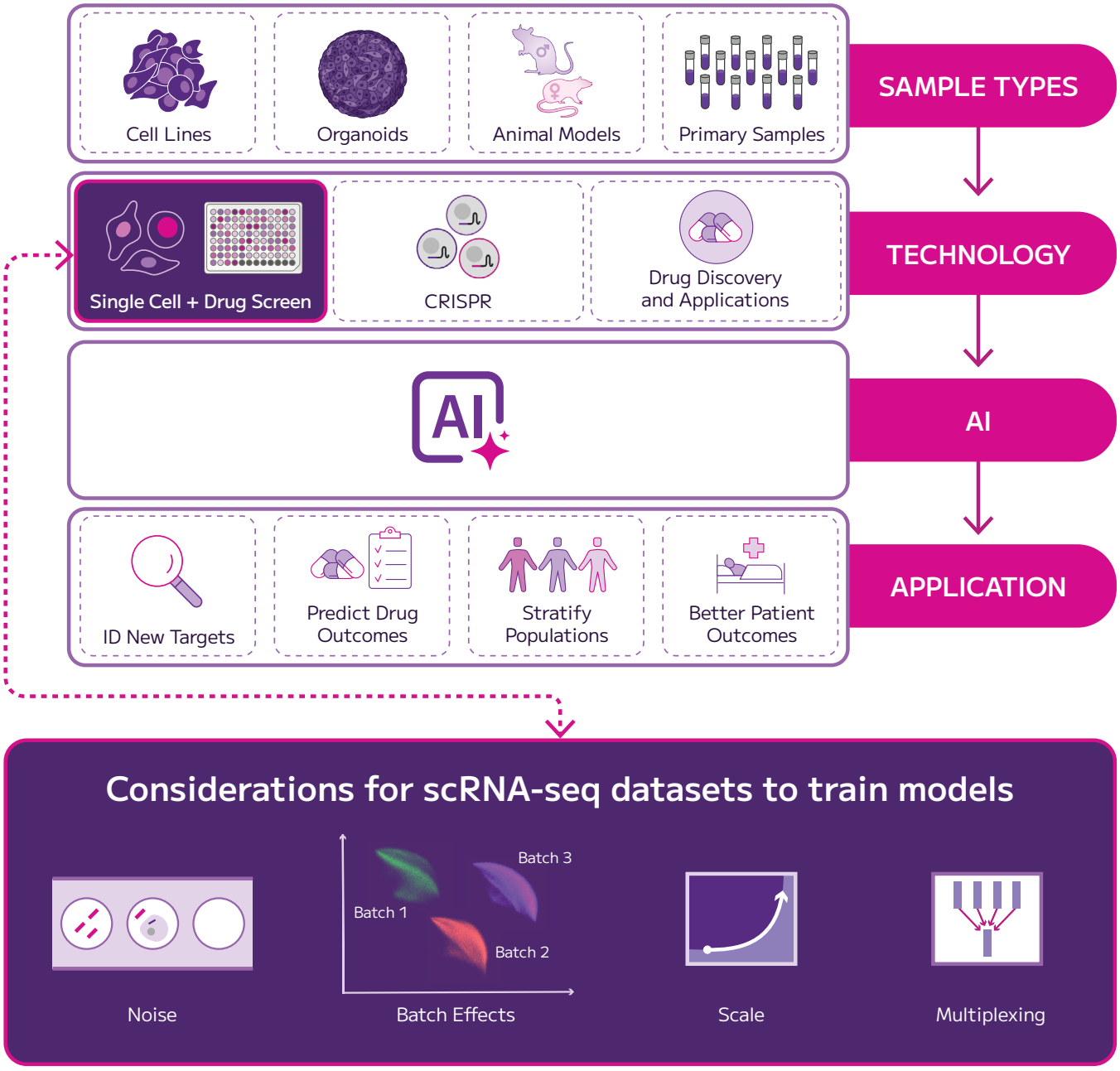
PARSE BIOSCIENCES

Single Cell Datasets for Training AI Models

Single Cell Datasets for Training AI Models

Single cell RNA sequencing (scRNA-seq) reveals cellular diversity at unprecedented resolution.

Artificial Intelligence (AI) models transform these complex data into biological insights, simulate perturbations, and generate synthetic cells to accelerate discovery.



What scRNA-seq and AI can do Together

ID NEW TARGETS

scRNA-seq captures the transcriptomes of thousands to hundreds of millions of individual cells. This reveals subtle subpopulations, including rare cell types, transient developmental intermediates, or malignant subclones.

AI foundation models integrate multi-omic data to map molecular networks and cell-cell interactions, and to predict how perturbations reshape cellular states beyond the original datasets they were trained on.

In 2024, Tahoe Biosciences released Tahoe-100M, the largest single-cell dataset to date: 100 million cells from 50 cancer lines exposed to 1,100 small-molecule perturbations (Figure 1). Parse Biosciences scientists at the Parse Gigalab fixed and barcoded the sample, and generated the count matrices, creating an open resource for AI models to study gene regulation and cellular dynamics.

Only a few months later, scientists developed STATE, a transformer model trained on Tahoe-100M plus the Parse PBMC 10M dataset. It captured subtle cell-cell variations across perturbations and transcriptional responses, and was able to generalize and predict cell perturbation in other models.

Outcome:
By training on large perturbation datasets, AI models can learn to identify and then predict dysregulated pathways, regulatory nodes, and lineage-specific vulnerabilities leading to novel, precise drug targets.

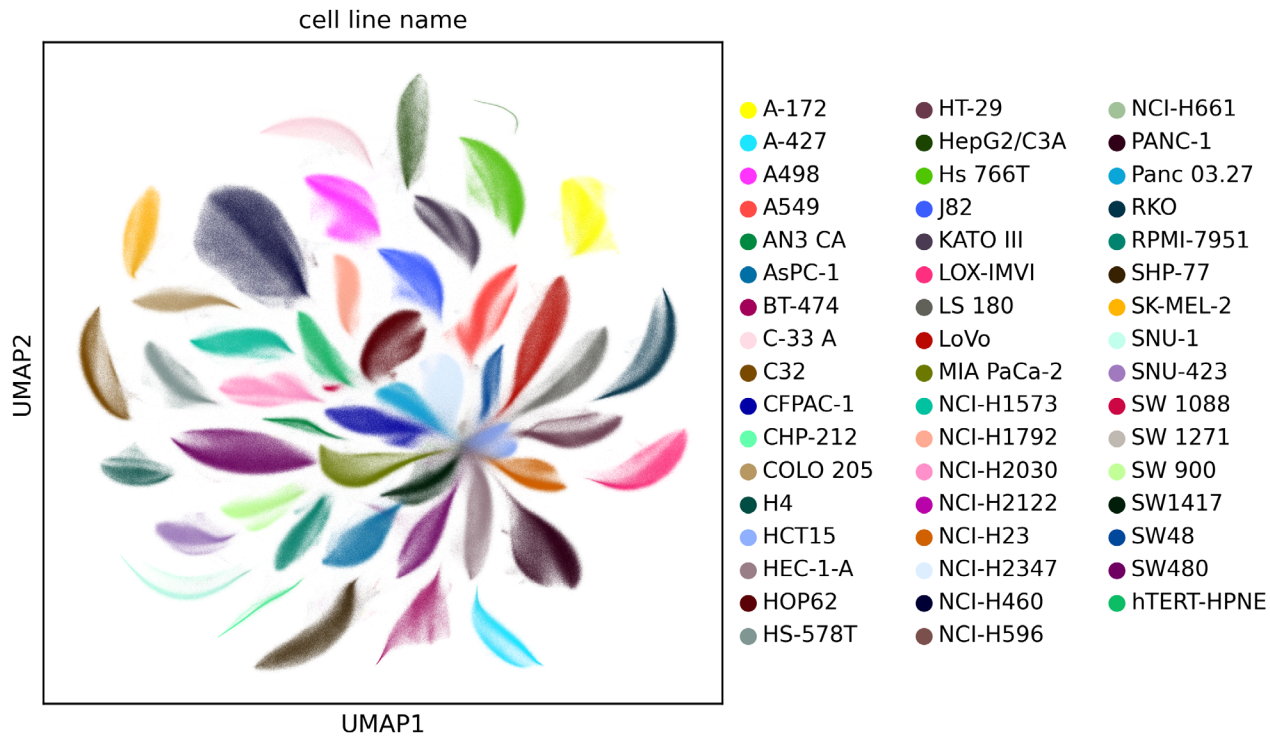


Figure 1. A representative UMAP of a 10M subset showing a clear separation of clusters reflects accurate demultiplexing.

PREDICT DRUG OUTCOMES

Comparing untreated and drug-treated cells with scRNA-seq reveals shifts in gene expression, pathways, and cell identities. AI models can then link drug exposure to responses such as pathway activation or suppression, resistant vs. sensitive subpopulations, and altered cell–cell interactions.

Not all compounds act in predictable ways. Different drugs can modulate the same pathway to varying degrees, and even drugs with unclear mechanisms of action may exert strong and unexpected effects.

For instance, Interferon-alpha (IFN-alpha) response and MHC-I antigen processing pathways are particularly significant for cancer immunotherapy. When asked to identify compounds that upregulate both the IFN-alpha response and MHC-I antigen processing pathways, the Tahoe-100M single-cell perturbation map highlighted 68 drugs that activated both simultaneously, including repurposable FDA approved drugs.

Outcome:
Combining scRNA-seq and AI enables us to predict how unknown compounds will impact diverse cell types, gain mechanistic insights that inform effective drug development, and create a framework that accelerates both discovery and repurposing of therapeutics.

STRATIFY PATIENT POPULATIONS TO UNCOVER DRUG RESPONSE PATTERNS

ScRNA-seq can detect rare resistant subclones and reveal pathways that drive therapy failure. Comparing responders and non-responders identifies biomarkers linked to positive outcomes or predict disease progression.

AI models trained on these integrated datasets can forecast which patients are most likely to respond or worsen, guiding therapy selection and adaptation.

For example, to predict a cancer therapy response, an AI framework like TahoeDive defines "positive" outcomes by pathway activity, builds balanced datasets, and integrates features like cell line profiles, gene status, and drug targets. It generates pathway activity scores, benchmarks performance against biologically grounded models, identifies key features driving response, and delivers drug effectiveness predictions with confidence scores.

Outcome:
A pathway toward personalized medicine in which predictive models not only anticipate therapeutic response and resistance but also guide the dynamic adjustment of treatment strategies.

How scRNA-seq Generates Data to Train AI Models

HIGH-THROUGHPUT SCREENING

scRNA-seq captures how individual cells respond to genetic or chemical perturbations, providing a high-resolution view of cellular dynamics.

A practical example of such a dataset is the 10M Human perturbed PBMCs study. In this experiment, PBMCs from 12 healthy donors were treated with 90 different cytokines for 24 hours, generating 1,092 unique experimental conditions. These were multiplexed, barcoded, and sequenced together in a single large-scale run.

This resource is now being used by scientists at Helmholtz Munich to train new AI models that can learn patterns of cellular change. Such models extend beyond descriptive analysis: they can predict how novel genes or drugs will reshape cellular states, offering a powerful framework to anticipate therapeutic outcomes and disease trajectories.

CRISPR

CRISPR-based approaches offer a major advantage over traditional drug screens: precision. Gene knockouts directly link function to phenotype, while CRISPRi and CRISPRa allow controlled suppression or activation of gene expression. This fine-tuned modulation more closely mirrors how drugs act, which often reduce rather than eliminate gene activity.

When combined with scRNA-seq, CRISPR perturbations can be traced to specific cell types and states at single cell resolution. Importantly, these experiments can be scaled to generate the rich datasets needed to train AI models. Tools such as CPA use deep learning to learn from these perturbations, separating the effects of dosage, time, drugs, and genetic edits. By capturing the complex, non-linear responses of biological systems, CPA can extrapolate to new conditions, predict novel gene–drug interactions, and uncover synthetic lethalities.

DRUG RESPONSE ATLASES

Large-scale drug screens are often analyzed with bulk RNA-seq, which only reveals average population-level responses. By integrating scRNA-seq, researchers can identify resistant subpopulations and uncover their resistance mechanisms.

Transfer learning further bridges bulk and single cell data: models such as autoencoders learn gene relationships from bulk profiles and apply this knowledge to denoise sparse single cell data.

Large sample size is as important as more cells. ScRNA-seq platforms make large-scale screens practical even without millions of cells. A single scientist can profile hundreds or thousands of conditions in parallel within 4 days.

For example, a study tested 88 anti-diabetic drugs across four time points (>350 conditions), collecting and freezing cells for batch analysis (Figure 2). This approach captured both strong and subtle perturbations and revealed distinct transcriptional responses even among drugs with the same mechanism of action.

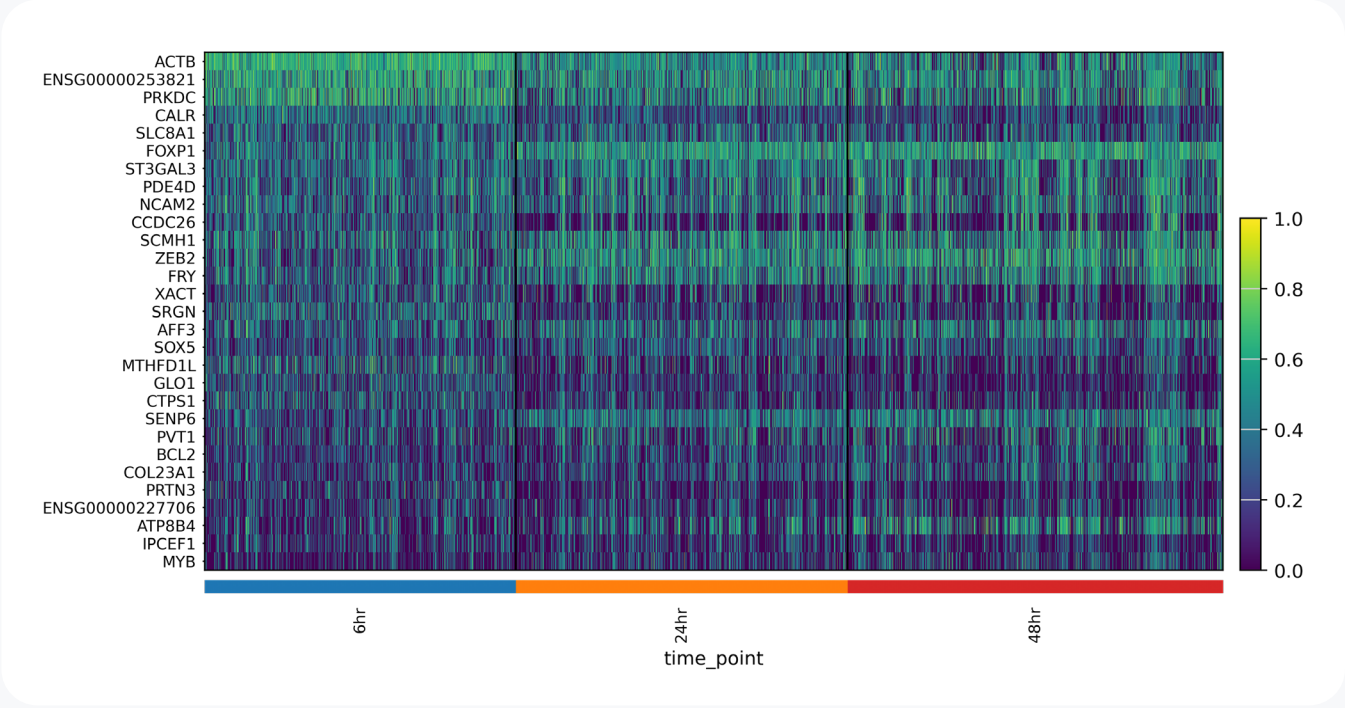


Figure 2: Time-dependent transcriptional response to bromodomain inhibitors in THP-1 cells. Heatmap of top differentially expressed genes across 6, 24, and 48 hr compared to DMSO controls. The major shift occurs between 6 and 24 hr, with expression stabilizing by 48 hr.

COMBINATION DRUG DISCOVERY

Different cell subsets within the same tissue often respond differently to a single drug. Combination drug discovery seeks to improve efficacy, overcome resistance, and reduce toxicity by identifying drug pairs or cocktails that act synergistically. While a single agent may only partially inhibit a pathway or encounter resistant subpopulations, combinations can achieve more comprehensive responses. However, the vast number of possible combinations makes exhaustive testing impractical, and outcomes are often context-dependent, varying by cell type, genetic background, and microenvironment.

For example, KRAS-mutant lung cancer is highly resistant to single-agent therapies, there is already preclinical evidence supporting the potential of combined MEK, PI3K, and CDK4/6 inhibition for KRAS-mutant cancers. Analysis of the Tahoe-100M dataset supports a triple combination of MEK, PI3K, and CDK4/6 inhibitors. This strategy overcomes MEK inhibition limits, driving strong cell cycle arrest, dual oncogene suppression, and enhanced apoptosis.

Supported by preclinical data, clinical trials, and the FDA approval of each agent, this precision medicine approach offers a promising strategy for KRAS-driven cancers across multiple tumor types.

From Noise to Scale: Addressing Challenges in scRNA-seq Model Training

SPARSITY AND NOISE

Like any technology, scRNA-seq has inherent limitations. Some are biological, such as stochastic transcriptional fluctuations, where cells with the same genome and environment still produce different transcript counts purely by chance.

More often, noise arises from technical factors.

Sparsity is caused by dropout events where transcripts present in a cell go undetected. This results in many zeros in the expression matrix, making it difficult to tell whether a gene is truly absent or just missed. Additional variability from sequencing depth, amplification bias, and RNA capture efficiency further masks biological patterns.

To address noise, researchers rely on statistical and deep learning methods. Statistical approaches such as log-normalization or SCTransform stabilize expression values for better comparison across cells. Models like ZINB-WaVE go further, explicitly modeling “extra zeros” to distinguish biological absence from dropout.

Deep learning approaches, including denoising autoencoders and other generative models, tackle noise from a data-driven perspective and learn to reconstruct clean signals by introducing noise during training and recovering structured patterns through latent representations.

BATCH EFFECTS

Large-scale, heterogeneous datasets are central to modern drug discovery, but they introduce the challenge of batch effects. Historically, balancing throughput, reproducibility, and multi-timepoint sampling was difficult while keeping batch variability under control.

Recent experimental and computational innovations have made this more manageable. Fixation-based scRNA-seq technologies halt biological processes at collection, allowing samples to be stored long-term and processed in large batches. This enables the design of large-scale experiments with hundreds or thousands of conditions, processed together at convenient times.

Combined with high-throughput and automated technologies, millions of data points can be generated quickly with less personnel, while minimizing batch-driven artifacts.

SCALE

Atlases are rapidly expanding with datasets now reaching billions of cells. A landmark atlas containing 100 million single cells has already been produced and is actively being used to train AI systems capable of answering diverse questions. This dataset was efficiently generated using plate-based combinatorial barcoding, a scalable and efficient platform for gigascale atlases.

To match this order of magnitude, computational algorithms must not only scale effectively but also remain resilient to batch effects, rare populations, and heterogeneous samples.

Multiple AI solutions address this:

- scPoli is a next-generation semi-supervised conditional generative model and open-world learner. Learns representations of cells and samples, integrates across diverse studies, transfers labels from annotated datasets and maps new cells to reference atlases while leaving room for discovery of novel populations.
- scVI, totalVI are generative models for integration and multimodal analysis.
- scBERT, scGPT are pretrained transformer models designed for transfer learning and rare cell detection.
- Geneformer, scFoundation: are foundation transformers pretrained on >30M, up to 100M single cell profiles, capable of transferring learning like cell-type annotation, perturbation and drug response prediction across downstream tasks.
- Harmony, Scanorama, Azimuth are efficient frameworks for batch correction and reference mapping.

Together, these methods allow atlas-scale analysis, scaling from millions to billions of cells while still capturing fine-grained biology and rare states.

MULTIPLEXING AND DEMULTIPLEXING

Pooling cells from multiple samples before sequencing increases throughput and reduces batch effects, often by tagging them with unique identifiers. Demultiplexing is the subsequent computational step that uses these tags to determine each cell's original sample of origin, allowing for the correct sample assignment and identification of technical artifacts like multiplets.

Greater multiplexing enhances both scalability and modularity. More samples can be processed per run, lowering costs, reducing batch variability, and enabling atlas-scale datasets.

At the same time, experiments become more flexible: researchers can design complex perturbation or time-course studies, expand projects iteratively, and integrate data across studies more easily.

Split-pool scRNA-seq platforms introduce combinatorial barcodes directly during library preparation. This enables efficient multiplexing of hundreds or even thousands of samples, with demultiplexing handled computationally through these barcodes.

Different scRNA-seq technologies typically provide their own dedicated pipelines for this step.

Conclusions

The integration of scRNA-seq and AI is reshaping how we study biology and discover therapies. scRNA-seq provides an unprecedented view of cellular diversity, while AI transforms these massive, noisy datasets into predictive models that capture gene regulation, drug response, and disease mechanisms. Together, they make it possible to move beyond descriptive catalogs of cell states toward actionable insights: identifying novel drug targets, predicting therapeutic outcomes, personalizing treatments, and designing rational drug combinations.

As datasets scale from millions to billions of cells and multiplexing technologies make complex experiments routine, the synergy between scRNA-seq and AI will only deepen.

With increasingly powerful computational models trained on ever-larger single-cell atlases, biology can now be explored at a resolution and scale that was previously unattainable, transforming drug discovery into a precise, data-driven science.