FROM INDEPENDENT TO CONTEXT-AWARE INFERENCE VIA TABULAR FOUNDATION MODELS FOR ENERGY PREDICTION ON 3D ATOMIC STRUCTURES

Anonymous authors
Paper under double-blind review

ABSTRACT

Machine Learning Interatomic Potentials (MLIPs) provide an efficient approach to approximating Density Functional Theory (DFT) energy calculations on 3D atomic structures. However, existing MLIPs largely follow an independent inference paradigm, leaving previously computed DFT observations unused and limiting potential performance gains from such observations. We introduce In-Context MLIPs (IC-MLIPs), a new inference paradigm that augments MLIPs with In-Context Learning (ICL) for context-aware energy prediction. Specifically, we attach Tabular Foundation Models (TFMs) as contextual predictors on top of pretrained MLIPs, conditioning predictions on atomic structures with previously computed energy targets without additional training or fine-tuning of the base MLIPs. Experiments across six large-scale datasets spanning diverse domains demonstrate substantial performance gains from ICL-based inference. To address the unique challenges arising when applying TFMs to IC-MLIPs, we further introduce Structure-Global Information Enrichment (SGIE) to enrich atom-wise MLIP representations with structure-global information, and Dynamic Context Construction to adaptively construct effective contexts during TFMs training. Together, these components yield further performance improvements.

1 INTRODUCTION

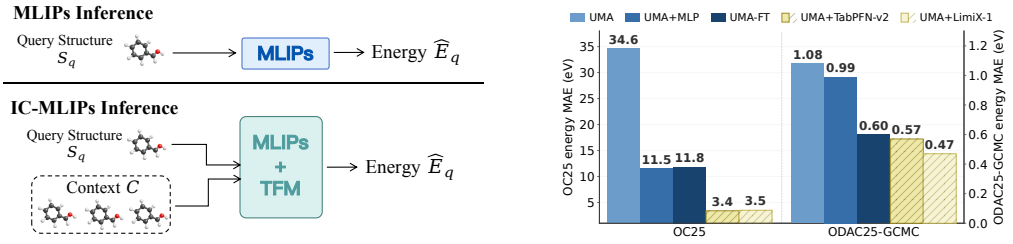
Density Functional Theory (DFT) provides accurate quantum-mechanical descriptions of atomic systems but is computationally expensive. MLIPs offer an efficient alternative by learning to approximate DFT from 3D atomic structures, with recent models such as MatterSim Yang et al. (2024) and UMA Wood et al. (2026) leveraging large-scale pretraining across diverse atomic systems.

Despite these advances, existing MLIPs largely follow an **independent inference paradigm**, predicting energy of each query structure in isolation even when DFT energies for other structures are already available. This independent inference paradigm leaves available DFT observations unused, limiting potential performance gains. A few attempts Fifty et al. (2023); Li et al. (2026) explore ICL for molecular or materials property prediction, showing that contextual information can improve prediction. However, they focus on small-scale, domain-specific settings and do not address context construction for atomic-structure data, which lack predefined contexts.

We introduce **In-Context Machine Learning Interatomic Potentials (IC-MLIPs)**, which augment pretrained MLIPs with ICL for context-aware energy prediction. Given a query structure and a set of structure–energy pairs, IC-MLIPs condition the prediction on these contextual observations, enabling performance improvements without updating the base MLIPs parameters, as shown in Figure 1a. This new inference paradigm differs from independent inference: while existing MLIPs predict each query independently, IC-MLIPs make predictions in a context-aware manner.

To realize large-scale, cross-domain IC-MLIPs inference, we use a frozen MLIP as a atomic structure encoder and a Tabular Foundation Model (TFM) as a contextual predictor. MLIP representations and energy labels are organized as tabular feature–label examples for the TFM. We choose TFMs for their powerful, well-established ICL capabilities and relatively lightweight architectures Zhang et al. (2025); Grinsztajn et al. (2025). The TFM can either be directly applied for ICL using its pretrained parameters (**TFM-ICL**) or continually pretrained on 3D atomic-structure tables before ICL (**TFM-**

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.



(a) (b)

Figure 1: (a) Comparison of MLIPs and IC-MLIPs. (b) Effectiveness of IC-MLIPs.

CPT), adapting the contextual predictor to atomic-structure data while retaining the broad coverage of the pretrained MLIPs. Table 1 summarizes these two IC-MLIPs settings alongside MLIPs direct inference and fine-tuning.

Aspect	MLIP		IC-MLIP	
	Direct	MLIP-FT	TFM-ICL	TFM-CPT
Inference paradigm	Independent	Independent	In-context	In-context
Prediction inputs	query structures	query structures	query and context structures	query and context structures
Observations utilization	✗	✓	✓	✓
Parameter updates	None	MLIP	None	TFM only

Table 1: Comparison of MLIPs and IC-MLIPs.

Figure 1b presents a validation comparison between two inference paradigms on OC25 and ODAC25-GCMC datasets: conventional MLIPs baselines based on UMA (UMA, UMA-FT, UMA+MLP) and TFM-ICL variants (UMA+TabPFN, UMA+LimiX). The context-aware variants achieve considerably lower MAEs than UMA baselines, with no additional training on atomic-structure data. The performance advantage is most prominent on OC25, which was withheld from UMA pretraining, highlighting the effectiveness of the TFM-ICL.

However, TFM-ICL and TFM-CPT face two key challenges: **(C1) Representation-side challenge**: limited structure-global information in MLIP representations for TFM contextual attention; and **(C2) Context-side challenge**: the lack of a principled context construction mechanism during TFM training, as atomic-structure data lack an intrinsic context organization as in tabular data, and mismatched contexts can substantially degrade TFM performance.

To address these challenges, we introduce two complementary components: *Structure-Global Information Enrichment* (SGIE), which enriches MLIP representations with structure-global information, and *Dynamic Context Construction*, which adaptively constructs contexts during the training of TFM-CPT. Experiments across six large-scale datasets, including a transfer setting excluded from MLIP pretraining, demonstrate that TFM-ICL reduces MAE by up to 87.6% over conventional MLIP inference while maintaining a comparable inference throughput, and TFM-CPT reduces MAE by up to 69.9% over MLIP fine-tuning. Ablation studies verify that the proposed SGIE enhances the contextual attention of TFMs, and Dynamic Context Construction strengthens the model's robustness against mismatched context. These two core designs jointly optimize the context-aware inference and effectively yield substantial performance improvements.

2 Related Work

Machine learning interatomic potentials. Machine Learning Interatomic Potentials (MLIPs) efficiently approximate first-principles DFT calculations by learning potential energy surfaces, with recent advances moving toward universal models pretrained on large and diverse datasets Jacobs et al. (2025); Yuan et al. (2026); Lysogorskiy et al. (2026). Despite this progress, MLIPs typically follow a structure-independent inference paradigm, predicting each atomic structure individually without incorporating previously computed DFT observations Jacobs et al. (2025); Yuan et al. (2026); Kim & Cheng (2026). Existing efforts mainly improve representations, data coverage, and pretraining Kim & Cheng (2026); Allam et al. (2026); Ikeda et al. (2026); Yuan et al. (2026); Lyso-

gorskiy et al. (2026). We instead explore a complementary direction by using previously computed DFT observations as inference-time context for context-aware prediction.

In-context learning for scientific prediction. Recent studies explore ICL for molecular and biological prediction, including molecular property regression Fifty et al. (2023); Li et al. (2026); Joe et al. (2026), drug design Deng et al. (2025), and antibody affinity ranking Chen et al. (2026). These studies demonstrate the potential of leveraging contextual observations for scientific prediction, but focus on task-specific settings and representations such as SMILES, assay records, or protein structures. In contrast, we investigate ICL for MLIPs, where 3D atomic structures span diverse domains, including molecules, catalysts, and adsorbates, toward general-purpose energy prediction.

Tabular foundation models. Tabular Foundation Models (TFMs) have emerged as pretrained models that perform prediction by conditioning on labeled examples provided at inference time, enabling adaptation to new tasks without task-specific parameter updates. On the one hand, recent developments have explored scaling model size and pretraining data, as exemplified by TabICLv2 Qu et al. (2026), TabDPT Ma et al. (2026) and TabStar Arazi et al. (2026). On the other hand, a series of lightweight TFMs have been developed for efficient and practical in-context prediction, with representative models including TabPFN-v2 Hollmann et al. (2022) and LimiX-1 Zhang et al. (2025). Their strong in-context learning capabilities and relatively compact architectures make them suitable contextual predictors for IC-MLIPs, where repeated prediction over context–query is required.

3 Problem Formulation

Let $\mathcal{S} = \{S_1, \ldots, S_{|\mathcal{S}|}\}$ denote a collection of atomic structures, where each structure $S = (X, R)$ consists of atomic numbers $X \in \mathbb{R}^N$ and 3D coordinates $R \in \mathbb{R}^{N\times 3}$ for its N atoms, with DFT energy $E \in \mathbb{R}$. A pretrained MLIPs encoder Φ maps S to d-dimensional atom-wise embeddings $H \in \mathbb{R}^{N\times d}$, where $H_j \in \mathbb{R}^d$ denotes the embedding of atom j. Conventional MLIPs perform independent inference with a predictor f_{ind} that outputs per-atom energies and aggregates them:

$$H = \Phi(S), \quad e_j = f_{\text{ind}}(H_j), \quad \hat{E} = \sum_{j=1}^{N} e_j. \tag{1}$$

In contrast, TFM-ICL and TFM-CPT perform context-aware inference based on a context of previously computed DFT observations $\mathcal{C}$ and structure-level embedding $Z \in \mathbb{R}^d$:

$$Z = \sum_{j=1}^{N} H_j, \quad \mathcal{C} = (Z_k, E_k)_{k=1}^{|\mathcal{C}|}, \quad \hat{E} = f_{\text{tfm}}(Z, \mathcal{C}), \tag{2}$$

where f_{tfm} denotes a TFM and the MLIPs encoder Φ remains frozen.

4 Tabular Foundation Model based IC-MLIP Framework

TFM-ICL conducts direct in-context prediction with pretrained TFM and MLIP, while TFM-CPT uses a TFM continually pretrained on 3D atomic-structure tables. Given a context $\mathcal{C}$ and a query set $\mathcal{Q} = (Z_q, E_q)_{q=1}^{|\mathcal{Q}|}$ for training, the continue pretraining loss of a TFM model g is defined as

$$\hat{E}_q = g(Z_q, \mathcal{C}), \qquad \mathcal{L}_{\text{energy}} = \frac{1}{|\mathcal{Q}|} \sum_{q=1}^{|\mathcal{Q}|} \left| \hat{E}_q - E_q \right|. \tag{3}$$

To address the two challenges identified above, we introduce two complementary components for adapting TFM-based inference to 3D atomic structures. *Structure-Global Information Enrichment* (SGIE) addresses the representation-side challenge (**C1**) by enriching MLIP representations with structure-global information, while *Dynamic Context Construction* addresses the context-side challenge (**C2**) via our proposed ClusterDrift strategy. ClusterDrift adaptively updates contextual samples during CPT to improve the mismatched context robustness of TFMs. The overall framework is illustrated in Figure 2.

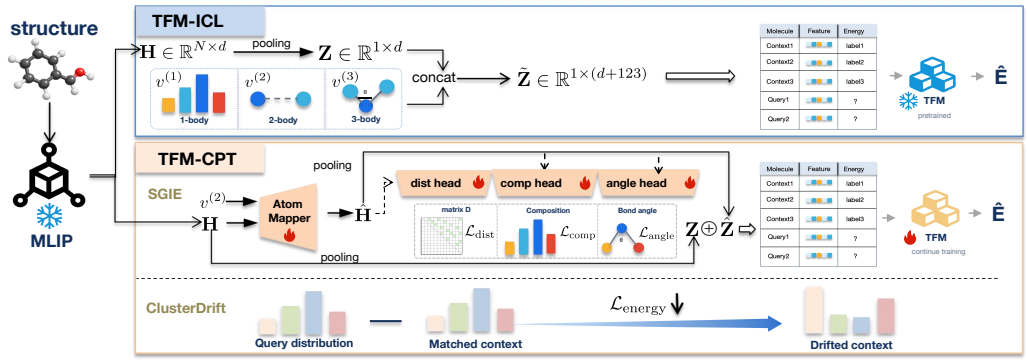


Figure 2: Overview of the proposed TFM based IC-MLIP framework with TFM-ICL and TFM-CPT. TFM-CPT further incorporates SGIE and Dynamic Context Construction

4.1 Structure-Global Information Enrichment

In TFM-based ICL, the query representation interacts with context representations via attention to incorporate contextual information into prediction Hollmann et al. (2022); Zhang et al. (2025). Since each representation corresponds to a structure, effective contextual inference requires informative structure-level representations for contextual attention mechanism of TFMs. However, MLIPs encoders primarily produce atom-wise representations that characterize individual atoms and their local environments. Simply sum-pooling atom-wise representations does not encode information describing the structure as a whole. Therefore, we propose *Structure-Global Information Enrichment* (SGIE) to enrich MLIP representations with structure-global descriptors. For TFM-ICL, these descriptors are directly concatenated with the sum-pooled MLIP representations, avoiding parameter updates to MLIPs and TFMs; for TFM-CPT, a lightweight Self-Supervised Learning (SSL) module jointly trained with the TFM learns structure-global representations by regressing these descriptors.

Structure-global descriptors. For each structure S, SGIE computes one-/two-/three-body descriptors, encoding complementary structure-global information at composition-, pair-, and angle-level granularity under the standard many-body hierarchy(Huo & Rupp, 2022; Drautz, 2019; Behler, 2011). We restrict the descriptors to three-body statistics to balance structural expressiveness and computational cost.

The one-body descriptor $v^{(1)} \in \mathbb{R}^{57}$ encodes global compositional and structural statistics. The first 54 dimensions represent the elemental fractions for atomic numbers 1–54, covering the elements from H to Xe while avoiding dimensional sparsity induced by rarely occurring heavy elements. The remaining three dimensions correspond to the normalized structure size $\log(1+N)/\log(1+1200)$, the non-hydrogen atom fraction N_{nonH}/N, and the ratio of the number of distinct elements to the total number of atoms, N_{elem}/N, respectively.

For ICL, $v^{(2)} \in \mathbb{R}^{30}$ is constructed as a normalized histogram of pairwise distances $d_{ij} \leq 6$Å, with the histogram entries summing to one. For CPT, the two-body representation is given by an upper-triangular matrix $D \in \mathbb{R}^{N\times N}$, where $D_{ij} = 1/d_{ij}$ for $i < j$ and $d_{ij} \leq 6$Å, and $D_{ij} = 0$ otherwise. Here, d_{ij} denotes the distance between atoms i and j, computed under the minimum image convention for periodic systems. We adopt the 6Å cutoff used by UMA Wood et al. (2026) to maintain consistency with the spatial scale of its atomic representations.

The three-body descriptor $v^{(3)} \in \mathbb{R}^{36}$ is a normalized histogram of bond angles over $[0°, 180°]$ with 36 bins. For each center atom j, an angle i–j–k is included when both neighboring atoms satisfy $d_{ij} \leq 6$Å and $d_{jk} \leq 6$Å. The resulting histogram is normalized to sum to one.

SGIE for ICL. SGIE is applied to both context and query structures before they are jointly provided to the frozen TFM for ICL-based energy prediction. Specifically, the pooled MLIP representation $Z \in \mathbb{R}^d$ is directly concatenated with descriptors to obtain

$$\tilde{Z} = \left[Z;\, v^{(1)};\, v^{(2)};\, v^{(3)}\right] \in \mathbb{R}^{d+123}. \tag{4}$$

SGIE for CPT. During TFM continued pretraining (CPT), SGIE learns structure-global information from MLIP representations through self-supervised learning (SSL). Given atom-wise representations H and the two-body descriptor $v^{(2)}$, a lightweight mapper produces an enhanced representation $\hat{H}$. The enhanced representation is trained to reconstruct three descriptors, providing energy-independent structural supervision beyond the original MLIP representation. Specifically, $\hat{H}$ is used to reconstruct the three descriptors through dedicated heads f_{comp}, f_{dist}, and f_{angle}:

$$\begin{aligned} \hat{H} = f_{\text{map}}(H, v^{(2)}), \quad H' = f_{\text{dist}}(\hat{H}), \quad \hat{D} = H'H'^{\mathsf{T}}, \quad \hat{H}, H' \in \mathbb{R}^{N\times d}, \\ \hat{Z} = \sum_{i=1}^{N} \hat{\mathbf{h}}_i, \quad \hat{v}^{(1)} = f_{\text{comp}}(\hat{Z}), \quad \hat{v}^{(3)} = f_{\text{angle}}(\hat{Z}), \quad \hat{Z} \in \mathbb{R}^{d}. \end{aligned} \tag{5}$$

The corresponding reconstruction objectives are jointly defined as

$$\mathcal{L}_{\text{SSL}} = \frac{1}{3}\left(\left\|v^{(1)} - \hat{v}^{(1)}\right\|_1 + \left\|D - \hat{D}\right\|_1 + \left\|v^{(3)} - \hat{v}^{(3)}\right\|_1\right). \tag{6}$$

The trained structure-global representation $\hat{Z}$ is fused with the original MLIP representation Z and provided to the TFM. The SSL objective is incorporated into TFM-CPT as

$$Z_{\text{fused}} = f_{\text{fuse}}(Z, \hat{Z}), \qquad \mathcal{L}_{\text{CPT}} = \mathcal{L}_{\text{energy}} + \lambda_{\text{SSL}}\left(\mathcal{L}_{\text{SSL}} + \mathcal{L}_{\text{cos}}\right), \tag{7}$$

where f_{fuse} denotes the fusion module based on MLP and $\mathcal{L}_{\text{cos}}$ is a bounded cosine regularization between $\hat{Z}$ and the frozen base representation Z. Detailed designs are provided in the appendix C.

4.2 Dynamic Context Construction

Unlike tabular data, atomic-structure datasets do not provide predefined contexts for query structures, and the composition of the context affects ICL performance Rubin et al. (2022). For example, a context dominated by metallic structures may provide less useful information for a query set dominated by non-metallic structures. To address this challenge, we propose *ClusterMatch* to construct query-aligned contexts first, and further develop *ClusterDrift* to progressively introduce distributional mismatch during training, improving robustness to such mismatch at inference time.

ClusterMatch: Query-aligned context construction. ClusterMatch aligns the cluster composition of context structures with that of the query structures. Specifically, we first partition the training structures into K clusters using their fixed pretrained MLIP embeddings. For each query table for CPT, let p_k^q denote the proportion of query structures assigned to cluster k, which is used as the target probability for sampling context structures from the entire training pool:

$$p_k^q = p_k^{\text{ctx}} = \frac{1}{N_q}\sum_{i=1}^{N_q} \mathbb{I}(k_i = k), \qquad k = 1, \ldots, K. \tag{8}$$

Here, N_q is the number of query structures and k_i denotes the cluster assignment of the i-th query structure. We then sample a fixed-size context from the training pool according to p_k^{ctx}.

ClusterDrift: Context construction for inference robustness. To improve robustness against contexts mismatch, we progressively increase the mismatch between context and query as CPT loss decreases, forcing TFMs to learn drift-robust contextual reasoning. Building on ClusterMatch, ClusterDrift controls this process according to the relative reduction in CPT loss. Given the initial CPT loss L_0 and the loss ℓ_t of current epoch t, we measure training progress as

$$\begin{aligned} \delta_t = \frac{L_0 - \ell_t}{L_0}, \qquad \tilde{\alpha}_t = \alpha_{\max}\,\text{clip}\left(\frac{\delta_t}{s}, 0, 1\right), \\ \alpha_t = \max(\alpha_{t-1}, \tilde{\alpha}_t), \qquad p_k^{\text{tgt}} \propto \left(p_k^q\right)^{1-2\alpha_t}. \end{aligned} \tag{9}$$

where α_t is drift coefficient and $\alpha_{\max}$ controls the maximum drift and s determines the loss-reduction scale at which it is reached. Thus, a larger reduction in CPT loss leads to a larger drift coefficient. When $\alpha_t = 0$, ClusterDrift reduces to ClusterMatch; as α_t increases, the context distribution gradually departs from the query distribution, providing progressively more diverse and mismatched contexts while keeping the context size fixed.

Table 2: Performance Comparison of MLIPs and TFM-ICL (mean ± std [10^{-2}]).

Type	Models	OMol25	OMC25	ODAC25 GCMC	ODAC25 MOF	ODAC25 MOF+Ads	OC25
MLIPs-Direct	eSEN-s-OMC25	—	0.105±0.02	—	—	—	—
	eSEN-sm-ODAC25	—	—	2.392±2.7	0.929±0.3	1.105±0.5	—
	UMA-s-1.1	0.062±0.02	0.109±0.02	1.092±1.9	0.389±0.02	0.344±0.08	34.64±1.4
	UMA-s-1.2	**0.044±0.01**	0.110±0.01	1.516±3.4	0.512±0.01	0.475±0.01	—
	UMA-m-1.1	<u>0.049±0.01</u>	**0.092±0.01**	1.185±0.023	0.446±0.1	0.479±0.1	34.70±0.6
TFM-ICL	LimiX-ICL	0.354±0.2	<u>0.10±0.01</u>	**0.489±2.2**	<u>0.179±0.12</u>	<u>0.163±0.15</u>	**5.88±4.5**
	TabPFN-ICL	0.379±1.2	0.146±0.1	<u>0.866±5.4</u>	**0.092±0.27**	**0.067±0.29**	<u>6.66±5.8</u>

5 EXPERIMENTS

We evaluate TFM-ICL and TFM-CPT in terms of prediction performance and inference efficiency, followed by ablations of the proposed components and descriptors. We further analyze contextual inference through attention, mismatched-context robustness, and case studies. Additional experiments with other TFMs and parameter sensitivity analyses are provided in Appendix A and Appendix B.

5.1 EXPERIMENTAL SETUP

Datasets and evaluation. We evaluate TFM-ICL and TFM-CPT on six recent large-scale datasets: **OMol25** Levine et al. (2026), **OC25** Sahoo et al. (2025), **OMC25** Gharakhanyan et al. (2026), and three **ODAC25** subsets— **GCMC**, **MOF**, and **MOF+Ads** Sriram et al. (2025), using energy MAE in eV as the evaluation metric. TFM-CPT is performed on about 7.4M structures drawn from these datasets. **OC25** is excluded from both UMA-s-1.1 pretraining and TFM-CPT, providing a transfer setting. Detailed dataset statistics and evaluation settings are provided in Appendix C.

Models and implementation. We use frozen UMA-s-1.1 Wood et al. (2026) as the MLIP encoder and instantiate the contextual predictor with two lightweight TFMs, LimiX-1 Zhang et al. (2025) and TabPFN-v2 Hollmann et al. (2022). TFM-ICL directly uses their pretrained checkpoints, while TFM-CPT continually pretrains the TFMs on atomic-structure data with SGIE and ClusterDrift.

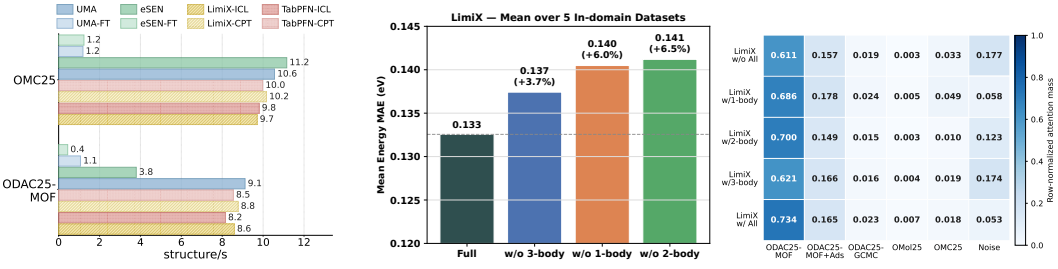
Evaluation protocol and baselines. For large-scale evaluation, each dataset is divided into non-overlapping windows, within which context structures and query structures are constructed for TFM inference. We compare against five MLIP baselines, including two dataset-specific eSEN models Fu et al. (2025) and three general-purpose models Wood et al. (2026), under Direct and Fine-tuning settings. More detailed are provided in Appendix C.

5.2 PERFORMANCE AND EFFICIENCY COMPARISON

Performance of TFM-ICL. Without any training on atomic-structure data, TFM-ICL directly applies pretrained TFMs for in-context inference, with 30% of the structures in each window randomly selected as context, substantially improving over conventional MLIP inference without updating the underlying MLIP parameters. As shown in Table. 2, the best and second-best results are indicated by **bold** and <u>underline</u>, respectively. Reported mean ± std values are computed over three fixed query splits with three different random seeds. LimiX-ICL achieves an MAE of 0.489 eV on ODAC25-GCMC, less than half of MLIPs models. The improvement is even more pronounced in the transfer setting: on the unseen OC25 dataset, both LimiX-ICL and TabPFN-ICL reduce the MAE of UMA-s-1.1 from 34.60 eV to 5.88 and 6.66 eV, respectively, reducing the error to roughly one-fifth to one-sixth of the MLIP. On OMol25, TFM-ICL models underperform UMA-s-1.1, which is pretrained on 101 millions OMol25 structures, whereas TFM-ICL has no atomic-structure pretraining and relies exclusively on in-context learning. Since TFM-ICL models use the frozen UMA-s-1.1 encoder, the improvements compared to UMA-s-1.1 suggest that the gains primarily arise from the context-aware inference paradigm, which enables previously computed DFT observations to inform predictions.

Table 3: Performance Comparison of MLIPs and TFM-CPT.

Type	Models	ctx (%)	OMol25	OMC25	ODAC25 GCMC	ODAC25 MOF	ODAC25 MOF+Ads	OC25
MLIPs-FT	eSEN-s-OMC25-FT	10	—	0.107	—	—	—	—
		30	—	0.104	—	—	—	—
		50	—	0.104	—	—	—	—
	eSEN-sm-ODAC25-FT	10	—	—	1.039	0.163	0.107	—
		30	—	—	0.558	0.141	0.096	—
		50	—	—	0.550	0.132	0.098	—
	UMA-s-1.1-FT	10	0.139	0.137	0.669	0.152	0.116	12.73
		30	0.135	0.130	0.600	0.147	0.113	11.75
		50	0.138	0.128	0.624	0.131	0.099	10.83
TFM-CPT	LimiX-CPT	10	0.123	0.075	0.288	0.140	0.131	12.63
		30	0.116	0.071	0.205	0.138	0.130	11.15
		50	0.115	0.077	0.196	0.138	0.129	10.75
	TabPFN-CPT	10	0.103	0.073	0.250	0.073	0.048	11.48
		30	0.097	0.062	**0.180**	0.069	0.047	8.88
		50	**0.095**	**0.059**	0.188	**0.068**	**0.047**	**8.04**



(a) (b) (c)

Figure 3: (a) Inference speed comparison, (b) ablation of self-supervised objectives in SGIE, (c) attention distribution study.

Performance of TFM-CPT. We systematically evaluate MLIPs-FT and TFM-CPT under three context ratios of 10%, 30%, and 50%, with the results reported in Table. 3, where **ctx (%)** denotes the ratio of context structures randomly sampled from each window. Overall, TFM-CPT generally outperforms UMA-s-1.1-FT across the six datasets and context ratios. At a 50% context ratio, for example, TabPFN-CPT reduces the MAE of UMA-s-1.1-FT by 69.9% on ODAC25-GCMC, 48.1% on ODAC25-MOF, and 52.5% on ODAC25-MOF+Ads, while achieving a 25.7% reduction on the transfer dataset OC25. These results support TFM-CPT as an effective alternative to conventional MLIP fine-tuning, achieving stronger performance across diverse datasets while keeping the underlying MLIP encoder frozen and leveraging previously computed DFT observations through context-aware inference. Additionally, we note that UMA-s-1.1-FT underperforms direct UMA-s-1.1 inference on OMol25, which we attribute to the limited context subset used for fine-tuning relative to the highly diverse OMol25 train set. We also note that TFM-CPT underperforms TFM-ICL on OC25, suggesting a trade-off between continual adaptation to atomic-structure energy prediction and preservation of the pretrained TFM's original ICL capability.

Efficiency of TFM-ICL and TFM-CPT. Figure 3a compares the inference throughput of MLIPs, MLIPs-FT, TFM-ICL and TFM-CPT, measured in structures/s. TFM-ICL and TFM-CPT achieves throughput comparable to direct MLIPs, indicating limited inference overhead from context-aware prediction. In contrast, MLIP-FT methods are substantially slower because they fine-tuning for new structures. These results show that IC-MLIPs can leverage previously computed DFT observations while largely retaining the inference efficiency of conventional MLIPs.

5.3 Ablation Experiment

Effect of SGIE and ClusterDrift. We report the performance of four TFM-CPT variants, including CPT-Only, CPT w/ SGIE, CPT w/ CD, and CPT w/ both, for both LimiX and TabPFN in

Table 4: Ablation study of Structure-Global Information Enrichment (SGIE) and ClusterDrift (CD).

Dataset	TabPFN				LimiX			
	CPT Only	CPT w/ SGIE	CPT w/ CD	CPT w/ Both	CPT Only	CPT w/ SGIE	CPT w/ CD	CPT w/ Both
OMol25	0.117	0.106	0.105	**0.097**	0.112	0.118	0.143	**0.100**
OMC25	0.058	**0.056**	0.060	0.062	0.081	0.080	0.080	**0.072**
ODAC25-GCMC	0.209	**0.171**	0.223	0.180	0.265	0.276	0.274	**0.231**
ODAC25-MOF	0.076	0.070	0.071	**0.069**	0.158	0.159	0.160	**0.149**
ODAC25-MOF+Ads	0.053	0.048	**0.043**	0.047	0.139	0.143	0.151	**0.138**
OC25	8.55	9.38	**8.00**	8.88	13.37	14.14	**12.18**	14.00

Table 5: Ablation study of structure-global descriptors.

Dataset	LimiX					TabPFN				
	w/o All	w/ 1-body	w/ 2-body	w/ 3-body	w/ All	w/o All	w/ 1-body	w/ 2-body	w/ 3-body	w/ All
OMol25	0.356	0.324	**0.323**	0.324	0.325	0.392	0.391	0.392	0.391	**0.390**
OMC25	0.094	**0.091**	**0.091**	0.096	0.093	0.146	0.146	0.145	**0.144**	0.145
ODAC25-GCMC	**0.473**	0.495	0.510	0.493	0.552	0.860	**0.857**	0.859	0.858	0.858
ODAC25-MOF	0.176	0.164	0.151	0.154	**0.138**	0.083	0.084	**0.081**	0.082	0.082
ODAC25-MOF+Ads	0.163	0.164	0.141	0.148	**0.135**	0.066	0.066	0.065	**0.064**	0.066
OC25	3.54	2.84	3.06	3.29	**2.67**	3.93	3.22	3.68	3.73	**3.14**

Table. 4. The combination of SGIE and CD provides the strongest overall performance: TabPFN w/ both achieves the best and second-best results on two datasets each, while LimiX w/ both achieves the best results on four of the five in-domain datasets. We further note that CPT w/ CD achieves the best performance on the transfer setting, OC25, for both TFM backbones. We attribute this improvement to the mismatch contexts introduced by CD during CPT, which exposes the TFM to distributionally mismatched contexts and improves its transfer capability.

Effect of Structure-Global Descriptors. Table 5 reports the performance of LimiX-ICL and TabPFN-ICL with different combinations of structure-global descriptors. The results show that the two TFMs exhibit different preferences for descriptors across datasets. For example, TabPFN-ICL w/ 2-body achieves the lowest MAE on ODAC25-MOF, whereas LimiX-ICL w/ All achieves the lowest MAE on the same dataset. Despite these dataset- and TFM-specific preferences, the w/ All variant achieves the strongest overall performance for both TFMs, indicating that combining all three descriptors provides complementary structural information.

Effect of Self-Supervised Objectives. Figure 3b compares the average MAE of LimiX across the five datasets for three variants that individually remove the 1-body, 2-body, or 3-body objective. The full model achieves the lowest MAE of 0.133, compared with 0.140, 0.141, and 0.137 when removing 1-body, 2-body, and 3-body supervision, respectively, indicating that the three complementary objectives jointly provide more effective structural supervision for SGIE.

5.4 Analysis of Contextual Inference

Attention analysis of structure-global information. To examine whether SGIE improves TFM contextual attention, we analyze the sample attention of the LimiX-ICL under a mixed-domain context, as shown in Figure 3c. Using ODAC25-MOF structures as queries, we construct a mixed-domain context comprising structures from five datasets, together with a perturbed MOF context obtained by injecting random noise into the input features while keeping the corresponding energy labels unchanged. Without structure-global information, LimiX assigns 0.177 attention mass to the perturbed MOF context. Adding any individual structure-global descriptor consistently reduces this attention, while combining all descriptors produces the strongest effect, reducing it to 0.053 and increasing attention to the clean MOF context from 0.611 to 0.734. This observation demonstrates that structure-global information improves cross-structure comparison, effectively addressing the **(C1) Representation-side challenge** by enabling more informative contextual attention.

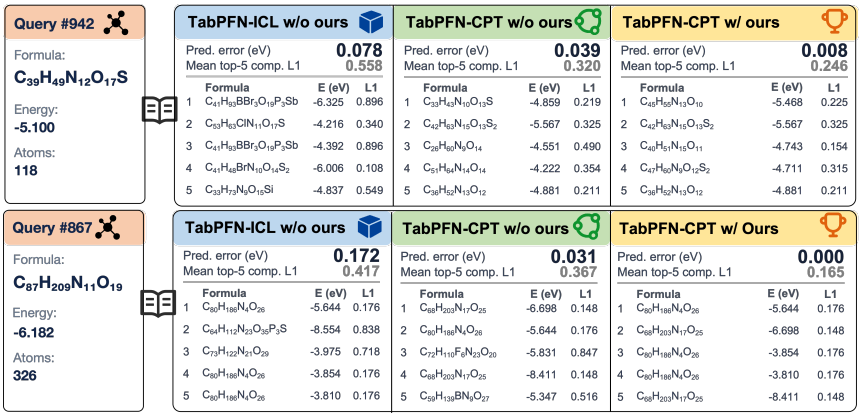


Figure 5: Case study of context attention for query structures across TabPFN LimiX variants.

Robustness to mismatched context. We examine whether ClusterDrift improves contextual inference under mismatched context. We conduct the analysis on 60,000 structures from each of OMol25 and ODAC25-GCMC, constructing contexts with increasing cluster mismatch from the query distribution while keeping the query set and context size fixed. We compare TFM-CPT trained with randomly constructed contexts against TFM-CPT with ClusterDrift. As shown in Figure 4, on OMol25, the random-context model degrades from 0.090 to 0.095 MAE as the mismatch coefficient increases, whereas the dynamically trained model remains stable. On ODAC25-GCMC, ClusterDrift likewise achieves lower MAE across all mismatch coefficients, reaching 0.16 compared with 0.20 for random context at the largest mismatch. Overall, ClusterDrift improves the robustness of TFM to mismatched context, effectively addressing the **(C2) Context-side challenge**.

Case study. We further provide a qualitative case study to illustrate how our proposed components affect contextual inference. For two representative OMol25 queries, we identify the top-5 context structures receiving the highest attention during inference under three TabPFN variants. After incorporating our components, the top-5 attended structures become more similar to the query in composition, as reflected by a decreasing mean composition L1 distance, accompanied by lower prediction error. For instance, for Query#942, TabPFN-CPT w/ ours identifies the most compositionally similar top-5 attended structures, achieving the lowest mean composition L1 distance of 0.246 and the lowest prediction MAE of 0.008. This suggests that the proposed components help TFM identify relevant context structures for a query, enabling more accurate energy prediction.

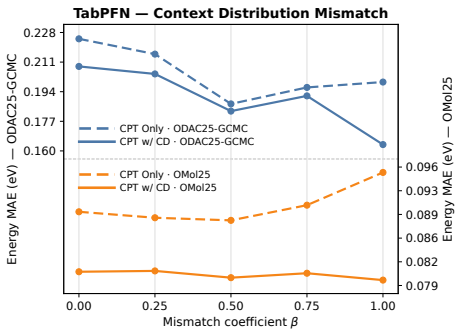


Figure 4: Mismatched context comparison on OMol25 and ODAC25-GCMC.

6 Conclusion

In this paper, we introduce **In-Context MLIPs (IC-MLIPs)**, a new inference paradigm for MLIPs that leverages previously computed DFT observations for context-aware energy prediction, instantiated with lightweight TFMs as contextual predictors. We propose **Structure-Global Information Enrichment (SGIE)** to address limited structure-global information in MLIP representations and **Dynamic Context Construction** to address the lack of principled context construction, improving contextual attention and robustness to mismatched context, respectively. Experiments across six large-scale datasets, including a transfer setting excluded from MLIP pretraining, show that TFM-ICL and TFM-CPT reduce MAE by up to 87.6% and 69.9% over conventional MLIP inference and fine-tuning, respectively.

7 AI USE STATEMENT

In this work, we used generative AI tools for assist in the writing of proofs, providing feedback on research methodology or experiments, assist with translation, support qualitative and thematic data analysis. We have not used generative AI tools for formulating mathematical claims, providing critical ingredients for proving mathematical claims, designing feedback on research methodology or experiments, implementing methods, and generating synthetic data sets, help develop theoretical models or conceptual frameworks, proposing or refining hypotheses, cleaning and reformatting dataset, supporting qualitative and thematic data analysis, interpreting results, are not applicable to this work. Additionally, we used generative AI tools for summarizing existing literature, sourcing/searching for information. We have reviewed all AI-assisted work. we checked LLM-generated literature summarizations and information through a manual literature survey. LLM-generated translation was verified and tested for correctness by 2 authors. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

8 REPRODUCIBILITY STATEMENT

We provide detailed descriptions of dataset, evaluation protocols, model configurations, training objectives, and hyperparameter settings in the appendix C. All methods are evaluated with fixed support/query splits under a unified protocol. The pretrained MLIP encoder remains frozen, while only the proposed adaptation modules and TFM predictor are optimized. We further report sensitivity analyses for key design choices, including descriptor configurations, evaluation window size, and clustering settings in the appendix A. Code, configuration files, and evaluation scripts will be released upon publication.

REFERENCES

Omar Allam, Brook Wander, SungYeon Kim, Rudi Plesch, Tyler Sours, Jia-Min Chu, Thomas Ludwig, Jiyoon Kim, Rodrigo Wang, Shivang Agarwal, et al. Aqcat25: unlocking spin-aware, high-fidelity machine learning potentials for heterogeneous catalysis. *npj Computational Materials*, 2026.

Alan Arazi, Eilam Shapira, and Roi Reichart. Tabstar: A tabular foundation model for tabular data with text fields. *Advances in Neural Information Processing Systems*, 38:172108–172161, 2026.

Jörg Behler. Atom-centered symmetry functions for constructing high-dimensional neural network potentials. *The Journal of Chemical Physics*, 134(7):074106, 2011. doi: 10.1063/1.3553717.

Zhiyuan Chen, Jing Hu, Junzhe Wang, Yueyang Huang, Xinyi Yang, Zhaoyang Wang, and Feng Zhu. Abicl: In-context learning for antigen-specific antibody affinity ranking. *arXiv preprint arXiv:2607.05846*, 2026.

Yifan Deng, Spencer S Ericksen, and Anthony Gitter. Assay2mol: large language model-based drug design using bioassay context. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pp. 27325–27350, 2025.

Ralf Drautz. Atomic cluster expansion for accurate and transferable interatomic potentials. *Physical Review B*, 99(1):014104, 2019. doi: 10.1103/PhysRevB.99.014104.

Christopher Fifty, Jure Leskovec, and Sebastian Thrun. In-context learning for few-shot molecular property prediction. *arXiv preprint arXiv:2310.08863*, 2023.

Xiang Fu, Brandon M Wood, Luis Barroso-Luque, Daniel S. Levine, Meng Gao, Misko Dzamba, and C. Lawrence Zitnick. Learning smooth and expressive interatomic potentials for physical property prediction. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=R0PBjxIbgm.

Vahe Gharakhanyan, Luis Barroso-Luque, Yi Yang, Muhammed Shuaibi, Kyle Michel, Daniel S. Levine, Misko Dzamba, Xiang Fu, Meng Gao, Xingyu Liu, Haoran Ni, Keian Noori, Brandon M. Wood, Matt Uyttendaele, Arman Boromand, C. Lawrence Zitnick, Noa Marom, Zachary W.

Ulissi, and Anuroop Sriram. Open molecular crystals 2025 (omc25) dataset and models. *Scientific Data*, 13:354, 2026. doi: 10.1038/s41597-026-06628-2.

Léo Grinsztajn, Klemens Flöge, Oscar Key, Felix Birkel, Philipp Jund, Brendan Roof, Benjamin Jäger, Dominik Safaric, Simone Alessi, Adrian Hayler, et al. Tabpfn-2.5: Advancing the state of the art in tabular foundation models. *arXiv preprint arXiv:2511.08667*, 2025.

Noah Hollmann, Samuel Müller, Katharina Eggensperger, and Frank Hutter. Tabpfn: A transformer that solves small tabular classification problems in a second. *arXiv preprint arXiv:2207.01848*, 2022.

Haoyan Huo and Matthias Rupp. Unified representation of molecules and crystals for machine learning. *Machine Learning: Science and Technology*, 3(4):045017, 2022. doi: 10.1088/2632-2153/aca005.

Yuji Ikeda, Xiang Xu, Pranav Kumar, Jörg Neugebauer, Li-Fang Zhu, Xi Zhang, Fritz Körmann, and Blazej Grabowski. Atomistic simulations of high-entropy alloys: from density functional theory to machine-learning interatomic potentials. *Journal of Materials Science*, pp. 1–47, 2026.

Ryan Jacobs, Dane Morgan, Siamak Attarian, Jun Meng, Chen Shen, Zhenghao Wu, Clare Yijia Xie, Julia H Yang, Nongnuch Artrith, Ben Blaiszik, et al. A practical guide to machine learning interatomic potentials–status and future. *Current Opinion in Solid State and Materials Science*, 35:101214, 2025.

Chan Young Joe, Kyungwoo Song, and Rakwoo Chang. Evaluating in-context learning in large language models for molecular property regression. *Journal of Computational Chemistry*, 47(2): e70308, 2026.

Dongjin Kim and Bingqing Cheng. Long-range electrostatics for machine learning interatomic potentials is easier than we thought. *The Journal of chemical physics*, 164(6), 2026.

Daniel S. Levine, Muhammed Shuaibi, Evan Walter Clark Spotte-Smith, Michael G. Taylor, Muhammad R. Hasyim, Kyle Michel, Ilyes Batatia, Gábor Csányi, Misko Dzamba, Peter Eastman, Nathan C. Frey, Xiang Fu, Vahe Gharakhanyan, Aditi S. Krishnapriyan, Joshua A. Rackers, Sanjeev Raja, Ammar Rizvi, Andrew S. Rosen, Zachary Ulissi, Santiago Vargas, C. Lawrence Zitnick, Samuel M. Blau, and Brandon M. Wood. The open molecules 2025 (omol25) dataset, evaluations, and models, 2026.

Qinyang Li, Rongzhi Dong, Nicholas Miklaucic, Jeffrey Hu, Sadman Sadeed Omee, Lai Wei, Sourin Dey, Ming Hu, and Jianjun Hu. In context learning foundation models for materials property prediction with small datasets. *npj Computational Materials*, 2026.

Stuart P. Lloyd. Least squares quantization in pcm. *IEEE Transactions on Information Theory*, 28 (2):129–137, 1982. doi: 10.1109/TIT.1982.1056489.

Yury Lysogorskiy, Anton Bochkarev, and Ralf Drautz. Graph atomic cluster expansion for foundational machine learning interatomic potentials. *npj Computational Materials*, 12(1):114, 2026.

Junwei Ma, Valentin Thomas, Rasa Hosseinzadeh, Alex Labach, Jesse Cresswell, Keyvan Golestan, Guangwei Yu, Anthony L Caterini, and Maks Volkovs. Tabdpt: Scaling tabular foundation models on real data. *Advances in Neural Information Processing Systems*, 38:172692–172722, 2026.

Jingang Qu, David Holzmüller, Gaël Varoquaux, and Marine Le Morvan. TabICLv2: A better, faster, scalable, and open tabular foundation model. *arXiv preprint arXiv:2602.11139*, 2026.

Ohad Rubin, Jonathan Herzig, and Jonathan Berant. Learning to retrieve prompts for in-context learning. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 2655–2671. Association for Computational Linguistics, 2022. doi: 10.18653/v1/2022.naacl-main.191.

Sushree Jagriti Sahoo, Mikael Maraschin, Daniel S. Levine, Zachary Ulissi, C. Lawrence Zitnick, Joel B. Varley, Joseph A. Gauthier, Nitish Govindarajan, and Muhammed Shuaibi. The open catalyst 2025 (oc25) dataset and models for solid-liquid interfaces. *arXiv preprint arXiv:2509.17862*, 2025.

David Sculley. Web-scale k-means clustering. In *Proceedings of the 19th International Conference on World Wide Web*, pp. 1177–1178, 2010. doi: 10.1145/1772690.1772862.

Anuroop Sriram, Logan M. Brabson, Xiaohan Yu, Sihoon Choi, Kareem Abdelmaqsoud, Elias Moubarak, Pim de Haan, Sindy Löwe, Johann Brehmer, John R. Kitchin, Max Welling, C. Lawrence Zitnick, Zachary Ulissi, Andrew J. Medford, and David S. Sholl. The open dac 2025 dataset for sorbent discovery in direct air capture. *arXiv preprint arXiv:2508.03162*, 2025.

Brandon Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John Kitchin, Daniel Levine, et al. Uma: A family of universal models for atoms. *Advances in Neural Information Processing Systems*, 38:129391–129427, 2026.

Han Yang, Chenxi Hu, Yichi Zhou, Xixian Liu, Yu Shi, Jielan Li, Guanzhi Li, Zekun Chen, Shuizhou Chen, Claudio Zeni, et al. Mattersim: A deep learning atomistic model across elements, temperatures and pressures. *arXiv preprint arXiv:2405.04967*, 2024.

Eric C-Y Yuan, Yunsheng Liu, Junmin Chen, Peichen Zhong, Sanjeev Raja, Tobias Kreiman, Santiago Vargas, Wenbin Xu, Martin Head-Gordon, Chao Yang, et al. Foundation models for atomistic simulation of chemistry and materials. *Nature Reviews Chemistry*, 10(3):212–230, 2026.

Xingxuan Zhang, Gang Ren, Han Yu, Hao Yuan, Hui Wang, Jiansheng Li, Jiayun Wu, Lang Mo, Li Mao, Mingchao Hao, et al. Limix: Unleashing structured-data modeling capability for generalist intelligence. *arXiv preprint arXiv:2509.03505*, 2025.

Xingxuan Zhang, Gang Ren, Hao Yuan, Hao Zou, Hongze Tan, Hui Wang, Jianhao Song, Jiansheng Li, Jiayao Zhang, Jinghan Zhang, Kaifang Li, Lang Mo, Li Mao, Mingchao Hao, Nuo Xu, Rui Ding, Ruiji Zhang, Shuyang Li, Siyu Mei, Tianyang Zhang, Weiyang Mu, Yancheng Dong, Yongxian Wei, Yuan Xue, Yuanrui Wang, Yue He, Zijia Yang, Ziyun Li, Dongzhe Li, Fuqiang Wang, Jiandong Liu, Jiawei Chen, Jiaxin Du, Kaijie Cheng, Kehan Li, Lei Sun, Linjun Zhou, Ningbo Dai, Qi Wang, Renzhe Xu, Shaoxing Du, Shumeng Yang, Wang Lu, Wenjing Chu, Xiannan Huang, Xiaoyu Lin, Xing Ai, Xinyan Han, Xuanyue Li, Xuanyue Su, Xukun Zhang, Yan Lu, Yaxin Zhang, Yi Qin, Yifei Huang, Yihan Xu, Yongle Lv, Yuanyuan Jiang, Yushan Han, and Peng Cui. Limix-2: A contextual mechanism network towards general structured-data intelligence, 2026.

A Parameter Sensitivity Analysis

A.1 Structure-Global Descriptor Configuration

We first examine whether TFM-ICL is sensitive to the hyperparameters used to construct structure-global descriptors. Using the pretrained LimiX with the same feature-concatenation setting as in the main experiments, we independently vary the configuration of each descriptor order: the element vocabulary ($Z54/Z83$) for one-body information, the number of pair-distance bins ($15/30/60$) with a fixed 6 Å cutoff for two-body information, and the number of angular bins ($18/36/72$) for three-body information. Each configuration is evaluated over the first four validation windows of each dataset following the $r = 0.3$ context protocol, and the resulting MAE is averaged across the four windows.

Since the absolute MAE scales differ substantially across datasets, we report the relative MAE increase with respect to the best configuration within each descriptor order:

$$\mathrm{RelInc}_{d,c} = \frac{\mathrm{MAE}_{d,c} - \min_{c' \in \mathcal{C}_g} \mathrm{MAE}_{d,c'}}{\min_{c' \in \mathcal{C}_g} \mathrm{MAE}_{d,c'}},$$

where d denotes the dataset and $\mathcal{C}_g$ contains the configurations belonging to descriptor order g. The comparison therefore measures sensitivity within each descriptor family.

As shown in Figure 6, TFM-ICL remains generally robust to the descriptor configurations across the evaluated domains. Varying the one-body element vocabulary, the binning resolution of the two-body distance descriptor, or the angular resolution of the three-body descriptor results in only limited performance variation in most cases. This suggests that the effectiveness of structure-global enrichment does not rely on a narrowly tuned descriptor configuration. Some domain-specific preferences can still be observed, such as the relatively higher sensitivity of ODAC25-GCMC to two-body binning and the preference of the MOF-related domains for finer angular resolution, but these variations do not alter the overall robustness across the tested parameter ranges.

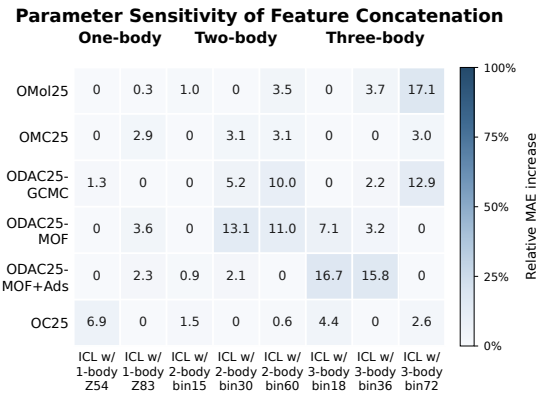


Figure 6: Sensitivity of TFM-ICL to structure-global descriptor configurations.

A.2 Sensitivity to Evaluation Window Size

We further evaluate the sensitivity to the number of structures jointly processed within each ICL window. We use the same first 60K validation structures and construct fixed 1K base partitions, which are subsequently grouped into evaluation windows of 2K, 5K, 10K, 12K, and 15K structures. The context and query ratios are fixed at 30% and 20%, respectively, such that the underlying samples and their context/query assignments remain unchanged while only the number of structures jointly processed in each ICL forward pass varies.

As shown in Table 6, both TFM backbones remain relatively stable across a broad range of window sizes on most in-domain datasets, without a consistent tendency for larger windows to improve prediction accuracy. The preferred window size varies across domains, while OC25 exhibits relatively greater sensitivity and generally favors smaller windows. Overall, these results indicate that IC-MLIP performance does not rely on a narrowly tuned evaluation window size, and increasing the amount of context alone does not necessarily improve contextual inference.

A.3 Sensitivity to the Number of Clusters

Table 6: Sensitivity to evaluation window size. Energy MAE (eV) is reported for different numbers of structures per ICL window.

Dataset	LimiX					TabPFN				
	2K	5K	10K	12K	15K	2K	5K	10K	12K	15K
OMol25	0.10	0.10	0.09	0.09	0.09	0.08	0.08	0.08	0.08	0.08
OMC25	0.05	0.06	0.07	0.07	0.08	0.04	0.04	0.06	0.06	0.06
ODAC25-GCMC	0.27	0.25	0.25	0.26	0.26	0.17	0.18	0.17	0.19	0.19
ODAC25-MOF	0.13	0.12	0.11	0.10	0.11	0.05	0.05	0.05	0.05	0.05
ODAC25-MOF+Ads	0.16	0.18	0.20	0.17	0.15	0.05	0.05	0.06	0.06	0.06
OC25	9.76	10.23	11.45	11.86	12.26	8.60	8.46	9.39	9.79	10.25

We further examine the sensitivity of Dynamic Context Construction to the number of clusters C on ODAC25-GCMC. We vary $C \in \{3, 5, 10\}$ while keeping the remaining training settings fixed. All experiments use the same ClusterDrift configuration with $\alpha_{\max} = 0.3$, $s = 2.0$, and a non-monotonic update. To isolate the effect of context construction, SGIE is excluded in this experiment.

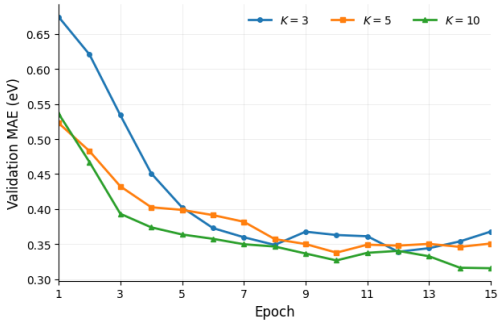


Figure 7: Sensitivity to the number of clusters K on ODAC25-GCMC. Curves show validation energy MAE with light three-epoch moving-average smoothing.

As shown in Figure 7, all three settings exhibit similar convergence behavior and reach a comparable MAE range after the initial training stage. Increasing the cluster granularity does not substantially change the overall optimization trend, although $C = 10$ achieves slightly lower validation errors during the later epochs. These results indicate that Dynamic Context Construction is generally robust to the choice of C within the evaluated range and does not rely on a narrowly tuned cluster granularity.

B Generalization to Recent Tabular Foundation Models

To examine whether IC-MLIPs generalizes beyond the TFM backbones used in the main experiments, we further evaluate TFM-ICL with two recent tabular foundation models, TabICL-v2 and LimiX-2. We follow the same TFM-ICL pipeline and compare the original UMA representation with its structure-global enriched counterpart under a 10% context ratio.

Table 7: TFM-ICL with recent TFMs under 10% context.

Dataset	TabICL-v2		LimiX-2	
	UMA	w/ SGIE	UMA	w/ SGIE
OMol25	0.310	0.311	0.101	0.0964
OMC25	0.0401	0.0390	0.0265	0.0257
ODAC25-GCMC	0.244	0.250	0.0626	0.0650
ODAC25-MOF	0.0383	0.0281	0.0151	0.0139
ODAC25-MOF+Ads	0.0198	0.0150	0.0062	0.0062
OC25	7.55	6.72	2.63	1.68

As shown in Table 7, structure-global enrichment remains effective with both newer TFMs. The improvement is particularly evident on the MOF-related domains and the transfer setting. For example, SGIE reduces the MAE from 0.0383 to 0.0281 on ODAC25-MOF and from 7.55 to 6.72 on OC25 with TabICL-v2 Qu et al. (2026), while the corresponding improvements with LimiX-2 Zhang et al. (2026) are from 0.0151 to 0.0139 and from 2.63 to 1.68. Other domains remain largely comparable to their UMA-only counterparts, with small regressions on a few settings. Overall, these results suggest that the benefit of structure-global enrichment is not tied to the specific TFM backbones adopted in the main experiments.

We further apply the complete TFM-CPT pipeline to TabICL-v2 on OMol25, using the same training data and training protocol as in the main experiments. Under the 30% context setting, TabICL-v2

obtains an MAE of **0.194** with direct ICL, which is reduced to **0.0928** after CPT with our proposed training framework, corresponding to a 52.2% reduction. This result further demonstrates that IC-MLIPs can extend beyond inference-time feature enrichment: the proposed CPT framework can also effectively adapt a newer TFM to molecular energy prediction while keeping the MLIP backbone frozen.

C Implementation Details

C.1 Datasets and Preprocessing

We construct the CPT training pool from OMol25, OMC25, and the three ODAC25 subsets, namely GCMC, MOF, and MOF+Ads. OC25 is excluded from continual pretraining and is used only as an unseen-domain transfer benchmark. For OMC25 and each ODAC25 subset, we randomly sample one million structures from the corresponding official training set. For OMol25, we use the full OMol-4M training subset containing 3,886,754 structures. From each sampled training pool, we further reserve 100K structures as an internal hold-out subset. These hold-out structures are not used for model selection or reported evaluation, since all reported models follow a fixed training budget of 15 epochs. Consequently, the effective CPT training set contains 900K structures for each of OMC25, ODAC25-GCMC, ODAC25-MOF, and ODAC25-MOF+Ads, and 3,786,754 structures for OMol25.

For all structures, we use the 128-dimensional hidden features immediately before the final energy prediction layer of UMA-s-1p1, which keeps frozen during our training. Following the official UMA preprocessing pipeline, the physical DFT total energies are transformed using the corresponding normalization and reference-energy correction to obtain the training targets. In addition, to better accommodate the prediction characteristics of TFMs, the energy labels are further z-normalized within each tabular forward pass using statistics computed from the context labels. Predictions are transformed back through the inverse normalization before computing energy MAE in eV.

C.2 Evaluation Protocol

We adopt a unified evaluation protocol for all ICL, CPT, and energy-head fine-tuning methods. For each dataset, the official validation split is partitioned into non-overlapping windows of 10,000 structures. Within each window, we perform a fixed random split with seed 42, yielding 8,000 structures as the labeled support candidate pool and 2,000 structures as queries. The same split is shared across all methods and context settings.

Given a support ratio $r \in \{0.1, 0.3, 0.5\}$, we select $k = \lfloor 10{,}000r \rfloor$ structures from the fixed support pool, corresponding to 1,000, 3,000, and 5,000 labeled context structures respectively. The query set remains fixed across different support ratios. For TFM-based methods, the selected structures serve directly as context for in-context prediction, without parameter updates at evaluation time. For the MLIP fine-tuning baselines, the same labeled structures are used to optimize the energy head while keeping their backbone frozen; the adapted head is then evaluated on the same query set and reset before processing the next window.

We report the mean absolute error (MAE) in eV on the query structures. All CPT-based models use the checkpoint obtained after a fixed training budget of 15 epochs. Dynamic Context Construction is used only during CPT training; at evaluation time, all TFM-based methods follow the fixed support protocol described above.

C.3 Training Details

All TFM-CPT models are jointly trained on the five CPT domains with a fixed budget of 15 epochs. UMA-s-1.1 remains frozen throughout training, and only the structure-global mapper, fusion and alignment modules, self-supervised heads, and the TFM energy head are optimized. Training steps from different datasets are shuffled across each epoch. Each step contains 1,000 labeled context structures and 250 query structures sampled from a single dataset, with the query structures explicitly

excluded from the context set. The overall training objective is:

$$\mathcal{L} = \mathcal{L}_{\text{energy}} + \lambda_{\text{ssl}} \left(\frac{\mathcal{L}_{\text{comp}} + \mathcal{L}_{\text{dist}} + \mathcal{L}_{\text{angle}}}{3} + \mathcal{L}_{\cos} \right).$$

where $\mathcal{L}_{\text{energy}}$ is the L1 loss on context-normalized energy targets, $\mathcal{L}_{\text{comp}}$, $\mathcal{L}_{\text{dist}}$, and $\mathcal{L}_{\text{angle}}$ denote the three structure-global self-supervised objectives, and $\mathcal{L}_{\cos}$ is the cosine regularization term. The cosine regularization constrains the directional similarity between g_{ssl} and the frozen base representation g_{base}. Let

$$c = \frac{\hat{z}^\top \operatorname{sg}(z)}{\|\hat{z}\| \|\operatorname{sg}(z)\|},$$

where $\operatorname{sg}(\cdot)$ denotes stop-gradient. We use a two-sided hinge loss

$$\mathcal{L}_{\cos} = \max(0, \tau_{\text{low}} - c) + \max(0, c - \tau_{\text{high}}), \qquad [\tau_{\text{low}}, \tau_{\text{high}}] = [0.1, 0.2].$$

Thus, $\mathcal{L}_{\cos} = 0$ when the cosine similarity lies inside this interval, and gradients are introduced only when it falls below the lower bound or exceeds the upper bound. The lower bound prevents g_{ssl} from becoming nearly orthogonal or oppositely directed to the UMA representation, which would inject information along directions largely unsupported by the base representation. The upper bound, in contrast, prevents g_{ssl} from collapsing into a scaled copy of g_{base}. This bounded similarity therefore encourages the SSL representation to remain compatible with the pretrained UMA feature space while preserving complementary composition, distance, and angular information.

We set $\lambda_{\text{ssl}} = 0.05$. Training uses AdamW with a weight decay of 10^{-4} and bfloat16 mixed precision. For LimiX, we initialize the energy head from the official LimiX-16M checkpoint and use a learning rate of 3×10^{-5} for all trainable components. For TabPFN, we initialize from the TabPFN-v2 regressor checkpoint and use a learning rate of 1×10^{-5} for all trainable components. Training is conducted on 6 compute nodes, each equipped with 8 NVIDIA RTX PRO 5000 GPUs.

C.4 Structure-Global Representation Fusion

The SSL objectives produce a structure-global representation $\hat{Z} \in \mathbb{R}^{128}$ through the shared mapper, while $Z \in \mathbb{R}^{128}$ denotes the sum-pooled frozen UMA representation. Rather than replacing the pretrained representation with $\hat{Z}$, we treat it as a complementary signal and fuse the two representations as

$$Z_{\text{fused}} = Z + v_d \odot \hat{Z} + \delta_d, \qquad \delta_d = \text{GELU}\left(W_d [Z; \hat{Z}] + b_d \right),$$

where d denotes the dataset/domain, $v_d \in \mathbb{R}^{128}$ is a learnable channel-wise scaling vector, and $[\,\cdot\,;\,\cdot\,]$ denotes concatenation. The correction branch consists of a single linear layer followed by GELU. Its parameters are zero-initialized, so the fusion starts close to the original UMA representation and gradually introduces structure-global information during training.

The two residual branches serve complementary roles. The vector term $v_d \odot \hat{Z}$ provides a low-capacity, channel-wise injection of the SSL representation, whereas δ_d allows cross-channel interactions between the pretrained and structure-global features. Both are dataset-specific to accommodate domain-dependent feature distributions while preserving the shared UMA representation space. The resulting Z_{fused} is subsequently projected by the alignment module to the input space of the corresponding TFM.

C.5 Structure-Global Descriptor Construction

TFM-ICL and TFM-CPT share the same definitions of structure-global descriptors, while using them differently. TFM-ICL directly concatenates these descriptors with the pooled MLIP representation, whereas TFM-CPT uses them to initialize the structure-global channel or as self-supervised targets. For periodic systems, all geometric quantities are computed using minimum-image-convention (MIC) distances, with a uniform cutoff of 6 Å.

The **composition descriptor** is a 57-dimensional vector. The first 54 dimensions encode the normalized elemental composition over a fixed vocabulary of 54 elements, where each entry is the fraction

of atoms of the corresponding element. Three additional features characterize the system scale and compositional complexity:

$$\frac{\log(1+N)}{\log(1+1200)}, \qquad \frac{N_{\mathrm{nonH}}}{N}, \qquad \frac{N_{\mathrm{elem}}}{N},$$

where N is the number of atoms, N_{nonH} is the number of non-hydrogen atoms, and N_{elem} is the number of distinct element types. The constant 1200 corresponds to the maximum number of atoms observed across the training structures of the five training datasets and is used to normalize the system-size feature.

The **two-body descriptor** is a normalized histogram of pairwise distances $d_{ij} \leq 6$ Å. All valid atom pairs within the cutoff are assigned to 30 uniformly spaced distance bins, and the resulting histogram entries are normalized to sum to one. The **three-body descriptor** characterizes angular statistics. For each central atom j, neighboring atoms within 6 Å are collected, and the bond angle $\angle ijk$ is computed for each neighbor pair (i, k). The resulting angles over $[0°, 180°]$ are accumulated into 36 uniformly spaced bins and normalized to form a 36-dimensional distribution.

We distinguish the molecule-level pairwise-distance descriptor from the distance-prediction objective used during CPT. The former is a 30-dimensional histogram: it is directly used as an input descriptor in TFM-ICL and initializes the structure-global channel in TFM-CPT, but is never used as a regression target. In contrast, the distance SSL task operates on individual atom pairs. For each upper-triangular pair satisfying $i < j$ and $0 < d_{ij} \leq 6$ Å, the atom-level SSL representation predicts the inverse distance $1/d_{ij}$. The composition and angle descriptors instead serve as molecule-level SSL targets after aggregating the atom-level SSL representations.

C.6 Dynamic Context Construction Details

For each training dataset, we first cluster the fixed pretrained UMA representations and keep the resulting cluster assignments fixed throughout CPT. We use $K = 5$ for OMol25, OMC25, and ODAC25-MOF+Ads, and $K = 3$ for ODAC25-GCMC and ODAC25-MOF. K-means(Lloyd, 1982). is used for clustering, with MiniBatchKMeans(Sculley, 2010) adopted for OMol25 due to its larger clustering set. At each training step, the context and query sets are sampled without overlap. Let $p_q(c)$ denote the cluster distribution of the sampled queries. ClusterMatch initially allocates context structures according to p_q, providing a query-matched context distribution.

ClusterDrift then progressively modifies this distribution according to the optimization progress of each dataset. Let L_0 denote the EMA of the z-normalized energy loss at the end of warmup and L_t the current EMA. We define the relative loss reduction as

$$\delta_t = \frac{L_0 - L_t}{L_0},$$

and map it to a drift coefficient by

$$\tilde{\alpha}_t = \alpha_{\max} \cdot \mathrm{clip}\left(\frac{\delta_t}{s},\, 0,\, 1\right),$$

where $\alpha_{\max}$ controls the maximum drift and s denotes the loss-drop span. The target context distribution is then

$$p_t^{\mathrm{target}}(c) = \frac{p_q(c)^{\,1-2\alpha_t}}{\sum_{c'} p_q(c')^{\,1-2\alpha_t}}.$$

When $\alpha_t = 0$, the target distribution recovers the query-matched distribution p_q. As α_t increases, the context allocation progressively departs from the query distribution; $\alpha_t = 0.5$ yields a uniform distribution over non-empty clusters, while larger values increasingly emphasize clusters that are less represented in the queries. The resulting target distribution determines the context quota assigned to each cluster.

We use two variants for updating α_t. Under the monotonic schedule,

$$\alpha_t = \max\left(\alpha_{t-1}, \tilde{\alpha}_t\right),$$

so that the achieved drift level does not decrease when the energy loss fluctuates. Under the non-monotonic schedule, we directly set

$$\alpha_t = \tilde{\alpha}_t,$$

allowing the context distribution to move back toward the query distribution when the energy loss increases.