NOISE-ROBUST COUNTERFACTUAL EXPLANATIONS UNDER THE RASHOMON EFFECT VIA HOMOTOPY PATH-FOLLOWING

Anonymous authors
Paper under double-blind review

ABSTRACT

Although there has been extensive research on robust counterfactual explanations (RCEs) addressing the Rashomon Effect, when training data contains noisy labels, the Hessian matrix — which characterises the geometry of the ellipsoid — becomes corrupted under label noise for RCE methods based on Rashomon uncertainty sets. This corruption distorts the geometry of the uncertainty set, causing robust optimisation to yield unreliable RCEs. We find that the perturbation ΔH introduced by noisy labels to the Hessian matrix exhibits a low-rank structure, leading to a deflection of the principal directions of the Rashomon ellipsoid defined by the Hessian; this phenomenon is observable across different noise rates, regardless of whether the label noise is random flipping or systematic flipping. Our theoretical analysis further establishes that this conclusion generalises to model uncertainty sets defined by general quadratic forms. To mitigate the impact of noisy labels on RCE, we propose a path-following strategy based on homotopy: starting from a conservative uncertainty set that is independent of labels, label information is gradually incorporated along a continuously deformed path, with the homotopy parameter controlling the degree of trust placed in the training labels. We prove that the robust CE solution is continuously differentiable along the homotopy path, enabling stable path-following and principled selection of the homotopy parameter. Experimental results on five diverse datasets demonstrate that this new framework produces counterfactual explanations that are both robust and computationally efficient across model uncertainty sets defined by different quadratic forms.

1 INTRODUCTION

Automated models and algorithms have been widely applied across diverse domains such as finance and justice to assist people in decision-making(Brennan et al., 2009; Bureau, 2022). When outcomes prove unfavourable, a natural question arises: *What actions can be taken to reverse the result?* Searching for counterfactual explanations (CEs) can help individuals understand predictive outcomes and take actionable steps. In practice, however, updates to the underlying model — whether induced by retraining or distributional shifts in training data — may undermine the validity of the recommended actions. This has motivated a growing research trend toward finding robust counterfactual explanations (RCEs) that remain valid across a set of models characterised by uncertain parameters. Existing methods can be roughly divided into two categories. The first defines the uncertainty set via bounded perturbations in parameter space: a norm ball centred at the current model parameters (Upadhyay et al., 2021; Jiang et al., 2024; 2023). The second, exemplified by T-Rex (Hamman et al., 2023), allows arbitrary parameter variation but requires that the new model's expected predictions over the data manifold remain consistent with those of the original model. This better reflects real-world retraining scenarios, and its core requirement directly connects to the Rashomon Effect — the phenomenon of a set of models that exhibit similar predictive performance over the data manifold.

In the machine learning context, the Rashomon Effect implies that for a given dataset, multiple models with nearly identical predictive performance yet giving different predictions on certain inputs. Motivated by this observation, ElliCE (Turbal et al., 2026) was proposed. It approximates

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

the Rashomon Set — the set of all model parameters whose loss is close to the optimal — as an ellipsoidal uncertainty set via the Hessian matrix of the loss function, and then seeks an RCE that remains valid under the worst-case model within this ellipsoid. However, we argue that robustness remains underachieved. The current approaches share an implicit assumption that the model parameter uncertainty set itself is reliable. In reality, datasets are imperfect and contain randomly mislabeled instances(Song et al., 2022; Piktus et al., 2023) or even systematically biased labels (Eckhouse, 2017). Noisy labels can deflect the principal directions of the Hessian-based ellipsoid used to approximate the Rashomon Set; any CE found within such a geometrically misaligned ellipsoid carries only spurious robustness guarantees and thus offers limited practical utility. This motivates a stronger requirement: the counterfactual explanation must be robust not only to all models within a fixed uncertainty set, but also to estimation errors in the uncertainty set itself — that is, the validity of a CE should not degrade substantially under mild perturbations to the uncertainty set.

In this paper, we aim to answer the following research question: *How can we prevent the approximately optimal model set, contaminated by noisy labels, from steering counterfactual explanations toward spurious optima?* We begin by presenting a Hessian spectral study. We design two noise regimes: one applies random label flipping across the entire dataset, while the other targets only the negative class of a binary classification dataset, thereby synthesising randomly noisy and systematically noisy datasets respectively. We observed that, first, the ellipsoidal uncertainty set geometrically resembles a low-rank structure. Second, noise does not alter this low-rank geometry but instead rotates its principal directions, and this rotation is driven by a low-rank perturbation ΔH; this finding holds consistently across both noise regimes and varying noise levels. These results indicate that the corrupting effect of noisy labels on the Rashomon ellipsoid is both structured and predictable. Essencially, the perturbation mechanism has only a few degrees of freedom, concentrated along the leading directions of the noisy samples in gradient space. We further establish theoretically that this corruption mechanism generalises to arbitrary ellipsoidal uncertainty sets defined by quadratic forms.

Inspired by these findings, we propose a Homotopy Path-Following strategy. Starting from a conservative ellipsoid that is label-agnostic (i.e., the prediction confidence for all samples is set to 50%, showing no preference for any sample), we use a single homotopy parameter t to control the degree of trust placed in the training labels, enabling a smooth transition along a continuous path toward the noisy model's ellipsoid. Along this path, we identify a CF that is robust to directional shifts. Since the shape matrix remains positive definite throughout the path, the robust CE problem is well-posed at every t; we further prove that the solution varies continuously differentiably along the path, and the framework generalises to arbitrary uncertainty sets defined by quadratic forms.

The main contributions of this paper are as follows:

- Through a Hessian spectral analysis under varying noisy label conditions, we observe that noise does not alter the low-rank structure of the ellipsoid but only rotates its orientation, with this rotation driven by a low-rank ΔH. We further prove theoretically that this finding generalizes to uncertainty sets defined by general quadratic forms. This analysis reveals that the corrupting effect of noisy labels on the ellipsoid is predictable rather than random.
- We propose a Homotopy Path-Following strategy applicable to uncertainty sets defined by general quadratic forms. Via a single homotopy parameter t, the strategy traces a continuous path from a conservative ellipsoid to the noisy model's ellipsoid, identifying an CE that is robust to directional shifts of the ellipsoid, with theoretical guarantees on the continuous existence of solutions along the path.
- We conduct extensive experiments on five datasets to evaluate the proposed strategy against baseline methods in terms of robustness and computational efficiency. Results demonstrate that, compared to existing approaches, our framework can efficiently generate RCEs with relatively lower recourse cost and higher validity under noisy label conditions.

2 PRELIMINARIES

Problem setup. Consider a binary classification task with input space $\mathcal{X} \subseteq \mathbb{R}^d$ and label space $\mathcal{Y} = \{0, 1\}$. We work with multilayer perceptrons (MLPs) of the form $f_{\boldsymbol{\theta}}(\mathbf{x}) = \boldsymbol{\theta}^\top \tilde{h}(\mathbf{x})$, where $\tilde{h}(\mathbf{x}) = [h(\mathbf{x})^\top, 1]^\top \in \mathbb{R}^{p+1}$ is the bias-appended penultimate-layer embedding and $\boldsymbol{\theta} \in \mathbb{R}^{p+1}$ are

the last-layer parameters. The model is trained by minimising the ℓ_2-regularised empirical logistic loss $\hat{L}(\boldsymbol{\theta}) = \frac{1}{N}\sum_{i=1}^{N}\phi(f_{\boldsymbol{\theta}}(\mathbf{x}_i), y_i) + \frac{\lambda}{2}\|\boldsymbol{\theta}\|_2^2$, and we denote the empirical risk minimiser by $\hat{\boldsymbol{\theta}}$.

Counterfactual explanations. For a query instance $\mathbf{x}_0$ with predicted label $\hat{y}_0 = 0$, a counterfactual explanation (CE) is a point $\mathbf{x}_c \in \mathcal{X}$ such that $f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_c) \geq \tau$ for some target threshold τ, while minimising the cost $\|\mathbf{x}_c - \mathbf{x}_0\|_2^2$. We focus on generating CEs that flip the predicted label from 0 to 1.

Rashomon set and ellipsoidal approximation. The ϵ-Rashomon set is the collection of model parameters whose training loss is close to optimal:

$$\mathcal{R}(\epsilon) = \left\{\boldsymbol{\theta} : \hat{L}(\boldsymbol{\theta}) \leq \hat{L}(\hat{\boldsymbol{\theta}}) + \epsilon\right\}. \tag{1}$$

Following Turbal et al. (2026), we approximate $\mathcal{R}(\epsilon)$ by a Hessian-defined ellipsoid:

$$\hat{\mathcal{R}}(\epsilon) = \left\{\boldsymbol{\theta} : \frac{1}{2}(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}})^\top H(\boldsymbol{\theta} - \hat{\boldsymbol{\theta}}) \leq \epsilon\right\}, \tag{2}$$

where $H = \frac{1}{N}\Phi^\top W\Phi + \lambda I_{p+1}$ is the Hessian of $\hat{L}$ at $\hat{\boldsymbol{\theta}}$, $\Phi \in \mathbb{R}^{N\times(p+1)}$ is the penultimate-layer feature matrix, and $W = \mathrm{diag}(\sigma_i(1-\sigma_i))$ with $\sigma_i = \sigma(f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_i))$.

Robust counterfactual explanations. A CE $\mathbf{x}_c$ is robust with respect to $\hat{\mathcal{R}}(\epsilon)$ if it satisfies the target threshold under the worst-case model within the ellipsoid. By a Cauchy–Schwarz argument, this worst-case prediction admits a closed form (Turbal et al., 2026), and the robust CE problem reduces to:

$$\min_{\mathbf{x}_c} \|\mathbf{x}_c - \mathbf{x}_0\|_2^2 \quad \text{s.t.} \quad \hat{\boldsymbol{\theta}}^\top\tilde{h}(\mathbf{x}_c) - \sqrt{2\epsilon\,\tilde{h}(\mathbf{x}_c)^\top H^{-1}\tilde{h}(\mathbf{x}_c)} \geq \tau, \tag{3}$$

where the robustness penalty $\sqrt{2\epsilon\,\tilde{h}(\mathbf{x}_c)^\top H^{-1}\tilde{h}(\mathbf{x}_c)}$ is a Mahalanobis norm that measures the worst-case prediction shift within the ellipsoid. When training labels are corrupted, $\hat{\boldsymbol{\theta}}$ is estimated from noisy data and W inherits the label corruption, distorting the geometry of $\hat{\mathcal{R}}(\epsilon)$ and undermining the robustness guarantee. We address this in Section 4.1.

3 Hessian spectral analysis under varying noisy label conditions

In the Hessian spectral analysis, we conduct experiments on two datasets, which are COMPAS(Brennan et al., 2009) and FICO(Holter et al., 2018), under two distinct noise regimes. The first applies random label flipping across the entire dataset; the second restricts flipping to the negative class of the binary classification dataset. These two regimes are used to synthesise symmetrically and asymmetrically noisy datasets, respectively. We further introduce four noise levels (0%, 5%, 10%, and 20%, where 0% denotes the clean dataset) to facilitate a more detailed examination of how the Hessian matrix evolves under label corruption. Our analysis focuses on three properties of the Hessian matrix: (1) **Effective Rank.** It measures the degree to which the energy of the Hessian is concentrated along a small number of directions. When the energy is highly concentrated along a single eigendirection, the corresponding ellipsoid resembles a needle and the effective rank approaches 1; when the energy is spread uniformly across multiple directions, the ellipsoid approximates a sphere and the effective rank approaches the parameter dimensionality. (2) **Principal Direction Deflection Angle.** This is the angle between the leading eigenvectors of the Hessian matrices of the noisy and clean models, respectively. Intuitively, it directly quantifies the degree to which label noise rotates the orientation of the ellipsoid. (3) **Rank-1 Dominance Ratio(R1DR).** This quantity measures the extent to which the noise-induced perturbation to the Hessian, denoted ΔH, is dominated by a single direction. A value approaching 1 indicates that the perturbation is

well approximated by a pure rank-1 matrix, implying that the corrupting effect of label noise on the ellipsoid is almost entirely concentrated along one direction. Formal definitions of all three quantities are provided in the Appendix B. We additionally repeat the same analysis with one- and three-hidden-layer MLPs, and on three further datasets; these complementary results are reported in Appendix B. Our observations are as follows.

First, the ellipsoidal uncertainty set maintains a low-rank structure under label noise. Across both datasets and both noise modes, the effective rank remains stable throughout the entire noise range considered. For COMPAS it stays within 1.74–1.92, and for FICO within 1.98–2.25, with changes of typically within 0.2 relative to the clean baseline. The fact that the effective rank remains around 2 stably across all noise rates, indicates that label noise does not alter the fundamental geometry of the Rashomon ellipsoid.

Second, label noise induces a structured rotation of the principal direction of the ellipsoid. The principal direction rotates monotonically with noise rate η under symmetric noise on both datasets: from 5.0° to 13.6° on COMPAS and from 3.9° to 12.1° on FICO as η increases from 5% to 20%. Asymmetric noise produces comparable rotations that likewise increase monotonically with η, and on both datasets the rotation is even larger at $\eta = 20\%$ (18.8° on COMPAS and 19.2° on FICO). We argue that these experimental results confirm that the rotation is a structured, predictable effect.

Last, the perturbation ΔH is low-rank. One might expect symmetric noise to produce a less structured perturbation than asymmetric noise, as randomly flipping labels from both classes introduces gradient contributions of opposing signs that partially cancel, potentially dispersing energy across multiple directions rather than concentrating it along a few leading directions. Contrary to this expectation, the R1DR of ΔH remains consistently high across both datasets and both noise regimes, ranging from 0.76 to 0.83 on COMPAS and from 0.79 to 0.86 on FICO throughout the entire noise range. This indicates that the entire perturbation to the ellipsoid is concentrated in a low-dimensional subspace, explaining why the rotation observed in Observation 2 is predictable: the corruption mechanism only acts along the leading directions in gradient space.

In summary, label noise induces a structured rotation of the low-rank ellipsoid, yet the underlying corruption mechanism has only a few degrees of freedom, acting along the leading directions in gradient space. This insight suggests that, although robust counterfactual explanations must account for the influence of label noise, a targeted correction along these leading directions is sufficient to mitigate the corruption.

4 METHOD

Overview. Motivated by the findings of the Hessian spectral analysis, we design an effective corrective mechanism, namely the Homotopy Path-Following (HPF) strategy, to mitigate the influence of label noise and obtain robust counterfactual explanations (RCEs) within model parameter uncertainty sets approximated by Rashomon ellipsoids. In this section, we first demonstrate how the proposed strategy operates on the ellipsoidal approximation of the Rashomon Set whose geometry is governed by the Hessian matrix. We then provide a theoretical analysis establishing that the HPF strategy generalises to uncertainty sets defined by general quadratic forms with an arbitrary positive definite shape matrix Q. Finally, we present two concrete instantiations employing distinct positive definite matrices: an ellipsoidal uncertainty set based on the empirical Fisher Information Matrix (FIM), and one derived from the bootstrap parameter covariance.

4.1 HOMOTOPY PATH-FOLLOWING FOR NOISE-ROBUST COUNTERFACTUAL EXPLANATION

Our spectral analysis (Section 3) shows that this causes a low-rank perturbation of H (denoted ΔH) that rotates its principal directions, distorting the geometry of the Rashomon ellipsoid and causing the worst-case model to be estimated incorrectly.

Label-agnostic baseline. To obtain a reference ellipsoid that is immune to label corruption, we set all predicted probabilities to $1/2$, yielding constant weights $w_i = 1/4$:

$$H_0 = \frac{1}{4N}\Phi^\top\Phi + \lambda I_{p+1}. \tag{4}$$

Because H_0 does not depend on $\hat{\boldsymbol{\theta}}$ or any training label, it serves as a label-agnostic, conservative approximation of the Rashomon ellipsoid. The corresponding uncertainty set $\hat{\mathcal{R}}_0(\epsilon)$ is shaped solely by the feature covariance $\Phi^\top\Phi$, with no label-dependent weighting; its geometry is therefore insensitive to label noise by construction.

Homotopy construction. We interpolate between the label-agnostic baseline H_0 and the label-conditioned Hessian $H_1 = \frac{1}{N}\Phi^\top W\Phi + \lambda I_{p+1}$ via a scalar path parameter $t \in [0,1]$:

$$H_t = (1-t)H_0 + tH_1, \quad t \in [0,1]. \tag{5}$$

Each H_t defines an approximated Rashomon ellipsoid $\hat{\mathcal{R}}_t(\epsilon) = \left\{\boldsymbol{\theta} : \frac{1}{2}(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top H_t(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \le \epsilon\right\}$. The path starts at the noise-robust baseline ($t=0$) and gradually incorporates label information as t increases toward 1. Since both H_0 and H_1 are positive definite (as $\lambda > 0$), H_t is positive definite for all $t \in [0,1]$, ensuring $\hat{\mathcal{R}}_t(\epsilon)$ is a well-defined bounded ellipsoid.

In practice, we estimate t automatically by treating it as a learnable scalar optimised to maximise the worst-case robustness of a set of candidate counterfactuals. Specifically, given a set of K candidate counterfactuals $\{\mathbf{x}_c^{(k)}\}_{k=1}^K$ generated at $t=1$, we solve:

$$\max_{t\in[t_{\min},1]} \frac{1}{K}\sum_{k=1}^{K} \sigma\left(\hat{\boldsymbol{\theta}}^\top\tilde{h}(\mathbf{x}_c^{(k)}) - \sqrt{2\epsilon\tilde{h}(\mathbf{x}_c^{(k)})^\top H_t^{-1}\tilde{h}(\mathbf{x}_c^{(k)})}\right) \tag{6}$$

where $\sigma(\cdot)$ is the sigmoid function. The objective measures how confidently each candidate satisfies the target threshold under the worst-case model within $\hat{\mathcal{R}}_t(\epsilon)$. Since H_t depends affinely on t, the entire path $t \to H_t \to H_t^{-1/2} \to$ worst-case logit is differentiable, and we optimise 6 via gradient ascent. This procedure is a single-level approximation to the bilevel problem of jointly optimising t and the counterfactuals; we fix the candidate set at $t=1$ as a proxy and subsequently regenerate counterfactuals using the learned t. We then denote the solution of the robust CE problem 2 by $\mathbf{x}_c^*(t)$.

Theorem 2 (Path continuity of robust recourse)**.** *Suppose the robust CE problem in equation , with H replaced by H_t is feasible for all $t \in [0,1]$, and that $\nabla_{\mathbf{x}_c} g(\mathbf{x}_c^*(t), t) \neq \mathbf{0}$ for all t, where $g(\mathbf{x}_c, t) := \hat{\boldsymbol{\theta}}^\top\tilde{h}(\mathbf{x}_c) - \sqrt{2\epsilon\,\tilde{h}(\mathbf{x}_c)^\top H_t^{-1}\tilde{h}(\mathbf{x}_c)} - \tau$. Then the solution map $t \mapsto \mathbf{x}_c^*(t)$ is C^1 on $[0,1]$.*

We prove Theorem 2 in Appendix C. As the solution map $t \mapsto \mathbf{x}_c^*(t)$ is continuously differentiable, it can be guaranteed that the HPF strategy does not produce abrupt changes in the recommended recourse as t is varied, enabling stable path-following and principled selection of t.

4.2 Generalisation to Quadratic Uncertainty Sets

The homotopy construction in Section 4.1 relies only on the positive definiteness of H_0 and H_1, not on their specific form as Hessians. We now show that the entire framework extends to uncertainty sets defined by an arbitrary positive definite matrix Q, and reveal that the robust margin is a Mahalanobis distance in parameter space.

General quadratic uncertainty set. Let $Q \in \mathbb{R}^{q\times q}$ be any positive definite matrix. Define the ellipsoidal uncertainty set:

$$\mathcal{U}_Q(\epsilon) = \left\{\boldsymbol{\theta} : (\boldsymbol{\theta}-\hat{\boldsymbol{\theta}})^\top Q\,(\boldsymbol{\theta}-\hat{\boldsymbol{\theta}}) \le 2\epsilon\right\}, \tag{7}$$

which is a ball of radius $\sqrt{2\epsilon}$ in the Mahalanobis metric induced by Q on the parameter space. The Hessian-based Rashomon ellipsoid $\hat{\mathcal{R}}(\epsilon)$ is the special case $Q = H$. By a Cauchy–Schwarz argument, the worst-case prediction over $\mathcal{U}_Q(\epsilon)$ admits the closed form:

$$\min_{\boldsymbol{\theta}\in\mathcal{U}_Q(\epsilon)} \boldsymbol{\theta}^\top\tilde{h}(\mathbf{x}_c) \;=\; \hat{\boldsymbol{\theta}}^\top\tilde{h}(\mathbf{x}_c) - \sqrt{2\epsilon}\,\|\tilde{h}(\mathbf{x}_c)\|_{Q^{-1}}, \tag{8}$$

where $\|\tilde{h}(\mathbf{x}_c)\|_{Q^{-1}} := \sqrt{\tilde{h}(\mathbf{x}_c)^\top Q^{-1} \tilde{h}(\mathbf{x}_c)}$ is the Mahalanobis norm of the penultimate-layer representation of $\mathbf{x}_c$ under the metric Q^{-1}. Intuitively, $\|\tilde{h}(\mathbf{x}_c)\|_{Q^{-1}}$ measures how far the counterfactual $\mathbf{x}_c$ can shift the prediction in the direction of greatest uncertainty within $\mathcal{U}_Q(\epsilon)$. The RCE problem becomes:

$$\min_{\mathbf{x}_c} \ \|\mathbf{x}_c - \mathbf{x}_0\|_2^2 \quad \text{s.t.} \quad \hat{\boldsymbol{\theta}}^\top \tilde{h}(\mathbf{x}_c) - \sqrt{2\epsilon}\, \|\tilde{h}(\mathbf{x}_c)\|_{Q^{-1}} \ \geq \ \tau. \tag{9}$$

Different choices of Q thus induce different Mahalanobis metrics, each reflecting a distinct notion of model uncertainty: the Hessian H captures local curvature of the loss, the Fisher Information Matrix (FIM) weights directions of residual uncertainty, and the bootstrap precision matrix reflects empirical parameter variability across resamples.

General homotopy. Given a label-agnostic baseline $Q_0 \succ 0$ and a label-conditioned matrix $Q_1 \succ 0$, define:

$$Q_t = (1-t)\, Q_0 + t\, Q_1, \qquad t \in [0,1]. \tag{10}$$

Since $Q_0,\, Q_1 \succ 0$, Q_t is positive definite for all $t \in [0,1]$, and problem equation 9 with $Q = Q_t$ is well-defined at every t. The homotopy path thus interpolates between two Mahalanobis metrics, gradually shifting the geometry of the uncertainty set from the noise-robust baseline to the label-conditioned estimate. Moreover, Theorem 2 applies directly with H_t replaced by Q_t: the solution map $t \mapsto \mathbf{x}_c^*(t)$ is continuously differentiable on $[0,1]$.

Two instantiations. We consider two choices of Q_1 beyond the Hessian.

Empirical Fisher Information Matrix (FIM).

$$Q_{\mathrm{FIM}} = \frac{1}{N} \sum_{i=1}^{N} (y_i - \sigma_i)^2\, \tilde{h}(\mathbf{x}_i) \tilde{h}(\mathbf{x}_i)^\top + \lambda I, \tag{11}$$

where the squared residual weights $(y_i - \sigma_i)^2$ up-weight samples where the model prediction is inconsistent with the observed label, making $\|\tilde{h}(\mathbf{x}_c)\|_{Q_{\mathrm{FIM}}^{-1}}$ sensitive to directions corrupted by label noise.

Bootstrap precision matrix. Let $\{\hat{\boldsymbol{\theta}}_k\}_{k=1}^K$ be last-layer parameters from K bootstrap resamples, and let $\bar{\boldsymbol{\theta}} = \frac{1}{K} \sum_k \hat{\boldsymbol{\theta}}_k$. It can be represented as follows:

$$\Sigma_{\mathrm{boot}} = \frac{1}{K} \sum_{k=1}^{K} (\hat{\boldsymbol{\theta}}_k - \bar{\boldsymbol{\theta}})(\hat{\boldsymbol{\theta}}_k - \bar{\boldsymbol{\theta}})^\top, \qquad Q_{\mathrm{boot}} = (\Sigma_{\mathrm{boot}} + \lambda I)^{-1}. \tag{12}$$

The Mahalanobis norm $\|\tilde{h}(\mathbf{x}_c)\|_{Q_{\mathrm{boot}}^{-1}} = \|\tilde{h}(\mathbf{x}_c)\|_{\Sigma_{\mathrm{boot}} + \lambda I}$ penalises directions of high parameter variance across resamples, so the robust margin reflects genuine sampling uncertainty rather than label-dependent curvature.

For both instantiations, Q_1 is trace-normalised to match Q_0: we replace $Q_1 \leftarrow \frac{\mathrm{tr}(Q_0)}{\mathrm{tr}(Q_1)}\, Q_1$, ensuring that both endpoints of the homotopy path operate at the same overall scale.

5 Experiments

Datasets. We validate our proposed strategy on five datasets: (1) **COMPAS.** (Brennan et al., 2009) The task is to predict whether a defendant will re-offend within two years. (2) **FICO.** (Holter et al., 2018) It is widely used in explainable machine learning and fair lending research due to its mix of credit history features and potential for demographic disparity analysis. (3) **German Credit.** (Hofmann, 1994) A classic benchmark for binary credit risk classification, with the task of labelling customers as good or bad credit risks. (4) **Australian Credit Approval.** (Quinlan, 1987)

Table 1: Results using AWP evaluator. Evaluation metrics: (1) "Val" denotes Validity; (2) "RF" denotes robust fraction; (3) "L2" denotes ℓ_2 distance; (4) "LOF" denotes local outlier factor. Best per column in **bold**. See Appendix E for results on the other evaluator and on other datasets.

AWP - COMPAS

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	0.72 ± 0.45	1.76 ± 1.43	2.33 ± 17.10	0.72 ± 0.45	0.67 ± 0.47	1.45 ± 1.25	2.40 ± 17.08	0.64 ± 0.48	0.81 ± 0.39	1.92 ± 1.27	2.49 ± 20.27	0.74 ± 0.44
ROAR	$\mathbf{1.00 \pm 0.00}$	1.94 ± 0.68	1.59 ± 0.85	0.70 ± 0.46	$\mathbf{1.00 \pm 0.00}$	2.15 ± 0.72	1.63 ± 0.78	0.65 ± 0.48	1.00 ± 0.02	1.68 ± 0.55	1.45 ± 0.85	0.77 ± 0.42
T-Rex	0.67 ± 0.47	$\mathbf{0.80 \pm 0.17}$	1.44 ± 1.02	0.50 ± 0.50	0.56 ± 0.50	$\mathbf{0.63 \pm 0.20}$	1.67 ± 17.14	0.70 ± 0.46	0.68 ± 0.47	$\mathbf{0.73 \pm 0.20}$	1.36 ± 1.20	0.47 ± 0.50
ProPlace	0.43 ± 0.50	1.22 ± 0.65	$\mathbf{1.33 \pm 0.76}$	0.29 ± 0.45	0.40 ± 0.49	1.24 ± 0.65	$\mathbf{1.33 \pm 0.81}$	0.22 ± 0.41	0.39 ± 0.49	1.14 ± 0.60	$\mathbf{1.28 \pm 0.49}$	0.53 ± 0.50
HPF (Hessian)	1.00 ± 0.03	2.51 ± 1.39	1.95 ± 0.90	0.83 ± 0.37	0.99 ± 0.10	2.92 ± 1.59	1.84 ± 0.88	1.00 ± 0.07	0.96 ± 0.19	2.22 ± 1.38	1.75 ± 0.75	0.90 ± 0.30
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	3.43 ± 1.99	2.81 ± 1.61	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	4.23 ± 2.10	2.35 ± 1.17	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.33 ± 1.78	2.54 ± 1.29	$\mathbf{1.00 \pm 0.00}$
HPF (FIM)	1.00 ± 0.02	3.16 ± 1.74	2.89 ± 1.52	1.00 ± 0.03	1.00 ± 0.02	3.76 ± 2.00	2.49 ± 1.30	0.99 ± 0.08	0.99 ± 0.12	3.02 ± 1.63	2.63 ± 1.26	0.99 ± 0.12

AWP - FICO

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	0.99 ± 0.08	5.12 ± 1.91	1.37 ± 0.27	0.99 ± 0.08	1.00 ± 0.06	5.28 ± 1.71	1.37 ± 0.26	0.99 ± 0.07	0.98 ± 0.13	4.79 ± 1.74	1.33 ± 0.24	0.98 ± 0.13
ROAR	0.71 ± 0.45	2.11 ± 0.91	1.20 ± 0.14	0.42 ± 0.49	0.91 ± 0.29	2.28 ± 0.88	1.21 ± 0.13	0.55 ± 0.50	0.54 ± 0.50	1.79 ± 0.94	1.19 ± 0.13	0.33 ± 0.47
T-Rex	0.32 ± 0.47	$\mathbf{0.65 \pm 0.21}$	1.11 ± 0.10	0.14 ± 0.35	0.28 ± 0.45	$\mathbf{0.50 \pm 0.22}$	1.10 ± 0.10	0.01 ± 0.10	0.33 ± 0.47	$\mathbf{0.54 \pm 0.24}$	1.10 ± 0.10	0.06 ± 0.24
ProPlace	0.09 ± 0.28	2.19 ± 0.60	$\mathbf{1.02 \pm 0.04}$	0.79 ± 0.41	0.08 ± 0.27	2.24 ± 0.59	$\mathbf{1.02 \pm 0.04}$	0.54 ± 0.50	0.09 ± 0.29	2.14 ± 0.58	1.03 ± 0.05	0.64 ± 0.48
HPF (Hessian)	$\mathbf{1.00 \pm 0.00}$	4.04 ± 1.66	1.30 ± 0.20	1.00 ± 0.03	$\mathbf{1.00 \pm 0.00}$	4.64 ± 1.76	1.34 ± 0.25	0.96 ± 0.21	1.00 ± 0.02	4.08 ± 1.57	1.28 ± 0.18	1.00 ± 0.02
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	3.83 ± 1.54	1.32 ± 0.20	$\mathbf{1.00 \pm 0.00}$	1.00 ± 0.01	4.71 ± 1.76	1.37 ± 0.24	$\mathbf{1.00 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	3.89 ± 1.54	1.31 ± 0.19	$\mathbf{1.00 \pm 0.00}$
HPF (FIM)	$\mathbf{1.00 \pm 0.00}$	3.74 ± 1.59	1.29 ± 0.19	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	4.44 ± 1.76	1.33 ± 0.24	1.00 ± 0.03	$\mathbf{1.00 \pm 0.00}$	6.70 ± 1.58	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$

AWP - Diabetes

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	$\mathbf{1.00 \pm 0.00}$	3.33 ± 1.03	1.07 ± 0.07	$\mathbf{1.00 \pm 0.00}$	0.97 ± 0.16	2.96 ± 1.16	1.05 ± 0.07	0.97 ± 0.16	$\mathbf{1.00 \pm 0.00}$	3.23 ± 0.95	1.05 ± 0.07	$\mathbf{1.00 \pm 0.00}$
ROAR	$\mathbf{1.00 \pm 0.00}$	1.88 ± 0.76	1.10 ± 0.12	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	1.77 ± 0.66	1.12 ± 0.16	0.73 ± 0.44	$\mathbf{1.00 \pm 0.00}$	1.58 ± 0.63	1.11 ± 0.15	0.96 ± 0.19
T-Rex	0.76 ± 0.42	$\mathbf{1.04 \pm 0.25}$	1.09 ± 0.16	0.66 ± 0.47	0.70 ± 0.46	$\mathbf{0.82 \pm 0.35}$	1.10 ± 0.18	0.59 ± 0.49	0.73 ± 0.45	$\mathbf{0.90 \pm 0.31}$	1.10 ± 0.18	0.62 ± 0.49
ProPlace	0.97 ± 0.16	1.72 ± 0.72	$\mathbf{1.04 \pm 0.08}$	$\mathbf{1.00 \pm 0.00}$	0.99 ± 0.11	1.86 ± 0.92	$\mathbf{1.04 \pm 0.09}$	0.69 ± 0.46	0.98 ± 0.13	1.75 ± 0.89	$\mathbf{1.03 \pm 0.08}$	0.84 ± 0.37
HPF (Hessian)	$\mathbf{1.00 \pm 0.00}$	2.45 ± 1.02	1.11 ± 0.13	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.04 ± 1.22	1.10 ± 0.10	0.95 ± 0.23	$\mathbf{1.00 \pm 0.00}$	2.56 ± 1.11	1.11 ± 0.16	$\mathbf{1.00 \pm 0.00}$
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	2.37 ± 1.08	1.12 ± 0.14	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.00 ± 1.26	1.10 ± 0.12	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.45 ± 0.81	1.12 ± 0.09	$\mathbf{1.00 \pm 0.00}$
HPF (FIM)	$\mathbf{1.00 \pm 0.00}$	2.23 ± 1.05	1.11 ± 0.15	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.84 ± 1.25	1.09 ± 0.10	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.37 ± 1.10	1.10 ± 0.15	$\mathbf{1.00 \pm 0.00}$

This dataset contains anonymised credit card application records and the task is to predict whether an application will be approved. (5) **Diabetes (Pima Indians).** (Smith et al., 1988) This dataset contains medical examination records and the task is to predict whether a patient has diabetes. Full details of data processing are provided in Appendix D.

Synthetic data. To evaluate robustness under label noise, we inject noise into the training labels at a fixed rate of $\eta = 10\%$ under two modes. **Symmetric noise**: each training label is independently flipped with probability η, regardless of class. **Asymmetric noise**: labels are flipped with probability η only within the negative class.

Baselines. We compare HPF against four baselines. (1) **ElliCE** (Turbal et al., 2026) approximates the Rashomon set as a Hessian-defined ellipsoid and solves a closed-form QCQP to generate robust CEs, serving as our closest baseline. (2) **ROAR** (Upadhyay et al., 2021) formulates a minimax objective under a worst-case ℓ_p-bounded parameter perturbation, optimised via an adversarial training procedure. (3) **PROPLACE**(Jiang et al., 2024) adopts an ℓ_∞ parameter ball and solves the robust optimisation problem exactly via MILP with interval neural network abstractions. (4) **T-Rex**(Hamman et al., 2023) characterises model uncertainty through naturally-occurring prediction changes on the data manifold and provides probabilistic validity guarantees via a Stability measure.We evaluate three variants of our proposed method, differing in how the ellipsoid matrix Q_1 is constructed. **HPF(Hessian)** uses the label-dependent Hessian H_1^{-1} directly. **HPF(bootstrap)** estimates Q_1 from the empirical covariance of parameters obtained by bootstrap resampling. **HPF(FIM)** constructs Q_1 as a Fisher information matrix weighted by squared residuals. In all variants, the homotopy parameter t is estimated via gradient ascent on proxy counterfactual explanations generated from training negatives.

Metrics. We consider four metrics in our evaluation: (1) *Validity* measures the fraction of counterfactuals that achieve the target class prediction under the original trained model. A counterfactual $\mathbf{x}_c$ is valid if $f_{\hat{\boldsymbol{\theta}}}(\mathbf{x}_c) \geq 0$. (2) *Robust fraction* measures the fraction of valid counterfactuals that retain their target-class prediction when evaluated against a set of alternative models within a bounded neighbourhood of the parameter space. (3) *L2 distance* measures the proximity of a counterfactual to the original instance as $\|\mathbf{x}_c - \mathbf{x}_0\|_2$; smaller values indicate less costly recourse. (4) *Local Outlier Factor (LOF)* (Jiang et al., 2024) measures the plausibility of a counterfactual by comparing its local density to that of its neighbours in the training data; values close to 1 indicate a realistic, in-distribution point.

Table 2: Averaged single-query counterfactual generation time (second) over 4 repeated runs on a single fixed query per dataset.

Speed					
Method	COMPAS	FICO	Diabetes	German	Australian
ElliCE	0.08 ± 0.00	0.05 ± 0.00	0.10 ± 0.02	0.13 ± 0.04	0.07 ± 0.00
ROAR	0.01 ± 0.00	0.01 ± 0.00	0.01 ± 0.00	0.02 ± 0.00	0.01 ± 0.00
T-Rex	0.09 ± 0.02	0.09 ± 0.03	0.09 ± 0.02	0.09 ± 0.01	0.09 ± 0.01
ProPlace	259.53 ± 519.06	398.12 ± 796.24	24.59 ± 49.18	26.42 ± 52.85	22.36 ± 44.72
HPF (Hessian)	0.10 ± 0.01	0.07 ± 0.01	0.12 ± 0.02	0.17 ± 0.10	0.07 ± 0.00
HPF (Bootstrap)	0.12 ± 0.01	0.08 ± 0.01	0.11 ± 0.02	0.23 ± 0.10	0.01 ± 0.00
HPF (FIM)	0.13 ± 0.00	0.12 ± 0.01	0.09 ± 0.02	0.12 ± 0.08	0.06 ± 0.00

Experimental Setting. We train a two-hidden-layer MLP with early stopping. We use binary cross entropy loss and the Adam optimiser in the training. All experiments use 4-fold stratified cross-validation (fixed seed): in each fold, the base model is trained on three folds (75% of data) and the held-out fold is used for counterfactual generation and evaluation; evaluation metrics are averaged across folds. We focus on generating counterfactuals for instances predicted as class 0, targeting class 1. Counterfactuals are generated via continuous optimisation in feature space. To simulate the Rashomon effect, we employ two evaluators: (1)The **retrain** evaluator generates 100 alternative models by retraining from scratch with different random seeds on the same training fold.(2) The **AWP** (Hsu & Calmon, 2022) evaluator applies adversarial weight perturbation with step size 10^{-3} for 50 steps to produce 100 worst-case model perturbations. Both with Rashomon tolerance $\epsilon_{\text{eval}} = 1.1\,\hat{\mathcal{L}}$, where $\hat{\mathcal{L}}$ is the training loss. The Rashomon radius is $\epsilon = 0.01$, regularisation coefficient $\lambda = 10^{-3}$, and label noise rate $\eta = 10\%$. For HPF, t is selected via gradient ascent (200 steps, lr $= 0.05$) with a lower bound $t_{\min} = 0.6$ to ensure the uncertainty set retains sufficient label information and avoid degenerate solutions where the ellipsoid is entirely decoupled from the learned decision boundary. Our experiments were conducted on Apple M4 Mac mini (16 GB RAM) for development.

5.1 Results

Tables 1 show the results comparing our proposed models to other baselines. Across all five datasets and three noise regimes, HPF(bootstrap) and HPF(FIM) achieve AWP robust fraction (RF) ≈ 1.00 throughout, and HPF(Hessian) closely follows; all HPF variants simultaneously maintain validity at 1.00 on four of five datasets. Among baselines, ROAR performs best under clean labels, achieving RF ≥ 0.97 on Australian, Diabetes, and German, but its reliance on nearest-neighbour structures renders it sensitive to label corruption: under asymmetric noise, RF collapses to 0.53 on German (0.98 clean), as corrupted negative-class neighbours directly invalidate its recourse directions. While under both symmetric and asymmetric noise, HPF shows no meaningful degradation. HPF(Hessian) shows a modest gap relative to the other HPF variants on COMPAS (RF 0.83 clean, 0.90 asymmetric noise), yet remains strictly above all baselines on that dataset. We attribute this gap to the quality of Q_1: unlike HPF(bootstrap) and HPF(FIM), whose precision matrices directly capture empirical parameter uncertainty through bootstrap resampling and residual weighting respectively, HPF(Hessian) relies on H_1^{-1}, which is a local quadratic approximation of the loss landscape that is more sensitive to the biased and socially complex label structure (Eckhouse, 2017) of COMPAS. At the other extreme, T-Rex achieves the smallest ℓ_2 distances but collapses in robustness, with RF reaching 0.01 on FICO under symmetric noise, exposing a fundamental incompatibility between proximity optimisation and Rashomon-set robustness. ProPlace fails more severely still, unable to generate valid counterfactuals on COMPAS (validity ≈ 0.40) and FICO (validity ≈ 0.09) regardless of noise regime.

5.2 Computational Efficiency and Robustness-Cost Trade-offs

Tables 2 reports per-query generation time for all methods. Per-query generation time for HPF variants is comparable to ElliCE and T-Rex. ROAR is the fastest method overall, though this speed advantage comes at a significant cost in robustness. ProPlace is wholly impractical: per-query time reaches 398s on FICO with a standard deviation of 796s which is twice the mean, indicating

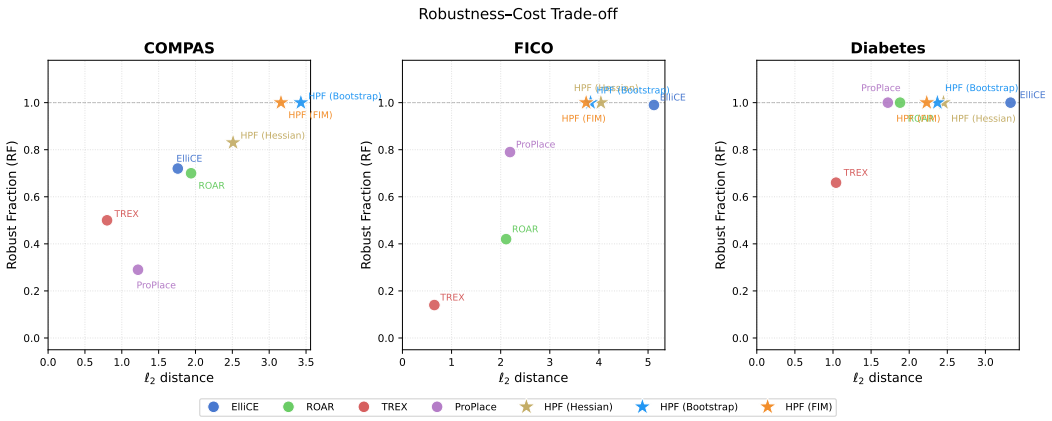

Figure 1: Robustness–cost trade-off on three datasets under the AWP evaluator using original labels. Methods in the upper-left region achieve both high robustness and low cost.

severe convergence failures on a substantial fraction of instances, consistent with its low validity scores on that dataset. Figure 1 reveals a clear robustness-cost trade-off across methods. T-Rex represents the most extreme case: it achieves the lowest ℓ_2 distance on all datasets but also the low AWP robust fraction, confirming that counterfactual explanations(CEs) placed closest to the decision boundary are most fragile under model perturbation. ROAR occupies an inconsistent position in the robustness–proximity space, achieving RF $= 1.00$ on Diabetes at $\ell_2 = 1.88$ but only RF $= 0.42$ on FICO at $\ell_2 = 2.11$, suggesting its robustness is highly dataset-dependent. HPF variants consistently lie on the Pareto frontier: on FICO, HPF(FIM) and HPF(bootstrap) achieve RF $= 1.00$ at $\ell_2 \approx 3.7$–3.8, while ElliCE requires $\ell_2 = 5.12$ to reach RF $= 0.99$. On COMPAS, HPF(bootstrap) and HPF(FIM) attain RF $= 1.00$ at $\ell_2 \approx 3.2$–3.4, whereas ROAR reaches only RF $= 0.70$ at $\ell_2 = 1.94$. These results demonstrate that HPF achieves high robustness at a proximity cost comparable to ElliCE, offering a more favourable robustness-cost balance than any baseline method.

6 Conclusion and Limitations

Existing robust CE methods implicitly assume that training labels are clean; in practice, noisy or biased labels corrupt the estimation of the model parameter uncertainty set, rendering the resulting robustness guarantees spurious. We proposed Homotopy Path-Following (HPF), which interpolates between a label-agnostic and a label-dependent ellipsoid via a scalar parameter t, preserving positive definiteness throughout and generalising to arbitrary quadratic uncertainty sets. In our experiments, HPF consistently outperforms all baselines in robustness across multiple datasets and three noise regimes, while maintaining competitive proximity and plausibility. To the best of our knowledge, this work is among the first to study the effect of label noise on the Rashomon uncertainty set and its impact on robust CE generation. Our approach has several limitations. First, geometric regularisation does not guarantee recovery of the uncertainty set that would be obtained from uncorrupted labels. The centre and hidden-layer parameters remain fixed, leaving noise-induced changes in these quantities unaddressed. Second, our method and main evaluation focus on MLPs. This limits its interpretation as a universal correction principle and motivates further evaluation beyond MLPs. Finally, The primary limitation is t estimation: the gradient-based variant can be less stable when the decision boundary is ambiguous, and a uniform t per fold may be suboptimal for individual queries. Extending the approach to these settings remains a direction for future work.

AI Use Statement

Generative AI tools were used to assist with manuscript writing and editing, literature retrieval and discovery, drafting and revising portions of the paper, and executing code-based experimental tasks through AI agents.

The AI agents were used to assist with implementing and running experimental workflows based on specifications provided by the authors. Their outputs, including code changes and experimental execution, were reviewed and verified by the authors before being incorporated into this work. The authors independently determined the research methodology, experimental design, mathematical formulation, analysis, and interpretation of results, and take full responsibility for the content of this paper.

REFERENCES

Emily Black, Zifan Wang, Matt Fredrikson, and Anupam Datta. Consistent counterfactuals for deep models. arXiv preprint arXiv:2110.03109, 2021.

Zachery Boner, Harry Chen, Lesia Semenova, Ronald Parr, and Cynthia Rudin. Using noise to infer aspects of simplicity without learning. Advances in Neural Information Processing Systems, 37: 131824–131858, 2024.

Leo Breiman. Statistical modeling: The two cultures (with comments and a rejoinder by the author). Statistical science, 16(3):199–231, 2001.

Tim Brennan, William Dieterich, and Beate Ehret. Evaluating the predictive validity of the compas risk and needs assessment system. Criminal Justice and behavior, 36(1):21–40, 2009.

Consumer Financial Protection Bureau. Cfpb targets unfair discrimination in consumer finance. Press release, Consumer Financial Protection Bureau, Washington, DC, 2022.

Gordon Dai, Pavan Ravishankar, Rachel Yuan, Emily Black, and Daniel B Neill. Be intentional about fairness!: Fairness, size, and multiplicity in the rashomon set. In Proceedings of the 5th ACM Conference on Equity and Access in Algorithms, Mechanisms, and Optimization, pp. 42–73, 2025.

Alexander D'Amour, Katherine Heller, Dan Moldovan, Ben Adlam, Babak Alipanahi, Alex Beutel, Christina Chen, Jonathan Deaton, Jacob Eisenstein, Matthew D Hoffman, et al. Underspecification presents challenges for credibility in modern machine learning. Journal of Machine Learning Research, 23(226):1–61, 2022.

Jon Donnelly, Srikar Katta, Cynthia Rudin, and Edward Browne. The rashomon importance distribution: Getting rid of unstable, single model-based variable importance. Advances in Neural Information Processing Systems, 36:6267–6279, 2023.

Laurel Eckhouse. Big data may be reinforcing racial bias in the criminal justice system. The Washington Post, 2017.

Faisal Hamman, Erfaun Noorani, Saumitra Mishra, Daniele Magazzeni, and Sanghamitra Dutta. Robust counterfactual explanations for neural networks with probabilistic guarantees. In International Conference on Machine Learning, pp. 12351–12367. PMLR, 2023.

Hans Hofmann. Statlog (German Credit Data). UCI Machine Learning Repository, 1994. DOI: https://doi.org/10.24432/C5NC77.

Steffen Holter, Oscar Gomez, and Enrico Bertini. Fico explainable machine learning challenge. FICO COmmunity, 2018.

Hsiang Hsu and Flavio Calmon. Rashomon capacity: A metric for predictive multiplicity in classification. Advances in Neural Information Processing Systems, 35:28988–29000, 2022.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Formalising the robustness of counterfactual explanations for neural networks. In Proceedings of the AAAI conference on artificial intelligence, volume 37, pp. 14901–14909, 2023.

Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. Provably robust and plausible counterfactual explanations for neural networks via robust optimisation. In Asian Conference on Machine Learning, pp. 582–597. PMLR, 2024.

Lucas Langlade, Julien Ferry, Gabriel Laberge, and Thibaut Vidal. Fairness and sparsity within rashomon sets: Enumeration-free exploration and characterization. In Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society, volume 8, pp. 1536–1547, 2025.

Charles Marx, Flavio Calmon, and Berk Ustun. Predictive multiplicity in classification. In International conference on machine learning, pp. 6765–6774. PMLR, 2020.

Jorge Nocedal and Stephen J Wright. Numerical optimization. Springer, 2006.

Aleksandra Piktus, Christopher Akiki, Paulo Villegas, Hugo Laurençon, Gérard Dupont, Sasha Luccioni, Yacine Jernite, and Anna Rogers. The roots search tool: Data transparency for llms. In Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations), pp. 304–314, 2023.

Ross Quinlan. Statlog (Australian Credit Approval). UCI Machine Learning Repository, 1987. DOI: https://doi.org/10.24432/C59012.

Stephan R Sain. The nature of statistical learning theory, 1996.

Lesia Semenova, Cynthia Rudin, and Ronald Parr. On the existence of simpler machine learning models. In Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency, pp. 1827–1858, 2022.

Lesia Semenova, Harry Chen, Ronald Parr, and Cynthia Rudin. A path to simpler models starts with noise. Advances in neural information processing systems, 36:3362–3401, 2023.

Jack W Smith, James E Everhart, William C Dickson, William C Knowler, and Robert Scott Johannes. Using the adap learning algorithm to forecast the onset of diabetes mellitus. In Proceedings of the annual symposium on computer application in medical care, pp. 261, 1988.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. Learning from noisy labels with deep neural networks: A survey. IEEE transactions on neural networks and learning systems, 34(11):8135–8153, 2022.

Bohdan Turbal, Iryna Voitsitska, and Lesia Semenova. Ellice: Efficient and provably robust algorithmic recourse via the rashomon sets. arXiv preprint arXiv:2602.07674, 2026.

Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse. Advances in Neural Information Processing Systems, 34:16926–16937, 2021.

Rui Xin, Chudi Zhong, Zhi Chen, Takuya Takagi, Margo Seltzer, and Cynthia Rudin. Exploring the whole rashomon set of sparse decision trees. Advances in neural information processing systems, 35:14071–14084, 2022.

Chudi Zhong, Zhi Chen, Jiachang Liu, Margo Seltzer, and Cynthia Rudin. Exploring and interacting with the set of good sparse generalized additive models. Advances in neural information processing systems, 36:56673–56699, 2023.

A Related Work

Robust counterfactual explanations. Counterfactual explanations (CEs) are widely used to provide actionable recommendations to individuals affected by automated decisions. A critical challenge is that CEs generated for a fixed model can become invalid once the model is retrained or updated, as small shifts in data or hyperparameters can substantially alter the decision boundary (Black et al., 2021). This motivates a growing body of work on robust CEs (RCEs) that remain valid across plausible model changes. One line of research characterises model uncertainty as a bounded perturbation in parameter space. ROAR (Upadhyay et al., 2021) defines the uncertainty set as an ℓ_p-norm ball over parameter or gradient perturbations and optimises a minimax objective that jointly minimises CE cost and maximises worst-case validity, solved via an adversarial training procedure that alternates between finding the worst-case perturbation and updating the CE. PROPLACE (Jiang et al., 2024) extends this by adopting an ℓ_∞ parameter ball within a bi-level optimisation framework: the outer level minimises CE cost and the inner level finds the worst-case model shift, with

both levels solved via MILP and iterated to convergence. Crucially, PROPLACE restricts the CE search space to the convex hull of k Δ-robust nearest neighbours of the query, explicitly enforcing plausibility alongside robustness. Both approaches require the user to pre-specify a perturbation radius that is decoupled from the training data and does not reflect the actual geometry of near-optimal models. A complementary line of work, exemplified by T-Rex, abandons bounded parameter perturbations and instead characterises model uncertainty through the notion of naturally-occurring model change: the updated model may differ arbitrarily in parameter space, but its expected predictions on the data manifold remain consistent with those of the original model and its local output variance is bounded. T-Rex (Hamman et al., 2023) quantifies CE robustness via a Stability measure that rewards CEs located in high-confidence, low-variability regions of the model output, and proves via Gaussian concentration bounds that CEs with sufficiently high Stability remain valid under naturally-occurring model change with high probability. While this captures a more realistic class of model changes, the guarantees are probabilistic rather than deterministic, and the Stability measure depends on the local Lipschitz constant of the model, which is generally unknown and must be approximated in practice. ElliCE (Turbal et al., 2026) takes a different approach by approximating the Rashomon set as an ellipsoid defined by the Hessian of the regularised training loss, yielding a data-driven, closed-form characterisation of model uncertainty that is both tractable and theoretically grounded. However, ElliCE computes the Hessian at a fixed label configuration, making it sensitive to label noise: when training labels are corrupted, the Hessian geometry is distorted, and the resulting ellipsoid may fail to faithfully approximate the true Rashomon set. Our work addresses this limitation by introducing a Homotopy path between a label-agnostic and a label-dependent Hessian, allowing the ellipsoid geometry to adapt to the degree and type of label noise while preserving ElliCE's closed-form tractability and deterministic robustness guarantees. More broadly, all existing approaches share an implicit assumption that the signals used to define the uncertainty set are themselves reliable. In practice, training labels are often noisy or systematically biased, which corrupts the estimated uncertainty set and renders the resulting robustness guarantees spurious. Our work is the first to explicitly address this blind spot in the context of Noise-RCEs generation over Rashomon set.

Rashomon Effect. The Rashomon effect (Breiman, 2001; Sain, 1996; D'Amour et al., 2022) refers to the phenomenon that many distinct models achieve near-equivalent predictive performance on the same dataset. Formally, the Rashomon set is the set of models whose empirical loss lies within a small tolerance ε of the best model (Semenova et al., 2022). This perspective has motivated work across several areas: enumerating near-optimal interpretable models such as sparse decision trees (Xin et al., 2022) and GAMs (Zhong et al., 2023); studying fairness within the Rashomon set (Dai et al., 2025; Langlade et al., 2025); and quantifying predictive multiplicity, defined as the degree to which near-optimal models assign conflicting predictions to the same individual (Marx et al., 2020; Hsu & Calmon, 2022; Donnelly et al., 2023). The relationship between label noise and the Rashomon set has received recent theoretical attention. Semenova et al. (2023) show empirically that label noise is a key driver of large Rashomon sets and model simplicity, while Boner et al. (2024) formally prove that training under noisy labels is equivalent to optimising on clean data with amplified regularisation, implying that noisy Rashomon sets tend to contain simpler models. However, neither line of work considers how noisy labels distort the geometric approximation of the Rashomon set used for downstream tasks such as robust counterfactual explanation generation, which is the gap our work addresses.

B Spectral Properties of the Hessian

The same protocol is repeated with one- and three-hidden-layer MLPs. These complementary results are reported below.

Three metrics are used to characterise how label noise distorts the Hessian geometry.

Effective Rank. Let $\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_d$ be the eigenvalues of H. The effective rank is defined as

$$\mathrm{EffRank}(H) = \frac{\left(\sum_i |\lambda_i|\right)^2}{\sum_i \lambda_i^2}, \tag{13}$$

Table 3: Hessian spectral analysis on all five datasets under a two-hidden-layer MLP $[32, 32]$ ($\eta \in \{0\%, 5\%, 10\%, 20\%\}$; means over three seeds). Evaluation metrics: (1) "EffRank" denotes Effective Rank; (2) "PDD Angle" denotes Principal Direction Deflection Angle; (3) "R1DR" denotes Rank-1 Dominance Ratio.

Symmetric Noise

	EffRank				PDD Angle (°)				R1DR			
Dataset	0%	5%	10%	20%	0%	5%	10%	20%	0%	5%	10%	20%
COMPAS	1.91	1.91	1.86	1.84	–	5.0	6.6	13.6	–	0.81	0.77	0.80
FICO	2.15	2.17	2.19	2.20	–	3.9	6.2	12.1	–	0.81	0.80	0.81
German	1.59	1.66	1.64	1.58	–	10.9	13.0	19.6	–	0.88	0.82	0.84
Australian	2.11	2.10	2.17	2.13	–	6.9	9.5	13.7	–	0.86	0.78	0.76
Diabetes	2.05	2.09	2.13	2.28	–	5.4	9.3	17.3	–	0.70	0.77	0.87

Asymmetric Noise

	EffRank				PDD Angle (°)				R1DR			
Dataset	0%	5%	10%	20%	0%	5%	10%	20%	0%	5%	10%	20%
COMPAS	1.91	1.87	1.83	1.74	–	4.1	9.7	18.8	–	0.83	0.81	0.76
FICO	2.15	2.21	2.20	2.09	–	5.4	9.9	19.2	–	0.79	0.83	0.86
German	1.59	1.58	1.62	1.72	–	8.3	11.2	17.9	–	0.86	0.85	0.93
Australian	2.11	2.13	2.11	2.13	–	4.0	9.4	12.3	–	0.75	0.69	0.65
Diabetes	2.05	2.09	2.13	2.15	–	7.8	9.6	16.5	–	0.79	0.80	0.80

Table 4: All five datasets under a one-hidden-layer MLP $[32]$ (means over three seeds).

Symmetric Noise

	EffRank				PDD Angle (°)				R1DR			
Dataset	0%	5%	10%	20%	0%	5%	10%	20%	0%	5%	10%	20%
COMPAS	2.35	2.38	2.41	2.53	–	1.4	2.8	7.6	–	0.86	0.81	0.80
FICO	3.89	3.95	4.04	4.13	–	1.6	3.4	6.2	–	0.64	0.67	0.67
German	1.95	2.06	2.23	2.44	–	9.9	13.0	16.9	–	0.82	0.88	0.93
Australian	3.87	3.91	3.99	3.96	–	2.8	4.5	6.7	–	0.60	0.67	0.69
Diabetes	3.63	3.62	3.67	3.70	–	2.8	3.7	5.9	–	0.59	0.66	0.62

Asymmetric Noise

	EffRank				PDD Angle (°)				R1DR			
Dataset	0%	5%	10%	20%	0%	5%	10%	20%	0%	5%	10%	20%
COMPAS	2.35	2.36	2.38	2.42	–	3.4	5.8	12.8	–	0.84	0.83	0.85
FICO	3.89	3.86	3.81	3.90	–	3.1	6.4	9.9	–	0.76	0.77	0.72
German	1.95	2.00	2.13	2.26	–	6.4	14.6	20.0	–	0.80	0.83	0.90
Australian	3.87	3.87	3.84	3.81	–	4.6	5.3	9.7	–	0.68	0.65	0.68
Diabetes	3.63	3.64	3.65	3.68	–	2.9	5.3	8.7	–	0.67	0.64	0.64

i.e., the squared ratio of the nuclear norm to the Frobenius norm of H. Values close to 1 indicate that energy is concentrated along a single eigendirection (a highly anisotropic ellipsoid); values close to d indicate isotropy.

Table 5: All five datasets under a three-hidden-layer MLP $[32, 32, 32]$ (means over three seeds).

Symmetric Noise

	EffRank				PDD Angle (°)				R1DR			
Dataset	0%	5%	10%	20%	0%	5%	10%	20%	0%	5%	10%	20%
COMPAS	1.78	1.78	1.72	1.62	–	6.4	7.5	13.7	–	0.84	0.83	0.84
FICO	1.84	1.88	1.86	1.89	–	5.0	7.5	15.8	–	0.94	0.91	0.91
German	1.61	1.70	1.62	1.66	–	11.5	15.6	17.6	–	0.86	0.89	0.85
Australian	1.91	1.96	1.99	1.91	–	9.4	10.4	17.4	–	0.76	0.80	0.75
Diabetes	1.61	1.63	1.76	1.99	–	7.2	11.8	19.8	–	0.84	0.90	0.96

Asymmetric Noise

	EffRank				PDD Angle (°)				R1DR			
Dataset	0%	5%	10%	20%	0%	5%	10%	20%	0%	5%	10%	20%
COMPAS	1.78	1.78	1.72	1.58	–	7.0	13.4	24.4	–	0.86	0.81	0.88
FICO	1.84	1.87	1.89	1.84	–	7.1	13.5	27.4	–	0.80	0.83	0.85
German	1.61	1.59	1.72	1.77	–	11.3	17.3	22.7	–	0.87	0.87	0.87
Australian	1.91	1.93	1.99	2.00	–	7.6	12.6	17.9	–	0.76	0.77	0.75
Diabetes	1.61	1.69	1.73	1.70	–	8.1	10.0	18.2	–	0.85	0.84	0.83

Principal Direction Deflection Angle. Let $v_1^{(clean)}$ and $v_1^{(\eta)}$ denote the leading eigenvectors of the clean Hessian H_{clean} and the noisy Hessian H_η, respectively. The deflection angle is

$$\theta = \arccos\left(\left|{v_1^{(clean)}}^\top v_1^{(\eta)}\right|\right). \tag{14}$$

The absolute value removes sign ambiguity, giving $\theta \in [0°, 90°]$.

Rank-1 Dominance Ratio. Let $\Delta H = H_\eta - H_{clean}$ and let μ_1 be its largest eigenvalue. The rank-1 dominance ratio is

$$\mathrm{R1DR}(\Delta H) = \frac{|\mu_1|}{\|\Delta H\|_F}, \tag{15}$$

where $\|\cdot\|_F$ denotes the Frobenius norm. A value approaching 1 indicates that $\Delta H \approx \mu_1 v_1 v_1^\top$, i.e., the noise-induced perturbation to the Hessian is well approximated by a rank-1 matrix.

C FOR THEOREM 2

Proof. Since $H_0, H_1 \succeq \lambda I \succ 0$ and $H_t = (1-t)H_0 + tH_1$, we have $H_t \succeq \lambda I \succ 0$ for all $t \in [0, 1]$. The map $A \mapsto A^{-1}$ is C^∞ on positive definite matrices, so $t \mapsto H_t^{-1}$ is C^∞, and consequently $g(\mathbf{x}_c, t) := \hat{\boldsymbol{\theta}}^\top \tilde{h}(\mathbf{x}_c) - \sqrt{2\epsilon\, \tilde{h}(\mathbf{x}_c)^\top H_t^{-1} \tilde{h}(\mathbf{x}_c)} - \tau$ is C^∞ in $(\mathbf{x}_c, t)$. Since $\mathbf{x}_0$ belongs to the negative class, $g(\mathbf{x}_0, t) < 0$, so the unconstrained minimiser is infeasible and the constraint is active at any feasible optimum: $g(\mathbf{x}_c^*(t), t) = 0$. The feasible set $\mathcal{F}(t) = \{\mathbf{x}_c : g(\mathbf{x}_c, t) \geq 0\}$ is convex (as $g(\cdot, t)$ is concave), and the strictly convex objective ensures a unique solution $\mathbf{x}_c^*(t)$ at each t. Under the assumption $\nabla_{\mathbf{x}_c} g \neq \mathbf{0}$ (LICQ), there exists a unique multiplier $\mu^*(t) > 0$, and the KKT system

$$F(\mathbf{x}_c, \mu, t) := \begin{pmatrix} 2(\mathbf{x}_c - \mathbf{x}_0) - \mu \nabla_{\mathbf{x}_c} g(\mathbf{x}_c, t) \\ g(\mathbf{x}_c, t) \end{pmatrix} = \mathbf{0}$$

is C^1 in all arguments. The Jacobian of F with respect to $(\mathbf{x}_c, \mu)$ is

$$J = \begin{pmatrix} 2I - \mu^* \nabla^2_{\mathbf{x}_c \mathbf{x}_c} g & -\nabla_{\mathbf{x}_c} g \\ (\nabla_{\mathbf{x}_c} g)^\top & 0 \end{pmatrix}.$$

Since $\sqrt{2\epsilon\, \mathbf{x}_c^\top H_t^{-1} \mathbf{x}_c}$ is convex, $-\nabla^2_{\mathbf{x}_c \mathbf{x}_c} g \succeq 0$, and hence $2I - \mu^* \nabla^2_{\mathbf{x}_c \mathbf{x}_c} g \succeq 2I \succ 0$, which implies J is invertible (Nocedal & Wright, 2006). By the implicit function theorem, there exists a

unique C^1 local solution curve around each $t_0 \in [0, 1]$. Since the global solution is unique, these local curves patch together into a global C^1 map $t \mapsto \mathbf{x}_c^*(t)$ on $[0, 1]$. □

D Dataset Preprocessing

Table 6: Dataset summary. For each dataset we list the preprocessing steps applied, the resulting size after balancing, and the immutable features excluded from counterfactual modification. All datasets undergo majority-class under-sampling and per-fold standardisation.

Dataset	**Additional preprocessing**	**Size**	**Immutable features**
COMPAS	Select 8 columns; dropna; binary-encode sex, charge degree; one-hot encode race	6392×12	race, sex
FICO	Drop 3 high-missing columns; replace $-7/-8/-9$ with NaN; drop rows $>$40% missing; fill NaN $\to -1$	9470×20	—
German	One-hot encode all categorical features	600×61	personal status, foreign worker
Australian	Drop identifier column; dropna; label-encode categorical features	$\approx 530 \times 14$	—
Diabetes	Retain zero values as-is	536×8	DiabetesPedigreeFunction

E Additional Experiments

Table 7: Results using AWP evaluator. Evaluation metrics: (1) "Val" denotes Validity; (2) "RF" denotes robust fraction; (3) "L2" denotes ℓ_2 distance (4) "LOF" denotes local outlier factor. Best per column in **bold**.

AWP - COMPAS

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	0.72 ± 0.45	1.76 ± 1.43	2.33 ± 17.10	0.72 ± 0.45	0.67 ± 0.47	1.45 ± 1.25	2.40 ± 17.08	0.64 ± 0.48	0.81 ± 0.39	1.92 ± 1.27	2.49 ± 20.27	0.74 ± 0.44
ROAR	$\mathbf{1.00 \pm 0.00}$	1.94 ± 0.68	1.59 ± 0.85	0.70 ± 0.46	$\mathbf{1.00 \pm 0.00}$	2.15 ± 0.72	1.63 ± 0.78	0.65 ± 0.48	1.00 ± 0.02	1.68 ± 0.55	1.45 ± 0.85	0.77 ± 0.42
T-Rex	0.67 ± 0.47	$\mathbf{0.80 \pm 0.17}$	1.44 ± 1.02	0.50 ± 0.50	0.56 ± 0.50	$\mathbf{0.63 \pm 0.20}$	1.67 ± 17.14	0.70 ± 0.46	0.68 ± 0.47	$\mathbf{0.73 \pm 0.20}$	1.36 ± 1.20	0.47 ± 0.50
ProPlace	0.43 ± 0.50	1.22 ± 0.65	$\mathbf{1.33 \pm 0.76}$	0.29 ± 0.45	0.40 ± 0.49	1.24 ± 0.65	$\mathbf{1.33 \pm 0.81}$	0.22 ± 0.41	0.39 ± 0.49	1.14 ± 0.60	$\mathbf{1.28 \pm 0.49}$	0.53 ± 0.50
HPF (Hessian)	1.00 ± 0.03	2.51 ± 1.39	1.95 ± 0.90	0.83 ± 0.37	0.99 ± 0.10	2.92 ± 1.59	1.84 ± 0.88	1.00 ± 0.07	0.96 ± 0.19	2.22 ± 1.38	1.75 ± 0.75	0.90 ± 0.30
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	3.43 ± 1.99	2.81 ± 1.61	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	4.23 ± 2.10	2.35 ± 1.17	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.33 ± 1.78	2.54 ± 1.29	$\mathbf{1.00 \pm 0.00}$
HPF (FIM)	1.00 ± 0.02	3.16 ± 1.74	2.89 ± 1.52	1.00 ± 0.03	1.00 ± 0.02	3.76 ± 2.00	2.49 ± 1.30	0.99 ± 0.08	0.99 ± 0.12	3.02 ± 1.63	2.63 ± 1.26	0.99 ± 0.12

AWP - FICO

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	0.99 ± 0.08	5.12 ± 1.91	1.37 ± 0.27	0.99 ± 0.08	1.00 ± 0.06	5.28 ± 1.71	1.37 ± 0.26	0.99 ± 0.07	0.98 ± 0.13	4.79 ± 1.74	1.33 ± 0.24	0.98 ± 0.13
ROAR	0.71 ± 0.45	2.11 ± 0.91	1.20 ± 0.14	0.42 ± 0.49	0.91 ± 0.29	2.28 ± 0.88	1.21 ± 0.13	0.55 ± 0.50	0.54 ± 0.50	1.79 ± 0.94	1.19 ± 0.13	0.33 ± 0.47
T-Rex	0.32 ± 0.47	$\mathbf{0.65 \pm 0.21}$	1.11 ± 0.10	0.14 ± 0.35	0.28 ± 0.45	$\mathbf{0.50 \pm 0.22}$	1.10 ± 0.10	0.01 ± 0.10	0.33 ± 0.47	$\mathbf{0.54 \pm 0.24}$	1.10 ± 0.10	0.06 ± 0.24
ProPlace	0.09 ± 0.28	2.19 ± 0.60	$\mathbf{1.02 \pm 0.04}$	0.79 ± 0.41	0.08 ± 0.27	2.24 ± 0.59	$\mathbf{1.02 \pm 0.04}$	0.54 ± 0.50	0.09 ± 0.29	2.14 ± 0.58	1.03 ± 0.05	0.64 ± 0.48
HPF (Hessian)	$\mathbf{1.00 \pm 0.00}$	4.04 ± 1.66	1.30 ± 0.20	1.00 ± 0.03	$\mathbf{1.00 \pm 0.00}$	4.64 ± 1.76	1.34 ± 0.25	0.96 ± 0.21	1.00 ± 0.02	4.08 ± 1.57	1.28 ± 0.18	1.00 ± 0.02
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	3.83 ± 1.54	1.32 ± 0.20	$\mathbf{1.00 \pm 0.00}$	1.00 ± 0.01	4.71 ± 1.76	1.37 ± 0.24	$\mathbf{1.00 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	3.89 ± 1.54	1.31 ± 0.19	$\mathbf{1.00 \pm 0.00}$
HPF (FIM)	$\mathbf{1.00 \pm 0.00}$	3.74 ± 1.59	1.29 ± 0.19	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	4.44 ± 1.76	1.33 ± 0.24	1.00 ± 0.03	$\mathbf{1.00 \pm 0.00}$	6.70 ± 1.58	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$

AWP - Diabetes

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	$\mathbf{1.00 \pm 0.00}$	3.33 ± 1.03	1.07 ± 0.07	$\mathbf{1.00 \pm 0.00}$	0.97 ± 0.16	2.96 ± 1.16	1.05 ± 0.07	0.97 ± 0.16	$\mathbf{1.00 \pm 0.00}$	3.23 ± 0.95	1.05 ± 0.07	$\mathbf{1.00 \pm 0.00}$
ROAR	$\mathbf{1.00 \pm 0.00}$	1.88 ± 0.76	1.10 ± 0.12	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	1.77 ± 0.66	1.12 ± 0.16	0.73 ± 0.44	$\mathbf{1.00 \pm 0.00}$	1.58 ± 0.63	1.11 ± 0.15	0.96 ± 0.19
T-Rex	0.76 ± 0.42	$\mathbf{1.04 \pm 0.25}$	1.09 ± 0.16	0.66 ± 0.47	0.70 ± 0.46	$\mathbf{0.82 \pm 0.35}$	1.10 ± 0.18	0.59 ± 0.49	0.73 ± 0.45	$\mathbf{0.90 \pm 0.31}$	1.10 ± 0.18	0.62 ± 0.49
ProPlace	0.97 ± 0.16	1.72 ± 0.72	$\mathbf{1.04 \pm 0.08}$	$\mathbf{1.00 \pm 0.00}$	0.99 ± 0.11	1.86 ± 0.92	$\mathbf{1.04 \pm 0.09}$	0.69 ± 0.46	0.98 ± 0.13	1.75 ± 0.89	$\mathbf{1.03 \pm 0.08}$	0.84 ± 0.37
HPF (Hessian)	$\mathbf{1.00 \pm 0.00}$	2.45 ± 1.02	1.11 ± 0.13	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.04 ± 1.22	1.10 ± 0.10	0.95 ± 0.23	$\mathbf{1.00 \pm 0.00}$	2.56 ± 1.11	1.11 ± 0.16	$\mathbf{1.00 \pm 0.00}$
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	2.37 ± 1.08	1.12 ± 0.14	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.00 ± 1.26	1.10 ± 0.12	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.45 ± 0.81	1.12 ± 0.09	$\mathbf{1.00 \pm 0.00}$
HPF (FIM)	$\mathbf{1.00 \pm 0.00}$	2.23 ± 1.05	1.11 ± 0.15	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.84 ± 1.25	1.09 ± 0.10	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.37 ± 1.10	1.10 ± 0.15	$\mathbf{1.00 \pm 0.00}$

AWP - Australian

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	$\mathbf{1.00 \pm 0.00}$	3.94 ± 1.08	1.03 ± 0.07	$\mathbf{1.00 \pm 0.00}$	0.99 ± 0.08	4.14 ± 1.00	1.04 ± 0.07	0.99 ± 0.08	0.99 ± 0.10	3.93 ± 1.01	1.03 ± 0.06	0.99 ± 0.10
ROAR	$\mathbf{1.00 \pm 0.00}$	2.18 ± 0.49	1.10 ± 0.15	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.17 ± 0.51	1.08 ± 0.09	1.00 ± 0.06	$\mathbf{1.00 \pm 0.00}$	1.84 ± 0.41	1.08 ± 0.14	0.96 ± 0.20
T-Rex	0.82 ± 0.39	$\mathbf{1.40 \pm 0.36}$	1.08 ± 0.14	0.79 ± 0.40	0.52 ± 0.50	$\mathbf{0.89 \pm 0.36}$	1.07 ± 0.14	0.44 ± 0.50	0.80 ± 0.40	$\mathbf{1.19 \pm 0.36}$	1.07 ± 0.09	0.77 ± 0.42
ProPlace	$\mathbf{1.00 \pm 0.00}$	2.85 ± 0.85	$\mathbf{1.03 \pm 0.05}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.83 ± 0.70	$\mathbf{1.02 \pm 0.04}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.70 ± 0.79	$\mathbf{1.03 \pm 0.05}$	0.99 ± 0.12
HPF (Hessian)	$\mathbf{1.00 \pm 0.00}$	2.92 ± 0.86	1.12 ± 0.10	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.22 ± 0.89	1.11 ± 0.10	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.85 ± 0.86	1.11 ± 0.09	$\mathbf{1.00 \pm 0.00}$
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	2.54 ± 0.79	1.13 ± 0.10	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.94 ± 0.86	1.12 ± 0.09	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.45 ± 0.81	1.12 ± 0.09	$\mathbf{1.00 \pm 0.00}$
HPF (FIM)	$\mathbf{1.00 \pm 0.00}$	2.56 ± 0.82	1.12 ± 0.10	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.02 ± 0.87	1.12 ± 0.09	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.49 ± 0.86	1.11 ± 0.09	$\mathbf{1.00 \pm 0.00}$

AWP - German

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	$\mathbf{1.00 \pm 0.00}$	7.20 ± 1.70	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	7.95 ± 1.60	1.02 ± 0.07	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	7.21 ± 1.60	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$
ROAR	0.99 ± 0.10	3.65 ± 1.18	1.13 ± 0.12	0.98 ± 0.15	0.97 ± 0.16	3.76 ± 1.62	1.14 ± 0.13	0.97 ± 0.18	0.51 ± 0.50	2.71 ± 1.88	1.12 ± 0.12	0.53 ± 0.50
T-Rex	0.59 ± 0.49	$\mathbf{0.91 \pm 0.27}$	1.10 ± 0.12	0.57 ± 0.49	0.52 ± 0.50	$\mathbf{0.65 \pm 0.28}$	1.10 ± 0.13	0.51 ± 0.50	0.58 ± 0.49	$\mathbf{0.76 \pm 0.25}$	1.10 ± 0.12	0.57 ± 0.50
ProPlace	0.96 ± 0.19	6.97 ± 1.65	1.02 ± 0.06	$\mathbf{1.00 \pm 0.00}$	0.97 ± 0.18	7.17 ± 1.50	1.03 ± 0.07	$\mathbf{1.00 \pm 0.00}$	0.96 ± 0.19	6.89 ± 1.60	1.03 ± 0.07	$\mathbf{1.00 \pm 0.00}$
HPF (Hessian)	$\mathbf{1.00 \pm 0.00}$	3.26 ± 0.59	1.09 ± 0.05	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.28 ± 0.73	1.08 ± 0.05	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.11 ± 0.59	1.08 ± 0.05	$\mathbf{1.00 \pm 0.00}$
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	6.90 ± 1.65	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	7.48 ± 1.87	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	6.78 ± 1.69	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$
HPF (FIM)	$\mathbf{1.00 \pm 0.00}$	6.81 ± 1.64	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	7.38 ± 1.52	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	6.70 ± 1.58	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$

Table 8: Results using Retrain evaluator. Evaluation metrics: (1) "Val" denotes Validity; (2) "RF" denotes robust fraction; (3) "L2" denotes ℓ_2 distance (4) "LOF" denotes local outlier factor. Best per column in **bold**.

Retrain - COMPAS

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	0.65 ± 0.48	1.62 ± 1.39	2.17 ± 16.98	0.60 ± 0.45	0.66 ± 0.47	1.49 ± 1.28	2.27 ± 17.06	0.59 ± 0.45	0.79 ± 0.40	1.78 ± 1.17	2.03 ± 1.31	0.69 ± 0.42
ROAR	$\mathbf{1.00 \pm 0.00}$	1.94 ± 0.66	1.61 ± 0.88	0.99 ± 0.04	$\mathbf{1.00 \pm 0.00}$	2.13 ± 0.69	1.65 ± 0.66	$\mathbf{1.00 \pm 0.03}$	$\mathbf{1.00 \pm 0.00}$	1.70 ± 0.55	1.51 ± 0.91	0.99 ± 0.04
T-Rex	0.67 ± 0.47	$\mathbf{0.80 \pm 0.17}$	1.44 ± 1.02	0.60 ± 0.44	0.56 ± 0.50	$\mathbf{0.63 \pm 0.20}$	1.67 ± 17.14	0.51 ± 0.46	0.68 ± 0.47	$\mathbf{0.73 \pm 0.20}$	1.36 ± 1.20	0.62 ± 0.44
ProPlace	0.42 ± 0.49	1.23 ± 0.67	$\mathbf{1.34 \pm 0.53}$	0.98 ± 0.05	0.40 ± 0.49	1.23 ± 0.64	$\mathbf{1.34 \pm 0.61}$	0.99 ± 0.04	0.46 ± 0.50	1.06 ± 0.54	$\mathbf{1.28 \pm 0.79}$	0.99 ± 0.04
HPF (Hessian)	0.99 ± 0.12	2.62 ± 1.58	2.09 ± 0.88	0.99 ± 0.12	0.94 ± 0.24	2.64 ± 1.48	1.91 ± 0.92	0.94 ± 0.24	0.94 ± 0.23	2.22 ± 1.40	1.85 ± 0.75	0.94 ± 0.23
HPF (Bootstrap)	1.00 ± 0.02	3.27 ± 1.71	2.95 ± 1.60	1.00 ± 0.02	1.00 ± 0.05	4.00 ± 1.68	2.31 ± 1.23	1.00 ± 0.06	0.98 ± 0.13	3.21 ± 1.66	2.76 ± 1.35	0.98 ± 0.13
HPF (FIM)	$\mathbf{1.00 \pm 0.00}$	3.03 ± 1.54	2.95 ± 1.47	$\mathbf{1.00 \pm 0.01}$	0.99 ± 0.10	3.70 ± 1.77	2.57 ± 1.37	0.99 ± 0.10	1.00 ± 0.06	2.97 ± 1.46	2.81 ± 1.26	$\mathbf{1.00 \pm 0.06}$

Retrain - FICO

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	1.00 ± 0.04	5.10 ± 1.69	1.35 ± 0.26	0.97 ± 0.11	0.98 ± 0.13	5.07 ± 1.60	1.37 ± 0.25	0.91 ± 0.22	0.99 ± 0.10	4.76 ± 1.48	1.34 ± 0.24	0.95 ± 0.17
ROAR	0.82 ± 0.39	1.98 ± 0.86	1.20 ± 0.14	0.53 ± 0.30	0.97 ± 0.17	2.18 ± 0.86	1.20 ± 0.13	0.75 ± 0.26	0.20 ± 0.40	1.51 ± 0.85	1.20 ± 0.14	0.17 ± 0.21
T-Rex	0.32 ± 0.47	$\mathbf{0.65 \pm 0.21}$	1.11 ± 0.10	0.27 ± 0.41	0.28 ± 0.45	0.50 ± 0.22	1.10 ± 0.10	0.24 ± 0.38	0.33 ± 0.47	$\mathbf{0.54 \pm 0.24}$	1.10 ± 0.10	0.29 ± 0.41
ProPlace	0.08 ± 0.26	2.19 ± 0.59	1.02 ± 0.04	1.00 ± 0.03	0.07 ± 0.25	2.24 ± 0.58	$\mathbf{1.02 \pm 0.05}$	$\mathbf{1.00 \pm 0.01}$	0.09 ± 0.29	2.16 ± 0.55	$\mathbf{1.03 \pm 0.05}$	1.00 ± 0.02
HPF (Hessian)	$\mathbf{1.00 \pm 0.00}$	4.32 ± 1.86	1.31 ± 0.22	1.00 ± 0.01	1.00 ± 0.01	4.94 ± 1.85	1.37 ± 0.26	1.00 ± 0.05	$\mathbf{1.00 \pm 0.00}$	4.23 ± 1.72	1.29 ± 0.20	1.00 ± 0.01
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	3.91 ± 1.67	1.32 ± 0.21	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	4.65 ± 1.76	1.38 ± 0.25	$\mathbf{1.00 \pm 0.01}$	$\mathbf{1.00 \pm 0.00}$	3.87 ± 1.61	1.31 ± 0.19	1.00 ± 0.01
HPF (FIM)	$\mathbf{1.00 \pm 0.00}$	3.96 ± 1.77	1.29 ± 0.20	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	4.68 ± 1.84	1.34 ± 0.25	1.00 ± 0.04	$\mathbf{1.00 \pm 0.00}$	3.92 ± 1.63	1.28 ± 0.18	$\mathbf{1.00 \pm 0.00}$

Retrain - Diabetes

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	$\mathbf{1.00 \pm 0.00}$	3.36 ± 1.04	1.06 ± 0.06	$\mathbf{1.00 \pm 0.00}$	0.95 ± 0.23	2.95 ± 1.28	$\mathbf{1.03 \pm 0.06}$	0.93 ± 0.21	$\mathbf{1.00 \pm 0.00}$	3.40 ± 1.07	1.04 ± 0.05	$\mathbf{1.00 \pm 0.00}$
ROAR	$\mathbf{1.00 \pm 0.00}$	1.84 ± 0.72	1.10 ± 0.12	0.99 ± 0.06	$\mathbf{1.00 \pm 0.00}$	1.69 ± 0.58	1.10 ± 0.13	0.87 ± 0.25	$\mathbf{1.00 \pm 0.00}$	1.56 ± 0.59	1.10 ± 0.14	0.96 ± 0.07
T-Rex	0.76 ± 0.42	$\mathbf{1.04 \pm 0.25}$	1.09 ± 0.16	0.64 ± 0.43	0.70 ± 0.46	$\mathbf{0.82 \pm 0.35}$	1.10 ± 0.18	0.58 ± 0.43	0.73 ± 0.45	$\mathbf{0.90 \pm 0.31}$	1.10 ± 0.18	0.64 ± 0.43
ProPlace	0.98 ± 0.15	1.72 ± 0.73	$\mathbf{1.04 \pm 0.08}$	0.90 ± 0.20	0.99 ± 0.11	1.76 ± 0.77	1.04 ± 0.07	0.88 ± 0.18	0.98 ± 0.15	1.64 ± 0.79	$\mathbf{1.03 \pm 0.09}$	0.87 ± 0.19
HPF (Hessian)	$\mathbf{1.00 \pm 0.00}$	2.62 ± 1.03	1.14 ± 0.15	$\mathbf{1.00 \pm 0.00}$	1.00 ± 0.07	2.77 ± 1.11	1.12 ± 0.12	0.96 ± 0.19	$\mathbf{1.00 \pm 0.00}$	2.45 ± 1.02	1.11 ± 0.15	$\mathbf{1.00 \pm 0.00}$
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	2.48 ± 1.02	1.13 ± 0.14	1.00 ± 0.01	$\mathbf{1.00 \pm 0.00}$	2.86 ± 1.21	1.10 ± 0.12	1.00 ± 0.01	$\mathbf{1.00 \pm 0.00}$	2.30 ± 0.98	1.12 ± 0.15	1.00 ± 0.01
HPF (FIM)	$\mathbf{1.00 \pm 0.00}$	2.35 ± 1.00	1.13 ± 0.14	1.00 ± 0.02	$\mathbf{1.00 \pm 0.00}$	2.77 ± 1.20	1.09 ± 0.12	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.19 ± 0.92	1.09 ± 0.14	1.00 ± 0.01

Retrain - Australian

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	$\mathbf{1.00 \pm 0.00}$	4.01 ± 1.09	1.04 ± 0.07	1.00 ± 0.04	0.98 ± 0.15	4.04 ± 1.16	1.04 ± 0.07	0.97 ± 0.16	0.99 ± 0.12	3.83 ± 1.07	1.04 ± 0.07	0.98 ± 0.12
ROAR	$\mathbf{1.00 \pm 0.00}$	2.33 ± 0.54	1.10 ± 0.15	1.00 ± 0.01	$\mathbf{1.00 \pm 0.00}$	2.29 ± 0.47	1.09 ± 0.15	0.97 ± 0.08	$\mathbf{1.00 \pm 0.00}$	2.02 ± 0.45	1.09 ± 0.14	0.99 ± 0.03
T-Rex	0.82 ± 0.39	$\mathbf{1.40 \pm 0.36}$	1.08 ± 0.14	0.69 ± 0.39	0.52 ± 0.50	$\mathbf{0.89 \pm 0.36}$	1.07 ± 0.14	0.40 ± 0.42	0.80 ± 0.40	$\mathbf{1.19 \pm 0.36}$	1.07 ± 0.09	0.68 ± 0.40
ProPlace	$\mathbf{1.00 \pm 0.00}$	2.84 ± 0.86	$\mathbf{1.03 \pm 0.05}$	0.99 ± 0.06	$\mathbf{1.00 \pm 0.00}$	2.89 ± 0.77	$\mathbf{1.02 \pm 0.05}$	0.99 ± 0.04	$\mathbf{1.00 \pm 0.00}$	2.68 ± 0.78	$\mathbf{1.03 \pm 0.05}$	1.00 ± 0.02
HPF (Hessian)	$\mathbf{1.00 \pm 0.00}$	3.00 ± 0.89	1.12 ± 0.11	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.43 ± 1.04	1.10 ± 0.11	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.97 ± 0.86	1.11 ± 0.09	$\mathbf{1.00 \pm 0.00}$
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	2.55 ± 0.86	1.13 ± 0.11	0.96 ± 0.10	$\mathbf{1.00 \pm 0.00}$	3.25 ± 0.90	1.15 ± 0.10	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	2.50 ± 0.79	1.12 ± 0.09	0.97 ± 0.07
HPF (FIM)	$\mathbf{1.00 \pm 0.00}$	2.54 ± 0.88	1.13 ± 0.11	0.96 ± 0.12	$\mathbf{1.00 \pm 0.00}$	3.26 ± 0.89	1.14 ± 0.10	1.00 ± 0.01	$\mathbf{1.00 \pm 0.00}$	2.55 ± 0.83	1.12 ± 0.09	0.99 ± 0.04

Retrain - German

	Clean				Symmetric Noise				Asymmetric Noise			
Method	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑	Val↑	L2↓	LOF↓	RF↑
ElliCE	$\mathbf{1.00 \pm 0.00}$	7.36 ± 1.68	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$	0.99 ± 0.08	7.63 ± 1.69	$\mathbf{1.01 \pm 0.07}$	0.99 ± 0.09	$\mathbf{1.00 \pm 0.00}$	7.05 ± 1.63	$\mathbf{1.01 \pm 0.07}$	$\mathbf{1.00 \pm 0.00}$
ROAR	0.67 ± 0.47	2.33 ± 1.17	1.11 ± 0.13	0.50 ± 0.35	0.29 ± 0.46	1.42 ± 1.39	1.12 ± 0.15	0.24 ± 0.28	0.01 ± 0.09	0.83 ± 1.07	1.11 ± 0.12	0.14 ± 0.21
T-Rex	0.59 ± 0.49	$\mathbf{0.91 \pm 0.27}$	1.10 ± 0.12	0.44 ± 0.43	0.52 ± 0.50	$\mathbf{0.65 \pm 0.28}$	1.10 ± 0.13	0.39 ± 0.39	0.58 ± 0.49	$\mathbf{0.76 \pm 0.25}$	1.10 ± 0.12	0.45 ± 0.42
ProPlace	0.96 ± 0.20	6.94 ± 1.64	1.02 ± 0.06	1.00 ± 0.03	0.97 ± 0.17	7.13 ± 1.63	1.03 ± 0.07	0.98 ± 0.07	0.96 ± 0.20	6.83 ± 1.65	1.02 ± 0.07	1.00 ± 0.01
HPF (Hessian)	$\mathbf{1.00 \pm 0.00}$	3.34 ± 0.68	1.08 ± 0.05	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.20 ± 0.69	1.08 ± 0.05	$\mathbf{1.00 \pm 0.00}$	$\mathbf{1.00 \pm 0.00}$	3.15 ± 0.61	1.08 ± 0.05	$\mathbf{1.00 \pm 0.00}$
HPF (Bootstrap)	$\mathbf{1.00 \pm 0.00}$	7.05 ± 1.67	$\mathbf{1.01 \pm 0.07}$	1.00 ± 0.02	$\mathbf{1.00 \pm 0.00}$	7.15 ± 1.66	$\mathbf{1.01 \pm 0.07}$	1.00 ± 0.03	$\mathbf{1.00 \pm 0.00}$	6.79 ± 1.71	$\mathbf{1.01 \pm 0.07}$	1.00 ± 0.03
HPF (FIM)	$\mathbf{1.00 \pm 0.00}$	6.84 ± 1.61	$\mathbf{1.01 \pm 0.07}$	0.99 ± 0.04	$\mathbf{1.00 \pm 0.00}$	7.09 ± 1.77	$\mathbf{1.01 \pm 0.07}$	0.99 ± 0.05	$\mathbf{1.00 \pm 0.00}$	6.79 ± 1.56	$\mathbf{1.01 \pm 0.07}$	1.00 ± 0.02