Machine Learning Approaches to Accelerate Molecular Design

Abstract

Machine learning (ML) has emerged as a transformative tool in computational chemistry, offering new approaches to accelerate and enhance MD simulations. By leveraging the power of ML algorithms, researchers can develop more efficient force fields, analyze complex Molecular Dynamics (MD) trajectories, and predict properties of molecules quickly and efficiently which enables the quick design of novel materials with desired properties. This paper analyses three such works on using ML models in computational chemistry. The first work is on reproducing Molecular DYnamics trajectoris using a Long Short Term Memeory model (LSTM). The second is on using Reinforcement Learning (RL) to predict Conformers of molecules. The final paper is on using the method of macimizing mutual information to pretrain a 2D Graph Neural Network (GNN) to learn 3D gemotric features that is learned by a 3D GNN.

1. Introduction:

Machine learning has become a valuable tool in various areas of scientific research, including computational chemistry. These machine learning methods offer new approaches to accelerate and analyze complex Molecular Dynamics trajectories, and design new materials with desired properties.

Molecular dynamics (MD) simulations have emerged as a powerful tool for understanding the behavior of molecules at the atomic level. By tracking the motions of atoms over time, MD simulations can provide molecular insights. However, traditional MD simulations are often computationally expensive, limiting their applicability to large or complex systems. Similarly, many properties of molecules that have to be calculated have a strong correlation with the different possible shapes or conformations of a molecule. Once again the traditional methods used to find different conformers are computationally expensive. Finally, many methods that are used to calculate the various properties of these moleculeas also rely on time consuming quantum chemical calculations.

Machine learning offers several advantages over traditional methods for predicting molecular properties. ML algorithms can learn from large datasets of molecular structures and their corresponding energy levels and other properties, enabling them to aid in designing new materials and molecules efficiently.

2. Paper-1: Learning molecular dynamics with simple language model built upon long short-term memory neural network

2.1. Storyline:

High Level Motivation: In this paper, the authors propose a simple language model based on Long Short-Term Memory (LSTM) neural networks to learn the temporal evolution of molecular dynamics trajectories. They show that their model can capture the Boltzmann statistics of the system and reproduce kinetics across a spectrum of timescales. The authors' approach is motivated by the fact that molecular dynamics trajectories can be mapped into a sequence of characters by discretizing the spatial coordinates of the particles. This allows them to use a character-level language model to learn the dynamics of the system. The authors train their model on a dataset of molecular dynamics trajectories generated from a variety of systems, including a simple 4-state model potential which is the work that has been implemented here. They show that their model can accurately predict the future states of the system.

Prior work on this problem: There has been a growing interest in using machine learning to predict molecular dynamics trajectories in recent years. This is motivated by the fact that MD simulations can be computationally expensive, especially for large systems and long timescales. Machine learning models can be used to accelerate MD simulations by predicting the future states of the system, thereby reducing the need to simulate the system for its full length. One of the approaches to using machine learning for MD prediction. One of the most common approach is to use graph neural networks. Graph neural networks are able to learn from the structural information of the system, which can be useful for predicting the dynamics of complex systems.

Research Gap: The research gap that the paper addresses is the need for simple and efficient machine learning methods for predicting molecular dynamics trajectories. Previous approaches to using machine learning for MD prediction have been either inaccurate or computationally expensive.

Contributions: The authors of this paper propose a simple language model based on Long Short-Term Memory (LSTM) neural networks that is both accurate and efficient. Their model is able to capture the Boltzmann statistics of the system and reproduce kinetics across a spectrum of timescales.

2.2. Proposed Solution:

The Porposed soultion is the application of a LSMT model to predict MD trajectries. This paper doesn't involve any new algorithms but the novelty is the application of a tried and tested model for a completely new numerical methods which is molecular dynamics. The approach is motivated by the fact that molecular dynamics trajectories can be mapped into a sequence of characters by discretizing the spatial coordinates of the particles. This allows them to use a character-level language model to learn the dynamics of the system.

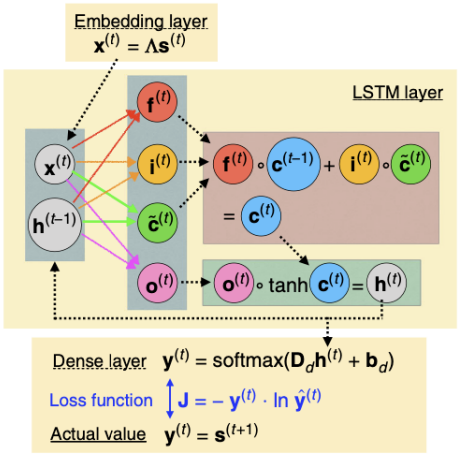

Figure 1. Neural Network Architecture

2.3. Claim And Evidence:

Claim 1: In this paper, the authors claim a simple language model based on Long Short-Term Memory (LSTM) neural networks can learn the temporal evolution of molecular dynamics trajectories. Further. they claim that their model can capture the Boltzmann statistics of the system and reproduce kinetics across a spectrum of timescales.

Evidence: The authors first test their LSTM model's ability to capture the Boltzmann weighted statistics for the different states in each model potential. This is the probability distribution P or equivalently the related free energy

$$F = (-1/\beta) log P$$

and can be calculated by directly counting the probabilities from the trajectory. As can be seen in Fig. 2, the LSTM does an excellent job of recovering the Boltzmann probability within error bars.

Claim 2: One can represent molecular dynamics trajectories as a bunch of characters.

Evidence: One can project the high-dimensional data along many different possible low-dimensional order parameters, for instance x, y, or a combination thereof. However, most such projections will end up not being kinetically truthful and give a wrong impression of how distant the metastable states actually are from each other in the underlying high-dimensional space. However, if we have discretize the potential energy surface based on there low-dimensional order parameters, we can assign a characters to each of these discretized buckets and use that as input to the LSTM.

Claim 3: In this paper, the authors further claim that their LSTM model is agnostic to the quality of the low-dimensional projection in capturing accurate kinetics.

Evidence: This model is aganostic to projection because the molecular dynamics trajectory is not projected into a lower dymension suchas X or Y axis rater it gets mapped on to a lower state space as in the new dimension encodes the state in which the system is. For example eiter state A or B or C in figure 1 a,b,c. Since this lower dimensional mapping into the state sapce doesn't have any loss of information the authors are able to reproduce accurate kinetics.

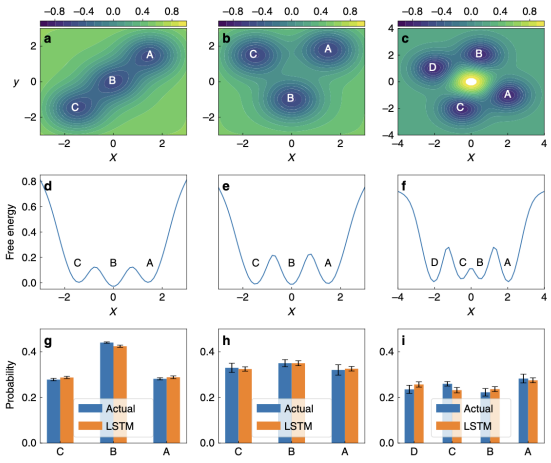

Figure 2. Reproduction of Boltzmann Statisics for the 3 different system.

3. Paper-2: TorsionNet: A Reinforcement Learning Approach to Sequential Conformer Search

3.1. Storyline:

High Level Motivation: Predicting the 3D shapes of flexible molecules is a key goal of computational chemistry, with wide applications in drug design. However, this is a very challenging problem because the number of possible conformers grows exponentially with the size of the molecule. While the conformational space of a molecule is continuous, there are a finite number of stable, low-energy conformers that correspond to local minima on the energy surface. Accurately predicting these stable conformers is essential for understanding and predicting the properties of molecules.

Prior work on this problem: Recently there has been some work on using to predict conformer distribution. Supervised approaches require a target dataset of empirically measured molecule shapes, utilizing scarce data generated by expensive experimental methods such as X-ray crystallography or computationally expensive MD simulation moreover these studies are limited to a limited class of small molecules.

Research Gap: The research gap that this paper addresses is the need for more efficient and scalable methods for conformer search. Traditional methods, such as Monte Carlo and Molecular Dynamics, can be computationally expensive for large and complex molecules. Further, Traditional methods for conformer search can be biased towards certain regions of the conformational space. TorsionNet is a reinforcement learning-based approach that has been shown to outperform pyhsics based methods for a variety of molecules, while also being significantly more computationally efficient.

Contributions: TorsionNet is a reinforcement learning-based approach to sequential conformer search. It is a more efficient and scalable alternative to traditional methods, such as Monte Carlo and Molecular Dynamics, which can be computationally expensive for large and complex molecules. TorsionNet has been shown to outperform physics based methods for a variety of molecules while being significantly more computationally efficient. Overall, TorsionNet is a promising new approach to conformer search that has the potential to revolutionize the field of computational chemistry.

3.2. Proposed Solution:

Since pyhsics based menthods for conformer generation are computationally expensive or don't sample the conformer space completely. The authors have proposed a Reinforcement Learning methodscalled TorchNet to address these research gaps.

The author use Message Passing Neural Networks (MPNN) with Long Short Term Memory (LSTM) to generate independent torsion sampling distributions for all torsions at every timestep. Further, they employ curriculum learning, a learning strategy that trains a model on simpler tasks and then gradually increases the task difficulty to train the model. Finally, they developed a novel reward function called the Gibbs Score

The Gibbs Score reward function is a thermodynamic metric that measures the stability of a molecule. The model sequentially rotates a molecule's rotatable bonds, observing the change in Gibbs Score at each step. It is rewarded for selecting torsions that lower the Gibbs Score, indicating more stable conformers.

The TorsionNet model consists of a graph network for node embeddings. This layer essentially encodes the atom and bonding information. The embeded vectors are passed to a LSTM module. This is done to include history of the conformation of the molecule. That is it learns to predict new conformation based on the previously seen conformations. The output of the LSTM is passsed to a fully connected linear layer which gives the action output that is the new angle changes of the molecules that will give a new conformer. Finally this conformer is taken and its energy is calculated using a inexpensive methods and the gibbs score is used to calculate the reward.

3.3. Claim And Evidence:

Claim 1: The authors claim that posing conformer search as a reinforcement learning problem has several benefits over alternative formulations including generative models.

Evidence: Posing conformer generation as an RL problem does have many benefits in that its computationally inexpensive. Additionally, as shown in figure 3 TorsionNet is much faster which also providing superior gibbs score in comparision to the physics based methods.

Table 1: Method comparison of both score and speed on two branched alkane benchmark molecules. All methods sample exactly 200 conformers. Standard errors produced over 10 runs.

Method	11 torsion alkane		22 torsion alkane	
	Gibbs Score	Wall Time (s)	Gibbs Score	Wall Time (s)
RDKit	1.14 ± 0.16	11.41 ± 0.11	1.22 ± 0.43	68.72 ± 0.08
Confab	0.10 ± 0.01	$\mathbf{10.25 \pm 0.02}$	$\leq 10^{-4}$	$\mathbf{26.04 \pm 0.12}$
TorsionNet	$\mathbf{2.38 \pm 0.25}$	15.69 ± 0.03	$\mathbf{4.48 \pm 1.86}$	35.23 ± 0.06

Figure 3. Results of TorsionNet in comprison to confab and rdkit which are 2 physics based methods.

Claim 2: The authors claim that pthe use of Gibbs Score as reward function enables the model to predict conformers better.

Evidence: The authors introduce a novel metric called

the Gibbs Score. The Gibbs Score is a measure of how well a set of conformers represents the Gibbs distribution of conformers for a given molecule. This means that it takes into account both the energies of the conformers and their relative frequencies. The Gibbs Score has a number of advantages over other metrics for conformer generation. First, it takes into account both the energies and frequencies of conformers. Second, it can be used to directly compare the quality of different conformer sets. Third, it guarantees a level of inter-conformer diversity.

Claim 3: The authors also claim that TorsionNet has learned to detect important conformational regions.

Evidence: Although TorsionNet sampled 10x fewer conformers than SGMD, it produces a Gibbs scores higher than physics based methods which demonstrates that TorsionNet sampled low-energy unique conformers far more frequently. Figure 4 shows the correlated motion of molecule. The highest contributions to molecular motion from physics based methods (shown in left matrix) are mostly localized and found along the diagonal, middle, lower right sections of the matrix. With TorsionNet, we can note high correlations in similar regions, specifically the middle and lower right parts of the matrix. This shows that TorsionNet preferred to manipulate torsions in regions that physics based method also deemed to be conformationally significant.

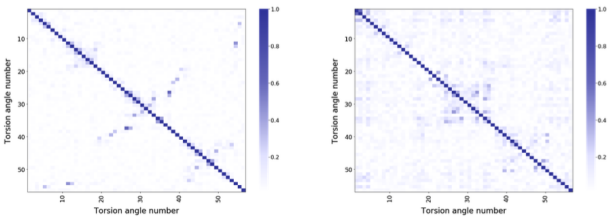

Figure 4. Results of Physics based method (Molecular Dynamics) (Left) vs TorsionNet (Right).

3.4. Critique and Discussion:

The paper introduces a novel application of reinforcement learning (RL) to sequential conformer search, a challenging task in computational chemistry. The proposed RL-based approach, TorsionNet, demonstrates significant performance improvements over traditional methods, particularly for large and flexible molecules. The authors provide a detailed explanation of the implementation of the RL algorithm, including the reward function and state representation. Moreover, they thoroughly evaluate the effectiveness of TorsionNet on a variety of benchmark datasets. One potential limitation of TorsionNet lies in its reliance on a fixed reward function. This could restrict the algorithm's ability to learn optimal conformers for molecules with properties or conformations that differ from those encountered in the training data. Despite these limitations, the paper presents a promising approach to sequential conformer search.

4. Paper-3: 3D Infomax improves GNNs for Molecular Property Prediction

4.1. Storyline:

High Level Motivation: The motivation for this paper is to develop a new method for pre-training GNNs to reason about 3D molecular structure, even when 3D structures are not available. The authors show that 3D Infomax pre-training provides significant improvements for a wide range of molecular properties, including quantum mechanical properties, thermodynamic properties, and spectroscopic properties. They also show that the learned representations are highly generalizable, and can be effectively transferred between datasets in different molecular spaces.

Prior work on this problem: GNNs typically only reason about the local structure of molecules, and do not explicitly model the 3D geometry of molecules. In previous work to incorporate 3D geometry information. Classical methods have been used to generate 3D data which was then fed to the machine learning model, this is a bottleneck.

Research Gap: This is a limitation, as many molecular properties are strongly influenced by the 3D geometry of the molecule. However, computing the 3D geometry of a molecule using classical molecular dynamics simulations is computationally intractable for large-scale applications.

Contributions: This paper proposes a new method for pre-training graph neural networks (GNNs) to improve their performance on molecular property prediction tasks. The authors argue that 3D molecular structure is essential for many molecular properties, but it is often infeasible to obtain 3D structures at the scale required by real-world applications. To address this challenge, the authors propose to pre-train a GNN to reason about the geometry of molecules given only their 2D molecular graphs. They do this by using a self-supervised learning approach called 3D Infomax, which maximizes the mutual information between learned 3D summary vectors and the representations of the GNN.

4.2. Proposed Solution:

The proposed solution is a self-supervised learning method called 3D Infomax. The method pre-trains GNNs to predict/or incorporate information about the 3D geometry of molecules when 3D structures are not available. The method tries to maximizing the mutual information between learned 3D summary vectors and the representations of the GNN. This forces the GNN to learn to encode information about the 3D geometry of the molecule into its representations.

Once the GNN has been pre-trained with 3D Infomax, it can be used to predict the properties of molecules without the need for 3D structures.

The method uses 2 different models. First is a 2D GNN and the second is a 3D GNN and the 3D information learned by the 3D model is used by the 2D model to improve its ability to implicitly learn the 3D geometry information from just the 2D structure. First the 2D model is pre-trained by maximizing the mutual information (MI) between its representation za of a molecular graph G and a 3D representation zb produced by the 3D model. This is done using contrastive learning.

In contrastive learning, we want to make the representations of similar molecules (positive pairs) more similar, and the representations of different molecules (negative pairs) more dissimilar. This is achieved by minimizing the NTXent loss shown in figure 5, which measures the difference in similarity between positive and negative pairs.

$$\mathcal{L} = -\frac{1}{N}\sum_{i=1}^{N}\left[\log \frac{e^{sim(z_i^a, z_i^b)/\tau}}{\sum_{\substack{k=1 \\ k\neq i}}^{N} e^{sim(z_i^a, z_k^b)/\tau}}\right]$$

Figure 5. NTXloss.

4.3. Claim And Evidence:

Target	Rand Init	Our 3D Infomax QM9	Drugs	QMugs
μ	0.4133±0.003	**0.3507**	**0.3512**	0.3668
α	0.3972±0.014	0.3268	0.2959	**0.2807**
HOMO	82.10±0.33	**68.96**	70.78	70.77
LUMO	85.72±1.62	**69.51**	71.38	78.10
GAP	123.08±3.98	**101.71**	102.59	103.85
R2	22.14±0.21	**17.39**	18.96	18.00
ZPVE	15.08±2.83	7.966	9.677	12.06
c_v	0.1670±0.004	0.1306	0.1409	**0.1208**

Figure 6. 3D Infomax performace on predicting various molecular properties from 3 datasets of different molecules which are QM9, DRUGS, QMUGS.

Claim 1: 3D geometry information can better the prediction of molecula rproperties.

Evidence: Graph neural networks (GNNs) that operate on 2D graphs have been successful, many tasks on molecules can be improved by using 3D information. Even simple approaches to encoding geometric information such as those methods that use bond lengths as edge features have been shown to perform better.

Claim 2: A 3D pre-training method that enables GNNs to reason about the geometry of molecules given only their 2D molecular graphs, which improves predictions.

Evidence: As Shown in figure 6, the 2D model pretrained with the 3D geometry information performs much better in all the 3 datasets in predicting various molecular property as compared to random initialization.

Claim 3: The molecular embeddings is generalized over the chemical space.

Evidence: The 2D model is pretrained on a subset of each dataset and the pre trained 2D model is used to predict properties of molecules that doesn't overlap with those in the pretraining set. Even though the 2D model wasn't pretrained on the 3D geometry of all of chemical space it still is ables to perform better than random initialization. This shows that the embedding of 3D geometry information using mutual information maximization is generalizable across chmical space.

4.4. Critique and Discussion:

The paper proposes a novel method for pre-training graph neural networks (GNNs) to improve their performance on molecular property prediction tasks. The method, called 3D Infomax, leverages 3D information to generate better learned embeddings and improves performance on downstream prediction tasks where 3D information is not available. The paper should provide some evidence that the 2D model does indeed learn the 3D embedding information or if the performace boos is just an artifact of better initialization that leads to better optimization of the model. One approach that could be applied is to map the latent vector of the 2D model back to a 3D goemetry and see how well it matches with the original 3D geometry of the molecule. Further, the paper could be strengthened by including a more detailed comparison of the proposed approach to other methods for molecular property prediction. This is becuase the performace advantage shown with this methods appreas to be only narginally better.

5. Implementation:

5.1. Implementation motivation:

I am implementing the first paper on using LSTM to reproduce MD trjectories. The motivation behind implementing the code for reproducing MD trajectories with LSTM networks is to provide a more efficient and accurate alternative to traditional MD simulations. Traditional MD simulations are often computationally expensive, limiting their applica-

bility to large or complex systems. LSTMs, on the other hand are well-suited for modeling sequential data, such as MD trajectories. By using an LSTM to model MD trajectories, researchers can significantly reduce the computational cost of simulations while maintaining high accuracy.

5.2. Implementation Setup:

The LSTM itself consists of the following elements: the input gate i(t), the forget gate f(t), the output gate o(t) the cell state c(t), and h(t) which is the hidden state vector and the final output from the LSTM. Each gate processes information in different aspects. Briefly, the input gate decides which information to be written, the forget gate decides which information to be erased, and the output gate decides which information to be read from the cell state to the hidden state. An addition that is not part of the standart LSTM model is the embedding layer. The embedding layer is a linear layer which multiplies the one-hot input s(t) by a matrix and produces an embedding vector x(t)

The Model used is the standard LSTM model wherein all time series were batched into sequences with a sequence length of 100 and the batch size of 64. The model was trained for 3000 epochs with Adam optimizer and a learning rate of 0.01.

The LSTM and the data generation codes were written by me. The actual molecular dynamics data was taken from the gihub repo: https://github.com/tiwarylab/LSTM-predict-MD. The binning of the data was sone using an in-house code that binns the data based on its eucledian distance to the basin center.

Figure 7 (A) shows the heat map of the MD trajectory. The hotter the regoin the more probable that the molecule in that particular region or state. The Figure 7 (B) shows the binned time series data. You can observe the states on the Y axis (1,2,3,4) and on the x azis is the MD trajectory step

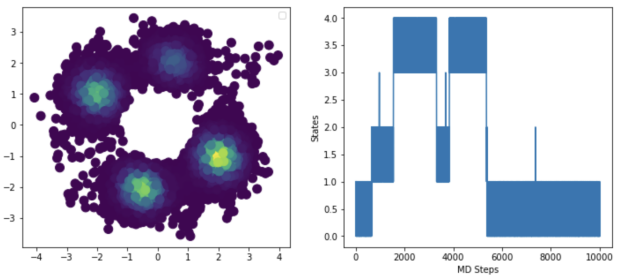

Figure 7. (A) The molecular Dynamics Trajectory Heat Map. (B) Binned Data

5.3. Results and Interpretation:

Figure 8 shows the Training Loss vs Epochs of the LSTM model. We can see that the Loss is decreasing therefore the problem of predicting MD trajectories is learnable and a simple LSTM model is able to achieve it.

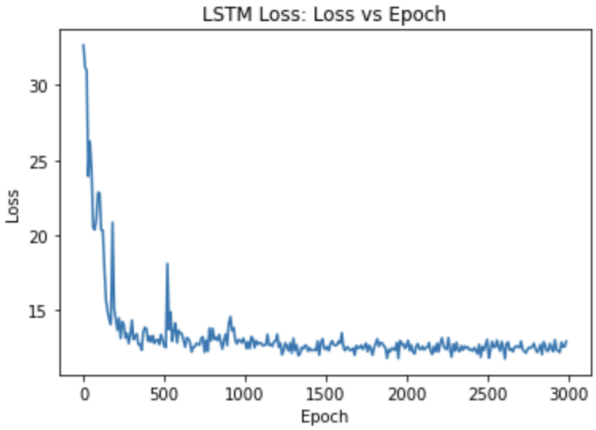

Figure 8. LSTM Loss vs Epochs.

Further in figure 9 The Boltzmann statistics of finding the molecule in a particula state. This is just the count of the time a molecule spends in a particular state divided by the total number of MD steps. We ober a very low MAE of just 0.03 there fore we can conclude that the LSTM model is able to achieve reproduction of the Boltzmann statistics.

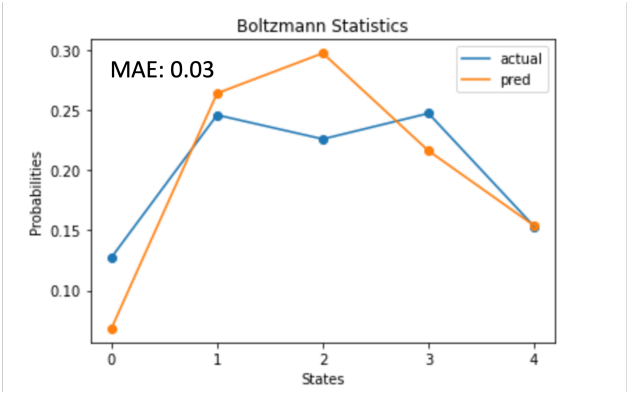

Figure 9. Boltzmann Statistics. The states are 0,1,2,3,4 which corresponds to A,B,C,D in figure 2

5.4. Critique and Discussion:

This paper presents a novel approach to molecular dynamics (MD) using a character-level language model based on long short-term memory (LSTM) neural networks. The authors demonstrate that their approach can capture the Boltzmann

statistics of a system and reproduce kinetics across a spectrum of timescales. One potential limitation of the proposed approach is that it relies on a character-level representation of the molecular system. This representation may not be able to capture all of the important physical properties of the system. Another potential limitation is that the LSTM model is trained on a dataset of MD trajectories. This means that the model may not be able to generalize to systems that are not well represented in the training data. The paper could be strengthened by including a more detailed discussion of the limitations of the proposed approach. Additionally, the authors could provide more evidence that the model can generalize to systems that are not well represented in the training data.

6. Conclusion:

Machine learning (ML) has emerged as a transformative force in computational chemistry, offering new and powerful approaches to accelerate and enhance molecular design. By harnessing the capabilities of ML algorithms, researchers can develop more efficient force fields, analyze complex MD trajectories, and predict molecular properties with great accuracy. This rapid progress in ML-driven computational chemistry paves the way for the accelerated design of novel materials with precisely tailored properties.

This paper presents a comprehensive overview of three applications of ML models in computational chemistry:

Reproducing Molecular Dynamics Trajectories with Long Short Term Memory (LSTM): This study employed an LSTM model to effectively reproduce MD trajectories, along with reproducing accuate kiectics and boltzmann statistics. This study was the basis of my code implementation.

Predicting Molecular Conformers with Reinforcement Learning (RL): This innovative work utilizes RL to predict molecular conformers, a crucial step in understanding molecular behavior and predicting chemical reactivity. RL's inherent adaptability makes it a powerful tool for exploring diverse conformational landscapes and identifying optimal conformations.

Pretraining Graph Neural Networks (GNNs) with Mutual Information Maximization: This novel research introduces a method that leverages mutual information maximization to pretrain 2D GNNs, enabling them to learn 3D geometric features. This approach significantly enhances GNNs' ability to capture molecular properties and accurately predict their behavior.

These innovative applications demonstrate the transformative potential of ML in computational chemistry. ML is poised to revolutionize the field by empowering researchers to tackle complex problems with greater efficiency and accuracy, ultimately leading to the discovery of novel materials and molecules with transformative properties that will shape the future of science and technology.