Integration Methods for Single-Cell Analysis

Anchor-based Methods

Seurat, LIGER find correspondence between datasets

Deep Learning Approaches

scVI, scGAN learn shared latent space

Reference Building

Create comprehensive cell atlases

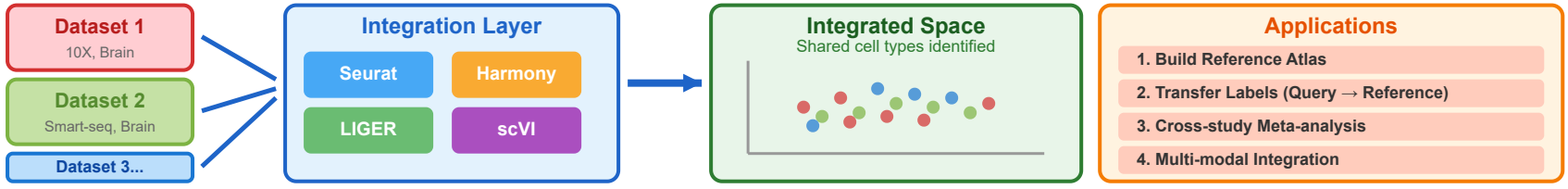
Query Mapping

Project new data onto reference

Performance Metrics

Biological conservation vs batch mixing

💡 Integration enables meta-analysis and transfer learning



Detailed Method Descriptions

1 Anchor-based Methods

Overview

Anchor-based methods identify mutual nearest neighbors (MNNs) or "anchors" between datasets to align cells across different batches or technologies. These anchors serve as reference points for integration.

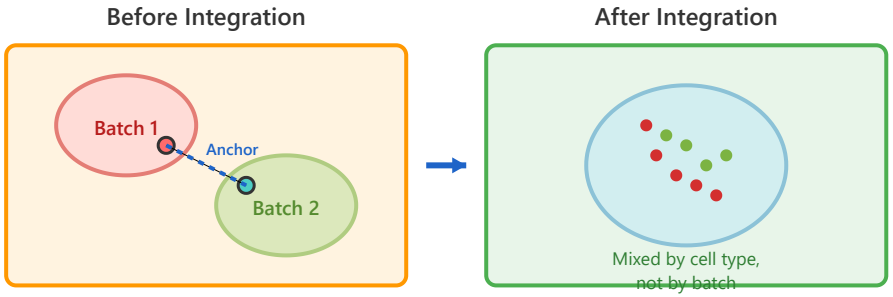
Key Algorithms

- **Seurat v3/v4:** Uses canonical correlation analysis (CCA) to find shared correlation structures, then identifies anchors via mutual nearest neighbors
- **LIGER:** Uses integrative non-negative matrix factorization (iNMF) to discover shared and dataset-specific factors
- **Harmony:** Iteratively clusters cells and corrects for batch effects using soft k-means clustering

Advantages

- Computationally efficient for large datasets
- Preserves biological variation while removing batch effects
- Well-validated and widely adopted

Anchor-based Integration Process



Algorithm Steps

1. **Feature Selection**
   Identify highly variable genes across datasets
2. **Anchor Identification**
   Find mutual nearest neighbors (MNNs) between datasets
3. **Integration**
   Use anchors to align and integrate datasets into shared space

⭐ Key Points

- Anchors are pairs of cells from different datasets that represent the same cell type or state
- MNN approach ensures bidirectional nearest neighbors, reducing false positives

- Works well when datasets share common cell types but differ in technology or batch

2 Deep Learning Approaches

Overview

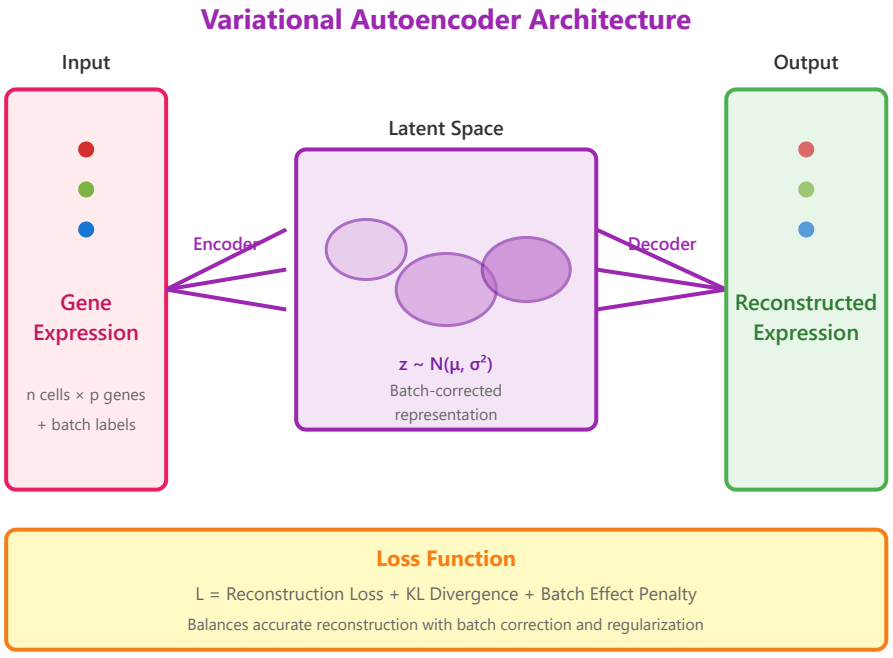
Deep learning methods use neural networks to learn a shared latent representation that captures biological variation while removing technical effects. These models can handle complex non-linear relationships.

Key Algorithms

- **scVI (Single-Cell Variational Inference):** Uses variational autoencoders (VAE) to model gene expression with explicit batch correction
- **scGAN:** Generative adversarial network that learns batch-invariant representations
- **scANVI:** Semi-supervised version of scVI that incorporates cell type labels
- **SAUCIE:** Autoencoder with explicit regularization for batch effects

Advantages

- Captures complex non-linear relationships
- Probabilistic framework provides uncertainty estimates
- Can incorporate multiple data modalities



- Scales to millions of cells

Key Points

- VAE learns a probabilistic latent representation with explicit uncertainty quantification
- Batch information is provided as input but removed from latent space through adversarial training or regularization
- Can impute missing values and denoise expression data as part of the integration process
- Requires GPU acceleration for large datasets but provides state-of-the-art performance

3 Reference Building

Overview

Reference building involves creating comprehensive cell atlases by integrating multiple high-quality datasets. These references serve as standardized maps of cell types and states for a given tissue or organism.

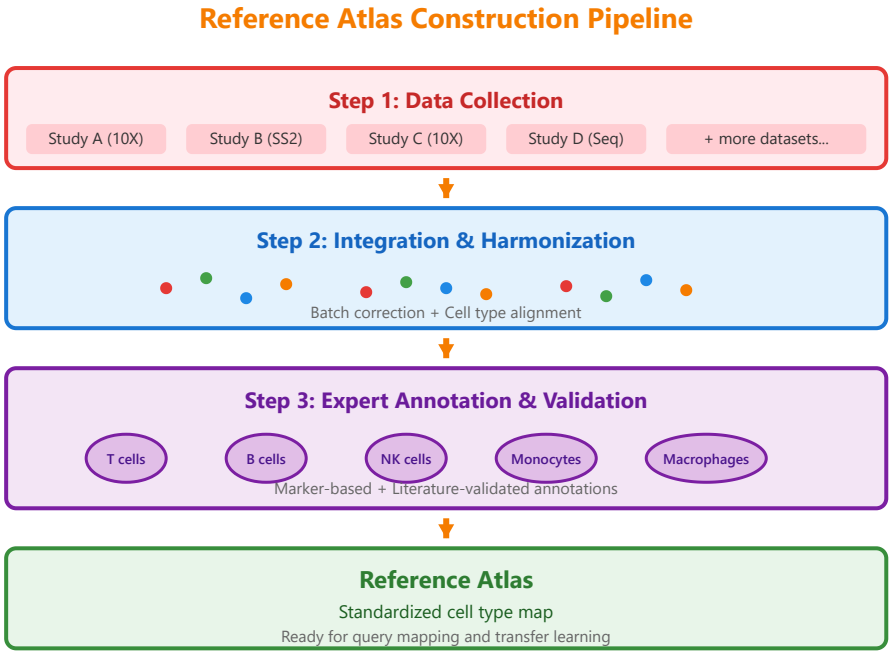
Construction Process

- **Data Collection:** Aggregate datasets from multiple sources, technologies, and conditions
- **Quality Control:** Rigorous filtering and validation of cells and annotations
- **Integration:** Harmonize datasets while preserving biological heterogeneity
- **Annotation:** Comprehensive cell type labeling by expert curators

- **Validation:** Cross-validation and benchmarking against known markers

Major Reference Atlases

- **Human Cell Atlas (HCA):** Comprehensive map of human cells
- **Tabula Sapiens:** Reference for human organ systems
- **Mouse Cell Atlas:** Complete mouse cell type map



⭐ Key Points

- References require diverse, high-quality datasets spanning multiple conditions and technologies
- Expert curation ensures accurate and consistent cell type annotations
- Regular updates incorporate new data and refined annotations
- Enables standardized analysis and comparison across studies

4 Query Mapping

Overview

Query mapping projects new datasets onto existing reference atlases, enabling rapid cell type annotation and

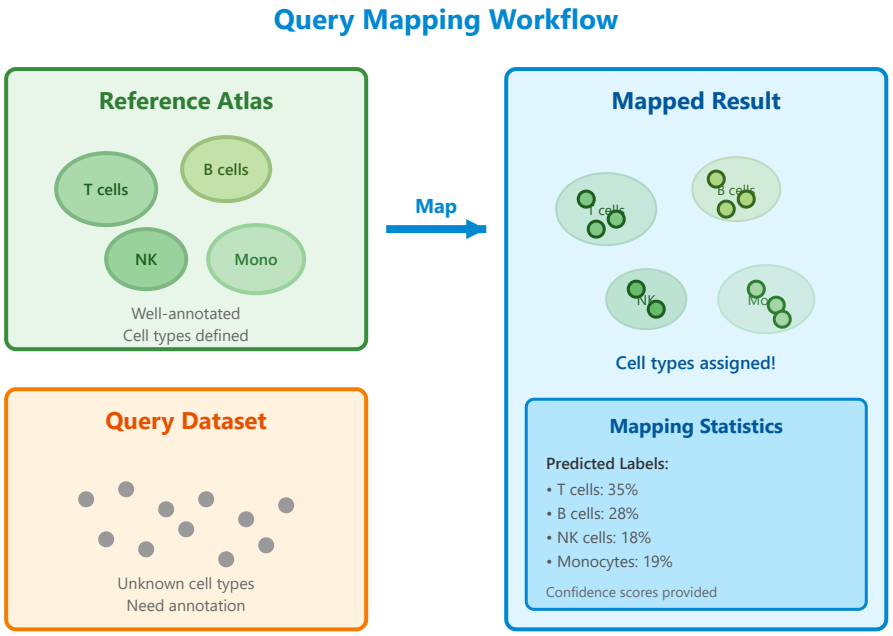
comparison without full re-integration. This is a form of transfer learning for single-cell analysis.

Mapping Methods

- **Seurat Reference Mapping:** Projects query onto PCA space of reference using supervised anchors
- **Symphony:** Fast reference mapping using harmonized coordinates
- **scArches:** Transfer learning with neural networks, updates model with query data
- **SingleR:** Correlation-based cell type assignment from reference

Applications

- Rapid annotation of new datasets
- Cross-study comparisons
- Disease vs. healthy comparison
- Time-series or perturbation studies



Key Points

- Much faster than full integration (seconds to minutes vs. hours)
- Provides confidence scores for cell type predictions
- Works best when query and reference share common cell types
- Novel cell types in query may be assigned to nearest reference type (limitation)
- Enables large-scale studies by reusing well-curated references

5 Performance Metrics

Overview

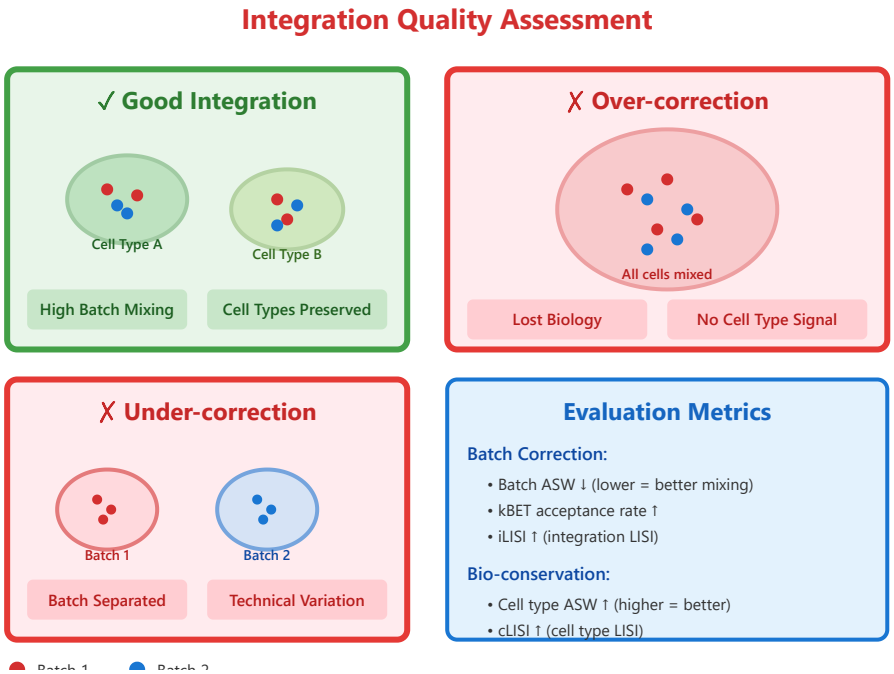
Evaluating integration quality requires balancing two competing objectives: removing technical variation (batch mixing) while preserving biological variation (cell type separation).

Key Metrics

- **Batch Mixing Metrics:**
  - Batch ASW (Average Silhouette Width): Measures batch separation
  - kBET (k-nearest neighbor Batch Effect Test): Tests batch mixing
  - LISI (Local Inverse Simpson's Index): Quantifies local diversity
- **Bio-conservation Metrics:**
  - Cell type ASW: Measures cell type separation
  - ARI (Adjusted Rand Index): Compares clustering to labels
  - NMI (Normalized Mutual Information): Information preservation
- **Trajectory Preservation:** Ensures developmental/temporal relationships maintained

Benchmarking Studies

- No single method dominates all scenarios



- Performance depends on dataset characteristics and goals

⭐ Key Points

- Integration is a balancing act: remove batch effects without losing biology
- Multiple metrics needed to assess both batch mixing and bio-conservation
- Over-correction merges distinct cell types; under-correction leaves technical variation
- Optimal method depends on dataset characteristics (number of batches, cell types, technologies)
- Visual inspection (UMAP plots) combined with quantitative metrics provides best assessment