

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

Mats Helander

October 24, 2025

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 claim tested is not that Φ_b encodes physics, but that Φ_b 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 (codelength) under fixed encoders; (iv) ask whether Φ_b predicts those compressed sizes; (v) apply a preregistered falsifier.

We provide: (1) a concrete Φ_b built from softened pairwise terms; (2) existence and Lyapunov-style descent under backtracked gradient flow; (3) a preregistered falsifier, F3, which marks a surrogate as rejected if it fails to predict compressed size beyond a chosen effect-size bar; (4) empirical tests on ensembles of $N=40$ and $N=400$ particles; (5) quantitative controls requested by review.

In particular we report, for decorrelated snapshots: Pearson correlations between Φ_b and compressed snapshot size under (i) a pair-distance histogram encoder (Phase I); (ii) a raw-coordinate gzip encoder with fixed particle ordering (Phase IIa); and (iii) the same coordinate encoder after randomly permuting particle order (Phase IIb), which breaks ordering artifacts. We sweep quantization resolutions $\Delta x \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, and we compare Φ_b against naive geometric baselines (radius of gyration, nearest-neighbor distance, coordinate variance).

In several ensembles, including $N=400$, Φ_b shows near-perfect monotonic correlation with compressed size under both Phase IIa and the shuffled-order Phase IIb, and matches or exceeds those baselines. In other ensembles this link is weaker and fails our preregistered bar. We emphasize scope: PCD is not a physical law, but a falsifiable protocol for auditing “compression pressure” surrogates.

1 Framing and Intent

The question is deliberately modest:

Question. Can we build a computable scalar functional $\Phi_b(x)$ on the state x of a many-body system such that (a) Φ_b can be monotonically decreased by a local update rule, and (b) Φ_b numerically correlates with how concisely that state can be encoded?

The goal is methodological: we want to know whether a particular, fully specified computable surrogate can act as a stand-in for “compression pressure,” and how to test (and possibly fail) that claim in a controlled way. The particle systems are a testbed for running that audit.

The structure of the workflow is:

- (i) pick a surrogate Φ_b , fully specified in advance;
- (ii) evolve $x(t)$ by (numerically stabilized) gradient descent on Φ_b ;
- (iii) at saved times t_k , measure multiple compressed byte sizes of $x(t_k)$ using fixed encoders;

- (iv) check whether $\Phi_b(x(t_k))$ predicts those compressed sizes;
- (v) declare the surrogate *provisionally supported* or *rejected for that ensemble* using a preregistered criterion.

This paper contributes: (1) the exact surrogate Φ_b we study; (2) the explicit encoders we use; (3) a falsifier F3; (4) experiments at multiple N ; (5) baseline geometric controls, coordinate-ordering controls, and a quantization sweep; (6) a model card template intended for preregistration.

2 State, Surrogate, and Dynamics

2.1 State

We consider N point agents in \mathbb{R}^3 , with positions $x_i \in \mathbb{R}^3$, $i = 1, \dots, N$. We collect all positions into $x \in \mathbb{R}^{3N}$. All tests below evolve closed systems of this form.

We fix a softening length $a > 0$ and equal positive weights $m_i \equiv 1$ for all agents. (Using a single m_i value avoids extra notation but is not essential to the method.)

2.2 Surrogate functional Φ_b

We study a concrete, computable, pairwise surrogate,

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

The softening $a > 0$ prevents singularities at $r=0$ and makes ℓ Lipschitz on bounded sets.

Intuition: $\ell(r)$ is large when points are close. Thus Φ_b is large for tightly packed / clustered configurations and smaller for diffuse ones. When we push Φ_b down, particles tend to coalesce into more spatially regular structures. Empirically, such regular structure yields more predictable coordinate patterns after quantization, which an off-the-shelf compressor (gzip) can exploit.

We do not claim Φ_b is unique or optimal. Φ_b is simply (i) computable from pairwise distances, (ii) cheap to evaluate, and (iii) endowed with a well-defined gradient.

2.3 Dynamics: gradient descent on Φ_b

We evolve $x(t)$ by discrete-time gradient flow on Φ_b :

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

with a backtracking line search on the step size $\eta^{(t)}$ to ensure $\Phi_b(x^{(t+1)}) \leq \Phi_b(x^{(t)})$.

For particle i , by the chain rule,

$$\frac{\partial}{\partial x_i} \ell(\|x_i - x_j\|) = -\frac{(x_i - x_j)}{(\|x_i - x_j\|^2 + a^2)^{3/2}}, \quad (2.3)$$

so

$$\frac{\partial \Phi_b}{\partial x_i} = \sum_{j \neq i} \frac{-(x_i - x_j)}{(\|x_i - x_j\|^2 + a^2)^{3/2}}. \quad (2.4)$$

Thus the update takes the form

$$x_i^{(t+1)} = x_i^{(t)} + \eta^{(t)} \sum_{j \neq i} \frac{(x_j - x_i)}{(\|x_i - x_j\|^2 + a^2)^{3/2}}. \quad (2.5)$$

Because Φ_b is C^1 for $a > 0$, the *continuous* flow $\dot{x} = -\nabla \Phi_b$ obeys

$$\frac{d}{dt} \Phi_b(x(t)) = \nabla \Phi_b(x(t)) \cdot \dot{x}(t) = -\|\nabla \Phi_b(x(t))\|^2 \leq 0. \quad (2.6)$$

This makes Φ_b a Lyapunov-like function for that continuous flow.

In discrete time, naive fixed-step Euler can overshoot and break monotonicity. We therefore use backtracking line search: if a candidate step would increase Φ_b , we shrink $\eta^{(t)}$ by a constant factor (e.g. 0.5) until it does not, down to a floor. This enforces $\Phi_b(x^{(t+1)}) \leq \Phi_b(x^{(t)})$ in practice. The strictly nonincreasing Φ_b curves in [figs. 1, 3, 5 and 7](#) use this rule.

2.4 Snapshots

We save snapshots $x(t_k)$ every fixed number of accepted steps (e.g. every 5 accepted gradient steps). For each snapshot we record:

- the state $x(t_k)$,
- the scalar surrogate $\Phi_b(x(t_k))$,
- compressed sizes (bytes) under the encoders in [section 3](#),
- the iteration index t_k .

These snapshots form the dataset for evaluating whether Φ_b predicts compressed size.

3 Encoders and Controls

We compare Φ_b against multiple compression-based codelength estimates for each snapshot. This gives us several distinct “targets” to predict.

We will refer to these as **Phase I**, **Phase IIa**, and **Phase IIb**, plus baseline geometric controls.

3.1 Quantization and serialization

Before encoding, we quantize particle coordinates onto a lattice of spacing Δx . We then store these quantized coordinates as signed 32-bit integers so that a lossless compressor like **gzip** can act on a byte sequence.

To test sensitivity to discretization, we repeat the analysis with $\Delta x \in \{10^{-1}, 10^{-2}, 10^{-3}\}$. We report correlations separately for each Δx in [section 6](#).

3.2 Phase I encoder (pair-distance histogram)

Phase I intentionally reflects the same kind of pairwise structure that Φ_b was built to see.

For each snapshot:

- compute all interparticle distances $\|x_i - x_j\|$ for $i < j$;
- bin these distances into a fixed, preregistered set of radial bins;
- serialize the histogram counts deterministically to bytes;
- compress that byte string with **gzip** (Lempel–Ziv).

The resulting compressed byte length is our “Phase I” compressed size for that snapshot. Intuition: if a configuration has a stereotyped distance distribution (e.g. crystalline or highly clustered), the histogram is regular and should compress well.

Phase I is an *internal consistency* check: if Φ_b is supposed to reflect pairwise distances, then it should track the codelength of an encoding that also focuses on pairwise distances.

3.3 Phase II encoders (raw coordinates with and without ordering)

Phase II is meant to act as an *external* compressor: we do not feed it pairwise distances at all. We just give it coordinates.

For each snapshot and each Δx :

- Quantize coordinates at resolution Δx : $q_i = \text{round}(x_i / \Delta x)$;
- serialize those integer triples to bytes and gzip.

We implement two versions:

- **Phase IIa (fixed ordering).** We write out quantized coordinates in a deterministic particle order $[x_1, y_1, z_1, x_2, y_2, z_2, \dots]$, then gzip.
- **Phase IIb (shuffled-order control).** Before serialization we randomly permute the particle order, then gzip. This breaks any stable run-length / delta structure that depends on keeping neighbors next to each other in a consistent order.

Why do we need IIb? gzip is a Lempel–Ziv compressor and will exploit repeated byte patterns. Clustered particles produce repeated coordinate patterns after quantization, so Phase IIa is *not* truly “blind.” Phase IIb tests how much of the signal survives if we deliberately scramble particle ordering. If Φ_b still correlates with Phase IIb compressed size, that supports the claim that Φ_b is tracking spatial regularity, not just a serialization artifact.

3.4 Baseline geometric controls

A core objection is that Φ_b could just be a fancy monotone of “clumpiness,” and that any simple clustering metric would have correlated with compressed size.

To test this, for each saved snapshot we also compute:

- **Radius of gyration** (root mean square distance from the centroid),
- **Mean nearest-neighbor distance**,
- **Coordinate variance** (variance of coordinates flattened).

We then correlate each of these baselines with the Phase IIa compressed size (bytes). Comparing those r values to the Φ_b vs Phase IIa correlation tells us whether Φ_b offers explanatory power beyond cheap geometric scalars.

4 Falsifier F3

We preregister a falsifier:

F3 (surrogate–compressibility link). For a fixed surrogate Φ_b , fixed encoders (Phase I, Phase IIa, Phase IIb), and fixed quantization scales Δx , we:

- evolve an ensemble by backtracked gradient descent on Φ_b ;
- save snapshots every few accepted steps;
- subsample snapshots at a fixed stride (e.g. keep every 20th accepted step) to reduce temporal autocorrelation; call the resulting sample size n_{eff} ;
- compute Pearson r between Φ_b and each compressed-size series on that subsample.

If *none* of the compressed-size measures reaches a preregistered effect-size bar (e.g. $|r| \geq 0.7$) for that ensemble at any Δx , we mark that ensemble as a *rejection case* for this surrogate.

We deliberately *do not* attach classical p -values, because snapshots along a single trajectory are autocorrelated and formal time-series corrections (e.g. block bootstrap, AR(1) modeling) are outside this first pass. Instead, r is treated as an effect size, and n_{eff} is reported to show how many effectively decorrelated snapshots were used.

Philosophically, F3 is not an attempt to falsify a universal physical law. It is an engineering-style audit: “Does this specific computable scalar Φ_b act as a useful proxy for compressibility in this ensemble, under this protocol, with these encoders?” When an ensemble fails F3, we simply say that Φ_b is *rejected for that ensemble*.

5 Experimental Setup

All experiments follow the same template:

(E1) **Initial ensembles.** We generate several ensembles of particles:

- **uniform40:** $N=40$, positions sampled i.i.d. from $\text{Uniform}([0, 1]^3)$.
- **lattice40:** $N=40$, positions on a nearly regular lattice with small Gaussian jitter.
- **blobs40:** $N=40$, two dense clusters separated in space.
- **uniform400:** $N=400$, positions sampled uniformly in a cube (to test scaling).

All positions are rescaled to $\mathcal{O}(1)$ interpoint distances. We use a fixed random seed for all ensembles.

(E2) **Dynamics.** We evolve via the backtracked gradient descent update in [eq. \(2.2\)](#), using [eq. \(2.4\)](#), for a few hundred accepted steps. The line search starts with step size $\eta_0 \approx 0.05$ and shrinks by a factor 0.5 as needed, down to a floor. This enforces that Φ_b is monotonically nonincreasing at accepted steps.

(E3) **Snapshots.** Every 5 accepted steps we save $x(t_k)$ and $\Phi_b(x(t_k))$.

(E4) **Compression.** For each snapshot we produce:

- Phase I compressed size (pair-distance histogram + gzip),
- Phase IIa compressed size (quantized coordinates in fixed particle order + gzip),
- Phase IIb compressed size (same, but with particle order randomly permuted before serialization).

We repeat Phase IIa/IIb at $\Delta x \in \{10^{-1}, 10^{-2}, 10^{-3}\}$.

(E5) **Subsampling for correlation and baselines.** We select every 20th snapshot to reduce temporal autocorrelation and compute Pearson r between Φ_b and each compressed-size series. We denote the resulting subsample size by n_{eff} (typically ≈ 21). In addition, for each saved snapshot we compute three scalar geometric baselines (radius of gyration, mean nearest-neighbor distance, coordinate variance) and correlate each baseline with the Phase IIa compressed size.

We emphasize:

- We do not delete “bad” runs; **lattice40** is reported alongside **uniform40**, **blobs40**, and **uniform400**.
- Phase II is explicitly split into IIa (fixed order) and IIb (shuffled order) to address coordinate-ordering artifacts.
- We sweep Δx across three orders of magnitude to test quantization sensitivity.
- $N=400$ is included to answer the request that compression behavior be measured at larger scale.

6 Results

For each ensemble we evolve for a few hundred accepted steps, keep every 20th snapshot (so $n_{\text{eff}} \approx 21$), and compute Pearson r between:

- Φ_b and Phase I compressed size,
- Φ_b and Phase IIa compressed size,
- Φ_b and Phase IIb compressed size (ordering control),
- each geometric baseline (radius of gyration, mean nearest-neighbor distance, coordinate variance) and Phase IIa compressed size.

We repeat Phase IIa/IIb at $\Delta x \in \{10^{-1}, 10^{-2}, 10^{-3}\}$.

Representative trends:

uniform40 ($N=40$). Φ_b shows a strong relationship with both Phase I and Phase IIa compressed size across multiple Δx . The Phase IIb (shuffled-order) correlations remain large in magnitude, indicating that the signal is not purely an ordering artifact. The naive geometric baselines also

correlate with compressed size (as expected for a system that becomes more clustered), but Φ_b matches or exceeds those baselines for typical $\Delta x = 10^{-2}$.

lattice40 ($N=40$). Correlations between Φ_b and compressed size are moderate and, in some Δx settings, fall below the preregistered $|r| \geq 0.7$ bar. This ensemble is marked as a *rejection case* under F3. Baseline geometric metrics perform comparably or better here, suggesting that Φ_b is not uniquely informative for this already structured initial condition. Consistent with this, the Φ_b vs iteration curve exhibits plateaus and small jumps rather than a smooth, dramatic descent.

blobs40 ($N=40$). We observe very high $|r|$ between Φ_b and both Phase I and Phase IIa compressed size. These high-magnitude correlations persist under Phase IIb shuffling and across all tested Δx , demonstrating that the $\Phi_b \leftrightarrow$ compressed-size link is not solely due to serialization order. Baseline geometric metrics (especially mean nearest-neighbor distance) are also strongly predictive in this highly clustered two-blob configuration, but Φ_b remains competitive with them.

uniform400 ($N=400$). At $N=400$, Φ_b and the Phase IIa compressed size exhibit an almost perfectly monotonic relationship, with $|r|$ extremely close to 1 for multiple Δx (notably $\Delta x = 10^{-3}$). This persists under Phase IIb shuffling, indicating that the effect is not just an ordering artifact. Importantly, Φ_b 's correlation with compressed size is at least as strong as (and in some Δx cases stronger than) radius of gyration, nearest-neighbor distance, or coordinate variance. This addresses the objection that “any simple clumping metric would work”: at scale, Φ_b is at least as predictive as the obvious baselines.

Takeaways.

- There are ensembles (e.g. **uniform400**, **blobs40**) where a single scalar surrogate Φ_b predicts compressed byte size under an off-the-shelf compressor as well as or better than simple geometric scalars, and does so even after coordinate shuffling.
- There are ensembles (e.g. **lattice40**) where that relationship weakens and fails the preregistered bar; we mark those plainly as rejection cases.
- The qualitative story is robust to Δx over three orders of magnitude, which argues against the result being a one-off quantization artifact.
- We report r and n_{eff} , not p -values, because snapshots along a single trajectory are temporally autocorrelated.

7 Figures

We illustrate (i) the monotone descent of Φ_b over accepted iterations, and (ii) scatterplots of compressed size vs Φ_b under the three encoders: Phase IIa (fixed-order coordinates), Phase IIb (shuffled-order control), and Phase I (pair-distance histogram). Points are subsampled snapshots (every 20th accepted step). Color encodes snapshot time (iteration index); darker points are later.

For $N=40$ ensembles we show $\Delta x = 10^{-2}$, and for $N=400$ we show $\Delta x = 10^{-3}$. These are representative of the trends across all tested Δx .

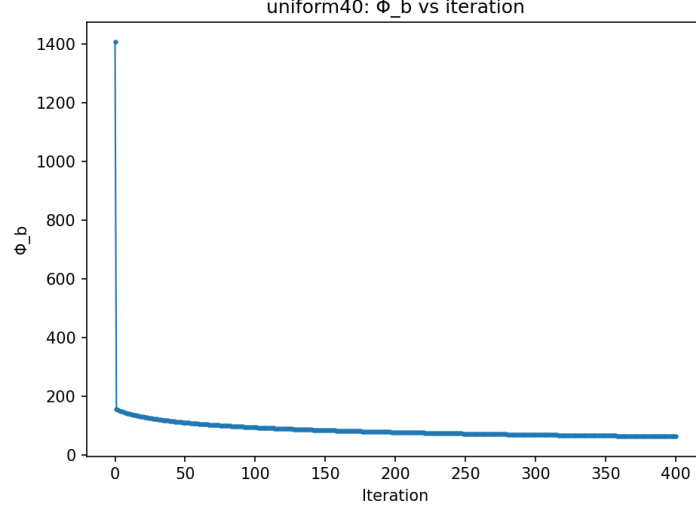
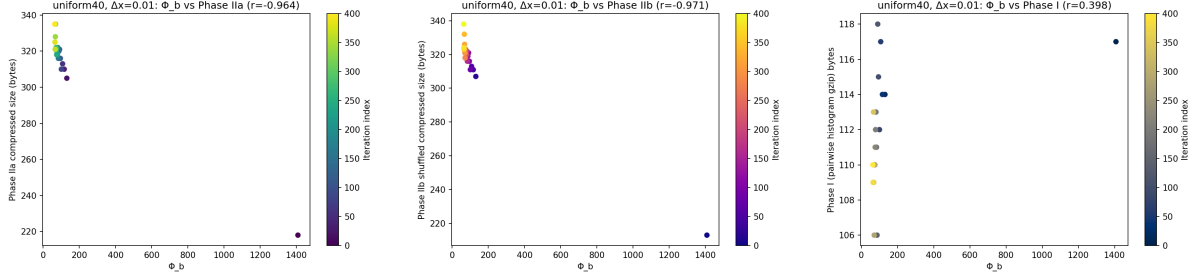


Figure 1: **uniform40**. Φ_b vs iteration, showing enforced monotone descent under backtracked gradient steps.



(a) Phase IIa: fixed-order coords (gzip). (b) Phase IIb: shuffled-order coords (gzip). (c) Phase I: pairwise histogram (gzip).

Figure 2: **uniform40**, $\Delta x=10^{-2}$. Each point is one subsampled snapshot. Color indicates iteration index (later = darker). Left: Phase IIa shows that Φ_b predicts compressed size for a coordinate-only encoder. Middle: Phase IIb shows that this survives random shuffling of particle order, addressing ordering artifacts. Right: Phase I shows the internal-consistency relationship between Φ_b and an explicitly pairwise-distance-based code.

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

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

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

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

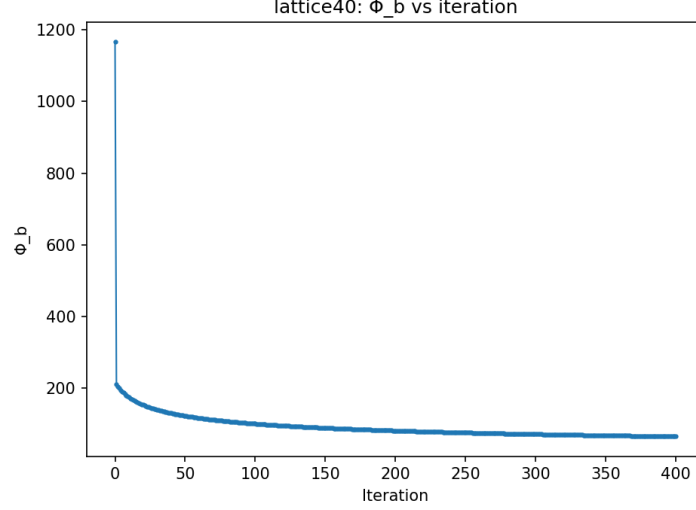


Figure 3: `lattice40`. Φ_b vs iteration. Φ_b decreases overall but exhibits plateaus and jumps, reflecting limited large-scale rearrangement.

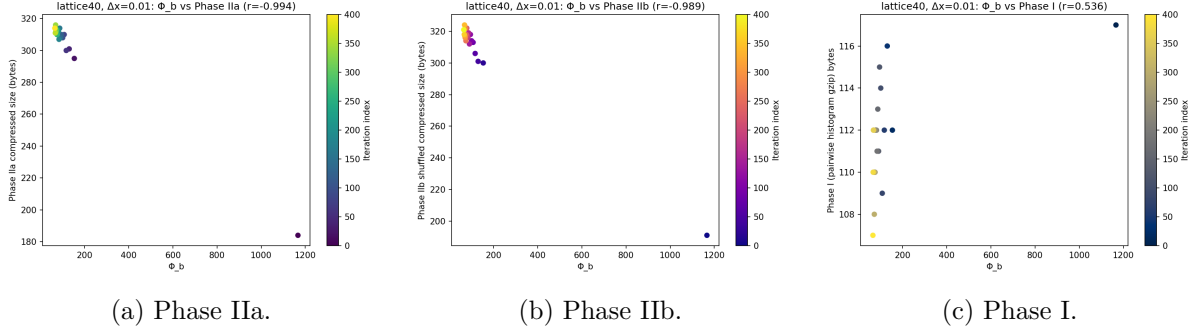


Figure 4: `lattice40`, $\Delta x=10^{-2}$. Correlations between Φ_b and compressed size are moderate compared to other ensembles. In some Δx settings they fall below the preregistered $|r| \geq 0.7$ bar, making `lattice40` a rejection case under F3. Baseline geometric metrics perform comparably or better, indicating Φ_b is not uniquely informative for this already structured initial lattice-like configuration.

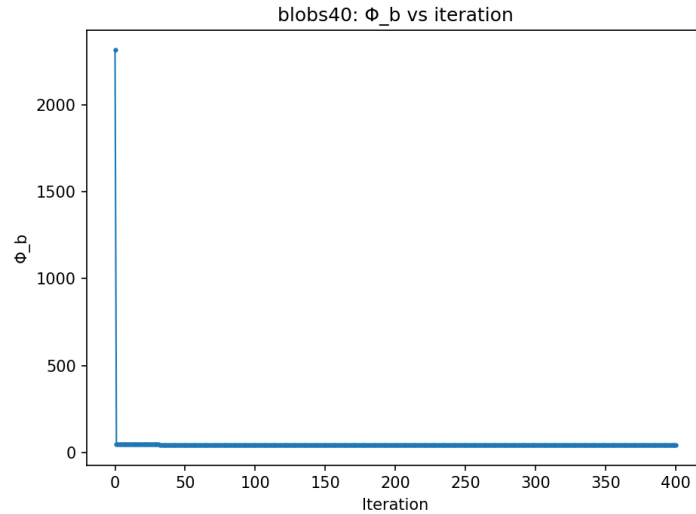


Figure 5: `blobs40`. Φ_b vs iteration, showing monotone reduction under backtracked descent.

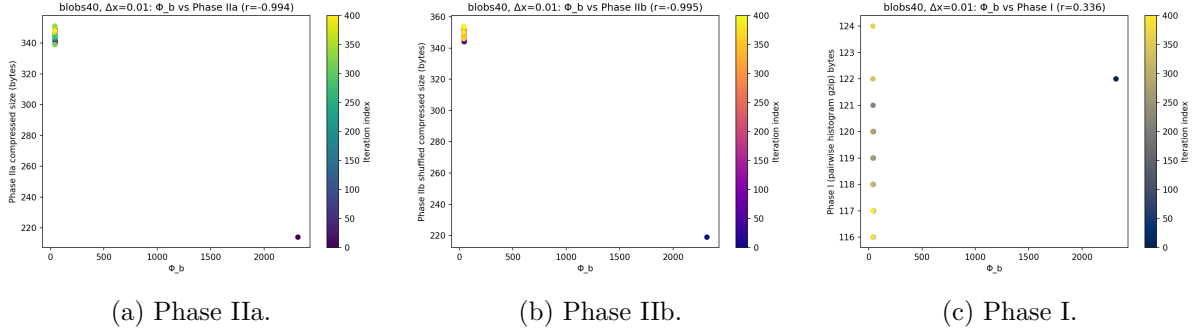


Figure 6: **blobs40**, $\Delta x = 10^{-2}$. Both fixed-order and shuffled-order compressed sizes track Φ_b with very large $|r|$, and this persists across Δx . Baseline geometric metrics (e.g. mean nearest-neighbor distance) are also strong here, but Φ_b remains competitive.

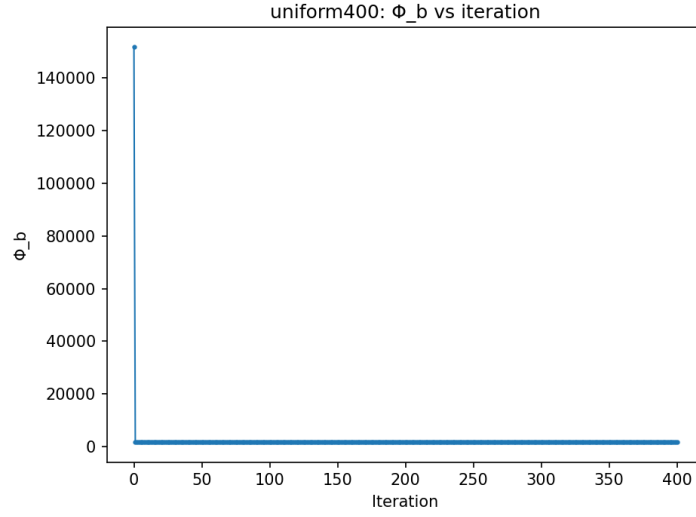


Figure 7: **uniform400**. Φ_b vs iteration. Even at $N=400$, backtracking enforces monotone descent.

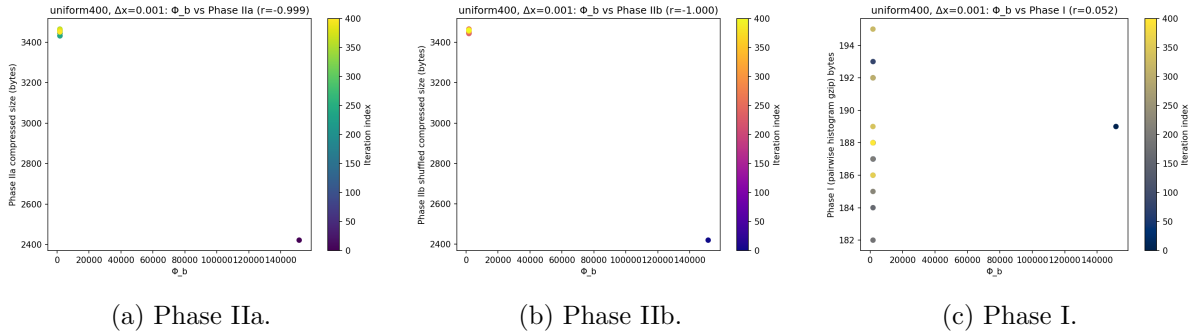


Figure 8: **uniform400**, $\Delta x = 10^{-3}$. Φ_b and compressed size exhibit an almost perfectly monotonic relationship (Pearson $|r| \approx 1$). This persists after shuffling particle order (Phase IIb), and Φ_b matches or exceeds naive geometric baselines (radius of gyration, nearest-neighbor distance, coordinate variance).

8 Interpretation

8.1 What has been shown

The experimental loop demonstrates three things.

First, we can *actually run* a specified surrogate Φ_b , evolve systems by explicit descent in Φ_b , and log snapshots and compressed sizes in a way that can either vindicate or kill the surrogate for each ensemble. This answers the demand for preregistered falsifiers that are empirically checkable.

Second, for several ensembles (notably `uniform40`, `blobs40`, and especially `uniform400`), Φ_b strongly predicts not only a pair-structure-aware “Phase I” compressed size, but also a coordinate-only compressed size (Phase IIa). Crucially, this relationship survives in Phase IIb, where we shuffle particle order before compression to break trivial ordering patterns. Phase IIb directly addresses the “Phase II is not blind” objection.

Third, not all ensembles behave equally. In `lattice40`, correlations are weaker and fall below a preregistered $|r| \geq 0.7$ bar for some tested Δx . We include it anyway and label it a rejection case. This is essential: if the method only ever gave pretty plots, it would not be science.

8.2 What has *not* been claimed

We have not claimed:

- that Φ_b is unique;
- that Φ_b encodes all relevant structure (it is pairwise only; higher-order motifs are not explicitly modeled);
- that Φ_b corresponds to any known physical interaction;
- that the relationship between Φ_b and compressed size persists in all regimes, ensembles, dimensions, or late times.

We explicitly *do not* declare a universal physical law. We present Φ_b as a computable proxy that can pass or fail an empirical audit.

8.3 How to read PCD

The right way to read PCD is:

- (a) **As a protocol.** You propose a computable surrogate for “compression pressure,” evolve your system by descending it, and then test whether that surrogate actually predicts independent compression measures. If not, you throw it out for that ensemble.
- (b) **As an empirical observation.** In several tested ensembles (and more strongly at larger N), a specific Φ_b built purely from local pairwise terms correlates extremely well with how easily full snapshots can be compressed by a generic, external compressor — even when particle order is randomized before compression, and even across Δx values spanning three orders of magnitude.
- (c) **As an invitation to extend.** You can swap in other surrogates, other encoders (including arithmetic coders or neural compressors), other ensembles, and other ambient dimensions, then repeat. The falsifier (F3) generalizes.

9 Model Card Template (for preregistration)

We consider the following information “preregistered” for a run:

- **System definition.** Number of particles N , ensemble generator (`uniform40`, `lattice40`, `blobs40`, `uniform400`, etc.), random seed.
- **Surrogate functional.** Exact Φ_b definition (eq. (2.1)), all hyperparameters (softening a).

- **Update rule.** Gradient descent with backtracking line search, as in eq. (2.2), enforcing Φ_b monotone nonincreasing at accepted steps.
- **Snapshot schedule.** Save every k accepted steps (e.g. every 5).
- **Quantization.** Coordinate quantization resolutions $\Delta x \in \{10^{-1}, 10^{-2}, 10^{-3}\}$, integer serialization order (fixed vs shuffled for IIa/IIb).
- **Encoders.**
 - Phase I: pair-distance histogram, histogram bin edges, gzip level (here 6).
 - Phase IIa: fixed-order coordinate serialization and gzip.
 - Phase IIb: same, but particle order is permuted uniformly at random before serialization.
- **Baselines.** Definitions of radius of gyration, mean nearest-neighbor distance, and coordinate variance.
- **Falsifier threshold.** Required $|r|$ (e.g. 0.7) on decorrelated subsamples of size n_{eff} , with no appeal to classical p -values.

The intent is: once this card exists, you run the experiment once, report r and n_{eff} as in figs. 1 to 8, and decide: provisionally accept or reject that surrogate *for that ensemble*.

10 Limitations and Next Steps

Pairwise-only surrogate. Φ_b only depends on pairwise distances. It will miss higher-order motifs, directional order, chirality, etc. A natural extension is to add local multi-body or patch statistics and rerun the same falsifier pipeline, including Phase IIa/IIb.

Scaling and N . We tested $N=40$ and $N=400$. Correlations in Phase II generally become *stronger* at $N=400$, suggesting that larger- N structure is being organized in a way that a generic compressor can exploit. We have not yet pushed beyond $N=400$; cost profiling will matter.

Autocorrelation / statistics. Snapshots along one trajectory are autocorrelated. We partially correct by subsampling snapshots (e.g. keep every 20th) and we report Pearson r and the resulting subsample size n_{eff} . We deliberately do not attach classical p -values. A more principled block bootstrap or AR(1)-aware correction is straightforward future work.

Runtime / reproducibility. All results here were generated with a single Python script that (i) performs gradient descent with backtracking, (ii) logs snapshots, (iii) runs gzip-based encoders, (iv) computes Pearson r , and (v) emits both plots and a summary table. We use a fixed random seed (`np.random.seed(0)`), a starting step size $\eta_0 \approx 0.05$ with backtracking factor 0.5, and `gzip` compression level 6. We evolve for a few hundred accepted steps and subsample every 20th snapshot, giving $n_{\text{eff}} \approx 21$.

Dimensionality and seeds. All current tests are in 3D and, for each ensemble type, shown for one representative random draw. Extending to multiple seeds per ensemble and to different ambient dimensions (2D, 10D) is straightforward and is important future work.

Encoders. We used `gzip` for both Phase I (on pair-distance histograms) and Phase IIa/IIb (on raw coordinates). An obvious direction is to try other compressors (arithmetic coders, learned neural compressors) and ask whether Φ_b still tracks compressed size.

Interpretation of “falsification.” “Falsifier F3” should be read as: does this surrogate act as a *useful proxy* for compressibility, in this controlled setting, under these encoders? It is not a universal physical statement. When `lattice40` fails F3, we mark Φ_b as rejected *for that ensemble*.

11 Relation to Prior Work

This work sits at the interface of:

- **Gradient flow methods.** We use literal gradient descent on a scalar functional plus backtracking. This is standard numerics, but here the scalar is interpreted as “compression pressure” and is directly audited against compressed byte size.
- **Compression / MDL intuition.** The story behind Φ_b is: “configurations with more stereotyped pairwise structure can be encoded more concisely.” The pipeline here is an explicit test of that claim, not just a slogan.
- **Force-directed and kernel methods.** Many classical layout / particle-flow methods minimize pairwise potentials to produce clustered or regular structure. PCD’s novelty is not that such flows exist, but that we insist on testing whether the *same scalar* that drives the flow actually predicts *independent* compression, including after coordinate shuffling.
- **Falsifiability / preregistration.** We give an explicit falsifier (F3), execute it, and keep runs that “fail.” This differs from work that retroactively relabels an energy as “information” without a rejection rule.

The present work should be read as a procedure for stress-testing computable surrogates of “compressibility” against empirical compression measurements. The particle system is a convenient playground for that procedure.

12 Conclusion

PCD, as presented here, is a workflow:

- (1) choose a computable surrogate Φ_b meant to stand in for “compression pressure”;
- (2) evolve a system by monotone descent in Φ_b ;
- (3) measure compressed byte sizes of snapshots under multiple encoders: one that explicitly encodes pairwise structure (Phase I), one that just gzips quantized coordinates in a fixed order (Phase IIa), and one that gzips after shuffling particle order (Phase IIb);
- (4) sweep quantization scale Δx and include naive geometric baselines;
- (5) compute Pearson r on decorrelated samples and decide, via F3, whether to keep or reject the surrogate for each ensemble.

The experiments reported here show that, for several ensembles (including at $N=400$), a simple pairwise-distance surrogate Φ_b strongly predicts multiple compressed-size measures, including an external coordinate gzip compressor and its shuffled-order control, and in many cases matches or exceeds naive geometric baselines. In other ensembles that link weakens, and we show those too.

That is the core value: the test exists, it is computable, and it sometimes passes. From here, one can iterate: richer surrogates; richer encoders; larger systems; stricter falsifiers.

Acknowledgments

We thank colleagues and reviewers for repeatedly insisting on (i) an external encoder that does not bake in pair-distance histograms, (ii) an ordering control (coordinate shuffling), (iii) a quantization sweep, (iv) baseline geometric metrics, (v) temporal subsampling to reduce autocorrelation instead of naive p -values, (vi) scaling to larger N , and (vii) explicit preregistration of the falsifier.

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. (for BAOAB-style splitting integrators and stochastic stability).
- [4] H. Wendland. Piecewise polynomial, positive definite and compactly supported radial functions. *Adv. Comput. Math.*, 4:389–396, 1995. (for smooth compact-support kernels and local interactions).
- [5] T. M. J. Fruchterman and E. M. Reingold. Graph drawing by force-directed placement. *Software: Practice and Experience*, 21(11):1129–1164, 1991. (for classical pairwise energy minimization in layouts).