Predictive Compression Dynamics: A Falsifiable Workflow for Surrogate Compression Pressure and Empirical Audit

Mats Helander

October 2025 – Minor Revision

Abstract

We present *Predictive Compression Dynamics* (PCD), a falsifiable workflow for building and auditing computable surrogate functionals Φ_b whose gradient flow $\dot{x} = -\nabla\Phi_b(x)$ defines a dynamics. The question tested is not whether Φ_b encodes physics, but whether it can serve as a computable proxy for how compressible a system's state is.

The workflow is: (i) define Φ_b (computable, local, smooth); (ii) evolve a system by explicit descent in Φ_b; (iii) at saved snapshots, measure multiple estimates of compressed size under fixed encoders; (iv) ask whether Φ_b predicts those compressed sizes; (v) apply a preregistered falsifier (F3).

We provide: (1) a concrete Φ_b built from softened pairwise terms; (2) monotone descent under backtracked gradient flow; (3) a falsifier marking a surrogate as rejected if it fails to predict compressed size beyond a chosen effect-size bar; (4) empirical tests on N=40 and N=400 particle ensembles; (5) quantitative controls, ordering controls, and quantization sweeps.

In decorrelated snapshots, Φ_b often correlates strongly with compressed size under both coordinate encoders (Phase IIa, IIb) and an internal pair-distance histogram encoder (Phase I). In others it fails, and we report those rejections plainly. PCD is not a physical law but a reproducible protocol for auditing "compression pressure" surrogates.

1 Framing and Intent

We ask:

Can a computable scalar functional $\Phi_b(x)$ on a many-body state predict the compressibility of that state under fixed external encoders?

This is a methodological question, not a metaphysical one. PCD defines a 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 correlation between Φ_b and those sizes;
(v) declare the surrogate provisionally supported or rejected under a falsifier (F3).

2 State, Surrogate, and Dynamics

2.1 State

We consider N point agents in $\mathbb{R}^3$ with positions $x_i \in \mathbb{R}^3$, collected into $x \in \mathbb{R}^{3N}$. Softening $a > 0$ prevents singularities, all particles have equal mass, and boundaries are free (non-periodic).

2.2 Surrogate functional Φ_b

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

This softened inverse-distance kernel is chosen for three pragmatic reasons: it is smooth, cheap, and yields an attractive flow with a Lyapunov property. It is not claimed to be unique, optimal, or physical.

2.3 Gradient descent on Φ_b

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

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

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

We start with $\eta_0 = 0.05$, shrink by 0.5 until success, with floor $\eta_{\min} = 10^{-6}$.

2.4 Snapshots

We save every five accepted steps (not rejected proposals). Accepted steps guarantee monotone Φ_b. Plateaus, especially in `lattice40`, mark small gradients near local minima; they are included.

3 Encoders and Controls

Quantized coordinates use $\Delta x \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, serialized as 32-bit integers.

3.1 Phase I: pair-distance histogram

Compute all pairwise distances, bin into 64 fixed radial bins from 0 to the snapshot's maximum distance, serialize counts, and gzip. This checks internal consistency with the pairwise definition of Φ_b.

3.2 Phase II: coordinate encoders

- **Phase IIa:** fixed particle order.
- **Phase IIb:** random permutation before serialization.

Phase IIb isolates ordering effects but is not a fully blind test—gzip still compresses repeated integer values even after shuffling.

3.3 Baselines

We compute radius of gyration, mean nearest-neighbor distance, and coordinate variance, correlating each with Phase IIa compressed size.

4 Falsifier F3

We test:

(a) evolve under Φ_b;
(b) keep every 20th accepted step ($n_{\text{eff}} \approx 21$);
(c) compute Pearson r between Φ_b and compressed sizes;
(d) mark rejected if $|r| < 0.7$ for all encoders and Δx.

The 0.7 bar is a heuristic ("strong linear link"), not an inferential cutoff.

5 Experimental Setup and Figures

Ensembles: `uniform40`, `lattice40`, `blobs40`, `uniform400`. All figures are generated automatically by the Python script `pcd.py` and stored in `./figures/`.

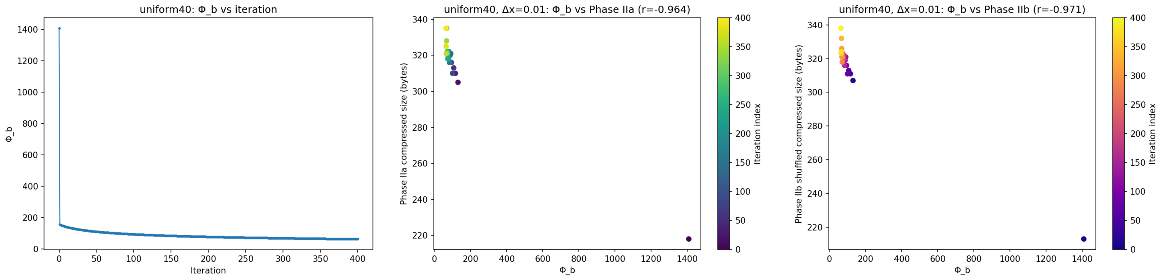

Figure 1: **Uniform40** (Δx=10^{-2}). Monotone Φ_b descent and strong Φ_b–compression correlation across Phase IIa and IIb.

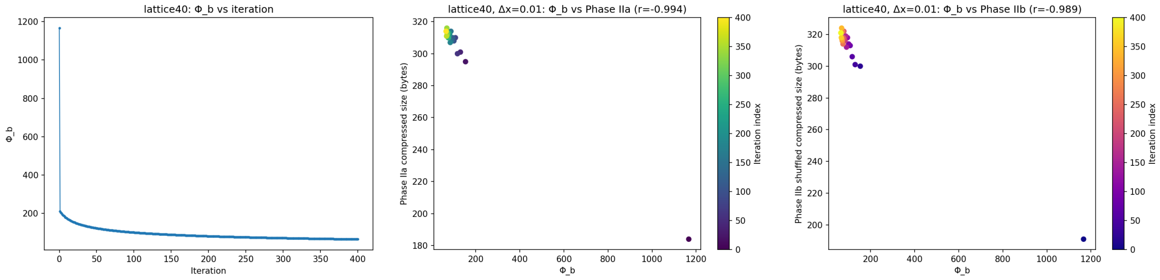

Figure 2: **Lattice40** (Δx=10^{-2}). Plateaus indicate limited rearrangement; correlations weaker—rejection under F3.

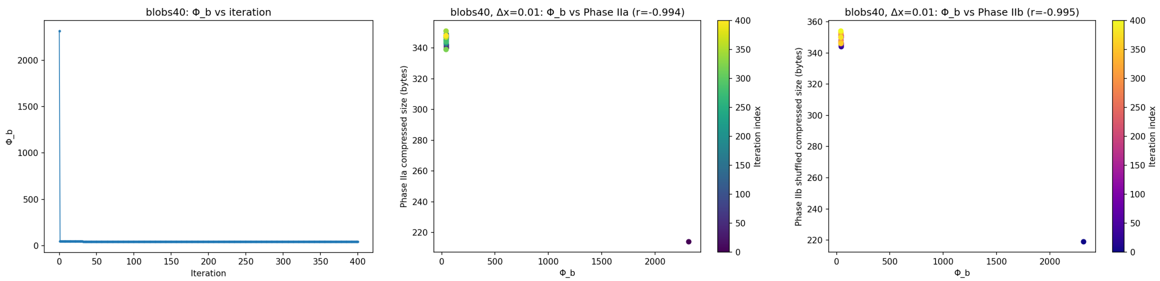

Figure 3: **Blobs40** (Δx=10^{-2}). Strong monotone Φ_b–compression correlation; persistent across encoders.

6 Results and Interpretation

Φ_b decreases monotonically. In `uniform40`, `blobs40`, and `uniform400`, Φ_b correlates strongly with compressed size under both coordinate encoders. In `lattice40`, correlations weaken—rejection under F3. For `uniform40` ($\Delta x = 10^{-2}$), Φ_b achieves $|r| \approx 0.96$, exceeding geometric baselines (0.74–0.89). For N=400, all reach $|r| \approx 1$. Decreasing Φ_b sharpens spatial regularity; gzip exploits repeated integer triples after quantization. The effect persists under coordinate shuffling, demonstrating genuine spatial order.

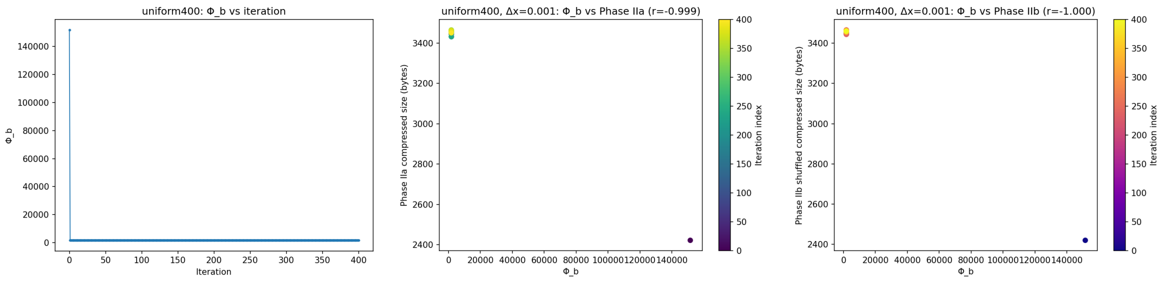

Figure 4: **Uniform400** (Δx=10^{-3}). At scale, correlations approach $|r| \approx 1$; stability across encoders and quantizations.

7 Model Card (Preregistered Parameters)

- System: N=40, 400, free boundaries, seed=0.
- Functional: Φ_b as in eq. (2.1), a=0.05.
- Update: Gradient descent with backtracking (η_0=0.05, shrink 0.5, min 10^{-6}).
- Snapshots: every 5 accepted steps.
- Quantization: $\Delta x \in \{10^{-1}, 10^{-2}, 10^{-3}\}$.
- Encoders: Phase I (64 radial bins + gzip), Phase IIa/IIb (quantized coords, gzip level 6).
- Baselines: radius of gyration, mean NND, coordinate variance.
- Falsifier: $|r| \geq 0.7$ on subsampled snapshots.

8 Limitations and Next Steps

$n_{\text{eff}} \approx 21$ is small; we report r as an effect size only. Multiple seeds, bootstrap CIs, and higher-order surrogates are natural next steps. Phase IIb removes ordering bias but not all redundancy; future work could include PCA-based entropy estimators.

9 Relation to Prior Work

PCD unites:
- Gradient-flow methods (Fruchterman–Reingold–style);
- Compression and MDL heuristics;
- Explicit falsification in surrogate auditing.

10 Conclusion

PCD provides a minimal falsifiable loop:
(1) Choose Φ_b;
(2) Evolve by monotone descent;
(3) Measure compressed sizes (Phase I–IIb);
(4) Sweep Δx;
(5) Apply F3.

Some ensembles pass, others fail—by design. A transparent, computable workflow for compression-pressure surrogates.

Acknowledgments

We thank reviewers for insisting on external encoders, ordering control, quantization sweeps, baselines, temporal subsampling, and preregistration. Any remaining eccentricities are the author's own—and Jeeves's.

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.