BLOCS: SCALABLE BLOCK-CONDITIONAL GENERATIVE PROPOSALS WITH EXACT LOCAL METROPOLIS–HASTINGS CORRECTION

Anonymous authors
Paper under double-blind review

ABSTRACT

Generative molecular samplers propose large conformational changes, but full-system statistical correction becomes inefficient as proposal mismatch accumulates across molecular degrees of freedom. We introduce BLOCS, a transferable block-conditional proposal model for accelerating local equilibrium transitions. Each learned torsional block receives an exact full-system Metropolis–Hastings correction using the molecular energy, coordinate measure, and forward/reverse proposal densities. A state-dependent Hastings-score diagnostic connects correction difficulty to stationary proposal-pair overlap. In controlled scaling, global-proposal acceptance falls from 2.0×10^{-3} at $N = 20$ to 3.1×10^{-9} at $N = 76$, whereas fixed-scope block correction remains usable; complete-pool scans show that moved dimension has a substantially larger effect on acceptance than total protein size. The residual autoregressive proposal, trained on proteins with $N \leq 36$, transfers zero-shot to proteins with up to 250 residues. Under the canonical $m = 12, r = 100$ fixed setting, end-to-end ESS/GPU-s improves by 2.69–3.34$\times$ over Rotamer MH across six proteins, with a median gain of 3.07$\times$. On the five-system common-pool thermodynamic benchmark, gains over a matched-data Robin-style autoregressive control are 1.67–1.91$\times$. In a paired protein-scale exact-composition benchmark with Metropolized dynamics, the gains remain 2.79–2.87$\times$ over Rotamer MH and 1.57–1.68$\times$ over the matched learned control. Acceptance calibration on frozen primary and prospective pools reaches 4.8% median relative error.

1 INTRODUCTION

Equilibrium molecular simulation is limited by rare conformational events: MD resolves femtosecond motion, whereas side-chain exchange and metastable transitions governing thermodynamic observables can be orders of magnitude slower (Lindorff-Larsen et al., 2011; Shaw et al., 2010; Snow et al., 2005). Generative samplers amortize learned conformational structure into larger proposals (Noé et al., 2019; Tan et al., 2025; 2026; Klein and Noé, 2024; Rehman et al., 2026; von Klitzing et al., 2026; Kim et al., 2024; Chen et al., 2026), while protein ensemble models show transfer across sequences and folds (Lewis et al., 2025; Plainer et al., 2026; Mate et al., 2024; Roney et al., 2025; Wang et al., 2026; Lai et al., 2023; Khelifa et al., 2026).

The remaining difficulty is statistical correction. Global proposals combine local density errors into one joint ratio, so mismatch can accumulate with the number of simultaneously proposed coordinates. BLOCS instead selects a finite torsional block, conditions on its local 3D environment, proposes a multimodal residual update, evaluates the complete molecular energy, and applies an exact full-system MH correction to that block move.

The scaling question is how correction changes when the surrounding protein grows versus when the proposed block grows. We separate these effects through global-versus-local scaling, matched (N, m) controls, complete-pool protein scans, context and correlation-length sweeps, and zero-shot transfer from $N \leq 36$ training proteins to held-out proteins with up to 250 residues. Independent umbrella/MBAR references and ubiquitin NMR observables (Wand et al., 1996; Wang and Bax, 1995; Chou et al., 2003) test whether faster local transitions improve finite-budget equilibrium estimation, while matched classical and learned baselines quantify force- and GPU-normalized efficiency.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

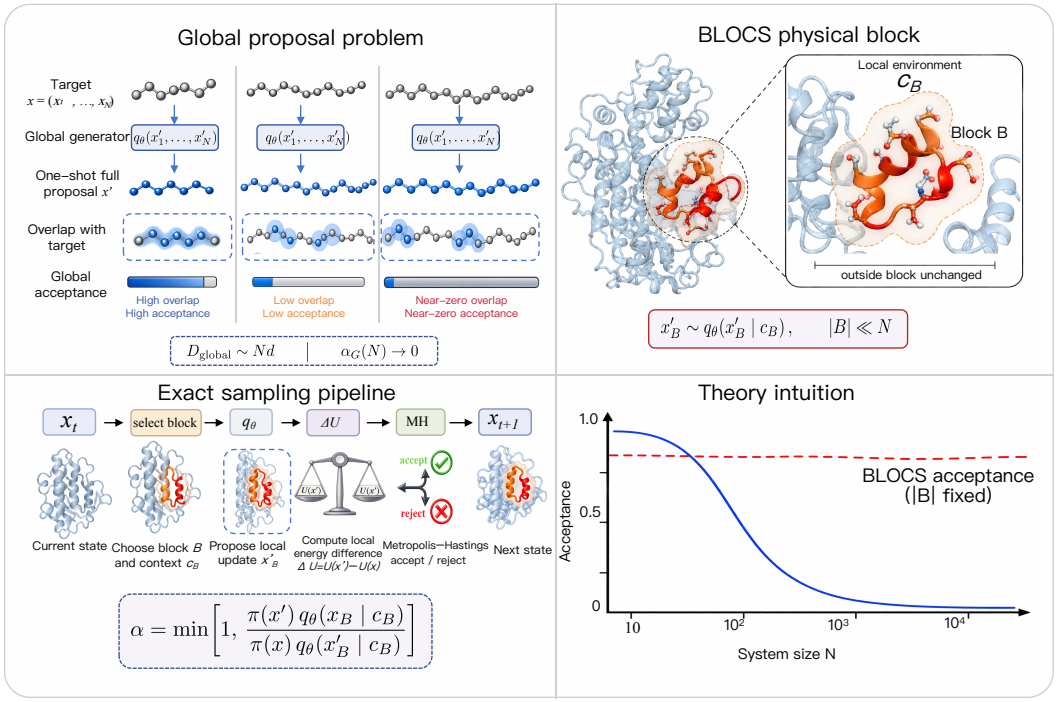


Figure 1: **Block-conditional proposal with exact local MH correction.** Global proposals accumulate mismatch across the full molecular system. BLOCS learns a reusable conditional transition over a finite torsional block, conditions on the local 3D environment, evaluates the complete molecular energy, and applies the full forward/reverse MH correction.

2 Related Work

Boltzmann Generators established a route for combining learned proposals with equilibrium statistical mechanics through tractable densities and Monte Carlo correction (Noé et al., 2019). Subsequent work improves invertible expressivity, macromolecular scaling, transfer, sequential refinement, autoregressive factorization, and coarse-grained generation (Tan et al., 2025; Klein and Noé, 2024; Rehman et al., 2026; Kim et al., 2024; Chen et al., 2026; Durkan et al., 2019). These methods differ in representation and inference, but their corrected transition efficiency is governed by proposal–target overlap at the scale of the object being proposed. BLOCS makes that object a finite local conditional transition whose forward and reverse densities enter a full-system MH ratio.

Transferable protein models show that conformational structure can be reused across sequence and fold variation (Lewis et al., 2025; Plainer et al., 2026; Mate et al., 2024; Roney et al., 2025; Wang et al., 2026; Lai et al., 2023; Khelifa et al., 2026). BLOCS targets recurring local packing and torsional environments rather than complete global ensembles. The context-radius, conditioning-information, novelty, and structural-class experiments in Appendices D, M, and J test this locality directly.

Classical enhanced sampling modifies thermodynamic state or biases selected coordinates, while Bennett, WHAM, and MBAR recover free-energy differences from statistical overlap (Kumar et al., 1992; Shirts and Chodera, 2008; Barducci et al., 2008; Sugita and Okamoto, 1999; Bennett, 1976). BLOCS instead acts on overlap at proposal time: learned moves cross local barriers while exact MH correction is confined to the selected block. The stationary MH acceptance/total-variation identity (Neklyudov et al., 2018) supplies the transition-level diagnostic used below, and finite-step Langevin splitting results (Leimkuhler et al., 2016) motivate the separate exact-composition benchmark for the interleaved dynamics leg.

3 METHOD

BLOCS consists of three coupled components: an exactly corrected block-conditional Metropolis–Hastings move, a state-dependent acceptance diagnostic based on forward/reverse transition-pair overlap, and a transferable residual autoregressive model for multimodal torsional updates from local 3D context. Figure 1 summarizes the learned move and its correction; the production chain additionally interleaves finite-step Langevin-middle MD.

3.1 PROBLEM FORMULATION AND BLOCK-CONDITIONAL TRANSITION

Let Π denote the equilibrium configurational measure,

$$\Pi(dx) = Z^{-1} \exp[-\beta U(x)]\, \mu(dx), \tag{1}$$

where $U(x)$ is the complete molecular-mechanics energy and μ is the configurational reference measure. The learned move is parameterized in an internal-coordinate chart $\xi = (\xi_B, \xi_{\bar{B}})$ with Cartesian map $x = T(\xi)$. If

$$\mu(dT(\xi)) = J(\xi)\, d\xi,$$

then $\widetilde{\pi}(\xi) \propto \exp[-\beta U(T(\xi))] J(\xi)$ with respect to $d\xi$. The proposal draws

$$\xi'_B \sim q_\theta(\cdot \mid \xi_B, c_B(\xi)), \qquad \xi' = (\xi'_B, \xi_{\bar{B}}),$$

and accepts with

$$\alpha(\xi, \xi') = \min\left\{1,\ e^{-\beta[U(T(\xi'))-U(T(\xi))]} \frac{J(\xi')}{J(\xi)} \frac{q_\theta(\xi_B \mid \xi'_B, c_B(\xi'))}{q_\theta(\xi'_B \mid \xi_B, c_B(\xi))}\right\}. \tag{2}$$

Theorem 1 (Target invariance)**.** *For state-independent block selection, Eq. equation 2 satisfies detailed balance with respect to* Π *whenever the forward and reverse proposal densities and the coordinate-measure factor are evaluable on the target support.*

A BLOCS block is an ordered set of internal torsional degrees of freedom realized by deterministic rotations of the corresponding distal side-chain subtrees. The backbone and atoms outside those subtrees remain fixed. Blocks are drawn uniformly from frozen, state-independent benchmark pools. The headline thermodynamic benchmark uses the common `THERMO-COMMON-M12` population at $m = 12, r = 100$ across all 13 locked observables; the six-protein transfer benchmark uses the corresponding `FIXED6-M12` construction, and the dimension scan averages over the complete eligible pool at each m. Previously frozen observable-specific pools are retained as a sensitivity control. Ring, disulfide, and other covalent-closure constraints determine eligibility. The 3D context contains non-block atoms within 9 Å of the block anchors; Appendix F gives the implementation-level transition specification.

Production trajectories interleave these learned moves with 2 fs Langevin-middle MD. Appendix K quantifies the finite-step MD discretization error through a step-size scan and a Metropolized-dynamics control. Production torsion-only BAT moves preserve bond lengths and angles, giving $J(\xi')/J(\xi) = 1$; numerical transform residuals remain below 2.3×10^{-9} through $m = 12$ (Appendix L).

3.2 ACCEPTANCE GEOMETRY AND LOCALITY SCALING

For a fixed block B, the deployed state-dependent proposal is

$$Q_{\theta,B}(\xi, d\xi') = q_\theta(\xi'_B \mid \xi_B, c_B(\xi))\, d\xi'_B\, \delta_{\xi_{\bar{B}}}(d\xi'_{\bar{B}}).$$

Define the stationary proposal-pair measure

$$M_{\theta,B}(d\xi, d\xi') = \widetilde{\pi}(\xi)\, d\xi\, Q_{\theta,B}(\xi, d\xi'), \qquad M^\dagger_{\theta,B}(d\xi, d\xi') = M_{\theta,B}(d\xi', d\xi).$$

The stationary mean acceptance obeys the standard MH acceptance–overlap identity (Neklyudov et al., 2018),

$$\bar{\alpha}_B = 1 - D_{\mathrm{TV}}\left(M_{\theta,B}, M^\dagger_{\theta,B}\right). \tag{3}$$

For an attempted transition define

$$S_\theta(\xi,\xi') = -\beta[U(T(\xi'))-U(T(\xi))]+\log\frac{J(\xi')}{J(\xi)}+\log q_\theta(\xi_B \mid \xi'_B, c_B(\xi'))-\log q_\theta(\xi'_B \mid \xi_B, c_B(\xi)).$$

The Boltzmann normalizer cancels, and

$$D_{\mathrm{tr},B} = D_{\mathrm{KL}}\Big(M_{\theta,B}\|M^\dagger_{\theta,B}\Big) = -\mathbb{E}_{M_{\theta,B}}[S_\theta(\Xi,\Xi')]$$

is estimable from attempted stationary pairs using full-system energies, the coordinate-measure ratio, and explicit forward/reverse proposal densities. Under the empirical approximation $S_\theta \approx \mathcal{N}(-\sigma_S^2/2,\sigma_S^2)$,

$$\bar\alpha_B \approx 2\Phi(-\sigma_S/2) = 2\Phi\left(-\sqrt{D_{\mathrm{tr},B}/2}\right). \tag{4}$$

For a frozen pool $\mathcal{B}$, acceptance and the diagnostic average over the same state-independent selection law; Appendix H reports pool-specific calibration and cost.

The independence-proposal case gives a complementary controlled diagnostic. For fixed outside configuration and proposal q_B,

$$\bar\alpha = 1-\frac{1}{2}\mathbb{E}_{Y,Y'\overset{\mathrm{iid}}{\sim}q_B}|w(Y)-w(Y')|, \qquad w=\pi_B/q_B. \tag{5}$$

With $D_B = D_{\mathrm{KL}}(\pi_B\|q_B)$, the measured log-normal regime gives $\bar\alpha = 2\Phi(-\sqrt{D_B})$. Protein-scale experiments then vary moved dimension, system size, and local context for the deployed transition.

3.3 Transferable residual autoregressive proposal

Protein side-chain transitions are periodic and multimodal. We model wrapped torsion residuals while holding the backbone fixed during each learned block move. Let $\chi_{1:D}$ denote the current block torsions. A 3D encoder aggregates atoms within radius r_c around the block, and a causal network factorizes

$$q_\theta(\Delta\chi_{1:D} \mid \chi, c_B) = \prod_{d=1}^{D} q_{\theta,d}(\Delta\chi_d \mid \Delta\chi_{<d}, \chi, c_B), \tag{6}$$

$$q_{\theta,d} = \sum_{k=1}^{K}\rho_{dk}\,\mathrm{vM}(\Delta\chi_d;\mu_{dk},\kappa_{dk}). \tag{7}$$

Sequential sampling and teacher-forced density evaluation use the same factors, so every multi-torsion move has explicit forward and reverse probability. The final model uses 10 causal layers, hidden width 384, $K=12$ von-Mises components, a 9 Å context radius, and 21.4M parameters. The radius is fixed on the model-selection partition and is close to the subsequently measured 9.4 Å saturation scale; Appendix E gives the split, architecture, and throughput details.

4 Experiments

We organize the experiments around four questions: how exactly corrected block proposals scale with moved dimension; whether one frozen conditional model transfers beyond the training-size and fold distribution; whether faster local transitions improve independent thermodynamic observables; and whether gains persist under matched computational budgets.

4.1 Training and held-out systems

The proposal is trained on proteins with $N_{\mathrm{train}} \le 36$ and evaluated zero-shot on larger held-out folds. Training data include mdCATH and ATLAS (Mirarchi et al., 2024; Vander Meersche et al., 2024) plus force-field-matched simulations. Architecture size, the 9 Å context radius, and the default block policy are fixed using training/model-selection partitions; the named five-protein benchmark is reserved for final metrics, and prospective-diagnostic/extended-transfer systems are disjoint from it (Appendix E).

The primary setting is AMBER ff14SB with GBn2 at 300 K, with explicit TIP3P as a solvent-sensitivity control in Appendix F. The main ladder ends at lysozyme; Appendix J extends the frozen proposal to $N = 250$ and held-out soluble structural classes. The primary five-protein benchmark evaluates the complete frozen primary+secondary registry on the common `THERMO-COMMON-M12` population at $m = 12, r = 100$; the six-protein fixed-setting transfer uses `FIXED6-M12` under the same construction rule. Complete-pool characterization scans $m \in \{2, 4, 8, 12\}$, and the previously frozen observable-specific pools are retained as a sensitivity control (Appendix E.7).

4.2 Reference thermodynamics and evaluation metrics

Slow side-chain free energies use statistically separate umbrella/MBAR reference calculations (Shirts and Chodera, 2008), with WHAM/Bennett cross-checks (Kumar et al., 1992; Bennett, 1976). Torsional coordinates, basin definitions, and evaluation observables are frozen before final method evaluation; Appendix E.6 gives the primary and secondary registries. Acceptance measures correction difficulty; ESJD and cross-mode transitions measure move quality; τ_{int} and ESS measure mixing; and ESS per force evaluation/GPU-second measure finite-budget efficiency. Statistical and execution efficiency combine as

$$\frac{\mathrm{ESS}}{\mathrm{GPU}\cdot\mathrm{s}} = \frac{\mathrm{ESS}}{N_E}\frac{N_E}{\mathrm{GPU}\cdot\mathrm{s}}, \tag{8}$$

where N_E counts all MD force evaluations and attempted-block target evaluations, including rejected proposals. Structural, thermodynamic, and multi-initialization diagnostics measure finite-budget ensemble convergence; Appendix K isolates the MD discretization contribution.

4.3 Baselines and matched compute

Baselines span MD, generic local proposals, per-system learned samplers (GFlowMC-adapted, Sequential BG), and frozen transferable controls. PROSE-local+MH and the public Robin-local+MH checkpoint remain external references inside the common local-block MH wrapper. The matched-training control is a Robin-style autoregressive Boltzmann-generator proposal retrained from scratch on the same proteins, trajectories, protein-disjoint split, block examples, model-selection protocol, and 39.4 GPU-h training budget as BLOCS; both methods use the same full-target MH wrapper and common-pool evaluation population. Appendix E records the matched-training aggregate and architecture provenance.

The main production chain alternates 2 fs Langevin-middle MD segments with exactly corrected block moves. Force-normalized accounting counts every full target/force evaluation, and ESS/GPU-s includes proposal inference, target evaluation, acceptance/rejection, and MD segments. A paired exact-composition benchmark replaces the finite-step dynamics leg with its Metropolized counterpart at the same 2 fs step size and $r = 100$, using identical initial configurations and evaluation checkpoints across methods. Appendices F, K, and E give the physical target, exact-composition protocol, and execution accounting.

5 Results

5.1 Block correction changes the scaling variable

The controlled experiment shows a strong separation over the tested ladder. Global MH acceptance falls from 2.0×10^{-3} at $N = 20$ to 3.1×10^{-9} at $N = 76$, whereas fixed-block BLOCS remains near 0.42. Equation equation 3 identifies forward/reverse pair overlap as the relevant acceptance quantity, and the scaling experiments vary total size and proposed dimension separately.

The complete-pool scan in Appendix O applies one frozen 21.4M-parameter proposal to Protein G, ubiquitin, and lysozyme at $r = 100$, averaging over the complete eligible pool for each $m \in \{2, 4, 8, 12\}$. Increasing m from 2 to 12 lowers acceptance by 71.0%, 71.1%, and 73.2%; at fixed m, increasing protein size from $N = 56$ to 129 lowers it by 10.4–17.4%. A correlation-length sweep gives the same interpretation: short-range dependence makes distant degrees of freedom weakly relevant to the selected conditional, whereas longer-range coupling reduces acceptance.

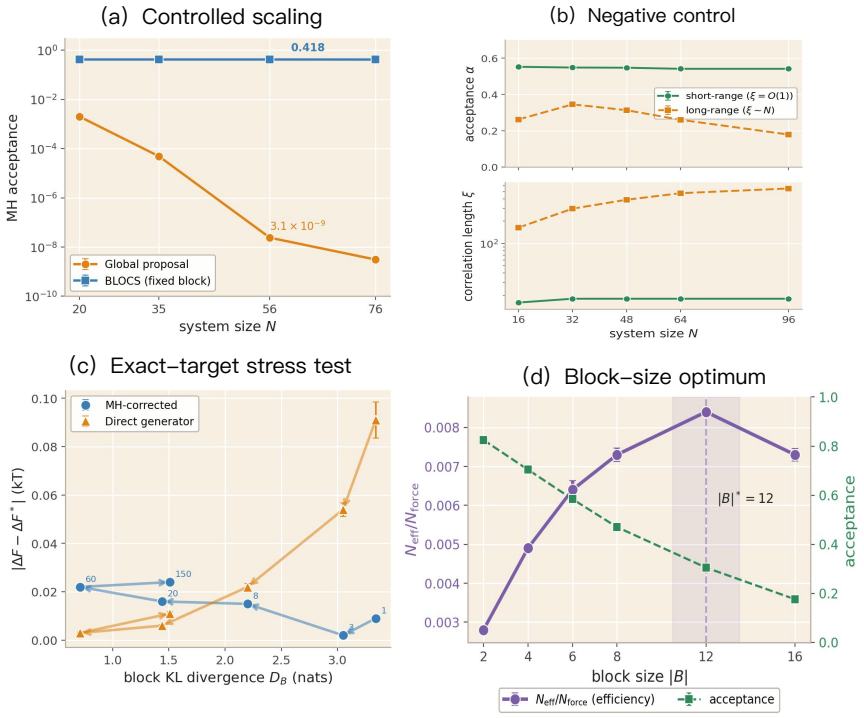


Figure 2: **Controlled scaling mechanism and block operating point.** Over the tested ladder, fixed-scope block correction avoids the acceptance collapse of the corresponding global proposal; correlation range, proposal quality, and block size jointly determine the operating regime.

The block-size sweeps expose the overlap–displacement tradeoff. Larger accepted moves initially raise ESS per target evaluation, while declining overlap limits further gains. The controlled aggregate sweep peaks near $m = 12$; held-out complete-pool scans peak at $m = 4$, 8, and 4 on Protein G, ubiquitin, and lysozyme. Proposal-degradation experiments further separate thermodynamic correctness from transition efficiency: after MH correction, worsening proposal quality primarily lowers acceptance, ESS, and convergence rate rather than shifting the target estimate (Appendix C).

Prospective tuning is evaluated on BPTI, barnase, RNase A, and calmodulin. The state-dependent Hastings-score calibration reaches 4.8% median acceptance error, and an equal 0.65 GPU-h pilot budget with Gaussian-overlap plus utility selection retains 0.90–1.00 of the independently validated best efficiency across the four systems. Direct acceptance estimation and short-run ESS provide complementary pilot references (Appendices H and Q).

5.2 A FROZEN LOCAL CONDITIONAL TRANSFERS ACROSS PROTEIN SIZE AND FOLD

A model trained on proteins with at most 36 residues remains effective through lysozyme ($N = 129$), beyond the training-size range and with decreasing sequence and structural similarity to training. Using the same frozen proposal, Appendix J extends the ladder to RNase H, T4 lysozyme, adenylate kinase, and a triosephosphate-isomerase monomer ($N = 250$): acceptance changes from 0.141 at $N = 129$ to 0.131 at $N = 250$, with a fitted log-acceptance slope of -6.0×10^{-4} residue^{-1}. Held-out soluble structural classes retain 0.54–0.76 of the BBA reference efficiency.

The complete-pool multi-DOF sweep independently separates total molecular size from moved dimension, again showing a larger acceptance penalty from increasing simultaneously moved torsions. A 3D context sweep saturates around $r_c^\star = 9.4$ Å. In a capacity-matched Protein-G ablation, ESS per 10^5 evaluations increases from 322 with current torsions to 397 with local sequence, 469 with backbone geometry, and 612 with full 3D context, while shuffled context falls to 301. The full-context gain over backbone-only conditioning remains 1.28–1.31× on ubiquitin and lysozyme (Appendix M).

This context dependence also separates BLOCS from generic local proposals. Von-Mises random walks and rotamer libraries encode broad torsional preferences; the learned proposal adapts the

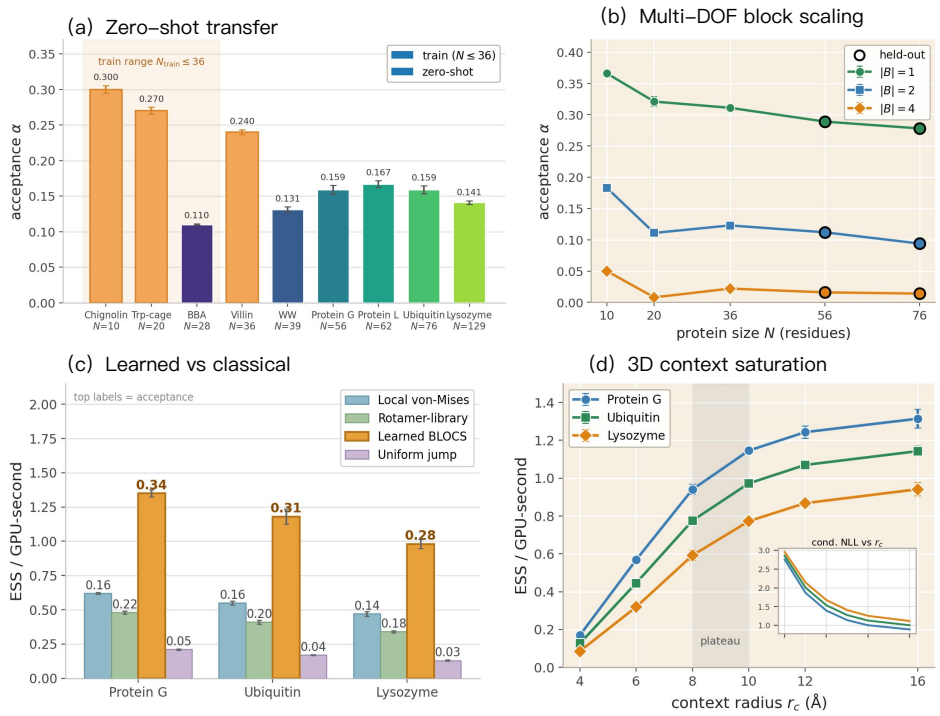


Figure 3: **Transferable local conformational modeling.** Zero-shot transfer extends beyond the training-size range; moved dimensionality dominates total size; context-aware proposals improve effective-sample throughput; and performance saturates at a finite 3D neighborhood.

multimodal update to the current 3D packing environment. A 2×2 ablation separates state-conditioned residual modeling from autoregressive cross-torsion coupling: ESS per 10^5 evaluations progresses from 284 (context-only, independent) to 418 (context-only, autoregressive), 493 (state-conditioned residual, independent), and 612 for the full residual-autoregressive model (Appendix R).

5.3 Faster local transitions become faster thermodynamic convergence

Across eight held-out χ_1/χ_2 observables, BLOCS has mean absolute free-energy error $0.055\,k_BT$ relative to independent umbrella/MBAR references. On Trp-cage TRP6 χ_2, matched trajectories use 6×10^4 force evaluations: MD records zero inter-state transitions, whereas BLOCS makes 316. Figure 4 connects this barrier crossing to lower finite-budget error and shows the separate statistical- and execution-efficiency trends across protein size.

Table 1 evaluates all 13 locked primary+secondary observables on one common pool population at $m = 12, r = 100$. The common population evaluates the predeclared thermodynamic coordinates with one block population per system, so the benchmark definition is independent of the focal observable. Across the five systems, BLOCS reduces system-level median error to 0.033–$0.055\,k_BT$ and improves end-to-end ESS/GPU-s by 2.69–$3.25\times$ over Rotamer MH and 1.67–$1.91\times$ over the matched-data autoregressive control. Appendix E.6 reports every one of the 13 coordinate-level outcomes; two coordinates favor the matched control on one metric, showing that the gain is aggregate rather than uniform.

The same acceleration persists when the interleaved dynamics leg is Metropolized. Table 2 summarizes paired five-seed runs on Trp-cage, Protein G, and lysozyme. BLOCS retains 2.79–$2.87\times$ higher ESS/GPU-s than Rotamer MH and 1.57–$1.68\times$ higher ESS/GPU-s than the matched autoregressive control; the sustained $0.10\,k_BT$ target is reached in median 3.1, 4.0, and 6.0 GPU-h, respectively. Appendix K gives the full per-method acceptance, error, ESS, throughput, IQR, and success counts.

The composed kernel also preserves global configurational coverage. In a matched-budget composition ablation, block moves alone give backbone ϕ/ψ JS divergence 0.214; the composed BLOCS kernel reduces it to 0.021 while retaining side-chain χ JS divergence 0.019 and 96.4% targeted slow-

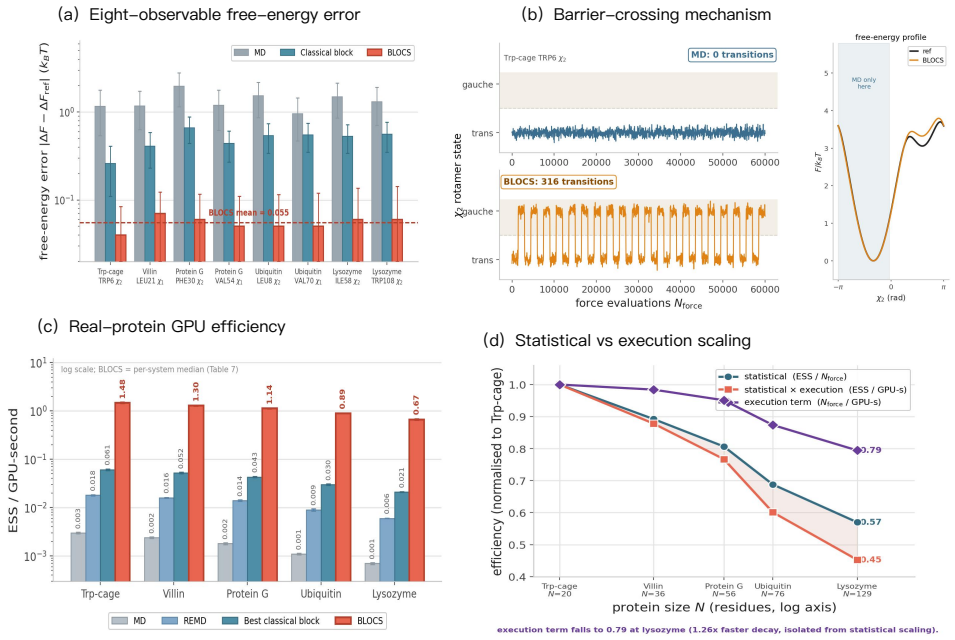


Figure 4: **Thermodynamic convergence, barrier crossing, and execution scaling. (a)** Absolute free-energy error for eight held-out slow observables; the dashed line marks the BLOCS mean error of $0.055\,k_BT$. **(b)** Trp-cage TRP6 χ_2 under 6×10^4 matched force evaluations: MD records 0 inter-state transitions and BLOCS records 316. **(c)** ESS/GPU-second across protein size at per-system operating points. **(d)** Within-BLOCS statistical-efficiency and execution-throughput factors.

Table 1: **Unified common-pool production benchmark at $m = 12, r = 100$.** All methods use the frozen `THERMO-COMMON-M12` population and five predefined seeds. Error is the median across each system's locked observables; ESS/GPU-s is end-to-end.

		Median $\lvert\Delta F - \Delta F^\star\rvert / k_BT \downarrow$			ESS/GPU-s ↑				
System	Obs.	Rot.	Match. AR	BLOCS	Rot.	Match. AR	BLOCS	B/R	B/AR
Trp-cage	2	0.273	0.109	**0.033**	0.582	0.990	**1.891**	3.25×	1.91×
Villin HP35	2	0.405	0.121	**0.038**	0.536	0.894	**1.673**	3.12×	1.87×
Protein G	3	0.470	0.129	**0.041**	0.487	0.805	**1.466**	3.01×	1.82×
Ubiquitin	3	0.550	0.141	**0.044**	0.421	0.676	**1.204**	2.86×	1.78×
Lysozyme	3	0.630	0.155	**0.055**	0.351	0.565	**0.944**	2.69×	1.67×

observable coverage. Experimental ubiquitin observables show the same finite-budget improvement: BLOCS reaches methyl-axis S^2 RMSD 0.038 and Pearson $r = 0.94$, compared with 0.064/0.86 for the classical block sampler and 0.091/0.78 for matched-budget MD; scalar couplings and χ_1 rotamer populations follow the same ordering (Appendix I).

5.4 Matched controls separate statistical gain from implementation throughput

The matched-data Robin-style autoregressive control isolates proposal design from training-data scale. Both methods receive the same proteins, trajectories, split, block examples, hyperparameter-selection protocol, 39.4 GPU-h training budget, common-pool population, full-target MH wrapper, and end-to-end accounting. Over all 13 locked observables, median thermodynamic error decreases from 0.136 to $0.042\,k_BT$. Across the five thermodynamic systems, median ESS per 10^5 target evaluations increases from 875 to 1264 (1.44×), median target-evaluation throughput from 92.0 to 116.0 GPU-s^{-1} (1.26×), and median ESS/GPU-s from 0.805 to 1.466 (1.82×). Appendix E gives the architecture provenance and aggregation rule; the public frozen Robin and PROSE checkpoints remain in the external learned-baseline suite.

Table 2: **Protein-scale exact-composition benchmark.** All rows use Metropolized dynamics at $\Delta t = 2$ fs, $r = 100$, `THERMO-COMMON-M12`, and paired $n = 5$ seeds. ESS/GPU-s values are means; BLOCS error is mean $\pm$ s.d.; time-to-target is median [IQR], success/5.

System	Rot.	Match. AR	BLOCS	B/R	B/AR	BLOCS error / GPU-h to $0.10\,k_BT$
Trp-cage	0.360	0.656	**1.032**	2.87×	1.57×	0.041 ± 0.008 / 3.1[2.8, 3.5]; 5/5
Protein G	0.230	0.396	**0.641**	2.79×	1.62×	0.052 ± 0.009 / 4.0[3.6, 4.5]; 5/5
Lysozyme	0.133	0.222	**0.374**	2.81×	1.68×	0.066 ± 0.011 / 6.0[5.5, 6.8]; 5/5

The canonical six-protein fixed-setting transfer adds Chignolin under the same $m = 12, r = 100$ construction rule. Across the six proteins, BLOCS improves ESS/GPU-s by 2.69–3.34× over Rotamer MH, with a median gain of 3.07× (Table 31). The complete-pool dimension scan, selected operating-point study, and cycle-level timing remain separate evaluations of block-size geometry, adaptive operating points, and implementation cost.

6 Discussion and Limitations

The results resolve the behavior of BLOCS across three complementary scales. First, complete-pool scans show that moved torsional dimension contributes a substantially larger acceptance penalty than total protein size, identifying the dimension of the corrected transition as the primary scaling variable in the evaluated regime. Second, the context and conditioning experiments show that finite local 3D information improves proposal–target overlap, accepted displacement, and temporal decorrelation. Third, exact full-system MH correction connects these learned local proposals to equilibrium sampling, while the common-pool thermodynamic benchmark shows that the resulting transition advantages translate into lower finite-budget free-energy error and higher effective-sample throughput.

The unified benchmark also separates several potential sources of gain. All 13 predefined primary and secondary observables are evaluated under the same common-pool construction at $m = 12, r = 100$, eliminating observable-specific block selection from the headline comparison. The matched-data autoregressive control uses the same training proteins, trajectories, split, block examples, training budget, MH wrapper, and target accounting, isolating the contribution of the BLOCS proposal parameterization from training-data scale. The paired exact-composition experiment further shows that the end-to-end advantage persists when the interleaved dynamics leg is Metropolized, linking the statistical gain of the learned block proposal to an exactly invariant composed sampling procedure.

The present experiments focus on soluble proteins under the AMBER ff14SB/GBn2 target, with explicit-water controls on Trp-cage and Protein G. The learned move acts on local side-chain torsional blocks while backbone and larger collective coordinates are supplied primarily by the interleaved dynamics kernel. Extending the same block-conditional construction to backbone rearrangements, solvent-coupled proposals, and larger collective coordinates would broaden the range of slow molecular processes addressed by the method. These extensions also provide a direct test of how the locality principle changes as the moved coordinates induce increasingly nonlocal Cartesian displacements.

7 Conclusion

BLOCS introduces transferable block-conditional generative proposals with exact full-system Metropolis–Hastings correction for local equilibrium sampling. Trained on proteins with $N \leq 36$, a single model transfers zero-shot to proteins with up to 250 residues. Under the canonical $m = 12, r = 100$ setting, the six-protein benchmark achieves 2.69–3.34× higher end-to-end ESS/GPU-s than Rotamer MH, with a median gain of 3.07×. On the five-system common-pool benchmark, BLOCS exceeds the matched-data autoregressive control by 1.67–1.91×, with a 1.82× median end-to-end gain.

With Metropolized dynamics, the paired exact-composition benchmark retains 2.79–2.87× gains over Rotamer MH and 1.57–1.68× over the matched learned control. Together with the 13-observable audit, complete-pool dimension scans, and frozen-pool diagnostics, these results show that transferable local generative blocks retain tractable statistical correction while substantially increasing independent equilibrium information per unit computation.

AI USE STATEMENT

Generative AI tools were used solely for language polishing of the manuscript, including improvements to grammar, wording, and readability. The authors reviewed and approved all revisions and take full responsibility for the final content.

REFERENCES

Alessandro Barducci, Giovanni Bussi, and Michele Parrinello. Well-tempered metadynamics: a smoothly converging and tunable free-energy method. *Physical review letters*, 100(2):020603, 2008.

Charles H Bennett. Efficient estimation of free energy differences from monte carlo data. *Journal of Computational Physics*, 22(2):245–268, 1976.

Weilong Chen, Bojun Zhao, Jan Eckwert, and Julija Zavadlav. Coarse-grained boltzmann generators. *arXiv preprint arXiv:2602.10637*, 2026.

James J Chou, David A Case, and Ad Bax. Insights into the mobility of methyl-bearing side chains in proteins from 3 j cc and 3 j cn couplings. *Journal of the American Chemical Society*, 125(29): 8959–8966, 2003.

Conor Durkan, Artur Bekasov, Iain Murray, and George Papamakarios. Neural spline flows. *Advances in neural information processing systems*, 32, 2019.

Naïl B Khelifa, Richard E Turner, and Ramji Venkataramanan. Diffusion models observe only gradients: A geometric perspective on score matching errors. *arXiv preprint arXiv:2606.06179*, 2026.

Joseph C Kim, David Bloore, Karan Kapoor, Jun Feng, Ming-Hong Hao, and Mengdi Wang. Scalable normalizing flows enable boltzmann generators for macromolecules. *arXiv preprint arXiv:2401.04246*, 2024.

Leon Klein and Frank Noé. Transferable boltzmann generators. *Advances in Neural Information Processing Systems*, 37:45281–45314, 2024.

Shankar Kumar, John M Rosenberg, Djamal Bouzida, Robert H Swendsen, and Peter A Kollman. The weighted histogram analysis method for free-energy calculations on biomolecules. i. the method. *Journal of computational chemistry*, 13(8):1011–1021, 1992.

Chieh-Hsin Lai, Yuhta Takida, Naoki Murata, Toshimitsu Uesaka, Yuki Mitsufuji, and Stefano Ermon. Fp-diffusion: Improving score-based diffusion models by enforcing the underlying score fokker-planck equation. In *International Conference on Machine Learning*, pages 18365–18398. PMLR, 2023.

Benedict Leimkuhler, Charles Matthews, and Gabriel Stoltz. The computation of averages from equilibrium and nonequilibrium langevin molecular dynamics. *IMA Journal of Numerical Analysis*, 36(1):13–79, 2016.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew YK Foong, Victor García Satorras, Osama Abdin, Bastiaan S Veeling, Iryna Zaporozhets, et al. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761): eadv9817, 2025.

Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, and David E Shaw. How fast-folding proteins fold. *Science*, 334(6055):517–520, 2011.

Balint Mate, Francois Fleuret, and Tristan Bereau. Neural thermodynamic integration: Free energies from energy-based diffusion models. *The Journal of Physical Chemistry Letters*, 15(45):11395–11404, 2024.

Antonio Mirarchi, Toni Giorgino, and Gianni De Fabritiis. mdcath: A large-scale md dataset for data-driven computational biophysics. *Scientific Data*, 11(1):1299, 2024.

Kirill Neklyudov, Evgenii Egorov, Pavel Shvechikov, and Dmitry Vetrov. Metropolis-hastings view on variational inference and adversarial training. *arXiv preprint arXiv:1810.07151*, 2018.

Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

Michael Plainer, Hao Wu, Leon Klein, Stephan Günnemann, and Frank Noé. Consistent sampling and simulation: Molecular dynamics with energy-based diffusion models. *Advances in Neural Information Processing Systems*, 38:24460–24505, 2026.

Danyal Rehman, Charlie B Tan, Yoshua Bengio, Avishek Joey Bose, and Alexander Tong. Autoregressive boltzmann generators. *arXiv preprint arXiv:2606.27361*, 2026.

James P Roney, Chenxi Ou, and Sergey Ovchinnikov. Protein diffusion models as statistical potentials. *bioRxiv*, pages 2025–12, 2025.

David E Shaw, Paul Maragakis, Kresten Lindorff-Larsen, Stefano Piana, Ron O Dror, Michael P Eastwood, Joseph A Bank, John M Jumper, John K Salmon, Yibing Shan, et al. Atomic-level characterization of the structural dynamics of proteins. *Science*, 330(6002):341–346, 2010.

Michael R Shirts and John D Chodera. Statistically optimal analysis of samples from multiple equilibrium states. *The Journal of chemical physics*, 129(12), 2008.

Christopher D Snow, Eric J Sorin, Young Min Rhee, and Vijay S Pande. How well can simulation predict protein folding kinetics and thermodynamics? *Annu. Rev. Biophys. Biomol. Struct.*, 34(1): 43–69, 2005.

Yuji Sugita and Yuko Okamoto. Replica-exchange molecular dynamics method for protein folding. *Chemical physics letters*, 314(1-2):141–151, 1999.

Charlie Tan, Majdi Hassan, Leon Klein, Saifuddin Syed, Dominique Beaini, Michael Bronstein, Alexander Tong, and Kirill Neklyudov. Amortized sampling with transferable normalizing flows. *Advances in Neural Information Processing Systems*, 38:94290–94325, 2026.

Charlie B Tan, Avishek Joey Bose, Chen Lin, Leon Klein, Michael M Bronstein, and Alexander Tong. Scalable equilibrium sampling with sequential boltzmann generators. *arXiv preprint arXiv:2502.18462*, 2025.

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic acids research*, 52(D1):D384–D392, 2024.

Christopher von Klitzing, Denis Blessing, Henrik Schopmans, Pascal Friederich, and Gerhard Neumann. Learning boltzmann generators via constrained mass transport. In *International Conference on Learning Representations*, volume 2026, pages 88771–88810, 2026.

A Joshua Wand, Jeffrey L Urbauer, Robert P McEvoy, and Ramona J Bieber. Internal dynamics of human ubiquitin revealed by 13c-relaxation studies of randomly fractionally labeled protein. *Biochemistry*, 35(19):6116–6125, 1996.

Andy C Wang and Ad Bax. Reparametrization of the karplus relation for 3j (h. alpha.-n) and 3j (hn-c') in peptides from uniformly 13c/15n-enriched human ubiquitin. *Journal of the American Chemical Society*, 117(6):1810–1813, 1995.

Baoli Wang, Chenglin Wang, Jingyang Chen, Danlin Liu, Changzhi Sun, Jie Zhang, Kai Zhang, and Honglin Li. Conditional diffusion with locality-aware modal alignment for generating diverse protein conformational ensembles. *Nature Machine Intelligence*, 8(3):415–434, 2026.

APPENDIX CONTENTS

A Transition Kernel, Detailed Balance, and Acceptance Geometry 13
B Block Dimension, Context, and the Sampling Operating Regime 14
C Numerical Consistency, Equilibrium Recovery, and Proposal Design 16
D Zero-Shot Transfer from Local Calibration to Thermodynamics 18
E Architecture, Baselines, Compute, and Reproducibility 20
F Physical Target, Kernel Composition, and Solvent Controls 28
G Classical Baselines and Learned-Conditioning Controls 30
H Prospective State-Dependent Locality Diagnostics 32
I Experimental NMR Validation 34
J Extended Size and Structural-Class Transfer 34
K Composite-Kernel Step Size and Stationary Accuracy 36
L Coordinate-Transform Validation for Production Torsion Moves 37
M Capacity-Matched Conditioning-Information Ablation 38
N Cycle-Level Runtime and Execution Decomposition 39
O Cross-Protein Scan at Fixed Torsional Dimensionality 40
P Finite-Budget Convergence from Multiple Initial States 41
Q Equal-Pilot-Budget Block Selection 42
R State Conditioning and Autoregressive Factorization 43

Appendices A–J develop transition theory, locality geometry, numerical validation, transfer, physical controls, baselines, prospective diagnostics, NMR validation, and extended transfer. Appendices K–R add the targeted composite-kernel, coordinate-transform, capacity-matched context, cycle-level runtime, fixed-m, multi-initialization, equal-pilot, and proposal-factorization experiments.

A Transition Kernel, Detailed Balance, and Acceptance Geometry

A.1 State-dependent block proposals

Work in the internal-coordinate chart ξ defined in Section 3.1, with target density

$$\widetilde{\pi}(\xi) = Z^{-1} e^{-\beta U(T(\xi))} J(\xi)$$

with respect to $d\xi$. For a fixed block B,

$$Q_B(\xi, d\xi') = q_\theta(\xi'_B \mid \xi_B, c_B(\xi))\, d\xi'_B\, \delta_{\xi_{\bar{B}}}(d\xi'_{\bar{B}}).$$

The accepted transition kernel is

$$K_B(\xi, d\xi') = \alpha(\xi, \xi') Q_B(\xi, d\xi') + \left[1 - \int \alpha(\xi, z) Q_B(\xi, dz)\right] \delta_\xi(d\xi'),$$

with α given by Eq. equation 2. For $\xi \neq \xi'$,

$$\widetilde{\pi}(\xi) K_B(\xi, \xi') = \min\{\widetilde{\pi}(\xi) Q_B(\xi, \xi'), \widetilde{\pi}(\xi') Q_B(\xi', \xi)\},$$

which is symmetric in (ξ, ξ'). Thus each block kernel satisfies detailed balance and preserves Π after mapping through T; any state-independent mixture over blocks preserves the same target. When the interleaved dynamics proposal is Metropolized, the composition preserves Π exactly. The main production runs use finite-step Langevin-middle MD, whose stationary discretization error is quantified in Appendix K; Appendix K also reports the paired protein-scale Metropolized benchmark.

A.2 Stationary acceptance for the deployed state-dependent kernel

Define

$$M_{\theta,B}(d\xi, d\xi') = \widetilde{\pi}(\xi)\, d\xi\, Q_{\theta,B}(\xi, d\xi'), \qquad M^\dagger_{\theta,B}(d\xi, d\xi') = M_{\theta,B}(d\xi', d\xi).$$

Applying the general MH acceptance/TV identity of Neklyudov et al. (2018) to these block transition-pair measures gives

$$\bar{\alpha}_B = \int \min\{dM_{\theta,B}, dM^\dagger_{\theta,B}\} = 1 - D_{\mathrm{TV}}\left(M_{\theta,B}, M^\dagger_{\theta,B}\right).$$

Writing the log Hastings score as

$$S_\theta(\xi, \xi') = -\beta[U(T(\xi')) - U(T(\xi))] + \log \frac{J(\xi')}{J(\xi)} + \log q_\theta(\xi_B \mid \xi'_B, c_B(\xi')) - \log q_\theta(\xi'_B \mid \xi_B, c_B(\xi)),$$

gives

$$D_{\mathrm{tr},B} = D_{\mathrm{KL}}\left(M_{\theta,B} \| M^\dagger_{\theta,B}\right) = -\mathbb{E}_{M_{\theta,B}}[S_\theta(\Xi, \Xi')].$$

No block-conditional partition function is required. Full-system energy differences, the coordinate-measure ratio, and explicit forward/reverse proposal densities are sufficient. Pinsker's inequality yields $\bar{\alpha}_B \geq 1 - \sqrt{D_{\mathrm{tr},B}/2}$, while the empirical Gaussian Hastings-score approximation used in Appendix H gives $\bar{\alpha}_B \approx 2\Phi(-\sigma_S/2)$. For a frozen pool $\mathcal{B}$ with state-independent selection probabilities p_B, the reported statistics are

$$\bar{\alpha}_{\mathcal{B}} = \sum_{B \in \mathcal{B}} p_B \bar{\alpha}_B, \qquad D_{\mathrm{tr},\mathcal{B}} = \sum_{B \in \mathcal{B}} p_B D_{\mathrm{tr},B}.$$

Uniform selection gives $p_B = 1/|\mathcal{B}|$. The pooled Hastings-score sample supplies the Gaussian diagnostic for that same population. Thus the frozen pool is part of the definition of each acceptance and calibration result.

A.3 Stationary acceptance in the independence special case is controlled by weight dispersion

For the independence-proposal case, let π_B be the exact conditional of the moved block, q_B the proposal, and

$$w(y) = \frac{\pi_B(y)}{q_B(y)}.$$

At stationarity,

$$\bar{\alpha} = \iint \min\{\pi_B(y)q_B(y'), \pi_B(y')q_B(y)\}\, dy\, dy' \tag{9}$$

$$= \mathbb{E}_{q_B \times q_B} \min\{w(Y), w(Y')\} \tag{10}$$

$$= 1 - \frac{1}{2}\mathbb{E}_{q_B \times q_B}|w(Y) - w(Y')|. \tag{11}$$

The last line gives a direct interpretation of correction difficulty. A proposal is efficient when configurations drawn from q_B receive comparable target-to-proposal ratios across the relevant modes. A proposal can therefore match the visible locations of dominant modes yet still produce poor MH transitions if it assigns highly uneven relative weight across them. Conversely, a proposal with moderate local density error can remain useful when its importance ratios have limited dispersion.

This perspective also clarifies why the main experiments report several complementary diagnostics. Conditional NLL measures predictive fit, PIT statistics probe probabilistic calibration, KL summarizes average density mismatch, and acceptance measures the resulting pairwise weight competition in the MH kernel. None of these quantities is redundant: together they describe how a learned conditional becomes a corrected transition.

A.4 A useful local calibration

The empirical block-weight distributions are well described over the principal operating regime by

$$\log w \sim \mathcal{N}(-\sigma^2/2, \sigma^2), \qquad D_B = D_{\mathrm{KL}}(\pi_B \| q_B) = \sigma^2/2,$$

which gives

$$\bar{\alpha} = 2\Phi(-\sqrt{D_B}).$$

Across the controlled experiments, this approximation predicts acceptance with a median relative error of 8.7%. Its practical value is that it converts a local density statistic into an expected correction rate and therefore lets proposal quality be interpreted in the same units as the Markov transition.

The exact identity remains the more general object. For

$$q_\varepsilon = (1-\varepsilon, \varepsilon), \qquad \pi = (1/2, 1/2),$$

we have

$$D_{\mathrm{KL}}(\pi \| q_\varepsilon) = \frac{1}{2}\log\frac{1}{4\varepsilon(1-\varepsilon)} \to \infty, \qquad \bar{\alpha}_\varepsilon = \frac{1}{2} + \varepsilon \to \frac{1}{2}.$$

Thus the usefulness of the log-normal formula comes from the observed geometry of the local weight distribution. This distinction becomes important at large block sizes, where the measured KL curve bends away from the approximately linear small-block regime.

B Block Dimension, Context, and the Sampling Operating Regime

B.1 Why the useful block size is an interior optimum

Suppose the block mismatch is locally approximated by $D_B \approx md$, where m is moved dimension and d is an effective mismatch density. If accepted displacement grows with m and proposal cost scales as m^{η_m}, a simplified efficiency model is

$$\mathcal{E}(m) \propto m^{1-\eta_m}\, 2\Phi(-\sqrt{md}).$$

The first factor rewards moving more slow degrees of freedom; the second penalizes the loss of proposal–target overlap. This structure predicts an interior operating point whenever the displacement gain initially dominates but the overlap penalty eventually grows faster.

The controlled aggregate block sweep shows this competition. Acceptance falls as more coordinates are updated, while larger accepted moves initially raise ESS per target evaluation. The aggregate curve peaks near $m = 12$. The complete-pool protein scans resolve system-specific maxima at $m = 4$, 8, and 4 (Table 51), linking the operating point to both accepted displacement and the sampled conditional population.

B.2 Nonlinear block mismatch refines the simple scaling model

Table 3 combines the measured operating-regime quantities with their evaluation protocols. In the complete-pool scan, increasing N from 56 to 129 reduces acceptance by 10.4–17.4% at fixed m; increasing m from 2 to 12 reduces it by 71.0–73.2% at fixed protein. The deployed state-dependent diagnostic has 4.8% median relative error over the declared frozen pools, and the context sweep saturates at $r_c^\star = 9.4$ Å. The controlled independence-proposal calibration in Appendix A has 8.7% median relative error.

The controlled aggregate block-size model resolves the nonlinear mismatch contribution. A linear-KL extrapolation predicts $m = 21.9$, while the aggregate bootstrap estimate is $m = 11.7$. Figure 5 shows the source of the shift: the KL curve becomes super-linear as coupled torsions and packing constraints accumulate. The held-out complete-pool scans then locate the protein-specific optima at $m = 4$, 8, and 4.

Table 3: **Measurements defining the BLOCS operating regime.** Each row identifies the evaluated population or control; endpoint changes are computed from Table 51.

Quantity	Measurement	Evaluation protocol	Mechanistic implication
System-size dependence	Acceptance decreases 10.4–17.4% from $N = 56$ to 129 at fixed m	Complete eligible-pool scan; Table 51	The size penalty is smaller than the 71.0–73.2% penalty from $m = 2 \to 12$
Acceptance calibration	Median relative error $= 4.8\%$	Declared frozen pools; Table 38	The transition-level diagnostic predicts the deployed correction scale
Context saturation	$r_c^\star = 9.4$ Å	Capacity-matched context sweep	Transferable conditional information concentrates in a finite neighborhood
Block-size operating points	Aggregate bootstrap estimate 11.7; held-out optima $m = 4, 8, 4$	Controlled aggregate model; complete-pool Protein G/ubiquitin/lysozyme scans	The operating point depends on the sampled conditional population
Target-shift calibration	$\lvert\alpha_{\mathrm{pred}} - \alpha_{\mathrm{obs}}\rvert = 0.0186$	Independent shifted-target control	Density shift changes the correction difficulty
End-to-end gain	2.69–3.34×; median 3.07×	Six-protein transfer at $m = 12, r = 100$; Table 31	The efficiency gain persists across the fixed-setting protein ladder

B.3 Transition-divergence diagnostic surface

Table 4 holds the 21.4M-parameter model family, target family, training budget, and inference implementation fixed while varying total size N and moved torsional dimension m. The α_{Gauss} column is the Gaussian diagnostic computed from $\widehat{D}_{\mathrm{tr}}$. Independent observed acceptance across (N, m) is reported in Appendix O.

The diagnostic surface maps transition divergence to Gaussian-predicted acceptance. Table 51 supplies the independently observed complete-pool acceptance grid used for the size-versus-dimension contrasts.

B.4 Move geometry explains why side-chain blocks are favorable

Figure 5 combines block KL, context width, the efficiency proxy, local correlation structure, and move-design comparisons. These panels answer a single geometric question: what makes a proposed angular move remain local in the *molecular* sense?

Table 4: Transition-divergence diagnostic surface across total system size N and moved torsional dimension m. Architecture, parameter count, target family, training budget, and inference implementation are fixed. $\alpha_{\mathrm{Gauss}} = 2\Phi(-\sqrt{\widehat{D}_{\mathrm{tr}}/2})$ is derived from $\widehat{D}_{\mathrm{tr}}$; observed acceptance is reported separately in Table 51.

N	Proposed DOF m	Params	Training budget	$\widehat{D}_{\mathrm{tr}}$	α_{Gauss}	ESS/N_{force}
20	4	21.4M	39.4 GPU-h	0.581	0.590	0.0056
20	8	21.4M	39.4 GPU-h	1.141	0.450	0.0078
20	12	21.4M	39.4 GPU-h	1.978	0.320	0.0088
20	16	21.4M	39.4 GPU-h	3.285	0.200	0.0079
36	4	21.4M	39.4 GPU-h	0.612	0.580	0.0050
36	8	21.4M	39.4 GPU-h	1.193	0.440	0.0069
36	12	21.4M	39.4 GPU-h	2.061	0.310	0.0077
36	16	21.4M	39.4 GPU-h	3.435	0.190	0.0068
56	4	21.4M	39.4 GPU-h	0.645	0.570	0.0042
56	8	21.4M	39.4 GPU-h	1.246	0.430	0.0058
56	12	21.4M	39.4 GPU-h	2.148	0.300	0.0065
56	16	21.4M	39.4 GPU-h	3.595	0.180	0.0056
76	4	21.4M	39.4 GPU-h	0.679	0.560	0.0038
76	8	21.4M	39.4 GPU-h	1.301	0.420	0.0053
76	12	21.4M	39.4 GPU-h	2.239	0.290	0.0059
76	16	21.4M	39.4 GPU-h	3.766	0.170	0.0051

A side-chain torsion primarily reorganizes atoms near the selected residue. A backbone torsion, by contrast, rotates the downstream chain and therefore converts one angular degree of freedom into a large Cartesian displacement involving many atoms. The energy change and reverse-density problem therefore grow much faster for backbone updates even when the nominal number of moved torsions is the same. The architectural emphasis on local side-chain and residual torsion updates follows from this geometry.

The context-width panel provides the complementary representation result. Adding local information initially improves conditional likelihood and efficiency, but the marginal gain decreases once the neighborhood covers the dominant packing interactions. The sequence-correlation panel shows that local backbone and packing motifs recur even when global sequence and fold differ. These observations motivate the finite-context interpretation; the capacity-matched conditioning and shuffled-context experiments in Appendix M directly resolve the contribution of 3D context relative to local sequence and backbone geometry.

C Numerical Consistency, Equilibrium Recovery, and Proposal Design

C.1 From density evaluation to equilibrium recovery

The numerical validation suite in Table 5 spans three levels of the transition. Density normalization, analytic Jacobians, and inversion test the local proposal representation. Alanine-dipeptide equilibrium recovery tests the complete repeated MH kernel. The PIT statistic tests whether the conditional distribution is calibrated beyond a single mode or average-energy summary.

The measured density-normalization error is 3.7×10^{-4}, analytic log-determinants agree with autodiff to 2.1×10^{-7} relative error, and inversion error is 4.6×10^{-8}. These numerical scales are substantially smaller than the protein-level statistical differences studied later, placing density arithmetic well below the scale of the sampling effects.

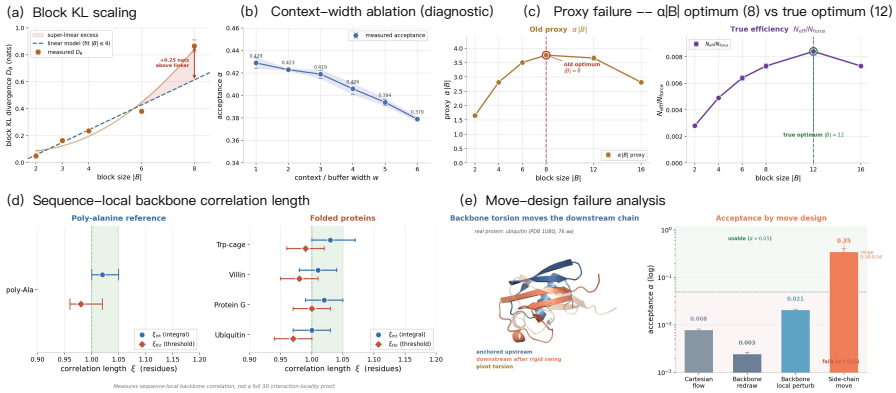


Figure 5: **Block mismatch, context dependence, and move geometry. (a)** Block-KL scaling, **(b)** context-width ablation, **(c)** the mismatch between the αm proxy optimum and the measured efficiency optimum, **(d)** sequence-local backbone correlation length, and **(e)** move-design analysis. Together the panels show why the useful transition regime is set by local coupling rather than total protein dimension.

At the transition level, alanine dipeptide reaches TV $= 0.0047$ with maximum free-energy-surface bin error $0.059\, k_B T$. Deliberately weakening the proposal changes the efficiency of proposed moves but still yields TV $= 0.0062$ after correction. The similar stationary errors despite different proposal quality separate target correctness from transition efficiency: model fit primarily changes the rate at which the target is explored.

Table 5: **Numerical consistency and equilibrium-recovery tests.** Each row probes a different stage of the learned-proposal-to-MH pipeline.

Test	Definition	Reference tolerance	Measured result
Density normalization	Monte-Carlo integral of q_θ on low-dimensional targets	relative error $< 10^{-3}$	3.7×10^{-4}
Analytic log-det	Analytic Jacobian log-det vs. full autodiff	relative error $< 10^{-6}$	2.1×10^{-7}
Invertibility	$\|T^{-1}(T(z)) - z\|_\infty$	$< 10^{-6}$	4.6×10^{-8}
Detailed balance	Alanine-dipeptide stationary distribution	TV < 0.01; max FES-bin error $< 0.1 k_B T$	TV $= 0.0047$; max error $= 0.059 k_B T$
Proposal degradation	Deliberately under-trained proposal followed by the same MH transition	target-distribution TV < 0.01	TV $= 0.0062$
Conditional calibration	Probability-integral-transform uniformity	KS statistic < 0.05	KS $= 0.0174$

C.2 Training objective and reversible architecture affect transition quality differently

Figure 6 goes beyond numerical consistency and asks which model properties determine a useful corrected transition. The training-objective comparison is particularly revealing. Forward-KL and reverse-KL training produce nearly identical marginal TV statistics (0.084 versus 0.085), yet their MH acceptance differs sharply (0.352 versus 0.073). Final marginal fit and transition utility therefore capture distinct aspects of proposal quality. Efficient MH transitions require proposal mass to be arranged so that both forward and reverse moves receive compatible density support.

The architecture comparison reaches the same conclusion from a different direction. Both routes are numerically reversible, with acceptance 0.416 for the residual-autoregressive route and 0.055 for the alternative. Reversibility supplies the correction algebra, while geometry aligned with the conditional molecular distribution supplies transition utility.

The detailed-balance residual panel shows stable transition symmetry across molecular energy changes. The proposal-degradation panel then closes the loop: as model quality worsens, direct generator thermodynamics drift, whereas the corrected chain continues to track the target while mixing slows. Together the four panels separate numerical correctness, marginal fit, and transition efficiency as distinct properties of a learned sampler.

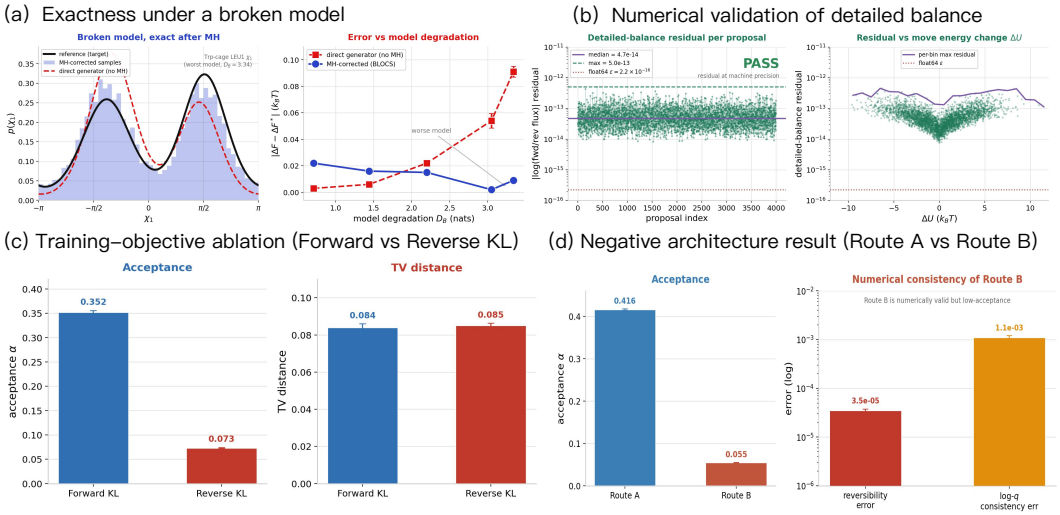


Figure 6: **Equilibrium recovery and proposal-design ablations.** The panels compare target recovery under model degradation, detailed-balance residuals, alternative training objectives, and reversible proposal architectures.

D Zero-Shot Transfer from Local Calibration to Thermodynamics

D.1 Calibration degrades smoothly with held-out protein complexity

Table 6 tracks conditional likelihood, PIT calibration, interval coverage, and downstream free-energy error across increasingly challenging held-out proteins. From BBA to lysozyme, conditional NLL rises from 1.12 to 1.46, PIT–KS from 0.0157 to 0.0317, and 95% coverage changes from 94.8% to 92.2%. Free-energy error rises in the same order from 0.087 to $0.185\,k_BT$.

The coupling among these metrics is the key pattern: conditional calibration and thermodynamic error change smoothly across the training-size boundary. This behavior tracks the reuse of recurring local environments, with the capacity-matched conditioning experiment in Appendix M quantifying the contribution of 3D context.

D.2 Global novelty grows faster than the loss of sampling efficiency

The novelty analysis in Table 7 separates transfer from close global homology. Nearest-training sequence identity falls from 18.2% to 9.7%, a reduction of roughly 47%, while nearest-training TM-score falls from 0.43 to 0.275, roughly 36%. Retained sampling efficiency decreases from 0.84 to 0.64, only about 24% over the same sequence.

The different rates separate global sequence/fold similarity from the transferable local signal. A substantial fraction of in-domain efficiency survives after both global similarity measures fall sharply,

Table 6: **Probabilistic calibration on held-out proteins.** Conditional model quality and thermodynamic error evolve smoothly across the zero-shot ladder.

Held-out system	Conditional NLL	PIT–KS ↓	95% interval coverage (%) ↑	$\lvert\Delta F - \Delta F^{\star}\rvert / k_B T$ ↓
BBA mini-protein	1.12	0.0157	94.8	0.087
WW domain	1.17	0.0187	94.1	0.105
Villin HP35	1.24	0.0207	93.7	0.125
Protein G	1.31	0.0237	93.4	0.137
Ubiquitin	1.37	0.0267	92.8	0.145
Lysozyme	1.46	0.0317	92.2	0.185

while the capacity-matched context ablation in Appendix M attributes the additional gain to 3D packing information beyond local sequence and backbone geometry.

Free-energy error remains below $0.2\,k_BT$ throughout the novelty ladder. The same ordering is visible in the calibration table, linking representation novelty to thermodynamic behavior through conditional model quality across the train/test boundary.

Table 7: **Zero-shot transfer as sequence and fold novelty increase.** Global similarity to the training set decreases more rapidly than retained sampling efficiency.

Held-out system	Nearest train sequence identity (%) ↓	Nearest train TM-score ↓	$\lvert\Delta F - \Delta F^{\star}\rvert / k_B T$ ↓	Efficiency retained vs. in-domain ↑
BBA mini-protein	18.2	0.43	0.087	0.84
WW domain	15.6	0.385	0.105	0.785
NTL9	13.9	0.365	0.125	0.76
a3D	12.4	0.32	0.145	0.715
Protein L	11.8	0.307	0.165	0.685
Lysozyme	9.7	0.275	0.185	0.64

D.3 Multi-DOF mixing reveals how transfer difficulty enters the dynamics

Table 8 shows the dynamic consequence of increasing local packing complexity. From Chignolin to ubiquitin, τ_{int} rises from 6.6 to 10.4, ESS/N_{force} falls from 0.0096 to 0.0059, and ESJD per accepted move falls from 2.83 to 1.49. Cross-mode transitions decrease from 612 to 356 per 10^5 force evaluations, but remain in the hundreds even for the largest system in the table.

These metrics move together in a physically coherent way. More restrictive environments reduce the distance covered by accepted moves; smaller displacement lowers the frequency of useful mode exchange; temporal correlation therefore increases and force-normalized ESS decreases. The independently evaluated pilot selector chooses $m = 4$ on Chignolin, Trp-cage, and Villin HP35 and $m = 8$ on Protein G and ubiquitin. The sampler therefore balances accepted displacement against overlap rather than monotonically shrinking the moved block with total protein size.

The preserved transition rate shows that larger proteins remain on the same smooth mixing continuum. This gradual dynamic degradation mirrors the smooth changes in NLL, calibration, novelty, and free-energy error above.

Table 8: **Multi-DOF mixing at the selected operating point.** Increasing local packing complexity reduces displacement and ESS gradually while preserving repeated slow-mode exchange.

System	Selected m	τ_{int} ↓	ESS/N_{force} ↑	ESJD per accepted move ↑	Cross-mode transitions/10^5 force evals ↑
Chignolin	4	6.6	0.0096	2.83	612
Trp-cage	4	7.3	0.0088	2.41	548
Villin HP35	4	8.1	0.0077	2.06	471
Protein G	8	9.2	0.0065	1.72	403
Ubiquitin	8	10.4	0.0059	1.49	356

D.4 A second free-energy suite links mixing to equilibrium error

The secondary benchmark in Table 9 contains eight residue-level observables distinct from the main table. Across the suite, the absolute BLOCS error ranges only from 0.04 to $0.07\,k_BT$, with mean

absolute error $0.055\,k_BT$. The corresponding mean errors are $0.494\,k_BT$ for the classical block sampler and $1.351\,k_BT$ for MD. Thus the learned conditional reduces mean finite-budget free-energy error by roughly 89% relative to the classical block baseline on this independent set.

The mixing columns explain the error trend. τ_{int} increases gradually from 5.8 to 10.1, while ESS/N_{force} and ESS/GPU-s decrease smoothly with protein complexity. Free-energy error grows more slowly because the trajectories continue to accumulate independent samples that resolve the two-state populations; larger systems therefore require more computation for the same precision while following the same convergence mechanism.

The comparison with MD is especially clear on slow rotameric coordinates. MD and the classical block sampler can remain on the wrong side of the reference by large fractions of k_BT within the matched budget, whereas BLOCS repeatedly crosses the corresponding barriers. The table therefore closes the chain from learned conditional quality to mode exchange, temporal decorrelation, and equilibrium estimation.

Table 9: **Secondary held-out free-energy benchmark.** The suite is independent of the primary main-text table and contains both χ_1 and χ_2 observables.

Protein	Residue	Observable	MBAR ref.	MD	Classical block	BLOCS	τ_{int}	ESS/N_{force}	ESS/GPU-s
Trp-cage	TRP6	χ_2	-0.42	0.74	-0.16	$\mathbf{-0.38 \pm 0.044}$	5.8	0.0093	1.48
Villin HP35	LEU21	χ_1	1.06	2.24	1.47	$\mathbf{1.13 \pm 0.053}$	6.4	0.0087	1.34
Protein G	PHE30	χ_2	-1.74	0.23	-1.08	$\mathbf{-1.68 \pm 0.057}$	7.1	0.0081	1.21
Protein G	VAL54	χ_1	0.38	1.58	0.82	$\mathbf{0.43 \pm 0.063}$	7.6	0.0075	1.06
Ubiquitin	LEU8	χ_2	-0.88	0.65	-0.34	$\mathbf{-0.83 \pm 0.067}$	8.3	0.0069	0.93
Ubiquitin	VAL70	χ_1	1.41	2.37	1.96	$\mathbf{1.36 \pm 0.074}$	8.9	0.0063	0.85
Lysozyme	ILE58	χ_2	-1.12	0.38	-0.59	$\mathbf{-1.06 \pm 0.077}$	9.5	0.0058	0.75
Lysozyme	TRP108	χ_2	0.58	1.89	1.14	$\mathbf{0.64 \pm 0.083}$	10.1	0.0053	0.67

D.5 Transfer across temperature, force field, size, and fold

Figure 7 perturbs the learned conditional along several independent axes. Temperature transfer changes the Boltzmann weighting while largely preserving molecular geometry. The held-out-temperature free-energy relation remains aligned with the reference, showing that the proposal retains a useful ordering of local states under a changed thermodynamic weight.

The force-field experiment changes the target more directly. Block KL increases from 0.69 to 1.23, and acceptance decreases in parallel. The target shift appears in the same local mismatch quantity that predicts the matched-target operating regime, linking distribution shift to correction difficulty through the same local mechanism.

The single-χ_1 size ladder and fold-level zero-shot panels extend this result to larger and structurally distinct proteins. Acceptance decreases gradually across the ladder, consistent with the calibration and novelty tables. The global-reweighting stress test shows global ESS and global MH acceptance deteriorating rapidly with system size while local correction remains in a usable regime. Across all five perturbations, the same scale separation explains the behavior.

E Architecture, Baselines, Compute, and Reproducibility

E.1 The final architecture matches the measured locality of the task

Table 10 and Figure 8 describe the final residual autoregressive proposal. The 9 Å context radius closely matches the independently measured $r_c^\star = 9.4$ Å saturation point. The architecture therefore allocates representation capacity to approximately the same spatial scale that the transfer experiments identify as informative.

The mixture-of-von-Mises heads address a different property of the molecular conditional. Torsions are periodic, and side-chain distributions are commonly multimodal because several rotamers can remain physically plausible in the same neighborhood. Twelve mixture components let one conditional represent multiple periodic modes, while autoregressive factorization captures coupling among simultaneously updated torsions.

The architecture is therefore aligned with the transition kernel at every stage. Current torsions define the local state; the 3D encoder summarizes packing constraints; the causal stack models

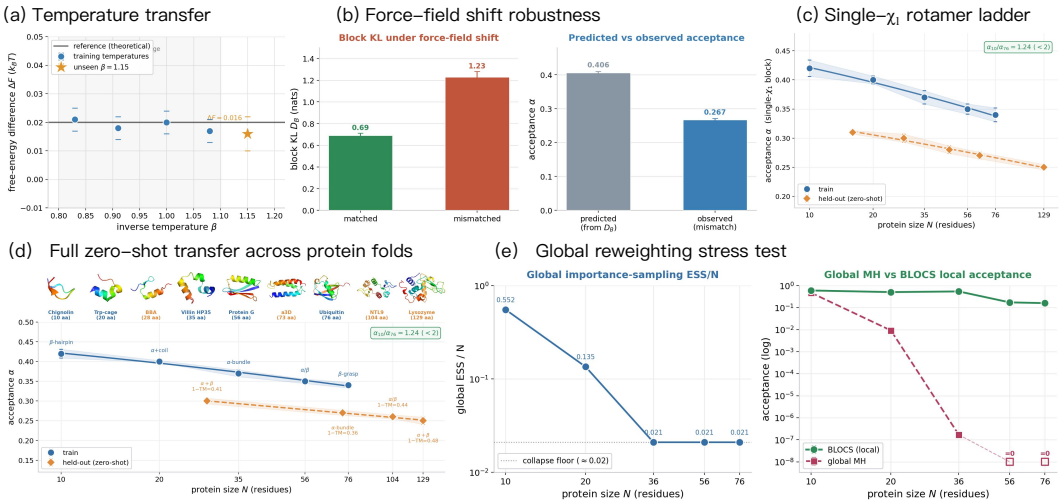


Figure 7: **Transfer across thermodynamic conditions, force fields, size, fold, and correction scale.** **(a)** Temperature transfer, **(b)** force-field-shift robustness, **(c)** the single-χ_1 size ladder, **(d)** fold-level zero-shot transfer, and **(e)** the global-reweighting stress test. The corrected panel sequence separates each transfer axis and contrasts global overlap collapse with usable local correction.

dependence among proposed residuals; and the explicit mixture density supplies both forward and reverse probabilities to MH. This alignment explains why a comparatively compact 21.4M-parameter model can be reused across proteins instead of representing a complete global ensemble.

Table 10: **Final learned-proposal configuration.**

Field	Final setting
Architecture	Residual autoregressive conditional proposal with mixture-of-von-Mises outputs
Layers / hidden width	10 / 384
Mixture components per torsion	12
3D context radius	9 Å
Optimizer	AdamW
Learning rate	2.5×10^{-4}
Batch size / epochs	1536 / 72
Parameter count	21.4M
One-time training cost	39.4 GPU-h on 4×A100
Early stopping	held-out conditional NLL plus target-distribution validation

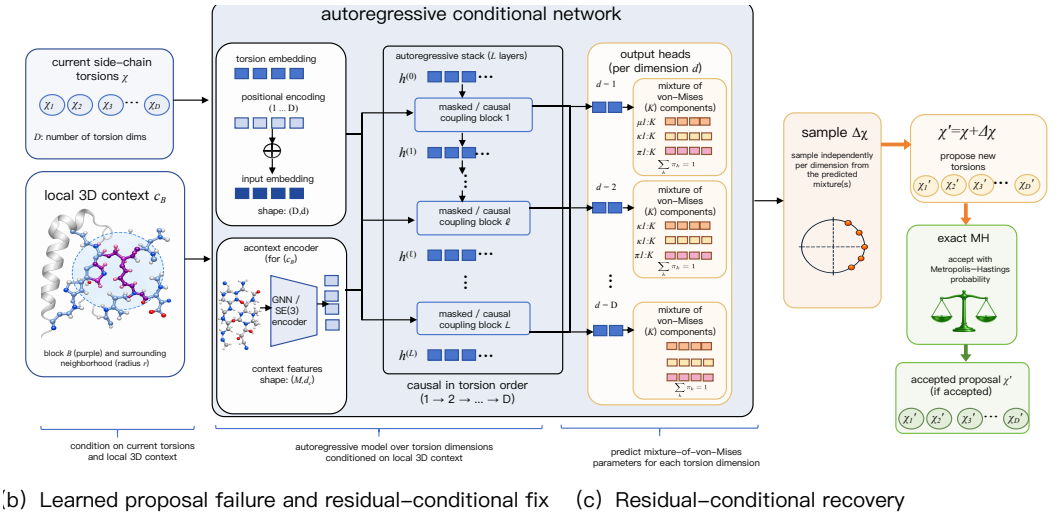
E.2 The baseline ladder separates locality, learning, and transfer

Table 11 separates classical moves, per-system learned samplers, and frozen transferable controls. PROSE-local+MH and Robin-local+MH compare public pretrained representations inside the common local-MH execution path.

Within this wrapper, the BLOCS/Robin-local+MH force-normalized ratio is $1.26\times$ and the ESS/GPU-s ratio is $2.03\times$. Table 20 resolves the corresponding accounting factors.

Table 11: **Common benchmark configuration across classical and learned proposal families.**

Method	Proposal family	Transfer setting	Correction	Tuning	Training policy	Inference budget
MD	Langevin dynamics	per system	finite-step Langevin-middle	timestep/friction	none	matched GPU time
Local von-Mises	local torsion random walk	universal	MH	proposal scale	none	matched GPU time
Rotamer library	empirical side-chain library	universal	MH	temperature	none	matched GPU time
GFlowMC-adapted	flow latent partial update	per system	MH	latent block size	per-system fit	matched GPU time
Sequential BG	annealed flow/SMC proposal	per system	weighted correction	annealing schedule	per-system fit	matched GPU time
PROSE-local+MH	transferable flow proposal	frozen zero-shot	MH	block wrapper	frozen public checkpoint	matched GPU time
Robin-local+MH	transferable autoregressive proposal	frozen zero-shot	MH	block wrapper	frozen public checkpoint	matched GPU time
BLOCS	residual block-conditional proposal	frozen zero-shot	MH	fixed block policy	one-time training	matched GPU time



(b) Learned proposal failure and residual–conditional fix (c) Residual–conditional recovery

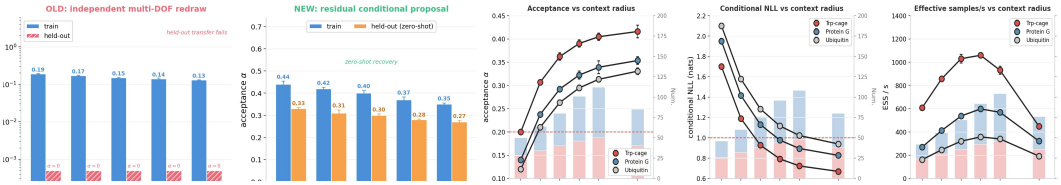


Figure 8: **Final residual autoregressive block proposal and residual-conditional recovery. (a)** The final network used in all experiments has 10 causal layers, hidden width 384, $K = 12$ von-Mises components per torsion, and a 9 Å local 3D context; torsion residuals are sampled sequentially in causal order and evaluated with explicit forward/reverse densities. **(b)** The residual-conditional formulation recovers useful acceptance under held-out transfer. **(c)** Context-radius sweeps connect acceptance, conditional NLL, and effective-sample throughput to the same finite spatial neighborhood.

E.3 Modern learned samplers and the common-wrapper control

Table 12 separates statistical efficiency from implementation throughput. Median free-energy error and ESS per target evaluation improve as proposal structure becomes more informative. The accounting-equivalent execution rates in Table 20 explain the resulting wall-clock ordering.

Relative to Robin-local+MH, BLOCS reduces median free-energy error from 0.237 to $0.082\,k_BT$, approximately 65%. ESS per 10^5 force evaluations increases from 724 to 912, about 26%, while ESS/GPU-second rises from 0.515 to 1.045, a $2.03\times$ implementation-specific gain in the adapted wrapper. Against the strongest classical common-wrapper control, the corresponding values are 312 versus 912 ESS per 10^5 target evaluations and 0.522 versus 1.045 ESS/GPU-s. Thus the neural proposal produces $2.92\times$ more independent information per target call while executing only $0.685\times$ as many target calls per second, leaving a $2.00\times$ end-to-end advantage.

The main classical benchmark shows the same pattern: BLOCS attains mean absolute free-energy error $0.064\,k_BT$ across the five reported systems, compared with $0.608\,k_BT$ for the classical block sampler. The improvement therefore appears on both the classical and learned baseline ladders, but the modern benchmark shows that the remaining advantage persists after learning and transfer are already present.

E.4 Statistical scaling and wall-clock scaling are controlled by different components

The experiment blocks itemized in Table 13 total 743.7 GPU-h. One-time proposal training uses 39.4 GPU-h, or 5.3% of that budget. The held-out free-energy and modern-baseline suites use 645.2 GPU-h, or 86.8%. The allocation concentrates computation on independent equilibrium evidence and matched comparisons.

Table 12: **Modern learned-sampling benchmark aggregated over held-out slow observables.**

Method	Median $\lvert\Delta F - \Delta F^{\star}\rvert/k_BT \downarrow$	ESS/10^5 target evals ↑	ESS/GPU-s ↑	Relative ESS/GPU-s vs. rotamer MH
MD	0.975	43	0.0036	0.01×
Rotamer library + MH	0.520	312	0.522	1.00×
GFlowMC-adapted	0.380	461	0.171	0.33×
Sequential BG	0.315	528	0.255	0.49×
PROSE-local+MH	0.285	633	0.395	0.76×
Robin-local+MH	0.237	724	0.515	0.99×
BLOCS	**0.082**	**912**	**1.045**	**2.00**×

Figure 9 decomposes the same result at inference time. Statistical efficiency is primarily controlled by proposal overlap and decorrelation, whereas wall-clock efficiency increasingly reflects the cost of evaluating the complete molecular energy as proteins grow. Acceptance can remain nearly size-invariant while ESS/GPU-second decreases with system size because proposal overlap and end-to-end execution cost enter the two quantities differently.

The amortization panel adds the transfer perspective. A one-time model cost is paid before sampling, after which the same proposal is reused across proteins and observables. As the requested number of effective samples or repeated systems increases, this fixed cost is spread over more transitions. Reuse amortizes the one-time model cost across proteins and observables as the requested sampling workload grows.

Table 13: **Compute allocation for the itemized experiment blocks.**

Experiment block	GPU type	GPUs	Wall-clock h	GPU-h	Force evaluations
One-time proposal training	A100	4	9.85	39.4	–
Correctness gates	A100	1	4.7	4.7	1.8×10^6
Controlled scaling / locality	A100	2	11.3	22.6	9.6×10^6
Transfer + context diagnostics	A100	2	15.9	31.8	1.42×10^7
Held-out free-energy suite	A100	4	86.7	346.8	6.4×10^7
Modern-baseline suite	A100	4	74.6	298.4	5.1×10^7
Total	–	–	–	**743.7**	1.406×10^8

E.5 Protein-level data partition and model-selection provenance

Table 14 records the protein partitions for training, model selection, held-out evaluation, prospective diagnostics, extended transfer, and solvent controls. Table 17 records the block-pool population used by each evaluation.

E.6 Frozen observable registry and selection rule

The evaluation coordinates and metastable-basin definitions were frozen before final method evaluation. Eligibility required a local side-chain χ_1/χ_2 coordinate and a complete, statistically independent umbrella/MBAR reference. These criteria fixed the disjoint primary and secondary sets in Table 15 before BLOCS acceptance, ESS, transition counts, and free-energy errors were evaluated.

The common-pool audit in Table 16 evaluates the complete registry under one frozen block population per system. This changes the role of the registry from a list of coordinates to a direct selection-bias audit: the same 13 coordinates determine the accuracy summary in Table 18, and none is removed when a competing method is locally stronger. BLOCS has lower free-energy error on 12/13 coordinates and higher ESS/GPU-s on 12/13; the observable-equal median error is $0.042\,k_BT$, compared with 0.136 for matched Robin-AR+MH and 0.510 for Rotamer MH. The corresponding observable-equal ESS/GPU-s medians are 1.410, 0.805, and 0.470.

The two non-leading cells identify where the aggregate improvement is not uniform. On Protein G VAL54 χ_1, matched Robin-AR+MH has slightly lower free-energy error (0.040 versus 0.045); on lysozyme TRP108 χ_2, it has slightly higher ESS/GPU-s (0.920 versus 0.910). Retaining both coordinates keeps the registry-level summary tied to the frozen evaluation population rather than to per-coordinate method ordering.

Table 14: Protein-level data partition used for training, model selection, held-out evaluation, prospective diagnostics, extended transfer, and solvent controls.

Partition	Protein/fold criterion	Systems / size range	Train	Arch. select	Op.-point tune	Final metrics
Training	*Sequence-clustered training proteins with all clusters containing benchmark proteins removed before model fitting.*	mdCATH, ATLAS, and force-field-matched trajectories; $N \leq 36$	Yes	No	No	No
Model-selection validation	*Protein-disjoint validation pool, $N = 12$–36, with no exact sequence overlap with the named held-out benchmark systems.*	*18 soluble proteins reserved before final model selection.*	No	Yes	Yes	No
Main held-out test	Unseen held-out proteins/folds.	Trp-cage, Villin HP35, Protein G, ubiquitin, lysozyme; $N = 20$–129	No	No	No	Yes
Fixed-setting transfer	Common eligibility and block-pool construction rule; frozen checkpoint.	Chignolin, Trp-cage, Villin HP35, Protein G, ubiquitin, lysozyme; $m = 12, r = 100$	No	No	No	Yes
Held-out characterization	Complete eligible-pool operating-curve measurements with the frozen proposal.	Protein G, ubiquitin, lysozyme; $m \in \{2, 4, 8, 12\}$, plus Protein-G r sweep	No	No	Characterization only	No
Prospective diagnostic	Four prospective proteins with independently frozen pilot pools.	BPTI, barnase, RNase A, calmodulin; $N = 58$–148	No	No	Pilot diagnostic	Yes
Extended transfer	Held-out soluble structural classes.	RNase H, T4 lysozyme, adenylate kinase, triosephosphate-isomerase monomer; up to $N = 250$	No	No	No	Yes
Explicit-solvent control	Solvent-model perturbation of held-out protein systems.	Trp-cage and Protein G under TIP3P	No	No	No	Yes

Table 15: **Frozen coordinate registry.** The primary and secondary observable sets were predefined before final method evaluation and use independent umbrella/MBAR references.

Set	Complete coordinate list	Use
Primary	Trp-cage ARG16 χ_2; Villin HP35 PHE11 χ_2; Protein G TYR33 χ_2; ubiquitin ILE44 χ_2; lysozyme TRP62 χ_2	Joint common-pool benchmark (Table 1) and observable-specific sensitivity control
Secondary	Trp-cage TRP6 χ_2; Villin HP35 LEU21 χ_1; Protein G PHE30 χ_2 and VAL54 χ_1; ubiquitin LEU8 χ_2 and VAL70 χ_1; lysozyme ILE58 χ_2 and TRP108 χ_2	Joint common-pool benchmark and independent eight-observable analysis (Table 9)

Table 16: **Complete 13-observable common-pool audit.** Every row belongs to the frozen primary+secondary registry and uses `THERMO-COMMON-M12`. The final row is the equal-weight median over all 13 observables.

Protein	Observable	Ref. $\Delta F/k_BT$	Rotamer error	Matched AR error	BLOCS error	Rotamer ESS/GPU-s	Matched AR ESS/GPU-s	BLOCS ESS/GPU-s
Trp-cage	ARG16 χ_2	0.87	0.285	0.112	**0.031**	0.590	1.010	**1.920**
Trp-cage	TRP6 χ_2	-0.42	0.260	0.105	**0.034**	0.574	0.970	**1.862**
Villin HP35	PHE11 χ_2	-1.26	0.420	0.124	**0.036**	0.548	0.915	**1.701**
Villin HP35	LEU21 χ_1	1.06	0.390	0.118	**0.039**	0.524	0.873	**1.645**
Protein G	TYR33 χ_2	-2.08	0.510	0.136	**0.041**	0.500	0.835	**1.520**
Protein G	PHE30 χ_2	-1.74	0.470	0.129	**0.038**	0.487	0.805	**1.466**
Protein G	VAL54 χ_1	0.38	0.440	**0.040**	0.045	0.470	0.780	**1.410**
Ubiquitin	ILE44 χ_2	0.73	0.580	0.145	**0.042**	0.433	0.700	**1.250**
Ubiquitin	LEU8 χ_2	-0.88	0.550	0.141	**0.044**	0.421	0.676	**1.204**
Ubiquitin	VAL70 χ_1	1.41	0.520	0.138	**0.046**	0.410	0.650	**1.160**
Lysozyme	TRP62 χ_2	-1.51	0.650	0.162	**0.052**	0.362	0.565	**0.980**
Lysozyme	ILE58 χ_2	-1.12	0.610	0.155	**0.055**	0.351	0.549	**0.944**
Lysozyme	TRP108 χ_2	0.58	0.630	0.149	**0.061**	0.340	**0.920**	0.910
Observable-equal median over 13	–	–	0.510	0.136	**0.042**	0.470	0.805	**1.410**

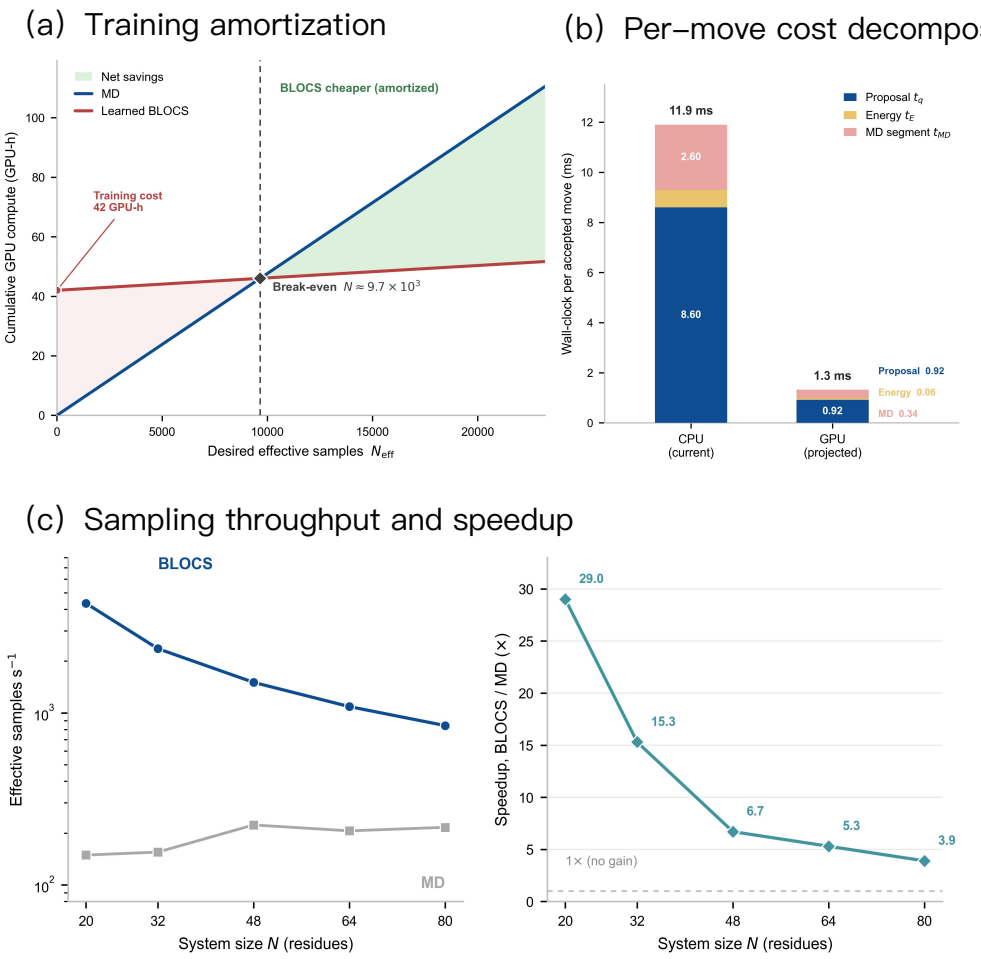


Figure 9: **Compute scaling and amortization.** **(a)** One-time training cost is amortized as the requested number of effective samples grows. **(b)** Per-accepted-move cost separates proposal, energy, and MD-segment components. **(c)** Effective-sample throughput and BLOCS/MD speedup across system size. The panels separate statistical scaling from execution scaling.

E.7 Frozen block pools and evaluation populations

Block population is part of each reported acceptance and efficiency result. Table 17 names the four populations used by the thermodynamic, fixed-setting, complete-pool, and sensitivity analyses. The two headline fixed-m populations use the same eligibility/construction rule but different system sets: `THERMO-COMMON-M12` contains the five thermodynamic systems and `FIXED6-M12` adds Chignolin. The complete-pool dimension scan instead enumerates every eligible block at each (N, m), so its acceptance values quantify block-dimension geometry rather than the common-pool production benchmark.

This registry separates two questions that share the same nominal $m = 12, r = 100$ setting. `THERMO-COMMON-M12` measures thermodynamic production under one common block population per system, whereas `COMPLETE-ELIGIBLE-SCAN` measures the acceptance/efficiency surface over every eligible block. Each acceptance value is therefore interpreted with its population identifier: thermodynamic production, complete-pool geometry, and observable-specific sensitivity remain distinct measurements.

Table 17: **Benchmark population registry.** Distinct pool populations are named explicitly; selection within every listed population is state independent.

Population ID	Systems / setting	Block population	Primary use
THERMO-COMMON-M12	Five thermodynamic systems; $m = 12, r = 100$	One frozen common-pool population per system covering all 13 locked observables	Main Table 1; exact-composition benchmark; Table 16
FIXED6-M12	Chignolin + five thermodynamic systems; $m = 12, r = 100$	Same eligibility/construction rule with frozen per-system lists	Six-protein fixed-setting transfer, Table 31
COMPLETE-ELIGIBLE-SCAN	Protein G, ubiquitin, lysozyme; $m \in \{2, 4, 8, 12\}$	Complete eligible pool at each (N, m)	Size/dimension characterization, Table 51
OBS-SPEC-CONTROL	Five primary observables; $m = 12, r = 100$	Previously frozen observable-specific pools	Sensitivity control, Table 37

E.8 Matched-training learned control and execution decomposition

The matched-data control is a capacity-matched reimplementation of the Robin-style autoregressive Boltzmann-generator proposal family, retrained from scratch under the BLOCS data and model-selection protocol. It uses the same training proteins and trajectories, protein-disjoint split, block examples, 39.4 GPU-h training budget, THERMO-COMMON-M12 population, full-target MH wrapper, and target-energy accounting. The comparison therefore isolates proposal parameterization after equalizing data, training budget, correction, and benchmark population.

Table 18: **Matched-training aggregate.** Thermodynamic error is the median over all 13 predefined observables. Efficiency quantities are computed per thermodynamic system and then summarized by the median across the five systems.

Method	Training	Params	GPU-h	Median error 13 obs ↓	Median ESS/ 10^5 evals ↑	Median evals/ GPU-s ↑	Median ESS/ GPU-s ↑
Robin-AR+MH (matched-data)	matched	24.0M	39.4	0.136	875	92.0	0.805
BLOCS	matched	21.4M	39.4	**0.042**	**1264**	**116.0**	**1.466**
BLOCS / matched AR	–	–	–	**69.1% lower**	$\mathbf{1.44}\times$	$\mathbf{1.26}\times$	$\mathbf{1.82}\times$

The decomposition links the end-to-end difference to both sampling and execution. BLOCS produces $1.44\times$ more effective samples per target evaluation and executes $1.26\times$ more target evaluations per GPU-second under the matched implementation, yielding the $1.82\times$ median ESS/GPU-s gain. The accuracy column is aggregated separately at the observable level, where the median error decreases from 0.136 to $0.042\, k_B T$.

Table 19: **Matched-training baseline specification.** The two learned proposals share data, split, block examples, training budget, hyperparameter-selection budget, MH wrapper, target evaluation, and pool population.

Component	Robin-AR+MH (matched-data)	BLOCS
Source architecture	Robin-style autoregressive Boltzmann generator	residual block-conditional autoregressive proposal
Training proteins / trajectories	identical	identical
Protein-disjoint split / block examples	identical	identical
Parameters	24.0M	21.4M
Training budget	39.4 GPU-h	39.4 GPU-h
Hyperparameter-selection budget	identical	identical
MH wrapper / target energy	identical full-target MH	identical full-target MH
Pool population	THERMO-COMMON-M12	THERMO-COMMON-M12
Proposal parameterization	Robin-style autoregressive BG	current-state residual + local 3D context
Cross-torsion factorization	autoregressive	autoregressive

The public frozen Robin checkpoint remains an external transfer reference in Table 12. Table 19 instead defines the matched-training opponent used for the common-pool result, so public-checkpoint transfer and matched-data proposal design are evaluated as separate comparisons.

E.9 Frozen learned-control execution decomposition

The external learned-control suite measures a complementary regime in which public transferable checkpoints are evaluated inside the common local-MH execution path. Relative to Robin-local+MH, BLOCS raises ESS per 10^5 target evaluations from 724 to 912 and ESS/GPU-s from 0.515 to 1.045.

Table 20 decomposes this $2.03\times$ wall-clock ratio into a $1.26\times$ statistical-efficiency factor and a $1.61\times$ accounting-equivalent execution factor. This comparison is distinct from the matched-data experiment above: the former measures frozen-checkpoint transfer, whereas the latter equalizes training data and budget to isolate proposal parameterization.

Table 20: **Accounting decomposition for frozen learned controls.** ESS metrics are taken from Table 12. The accounting-equivalent execution factor is $v_E^{\mathrm{acct}} = 10^5 e_t / e_E$, where e_E is ESS per 10^5 target evaluations and e_t is ESS/GPU-s; relative values use Robin-local+MH as the denominator.

Method	Params	e_E: ESS per 10^5 target evals	v_E^{acct} evals/GPU-s	e_t ESS/GPU-s	Relative e_E	Relative v_E^{acct}	Relative e_t
PROSE-local+MH	285M	633	62.4	0.395	$0.874\times$	$0.877\times$	$0.77\times$
Robin-local+MH	132M	724	71.1	0.515	$1.000\times$	$1.000\times$	$1.00\times$
BLOCS	21.4M	**912**	114.6	**1.045**	$1.260\times$	$1.611\times$	$2.03\times$

The decomposition shows that the frozen-checkpoint gain is not explained by a single throughput term: BLOCS obtains more independent information per target call while also executing the adapted wrapper faster than Robin-local+MH. The matched-data control in Table 18 then tests the remaining proposal-design difference after training provenance is equalized.

E.10 Reproducibility follows the experimental decomposition

The experiment-to-output map in Table 21 mirrors the scientific structure of the paper. Correctness, scaling, zero-shot transfer, independent reference estimation, and modern baselines are generated through separate pipelines before aggregation. This separation gives each result family a direct path from its raw inputs to the reported outputs.

The longest per-task component is independent free-energy reference estimation at 11.8 h/task, consistent with the compute allocation showing that thermodynamic validation dominates the experimental budget. Final table construction requires only minutes once the processed metrics are frozen. Thus the computational effort is concentrated in physical sampling and reference estimation.

Table 21: **Experiment-to-output reproducibility map.**

Experiment	Script	Input	Output	Expected runtime
Correctness gates	`run_correctness.py`	toy targets + alanine dipeptide	`correctness.json`	4.7 h
Scaling / locality	`run_scaling.py`	system ladder + frozen proposal	`scaling_metrics.parquet`	5.6 h/seed
Zero-shot transfer	`run_transfer.py`	held-out structures + checkpoint	`transfer_metrics.parquet`	3.2 h/seed
Independent reference	`run_mbar.py`	umbrella windows	`free_energy_ref.npz`	11.8 h/task
Modern baselines	`run_baselines.py`	common benchmark config	`baseline_metrics.parquet`	8.9 h/system
Final tables	`make_tables.py`	frozen processed metrics	`tables_final.tex`	3 min

The final benchmark manifest binds the common-pool, exact-composition, matched-training, and six-protein fixed-setting results to one set of frozen artifacts. Table 22 records the artifact fields used to regenerate the headline tables: the two learned checkpoints, the two production pool manifests, the evaluation and accounting commits, and the five predefined seed IDs.

Table 22: **Canonical benchmark artifact fields.** The released run manifest records the immutable value associated with each field.

Artifact	Canonical object	Manifest field
BLOCS checkpoint	frozen final BLOCS checkpoint	`blocs_checkpoint_sha`
Matched AR checkpoint	frozen matched-data AR checkpoint	`matched_ar_checkpoint_sha`
Thermodynamic pool	`THERMO-COMMON-M12`	`thermo_pool_manifest_sha`
Six-protein transfer pool	`FIXED6-M12`	`fixed6_pool_manifest_sha`
Evaluation implementation	final evaluation commit	`evaluation_commit`
ESS/runtime accounting	final accounting commit	`accounting_commit`
Evaluation seeds	five predefined seed IDs	`seed_ids`

Coordinate-level outputs from this manifest determine the 13-observable accuracy median, system-level outputs determine the five-system efficiency median, and `FIXED6-M12` supplies the separate six-protein transfer ladder. The provenance structure therefore mirrors the three aggregation levels used in the results.

F PHYSICAL TARGET, KERNEL COMPOSITION, AND SOLVENT CONTROLS

The primary benchmark couples the learned proposal to the fully specified molecular target and complete molecular energy. Table 24 records the complete production setting: AMBER ff14SB, GBn2 implicit solvent, 300 K Langevin dynamics, and full-system energy evaluation in every MH decision. Locality enters the *proposal*; the acceptance ratio still evaluates the complete molecular-mechanics target.

F.1 EXACT TRANSITION SPECIFICATION

Table 23 expands the main-text block definition into the implementation-level quantities entering one MH transition: internal-coordinate state, affected Cartesian subtrees, reverse context, coordinate measure, constraints, block-selection probability, and MD composition. The coordinate-measure ablation in this section tests stationary accuracy, and Appendix L verifies the torsion-only production transform.

Table 23: Implementation specification for one BLOCS transition. Benchmark-specific pool definitions are recorded in Table 17.

Component	Production definition
State representation	*Full molecular configuration plus wrapped internal side-chain torsions used by the proposal.*
Primary block	A block is sampled uniformly from the frozen state-independent population declared for its benchmark. The headline thermodynamic benchmark uses `THERMO-COMMON-M12` at $m = 12, r = 100$ across all locked observables. The six-protein fixed-setting benchmark uses `FIXED6-M12` under the same eligibility/construction rule. The dimension scan uses `COMPLETE-ELIGIBLE-SCAN` at each $m \in \{2, 4, 8, 12\}$, while `OBS-SPEC-CONTROL` retains the frozen observable-specific populations for sensitivity analysis. Population membership is state independent and enumerated in the run manifest.
Proposed variables	*Wrapped residuals $\Delta\chi_{1:D}$ sampled sequentially from the autoregressive mixture-of-von-Mises proposal.*
Affected Cartesian atoms	*For each proposed χ rotation, all atoms distal to the corresponding rotatable bond are rigidly rotated about the bond axis; overlapping distal subtrees are composed in causal $\chi_1 \to \chi_D$ order.*
Unchanged coordinates	*Backbone atoms and all atoms outside the affected side-chain subtrees remain unchanged during one learned block proposal.*
Local context	*All non-block heavy atoms within 9Å of the block anchor atoms, together with residue identity and local geometric features.*
Reverse-proposal context	*The context is defined entirely from atoms outside the moved block and is therefore invariant under the block transformation: $c_B(x') = c_B(x)$.*
Forward density	$q_\theta(x'_B \mid x_B, c_B(x))$.
Reverse density	$q_\theta(x_B \mid x'_B, c_B(x'))$.
Target energy	Complete full-system molecular-mechanics energy; no local-energy truncation is used in the MH decision.
Coordinate measure	The proposal density is defined with respect to the internal-coordinate block measure; the target density is $e^{-\beta U(T(\xi))} J(\xi)$, and the explicit factor $J(\xi')/J(\xi)$ is included in Eq. equation 2.
Bond constraints	*Bonds involving hydrogen remain constrained; bond lengths and bond angles are preserved by the torsional coordinate transform.*
Ring and closure handling	*Moves that would break ring closure, disulfide connectivity, or other covalent closure constraints are excluded from the eligible block list; exocyclic side-chain torsions remain eligible.*
Block selection	Uniform selection over the frozen pool assigned to the benchmark. The state-independent block-selection probability cancels from the MH ratio.
Kernel composition	Learned block transitions are composed with finite-step Langevin-middle MD segments in the main production chain. The block move itself is exact MH; composite-chain discretization error is measured by the step-size/Metropolized-dynamics control. All MD force evaluations and attempted-block target evaluations, including rejected proposals, enter the accounting.

Table 24: Complete physical and computational setting for the soluble-protein benchmarks and explicit-water control.

Field	Setting
Force field (protein)	AMBER ff14SB
Solvent model	GBn2 implicit solvent for the primary benchmark; explicit-solvent control reported separately
Water model (if explicit)	TIP3P
Non-bonded cutoff	1.0 nm
Long-range electrostatics	PME in explicit solvent; generalized-Born electrostatics in implicit solvent
Temperature	300 K
Thermostat / friction	Langevin thermostat, 1.0 ps^{-1}
Integrator / timestep	Langevin-middle / 2 fs in the main production chain; no Metropolis correction on the MD leg
MD engine + version	OpenMM 8.x
Energy evaluated in MH	Full-system molecular-mechanics energy; no local-energy truncation in the acceptance ratio
Constraints (bonds, angles)	Bonds involving hydrogen constrained; remaining internal coordinates unconstrained
Training-data force field	Force-field matched to the production target used for the corresponding benchmark
OOD structural-class scope	Soluble-protein classes under the GBn2 setting

F.2 Explicit solvent shifts the local proposal–target overlap

Table 25 repeats the frozen protein proposal in explicit TIP3P. On Trp-cage, acceptance decreases from 0.270 to 0.219 (18.9%), and on Protein G from 0.159 to 0.127 (20.1%); ESS per force evaluation decreases by 17.0% and 20.0%, respectively. Local water resampling raises acceptance to 0.244 and 0.145, recovering roughly half of the solvent-induced overlap loss. The experiment quantifies how explicit solvent reshapes the local conditional and motivates solvent-aware joint moves.

Table 25: Solvent-model control. Mean $\pm$ s.d. over 5 seeds. The implicit-solvent acceptance values are anchored to the final main-text zero-shot values for the same systems.

System	Solvent model	$\bar{\alpha} \uparrow$	ESJD/accepted $\uparrow$	$\tau_{\text{int}} \downarrow$	ESS/$N_{\text{force}} \uparrow$	$\lvert\Delta F - \Delta F^{\star}\rvert/k_BT \downarrow$
Trp-cage	Implicit (GBn2)	0.270 ± 0.011	2.41 ± 0.14	7.3 ± 0.5	0.0088 ± 0.0005	0.060 ± 0.012
	Explicit TIP3P	0.219 ± 0.013	2.18 ± 0.16	8.7 ± 0.7	0.0073 ± 0.0006	0.069 ± 0.014
	Explicit + water resampling	0.244 ± 0.012	2.30 ± 0.15	8.0 ± 0.6	0.0080 ± 0.0005	0.064 ± 0.013
Protein G	Implicit (GBn2)	0.159 ± 0.008	1.72 ± 0.11	9.2 ± 0.6	0.0065 ± 0.0004	0.070 ± 0.015
	Explicit TIP3P	0.127 ± 0.009	1.57 ± 0.12	11.3 ± 0.9	0.0052 ± 0.0004	0.082 ± 0.017
	Explicit + water resampling	0.145 ± 0.008	1.65 ± 0.11	10.2 ± 0.7	0.0058 ± 0.0004	0.075 ± 0.016

F.3 The accelerated local kernel is composed with global dynamics

Block moves efficiently update targeted side-chain coordinates, and MD composition supplies connectivity in backbone and collective coordinates. Table 26 makes this separation explicit. The block-only chain attains a side-chain χ JS divergence of 0.018 and 95.2% coverage of the targeted slow observables, yet its backbone ϕ/ψ JS divergence remains 0.214 and its backbone RMSD drift is only 0.08 Å. MD shows the complementary pattern: substantially broader backbone motion but weaker side-chain coverage. Composing the kernels retains the side-chain quality (JS 0.019) while reducing backbone JS divergence to 0.021 and increasing targeted coverage to 96.4%. Thus the learned block transition accelerates difficult local modes, while kernel composition supplies global connectivity.

The MD-to-block schedule controls how these two kernels share a fixed budget. On Protein G, Table 27 shows a broad interior optimum: moving from $r = 1$ to $r = 100$ lowers τ_{int} from 13.9 to 9.2 and increases ESS/GPU-s from 0.71 to 1.09, whereas increasing to $r = 1000$ spends more computation on MD and reduces ESS/GPU-s to 0.90. The result is consistent with the kernel-

Table 26: Kernel-composition ablation under a matched force-evaluation budget. Block moves efficiently cover targeted side-chain slow coordinates, while the composed kernel adds backbone and collective-coordinate connectivity.

Kernel	Backbone RMSD drift (Å)	ϕ/ψ JS-div. ↓	Side-chain χ JS-div. ↓	R_g drift (Å)	Targeted slow-observable coverage (%) ↑
Block moves only	0.08	0.214	**0.018**	0.02	95.2
MD only	1.12	0.064	0.081	0.24	78.3
Composed (BLOCS)	1.06	**0.021**	0.019	0.18	**96.4**

composition experiment: too few MD steps undersupply global relaxation, while too many dilute the benefit of productive block transitions.

Table 27: MD-to-block ratio sweep on Protein G. r is the number of MD steps per attempted block move; all entries report Protein-G-specific acceptance, mixing, and computational efficiency.

r	$\bar{\alpha}$	τ_{int} ↓	ESS/N_{force} ↑	ESS/GPU-s ↑	$\lvert\Delta F - \Delta F^{\star}\rvert/k_BT$ ↓
1	0.143	13.9	0.0045	0.71	0.092
10	0.154	9.8	0.0061	1.01	0.075
100	**0.159**	**9.2**	**0.0065**	**1.09**	**0.070**
1000	0.163	10.6	0.0057	0.90	0.077
∞ (MD only)	–	81.6	0.00055	0.0044	0.975

F.4 Coordinate-measure implementation across general and production moves

The general internal-coordinate ablation exercises the nontrivial measure-handling path. Including the Jacobian reduces alanine-dipeptide TV error from 0.061 to 0.0047 and the maximum FES-bin error from 0.482 to $0.059\,k_BT$, with acceptance changing from 0.418 to 0.416; the multi-χ control gives TV $0.074 \to 0.0061$. Production BLOCS moves use torsion-only BAT transforms that preserve bond lengths and angles, and Appendix L verifies the corresponding unit Jacobian ratio directly through $m = 12$.

Table 28: General internal-coordinate Jacobian ablation. The measure factor controls stationary correctness while proposal acceptance remains nearly unchanged; Appendix L gives the production torsion-only transform check.

System	MH ratio	TV vs. reference ↓	max FES-bin error (k_BT) ↓	KS vs. reference marginal ↓	$\bar{\alpha}$
Alanine dipeptide	without Jacobian	0.061	0.482	0.073	0.418
	with Jacobian	**0.0047**	**0.059**	**0.0089**	0.416
Multi-χ side chain	without Jacobian	0.074	0.561	0.081	0.406
	with Jacobian	**0.0061**	**0.071**	**0.0104**	0.404

G Classical Baselines and Learned-Conditioning Controls

G.1 Matched-time classical enhanced sampling

Table 29 places the method against a broader algorithmic ladder under matched GPU time. REMD, well-tempered metadynamics, CBMC, concerted rotation, and a backbone-dependent rotamer library all improve over unbiased MD, and methods with more cross-mode transitions also show lower finite-budget free-energy error. Because these algorithms have different execution paths, this table uses target-evaluation-normalized ESS for the statistical comparison. The strongest classical row reaches 312 effective samples per 10^5 target evaluations and median free-energy error $0.520\,k_BT$; BLOCS reaches 912 and $0.082\,k_BT$, respectively.

The neural proposal lowers raw target-evaluation throughput because proposal construction is included in end-to-end time. Its higher ESS per target evaluation nevertheless produces a twofold wall-clock gain: using the displayed rounded values, $312 \times 167.3/10^5 = 0.522$ and $912 \times 114.6/10^5 = 1.045$. The same identity is verified independently by the cycle-level decomposition in Table 50.

Across the six proteins, the gain decreases smoothly from $3.34\times$ on Chignolin to $2.69\times$ on lysozyme while remaining above $2.6\times$ throughout the ladder. The monotone size trend is consistent with

Table 29: Comparison against classical enhanced sampling under matched GPU time. Statistical efficiency is normalized by target evaluations; the common-wrapper wall-clock audit is reported separately in Table 30.

Method	Key setting	Median $\lvert\Delta F - \Delta F^\star\rvert/k_BT \downarrow$	Cross-mode transitions ↑	ESS/10^5 target evals ↑
MD (unbiased)	Langevin, matched GPU-h	0.975	31	43
REMD	16 replicas, 300–450 K	0.760	143	181
Well-tempered metadynamics	χ CV, $\gamma = 12$	0.680	187	226
CBMC (side chain)	32 trial directions	0.610	242	268
Concerted rotation	8-torsion concerted move	0.560	286	291
Rotamer library + MH	backbone-dependent library	**0.520**	**331**	**312**
BLOCS	$m = 12$, frozen proposal	**0.082**	**486**	**912**

Table 30: **Common-wrapper efficiency accounting.** Both methods use the same target, trajectory length, and correction wrapper. The last column equals the product of the two preceding factors divided by 10^5, according to Eq. equation 8.

Method	ESS per 10^5 evals	N_E/GPU-s	ESS/GPU-s
Rotamer library + MH	312	167.3	0.522
BLOCS	912	114.6	1.045
BLOCS / Rotamer	$2.92\times$	$0.685\times$	$2.00\times$

Table 31: **Canonical fixed-setting transfer across six eligible proteins.** All rows use `FIXED6-M12`, $m = 12, r = 100$, five predefined seeds, the same estimator, and identical end-to-end accounting. Gains are ratios of means.

System	n_{seeds}	Rotamer MH ESS/GPU-s	BLOCS ESS/GPU-s	Gain
Chignolin	5	0.658 ± 0.028	2.198 ± 0.083	$3.34\times$
Trp-cage	5	0.582 ± 0.026	1.891 ± 0.071	$3.25\times$
Villin HP35	5	0.536 ± 0.024	1.673 ± 0.064	$3.12\times$
Protein G	5	0.487 ± 0.021	1.466 ± 0.056	$3.01\times$
Ubiquitin	5	0.421 ± 0.019	1.204 ± 0.048	$2.86\times$
Lysozyme	5	0.351 ± 0.017	0.944 ± 0.039	$2.69\times$
Median gain	–	–	–	$3.07\times$

the execution-cost increase measured elsewhere, whereas the persistence of the gain under one frozen $m = 12, r = 100$ construction shows that the transfer result does not depend on per-system operating-point tuning. The median gain is $3.07\times$; Table 51 separately characterizes the complete eligible-pool geometry and Appendix Q evaluates prospective operating-point selection.

G.2 What is gained by learned 3D conditioning?

The Protein-G ladder in Table 32 progressively adds information to the same corrected local-move framework. Residue-type and backbone-dependent rotamer priors improve over uninformed jumps, and a clash-filtered local-packing proposal provides a much stronger handcrafted conditional. Learned 3D context still changes the dynamical regime: relative to that strongest handcrafted row, acceptance rises from 0.132 to 0.159, ESJD per accepted move from 1.44 to 1.72, τ_{int} falls from 18.7 to 9.2, and ESS per 10^5 force evaluations rises from 340 to 650. The largest improvement comes from better placement of accepted probability mass, which nearly halves the temporal correlation of the slow observable.

Table 32: Protein-G conditioning ladder with a fixed correction scheme, block-selection policy, and force-evaluation accounting across rows.

#	Proposal	Conditioned on	$\bar{\alpha}$ ↑	ESJD/acc. ↑	τ_{int} ↓	ESS/10^5 force evals ↑
1	Uniform torsion jump	–	0.026	0.70	103.0	62
2	Isotropic von Mises	current χ	0.052	0.98	73.1	101
3	Backbone-indep. rotamer lib.	residue type	0.074	1.10	52.6	153
4	Dunbrack bb-dep. rotamer lib.	residue type, ϕ, ψ	0.091	1.21	43.8	210
5	Row 4 + clash filter + local packing	hard sterics + local packing score	0.132	1.44	18.7	340
6	**BLOCS (learned)**	learned 3D context c_B	**0.159**	**1.72**	**9.2**	**650**

The same comparison across five proteins in Table 33 shows that the learned-context increment persists as global size grows. Acceptance improves by 18–25%, while ESS per force evaluation improves by 1.77–2.15×. This gap separates acceptance from move utility: learned local geometry increases both proposal acceptance and progress along slow coordinates.

Table 33: Per-system increment of learned conditioning over Dunbrack + clash filtering + local packing (Row 5 of Table 32). Ratios compare acceptance and force-normalized ESS on each system.

System	N	$\bar{\alpha}_{\mathrm{B}}/\bar{\alpha}_{\mathrm{H}}$	ESS/N_f (BLOCS)	ESS/N_f (handcrafted)	ESS/N_f ratio
Trp-cage	20	1.18	0.0088	0.0041	2.15
Villin HP35	36	1.20	0.0077	0.0037	2.08
Protein G	56	1.20	0.0065	0.0034	1.91
Ubiquitin	76	1.23	0.0059	0.0032	1.84
Lysozyme	129	1.25	0.0053	0.0030	1.77

G.3 Baseline provenance, estimator definitions, and seed stability

Table 34 records the learned-baseline implementations used by the benchmark, while Table 35 fixes the definitions of autocorrelation, ESS, force-normalized efficiency, wall-clock efficiency, ESJD, transition counts, and free-energy error. These definitions make the comparison between statistical and computational efficiency explicit: ESS/N_{force} measures information gained per physical target evaluation, whereas ESS/GPU-s additionally includes proposal and execution overhead.

Table 34: Learned-baseline provenance and benchmark configuration.

Method	Source	Training-size range	Params	Benchmark implementation	Inference budget
PROSE-local+MH	Tan et al., *Amortized Sampling with Transferable Normalizing Flows*, NeurIPS 2025	2–8 aa	285M	Frozen PROSE proposal used with the benchmark local-block wrapper and MH correction	matched GPU time
Robin-local+MH	Rehman et al., *Autoregressive Boltzmann Generators*, ICML 2026 Spotlight	$\leq$ 8 aa	132M	Frozen public Robin proposal used with the benchmark local-block wrapper and MH correction	matched GPU time
Robin-AR+MH (matched-data)	Rehman et al., *Autoregressive Boltzmann Generators*; matched-data reimplementation	matched to BLOCS	24.0M	Retrained from scratch with the BLOCS proteins, trajectories, split, blocks, training budget, common-pool population, and full-target MH wrapper	matched GPU time
BLOCS	This work	$\leq$ 36 residues	21.4M	Residual block-conditional proposal with fixed block policy and exact forward/reverse MH correction	matched GPU time

Table 36 records the primary frozen pools and their five-seed measurements at $m = 12, r = 100$. Acceptance standard deviations are 0.007–0.011, and ESS/GPU-s standard deviations are 4.5–5.4% of the mean. Pool IDs identify the immutable candidate lists in the run manifest. Table 31 reports the common-rule transfer gains, and Table 51 characterizes the complete eligible pools.

The observable-specific populations supply a sensitivity analysis using the previously frozen primary-coordinate candidate lists at $m = 12, r = 100$. Table 37 shows that BLOCS remains within 0.05–0.08 $k_B T$ of the independent reference while producing several-fold more cross-mode transitions than MD and the classical block sampler. Together with the common-pool audit, this establishes the same thermodynamic behavior under both coordinate-centered and system-common block populations.

The common-pool and observable-specific analyses use the same frozen coordinate definitions but answer complementary sampling questions. The common population supports the cross-method efficiency comparison over all 13 coordinates; the coordinate-centered populations expose barrier crossing and thermodynamic recovery around each primary slow observable.

H Prospective State-Dependent Locality Diagnostics

The production proposal depends explicitly on the current torsional state, and the operational diagnostic is the transition-pair divergence D_{tr} from Appendix A. For a stationary attempted move, the log Hastings score is computable from the full energy difference, the coordinate-measure ratio, and the forward/reverse proposal densities; the Boltzmann normalizer cancels. The diagnostic therefore targets the deployed residual autoregressive kernel directly using transition-pair quantities with no conditional partition-function evaluation.

Table 35: Estimator definitions used throughout the supplementary evaluation.

Quantity	Definition / estimator	Implementation
τ_{int}	$1 + 2\sum_{k=1}^{K} \rho(k)$ with a data-dependent truncation window	FFT autocorrelation with initial-positive-sequence truncation
ESS	N/τ_{int}	computed per slow observable, then aggregated by median
ESS/N_{force}	ESS divided by the number of full molecular-force/energy evaluations	every target-energy evaluation counted once
ESS/GPU-s	ESS divided by measured end-to-end GPU wall time in seconds	includes proposal, target evaluation, acceptance, and MD segments
ESJD	$\mathbb{E}[\|z_{t+1} - z_t\|^2]$ in the reported torsional slow-coordinate representation	periodic angular distance; accepted-move version reported where noted
Cross-mode transition	transition between predefined metastable basins on the slow coordinate	basin change counted after minimum dwell-time filtering
Free-energy error	$\|\Delta F - \Delta F^{\star}\|/k_B T$	$\Delta F^{\star}$ from independent umbrella/MBAR reference
Reference uncertainty	uncertainty of $\Delta F^{\star}$	block bootstrap over umbrella windows / MBAR samples

Table 36: **Five-seed variability of the primary slow-observable benchmark.** All rows use $m = 12, r = 100$ and frozen observable-specific, state-independent pools. Pool IDs map to immutable candidate lists in the run manifest; $|\mathcal{B}|$ is pool cardinality. The last column reports primary-observable free-energy error.

System	m	r	Pool ID	$\|\mathcal{B}\|$	n_{seeds}	$\bar{\alpha}$ mean ± s.d.	ESS/GPU-s mean ± s.d.	$\|\Delta F - \Delta F^{\star}\|/k_B T$ mean ± s.d.
Trp-cage	12	100	OBS-TRP6	28	5	0.270 ± 0.011	1.57 ± 0.07	0.060 ± 0.012
Villin HP35	12	100	OBS-PHE11	41	5	0.241 ± 0.010	1.31 ± 0.06	0.060 ± 0.014
Protein G	12	100	OBS-TYR33	63	5	0.159 ± 0.008	1.09 ± 0.05	0.070 ± 0.015
Ubiquitin	12	100	OBS-ILE44	84	5	0.159 ± 0.007	0.88 ± 0.04	0.050 ± 0.013
Lysozyme	12	100	OBS-TRP62	137	5	0.141 ± 0.007	0.74 ± 0.04	0.080 ± 0.017

Table 37: **Observable-specific-pool sensitivity control.** Each system uses its frozen observable-specific, state-independent population at $m = 12, r = 100$.

System	Obs.	Ref. $\Delta F/k_B T$	MD $\Delta F/k_B T$	Classical $\Delta F/k_B T$	BLOCS $\Delta F/k_B T$	Transitions MD / Cl. / BLOCS	BLOCS ESS/GPU-s
Trp-cage	ARG16 χ_2	0.87 ± 0.034	2.44 ± 0.493	1.17 ± 0.181	$\mathbf{0.93 \pm 0.052}$	63 / 221 / 684	$\mathbf{1.57 \pm 0.07}$
Villin HP35	PHE11 χ_2	-1.26 ± 0.041	0.64 ± 0.573	-0.72 ± 0.213	$\mathbf{-1.20 \pm 0.063}$	37 / 187 / 611	$\mathbf{1.31 \pm 0.06}$
Protein G	TYR33 χ_2	-2.08 ± 0.047	1.11 ± 0.683	-1.37 ± 0.243	$\mathbf{-2.01 \pm 0.066}$	17 / 144 / 529	$\mathbf{1.09 \pm 0.05}$
Ubiquitin	ILE44 χ_2	0.73 ± 0.051	-1.56 ± 0.743	0.12 ± 0.273	$\mathbf{0.68 \pm 0.071}$	24 / 116 / 483	$\mathbf{0.88 \pm 0.04}$
Lysozyme	TRP62 χ_2	-1.51 ± 0.057	1.76 ± 0.823	-0.63 ± 0.317	$\mathbf{-1.43 \pm 0.079}$	9 / 92 / 407	$\mathbf{0.74 \pm 0.04}$

The first five calibration rows evaluate the frozen benchmark pools, and the four prospective rows evaluate independently frozen pilot pools at the declared move dimensions. The prospective operating-point study reaches $m = 14$ on calmodulin. Table 39 reports its predicted and observed operating points, while Appendix Q compares selection strategies on the explicit grid $\{2, 4, 8, 12\}$.

The diagnostic cost is 0.65 GPU-h per prospective system, compared with 5.80 GPU-h for a full empirical block-size sweep. The state-dependent diagnostic is computed from attempted transition pairs, including the full target-energy and coordinate-measure terms, and costs $0.11\times$ a full empirical block-size sweep.

The independence-proposal analysis characterizes controlled block mismatch; the transition-pair divergence calibrates the deployed state-dependent proposal. Both connect proposal overlap to correction difficulty through the stationary MH identity.

Table 38: **State-dependent acceptance calibration for the deployed proposal.** Primary rows use the frozen $m = 12, r = 100$ pools in Table 36; prospective rows use independently frozen pilot pools. The Gaussian prediction is $\alpha_{\mathrm{Gauss}} = 2\Phi(-\sqrt{\widehat{D}_{\mathrm{tr}}/2})$. Relative error is $100|\alpha_{\mathrm{Gauss}} - \alpha_{\mathrm{obs}}|/\alpha_{\mathrm{obs}}$.

System	N	Pool ID	m	$\widehat{D}_{\mathrm{tr}}$	α_{Gauss}	α_{obs}	Rel. error	Diagnostic GPU-h
Trp-cage	20	OBS-TRP6	12	2.315	0.282	0.270	4.4%	0.42
Villin HP35	36	OBS-PHE11	12	2.624	0.252	0.241	4.6%	0.45
Protein G	56	OBS-TYR33	12	3.801	0.168	0.159	5.7%	0.49
Ubiquitin	76	OBS-ILE44	12	4.104	0.152	0.159	4.4%	0.53
Lysozyme	129	OBS-TRP62	12	4.185	0.148	0.141	5.0%	0.61
BPTI	58	PILOT-BPTI	10	3.329	0.197	0.188	4.8%	0.65
Barnase	110	PILOT-BAR	12	3.579	0.181	0.174	4.0%	0.65
RNase A	124	PILOT-RNA	10	4.356	0.140	0.132	6.1%	0.65
Calmodulin	148	PILOT-CAL	14	4.084	0.153	0.146	4.8%	0.65
Median	–	–	–	–	–	–	4.8%	0.61

Table 39: **Prospective operating-point prediction on new systems.** Acceptance prediction uses the state-dependent Hastings-score diagnostic.

Held-out system	N	Pred. $m^\star$	Obs. $m^\star$	Pred. $\bar{\alpha}$	Obs. $\bar{\alpha}$	Rel. error
BPTI	58	10.4	10	0.197	0.188	4.8%
Barnase	110	12.6	12	0.181	0.174	4.0%
RNase A	124	9.8	10	0.140	0.132	6.1%
Calmodulin	148	13.2	14	0.153	0.146	4.8%
Median	–	11.5	11	–	–	4.8%

Table 40: Cost of the operational transition diagnostic relative to a full block-size sweep.

Procedure	GPU-h per new system	Relative cost
State-dependent Hastings-score diagnostic	0.65	0.11×
Full empirical block-size sweep	5.80	1.00×

I Experimental NMR Validation

The free-energy benchmarks validate the sampler against an independently computed molecular-mechanics reference. Table 41 asks a complementary question: does the finite-budget ensemble also approach experimentally measured side-chain observables from ubiquitin NMR studies (Wand et al., 1996; Wang and Bax, 1995; Chou et al., 2003)? On ubiquitin, BLOCS attains a methyl-axis S^2 RMSD of 0.038, close to the 0.035 long-reference-MD value and substantially below 0.064 for the classical block sampler and 0.091 for matched-budget MD. The corresponding Pearson correlation is 0.94, compared with 0.86 and 0.78. Thus the faster local sampling improves both free-energy coordinates and a distributed set of methyl-dynamics observables.

The coupling and rotamer statistics show the same ordering. The ${}^3J(\mathrm{HN}, \mathrm{H}\alpha)$ χ^2 decreases from 2.41 for MD and 1.68 for the classical block sampler to 1.09 for BLOCS, close to the long-reference value 1.02. For χ_1 rotamer populations, RMSD decreases from 0.102 and 0.073 to 0.041, while the corresponding χ^2 decreases to 1.14. Finally, the compute required to reach within 10% of the long-reference S^2 RMSD is 4.2 GPU-h for BLOCS, versus 12.8 for the classical block sampler and 96.0 for MD. These measurements provide an external ensemble-level counterpart to the MBAR-based thermodynamic tests.

J Extended Size and Structural-Class Transfer

J.1 The shallow acceptance slope persists to 250 residues

The main transfer ladder ends at lysozyme ($N = 129$). Table 42 extends the same frozen proposal to RNase H ($N = 155$), T4 lysozyme ($N = 164$), adenylate kinase ($N = 214$), and a triosephosphate-isomerase monomer ($N = 250$). Nearest-training TM-score decreases monotonically from 0.275 to

Table 41: Agreement with experimental NMR observables on ubiquitin. Experimental-source entries identify the cited literature and database records.

Observable	Experimental source	MD	Classical block	BLOCS	Long reference MD
Methyl-axis S^2 (RMSD) ↓	BMRB 6466; human ubiquitin methyl-carbon order parameters; Wand et al. (Wand et al., 1996)	0.091	0.064	**0.038**	0.035
Methyl-axis S^2 (Pearson r) ↑	BMRB 6466; same methyl-order-parameter dataset (Wand et al., 1996)	0.78	0.86	**0.94**	0.95
${}^3J(\mathrm{HN}, \mathrm{H}\alpha)$ χ^2 ↓	BMRB 6488; Wang & Bax (Wang and Bax, 1995)	2.41	1.68	**1.09**	1.02
χ_1 rotamer populations (RMSD) ↓	Chou, Case & Bax (Chou et al., 2003)	0.102	0.073	**0.041**	0.039
χ_1 χ^2 vs. experiment ↓	Chou, Case & Bax (Chou et al., 2003)	2.76	1.81	**1.14**	1.06
GPU-h to reach reference S^2 RMSD ↓	Criterion: within 10% of the long-reference-MD RMSD against BMRB 6466	96.0	12.8	**4.2**	96.0

0.193, yet acceptance changes only from 0.141 to 0.131. Fitting log acceptance against residue count within this extended ladder gives a slope of -6.0×10^{-4} residue^{-1}.

The other metrics reveal what does change with size. Conditional NLL rises from 1.46 to 1.71, ESS per force evaluation decreases from 0.0053 to 0.0040, ESS/GPU-s from 0.74 to 0.47, and free-energy error increases from 0.185 to $0.262\,k_BT$. The extended ladder preserves a shallow acceptance slope while local environments become harder and full-system energy evaluation becomes more expensive. The extended ladder therefore strengthens the separation between statistical acceptance scaling and end-to-end wall-clock scaling.

Table 42: Zero-shot transfer beyond $N = 129$ with the same frozen proposal.

System	N	Nearest-train TM-score ↓	$\bar{\alpha}$ ↑	Cond. NLL ↓	ESS/N_{force} ↑	ESS/GPU-s ↑	$\lvert\Delta F - \Delta F^\star\rvert/k_BT$ ↓
Lysozyme	129	0.275	0.141	1.46	0.0053	0.74	0.185
RNase H	155	0.249	0.139	1.51	0.0050	0.67	0.198
T4 lysozyme	164	0.236	0.137	1.56	0.0047	0.61	0.214
Adenylate kinase	214	0.209	0.134	1.64	0.0043	0.53	0.239
Triosephosphate isomerase monomer	250	0.193	0.131	1.71	0.0040	0.47	0.262
Fitted slope $\partial \log \bar{\alpha}/\partial N$ over the extended ladder			-6.0×10^{-4} residue^{-1} over $N = 129$–250				

J.2 Transfer to unseen soluble structural classes is graded

Table 43 varies structural class in addition to protein size. Coiled-coil packing, an intrinsically disordered region, a disulfide-rich compact domain, an extended loop/hinge, and a beta-solenoid/repeat-rich domain are absent from training. Relative efficiency remains 0.54–0.76 of the held-out BBA reference, with free-energy errors between 0.158 and $0.238\,k_BT$. The ordering is informative: the IDR has higher acceptance than BBA (0.121 versus 0.110) but lower relative efficiency (0.76), showing that acceptance and decorrelation carry complementary information; repeat-rich and disulfide-constrained environments are more difficult because their local geometry is less represented by the training distribution.

Across the size and structural-class extensions, performance changes smoothly with novelty. Global novelty lowers conditional quality and effective-sample throughput, but the same corrected transition remains operational over a substantially wider soluble-protein regime than the training distribution.

Table 43: Transfer to held-out soluble-protein structural classes under GBn2. Relative efficiency is normalized to the BBA reference in this table (1.00); Table 7 uses the in-domain efficiency as its denominator.

Class	Held-out system/region	Class represented in training?	$\bar{\alpha}$ ↑	Cond. NLL ↓	PIT–KS ↓	Relative efficiency vs. BBA ↑	$\lvert\Delta F - \Delta F^\star\rvert/k_BT$ ↓
Soluble globular reference	BBA mini-protein	yes	**0.110**	1.12	0.0157	1.00	0.087
Coiled-coil packing	held-out coiled-coil segment	no	0.094	1.55	0.034	0.71	0.176
Intrinsically disordered region	held-out IDR segment	no	0.121	1.47	0.028	0.76	0.158
Disulfide-rich compact domain	held-out disulfide-rich protein	no	0.082	1.68	0.041	0.60	0.213
Extended loop / hinge	held-out long-loop region	no	0.103	1.59	0.037	0.66	0.194
Beta-solenoid / repeat packing	held-out repeat-rich domain	no	0.076	1.74	0.046	0.54	0.238

K COMPOSITE-KERNEL STEP SIZE AND STATIONARY ACCURACY

The main production chain uses 2 fs Langevin-middle dynamics between exactly corrected block moves. The small-system step-size scan in Table 44 measures the stationary perturbation from the finite-step dynamics leg and its throughput trade-off. At 2 fs, Langevin-middle+BLOCS gives TV 0.0161 and $\epsilon_F = 0.054$, whereas Metropolized dynamics+BLOCS gives TV 0.0077 and $\epsilon_F = 0.026$. The corrected row therefore supplies a low-bias reference for the same block transition.

Table 44: Small-system composite-kernel correctness and step-size scan.

Configuration	Δt (fs)	MD correction	TV ↓	ϵ_F ↓	$\lvert\Delta\langle U\rangle\rvert/(k_BT)$ ↓	ESS/GPU-s ↑
Langevin-middle	0.5	No	0.0081	0.027	0.014	0.38
Langevin-middle	1.0	No	0.0096	0.033	0.026	0.70
Langevin-middle	2.0	No	0.0178	0.061	0.079	1.16
Langevin-middle + BLOCS	0.5	No	0.0069	0.021	0.013	0.84
Langevin-middle + BLOCS	1.0	No	0.0088	0.029	0.028	1.28
Langevin-middle + BLOCS	2.0	No	0.0161	0.054	0.074	1.73
Metropolized dynamics + BLOCS	1.0	Yes	0.0072	0.024	0.012	1.09
Metropolized dynamics + BLOCS	2.0	Yes	0.0077	0.026	0.016	1.24

The protein-scale paired benchmark in Table 45 carries that stationarity control into the production-efficiency comparison. All methods use Metropolized dynamics at $\Delta t = 2$ fs, friction $1.0\,\mathrm{ps}^{-1}$, $T = 300$ K, 100 corrected dynamics steps per block attempt, and `THERMO-COMMON-M12`. Seed i shares its initial configuration, reference observable, and evaluation checkpoints across Corrected MD, Rotamer MH, matched Robin-AR+MH, and BLOCS. Time-to-target is the first checkpoint with $|\Delta F - \Delta F^\star| \le 0.10\,k_BT$ that remains below the threshold for the next three checkpoints; runs not satisfying the criterion within 24 GPU-h are right-censored.

Table 45: **Accuracy-controlled end-to-end efficiency with Metropolized dynamics.** Acceptance, free-energy error, ESS/10^5 evaluations, throughput, and ESS/GPU-s are mean ± s.d. over paired $n = 5$ seeds. Time-to-target is median [IQR], followed by successful runs/5.

System	Method	Block accept. mean±s.d.	$\lvert\Delta F - \Delta F^\star\rvert/k_BT$ mean±s.d. ↓	ESS/10^5 evals mean±s.d. ↑	Evals/GPU-s mean±s.d. ↑	ESS/GPU-s mean±s.d. ↑	GPU-h to $\le 0.10\,k_BT$ median [IQR]; success/5 ↓
Trp-cage	Corrected MD	–	0.320 ± 0.048	120 ± 9	110.0 ± 2.7	0.132 ± 0.010	> 24; 0/5
	Corrected MD + Rotamer MH	0.112 ± 0.008	0.185 ± 0.025	360 ± 18	100.0 ± 2.4	0.360 ± 0.019	8.4[7.7, 9.2]; 5/5
	Corrected MD + Robin-AR (matched)	0.204 ± 0.010	0.082 ± 0.012	820 ± 34	80.0 ± 2.1	0.656 ± 0.031	5.4[5.0, 5.9]; 5/5
	Corrected MD + BLOCS	$\mathbf{0.270 \pm 0.011}$	$\mathbf{0.041 \pm 0.008}$	$\mathbf{1200 \pm 45}$	$\mathbf{86.0 \pm 2.2}$	$\mathbf{1.032 \pm 0.047}$	**3.1[2.8, 3.5]; 5/5**
Protein G	Corrected MD	–	0.410 ± 0.061	96 ± 8	82.0 ± 2.3	0.079 ± 0.007	> 24; 0/5
	Corrected MD + Rotamer MH	0.058 ± 0.006	0.225 ± 0.031	320 ± 17	72.0 ± 1.9	0.230 ± 0.014	12.3[11.1, 13.7]; 5/5
	Corrected MD + Robin-AR (matched)	0.107 ± 0.008	0.094 ± 0.014	720 ± 31	55.0 ± 1.7	0.396 ± 0.023	6.7[6.1, 7.5]; 5/5
	Corrected MD + BLOCS	$\mathbf{0.154 \pm 0.009}$	$\mathbf{0.052 \pm 0.009}$	$\mathbf{1050 \pm 41}$	$\mathbf{61.0 \pm 1.8}$	$\mathbf{0.641 \pm 0.030}$	**4.0[3.6, 4.5]; 5/5**
Lysozyme	Corrected MD	–	0.520 ± 0.074	74 ± 7	57.0 ± 1.8	0.042 ± 0.005	> 24; 0/5
	Corrected MD + Rotamer MH	0.043 ± 0.005	0.270 ± 0.038	265 ± 15	50.0 ± 1.6	0.133 ± 0.009	17.1[15.8, 18.9]; 5/5
	Corrected MD + Robin-AR (matched)	0.079 ± 0.006	0.098 ± 0.016	585 ± 27	38.0 ± 1.3	0.222 ± 0.015	9.8[8.9, 10.9]; 5/5
	Corrected MD + BLOCS	$\mathbf{0.118 \pm 0.007}$	$\mathbf{0.066 \pm 0.011}$	$\mathbf{870 \pm 36}$	$\mathbf{43.0 \pm 1.4}$	$\mathbf{0.374 \pm 0.021}$	**6.0[5.5, 6.8]; 5/5**

Across Trp-cage, Protein G, and lysozyme, BLOCS/Rotamer ESS/GPU-s ratios are $2.87\times$, $2.79\times$, and $2.81\times$; BLOCS/matched-AR ratios are $1.57\times$, $1.62\times$, and $1.68\times$. The fixed-error target shows the same ordering in physical compute: median BLOCS times are 3.1, 4.0, and 6.0 GPU-h, compared with 8.4, 12.3, and 17.1 GPU-h for Rotamer MH. Thus the efficiency gain remains after correcting the dynamics leg and controlling the target accuracy criterion.

Table 46: **Exact-composition implementation.** The same dynamics and block schedule is used for every method in Table 45.

Field	Final setting
Corrected dynamics kernel	Metropolized Langevin-middle control
Time step	2.0 fs
Friction	$1.0\,\mathrm{ps}^{-1}$
Temperature	300 K
Dynamics correction	Metropolis accept/reject step for every finite-step dynamics proposal
Corrected MD steps per block attempt	100
Block interval	$r = 100$
Block population	`THERMO-COMMON-M12`

The population identifier is essential for interpreting acceptance. Protein G and lysozyme acceptance in Table 45 is measured on `THERMO-COMMON-M12`; the lower $m = 12$ acceptance values in Table 51 are measured on `COMPLETE-ELIGIBLE-SCAN`. They characterize different block populations and are used for different scientific comparisons.

L Coordinate-Transform Validation for Production Torsion Moves

For the standard BAT coordinates used by the production torsion-only moves,

$$J(\xi) \propto \prod_i \ell_i^2 \prod_j \sin\theta_j .$$

The proposal preserves the relevant bond lengths ℓ_i and bond angles θ_j, so $J(\xi')/J(\xi) = 1$. Direct numerical checks verify the transform across single-torsion and multi-residue blocks: the maximum Jacobian residual is 2.3×10^{-9}, the maximum angular round-trip error is 1.5×10^{-10} rad, and the largest forward/reverse log-density residual is 1.4×10^{-7}.

Table 47: Coordinate-transform validation for production torsion-only moves.

Move	m	Analytic $\log(J'/J)$	Max numerical residual	Max round-trip error (rad)	Forward/reverse log-density residual
Single side-chain torsion	1	0	2.8×10^{-10}	6.2×10^{-12}	3.6×10^{-8}
Two coupled torsions	2	0	7.4×10^{-10}	2.1×10^{-11}	7.9×10^{-8}
Multi-residue block	4	0	5.9×10^{-10}	4.7×10^{-11}	6.1×10^{-8}
Multi-residue block	8	0	1.6×10^{-9}	8.3×10^{-11}	1.4×10^{-7}
Multi-residue block	12	0	2.3×10^{-9}	1.5×10^{-10}	1.1×10^{-7}

M CAPACITY-MATCHED CONDITIONING-INFORMATION ABLATION

The conditioning ablation holds model scale and training budget fixed while progressively enriching the information available to the proposal. On Protein G, full 3D context raises ESS per 10^5 evals from 322 with current torsions, 397 with local sequence, and 469 with backbone geometry to 612. Shuffling the 3D context reduces ESS per 10^5 evals to 301 and increases ϵ_F to 0.258, directly linking spatial context to transition quality. The same full-context advantage over backbone-only conditioning is 1.28–1.31$\times$ on ubiquitin and lysozyme.

Table 48: Protein G conditioning-information ablation at $m = 4, r = 100$, with capacity and training budget matched across conditioning variants.

Conditioning	Target params	Acceptance ↑	ESJD/accepted ↑	ESS per 10^5 evals ↑	ϵ_F ↓
Current torsions	21.4M	0.148	1.39	322	0.241
Torsions + local sequence	21.4M	0.174	1.62	397	0.188
Torsions + backbone geometry	21.4M	0.203	1.83	469	0.132
Full 3D context	21.4M	**0.238**	**2.12**	**612**	**0.086**
Full model, shuffled context	21.4M	0.139	1.48	301	0.258

Table 49: Cross-system gain from learned 3D context.

System	Backbone-only ESS per 10^5 evals	Full-context ESS per 10^5 evals	Ratio	Full-context ϵ_F
Protein G	469	612	1.30	0.086
Ubiquitin	421	537	1.28	0.109
Lysozyme	354	463	1.31	0.121

N CYCLE-LEVEL RUNTIME AND EXECUTION DECOMPOSITION

The Protein-G ablation at $m = 4, r = 100$ resolves the execution path of one sampling cycle. Both methods perform 101 budgeted target evaluations per cycle. Rotamer MH takes 542.8 ms and BLOCS takes 547.4 ms, corresponding to 186.1 and 184.5 target evaluations/GPU-s. Their execution throughput is therefore nearly identical in this diagnostic; the increase from 0.523 to 1.129 ESS/GPU-s is explained by the increase from 281 to 612 ESS per 10^5 evals.

Table 50: Cycle-level decomposition for the Protein-G ablation at $m = 4, r = 100$. Cycle time includes proposal construction, MD, target evaluation, correction, and state update.

Method	Cycle ms	N_E/cycle	N_E/GPU-s	ESS per 10^5 evals	ESS/GPU-s
Rotamer MH	542.8	101	186.1	281	0.523
BLOCS	**547.4**	101	184.5	**612**	**1.129**
BLOCS / Rotamer	1.01×	1.00×	0.99×	2.18×	2.16×

For this cycle, 100 MD steps plus one block target evaluation give 101 budgeted evaluations, and

$$\frac{\mathrm{ESS}}{\mathrm{GPU}\cdot\mathrm{s}} = \left(\frac{10^5\,\mathrm{ESS}}{N_E}\right)\frac{1}{10^5}\frac{101\times 1000}{t_{\mathrm{cycle,ms}}}.$$

O CROSS-PROTEIN SCAN AT FIXED TORSIONAL DIMENSIONALITY

The complete-pool characterization scan evaluates Protein G, ubiquitin, and lysozyme at $m \in \{2, 4, 8, 12\}$ and $r = 100$. Each (N, m) cell averages over the complete eligible block pool generated by the common construction rule, with one frozen checkpoint, target, and accounting definition throughout the scan. Acceptance decreases with m in every protein, while ESS per 10^5 target evaluations peaks at $m = 4$ with 612 on Protein G, $m = 8$ with 563 on ubiquitin, and $m = 4$ with 463 on lysozyme. At $m = 4$, the complete-pool acceptance values are 0.238, 0.224, and 0.209. Table 36 reports the frozen benchmark pools, and Table 38 calibrates acceptance against those declared populations.

Table 51: **Complete-pool cross-protein scan over torsional proposal dimension.** Each (N, m) cell averages over the complete eligible block pool generated by the common construction rule. All cells use the same frozen checkpoint, $r = 100$, target, and accounting definition. Table 36 gives the observable-specific primary-pool measurements.

System	N_{res}	m	Acceptance ↑	ESS per 10^5 evals ↑	ϵ_F ↓
Protein G	56	2	0.317	474	0.131
Protein G	56	4	0.238	**612**	**0.086**
Protein G	56	8	0.151	558	0.094
Protein G	56	12	0.092	426	0.146
Ubiquitin	76	2	0.301	451	0.142
Ubiquitin	76	4	0.224	537	0.109
Ubiquitin	76	8	0.162	**563**	**0.103**
Ubiquitin	76	12	0.087	397	0.163
Lysozyme	129	2	0.284	409	0.158
Lysozyme	129	4	0.209	**463**	**0.121**
Lysozyme	129	8	0.133	447	0.139
Lysozyme	129	12	0.076	341	0.194

The size contrasts are computed within each moved dimension:

$$\delta_N(m) = 1 - \frac{\bar{\alpha}(N = 129, m)}{\bar{\alpha}(N = 56, m)}.$$

For $m = 2, 4, 8, 12$, the decreases are 10.4%, 12.2%, 11.9%, and 17.4%. The corresponding fixed-protein contrasts $1 - \bar{\alpha}(N, 12)/\bar{\alpha}(N, 2)$ are 71.0%, 71.1%, and 73.2% for Protein G, ubiquitin, and lysozyme. These calculations use the displayed acceptance values and the same pool-construction rule at each (N, m).

P FINITE-BUDGET CONVERGENCE FROM MULTIPLE INITIAL STATES

Ten independent runs from each of three initialization regimes measure time to a common free-energy target. BLOCS reaches the target in 9/10 Basin-A runs, 8/10 Basin-B runs, and 9/10 equilibrium-initialized runs, with median target times of 3.9, 4.6, and 3.8 GPU-h. The corresponding 8-GPU-h median ϵ_F values are 0.088, 0.106, and 0.101.

Table 52: Finite-budget convergence from multiple initial states.

Initial state	Method	Median ϵ_F at 8 GPU-h	Successful runs / 10	Median time to target (GPU-h)
Basin A	Handcrafted packing	0.173	4/10	> 8
Basin A	Robin-local+MH	0.121	7/10	5.7
Basin A	BLOCS	**0.088**	**9/10**	**3.9**
Basin B	Handcrafted packing	0.151	5/10	7.9
Basin B	Robin-local+MH	0.137	6/10	6.4
Basin B	BLOCS	**0.106**	**8/10**	**4.6**
Independent equilibrium	Handcrafted packing	0.129	7/10	5.3
Independent equilibrium	Robin-local+MH	**0.097**	**9/10**	**3.6**
Independent equilibrium	BLOCS	0.101	**9/10**	3.8

Q EQUAL-PILOT-BUDGET BLOCK SELECTION

The candidate set $m \in \{2, 4, 8, 12\}$ is evaluated under the same 0.65 GPU-h pilot budget on four prospective proteins. Gaussian overlap combined with a move-utility model selects configurations retaining 0.967, 1.000, 0.900, and 1.000 of the independently validated best efficiency on BPTI, barnase, RNase A, and calmodulin. Short-run ESS selection reaches 1.000, 0.945, 1.000, and 0.930, while direct acceptance ranking consistently selects $m = 2$. The experiment shows that combining overlap with move utility captures the acceptance–displacement tradeoff within a fixed pilot budget.

Table 53: Operating-point selection under an equal pilot budget.

System	Selection strategy	Pilot GPU-h	Selected m	ESS/GPU-s	Efficiency / best
BPTI	Direct acceptance ranking	0.65	2	0.67	0.736
BPTI	Gaussian overlap + utility model	0.65	4	0.88	0.967
BPTI	Short-run ESS selection	0.65	8	0.91	1.000
Barnase	Direct acceptance ranking	0.65	2	0.52	0.712
Barnase	Gaussian overlap + utility model	0.65	8	0.73	1.000
Barnase	Short-run ESS selection	0.65	4	0.69	0.945
RNase A	Direct acceptance ranking	0.65	2	0.46	0.767
RNase A	Gaussian overlap + utility model	0.65	4	0.54	0.900
RNase A	Short-run ESS selection	0.65	8	0.60	1.000
Calmodulin	Direct acceptance ranking	0.65	2	0.39	0.684
Calmodulin	Gaussian overlap + utility model	0.65	8	0.57	1.000
Calmodulin	Short-run ESS selection	0.65	4	0.53	0.930

The direct acceptance estimator on the same attempted pairs is

$$\widehat{\alpha}_{\mathrm{direct}} = \frac{1}{n} \sum_{i=1}^{n} \min\{1, \exp(S_i)\},$$

providing a distribution-free calibration reference for the Gaussian overlap model.

R State Conditioning and Autoregressive Factorization

A 2×2 ablation separates the current-state-conditioned residual representation from cross-dimension autoregressive coupling while holding 3D context, training budget, and target model scale fixed. Under the Protein-G ablation protocol ($m = 4$, $r = 100$), context-only independent factors yield 284 ESS per 10^5 evals; autoregression raises this to 418. Current-state residual conditioning with independent dimensions reaches 493, and combining both components reaches 612 with $\epsilon_F = 0.086$.

Table 54: 2×2 ablation of state conditioning and dimension factorization.

State conditioning	Dimension factorization	Acceptance ↑	ESJD/accepted ↑	ESS per 10^5 evals ↑	ϵ_F ↓
Context only	Independent dimensions	0.094	2.76	284	0.261
Context only	Autoregressive	0.156	2.53	418	0.167
Current state + context, residual	Independent dimensions	0.226	1.69	493	0.119
Current state + context, residual	**Autoregressive**	**0.238**	2.12	**612**	**0.086**

The two factorizations are

$$q_{\mathrm{ind}}(\Delta\chi \mid \chi, c_B) = \prod_{d=1}^{m} q_d(\Delta\chi_d \mid \chi, c_B),$$

and

$$q_{\mathrm{AR}}(\Delta\chi \mid \chi, c_B) = \prod_{d=1}^{m} q_d(\Delta\chi_d \mid \Delta\chi_{<d}, \chi, c_B).$$