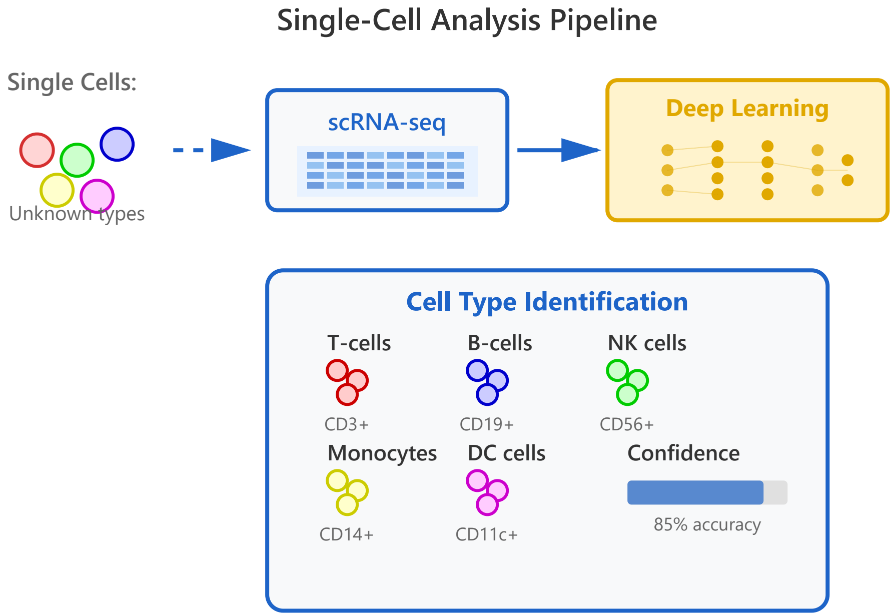
Cell Type Classification



Zero-shot Learning: Can identify novel cell types without prior training

Reference mapping → Batch correction → Uncertainty estimation

Single-cell models

scBERT, Geneformer architectures

Reference mapping

Atlas-based annotation

Zero-shot learning

Novel cell type discovery

Batch correction

Remove technical variation

Uncertainty estimation

Confidence scoring

1 Single-cell Models

Overview

Single-cell foundation models are deep learning architectures specifically designed to understand and analyze gene expression patterns at the individual cell level. These models leverage transformer-based architectures, similar to those used in natural language processing, to learn meaningful representations of cellular states.

Key Architectures

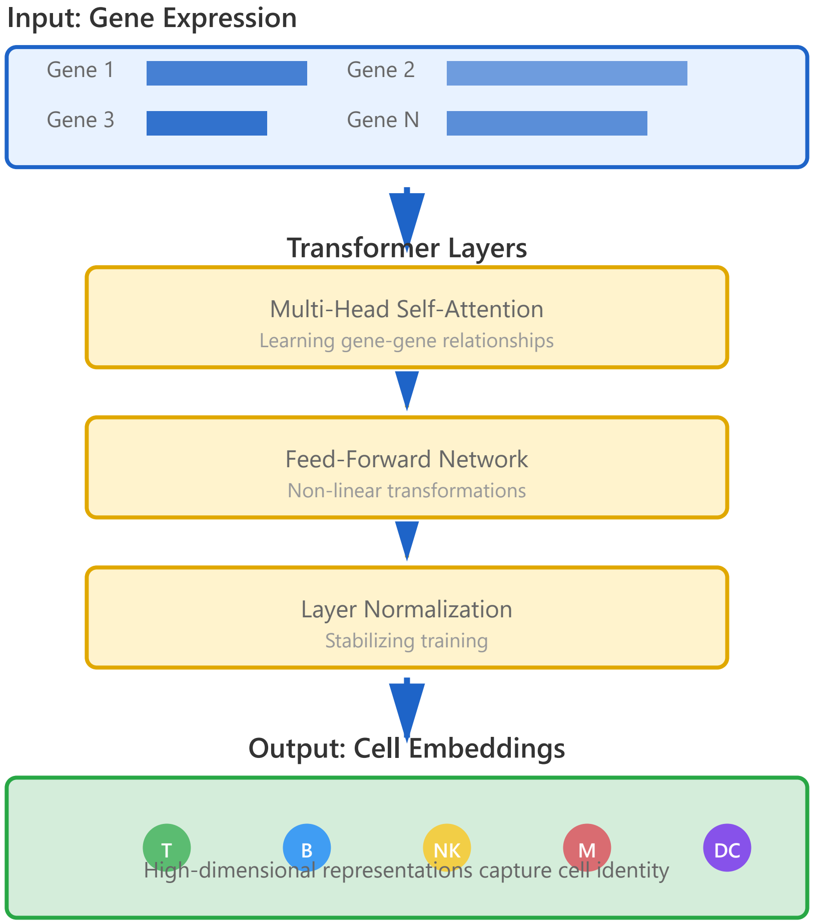
scBERT (Single-cell BERT): Adapts the BERT architecture for single-cell RNA-seq data, treating genes as "words" and cells as "sentences". The model learns contextual relationships between genes through self-supervised pre-training on large-scale datasets.

Geneformer: A transformer model that processes genes ranked by expression level, enabling the model to capture gene regulatory networks and cellular hierarchies. It can predict cell fate and identify key regulatory genes.

Key Capabilities:

- Learn universal gene expression patterns across millions of cells
- Transfer learning to new datasets with minimal fine-tuning
- Identify cell type-specific gene signatures automatically
- Predict cell states and developmental trajectories

Single-cell Model Architecture



Input: Gene Expression

Gene 1

Gene 2

Gene 3

Gene N

Transformer Layers

Multi-Head Self-Attention

Learning gene-gene relationships

Feed-Forward Network

Non-linear transformations

Layer Normalization

Stabilizing training

Output: Cell Embeddings

T

B

NK

M

DC

High-dimensional representations capture cell identity

Why It Matters

Foundation models enable accurate cell type classification even with limited labeled data, dramatically reducing the need for manual annotation and improving consistency across different studies.

2 Reference Mapping

Overview

Reference mapping is a transfer learning approach that annotates new single-cell datasets by comparing them to well-characterized reference atlases. This method leverages comprehensive, manually curated cell type annotations from large-scale projects to automatically classify cells in new experiments.

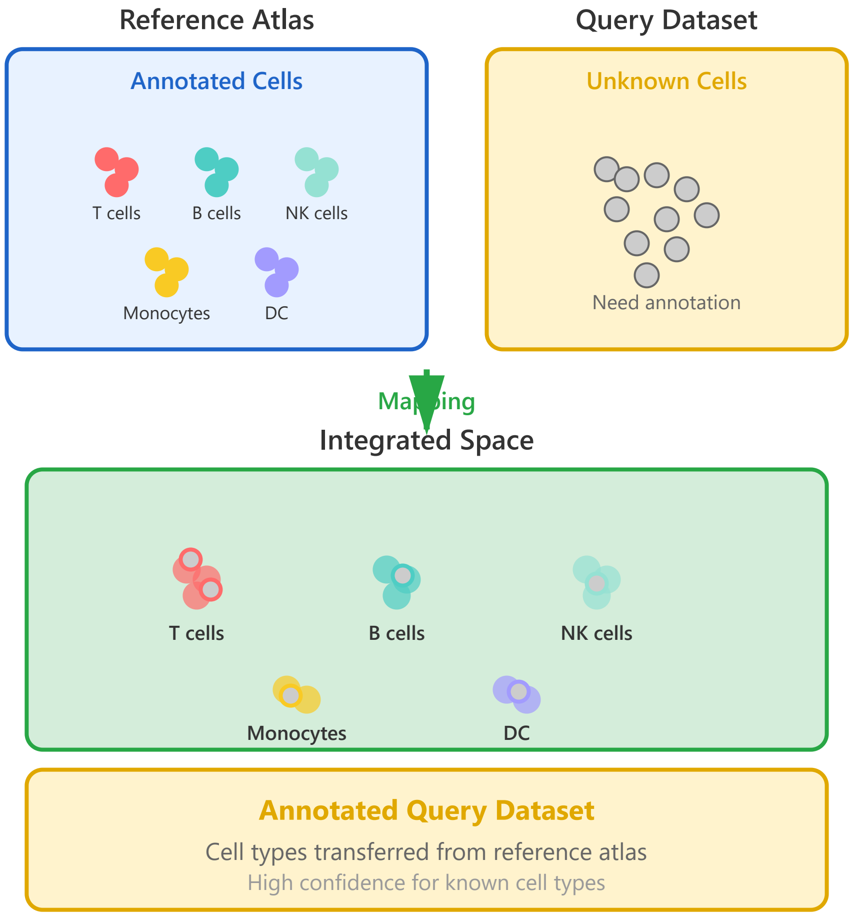
How It Works

The process involves projecting query cells into the same embedding space as reference cells, then transferring labels based on similarity. Advanced methods like Seurat's reference mapping and Symphony use canonical correlation analysis (CCA) and harmony integration to align datasets while preserving biological variation.

Key Advantages:

- Leverages expert knowledge from reference atlases (e.g., Human Cell Atlas)
- Consistent annotations across different studies and laboratories
- Fast inference without requiring model training

Reference Mapping Pipeline



- Works well for common, well-characterized cell types

Limitations:

- May struggle with novel or rare cell types not in reference
- Depends on quality and comprehensiveness of reference atlas
- Can be affected by batch effects between datasets

Real-World Application

Reference atlases like Tabula Sapiens contain millions of annotated human cells across 24 tissues, enabling researchers to classify cells in new disease studies rapidly and accurately.

3 Zero-shot Learning

Overview

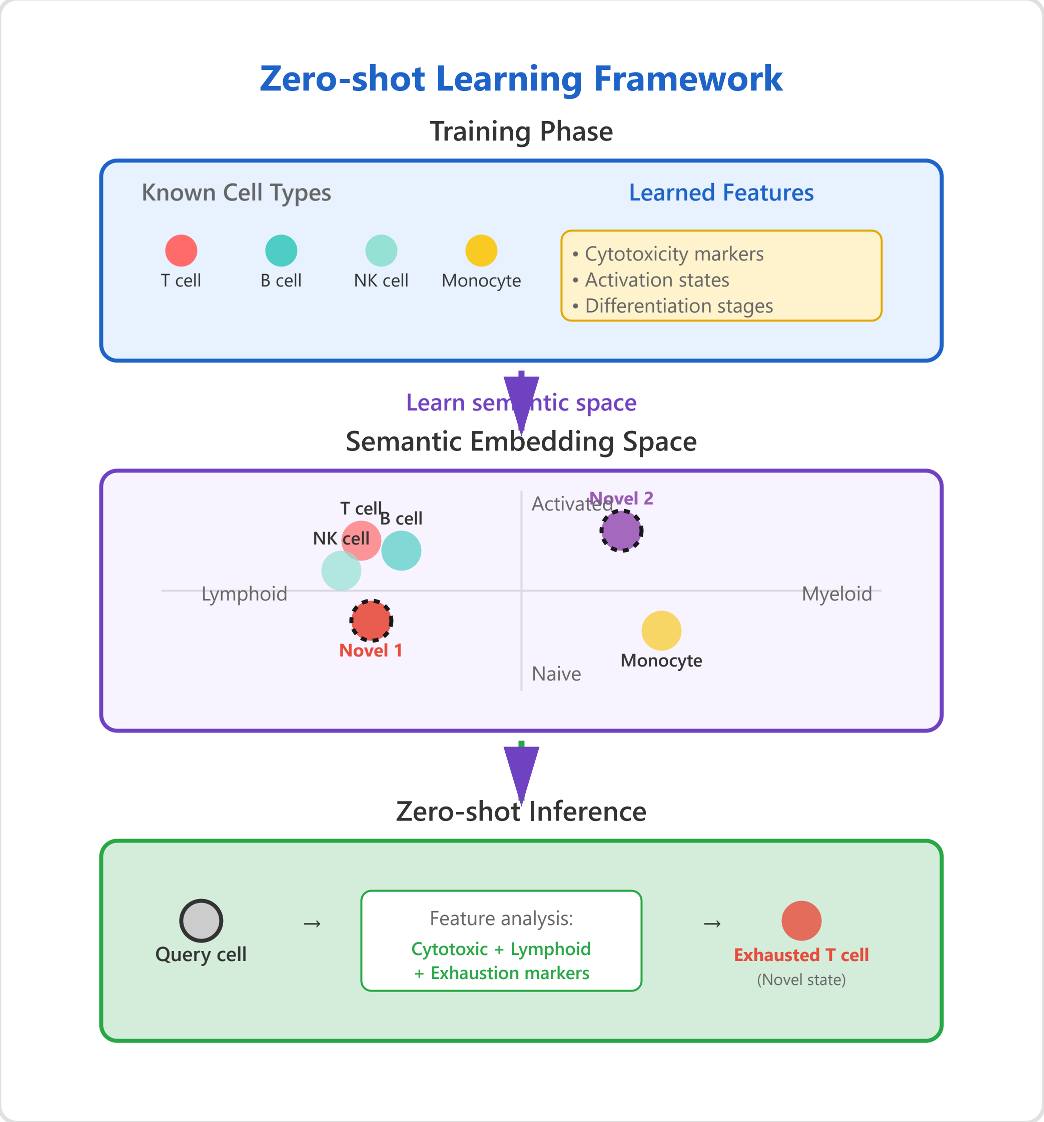
Zero-shot learning enables AI models to identify and classify cell types that were never seen during training. This revolutionary capability is crucial for discovering novel cell populations, rare cell states, and disease-specific cell types that don't exist in healthy reference atlases.

Mechanism

Zero-shot models learn a semantic embedding space where cell types are represented based on their functional properties and gene expression characteristics rather than explicit labels. The model can recognize new cell types by understanding the relationships between genes and cellular functions, similar to how humans can identify an unfamiliar animal by recognizing its features.

Approaches

Semantic Embeddings: Models learn to associate gene expression patterns with cell type descriptions or functional annotations, enabling classification based on textual descriptions.



Zero-shot Learning Framework

Training Phase

Known Cell Types

Learned Features

- Cytotoxicity markers
- Activation states
- Differentiation stages

Learn semantic space

Semantic Embedding Space

Zero-shot Inference

Feature analysis:

Cytotoxic + Lymphoid
+ Exhaustion markers

Compositional Learning: Breaking down cell types into fundamental properties (e.g., "cytotoxic" + "lymphocyte" = "cytotoxic T cell"), allowing recognition of novel combinations.

Key Capabilities:

- Identify disease-specific cell states not in healthy references
- Discover rare or transitional cell populations
- Classify cells in non-model organisms with limited annotations
- Adapt to emerging cell type nomenclature

Breakthrough Impact

Zero-shot learning was instrumental in identifying novel immune cell states in COVID-19 patients and discovering rare developmental intermediates in embryonic development studies.

4 Batch Correction

Overview

Batch effects are systematic technical variations that arise from differences in experimental conditions, reagents, sequencing platforms, or processing times. These non-biological variations can obscure true biological signals and lead to incorrect cell type classifications if not properly addressed.

Sources of Batch Effects

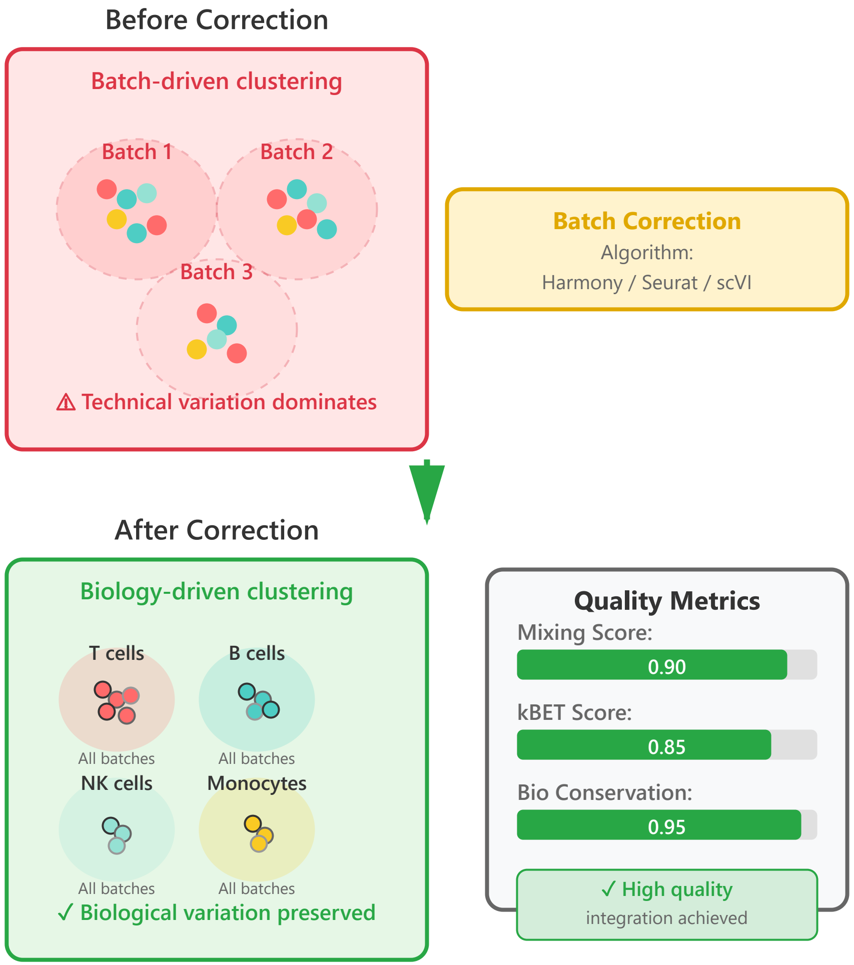
Common sources include differences in cell capture efficiency, library preparation protocols, sequencing depth, ambient RNA contamination, and even the laboratory or technician performing the experiment. These effects can be so strong that cells cluster by experimental batch rather than biological cell type.

Correction Methods

Harmony: A fast integration method that iteratively corrects batch effects while preserving biological variation using soft k-means clustering in PCA space.

Seurat Integration: Uses canonical correlation analysis (CCA) to identify shared correlation structures across

Batch Correction Process



batches, anchoring datasets together based on mutual nearest neighbors.

scVI (Single-cell Variational Inference): A deep learning approach using variational autoencoders to model both biological variation and batch effects simultaneously, learning a corrected latent representation.

Best Practices:

- Always visualize data before and after correction with UMAP/t-SNE
- Verify that biological variation is preserved, not removed
- Use multiple quality metrics (mixing metrics, kBET, LISI)
- Consider whether correction is necessary - some "batches" may have real biology

> #### Critical Consideration
>
> Over-correction can remove genuine biological differences. For example, disease-control comparisons should preserve disease-specific cell states while removing technical variation.

5 Uncertainty Estimation

Overview

Uncertainty estimation quantifies the confidence of cell type predictions, distinguishing between cells that are confidently classified and those in ambiguous states. This is crucial for identifying transitional cells, doublets (two cells captured together), low-quality cells, and truly novel cell populations that require manual curation.

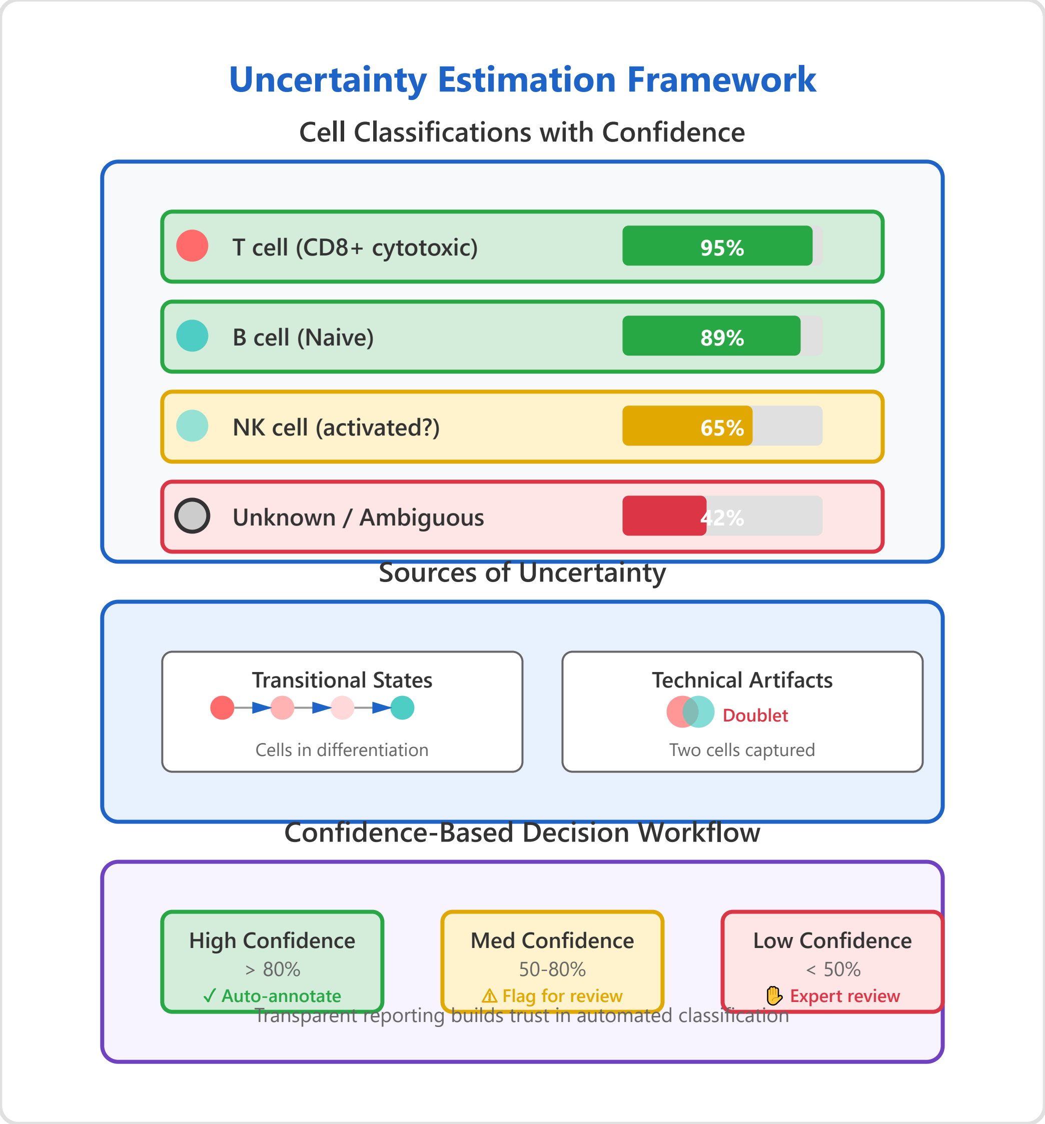
Types of Uncertainty

Aleatoric Uncertainty: Inherent noise in the data due to technical limitations like low gene capture efficiency or stochastic gene expression. This is irreducible uncertainty in the measurement itself.

Epistemic Uncertainty: Uncertainty arising from model limitations or lack of training data. This can be reduced with more data or better models and indicates cells that are far from known training examples.

Estimation Methods

Probabilistic Classifiers: Models output probability distributions over cell types rather than hard labels, with entropy serving as an uncertainty measure.



Uncertainty Estimation Framework

Cell Classifications with Confidence

T cell (CD8+ cytotoxic) 95%

B cell (Naive) 89%

NK cell (activated?) 65%

Unknown / Ambiguous 42%

Sources of Uncertainty

Transitional States

Cells in differentiation

Technical Artifacts

Doublet

Two cells captured

Confidence-Based Decision Workflow

High Confidence

> 80%

✓ Auto-annotate

Med Confidence

50-80%

⚠ Flag for review

Low Confidence

< 50%

✋ Expert review

Transparent reporting builds trust in automated classification

Monte Carlo Dropout: Running multiple predictions with dropout enabled to create a distribution of predictions, estimating uncertainty from variance.

Ensemble Methods: Training multiple models and measuring disagreement between their predictions.

Practical Applications:

- Flag ambiguous cells for manual review by experts
- Identify potential novel cell states requiring further investigation
- Detect technical artifacts (doublets, damaged cells)
- Prioritize cells for validation experiments
- Provide honest assessment of classification reliability

Clinical Relevance

In clinical applications, uncertainty estimation is critical. High-confidence predictions can guide treatment decisions, while low-confidence cases can be flagged for additional testing or expert review.