MDG-IS: AN ITERATIVE SAMPLING FRAMEWORK FOR THEORY-EXPERIMENT DATA ALIGNMENT IN ARTIFICIAL INTELLIGENCE CHEMISTRY

Anonymous authors
Paper under double-blind review

ABSTRACT

AI4Chemistry faces a fundamental bottlewneck: high-fidelity experimental data are scarce and heterogeneous, while large low-fidelity theoretical datasets are abundant but subdomain-biased. This mismatch leads to distribution misalignment and frequent negative transfer, limiting the effectiveness of existing transfer learning and multi-fidelity approaches that assume global theory–experiment consistency. We propose **M**ulti-Sub**D**omain **G**MM-guided **I**terative **S**ampling (**MDG-IS**), a fidelity-aware data selection framework that identifies which theoretical samples are actually useful for learning from real experimental data. MDG-IS models experimental subdomains through a Gaussian Mixture Model and performs local KL-based alignment to iteratively extract a compact, well-aligned subset of theoretical data. Across molecular, reaction, and materials benchmarks, MDG-IS markedly reduces computational cost, mitigates negative transfer, and consistently improves prediction accuracy in real scenarios.

1 INTRODUCTION

Artificial intelligence is rapidly transforming the discovery of molecules and materials by closely integrating machine learning with computational and experimental chemistry Goh et al. (2017); Divya et al. (2024); Rial (2024); Ananikov (2024a), particularly in the areas of molecular property prediction Walters & 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, despite these achievements, the performance and reliability of AI4Chemistry models remain fundamentally constrained by the quality of the data: models require large datasets for generalization, but the only data available at scale deviate substantially from the target experimental distribution. McCready et al. (2006); Du et al. (2025); Ananikov (2024b).

High-fidelity experimental measurements—whether obtained through state-of-the-art quantum chemical calculations, precise spectroscopy, or rigorously controlled laboratory experiments—are scarce and costly. In contrast, low-fidelity theoretical data generated from approximate quantum chemistry, large-scale molecular dynamics, 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). This imbalance has become a critical bottleneck limiting the scalability, robustness, and generalizability of AI-driven chemical modeling Kim et al. (2024); Vinod & Zaspel (2025); Gardner et al. (2025).

To address this imbalance, several strategies have emerged. Transfer learning leverages large theoretical datasets for pre-training and then fine-tunes on scarce experimental data, which is effective when the mapping relationship between upstream and downstream data is stable Goh et al. (2018); Li & Jiang (2021); Wu et al. (2024b); Vermeire & Green (2021); Buterez et al. (2024). Multi-fidelity learning jointly models data from multiple sources, assuming that low-fidelity and high-fidelity measurements share compatible structural relationships Ramakrishnan et al. (2015); Fare et al. (2022). Data augmentation expands the coverage of experimental data through synthetic generation or more chemically aware enhancements You et al. (2020); Xia et al. (2023); Wang et al. (2023). Overall, these approaches perform well in contexts where theoretical data serve as a proxy for global information, such as small, rigid molecules or systems with consistent electronic structures. In these scenarios, the

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

theoretical and experimental distributions are sufficiently aligned, allowing these methods to extract transferable knowledge Vermeire & Green (2021); Chung & Green (2024); Gardner et al. (2025); Shiota et al. (2024).

However, real-world scientific repositories rarely satisfy the assumptions underlying existing methods. Instead, they present three key challenges. First, theoretical data are generated under diverse computational protocols, chemical families, and reaction mechanisms, resulting in subdomain-dependent systematic biases. Consequently, samples that are highly informative in one chemical subdomain may become misleading in another Griffiths et al. (2021); Weymuth & Reiher (2022). Second, experimental observations are extremely scarce and sparsely distributed across chemical subdomains. Their limited coverage makes global alignment unreliable and local distribution estimation substantially more challenging Rosser et al. (2024). Third, blindly utilizing ever-growing theoretical repositories substantially increases computational cost while introducing noisy or weakly relevant supervision, often causing negative transfer rather than performance gains Qin et al. (2024); Chen et al. (2024); Na (2025). Recent studies further suggest that carefully selected subsets can outperform substantially larger but weakly aligned datasets Rong et al. (2020); Xia et al. (2023); Luong & Singh (2023); Ji et al. (2024).

These observations suggest that the key challenge is not how to leverage more theoretical data, but how to identify a compact, representative subset that is well aligned with the target experimental distribution. This problem fundamentally differs from transfer learning and multi-fidelity learning because the objective is no longer representation learning, but target-aware data selection.

Thus, we introduce the **M**ulti-Sub**D**omain **G**MM-guided **I**terative **S**ampling **(MDG-IS)** framework, a target-domain-aware strategy for selecting informative theoretical data under distributional heterogeneity. MDG-IS models scarce experimental data with a Gaussian Mixture Model (GMM) Rasmussen (1999) to identify heterogeneous chemical subdomains, and evaluates theoretical samples through local KL-divergence alignment. An iterative selection-and-pruning procedure progressively retains aligned samples and removes misleading regions, with the subset size automatically determined by convergence. This yields a compact, subdomain-consistent coreset that reduces negative transfer and computational overhead while remaining tailored to the experimental domain. A concurrent work in the supplementary material instead performs theoretical-domain selection based on label tolerance intervals to decouple accidental noise from cognitive uncertainty.

Our contributions can be summarized as follows:

- **Fidelity-aware and efficient data selection.** MDG-IS filters massive theoretical datasets to retain distributionally aligned samples, reducing storage and training costs.
- **Subdomain-aware alignment.** GMM-based modeling and local KL divergence mitigate heterogeneity bias and negative transfer by filtering unreliable regions.
- **Broad applicability.** Experiments across molecules, reactions, and materials demonstrate consistent gains under diverse fidelity gaps and data regimes.

2 METHODOLOGY

2.1 PROBLEM FORMULATION

Task definition. Under conditions of limited and heterogeneous experimental data$\mathcal{T}$, we aim to select a minimal yet most informative subset $\mathcal{C}$ from the vast theoretical data $\mathcal{S}$ exhibits consistent systematic biases in the lungs.

$$\mathcal{C} \subseteq \mathcal{S}, \tag{1}$$

This ensures that the induced empirical distribution can closely approximate the latent structure of the target experimental domain.

Feature Representation and Latent Alignment. The theoretical domain and the target domain are respectively

$$\mathcal{S} = \{x_i^{(s)}\}_{i=1}^{N_s}, \qquad \mathcal{T} = \{x_j^{(t)}\}_{j=1}^{N_t}, \tag{2}$$

We use a pre-trained proxy model $f_{\text{pre}}$to extract feature representations:

$$h_i = f_{\text{pre}}(x_i). \tag{3}$$

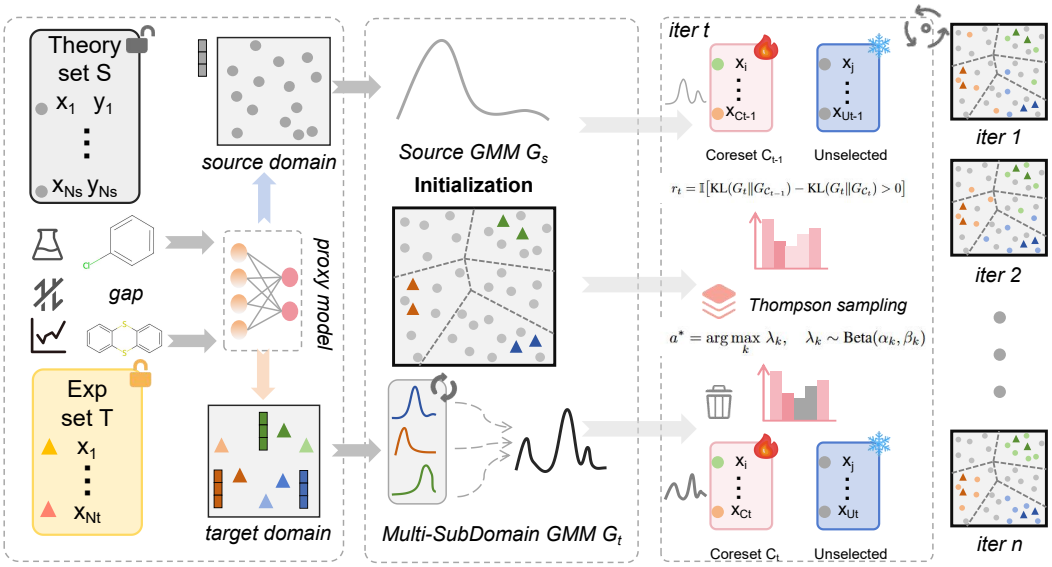


Figure 1: Model framework: Extract features of $\mathcal{S}$ and $\mathcal{T}$ and perform GMM modeling, then according to Thompson sampling method, iteratively update and delete the core set $\mathcal{C}$ aligned with $\mathcal{T}$.

For multi entity samples, dot product aggregation is used to obtain the overall representation.

Due to the scarcity of target domain samples and significant differences in noise from various sources. Directly estimating GMM in high-dimensional feature space can lead to problems such as singular covariance, which in turn affects the reliable characterization of the target domain structure. PCA Maćkiewicz & Ratajczak (1993) can effectively compress dimensions and improve the numerical stability of GMM parameter estimation. In order to facilitate comparable density estimation between the theoretical domain and the target domain, we perform PCA on $H_s \cup H_t$to a shared latent space, denoted as Z_s, Z_t.

Distributional Objective. We use KL divergence metric Hershey & Olsen (2007) to measure the difference between the mixture Gaussian distribution $G_{\mathcal{C}}$induced by subsets and the GMM distribution $G_t$in the target domain:

$$\mathrm{KL}(G_t \| G_{\mathcal{C}}) = \mathbb{E}_{z \sim G_t}\left[\log \frac{G_t(z)}{G_{\mathcal{C}}(z)}\right]. \tag{4}$$

Therefore, our goal is to select the smallest theoretical subset that has the highest logarithmic likelihood on the target domain samples. These observations share a common motivation: we need a structure aware target domain density model to reliably distinguish between useful and misleading theoretical samples.

2.2 The modules of MDG-IS

MDG-IS consists of three core modules: (1) Modeling of soft subdomain structure for target domain; (2) Data selection based on KL iterative theory; (3) Prune the structural consistency of the selected subset. This process constitutes a closed-loop mechanism that aligns theory with experiment.

2.3 Multi-SubDomain GMM

Given the potential representation $Z_t = \{z_j\}$in the target domain, we construct a Gaussian mixture model with a soft subdomain structure.

The goal is to construct a structure-guided mixture model that captures both global coverage and local coherence. To this end, we progressively establish the following parameter hierarchy.

Local affinity and density estimation. We first estimate pairwise local affinities among target embeddings to characterize the neighborhood structure:

$$A_{ij} = \exp\left(-\frac{\|z_i - z_j\|^2}{\tau^2}\right), \rho_j = \sum_{i=1}^{N_t} A_{ij}. \tag{5}$$

where τ controls the sensitivity to local scale. The local density of each sample is then computed as ρ_j serves as a continuous density indicator derived from the affinity matrix A_{ij}; points with larger ρ_j correspond to denser regions in the target manifold.

Anchor identification. Based on the density map $\{\rho_j\}$, we identify M representative subdomain anchors:

$$c_m = \arg\max_j \rho_j, \qquad m = 1, \ldots, M. \tag{6}$$

each anchor c_m thus depends implicitly on the local density field determined by A_{ij}. These anchors serve as coarse prototypes of subdomains and will guide subsequent responsibility assignment.

Continuous subdomain responsibilities. Given anchors $\{c_m\}$, each target sample z_j is softly assigned to all subdomains through a Gaussian kernel Keerthi & Lin (2003) of local bandwidth σ_m:

$$\gamma_{jm} = \frac{\exp\big(-\|z_j - c_m\|^2/\sigma_m^2\big)}{\sum_{m'=1}^{M} \exp\big(-\|z_j - c_{m'}\|^2/\sigma_{m'}^2\big)}, \tag{7}$$

$\gamma_{jm} \in (0, 1)$ is a deterministic function of the sample position z_j, the anchor set $\{c_m\}$, and the local spreads $\{\sigma_m\}$, which themselves are derived from the affinity-induced densities ρ_j. The soft responsibility γ_{jm} measures the degree to which sample z_j belongs to subdomain m.

The global subdomain prior weight is defined as the normalized aggregate responsibility:

$$w_m = \frac{\sum_{j=1}^{N_t} \gamma_{jm}}{\sum_{m'=1}^{M} \sum_{j=1}^{N_t} \gamma_{jm'}}. \tag{8}$$

Subdomain-aware GMM formulation. With the soft structural prior $\{\gamma_{jm}, w_m\}$ fixed, we define within each subdomain m a local Gaussian mixture:

$$p_m(z_j) = \sum_{k=1}^{K_m} \pi_{mk}\, \mathcal{N}(z_j \mid \mu_{mk}, \Sigma_{mk}), \tag{9}$$

where π_{mk}, μ_{mk}, and Σ_{mk} are the internal mixture coefficients, mean, and covariance of subdomain m.

Given γ_{jm} as a structural prior, the posterior responsibility of sample z_j to component (m, k) becomes

$$r_{j,mk} = \frac{\gamma_{jm}\, \pi_{mk}\, \mathcal{N}(z_j \mid \mu_{mk}, \Sigma_{mk})}{\sum_{m'=1}^{M} \sum_{k'=1}^{K_{m'}} \gamma_{jm'}\, \pi_{m'k'}\, \mathcal{N}(z_j \mid \mu_{m'k'}, \Sigma_{m'k'})}. \tag{10}$$

which is used to update the GMM parameters.

Global target distribution. The complete target density model is expressed as

$$G_t(z_j) = \sum_{m=1}^{M} w_m\, p_m(z_j). \tag{11}$$

where each w_m reflects the prior mass of subdomain m estimated from soft responsibilities.

2.4 KL-Guided Iterative Coreset Selection

To select an informative subset from the source domain that aligns with the target distribution, let $X_s = \{x_i\}_{i=1}^{N_s}$ denote source-domain samples. We aim to select a compact subset $\mathcal{C} \subseteq X_s$ whose $G_{\mathcal{C}}$ best matches G_t.

KL Divergence as Objective The KL divergence between the target model and the Coreset model is approximated by Monte-Carlo James (1980):

$$\begin{aligned} \mathrm{KL}(G_t \| G_{\mathcal{C}}) &= \mathbb{E}_{z\sim G_t}\left[\log \frac{G_t(z)}{G_{\mathcal{C}}(z)}\right] \\ &\approx \frac{1}{N_t}\sum_{j=1}^{N_t}\left[\log G_t(z_j) - \log G_{\mathcal{C}}(z_j)\right], \end{aligned} \tag{12}$$

this KL divergence quantifies distribution mismatch and drives the selection procedure.

Bandit Formulation. We first cluster the source domain into K groups via GMM:

$$C_k = \{i \mid x_i \text{belongs to cluster } k\}, \tag{13}$$

each cluster C_k is treated as an arm of a multi-armed bandit Slivkins et al. (2019).At each iteration, we employ a stochastic multi-armed bandit mechanism: each arm corresponds to a source cluster, and its reward is determined by whether the inclusion of its samples reduces the KL divergence between G_t and $G_{\mathcal{C}}$:

$$r_t = \mathbb{I}\left[\mathrm{KL}(G_t \| G_{\mathcal{C}_{t-1}}) - \mathrm{KL}(G_t \| G_{\mathcal{C}_t}) > 0\right], \tag{14}$$

we assign reward $r_t = 1$, otherwise $r_t = 0$. r_t indicates whether sampling from cluster k benefits global alignment.

The posterior parameters (α_k, β_k) of each arm are updated:

$$\alpha_k \leftarrow \alpha_k + r_t, \quad \beta_k \leftarrow \beta_k + (1 - r_t), \tag{15}$$

The next arm is chosen via Thompson sampling:

$$a^* = \arg\max_k \lambda_k, \quad \lambda_k \sim \mathrm{Beta}(\alpha_k, \beta_k). \tag{16}$$

This adaptive scheme naturally balances exploration and exploitation, progressively refining the selected subset toward optimal distributional alignment.

Consistency of alignment and pruning. To further prevent overfitting caused by rare subspace over-coverage, we introduce an alignment-based pruning mechanism that removes inconsistent samples from the coreset. For each sample x, we compute an alignment deviation score

$$\Delta(x) = \lambda \log G_s(x) + \log G_t(x), \tag{17}$$

and exclude samples whose $\Delta(x)$ exceeds one standard deviation above the mean, thus maintaining the consistency at the subspace-level and improving the learnability of the final model.

Termination criterion. The iterative process terminates once the relative improvement in KL divergence falls below a small tolerance ε, i.e.,

$$\left|\mathrm{KL}_t - \mathrm{KL}_{t-1}\right| < \varepsilon. \tag{18}$$

The final output is the optimized Coreset, $\mathcal{C}^*$ is a compact, noise suppressed theoretical data subset that is consistent with the target experimental domain structure.:

$$\mathcal{C}^* = \arg\min_{\mathcal{C}} \mathrm{KL}(G_t \| G_{\mathcal{C}}). \tag{19}$$

2.5 Coreset Pretraining and Fine-Tuning

Coreset-Guided Pretraining. Given $\mathcal{C}^*$, we pretrain a representation model $f_\theta(\cdot)$ by solving

$$\theta^* = \arg\min_\theta \mathcal{L}_{\mathrm{pre}}(\mathcal{C}^*; \theta), \tag{20}$$

where $\mathcal{L}_{\mathrm{pre}}$ denotes a self-supervised or weakly supervised molecular learning objective. Since $\mathcal{C}^*$ is optimized to minimize divergence from G_t, pretraining on it directly exposes the model to a distributionally aligned and noise-reduced approximation of the target domain, providing a target-aware initialization.

Target-Domain Fine-Tuning. The pretrained model is further adapted to X_t by minimizing

$$\hat{\theta} = \arg\min_\theta \mathcal{L}_{\mathrm{ft}}(X_t; \theta^*), \tag{21}$$

such that the already target-aligned parameters require only small adjustments to fit the experimental data, resulting in improved task performance and better data efficiency. From a distributional point of view, fine-tuning reduces the remaining gap in $\mathrm{KL}(G_t \,\|\, G_{f_\theta})$.

3 EXPERIMENTAL SCENARIO INTRODUCTION

To evaluate cross-domain generalization, we consider four representative tasks.

Thermochemical Property (Molecular enthalpy) Prediction: We use two datasets for Molecular standard enthalpy of formation prediction. **Wudily_Cho** Bruce et al. (2025) contains 6,074 high-level quantum chemical calculations for hydrocarbons, oxygenates, PAHs and radicals, calculated using CBS-QB3, CBS-APNO, G3 and G4 methods. **ATcT** Laboratory (2025) provides 358 high-accuracy experimental gas-phase enthalpies obtained from the Active Thermochemical Tables; SMILES were generated from molecular formulas and removed if $\Delta_f H^\circ > \pm 500$ kJ·mol^{-1}

Drug-target Binding Affinity (DTA) Prediction: Two datasets are used: **Davis** Davis et al. (2011), a low-precision dataset with 24,956 drug-target pairs, where binding affinities (Kd) are measured via kinase inhibition assays and converted to pKd $(5.0–10.8)$; **CASF-2016** Su et al. (2018), a high-precision benchmark of 285 complexes from PDBbind2016 refined set, includes experimentally measured affinities and high-resolution X-ray structures ($\leq 2.5\mathring{A}$).

Gibbs Free Energy (Molecular dissolution) Prediction: We consider **CombiSolv-QM** Vermeire & Green (2021), a theoretical dataset of 1 million solute–solvent pairs with solvation free energies computed via COSMOtherm, and **CombiSolv-Exp** Vermeire & Green (2021), an experimental dataset of 8,780 publicly available measurements collected from MiniSolv, FreeSolv, CompSol, and Abraham databases.

Material Band Gap Prediction: For band gap prediction, **theoretical dataset** National Renewable Energy Laboratory (2024) contains 546 PBE-computed values (low-precision), while **experimental dataset** National Renewable Energy Laboratory (2024) comprises 52 high-accuracy values obtained via GWVD calculations.

Table 1: Experimental results of molecular enthalpy and drug-target binding affinity prediction with baselines under different coreset size. The Full dataset means we use all theoretical data for pretraining

Dataset	Wudily_Cho→ATcT (MAE↓)				Davis→CASF-2016 (MAE↓)			
Backbone	GIN				GIN+Transformer			
Coreset size %	20	40	60	80	20	40	60	80
Static Coreset Selection								
Hard Random	$40.38_{\pm 2.80}$	$38.75_{\pm 1.82}$	$38.63_{\pm 1.93}$	$34.74_{\pm 1.72}$	$2.84_{\pm 0.26}$	$2.53_{\pm 0.25}$	$1.95_{\pm 0.22}$	$1.80_{\pm 0.24}$
Herding	$42.33_{\pm 2.17}$	$41.56_{\pm 1.78}$	$39.12_{\pm 1.44}$	$35.77_{\pm 1.53}$	$2.91_{\pm 0.25}$	$2.12_{\pm 0.20}$	$1.87_{\pm 0.23}$	$1.75_{\pm 0.19}$
K-Means	$39.26_{\pm 2.70}$	$40.15_{\pm 1.63}$	$40.27_{\pm 1.61}$	$36.83_{\pm 1.49}$	$2.70_{\pm 0.23}$	$1.93_{\pm 0.18}$	$1.82_{\pm 0.14}$	$1.76_{\pm 0.16}$
Entropy	$37.63_{\pm 2.03}$	$36.55_{\pm 1.75}$	$34.57_{\pm 1.78}$	$34.62_{\pm 1.31}$	$2.42_{\pm 0.21}$	$1.88_{\pm 0.18}$	$1.87_{\pm 0.17}$	$1.81_{\pm 0.16}$
Least Confidence	$38.19_{\pm 1.98}$	$37.44_{\pm 1.58}$	$34.87_{\pm 1.94}$	$33.96_{\pm 1.64}$	$1.95_{\pm 0.19}$	$1.92_{\pm 0.18}$	$1.73_{\pm 0.16}$	$1.79_{\pm 0.15}$
GraNd-20	$39.12_{\pm 2.12}$	$37.98_{\pm 1.57}$	$36.55_{\pm 1.50}$	$34.73_{\pm 1.42}$	$1.83_{\pm 0.23}$	$1.85_{\pm 0.17}$	$1.74_{\pm 0.16}$	$1.70_{\pm 0.15}$
Glister	$40.37_{\pm 1.80}$	$39.54_{\pm 1.45}$	$37.18_{\pm 1.36}$	$35.23_{\pm 1.58}$	$2.51_{\pm 0.23}$	$2.12_{\pm 0.14}$	$1.89_{\pm 0.21}$	$1.76_{\pm 0.16}$
Influence	$38.90_{\pm 2.12}$	$37.61_{\pm 1.60}$	$36.33_{\pm 1.52}$	$34.18_{\pm 1.45}$	$2.03_{\pm 0.20}$	$1.79_{\pm 0.17}$	$1.76_{\pm 0.13}$	$1.86_{\pm 0.17}$
EL2N-2	$37.56_{\pm 2.58}$	$37.44_{\pm 1.65}$	$35.71_{\pm 1.48}$	$33.63_{\pm 1.37}$	$2.11_{\pm 0.21}$	$1.83_{\pm 0.18}$	$1.86_{\pm 0.17}$	$1.66_{\pm 0.19}$
DP	$38.02_{\pm 1.95}$	$36.89_{\pm 1.77}$	$35.07_{\pm 1.47}$	$33.34_{\pm 1.40}$	$2.52_{\pm 0.24}$	$2.07_{\pm 0.20}$	$1.80_{\pm 0.21}$	$1.69_{\pm 0.14}$
Dynamic Coreset Selection								
Soft Random	$37.55_{\pm 2.60}$	$36.11_{\pm 2.33}$	$36.12_{\pm 1.48}$	$34.76_{\pm 1.43}$	$2.61_{\pm 0.24}$	$1.99_{\pm 0.17}$	$1.85_{\pm 0.14}$	$1.77_{\pm 0.13}$
ϵ-greedy	$38.40_{\pm 2.67}$	$35.34_{\pm 2.16}$	$35.33_{\pm 1.49}$	$33.87_{\pm 1.43}$	$2.19_{\pm 0.21}$	$2.02_{\pm 0.19}$	$1.84_{\pm 0.17}$	$1.83_{\pm 0.16}$
UCB	$37.62_{\pm 2.59}$	$36.58_{\pm 2.32}$	$34.86_{\pm 1.47}$	$34.05_{\pm 1.50}$	$2.33_{\pm 0.22}$	$1.92_{\pm 0.18}$	$1.75_{\pm 0.16}$	$1.74_{\pm 0.11}$
InfoBatch	$38.07_{\pm 2.61}$	$36.12_{\pm 1.53}$	$34.56_{\pm 1.46}$	$33.12_{\pm 1.39}$	$1.97_{\pm 0.19}$	$1.76_{\pm 0.25}$	$1.73_{\pm 0.13}$	$1.66_{\pm 0.14}$
Cross-domain Coreset Selection								
MolPeg	$37.22_{\pm 1.58}$	$35.48_{\pm 1.52}$	$34.93_{\pm 1.48}$	$32.37_{\pm 1.49}$	$1.85_{\pm 0.18}$	$1.78_{\pm 0.23}$	$1.81_{\pm 0.16}$	$1.76_{\pm 0.17}$
CSIE	$36.14_{\pm 1.87}$	$35.69_{\pm 1.31}$	$33.49_{\pm 1.33}$	$32.50_{\pm 1.51}$	$1.93_{\pm 0.19}$	$1.85_{\pm 0.21}$	$1.87_{\pm 0.13}$	$1.73_{\pm 0.15}$
MDG-IS	**30.36**$_{\pm 1.20}$ (Core Set Size 21.5%)				**1.61**$_{\pm 0.13}$(Core Set Size 16.8%)			
Full dataset	$33.59_{\pm 1.24}$				$1.71_{\pm 0.16}$			

4 EXPERIMENTS

4.1 EXPERIMENTS SETUP

Baselines. The static Coreset filtering method includes Hard Random, Herding Welling (2009), K-Means Sorscher et al. (2022), Entropy Coleman et al. (2019), Least Confidence Coleman et al. (2019), GraNd-20 Paul et al. (2021), Glister Killamsetty et al. (2021), Influence Koh & Liang (2017), EL2N-2 Toneva et al. (2018) and DP Yang et al. (2022). These methods pre screen the Coreset before training. The dynamic Coreset filtering method includes Soft Random, ϵ-greedy Raju et al. (2021), UCB Raju et al. (2021), InfoBacth Qin et al. (2024).These methods dynamically adjust the participating samples during the training process. And MolPeg Chen et al. (2024) and CSIE Wang et al. are coreset selection methods specifically developed for cross-domain problems.

Experimental Settings. For solvation free energy prediction, the model is trained using the Adam optimizer Kingma & Ba (2014) with an initial learning rate of $1e^{-4}$. Batch size is set to 64. The training is stopped either when the validation loss does not improve for 50 epochs or when a maximum of 1000 epochs is reached. We implement the model in PyTorch and train on a Tesla A100 80GB GPU.

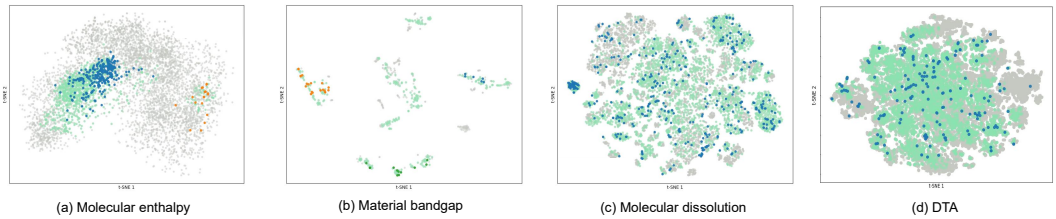
Evaluation Metrics. As a regression task, unless there are other recognized metrics, we use Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) to evaluate model performance. Each experiment is repeated three times, and we report the mean and standard deviation.

4.2 PERFORMANCE ANALYSIS

Obs.❶. MDG-IS consistently maintains excellent performance in various settings. Taking molecular enthalpy prediction and DTA prediction tasks as examples, the performance of MDG-IS is significantly better than all baseline methods. It is worth noting that in the enthalpy prediction task, MDG-IS achieved an improvement of 9.62% on MAE using only complete pre training data, surpassing CSIE and other subset based pre training strategies on all subset sizes. A similar trend was observed in the DTA task, where MDG-IS improved prediction performance by 5.84% compared to using a complete dataset. These results highlight the effectiveness of our coreset selection mechanism: a compact but highly aligned coreset can produce better training performance than a larger but weakly correlated subset.

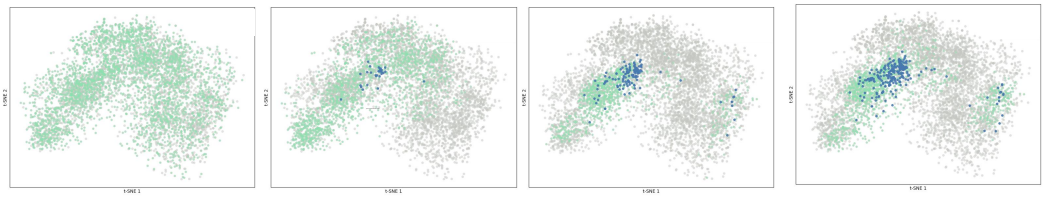
Table 2: Experimental results of predicting material bandgap and molecular dissolution with baselines under different coreset size. The Full dataset means we use all theoretical data for pretraining

Dataset	NREL-MatDB PBE →GWVD (MAE↓)				CombiSolv QM→Exp (RMSE↓)			
Backbone	MLP				D-MPNN			
Coreset size %	20	40	60	80	5	10	30	60
Static Coreset Selection								
Hard Random	$0.652_{\pm 0.072}$	$0.643_{\pm 0.041}$	$0.541_{\pm 0.029}$	$0.498_{\pm 0.095}$	$0.736_{\pm 0.048}$	$0.690_{\pm 0.023}$	$0.691_{\pm 0.012}$	$0.685_{\pm 0.031}$
Herding	$0.593_{\pm 0.088}$	$0.582_{\pm 0.067}$	$0.475_{\pm 0.039}$	$0.482_{\pm 0.054}$	$0.752_{\pm 0.020}$	$0.742_{\pm 0.041}$	$0.720_{\pm 0.018}$	$0.689_{\pm 0.027}$
K-Means	$0.587_{\pm 0.037}$	$0.490_{\pm 0.094}$	$0.483_{\pm 0.061}$	$0.479_{\pm 0.022}$	$0.741_{\pm 0.055}$	$0.711_{\pm 0.030}$	$0.689_{\pm 0.015}$	$0.692_{\pm 0.048}$
Entropy	$0.511_{\pm 0.062}$	$0.486_{\pm 0.028}$	$0.499_{\pm 0.053}$	$0.487_{\pm 0.071}$	$0.697_{\pm 0.036}$	$0.690_{\pm 0.014}$	$0.685_{\pm 0.023}$	$0.683_{\pm 0.041}$
Least Confidence	$0.520_{\pm 0.045}$	$0.506_{\pm 0.079}$	$0.480_{\pm 0.032}$	$0.461_{\pm 0.068}$	$0.713_{\pm 0.059}$	$0.705_{\pm 0.033}$	$0.698_{\pm 0.026}$	$0.701_{\pm 0.019}$
GraNd-20	$0.536_{\pm 0.051}$	$0.512_{\pm 0.044}$	$0.506_{\pm 0.083}$	$0.479_{\pm 0.037}$	$0.708_{\pm 0.021}$	$0.675_{\pm 0.060}$	$0.685_{\pm 0.027}$	$0.678_{\pm 0.016}$
Glister	$0.573_{\pm 0.099}$	$0.496_{\pm 0.047}$	$0.488_{\pm 0.072}$	$0.485_{\pm 0.034}$	$0.753_{\pm 0.041}$	$0.712_{\pm 0.025}$	$0.681_{\pm 0.044}$	$0.670_{\pm 0.014}$
Influence	$0.539_{\pm 0.064}$	$0.514_{\pm 0.053}$	$0.487_{\pm 0.087}$	$0.482_{\pm 0.058}$	$0.685_{\pm 0.011}$	$0.686_{\pm 0.036}$	$0.673_{\pm 0.020}$	$0.670_{\pm 0.055}$
EL2N-2	$0.538_{\pm 0.035}$	$0.491_{\pm 0.076}$	$0.490_{\pm 0.048}$	$0.484_{\pm 0.063}$	$0.697_{\pm 0.028}$	$0.692_{\pm 0.017}$	$0.689_{\pm 0.033}$	$0.663_{\pm 0.046}$
DP	$0.552_{\pm 0.081}$	$0.487_{\pm 0.039}$	$0.465_{\pm 0.059}$	$0.481_{\pm 0.072}$	$0.733_{\pm 0.062}$	$0.700_{\pm 0.042}$	$0.688_{\pm 0.025}$	$0.672_{\pm 0.018}$
Dynamic Coreset Selection								
Soft Random	$0.576_{\pm 0.068}$	$0.513_{\pm 0.083}$	$0.492_{\pm 0.027}$	$0.497_{\pm 0.060}$	$0.783_{\pm 0.049}$	$0.704_{\pm 0.012}$	$0.681_{\pm 0.039}$	$0.675_{\pm 0.021}$
ϵ-greedy	$0.544_{\pm 0.033}$	$0.493_{\pm 0.066}$	$0.471_{\pm 0.075}$	$0.469_{\pm 0.058}$	$0.696_{\pm 0.018}$	$0.682_{\pm 0.029}$	$0.681_{\pm 0.055}$	$0.664_{\pm 0.036}$
UCB	$0.531_{\pm 0.044}$	$0.507_{\pm 0.091}$	$0.482_{\pm 0.033}$	$0.460_{\pm 0.050}$	$0.706_{\pm 0.026}$	$0.685_{\pm 0.015}$	$0.673_{\pm 0.040}$	$0.660_{\pm 0.047}$
InfoBatch	$0.493_{\pm 0.030}$	$0.471_{\pm 0.055}$	$0.462_{\pm 0.066}$	$0.459_{\pm 0.088}$	$0.687_{\pm 0.053}$	$0.674_{\pm 0.022}$	$0.652_{\pm 0.035}$	$0.669_{\pm 0.012}$
Cross-domain Coreset Selection								
MolPeg	$0.512_{\pm 0.041}$	$0.465_{\pm 0.049}$	$0.473_{\pm 0.072}$	$0.455_{\pm 0.057}$	$0.685_{\pm 0.036}$	$0.683_{\pm 0.031}$	$0.671_{\pm 0.029}$	$0.668_{\pm 0.032}$
CSIE	$0.477_{\pm 0.037}$	$0.481_{\pm 0.054}$	$0.460_{\pm 0.057}$	$0.458_{\pm 0.061}$	$0.693_{\pm 0.052}$	$0.684_{\pm 0.038}$	$0.681_{\pm 0.030}$	$0.652_{\pm 0.028}$
MDG-IS	**0.436**$_{\pm 0.021}$ (Core Set Size 63.3%)				**0.628**$_{\pm 0.033}$(Core Set Size 35.2%)			
Full dataset	$0.463_{\pm 0.043}$				$0.652_{\pm 0.041}$			



(a) Molecular enthalpy (b) Material bandgap (c) Molecular dissolution (d) DTA

Figure 2: Distribution maps of theoretical data, Coreset, and experimental data of four datasets. Here, gray for theoretical domain samples, green for coreset samples, and other colors like orange and blue for different subdomain real samples.



(a) No experimental data available (b) 10% experimental data available (c) 40% experimental data available (d) 80% experimental data available

Figure 3: In real-life scenarios, real data constantly changes and the coreset also changed accordingly. Here, gray for theoretical domain samples, green for coreset samples, and blue for real domain samples.

Obs.❷. MDG-IS adaptively determines core-set size. Prior core-set methods depend on repeated threshold trials, yielding limited gains but large computational overhead. MDG-IS avoids this by using a unified sampling–deletion–convergence mechanism that relies solely on distributional alignment. Through KL-based filtering, it automatically isolates compact yet representative coresets (e.g., 21% on Wudily_Cho, 15.8% on Davis), which accelerates training and reduces memory without sacrificing accuracy.

Obs.❸. The size of the coreset selected by MDG-IS largely depends on the task and data scale. In four cases, we observed a clear pattern: tasks with scarce experimental data but abundant theoretical data (Molecular enthalpy, DTA) require only a small subset of theory to achieve powerful performance; Tasks with limited data in two domains (Bandgap) exhibit unstable baseline behavior, forcing MDG-IS to adopt higher sampling rates; Tasks with sufficient data in both domains (Dissolution) exhibit performance closely related to the size of the pretrained subset, and MDG-IS accordingly chooses a larger core. This indicates that the optimal alignment of the coreset size depends on the effective chemical spatial coverage between the theoretical and experimental domains, where any shortage on either side will alter the sampling rate.

4.3 Visual Analysis

To further assess the effectiveness of MDG-IS in sample screening, we visualize the final core-set selections in all four scenarios. As shown in Figure 2, we apply UMAP McInnes et al. (2018) to jointly project theoretical data, the selected theoretical coreset, and experimental data. Figure 3 further uses the molecular enthalpy task to simulate real-world settings where experimental data gradually increases over time.

Obs.❹. With heterogeneous experimental data, MDG-IS effectively separates subdomains. In the molecular enthalpy and band-gap tasks (Figure 2a, b), scarce and highly diverse experimental samples lead to strong heterogeneity. MDG-IS leverages subdomain-aware GMM modeling to identify these regions and select highly aligned theoretical data, thereby narrowing domain discrepancies. In DTA and dissolution tasks, where experimental data is abundant and broadly distributed, MDG-IS naturally collapses the structure into a near-single distribution and screens theoretical samples across the full chemical space.

Obs.❺. With zero experimental samples, MDG-IS degenerates into single-domain core-set selection. In Figure 3a, we simulate the initial stage with no experimental data. MDG-IS uniformly covers the theoretical space and, via the control in Eq. equation 17, constructs a coarse GMM representation of the entire theoretical domain using only a small subset. This achieves broad feature-space coverage with minimal cost, showing that the algorithm remains functional even under zero-sample conditions.

Obs.❻. In dynamic real-world scenarios, MDG-IS avoids repeated retraining. As illustrated in Figure 3, conventional approaches must either retrain from scratch when the target domain shifts—incurring high cost—or reuse a fixed pretrained model, often at the expense of accuracy. MDG-IS instead incrementally adds or removes source-domain samples and updates the coreset to match the new experimental distribution. Only the pretrained model needs light adjustment to accommodate the evolving coreset, achieving a favorable balance between adaptability, cost, and performance.

Table 3: Ablation study across four evaluation scenarios.

	Molecular enthalpy		DTA		Material bandgap		Dissolution reaction	
Method	**Coreset size**	**MAE**	**Coreset size**	**MAE**	**Coreset size**	**MAE**	**Coreset size**	**RMSE**
w/o Subspace-weighted GMM	26.6%	$32.54_{\pm 1.22}$	23.2%	$1.83_{\pm 0.08}$	58.7%	$0.467_{\pm 0.021}$	40.1%	$0.761_{\pm 0.072}$
w/o KL-guided bandit selection	27.9%	$31.79_{\pm 1.39}$	15.4%	$1.69_{\pm 0.09}$	75.2%	$0.452_{\pm 0.021}$	34.7%	$0.732_{\pm 0.085}$
w/o Coreset purification	37.2%	$34.36_{\pm 2.01}$	21.3%	$1.89_{\pm 0.12}$	75.1%	$0.515_{\pm 0.033}$	36.2%	$0.951_{\pm 0.121}$
MDG-IS	21.5%	$30.36_{\pm 1.40}$	16.8%	$1.61_{\pm 0.08}$	63.3%	$0.436_{\pm 0.021}$	35.2%	$0.663_{\pm 0.081}$

4.4 Ablation & Hyperparameter Analysis

Obs.❼. Ablation study. We conduct ablations by removing: (1) **Subspace-weighted GMM**, replacing subdomain modeling with a single GMM; (2) **KL-guided bandit selection**, replacing KL-based scoring with $G_t(z_j)$; and (3) **Core-set purification**, disabling iterative add–delete refinement. As shown in Table 3, the full MDG-IS achieves the best performance. Removing core-set purification causes the largest drop, highlighting its role in suppressing theoretical-data noise, followed by subspace-weighted GMM, which captures experimental-data heterogeneity. KL-based scoring has the smallest impact, suggesting that its role can be partially substituted by alternative metrics.

Table 4: Relationship between λ and real data size in the molecular enthalpy scenario.

Size (%)	Metrics	$\lambda = 0.1$	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 1.0$
20	MAE	**35.43**	38.62	39.31	47.54
	RMSE	**46.77**	48.16	49.50	58.31
60	MAE	33.48	**32.53**	35.45	35.65
	RMSE	44.62	**43.71**	49.53	49.44
100	MAE	30.36	30.53	**28.40**	32.62
	RMSE	40.43	42.97	**38.82**	44.77

Obs.❽. Hyperparameter sensitivity. As stated in Obs. ❺., Eq. equation 17 controls λ, balancing real-domain alignment and theoretical-domain exploration. Its optimal value increases with real-data availability: a smaller λ is preferred with scarce data (20%), a medium value at 60%, and a larger value with sufficient data (100%). This indicates that more real data favors stronger alignment with the real distribution, whereas scarce data benefits from broader theoretical exploration.

5 Conclusion

We introduce MDG-IS, a target domain-aware data selection framework that mitigates the persistent distribution mismatch between scarce experimental measurements and biased theoretical datasets in AI4Chemistry. MDG-IS performs fine-grained subdomain-level modeling of real data using a multi-subdomain GMM, followed by KL-guided iterative pruning between real and theoretical data to identify theoretical samples that truly align with the experimental distribution. MDG-IS consistently improves prediction accuracy across benchmarks while using significantly fewer theoretical samples, suggesting that distribution alignment is a key driver of transfer. The method adapts to varying fidelity gaps, reduces negative transfer, and remains robust even under extreme data scarcity.

AI USE STATEMENT

In this study, we used generative AI tools to polish the writing of the paper. We have reviewed all work assisted by AI to ensure that the polished manuscript contains no new hallucinations and that the original meaning remains unchanged.

ETHICS STATEMENT

This article does not involve ethical issues.

REPRODUCIBILITY STATEMENT

Code is available at: https://MDG-IS. For further experiments, datasets, and parameter settings, please refer to the Appendix C SUPPLEMENTARY EXPERIMENTS..

REFERENCES

Valentine P Ananikov. Artificial intelligence chemistry. *Artificial Intelligence*, 2:100075, 2024a.

Valentine P Ananikov. Top 20 influential ai-based technologies in chemistry. *Artificial Intelligence Chemistry*, pp. 100075, 2024b.

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*, 384:133999, 2025.

David Buterez, Jon Paul Janet, Steven J Kiddle, Dino Oglic, and Pietro Lió. Transfer learning with graph neural networks for improved molecular property prediction in the multi-fidelity setting. *Nature communications*, 15(1):1517, 2024.

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.

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

Yunsie Chung and William H Green. Machine learning from quantum chemistry to predict experimental solvent effects on reaction rates. *Chemical Science*, 15(7):2410–2424, 2024.

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. *arXiv preprint arXiv:1906.11829*, 2019.

Mindy I Davis, Jeremy P Hunt, Sanna Herrgard, 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(11):1046–1051, 2011.

V Divya, V Aparna, Sunitha Ashna Rasheed, Lakshmy M V R, and K Anu. The transformative role of artificial intelligence in advancing chemistry: A comprehensive review. *Library Progress International*, 44(3):23917–23927, 2024.

Wenjie Du, Shuai Zhang, Zhaohui Cai, Zhiyuan Liu, Junfeng Fang, Jianmin Wang, and Yang Wang. Molecular merged hypergraph neural network for explainable solvation free energy prediction. *Research*, 8:0740, 2025. doi: 10.34133/research.0740. URL https://spj.science.org/doi/abs/10.34133/research.0740.

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.

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.

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*, pp. 302–310, 2018.

Ryan-Rhys Griffiths, Philippe Schwaller, and Alpha A Lee. Dataset bias in the natural sciences: a case study in chemical reaction prediction and synthesis design. *arXiv preprint arXiv:2105.02637*, 2021.

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.

Herim Han and Sunghwan Choi. Transfer learning from simulation to experimental data: Nmr chemical shift predictions. *The Journal of Physical Chemistry Letters*, 12(14):3662–3668, 2021.

John R Hershey and Peder A Olsen. Approximating the kullback leibler divergence between gaussian mixture models. In *2007 IEEE International Conference on Acoustics, Speech and Signal Processing-ICASSP'07*, volume 4, pp. IV–317. IEEE, 2007.

Frederick James. Monte carlo theory and practice. *Reports on progress in Physics*, 43(9):1145, 1980.

Xiaohong Ji, Zhen Wang, Zhifeng Gao, Hang Zheng, Linfeng Zhang, Guolin Ke, et al. Uni-mol2: Exploring molecular pretraining model at scale. *arXiv preprint arXiv:2406.14969*, 2024.

S Sathiya Keerthi and Chih-Jen Lin. Asymptotic behaviors of support vector machines with gaussian kernel. *Neural computation*, 15(7):1667–1689, 2003.

Krishnateja Killamsetty, Durga Sivasubramanian, Ganesh Ramakrishnan, and Rishabh Iyer. Glister: Generalization based data subset selection for efficient and robust learning. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pp. 8110–8118, 2021.

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.

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

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

Tobias Kreiman and Aditi S Krishnapriyan. Understanding and mitigating distribution shifts for machine learning force fields. *arXiv preprint arXiv:2503.08674*, 2025.

Argonne National Laboratory. Active thermochemical tables (atct) — thermochemical data resource, 2025. URL `https://atct.anl.gov/`.

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, 2021.

Kha-Dinh Luong and Ambuj K Singh. Fragment-based pretraining and finetuning on molecular graphs. *Advances in Neural Information Processing Systems*, 36:17584–17601, 2023.

Andrzej Maćkiewicz and Waldemar Ratajczak. Principal components analysis (pca). *Computers & Geosciences*, 19(3):303–342, 1993.

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.

Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

Antonia Mera, Martin Vogt, and Jürgen Bajorath. A meta-learning framework to mitigate negative transfer in transfer learning applicable to drug design. *Scientific Reports*, 15(1):35236, 2025.

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

National Renewable Energy Laboratory. Nrel materials database. Accessed: 2025-05-06, 2024. `https://materials.nrel.gov/queryStd`.

Seyedmehdi Orouji, Martin C Liu, Tal Korem, and Megan AK Peters. Domain adaptation in small-scale and heterogeneous biological datasets. *Science Advances*, 10(51):eadp6040, 2024.

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.

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. URL `https://openreview.net/forum?id=C61sk5LsK6`.

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.

Carl Rasmussen. The infinite gaussian mixture model. *Advances in neural information processing systems*, 12, 1999.

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

Yu Rong, Yatao Bian, Tingyang Xu, Weiyang Xie, Ying Wei, Wenbing Huang, and Junzhou Huang. Self-supervised graph transformer on large-scale molecular data. *Advances in neural information processing systems*, 33:12559–12571, 2020.

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.

Darik A Rosser, Brianna R Farris, and Kevin C Leonard. Predictive machine learning models trained on experimental datasets for electrochemical nitrogen reduction. *Digital Discovery*, 3(4):667–673, 2024.

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.

Tomoya Shiota, Kenji Ishihara, Tuan Minh Do, Toshio Mori, and Wataru Mizukami. Taming multi-domain,-fidelity data: towards foundation models for atomistic scale simulations. *arXiv preprint arXiv:2412.13088*, 2024.

Aleksandrs Slivkins et al. Introduction to multi-armed bandits. *Foundations and Trends® in Machine Learning*, 12(1-2):1–286, 2019.

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.

Mariya Toneva, Alessandro Sordoni, Remi Tachet des Combes, Adam Trischler, Yoshua Bengio, and Geoffrey J Gordon. An empirical study of example forgetting during deep neural network learning. *arXiv preprint arXiv:1812.05159*, 2018.

Florence H Vermeire and William H Green. Transfer learning for solvation free energies: From quantum chemistry to experiments. *Chemical Engineering Journal*, 418:129307, 2021.

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.

Vivin Vinod, Ulrich Kleinekathöfer, and Peter Zaspel. Optimized multifidelity machine learning for quantum chemistry. *Machine Learning: Science and Technology*, 5(1):015054, 2024.

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.

Hanchen Wang, Jean Kaddour, Shengchao Liu, Jian Tang, Joan Lasenby, and Qi Liu. Evaluating self-supervised learning for molecular graph embeddings. *Advances in Neural Information Processing Systems*, 36:68028–68060, 2023.

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. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*.

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

Thomas Weymuth and Markus Reiher. The transferability limits of static benchmarks. *Physical Chemistry Chemical Physics*, 24(24):14692–14698, 2022.

Hongjie Wu, Junkai Liu, Tengsheng Jiang, Quan Zou, Shujie Qi, Zhiming Cui, Prayag Tiwari, and Yijie Ding. Attentionmgt-dta: A multi-modal drug-target affinity prediction using graph transformer and attention mechanism. *Neural Networks*, 169:623–636, 2024a.

Lirong Wu, Yijun Tian, Yufei Huang, Siyuan Li, Haitao Lin, Nitesh V Chawla, and Stan Z Li. Mape-ppi: Towards effective and efficient protein-protein interaction prediction via microenvironment-aware protein embedding. *arXiv preprint arXiv:2402.14391*, 2024b.

Jun Xia, Chengshuai Zhao, Bozhen Hu, Zhangyang Gao, Cheng Tan, Yue Liu, Siyuan Li, and Stan Z Li. Mole-bert: Rethinking pre-training graph neural networks for molecules. 2023.

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.

Shuo Yang, Zeke Xie, Hanyu Peng, Min Xu, Mingming Sun, and Ping Li. Dataset pruning: Reducing training data by examining generalization influence. *arXiv preprint arXiv:2205.09329*, 2022.

Yuning You, Tianlong Chen, Yongduo Sui, Ting Chen, Zhangyang Wang, and Yang Shen. Graph contrastive learning with augmentations. *Advances in neural information processing systems*, 33: 5812–5823, 2020.

A APPENDIX

You may include other additional sections here.

B RELATED WORK.

The deployment in the real world has exposed the fundamental challenge posed by the system mismatch between theoretical and experimental distributions. Firstly, theoretical bias typically depends on subdomains: different chemical families, reaction categories, or material categories generate inconsistent systematic errors in calculating labels, so a single global pretraining of all available theoretical data may inject misleading inductive bias into some downstream targets. Secondly, the data measured in experiments is extremely scarce and heterogeneous, collected under different conditions and protocols. Therefore, simple fine-tuning is prone to overfitting and may suffer from negative transfer. These concerns have been identified in several empirical studies and position papers, calling for more careful handling of comparisons and data selection Griffiths et al. (2021); Weymuth & Reiher (2022); Rosser et al. (2024).

Transfer Learning. Transfer learning leverages abundant theoretical data by pretraining models on large simulation corpora before fine-tuning on scarce experimental observations. This paradigm has been successfully applied to molecular property prediction, NMR chemical shift estimation, and reaction modeling when theoretical and experimental distributions exhibit relatively stable correspondence Han & Choi (2021); Buterez et al. (2024). However, its effectiveness relies on the assumption that knowledge learned from the source domain is broadly transferable. When theoretical data contain heterogeneous biases or weak correlations with downstream experimental tasks, large-scale pretraining may introduce irrelevant inductive biases, increase computational overhead, and even lead to negative transfer.

Multi-fidelity Learning. Multi-fidelity learning explicitly models the relationship between low- and high-fidelity observations to jointly exploit inexpensive simulations and scarce experimental measurements. Representative approaches, including Δ-learning and hierarchical multi-fidelity regression, have demonstrated strong performance when low-fidelity signals provide reliable approximations to high-fidelity targets Ramakrishnan et al. (2015); Fare et al. (2022); Vinod et al. (2024); Vinod & Zaspel (2025). Nevertheless, these methods generally assume stable cross-fidelity correlations across the entire data space. Their performance deteriorates when fidelity gaps vary across chemical subdomains or exhibit strong non-stationary behavior.

Domain Adaptation. Domain adaptation methods formulate the discrepancy between theoretical and experimental data as a distribution shift and employ feature alignment, importance weighting, or calibration techniques to reduce domain mismatch Kreiman & Krishnapriyan (2025); Orouji et al. (2024). Recent studies have further combined uncertainty estimation with domain adaptation to improve robustness under heterogeneous experimental conditions. However, most existing methods focus on learning domain-invariant representations after all source samples have been incorporated, rather than identifying which theoretical samples should be retained before model training.

Data-centric Data Selection. Recent advances in data-centric AI suggest that data quality often outweighs data quantity for downstream model performance. Representative approaches, including dataset pruning, data valuation, subset selection, and active learning, demonstrate that carefully curated training data can outperform substantially larger but weakly related datasets while reducing computational cost Rong et al. (2020); Xia et al. (2023); Luong & Singh (2023); Ji et al. (2024). Nevertheless, these methods typically estimate sample importance independently of the target experimental distribution and therefore do not explicitly address heterogeneous fidelity gaps or subdomain-dependent distribution mismatch.

Meta Learning. Meta-learning approaches alleviate negative transfer by learning source-example or source-task weighting strategies that maximize downstream adaptation performance Mera et al. (2025). Such methods improve knowledge transfer under distribution shift by optimizing sample importance during training. However, they generally require repeated bi-level optimization and focus on model adaptation rather than constructing compact, representative datasets prior to training, limiting their scalability for massive scientific repositories.

Despite their different formulations, existing approaches share two common limitations. First, they generally assume that theoretical samples are globally informative and therefore focus on representation learning after incorporating nearly all available data. Second, few methods explicitly consider that the usefulness of a theoretical sample depends on its compatibility with the local target experimental distribution. Consequently, they are unable to effectively address subdomain-dependent systematic bias while simultaneously reducing computational overhead. These limitations motivate a target-aware data selection framework that explicitly models heterogeneous experimental distributions and constructs compact, alignment-consistent theoretical subsets before downstream learning.

C SUPPLEMENTARY EXPERIMENTS.

C.1 BASELINES.

We compare our method against a comprehensive set of static and dynamic coreset selection strategies, as well as approaches specifically designed for cross-domain scenarios.

Hard Random randomly extracts subsets from the training set without relying on any model signals or feature statistics. It serves as the most fundamental unbiased baseline for measuring the relative returns of other selection strategies.

Herding Welling (2009) is based on the idea of feature mean matching, greedily selecting samples that make the subset feature mean close to the overall data feature mean. This method is commonly used to construct representative small-scale datasets.

K-Means Sorscher et al. (2022) performs clustering in the feature space by selecting points near each cluster center to obtain representative samples. It can cover the main patterns of data and reduce redundancy.

Entropy Coleman et al. (2019) sorts the predicted entropy based on the model and selects the most uncertain sample for prediction. This method is suitable for selecting the most informative data points in active learning or data pruning.

Least Confidence Coleman et al. (2019) uses the reciprocal of the maximum class probability as the uncertainty measure to select the sample with the least confidence in the model. Compared to entropy, it is often simpler but equally effective.

GraNd-20 Paul et al. (2021) uses the gradient norm from early training to score the samples, retaining the samples with the highest gradients as they have the greatest impact on model updates. This method embodies the principle of 'difficult samples are more important'.

EL2N-2 Toneva et al. (2018) evaluates sample difficulty based on the initial prediction error (L2 norm) during training, with greater errors being prioritized for retention. It can effectively identify key samples that are crucial for model convergence.

Influence Koh & Liang (2017) approximates the impact of a single training sample on validation loss and selects the sample with the greatest impact. It reflects the direct causal relationship between training data and validation performance.

Glister Killamsetty et al. (2021) models subset selection as a two-layer optimization, selecting data by maximizing the improvement in validation performance. This method has strong theoretical basis and combines information coverage and gradient value.

DP Yang et al. (2022) estimates sample importance based on early loss or gradient statistics and prunes low importance data. The core idea is to improve training efficiency by removing redundant or easy to learn samples.

ϵ-greedy Raju et al. (2021) uses an exploration exploitation strategy to dynamically select data, balancing between high score samples and random samples. This can avoid getting stuck in local optima and continuously discover potential valuable samples.

UCB Raju et al. (2021) introduces an upper bound on uncertainty and combines sample value with selection confidence interval for selection. It encourages exploration on samples with unclear value but potential returns.

InfoBatch Qin et al. (2024) adjusts the sampling probability based on sample loss statistics and scales the loss through probability loss measurement. This method can dynamically filter low information samples and accelerate training.

MolPeg Chen et al. (2024) monitors the real-time loss of samples during the target domain update process and dynamically selects the most valuable data for the current task. It is particularly suitable for transfer learning or cross domain scenarios.

CSIE Wang et al. selects the core subset by exploring the maximum generalizable space, ensuring that the subset covers the critical decision areas of the task.

C.2 Datasets.

C.2.1 Thermodynamic Property Prediction: Standard Enthalpy of Formation $\Delta_f H^\circ(298.15\,\mathrm{K})$

Theoretical Dataset: wudily_cho. This dataset originates by Bruce et al **?** It compiles computational thermochemical results from multiple research groups and covers hydrocarbons, oxygenates, polycyclic aromatic hydrocarbons (PAHs), oxygenated PAHs (OPAHs), and their corresponding radicals. The data are primarily derived from high-accuracy single-point energy calculations performed using composite methods such as CBS-QB3, CBS-APNO, G3, and G4. The dataset contains 6074 entries and is fully consolidated, allowing direct use without additional preprocessing.

Experimental Dataset: ATcT. The Active Thermochemical Tables (ATcT), developed at Argonne National Laboratory, constitute a dynamically updated platform designed to provide accurate and internally consistent thermochemical data. ATcT determines property values by constructing and globally optimizing an extensive thermochemical network that integrates all available experimental and theoretical information. This network-based approach effectively resolves the inconsistencies typically encountered in traditional thermochemical methods. The typical uncertainty of ATcT values is within ± 0.1–$1\,\mathrm{kJ\ mol^{-1}}$, and for small molecules it can reach $\pm 0.05\,\mathrm{kJ\ mol^{-1}}$.

Version 1.220 contains 3423 entries. Because SMILES representations are not available in the original database, we generated them from the Preferred Formula using a large language model and subsequently verified their validity. Additional validation and filtering of invalid or unparsable SMILES strings were performed using the `smiles2graph` utility in the OGB library. We retained only gas-phase entries and removed outliers with $\Delta_f H^\circ(298.15\,\mathrm{K})$ exceeding $\pm 500\,\mathrm{kJ\ mol^{-1}}$, resulting in a final curated dataset of 358 samples.

Processing Pipeline. SMILES strings were converted into molecular graphs using the OGB library. Molecular features were extracted using the `GINEConv` architecture from PyTorch Geometric, followed by an MLP to produce scalar regression outputs. Training employed mean squared error (MSE) loss, and model performance was evaluated using MAE and RMSE.

C.2.2 Drug–Target Affinity Prediction Datasets

Low-Precision Dataset: Davis. We employ the processed Davis dataset from Wu et al. Wu et al. (2024a). The dataset contains 68 drugs and 361 targets. Each drug–target pair is annotated with an affinity value measured through the kinase activity inhibition assay, expressed as the dissociation constant K_d. These values are transformed into the negative logarithmic scale (pK_d), resulting in a range from 5.0 to 10.8. After preprocessing, the dataset consists of 24,956 valid interaction pairs.

High-Precision Benchmark Dataset: CASF-2016. CASF-2016 Su et al. (2018) is a high-accuracy benchmark constructed from the Refined Set of PDBbind v2016. It comprises 285 protein–ligand complexes selected based on high-resolution X-ray crystal structures (resolution ≤ 2.5 Å). Each entry includes experimentally measured binding affinities together with binding pocket information and ligand 3D coordinates. Compared with large-scale high-throughput datasets such as Davis and KIBA, CASF-2016 offers substantially higher measurement fidelity. To ensure comparability with the Davis dataset, we simplify the original 3D structural data into molecular SMILES for ligands and amino-acid sequences for proteins.

Processing Pipeline. SMILES strings are processed using the OGB and PyTorch Geometric frameworks, while protein sequences are encoded with a TransformerEncoder. The resulting representations are concatenated and fed into an MLP for affinity regression. Model training uses mean squared error (MSE) loss.

C.2.3 Solvation Reaction Gibbs Free Energy Prediction

Theoretical Dataset: CombiSolv-QM. CombiSolv-QM, introduced by Vermeire et al. Vermeire & Green (2021) in 2021, is designed to support quantum-chemistry–based transfer learning models aimed at improving prediction accuracy on experimental solvation data. The dataset contains approximately one million solute-solvent pairs, each annotated with a computed solvation free energy value. All entries are generated using the COSMOtherm framework, providing a large-scale and internally consistent quantum-chemical benchmark.

Experimental Dataset: CombiSolv-Exp. CombiSolv-Exp aggregates experimental solvation free energy measurements from several public resources, including MiniSolv, FreeSolv, CompSol, and Abraham databases. The full collection consists of about 10,145 measurements, with 8,780 entries publicly available and suitable for model training and evaluation.

Processing Pipeline. We employ a graph neural network based on the directed message passing neural network architecture to encode molecular structures and perform regression on solvation free energies.

C.2.4 Material Band Gap Prediction

Datasets. We evaluate our models on two band gap datasets from the NREL-MatDB National Renewable Energy Laboratory (2024) repository. The low-precision theoretical band gap dataset contains 546 band gap values computed using the PBE functional. The high-precision experimental band gap dataset consists of 52 band gap values obtained from GWVD calculations. Together, these datasets provide complementary sources of approximate and accurate electronic-structure information.

Experimental Method. Each material's chemical formula is featurized using `matminer`, from which we extract a set of key descriptors including atomic number, electronegativity, atomic radius, ionization energy, atomic packing efficiency, and valence and conduction band centers. These features are passed into a multilayer MLP to predict the corresponding molecular band gap.

C.3 Backbone Analysis.

As shown in Table 5 and Table 6, compared with the main text, a stronger backbone model will have better performance, which is in line with our expectations, but the effect of the data selection method itself on the model still exists. Similar to previous observations, dynamic data filtering methods have better results, and our method achieves the best results on different data sets and different backbones, which illustrates the superiority of our method.

C.4 Complexity Analysis.

Complexity Observation. Table 7 summarizes the computational time required by the MDG-IS framework across four representative datasets with varying sizes and modalities. As expected, the runtime scales approximately linearly with the dataset volume: small datasets such as NREL-MatDB PBE (546 entries) can be processed in only 0.1 minutes, while medium-sized datasets like Davis (24,956 entries) and Wudily_Cho (6,074 entries) require 13.5 and 2.4 minutes, respectively. The largest dataset, CombiSolv-QM, containing 1 million entries, requires 7.6 hours, reflecting the significant computational demand for high-volume solvation free energy data. These observations confirm that MDG-IS can handle large-scale datasets efficiently, although the absolute runtime is dominated by dataset size and molecular complexity.

Table 5: Experimental results of molecular enthalpy and drug-target binding affinity prediction with baselines under different coreset size. The Full dataset means we use all theoretical data for pretraining

Dataset	Wudily_Cho→ATcT (MAE↓)				Davis→CASF-2016 (MAE↓)			
Backbone	GraphSAGE				GraphSAGE+Transformer			
Coreset size %	20	40	60	80	20	40	60	80
Static Coreset Selection								
Hard Random	$37.55_{\pm 2.38}$	$36.07_{\pm 1.57}$	$35.85_{\pm 1.64}$	$32.21_{\pm 1.46}$	$2.62_{\pm 0.22}$	$2.34_{\pm 0.21}$	$1.80_{\pm 0.19}$	$1.67_{\pm 0.20}$
Herding	$39.10_{\pm 1.84}$	$39.42_{\pm 1.52}$	$36.20_{\pm 1.22}$	$33.12_{\pm 1.30}$	$2.68_{\pm 0.21}$	$1.95_{\pm 0.17}$	$1.72_{\pm 0.20}$	$1.61_{\pm 0.16}$
K-Means	$38.42_{\pm 2.30}$	$37.14_{\pm 1.39}$	$37.25_{\pm 1.37}$	$34.01_{\pm 1.27}$	$2.49_{\pm 0.20}$	$1.78_{\pm 0.15}$	$1.67_{\pm 0.12}$	$1.62_{\pm 0.14}$
Entropy	$34.80_{\pm 1.73}$	$33.80_{\pm 1.49}$	$32.15_{\pm 1.51}$	$32.15_{\pm 1.11}$	$2.24_{\pm 0.18}$	$1.73_{\pm 0.15}$	$1.72_{\pm 0.14}$	$1.67_{\pm 0.14}$
Least Confidence	$35.35_{\pm 1.68}$	$34.71_{\pm 1.34}$	$36.26_{\pm 1.65}$	$31.42_{\pm 1.40}$	$1.80_{\pm 0.16}$	$1.77_{\pm 0.15}$	$1.60_{\pm 0.14}$	$1.65_{\pm 0.13}$
GraNd-20	$36.20_{\pm 1.81}$	$35.20_{\pm 1.34}$	$33.80_{\pm 1.28}$	$32.20_{\pm 1.21}$	$1.69_{\pm 0.20}$	$1.71_{\pm 0.14}$	$1.61_{\pm 0.14}$	$1.58_{\pm 0.13}$
Glister	$35.44_{\pm 1.53}$	$36.68_{\pm 1.23}$	$34.30_{\pm 1.15}$	$32.51_{\pm 1.34}$	$2.32_{\pm 0.20}$	$1.95_{\pm 0.12}$	$1.74_{\pm 0.18}$	$1.63_{\pm 0.14}$
Influence	$36.00_{\pm 1.81}$	$34.80_{\pm 1.36}$	$33.63_{\pm 1.30}$	$31.55_{\pm 1.23}$	$1.88_{\pm 0.17}$	$1.65_{\pm 0.14}$	$1.62_{\pm 0.11}$	$1.72_{\pm 0.14}$
EL2N-2	$34.81_{\pm 2.20}$	$34.71_{\pm 1.43}$	$32.16_{\pm 1.26}$	$31.14_{\pm 1.16}$	$1.95_{\pm 0.18}$	$1.69_{\pm 0.15}$	$1.72_{\pm 0.14}$	$1.53_{\pm 0.16}$
DP	$35.18_{\pm 1.66}$	$34.17_{\pm 1.51}$	$32.56_{\pm 1.25}$	$30.81_{\pm 1.19}$	$2.33_{\pm 0.20}$	$1.91_{\pm 0.17}$	$1.66_{\pm 0.18}$	$1.57_{\pm 0.12}$
Dynamic Coreset Selection								
Soft Random	$34.80_{\pm 2.21}$	$33.48_{\pm 1.98}$	$32.63_{\pm 1.26}$	$32.25_{\pm 1.21}$	$2.41_{\pm 0.20}$	$1.84_{\pm 0.14}$	$1.71_{\pm 0.12}$	$1.64_{\pm 0.11}$
ϵ-greedy	$35.50_{\pm 2.27}$	$32.73_{\pm 1.84}$	$32.73_{\pm 1.27}$	$31.39_{\pm 1.22}$	$2.03_{\pm 0.18}$	$1.86_{\pm 0.16}$	$1.71_{\pm 0.14}$	$1.70_{\pm 0.15}$
UCB	$34.87_{\pm 2.20}$	$33.80_{\pm 1.97}$	$32.27_{\pm 1.25}$	$31.40_{\pm 1.27}$	$2.15_{\pm 0.19}$	$1.77_{\pm 0.15}$	$1.62_{\pm 0.14}$	$1.61_{\pm 0.09}$
InfoBatch	$35.18_{\pm 2.22}$	$33.59_{\pm 1.31}$	$32.16_{\pm 1.24}$	$30.61_{\pm 1.18}$	$1.82_{\pm 0.16}$	$1.63_{\pm 0.21}$	$1.60_{\pm 0.11}$	$1.54_{\pm 0.12}$
Cross-domain Coreset Selection								
MolPeg	$34.50_{\pm 1.35}$	$32.90_{\pm 1.29}$	$32.50_{\pm 1.26}$	$29.80_{\pm 1.27}$	$1.71_{\pm 0.15}$	$1.65_{\pm 0.21}$	$1.67_{\pm 0.14}$	$1.63_{\pm 0.14}$
CSIE	$33.50_{\pm 1.60}$	$33.10_{\pm 1.11}$	$31.00_{\pm 1.13}$	$29.90_{\pm 1.28}$	$1.78_{\pm 0.16}$	$1.71_{\pm 0.18}$	$1.73_{\pm 0.11}$	$1.60_{\pm 0.13}$
MDG-IS	**28.20**$_{\pm 1.02}$				**1.49**$_{\pm 0.11}$			
Full dataset	$31.00_{\pm 1.05}$				$1.58_{\pm 0.14}$			

Table 6: Experimental results of predicting material bandgap and molecular dissolution.

Dataset	NREL-MatDB PBE →GWVD (MAE↓)				CombiSolv QM→Exp (RMSE↓)			
Backbone	Gaussian Process Regression (GPR)				GAT			
Coreset size %	20	40	60	80	5	10	30	60
Static Coreset Selection								
Hard Random	$0.799_{\pm 0.088}$	$0.781_{\pm 0.050}$	$0.660_{\pm 0.035}$	$0.605_{\pm 0.116}$	$0.677_{\pm 0.041}$	$0.635_{\pm 0.020}$	$0.636_{\pm 0.010}$	$0.630_{\pm 0.027}$
Herding	$0.723_{\pm 0.107}$	$0.710_{\pm 0.082}$	$0.579_{\pm 0.048}$	$0.587_{\pm 0.066}$	$0.692_{\pm 0.017}$	$0.683_{\pm 0.035}$	$0.662_{\pm 0.015}$	$0.634_{\pm 0.023}$
K-Means	$0.716_{\pm 0.045}$	$0.596_{\pm 0.115}$	$0.588_{\pm 0.075}$	$0.583_{\pm 0.027}$	$0.682_{\pm 0.047}$	$0.654_{\pm 0.026}$	$0.634_{\pm 0.013}$	$0.636_{\pm 0.041}$
Entropy	$0.624_{\pm 0.076}$	$0.592_{\pm 0.034}$	$0.608_{\pm 0.065}$	$0.593_{\pm 0.087}$	$0.641_{\pm 0.031}$	$0.635_{\pm 0.012}$	$0.630_{\pm 0.020}$	$0.628_{\pm 0.035}$
Least Confidence	$0.635_{\pm 0.055}$	$0.617_{\pm 0.096}$	$0.585_{\pm 0.039}$	$0.562_{\pm 0.083}$	$0.656_{\pm 0.050}$	$0.650_{\pm 0.028}$	$0.642_{\pm 0.022}$	$0.651_{\pm 0.016}$
GraNd-20	$0.654_{\pm 0.062}$	$0.624_{\pm 0.054}$	$0.618_{\pm 0.101}$	$0.584_{\pm 0.045}$	$0.651_{\pm 0.018}$	$0.621_{\pm 0.051}$	$0.630_{\pm 0.023}$	$0.624_{\pm 0.014}$
Glister	$0.700_{\pm 0.121}$	$0.603_{\pm 0.057}$	$0.595_{\pm 0.088}$	$0.591_{\pm 0.041}$	$0.693_{\pm 0.035}$	$0.655_{\pm 0.022}$	$0.626_{\pm 0.037}$	$0.616_{\pm 0.012}$
Influence	$0.658_{\pm 0.078}$	$0.627_{\pm 0.065}$	$0.594_{\pm 0.106}$	$0.588_{\pm 0.071}$	$0.630_{\pm 0.009}$	$0.631_{\pm 0.031}$	$0.619_{\pm 0.017}$	$0.616_{\pm 0.047}$
EL2N-2	$0.657_{\pm 0.043}$	$0.598_{\pm 0.093}$	$0.597_{\pm 0.059}$	$0.590_{\pm 0.077}$	$0.641_{\pm 0.024}$	$0.637_{\pm 0.015}$	$0.634_{\pm 0.028}$	$0.609_{\pm 0.039}$
DP	$0.673_{\pm 0.099}$	$0.593_{\pm 0.048}$	$0.567_{\pm 0.072}$	$0.586_{\pm 0.088}$	$0.674_{\pm 0.053}$	$0.644_{\pm 0.036}$	$0.633_{\pm 0.021}$	$0.618_{\pm 0.015}$
Dynamic Coreset Selection								
Soft Random	$0.701_{\pm 0.083}$	$0.625_{\pm 0.101}$	$0.600_{\pm 0.033}$	$0.605_{\pm 0.073}$	$0.722_{\pm 0.042}$	$0.650_{\pm 0.010}$	$0.626_{\pm 0.033}$	$0.621_{\pm 0.018}$
ϵ-greedy	$0.663_{\pm 0.040}$	$0.601_{\pm 0.081}$	$0.574_{\pm 0.092}$	$0.572_{\pm 0.071}$	$0.640_{\pm 0.015}$	$0.627_{\pm 0.025}$	$0.626_{\pm 0.047}$	$0.611_{\pm 0.031}$
UCB	$0.648_{\pm 0.054}$	$0.619_{\pm 0.111}$	$0.588_{\pm 0.040}$	$0.561_{\pm 0.061}$	$0.650_{\pm 0.022}$	$0.630_{\pm 0.013}$	$0.619_{\pm 0.034}$	$0.607_{\pm 0.040}$
InfoBatch	$0.600_{\pm 0.037}$	$0.574_{\pm 0.067}$	$0.563_{\pm 0.081}$	$0.560_{\pm 0.107}$	$0.632_{\pm 0.045}$	$0.620_{\pm 0.019}$	$0.598_{\pm 0.030}$	$0.615_{\pm 0.010}$
Cross-domain Coreset Selection								
MolPeg	$0.625_{\pm 0.050}$	$0.567_{\pm 0.060}$	$0.577_{\pm 0.088}$	$0.555_{\pm 0.070}$	$0.630_{\pm 0.031}$	$0.628_{\pm 0.027}$	$0.617_{\pm 0.025}$	$0.614_{\pm 0.027}$
CSIE	$0.583_{\pm 0.045}$	$0.587_{\pm 0.066}$	$0.561_{\pm 0.070}$	$0.560_{\pm 0.075}$	$0.638_{\pm 0.044}$	$0.629_{\pm 0.033}$	$0.626_{\pm 0.025}$	$0.598_{\pm 0.024}$
MDG-IS	**0.531**$_{\pm 0.026}$				**0.578**$_{\pm 0.028}$			
Full dataset	$0.566_{\pm 0.053}$				$0.598_{\pm 0.035}$			

D LIMITATIONS.

While MDG-IS demonstrates robust performance across multiple cross-domain chemistry tasks, several limitations remain, offering directions for future research:

Reliance on low-dimensional latent space modeling: The multi-subdomain structure in MDG-IS depends on PCA-reduced GMM representations. When the experimental domain contains extremely few samples, the estimation of GMM covariances may be unstable, and subdomain structures may

Table 7: Time complexity of MDG-IS framework.

	Wudily_Cho	Davis	CombiSolv-QM	NREL-MatDB PBE
DATA VOLUME	6,074	24,956	1,000,000	546
TIME	2.4 min	13.5 min	7.6 h	0.1 min

shift due to noise, potentially affecting KL-guided alignment and iterative sampling. Future work could explore *non-parametric modeling* or *Bayesian hierarchical structures* to improve robustness under small-sample conditions.

Assumption of comparable local densities for KL alignment: $\mathrm{KL}(G_t \parallel G_C)$ is used to measure alignment between the theoretical coreset and experimental distribution, implicitly assuming that the local density scales are comparable. In practice, theoretical domains often vastly outnumber experimental samples, and the feature variances differ, which may bias the bandit sampling toward certain clusters. Future research may consider *contrastive divergence* or *density-ratio estimation* to mitigate such bias.

Heuristic threshold in coreset purification: The alignment deviation score (Eq. 17) uses a mean $\pm 1\sigma$ threshold to remove inconsistent samples. While effective, this heuristic may not be globally optimal; in noisy theoretical datasets, some boundary samples may be mistakenly removed, and different tasks may require different scales. Future improvements could involve *Bayesian purification* or *learnable thresholds*.