1 Introduction

R. Heller · Date: 2026-07-29 · Last changed: 2026-09-07

Why should a crease pattern tell you anything about a gene expression matrix?

1.1 What this chapter answers

It should not, and this book never argues that it does. A gene expression matrix is not a folded sheet of paper, and no chapter here claims a resemblance between them.

What a crease pattern has is a property no real dataset has. Its correct two-dimensional structure is known exactly, in closed form, before a single point is sampled. Rigid folding bends a flat sheet along its creases while the facets stay flat and unstretched, so the folded surface is isometric to the sheet it came from. The flat pattern is not an approximation to the folded surface's intrinsic geometry. It **is** that geometry.

One thing follows, and the book is built on it: an evaluator can be handed the right answer and asked whether it recognises it. A metric that ranks the correct embedding below a plainly wrong one where the correct embedding is *known* has not earned trust where it is not. That is a necessary condition rather than a resemblance, and it is the only transfer to real data this book asserts.

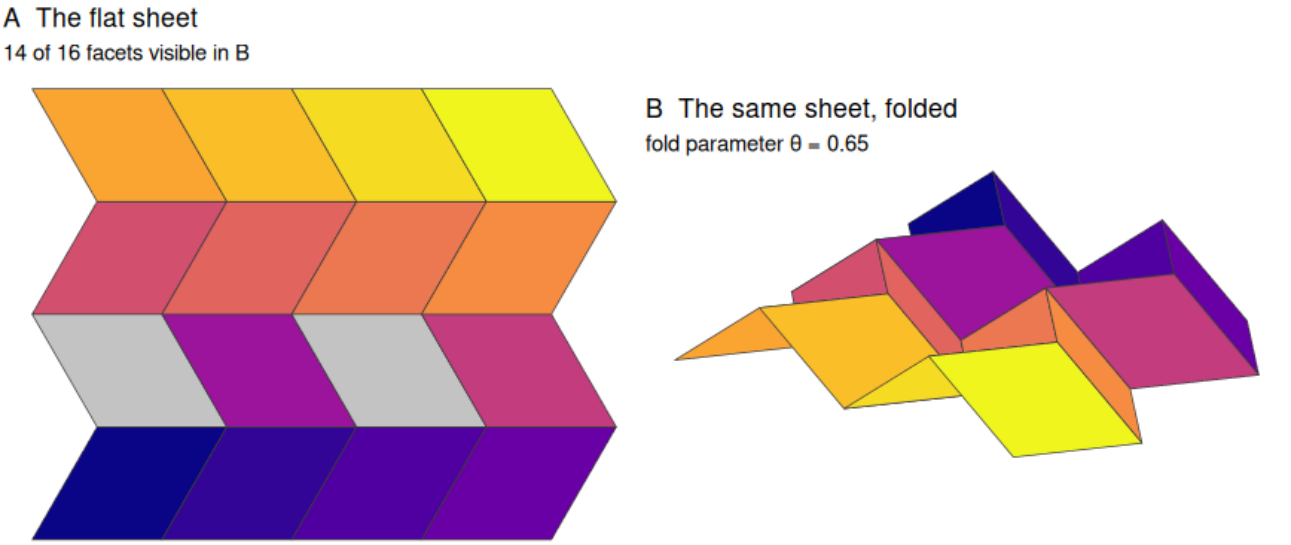


Figure 1.1: A flat crease pattern and the same sheet folded. Turning the folded sheet changes which facets you can see; it does not change the pattern.

The pattern in panel A is the same however panel B is turned. What changes is how much of it any one view exposes — and a dimension-reduction method sees something closer to B than to A. The gap between the two is what the rest of this book measures.

1.2 The argument this book joins

Chari and Pachter argued that the two-dimensional embeddings on which single-cell biology has come to depend distort their data badly enough that conclusions drawn from the pictures are unsafe (1). The reply from Lause, Berens and Kobak was not that the distortion is small. It was that the case against the embeddings rested on metrics that "focused on preservation of distances, where 2D PCA was unsurprisingly the best" (2) — that the disagreement had moved from the embeddings to the instruments used to judge them.

That framing is not this book's, and saying so belongs here rather than in a footnote. It is the published position of one of the two parties to the dispute. What this book adds is not the position but a measurement of it: how far each evaluator sits from an evaluator that has the answer key.

Machado, Behrisch and Telea supply the other half of the reason to look (3). They construct projections that score well on trustworthiness and continuity while displaying structure the data does not contain. That is the false-*positive* direction, and it needs no ground truth: to show that a bad embedding scores high, it is enough to know the embedding is bad. The complement — that the *right* embedding scores low — cannot be shown without an answer key. It is rarely shown for that reason, and it is what Chapter 9 does.

```
audit ← read_run("data/processed/evaluator-audit.rds", "d7982cf7eb3afe7d")
prov  ← attr(audit, "provenance")

# knitr rounds an inline numeric to options(digits), which turns 4e-16 into 0.
# These return character, so a small number survives the trip into prose.
# trimws() because formatC() pads "fg" output to a common width, and a padded
# number lands in the prose with its leading spaces intact.
fmt ← function(x, d = 3L) trimws(formatC(signif(x, d), format = "fg", digits = d))
sci ← function(x, d = 2L) formatC(x, format = "e", digits = d - 1L)

amb ← audit$inversions[audit$inversions$reference == "ambient" &
                       audit$inversions$n == max(prov$n), ]
inverts    ← tapply(amb$inverted, amb$metric, mean)
trust_span ← range(c(amb$truth[amb$metric == "trust"],
                     amb$pca2[amb$metric == "trust"]))

perfect ← audit$scores[audit$scores$reference == "ambient" &
                       audit$scores$candidate == "truth" &
                       audit$scores$k == K_DEFAULT &
                       audit$scores$n == max(prov$n), ]
ceiling ← aggregate(cbind(knn, qnx) ~ theta, perfect, mean)
```

The design is one sentence. Hand each evaluator the exact chart as though it were a submission, alongside a two-component PCA of the same folded cloud, and record which of the two it prefers. Across the audit's ambient-referenced cells at the larger sample size, continuity ranks the wrong answer above the right one in 41.8 per cent of them. Trustworthiness almost never inverts, and only because it barely moves: over the whole sweep it rates every candidate in the audit, the exact chart and the wrong one alike, at

0.984 or above. At the last fold angle, kNN preservation rates the *exact* unfolding 0.58 and Q_{NX} rates it 0.714. A reader shown the first of those would call the embedding a partial failure. It is the answer.

The law that chapter set out to state does not exist, and this introduction will not promise it. The plan of record specified a threshold in the fold angle: rank metrics are invariant to a monotone contraction of distances until neighbour identity changes, so they should invert only past some critical θ. They do invert, and the mechanism is the one predicted. The threshold is not. It slides with the neighbourhood size across most of the available range (Figure 9.6), so there is no threshold in θ alone. What survives is a threshold surface in θ and the neighbourhood *fraction* k/n, and it survives for continuity only. That is a smaller finding than the one budgeted for it, and Chapter 9 reports it as the smaller thing it is.

None of this arbitrates the dispute above. Espadoto and colleagues survey what dimension-reduction evaluations actually report (4); the claim in this book is about those reports, and it requires no position on what a two-dimensional embedding is for.

1.3 The answer key

The construction is one pattern family and it is not new. Miura's report describes the tessellation this book folds as "the isometric transfer of an infinite plane subject to biaxial shortening" (5), which is the whole claim in the words of the person who first wrote it down. Chapter 2 derives the folding rather than citing it, because the derivation is what makes the answer key exact instead of approximate.

```
p ← miura_ori(nx = PATTERN_GRID$miura[["nx"]], ny = PATTERN_GRID$miura[["ny"]])
m ← sample_manifold(p, theta = FIG_INTRO_THETA, n = N_DEFAULT,
                    seed = BENCH_SEEDS[1])

iso  ← facet_isometry_error(p, FIG_INTRO_THETA)
pca2 ← stats::prcomp(m$X, rank. = EMBED_DIM)$x[, seq_len(EMBED_DIM)]
err  ← c(chart = reconstruction_error(m$truth, m$truth),
         pca   = reconstruction_error(pca2, m$truth))
qn   ← c(chart = qnx(m$truth, reference_dist(m, "chart"), K = K_QNX),
         pca   = qnx(pca2,    reference_dist(m, "chart"), K = K_QNX))
```

Two numbers fix what "answer key" means here. Folding preserves every pairwise distance within every facet to 4.4e-16 — floating-point noise on a closed form, not a tolerance anyone chose. And handing the chart back as though it were a submission scores 1.7e-16 against itself under the book's headline error, where a two-component PCA of the same folded cloud scores 0.0775.

θ is a parameter, not an angle. It runs on the unit interval: zero is the flat sheet, one is flat-folded, and the dihedral angle at a crease is derived from it rather than being another name for it. Because it is continuous and the geometry moves monotonically along it, every result in this book is a curve rather than a number — standing rule 2, and the reason a benchmark with one fixed geometry cannot produce the observations Part III reports.

Two headline numbers per cell, and the pair is deliberate. The first is a normalised Procrustes RMSE: the isometry-recovery score, admitting translation, rotation, reflection and one isotropic scale, and charging for everything else. The second is Q_{NX} computed with the exact chart as its reference geometry, which is invariant under any diffeomorphism of the plane and therefore neutral between preserving distances and preserving neighbourhoods. On the sample above the two disagree about PCA: an error of 0.0775 against a Q_{NX} of 0.899, where the exact chart scores 1 on the second and 1.7e-16 on the first. The gap between them, as a function of θ, is a measurement of how much observed error is genuine structural loss and how much is harmless reparameterisation. Reporting only the first would build the book's scoring function out of one side of the dispute it claims to measure.

Folding preserves one kind of distance and destroys the other. Distances along the surface are unchanged, exactly; distances through the surrounding space contract by a factor that depends only on the dihedral angle. Every method in Part II consumes the second kind and is scored against the first. That mismatch is not a flaw in the benchmark, it is the benchmark's subject — and it comes with a cost worth knowing before Part III: a crease has reach exactly zero, so the sampling guarantees usually invoked to justify these methods do not apply to this book's manifolds at all — not weakly, not approximately. Chapter 2 says what replaces them.

1.4 What is claimed, and what is not

Most of the methods under test never claimed what the headline metric measures. Isomap consumes geodesic distances, and its assumption is satisfied exactly on a crease pattern (6); it is the one method in Part II trying to do the thing a Procrustes score scores. t-SNE optimises a divergence between neighbour probability distributions (7), UMAP a cross-entropy on a fuzzy simplicial set (8), and neither claims to reproduce distances at all — a t-SNE embedding's very extent is a function of its perplexity and its iteration count. Local linear embedding, the Laplacian eigenmap and an autoencoder are in the same position. Scoring them against isometric truth and reporting that they lose is not a finding about them; it is a category error dressed as a measurement. The second headline number exists so that a method preserving neighbourhoods without preserving distances can be seen doing it, and each Part II chapter restates the point for the method it covers.

```r
e1  <- read_run("data/processed/e1-controlled.rds", "01d82cfe8f45bb44")
ctl <- local({                          # both difficulty axes inside the joint range
  s <- e1[is.finite(e1$rmse), ]; s$lg <- log(s$gs)
  s[s$lg >= max(tapply(s$lg, s$kind, min)) & s$lg <= min(tapply(s$lg, s$kind, max)) &
    s$np >= max(tapply(s$np, s$kind, min)) & s$np <= min(tapply(s$np, s$kind, max)), ]
})
w <- reshape(ctl[, c("kind", "label", "seed", "method", "rmse")],
             idvar = c("kind", "label", "seed"), timevar = "method",
             direction = "wide")
w$spread <- abs(w$rmse.pca - w$rmse.isomap)
sp    <- tapply(w$spread, w$kind, mean)
ratio <- sp[["classic"]] / sp[["crease"]]

set.seed(11)                            # cluster bootstrap, resampling whole settings
ci <- quantile(replicate(2000, {
  r <- vapply(split(w, w$kind), function(d) {
    l <- sample(unique(d$label), length(unique(d$label)), replace = TRUE)
    mean(unlist(lapply(l, function(x) d$spread[d$label == x])))
  }, numeric(1))
  r[["classic"]] / r[["crease"]]
}), c(0.025, 0.975))

fit   <- lm(spread ~ I(kind == "classic"), data = w)   # cluster-robust by setting
X     <- model.matrix(fit); u <- residuals(fit); bread <- solve(crossprod(X))
meat  <- Reduce(`+`, lapply(unique(w$label), function(l) {
  i <- w$label == l
  tcrossprod(crossprod(X[i, , drop = FALSE], u[i]))
}))
tval  <- coef(fit)[2] / sqrt((bread %*% meat %*% bread)[2, 2])
pval  <- 2 * pt(-abs(tval), df = length(unique(w$label)) - 2)
```

The Swiss roll discriminates better than this benchmark does. The claim this project set out to make was that a crease pattern is a *harder* test than the classical surfaces. Experiment E1 was run before Chapter 11 was drafted, precisely so that the prose could not outlive a refutation, and it refuted it. Matched on both axes of difficulty — branch separation and ambient non-planarity — over 9 crease settings and 5 Swiss rolls, the mean gap between PCA and Isomap is 0.574 on the classical side against 0.113 on the crease side: a ratio of 5.06, with a cluster bootstrap interval over settings of 1.7 to 9.8 and a cluster-robust p of 0.02. Inside the overlap the rankings are identical — Isomap first on both families, the linear methods behind it on both — so nothing flips. Crease patterns are not a harder benchmark. They are an easier one that tells methods apart about five times less well, and a reader choosing between the two for a straight comparison of methods should choose the roll. E1 pre-registered three outcomes and none of them occurred; the replacement claim is in this book because it fits the evidence, not because its antecedent was met.

The difficulty axis stops short of the failure it was expected to reach. Folding corrupts the neighbourhood graph a method consumes, and Chapter 3 measures the corruption growing along the whole sweep. What it does not find, anywhere on the committed grid, is a single graph edge joining facets two or more cells apart — the sheet-to-sheet short circuit that classically breaks Isomap (9) lives past the end of the axis, because the sweep stops before the sheet begins collapsing onto itself. This benchmark measures the graded corruption that precedes that failure. It does not reach the failure.

Exactness alone is not the differentiator. An arc-length-charted Swiss roll also has an exact isometric key, and it is published prior art: Schoeneman and colleagues noticed that the customary answer key is not the isometric one and replaced the Archimedean spiral with a unit-speed one (10). Claiming exact ground truth as the novelty would be wrong. What a crease pattern has that theirs does not is reach exactly zero rather than positive, so that it sits outside the smoothness theory rather than at its edge; intrinsic dimension above two while the chart stays in closed form, which is what makes an irreducible-loss bound computable at all; and a difficulty parameter whose critical value can be derived rather than found by sweeping and looking.

One family, and two documented negative results. Two further pattern families were specified and both were withdrawn on evidence, which Chapter 8 reports as results. A benchmark built on one family cannot support claims about pattern variety, and none are made.

1.5 How the book is organised

Part I builds the object. Chapter 2 derives the folding and proves what it preserves and what it destroys, including a proposition that an ambient-referenced stress must rank the exact unfolding below a two-component PCA — a theorem rather than a simulation. Chapter 3 turns the surface into data: sampling, noise models, and the four measures of difficulty that the θ axis moves.

Part II tests the methods one family at a time, each against the same answer key. Chapter 4 takes the linear projections, Chapter 5 the geodesic methods and the short-circuit failure they are prone to, Chapter 6 the neighbour embeddings, and Chapter 7 the learned folds — where the honest result is that the autoencoder row is declared unavailable rather than quietly dropped.

Part III is the benchmark. Chapter 8 builds the manifolds and the irreducible-loss bound that says how much of an error was unavoidable. Chapter 9 runs the known-answer test on the evaluators themselves. Chapter 10 reads the full grid, and asks whether the benchmark separates methods before asking which method wins. Chapter 11 puts the construction against the Swiss roll, the S-curve and the severed sphere, and reports that it loses on discrimination.

Part IV spends the result. Chapter 12 uses the grid to choose a method and its settings for a real dataset, under a rule fixed in code before the grid existed, and reports what happens.

Two appendices carry the reference material: A fixes every symbol, and B documents the committed artefacts. Every artefact is read with its digest, so a regenerated run is a build failure rather than a silently changed sentence, and the three standing rules the chapters do not restate are set out in *How to use this book*.

Not every reader wants the whole path. Someone who came for the geometry can read Chapters 2 and 3 and stop; someone who came for the numbers can start at Chapter 10 and work backwards through the two chapters it depends on; and someone deciding whether to use this benchmark at all should read Chapter 11 first, because that is where the case against it is made.

1.6 Further reading

Where to start depends on what you came for. For the origami, Hull's *Origametry* (11) is the single best entry point and covers both the classical flat-foldability theorems this book uses as checks. For the methods and for the rank metrics Part III audits, Lee and Verleysen (12) is the textbook treatment of everything Part II fits.

For the dispute, read Chari and Pachter (1) and the reply from Lause, Berens and Kobak (2) in that order; they are short, and between them they define the question this book answers with measurement rather than with argument. Espadoto and colleagues (4) survey what evaluations report in practice, and Machado, Behrisch and Telea (3) is the direct complement to Chapter 9. For the nearest prior benchmark with an exact answer key, Schoeneman and colleagues (10) — read alongside Chapter 11, which measures this construction against theirs and reports that theirs separates methods better.

Miura's own report (5) is worth finding for the sentence quoted above, and each chapter closes with a *Further reading* section of its own that goes deeper on its own subject than this one can.