Conformal Candidate Certification for Offline Model-Based Optimization

Anonymous authors
Paper under double-blind review

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. Existing methods address this issue by regularizing the surrogate or proposal mechanism, but they do not provide a per-candidate statistical certificate that a proposed design meets a user-specified performance 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. The key challenge is covariate shift: calibration data follow the historical distribution, while candidates follow the proposal distribution. We show that entropy-regularized surrogate maximization induces a *Gibbs-tilted proposal distribution*, allowing the same surrogate that drives optimization to supply importance weights for weighted conformal prediction, without a separate density-ratio estimation step. Under oracle weights and strict data splitting, CCC satisfies finite-sample marginal lower-bound validity. Experiments on a synthetic stress test and the superconductivity dataset of Hamidieh (2018) ($n = 17{,}011$ compounds) show selective certification with empirical coverage at or above the nominal level, a 9.5 K gain in mean certified critical temperature over the naive surrogate rule, and a reduction in the false-acceptance rate by more than half.

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. Sequential Bayesian optimization maintains a surrogate of f, proposes a candidate via an acquisition function, runs the experiment, and updates the surrogate. In practice, however, the sequential query-evaluate-update loop is often impractical: a single synthesis-and-assay cycle in antibody discovery can cost $10,000 and four weeks of lead time, permitting two to five experimental rounds at most.

Offline model-based optimization (MBO) addresses this in a single offline phase: 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 training (COM, RoMA), invariant representations (IOM), and conditional generative models (CbAS, DDOM, BONET), address this by modifying the surrogate or proposal mechanism. None of them answers the downstream question: given a proposed candidate x^*, 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 commit to synthesis) are distinct tasks. A conservatism

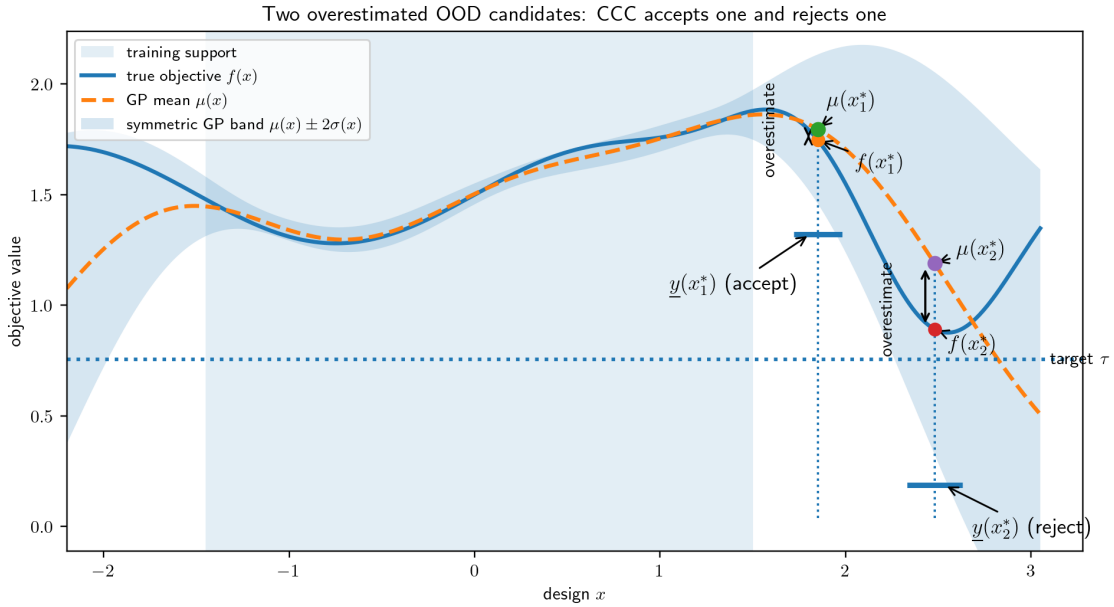

Figure 1: **Surrogate overestimation and selective certification by CCC.** Both candidates lie outside the training support and are overestimated: $\mu(x_j^*) > f(x_j^*)$. A surrogate-threshold rule advances both because both GP means exceed τ. CCC computes $\underline{y}(x_j^*)$ via a weighted conformal quantile (defined in Section 4) 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).

penalty or guidance scale can reduce overestimation on average, but it does not tell a practitioner which specific candidates are safe to synthesize.

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.

The key challenge is covariate shift. Calibration inputs are drawn from the historical distribution P_X, while candidates are drawn from the proposal distribution Q_X, concentrated in high-surrogate regions. We resolve this via the scheme of importance-weighted conformal prediction (Tibshirani et al., 2019). The critical insight is that the shift $P_X \to Q_X$ is *algorithm-induced*, which is not the same as the standard covariate shift. Entropy-regularized surrogate maximization yields an exponential tilt of the base distribution P_X, with the surrogate as the energy function — a *Gibbs tilt* $Q_X(dx) \propto P_X(dx) \exp\{\widehat{f}_\theta(x)/T\}$, whose density ratio $w(x) \propto \exp\{\widehat{f}_\theta(x)/T\}$ is determined by the same surrogate that created the shift, eliminating a separate density-ratio estimation step.

CCC is post-hoc. The certification step wraps any proposal algorithm without modifying it. Exact finite-sample validity requires the proposal distribution to be absolutely continuous with respect to the calibration distribution. The Gibbs-weight implementation is exact when the proposal arises from entropy-regularized surrogate maximization, which yields the Gibbs tilt $Q_T \propto P_X \exp\{\widehat{f}(x)/T\}$ derived in Section 4. For other proposal mechanisms the weights are approximate and empirical coverage should be monitored.

The contributions of this paper are summarized below.

- We identify that offline MBO creates an *algorithm-induced* covariate shift whose structure is determined by the optimizer's own surrogate, and formulate post-hoc candidate certification as an importance-weighted conformal prediction problem that exploits this structure, separating certification from proposal.
- We show that entropy-regularized surrogate maximization yields an exponential tilt of P_X — a Gibbs-tilt proposal $Q_T \propto P_X \exp\{\widehat{f}_\theta(x)/T\}$ — making the algorithm-induced covariate shift analytically tractable: the density ratio is determined by the surrogate itself, eliminating the separate density-ratio estimation step required by standard covariate-shift conformal methods.
- We prove finite-sample marginal lower-bound validity under oracle weights (Theorem 1); the required measurability condition follows directly from the strict data-splitting discipline of Assumption 2.
- A controlled synthetic stress test and a real materials experiment on the superconductivity dataset (Hamidieh, 2018) ($n = 17{,}011$ compounds) show that CCC achieves empirical coverage at or above the nominal level, and raises the mean certified critical temperature by 9.5 K over the naive surrogate rule while reducing the false-acceptance rate from 7.3% to 1.8%.

The remainder of the paper is organized as follows. Section 2 reviews related work. Section 3 states the problem and assumptions. Section 4 presents the full CCC pipeline. Section 5 states the main validity result. Section 6 discusses the relation to conformalized selection. Section 7 reports experiments. Section 8 discusses limitations and future work.

2 Related Work

Offline model-based optimization. Offline MBO methods address surrogate overestimation by modifying the surrogate, the proposal mechanism, or the feature space. Conservative surrogate methods include COM (Trabucco et al., 2021b), which adds an adversarial penalty that forces the surrogate below the training maximum at gradient-ascent OOD points, and RoMA (Yu et al., 2021), which applies sharpness-aware minimization to flatten the surrogate near training data. IOM (Qi et al., 2022) frames offline MBO as domain adaptation and minimizes maximum mean discrepancy between historical and candidate feature distributions. Generative and inverse-model approaches include MINs (Kumar & Levine, 2020), which learn an inverse mapping from desired objective values to designs via reweighted conditional generative adversarial networks; CbAS (Brookes et al., 2019), which iteratively retrains a VAE toward high-surrogate regions using importance-weighted ELBO updates; DDOM (Krishnamoorthy et al., 2023), which combines score-based generative modeling with classifier-free guidance; and BONET (Mashkaria et al., 2023), which autoregressively unrolls candidates from trajectory prefixes sorted by objective value. The Design-Bench benchmark (Trabucco et al., 2021a) provides standardized continuous and discrete tasks and baselines for systematic comparison.

Despite their diversity, all of these methods share the same structural gap: they tune a heuristic hyperparameter (a conservatism penalty, a guidance scale, or a trust-region radius) whose setting has no principled mapping to a per-candidate false-acceptance rate at a user-specified threshold τ. A practitioner setting $\lambda = 1.0$ in COM has no way to know whether this reduces the false-acceptance rate to 10% or to 50% on their task. CCC is complementary to all of these methods: it wraps any proposal algorithm and answers the certification question directly, without retraining or modifying the optimizer.

Conformal prediction and covariate shift. Split conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2021) provides finite-sample marginal coverage guarantees under exchangeability by inverting a nonconformity score on a held-out calibration set. For regression, the one-sided score $s(x, y) = \mu(x) - y$ yields a lower prediction bound $\underline{y}(x) = \mu(x) - \widehat{q}_{1-\alpha}$, which is precisely the form CCC uses. Tibshirani et al. (2019) extended split conformal prediction to covariate shift via importance-weighted conformal prediction (IW-CP): calibration scores are reweighted by the density ratio $w(x) = dQ/dP(x)$, restoring marginal validity when test inputs follow a different distribution Q from the calibration distribution P. IW-CP is the theoretical

backbone of CCC. The key departure in CCC is that the density ratio is not estimated from a separate sample: entropy-regularized surrogate maximization *induces* the shift as a Gibbs tilt, so the density ratio is determined analytically by the same surrogate that drives optimization. Sugiyama et al. (2012) provides the discriminative density-ratio estimation fallback used for proposal mechanisms that are not well described by a Gibbs tilt.

Conformalized selection and FDR control. Jin & Candès (2023) introduced *conformalized selection* (CS), which approaches the candidate screening problem from a multiple-testing perspective. For each candidate x_j^*, CS constructs a conformal p-value for the one-sided null hypothesis $H_j : f(x_j^*) \leq \tau$ by asking: how extreme is x_j^* relative to a held-out calibration set under the fitted surrogate? The resulting p-values are valid under exchangeability, and the Benjamini–Hochberg procedure is then applied to control the false discovery rate (FDR) over the selected set $\widehat{\mathcal{C}}$. The guarantee is a bound on the *proportion* of falsely certified candidates in expectation, not a per-candidate lower bound on $f(x_j^*)$. Jin & Candès (2026) extended CS to covariate shift by constructing weighted conformal p-values using an externally estimated density ratio, enabling FDR control when calibration and candidate designs follow different distributions.

CCC differs from CS in two respects. First, CCC provides a *quantitative* lower certificate $\underline{y}(x_j^*)$: the user learns not only whether a candidate is certified but also the lower bound on its true objective, which conveys how conservatively the guarantee was set. CS provides only a binary certified/rejected label. Second, CCC derives importance weights from the Gibbs-tilt proposal model, eliminating the need for a separate density-ratio estimator and connecting the weighting directly to the MBO algorithm that generated the candidates. The two frameworks are complementary: FDR control over $\widehat{\mathcal{C}}$ can be obtained by passing CCC's per-candidate p-values into the weighted CS procedure of Jin & Candès (2026), as described in Section 6.

3 Problem Setup and Assumptions

Let $\mathcal{X} \subseteq \mathbb{R}^d$ be a design space and $f : \mathcal{X} \to \mathbb{R}$ an unknown black-box objective whose evaluation at a single design requires wet-lab synthesis and assay. Let P denote the historical data-generating distribution over (X, Y), with design marginal P_X and conditional response $P(Y \mid X)$. We observe a static dataset $\mathcal{D} = \{(x_i, y_i)\}_{i=1}^n$ with $y_i = f(x_i) + \varepsilon_i$ and $\varepsilon_i \overset{\text{iid}}{\sim} \mathcal{N}(0, \sigma_\varepsilon^2)$.

The 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.

Assumption 1 is standard in offline MBO: the true objective function and noise model are fixed properties of the biology or physics, independent of which designs were historically screened. Assumption 2 is enforced by strict data splitting; it is the analog of the independence between calibration scores and the conformal threshold that underlies all split conformal methods. Assumption 3 is standard in importance-weighted conformal prediction (Tibshirani et al., 2019); it fails for candidates that are infinitely far from the training support, in which case the IW quantile becomes $+\infty$ and CCC correctly returns a vacuous bound.

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

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

The guarantee (2) is marginal over a fresh candidate from Q_X; it does not imply simultaneous coverage for all candidates in $\mathcal{X}_{\text{cand}}$ nor conditional coverage given selection into $\widehat{\mathcal{C}}$. Finite-sample FDR control over $\widehat{\mathcal{C}}$ requires an additional conformalized selection layer (Jin & Candès, 2023; 2026), as discussed in Section 6.

4 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, which is not satisfied in offline MBO. 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 shows how the algorithm-induced covariate shift can be corrected using the surrogate itself, culminating in the lower bound $\underline{y}(x^*)$.

4.1 Gibbs-Tilt Proposal Model and Importance Weights

A natural proposal distribution for offline MBO should concentrate probability mass on high-surrogate designs while staying close to P_X, the distribution on which the surrogate is calibrated. This trade-off is formalized by entropy-regularized surrogate maximization:

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

where $T > 0$ is a temperature hyperparameter: smaller T gives a more aggressive concentration on high-$\widehat{f}_\theta$ regions; as $T \to \infty$ the proposal approaches P_X. The closed-form solution (derived in Appendix A) is an *exponential tilt* of P_X, with the surrogate playing the role of an energy function. Following the terminology of Gibbs measures in statistical physics, we call this the *Gibbs tilt*:

$$q_T(x) = \frac{p_X(x) \exp\{\widehat{f}_\theta(x)/T\}}{Z_T}, \qquad Z_T = \mathbb{E}_{P_X}\Big[\exp\{\widehat{f}_\theta(X)/T\}\Big]. \tag{4}$$

The corresponding density ratio is

$$w(x) = \frac{q_T(x)}{p_X(x)} = \frac{\exp\{\widehat{f}_\theta(x)/T\}}{Z_T}. \tag{5}$$

The key property is that Z_T cancels in the normalized IW-CP atoms (below), so the surrogate alone determines the importance weights — no separate density-ratio estimation step is needed. Equation (4) holds exactly for proposals derived from entropy-regularized surrogate maximization, and provides a principled approximation for other MBO algorithms that concentrate candidates in high-predicted-value regions.

4.2 Temperature Estimation

The temperature T controls the strength of the Gibbs tilt. We estimate it by moment-matching: find T such that the tilted mean surrogate score over a reference set equals the mean surrogate score of the candidate set, $\bar{f}_{\text{cand}} = K^{-1} \sum_{j=1}^{K} \widehat{f}_\theta(x_j^*)$.

The choice of reference set matters. Using only the validation set can underestimate T when candidates are OOD relative to $\mathcal{D}_{\text{val}}$: the moment-matching objective may collapse the tilt onto the few high-$\widehat{f}_\theta$ validation points near the candidate region to match $\bar{f}_{\text{cand}}$, producing degenerate importance weights. For a finite returned candidate set, it is therefore useful for the reference used to estimate T to include information about the upper-tail region explored by the proposal, rather than only the bulk of the historical validation distribution. This is a practical stabilization device for plug-in weight construction; it is not needed for the oracle validity theorem, and the resulting weights remain plug-in estimates.

Pooled reference (primary estimator). Since the candidate covariates $\{x_j^*\}$ are always observed at estimation time, we pool them with the validation set to form the reference. Let $\mathcal{R}_\lambda = \mathcal{D}_{\text{val}} \cup \mathcal{X}_{\text{cand}}^\lambda$ where $\mathcal{X}_{\text{cand}}^\lambda$ is a random subsample of $\min\{K, \lfloor \frac{\lambda}{1-\lambda} |\mathcal{D}_{\text{val}}| \rfloor\}$ candidate surrogate scores. Thus $\lambda \in (0,1)$ is a target candidate fraction in the pooled reference; when the candidate set is too small to attain that target, all candidates are included (default $\lambda = 0.5$):

$$\widehat{T} = \underset{T \in [T_{\min}, T_{\max}]}{\arg\min} \left(\frac{\sum_{r \in \mathcal{R}_\lambda} \exp\{\widehat{f}_\theta(r)/T\}\, \widehat{f}_\theta(r)}{\sum_{r \in \mathcal{R}_\lambda} \exp\{\widehat{f}_\theta(r)/T\}} - \bar{f}_{\text{cand}} \right)^2. \tag{6}$$

Pooling anchors the upper tail of the reference distribution with candidate scores, allowing the optimizer to reach $\bar{f}_{\text{cand}}$ at a moderate T without collapsing. The default target fraction $\lambda = 0.5$ gives equal weight to historical and OOD information when enough candidates are available, and works well in the ablations in Appendix C. This estimator should be viewed as a finite-sample heuristic for stabilizing the Gibbs weights, not as a proof that the pooled empirical distribution equals the true base measure of the proposal. When K is large, the full candidate set can replace the subsample. The domain $[T_{\min}, T_{\max}]$ (e.g. $[0.01, 100]$ on the surrogate scale) is searched by bisection.

When the MBO algorithm exposes its sampling support (e.g. an explicit Gibbs proposal with known bounds $[a, b]$), the reference can be replaced by a dense grid $\mathcal{G}$ over that support. This provides a useful support-aware alternative to validation-only moment matching, although it is still a plug-in approximation unless the support-restricted base measure is known.

Validation set only with ESS constraint. When only $\mathcal{D}_{\text{val}}$ is available as a reference, define tilt weights and their effective sample size:

$$\pi_i(T) = \frac{\exp\{\widehat{f}_\theta(x_i)/T\}}{\sum_{\ell \in \mathcal{D}_{\text{val}}} \exp\{\widehat{f}_\theta(x_\ell)/T\}}, \qquad \text{ESS}_{\mathcal{D}_{\text{val}}}(T) = \frac{1}{\sum_{i \in \mathcal{D}_{\text{val}}} \pi_i(T)^2}. \tag{7}$$

Add a floor $\text{ESS}_{\mathcal{D}_{\text{val}}}(T) \geq m_{\min} = \max(20, 0.05|\mathcal{D}_{\text{val}}|)$ as a hard constraint:

$$\widehat{T}^{\text{ess}} = \underset{T\,:\,\text{ESS}_{\mathcal{D}_{\text{val}}}(T) \geq m_{\min}}{\arg\min} \left(\frac{\sum_{i \in \mathcal{D}_{\text{val}}} \exp\{\widehat{f}_\theta(x_i)/T\}\, \widehat{f}_\theta(x_i)}{\sum_{i \in \mathcal{D}_{\text{val}}} \exp\{\widehat{f}_\theta(x_i)/T\}} - \bar{f}_{\text{cand}} \right)^2. \tag{8}$$

Because $\text{ESS}_{\mathcal{D}_{\text{val}}}(T)$ is monotonically non-decreasing in T, the feasible set is $[T_{\text{ess}}, T_{\max}]$ found by bisection. This fallback prevents weight collapse but retains a downward bias when the candidate support is substantially OOD.

In all variants $\widehat{T}$ is computed from surrogate evaluations and candidate or validation covariates only, with no access to calibration responses in $\mathcal{D}_{\text{cal}}$, so Assumption 2 is satisfied.

IW conformal atoms. For a candidate x^*, IW-CP constructs a weighted empirical distribution over the m calibration scores and one atom at $+\infty$ representing the unknown test score. Substituting the Gibbs ratio, Z_T cancels and the normalized atoms are

$$\widehat{p}_i^w(x^*) = \frac{\exp\{\widehat{f}_\theta(x_i)/\widehat{T}\}}{\sum_{\ell=1}^m \exp\{\widehat{f}_\theta(x_\ell)/\widehat{T}\} + \exp\{\widehat{f}_\theta(x^*)/\widehat{T}\}}, \quad i = 1, \ldots, m, \tag{9}$$

$$\widehat{p}_{m+1}^w(x^*) = \frac{\exp\{\widehat{f}_\theta(x^*)/\widehat{T}\}}{\sum_{\ell=1}^m \exp\{\widehat{f}_\theta(x_\ell)/\widehat{T}\} + \exp\{\widehat{f}_\theta(x^*)/\widehat{T}\}}, \tag{10}$$

so that $\sum_{i=1}^{m+1} \widehat{p}_i^w(x^*) = 1$. The atom $\widehat{p}_{m+1}^w(x^*)$ is placed at $+\infty$; when x^* is far from the calibration support, $\widehat{p}_{m+1}^w(x^*)$ dominates and the weighted quantile becomes infinite, so CCC returns a vacuous bound. This is the correct behavior: the calibration data contain too little information about such a candidate to support certification. For generative MBO methods whose proposal is poorly described by (4), a discriminative fallback fits a classifier to distinguish $\mathcal{D}_{\text{val}}$ from $\mathcal{X}_{\text{cand}}$ and uses the odds-ratio density-ratio identity (Sugiyama et al., 2012).

Algorithm 1 Conformal Candidate Certification (CCC)

Require: Dataset $\mathcal{D}$, candidate set $\mathcal{X}_{\text{cand}}$, level α, threshold τ.
1: Split $\mathcal{D}$ into $\mathcal{D}_{\text{tr}}$, $\mathcal{D}_{\text{val}}$, $\mathcal{D}_{\text{cal}}$.
2: Train surrogate on $\mathcal{D}_{\text{tr}}$; obtain μ, σ, $\widehat{f}_\theta$.
3: Estimate $\widehat{T}$ via pooled reference (6) (or grid/ESS fallback (8)).
4: **for** each $x^* \in \mathcal{X}_{\text{cand}}$ **do**
5: Compute Gibbs atoms $\widehat{p}_i^w(x^*)$ and $\widehat{p}_{m+1}^w(x^*)$ by (9).
6: Compute calibration scores $s_i(x^*) = s(x_i, y_i)$, $i \in \mathcal{D}_{\text{cal}}$.
7: Compute $\widehat{q}_{1-\alpha}^w(x^*)$ by (12).
8: Compute $\underline{y}(x^*)$ by (13).
9: **end for**
10: **Return** $\widehat{\mathcal{C}} = \{x^* \in \mathcal{X}_{\text{cand}} : \underline{y}(x^*) \geq \tau\}$.

4.3 One-Sided Score

We use a surrogate trained on $\mathcal{D}_{\text{tr}}$, 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

$$s(x, y) = \frac{\mu(x) - y}{\sigma(x)}, \tag{11}$$

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

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.

4.4 Weighted Conformal Quantile and Lower Bound

For a fixed candidate x^*, compute calibration scores

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

and the IW quantile

$$\widehat{q}_{1-\alpha}^w(x^*) = \text{Quantile}_{1-\alpha}\left(\sum_{i=1}^{m} \widehat{p}_i^w(x^*)\, \delta_{s_i(x^*)} + \widehat{p}_{m+1}^w(x^*)\, \delta_{+\infty}\right). \tag{12}$$

If $\widehat{p}_{m+1}^w(x^*) > \alpha$, the finite-mass part sums to less than $1 - \alpha$, so $\widehat{q}_{1-\alpha}^w(x^*) = +\infty$ and the lower bound is vacuous: the correct behavior when the calibration data cannot support certification.

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

$$\underline{y}(x^*) = \mu(x^*) - \widehat{q}_{1-\alpha}^w(x^*)\, \sigma(x^*). \tag{13}$$

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

5 Validity

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 for offline MBO:

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 $s(\cdot,\cdot)$ (11) 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.

By Assumption 2 and the construction of the CCC pipeline, the score $s(\cdot,\cdot)$ (11) is fully determined before any calibration response in $\mathcal{D}_{\mathrm{cal}}$ is revealed: μ and σ are computed from $\mathcal{D}_{\mathrm{tr}}$, while $\widehat{T}$ and the plug-in atoms use only surrogate evaluations at validation, calibration, and candidate covariates. No calibration response is used before the final conformal quantile is formed. This response-independence is the key measurability condition that allows the IW-CP theorem to be applied marginally.

Theorem 1 (Oracle-weight marginal lower-bound validity)**.** *Suppose Assumptions 1, 2, and 3 hold. Let* $\underline{y}^{\mathrm{oracle}}(x^*)$ *denote the lower bound* (13) *computed with the* oracle *density ratio* $w = dQ_X/dP_X$ *in place of the plug-in Gibbs weights. Then for a fresh candidate* $X^* \sim Q_X$ *and response* $Y^* \sim P(\cdot \mid X^*)$*,*

$$\mathbb{P}\big\{Y^* \geq \underline{y}^{\mathrm{oracle}}(X^*)\big\} \geq 1-\alpha. \tag{14}$$

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

Proof. See Appendix B. □

Two scope qualifications on Theorem 1 are worth stating explicitly. The guarantee is *marginal* over a single fresh candidate $X^* \sim Q_X$, meaning 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 $Q_X \ll P_X$ (Assumption 3). It is also an *oracle* statement, as the two remarks below clarify.

Remark 2 (Plug-in weights and empirical coverage). Theorem 1 is stated for the lower bound $\underline{y}^{\mathrm{oracle}}(x^*)$ computed with the true importance weight function $w(x) = \exp\{\widehat{f}_\theta(x)/T\}/Z_T$. The practical CCC algorithm substitutes plug-in Gibbs weights $\widehat{p}_i^{\,w}(x^*)$, so the implemented bound $\underline{y}(x^*)$ does not inherit the exact finite-sample guarantee. Empirical coverage above the nominal level in repeated experiments is consistent with the oracle target but should be interpreted as diagnostic evidence. When the Gibbs proposal model is correctly specified and $\widehat{T}$ consistently estimates the true temperature, the plug-in atoms approach the oracle atoms and, under standard quantile-stability conditions, the implemented bound approaches the oracle bound as the calibration size grows.

Remark 3 (Noisy measurements vs. latent objective). When wet-lab measurements are noisy ($Y = f(X) + \varepsilon$), 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, repeated measurements) are needed to translate the Y^*-coverage guarantee into an f-guarantee. In the noiseless case $Y = f(X)$ the two coincide.

6 Relationship to Conformalized Selection

Jin & Candès (2023) introduced *conformalized selection* (CS), which constructs conformal p-values for the hypothesis $H_j : f(x_j^*) \leq \tau$ for each candidate (where f denotes the latent objective; under noise this differs from the measured response Y^*) and applies the Benjamini–Hochberg procedure to control the false discovery rate over the accepted set. Jin & Candès (2026) extended this to covariate shift via weighted conformal p-values

$$p_j^w = \frac{w(x_j^*) + \sum_{i\in\mathcal{D}_{\mathrm{cal}}} w(x_i)\,\mathbb{1}\{s_i \geq s_j^\tau\}}{w(x_j^*) + \sum_{i\in\mathcal{D}_{\mathrm{cal}}} w(x_i)}, \tag{15}$$

where $w = dQ_X/dP_X$ and s_j^τ is the score at the threshold τ.

Table 1 summarizes the comparison. CCC and CS/WCS differ in inferential target and multiplicity treatment. CCC produces a *lower bound* $\underline{y}(x^*)$ for each candidate, which directly addresses the wet-lab question "is this candidate certifiably above τ?" and provides richer information than a binary accept/reject. CS/WCS produces a set $\widehat{\mathcal{C}}$ whose false-discovery rate is controlled at the nominal level under oracle weights, but does not attach a numerical certificate to each member.

Table 1: CCC versus conformalized selection (CS/WCS). Both use importance-weighted conformal calibration; they differ in the inferential target and multiplicity treatment.

	CCC (this work)	**CS/WCS (Jin & Candès, 2023; 2026)**
Guarantee type	Marginal coverage	FDR control over selected set
Error metric	$\mathbb{P}\{Y^* \geq \underline{y}(X^*)\} \geq 1-\alpha$	$\mathrm{FDR}(\widehat{\mathcal{C}}) \leq q$
Output	Lower bound per candidate	Accept/reject per candidate
Multiple candidates	Uncontrolled false accepts	FDR-controlled
Covariate shift	Yes: oracle IW-CP; plug-in in practice	Yes: oracle WCS; estimated weights in practice
One-sided lower bound	Yes (certification target)	Not built in

The natural synthesis of the two frameworks is to layer WCS on top of CCC's per-candidate lower bounds. For each candidate x_j^*, define the threshold score

$$s_j^\tau = s(x_j^*, \tau) = \frac{\mu(x_j^*) - \tau}{\sigma(x_j^*)},$$

and the weighted conformal p-value for the one-sided threshold hypothesis:

$$p_j^w(\tau) = \widehat{p}_{m+1}^w(x_j^*) + \sum_{i=1}^{m} \widehat{p}_i^w(x_j^*)\, \mathbb{1}\{s_i(x_j^*) \geq s_j^\tau\}. \tag{16}$$

Under oracle weights, and in the noiseless setting ($Y = f(X)$) or when the null is interpreted for the future measured response, $p_j^w(\tau)$ is a valid weighted conformal p-value (Jin & Candès, 2026). Note that $\widehat{p}_{m+1}^w(x_j^*)$ in (16) is the test-point atom weight (part of the p-value formula), *not* itself a conformal p-value. Applying the weighted conformalized selection procedure of Jin & Candès (2026) to $\{p_j^w(\tau)\}_{j=1}^K$ would yield FDR-controlled certification over $\widehat{\mathcal{C}}$, combining a numerical lower bound from CCC with set-level multiplicity control from WCS. This synthesis is a direct avenue for future work.

7 Experiments

We present two experiments. Section 7.1 is the controlled synthetic stress test that isolates the core CCC mechanism under known ground truth. Section 7.2 applies CCC to the superconductivity dataset (Hamidieh, 2018), a real materials-science offline MBO task with 17,011 measured compounds.

7.1 Synthetic Selective-Certification Study

Setup. The historical distribution is concentrated mostly on $[-1.5, 1.5]$, with a small 10% overlap component in the candidate region $[1.55, 3.0]$; candidates are proposed from the OOD-enriched region $[1.55, 3.0]$. The true objective is

$$f(x) = 1.15 + 0.65 \exp\left\{-\frac{(x-1.85)^2}{2(0.28)^2}\right\} - 0.80 \exp\left\{-\frac{(x-2.58)^2}{2(0.32)^2}\right\} + 0.05 \sin(3x), \tag{17}$$

containing a genuinely high-value neighborhood near $x \approx 1.85$ and an overestimated low-value neighborhood near $x \approx 2.58$. The surrogate

$$\mu(x) = f(x) + 0.08 \sin(4x) + 0.14(x-1.5)_+ + 1.25 \exp\left\{-\frac{(x-2.58)^2}{2(0.38)^2}\right\} \tag{18}$$

is intentionally optimistic in the overestimated region, with GP-like uncertainty scale

$$\sigma(x) = 0.08 + 0.22(x-1.5)_+ + 0.12 \exp\left\{-\frac{(x-2.60)^2}{2(0.35)^2}\right\}. \tag{19}$$

Table 2: Synthetic selective-certification ($\tau = 1.45$, $\alpha = 0.10$, $K = 300$, 40 seeds).

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 3: Calibration of CCC across nominal levels (synthetic, 40 seeds).

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

Candidates are sampled from a Gibbs proposal $\propto \exp\{\mu(x)/T_{\text{prop}}\}$ with $T_{\text{prop}} = 0.35$ over $[1.55, 3.0]$, $\tau = 1.45$, $K = 300$, $\alpha = 0.10$, averaged over 40 seeds. Since the proposal temperature is known by construction, we set $\widehat{T} = T_{\text{prop}} = 0.35$ directly rather than estimating it by moment-matching; this is the oracle proposal-temperature setting for the synthetic stress test, although the resulting Gibbs atoms are still an approximation to the exact density ratio because the proposal is restricted to the candidate interval. We compare three rules: (1) **Naive Surrogate** (accept if $\mu(x^*) \geq \tau$); (2) **Unweighted CP** (standard split conformal with equal weights, ignoring the Gibbs covariate shift); (3) **CCC** (IW-weighted conformal quantile with the one-sided score (11)). The main table uses the oracle temperature; the impact of the pooled MM estimator (6) on CCC performance is reported in Table 7 of Appendix C.

Results. Table 2 shows CCC performs selective certification rather than trivial abstention. The naive rule accepts all 300 candidates yet only 41% are truly above the target. The key finding is that Unweighted CP, which ignores the Gibbs covariate shift, dramatically under-covers: empirical coverage is 0.416, far below the nominal 0.90, with nearly 23% false certifications. This suggests that the Gibbs-weight correction is essential when candidates are drawn from a shifted distribution. CCC certifies $\sim 17\%$ of the pool with pool with empirical coverage 0.990 and false-acceptance rate below 0.5%. Table 3 shows that empirical coverage tracks $1 - \alpha$ across all levels.

7.2 Superconductivity Dataset

Dataset and MBO framing. The superconductivity dataset of Hamidieh (2018) contains 21,263 compounds described by 81 chemical formula statistics (mean atomic mass, mean electron affinity, mean valence, etc.) with measured critical temperature T_c (K) as the objective. Following standard offline MBO practice we apply a 20th-percentile floor clip ($T_c \geq 4$ K), retaining $n = 17{,}011$ compounds. The offline MBO framing is: a prior experimental campaign has generated a historical dataset of measured compounds; we propose new high-T_c candidates from the remainder and certify which exceed a target threshold τ.

The scientific challenge is pronounced: the majority of compounds have $T_c < 30$ K, so a surrogate trained on this skewed distribution overestimates T_c for rare high-temperature superconductors (cuprates, iron-based compounds). This is precisely the overestimation failure mode CCC is designed to address.

Setup. Each seed randomly partitions the 17,011 compounds into a historical pool (60%, $n \approx 10{,}207$), an auxiliary pool (20%, $n \approx 3{,}402$), and a candidate pool (20%, $n \approx 3{,}402$). The surrogate is trained on the historical pool. The auxiliary pool is split into $\mathcal{D}_{\text{val}}$ and $\mathcal{D}_{\text{cal}}$ with fractions $\rho_{\text{val}} = 0.20$ and $\rho_{\text{cal}} = 0.45$ of the auxiliary-pool size, respectively; the remaining auxiliary examples are unused. We use a **gradient boosting regressor** (200 trees, depth 4, subsample 0.8, learning rate 0.05) as the surrogate in place of a GP: a GP on 81 tabular features produces posterior standard deviations of 15–30 K, which exceeds the certification

Table 4: Superconductivity dataset experiment ($\tau \approx 80$ K, $\alpha = 0.10$, $K = 200$, $n = 17{,}011$, 30 seeds). CCC uses the pooled reference estimator (6) with $\lambda = 0.5$. Coverage undefined (–) for Naive Surrogate.

Method	Pass	Cov.	Prec.	False	Mean T_c (K)
Naive Surrogate	0.832±0.021	–	0.927±0.015	0.073±0.015	105.2±1.2
Unweighted CP	0.553±0.037	0.893±0.033	0.979±0.012	0.021±0.012	113.9±1.3
CCC	0.519±0.042	0.922±0.026	0.981±0.012	0.019±0.012	114.7±1.3

Table 5: Calibration of CCC (pooled MM, $\lambda = 0.5$) across nominal levels (superconductivity dataset, 30 seeds). Empirical coverage is at or above nominal at every CCC level; Unweighted CP under-covers at $\alpha = 0.10$ (shaded).

α	Nominal	CCC		Unweighted CP	
		Emp. cov.	Pass	Emp. cov.	Pass
0.05	0.95	0.978±0.020	0.400±0.049	0.954±0.025	0.452±0.038
0.10	0.90	0.922±0.026	0.519±0.042	0.893±0.033	0.553±0.037
0.15	0.85	0.861±0.036	0.604±0.050	0.843±0.035	0.590±0.035
0.20	0.80	0.828±0.037	0.704±0.048	0.815±0.033	0.684±0.036

gap and makes all conformal bounds vacuous. The nonconformity score is the **signed one-sided residual** $s(x, y) = \mu(x) - y$ (i.e. $\sigma(x) = 1$), which requires no surrogate uncertainty estimate.

Candidates are drawn from the candidate pool via a Gibbs tilt $Q \propto \exp\{\mu(x)/T_{\text{prop}}\}$ with $T_{\text{prop}} = 10$ K: $K = 200$ candidates per seed, selected without replacement. The threshold τ is the 80th percentile of training T_c values ($\tau \approx 80$ K, well above the dataset mean of 42 K but achievable for cuprate-class compounds). Temperature is estimated by the pooled reference estimator (6) with $\lambda = 0.5$, giving $\widehat{T} \approx 25$ K across seeds. A comparison against val-only moment-matching ($\widehat{T} \approx 15$ K) is provided in Appendix C. All results are mean±std over 30 random seeds.

Results. Table 4 shows the main results at $\alpha = 0.10$. Naive Surrogate accepts 83.2% of candidates but 7.3% are spurious (below τ). Unweighted CP reduces the pass rate to 55.3% but under-covers: empirical coverage is 0.893, below the nominal 0.90, suggesting that ignoring the Gibbs covariate shift is unsafe in this experiment. CCC achieves empirical coverage 0.922 (above nominal) with precision 0.981 and a pass rate of 51.9%. The mean certified critical temperature rises from 105.2 K (Naive) to 114.7 K (CCC), a gain of 9.5 K over the naive surrogate rule while reducing the false-acceptance rate from 7.3% to 1.9%.

Calibration across nominal levels. Table 5 reports empirical coverage across $\alpha \in \{0.05, 0.10, 0.15, 0.20\}$. CCC coverage is at or above the nominal level at all four settings ($0.978 \geq 0.95$, $0.922 \geq 0.90$, $0.861 \geq 0.85$, $0.828 \geq 0.80$), consistent with the oracle-weight guarantee of Theorem 1. Because the importance weights are plug-in estimates, this experiment should be interpreted as diagnostic evidence rather than a finite-sample guarantee. Unweighted CP under-covers at $\alpha = 0.10$ ($0.893 < 0.90$), while CCC tracks the nominal level throughout. Pass rates increase monotonically with α, from 0.400 at $\alpha = 0.05$ to 0.704 at $\alpha = 0.20$, showing that CCC gracefully adapts to any user-chosen threshold without recalibration.

Practical recommendations. *(i)* Empirical coverage is consistent with the oracle-weight guarantee of Theorem 1; note that surrogate ranking quality (Spearman rank-correlation 0.88–0.93 across seeds) affects pass rate and certified mean T_c but is not a requirement for the marginal coverage theorem itself. *(ii)* When GP posterior variance is unreliable in high dimensions, use $\sigma(x) = 1$ (signed one-sided residual score); the conformal quantile absorbs the scale and the guarantee is unchanged. *(iii)* Always report empirical coverage across multiple α levels rather than at a single level; the calibration sweep in Table 5 is essential for checking that CCC is not accidentally conservative at one level only. *(iv)* The Gibbs temperature T_{prop} controls how aggressively the proposal concentrates on high-surrogate compounds. The pooled reference estimator (6)

with $\lambda = 0.5$ gives $\widehat{T} \approx 25$ K across seeds, somewhat higher than the design value $T_{\text{prop}} = 10$ K; the val-only estimator gives $\widehat{T} \approx 15$ K (see Appendix C).

8 Discussion and Conclusion

CCC separates offline MBO into two linked but distinct problems, namely proposal and certification. An offline MBO algorithm can be aggressive in proposing promising candidates. CCC then determines which have enough empirical support to justify costly wet-lab evaluation. The accepted set $\widehat{\mathcal{C}}$ becomes the bridge to the next experimental round, where new measurements can be added and the certification step repeated.

The main limitation is weight estimation. The finite-sample guarantee assumes oracle weights, and plug-in weights must be fixed independently of the calibration responses to avoid leakage, noting that the Gibbs tilt is exact only for entropy-regularized proposals. Reassuringly, underestimating T acts conservatively rather than anti-conservatively, in that a smaller temperature concentrates the Gibbs weights more aggressively, widens the conformal penalty, and certifies fewer candidates while maintaining coverage. The superconductivity experiment confirms this on real tabular data with a non-GP surrogate, where plug-in CCC achieves empirical coverage at or above the nominal level at every α, with a mean certified critical temperature 9.5 K higher than the naive surrogate rule and the false-acceptance rate reduced by more than half.

Future work should derive coverage-degradation bounds that quantify the gap between plug-in and oracle coverage in terms of density-ratio estimation error, and extend CCC to simultaneous FDR control for the accepted set via the per-candidate p-values (16) and online updates after certified experimental rounds. In summary, CCC provides a principled post-hoc answer to the question of which candidates proposed by an offline optimizer are sufficiently trustworthy to test. Under oracle weights the answer carries marginal conformal validity, while with estimated weights it gives an empirically testable certification layer for offline-to-online scientific decision-making.

References

Anastasios N. Angelopoulos and Stephen Bates. A gentle introduction to conformal prediction and distribution-free uncertainty quantification, 2021. *Preprint arXiv:2107.07511*.

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

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

Kam Hamidieh. A data-driven statistical model for predicting the critical temperature of a superconductor. *Computational Materials Science*, 154:346–354, 2018.

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

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

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

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

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

Han Qi, Yi Su, Aviral Kumar, and Sergey Levine. Data-driven offline decision-making via invariant representation learning. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022.

Masashi Sugiyama, Taiji Suzuki, and Takafumi Kanamori. *Density Ratio Estimation in Machine Learning*. Cambridge University Press, 2012.

Ryan J. Tibshirani, Rina Foygel Barber, Emmanuel J. Candès, and Aaditya Ramdas. Conformal prediction under covariate shift. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2019.

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

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

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic Learning in a Random World*. Springer, 2005.

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

A Derivation of the Gibbs Tilt

We derive the closed-form solution to the entropy-regularized surrogate maximization problem

$$Q_T = \arg\max_{Q \ll P_X} \Big\{ \mathbb{E}_{X\sim Q}[\widehat{f}_\theta(X)] - T\ \mathrm{D}_{\mathrm{KL}}(Q\|P_X) \Big\}, \tag{20}$$

where $T > 0$. Write

$$r(x) = \frac{dQ}{dP_X}(x)$$

for the Radon–Nikodym derivative of Q with respect to P_X. Then the objective becomes

$$\int r(x)\widehat{f}_\theta(x)\, P_X(dx) - T \int r(x) \log r(x)\, P_X(dx),$$

subject to $r(x) \geq 0$ and

$$\int r(x)\, P_X(dx) = 1.$$

Assume

$$0 < Z_T = \mathbb{E}_{P_X}\Big[\exp\{\widehat{f}_\theta(X)/T\}\Big] < \infty.$$

Forming the Lagrangian with multiplier ν for the normalization constraint gives

$$\mathcal{L}(r,\nu) = \int r(x)\widehat{f}_\theta(x)\, P_X(dx) - T \int r(x) \log r(x)\, P_X(dx) - \nu \left\{ \int r(x)\, P_X(dx) - 1 \right\}.$$

Taking the functional derivative with respect to $r(x)$ yields

$$\widehat{f}_\theta(x) - T\{\log r(x) + 1\} - \nu = 0.$$

Therefore

$$r(x) = \exp\left\{ \frac{\widehat{f}_\theta(x) - \nu - T}{T} \right\} = C \exp\left\{ \frac{\widehat{f}_\theta(x)}{T} \right\},$$

where $C = \exp\{-(\nu + T)/T\}$. Enforcing

$$\int r(x)\, P_X(dx) = 1$$

gives

$$C = \frac{1}{\mathbb{E}_{P_X}[\exp\{\widehat{f}_\theta(X)/T\}]} = \frac{1}{Z_T}.$$

Hence the optimizer has Radon–Nikodym derivative

$$\frac{dQ_T}{dP_X}(x) = r_T(x) = \frac{\exp\{\widehat{f}_\theta(x)/T\}}{Z_T}. \tag{21}$$

Equivalently, if P_X admits a Lebesgue density p_X, then Q_T admits Lebesgue density

$$p_T(x) = \frac{p_X(x)\exp\{\widehat{f}_\theta(x)/T\}}{Z_T},$$

which is the Gibbs tilt in (4). Since the objective is strictly concave in r over the probability simplex—the first term is linear and the negative KL term is strictly concave—this stationary point is the unique maximizer, P_X-almost surely. The density ratio needed for IW-CP is therefore

$$w(x) = \frac{dQ_T}{dP_X}(x) = \frac{\exp\{\widehat{f}_\theta(x)/T\}}{Z_T},$$

which is equation (5). Since Z_T is constant in x, it cancels in normalized importance weights, and in particular in the normalized IW-CP atoms (9).

B Proof of Theorem 1

Proof. Let

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

be the final calibration split, with

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

and let

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

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 $\mathcal{D}_{\mathrm{tr}}$, $\mathcal{D}_{\mathrm{val}}$, candidate covariates, and calibration covariates only, with no dependence on any $Y_i \in \mathcal{D}_{\mathrm{cal}}$.

Under Assumption 1, the calibration and candidate distributions differ only in their design marginals: the conditional response distribution $P(Y \,|\, X)$ 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

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

Therefore

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

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

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

□

C Experimental Details

Synthetic study. Analytical surrogate $\mu(x)$ and $\sigma(x)$ given by (18)–(19); no surrogate is fitted on training data. Historical distribution $P_X = 0.9\,\mathrm{Unif}[-1.5, 1.5] + 0.1\,\mathrm{Unif}[1.55, 3.0]$; $n = 400$ points, noise std 0.05, split 60/20/20 into $\mathcal{D}_{\mathrm{tr}}/\mathcal{D}_{\mathrm{val}}/\mathcal{D}_{\mathrm{cal}}$ (training split unused since the surrogate is analytical). Candidates: Gibbs proposal $\propto \exp\{\mu(x)/0.35\}$ over $[1.55, 3.0]$, $K = 300$. Temperature: oracle value $\widehat{T} = T_{\mathrm{prop}} = 0.35$ in the main selective-certification table, used directly because the proposal temperature is known by construction. Score: signed one-sided residual (11). 40 seeds.

Temperature estimation ablation (synthetic). Table 6 compares five estimators on the synthetic stress test ($n_{\mathrm{val}} = 80$, $K = 300$, 40 seeds). Validation-only MM returns $\widehat{T} \approx 0.14$ because the validation set (mostly in $[-1.5, 1.5]$) and the candidate set ($[1.55, 3.0]$) have largely disjoint surrogate-score supports. All three improved estimators are always available in a CCC run: the pooled reference (6) uses only the val set and the already-observed candidate scores, the ESS-constrained fallback (8) uses only the val set with an ESS floor, and the grid-based estimator additionally requires the candidate support boundary.

Table 6: Temperature-estimation ablation on the synthetic stress test ($T_{\mathrm{prop}} = 0.35$, $n_{\mathrm{val}} = 80$, $K = 300$, 40 seeds). The pooled reference (6) with $\lambda = 0.5$ requires no knowledge of the candidate support boundary and achieves near-oracle temperature accuracy.

Estimator	$\widehat{T}$	MAE	Bias
MM on val set only	0.138±0.042	0.212	−0.212
MM+ESS on val set (8)	0.285±0.020	0.065	−0.065
MM on cand-support grid (known boundary)	0.352±0.056	0.042	+0.002
MM on pooled ref, $\lambda = 0.5$ (6)	0.315±0.025	0.038	−0.035
MM on pooled ref, $\lambda = 0.6$ (6)	0.357±0.027	0.022	+0.007
Oracle T_{prop}	0.350	0	0

The pooled estimator with $\lambda = 0.5$–0.6 achieves MAE of 0.022–0.038, substantially better than validation-only MM (0.212) and MM+ESS (0.065), and comparable to or better than the grid-based approach (0.042). The optimal λ lies near the zero-bias crossing point ($\lambda \approx 0.55$–0.60); $\lambda = 0.5$ is a conservative default that slightly underestimates T but avoids over-weighting the OOD region. Sensitivity to λ is provided in the supplementary code.

CCC performance with estimated temperature. Table 7 reports CCC pass rate, coverage, precision, and false-acceptance rate when $\widehat{T}$ is estimated by the pooled MM estimator (6) rather than set to the oracle value. At $\lambda = 0.5$ the pass rate drops from 0.167 (oracle) to 0.119, a reduction of about 29%, because the slightly underestimated temperature ($\widehat{T} \approx 0.314$) produces more conservative Gibbs weights and wider conformal intervals. Coverage rises to 1.000 and the false-acceptance rate falls to 0.000: estimated T errs in the safe direction. At $\lambda = 0.6$ ($\widehat{T} \approx 0.356$, near oracle) the pass rate recovers to 0.161 with coverage 0.996 and false rate 0.004, essentially matching the oracle result. The calibration rows show that empirical coverage tracks the nominal level at all four α values under both λ settings.

Superconductivity dataset experiment. Raw dataset: 21,263 compounds, 81 tabular features, objective T_c (K). Pre-processing: 20th-percentile floor clip ($T_c \geq 4$ K), retaining $n = 17{,}011$ compounds; features standardized per-training-split. Surrogate: `HistGradientBoostingRegressor` (200 trees, depth 4, learning rate 0.05, `random_state`=seed). Score: signed one-sided residual $s(x, y) = \mu(x) - y$ ($\sigma(x) = 1$). Data partition (per seed): 60% historical ($n \approx 10{,}207$), 20% auxiliary, 20% candidate pool. Auxiliary split: $\rho_{\mathrm{val}} = 0.20$, $\rho_{\mathrm{cal}} = 0.45$. Candidate pool: $K = 200$ compounds drawn without replacement, weighted by Gibbs tilt $\propto \exp\{\mu(x)/T_{\mathrm{prop}}\}$ with $T_{\mathrm{prop}} = 10$ K. Threshold: $\tau = 80$th percentile of training T_c ($\tau \approx 80$ K). Temperature (pooled MM, primary): pooled reference (6) with $\lambda = 0.5$, giving $\widehat{T} \approx 25$ K; $\lambda = 0.5$ and $\lambda = 0.6$ give identical estimates because $K = 200$ is smaller than $\lambda/(1-\lambda) \cdot |\mathcal{D}_{\mathrm{val}}| \approx 680$, so all 200 candidates are pooled in both cases. 30 seeds.

Table 7: CCC performance under oracle vs. estimated temperature ($\tau = 1.45$, $K = 300$, 40 seeds). Pooled MM (6) with $\lambda = 0.5$–0.6 produces conservative but valid results: lower pass rate than oracle but empirical coverage at or above nominal and near-zero false-acceptance rate.

T setting	$\widehat{T}$	Pass	Cov. (α=0.10)	Prec.	False
Oracle $T_{\text{prop}} = 0.35$	0.350	0.167±0.099	0.990±0.055	0.996±0.025	0.004
Pooled MM, $\lambda = 0.5$ (6)	0.314±0.018	0.119±0.069	1.000±0.000	1.000±0.000	0.000
Pooled MM, $\lambda = 0.6$ (6)	0.356±0.017	0.161±0.081	0.996±0.021	0.996±0.021	0.004
Val-only MM	0.133±0.051	0.015±0.038	1.000±0.000	1.000±0.000	0.000
Calibration sweep, pooled MM $\lambda = 0.5$:					
$\alpha = 0.05$, nom. 0.95		0.001±0.005	1.000±0.000	1.000±0.000	0.000
$\alpha = 0.10$, nom. 0.90		0.119±0.069	1.000±0.000	1.000±0.000	0.000
$\alpha = 0.15$, nom. 0.85		0.219±0.109	0.964±0.153	0.994±0.039	0.006
$\alpha = 0.20$, nom. 0.80		0.268±0.092	0.955±0.161	0.986±0.063	0.014

Temperature estimator ablation (superconductivity). Table 8 compares validation-only MM and pooled MM on the superconductivity experiment. Because val and candidate surrogate-score distributions overlap substantially in this experiment, validation-only MM already gives a reasonable estimate (≈ 15 K vs $T_{\text{prop}} = 10$ K). The pooled estimator raises $\widehat{T}$ to ≈ 25 K and produces marginal improvements across all metrics, suggesting that the pooled reference is the more stable choice, while the practical difference is small when support overlap is substantial.

Table 8: Temperature estimator ablation on the superconductivity dataset ($\alpha = 0.10$, 30 seeds). Both estimators give empirical coverage at or above nominal; differences are within one standard deviation.

Estimator	$\widehat{T}$ (K)	Pass	Cov.	Prec.	Mean T_c (K)
MM on val set only	15.18±1.64	0.531±0.040	0.914±0.033	0.982±0.011	114.4±1.4
MM pooled, $\lambda = 0.5$ (6)	25.18±1.08	0.519±0.042	0.922±0.026	0.981±0.012	114.7±1.3

Validation-size sensitivity. Table 9 varies the validation fraction $\rho_{\text{val}} \in \{0.10, \ldots, 0.50\}$ under the validation-only temperature estimator. Pass rate and empirical coverage are stable throughout, suggesting that the qualitative behavior is not an artifact of the split size. These results are reported as a sensitivity diagnostic; the main synthetic tables use either the oracle proposal temperature or the pooled estimator in Table 7.

Table 9: Validation-size sensitivity for CCC ($\tau = 1.45$, $\alpha = 0.10$, $K = 300$, 40 seeds).

ρ_{val}	$\widehat{T}$	Pass	Cov.	Prec.	# acc.
0.10	0.230±0.091	0.244±0.118	0.949±0.067	1.000±0.000	73.2±35.3
0.20	0.238±0.064	0.263±0.106	0.958±0.057	1.000±0.000	78.8±31.7
0.30	0.241±0.057	0.260±0.074	0.960±0.054	1.000±0.000	78.1±22.2
0.40	0.242±0.040	0.282±0.061	0.982±0.028	1.000±0.000	84.7±18.3
0.50	0.234±0.031	0.286±0.064	0.965±0.059	1.000±0.000	85.8±19.2