RELIABLE SELF-EVOLUTION WITH IMPERFECT PROXY REWARDS.

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language model (LLM)-based self-evolving search is a promising approach to scientific discovery. However, high-fidelity evaluation of every candidate is prohibitively expensive in some domains. Self-evolving systems in such settings therefore rely on low-cost but imperfect proxy rewards, which may assign high scores to infeasible candidates. These false positives may contaminate both the final output and the feedback used to guide subsequent generations. This motivates statistically calibrated reward intervals for more reliable self-evolving search. We propose Conformal Interval-Driven Self-Evolution (CISE), which constructs candidate-specific reward intervals using conditional conformal inference and iteration-wise online density-ratio estimation. CISE uses conservative interval-based rewards for evolutionary feedback and returns candidates only when all required property intervals lie entirely within their respective feasible regions. We derive fixed-iteration coverage results under explicit assumptions of independence and covariate shift. We evaluate CISE on three self-evolving search tasks in materials science. In our experiments, all candidates returned by CISE are true positives under high-fidelity evaluation, whereas the baselines return more candidates but include false positives. These results highlight the value of a smaller, more precise shortlist when downstream validation budgets are limited. Our anonymized repository is available at `https://github.com/anonoymouscise/CISE-Reliable-Self-Evolution.git`.

1 INTRODUCTION

Self-evolving search uses large language models (LLMs) to generate candidates and improve subsequent candidates using feedback from previous evaluations. This paradigm has recently emerged as an approach to scientific discovery (Romera-Paredes et al., 2024; Novikov et al., 2025; Liu et al., 2026). These methods retain previously evaluated candidates together with their rewards and reuse them to guide subsequent search. The effectiveness of self-evolution therefore depends on the reliability of the rewards that determine which candidates are treated as successful and reused in later iterations.

However, obtaining the true reward for every generated candidate is often slow or expensive in iterative search. In such settings, self-evolving systems often rely on computationally inexpensive proxy rewards that approximate the true reward. For example, in materials discovery, evaluating every candidate with high-fidelity density functional theory (DFT) calculations or experimental validation is prohibitively expensive. Property prediction models are therefore often used as proxy rewards during search (Abhyankar et al., 2026). Although proxy rewards make iterative search practical, prediction errors may produce false positives: candidates that satisfy the target constraints under the proxy reward but fail under the true reward. In self-evolving search, such errors affect more than the returned candidates. Once a false positive is assigned a favorable reward and stored as a successful example, it is reused to guide subsequent generations. This false-positive propagation may steer the search toward regions where the proxy reward is overly optimistic rather than toward candidates with genuinely high true rewards.

This motivates uncertainty-aware reward estimates rather than relying solely on point-valued proxy rewards. Conformal inference (Vovk et al., 2005; Angelopoulos & Bates, 2022) provides a principled way to construct such estimates by converting model predictions into calibrated intervals with statistical coverage guarantees. In particular, conditional conformal inference (Gibbs et al., 2025)

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.

incorporates candidate characteristics into calibration, which is useful when proxy-reward reliability varies across candidates. These calibrated intervals support conservative decisions about which candidates guide subsequent evolution and which candidates are returned as outputs.

We propose Conformal Interval-Driven Self-Evolution (CISE), a self-evolving framework that replaces point-valued proxy rewards with candidate-specific calibrated reward intervals. CISE uses conditional conformal inference to construct candidate-specific reward intervals with adaptive widths. The resulting reward interval is stored in the archive and used to guide subsequent generations, while only candidates whose entire reward intervals satisfy the task-specific constraints are returned as outputs. Because the candidate distribution can change as the archive, selected parents, and generation context evolve, CISE additionally performs iteration-wise online density-ratio estimation (Kanamori et al., 2009) to characterize the difference between the source distribution and the current candidate distribution and incorporates the estimated ratio into conditional conformal inference. As a result, calibrated reward intervals are used for both evolutionary feedback and final candidate selection.

Theoretically, we analyze fixed-iteration marginal coverage under explicit independence and covariate-shift assumptions and characterize how density-ratio estimation error affects the resulting guarantee. Empirically, we evaluate CISE on three self-evolving material-property search tasks and validate the returned candidates using high-fidelity DFT calculations. Across the reported experiments, CISE trades output yield for higher observed DFT-validated output precision: all CISE outputs with resolved DFT evaluations are true positives, whereas the baseline methods return false positives. We further show under a matched downstream validation budget that the reliability improvement cannot be explained solely by CISE returning fewer candidates. These results demonstrate the potential of calibrated reward intervals to make self-evolving search more reliable when accurate true rewards are too expensive to use throughout evolution. To the best of our knowledge, CISE is the first framework to use conformal proxy-reward intervals for both evolutionary feedback and final candidate selection in LLM-based self-evolving candidate search.

2 Related Work

Self-evolving search reuses previously generated candidates and their rewards to guide subsequent generations across domains, including program discovery (Romera-Paredes et al., 2024; Novikov et al., 2025), video-prompt optimization (Ji et al., 2024), scientific hypothesis generation (Gottweis et al., 2026), and materials discovery (Abhyankar et al., 2026). While some settings provide directly verifiable rewards, others rely on learned or LLM-generated proxy rewards to enable efficient iterative search. In such cases, proxy-favored candidates may not correspond to true improvements, and reusing them as successful examples can propagate reward errors across subsequent generations. This motivates accounting for proxy-reward uncertainty in both evolutionary feedback and final selection.

Recent work has sought to improve the reliability of proxy evaluations and rewards through calibration and uncertainty-aware optimization. For LLM-as-a-judge, post-hoc score calibration learns regression models from limited human feedback to adjust judge scores using their original scores and rationales (Sahoo et al., 2026), while linear probes on judge hidden states provide calibrated confidence estimates for the correctness of their judgments (Radharapu et al., 2025). Related work on learned reward models addresses uncertainty and overoptimization directly: ensemble-based conservative objectives use disagreement among reward models to reduce optimization against errors in an imperfect reward model (Coste et al., 2024). Process reward models have also been calibrated using quantile regression to better align their predictions with empirical success probabilities and support uncertainty-aware allocation of inference-time computation (Park et al., 2025). These approaches improve the calibration, interpretation, or robustness of model-derived rewards. However, they do not statistically control false-positive rewards under task-specific true-reward constraints in iterative self-evolving search.

Conformal inference provides a complementary framework for uncertainty quantification by converting model predictions into calibrated prediction sets or intervals with finite-sample coverage guarantees under explicit assumptions (Vovk et al., 2005; Angelopoulos & Bates, 2022). Conditional conformal inference formulates conditional validity through classes of covariate shifts and provides finite-sample guarantees for shifts represented by a finite-dimensional calibration function space (Gibbs et al., 2025). More recently, conformal inference has been applied to LLM-as-a-judge evaluation, replacing point-valued judge scores with calibrated intervals that explicitly represent

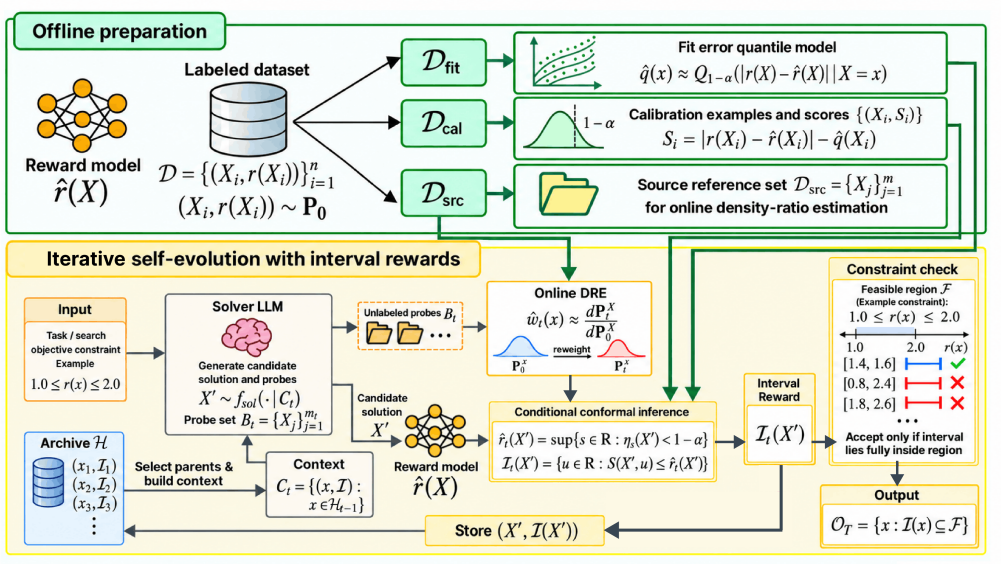


Figure 1: Overview of CISE. Offline, we fit the conditional proxy-error quantile using $\mathcal{D}_{\text{fit}}$, $\mathcal{D}_{\text{cal}}$ provides calibration scores, and $\mathcal{D}_{\text{src}}$ provides the source reference for density-ratio estimation. During evolution, unlabeled probes are used to estimate the current distribution shift, which is incorporated into conditional conformal inference to construct candidate-specific reward intervals. These intervals guide subsequent evolution and candidates are returned only when the entire interval satisfies the task constraint.

evaluation uncertainty (Sheng et al., 2025; Noh et al., 2026). Building on this line of work, we use conformal inference to construct calibrated intervals for proxy rewards and use these reward intervals in both evolutionary feedback and final candidate selection in self-evolving search.

3 Preliminaries: Self-Evolving Agent for Material Search with Proxy Reward

We consider a self-evolving agent that searches a candidate space $\mathcal{X}$ for solutions satisfying task-specific constraints. For each candidate $x \in \mathcal{X}$, $r(x) \in \mathbb{R}$ denotes a scalar true reward, and $\hat{r}(x) \in \mathbb{R}$ denotes an inexpensive and imperfect proxy reward. During search, generated candidates are evaluated using the proxy reward. The task-specific constraints are specified before evolution. Let $\mathcal{F}_r$ and $\mathcal{F}_{\hat{r}}$ denote the sets of candidates satisfying these constraints according to the true and proxy rewards, respectively. Membership in $\mathcal{F}_{\hat{r}}$ is determined by computing $\hat{r}(x)$ and checking it against the specified constraints. A candidate $x \in \mathcal{F}_{\hat{r}} \setminus \mathcal{F}_r$ is a false positive: it satisfies the constraints under the proxy reward but violates them under the true reward.

The archive $\mathcal{H}_t$ stores candidates generated through iteration t together with their proxy rewards. The parent-selection policy $\pi_t(\cdot \mid \mathcal{H}_t)$ selects K candidates from the archive to form the parent set $\mathcal{S}_t$, which guides candidate generation at the next iteration. We treat $\{\pi_t\}_{t=0}^T$ as a given sequence of selection rules; we do not consider a separate policy-learning step. As the archive and its stored rewards change, the selected parents and the resulting candidate generation can nevertheless change across iterations. Algorithm 2 summarizes the proxy-reward-driven search process. At iteration t, the agent uses the previous parents $\mathcal{S}_{t-1}$ and their stored rewards to generate a batch $\mathcal{C}_t = \{x_{t,1}, \ldots, x_{t,M}\}$ of M new candidates. The proxy reward is computed for each candidate, and the resulting candidate-reward pairs are added to the archive. Candidates whose proxy rewards satisfy the task-specific constraints are also added to the cumulative output set, giving $\mathcal{O}_t = \mathcal{O}_{t-1} \cup (\mathcal{C}_t \cap \mathcal{F}_{\hat{r}})$. The updated archive is then used to select the next parent set. The archive and output set serve different purposes: the archive retains generated candidates as search experience, whereas the output set contains only candidates that satisfy the proxy-reward constraints.

4 METHODOLOGY

4.1 PROBLEM SETUP

We are given a labeled source dataset $\mathcal{D} = \{(X_i, r(X_i))\}_{i=1}^n$ of independent candidate-reward pairs drawn from a joint distribution $\mathbf{P}_0$, with candidate marginal $\mathbf{P}_0^X$. We partition $\mathcal{D}$ into three mutually disjoint subsets: $\mathcal{D}_{\text{fit}}$, $\mathcal{D}_{\text{cal}}$, and $\mathcal{D}_{\text{src}}$. The fitting set $\mathcal{D}_{\text{fit}}$ is used to train a model $\hat{q}(x)$ that predicts an upper conditional quantile of the absolute proxy error $|r(x) - \hat{r}(x)|$. The calibration set $\mathcal{D}_{\text{cal}}$ is used to construct conformal reward intervals, while $\mathcal{D}_{\text{src}}$ provides the source reference for density-ratio estimation. Only the candidate information in $\mathcal{D}_{\text{src}}$ is used for density-ratio estimation. The proxy reward model, fitted quantile model, and candidate representation are fixed before evolution.

For a fixed iteration t, let $\mathbf{P}_t$ denote the joint distribution of a candidate and its true reward under the current archive, selected parents, and generation configuration. For a fresh candidate–reward pair $(X', r(X')) \sim \mathbf{P}_t$, CISE uses conditional conformal inference (Gibbs et al., 2025) to construct an iteration-specific prediction interval $\mathcal{I}_t(X')$ for the true reward. This interval represents uncertainty in the proxy reward and is used for both evolutionary feedback and output selection.

For evolutionary feedback, a candidate with a nonempty, bounded interval is evaluated conservatively over that interval using a task-specific rule. The resulting interval reward is stored with the candidate in $\mathcal{H}_t$. The parent-selection policy $\pi_t(\cdot \mid \mathcal{H}_t)$ uses these stored rewards to select the next parents, and subsequent generation uses the selected parents and their rewards as feedback. For output selection, a candidate is added to $\mathcal{O}_t$ only when its interval is nonempty and bounded and every reward value in the interval satisfies the task-specific constraints. A lower-bound constraint is checked using the lower endpoint, an upper-bound constraint using the upper endpoint, and a bounded-range constraint requires the entire interval to lie within the prescribed range. Thus, a candidate with a valid interval may be retained in the archive even when it does not satisfy the output criterion.

Our target is iteration-wise marginal coverage: for a fresh candidate–reward pair at a fixed iteration t, the true reward should lie in its prediction interval with probability at least $1 - \alpha$,

$$\mathbb{P}(r(X') \in \mathcal{I}_t(X')) \geq 1 - \alpha, \qquad t = 1, \ldots, T. \tag{1}$$

The probability is marginal over the source data and the fresh candidate–reward pair under the assumptions stated below. The guarantee is evaluated separately at each iteration and does not assert simultaneous coverage over the full search trajectory. We first analyze this target under exchangeability and then characterize the effect of covariate shift and density-ratio estimation error.

Let A_t denote acceptance at iteration t. Since acceptance requires the entire interval to satisfy the task constraint,

$$A_t \cap \{X' \notin \mathcal{F}_r\} \subseteq \{r(X') \notin \mathcal{I}_t(X')\},$$

and hence $\mathbb{P}(A_t, X' \notin \mathcal{F}_r) \leq \alpha$ whenever equation 1 holds. This controls the joint acceptance-and-false-positive event, not the false-positive probability conditional on acceptance.

4.2 CONFORMAL INTERVAL-DRIVEN SELF-EVOLUTION UNDER EXCHANGEABILITY

We first consider an exchangeable reference setting with $\mathbf{P}_t = \mathbf{P}_0$ for a fixed iteration t before considering covariate shift. Write the calibration pairs as $\{(X_i, r(X_i))\}_{i=1}^{n_{\text{cal}}}$, and let $(X', r(X')) \sim \mathbf{P}_t$ denote an independent fresh candidate-reward pair. The equality $\mathbf{P}_t = \mathbf{P}_0$ means that the calibration candidates and the newly generated candidate follow the same joint distribution over material structures and true rewards. We hold the archive, selected parents, and generation configuration fixed, and assume that the resulting candidate distribution is independent of $\mathcal{D}_{\text{cal}}$.

Let $S(x, u; \hat{r})$ denote a fixed nonconformity score measuring the discrepancy between a hypothetical reward value u and the proxy prediction $\hat{r}(x)$. Let $\hat{q}(x)$ denote a fitted conditional quantile of the absolute proxy error. We define the score as $S(x, u; \hat{r}) = |u - \hat{r}(x)| - \hat{q}(x)$, so that larger values indicate discrepancies that are unusually large relative to the prediction error expected at x. For each calibration pair $(X_i, r(X_i))$, we write $S_i = S(X_i, r(X_i); \hat{r})$. Let $\mathcal{G}$ be a finite-dimensional linear function space spanned by the constant function and fixed functions of the candidate representation $\phi(x)$. For example, $\mathcal{G}$ may be chosen as $\mathcal{G} = \text{span}\{1, \phi_1(x), \ldots, \phi_d(x)\}$. This space allows the calibration adjustment to vary with candidate characteristics.

Following Gibbs et al. (2025), we calibrate the scores through an augmented quantile regression that includes the fresh candidate. For a fixed X' and an imputed test score $s \in \mathbb{R}$, define

$$\widehat{q}_s \in \arg\min_{q \in \mathcal{G}} \left\{ \sum_{i=1}^{n_{\text{cal}}} \ell_\alpha(q(X_i), S_i) + \ell_\alpha(q(X'), s) \right\}, \tag{2}$$

where $\ell_\alpha(a, b) = (1-\alpha)(b-a)_+ + \alpha(a-b)_+$ is the pinball loss for the $(1-\alpha)$-quantile, with $(v)_+ = \max\{v, 0\}$. The additional term incorporates the fresh candidate into calibration using a hypothetical score, allowing the adjustment to account for X' before its true reward is observed.

Let $\eta_s(X') \in [-\alpha, 1-\alpha]$ denote the dual variable associated with the fresh candidate in equation 2. This variable is nondecreasing in s. The nonrandomized score cutoff and the resulting reward interval are

$$\begin{aligned} \widehat{\tau}_t(X') &= \sup\{s \in \mathbb{R} : \eta_s(X') < 1-\alpha\}, \\ \mathcal{I}_t(X') &= \{u \in \mathbb{R} : S(X', u) \le \widehat{\tau}_t(X')\}. \end{aligned} \tag{3}$$

The cutoff is computed from the dual boundary of the unpenalized augmented problem. Inverting the chosen absolute-error score yields an interval centered at $\hat{r}(X')$, with its width determined by the fixed normalization, fitted base quantile, and calibrated cutoff. An empty interval leads to abstention under the output rule in Section 4.1.

Theorem 1 (Iteration-wise marginal coverage under exchangeability)**.** *For a fixed iteration t, assume that the calibration pairs are i.i.d. draws from $\mathbf{P}_0$, and that the fresh pair $(X', r(X'))$ is an independent draw from $\mathbf{P}_t = \mathbf{P}_0$. Suppose that the score S and finite-dimensional linear space $\mathcal{G}$ are fixed independently of these pairs, and that $\mathcal{G}$ contains the constant function. If equation 2 is solved exactly with a dual-solution rule symmetric under permutations of the augmented observations, then the interval in equation 3 satisfies*

$$\mathbb{P}\left(r(X') \in \mathcal{I}_t(X')\right) \ge 1-\alpha.$$

At the true test score, the augmented observations are exchangeable. The constant function in $\mathcal{G}$ imposes a zero-sum constraint on the optimal dual variables. Together with their bounds, this limits the fraction reaching the upper boundary $1-\alpha$ to at most α. The result establishes equation 1 in the reference setting. The next subsection adapts the calibration function space to the current candidate distribution when $\mathbf{P}_t \neq \mathbf{P}_0$.

4.3 Conformal Interval-driven Self-Evolution under Evolutionary Covariate Shift

As evolution proceeds, changes in the archive, selected parents, and generation strategy alter the distribution of materials proposed by the agent. For a fixed iteration t, let $\mathbf{P}_t$ denote the joint distribution of a candidate and its true reward under the current archive, parent set, and generation configuration. We consider evolutionary covariate shift satisfying

$$\mathbf{P}_t(r(X) \mid X) = \mathbf{P}_0(r(X) \mid X), \qquad \mathbf{P}_t^X \ll \mathbf{P}_0^X, \qquad w_t(x) = \frac{d\mathbf{P}_t^X}{d\mathbf{P}_0^X}(x) \le W, \tag{4}$$

where $W < \infty$ is a bounded-overlap constant. Thus, evolution may change how frequently different material structures are generated while preserving the conditional relationship between a structure and its true reward. The density ratio w_t describes this change relative to the source distribution.

At iteration t, we generate an unlabeled probe batch $\mathcal{B}_t$ from $\mathbf{P}_t^X$ using the current parents and their stored rewards. The probe batch is used only for density-ratio estimation, and fresh candidates are subsequently generated for interval evaluation and archive updates. The archive, parents, and generation configuration remain fixed throughout both stages. We assume that the probe candidates and the subsequent candidates are independent draws from the same current distribution, with $\mathcal{D}_{\text{cal}}$ independent of both stages.

Let $\phi(X)$ be the fixed candidate representation, and let $\mathcal{W} = \{w_\theta : w_\theta(x) = h_\theta(\phi(x)) \in [0, W], \theta \in \Theta\}$ be a bounded density-ratio model class. We estimate the current ratio through least-squares

importance fitting (Kanamori et al., 2009):

$$\widehat{\theta}_t \in \arg\min_{\theta\in\Theta} \left\{ \frac{1}{2|\mathcal{D}_{\mathrm{src}}|} \sum_{(X_i,r(X_i))\in\mathcal{D}_{\mathrm{src}}} w_\theta(X_i)^2 - \frac{1}{|\mathcal{B}_t|} \sum_{X'\in\mathcal{B}_t} w_\theta(X') + \lambda\Omega(\theta) \right\}, \qquad \widehat{w}_t = w_{\widehat{\theta}_t}. \tag{5}$$

Here, $\lambda \geq 0$ controls regularization of the ratio estimator through a fixed penalty Ω. The population counterpart of the first two terms equals $\frac{1}{2}\mathbb{E}_{\mathbf{P}_0^X}[(w_\theta(X) - w_t(X))^2]$ up to a constant independent of θ. Equation 5 therefore approximates the current ratio using source candidates and unlabeled target candidates. Repeating this update with $\mathcal{B}_t$ at each iteration gives an online density-ratio estimation procedure for the sequence $\{\mathbf{P}_t^X\}_{t=1}^T$.

We incorporate the estimated ratio into CCI by extending the calibration function space from Section 4.2:

$$\mathcal{G}_t = \mathcal{G} + \mathrm{span}\{\widehat{w}_t\}.$$

The ratio enters as an additional basis function, while the augmented quantile regression retains its unweighted pinball loss and unpenalized calibration objective. For each fresh candidate x, we solve equation 2 over $\mathcal{G}_t$ and construct $\mathcal{I}_t(x)$ through equation 3. Finite-dimensional CCI provides coverage under covariate shifts represented by nonnegative functions in its calibration space (Gibbs et al., 2025). Adding $\widehat{w}_t$ therefore connects the calibration adjustment to the current generation distribution.

For the coverage analysis, suppose that $\mathbb{E}_{\mathbf{P}_0^X}[\widehat{w}_t(X)] > 0$ and define

$$\epsilon_t = \frac{1}{2}\mathbb{E}_{\mathbf{P}_0^X}\left[\left| w_t(X) - \frac{\widehat{w}_t(X)}{\mathbb{E}_{\mathbf{P}_0^X}[\widehat{w}_t(X)]} \right|\right]. \tag{6}$$

The normalized ratio defines a candidate distribution relative to $\mathbf{P}_0^X$, and ϵ_t is its total variation distance from $\mathbf{P}_t^X$. This term captures density-ratio estimation error and misspecification of the representation or ratio model. Normalization preserves membership in $\mathcal{G}_t$, since multiplying a basis function by a positive constant leaves its linear span unchanged.

Algorithm 1 summarizes the resulting search process. The current probe batch determines $\widehat{w}_t$ and $\mathcal{G}_t$, while each fresh candidate receives its own test-point calibration. The archive stores the resulting interval rewards, and the cumulative output contains candidates whose entire intervals satisfy the task-specific constraints.

The following result combines the finite-dimensional CCI guarantee with the distribution discrepancy in Equation 6.

Theorem 2 (Iteration-wise marginal coverage under evolutionary covariate shift)**.** *For a fixed iteration t, condition on the current generation state, $\mathcal{D}_{\mathrm{fit}}$, $\mathcal{D}_{\mathrm{src}}$, and $\mathcal{B}_t$. Assume that, conditionally on these quantities, the calibration pairs remain i.i.d. draws from $\mathbf{P}_0$, and the fresh pair $(X', r(X'))$ is an independent draw from $\mathbf{P}_t$ satisfying equation 4. Suppose that $\widehat{w}_t$ is nonnegative with positive source expectation, and that the score S and finite-dimensional space $\mathcal{G}_t$ are fixed under this conditioning. If the augmented calibration over $\mathcal{G}_t$ is solved exactly with the symmetric dual-solution rule in Theorem 1, then*

$$\mathbb{P}\left(r(X') \in \mathcal{I}_t(X')\right) \geq 1 - \alpha - \epsilon_t,$$

where ϵ_t is defined in equation 6. The probability is conditional on the quantities held fixed and marginal over the calibration data and the fresh candidate–reward pair.

The normalized estimated ratio belongs to $\mathcal{G}_t$, so the finite-dimensional CCI result gives coverage of at least $1 - \alpha$ under its induced candidate distribution. The calibration distribution remains $\mathbf{P}_0$ in both settings, and the conditional reward law is shared. Replacing the induced candidate distribution by $\mathbf{P}_t^X$ therefore changes the coverage probability by at most ϵ_t, establishing Theorem 2. Averaging over the estimation data gives the corresponding marginal bound with $\mathbb{E}[\epsilon_t]$. The same $1-\alpha$ coverage follows directly whenever $w_t \in \mathcal{G}_t$. By the output rule in Section 4.1, the probability of returning a false positive is bounded by $\alpha + \epsilon_t$ under the same fixed-iteration setting.

Algorithm 1 CISE

Require: Initial archive $\mathcal{H}_0$; evolving agent; proxy reward $\hat{r}$; data splits $\mathcal{D}_{\text{fit}}$, $\mathcal{D}_{\text{cal}}$, and $\mathcal{D}_{\text{src}}$; base calibration space $\mathcal{G}$; density-ratio model class $\mathcal{W}$; parent-selection policies $\{\pi_t\}_{t=0}^{T}$; task-specific reward constraints; miscoverage level α; probe budget; candidate budget M; parent budget K; number of iterations T

Ensure: Interval-qualified candidate set $\mathcal{O}_T$

Fit the base quantile model on $\mathcal{D}_{\text{fit}}$ and fix the score S
Compute $S(X_i, r(X_i))$ for every $(X_i, r(X_i)) \in \mathcal{D}_{\text{cal}}$
Initialize the stored rewards in $\mathcal{H}_0$ using the intervals from Section 4.2, and set $\mathcal{O}_0 \leftarrow \emptyset$
Select K initial parents according to $\pi_0(\cdot \mid \mathcal{H}_0)$ and denote them by $\mathcal{S}_0$
for $t = 1, \ldots, T$ **do**
  Fix the current generation configuration using $\mathcal{S}_{t-1}$ and their stored rewards
  Generate an unlabeled probe batch $\mathcal{B}_t$ under this configuration without updating the archive
  Estimate $\widehat{w}_t$ from $\mathcal{D}_{\text{src}}$ and $\mathcal{B}_t$ using equation 5
  Set $\mathcal{G}_t \leftarrow \mathcal{G} + \text{span}\{\widehat{w}_t\}$
  Propose fresh candidates $\mathcal{C}_t = \{x_{t,1}, \ldots, x_{t,M}\}$ independently of $\mathcal{B}_t$ under the same generation configuration
  Initialize $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1}$ and $\mathcal{O}_t \leftarrow \mathcal{O}_{t-1}$
  for all $x \in \mathcal{C}_t$ **do**
    Evaluate $\hat{r}(x)$ and construct $\mathcal{I}_t(x)$ using equation 2 and equation 3 with $\mathcal{G}_t$
    if $\mathcal{I}_t(x)$ is nonempty and bounded **then**
      Evaluate the candidate conservatively over $\mathcal{I}_t(x)$ and store it in $\mathcal{H}_t$
      if every reward value in $\mathcal{I}_t(x)$ satisfies the task-specific constraints **then**
        Update $\mathcal{O}_t \leftarrow \mathcal{O}_t \cup \{x\}$
      end if
    end if
  end for
  Select K parents according to $\pi_t(\cdot \mid \mathcal{H}_t)$ and denote them by $\mathcal{S}_t$
end for
return $\mathcal{O}_T$

5 Experiments

5.1 Experimental Setup

We evaluate CISE on three multi-constraint material-discovery tasks from the LLEMA benchmark (Abhyankar et al., 2026): wide-bandgap semiconductors (WBG), which require $E_g \geq 2.5$ eV and $E_{\text{form}} \leq -1.0$ eV/atom; solid-state electrolytes (SSE), which require $E_{\text{form}} \leq -1.0$ eV/atom, $E_g \geq 2.0$ eV, and at least one mobile species from $\{\text{Li}, \text{Na}, \text{K}, \text{Mg}, \text{Ca}, \text{Al}\}$; and photovoltaic absorbers (PV), which require $0.7 \leq E_g \leq 2.0$ eV, $E_{\text{form}} \leq 0.0$ eV/atom, and compositions restricted to earth-abundant, non-toxic elements. A candidate is counted as a task hit only when all corresponding constraints are satisfied. We provide detailed descriptions of each task in Appendix A.2. We set $\alpha = 0.1/\#\text{constraint}$ for each property. For example, we set $\alpha = 0.05$ for both E_g and E_{form} in the WBG task.

We compare CISE against MatterGen (Zeni et al., 2025), LLMatDesign (Jia et al., 2024), and LLEMA (Abhyankar et al., 2026). CISE modifies only the archive-admission and output-selection steps of LLEMA; its generator and evolutionary operators remain unchanged. For proxy evaluation, we use MGT in place of CGCNN for band-gap prediction, while otherwise following the LLEMA experimental protocol, including the task definitions, hierarchical proxy-property pipeline, proxy-reward construction, candidate budgets, and method-specific generation settings. For tasks involving multiple predicted properties, we apply CISE separately to each property and return only candidates that satisfy all property-specific CISE criteria. During search, surrogate predictions provide the proxy properties used to score and qualify candidates. We then re-evaluate the candidates returned by each method using DFT implemented in Quantum ESPRESSO (Giannozzi et al., 2009). A proxy hit is a candidate satisfying the task constraints under surrogate predictions, whereas a DFT hit satisfies the same physical-property constraints under Quantum ESPRESSO calculations, together with any deterministic compositional restrictions. A proxy hit that fails this final evaluation is recorded as a

Table 1: Comparison of output reliability across three material-discovery tasks. Each method generates 200 candidate structures. **Output** denotes candidates satisfying the method-specific output criterion. **TP** and **FP** denote output candidates that satisfy or violate the target constraints under DFT evaluation, respectively. Values in parentheses denote the fraction among output candidates.

	Wide-bandgap (WBG)			**Solid-state electrolytes (SSE)**			**Photovoltaic absorbers (PV)**		
Method	Output	TP	FP	Output	TP	FP	Output	TP	FP
MatterGen	3/200	2 (0.67)	1 (0.33)	2/200	1 (0.50)	1 (0.50)	0/200	–	–
LLMatDesign	37/200	0 (0.00)	37 (1.00)	2/200	0 (0.00)	2 (1.00)	18/200	0 (0.00)	18 (1.00)
LLEMA	102/200	94 (0.92)	8 (0.08)	104/200	71 (0.68)	33 (0.32)	60/200	39 (0.65)	21 (0.35)
CISE	**41/200**	**41 (1.00)**	**0 (0.00)**	**51/200**	**51 (1.00)**	**0 (0.00)**	**6/200**	**6 (1.00)**	**0 (0.00)**

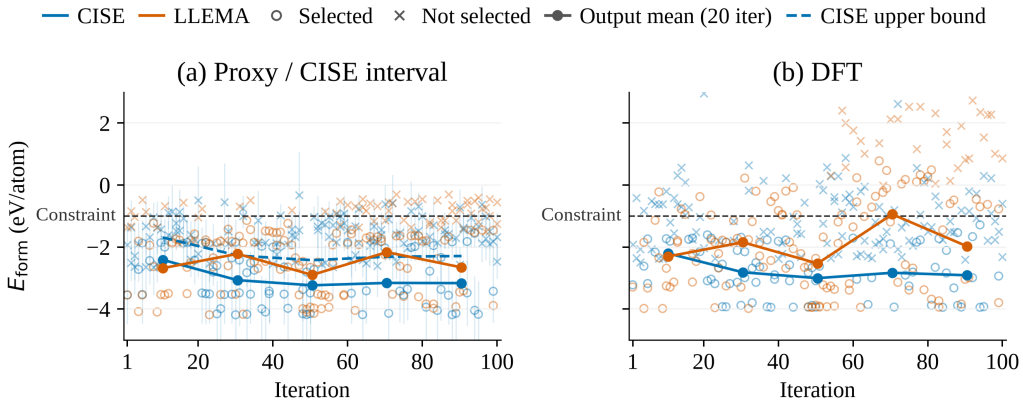


Figure 2: Evolution of formation energy on the solid-state electrolyte (SSE) task, where success requires $E_{\text{form}} \leq -1.0$ eV/atom. (a) Proxy rewards for LLEMA and reward intervals for CISE; vertical blue bars denote CISE intervals and the dashed blue curve shows their mean upper bound. (b) Corresponding DFT evaluations. Circles and crosses denote selected and non-selected candidates, respectively, and solid curves show the 20-iteration output mean. In later iterations, LLEMA produces a pronounced concentration of candidates above the DFT success threshold, visible in the upper-right of (b), including selected false positives. In contrast, CISE shows no comparable accumulation and its output mean remains within the feasible region throughout evolution.

false positive. We use gpt-5-mini (OpenAI, 2026) as a solver LLM to generate candidates. More detailed experimental settings are provided in Appendix A.

5.2 Empirical results

Main results. Table 1 shows a trade-off between the number of returned structures and their DFT results. This contrast is particularly clear between LLEMA and CISE on WBG and SSE. LLEMA returns 102 and 104 candidates, respectively, but 7.84% and 31.73% fail DFT validation. In contrast, CISE returns 41 WBG candidates and 51 SSE candidates, with no observed false positives among candidates with resolved DFT evaluations: all 41 WBG candidates and 51 SSE candidates are confirmed true positives. A smaller but more reliable shortlist can be valuable when downstream experimental validation is the bottleneck. The objective is not simply to maximize the number of structures satisfying proxy constraints, but to prioritize candidates worth further investigation. Under a limited validation budget, higher output precision can reduce the fraction of resources allocated to candidates that do not satisfy the intended properties. We therefore view the WBG and SSE results as a reliability-yield trade-off. This trade-off is consistent with CISE's joint interval-based selection rule. Whereas proxy-based selection checks whether each point prediction satisfies its constraint, CISE requires the entire prediction interval of every required property to lie within the corresponding

Table 2: Equal post-evaluation budget comparison on the photovoltaic absorber (PV) task. To match the six candidates returned by CISE, we select six candidates from the LLEMA output whose proxy-predicted band gaps are closest to the target value of 1.2 eV. Importantly, this ranking is performed before DFT evaluation and therefore does not use the TP/FP labels. Under the same validation budget, three of the six LLEMA candidates are false positives under DFT, whereas all six candidates returned by CISE are true positives.

	LLEMA: top-6 closest to 1.2 eV				CISE: six returned candidates			
Rank	Material	Proxy E_g	DFT E_g	Result	Material	CISE interval	DFT E_g	Result
1	SrO	1.24	2.99	FP	RbLiO	$[0.73, 1.98]$	1.24	TP
2	ZnF_2	1.25	2.32	FP	$RbNa_2OF$	$[1.34, 1.51]$	1.32	TP
3	$MgTiO_3$	1.37	1.65	TP	KNa_2OF	$[1.19, 1.62]$	1.50	TP
4	MgO	1.54	1.39	TP	$Na_3Li_3P_2$	$[0.92, 1.72]$	1.08	TP
5	$MgZrO_3$	1.57	1.07	TP	$NaLi(MgP)_2$	$[0.83, 1.86]$	1.66	TP
6	CaO	1.75	3.06	FP	KLi_5P_2	$[0.89, 1.73]$	1.00	TP
DFT outcome			3 TP / 3 FP			6 TP / 0 FP		

feasible region. A candidate can therefore satisfy all point constraints yet be excluded because at least one interval extends beyond an admissible boundary.

Calibrated interval rewards facilitate reliable LLM evolution. Figure 2 shows a clear difference in the search trajectories of LLEMA and CISE on the SSE task. Although LLEMA continues to assign favorable proxy rewards, its DFT-evaluated candidates increasingly populate the failure region in later iterations. This is particularly visible in the upper-right of Figure 2(b), where many late-iteration LLEMA candidates lie above the $E_{\mathrm{form}} = -1.0$ threshold, including selected false positives. In contrast, CISE shows no comparable accumulation, and its output mean remains within the feasible region throughout evolution. This pattern suggests that calibrated interval rewards help prevent overly optimistic proxy evaluations from being repeatedly reinforced during LLM evolution.

Comparison of fixed-budget shortlists of LLEMA and CISE. To examine whether the higher reliability of CISE is simply a consequence of returning fewer candidates, we perform an equal-budget comparison on the PV task. CISE returns six candidates in this experiment, so we restrict LLEMA to the same downstream validation budget of six structures. Among the proxy-qualified LLEMA outputs, we select the six candidates whose proxy-predicted band gaps are closest to 1.2 eV, representing a natural point-estimate-based prioritization strategy for PV screening. Importantly, this selection is performed before DFT evaluation, so the true-positive and false-positive labels are unavailable when the candidates are ranked. As shown in Table 2, only three of the six selected LLEMA candidates satisfy the PV band-gap constraint after DFT evaluation, while the remaining three are false positives. In contrast, all six candidates returned by CISE are confirmed as true positives. This result indicates that the reliability gain of CISE cannot be explained solely by its smaller output set: even under the same validation budget and a stricter post-hoc truncation of the LLEMA output, point-estimate-based prioritization can still select candidates whose apparent quality is caused by proxy error, whereas interval-based selection excludes such uncertain candidates before expensive validation.

6 Conclusion

In this work, we study reliable self-evolution when accurate true rewards are too expensive to obtain throughout iterative search. We propose CISE, which replaces point-valued proxy rewards with calibrated reward intervals constructed using conditional conformal inference and iteration-wise online density-ratio estimation. These intervals are used for both evolutionary feedback and final candidate selection, allowing proxy-reward uncertainty to directly influence the search process. Under explicit independence and covariate-shift assumptions, we provide fixed-iteration coverage guarantees and characterize the effect of density-ratio estimation error. Across three material-discovery tasks, CISE consistently returns a smaller but more reliable set of candidates, with no observed false positives among its DFT-validated outputs and improved reliability under a matched validation budget. Our results suggest that calibrated interval rewards provide a practical route toward more reliable self-evolving agents when true rewards are costly to obtain.

REFERENCES

Nikhil Abhyankar, Sanchit Kabra, Saaketh Desai, and Chandan K. Reddy. Llema: Evolutionary search with llms for multi-objective materials discovery, 2026. URL https://arxiv.org/abs/2510.22503.

Anastasios N. Angelopoulos and Stephen Bates. A Gentle Introduction to Conformal Prediction and Distribution-Free Uncertainty Quantification, 2022. *arXiv preprint arXiv:2107.07511*.

Kamal Choudhary. The jarvis infrastructure is all you need for materials design. *Computational Materials Science*, 259:114063, 2025. ISSN 0927-0256. doi: 10.1016/j.commatsci.2025.114063. URL http://dx.doi.org/10.1016/j.commatsci.2025.114063.

Thomas Coste, Usman Anwar, Robert Kirk, and David Krueger. Reward model ensembles help mitigate overoptimization, 2024. URL https://arxiv.org/abs/2310.02743.

Paolo Giannozzi, Stefano Baroni, Nicola Bonini, Matteo Calandra, Roberto Car, Carlo Cavazzoni, Davide Ceresoli, Guido L Chiarotti, Matteo Cococcioni, Ismaila Dabo, Andrea Dal Corso, Stefano de Gironcoli, Stefano Fabris, Guido Fratesi, Ralph Gebauer, Uwe Gerstmann, Christos Gougoussis, Anton Kokalj, Michele Lazzeri, Layla Martin-Samos, Nicola Marzari, Francesco Mauri, Riccardo Mazzarello, Stefano Paolini, Alfredo Pasquarello, Lorenzo Paulatto, Carlo Sbraccia, Sandro Scandolo, Gabriele Sclauzero, Ari P Seitsonen, Alexander Smogunov, Paolo Umari, and Renata M Wentzcovitch. Quantum espresso: a modular and open-source software project for quantum simulations of materials. *Journal of Physics: Condensed Matter*, 21(39):395502, 2009. ISSN 1361-648X. doi: 10.1088/0953-8984/21/39/395502. URL http://dx.doi.org/10.1088/0953-8984/21/39/395502.

Isaac Gibbs, John J Cherian, and Emmanuel J Candès. Conformal prediction with conditional guarantees. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 87(4): 1100–1126, 2025. ISSN 1369-7412. doi: 10.1093/jrsssb/qkaf008.

Juraj Gottweis, Wei-Hung Weng, Alexander Daryin, Tao Tu, Petar Sirkovic, Artiom Myaskovsky, Grzegorz Glowaty, Felix Weissenberger, Alessio Orlandi, Dan Popovici, Anil Palepu, Keran Rong, Ryutaro Tanno, Khaled Saab, Fan Zhang, Jacob Blum, Andrew Carroll, Kavita Kulkarni, Nenad Tomašev, Dina Zverinski, Ivor Rendulic, Elahe Vedadi, Florian Hasler, Luka Rimanic, Marina Boia, Ivan Budiselic, Ben Feinstein, Mathias Bellaiche, Tom Sheffer, Jan Freyberg, Jeremy Ratcliff, Ottavia Bertolli, Katherine Chou, Avinatan Hassidim, Burak Gokturk, Amin Vahdat, Yuan Guan, Vikram Dhillon, Eeshit Dhaval Vaishnav, Byron Lee, Tiago R. D. Costa, José R. Penadés, Gary Peltz, Yossi Matias, James Manyika, Demis Hassabis, Yunhan Xu, Pushmeet Kohli, Annalisa Pawlosky, Alan Karthikesalingam, and Vivek Natarajan. Accelerating scientific discovery with co-scientist. *Nature*, 655(8122):487–496, May 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10644-y. URL http://dx.doi.org/10.1038/s41586-026-10644-y.

Yatai Ji, Jiacheng Zhang, Jie Wu, Shilong Zhang, Shoufa Chen, Chongjian GE, Peize Sun, Weifeng Chen, Wenqi Shao, Xuefeng Xiao, Weilin Huang, and Ping Luo. Prompt-a-video: Prompt your video diffusion model via preference-aligned llm, 2024. URL https://arxiv.org/abs/2412.15156.

Shuyi Jia, Chao Zhang, and Victor Fung. Llmatdesign: Autonomous materials discovery with large language models, 2024. URL https://arxiv.org/abs/2406.13163.

Takafumi Kanamori, Shohei Hido, and Masashi Sugiyama. A least-squares approach to direct importance estimation. *Journal of Machine Learning Research*, 10(48):1391–1445, 2009. URL http://jmlr.org/papers/v10/kanamori09a.html.

Shu Liu, Shubham Agarwal, Monishwaran Maheswaran, Mert Cemri, Zhifei Li, Qiuyang Mang, Ashwin Naren, Ethan Boneh, Audrey Cheng, Melissa Z. Pan, Alexander Du, Kurt Keutzer, Alvin Cheung, Alexandros G. Dimakis, Koushik Sen, Matei Zaharia, and Ion Stoica. Evox: Meta-evolution for automated discovery, 2026. URL https://arxiv.org/abs/2602.23413.

Kangjun Noh, Ryunyi LEE, Heejin Jung, Young Jin Kim, and Kyungwoo Song. Reliable LLM-as-a-judge evaluation via transferable calibration. In *ICML 2026 Statistical Frameworks for Uncertainty in Agentic Systems*, 2026. URL https://openreview.net/forum?id=AIBoVLkc31.

Alexander Novikov, Ngân Vũ, Marvin Eisenberger, Emilien Dupont, Po-Sen Huang, Adam Zsolt Wagner, Sergey Shirobokov, Borislav Kozlovskii, Francisco J. R. Ruiz, Abbas Mehrabian, M. Pawan Kumar, Abigail See, Swarat Chaudhuri, George Holland, Alex Davies, Sebastian Nowozin, Pushmeet Kohli, and Matej Balog. Alphaevolve: A coding agent for scientific and algorithmic discovery, 2025. URL https://arxiv.org/abs/2506.13131.

OpenAI. Openai gpt-5 system card, 2026. URL https://arxiv.org/abs/2601.03267.

Young-Jin Park, Kristjan Greenewald, Kaveh Alim, Hao Wang, and Navid Azizan. Know what you don't know: Uncertainty calibration of process reward models, 2025. URL https://arxiv.org/abs/2506.09338.

Bhaktipriya Radharapu, Eshika Saxena, Kenneth Li, Chenxi Whitehouse, Adina Williams, and Nicola Cancedda. Calibrating llm judges: Linear probes for fast and reliable uncertainty estimation, 2025. URL https://arxiv.org/abs/2512.22245.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475, 2024.

Aishwarya Sahoo, Jeevana Kruthi Karnuthala, Tushar Parmanand Budhwani, Pranchal Agarwal, Sankaran Vaidyanathan, Alexa Siu, Franck Dernoncourt, Jennifer Healey, Nedim Lipka, Ryan A. Rossi, Uttaran Bhattacharya, and Branislav Kveton. Quantitative LLM judges using post-hoc score calibration. *Transactions on Machine Learning Research*, 2026. ISSN 2835-8856. URL https://openreview.net/forum?id=isUKzVxmOB.

Huanxin Sheng, Xinyi Liu, Hangfeng He, Jieyu Zhao, and Jian Kang. Analyzing uncertainty of LLM-as-a-judge: Interval evaluations with conformal prediction. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 11286–11328, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.569. URL https://aclanthology.org/2025.emnlp-main.569/.

Vladimir Vovk, Alexander Gammerman, and Glenn Shafer. *Algorithmic learning in a random world*. Springer, 2005.

Claudio Zeni, Robert Pinsler, Daniel Zügner, Andrew Fowler, Matthew Horton, Xiang Fu, Zilong Wang, Aliaksandra Shysheya, Jonathan Crabbé, Shoko Ueda, et al. A generative model for inorganic materials design. *Nature*, 639(8055):624–632, 2025.

Algorithm 2 PROXY-REWARD-DRIVEN SELF-EVOLVING SEARCH

Require: Initial archive $\mathcal{H}_0$; evolving agent; proxy reward $\hat{r}$; parent-selection policies $\{\pi_t\}_{t=0}^{T}$; proxy-reward constraints; candidate budget M; parent budget K; number of iterations T
Ensure: Proxy-qualified candidate set $\mathcal{O}_T$
1: Initialize $\mathcal{O}_0 \leftarrow \emptyset$
2: Select K initial parents according to $\pi_0(\cdot \mid \mathcal{H}_0)$ and denote them by $\mathcal{S}_0$
3: **for** $t = 1, \ldots, T$ **do**
4: Propose $\mathcal{C}_t = \{x_{t,1}, \ldots, x_{t,M}\}$ using $\mathcal{S}_{t-1}$ and their stored proxy rewards
5: Evaluate $\hat{r}(x)$ for every $x \in \mathcal{C}_t$
6: Update $\mathcal{H}_t \leftarrow \mathcal{H}_{t-1} \cup \{(x, \hat{r}(x)) : x \in \mathcal{C}_t\}$
7: Update $\mathcal{O}_t \leftarrow \mathcal{O}_{t-1} \cup (\mathcal{C}_t \cap \mathcal{F}_{\hat{r}})$
8: Select K parents according to $\pi_t(\cdot \mid \mathcal{H}_t)$ and denote them by $\mathcal{S}_t$
9: **end for**
10: **return** $\mathcal{O}_T$

AI USE STATEMENT

We used large language models (LLMs) throughout the writing process to edit text, improve clarity, and generate method illustrations; we also used coding agents.

A DETAILED EXPERIMENTAL SETTINGS

A.1 IMPLEMENTATIONS

Calibration set. We construct the calibration set $\mathcal{D}_{\text{cal}}$ using 10,000 material structures from JARVIS-QE-TB (QETB) (Choudhary, 2025). The calibration samples are separated from the model-fitting data at the composition-group level to avoid overlap between the two sets.

Model-fitting set. We use $\mathcal{D}_{\text{fit}}$ consisting of 423,205 structures from QETB to train the conditional quantile regression model. The proxy models remain fixed throughout the evolution process.

Probe generation. At each generation context, we generate eight valid probe structures using the same parent structures and prompt context as the main candidate generation. These probes are used to characterize the current candidate distribution for online density-ratio estimation.

Conditional quantile regression. We implement the conditional quantile regression model using a gradient-boosted regression model (HistGradientBoostingRegressor) with quantile loss. The model is trained to predict the conditional 0.95 quantile of absolute proxy error. Its input consists of 27 structural features together with the corresponding proxy prediction.

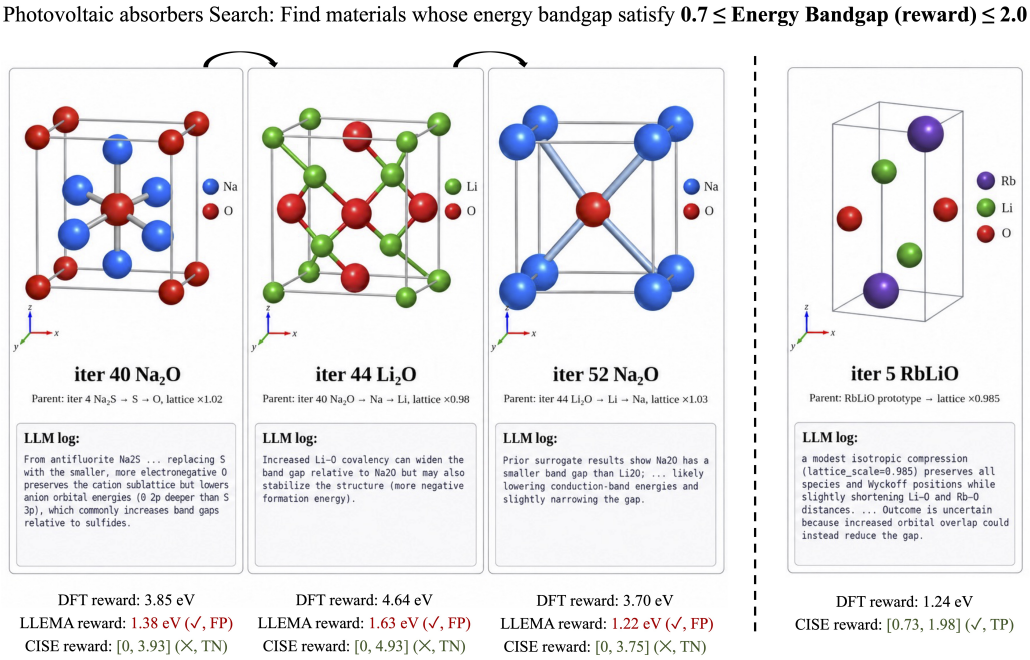
CCI features. For conditional conformal inference, the structural features are used to capture heterogeneous proxy uncertainty across materials. The online density ratio estimated from the probe structures and a fixed source set is additionally incorporated into the finite-dimensional Gibbs calibration basis, allowing the calibration procedure to adapt to the candidate distribution at the current generation step.

DFT evaluation. Selected material structures are evaluated using Quantum ESPRESSO under a PBEsol fixed-structure protocol consistent with the QETB calculations. We use the Modified GBRV PBEsol pseudopotentials and evaluate the target material properties independently from the proxy models.

A.2 DETAILED TASK DESCRIPTIONS

- **Wide-bandgap semiconductors (WBG):** the target constraints are $E_g \geq 2.5$ eV and $E_{\text{form}} \leq -1.0$ eV/atom. Here, $\hat{r}(X) = (\widehat{E}_g, \widehat{E}_{\text{form}})$ and $r(X) = (E_g, E_{\text{form}})$, where the latter quantities are obtained by DFT. Since a larger predicted E_g and a smaller predicted E_{form} can make a candidate appear more favorable,

$$Y(X) = \mathbb{I}\left\{\widehat{E}_g - E_g \leq \delta_{E_g},\ E_{\text{form}} - \widehat{E}_{\text{form}} \leq \delta_{E_{\text{form}}}\right\}.$$



(a) False-positive propagation of proxy reward-based evolution

(b) Output example of CISE

Figure 3: Qualitative comparison of proxy-driven evolution and CISE on the photovoltaic absorber task, with target band gap $0.7 \le E_g \le 2.0$ eV. (a) LLEMA reuses proxy-qualified candidates along the parent–child lineage $\mathrm{Na_2O} \to \mathrm{Li_2O} \to \mathrm{Na_2O}$. All three candidates satisfy the point-estimate constraint but fail DFT validation (FP). The recorded LLM design justifications illustrate how previous surrogate evaluations inform subsequent proposals, allowing false positives to become reusable search feedback. The displayed CISE intervals extend beyond the target range, so these candidates would be excluded by the interval-based output rule. (b) A RbLiO candidate from a separate CISE run has its entire band-gap interval within the target range and also satisfies the constraint under DFT evaluation. Arrows indicate parent–child relations; checkmarks and crosses indicate acceptance and rejection, respectively.

- **Solid-state electrolytes (SSE):** the constraints are $E_{\mathrm{form}} \le -1.0$ eV/atom, $E_g \ge 2.0$ eV, and the presence of at least one mobile species from $\{\mathrm{Li}, \mathrm{Na}, \mathrm{K}, \mathrm{Mg}, \mathrm{Ca}, \mathrm{Al}\}$. We use $\hat{r}(X) = (\widehat{E}_g, \widehat{E}_{\mathrm{form}})$ and $r(X) = (E_g, E_{\mathrm{form}})$, with
$$Y(X) = \mathbb{I}\left\{\widehat{E}_g - E_g \le \delta_{E_g},\ E_{\mathrm{form}} - \widehat{E}_{\mathrm{form}} \le \delta_{E_{\mathrm{form}}}\right\}.$$
The mobile-species requirement is determined directly from the composition and therefore does not require proxy calibration.
- **Photovoltaic absorbers (PV):** the constraints are $0.7 \le E_g \le 2.0$ eV, $E_{\mathrm{form}} \le 0.0$ eV/atom, and compositions restricted to earth-abundant, non-toxic elements. We again use $\hat{r}(X) = (\widehat{E}_g, \widehat{E}_{\mathrm{form}})$ and $r(X) = (E_g, E_{\mathrm{form}})$. Because E_g is constrained on both sides, its proxy error is controlled in both directions:
$$Y(X) = \mathbb{I}\left\{|\widehat{E}_g - E_g| \le \delta_{E_g},\ E_{\mathrm{form}} - \widehat{E}_{\mathrm{form}} \le \delta_{E_{\mathrm{form}}}\right\}.$$
The elemental restriction is evaluated directly from the composition.

B Additional Empirical Results

Interval rewards mitigate false-positive propagation. Figure 3 qualitatively illustrates how proxy errors can propagate through self-evolving search and how CISE changes this feedback process. In

the LLEMA trajectory in Figure 3(a), the agent repeatedly reuses proxy-qualified candidates along the parent-child lineage $\mathrm{Na_2O} \to \mathrm{Li_2O} \to \mathrm{Na_2O}$. Although all three candidates satisfy the photovoltaic band-gap constraint according to their proxy predictions, DFT evaluation shows that each is a false positive. Importantly, the recorded LLM rationales explicitly refer to previous surrogate evaluations when proposing subsequent candidates, showing that an erroneous proxy reward can become reusable feedback rather than remaining an isolated output error. In contrast, the corresponding CISE intervals extend beyond the feasible band-gap range, preventing these candidates from being treated as reliable successes under the interval-based criterion. Figure 3(b) shows the complementary case: the full CISE interval for RbLiO lies within the target range, and the candidate is subsequently confirmed as a true positive by DFT. Together, these examples illustrate how interval rewards can suppress the propagation of uncertain proxy successes while retaining candidates whose uncertainty-aware rewards remain feasible.

C PROOFS

Throughout this section, we fix an iteration t and write $N = n_{\mathrm{cal}} + 1$. All scores, basis functions, and solution rules are taken to be measurable. Symmetry of the dual-solution rule means that permuting the augmented observations permutes the selected dual coordinates in the same way. We interpret the supremum in Equation 3 in the extended real line, with $\sup \emptyset = -\infty$.

C.1 PROOF OF THEOREM 1

Proof. Condition on the independently fixed score S and calibration space $\mathcal{G}$. Set $X_N = X'$ and write $S_i = S(X_i, r(X_i))$ for $i = 1, \ldots, N$. At the true test score S_N, the augmented calibration problem in equation 2 becomes

$$\min_{q \in \mathcal{G}} \sum_{i=1}^{N} \ell_\alpha(q(X_i), S_i).$$

The pinball loss admits the representation

$$\ell_\alpha(a, b) = \max_{-\alpha \le \eta \le 1-\alpha} \eta(b - a).$$

Since $\mathcal{G}$ is a finite-dimensional linear space, the augmented problem is a finite-dimensional linear program. It is feasible and has a finite optimal value: its objective is nonnegative, and the choice $q = 0$ gives a finite value. Linear-programming duality therefore yields the dual problem

$$\begin{aligned} \max_{\eta \in \mathbb{R}^N} \quad & \sum_{i=1}^{N} \eta_i S_i \\ \text{subject to} \quad & -\alpha \le \eta_i \le 1 - \alpha, \qquad i = 1, \ldots, N, \\ & \sum_{i=1}^{N} \eta_i g(X_i) = 0, \qquad \text{for every } g \in \mathcal{G}. \end{aligned} \tag{7}$$

The last constraint follows because the coefficients of $q \in \mathcal{G}$ are unrestricted. The dual feasible set is nonempty and compact, so an optimal dual solution exists.

Let $\eta = (\eta_1, \ldots, \eta_N)$ be the optimal dual solution selected by the symmetric rule. Since the constant function belongs to $\mathcal{G}$, equation 7 implies

$$\sum_{i=1}^{N} \eta_i = 0.$$

The box constraints also give, for every i,

$$\mathbf{1}\{\eta_i = 1 - \alpha\} \le \eta_i + \alpha.$$

Consequently,

$$\sum_{i=1}^{N} \mathbf{1}\{\eta_i = 1 - \alpha\} \le \sum_{i=1}^{N} (\eta_i + \alpha) = N\alpha.$$

Under the assumptions of the theorem, the N augmented pairs are independent draws from $\mathbf{P}_0$. The fixed score and symmetric dual-solution rule therefore make the selected dual coordinates exchangeable. It follows that

$$\begin{aligned}\mathbb{P}(\eta_N = 1-\alpha) &= \frac{1}{N}\mathbb{E}\left[\sum_{i=1}^{N}\mathbf{1}\{\eta_i = 1-\alpha\}\right] \\ &\leq \alpha.\end{aligned}$$

At the true test score, $\eta_N = \eta_{S_N}(X_N)$. If $\eta_N < 1-\alpha$, then S_N belongs to the set whose supremum defines $\widehat{\tau}_t(X_N)$. Thus, $S_N \leq \widehat{\tau}_t(X_N)$, which implies $r(X_N) \in \mathcal{I}_t(X_N)$ by equation 3. We therefore have the pointwise inequalities

$$\mathbf{1}\{r(X_N) \notin \mathcal{I}_t(X_N)\} \leq \mathbf{1}\{\eta_N = 1-\alpha\} \leq \eta_N + \alpha. \tag{8}$$

Combining the first inequality with the boundary-probability bound gives

$$\mathbb{P}\big(r(X') \notin \mathcal{I}_t(X')\big) \leq \alpha.$$

Averaging over any independently fitted quantities conditioned on above proves the stated guarantee.

Only an inclusion between the miscoverage event and the dual boundary event is needed. In particular, the argument remains valid when the true score equals the cutoff or when optimal solutions are not unique, provided that the selected dual solution follows the stated symmetric rule. □

We also verify the monotonicity of $\eta_s(X')$ used in Section 4.2. Fix the calibration observations, the fresh candidate X', and the calibration space. For $s_1 < s_2$, let $\eta^{(1)}$ and $\eta^{(2)}$ be optimal dual solutions with imputed test scores s_1 and s_2, respectively. The dual feasible set in equation 7 does not depend on the imputed score. Optimality therefore gives

$$\begin{aligned}\sum_{i=1}^{N-1} S_i\eta_i^{(1)} + s_1\eta_N^{(1)} &\geq \sum_{i=1}^{N-1} S_i\eta_i^{(2)} + s_1\eta_N^{(2)}, \\ \sum_{i=1}^{N-1} S_i\eta_i^{(2)} + s_2\eta_N^{(2)} &\geq \sum_{i=1}^{N-1} S_i\eta_i^{(1)} + s_2\eta_N^{(1)}.\end{aligned}$$

Adding these inequalities yields

$$(s_2 - s_1)\big(\eta_N^{(2)} - \eta_N^{(1)}\big) \geq 0.$$

Hence $\eta_{s_2}(X') \geq \eta_{s_1}(X')$. This argument applies to any choices of optimal dual solutions and remains valid when $\mathcal{G}$ is replaced by a fixed $\mathcal{G}_t$.

C.2 Proof of Theorem 2

Proof. Condition on the current generation state, $\mathcal{D}_{\mathrm{fit}}$, $\mathcal{D}_{\mathrm{src}}$, and $\mathcal{B}_t$, as specified in the theorem. All probabilities and expectations below are conditional on these quantities unless stated otherwise. In particular, the score S, estimated ratio $\widehat{w}_t$, and calibration space $\mathcal{G}_t$ are fixed. By assumption, the calibration pairs retain their independent $\mathbf{P}_0$ laws under this conditioning.

Define

$$\mu_t = \mathbb{E}_{\mathbf{P}_0^X}[\widehat{w}_t(X)], \qquad \widetilde{w}_t(x) = \frac{\widehat{w}_t(x)}{\mu_t}.$$

The bounded ratio model and the positive-source-expectation assumption give $0 < \mu_t < \infty$. Therefore, $\widetilde{w}_t$ is nonnegative and satisfies

$$\mathbb{E}_{\mathbf{P}_0^X}[\widetilde{w}_t(X)] = 1, \qquad \widetilde{w}_t \in \mathcal{G}_t.$$

The latter inclusion follows from $\widehat{w}_t \in \mathcal{G}_t$ and linearity of $\mathcal{G}_t$.

Let $\widetilde{\mathbf{P}}_t$ be the joint distribution defined by

$$\frac{d\widetilde{\mathbf{P}}_t}{d\mathbf{P}_0}(x,u) = \widetilde{w}_t(x).$$

This is a probability distribution because $\widetilde{w}_t$ is nonnegative and has source expectation one. It changes only the candidate marginal while preserving the conditional reward law of $\mathbf{P}_0$.

First consider an auxiliary experiment in which all N augmented pairs are independent draws from $\mathbf{P}_0$. Let $\mathbb{E}_0$ denote expectation in this experiment, with the conditioned quantities above still held fixed. For each realization of the augmented sample, solve the same calibration problem over $\mathcal{G}_t$ at the true test score. Let η denote the selected optimal dual solution.

The dual formulation in equation 7 applies with $\mathcal{G}_t$ in place of $\mathcal{G}$. Since $\widetilde{w}_t \in \mathcal{G}_t$, dual feasibility gives

$$\sum_{i=1}^{N} \widetilde{w}_t(X_i)\eta_i = 0.$$

Under the auxiliary source experiment, the augmented pairs are exchangeable. Since $\widetilde{w}_t$ is fixed and the dual-solution rule is symmetric, the pairs $(\widetilde{w}_t(X_i), \eta_i)$ are also exchangeable. Consequently,

$$\begin{aligned}\mathbb{E}_0[\widetilde{w}_t(X_N)\eta_N] &= \frac{1}{N}\mathbb{E}_0\left[\sum_{i=1}^{N} \widetilde{w}_t(X_i)\eta_i\right] \\ &= 0.\end{aligned}$$

These expectations are well-defined because the dual variables are bounded and $\mathbb{E}_0[\widetilde{w}_t(X_N)] = 1$.

The pointwise bound in equation 8 also applies to the calibration problem over $\mathcal{G}_t$. Multiplying it by $\widetilde{w}_t(X_N) \geq 0$ and taking expectations gives

$$\begin{aligned}&\mathbb{E}_0\left[\widetilde{w}_t(X_N)\mathbf{1}\{r(X_N) \notin \mathcal{I}_t(X_N)\}\right] \\ &\quad \leq \mathbb{E}_0\left[\widetilde{w}_t(X_N)(\eta_N + \alpha)\right] = \alpha.\end{aligned} \tag{9}$$

Changing measure for the last observation only, the left-hand side of equation 9 is exactly the miscoverage probability when the calibration pairs are independent draws from $\mathbf{P}_0$ and the independent fresh pair has law $\widetilde{\mathbf{P}}_t$. Thus, the prediction interval has miscoverage at most α under this auxiliary target law. The fitted score, estimated ratio, calibration space, and interval-construction rule remain unchanged in this comparison.

It remains to replace $\widetilde{\mathbf{P}}_t$ by the actual target distribution $\mathbf{P}_t$. The covariate-shift assumption in equation 4 implies

$$\frac{d\mathbf{P}_t}{d\mathbf{P}_0}(x, u) = w_t(x).$$

Since both target laws preserve the conditional reward law of $\mathbf{P}_0$, their total variation distance is

$$\begin{aligned}d_{\mathrm{TV}}\big(\mathbf{P}_t, \widetilde{\mathbf{P}}_t\big) &= \frac{1}{2}\mathbb{E}_{\mathbf{P}_0^X}\left[|w_t(X) - \widetilde{w}_t(X)|\right] \\ &= \epsilon_t,\end{aligned}$$

where the last equality follows from equation 6.

The calibration sample has the same independent law under both target distributions. Therefore,

$$\begin{aligned}&d_{\mathrm{TV}}\left(\mathbf{P}_0^{\otimes n_{\mathrm{cal}}} \otimes \mathbf{P}_t,\, \mathbf{P}_0^{\otimes n_{\mathrm{cal}}} \otimes \widetilde{\mathbf{P}}_t\right) \\ &\quad = d_{\mathrm{TV}}\big(\mathbf{P}_t, \widetilde{\mathbf{P}}_t\big) = \epsilon_t.\end{aligned}$$

Applying the defining total-variation bound to the miscoverage event shows that replacing the auxiliary target law by $\mathbf{P}_t$ increases its probability by at most ϵ_t. Combining this with equation 9 yields

$$\mathbb{P}\big(r(X') \notin \mathcal{I}_t(X')\big) \leq \alpha + \epsilon_t,$$

or equivalently,

$$\mathbb{P}\big(r(X') \in \mathcal{I}_t(X')\big) \geq 1 - \alpha - \epsilon_t.$$

This is the conditional guarantee stated in the theorem.

When the stated conditional assumptions hold almost surely, averaging over all quantities conditioned on above gives the unconditional bound

$$\mathbb{P}\big(r(X') \in \mathcal{I}_t(X')\big) \geq 1 - \alpha - \mathbb{E}[\epsilon_t].$$

Finally, if $w_t \in \mathcal{G}_t$, the weighted argument can be applied directly with w_t in place of $\widetilde{w}_t$. Indeed, w_t is nonnegative and $\mathbb{E}_{\mathbf{P}_0^X}[w_t(X)] = 1$, while dual feasibility gives $\sum_{i=1}^N w_t(X_i)\eta_i = 0$. Changing measure then leads directly to the actual target law $\mathbf{P}_t$, so no total-variation comparison is needed. The resulting coverage is at least $1 - \alpha$. In particular, this applies when $\widehat{w}_t = w_t$, in which case $\epsilon_t = 0$. □

For completeness, the false-positive implication follows directly from the output rule. Let A_t be the acceptance event defined in Section 4.1. On A_t, every reward value in $\mathcal{I}_t(X')$ satisfies the task-specific constraints. If $r(X')$ also belongs to this interval, then $X' \in \mathcal{F}_r$. Hence

$$A_t \cap \{X' \notin \mathcal{F}_r\} \subseteq \{r(X') \notin \mathcal{I}_t(X')\}.$$

Under the assumptions of Theorem 1, this gives $\mathbb{P}(A_t,\ X' \notin \mathcal{F}_r) \leq \alpha$. Under the conditioning and assumptions of Theorem 2, it gives

$$\mathbb{P}(A_t,\ X' \notin \mathcal{F}_r) \leq \alpha + \epsilon_t.$$

These bounds concern the joint event that a fresh proposal is accepted and is a false positive. They do not assert the same bounds conditional on acceptance or simultaneously over the full search trajectory.

Table 3: Notation used for proxy-reward-driven material evolution and conditional conformal inference.

Notation	Description
Material search and rewards	
$\mathcal{X}$; X, x; X'	Space of material candidates; a random candidate and its realization; a fresh candidate at the current iteration.
$\hat{r}(x)$; $r(x)$	Computationally inexpensive proxy reward and expensive true reward of candidate x.
$\mathcal{H}_t$	Archive after iteration t, containing candidates and their stored rewards. Algorithm 2 stores proxy rewards, while Algorithm 1 stores rewards evaluated conservatively over prediction intervals.
$\pi_t(\cdot \mid \mathcal{H}_t)$; $\mathcal{S}_t$	Parent-selection policy applied to $\mathcal{H}_t$ and the resulting set of K parents used to guide the following iteration.
$\mathcal{C}_t$; $\mathcal{O}_t$	Candidate batch proposed at iteration t and cumulative output set through that iteration.
$\mathcal{F}_{\hat{r}}$; $\mathcal{F}_r$	Sets of candidates satisfying the task-specific constraints according to the proxy reward and the true reward, respectively.
M; K; T	Number of fresh candidates evaluated per iteration, number of selected parents, and number of iterations.
Conditional conformal inference	
$\mathcal{D}$; n; n_{cal}	Labeled source dataset $\mathcal{D} = \{(X_i, r(X_i))\}_{i=1}^{n}$, its size, and the number of calibration pairs.
$\mathcal{D}_{\mathrm{fit}}$; $\mathcal{D}_{\mathrm{cal}}$; $\mathcal{D}_{\mathrm{src}}$	Mutually disjoint subsets used to fit the base quantile model, calibrate prediction intervals, and provide the source reference for density-ratio estimation.
$\mathcal{I}_t(x)$	Prediction interval for $r(x)$ constructed at iteration t. A candidate is returned only when its interval is nonempty, bounded, and satisfies the task-specific constraints throughout.
α	Nominal marginal miscoverage level for a fresh candidate, evaluated separately at each iteration.
$S(x, u)$; S_i	Fixed nonconformity score for a hypothetical reward u and calibration score $S_i = S(X_i, r(X_i))$.
$\mathcal{G}$; $\mathcal{G}_t$	Base finite-dimensional calibration space containing the constant function and its iteration-specific extension $\mathcal{G}_t = \mathcal{G} + \mathrm{span}\{\widehat{w}_t\}$.
s; $\widehat{q}_s$	Imputed test score and fitted calibration function obtained from augmented quantile regression with that score.
$\ell_\alpha(a, b)$	Pinball loss for the $(1-\alpha)$-quantile, $\ell_\alpha(a,b) = (1-\alpha)(b-a)_+ + \alpha(a-b)_+$.
$\eta_s(X')$	Dual variable associated with the fresh candidate in the augmented calibration problem, taking values in $[-\alpha, 1-\alpha]$.
$\widehat{\tau}_t(x)$	Candidate-specific score cutoff obtained from the augmented dual boundary and inverted to construct $\mathcal{I}_t(x)$.
Evolutionary covariate-shift correction	
$\mathbf{P}_0$; $\mathbf{P}_t$	Joint distributions of a candidate and its true reward in the source data and under the fixed generation configuration at iteration t, respectively.
$\mathbf{P}_0^X$; $\mathbf{P}_t^X$	Marginal distributions over material candidates under $\mathbf{P}_0$ and $\mathbf{P}_t$.
$\phi(x)$	Fixed material representation used in the calibration basis and density-ratio estimation.
$\mathcal{B}_t$	Unlabeled probe batch generated under the current fixed configuration and used only for density-ratio estimation.
$w_t(x)$; $\widehat{w}_t(x)$	Oracle density ratio $w_t(x) = \frac{d\mathbf{P}_t^X}{d\mathbf{P}_0^X}(x)$ and its online estimate from $\mathcal{D}_{\mathrm{src}}$ and $\mathcal{B}_t$.
$\mathcal{W}$; W	Density-ratio model class with values in $[0, W]$ and the bounded-overlap constant satisfying $w_t(x) \le W$.
ϵ_t	Total variation distance between $\mathbf{P}_t^X$ and the candidate distribution induced by the normalized estimated ratio, as defined in equation 6.