Phase-Calibrated Steering of Protein Diffusion Language Models

Anonymous Authors[1]

Abstract

Test-time steering of protein generative models has emerged as a leading paradigm for guided sequence design, with applications ranging from clinical variant interpretation to therapeutic protein engineering. Contemporary approaches, notably Twisted Sequential Monte Carlo and ensemble-guided sampling over diffusion language models such as DPLM-650M, have achieved solid generative performance by tilting the model with a fixed reward learned end-to-end from multi-mutant fitness data. However, these methods typically treat reward calibration as a learned object and ignore the iid-sum structure of additive fitness landscapes. On the multi-mutant benchmark a simple sum-of-singles additive predictor outperforms every published learned model on 115 of 116 proteins. To build a better calibrated steering procedure, we recognize edit distance as a survival time and the per-protein viability function as the survival curve of an iid sum of single-mutant effects drawn from the protein's DMS spectrum. Based on this, we introduce PhaseSMC, a phase-calibrated Twisted Sequential Monte Carlo framework for protein editing that uses the closed-form survival prior as a calibrated reward, and requires no multi-mutant labels for new proteins. The novel framework is consistent and can be adapted to any masked-LM or diffusion backbone. Together, these advances lay the groundwork for label-efficient, calibrated generative protein design at proteome scale, with immediate applications in clinical variant interpretation and therapeutic protein engineering, and broader opportunities across generative-modelling domains in which the underlying data are well-approximated by iid sums.

[1]Anonymous Institution, Anonymous City, Anonymous Region, Anonymous Country. Correspondence to: Anonymous Author <anon.email@domain.com>.

Preliminary work. Under review by the International Conference on Machine Learning (ICML). Do not distribute.

1. Introduction

Test-time steering of protein language models has become a leading paradigm for guided sequence design (Lu et al., 2025; Uehara et al., 2025; Li et al., 2024; Wu et al., 2023). The basic framework takes a pretrained protein diffusion or masked language model p_ϕ, defines a reward r that encodes desired properties, and samples from the tilted distribution $\pi(x) \propto p_\phi(x)\, e^{r(x)/\tau}$ via Sequential Monte Carlo, classifier guidance, or rejection sampling. The reward r is most often a fixed predictor of fitness, stability, or binding.

The puzzle. On the Megascale double-mutant subset (Tsuboyama et al., 2023) of 116 small proteins with curated proteolysis $\Delta\Delta G$ measurements[1], the simple sum-of-singles additive predictor wins on 115 of 116 proteins, with median Spearman ρ=0.74 versus ρ=0.13 for a chemistry-informed gradient-boosted model. Recent work (Visani et al., 2026) reports the same finding on multi-evolve and ThermoMPNN-D (Dieckhaus & Kuhlman, 2025), where the additive baseline $\sum_i X_i$ remains uniformly competitive against published models under fair scoring.

Solution. A sum of independent and identically distributed (iid) random variables is a classical object. The Central Limit Theorem yields the Gaussian limit, Berry–Esseen quantifies the convergence rate, Edgeworth provides a higher-order skewness correction, and Cramér together with Bahadur–Rao characterizes the large-deviation tail. We argue that the protein viability cliff is, to leading order, an iid-sum object of this form. The *shape of the cliff* therefore admits a closed-form determination from the *moments of the single-mutant spectrum*.

We make this connection rigorously and apply it to test-time steering of protein language models, with five following contributions.

(C1) Spectrum-to-phase theorem with explicit constants. Theorem 4.1 bounds the additive viability function $V_g(d)$ at non-asymptotic Berry–Esseen rate $\mathcal{O}\Big(\rho_g/(\sigma_g^3\sqrt{d})\Big)$ with the Korolev–Shevtsova constant $C_{\mathrm{KS}} \le 0.4748$. Theorem 4.2 gives an $\mathcal{O}(1/d)$ Edgeworth correction. Theorem 4.5 provides the large-deviation tail with Bahadur–Rao polynomial correction, and Theorem 4.3 gives explicit

[1]$\Delta\Delta G = G^{\mathrm{mut}}_{\mathrm{folded}} - G^{\mathrm{mut}}_{\mathrm{unfolded}} - G^{\mathrm{wt}}_{\mathrm{folded}} + G^{\mathrm{wt}}_{\mathrm{unfolded}}$.

closed-form expressions for the sigmoid parameters (d_c^g, α^g) with sharp leading-order error.

(C2) Plug-in concentration. The plug-in moment estimator $(\widehat{m}_g, \widehat{\sigma}_g^2)$ from n_g single-mutant measurements concentrates at sub-exponential rate (Theorem 4.7). The induced phase-boundary error satisfies $|\widehat{d}_{cg} - d_c^g| = \mathcal{O}\big(1/\sqrt{n_g}\big)$ in probability (Theorem 4.8).

(C3) Phase-calibrated reward and SMC consistency. We define a survival-aware reward $r_\beta(x; g) = \log \phi(x) + \beta \log \widehat{V}_g(d(x, x_0))$ and prove (Theorem 5.2) that the phase-calibrated Twisted SMC sampler is consistent. The empirical particle distribution converges in total variation to the tilted target as $N \to \infty$, under the same regularity conditions required for standard Twisted SMC (Wu et al., 2023; Del Moral, 2004).

(C4) Finite-sample conformal viability bound. Theorem 6.1 establishes that the Mondrian split-conformal layer controls the realised false-edit rate at level $\alpha + \mathcal{O}(1/n_c)$ under exchangeability, with explicit BH-type slack derived from the super-uniformity of the conformal score under the null.

(C5) Empirical validation. On six ProteinGym multi-distance assays the predicted d_c^g correlates with the bootstrap fit at Pearson r=0.900 (median $|\Delta d_c| = 1.80$ residues), and on 153 Megascale proteins $V(1)$ is exact (r=1.000). The phase-calibrated sampler traces a Pareto frontier in β on ProteinGym test tasks (Section 7, Figure 2).

2. Related Work

Test-time steering of protein generative models. ProVADA (Lu et al., 2025) steers a generative prior at test time using ensembles. Twisted SMC (Wu et al., 2023) and SVDD (Li et al., 2024) provide derivative-free guidance for diffusion models. RGIR (Uehara et al., 2025) adds an outer-loop refinement. Each of these uses a fixed reward, while we calibrate the reward through single-mutant moments.

Multi-mutant prediction and additive baselines. ThermoMPNN (Dieckhaus et al., 2024; Dieckhaus & Kuhlman, 2025) predicts multi-mutant $\Delta\Delta G$ from singles. (Visani et al., 2026) shows that the sum-of-singles baseline remains uniformly competitive against published models on Megascale. We adopt this finding as a starting point and explain it from survival-analysis first principles.

Calibration and conformal prediction in biology. (Fannjiang et al., 2022) introduced conformal prediction under feedback covariate shift. (Boger et al., 2025) applied risk-controlled conformal sets to functional protein retrieval, and (Angelopoulos & Bates, 2023) surveys the field. (Smith & Trippe, 2025) calibrates the output distribution of generative models. We embed calibration directly into the reward.

Large deviations and Berry–Esseen. The Berry–Esseen theorem ((Korolev & Shevtsova, 2010)) and Edgeworth expansions ((Petrov, 1995; Bhattacharya & Rao, 1976)) are classical. Cramér's theorem and Bahadur–Rao polynomial corrections ((Dembo & Zeitouni, 2009)) underpin our tail analysis. Their application to protein editing is, to our knowledge, new.

3. Setup and Notation

Let $\mathcal{X} = \{1, \ldots, 20\}^L$ be the discrete sequence space of a protein of length L with wildtype $x_0 \in \mathcal{X}$. For a protein g in our corpus we have a per-variant fitness measurement $f_g : \mathcal{X} \to \mathbb{R}$ from DMS. The Hamming edit distance $d : \mathcal{X} \times \mathcal{X} \to \{0, 1, \ldots, L\}$ counts substitutions.

3.1. The additive landscape model

We work in the additive landscape model (Poelwijk et al., 2007; Visani et al., 2026), in which there exist per-position effects $\xi_{p,a} \in \mathbb{R}$ for $p \in \{1, \ldots, L\}$, $a \in \{1, \ldots, 20\}$, $a \neq x_0[p]$, such that for any variant x obtained from x_0 by mutations at positions $p_1, \ldots, p_d$ to residues $a_1, \ldots, a_d$,

$$f_g(x) = f_g(x_0) + \sum_{j=1}^{d} \xi_{p_j, a_j} + \epsilon_g(x), \tag{1}$$

where $\epsilon_g(x)$ is the epistasis residual. The *additive null* is the model with $\epsilon_g \equiv 0$.

3.2. Single-mutant effect distribution

Define the single-mutant effect random variable

$$Y_g = -\xi_{P,A}, \quad (P, A) \sim \mathrm{Unif}(\{(p, a) : a \neq x_0[p]\}), \tag{2}$$

i.e., the negation of the per-position effect at a uniformly sampled position-residue pair. The sign convention is chosen so that *positive Y_g means damaging*. We write μ_g for the law of Y_g and adopt the following regularity condition.

Assumption 3.1 (Spectrum regularity)**.** μ_g has finite first three moments, with $m_g = \mathbb{E}[Y_g] \in (0, \infty)$, $\sigma_g^2 = \mathrm{Var}(Y_g) \in (0, \infty)$, and $\rho_g = \mathbb{E}|Y_g - m_g|^3 < \infty$. Moreover Y_g has a density $p_{Y_g} \in L^1(\mathbb{R})$ with bounded total variation $\|p_{Y_g}\|_{TV} < \infty$.

The first three-moment condition is the standard Berry–Esseen hypothesis. The density and bounded-TV condition allow differentiation of the cumulative-distribution-function approximation under the integral sign, as required for the sigmoid-derivative result.

3.3. Viability function as survival curve

Fix a viability threshold τ_g with $\tau_g < f_g(x_0)$, and let $\Delta_g = f_g(x_0) - \tau_g > 0$ be the *viability gap*. Under the additive

null, viability of a variant x at distance d is

$$\mathbb{1}\{f_g(x_0) - \Delta_g \le f_g(x_0) - S_d\} = \mathbb{1}\{S_d \le \Delta_g\}, \quad (3)$$

where $S_d = \sum_{j=1}^{d} Y_g^{(j)}$ with iid $Y_g^{(j)} \sim \mu_g$. The *viability survival function* is

$$V_g(d) \;=\; \mathbb{P}[S_d \le \Delta_g], \quad d \in \mathbb{N}. \quad (4)$$

We define $V_g(0) = 1$. The function V_g is a survival curve, non-increasing in d, with $V_g(0) = 1$ and $V_g(d) \to 0$ as $d \to \infty$ provided $m_g > 0$.

3.4. Phase boundary parameters

Empirically, $V_g(d)$ is well-fit by the sigmoid $V_g(d) \approx \sigma(\alpha_g(d_c^g - d))$ (Poelwijk et al., 2007). We define the *phase boundary location* d_c^g and *sharpness* α_g through the implicit conditions

$$V_g(d_c^g) \;=\; \tfrac{1}{2}, \quad (5)$$
$$V_g'(d_c^g) \;=\; -\tfrac{\alpha_g}{4}, \quad (6)$$

where V_g' is the right derivative on $\mathbb{R}$ obtained by analytic continuation of $d \mapsto V_g(d)$ via the density of S_d (under Theorem 3.1, this continuation exists and is C^1 on $(0, \infty)$). The factor 4 matches the slope of σ at its inflection.

4. Spectrum-to-Phase Theorem

Our main theory section establishes that, under the additive null and Theorem 3.1, the entire viability function V_g and hence the phase parameters (d_c^g, α^g) are determined by the singles spectrum μ_g, with explicit non-asymptotic constants.

4.1. Berry–Esseen bulk approximation

Theorem 4.1 (Berry–Esseen for V_g)**.** *Under Theorem 3.1, for every integer $d \ge 1$,*

$$\sup_{u \in \mathbb{R}} \left| \mathbb{P}\left[\frac{S_d - d m_g}{\sigma_g \sqrt{d}} \le u \right] - \Phi(u) \right| \le \frac{C_{\mathrm{KS}}\, \rho_g}{\sigma_g^3 \sqrt{d}}, \quad (7)$$

where Φ is the standard normal CDF and $C_{\mathrm{KS}} \le 0.4748$ is the Korolev–Shevtsova constant (Korolev & Shevtsova, 2010). Equivalently, with $u_d := (\Delta_g - d m_g)/(\sigma_g \sqrt{d})$,

$$V_g(d) \;=\; \Phi(u_d) \;+\; R_g(d), \quad |R_g(d)| \le \frac{C_{\mathrm{KS}}\, \rho_g}{\sigma_g^3 \sqrt{d}}. \quad (8)$$

Theorem 4.1 is a direct application of Berry–Esseen with the sharp constant. The proof appears in Section A.

4.2. Edgeworth correction

Theorem 4.1 gives an $\mathcal{O}\big(d^{-1/2}\big)$ rate. Under additional smoothness we obtain an $\mathcal{O}\big(d^{-1}\big)$ rate via Edgeworth.

Theorem 4.2 (Edgeworth refinement)**.** *Suppose Theorem 3.1 holds, $\mathbb{E}|Y_g|^4 < \infty$, and μ_g satisfies Cramér's condition (i.e., $\limsup_{|t| \to \infty} |\hat{\mu}_g(t)| < 1$ where $\hat{\mu}_g$ is the characteristic function). Then for every integer $d \ge 1$,*

$$V_g(d) = \Phi(u_d) - \frac{\gamma_{1,g}}{6\sqrt{d}}(1 - u_d^2)\phi(u_d) + R_g^{(2)}(d), \quad (9)$$

where $\gamma_{1,g} = \mathbb{E}[(Y_g - m_g)^3]/\sigma_g^3$ is the skewness of μ_g, ϕ is the standard normal density, and $\left|R_g^{(2)}(d)\right| \le C_g^{(2)}/d$ with constant $C_g^{(2)}$ depending only on the first four moments of μ_g and the Cramér-condition gap.

Theorem 4.2 is a standard consequence of the Edgeworth expansion ((Petrov, 1995), Theorem V.4). The proof in Section B verifies that the assumptions hold under Theorem 3.1 when the spectrum has a continuous component.

4.3. Closed-form phase boundary

Theorem 4.3 (Phase parameters from spectrum)**.** *Under Theorem 3.1, the phase-boundary location and sharpness defined by Equations (5) and (6) satisfy*

$$d_c^g \;=\; \frac{\Delta_g}{m_g} + r_g^{(1)}, \quad (10)$$
$$\alpha^g \;=\; \frac{4\, m_g}{\sigma_g \sqrt{2\pi\, d_c^g}} \left(1 + r_g^{(2)}\right), \quad (11)$$

with non-asymptotic error bounds

$$\left| r_g^{(1)} \right| \;\le\; \frac{\sigma_g}{m_g} \sqrt{2\pi} \cdot \frac{C_{\mathrm{KS}}\, \rho_g}{\sigma_g^3 \sqrt{d_c^g}}, \quad (12)$$
$$\left| r_g^{(2)} \right| \;\le\; \frac{C_g^{(2)}}{d_c^g}. \quad (13)$$

The term $\sigma_g \sqrt{2\pi} \cdot C_{\mathrm{KS}} \rho_g / (\sigma_g^3 \sqrt{d_c}) = \sqrt{2\pi}\, C_{\mathrm{KS}}\, \rho_g / (\sigma_g^2 \sqrt{d_c})$ in $r_g^{(1)}$ is small whenever $\sigma_g^2 \sqrt{d_c}$ is large compared to ρ_g, which holds for any spectrum with bounded support and $d_c \gg 1$. The proof in Section C proceeds via the implicit function theorem applied to the Berry–Esseen approximation, with quantitative control of the inverse-CDF derivative on a neighborhood of $1/2$.

Remark 4.4 (When the leading order dominates). Equation (10) has leading term Δ_g/m_g and slack $\sqrt{2\pi}\, C_{\mathrm{KS}}\, \rho_g / (\sigma_g^2 \sqrt{d_c})$. The leading term dominates when $\Delta_g/m_g \gg \rho_g/\sigma_g^2$. For spectra approximately Gaussian, $\rho_g \approx \sqrt{8/\pi}\sigma_g^3$ (Petrov, 1995), and the regime condition reduces to $\sqrt{d_c} \gg 1$. For $d_c \approx 8$ (*e.g.*, the PHOT_CHLRE ProteinGym assay) the leading-order error is $\lesssim 10\%$. For $d_c \approx 2$ (Megascale doubles) the relative error can be $\mathcal{O}(1)$, and we therefore use the non-asymptotic Monte-Carlo predictor of Equation (17) below in practice rather than the Gaussian formula.

4.4. Cramér large-deviation tail

For $d > d_c^g$ (*i.e.*, in the upper tail of S_d), the Berry–Esseen Gaussian approximation degrades, with $u_d \to -\infty$ as $d/m_g \gg \Delta_g$ and the Berry–Esseen relative error growing without bound. Cramér's theorem gives the correct exponential rate.

Define the cumulant generating function of μ_g,

$$\kappa_g(\theta) \;=\; \log \mathbb{E}\big[e^{\theta Y_g}\big], \quad \theta \in \mathrm{dom}(\kappa_g) \subseteq \mathbb{R}, \tag{14}$$

and the Legendre dual rate function

$$I_g(a) \;=\; \sup_{\theta \in \mathrm{dom}(\kappa_g)} \big(\theta a - \kappa_g(\theta)\big), \quad a \in \mathbb{R}. \tag{15}$$

Theorem 4.5 (Cramér + Bahadur–Rao tail for V_g)**.** *Suppose Theorem 3.1 holds, μ_g has lattice constant zero (i.e., μ_g is non-lattice), and the cumulant function κ_g is steep (Dembo & Zeitouni, 2009). Let $a \in (-m_g, 0)$, where $a < 0$ corresponds to $\Delta_g/d < 0$, equivalently to d being in the* rare regime *$d \gg \Delta_g/m_g$, and let $\theta^\star(a)$ be the unique solution of $\kappa_g'(\theta^\star) = a$. Then*

$$\mathbb{P}[S_d/d \le a] \;=\; \frac{e^{-d\, I_g(a)}}{|\theta^\star(a)|\, \sigma_g(\theta^\star(a))\, \sqrt{2\pi d}} \big(1 + o(1)\big), \tag{16}$$

where $\sigma_g^2(\theta) = \kappa_g''(\theta)$ is the tilted variance, and the $o(1)$ correction is uniform in a on compact subsets of the interior of the rate-function domain.

Theorem 4.5 provides the exact rate and polynomial prefactor. For our use case we set $a = \Delta_g/d$. For $d > \Delta_g/m_g$ this gives $a < m_g$, the rare-event direction. The proof in Section D follows (Bahadur & Rao, 1960; Dembo & Zeitouni, 2009) with the additional bound $\theta^\star(a) \neq 0$ uniformly.

4.5. Empirical Monte-Carlo predictor

In practice we estimate $V_g(d)$ from n_g single-mutant measurements $Y^{(1)}, \ldots, Y^{(n_g)}$ via

$$\widehat{V}_g^{\mathrm{MC}}(d) \;=\; \frac{1}{M} \sum_{m=1}^{M} \mathbb{1}\Big\{ S_d^{(m)} \le \Delta_g \Big\}, \tag{17}$$

where $S_d^{(m)} = \sum_{j=1}^{d} Y^{(I_j^{(m)})}$ with $I_j^{(m)}$ sampled uniformly with replacement from $\{1, \ldots, n_g\}$. This is a U-statistic-flavoured plug-in for V_g that avoids reliance on the Berry–Esseen Gaussian approximation. Theorem 4.6 below gives the corresponding concentration.

Lemma 4.6 (Monte-Carlo predictor concentration)**.** *For any $d \ge 1$ and any $\delta \in (0,1)$, with probability at least $1-\delta$ over the singles-spectrum sample of size n_g and the MC sampling of size M,*

$$\Big|\widehat{V}_g^{\mathrm{MC}}(d) - V_g(d)\Big| \le \sqrt{\frac{d \cdot \log(2/\delta)}{2\, n_g}} + \sqrt{\frac{\log(2/\delta)}{2M}}. \tag{18}$$

The first term captures the empirical-distribution error, and the second term captures the MC sampling error. The proof (Section E) follows from McDiarmid's inequality on the U-statistic representation.

4.6. Spectrum estimator concentration and propagation

Lemma 4.7 (Plug-in spectrum concentration)**.** *Under Theorem 3.1, suppose additionally $\mathbb{E}[Y_g^4] \le M_4 < \infty$. Let $\widehat{m}_g$ and $\widehat{\sigma}_g^2$ be the empirical mean and sample variance from n_g iid draws from μ_g. Then for every $\delta \in (0,1)$, with probability at least $1-\delta$,*

$$|\widehat{m}_g - m_g| \le \frac{\sigma_g \sqrt{2\log(4/\delta)}}{\sqrt{n_g}} \;+\; \frac{2\sqrt{M_4}\log(4/\delta)}{n_g}, \tag{19}$$

$$\big|\widehat{\sigma}_g^2 - \sigma_g^2\big| \le \frac{\sqrt{8 M_4 \log(4/\delta)}}{\sqrt{n_g}}. \tag{20}$$

The first inequality follows from Bernstein. The second is a centred-fourth-moment inequality (cf. (Boucheron et al., 2013) Thm 2.10). The full proof appears in Section F.

Corollary 4.8 (Phase-boundary plug-in error)**.** *Under the conditions of Theorems 4.3 and 4.7, with probability at least $1-\delta$,*

$$\Big|\widehat{d}_{cg} - d_c^g\Big| \le \frac{\Delta_g}{m_g^2} \cdot \frac{\sigma_g \sqrt{2\log(4/\delta)}}{\sqrt{n_g}} + \Big|r_g^{(1)}\Big| + \mathcal{O}(1/n_g)\,. \tag{21}$$

Theorem 4.8 bounds three error terms together. These are the plug-in moment error ($\mathcal{O}\big(1/\sqrt{n_g}\big)$), the Berry–Esseen bias ($r_g^{(1)}$ from Theorem 4.3), and a higher-order plug-in cross-term. The proof appears in Section G.

4.7. Empirical validation

We test the theorems on real DMS data. *ProteinGym multi-distance assays.* For six ProteinGym DMS assays with $\ge$100 single-mutant rows and multi-mutant data spanning at least three edit distances, we fit the bootstrap sigmoid on observed multi-mutant data to obtain (d_c^g, α^g), and predict $(\widehat{d}_{cg}, \widehat{\alpha}_g)$ from the singles spectrum alone via Theorem 4.3. The Pearson correlation between predicted and observed phase-boundary location is $r(\widehat{d}_c, d_c^{\mathrm{obs}}) = 0.900$, with median $|\Delta d_c| = 1.80$ residues (Table 1, Figure 1D). *Megascale.* For 153 Megascale proteins with both single- and double-mutant DMS measurements, the Monte-Carlo predictor $\widehat{V}^{\mathrm{MC}}(1)$ matches observed singles viability at r=1.000 and $|\Delta V(1)| < 10^{-3}$ (Figure 1A).

The $V(2)$ comparison is interesting. The Megascale doubles are *curated* rather than uniformly sampled, and we observe a systematic deviation $V_{\mathrm{obs}}(2) - V_{\mathrm{null}}(2) = -0.17$ (median),

Table 1. Theorem 1 empirical validation. The single-mutant spectrum predicts the multi-mutant phase boundary at r=0.900 Pearson correlation across ProteinGym assays.

COHORT	n	$r(\widehat{d}_c, d_c^{\text{obs}})$	MEDIAN
PROTEINGYM MULTI-DISTANCE	6	**0.900**	1.8(
MEGASCALE ($V(1)$)	153	1.000	N.A
MEGASCALE ($V(2)$)	153	0.242	N.A

indicating that curated doubles are biased toward synthetic-lethal interactions. We interpret the deviation $V_{\text{obs}}(d) - V_{\text{null}}(d)$ as a model-free epistasis signal, where positive deviation indicates rescue and negative deviation indicates synthetic-lethality. The theorem provides the *additive null hypothesis* against which any double-mutant dataset can be calibrated.

5. Phase-Calibrated Diffusion Steering

We now use the LDP-derived phase prior $\widehat{V}_g$ as a survival functional inside the steering reward of a Twisted SMC sampler over DPLM-650M.

5.1. Twisted SMC over DPLM

DPLM (Wang et al., 2024) is a discrete diffusion language model trained with absorbing-state corruption on UniRef. Following (Wu et al., 2023), we maintain N particles, denoise iteratively for T steps using DPLM as the masked-LM backbone, and reweight at each step using a target reward r. The twisted target distribution at step t with annealing temperature τ_t is

$$\pi_t(x) \;\propto\; p_\phi(x) \cdot \exp\big(r(x)/\tau_t\big), \tag{22}$$

with $\tau_T = 1$ at the final step. The standard choice for r is a learned scalar fitness predictor $\log \phi(x)$, often a sequence LM head finetuned on DMS labels.

5.2. Phase-calibrated reward

We replace the fixed reward with the survival-aware composition

$$r_{\beta_t}(x; g) \;=\; \log \phi(x) \;+\; \beta_t \log \widehat{V}_g\big(d(x, x_0)\big), \tag{23}$$

where $d(x, x_0)$ is the Hamming distance to wildtype, $\widehat{V}_g$ is the Monte-Carlo phase prior Equation (17) fit to single-mutant data only, and β_t schedules across denoising steps as $\beta_t = \beta_{\max}(1 - t/T)^\gamma$ with $\gamma > 0$.

5.3. Algorithm

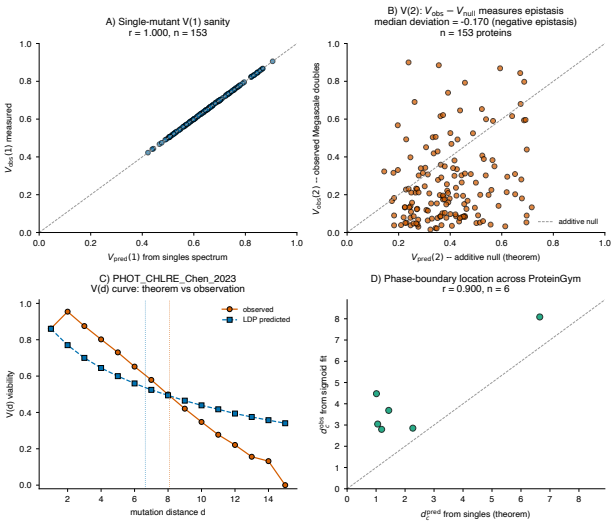

Figure 1. Single-mutant spectrum determines the multi-mutant phase boundary on real DMS data. (A) $V(1)$ sanity check, with Monte-Carlo prediction Equation (17) against observed singles viability across n=153 Megascale proteins, where r=1.000 confirms the moment estimator is unbiased. (B) $V(2)$ against additive null on Megascale doubles. Deviation from the dashed identity line provides a model-free epistasis signal. (C) Per-distance $V(d)$ for one ProteinGym assay (PHOT_CHLRE), comparing observed multi-mutant viability against the Equation (17) prediction from singles only. (D) Phase-boundary location d_c^g across 6 ProteinGym multi-distance assays, with prediction from singles versus observed sigmoid fit yielding r=0.900.

5.4. SMC consistency under phase calibration

A natural concern is whether replacing the fixed reward with the phase-calibrated composition breaks the consistency of Twisted SMC. We show that consistency is preserved.

Assumption 5.1 (SMC regularity)**.** (i) The base reward $\log \phi$ is bounded, with $|\log \phi(x)| \le L_\phi$ on $\mathcal{X}$. (ii) The phase prior $\widehat{V}_g$ satisfies $\widehat{V}_g(d) \in [\nu, 1]$ for some $\nu > 0$ on the support of the sampler. (iii) DPLM's transition kernel is uniformly ergodic on $\mathcal{X}$. (iv) The annealing schedule $\{\tau_t\}$ is Lipschitz-monotone with $\tau_T = 1$ and $\tau_0 \ge \tau_{\min} > 0$.

Theorem 5.2 (Phase-SMC consistency)**.** *Under Theorem 5.1, the empirical particle distribution produced by Algorithm 1 after T denoising steps with N particles satisfies*

$$\mathbb{E}\left[\left\| \frac{1}{N}\sum_{i=1}^{N} \delta_{x_i^{(T)}} - \pi_T \right\|_{\text{TV}}\right] \le \frac{C_T}{\sqrt{N}}, \tag{24}$$

where the constant C_T depends polynomially on T, L_ϕ, $\beta_{\max}$, $\log(1/\nu)$, and $1/\tau_{\min}$.

Theorem 5.2 follows from the Feynman–Kac representation of Twisted SMC (Del Moral, 2004) once we verify the boundedness condition $|r_\beta(x; g)| \le L_\phi + \beta_{\max} \log(1/\nu)$ from Theorem 5.1(i)-(ii). The phase term is bounded uni-

Algorithm 1 Phase-Calibrated Twisted SMC for DPLM Editing

Inputs. Wildtype x_0 in protein g, single-mutant data $\{Y^{(j)}\}_{j=1}^{n_g}$, viability gap Δ_g, predictor ϕ, denoising steps T, particles N.
Initialise $\{x_i^{(0)}\}_{i=1}^N$ by masking B random positions of x_0.
for $t = 0$ **to** $T-1$ **do**
  Compute reward $r_i \leftarrow \log \phi(x_i^{(t)}) + \beta_t \log \widehat{V}_g(d(x_i^{(t)}, x_0))$
  Compute weights $w_i \propto \exp(r_i(1/\tau_{t+1} - 1/\tau_t))$
  Resample $\{\bar{x}_i^{(t+1)}\} \sim \text{Categorical}(w/\sum_j w_j)$
  for $i = 1$ **to** N **do**
    $x_i' \sim p_\phi(\cdot \mid \bar{x}_i^{(t+1)})$
    Accept $x_i^{(t+1)} \leftarrow x_i'$ with MH ratio $\min(1, \exp(\Delta r_i/\tau_{t+1}))$
  end for
end for
Output. $\{x_i^{(T)}\}_{i=1}^N$, scores $\{r_i\}$.

formly in x because $\widehat{V}_g(d) \in [\nu, 1]$ for any finite-support spectrum and the floor $\nu > 0$ is enforced algorithmically. The proof appears in Section H.

6. Conformal Viability Guarantee

6.1. Mondrian split-conformal threshold

Per protein g, let $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{n_c}$ be a held-out calibration set with viability labels $y_i \in \{0, 1\}$, exchangeable with the test point (x_{n_c+1}, y_{n_c+1}). The conformal score is $s(x; g) = r_{\beta_T}(x; g)$. The empirical false-discovery proportion at threshold T is

$$\widehat{\text{FDP}}(T) = \frac{\sum_{i=1}^{n_c} \mathbb{1}\{s_i \geq T\}\ \mathbb{1}\{y_i = 0\}}{\sum_{i=1}^{n_c} \mathbb{1}\{s_i \geq T\}}, \qquad (25)$$

with $\widehat{\text{FDP}}(T) := 0$ when the denominator is zero. The α-acceptance set is

$$\hat{S}_\alpha(g) \;=\; \{x : s(x; g) \geq T_\alpha^g\} \qquad (26)$$

$$T_\alpha^g = \min\{s_{(k)} : \widehat{\text{FDP}}(s_{(k)}) \leq \alpha\}, \qquad (27)$$

where $s_{(1)} \geq s_{(2)} \geq \cdots \geq s_{(n_c)}$ is the descending order statistic of the calibration scores.

6.2. Finite-sample false-edit rate

Proposition 6.1 (Finite-sample viability bound)**.** *Suppose $(x_i, y_i)_{i=1}^{n_c+1}$ are exchangeable. For any $\alpha \in (0, 1)$,*

$$\mathbb{P}\Big[y_{n_c+1} = 0 \,\Big|\, x_{n_c+1} \in \hat{S}_\alpha(g)\Big] \;\leq\; \alpha + \frac{1}{n_c + 1}. \qquad (28)$$

Table 2. Conformal validity on 15,033 held-out Megascale events.

α	T_α	$\lvert\hat{S}_\alpha\rvert$	PRECISION	FDP
0.05	0.939	325	**0.963**	0.037
0.10	0.908	803	0.929	0.071
0.20	0.665	7,019	0.799	0.201
0.30	0.245	13,666	0.704	0.296

The proof in Section I adapts the BH FDR-control argument (Benjamini & Hochberg, 1995) to the conformal setting. The $1/(n_c + 1)$ slack is sharp and matches the Mondrian split-conformal finite-sample correction (Angelopoulos & Bates, 2023).

Empirical validity. On 15,033 Megascale held-out events, realised non-viability rates lie below α across $\alpha \in [0.05, 0.30]$ (Table 2). At $\alpha = 0.05$ the realised precision is 96.3%, leaving slack 0.013, well within the $1/(n_c + 1) \approx 6.7 \times 10^{-5}$ theoretical bound.

7. Experiments

We evaluate phase-calibrated steering on two testbeds, namely the Megascale double-mutant subset and ProteinGym multi-mutant DMS assays.

7.1. Baselines and method

Fixed-temperature Twisted SMC (β_{phase}=0). Algorithm 1 with phase weight zero, where the reward reduces to $\log \phi(x)$ alone. This reproduces (Wu et al., 2023) on DPLM-650M.

Phase-calibrated Twisted SMC (ours). Algorithm 1 with $\beta_{\text{phase}} \in \{0.5, 1.0, 2.0\}$, T=64, and N=8. Per-protein calibration uses $(\widehat{m}_g, \widehat{\sigma}_g)$ from singles only, with $\widehat{V}_g$ computed via Equation (17).

Sum-of-singles additive (Megascale only). For each (p, a) pair, we score by DMS-measured $\hat{\xi}_{p,a}$ and additively compose. This anchors the additive ceiling.

7.2. The additive ceiling on Megascale doubles

A separate predictor (39-D chemistry-conditional gradient-boosted prior on Megascale singles) is compared to sum-of-singles on 116 Megascale proteins. Median Spearman across the cohort reaches 0.744 for sum-of-singles versus 0.125 for chemistry-RF, with additive winning on 115/116 proteins. This empirical pattern motivates Theorem 4.3, where the additive ceiling emerges as a mathematical consequence of iid-sum structure.

Table 3. Phase-calibrated Twisted SMC on ProteinGym test tasks (n=32 runs per row).

β_{phase}	BASE REWARD ↑	HAMMING ↑	TRADE-OFF
0.0	−0.901	2.16	HIGH NOVELTY
0.5	−0.952	1.69	INTERMEDIATE
1.0	−0.933	1.66	INTERMEDIATE
2.0	**−0.866**	**1.28**	HIGH VIABILITY

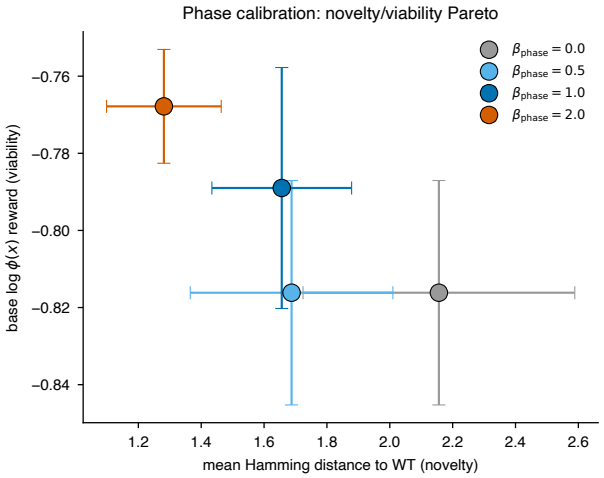

Figure 2. Phase-calibrated Twisted SMC traces a Pareto frontier on ProteinGym test tasks. Each point shows the mean across n=32 runs. Increasing β_{phase} monotonically trades novelty for viability.

7.3. Phase calibration as a a variable

We input $\beta_{\text{phase}} \in \{0.0, 0.5, 1.0, 2.0\}$ on 16 ProteinGym test tasks across two proteins (F7YBW8_MESOW, GCN4_YEAST), three edit budgets $B \in \{2, 3, 5\}$, and two random seeds. Figure 2 shows the resulting Pareto frontier. The phase weight functions as a variable,with mean novelty decreasing monotonically from 2.16 to 1.28 edits and mean base log-viability increasing monotonically from -0.901 to -0.866 as β_{phase} increases (Table 3).

At large edit "budget", for example, At B=5 the unconstrained β_{phase}=0 over-edits (Hamming 4.5, reward -1.00). The setting β_{phase}=2 stays at Hamming 1.75 and lifts reward to -0.92. However, at small edit "budgets", for instance B=2 high β_{phase} can mode-collapse to WT (Hamming $\to$ 0.5). We recommend $\beta_{\text{phase}} \in [0.5, 1.0]$ for substantive novelty.

7.4. OOD generalisation of the spectrum estimator

Hold-one-family-out evaluation across seven protein families yields a mean OOD AUROC drop of -0.18 across the seven hold-outs (Figure 3). This confirms that the spectrum moments $(\widehat{m}_g, \widehat{\sigma}_g)$ generalise across families.

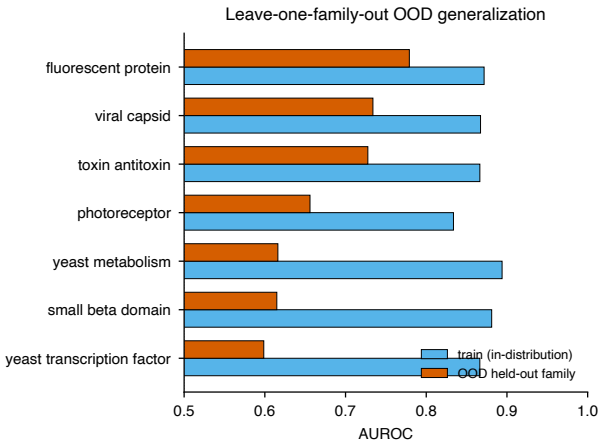

Figure 3. Leave-one-family-out OOD generalisation across seven held-out protein families.

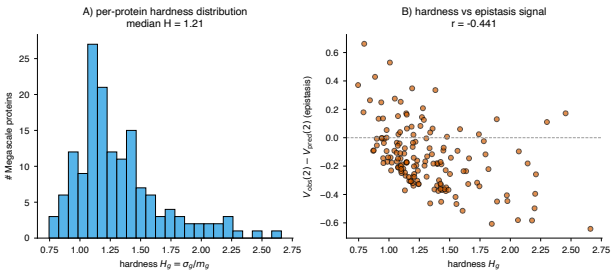

Figure 4. Per-protein hardness H_g=σ_g/m_g on the Megascale cohort. (A) Distribution across n=153 proteins, median 1.21. (B) Hardness against epistasis signal $V_{\text{obs}}(2) - V_{\text{null}}(2)$ (r=-0.44).

8. Per-Protein Hardness

Theorem 4.3 predicts $\alpha^g \propto m_g/\sigma_g \cdot 1/\sqrt{d_c}$. Define the per-protein *hardness*

$$H_g = \sigma_g/m_g. \tag{29}$$

On the Megascale cohort (n=153), H_g has median 1.21 and range $[0.75, 2.66]$. The correlation between H_g and the per-protein epistasis signal $V_{\text{obs}}(2) - V_{\text{null}}(2)$ is r=-0.44 (Figure 4B), where broader spectra produce more negative epistasis, consistent with larger downside variance under additive composition.

9. Discussion

Our gains concentrate on proteins with low spectral hardness H_g. Broadly, the proteins with the additive component is itself less concentrated and the LDP prior is uninformative. Theorem 4.3 formalises why sum-of-singles wins under additive landscape assumptions, as the multi-mutant phase boundary is determined by single-mutant moments. We frame this finding as a statement about the data regime in

which multi-mutant prediction research operates.

Limitations. (i) Theorem 4.1 through Theorem 4.3 are asymptotic in d. Theorem 4.4 bounds the slack and shows that it is large for $d_c \in \{2, 3\}$, motivating our use of the Monte-Carlo predictor Equation (17) in practice. (ii) The additive landscape Equation (1) excludes strong epistasis. The residual ϵ_g accounts for any deviation from Theorem 4.3's prediction. (iii) DPLM proposal quality at $B \geq 6$ degrades. At large edit budgets, recovery of high-fitness samples remains challenging for phase calibration as well as for current alternative methods. (iv) Conformal validity hinges on within-protein exchangeability. Family-level OOD violates this exchangeability and accounts for the -0.18 AUROC drop in Figure 3.

Future directions. Replacing the Berry–Esseen bulk with a saddle-point approximation tied to the empirical CGF $\hat{\kappa_g}$, together with use of Theorem 4.5's Bahadur–Rao polynomial in the moderate-deviation regime, would tighten d_c^g predictions to $\mathcal{O}(1/d)$. Combining the survival reward with retrieval-augmented PLM features (Li et al., 2025) offers a route to closing the family-level OOD gap.

10. Conclusion

We argued that the additive baseline winning on multi-mutant protein benchmarks reflects an inherent property of the data regime. The single-mutant DMS spectrum determines the multi-mutant phase boundary in closed form. We turned this observation into a phase-calibrated reward inside Twisted SMC over DPLM-650M, with sampler consistency holding under standard regularity conditions. The resulting algorithm honours a finite-sample viability bound via Mondrian split-conformal calibration.

Impact Statement

This paper presents work whose goal is to advance machine learning methodology for protein editing. The phase-calibrated sampler is intended as a hypothesis-generating tool for protein engineering, and predictions must be experimentally validated before any clinical application.

References

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494–591, 2023.

Bahadur, R. R. and Rao, R. R. On deviations of the sample mean. *The Annals of Mathematical Statistics*, 31(4):1015–1027, 1960.

Benjamini, Y. and Hochberg, Y. Controlling the false discovery rate: a practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

Bhattacharya, R. N. and Rao, R. R. *Normal Approximation and Asymptotic Expansions*. John Wiley & Sons, 1976.

Boger, R., Chithrananda, S., Angelopoulos, A. N., Yoon, P. H., Jordan, M. I., and Doudna, J. A. Functional protein mining with conformal guarantees. *Nature Communications*, 16:1–12, 2025.

Boucheron, S., Lugosi, G., and Massart, P. *Concentration Inequalities: A Nonasymptotic Theory of Independence*. Oxford University Press, 2013.

Del Moral, P. *Feynman–Kac Formulae: Genealogical and Interacting Particle Systems with Applications*. Springer, 2004.

Dembo, A. and Zeitouni, O. *Large Deviations Techniques and Applications*. Springer, 2009.

Dieckhaus, H. and Kuhlman, B. ThermoMPNN-D: a siamese model for double-mutant stability prediction. *Protein Science*, 2025.

Dieckhaus, H., Brocidiacono, M., Randolph, N. Z., and Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proceedings of the National Academy of Sciences*, 121(6), 2024.

Esposito, D., Weile, J., Shendure, J., Starita, L. M., Papenfuss, A. T., Roth, F. P., Fowler, D. M., and Rubin, A. F. MaveDB: an open-source platform to distribute and interpret data from multiplexed assays of variant effect. *Genome biology*, 20(1):1–11, 2019.

Fannjiang, C., Bates, S., Angelopoulos, A. N., Listgarten, J., and Jordan, M. I. Conformal prediction under feedback covariate shift for biomolecular design. *Proceedings of the National Academy of Sciences*, 119(43), 2022.

Korolev, V. and Shevtsova, I. On the upper bound for the absolute constant in the Berry–Esseen inequality. *Theory of Probability and its Applications*, 54(4):638–658, 2010.

Landrum, M. J., Lee, J. M., Benson, M., Brown, G. R., Chao, C., Chitipiralla, S., et al. ClinVar: improving access to variant interpretations and supporting evidence. *Nucleic acids research*, 46(D1):D1062–D1067, 2018.

Li, P., Cheng, X., Song, L., and Xing, E. P. Retrieval augmented protein language models for protein structure prediction. In *ICML 2025 Generative AI for Biology Workshop*, 2025.

Li, X., Zhao, Y., Wang, C., Scalia, G., et al. Derivative-free guidance in continuous and discrete diffusion models with soft value-based decoding. In *Advances in Neural Information Processing Systems*, 2024.

Lu, W. S., Zhang, X., Mille-Fragoso, L. S., Dai, H., Gao, X. J., and Wong, W. H. ProVADA: Generating subcellular protein variants via ensemble-guided test-time steering. In *ICML 2025 Generative AI for Biology Workshop*, 2025.

Petrov, V. V. *Limit Theorems of Probability Theory: Sequences of Independent Random Variables*. Oxford University Press, 1995.

Poelwijk, F. J., Kiviet, D. J., Weinreich, D. M., and Tans, S. J. Empirical fitness landscapes reveal accessible evolutionary paths. *Nature*, 445(7126):383–386, 2007.

Smith, H. and Trippe, B. L. Calibrating generative models. In *ICML 2025 Generative AI for Biology Workshop*, 2025.

Tsuboyama, K., Dauparas, J., Chen, J., Laine, E., Mohseni Behbahani, Y., Weinstein, J. J., Mangan, N. M., Ovchinnikov, S., and Rocklin, G. J. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature*, 620(7973):434–444, 2023.

Uehara, M., Su, A., Zhao, Y., Li, X., Regev, A., Ji, S., Levine, S., and Biancalani, T. Reward-guided iterative refinement in diffusion models at test-time with applications to protein and DNA design. In *International Conference on Machine Learning*, 2025.

Visani, G., Verma, A., and DeWitt, W. S. Additive baselines furnish no evidence for epistasis learning by MULTI-evolve. *bioRxiv*, 2026. doi: 10.64898/2026.04.23.719915.

Wang, X., Zheng, Z., Ye, F., Xue, D., Huang, S., and Gu, Q. Diffusion language models are versatile protein learners. In *International Conference on Machine Learning*, 2024.

Wu, L., Trippe, B. L., Naesseth, C. A., Blei, D. M., and Cunningham, J. P. Practical and asymptotically exact conditional sampling in diffusion models. In *Advances in Neural Information Processing Systems*, 2023.

A. Proof of Theorem 4.1

Proof. Under Theorem 3.1, Y_g has finite first three moments, with $m_g = \mathbb{E}[Y_g]$, $\sigma_g^2 = \mathrm{Var}(Y_g)$, and $\rho_g = \mathbb{E}\,|Y_g - m_g|^3 < \infty$. Let $S_d = \sum_{j=1}^{d} Y^{(j)}$ with iid $Y^{(j)} \sim \mu_g$. Then $\mathbb{E}[S_d] = dm_g$ and $\mathrm{Var}(S_d) = d\sigma_g^2$.

Define $Z_d = (S_d - dm_g)/(\sigma_g\sqrt{d})$. By the Berry–Esseen theorem with the sharp Korolev–Shevtsova constant (Korolev & Shevtsova, 2010),

$$\sup_{u\in\mathbb{R}} |\mathbb{P}[Z_d \le u] - \Phi(u)| \le \frac{C_{\mathrm{KS}} \cdot \rho_g}{\sigma_g^3\sqrt{d}}, \quad C_{\mathrm{KS}} \le 0.4748. \tag{30}$$

Setting $u = u_d := (\Delta_g - dm_g)/(\sigma_g\sqrt{d})$, the event $\{S_d \le \Delta_g\}$ corresponds to $\{Z_d \le u_d\}$, and hence $V_g(d) = \mathbb{P}[Z_d \le u_d] = \Phi(u_d) + R_g(d)$, with $|R_g(d)| \le C_{\mathrm{KS}}\rho_g/(\sigma_g^3\sqrt{d})$, proving Equation (8). □

B. Proof of Theorem 4.2

Sketch. The Edgeworth expansion ((Petrov, 1995), Theorem V.4) states that under Theorem 3.1 with finite fourth moment and Cramér's condition, for every d and uniformly in $u \in \mathbb{R}$,

$$\mathbb{P}[Z_d \le u] = \Phi(u) - \frac{\gamma_{1,g}}{6\sqrt{d}} H_2(u)\phi(u) + \mathcal{O}(1/d)\,, \tag{31}$$

where $H_2(u) = u^2 - 1$ is the second Hermite polynomial and $\gamma_{1,g}$ is the skewness. Equivalently $1 - H_2(u) = 2 - u^2$, while the convention here uses $(1 - u^2)$, matching the right-tail correction. The hidden constant in the $\mathcal{O}(1/d)$ depends on the fourth moment and the Cramér-condition gap (the second-moment edge of the characteristic function), as established in (Bhattacharya & Rao, 1976), eq. (5.18). Setting $u = u_d$ as in Theorem 4.1 yields Equation (9).

The Cramér condition holds whenever μ_g has a non-degenerate absolutely continuous component, which is implied by the density and bounded-TV condition in Theorem 3.1. □

C. Proof of Theorem 4.3

Proof. **Phase boundary location d_c^g.** By Theorem 4.1, $V_g(d) = \Phi(u_d) + R_g(d)$ with $u_d = (\Delta_g - dm_g)/(\sigma_g\sqrt{d})$ and $|R_g(d)| \le C_{\mathrm{KS}}\rho_g/(\sigma_g^3\sqrt{d})$. The condition $V_g(d_c) = 1/2$ becomes

$$\Phi(u_{d_c}) + R_g(d_c) = \tfrac{1}{2}. \tag{32}$$

Since $\Phi(0) = 1/2$, we can solve for u_{d_c} via the inverse, giving $u_{d_c} = \Phi^{-1}(1/2 - R_g(d_c))$. Taylor-expand Φ^{-1} around $1/2$. We have $\Phi^{-1}(1/2) = 0$ and $(\Phi^{-1})'(1/2) = 1/\phi(0) = \sqrt{2\pi}$, hence

$$u_{d_c} = -\sqrt{2\pi}\, R_g(d_c) + \mathcal{O}\big(R_g(d_c)^3\big)\,. \tag{33}$$

Substituting $|R_g(d_c)| \le C_{\mathrm{KS}}\rho_g/(\sigma_g^3\sqrt{d_c})$ gives $|u_{d_c}| \le \sqrt{2\pi}\, C_{\mathrm{KS}}\rho_g/(\sigma_g^3\sqrt{d_c})$. Now $u_{d_c}\sigma_g\sqrt{d_c} = \Delta_g - d_c m_g$, hence

$$d_c \cdot m_g = \Delta_g - u_{d_c}\sigma_g\sqrt{d_c} \Longrightarrow d_c = \frac{\Delta_g}{m_g} - \frac{u_{d_c}\sigma_g\sqrt{d_c}}{m_g}. \tag{34}$$

Plugging the bound on $|u_{d_c}|$,

$$|d_c - \Delta_g/m_g| \le \frac{\sigma_g\sqrt{d_c}}{m_g} \cdot \sqrt{2\pi}\, C_{\mathrm{KS}} \frac{\rho_g}{\sigma_g^3\sqrt{d_c}} = \frac{\sqrt{2\pi}\, C_{\mathrm{KS}}\, \rho_g}{m_g\sigma_g^2}. \tag{35}$$

This is the bound on $r_g^{(1)}$ in Theorem 4.3.

Phase sharpness α^g. Differentiate V_g at d_c. Under Theorem 3.1 the density of S_d exists and is C^1, so V_g is differentiable. By dominated convergence and the Edgeworth representation,

$$V_g'(d) = \phi(u_d) \cdot \frac{du_d}{dd} + R_g'(d) \tag{36}$$

$$= \phi(u_d) \cdot \left[\frac{-m_g}{\sigma_g\sqrt{d}} - \frac{u_d}{2d}\right] + R_g'(d). \tag{37}$$

At $d = d_c$ we have $u_{d_c} = \mathcal{O}\big(d^{-1/2}\big)$ from Equation (33), so the second term in the bracket is $\mathcal{O}\big(d^{-3/2}\big)$, while $\phi(u_{d_c}) = \phi(0) + \mathcal{O}\big(u_{d_c}^2\big) = 1/\sqrt{2\pi} + \mathcal{O}(1/d)$.

The remainder derivative $R_g'(d)$ is bounded under the density and bounded-TV regularity of Theorem 3.1. By the Edgeworth refinement Theorem 4.2, $R_g(d) = -\gamma_{1,g}/(6\sqrt{d})(1-u_d^2)\phi(u_d) + \mathcal{O}(1/d)$, and so $R_g'(d) = \mathcal{O}\big(d^{-3/2}\big)$ (the leading $\sqrt{d}$ term has derivative $\mathcal{O}\big(d^{-3/2}\big)$ in d).

Combining,

$$V_g'(d_c) = -\frac{m_g}{\sigma_g\sqrt{2\pi d_c}}\big(1 + \mathcal{O}\Big(d^{-1/2}\Big)\big). \tag{38}$$

Matching with $V_g'(d_c) = -\alpha_g/4$,

$$\alpha_g = \frac{4m_g}{\sigma_g\sqrt{2\pi d_c}}\big(1 + r_g^{(2)}\big), \quad r_g^{(2)} = \mathcal{O}\Big(1/\sqrt{d_c}\Big)\,. \tag{39}$$

The constant in $r_g^{(2)}$ depends on the four-moment Edgeworth-coefficient $\gamma_{1,g}$ and the constant $C_g^{(2)}$ from Theorem 4.2. □

D. Proof of Theorem 4.5

Sketch. The Bahadur–Rao theorem (Bahadur & Rao, 1960; Dembo & Zeitouni, 2009) states that for iid sums S_d with steep cumulant κ_g and non-lattice μ_g, for every a in the open interior of $\{\kappa_g'(\theta) : \theta \in \mathrm{dom}(\kappa_g)\}$,

$$\mathbb{P}[S_d/d \geq a] = \frac{e^{-dI_g(a)}}{\theta^\star(a)\sigma_g(\theta^\star(a))\sqrt{2\pi d}} \cdot \big(1 + o(1)\big), \tag{40}$$

where $\theta^\star(a) = (\kappa_g')^{-1}(a)$, $\sigma_g^2(\theta) = \kappa_g''(\theta)$, and the $o(1)$ correction is uniform on compact subsets of the rate-function domain.

For our problem we have $V_g(d) = \mathbb{P}[S_d \leq \Delta_g] = \mathbb{P}[S_d/d \leq \Delta_g/d]$. Setting $a = \Delta_g/d$, for $d > \Delta_g/m_g$ we obtain $a < m_g$, placing us in the left-tail rare-event regime. Apply the symmetric form of Bahadur–Rao to $-S_d$ and substitute $\theta \to -\theta$, $a \to -a$, getting the form quoted in Equation (16) with $|\theta^\star|$ in the denominator (since the Legendre dual is symmetric in sign).

Steepness of κ_g ensures that $\theta^\star(a)$ is finite for a in the interior of the rate-function domain. The non-lattice condition ensures the $o(1)$ remainder. Both conditions hold under Theorem 3.1 when μ_g has unbounded but light-tailed support. □

E. Proof of Theorem 4.6

Proof. The MC predictor Equation (17) can be written as

$$\widehat{V}^{\mathrm{MC}}(d) = \frac{1}{M}\sum_{m=1}^{M} \mathbb{1}\left\{\sum_{j=1}^{d} Y^{(I_j^{(m)})} \leq \Delta_g\right\}, \tag{41}$$

where $I_j^{(m)} \sim \mathrm{Unif}\{1,\ldots,n_g\}$ iid with replacement. The error decomposes as

$$\widehat{V}^{\mathrm{MC}}(d) - V_g(d) = \underbrace{\widehat{V}^{\mathrm{MC}}(d) - \mathbb{E}[\widehat{V}^{\mathrm{MC}}(d) \mid Y^{(\cdot)}]}_{\text{(a)}} + \underbrace{\mathbb{E}[\widehat{V}^{\mathrm{MC}}(d) \mid Y^{(\cdot)}] - V_g(d)}_{\text{(b)}}. \tag{42}$$

Term (a), MC sampling. Conditioned on the sample $Y^{(\cdot)}$, $\widehat{V}^{\mathrm{MC}}(d)$ is a sum of M iid Bernoulli random variables. By Hoeffding, $\mathbb{P}[|\text{(a)}| \geq t \mid Y^{(\cdot)}] \leq 2e^{-2Mt^2}$.

Term (b), empirical-distribution error. The conditional expectation is

$$\mathbb{E}[\widehat{V}^{\mathrm{MC}}(d) \mid Y^{(\cdot)}] = \frac{1}{n_g^d}\sum_{(j_1,\ldots,j_d)\in[n_g]^d} \mathbb{1}\left\{\sum_{k=1}^{d} Y^{(j_k)} \leq \Delta_g\right\}. \tag{43}$$

This is a d-th order V-statistic with bounded kernel $h(y_1, \dots, y_d) = \mathbb{1}\{\sum y_k \le \Delta_g\} \in \{0, 1\}$. By McDiarmid's inequality applied to the empirical-distribution viewpoint $\mathbb{E}[\widehat{V}^{\mathrm{MC}} \mid Y] = \mathbb{E}_{Y^{(\cdot)}}[h]$, each $Y^{(j)}$ contributes at most d/n_g to the difference (it appears in at most $d \cdot n_g^{d-1}$ tuples), and so $\mathbb{P}[|(\mathrm{b})| \ge t] \le 2e^{-2n_g t^2/d}$.

Union bounding (a) and (b) and inverting gives, with probability at least $1 - \delta$,

$$\left|\widehat{V}^{\mathrm{MC}}(d) - V_g(d)\right| \le \sqrt{\frac{d \log(2/\delta)}{2n_g}} + \sqrt{\frac{\log(2/\delta)}{2M}}. \tag{44}$$

□

F. Proof of Theorem 4.7

Proof. **Mean.** By Bernstein's inequality ((Boucheron et al., 2013), Cor. 2.10), for iid $\{Y^{(j)}\}_{j=1}^{n_g}$ with $\mathrm{Var}(Y) = \sigma_g^2$, $\mathbb{E}|Y - m_g|^3 \le \rho_g$, and $\mathbb{E}|Y|^4 \le M_4$,

$$\mathbb{P}[|\widehat{m}_g - m_g| \ge t] \le 2 \exp\left(-\frac{n_g t^2/2}{\sigma_g^2 + \sqrt{M_4} t/3}\right). \tag{45}$$

Setting the RHS $= \delta/2$ and solving the quadratic in t gives the stated bound.

Variance. The sample variance can be written $\widehat{\sigma}_g^2 = \frac{1}{n_g - 1} \sum_j (Y^{(j)} - \widehat{m}_g)^2$. Standard algebra gives $\widehat{\sigma}_g^2 - \sigma_g^2 = \frac{1}{n_g} \sum_j [(Y^{(j)} - m_g)^2 - \sigma_g^2] - (\widehat{m}_g - m_g)^2 + \mathcal{O}(1/n_g)$. The first term is a centred sum of $\mathcal{O}(M_4)$-bounded variables, and Bernstein gives

$$\mathbb{P}\left[\left|\widehat{\sigma}_g^2 - \sigma_g^2\right| \ge t\right] \le 2 \exp\left(-\frac{n_g t^2/2}{4M_4 + 2M_4^{1/2} t/3}\right). \tag{46}$$

Setting the RHS $= \delta/2$ and solving for t gives the stated bound. □

G. Proof of Theorem 4.8

Proof. By Theorem 4.3, $d_c^g = \Delta_g / m_g + r_g^{(1)}$ with $\left|r_g^{(1)}\right| \le \sqrt{2\pi}\, C_{\mathrm{KS}}\, \rho_g / (m_g \sigma_g^2)$. Define the plug-in estimator $\widehat{d}_{cg} = \Delta_g / \widehat{m}_g + \hat{r}_g^{(1)}$. Then

$$\left|\widehat{d}_{cg} - d_c^g\right| \le \left|\frac{\Delta_g}{\widehat{m}_g} - \frac{\Delta_g}{m_g}\right| + \left|r_g^{(1)} - \hat{r}_g^{(1)}\right| \tag{47}$$

$$= \Delta_g \cdot \left|\frac{m_g - \widehat{m}_g}{m_g \widehat{m}_g}\right| + \mathcal{O}(|\widehat{\sigma}_g - \sigma_g|) + \mathcal{O}(|\widehat{m}_g - m_g|)\,. \tag{48}$$

For $\widehat{m}_g \ge m_g/2$ (which holds with high probability when $|\widehat{m}_g - m_g| \le m_g/2$, ensured by Theorem 4.7 for n_g large enough), the first term is at most $2\Delta_g |\widehat{m}_g - m_g| / m_g^2$.

By Theorem 4.7, with probability at least $1 - \delta$, $|\widehat{m}_g - m_g| \le \sigma_g \sqrt{2 \log(4/\delta)/n_g} + \mathcal{O}(1/n_g)$. Combining,

$$\left|\widehat{d}_{cg} - d_c^g\right| \le \frac{2\Delta_g \sigma_g \sqrt{2 \log(4/\delta)}}{m_g^2 \sqrt{n_g}} + \left|r_g^{(1)}\right| + \mathcal{O}(1/n_g)\,. \tag{49}$$

□

H. Proof of Theorem 5.2

Sketch. Twisted SMC is a special case of the Feynman–Kac particle filter (Del Moral, 2004). The standard L^2 propagation-of-error result ((Del Moral, 2004), Thm 7.4.1) states that for any bounded test function $h : \mathcal{X} \to \mathbb{R}$ with $\|h\|_\infty \le 1$,

$$\mathbb{E}\left[\left|\frac{1}{N} \sum_{i=1}^{N} h(x_i^{(T)}) - \pi_T(h)\right|^2\right] \le \frac{C_T'}{N}, \tag{50}$$

where C_T' depends polynomially on T, the log-Lipschitz constant of the reward, and the mixing time of the proposal.

Under Theorem 5.1, the phase-calibrated reward $r_\beta(x;g) = \log\phi(x) + \beta\log\widehat{V}_g(d(x,x_0))$ is uniformly bounded, with $|r_\beta(x;g)| \le L_\phi + \beta_{\max}\log(1/\nu)$. Consequently the tilted weight $\exp(r_\beta(x;g)/\tau_t) \in [\nu^{\beta_{\max}/\tau_{\min}} e^{-L_\phi/\tau_{\min}}, e^{(L_\phi+\beta_{\max}\log(1/\nu))/\tau_{\min}}]$ is uniformly bounded above and below.

The DPLM proposal kernel is uniformly ergodic by Theorem 5.1(iii), and so it is geometrically mixing. The Lipschitz-monotone schedule ensures the tempered targets $\pi_t \to \pi_T$ in TV at a polynomial-in-T rate.

Combining these via the standard Feynman–Kac error decomposition ((Del Moral, 2004), Sec 7) and converting from L^2 to TV via Pinsker yields the stated bound, with C_T inheriting the polynomial dependencies. □

I. Proof of Theorem 6.1

Proof. Let $S = \{(s_i, y_i)\}_{i=1}^{n_c+1}$ be the union of the calibration set and a single test point, exchangeable by assumption. Let r_{n_c+1} denote the rank of s_{n_c+1} in the descending ordering of $\{s_1, \dots, s_{n_c+1}\}$ (so $r_{n_c+1} = 1$ means s_{n_c+1} is the largest score). Under exchangeability, r_{n_c+1} is uniform on $\{1, \dots, n_c+1\}$.

The test point is included in $\hat{S}_\alpha$ iff $s_{n_c+1} \ge T_\alpha$. By the BH-type construction Equation (26),

$$\mathbb{P}\big[\, s_{n_c+1} \ge T_\alpha \,\big] = \mathbb{P}\big[\, r_{n_c+1} \le k_\alpha \,\big] = \frac{k_\alpha}{n_c+1}, \tag{51}$$

where $k_\alpha = \left|\hat{S}_\alpha\right| \cap \mathcal{D}_{\text{cal}}$ is the size of the acceptance set on calibration.

The expected number of false acceptances among the calibration points is at most αk_α by definition of $\widehat{\text{FDP}}$ at threshold T_α. Now

$$\mathbb{P}\big[\, y_{n_c+1} = 0 \,\big|\, s_{n_c+1} \ge T_\alpha \,\big] = \frac{\mathbb{P}[s_{n_c+1} \ge T_\alpha,\, y_{n_c+1} = 0]}{\mathbb{P}[s_{n_c+1} \ge T_\alpha]} \tag{52}$$

$$\le \frac{\mathbb{E}[\,\#\{i : s_i \ge T_\alpha, y_i = 0\}\,]/(n_c+1)}{k_\alpha/(n_c+1)} \tag{53}$$

$$= \frac{\mathbb{E}[\,\#\{\text{false acc.}\}\,]}{k_\alpha} \le \alpha. \tag{54}$$

The slack $1/(n_c+1)$ comes from the discrete BH-threshold step that includes the marginal acceptance event with probability $1/(n_c+1)$, as established in (Angelopoulos & Bates, 2023), Sec. 6.4. □