Unifying Force Prediction and Molecular Conformation Generation Through Representation Alignment

Lucas Pinede [*] [1] [2] **Soojung Yang** [*] [3] **Juno Nam** [2] **Rafael Gómez-Bombarelli** [2]

Abstract

Sampling equilibrium conformational ensemble is essential for understanding biomolecular functions. Although i.i.d. ensemble samplers (i.e., Boltzmann emulators) hold great promise, their training strategies have been underexplored. We hypothesize that machine learning interatomic potentials (MLIPs), trained on abundant non-equilibrium data, already capture geometric and energetic information needed for ensemble generation. Inspired by Yu et al. (2025), we introduce a representation alignment loss that regularizes the emulator's hidden states to be similar to those of pretrained MLIP. This simple addition only requires a one-time inference cost of samples yet shortens the training time by 1.5× and yields better sampling performances.

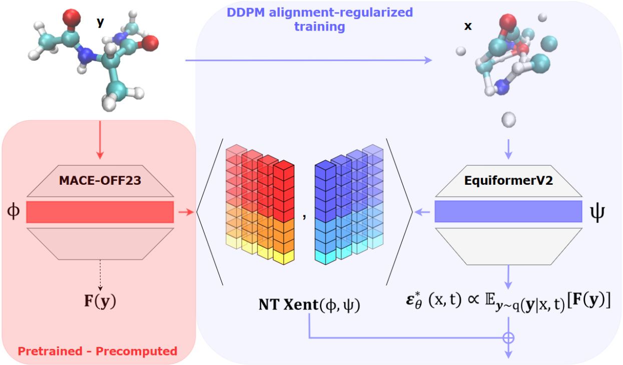

y
DDPM alignment-regularized training
x
MACE-OFF23
ϕ
F(y)
Pretrained - Precomputed
EquiformerV2
ψ
NT Xent(ϕ, ψ)
$\varepsilon_\theta^*(x,t) \propto \mathbb{E}_{y \sim q(\mathbf{y}|x,t)}[\mathbf{F}(\mathbf{y})]$

Figure 1. Overview of our framework. We accelerate the training of Boltzmann emulators via representation alignment (REPA) with pretrained, foundational MLIP model embeddings.

1. Introduction

Sampling the conformational ensemble of biomolecules is fundamental to understanding biological functions. These ensembles enable free energy calculations, uncover mechanisms of binding and folding, and reveal rare but functionally important states. However, most experimental techniques can be costly, and often yield only static snapshots or low-resolution ensemble averages. When a force field is available, the ensemble can be computationally sampled using molecular dynamics (MD), which integrates Newton's equations of motion based on interatomic or coarse-grained forces. In practice, however, MD is hindered by slow transitions between metastable states, making exhaustive sampling computationally prohibitive for many biomolecular systems.

Deep generative models offer a promising alternative because they are capable of i.i.d sampling of equilibrium structures, bypassing the challenge of metastability. A recently emerging class of models known as *Boltzmann emulators* (or simply *emulators*) — including BioEmu (Lewis et al., 2024), AlphaFlow (Jing et al., 2024), and ETFlow (Hassan et al., 2024) — leverages the expressiveness of score-based and flow-based generative frameworks to directly model equilibrium (Boltzmann) distributions.

Training Boltzmann emulators suffer from lack of data because they require near-equilibrium ensembles. By contrast, machine learning interatomic potentials (MLIPs) (Batatia et al., 2023b; Liao et al., 2024b) can exploit more abundant non-equilibrium structures and have better access to a broader transferability from the same amount of data because force prediction is largely local. Recent MLIPs are becoming more "foundational", trained on chemically diverse datasets (Mazitov et al., 2025), aiming to learn universal chemistry.

Yu et al. (2025) recently proposed that the major bottleneck in training of visual diffusion models is the quality of the learned hidden representation. They showed that adding an auxiliary loss to align a diffusion model's hidden states to a pretrained image encoder accelerated the training by 17.5×. Since pretrained MLIPs already encode rich information about molecular geometry and energetics, if those pretrained

[*] Equal contribution [1] Chimie ParisTech, PSL University, Paris, France [2] MIT, Department of Materials Science and Engineering, Cambridge, MA, United States [3] MIT, Computational and Systems Biology, Cambridge, MA, United States. Correspondence to: Rafael Gómez-Bombarelli <rafagb@mit.edu>.

Proceedings of the Workshop on Generative AI for Biology at the 42^{nd} *International Conference on Machine Learning*, Vancouver, Canada. PMLR 267, 2025. Copyright 2025 by the author(s).

representations could be leveraged, emulator training could be cheaper and faster.

The mathematical link between scores and forces further strengthens the emulator–MLIP connection. In score-based diffusion models — the backbone of most emulators — at sufficiently low noise level, the score approximates the negative gradient of the potential energy, which is the physical force. Arts et al. (2023) leveraged this property to extract a coarse-grained force field directly from a pretrained diffusion model. Several recent MLIPs are trained with a denoising auxiliary task (e.g., DeNS (Liao et al., 2024a)) that reconstructs noiseless structures from noisy structures and force labels. Together, these studies point to a connection between force prediction and conformer generation, suggesting that knowledge for one task can be leveraged to improve the other.

The opposite direction, however — turning a pretrained force predictor into an i.i.d. generator of equilibrium structures — has barely been explored. Because modern MLIPs are trained with massive amount of non-equilibrium data and therefore generalize well, transferring their knowledge to emulators could sidestep the scarcity of equilibrium samples.

We therefore propose a simple yet effective training strategy: align the representation spaces of the two models. If force prediction and conformer generation rely on a shared representation of molecular geometry and energy, aligning an emulator's hidden states toward those of a pretrained MLIP should provide an informative prior. With our experiments, we demonstrate that adding an alignment loss delivers more accurate emulators at much cheaper cost.

Using alanine dipeptide as our test system, we:

- **Quantify representation overlap between force prediction and conformer generation task.** We track the alignment between the hidden states of an emulator and those of a pretrained MLIP during training. Alignment rises with the conformer generation performance, suggesting that the generative model training is a search for suitable representations, and that the pretrained MLIP provides one almost "for free".
- **Accelerate training.** Adding a simple alignment loss accelerates the convergence by **1.5**× for the emulator to reproduce the Boltzmann distribution.

2. Method

2.1. Force–score relationship

To learn the Boltzmann distribution over the space of 3D atomic coordinates, we leverage the denoising score matching framework (Song et al., 2021). The forward SDE gradually injects noise into the positions through a Wiener process W:

$$d\mathbf{x} = f(\mathbf{x}, t)dt + g(t)dW. \tag{1}$$

The reverse SDE maps back the Gaussian distribution to the data distribution by following a score function:

$$d\mathbf{x} = [f(\mathbf{x}, t) - g^2(t)\nabla_{\mathbf{x}} \log p_t(\mathbf{x})]dt + g(t)d\bar{W}. \tag{2}$$

The score is approximated by a neural network: $s_\theta(\mathbf{x}_t, t) \approx \nabla_{\mathbf{x}} \log p_t(\mathbf{x})$.

As pointed out in Arts et al. (2023), if the target distribution follows the Boltzmann distribution $q(\mathbf{x}) \propto e^{-\beta V(\mathbf{x})}$, where $\beta = (k_B T)^{-1}$ and $V(\mathbf{x})$ is potential energy, then at the lowest noise level, the optimal score will capture the forces: $s^*_\theta(\mathbf{x}, t = 0) \approx \nabla_{\mathbf{x}} \log q(\mathbf{x}) = -\beta \nabla V(\mathbf{x}) = \beta \mathbf{F}$.

With Gaussian perturbation, the density at noise level t of Variance Preserving SDE (VP SDE) is $p_t(\mathbf{x}) = \int q(\mathbf{y})\mathcal{N}(\mathbf{x}; \alpha_t \mathbf{y}, \sigma_t^2 \mathbf{I})d\mathbf{y}$. Then, the optimal score at t becomes

$$s^*_\theta(\mathbf{x}, t) \propto \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y}|\mathbf{x}, t)}[\mathbf{F}(\mathbf{y})] \tag{3}$$

, where $q(\mathbf{y}|\mathbf{x}, t) \propto q(\mathbf{y})\mathcal{N}(\mathbf{x}; \alpha_t \mathbf{y}, \sigma_t^2 \mathbf{I})$, and the expectation of force $\mathbf{F}$ is taken over clean conformers $\mathbf{y}$ that could have produced the noisy sample $\mathbf{x}$. We provide a derivation based on Bortoli et al. (2024)'s formulation, in Appendix.

2.2. MLIP alignment as an emulator training strategy

Yu et al. (2025) showed that aligning a diffusion model's hidden representations with those of a pretrained vision encoder (**REPA** framework) accelerates convergence by about 18×. Motivated by this result, we apply the REPA learning objective for emulator training. We align our emulator's hidden states with the representations of a foundation MLIP, distilling its pretrained knowledge of molecular geometry and energy into a generative model.

2.2.1. REPRESENTATION ALIGNMENT (REPA) LEARNING OBJECTIVE

Given a hidden state $\mathbf{h}_{\theta,t} = f_\theta(\mathbf{x}_t) \in \mathbb{R}^{N \times D}$ from the denoising model and a representation $\mathbf{z} = f(\mathbf{y}) \in \mathbb{R}^{N \times D'}$ computed from the noiseless conformation $\mathbf{y}$ by the pretrained MLIP model, we define the alignment loss following (Yu et al., 2025):

$$\mathcal{L}_{\text{REPA}}(\theta, \phi) := -\mathbb{E}_{\mathbf{y},\epsilon,t}\left[\frac{1}{N}\sum_{n=1}^{N} \text{sim}(\mathbf{z}^{[n]}, h_\phi(\mathbf{h}_{\theta,t}^{[n]}))\right], \tag{4}$$

where h_ϕ is a projection feed forward layer that matches the dimensionalities between the intermediate representations while providing flexibility. The final loss is

$$\mathcal{L} := \mathcal{L}_{\text{DDPM}} + \lambda \mathcal{L}_{\text{REPA}} \tag{5}$$

where λ is a hyperparameter.

For the similarity metric, we chose to use a contrastive loss that has been previously used for molecular property prediction task (Stärk et al., 2022), which is a derivative of the NT-Xent loss (Chen et al., 2020).

$$\mathcal{L}_{\text{REPA}} = -\frac{1}{N}\sum_{n=1}^{N}\left[\log\frac{\exp\left(\text{sim}(\mathbf{z}^{[n]}, h_\phi(\mathbf{h}_{\theta,t}^{[n]}))/\tau\right)}{\sum_{\substack{k=1\\k\neq i}}^{N}\exp\left(\text{sim}(\mathbf{z}^{[k]}, h_\phi(\mathbf{h}_{\theta,t}^{[n]}))/\tau\right)}\right] \tag{6}$$

2.2.2. Aligning MACE with an EquiformerV2 emulator

MACE (Batatia et al., 2023b) is a E(3)-equivariant message passing network that predicts per-atom energies and forces from 3D coordinates. For our alanine dipeptide experiments, we align our emulator with **MACE-OFF23** model (Kovács et al., 2025), trained on 85% of the SPICE dataset (Najibi & Goerigk, 2018) containing drug-like molecules, peptides, and larger organic molecules from QMugs dataset (Isert et al., 2021). We also test our method with **MACE-MP0** model (Batatia et al., 2023a), trained on bulk crystal structures of the MPTrj dataset (Deng et al., 2023).

We adopt **EquiformerV2** model as our score model architecture (Liao et al., 2024b). Since EquiformerV2 is originally an MLIP model, we add (i) time (noise level) conditioning and (ii) positional encoding of atom orders so that the network can serve as a score model. Our score model takes the atomic positions and atomic numbers as input, initializes a radius graph based on a predefined cutoff, and performs graph attention (Brody et al., 2022) on *irreps* using equivariant transformations. The predicted type-$l = 1$ vector lies in $\mathbb{R}^3$ and approximates the score $\nabla_x \log p_t(x)$ at the current noise level t.

Both the pretrained and the denoising model are at least $SE(3)$ equivariant graph neural networks, meaning that for $x \in X$ and $f(x) \in Y$, then $f(D_X(g)x) = D_Y(g)f(x), \forall g \in SE(3)$, where $D_X(g)$ is the representation of the element $g \in SE(3)$ in X space. The models use irreducible representations (*irreps*) as internal representation for the data. This corresponds to expressing the hidden features in the basis where the representations of $SO(3)$ transformations can be expressed as a block diagonal matrix made of irreducible matrices called Wigner-D matrices. The resulting space can be expressed as a direct sum of type$-l$ vectors $\bigoplus_0^{L_{max}} v^l$ each $(2l+1)$ dimensional vectors. This projection is performed with spherical harmonics of up to a fixed maximum order L_{max}, which we used the identical value $L_{max} = 2$ for both models. Information is aggregated along the radius graph edges with the Clebsch-Gordan tensor product, which gives the decomposition of the resulting tensor into *irreps* through the Clebsch-Gordan coefficients.

To reconcile the mismatch in feature dimensions during alignment, Yu et al. (2025) used feed forward layers to project the pretrained encoder representation onto the diffusion model's hidden dimensions. To preserve the equivariance during projection, we use an equivariant feed forward layer from the EquiformerV2 architecture. Notably, type-l vectors are transformed separately and S^2 activations are used.

2.3. Experimental setting

We chose alanine dipeptide as our benchmark system. A 250 ns, fully equilibrated MD trajectory was downloaded from the **mdshare** repository (Nüske et al., 2017). From this trajectory we randomly selected 5k frames for training. Before training begins, we compute MACE embeddings for all noiseless training frames. This requires only a one-time inference cost.

We train emulators with and without the REPA alignment. Throughout training, we monitored the MLIP–emulator representation alignment with Centered Kernel Alignment (CKA) metric (Kornblith et al., 2019), which is widely used for quantifying representation alignment. Generative performance was assessed by drawing 100k samples from each emulator and comparing their ϕ backbone dihedral distribution to the 250k-frame reference ensemble. We quantified the discrepancy with Jensen–Shannon divergence (JSD).

3. Results

3.1. Do force prediction and conformer generation task share similar representation space?

Figure 1 tracks the CKA alignment between the Boltzmann emulator score model and the pretrained MACE representations during training. Even without the representation alignment, the alignment grows steadily, implying that the representations required for conformation generation converge toward those used for force prediction. As a result of adding the REPA auxiliary loss $\mathcal{L}_{\text{REPA}}$, alignment improves more rapidly and converges at a higher value.

We further analyze the representation alignment between the score model and the pretrained MACE model as a function of noise level in the diffusion process. As expected, regardless of the REPA regularization, alignment is strongest at lower noise levels, where the score of the diffusion model more closely approximates the true physical force. Alignment also generally improves with more training steps. Notably, without REPA the trend is reversed at zero noise (clean samples) — more training reduces the alignment.

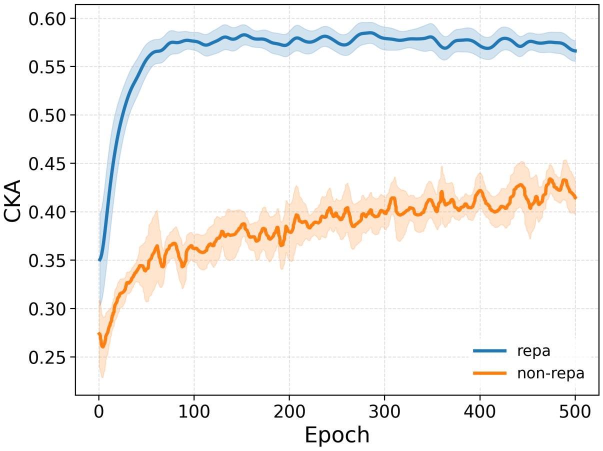


Figure 2. Train epoch vs. CKA alignment between the score model and the pretrained MACE.

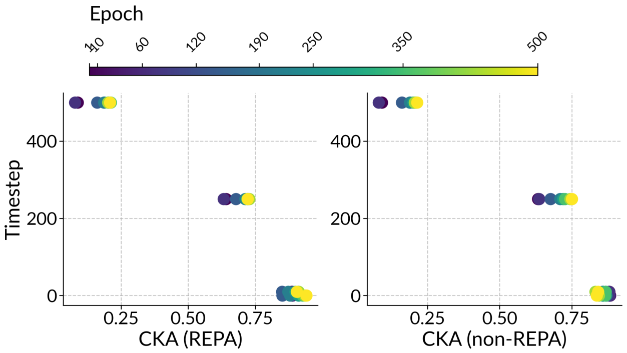


Figure 3. Noise level vs. CKA alignment.

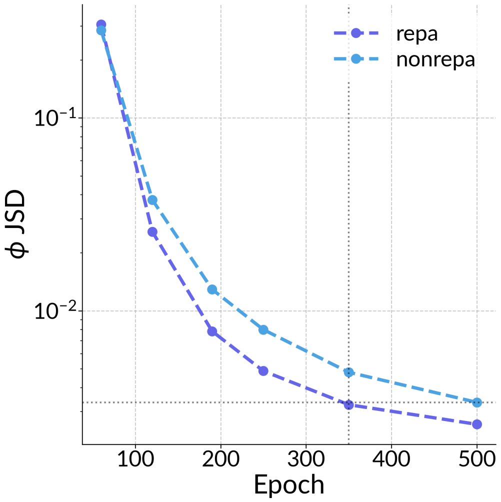


Figure 4. epoch vs. JSD (ϕ). The emulator trained with the REPA loss (**repa**) shows faster and better performance gain than the baseline without alignment (**nonrepa**). **repa** reaches the final performance level of **nonrepa** in roughly two-thirds of the training time (1.5× faster).

3.2. Does representation alignment accelerate the generative model convergence?

To show how the accelerated and strengthened alignment contribute to the performance gain, we monitor the quality of the generated conformational ensemble. Incorporating the REPA loss drives a faster drop in JSD and achieves a lower final JSD than score-matching alone. Ablation studies over the $\mathcal{L}_{\text{REPA}}$ hyperparameters λ and τ, as well as the layer and the number of dimensions for the alignment are present in Appendix. In every setting, REPA regularization only improves the performance, except when we use an extremely large λ weight for $\mathcal{L}_{\text{REPA}}$.

Notably, the temperature hyperparameter τ in the NT-Xent loss has a strong impact on the performance of the method. We propose to partially align the embeddings to incorporate more flexibility in the alignment and achieve robustness to τ by slicing $3/4$ of the channels for alignment instead of taking the whole hidden state. Then, given the significant dependency of the alignment on the noise level (see Figure 3), we directly allow the model to learn which channel to align based on a time-conditioned gating mechanism on the channels, thus removing the need to chose the portion of the embeddings to align. In this setting, equation 6 becomes:

$$\mathcal{L}_{\text{REPA}} = -\frac{1}{N}\sum_{n=1}^{N}\left[\log \frac{\exp\left(\text{sim}(\mathbf{z}^{[n]}, h_\phi(g_\psi(t^{[n]}) * \mathbf{h}_{\theta,t}^{[n]}))/\tau\right)}{\sum_{\substack{k=1\\k\neq i}}^{N} \exp\left(\text{sim}(\mathbf{z}^{[k]}, h_\phi((g_\psi(t^{[n]}) * \mathbf{h}_{\theta,t}^{[n]}))/\tau\right)}\right] \tag{7}$$

where g_ψ denotes a sigmoid-activated Multi-layer Perceptron that takes the timestep $t^{[n]}$ as input, and $*$ denotes an element-wise multiplication.

The results shown Figure S6 also demonstrates a slight increase in convergence speed with this method. This simple trick makes the REPA regularization more robust to λ within mild alignment conditions as illustrated in Figure S7.

Finally, we evaluate the method's ability to transfer meaningful information across distinct chemical domains by aligning the denoising model's representation with that of **MACE-MP0**, trained on crystal structures (Figure S8). While the resulting speedup is reduced by approximately 20%, this likely reflects the domain specificity of the MLIP embeddings due to their training on specialized datasets. Nonetheless, the performance gain remains appreciable, suggesting that the representations learned by **MACE-MP0** still encode useful information that our method can leverage to model distributions over organic molecules. Looking forward, the availability of larger and more chemically diverse datasets is expected to enhance the cross-domain transferability of our approach.

4. Discussion

We have shown that Boltzmann emulators train faster and achieve better performance when we align their hidden states with representations from a pretrained MLIP. This result supports the view that force prediction and conformation generation share common representation space and that one task can inform the other with appropriate training strategies. The additional computational cost is modest because MACE representations are computed only once at the start of training for noiseless training frames.

A natural next step is to finetune the MLIP itself as a Boltzmann emulator, leveraging not only its representations but also its weights. Still, for scenarios where MLIP inference is too slow for denoising sampling, representation alignment could be useful as a distillation method. We plan to compare direct fine-tuning and REPA-based distillation in future work.

Alanine dipeptide is a simple system, which likely limits the observable performance gain. As a future work, we will test our method with larger peptides such as chignolin and diverse multi-molecule benchmarks like GEOM-DRUGS (Axelrod & Gómez-Bombarelli, 2022), where representation alignment can have a more pronounced impact. One key question would be whether representation alignment also improves data efficiency, since the scarcity of high quality equilibrium data remains the primary bottleneck for training Boltzmann emulators.

Acknowledgments

The authors thank the reviewers for their helpful comments and suggestions. The authors thank Maurice Weiler for his insight on equivariance in our approach. The authors acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing (HPC, database, consultation) resources that have contributed to the research results reported within this paper.

Impact Statement

Our paper introduces a training strategy for Boltzmann emulators, which can contribute to studying biomolecular functions and designs. We do not expect any direct negative societal impacts from the methods presented here.

References

Arts, M., Garcia Satorras, V., Huang, C.-W., Zügner, D., Federici, M., Clementi, C., Noé, F., Pinsler, R., and van den Berg, R. Two for one: Diffusion models and force fields for coarse-grained molecular dynamics. *Journal of Chemical Theory and Computation*, 19(18):6151–6159, 2023. doi: 10.1021/acs.jctc.3c00702. URL https://doi.org/10.1021/acs.jctc.3c00702. PMID: 37688551.

Axelrod, S. and Gómez-Bombarelli, R. Geom, energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022. doi: 10.1038/s41597-022-01288-4. URL https://doi.org/10.1038/s41597-022-01288-4.

Batatia, I., Benner, P., Chiang, Y., Elena, A. M., Kovács, D. P., Riebesell, J., Advincula, X. R., Asta, M., Baldwin, W. J., Bernstein, N., Bhowmik, A., Blau, S. M., Cărare, V., Darby, J. P., De, S., Pia, F. D., Deringer, V. L., Elijošius, R., El-Machachi, Z., Fako, E., Ferrari, A. C., Genreith-Schriever, A., George, J., Goodall, R. E. A., Grey, C. P., Han, S., Handley, W., Heenen, H. H., Hermansson, K., Holm, C., Jaafar, J., Hofmann, S., Jakob, K. S., Jung, H., Kapil, V., Kaplan, A. D., Karimitari, N., Kroupa, N., Kullgren, J., Kuner, M. C., Kuryla, D., Liepuoniute, G., Margraf, J. T., Magdău, I.-B., Michaelides, A., Moore, J. H., Naik, A. A., Niblett, S. P., Norwood, S. W., O'Neill, N., Ortner, C., Persson, K. A., Reuter, K., Rosen, A. S., Schaaf, L. L., Schran, C., Sivonxay, E., Stenczel, T. K., Svahn, V., Sutton, C., van der Oord, C., Varga-Umbrich, E., Vegge, T., Vondrák, M., Wang, Y., Witt, W. C., Zills, F., and Csányi, G. A foundation model for atomistic materials chemistry, 2023a.

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.

Bortoli, V. D., Hutchinson, M., Wirnsberger, P., and Doucet, A. Target score matching, 2024. URL https://arxiv.org/abs/2402.08667.

Brody, S., Alon, U., and Yahav, E. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=F72ximsx7C1.

Chen, T., Kornblith, S., Norouzi, M., and Hinton, G. A simple framework for contrastive learning of visual representations. In III, H. D. and Singh, A. (eds.), *Proceedings of the 37th International Conference on Machine Learning*, volume 119 of *Proceedings of Machine Learning Research*, pp. 1597–1607. PMLR, 13–18 Jul 2020. URL https://proceedings.mlr.press/v119/chen20j.html.

Deng, B., Zhong, P., Jun, K., Riebesell, J., Han, K., Bartel, C. J., and Ceder, G. Chgnet: Pretrained universal neural network potential for charge-informed atomistic modeling, 2023. URL https://arxiv.org/abs/2302.14231.

Hassan, M., Shenoy, N., Lee, J., Stark, H., Thaler, S., and Beaini, D. Et-flow: Equivariant flow-matching for molecular conformer generation, 2024. URL https://arxiv.org/abs/2410.22388.

Ho, J., Jain, A., and Abbeel, P. Denoising diffusion probabilistic models, 2020. URL https://arxiv.org/abs/2006.11239.

Huh, M., Cheung, B., Wang, T., and Isola, P. The platonic representation hypothesis. In *International Conference on Machine Learning*, 2024.

Isert, C., Atz, K., Jiménez-Luna, J., and Schneider, G. Qmugs: Quantum mechanical properties of drug-like molecules, 2021. URL https://arxiv.org/abs/2107.00367.

Jing, B., Berger, B., and Jaakkola, T. Alphafold meets flow matching for generating protein ensembles. In *Forty-first International Conference on Machine Learning*, 2024.

Kornblith, S., Norouzi, M., Lee, H., and Hinton, G. Similarity of neural network representations revisited. In Chaudhuri, K. and Salakhutdinov, R. (eds.), *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pp. 3519–3529. PMLR, 09–15 Jun 2019. URL https://proceedings.mlr.press/v97/kornblith19a.html.

Kovács, D. P., Moore, J. H., Browning, N. J., Batatia, I., Horton, J. T., Pu, Y., Kapil, V., Witt, W. C., Magdău, I.-B., Cole, D. J., and Csányi, G. Mace-off: Transferable short range machine learning force fields for organic molecules, 2025. URL https://arxiv.org/abs/2312.15211.

Lewis, S., Hempel, T., Jiménez-Luna, J., Gastegger, M., Xie, Y., Foong, A. Y. K., Satorras, V. G., Abdin, O., Veeling, B. S., Zaporozhets, I., Chen, Y., Yang, S., Schneuing, A., Nigam, J., Barbero, F., Stimper, V., Campbell, A., Yim, J., Lienen, M., Shi, Y., Zheng, S., Schulz, H., Munir, U., Clementi, C., and Noé, F. Scalable emulation of protein equilibrium ensembles with generative deep learning. *bioRxiv*, 2024. doi: 10.1101/2024.12.05.626885.

Liao, Y.-L., Smidt, T., Shuaibi*, M., and Das*, A. Generalizing denoising to non-equilibrium structures improves equivariant force fields. *arXiv preprint arXiv:2403.09549*, 2024a.

Liao, Y.-L., Wood, B., Das*, A., and Smidt*, T. EquiformerV2: Improved Equivariant Transformer for Scaling to Higher-Degree Representations. In *International Conference on Learning Representations (ICLR)*, 2024b. URL https://openreview.net/forum?id=mCOBKZmrzD.

Lu, C., Zhou, Y., Bao, F., Chen, J., Li, C., and Zhu, J. Dpm-solver: A fast ode solver for diffusion probabilistic model sampling in around 10 steps, 2022. URL https://arxiv.org/abs/2206.00927.

Mazitov, A., Bigi, F., Kellner, M., Pegolo, P., Tisi, D., Fraux, G., Pozdnyakov, S., Loche, P., and Ceriotti, M. Pet-mad, a universal interatomic potential for advanced materials modeling, 2025. URL https://arxiv.org/abs/2503.14118.

Murphy, A., Zylberberg, J., and Fyshe, A. Correcting biased centered kernel alignment measures in biological and artificial neural networks, 2024. URL https://arxiv.org/abs/2405.01012.

Najibi, A. and Goerigk, L. The nonlocal kernel in van der waals density functionals as an additive correction: An extensive analysis with special emphasis on the b97m-v and ωb97m-v approaches. *Journal of Chemical Theory and Computation*, 14(11):5725–5738, 2018. doi: 10.1021/acs.jctc.8b00842. URL https://doi.org/10.1021/acs.jctc.8b00842. PMID: 30299953.

Nichol, A. and Dhariwal, P. Improved denoising diffusion probabilistic models, 2021. URL https://arxiv.org/abs/2102.09672.

Nüske, F., Wu, H., Prinz, J.-H., Wehmeyer, C., Clementi, C., and Noé, F. Markov state models from short non-equilibrium simulations—analysis and correction of estimation bias. *The Journal of Chemical Physics*, 146 (9), March 2017. ISSN 1089-7690. doi: 10.1063/1.4976518. URL http://dx.doi.org/10.1063/1.4976518.

Song, L., Smola, A., Gretton, A., Borgwardt, K., and Bedo, J. Supervised feature selection via dependence estimation, 2007. URL https://arxiv.org/abs/0704.2668.

Song, Y., Sohl-Dickstein, J., Kingma, D. P., Kumar, A., Ermon, S., and Poole, B. Score-based generative modeling through stochastic differential equations, 2021. URL https://arxiv.org/abs/2011.13456.

Stärk, H., Beaini, D., Corso, G., Tossou, P., Dallago, C., Günnemann, S., and Lió, P. 3D infomax improves GNNs for molecular property prediction. In Chaudhuri, K., Jegelka, S., Song, L., Szepesvari, C., Niu, G., and Sabato,

S. (eds.), *Proceedings of the 39th International Conference on Machine Learning*, volume 162 of *Proceedings of Machine Learning Research*, pp. 20479–20502. PMLR, 17–23 Jul 2022. URL https://proceedings.mlr.press/v162/stark22a.html.

Yu, S., Kwak, S., Jang, H., Jeong, J., Huang, J., Shin, J., and Xie, S. Representation alignment for generation: Training diffusion transformers is easier than you think. In *International Conference on Learning Representations*, 2025.

A. Implementation details

A.1. Alanine dipeptide

We use the AdamW optimizer with constant learning rate 6.10^{-4}, 0.001 weight decay. We use an 0.999 Exponential Moving Average (EMA) decay with an effective batch size of 128, and train for 500 epochs. The λ factor in equation 5 follows a cosine schedule that eventually decreases to 0.0 after warming up to a maximum value for 200 steps from 0.0.

For all experiments, we use a three-layer EquiformerV2 architecture with four attention heads and representations with 64 channels. The radius graph is computed with $r_{cutoff} = 5.0$ Angstrom. Unless specified otherwise, the alignment is performed on the second layer of the denoising model. This was initially motivated by the observations from (Yu et al., 2025). However, radius graph–based equivariant GNNs may behave quite differently from vision and text transformers. These models are inherently local, with an effective cutoff (or receptive field) that gradually increases with network depth. In the context of denoising, access to global context may be crucial for determining the correct score directions, since global semantics tend to degrade more slowly under noise. This is particularly relevant for molecules, where geometric structures can deteriorate rapidly. We thus provide an ablation study for the aligned layer below.

For the time-conditioned gating, we use a 2-layer MLP with a sigmoid output activation that takes the embedded timestep as input. The MLP has the same width as the denoising model.

As MACE has two message passing layers and the last layer contains only invariant *irreps* (energy prediction), we extract the full first layer with all its *irreps*. The dot product operation is performed on the flattened full higher order tensors.

The diffusion is performed with a maximum of $T = 1000$ steps. While sampling, we use DPM Solver-3 for integration with 150 score evaluation steps, and sample $100K$ configurations.

We compute the Jensen-Shannon Divergence (JSD) for the backbone dihedral angle distribution against $250k$ samples from the reference MD simulation with 61 bins.

For the representation alignment, the best performing hyperparameter settings are shown in Table 1

Table 1. Best hyperparameter settings used in our experiments.

Hyperparameter	Fastest JSD drop	Best final JSD
λ	1.0	5.0
τ	0.75	0.75
Similarity Metric	NT-Xent	NT-Xent
Layer Aligned	2	2

B. Ablation study on Alanine dipeptide

We study the effect of the temperature τ from equation 6 in Figure S1, fixing $\lambda = 1.0$. We found that $\tau = 0.75$ was best performing. We note that the better results for $\tau = 0.75$ correlates with a better alignment of about 5% for the validation CKA compared to its $\tau = 0.5$ counterpart. We also perform an ablation study on λ. Figure S2 shows that within reasonable scale of regularization, aligning the higher order tensors from the EquiformerV2 score model with pretrained MACE embeddings consistently improves the performance. Using different score model layers for alignment does not have a significant impact as shown in Figure S3. Therefore we used the second layer. We note that, this behavior might differ when facing larger systems necessitating deeper models.

From our initial observations with constant λ, REPA was improving early-to-mid epoch convergence but struggled to converge to the optimal JSD value. Therefore, we used a cosine schedule for λ as mentioned above. As another way of loosening the constraint from the regularization, we also tried aligning only the $3/4$ of score model representation dimensions and freed the other $1/4$ from REPA. Results can be seen Figure S4 along the ablation study of λ for $\tau = 0.5$, which show a significant improvement achieved by partial alignment this approach. A τ ablation study (see Figure S5) for the $3/4$ partially aligned model reveal that the method is robust to this hyperparameter. As discussed in section 3, we introduced a time-conditioned gate to learn a mask over the intermediate latents that are aligned to the pretrained MLIP embeddings. We provide a performance comparison with our best performing methods to demonstrate its effectiveness

Figure S6. We also provide a λ ablation study for the partial gated alignment showing more robustness compared to the standard approach.

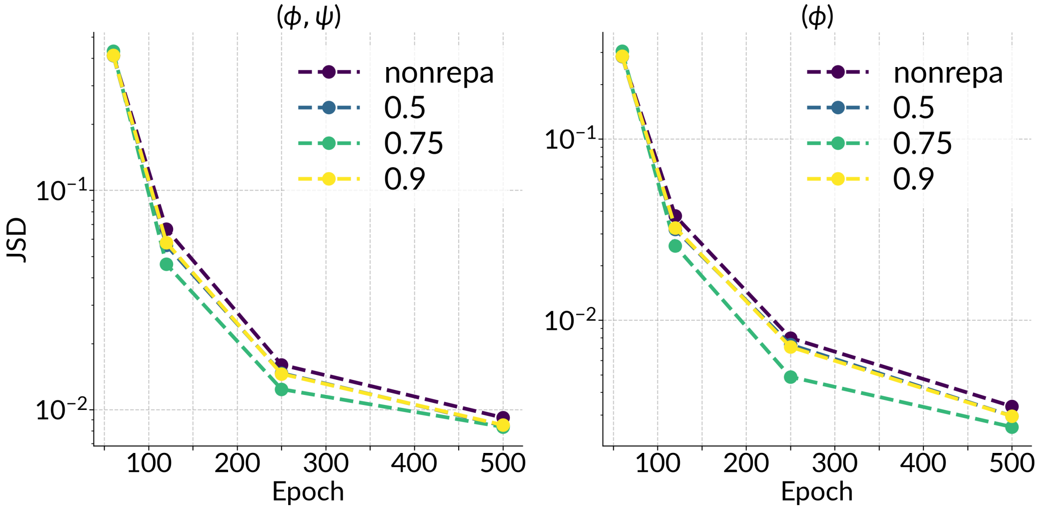


Figure S1. JSD of the sampled distributions against reference for different τ values (left: (ϕ, ψ), right: ϕ).

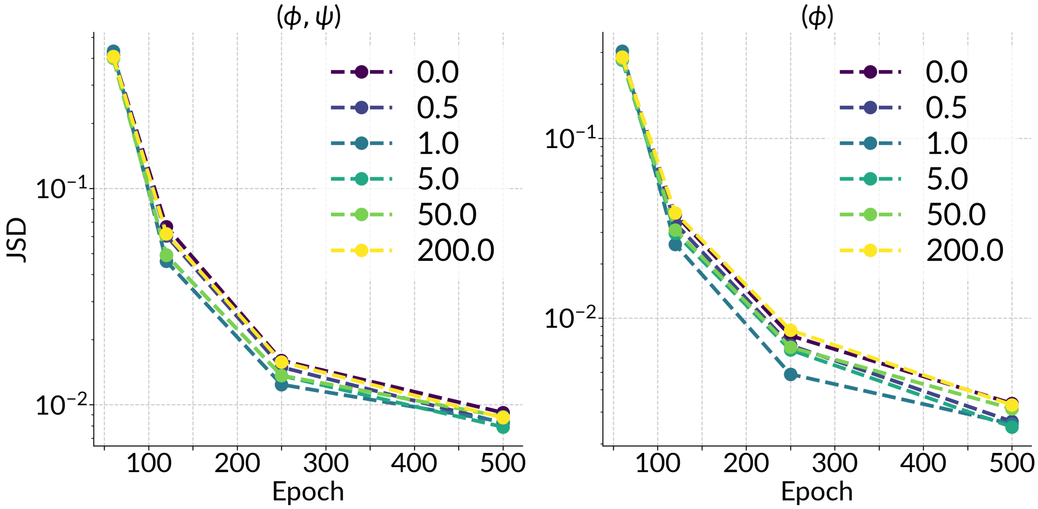


Figure S2. JSD of the sampled distributions against reference for different λ values (left: (ϕ, ψ), right: ϕ). τ is fixed at 0.75.

C. Ramachandran plots

We show Ramachandran plots at different epochs during training in Figure S10 without regularization, and in Figure S11 with REPA. The improvement in convergence and final performance is clearly visible especially at Epoch 120, with smoother surfaces and transition regions. To visualize the sampling quality, we plot potential of mean force with ϕ angle as the collective variable S9a.

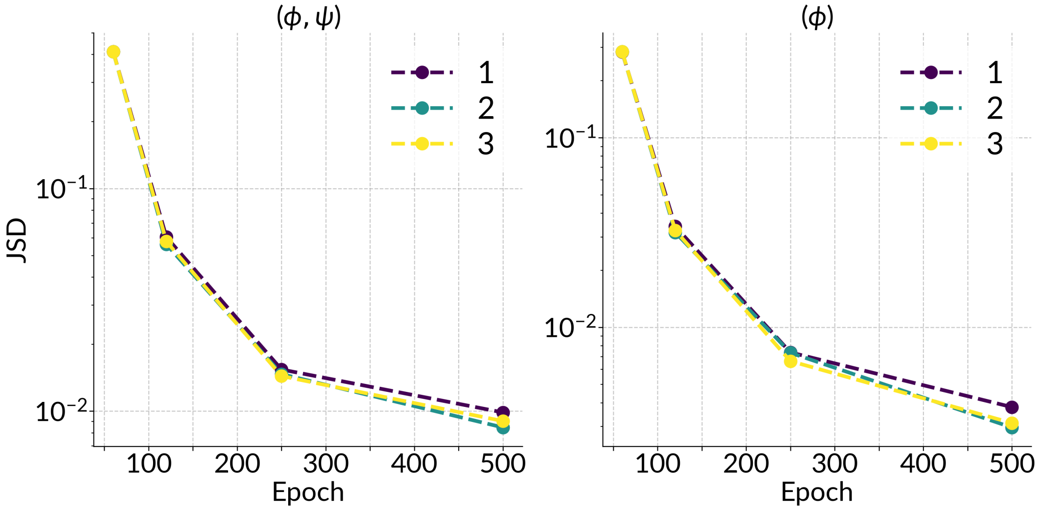


Figure S3. JSD of the sampled distributions against reference for different EquiformerV2 layers being aligned (left: (ϕ, ψ), right: ϕ). τ is fixed at 0.5.

D. Methods

D.1. Denoising Diffusion Probabilistic Models

We use the formulation from Denoising Diffusion Probabilistic Models (Ho et al., 2020) (DDPM) that leverages a discrete Markov chain $p(x_i|x_{i-1}) = \mathcal{N}(x_i; \sqrt{1-\beta_i}x_{i-1}, \beta_i\mathbf{I})$ and the closed-form marginalization to add noise to the data, yielding:

$$x_t = \sqrt{\bar{\alpha}_t}x_0 + \sqrt{1-\bar{\alpha}_t}\epsilon, \quad \epsilon \sim \mathcal{N}(0, I). \tag{8}$$

where $\alpha_t = 1 - \beta_t$, $\bar{\alpha}_t = \prod_{s=1}^{t}\alpha_s$, and $i \in \{1, ..., T\}$ specifies the timestep. The reverse diffusion process can then be sampled by performing, from the Gaussian prior, successive steps of the reverse chain $p_\theta(x_{i-1}|x_i) = \mathcal{N}(x_{i-1}; \frac{1}{\sqrt{\alpha_i}}\left(\mathbf{z}_i - \frac{\beta_i}{\sqrt{1-\bar{\alpha}_i}}\epsilon_\theta(\mathbf{z}_i, i)\right), \beta_i\mathbf{I})$. We optimize for the conditional denoising score matching objective

$$\mathcal{L}_{DDPM} = \mathbb{E}[||\epsilon - \epsilon_\theta(x_t, t)||^2] \tag{9}$$

where x_t follows equation 8. We use a cosine variance schedule β_i following the observations of iDDPM (Nichol & Dhariwal, 2021). At inference, we sample the initial atomic positions from the Gaussian distribution. We then take advantage of the ODE formulation

$$\frac{d\mathbf{x}_t}{dt} = f(t)\mathbf{x}_t - \frac{1}{2}g^2(t)\nabla_\mathbf{x}\log p_t(\mathbf{x}_t), \quad \mathbf{x}_T \sim p_T(\mathbf{x}_T), \tag{10}$$

as well as higher order ODE solver to obtain good quality generation with reasonable sampling time. We use DPM-Solver-3 (Lu et al., 2022) to perform a third order solve.

D.2. Derivation of the force-score relationship

At $t = 0$. For a Boltzmann target $q(\mathbf{y}) \propto e^{-\beta V(\mathbf{y})}$ with force $\mathbf{F} = -\nabla_\mathbf{y}V(\mathbf{y})$, the zero-noise score is

$$s^*_\theta(\mathbf{x}, 0) = \nabla_\mathbf{x}\log q(\mathbf{x}) = \beta\,\mathbf{F}(\mathbf{x}).$$

At finite $t > 0$. With Gaussian noise perturbation $p_t(\mathbf{x}) = \int q(\mathbf{y})\,\mathcal{N}(\alpha_t\mathbf{x}; \mathbf{y}, \sigma_t^2\mathbf{I})\,d\mathbf{y}$,

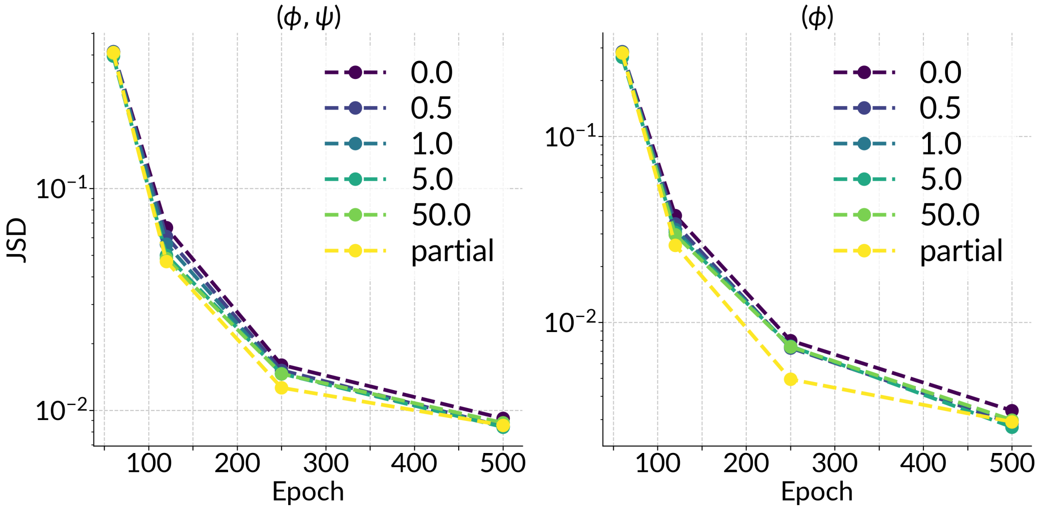


Figure S4. JSD of the sampled distributions against reference for different λ values (left: (ϕ, ψ), right: ϕ). The 'partial' condition corresponds to aligning only $3/4$ of score model representation dimensions. τ is fixed at 0.5

$$\begin{aligned}
\nabla_{\mathbf{x}} p_t(\mathbf{x}) &= \int q(\mathbf{y})\, \nabla_{\mathbf{x}} \mathcal{N}(\mathbf{x}; \alpha_t \mathbf{y}, \sigma_t^2 \mathbf{I})\, d\mathbf{y} \\
&= -\frac{1}{\alpha_t} \int q(\mathbf{y})\, \nabla_{\mathbf{y}} \mathcal{N}(\mathbf{x}; \alpha_t \mathbf{y}, \sigma_t^2 \mathbf{I})\, d\mathbf{y} \quad \text{(since } \nabla_{\mathbf{x}} = -(1/\alpha_t)\nabla_{\mathbf{y}} \text{ for the Gaussian)} \\
&= -\frac{1}{\alpha_t} \Big[q(\mathbf{y})\, \mathcal{N}(\mathbf{x}; \alpha_t \mathbf{y}, \sigma_t^2 \mathbf{I}) \Big]_{\mathbf{y}=-\infty}^{\mathbf{y}=\infty} + \frac{1}{\alpha_t} \int \mathcal{N}(\mathbf{x}; \alpha_t \mathbf{y}, \sigma_t^2 \mathbf{I})\, \nabla_{\mathbf{y}} q(\mathbf{y})\, d\mathbf{y} \quad \text{(integration by parts)} \\
&= \frac{1}{\alpha_t} \int \mathcal{N}(\mathbf{x}; \alpha_t \mathbf{y}, \sigma_t^2 \mathbf{I})\, \nabla_{\mathbf{y}} q(\mathbf{y})\, d\mathbf{y},
\end{aligned}$$

Because $\nabla_{\mathbf{y}} q = \beta q\, \mathbf{F}(\mathbf{y})$,

$$\nabla_{\mathbf{x}} p_t(\mathbf{x}) = \frac{\beta}{\alpha_t} \int q(\mathbf{y})\, \mathcal{N}(\mathbf{x}; \alpha_t \mathbf{y}, \sigma_t^2 \mathbf{I})\, \mathbf{F}(\mathbf{y})\, d\mathbf{y}.$$

Divide by p_t to obtain the optimal score:

$$s_\theta^*(\mathbf{x}, t) \propto \mathbb{E}_{\mathbf{y} \sim q(\mathbf{y}|\mathbf{x}, t)}[\mathbf{F}(\mathbf{y})]$$

, where $q(\mathbf{y}|\mathbf{x}, t) = q(\mathbf{y})\mathcal{N}(\mathbf{x}; \alpha_t \mathbf{y}, \sigma_t^2 \mathbf{I})/p_t(\mathbf{x})$, and the expectation is taken over clean conformers $\mathbf{y}$ that could have produced the noisy sample $\mathbf{x}$.

D.3. Alignment metric

The Centered Kernel Alignment (CKA) (Kornblith et al., 2019) measures global similarity from all data pairs by computing a normalized Hilbert-Schmidt Independent Criterion estimator:

$$\mathrm{CKA}(\mathbf{K}, \mathbf{L}) = \frac{\mathrm{HSIC}(\mathbf{K}, \mathbf{L})}{\sqrt{\mathrm{HSIC}(\mathbf{K}, \mathbf{K})\, \mathrm{HSIC}(\mathbf{L}, \mathbf{L})}}, \tag{11}$$

Following the original notation from (Huh et al., 2024), with ϕ_i, ϕ_j the intermediate representations extracted from one network and ψ_i, ψ_j the ones extracted from the other model, and with the kernel matrices $\mathbf{K}_{i,j} = \mathcal{K}(\phi_i, \phi_j)$, and

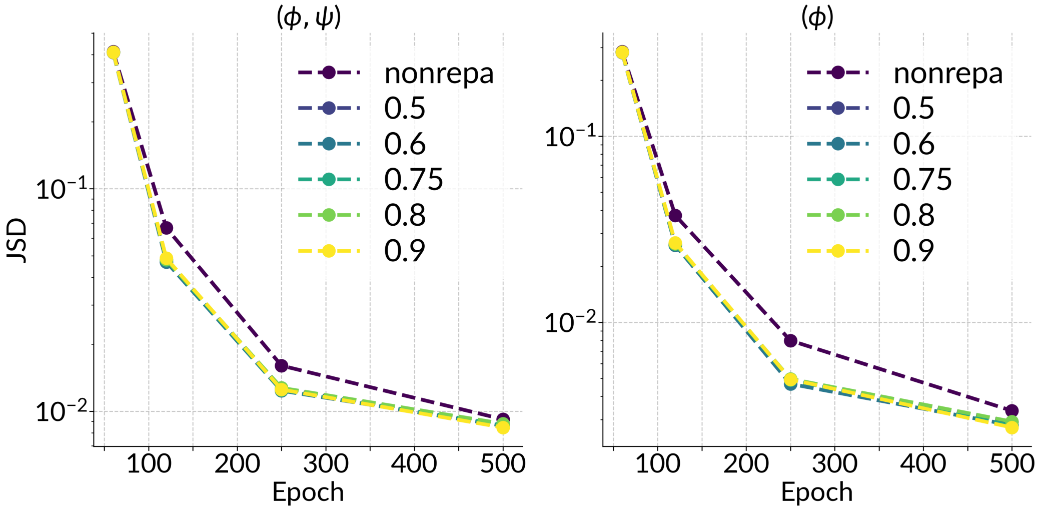


Figure S5. JSD of the sampled distributions against reference for different τ values for partial alignment (left: (ϕ, ψ), right: ϕ). 3/4 of the channels are sliced for alignment

$\mathbf{K}_{i,j} = \mathcal{K}(\psi_i, \psi_j)$:

$$\mathrm{HSIC}(\mathbf{K}, \mathbf{L}) = \frac{1}{(n-1)^2} \left(\sum_i \sum_j \left(\langle \phi_i, \phi_j \rangle - \mathbb{E}_l[\langle \phi_i, \phi_l \rangle] \right) \left(\langle \psi_i, \psi_j \rangle - \mathbb{E}_l[\langle \psi_i, \psi_l \rangle] \right) \right) \tag{12}$$

More specifically, as Murphy et al. (2024) pointed out, the standard biased estimator of CKA is highly dependent on the number of points taken to compute the metric. Therefore, we use the more reliable unbiased estimator from Song et al. (2007) to compute HSIC.

E. Additional discussion

E.1. The discrepancy between the pretrained MLIP being aligned and the force field used to sample the training data

Another notable point is the discrepancy between the force field used to generate the training data and the one used for alignment. In our experiments, the reference conformational ensemble was generated via MD simulations using the classical AmberFF99SB force field, while the alignment was performed using representations from the pretrained MACE model. This mismatch may have provided less ideal results, particularly in high-energy regions where different force fields tend to diverge. In future work, we will systematically investigate how such discrepancies affect the quality of alignment and overall model performance.

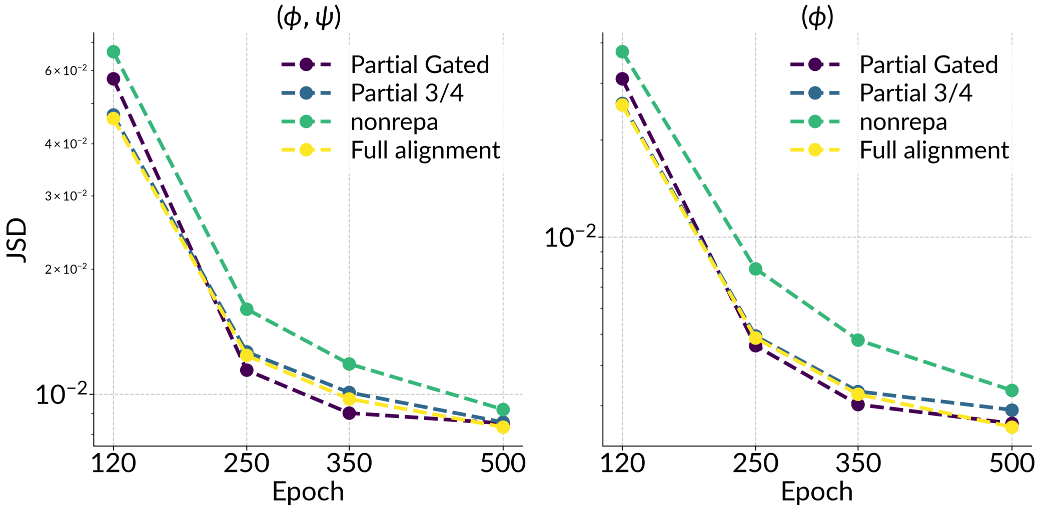


Figure S6. JSD of the sampled distributions against reference for the best performing methods against the partial gated alignment method (left: (ϕ, ψ), right: ϕ). The full alignment is with $\tau = 0.75$. All results are obtained with $\lambda = 1$ except for the non aligned one.

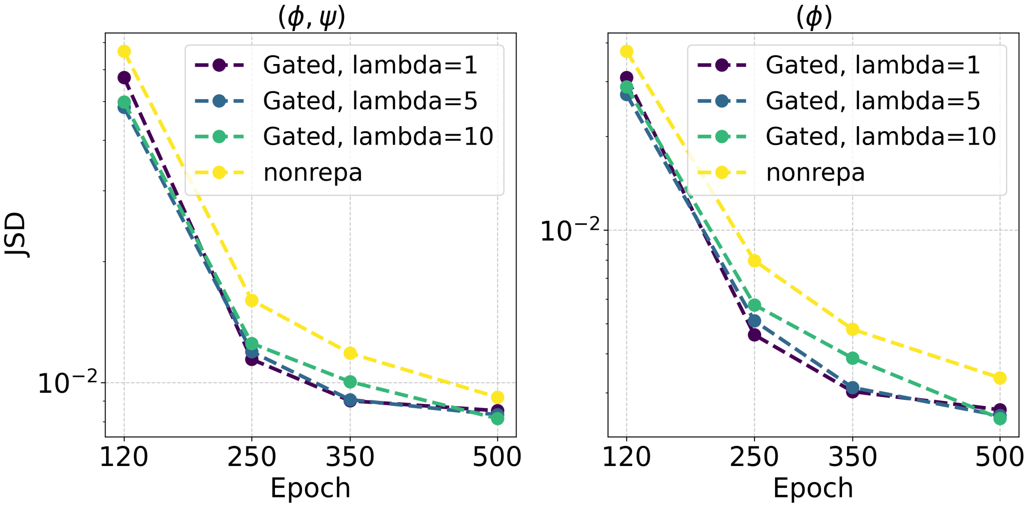


Figure S7. JSD of the sampled distributions against reference for different λ values for partial gated alignment (left: (ϕ, ψ), right: ϕ)

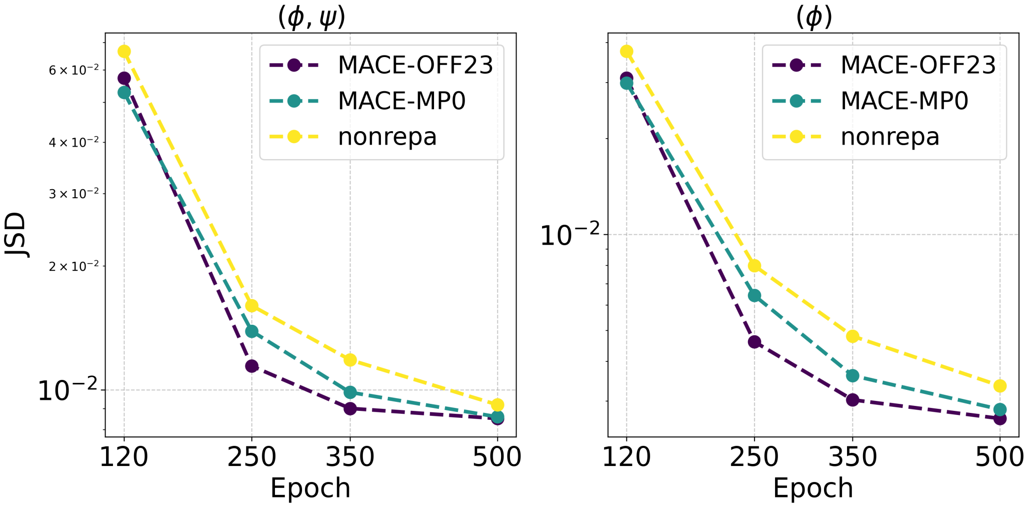
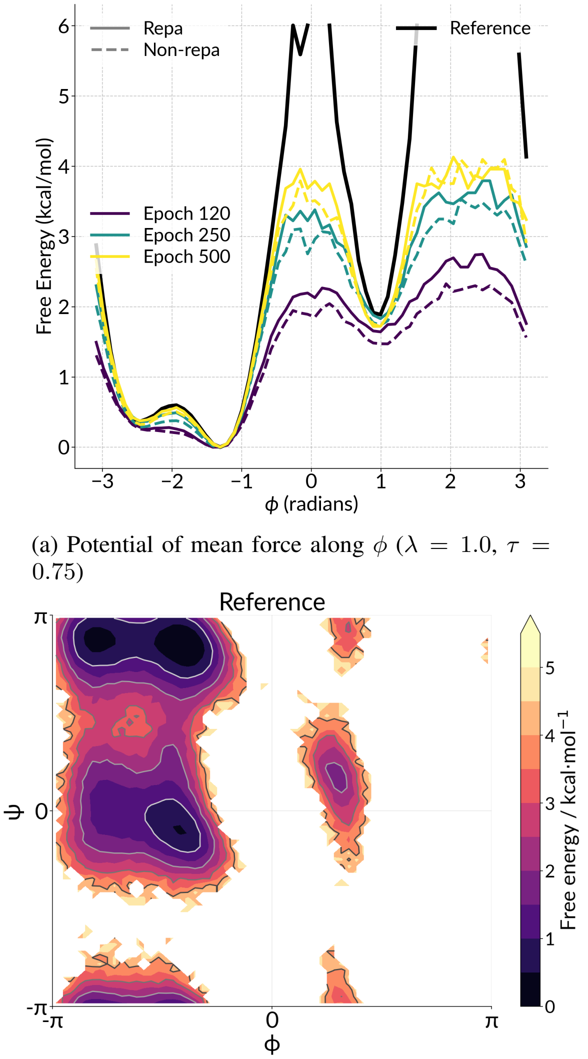


Figure S8. JSD of the sampled distributions against reference using **MACE-OFF23** or **MACE-MP0** as the pretrained MLIP (left: (ϕ, ψ), right: ϕ). The REPA results are obtained with the time-conditioned gate and $\lambda = 1$.



(a) Potential of mean force along ϕ ($\lambda = 1.0$, $\tau = 0.75$)

(b) Reference Ramachandran plot

Figure S9. (a) PMF along ϕ and (b) reference Ramachandran plot

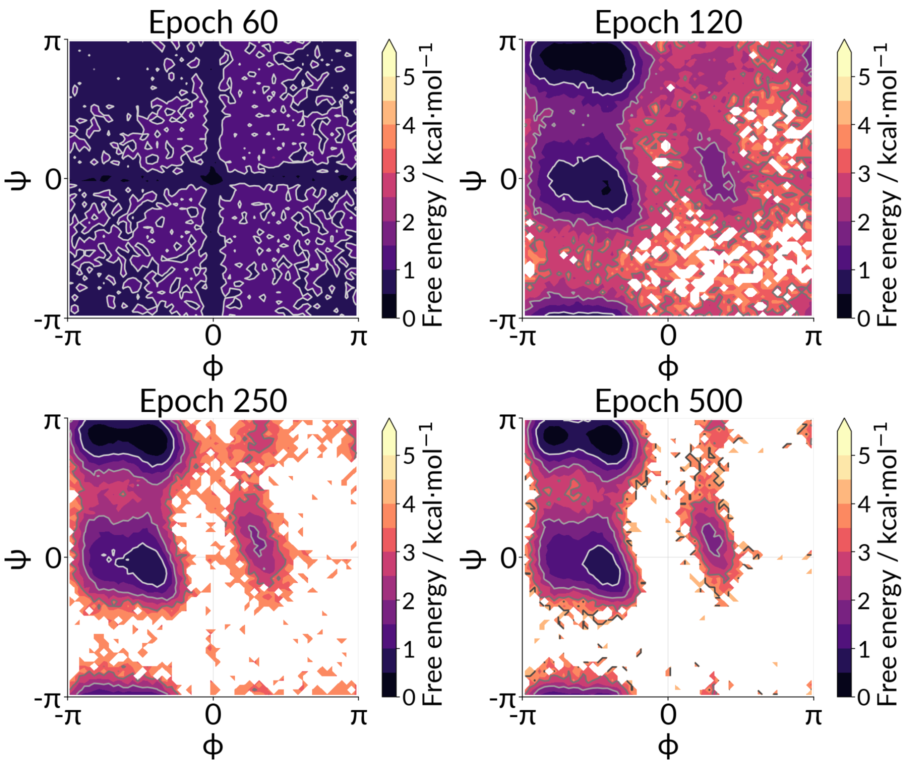


Figure S10. Ramachandran plots of 100k samples generated from the model trained without representation alignment ($\lambda = 1.0$, $\tau = 0.75$).

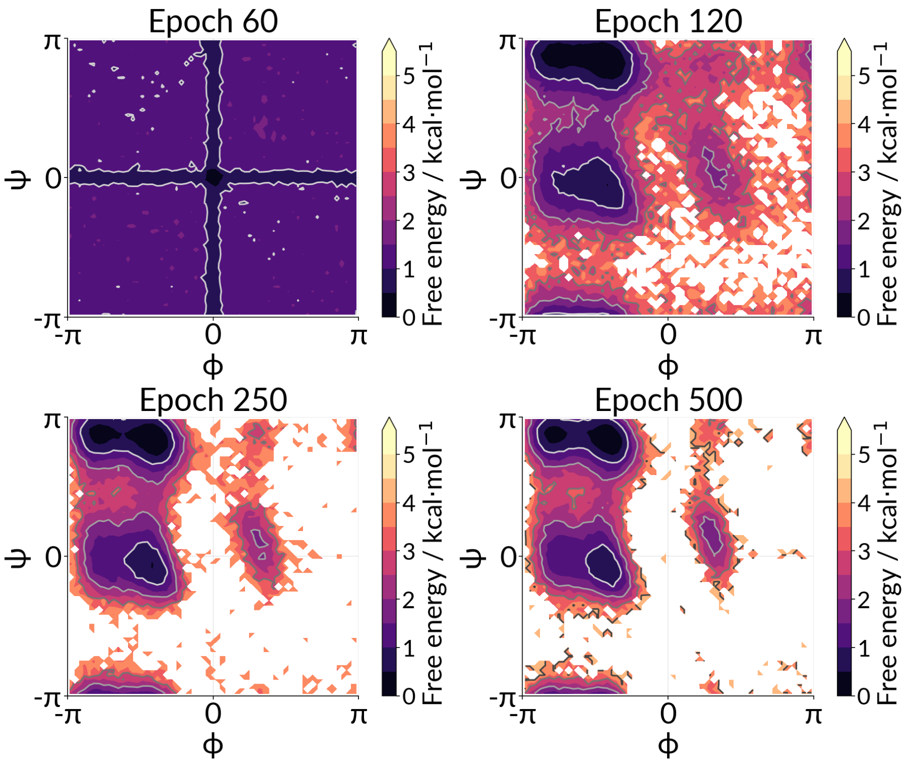


Figure S11. Ramachandran plots of 100k samples generated from the model trained with representation alignment ($\lambda = 1.0$, $\tau = 0.75$).