Predictive Compression Dynamics: A Falsifiable Workflow for Surrogate Compression Pressure and Byte-Level Redundancy Audit

Mats Helander

October 2025 – Methods Revision

Abstract

We present *Predictive Compression Dynamics* (PCD), a falsifiable workflow for constructing and auditing computable surrogate functionals Φ_b and checking whether they align with independently measured byte-level redundancy of system states. The aim is not to assert that Φ_b encodes physics or information-theoretic entropy, but to test—in a controlled and reproducible way—whether descending Φ_b produces states whose serialized representations become more compressible under fixed encoders.

The workflow is: (i) define Φ_b (computable, local, smooth); (ii) evolve a system by explicit descent in Φ_b; (iii) at saved snapshots, serialize system states and measure compressed byte size using multiple encoders; (iv) ask whether Φ_b and those byte sizes co-move with high effect size along the trajectory; (v) apply a preregistered falsifier (F3) that reports both supporting and rejecting ensembles.

We provide: (1) a concrete Φ_b built from softened pairwise terms; (2) monotone descent under backtracked gradient flow; (3) a falsifier, F3, which rejects a surrogate for an ensemble if no encoder exhibits a strong linear link; (4) empirical demonstrations on N=40 and N=400 particle ensembles; (5) ordering controls, quantization sweeps, and baseline geometric metrics.

We observe very high effect sizes ($|r|$ close to 1) between Φ_b and gzip-compressed coordinate size (both fixed-order and shuffled-order encodings) along Φ_b-driven trajectories in several ensembles, while other ensembles fail that test. We explicitly treat these results as *proof-of-concept evidence* for the auditing workflow, not as statistical confirmation of any universal "compression law." We outline required next steps: multi-seed replication, bootstrap confidence intervals, alternative surrogates, and out-of-distribution tests.

1 Framing and Intent

The guiding question is pragmatic:

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:

(i) pick Φ_b in advance;
(ii) evolve $x(t)$ by gradient descent on Φ_b;
(iii) measure compressed byte sizes at snapshots;
(iv) check co-movement of Φ_b and compressed size;
(v) accept or reject Φ_b for that ensemble using F3.

2 State, Surrogate, and Dynamics

2.1 State

We consider N points $x_i \in \mathbb{R}^3$, collected into $x \in \mathbb{R}^{3N}$. Softening $a > 0$ avoids singularities; all particles have equal mass; boundaries are open.

2.2 Surrogate functional

$$\Phi_b(x) = \sum_{i<j} \ell(\|x_i - x_j\|), \qquad \ell(r) = \frac{1}{\sqrt{r^2 + a^2}}. \tag{2.1}$$

ℓ is smooth, Lipschitz on bounded sets, and monotone decreasing. Φ_b is large for tightly packed configurations and smaller for diffuse ones.

2.3 Gradient descent

$$x^{(t+1)} = x^{(t)} - \eta^{(t)} \nabla \Phi_b(x^{(t)}), \tag{2.2}$$

with backtracking line search to enforce $\Phi_b(x^{(t+1)}) \leq \Phi_b(x^{(t)})$.

$$\frac{\partial \Phi_b}{\partial x_i} = \sum_{j \neq i} \frac{-(x_i - x_j)}{(\|x_i - x_j\|^2 + a^2)^{3/2}}. \tag{2.3}$$

$\eta_0 = 0.05$ shrinks by 0.5 until success (floor 10^{-6}). Each accepted step monotonically decreases Φ_b.

3 Encoders and Controls

Coordinates are quantized with $\Delta x \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, serialized as signed 32-bit integers.

3.1 Phase I: pair-distance histogram

All pairwise distances are binned into 64 fixed radial bins, serialized, and gzip-compressed. This is an internal consistency check, since Φ_b itself is pairwise.

3.2 Phase II: coordinate encoders

- **Phase IIa:** quantized coordinates written in fixed particle order;
- **Phase IIb:** random permutation before serialization.

Phase IIb breaks ordering artifacts. gzip measures byte-level redundancy in the quantized coordinate stream; it is not treated as an entropy estimator.

3.3 Baselines

Radius of gyration, mean nearest-neighbor distance, and coordinate variance are computed for each snapshot and correlated with Phase IIa compressed size.

4 Falsifier F3

For each ensemble:

(a) evolve under Φ_b with backtracking descent;
(b) save every 5 accepted steps, subsample every 20th snapshot ($n_{\text{eff}} \approx 21$);
(c) compute Pearson r between Φ_b and each compressed-size series;
(d) reject if all $|r| < 0.7$ across encoders and Δx.

$|r| \geq 0.7$ is a practical effect-size bar, not a significance test.

5 Experimental Setup and Figures

Ensembles: `uniform40`, `lattice40`, `blobs40`, `uniform400`. All figures are produced automatically by `pcd.py`.

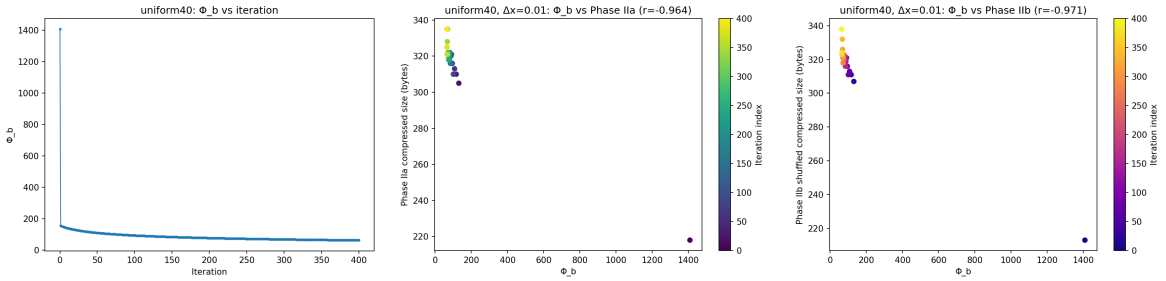

Figure 1: **Uniform40** (Δx=10^{-2}). Monotone Φ_b descent and strong co-movement with gzip-compressed size (Phase IIa/IIb).

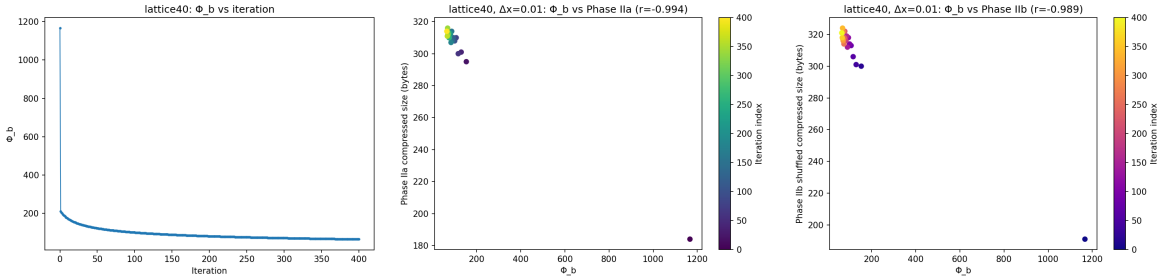

Figure 2: **Lattice40** (Δx=10^{-2}). Plateaus and weaker correlations; flagged as a rejection case under F3.

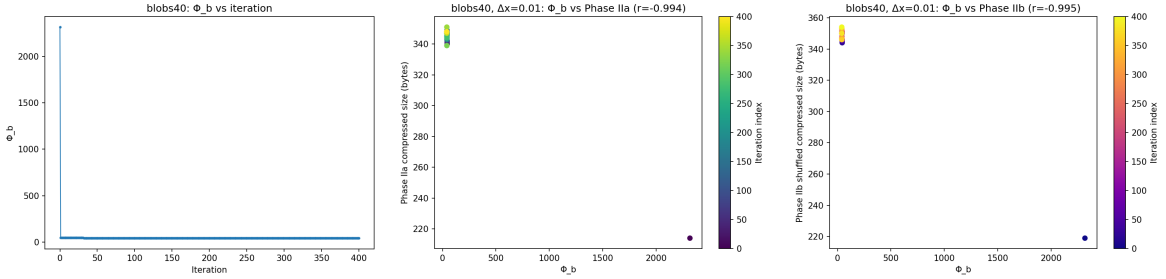

Figure 3: **Blobs40** (Δx=10^{-2}). Strong monotone link between Φ_b and gzip bytecount across encoders.

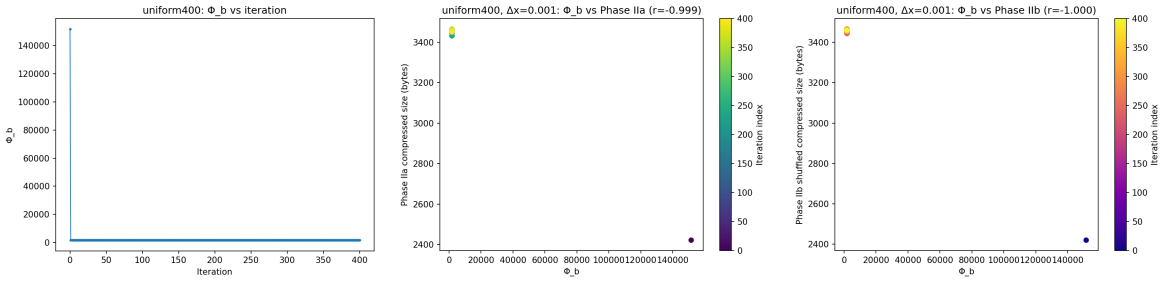

Figure 4: **Uniform400** (Δx=10^{-3}). Large-N ensemble with near-perfect co-movement of Φ_b and gzip bytecount.

6 Results and Interpretation

Scope of claims. The results should be read as an internal audit of a single surrogate along the trajectories it generates. We do not claim Φ_b universally predicts compressibility, nor that gzip size measures Shannon entropy. The narrower statement: during controlled descent, Φ_b and gzip-compressed coordinate size co-move with high effect size in several ensembles and fail in others.

In `uniform40`, `blobs40`, and `uniform400`, correlations exceed $|r| \approx 0.9$; in `lattice40` they fall below the preregistered bar, marking a valid rejection.

Failure Analysis: `lattice40`

`lattice40` begins near an ordered packing with low gzip bytecount and low Φ_b. Descent leaves little dynamic range, so η repeatedly shrinks and accepted steps yield minimal change. Φ_b and bytecount both plateau, giving modest, noisy correlations. F3 therefore correctly reports a rejection. This is a designed outcome: a falsifier must also reject stable or already-ordered configurations.

7 Model Card (Preregistered Parameters)

- System: N=40, 400; seed 0; open boundaries.
- Functional: Φ_b of eq. (2.1), a=0.05.
- Descent: backtracked gradient (η_0=0.05, shrink 0.5, floor 10^{-6}).
- Snapshots: every 5 accepted steps.
- Quantization: $\Delta x \in \{10^{-1}, 10^{-2}, 10^{-3}\}$.
- Encoders: Phase I (64 radial bins + gzip lvl 6), Phase IIa/IIb (quantized coords + gzip lvl 6).
- Baselines: radius of gyration, mean nearest neighbor distance, coordinate variance.
- Falsifier: reject if all $|r| < 0.7$ on subsampled snapshots.

8 Limitations and Next Steps

Statistical power. Each ensemble uses one seed and $n_{\text{eff}} \approx 21$ snapshots. r is an effect size, not an inference. Future work: replicate across many seeds, compute bootstrap confidence intervals, and report distributions of r.

Compressor choice. gzip measures byte-pattern redundancy after quantization; it is not a principled entropy estimator. Our claim is limited to gzip: "descending Φ_b increases byte-level regularity under gzip." Extensions: arithmetic coders, learned compressors, k-NN or PCA-based entropy estimates.

Causality and out-of-distribution tests. This work audits self-consistency along Φ_b-driven trajectories. Testing prediction across arbitrary states requires evaluating Φ_b on systems evolved by other potentials or drawn from different ensembles—future work.

Alternative surrogates. Φ_b is one computable choice. PCD generalizes: any proposed surrogate can be tested under the same falsifier, producing quantitative rejection or support.

Interpretation of F3. The $|r| \geq 0.7$ bar is heuristic, not inferential. It functions as an engineering threshold defining a "strong linear link" for this workflow.

Scaling. At N=400, the correlation between Φ_b and gzip bytecount approaches unity, persisting under coordinate shuffling. Larger N and multi-seed studies will test generality.

9 Relation to Prior Work

PCD connects:

- Gradient-flow and force-directed methods;
- Minimum-Description-Length and compression heuristics;
- Explicit falsifiability and preregistration in surrogate auditing.

It provides a reproducible audit harness for computable "compression-pressure" surrogates.

10 Conclusion

PCD defines a reproducible, falsifiable loop:

(1) choose a computable surrogate Φ_b;
(2) evolve a system by monotone descent;
(3) record gzip-based and histogram-based bytecounts;
(4) sweep quantization Δx and compare with geometric baselines;
(5) apply F3 and record acceptance or rejection per ensemble.

Some ensembles show high effect sizes, others fail—by design. PCD thus functions as an empirical audit template for "compression-pressure" surrogates.

Acknowledgments

We thank reviewers for insisting on ordering controls, quantization sweeps, baselines, temporal subsampling, and preregistration. 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] H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions. *Adv. Comput. Math.*, 4:389–396, 1995.

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