

---

# Active Learning for Machine Learning Driven Molecular Dynamics

---

**Kevin Bachelor**

University of California, Santa Cruz  
kwbachel@ucsc.edu

**Sanya Murdeshwar**

University of California, Santa Cruz  
smurdesh@ucsc.edu

**Daniel Sabo**

University of California, Santa Cruz  
dsabo@ucsc.edu

**Ashwin Lokapally**

GiwoTech Inc.  
ashwin@giwotech.com

**Razvan Marinescu**

University of California, Santa Cruz  
ramarine@ucsc.edu

## Abstract

Machine-learned coarse-grained (CG) potentials are fast, but degrade over time when simulations reach under-sampled bio-molecular conformations, and generating widespread all-atom (AA) data to combat this is computationally infeasible. We propose a novel active learning (AL) framework for CG neural network potentials in molecular dynamics (MD). Building on the CGSchNet model [17], our method employs root mean squared deviation (RMSD)-based frame selection from MD simulations in order to generate data on-the-fly by querying an oracle during the training of a neural network potential. This framework preserves CG-level efficiency while correcting the model at precise, RMSD-identified coverage gaps. By training CGSchNet—a coarse-grained neural network potential—we empirically show that our framework explores previously unseen configurations and trains the model on unexplored regions of conformational space. Our active learning framework enables a CGSchNet model trained on the Chignolin protein to achieve a 33.05% improvement in the Wasserstein-1 (W1) metric in Time-lagged Independent Component Analysis (TICA) [15] space on an in-house benchmark suite.

## 1 Introduction

Advancements in the realm of MD have allowed for the effective exploration of complex conformational landscapes of biomolecular structures such as proteins, but such methods remain limited by the immense computational overhead associated with system size and the granularity of the atomic representation [9]. Multiple generations of AA neural network potentials have progressed high-resolution MD by taking atomic coordinates,  $\mathbf{r} = (\mathbf{r}_1, \dots, \mathbf{r}_N) \in \mathbb{R}^{3N}$ , and predicting the energy  $E = U_\theta(\mathbf{r})$  with forces  $\mathbf{F}(\mathbf{r}) = -\nabla_{\mathbf{r}}U_\theta(\mathbf{r})$ , surpassing classical energy functionals in terms of accuracy [3, 6, 18, 19, 2, 4, 5]. However, such high-dimensional models remain slow at AA resolution and are often more computationally expensive than computing classical force-field terms. [9] Consequently, simulating large structures takes a prohibitive amount of time to reach critical and relevant regions, thereby making speedup and efficiency key goals in MD development.

A renowned approach for achieving longer timescales with neural network potentials in MD is to reduce the size of the conformational space through coarse-graining, in which groups of atoms are represented as single beads [13]. Generally, such models have been trained via force matching, which fits the CG potential to match reference AA forces on mapped configurations to minimize the difference between CG forces and mapped AA forces [10]. However, because most training data samples single metastable states in proteins, they struggle to generalize to unseen states and fully explore the conformational space, particularly during transitions between metastable states, making simulations prone to conformational "explosion" or "implosion" anomalies. These divergences result from the network generating physically inconsistent forces upon encountering configurations significantly different from those in the initial training data. This issue is especially prevalent for large proteins with complex folding structures.

To address this, we propose an active learning framework that uses targeted, on-the-fly data acquisition to obtain data beyond the regions normally sampled by MD by selecting specific out-of-distribution frames during simulation and querying an AA ground-truth oracle, thereby patching coverage gaps at minimal cost. For data acquisition, we identify frames most different from the training set quantified by distance-based proxies such as RMSD, enabling data generation in the least-explored regions of the protein conformational space. We evaluate our work using a recently-published benchmark for AI-driven dynamics [1], which uses a free-energy matching loss in TICA space. Our active learning framework achieves an improvement of 33.05% in TICA space compared to a standard training setup without AL. In addition, we find the AL framework results in significant qualitative exploration of the conformational space as visualized in the TICA plots.

### 1.1 Related Works

Previous approaches to improve data efficiency in MD simulations have involved both active learning strategies as well as CG neural network potentials, but these avenues have largely been explored separately. On the active learning front, the Deep Potential Generator (DP-GEN) serves as an early active learning framework for AA neural network potentials for materials simulations [21]. More recently, a framework proposed by Jung et al. introduces another active learning strategy for AA neural network potentials, demonstrating the benefits of uncertainty-guided sampling in minimizing redundant data generation during molecular simulations [12]. These frameworks, however, were designed for AA systems and do not address the CG perspective, which inherently imposes additional limitations. More recently, Duschatko et al. [7] propose a Bayesian uncertainty-aware active learning framework that does address CG free energy models, in which CG configurations with high uncertainty trigger active learning, though this workflow does not clearly refer to a full all-atom oracle for every sampled configuration.

In parallel, coarse-graining has long been used to increase simulation timescales by reducing model resolution [10, 13, 14, 20]. Classical top-down CG methods later emerged such as the Martini force field, reproducing experimental observations through parametric model tuning [14]. In contrast, bottom-up CG approaches derive CG potentials from AA simulations to retain physical accuracy, usually by employing force matching [11]. In recent years, machine learning has been applied to bottom-up CG approaches with notable success. For instance, Wang et al. [20] demonstrates that a neural network (CGNet) is able to learn a CG free energy surface via force matching, though their framework requires manual feature selection for molecules. With this foundation, developments such as CGSchNet [10], building upon the SchNet model [17], have significantly advanced coarse-graining via a graph neural network architecture, achieving strong thermodynamical accuracy. However, such models are trained on fixed datasets and therefore lack a mechanism for adaptive data acquisition, leading to problematic behavior when simulations enter under-sampled regions of conformational space.

## 2 Methodology

In Fig. 1 we show an overview of our active learning methodology. At each active learning iteration, the model is trained on MD trajectory data and then used to simulate the CG protein system. We then select the frames with the largest RMSD discrepancies, indicating the greatest difference between the configuration and the training dataset, ultimately targeting the least-covered configurations. These frames are then backmapped into AA space, used to query the oracle by seeding OpenMM [8]

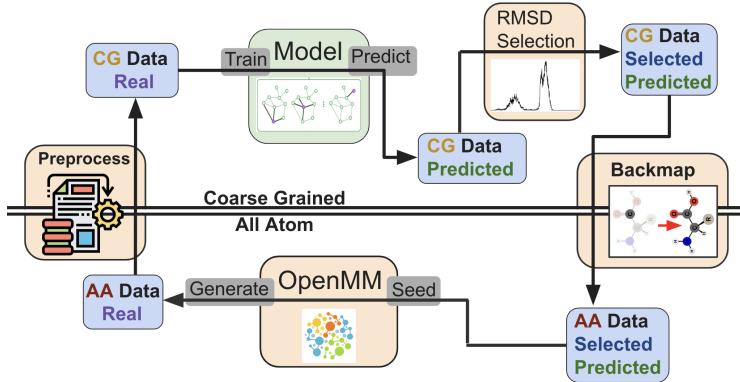


Figure 1: Our active learning pipeline showing the model-training  $\Leftrightarrow$  data-generation loop. Each round trains a coarse-grained (CG) model, runs a CG simulation, selects least-covered frames via RMSD distance proxy, backmaps them to all-atom (AA) space, runs short OpenMM simulations, and projects AA  $\rightarrow$  CG to augment the dataset and retrain. By querying the oracle only where coverage is poor, the loop increases conformational coverage at minimal AA simulation cost.

all-atom simulations, projected back into CG space, and appended to the dataset. The model is retrained, and the loop repeats, querying the oracle only where coverage is low to minimize usage of AA simulations.

## 2.1 Neural network potential (CGSchNet)

Our neural network leverages a graph neural network (GNN) architecture based on CGSchNet [17], using continuous filter convolutions to build features based on inter-bead distances. Let us define the CG bead coordinates as  $\mathbf{R} = (\mathbf{R}_1, \dots, \mathbf{R}_M) \in \mathbb{R}^{3M}$ , and the inter-bead distance as  $r_{ij} = \|\mathbf{R}_i - \mathbf{R}_j\|$ . Based on these distances, the network outputs a scalar energy potential  $\mathbf{U}_\theta(\mathbf{R})$  which accounts for the system's geometry rather than just the absolute position or orientation. This architecture thereby ensures that  $\mathbf{U}_\theta$  remains invariant to global translations and rotations, and the resultant forces  $\mathbf{F}_\theta(\mathbf{R}) = -\nabla_{\mathbf{R}} \mathbf{U}_\theta(\mathbf{R})$  are translation-invariant and rotation-equivariant as well. The objective of force matching is done by minimizing the mean-squared error between the predicted forces and the target CG forces projected from AA space. The force matching loss function  $\mathcal{L}_{\mathcal{F}\mathcal{M}}(\theta)$  is given as:

$$\mathcal{L}_{\mathcal{F}\mathcal{M}}(\theta) = \frac{1}{T} \sum_{t=1}^T \frac{1}{3M^{(t)}} \left\| \mathbf{F}_\theta(\mathbf{R}^{(t)}) - \mathbf{F}_{\text{CG}}^{(t)} \right\|_F^2$$

where  $\mathbf{R}^{(t)} \in \mathbb{R}^{3M^{(t)}}$  represents the CG bead coordinates for  $M^{(t)}$  beads in frame  $t$ ,  $F_{\text{CG}}^{(t)}$  represents the AA  $\rightarrow$  CG projected CG forces and  $T$  tells us the number of training frames.

## 2.2 AA $\rightarrow$ CG and CG $\rightarrow$ AA projection

Our neural network potential is compatible with CG data, while the oracle simulator works only with AA data. Thus, we require a bidirectional bridge between the spaces. We convert AA to CG by mapping atomic coordinates to carbon-alpha ( $C\alpha$ ) beads via a pre-processing script using a linear operator

$$\Xi : \mathbb{R}^{3N} \rightarrow \mathbb{R}^{3M}, \quad \mathbf{R} = \Xi \mathbf{r}.$$

We map the forces by projecting AA forces onto CG degrees of freedom via force projection operators [17, 10]:

$$\begin{aligned} \mathbf{F}_{\text{CG}} &= \Xi_F \mathbf{f}_{\text{AA}} \\ \Xi_F &= (\Xi \Xi^\top)^{-1} \Xi \end{aligned}$$

To convert from CG to AA, we utilize the PULCHRA [16] backmappier to reconstruct the missing, non- $C\alpha$  atoms into statistically likely positions.

### 2.3 Frame selection

For frame selection, we use the RMSD distance proxy for concrete visualization of coverage gaps. To select frames, during each active learning iteration we compute histograms of the RMSD values between simulated frames and the training set. We then select the frames with the largest simulation-minus-training RMSD discrepancies as candidates, signifying abnormal gaps in coverage. We then filter further by removing any frames with RMSD value outside of a cutoff to avoid getting frames that are part of implosions or explosions. These frames are then backmapped to AA space in order to seed the OpenMM [8] oracle which generates robust data that is then projected back into CG space to begin retraining.

For readers unfamiliar with high-dimensional stochastic data, it may seem that a distance-based acquisition function would bias selection toward later parts of the trajectory, since the simulation begins in-distribution and would generally drift outwards. In practice, this does not occur. Trajectories typically spend long intervals fluctuating around local minima before occasionally jumping to new basins, where they again remain for many steps. A newly-reached RMSD region may be previously explored or be entirely new. As a result, variance in the projected metric does not grow linearly with time, and RMSD-based acquisition does not inherently favor later frames.

## 3 Results

To test the efficacy of our active learning framework, we trained a base CGSchnet model on a small subset of data from the Chignolin protein, and used an in-house benchmark suite to measure the (i) conformation exploration along slow modes of motion via TICA projections [15], (ii) reaction coordinates representing the fraction of time certain atoms are within close proximity, and the distributions of (iii) bond lengths, (iv) bond angles, and (v) dihedral angles. Model-generated distributions were compared against extensive Chignolin ground-truth references, depicted in Figure A.1 for the base model, and in Figure A.2 for the model after applying the active learning loop. We summarize the outcomes quantitatively with the W1 distance metric between distributions, representing the minimal "work" required to transform the model distribution into the reference, where lower values are favorable. In Table 1 we detail the results of the CGSchnet[10] model before applying the active learning, the same model after our active learning have been applied, as well as another recent model that utilizes a metalearning framework for a Chignolin trained CGSchnet model as reference. The table details the W1 distance between the ground truth, and each model's simulations, for the distributions over the 5 measurements we took for Chignolin.

Table 1: Wasserstein-1 (W1) before and after active learning (lower is better).

Plot	W1↓		
	Before AL	After AL	Match Energy[1]
TICA KDE	1.15023	<b>0.77003</b>	1.17342
Reaction coordinate	<b>0.15141</b>	0.38302	0.65711
Bond length	0.00043	<b>0.00022</b>	0.00057
Bond angle	0.11036	<b>0.10148</b>	0.24846
Dihedral	0.25472	0.36378	<b>0.21482</b>

These benchmarks show significant improvements in exploration capability and local accuracy measures. In TICA space, the W1 distance to the reference dropped by 33.05%. Additionally, it produced more stable trajectories, with bond-length and bond-angle distributions aligning more closely with the ground truth with a 48.84% and 8.05% decrease in W1 distance, respectively.

The dihedral and RC metrics did not improve, which is expected: the AL dihedral benchmark graph – shown in A.2 – visually matches well, given that dihedral angles are inherently noisy. Furthermore, the RC metric captures the PDF of a single atom-pair distance, representing a highly localized, magnified view of the configuration space, and therefore is extremely sensitive to global changes. These localized deviations do not contradict the broader trend, as the other metrics that track global conformational structure all show strong improvement.

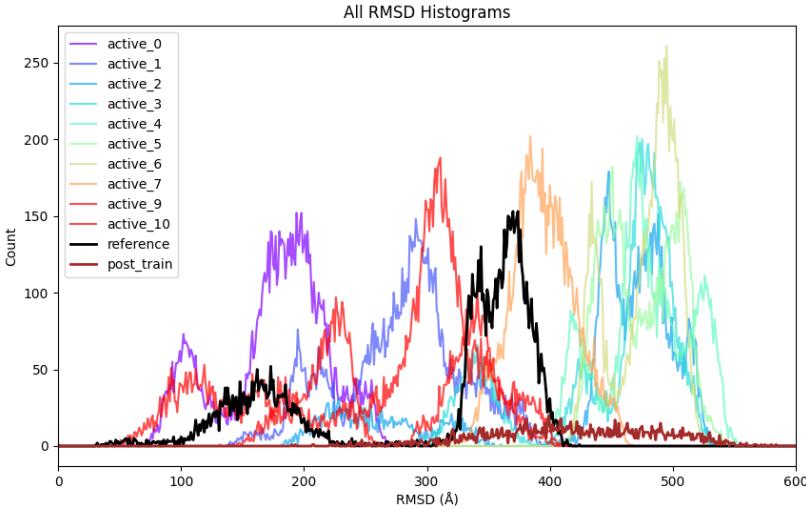


Figure 2: Base model output (black), model output after active learning (brown), and training data per active learning iteration (purple → red), plotted as a histogram of RMSD values from the reference frame.

Improvements in exploration are also evident in the RMSD histograms before and after active learning (Figure 2). The base model (black) shows a bimodal RMSD density distribution, indicating sampling concentrated in two conformational states. The additional histograms represent the active learning iterations (purple → red), reflecting the data that was used to train the model at each iteration. The substantial variance across iterations indicates that each round, the model targets different under-sampled regions, with the active learning loop supplying training data to support it across diverse conformational states. The post-active learning model (brown) shows a much broader RMSD distribution, further illustrating improvements in phase-space coverage aligning with the TICA W1 reductions reported above.

## 4 Discussion

Since we query OpenMM only when the RMSD histograms indicate poorly covered regions, we retain the speed and efficiency of CG simulations while correcting the model in those under-sampled states on-the-fly, restoring accuracy and thoroughness in coverage. Evaluated on Chignolin’s complex structure, pure CG simulation without using our active learning framework often resulted in explosions or implosions due to limited coverage and indicative of poor generalization over the conformational space of the protein. Empirically, we found that our active learning method explored more of the conformational space and improved accuracy across the bond, bond-angle, dihedral, and TICA space metrics in the standardized benchmark. This indicates that (*i*) long CG trajectories are stabilized and explosion and implosion anomalies are avoided, and (*ii*) more of the conformation space is explored, which is crucial for the ultimate goal of generalization.

These results demonstrate that targeted active learning in MD stabilizes CG simulations and expands conformational coverage within reasonable compute, while also providing paths for future improvements upon mechanisms of our framework that impose limitations on its efficacy. Stronger backmapping methodologies would allow for increased accuracy in AA reconstructions while reducing relaxation cost. Furthermore, more rigorous distance proxies or measures may further optimize how frames are prioritized and, in turn, reduce excessive oracle usage. Ultimately, our active learning framework’s targeted, precise, oracle-backed patching of phase-space gaps can significantly expand the training data for ML models in molecular dynamics at a practical cost, and provides clear paths for improvements in selection, backmapping, and representation bridging.

## Impact statement

Accurate yet affordable exploration of protein conformational spaces is a critical bottleneck in ML-driven drug discovery, where rare conformational states and transitions govern crucial protein mechanisms such as binding ability. Our active learning framework unifies high-speed CG neural network potentials with precise all-atom oracle queries to explore conformational spaces more rigorously while keeping the computational cost of labeling low. Our result enables increased capability in viewing protein behavior, which may play a role in revealing unique binding opportunities or promising compounds far sooner, minimizing the time spent in extensive trial-and-error in drug discovery laboratories. Because our framework is model-agnostic and bridges CG $\leftrightarrow$ AA reliably, it can be used to augment current or future state-of-the-art ML force fields easily rather than replacing them altogether, resulting in confident observations and shorter paths from trial-and-error to effective treatments.

## References

- [1] Alexander Aghili, Andy Bruce, Daniel Sabo, and Razvan Marinescu. Tica-based free energy matching for machine-learned molecular dynamics, 2025.
- [2] Albert P. Bartók, Mike C. Payne, Risi Kondor, and Gábor Csányi. Gaussian approximation potentials: The accuracy of quantum mechanics, without the electrons. *Physical Review Letters*, 104(13):136403, April 2010.
- [3] Jörg Behler. Four generations of high-dimensional neural network potentials. *Chemical Reviews*, 121(16), 2021.
- [4] Jörg Behler. Neural network potential-energy surfaces in chemistry: a tool for large-scale simulations. *Physical Chemistry Chemical Physics*, 13(40):17930–17955, October 2011.
- [5] Jörg Behler and Michele Parrinello. Generalized neural-network representation of high-dimensional potential-energy surfaces. *Physical Review Letters*, 98(14):146401, April 2007.
- [6] Timothy T. Duignan. The potential of neural network potentials. *ACS Physical Chemistry Au*, 4(3):232–241, 2024.
- [7] Blake R. Duschatko, Jonathan Vandermause, Nicola Molinari, and Boris Kozinsky. Uncertainty driven active learning of coarse grained free energy models. *npj Computational Materials*, 10(1):9, January 2024.
- [8] Peter Eastman, Raimondas Galvelis, Raúl P. Peláez, Charles R. A. Abreu, Sarah E. Farr, Emilio Gallicchio, Anton Gorenko, Matthew M. Henry, Feng Hu, Jing Huang, Andreas Krämer, Julien Michel, Joshua A. Mitchell, Vijay S. Pande, João P. G. L. M. Rodrigues, Jaime Rodriguez-Guerra, Andrew C. Simmonett, Sukrit Singh, Jason Swails, Peter Turner, Yuanqing Wang, Ivy Zhang, John D. Chodera, Gianni De Fabritiis, and Thomas E. Markland. Openmm 8: Molecular dynamics simulation with machine learning potentials. *J. Phys. Chem. B*, 128(1):109–116, 2023.
- [9] Scott A. Hollingsworth and Ron O. Dror. Molecular dynamics simulation for all. *Neuron*, 99(6):1129–1143, 2018.
- [10] Brooke E. Husic, Nicholas E. Charron, Dominik Lemm, Jiang Wang, Adrià Pérez, Maciej Majewski, Andreas Krämer, Yaoyi Chen, Simon Olsson, Gianni de Fabritiis, Frank Noé, and Cecilia Clementi. Coarse graining molecular dynamics with graph neural networks. *The Journal of Chemical Physics*, 153(19), November 2020.
- [11] Jaehyeok Jin, Alexander J. Pak, Aleksander E. P. Durumeric, Timothy D. Loose, and Gregory A. Voth. Bottom-up coarse-graining: Principles and perspectives. *Journal of Chemical Theory and Computation*, 18(10):5759–5791, October 2022.
- [12] Gang Seob Jung, Jong Youl Choi, and Sangkeun Matthew Lee. Active learning of neural network potentials for rare events. *Digital Discovery*, 3:514–527, 2024.

- [13] Sebastian Kmiecik, Dominik Gront, Michal Kolinski, Lukasz Witeska, Aleksandra Elzbieta Dawid, and Andrzej Kolinski. Coarse-grained protein models and their applications. *Chemical Reviews*, 116(14):7898–7936, 2016.
- [14] Siewert J. Marrink, H. Jelger Risselada, Serge Yefimov, D. Peter Tieleman, and Alex H. de Vries. The martini force field: coarse grained model for biomolecular simulations. *The Journal of Physical Chemistry B*, 111(27):7812–7824, 2007.
- [15] Frank Noé and Cecilia Clementi. Kinetic distance and kinetic maps from molecular dynamics simulation. *Journal of Chemical Theory and Computation*, 11(10):5002–5011, 2015. PMID: 26574285.
- [16] Piotr Rotkiewicz and Jeffrey Skolnick. Fast procedure for reconstruction of full-atom protein models from reduced representations. *Journal of Computational Chemistry*, 29(9):1460–1465, July 2008.
- [17] K. T. Schütt, P.-J. Kindermans, H. E. Sauceda, S. Chmiela, A. Tkatchenko, and K.-R. Müller. Schnet: a continuous-filter convolutional neural network for modeling quantum interactions. In *Proceedings of the 31st International Conference on Neural Information Processing Systems, NIPS’17*, page 992–1002, Red Hook, NY, USA, 2017. Curran Associates Inc.
- [18] Bobby G. Sumpter and Donald W. Noid. Potential energy surfaces for macromolecules. a neural network technique. *Chemical Physics Letters*, 192(5):455–462, May 1992.
- [19] Oliver T. Unke and Markus Meuwly. Physnet: A neural network for predicting energies, forces, dipole moments, and partial charges. *Journal of Chemical Theory and Computation*, 15(6):3678–3693, 2019.
- [20] Jiang Wang, Simon Olsson, Christoph Wehmeyer, Adrià Pérez, Nicholas E. Charron, Gianni de Fabritiis, Frank Noé, and Cecilia Clementi. Machine learning of coarse-grained molecular dynamics force fields. *ACS Central Science*, 5(5):755–767, May 2019.
- [21] Linfeng Zhang, De-Ye Lin, Han Wang, Roberto Car, and Weinan E. Active learning of uniformly accurate interatomic potentials for materials simulation. *Phys. Rev. Mater.*, 3:023804, Feb 2019.

## A Technical Appendices and Supplementary Material

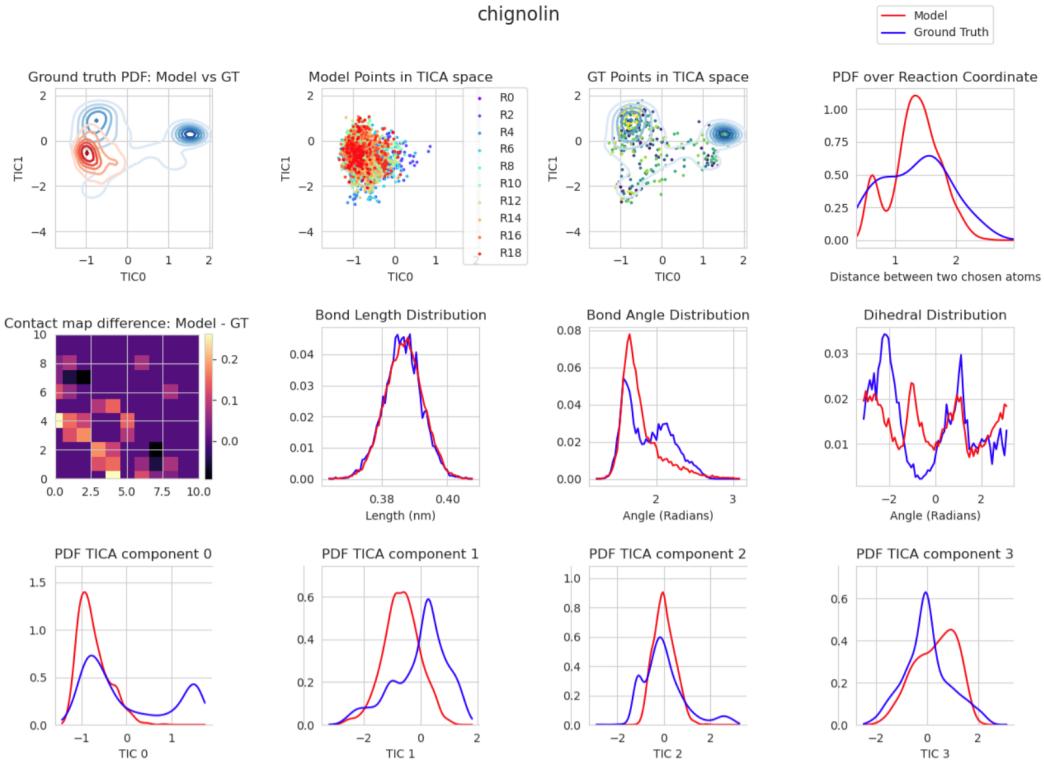


Figure A.1: Benchmark suite evaluating the base model’s performance before applying the active learning loop.

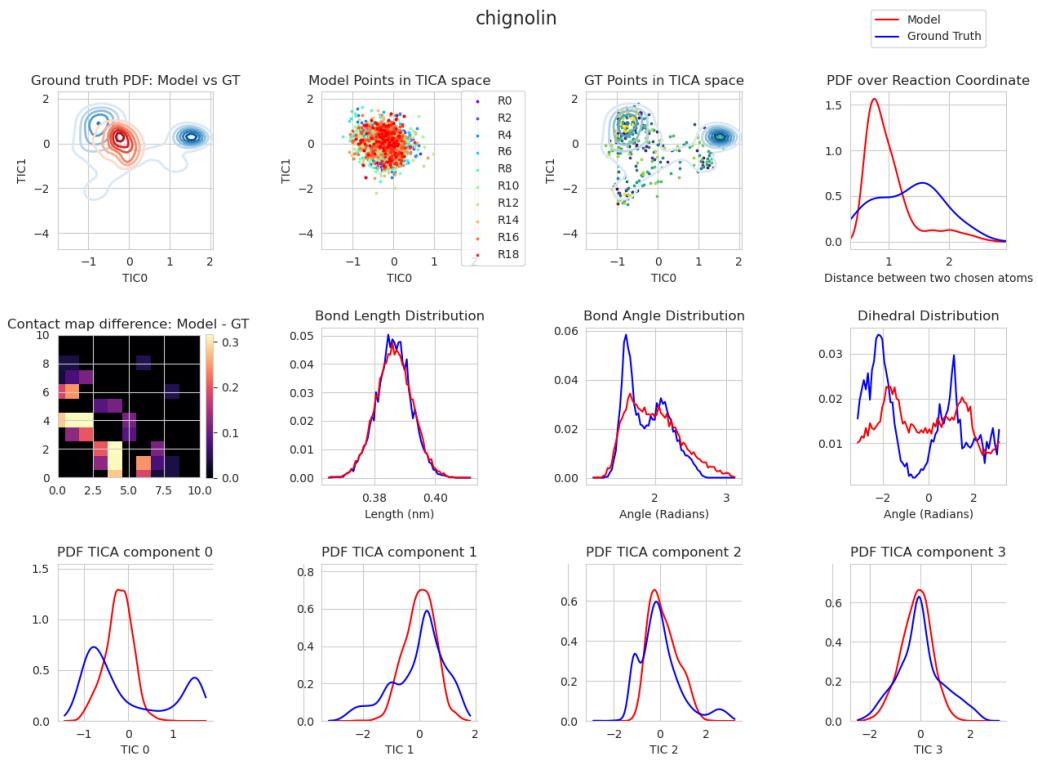


Figure A.2: Benchmark suite evaluating the base model’s performance after applying the active learning loop. Note: Without an unreasonable amount of computation time, the benchmark is unable to explore the entire space, and therefore the first graph won’t reach the entire ground truth distribution, however the improvement is still visible, as this post active learning benchmark sees a more spread out distribution, implying more exploration and coverage.