FEATURE INFORMED BATCH SELECTION MAY ACCELERATE TRAINING AND TUNING OF CHEMICAL FOUNDATION MODELS

Benjamin du Pont*, **Omar Allam***†, **Aayush Singh, & Ang Xiao**
SandboxAQ, 780 High Street Palo Alto, CA, 94301 USA
omar.allam@sandboxaq.com

ABSTRACT

Chemical foundational models pretrained on expansive materials databases have the potential to significantly accelerate materials discovery relative to traditional quantum-mechanical calculations. However, training and even fine-tuning these models remains expensive and not widely accessible due to the vast amount of data typically required and the complexity of optimization. To address this, we propose a framework for improving the efficiency of the training and fine-tuning of foundational models by prioritizing the most informative training samples and density functional theory (DFT) calculations through **F**eature **I**nformed **B**atch Selection - FIBonAQi. Specifically, by using online batch selection strategies, such as **Div**ersified **B**atch **S**election (DivBS) (Hong et al., 2024), originally tested on vision and natural language processing models, FIBonAQi aims to make training and tuning of foundation ML models in chemistry more data efficient relative to conventional uniform sampling strategies. We evaluate the proposed approach both by training from scratch and fine-tuning scenarios. While more extensive testing is needed, preliminary results suggest that online batch selection strategies such as FIBonAQi-DivBS may be able to improve data efficiency in chemical foundation model training.

1 INTRODUCTION

Atomistic simulations have experienced a significant paradigm shift with recent developments in chemical graph neural networks (GNNs). Machine learning interatomic potentials (MLIPs) are one family of such models designed to estimate the properties of chemical systems; these frequently include energies, interatomic forces, and adsorption energies (Chanussot et al., 2021). MLIPs significantly accelerate materials discovery workflows and reduce their cost by offering a cheaper alternative to traditional quantum mechanical calculations such as density functional theory (DFT). Further, "foundation" MLIPs trained on diverse materials databases offer a more universal applicability and generalization. However, training and fine-tuning graph-based MLIPs is expensive, due to the large number of parameters that must be optimized, and the amount of training data they require for accurate inference, which is frequently generated by such quantum mechanical calculations. This is particularly true of GNNs, which are notoriously data-hungry. Active learning loops attempt to circumvent this problem by generating training samples only when they are requested by the model. Yet this remains costly, as there are few widespread methods to identify samples worth generating via DFT. One such method is DIRECT sampling (Qi et al., 2024), which selects a diverse subset of training data by clustering structures in a reduced feature space before labeling them with DFT.

In addition to selecting a training set, efficient sample selection can also be applied dynamically during training. Specifically, online batch selection algorithms, where batches are curated based on some criteria with the object of improving or accelerating training, have seen success in other applications of machine learning. For instance, Loshchilov & Hutter (2016) find that exponentially weighting a training sample's selection probability significantly accelerates model training on

*These authors contributed equally.
†Correspondence to Omar Allam.

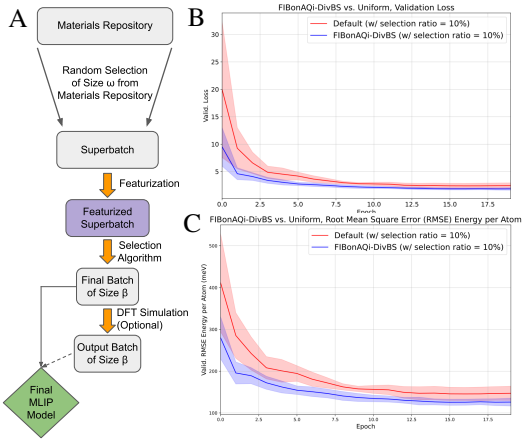

Figure 1: **A:** Schematic of the FIBonAQi framework. **B, C.)** A comparison of FIBonAQi-DivBS sampling to standard uniform sampling on the task of fine-tuning the small MACE-MP-0 foundation model on the OC20NEB, both with a selection ratio of 10% and a training batch size of 10. Solid lines show means (10 seeds) and shaded areas indicate one standard deviation. One epoch is the interval over which the superbatches have encompassed the entire dataset.

MNIST classification tasks when optimized via the ADAM and AdaDelta algorithms. More recently, Hong et al. (2024) proposed an online batch selection scheme, termed **Div**ersified **B**atch **S**election (DivBS), which creates batches by approximately and stochastically maximizing their orthagonalized representativeness w.r.t. a chosen sample featurization scheme. They find that the gradients of a sample's loss w.r.t. the model's last layer parameters serve as a particularly effective featurization scheme, ameliorating training on both computer vision and natural language processing tasks.

Though batch curation for chemical MLIP active learning has been studied (Zaverkin et al., 2022; Bailey et al., 2023), we note that there is a relative dearth of literature on the application of online batch selection for training and fine-tuning in the chemical domain. Building on the success of online batch selection algorithms in other machine learning applications, and in specific chemical contexts, we propose a unified framework for both accelerating the fine-tuning of graph-based MLIPs, and for designing more effective active learning loops. We term this **F**eature **I**nformed **B**atch Selection (FIBonAQi). Our approach addresses three key objectives:

The unification active learning and online batch selection: FIBonAQi bridges the previously disparate domains of active learning loop improvement and batch selection by providing a general means of enhancing MLIP training; we believe that these problems may be somewhat isomorphic, and the presentation of a unified framework for such may therefore have utility.

The highlighting of the need for research in diverse chemical featurization schema: By proposing FIBonAQi, we draw attention to the need for more research on chemical featurization algorithms. Materials science is one of the few domains where large, well-developed foundation models are being more widely available, removing the constraint of model-free featurization faced by Hong et al. (2024). As such, diverse featurization strategies may yield context-dependent improvements in MLIP training.

The facilitation of future innovations: FIBonAQi provides a common set of theoretical interfaces which encompass existing batch selection algorithms and active learning strategies, and elucidates ways these may be combined to ameliorate MLIP training.

2 FIBONAQI

FIBonAQi is an online batch selection framework designed to accelerate MLIP finetuning, improve model performance, and ameliorate active learning loops via efficient, on-line batch selection. The framework is summarized by the following steps, and is illustrated graphically in Fig. 1A.

Superbatch Selection: Sets of samples of size ω, termed superbatches, are randomly drawn from the selected materials repository; we have termed these superbatches to reflect their role as "batches" from which the final training batches are drawn. Usually, instances of this framework are initialized with some selection ratio, r. This is used to determine ω, given the desired final batch size β, via the equation $\omega = \frac{\beta}{r} : r \in (0, 1]; \omega, \beta \in \mathbb{Z}$ (1), where allowed values of r are restricted to those that yield an integer ω.

Superbatch Featurization: All samples within the current superbatch are vectorized via a user-selected featurization scheme. In the case of DivBS, this is simply the gradients of the model's loss function w.r.t. the weights and biases of the last layer of the model being trained.

Batch Selection: Batches of size β are drawn from the superbatch according to a user-chosen selection algorithm, which accepts the featurized superbatch as an input. We focus this paper on the greedy selection algorithm proposed by Hong et al. (2024).

(Optional) DFT labeling: If the used materials repository is composed of data generated by another pretrained MLIP, FIBonAQi then selects the most informative batches, which are subsequently labeled using DFT. These "up-scaled" data points then become the final batch which is provided to the model. This setup can naturally extend to an active learning loop (Qi et al., 2024; Zaverkin et al., 2022), where the MLIP iteratively improves with newly acquired high-fidelity data.

The DivBS algorithm can be considered as an instance of the feature-informed batch selection approach used in FIBonAQi, where sample featurization is performed using any desired means, and the selection scheme is the approximate orthogonalized representativeness maximization algorithm described by Hong et al. (2024). This method is designed to select an array of sample structures from a chemical repository (which can then be optionally refined via DFT simulation).

3 Experiments

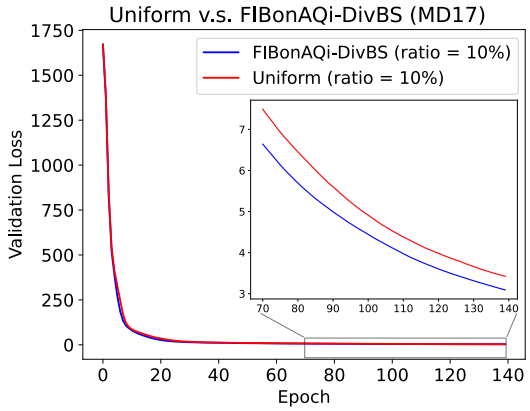

Figure 2: FIBonAQi-DivBS vs. uniform sampling for training MACE on MD17@CCSD over 136 epochs (batch size 10, 4750 samples). Inset: final half of training (1 seed).

We conduct experiments on two datasets: MD17@CCSD (Chmiela et al., 2017) and OC20NEB (Wander et al., 2024). FIBonAQi is tested both on from-scratch training and fine-tuning a pretrained model, for the MD17@CCSD and OC20NEB datasets, respectively. We also train and validate a MACE model from scratch on exclusively benzene and aspirin from the MD17@CCSD dataset, to further probe the versatility of FIBonAQi-DivBS. Lastly, as done by Hong et al. (2024), we vary the selection ratio r of FIBonAQi-DivBS to explore its effect on training efficacy. We perform this experiment using benzene samples drawn from the MD17@CCSD dataset.

MD17@CCSD. The MD17@CCSD dataset consists of small organic molecules including aspirin (at CCSD level), as well as benzene, malonaldehyde, toluene, and ethanol (at CCSD(T) level). It provides atomic coordinates, energies, and interatomic forces. Similar to conventional splits in the literature for this dataset, we train MACE (Batatia et al., 2023) from scratch on this dataset using 950 samples per molecule for training and 50 for validation, all selected uniformly. This training was

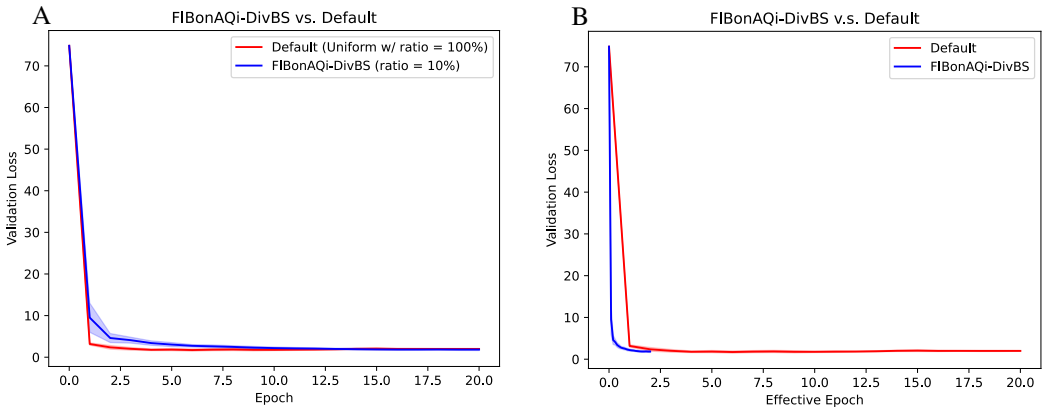

Figure 3: **A:** Uniform (100%) vs. FIBonAQi-DivBS (10%) tuning of MACE-MP-0 on 800 OC20NEB samples. **B:** Validation loss as shown in (A) with an adjusted, "effective" epoch, such that one effective epoch marks training on a number of samples equivalent to the size of the entire training dataset. (10 seeds).

done with an energy and force loss weight ratio of 1 to 100, a radial cutoff of 5.0 Å, 128 channels, hidden irreps of 128x0e + 128x1o, and a learning rate of 0.001.

Benzene & Aspirin, MD17@CCSD. As described above, we train MACE from scratch using 950 samples drawn randomly from the benzene and aspirin molecules in the MD17@CCSD dataset. We similarly use 50 samples uniformly drawn from each molecule as validation for that molecule. This training was done with an energy and force loss weight ratio of 1 to 100, a radial cutoff of 5.0 Å, 256 channels, hidden irreps of 256x0e+256x1o+256x2e, and a learning rate of 0.01 (see Fig. A1).

OC20NEB. To test FIBonAQi in a realistic fine-tuning challenge, we use OC20NEB, which consists of near-equilibrium and transition-state structures relevant to catalysis. Fine-tuning on this dataset evaluates whether FIBonAQi can efficiently guide batch selection in a setting where a model has already learned a general representation and must adapt to a new but related distribution. We fine-tune the small MACE-MP-0 (Batatia et al., 2024) on a small subset of 800 randomly sampled OC20NEB configurations and evaluate its performance using a separate 100-sample validation set.

Varying Selection Ratio. To study how the selection ratio r impacts training efficacy, we train several MACE models from scratch on 950 samples drawn uniformly from the MD17@CCSD dataset. We use the same settings as when training on aspirin and benzene alone, except that we vary the selection ratio r between trials (see Fig. A2).

Preliminary results suggest that FIBonAQi-DivBS may consistently outperform standard uniform batch selection when provided with similar data selection ratios. This may be seen by the blue line in Fig. 1B, which represents the validation loss of MACE-MP-0 fine-tuning via the FIBonAQi-DivBS algorithm, whose mean lies strictly below the red line, which represents uniform sampling. This trend appears to hold across different datasets, as evident in Fig. 2, which illustrates the training of Mace on 4750 samples drawn from the MD17@CCSD. Of note is Fig 3A, which initially seems to illustrate a failure of FIBonAQi-DivBS to outperform standard selection. However, one must consider that during each "epoch" depicted, the model has access to only 10% of the data under the FIBonAQI-DivBS scheme that the model has under its default scheme. As such, this behavior is somewhat expected. When one adjusts for this by scaling relative to the amount of samples the model has been trained on, as illustrated in Fig. 3B, one observes comparative performance broadly similar to that shown in Fig. 1B. This behavior is not, innately, surprising. The gradient of the loss of a particular sample w.r.t. a model's last layer directly corresponds to how the model's output space must adjust. As such, gradient-based featurization schemes may perform better than input-based ones at loss minimization.

Of concern is FIBonAQi's computational overhead; each FIBonAQi-DivBS-based batch selection requires a forward pass and a backward pass through the weights and biases of the model's last layer for each sample in the superbatch. The theoretical overhead of such back passes is small compared to

training the model. Yet, the need for forward passes limits the gains available to last-layer gradient-based sample selection methods, like DivBS, in our framework (Hong et al., 2024). As increasing the selection ratio r directly reduces superbatch size ω, per Eqn. 1, and thus the computational overhead of FIBonAQi-DivBS, we therefore study the effect of r on model performance. Preliminary results suggest model performance may decrease as r increases, per the results shown in A2. However, many more seeds must be averaged over for this result to be definitive. Interestingly, all values of r seem to yield approximately better results than uniform selection. However, we cannot yet rule out the effect of statistical noise due to a dearth of data.

4 DISCUSSION

By incorporating the DivBS algorithm (Hong et al., 2024) into FIBonAQi, we propose a framework for more efficiently tuning chemical foundation models. FIBonAQi takes steps toward more efficient MLIP training by narrowing the focus to batches with greater predictive value, while simultaneously supporting a wide range of model architectures and optimization settings. While our results show promise, they remain preliminary and require further validation across different datasets, architectures, and seeds. Fig. A1 and A2 for instance, display that models trained with FIBonAQi-DivBS seem to outperform those trained by standard methods. Still, the impact of statistical noise cannot yet be excluded, especially in the case of Fig. A2, where the effect of varying the selection ratio remains unclear. Future work will extend our analysis to additional datasets, metrics, and model variants (Liao et al., 2024; Musaelian et al., 2023; Batzner et al., 2022; Chen & Ong, 2022).

While FIBonAQi-DivBS appears to be generally more data-efficient than uniform selection, a more rigorous examination of its computational overhead is required. Hong et al. (2024) find that DivBS is significantly superior to uniform selection in the performance-speedup trade-off; however, the difficulty of accessing per sample gradients with PyTorch-based MLIPs has required us to leave this to future work. We also intend to examine alternative featurization approaches for both batch selection and active learning loops. These could be better suited to a uniquely chemical context; for example, a M3GNet encoder could be employed, similar to what was done by Qi et al. (2024). Additional effort is required to reduce the computational overhead of FIBonAQi, which is not necessarily intensive in theory, but is so in its current incarnation. Still, FIBonAQi-DivBS requires the entire training materials repository used to be evaluated each epoch, representing a non-insignificant cost for most practical applications. As such, model distillation may be an interesting direction of future research. Namely, if a coarse approximation of the model's inference on each sample may be used to approximate the gradient of that loss w.r.t. the last layer's parameters, for example via model distillation, this would constitute a significant improvement over FIBonAQi-DivBS.

It should also be noted that if FIBonAQi generally yields similar convergent performance as uniform selection, it may be possible to inadvertently yield worse model validation performance in that high-epoch limit; this may be because FIBonAQi could only train the model some fixed ratio of the dataset per "epoch." As such, per effective epoch, it is probable that samples are repeated, especially as gradients approach zero. Including input featurization may help resolve this, as well as a selection ratio schedule. It should be noted that as the selection ratio approaches 100%, FIBonAQi's overhead typically decreases. This is because, in that limit, superbatch size decreases to meet the size of the final batch, meaning less featurization computation must occur. Moreover, even if FIBonAQi does not result in improved convergent performance, it may still find utility in architecture design settings, where it is desirable to train a model on significant amounts of data without achieving convergence to gain an approximate understanding of a model's capabilities. Lastly, it is widely understood that, in the large data limit, even marginal improvements in validation performance often necessitate much larger training datasets (Merchant et al., 2023). If, therefore, FIBonAQi yields only marginal improvements at convergence, it represents a significantly larger effective training set size.

REFERENCES

Michael Bailey, Saeed Moayedpour, Ruijiang Li, Alejandro Corrochano-Navarro, Alexander Kötter, Lorenzo Kogler-Anele, Saleh Riahi, Christoph Grebner, Gerhard Hessler, Hans Matter, Marc Bianciotto, Pablo Mas, Ziv Bar-Joseph, and Sven Jager. Deep batch active learning for drug discovery. *bioRxiv*, 2023. doi: 10.1101/2023.07.26.550653. URL https://www.biorxiv.org/content/early/2023/07/29/2023.07.26.550653.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. Mace: Higher order equivariant message passing neural networks for fast and accurate force fields, 2023. URL https://arxiv.org/abs/2206.07697.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Samuel M. Blau, Vlad Cărare, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Jad Jaafar, Stephan Hofmann, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry, 2024. URL https://arxiv.org/abs/2401.00096.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nat. Commun.*, 13(1):2453, May 2022.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (oc20) dataset and community challenges, 2021. URL https://doi.org/10.1021/acscatal.0c04525.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nat. Comput. Sci.*, 2(11):718–728, November 2022.

Stefan Chmiela, Alexandre Tkatchenko, Huziel E. Sauceda, Igor Poltavsky, Kristof T. Schütt, and Klaus-Robert Müller. Machine learning of accurate energy-conserving molecular force fields. *Science Advances*, 3(5), May 2017. ISSN 2375-2548. doi: 10.1126/sciadv.1603015. URL http://dx.doi.org/10.1126/sciadv.1603015.

Feng Hong, Yueming Lyu, Jiangchao Yao, Ya Zhang, Ivor W. Tsang, and Yanfeng Wang. Diversified batch selection for training acceleration, 2024. URL https://arxiv.org/abs/2406.04872.

Yi-Lun Liao, Brandon Wood, Abhishek Das, and Tess Smidt. Equiformerv2: Improved equivariant transformer for scaling to higher-degree representations, 2024. URL https://arxiv.org/abs/2306.12059.

Ilya Loshchilov and Frank Hutter. Online batch selection for faster training of neural networks, 2016. URL https://arxiv.org/abs/1511.06343.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

Albert Musaelian, Simon Batzner, Anders Johansson, Lixin Sun, Cameron J Owen, Mordechai Kornbluth, and Boris Kozinsky. Learning local equivariant representations for large-scale atomistic dynamics. *Nat. Commun.*, 14(1):579, February 2023.

Ji Qi, Tsz Wai Ko, Brandon C Wood, Tuan Anh Pham, and Shyue Ping Ong. Robust training of machine learning interatomic potentials with dimensionality reduction and stratified sampling. *Npj Comput. Mater.*, 10(1), February 2024.

Brook Wander, Muhammed Shuaibi, John R. Kitchin, Zachary W. Ulissi, and C. Lawrence Zitnick. Cattsunami: Accelerating transition state energy calculations with pre-trained graph neural networks, 2024. URL https://arxiv.org/abs/2405.02078.

Viktor Zaverkin, David Holzmüller, Ingo Steinwart, and Johannes Kästner. Exploring chemical and conformational spaces by batch mode deep active learning. *Digit. Discov.*, 1(5):605–620, 2022.

APPENDIX: SUPPLEMENTAL EXPERIMENTS

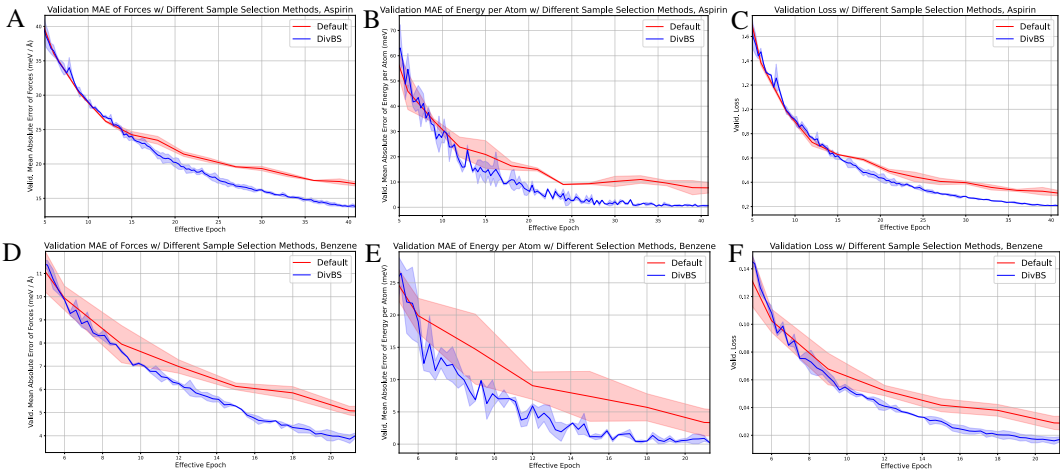

Figure A1: MACE training comparison on **A-C)** Aspirin and **D-F)** Benzene: Default sampling (Red) vs. FIBonAQi-DivBS (Blue, 10% selection ratio). Trained on 950 MD17@CCSD samples, with a batch size of 2, and validated on 50. Metrics shown vs. effective epochs, where one "effective" epoch is equivalent to training on a number of samples equal to that in the data set, as defined in Fig.3. **A,D)** Force MAE (meV/Å), **B,E)** Energy MAE, **C,F)** Validation Loss. Solid colored lines represent means, while shaded areas represent 1 standard deviation. Default results are averaged over 3 seeds, while those of FIBonAQi-DivBS herein are averaged over 2. The first 5 effective epochs are omitted from the figure.

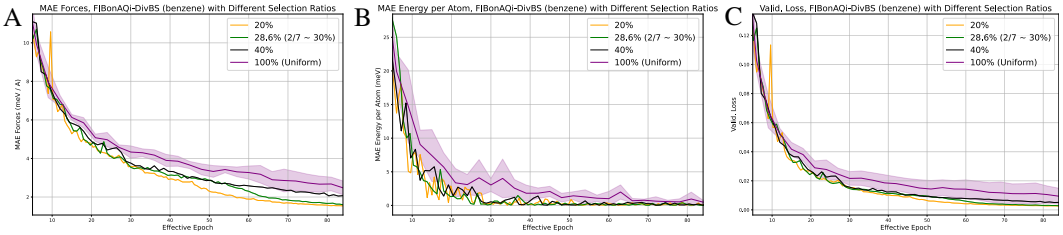

Figure A2: MACE training comparison on benzene, using FIBonAQi-DivBS, with various selection ratios r. Model performance is evaluated on **A.)** mean absolute error of forces (meV/Å), **B.)** mean absolute error of energy per atom (meV), and **C.)** validation loss. Uniform (r = 100%) selection was averaged over 3 seeds, whereas all other trials comprise of 1 seed. An "effective" epoch is defined as in Fig. 3, and refers to the duration over which the model has been trained on a number of samples equivalent to the total size of the training-set. The first 5 effective epochs are omitted from the figure, as their inclusion skews the Y axis scaling toward over emphasizing initial errors, which may be largely a function of model initialization.