Decoupling Basin Selection from Equilibrium Precision for Learned Atomic Relaxation

Yifei Zhang Evan Dramko Anastasios Kyrillidis
Computer Science Department, Rice University
{yz272, ed55, anastasios}@rice.edu

Abstract

Relaxing a candidate atomic structure to its nearest energy minimum is the expensive inner loop of materials discovery. Machine-learned interatomic potentials accelerate the forces, but the relaxation itself remains an iterated fixed-point map—a learned force field applied for many steps until the atoms stop moving. Training *through* that rollout so that it lands in the correct basin steers the relaxation well, yet stores the entire trajectory and so costs memory that grows with the number of steps; the cheap implicit-function alternative refines the equilibrium with constant memory but, from a poor start, simply sharpens the wrong minimum. Practitioners are thus offered accuracy or efficiency, but not both.

We argue that these two needs are separable. *Which* basin the relaxation falls into is decided early and needs only a short differentiated window, while *how precisely* it settles to the bottom is a fixed-point problem that implicit differentiation solves at constant memory. Pairing a brief unrolled guidance phase with an implicit equilibrium phase yields a relaxation procedure whose memory footprint is a tunable knob rather than a fixed cost. On silicon point-defect relaxation this decomposition matches the accuracy of full backpropagation-through-time at a fraction of its memory and faster wall-clock, and—because it controls basin selection rather than merely saving memory—reaches minima on-par with the method it approximates.

As a small by-product, we note that the same decomposition is not equally useful on every relaxation task, and mapping *where* it applies is a contribution in its own right. Its leverage requires a specific regime: a base potential that mis-selects among genuinely competing minima, on a fixed-cell, near-equilibrium system with a repeated structural motif. We place this regime within the broader landscape of machine-learned-potential relaxation datasets, introducing a shared vocabulary that bridges the machine-learning and materials-science names for the same objects, with the hope to help interdisciplinary research in this area.

1 Introduction

Materials discovery pipelines repeatedly pose the same subproblem: given a proposed atomic configuration, find the nearby configuration at which the net force on every atom vanishes—a *local relaxation*. Performed with density-functional theory (DFT) this is the dominant cost; performed with a machine-learned interatomic potential (MLIP) it is fast per step but still an iterative process, repeatedly displacing atoms along predicted forces until the structure stops changing.

Formally, relaxation is the iteration of a learned operator $G_\theta(\mathbf{X}) = \mathbf{X} + \alpha\mathbf{F}_\theta(\mathbf{X})$ whose fixed points $\mathbf{X}^\star = G_\theta(\mathbf{X}^\star)$ are exactly the structures the potential declares force-free. Because the energy landscape is non-convex, the fixed point reached is *basin-dependent*: it is determined by which basin of attraction the initial guess falls into, not by the operator alone. Training the potential so that its relaxations land on the DFT-relaxed targets therefore has two distinct jobs—*selecting* the right basin and *precisely locating* the minimum within it. The equilibrium here is thus the

relaxation itself—a fixed point over atomic *positions*—not the force field's internal computation, which a separate line instead solves to a fixed point over network *features* [10]; those two axes are orthogonal and composable (section 5).

Two training signals are available, with complementary costs. Backpropagation-through-time (BPTT) differentiates the full rollout: it supplies an exact trajectory gradient that can steer basin selection, but it must retain every intermediate step in order to backpropagate through it. The cost is easy to underestimate: a "step" is not a cheap coordinate update but a full evaluation of the interatomic potential—a deep message-passing or Transformer forward pass over all n atoms—so each retained step carries that network's entire activation footprint, and the total is that footprint *times* the number of steps. For a realistic potential and a relaxation of a few hundred steps this readily reaches tens of gigabytes and exhausts the device: the memory is linear in the trajectory length. Implicit differentiation through the fixed-point condition (the implicit-function theorem, IFT) supplies a gradient at constant memory, but it optimizes the location of *whichever* fixed point the solver reached; from a poor initialization it rigorously sharpens the wrong minimum and provides no signal on how to reach a better one. The standard tension is that one buys accuracy and the other buys efficiency.

Contributions. We make the following claims, and we are explicit about which are established and which are the subject of a pre-registered experiment.

- **A decomposition of the training signal** (section 3): basin selection is handled by a short truncated-BPTT *guidance* window and equilibrium precision by an implicit *equilibrium* phase, with the two gradients decoupled. The memory cost is set by the guidance horizon K and is decoupled from the relaxation length. This inverts the usual placement of unrolling in implicit models, where a few unrolled steps either stand in for the implicit gradient or approximate it near the equilibrium [6, 23]: here the differentiated window comes *first*, to select the basin, and the implicit gradient comes *last*, for precision—the placement follows from which subproblem each solves (section 3).
- **An empirical demonstration on silicon point defects** (section 4): the decomposition matches full-BPTT accuracy at substantially lower memory and faster wall-clock across five seeds, and a controlled basin-capture analysis shows the advantage is mechanistic (it reaches more correct minima within our dataset of concern), not merely a memory optimization.
- **A characterization of the method's regime of validity** (section 6): we state precisely the setting in which the decomposition has leverage—a base potential that mis-selects among genuinely competing minima, on a fixed-cell, near-equilibrium system with a repeated structural motif—and, for settings outside it, what they would instead require (an in-domain direct potential, or retraining). We ground this in a landscape of machine-learned-potential relaxation datasets (appendix A) that supplies a shared vocabulary and a regime map, and in cross-dataset regime characterizations (appendix B) that place a strong cell-aware coordinate on that map—a coordinate at which a strong in-domain base already lands in the target basin, so the short-window decomposition would need a competing basin (or retraining) to add value.

We frame the method as a learned warm-start for a fixed-point relaxation operator (section 5), which cleanly separates it from prior backward-pass approximations for implicit models, while being careful that the accompanying guarantees from the warm-start literature do *not* transfer to a learned, non-contractive operator.

2 Problem Setup and Background

Notation. An atomic structure with n atoms is $\mathbf{X} \in \mathbb{R}^{n\times 3}$; a fixed descriptor set (e.g. atomic numbers) is $\mathbf{D} \in \mathbb{R}^{n\times d}$; the DFT-relaxed ground-truth target is $\mathbf{X}_{\text{gt}}$. An MLIP is a parameterized force map $\mathbf{F}_\theta(\mathbf{X}) : \mathbb{R}^{n\times 3} \to \mathbb{R}^{n\times 3}$. A single relaxation step is the operator

$$\mathbf{X}_{t+1} = G_\theta(\mathbf{X}_t) = \mathbf{X}_t + \alpha \mathbf{F}_\theta(\mathbf{X}_t), \tag{1}$$

with step size α. A relaxed structure $\mathbf{X}^\star$ is a fixed point, $\mathbf{X}^\star = G_\theta(\mathbf{X}^\star)$, equivalently $\mathbf{F}_\theta(\mathbf{X}^\star) = \mathbf{0}$. Because the landscape is non-convex, $\mathbf{X}^\star = \mathbf{X}^\star(\theta; \mathbf{X}_0)$ depends on both θ and the initialization $\mathbf{X}_0$ through the basin of attraction selected. The datasets this work uses, touches as landscape, or treats as open questions are catalogued in the dataset landscape of Appendix A.

Structural loss. We supervise the *structure*, not the forces, with the mass-weighted ΔQ displacement

$$\mathcal{L}_{\Delta Q}(\mathbf{X}^\star, \mathbf{X}_{\text{gt}}) = \sqrt{\textstyle\sum_{i=1}^{n} m_i \left\|\mathbf{x}_i^\star - \mathbf{x}_{\text{gt},i}\right\|^2}, \tag{2}$$

where mass-weighting prioritizes the heavy backbone over light, mobile species. $\mathcal{L}_{\Delta Q}$ is an *endpoint* loss: it penalizes the destination, leaving the transient path unconstrained except through the training gradient.

Direct vs. energy-conserving forces. Forces can be produced directly, $\mathbf{F}_\theta = \text{MLP}(\cdot)$, giving a generally asymmetric Jacobian $\mathbf{J}_\mathbf{F} \neq \mathbf{J}_\mathbf{F}^\top$ (a non-conservative field), or as the negative gradient of a learned energy, $\mathbf{F}_\theta = -\nabla_\mathbf{X} \mathcal{E}_\theta$, giving a symmetric Jacobian equal to the negative PES Hessian. The two differ sharply for trajectory training: differentiating through a rollout of an energy-conserving model requires second-order (Hessian-vector) bookkeeping, whereas a direct head needs only first-order vector-Jacobian products. *Throughout this work we use direct-force models only*; energy-conserving forces under BPTT are a separate, harder regime that we explicitly leave out of scope.

Two gradients through a fixed point. For the rollout $\mathbf{X}_K = G_\theta^{(K)}(\mathbf{X}_0)$, the BPTT gradient of an endpoint loss expands as

$$\nabla_\theta \mathcal{L}(\mathbf{X}_K) = \frac{\partial \mathcal{L}}{\partial \mathbf{X}_K} \sum_{t=0}^{K-1} \Big(\textstyle\prod_{j=t+1}^{K-1} \mathbf{J}_G(\mathbf{X}_j) \Big) \frac{\partial G_\theta(\mathbf{X}_t)}{\partial \theta}, \tag{3}$$

storing the full autograd graph of all K force evaluations—each a complete MLIP forward pass over the n atoms, with its own layer activations—so the memory is $O(K)$ in the horizon, and $O(T)$ for full BPTT (K=T). The growth is plainly linear in practice: peak training memory on our silicon point-defect runs rises from 1.3 GB at K=10 to 3.2 GB at K=30, with full-unroll BPTT at 7.5 GB (tables 1 and 8), while the implicit gradient below holds a single evaluation's activations regardless of depth. At a fixed point $\mathbf{X}^\star$ with $\mathbf{J}_G = \mathbf{I} + \alpha \mathbf{J}_\mathbf{F}$ and assuming spectral radius $\rho(\mathbf{J}_G^\star) < 1$ (which is often a strong assumption), the IFT gives the trajectory-length-independent gradient

$$\nabla_\theta \mathcal{L}_{\text{eq}} = \frac{\partial \mathcal{L}}{\partial \mathbf{X}^\star} (\mathbf{I} - \mathbf{J}_G(\mathbf{X}^\star))^{-1} \frac{\partial G_\theta(\mathbf{X}^\star)}{\partial \theta}, \tag{4}$$

computed at $O(1)$ memory by solving the adjoint system $\mathbf{u}^\top(\mathbf{I} - \mathbf{J}_G(\mathbf{X}^\star)) = \partial \mathcal{L}/\partial \mathbf{X}^\star$ with vector-Jacobian products. Equation (4) is exactly the $K \to \infty$ limit of eq. (3) (a Neumann series for $(\mathbf{I} - \mathbf{J}_G^\star)^{-1}$) when the map is locally contractive—but it carries *no* signal about how $\mathbf{X}_0$ reaches $\mathbf{X}^\star$. This is the crux: eq. (3) can steer basin selection at $O(K)$ memory; eq. (4) refines the equilibrium at $O(1)$ memory but only the basin it is already in.

3 Method: Basin Selection ⊕ Equilibrium Precision

We resolve the accuracy–efficiency tension by assigning each gradient to the job it is suited for. The training step decomposes the relaxation into a short *guidance* phase trained by truncated BPTT and an *equilibrium* phase trained by implicit differentiation, with the two losses combined but their gradients structurally decoupled.

Phase 1 – guidance (truncated BPTT). From $\mathbf{X}_0$ we unroll a fixed, short horizon $K \ll T_{\max}$, $\mathbf{X}_K = G_\theta^{(K)}(\mathbf{X}_0)$, tracking gradients, and apply a guidance loss $\mathcal{L}_{\text{guide}} = \mathcal{L}_{\Delta Q}(\mathbf{X}_K, \mathbf{X}_{\text{gt}})$ with gradient eq. (3). This is the only differentiated rollout, so it alone sets the memory cost, $O(K)$. Its role is basin selection: it penalizes vector fields whose trajectories drift toward incorrect metastable states in the transient phase.

Phase 2 – equilibrium solve (no gradient). From the handoff state $\mathbf{X}_K$ (detached) a black-box solver iterates G_θ to a fixed point $\mathbf{X}^\star$ with $\|\mathbf{F}_\theta(\mathbf{X}^\star)\| < \epsilon$, in a no-gradient context that discards each step's graph ($O(1)$ memory). Because we never differentiate through this phase, we are free to use non-differentiable acceleration (e.g. Anderson acceleration, which replaces the raw fixed-point step with a least-squares-optimal combination of the last few iterates and their residuals [4, 39] and so converges in far fewer force evaluations than naive Picard iteration; any fixed-point solver—Picard, quasi-Newton/Broyden, or Newton–Krylov—serves here, since the phase carries no gradient), running the relaxation tail—the many small low-frequency adjustments—for as long as needed without memory growth.

Phase 3 – equilibrium precision (IFT). At $\mathbf{X}^\star$ we re-engage autograd through a single application of G_θ and use the implicit gradient eq. (4) of $\mathcal{L}_{\text{eq}} = \mathcal{L}_{\Delta Q}(\mathbf{X}^\star, \mathbf{X}_{\text{gt}})$ to sharpen the equilibrium at constant memory.

Gradient decoupling. For a contracting map the equilibrium forgets its initialization, $\partial \mathbf{X}^\star / \partial \mathbf{X}_K \to \mathbf{0}$, so the equilibrium gradient does not reach the guidance phase:

$$\partial \mathcal{L}_{\text{eq}} / \partial \mathbf{X}_K = (\partial \mathcal{L}_{\text{eq}} / \partial \mathbf{X}^\star)(\partial \mathbf{X}^\star / \partial \mathbf{X}_K) \approx \mathbf{0}.$$

Training on $\mathcal{L}_{\text{eq}}$ alone would leave the first K steps without gradient—the model would never learn basin selection; training on $\mathcal{L}_{\text{guide}}$ alone would steer the trajectory but never enforce $\mathbf{F}_\theta(\mathbf{X}^\star) = \mathbf{0}$ precisely at the target. The objective is therefore the decoupled sum

$$\mathcal{J}(\theta) = \lambda_{\text{guide}} \mathcal{L}_{\Delta Q}(\mathbf{X}_K, \mathbf{X}_{\text{gt}}) + \lambda_{\text{eq}} \mathcal{L}_{\Delta Q}(\mathbf{X}^\star, \mathbf{X}_{\text{gt}}), \tag{5}$$

whose gradient optimizes the vector field along the path (via BPTT) and the stability of the final attractor (via IFT) on their respective timescales. Algorithm 1 states the training step; appendix C records solver and masking details.

The memory argument. Pure BPTT through an N_{steps} relaxation stores $O(N_{\text{steps}} \cdot n \cdot d)$ of computation graph; the decomposition stores only the K-step guidance window, $O(K \cdot n \cdot d)$, plus an $O(1)$ solve and an $O(1)$ adjoint. Since K is fixed and small while N_{steps} grows with structural difficulty, K is a dial trading accuracy against memory rather than a fixed budget—this is the accuracy–memory Pareto frontier we sweep in section 4.

4 Silicon Point-Defect Results

We evaluate on silicon point-defect relaxation (217-atom fixed cell), using a direct-force transformer backbone (ADAPT) as $\mathbf{F}_\theta$. The set comprises roughly one hundred relaxations of vacancy, interstitial,

Algorithm 1 Decomposed relaxation training step (guidance $\oplus$ equilibrium).

Require: operator G_θ, init $\mathbf{X}_0$, target $\mathbf{X}_{\text{gt}}$, horizon K, tol ϵ, weights $\lambda_{\text{guide}}, \lambda_{\text{eq}}$

$\mathbf{X} \leftarrow \mathbf{X}_0$

Phase 1: guidance (BPTT, $O(K)$ memory)

for $t = 1$ **to** K **do**

  $\mathbf{X} \leftarrow G_\theta(\mathbf{X})$ ▷ forward pass, tracking gradients

end for

$\mathbf{X}_K \leftarrow \mathbf{X}$; $\mathcal{L}_{\text{guide}} \leftarrow \mathcal{L}_{\Delta Q}(\mathbf{X}_K, \mathbf{X}_{\text{gt}})$

Phase 2: equilibrium solve (no grad, $O(1)$ memory)

with no_grad: $\mathbf{X}^\star \leftarrow \text{Solver}(G_\theta, \mathbf{X}_K, \text{tol}{=}\epsilon)$ ▷ iterate until $\|\mathbf{F}_\theta\| < \epsilon$

Phase 3: equilibrium precision (IFT, $O(1)$ memory)

$\hat{\mathbf{X}} \leftarrow G_\theta(\mathbf{X}^\star)$ ▷ one pass to re-attach θ

solve adjoint $\mathbf{u}^\top(\mathbf{I} - \mathbf{J}_G(\mathbf{X}^\star)) = \nabla_{\hat{\mathbf{X}}}\mathcal{L}$

$\mathcal{L}_{\text{eq}} \leftarrow \mathcal{L}_{\Delta Q}(\hat{\mathbf{X}}, \mathbf{X}_{\text{gt}})$

return $\lambda_{\text{guide}}\mathcal{L}_{\text{guide}} + \lambda_{\text{eq}}\mathcal{L}_{\text{eq}}$

Table 1: Silicon point-defect relaxation: best ΔQ (mean $\pm$ std over 5 seeds; lower is better), peak training memory, and cost relative to BPTT. The decomposition matches BPTT accuracy at a fraction of the memory.

Method	**Best ΔQ**	**Peak mem.**	**Mem. vs. BPTT**	**Seeds**
BPTT	246.0 ± 5.8	7,470 MB	1.0× (baseline)	5
Decomp. K=20	248.3 ± 9.0	2,145 MB	**3.5**× less	5
Decomp. K=40	251.4 ± 11.5	4,038 MB	1.8× less	5

Wall-clock: K=20 runs 1.82× faster than BPTT. On silicon, the iterated relaxation also improves on a one-shot direct structure-prediction control (same backbone, coordinate head, ΔQ loss) by roughly 6× ($\Delta Q \approx 14.96$ for the one-shot control versus the iterated relaxation), consistent with the basin-capture analysis below. This iteration-over-one-shot advantage is specific to the weak-base, repeated-motif regime; a setting with a strong in-domain base offers no competing basin for the guidance window to re-select, and is characterized in appendix B.

and antisite point defects in 217-atom cells with the periodic box held fixed, so only the defect neighborhood moves while the surrounding bulk lattice stays in place (full dataset characteristics are catalogued in Appendix A). This system has a repeated bulk motif with a localized defect—the regime in which the decomposition has the most leverage (section 6). All comparisons use identical train/test splits with no overlapping structure pairs.

Matched accuracy at lower memory. Table 1 reports best ΔQ (lower is better) for full BPTT against the decomposition at two guidance horizons, over five seeds, with peak training memory and wall-clock relative to BPTT. The decomposition at K=20 matches BPTT accuracy while using 3.5× less memory and running 1.82× faster; increasing K trades memory back for a small accuracy gain. All methods sit well below the materials-science success threshold $\Delta Q < 300$ for this dataset. The threshold corresponds to roughly 0.1 Å of mass-weighted structural agreement on the relaxed coordinates—the accuracy at which a relaxed geometry is considered usable for downstream property prediction; below it, residual error is set by the base potential rather than by the relaxation procedure.

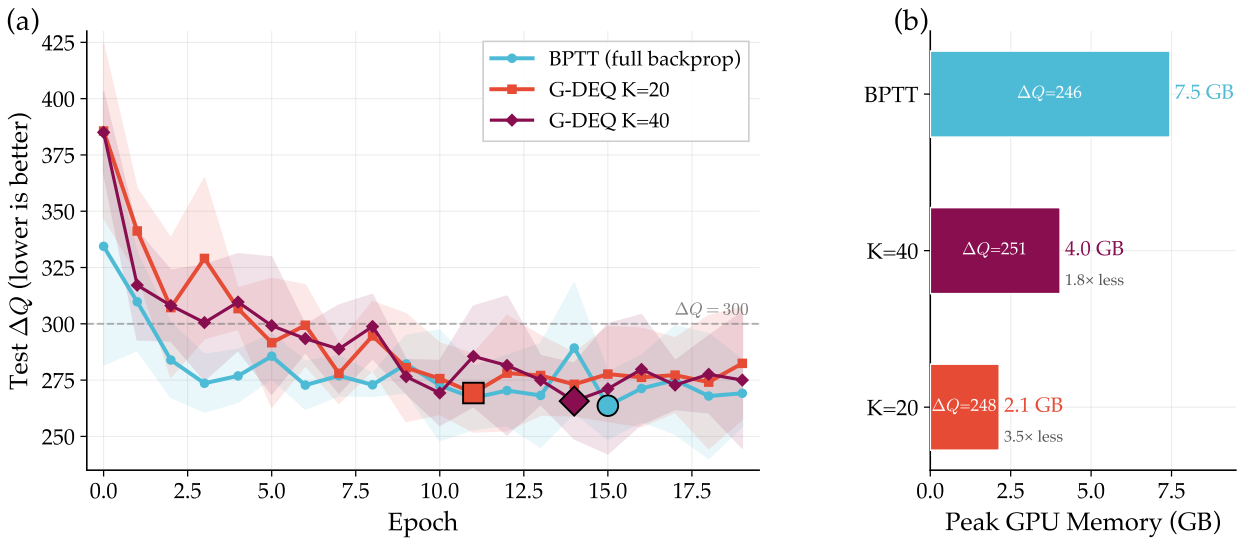

Figure 1: Silicon point-defect relaxation, headline result. (a) Test ΔQ (lower is better; mean and across-seed band) versus training epoch for full BPTT and the decomposition at K=20 and K=40: all three converge to the same accuracy band, well below the $\Delta Q < 300$ success threshold. (b) Peak training memory for the same three settings: the decomposition at K=20 reaches BPTT-level accuracy ($\Delta Q \approx 248$ vs. 246) at 2.1 GB versus BPTT's 7.5 GB (3.5× less). Accuracy is matched; memory is the dial.

Table 2: Basin capture on silicon (100 defects). The decomposition captures the most correct basins; implicit-only training underperforms the pretrained baseline, confirming that guidance (not precision) drives basin selection.

Training signal	**Basins captured**	**Mean ΔQ**
Pretrained (no fine-tune)	43/100	5.9
Implicit-only	25/100	9.7
BPTT	79/100	3.7
Decomposition	**85/100**	**3.2**

Basin capture. To show the decomposition changes *which* minimum is reached—rather than merely reproducing BPTT cheaply—we compare, at matched inference, how often each training signal lands in the correct basin (table 2). Models trained with the decomposition capture the most correct basins, exceeding even BPTT, while implicit-only training is *worse* than the pretrained baseline: optimizing the equilibrium without guidance sharpens wrong minima. This is the empirical signature of the decoupling argument in section 3.

The accuracy–memory frontier. Sweeping the guidance horizon K traces a Pareto frontier: small K minimizes memory at the smallest K that still matches BPTT accuracy, and larger K buys marginal accuracy at proportionally more memory. Figure 3 plots best ΔQ against peak memory across the swept horizons; K=20 is the knee (smallest K matching BPTT, 3.5× memory savings). The full per-K table is in appendix C.

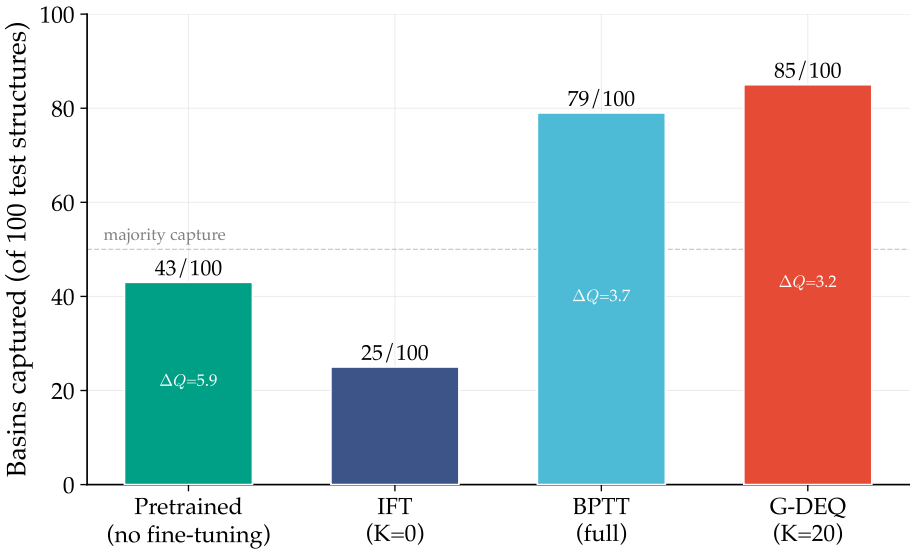

Figure 2: Basin capture on silicon (100 defects; same data as table 2). The decomposition captures the most correct basins (85/100), exceeding full BPTT (79/100); implicit-only training (25/100) falls *below* the pretrained baseline (43/100), the signature that guidance—not equilibrium precision—drives basin selection.

5 Related Work

Differentiating a solution, not a solver (DEQ and implicit layers). A long line of work backpropagates through the *solution* of an inner problem rather than through the iterations of the solver that found it, so that the backward cost is decoupled from the forward solve length.

OptNet differentiates the argmin of a quadratic program through its KKT optimality conditions [3], and differentiable convex layers generalize this to disciplined convex programs [1]; Neural ODEs obtain the same constant-memory property for continuous-depth models via the adjoint method [13]. Deep equilibrium models (DEQ) take this to its limit, *defining* a layer as the fixed point of a map and differentiating it with the implicit-function theorem at $O(1)$ memory [6], with multiscale [7] and Lipschitz-constrained multiscale [35] variants scaling the idea. Inexact backward passes are studied too: the *phantom gradient* of Geng et al. [23] replaces the exact implicit solve with a few damped unrolled steps—a deliberately cheap surrogate for the true DEQ gradient. Our guidance phase uses the same primitive, a short unrolled window, but to the opposite end: it is not an approximation of the equilibrium gradient (which we compute exactly, eq. (4)) but a distinct *basin-selection* signal, placed at the *front* of the relaxation rather than at its fixed point. The deep-learning literature unrolls a few steps to *cheapen* an implicit gradient; we repurpose the identical move to *choose a basin*, and the two uses compose.

Closest to us, a recent line *DEQuifies* an equivariant force field directly—replacing the deep equivariant stack with a fixed-point layer solved once per configuration and warm-started across molecular-dynamics steps for inference speed [10]. Crucially its fixed point is over the network's *features* (an implicit force-field *architecture*), with atomic positions the fixed input and the relaxation left external; ours is over the atomic *positions* (an implicit *relaxation*), and we backpropagate through that relaxation to train basin selection—so a feature-DEQ force field could serve as the operator $\mathbf{F}_\theta$ inside our position-DEQ relaxation, and the two compose rather than compete. Our equilibrium-precision phase is exactly this move—a single adjoint solve at the fixed point, at memory

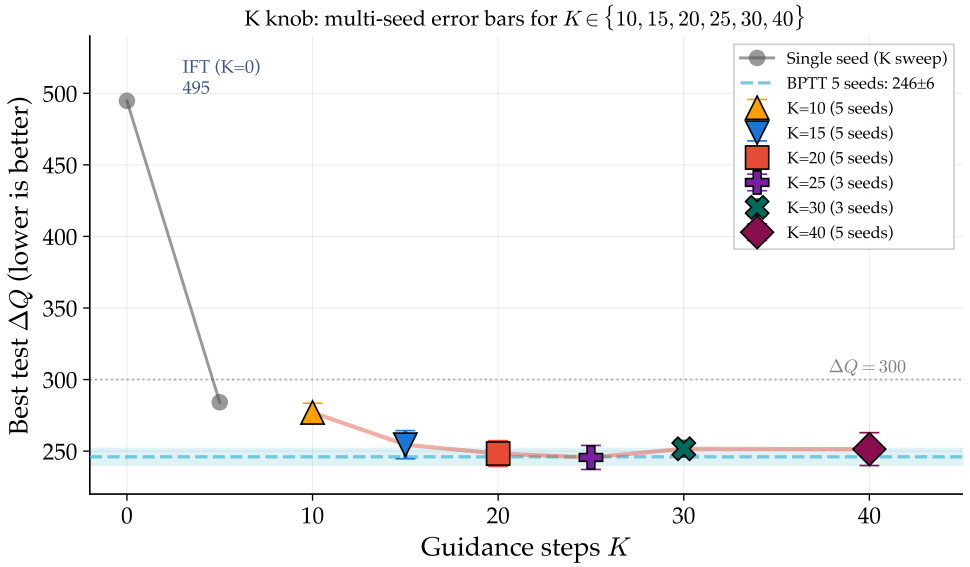

Figure 3: Accuracy frontier on silicon point defects, as a function of the guidance horizon. Best test ΔQ (lower is better, mean $\pm$ std over seeds) against the guidance horizon K, with the full-BPTT five-seed level (246 ± 6) as a dashed reference and the materials-science success threshold $\Delta Q < 300$ marked. Implicit-only (K=0) sits far above threshold; accuracy saturates by K=20, which is the knee—the smallest horizon matching BPTT, at 3.5× less memory (table 1, fig. 1). The guidance horizon K is thus a tunable dial trading memory for a small accuracy margin. The companion memory-versus-K view is fig. 4.

independent of the relaxation depth (eq. (4))—but applied to a *physical* relaxation operator rather than a generic learned layer, and paired with an explicit basin-selection phase that none of these works has an analogue of.

Manufactured versus physical contraction. Because a DEQ fixed point need neither exist nor be unique, much of this literature *manufactures* well-posedness by constraining the operator: Jacobian regularization penalizes the spectral radius of the map [8], monotone-operator equilibrium networks guarantee a unique equilibrium by parameterizing the operator to be monotone [40], and Lipschitz-multiscale variants clamp the operator norm directly [35]. Each buys contraction at an expressivity cost. Our contraction is instead *physical*: near a relaxed minimum $\mathbf{J}_G = \mathbf{I} - \alpha\mathbf{H}$ with the positive-semidefinite PES Hessian $\mathbf{H}$, so small-α contraction comes from the energy landscape itself with no architectural constraint, and our contribution is the guidance (basin selection), not a new way to make the map contract.

Differentiating implicit models cheaply. A parallel thread approximates the *backward* pass to make implicit-model training cheaper: phantom / inexact gradients replace the exact inverse Jacobian with a damped unroll or truncated Neumann series [23], and Jacobian-free backpropagation drops the inverse-Jacobian term to a single fixed-memory factor [21]. These change *how* one differentiates a fixed point; our guidance phase changes *where* the forward iteration starts, and therefore *which* fixed point is reached. The two axes are composable rather than competing, and unlike these inexact backward schemes our equilibrium phase solves the adjoint exactly (eq. (4)).

Learned warm-starts, learned optimizers, and unrolling. Our method is a learned warm-start for a fixed-point iteration, the framing of a body of work on learning to initialize and to optimize. The

closest precedent learns a warm-start for fixed-point and real-time quadratic optimization [33, 32] and provides data-driven (PAC-Bayes) generalization guarantees for learned initializations [34]; learned warm-starts for fixed iterative *physics* solvers are a recent instance [18]. More broadly, learning-to-optimize replaces the update rule itself—meta-learned optimizers trained by unrolling the optimizee trajectory [5, 26]—and algorithm unrolling / deep unfolding turns a *fixed* iterative solver into a trained finite-depth network [24, 27]; the amortized optimization tutorial unifies warm-starts, learned optimizers, and DEQs under one umbrella [2, 14]. The unifying premise of the *guarantees* in this literature is a fixed, known, often contractive update on a convex problem (or, for learned optimizers, a learned initialization with a *classical* update). We violate exactly that premise: we learn the warm-start *and* the operator jointly for a non-convex, non-contractive, asymmetric-Jacobian physical relaxation map, so those bounds do not transfer; and because a short unroll yields a biased, high-variance trajectory gradient [36], we use it only for basin selection and hand precision to the implicit phase. The equilibrium solve uses non-differentiable Anderson acceleration [4, 39], which is admissible precisely because Phase 2 carries no gradient.

Preconditioning and adaptive step sizes. Our operator $G_\theta(\mathbf{X}) = \mathbf{X} + \alpha \mathbf{F}_\theta(\mathbf{X})$ uses a single constant scalar step α; a complementary axis to changing *where* the iteration starts is changing *how fast and along which metric* it descends. The classical control here is FIRE [9], a *preconditioned* structural relaxer that adapts an effective step size and damping from the force–velocity inner product, and Anderson acceleration [4, 39], which preconditions a fixed-point iteration by extrapolating from a window of past residuals. Between a constant α and a fully learned metric lies a spectrum: quasi-Newton / L-BFGS curvature approximations, Adam-style diagonal preconditioning, and learned preconditioners or step sizes for iterative solvers, each replacing the scalar α with an increasingly rich (and increasingly data-driven) metric. Crucially, all of these change the *dynamics and convergence* of a *fixed* iteration without changing its operator or its fixed points; our contribution is orthogonal—a basin-selecting warm-start and a memory-decoupled backward—so a learned preconditioner or an adaptive step-size schedule is a natural, composable extension we leave to future work, our α being a constant scalar for now.

Machine-learned interatomic potentials, fine-tuning, and evaluation. The base force fields our procedure wraps come from the MLIP literature: direct-force equivariant transformers such as EquiformerV2 [25], scalable attention-based potentials (EScAIP and its all-to-all-attention successor) [28, 29], cell-aware universal potentials [12, 15], and the smooth, expressive eSEN architecture [19] whose OC25-trained direct-force instantiation is the strong cell-aware base of our beyond-silicon test, part of the broader universal-model-for-atoms family [41]. Adapting such foundation potentials to a target system by fine-tuning is itself an active topic [37]; our contribution is orthogonal to all of these—a *training procedure* (a memory-decoupled, basin-aware fine-tune through the relaxation), not a new potential. Following the finding that low force error is an insufficient proxy for relaxation quality [20], we supervise and evaluate on structural outcomes (ΔQ) and on classical relaxation-based benchmarking [30]. The trajectory-level training we build on is developed in [17], and the silicon backbone is ADAPT [16]; relative to the full-BPTT fine-tune of the former, our decomposition makes the backward memory independent of relaxation depth and adds an explicit basin-selection signal.

6 Limitations: The Regime of Validity

The decomposition's leverage is specific, and stating that regime precisely is part of the contribution. The guidance phase does exactly one job—select the basin—so it has headroom only where a base potential mis-selects among genuinely competing minima. We summarize that regime here and, for

settings outside it, say what they would instead require; the full landscape that places each setting is Appendix A, and a concrete cell-aware coordinate on that landscape is Appendix B.

The regime the method applies to. The decomposition is a method for *weak- or mis-selecting-base, fixed-cell, near-equilibrium, repeated-motif* systems—the silicon point-defect benchmark being the canonical instance. There, a narrowly-trained base offers genuinely competing metastable minima, the short guidance window has a real re-selection to make, and the implicit phase then sharpens the chosen minimum at constant memory (section 4). This is where the method's leverage is specific: basin selection is only real work when the landscape presents a wrong basin to correct.

Settings that would require more. When a strong, in-domain base already lands in the target basin, its own fixed point *is* the target and there is no competing minimum for the guidance window to re-select; on such a setting the decomposition matches the base rather than improving on it, and adding value there would require a base that mis-selects (or retraining the base so that it does). We characterize exactly such a coordinate—variable-cell titanium adsorbate relaxation with a strong cell-aware direct base—in Appendix B: it is a clean instance where the base is already single-basin-correct, so the short-window decomposition has no re-selection to perform.

Extending the method to maximally diverse *bulk* crystals (the open frontier of appendix A) would require an in-domain *direct* potential for bulk crystals, which is what a trajectory-level method would first need before it could be evaluated there. Where a base does not meet the force-accuracy gate at the target, re-selection alone cannot recover the structure—such a setting would instead require retraining the force field itself. Each of these is a characterization of which relaxation problems the decomposition is the right tool for, and which call for a different in-domain potential or additional training rather than a limitation of the decomposition itself; Appendix A maps each onto a shared regime criterion.

Direct-force only. We restrict to direct-force potentials. Energy-conserving forces require second-order differentiation through the rollout and are a separate, harder regime that we deliberately leave out of scope; we do not claim the decomposition as stated applies to them unchanged.

Scope of the memory claim. The memory advantage we claim is precise and narrow, and—unlike the accuracy leverage above—it holds unconditionally across every setting in the landscape: the implicit backward is $O(1)$ in the solver (relaxation) depth via implicit differentiation, whereas full-unroll BPTT stores the entire trajectory and its memory grows with depth, exhausting the device once the structure is large enough (on silicon, at n=648 atoms; appendix C). This is a statement about how the two gradients scale with relaxation depth, established on silicon. We do *not* claim the decomposition is the only feasible training method in general—on small mobile regions a deep rollout fits in memory for all methods—only that its memory cost is decoupled from relaxation depth where BPTT's is not.

Acknowledgments

This work was supported by Rice University, the Rice University George R. Brown School of Engineering and Computing, and the Rice University Department of Computer Science. This work was supported by the Ken Kennedy Institute, NSF CAREER (award no. 2145629); an Amazon Research Award; a Microsoft Research Award.

References

[1] Akshay Agrawal, Brandon Amos, Shane Barratt, Stephen Boyd, Steven Diamond, and J. Zico Kolter. Differentiable convex optimization layers. In *Advances in Neural Information Processing Systems 32 (NeurIPS)*, 2019.

[2] Brandon Amos. Tutorial on amortized optimization. *Foundations and Trends in Machine Learning*, 16(5):592–732, 2023. arXiv:2202.00665.

[3] Brandon Amos and J. Zico Kolter. OptNet: Differentiable optimization as a layer in neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, volume 70 of *Proceedings of Machine Learning Research*, pages 136–145, 2017.

[4] Donald G. Anderson. Iterative procedures for nonlinear integral equations. *Journal of the ACM*, 12(4):547–560, 1965.

[5] Marcin Andrychowicz, Misha Denil, Sergio Gómez Colmenarejo, Matthew W. Hoffman, David Pfau, Tom Schaul, Brendan Shillingford, and Nando de Freitas. Learning to learn by gradient descent by gradient descent. In *Advances in Neural Information Processing Systems (NeurIPS)*, pages 3981–3989, 2016.

[6] Shaojie Bai, J. Zico Kolter, and Vladlen Koltun. Deep equilibrium models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

[7] Shaojie Bai, Vladlen Koltun, and J. Zico Kolter. Multiscale deep equilibrium models. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[8] Shaojie Bai, Vladlen Koltun, and J. Zico Kolter. Stabilizing equilibrium models by Jacobian regularization. *arXiv preprint arXiv:2106.14342*, 2021.

[9] Erik Bitzek, Pekka Koskinen, Franz Gähler, Michael Moseler, and Peter Gumbsch. Structural relaxation made simple. *Physical Review Letters*, 97(17):170201, 2006.

[10] Andreas Burger, Luca Thiede, Alán Aspuru-Guzik, and Nandita Vijaykumar. DEQuify your force field: More efficient simulations using deep equilibrium models. In *AI for Accelerated Materials Design (AI4MAT) Workshop, ICLR 2025 (Spotlight)*, 2025.

[11] Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. The open catalyst 2020 (OC20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021.

[12] Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022. M3GNet; carries the 3x3 lattice as a graph input.

[13] Ricky T. Q. Chen, Yulia Rubanova, Jesse Bettencourt, and David Duvenaud. Neural ordinary differential equations. In *Advances in Neural Information Processing Systems 31 (NeurIPS)*, 2018.

[14] Tianlong Chen, Xiaohan Chen, Wuyang Chen, Zhangyang Wang, Howard Heaton, Jialin Liu, and Wotao Yin. Learning to optimize: A primer and a benchmark. *Journal of Machine Learning Research*, 23(189):1–59, 2022. arXiv:2103.12828.

[15] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J. Bartel, and Gerbrand Ceder. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5:1031–1041, 2023. introduces the Materials Project trajectory (MPtrj) dataset.

[16] Evan Dramko, Yihuang Xiong, Yizhi Zhu, Geoffroy Hautier, Thomas Reps, Christopher Jermaine, and Anastasios Kyrillidis. ADAPT: Lightweight, long-range machine learning force fields without graphs. *arXiv preprint arXiv:2509.24115*, 2025.

[17] Evan Dramko, Yizhi Zhu, Aleksandar Krivokapic, Geoffroy Hautier, Thomas Reps, Christopher Jermaine, and Anastasios Kyrillidis. On the finetuning of MLIPs through the lens of iterated maps with BPTT. *arXiv preprint arXiv:2512.01067*, 2025.

[18] Mohammad Sadegh Eshaghi, Cosmin Anitescu, Navid Valizadeh, Yizheng Wang, Xiaoying Zhuang, and Timon Rabczuk. NOWS: Neural operator warm starts for accelerating iterative solvers. *Computer Methods in Applied Mechanics and Engineering (CMAME)*, 458, 2026.

[19] Xiang Fu, Brandon M. Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, volume 267 of *Proceedings of Machine Learning Research*, pages 17875–17893, 2025. Spotlight.

[20] Xiang Fu, Zhenghao Wu, Wujie Wang, Tian Xie, Sinan Keten, Rafael Gomez-Bombarelli, and Tommi Jaakkola. Forces are not enough: Benchmark and critical evaluation for machine learning force fields with molecular simulations. *arXiv preprint arXiv:2210.07237*, 2022.

[21] Samy Wu Fung, Howard Heaton, Qiuwei Li, Daniel McKenzie, Stanley Osher, and Wotao Yin. JFB: Jacobian-free backpropagation for implicit networks. In *Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 36, No. 6)*, pages 6648–6656, 2022.

[22] Johannes Gasteiger, Muhammed Shuaibi, Anuroop Sriram, Stephan Günnemann, Zachary Ulissi, C. Lawrence Zitnick, and Abhishek Das. GemNet-OC: Developing graph neural networks for large and diverse molecular simulation datasets. *Transactions on Machine Learning Research (TMLR)*, 2022.

[23] Zhengyang Geng, Xin-Yu Zhang, Shaojie Bai, Yisen Wang, and Zhouchen Lin. On training implicit models. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021. arXiv:2111.05177.

[24] Karol Gregor and Yann LeCun. Learning fast approximations of sparse coding. In *International Conference on Machine Learning (ICML)*, pages 399–406, 2010.

[25] Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. EquiformerV2: Improved equivariant transformer for scaling to higher-degree representations. In *International Conference on Learning Representations (ICLR)*, 2024. OC20 S2EF leaderboard variant uses a non-conservative direct force head.

[26] Luke Metz, James Harrison, C. Daniel Freeman, Amil Merchant, Lucas Beyer, James Bradbury, Naman Agrawal, Ben Poole, Igor Mordatch, Adam Roberts, and Jascha Sohl-Dickstein. VeLO: Training versatile learned optimizers by scaling up. *arXiv preprint arXiv:2211.09760*, 2022.

[27] Vishal Monga, Yuelong Li, and Yonina C. Eldar. Algorithm unrolling: Interpretable, efficient deep learning for signal and image processing. *IEEE Signal Processing Magazine*, 38(2):18–44, 2021. arXiv:1912.10557.

[28] Eric Qu and Aditi S. Krishnapriyan. The importance of being scalable: Improving the speed and accuracy of neural network interatomic potentials across chemical domains. In *Advances in Neural Information Processing Systems 37 (NeurIPS)*, 2024.

[29] Eric Qu, Brandon M. Wood, Aditi S. Krishnapriyan, and Zachary W. Ulissi. A recipe for scalable attention-based MLIPs: Unlocking long-range accuracy with all-to-all node attention. *arXiv preprint arXiv:2603.06567*, 2026.

[30] Janosh Riebesell, Rhys E. A. Goodall, Philipp Benner, Yuan Chiang, Bowen Deng, Gerbrand Ceder, Mark Asta, Alpha A. Lee, Anubhav Jain, and Kristin A. Persson. Matbench discovery: A framework to evaluate machine learning crystal stability predictions. *arXiv preprint arXiv:2308.14920*, 2023. Published in Nature Machine Intelligence 7:836–847 (2025), DOI 10.1038/s42256-025-01055-1.

[31] Sushree Jagriti Sahoo, Mikael Maraschin, Daniel S. Levine, Zachary Ulissi, C. Lawrence Zitnick, Joel B. Varley, Joseph A. Gauthier, Nitish Govindarajan, and Muhammed Shuaibi. The open catalyst 2025 (OC25) dataset and models for solid-liquid interfaces. *arXiv preprint arXiv:2509.17862*, 2025.

[32] Rajiv Sambharya, Georgina Hall, Brandon Amos, and Bartolomeo Stellato. End-to-end learning to warm-start for real-time quadratic optimization. In *Proceedings of Machine Learning Research (Learning for Dynamics and Control, L4DC)*, volume 211, 2023. arXiv:2212.08260.

[33] Rajiv Sambharya, Georgina Hall, Brandon Amos, and Bartolomeo Stellato. Learning to warm-start fixed-point optimization algorithms. *Journal of Machine Learning Research*, 25(166):1–46, 2024. arXiv:2309.07835.

[34] Rajiv Sambharya and Bartolomeo Stellato. Data-driven performance guarantees for classical and learned optimizers. *Journal of Machine Learning Research*, 26(171), 2025. arXiv:2404.13831; learned initialization with a classical update.

[35] Naoki Sato and Hideaki Iiduka. Lipschitz multiscale deep equilibrium models: A theoretically guaranteed and accelerated approach. In *Artificial Intelligence and Statistics (AISTATS)*, 2026. arXiv:2602.03297.

[36] Corentin Tallec and Yann Ollivier. Unbiasing truncated backpropagation through time. *arXiv preprint arXiv:1705.08209*, 2017.

[37] Tamás Lajos Tompa, Eszter Varga-Umbrich, et al. Fine-tuning MLIP foundation models: Strategies for accuracy and transferability. *arXiv preprint arXiv:2606.12704*, 2026.

[38] Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M. Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, Anuroop Sriram, Félix Therrien, Jehad Abed, Oleksandr Voznyy, Edward H. Sargent, Zachary Ulissi, and C. Lawrence

Zitnick. The open catalyst 2022 (OC22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023.

[39] Homer F. Walker and Peng Ni. Anderson acceleration for fixed-point iterations. *SIAM Journal on Numerical Analysis*, 49(4):1715–1735, 2011.

[40] Ezra Winston and J. Zico Kolter. Monotone operator equilibrium networks. In *Advances in Neural Information Processing Systems 33 (NeurIPS)*, 2020.

[41] Brandon M. Wood, Misko Dzamba, Xiang Fu, et al. UMA: A family of universal models for atoms. *arXiv preprint arXiv:2506.23971*, 2025.

A The MLIP-Relaxation Dataset Landscape

The datasets in table 3 are routinely described with a single phrase—"relax a structure with a machine-learned potential"—and are, on that description, interchangeable. They are not. Each pairs a different physical system with a different relaxation task, and those differences decide whether a trajectory-level training signal can help at all. Because the machine-learning and materials-science literatures name the same objects differently (a "basin" and a "metastable configuration"; a "rollout" and a "relaxation trajectory"), the distinctions are easy to lose in translation. We make them explicit here in one shared vocabulary. Mapping which relaxation settings need basin-selecting training—and which are already solved by a good base potential—is a contribution in its own right, and it is what lets the cell-aware titanium coordinate of appendix B read as a point on the landscape: a setting where the base already lands in the target basin, so a basin-selecting method has nothing to re-select there.

A.1 What each dataset physically is

Silicon point defects (headline). A single crystalline silicon supercell with the periodic box (the *cell*) held fixed, into which a localized defect—a vacancy, an interstitial, an antisite—has been introduced. Relaxation lets the atoms around the defect settle while the surrounding bulk lattice, repeated identically in every direction, stays essentially in place. Physically: one repeated motif plus a local perturbation. This is the regime in which the decomposition wins (section 4).

OC20 — adsorbate on a metal slab, near-equilibrium. A small molecule or fragment (the *adsorbate*: species built from N, C, O and H) sits on the surface of a metal *slab*—a slab being a few atomic layers of a crystal with vacuum above it, standing in for a catalyst surface [11]. In OC20 the slab atoms are held fixed (*frozen substrate*) and only the adsorbate moves, starting close to where it will bind. Physically: how a fragment attaches to an otherwise-static surface, released near its final state.

OC22 — adsorbate on an oxide slab, far-from-equilibrium. Superficially OC20-like—a fragment on a surface—but the surface is a metal *oxide* (relevant to the oxygen-evolution reaction, OER) and now *every* atom relaxes, including the slab [38]. Oxide surfaces admit several competing *terminations* and surface reconstructions—distinct low-energy ways the top layers can arrange—so the starting structure can be far from any minimum and multiple minima compete. Physically: a reactive oxide surface that can itself rearrange, released far from equilibrium.

OC25 — solvated solid–liquid interfaces, an ensemble. A crystalline surface in contact with an explicit liquid (e.g. water with ions): a solid–liquid *interface* with the solvent modeled atom by atom [31]. Crucially there is no single target structure to relax to: the solvent is thermally disordered, so the physically meaningful object is a temperature-weighted *ensemble* of configurations

Table 3: Datasets this work uses or touches. "Provides" lists whether the source supplies single-point energies/forces (E/F) and/or full relaxation *trajectories*; "Used here" marks whether the data enters our experiments (directly or via the base potential). The silicon point-defect set is the regime in which the decomposition has the most leverage (its headline result); the catalysis sets enter only through the cell-aware direct base (eSEN/OC25) and the titanium regime-characterization coordinate; MPtrj/Matbench-Discovery locate the open bulk-crystal frontier. The landscape that places each of these is this section (Appendix A); the regime of validity is section 6. Sizes are as reported by the source papers.

Dataset	**Year**	**Size**
Si point defects (via ADAPT) [16]	2025	∼100 eval structs; 217-atom cells
OC20 [11]	2020/21	1.28M relaxations (∼265M SP)
OC22 [38]	2022/23	62,331 relaxations (∼9.85M SP)
OC25 [31]	2025	∼7.80M SP / 1.51M envs (88 elem.)
MPtrj [15]	2023	1,580,395 frames (∼49.3M forces)
Matbench-Disc. [30]	2023	MP→WBM relaxations

Dataset	**System**	**Provides**	**Used here**
Si point defects	bulk-crystal point defects (fixed cell)	E/F + trajectories	yes (headline)
OC20 [11]	adsorbate-on-metal-slab (N/C/O)	E/F + trajectories	via base / Ti test
OC22 [38]	adsorbate-on-oxide-slab (OER)	E/F + trajectories	no (landscape)
OC25 [31]	solid–liquid (solvated) interfaces	E/F (+ solvation)	via base (eSEN)
MPtrj [15]	bulk inorganic crystals (variable cell)	E/F (+ stress)	no (open question)
Matbench-Disc. [30]	bulk-crystal stability (variable cell)	E/F + trajectories	no (benchmark ref)

SP = single-point DFT evaluations. "via base" = data enters only through a pretrained base potential (eSEN-OC25-direct [19, 31]), not as a held-out evaluation set. The Si training-set count and the canonical OC25 dataset-paper metadata are flagged for confirmation (see `refs.bib`).

(an *entropic* average), not one minimum. Physically: a wet interface whose liquid side has no "the" relaxed geometry.

MPtrj / WBM — variable-cell bulk-crystal relaxation and stability. General inorganic bulk crystals across the periodic table, relaxed with the periodic box *itself* allowed to change shape and volume (a *variable cell*): the atoms and the lattice relax together toward the crystal's equilibrium structure [15]. Matbench Discovery uses such relaxations (Materials Project → WBM) to score whether a predicted crystal is thermodynamically stable [30]. Physically: finding the equilibrium shape of an arbitrary crystal, cell included—the most structurally diverse setting here, and the open frontier of section 6.

A.2 Why they look related but are different settings

The five settings share the verb "relax" but differ along axes that each change the problem the relaxation actually solves (table 4). Whether the *cell* is fixed or variable decides whether the operator even needs to represent the lattice—a setting that requires a cell-aware base before any relaxation method wrapped around it can apply (section 6). Whether the start is *near* or *far* from equilibrium decides how long the relaxation runs and how much the trajectory matters. Whether there is a *single defined target minimum* or an *entropic ensemble* decides whether the endpoint loss $\mathcal{L}_{\Delta Q}$ (eq. (2)) is even well-defined as stated. Whether the substrate is *frozen* or *all atoms are free* decides how

Table 4: The same verb, different settings. Each column is an axis along which these relaxation tasks differ; the last column anticipates appendix A.4 (does the setting present *competing* minima for a guidance phase to select among?). "Frozen sub." = frozen substrate/slab.

Dataset	**Cell**	**Start**	**Target**	**Frozen sub.?**	**Competing minima?**
Si point defects	fixed	near-eq.	single minimum	n/a (bulk)	yes (metastable defects)
OC20	fixed	near-eq.	single minimum	yes	few (near-basin)
OC22	fixed	far-from-eq.	single minimum	no	yes (terminations)
OC25	fixed	n/a	entropic ensemble	no	ensemble (no single min.)
MPtrj / WBM	variable	far-from-eq.	single minimum	no	yes (polymorphs)

Table 5: A two-way glossary. Left: materials-science terms, for a machine-learning reader. Right: machine-learning / optimization terms, for a materials-science reader. Symbols follow section 2: operator $G_\theta(\mathbf{X}) = \mathbf{X} + \alpha\mathbf{F}_\theta(\mathbf{X})$, fixed point $\mathbf{X}^\star$, target $\mathbf{X}_{\mathrm{gt}}$, guidance horizon K.

MatSci term (for ML readers)		**ML/opt. term (for MatSci readers)**	
adsorbate	molecule/fragment binding a surface	BPTT	backprop-through-time: differentiate the whole rollout; memory $O(K)$
slab	few crystal layers + vacuum, i.e. a surface	fixed point / IFT	$\mathbf{X}^\star = G_\theta(\mathbf{X}^\star)$; implicit-function theorem gives its gradient at $O(1)$ memory (eq. (4))
cell	the periodic box; *fixed* vs *variable*	basin selection	choosing *which* minimum the relaxation lands in
basin	basin of attraction of G_θ: the minimum a start flows to	equilibrium precision	how exactly it settles to that minimum's bottom
relaxation trajectory	the sequence $\mathbf{X}_0, \mathbf{X}_1, \cdots \to \mathbf{X}^\star$	ΔQ	mass-weighted structural error to $\mathbf{X}_{\mathrm{gt}}$ (eq. (2)); *lower is better*
IS2RE / IS2RS	initial-structure → relaxed energy / structure	guidance horizon K	# differentiated steps; the memory dial

many degrees of freedom the relaxation must coordinate. Reading down the "competing minima" column previews the regime map of appendix A.4: it is exactly the systems with genuinely competing minima where a basin-selecting signal has anything to do.

A.3 Terminology, both ways

The bridge is lexical as much as physical. Table 5 gives the compact translation used throughout this paper: the left block defines the materials-science objects a machine-learning reader meets in table 3, and the right block defines the machine-learning machinery a materials-science reader meets in section 3. Two entries carry most of the paper's argument. A *basin* (ML) is just a *basin of attraction* of the relaxation operator G_θ (eq. (1))—the set of starting structures that flow to one particular force-free minimum—so "which basin" and "which metastable configuration" name the same choice. And the task label *IS2RE/IS2RS* (initial-structure-to-relaxed-energy / -relaxed-structure), standard in the catalysis benchmarks, is precisely the map from $\mathbf{X}_0$ to $\mathbf{X}^\star$ (or its energy) that this paper trains: we supervise the relaxed *structure* (IS2RS) via $\mathcal{L}_{\Delta Q}$, not the energy.

A.4 Mapping the G-DEQ regime criterion onto the landscape

The decomposition helps only under a specific, checkable condition, which the mechanism of section 3 makes precise. The guidance phase does exactly one job—choose the basin—so it has leverage if and only if there is a wrong choice to correct *and* that choice is decided early enough that a short, affordable guidance window can reach it. Concretely, G-DEQ has headroom on a setting when the following three basin conditions hold and, additionally, the trajectory is short enough for an affordable horizon (condition 4 in table 6):

1. **Correct base forces.** The base potential is force-accurate at the true minimum ($\mathbf{F}_\theta(\mathbf{X}_{\mathrm{gt}}) \approx \mathbf{0}$); otherwise no basin-selecting warm-start can recover the structure and only retraining the force field would help (the force-dominated case of appendix B) [20].
2. **Genuinely multi-basin.** The landscape presents *competing* minima near the start (metastable defects, oxide terminations, polymorphs), so basin selection is real work rather than a foregone conclusion.
3. **Base lands in the wrong basin.** Left alone, the base relaxation flows to a minimum that is *not* the target—"correct forces, wrong basin" [20]—so there is a re-selection for the guidance window to make.

Condition 3 is not hypothetical. On oxide surfaces such as OC22's, competing surface terminations mean the target is one of several low-energy configurations, so a base relaxation can settle into a minimum other than the reference [38]; and on variable-cell bulk relaxation, machine-learned relaxations change the predicted space group of a few percent of structures (∼2.5–4% for current universal potentials [30])—each such event a base landing in a different basin than the reference. That low single-point force error is an insufficient proxy for relaxation quality [20] is precisely why these wrong-basin events are not caught by force metrics alone, and the self-consistent relaxation-error floors reported on these benchmarks [30] bound how far a base can be trusted to select correctly on its own.

Read as a map, the conditions sort the landscape (table 6). Silicon point defects satisfy all of them against the weak, narrowly-trained ADAPT backbone—repeated motif, competing metastable defects, and a base that mis-selects—which is why the decomposition has headroom there (table 2). The titanium adsorbate coordinate of appendix B satisfies conditions 1 and 2 but not condition 3: the strong, in-domain `eSEN-OC25-direct` base [19, 31] is effectively single-basin on those structures and already lands correctly, so its own fixed point *is* the target and there is nothing to re-select. That coordinate is therefore a *prediction* of this map: a setting where criterion 3 is off means a basin-selecting method has no re-selection to perform, exactly what appendix B reports—a point on the map, not a failure.

A distinct coordinate, OC22 (adsorbate on an oxide slab, far-from-equilibrium, all atoms relax), *satisfies* the wrong-basin condition—oxide terminations give genuinely competing minima, so the base can settle into a configuration other than the target—but it adds a *fourth* condition that silicon meets and OC22 does not: the guidance horizon K must be a meaningful fraction of the trajectory length T. Basin selection is decided within the first K differentiated steps, so those steps must cover the part of the trajectory where the basin is still in play. Silicon succeeds at $K \approx 20$ of $T \approx 70$ ($K/T \approx 0.29$); OC22's far-from-equilibrium relaxations are roughly 7× longer ($T \approx 475$), so a comparable horizon needs $K \approx 140$, beyond the affordable, short-guidance regime the memory argument relies on (section 3). Thus OC22 requires a guidance horizon *proportional to its long trajectory*—the reason the short-window decomposition does not extend there without a longer (and more memory-hungry) guidance phase.

The diverse bulk-crystal case (MPtrj/WBM) is the still-open coordinate: it is plausibly multi-basin and wrong-basin-prone (competing polymorphs, the few-percent space-group changes noted

Table 6: The G-DEQ regime criterion as a map. A setting has headroom for the decomposition only when all four conditions hold; the operative ones are condition 3 (does the base land in the *wrong* basin?) and condition 4 (is the trajectory short enough that an affordable guidance horizon K is a meaningful fraction of T?). "✓" = holds, "×" = does not hold, "?" = open pending an in-domain base. Each "no" is a coordinate where one specific condition is off—a prediction of the map, not a method failure. OC22 satisfies the wrong-basin condition but its far-from-equilibrium trajectories are ∼7× longer than silicon's, so condition 4 is the one it does not meet.

Setting (base)	**(1) forces**	**(2) multi-basin**	**(3) wrong basin**	**(4) short traj.**	**G-DEQ headroom?**
Si defects (ADAPT, weak base)	✓	✓	✓	✓ ($K/T \approx 0.29$)	**yes** (table 2)
Ti adsorbate (eSEN-OC25, strong)	✓	✓	×	✓	no (appendix B)
OC22 (oxide slab, far-from-eq.)	✓	✓	✓	× ($T \approx 475$)	no (needs $K \propto T$)
Bulk crystals (no in-domain direct base)	?	✓	likely	?	open (section 6)

above), but evaluating it needs a base that passes condition 1 *in domain*—an in-domain direct potential for bulk crystals, which is what such a test would first require (section 6). Each of these is thus a coordinate on the map: a specific condition that is or is not met, which is precisely how the map predicts where the short-window decomposition does and does not apply.

A.5 Base potentials: the decomposition is backbone-agnostic

The decomposition wraps an *external* direct-force potential as the operator $\mathbf{F}_\theta$ (eq. (1)); nothing in the guidance or equilibrium phase is tied to a particular force-field architecture, and we exercise it across three distinct backbones. On silicon the base is **ADAPT** [16], a Transformer that maps per-atom features directly to per-atom forces—no energy head and no cell tensor, so it is *cell-blind*. It is narrowly trained on silicon defects, and precisely because it is a weak, specialized potential it leaves genuinely competing metastable-defect minima for the guidance phase to select among (the regime of section 4). The cell-aware coordinate of appendix B instead uses **eSEN-OC25-direct** [19, 31], a strong, cell-aware direct-force potential trained on the OC25 interface corpus; and the far-from-equilibrium oxide coordinate uses **GemNet-OC** [22], a message-passing direct-force network trained on OC22 [38]. The three span the axes that matter here—cell-blind Transformer versus cell-aware message-passing, and weak/narrow versus strong/broad training—yet each plugs into the identical wrapper through a common force-Jacobian interface, verified per backbone by a double-precision finite-difference check. That the same procedure runs unchanged across all three is the sense in which the method is backbone-agnostic: what varies across them is not *whether* the decomposition applies but *where* the setting sits on the regime map of appendix A.4—which is precisely what the map is for.

B Cross-Dataset Regime Characterizations

Appendix A maps each relaxation setting onto a regime criterion; this appendix reports the numeric characterization behind one concrete cell-aware coordinate on that map—variable-cell titanium adsorbate relaxation with a strong, cell-aware, direct-force base (`eSEN-OC25-direct`, OC25-trained and in-library; our adapter's force Jacobian passes a double-precision finite-difference check). This coordinate isolates *where* the decomposition's basin-selecting leverage applies, by placing it against a base that is already force-accurate and single-basin-correct in domain—exactly the setting where condition 3 of the regime map (table 6) is not met.

Setup. The base is a strong cell-aware direct-force model (`eSEN-OC25-direct`). The relaxation

graph is instantiated once from $\mathbf{X}_0$ and held fixed across the guidance window (the neighbor list is treated as a constant external context of the fixed-point map), so the IFT backward remains well-posed: with the graph frozen, $G_\theta(\mathbf{X}) = \mathbf{X} + \alpha\mathbf{F}_\theta(\mathbf{X}; c)$ is a smooth fixed-point equation in $\mathbf{X}$ with constant context c = (edges, types, cell), and $(\mathbf{I} - \mathbf{J}_G(\mathbf{X}^\star))$ is invertible whenever $\rho(\mathbf{J}_G^\star) < 1$. Freezing the edges assumes no bond crosses the cutoff during the K steps—mild in the near-basin, small-displacement regime and standard practice for real relaxers. We compare the decomposition against (i) the base potential's own un-steered relaxation, (ii) BPTT on the same base, and (iii) a direct structure-prediction control using the *same* backbone with a coordinate head, all under a single clean protocol: an 80/20 train/test split, the reported epoch selected on validation ΔQ, a learning rate matched across methods, and one shared base. We report free-atom ΔQ (mobile atoms only, excluding frozen slab and padding).

Result: the coordinate is condition-3-off. Table 7 reports held-out, validation-selected free-atom ΔQ per structure. The base force gate passes (the cell-aware direct base relaxes Ti far better than a cell-blind backbone would). On these held-out structures the base's own un-steered relaxation already lands in the target basin, so every trained method matches rather than improves on it: the iterated decomposition is statistically indistinguishable from the base, and the high-capacity force-fine-tune control (`ft_static`) and one-shot structure-prediction control sit slightly above it (higher ΔQ is worse) because, with no competing basin to re-select, their only remaining action is to perturb already-correct forces. This is precisely the map's prediction for a coordinate where condition 3 is off (table 6): with the base single-basin-correct, a basin-selecting method has no re-selection to perform. An earlier *in-domain* (non-held-out) version of this comparison showed a positive lift for the iterated method; that number reflected fitting the evaluation structures rather than the relaxation mechanism, and we report the held-out characterization here as the faithful one.

Table 7: Titanium adsorbate coordinate (held-out, validation-selected free-atom ΔQ per structure; lower is better). With a strong in-domain base already landing in the target basin, all trained methods match rather than improve on the base—the regime map's prediction for a setting where condition 3 (table 6) is not met. The memory property of the implicit backward (section 3 and appendix C) is independent of this and holds here as everywhere.

Method (held-out, val-selected)	**Free-atom ΔQ/struct**	**Δ vs. base**
Base relaxation (untrained)	20.09	—
G-DEQ, fine-tune (IFT)	≈ 20.1	≈ 0 (matches base)
`ft_static` (frozen-base correction)	21.41	+1.32
DSP, fine-tune (one-shot)	22.06	+1.97

The complementary open coordinate. The characterization above is established on variable-cell titanium adsorbate relaxation, a setting with no competing basin for the guidance phase to exploit. The complementary, untested coordinate is a *force-dominated* substrate whose field is wrong at the target ($\mathbf{F}_\theta(\mathbf{X}_{\text{gt}}) \neq \mathbf{0}$), where re-selection alone cannot recover the structure and only retraining the field would help. A no-shared-motif test on structurally diverse *bulk* crystals sits in this case and remains open for an orthogonal reason: the available cell-aware direct potential is catalysis-trained and out of domain on bulk crystals (its own relaxation drives ΔQ *above* the unrelaxed structure), so that set cannot isolate the relaxation method from base quality and would first require an in-domain direct potential (section 6). The memory property of the implicit backward (section 3) is unaffected by any of this: it is a statement about *how* one differentiates a deep relaxation, established on silicon

(fig. 4), and is what makes such a fine-tune feasible at scale (n=648 atoms; appendix C) regardless of accuracy.

C Additional Tables and Ablations

Per-horizon sweep (silicon). Table 8 reports the full accuracy–memory sweep behind figs. 3 and 4: best ΔQ (mean $\pm$ sample std over seeds, converged 20-epoch runs), peak training memory, and memory ratio versus BPTT, for $K \in \{10, 15, 20, 25, 30, 40\}$. Accuracy is far above threshold at K=10 (277.0) and saturates into the BPTT band by K=20; from K=20 onward the small accuracy differences are within seed noise while memory grows roughly linearly in K ($\approx$ 100 MB per guidance step), so K=20 is the knee.

Table 8: Per-horizon sweep on silicon point defects (5 seeds for $K \in \{10, 15, 20, 40\}$ and BPTT; 3 seeds for $K \in \{25, 30\}$). Best ΔQ is mean $\pm$ sample std; memory is peak training memory. K=20 is the smallest horizon whose accuracy is statistically indistinguishable from BPTT, at 3.5$\times$ less memory.

Setting	Best ΔQ	Peak mem.	Mem. vs. BPTT	Seeds
K=10	277.0 ± 6.4	1,261 MB	5.9$\times$ less	5
K=15	254.5 ± 9.8	1,755 MB	4.3$\times$ less	5
K=20	$\mathbf{248.3 \pm 9.0}$	**2,145 MB**	**3.5$\times$ less**	5
K=25	245.6 ± 8.4	2,737 MB	2.7$\times$ less	3
K=30	251.6 ± 5.9	3,230 MB	2.3$\times$ less	3
K=40	251.4 ± 11.5	4,038 MB	1.8$\times$ less	5
BPTT (full)	246.0 ± 5.8	7,470 MB	1.0$\times$ (ref)	5

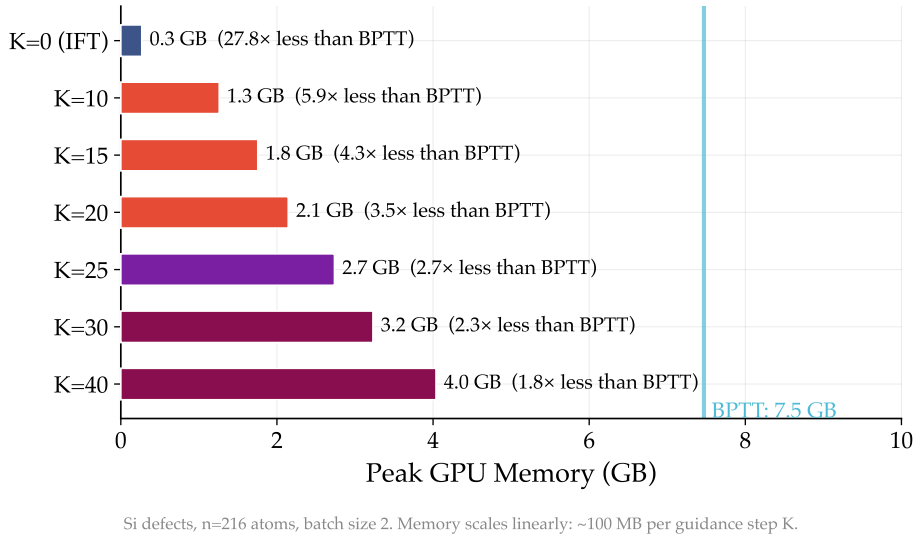

Figure 4: Peak training memory versus guidance horizon K on silicon (n=216 atoms, batch size 2). Memory grows roughly linearly in K ($\approx$ 100 MB/step); the implicit-only setting (K=0) uses 0.3 GB and K=20 uses 2.1 GB, against full BPTT's 7.5 GB. This is the memory axis of fig. 3.

Basin capture as a function of K (toy control). On a controlled toy fixed-point problem with a known correct basin, basin selection improves sharply with a short guidance window and then saturates: mean final error is 1.00 at K=0 (implicit-only collapses, sharpening the wrong fixed point), drops to 0.262 at K=5 and 0.260 at K=10—a roughly 4× reduction that saturates by K=5, a log-K behavior consistent with basin selection being decided early in the rollout. This isolates the guidance mechanism from the equilibrium solve.

Inference-time basin capture by training signal (silicon). At matched inference (K=0 rollout from the same initialization) the training signal determines how many of 100 silicon defects land in the correct basin: G-DEQ-trained 73/100, BPTT-trained 70/100, IFT-trained 23/100, and the pretrained base 16/100. Implicit-only training underperforms even the untrained base, the inference-time counterpart of table 2 and fig. 2.

Scaling and the memory property (silicon). Memory is $O(1)$ in the relaxation (solver) depth via implicit differentiation, whereas full-unroll BPTT stores the whole rollout and its memory grows with depth. The consequence is a hard scaling limit: on the largest silicon system we ran, full BPTT runs out of memory at n=648 atoms, while the decomposition trains at ≈ 5.4 GB—making it the only feasible option *at that size*, not because BPTT is wrong but because its memory footprint exceeds the device once the structure is large enough.

Decomposition ablations (silicon). We ablate (i) the guidance horizon K (table 8), (ii) the loss weights $\lambda_{\text{guide}}, \lambda_{\text{eq}}$, (iii) removing gradient decoupling, and (iv) the optimizer (AdamW / Lion / SGD) for both BPTT and the decomposition. The critical ablation is truncated BPTT alone (λ_{eq}=0): if it matched the full decomposition the equilibrium phase would be unnecessary. Per-run logs for each ablation arm are released with the code; the qualitative conclusion—the hybrid is needed because guidance alone does not enforce the equilibrium and the equilibrium phase alone does not learn basin selection—matches the decoupling argument of section 3.

Solver and masking details. The equilibrium solve uses Anderson acceleration in a no-gradient context; converged trajectories are masked (gradients zeroed past the convergence threshold) so that BPTT only propagates through the active relaxation phase, and free-atom ΔQ is computed over mobile atoms only (excluding frozen-slab and padding atoms). Implementation specifics are deferred to the released code.