Conformal Candidate Certification for Offline Model-Based Optimization

Seungjin Choi [1]

Abstract

Offline model-based optimization (MBO) proposes candidates by optimizing a surrogate trained on a fixed historical dataset. Because candidates are deliberately out-of-distribution, surrogate rankings are least reliable exactly where the optimizer is most aggressive, yet existing methods provide no per-candidate statistical certificate that a design meets a target threshold. We propose *Conformal Candidate Certification* (CCC), a post-hoc wrapper that attaches a calibrated one-sided lower bound to each candidate and advances only those whose bound exceeds the target. We show that entropy-regularized surrogate maximization induces a Gibbs-tilted proposal, so the same surrogate supplies importance weights for weighted conformal prediction without a separate density-ratio estimation step. In a controlled synthetic study, CCC certifies 16.7% of an aggressive proposal pool with empirical coverage 0.990 at nominal 0.90, while standard conformal prediction ignoring the covariate shift collapses to 0.416 coverage.

1. Introduction

Protein engineering, molecular design, and materials optimization require selecting designs $x \in \mathcal{X}$ with high experimental response $f(x)$ under severe evaluation budgets. Offline MBO addresses this in a single shot: given only a static dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^{n}$, propose a small candidate set $\mathcal{X}_{\text{cand}}$ without further oracle access (Kumar & Levine, 2020; Trabucco et al., 2021a). The usual pipeline trains a surrogate $\widehat{f}_\theta$ on $\mathcal{D}$ and optimizes or conditions on $\widehat{f}_\theta$ to obtain candidates.

Offline MBO methods intentionally search high-predicted-value regions that lie away from the historical distribution, where surrogates extrapolate unreliably. The result is *surrogate overestimation*: candidates look excellent under $\widehat{f}_\theta$ yet perform poorly when measured. Existing methods (Trabucco et al., 2021b; Yu et al., 2021; Qi et al., 2022; Krishnamoorthy et al., 2023; Mashkaria et al., 2023; Brookes et al., 2019), including conservative surrogate learning (COM, RoMA), invariant representations (IOM), and conditional generative models (CbAS, DDOM, BONET), address this by modifying the *proposal* mechanism. None of them answers the downstream question, “Given a proposed candidate, does the calibration data support the claim that $f(x^*) \geq \tau$ for a target level τ?” Proposal (generate high-predicted-value candidates) and certification (decide which are trustworthy enough to test) are distinct tasks; conflating them means a conservatism penalty or diffusion-guidance parameter never translates into a per-candidate coverage guarantee.

We propose **Conformal Candidate Certification** (CCC), a post-hoc certification layer that wraps any offline MBO algorithm. Given any candidate set $\mathcal{X}_{\text{cand}}$, CCC attaches a conformal lower bound $\underline{y}(x^*)$ to each candidate and returns

$$\widehat{\mathcal{C}} = \{x^* \in \mathcal{X}_{\text{cand}} : \underline{y}(x^*) \geq \tau\}. \tag{1}$$

Formally, the finite-sample guarantee is for a future measured response Y^* at the candidate: $\mathbb{P}\{Y^* \geq \underline{y}(X^*)\} \geq 1 - \alpha$. Under noiseless observations ($Y = f(X)$) this is equivalent to a guarantee on the latent objective $f(x^*)$; under measurement noise, additional assumptions on the noise model are needed to certify $f(x^*) \geq \tau$ directly. Figure 1 shows two OOD candidates both overestimated by the surrogate; CCC certifies the nearer one (smaller uncertainty penalty) and rejects the farther one (penalty drives the bound below τ), while a surrogate-threshold rule advances both.

Contributions. **(i)** We formulate post-hoc candidate certification for offline MBO as a covariate-shift conformal prediction problem, separating certification from proposal for the first time. **(ii)** We model the proposal distribution as a Gibbs tilt $Q_T \propto P_X \exp\{\widehat{f}_\theta(x)/T\}$, derived from entropy-regularized surrogate maximization; the same surrogate that creates the shift supplies the importance weights, eliminating a separate density-ratio estimation step. **(iii)** We prove finite-sample marginal lower-bound validity under oracle weights: $\mathbb{P}\{Y^* \geq \underline{y}(X^*)\} \geq 1 - \alpha$ for a fresh candidate from the proposal distribution. The required

---

[1] CROID Research and aSSIST University, Seoul, Korea. Correspondence to: Seungjin Choi <seungjin.choi.mlg@gmail.com>.

ICML 2026 Workshop on Decision-Making from Offline Datasets to Online Adaptation: Black-Box Optimization to Reinforcement Learning, Seoul, Korea. Copyright 2026 by the author(s).

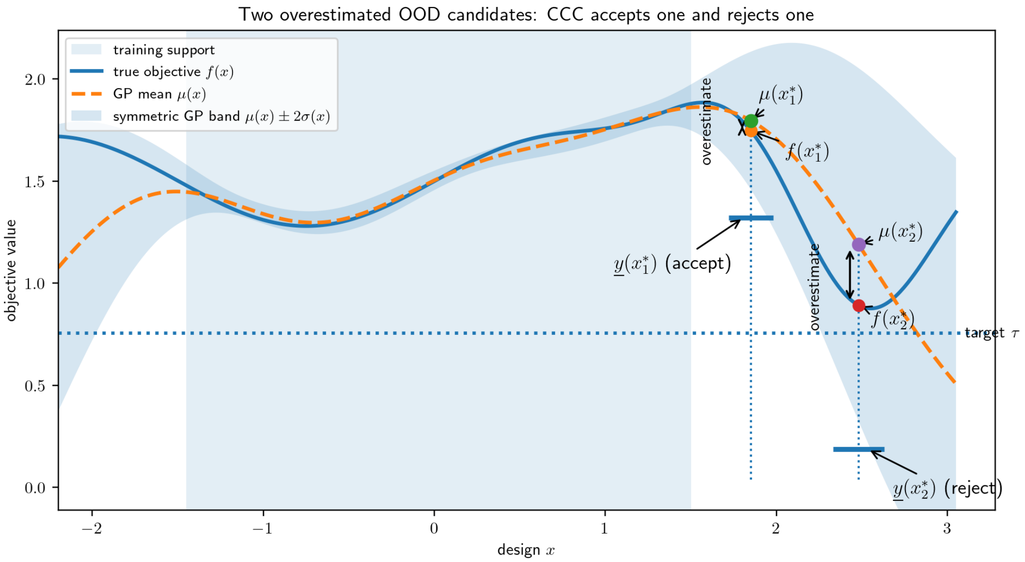

Figure 1. **Surrogate overestimation and selective certification by CCC.** Both candidates lie outside the training support and are overestimated: $\mu(x_j^*) > f(x_j^*)$. A surrogate-threshold rule advances both because both GP means exceed τ. CCC computes $\underline{y}(x_j^*)$ via (10) and accepts only if $\underline{y}(x_j^*) \geq \tau$. It accepts x_1^* (smaller conformal margin) but rejects x_2^* (larger margin pushes the bound below τ).

score-measurability condition is verified directly from the strict data-splitting discipline of Assumption 2. **(iv)** A synthetic study shows CCC certifies 16.7% of an aggressive Boltzmann proposal pool with near-zero false certifications and empirical coverage 0.990 at nominal 0.90, while Unweighted CP collapses to 0.416 coverage under the Boltzmann covariate shift.

A closely related framework is *conformalized selection* (Jin & Candès, 2023; 2026), which constructs conformal p-values controlling the false discovery rate over a selected set under (weighted) exchangeability. CCC differs in two ways: it targets a *per-candidate* lower certificate $\underline{y}(x^*)$ rather than an FDR rule, and it models the covariate shift as algorithm-induced via the Gibbs tilt rather than requiring an external density-ratio estimate. The two frameworks are complementary: FDR control over $\widehat{\mathcal{C}}$ can be obtained by passing CCC's conformal p-values $\widehat{p}^{w}_{m+1}(x_j^*)$ into the weighted conformalized selection procedure of Jin & Candès (2026).

2. Problem Setup

Let P denote the historical data-generating distribution over (X, Y), with design marginal P_X and conditional response $P(Y \mid X)$. The offline dataset $\mathcal{D}$ is partitioned into three disjoint splits: $\mathcal{D}_{\text{tr}}$ for surrogate fitting and proposal, $\mathcal{D}_{\text{val}}$ for estimating plug-in weights, and $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^{m}$ for final conformal calibration. An offline MBO algorithm, seeing only $\mathcal{D}_{\text{tr}}$, returns a candidate set $\mathcal{X}_{\text{cand}} = \{x_j^*\}_{j=1}^{K}$, which we model as approximate draws from a proposal distribution Q_X over designs.

Assumption 1 (Covariate shift)**.** *The conditional response $P(Y \mid X = x)$ is the same for historical and candidate designs. Only the design marginal shifts from P_X to Q_X.*

Assumption 2 (No calibration leakage)**.** *The proposal algorithm uses only $\mathcal{D}_{\text{tr}}$. All plug-in quantities (the temperature $\widehat{T}$ and importance weights) are fixed from $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{val}}$, the calibration covariates $\{X_i : i \in \mathcal{D}_{\text{cal}}\}$, and the candidate covariates before any response in $\mathcal{D}_{\text{cal}}$ is accessed.*

Assumption 3 (Bounded oracle weights)**.** *The oracle density ratio $w(x) = dQ_X/dP_X(x)$ exists and satisfies $w(x) \leq w_{\max} < \infty$ for P_X-almost all x.*

The certification target is a lower bound $\underline{y}(X^*)$ satisfying, for a fresh candidate $X^* \sim Q_X$ and response $Y^* \sim P(\cdot \mid X^*)$,

$$\mathbb{P}\{Y^* \geq \underline{y}(X^*)\} \geq 1 - \alpha. \tag{2}$$

The guarantee (2) is for a single fresh candidate $X^* \sim Q_X$; it does not imply that the accepted set $\widehat{\mathcal{C}}$ has a bounded false-certification rate. Finite-sample FDR control over $\widehat{\mathcal{C}}$ can be obtained by treating $\widehat{p}^{w}_{m+1}(x_j^*)$ as a conformal p-value for each candidate and applying the weighted conformalized selection procedure of Jin & Candès (2026).

3. Conformal Candidate Certification

Conformal prediction provides distribution-free uncertainty quantification by calibrating conformal scores on held-out data (Vovk et al., 2005; Angelopoulos & Bates, 2023). Standard split conformal prediction requires exchangeability between calibration and test points, but offline MBO breaks this: calibration inputs are drawn from the historical distribution P_X while candidates are drawn from the proposal distribution Q_X. Importance-weighted conformal prediction (IW-CP) restores marginal validity when the density ratio $w = dQ_X/dP_X$ is known (Tibshirani et al., 2019). The rest of this section builds the complete CCC pipeline: derive w, estimate its temperature, form the IW atoms, construct the one-sided score, and invert the weighted quantile to obtain the lower bound.

3.1. Gibbs-tilt proposal model and importance weights

The proposal distribution should concentrate on high-surrogate designs while staying close to P_X, where the surrogate is calibrated. This yields an entropy-regularized optimization:

$$Q_T = \arg\max_{Q \ll P_X} \Big\{ \mathbb{E}_{X\sim Q}\big[\widehat{f}_\theta(X)\big] - T\ \mathrm{D}_{\mathrm{KL}}(Q\|P_X) \Big\}, \quad (3)$$

whose closed-form solution is the Gibbs tilt

$$\begin{aligned} q_T(x) &= \frac{p_X(x)\exp\{\widehat{f}_\theta(x)/T\}}{Z_T}, \\ Z_T &= \mathbb{E}_{P_X}\exp\{\widehat{f}_\theta(X)/T\}, \end{aligned} \quad (4)$$

giving density ratio $w(x) = \exp\{\widehat{f}_\theta(x)/T\}/Z_T$. The identity (4) is exact for Boltzmann or softmax proposals and serves as an interpretable working model for gradient-ascent or generative MBO. Its key property is that Z_T cancels in the normalized IW-CP atoms (below), so the surrogate alone determines the weights.

Temperature estimation. We estimate T by moment-matching: find $\widehat{T}$ such that the tilted mean surrogate score over a reference set equals the mean candidate score $\bar{f}_{\mathrm{cand}} = K^{-1}\sum_{j=1}^{K}\widehat{f}_\theta(x_j^*)$. Using $\mathcal{D}_{\mathrm{val}}$ alone as the reference can severely underestimate T when candidates are OOD relative to $\mathcal{D}_{\mathrm{val}}$: the tilt collapses onto the few high-$\widehat{f}_\theta$ validation points, producing degenerate weights and near-vacuous bounds. We therefore pool $\mathcal{D}_{\mathrm{val}}$ with the candidate surrogate scores to anchor the upper tail of the reference:

$$\widehat{T} = \arg\min_{T\in[T_{\min},T_{\max}]} \left(\frac{\sum_{r\in\mathcal{R}}\exp\{\widehat{f}_\theta(r)/T\}\widehat{f}_\theta(r)}{\sum_{r\in\mathcal{R}}\exp\{\widehat{f}_\theta(r)/T\}} - \bar{f}_{\mathrm{cand}} \right)^2, \quad (5)$$

where $\mathcal{R} = \mathcal{D}_{\mathrm{val}} \cup \{\widehat{f}_\theta(x_j^*)\}_{j=1}^{K}$ pools the validation set with all candidate surrogate scores, and $[T_{\min}, T_{\max}]$ (e.g. $[0.01, 100]$ on the surrogate scale) is searched by bisection. Candidate scores are always available at estimation time and do not include calibration responses, so Assumption 2 is satisfied. When the proposal temperature is known (e.g. an explicit Boltzmann proposal), $\widehat{T}$ may be set to the oracle value directly.

IW conformal atoms. For a candidate x^*, IW-CP places atoms over the m calibration points and the (unknown) test point:

$$\begin{aligned} p_i^w(x^*) &= \frac{w(x_i)}{\sum_{\ell=1}^{m} w(x_\ell) + w(x^*)}, \\ p_{m+1}^w(x^*) &= \frac{w(x^*)}{\sum_{\ell=1}^{m} w(x_\ell) + w(x^*)}. \end{aligned} \quad (6)$$

Substituting the Gibbs ratio, Z_T cancels and the practical atoms are

$$\widehat{p}_i^w(x^*) = \frac{\exp\{\widehat{f}_\theta(x_i)/\widehat{T}\}}{\sum_{\ell=1}^{m}\exp\{\widehat{f}_\theta(x_\ell)/\widehat{T}\} + \exp\{\widehat{f}_\theta(x^*)/\widehat{T}\}}, \quad (7)$$

with an analogous expression for $\widehat{p}_{m+1}^w(x^*)$. Each candidate induces its own augmented weighted empirical distribution. For generative MBO methods whose proposal is poorly described by (4), a discriminative fallback fits a classifier to distinguish $\mathcal{D}_{\mathrm{val}}$ from $\mathcal{X}_{\mathrm{cand}}$ and uses the odds-ratio identity (Sugiyama et al., 2012).

3.2. One-sided score

We use a surrogate trained on $\mathcal{D}_{\mathrm{tr}}$, with predicted mean $\mu(x)$ and uncertainty scale $\sigma(x) > 0$. Since certification is a one-sided problem (we advance x^* only if $f(x^*)$ is certifiably above τ), a symmetric score wastes calibration probability on the upper tail. The nonconformity score is the signed one-sided residual

$$s(x, y) = \frac{\mu(x) - y}{\sigma(x)}, \quad (8)$$

which is large when the surrogate overestimates. Inverting $s(x, y) \le q$ gives directly $y \ge \mu(x) - q\sigma(x)$.

Remark 1 (Surrogate and uncertainty choice)**.** For GP surrogates, $\mu(x)$ and $\sigma(x)$ are the predictive mean and standard deviation. For non-GP surrogates (e.g. gradient boosting), set $\sigma(x) = 1$, reducing the score to the signed one-sided residual $\mu(x) - y$. The conformal quantile absorbs the scale, so the lower bound remains valid.

A natural extension is to replace $\mu(x)$ with a proposal-weighted local surrogate mean that adjusts the score center to reflect the proposal's concentration near x^*; we set this correction to zero throughout (the construction and conditions under which it helps are discussed in the supplementary material of an extended version).

3.3. Weighted conformal quantile and lower bound

For a fixed candidate x^*, compute calibration scores

$$s_i(x^*) = s(x_i, y_i) = \frac{\mu(x_i) - y_i}{\sigma(x_i)}, \qquad i \in \mathcal{D}_{\mathrm{cal}},$$

Algorithm 1 Conformal Candidate Certification (CCC)

```
Require: Dataset $\mathcal{D}$, candidate set $\mathcal{X}_{\text{cand}}$, level $\alpha$, threshold $\tau$.
Split $\mathcal{D}$ into $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{val}}$, $\mathcal{D}_{\text{cal}}$.
Train surrogate on $\mathcal{D}_{\text{tr}}$; obtain $\mu$, $\sigma$, and $\widehat{f}_\theta$.
Estimate $\widehat{T}$ on $\mathcal{D}_{\text{val}}$ by (5).
for each $x^* \in \mathcal{X}_{\text{cand}}$ do
    Compute Gibbs atoms $\widehat{p}^w_i(x^*)$ and $\widehat{p}^w_{m+1}(x^*)$ by (7).
    Compute calibration scores $s_i(x^*) = s(x_i, y_i)$, $i \in \mathcal{D}_{\text{cal}}$.
    Compute $\widehat{q}^w_{1-\alpha}(x^*)$ by (9).
    Compute $\underline{y}(x^*)$ by (10).
end for
Return $\widehat{\mathcal{C}} = \{x^* \in \mathcal{X}_{\text{cand}} : \underline{y}(x^*) \geq \tau\}$.
```

and the IW quantile

$$\widehat{q}^w_{1-\alpha}(x^*) = \text{Quantile}_{1-\alpha}\Bigg(\sum_{i=1}^{m} \widehat{p}^w_i(x^*)\,\delta_{s_i(x^*)} + \widehat{p}^w_{m+1}(x^*)\,\delta_{+\infty}\Bigg). \tag{9}$$

The $+\infty$ atom accounts for the unknown test score. Whenever $\widehat{p}^w_{m+1}(x^*) > \alpha$, the finite-mass part of the distribution sums to less than $1-\alpha$, so $\widehat{q}^w_{1-\alpha}(x^*) = +\infty$ and the lower bound is vacuous: the correct behavior when the calibration data cannot support certification of x^*.

Since $s(x^*, y)$ is strictly decreasing in y, inverting $s(x^*, y^*) \leq \widehat{q}^w_{1-\alpha}(x^*)$ gives the CCC lower bound:

$$\underline{y}(x^*) = \mu(x^*) - \widehat{q}^w_{1-\alpha}(x^*)\,\sigma(x^*). \tag{10}$$

The two terms are transparent: surrogate center and conformal uncertainty penalty. A candidate passes when $\underline{y}(x^*) \geq \tau$.

4. Validity and Diagnostic Properties

Theorem 1 applies the importance-weighted conformal prediction framework of Tibshirani et al. (2019) to the offline MBO setting. The contribution of CCC lies in how this framework is instantiated: the algorithm-induced covariate shift is governed by the Gibbs-tilt proposal Q_T, which supplies analytically tractable importance weights from the surrogate itself, and the one-sided score $s(\cdot,\cdot)$ (8) can be inverted into a lower certificate on the objective value. The theorem makes precise the validity guarantee that CCC inherits when these weights are oracle and the score is fixed before the calibration responses are accessed.

Theorem 1 (Oracle-weight marginal lower-bound validity)**.** *Suppose Assumptions 1, 2, and 3 hold, and suppose the weighted conformal quantile uses the true density ratio* $w = dQ_X/dP_X$. *Let* $\underline{y}^{\text{oracle}}(x^*)$ *denote the lower bound (10) computed with these oracle weights. Then for a fresh candidate* $X^* \sim Q_X$ *and response* $Y^* \sim P(\cdot \mid X^*)$,

$$\mathbb{P}\{Y^* \geq \underline{y}^{\text{oracle}}(X^*)\} \geq 1-\alpha. \tag{11}$$

Under noiseless observations $Y = f(X)$, *the same bound holds with* $f(X^*)$ *in place of* Y^*.

Theorem 1 should be read carefully. It guarantees validity for the *oracle* lower bound $\underline{y}^{\text{oracle}}(x^*)$ computed with the true density ratio. The practical CCC algorithm uses plug-in Gibbs weights, so the implemented $\underline{y}(x^*)$ does not inherit the exact finite-sample guarantee. The theorem is marginal over a fresh candidate $X^* \sim Q_X$: it does not hold simultaneously for all K proposed candidates, does not imply post-selection validity conditional on $x^* \in \widehat{\mathcal{C}}$, and requires $Q_X \ll P_X$.

Remark 2 (Plug-in weights and empirical coverage)**.** When Gibbs weights are estimated from data, the finite-sample guarantee no longer holds exactly. Empirical coverage above the nominal level in repeated experiments is consistent with the oracle target, but should be interpreted as diagnostic evidence rather than a finite-sample guarantee.

Remark 3 (Noisy measurements vs. latent objective)**.** Theorem 1 certifies the *future measured response* Y^*, not the latent objective $f(X^*)$. If the scientific target is $f(x^*) \geq \tau$, additional assumptions (e.g. known noise variance) are needed. In the noiseless case $Y = f(X)$ the two coincide.

5. Experiments: Synthetic Validation

We present a controlled synthetic stress test designed to address the main failure mode in offline MBO: an aggressive surrogate proposal can rank overestimated OOD candidates above the experimental target, and a useful certification rule should select a nontrivial subset rather than return the empty set.

5.1. Setup

The training support is $[-1.5, 1.5]$, while candidates are proposed in the OOD region $[1.55, 3.0]$. The true objective is

$$f(x) = 1.15 + 0.65\exp\left\{-\frac{(x-1.85)^2}{2(0.28)^2}\right\} - 0.80\exp\left\{-\frac{(x-2.58)^2}{2(0.32)^2}\right\} + 0.05\sin(3x). \tag{12}$$

Thus the OOD region contains a genuinely high-value neighborhood near $x \approx 1.85$ and an overestimated low-value neighborhood near $x \approx 2.58$. We define an intentionally optimistic surrogate center

$$\begin{aligned}\mu(x) &= f(x) + 0.08\sin(4x) + 0.14(x-1.5)_+ \\ &+ 1.25\exp\left\{-\frac{(x-2.58)^2}{2(0.38)^2}\right\}, \end{aligned} \tag{13}$$

with GP-like uncertainty scale

$$\sigma(x) = 0.08+0.22(x-1.5)_+ + 0.12\exp\left\{-\frac{(x-2.60)^2}{2(0.35)^2}\right\}. \tag{14}$$

The target threshold is $\tau = 1.45$. The surrogate center exceeds τ throughout much of the OOD region, including points whose true objective falls below τ. This creates a nontrivial screening problem: a naive surrogate threshold rule over-accepts, while a useful CCC rule should certify only the subset whose lower certificates clear the target.

Candidates are sampled from a Gibbs proposal proportional to $\exp\{\mu(x)/T_{\text{prop}}\}$ with $T_{\text{prop}} = 0.35$ over $[1.55, 3.0]$, which produces a mixture of high-value and overestimated OOD candidates. Since the proposal temperature is known by construction, we set $\widehat{T} = T_{\text{prop}} = 0.35$ directly; moment-matching on $\mathcal{D}_{\text{val}}$ underestimates T and renders most bounds vacuous in this setup. Auxiliary data are split into $\mathcal{D}_{\text{val}}$ and $\mathcal{D}_{\text{cal}}$; the calibration split is reserved for the final weighted conformal quantile. We use $\alpha = 0.10$ for the main comparison, $K = 300$ candidates per seed, and average over 40 independent seeds.

We compare three rules: (1) **Naive Surrogate** accepts candidates with $\mu(x^*) \geq \tau$; (2) **Unweighted CP** applies standard split conformal with equal weights, ignoring the Boltzmann covariate shift; (3) **CCC** uses the IW-weighted conformal quantile with the one-sided score (8). We report pass rate, empirical lower-bound coverage, precision, false-acceptance rate, and mean true objective among accepted candidates.

5.2. Selective Certification Results

Table 1 shows CCC performs selective certification rather than trivial abstention. The naive surrogate rule accepts all candidates yet only 41% are truly above τ. The key finding is that Unweighted CP, which ignores the Boltzmann covariate shift, dramatically under-covers: empirical coverage is 0.416, far below the nominal 0.90, with a 22.7% false-acceptance rate. This confirms that the Gibbs-weight correction is essential when candidates are drawn from a shifted distribution. CCC certifies 16.7% of the pool with empirical coverage 0.990 and false-acceptance rate below 0.5%.

To distinguish CCC from a merely conservative abstention rule, we evaluate coverage across multiple nominal levels. Table 2 reports empirical coverage for $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$. CCC maintains empirical coverage at or above the nominal level while certifying a nontrivial fraction of candidates at every α.

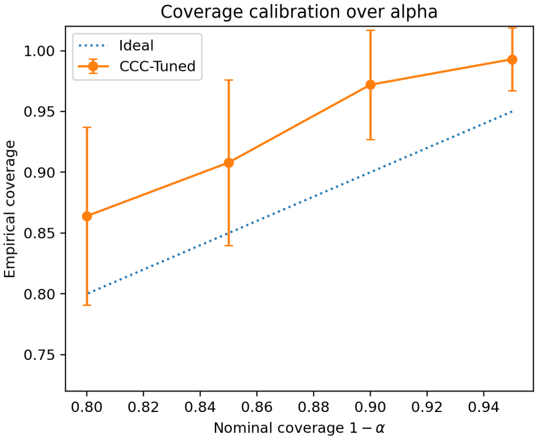

Figure 2. Coverage calibration of CCC across $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$. The dotted line is ideal calibration. CCC maintains empirical coverage at or above the nominal level while certifying a nontrivial subset at each α.

Figure 2 visualizes the same calibration diagnostic for CCC. CCC is conservative but not vacuous: increasing α raises the pass rate while empirical lower-bound coverage remains above the nominal line.

6. Discussion and Conclusion

CCC separates offline MBO into two distinct tasks: proposal and certification. Any proposal algorithm can remain aggressive; CCC then determines which candidates have enough empirical support to justify costly evaluation, and the accepted set $\widehat{\mathcal{C}}$ seeds the next experimental round.

The main limitation is weight estimation. The finite-sample guarantee assumes oracle weights; the practical Gibbs tilt is exact only for entropy-regularized proposals and approximate otherwise. The failure mode is conservative rather than anti-conservative: underestimating T widens the conformal penalty and certifies fewer candidates while maintaining coverage. High-risk discovery may still require a separate exploration budget for uncertified candidates.

Future work includes coverage-degradation bounds under plug-in weights, set-level FDR control by combining CCC's per-candidate p-values with Jin & Candès (2026), and online updates after certified experimental rounds. CCC provides a principled post-hoc answer to the question offline MBO leaves unanswered: which proposed candidates are trustworthy enough to test? Under oracle weights the answer carries marginal conformal validity; with estimated weights it gives an empirically testable certification layer for offline-to-online decision-making.

Table 1. Selective-certification experiment ($\tau = 1.45$, $\alpha = 0.10$, $K = 300$, 40 seeds). CCC certifies a nontrivial subset of the aggressive proposal pool while maintaining empirical lower-bound coverage.

Method	Pass	Cov.	Prec.	False	Mean f
Naive Surrogate	1.000±0.000	–	0.410±0.029	0.590±0.029	1.150±0.023
Unweighted CP	0.500±0.108	0.416±0.300	0.773±0.126	0.227±0.126	1.555±0.035
CCC	0.167±0.099	0.990±0.055	0.996±0.025	0.004±0.025	1.688±0.019

Table 2. Calibration of CCC across nominal levels (40 seeds, same run as Table 1).

α	Nominal	Emp. cov.	Pass	Prec.
0.05	0.95	1.000±0.000	0.003±0.008	1.000±0.000
0.10	0.90	0.990±0.055	0.167±0.099	0.996±0.025
0.15	0.85	0.964±0.152	0.237±0.098	0.986±0.063
0.20	0.80	0.938±0.174	0.285±0.095	0.975±0.072

References

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends® in Machine Learning*, 16(4):494–591, 2023.

Brookes, D. H., Park, H., and Listgarten, J. Conditioning by adaptive sampling for robust design. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2019.

Jin, Y. and Candès, E. J. Selection by prediction with conformal p-values. *Journal of Machine Learning Research*, 24:1–41, 2023.

Jin, Y. and Candès, E. J. Model-free selective inference under covariate shift via weighted conformal p-values. *Biometrika*, 113(1), 2026.

Krishnamoorthy, S., Mashkaria, S., and Grover, A. Diffusion models for black-box optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Kumar, A. and Levine, S. Model inversion networks for model-based optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020.

Mashkaria, S., Krishnamoorthy, S., and Grover, A. Generative pretraining for black-box optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

Qi, H., Su, Y., Kumar, A., and Levine, S. Data-driven offline decision-making via invariant representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Sugiyama, M., Suzuki, T., and Kanamori, T. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Tibshirani, R. J., Barber, R. F., Candès, E. J., and Ramdas, A. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

Trabucco, B., Geng, X., Kumar, A., and Levine, S. Design-bench: Benchmarks for data-driven offline model-based optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021a.

Trabucco, B., Kumar, A., Geng, X., and Levine, S. Conservative objective models for effective offline model-based optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, 2021b.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World.* Springer, 2005.

Yu, S., Ahn, S., Song, L., and Shin, J. RoMA: Robust model adaptation for offline model-based optimization. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2021.

A. Proof of Theorem 1

Proof. Let

$$\mathcal{D}_{\text{cal}} = \{(X_i, Y_i)\}_{i=1}^m$$

be the final calibration split, with

$$(X_i, Y_i) \sim P_X(dx)P(dy \mid x), \qquad i = 1, \dots, m,$$

and let

$$(X^*, Y^*) \sim Q_X(dx)P(dy \mid x)$$

be an independent fresh candidate-response pair. By Assumption 2 and the construction of the CCC pipeline, the candidate-specific score map

$$(x, y) \mapsto s(x, y)$$

is fixed before the final calibration responses are used: μ, σ, and $\widehat{T}$ are all determined from $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{val}}$, candidate covariates, and calibration covariates only, with no dependence on any $Y_i \in \mathcal{D}_{\text{cal}}$.

Under Assumption 1, the calibration and candidate distributions differ only in their design marginals: the conditional response distribution $P(Y \mid X)$ is the same. Under Assumption 3, the oracle density ratio

$$w(x) = \frac{dQ_X}{dP_X}(x)$$

exists and is finite. Applying the covariate-shift conformal theorem of Tibshirani et al. (2019) to this fixed candidate-specific score gives the following marginal statement. With oracle atoms

$$p_i^w(X^*) = \frac{w(X_i)}{\sum_{\ell=1}^m w(X_\ell) + w(X^*)}, \qquad p_{m+1}^w(X^*) = \frac{w(X^*)}{\sum_{\ell=1}^m w(X_\ell) + w(X^*)},$$

and weighted quantile $q_{1-\alpha}^w(X^*)$ formed from

$$\sum_{i=1}^m p_i^w(X^*)\, \delta_{s(X_i, Y_i)} + p_{m+1}^w(X^*)\, \delta_{+\infty},$$

we have

$$\mathbb{P}\big\{s(X^*, Y^*) \le q_{1-\alpha}^w(X^*)\big\} \ge 1 - \alpha.$$

This probability is marginal over the calibration sample and the fresh candidate-response pair. It is not a conditional coverage statement for every realized set of covariates.

It remains only to invert the one-sided score. Since $\sigma(X^*) > 0$,

$$s(X^*, Y^*) \le q_{1-\alpha}^w(X^*)$$

is equivalent to

$$Y^* \ge \mu(X^*) - q_{1-\alpha}^w(X^*)\, \sigma(X^*) = \underline{y}^{\text{oracle}}(X^*).$$

Therefore

$$\mathbb{P}\{Y^* \ge \underline{y}^{\text{oracle}}(X^*)\} \ge 1 - \alpha.$$

If the experiment is noiseless, $Y = f(X)$ almost surely, the same argument gives

$$\mathbb{P}\{f(X^*) \ge \underline{y}^{\text{oracle}}(X^*)\} \ge 1 - \alpha.$$

□