Batch Effect Correction

Comprehensive Guide to Single-Cell Integration Methods

MNN Correction

Mutual nearest neighbors for batch alignment

Harmony Algorithm

Iterative clustering and correction

LIGER Integration

Integrative non-negative matrix factorization

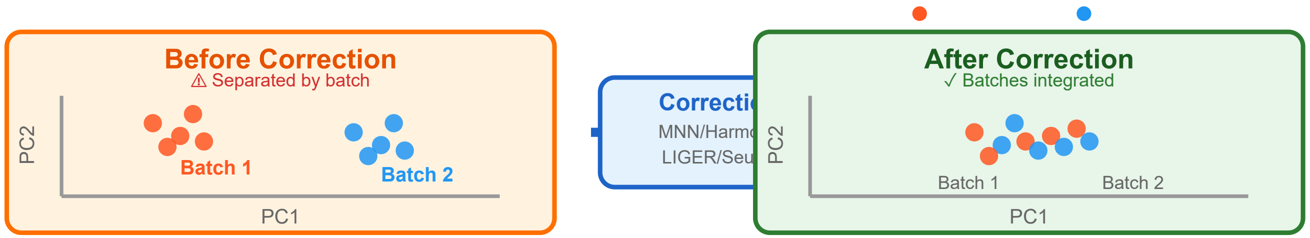
Seurat Integration

Canonical correlation analysis + anchors

Benchmark Studies

Compare methods on simulated and real data

💡 Critical for multi-sample and multi-technology integration



Integration Goals

- ✓ Mix batches properly
- ✓ Preserve biological variation
- ✓ Maintain cell type identity
- ✓ Remove technical artifacts

1 MNN Correction (Mutual Nearest Neighbors)

Overview

MNN correction identifies pairs of cells from different batches that are mutual nearest neighbors in high-dimensional space. These pairs represent cells of the same type across batches and are used to calculate correction vectors.

How It Works

- **Step 1:** Identify mutual nearest neighbors between batch pairs
- **Step 2:** Calculate correction vectors from MNN pairs
- **Step 3:** Apply correction to all cells using weighted averaging
- **Step 4:** Preserve local structure within each batch

> **Key Concept:** If cell A in batch 1 is among the nearest neighbors of cell B in batch 2, AND cell B is among the nearest neighbors of cell A, they form an MNN pair.

Best Used For

- Integration of datasets with similar cell type compositions
- Scenarios where batch effects are relatively mild

- When preserving rare cell populations is important

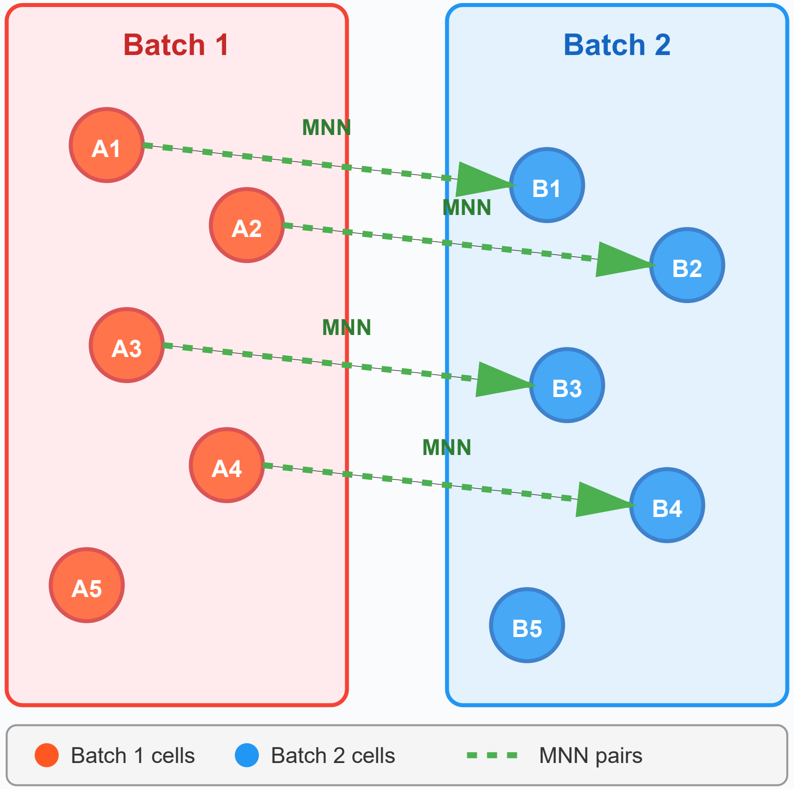
✓ Advantages

- Preserves biological variation well
- No overcorrection of rare populations
- Computationally efficient

✗ Limitations

- Requires overlapping cell types
- Less effective for strong batch effects
- May struggle with very different batches

MNN Correction Process



2 Harmony Algorithm

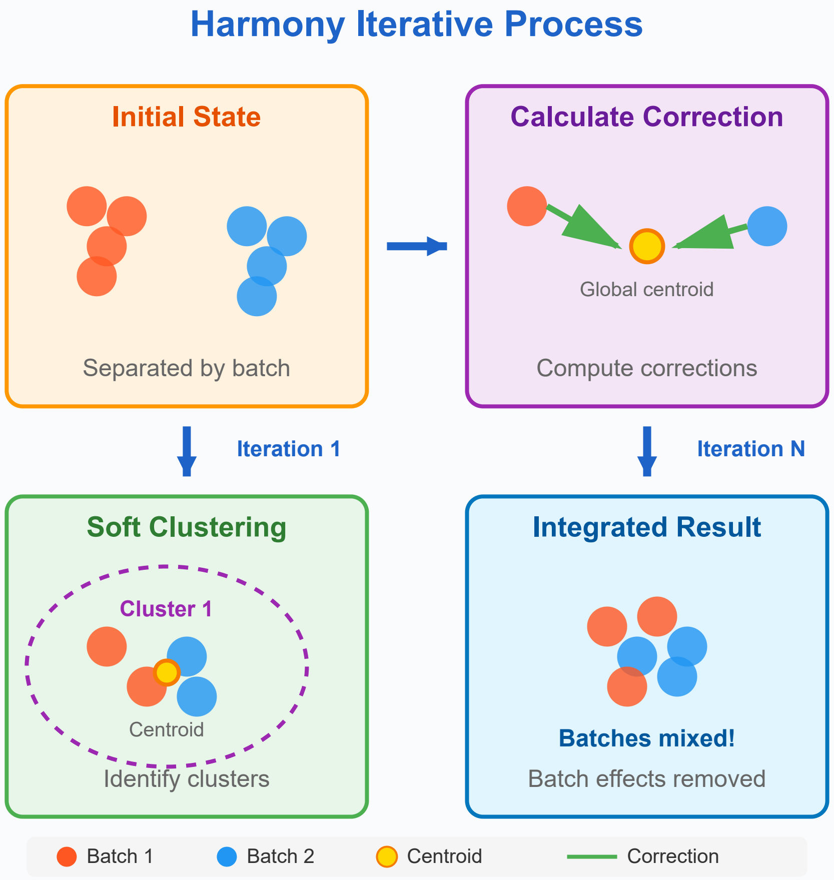
Overview

Harmony is an iterative algorithm that performs batch correction by soft clustering cells and then correcting their positions to remove batch-specific effects while preserving biological structure. It works directly on PCA embeddings.

How It Works

- **Step 1:** Start with PCA-reduced data
- **Step 2:** Perform soft clustering to identify cell groups
- **Step 3:** Calculate batch-specific centroids for each cluster
- **Step 4:** Correct cell positions toward global centroids
- **Step 5:** Iterate until convergence

> **Key Advantage:** Harmony uses a diversity penalty to ensure that clusters are balanced across batches, preventing overclustering of any single batch.



Best Used For

- Large-scale datasets with multiple batches
- Strong batch effects requiring aggressive correction
- Fast integration of many samples (computationally efficient)
- Integration across different sequencing platforms

✓ Advantages

- Very fast and scalable
- Works on existing PCA embeddings

✗ Limitations

- May overcorrect biological variation

- Handles strong batch effects well
- Simple to implement

- Less control over correction strength
- Can merge distinct cell states

3 LIGER Integration (iNMF)

Overview

LIGER (Linked Inference of Genomic Experimental Relationships) uses integrative non-negative matrix factorization (iNMF) to identify shared and dataset-specific factors. It decomposes gene expression into shared biological signals and batch-specific technical variations.

How It Works

- **Step 1:** Factorize each dataset's expression matrix
- **Step 2:** Identify shared factors (W) across datasets
- **Step 3:** Learn dataset-specific factors (V_1, V_2, ...)
- **Step 4:** Use shared factors for integration
- **Step 5:** Quantile normalize factor loadings

Mathematical Framework: For dataset i: $X_i \approx W \times H_i + V_i \times H_i$, where W contains shared factors and V_i contains dataset-specific factors.

Best Used For

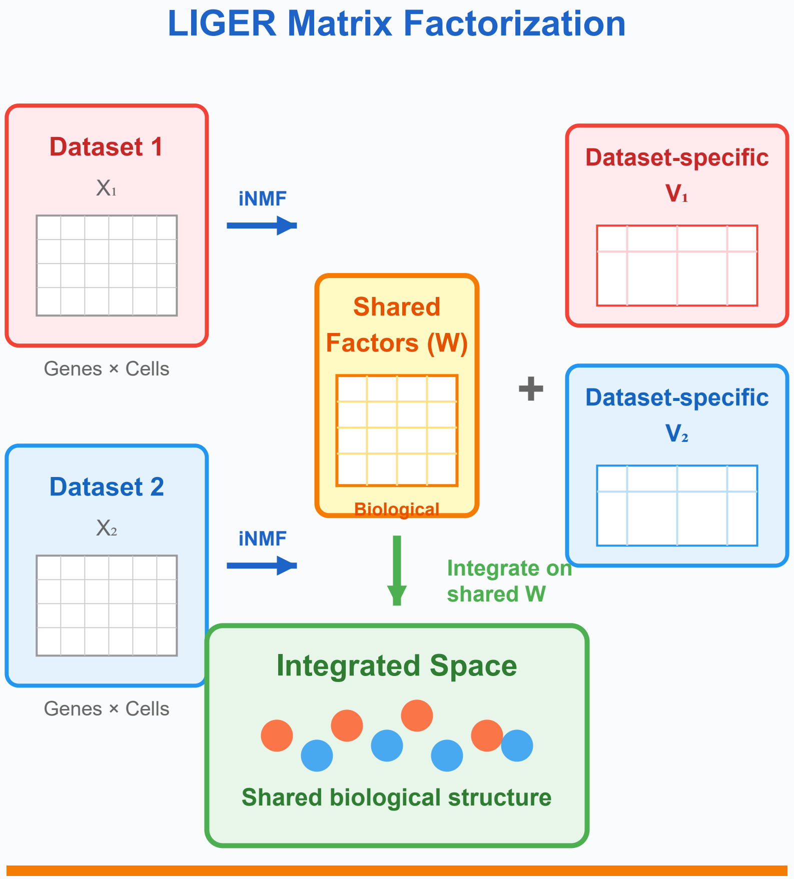
- Multi-modal data integration (e.g., scRNA-seq + scATAC-seq)
- Cross-species comparisons
- Datasets with fundamentally different feature sets
- When you need to identify shared vs. unique biological signals

✓ Advantages

- Works with different feature spaces
- Identifies shared biology explicitly
- Great for multi-modal integration
- Preserves dataset-specific signals

✗ Limitations

- Computationally intensive
- Requires parameter tuning (k factors)
- More complex to implement
- Memory intensive for large datasets



4 Seurat Integration (CCA + Anchors)

Overview

Seurat integration uses Canonical Correlation Analysis (CCA) to identify shared correlation structures between datasets, then finds "anchor" cells that represent correspondences across batches. These anchors guide the integration process.

How It Works

- **Step 1:** Identify highly variable genes in each dataset
- **Step 2:** Perform CCA to find shared correlation structures
- **Step 3:** Identify mutual nearest neighbors as "anchors"
- **Step 4:** Score and filter anchors based on similarity
- **Step 5:** Use anchors to harmonize datasets

> **Key Innovation:** Anchors are high-confidence cell pairs that serve as reference points for integration, allowing precise correction while preserving biological heterogeneity.

Best Used For

- Standard scRNA-seq integration workflows

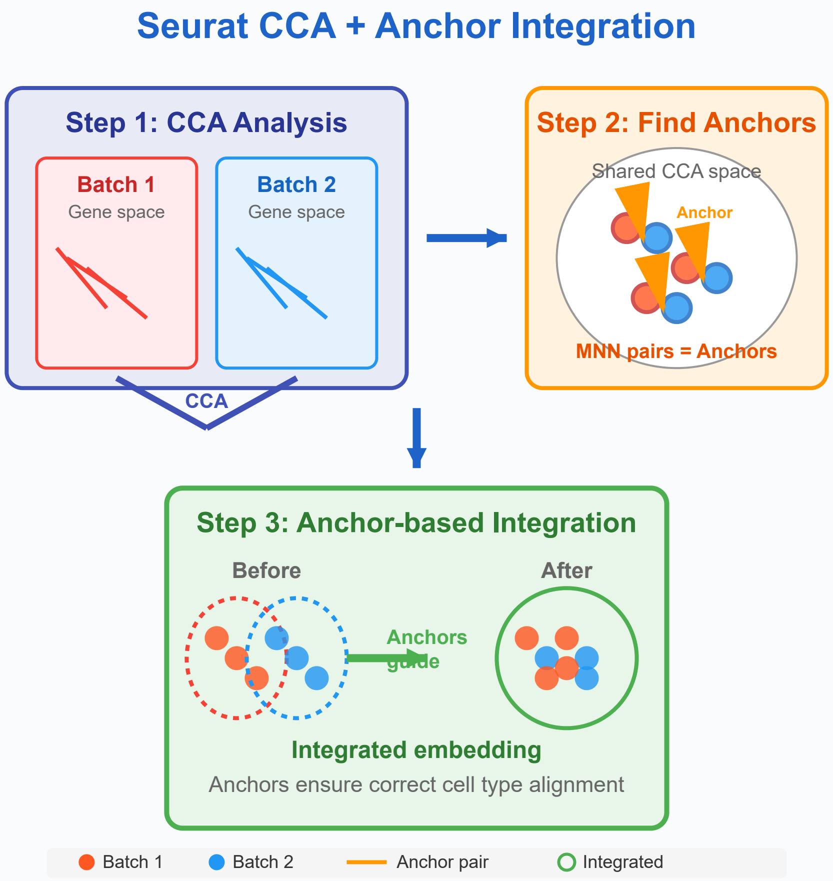
- Datasets with good cell type overlap
- When you want fine control over integration
- Reference-based integration (map query to reference)

✓ Advantages

- Well-established and validated
- Excellent documentation
- Works well with Seurat ecosystem
- Flexible anchor filtering

✗ Limitations

- Computationally demanding
- Memory intensive for large datasets
- Requires overlapping cell populations
- Can be slow with many datasets



5 Benchmark Studies & Method Comparison

Overview

Benchmark studies systematically compare integration methods using standardized metrics on both simulated and real datasets.

These studies help researchers choose the most appropriate method for their specific use case.

Key Evaluation Metrics

- **Batch Mixing:** How well cells from different batches are mixed (e.g., kBET, LISI)
- **Bio-conservation:** Preservation of biological variation (e.g., ARI, NMI, ASW)
- **Cell Type Purity:** Maintenance of distinct cell populations
- **Computational Efficiency:** Runtime and memory usage
- **Scalability:** Performance with increasing dataset size

> **Important Finding:** No single method is universally best. Method selection depends on batch effect strength, dataset characteristics, and analysis goals.

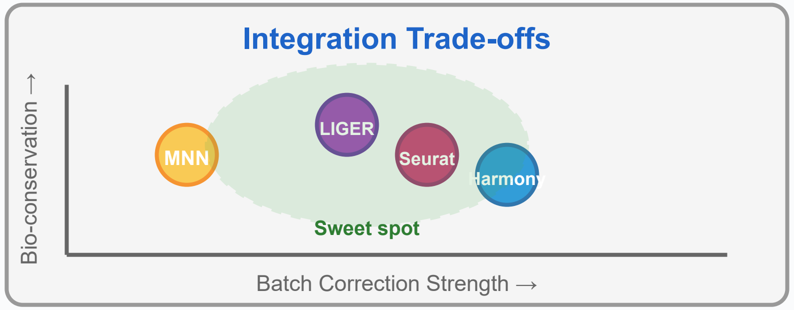
General Guidelines

- **Mild batch effects:** MNN, fastMNN
- **Strong batch effects:** Harmony, Seurat
- **Multi-modal data:** LIGER, MultiVI
- **Large-scale data:** Harmony (fastest), scVI
- **Reference mapping:** Seurat, Symphony

Notable Benchmark Papers

Method Performance Comparison

Method	Batch Mixing	Bio-conservation	Speed
MNN	75%	90%	70%
Harmony	95%	80%	95%
LIGER	85%	95%	60%
Seurat	90%	85%	55%



- Luecken et al. (2021) - Comprehensive scRNA-seq integration comparison
- Tran et al. (2020) - Evaluation across 77 batches
- Chazarra-Gil et al. (2021) - Flexible benchmarking framework

✓ Best Practices

- Test multiple methods on your data
- Check both mixing AND biology
- Visualize before/after integration
- Use appropriate metrics for evaluation

⚠ Common Pitfalls

- Overcorrection removes biology
- Undercorrection leaves batch effects
- Ignoring method assumptions
- Not validating integration quality

💡 Remember: Always validate your integration results by checking both batch mixing and biological signal preservation!