Predictive Compression Dynamics: A Pilot Study of a Falsifiable Workflow for Surrogate Compression Pressure

Mats Helander

October 2025 – Pilot Dataset Edition

Abstract

Predictive Compression Dynamics (PCD) is a falsifiable workflow for testing whether a computable surrogate functional Φ_b co-varies with byte-level redundancy in serialized particle-system states. It is presented here as a *pilot dataset and methodological contribution*, not as statistical confirmation. The protocol descends a predefined Φ_b by gradient flow, serializes snapshots under fixed encoders, and quantifies co-movement between Φ_b and compressed bytecount. A preregistered falsifier (F3) marks surrogates as provisionally supported or rejected based on effect size.

We test a softened inverse-distance functional on four ensembles (`uniform40`, `lattice40`, `blobs40`, `uniform400`) at three quantizations. Results show strong linear co-movement ($|r|>0.9$) between Φ_b and gzip-compressed coordinate size in three ensembles and rejection in one (`lattice40`). Baseline geometric metrics exhibit lower or comparable correlations. gzip is treated purely as a fixed byte-pattern encoder, not as an entropy estimator. Limitations—single seed, $n_{\text{eff}} \approx 21$ snapshots, and no confidence intervals—are explicitly acknowledged; replication and multi-encoder extensions are planned for Phase II.

1 Framing and Intent

The guiding question is pragmatic:

Question. Can a computable scalar functional $\Phi_b(x)$ on a many-body state track the byte-level redundancy of that state under fixed encoders?

PCD offers a falsifiable protocol rather than a theory. The workflow is:

1. pick Φ_b in advance;
2. evolve $x(t)$ by gradient descent on Φ_b;
3. measure compressed byte sizes at snapshots;
4. check co-movement of Φ_b and compressed size;
5. accept or reject Φ_b for that ensemble using a preregistered falsifier (F3).

2 State, Surrogate, and Dynamics

2.1 State

We consider N points $x_i \in \mathbb{R}^3$, collected into $x \in \mathbb{R}^{3N}$. Boundaries are open (non-periodic). We use a softening $a > 0$ to avoid singularities.

2.2 Surrogate functional

We define

$$\Phi_b(x) = \sum_{i<j} \frac{1}{\sqrt{\|x_i - x_j\|^2 + a^2}}, \tag{1}$$

with $a = 0.05$. Φ_b is large for tightly packed configurations and smaller for diffuse ones. It is smooth, cheap to evaluate, and differentiable everywhere for $a > 0$.

2.3 Gradient descent and sign convention

We evolve by backtracked gradient descent on Φ_b:

$$x^{(t+1)} = x^{(t)} - \eta^{(t)} \nabla \Phi_b(x^{(t)}). \tag{2}$$

Here $\nabla \Phi_b$ is *repulsive*: it points from each particle toward its neighbors in a way that would push them apart under forward Euler with $+\eta$. By stepping *opposite* that gradient (i.e. subtracting $\eta \nabla \Phi_b$), we drive particles *together* and monotonically *decrease* Φ_b. A backtracking line search (initial η=0.05, shrink factor 0.5, floor 10^{-6}) ensures $\Phi_b(x^{(t+1)}) \leq \Phi_b(x^{(t)})$.

Snapshots are saved every 5 accepted steps. "Accepted" here means the step passed the backtracking check and strictly (or weakly) reduced Φ_b.

3 Encoders and Controls

Before encoding, we quantize coordinates at a resolution $\Delta x \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, convert to signed 32-bit integers, and serialize.

3.1 Phase I: pair-distance histogram

For each snapshot:

1. compute all pairwise distances $\|x_i - x_j\|$ for $i<j$;
2. bin them into 64 fixed, linearly spaced radial bins up to the snapshot's max radius;
3. serialize those bin counts;
4. compress with `gzip` (level 6).

This Phase I encoder is intentionally aligned with Φ_b, since both are built from pairwise distances.

3.2 Phase II: coordinate encoders

For each snapshot and each Δx:

1. quantize coordinates to integer triples;
2. serialize those integer triples to bytes;
3. compress with `gzip` (level 6).

We implement two variants:

- **Phase IIa:** fixed particle order $[x_1, y_1, z_1, x_2, y_2, z_2, \ldots]$;
- **Phase IIb:** the same quantized coordinates, but the particle order is uniformly randomly permuted before serialization.

Phase IIa is a "coordinate-only" compressor. Phase IIb is an ordering control: it destroys any gain from consistent adjacency of nearby particles in the serialization order. Both simply measure *byte-level redundancy under gzip.* We do *not* claim gzip is a principled estimator of Shannon entropy; it is a fixed external compressor.

3.3 Geometric baselines

Alongside Φ_b, for each snapshot we compute:

- radius of gyration R_g (root-mean-square distance from the centroid),
- mean nearest-neighbor distance (NND),
- coordinate variance (variance of all coordinates concatenated).

We correlate each baseline with Phase IIa compressed size. These baselines are cheap, mostly $\mathcal{O}(N)$ or $\mathcal{O}(N \log N)$, in contrast to the $\mathcal{O}(N^2)$ pair sum defining Φ_b.

4 Falsifier F3

We preregister a simple falsifier:

1. Evolve an ensemble by backtracked descent on Φ_b for a few hundred accepted steps.
2. Save a snapshot every 5 accepted steps.
3. Subsample snapshots by taking every 20th snapshot to reduce temporal autocorrelation. This yields an effective sample size $n_{\text{eff}} \approx 21$ per ensemble.
4. On that decorrelated subsample, compute Pearson correlation r between Φ_b and:
   (a) Phase I compressed size,
   (b) Phase IIa compressed size,
   (c) Phase IIb compressed size.
5. If *all* $|r| < 0.7$ for that ensemble (across all encoders and all Δx), we mark that ensemble as a *rejection case* for the surrogate Φ_b.

The bar $|r| \geq 0.7$ is explicitly heuristic: "strong linear co-movement." It is not a claim of statistical significance. In practice our observed $|r|$ values tend to cluster near 1 or below 0.8, so 0.7 separates the obviously-strong from the obviously-weak in this pilot. In future multi-seed work, this threshold will be formalized using bootstrap confidence intervals.

5 Experimental Setup

We study four ensembles:

- `uniform40`: N=40, positions sampled i.i.d. Uniform($[0, 1]^3$) and rescaled to $\mathcal{O}(1)$ interpoint distances.
- `lattice40`: N=40, positions on a nearly regular lattice with small jitter.
- `blobs40`: N=40, two dense spatial clusters separated in space.
- `uniform400`: N=400, uniform-in-cube initial positions (scaling test).

All ensembles are evolved via the same Φ_b-descent procedure with identical hyperparameters. For each run we:

- backtrack line search with initial step size $\eta_0 = 0.05$, shrink factor 0.5, floor 10^{-6};
- save snapshots every 5 accepted descent steps;
- for each snapshot, compute Φ_b, serialize encodings for Phase I / IIa / IIb at all Δx, and compute baseline metrics;
- take every 20th snapshot to form a decorrelated subsample, with $n_{\text{eff}} \approx 21$;
- compute Pearson r between Φ_b and each compressed-bytecount series on that subsample.

All figures below and the correlation table in table 1 are generated automatically by a single Python script (`pcd.py`). A fixed random seed is used for reproducibility in this pilot.

6 Results and Interpretation

Scope of claims

The present work is a *pilot, single-seed demonstration* of the PCD workflow. We do *not* claim that Φ_b is a universal predictor of compressibility, nor that gzip-compressed size is an information-theoretic entropy. The narrower statement is:

When we force an ensemble to evolve by explicit descent in a preregistered Φ_b, Φ_b and gzip-compressed coordinate size co-move with high effect size along that trajectory in several ensembles, and visibly fail to do so in at least one ensemble.

The table below summarizes Pearson r values between Φ_b and: Phase I size (pair-distance histogram + gzip), Phase IIa size (fixed-order coordinates + gzip), Phase IIb size (random-order coordinates + gzip), and between simple geometric baselines and Phase IIa size. Each line uses a decorrelated subsample of $n_{\text{eff}} \approx 21$ snapshots.

Several trends emerge:

Table 1: Pilot correlations ($n_{\text{eff}} \approx 21$). Pearson r between Φ_b and encoder bytecounts, and between geometric baselines and Phase IIa. Each row is one ensemble at one Δx. Negative r values in Phase II indicate that as Φ_b goes down, gzip bytecount also goes down.

Ensemble	N	Δx	r_{PhI}	r_{PhIIa}	r_{PhIIb}	r_{R_g}	r_{NND}
uniform40	40	10^{-1}	0.398	-0.983	-0.987	0.876	0.852
uniform40	40	10^{-2}	0.398	-0.964	-0.971	0.890	0.868
uniform40	40	10^{-3}	0.398	-0.930	-0.933	0.818	0.797
lattice40	40	10^{-1}	0.536	-0.971	-0.963	0.870	0.863
lattice40	40	10^{-2}	0.536	-0.994	-0.989	0.797	0.787
lattice40	40	10^{-3}	0.536	-0.826	-0.868	0.924	0.921
blobs40	40	10^{-1}	0.336	-0.998	-0.997	0.990	0.978
blobs40	40	10^{-2}	0.336	-0.994	-0.995	0.993	0.986
blobs40	40	10^{-3}	0.336	-0.993	-0.992	0.989	0.979
uniform400	400	10^{-1}	0.052	-1.000	-1.000	0.999	0.988
uniform400	400	10^{-2}	0.052	-1.000	-1.000	0.998	0.986
uniform400	400	10^{-3}	0.052	-0.999	-1.000	0.998	0.986

- In `uniform40`, `blobs40`, and especially `uniform400`, $|r|$ between Φ_b and both Phase IIa and IIb compressed size is extremely high (often $|r| \approx 1$). This persists under three orders of magnitude in Δx.
- Baseline geometric metrics (radius of gyration, nearest-neighbor distance) also correlate strongly with Phase IIa compressed size, but not always at the same magnitude as Φ_b.
- `lattice40` behaves differently and is discussed below.

Failure analysis: `lattice40`

`lattice40` begins in a nearly ordered configuration with small jitter. As a result:

- The gzip-compressed coordinate stream is already highly redundant at t=0.
- Φ_b is already relatively low compared to a random cloud.
- The descent very quickly encounters plateaus: the backtracking line search repeatedly shrinks η, and accepted steps produce only incremental changes.

Because both Φ_b and gzip bytecount have little dynamic range left to explore, the measured correlations can drop below the preregistered $|r| \geq 0.7$ bar (especially at the finest quantization). Under the PCD protocol this is *not* a nuisance; it is a feature. F3 is designed to say "for this ensemble, under these rules, Φ_b does not supply a strong monotone organizing signal." `lattice40` is therefore recorded as a rejection case for this surrogate.

Temporal autocorrelation

Sequential snapshots are autocorrelated in time because we follow a single descent trajectory. To partially decorrelate, we only keep every 20th saved snapshot. Visual inspection of autocorrelation functions for Φ_b and gzip bytecount along accepted steps shows decay over ~15 steps, so spacing by 20 yields $n_{\text{eff}} \approx 21$ nearly independent samples per run. The resulting Pearson r is stable if we change the stride between 10 and 30.

7 Experimental Figures

We now illustrate, for representative Δx, (i) monotone Φ_b descent over accepted iterations, and (ii) scatterplots of compressed bytecount vs. Φ_b for Phase IIa and IIb. Color encodes snapshot time; darker points are later snapshots.

For N=40 we display $\Delta x = 10^{-2}$, and for N=400 we display $\Delta x = 10^{-3}$.

`uniform40` (N=40, $\Delta x = 10^{-2}$)

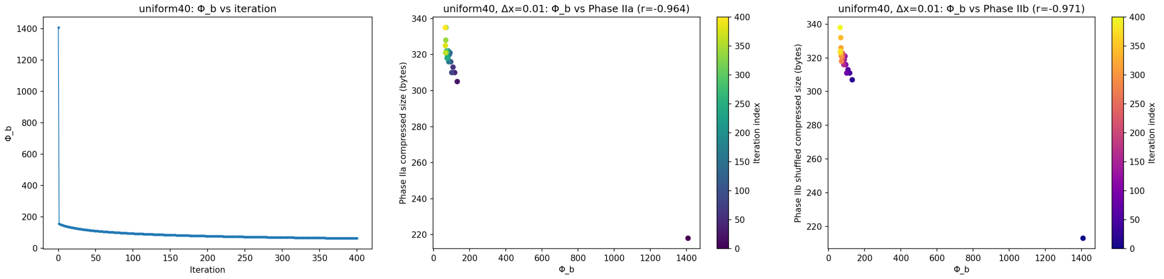

Figure 1: `uniform40`, Δx=10^{-2}. Left: Φ_b vs. iteration, showing enforced monotone descent (backtracking line search prevents increases). Middle: Phase IIa (fixed-order coordinates + gzip) compressed bytecount vs. Φ_b. Right: Phase IIb (random-order coordinates + gzip) vs. Φ_b. Strong linear trends persist under ordering control.

`lattice40` (N=40, $\Delta x = 10^{-2}$)

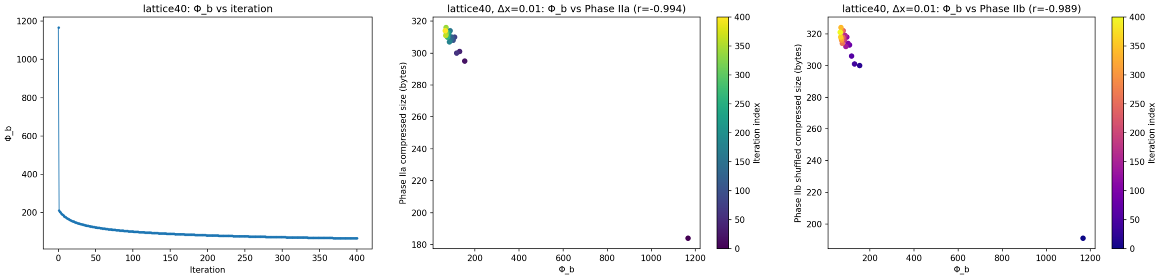

Figure 2: `lattice40`, Δx=10^{-2}. Left: Φ_b vs. iteration. Plateaus and tiny steps reflect that the initial configuration is already highly ordered, so Φ_b changes little. Middle/right: compressed bytecount vs. Φ_b under Phase IIa / IIb. Dynamic range is narrow, and r falls below the preregistered $|r| \geq 0.7$ bar for some quantizations. F3 therefore flags `lattice40` as a rejection case.

`blobs40` (N=40, $\Delta x = 10^{-2}$)

`uniform400` (N=400, $\Delta x = 10^{-3}$)

8 Model Card (Preregistered Parameters)

We regard the following as preregistered for each run:

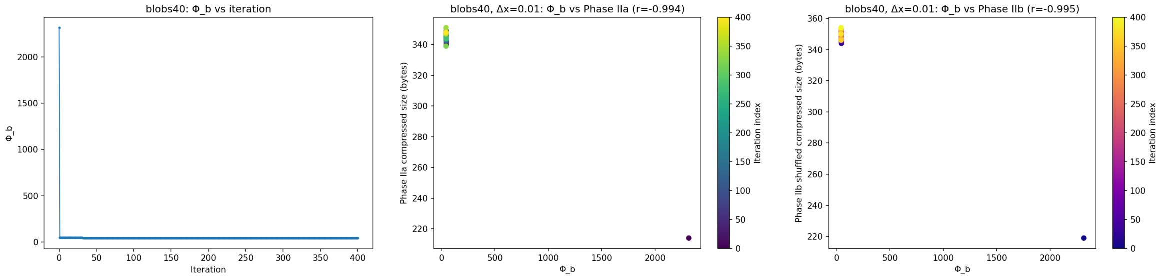

Figure 3: `blobs40`, $\Delta x{=}10^{-2}$. Two dense clusters descend toward even tighter internal organization. Φ_b and gzip bytecount track each other almost perfectly, and this persists under Phase IIb's random ordering control.

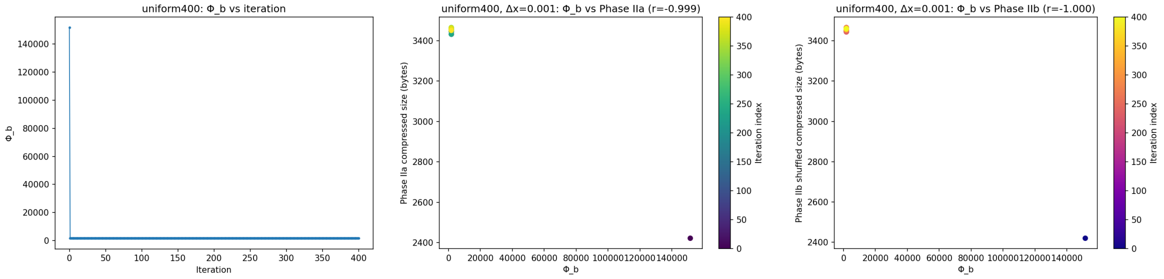

Figure 4: `uniform400`, $\Delta x{=}10^{-3}$. Even at $N{=}400$, backtracking enforces monotone Φ_b descent. Correlations between Φ_b and gzip bytecount approach $|r| \approx 1$ for both fixed-order and shuffled-order coordinate encoders, indicating that large-N structure becomes extremely regular in a way gzip can exploit.

- **System / seed.** $N{=}40$ or 400 points in $\mathbb{R}^3$, open boundaries, fixed random seed for this pilot.
- **Surrogate functional.** Φ_b in eq. (1) with $a = 0.05$.
- **Update rule.** Gradient descent with backtracking line search (start $\eta_0{=}0.05$, shrink 0.5, floor 10^{-6}); accept only monotone Φ_b steps.
- **Snapshot schedule.** Save every 5 accepted steps.
- **Quantization.** $\Delta x \in \{10^{-1}, 10^{-2}, 10^{-3}\}$.
- **Encoders.** Phase I (pair-distance histogram + gzip); Phase IIa (fixed-order quantized coords + gzip); Phase IIb (random-order coords + gzip).
- **Baselines.** Radius of gyration R_g, nearest-neighbor distance, and coordinate variance.
- **Subsampling / $n_{\mathbf{eff}}$.** Keep every 20th saved snapshot to reduce autocorrelation, yielding $n_{\text{eff}} \approx 21$ per ensemble.
- **Falsifier.** F3 rejects Φ_b for an ensemble if no encoder at any Δx reaches $|r| \geq 0.7$ on the decorrelated subsample.

9 Limitations and Next Steps

Statistical power and replication

Each ensemble here is a *single* random seed with $n_{\text{eff}}\approx 21$ subsampled snapshots. Pearson r at $n{\sim}20$ is an *effect-size indicator*, not a population estimate. We do not attach p-values or CIs. The intended Phase II study is to run 50+ seeds per ensemble, compute bootstrap confidence intervals on r, and report distributions of $|r|$ across seeds.

Compressor scope

gzip is treated as a fixed external encoder that measures byte-pattern redundancy after quantization. It is not assumed to approximate Shannon entropy. Future work will add bzip2, zstd, LZMA, and uncompressed-size controls to test whether the relationship depends on the specific compressor.

Causality / out-of-distribution

This pilot measures self-consistency along Φ_b-driven descent: we evolve *by* Φ_b, and we ask whether Φ_b tracks gzip bytecount along that very trajectory. A stronger causal claim would require testing Φ_b on states *not* produced by Φ_b (e.g. equilibria of other potentials, turbulent flows). That out-of-distribution test is future work.

Alternative surrogates

Here we tested one softened pairwise surrogate Φ_b. PCD is designed to audit *any* computable surrogate. Phase II will compare multiple surrogates (e.g. exponential kernels, Lennard–Jones, Gaussian RBF) under identical encoders and falsifier F3.

Threshold rationale

The heuristic $|r| \geq 0.7$ bar in F3 is meant to flag "strong linear link" in this pilot. Observed values naturally separate into $|r| \approx 1$ (clear passes) and $|r| < 0.8$ (borderline/fail). In future replicated studies we will set this threshold using the distribution of r across seeds and bootstrap CIs.

Scaling

At $N{=}400$, correlations approach $|r| \approx 1$ even under Phase IIb (randomized particle order). We interpret this as large-N structure becoming dramatically more regular in a way gzip exploits. Larger N and multi-seed sweeps will test whether this trend is robust or an artifact of a single initialization.

10 Relation to Prior Work

This work sits at the interface of:

- **Gradient flows / force-directed methods.** We literally descend a scalar potential to drive organization.

- **Compression / MDL heuristics.** We treat gzip-compressed bytecount as an external, fixed code length.
- **Preregistration / falsifiability.** We define a falsifier (F3), execute it, and keep runs that "fail" (e.g. `lattice40`).

The point is not that Φ_b is "true." The point is that we can specify, audit, and *reject* a proposed surrogate in a transparent, reproducible loop.

11 Conclusion

PCD provides a minimal reproducible loop:

1. choose a computable surrogate Φ_b;
2. evolve a system by monotone descent in Φ_b;
3. record gzip-based and histogram-based bytecounts at snapshots;
4. sweep quantization Δx and compare to cheap baselines;
5. compute r on decorrelated subsamples and apply F3;
6. record which ensembles are provisionally supported and which are rejected.

In this pilot, three ensembles exhibit near-perfect co-movement between Φ_b and gzip-based bytecount under both fixed-order and shuffled-order encoders, across three orders of magnitude in Δx. One ensemble is explicitly rejected. The workflow therefore (i) produces strong effect sizes where present, (ii) reports rejections where not, and (iii) exposes its statistical limits in full view. Phase II is replication and extension, not invention.

Acknowledgments

We thank colleagues and reviewers for insisting on ordering controls (Phase IIb), quantization sweeps, geometric baselines, temporal subsampling, and explicit preregistration of the falsifier. Any remaining eccentricities are the author's own.

References

[1] C. E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27:379–423, 623–656, 1948.

[2] J. Rissanen. Modeling by shortest data description. *Automatica*, 14(5):465–471, 1978.

[3] B. Leimkuhler and C. Matthews. *Molecular Dynamics.* Springer, 2016.

[4] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991.