PROTEIN COUNTERFACTUALS VIA DIFFUSION-GUIDED LATENT OPTIMIZATION

Weronika Kłos[1,2] **Sidney Bender**[1,2] **Lukas Kades**[3]

[1]Machine Learning Group, Technische Universität Berlin, Berlin, Germany
[2]Berlin Institute for the Foundations of Learning and Data (BIFOLD)
[3]BASF Digital Solutions GmbH, Ludwigshafen am Rhein, Germany
{w.klos,s.bender}@tu-berlin.de lukas.kades@basf.com

ABSTRACT

Deep learning models can predict protein properties with unprecedented accuracy but rarely offer mechanistic insight or actionable guidance for engineering improved variants. When a model flags an antibody as unstable, the protein engineer is left without recourse: which mutations would rescue stability while preserving function? We introduce **Manifold-Constrained Counterfactual Optimization for Proteins (MCCOP)**, a framework that computes minimal, biologically plausible sequence edits that flip a model's prediction to a desired target state. MCCOP operates in a continuous joint sequence–structure latent space and employs a pretrained diffusion model as a manifold prior, balancing three objectives: validity (achieving the target property), proximity (minimizing mutations), and plausibility (producing foldable proteins). We evaluate MCCOP on three protein engineering tasks – GFP fluorescence rescue, thermodynamic stability enhancement, and E3 ligase activity recovery – and show that it generates sparser, more plausible counterfactuals than both discrete and continuous baselines. The recovered mutations align with known biophysical mechanisms, including chromophore packing and hydrophobic core consolidation, establishing MCCOP as a tool for both model interpretation and hypothesis-driven protein design. Our code is publicly available at github.com/weroks/mccop.

1 INTRODUCTION

Deep learning has transformed computational protein science. Structure prediction models achieve near-experimental accuracy (Jumper et al., 2021; Abramson et al., 2024), protein language models capture evolutionary grammar (Lin et al., 2023; Team et al., 2024), and generative frameworks design novel folds from scratch (Watson et al., 2023; Ingraham et al., 2023). Yet these models remain oracles rather than guides: when a predictor flags a candidate as "aggregation-prone", the engineer receives no indication of which mutations would resolve the problem.

This paper addresses the need for algorithmic recourse: given a protein P predicted to lack a desired property y_{target}, what is the minimal modification such that the prediction changes? This maps directly to *counterfactual explanations* (Wachter et al., 2017). Applied to a model of uncertain quality, counterfactuals expose reliance on spurious correlations; applied to a robust model, they generate testable hypotheses for wet-lab validation.

Translating counterfactual methods to proteins introduces two fundamental challenges. First, the *manifold constraint*: Unlike images, proteins are governed by strict epistatic constraints – a single core mutation can abolish folding while a compensatory mutation restores it. Naive gradient optimization produces adversarial or invalid examples that satisfy the predictor but correspond to unfoldable proteins. Second, discreteness and geometry: Proteins are *discrete sequences* whose function emerges from *continuous 3D geometry*. Gradient-based methods require continuous relaxation, while naively treating them as sequences ignores spatial relationships: a mutation can be compensatory for another one only if the residues are proximal in 3D space, a property not directly apparent from the 1D sequence.

We address both challenges with **MCCOP**, a gradient-based framework operating in a continuous joint sequence–structure embedding space that uses a pretrained diffusion model as a manifold prior. Our contributions are:

1. **Framework.** MCCOP combines predictor-guided gradient descent with diffusion-based manifold projection and gradient-sensitivity masking to produce sparse, valid, and plausible protein counterfactuals, without task-specific retraining of the generative model.
2. **Quantitative evaluation.** On three benchmarks, MCCOP achieves near-perfect success rates with 3–5$\times$ fewer mutations than discrete baselines and near-zero adversarial rates.
3. **Mechanistic interpretability.** MCCOP rediscovers known functional motifs and in several cases exactly recovers ground-truth counterfactual sequences from held-out test data.

An overview of our approach is depicted in Figure 1.

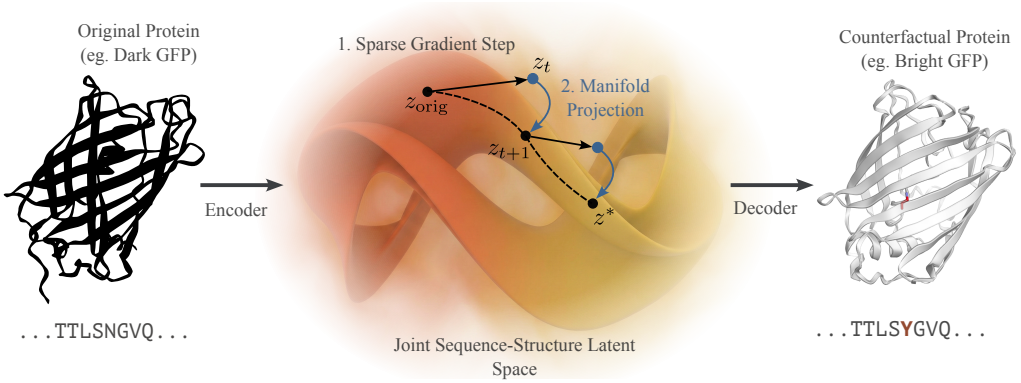

Figure 1: Overview of MCCOP. A non-fluorescent GFP variant is mapped to a continuous joint sequence–structure latent space via a pretrained autoencoder. After smoothing the classification boundary, the counterfactual embedding is optimized by alternating between (1) a sparse gradient step maximizing target class probability and (2) manifold projection using a pretrained diffusion model (DiMA). The counterfactual shown is a rediscovered sample from the held-out test set (red sticks denote the mutated residue).

2 Related Work

Protein language models and embeddings. Models such as ESM-2 (Lin et al., 2023) and ESM-C (Team et al., 2024) learn unsupervised representations from millions of sequences. Recent multimodal embeddings go further: CHEAP (Lu et al., 2025) compresses ESMFold (Lin et al., 2023) activations into a joint sequence–structure representation whose decoder maps back to both amino acid sequences and atomistic coordinates. This bidirectional mapping is central to our approach.

Diffusion and generative protein design. EvoDiff (Alamdari et al., 2023) and DiMA (Meshchaninov et al., 2024) apply diffusion to discrete sequences or continuous embeddings; RFdiffusion (Watson et al., 2023) and folding diffusion (Wu et al., 2024) operate in $SE(3)$ space (see Wang et al. (2025) for an overview). Most generative methods focus on conditional or unconditional sampling. MCCOP differs by using diffusion not for generation but as a regularizer within an optimization loop – conceptually an inversion of classifier guidance.

Explainability in the protein domain. Prior work relies on attention visualization (Vig et al., 2020), feature attribution (Sibli et al., 2025; Dickinson & Meyer, 2022), gradient-based structure perturbation (Tan & Zhang, 2023), or sparse autoencoders applied to pLMs (Gujral et al., 2025). Unlike passive attribution, MCCOP provides *active recourse*: not just why a protein is predicted to fail, but how to rescue it.

Counterfactual explanations. Counterfactuals, formalized for ML by Wachter et al. (2017), seek minimal input modifications that change a model's output – a concept not to be confused with causal counterfactual inference in the structural causal model (SCM) sense (Pearl, 2009). Methods for tabular data (Mothilal et al., 2020; Russell, 2019) are well established. For high-dimensional inputs, diffusion-based approaches (DVCE (Augustin et al., 2022), DIME (Jeanneret et al., 2022), Diff-ICE (Pegios et al., 2025), FastDiME (Weng et al., 2024), ACE (Jeanneret et al., 2023)) generate on-manifold counterfactuals via guided denoising, with extensions to diverse sets (Bender et al., 2025; Bender & Morik, 2026), graphs (Bechtoldt & Bender, 2026; Chen et al., 2023), and text (Sarkar, 2024). GAN/VAE-based predecessors include DiVE (Rodriguez et al., 2021) and Diffeomorphic Counterfactuals (Dombrowski et al., 2024). To our knowledge, no prior work applies diffusion-guided counterfactual optimization to proteins. The closest biological relatives – latent fitness optimization (Ngo et al., 2024; Castro et al., 2022) – seek global optima rather than minimal edits and train task-specific generative models.

3 METHODS

We now describe each component of MCCOP: the latent representation, predictor smoothing, and the counterfactual optimization loop itself.

3.1 PROBLEM FORMULATION

Let $\mathcal{M} \subset \mathbb{R}^{L' \times D}$ denote the manifold of biologically plausible protein embeddings. Given a predictor $f_\theta : \mathcal{M} \to \mathcal{Y}$ and an input embedding $z_0 \in \mathcal{M}$ with prediction $y_0 = f_\theta(z_0)$, we seek:

$$z^* = \arg\min_{z \in \mathcal{M}} \left[\mathcal{L}_{\text{task}}(f_\theta(z),\, y_{\text{target}}) + \lambda\, d(z,\, z_0) \right], \tag{1}$$

where d enforces proximity to z_0 and $z \in \mathcal{M}$ ensures plausibility. Without the manifold constraint, optimization yields adversarial examples (Dombrowski et al., 2024). We enforce it implicitly using the score function of a diffusion model trained on protein embeddings, whose denoising step acts as a projection $\Pi_{\mathcal{M}}$, interleaved with gradient steps on Eq. 1.

3.2 LATENT REPRESENTATION

We map sequences $S \in \mathcal{A}^L$ to continuous representations $z \in \mathbb{R}^{L' \times D}$ using CHEAP (Lu et al., 2025). The encoder $\mathcal{E}$ compresses ESMFold activations into embeddings jointly capturing evolutionary and structural information. The decoder $\mathcal{D}$ maps z back to both a sequence $\hat{S} = \mathcal{D}_{\text{seq}}(z)$ and backbone coordinates $\hat{\Omega} = \mathcal{D}_{\text{struct}}(z)$, with near-perfect round-trip reconstruction (>99% residue accuracy). Crucially, $\mathcal{D}$ is a position-wise MLP – each token $\hat{S}_i$ depends only on z_i – enabling sequence-level sparsity via row-wise latent masking (§3.4.2). Both encoder and decoder are frozen throughout.

3.3 PREDICTOR SMOOTHING

Our framework is model-agnostic: any differentiable predictor on CHEAP embeddings can be used. As a test-bed we train a shallow MLP f_θ on flattened embeddings (architecture details in Appendix A).

A non-smooth f_θ produces high-frequency gradients guiding optimization toward adversarial perturbations. Motivated by the observation of Bender et al. (2025), that smooth classifiers yield more reliable counterfactual optimization, we smooth f_θ via four complementary mechanisms: (1) **spectral normalization** (Miyato et al., 2018) on all linear layers; (2) **Jacobian regularization** (Jakubovitz & Giryes, 2018), penalizing $\|\nabla_z f_\theta(z)\|_F^2$; (3) **Softplus activations** ($\beta = 1$); and (4) **embedding-space adversarial augmentation** via FGSM (Goodfellow et al., 2014), where perturbations decoding to the original sequence are added with the original label, teaching invariance to semantically null perturbations. As shown in Table 1, this reduces gradient norms by up to 4× while maintaining or improving AUROC.

Algorithm 1 MCCOP: Manifold-Constrained Counterfactual Optimization for Proteins

Require: Embedding z_{orig}, predictor f_θ, diffusion projector Π_ϕ, target label $\tilde{y}$, sparsity k, projection strength α, margin m, learning rate η, max steps $T_{\max}$, confidence threshold τ

$z_0 \leftarrow z_{\text{orig}}$
for $t = 0, 1, \ldots, T_{\max} - 1$ **do**
  Compute $\mathcal{L}_{\text{CF}}(z_t)$ via Eq. 2
  Compute per-position sensitivity: $s_i = \|\nabla_{z_i} \mathcal{L}_{\text{CF}}\|_2$
  Construct top-k mask: $M_i = \mathbf{1}[s_i \geq s_{(k)}]$
  Gradient step: $z'_t = z_t - \eta \cdot (M \odot \nabla_z \mathcal{L}_{\text{CF}})$
  Hard reset: $z'_t[i] \leftarrow z_{\text{orig}}[i]$ for all i where $M_i = 0$
  Manifold projection: $z_{t+1} = (1 - \alpha)\, z'_t + \alpha\, \Pi_\phi(z'_t)$
  if $\sigma(\tilde{y} \cdot f_\theta(z_{t+1})) \geq \tau$ **and** $\mathcal{D}_{\text{seq}}(z_{t+1}) \neq S_{\text{orig}}$ **then**
    return z_{t+1} {Early stopping: valid counterfactual found}
  end if
end for
return $z_{T_{\max}}$ {Return best attempt}

3.4 Counterfactual Optimization

Given embedding z_{orig} with predicted class y_{orig}, we seek z^* such that $f_\theta(z^*) = y_{\text{target}} \neq y_{\text{orig}}$, minimizing decoded mutations while staying on $\mathcal{M}$. Algorithm 1 summarizes the procedure.

3.4.1 Objective Function

At step t, we minimize:

$$\mathcal{L}_{\text{CF}}(z_t) = \underbrace{\log\big(1 + \exp(m - \tilde{y} \cdot f_\theta(z_t))\big)}_{\mathcal{L}_{\text{margin}}} + \lambda_{\text{dist}} \underbrace{\|z_t - z_{\text{orig}}\|_2^2}_{\mathcal{L}_{\text{prox}}} \tag{2}$$

where $\tilde{y} \in \{-1, +1\}$ is the signed target label, $m > 0$ is a confidence margin, and λ_{dist} controls the proximity-validity trade-off.

3.4.2 Gradient-Based Sparsity Masking

We compute per-position sensitivity $s_i = \|\nabla_{z_i} \mathcal{L}_{\text{CF}}\|_2$ and construct a binary mask selecting the top-k positions:

$$M_i = \mathbf{1}\big[s_i \geq s_{(k)}\big]. \tag{3}$$

Gradients are applied only at masked positions; non-masked positions are hard-reset to z_{orig}. Because $\mathcal{D}$ is position-wise, row-wise masking in latent space directly enforces sequence-space sparsity. The mask can alternatively be user-defined for constrained editing (e.g., fixing catalytic residues).

3.4.3 Manifold Projection

We regularize the trajectory using DiMA (Meshchaninov et al., 2024) as an implicit manifold prior. At each step, we partially diffuse to noise level t_{diff}, denoise to obtain $\Pi_\phi(z'_t)$, and blend:

$$z_{t+1} = (1 - \alpha)\, z'_t + \alpha\, \Pi_\phi(z'_t), \tag{4}$$

where $\alpha \in [0, 1]$ controls projection strength ($\alpha = 0$: unconstrained; $\alpha = 1$: full projection, which destabilizes optimization). We use $\alpha = 0.3$ in practice (ablation in Appendix D).

3.5 Experimental Setup

3.5.1 Datasets

We evaluate on three datasets with diverse physical origins (statistics in Appendix G): **(1) TAPE Fluorescence** (Sarkisyan et al., 2016; Rao et al., 2019): GFP homologs with bimodal fluorescence, binarized into bright/dark classes (optimize dark→bright). **(2) TAPE Stability** (Rocklin et al.,

Table 1: Predictor AUROC and average L_2 gradient norm before and after smoothing (mean $\pm$ std, 3 seeds).

Dataset	AUROC (↑)		Avg. L_2 norm (↓)	
	Before	*After*	*Before*	*After*
Fluorescence	0.99 ± 0.00	0.99 ± 0.00	2.21 ± 0.19	1.10 ± 0.10
Stability	0.94 ± 0.00	0.98 ± 0.01	1.38 ± 0.13	0.36 ± 0.08
Activity	0.82 ± 0.00	0.93 ± 0.01	0.36 ± 0.06	0.33 ± 0.04

2017): proteolysis-based stability measurements; we remove the middle 33% quantile to create stable/unstable classes (optimize unstable→stable). **(3) Ube4b Activity** (Starita et al., 2013): ∼100k mutations in the U-box domain mapped to auto-ubiquitination activity; middle 33% removed, active/inactive classes defined by top/bottom quantiles (optimize inactive→active).

3.5.2 Baselines

We compare against: (1) **Stochastic Hill Climbing**: greedy random single-site mutations; (2) **Genetic Algorithm**: population-based evolution with edit-distance-penalized fitness; (3) **Gradient Descent**: unconstrained latent optimization without smoothing or manifold projection. Details in Appendix E.

3.5.3 Evaluation Metrics

We assess **validity and sparsity** via success rate (fraction achieving target class), Hamming distance (number of mutations), and adversarial rate (fraction of successful counterfactuals corresponding to the same sequence). **Structural plausibility** is evaluated using ESM3-predicted pLDDT confidence and radius of gyration (R_g). **Physicochemical plausibility** is monitored via GRAVY hydrophobicity, instability index, and a binary solubility proxy.

4 Results

We evaluate on complete test sets, excluding samples misclassified by the predictor. This results in $n = 2093$, 2209, and 2600 samples for the stability, fluorescence, and activity datasets respectively. Results are mean $\pm$ std over three seeds.

4.1 Predictor Smoothing Improves Robustness Without Sacrificing Accuracy

Table 1 shows that smoothing reduces gradient norms by up to 4× while maintaining or improving AUROC. The largest gain is on the activity dataset (AUROC: 0.82→0.93), likely because Jacobian regularization and adversarial augmentation reduce overfitting to noisy labels.

4.2 MCCOP Produces Valid and Sparse Counterfactuals

Table 2 reveals three key findings. **(1) Unconstrained gradient optimization is entirely adversarial:** every counterfactual decodes to the original sequence, confirming exploitable high-frequency artifacts and validating our smoothing and projection pipeline. **(2) MCCOP achieves high success with minimal edits:** 100% success on stability and activity with 2.3–2.5 mutations versus 6.2–10.9 for discrete baselines. MCCOP reaches early stopping after a median of 2–10 steps, while hill climbing exhausts the budget in >95% of cases. **(3) Fluorescence is harder:** MCCOP's 19% success rate reflects the requirement for precise chromophore geometry potentially exceeding our sparsity budget ($k = 5$), yet successful counterfactuals are the sparsest (1.4 mutations) with near-zero adversarial rate.

Edit distances for MCCOP and the genetic algorithm are tunable via k/λ_{dist} and fitness weighting, respectively; MCCOP's advantage lies in the favorable trade-off between success rate, sparsity, and plausibility.

Table 2: Success rate, adversarial rate, and edit distance (mean $\pm$ std, 3 seeds). Edit distance computed on successful counterfactuals only (confidence ≥ 0.95, edit distance ≥ 1). Discrete methods cannot produce adversarial examples by construction, so no values are bolded in this column. †Gradient Descent achieves 100% adversarial rate; edit distance is undefined.

Dataset	Method	Success Rate (↑)	Adv. Rate (↓)	Edit Dist. (↓)
Stability	Genetic Algorithm	0.55 ± 0.01	0.00 ± 0.00	7.76 ± 0.06
	Gradient Descent†	1.00 ± 0.00	1.00 ± 0.00	–
	Stochastic Hill Climb	0.23 ± 0.00	0.00 ± 0.00	9.46 ± 0.18
	MCCOP (ours)	$\mathbf{1.00} \pm 0.00$	0.03 ± 0.00	$\mathbf{2.32} \pm 0.01$
Fluorescence	Genetic Algorithm	$\mathbf{0.36} \pm 0.30$	0.00 ± 0.00	5.37 ± 3.09
	Gradient Descent†	1.00 ± 0.00	1.00 ± 0.00	–
	Stochastic Hill Climb	0.13 ± 0.00	0.00 ± 0.00	7.79 ± 0.30
	MCCOP (ours)	0.19 ± 0.00	0.01 ± 0.00	$\mathbf{1.37} \pm 0.01$
Activity	Genetic Algorithm	0.17 ± 0.15	0.00 ± 0.00	6.24 ± 3.67
	Gradient Descent†	1.00 ± 0.00	1.00 ± 0.00	–
	Stochastic Hill Climb	0.03 ± 0.00	0.00 ± 0.00	10.91 ± 0.02
	MCCOP (ours)	$\mathbf{1.00} \pm 0.00$	0.02 ± 0.02	$\mathbf{2.46} \pm 0.33$

4.3 Structural and Physicochemical Plausibility

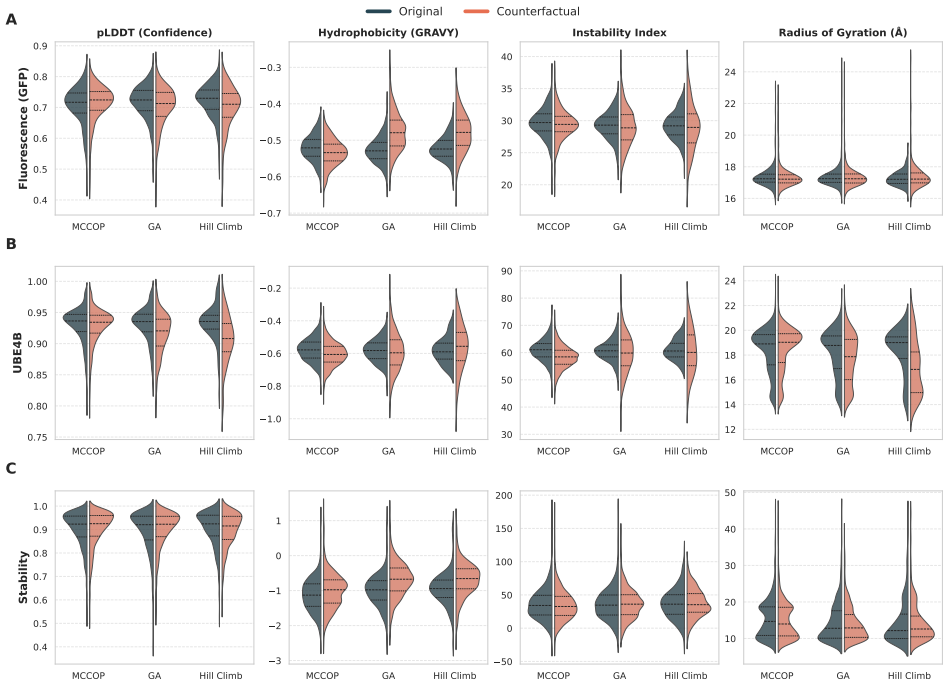

Figure 2: Physicochemical plausibility across benchmarks (columns: pLDDT, GRAVY, instability index, R_g; rows: fluorescence, activity, stability). MCCOP (orange) closely matches the original distribution (gray); discrete baselines show broader shifts. Statistical comparisons via Kruskal-Wallis/Dunn's tests with Benjamini-Hochberg correction: MCCOP achieves significantly higher pLDDT than both baselines across all tasks (adjusted $p < 0.02$).

Figure 2 shows that MCCOP counterfactuals are nearly indistinguishable from the original distribution across all metrics, occasionally shifting toward more favorable values. Discrete baselines introduce broader shifts, especially in hydrophobicity and instability index, as they explore sequence

space without structural priors. A controlled comparison at fixed edit distance (Appendix C) confirms these trends.

4.4 MCCOP Rediscovers Known Biophysical Mechanisms

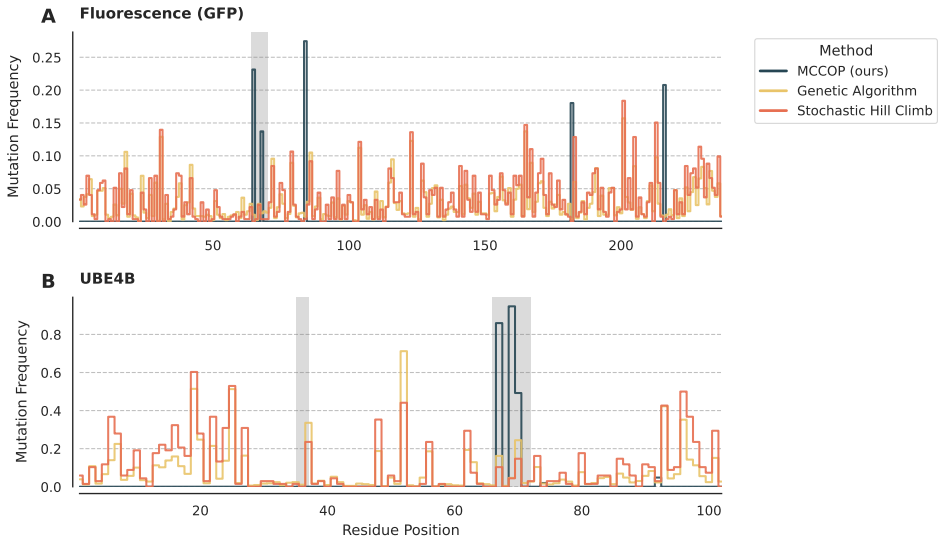

Figure 3: Per-residue mutation frequency for fluorescence (**A**) and Ube4b activity (**B**). MCCOP (blue) concentrates mutations in functionally relevant regions – chromophore-proximal residues for GFP, E2-binding interface for Ube4b – while baselines distribute mutations nearly uniformly. Shaded regions: known functional motifs (Sarkisyan et al., 2016; Starita et al., 2013).

GFP fluorescence. MCCOP concentrates mutations in the chromophore-proximal region (residues 63–69) and β-barrel strands forming the chromophore cavity (Figure 3A), consistent with the requirement for tight packing to suppress non-radiative decay (Sarkisyan et al., 2016). A small number of distal mutations (e.g., residues 181, 216) may represent novel compensatory interactions or predictor artifacts, requiring experimental follow-up.

Ube4b activity. Mutations cluster at the E2-binding interface (residues 66–71; Figure 3B), through which Ube4b recruits UbcH5c for ubiquitin transfer (Starita et al., 2013).

Thermodynamic stability. The stability dataset spans diverse topologies (Rocklin et al., 2017), so no universal residue positions dominate. However, MCCOP frequently targets core-facing residues, suggesting hydrophobic core consolidation as a general stabilization strategy (Appendix F).

Recovery of ground-truth counterfactuals. MCCOP exactly recovers existing opposite-label sequences in 16 (fluorescence), 18 (activity), and 4 (stability) cases – several from the held-out test set. Figure 4 shows structural alignments confirming that recovered mutations localize to functionally relevant regions.

5 Discussion

MCCOP generates sparse, on-manifold counterfactual explanations achieving near-perfect success rates with 1–3 mutations on average (vs. 8–11 for discrete baselines) while maintaining structural and physicochemical plausibility. Recovered mutations align with established biophysical mechanisms, suggesting that the underlying predictors have learned meaningful sequence-function relationships.

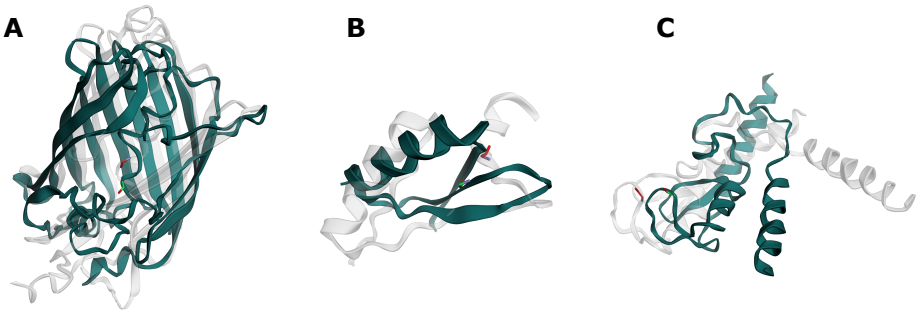

Figure 4: Structural alignments between original (gray) and counterfactual (colored) proteins for rediscovered ground-truth examples. (**A**) GFP: mutations near the chromophore. (**B**) Stability: core-facing mutations. (**C**) Ube4b: E2-binding interface mutations. Structures predicted by ESM3.

Explanation versus editing. MCCOP's primary goal is model interpretation, not direct engineering. A counterfactual is only as trustworthy as the predictor it explains: if the model has learned spurious correlations, the counterfactual reflects them faithfully – which is itself diagnostic. When the predictor is robust, MCCOP's outputs become candidates for experimental validation.

From correlation to causation. Our framework identifies correlational, not causal, relationships. Establishing true causal links requires interventional experiments, but MCCOP's sparse suggestions (2 mutations vs. thousands of directed-evolution variants) are directly amenable to such follow-up.

Limitations. (1) Plausibility evaluation relies on computational proxies (ESM3 pLDDT, R_g, physicochemical indices) rather than experimental validation. (2) The CHEAP encoder–decoder introduces reconstruction error that may produce artifacts for proteins distant from ESMFold's training distribution. (3) We evaluate only binary tasks; extending to continuous regression targets requires replacing the margin loss with MSE or quantile losses.

On the manifold and smoothness assumptions. Two assumptions embedded in our framework deserve scrutiny. *First*, MCCOP operates in a continuous latent space under the implicit premise that plausible protein sequences concentrate near a low-dimensional manifold. The same *manifold hypothesis* is routinely invoked in computer vision, where natural images are assumed to populate a thin subspace of pixel space, yet to the best of our knowledge no formal proof of this assertion exists for images or for proteins. Fefferman et al. develop statistical tests for the hypothesis but do not establish it for any natural data distribution (Fefferman et al., 2016); empirically, the evidence is consistent with data concentrating on disconnected clusters or "blobs" rather than a single smooth manifold (Bengio et al., 2013). For proteins, the situation is arguably more fraught: functional sequences are constrained by folding, stability, and epistasis, producing a viable sequence space that may be fragmented and topologically complex rather than smoothly connected. *Second*, MCCOP's Gaussian smoothing of the latent space presupposes that the underlying sequence–function mapping varies smoothly, so that local perturbations yield gradual changes in the predicted phenotype. However, protein fitness landscapes are known to be rugged: higher-order epistasis creates abrupt fitness transitions even between sequences that differ by a single residue (Weinreich et al., 2006; Sarkisyan et al., 2016), and there is no *a priori* reason to expect the predictor's decision surface, which reflects these landscapes, to be smooth either. An alternative to smoothing might be signal filtering tuned to a desired frequency, which would suppress high-frequency noise without globally flattening the landscape; we opted for Gaussian smoothing as a pragmatic engineering choice that made the gradient-based optimization tractable, rather than as a theoretically motivated operation. We flag these points not because they invalidate the results – MCCOP's strong empirical performance suggests the assumptions are serviceable in practice – but because they circumscribe the regime in which the method's outputs should be trusted and highlight opportunities for more principled geometric and spectral approaches in future work.

Future directions. (1) *Multi-objective counterfactuals*: jointly optimizing stability and binding affinity by combining predictor gradients. (2) *Experimental validation*: synthesizing top-ranked variants for closed-loop validation. (3) *Diverse counterfactual sets* (Mothilal et al., 2020; Bender & Morik, 2026): revealing alternative mutational strategies and enriching fitness landscape understanding.

ACKNOWLEDGMENTS

We would like to thank Marvin Sextro for many valuable pieces of advice and proof-reading, as well as Klaus-Robert Müller, Adrian Hill, and Stefan Chmiela for interesting and fruitful discussions. We also would like to thank Dominik Kühne for maintaining our HPC cluster hydra and being always there to help in case of technical difficulties. We used GitHub Copilot for assistance with code development and editing of paper text. All AI-generated content was reviewed, verified, and revised by the authors, who take full responsibility for the final manuscript.

This work was supported by the German Ministry for Education and Research (BMBF) under Grant 01IS18037A, and by BASLEARN – TU Berlin/BASF Joint Laboratory, co-financed by TU Berlin and BASF SE.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J Ballard, Joshua Bambrick, et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630(8016):493–500, 2024.

Sarah Alamdari, Nitya Thakkar, Rianne Van Den Berg, Neil Tenenholtz, Robert Strome, Alan M Moses, Alex X Lu, Nicolò Fusi, Ava P Amini, and Kevin K Yang. Protein generation with evolutionary diffusion: sequence is all you need. *BioRxiv*, pp. 2023–09, 2023.

Maximilian Augustin, Valentyn Boreiko, Francesco Croce, and Matthias Hein. Diffusion visual counterfactual explanations. *Advances in Neural Information Processing Systems*, 35:364–377, 2022.

David Bechtoldt and Sidney Bender. Graph diffusion counterfactual explanation. *ESANN*, 2026.

Sidney Bender and Marco Morik. Visual disentangled diffusion autoencoders. *arXiv preprint arXiv:2601.21851*, 2026.

Sidney Bender, Jan Herrmann, Klaus-Robert Müller, and Grégoire Montavon. Towards desiderata-driven design of visual counterfactual explainers. *Pattern Recognition*, 2025.

Yoshua Bengio, Aaron Courville, and Pascal Vincent. Representation learning: A review and new perspectives. *IEEE transactions on pattern analysis and machine intelligence*, 35(8):1798–1828, 2013.

Egbert Castro, Abhinav Godavarthi, Julian Rubinfien, Kevin B Givechian, Dhananjay Bhaskar, and Smita Krishnaswamy. Relso: a transformer-based model for latent space optimization and generation of proteins. *arXiv preprint arXiv:2201.09948*, 2022.

Jialin Chen, Shirley Wu, Abhijit Gupta, and Rex Ying. D4explainer: In-distribution explanations of graph neural network via discrete denoising diffusion. *Advances in Neural Information Processing Systems*, 36:78964–78986, 2023.

Quinn Dickinson and Jesse G Meyer. Positional shap (poshap) for interpretation of machine learning models trained from biological sequences. *PLOS Computational Biology*, 18(1):e1009736, 2022.

Ann-Kathrin Dombrowski, Jan E Gerken, Klaus-Robert Müller, and Pan Kessel. Diffeomorphic counterfactuals with generative models. *IEEE Transactions on Pattern Recognition and Machine Intelligence*, 46(5):3257–3274, 2024.

Charles Fefferman, Sanjoy Mitter, and Hariharan Narayanan. Testing the manifold hypothesis. *Journal of the American Mathematical Society*, 29(4):983–1049, 2016.

Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. *arXiv preprint arXiv:1412.6572*, 2014.

Onkar Gujral, Mihir Bafna, Eric Alm, and Bonnie Berger. Sparse autoencoders uncover biologically interpretable features in protein language model representations. *Proceedings of the National Academy of Sciences*, 122(34):e2506316122, 2025.

John B Ingraham, Max Baranov, Zak Costello, Karl W Barber, Wujie Wang, Ahmed Ismail, Vincent Frappier, Dana M Lord, Christopher Ng-Thow-Hing, Erik R Van Vlack, et al. Illuminating protein space with a programmable generative model. *Nature*, 623(7989):1070–1078, 2023.

Daniel Jakubovitz and Raja Giryes. Improving dnn robustness to adversarial attacks using jacobian regularization. In *Proceedings of the European conference on computer vision (ECCV)*, pp. 514–529, 2018.

Guillaume Jeanneret, Loïc Simon, and Frederic Jurie. Diffusion models for counterfactual explanations. In *Proceedings of the Asian Conference on Computer Vision*, pp. 858–876, 2022.

Guillaume Jeanneret, Loïc Simon, and Frederic Jurie. Adversarial counterfactual visual explanations. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pp. 16425–16435, 2023.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, et al. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023.

Amy X Lu, Wilson Yan, Kevin K Yang, Vladimir Gligorijevic, Kyunghyun Cho, Pieter Abbeel, Richard Bonneau, and Nathan C Frey. Tokenized and continuous embedding compressions of protein sequence and structure. *Patterns*, 6(6), 2025.

Viacheslav Meshchaninov, Pavel Strashnov, Andrey Shevtsov, Fedor Nikolaev, Nikita Ivanisenko, Olga Kardymon, and Dmitry Vetrov. Diffusion on language model encodings for protein sequence generation. *arXiv preprint arXiv:2403.03726*, 2024.

Takeru Miyato, Toshiki Kataoka, Masanori Koyama, and Yuichi Yoshida. Spectral normalization for generative adversarial networks. *arXiv preprint arXiv:1802.05957*, 2018.

Ramaravind K Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 conference on fairness, accountability, and transparency*, pp. 607–617, 2020.

Nhat Khang Ngo, Thanh VT Tran, Viet Thanh Duy Nguyen, and Truong Son Hy. Latent-based directed evolution accelerated by gradient ascent for protein sequence design. *bioRxiv*, pp. 2024–04, 2024.

Judea Pearl. *Causality*. Cambridge university press, 2009.

Paraskevas Pegios, Manxi Lin, Nina Weng, Morten Bo Søndergaard Svendsen, Zahra Bashir, Siavash Bigdeli, Anders Nymark Christensen, Martin Tolsgaard, and Aasa Feragen. Diffusion-based iterative counterfactual explanations for fetal ultrasound image quality assessment. In *International Workshop on Advances in Simplifying Medical Ultrasound*, pp. 174–184. Springer, 2025.

Roshan Rao, Nicholas Bhattacharya, Neil Thomas, Yan Duan, Peter Chen, John Canny, Pieter Abbeel, and Yun Song. Evaluating protein transfer learning with tape. *Advances in neural information processing systems*, 32, 2019.

Gabriel J Rocklin, Tamuka M Chidyausiku, Inna Goreshnik, Alex Ford, Scott Houliston, Alexander Lemak, Lauren Carter, Rashmi Ravichandran, Vikram K Mulligan, Aaron Chevalier, et al. Global analysis of protein folding using massively parallel design, synthesis, and testing. *Science*, 357 (6347):168–175, 2017.

Pau Rodriguez, Massimo Caccia, Alexandre Lacoste, Lee Zamparo, Issam Laradji, Laurent Charlin, and David Vazquez. Beyond trivial counterfactual explanations with diverse valuable explanations. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pp. 1056–1065, 2021.

Chris Russell. Efficient search for diverse coherent explanations. In *Proceedings of the conference on fairness, accountability, and transparency*, pp. 20–28, 2019.

Advait Sarkar. Large language models cannot explain themselves. *arXiv preprint arXiv:2405.04382*, 2024.

Karen S Sarkisyan, Dmitry A Bolotin, Margarita V Meer, Dinara R Usmanova, Alexander S Mishin, George V Sharonov, Dmitry N Ivankov, Nina G Bozhanova, Mikhail S Baranov, Onuralp Soylemez, et al. Local fitness landscape of the green fluorescent protein. *Nature*, 533(7603):397–401, 2016.

Sabbir Ahmed Sibli, Vlasios Panagiotis Panagiotou, and Christos Makris. Enhancing protein structure predictions: Deepshap as a tool for understanding alphafold2. *Expert Systems with Applications*, pp. 127853, 2025.

Lea M Starita, Jonathan N Pruneda, Russell S Lo, Douglas M Fowler, Helen J Kim, Joseph B Hiatt, Jay Shendure, Peter S Brzovic, Stanley Fields, and Rachel E Klevit. Activity-enhancing mutations in an e3 ubiquitin ligase identified by high-throughput mutagenesis. *Proceedings of the National Academy of Sciences*, 110(14):E1263–E1272, 2013.

Juntao Tan and Yongfeng Zhang. Explainablefold: Understanding alphafold prediction with explainable ai. In *Proceedings of the 29th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pp. 2166–2176, 2023.

ESM Team et al. Esm cambrian: Revealing the mysteries of proteins with unsupervised learning. *Evolutionary Scale Website https://www. evolutionaryscale. ai/blog/esm-cambrian*, 2024.

Jesse Vig, Ali Madani, Lav R Varshney, Caiming Xiong, Richard Socher, and Nazneen Fatema Rajani. Bertology meets biology: Interpreting attention in protein language models. *arXiv preprint arXiv:2006.15222*, 2020.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harv. JL & Tech.*, 31:841, 2017.

Chentong Wang, Sarah Alamdari, Carles Domingo-Enrich, Ava P Amini, and Kevin K Yang. Toward deep learning sequence–structure co-generation for protein design. *Current Opinion in Structural Biology*, 91:103018, 2025.

Joseph L Watson, David Juergens, Nathaniel R Bennett, Brian L Trippe, Jason Yim, Helen E Eisenach, Woody Ahern, Andrew J Borst, Robert J Ragotte, Lukas F Milles, et al. De novo design of protein structure and function with rfdiffusion. *Nature*, 620(7976):1089–1100, 2023.

Daniel M Weinreich, Nigel F Delaney, Mark A DePristo, and Daniel L Hartl. Darwinian evolution can follow only very few mutational paths to fitter proteins. *science*, 312(5770):111–114, 2006.

Nina Weng, Paraskevas Pegios, Eike Petersen, Aasa Feragen, and Siavash Bigdeli. Fast diffusion-based counterfactuals for shortcut removal and generation. In *European Conference on Computer Vision*, pp. 338–357. Springer, 2024.

Kevin E Wu, Kevin K Yang, Rianne van den Berg, Sarah Alamdari, James Y Zou, Alex X Lu, and Ava P Amini. Protein structure generation via folding diffusion. *Nature communications*, 15(1): 1059, 2024.

A Predictor Training and Smoothing Details

Architecture. The property predictor f_θ is a three-layer MLP with hidden dimensions [512, 256], each followed by spectral normalization (Miyato et al., 2018) and Softplus activation ($\beta = 1$). The final layer outputs a single logit. Input embeddings are flattened resulting in an input dimension of sequence length L times embedding dimension D (masking padding tokens) before being passed to the MLP.

Training protocol. We train on 80% of each dataset, reserving 10% for validation and 10% for testing, stratified by label. We use Adam with a learning rate of 1×10^{-5} and a dropout rate of 0.3. Early stopping is applied with a patience of 5 epochs based on validation AUROC.

Smoothing mechanisms. Further details concerning the smoothing mechanisms include:

1. **Spectral normalization** (Miyato et al., 2018): applied to all linear layers, constraining the Lipschitz constant of each layer to approximately 1.
2. **Jacobian regularization** (Jakubovitz & Giryes, 2018): we add a penalty term $\lambda_J \|\nabla_z f_\theta(z)\|_F^2$ to the training loss, with $\lambda_J = 10^{-3}$. The Frobenius norm is estimated via a Hutchinson trace estimator with 5 random projections per batch for computational efficiency.
3. **Adversarial data augmentation**: for each training sample (z_i, y_i), we generate an adversarial embedding z_i^{adv} via an FGSM attack ($\epsilon = 0.01$ in embedding space) targeting the opposite class. Adversarial samples that decode to the *same* amino acid sequence as the original (i.e., $\mathcal{D}(z_i^{\text{adv}}) = \mathcal{D}(z_i)$) are added to the training set with the original label y_i, teaching the model to be invariant to semantically null perturbations.

Smoothness quantification. We report the average L_2 gradient norm $\mathbb{E}_{z \sim \mathcal{D}_{\text{test}}}[\|\nabla_z f_\theta(z)\|_2]$ computed over the full test set. Lower values indicate a smoother decision boundary.

B Computational Cost Analysis

Figure 5 reports the average wall-clock time per sample for MCCOP and the two discrete baselines across all three benchmarks. The genetic algorithm is the most expensive method by roughly an order of magnitude due to its population-based evaluation. MCCOP and stochastic hill climbing have comparable per-sample execution times.

An important caveat applies to this comparison. The discrete baselines operate in sequence space but must evaluate candidates using the same embedding-space predictor; every proposed mutation therefore requires a full re-encoding through the CHEAP encoder (backed by ESMFold), which constitutes 97% and 94% of total computation time for hill climbing and the genetic algorithm respectively. This overhead is not intrinsic to those algorithms but arises from the requirement of a shared evaluation protocol. MCCOP, by contrast, operates natively in embedding space and avoids this round-trip entirely. Its dominant cost is the diffusion-based manifold projection, which accounts for 99% of computation time. For performance-critical applications we would recommend executing the diffusion-based projection step only every n optimization steps.

C Controlled Edit Distance Comparison

To ensure a fair comparison across methods, we filter all successful counterfactuals to those with exactly three mutations (edit distance = 3), which represents the bin with the highest overlap across MCCOP, the genetic algorithm, and stochastic hill climbing. Figure 6 shows the same physicochemical property distributions as Figure 2 in the main text, restricted to this subset. The trends observed in the main text are preserved: MCCOP-generated counterfactuals remain within the distribution of the original test set for pLDDT, GRAVY, instability index, and radius of gyration, whereas the discrete baselines show broader deviations, particularly in instability index and GRAVY.

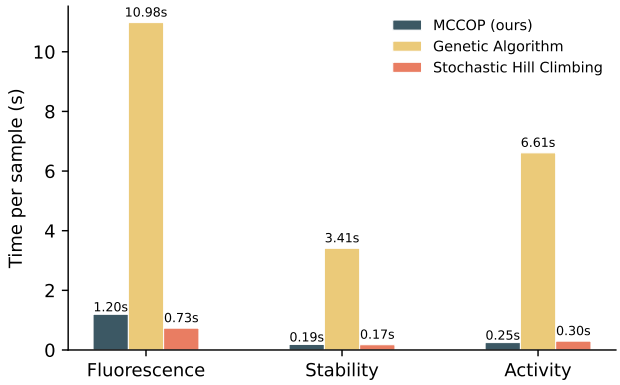

Figure 5: Average wall-clock time per sample across the complete test set. The genetic algorithm is the most expensive method. MCCOP and stochastic hill climbing have comparable execution times, though their computational profiles differ: discrete baselines spend >94% of time on re-encoding candidate sequences, while MCCOP spends 99% on diffusion-based manifold projection.

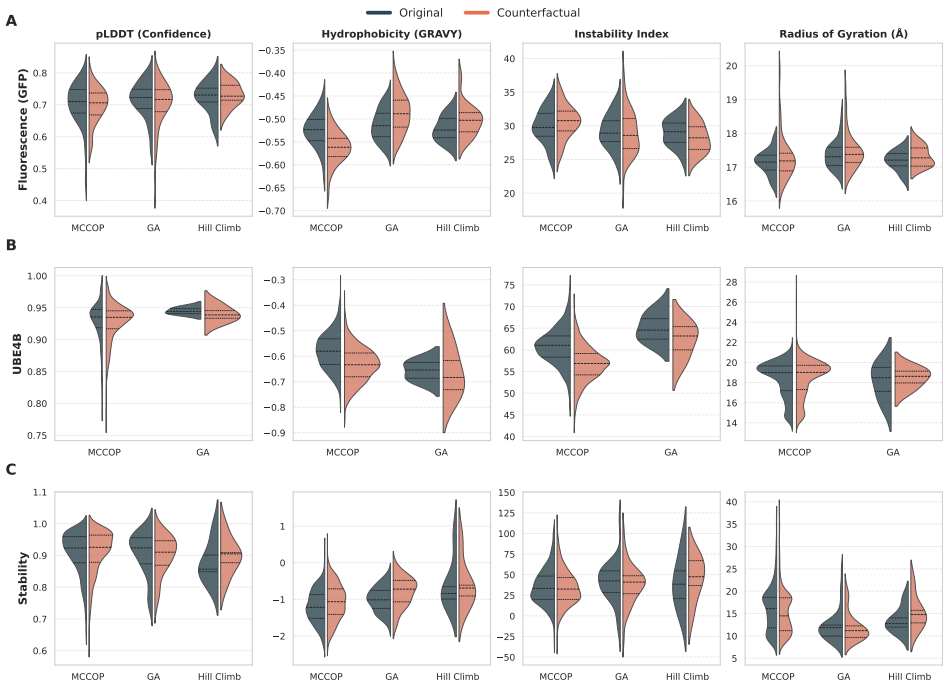

Figure 6: Physicochemical property distributions for counterfactuals with exactly 3 mutations. MCCOP closely tracks the original test set distribution, while the genetic algorithm and stochastic hill climbing show greater deviation, particularly for GRAVY and instability index.

D Hyperparameter Sensitivity

Table 3 lists the primary hyperparameters for MCCOP and the values used. We use the same set of hyperparameters across all three datasets.

We perform an ablation over all combinations of the previously mentioned smoother components (including no smoothing) as well as the manifold projection step and the masking value k. No smoothing, no masking and no manifold projection corresponds exactly to the gradient descent baseline. Lower k values proved to be too restrictive and caused low success rates while higher ones

Table 3: MCCOP hyperparameters and their values across benchmarks.

Symbol	Description	Fluorescence	Stability	Activity
k	Top-k masked positions	5	5	5
λ_{dist}	L_2 distance weight	0.1	0.1	0.1
m	Margin in $\mathcal{L}_{\text{margin}}$	2.2	2.2	2.2
α	Projection strength	0.3	0.3	0.3
t_{diff}	Diffusion noise level	100	100	100
η	Learning rate	0.5	0.5	0.5
$T_{\max}$	Max optimization steps	50	50	50

provided only a marginal increase in success rate. No smoothing significantly increases adversarial rates, while no manifold projection significantly reduces pLDDT scores of generated counterfactuals.

E Baselines Implementation Details

We compare our method against three baseline counterfactual explanation strategies, each operating over protein sequences and their corresponding embeddings. All baselines share a common interface and are evaluated using the same predictor and confidence threshold ($\tau = 0.95$ by default). Below we describe each baseline along with its hyperparameters.

E.1 Gradient Descent

This baseline performs standard gradient descent directly in the continuous embedding space. Given an input embedding $\mathbf{x}$, a differentiable copy $\mathbf{x}'$ is optimized to maximize the predictor's probability of the target (flipped) class via binary cross-entropy loss. The Adam optimizer is used to update $\mathbf{x}'$ over a fixed number of steps. At each step, the candidate counterfactual with the highest confidence toward the target class is retained.

Table 4: Hyperparameters for the gradient descent baseline.

Hyperparameter	Symbol	Value
Learning rate	η	1×10^{-2}
Gradient steps	T	50
Confidence threshold	τ	0.95
Optimizer	–	Adam
Loss function	$\mathcal{L}$	BCEWithLogitsLoss

Notably, this baseline operates entirely in embedding space and does not enforce any manifold constraints or discrete sequence validity, making it a purely continuous relaxation approach.

E.2 Random Mutation

This baseline performs a stochastic hill-climbing search in discrete sequence space. At each step, a single random point mutation is applied to each unsolved sequence: a uniformly random position is selected and replaced with a uniformly random amino acid from the standard 20-letter alphabet. The mutated sequence is re-encoded into embedding space using a lightweight encoder, and the predictor evaluates the new embedding. If the target-class confidence improves, the mutation is accepted; otherwise, the sequence reverts to the previous best. The process repeats for a fixed number of steps.

E.3 Genetic Algorithm

This baseline employs a population-based evolutionary strategy operating in discrete sequence space. For each input sequence, an initial population is constructed by applying random mutations to

Table 5: Hyperparameters for the Random Mutation baseline.

Hyperparameter	Symbol	Value
Maximum steps	T	50
Confidence threshold	τ	0.95
Amino acid alphabet	$\mathcal{A}$	Standard 20
Mutations per step	–	1

the original. At each generation, individuals are evaluated by encoding them into embedding space and computing a fitness score defined as the predictor's target-class confidence, optionally penalized by the Hamming distance to the original sequence:

$$f(\mathbf{s}) = \text{conf}(\mathbf{s}) - \lambda \cdot d_H(\mathbf{s}, \mathbf{s}_{\text{orig}}), \tag{5}$$

where $\text{conf}(\mathbf{s})$ is the predictor's confidence on the target class for sequence $\mathbf{s}$, d_H denotes the Hamming distance, and λ is the edit distance penalty. Selection uses tournament selection with tournament size 3. The top 20% of the population is preserved as elites. Offspring are generated via single-point crossover and random point mutation (1–2 mutations per offspring). Evolution proceeds for a fixed number of generations or until all samples exceed the confidence threshold.

Table 6: Hyperparameters for the Genetic Algorithm baseline.

Hyperparameter	Symbol	Value
Population size	N	40
Generations	G	30
Crossover rate	p_c	0.5
Edit distance penalty	λ	0.02
Confidence threshold	τ	0.95
Elite fraction	–	20%
Tournament size	k	3
Mutations per offspring	–	1–2
Maximum batch size	–	8

F Additional Structural Visualizations

We provide two additional structure visualizations for the stability dataset as we could not verify hydrophobic core packing by investigating mutation frequencies per residue.

Figure 7: Structural alignment of original (gray) and counterfactual (cyan) stability variants across three topologies. Core-facing mutated residues shown as sticks.

Table 7: Dataset statistics after preprocessing.

Dataset	Sequences	Positive	Negative	Avg. Length	Binarization
Fluorescence	54,025	30,697	23,328	236.96	Bimodal split
Stability	45,901	22,694	23,207	45.06	Remove middle 33%
Activity	60,692	30,293	30,399	102	Remove middle 33%

G Dataset Statistics and Preprocessing

For the fluorescence dataset, we exploit the natural bimodality of the log-fluorescence distribution and determine the optimal threshold using Otsu's method. For the stability and activity datasets, removing the middle tercile ensures a clear margin between classes, reducing label noise near the decision boundary. All embeddings are computed using the CHEAP encoder with ESMFold as the backbone, producing representations of dimension $D = 1024$ with no compression along the length dimension.