Physical Education, Health and Social Sciences

https://journal-of-social-education.org

E-ISSN: 2958-5996

P-ISSN: 2958-5988

AI-Driven Protein Engineering and Molecular Dynamics-Based Optimization of Enzyme Stability and Catalytic Activity

Sudeer Khan[1], Aiman khan[2]*

[1] Department of Bio-Chemistry, University of Malakand, Email: sudeeruom@gmail.com
[2] Corresponding Author, Email: aimanmuhammad930@gmail.com

DOI: https://doi.org/10.63163/jpehss.v4i1.1778

Abstract

Engineering enzymes that are both stable and highly active is limited by a trade-off in which stabilizing mutations often depress catalysis. We hypothesize that this trade-off arises from how mutations redistribute the conformational *ensemble*, not from static energetics alone. Current AI-guided pipelines optimize static proxies (predicted ΔΔG, language-model likelihood) and use molecular dynamics (MD) only for post hoc inspection. We present MF-DynOpt, a closed-loop framework that (i) generates variants with protein language models and inverse folding; (ii) routes them through a four-level fidelity cascade (language-model scoring, learned ΔΔG, adaptive-replica MD, alchemical free energy) whose escalation is governed by weighted conformal prediction; (iii) optimizes a new, theory-grounded objective, Productive Occupancy (PO), the probability that the enzyme–substrate complex populates a catalytically competent near-attack ensemble, jointly with stability using cost-aware multi-objective Bayesian optimization; and (iv) uses an Allosteric Mutation Prior (AMP) derived from dynamical networks to favor distal substitutions. We evaluate the framework in three tiers: retrospective validation on public data, benchmark replay against seven baselines at equal compute, and prospective experimental validation on [enzyme]. PO versus experimental kcat/KM (Spearman ρ = __, versus __ for ΔΔG); hypervolume gain of __% over the strongest baseline at equal compute; dual-success rate of __/__ variants (ΔTm ≥ +5 °C and kcat/KM ≥ 0.8 × parent) versus __/__ for baselines; GPU-hours per dual-success reduced by __%.]. The framework tests the claim that distal, dynamically decoupled mutations can raise global stability without depleting productive occupancy, and supplies falsifiable predictions (temperature dependence of kcat, HDX-MS) for that claim. Limitations include reliance on classical force fields and finite sampling of slow gating motions.

Keywords: Enzyme Engineering; Protein Language Models; Molecular Dynamics; Multi-Fidelity Bayesian Optimization; Productive Occupancy; Stability–Activity Trade-Off; Conformal Prediction

Introduction

Enzymes underpin biocatalysis, therapeutics and plastic depolymerization, yet natural enzymes rarely satisfy industrial requirements for thermal, solvent and operational stability. Directed evolution has delivered extraordinary successes [1] but explores sequence space locally and demands extensive screening. Machine-learning-guided directed evolution reduces this burden [2,3], and deep learning has transformed structure prediction [4], representation learning [5], sequence design [6,7] and stability prediction trained on mega-scale data [8,9]. Active-learning

methods that couple protein language models with small experimental batches now enable rapid in silico directed evolution [32].

Despite this progress, three problems remain. First, the stability–activity trade-off [10,11] is a property of the conformational ensemble: catalysis requires that the active site frequently samples near-attack conformations [12,13], and experimental studies show that evolution and directed evolution reshape enzyme dynamics to rescue or enhance catalysis [35,47–49]. Static objectives cannot see this. Second, MD, the natural tool for ensemble properties [43], is too expensive to apply to the vast candidate sets that generative models produce, and is therefore confined to validating a handful of finalists. Third, surrogate models are miscalibrated under distribution shift, so pipelines cannot decide rationally when a cheap prediction suffices and when simulation is warranted.

Here we address these problems with three contributions, each framed as a testable research question (RQ):

- **RQ1 (objective).** Does a theory-grounded, MD-derived Productive Occupancy (PO) predict experimental catalytic efficiency better than static stability and language-model proxies?
- **RQ2 (efficiency).** Does embedding MD as a conformally governed fidelity tier in a multi-objective active-learning loop improve the stability–activity Pareto front at equal compute, relative to the strongest sequential and single-fidelity baselines?
- **RQ3 (mechanism).** Do distal, dynamically coupled mutations selected by AMP raise stability while preserving PO and catalytic efficiency, and are the predicted dynamical signatures confirmed experimentally?

We deliberately limit our claims. To the best of our knowledge, and within the literature surveyed up to [search date], prior work has (a) demonstrated the importance of dynamics in enzyme evolution [35,47–49], (b) used PLMs, Bayesian optimization or learned $\Delta\Delta G$ for engineering [2,3,8,17,32], and (c) applied multi-fidelity or conformal methods in other domains [24,26–28]. We are not aware of a study that combines them with an ensemble-level catalytic objective under an explicit compute budget. Whether this combination outperforms its parts is the empirical question the paper answers.

Theoretical Framework

From Near-Attack Conformations to Productive Occupancy

Consider an enzyme E, substrate S and the Michaelis complex ES. Let NAC denote the set of ES microstates whose reactive geometry lies within a reference region $\mathcal{C}$ defined by the transition-state analogue or a QM-cluster optimized structure [12,13]. In a conformational-selection picture, and when either chemistry or NAC formation is rate-limiting, the apparent specificity constant is approximately proportional to

$$k_{cat}/K_M \propto \kappa \cdot k_{NAC} \cdot P(ES) \cdot P(NAC \mid ES) \qquad (1)$$

where k_{NAC} is the intrinsic rate of the chemical step from a competent geometry, κ a transmission coefficient and P(ES) the equilibrium occupancy of the productively bound state at reference substrate concentration. We define the **Productive Occupancy**

$$PO(s, T) = P_s(ES; T) \cdot P_s(NAC \mid ES; T) \qquad (2a)$$

$$\Delta G_{PO}(s, T) = -RT \ln PO(s, T) \qquad (2b)$$

and, for a variant s relative to parent s_0, $\Delta\Delta G_{PO} = \Delta G_{PO}(s) - \Delta G_{PO}(s_0)$. Taking logarithms of Eq. (1) for variant versus parent gives the working relation

$$\Delta \log_{10}(k_{cat}/K_M) \approx -\Delta\Delta G_{PO}/(RT \ln 10) + \Delta \log_{10} k_{NAC} \qquad (3)$$

Assumptions and scope. Eq. (3) is an approximation, not an identity. It assumes (i) a single dominant catalytic geometry, (ii) a conformational-selection-dominated mechanism, (iii) k_{NAC} unchanged by mutations that are distant from the reactive centre, and (iv) adequate force-field

representation of NAC populations. Violations (for example, mutations that alter the chemical barrier at fixed geometry, or product-release-limited enzymes) are expected to appear as systematic residuals in Eq. (3), and we test for them explicitly (§4.1). For finalists we estimate $\Delta \log k_{NAC}$ with QM/MM or empirical valence bond calculations [37], so the two terms are separated.

The Stability–Activity Trade-Off as Ensemble Tension

Global stabilization typically narrows the folded-state ensemble. If the narrowing preferentially removes substates that are productive (for example open-loop or gate-open states), $\Delta\Delta G_{PO}$ becomes positive, and Eq. (3) predicts a loss of efficiency *even when the static active-site geometry is unchanged*. Conversely, if stabilization is achieved by mutations whose effect on the ensemble is dynamically decoupled from the reactive centre, $\Delta\Delta G_{PO} \approx 0$ and activity is preserved. The framework therefore replaces the vague claim 'rigidity hurts activity' with a quantitative one: rigidity hurts activity only to the extent that it reduces PO.

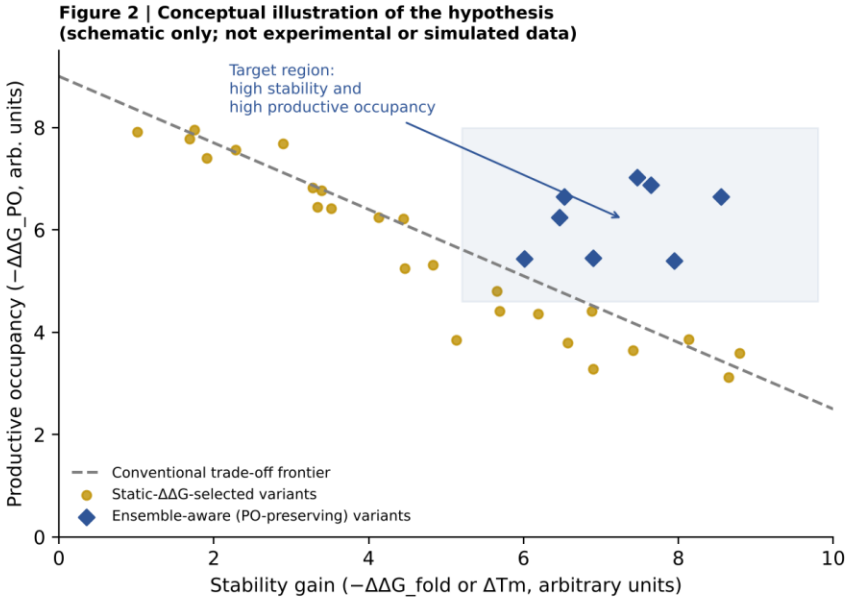


Figure 2. Conceptual illustration of the hypothesis. Static-ΔΔG-based selection is expected to track a trade-off frontier (dashed), whereas PO-preserving distal mutations should occupy the high-stability, high-PO region. Schematic only; not experimental or simulated data.

Temperature dependence

Because both stability and catalysis depend on temperature, a single-temperature objective can mislead. The Equilibrium Model [33] attributes the decline of activity above T_{opt} to a reversible, rapidly interconverting active–inactive equilibrium that precedes irreversible denaturation, while macromolecular rate theory [34] relates the curvature of ln k versus T to a non-zero heat-capacity change of activation, $\Delta C_p‡$, which reports on changes in the conformational ensemble between ground and transition states. We therefore compute PO(T) at the assay temperature and at $T + \Delta T$, and treat (i) T_{opt}, (ii) kinetic stability (half-life $t_{½}$ at a challenge temperature) and (iii) $\Delta C_p‡$ as experimental readouts.

Falsifiable corollaries

ID	Prediction	What would falsify it
C1	Across variants, Δlog(kcat/KM) correlates with $-\Delta\Delta G_{PO}$ with slope ≈ 1 (Eq. 3) for mutations remote from the reactive centre.	Slope indistinguishable from 0, or PO explains no more variance than ΔΔG or active-site RMSD.
C2	High-AMP distal mutations yield larger ΔTm per unit loss of PO than active-site-proximal mutations.	No difference, or AMP not better than distance-to-active-site or conservation.
C3	Variants with preserved PO show unchanged ΔCp‡; variants with reduced PO show altered ΔCp‡.	ΔCp‡ uncorrelated with ΔPO.
C4	Loop rigidification (negative ΔRMSF at gating loops) reduces activity only when PO falls.	Activity loss with unchanged PO, or PO loss without activity loss.
C5	Weighted conformal escalation lowers GPU-hours per dual-success versus MD-for-all without degrading the front.	No cost saving, or front degradation beyond a pre-set tolerance.

Methods

Problem Formulation

Let s be a variant of parent s_0 with at most k substitutions, excluding a protected set of catalytic and substrate-binding residues unless explicitly enabled. We seek the Pareto set of $f_1(s, T) = -\Delta\Delta G_{fold}(s, T)$ (stability; experimental surrogates ΔTm and t½) and $f_2(s, T) = -\Delta\Delta G_{PO}(s, T)$ (productive occupancy), subject to a compute budget B (GPU-hours on a stated device) and an experimental budget N_{exp} variants per round for R rounds. Each objective is observable at fidelity $\ell \in \{0,1,2,3\}$ with noise variance $\sigma^2_\ell(s)$ and cost c_ℓ.

System Preparation and Variant Modeling

The parent structure is a high-resolution holo (or transition-state-analogue-bound) crystal structure of [enzyme, PDB ID]. Protonation states are assigned at the assay pH with PROPKA/PDB2PQR and catalytic residues are verified manually. Variant structures are not generated by AlphaFold2, because co-folding networks are weakly sensitive to single substitutions. Instead, each variant is built by side-chain substitution on the parent holo structure, rotamer repacking within 8 Å of the site, restrained minimization, and subsequent MD equilibration; the same protocol is used for all variants so that errors are comparable. AlphaFold2/ESMFold models [4,5] are used only as quality control (flagging large predicted backbone departures) and for variants with ≥5 substitutions, where backbone changes are plausible.

Fidelity Hierarchy

Level	Method	Output	Indicative cost
L0	ESM-2 pseudo-log-likelihood [5]; ProteinMPNN conditional likelihood [6]	Zero-shot fitness prior	Seconds (GPU/CPU)
L1	Learned ΔΔG (ThermoMPNN / Stability Oracle type [8,31]); PO-Net surrogate (§3.5)	$\hat{f}_1$, $\hat{f}_2$ with ensemble uncertainty	Seconds–minutes

Level	Method	Output	Indicative cost
L2	Explicit-solvent MD, adaptive replicas (≥3, 50–200 ns each) at T and T+ΔT, substrate-bound and apo	PO, RMSF, H-bond persistence, dynamical network, native contacts Q	Tens of GPU-hours
L3	Alchemical FEP/TI with MBAR [22,23] for $\Delta\Delta G_{fold}$ and $\Delta\Delta G_{bind}$; QM/MM or EVB for k_{NAC} of finalists; REST2/MSM sampling of gating loops [21,45,46]	Reference-grade ΔΔG; $\Delta\log k_{NAC}$; slow-mode populations	Hundreds of GPU-hours

Estimation of PO and its Uncertainty

Reference region 𝒞 (frozen* a priori*). Reactive descriptors d (nucleophile–electrophile distance, attack angle, catalytic hydrogen-bond distances, oxyanion-hole contacts) and their thresholds τ are defined from a QM-cluster or transition-state-analogue geometry and from published NAC criteria [13], and are registered before any activity data are examined. They are not tuned to maximize correlation with kcat/KM. We report three estimators: (a) a hard indicator, (b) a soft logistic-gated estimator, and (c) a threshold-free estimator based on the Wasserstein distance between the variant's descriptor distribution and the reference NAC distribution.

Binding Term. P(ES) is estimated from the stability of the productive pose in substrate-bound MD (pose-retention fraction) and, for finalists, from alchemical $\Delta\Delta G_{bind}$, which prevents high NAC occupancy from masking weakened binding.

Sampling and Error. Production runs use OPC (AMBER ff19SB) or TIP3P/CHARMM36m, a 4 fs time step with hydrogen-mass repartitioning, and ≥3 independent replicas started from independently equilibrated frames. Equilibration is detected automatically [42]; statistical inefficiency, effective sample size and block-bootstrap confidence intervals are reported for every PO estimate; replicas are added adaptively until the bootstrap 95% CI half-width falls below a pre-set tolerance or a replica cap is reached. Because classical force fields can bias conformational populations [44], PO is recomputed with a second force field for all finalists, and a variant is retained only if both agree in rank within a tolerance. For finalists, enhanced sampling (REST2) and Markov-state models [45,46] resolve slow loop gating, and we report implied timescales. We do not claim to simulate unfolding: elevated-temperature MD provides only relative proxies (native-contact loss, RMSF), which are validated against experimental ΔTm in Tier A.

PO-Net Surrogate

PO-Net is an SE(3)-equivariant graph neural network that predicts PO and its predictive variance from the sequence and a single modeled structure. Training data come from L2 trajectories accumulated during the loop. To avoid leakage, evaluation uses **homology-aware cluster splits** (<30% sequence identity between train and test) and a held-out enzyme family. Calibration of the variance head is audited (§3.7).

Allosteric Mutation Prior (AMP)

From parent MD we compute the dynamical cross-correlation matrix, build a residue network weighted by $1 - |C_{ij}|$, and identify communities and shortest paths from catalytic residues. For each non-active-site position i,

$$AMP_i = \alpha \cdot coupling(i \rightarrow active\ site) + \beta \cdot (1 - rigidity\ cost_i) + \gamma \cdot tolerance_i \qquad (4)$$

where tolerance_i is the PLM entropy at i. Weights (α, β, γ) are fitted by **nested cross-validation** on a training enzyme set and then frozen before prospective use. AMP shapes the proposal distribution but never excludes sites. It is benchmarked against four site-selection alternatives: distance-to-active-site, evolutionary conservation, direct-coupling/statistical-coupling analysis, and random selection.

Closed-Loop Multi-Fidelity, Multi-Objective Optimization

Candidate Generation. PLM masked sampling, ProteinMPNN design with catalytic residues fixed [6], and AMP-weighted site-saturation.

Surrogate. An autoregressive multi-fidelity Gaussian-process model [24], $y_\ell = \rho_\ell\, y_{\ell-1} + \delta_\ell$, with heteroscedastic noise $\sigma^2_\ell(s)$ taken from the MD bootstrap estimates, jointly modeling (f_1, f_2). Correlation between fidelities is *checked*, not assumed: if the empirical rank correlation between adjacent levels falls below 0.4 on the calibration set, the cascade skips the offending level.

Calibrated Escalation. Prediction intervals use split conformal prediction [26] adapted to the active-learning setting by likelihood-ratio weighting for covariate shift [27] and conformalized quantile regression [28]. Joint coverage for the two objectives is controlled by Bonferroni correction. These guarantees are approximate, because weights are estimated; we therefore audit empirical coverage on a held-out calibration pool each round and report it (Fig. 5).

Acquisition. Cost-aware expected hypervolume improvement [25], $\alpha(s, \ell) = \text{EHVI}(s \mid \mathcal{D}) / c_\ell$. A variant is escalated from ℓ to ℓ+1 if its conformal box overlaps the current Pareto front and α(s, ℓ+1) exceeds the batch median; otherwise it is pruned. Batches of N_{exp} variants are chosen by determinantal-point-process sampling for diversity [39].

Epistasis. Multi-mutants are first scored additively; pairwise epistasis predicted by the surrogate is *not* trusted without evidence, and a subset of double mutants is measured to build a calibration curve of predicted versus observed epistasis [41] (§3.10).

Experimental Validation

Genes are codon-optimized, synthesized and expressed in *E. coli* [strain]; proteins are purified by IMAC and SEC and verified by intact-mass spectrometry. Tm is measured by nanoDSF; kinetics by [substrate-specific assay] at T and T+ΔT; half-life t½ at [challenge temperature]; secondary structure by CD. All measurements use ≥3 biological and 2 technical replicates; mean ± SD and 95% CIs are reported.

Evidence Tiers, Benchmarks, Baselines, Metrics and Statistics

Tier A – Retrospective Validation (public data, no wet lab). (i) Stability: held-out domains from the mega-scale dataset [9] and the S669 set [30] to validate ΔΔG predictors and the elevated-temperature MD proxy; (ii) Fitness: enzyme assays in ProteinGym [29]; (iii) PO versus kcat/KM on curated literature variant series for [enzyme] (≥[40] variants with measured kinetics) with thresholds frozen beforehand.

Tier B – Benchmark Replay. The optimization loop is run *in silico* against a lookup 'oracle' built from enzyme deep-mutational-scanning or variant datasets that provide orthogonal stability and activity readouts ([datasets to be selected]); the oracle is revealed only through the fidelity

interfaces, with simulated costs equal to measured GPU-hours. This tests RQ2 at scale (10–30 seeds) without new experiments.

Tier C – Prospective Experiments. R = [3] rounds of N_{exp} = [24] variants on [enzyme], plus two held-out enzymes from different families with frozen hyperparameters.

Baselines (all at equal compute, including training and every fidelity level).

ID	Baseline	Rationale
B1	Rosetta / FoldX stability ranking [14,15]	Classical physics-based reference
B2	ThermoMPNN / Stability Oracle top-k [8,31]	State-of-the-art learned ΔΔG
B3	ESM-1v zero-shot, greedy [16]	Language-model prior
B4	PLM active learning (EVOLVEpro-type) [32]	State-of-the-art low-N active learning
B5	Single-fidelity multi-objective BO (ΔΔG, PLM)	Isolates multi-fidelity and PO
B6	Sequential pipeline: maximize stability (B2), then filter by MD	The strongest 'obvious' alternative; isolates joint optimization
B7	MD-for-all (no cascade)	Upper bound on information, lower bound on efficiency
B8	Random substitutions with matched mutation counts	Null model

Negative Controls. (a) PO computed after shuffling descriptor-to-frame assignments; (b) catalytically dead variants; (c) a scrambled-AMP prior; (d) PO-Net trained on permuted labels. A claim is supported only if it exceeds all null distributions.

Metrics. Dominated hypervolume in (f_1, f_2); *dual-success rate* ($\Delta Tm \geq +5$ °C and $k_{cat}/K_M \geq 0.8 \times$ parent at the assay temperature); simple regret relative to the best known variant; GPU-hours per dual-success; empirical coverage of 90% intervals; Spearman and Kendall correlations for PO versus k_{cat}/K_M.
Ablations. Remove (a) PO, (b) AMP, (c) conformal escalation, (d) multi-fidelity cascade, (e) second-force-field agreement, (f) heteroscedastic noise propagation, (g) epistasis calibration.

Statistics. Optimization seeds are *not* treated as independent biological replicates: *in silico* comparisons use hierarchical bootstrap over seeds and enzyme systems, and linear mixed-effects models with system as a random effect; effect sizes are reported with 95% bootstrap CIs. Multiple comparisons are corrected by Holm's method. **Power analysis for Tier C:** to detect a dual-success rate of 40% versus 15% (two-sided $\alpha = 0.05$, power 0.80), approximately 49 variants per arm are required by the normal-approximation formula, so we plan ≥ 50 variants per method arm pooled across rounds; hit rates are reported with Wilson intervals [40].

Mechanistic Validation (pre-specified)
For ≥2 variants and the parent we collect: (i) $k_{cat}(T)$ across ≥8 temperatures to estimate $\Delta C_p‡$ by macromolecular rate theory [34]; (ii) HDX-MS or NMR relaxation to compare loop dynamics;

(iii) a crystal or cryo-EM structure where feasible; and (iv) a double-mutant-cycle subset to calibrate predicted epistasis.

Reproducibility and Pre-registration

Hypotheses C1–C5, success criteria, baselines, thresholds τ, sample sizes and analysis code are pre-registered at [OSF DOI] **before** Tier B/C data are generated. Software is released as a container with pinned versions (OpenMM [20] / GROMACS [19], PyTorch, BoTorch, MDAnalysis, deeptime), random seeds are logged, and reduced trajectories, input files and surrogate weights are deposited at [Zenodo DOI]. Compute is reported in GPU-hours on [GPU model] for every method.

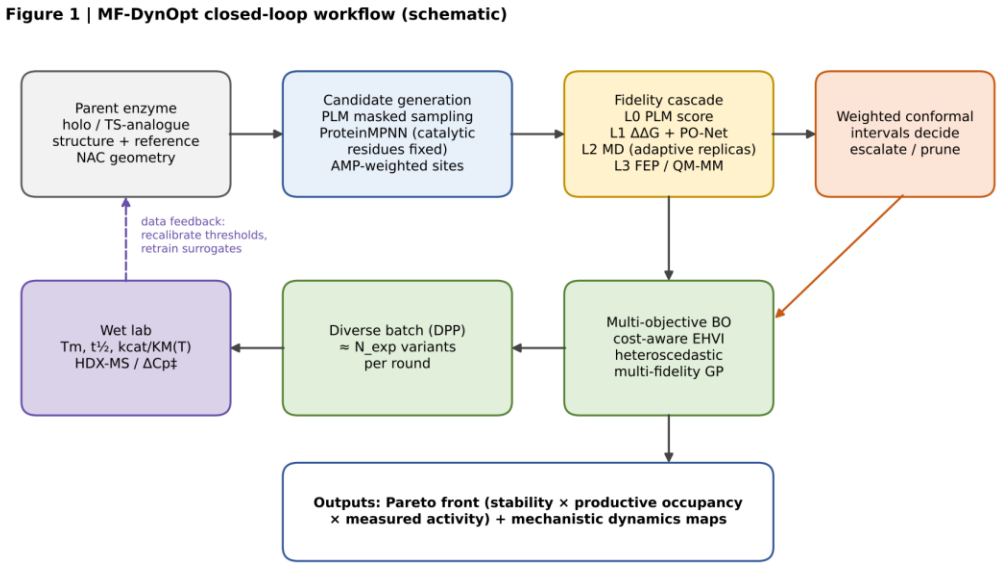


Figure 1. MF-DynOpt closed-loop workflow. Candidates from language-model and inverse-folding generators pass through a fidelity cascade governed by weighted conformal intervals; a cost-aware multi-objective optimizer selects diverse batches for experiments, whose results recalibrate thresholds and surrogates. Schematic.

Results

All values below are placeholders to be completed from your analyses. For each subsection we state the pre-specified criterion that determines whether the claim is supported, so that negative or partial outcomes are reported honestly.

PO Predicts Catalytic Efficiency Beyond Static Proxies (RQ1, C1)

Criterion: PO supported if Spearman ρ(PO, kcat/KM) exceeds that of every static proxy with a non-overlapping 95% bootstrap CI, **and** exceeds all negative controls.

Table 1. Association with experimental kcat/KM on [N] literature variants (thresholds frozen *a priori*).

Predictor	**Spearman ρ (95% CI)**	**Kendall τ**	**Residual slope vs Eq. (3)**
ΔΔG (FoldX)	[RUN]	[RUN]	n/a

Predictor	Spearman ρ (95% CI)	Kendall τ	Residual slope vs Eq. (3)
ΔΔG (ThermoMPNN / Stability Oracle)	[RUN]	[RUN]	n/a
ESM-1v score	[RUN]	[RUN]	n/a
Active-site RMSD	[RUN]	[RUN]	n/a
PO – hard / soft / threshold-free	[RUN]	[RUN]	[RUN]
Negative control: shuffled PO	[RUN]	[RUN]	n/a

Report stratification by distance from the reactive centre: Eq. (3) is predicted to hold with slope ≈ 1 for remote mutations (C1) and to show residuals for mutations that change k_{NAC}.

4.2 Validity of the MD proxy and of PO-Net

Report (i) correlation between elevated-temperature MD proxies and experimental ΔTm on Tier A stability data; (ii) PO-Net R^2, MAE and reliability diagrams on homology-split test sets; (iii) MD convergence diagnostics (effective sample size, bootstrap CI widths, cross-force-field rank agreement). **[RUN]**

Pareto Front at Equal Compute (RQ2, Tier B replay)

Criterion: supported if MF-DynOpt exceeds B6 (sequential pipeline) and B4 in hypervolume with a hierarchical-bootstrap CI excluding zero, after Holm correction.

Table 2. Tier B replay at equal compute (mean over seeds and systems, with 95% CI).

Method	Hypervolume	Dual-success rate	GPU-h per dual-success	Simple regret
B1 Rosetta/FoldX	[RUN]	[RUN]	[RUN]	[RUN]
B2 ThermoMPNN / Stability Oracle	[RUN]	[RUN]	[RUN]	[RUN]
B3 ESM-1v greedy	[RUN]	[RUN]	[RUN]	[RUN]
B4 PLM active learning	[RUN]	[RUN]	[RUN]	[RUN]
B5 Single-fidelity BO	[RUN]	[RUN]	[RUN]	[RUN]
B6 Sequential stability → MD	[RUN]	[RUN]	[RUN]	[RUN]
B7 MD-for-all	[RUN]	[RUN]	[RUN]	[RUN]
B8 Random matched	[RUN]	[RUN]	[RUN]	[RUN]
MF-DynOpt	[RUN]	[RUN]	[RUN]	[RUN]

Ablations and Calibration (C5)

Table 3. Ablations (change relative to full MF-DynOpt).

Component removed	Δ hypervolume	Δ cost (GPU-h)	Δ coverage of 90% boxes
PO objective	[RUN]	[RUN]	[RUN]
AMP prior	[RUN]	[RUN]	[RUN]

Component removed	Δ hypervolume	Δ cost (GPU-h)	Δ coverage of 90% boxes
Weighted conformal escalation	[RUN]	[RUN]	[RUN]
Multi-fidelity cascade (MD-for-all)	[RUN]	[RUN]	[RUN]
Second-force-field agreement	[RUN]	[RUN]	[RUN]
Heteroscedastic noise propagation	[RUN]	[RUN]	[RUN]

Expected figures: Fig. 3 (Pareto fronts per method), Fig. 4 (hypervolume versus cumulative GPU-hours), Fig. 5 (empirical versus nominal conformal coverage by round).

4.5 AMP versus alternative site-selection rules (C2)

Report ΔTm per unit ΔPO loss for mutations chosen by AMP, distance-to-active-site, conservation, DCA/SCA and random selection (Fig. 6). *Criterion:* AMP supported if it yields a significantly larger ratio than all four alternatives. **[RUN]**

Prospective Experimental Validation (RQ2–RQ3)

Table 4. Characterization of selected variants (mean ± SD, n = 3 biological replicates).

Variant	Substitutions	ΔTm (°C)	t½ (min) at [T]	kcat/KM ($M^{-1}s^{-1}$) at T / T+ΔT	ΔCp‡	Dual success
Parent	–	0	[RUN]	[RUN]	[RUN]	–
Best baseline (B6)	[RUN]	[RUN]	[RUN]	[RUN]	[RUN]	[RUN]
MF-DynOpt-1	[RUN]	[RUN]	[RUN]	[RUN]	[RUN]	[RUN]
MF-DynOpt-2	[RUN]	[RUN]	[RUN]	[RUN]	[RUN]	[RUN]
…	…	…	…	…	…	…

Report dual-success counts per method with Wilson 95% intervals and the power-analysis target (≥ 50 variants per arm). **[RUN]**

Mechanistic Interpretation (RQ3; C3, C4)

Fig. 7: per-residue ΔRMSF maps, dynamical-network rewiring and hydrogen-bond persistence for the best variant versus the parent; ΔCp‡ from kcat(T) fits; HDX-MS protection differences. *Criteria:* C3 supported if ΔCp‡ tracks ΔPO; C4 supported if loop rigidification predicts activity loss only when PO decreases. **[RUN]**

Generalization

Frozen pipeline applied to two held-out enzyme families without retuning: hypervolume gain versus B4 and B6, and PO versus kcat/KM correlation. **[RUN]**

Discussion

Interpretation and Relation to Prior Work

The framework formalizes a view that experimental enzymology has long suggested: catalytic efficiency depends on how often the enzyme visits productive states, and mutations that act far from the active site can retune this population by altering networks of correlated motions [35,36,47–49]. By expressing this view as an optimizable objective with an explicit, testable link to kinetics (Eq. 3), the study moves beyond descriptive MD analysis. The cascade design follows the principle that physics should be spent where information per unit cost is highest, and the baselines B4 and B6 test whether that principle outperforms both modern active learning and the intuitive sequential pipeline. [Discuss observed outcomes here; if PO fails to outperform static proxies, report this as a substantive negative finding that bounds the scope of ensemble-based objectives.]

Limitations and Threats to Validity

- **Construct Validity.** PO is an occupancy, not a rate. Eq. (3) holds only under stated assumptions; for mechanisms limited by product release, multi-step chemistry or large-scale domain motions, a multi-state or committor-based extension is required.
- **Force-Field Dependence.** Conformational populations depend on the force field [44]. Dual force-field agreement mitigates but does not remove this risk, and classical MD cannot describe bond making and breaking.
- **Sampling.** Slow gating motions may exceed accessible timescales. We report effective sample sizes and use REST2/MSM for finalists, but microsecond-to-millisecond conformational selection remains partly unresolved.
- **Thermal-Stability Proxy.** MD at elevated temperature does not reproduce unfolding kinetics; it is used as a relative proxy and validated against experimental ΔTm.
- **Internal Validity.** Active-learning evaluation can leak information through repeated reuse of a calibration pool; we use rolling holdouts and pre-registration.
- **External Validity.** Results on [enzyme(s)] may not transfer to metalloenzymes, cofactor-dependent or membrane enzymes; generalization is tested only on the held-out families reported.
- **Epistasis.** Surrogate accuracy for variants with more than about five substitutions is limited; measured double-mutant cycles bound the error but do not eliminate it.

Practical Implications

If the pre-specified criteria are met, practitioners obtain (i) a small, diverse set of variants with elevated probability of improving both properties, (ii) an explicit compute-versus-information policy for incorporating MD at scale, and (iii) mechanistic hypotheses (which distal regions to loosen or tighten) that can be tested by HDX-MS, NMR or time-resolved crystallography.

Ethics, Biosafety and Dual Use

The framework targets industrial and environmental enzymes. Enzyme-engineering methods are generally dual-use; we do not apply the framework to toxins, virulence factors or pathogen-relevant targets, and all work uses [biosafety level 1] organisms under institutional approval.

Future Directions

Extensions include multi-substrate specificity design, differentiable surrogates of PO for gradient-based design, generative backbone remodeling of gating loops conditioned on PO, committor-based objectives, and integration with automated self-driving laboratories.

Conclusions

We propose treating the stability–activity trade-off as a problem of ensemble occupancy and provide a theory-grounded objective (Productive Occupancy), a calibrated multi-fidelity optimization architecture and a pre-registered, three-tier evidence plan that makes the claim falsifiable.

References

Arnold, F. H. Directed evolution: bringing new chemistry to life. *Angew. Chem. Int. Ed.* 57, 4143–4148 (2018).

Yang, K. K., Wu, Z. & Arnold, F. H. Machine-learning-guided directed evolution for protein engineering. *Nat. Methods* 16, 687–694 (2019).

Hie, B. L. & Yang, K. K. Adaptive machine learning for protein engineering. *Curr. Opin. Struct. Biol.* 72, 145–152 (2022).

Jumper, J. et al. Highly accurate protein structure prediction with AlphaFold. *Nature* 596, 583–589 (2021).

Lin, Z. et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science* 379, 1123–1130 (2023).

Dauparas, J. et al. Robust deep learning–based protein sequence design using ProteinMPNN. *Science* 378, 49–56 (2022).

Watson, J. L. et al. De novo design of protein structure and function with RFdiffusion. *Nature* 620, 1089–1100 (2023).

Dieckhaus, H., Brocidiacono, M., Randolph, N. Z. & Kuhlman, B. Transfer learning to leverage larger datasets for improved prediction of protein stability changes. *Proc. Natl Acad. Sci. USA* 121, e2314853121 (2024).

Tsuboyama, K. et al. Mega-scale experimental analysis of protein folding stability in biology and design. *Nature* 620, 434–444 (2023).

Tokuriki, N. & Tawfik, D. S. Stability effects of mutations and protein evolvability. *Curr. Opin. Struct. Biol.* 19, 596–604 (2009).

Beadle, B. M. & Shoichet, B. K. Structural bases of stability–function tradeoffs in enzymes. *J. Mol. Biol.* 321, 285–296 (2002).

Bruice, T. C. & Lightstone, F. C. Ground state and transition state contributions to the rates of intramolecular and enzymatic reactions. *Acc. Chem. Res.* 32, 127–136 (1999).

Hur, S. & Bruice, T. C. The near attack conformation approach to the study of the chorismate to prephenate reaction. *Proc. Natl Acad. Sci. USA* 100, 12015–12020 (2003).

Schymkowitz, J. et al. The FoldX web server: an online force field. *Nucleic Acids Res.* 33, W382–W388 (2005).

Alford, R. F. et al. The Rosetta all-atom energy function for macromolecular modeling and design. *J. Chem. Theory Comput.* 13, 3031–3048 (2017).

Meier, J. et al. Language models enable zero-shot prediction of the effects of mutations on protein function. *Adv. Neural Inf. Process. Syst.* 34 (2021).

Romero, P. A., Krause, A. & Arnold, F. H. Navigating the protein fitness landscape with Gaussian processes. *Proc. Natl Acad. Sci. USA* 110, E193–E201 (2013).

Lu, H. et al. Machine learning-aided engineering of hydrolases for PET depolymerization. *Nature* 604, 662–667 (2022).

Abraham, M. J. et al. GROMACS: high performance molecular simulations through multi-level parallelism from laptops to supercomputers. *SoftwareX* 1–2, 19–25 (2015).

Eastman, P. et al. OpenMM 7: rapid development of high performance algorithms for molecular dynamics. *PLoS Comput. Biol.* 13, e1005659 (2017).

Sugita, Y. & Okamoto, Y. Replica-exchange molecular dynamics method for protein folding. *Chem. Phys. Lett.* 314, 141–151 (1999).

Wang, L. et al. Accurate and reliable prediction of relative ligand binding potency in prospective drug discovery by way of a modern free-energy calculation protocol and force field. *J. Am. Chem. Soc.* 137, 2695–2703 (2015).

Shirts, M. R. & Chodera, J. D. Statistically optimal analysis of samples from multiple equilibrium states. *J. Chem. Phys.* 129, 124105 (2008).

Kennedy, M. C. & O'Hagan, A. Predicting the output from a complex computer code when fast approximations are available. *Biometrika* 87, 1–13 (2000).

Daulton, S., Balandat, M. & Bakshy, E. Differentiable expected hypervolume improvement for parallel multi-objective Bayesian optimization. *Adv. Neural Inf. Process. Syst.* 33 (2020).

Angelopoulos, A. N. & Bates, S. A gentle introduction to conformal prediction and distribution-free uncertainty quantification. *arXiv* 2107.07511 (2021).

Tibshirani, R. J., Barber, R. F., Candès, E. J. & Ramdas, A. Conformal prediction under covariate shift. *Adv. Neural Inf. Process. Syst.* 32 (2019).

Romano, Y., Patterson, E. & Candès, E. J. Conformalized quantile regression. *Adv. Neural Inf. Process. Syst.* 32 (2019).

Notin, P. et al. ProteinGym: large-scale benchmarks for protein fitness prediction and design. *Adv. Neural Inf. Process. Syst.* (Datasets and Benchmarks) 36 (2023).

Pancotti, C. et al. Predicting protein stability changes upon single-point mutation: a thorough comparison of the available tools on a new dataset. *Brief. Bioinform.* 23, bbab555 (2022).

Diaz, D. J. et al. Stability Oracle: a structure-based graph-transformer framework for identifying stabilizing mutations. *Nat. Commun.* 15, 6170 (2024).

Jiang, K. et al. Rapid in silico directed evolution by a protein language model with active learning (EVOLVEpro). *Science* (2024/2025) [verify volume and pages].

Daniel, R. M. & Danson, M. J. A new understanding of how temperature affects the catalytic activity of enzymes. *Trends Biochem. Sci.* 35, 584–591 (2010).

Arcus, V. L. et al. On the temperature dependence of enzyme-catalyzed rates. *Biochemistry* 55, 1681–1688 (2016).

Henzler-Wildman, K. & Kern, D. Dynamic personalities of proteins. *Nature* 450, 964–972 (2007).

Boehr, D. D., Nussinov, R. & Wright, P. E. The role of dynamic conformational ensembles in biomolecular recognition. *Nat. Chem. Biol.* 5, 789–796 (2009).

Warshel, A. et al. Electrostatic basis for enzyme catalysis. *Chem. Rev.* 106, 3210–3235 (2006).

Gelman, A. & Rubin, D. B. Inference from iterative simulation using multiple sequences. *Stat. Sci.* 7, 457–472 (1992).

Kulesza, A. & Taskar, B. Determinantal point processes for machine learning. *Found. Trends Mach. Learn.* 5, 123–286 (2012).

Brown, L. D., Cai, T. T. & DasGupta, A. Interval estimation for a binomial proportion. *Stat. Sci.* 16, 101–133 (2001).

Starr, T. N. & Thornton, J. W. Epistasis in protein evolution. *Protein Sci.* 25, 1204–1218 (2016).

Chodera, J. D. A simple method for automated equilibration detection in molecular simulations. *J. Chem. Theory Comput.* 12, 1799–1805 (2016).

Hollingsworth, S. A. & Dror, R. O. Molecular dynamics simulation for all. *Neuron* 99, 1129–1143 (2018).

Lindorff-Larsen, K. et al. Systematic validation of protein force fields against experimental data. *PLoS ONE* 7, e32131 (2012).

Chodera, J. D. & Noé, F. Markov state models of biomolecular conformational dynamics. *Curr. Opin. Struct. Biol.* 25, 135–144 (2014).