Pretrained Model Representations as Acquisition Signals for Active Learning of MLIPs

Eszter Varga-Umbrich [1 2] **Shikha Surana**[† 1] **Paul Duckworth** [1] **Jules Tilly** [1] **Olivier Peltre** [1]
Zachary Weller-Davies[† 1]

Abstract

Training machine learning interatomic potentials (MLIPs) for reactive chemistry is often bottlenecked by the high cost of quantum chemical labels and the scarcity of transition state configurations in candidate pools. Active learning (AL) can mitigate these costs, but its effectiveness hinges on the acquisition rule. We investigate whether the latent space of a pretrained MLIP already contains the information necessary for effective acquisition, eliminating the need for auxiliary uncertainty heads, Bayesian training and fine-tuning, or committee ensembles. We introduce two acquisition signals derived directly from a pretrained MACE potential: a finite-width neural tangent kernel (NTK) and an activation kernel built from hidden latent space features. On reactive-chemistry benchmarks, both kernels consistently outperform fixed-descriptor baselines, committee disagreement, and random acquisition, reducing the data required to reach performance targets by an average of 38% for energy error and 28% for force error. We further show that the pretrained model induces similarity spaces that preserve chemically meaningful structure and provide more reliable residual uncertainty estimates than randomly initialised or fixed-descriptor-based kernels. Our results suggest that pretraining aligns latent-space geometry with model error, yielding a practical and sufficient acquisition signal for reactive MLIP fine-tuning.

1. Introduction

Machine-learning interatomic potentials (MLIPs) have become practical surrogates for quantum chemical calculations, enabling simulations at far lower cost than density functional theory (DFT) (Kohn et al., 1996) while retaining useful accuracy across molecular and materials systems (Jacobs et al., 2025; Wang et al., 2024; Li et al., 2025; Behler & Parrinello, 2007; Batatia et al., 2023b). Recent atomistic foundation models improve this further by learning reusable chemical and geometric representations from large datasets (Batatia et al., 2023a; Wood et al., 2026; Deng et al., 2023; Chen & Ong, 2022). However, even with strong pretraining, out-of-distribution (OOD) chemistry presents a fundamental bottleneck: accurate energies and forces require expensive reference calculations, and the model must be adapted to the specific chemical space of interest. Reactive chemistry, with its scarce transition-state configurations, is a canonical example of this challenge.

Active learning (AL) is a principled way to reduce this labelling burden (Settles, 2012). In each round, a model selects a small batch of unlabelled structures via an acquisition rule, obtains reference labels, and is fine-tuned on the enlarged training set. Sampling relevant unlabelled structures in reactive chemistry is itself difficult: molecular dynamics (MD) oversamples near-equilibrium configurations and undersamples transition states. More advanced samplers, such as metadynamics (Laio & Parrinello, 2002), umbrella sampling (Torrie & Valleau, 1977), nudged elastic band (NEB) methods (Jónsson et al., 1998; Henkelman & Jónsson, 2000; Henkelman et al., 2000), uncertainty-biased MD (Zaverkin et al., 2024), improve rare-event coverage but still encode the choices and limitations of the generator, and often require a potential accurate enough to support the exploration. Whichever generator is used, the resulting candidate pool is biased, and one must still decide which structures to label to achieve good performance on the target test distribution. We therefore study AL in the offline, pool-based setting, where a fixed candidate pool is given, and the acquisition rule alone determines labelling efficiency under this bias.

In the literature, AL for MLIPs often relies on committee disagreement or extrapolation criteria during simulation (Smith et al., 2018; Podryabinkin & Shapeev, 2017; Jinnouchi et al., 2019; Vandermause et al., 2020; Zhang et al., 2019; Achar et al., 2025; Khan et al., 2026). In contrast, representation-based batch AL treats selection as a geometry problem: construct a similarity space, then select structures that are

[†]Equal supervision.
[1]InstaDeep [2]University of Cambridge. Correspondence to: Zachary Weller-Davies <z.weller-davies@instadeep.com>.

uncertain or poorly covered in that space (Zaverkin et al., 2022; Holzmüller et al., 2023).

The central hypothesis of this work is that the latent space of a pretrained MLIP model already encodes sufficient information about model uncertainty for effective AL, removing the need for explicit uncertainty heads (Rhodes et al., 2025; Ho et al., 2025), Bayesian training and fine-tuning (Jinnouchi et al., 2019; Vandermause et al., 2020; Coscia et al., 2026), or committee ensembles (Smith et al., 2018; Schran et al., 2020; Peterson et al., 2017; Kahle & Zipoli, 2022).

We test this hypothesis primarily in pool-based active offline learning for reactive chemistry using pretrained MACE models (Batatia et al., 2023b). We introduce two model-based similarity metrics for AL: a finite-width energy neural tangent kernel (NTK) and a similarity kernel derived from hidden activation features. To our knowledge, finite-width NTK acquisition has previously been studied only for invariant-descriptor atomistic networks (Gaussian moment NNs) (Zaverkin et al., 2022), and was not adapted to SO(3)-equivariant pretrained MLIPs prior to the present work. Similarly, latent-space active learning has appeared only in a narrow materials setting (Ouyang et al., 2024) rather than in the context of modern pretrained foundation models. We compare these methods against fixed-descriptor baselines such as SOAP (smooth overlap of atomic positions) (Bartók et al., 2013; 2017; Himanen et al., 2020), Morgan fingerprints (Morgan, 1965; Rogers & Hahn, 2010; Ralaivola et al., 2005), as well as committee disagreement and random acquisition.

We evaluate primarily on the Transition1x (T1x) dataset (Schreiner et al., 2022), a reactive-chemistry benchmark containing DFT energies and forces along NEB reaction pathways, with additional evaluation on the RGD (Zhao et al., 2023) and PMechDB (Tavakoli et al., 2024) reactivity subsets of OMol (Levine et al., 2026). Our main empirical finding is that pretrained model-based kernels are the most effective acquisition signals among the methods evaluated in the reactive settings we study, and they consistently outperform random acquisition, fixed-descriptor baselines, and committee ensembles (see Figure 1 and Table 3). In particular, the NTK kernel coupled with the largest cluster maximum distance (LCMD) batch acquisition strategy reduces the number of acquisition rounds needed to reach a shared target by an average of 38.1%, 28.3%, 27.2%, and 8.3% for energy RMSE, force RMSE, energy MAE, and force MAE, respectively.

In our T1x case study, the accompanying diagnostics support a mechanistic interpretation: compared with randomly initialised neural kernels and fixed descriptors, pretrained model-based kernels yield better residual interpolation, better uncertainty calibration, and a similarity geometry that preserves both reaction-family structure and variation along reaction paths. Although we focus on a specific pretrained MACE model, our results suggest that pretrained model-based representations already encode uncertainty-relevant structure, can be used effectively for active learning, and merit broader study as practical acquisition signals for MLIPs.

2. Related Work

Active learning for MLIPs. Active learning has become a standard tool for reducing the cost of training MLIPs. Most prior work has focused on *online* active learning, in which candidate generation and acquisition are coupled: structures are generated on the fly and selected using uncertainty or extrapolation criteria. Representative examples include committee-based approaches (Smith et al., 2018; Schran et al., 2020), extrapolation-based methods for moment-tensor potentials (Podryabinkin & Shapeev, 2017), Bayesian force-field models (Jinnouchi et al., 2019; Vandermause et al., 2020; Coscia et al., 2026), concurrent learning (Zhang et al., 2019), and uncertainty-driven dynamics (Kulichenko et al., 2023). In contrast, offline pool-based active learning operates on a fixed set of candidate structures (Zaverkin et al., 2022; Zou & Marzouk, 2026), where acquisition must contend with both scale and distribution shift. This is particularly pronounced in reactive chemistry, where candidate pools are often strongly biased relative to the target distribution (Deng et al., 2025; Cui et al., 2025).

Pretrained representations as acquisition signals. Finite-width energy NTKs have previously been studied as active-learning signals in Gaussian Moment Networks (Zaverkin et al., 2022), which use invariant descriptors as inputs to a feedforward MLP. Other methods (Holzmüller et al., 2023) provide a broader benchmark for deep batch active learning in the regression setting. Likewise, uncertainty quantification based on hidden latent-space representations has been explored in Gaussian Moment Networks (Zhu et al., 2023), and related latent-space active learning has appeared in prior work (Ouyang et al., 2024), but only in a relatively narrow HfO_2 materials setting rather than in the context of modern pretrained foundation models. In a related context, recent studies have explored using hidden MLIP descriptors as inputs for downstream tasks such as property prediction (Shiota et al., 2024; Kim et al., 2026). By contrast, fixed descriptor spaces such as SOAP (Bartók et al., 2013; 2017) and fingerprint similarities (Rogers & Hahn, 2010; Ralaivola et al., 2005) provide model-independent notions of structural similarity and have also been used for atomistic data selection (Zou & Marzouk, 2026). More recently, uncertainty in MLIPs has also been approached through uncertainty-calibration or confidence heads (Tan et al., 2023; Ho et al., 2025; Rhodes et al., 2025), Bayesian networks (Coscia et al., 2026), and ensemble-based uncer-

tainty estimates. While committee-based uncertainty has been studied extensively in MLIPs, obtaining committee-style uncertainty from a pretrained model remains relatively underexplored. A notable recent exception is the multi-head committee approach of (Beck et al., 2025). Our work asks a different question: whether, for *pretrained* equivariant foundation models, uncertainty-relevant acquisition geometry can be extracted directly from the pretrained representation itself.

3. Methods

3.1. Offline Active Learning

We study pool-based active learning for machine-learned interatomic potentials. At round t, the learner has a labelled training set $\mathcal{T}^{(t)}$ and a fixed unlabelled candidate pool $\mathcal{P}^{(t)}$ containing $n_{\mathcal{P}}$ points. Here, a *candidate pool* is an unlabelled set of potential objects on which one can evaluate and decide whether to acquire the label. Each structure in the candidate pool is denoted by $\mathbf{x} = (\mathbf{z}, \mathbf{r})$, with atomic numbers $\mathbf{z} = (z_i)_{i=1}^{N(\mathbf{x})}$ and Cartesian coordinates $\mathbf{r} = (\mathbf{r}_i)_{i=1}^{N(\mathbf{x})}$, where $N(\mathbf{x})$ is the number of atoms in the structure.

The goal of a *batch acquisition rule* is to select the most informative candidates to label to efficiently improve downstream test error. For the case of MLIPs, a label for a structure $\mathbf{x}$ is the associated energies and forces of a DFT calculation

$$y = \left(E(\mathbf{x}), \{\mathbf{F}_i(\mathbf{x})\}_{i=1}^{N(\mathbf{x})} \right), \tag{1}$$

where $E(\mathbf{x}) \in \mathbb{R}$ is the total energy and $\mathbf{F}_i(\mathbf{x}) \in \mathbb{R}^3$ is the force on atom i, given by $\mathbf{F}_i(\mathbf{x}) = -\nabla_{\mathbf{r}_i} E(\mathbf{x})$.

Formally, a batch acquisition rule selects $\mathcal{A}^{(t)} \subseteq \mathcal{P}^{(t)}$ with $|\mathcal{A}^{(t)}| = B$. Reference labels are then revealed and the training and candidate pool sets are updated as

$$\mathcal{T}^{(t+1)} = \mathcal{T}^{(t)} \cup \{(\mathbf{x}, y) : \mathbf{x} \in \mathcal{A}^{(t)}\}, \tag{2}$$

and $\mathcal{P}^{(t+1)} = \mathcal{P}^{(t)} \setminus \mathcal{A}^{(t)}$.

The model is fine-tuned after each acquisition round on all of $\mathcal{T}^{(t+1)}$. All experiments use the MACE architecture (Batatia et al., 2023b) through a fork of the `mlip` library (Brunken et al., 2025). The MACE architecture is a parameterised map taking atomic structures to energy predictions $E_\theta(\mathbf{x})$, where θ denotes the parameters of the network; forces are obtained by auto-differentiation.

Models are initialised from the same internally trained SPICE-2 (Eastman et al., 2023; Levine et al., 2026) model; all labels are at the ωB97M-V/def2-TZVPD level of theory used in OMol25 (Levine et al., 2026). Further architecture and training details are given in Appendices A and B.

Table 1. Summary of acquisition methods considered in this work. Kernel-based methods first define a similarity kernel and then apply a batch selector such as posterior variance or LCMD. Direct methods assign acquisition scores without an intermediate kernel.

Method	Representation / signal	Score construction
Model-based		
Activation	Hidden activation features	Kernel-based
Energy NTK	Energy gradient features	Kernel-based
Feature-based		
SOAP	SOAP descriptor	Kernel-based
Tanimoto	Morgan fingerprint	Kernel-based (Tanimoto)
Uncertainty-based		
Committee-E	Energy disagreement	Direct score
Committee-F	Force disagreement	Direct score
Baseline		
Random	None	Uniform sampling

3.2. Acquisition Signals

The purpose of an acquisition signal is to rank candidates in the unlabelled pool and select those that are expected to most improve the model. We distinguish between two broad classes of acquisition methods.

Kernel-based methods: These imply constructing a similarity kernel over structures $k(\mathbf{x}, \mathbf{x}')$, typically from a structure-level embedding $\phi(\mathbf{x}) \in \mathbb{R}^d$, where d is the embedding dimension, and then apply a downstream batch selection rule such as posterior variance. Within the kernel-based family, we further distinguish between *model-based representations*, extracted from the pretrained force field itself, and *feature-based representations*, built from fixed molecular descriptors.

Direct uncertainty methods: These approaches assign acquisition scores directly, for example, through committee disagreement. Though a kernel can implicitly be used to quantify model-uncertainty, we think this distinction is important because a kernel provides pairwise information about similarity, coverage, and redundancy across the pool, whereas a direct uncertainty score is pointwise and does not by itself encode relations among candidates.

The central question of this work is whether pretrained *model-based representations* already contain information that is useful for AL. A summary of this taxonomy is given in Table 1, while Table 5 in Appendix C summarises the practical computational considerations for each method introduced in this section.

3.2.1. Kernel-Based Methods

For kernel-based methods, structures are first mapped to a representation $\phi(\mathbf{x})$, from which we construct a similarity kernel. For continuous embeddings, we use the cosine-

normalised kernel

$$k(\mathbf{x}, \mathbf{x}') = \tilde{\phi}(\mathbf{x})^\top \tilde{\phi}(\mathbf{x}'), \qquad \tilde{\phi}(\mathbf{x}) = \frac{\phi(\mathbf{x})}{\max(\|\phi(\mathbf{x})\|_2, \varepsilon)}. \tag{3}$$

Equation (3) is what we use in practice to construct similarity kernels between structures. The exception is for Morgan fingerprints, which are binary set-like descriptors and are therefore compared using Tanimoto similarity,

$$k_{\mathrm{Tan}}(\mathbf{x}, \mathbf{x}') = \frac{|\phi(\mathbf{x}) \cap \phi(\mathbf{x}')|}{|\phi(\mathbf{x}) \cup \phi(\mathbf{x}')|}. \tag{4}$$

We now introduce two model-based representations used to extract representations from pretrained models.

Model-based representations

Energy NTK features: The finite-width energy neural tangent kernel represents a structure by the local sensitivity of the model energy prediction (in this instance, MACE) to parameter perturbations,

$$\phi_{\mathrm{NTK}}(\mathbf{x}) = \nabla_\theta E_\theta(\mathbf{x}), \tag{5}$$

where dimension of ϕ_{NTK} is the number of parameters. Intuitively, two structures with similar NTK features cause similar perturbations to the model's predictions and therefore probe overlapping directions of parameter sensitivity. NTK feature similarity then becomes a natural measure of coverage between the candidate pool and the current training set, and a natural acquisition signal for how much a new structure would improve the model.

A key practical issue in pretrained networks is that one cannot reasonably form NTK features with respect to the entire parameter space: the resulting gradient representation is prohibitively large and too expensive for repeated scoring over a candidate pool (Zaverkin et al., 2022). It is therefore necessary to restrict the NTK to a parameter subspace θ_P. In this work, we use the embedding parameter blocks of MACE for θ_P (see Appendix A). This subset captures the model's chemical encoders explicitly, and keeps feature extraction tractable for scoring. For the MACE model used in this work, this yields a 1920-dimensional feature vector. A broader study of reduced parameter subsets is left for future work.

Activation features: As a cheaper model-dependent representation, we extract hidden MACE activations. Let $h_i^{(\ell)}(\mathbf{x})$ be the atom-wise hidden state at interaction layer ℓ. Because MACE features contain multiple rotation orders, we retain the invariant scalar channels and take the mean over atoms:

$$\phi_{\mathrm{act}}(\mathbf{x}) = \underset{\ell}{\mathrm{concat}} \frac{1}{N(\mathbf{x})} \sum_{i=1}^{N(\mathbf{x})} h_{i,L=0}^{(\ell)}(\mathbf{x}). \tag{6}$$

All experiments use a two-layer MACE model with 128 scalar hidden channels per layer, giving a 256-dimensional activation vector. Activation features are much quicker to compute than the NTK: they require only a single forward pass, whereas NTK features require reverse-mode differentiation of the energy with respect to θ_P.

Feature-based representations

SOAP: Smooth Overlap of Atomic Positions (SOAP) (Bartók et al., 2013) features are computed with `DScribe` (Himanen et al., 2020) using $r_{\mathrm{cut}} = 6$ Å, $n_{\max} = 8$, and $l_{\max} = 6$. Since SOAP features depend on the local atomic environment, we aggregate local SOAP descriptors to a structure-level representation by averaging over the power spectrum of different sites using the outer averaging mode of `DScribe`.

Morgan fingerprints: Morgan fingerprints (Morgan, 1965) use radius 3 and 2048 bits, and are paired with the Tanimoto kernel (Ralaivola et al., 2005). We compute Morgan fingerprints using the open source library `RDKit` (Landrum et al., 2025). They provide a cheap chemistry-only baseline, complementary to SOAP's local geometric descriptor and to the model-based MACE representations.

3.2.2. Direct Uncertainty Methods

Committee disagreement: Committee methods estimate uncertainty by training an ensemble of M MACE models $\{f_{\theta_m}\}_{m=1}^M$ on the current labelled set and scoring candidates by prediction disagreement. For energy acquisition (Committee-E), we use the standard deviation of predicted energies,

$$\alpha_{\mathrm{com}}^E(\mathbf{x}) = \mathrm{std}_{m=1}^M E_{\theta_m}(\mathbf{x}). \tag{7}$$

For force acquisition (Committee-F), we compute disagreement over force components. In our committee benchmarks, we use $M = 3$ models initialised from the same pretrained checkpoint, with diversity introduced by independent data-order seeds of $\mathcal{T}^{(t)}$. Committee methods provide a direct uncertainty score but increase training and evaluation cost approximately linearly with M. Prior work employs small ensembles (typically 3–5 models (Achar et al., 2025; Kulichenko et al., 2023; Khan et al., 2026)), and we adopt $M = 3$ as a minimal, cost-efficient configuration that yields a well-defined variance estimate for disagreement-based acquisition.

A practical complication in the pretrained-model setting is that meaningful ensemble diversity is difficult to obtain: all committee members begin from the same pretrained checkpoint, so diversity must be induced during fine-tuning rather than through independent pretraining. In our committee benchmarks in Appendix C.1, we explore two such

strategies: independent data-order shuffling and bootstrap resampling of the current labelled set $\mathcal{T}^{(t)}$. In expectation, each bootstrap sample contains about 63.2% unique examples from the original dataset. On T1x, we find that, under this committee construction, the shuffle variant performs better, and it is therefore the version reported in the main text. For completeness, we also consider committees trained from scratch with different initialisation seeds (Appendix C.2) and observe qualitatively similar results.

3.3. Batch Selection Rules

For kernel-based methods, the representation is first converted into a similarity kernel and then coupled to a batch selection rule. For direct uncertainty methods, such as committee disagreement, candidates are ranked directly by the acquisition score.

Posterior variance (PV): For kernel methods, the main selection rule is greedy Gaussian-posterior variance. Given a kernel k and current training set $\mathcal{T}^{(t)}$, a candidate $\mathbf{x}$ has posterior variance

$$\sigma_t^2(\mathbf{x}) = k(\mathbf{x},\mathbf{x}) - k_{\mathcal{T}^{(t)}}(\mathbf{x})^\top (k_{\mathcal{T}^{(t)}\mathcal{T}^{(t)}} + \lambda I)^{-1} k_{\mathcal{T}^{(t)}}(\mathbf{x}), \tag{8}$$

where $k_{\mathcal{T}^{(t)}}(\mathbf{x})$ is the vector of similarities between $\mathbf{x}$ and the training set. The batch is constructed greedily: after each selection, the newly selected point is added to the conditioning set before the next point is chosen (Rasmussen & Williams, 2005).

LCMD: We also use kernel coverage rules. Largest-cluster maximum-distance (LCMD) (Holzmüller et al., 2023) modifies greedy farthest-point sampling to favour dense but under-covered regions of the pool. Let S denote the set of points already selected into the current batch. At each step, every remaining candidate is assigned to its nearest centre in $S \cup \mathcal{T}^{(t)}$. For each cluster $\mathcal{C}_c$ with center $\mathbf{z}_c$, we compute the total squared distance mass

$$m_c = \sum_{\mathbf{x}\in\mathcal{C}_c} d_k^2(\mathbf{x},\mathbf{z}_c), \tag{9}$$

where the kernel-induced squared distance is

$$d_k^2(\mathbf{x},\mathbf{z}) = k(\mathbf{x},\mathbf{x}) + k(\mathbf{z},\mathbf{z}) - 2k(\mathbf{x},\mathbf{z}). \tag{10}$$

LCMD selects the cluster with the largest m_c and then adds the candidate in that cluster farthest from its centre,

$$\mathbf{x}_{\text{next}} = \arg\max_{\mathbf{x}\in\mathcal{C}_{c^\star}} d_k^2(\mathbf{x},\mathbf{z}_{c^\star}), \quad c^\star = \arg\max_c m_c. \tag{11}$$

4. Results

We build the case for pretrained model representations as acquisition signals in three steps: learning curves on a reactive-chemistry benchmark (Section 4.1), followed by two complementary diagnostics, kernel geometry (Section 4.2) and residual uncertainty calibration (Section 4.3), and finally generalisation to additional reactive datasets (Section 4.4). Together, these results support a unified interpretation: *pretraining shapes the latent space into a geometry that is already aligned with model error, making it a practical and sufficient acquisition signal*.

4.1. T1x Case Study

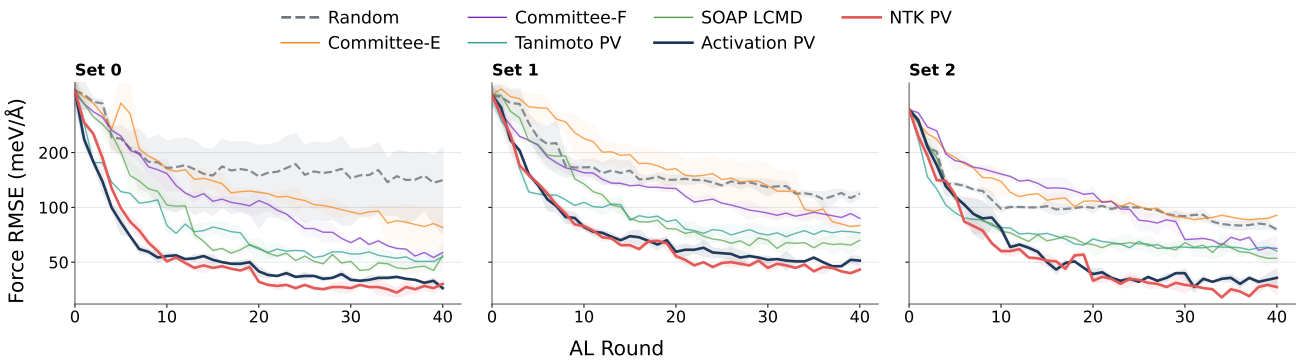
Transition1x (T1x) (Schreiner et al., 2022) contains DFT energies and forces along NEB and climbing-image NEB reaction paths. Each reaction is represented as a sequence of frames connecting reactant, transition-state-like, and product configurations, so the frame index provides a discrete reaction coordinate. This makes T1x a useful benchmark for testing whether acquisition strategies can distinguish different reaction pathways while also resolving variation along the reaction coordinate.

We construct three candidate pools, labelled Set 0, Set 1, and Set 2, containing approximately 1.7k, 4.2k, and 5.7k structures, respectively. Each pool is built from five randomly selected T1x reaction pathways, and we retain the natural imbalance of the sets, so the number of structures per reaction may differ. At each acquisition round, the active-learning method adds five new structures to the training set. All methods are evaluated on the same fixed balanced test set, containing 35 structures from each reaction pathway. Each active-learning run is repeated with two different seeds.

Figure 1 summarises the central active-learning result on T1x. For each method, we report only its best-performing selection rule; a full breakdown of results is provided in Appendix C, and the overall pattern remains the same. We also compare against training from scratch in Appendix C.2 and find the results are qualitatively the same, with all methods finding a consistent advantage in using a pretrained model.

Model-based kernel methods are the most sample-efficient acquisition signals. Activation-PV gives the best energy area under the curve (AUC), and NTK-PV gives the best force AUC, best final energy error, and best final force error among the reported methods. Both learned kernels substantially outperform random selection. Descriptor methods are useful but weaker: Tanimoto-PV improves over random, and SOAP-LCMD is competitive on final energy, but neither matches the neural kernels on force learning. Committee disagreement is less reliable, with committee-energy worse than random by force AUC, and committee-force trading improved force error for much worse energy error.

The learning-curve advantage shows pretrained represen-



Method	Energy AUC (↓)	Force AUC (↓)	Final Energy (↓)	Final Force (↓)
Random	648.33	6490.00	6.33	111.83
Committee-E	623.33	6636.67	3.50	82.50
Committee-F	915.33	5496.00	8.00	67.67
NTK (PV)	252.83	**3078.00**	**2.83**	**40.00**
Activation (PV)	**252.17**	3179.00	3.17	42.67
Tanimoto (PV)	275.33	3791.83	3.17	61.17
SOAP (LCMD)	278.83	4270.67	3.00	57.67

Figure 1. Force RMSE (meV Å^{-1}) under the natural T1x setting. Each acquisition step adds five structures. The table reports metrics averaged across the natural T1x pools; lower is better for all columns. Among the methods compared here, model-dependent kernels are the strongest acquisition signals: NTK-PV gives the best force AUC and final force error, while Activation-PV gives the best energy AUC.

tations induce a useful similarity space. We now probe the kernel itself: how distinctively does it separate structurally different configurations, and how much fine-grained resolution does it preserve?

4.2. Kernel Geometry Diagnostics

The acquisition results suggest that the pretrained model induces a useful similarity space for reactive structures. Figure 2 visualises this geometry on a five-reaction T1x subset. Structures are ordered first by reaction family and then by frame index along the reaction coordinate. A useful kernel should capture both levels of structure: it should separate different reaction families while also resolving meaningful variation within a single reaction path.

Among the kernels visualised in Figure 2, the pretrained NTK most clearly combines these two properties. It preserves reaction-family block structure, but does not collapse each family into a nearly homogeneous cluster; instead, it retains variation along the reaction coordinate. The randomly initialised NTK shows some related architectural signal, but is substantially more homogeneous, indicating that pretraining sharpens the representation geometry. SOAP recovers coarse reaction-family structure, but is comparatively saturated within individual reaction paths. Activation kernels show similar behaviour to the NTK. Full global and within-reaction kernel diagnostics for NTK, activations, SOAP, and Tanimoto are provided in Appendix C.

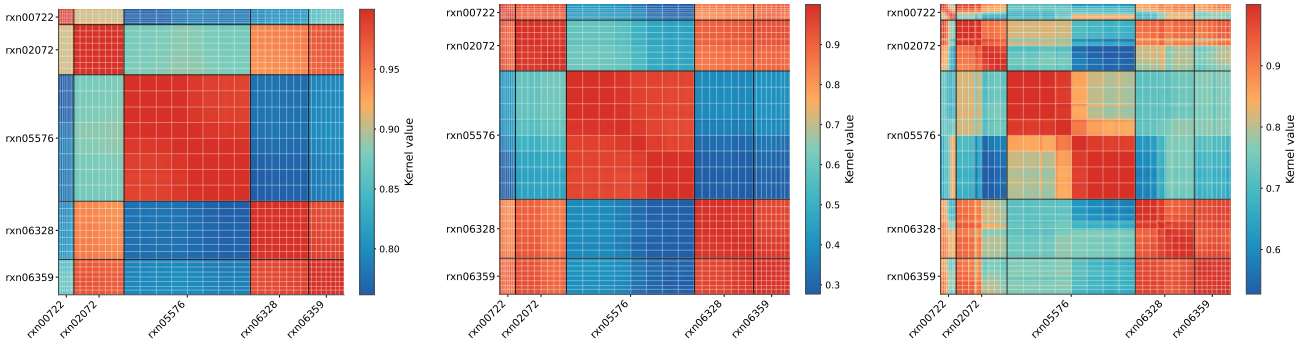
4.3. Residual-GP Uncertainty Calibration

As another diagnostic, we test whether the kernels used for acquisition also provide useful residual uncertainty estimates. Well-calibrated uncertainty is important for kernel-based acquisition strategies: it ensures that predictive variances correspond to true error frequencies, which, in turn, enables principled decision-making.

For each kernel, we hold the pretrained MACE prediction $b(\mathbf{x})$ fixed and fit a Gaussian process (GP) to the residual target $r(\mathbf{x}) = y - b(\mathbf{x})$. The GP therefore uses the kernel only to model the residual correction on top of the pretrained predictor. Because the kernel determines the geometry, but not the overall scale of the residual signal, we estimate a variance scale from the training residuals and evaluate Gaussian negative log-likelihood (NLL), predictive intervals, and calibration metrics using the resulting predictive variance. Details can be found in Appendix C.4.

We compare six residual kernels: pretrained activation features, randomly initialised activation features, pretrained NTK, randomly initialised NTK, SOAP, and Tanimoto. For each kernel, we use the Set 0 dataset from the T1x case study in Section 4.1 and replay posterior-variance acquisition from the same initial seed to obtain a nested sequence of training prefixes. The model-based kernels are computed once and held fixed; we do not fine-tune the model at each step. We do this to isolate the effects of pretraining on the model inductive bias, but it is also computationally cheaper since no training is required.

Figure 3 and Table 2 show that the pretrained model-based



(a) SOAP *(b)* NTK, random init *(c)* NTK, pretrained

Figure 2. Global kernel matrices on a five-reaction T1x subset. Structures are sorted by reaction family and frame index. The pretrained NTK preserves coarse reaction-family structure while retaining finer variation along reaction paths. Appendix C gives the full global and within-reaction kernel diagnostics for NTK, activations, SOAP, and Tanimoto kernels.

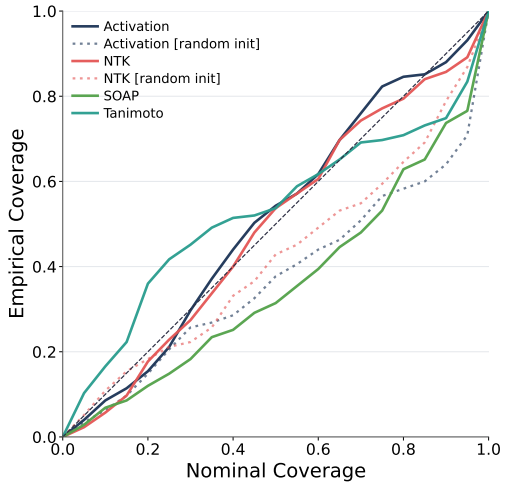


Figure 3. Nominal versus empirical coverage, for the selected kernels. Pretrained activation and pretrained NTK show the best calibration among the compared kernels, while random neural kernels and descriptor kernels are less accurate and less well calibrated.

Kernel	G. NLL	ECE	ENCE	90% Cov.	RMSE (meV)
Activation	**-1.679**	0.028	**0.237**	**0.880**	**45.44**
Activation (random)	0.137	0.106	0.796	0.640	155.83
NTK	-1.439	**0.020**	0.337	0.857	54.54
NTK (random init)	-0.917	0.071	0.294	0.789	88.95
SOAP	0.837	0.113	0.741	0.737	351.90
Tanimoto	2.741	0.069	1.573	0.749	253.86

Table 2. Residual-GP test metrics at the validation-selected stopping point. Lower is better for Gaussian NLL, RMSE, ECE, and ENCE, with ECE and ENCE equal to zero indicating perfect calibration. Empirical 90% coverage is best when closest to the nominal 90% level.

kernels provide the strongest residual uncertainty estimates. Pretrained activation features give the best Gaussian NLL, expected normalised calibration error (ENCE), 90% coverage, and root mean squared error (RMSE), while the pretrained NTK gives the lowest expected calibration error (ECE). Randomly initialised neural kernels are weaker than their pretrained counterparts, and descriptor kernels are weaker overall. These results are consistent with the mechanism suggested by the acquisition experiments: pretraining induces representation geometries that are better aligned with the residual errors left by the MACE bias model.

Note that because the GP is fit with a global structure kernel, its residual predictions should not be interpreted as a fully accurate energy model: the true energy depends on local atomic environments. We expect this limitation to become more pronounced for more difficult fine-tuning tasks beyond T1x. In such settings, the residual function may change substantially during fine-tuning. Thus, the residual GP should be interpreted as a diagnostic of how well the pretrained representation organises residual error, rather than as a fully adaptive uncertainty model for the fine-tuned MLIP.

4.4. Additional Reactivity Datasets

In this section we test whether model-based acquisition strategies outperform random acquisition across other reactivity benchmarks. These transfer experiments keep the same pretrained model family and therefore test dataset transfer rather than architecture transfer.

We evaluate the same pretrained model on PMechDB (Tavakoli et al., 2024), RGD (Zhao et al., 2023), and a larger T1x subset. RGD and PMechDB are random 5k and 10k splits of the corresponding OMol subsets (Levine et al., 2026), with PMechDB filtered to H, C, N, and O chemistry so we can use the pretrained SPICE 2 model. The larger T1x subset, which we call *T1x Mixed*, contains a random selection of 100 T1x reaction pathways.

For each dataset, we first reserve disjoint test and validation sets comprising 20% and 10% of the data, respectively. The

Method	Metric	PMechDB	RGD	T1X Mixed	Average
Activation LCMD	E RMSE	+15.0%	+22.2%	+77.8%	**+38.3%**
	F RMSE	+20.0%	+10.0%	+44.4%	**+24.8%**
	E MAE	+5.0%	+10.0%	+61.5%	**+25.5%**
	F MAE	+0.0%	+0.0%	+15.0%	**+5.0%**
NTK LCMD	E RMSE	+20.0%	+11.1%	+83.3%	**+38.1%**
	F RMSE	+25.0%	+10.0%	+50.0%	**+28.3%**
	E MAE	+5.0%	+15.0%	+61.5%	**+27.2%**
	F MAE	+0.0%	+0.0%	+25.0%	**+8.3%**
Committee Energy	E RMSE	-23.5%	+11.1%	+27.8%	**+5.1%**
	F RMSE	+25.0%	+15.0%	+0.0%	**+13.3%**
	E MAE	-31.6%	+5.0%	-13.3%	**-13.3%**
	F MAE	-5.0%	+0.0%	+0.0%	**-1.7%**

Table 3. Round gain relative to random acquisition across datasets. Positive values indicate earlier acquisition than random, and negative values indicate later acquisition.

remaining structures define the acquisition pool. From this pool, we initialise active learning with 50 labelled structures and then run 20 acquisition rounds, adding 150 structures per round. Each experiment is repeated with two random seeds and metrics are averaged over both seeds. We compare model-based acquisition families to committee and random-based acquisition. For kernel-based methods, we use LCMD because posterior-variance selection can over-prioritise isolated high-variance outliers, whereas LCMD balances distance from the labelled set with cluster mass, and is more stable on heterogeneous candidate pools.

Table 3 reports round gains relative to random acquisition. For each dataset, method, and metric, we define the best shared value as the lowest value reached by both the method and random, and we report the percentage gain in the number of rounds taken to reach this target. Positive values indicate that the method reaches the target earlier than random, whereas negative values indicate that it reaches the target later.

The model-based representation methods transfer most consistently, whereas the energy committee shows mixed behaviour across energy and force metrics. Specifically, both Activation-LCMD and NTK-LCMD improve force RMSE on all three datasets, and they also improve energy RMSE and energy MAE across all three benchmarks. They achieve the strongest average gains across metrics overall. The gains are more modest on RGD and PMechDB than on T1x, which we believe reflects the greater difficulty of those settings. The gains are also larger for RMSE than for MAE, consistent with active learning preferentially reducing high-error tail cases. Detailed learning curves are provided in Appendix D.

5. Discussion

Across the reactive-chemistry settings we study, kernels built directly from pretrained or fine-tuned MACE models, that is the energy NTK and hidden activations, give stronger active-learning performance than fixed molecular descriptors or committee disagreement based strategies. Kernel diagnostics support the same interpretation: pretrained neural representations preserve both reaction-family structure and within-reaction variation, the distinctions needed to select informative structures along reaction pathways. Concretely, this means that model-derived representations are robust under the kind of generator-induced pool bias that reactive chemistry intrinsically exhibits.

From a practical perspective, activation-based representations are especially attractive as they require only a forward pass, whereas NTK features require an additional backward pass but remain substantially cheaper than committee ensembles, which add M full training runs per round. Both representations are already present inside the model being fine-tuned and require no auxiliary training.

The binding constraint of our implementation is memory. Both PV and LCMD operate in kernel space and form the candidate–candidate Gram matrix $k_{\mathcal{PP}} \in \mathbb{R}^{n_{\mathcal{P}} \times n_{\mathcal{P}}}$, which costs $O(n_{\mathcal{P}}^2)$. At the moderate pool sizes considered here ($n_{\mathcal{P}} \leq 10$k), $k_{\mathcal{PP}}$ comfortably fits on a single GPU; an order of magnitude more candidates and the kernel can no longer be materialised on a single device, making kernel construction the dominant cost. Activation features are cheaper per byte (d=256, ~ 10 MB at $n_{\mathcal{P}}$=10k) but the kernel itself remains $O(n_{\mathcal{P}}^2)$, so they share the same large-pool ceiling.

Future work should aim to extend to larger candidate pools, and test whether these conclusions extend to other pretrained MLIP architectures and datasets, and consider alternative committee constructions in pretrained settings. It would also be valuable to study force-aware model representations, compare against newer uncertainty heads (Ho et al., 2025), Bayesian interatomic potentials (Coscia et al., 2026), and multi-head ensemble methods (Beck et al., 2025). We initiate the exploration of a force-aware NTK in (Varga-Umbrich et al., 2026).

More broadly, the same kernels could be useful beyond active learning, for training-set summarisation, dataset distillation and validation-set construction. Across these settings the underlying question is the same: whether pretrained molecular representations define a similarity space better aligned with downstream error than fixed descriptors alone. Our results suggest this direction is promising.

Acknowledgements

We thank Silvia Acosta Gutiérrez, Massimo Bortone, Heloise Chomet, Valentin Heyraud, Jack Simons, Tamás Lajos Tompa, Lucien Walewski, and Leon Wehrhan for insightful discussions. We also thank Scott Cameron for highlighting the potential usefulness of Neural Tangent Ker-

nels for uncertainty quantification.

References

Achar, S. K., Shukla, P. B., Mhatre, C. V., Bernasconi, L., Vinger, C. Y., and Johnson, J. K. Reactive active learning: An efficient approach for training machine learning interatomic potentials for reacting systems. *Journal of Chemical Theory and Computation*, 21(18):8889–8906, 2025. ISSN 1549-9626. doi: 10.1021/acs.jctc.5c00920. URL http://dx.doi.org/10.1021/acs.jctc.5c00920.

Bartók, A. P., Kondor, R., and Csányi, G. On representing chemical environments. *Physical Review B*, 87(18): 184115, 2013. doi: 10.1103/PhysRevB.87.184115.

Bartók, A. P., De, S., Poelking, C., Bernstein, N., Kermode, J. R., Csányi, G., and Ceriotti, M. Machine learning unifies the modeling of materials and molecules. *Science Advances*, 3(12), December 2017. ISSN 2375-2548. doi: 10.1126/sciadv.1701816. URL http://dx.doi.org/10.1126/sciadv.1701816.

Batatia, I., Benner, P., Chiang, Y., Elena, A. M., Kovács, D. P., Riebesell, J., Advincula, X. R., Asta, M., Avaylon, M., et al. A foundation model for atomistic materials chemistry, 2023a. URL https://arxiv.org/abs/2401.00096.

Batatia, I., Kovács, D. P., Simm, G. N. C., Ortner, C., and Csányi, G. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, 2023b. URL https://arxiv.org/abs/2206.07697.

Beck, H., Simko, P., Schaaf, L. L., Marsalek, O., and Schran, C. Multi-head committees enable direct uncertainty prediction for atomistic foundation models. *The Journal of Chemical Physics*, 163(23), December 2025. ISSN 1089-7690. doi: 10.1063/5.0302097. URL http://dx.doi.org/10.1063/5.0302097.

Behler, J. and Parrinello, M. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14):146401, 2007. doi: 10.1103/PhysRevLett.98.146401.

Brunken, C., Peltre, O., Chomet, H., Walewski, L., McAuliffe, M., Heyraud, V., Attias, S., Maarand, M., Khanfir, Y., Toledo, E., Falcioni, F., Bluntzer, M., Acosta-Gutiérrez, S., and Tilly, J. Machine learning interatomic potentials: library for efficient training, model development and simulation of molecular systems, 2025. URL https://arxiv.org/abs/2505.22397.

Chen, C. and Ong, S. P. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, November 2022. ISSN 2662-8457. doi: 10.1038/s43588-022-00349-3. URL http://dx.doi.org/10.1038/s43588-022-00349-3.

Coscia, D., de Haan, P., and Welling, M. Blips: Bayesian learned interatomic potentials, 2026. URL https://arxiv.org/abs/2508.14022.

Cui, T., Tang, C., Zhou, D., Li, Y., Gong, X., Ouyang, W., Su, M., and Zhang, S. Online test-time adaptation for better generalization of interatomic potentials to out-of-distribution data. *Nature Communications*, 16(1), February 2025. ISSN 2041-1723. doi: 10.1038/s41467-025-57101-4. URL http://dx.doi.org/10.1038/s41467-025-57101-4.

Deng, B., Zhong, P., Jun, K., Riebesell, J., Han, K., Bartel, C. J., and Ceder, G. CHGNet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5:1031–1041, 2023. doi: 10.1038/s42256-023-00716-3.

Deng, B., Choi, Y., Zhong, P., Riebesell, J., Anand, S., Li, Z., Jun, K., Persson, K. A., and Ceder, G. Systematic softening in universal machine learning interatomic potentials. *npj Computational Materials*, 11:9, 2025. doi: 10.1038/s41524-024-01500-6.

Eastman, P., Behara, P. K., Dotson, D. L., Galvelis, R., Herr, J. E., Horton, J. T., Mao, Y., Chodera, J. D., Pritchard, B. P., Wang, Y., De Fabritiis, G., and Markland, T. E. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10(1), January 2023. ISSN 2052-4463. doi: 10.1038/s41597-022-01882-6. URL http://dx.doi.org/10.1038/s41597-022-01882-6.

Henkelman, G. and Jónsson, H. Improved tangent estimate in the nudged elastic band method for finding minimum energy paths and saddle points. *The Journal of Chemical Physics*, 113(22):9978–9985, December 2000. ISSN 1089-7690. doi: 10.1063/1.1323224. URL http://dx.doi.org/10.1063/1.1323224.

Henkelman, G., Uberuaga, B. P., and Jónsson, H. A climbing image nudged elastic band method for finding saddle points and minimum energy paths. *The Journal of Chemical Physics*, 113(22):9901–9904, December 2000. ISSN 1089-7690. doi: 10.1063/1.1329672. URL http://dx.doi.org/10.1063/1.1329672.

Himanen, L., Jäger, M. O. J., Morooka, E. V., Federici Canova, F., Ranawat, Y. S., Gao, D. Z., Rinke, P., and Foster, A. S. DScribe: Library of descriptors

for machine learning in materials science. *Computer Physics Communications*, 247:106949, 2020. ISSN 0010-4655. doi: 10.1016/j.cpc.2019.106949. URL https://doi.org/10.1016/j.cpc.2019.106949.

Ho, C. H., Ortner, C., and Wang, Y. Flexible uncertainty calibration for machine-learned interatomic potentials, 2025. URL https://arxiv.org/abs/2510.00721.

Holzmüller, D., Zaverkin, V., Kästner, J., and Steinwart, I. A framework and benchmark for deep batch active learning for regression. *Journal of Machine Learning Research*, 24(164):1–81, 2023. URL https://www.jmlr.org/papers/v24/22-0937.html.

Jacobs, R., Morgan, D., Attarian, S., Meng, J., Shen, C., Wu, Z., Xie, C. Y., Yang, J. H., Artrith, N., Blaiszik, B., Ceder, G., Choudhary, K., Csanyi, G., Cubuk, E. D., Deng, B., Drautz, R., Fu, X., Godwin, J., Honavar, V., Isayev, O., Johansson, A., Kozinsky, B., Martiniani, S., Ong, S. P., Poltavsky, I., Schmidt, K., Takamoto, S., Thompson, A. P., Westermayr, J., and Wood, B. M. A practical guide to machine learning interatomic potentials – status and future. *Current Opinion in Solid State and Materials Science*, 35:101214, 2025. ISSN 1359-0286. doi: https://doi.org/10.1016/j.cossms.2025.101214. URL https://www.sciencedirect.com/science/article/pii/S1359028625000014.

Jinnouchi, R., Karsai, F., and Kresse, G. On-the-fly machine learning force field generation: Application to melting points. *Physical Review B*, 100(1):014105, 2019. doi: 10.1103/PhysRevB.100.014105.

Jónsson, H., Mills, G., and Jacobsen, K. W. Nudged elastic band method for finding minimum energy paths of transitions. In *Classical and Quantum Dynamics in Condensed Phase Simulations*, pp. 385–404. World Scientific, June 1998. doi: 10.1142/9789812839664_0016. URL http://dx.doi.org/10.1142/9789812839664_0016.

Kahle, L. and Zipoli, F. Quality of uncertainty estimates from neural network potential ensembles. *Physical Review E*, 105(1), January 2022. ISSN 2470-0053. doi: 10.1103/physreve.105.015311. URL http://dx.doi.org/10.1103/PhysRevE.105.015311.

Khan, M. A., D'Souza, A., and Choyal, V. Active learning strategies for efficient machine-learned interatomic potentials across diverse material systems, 2026. URL https://arxiv.org/abs/2601.06916.

Kim, S. Y., Park, Y. J., and Li, J. Leveraging neural network interatomic potentials for a foundation model of chemistry. *npj Computational Materials*, 2026. ISSN 2057-3960. doi: 10.1038/s41524-026-02167-x. URL http://dx.doi.org/10.1038/s41524-026-02167-x.

Kohn, W., Becke, A. D., and Parr, R. G. Density functional theory of electronic structure. *The Journal of Physical Chemistry*, 100(31):12974–12980, 1996. ISSN 0022-3654. doi: 10.1021/jp960669l. URL https://doi.org/10.1021/jp960669l.

Kulichenko, M., Barros, K., Lubbers, N., Li, Y. W., Messerly, R., Tretiak, S., Smith, J. S., and Nebgen, B. Uncertainty-driven dynamics for active learning of interatomic potentials. *Nature Computational Science*, 3:230–239, 2023. doi: 10.1038/s43588-023-00406-5.

Laio, A. and Parrinello, M. Escaping free-energy minima. *Proceedings of the National Academy of Sciences*, 99(20):12562–12566, September 2002. ISSN 1091-6490. doi: 10.1073/pnas.202427399. URL http://dx.doi.org/10.1073/pnas.202427399.

Landrum, G. et al. RDKit: Open-source cheminformatics Software, 2025. URL https://doi.org/10.5281/zenodo.17232453. Version 2025_09_1.

Levine, D. S., Shuaibi, M., Spotte-Smith, E. W. C., Taylor, M. G., Hasyim, M. R., Michel, K., Batatia, I., Csányi, G., Dzamba, M., Eastman, P., Frey, N. C., Fu, X., Gharakhanyan, V., Krishnapriyan, A. S., Rackers, J. A., Raja, S., Rizvi, A., Rosen, A. S., Ulissi, Z., Vargas, S., Zitnick, C. L., Blau, S. M., and Wood, B. M. The open molecules 2025 (omol25) dataset, evaluations, and models, 2026. URL https://arxiv.org/abs/2505.08762.

Li, Y., Zhang, X., Liu, M., and Shen, L. A critical review of machine learning interatomic potentials and hamiltonian. *Journal of Materials Informatics*, 5(4), 2025. ISSN 2770-372X. doi: 10.20517/jmi.2025.17. URL https://www.oaepublish.com/articles/jmi.2025.17.

Morgan, H. L. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, May 1965. ISSN 1541-5732. doi: 10.1021/c160017a018. URL http://dx.doi.org/10.1021/c160017a018.

Ouyang, X., Wang, Z., Jie, X., Zhang, F., Zhang, Y., Liu, L., and Wang, D. Latent space active learning with message passing neural network: The case of hfo_2. *Physical Review Materials*, 8(10):103804, October 2024. ISSN 2475-9953. doi: 10.1103/PhysRevMaterials.8.103804. URL https://doi.org/10.1103/PhysRevMaterials.8.103804.

Peterson, A. A., Christensen, R., and Khorshidi, A. Addressing uncertainty in atomistic machine learning. *Physical Chemistry Chemical Physics*, 19:10978–10985, 2017. doi: 10.1039/C7CP00375G.

Podryabinkin, E. V. and Shapeev, A. V. Active learning of linearly parametrized interatomic potentials. *Computational Materials Science*, 140:171–180, 2017. doi: 10.1016/j.commatsci.2017.08.031.

Ralaivola, L., Swamidass, S. J., Saigo, H., and Baldi, P. Graph kernels for chemical informatics. *Neural Networks*, 18(8):1093–1110, October 2005. ISSN 0893-6080. doi: 10.1016/j.neunet.2005.07.009. URL http://dx.doi.org/10.1016/j.neunet.2005.07.009.

Rasmussen, C. E. and Williams, C. K. I. *Gaussian Processes for Machine Learning*. The MIT Press, November 2005. ISBN 9780262256834. doi: 10.7551/mitpress/3206.001.0001. URL http://dx.doi.org/10.7551/mitpress/3206.001.0001.

Rhodes, B., Vandenhaute, S., Šimkus, V., Gin, J., Godwin, J., Duignan, T., and Neumann, M. Orb-v3: atomistic simulation at scale, 2025. URL https://arxiv.org/abs/2504.06231.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Schran, C., Brezina, K., and Marsalek, O. Committee neural network potentials control generalization errors and enable active learning. *The Journal of Chemical Physics*, 153(10):104105, 2020. doi: 10.1063/5.0016004.

Schreiner, M., Bhowmik, A., Vegge, T., Busk, J., and Winther, O. Transition1x - a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9(1), December 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01870-w. URL http://dx.doi.org/10.1038/s41597-022-01870-w.

Settles, B. *Active Learning*. Springer International Publishing, 2012. ISBN 9783031015601. doi: 10.1007/978-3-031-01560-1. URL http://dx.doi.org/10.1007/978-3-031-01560-1.

Shiota, T., Ishihara, K., and Mizukami, W. Universal neural network potentials as descriptors: towards scalable chemical property prediction using quantum and classical computers. *Digital Discovery*, 3(9):1714–1728, 2024. ISSN 2635-098X. doi: 10.1039/d4dd00098f. URL http://dx.doi.org/10.1039/d4dd00098f.

Smith, J. S., Nebgen, B., Lubbers, N., Isayev, O., and Roitberg, A. E. Less is more: Sampling chemical space with active learning. *The Journal of Chemical Physics*, 148 (24):241733, 2018. doi: 10.1063/1.5023802.

Tan, A. R., Urata, S., Goldman, S., Dietschreit, J. C. B., and Gómez-Bombarelli, R. Single-model uncertainty quantification in neural network potentials does not consistently outperform model ensembles. *npj Computational Materials*, 9(1), December 2023. ISSN 2057-3960. doi: 10.1038/s41524-023-01180-8. URL http://dx.doi.org/10.1038/s41524-023-01180-8.

Tavakoli, M., Miller, R. J., Angel, M. C., Pfeiffer, M. A., Gutman, E. S., Mood, A. D., Van Vranken, D., and Baldi, P. Pmechdb: A public database of elementary polar reaction steps. *Journal of Chemical Information and Modeling*, 64(6):1975–1983, March 2024. ISSN 1549-960X. doi: 10.1021/acs.jcim.3c01810. URL http://dx.doi.org/10.1021/acs.jcim.3c01810.

Torrie, G. and Valleau, J. Nonphysical sampling distributions in monte carlo free-energy estimation: Umbrella sampling. *Journal of Computational Physics*, 23 (2):187–199, February 1977. ISSN 0021-9991. doi: 10.1016/0021-9991(77)90121-8. URL http://dx.doi.org/10.1016/0021-9991(77)90121-8.

Vandermause, J., Torrisi, S. B., Batzner, S., Xie, Y., Sun, L., Kolpak, A. M., and Kozinsky, B. On-the-fly active learning of interpretable Bayesian force fields for atomistic rare events. *npj Computational Materials*, 6:20, 2020. doi: 10.1038/s41524-020-0283-z.

Varga-Umbrich, E., Weller-Davies, Z., Duckworth, P., Tilly, J., Peltre, O., and Surana, S. Force-aware neural tangent kernels for scalable and robust active learning of mlips, 2026. URL https://arxiv.org/abs/2605.13788.

Vazirani, V. V. *k-Center*, pp. 47–53. Springer Berlin Heidelberg, 2003. ISBN 9783662045657. doi: 10.1007/978-3-662-04565-7_5. URL http://dx.doi.org/10.1007/978-3-662-04565-7_5.

Wang, G., Wang, C., Zhang, X., Li, Z., Zhou, J., and Sun, Z. Machine learning interatomic potential: Bridge the gap between small-scale models and realistic device-scale simulations. *iScience*, 27(5):109673, 2024. ISSN 2589-0042. doi: https://doi.org/10.1016/j.isci.2024.109673. URL https://www.sciencedirect.com/science/article/pii/S2589004224008952.

Wood, B. M., Dzamba, M., Fu, X., Gao, M., Shuaibi, M., Barroso-Luque, L., Abdelmaqsoud, K., Gharakhanyan, V., Kitchin, J. R., Levine, D. S., Michel, K., Sriram, A., Cohen, T., Das, A., Rizvi, A., Sahoo, S. J., Ulissi, Z. W., and Zitnick, C. L. Uma: A family of universal models for atoms, 2026. URL https://arxiv.org/abs/2506.23971.

Zaverkin, V., Holzmüller, D., Steinwart, I., and Kästner, J. Exploring chemical and conformational spaces by batch mode deep active learning. *Digital Discovery*, 1:605–620, 2022. doi: 10.1039/D2DD00034B.

Zaverkin, V., Holzmüller, D., Christiansen, H., Errica, F., Alesiani, F., Takamoto, M., Niepert, M., and Kästner, J. Uncertainty-biased molecular dynamics for learning uniformly accurate interatomic potentials. *npj Computational Materials*, 10:83, 2024. doi: 10.1038/s41524-024-01254-1.

Zhang, L., Lin, D.-Y., Wang, H., Car, R., and E, W. Active learning of uniformly accurate interatomic potentials for materials simulation. *Physical Review Materials*, 3(2): 023804, 2019. doi: 10.1103/PhysRevMaterials.3.023804.

Zhao, Q., Vaddadi, S. M., Woulfe, M., Ogunfowora, L. A., Garimella, S. S., Isayev, O., and Savoie, B. M. Comprehensive exploration of graphically defined reaction spaces. *Scientific Data*, 10(1), March 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02043-z. URL http://dx.doi.org/10.1038/s41597-023-02043-z.

Zhu, A., Batzner, S., Musaelian, A., and Kozinsky, B. Fast uncertainty estimates in deep learning interatomic potentials. *The Journal of Chemical Physics*, 158(16):164111, 2023. doi: 10.1063/5.0136574.

Zou, J. and Marzouk, Y. Data curation for machine learning interatomic potentials by determinantal point processes, 2026. URL https://arxiv.org/abs/2603.22160.

A. MACE Architecture and Representation Extraction

We write a structure as $\mathbf{x} = (\mathbf{z}, \mathbf{r})$, with atomic numbers $\mathbf{z} = (z_i)_{i=1}^{N(\mathbf{x})}$ and Cartesian coordinates $\mathbf{r} = (\mathbf{r}_i)_{i=1}^{N(\mathbf{x})}$. MACE (Batatia et al., 2023b) constructs a radius-cutoff graph $G(\mathbf{x})$: nodes are atoms, and directed edges connect neighbours within cutoff $r_{\max}$. Node attributes are one-hot species vectors $\mathbf{a}_i$. For an edge $j \to i$, define the relative vector $\mathbf{r}_{ji} = \mathbf{r}_j - \mathbf{r}_i$, distance $d_{ji} = \|\mathbf{r}_{ji}\|$, and direction $\hat{\mathbf{r}}_{ji} = \mathbf{r}_{ji}/d_{ji}$.

The embedding stage has two parts: an atomic (node) embedding block and a radial (edge) embedding block.

Atomic embedding weights. The atomic embedding block maps the one-hot species attribute $\mathbf{a}_i$ of each atom to an initial scalar node feature. Concretely, it is a learnable lookup table

$$W^{\text{emb}} \in \mathbb{R}^{Z_{\max} \times c}, \qquad h_{i,L=0}^{(0)} \;=\; W^{\text{emb}}_{z_i,:}, \tag{12}$$

where $Z_{\max}$ is the number of supported atomic species and c is the number of scalar channels. The initial node feature of atom i is therefore the row of W^{emb} indexed by its atomic number z_i: it depends only on the atomic species, not on the local environment, the geometry, or the rest of the structure. Two atoms of the same species in different configurations enter the network with identical initial features, and differences in their downstream representations arise entirely from the geometric information injected by the interaction blocks. Equivalently, the embedding weights act as a per-species learnable bias that carries the chemical identity of the atom into the rest of the model.

Radial embedding. The radial embedding block maps each distance d_{ji} to edge features through a Bessel basis modulated by a smooth cutoff envelope. In the MACE checkpoints used in this work this basis is parameter-free, so the trainable parameters of the embedding stage are dominated by W^{emb}. Angular information is injected separately, and not at the embedding stage, through spherical harmonics of $\hat{\mathbf{r}}_{ji}$.

Interaction stack. Interaction blocks combine neighbour node features with radial and angular edge features to produce equivariant messages, and product blocks increase the effective body order through symmetric tensor products. After interaction layer ℓ, the hidden state of atom i can be written as

$$h_i^{(\ell)} = \bigoplus_{L=0}^{L_{\max}} h_{i,L}^{(\ell)}, \tag{13}$$

where $L = 0$ channels are invariant scalars and $L > 0$ channels transform as vectors or higher-order tensors under rotations.

MACE predicts energy through scalar readouts applied to the invariant channels. With T interaction layers, the total energy has the form

$$E_\theta(\mathbf{x}) = \sum_{i=1}^{N(\mathbf{x})} \sum_{\ell=0}^{T} E_i^{(\ell)}(\mathbf{x}), \tag{14}$$

and forces are analytic coordinate gradients,

$$\mathbf{F}_i(\mathbf{x}) = -\nabla_{\mathbf{r}_i} E_\theta(\mathbf{x}). \tag{15}$$

The activation representation in the main text pools the scalar $h_{i,L=0}^{(\ell)}$ channels over atoms and concatenates the pooled features from each message-passing layer. In our experiments, the MACE model uses angular features up to $L = 2$ and has two message-passing layers with 128 scalar hidden channels per layer, so this gives a 256-dimensional activation vector. The energy NTK representation differentiates $E_\theta(\mathbf{x})$ with respect to the node embedding parameters, which are the same MACE blocks that encode species identity. Although these parameters are themselves species-indexed and geometry-free, the corresponding gradient is computed by backpropagating the energy through the full interaction stack, so each row of $\nabla_{W^{\text{emb}}} E_\theta(\mathbf{x})$ aggregates per-atom sensitivities that depend on the entire local environment of every atom of that species. The embedding-NTK feature can therefore distinguish structures with identical composition but different geometries, while remaining substantially smaller than the gradient with respect to the full network.

B. Training Details

All active-learning rounds fine-tune the same SPICE 2 pretrained MACE checkpoint that is trained using the `mlip` library (Brunken et al., 2025). Weights are optimised on the current labelled set without freezing any model parameters. The

surrogate predicts total energies and atomic forces, and the training objective is a Huber loss on energy and force errors with equal weights $w_E = w_F = 10.0$.

For the T1x dataset, each round uses a dynamic schedule that keeps the approximate number of gradient updates comparable as the labelled set grows. If $|\mathcal{T}^{(t)}|$ is the current labelled-set size, the batch size B and learning rate η are

$$(B, \eta) = \begin{cases} (1, 10^{-3}), & |\mathcal{T}^{(t)}| \leq 20, \\ (2, 5 \times 10^{-3}), & 20 < |\mathcal{T}^{(t)}| \leq 100, \\ (4, 5 \times 10^{-3}), & |\mathcal{T}^{(t)}| > 100, \end{cases} \tag{16}$$

and the number of epochs is

$$E = \max\left(10, \left\lceil \frac{1000B}{|\mathcal{T}^{(t)}|} \right\rceil\right). \tag{17}$$

For all other datasets, we train for 50 epochs with a batch size of 16 and learning rate of 0.01. This was sufficient to get approximate convergence for all trainings.

The initial training sets, validation sets, test sets, and candidate pool are disjoint. Validation performance is used for model selection at each active-learning round.

C. Full T1x results

Table 4 reports the full T1x experimental results, including posterior variance (PV), largest-cluster maximum-distance (LCMD), and k-center (Kc) (Vazirani, 2003) variants. This is the full table underlying the abbreviated main-text summary in Figure 1.

Table 4. T1x summary metrics grouped by kernel family and method variant. Final columns are final-round RMSE in meV for energy and meV Å^{-1} for forces. AUC is the discrete sum of RMSE over acquisition steps, with units of the corresponding RMSE. Best (lowest) values in each metric are shown in bold. The reported values are averaged over the 3 sets.

Kernel	Method	Energy AUC (↓)	Force AUC (↓)	Final E RMSE (↓)	Final F RMSE (↓)
Random	–	648.33	6490.00	6.33	111.83
Committee	Energy	623.33	6636.67	3.50	82.50
	Force	951.33	5496.00	8.00	67.67
Activation	Kc	259.67	3208.67	3.00	43.83
	LCMD	**252.00**	3470.67	**2.83**	47.33
	PV	252.17	3179.00	3.17	42.67
NTK	Kc	273.50	3341.17	**2.83**	45.33
	LCMD	259.17	3452.00	**2.83**	45.67
	PV	252.83	**3078.00**	**2.83**	**40.00**
SOAP	Kc	273.17	4080.83	3.50	54.00
	LCMD	278.83	4270.67	3.00	57.67
	PV	331.83	6926.17	4.33	126.50
Tanimoto	Kc	311.17	4580.00	4.00	84.17
	LCMD	707.67	5209.50	5.67	78.17
	PV	275.33	3791.83	3.17	61.17

Table 5 provides a practical cost comparison of the acquisition methods, including the representation dimension, the number of forward and backward passes, the number of required models, and the overall evaluation cost in the pretrained setting.

C.1. Committee benchmarks

We benchmark committee acquisition under the T1x setting using the same active-learning splits and training schedule as the other methods. Each committee contains three MACE models initialised from the same pretrained checkpoint and trained on the current labelled set at each acquisition round.

Table 5. Operational cost comparison of the acquisition families used in this work. For each method we report the number of forward and backward passes per candidate, the number of models maintained, the representation dimension d, the binding peak-memory term in our kernel-space implementation, and an overall practical cost category. Kernel methods (PV, LCMD) form an $n_{\mathcal{P}} \times n_{\mathcal{P}}$ Gram matrix; this is the dominant memory cost and scales as $O(n_{\mathcal{P}}^2)$ regardless of d. Committees instead pay M full training runs per round.

Method	Fwd.	Bwd.	#Models	Dim. d	Peak memory	Practical cost
Activation	1	0	1	256	$O(n_{\mathcal{P}}^2)$ kernel	Low–Moderate
NTK	1	1	1	1,920	$O(n_{\mathcal{P}}^2)$ kernel + $O(n_{\mathcal{P}} d)$ feats	Moderate
SOAP	0	0	1	3,696	$O(n_{\mathcal{P}}^2)$ kernel	Low
Tanimoto	0	0	1	2,048 (binary)	$O(n_{\mathcal{P}}^2)$ kernel	Low
Committee-E	M	0	M	–	$M\times$ model state	High (M trainings/round)
Committee-F	M	0	M	–	$M\times$ model state	High (M trainings/round)
Random	0	0	1	–	–	Minimal

We compare two ways of inducing diversity across committee members. In the *shuffle* variant, all committee members see the same labelled set but use independent data-order seeds during training. In the *bootstrap* variant, each member is trained on a sample drawn with replacement from the current labelled set, so individual ensemble members see different empirical training distributions. Both variants are evaluated with energy and force disagreement scores.

Figure 4 shows the energy and force learning curves for these committee variants on the three T1x candidate pools. The main failure mode is an unstable energy–force trade-off. Energy committees can improve final energy error, but their acquisition scores are not well aligned with force improvement and they give weaker force learning curves than random acquisition. Force committees select structures that are more useful for reducing force error, but this comes at a large cost to energy accuracy: in Table 4, Committee-F has better final force error than Committee-E, but has the worst energy AUC and final energy error among the reported natural-bias methods.

The rank-correlation diagnostic in Figure 5 helps explain this behaviour. At each round, we compute the Spearman correlation between the committee standard deviation on pool candidates and the corresponding absolute prediction error. A useful uncertainty score should have a consistently positive correlation with held-out error. Instead, the committee correlations are weak across candidate pools, rounds, and randomization schemes. This indicates that the committee score is often ranking candidates by ensemble variability that does not correspond to the downstream error being optimized.

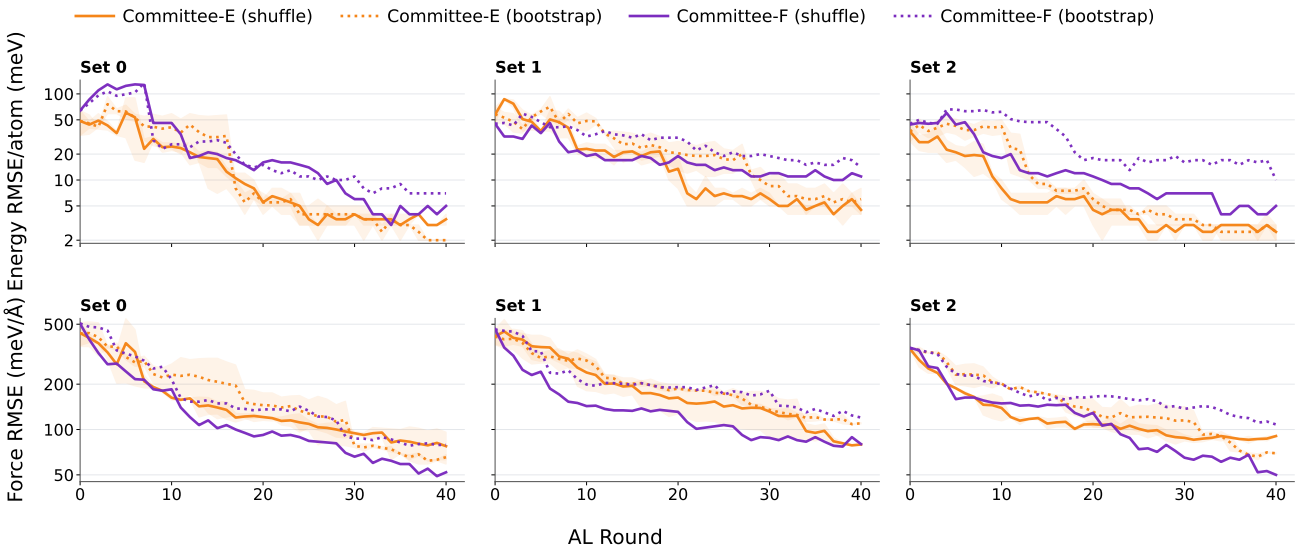


Figure 4. Energy and force RMSE across active-learning rounds for committee variants on the three T1x sets. We compare energy and force disagreement scores, each with shuffle- and bootstrap-based ensemble diversity, across the three natural-bias candidate pools.

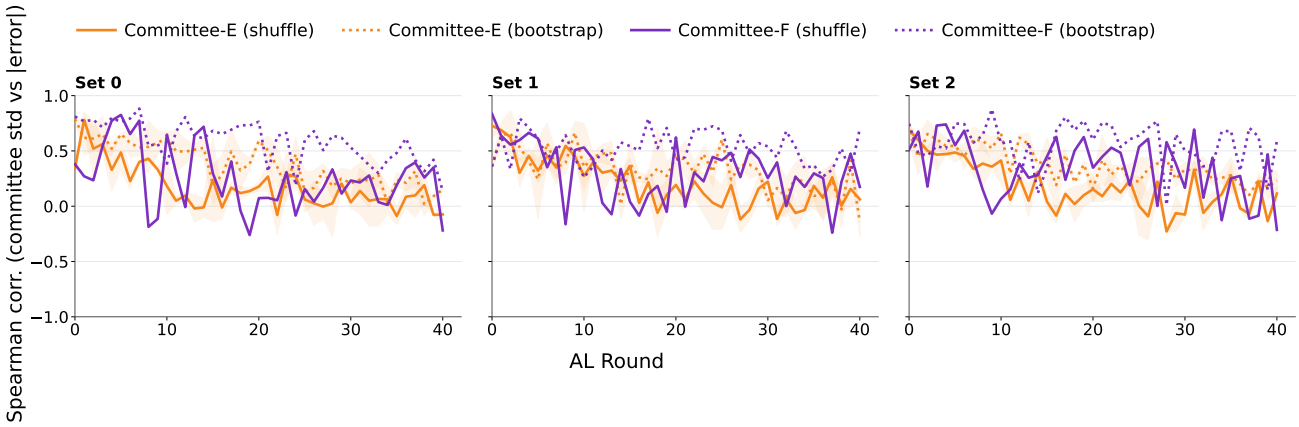


Figure 5. Spearman correlation between committee standard deviation and absolute prediction error across active-learning rounds. The weak correlations show that committee disagreement is not a consistently calibrated ranking signal.

C.2. Scratch training

In this section, we train models from scratch using the same MACE architecture with random initialization. For committee, this now involves using canonical MACE ensembles with different intialization seeds. Figure 6 shows the resulting learning curves. We observe qualitatively similar behavior to the pretrained setting: model-based acquisition strategies such as NTK and activations significantly outperform the alternatives. The full averaged results can be found in Table 6, which also indicates consistent improvements when using pretrained models compared to training from scratch.

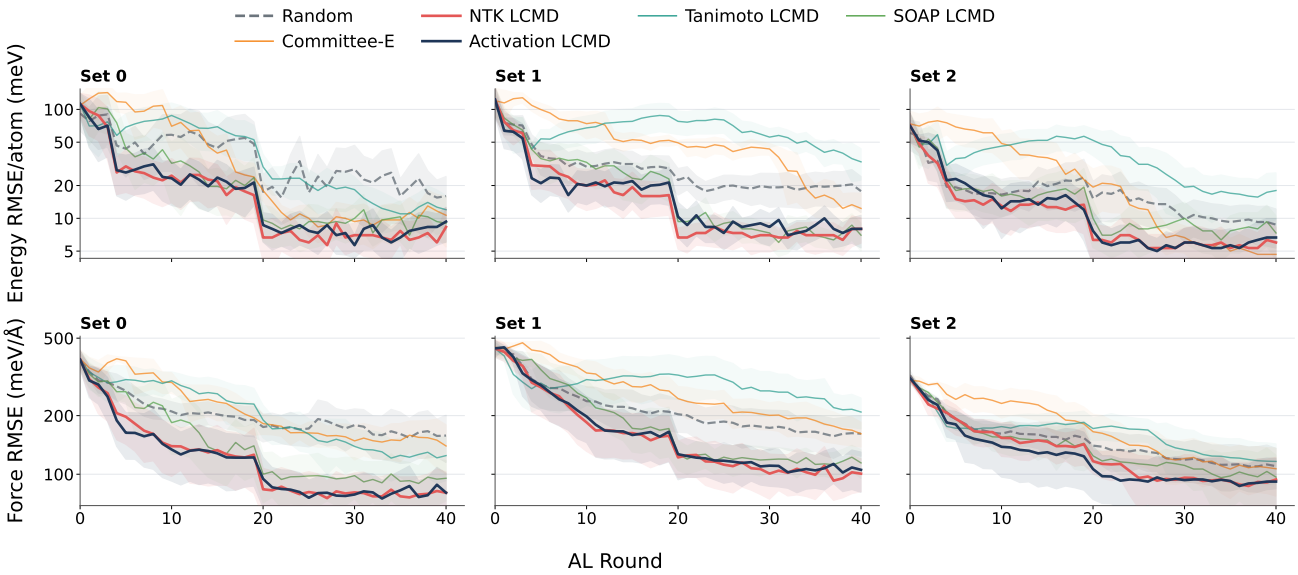


Figure 6. Force and energy learning curves across the T1x sets when training from scratch showing similar results to the pretrained case.

Table 6. Comparison of scratch and pretrained performance grouped by kernel family. All kernel methods are run with LCMD acquisition. Final columns are RMSE in meV for energy and meV Å^{-1} for forces. Best (lowest) values in each metric are shown in bold. The reported values are averaged over the 3 sets.

Kernel	Method	Energy AUC (↓)	Force AUC (↓)	Final E RMSE (↓)	Final F RMSE (↓)
Random	Scratch	1211.58	7953.25	14.17	143.17
	Pretrained	648.33	6490.00	6.33	111.83
Committee	Scratch	1803.44	9352.44	9.22	135.89
	Pretrained	623.33	6636.67	3.50	82.50
NTK	Scratch	743.22	5972.56	7.44	91.67
	Pretrained	259.17	**3452.00**	**2.83**	**45.67**
Activation	Scratch	764.22	5896.11	8.00	92.44
	Pretrained	**252.00**	3470.67	**2.83**	47.33
SOAP	Scratch	927.56	6724.44	7.78	102.11
	Pretrained	278.83	4270.67	3.00	57.67
Tanimoto	Scratch	2004.11	9201.11	21.00	149.89
	Pretrained	707.67	5209.50	5.67	78.17

C.3. Kernel visualisations

We now present the detailed kernel visualisations underlying Figure 2. The global plots order candidate structures by reaction family, making it possible to assess whether a representation separates different reaction classes. The frame-block plots isolate within-reaction structure by ordering frames along each reaction coordinate. Together, these views distinguish coarse reaction-family separation from sensitivity to geometric changes along a reaction path.

The global kernel matrices are computed on the five-reaction T1x subset set 0. For both activations and NTK kernels, we compare trained models to their randomly initialised counterparts, and also track their evolution over selected fine-tuning iterations (1, 20, and 40).

The within-reaction frame-block plots isolate structure by displaying kernels restricted to individual reactions. These typically evolve during training. For example, for the NTK kernels shown in Figure 8, iteration 1 shows some within-reaction variation across most reactions. By iterations 20 and 40, several reactions, most notably 00722, 05576, and 06328, develop clearer block patterns, where early frames become measurably less similar to later frames. In contrast, reactions 02072 and 06359 remain comparatively flat throughout, indicating weaker sensitivity to progression along the reaction coordinate.

Pretrained NTK Kernels

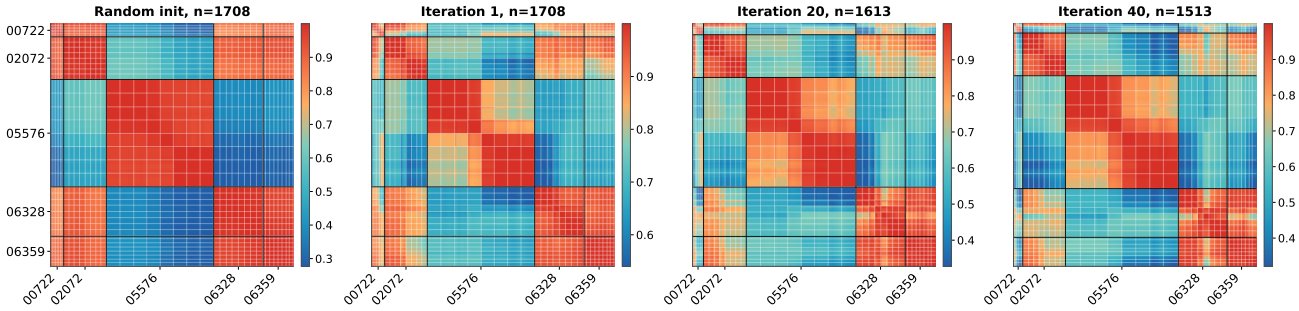


Figure 7. Global NTK kernel matrices on the T1x subset, with structures ordered by reaction family. The panels compare the random-initialised MACE NTK with NTK kernels computed after selected active-learning iterations.

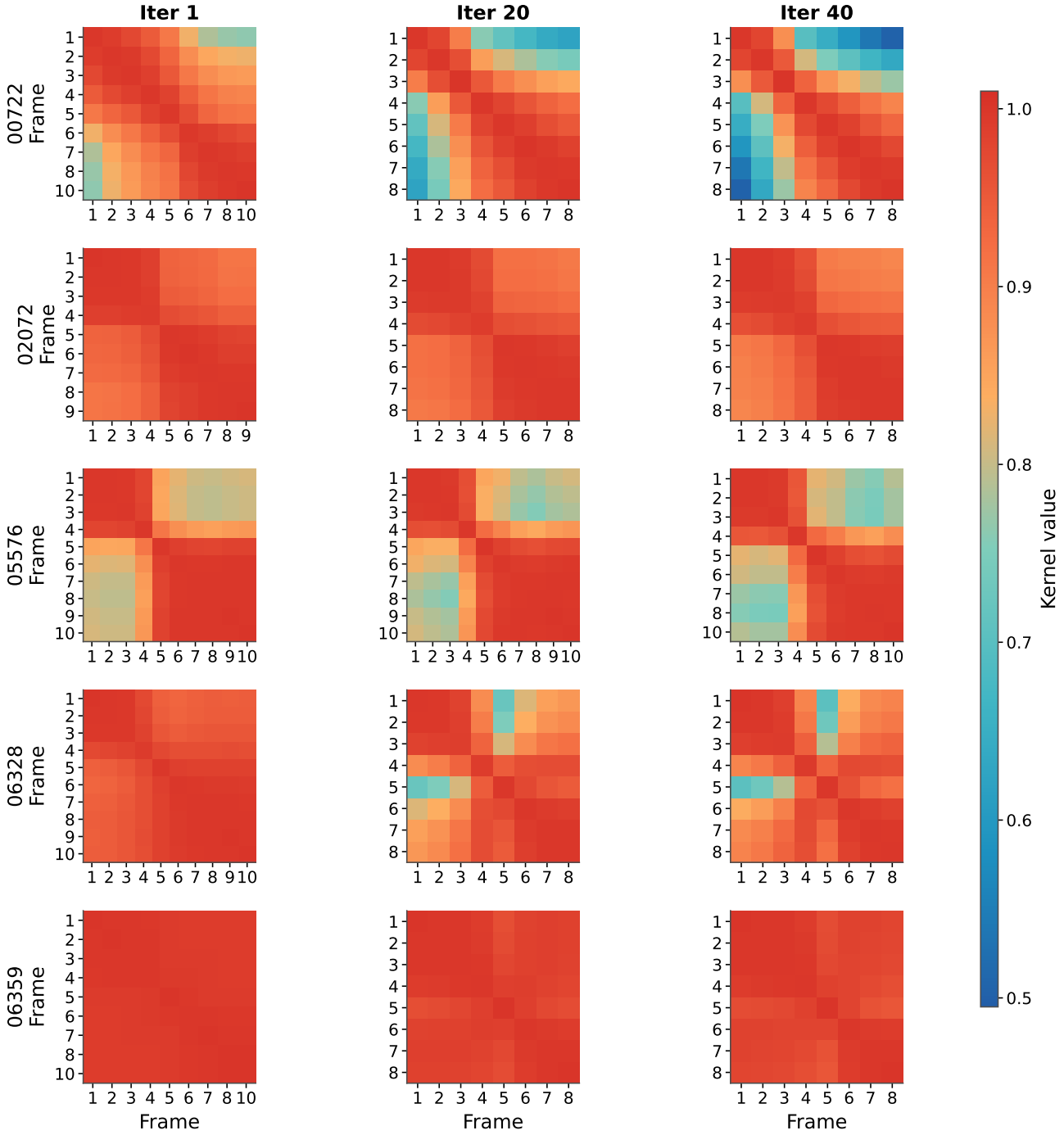


Figure 8. Within-reaction NTK frame-block kernels for the T1x subset after selected active-learning iterations. Each block orders structures by frame index along a reaction pathway.

Scratch NTK kernels

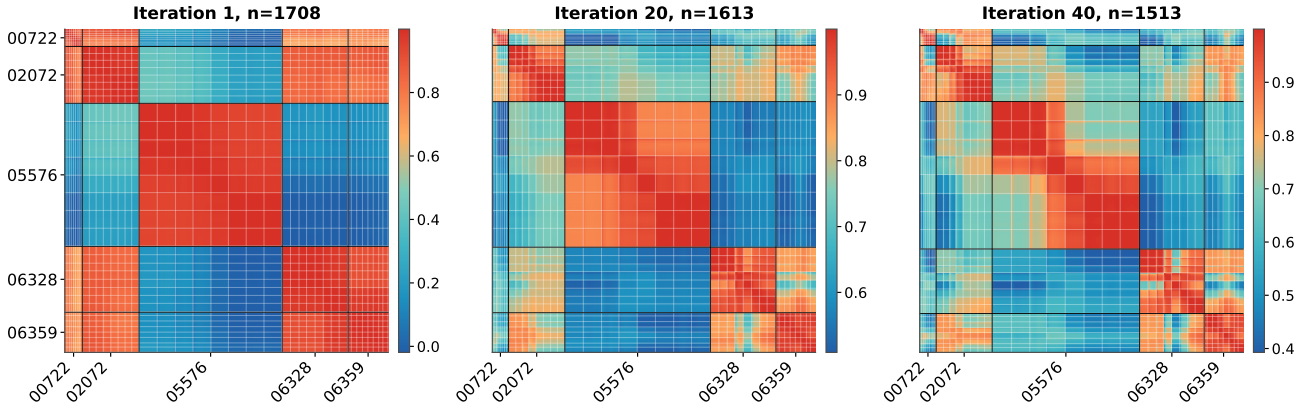


Figure 9. Global NTK kernel matrices on the T1x subset, with structures ordered by reaction family. The panels show how an untrained NTK evolves with AL iteration.

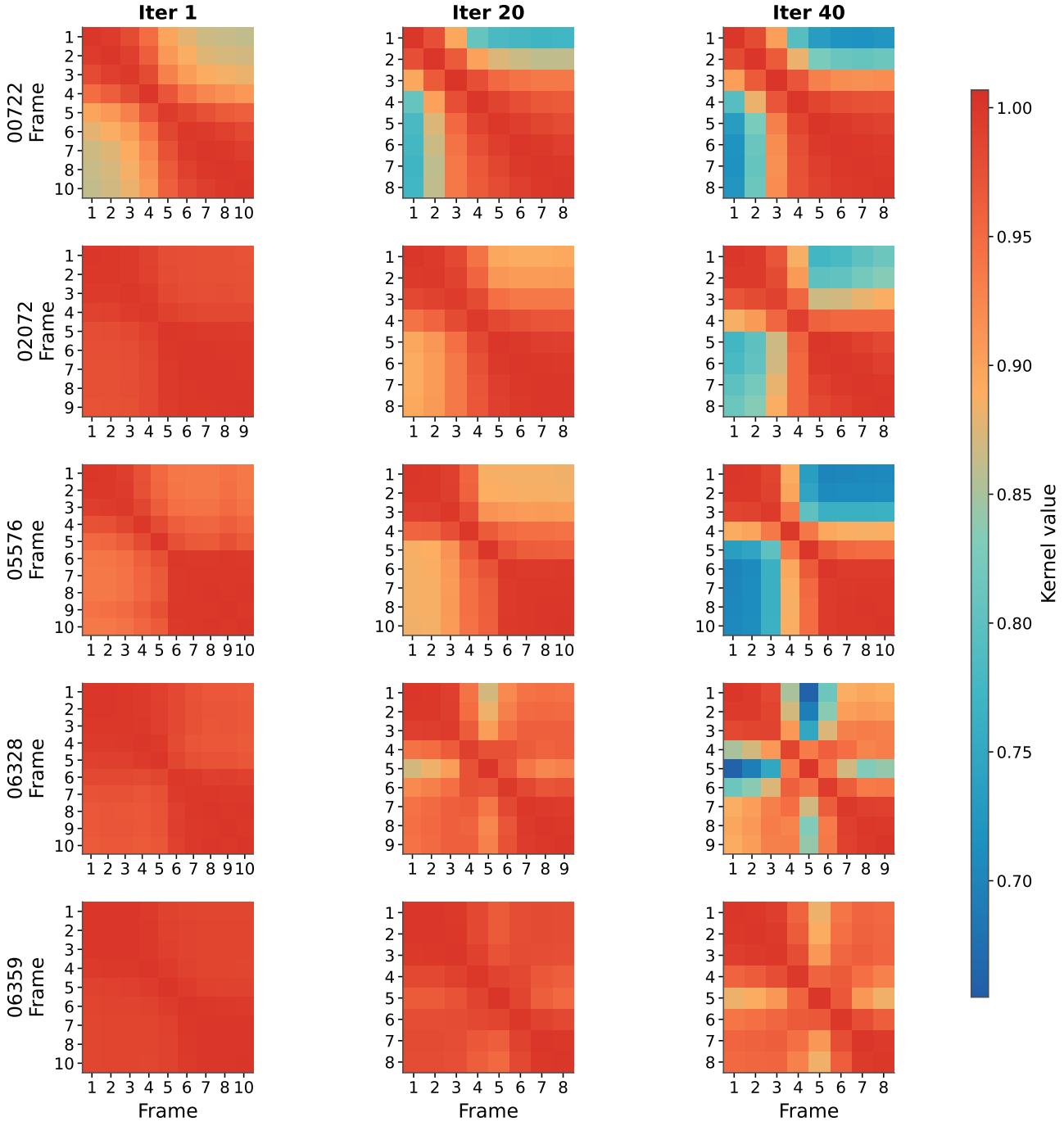


Figure 10. Within-reaction scratch NTK frame-block kernels for the T1x subset after selected active-learning iterations. Each block orders structures by frame index along a reaction pathway.

Pretrained Activation Kernels

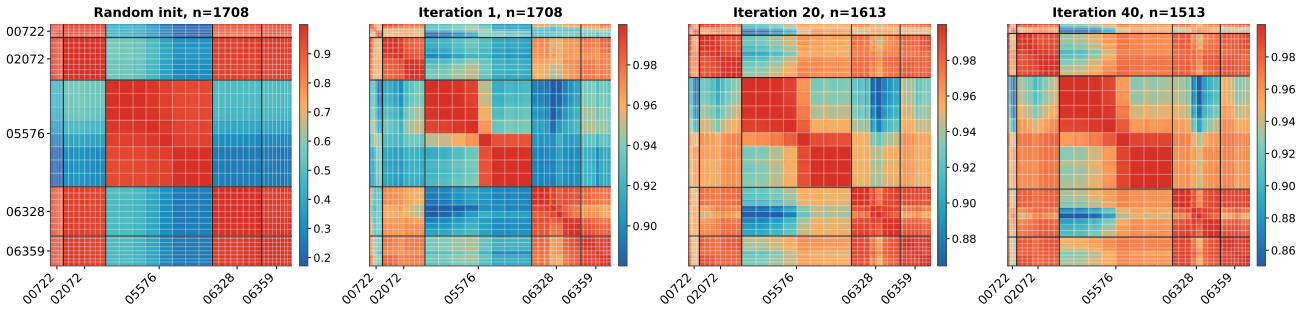


Figure 11. Global activation-kernel matrices on the T1x subset, with structures ordered by reaction family. The panels compare random-initialized and active-learning iteration checkpoints using pooled scalar MACE activations.

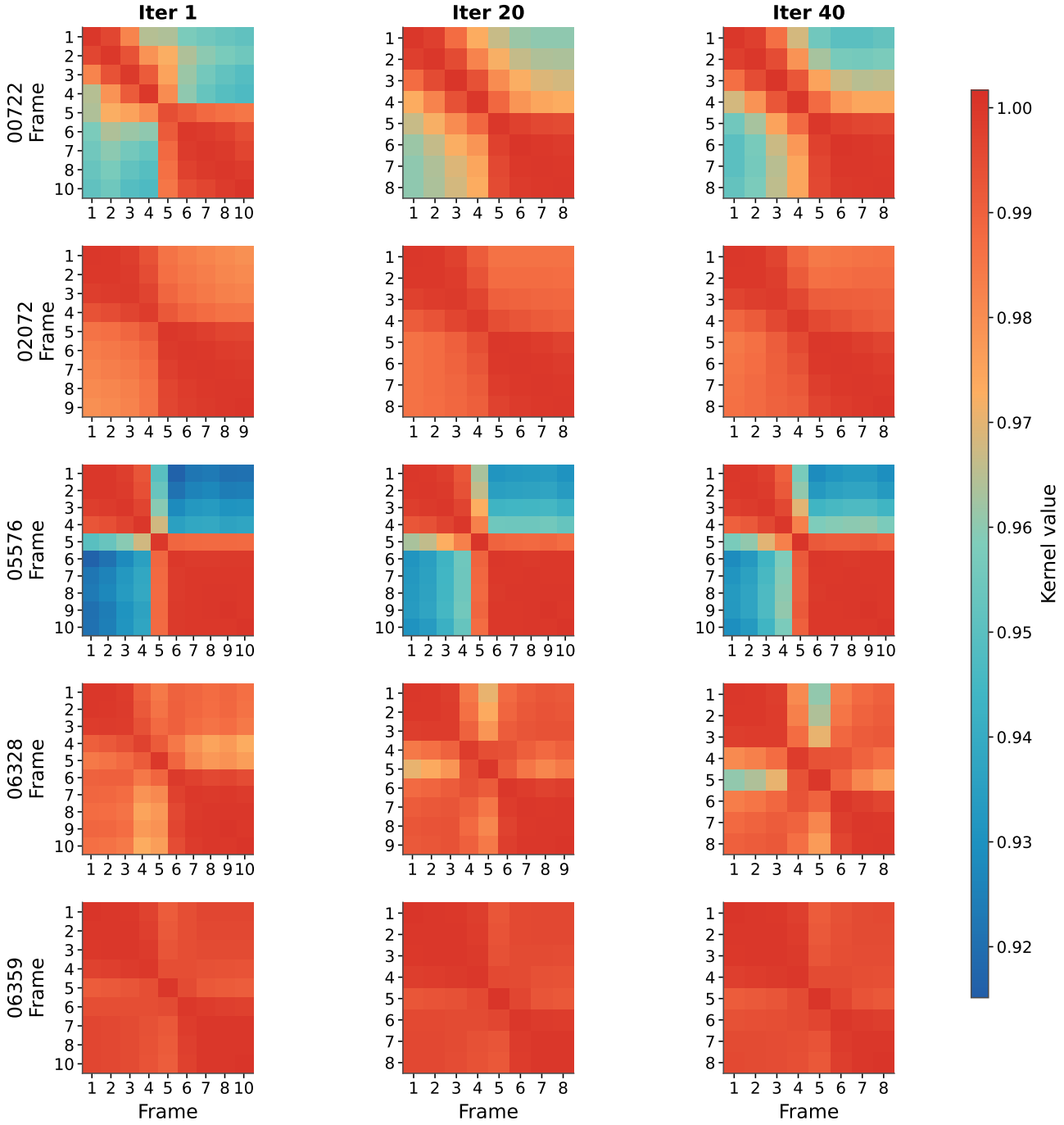


Figure 12. Within-reaction activation frame-block kernels for the T1x subset after selected active-learning iterations.

Scratch Activation Kernels

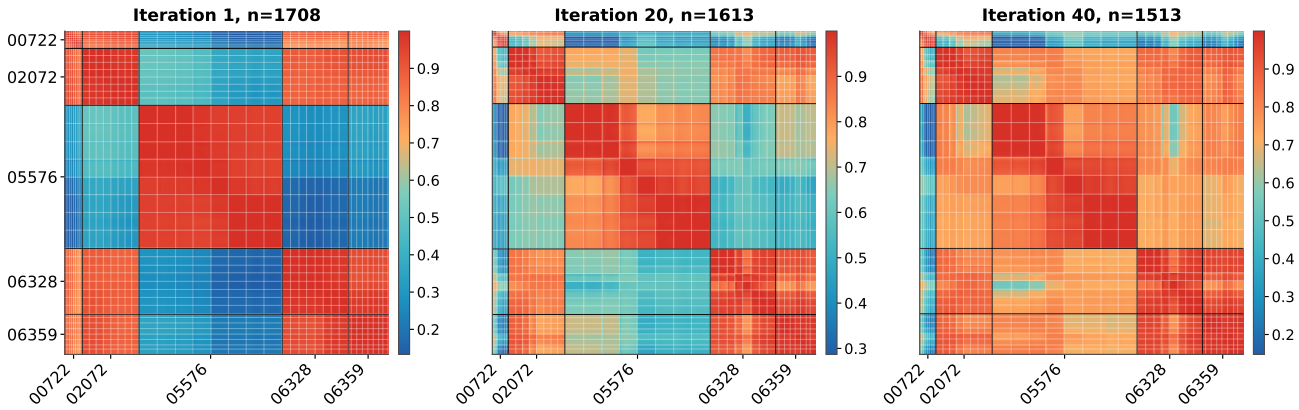


Figure 13. Global srctach activation kernel matrices on the T1x subset, with structures ordered by reaction family. The panels show how an untrained activation kernel evolves with AL iteration.

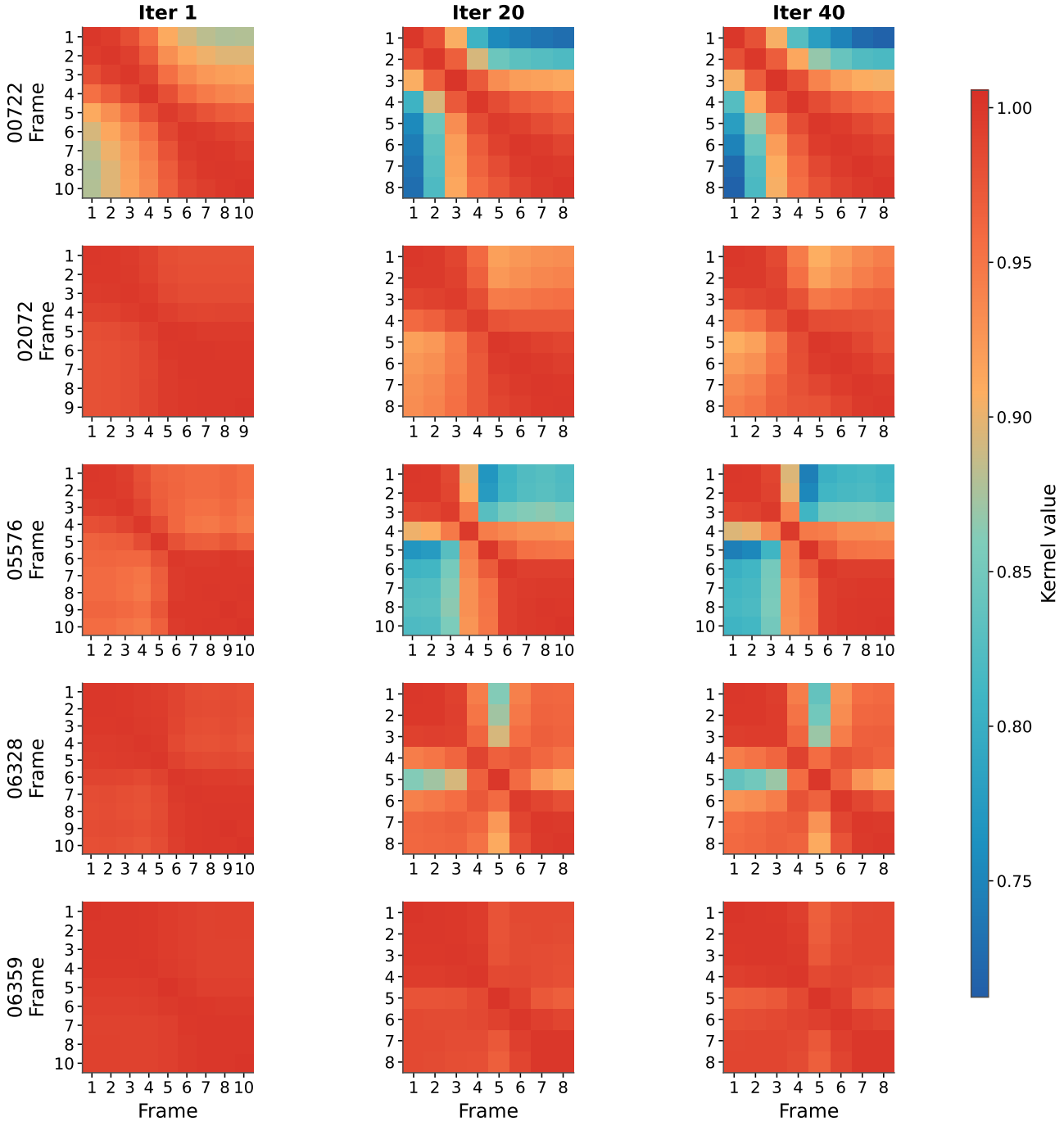
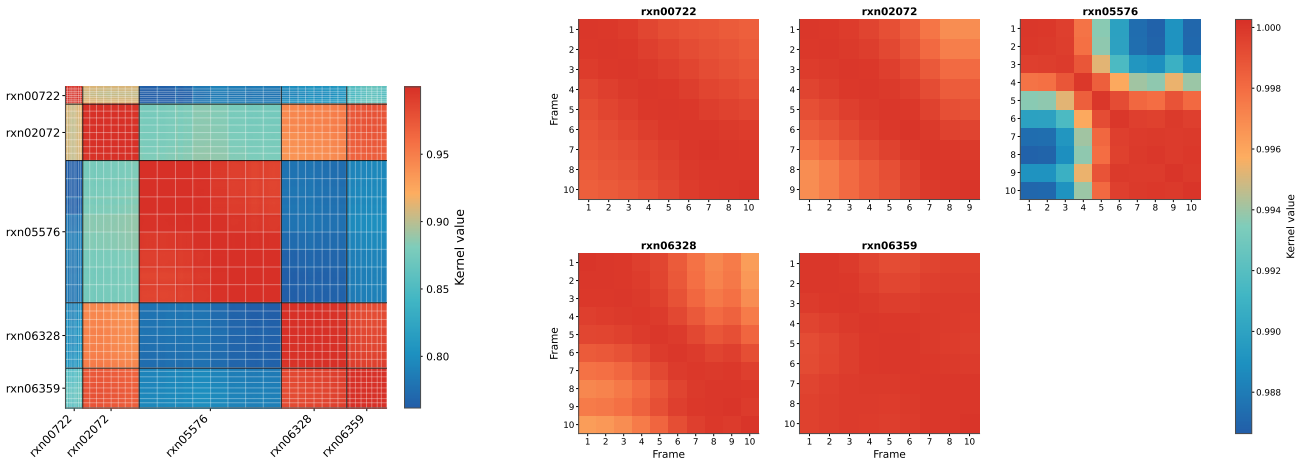


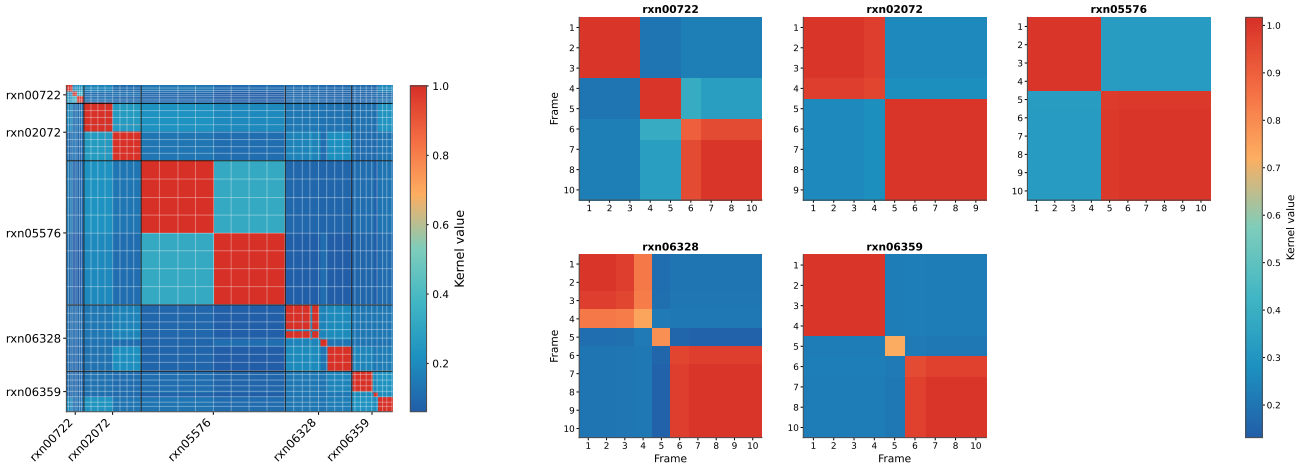
Figure 14. Within-reaction scratch activation frame-block kernels for the T1x subset after selected active-learning iterations.

Descriptor Kernels



(a) SOAP global kernel (b) SOAP within-reaction blocks

Figure 15. SOAP kernel diagnostics on the T1x subset. SOAP captures coarse reaction-family structure using fixed local-geometry descriptors, but many within-reaction similarities remain high.



(a) Tanimoto global kernel (b) Tanimoto within-reaction blocks

Figure 16. Tanimoto kernel diagnostics on the T1x subset. Morgan fingerprints mainly reflect molecular graph identity and are less sensitive to continuous geometry changes along a fixed reaction path and are not able to capture inter reaction similarities.

C.4. Residual-GP Calibration

For the residual GP experiment, for each kernel, we hold the pretrained MACE prediction $b(\mathbf{x})$ fixed and fit a Gaussian process to the residual target

$$r(\mathbf{x}) = y - b(\mathbf{x}).$$

Given a selected training set $\mathcal{T}$ and held-out structure $\mathbf{x}_i$, we center the training residuals by $\bar{r}_{\mathcal{T}} = |\mathcal{T}|^{-1} \sum_{j \in \mathcal{T}} r_j$ and use fixed kernel regularisation $\lambda = 10^{-3}$. The residual-GP predictive correction and raw latent variance are

$$\mu_i = \bar{r}_{\mathcal{T}} + k_{\mathcal{T}}(\mathbf{x}_i)^\top (K_{\mathcal{T}\mathcal{T}} + \lambda I)^{-1} (\mathbf{r}_{\mathcal{T}} - \bar{r}_{\mathcal{T}} \mathbf{1}), \tag{18}$$

$$s_i^2 = k(\mathbf{x}_i, \mathbf{x}_i) - k_{\mathcal{T}}(\mathbf{x}_i)^\top (K_{\mathcal{T}\mathcal{T}} + \lambda I)^{-1} k_{\mathcal{T}}(\mathbf{x}_i). \tag{19}$$

The corrected prediction is

$$\hat{y}_i = b(\mathbf{x}_i) + \mu_i.$$

Since the embedding or similarity definition fixes the kernel geometry but not the residual signal scale, we estimate a per-prefix variance scale from the training residuals,

$$\hat{\gamma} = \frac{(\mathbf{r}_{\mathcal{T}} - \bar{r}_{\mathcal{T}} \mathbf{1})^\top (K_{\mathcal{T}\mathcal{T}} + \lambda I)^{-1} (\mathbf{r}_{\mathcal{T}} - \bar{r}_{\mathcal{T}} \mathbf{1})}{|\mathcal{T}|}. \tag{20}$$

Gaussian negative log likelihood, predictive intervals, and calibration metrics are computed using predictive variance $\hat{\gamma} s_i^2$.

We compare six residual kernels: pretrained activation features, randomly initialised activation features, pretrained NTK, randomly initialized NTK, SOAP, and Tanimoto. The randomly initialized neural kernels use the same MACE architecture before pretraining, separating architectural inductive bias from representation structure learned during pretraining.

For each kernel, we replay posterior-variance acquisition from the same initial seed to produce a nested sequence of training prefixes $\mathcal{T}_1 \subset \mathcal{T}_2 \subset \cdots$. Each prefix is then used to fit a residual Gaussian process (GP). Because calibration can change as additional data is acquired, we select the calibrated GP on the validation split using the Gaussian negative log-likelihood (NLL).

D. Additional Datasets Plots

This appendix provides the detailed learning curves underlying the round-gain summary in Table 3. For each of the three datasets, PMechDB, RGD, and T1x Mixed, we report energy and force errors in both RMSE and MAE form across acquisition rounds, comparing model-based kernels (Activation-LCMD and NTK-LCMD), committee energy disagreement, and random acquisition. Figure 17 shows force RMSE, Figure 18 shows force MAE, Figure 19 shows energy RMSE, and Figure 20 shows energy MAE.

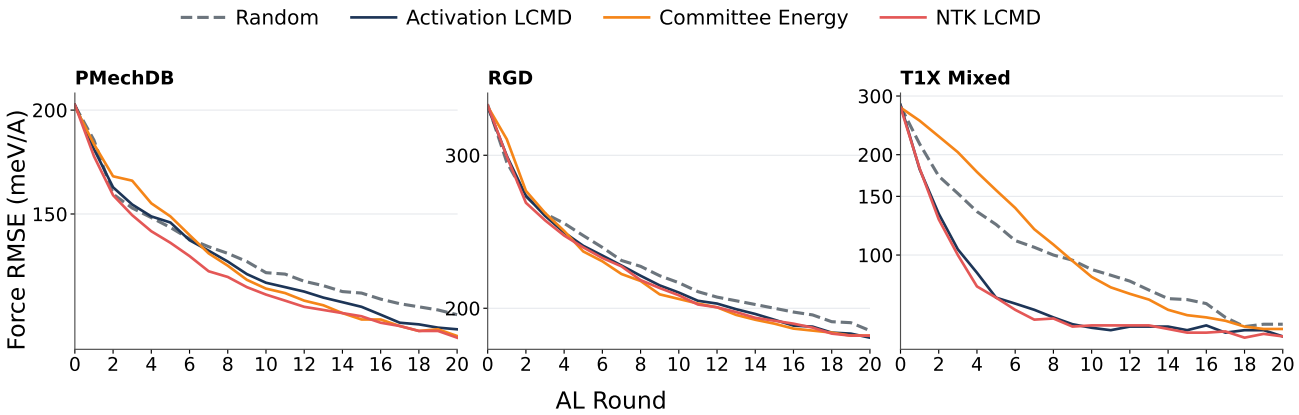


Figure 17. Additional transferability results on PMechDB, RGD, and T1x Mixed. The curves show force RMSE across active-learning rounds.

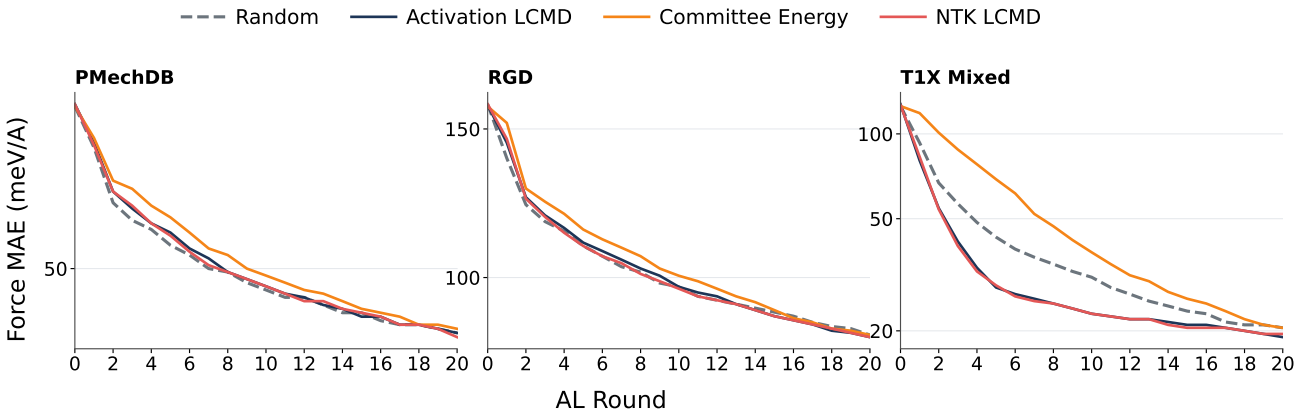


Figure 18. Additional transferability results on PMechDB, RGD, and T1x Mixed. The curves show force MAE across active-learning rounds.

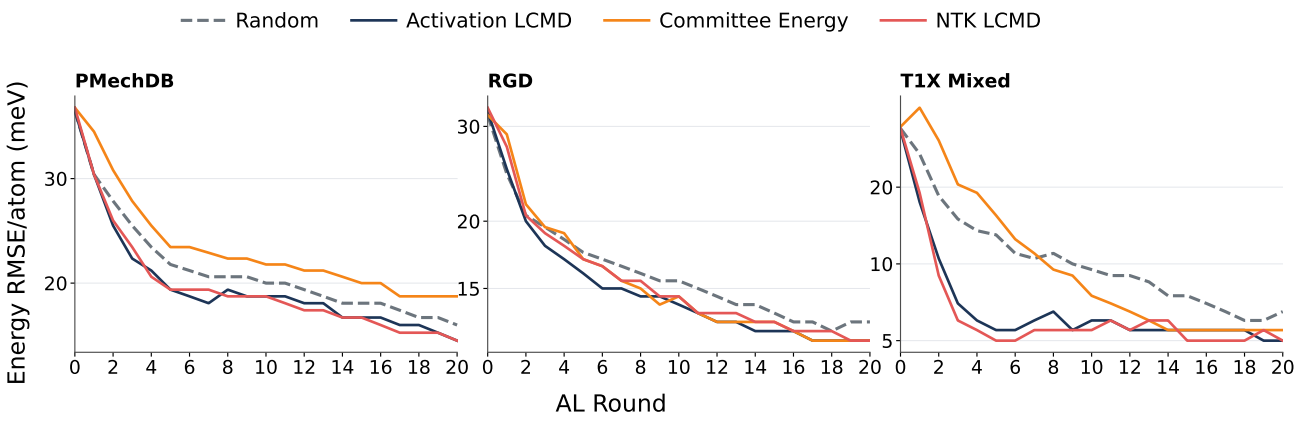


Figure 19. Additional transferability results on PMechDB, RGD, and T1x Mixed. The curves show energy RMSE across active-learning rounds.

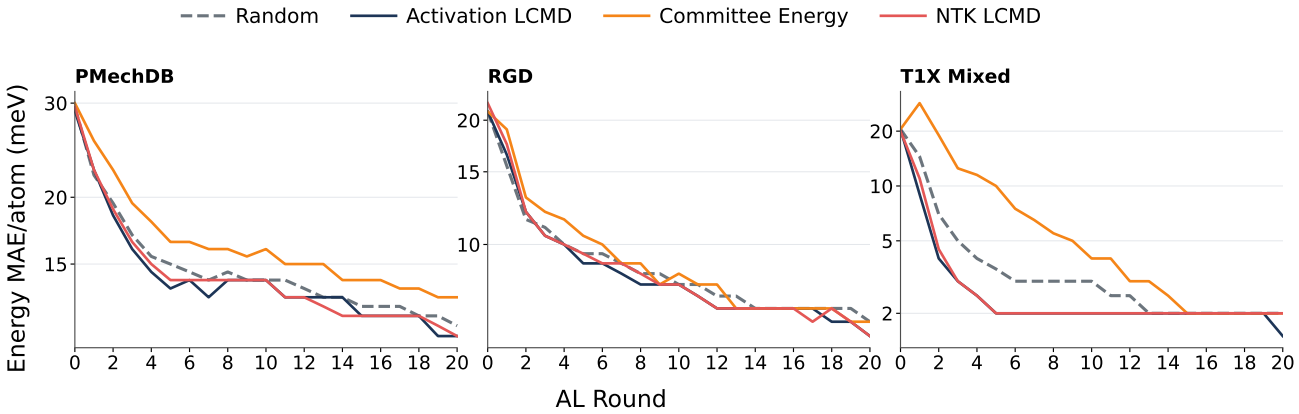


Figure 20. Additional transferability results on PMechDB, RGD, and T1x Mixed. The curves show energy MAE across active-learning rounds.