RECOURSE ON BORROWED TIME: COUNTERFACTUALS THAT SURVIVE A MODEL UPDATE

Anonymous authors
Paper under double-blind review

ABSTRACT

Recourse has to respect what a person can actually change: age and sex are fixed, income moves only one way, employment type is categorical. The methods that honour such constraints exactly are mixed-integer programs, and we examine how their recommendations fare when the model changes. The leading exact method asks only that the decision flip, and it shows: under its shipped validity threshold a third of its counterfactuals sit exactly on the decision boundary, and only about half survive an ordinary retrain of the deployed classifier. That threshold is a parameter, and raising it repairs both, at the cost of a mixed-integer solve per individual. We introduce **PACER**, which obtains the same guarantee at gradient cost. It compiles a dataset's metadata into a symbolic rule base, projects every gradient step onto it, and optimises to a confidence margin the user sets. Plausibility comes from a sum–product network distilled into a differentiable surrogate: the exact circuit scores every result while the surrogate only supplies the search direction. PACER is the only gradient-based method in our study whose counterfactuals are fully actionable on all three benchmarks, and it keeps 0.96–1.00 of them valid after a model update, at full coverage, in one to two seconds each. We report what this costs as well as what it buys, and release a harness that reproduces the exact baseline's published numbers cell for cell.

1 INTRODUCTION

Machine learning models increasingly gate access to credit, employment and housing, and the people affected need to know what to change (Dwivedi et al., 2023; Gunning et al., 2021). Counterfactual explanations answer that question directly: they describe a modified input that would receive the favourable outcome (Wachter et al., 2017). A usable recommendation must be *actionable* — it cannot ask someone to change their age or race, and it must respect admissible ranges and monotonic directions (Ustun et al., 2019; Karimi et al., 2021) — and it should be *plausible*, lying in a well-supported region of the data distribution (Karimi et al., 2022).

Recourse is almost always posed as a constrained minimisation: find the closest point that flips the decision (Ustun et al., 2019; Kanamori et al., 2020; Němeček et al., 2025). That objective has a well-known consequence — the solution lies *on* the decision boundary, where the classification margin is essentially zero — and a literature has grown around the fragility that follows (Hamman et al., 2023; Kostrzewa et al., 2026; Jiang et al., 2024c). That confidence helps is not our claim; it is the premise of that literature.

The leading exact method has not been examined. Real recourse is constrained — age and sex cannot change, income cannot fall if the advice is to save more, employment type is categorical — and advice that violates this is not recourse; DiCE reaches 0.95 validity at 0.01 actionability on Credit (§5.5). Mixed-integer programs encode such constraints exactly (Ustun et al., 2019; Jiang et al., 2024a; Němeček et al., 2025). Robustness work has mostly omitted them (Hamman et al., 2023; Kostrzewa et al., 2026); PROPLACE (Jiang et al., 2024a) adds them inside a mixed-integer program and ElliCE (Turbal et al., 2026) by gradient descent. Neither evaluates LiCE (Němeček et al., 2025), the exact method with circuit-based plausibility, and to our knowledge how its recommendations fare under model change has not been measured.

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.

What we find. The leading exact method for this setting fixes its validity constraint at $sf_{\boldsymbol{w}}(x') \geq 10^{-4}$, asking only that the decision flip. Under that setting roughly a third of its counterfactuals have signed margin exactly zero, and only about half survive a retrain (§5.2). The threshold is a parameter and raising it repairs both, at the cost of a mixed-integer solve per individual (§5.4). This paper asks whether the same guarantee is available at gradient cost, inside the actionable set, on mixed-type data.

Confidence as a first-class parameter. We make the confidence level explicit. **PACER** — *Projected Actionable Counterfactuals with Enforced Robustness* — sets the validity constraint at a user-specified margin m rather than at the boundary and solves the problem as one differentiable objective: a margin hinge under an adaptive penalty, a projection onto the actionable set after every gradient step, and a deterministic softmax relaxation for categorical actions. Plausibility follows what we call an *exact-score, distilled-gradient* design: a sum–product network's exact likelihood scores every counterfactual we report, while a surrogate distilled from it supplies only the search direction. The circuit thus steers a first-order method without being embedded in a mixed-integer program (Němeček et al., 2025) or replaced by a differentiable density that never yields its likelihood.

What the margin buys. The margin behaves as a dial with a measurable price: raising it moves retention under a model update towards one on all three benchmarks along a smooth frontier against proximity whose slope depends on the data (§5.2, §5.3). PACER is also the only gradient-based method in our study that returns fully actionable counterfactuals on all three benchmarks (§5.5).

Contributions. The two literatures that bear on this problem use incompatible machinery: robustness methods optimise unconstrained gradients over continuous features, while actionable methods solve integer programs. Our first contribution is a diagnosis of the second. The leading mixed-integer method for actionable recourse loses about half its counterfactuals to a routine retrain, and the fragility is not intrinsic to exact optimisation: it follows from one shipped default, a validity threshold of 10^{-4}, under which a third of its solutions sit on the boundary and none do once the threshold is raised (§5.2, §5.4). Alongside that measurement we report retention for eight methods with the population each figure is computed over, which separates robustness earned by abstaining from robustness across everyone. PACER itself contributes two mechanisms. A symbolic rule base, compiled from each dataset's metadata, drives a closed-form projection after every gradient step and is shared with every baseline so that actionability is measured under one feasible set; PACER matches the mixed-integer method at full actionability on all three benchmarks where the other baselines reach 0.15–0.71 (§3, §5.5). Plausibility follows an exact-score, distilled-gradient design: a sum–product network gives exact mixed-type likelihoods but no usable gradient, so rather than embed it in an integer program, as LiCE does, replace it with a flow that never yields the circuit's likelihood, as PPCEF and CROCE do, or restrict solutions to observed samples, as ElliCE does, we let a distilled surrogate supply the search direction while the exact circuit scores every reported number (§3.3, Appendix C).

2 Related Work

Actionable recourse. Actionable recourse restricts changes to user-controllable variables via integer programming (Ustun et al., 2019); causal formulations prefer minimal interventions (Karimi et al., 2021), surveyed in Karimi et al. (2022). The constraints that make a counterfactual enactable — immutability, monotone directions, admissible ranges, categorical value sets — are naturally integer constraints, and the methods that encode them exactly inside the optimisation are mixed-integer programs (Ustun et al., 2019; Němeček et al., 2025). Gradient-based and generative-model-based methods (Mothilal et al., 2020; Pawelczyk et al., 2020; Duong et al., 2023; Wielopolski et al., 2024; Dombrowski et al., 2024) are faster and vary in how fully they enforce such constraints; ElliCE (Turbal et al., 2026) projects onto immutability and range constraints during optimisation. §5.5 measures actionability for every method we evaluate under one shared constraint set.

Robustness of recourse to model change. That recourse can be invalidated by model change is established, and Jiang et al. (2024c) organise the area. Rawal et al. (2021) study how data and model shifts invalidate recourse, and Upadhyay et al. (2021) propose ROAR, which optimises recourse

Table 1: **The three closest methods.** ElliCE shares PACER's combination of actionability, mixed types and a robustness target without a solver; the methods differ in what the robustness target requires and in where plausibility comes from. Full comparison in Table 11.

Method	Action.	Mixed	Solver-free	Robustness target needs	Plausibility from
CROCE (Kostrzewa et al., 2026)	—	—	✓	40-model ensemble + flow	normalising flow
ElliCE (Turbal et al., 2026)	✓	✓	✓	training-loss Hessian	observed samples
LiCE (Němeček et al., 2025)	✓	✓	—	—(no target)	exact circuit, in a MIP
PACER	✓	✓	✓	classifier logit only	exact circuit, distilled gradient

to remain valid under bounded perturbations of the model parameters. Hamman et al. (2023) define *naturally-occurring* model change, which permits arbitrary changes in the parameters provided predictions on the data manifold change little, and test it by retraining with different weight initialisations among other changes. They prove a probabilistic validity guarantee for a stability measure, and their T-Rex methods accept a counterfactual once a practical relaxation of that measure, combining the model's output around it with its local variability, exceeds a user-set threshold; Stępka et al. (2025) give probabilistic guarantees over a space of admissible retrainings; and Kostrzewa et al. (2026) condition a normalising flow on ensemble consensus so that one parameter sets the robustness level at inference. Certified defences use robust optimisation (Jiang et al., 2024a), interval abstraction (Jiang et al., 2024b) and probabilistic guarantees (Marzari et al., 2026).

Our position. We do not claim that confidence confers robustness. Hamman et al. (2023) treat a higher model output as a desirable property, following Dutta et al. (2022), but caution that it does not suffice where the output varies locally; PACER's margin is pointwise, and §5.2 measures how much robustness it buys. Nor are we the first to combine robustness with actionability on mixed-type data: PROPLACE (Jiang et al., 2024a) does so inside a mixed-integer program, and ElliCE (Turbal et al., 2026), the closest prior work, optimises counterfactuals over an ellipsoidal approximation of the Rashomon set, projecting onto immutability and range constraints during gradient descent and relaxing categorical features with a Gumbel-softmax. Table 1 shows where we differ: a margin needs only the deployed classifier's logit where ElliCE's ellipsoid needs the training-loss Hessian, and our plausibility comes from a circuit's exact likelihood rather than observed samples. Above all, our evaluation is organised around the exact mixed-integer formulation for actionable recourse, which ElliCE does not compare against; a preliminary run of it in our harness is reported in Appendix M. Khotanlou et al. (2026) provide a reproducible evaluation framework, whereas our harness is verified against an external reference (§4).

3 PACER

PACER produces recourse that is valid to a user-specified confidence margin by minimising a single differentiable objective with projected gradient descent. Unlike exact combinatorial recourse it needs no mixed-integer solver, and it handles numeric and categorical actions jointly. Appendix A collects the notation.

3.1 Problem setup

Let $x^0 \in \mathcal{X}$ be a factual instance with D features split into numeric indices $\mathcal{N}$ and categorical indices $\mathcal{C}$. A fixed encoder $\phi : \mathcal{X} \to \mathbb{R}^p$ min–max normalises numeric features and one-hot encodes categorical ones. We are given a *frozen* binary classifier with logit $f_{\boldsymbol{w}} \circ \phi$, never updated during optimisation. For desired class $y^\star$ we set the target sign $s = +1$ if $y^\star = 1$ and $s = -1$ otherwise, and define the *signed margin* $s\, f_{\boldsymbol{w}}(x')$; the candidate x' receives the desired class iff $s\, f_{\boldsymbol{w}}(x') \geq 0$.

The actionable set, and where it comes from. Recourse must be actionable (Ustun et al., 2019):

$$\mathcal{A}(x^0) = \big\{\, x' :\; x'_j = x^0_j \;\forall j \in \mathcal{I};\;\; \ell_j \leq x'_j \leq u_j \;\forall j \in \mathcal{N};\;\; \rho_j (x'_j - x^0_j) \geq 0 \;\forall j \in \mathcal{N} \big\}, \qquad (1)$$

where $\mathcal{I}$ indexes immutable features, $[\ell_j, u_j]$ is the admissible range of numeric feature j, and $\rho_j \in \{-1, 0, +1\}$ encodes a monotonicity requirement. Stating (1) is easy; obtaining it for a real dataset and keeping a gradient method inside it is the part that has previously required integer programming.

We therefore build $\mathcal{A}(x^0)$ from a *symbolic rule base* $\mathcal{K}$: one rule per feature, compiled from the dataset's metadata rather than written by hand. There are four kinds. IMMUTABLE marks a feature that cannot change and drops it from the optimisation variables. NUMERIC carries $[\ell_j, u_j]$, the monotone direction ρ_j and an integer flag. CATEGORICAL holds the admissible values as a transition graph, and ORDINAL is its monotone special case for ranked categories such as education level. A feature's mutability flag, declared type and recorded bounds select among them, so the rule base is produced automatically for each dataset (Appendix D). Those four quantities are exactly what the optimiser needs: they make the closed-form projection of Eq. (10) possible and let the categorical relaxation of Eq. (3) range over the right value set. The same $\mathcal{K}$ is applied to PACER and to every re-scored baseline, so all methods share one feasible set. Because the projection is applied after every gradient step, *every iterate* is a change the person could enact — the property that separates PACER from the gradient baselines in Table 4.

The recourse problem. A counterfactual in its classical form is the closest point that flips the decision (Wachter et al., 2017); recourse adds actionability (Ustun et al., 2019) and, often, plausibility (Karimi et al., 2022). We replace the flip with a *margin*:

$$\min_{x' \in \mathcal{A}(x^0)} \; d(x^0, x') - \beta \log p_{\boldsymbol{\theta}}\big(\phi(x')\big) \qquad \text{s.t.} \qquad s\, f_{\boldsymbol{w}}(x') \geq m. \tag{2}$$

Here d is a distance (MAD-scaled ℓ_1; exact form in Eq. (11)), $p_{\boldsymbol{\theta}}$ is a plausibility density (§3.3), $\beta \geq 0$ weights it, and $m \geq 0$ is the confidence margin. Setting $m = 0$ recovers ordinary boundary crossing. The whole paper concerns what $m > 0$ buys and what it costs.

3.2 DIFFERENTIABLE PARAMETERISATION

PACER optimises $\boldsymbol{\Theta} = \big(\{\delta_j\}_{j \in \mathcal{N}}, \{\boldsymbol{z}^{(c)}\}_{c \in \mathcal{C}}\big)$: additive numeric perturbations and unconstrained categorical logits.

Numeric features. Each numeric feature is $x'_j = x^0_j + \delta_j$, where δ_j is a free perturbation. After every gradient step we project δ back onto $\mathcal{A}(x^0)$ in closed form using the quantities read from $\mathcal{K}$: immutable features are held fixed, monotonicity-violating moves are zeroed, and values are clipped to $[\ell_j, u_j]$ (Eq. (10)).

Categorical features. Discrete choices are the obstacle to a gradient method. We relax each categorical feature into a temperature-controlled distribution over its K_c values,

$$\boldsymbol{\pi}^{(c)} = \mathrm{softmax}\big(\boldsymbol{z}^{(c)}/\tau\big), \tag{3}$$

where $\boldsymbol{z}^{(c)} \in \mathbb{R}^{K_c}$ are learnable logits and $\tau > 0$ is a temperature. We place $\boldsymbol{\pi}^{(c)}$ in the one-hot block of $\phi(x')$ so gradients flow to $\boldsymbol{z}^{(c)}$. The relaxation is *deterministic*: unlike Gumbel–Softmax (Jang et al., 2016; Maddison et al., 2016) we draw no noise, trading a small bias for lower gradient variance. At decode time we take the hard category

$$x'_c = v_{c,k^\star}, \qquad k^\star = \arg\max_k \; \pi^{(c)}_k, \tag{4}$$

and re-verify validity on the hard-decoded instance, so no reported number depends on the relaxation. Initialising $\boldsymbol{z}^{(c)}$ at the factual category (Appendix B) starts optimisation from "no change", so categorical parsimony arises for free.

3.3 OBJECTIVE

Two quantities we care about are not differentiable as stated, so the objective uses surrogates for the *search direction* only; all *reported* metrics use the exact quantity on the hard-decoded counterfactual.

Proximity surrogate. The reported distance charges a per-change cost on categoricals whose gradient vanishes, so we differentiate only numeric displacement, in encoder units:

$$\tilde{d}\big(x^0, x'\big) = \frac{1}{|\mathcal{N}|} \sum_{j \in \mathcal{N}} \frac{|\delta_j|}{\sigma_j}, \tag{5}$$

where $\sigma_j > 0$ is the encoder's min–max scale for feature j and $|\mathcal{N}|$ the numeric-feature count. This drives δ toward small edits; the exact MAD distance d is what we report.

Algorithm 1 PACER: Projected Actionable Counterfactuals with Enforced Robustness (single counterfactual)

Require: factual x^0; target sign s; frozen $f_{\boldsymbol{w}}$; density surrogate $g_{\boldsymbol{\psi}}$; constraints $\mathcal{A}(x^0)$; margin m
hyperparameters: weights λ_d, λ_p; escalation $(\lambda_v, \eta_\lambda, \lambda_v^{\max})$; temperature τ; step α; iterations T
Ensure: actionable counterfactual x'
Initialise at the factual: $\delta_j \leftarrow 0$ for $j \in \mathcal{N}$; $\boldsymbol{z}^{(c)} \leftarrow$ factual category for $c \in \mathcal{C}$
for $t = 1, \ldots, T$ **do**
  build $\phi(x')$: numeric $x_j^0 + \delta_j$; categorical block $\leftarrow \mathrm{softmax}(\boldsymbol{z}^{(c)}/\tau)$
  $\mathcal{L} \leftarrow \lambda_v [m - s f_{\boldsymbol{w}}(x')]_+ + \lambda_d \tilde{d}(x^0, x') - \lambda_p g_{\boldsymbol{\psi}}(\phi(x'))$
  $\boldsymbol{\Theta} \leftarrow \boldsymbol{\Theta} - \alpha \nabla_{\boldsymbol{\Theta}} \mathcal{L}$ (Adam step)
  project δ onto $\mathcal{A}(x^0)$ (immutable / monotone / box, Eq. (10))
  $\lambda_v \leftarrow \min(\lambda_v + \eta_\lambda [m - s f_{\boldsymbol{w}}(x')]_+, \lambda_v^{\max})$
end for
decode: $\arg\max$ categoricals; sparsify + round numeric edits
return x' (validity re-verified on hard-decoded x')

Plausibility surrogate. Plausibility uses a sum–product network (SPN) (Poon & Domingos, 2011) in the mixed-type form of Molina et al. (2018), $p_{\boldsymbol{\theta}}$, which gives exact likelihoods on mixed data but is not differentiable through our autograd stack. We distil it into a small MLP $g_{\boldsymbol{\psi}}$ by least-squares regression onto $\log p_{\boldsymbol{\theta}}$ (Eq. (9)); $g_{\boldsymbol{\psi}}$ supplies the plausibility gradient, while every reported NLL uses the exact circuit.

Margin-controlled Lagrangian. PACER minimises

$$\mathcal{L}(\boldsymbol{\Theta}; \lambda_v) = \underbrace{\lambda_v \big[m - s\, f_{\boldsymbol{w}}(x')\big]_+}_{\text{validity}} + \underbrace{\lambda_d\, \tilde{d}(x^0, x')}_{\text{proximity}} - \underbrace{\lambda_p\, g_{\boldsymbol{\psi}}\big(\phi(x')\big)}_{\text{plausibility}}, \tag{6}$$

where $[t]_+ = \max(0, t)$ is a hinge, $\lambda_v \geq 0$ is the (adapted) validity weight, and $\lambda_d, \lambda_p \geq 0$ are fixed proximity and plausibility weights (λ_d=2, λ_p=0.1). The hinge charges a penalty *only* while the signed margin falls short of m; once the margin is met the validity term is inert and the objective simply keeps x' close and plausible.

Adaptive penalty escalation. Rather than fix λ_v, we grow it in proportion to the current violation:

$$\lambda_v \;\leftarrow\; \min\Big(\lambda_v + \eta_\lambda \big[m - s\, f_{\boldsymbol{w}}(x')\big]_+,\; \lambda_v^{\max}\Big), \tag{7}$$

with escalation rate $\eta_\lambda = 0.05$ and cap $\lambda_v^{\max} = 20$. Because the increment is a non-negative hinge, λ_v is monotonically non-decreasing: it rises while the margin is unmet, freezes once met, and saturates at the cap. We therefore call this *penalty escalation*, not dual ascent — a true dual step would use the signed violation and could relax λ_v on satisfied iterates, inducing oscillation between the validity and proximity terms. The monotone rule avoids that, while the cap prevents validity from overwhelming proximity and plausibility on instances that cannot reach m.

3.4 Algorithm and Decoding

Algorithm 1 assembles the pieces: each iteration builds the relaxed encoding, evaluates (6), takes an Adam (Kingma & Ba, 2014) step, projects the numeric perturbations, and escalates λ_v via (7). At decode we suppress numeric edits below a relative threshold $\kappa = 0.03$, round integer features, take the $\arg\max$ category (4), and re-verify validity (Appendix B, Eq. (8)). The loop is solver-free and uses only standard automatic differentiation.

4 Experimental Setup

We evaluate on GMSC (Fusion & Cukierski, 2011), Adult (Becker & Kohavi, 1996) and German Credit (Hofmann, 1994) with 5-fold cross-validation and 100 factual instances per fold; for each

factual the target is the complement of the classifier's predicted class. Every method is scored in one harness: the released pre-trained classifier used unmodified, the fitted sum–product network for plausibility, the MAD-scaled ℓ_1 distance of Eq. (11), and the shared actionable set of Eq. (1). Baselines — DiCE, VAE, CVAE, two FACE variants, PROPLACE, and the LiCE *optimize* and *median* operating points — are taken from LiCE's released counterfactuals and re-scored identically. Sparsity uses the same definition as the exact baseline, `isclose` on the normalised non-one-hot encoding, so that our numbers are directly comparable to the published ones.

Harness fidelity. Our re-scoring recovers the exact baseline's published results: validity $1.00/1.00/0.99$ against a reported $100.0/100.0/99.4$, sparsity $2.56/2.07/2.11$ against $2.6/2.1/2.3$, and exact-circuit NLL $25.6/18.1/28.9$ against $25.6/18.1/28.8$, on GMSC / Adult / Credit. Their published numbers are reproduced as Table 19; ours are the LiCE (opt.) rows of Table 12 and Table 4, where actionability is 1.00 on all three. Every comparison below therefore rests on a harness verified against an external reference rather than on a re-implementation. Appendix N states the artefact conditions required to reach this agreement, which we document because they are easy to violate and consequential: scoring boundary-hugging counterfactuals under an independently retrained classifier moves measured validity by up to 50 points.

5 Results

5.1 Boundary-targeted recourse has no headroom

A counterfactual is *valid at threshold* γ if the classifier assigns the target class with signed logit exceeding γ; $\gamma = 0$ recovers ordinary validity and $\gamma > 0$ demands a confident prediction. We re-score every method's released counterfactuals at increasing γ (Appendix F, Figure 3).

Exact recourse is valid for essentially every individual at the boundary and for almost none just beyond it: on GMSC all but a few percent of its counterfactuals fail once a quarter of a logit is required, and the decay on Adult and Credit is slower but of the same shape. PACER holds out to $\gamma \approx m$, which is what the margin is for. The contrast is in headroom, of which a boundary-targeted objective has none.

5.2 Recourse under a model update

Confidence thresholds are a proxy. The deployment question is whether recourse survives the model being retrained. Our protocol is an instance of the naturally-occurring model change of Hamman et al. (2023), who likewise retrain with different weight initialisations, and matches the *Random Retrain* evaluator of Turbal et al. (2026) except that they keep only retrained models whose loss stays inside a Rashomon bound, whereas we score against whatever the retrain produces. We train an independent classifier per fold — same data, same architecture, different seed — and re-score against it. Entries in Table 2 are the fraction of *originally valid* counterfactuals that remain valid.

Prior work establishes that recourse can be invalidated by model shift and proposes certificates against it. What has not been reported is how the dominant *objective* behaves under the most ordinary shift of all. Table 2 measures it: exact recourse, provably minimal and valid for every individual, retains roughly half after a retrain on identical data. The failure follows from targeting the boundary, not from any implementation.

The n column separates what the literature conflates. Exact recourse is the only family combining near-complete coverage with retention near chance; every method that retains well either covers fewer individuals — the certified robust baseline and the VAE cover about a quarter of the population on Credit — or sits far from the boundary, as DiCE does at three times PACER's distance on the mixed-type benchmarks (Table 12). Displacement alone is not sufficient: FACE has full coverage on GMSC and moves further than PACER at $m{=}0.2$, yet loses a quarter of its counterfactuals. PACER reaches high retention at full coverage, and is the only method here whose level is chosen in advance rather than inherited, with retention rising monotonically in m while coverage rises with it.

The spread matters as much as the mean. Per-fold retention for exact recourse on GMSC ranges from roughly a fifth to four fifths, so a practitioner cannot predict how much of a deployment's recourse survives a retrain, only that a large share will not; FACE is equally unstable. PACER's

Table 2: **Recourse under a model update.** Fraction of originally-valid counterfactuals that remain valid when the classifier is retrained on the same data with a different seed (mean $\pm$ std over five folds). n is the mean number of originally-valid counterfactuals per fold — the population each retention figure is computed over; cells with $n < 10$ are too small to estimate and are omitted. Red marks retention ≥ 0.995. Read together, the two columns separate retention earned by covering few individuals well from retention across the whole population, and the spread separates methods whose retention is predictable from those whose is not. The full six-margin sweep is in Appendix H.

	GMSC		Adult		Credit	
Method	retained	n	retained	n	retained	n
DiCE	1.00±0.00	100	0.93±0.07	100	0.80±0.07	95
VAE	–	1	0.87±0.12	80	1.00±0.00	28
CVAE	0.88±0.17	85	–	9	–	1
FACE$_k$	0.74±0.27	99	0.86±0.09	80	0.88±0.10	28
FACE$_\epsilon$	0.75±0.26	99	0.85±0.09	63	0.85±0.10	28
PROPLACE	1.00±0.00	99	1.00±0.00	80	1.00±0.00	28
LiCE (opt.)	0.48±0.28	100	0.58±0.12	100	0.55±0.09	99
LiCE (med.)	0.56±0.21	60	0.72±0.06	91	0.53±0.08	99
PACER (m=0.2)	0.84±0.16	96	0.94±0.03	80	0.95±0.01	95
PACER (m=1.0)	1.00±0.00	100	0.96±0.02	94	0.99±0.01	99
PACER (m=1.5)	1.00±0.00	100	0.98±0.03	98	1.00±0.00	100

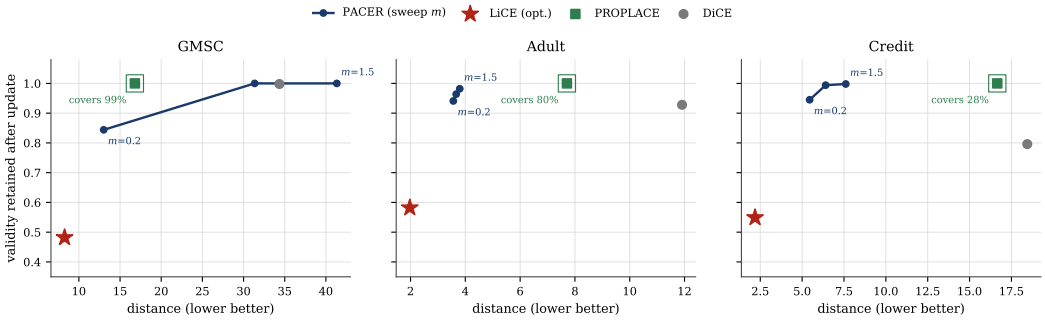


Figure 1: **The robustness/cost frontier.** Validity retained after a model update against MAD-scaled ℓ_1 distance. PACER is a sweep over $m \in [0.2, 1.5]$; on the two mixed-type benchmarks it reaches full retention at roughly half the distance of the certified robust baseline, at three times its coverage on Credit.

spread is an order of magnitude smaller at every margin except the smallest on GMSC, where the margin is too small to lift counterfactuals clear of the boundary.

5.3 The robustness/cost frontier

Figure 1 places every method on the robustness/cost plane. The frontier PACER traces is smooth and monotone in m on every dataset, which is what makes the parameter usable: a practitioner who knows what an invalid recommendation costs can read off the margin that matches it, rather than accepting whatever robustness a method happens to supply. Its slope varies with the data in a way that evaluations on continuous features alone cannot reveal. On Adult the entire sweep costs under twice the distance of exact recourse, consistent with the softmax relaxation spending the margin on a categorical change instead of stretching a continuous feature, while on the fully-continuous GMSC the same confidence must be bought with continuous displacement and the cost grows several-fold. Credit is also mixed-type yet sits between the two, so feature type is not the whole explanation; Figure 2 in Appendix F isolates the cost relative to exact recourse.

Table 12 in Appendix G reports the full within-harness comparison at the decision boundary. PACER matches exact recourse on validity over the fifteen dataset–fold pairs (paired Wilcoxon $p = 0.11$)

Table 3: **Sweeping the exact method's validity threshold** on German Credit (3 folds × 25 factuals, mean ± std). *On-boundary*: signed margin below 10^{-3}. *Retained*: still valid after a retrain with a different seed. Intermediate thresholds are in the text; one instance of 75 is infeasible at every threshold, so solve rate does not depend on τ.

τ	on-boundary	retained	distance	sec/instance
10^{-4} (shipped)	0.32±0.10	0.57±0.08	2.52±0.58	31±10
1.5	0.00±0.00	0.96±0.04	4.98±1.64	33±19
PACER (m=1.0)	0.00	0.99±0.01	6.40±0.28	**2.2**

and pays for the margin in proximity. At $\gamma = 0$ an exact solver is by construction the closest point available, so that comparison measures the price of the margin rather than the quality of the optimiser.

5.4 Inside the solver: sweeping the validity threshold

The mixed-integer formulation encodes validity as $sf_{\boldsymbol{w}}(x') \geq \tau$ and ships with $\tau = 10^{-4}$: cross the boundary, no further. Because τ is already a parameter of the program, the margin can be demanded of the exact method directly. We sweep it on German Credit (3 folds, 25 factuals per fold, HiGHS, 300 s cap) and score each returned counterfactual both on the deployed classifier and on the independently retrained one of §5.2 (Table 3).

Three things follow. The mechanism is visible in the solver's own output: at the shipped threshold about a third of solutions have signed margin indistinguishable from zero, and at every $\tau > 0$ that fraction is exactly zero. Boundary-activity is not inferred from re-scoring curves; it is what the program returns when asked only to cross. The repair is a single constant, with retention rising monotonically in τ at a predictable cost in proximity and no growth in solve time. And the sweep delimits our claim about PACER precisely: exact optimisation at a positive threshold is a valid route to margin-robust actionable recourse, and on this dataset a closer one than ours, bought with tens of seconds of mixed-integer solve per individual and a solver dependency. We therefore position PACER not as the only way to obtain a margin under actionability constraints, but as the way to obtain one at gradient cost. Timings use HiGHS and are not comparable to the Gurobi figures in Table 13.

5.5 Actionability

Validity alone is a weak notion of usefulness: a counterfactual that flips the decision by altering an immutable attribute is not recourse at all. Scoring every method against the shared rule base of Eq. (1) (Table 4; scatter view in Appendix I) separates the field sharply. PACER returns fully actionable counterfactuals on all three datasets, matching the mixed-integer method that enforces the constraints combinatorially; every other baseline lands between 0.15 and 0.71. A preliminary run of ElliCE (Turbal et al., 2026), the closest prior method, places it in the same band (0.16–0.67) for the same reason: its data-supported variant returns observed training points, which need not preserve the factual's immutable attributes (Appendix M). DiCE on Credit is the sharpest illustration: 0.95 valid, 0.01 actionable — it achieves near-perfect validity by changing attributes nobody can change. PACER reaches this by projecting onto the constraint set after every step; §6 shows which constraints matter. We do not compare against enforcing the same constraints post hoc.

6 Ablations

The rule base. Table 5 disables one rule family at a time. Immutability is the binding constraint everywhere; the bounds matter only on GMSC, where they also carry validity; monotonicity is inert. Not every rule family earns its place on these benchmarks.

The distilled circuit and the cost of a margin. Table 6 replaces the circuit with a Gaussian mixture or with nothing: on the mixed-type benchmarks the circuit leads at both margins, while on

Table 4: **Actionability** of valid counterfactuals under the shared rule base (5-fold means). Only PACER and the mixed-integer method are fully actionable everywhere.

Method	GMSC	Adult	Credit
DiCE	0.41	0.43	0.01
VAE	0.25	0.31	0.00
CVAE	0.25	0.71	0.50
FACE_k	0.15	0.57	0.40
FACE_ϵ	0.16	0.50	0.37
PROPLACE	0.35	0.69	0.34
LiCE (opt.)	1.00	1.00	1.00
LiCE (med.)	1.00	1.00	1.00
PACER	1.00	1.00	1.00

Table 5: **What each rule family contributes.** PACER at m=1.0 with one rule family disabled at a time; Act is measured against the *full* rule base. Immutability matters on every benchmark; bounds matter only on GMSC, where removing them costs validity as well; monotonicity is inert.

	GMSC		Adult		Credit	
Rule base	Act	Val	Act	Val	Act	Val
full $\mathcal{K}$	1.00	1.00	1.00	0.94	1.00	0.99
− immutability	0.71	1.00	0.78	0.94	0.39	1.00
− monotonicity	1.00	1.00	1.00	0.94	1.00	0.99
− bounds	0.17	0.71	1.00	0.94	1.00	1.00
no rules	0.14	0.71	0.79	0.93	0.40	1.00

Table 6: **The distilled circuit carries coverage.** Three plausibility terms at two margins (5-fold means; red: highest validity). All six margins in Appendix J.

		GMSC		Adult		Credit	
m	Plausibility term	Val	NLL	Val	NLL	Val	NLL
0.2	none (λ_p=0)	0.91	33.1	0.64	22.5	0.64	55.9
	Gaussian mixture	0.93	35.3	0.32	15.5	0.75	47.1
	distilled circuit	0.94	31.8	0.81	18.3	0.97	52.8
1.0	none (λ_p=0)	1.00	62.4	0.87	22.9	0.94	56.3
	Gaussian mixture	1.00	51.0	0.70	17.5	0.95	47.8
	distilled circuit	1.00	58.4	0.96	19.0	1.00	53.7

GMSC all three saturate. The decode threshold κ has a modest effect on sparsity (Table 8); cost is the more consequential property. PACER takes one to two seconds per counterfactual and the time does not change with m (Table 13), so the confidence level can be explored interactively – a practitioner, or the affected person, can move the margin and watch the proximity it costs. At tens of seconds per point for the exact formulation (§5.4) that exploration is impractical whatever the aggregate throughput, which is the substance of the speed comparison.

7 Scope and Limitations

The margin is imposed on the relaxed encoding and evaluated after hard $\arg\max$ decoding, so PACER targets a confidence level rather than certifying one: achieved margins (Table 16) track the request on GMSC and overshoot on the mixed-type datasets, leaving a minority short. Our comparison with ElliCE (Turbal et al., 2026) is preliminary, on one fold with a configuration of our own choosing (Appendix M); the threshold sweep covers one dataset and one solver; and model update means an independent retrain with a different seed, so distribution shift and continual updates are out of scope.

8 Conclusion

Robustness has been studied mostly without actionability constraints; recourse in deployment has them. There, the leading exact method asks only that the decision flip, and half its counterfactuals stop working after a retrain. PACER asks for a margin, at gradient cost, inside the actionable set.

Reproducibility Statement

Every number in this paper comes from a harness that re-scores all methods under one frozen classifier, one density model, one distance and one actionable set. Appendix N states the artefact-provenance conditions required to reproduce it and reports the verification that our re-scoring recovers the exact baseline's published validity, actionability and sparsity. The method is specified in §3 with all equations, Algorithm 1 and hyperparameters; the remaining implementation choices — projection operator, distance definition, surrogate distillation, hyperparameter sensitivity — are in Appendix C, and the actionability rule base and its construction from dataset metadata in Appendix D. Per-fold spread and paired significance tests are in Appendix L. Anonymised code and the reproduction harness are included in the supplementary material.

Ethics Statement

This work studies algorithmic recourse on public credit datasets containing sensitive attributes. To prevent harmful recommendations, our actionability constraints strictly treat demographic attributes (e.g., age, sex) as immutable. We highlight a practical ethical risk in recourse deployment: boundary-focused recourse often fails after routine model updates, misleading individuals who act in good faith. While our margin mechanism reduces this vulnerability, it provides a confidence target rather than a certified guarantee. No human subjects were involved and no new data were collected.

Use of AI Assistance

In this work, generative AI tools were used solely for drafting, editing, and polishing the text of the manuscript (including grammar correction, stylistic refinements, and LaTeX formatting). Generative AI was *not* used to formulate research hypotheses, design experimental methodologies, generate or analyze data, derive mathematical claims or proofs, or interpret experimental results. The authors have thoroughly reviewed, edited, and verified all text and take full responsibility for the final content of this manuscript.

References

Barry Becker and Ronny Kohavi. Adult, 1996. URL https://doi.org/10.24432/C5XW20.

Ann-Kathrin Dombrowski, Jan E. Gerken, Klaus-Robert Müller, and Pan Kessel. Diffeomorphic counterfactuals with generative models. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 46(5):3257–3274, 2024. doi: 10.1109/TPAMI.2023.3339980.

Tri Dung Duong, Qian Li, and Guandong Xu. CeFlow: A robust and efficient counterfactual explanation framework for tabular data using normalizing flows. In *Advances in Knowledge Discovery and Data Mining*, pp. 133–144. Springer Nature Switzerland, 2023. doi: 10.1007/978-3-031-33377-4_11.

Sanghamitra Dutta, Jason Long, Saumitra Mishra, Cecilia Tilli, and Daniele Magazzeni. Robust counterfactual explanations for tree-based ensembles. In *International conference on machine learning*, pp. 5742–5756. PMLR, 2022.

Rudresh Dwivedi, Devam Dave, Het Naik, Smiti Singhal, Rana Omer, Pankesh Patel, Bin Qian, Zhenyu Wen, Tejal Shah, Graham Morgan, et al. Explainable ai (xai): Core ideas, techniques, and solutions. *ACM Computing Surveys*, 55(9):1–33, 2023. doi: 10.1145/3561048.

Credit Fusion and Will Cukierski. Give me some credit, 2011. URL https://kaggle.com/competitions/GiveMeSomeCredit. Kaggle competition.

David Gunning, Eric Vorm, Yunyan Wang, and Matt Turek. Darpa's explainable ai (xai) program: A retrospective. *Applied AI Letters*, 2021. doi: 10.1002/ail2.61.

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), 1994. URL https://doi.org/10.24432/C5NC77.

Eric Jang, Shixiang Gu, and Ben Poole. Categorical reparameterization with gumbel-softmax. *arXiv preprint arXiv:1611.01144*, 2016.

Junqi Jiang, Jianglin Lan, Francesco Leofante, Antonio Rago, and Francesca Toni. Provably robust and plausible counterfactual explanations for neural networks via robust optimisation. In *Proceedings of the 15th Asian Conference on Machine Learning*, pp. 582–597. PMLR, 2024a.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Interval abstractions for robust counterfactual explanations. *Artificial Intelligence*, 336:104218, 2024b. ISSN 0004-3702. doi: 10.1016/j.artint.2024.104218.

Junqi Jiang, Francesco Leofante, Antonio Rago, and Francesca Toni. Robust counterfactual explanations in machine learning: A survey. *arXiv preprint arXiv:2402.01928*, 2024c.

Kentaro Kanamori, Takuya Takagi, Ken Kobayashi, and Hiroki Arimura. DACE: Distribution-aware counterfactual explanation by mixed-integer linear optimization. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence (IJCAI)*, pp. 2855–2862, 2020.

Amir-Hossein Karimi, Bernhard Schölkopf, and Isabel Valera. Algorithmic recourse: From counterfactual explanations to interventions. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21)*, pp. 353–362. Association for Computing Machinery, 2021. doi: 10.1145/3442188.3445899.

Amir-Hossein Karimi, Gilles Barthe, Bernhard Schölkopf, and Isabel Valera. A survey of algorithmic recourse: contrastive explanations and consequential recommendations. *ACM Computing Surveys*, 55(5):1–29, 2022. doi: 10.1145/3527848.

Zahra Khotanlou, Hashir Ahmed, Chenghao Tan, Ahmed Abdelaal, and Amir-Hossein Karimi. Recoursebench: A modular framework for reproducible algorithmic recourse evaluation. *arXiv preprint arXiv:2606.16113*, 2026.

Diederik P. Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

Marcin Kostrzewa, Maciej Zięba, and Jerzy Stefanowski. A probabilistic consensus-driven approach for robust counterfactual explanations. *arXiv preprint arXiv:2604.17494*, 2026.

Chris J. Maddison, Andriy Mnih, and Yee Whye Teh. The concrete distribution: A continuous relaxation of discrete random variables. *arXiv preprint arXiv:1611.00712*, 2016.

Luca Marzari, Francesco Leofante, Ferdinando Cicalese, and Alessandro Farinelli. Probabilistically robust counterfactual explanations under model changes. *Artificial Intelligence*, 351:104459, 2026. ISSN 0004-3702. doi: 10.1016/j.artint.2025.104459.

Alejandro Molina, Antonio Vergari, Nicola Di Mauro, Sriraam Natarajan, Floriana Esposito, and Kristian Kersting. Mixed sum-product networks: A deep architecture for hybrid domains. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 32, 2018.

Ramaravind K. Mothilal, Amit Sharma, and Chenhao Tan. Explaining machine learning classifiers through diverse counterfactual explanations. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 607–617. ACM, 2020. doi: 10.1145/3351095.3372850.

Jiří Němeček, Tomáš Pevnỳ, and Jakub Mareček. Generating likely counterfactuals using sum-product networks. In *International Conference on Learning Representations (ICLR)*, 2025.

Martin Pawelczyk, Klaus Broelemann, and Gjergji Kasneci. Learning model-agnostic counterfactual explanations for tabular data. In *Proceedings of The Web Conference 2020*, pp. 3126–3132, 2020. doi: 10.1145/3366423.3380087.

Hoifung Poon and Pedro Domingos. Sum-product networks: A new deep architecture. In *2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops)*, pp. 689–690, 2011. doi: 10.1109/ICCVW.2011.6130310.

Rafael Poyiadzi, Kacper Sokol, Raul Santos-Rodriguez, Tijl De Bie, and Peter Flach. FACE: Feasible and actionable counterfactual explanations. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pp. 344–350, 2020.

Kaivalya Rawal, Ece Kamar, and Himabindu Lakkaraju. Algorithmic recourse in the wild: Understanding the impact of data and model shifts, 2021. URL https://arxiv.org/abs/2012.11788.

Ignacy Stępka, Jerzy Stefanowski, and Mateusz Lango. Counterfactual explanations with probabilistic guarantees on their robustness to model change. In *Proceedings of the 31st ACM SIGKDD Conference on Knowledge Discovery and Data Mining V. 1*, pp. 1277–1288, 2025.

Bohdan Turbal, Iryna Voitsitska, and Lesia Semenova. Ellice: Efficient and provably robust algorithmic recourse via the rashomon sets. *Advances in Neural Information Processing Systems*, 38: 123872–123893, 2026.

Sohini Upadhyay, Shalmali Joshi, and Himabindu Lakkaraju. Towards robust and reliable algorithmic recourse, 2021. URL https://arxiv.org/abs/2102.13620. Advances in Neural Information Processing Systems 34, pp. 16926–16937.

Berk Ustun, Alexander Spangher, and Yang Liu. Actionable recourse in linear classification. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pp. 10–19. ACM, 2019. doi: 10.1145/3287560.3287566.

Sandra Wachter, Brent Mittelstadt, and Chris Russell. Counterfactual explanations without opening the black box: Automated decisions and the gdpr. *Harvard Journal of Law & Technology*, 31: 841, 2017. doi: 10.2139/ssrn.3063289.

Patryk Wielopolski, Oleksii Furman, Jerzy Stefanowski, and Maciej Zięba. Probabilistically plausible counterfactual explanations with normalizing flows. *arXiv preprint arXiv:2405.17640*, 2024. doi: 10.3233/FAIA240584.

A NOTATION

Table 7 collects the symbols used throughout the paper, in the order they are introduced in §3.

Table 7: Summary of notation.

Symbol	Meaning
x^0, x'	Factual; counterfactual instance
D, p	Feature count; encoded input dimension
$\mathcal{N}$, $\mathcal{C}$	Numeric / categorical index sets
$\mathcal{V}_c$, K_c	Value set; cardinality of categorical feature c
$\phi(\cdot)$	Fixed encoder (min–max normalised / one-hot)
$f_{\boldsymbol{w}}$	Classifier logit
$y^\star$, s	Desired class; target sign
m, γ	Confidence margin (set); evaluation threshold
$\mathcal{I}$, ρ_j, $[\ell_j, u_j]$	Immutable set; monotone direction; bounds
$\mathcal{A}(x^0)$	Actionable set, Eq. (1)
$\mathcal{K}$	Rule base / knowledge base (App. D)
d, $\tilde{d}$	MAD-scaled ℓ_1 distance; differentiable surrogate
r_j, a_j, ϵ_{MAD}	Range; floored MAD; MAD floor
$p_{\boldsymbol{\theta}}$, g_{ψ}	Exact density (SPN); distilled surrogate
β	Plausibility weight in Eq. (2)
$\boldsymbol{\Theta}$	Optimisation variables ($\{\delta_j\}, \{\boldsymbol{z}^{(c)}\}$)
δ_j	Numeric perturbation
σ_j	Encoder min–max scale of feature j, Eq. (5)
$\Pi_{\mathcal{N}}$	Projection onto $\mathcal{A}(x^0)$, Eq. (10)
$v_{c,k}$	k-th admissible value of categorical feature c
$\mathcal{D}_{\text{fit}}$	Sample used to distil the surrogate, Eq. (9)
$\boldsymbol{z}^{(c)}$, $\boldsymbol{\pi}^{(c)}$, τ	Categorical logits; softmax; temperature
ζ	Categorical initialisation scale
$\lambda_v, \lambda_d, \lambda_p$	Validity / proximity / plausibility weights
η_λ, $\lambda_v^{\max}$	Penalty escalation rate; cap
κ	Sparsification threshold
α, T	Step size; iterations

B CATEGORICAL INITIALISATION AND SPARSIFICATION

The categorical logits are initialised at the factual category, setting $z_k^{(c)} \leftarrow \zeta$ for the index k with $v_{c,k} = x_c^0$ and 0 otherwise, with $\zeta > 0$ (ζ=3 throughout), so that optimisation starts from "no change".[1] A category is therefore selected only when the validity and plausibility gradients are strong enough to overcome this initial margin. To promote sparse, human-interpretable recourse, we suppress numeric edits that fall below a relative threshold $\kappa \in [0, 1)$ at decode time:

$$x'_j = \begin{cases} x_j^0, & |\delta_j| < \kappa\,(u_j - \ell_j), \\ x_j^0 + \delta_j, & \text{otherwise}, \end{cases} \qquad j \in \mathcal{N}. \tag{8}$$

Integer-valued features are then rounded to the nearest admissible integer, and validity is re-checked on the resulting hard-decoded instance. Table 8 sweeps κ.

Table 8: **Sparsification threshold.** PACER at m=1.0 swept over the decode threshold κ (5-fold means); κ=0 disables suppression entirely.

	GMSC			**Adult**			**Credit**		
κ	Val	Dist	Spar	Val	Dist	Spar	Val	Dist	Spar
0.00	1.00	31.38	7.39	0.96	3.88	4.51	1.00	6.66	8.13
0.01	1.00	31.38	6.39	0.96	3.86	4.39	0.99	6.36	6.51
0.03	1.00	31.33	6.15	0.94	3.68	3.76	0.99	6.26	6.11
0.06	1.00	31.19	5.89	0.94	3.65	3.65	0.99	6.26	6.10
0.10	1.00	30.66	5.41	0.94	3.65	3.65	0.99	6.26	6.10

[1] We write ζ for this initialisation scale to avoid a collision with the evaluation threshold γ.

C IMPLEMENTATION DETAILS

All components are implemented in PyTorch; the method is solver-free and runs on CPU or GPU.

Frozen classifier and encoding. For each dataset and fold we load LiCE's released, pre-trained neural classifier and use it *unmodified* (evaluation mode, parameters frozen). The network outputs a single logit; the predicted positive-class probability is its sigmoid. Instances are mapped to the classifier's input space by LiCE's data handler, which is the encoder ϕ of §3.1. PACER optimises only the counterfactual variables $\{\delta_j\}$ and $\{\boldsymbol{z}^{(c)}\}$; the classifier's weights are never updated.

Exact plausibility via the circuit. Plausibility is scored by LiCE's fitted sum–product network. Because the circuit performs its own categorical encoding, we evaluate it by passing the *raw* mixed-type counterfactual together with the target label in its original form and reading off the exact negative log-likelihood. We verified that real training rows receive lower NLL than arbitrary factuals, confirming the scoring is correctly aligned. Every NLL reported in the paper uses this exact circuit.

Differentiable density surrogate. For the plausibility *gradient* only, we distil the circuit into a small multilayer-perceptron surrogate $g_{\boldsymbol{\psi}}$ trained by mean-squared error to regress the circuit's log-likelihood on a per-fold sample $\mathcal{D}_{\text{fit}}$ of training instances:

$$\boldsymbol{\psi}^{\star} = \arg\min_{\boldsymbol{\psi}} \; \mathbb{E}_{x \sim \mathcal{D}_{\text{fit}}} \Big[\big(g_{\boldsymbol{\psi}}(\phi(x)) - \log p_{\boldsymbol{\theta}}(\phi(x)) \big)^2 \Big]. \tag{9}$$

The surrogate supplies only the search direction; it is never used for evaluation, and its initialisation is seeded so that runs are reproducible. As an ablation we replace the circuit-derived plausibility term with a diagonal-covariance Gaussian mixture ($K{=}10$) fit to the encoded training data, whose log-density is available in closed form and is directly differentiable.

Surrogate fidelity. Because $g_{\boldsymbol{\psi}}$ influences PACER only through its gradient, we verify that it tracks the exact circuit on data it was not fitted on. For this check the surrogate is refit with the deployed recipe on 70% of each training fold and scored on the disjoint remainder ($n{=}500$ held-out instances on GMSC and Adult, $n{=}240$ on the smaller Credit). Agreement is Pearson $r = 0.921,\ 0.845,\ 0.895$ and Spearman $\rho = 0.834,\ 0.883,\ 0.893$ on Credit / GMSC / Adult, consistent across all five folds of every dataset (per-fold range 0.85–0.95 for r). The rank correlations are what matter for the descent direction: they indicate the surrogate preserves the ordering of high- versus low-density points that the optimiser follows. Every reported NLL uses the exact circuit, so residual surrogate error affects the optimisation path only, never the evaluation.

Actionability and projection. Numeric features are optimised as additive perturbations and projected after each gradient step:

$$\Pi_{\mathcal{N}}(\delta_j) = \text{clip}\big(x_j^0 + \tilde{\delta}_j,\ \ell_j,\ u_j\big) - x_j^0, \qquad \tilde{\delta}_j = \begin{cases} 0, & j \in \mathcal{I} \text{ or } \rho_j\, \delta_j < 0, \\ \delta_j, & \text{otherwise.} \end{cases} \tag{10}$$

Immutable features are held fixed, monotonicity-violating updates are zeroed, and values are clipped to per-feature box bounds. Categorical features are optimised through the deterministic temperature-softmax relaxation of their one-hot block (Eq. (3)); at decode the relaxation is projected to the $\arg\max$ category (Eq. (4)). Only categorical features whose one-hot columns are cleanly identifiable are made editable; the remainder are held at the factual value.

Distance and sparsity. Reported proximity uses the MAD-weighted ℓ_1 distance

$$d(x^0, x') = \sum_{j \in \mathcal{N}} \frac{\big| x'_j - x_j^0 \big|}{r_j\, a_j} \;+\; \sum_{c \in \mathcal{C}} \mathbf{1}\big[x'_c \neq x_c^0\big], \tag{11}$$

where r_j is the observed range of numeric feature j and a_j is the MAD of the range-normalised feature, floored by $\epsilon_{\text{MAD}}{=}0.05$ so that near-constant columns cannot inflate the distance; each categorical change costs one unit. During optimisation the differentiated proximity term is instead the surrogate $\tilde{d}$ (5), because the categorical indicator has zero gradient and the encoded space is where

the optimiser operates. Distance is computed with Eq. (11) on the hard-decoded counterfactual; sparsity uses LiCE's `isclose` rule on the normalised non-one-hot encoding, so that our numbers are comparable to theirs.

Hyperparameter sensitivity. PACER exposes one deliberate user-facing control, the confidence margin m. Its remaining hyperparameters were held *fixed across all three datasets* — no per-dataset tuning — and λ_v is not tuned but adapted online by (7). To confirm the method does not rest on a knife-edge setting, we varied each remaining hyperparameter one at a time around its default and measured the coefficient of variation (CV = std/mean) of validity, NLL and distance over the swept range (Table 9). Every coefficient is below 0.06, and the loss weights are below 0.03: over 2–4$\times$ ranges in λ_p, λ_d, τ, the learning rate and the iteration budget, the method's behaviour is essentially unchanged. The margin is therefore the only parameter a practitioner needs to reason about.

Table 9: Hyperparameter sensitivity: coefficient of variation of each metric as one hyperparameter is varied over a 2–4$\times$ range with all others fixed, on Adult (fold 0). No coefficient exceeds 0.06.

Hyperparameter (range)	CV(Val)	CV(NLL)	CV(Dist)
$\lambda_p \in [0.05, 0.4]$	0.027	0.008	0.009
$\lambda_d \in [1.0, 4.0]$	0.009	0.002	0.008
$\tau \in [0.3, 1.0]$	0.042	0.014	0.044
lr $\in [0.02, 0.1]$	0.053	0.021	0.053
steps $\in [100, 200]$	0.027	0.016	0.008

D THE ACTIONABILITY RULE BASE

The actionable set $\mathcal{A}(x^0)$ of (1) is not specified by hand per instance; it is compiled from a per-feature *rule base* — a symbolic knowledge base $\mathcal{K}$ that the same feasibility and projection code reads for every dataset. Each feature is assigned one rule describing its admissible one-step transitions, and PACER extracts from that rule exactly the four quantities its optimiser needs: whether the feature is immutable, its monotonic direction, its bounds, and whether it is integer-valued.

Rule types. Table 10 lists them. *Immutable* features (e.g. age, sex, foreign-worker status) never change; they populate $\mathcal{I}$ and are dropped from the optimisation variables outright. A *Numeric* rule carries the bounds $[\ell_j, u_j]$, a monotonic direction ρ_j and an integer flag; these parameterise the projection (10) and the rounding at decode. A *Categorical* rule stores the admissible target values as a transition graph (fully connected by default), determining both which categoricals are editable and the value set entering (3). *Ordinal* rules are the monotone special case — ranked categories such as education level or savings band, restricted to moves within $\pm$ one rank.

Construction from dataset metadata. The rule base is built programmatically from each dataset's LiCE data handler: a feature's mutability flag selects *Immutable* versus editable; its declared type selects *Categorical* versus *Numeric*; and its recorded bounds, monotonicity and discreteness fill in $[\ell_j, u_j]$, ρ_j and the integer flag. Binary features reduce to a two-node categorical flip. We describe the constraints at the level of the metadata they are read from rather than asserting a verified equivalence to LiCE's internally enforced feasible set. The resulting $\mathcal{K}$ is applied identically to PACER and to every re-scored baseline, so all methods share one feasible set.

Table 10: Rule types in the knowledge base $\mathcal{K}$ and the actionability information PACER reads from each.

Rule	Governs	PACER reads
Immutable	fixed features	$j \in \mathcal{I}$ (frozen)
Numeric	continuous / integer	$[\ell_j, u_j]$, ρ_j, integer
Categorical	unordered categories	editable set, value list
Ordinal	ranked categories	ordered values, $\pm$rank

E POSITIONING AGAINST PRIOR WORK

Table 11 places PACER against the methods of §2, following each work's own description of its properties.

F VALIDITY UNDER A CONFIDENCE THRESHOLD

Figure 3 plots the fraction of counterfactuals that remain valid as the required signed logit γ increases, and Figure 2 isolates the proximity cost of the margin relative to exact recourse.

Table 11: **The gap this paper addresses**, following each work's own description. **Action.**: per-feature actionability (immutability, monotone directions, bounds). **Mixed**: native mixed numeric/categorical support. **Plaus.**: explicit density term. **Free**: no mixed-integer or commercial solver. **Robustness**: the quantity made robust, † if set by the user at inference. A dash marks a property not claimed, not one shown to be absent. Most methods that target robustness do not enforce actionability; PROPLACE does so inside a mixed-integer program, and ElliCE, the closest to PACER, does so by gradient descent.

Method	Family	Action.	Mixed	Plaus.	Free	Robustness target
Robust counterfactual explanations						
T-Rex (Hamman et al., 2023)	gradient / NN	—	—	✓	✓	stability threshold†
PROPLACE (Jiang et al., 2024a)	robust opt.	✓	✓	✓	—	model parameters
Interval abs. (Jiang et al., 2024b)	MILP verif.	—	—	✓	—	model parameters
BetaRCE (Stępka et al., 2025)	post hoc	—	—	—	✓	admissible retrainings†
CROCE (Kostrzewa et al., 2026)	norm. flow	—	—	✓	✓	ensemble consensus†
ElliCE (Turbal et al., 2026)	Rashomon sets	✓	✓	✓	✓	Rashomon set
Actionable / plausible recourse						
DiCE (Mothilal et al., 2020)	gradient	✓	✓	—	✓	—
FACE (Poyiadzi et al., 2020)	graph search	✓	—	✓	✓	—
PPCEF (Wielopolski et al., 2024)	gradient	—	—	✓	✓	—
Actionable recourse (Ustun et al., 2019)	integer prog.	✓	✓	—	—	—
LiCE (Němeček et al., 2025)	MIO + SPN	✓	✓	✓	—	—
PACER (ours)	gradient + SPN	✓	✓	✓	✓	**conf. margin**†

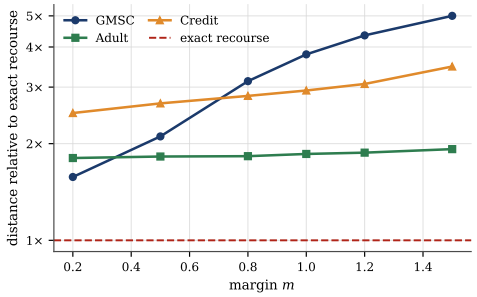


Figure 2: **What a margin costs.** Distance relative to exact recourse (log scale). On Adult the entire sweep costs 1.8–1.9×, because the softmax relaxation flips categorical features instead of stretching continuous ones. On the fully-continuous GMSC the same displacement must fall on continuous features, and the cost grows from 1.6× to 5.0×. Credit lies between the two, at 2.5–3.5×.

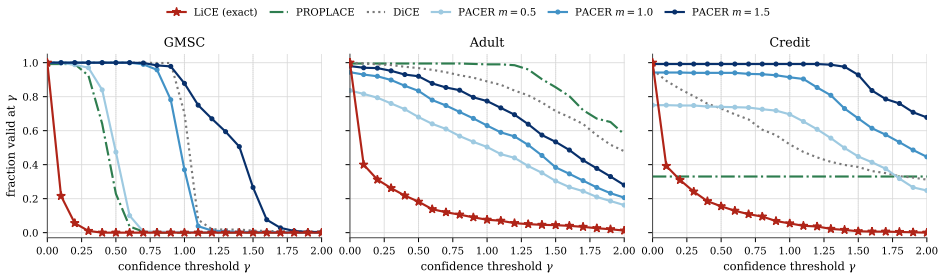


Figure 3: **Validity under a confidence threshold.** Fraction of counterfactuals that remain valid when the target class is required with signed logit above γ, pooled over five folds. Validity is measured from each method's optimisation-time signed margin; the appendix table re-encodes and re-classifies each decoded counterfactual, so values at $\gamma = 0$ can differ by a few points for methods using a relaxed categorical encoding.

G FULL WITHIN-HARNESS COMPARISON

Table 12 reports every method under the common harness with validity assessed at the decision boundary (γ=0), complementing the retention results of §5.2. The target is standardised before regression, which rescales the plausibility gradient and is absorbed into λ_p

Table 12: **Within-harness comparison** at the decision boundary (5-fold means). Val (↑) fraction of factuals receiving a valid counterfactual; NLL (↓) exact-circuit negative log-likelihood; Dist (↓) MAD-scaled ℓ_1 distance; Spar (↓) changed features under the baseline's own `isclose` definition. Red best, blue second-best per dataset and metric. The PACER block traces the price of the margin: distance grows with m at the rate plotted in Figure 2, while validity is statistically indistinguishable from exact recourse (paired Wilcoxon over 15 dataset–fold pairs, p=0.11; Appendix L).

	GMSC				**Adult**				**Credit**			
Method	Val	NLL	Dist	Spar	Val	NLL	Dist	Spar	Val	NLL	Dist	Spar
DiCE	1.00	37.2	34.35	6.42	1.00	29.7	11.91	5.17	0.95	70.4	18.45	9.38
VAE	0.01	17.8	24.03	7.88	0.80	19.3	15.33	5.36	0.28	48.4	17.69	10.78
CVAE	0.85	25.6	21.06	8.33	0.09	17.1	3.57	2.70	0.01	30.8	8.32	6.60
FACE$_k$	0.99	29.0	19.49	8.42	0.80	14.5	5.30	3.70	0.28	44.3	14.34	7.07
FACE$_\epsilon$	0.99	29.4	19.31	8.43	0.63	14.9	5.58	3.78	0.28	43.2	14.29	7.03
PROPLACE	0.99	27.9	16.80	6.39	0.80	15.4	7.71	4.79	0.28	38.2	16.66	8.94
LiCE (opt.)	1.00	25.6	8.26	2.56	1.00	18.1	1.97	2.07	0.99	28.9	2.19	2.11
LiCE (med.)	0.60	18.5	13.87	4.38	0.91	12.9	3.19	2.99	0.99	30.0	2.37	2.07
PACER (ours), density = SPN, swept over the margin m												
PACER (m=0.2)	0.94	31.8	13.01	4.28	0.81	18.3	3.56	3.60	0.97	52.8	5.45	5.45
PACER (m=0.5)	1.00	34.3	17.41	4.92	0.85	18.6	3.60	3.61	0.99	53.1	5.84	5.64
PACER (m=0.8)	1.00	46.6	25.87	5.74	0.93	18.7	3.61	3.70	0.99	53.6	6.16	5.90
PACER (m=1.0)	1.00	58.4	31.35	6.15	0.96	19.0	3.67	3.76	1.00	53.7	6.40	6.11
PACER (m=1.2)	1.00	70.1	35.92	6.39	0.98	19.3	3.70	3.85	1.00	54.1	6.71	6.31
PACER (m=1.5)	1.00	76.2	41.32	6.64	0.99	19.9	3.80	3.95	1.00	54.8	7.61	6.76

Table 13: **Wall-clock cost per counterfactual.** Order-of-magnitude cost, not a controlled same-hardware comparison; PACER on a single GPU, LiCE's released mixed-integer solve times on the authors' CPU.

Dataset	PACER (s)	LiCE (s)	Speedup
GMSC	1.08	124.75	**115**×
Adult	1.95	35.02	**18**×
Credit	2.20	3.47	1.6×

H ADDITIONAL RESULT TABLES

Tables 4 and 13 collect the result tables referenced from the main text but deferred for space: per-method actionability under the shared rule base, the rule-base ablation with distance and sparsity, and the wall-clock cost per counterfactual.

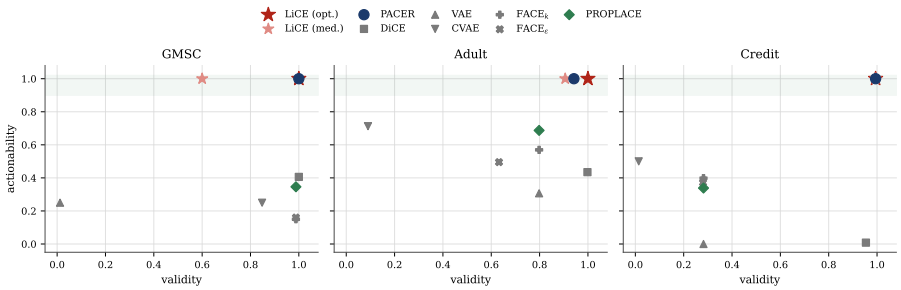


Figure 4: **Validity is not usefulness.** Actionability against validity under the shared rule base. Only PACER and the mixed-integer method occupy the top-right region.

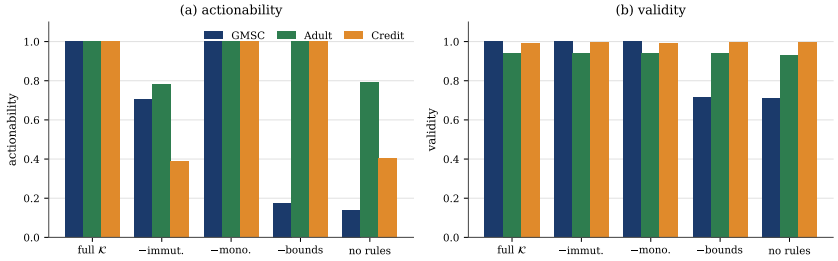


Figure 5: **The rule base is load-bearing.** Disabling one rule family at a time. Immutability is what the constraint set mostly buys; on GMSC the admissible ranges also carry validity, so there they help the optimiser rather than merely restricting it. Monotonicity is inert on these benchmarks.

I Additional ablation detail

This appendix reports the per-method actionability figures summarised in §5.5, the scatter view of validity against actionability, and the rule-base ablation with all four metrics rather than the two shown in the main text. Table 4 gives actionability for every method; Figure 4 plots it against validity, where only PACER and the mixed-integer method occupy the upper-right region; Table 5 and Table 14 report what each rule family contributes, with Figure 5 showing the same breakdown graphically.

Table 14: Role of the rule base, all four metrics (5-fold means).

	GMSC				**Adult**				**Credit**			
Rule base	Act	Val	Dist	Spar	Act	Val	Dist	Spar	Act	Val	Dist	Spar
full $\mathcal{K}$	1.00	1.00	31.33	6.15	1.00	0.94	3.68	3.76	1.00	0.99	6.26	6.11
− immutability	0.71	1.00	32.78	6.50	0.78	0.94	3.82	3.87	0.39	1.00	6.76	6.55
− monotonicity	1.00	1.00	31.28	6.14	1.00	0.94	3.69	3.77	1.00	0.99	6.28	6.12
− bounds	0.17	0.71	27.93	3.70	1.00	0.94	3.65	3.65	1.00	1.00	6.40	6.27
no rules	0.14	0.71	28.70	3.89	0.79	0.93	3.78	3.78	0.40	1.00	6.76	6.61

J FULL DENSITY ABLATION

Table 15 gives the density ablation of §6 over all six margins rather than the two reported in the main text. The tractable circuit leads on validity at every margin on both mixed-type benchmarks; on the fully-continuous GMSC all three variants saturate and the difference appears in likelihood instead.

Table 15: Density ablation, full sweep over the margin (5-fold means).

		GMSC		**Adult**		**Credit**	
m	Density	Val	NLL	Val	NLL	Val	NLL
0.2	none	0.91	33.1	0.64	22.5	0.64	55.9
0.2	GMM	0.93	35.3	0.32	15.5	0.75	47.1
0.2	SPN (ours)	0.94	31.8	0.81	18.3	0.97	52.8
0.5	none	1.00	35.8	0.73	22.5	0.75	55.5
0.5	GMM	1.00	37.0	0.45	16.7	0.84	48.2
0.5	SPN (ours)	1.00	34.3	0.85	18.6	0.99	53.1
0.8	none	1.00	48.9	0.81	22.6	0.89	55.7
0.8	GMM	1.00	45.2	0.61	17.2	0.90	48.0
0.8	SPN (ours)	1.00	46.6	0.93	18.7	0.99	53.6
1.0	none	1.00	62.4	0.87	22.9	0.94	56.3
1.0	GMM	1.00	51.0	0.70	17.5	0.95	47.8
1.0	SPN (ours)	1.00	58.4	0.96	19.0	1.00	53.7
1.2	none	1.00	71.6	0.89	23.1	0.97	56.5
1.2	GMM	1.00	54.1	0.77	17.9	0.97	48.1
1.2	SPN (ours)	1.00	70.1	0.98	19.3	1.00	54.1
1.5	none	1.00	75.5	0.92	23.8	0.99	57.0
1.5	GMM	1.00	63.2	0.81	19.1	0.99	50.1
1.5	SPN (ours)	1.00	76.2	0.99	19.9	1.00	54.8

Table 16: **Requested versus achieved margin.** Median achieved signed margin, 10th percentile, and the fraction of counterfactuals attaining the requested margin after hard decoding.

Dataset	requested m	median	10th pct	frac. $\geq m$
GMSC	0.5	0.50	0.37	0.47
	1.0	0.98	0.84	0.37
	1.5	1.40	0.98	0.27
Adult	0.5	1.02	−0.37	0.68
	1.0	1.31	0.30	0.63
	1.5	1.58	0.55	0.53
Credit	0.5	1.40	−0.48	0.74
	1.0	1.86	1.11	0.91
	1.5	2.32	1.53	0.93

K MARGIN CALIBRATION

The margin m is imposed on the relaxed encoding of Eq. (3) and evaluated only after hard $\arg\max$ decoding, so the achieved margin need not equal the requested one. Table 16 reports the distribution of achieved signed margins against the request. The median tracks the request closely on the fully-continuous GMSC and overshoots on the mixed-type benchmarks, where a categorical flip moves the decision further than needed; the tenth percentile shows the spread that leaves a minority of counterfactuals short of the target. This is the basis for the limitation stated in §7: PACER targets a confidence level rather than certifying one.

L FOLD VARIANCE AND SIGNIFICANCE

Table 17: Full within-harness comparison as mean $\pm$ std over the five folds. Gaps smaller than the spread should not be read as differences. PACER's NLL and distance on GMSC widen at large margins, where reaching the requested confidence on a fully-continuous feature space admits several distant solutions; on the mixed-type benchmarks the spread stays below 0.3 in distance throughout.

	GMSC				Adult				Credit			
Method	Val	NLL	Dist	Spar	Val	NLL	Dist	Spar	Val	NLL	Dist	Spar
DiCE	1.00±0.00	37.2±1.7	34.35±5.74	6.42±0.29	1.00±0.00	29.7±0.8	11.91±0.26	5.17±0.35	0.95±0.02	70.4±3.3	18.45±2.04	9.38±0.43
VAE	0.01±0.01	17.8±2.4	24.03±2.68	7.88±1.31	0.80±0.04	19.3±0.5	15.33±1.57	5.36±0.13	0.28±0.03	48.4±2.6	17.69±0.77	10.78±0.45
CVAE	0.85±0.05	25.6±1.1	21.06±0.70	8.33±0.04	0.09±0.03	17.1±0.2	3.57±0.76	2.70±0.39	0.01±0.01	30.8±3.2	8.32±1.87	6.60±0.82
FACE$_k$	0.99±0.01	29.0±1.4	19.49±0.81	8.42±0.10	0.80±0.04	14.5±0.4	5.30±0.30	3.70±0.10	0.28±0.03	44.3±4.3	14.34±1.29	7.07±0.25
FACE$_\epsilon$	0.99±0.01	29.4±1.3	19.31±0.96	8.43±0.19	0.63±0.05	14.9±0.3	5.58±0.41	3.78±0.13	0.28±0.03	43.2±1.3	14.29±1.36	7.03±0.29
PROPLACE	0.99±0.01	27.9±0.9	16.80±0.91	6.39±0.18	0.80±0.04	15.4±0.9	7.71±1.28	4.79±0.25	0.28±0.03	38.2±6.6	16.66±1.25	8.94±0.43
LiCE (opt.)	1.00±0.00	25.6±0.5	8.26±0.49	2.56±0.19	1.00±0.00	18.1±0.3	1.97±0.09	2.07±0.09	0.99±0.01	28.9±0.5	2.19±0.59	2.11±0.05
LiCE (med.)	0.60±0.09	18.5±0.0	13.87±1.09	4.38±0.34	0.91±0.04	12.9±0.1	3.19±0.17	2.99±0.14	0.99±0.01	30.0±0.6	2.37±0.87	2.07±0.09
PACER (m=0.2)	0.94±0.05	31.8±0.8	13.01±0.71	4.28±0.28	0.81±0.07	18.3±1.0	3.56±0.10	3.60±0.07	0.97±0.01	52.8±3.2	5.45±0.16	5.45±0.12
PACER (m=0.5)	1.00±0.00	34.3±1.1	17.41±2.16	4.92±0.36	0.85±0.07	18.6±0.9	3.60±0.11	3.61±0.10	0.99±0.02	53.1±3.0	5.84±0.22	5.64±0.13
PACER (m=0.8)	1.00±0.00	46.6±9.3	25.87±5.98	5.74±0.71	0.93±0.04	18.7±0.8	3.61±0.08	3.70±0.07	0.99±0.01	53.6±3.1	6.16±0.19	5.90±0.17
PACER (m=1.0)	1.00±0.00	58.4±17.7	31.35±7.99	6.15±0.71	0.96±0.04	19.0±0.6	3.67±0.09	3.76±0.06	1.00±0.00	53.7±3.2	6.40±0.28	6.11±0.16
PACER (m=1.2)	1.00±0.00	70.1±25.3	35.92±8.94	6.39±0.58	0.98±0.02	19.3±0.7	3.70±0.09	3.85±0.10	1.00±0.00	54.1±3.3	6.71±0.43	6.31±0.18
PACER (m=1.5)	1.00±0.00	76.2±25.8	41.32±9.36	6.64±0.42	0.99±0.01	19.9±0.5	3.80±0.08	3.95±0.11	1.00±0.00	54.8±4.1	7.61±1.68	6.76±0.52

A paired Wilcoxon signed-rank test over the 15 dataset–fold pairs, comparing PACER at m=1.0 against LiCE (opt.), gives: validity 0.987 vs 0.998 ($p = 0.11$, not significant); sparsity 5.34 vs 2.25 ($p = 0.0001$); distance 13.81 vs 4.14 ($p = 0.0001$); NLL 43.70 vs 24.23 ($p = 0.0001$).

M A PRELIMINARY COMPARISON WITH ELLICE

ElliCE (Turbal et al., 2026) is the closest prior method: it combines robustness to model change with per-feature actionability on mixed-type data without a solver (Table 11). We therefore ran its released implementation inside our harness. We report the result here rather than in the main comparison because its scope is limited, and we state that scope precisely.

Protocol. We use the authors' released code and its data-supported variant, scoring the counterfactuals it returns with our own metrics: validity against the same frozen classifier, distance under Eq. (11), and actionability against the shared rule base of Eq. (1). ElliCE requires a training support and a Rashomon tolerance ϵ; we label the support by the deployed classifier's own decisions, stratified so that both predicted classes are present, and use $\epsilon = 0.04$ throughout, with last-layer refitting enabled on GMSC and disabled elsewhere. These settings were chosen by us, not by its authors: below $\epsilon = 0.04$ the robust-candidate set grows and above it shrinks to empty, and on GMSC no candidates survive without refitting. Results are one fold and 20 factuals per dataset.

Table 18: **ElliCE in our harness** (fold 0, 20 factuals, $\epsilon = 0.04$; last-layer refit on GMSC). PACER is at m=1.0 from Table 12 and Table 4. ElliCE attains high validity on GMSC and Adult and is closer than PACER on GMSC, but its counterfactuals satisfy the shared rule base for only 0.16–0.67 of instances.

	GMSC		**Adult**		**Credit**	
Method	Val	Act	Val	Act	Val	Act
ElliCE (Turbal et al., 2026)	0.95	0.16	0.90	0.61	0.30	0.67
PACER (m=1.0)	1.00	1.00	0.96	1.00	1.00	1.00
distance (MAD-scaled ℓ_1, lower is better)						
ElliCE	17.70		6.00		15.05	
PACER (m=1.0)	31.35		3.67		6.40	

Reading. Proximity is mixed: ElliCE is closer than PACER on GMSC and further on the two mixed-type benchmarks. Actionability separates the methods. ElliCE's data-supported variant returns *observed training points* that its ellipsoid certifies as robust, and another individual's record need not preserve the factual's immutable attributes; the rule base then rejects it. ElliCE does enforce immutability and ranges inside its own gradient-based mode, so this reflects the variant and the configuration we ran. Credit's low validity has the same origin: the points its ellipsoid certifies are often ones this particular frozen classifier still rejects.

What this does and does not show. It is not a tuned comparison. ElliCE was designed around its own benchmarks and its own trained models, where the Rashomon ellipsoid is built from the training loss it has access to; our harness asks it to explain a third-party classifier we hold fixed. A configuration chosen by its authors would likely do better, and a full comparison over five folds and 100 factuals per dataset is left to future work. We report it to give the closest prior method a concrete place in our evaluation rather than only a description in Table 11.

N REPRODUCTION PROTOCOL

Our harness recovers the exact baseline's published numbers on validity, actionability and sparsity (Table 19). Three conditions are required to reach that agreement, and we state them because each is easy to violate and each changes the result materially.

Score against the released model artefacts. Released counterfactuals are solutions to an optimisation against specific weights. Regenerating the models with the authors' initialisation script — identical data splits, identical architecture, a different random seed — yields a classifier whose decision boundary differs slightly from the one the counterfactuals were solved against. Because boundary-targeted counterfactuals lie exactly on that boundary, scoring them under the regenerated model flips roughly half: measured validity moves from 1.00 to 0.50 on Adult and from 0.99 to 0.46 on Credit. Methods whose counterfactuals lie far from the boundary are almost unaffected (VAE 1.2% vs 1.2%; PROPLACE 98.8% vs 98.6% on GMSC). The asymmetry is itself a measurement of the phenomenon this paper addresses, and it is the reason we report model-update retention as a first-class metric in §5.2.

Compare feature values numerically, not as strings. Sparsity must be computed as `isclose` on the normalised encoding. Comparing raw values as strings miscounts unchanged features whose dtype differs after serialisation and inflates the exact method's reported sparsity by 3–$4\times$.

Check the candidate pools before choosing a scoring protocol. The released pools are largely degenerate — mean distinct candidates $1.28/1.00/2.97$ on GMSC / Adult / Credit — so best-of-ten and first-candidate scoring coincide on two of the three benchmarks, and the choice between them does not account for differences in reported validity.

Table 19: LiCE's *published* results: validity and actionability from their Table 3, and NLL, $\|\cdot\|_{1,\mathrm{MAD}}$ similarity and sparsity from their Table 4, both over 500 factuals (5-fold, 100 per fold) on valid generated counterfactuals. Methods able to return several candidates (MIO, LiCE, DiCE, VAE) are given ten, from which they select the likeliest valid one post hoc; our harness re-scores a single counterfactual per method, so the similarity column in particular is not cell-for-cell comparable to Table 12. "Credit" denotes German Credit.

Data	Method	Val%	Act%	NLL	MAD	Spar
GMSC	LiCE (optimize)	100.0	100.0	25.6	5.9	2.6
	LiCE (median)	60.0	60.0	18.3	11.0	4.4
	MIO (+spn)	100.0	100.0	27.9	5.9	2.1
	DiCE (+spn)	100.0	100.0	29.1	27.3	6.5
	VAE (+spn)	1.2	0.2	18.0	17.2	8.0
	C-CHVAE	84.8	21.2	25.6	18.0	8.3
	FACE (ϵ)	98.8	15.8	29.4	14.9	8.4
	FACE (knn)	98.8	14.8	29.0	15.0	8.4
	PROPLACE	98.8	34.2	27.9	12.9	6.4
Adult	LiCE (optimize)	100.0	100.0	18.1	5.6	2.1
	LiCE (median)	90.6	90.6	12.9	9.7	3.0
	MIO (+spn)	100.0	100.0	17.8	5.8	2.2
	DiCE (+spn)	99.8	65.4	21.0	26.7	4.5
	VAE (+spn)	79.8	22.6	18.4	37.1	5.4
	C-CHVAE	13.4	9.0	18.0	8.2	2.8
	FACE (ϵ)	63.2	31.2	14.9	15.2	3.8
	FACE (knn)	79.8	45.4	14.5	14.3	3.7
	PROPLACE	79.8	54.8	15.4	23.2	4.8
Credit	LiCE (optimize)	99.4	99.4	28.8	4.4	2.3
	LiCE (median)	99.4	99.4	29.9	4.4	2.1
	MIO (+spn)	99.4	99.4	43.6	4.4	2.3
	DiCE (+spn)	99.2	3.4	51.0	27.7	8.7
	VAE (+spn)	28.2	0.0	48.5	28.2	10.8
	C-CHVAE	11.6	9.6	32.3	12.9	6.6
	FACE (ϵ)	28.0	10.4	43.2	18.4	7.0
	FACE (knn)	28.2	11.2	44.2	18.6	7.1
	PROPLACE	28.2	9.6	38.4	24.4	9.0