B Datasets and codebook

R. Heller · Date: 2026-07-29 · Last changed: 2026-08-31

Generated data

All crease-pattern manifolds are generated, not collected. Every dataset in Parts I–III is reproducible from a pattern specification, a fold angle, a sample size, a noise specification, and a seed.

Precomputed results

`data/processed/benchmark-grid.rds` — the Chapter 10 grid. Provenance: `scripts/run-benchmark-grid.R`.

`data/processed/product-grid.rds` — the product suite behind Chapter 8's irreducible-loss bound, in two arms: a product of two Miura sheets and a product of two Swiss rolls. Provenance: `scripts/run-product-grid.R`.

`data/processed/e1-difficulty.rds`, `e1-controlled.rds`, `e1-armB.rds` — the three arms of E1, which retired the book's original headline claim. Provenance: `scripts/experiment-e1.R`.

`data/processed/part2-sweeps.rds` — the same cell as the main grid, swept over neighbourhood size instead of over noise. This is where Chapters 4, 5 and 6 answer how much of a method's result is the method and how much is the k it was handed. Provenance: `scripts/run-part2-sweeps.R`.

`data/processed/classic-grid.rds` — the Swiss roll, the S-curve and the severed sphere, scored by the same function, the same metrics and the same floor as the crease patterns. Chapter 11 needs it: a comparison in which the two sides are scored by two code paths is not a comparison. Provenance: `scripts/run-classic-grid.R`.

`data/processed/evaluator-audit-pilot.rds` — the narrow pilot that decided whether Chapter 9 had a subject. Provenance: `scripts/run-evaluator-audit.R --pilot`.

`data/processed/seed-budget.rds` — the within-cell spread of every method, and the seed count each one needs before a difference the size of the pre-registered reportable difference (`MIN_REPORTABLE`, fixed in `R/selection.R`) can be detected. This is what standing rule 1's floor is measured against rather than assumed. Provenance: `scripts/measure-seed-budget.R`.

Every one of these carries a provenance block recording the `R/` tree hash, the R version, the BLAS, the package versions, the wall clock and whether the tree was dirty when it ran. `read_run()` refuses one that does not, refuses one written by a `--quick` smoke test, and

refuses one whose contents are not the contents the chapter around the call was written against.

External data

One dataset, used only in Chapter 12.

Fresh 68k PBMCs (Donor A) — peripheral blood mononuclear cells from a single healthy donor, from Zheng et al. (39). Obtained from the 10x Genomics dataset repository, Cell Ranger 1.1.0 processing, Chromium Universal 3′ chemistry.

`scripts/prepare-single-cell.R` downloads the filtered matrix, applies the preprocessing recorded in that script, and writes the artefact Chapter 12 reads. It is not run in CI and not run at render time, on the same provenance convention as the benchmark grid: run it once, commit the result, keep the script as the record of how.

Licence, and what may be redistributed

The dataset is published under **CC BY 4.0**, which permits redistribution with attribution. So the processed matrix *may* ship in this repository, and the constraint on whether it does is size rather than licence — the full filtered matrix is far too large for git, and Chapter 12 does not need it. What is committed is the reduced matrix the chapter actually reads, produced by the script above, with the reduction stated there and in the chapter.

Label provenance, and why the comparison is not circular

This matters more than it looks, and it was an open question until it was checked. Chapter 12's only external evaluation channel is the cell-type labels. If those labels had themselves been derived by clustering one of the embeddings under test, the comparison would be circular and the chapter would be measuring its own assumptions.

They were not. Zheng et al. generated reference profiles by sequencing ten bead-enriched purified subpopulations from the same donor, separately, and assigned each cell in the 68k population the identity of the purified population with which it had the highest Spearman correlation. The labels come from a separate set of experiments, and no clustering or embedding of the matrix under test enters the assignment.

Two caveats belong in the chapter rather than in a footnote, because both bound what a per-label result can mean.

- The assignment is made in gene-expression space — the same space the embeddings are computed from. It is independent of the *embedding*, which is the circularity that would matter here, but it is not an independent measurement of cell identity, and the chapter should not describe it as ground truth. The book reserves that phrase for the crease patterns, where it is earned.
- The authors flag their own assignments as unreliable in specific places: the gap between the highest and second-highest correlation is small for some cells,

cytotoxic T against NK being the example they give, and the purified populations are nested — CD4+ T-helper contains all CD4+ cells — so a hand-written rule reassigns cells to the next-highest correlation within a nested family. Any per-label accuracy in Chapter 12 inherits that, and the pre-registered selection rule of Chapter 10 is what keeps it from being read as a ranking.