Fidelity-Concordance Reward Calibration for Reliable Generative Design

Wenhui Sophia Lu [1] **Xiaowei Zhang** [2 3] **Xiaojing Gao** [4] **Dominik Rothenhäusler** [1] **Wing Hung Wong** [1 5]

Abstract

Reward-guided generative design is widely used to steer protein models toward desired functional properties, but learned surrogate rewards can become unreliable in the high-score regions induced by optimization. In protein engineering, this unreliability carries direct experimental costs: candidates that optimize an assay-specific proxy may fail downstream wet-lab validation by losing activity, misfolding, or violating feasibility constraints required for biological function. We introduce *Fidelity-Concordance Steering* (FiCS), an uncertainty-aware reward-calibration framework that combines an inexpensive primary reward with sparse feedback from high-fidelity experimental or computational assessments. FiCS uses sparse high-fidelity feedback to calibrate the reward before optimization, replacing a single point-estimate surrogate with a reliability-adjusted steering signal. Across synthetic benchmarks and a renin *in silico* experiment, FiCS improves biological reliability by selecting candidates with higher oracle feasibility while preserving strong primary-reward performance relative to point-estimate and unweighted ensemble baselines. These gains are most pronounced in small-batch regimes where reliable candidate prioritization is crucial.

The central challenge in protein variant engineering is to discover sequences that improve complex functional traits such as catalytic efficiency, binding affinity, or thermal stability within a sequence space that grows as $20^{\text{sequence length}}$. While recent generative models have successfully captured the broad structure of natural proteins (Koh et al., 2025), they model the distribution of biologically plausible sequences rather than any specific engineering objective. Generative protein design therefore requires objective-specific optimization on top of pretrained models that primarily capture biological plausibility. Learned rewards provide this signal across pipelines such as fine-tuning, best-of-N selection, reranking, and test-time steering. The bottleneck studied here is not the choice of steering pipeline, but the reliability of the reward being optimized. Even when a surrogate performs well under ordinary predictive evaluation, optimization can amplify small reward errors until a surrogate artifact becomes the dominant selection signal. Thus, the problem is not simply how to steer a pretrained generator, but how to construct a reliable reward signal when high-fidelity biological feedback is too sparse to densely relabel the search space or reliably fine-tune the generator. In this low-oracle regime, sparse feedback is most naturally used to assess the reliability of the reward landscape rather than to learn a new generator.

This reward-reliability problem arises from two structural limitations of the data used to train surrogate models. First, predictive models are often fit using data from high-throughput experimental screening assays, where library construction restricts exploration to localized neighborhoods or low-order substitutions. Because protein fitness landscapes are rugged and shaped by pervasive epistatic interactions, models trained on these simple mutations cannot reliably extrapolate to complex, multi-site variants where mutational effects are non-additive (Wittmann et al., 2021). Second, biological assays suffer from restricted dynamic ranges: optimized to resolve highly functional variants, they often yield compressed or noisy measurements in low-fitness regimes. Consequently, these datasets provide an incomplete and structurally biased view of the fitness landscape.

These data limitations are further compounded by reward-guided steering. By concentrating samples in the high-score tail of the surrogate, optimization pushes candidates into out-of-distribution regions where the learned reward model must extrapolate beyond the regime supported by training data. In these regions, high surrogate scores may reflect spurious optimism rather than true biological fitness. The optimizer thereby exploits estimation error rather than discovering genuine high-fitness variants—a phenomenon known as reward hacking. Existing test-time alignment methods primarily

[1]Department of Statistics, Stanford University [2]Department of Bioengineering, Stanford University [3]Sarafan ChEM-H, Stanford University [4]Department of Chemical Engineering, Stanford University [5]Department of Biomedical Data Science, Stanford University. Correspondence to: Wing Hung Wong <whwong@stanford.edu>, Wenhui Sophia Lu <sophialu@stanford.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

address the downstream question of how to control sampling once a reward has been specified, without over-optimizing that reward or drifting too far from the pretrained model. In contrast, FiCS acts upstream of sampling by assessing whether the reward remains trustworthy in the regions that steering is likely to exploit.

Preventing this misalignment requires evaluating candidates against biological ground truth, either through rigorous experimental validation or through more faithful computational evaluators that capture additional biological constraints. Such high-fidelity evaluation is resource-intensive and, frequently, can be applied only to a sparse subset of candidates, making it insufficient for directly verifying every generated sequence or for reliably adapting the generator itself. In this regime, sparse high-fidelity feedback is most valuable as a calibration signal for the fitted reward, because it can identify which primary reward guides remain concordant with the high-fidelity evaluator in the high-reward regions that steering is likely to exploit.

To this end, we propose *Fidelity-Concordance Steering* (FiCS), an uncertainty-aware reward-calibration framework that combines an inexpensive primary reward with sparse high-fidelity feedback. FiCS replaces point-estimate rewards with an uncertainty-aware score that integrates three signals: resampling-based reward stability, concordance with sparse high-fidelity feedback, and base-model plausibility. The resulting objective favors candidates that score highly under the primary reward while remaining stable, concordant with high-fidelity feedback, and well supported by the base model.

Our main contributions are as follows:

1. We formulate reward-guided generative design with sparse high-fidelity feedback as constrained surrogate optimization, where candidates must achieve high primary reward while satisfying a feasibility threshold under a sparsely queried high-fidelity evaluator.
2. We propose FiCS, a modular reward-reliability framework that combines resampling-based ensemble stability, sparse oracle concordance, and base-model support into a reliability-adjusted reward usable across reward-guided pipelines.
3. We establish theoretical guarantees showing that (i) oracle-concordance scores concentrate under sparse feedback, (ii) the FiCS score is stable under estimation error, and (iii) FiCS-selected candidates match the oracle-optimal solution when the score gap at the selection boundary is sufficiently large.
4. We validate FiCS on synthetic benchmarks and protein localization tasks, demonstrating that oracle-concordance reweighting yields substantial improvements in ground-truth primary reward, feasibility rates, and constrained utility relative to point-estimate and unweighted ensemble baselines.

1. Background

Setup and Notation. Let x denote a candidate protein sequence or structure, and let $p_0(x)$ denote the density or probability mass function induced by a pretrained generative model. We treat p_0 as a reference distribution for biologically plausible candidates. Inference-time steering modifies sampling from p_0 to favor candidates with high values of an inexpensive primary reward or property of interest, denoted $R_1(x)$. In practice, R_1 may be available through a learned predictor; we write $\widehat{f}$ for a fitted point-estimate model of this primary reward and $\{\widehat{f}_k\}_{k=1}^{K}$ for resampled fitted primary reward models. The reward-training data are drawn from a distribution Q_{train}, with density q_{train} when it exists.

While the primary reward R_1 is inexpensive to evaluate, it provides only a partial picture of the true biological objective, as it is typically fit to a limited and structurally biased set of empirical measurements. In contrast, we assume the high-fidelity evaluator R_2 provides trustworthy assessments of candidate quality—such as wet-lab assay results, structural confidence scores, solubility measurements, or foldability predictions. However, because evaluating R_2 is resource-intensive, it cannot be applied to every generated candidate. Instead, we leverage sparse evaluations on a carefully selected subset to calibrate the relationship between the inexpensive primary reward and the higher-fidelity biological signal.

1.1. Related Work

A broad line of work in protein engineering uses predictive uncertainty to guide experimental design, including Bayesian optimization and active-learning methods that allocate assay budgets across variant libraries (Yang et al., 2025; Biswas et al., 2021; Stanton et al., 2022; Wittmann et al., 2021). A distinct line uses uncertainty or model averaging to improve objective reliability, for example by penalizing model uncertainty under distribution shift (Yu et al., 2020) or averaging reward models to mitigate reward fragility (Ramé et al., 2024). FiCS uses uncertainty for a different purpose. Rather than selecting which variants to assay next, it calibrates the reward used for candidate selection. This places FiCS closer to methods that make learned optimization objectives reliable than to active-learning methods for assay allocation, but with a steering-specific source of distribution shift. In reward-guided steering, optimizing a learned score can move candidate selection into regions poorly supported by the reward-training data, where high predicted scores may reflect spurious optimism rather than genuine biological utility. Unlike reward-model averaging, which improves robustness by aggregating predictors,

FiCS explicitly uses sparse high-fidelity feedback to assess whether the high-score tail remains concordant with the high-fidelity evaluator. Thus, uncertainty is used not to guide future assays, but to construct and calibrate a robust steering signal that accounts for reward instability while leveraging limited high-fidelity feedback.

Beyond protein design, DAS (Kim et al., 2025) uses tempered Sequential Monte Carlo to optimize a specified reward while staying close to the pretrained prior. DAS addresses how to optimize a trusted reward without excessive drift, whereas FiCS addresses whether that reward is reliable enough to optimize in the first place. The two approaches are therefore composable, with FiCS providing the reliability-adjusted reward used inside SMC-based test-time alignment and other reward-guided steering procedures.

1.2. Inference-Time Steering

Inference-time steering refers to a family of methods that guide the outputs of a pretrained generative model at test time without modifying the model's parameters. The objective is to bias generation toward candidates with high predicted reward, such as proteins with desired binding affinity, while preserving consistency with the base distribution p_0. Given p_0 and a learned reward model $\widehat{f}$, common strategies include best-of-N selection, classifier-guided generation (Dhariwal & Nichol, 2021), and distributional tilting (Jain et al., 2025; Viggiano et al., 2025). Best-of-N samples from p_0 and retains the top-scoring candidates; classifier-guided methods modify the generation trajectory using an external conditioning signal; and tilted-distribution methods target a reward-weighted law $\pi_\beta(x) \propto p_0(x)\exp\{\beta\widehat{f}(x)\}$, where β controls steering strength.

Reward Model Over-Optimization. While these approaches can substantially shift generation toward high-scoring candidates, they share a common vulnerability: they implicitly treat $\widehat{f}$ as a faithful surrogate for the desired objective. This assumption becomes increasingly fragile as steering intensifies and the induced distribution diverges from the support of the reward model's training data. The problem is especially pronounced when $\widehat{f}$ is a single point estimate trained on finite data from q_{train} that covers only a fraction of the broader biological population p_0. As steering pushes samples into poorly covered regions, the point estimate ignores epistemic uncertainty and can confidently assign high rewards to under-characterized mutations, causing the optimizer to exploit spurious peaks that would not be supported by alternative models trained on different subsamples. In protein engineering, training data such as deep mutational scans are often noisy and unevenly distributed across sequence space, leaving some regions well characterized and others poorly sampled. As a result, the generator may exploit estimation errors and concentrate probability mass on candidates that receive high predicted reward but are biologically implausible or experimentally inactive.

Moreover, high reward under an *in silico* primary reward does not guarantee high utility under the true biological objective. A predictor trained solely on specific assay data may assign high scores to candidates that ultimately fail critical downstream validations such as structural stability or experimental viability. The resulting designs may therefore optimize modeling artifacts rather than genuine biological function, a phenomenon known as reward hacking. This motivates uncertainty-aware steering: rather than trusting a single reward estimate unconditionally, we quantify signal reliability for each candidate, downweight unstable regions, and align the search with sparse high-fidelity feedback.

1.3. Decomposing Uncertainty in Inference-Time Steering

Reliable inference-time steering depends not only on predicted reward magnitude, but also on the trustworthiness of the steering signal and its validity when extrapolating beyond characterized biological regimes. We decompose steering uncertainty into three components: reward-model stochasticity, fidelity-alignment uncertainty, and base-support reliability. The first two components are primarily epistemic, reflecting reducible uncertainty about poorly characterized regions of the protein fitness landscape and about whether inexpensive proxy measurements remain aligned with true experimental viability. The third component captures distributional uncertainty, assessing whether p_0 assigns sufficient probability to a candidate for it to be considered biologically plausible.

Reward-Model Uncertainty. The first source of uncertainty arises from the learned primary model $\widehat{f}$. Whether $\widehat{f}$ is trained on targeted high-throughput assays or large-scale biological databases (e.g., UniProt or the PDB), its predictions are inherently constrained by the biases and coverage limitations of the underlying data. Due to uneven evolutionary sampling and experimental coverage, predictions in underrepresented sequence regions become highly sensitive to assay noise and stochastic elements of the training pipeline, including data subsampling, random initialization, and optimization noise. This variability is especially pronounced in regions with limited training support, where small perturbations to the training data or training procedure can yield substantially different extrapolations. Consequently, a novel protein variant predicted to be highly functional by one fitted model may score poorly under other equally plausible models trained on slightly different data partitions. This disagreement reflects genuine epistemic uncertainty in the steering signal when navigating beyond well-characterized biological regimes.

Fidelity-Alignment Uncertainty. The second source of uncertainty characterizes the alignment between a high predicted score under the inexpensive primary R_1 and true biological utility under a sparse, high-fidelity evaluator R_2. Because gold-standard biological validation, such as measuring target binding, enzymatic kinetics, or resolving structures by cryo-EM, is highly resource-intensive, R_2 can be observed for only a small fraction of generated candidates. This creates substantial uncertainty about whether sequence regions favored by a computational predictor or noisy high-throughput screen remain viable under real-world biological criteria. This issue becomes critical when steering pushes the generative model toward high-R_1 regions that lack high-fidelity feedback. In these unverified regimes, the mutations driving the primary score may actively conflict with the true scientific objective measured by R_2. For instance, a predictor might artificially inflate an *in silico* binding score by introducing hydrophobic residues at the binding interface that cause the protein to aggregate or misfold during experimental expression.

Base-Support Reliability. The third component captures whether steered candidates remain sufficiently supported by the base generative distribution p_0, which represents the landscape of evolutionarily and biophysically plausible proteins. When steering pushes candidates into low-support regions of p_0 such as unnatural sequence motifs or sterically clashing residue pairs, high predicted fitness likely reflects algorithmic exploitation of the reward model rather than discovery of a realistically viable protein.

Accordingly, reliable biological sequence steering requires tracking three complementary quantities: the stability of the primary model against training-data perturbations, the alignment between inexpensive primary scores and high-fidelity evaluation, and the evolutionary and biophysical consistency with the base generative model.

2. Fidelity-Concordance Steering

We propose Fidelity-Concordance Steering (FiCS), an uncertainty-aware reward-calibration framework for generative design under sparse high-fidelity feedback. Our method addresses two distinct yet interrelated challenges. First, the fitted primary reward $\widehat{f}$ may exhibit instability in regions inadequately represented in its training distribution. Second, R_1 might be misaligned with an expensive oracle evaluator R_2 that assesses critical properties such as feasibility or experimental viability but can be queried only sparingly. Given a base distribution p_0, FiCS constructs a reliability-adjusted reward score that prioritizes candidates with high predicted R_1 while systematically avoiding regions where the primary reward exhibits instability or demonstrates poor concordance with R_2. Because FiCS outputs a calibrated reward score S_{FiCS} rather than a new generator, it can be used inside best-of-N selection, reranking, SMC-style test-time alignment, or other reward-guided pipelines. The reward ensemble is trained once offline, after which FiCS requires neither generator retraining nor additional oracle queries beyond the calibration stage. The algorithm pseudocode is provided in Algorithm 1 and the full algorithm is provided in Appendix A. At a high level, the method proceeds in three stages:

1. **Resampling-based reward ensemble:** train a bootstrap ensemble of reward guides $\widehat{f}_1, \ldots, \widehat{f}_K$ on repeated subsamples of the reward-training data, using data perturbations and refitting to assess whether the induced high-reward regions remain stable under plausible variations of the training observations.
2. **Sparse high-fidelity calibration:** construct a small calibration set and evaluate it with the high-fidelity oracle R_2. These evaluations are used to reweight the reward guides according to their high-fidelity alignment, assigning larger weights to guides whose top-ranked candidates remain favorable under R_2.
3. **Pessimistic fidelity-concordance scoring:** score new candidates using a fidelity-aligned lower confidence bound that combines the weighted ensemble mean with a penalty for residual guide disagreement.

2.1. Resampling-Based Reward Construction for Shift Sensitivity

The first step of FiCS is to construct a family of fitted primary rewards that quantifies the sensitivity of the steering signal to finite-sample variation in the reward-training data. This sensitivity is critical because steering changes the evaluation regime. Instead of scoring typical samples from Q_{train}, the reward is used to rank and select candidates in the high-score tail, where training support may be limited and fitted rewards may depend disproportionately on a few observations. Because the distribution of candidates selected by steering is unknown in advance, we use random reweighting of the reward-training data to probe how sensitive the fitted reward landscape is to plausible perturbations of its empirical support.

Let W be a nonnegative weighting process satisfying $\mathbb{E}_{Q_{\text{train}}}[W(X,Y)] = 1$, and define $dQ^{\star}_{W}(x,y) = W(x,y)\, dQ_{\text{train}}(x,y)$. The corresponding target risk is $\mathcal{R}_{Q^{\star}_{W}}(f) = \mathbb{E}_{Q_{\text{train}}}[W(X,Y)\ell(f(X),Y)]$. Thus, each draw of W defines a plausible shifted reward-training environment, so disagreement across resampled reward guides captures sensitivity under random perturbations of the data-generating process rather than a single specified adversarial shift (Jeong & Rothenhäusler, 2025).

At the sample level, this random-shift model is implemented by drawing nonnegative sample weights and fitting

Algorithm 1: FiCS: Fidelity-Concordance Steering

Input: Training data $\mathcal{D}_1$, base distribution p_0, oracle R_2, ensemble size K, oracle budget b, threshold c, hyperparameters $\alpha, \gamma, \lambda, \eta$

Output: Steering score $S_{\mathsf{FiCS}}(\cdot)$

```
/* Stage 1: Resampled reward ensemble */
```

for $k = 1, \ldots, K$ **do**

  Draw weights $w^{(k)}$; fit $\widehat{f}_k$ by weighted ERM on $\mathcal{D}_1$;

end

```
/* Stage 2: Sparse oracle calibration */
```

Draw pool from p_0 and select b candidates to form the concordance set;

Evaluate R_2 on concordance set to obtain $\mathcal{D}_2 = \{(z_j, R_2(z_j))\}_{j=1}^{b}$;

for $k = 1, \ldots, K$ **do**

  $A_k \leftarrow$ feasibility rate of R_2 among top-α of $\mathcal{D}_2$ under $\widehat{f}_k$;

end

$\varphi_k \leftarrow \exp(\gamma A_k) \big/ \sum_\ell \exp(\gamma A_\ell)$;

```
/* Stage 3: Pessimistic concordance scoring */
```

for *each candidate* x **do**

  $S_{\mathsf{FiCS}}(x) \leftarrow \underbrace{\bar{f}_\varphi(x)}_{\text{aligned mean}} - \underbrace{\lambda\, \hat{\sigma}_\varphi(x)}_{\text{uncertainty}} - \underbrace{\eta\,(-\log p_0(x))}_{\text{support}}$;

end

return $S_{\mathsf{FiCS}}(\cdot)$

the corresponding weighted empirical risk minimizer. Let $\mathcal{D}_1 = \{(x_i, y_i)\}_{i=1}^{n}$ denote the training set for the inexpensive reward R_1. For each ensemble member k, we draw weights $w^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)})$ satisfying $\sum_i w_i^{(k)} = 1$, and fit

$$\widehat{f}_k \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^{n} w_i^{(k)} \ell(f(x_i), y_i).$$

Random reweighting provides a local perturbation model around the observed reward-training data. In well-supported regions, refits should produce similar reward landscapes and stable steering directions; in under-covered or distribution-shifted regions, small data perturbations may yield distinct extrapolations and high-reward regions. As such, ensemble disagreement is interpreted as a sensitivity diagnostic for steering-time vulnerability, rather than merely as classical sampling uncertainty around $\widehat{Q}_n$. We use the ordinary nonparametric bootstrap in the experiments, although the framework accommodates other nonnegative weighting schemes; see Appendix B.

2.2. Aligning Steering Signal with Oracle Feedback

In many scientific design problems, the inexpensive reward R_1 is used primarily to identify and prioritize promising candidates rather than to determine their ultimate acceptability. Final validity is instead assessed by a more reliable but expensive oracle evaluator R_2, such as experimental activity, structural plausibility, and solubility. Because evaluating R_2 for every generated candidate at each iteration is often computationally prohibitive (and in some settings practically infeasible), it cannot serve as a dense reward signal throughout the steering procedure. Instead, R_2 provides sparse but high-fidelity feedback on downstream feasibility.

Accordingly, we formalize the optimization problem as

$$\max_x \quad R_1(x) \qquad \text{subject to} \quad R_2(x) \geq c,$$

where c is a feasibility threshold determined based on prior biological knowledge or adaptively using empirical quantiles. FiCS allocates a budget of $b \ll n$ oracle queries to construct a steering-aware concordance set. Specifically, we first draw a large candidate pool from the base distribution p_0 and rank candidates using the ensemble mean $\bar{f}$ of the resampled primary reward models $\{\widehat{f}_k\}_{k=1}^{K}$. We then select a mixture of top-ranked candidates, which target the high-reward regions relevant to steering, and unfiltered samples from p_0, which provide background coverage of the generation space. Evaluating R_2 on these selected candidates yields the sparse concordance set $\mathcal{D}_2 = \{(z_j, R_2(z_j))\}_{j=1}^{b}$. This concordance set allows us to identify which fitted primary steering models remain most consistent with oracle feasibility in the regions of the candidate space most relevant to steering.

Aligning Primary Rewards by Oracle Concordance. For each fitted primary reward $\widehat{f}_k$, we assess whether the candidates it ranks most highly remain feasible under the expensive evaluator. Let $\mathrm{Top}_\alpha(\widehat{f}_k; \mathcal{D}_2)$ denote the top α-fraction of concordance candidates according to $\widehat{f}_k$. We then define the oracle-concordance score as

$$A_k = \frac{1}{|\mathrm{Top}_\alpha(\widehat{f}_k; \mathcal{D}_2)|} \sum_{z_j \in \mathrm{Top}_\alpha(\widehat{f}_k; \mathcal{D}_2)} \mathbb{1}\{R_2(z_j) \geq c\},$$

which estimates the feasibility rate among the candidates preferred by $\widehat{f}_k$. A fitted primary reward receives a high concordance score when its top-ranked candidates consistently satisfy the oracle constraint, and a low score when its preferred candidates concentrate in regions where R_2 falls below the feasibility threshold. By restricting attention to the top-α tail rather than measuring global association with R_2, this criterion focuses on precisely the candidates that would be selected in practice if $\widehat{f}_k$ were used for steering.

Oracle-Weighted Reward Ensemble. FiCS then converts the oracle-concordance scores into nonnegative ensem-

ble weights,

$$\varphi_k = \frac{\exp(\gamma A_k - \max_m \gamma A_m)}{\sum_{\ell=1}^{K} \exp(\gamma A_\ell - \max_m \gamma A_m)},$$

where $\gamma \geq 0$ controls the strength of oracle concordance. When $\gamma = 0$, all guides receive equal weight. As γ increases, the ensemble increasingly concentrates on reward guides whose top-ranked candidates align with the expensive evaluator.

The resulting oracle-aligned ensemble scores are

$$\bar{f}_\varphi(x) = \sum_{k=1}^{K} \varphi_k \widehat{f}_k(x),$$

$$\hat{\sigma}_\varphi(x) = \left(\sum_{k=1}^{K} \varphi_k \big(\widehat{f}_k(x) - \bar{f}_\varphi(x)\big)^2 \right)^{1/2}.$$

This weighting scheme shifts the steering signal toward fitted primary rewards that rank oracle-feasible candidates highly, while preserving disagreement among aligned rewards as an uncertainty estimate.

2.3. Uncertainty-Aware Steering via Fidelity-Concordance Scoring

After constructing the oracle-aligned ensemble, we define the pessimistic fidelity-concordance score

$$S_{\mathsf{FiCS}}(x) = \bar{f}_\varphi(x) - \lambda\, \hat{\sigma}_\varphi(x),$$

where $\lambda \geq 0$ controls the strength of the uncertainty penalty. The first term favors candidates with high oracle-aligned primary reward, while the second penalizes high-uncertainty candidates. Accordingly, FiCS prioritizes candidates that are both highly scored and stable across the oracle-aligned primary rewards. When FiCS is used as a reranking or best-of-N selection method, we may additionally include an explicit support penalty:

$$S_{\mathsf{FiCS}}(x) = \bar{f}_\varphi(x) - \lambda\, \hat{\sigma}_\varphi(x) - \eta\,(-\log p_0(x)).$$

This term discourages the selection of candidates that receive high primary reward but lie in low-density regions of the base model. The support penalty is most relevant for best-of-N or reranking procedures, where selection can concentrate on low-probability regions of p_0. For distribution-based steering methods that already constrain the sampling law to remain close to p_0, such as KL-regularized sampling or classifier-free guidance, this penalty may be unnecessary; we discuss this variant further in Appendix **??**.

Final candidates are selected by ranking samples according to S_{FiCS}. After the concordance stage, FiCS requires no additional oracle evaluations: R_2 is used only to determine the ensemble weights, and all subsequent candidate scoring is performed using the aligned primary rewards.

3. Theoretical Analysis

We analyze the theoretical properties of FiCS. Our first result shows that the oracle-concordance scores used to weight the resampled reward guides concentrate around their population counterparts under the concordance-set sampling rule. Thus, sparse high-fidelity feedback suffices to estimate each guide's population concordance under the calibration design. If two reward guides differ in their true oracle concordance by more than the finite-sample estimation error, then their empirical concordance scores preserve the correct ordering. The theory formalizes the statistical claim underlying FiCS, showing that a modest amount of sparse oracle feedback can stably identify which reward guides remain concordant with the oracle in the steering-relevant tail, and that the resulting concordance errors propagate stably to the final FiCS score. Full statements and proofs are deferred to Appendix F.

Theorem 3.1 (Concentration of oracle-concordance scores)**.** *Let $\mathcal{D}_2 = \{(z_j, R_2(z_j))\}_{j=1}^{b}$ denote the concordance set and let $n_\alpha = \lfloor \alpha b \rfloor$. For each guide $\widehat{f}_k$, let A_k denote the empirical feasibility rate among the top-α candidates in $\mathcal{D}_2$, and let*

$$A_k^* = \mathbb{P}(R_2(Z) \geq c \mid Z \in \mathrm{Top}_\alpha(\widehat{f}_k)), \qquad Z \sim P_{\mathrm{conc}},$$

denote the corresponding population concordance under the concordance-set sampling rule, where $\mathrm{Top}_\alpha(\widehat{f}_k)$ denotes the population top-α tail of $\widehat{f}_k$ under P_{conc}. Under standard regularity assumptions made precise in the appendix, for any $\delta \in (0,1)$, with probability at least $1-\delta$,

$$\max_{k=1,\dots,K} |A_k - A_k^*| \leq \sqrt{\frac{\log(4K/\delta)}{2n_\alpha}} + \frac{C}{\alpha}\sqrt{\frac{\log(Kb/\delta)}{b}},$$

for a universal constant $C > 0$.

The next result shows that finite-sample error in the oracle-concordance scores propagates stably to the final FiCS score. In particular, the inverse temperature parameter γ controls sensitivity to concordance-estimation error, while the uncertainty penalty λ controls how strongly weight uncertainty affects the disagreement term. A full proof is given in Theorem F.4 in the appendix.

Theorem 3.2 (Uniform score stability)**.** *Let S^* be the FiCS score computed with population concordance weights $\varphi^* = \mathrm{softmax}(\gamma A^*)$, and let $\hat{S}$ be the score computed with empirical weights $\hat{\varphi} = \mathrm{softmax}(\gamma \hat{A})$. Assume $\lambda \geq 0$, $\max_k \sup_x |\widehat{f}_k(x)| \leq B$, and $\|\hat{A} - A^*\|_\infty \leq \varepsilon_A$ on an event with probability at least $1-\delta$. Let $\rho = \min\{2, 2\gamma\varepsilon_A\}$. On this event,*

$$\sup_x |S^*(x) - \hat{S}(x)| \leq B\rho + \lambda B\sqrt{3\rho}.$$

As a direct consequence of Theorem 3.2, if the population-concordance FiCS score has a sufficiently large gap at the top-L selection boundary, then empirical FiCS yields the

same selected candidates as the oracle-optimal FiCS score; the formal statement is given in Theorem F.6.

4. Empirical Evaluation

We first evaluate FiCS on a controlled synthetic benchmark on $\mathbb{R}^2$ where the point reward is attractive precisely because it is wrong. The benchmark isolates two failure modes that motivate FiCS, namely instability of the learned primary reward in undersampled regions and misalignment between the primary reward and a high-fidelity oracle. The environment comprises three Gaussian mixture components, each representing a distinct reward regime. Region A is desirable, exhibiting both high true primary reward $R_1^\star$ and high oracle reward $R_2^\star$. Region B is deceptive, with moderately high $R_1^\star$ but low $R_2^\star$; it appears attractive under the inexpensive reward yet fails the oracle criterion. Region C is safe, offering moderate $R_1^\star$ alongside high $R_2^\star$. To induce misspecification, we corrupt the observed training labels with noise and systematically inflate values near region B, causing fitted reward guides to overvalue the deceptive region.

The base distribution p_0 and reward-training distribution q_{train} are chosen so that region B is substantially underrepresented in training relative to generation. We draw labeled samples from q_{train} and train a bootstrap ensemble of K kernel ridge regression guides to quantify reward uncertainty and enable oracle calibration. The oracle reward $R_2^\star$ is evaluated directly from the data-generating mechanism rather than fitted, serving as an expensive ground-truth oracle. We use a limited budget of oracle evaluations on candidates drawn from p_0 to calibrate the ensemble, then assess candidate selection on a large test pool from p_0. Additional details on the simulation setup are provided in Appendix D.

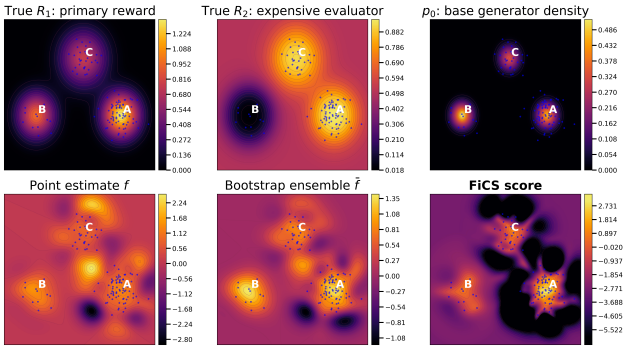

Figure 1. Learned scoring landscapes for a representative seed. **Top:** Ground-truth rewards $R_1^\star$, $R_2^\star$, and base density p_0; blue dots show the $n = 120$ training points undersampling region B. **Bottom:** Point estimate and ensemble mean score the deceptive region B highly, whereas FiCS correctly concentrates on desirable region A.

We compare four methods: a point-estimate reward model, the unweighted bootstrap ensemble mean, and FiCS with

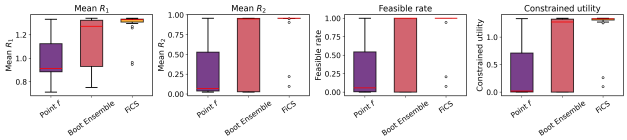

Figure 2. Performance across 30 seeds: mean $R_1^\star$, mean $R_2^\star$, feasibility rate, and constrained utility. The point estimate and bootstrap ensemble show high variance and frequent infeasibility. FiCS achieves near-perfect feasibility and the highest constrained utility with minimal variance.

Table 1. Simulation results across 30 random seeds. Values shown as mean (standard deviation). Blue indicates best performance; green indicates second-best.

Method	Mean $R_1^\star$	Mean $R_2^\star$	Feasible rate	Constr. utility
Point Est.	1.008 (.214)	0.370 (.418)	0.378 (.449)	0.457 (.606)
Bootstrapped Mean	1.191 (.192)	0.626 (.444)	0.646 (.476)	0.861 (.635)
FiCS (dense weights)	1.286 (.126)	0.863 (.295)	0.900 (.316)	1.193 (.419)
FiCS (row bootstrap)	**1.321** (.025)	**0.951** (.017)	**0.994** (.018)	**1.315** (.032)

two resampling schemes: (i) row bootstrap (standard nonparametric bootstrap) and (ii) dense weights (continuous Dirichlet-distributed perturbations). Across 30 random seeds, we evaluate four metrics: (i) mean $R_1^\star$, the average true primary reward among selected candidates; (ii) mean $R_2^\star$, the average oracle reward; (iii) feasibility rate, the fraction satisfying $R_2^\star \geq c$; and (iv) constrained utility, $R_1^\star(x)\mathbb{1}\{R_2^\star(x) \geq c\}$ averaged over selected candidates, which rewards high primary performance only among oracle-feasible selections.

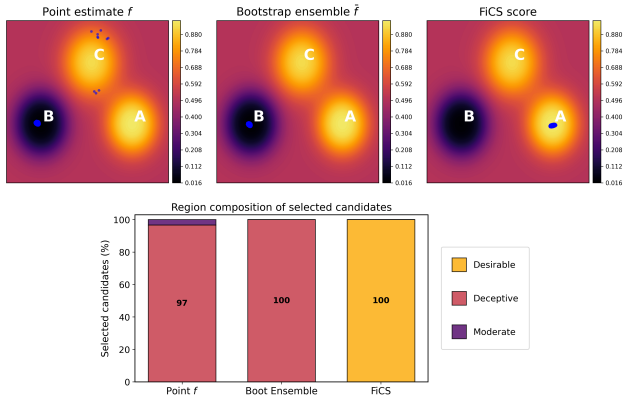

Figure 3. **Top:** Selected candidates (blue dots) by each steering function overlaid on the true oracle landscape $R_2^\star$ for a representative seed. Under the point estimate and bootstrap ensemble mean, most selected candidates fall in the deceptive region B where $R_2^\star$ is near zero. FiCS redirects all selections to region A, the only region with both high $R_1^\star$ and high $R_2^\star$. **Bottom:** Region composition of the top-L selected candidates. The point estimate selects almost entirely from region B (97%), the bootstrap ensemble mean selects 100% from region B, and FiCS selects 100% from the desirable region A.

Table 1 and Figure 2 summarize performance across all 30

seeds. Both baselines exhibit high variance and weaker oracle alignment, with a substantial fraction of selections falling below the feasibility threshold. In contrast, FiCS achieves consistently high values with minimal dispersion. The contrast is most pronounced in feasibility rate and constrained utility: both baselines frequently yield entirely infeasible selections, while FiCS under row bootstrap maintains near-perfect feasibility across all seeds. The constrained utility, which captures the joint desideratum of high $R_1^\star$ among oracle-feasible candidates, improves by approximately $3\times$ over the point estimate and $1.5\times$ over the bootstrap ensemble mean under FiCS with row bootstrap. This shows that oracle-concordance reweighting improves both feasibility and constrained utility.

Notably, the unweighted bootstrap ensemble mean provides only modest improvement over the point estimate (feasibility rate 0.646 vs. 0.378), confirming that averaging reward landscapes is insufficient when all ensemble members inherit the same systematic coverage gap. The oracle-concordance reweighting in FiCS is the decisive mechanism: it identifies which guides produce feasible top candidates and concentrates weight accordingly.

Figure 3 provides geometric intuition for a representative seed. Under both baselines, selected candidates concentrate in region B, where the learned reward overestimates due to training bias and sparse coverage. FiCS redirects selections entirely to region A, the only mode with jointly high $R_1^\star$ and $R_2^\star$. The region composition bar chart confirms this pattern is systematic rather than seed-specific.

5. Renin: Steering Toward Cytosolic Localization Under a Limited Reward Assay

In this section, we demonstrate the empirical performance of FiCS on a biologically motivated protein engineering task. Designing protein variants that remain functional across drastically different subcellular environments is a central challenge in protein engineering. Human renin illustrates this difficulty: it is a highly specific aspartic protease whose native fold requires disulfide bonds formed in the oxidizing secretory pathway and typically fails when expressed in the reducing cytosol. A cytosol-compatible renin variant would broaden the synthetic biology toolbox by providing an orthogonal, highly specific protease for intracellular use, but obtaining one requires steering generative models away from the evolutionary regime in which the wild-type was selected. Under a localization classifier trained on UniProt subcellular annotations, wild-type renin has a predicted cytosolic probability of only 0.036, so cytosolic steering must push generation far into the tails of p_0, precisely where finite-sample reward models are least reliable.

To study FiCS in this regime under controlled conditions, we construct R_1 and R_2 from the same architecture and prediction target but with different training-set coverage: R_1 is trained on n labeled examples drawn from a narrow region of sequence space, whereas R_2 is trained on the full UniProt localization dataset. This setup mirrors the common scenario in which a cheap, low-coverage predictor screens broadly while a high-fidelity oracle, limited by computational or wet-lab cost, is queried only sparsely for calibration. Although R_2 is not itself an experimental measurement, it serves as a surrogate for the assay feedback that would be available in practice. In typical FiCS deployments, R_2 would carry additional information beyond R_1 (e.g., AlphaFold structural confidence, experimental expression or activity readouts, or composite metrics combining multiple sources). This experiment represents a realistic low-oracle regime in which the available high-fidelity signal is too sparse to support reliable generator retraining, yet sufficiently informative to calibrate which reward guides remain trustworthy for candidate prioritization.

5.1. *In Silico* Experiment Setup

We use ProteinMPNN (Dauparas et al., 2022) as the base distribution p_0, conditioning on the AlphaFold 3 predicted structure of mature renin, and steer via best-of-N selection. Following ProVADA (Lu et al., 2025; Viggiano et al., 2025), we embed each sequence with ESM-2 (Lin et al., 2023) and use a dropout linear classifier trained on UniProt annotations to predict subcellular localization; the cytosolic-class probability serves as the steering target. R_2 is trained on the full UniProt localization dataset, while R_1 shares the same architecture but is trained on a random subsample of $n = 200$ sequences. The point-estimate baseline uses a single such classifier; the bootstrap-ensemble baseline and FiCS both employ an ensemble of $K = 30$ classifiers fit via nonparametric bootstrap on the same subsample. FiCS uses an oracle-calibration budget of $b = 30$ (75% sampled from the top under the ensemble mean, 25% sampled uniformly), with per-model top $b_k = 5$ and $\gamma = 10$. We generate $N = 250$ candidates and select the top L under each method.

5.2. Results

Table 2 reports performance on the top-L candidates. At $L = 5$, FiCS improves mean oracle cytosolic probability by 29% over the point estimate and 46% over the bootstrap ensemble, while achieving perfect feasibility against baseline rates of 0.600 and 0.400. At $L = 30$ the gap narrows but remains substantial. Notably, the unweighted bootstrap ensemble offers negligible improvement over the point estimate: averaging reward landscapes is insufficient when all members inherit the same coverage limitations. FiCS's gains are largest at small L, where reliable selection matters most. Figure 4 (top) confirms this across batch sizes: FiCS

Table 2. Oracle evaluation performance by selection size L. Blue indicates best method for each batch size.

Size	Method	Mean R_2	Median R_2	Feasible rate
$L = 5$	FiCS	**0.155**	**0.139**	**1.000**
	Point Est.	0.120	0.126	0.600
	Bootstrapped Mean	0.106	0.092	0.400
$L = 30$	FiCS	**0.124**	**0.118**	**0.700**
	Point Est.	0.114	0.113	0.600
	Bootstrapped Mean	0.111	0.114	0.633

maintains higher mean oracle R_2 and the highest feasibility rate across the evaluated selection sizes, with perfect feasibility in the small-batch $L = 5$ setting.

To rule out the possibility that FiCS merely re-ranks similar pools, Figure 4 (middle) partitions selections into those unique to FiCS, unique to a baseline, or shared. The true R_2 distribution for FiCS-only selections lies clearly above both baseline-only distributions, indicating that FiCS systematically identifies higher-quality candidates rather than marginal cases. Figure 4 (bottom) plots each method's selection score against true oracle R_2. The Pearson correlation is near zero for both baselines but reaches 0.362 for FiCS, showing that oracle-concordance reweighting converts a near-uninformative steering signal into one whose ordering tracks oracle quality. Additional results at $L = 30$ are provided in Appendix E.

6. Discussion

We introduce Fidelity-Concordance Steering (FiCS), a reward-calibration framework for reliable generative design under sparse high-fidelity feedback. The method addresses a pervasive failure mode in reward-guided generation: when steering concentrates candidates in regions where a finite-sample surrogate must extrapolate, high predicted scores often reflect estimation artifacts rather than genuine biological utility. Unlike test-time alignment methods that assume a trusted reward and limit over-optimization by keeping the steered distribution close to the pretrained generator (Kim et al., 2025), FiCS addresses the upstream problem of reward unreliability itself. It converts an unstable or biased surrogate into a reliability-adjusted signal before optimization amplifies its errors by integrating resampling-based uncertainty quantification with oracle-concordance calibration. Our empirical results demonstrate that FiCS yields substantial improvements in oracle feasibility and constrained utility, particularly in the small-batch regime where reliable selection is most critical. More broadly, our findings indicate that robust steering must jointly address two fundamental challenges: the epistemic instability of finite-sample surrogates and their imperfect alignment with the true scientific objective.

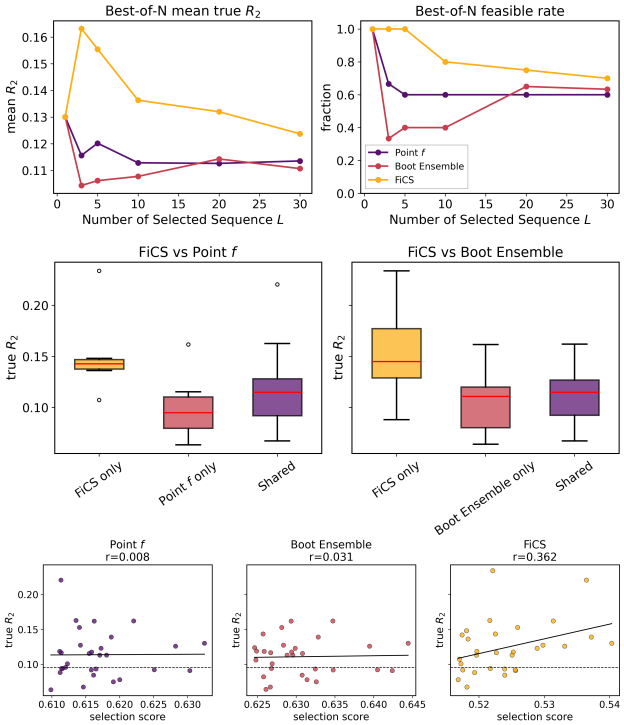

Figure 4. Renin steering results. **Top:** Best-of-N mean oracle R_2 and feasibility rate as a function of selection size L. FiCS maintains consistently higher mean oracle R_2 and the highest feasibility rate across the evaluated selection sizes; at $L = 5$ it improves mean R_2 by 29% over the point estimate and 46% over the bootstrap ensemble while achieving perfect feasibility. **Middle:** True R_2 distributions for candidates selected exclusively by each method versus shared selections. Candidates unique to FiCS consistently achieve higher oracle scores. **Bottom:** Selection score versus true R_2. Only FiCS exhibits meaningful correlation ($r = 0.362$) between its scoring function and oracle quality.

Looking ahead, FiCS should be evaluated in prospective design campaigns where the high-fidelity signal is a genuine experimental readout, such as expression, solubility, binding, or activity. By extracting reliable steering signal from a biased, inexpensive primary assay and correcting it with sparse high-fidelity feedback, FiCS could ease two experimental bottlenecks at once: reducing the need for large, evenly sampled training libraries and supporting objectives whose faithful evaluation is low-throughput or resource-intensive. In such campaigns, the concordance set could be refreshed as new labels arrive, extending the one-shot calibration studied here into an iterative procedure.

Acknowledgements

W.S.L. gratefully acknowledges support from NIH grant GM 134483. X.Z. is supported by the Stanford Interdisciplinary Graduate Fellowship affiliated with ChEM-H. D.R. gratefully acknowledges support as a David Huntington Faculty Scholar, Chamber Fellow, and from the Dieter Schwarz

Foundation. The authors gratefully acknowledge support from the Google Cloud Research Credits Program and the HAI Cloud Credits Grant.

Impact Statement

This paper presents work whose goal is to advance the field of Machine Learning. There are many potential societal consequences of our work, none of which we feel must be specifically highlighted here.

References

Biswas, S., Khimulya, G., Alley, E. C., Esvelt, K. M., and Church, G. M. Low-n protein engineering with data-efficient deep learning. *Nature Methods*, 18(4): 389–396, April 2021. ISSN 1548-7105. doi: 10.1038/s41592-021-01100-y. URL http://dx.doi.org/10.1038/s41592-021-01100-y.

Dauparas, J., Anishchenko, I., Bennett, N., Bai, H., Ragotte, R. J., Milles, L. F., Wicky, B. I. M., Courbet, A., de Haas, R. J., Bethel, N., Leung, P. J. Y., Huddy, T. F., Pellock, S., Tischer, D., Chan, F., Koepnick, B., Nguyen, H., Kang, A., Sankaran, B., Bera, A. K., King, N. P., and Baker, D. Robust deep learning–based protein sequence design using proteinmpnn. *Science*, 378(6615):49–56, October 2022. ISSN 1095-9203. doi: 10.1126/science.add2187. URL http://dx.doi.org/10.1126/science.add2187.

Dhariwal, P. and Nichol, A. Diffusion models beat gans on image synthesis. *Advances in neural information processing systems*, 34:8780–8794, 2021.

Jain, V., Sareen, K., Pedramfar, M., and Ravanbakhsh, S. Diffusion tree sampling: Scalable inference-time alignment of diffusion models. *arXiv preprint arXiv:2506.20701*, 2025.

Jeong, Y. and Rothenhäusler, D. Calibrated inference: statistical inference that accounts for both sampling uncertainty and distributional uncertainty. *Journal of Machine Learning Research*, 26(196):1–48, 2025.

Kim, S., Kim, M., and Park, D. Test-time alignment of diffusion models without reward over-optimization. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=vi3DjUhFVm.

Koh, H. Y., Zheng, Y., Yang, M., Arora, R., Webb, G. I., Pan, S., Li, L., and Church, G. M. Ai-driven protein design. *Nature Reviews Bioengineering*, 3:1034–1056, 2025.

Lin, Z., Akin, H., Rao, R., Hie, B., Zhu, Z., Lu, W., Smetanin, N., Verkuil, R., Kabeli, O., Shmueli, Y., dos Santos Costa, A., Fazel-Zarandi, M., Sercu, T., Candido, S., and Rives, A. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, March 2023. ISSN 1095-9203. doi: 10.1126/science.ade2574. URL http://dx.doi.org/10.1126/science.ade2574.

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 and Biology (GenBio) Workshop*, 2025.

Ramé, A., Vieillard, N., Hussenot, L., Dadashi, R., Cideron, G., Bachem, O., and Ferret, J. Warm: On the benefits of weight averaged reward models. *arXiv preprint arXiv:2401.12187*, 2024.

Stanton, S., Maddox, W., Gruver, N., Maffettone, P., Delaney, E., Greenside, P., and Wilson, A. G. Accelerating bayesian optimization for biological sequence design with denoising autoencoders, 2022. URL https://arxiv.org/abs/2203.12742.

Viggiano, B., Lu, W. S., Zhang, X., Mille-Fragoso, L. S., Gao, X. J., Ashley, E., and Wong, W. H. Steering protein generative models at test-time for guided aav2 capsid design. In *Biocomputing 2026: Proceedings of the Pacific Symposium*, pp. 438–451. World Scientific, 2025.

Wittmann, B. J., Yue, Y., and Arnold, F. H. Informed training set design enables efficient machine learning-assisted directed protein evolution. *Cell Systems*, 12(11): 1026–1045.e7, November 2021. ISSN 2405-4712. doi: 10.1016/j.cels.2021.07.008. URL http://dx.doi.org/10.1016/j.cels.2021.07.008.

Yang, J., Lal, R. G., Bowden, J. C., Astudillo, R., Hameedi, M. A., Kaur, S., Hill, M., Yue, Y., and Arnold, F. H. Active learning-assisted directed evolution. *Nature Communications*, 16(1), January 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-55987-8. URL http://dx.doi.org/10.1038/s41467-025-55987-8.

Yu, T., Thomas, G., Yu, L., Ermon, S., Zou, J. Y., Levine, S., Finn, C., and Ma, T. Mopo: Model-based offline policy optimization. *Advances in neural information processing systems*, 33:14129–14142, 2020.

A. Full Algorithmic Description of FiCS

Algorithm 2: FiCS: Fidelity-Concordance Steering

Input: Base distribution p_0, training data $\mathcal{D}_1 = \{(x_i, y_i)\}_{i=1}^n$, oracle R_2, ensemble size K, oracle budget b, feasibility threshold c, top-fraction α, inverse concordance temperature γ, uncertainty penalty λ, support penalty η, top-fraction for calibration ρ_{cal}, calibration pool size N_{cal}

Output: Fidelity-concordance steering score $S_{\mathsf{FiCS}}(\cdot)$ for inference-time guidance

/* Fit resampled primary reward models */

for $k = 1, \ldots, K$ **do**

  Draw resampling weights $w^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)})$;

  $\widehat{f}_k \leftarrow \arg\min_{f \in \mathcal{F}} \sum_{i=1}^n w_i^{(k)} \, \ell\big(f(x_i),\, y_i\big)$;

end

/* Construct steering-aware concordance set */

Draw candidate pool $\{u_1, \ldots, u_{N_{\text{cal}}}\} \sim p_0$;

Compute ensemble mean $\bar{f}(u) \leftarrow K^{-1} \sum_{k=1}^K \widehat{f}_k(u)$ for each u;

Select $\lfloor \rho_{\text{cal}}\, b \rfloor$ candidates with highest $\bar{f}$ and $b - \lfloor \rho_{\text{cal}}\, b \rfloor$ candidates uniformly at random;

Evaluate oracle on selected candidates to form $\mathcal{D}_2 = \{(z_j, R_2(z_j))\}_{j=1}^b$;

/* Compute oracle-concordance scores */

for $k = 1, \ldots, K$ **do**

  $\mathcal{T}_k \leftarrow \mathrm{Top}_\alpha(\widehat{f}_k;\, \mathcal{D}_2)$; // top-α fraction of $\mathcal{D}_2$ under $\widehat{f}_k$

  $A_k \leftarrow |\mathcal{T}_k|^{-1} \sum_{z_j \in \mathcal{T}_k} \mathbb{1}\{R_2(z_j) \geq c\}$;

end

/* Construct oracle-concordance ensemble */

$\varphi_k \leftarrow \exp(\gamma A_k) \big/ \sum_{\ell=1}^K \exp(\gamma A_\ell)$ for $k = 1, \ldots, K$;

/* Compute fidelity-concordance steering score */

for *each candidate* x **do**

  $\bar{f}_\varphi(x) \leftarrow \sum_{k=1}^K \varphi_k \, \widehat{f}_k(x)$;

  $\hat{\sigma}_\varphi(x) \leftarrow \Big(\sum_{k=1}^K \varphi_k \big(\widehat{f}_k(x) - \bar{f}_\varphi(x)\big)^2 \Big)^{1/2}$;

  $S_{\mathsf{FiCS}}(x) \leftarrow \bar{f}_\varphi(x) - \lambda\, \hat{\sigma}_\varphi(x) - \eta\, (-\log p_0(x))$;

end

return $S_{\mathsf{FiCS}}(\cdot)$

B. Alternative Random Reweighting Schemes

B.1. Distribution Shift

Distribution shift arises in steering because the primary reward is trained under Q_{train}, but is deployed on optimizer-selected candidates in high-reward regions that may be poorly represented in the reward-training data. Classical shift models posit a fixed target distribution $Q^\star$ and impose structural constraints on the likelihood ratio $r = dQ^\star / dQ_{\text{train}}$: covariate shift restricts r to depend only on x, label shift restricts it to depend only on y, bounded-likelihood-ratio models constrain $\|r\|_\infty$, and f-divergence models constrain $\mathbb{E}_{Q_{\text{train}}}[\phi(r)]$ for a convex function ϕ. In contrast, FiCS adopts a random-shift view in which the likelihood ratio is itself random, inducing a distribution over plausible reweightings of Q_{train}. The goal is therefore not to estimate a single shifted distribution, but to probe whether the steering-relevant reward landscape remains stable across plausible shifted environments.

B.2. Reweighting Schemes

The resampling ensemble in FiCS requires only a distribution over nonnegative weights $w^{(k)} = (w_1^{(k)}, \ldots, w_n^{(k)})$, with $\sum_{i=1}^n w_i^{(k)} = 1$, used to fit each primary reward model:

$$\widehat{f}_k \in \arg\min_{f \in \mathcal{F}} \sum_{i=1}^n w_i^{(k)} \ell(f(x_i), y_i).$$

If $\widehat{Q}_n = n^{-1} \sum_{i=1}^n \delta_{(x_i,y_i)}$ denotes the empirical reward-training distribution, each weight vector induces a perturbed empirical distribution $\widehat{Q}_n^{(k)} = \sum_{i=1}^n w_i^{(k)} \delta_{(x_i,y_i)}$. The resulting fitted rewards $\{\widehat{f}_k\}_{k=1}^K$ should be interpreted as a local sensitivity analysis around the observed reward-training data, rather than as a consistent estimator of the unknown steering-time environment. If the high-reward region induced by a guide is unstable under these empirical perturbations, then the guide should not be trusted for steering; conversely, agreement across refits indicates robustness to the particular training sample used to estimate R_1. Variability across the fitted guides therefore functions as a diagnostic for whether a high predicted reward reflects a stable signal or a finite-sample artifact: large variation indicates an unreliable steering direction, whereas agreement across refits supports the robustness of the reward landscape under local perturbations of the training distribution.

The ordinary nonparametric bootstrap draws multinomial counts $(N_1^{(k)}, \ldots, N_n^{(k)}) \sim \mathrm{Multinomial}\left(n; \frac{1}{n}, \ldots, \frac{1}{n}\right)$, where $N_i^{(k)}$ denotes the number of times example (x_i, y_i) appears in the k-th replicate, and sets $w_i^{(k)} = N_i^{(k)}/n$. This perturbs the empirical distribution by repeating some observations and omitting others. The m-out-of-n bootstrap provides a stronger perturbation model by drawing only $m < n$ samples with replacement:

$$(N_1^{(k)}, \ldots, N_n^{(k)}) \sim \mathrm{Multinomial}\left(m; \frac{1}{n}, \ldots, \frac{1}{n}\right), \qquad w_i^{(k)} = \frac{N_i^{(k)}}{m}.$$

Because each refit is trained under a smaller effective sample size, this scheme induces larger variation across fitted guides and can be used to stress-test whether high-reward regions remain stable under stronger perturbations of the reward-training environment. The ordinary bootstrap is recovered when $m = n$.

A Bayesian bootstrap instead draws continuous weights

$$w^{(k)} \sim \mathrm{Dirichlet}(1, \ldots, 1),$$

which smoothly reweights all observations and can be interpreted as posterior uncertainty over the empirical distribution. More generally, gamma multiplier weights draw and normalize

$$g_i^{(k)} \sim \mathrm{Gamma}(a, a), \qquad w_i^{(k)} = \frac{g_i^{(k)}}{\sum_{j=1}^n g_j^{(k)}},$$

where the shape parameter a controls the strength of the perturbation. Finally, covariate-dependent random weights allow the perturbation to depend on candidate features, for example

$$w_i^{(k)} \propto \exp\{\xi_k^\top \psi(x_i)\},$$

where $\psi(x_i)$ is a feature representation and ξ_k is randomly drawn. Such a weighting scheme emphasizes particular regions of the reward-training distribution rather than exchangeable observation-level noise. All of these choices lead to the same weighted empirical risk formulation, but encode different assumptions about plausible perturbations of the reward-training environment.

C. Variational Interpretation of Oracle-Concordance Weights

The oracle-concordance weights defined in Section 3 admit a natural variational interpretation that clarifies the role of the inverse temperature parameter γ and connects the weighting scheme to regularized optimization over the ensemble. Let Δ_K denote the probability simplex over the K guides, and let φ^0 be the prior weighting over guides before oracle calibration.

Proposition C.1 (Variational form of oracle-concordance weights)**.** *Assume $\gamma > 0$ and $\varphi_k^0 > 0$ for all k. The solution to*

$$\max_{\omega \in \Delta_K} \left\{ \sum_{k=1}^K \omega_k A_k - \frac{1}{\gamma} \mathrm{KL}(\omega \| \varphi^0) \right\}$$

is

$$\varphi_k = \frac{\varphi_k^0 \exp(\gamma A_k)}{\sum_{\ell=1}^{K} \varphi_\ell^0 \exp(\gamma A_\ell)}.$$

Proof. The objective is strictly concave on Δ_K because it is linear in ω minus a positive multiple of the KL divergence. Hence any interior stationary point is the unique maximizer. Introduce a Lagrange multiplier μ for the constraint $\sum_k \omega_k = 1$. The first-order condition is

$$A_k - \frac{1}{\gamma}\left(\log \frac{\omega_k}{\varphi_k^0} + 1\right) - \mu = 0.$$

Rearranging gives

$$\omega_k = \varphi_k^0 \exp\{\gamma A_k - 1 - \gamma\mu\}.$$

The factor $\exp\{-1 - \gamma\mu\}$ is common across k, so normalization over the simplex yields

$$\omega_k = \frac{\varphi_k^0 \exp(\gamma A_k)}{\sum_{\ell=1}^{K} \varphi_\ell^0 \exp(\gamma A_\ell)}.$$

Thus the calibrated weights satisfy $\varphi_k \propto \varphi_k^0 \exp(\gamma A_k)$, and strict concavity gives uniqueness. □

Remark C.2. In our default choice, $\varphi_k^0 = 1/K$, so this reduces to the softmax rule $\varphi_k \propto \exp(\gamma A_k)$. Thus, FiCS balances oracle concordance against deviation from the pre-oracle ensemble, favoring oracle-aligned guides while regularizing the weights when validation data are limited. As $\gamma \to 0$, the solution approaches the prior weights φ^0; as γ increases, the ensemble concentrates on guides with higher oracle-concordance scores.

D. Details on Simulation Experiment

We construct a synthetic environment on $\mathcal{X} = \mathbb{R}^2$ with three Gaussian mixture components defining distinct reward regions. Region A is desirable (high true primary reward $R_1^\star$ and high oracle reward $R_2^\star$); region B is deceptive (moderately high $R_1^\star$ but low $R_2^\star$); and region C is safe (moderate $R_1^\star$ and high $R_2^\star$). The true primary reward $R_1^\star$ is a sum of Gaussian bumps; observed training labels add Gaussian noise and a positive bias near region B, causing fitted primary reward guides to overestimate candidates in the deceptive region. The oracle reward $R_2^\star$ applies a sigmoid transformation to a latent score with a negative contribution near B, creating a sharp feasibility gap.

The base distribution p_0 is a Gaussian mixture with weights $(0.35, 0.40, 0.25)$ over regions A, B, and C, while the reward-training distribution q_{train} uses weights $(0.60, 0.05, 0.35)$, severely undersampling region B. We train a bootstrap ensemble of $K = 50$ kernel ridge regression guides on $n = 120$ labeled examples from q_{train}. We draw a calibration pool of 5,000 candidates from p_0 and select $b = 60$ for oracle evaluation: 75% are sampled from the top 15% of calibration candidates under the ensemble mean, and 25% are sampled uniformly from the remainder. Oracle-concordance scores A_k are computed on the top $\alpha = 20\%$ fraction with feasibility threshold $c = 0.55$, and oracle-calibrated weights use $\gamma = 8$.

For evaluation, we score a test pool of $N = 20{,}000$ candidates from p_0 and select the top $L = 20$. We compare a point-estimate reward model, the bootstrap ensemble mean, and full FiCS, reporting mean $R_1^\star$, mean $R_2^\star$, feasibility rate $(R_2^\star \geq c)$, and constrained utility across 30 random seeds.

E. Additional Results for Renin *In Silico* Experiment

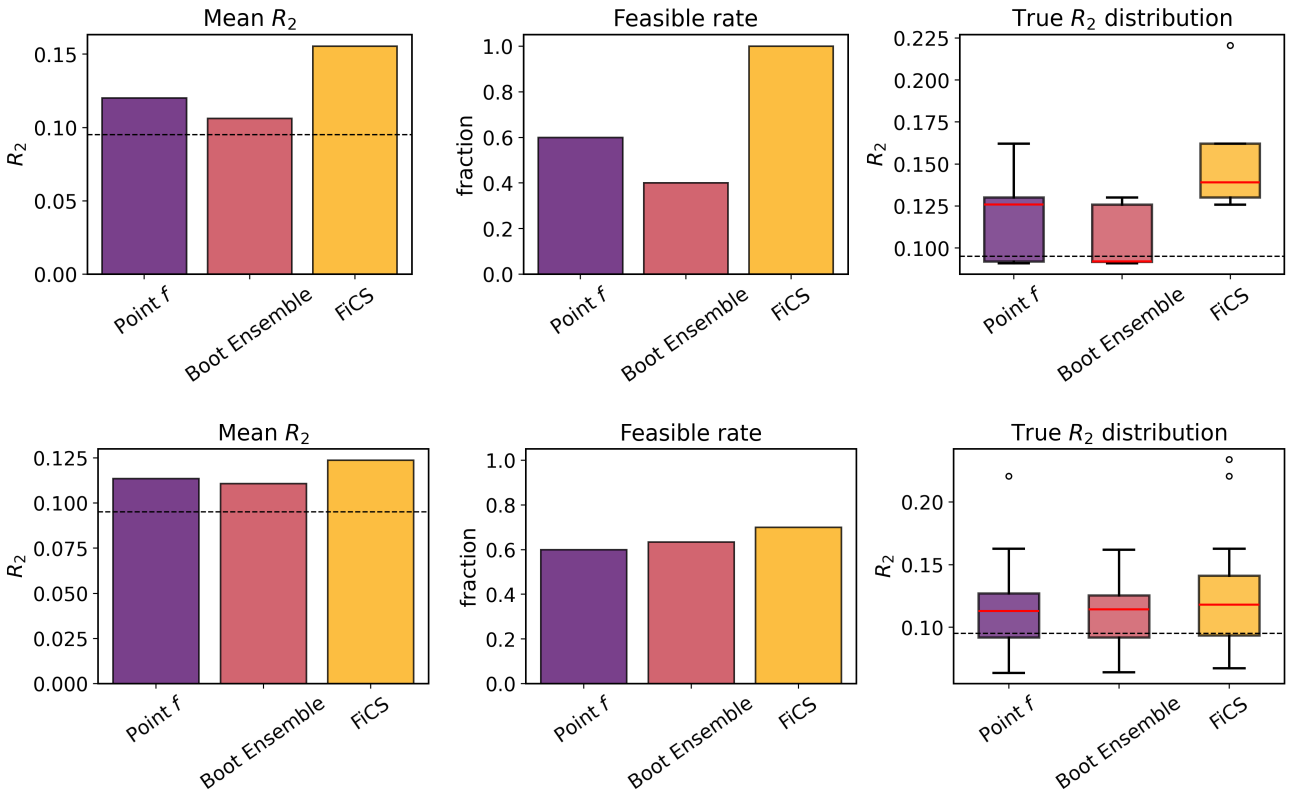

Figure 5. Oracle performance summary for the top-L selected sequences. **Top:** $L = 5$. **Bottom:** $L = 30$. Each row reports mean true R_2 (left), feasibility rate $R_2 \geq c$ (middle), and the distribution of true R_2 across selected candidates (right); the dashed line marks the feasibility threshold. At $L = 5$, FiCS achieves perfect feasibility and the highest mean R_2. The advantage persists at $L = 30$, where FiCS maintains the highest oracle score and feasibility rate among all methods.

F. Theoretical Results and Proofs

F.1. Finite-Sample Concentration of Oracle-Concordance Scores

The following result analyzes an idealized version of the concordance-set sampling rule. We assume that calibration inputs are sampled independently from the mixture distribution targeted by the top-ranked and random components of the concordance design. This idealization isolates the statistical concentration behavior of oracle-concordance scores from the finite-pool dependence effects that arise when sampling without replacement from a fixed proposal pool.

Theorem F.1 (Concentration of oracle-concordance scores)**.** *Let $\mathcal{D}_2 = \{(z_j, R_2(z_j))\}_{j=1}^{b}$ be the sparse concordance set from the main text, with candidate inputs $z_1, \ldots, z_b$ sampled independently from the idealized concordance-set sampling distribution P_{conc}. Fix the fitted guides $\{\widehat{f}_k\}_{k=1}^{K}$ and write $g(z) = \mathbb{P}(R_2(z) \geq c \mid z)$. Let $\widehat{\mathcal{T}}_k = \text{Top}_\alpha(\widehat{f}_k; \mathcal{D}_2)$ be the top $n_\alpha = \lfloor \alpha b \rfloor$ concordance candidates ranked by $\widehat{f}_k$. Define*

$$A_k = \frac{1}{n_\alpha} \sum_{z_j \in \widehat{\mathcal{T}}_k} \mathbb{1}\{R_2(z_j) \geq c\}.$$

Let τ_k be the population top-α threshold satisfying $\mathbb{P}(\widehat{f}_k(Z) \geq \tau_k) = \alpha$ for a population draw $Z \sim P_{\text{conc}}$, and define

$$A_k^* = \mathbb{E}[g(Z) \mid \widehat{f}_k(Z) \geq \tau_k].$$

Assume the following conditions hold:

(A1) $\alpha \in (0, 1)$ *and* $b \geq 2/\alpha$.

(A2) Oracle evaluations are conditionally independent given the concordance inputs and independent of the fitted guides.

(A3) The feasibility threshold c is either fixed before constructing $\mathcal{D}_2$ or estimated from data independent of $\mathcal{D}_2$.

(A4) For each k, the empirical top set $\widehat{\mathcal{T}}_k$ is almost surely representable as a threshold set $\{z_j : \widehat{f}_k(z_j) \geq \widehat{\tau}_k\}$ of cardinality n_α; for example, this holds if $\widehat{f}_k(Z)$ has a non-atomic distribution.

Additionally, assume the Dvoretzky–Kiefer–Wolfowitz and bounded Vapnik–Chervonenkis concentration inequalities hold for the threshold classes $\{z : \widehat{f}_k(z) \geq t\}$ and the bounded weighted threshold classes $\{z \mapsto g(z)\mathbb{1}\{\widehat{f}_k(z) \geq t\} : t \in \mathbb{R}\}$. Then, for any $\delta \in (0,1)$, with probability at least $1-\delta$ over the concordance inputs and oracle evaluations,

$$\max_{k=1,\dots,K} |A_k - A_k^*| \leq \sqrt{\frac{\log(4K/\delta)}{2n_\alpha}} + \frac{2}{\alpha}\left\{\sqrt{\frac{\log(8K/\delta)}{2b}} + \sqrt{\frac{32\log(32K(b+1)/\delta)}{b}}\right\} + \frac{4}{\alpha b}.$$

Proof. Define the finite-sample conditional concordance target

$$\bar{A}_k^* = \frac{1}{n_\alpha} \sum_{z_j \in \widehat{\mathcal{T}}_k} g(z_j).$$

We use the decomposition

$$|A_k - A_k^*| \leq |A_k - \bar{A}_k^*| + |\bar{A}_k^* - A_k^*|.$$

The first term is the conditional oracle-evaluation noise. Given the concordance inputs and fitted guides, the set $\widehat{\mathcal{T}}_k$ is fixed. The variables $\mathbb{1}\{R_2(z_j) \geq c\}$ for $z_j \in \widehat{\mathcal{T}}_k$ are independent, bounded in $[0,1]$, and have conditional means $g(z_j)$. Therefore, Hoeffding's inequality gives, for each fixed k,

$$\mathbb{P}\Big(|A_k - \bar{A}_k^*| > \varepsilon \mid z_1, \dots, z_b, \{\widehat{f}_k\}_{k=1}^K\Big) \leq 2\exp(-2n_\alpha\varepsilon^2).$$

A union bound over $k = 1, \dots, K$ gives

$$\mathbb{P}\left(\max_{k=1,\dots,K} |A_k - \bar{A}_k^*| > \varepsilon \mid z_1, \dots, z_b, \{\widehat{f}_k\}_{k=1}^K\right) \leq 2K\exp(-2n_\alpha\varepsilon^2).$$

Setting the right-hand side equal to $\delta/2$ yields, with conditional probability at least $1-\delta/2$,

$$\max_{k=1,\dots,K} |A_k - \bar{A}_k^*| \leq \sqrt{\frac{\log(4K/\delta)}{2n_\alpha}}.$$

It remains to compare $\bar{A}_k^*$ with the population top-tail concordance A_k^*. Expectations below are taken over a population draw $Z \sim P_{\text{conc}}$. For each threshold t, define

$$q_k(t) = \mathbb{E}[\mathbb{1}\{\widehat{f}_k(Z) \geq t\}], \qquad h_k(t) = \mathbb{E}[g(Z)\mathbb{1}\{\widehat{f}_k(Z) \geq t\}],$$

and let $q_{b,k}(t)$ and $h_{b,k}(t)$ denote the corresponding empirical quantities. By the empirical-threshold representation assumption, there exists $\widehat{\tau}_k$ such that $\widehat{\mathcal{T}}_k = \{z_j : \widehat{f}_k(z_j) \geq \widehat{\tau}_k\}$ and $q_{b,k}(\widehat{\tau}_k) = n_\alpha/b$. Since $n_\alpha = \lfloor \alpha b \rfloor$, we also have $|n_\alpha/b - \alpha| \leq b^{-1}$. Therefore,

$$\bar{A}_k^* = \frac{h_{b,k}(\widehat{\tau}_k)}{q_{b,k}(\widehat{\tau}_k)}, \qquad A_k^* = \frac{h_k(\tau_k)}{\alpha}.$$

By the Dvoretzky–Kiefer–Wolfowitz inequality applied to the empirical distribution function of $\widehat{f}_k(z_j)$, together with a union bound over k,

$$\mathbb{P}\left(\max_{k=1,\dots,K} \sup_t |q_{b,k}(t) - q_k(t)| > \sqrt{\frac{\log(8K/\delta)}{2b}}\right) \leq \frac{\delta}{4}.$$

For the weighted numerator term, fix k and consider the function class

$$\mathcal{H}_k = \{z \mapsto g(z)\mathbb{1}\{\widehat{f}_k(z) \geq t\} : t \in \mathbb{R}\}.$$

Every function in $\mathcal{H}_k$ is bounded in $[0,1]$. Moreover, on any b fixed points, the threshold sets $\{z : \widehat{f}_k(z) \geq t\}$ can produce at most $b+1$ distinct subsets as t varies, so the growth function of $\mathcal{H}_k$ is bounded by $b+1$. The standard Vapnik–Chervonenkis

uniform convergence inequality for bounded classes therefore gives, for each fixed k and every $\varepsilon > 0$,

$$\mathbb{P}\left(\sup_t |h_{b,k}(t) - h_k(t)| > \varepsilon\right) \le 8(b+1)\exp\left(-\frac{b\varepsilon^2}{32}\right).$$

A union bound over $k = 1, \dots, K$ yields

$$\mathbb{P}\left(\max_{k=1,\dots,K} \sup_t |h_{b,k}(t) - h_k(t)| > \varepsilon\right) \le 8K(b+1)\exp\left(-\frac{b\varepsilon^2}{32}\right).$$

Taking $\varepsilon = \sqrt{32\log(32K(b+1)/\delta)/b}$ gives

$$\mathbb{P}\left(\max_{k=1,\dots,K} \sup_t |h_{b,k}(t) - h_k(t)| > \sqrt{\frac{32\log(32K(b+1)/\delta)}{b}}\right) \le \frac{\delta}{4}.$$

On the event

$$\sup_t |q_{b,k}(t) - q_k(t)| \le \varepsilon_q, \qquad \sup_t |h_{b,k}(t) - h_k(t)| \le \varepsilon_h,$$

since $q_{b,k}(\widehat{\tau}_k) = n_\alpha/b$ and $|n_\alpha/b - \alpha| \le b^{-1}$,

$$|q_k(\widehat{\tau}_k) - \alpha| \le \varepsilon_q + b^{-1}.$$

Moreover,

$$|h_{b,k}(\widehat{\tau}_k) - h_k(\tau_k)| \le |h_{b,k}(\widehat{\tau}_k) - h_k(\widehat{\tau}_k)| + |h_k(\widehat{\tau}_k) - h_k(\tau_k)|.$$

Since threshold sets are nested and $0 \le g \le 1$,

$$|h_k(\widehat{\tau}_k) - h_k(\tau_k)| \le |q_k(\widehat{\tau}_k) - q_k(\tau_k)| \le \varepsilon_q + b^{-1}.$$

Thus

$$|h_{b,k}(\widehat{\tau}_k) - h_k(\tau_k)| \le \varepsilon_h + \varepsilon_q + b^{-1}.$$

Combining the numerator and denominator perturbations, and using $h_k(\tau_k) \le q_k(\tau_k) = \alpha$, yields

$$\begin{aligned} |\bar{A}_k^* - A_k^*| &= \left|\frac{h_{b,k}(\widehat{\tau}_k)}{q_{b,k}(\widehat{\tau}_k)} - \frac{h_k(\tau_k)}{\alpha}\right| \\ &\le \frac{|h_{b,k}(\widehat{\tau}_k) - h_k(\tau_k)|}{q_{b,k}(\widehat{\tau}_k)} + h_k(\tau_k)\left|\frac{1}{q_{b,k}(\widehat{\tau}_k)} - \frac{1}{\alpha}\right|. \end{aligned}$$

Because $b \ge 2/\alpha$, we have $q_{b,k}(\widehat{\tau}_k) = n_\alpha/b \ge \alpha/2$. Also, $|q_{b,k}(\widehat{\tau}_k) - \alpha| \le b^{-1}$. Therefore,

$$|\bar{A}_k^* - A_k^*| \le \frac{2}{\alpha}(\varepsilon_q + \varepsilon_h) + \frac{4}{\alpha b}.$$

Taking $\varepsilon_q = \sqrt{\log(8K/\delta)/(2b)}$ and $\varepsilon_h = \sqrt{32\log(32K(b+1)/\delta)/b}$ gives, with probability at least $1 - \delta/2$ over the concordance inputs,

$$\max_{k=1,\dots,K} |\bar{A}_k^* - A_k^*| \le \frac{2}{\alpha}\left\{\sqrt{\frac{\log(8K/\delta)}{2b}} + \sqrt{\frac{32\log(32K(b+1)/\delta)}{b}}\right\} + \frac{4}{\alpha b}.$$

Combining this event with the conditional Hoeffding event and applying the triangle inequality proves the result. □

Remark F.2. Theorem F.1 shows that the oracle-concordance scores used by FiCS can be estimated reliably from a sparse oracle budget. The bound separates oracle noise among the selected top-α candidates, scaling as $n_\alpha^{-1/2}$, from the error of using a finite concordance set to approximate the population top-α region, scaling as $b^{-1/2}$. Thus, FiCS does not require dense oracle access over the full candidate space; it only requires enough high-fidelity feedback to compare which resampled reward guides remain concordant with the oracle in the region relevant for steering. Because $\mathcal{D}_2$ is constructed to include high-reward candidates, A_k^* measures concordance in the decision-relevant portion of the search space rather than under the base distribution alone.

Proposition F.3 (Stability of softmax concordance weights)**.** *Let $A, \tilde{A} \in \mathbb{R}^K$ be two vectors of oracle-concordance scores, and let $\varphi = \mathrm{softmax}(\gamma A)$ and $\tilde{\varphi} = \mathrm{softmax}(\gamma \tilde{A})$ denote the corresponding concordance weights for some inverse temperature $\gamma > 0$. Then the induced weights are Lipschitz-stable with respect to perturbations in the concordance scores:*

$$\|\varphi - \tilde{\varphi}\|_1 \le 2\gamma\|A - \tilde{A}\|_\infty.$$

Proof. Define $g : \mathbb{R}^K \to \Delta_K$ by

$$g_k(u) = \frac{\exp(u_k)}{\sum_{\ell=1}^{K} \exp(u_\ell)}, \qquad k = 1, \ldots, K.$$

We bound the ℓ_1 norm of the Jacobian $J_g(u)$ applied to a perturbation vector and then integrate along the path from $\gamma\tilde{A}$ to γA.

The partial derivatives of g_k with respect to u_j are

$$\frac{\partial g_k}{\partial u_j} = \begin{cases} g_k(u)(1 - g_k(u)) & \text{if } j = k, \\ -g_k(u)g_j(u) & \text{if } j \neq k. \end{cases}$$

Equivalently,

$$J_g(u) = \operatorname{diag}(g(u)) - g(u)g(u)^\top.$$

For any vector $v \in \mathbb{R}^K$, the k-th component of $J_g(u)v$ is

$$[J_g(u)v]_k = g_k(u) \left(v_k - \sum_{\ell=1}^{K} g_\ell(u) v_\ell \right).$$

Let $\bar{v} = \sum_{\ell=1}^{K} g_\ell(u) v_\ell$. Since $g(u)$ is a probability vector, $|\bar{v}| \leq \|v\|_\infty$, and hence $|v_k - \bar{v}| \leq 2\|v\|_\infty$ for every k. Therefore,

$$\|J_g(u)v\|_1 = \sum_{k=1}^{K} g_k(u)|v_k - \bar{v}| \leq 2\|v\|_\infty \sum_{k=1}^{K} g_k(u) = 2\|v\|_\infty.$$

Thus $\|J_g(u)\|_{\infty\to 1} \leq 2$ for every u.

Now set $u = \gamma A$ and $\tilde{u} = \gamma\tilde{A}$, and define the path $u(t) = \tilde{u} + t(u - \tilde{u})$ for $t \in [0, 1]$. By the fundamental theorem of calculus,

$$g(u) - g(\tilde{u}) = \int_0^1 J_g(u(t))(u - \tilde{u})\, dt.$$

Taking the ℓ_1 norm and using the bound above,

$$\begin{aligned} \|g(u) - g(\tilde{u})\|_1 &\leq \int_0^1 \|J_g(u(t))(u - \tilde{u})\|_1\, dt \\ &\leq \int_0^1 2\|u - \tilde{u}\|_\infty\, dt \\ &= 2\|u - \tilde{u}\|_\infty. \end{aligned}$$

Since $u - \tilde{u} = \gamma(A - \tilde{A})$, we obtain

$$\|\varphi - \tilde{\varphi}\|_1 = \|g(\gamma A) - g(\gamma\tilde{A})\|_1 \leq 2\gamma\|A - \tilde{A}\|_\infty.$$

□

Theorem F.4 (Uniform score stability)**.** *Let $\varphi^* = \operatorname{softmax}(\gamma A^*)$ denote the oracle-optimal weights computed from the population concordance scores A^*, and let $\hat{\varphi} = \operatorname{softmax}(\gamma\hat{A})$ denote the weights computed from the empirical concordance scores $\hat{A}$. Define the oracle* FiCS *score $S^*(x) = \bar{f}_{\varphi^*}(x) - \lambda\hat{\sigma}_{\varphi^*}(x)$ and the estimated score $\hat{S}(x) = \bar{f}_{\hat{\varphi}}(x) - \lambda\hat{\sigma}_{\hat{\varphi}}(x)$, with $\lambda \geq 0$. Let*

$$B = \max_{k=1,\ldots,K} \sup_x |\widehat{f}_k(x)|$$

be a uniform bound on the guide predictions, and assume $B < \infty$. Suppose that, with probability at least $1 - \delta$,

$$\|\hat{A} - A^*\|_\infty \leq \varepsilon_A(\delta).$$

Set

$$\rho_\delta = \min\{2, 2\gamma\varepsilon_A(\delta)\}.$$

Then, with probability at least $1-\delta$*,*

$$\sup_x |S^*(x) - \hat{S}(x)| \le B\rho_\delta + \lambda B\sqrt{3\rho_\delta}.$$

Proof. Let $\Delta\varphi = \varphi^* - \hat{\varphi}$. On the event $\|\hat{A} - A^*\|_\infty \le \varepsilon_A(\delta)$, Proposition F.3 gives

$$\|\Delta\varphi\|_1 \le 2\gamma\varepsilon_A(\delta).$$

Since φ^* and $\hat{\varphi}$ are probability vectors, $\|\Delta\varphi\|_1 \le 2$ as well. Hence

$$\|\Delta\varphi\|_1 \le \rho_\delta.$$

For the weighted mean term, for any x,

$$\begin{aligned}|\bar{f}_{\varphi^*}(x) - \bar{f}_{\hat{\varphi}}(x)| &= \left|\sum_{k=1}^{K}(\varphi_k^* - \hat{\varphi}_k)\widehat{f}_k(x)\right| \\ &\le B\|\Delta\varphi\|_1.\end{aligned}$$

Next define

$$M_\varphi(x) = \sum_{k=1}^{K}\varphi_k\widehat{f}_k(x)^2, \qquad V_\varphi(x) = \hat{\sigma}_\varphi(x)^2 = M_\varphi(x) - \bar{f}_\varphi(x)^2.$$

Then

$$|M_{\varphi^*}(x) - M_{\hat{\varphi}}(x)| \le B^2\|\Delta\varphi\|_1.$$

Also, since $|\bar{f}_{\varphi^*}(x)|, |\bar{f}_{\hat{\varphi}}(x)| \le B$,

$$\begin{aligned}|\bar{f}_{\varphi^*}(x)^2 - \bar{f}_{\hat{\varphi}}(x)^2| &\le 2B|\bar{f}_{\varphi^*}(x) - \bar{f}_{\hat{\varphi}}(x)| \\ &\le 2B^2\|\Delta\varphi\|_1.\end{aligned}$$

Therefore,

$$|V_{\varphi^*}(x) - V_{\hat{\varphi}}(x)| \le 3B^2\|\Delta\varphi\|_1.$$

Using $|\sqrt{a} - \sqrt{b}| \le \sqrt{|a-b|}$ for $a, b \ge 0$, we obtain

$$|\hat{\sigma}_{\varphi^*}(x) - \hat{\sigma}_{\hat{\varphi}}(x)| \le B\sqrt{3\|\Delta\varphi\|_1}.$$

Combining the mean and standard-deviation perturbations gives

$$\begin{aligned}|S^*(x) - \hat{S}(x)| &\le |\bar{f}_{\varphi^*}(x) - \bar{f}_{\hat{\varphi}}(x)| + \lambda|\hat{\sigma}_{\varphi^*}(x) - \hat{\sigma}_{\hat{\varphi}}(x)| \\ &\le B\|\Delta\varphi\|_1 + \lambda B\sqrt{3\|\Delta\varphi\|_1} \\ &\le B\rho_\delta + \lambda B\sqrt{3\rho_\delta}.\end{aligned}$$

Taking the supremum over x completes the proof. □

Remark F.5. Theorem F.4 establishes that the guide-dependent component of the FiCS score is stable under finite-sample error in the oracle-concordance weights. If a common support penalty is included in both S^* and $\hat{S}$, it cancels exactly, and the bound applies to the full score. The weighted mean term exhibits Lipschitz continuity with respect to weight perturbations, while the weighted disagreement penalty satisfies a denominator-free square-root bound. This square-root dependence arises from bounding the difference of standard deviations directly, without normalizing by the ensemble variance, thereby avoiding assumptions on a quantity that may be arbitrarily small when ensemble agreement is high. The result clarifies the distinct roles of γ and λ: larger γ increases the sensitivity of concordance weights to oracle feedback, while larger λ amplifies the contribution of uncertainty quantification to the final score.

Theorem F.6 (Selection stability under score separation)**.** *Work under the assumptions and notation of Theorem F.4. Let* $x_{(1)}, \ldots, x_{(N)}$ *denote candidates sorted by* $S^*(x)$ *in decreasing order, and let* $\Delta_L = S^*(x_{(L)}) - S^*(x_{(L+1)})$ *denote the score gap at the selection boundary. If*

$$B\rho_\delta + \lambda B\sqrt{3\rho_\delta} < \frac{\Delta_L}{2},$$

then, with probability at least $1-\delta$*, the top-*L *candidates under* $\hat{S}$ *are identical to the top-*L *candidates under* S^**.*

Proof. Let $\varepsilon_S = \sup_x |S^*(x) - \hat{S}(x)|$. By Theorem F.4, with probability at least $1 - \delta$,

$$\varepsilon_S \leq B\rho_\delta + \lambda B\sqrt{3\rho_\delta} < \frac{\Delta_L}{2}.$$

On this event, for any $i \leq L$ and any $j > L$,

$$\hat{S}(x_{(i)}) - \hat{S}(x_{(j)}) \geq S^*(x_{(i)}) - S^*(x_{(j)}) - 2\varepsilon_S \geq \Delta_L - 2\varepsilon_S > 0.$$

Thus, every candidate ranked in the top L by S^* receives a higher estimated score than every candidate outside the top L. Therefore, the top-L selected sets under $\hat{S}$ and S^* coincide. □