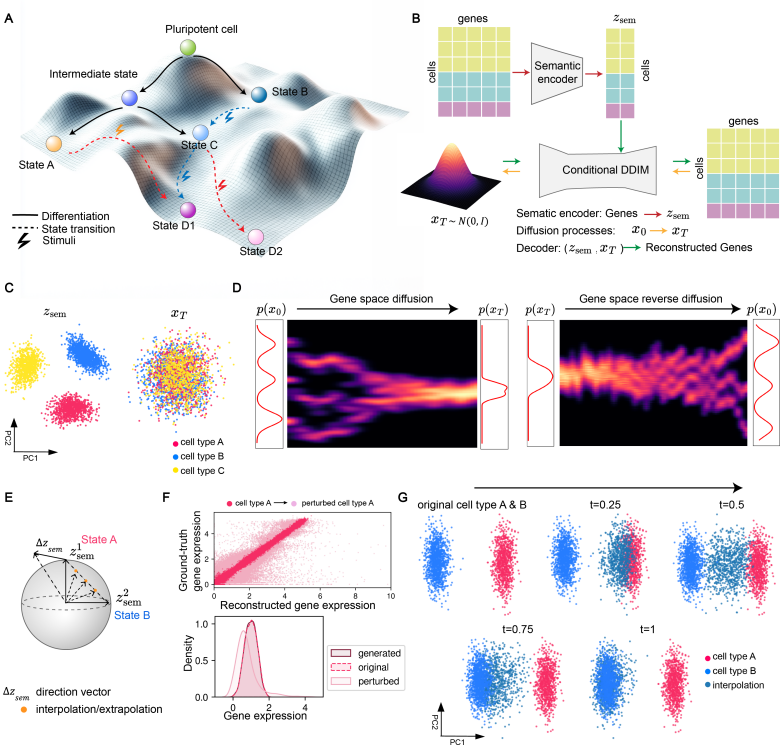



**Fig. 1**: **Overview of Squidiff and its performance on simulated data. A**, Waddington's landscape illustration depicting cell differentiation and transdifferentiation paths from a pluripotent cell to various states (A, B, C, D1, and D2). Solid black lines represent differentiation, while dashed lines indicate state transitions due to external stimuli. **B**, Diagram of the Squidiff model architecture, based on diffusion autoencoders [14]. The model consists of a semantic encoder to map single-cell RNA sequencing data into a semantic latent space ($z_{sem}$), and a diffusion process to generate $x_T$. The conditional denoising diffusion implicit model (DDIM) is employed to decode the latent variables ($z_{sem}$, $x_T$) back into reconstructed gene expressions. **C**, Principal component analysis (PCA) of latent representations: $z_{sem}$ (left) shows clustering of cell types A, B, and C, while $x_T$ (right) displays stochastic variations across the same cell types. **D**, Visualization of the forward diffusion process (left) and reverse diffusion process (right) in gene space. Probability distributions $p(x_0)$ and $p(x_T)$ illustrate the transition from the original gene space to a latent representation through noise, and back to the reconstructed space via denoising. **E**, Schematic of latent space manipulation for generating new cell states. Arrows represent direction vectors ($\Delta z_{sem}$) used for interpolation and extrapolation between states, with the spherical structure indicating the semantic latent space. **F**, Top: Correlation between ground-truth and reconstructed gene expression levels for cell type A and its perturbed state, showing high accuracy in predictions. Bottom: Density plot comparing the distribution of gene expression values between generated, original, and perturbed data, indicating successful reconstruction by Squidiff. **G**, UMAP visualization of cell types A and B, illustrating the temporal progression of interpolated states from $t = 0$ (original) to $t = 1$ (fully interpolated) . Intermediate time points ($t = 0.25$ , $t = 0.5$ , $t = 0.75$) show the gradual transition between cell types A and B in the latent space.

a perturbed direction $\Delta z_{sem}$. This results in a shift in the gene expression distribution, reflecting the effects of the perturbation (Fig. 1F). Interpolation, on the other hand, primarily uses Linear Interpolation (Lerp)[6], which computes intermediate points along a straight line between two vectors, offering computational efficiency. For instance, during interpolation, the blue-green cluster transitions gradually from the red to the blue cluster as time $t$ increases, with Lerp providing a direct path (Fig. 1G). We further validated Squidiff's performance across three biomedical scenarios, including cell differentiation or trans-differentiation, gene perturbation, and drug response predictions.

## Squidiff predicts cell differentiation

Squidiff demonstrates the ability to predict transcriptomic profiles of specific cell states when the transcriptome of a starting point and the latent embedding learned by Squidiff for the applied stimuli are known. This capability is particularly useful for investigating the evolution of cells or missing intermediate states in cellular differentiation and development. The approach leverages the generative capabilities of the diffusion model and its manipulation of semantic latent variables.

Applied to a public single-cell transcriptomics dataset of iPSC differentiation as an example[20], Squidiff effectively captures the state changes from iPSCs to mesendoderm and definitive endoderm cells over a period from day 0 to day 3 (Supplementary Fig. 2). The model was trained on a random split of data at day 0 and day 3, extracting semantic latent variables $z^0_{sem}$, and computing $\Delta z_{sem}$ by subtracting $z^3_{sem}$ from day 3 on $z^0_{sem}$. This difference represents the direction vector for the averaged stimuli over the period (Fig. 2A). Starting with the initial states at day 0 and applying the learned stimuli direction, Squidiff accurately predicts single-cell transcriptomics from day 1 to day 3 (Fig. 2B-C). Additionally, applying the stimuli vector to the predicted transcriptomics at day 1 results in an accurate prediction of cell states at day 2.

To further assess the biological relevance of the model's predictions, we conducted differential gene expression analysis, identifying day-specific genes with distinct signatures corresponding to mesendoderm and pre-definitive endoderm stages (Fig. 2D). For instance, *NANOG* emerged as a pluripotency marker at day 0[21], with its expression diminishing over subsequent days. In contrast, *GATA6*, a key transcription factor involved in the transition towards both mesodermal and endodermal fates[22], showed a progressive increase. By applying pseudotime analysis to the combined input data from day 0 and the model-predicted data from days 1 to 3, we observed matched fluctuations in cell proportions, with a continuous decline in *NANOG* and a corresponding rise in *GATA6* expression. We further examined the gene *T*, a mesodermal marker[23], which was enriched on days 1 and 2, indicating its role in early mesoderm differentiation (Fig. 2E-F). These findings underscore the nonlinearity of gene expression dynamics during these developmental processes and highlights the power of the diffusion model to learn the intricate gene relationships. It is worth noting that the latent variable embeddings from the ground-truth data of day 2 and day 3 do not exactly match the interpolated variables, likely due to the different growth factors applied each day in this experiments. This suggests that $z_{sem}$ represents an averaged trajectory.

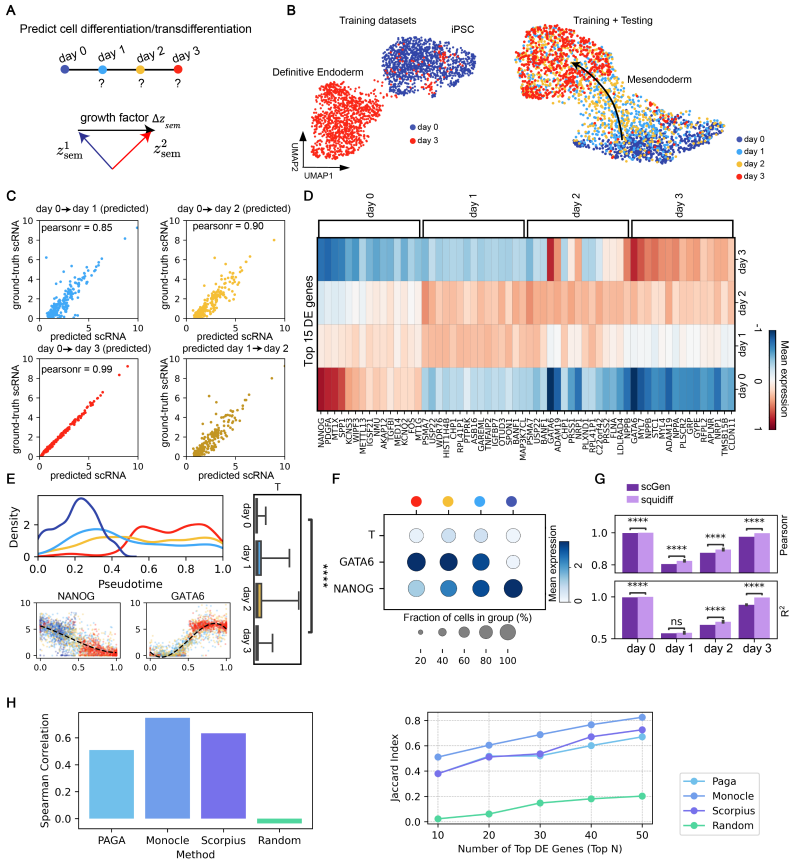



**Fig. 2**: **Squidiff predicts cell differentiation.** **A**, Schematic of cell differentiation/transdifferentiation prediction using Squidiff. The model extracts semantic latent variables ($z^1_{sem}$ and $z^2_{sem}$) and computes the difference ($\Delta z_{sem}$) to define the direction vector over the differentiation period. **B**, UMAP visualization of training datasets with iPSCs (day 0) and definitive endoderm (day 3) in blue and red, respectively. Combined training and testing datasets demonstrate differentiation from iPSCs to mesendoderm and definitive endoderm. **C**, Pearson correlation plots comparing model-predicted and ground-truth single-cell RNA-seq data across different days. High correlation coefficients indicate accurate predictions by Squidiff. **D**, Heatmap of the top 15 differentially expressed (DE) genes across days 0 to 3, illustrating the expression trends of key genes like NANOG and GATA6 during differentiation. **E**, Pseudotime analysis showing cell density distributions along pseudotime (top), with changes in expression patterns of *NANOG* and *GATA6* (middle and bottom). Density distribution of latent variables ($z_{sem}$) indicates differentiation progression. **F**, Dot plot showing the fraction of cells expressing *GATA6* and *NANOG* over different groups and days, reflecting their roles in cell state transitions. **G**, Bar plot of R-squared values comparing the prediction accuracy of Squidiff and scGen across all days, demonstrating Squidiff's superior predictive capability. Two-sided independent two-sample t-test was performed. $P < 1 \times 10^{-30}$, $= 1.48 \times 10^{-9}$, $5.79 \times 10^{-7}$, $< 1 \times 10^{-30}$, sequentially. $^{****}P < 0.0001$. NS, not significant. **H**, Left: Spearman correlation between predicted cell orders from trajectory methods (PAGA, Monocle, Scorpius), random orders, and actual day-time points used in Squidiff. Right: Jaccard index of top differentially expressed genes from day 0 to day 4 across pseudotime series identified by PAGA, Monocle, Scorpius, random orders, and the actual time points used in Squidiff.

Squidiff exhibited superior predictive performance compared to scGen [6], a previous state-of-the-art model that combines variational autoencoders and latent space vectors for high-dimensional single-cell gene expression data (Fig. 2G). When trajectory inference methods such as PAGA[24], Scorpius[25], and Monocle[26] were applied to the Squidiff output transcriptomics, we observed successful reconstruction of the dataset along a continuous pseudotime that closely matched the real discrete timepoints (days), surpassing random baselines. Moreover, there were significant overlaps between the genes identified based on real-time conditions and those predicted by these trajectory models (Fig. 2H).

## Squidiff predicts gene and drug perturbation

Another application of Squidiff is predicting cell responses to gene perturbations or drug treatments. Unlike scGen, which requires training to acquire the latent representation of given conditions, Squidiff can generalize a trained model across different conditions. For example, given a model trained exclusively on single-cell RNA transcriptomics perturbed by two genes, scGen cannot acquire the latent representation of control data without additional training. However, Squidiff can extract embedding and use the learned representation of individual gene perturbation to successfully predict the transcriptomic change of wild-type cells upon these perturbation (Supplementary Fig. 3).

Furthermore, given the semantic and stochastic latent variables for two-gene perturbations, Squidiff assumes that two-gene perturbations are represented as the sum of two latent variables (Fig. 3A). Genes interact synergistically to produce non-additive genetic effects upon multiple gene perturbations, which cannot be accurately modeled by additive approached. We tested Squidiff on an example involving non-additive perturbation using single-cell RNA transcriptomics of K562 cells perturbed by *ZBTB25* and *PTPN12*, as discussed by GEARS[8]. Despite not using graph knowledge as a model prior, as GEARS[8], Squidiff achieved highly accurate predictions (Fig. 3B). Comparative analysis with GEARS and scGen revealed that Squidiff consistently produced more robust results (Fig. 3C-D). Thus, Squidiff exhibits significant capability in predicting gene perturbations and elucidating underlying mechanisms.

Besides gene perturbation, we further tested Squidiff's performance by applying it to drug screening (Fig. 3E). Unlike gene perturbations, where the affected gene lists are known, drug effects are often complex, involving influences on multiple genes concurrently and the regulation of associated pathways. Thus, a model capable of predicting the overall cellular response to drugs is valuable for exploring drug mechanisms and identifying the most effective drugs with minimal side effects. A study involving drug tests on glioblastoma highlighted the distinct effects of Etopside and Panobinostat on tumor cells. We trained the Squidiff model on randomly selected myeloid cells with six drugs, untreated tumor cells, and oligodendrocytes (Fig. 3F, Supplementary Fig. 3). The model showed strikingly accurate predictions of the six drug effects on tumor cells and oligodendrocytes, demonstrating strong performance in predicting transcriptomic changes upon drug treatment (Fig. 3G). Squidiff also identified Panobinostat as having the strongest effect on tumor cells compared to other cell types among all screened

drugs, consistent with the finding from the reference (Fig. 3H)[27]. To further validate the drug effects on tumor cells, we plotted UMAP embedding of cells with and without treatment and observed the distinct effects of Etoposide and Panobinostat as previously reported (Supplementary Fig. 4).

## Squidiff predicts cell differentiation in blood vessel organoids

To evaluate Squidiff's unique capability in predicting the cascade of transcriptional programs during continuous cell state transitions, such as development and how they are altered with stimuli, we employed organoid technology as our experimental platform. Organoids, differentiated from human pluripotent stem cells, have emerged as a novel 3D biomimetic system for modeling human tissue development[28]. Over the last decade, organoids have been extensively used as disease models and drug testing platforms, including brain, liver, and lung organoids, among others[29]. However, as self-organized systems, organoids are inherently heterogeneous, and the interactions among cells within them are complex[29]. Current single-cell and spatial RNA sequencing techniques have been applied to unravel this cellular heterogeneity and profile gene expression patterns during their differentiation trajectory. Despite these advances, it remains challenging to fully characterize and quantify the entire developmental processes of organoids at any given time point. While pseudotime analysis of single-cell RNA sequencing provides snapshots of organoids development. at specific time points, real-time gene expression profiling across different organoids is challenging due to the labor-intensive workflows and technical expertise requirements for manipulation of these engineered models[29, 30], in addition to the costly sequencing steps. Here, Squidiff addresses these limitations by predicting cell transcriptomic changes during development and in response to stimuli.

Taking Blood vessel organoids (BVOs) as a case study for applying Squidiff, these organoids were differentiated from human iPSCs and directed to develop into endothelial and vascular lineages, enabling a vascular network-like structure to form as early as Day 6 (Supplementary Fig. 5). BVOs have emerged as a novel biomimetic system for modeling human vascular development and disease and serve as a powerful platform for drug discovery[18].

The fate and state transition of cells in blood vessel organoids have been studied in real-time[18], providing a ground truth for comparison. In these studies, progenitor states in early BVOs were observed to bifurcate into endothelial and mural fates, highlighting the potential of BVO mural cells to differentiate into endothelial lineages (Fig. 4A). To test whether Squidiff could replicate these critical transitions, we generated BVOs from healthy human iPSCs to create a training dataset. We fixed the organoids, dissociated them into single cells, and performed single-cell RNA sequencing on day 11. The BVOs were exposed to differentiation processes from day -1, and promoted to mesoderm aggregates from day 3 to day 5, with vascular formation starting on day 6. As a result, we identified fibroblasts, mural cells, endothelial cells, and progenitor cells in the organoids on day 11 (Fig. 4B, Supplementary Fig. 5A).

To predict cell fates and states during blood vessel development, we combined single-cell RNA data from iPSCs and BVOs at day 11 and trained the Squidiff model on these datasets. We then performed interpolation for days 1, 3, 5, 7, 9, 13, 15

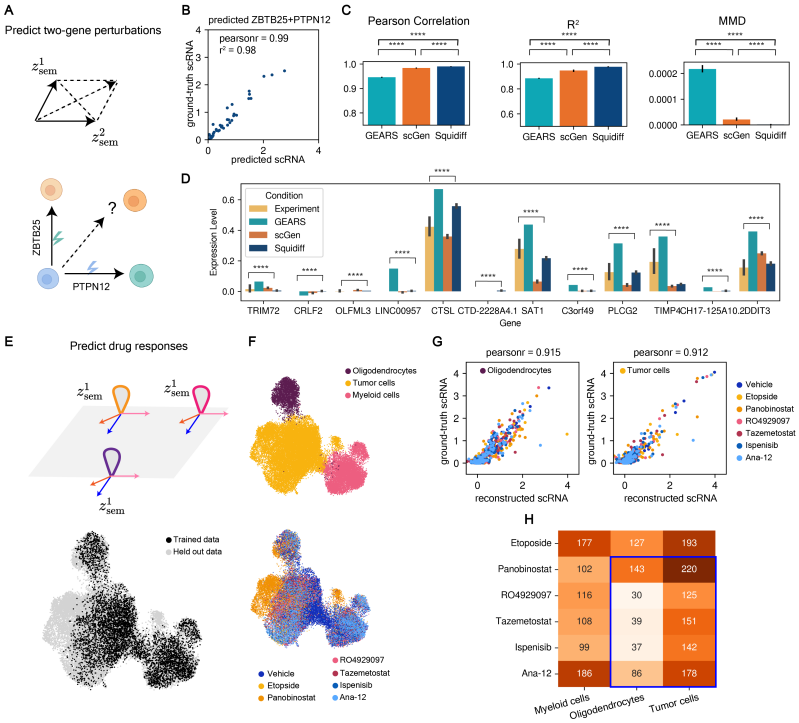



**Fig. 3**: **Squidiff predicts gene and drug perturbation.** **A**, Diagram illustrating the prediction of two-gene perturbations using Squidiff. The approach leverages semantic latent variables ($z^1_{sem}$ and $z^2_{sem}$) to explore combined effects of gene perturbations. **B**, Mean gene expression between Squidiff predicted and real perturbed data with $R^2$ value and Pearson correlation score. **C**, Bar plots comparing the prediction of ZBTB25+PTPN12 on control single-cell data with 203 genes by Squidiff, scGen, and GEARS in terms of Pearson correlation, $R^2$ values, and Maximum Mean Discrepancy (MMD). Squidiff outperforms other methods in both correlation and R-squared metrics, while achieving a lower MMD value. Two-sided independent two-sample t-test was performed. $P = 3.72 \times 10^{-23}$, $2.53 \times 10^{-26}$, $7.92 \times 10^{-11}$ for pearson correlation, $= 6.85 \times 10^{-17}$, $8.35 \times 10^{-25}$, $2.12 \times 10^{-12}$ for $R^2$ correlation, $= 9.28 \times 10^{-23}$, $9.70 \times 10^{-24}$, $5.87 \times 10^{-10}$ for maximum Mean Discrepancy. $^{****}P < 0.0001$. NS, not significant. **D**, Expression levels of key genes across different conditions: control, GEARS, scGen, and Squidiff. Squidiff demonstrates better alignment with the control group in capturing the expression of genes like SMIM1, ESPIN, and JUN, reflecting accurate predictions under perturbation scenarios. One-way ANOVA test across method for each gene was performed.$P < 10^{-30}$ for all tests. $^{****}P < 0.0001$. NS, not significant. **E**, Schematic representation of drug response predictions using Squidiff, highlighting the manipulation of semantic latent variables ($z^1_{sem}$) and capturing responses of held out data. The lower panel shows a UMAP plot of trained and held out data clusters. **F**, UMAP visualization of predicted drug responses across three cell types: oligodendrocytes, tumor cells, and myeloid cells. Different colors represent distinct drug treatments. Squidiff effectively separates cell clusters based on their responses to various drugs. **G**, Mean gene expression between Squidiff predicted and real perturbed data with Pearson correlation score. **H**, Heatmap displaying the response of each cell type (myeloid cells, oligodendrocytes, and tumor cells) to different drugs. The intensity values indicate the number of differentially expressed genes identified in each condition.

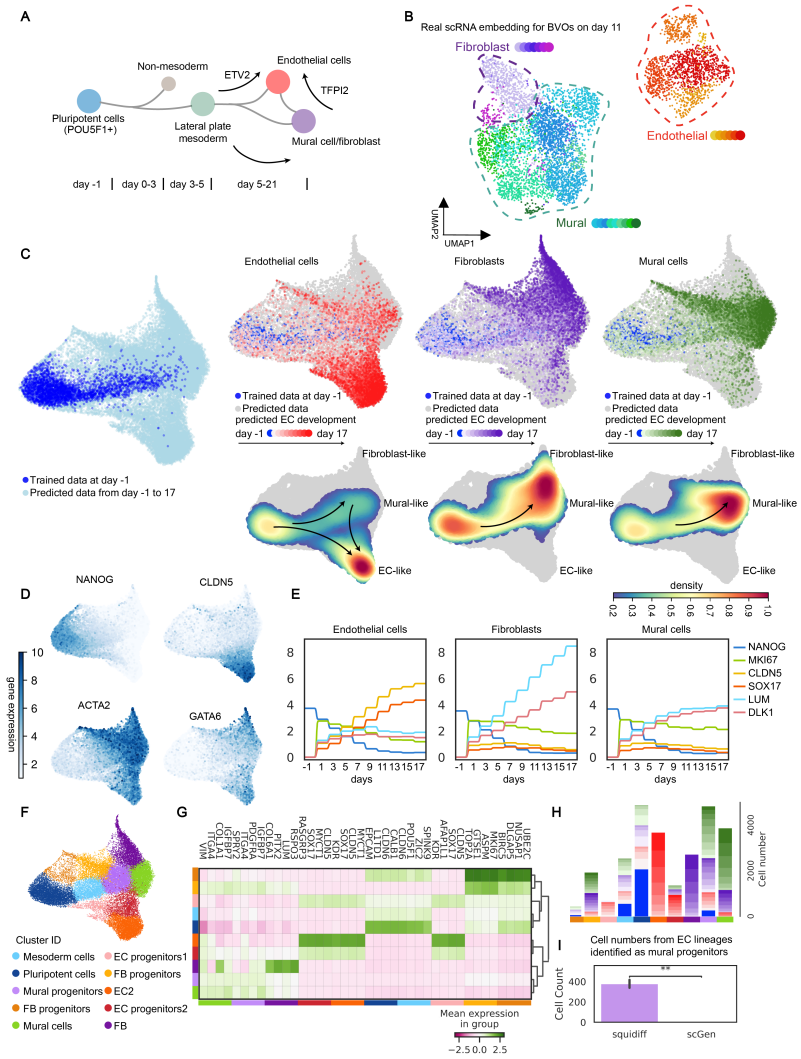



**Fig. 4**: **Squidiff predicts cell differentiation processes in blood vessel organoids. A**, Diagram for hypothesized differentiation processes of iPSC into cell components in blood vessel organoids. **B**, Embedding of single-cell RNA sequencing of blood vessel organoids on day 11 colored by subtypes of endothelial cells, fibroblast and mural cells. **C**, Joint UMAPs of trained scRNA-seq data on day -1, namely the iPSC cells and the predicted scRNA-seq data from day -1 to day 17, with two days intervals with the semantic meaning, which was learned by only endothelial cells at day -1 and day 11. **D**, Joint UMAPs of trained scRNA-seq data on day -1 and the predicted scRNA-seq data from day -1 to day 17, colored by the markers including pluripotent genes (*NANOG*, *MKI67*), endothelial genes (*CLDN5*, and *SOX17*), fibroblast and mural specific genes (*COL1A1*, *LUM*, and *ACTA2*), and the markers denoting differentiating into mesendoderm layer (*GATA6*). **E**, Stepwise expression of markers along with days in fibroblasts, endothelial cells, and mural cells. **F**, Joint UMAPs of trained scRNA-seq data on day -1 and the predicted scRNA-seq data from day -1 to day 17, colored by cell types. **G**, Heatmap of differential gene analysis for clustering of joint trained scRNA-seq data and predicted data. **H**, Stacked barplot of the number of cells colored by the major types and days across cell type clusters. **I**, Mean number of endothelial cell (EC) lineages in mural progenitor and mural cell clusters, comparing Squidiff and scGen. Two-sided independent two-sample t-test was performed. $P = 1.02 \times 10^{-3}$.

and 17 using Squidiff. The predictions captured the differentiation and development trajectories of endothelial and mural cells from iPSCs, as shown in the distribution of marker gene expression (Fig. 4C-D,Supplementary Fig. 6A). Interestingly, the predicted single-cell data revealed intermediate states where some endothelial cells exhibited properties akin to mural cells, while mural and fibroblast cells appeared to derive from mural progenitor cells, which is consistent with the discovery from a public study that experimentally sequenced real-time BVOs[18] (Fig. 4C). To quantify these observations, we tracked the expression of marker genes across days for endothelial, fibroblasts, and mural cells, identifying fluctuations in mural-specific genes, such as LUM and DLK1. These genes showed an increase before day 7 and a decrease after day 9, while they continued to rise in fibroblast and mural cells (Fig. 4E). A heatmap of differentially expressed genes across cell type clusters on the UMAP further confirmed cell type distinctions (Fig. 4F), and indicated the cell state transitions from iPSC to mature cell lineages (Fig. 4G). We observed a substantial proportion of endothelial cells in the mural progenitor population, which resembled the results of real single-cell RNA data[18] (Fig. 4H) . The existence of differentiating EC cells from day 5 to day 9 shown as mural progenitors features suggested the potential of EC differentiated from mural-like cells, a feature that scGEN methods failed to capture(Fig. 4I, Supplementary Fig. 6). This underscored Squidiff's enhanced capability in predicting cell heterogeneity and transient cell states, offering valuable insights that may expedite research efforts through in silico modeling.

## Structure damage and metabolic alterations in blood vessel organoids induced by neutron irradiation

Squidiff demonstrated the ability to predict the cellular responses to stimuli, including gene perturbations and drug perturbations above. We next evaluated whether Squidiff could predict physical perturbations such as irradiation, on different cell types. Ionizing radiation, whether from radiotherapy, nuclear-reactor accidents, or in anticipated deep space travel, presents a significant risk to human health[31]. Radiation exposures impact not only cancerous but also non-cancerous cells, particularly those that are rapidly proliferating, such as endothelial cells, blood and immune cells, and local parenchymal cells within the radiated area[32]. High-energy particles, such as neutrons, can penetrate biological tissues, causing damage at the cellular and molecular levels[33]. Understanding the effects of radiation on human tissues can help develop effective countermeasures and treatments for radiation-induced damage[34]. Degenerative vascular changes have been identified as one of the main radiation health risks to astronaut crews on exploration missions[35]. Microvascular endothelial cells are susceptible to ionizing radiation, and radiation-induced alterations in endothelial cell function are thought to be critical factors in organ damage through endothelial cell activation, enhanced leukocyte-endothelial cell interactions, increased barrier permeability, and initiation of apoptotic pathways[32, 36]. Blood vessel organoids, which mimic the human vascular system, provide an excellent model for studying these effects.

During the BVO development stage on day 5, we exposed the cultures to either neutrons or protons radiation and continued the BVO culture until day 11 (Fig. 5A).

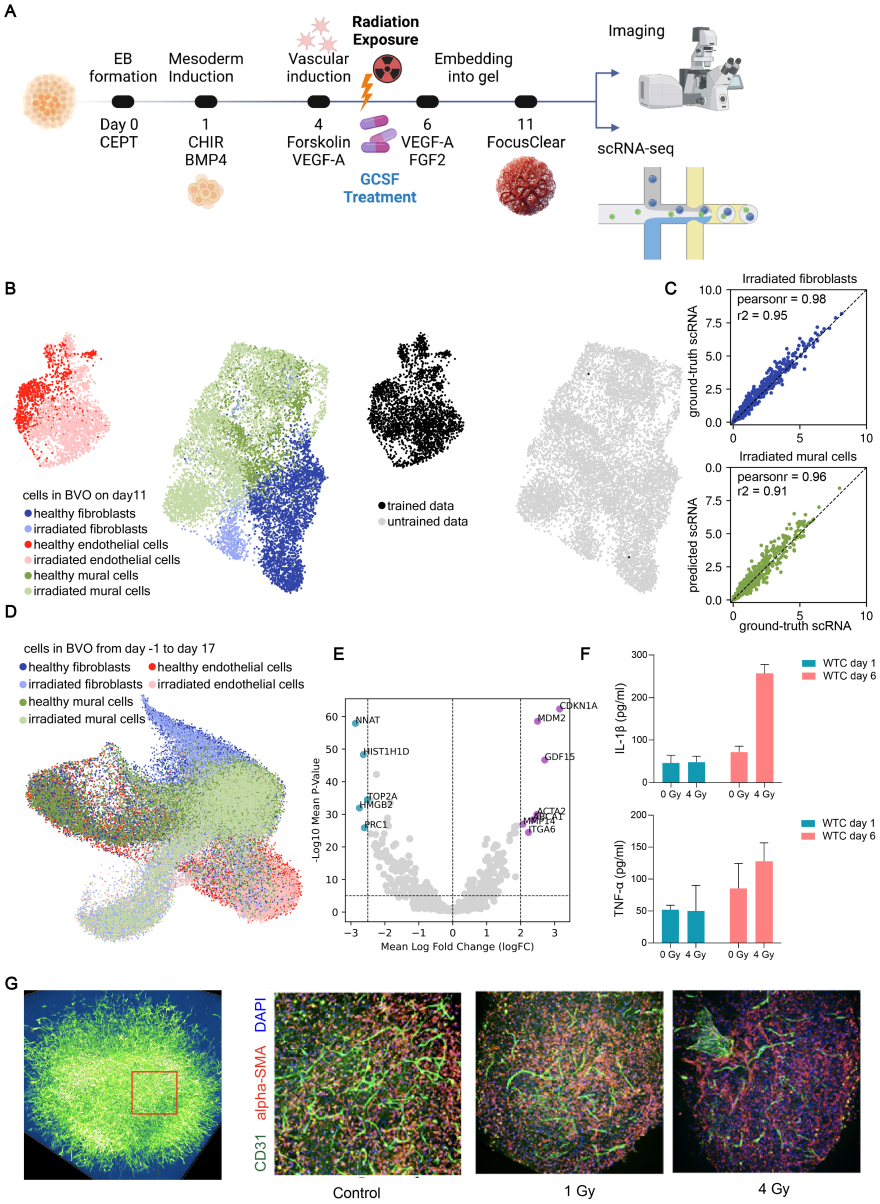



**Fig. 5: Structure damage and metabolic phenotypes alteration in blood vessel organoids induced by neutron irradiation. A**, Experimental workflow for differentiating blood vessel organoids (BVOs) from human pluripotent stem cells. Key steps include embryoid body (EB) formation, mesoderm induction, vascular induction, embedding into gel, and subsequent neutron irradiation or GCSF treatment. Imaging and single-cell RNA sequencing (scRNA-seq) were performed on day 11. **B**, UMAP visualization of single-cell RNA-seq data from BVOs on day 11, showing different cell types: healthy fibroblasts, irradiated fibroblasts, healthy endothelial cells, irradiated endothelial cells, healthy mural cells, and irradiated mural cells. Trained and untrained data sets are indicated. **C**, Scatter plots showing the correlation between ground-truth scRNA-seq data and Squidiff-predicted data for irradiated fibroblasts (top) and irradiated mural cells (bottom), with Pearson correlation coefficients and R-squared values. **D**, UMAP visualization of single-cell RNA-seq data from BVOs across different developmental days (day -1 to day 17), showing the distribution of healthy and irradiated cell types over time. **E**, Volcano plot highlighting differentially expressed genes upon neutron irradiation. Key genes such as CDKN1A and MDM2 are indicated. **F**, ELISA results showing the levels of IL-1$\beta$ and TNF-$\alpha$ in the culture medium of BVOs at day 1 and day 6, under different irradiation conditions (0 Gy, 4 Gy). **G**, Immunofluorescence images of BVOs showing structural changes under control conditions, 1 Gy, and 4 Gy photon irradiation. CD31 (endothelial cell marker) in green, alpha-SMA (smooth muscle cell marker) in red, and DAPI (nuclear stain) in blue.

At day 11, we profiled the cell states using single-cell RNA sequencing for both healthy BVOs and those exposed to neutron irradiation. The single-cell RNA sequencing revealed the presence of endothelial cells, mural cells, and fibroblasts. However, the reduced overlap between the irradiated and non-irradiated cells indicated substantial alterations in the molecular landscape (Fig. 5B). Squidiff effectively learned radiation exposure as a vector in the semantic space, enabling the generation of perturbed transcriptomes for different cell states in response to radiation.

To validate Squidiff's capability in data prediction and biological exploration, we masked the mural cells and fibroblasts entirely during Squidiff training, considering only the endothelial cells under irradiation and control conditions (Fig. 5B). The model successfully generated the transcriptomic data under irradiation with high accuracy (Fig. 5C). We then applied the model to the predicted data interpolated from the healthy sample to acquire the predicted irradiated data for endothelial cells, mural cells, and fibroblasts. The UMAP of all the data indicated these cells at the early development stage might be the most affected stage (Fig. 5D), likely due to the immaturity of cell types and the lack of tissue structure and cell interactions. The ranking of the most differentiated genes upon irradiation across cells indicated continuous changes along the course of BVO developmental days (Supplementary Fig. 7).

Intriguingly, genes behaved differently over days and across cell types. For instance, the importance of *MYCT1* in endothelial cells decreased in the middle period and resumed at the late stage of development, while *MYCT1* remained one of the most differentiated genes in mural cells and fibroblasts (Supplementary Fig. 7E). The most affected genes upon irradiation across all these cells included upregulation of genes such as *CDKN1A, MDM2, GDF15, ACTA2, ABCA1, MMP14*, and *ITGA6*, and downregulation of genes such as *NNAT, HIST1H1D, TOP2A, HMGB2,* and *PRC1* (Fig. 5E). These genes are involved in cellular responses to ionizing radiation, cell motility, DNA damage through P53 pathways, apoptosis, TNF-$\alpha$ signaling via NF-$k$B, KRAS pathways, protein kinase B signaling, and inflammation responses in gene enrichment analysis (Supplementary Fig. 8).

Notably, *CDKN1A* is a key gene implicated in radiation response, known for its role in cell cycle regulation and DNA damage repair. *MDM2* is another significant gene, functioning as a negative regulator of the p53 tumor suppressor, thus playing a critical role in controlling cell proliferation and apoptosis[37]. Using a live-dead assay, the percentage of dead cell was much higher in irradiated BVOs (Supplementary Fig. 9), consistent with the apoptosis process triggered by the radiation effects. The size of BVOs immediately after neutron irradiation and before the vascular induction showed a decreased size (Supplementary Fig. 9A). We found that the culture supernatant post-irradiated was enriched in IL1-beta and TNF-alpha, indicating enhanced induction of inflammation (Fig. 5F), consistent with the increased expression of *GDF15*, a stress response gene associated with the regulation of inflammation and cell survival. These metabolic changes overall resulted in downstream failures in vascular formation and abnormal distribution patterns in BVOs (Fig. 5G), causing earlier vascular sprouting in the radiation BVOs as compared to healthy controls. Overall, Squidiff effectively generated data with biological relevance and predicted the complex molecular responses to radiation exposure.

### Treatment potentials of G-CSF against radiation disruption in blood vessel organoids

In recognition of the disruption of vascular structures, the apoptosis of cells triggered by the p53 pathway, and protein kinase B signaling, we attempted to understand whether the FDA-approved radioprotective drug, granulocyte colony-stimulating factor (G-CSF), could rescue the defects in vasculogenesis *in vitro*. G-CSF is a glycoprotein that plays a crucial role in the production, differentiation, and function of granulocytes, a type of white blood cell essential for fighting infections[38]. It has been used in clinical settings to stimulate the bone marrow to produce more white blood cells, particularly after chemotherapy or bone marrow transplants. Recent studies have suggested that G-CSF may also have protective effects on vascular tissues, promoting endothelial cell survival and enhancing tissue repair processes[39]. A known effect of G-CSF is its activation of MEK1/2, PI3K/Akt and NF-$\kappa$B pathways (Fig. 6A)[40], consistent with findings on the regulation of NF-$\kappa$B and kinase B signaling on radiation injury (Supplementary Fig. 8). However, studies on G-CSF effect on fibroblasts and mural cells in the vasculature remain limited.

Squidiff offers a way to predict G-CSF effect on other cell types when only one cell type (e.g. endothelial) responses are known. We sequenced the single-cell RNA expression in GCSF-treated BVOs and considered it as ground truth data for the model as a case.

Given the training data on endothelial cells only (Fig. 6B), Squidiff successfully generated single-cell transcriptomics for fibroblast and mural cells (Fig. 6C, Supplementary Fig. 10). To further validate the biological relevance of these gene expression predictions, we conducted differential gene expression analysis comparing real irradiated data with predicted G-CSF-treated data. Notably, we identified distinct sets of differentially expressed (DE) genes for each cell type, with some genes shared across multiple types (Supplementary Fig. 10B-C). The biological processes identified for enriched genes revealed unique roles for each cell type in response to G-CSF treatment. Fibroblasts were primarily associated with processes involved in blood vessel morphogenesis and vasculogenesis, indicating their critical role in structural development. Endothelial cells showed enrichment in pathways regulating apoptosis and the cell cycle, highlighting their involvement in maintaining cell survival and proliferation. Mural cells, in contrast, were linked to pathways related to genomic stability, cell cycle progression, and mitosis, indicating their role in maintaining genomic integrity. These findings aligned with the observed reduction in cell death percentages in G-CSF-treated samples compared to those affected by radiation (Supplementary Fig. 9B). This study demonstrates Squidiff's capability to predict cell type-specific responses to drug treatments and provides valuable insights into the effects of G-CSF on irradiated blood vessel systems.

## Discussion

In this study, we present Squidiff, a conditional diffusion model specifically designed to predict transcriptomic changes across diverse cell types in response to a spectrum of environmental changes. Our findings demonstrate that Squidiff provides a robust and

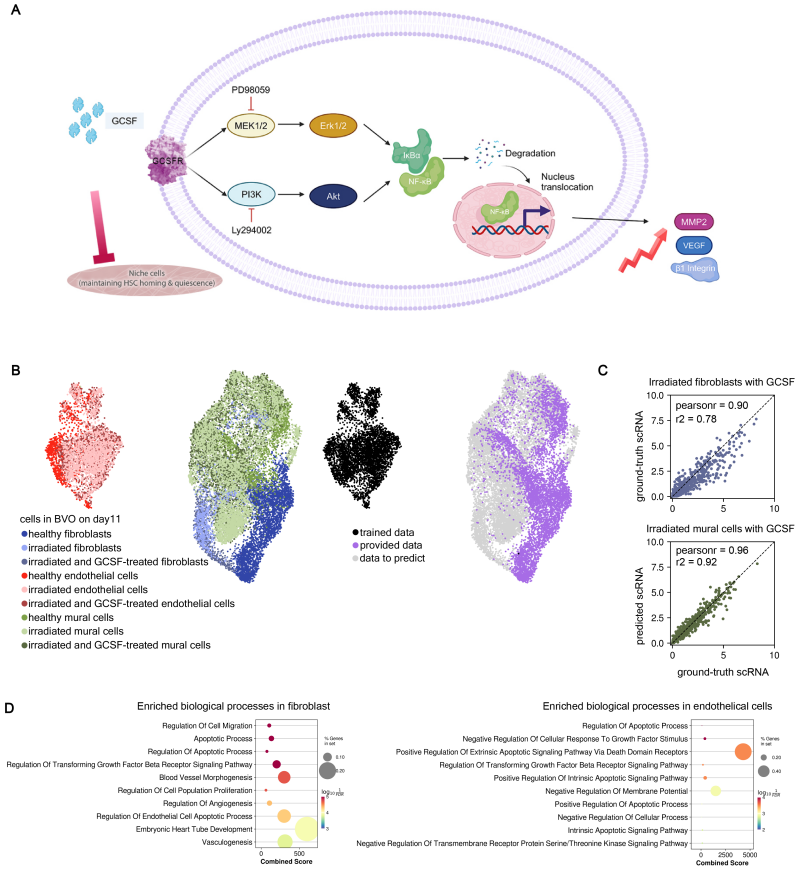



**Fig. 6**: **Treatment potentials of GCSF in securing against radiation disruption in blood vessel organoids. A**, Schematic illustration of the signaling pathways activated by GCSF treatment. GCSF binds to its receptor (GCSFR) on myeloid progenitor cells, triggering downstream signaling cascades involving MEK1/2, Erk1/2, PI3K, and Akt. This leads to the degradation of I$\kappa$B$\alpha$ and nuclear translocation of NF-$\kappa$B, promoting the expression of target genes such as MMP2, VEGF, and $\beta$1 integrin, which are involved in cell survival, proliferation, and migration. **B**, UMAP visualization of single-cell RNA-seq data from BVOs on day 11, showing the distribution of healthy, irradiated, and GCSF-treated cell types. The trained data includes healthy fibroblasts, endothelial cells, and mural cells. Provided data includes irradiated fibroblasts, endothelial cells, and mural cells. Data to predict includes irradiated and GCSF-treated fibroblasts, endothelial cells, and mural cells. **C**, Scatter plots comparing ground-truth scRNA-seq data with Squidiff-predicted data for irradiated fibroblasts (top) and irradiated mural cells (bottom) treated with GCSF. The Pearson correlation coefficients (r) and R-squared values ($r^2$) indicate high prediction accuracy.

accurate framework for estimating cell development trajectories and environmental responses based on high-dimensional cell-state representations. By capturing dynamic changes in cell states over varying conditions and time points through the generation of single-cell transcriptomics, Squidiff offers key insights into cell differentiation and therapeutic responses, positioning itself as a valuable tool in precision medicine.

Compared to existing generative models such as scGen, variational autoencoders (VAEs), Squidiff surpasses these approaches in its ability to model complex cell-state transitions and predict responses to gene and drug perturbations with high temporal resolution. While previous models like scGen rely on latent variable interpolation, their generative processes are often less effective in capturing smooth state transitions, particularly for rare or transient states. In contrast, Squidiff leverages diffusion processes to generate high-fidelity predictions of intermediate and transient cell states, which are crucial for understanding differentiation trajectories and the effects of environmental perturbations. Furthermore, our study highlights Squidiff's capacity to integrate conditional factors, enabling precise control over conditions and making it adaptable across various biological scenarios.

We applied Squidiff to blood vessel organoids to investigate the impact of neutron irradiation and subsequent G-CSF treatment, a radioprotective intervention. The human vascular system plays a central role in maintaining blood flow, nutrient distribution, and metabolic balance across all organs, making vascular health critical for understanding long-term risks in deep space missions. Exposure to cosmic radiation is anticipated to disrupt the sensitive endothelium, leading to structural damage, altered cell phenotypes, and diverse pro-inflammatory or pro-regenerative responses. Previous studies have already suggested the risk of vascular malformations and excessive angiogenesis under these conditions, with implications for cancer and other diseases [41–43].

In our work, we utilized advanced in vitro 3D models consisting of iPSC-derived organoids alongside Squidiff's predictive capabilities to elucidate the temporal dynamics of cellular responses. One key advantage of Squidiff is its ability to predict and validate cell type-specific responses using a single timepoint dataset, thus circumventing the high costs and sample heterogeneity challenges of repeated single-cell sequencing across multiple timepoints.

Using human BVOs as a model system, Squidiff accurately identified radiation-induced damage to endothelial cells and elucidated the mechanisms underlying G-CSF treatment in promoting endothelial cell maturation over stromal cell differentiation post-radiation injury. Specifically, we found that BVOs subjected to neutron irradiation exhibited significant shifts in differentiation trajectories, leading to an increased generation of mural cell populations that disrupted vascular formation during organoid maturation. Despite single-cell RNA sequencing being performed one week post-irradiation, Squidiff successfully reconstructed the precise cell-specific changes occurring during the initial acute phase of injury. This capability allowed us to identify temporal gene responses that could serve as promising drug targets for radioprotection.

Moreover, our analysis revealed the potential therapeutic effects of G-CSF proteins on vascular tissues, providing insights into enhancing G-CSF delivery efficacy and therapeutic index in vivo. The discovery of these radiation countermeasures opens avenues for future studies exploring G-CSF and other protective agents in space medicine and related fields.

Despite the promising results, there are several limitations in the application of diffusion models like Squidiff. The training process requires introducing Gaussian noise into the data, which leads to prolonged training times. Additionally, the computational resource requirements for diffusion models are higher compared to other generative frameworks such as VAEs or GANs. These challenges underscore the need for optimizing training protocols and exploring more efficient implementations of diffusion models.

In the future, improving the scalability and speed of Squidiff will be essential to broadening its application to large-scale single-cell datasets. Additionally, while Squidiff proved effective in identifying cell-state transitions and responses to specific perturbations, further validation using in vivo models and more diverse datasets would strengthen its applicability to clinical settings. Integrating Squidiff with multi-modal omics data, such as proteomics and epigenomics, could also enhance its predictive power and uncover novel regulatory mechanisms driving cell fate decisions.

# Methods

Diffusion probabilistic models (DPMs) have proven to be powerful methods for generating new data in various domains, such as images[12], videos[44], and protein structures[45] in biomedicine. These models have also shown high potential in single-cell reconstruction and inference[13], although their actual implementation in genetics is still limited. This limitation may be due to the unique challenges of applying DPMs to generate gene expression data, particularly the need for semantic manipulation and meaningful interpretation of the generated expression profiles. Recent advances in combining DPMs and variational autoencoders (VAEs) have demonstrated effective representation learning in the image domain[14]. Building on these advancements, we adapted the diffusion autoencoder model, also a conditional denoising diffusion implicit model (DDIM) model for single-cell gene expression prediction under various perturbations. The goal is to learn a semantically rich latent space that allows smooth interpolation while maintaining the reconstruction capability that the diffusion model excels at.

## Squidiff model

Similar to other diffusion autoencoder models, the goal of Squidiff is to model the target distribution of perturbed gene expression by learning a denoising process at varying noise levels. The diffusion process and reverse process are the two major components of the squiff model.

### *Diffusion process*

Given input data, the gene expression vector in an individual cell is denoted as $x_0$. We first define the Gaussian diffusion process, which increasingly adds noise to $x_0$ at time t, t$\in$ {0,1,2,3,..., T}. The forward diffusion process is defined as: $q(x_t|x_{t-1}) = N(\sqrt{1-\beta_t}x_{t-1}, \beta_t I)$, where $\beta_t$ are hyperparameters representing the noise levels. This leads to $q(x_t|x_0) = N(\sqrt{\alpha_t}x_0, (1-\alpha_t)I)$, and $\alpha_t = \Pi_{s=1}^{t}(1-\beta_s)$. In our implementation, these parameters are defined as follows:

- $\beta_s$ values are linearly spaced between 0.001 and 0.01.
- The cumulative product of $\beta_s$ values is computed to obtain $\alpha_t$
- The square root of these cumulative products and their complements are used to scale $x_0$ and the noise, respectively.

The forward diffusion step, which adds noise to the gene expression vector, follows the equation: $x_t = \sqrt{\alpha_t}x_0 + \sqrt{(1-\alpha_t)}\epsilon$, where $\epsilon$ is Gaussian noise.

### *Reverse process*

The reverse process of DPMs aims to learn the noise distribution $p(x_{t-1}|x_t)$, which can be approximated as $\sim N(\mu_\theta(x_t,t), \sigma_t^2 I)$. The mean $\mu_\theta(x_t,t)$ is usually defined as: $\mu_\theta(x_t,t) = \frac{1}{\sqrt{\alpha_t}}(x_t - \frac{\beta_t}{\sqrt{1-\alpha_t}}\epsilon_\theta(x_t,t))$, where $\epsilon_\theta(x_t,t)$ is the noise predicted by the model. DDIM suggests a deterministic generative process:

$$p(x_{t-1}|x_t) = \mathcal{N}(\sqrt{\alpha_{t-1}}(\frac{x_t - \sqrt{1-\alpha_t}\epsilon_\theta^t(x_t)}{\sqrt{\alpha_t}}), 0)$$

And derived to

$$p(x_{t-1}|x_t, x_0) = \mathcal{N}(\sqrt{\bar{\alpha}_{t-1}}x_0 + \sqrt{1-\bar{\alpha_{t-1}}}\frac{x_t - \sqrt{\bar{\alpha_t}x_0}}{\sqrt{1-\bar{\alpha}_t}}, 0)$$

where $x_0$ is the predicted clean imaging from $x_t$ using the noise estimate $\epsilon_\theta(x_t,t)$.

This ensures a deterministic transformation from $x_t$ back to $x_{t-1}$, and also enables a backward reverse process, allowing the encode of x0 to latent code $x_T$. However, $x_T$ typically does not contain high-level semantics. In other words, the interpolation between two latent variables $x_T$ does not provide a semantically-smooth change in the generative process, as validated in generated images[14].

To mitigate the high dependency on the input $x_0$ for $x_T$, the reverse process for generating new transcriptome in Squidiff is conditioned on an extra latent variable $z_{sem}$.The objective is to learn $p(x_{t-1}|x_t, z_{sem})$, where $z_{sem}$ is derived from a semantic encoder of variational autoencoder (VAE). Thus the conditional DDIM decoder takes $z$=($z_{sem}$, $x_T$) as input, with the entire reverse process described as:

$$\begin{aligned}
&p_\theta(x_{0:T}|z_{sem}) = p(x_T)\Pi_{t=1}^{T}p_\theta(x_{t-1}|x_t, z_{sem}),\\
&p_\theta(x_{t-1}|x_t, z_{sem}) = \mathcal{N}(f_\theta(x_1, 1, z_{sem}, 0)), if t = 1,\\
&p_\theta(x_{t-1}|x_t, z_{sem}) = q(x_{t-1}|x_t, f_\theta(x_t, t, z_{sem}))otherwise
\end{aligned}$$

The function $f_\theta$ denotes the deterministic transformation based on both the latent variable $z_{sem}$ and the current state $x_t$. The function $f_\theta$ is parameterized by $\epsilon_\theta(x_t, t, z_{sem})$:

$$f_\theta(x_t, t, z_{sem}) = \frac{1}{\sqrt{\bar{\alpha}_t}}(x_t - \sqrt{1-\bar{\alpha}_t}\epsilon_\theta(x_t, t, z_{sem})) \tag{1}$$

### *Semantic encoder*

The goal of the semantic encoder is to provide $z_{sem}$ for the reverse process to generate transcriptomics conditioned on $z_{sem}$. The semantic encoder is designed to capture high-level semantic features from the input data, which are then used to condition the reverse diffusion process, ensuring that the generated transcriptomics data adheres to the desired semantic characteristics.

Our semantic encoder is implemented as a Multilayer Perceptron (MLP) with conditional batch normalization layers and a residual connection for embedding time and semantic features. The architecture consists of two linear layers with ReLU activations, each followed by a conditional batch normalization layer. The conditional batch normalization layers adaptively normalize the input features based on the batch statistics, which is particularly useful when the batch size is small or when there is variability in the input data distribution.

In addition, a residual connection is incorporated to embed both the time and semantic features, enhancing the model's ability to capture temporal dynamics and complex semantic relationships in the data. This residual connection helps maintain the integrity of the original input features while allowing the model to learn more abstract representations.

The semantic encoder is trained as part of the overall model. During training, the semantic encoder takes the input data and produces $z_{sem}$, which is then used in the reverse diffusion process to conditionally generate the transcriptomics data. The training process ensures that the generated data is semantically coherent and adheres to the desired characteristics specified by $z_{sem}$. $z_{sem} = Enc(x_0)$.

### *Sinusoidal position embeddings*

To incorporate temporal information, we use sinusoidal position embeddings. These embeddings use sine and cosine functions of varying frequencies to encode positional information, providing scale invariance, smoothness, and a wide range of frequencies. This method is efficient and parameter-free, ensuring that the positional information is effectively captured without additional learned parameters.

### *Training*

The training is done by optimizing $L = \sum_{t=1}^{T} E_{x_0,\epsilon_t}[||\epsilon_\theta(x_t, t, z_{sem}) - \epsilon_t||]_2^2, \epsilon_t \sim N(0, 1)$ The noise predictive function is MLP for single-cell transcriptomics. The MLP is trained using the Adam optimizer with a learning rate of 1e-3.

### *Stochastic encoder*

The backward of the deterministic generative process enables the encode of input gene expression $x_0$ to the stochastic code $x_T$, followed by: $x_{t+1} = \sqrt{\bar{\alpha}_{t+1}} f_\theta(x_t, t, z_{sem}) + \sqrt{1 - \bar{\alpha}_{t+1}} \epsilon_\theta(x_t, t, z_{sem})$ To note, the stochastic encoder is not necessary for generating meaningful biological transcriptomics data, but the use of a stochastic encoder in the interpolation produces a more precise prediction of single-cell RNA transcriptomics due to the learning of low-level latent embedding.

## Processing of single-cell RNA sequencing data for model training

In the section on the prediction of iPSC differentiation, we first log normalized the gene counts and selected the top 500 variable genes for the task considering the computing capabilities. The training data are split into 60%, 30%, and 10% for the train set, test set, and validation set, respectively, which turns out to be 4800 cells for training when the data on day 1 and day 2 were not masked out, 2400 cells for testing, and 800 cells for validation. We then add Gaussian noise to the training data to determine the number of steps in the diffusion process (Figure S2). The results indicate the optimal number of diffusion steps. The whole model is trained on one H100 with 80 GB memory.

## Reprogramming and culture of human pluripotent stem cells (iPSCs)

We cultivate the wild-type iPSC line (WTC) using mTeSR plus medium on growth factor-reduced Matrigel-coated 6-well plates. The cells are usually passaged every 3-4 days based on the proliferation rate.

## Generation of blood vessel organoids

To model the vascular malformations seen in PS, we generated blood vessel organoids using previously established protocol[37, 46]. The embryoid body (EB) was first formed in Aggrewell® from iPSCs obtained one day prior to the induction of differentiation (Day -1). On day 0, CHIR and BMP4 were added to induce the differentiation of embryoids into the mesoderm germ layer. On day 3, vascular differentiation was induced in the organoids by applying VEGFA and forskolin. The differentiated organoids were then embedded in Matrigel® on day 5 and cultured in medium containing VEGFA, FGF2, and FBS. The organoids were fixed on day 11 and stained for light sheet microscopic imaging[47]. We stained for endothelial and mural cell markers (CD31+, PDGFR$\beta$+, $\alpha$SMA+, respectively) to confirm the generation of a vascular network constructed by endothelial tubes and mural cells, with visualization of the spatial arrangement of these cells in the vasculature.

## Radiation exposure

On Day 5 of BVO generation, we performed radiation exposure with either 1 Gy neutrons or 4 Gy photons and the BVOs continued to be cultured until day 11. Neutron irradiations were performed at the Columbia University Radiological Research Accelerator Facility (RARAF), using an accelerator-based neutron irradiator[48], as described previously[49]. The system was designed to mimic the neutron energy spectrum from an Improvised Nuclear Device[48, 50], but has also been used to model the high linear energy transfer components of galactic cosmic radiation[51]. Dosimetry was performed on the day of irradiation, as previously described, using a custom-built Tissue Equivalent Proportional Counter (TEPC) which measured the total dose, and a compensated Geiger–Mueller dosimeter, which had a very low response to neutrons and thus measures only the photon component. Photon irradiations were performed at Columbia University's Center for Radiological Research using a Gammacell 40 137Cs irradiator (Atomic Energy of Canada Ltd) with a dose rate of 0.68 Gy/min for varying low-energy doses (i.e., 4 Gy).

## Drug treatment

Following radiation exposure, the medium from the blood vessel organoids was collected and replaced with fresh vessel organoid medium. GCSF at a concentration of 80 ng/mL was then administered to the samples immediately.

## Cytokine testing

Culture medium was collected on days 0, 3, and 6 post-radiation to assess cytokine levels. The concentrations of TNF-$\alpha$ and IL-1$\beta$ in the irradiated blood vessel organoid (BVO) medium were measured using ELISA kits (Thermo Fisher Scientific). The wash buffer was prepared by diluting the concentrated solution with distilled water according to the instructions provided with the ELISA kits. Standard cytokine solutions were prepared by serially diluting the provided standards with the assay diluent to generate a standard curve. Irradiated BVO medium samples were added to each well of the 96-well microtiter plates. The plates were covered and incubated for 2 hours at room temperature. Following incubation, the detection antibody was added, and subsequent steps were performed according to the manufacturer's instructions. The absorbance of each well was measured at 450 nm using a plate reader.

## Whole-mount Immunocytochemistry

The blood vessel organoids were permeabilized with blocking buffer (1.5 mL FBS, 0.5 g of BSA, 250 µL of Triton X-100, 250 µL of Tween 20, and 47.5 mL of PBS in 50 mL buffer) and incubated with primary antibody overnight at 4°C. The samples were then washed with PBS-T solution and embedded in mounting medium.

## Confocal imaging

The blood vessel organoids were permeabilized with blocking buffer (1.5 ml FBS, 0.5 g of BSA, 250 $\mu$L of Triton X-100, 250 $\mu$L Tween 20 and 47.5 mL of PBS in 50

mL buffer) and incubated with primary antibody overnight at 4°C. The samples were then washed with PBS-T solution and embedded in mounting medium. A confocal microscope with a 10X objective was used to image the vascular organoids.

## Single-cell sequencing

Single-cell RNA sequencing was performed in BVOs on day 11. Blood vessel organoids were dissociated into single cells using enzymatic digestion with a dissociation buffer containing trypsin or collagenase. The dissociated cells were counted using an automated cell counter, and cell viability was assessed using trypan blue exclusion. Single-cell suspensions were then loaded into a microfluidic device, such as the 10x Genomics Chromium Controller, to capture individual cells in nanoliter-scale droplets containing reagents for cell lysis and reverse transcription. Unique molecular identifiers (UMIs) and cell barcodes were incorporated to enable accurate identification and quantification of transcripts from individual cells. After reverse transcription, the cDNA was amplified using polymerase chain reaction (PCR) and purified. Sequencing libraries were constructed from the amplified cDNA using a standard library preparation kit and indexed with unique sample barcodes for multiplexing. The libraries were sequenced on a high-throughput platform Illumina NovaSeq, to generate paired-end reads optimized for comprehensive coverage and accurate quantification of gene expression.

The raw sequencing data were processed using the Cell Ranger pipeline (version 6.1.2) to demultiplex reads, collapse duplicates based on UMIs, align reads to the GRCh38 reference transcriptome, and quantify gene expression levels for each cell. Quality control steps removed low-quality cells and potential doublets, and the high-quality single-cell RNA-seq data were analyzed for normalization, clustering, differential expression, and trajectory inference to elucidate the transcriptomic changes in response to different treatments.

**Supplementary information.** Figures S1-S10

**Acknowledgements.** We thank Konpat Preechakul for fruitful discussions in diffusion model. We also thank Sabrina Wang, Jong Ha Lee, and Rayna Batya-Leia Berris for help with single-cell RNA sequencing analysis and cell culture. We thank the staff at Columbia's Center for Radiological Research for assistance in running ionizing radiation equipment. These studies used the resources of the Herbert Irving Comprehensive Cancer Center Confocal and Specialized Microscopy Shared Resource funded in part through Center Grant P30CA013696. The Columbia IND neutron Facility was developed under NIAID grant number U19 AI067773. We gratefully acknowledge the funding support of this work by the Translational Research Institute for Space Health (TRISH/NASA) agency (RAD0104 and NNX16A069A to G.V.-N., D.J.B., K.L.).

**Author contributions.** K.L., E.A. and J.Z. conceived the study and provided overall supervision of the study. S.H., Y.Z., D.T. designed the blood vessel organoids study and performed the experiments. S.H. and H.Y. designed and developed Squidiff. S.H., Y.Z., D.T., H.Y., Z.Z., C.X., S.C. analyzed and interpreted data. G. G

performed irradiation experiments. R.T., S.Q. and G.V.N. provided additional supervision. S.H., Y.Z., D.T., H.Y., J.Z., E.A. and K.L. wrote the paper. All authors reviewed, contributed to and approved the paper.

**Declaration of interests.** The authors do not have any personal or financial relationships that would constitute conflict of interest.

**Data availability.** The raw data discussed in this manuscript will be deposited in the National Center for Biotechnology Information's Gene Expression Omnibus.

**Code availability.** The Squidiff package and code to reproduce the results in this manuscript is available on the GitHub repositories (https://github.com/siyuh/squidiff_public).

# References


[1] Schneider, G., Schmidt-Supprian, M., Rad, R., Saur, D.: Tissue-specific tumorigenesis: context matters. Nature Reviews Cancer **17**(4), 239–253 (2017)

[2] Potente, M., Mäkinen, T.: Vascular heterogeneity and specialization in development and disease. Nature Reviews Molecular Cell Biology **18**(8), 477–494 (2017)

[3] Rafelski, S.M., Theriot, J.A.: Establishing a conceptual framework for holistic cell states and state transitions. Cell **187**(11), 2633–2651 (2024)

[4] Gullapalli, R.R., Desai, K.V., Santana-Santos, L., Kant, J.A., Becich, M.J.: Next generation sequencing in clinical medicine: Challenges and lessons for pathology and biomedical informatics. Journal of pathology informatics **3**(1), 40 (2012)

[5] Atanasov, A.G., Zotchev, S.B., Dirsch, V.M., Supuran, C.T.: Natural products in drug discovery: advances and opportunities. Nature reviews Drug discovery **20**(3), 200–216 (2021)

[6] Lotfollahi, M., Wolf, F.A., Theis, F.J.: scgen predicts single-cell perturbation responses. Nature methods **16**(8), 715–721 (2019)

[7] Bunne, C., Stark, S.G., Gut, G., Del Castillo, J.S., Levesque, M., Lehmann, K.-V., Pelkmans, L., Krause, A., Rätsch, G.: Learning single-cell perturbation responses using neural optimal transport. Nature methods **20**(11), 1759–1768 (2023)

[8] Roohani, Y., Huang, K., Leskovec, J.: Predicting transcriptional outcomes of novel multigene perturbations with gears. Nature Biotechnology **42**(6), 927–935 (2024)

[9] Chen, Y.T., Zou, J.: Genepert: Leveraging genept embeddings for gene perturbation prediction. bioRxiv, 2024–10 (2024)

[10] Kingma, D.P.: Auto-encoding variational bayes. arXiv preprint arXiv:1312.6114 (2013)

[11] Yang, L., Zhang, Z., Song, Y., Hong, S., Xu, R., Zhao, Y., Zhang, W., Cui, B., Yang, M.-H.: Diffusion models: A comprehensive survey of methods and applications. ACM Computing Surveys **56**(4), 1–39 (2023)

[12] Ho, J., Jain, A., Abbeel, P.: Denoising diffusion probabilistic models. Advances in neural information processing systems **33**, 6840–6851 (2020)

[13] Guo, Z., Liu, J., Wang, Y., Chen, M., Wang, D., Xu, D., Cheng, J.: Diffusion models in bioinformatics and computational biology. Nature reviews bioengineering **2**(2), 136–154 (2024)

[14] Preechakul, K., Chatthee, N., Wizadwongsa, S., Suwajanakorn, S.: Diffusion autoencoders: Toward a meaningful and decodable representation. In: Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, pp. 10619–10629 (2022)

[15] Bunne, C., Roohani, Y., Rosen, Y., Gupta, A., Zhang, X., Roed, M., Alexandrov, T., AlQuraishi, M., Brennan, P., Burkhardt, D.B., et al.: How to build the virtual cell with artificial intelligence: Priorities and opportunities. arXiv preprint arXiv:2409.11654 (2024)

[16] Song, J., Meng, C., Ermon, S.: Denoising diffusion implicit models. arXiv preprint arXiv:2010.02502 (2020)

[17] Yahaya, B.H.: Organoid Technology for Disease Modelling and Personalized Treatment. Springer, ??? (2022)

[18] Nikolova, M.T., He, Z., Wimmer, R.A., Seimiya, M., Nikoloff, J.M., Penninger, J.M., Camp, J.G., Treutlein, B.: Fate and state transitions during human blood vessel organoid development. bioRxiv, 2022–03 (2022)

[19] Zappia, L., Phipson, B., Oshlack, A.: Splatter: simulation of single-cell rna sequencing data. Genome biology **18**(1), 174 (2017)

[20] Cuomo, A.S., Seaton, D.D., McCarthy, D.J., Martinez, I., Bonder, M.J., Garcia-Bernardo, J., Amatya, S., Madrigal, P., Isaacson, A., Buettner, F., *et al.*: Single-cell rna-sequencing of differentiating ips cells reveals dynamic genetic effects on gene expression. Nature communications **11**(1), 810 (2020)

[21] Chambers, I., Silva, J., Colby, D., Nichols, J., Nijmeijer, B., Robertson, M., Vrana, J., Jones, K., Grotewold, L., Smith, A.: Nanog safeguards pluripotency and mediates germline development. Nature **450**(7173), 1230–1234 (2007)

[22] Schrode, N., Saiz, N., Di Talia, S., Hadjantonakis, A.-K.: Gata6 levels modulate primitive endoderm cell fate choice and timing in the mouse blastocyst. Developmental cell **29**(4), 454–467 (2014)

[23] Wilson, V., Beddington, R.: Expression of t protein in the primitive streak is necessary and sufficient for posterior mesoderm movement and somite differentiation. Developmental biology **192**(1), 45–58 (1997)

[24] Wolf, F.A., Hamey, F.K., Plass, M., Solana, J., Dahlin, J.S., Göttgens, B., Rajewsky, N., Simon, L., Theis, F.J.: Paga: graph abstraction reconciles clustering with trajectory inference through a topology preserving map of single cells. Genome biology **20**, 1–9 (2019)

[25] Cannoodt, R., Saelens, W., Sichien, D., Tavernier, S., Janssens, S., Guilliams, M., Lambrecht, B., Preter, K.D., Saeys, Y.: Scorpius improves trajectory inference and identifies novel modules in dendritic cell development. biorxiv, 079509 (2016)

[26] Qiu, X., Mao, Q., Tang, Y., Wang, L., Chawla, R., Pliner, H.A., Trapnell, C.: Reversed graph embedding resolves complex single-cell trajectories. Nature methods **14**(10), 979–982 (2017)

[27] Zhao, W., Dovas, A., Spinazzi, E.F., Levitin, H.M., Banu, M.A., Upadhyayula, P., Sudhakar, T., Marie, T., Otten, M.L., Sisti, M.B., *et al.*: Deconvolution of cell type-specific drug responses in human tumor tissue with single-cell rna-seq. Genome medicine **13**(1), 82 (2021)

[28] Hofer, M., Lutolf, M.P.: Engineering organoids. Nature Reviews Materials **6**(5), 402–420 (2021)

[29] Yang, S., Hu, H., Kung, H., Zou, R., Dai, Y., Hu, Y., Wang, T., Lv, T., Yu, J., Li, F.: Organoids: The current status and biomedical applications. MedComm **4**(3), 274 (2023)

[30] Huang, Y., Liu, Z., Li, M., Wang, D., Ye, J., Hu, Q., Zhang, Q., Lin, Y., Chen, R., Liang, X., *et al.*: Deciphering the impact of aging on splenic endothelial cell heterogeneity and immunosenescence through single-cell rna sequencing analysis. Immunity & Ageing **21**(1), 48 (2024)

[31] Mettler Jr, F.A., Voelz, G.L.: Major radiation exposure—what to expect and how to respond. New England Journal of Medicine **346**(20), 1554–1561 (2002)

[32] Kameni, L.E., Januszyk, M., Berry, C.E., Downer Jr, M.A., Parker, J.B., Morgan, A.G., Valencia, C., Griffin, M., Li, D.J., Liang, N.E., *et al.*: A review of radiation-induced vascular injury and clinical impact. Annals of Plastic Surgery **92**(2), 181–185 (2024)

[33] Durante, M.: New challenges in high-energy particle radiobiology. The British journal of radiology **87**(1035), 20130626 (2014)

[34] Tavakol, D.N., Nash, T.R., Kim, Y., He, S., Fleischer, S., Graney, P.L., Brown, J.A., Liberman, M., Tamargo, M., Harken, A., *et al.*: Modeling and countering

the effects of cosmic radiation using bioengineered human tissues. Biomaterials **301**, 122267 (2023)

[35] Chancellor, J.C., Scott, G.B., Sutton, J.P.: Space radiation: the number one risk to astronaut health beyond low earth orbit. Life **4**(3), 491–510 (2014)

[36] Wijerathne, H., Langston, J.C., Yang, Q., Sun, S., Miyamoto, C., Kilpatrick, L.E., Kiani, M.F.: Mechanisms of radiation-induced endothelium damage: Emerging models and technologies. Radiotherapy and Oncology **158**, 21–32 (2021)

[37] Oliner, J.D., Saiki, A.Y., Caenepeel, S.: The role of mdm2 amplification and overexpression in tumorigenesis. Cold Spring Harbor perspectives in medicine **6**(6), 026336 (2016)

[38] Ruef, C., Coleman, D.L.: Granulocyte-macrophage colony-stimulating factor: pleiotropic cytokine with potential clinical usefulness. Reviews of infectious diseases **12**(1), 41–62 (1990)

[39] Ping, S., Qiu, X., Gonzalez-Toledo, M.E., Liu, X., Zhao, L.-R.: Stem cell factor in combination with granulocyte colony-stimulating factor reduces cerebral capillary thrombosis in a mouse model of cadasil. Cell Transplantation **27**(4), 637–647 (2018)

[40] Furmento, V.A., Marino, J., Blank, V.C., Roguin, L.P.: The granulocyte colony-stimulating factor (g-csf) upregulates metalloproteinase-2 and vegf through pi3k/akt and erk1/2 activation in human trophoblast swan 71 cells. Placenta **35**(11), 937–946 (2014)

[41] Hughson, R.L., Helm, A., Durante, M.: Heart in space: effect of the extraterrestrial environment on the cardiovascular system. Nature Reviews Cardiology **15**(3), 167–180 (2018)

[42] Delp, M.D., Charvat, J.M., Limoli, C.L., Globus, R.K., Ghosh, P.: Apollo lunar astronauts show higher cardiovascular disease mortality: possible deep space radiation effects on the vascular endothelium. Scientific reports **6**(1), 29901 (2016)

[43] Barcellos-Hoff, M.H., Blakely, E.A., Burma, S., Fornace Jr, A.J., Gerson, S., Hlatky, L., Kirsch, D.G., Luderer, U., Shay, J., Wang, Y., *et al.*: Concepts and challenges in cancer risk prediction for the space radiation environment. Life Sciences in Space Research **6**, 92–103 (2015)

[44] Ho, J., Salimans, T., Gritsenko, A., Chan, W., Norouzi, M., Fleet, D.J.: Video diffusion models. Advances in Neural Information Processing Systems **35**, 8633–8646 (2022)

[45] Wu, K.E., Yang, K.K., Berg, R., Alamdari, S., Zou, J.Y., Lu, A.X., Amini, A.P.:

Protein structure generation via folding diffusion. Nature communications **15**(1), 1059 (2024)

[46] Werschler, N., Penninger, J.: Generation of human blood vessel organoids from pluripotent stem cells. JoVE (Journal of Visualized Experiments) (191), 64715 (2023)

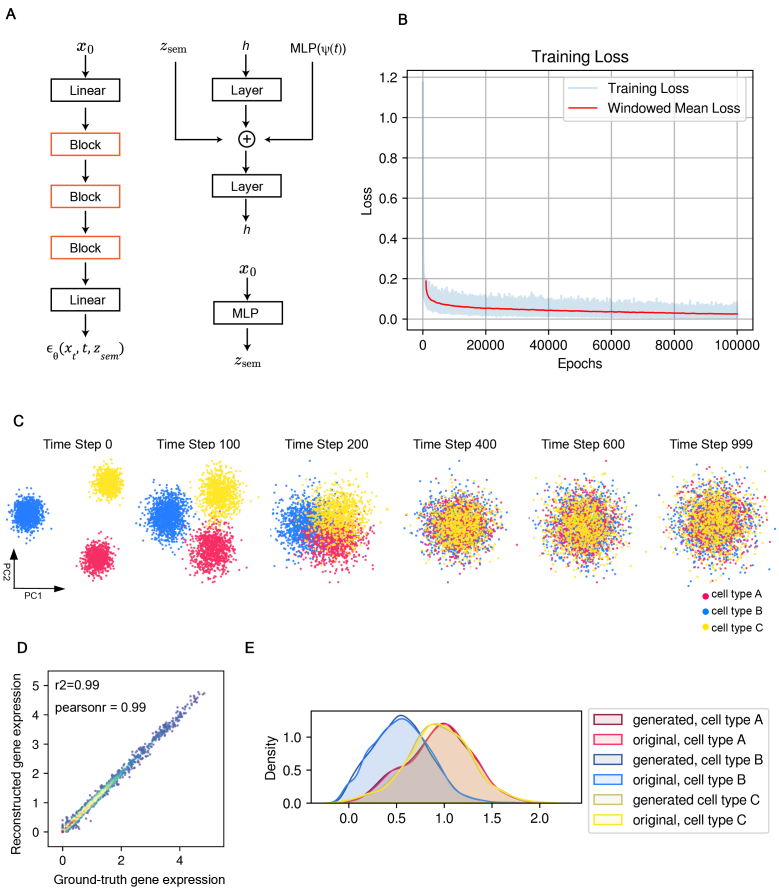
[47] Tomer, R., Lovett-Barron, M., Kauvar, I., Andalman, A., Burns, V.M., Sankaran, S., Grosenick, L., Broxton, M., Yang, S., Deisseroth, K.: Sped light sheet microscopy: fast mapping of biological system structure and function. Cell **163**(7), 1796–1806 (2015)

[48] Xu, Y., Randers-Pehrson, G., Turner, H.C., Marino, S.A., Geard, C.R., Brenner, D.J., Garty, G.: Accelerator-based biological irradiation facility simulating neutron exposure from an improvised nuclear device. Radiation research **184**(4), 404–410 (2015)

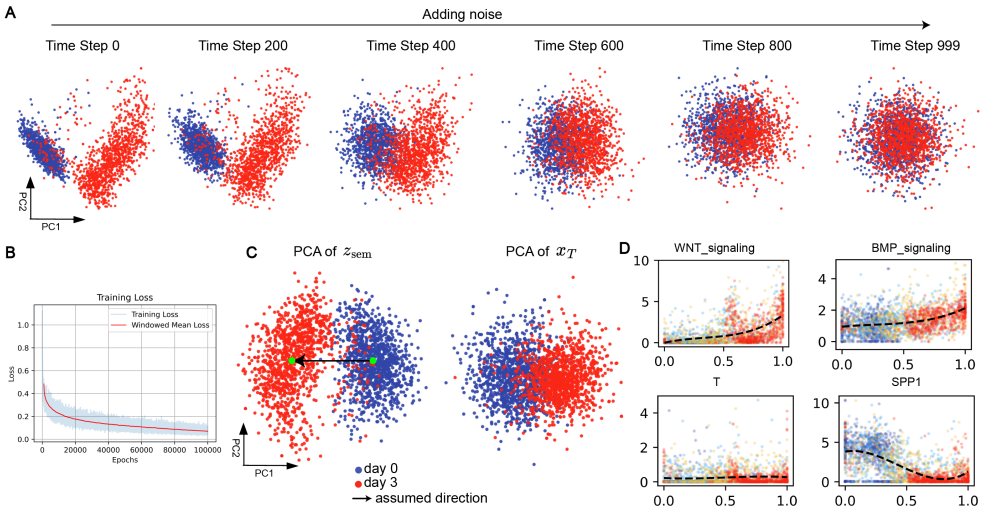
[49] Tavakol, D.N., Nash, T.R., Kim, Y., Graney, P.L., Liberman, M., Fleischer, S., Lock, R.I., O'Donnell, A., Andrews, L., Ning, D., et al.: Modeling the effects of protracted cosmic radiation in a human organ-on-chip platform. Advanced Science, 2401415 (2024)

[50] Xu, Y., Randers-Pehrson, G., Marino, S.A., Garty, G., Harken, A., Brenner, D.J.: Broad energy range neutron spectroscopy using a liquid scintillator and a proportional counter: Application to a neutron spectrum similar to that from an improvised nuclear device. Nuclear Instruments and Methods in Physics Research Section A: Accelerators, Spectrometers, Detectors and Associated Equipment **794**, 234–239 (2015)

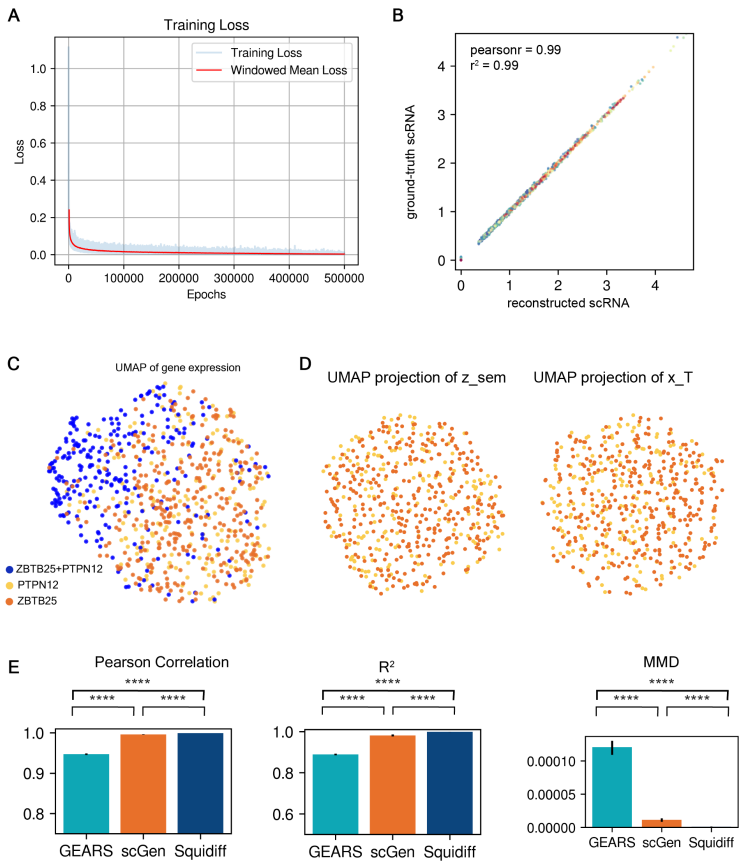
[51] Holden, S., Perez, R., Hall, R., Fallgren, C.M., Ponnaiya, B., Garty, G., Brenner, D.J., Weil, M.M., Raber, J.: Effects of acute and chronic exposure to a mixed field of neutrons and photons and single or fractionated simulated galactic cosmic ray exposure on behavioral and cognitive performance in mice. Radiation research **196**(1), 31–39 (2021)



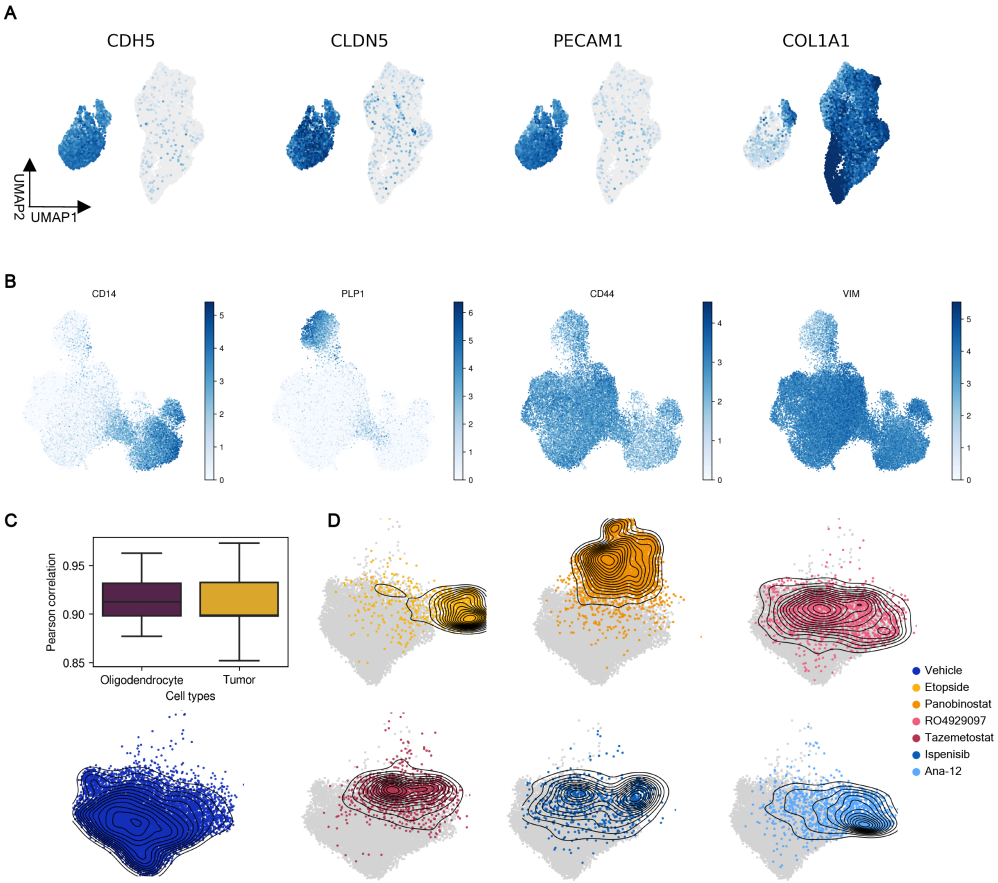

Supplementary Fig. 1: **Simulation of diffusion processes and training performance**. **A**, Schematic of a residual connection architecture for learning noise and the semantic encoder in the Squidiff model. The model processes initial inputs ($x_0$) through linear layers and block modules to generate the noise.Each block consists of residual connection structures that help stabilize and refine the learning process. Semantic latent variables ($z_{sem}$) are integrated with a multilayer perceptron (MLP) conditioned on time $t$. The final output combines these features through additional layers to reconstruct the gene expression state. The semantic encoder is an MLP that encodes $x_0$ to $z_{sem}$. **B**, Training loss curve showing the training loss (blue) and windowed mean loss (red) over 100,000 epochs. The decreasing trend in loss indicates successful model convergence during training. **C**,Visualization of the diffusion process at various time steps (0, 100, 200, 300, 400, and 499). The diffusion model gradually introduces noise, transforming the distinct clusters (red, blue, yellow) into a more dispersed distribution, demonstrating the model's capability to learn and reverse complex data distributions. **D**, Correlation between reconstructed and ground-truth gene expression. **E**, Density plots of original and generated cell types. The density curves compare original cell types (solid lines) with generated cell types (dashed lines), showing alignment across types A, B, and C, indicating Squidiff's ability to preserve cell-type-specific gene expression patterns.



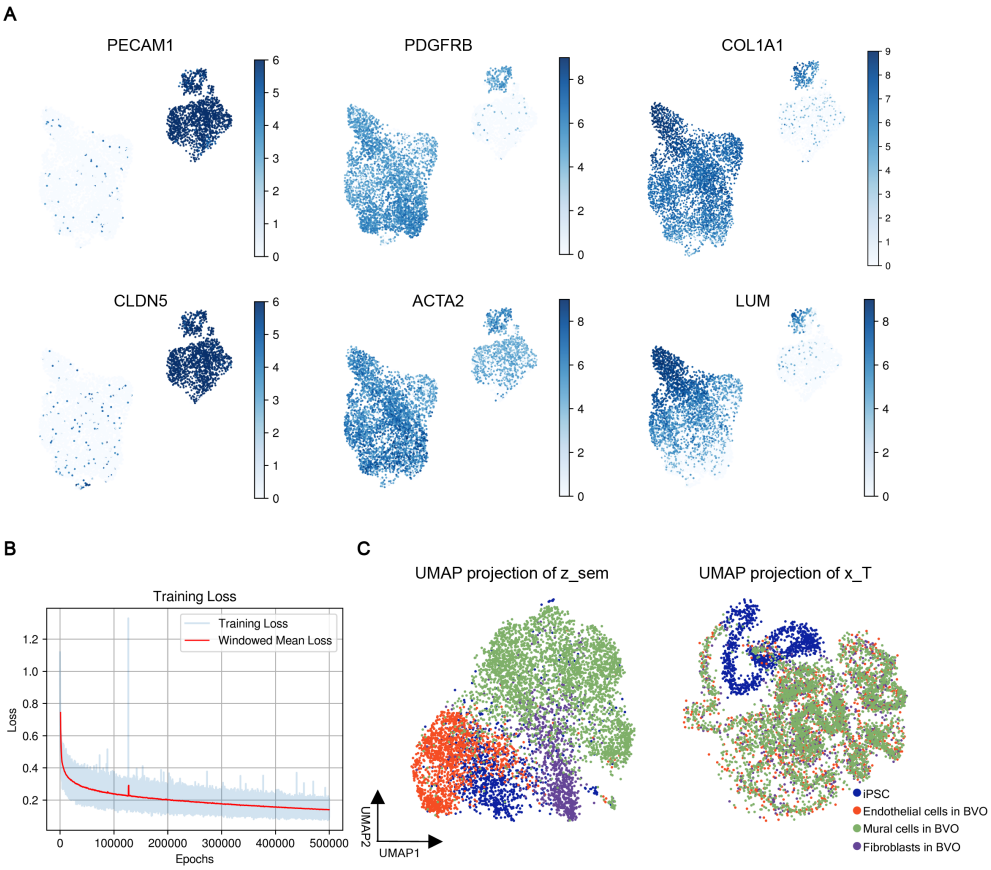

Supplementary Fig. 2: **Squidiff predicts cell differentiation from iPSCs to mesendoderm and definitive endoderm cells**. **A**, Visualization of the diffusion process at various time steps (0, 200, 400, 600, and 800). The model gradually introduces noise, transforming the distinct clusters (red, blue, yellow) into more dispersed distributions, demonstrating the model's ability to learn and reverse complex data distributions during differentiation. **B**, Training loss over 500,000 epochs, showing the convergence of the model. The loss decreases significantly in the initial stages and stabilizes, indicating effective training and convergence of the diffusion model. **C**, UMAP projection of $z_{sem}$ (left) and $x_T$(right) from single-cell RNA-seq data of iPSCs differentiating to mesendoderm and definitive endoderm on day 0 (red) and day 3 (blue). Arrows indicate the assumed direction of differentiation. **D**, Gene expression trends for key signaling pathways (WNT, T, BMP, SPP1) over time, visualized as scatter plots with corresponding gene expression levels.



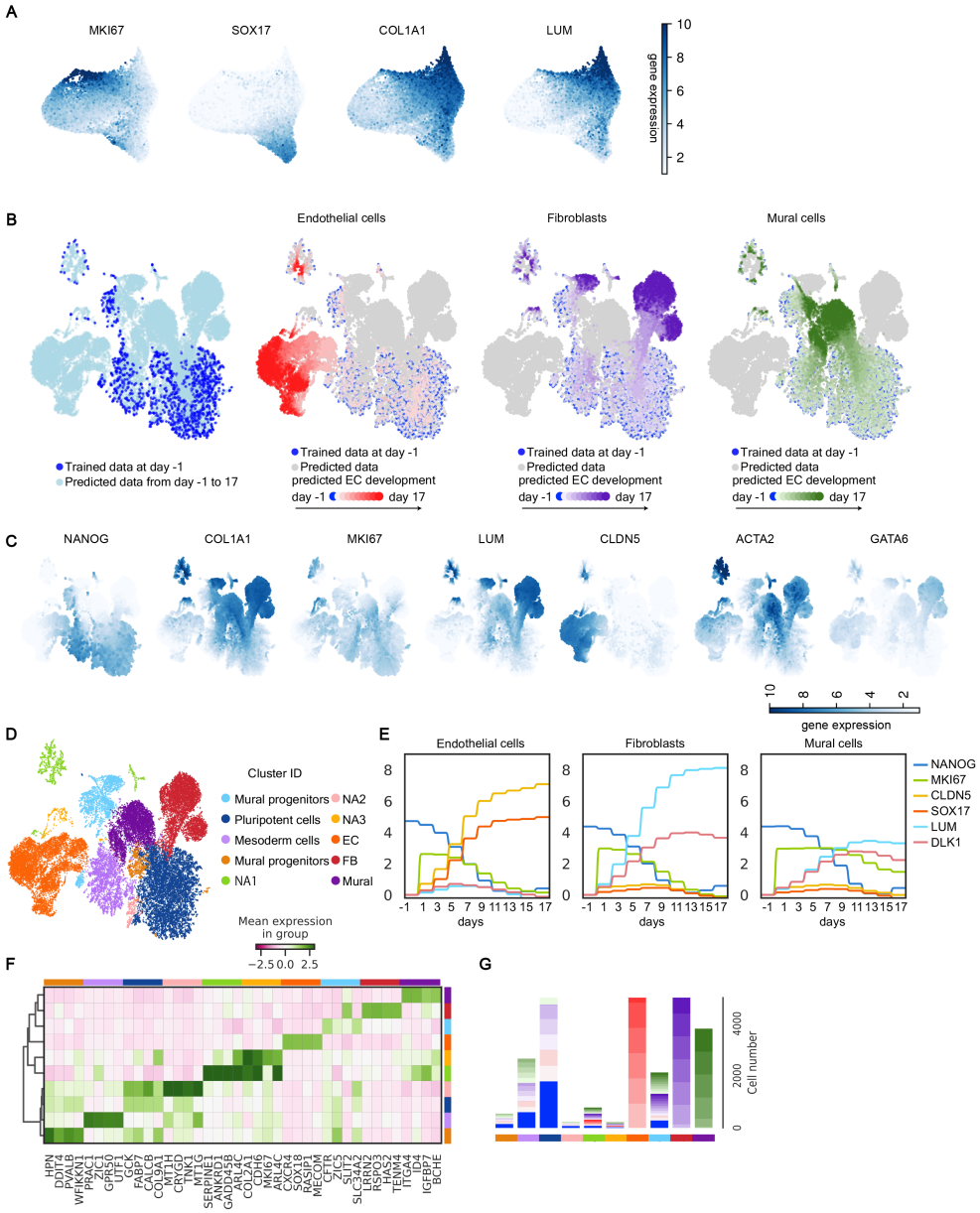

Supplementary Fig. 3: **Squidiff predicts gene perturbation**. **A**, Training loss over 500,000 epochs, showing the convergence of the model. The loss decreases significantly in the initial stages and stabilizes, indicating effective training and convergence of the diffusion model. **B**, Scatter plot comparing reconstructed single-cell RNA-seq data with ground-truth data, illustrating the high accuracy of Squidiff in reconstructing gene expression data. The Pearson correlation coefficient and R-squared values are both 0.99, indicating a strong correlation between the reconstructed and ground-truth values. **C**, UMAP projection of gene expression data for single-cell RNA-seq data perturbed by *ZBTB25* (orange), *PTPN12* (yellow), and the combination of *ZBTB25+PTPN12* (blue). The visualization demonstrates Squidiff's ability to predict non-additive genetic effects accurately. **D**, UMAP projections of semantic latent variables ($z_{sem}$) and stochastic latent variables ($x_T$) for the perturbed single-cell RNA-seq data. The projections show distinct clusters for each perturbation, indicating the model's capability to capture and represent the underlying genetic perturbations. **E**, Bar plots comparing the prediction of ZBTB25 on control single-cell data with 203 genes by Squidiff, scGen, and GEARS in terms of Pearson correlation, $R^2$ values, and Maximum Mean Discrepancy (MMD). Two-sided independent two-sample t-test was performed. $P = 2.49 \times 10^{-24}$, $1.56 \times 10^{-26}$, $7.46 \times 10^{-09}$ for pearson correlation, $= 4.15 \times 10^{-23}$, $1.01 \times 10^{-26}$, $2.94 \times 10^{-11}$ for $R^2$ correlation, $= 1.04 \times 10^{-21}$, $5.64 \times 10^{-22}$, $1.87 \times 10^{-4}$ for maximum Mean Discrepancy. $^{****}P < 0.0001$. NS, not significant.



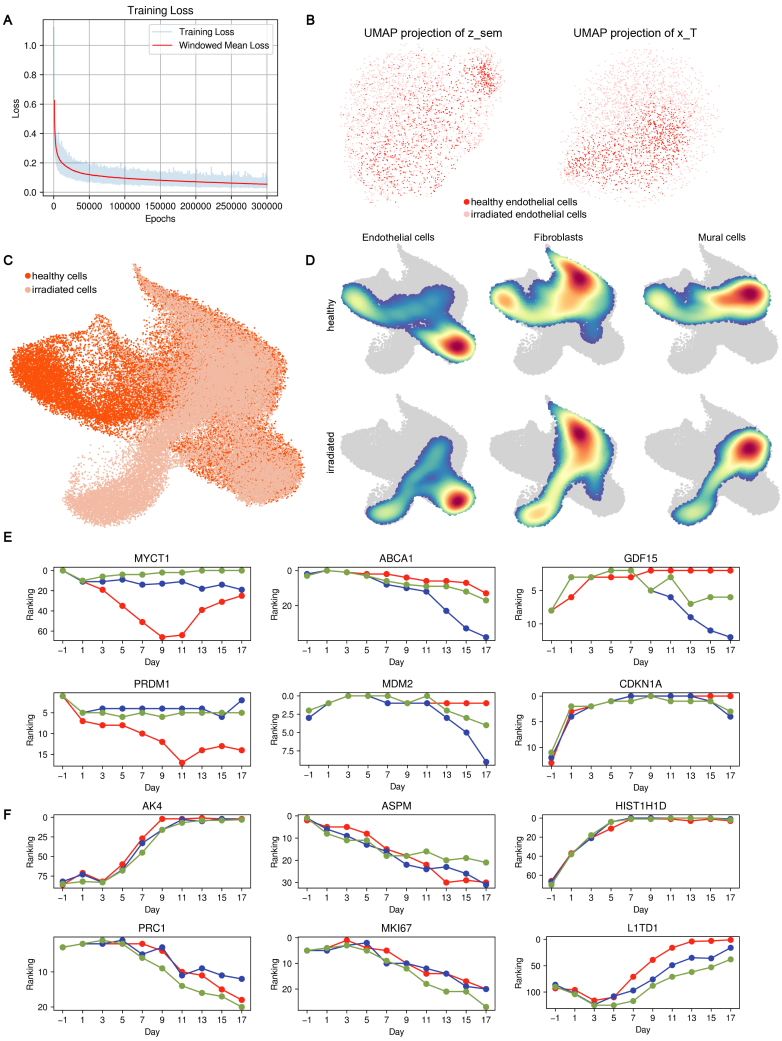

Supplementary Fig. 4: **Squidiff predicts drug perturbation effects on tumor cells and oligodendrocytes**. **A**, UMAP visualization of single-cell RNA-seq data showing the expression of marker genes (*CDH5, CLDN5, PECAM1*, and *COL1A1*) in oligodendrocytes and tumor cells under different conditions. The gradient color indicates the expression level of each marker gene. **B**, UMAP visualization of single-cell RNA-seq data showing the expression of additional marker genes (*CD14*, *PLP1*, *CD44*, and *VIM*) in oligodendrocytes and tumor cells. The gradient color indicates the expression level of each marker gene. **C**, Box plot comparing the Pearson correlation coefficients between reconstructed and ground-truth scRNA-seq data for oligodendrocytes and tumor cells. The high correlation indicates the accuracy of Squidiff in predicting gene expression changes upon drug treatment. **D**, Contour plots on UMAP visualizations highlighting the density distribution of oligodendrocytes and tumor cells under different drug treatments. Each contour plot corresponds to a different cell type or treatment condition, demonstrating the distinct effects of drug perturbations as predicted by Squidiff. The colors of contour plot represent drugs.




Supplementary Fig. 5: **Squidiff predicts cell differentiation in blood vessel organoids (BVOs)**. **A**, UMAP visualization of single-cell RNA-seq data showing the expression of marker genes (PECAM1, PDGFRB, COL1A1, CLDN5, ACTA2, LUM) in various cell types within BVOs. The gradient color indicates the expression level of each marker gene, highlighting the differentiation states of endothelial cells, mural cells, and fibroblasts. **B**, Training loss over 500,000 epochs, showing the convergence of the model. The loss decreases significantly in the initial stages and stabilizes, indicating effective training and convergence of the diffusion model. **C**, UMAP projections of semantic latent variables ($z_{sem}$) and stochastic latent variables ($x_T$) for single-cell RNA-seq data. The projections show distinct clusters for iPSCs, endothelial cells, mural cells, and fibroblasts in BVOs, indicating the model's capability to capture and represent the differentiation pathways.

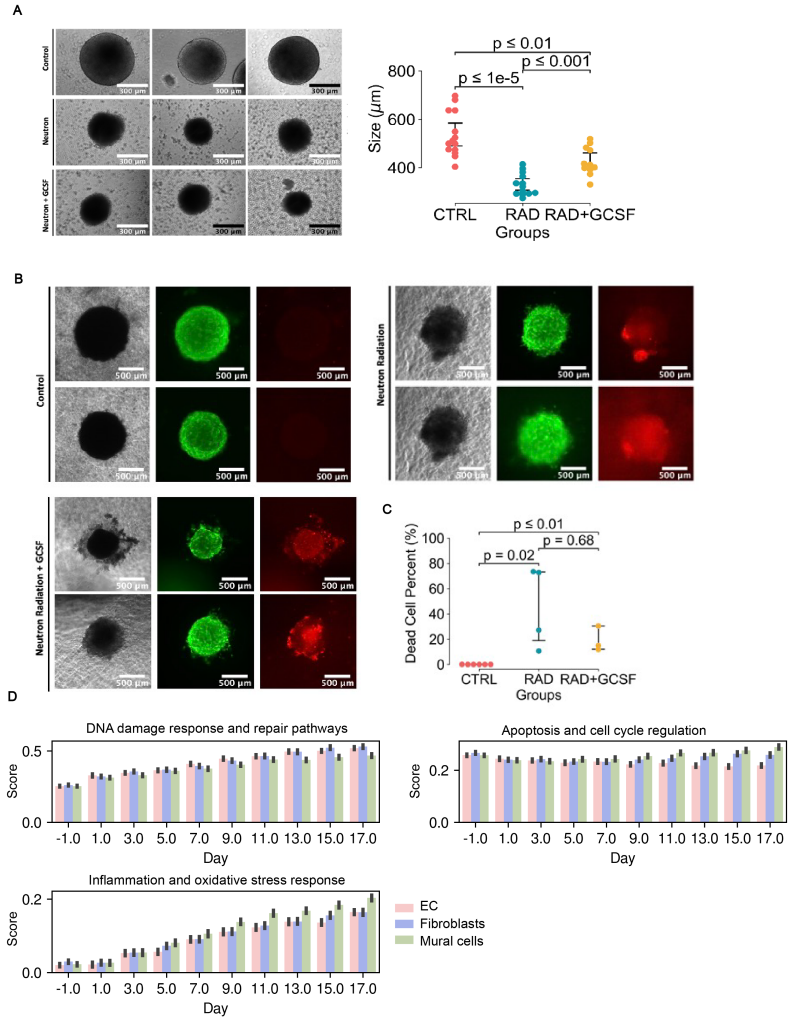



Supplementary Fig. 6: **Squidiff predicts cell differentiation processes in blood vessel organoids. A**, Joint UMAPs of trained scRNA-seq data from day -1 and predicted data from days -1 to 17, highlighting marker expression, including pluripotency genes (MKI67), endothelial markers (SOX17), fibroblast and mural-specific markers (COL1A1, LUM). **B**, UMAP visualization of single-cell RNA-seq data predicted by scGen showing the differentiation trajectories of endothelial cells, fibroblasts, and mural cells within blood vessel organoids (BVOs) from day -1 to day 17. The blue dots represent the trained data at day -1, while the grey dots represent the predicted data from day -1 to day 17. **C**, Gene expression levels of key markers (*NANOG, COL1A1, MKI67, LUM, CLDN5, ACTA2, GATA6*) across different developmental stages in BVOs, as predicted by scGen. The gradient color indicates the expression level of each marker gene. **D**, Joint UMAPs of trained scRNA-seq data from day -1 and predicted data from days -1 to 17, color-coded by cell types. **E**, Stepwise expression patterns of markers across days in fibroblasts, endothelial cells, and mural cells predicted by scGen. **F**, Heatmap showing differential gene expression analysis for clustering across joint trained and predicted scRNA-seq data. **G**, Stacked bar plot showing the number of cells by major cell types across different days within cell type clusters.

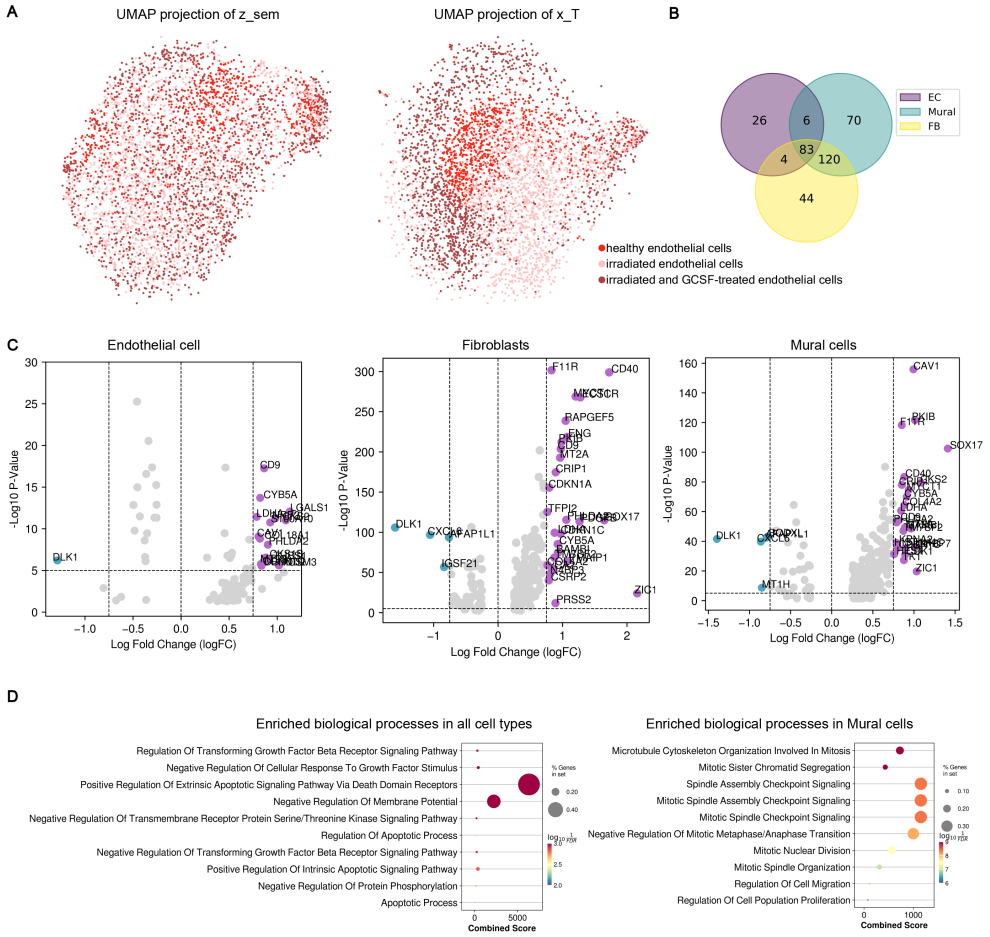



Supplementary Fig. 7: **Squidiff predicts the effects of neutron irradiation on blood vessel organoids. A**, Training loss over 300,000 epochs, showing the convergence of the model. The loss decreases significantly in the initial stages and stabilizes, indicating effective training and convergence of the diffusion model. **B**, UMAP projections of semantic latent variables ($z_{sem}$) and stochastic latent variables ($x_T$) for single-cell RNA-seq data, showing healthy endothelial cells (red) and irradiated endothelial cells (light red). The distinct clustering indicates the model's ability to capture the effects of irradiation. **C**, UMAP visualization comparing healthy cells (red) and irradiated cells (light red) within the blood vessel organoids, indicating the global transcriptomic changes due to irradiation. **D**, Density plots on UMAP visualizations highlighting the distribution of endothelial cells, fibroblasts, and mural cells under healthy and irradiated conditions. Each density plot demonstrates the distinct distribution of cell types as predicted by Squidiff. **E**, Temporal gene expression trends for key genes (*MYCT1, ABCA1, GDF15, PRDM1, MDM2, CDKN1A*) over the developmental timeline from day -1 to day 17, comparing healthy and irradiated conditions. The rankings indicate the changes in gene expression levels due to irradiation. **F**, Additional temporal gene expression trends for key genes (*AK4, ASPM, HIST1H1D, PRC1, MKI67, L1TD1*) over the same timeline, further illustrating the impact of irradiation on the molecular landscape of the cells.

Supplementary Fig. 8: **Gene ontology (GO) and pathway enrichment analysis of differentially expressed genes in blood vessel organoids (BVOs) following neutron irradiation. A**, GO Biological Process enrichment analysis for endothelial cells, fibroblasts, and mural cells. The dot plots show the top enriched GO terms for each cell type. The size of the dots represents the percentage of genes in each term, and the color indicates the log-transformed adjusted p-value. The combined score represents the enrichment score of the GO term. **B**, Hallmark pathway enrichment analysis for endothelial cells, fibroblasts, and mural cells. The dot plots show the top enriched hallmark pathways for each cell type. The size of the dots represents the percentage of genes in each term, and the color indicates the log-transformed adjusted p-value. The combined score represents the enrichment score of the pathway.




Supplementary Fig. 9: **Effects of neutron irradiation and GCSF treatment on blood vessel organoids (BVOs). A**, Representative images and size measurements of BVOs under control, neutron irradiation, and neutron irradiation with GCSF treatment conditions. The scatter plot shows the size of BVOs in each group, with statistical significance indicated (p-values). **B**, Live/dead cell assay images of BVOs under control, neutron irradiation, and neutron irradiation with GCSF treatment conditions. Green indicates live cells, and red indicates dead cells. **C**, Quantification of dead cell percentages in BVOs under control, neutron irradiation, and neutron irradiation with GCSF treatment conditions. The scatter plot shows the percentage of dead cells in each group, with statistical significance indicated (p-values). **D**, Bar plots showing the scores of pathway associated to irradiation exposure over the developmental timeline (day -1 to day 17) in BVOs across cell types.




Supplementary Fig. 10: **Effects of neutron irradiation and GCSF treatment on blood vessel organoids (BVOs). A**, UMAP projections of semantic latent variables ($z_{sem}$, left) and stochastic variables ($x_T$, right) for endothelial cells. Colors indicate healthy (red), irradiated (dark red), and irradiated G-CSF-treated (light red) endothelial cells. **B**, Venn diagram showing overlap of differentially expressed (DE) genes among endothelial cells (EC, purple), fibroblasts (FB, yellow), and mural cells (green), highlighting shared and cell-type-specific responses. **C**, Volcano plots of DE genes in endothelial cells, fibroblasts, and mural cells. Log2 fold change (x-axis) versus -log10 p-value (y-axis). Purple points denote upregulated genes, and blue points indicate downregulated genes post-G-CSF treatment. **D**, Enriched biological processes across all cell types (left) and mural cells (right). Bubble size represents the percentage of genes, and color indicates statistical significance (log10 FDR). Key pathways include TGF-$\beta$ signaling, apoptosis, and mitotic processes.