Predicting Molecular Dynamics using SchNet and sGDML

Anurag Singh (23110035)

Abstract– *This project aims to build and extend the observations from the study by Vassilev-Galindo, Valentin, et al., titled "Challenges for machine learning force fields in reproducing potential energy surfaces of flexible molecules" (The Journal of Chemical Physics, 2021). The focus is on evaluating two state-of-the-art machine learning models, SchNet and sGDML, for their ability to predict molecular forces and potential energy surfaces. Using the same dataset as the referenced study, this work explores the implementation challenges, computational requirements, and model performance, highlighting key observations and insights related to molecular dynamics predictions.*

I. INTRODUCTION

Molecular dynamics simulations are pivotal in computational chemistry, enabling the prediction of molecular properties and behaviours with high accuracy. Traditional methods, such as density functional theory (DFT), can be computationally intensive, prompting the development of machine learning models like SchNet and sGDML. These models have shown promise in efficiently predicting forces and potential energy surfaces for various molecular systems. This project focuses on implementing and evaluating SchNet and sGDML, comparing their performance with results reported in the referenced study.

II. USED DATASETS

The dataset utilised in this project was obtained from the supplementary materials provided by the authors of the referenced journal. It comprises multiple molecular geometries stored in `.xyz` files, each containing detailed information about atomic positions, forces, and energies for various molecular configurations.

Dataset Characteristics:

- **File Format:** `.xyz` files
- **Molecules present:** The molecules present in the dataset consist of Azobenze and Glycine. There are two datasets for the inversion and rotation geometries of Azobenzene.
- **Features**:
  - **Atomic Coordinates**: 3D Cartesian coordinates of each atom in the molecule.
  - **Forces**: Vector forces acting on each atom for different configurations.
  - **Energies**: Potential energy corresponding to each molecular geometry.

The dataset provides a robust foundation for training and evaluating machine learning models in predicting molecular properties.

III. MODELS AND LIBRARIES

A. SchNet:

SchNet is a deep neural network architecture specifically designed for predicting molecular properties using continuous-filter convolutional layers. It is capable of learning complex representations of molecular interactions and is particularly effective in handling flexible molecules with varying geometries.

Implementation Details:

- The model was initialised with standard hyperparameters such as learning rate, embedding dimension, cutoff radius, etc.
- Challenges in implementation included the limited number of provided examples in the documentation, necessitating an in-depth analysis of the model architecture and extensive hyperparameter tuning.

B. sGDML:

sGDML (symmetrised Gaussian Process Regression with gradient-domain machine learning) is an extension of the GDML model designed to capture the symmetries inherent in molecular dynamics data. It emphasises learning the force field of a molecule through a kernel-based approach, making it highly suitable for predicting forces and energies.

Implementation Details:

- The model was initially tested using the command-line interface (CLI), which provided a streamlined setup but did not save trained models consistently.
- Subsequently, the Python API was employed for greater control and customisation. This shift required significant modifications to the provided example in the documentation to align with the project's specific requirements.

IV. CODE AND COMPUTATION

The code development phase involved adapting example scripts from the respective model documentation. Given the limited use-case scenarios covered in the examples, several modifications were necessary to suit the specific dataset and task requirements. There were weeks of debugging and code writing before making the code to be able to work.

SchNet:

- Data preprocessing steps were implemented to ensure compatibility with the SchNet input format, including normalization of atomic features and splitting data into training and validation sets.
- Multiple hyperparameters, such as learning rate, batch size, and the number of training epochs, were adjusted iteratively to optimize model performance.

sGDML:

- Data preprocessing included extracting forces and energies from `.xyz` files and formatting them for input into the sGDML model.
- Both CLI and Python API implementations were tested, with the latter offering greater flexibility for hyperparameter tuning and evaluation.

The high computational demands of training SchNet and sGDML models necessitated the use of high-performance computing (HPC) systems. The following resources were utilised:

Vega, Sabarmati, and Aneesur Servers:

- These HPC systems provided access to GPUs and multi-core CPUs, significantly reducing computation time.
- Despite the powerful hardware, some training runs required up to 48 hours, highlighting the intensive nature of molecular dynamics simulations.

The complete code is uploaded on Github with a readme describing the workflow.

V. RESULTS

The performance of the models was evaluated based on their ability to predict atomic forces and potential energies. The Root Mean Squared Error (RMSE) and Mean Absolute Error (MAE) metrics were used to quantify prediction accuracy.

Schnet (Glycine):

No. of training points	MSE (Forces)	MAE (Forces)	MSE (Energy)	MAE (Energy)
200	6.148072	4.347779	3.0081063	2.3587982
400	4.176167	3.0410645	2.5985936	2.0532281
600	3.9304636	2.8016056	3.268648	2.5794214
800	3.2991348	2.4957296	2.1624171	1.75785734
1000	2.6808451	1.954286	1.959626	1.5336399

Schnet (Azobenzene rotation):

No. of training points	MSE (Forces)	MAE (Forces)	MSE (Energy)	MAE (Energy)
200	6.914219	4.4147671	9.9833332	7.6655284
400	5.932846	3.9486495	8.1492629	7.0055811
600	5.504	3.741273	6.7725578	5.7571779
800	4.833179	3.1448204	6.3032817	5.4517228
1000	4.704212	2.9917579	4.681798	3.5092268

Schnet (Azobenzene inversion):

No. of training points	MSE (Forces)	MAE (Forces)	MSE (Energy)	MAE (Energy)

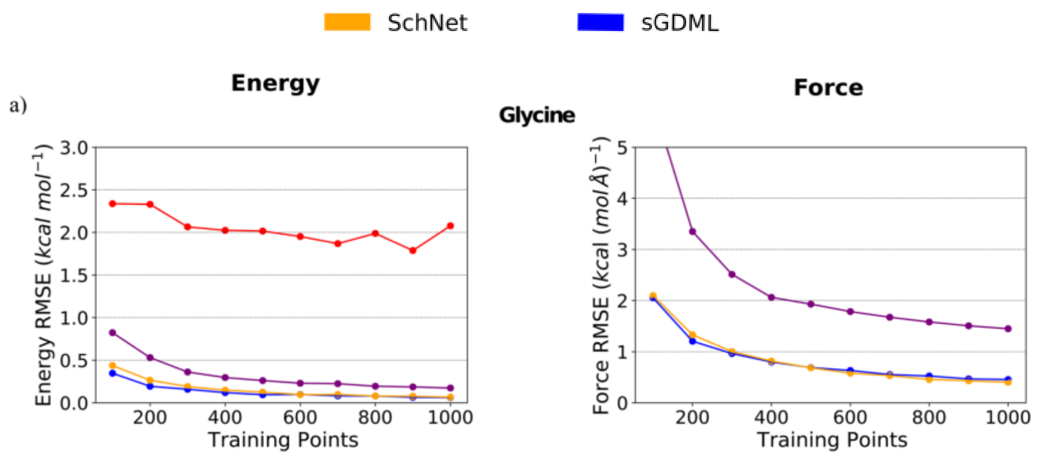
200	6.173528	4.068932	9.8974206	7.5639915
400	5.632793	3.821444	7.3582984	6.1492532
600	4.298277	2.844662	7.3029098	5.6217365
800	4.304237	2.894738	4.9979429	4.3092817
1000	4.018503	2.664945	4.3077302	3.8609439

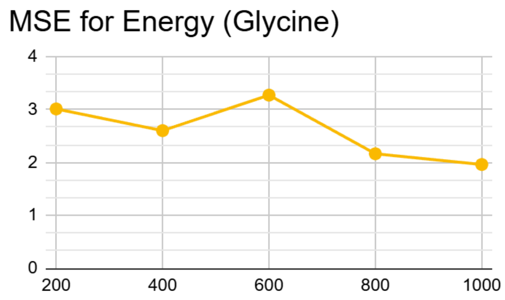
sGDML (Azobenzene inversion):

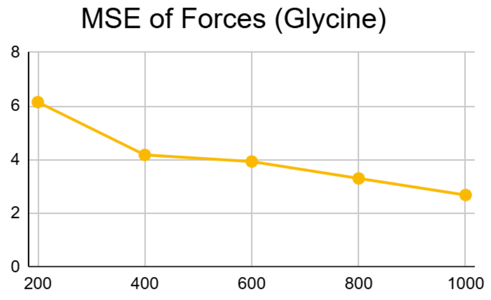
No.of training points	MSE (Energy)	MAE (Energy)	MSE (Forces)	MAE (Forces)
200	11.598008	9.223332	1.766122	1.022714
400	11.5990345	9.2249125	1.285753	0.71735
600	11.6017055	9.2176335	1.037922	0.673556
800	11.611148	9.2346615	0.939713	0.51058
1000	11.5961775	9.2240495	0.853141	0.458767

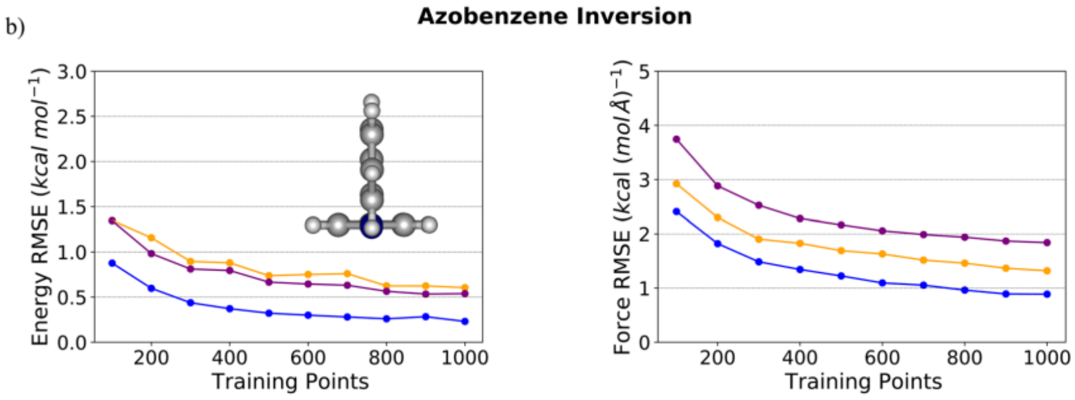
sGDML (Azobenzene rotation):

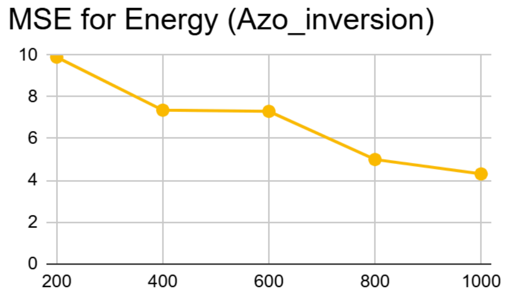
No. of training points	MSE (Energy)	MAE (Energy)	MSE (Forces)	MAE (Energy)
200	12.832626	10.53457	2.848774	1.628703
400	12.42872	9.7031415	2.041232	1.160382
600	12.242866	9.70858	1.656433	0.918274
800	12.203339	9.675524	1.449377	0.804371
1000	12.01032	9.760272	1.294717	0.698458

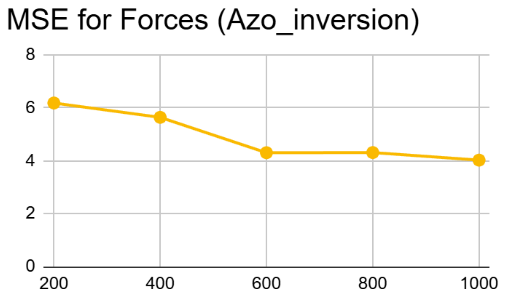


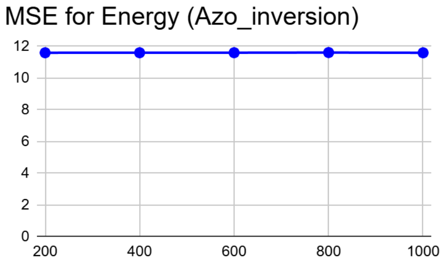


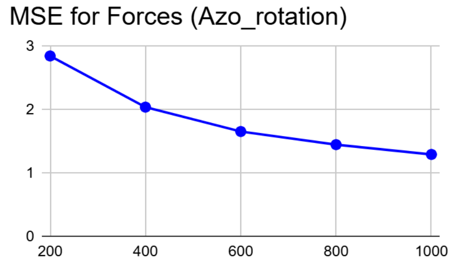


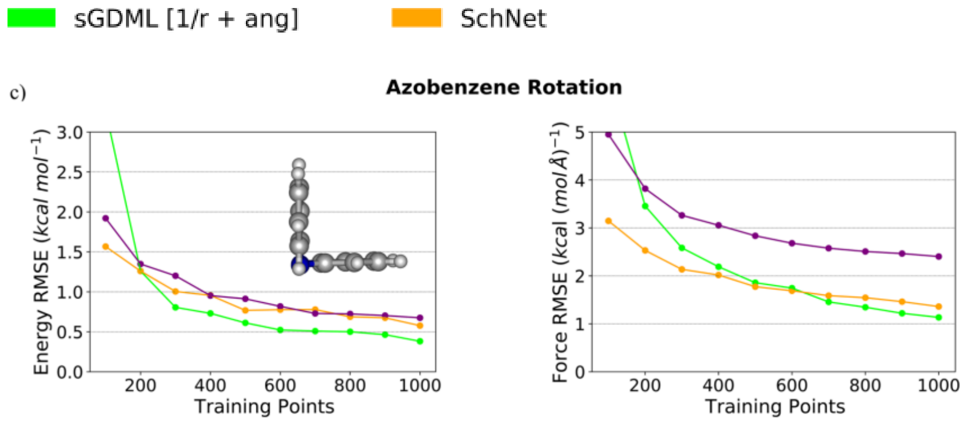


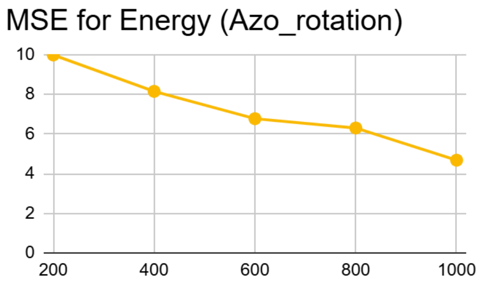


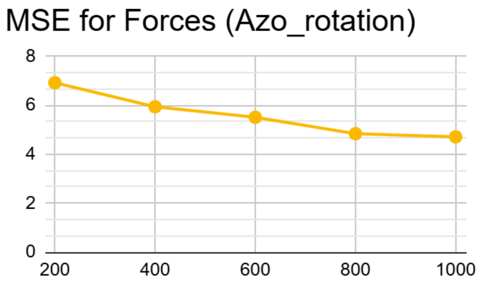


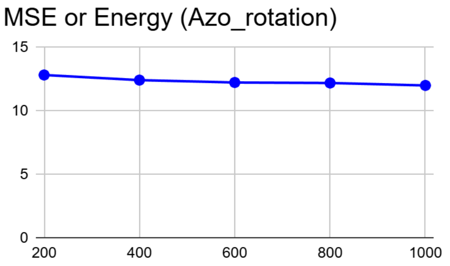


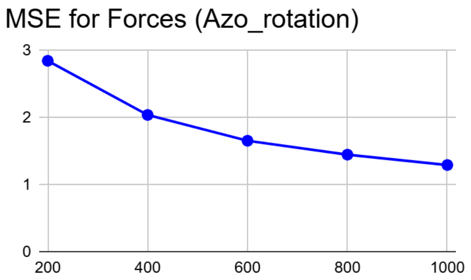












Observations for Schnet

- The RMSE and MAE of forces calculated by Schnet are close to the values reported in the Journal. However, there is a difference of about 3 units in both the RMSE.
- The RMSE and MAE initially came out to be very high it was primarily because of the fact that the value of energy is very large, so reducing the loss doesn't converge the values enough. So, I normalised the values of energy for all three molecules before training and testing. This reduced the RMSE from about 2000 to become less than 10. Even after this, the values reported by the Journal are about 5-6 units or 3-4 times smaller than the RMSE I reported.
- The slope of reduction in RMSE for both forces and energies was higher for my predictions than for the Journal.
- The issue of the higher RMSE values should be resolved after tuning and changing various other hyperparameters or after running more epochs. I primarily focused on 2-3 hyperparameters like embedding dimension, learning rate, and batch size. There are other hyperparameters like cutoff radius, no. of interactions, no. of Gaussian basis, etc., which can be tuned to get better results.

Observations for sGDML:

- The model was not able to load the energy in the Glycine dataset, and thus, I was unable to train the model for Glycine.
- sGDML demonstrated superior performance in force predictions, attributable to its focus on the symmetric nature of the molecular data, thus better capturing underlying physical properties. The RMSE is almost the same as reported by the Journal.
- The RMSE for energy is about 10-12 units higher than the value observed in the Journal. Moreover, the RMSE doesn't reduce significantly even when the number of training points is increased, indicating that the model doesn't focus enough on reducing the loss of energy.

VI. CONCLUSION

This project successfully implements and adapts the SchNet and sGDML models to replicate the results from the Journal of Chemical Physics (Vassilev-Galindo et al., 2021). The results confirm that the sGDML model can be directly applied for accurate force predictions, closely aligning with the RMSE values reported in the journal. However, the model's performance in energy predictions was subpar, with RMSE higher than expected, indicating that sGDML requires further adjustments to capture energy dynamics effectively. For SchNet, while the model performs well in both predictions, further hyperparameter tuning is required to reduce the RMSE, suggesting that additional optimisation of parameters is necessary.

Thus, this project creates a comprehensive workflow for processing molecular data, starting from handling raw data in the form of .xyz files and culminating in making predictions using advanced machine learning models. The workflow involves data preprocessing steps such as normalising atomic features and extracting relevant information, ensuring compatibility with the input formats required by both SchNet and sGDML. The code developed not only facilitates the preparation of datasets but also automates the training and testing of these models, providing a streamlined process for replicating and validating results.

Overall, the project establishes a robust framework for utilising machine learning models in molecular dynamics simulations, offering a foundation for further refinement and application to more complex systems.

VII. REFERENCES

[1] V. Vassilev-Galindo, M. E. Tuckerman, T. Müller, and J. Behler, "Challenges for machine learning force fields in reproducing potential energy surfaces of flexible molecules," *The Journal of Chemical Physics*, vol. 154, no. 9, pp. 094118, Mar. 2021. doi: 10.1063/5.0036522.

[2] K. T. Schütt, P. K. T. Unke, and M. Gastegger, "SchNet: A continuous-filter convolutional neural network for modeling quantum interactions," *The Journal of Chemical Theory and Computation*, vol. 14, no. 11, pp. 5820–5831, Nov. 2018. doi: 10.1021/acs.jctc.8b00908.

[3] S. Chmiela, H. E. Sauceda, K. R. Müller, and A. Tkatchenko, "sGDML: Constructing accurate and data-efficient molecular force fields using machine learning," *Nature Communications*, vol. 9, no. 1, pp. 1–10, Oct. 2018. doi: 10.1038/s41467-018-06169-2.

[4] A. Tkatchenko, S. Chmiela, and H. E. Sauceda, "Symmetric gradient-domain machine learning (sGDML) - An improved machine learning method for molecular dynamics," *GitHub Repository*, Accessed: Nov. 2024. [Online]. Available: https://github.com/stefanch/sGDML

[5] K. T. Schütt, M. Gastegger, A. Tkatchenko, K. R. Müller, and R. J. Maurer, "SchNetPack: A deep learning toolbox for atomistic systems," *GitHub Repository*, Accessed: Nov. 2024. [Online]. Available: https://github.com/atomistic-machine-learning/schnetpack