DETS: An Interval-Censored Evidential Sampling Framework for Cross-Domain Scientific Discovery

Jiahang Shen
School of Software Engineering, USTC
Suzhou, China
shenjiahang@mail.ustc.edu.cn

Hongxin Xiang
College of Computer Science and Technology, Hunan University
Changsha, China
xianghx@hnu.edu.cn

Wenjie Du*
School of Software Engineering, USTC
Suzhou, China
duwenjie@mail.ustc.edu.cn

Abstract

Scientific measurements inherently contain variance. Standard regression, however, treats noisy data as absolute ground truth, leading models to memorize noise and weaken robustness. This overfitting hampers generalization, particularly when transferring from abundant theoretical proxies to scarce experiments. We introduce DETS, a framework replacing rigid point estimation with Physical Tolerance Modeling. Our Interval-Censored Evidential Engine (ICEE) maximizes probability mass within acceptable error margins, alleviate aleatoric noise from epistemic uncertainty. Using this filtered uncertainty signal, a Thermodynamic Sampling strategy dynamically selects theoretical data to align source and target domains. Experiments across thermodynamics, drug affinity, solvation free energy, and bandgap prediction show that DETS consistently outperforms other methods, achieving up to 94.87% Acc@10 and improving full-data training by 14.92% on solvation free energy. Our code can be found at https://github.com/jiahangshen/DETS.

1 Introduction

Artificial intelligence is rapidly changing the landscape of molecular and material discovery by closely integrating machine learning with computational and experimental chemistry [Goh et al., 2017, Yadav et al., 2025, Rial, 2024]. This paradigm shift is particularly evident in molecular property prediction [Walters and Barzilay, 2020, Cai et al., 2024], catalytic simulations [Eyke et al., 2021, Xu et al., 2023], and the design of crystalline materials such as metal-organic frameworks [Chong et al., 2020, Rosen et al., 2021]. However, a pervasive phenomenon in AI4Chemistry is that the data available at scale deviates significantly from the target experimental distribution [McCready et al., 2006, Du et al., 2025, Ananikov, 2024]. High-precision experimental measurements are scarce and costly. The crucial point is that the experimental measurement data inherently carry their own measurement variance. As illustrated in Figure 1, when the same experiment is conducted under different conditions or by different individuals, the resulting measurement results will inevitably fluctuate. This variance is acceptable in

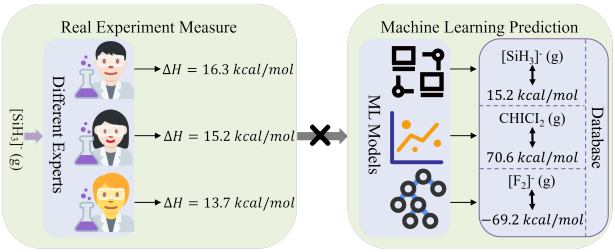


Figure 1: Comparison between variable experimental measurements and deterministic machine learning predictions. Real-world data often contains noise and expert variance, which traditional point-wise deterministic regression fail to capture.

*Corresponding author.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

science, but it is often misinterpreted by current regression indicators as a learnable signal, thereby leading to overfitting. On the other hand, low-precision theoretical data generated from approximate quantum chemistry or simplified reaction conditions are abundant but often contain considerable noise and systematic biases [Ruddigkeit et al., 2012, Visini et al., 2017, Goh et al., 2017, Guzman-Pando et al., 2024]. Overfitting to experimental fluctuations while struggling with theoretical distribution shifts creates a critical barrier to reliable scientific discovery [Kim et al., 2024, Vinod and Zaspel, 2025, Gardner et al., 2025].

To better extract knowledge from theoretical proxies and mitigate data scarcity, strategies such as transfer learning [Goh et al., 2018, Li and Jiang, 2021b] and multi-fidelity learning [Ramakrishnan et al., 2015, Fare et al., 2022] have been widely adopted. Under these frameworks, sampling-based pre-training on theoretical data has become a key paradigm [Na, 2025, Ji et al., 2024]. This strategy aims to initialize the model by screening theoretical samples that are consistent with the potential structure of the target domain, thereby alleviating the distribution shift between theoretical simulations and real experiments while utilizing massive data. Simultaneously, to account for experimental variance, Bayesian neural networks and uncertainty quantification methods have been proposed to estimate prediction confidence and manage data noise [Ryu et al., 2019, Scalia et al., 2020, Soleimany et al., 2021].

Several uncertainty-aware models estimate epistemic uncertainty [Ryu et al., 2019, Lakshminarayanan et al., 2016], and uncertainty-based data-selection methods use it to guide sampling [Coleman et al., 2020]. However, the reliability of this sampling signal depends on the training objective used by the uncertainty model [Kendall and Gal, 2017, Scalia et al., 2020, Soleimany et al., 2021]. Under pointwise supervision, every deviation from an experimental label contributes to the training loss, including deviations within the accepted measurement tolerance. Losses induced by these small residuals can unnecessarily inflate both epistemic and aleatoric uncertainty even when the model has captured the underlying physical relationship. When such contaminated epistemic uncertainty drives sampling, the sampler may repeatedly prioritize regions whose apparent uncertainty arises from label fluctuations, reducing attention to source samples that complement target-domain knowledge. DETS applies interval-censored learning to reduce the influence of residuals within the tolerance interval on uncertainty estimation. The resulting epistemic uncertainty provides a more stable signal for source sampling, thereby alleviating the distribution shift between theoretical and experimental domains [Kirschbaum and Bande, 2024, Kreiman and Krishnapriyan, 2026].

Thus, we introduce the Dual-Evidence Tolerance Sampling (DETS) framework, which integrates interval-likelihood estimation with dynamic source sampling guided by epistemic uncertainty. First, DETS employs the Interval-Censored Evidential Engine (ICEE) to reduce the sensitivity of the training loss and uncertainty estimates to experimental label fluctuations. Standard deep evidential regression optimizes a pointwise likelihood, causing residuals within a physically acceptable range to continue affecting the training loss and uncertainty estimates [Amini et al., 2020, Soleimany et al., 2021]. ICEE uses an Interval-Censored Likelihood (ICL) to maximize predictive probability mass within a physically admissible interval around each observation. This objective limits the influence of residuals within the tolerance interval on the training loss and uncertainty estimates, reducing unnecessary inflation of both epistemic and aleatoric uncertainty. Second, DETS combines target-domain prototypes with the epistemic uncertainty estimated by ICEE to update source-sampling weights. Higher epistemic uncertainty increases the effective sampling bandwidth, thereby assigning higher weights to a broader range of source samples. As epistemic uncertainty decreases, the weights become concentrated on source samples near the target prototypes.

Overall, our contributions can be summarized as follows:

- **Physical tolerance-based interval learning paradigm.** Addressing overfitting caused by measurement fluctuations, we reframe chemical property prediction from point regression to interval learning with physical tolerance. By reducing the penalty on residuals within the tolerance interval, this paradigm mitigates overfitting to pointwise experimental labels and yields more reliable estimates of epistemic uncertainty.
- **The DETS framework for uncertainty decoupling and dynamic sampling.** We propose the DETS framework, which combines an Interval-Censored Likelihood (ICL), tolerance-aware regularization, and dynamic source selection. The resulting epistemic uncertainty and target-domain prototypes jointly determine the source sampling weights.

- **Cross-task scientific evaluation.** Experiments demonstrate that DETS overcomes bottlenecks derived from fitting experimental fluctuations. By aligning data within a tolerance range, our method achieves state-of-the-art results in data-scarce, noisy scenarios, offering a robust solution to the theory-experiment distribution mismatch.

2 Methodology

2.1 Problem Formulation

Cross-domain scientific regression. We consider a regression problem with a scarce but high-fidelity target experimental dataset and a large auxiliary source dataset

$$\mathcal{T} = \{(x_j^{(t)}, y_j^{(t)})\}_{j=1}^{N_t}, \quad \mathcal{S} = \{(x_i^{(s)}, y_i^{(s)})\}_{i=1}^{N_s}, \quad N_s \gg N_t. \tag{1}$$

The target domain represents the experimental distribution. The statistical goal is to estimate the target conditional distribution $P_{\mathcal{T}}(Y|X)$. Target labels are noisy experimental measurements with fluctuations around the underlying values, so we assume that:

$$Y^{(t)} = f^\star(X^{(t)}) + \varepsilon_t, \qquad \mathbb{E}[\varepsilon_t | X^{(t)}] = 0, \tag{2}$$

where $f^\star(\cdot)$ denotes the true physical property function and ε_t denotes irreducible experimental measurement fluctuations. In contrast, source labels are generated from low-fidelity calculations or high-throughput assays, and may contain systematic response bias: $Y^{(s)} = f^\star(X^{(s)}) + \Delta(X^{(s)})$, where $\Delta(x)$ denotes the source-domain systematic bias relative to the corresponding true physical value. Equivalently, for the same input x, the source and target labels satisfy $Y^{(s)} = Y^{(t)} - \varepsilon_t + \Delta(x)$, which shows that the source-target label gap contains systematic source bias, while the target label itself contains experimental measurement fluctuations. Therefore, the problem involves both covariate shift and conditional distribution mismatch:

$$P_{\mathcal{S}}(X) \neq P_{\mathcal{T}}(X), \quad P_{\mathcal{S}}(Y|X) \neq P_{\mathcal{T}}(Y|X) \quad \text{when } \Delta(x) \neq 0. \tag{3}$$

Thus, using all source samples as equally valid supervision may introduce negative transfer.

Target expected risk. In scientific measurements, experimental fluctuation should not be treated as learnable. Since repeated measurements are usually unavailable, we center the observation interval at the reported value y and set its radius as $\epsilon_{\text{tol}}(y) = \max(\epsilon_{\text{abs}}, \rho|y|)$, where ϵ_{abs} encodes known or validation-selected tolerance and ρ is a relative tolerance. The corresponding valid interval is:

$$I(y) = [y - \epsilon_{\text{tol}}(y),\, y + \epsilon_{\text{tol}}(y)]. \tag{4}$$

A scientifically reliable predictor should maximize probability mass within this interval and treat the experimental label as an uncertain observation. Let $q_\theta(y|x)$ be a probabilistic predictor. The target-domain expected risk is defined as

$$\mathcal{R}_{\mathcal{T}}(\theta) = \mathbb{E}_{(X^{(t)}, Y^{(t)}) \sim P_{\mathcal{T}}} \left[-\log \int_{Y^{(t)} - \epsilon_{\text{tol}}(Y^{(t)})}^{Y^{(t)} + \epsilon_{\text{tol}}(Y^{(t)})} q_\theta(u|X^{(t)})\, du \right]. \tag{5}$$

This objective regards predictions within the physical tolerance interval as valid and avoids forcing the model to fit measurement fluctuations.

Dynamic source sampling. Since source labels may be systematically biased, the source dataset should be assigned sample-specific selection weights. Let $\epsilon_{\text{num}} > 0$ be a small numerical constant. We formulate the general learning objective as

$$\min_\theta \; \widehat{\mathcal{R}}_{\mathcal{T}}(\theta) + \lambda_s \frac{\sum_{i=1}^{N_s} w_i \ell_{\text{src}}(q_\theta(\cdot|x_i^{(s)}), y_i^{(s)})}{\sum_{i=1}^{N_s} w_i + \epsilon_{\text{num}}}, \qquad 0 \leq w_i \leq 1. \tag{6}$$

where λ_s controls the contribution of the source objective, $q_\theta(\cdot|x_i^{(s)})$ denotes the predictive distribution conditioned on the source input, and $\ell_{\text{src}}(q_\theta(\cdot|x_i^{(s)}), y_i^{(s)})$ is the corresponding source-domain training loss. The weight $w_i \in [0, 1]$ measures the utility of this source sample for the target task.

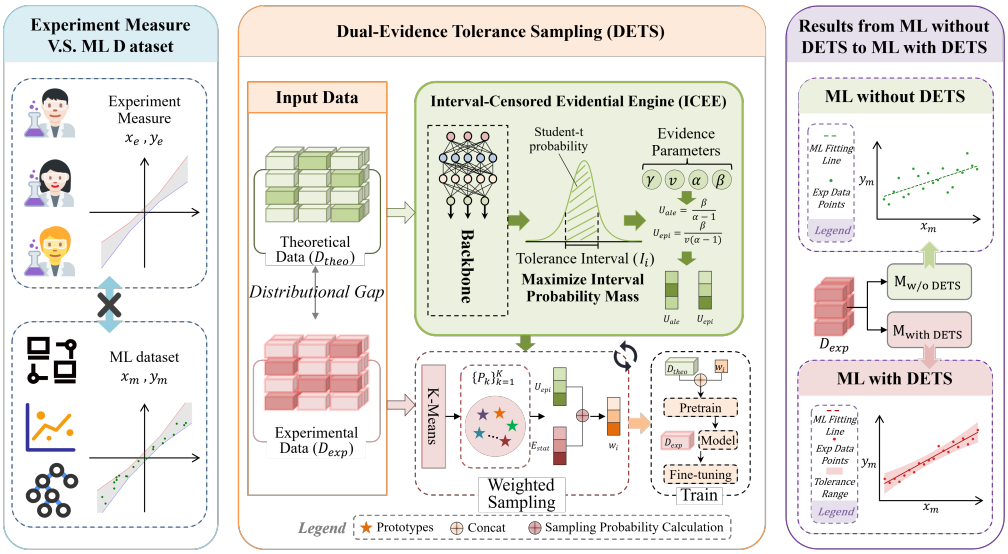


Figure 2: **Schematic of the DETS framework.** To address the distributional gap between noisy experimental measures (Left) and theoretical data (D_{theo}), DETS employs an Interval-Censored Evidential Engine (ICEE). This engine decouples uncertainties to drive a prototype-based Weighted Sampling strategy.

Meaning of overfitting in this work. In this paper, "overfitting" describes the learned response to irreducible experimental measurement fluctuations.

2.2 Interval-Censored Evidential Engine (ICEE)

2.2.1 Standard DER and its Limitations

Following established evidential regression frameworks [Soleimany et al., 2021], the evidential head outputs $\mathbf{m} = (\gamma, \nu, \alpha, \beta)$. We model the target property y using a Gaussian with latent statistics (μ, σ^2), placed under a Normal-Inverse-Gamma (NIG) conjugate prior:

$$p(\mu, \sigma^2|\mathbf{m}) = \text{NIG}(\mu, \sigma^2|\gamma, \nu, \alpha, \beta). \tag{7}$$

Marginalizing over latent μ and σ^2 yields a Student-t posterior predictive distribution with scale $\sigma_{st}^2 = \beta(1+\nu)/(\nu\alpha)$:

$$p(y|\mathbf{m}) = \text{St}\left(y; \gamma, \frac{\beta(1+\nu)}{\nu\alpha}, 2\alpha\right). \tag{8}$$

Standard DER minimizes the negative log-likelihood loss of this distribution:

$$\mathcal{L}(\mathbf{m}, y) = \frac{1}{2}\log\left(\frac{\pi}{\nu}\right) - \alpha\log\Omega + \frac{2\alpha+1}{2}\log\left((y-\gamma)^2\nu + \Omega\right) + \log\left(\frac{\Gamma(\alpha)}{\Gamma(\alpha+1/2)}\right), \tag{9}$$

where $\Omega = 2\beta(1+\nu)$. See the Appendix E.1 for the detailed formulation derivation.

Statistical Mismatch. The premise for minimizing the Eq. (9) is to assume that the true label is an exact Dirac delta function, $p(y_{true}) = \delta(y - y_i)$. However, in experimental chemistry, measurement fluctuations place valid observations within a range around the true value y^*. By forcing the model to maximize the point density through Eq. (9), two key drawbacks arise: overfitting to random noise, which causes the model to memorize random experimental residuals and obscures the underlying physical laws; and evidence collapse, in which the model unnecessarily increases epistemic uncertainty in response to measurement fluctuations. To eliminate these defects, the learning mode must shift from enforcing an exact point match to accepting an allowable error range. Therefore, the concept of Interval Censored Evidence Engine (ICEE) was proposed, aiming to transform the regression task from the goal of maximizing point density to maximizing interval probability quality, separating random noise from the learning.

2.2.2 From Point Estimation to Interval Likelihood

Interval observation model. Let y_i^{phys} denote the label in the original physical unit and let $y_i = \frac{y_i^{\text{phys}} - \mu_{\mathcal{T}}}{\sigma_{\mathcal{T}}}$ be the normalized label used for training, where $\mu_{\mathcal{T}}$ and $\sigma_{\mathcal{T}}$ are computed from the target experimental dataset. The physical tolerance radius is $\epsilon_{\text{tol}}(y_i^{\text{phys}}) = \max(\epsilon_{\text{abs}}, \rho|y_i^{\text{phys}}|)$.

After normalization, the corresponding tolerance radius $\delta_i = \frac{\epsilon_{\text{tol}}(y_i^{\text{phys}})}{\sigma_{\mathcal{T}}}$. Thus, the physically valid interval in the normalized training space is $I_i = [L_i, U_i] = [y_i - \delta_i,\ y_i + \delta_i]$, where L_i and U_i denote its lower and upper bounds, respectively.

Given the Student-t predictive distribution in Eq. 8, the probability assigned to the valid interval is

$$P(y^\star \in I_i | \mathbf{m}_i) = \int_{L_i}^{U_i} \text{St}(t; \gamma_i, \sigma_{st,i}^2, 2\alpha_i)\, dt = \Phi_{2\alpha_i}(Z_{upper}) - \Phi_{2\alpha_i}(Z_{lower}), \tag{10}$$

where $\Phi_{2\alpha_i}$ is the standard Student-t CDF and $Z_{upper} = \frac{U_i - \gamma_i}{\sigma_{st,i}}$, $Z_{lower} = \frac{L_i - \gamma_i}{\sigma_{st,i}}$.

The Interval-Censored Likelihood loss is therefore:

$$\mathcal{L}_{ICL} = -\log\left(\Phi_{2\alpha_i}(Z_{upper}) - \Phi_{2\alpha_i}(Z_{lower}) + \epsilon_{\text{num}}\right), \tag{11}$$

Geometrically, this objective maximizes the predictive probability mass inside the physically admissible interval. See the Appendix E.3 for the detailed formulation derivation.

The standard regularization term $\mathcal{L}_{REG} = |y - \gamma| \cdot (2\nu + \alpha)$ penalizes tolerable measurement fluctuations. Therefore, relative tolerance regularization was proposed. The tolerance regularization is defined as follows:

$$\mathcal{L}_{REG}^{Rel}(\mathbf{m}_i, y_i) = \underbrace{\text{Sp}\left(|y_i - \gamma_i| - \delta_i\right)}_{\text{Rectified Error}} \cdot \underbrace{(2\nu_i + \alpha_i)}_{\text{Total Evidence}}, \tag{12}$$

where $\text{Sp}(x) = \log(1 + \exp(x))$ smoothly approximates the positive part. When $|y_i - \gamma_i| \leq \delta_i$, the penalty is suppressed for tolerated deviations. Outside the interval, the penalty increases with both the excess residual and the total evidence, discouraging overconfident invalid predictions. The final objective is:

$$\mathcal{L}_{total}(\mathbf{m}, y) = \mathcal{L}_{ICL}(\mathbf{m}, y) + \lambda_t \cdot \mathcal{L}_{REG}^{Rel}(\mathbf{m}, y). \tag{13}$$

Here, λ_t uses an annealing strategy to stabilize convergence.

Uncertainty Decomposition. ICEE allows analytical decomposition of total variance:

$$\underbrace{\text{Var}[y]}_{\text{Uncertainty}} = \mathbb{E}[\sigma^2] + \text{Var}[\mu] = \underbrace{\frac{\beta}{\alpha - 1}}_{\text{Aleatoric}} + \underbrace{\frac{\beta}{\nu(\alpha - 1)}}_{\text{Epistemic}}. \tag{14}$$

The epistemic term $U_{epi} = \frac{\beta}{\nu(\alpha-1)}$, estimated under the tolerance-aware objective, serves as the uncertainty signal for sampling. See Appendix E.2 for the detailed derivation.

2.3 Dual-Evidence Tolerance Sampling (DETS)

Let g_θ denote the encoder before the evidential regression head. At training epoch e, target and source representations are $z_{j,e}^{(t)} = g_{\theta_e}(x_j^{(t)})$, $z_{i,e}^{(s)} = g_{\theta_e}(x_i^{(s)})$. To avoid manually fixing the number of target prototypes, we select K by maximizing the silhouette score on target representations. K-Means is then applied in the latent representation space:

$$\min_{\mathcal{P}_e} \sum_{j=1}^{N_t} \sum_{k=1}^{K^\star} \mathbb{I}(c_j = k) \|z_{j,e}^{(t)} - p_{k,e}\|_2^2. \tag{15}$$

where $p_{k,e}$ denotes the k-th target prototype at epoch e, and c_j is the cluster assignment of the j-th target sample. The resulting prototypes $\mathcal{P}_e = \{p_{k,e}\}_{k=1}^{K^\star}$ serve as high-fidelity anchors for the experimental target manifold. In practice, prototypes are refreshed once every R_{proto} epochs,

reducing the overhead associated with per-update refreshes. For each theoretical sample $x_i^{(s)}$ with potential embedding $z_i^{(s)}$, the similarity to the target domain is evaluated. For each source sample, we compute its structural evidence as the maximum soft Tanimoto similarity to the target prototypes:

$$E_{stat}^{(i)} = \max_{p_k \in \mathcal{P}} \frac{z_i^{(s)} \cdot p_k}{\|z_i^{(s)}\|^2 + \|p_k\|^2 - z_i^{(s)} \cdot p_k}, \tag{16}$$

where $\cdot$ denotes the dot product and $\|\cdot\|$ is the L_2 norm. The value of $E_{stat}^{(i)}$ lies within the range [0, 1]. A larger value indicates that the source samples are closer to the target-domain distribution.

We normalize epistemic uncertainty within each epoch's theoretical samples $\mathcal{B}_s$ for stability:

$$\bar{U}_{epi}(x_i) = \frac{U_{epi}(x_i) - \min_{x \in \mathcal{B}_s} U_{epi}(x)}{\max_{x \in \mathcal{B}_s} U_{epi}(x) - \min_{x \in \mathcal{B}_s} U_{epi}(x) + \epsilon_{\text{num}}}. \tag{17}$$

The dynamic bandwidth σ_{eff} is defined to adjust the strictness of sampling:

$$\sigma_{eff}^{(i)} = \sigma_0 \cdot (1 + \eta \cdot \bar{U}_{epi}(x_i)). \tag{18}$$

Here, σ_0 is the base bandwidth and η is a sensitivity hyperparameter. We conceptualize selection as a thermodynamic system. The weight w_i is computed via a Boltzmann-like distribution:

$$w_i = \alpha_0 \cdot \exp\left(-\frac{(1 - E_{stat}^{(i)})^2}{2(\sigma_{eff}^{(i)})^2}\right). \tag{19}$$

Thus, the epistemic uncertainty from Eq. 14 controls the bandwidth in Eq. 18, which in turn determines the source-sample weight in Eq. 19. Here, $E_{stat}^{(i)}$ is the normalized structural similarity to the nearest prototype, with larger values indicating greater similarity. The term $(1 - E_{stat}^{(i)})^2$ represents the system energy associated with structural mismatch, while $\bar{U}_{epi}$ modulates the temperature through σ_{eff}.

When $\bar{U}_{epi}$ is low, the sampling bandwidth is narrow, and high weights are concentrated on source samples close to the target prototypes. As $\bar{U}_{epi}$ increases, the bandwidth becomes wider, allowing a wider range of source structures to receive higher weights. In this way, epistemic uncertainty controls how broadly the source data are sampled.

2.4 Training Protocol and Complexity

Before constructing target prototypes and enabling dynamic source sampling, DETS performs a warm-up stage.

Phase 1: Dynamic Source Pre-training via Coreset Sampling. We iteratively train on $\mathcal{S}$. In each iteration t, we calculate weights w_i and construct a dynamic coreset $\mathcal{C}_t$ by stochastic sampling:

$$P(x_i \in \mathcal{C}_t) \propto w_i(x_i), \quad \mathcal{C}_t \subseteq \mathcal{B}_s. \tag{20}$$

Parameters are updated solely on this subset:

$$\theta_{t+1} \leftarrow \theta_t - \eta_{lr} \nabla_\theta \frac{1}{|\mathcal{C}_t|} \sum_{(x_j, y_j) \in \mathcal{C}_t} \mathcal{L}_{total}(x_j, y_j). \tag{21}$$

Phase 2: Target Fine-tuning. We employ a two-stage strategy on $\mathcal{T}$: (1) Head Alignment (freeze backbone, calibrate uncertainty); (2) Full Adaptation (unfreeze all). Note that we continue to use $\mathcal{L}_{total}$ (with $\mathcal{L}_{ICL}$) to maintain noise tolerance on the experimental set.

Complexity Analysis. DETS adds $\mathcal{O}(N_t K D)$ prototype refresh every R_{proto} epochs and $\mathcal{O}(|\mathcal{B}_s| K D)$ source-batch weighting. Epistemic uncertainty is obtained in the same forward pass, keeping the overhead lightweight.

3 Experiments

In this section, we conduct experiments to answer four research questions: **(RQ1)** Does DETS improve robustness under noisy measurements? **(RQ2)** Can thermodynamic sampling adapt the source distribution during training? **(RQ3)** Do the ICEE components effectively decouple aleatoric and epistemic uncertainty? **(RQ4)** How sensitive is DETS to the tolerance threshold ρ?

Table 1: Comprehensive evaluation on molecular enthalpy and drug-target binding affinity prediction. We report **MAE**, **RMSE**, **MAPE** (%), and **Acc@10%** (↑) for each method. The values represent the best performance achieved by each baseline across different sampling rates.

Dataset	**Wudily_Cho → ATcT**				**Davis → CASF-2016**			
Metric	**MAE** (↓)	**RMSE** (↓)	**MAPE** (%, ↓)	**Acc@10** (%, ↑)	**MAE** (↓)	**RMSE** (↓)	**MAPE** (%, ↓)	**Acc@10** (%, ↑)
Static Coreset Selection								
Hard Random	26.14±1.72	35.72±2.15	83.00±0.6	19.44±0.3	1.3980±0.24	1.7368±0.28	22.75±1.1	31.67±0.4
Herding	26.65±1.53	37.18±1.90	94.20±0.5	16.67±0.2	1.3748±0.19	1.7367±0.22	24.17±0.9	30.00±0.3
K-Means	25.26±1.49	35.05±1.85	88.59±0.6	16.67±0.3	1.3076±0.16	1.6996±0.20	25.43±0.8	31.67±0.2
Entropy	24.56±1.31	32.49±1.62	74.73±0.5	13.89±0.2	1.5318±0.16	1.9160±0.19	24.79±0.7	20.00±0.3
Least Confidence	23.68±1.64	32.37±2.01	90.19±0.7	16.67±0.4	1.3418±0.15	1.8556±0.18	23.68±0.8	36.67±0.3
GraNd-20	27.65±1.42	36.71±1.75	106.29±0.5	13.89±0.2	1.3627±0.15	1.7465±0.19	22.95±0.7	26.67±0.2
Glister	26.22±1.58	35.77±1.92	90.64±0.6	19.44±0.3	1.3854±0.16	1.6920±0.21	23.87±0.8	26.67±0.3
Influence	26.72±1.45	38.06±1.80	80.20±0.5	16.67±0.2	1.3761±0.13	1.7883±0.17	23.28±0.6	25.00±0.2
EL2N-2	25.75±1.37	35.27±1.68	102.70±0.5	8.33±0.2	1.3627±0.19	1.7465±0.24	22.34±0.9	28.33±0.3
DP	26.04±1.40	35.67±1.72	97.19±0.5	16.67±0.2	1.3967±0.14	1.7992±0.18	22.55±0.7	21.67±0.2
Dynamic Coreset Selection								
Soft Random	26.38±1.43	35.93±1.78	85.57±0.5	8.33±0.3	1.3982±0.13	1.7040±0.16	23.93±0.6	26.67±0.2
ϵ-greedy	26.36±1.43	37.92±1.75	87.44±0.5	16.67±0.3	1.4275±0.16	1.8522±0.20	23.66±0.8	28.33±0.3
UCB	27.39±1.50	36.38±1.84	95.14±0.6	11.11±0.3	1.4487±0.11	1.8480±0.14	24.27±0.5	21.67±0.2
InfoBatch	25.56±1.39	36.03±1.70	89.90±0.5	13.89±0.2	1.4735±0.14	1.8290±0.18	25.00±0.7	20.00±0.2
Cross-domain Coreset Selection								
MolPeg	26.28±1.49	37.74±1.85	79.14±0.5	22.22±0.3	1.5156±0.17	1.8754±0.21	24.83±0.8	23.33±0.3
CSIE	25.14±1.51	34.79±1.90	83.30±0.6	19.44±0.3	1.5955±0.15	1.9304±0.19	26.43±0.7	18.33±0.3
DETS (Ours)	**23.11**±1.20	**25.71**±1.55	**73.07**±0.4	**25.00**±0.2	**1.2734**±0.13	**1.6419**±0.16	**21.90**±0.6	**38.33**±0.2
Full dataset	34.59±1.24	44.11±1.58	106.54±0.4	5.56±0.2	1.4441±0.16	1.8136±0.20	27.12±0.7	30.00±0.3

3.1 Experiments Setup

Dataset. We conducted experiments on four representative attribute prediction tasks: molecular formation enthalpy, drug-target binding affinity (DTA), solute-solvent free energy, and material band gap prediction. Each task required transferring knowledge from large-scale theoretical or low-fidelity source domains (*Wudily_Cho* [Bruce et al., 2025], *Davis* [Davis et al., 2011], *CombiSolv-QM* [Vermeire and Green, 2020], and *NREL-PBE* [Kim et al., 2020]) to scarce, high-fidelity experimental target domains (*ATcT* [Ruscic and Bross, 2025], *CASF-2016* [Su et al., 2018], *CombiSolv-Exp* [Vermeire and Green, 2020], and *NREL-GWVD* [Kim et al., 2020]). The Davis → CASF-2016 task evaluates DTA transfer across data sources and measurement protocols. Details of datasets and preprocessing procedures can be found in Appendix B.

Baselines. The comparison methods include: (1) Static strategies (Hard Random, Herding [Welling, 2009], K-Means [Sorscher et al., 2022], Entropy [Coleman et al., 2020], Least Confidence [Coleman et al., 2020], GraNd-20 [Paul et al., 2021], Glister [Killamsetty et al., 2021], Influence [Koh and Liang, 2017], EL2N-2 [Paul et al., 2021], DP [Yang et al., 2023]), which fix the coreset before training initialization; (2) Dynamic strategies (Soft Random, ϵ-greedy [Raju et al., 2021], UCB [Raju et al., 2021], InfoBatch [Qin et al., 2024]), which optimize samples in real-time during the training process; (3) Cross-domain strategies (MolPeg [Chen et al., 2024], CSIE [Wang et al., 2026]), which are algorithms designed for domain shift scenarios. (4) Uncertainty baselines (MC-Dropout [Gal and Ghahramani, 2015] and Deep Ensemble [Lakshminarayanan et al., 2016]), which estimate uncertainty through stochastic inference and model ensembles, respectively. We report the predictive performance and computational overhead of these uncertainty-estimation baselines in Appendix D.4. Static and dynamic baselines use source-domain information during source selection. Cross-domain methods may use unlabeled target features, while no method uses per-sample target labels in its sampling weights. All methods follow the same target fine-tuning protocol.

Evaluation Metrics. We use MAE and RMSE to evaluate the absolute accuracy, and combine MAPE and tolerance accuracy rate (Acc@10%) to deal with the measurement noise of the experimental data. Specifically, Acc@10% quantifies the proportion of predictions within the 10% relative error range, directly reflecting the scientific credibility under the noise constraint. All results are reported as the mean and standard deviation of three independent experiments. To avoid the instability of the numerical values caused by extremely small values leading to an inflated MAPE, we applied a threshold mechanism. (See Appendix C.1).

Implementation Details. All the experiments were implemented using PyTorch on NVIDIA A100 GPU. DETS utilized a specific architecture tailored for the task: a 3-layer GIN for molecular properties (enthalpy, solvation), a hybrid GIN-Transformer for DTA, and a multi-layer perceptron for bandgap prediction. We adopted a two-stage Adam optimization strategy with cosine annealing.

Table 2: Comprehensive evaluation on material bandgap and molecular dissolution prediction. We report **MAE**, **RMSE**, **MAPE** (%), and **Acc@10%** (↑). Values indicate the best performance across sampling rates.

Dataset	**NREL-MatDB PBE → GWVD**				**CombiSolv QM → Exp**			
Metric	**MAE** (↓)	**RMSE** (↓)	**MAPE** (%, ↓)	**Acc@10** (%, ↑)	**MAE** (↓)	**RMSE** (↓)	**MAPE** (%, ↓)	**Acc@10** (%, ↑)
Static Coreset Selection								
Hard Random	0.6007±0.09	0.7685±0.12	39.01±1.2	26.70±0.03	0.3008±0.03	0.5236±0.03	8.46±0.6	88.95±0.02
Herding	0.6031±0.04	0.7465±0.06	39.10±0.9	25.56±0.02	0.3204±0.02	0.5143±0.03	8.96±0.5	87.24±0.03
K-Means	0.5864±0.02	0.7655±0.04	39.01±0.8	26.70±0.02	0.2699±0.02	0.4476±0.02	7.66±0.4	91.23±0.02
Entropy	0.6187±0.03	0.8011±0.05	52.80±0.8	28.41±0.02	0.2874±0.03	0.4662±0.04	8.37±0.7	90.21±0.03
Least Confidence	0.5886±0.07	0.7488±0.09	37.95±1.1	28.40±0.04	0.2974±0.02	0.5199±0.03	8.09±0.6	91.91±0.02
GraNd-20	0.6049±0.04	0.7783±0.06	43.11±0.9	29.55±0.02	0.2881±0.02	0.4840±0.02	8.12±0.5	90.43±0.02
Glister	0.5986±0.03	0.7427±0.05	36.83±0.7	26.14±0.02	0.2626±0.03	0.4777±0.01	7.19±0.6	92.37±0.02
Influence	0.5734±0.06	0.7655±0.08	42.71±1.0	30.68±0.03	0.2770±0.03	0.4515±0.05	7.32±0.8	91.80±0.03
EL2N-2	0.5859±0.06	0.7643±0.09	41.82±1.0	30.11±0.03	0.2724±0.03	0.4492±0.05	7.53±0.7	91.34±0.02
DP	0.5856±0.06	0.7447±0.08	38.87±0.9	29.54±0.03	0.2881±0.02	0.4515±0.02	8.73±0.5	88.84±0.02
Dynamic Coreset Selection								
Soft Random	0.6081±0.03	0.7421±0.05	40.41±0.7	26.70±0.02	0.2730±0.02	0.4400±0.02	7.78±0.4	91.80±0.02
ϵ-greedy	0.5711±0.06	0.7244±0.08	39.09±0.9	27.84±0.03	0.2654±0.03	0.4299±0.04	7.29±0.6	92.48±0.02
UCB	0.5842±0.05	0.7389±0.07	38.85±0.8	30.11±0.03	0.2513±0.04	0.4596±0.05	6.75±0.7	93.28±0.03
InfoBatch	0.5724±0.09	0.7278±0.12	36.76±1.3	31.25±0.04	0.2654±0.02	0.4615±0.01	7.10±0.4	93.17±0.02
Cross-domain Coreset Selection								
MolPeg	0.5812±0.06	0.7403±0.08	42.27±0.9	26.70±0.03	0.2654±0.03	0.4299±0.03	7.29±0.5	92.48±0.02
CSIE	0.5917±0.06	0.7614±0.08	38.67±0.9	26.14±0.03	0.2779±0.02	0.4970±0.03	7.07±0.6	92.14±0.02
DETS (Ours)	**0.5413**±0.02	**0.6885**±0.03	**26.01**±0.6	**32.39**±0.02	**0.2258**±0.02	**0.4155**±0.03	**5.48**±0.5	**94.87**±0.02
Full dataset	0.7207±0.04	0.9566±0.06	47.96±0.8	22.72±0.02	0.4253±0.03	0.6960±0.04	12.70±0.7	79.95±0.03

Source pre-training on theoretical data used $lr = 10^{-3}$, followed by target fine-tuning on experimental data with $lr = 10^{-5}$.

Hyperparameters. The default settings include tolerance threshold $\rho = 0.05$, regularization coefficient $\lambda = 0.01$, sampling bandwidth $\sigma_0 = 0.5$, and confidence threshold $\tau_{\text{high}} = 0.9$. When a reliable experimental resolution or accepted measurement error is available, it determines the absolute tolerance. Otherwise, we set the absolute tolerance to the 10th percentile of residuals from a source-trained MSE model on a fixed, labeled target validation set selected before test evaluation. This data-driven rule requires labeled target validation data. Detailed configurations and four-task results are provided in Appendices C.6 and C.2.

3.2 Superiority of DETS across Experiment Domains

To answer **RQ1**, we conducted a comprehensive comparison of DETS with 10 widely used static pruning methods, 4 dynamic pruning methods, and 2 cross-domain methods, as shown in Tables 1 and 2, and more detailed explanations can be found in Appendix D. We followed CSIE experimental setup and set a fixed dataset retention ratio [Wang et al., 2026]. Specifically, we set the dataset retention ratio (1 - s)% to {20%, 40%, 60%, 80%}. For each method, we selected the best experimental results from these four retention ratios. Our observation results (**Obs.**) are summarized as follows:

Obs.①. DETS demonstrates superior noise robustness. Tables 1 and 2 summarize the performance across four distinct scientific domains. DETS consistently surpasses state-of-the-art baselines, including the domain-adaptation methods MolPeg and CSIE. Most notably, our method achieves substantial gains in the Acc@10% metric. For instance, in the molecular enthalpy task, DETS improves tolerance accuracy by a significant margin compared to the best baseline. This result directly addresses the limitation of traditional models, which tend to overfit measurement noise. The Interval-Censored Evidential Engine (ICEE) maximizes probability mass within a valid tolerance window, reducing the influence of individual noisy points. As a result, the model can capture the underlying physical relationships and provide more reliable epistemic uncertainty estimates under experimental fluctuations. We also report the relative improvement over the second-best baseline in each scenario in Appendix D.5. The gains are consistent across all four domains, with improvements of 2.47%–11.29% in MAE and 1.59–8.33 percentage points in Acc@10%, demonstrating benefits across all datasets.

3.3 Dynamic Evolution of Coreset Sampling

To answer **RQ2**, we conducted a trajectory analysis of the core set selection at different training stages. Specifically, we visualized the dynamic evolution of sample weights in the latent feature

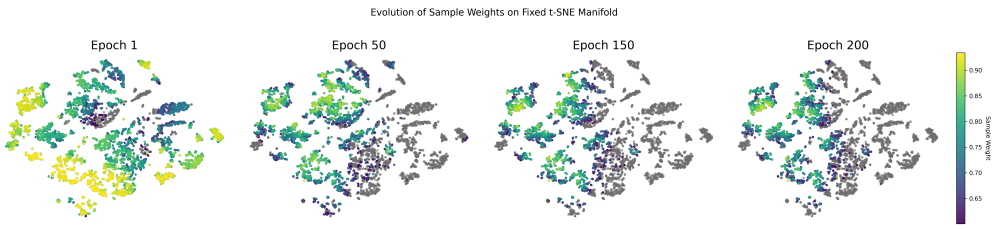


Figure 3: **Visualization of the dynamic coreset evolution in t-SNE space. (Left)** In the early phase, high-weight samples (yellow) are widely dispersed, reflecting broad knowledge exploration. **(Right)** As training converges, irrelevant source samples (gray) are effectively filtered.

space to study how DETS guides the model to shift from global exploration to local refinement. The visualization results and corresponding analysis are as follows:

Obs.②. Adaptability of the Sampling Strategy. To verify the dynamic characteristics of thermodynamic sampling, we visualized the selected coreset during the training process (Figure 3) in a fixed two-dimensional t-SNE space. These projections revealed a clear change corresponding to our thermodynamic assumptions. In the early stages (ranging from the 1st to the 50th epoch), high-weight samples (yellow) were dispersed throughout the feature space. This indicates a state driven by high cognitive uncertainty, where the sampling bandwidth remained relatively large. Therefore, the model actively explored various chemical structures in the source domain to build knowledge boundaries. This mechanism effectively avoids the common premature discarding of potential knowledge in hard pruning methods. As the model converged (ranging from the 150th to the 200th epoch), the uncertainty decreased, and the sampling bandwidth narrowed. The distribution of active samples significantly contracted. High weights concentrated in specific clusters closely corresponding to the experimental target anchor points, while irrelevant source samples (gray) were filtered out. This evolution confirmed that DETS can autonomously balance exploration and exploitation, and can transition from a broad knowledge acquisition stage to a targeted refinement process.

3.4 Ablation & Hyperparameter Analysis

To answer **RQ3** and **RQ4**, we conducted a series of detailed analyses. First, we quantitatively evaluated the contribution of each core component to the overall performance through ablation experiments, and the results are summarized in Table 3(a). Subsequently, we examined the model's sensitivity to key hyperparameters, the physical tolerance threshold, and the results are shown in Table 3(b). Our observations are as follows:

Table 3: Ablation and tolerance-threshold sensitivity analyses.

(a) Ablation study

Method	Enthalpy		DTA		Bandgap		Solvation	
	MAE↓	Acc↑	MAE↓	Acc↑	MAE↓	Acc↑	MAE↓	Acc↑
w/o ICEE	26.85	10.53	1.49	25.00	0.682	22.15	0.285	89.40
w/o Thermo.	26.14	15.20	1.47	30.12	0.665	26.80	0.293	90.10
w/o Rel. Reg.	25.90	22.90	1.34	35.50	0.651	30.50	0.270	92.20
DETS	**23.11**	**25.00**	**1.27**	**38.33**	**0.541**	**32.39**	**0.226**	**94.87**

(b) Tolerance threshold sensitivity

Metric	1%	3%	5%	10%	15%	20%
MAE↓	0.381	0.295	**0.226**	0.368	0.392	0.412
RMSE↓	0.585	0.492	**0.416**	0.479	0.505	0.525
MAPE↓	7.12	6.25	**5.48**	6.85	8.10	9.20
Acc@10↑	87.50	91.20	**94.87**	89.90	85.50	82.15

Obs.③. Uncertainty Decoupling via ICEE Components. We analyze the contribution of key components in Table 3(a). Replacing the interval-censored objective with standard NLL (**w/o ICEE**) results in the most significant performance drop, particularly in the Acc@10% metric. For enthalpy, accuracy plummets from 25.00% to 10.53%. This confirms that standard regression forces the model to memorize data noise. Without ICEE, the epistemic uncertainty becomes contaminated by aleatoric noise, misguiding the sampling process. Replacing dynamic weighting with uniform sampling (**w/o Thermo.**) also degrades performance across all tasks. This indicates that static strategies fail to handle the distribution shift. The thermodynamic mechanism is essential for transitioning the model from broad theoretical exploration to focused experimental exploitation. Removing Relative Regularization (**w/o Rel. Reg.**) leads to minor deterioration, verifying its role in preventing overconfidence in out-of-distribution regions. The target-only comparison in Appendix D.6 separates the contribution of ICEE from source sampling, and Appendix C.7 reports sensitivity to λ and σ_0 across all four tasks.

Obs.④. Impact of Tolerance Threshold on Model Fitting. The tolerance threshold ρ defines the width of the physical validity interval. Table 3(b) illustrates the model's sensitivity to ρ on the

molecular dissolution prediction task. Performance peaks when ρ is set to 5%, which aligns with the inherent experimental error margin. Setting ρ too low (1%) degrades DETS to standard point regression, causing overfitting to measurement noise. Conversely, an excessively large threshold (20%) relaxes constraints too much, leading to underfitting and loss of precision. This validates explicit physical-tolerance modeling, which distinguishes admissible fluctuations from substantive errors.

4 Limitations

DETS relies on a meaningful tolerance threshold. The data-driven rule requires a small labeled target validation set when measurement resolution is unknown. The current evidence covers cross-domain regression tasks in chemistry and materials science; classification and structured generation require different observation models and objectives.

5 Broader Impact and Future Work

DETS is designed to improve scientific regression under experimental fluctuations and bias in low-fidelity source data. More reliable source selection may reduce data and experimental costs in chemistry, drug discovery, and materials science. Predictions used in safety-critical settings still require expert assessment and experimental verification.

The same principle can be applied to sampling problems in which several outputs are valid or labels vary within an acceptable range. Examples include personalized and pluralistic alignment, where preferences differ across users and more than one response can be acceptable [Li et al., 2026]; reasoning-model post-training, where multiple candidate responses can supply relative preference signals and different solution traces may reach the same answer [Guo et al., 2025, Kou et al., 2026]; test-time scaling, where computation is allocated across sampled reasoning paths [Snell et al., 2025]; and web-agent training, where different action trajectories can complete the same task [Wei et al., 2025]. In these settings, a task-specific acceptable set can represent valid responses, verified reasoning traces, or successful action trajectories. After accounting for variation within this set, epistemic uncertainty can prioritize prompts and tasks where the model has limited knowledge and allocate sampling, annotation, or interaction budgets. Future work will study these extensions and define acceptable sets for different output structures.

6 Conclusion

This paper presents DETS, an interval-censored evidential sampling framework for noisy cross-domain scientific regression. ICEE maximizes predictive probability mass within a physical tolerance interval and estimates tolerance-aware epistemic uncertainty. DETS combines this uncertainty with target-domain prototypes to update source sampling weights. Experiments on thermochemistry, drug-target affinity, solvation free energy, and material bandgap prediction show consistent improvements in predictive error and tolerance accuracy.

7 Acknowledgment

This paper is partially supported by the National Natural Science Foundation of China (No.62502491). USTC School of Software Youth Innovation Fund Project (YN2260080011).

References

Alexander Amini, Wilko Schwarting, Ava Soleimany, and Daniela Rus. Deep evidential regression. In H. Larochelle, M. Ranzato, R. Hadsell, M. F. Balcan, and H. Lin, editors, *Advances in Neural Information Processing Systems*, volume 33, pages 14927–14937. Curran Associates, Inc., 2020.

Valentine P. Ananikov. Top 20 influential ai-based technologies in chemistry. *Artificial Intelligence Chemistry*, 2(2):100075, 2024. ISSN 2949-7477. doi: https://doi.org/10.1016/j.aichem.2024.100075.

Charles Blundell, Julien Cornebise, Koray Kavukcuoglu, and Daan Wierstra. Weight uncertainty in neural network. In Francis Bach and David Blei, editors, *Proceedings of the 32nd International Conference on Machine Learning*, volume 37 of *Proceedings of Machine Learning Research*, pages 1613–1622, Lille, France, 2015. PMLR.

Pedro Borlido, Thorsten Aull, Ahmad W. Huran, Fabien Tran, Miguel A. L. Marques, and Silvana Botti. Large-scale benchmark of exchange–correlation functionals for the determination of electronic band gaps of solids. *Journal of Chemical Theory and Computation*, 15:5069 – 5079, 2019.

Frederick Nii Ofei Bruce, Di Zhang, Xin Bai, Siwei Song, Fang Wang, Qingzhao Chu, Dongping Chen, and Yang Li. Machine learning predictions of thermochemical properties for aliphatic carbon and oxygen species. *Fuel*, 2025.

Xibao Cai, Houtim Lai, Xing Wang, Longyue Wang, Wei Liu, Yijun Wang, Zixu Wang, Dongsheng Cao, and Xiangxiang Zeng. Comprehensive evaluation of molecule property prediction with chatgpt. *Methods*, 222:133–141, 2024.

Dingshuo Chen, Zhixun Li, Yuyan Ni, Guibin Zhang, Ding Wang, Qiang Liu, Shu Wu, Jeffrey Yu, and Liang Wang. Beyond efficiency: Molecular data pruning for enhanced generalization. *Advances in Neural Information Processing Systems*, 37:18036–18061, 2024.

Seyone Chithrananda, Gabriel Grand, and Bharath Ramsundar. Chemberta: Large-scale self-supervised pretraining for molecular property prediction. *ArXiv*, abs/2010.09885, 2020.

Sanggyu Chong, Sangwon Lee, Baekjun Kim, and Jihan Kim. Applications of machine learning in metal-organic frameworks. *Coordination Chemistry Reviews*, 423:213487, 2020.

Cody Coleman, Christopher Yeh, Stephen Mussmann, Baharan Mirzasoleiman, Peter Bailis, Percy Liang, Jure Leskovec, and Matei Zaharia. Selection via proxy: Efficient data selection for deep learning. In *International Conference on Learning Representations*, 2020.

Larry A Curtiss, Paul C Redfern, and Krishnan Raghavachari. Gaussian-4 theory. *The Journal of chemical physics*, 126(8), 2007.

Mindy I. Davis, Jeremy P Hunt, Sanna Herrgård, Pietro Ciceri, Lisa M. Wodicka, Gabriel Pallares, Michael Hocker, Daniel K. Treiber, and Patrick P. Zarrinkar. Comprehensive analysis of kinase inhibitor selectivity. *Nature Biotechnology*, 29:1046–1051, 2011.

Wenjie Du, Shuai Zhang, Zhaohui Cai, Zhiyuan Liu, Junfeng Fang, Jianmin Wang, and Yang Wang. Molecular merged hypergraph neural network for explainable solvation gibbs free energy prediction. *Research*, 8, 2025.

Natalie S Eyke, Brent A Koscher, and Klavs F Jensen. Toward machine learning-enhanced high-throughput experimentation. *Trends in Chemistry*, 3(2):120–132, 2021.

Clyde Fare, Peter Fenner, Matthew Benatan, Alessandro Varsi, and Edward O Pyzer-Knapp. A multi-fidelity machine learning approach to high throughput materials screening. *npj Computational Materials*, 8(1):257, 2022.

Yarin Gal and Zoubin Ghahramani. Dropout as a bayesian approximation: Representing model uncertainty in deep learning. In *International Conference on Machine Learning*, 2015.

John LA Gardner, Hannes Schulz, Jean Helie, Lixin Sun, and Gregor NC Simm. Understanding multi-fidelity training of machine-learned force-fields. *arXiv preprint arXiv:2506.14963*, 2025.

Justin Gilmer, Samuel S. Schoenholz, Patrick F. Riley, Oriol Vinyals, and George E. Dahl. Neural message passing for quantum chemistry. In *International Conference on Machine Learning*, 2017.

Garrett B Goh, Nathan O Hodas, and Abhinav Vishnu. Deep learning for computational chemistry. *Journal of computational chemistry*, 38(16):1291–1307, 2017.

Garrett B Goh, Charles Siegel, Abhinav Vishnu, and Nathan Hodas. Using rule-based labels for weak supervised learning: a chemnet for transferable chemical property prediction. In *Proceedings of the 24th ACM SIGKDD international conference on knowledge discovery & data mining*, pages 302–310, 2018.

Daya Guo et al. DeepSeek-R1 incentivizes reasoning in LLMs through reinforcement learning. *Nature*, 645(8081):633–638, 2025. doi: 10.1038/s41586-025-09422-z.

Abimael Guzman-Pando, Graciela Ramirez-Alonso, Carlos Arzate-Quintana, and Javier Camarillo-Cisneros. Deep learning algorithms applied to computational chemistry. *Molecular Diversity*, 28 (4):2375–2410, 2024. doi: 10.1007/s11030-023-10771-y.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, and Weinan E. Exploring molecular pretraining model at scale. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 46956–46978. Curran Associates, Inc., 2024. doi: 10.52202/079017-1489.

Alex Kendall and Yarin Gal. What uncertainties do we need in bayesian deep learning for computer vision? In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. GLISTER: Generalization based data subset selection for efficient and robust learning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(9):8110–8118, 2021. doi: 10.1609/aaai.v35i9.16988.

Jaesun Kim, Jisu Kim, Jaehoon Kim, Jiho Lee, Yutack Park, Youngho Kang, and Seungwu Han. Data-efficient multifidelity training for high-fidelity machine learning interatomic potentials. *Journal of the American Chemical Society*, 147(1):1042–1054, 2024.

Sangtae Kim, Miso Lee, Changho Hong, Youngchae Yoon, Hyungmin An, Dongheon Lee, Wonseok Jeong, Dongsun Yoo, Youngho Kang, Yong Youn, et al. A band-gap database for semiconducting inorganic materials calculated with hybrid functional. *Scientific Data*, 7(1):387, 2020.

Thomas N. Kipf and Max Welling. Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, 2017.

Thorren Kirschbaum and Annika Bande. Transfer learning for molecular property predictions from small data sets. *ArXiv*, abs/2404.13393, 2024.

Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In *International conference on machine learning*, pages 1885–1894. PMLR, 2017.

Zhiqiang Kou, Jing Wang, Jiawei Tang, Yuheng Jia, Boyu Shi, and Xin Geng. Exploiting multi-label correlation in label distribution learning. In Kate Larson, editor, *Proceedings of the Thirty-Third International Joint Conference on Artificial Intelligence, IJCAI-24*, pages 4326–4334. International Joint Conferences on Artificial Intelligence Organization, 8 2024. doi: 10.24963/ijcai.2024/478. Main Track.

Zhiqiang Kou, Si Qin, Hailin Wang, Jing Wang, Mingkun Xie, Shuo Chen, Yuheng Jia, Tongliang Liu, Masashi Sugiyama, and Xin Geng. Label distribution learning with biased annotations assisted by multi-label learning. In James Kwok, editor, *Proceedings of the Thirty-Fourth International Joint Conference on Artificial Intelligence, IJCAI-25*, pages 5545–5553. International Joint Conferences on Artificial Intelligence Organization, 8 2025a. doi: 10.24963/ijcai.2025/617. Main Track.

Zhiqiang Kou, Jing Wang, Yuheng Jia, Biao Liu, and Xin Geng. Instance-dependent inaccurate label distribution learning. *IEEE Transactions on Neural Networks and Learning Systems*, 36(1): 1425–1437, 2025b. doi: 10.1109/TNNLS.2023.3329870.

Zhiqiang Kou, Yucheng Xie, Hailin Wang, Junyang Chen, Jingq Wang, Ming-Kun Xie, Shuo Chen, Yuheng Jia, Tongliang Liu, and Xin Geng. Rankmatch: A novel approach to semi-supervised label distribution learning leveraging rank correlation between labels. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen, editors, *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pages 158074–158095. Curran Associates, Inc., 2025c. doi: 10.52202/085713-5282.

Zhiqiang Kou, Junyang Chen, Xin-Qiang Cai, Xiaobo Xia, Ming-Kun Xie, Dong-Dong Wu, Biao Liu, Yuheng Jia, Xin Geng, Masashi Sugiyama, and Tat-Seng Chua. Positive–unlabeled reinforcement learning distillation for on-premise small models. In Tong Zhang, Miroslav Dudik, Martin Jaggi, Alekh Agarwal, Sharon Li, Dale Schuurmans, Jerry Zhu, Felix Berkenkamp, Hanze Dong, and Alberto Bietti, editors, *Proceedings of the 43rd International Conference on Machine Learning*, volume 306 of *Proceedings of Machine Learning Research*, pages 60447–60479. PMLR, 06–11 Jul 2026.

Christian Kramer, Tuomo Kalliokoski, Peter Gedeck, and Anna Vulpetti. The experimental uncertainty of heterogeneous public k(i) data. *Journal of medicinal chemistry*, 55(11):5165—5173, June 2012. ISSN 0022-2623. doi: 10.1021/jm300131x.

Tobias Kreiman and Aditi S. Krishnapriyan. Understanding and mitigating distribution shifts for universal machine learning interatomic potentials. *Digital Discovery*, 5(1):415–439, 2026. doi: 10.1039/D5DD00260E.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and scalable predictive uncertainty estimation using deep ensembles. In *Neural Information Processing Systems*, 2016.

Jia-Nan Li, Jian Guan, Songhao Wu, Wei Wu, and Rui Yan. From 1,000,000 users to every user: Scaling up personalized preference for user-level alignment. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens, editors, *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 30144–30168, San Diego, California, United States, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.acl-long.1391.

Juncai Li and Xiaofei Jiang. Mol-bert: An effective molecular representation with bert for molecular property prediction. *Wirel. Commun. Mob. Comput.*, 2021:7181815:1–7181815:7, 2021a.

Juncai Li and Xiaofei Jiang. Mol-bert: An effective molecular representation with bert for molecular property prediction. *Wireless Communications and Mobile Computing*, 2021(1):7181815, 2021b.

Stephanie McCready, Gavin F Birch, and Edward R Long. Metallic and organic contaminants in sediments of sydney harbour, australia and vicinity—a chemical dataset for evaluating sediment quality guidelines. *Environment International*, 32(4):455–465, 2006.

David L Mobley and J Peter Guthrie. Freesolv: a database of experimental and calculated hydration free energies, with input files. *Journal of computer-aided molecular design*, 28:711–720, 2014.

Gyoung S Na. Cross-modality material embedding loss for transferring knowledge between heterogeneous material descriptors. *npj Computational Materials*, 11(1):235, 2025.

Mansheej Paul, Surya Ganguli, and Gintare Karolina Dziugaite. Deep learning on a data diet: Finding important examples early in training. *Advances in neural information processing systems*, 34:20596–20607, 2021.

Ghanshyam Pilania, James E. Gubernatis, and Turab Lookman. Multi-fidelity machine learning models for accurate bandgap predictions of solids. *Computational Materials Science*, 129:156–163, 2017.

Ziheng Qin, Kai Wang, Zangwei Zheng, Jianyang Gu, Xiangyu Peng, xu Zhao Pan, Daquan Zhou, Lei Shang, Baigui Sun, Xuansong Xie, and Yang You. Infobatch: Lossless training speed up by unbiased dynamic data pruning. In *The Twelfth International Conference on Learning Representations*, 2024.

Ravi S Raju, Kyle Daruwalla, and Mikko Lipasti. Accelerating deep learning with dynamic data pruning. *arXiv preprint arXiv:2111.12621*, 2021.

Raghunathan Ramakrishnan, Pavlo O Dral, Matthias Rupp, and O Anatole Von Lilienfeld. Big data meets quantum chemistry approximations: The Δ-machine learning approach. *Journal of Chemical Theory and Computation*, 11(5):2087–2096, 2015. doi: 10.1021/acs.jctc.5b00099.

Rafael Cardoso Rial. Ai in analytical chemistry: Advancements, challenges, and future directions. *Talanta*, 274:125949, 2024.

Andrew S Rosen, Shaelyn M Iyer, Debmalya Ray, Zhenpeng Yao, Alan Aspuru-Guzik, Laura Gagliardi, Justin M Notestein, and Randall Q Snurr. Machine learning the quantum-chemical properties of metal–organic frameworks for accelerated materials discovery. *Matter*, 4(5):1578–1597, 2021.

Lars Ruddigkeit, Ruud Van Deursen, Lorenz C Blum, and Jean-Louis Reymond. Enumeration of 166 billion organic small molecules in the chemical universe database gdb-17. *Journal of chemical information and modeling*, 52(11):2864–2875, 2012.

Branko Ruscic and David H. Bross. Active thermochemical tables (ATcT) values based on ver. 1.220 of the thermochemical network, 2025.

Seongok Ryu, Yongchan Kwon, and Woo Youn Kim. A bayesian graph convolutional network for reliable prediction of molecular properties with uncertainty quantification. *Chemical Science*, 10: 8438 – 8446, 2019.

Gabriele Scalia, Colin A. Grambow, Barbara Pernici, Yi-Pei Li, and William H. Green. Evaluating scalable uncertainty estimation methods for deep learning-based molecular property prediction. *Journal of chemical information and modeling*, 2020.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. Evidential deep learning to quantify classification uncertainty. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31. Curran Associates, Inc., 2018.

Charlie Snell, Jaehoon Lee, Kelvin Xu, and Aviral Kumar. Scaling LLM test-time compute optimally can be more effective than scaling model parameters. In *International Conference on Learning Representations*, 2025.

Ava P. Soleimany, Alexander Amini, Samuel Goldman, Daniela Rus, Sangeeta N. Bhatia, and Connor W. Coley. Evidential deep learning for guided molecular property prediction and discovery. *ACS Central Science*, 7:1356 – 1367, 2021.

Ben Sorscher, Robert Geirhos, Shashank Shekhar, Surya Ganguli, and Ari Morcos. Beyond neural scaling laws: beating power law scaling via data pruning. *Advances in Neural Information Processing Systems*, 35:19523–19536, 2022.

Minyi Su, Qifan Yang, Yu Du, Guoqin Feng, Zhihai Liu, Yan Li, and Renxiao Wang. Comparative assessment of scoring functions: The casf-2016 update. *Journal of chemical information and modeling*, 59 2:895–913, 2018.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Neural Information Processing Systems*, 2017.

Florence H. Vermeire and William H. Green. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal*, 2020.

Vivin Vinod and Peter Zaspel. Benchmarking data efficiency in δ-ml and multifidelity models for quantum chemistry. *The Journal of Chemical Physics*, 163(2), 2025.

Ricardo Visini, Mahendra Awale, and Jean-Louis Reymond. Fragment database fdb-17. *Journal of chemical information and modeling*, 57(4):700–709, 2017.

W Patrick Walters and Regina Barzilay. Applications of deep learning in molecule generation and molecular property prediction. *Accounts of chemical research*, 54(2):263–270, 2020.

Sihan Wang, Wenjie Du, Qing Zhu, and Yang Wang. Bridging the gap between cross-domain theory and practical application: A case study on molecular dissolution. *Advances in Neural Information Processing Systems*, 38:129428–129459, 2026.

Zirui Wang, Zihang Dai, Barnabás Póczos, and Jaime G. Carbonell. Characterizing and avoiding negative transfer. *2019 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 11285–11294, 2018.

Zhepei Wei, Wenlin Yao, Yao Liu, Weizhi Zhang, Qin Lu, Liang Qiu, Changlong Yu, Puyang Xu, Chao Zhang, Bing Yin, Hyokun Yun, and Lihong Li. WebAgent-R1: Training web agents via end-to-end multi-turn reinforcement learning. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 7909–7928, Suzhou, China, 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.emnlp-main.401.

Max Welling. Herding dynamical weights to learn. In *Proceedings of the 26th annual international conference on machine learning*, pages 1121–1128, 2009.

Keyulu Xu, Weihua Hu, Jure Leskovec, and Stefanie Jegelka. How powerful are graph neural networks? In *International Conference on Learning Representations*, 2019.

Yougen Xu, Yadong Gao, Lebin Su, Haiting Wu, Hao Tian, Majian Zeng, Chunqiu Xu, Xinwei Zhu, and Kuangbiao Liao. High-throughput experimentation and machine learning-assisted optimization of iridium-catalyzed cross-dimerization of sulfoxonium ylides. *Angewandte Chemie International Edition*, 62(48):e202313638, 2023.

Suryanarayan Kumar Yadav, Parinita Tripathy, and Sanyogita Shahi. Artificial intelligence in chemistry: A transformative review. *Journal for Research in Applied Sciences and Biotechnology*, 2025.

Kevin Yang, Kyle Swanson, Wengong Jin, Connor W. Coley, Philipp Eiden, Hua Gao, Angel Guzman-Perez, Timothy Hopper, Brian P. Kelley, Miriam Mathea, Andrew Palmer, Volker Settels, T. Jaakkola, Klavs F. Jensen, and Regina Barzilay. Analyzing learned molecular representations for property prediction. *Journal of Chemical Information and Modeling*, 59:3370 – 3388, 2019.

Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. Dataset pruning: Reducing training data by examining generalization influence. In *International Conference on Learning Representations*, 2023.

A Related Work

This section reviews five areas related to DETS: machine learning for molecular sciences, learning with ambiguous and inaccurate supervision, transfer and multi-fidelity learning, uncertainty quantification, and data selection.

A.1 Machine Learning for Molecular Sciences

Machine learning is widely used in molecular discovery and materials design. Graph neural networks represent molecular topology and chemical interactions, with GIN, GCN, and MPNN serving as common backbones for molecular property prediction [Gilmer et al., 2017, Yang et al., 2019, Xu et al., 2019, Kipf and Welling, 2017]. Transformer architectures process protein sequences and SMILES strings through self-attention [Vaswani et al., 2017, Chithrananda et al., 2020]. These models are commonly trained against individual reported labels. For experimental data, small deviations from a label may fall within accepted measurement fluctuations, motivating an interval-aware training objective.

A.2 Learning with Ambiguous and Inaccurate Supervision

Label distribution learning represents the relevance of multiple labels to each instance as a distribution and provides a structured treatment of label ambiguity. Recent studies exploit multi-label correlations in label distribution learning [Kou et al., 2024] and recover label distributions from instance-dependent inaccuracies or biased annotations [Kou et al., 2025b,a]. In the semi-supervised setting, RankMatch preserves inter-label rankings while relaxing exact pseudo-label margins [Kou et al., 2025c]. DETS addresses a related supervision problem in continuous scientific regression by defining acceptable intervals around experimental labels and using tolerance-aware epistemic uncertainty to control source sampling.

A.3 Bridging the Fidelity Gap: Transfer and Multi-fidelity Learning

Transfer learning commonly pre-trains a model on large proxy datasets, such as DFT calculations, and fine-tunes it on experimental data [Goh et al., 2017, Li and Jiang, 2021a]. Multi-fidelity methods, including Δ-machine learning, model relationships between low- and high-fidelity properties [Ramakrishnan et al., 2015, Pilania et al., 2017]. Their performance depends on the correspondence between source and target domains. Systematic source bias can produce negative transfer when source patterns do not generalize to the target task [Wang et al., 2018].

A.4 Uncertainty Quantification in Deep Learning

Bayesian neural networks estimate predictive uncertainty through a distribution over model parameters [Blundell et al., 2015]. Monte Carlo dropout [Gal and Ghahramani, 2015] and deep ensembles [Lakshminarayanan et al., 2016] provide scalable approximations but require repeated inference or multiple trained models. Evidential deep learning instead predicts the parameters of a conjugate prior in one forward pass [Sensoy et al., 2018, Amini et al., 2020]. Standard evidential regression analytically separates aleatoric and epistemic components, while its pointwise likelihood still penalizes every residual. Residuals within an accepted tolerance can therefore affect the learned evidence and the resulting epistemic estimate [Soleimany et al., 2021].

A.5 Data Selection and Coreset Sampling

Data selection aims to identify the most informative subset of data to maximize model performance under computational constraints. Existing methods can be broadly categorized into static and dynamic strategies. Static methods, such as Herding [Welling, 2009] and K-Means [Sorscher et al., 2022], select representative samples based on geometric diversity in the feature space before training. Dynamic methods, such as active learning strategies based on entropy or least confidence [Coleman et al., 2020], and recent approaches like InfoBatch [Qin et al., 2024], adjust sampling probabilities during training based on loss or gradient information. In the context of cross-domain learning, methods like MolPeg [Chen et al., 2024] and CSIE [Wang et al., 2026] employ mixup or domain adaptation techniques to bridge distribution shifts.

B Dataset Details and Preprocessing

B.1 Data Source and Statistics

We evaluate DETS on four regression tasks spanning thermodynamics, biochemistry, and materials science. Each task contains a source–target fidelity gap.

For the **Thermochemical Property (Enthalpy)** task, *Wudily_Cho* [Bruce et al., 2025] provides calculated formation enthalpies for 6,074 hydrocarbons, oxygenates, and radicals. The values were obtained with composite methods including CBS-QB3, CBS-APNO, G3, and G4. We removed samples with enthalpy values outside ± 500 kJ/mol. The target *Active Thermochemical Tables (ATcT)* [Ruscic and Bross, 2025] dataset contains 358 gas-phase enthalpies derived from a statistical analysis of thermochemical networks.

For **Drug-Target Binding Affinity (DTA)**, the low-fidelity *Davis* dataset [Davis et al., 2011] contains 24,956 interaction pairs between 68 kinase inhibitors and 442 proteins. Its K_d values come from kinase inhibition assays and may vary with assay conditions. The target *CASF-2016* benchmark [Su et al., 2018] is a curated subset of PDBbind with 285 protein–ligand complexes, X-ray structures at resolutions of 2.5 Å or better, and associated pK_d values. This task evaluates transfer across data sources and measurement protocols.

For **Solvation Free Energy**, the source *CombiSolv-QM* dataset [Vermeire and Green, 2020] contains one million solute–solvent pairs calculated with COSMO-RS. The target *CombiSolv-Exp* dataset [Vermeire and Green, 2020] contains 8,780 experimental measurements collected from MiniSolv, FreeSolv, the Abraham database, and other sources. The task transfers from continuum-model calculations to experimental measurements.

For **Material Bandgap** prediction, the source *NREL-MatDB (PBE)* dataset [Kim et al., 2020] contains 7,909 inorganic crystalline materials calculated with the PBE functional. PBE commonly underestimates electronic bandgaps, creating a systematic source bias. The target *NREL-GWVD* dataset [Kim et al., 2020] contains GW bandgaps for 881 materials. This task evaluates transfer between two electronic-structure methods with different fidelities.

B.2 Feature Engineering and Preprocessing

We process raw chemical data into model-ready formats using domain-specific strategies.

For molecular tasks (enthalpy and solvation) and the drug component in DTA, we use `RDKit` and `OGB` to convert SMILES strings into graphs. Atom features include atomic number, chirality, degree, formal charge, hydrogen count, hybridization, aromaticity, and atomic mass. Bond features include bond type, conjugation, and ring membership. Heavy-atom count and molecular weight are normalized by 30 and 400, respectively, and added as global inputs to the GIN.

For the protein component in DTA, we use a 23-token amino-acid vocabulary that includes `X`, `-`, and `?`. Sequences are mapped to integer IDs and padded or truncated to length 1024. An attention mask excludes padding tokens, and the transformer encoder processes the resulting sequence.

For the material bandgap task, we use the `matminer` library to generate compositional features from chemical formulas. We extract element-fraction vectors and compute statistics with the `Magpie` preset, including atomic radii, electronegativity, and valence electrons. The resulting dense vector represents material stoichiometry and elemental properties and serves as input to the MLP.

C Implementation and Experimental Setup

C.1 Evaluation Metrics and Physical Threshold Mechanism

Mean absolute percentage error (MAPE) becomes unstable when a reported target value is close to zero. We therefore bound its denominator with the absolute tolerance ϵ_{abs}, which is set from domain knowledge or target-validation residuals. The bounded metric is

$$\mathrm{MAPE}_{\epsilon_{\mathrm{abs}}} = \frac{1}{N}\sum_{i=1}^{N}\frac{|y_i - \hat{y}_i|}{\max(|y_i|, \epsilon_{\mathrm{abs}})}, \tag{22}$$

where y_i is the reported target value, $\hat{y}_i$ is the prediction, and N is the number of samples. We apply the same absolute tolerance to Acc@10%: a prediction is counted as accurate when $|y_i - \hat{y}_i| \leq \max(0.1|y_i|, \epsilon_{\mathrm{abs}})$. This definition keeps both metrics finite and interpretable for small-magnitude targets.

We set ϵ_{abs} separately for each task. For enthalpy, chemical accuracy gives $1\ \mathrm{kcal/mol} = 4.184\ \mathrm{kJ/mol}$ [Curtiss et al., 2007]. We use 0.1 pKd for DTA [Kramer et al., 2012], 0.1 kcal/mol for solvation free energy [Mobley and Guthrie, 2014], and 0.1 eV for bandgap prediction [Borlido et al., 2019]. These values define the task-specific absolute tolerances used by the interval loss and the bounded evaluation metrics.

C.2 Data-driven Selection of the Threshold

When reliable domain priors are available, we set ϵ_{abs} from known experimental resolution or an accepted accuracy threshold. Without such priors, we train an MSE baseline on the source domain and set ϵ_{abs} to the 10th percentile of its absolute residuals on a held-out target validation set:

$$\epsilon_{\mathrm{abs}} = \mathrm{Quantile}_{0.10}\left(\left\{|y_j^{(t)} - \hat{y}_j^{(t)}|\right\}_{j\in\mathcal{V}_t}\right). \tag{23}$$

This rule defines an operational tolerance from the lower tail of the validation residual distribution.

Table 4: Prior-based and data-driven tolerance thresholds. MAE and Acc@10 independently report the best result over the candidate retention ratios and may therefore correspond to different ratios.

Task	Threshold Type	MAE↓	Acc@10↑	Threshold
Enthalpy	Domain prior	**23.11**	**25.00**	4.184 kJ/mol
	Data-driven	24.18	19.44	3.870 kJ/mol
DTA	Domain prior	**1.2734**	**38.33**	0.1000 pKd
	Data-driven	1.3105	36.67	0.2159 pKd
Solvation	Domain prior	**0.2258**	**94.87**	0.1000 kcal/mol
	Data-driven	0.2366	93.13	0.0697 kcal/mol
Bandgap	Domain prior	**0.5413**	**32.39**	0.100 eV
	Data-driven	0.5687	29.54	0.190 eV

The data-driven rule is fixed before test evaluation and requires a labeled target validation set. Domain knowledge or an external tolerance specification is required when such validation data are unavailable.

C.3 DETS Training Algorithm

Algorithm 1 summarizes the training protocol. Before prototype construction and dynamic source sampling, DETS performs a warm-up stage. The warmed-up encoder extracts target representations, selects $K^\star$ by silhouette score, and constructs K-Means prototypes. During source pre-training, DETS computes epistemic uncertainty and structural similarity, samples a weighted coreset, and updates the model on that subset.

After source pre-training, target fine-tuning proceeds in two steps. Head alignment freezes the encoder and updates the evidential head on target data. Full adaptation then unfreezes all parameters and continues end-to-end optimization at the lower fine-tuning learning rate. Both steps use the interval-censored objective defined in Eq. 13.

Algorithm 1 Dual-Evidence Tolerance Sampling (DETS)

Input: source dataset $\mathcal{S} = \{x_i^{(s)}, y_i^{(s)}\}$, target dataset $\mathcal{T} = \{x_j^{(t)}, y_j^{(t)}\}$.
Hyperparameters: Tolerances ϵ_{abs} and ρ, coefficient λ_t, bandwidth parameters σ_0 and η, candidate prototype set $\mathcal{K}$, refresh period R_{proto}, retention ratio r.
Initialization: Initialize encoder g_θ and evidential head h_ϕ.
Warm up g_θ and h_ϕ before prototype construction and dynamic sampling.
Extract target features $Z_t = g_\theta(\mathcal{T})$.
Select $K^\star$ by silhouette score over $\mathcal{K}$ and initialize target prototypes $\mathcal{P}$ by K-Means.
Phase 1: Dynamic Source Pre-training
for epoch $e = 1$ to E_{pre} **do**
  if $e \bmod R_{\text{proto}} = 0$ **then**
    Refresh $Z_t = g_\theta(\mathcal{T})$ and update prototypes $\mathcal{P}$ by K-Means with $K^\star$.
  end if
  for batch $B_s \subset \mathcal{S}$ **do**
    Forward pass: $\mathbf{m}_i = (\gamma_i, \nu_i, \alpha_i, \beta_i) = h_\phi(g_\theta(x_i))$, $z_i = g_\theta(x_i)$.
    Compute epistemic uncertainty $U_{epi}^{(i)} = \beta_i/(\nu_i(\alpha_i - 1))$.
    Compute structural evidence $E_{stat}^{(i)}$ using Eq. 16.
    Normalize uncertainty $\bar{U}_{epi}^{(i)}$ within the current batch.
    Compute $\sigma_{eff}^{(i)} = \sigma_0(1 + \eta\bar{U}_{epi}^{(i)})$.
    Compute sampling probability w_i using Eq. 19.
    Select coreset $\mathcal{C}_e \subset B_s$ by weighted sampling according to w_i.
    Update θ, ϕ by minimizing $\mathcal{L}_{total}$ on $\mathcal{C}_e$.
  end for
end for
Phase 2: Target Fine-tuning
Step 2.1: Freeze g_θ, update h_ϕ on $\mathcal{T}$ for 200 epochs.
Step 2.2: Unfreeze all, update g_θ, h_ϕ on $\mathcal{T}$ with a small learning rate.
Return: Trained model.

C.4 Model Architectures

We use one backbone architecture for each data modality. For molecular property prediction (enthalpy and solvation), we use a three-layer graph isomorphism network (GIN) with a hidden dimension of 256. Each layer is followed by batch normalization and ReLU activation. Global mean pooling aggregates node features into a graph-level embedding.

For drug-target affinity, a three-layer GIN with hidden dimension 128 processes the drug molecule. A two-layer transformer with four attention heads processes 128-dimensional amino-acid embeddings. The drug and protein embeddings are concatenated and projected to a joint latent space.

For material bandgap prediction, a multilayer perceptron takes a 145-dimensional vector of compositional descriptors. Its hidden dimensions are 256, 128, and 64, with batch normalization and ReLU activation.

All backbones feed into the same evidential regression head, which outputs the four Normal-Inverse-Gamma parameters γ, ν, α, and β. Softplus enforces positive ν and β, and Softplus plus one enforces $\alpha > 1$.

C.5 Baselines Implementation Details

All baselines use the same task-specific backbone and target fine-tuning protocol as DETS. Static and dynamic methods compute their selection statistics from source-domain data. Cross-domain methods may use unlabeled target features, and no method uses individual target labels to set source-sampling weights.

Static Coreset Selection. Hard Random samples a fixed source budget uniformly without replacement. **Herding** [Welling, 2009] selects source samples whose mean feature approaches the mean of the full source set, and **K-Means** [Sorscher et al., 2022] selects source samples nearest to cluster

centroids. **Entropy** and **Least Confidence** [Coleman et al., 2020] rank source samples by predictive uncertainty. **GraNd-20** and **EL2N-2** [Paul et al., 2021] use early-training gradient norms and error norms, respectively. **Glister** [Killamsetty et al., 2021] uses bilevel subset selection with a fixed source validation split, and **Influence** [Koh and Liang, 2017] estimates the effect of removing a source sample on the source validation loss. **DP** [Yang et al., 2023] selects a diverse subset through the determinant of a source-feature kernel matrix.

Dynamic Coreset Selection. Soft Random assigns source-sampling probabilities with a Boltzmann distribution over the current source losses. **ϵ-greedy** and **UCB** [Raju et al., 2021] use bandit policies based on source loss or uncertainty. **InfoBatch** [Qin et al., 2024] updates source-sampling probabilities from the preceding epoch's loss distribution.

Cross-Domain Adaptation. MolPeg [Chen et al., 2024] applies graph-level manifold mixup between source and target representations with coefficients drawn from $\mathrm{Beta}(0.2, 0.2)$. **CSIE** [Wang et al., 2026] uses a gradient-reversal layer and a two-layer MLP domain discriminator to learn representations with reduced source–target discrepancy.

C.6 Hyperparameter Configuration

We select hyperparameters on the validation set. The search uses $\rho \in \{0.01, 0.05, 0.10\}$ and $\lambda \in [10^{-3}, 10^{-1}]$. Adam with cosine annealing uses a learning rate of 10^{-3} for source pre-training and 10^{-5} for target fine-tuning. The sampling bandwidth σ_0 is selected from 0.2 to 0.5, and τ_{high} is 0.9. The candidate prototype counts are 10 and 20; $K^\star$ is chosen by the silhouette score. Table 5 lists the selected settings.

Table 5: Hyperparameter settings for DETS. The units of ϵ_{abs} are kJ/mol, pKd, kcal/mol, and eV for enthalpy, DTA, solvation, and bandgap, respectively.

Parameter	Symbol	Enthalpy	DTA	Solvation	Bandgap
Relative tolerance	ρ	0.05	0.05	0.05	0.05
Absolute tolerance	ϵ_{abs}	4.184	0.1	0.1	0.1
Regularization coeff.	λ_t	0.01	0.01	0.01	0.01
Base Bandwidth	σ_0	0.5	0.25	0.2	0.5
Conf. Threshold	τ_{high}	0.9	0.9	0.9	0.9
Prototypes	$K^\star$	10	10	20	10
Pre-train LR	lr_{pre}	1e-3	1e-3	1e-3	1e-3
Fine-tune LR	lr_{ft}	1e-5	1e-5	1e-5	1e-5
Source Batch Size	B_s	128	64	256	32
Target Batch Size	B_t	16	32	64	32
Pre-train Epochs	E_{pre}	200	1000	100	2000

C.7 Sensitivity to λ and σ_0

We vary one parameter at a time while keeping the remaining training settings fixed. Tables 6 and 7 report the three-seed mean $\pm$ sample standard deviation and the mean sampling weight over all source candidates.

D Extensive Experimental Results

D.1 Impact of Coreset Retention Ratios

We evaluate retention ratios $r \in \{20\%, 40\%, 60\%, 80\%\}$, where r is the fraction of source samples retained for each epoch. Tables 8–11 report all methods and ratios. Figure 4 shows representative methods from each baseline category for readability. DETS remains effective at low retention ratios because the ICEE epistemic estimate adjusts the sampling bandwidth.

Performance varies across retention ratios because additional source samples can increase both coverage and source–target mismatch. Across the four tasks, DETS with $r = 40\%$ outperforms full-

Table 6: Sensitivity to the regularization coefficient λ across four tasks.

Task	λ	MAE↓	Acc@10↑	Mean weight
Enthalpy	0.001	23.7781 ± 1.5994	24.9444 ± 0.2778	0.9562
	0.005	23.7747 ± 1.3931	23.0185 ± 0.2912	0.9562
	0.010	24.0549 ± 1.2258	25.3333 ± 0.2556	0.9566
	0.050	23.8861 ± 1.2982	25.6481 ± 0.3647	0.9614
	0.100	24.3704 ± 1.7055	25.9444 ± 0.0077	0.9671
DTA	0.001	1.2946 ± 0.2600	35.9487 ± 0.2923	0.6470
	0.005	1.2718 ± 0.1834	36.3675 ± 0.2335	0.6446
	0.010	1.2723 ± 0.2434	38.0769 ± 0.2840	0.6440
	0.050	1.2784 ± 0.2020	37.2222 ± 0.3335	0.6563
	0.100	1.2760 ± 0.2777	38.5128 ± 0.2840	0.6846
Solvation	0.001	0.2221 ± 0.0398	92.6234 ± 2.3920	0.6050
	0.005	0.2277 ± 0.0291	91.6743 ± 1.4970	0.5668
	0.010	0.2113 ± 0.0294	91.4465 ± 2.3782	0.5813
	0.050	0.2275 ± 0.0276	92.6044 ± 1.8526	0.7756
	0.100	0.2268 ± 0.0296	92.4146 ± 2.0838	0.7803
Bandgap	0.001	0.5405 ± 0.0297	32.3258 ± 1.7477	0.9011
	0.005	0.5430 ± 0.0146	31.1894 ± 3.4716	0.8975
	0.010	0.5461 ± 0.0293	31.7576 ± 3.4716	0.8970
	0.050	0.5508 ± 0.0242	32.2727 ± 2.0486	0.9332
	0.100	0.5419 ± 0.0239	32.0833 ± 2.2963	0.9619

Table 7: Sensitivity to the base sampling bandwidth σ_0 across four tasks.

Task	σ_0	MAE↓	Acc@10↑	Mean weight
Enthalpy	0.10	23.9503 ± 0.6891	25.7222 ± 0.2747	0.3686
	0.20	24.3691 ± 2.0012	24.4815 ± 0.1216	0.7632
	0.25	23.5931 ± 1.0332	25.4074 ± 0.1912	0.8396
	0.40	23.8603 ± 1.4118	23.0926 ± 0.2038	0.9332
	0.60	23.4388 ± 1.6770	24.9444 ± 0.2747	0.9696
DTA	0.10	1.3028 ± 0.2323	37.6581 ± 0.3627	0.0941
	0.20	1.2967 ± 0.2254	37.9487 ± 0.1282	0.5081
	0.25	1.2723 ± 0.2434	38.0769 ± 0.3840	0.6440
	0.40	1.3112 ± 0.2875	36.9487 ± 0.2564	0.8397
	0.60	1.3493 ± 0.2817	36.6581 ± 0.2048	0.9249
Solvation	0.10	0.2411 ± 0.0221	91.4655 ± 2.5091	0.2285
	0.20	0.2413 ± 0.0294	91.4465 ± 2.3782	0.5813
	0.25	0.2353 ± 0.0189	91.4085 ± 2.2058	0.6897
	0.40	0.2293 ± 0.0283	91.4465 ± 2.0462	0.8542
	0.60	0.2285 ± 0.0235	92.2817 ± 1.5969	0.9303
Bandgap	0.10	0.5517 ± 0.0255	31.7045 ± 3.7258	0.1876
	0.20	0.5717 ± 0.0474	31.2424 ± 2.5621	0.5610
	0.25	0.5497 ± 0.0300	32.5152 ± 3.3292	0.6748
	0.40	0.5581 ± 0.0356	31.6212 ± 2.3702	0.8468
	0.60	0.5638 ± 0.0303	31.5152 ± 0.6561	0.9264

source training on the reported metrics. The result shows that selective source sampling can use less source data while reducing the influence of shifted source samples.

D.2 Computational Efficiency and Scalability

We measure wall-clock training time per epoch on the molecular enthalpy task. Random sampling has no selection step and serves as the 1.0x reference.

DETS requires 1.49x the reference time. The silhouette score selects $K^\star$ once, and prototypes are refreshed every R_{proto} epochs. Sampling weights are computed with vectorized operations on latent embeddings.

Glister and GraNd exceed 10x the reference time because they use bilevel optimization or per-sample gradient norms with additional forward and backward passes. DETS adds 49% overhead without these per-sample gradient computations.

D.3 Uncertainty Calibration via ICEE

We assess whether U_{epi} ranks target samples according to prediction error by partitioning them into ten deciles of estimated epistemic uncertainty.

Table 8: Results for **Molecular Enthalpy** across all retention ratios.

Method	MAE (↓)				RMSE (↓)				MAPE (↓)				Acc@10 (↑)			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
DETS (Ours)	24.57	**23.11**	24.63	23.87	28.83	**25.71**	29.13	28.46	73.24	**73.07**	75.63	74.18	**25.00**	22.67	24.33	22.22
Full dataset	34.59	34.59	34.59	34.59	44.11	44.11	44.11	44.11	106.54	106.54	106.54	106.54	5.56	5.56	5.56	5.56
Hard Random	27.73	29.84	29.98	24.56	38.13	41.87	40.75	35.72	92.15	96.59	125.50	83.00	13.89	8.33	11.11	19.44
Herding	29.97	28.52	26.65	27.63	41.29	38.96	37.18	37.53	108.47	105.98	94.20	100.79	13.89	16.67	16.67	16.67
K-means	26.65	25.26	30.59	30.92	37.76	35.05	41.09	40.15	88.59	94.83	111.51	97.43	16.67	16.67	5.56	0.00
Entropy	23.11	30.42	28.38	25.56	32.49	41.03	39.14	36.6	74.73	113.28	108.37	101.42	13.89	8.33	13.89	5.56
Least Confidence	27.98	23.68	27.43	25.28	38.64	32.37	38.07	34.85	91.07	90.19	111.48	116.33	13.89	11.11	11.11	16.67
GraNd-20	28.53	27.66	27.65	27.68	39.59	36.98	37.78	36.71	108.54	109.87	106.29	108.76	11.11	13.89	13.89	8.33
Glister	27.69	26.22	30.06	27.55	40.59	35.77	40.46	38.35	116.04	90.64	109.91	102.37	19.44	11.11	5.56	13.89
Influence	29.72	26.8	29.27	26.72	41.05	38.28	38.76	38.06	104.37	96.58	112.25	80.20	11.11	13.89	2.78	16.67
EL2N-2	28.14	26.94	25.75	28.09	37.34	38.24	35.27	38.99	108.91	103.99	122.20	102.70	5.56	25.00	19.44	11.11
DP	27.88	26.04	26.77	28.92	36.76	35.67	37.58	39.8	97.19	106.55	100.30	106.96	13.89	16.67	13.89	8.33
Soft Random	29.19	29.36	27.72	26.38	40.86	38.81	36.97	35.93	102.41	100.11	114.56	85.57	2.78	0.00	5.56	8.33
ϵ-greedy	28.58	27.6	27.78	26.36	38.95	38.27	38.43	25.71	87.44	98.03	100.80	101.04	5.56	8.33	16.67	5.56
UCB	28.99	28.63	29.08	27.39	40.19	39.3	40.99	36.38	110.57	119.55	109.63	95.14	8.33	11.11	8.33	5.56
InfoBatch	27.22	25.56	28.34	26.57	39.63	36.03	37.93	36.7	106.81	89.90	116.27	96.72	13.89	11.11	8.33	8.33
MolPeg	29.06	29.07	27.13	26.28	40.59	39.79	38.94	37.74	95.69	109.77	96.10	73.07	11.11	13.89	11.11	22.22
CSIE	26.53	29.54	27.75	25.14	37.4	40.29	38.01	34.79	124.78	117.22	110.47	83.30	13.89	11.11	11.11	19.44

Table 9: Results for **DTA** across all retention ratios.

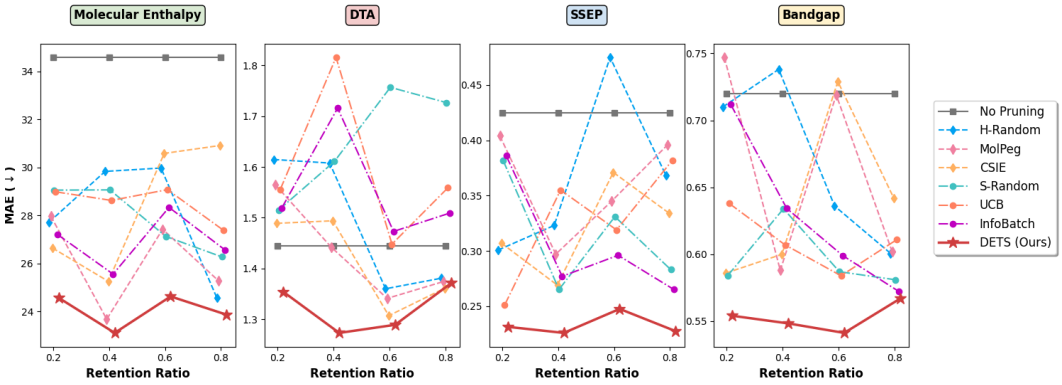
Method	MAE (↓)				RMSE (↓)				MAPE (↓)				Acc@10 (↑)			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
DETS (Ours)	1.354	**1.273**	1.289	1.371	**1.6419**	1.7191	1.6654	1.6837	22.46	**21.90**	23.15	22.11	**38.33**	35.00	33.23	32.14
Full dataset	1.4441	1.4441	1.4441	1.4441	1.8136	1.8136	1.8136	1.8136	27.12	27.12	27.12	27.12	30.0	30.0	30.0	30.0
Hard Random	1.6148	1.6078	1.3600	1.3809	1.9248	2.0086	1.8533	1.7368	27.80	27.98	22.97	22.75	25.0	30.0	40.0	25.0
Herding	1.3748	1.5336	1.5230	1.4456	1.7367	1.8617	1.8337	1.8172	24.38	26.33	26.57	24.17	30.0	35.0	25.0	30.0
K-means	1.4890	1.4940	1.3076	1.3614	1.8736	1.8891	1.6996	1.724	25.64	25.49	21.90	22.62	30.0	30.0	35.0	30.0
Entropy	1.5318	1.6983	1.5741	1.5679	1.916	2.0023	1.9809	2.002	26.47	28.22	25.00	24.79	20.0	15.0	20.0	25.0
Least Confidence	1.5654	1.4428	1.3418	1.3741	1.9348	1.9123	1.8556	1.8749	25.40	23.92	23.68	24.19	30.0	35.0	35.0	40.0
GraNd-20	1.563	1.3832	1.5031	1.3627	1.9413	1.7613	1.9704	1.7465	26.64	22.95	24.71	25.43	25.0	25.0	30.0	25.0
Glister	1.3854	1.4206	1.6673	1.5872	1.692	1.7854	2.169	1.9346	23.87	24.14	28.58	25.94	35.0	25.0	30.0	25.0
Influence	1.5191	1.3761	1.7129	1.7102	1.8728	1.7883	2.3268	2.0941	25.32	23.28	29.23	28.19	25.0	35.0	20.0	20.0
EL2N-2	1.571	1.7563	1.2734	1.3121	1.9301	2.2469	1.6419	1.6708	25.56	30.18	22.34	23.02	35.0	25.0	20.0	25.0
DP	1.5012	1.3967	1.6636	1.6253	1.8734	1.7992	2.0574	2.0329	25.77	22.55	28.00	27.25	25.0	25.0	15.0	25.0
Soft Random	1.5413	1.5032	1.4501	1.3982	1.9138	1.8945	1.7907	1.704	25.17	24.67	24.87	23.93	25.0	25.0	30.0	30.0
ϵ-greedy	1.5443	1.6491	1.4275	1.4387	1.9005	2.029	1.8585	1.8522	25.44	27.07	24.34	23.66	35.0	20.0	30.0	35.0
UCB	1.5559	1.8162	1.4487	1.5598	1.934	2.1918	1.8480	1.941	25.56	30.08	24.27	27.79	30.0	20.0	25.0	20.0
InfoBatch	1.5191	1.7161	1.4735	1.5091	1.8405	2.0324	1.829	1.8913	25.33	28.06	25.00	25.87	25.0	10.0	30.0	25.0
MolPeg	1.5156	1.6113	1.7571	1.7276	1.8754	1.9368	2.2662	2.2759	24.83	26.80	29.04	29.78	25.0	20.0	20.0	30.0
CSIE	1.6483	1.629	1.5955	1.6242	1.9304	2.0186	2.0159	2.0057	27.35	27.04	26.43	27.76	10.0	20.0	15.0	20.0

Table 10: Results for **Solvation Free Energy** across all retention ratios.

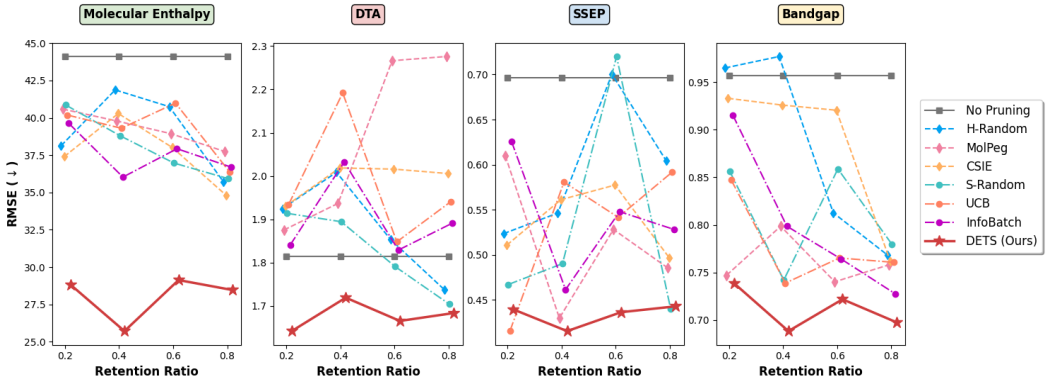
Method	MAE (↓)				RMSE (↓)				MAPE (↓)				Acc@10 (↑)			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
DETS (Ours)	0.2313	**0.2258**	0.2475	0.2274	0.4397	**0.4155**	0.4365	0.4429	6.34	**5.48**	7.84	5.79	92.11	93.42	**94.87**	92.23
Full dataset	0.4253	0.4253	0.4253	0.4253	0.696	0.696	0.696	0.696	12.7	12.7	12.7	12.7	79.95	79.95	79.95	79.95
Hard Random	0.3008	0.323	0.475	0.3681	0.5236	0.5464	0.7003	0.6046	8.46	8.95	13.82	10.45	88.95	88.50	73.92	86.10
Herding	0.3204	0.3319	0.3265	0.3531	0.5251	0.5207	0.5143	0.5794	8.96	9.45	9.28	10.87	87.24	85.65	87.24	85.08
K-means	0.3078	0.2699	0.371	0.3341	0.5061	0.4476	0.6095	0.5492	8.99	7.66	10.05	10.23	88.61	91.23	85.99	87.02
Entropy	0.4215	0.5163	0.2874	0.2874	0.6848	0.7395	0.4662	0.4662	12.00	15.66	8.37	8.37	80.64	70.05	90.21	90.21
Least Confidence	0.4041	0.2974	0.3454	0.3962	0.6138	0.5199	0.576	0.6039	12.12	8.09	9.60	11.29	80.52	91.91	87.47	82.00
GraNd-20	0.3796	0.3319	0.3387	0.2881	0.6481	0.5351	0.54	0.484	10.49	9.27	10.00	8.12	82.92	86.45	86.22	90.43
Glister	0.2626	0.2834	0.3453	0.2771	0.4777	0.4914	0.612	0.4781	7.19	8.09	9.70	7.58	91.91	91.23	86.45	92.37
Influence	0.3647	0.2881	0.2826	0.277	0.6029	0.4515	0.4775	0.4515	9.07	8.78	7.32	7.60	87.02	88.84	91.80	91.57
EL2N-2	0.3491	0.3486	0.3699	0.2724	0.5504	0.585	0.6112	0.4492	9.57	9.57	10.21	7.53	86.10	85.99	85.08	91.34
DP	0.3571	0.3066	0.2881	0.3537	0.5394	0.5654	0.4515	0.6059	9.70	8.73	8.78	9.83	85.19	87.70	88.84	85.08
Soft Random	0.2822	0.2787	0.4098	0.273	0.4672	0.4903	0.7199	0.44	8.13	8.06	10.90	7.78	90.66	91.80	84.17	91.57
ϵ-greedy	0.3016	0.3475	0.4594	0.2654	0.5332	0.5768	0.7043	0.4299	8.69	9.32	12.98	7.29	89.98	86.90	78.02	92.48
UCB	0.2513	0.355	0.3197	0.3826	0.4155	0.5807	0.5417	0.5918	6.75	8.97	9.35	10.34	93.28	87.24	88.27	84.97
InfoBatch	0.3869	0.2771	0.2965	0.2654	0.6259	0.4615	0.5479	0.5283	10.97	7.44	8.18	7.10	84.40	91.57	90.66	93.17
MolPeg	0.3821	0.2654	0.3317	0.283	0.61	0.4299	0.5282	0.4857	10.72	7.29	9.89	8.49	83.49	92.48	87.36	90.32
CSIE	0.3188	0.3459	0.3582	0.2779	0.5108	0.5612	0.5774	0.497	9.25	10.38	9.92	7.07	87.59	86.22	85.88	92.14

Table 11: Results for **Bandgap** across all retention ratios.

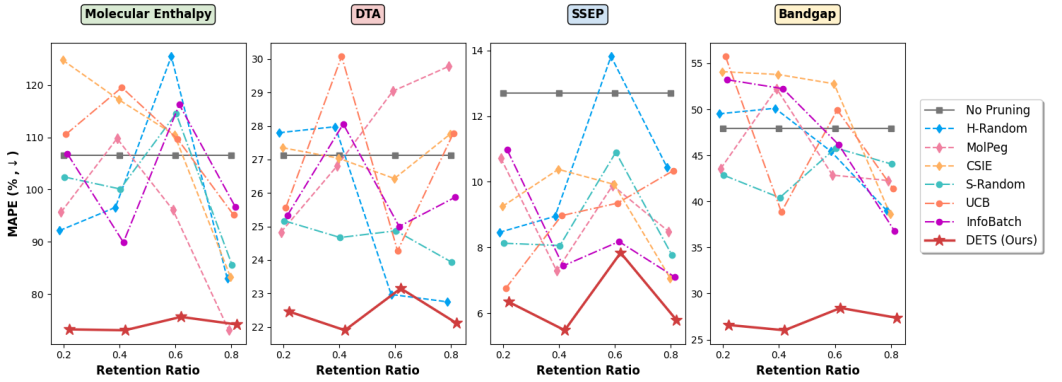
Method	MAE (↓)				RMSE (↓)				MAPE (↓)				Acc@10 (↑)			
	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8	0.2	0.4	0.6	0.8
DETS (Ours)	0.5541	0.5484	**0.5413**	0.5669	0.7385	**0.6885**	0.7219	0.6977	26.58	**26.01**	28.43	27.37	31.98	**32.39**	30.41	31.26
Full dataset	0.7207	0.7207	0.7207	0.7207	0.9566	0.9566	0.9566	0.9566	47.96	47.96	47.96	47.96	22.72	22.72	22.72	22.72
Hard Random	0.71	0.7381	0.636	0.6007	0.9648	0.9768	0.8127	0.7685	49.52	50.12	45.45	39.01	26.70	22.72	23.86	23.86
Herding	0.6099	0.7163	0.6031	0.7637	0.7765	0.9482	0.7465	1.013	43.52	53.58	39.10	57.58	25.00	25.56	24.43	24.43
K-means	0.5864	0.6007	0.7298	0.6427	0.7655	0.7685	0.97	0.8101	44.00	39.01	54.87	49.89	26.70	23.86	25.57	23.86
Entropy	0.7116	0.8164	0.7637	0.6187	0.9559	1.1004	1.013	0.8011	57.19	64.89	57.58	52.80	24.43	19.88	24.43	28.41
Least Confidence	0.7479	0.5886	0.7193	0.6024	0.9899	0.7488	0.9339	0.7673	57.57	37.95	53.66	42.91	25.56	27.27	21.59	28.40
GraNd-20	0.8229	0.761	0.6049	0.7351	1.0849	0.9696	0.7783	0.9503	60.30	67.87	43.11	56.31	19.88	22.15	29.55	16.47
Glister	0.5986	0.6668	0.6627	0.7547	0.7427	0.8255	0.8323	1.0463	36.83	51.48	55.73	68.61	26.14	18.18	18.75	23.29
Influence	0.6077	0.6031	0.5734	0.6328	0.7655	0.7681	0.7656	0.8241	42.71	47.87	43.49	56.06	22.72	27.84	30.68	27.27
EL2N-2	0.7547	0.6971	0.5859	0.66	0.9538	0.9127	0.7643	0.8534	55.82	51.09	41.82	43.52	17.04	25.00	30.11	22.72
DP	0.5856	0.6599	0.663	0.6133	0.7447	0.87	0.8429	0.8245	38.87	47.77	45.10	56.92	26.70	29.54	22.15	28.40
Soft Random	0.6573	0.6125	0.6616	0.6081	0.8564	0.7421	0.8588	0.7795	42.84	40.41	45.81	44.06	22.15	19.32	25.56	26.70
ϵ-greedy	0.6482	0.6077	0.5844	0.5711	0.8197	0.7693	0.749	0.7244	53.99	39.09	42.55	40.54	22.16	27.84	27.27	26.13
UCB	0.6381	0.6073	0.5842	0.6115	0.8471	0.7389	0.7653	0.7607	55.76	38.85	49.94	41.38	25.56	21.59	30.11	26.70
InfoBatch	0.7118	0.6343	0.5991	0.5724	0.9154	0.7986	0.7643	0.7278	53.22	52.25	46.14	36.76	18.18	23.29	31.25	30.68
MolPeg	0.5847	0.6343	0.5872	0.5812	0.7472	0.7986	0.7403	0.7581	43.56	52.25	42.86	42.27	25.56	23.29	25.56	26.70
CSIE	0.7159	0.7141	0.6970	0.5917	0.9329	0.9258	0.9204	0.7614	54.10	53.79	52.78	38.67	21.59	21.02	26.13	26.14



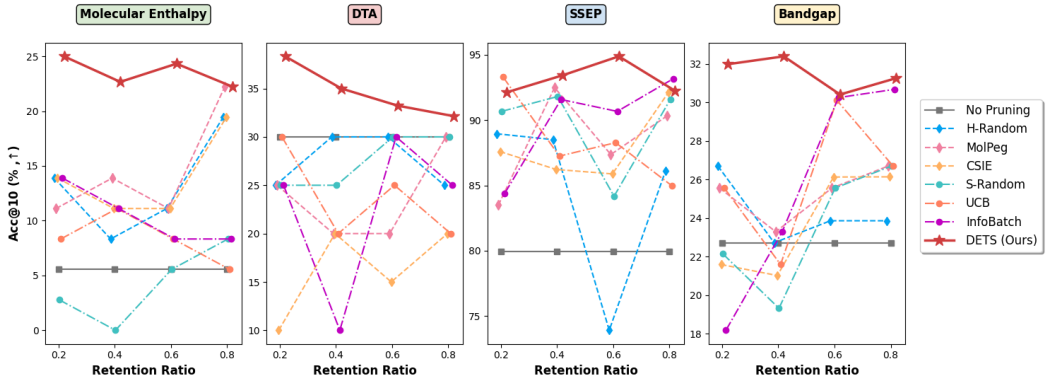
(a) MAE trends across different retention ratios.



(b) RMSE trends across different retention ratios.



(c) MAPE trends across different retention ratios.



(d) Acc@10% trends across different retention ratios.

Figure 4: **Performance across retention ratios.** The panels show (a) MAE, (b) RMSE, (c) MAPE, and (d) Acc@10% for representative methods on all tasks. DETS is shown in red.

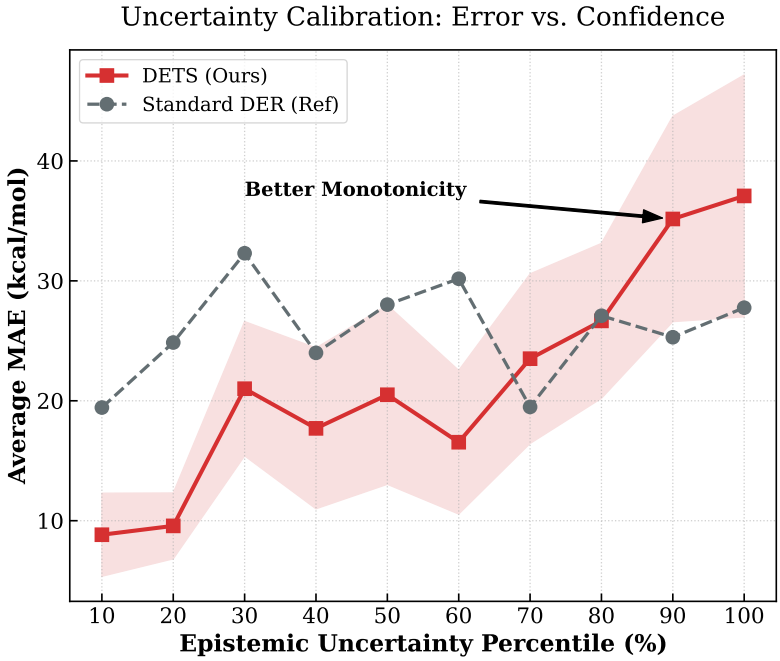


Figure 5: **Epistemic uncertainty and prediction error.** Target samples are grouped by estimated epistemic-uncertainty decile. The y-axis reports the mean MAE in each group; increasing MAE across deciles indicates agreement between the uncertainty ranking and prediction error.

Figure 5 shows a near-monotonic increase in MAE with the estimated uncertainty decile. In this experiment, the ICEE estimate ranks prediction errors more consistently than standard DER.

D.4 Comparison with Uncertainty Estimation Baselines

We further compare DETS with two widely used uncertainty estimation baselines: MC-Dropout and Deep Ensemble. Both baselines use the same backbone as DETS. For MC-Dropout, we perform $M = 5$ stochastic forward passes at inference time. For Deep Ensemble, we train $M = 5$ independently initialized models. On the molecular enthalpy task, Table 12 shows lower prediction errors for DETS and a shorter runtime than Deep Ensemble.

Table 12: Comparison with uncertainty estimation baselines on the molecular enthalpy task.

Method	**MAE↓**	**RMSE↓**	**MAPE↓**	**Acc@10↑**	**Time (s)**	**Memory (MB)**
Full dataset	34.59	44.11	106.54	5.56	5.57	57.71
MC-Dropout	28.66	37.95	95.30	11.11	24.90	62.08
Deep Ensemble	25.27	38.01	93.17	16.67	28.69	82.56
DETS	**23.11**	**25.71**	**73.07**	**25.00**	12.61	61.05

Table 13 extends the comparison to DTA, solvation, and bandgap. All methods use the same backbone and data split. Time is the three-seed mean of complete training and final uncertainty inference, and memory is the maximum `torch.cuda.max_memory_allocated` value across the three runs.

Table 13: Uncertainty-estimation baselines on the three additional tasks.

Task	Method	MAE↓	Acc@10↑	Time (s)	Memory (MB)
DTA	DETS	**1.2734**	**38.33**	23.16	21.30
	MC-Dropout	1.4344	25.38	17.78	21.29
	Deep Ensemble	1.4136	28.80	47.51	24.07
Solvation	DETS	**0.2258**	**94.87**	30.37	22.56
	MC-Dropout	0.3187	88.14	14.02	22.55
	Deep Ensemble	0.3039	89.02	65.24	26.70
Bandgap	DETS	**0.5413**	**32.39**	10.47	20.50
	MC-Dropout	0.6341	26.14	6.17	20.48
	Deep Ensemble	0.7144	27.65	21.61	22.39

D.5 Improvement over the Second-Best Baseline

Table 14 compares DETS with the second-best baseline for each task and metric. DETS has lower MAE and higher tolerance accuracy across the four tasks.

Table 14: Performance improvement of DETS over the second-best baseline in each scenario.

Scenario	Enthalpy MAE↓	Enthalpy Acc@10↑	DTA MAE↓	DTA Acc@10↑	Bandgap MAE↓	Bandgap Acc@10↑	Solvation MAE↓	Solvation Acc@10↑
DETS	23.11	25.00	1.2734	38.33	0.5413	32.39	0.2258	94.87
Second-best	23.68	16.67	1.3076	31.67	0.5711	27.84	0.2513	93.28
Improvement	2.47%	8.33 pp	2.68%	6.66 pp	5.51%	4.55 pp	11.29%	1.59 pp

D.6 Target-only ICEE Ablation

To isolate the effect of ICEE from the effect of source-domain sampling, we evaluate a target-only variant that trains only on the experimental target domain without using any source data. Table 15 shows lower MAE and higher Acc@10 for target-only ICEE than for target-only MSE. Adding uncertainty-guided source sampling gives the best result in this ablation.

Table 15: Ablation separating the effect of ICEE from source-domain sampling on the molecular enthalpy task.

Method	MAE↓	Acc@10↑
Target-domain MSE only	26.52	8.33
Target-domain ICEE only	24.27	13.89
DETS: ICEE + Sampling	**23.11**	**25.00**

D.7 Early-stage Sampling Comparison

We compare DETS with Hard Random and Least Confidence after 10 training iterations. Table 16 reports lower MAE and higher Acc@10 for DETS at this stage.

Table 16: Early-stage comparison after 10 training iterations.

Method	MAE↓	Acc@10↑
Hard Random	30.20	3.03
Least Confidence	30.22	2.73
DETS	**29.32**	**4.24**

D.8 Robustness to the Choice of Clustering Algorithm

DETS clusters continuous encoder representations. We replace K-Means with DBSCAN in the same latent space on the molecular enthalpy task. Table 17 shows similar results, with K-Means performing slightly better on all four metrics.

Table 17: Robustness of DETS to the clustering algorithm on the molecular enthalpy task.

Clustering Method	MAE↓	RMSE↓	MAPE↓	Acc@10↑
K-Means	**23.11**	**25.71**	**73.07**	**25.00**
DBSCAN	23.43	26.08	75.81	24.11

We repeat the replacement on DTA with the target split, ICEE loss, soft-Tanimoto weights, training epochs, and other hyperparameters fixed. Table 18 shows a small MAE difference and a larger Acc@10 difference in favor of K-Means.

Table 18: Clustering-algorithm robustness on DTA.

Clustering Method	MAE↓	Acc@10↑
K-Means	**1.2734**	**38.33**
DBSCAN	1.2796	35.51

D.9 Activity-cliff Evaluation

An activity-cliff pair is defined as two compounds from the same protein with different SMILES, Morgan fingerprint similarity of at least 0.70, and an absolute pKd difference of at least 1.0. We use radius-2, 256-bit Morgan fingerprints. This definition yields 299 cliff pairs, 588 unique endpoints, and 266 proteins. For each seed, 20% of proteins are held out by `PROTEIN_ID`, so test proteins do not appear among the source candidates. Table 19 reports the three-seed mean on the cliff endpoints; each metric independently uses the best candidate retention ratio.

Table 19: Protein-disjoint activity-cliff results on Davis.

Method	Cliff MAE↓	Cliff RMSE↓	Cliff Acc@10↑
Full-source training	1.9442	2.1519	15.40
Random sampling	3.1451	3.3211	4.15
EL2N-2	2.1314	2.3567	14.24
DETS	**1.2634**	**1.8223**	**28.77**

D.10 Controlled Structural-shift Evaluation

We use the paired bandgap PBE/GW data to evaluate controlled levels of structural shift. Source and target materials are separated by material ID, and fixed composition descriptors are clustered to adjust cluster and element coverage. The PBE $\rightarrow$ GW protocol, sample size, and training budget remain fixed. Distance measures the distributional difference between source and target materials in the predefined composition-descriptor space. Table 20 shows that DETS has lower MAE than full-source training at all four shift levels.

Table 20: Bandgap results under controlled structural shift.

Shift	Distance	Full-source MAE↓	DETS MAE↓
None	5.4906	0.6831	**0.5413**
Weak	6.8879	0.7211	**0.5598**
Moderate	8.1262	0.6712	**0.5613**
Strong	11.7359	0.7067	**0.5651**

E Mathematical Derivations

This section derives the Student-t predictive distribution, the NIG uncertainty decomposition, and the interval-censored likelihood used by DETS.

E.1 Derivation of the Student-t Posterior Predictive Distribution

Eq. 8 follows by marginalizing a Gaussian likelihood under a Normal-Inverse-Gamma (NIG) prior.

Setup. Let the target y follow a Gaussian distribution with unknown mean μ and variance σ^2:

$$y \sim \mathcal{N}(\mu, \sigma^2). \tag{24}$$

We place a conjugate Normal-Inverse-Gamma (NIG) prior on the parameters μ and σ^2:

$$p(\mu, \sigma^2|\mathbf{m}) = \text{NIG}(\mu, \sigma^2|\gamma, \nu, \alpha, \beta) = \mathcal{N}(\mu|\gamma, \sigma^2\nu^{-1}) \cdot \text{Inv-Gamma}(\sigma^2|\alpha, \beta). \tag{25}$$

The posterior predictive distribution marginalizes the latent parameters μ and σ^2:

$$p(y|\mathbf{m}) = \int_0^\infty \int_{-\infty}^\infty p(y|\mu, \sigma^2) \cdot p(\mu, \sigma^2|\mathbf{m})\, d\mu\, d\sigma^2. \tag{26}$$

First, we integrate over μ. Since both the likelihood $p(y|\mu, \sigma^2)$ and the conditional prior $p(\mu|\sigma^2)$ are Gaussian, their convolution is analytically tractable:

$$\begin{aligned} p(y|\sigma^2) &= \int_{-\infty}^\infty \mathcal{N}(y|\mu, \sigma^2) \cdot \mathcal{N}\left(\mu|\gamma, \frac{\sigma^2}{\nu}\right) d\mu \\ &= \mathcal{N}\left(y\middle|\gamma, \sigma^2 + \frac{\sigma^2}{\nu}\right) = \mathcal{N}\left(y\middle|\gamma, \sigma^2\frac{1+\nu}{\nu}\right). \end{aligned} \tag{27}$$

Next, we integrate over σ^2 using the Inverse-Gamma prior. Substituting the probability density functions (PDFs) into the integral:

$$p(y|\mathbf{m}) = \int_0^\infty \underbrace{\frac{1}{\sqrt{2\pi\sigma^2\frac{1+\nu}{\nu}}} \exp\left(-\frac{(y-\gamma)^2}{2\sigma^2\frac{1+\nu}{\nu}}\right)}_{\mathcal{N}(y|\dots)} \cdot \underbrace{\frac{\beta^\alpha}{\Gamma(\alpha)}(\sigma^2)^{-\alpha-1}\exp\left(-\frac{\beta}{\sigma^2}\right)}_{\text{Inv-Gamma}(\sigma^2|\dots)} d\sigma^2. \tag{28}$$

Rearranging the terms to group those involving σ^2:

$$p(y|\mathbf{m}) \propto \int_0^\infty (\sigma^2)^{-(\alpha+\frac{1}{2})-1} \exp\left(-\frac{1}{\sigma^2}\left[\beta + \frac{\nu(y-\gamma)^2}{2(1+\nu)}\right]\right) d\sigma^2. \tag{29}$$

The integrand is an unnormalized Inverse-Gamma kernel, giving

$$p(y|\mathbf{m}) \propto \left(\beta + \frac{\nu(y-\gamma)^2}{2(1+\nu)}\right)^{-(\alpha+\frac{1}{2})}. \tag{30}$$

To match the standard Student-t density form $p(t) \propto (1 + \frac{1}{df}(\frac{t-\text{loc}}{\text{scale}})^2)^{-\frac{df+1}{2}}$, we factor out β:

$$p(y|\mathbf{m}) \propto \left(1 + \frac{\nu(y-\gamma)^2}{2\beta(1+\nu)}\right)^{-(\alpha+\frac{1}{2})}. \tag{31}$$

Matching terms gives degrees of freedom $df = 2\alpha$, location γ, and squared scale $\sigma_{st}^2 = \frac{\beta(1+\nu)}{\nu\alpha}$. The predictive distribution is

$$p(y|\mathbf{m}) = \text{St}\left(y; \text{loc} = \gamma, \text{scale} = \sqrt{\frac{\beta(1+\nu)}{\nu\alpha}}, \text{df} = 2\alpha\right). \tag{32}$$

The term $\Omega = 2\beta(1+\nu)$ used in the NLL loss (Eq. 9 in the main text) relates to our derived scale parameter by $\Omega = 2\nu\alpha\sigma_{st}^2$.

E.2 Derivation of Uncertainty Decomposition

In Eq. 14, we decomposed the total predictive variance into aleatoric and epistemic components using the *law of total variance*:

$$\underbrace{\text{Var}[y]}_{\text{Total}} = \underbrace{\mathbb{E}_{\mu,\sigma^2}[\text{Var}[y|\mu,\sigma^2]]}_{\text{Aleatoric}} + \underbrace{\text{Var}_{\mu,\sigma^2}[\mathbb{E}[y|\mu,\sigma^2]]}_{\text{Epistemic}}. \tag{33}$$

Aleatoric component. The conditional variance is $\text{Var}[y|\mu,\sigma^2] = \sigma^2$. Its expectation under the Inverse-Gamma prior is

$$\mathbb{E}[\sigma^2] = \text{Mean of Inv-Gamma}(\alpha,\beta) = \frac{\beta}{\alpha-1}, \quad \text{for } \alpha > 1. \tag{34}$$

Epistemic component. The conditional expectation is $\mathbb{E}[y|\mu,\sigma^2] = \mu$, so the epistemic term is the variance of μ under the NIG prior. Applying the law of total variance to μ gives

$$\begin{aligned}
\text{Var}[\mu] &= \mathbb{E}_{\sigma^2}[\text{Var}[\mu|\sigma^2]] + \text{Var}_{\sigma^2}[\mathbb{E}[\mu|\sigma^2]] \\
&= \mathbb{E}_{\sigma^2}\left[\frac{\sigma^2}{\nu}\right] + \text{Var}_{\sigma^2}[\gamma] \quad (\text{since } \mu|\sigma^2 \sim \mathcal{N}(\gamma, \sigma^2/\nu)) \\
&= \frac{1}{\nu}\mathbb{E}[\sigma^2] + 0 \\
&= \frac{1}{\nu}\cdot\frac{\beta}{\alpha-1} = \frac{\beta}{\nu(\alpha-1)}.
\end{aligned} \tag{35}$$

The total predictive variance is therefore

$$\text{Var}[y] = \underbrace{\frac{\beta}{\alpha-1}}_{\text{Aleatoric}} + \underbrace{\frac{\beta}{\nu(\alpha-1)}}_{\text{Epistemic}}. \tag{36}$$

The two terms have distinct interpretations and can be computed separately from the NIG parameters. They are not statistically independent because both depend on α and β. ICEE changes their estimates through the tolerance-aware training objective, which reduces the influence of within-tolerance residuals on the learned evidence. DETS uses $U_{epi} = \frac{\beta}{\nu(\alpha-1)}$ as its source-sampling signal. The variance of a Student-t distribution is given by $\text{scale}^2 \cdot \frac{df}{df-2}$. Substituting our derived parameters:

$$\text{Var}[y] = \frac{\beta(1+\nu)}{\nu\alpha}\cdot\frac{2\alpha}{2\alpha-2} = \frac{\beta(1+\nu)}{\nu(\alpha-1)}. \tag{37}$$

This matches the sum in Eq. 36, $\frac{\beta}{\alpha-1} + \frac{\beta}{\nu(\alpha-1)}$.

E.3 Integration of Interval-Censored Likelihood (ICL)

ICEE maximizes predictive probability mass within the tolerance interval $I_i = [L_i, U_i]$.

For the Student-t predictive distribution $p(y|\mathbf{m})$, the probability assigned to I_i is

$$P(Y \in I_i|\mathbf{m}) = \int_{L_i}^{U_i} \text{St}(t;\gamma,\sigma_{st}^2,2\alpha)\,dt. \tag{38}$$

With the cumulative distribution function $\Phi(x) = \int_{-\infty}^{x} p(t)\,dt$, this probability is the difference of two CDF values:

$$\int_{L_i}^{U_i} p(t)dt = \Phi_{2\alpha}\left(\frac{U_i-\gamma}{\sigma_{st}}\right) - \Phi_{2\alpha}\left(\frac{L_i-\gamma}{\sigma_{st}}\right), \tag{39}$$

where $\Phi_{2\alpha}$ is the standard Student-t CDF with 2α degrees of freedom, and σ_{st} is the scale parameter derived in Appendix E.1.

Substituting $Z_{upper} = \frac{U_i-\gamma}{\sigma_{st}}$ and $Z_{lower} = \frac{L_i-\gamma}{\sigma_{st}}$, we obtain the objective function maximized by DETS:

$$\mathcal{L}_{ICL} = -\log\left(\Phi_{2\alpha}(Z_{upper}) - \Phi_{2\alpha}(Z_{lower}) + \epsilon_{\text{num}}\right). \tag{40}$$

The loss depends on the total probability mass inside $[L_i, U_i]$, reducing the penalty for residuals that remain within the tolerance interval.

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state the main contributions of DETS, including interval-censored evidential learning, uncertainty-guided source sampling, and evaluation on cross-domain scientific regression tasks. The claims are supported by the main experiments in Tables 1–3 and additional results in Appendix D.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: We discuss limitations in Section 4, including the dependence on meaningful tolerance thresholds, the assumption of partial source-target overlap, and the current focus on regression tasks in chemistry and materials science.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [Yes]

Justification: The probabilistic assumptions are stated in the methodology, and the derivations of the Student-t posterior predictive distribution, uncertainty decomposition, and interval-censored likelihood are provided in Appendix E.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: We describe the datasets, preprocessing, model architectures, baselines, training protocol, hyperparameters, and evaluation metrics in Sections 3 and Appendix B – C. The anonymized code repository is also provided for reproducing the experiments.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).

(d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: We provide an anonymized code repository in the abstract. The datasets used in the experiments are public datasets and are cited in the paper, with preprocessing details described in Appendix B.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 3.1 describes the experimental setup, including datasets, backbones, optimizers, learning rates, baselines, and evaluation metrics. Appendix C provides additional implementation details and hyperparameter configurations.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: Main results are reported as the mean and standard deviation over three independent runs. The tables include error bars for the main comparisons, and the text specifies that the reported variability corresponds to independent experimental runs.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: We report the GPU type used for experiments in Section 3.1 and provide additional computational efficiency analysis in Appendix D.2, including wall-clock overhead and comparison with more expensive selection methods.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The work uses public scientific datasets and does not involve human subjects, private data, or safety-sensitive personal information. We have reviewed the NeurIPS Code of Ethics and do not identify any violation.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: We discuss broader impacts in the Appendix 5 including the potential to reduce experimental cost and improve scientific data efficiency, as well as the need for experimental validation in safety-critical applications.

    Guidelines:

    - The answer [N/A] means that there is no societal impact of the work performed.
    - If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

    Answer: [N/A]

    Justification: The work does not release high-risk models, scraped datasets, personal data, or generative systems that require special safeguards.

    Guidelines:

    - The answer [N/A] means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [Yes]

Justification: All existing datasets and baseline methods used in this work are cited in the paper. The experiments are based on publicly available scientific datasets, and we do not claim ownership of these external assets.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

Answer: [Yes]

Justification: We release an anonymized code repository containing the implementation of DETS, training scripts, and instructions for reproducing the main experiments. No new dataset is introduced.

Guidelines:

- The answer [N/A] means that the paper does not release new assets.
- Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
- The paper should discuss whether and how consent was obtained from people whose asset is used.
- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

Answer: [N/A]

Justification: This work does not involve crowdsourcing, user studies, or research with human subjects.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
- According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [N/A]

    Justification: This work does not involve human subjects, so IRB approval is not required.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

    Answer: [N/A]

    Justification: LLMs are not used as an important, original, or non-standard component of the proposed method, experiments, or scientific claims.

    Guidelines:

    - The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.