RECOMB 2026

CAMP: Coreset Accelerated Metacell Partitioning enables scalable analysis of single-cell data

Danrong Li, Young Kun Ko, Stefan Canzar

Pennsylvania State University, University of Regensburg

Problem: metacells help, but construction is now the bottleneck

Single-cell atlases

- Hundreds of thousands to millions of sparse cell profiles
- High redundancy: many cells occupy similar transcriptional states
- Need aggregation that denoises without erasing local structure

Existing metacell methods

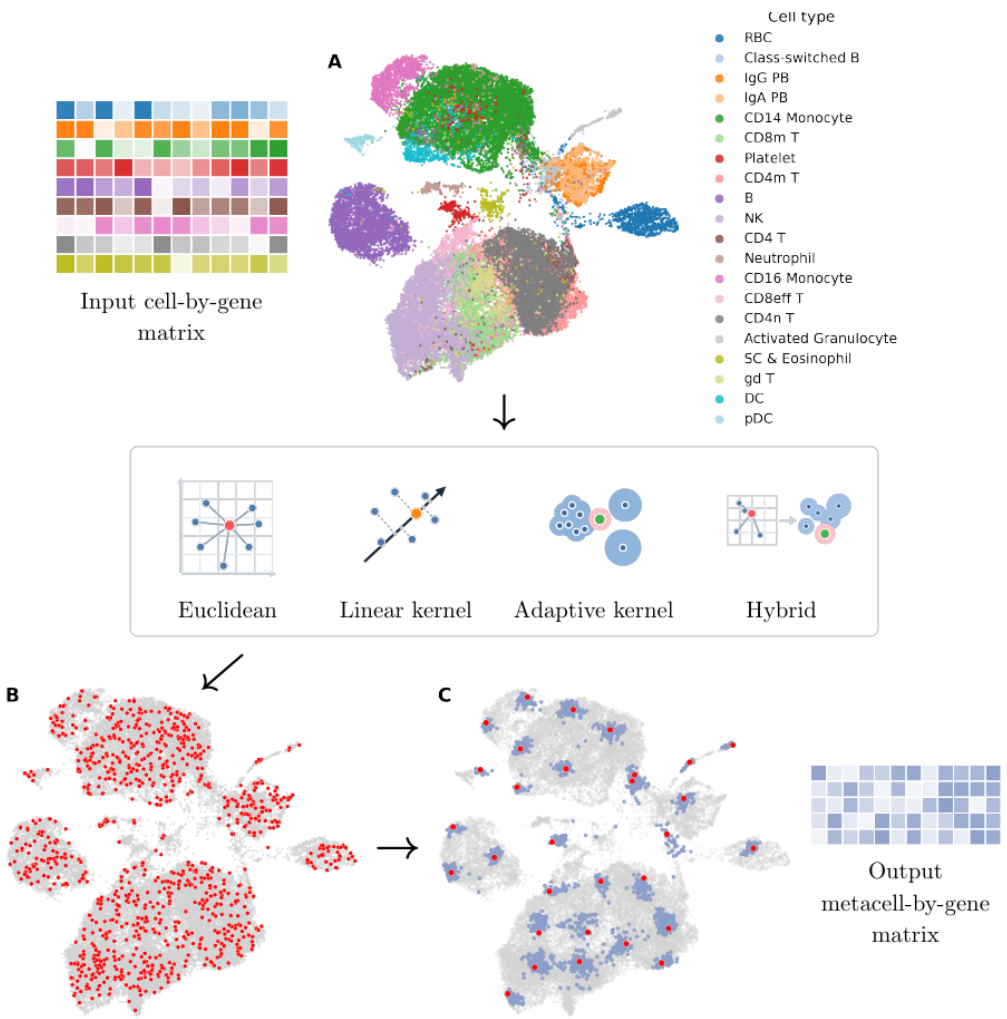
- Graph or kernel similarity structures
- Iterative refinement or model training
- Runtime and memory stress at atlas scale

What we want

- Local biological coherence
- Fast CPU-only construction
- Practical behavior on 500k+ cells

Can we replace global iterative optimization with representative cell anchoring, without losing biological signal?

Key idea: coreset anchors + assignment



1. Sample anchors

Select representative cells using a lightweight coreset distribution.

2. Treat as archetypes

The sampled cells serve as metacell centers.

3. Assign cells

Nearest anchor or highest kernel similarity gives the metacell partition.

Practical effect: avoid iterative archetypal optimization on the full dataset.

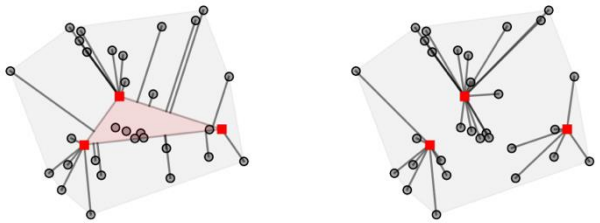
Algorithm: replace optimization with sampling and assignment

Weighted coreset sampling

For cell i, let d_i be distance to the mean of all n cells.

$$q_i = \frac{1}{2} \cdot \frac{1}{n} + \frac{1}{2} \cdot \frac{d_i^2}{\sum_{r=1}^{n} d_r^2}$$

- Designed for the k-means objective
- Every coreset for k-means is also a coreset for archetypal analysis
- In practice, sample approximately k anchors for k metacells



[Bachem et al. 2018, Mair & Brefeld 2019]

One-pass CAMP construction

1. Sample coreset C = {c1, ..., ck}
2. Initialize each selected cell as an anchor
3. Assign every cell to nearest or most similar anchor

Output: metacell membership map and aggregated expression profiles.

CAMP variants: default speed vs robust purity

CAMP is a family of four variants; the main practical choices are CAMP1 as the default and CAMP4 as the robust hybrid.

Variant	Space	Construction	Advantage	Use case
CAMP1	Euclidean	sample + nearest anchor	fast, simple, broad recovery	**default**
CAMP2	linear kernel	kernel alternative	directional similarity	specialized
CAMP3	adaptive kernel	density-aware alternative	nonlinear similarity	specialized
CAMP4	hybrid	Euclidean sampling + kernel assignment	highest purity, robust	**robust**

Practical guide: use CAMP1 as default; use CAMP4 when purity or robustness matters; keep CAMP2/3 as specialized kernel alternatives.

Benchmark design: quality, speed, and downstream signal

Datasets

44,721
PBMC cells

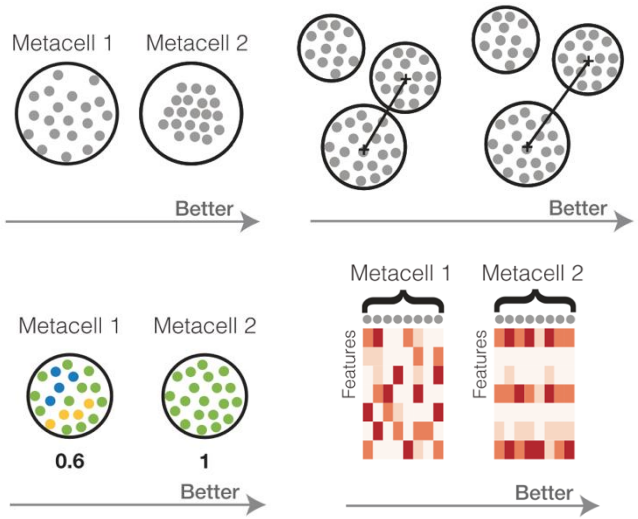
504,028
human fetal atlas cells

Standard preprocessing: normalization by total counts of 10,000 per cell, log(1+p), 2000 highly variable genes, and 50 principal components.

Competitors: SEACells, SuperCell, MetaCell, MetaCell2, MetaQ

Intrinsic metrics

- Compactness: cells within a metacell stay close
- Separation: nearby metacells remain distinct
- SC ratio balances both objectives above
- Purity and INV test biological homogeneity



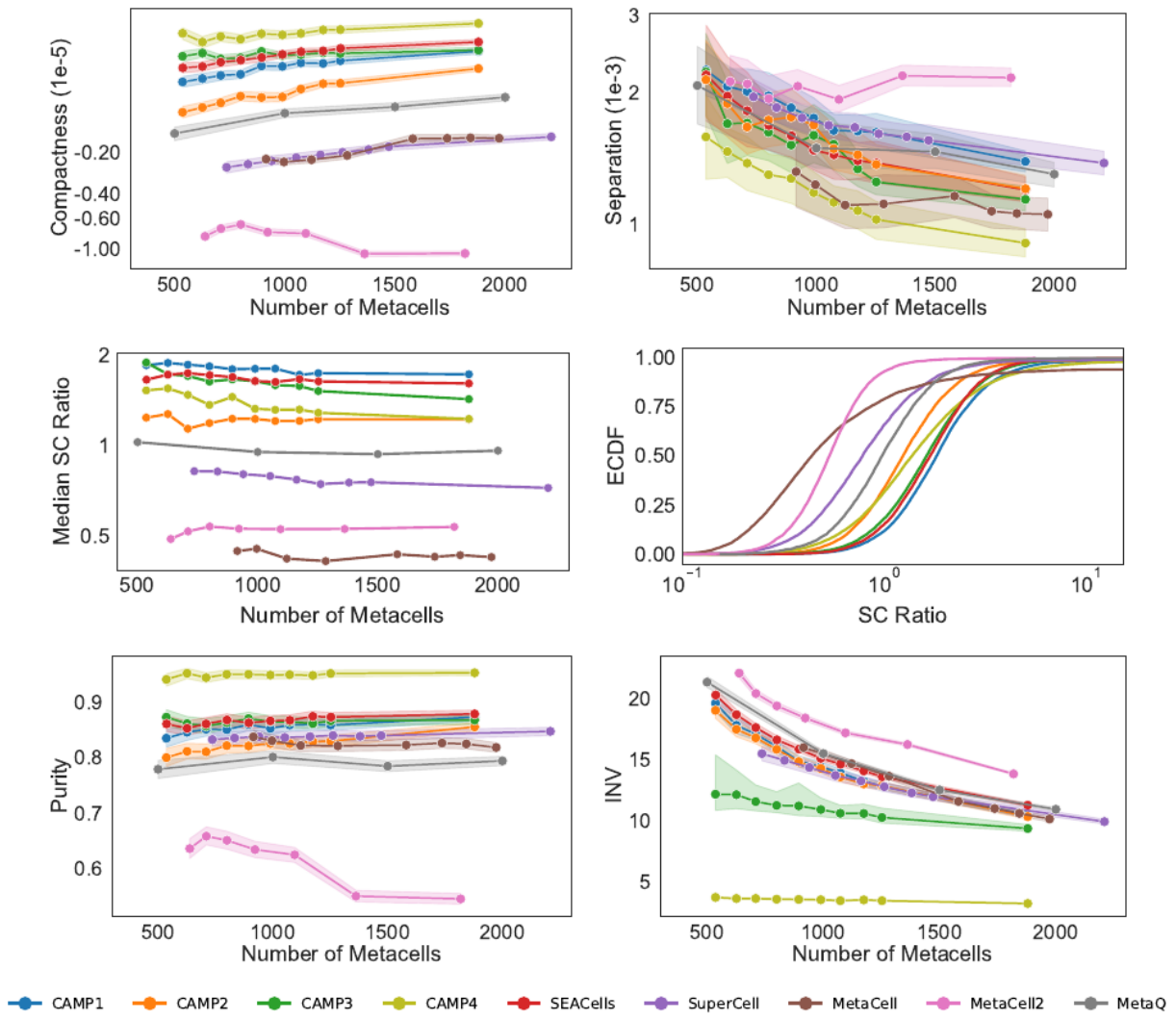
[Bilous et al. 2024]

Downstream test

- Aggregate metacell expression
- Assign majority cell type label
- Train CellTypist on metacells
- Predict original single cells

Metric: balanced accuracy over the ten most abundant cell types.

Result 1: High quality metacells (compact, well-separated, pure)

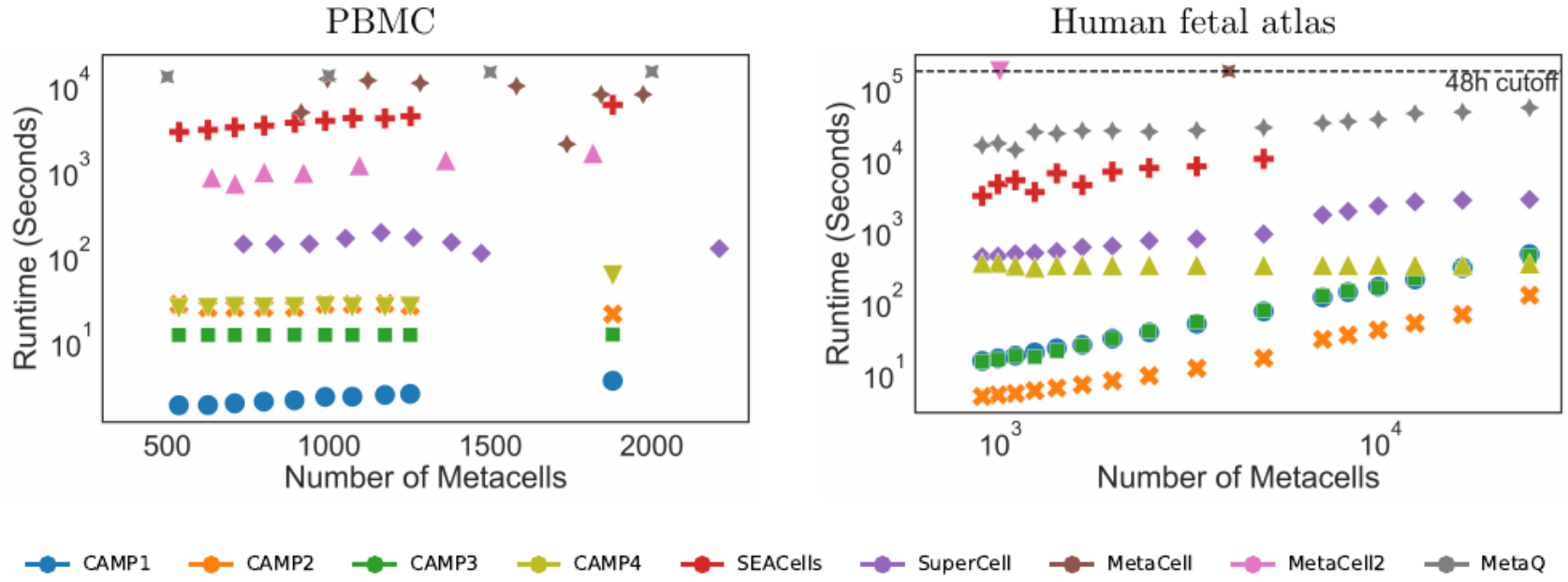


CAMP1 balances compactness and separation

CAMP4 gives the strongest purity / INV profile

SEACells can be competitive in quality, but at much higher cost

Result 2: Fast and memory-efficient metacell construction



< 8 min

CAMP on 504,028 cells

48 h limit

MetaCell and MetaCell2

10,189 s

SEACells at one HFA setting

52,971 s

MetaQ at 25,000 metacells

Minutes rather than hours, without GPU acceleration.

Result 3: compression preserves downstream cell type signal

Train CellTypist on metacell profiles, then predict labels on the original single-cell matrix.

Dataset	CAMP1	CAMP2	CAMP3	CAMP4	SEACells	SuperCell	MetaCell	MetaCell2	MetaQ
PBMC	92.29	88.57	88.52	**93.62**	89.17	89.42	84.44	76.40	83.67
HFA	74.90	73.45	72.55	**75.37**	65.16	51.75	--	--	74.95

Balanced accuracy (%)

PBMC

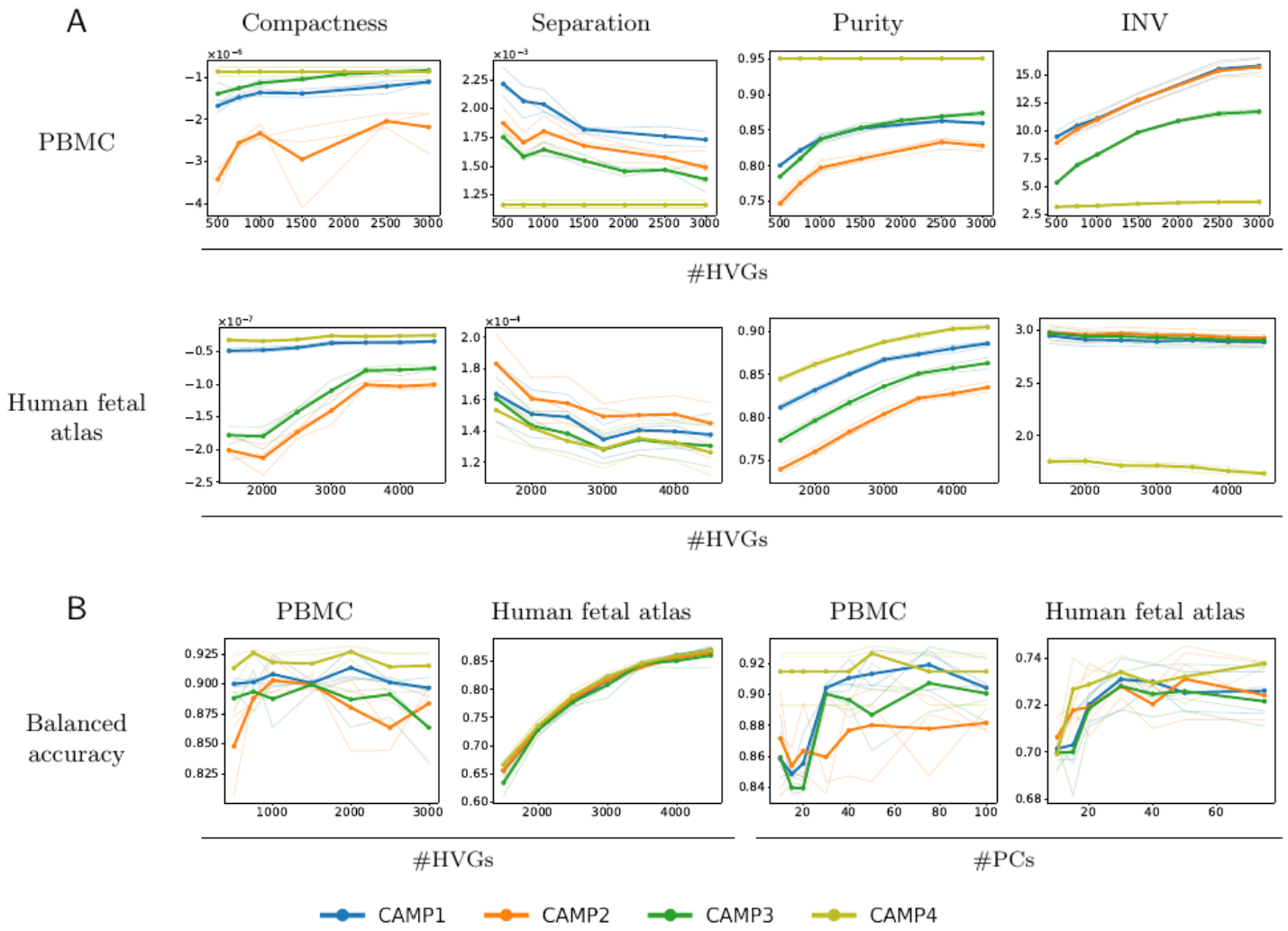
CAMP1/4 are the top two methods; CAMP4 reaches 93.62%.

Human fetal atlas

CAMP4, CAMP1, and MetaQ cluster near the best accuracy, but CAMP is much faster.

The runtime gain does not come at the cost of downstream biological signal.

Additional validation: CAMP is robust to preprocessing choices



Perturbations

HVGs and PCs varied around the default pipeline.

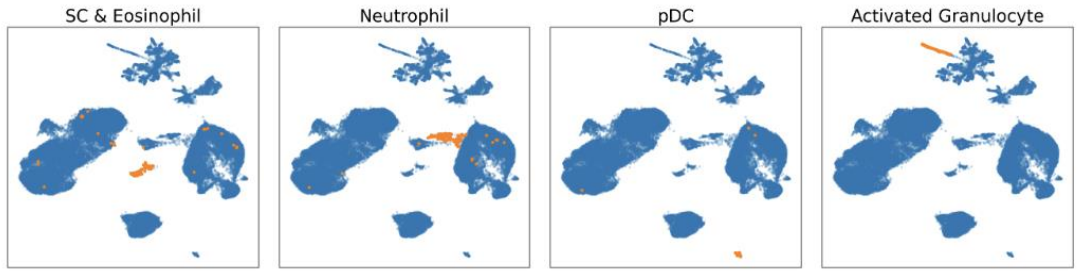
Stable region

Default: 2,000 HVGs and 50 PCs sits in a stable range.

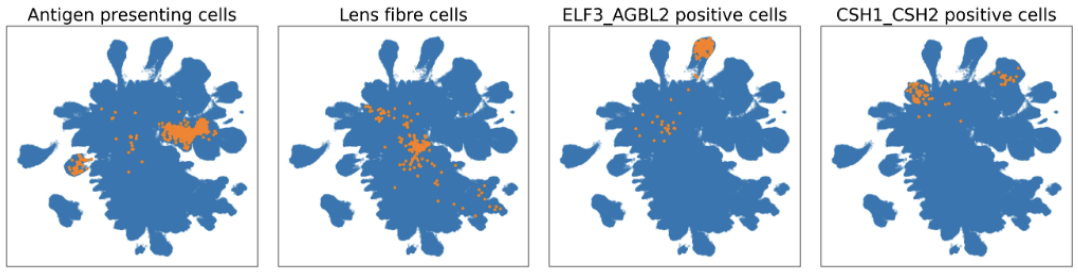
Main claim

CAMP4 is the most stable variant across metrics.

Additional validation: rare cell behaviors across datasets



PBMC



Human fetal atlas

CAMP1

captures broad rare populations (when well-separated)

CAMP4

best for purer, less fragmented rare cell metacells

Human Fetal atlas

harder: rare types are less separated, so no method dominates recall and precision

Take-home: representative cell anchoring is enough for atlas-scale metacells

1 Algorithmic simplification

CAMP replaces expensive global iterative metacell optimization with coreset anchors and direct assignment.

2 Speed at scale

All CAMP variants finish the 504k-cell fetal atlas in minutes on CPU-only hardware.

3 Biological fidelity

Metacells remain compact, pure, and useful for downstream cell-type classification.

Variant guidance: use CAMP1 as the default; use CAMP4 when purity, preprocessing robustness, or low rare cell fragmentation matter most.