# Extended Abstract

**Motivation** Molecular dynamics (MD) simulations provide the framework to explore important biomolecular processes at the atomic-level but are limited by prohibitive computational costs when capturing long-timescale events like ligand binding. Bridging this gap with efficient, accurate models that preserve thermodynamic and kinetic realism is critical to advance drug discovery. This motivates the development of reinforcement learning methods that can learn surrogate dynamics, enabling fast generation of realistic protein-ligand trajectories over biologically relevant timescales.

**Method** Pipeline begins by preparing a protein-ligand system through structural cleanup, protonation adjustment, and ligand placement within a scaled bounding sphere around the protein. The combined system is solvated, ionized, and subjected to energy minimization and NPT equilibration, followed by production MD using Langevin dynamics to generate time-resolved trajectories. Post-processing filters out the solvent and surrounding ions, and converts the remaining protein and ligand structures into graph representations and encodes the trajectory dynamics as state-action pairs. Two graph convolutional neural network encoders, trained with supervised imitation learning, models ligand displacement trajectories, with policy inference performed using both a multilayer perceptron (MLP) policy and a diffusion-based policy.

**Implementation** Simulations are performed on the Human Serum Albumin (HSA)–Ibuprofen complex in explicit solvent, using a scaled bounding sphere and water box with physiological ionic strength and temperature settings. Langevin dynamics with a $ps^{-1}$ friction coefficient and 1 femtosecond integration timestep are used to generate time-resolved MD trajectories. The model architecture employs graph convolutional neural network (GCNN) encoders for both the protein and ligand, with policy learning explored via an MLP and a diffusion-based policy, each with hidden dimension 128. Training uses an AdamW optimizer with a learning rate of $1 \times 10^{-3}$ and mean squared error loss, without applying a learning rate scheduler.

**Results** Simulations of both charged and uncharged ibuprofen validates the MD pipeline, with only the deprotonated form exhibiting stable binding to Human Serum Albumin (HSA), in line with experimental observations. Subsequent model evaluation compared an MLP policy against a diffusion-based policy. The MLP policy produced unstable and unphysical ligand motions, highlighting its inability to model the complex near-field energy landscape. In contrast, the diffusion-based policy generated smooth trajectories with stable binding configurations and an average prediction error of 0.93 Å, comparable to experimental and AlphaFold baselines. Additionally, the learned surrogate model provided 100–1000x faster inference speeds than conventional MD, enabling efficient simulation of ligand dynamics. Together, these results demonstrate that pairing reinforcement learning with a diffusion-based policy yields a physically accurate and computationally efficient model of protein-ligand interactions.

**Discussion** While the diffusion-based policy achieved strong accuracy and physical realism, the current MD training data's 1 picosecond resolution may miss faster molecular fluctuations critical to binding kinetics. Future work will incorporate higher-resolution simulations, richer state representations, more expressive architectures (e.g., LLMs, equivariant models), and offline algorithms (IQL, CQL) to further improve surrogate model fidelity and efficiency.

**Conclusion** This work demonstrates that reinforcement learning with diffusion-based policies can efficiently model realistic protein-ligand dynamics from MD data. The learned surrogate achieves substantial acceleration while preserving key thermodynamic and kinetic properties of binding. Future enhancements in simulation resolution, model expressivity, and learning algorithms will further expand its utility for drug discovery applications.

# Protein-Agent – an RL Surrogate for Atomistic Molecular Dynamics

**Chetan Chilkunda**
Department of Chemical Engineering
Stanford University
cchilkun@stanford.edu

## Abstract

Molecular dynamics (MD) simulations enable atomic-level exploration of biomolecular processes but remain computationally prohibitive for capturing long-timescale events such as ligand binding. To address this, I develop a reinforcement learning framework that learns surrogate dynamics for efficient generation of realistic protein-ligand trajectories. The pipeline prepares protein-ligand systems via structural cleanup, solvation, equilibration, and production MD, converting resulting trajectories into graph-based state-action pairs. Graph convolutional neural network (GCNN) encoders, trained through supervised imitation learning, support both multilayer perceptron (MLP) and diffusion-based policy inference. Implemented on the Human Serum Albumin (HSA)–Ibuprofen complex, the diffusion-based policy outperformed the MLP and is comparable to AlphaFold3 and experimental PDB binding poses, generating smooth, physically realistic ligand trajectories with an average prediction error of 0.93 Å and 100–1000x faster inference than MD, while preserving binding thermodynamics. Results demonstrate that diffusion-based RL policies offer a computationally efficient, physically accurate alternative to conventional MD for modeling protein-ligand interactions. Future work will incorporate higher-resolution data, richer architectures, and advanced offline RL algorithms to further enhance surrogate fidelity and accelerate drug discovery workflows.

## 1 Introduction

Molecular dynamics (MD) simulations capture the behavior of proteins and other biomolecules over time and predict how each atom will interact with its local environment based on the governing physics and interatomic action potentials. Karplus and McCammon [2002] These simulations can model diverse and important biomolecular processes such as conformational changes, ligand binding, protein folding, and can predict how biomolecules will *respond* to perturbations at the atomic level and at a femtosecond temporal resolution. MD simulations have become increasingly more popular and visible in recent years, with a key breakthrough in structural biology in AlphaFold2 Jumper et al. [2021] for protein structure prediction and the advancement in GPU processors. Salomon-Ferrer et al. [2013], Hollingsworth and Dror [2018]

An MD simulation takes the positions of all the atoms in a biomolecular system (e.g., a protein in aqueous solution or positioned in a lipid bilayer) and calculates the force exerted on each atom by all the other atoms. This computed, atom-specific net force is then stepped with Newton's Laws of motion to advance the simulation system. By iteratively updating the atomic positions and velocities, MD generates a time-ordered series of molecular configurations — a *trajectory* — that describes the dynamical evolution of the system. This trajectory captures rich temporal information about the biomolecular system and can be analyzed to extract insights into structural dynamics, interaction networks, thermodynamics, and kinetics. de Oliveira et al. [2006], de Groot and

Grubmüller [2001] Modern MD simulation tools paired with GPU hardware speedups allow for the generation of trajectories spanning hundreds of nanoseconds to milliseconds, making it possible to observe biologically relevant timescales and phenomena that are often inaccessible via experimental methods. Borhani and Shaw [2012]

MD simulations are especially relevant to modern drug discovery, where they can complement and inform experimental screening by providing mechanistic insights into dynamic phenomena such as ligand binding pathways, induced fit, allosteric modulation, and transient intermediate states. Dror et al. [2011] Unlike traditional experimental assays, which typically offer static views of molecular interactions, MD trajectories yield atomically detailed, time-resolved perspectives on binding events. Moreover, when combined with enhanced sampling and free energy methods, MD can quantitatively predict binding affinities and elucidate the the kinetics of molecular recognition. Abel et al. [2017]

A major limitation of MD, however, is its inherent computational inefficiency. While simulations operate at femtosecond time steps to maintain physical accuracy and numerical stability, many biologically relevant processes — such as ligand binding and unbinding — occur on timescales of microseconds to seconds or longer. Consequently, observing such rare events through brute-force MD requires prohibitively long simulations that demand extensive computational resources. Hollingsworth and Dror [2018] This time-scale gap severely limits the practicality of MD for routine exploration of protein-ligand interactions, especially in the context of high-throughput drug screening and discovery.

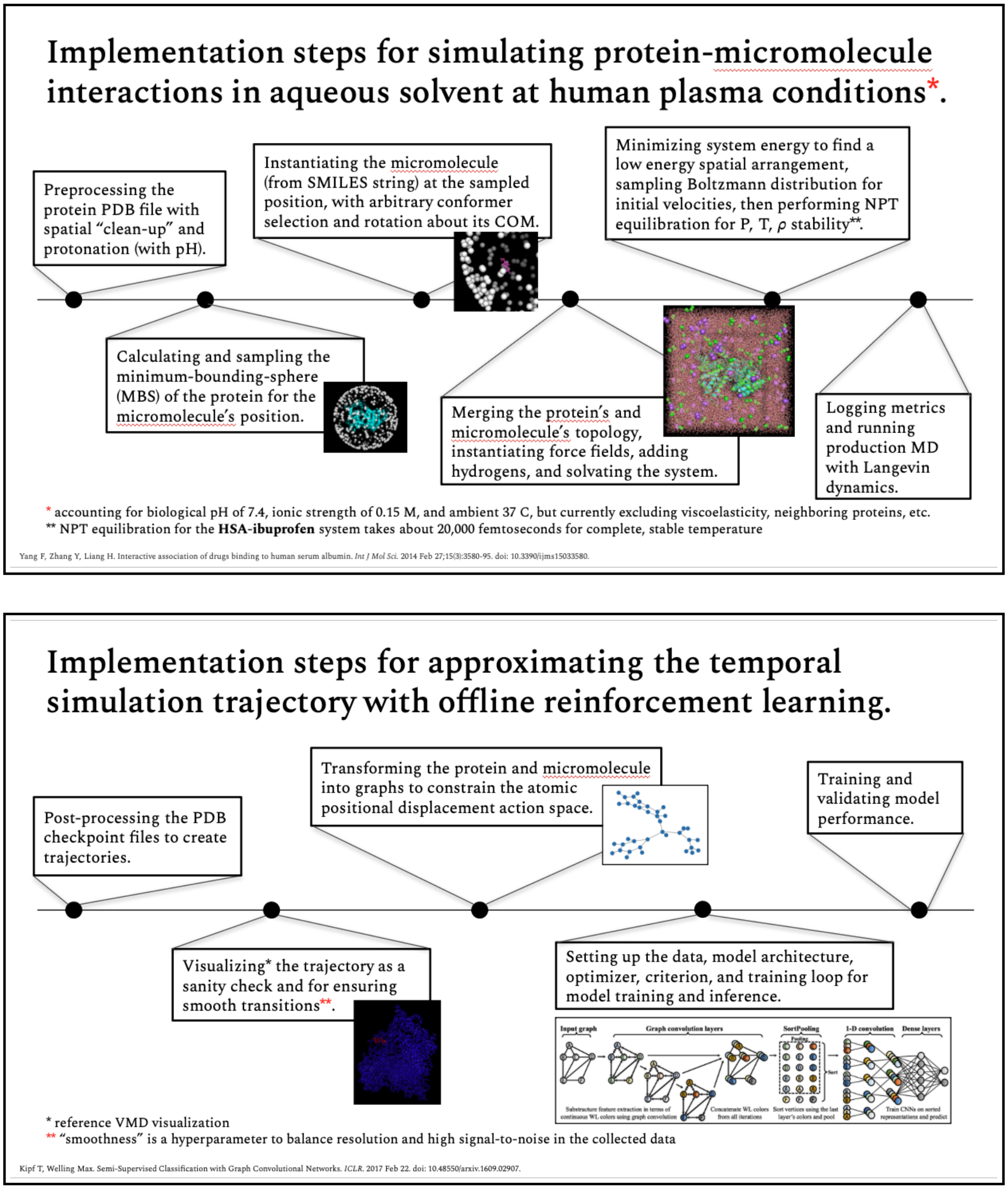
This work addresses the problem of computational inefficiency in MD simulations by employing reinforcement learning (RL) to learn a surrogate model of the protein's near-field potential energy landscape. The core idea is to represent the MD trajectory as a sequence of state-action pairs, where the state encodes the spatial configuration of the protein-ligand system and the action specifies the ligand's displacement through space and time. In this framework, the protein implicitly defines the energy landscape over which the ligand moves, and the RL model learns to predict these dynamics. The resulting surrogate aims to generate realistic protein-ligand trajectories that preserve key thermodynamic and kinetic properties, while achieving orders-of-magnitude improvements in computational efficiency over brute-force MD.

## 2 Related Work

Existing work in predicting protein-ligand interactions include more conventional deep learning approaches like AlphaFold3 Abramson et al. [2024], Diff Dock Corso et al. [2022], among other docking models trained on crystallographic 3D structures in the Protein Data Bank. Fan et al. [2019] Recently, reinforcement learning (RL) has also been applied to *rigid* ligand docking prediction (where the protein structure is fixed). Wang et al. [2022] developed an asynchronous advantage actor–critic model for protein–ligand docking within a unified framework for pose exploration and scoring. This RL method addresses static pose optimization but does not capture the temporal evolution of binding and unbinding events. Wohlwend et al. [2024], Passaro et al. [2025] leverage similar architecturial components to that in the AlphaFold3 model but focus more specifically on protein-ligand binding with Passaro et al. [2025] using *equilibrium* MD to train the model over an ensemble of bound crystallographic structures.

These works demonstrate the potential of deep learning and RL-based methods to improve pose prediction and binding affinity estimation. However, they largely rely on static crystallographic structures or equilibrium MD snapshots, and do not explicitly model the full dynamical trajectory of protein-ligand interactions over time. They do not address the challenge of learning the continuous, temporally coherent dynamics of ligand binding and unbinding pathways across the protein's complex potential energy landscape. Furthermore, while RL has shown promise in static docking optimization, its application to learning dynamic, time-resolved ligand trajectories remains largely unexplored.

Learning a protein's potential energy landscape and predicting full dynamic trajectories are important for a multitude of reasons including (1) accurately predicting equilibrium dissociation constants ($K_d$), which inherently depend on both thermodynamic and kinetic properties; (2) enabling the simulation of complex biological systems involving multiple proteins, co-factors, and regulatory elements; and (3) facilitating the construction of comprehensive virtual cell models, where understanding the temporal behavior of molecular interactions is essential.



## Implementation steps for simulating protein-micromolecule interactions in aqueous solvent at human plasma conditions*.

Preprocessing the protein PDB file with spatial "clean-up" and protonation (with pH).

Calculating and sampling the minimum-bounding-sphere (MBS) of the protein for the micromolecule's position.

Instantiating the micromolecule (from SMILES string) at the sampled position, with arbitrary conformer selection and rotation about its COM.

Merging the protein's and micromolecule's topology, instantiating force fields, adding hydrogens, and solvating the system.

Minimizing system energy to find a low energy spatial arrangement, sampling Boltzmann distribution for initial velocities, then performing NPT equilibration for P, T, $\rho$ stability**.

Logging metrics and running production MD with Langevin dynamics.

* accounting for biological pH of 7.4, ionic strength of 0.15 M, and ambient 37 C, but currently excluding viscoelasticity, neighboring proteins, etc.
** NPT equilibration for the **HSA-ibuprofen** system takes about 20,000 femtoseconds for complete, stable temperature

Yang F, Zhang Y, Liang H. Interactive association of drugs binding to human serum albumin. *Int J Mol Sci.* 2014 Feb 27;15(3):3580-95. doi: 10.3390/ijms15033580.

## Implementation steps for approximating the temporal simulation trajectory with offline reinforcement learning.

Post-processing the PDB checkpoint files to create trajectories.

Visualizing* the trajectory as a sanity check and for ensuring smooth transitions**.

Transforming the protein and micromolecule into graphs to constrain the atomic positional displacement action space.

Setting up the data, model architecture, optimizer, criterion, and training loop for model training and inference.

Training and validating model performance.

Input graph
Graph convolution layers
SortPooling
Pooling
Sort
1-D convolution
Dense layers
Substructure feature extraction in terms of continuous WL colors using graph convolution
Concatenate WL colors from all iterations
Sort vertices using the last layer's colors and pool
Train CNNs on sorted representations and predict

* reference VMD visualization
** "smoothness" is a hyperparameter to balance resolution and high signal-to-noise in the collected data

Kipf T, Welling Max. Semi-Supervised Classification with Graph Convolutional Networks. *ICLR.* 2017 Feb 22. doi: 10.48550/arxiv.1609.02907.

Figure 1: Method Overview. Part 1 (top) refers to setting up the molecular dynamics simulation for data generation, Part 2 (bottom) refers to processing the collected data and RL model development.

This represents a key gap in the literature: there is currently no established framework for using RL to generate realistic, time-continuous ligand trajectories that preserve both the thermodynamic and kinetic properties of molecular recognition—while achieving orders-of-magnitude speedup relative to conventional molecular dynamics simulations. Addressing this gap opens the door to scalable, high-throughput simulation pipelines that can bridge the accuracy of MD with the efficiency required for modern computational drug discovery. Yang et al. [2025]

# 3 Method

In developing the **Protein-Agent**, which is aimed to learn the near-field potential energy landscape of a specified protein with respect to surrounding water molecules and ions, the following molecular dynamics simulation pipeline is established to generate high-quality training data and system configurations. This approach addresses notable challenges arising from existing molecular dynamics (MD) datasets, which often suffer from inconsistencies such as varied sampling frequencies and inherent biases toward specific protein–micromolecule complexes or binding pockets. Such variability in sampling frequency leads to large discontinuities between consecutive state-action pairs, disconnecting the underlying physical behavior from the training data. Additionally, the limited and biased availability of MD data necessitates generating trajectory data to effectively train the model.

### Method Part 1: Simulation Data Generation

The pipeline begins with preprocessing the input protein structure file (PDB), where the protein undergoes spatial 'clean up' to resolve missing atoms, incomplete residues, and other structural irregularities. Protonation states of ionizable residues are assigned based on a specified pH to accurately reflect physiological or experimental conditions, ensuring realistic electrostatics for the downstream simulation.

Next, a minimum bounding sphere is computed around the protein structure to spatially constrain the region of interest. Starting from the protein's center of mass (COM), the smallest radius encompassing all protein atoms is determined, and then this sphere is scaled to define a sampling volume for the ligand's initial placement. This approach ensures the ligand's initial position lies within a biologically relevant vicinity of the protein, avoiding unrealistic initial configurations.

The small molecule ligand is instantiated from its SMILES representation and placed at a randomly sampled position within the scaled bounding sphere. To capture conformational diversity and eliminate bias, a random conformer is generated and an arbitrary rotation is applied about the ligand's COM prior to insertion. This step ensures a physically plausible starting geometry and orientation.

Subsequently, the protein and ligand topologies are merged into a single system topology, parameterized with an appropriate force field to model bonding and nonbonding interactions. Hydrogens are added to complete valences, and the complex is solvated explicitly with a water box to mimic physiological conditions forced with periodic boundary conditions. Ions are included to neutralize the system and maintain ionic strength. This assembly produces a unified simulation system with all molecular components and parameters.

Prior to production MD simulations, the system undergoes energy minimization to resolve steric clashes or unfavorable contacts, constrained by an interatomic cutoff of 1 nanometer. Atomic velocities are then initialized by sampling from the Boltzmann distribution at the target temperature, establishing thermodynamic fidelity. A subsequent equilibration phase under the NPT ensemble (constant number of particles, pressure, and temperature) stabilizes density, temperature, and pressure, allowing the system to relax to spatial and thermal equilibrium.

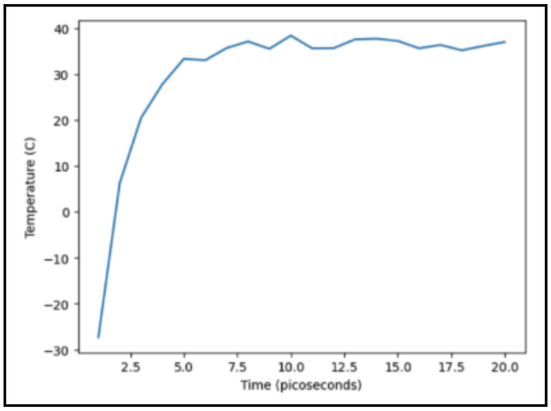


Figure 2: Spatial equilibration of the system to 37 °C for 20,000 femtoseconds.

Finally, production MD is performed using Langevin dynamics (see Equation (1)) to incorporate stochastic thermal fluctuations and maintain temperature control. Throughout these runs, system metrics like potential and kinetic energy, temperature, and pressure are logged to monitor simulation stability and convergence. The resulting trajectories capture the time-resolved motions of the protein-ligand system, providing detailed insight into binding dynamics, conformational changes, and interaction networks – data for training the agent to accurately model the energy landscape and dynamics.

$$m\frac{d^2\mathbf{r}}{dt^2} = -\nabla U(\mathbf{r}) - \gamma m\frac{d\mathbf{r}}{dt} + \sqrt{2\gamma m k_B T}\,\mathbf{R}(t) \tag{1}$$

where $m$ is the mass of the particle, $\mathbf{r}$ its position vector, $U(\mathbf{r})$ the potential energy, $\gamma$ the frictional decay coefficient, $k_B$ the Boltzmann constant, $T$ the temperature, and $\mathbf{R}(t)$ represents a Gaussian thermal fluctuations with zero mean and unit variance.

### Method Part 2: Simulation Data Post-Procesing and Model Development

Following the production MD simulations, the saved protein-ligand checkpoint structures (PDB files) are post-processed to construct temporally ordered trajectories. These trajectories are first visualized as a sanity check, ensuring the structural continuity and physical realism of the system's evolution over time. To prepare data suitable for learning, each checkpoint frame is filtered to retain only the protein and small molecule components, discarding solvent and surrounding ions.

The filtered structures are then converted into graph representations for both the protein and the ligand. In these graphs, atoms act as nodes characterized by the features of atom type and 3D coordinates, while covalent bonds define the edges that capture the molecular topology. The temporal trajectory is reformulated as a sequence of state-action pairs: each state comprises of the protein and ligand graphs at a specific time step, and the corresponding action is the displacement vector of the ligand's atoms from the current to the next time step, effectively encoding the system's dynamic behavior.

The next stage of the pipeline involves constructing the model architecture to learn from this data. Graph convolutional neural networks (GCNNs) are employed to encode the structural inputs of both protein and ligand. The model is trained with supervised imitation learning to predict the small molecule's atomic-level dispacement vector and policy inference is carried out with both a multilayer perceptron (MLP) policy and a diffusion-based policy. The action space is three-dimensional, representing the *(x, y, z)* coordinates of the displacement vector.

Finally, the training setup is completed by initializing an optimizer, defining the regression loss criterion, and establishing a training loop to iteratively update the model. Training is performed on a single prepared trajectory, with the first 80% of the trajectory used for training and the remaining 20% reserved for validation. This approach enables the **Protein-Agent** to learn and generate physically plausible ligand trajectories within the protein's near-field environment, as desired.

## 4 Experimental Setup

The experiment is instantiated using the well-studied protein Human Serum Albumin (HSA) and its specific interacting drug, Ibuprofen. This system occurs in human blood (where Ibuprofen works as a pain reliever!), so its biological conditions are closely mimicked in the simulation. The minimum bounding sphere is scaled to 1.25 times its smallest enclosing radius, and the explicit water box is padded with a 25 Å margin. Simulations are conducted at 37 °C, with an ionic strength of 0.15 M, and a Langevin dynamics (1) friction coefficient of 1 $ps^{-1}$. The simulation logging time resolution is 1 picosecond, with integration performed using a 1 femtosecond timestep.

The model architecture uses a graph convolutional neural network (GCNN) encoder block for both the protein and the ligand. Policy learning is explored with two configurations: a simple multilayer perceptron (MLP) with hidden dimension 128, and a diffusion-based policy architecture composed of a timestep embedding MLP and a noise predictor MLP, each with hidden dimension 128. The optimizer used is AdamW with an initial learning rate of $1 \times 10^{-3}$ and mean squared error (MSE) loss is the regression training criterion. No learning rate scheduler is applied.

## 5 Results

To validate the MD simulation pipeline, simulations were performed on both charged and uncharged forms of ibuprofen under biologically relevant conditions. Ibuprofen, a weak acid with a pKa of

4.91, predominantly exists in its negatively charged (deprotonated) form at physiological pH 7.4. Yamin et al. [2024] Visualization of the resulting trajectories revealed stable binding interactions between the deprotonated ibuprofen and Human Serum Albumin (HSA), consistent with experimental expectations and confirming the accuracy of the simulation setup. In contrast, the neutral (uncharged) ibuprofen failed to exhibit specific and stable binding events, further supporting the validity of the force field parameterization and simulation conditions. This preliminary analysis ensured that the data used to train the model reflected physically realistic interactions.

Subsequent experiments evaluated the performance of the **Protein-Agent** model with two distinct policy architectures: a multilayer perceptron (MLP) policy and a diffusion-based policy. The MLP policy demonstrated significant limitations in modeling the complex, multimodal potential energy landscape characteristic of protein-ligand systems. Trajectories generated by the MLP policy displayed irregular motions, unphysical ligand displacements, and a tendency to escape the protein's near-field environment. Such artifacts reflect the inability of simple feedforward architectures to capture the interatomic forces and temporal coherence required for a realistic simulation.

In contrast, the diffusion-based policy yielded far better results. The generated trajectories exhibited smooth, continuous ligand motions that closely mirrored the corresponding MD trajectories. The diffusion model produced an average mean squared error of approximately 1 Å relative to the ground-truth MD trajectories, a level of accuracy comparable to the resolution of AlphaFold2 and 3 Jumper et al. [2021], Abramson et al. [2024] and experimental PDB crystallographic structures. Importantly, the diffusion-based model demonstrated specific and stable binding configurations to the true binding sites. This emphasizes the gained expressivity with the diffusion policy to capture the (1) binding and unbinding equilibrium fluctuations and (2) the near-field potential energy minimum near the true ibuprofen binding site.

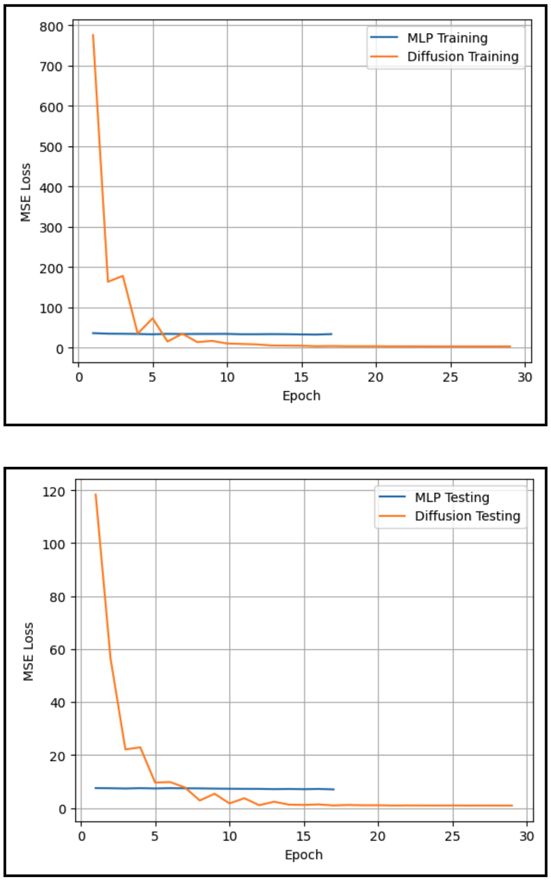


Figure 3: Training and Testing Logs. Training (top) plots the training MSE loss for both the MLP and diffusion policy, Testing (bottom) plots the testing MSE loss for both policies.

Table 1: Performance Comparison (Inference-Time)

| Method | Binding Site Prediction Error (Å) |
|---|---|
| PDB Experimental Baseline | 1.00 |
| AlphaFold3 | 0.96 |
| MLP Policy | 2.63 |
| **Diffusion Policy** | **0.93** |

In addition to improved physical fidelity, the surrogate model achieved significant computational acceleration. Once trained, the **Protein-Agent** generated ligand trajectories several orders of magnitude faster than conventional MD. Inference speeds were approximately 100-1000 times faster than MD simulations executed on identical GPU hardware. This acceleration enables near real-time simulation of ligand dynamics and opens new possibilities for high-throughput virtual drug screening and rapid exploration of protein-ligand interaction landscapes.

These results demonstrate that reinforcement learning, when paired with a diffusion-based policy architecture and supervised with high-quality MD data, can learn a surrogate model of a protein's near-field potential energy landscape. The resulting model preserves both the thermodynamic and kinetic features of ligand binding trajectories while dramatically reducing computational cost, representing an exciting direction for accelerating structure-based drug discovery.

## 6 Discussion

While the diffusion-based policy demonstrated great accuracy and physical fidelity, several limitations remain. The current MD training data is constrained to a temporal resolution of 1 picosecond, which may obscure faster protein and solvent fluctuations that influence binding kinetics. Improving simulation resolution (e.g., sub-picosecond time steps) will be critical for capturing high-frequency molecular motions and providing richer supervision for learning more expressive policies. Future work will also explore enhancing state representations with more atom-specific properties, and increasing policy expressivity via an LLM or equivariant architectures. Additionally, offline algorithms will be employed like IQL and CQL. These improvements aim to further close the gap between surrogate model trajectories and full-scale MD, enabling even more accurate and efficient simulation of protein-ligand dynamics.

Table 2: Summary of Current Implementation

| Component | Current Implementation |
|---|---|
| Data Resolution | One picosecond resolution (1,000 femtoseconds) |
| Node-level Graph Attributes | PDB residue label and (x, y, z) atomic positions |
| State Representation | Graph for both the protein and micromolecule only |
| Action Representation | Displacement information from current state to next state |
| Policy Expressivity | MLP and diffusion model PDF per-state |

Some of the major challenges in developing the pipeline appeared in the simulation and data processing stages. Determining the appropriate sampling resolution required careful experimentation to balance computational cost with the need to capture relevant molecular dynamics. Setting up the MD pipeline and tuning the various simulation parameters – such as temperature control, solvent models, and force field settings—also demanded significant iteration. Additionally, efficiently handling the large volume of generated data posed technical hurdles. Processing over 5,000 trajectory files required developing optimized file I/O pipelines and post-processing workflows to extract usable training data at scale. Addressing these challenges were important for ensuring the quality and utility of the resulting data for policy learning.

## 7 Conclusion

This work demonstrates that reinforcement learning, when combined with diffusion-based policies and high-quality MD supervision, can produce fast, accurate surrogate models of protein-ligand dynamics. The resulting models preserve essential thermodynamic and kinetic properties while achieving orders-of-magnitude computational speedup. These capabilities open new opportunities for high-throughput virtual screening and dynamic modeling of biomolecular interactions. Continued improvements in simulation resolution, model architecture, and learning algorithms will further enhance the fidelity and applicability of this approach in structure-based drug discovery.

## 8 Team Contributions

- **Chetan Chilkunda:** As a solo team, I ideated and implemented all aspects of the MD-RL pipeline, with feedback from TA Jensen Gao over simplifying some of the imitation learning steps to meet the time constraints of the course.

**Changes from Proposal** A major change from the initial project proposal was in the data used to train the model. Upon evaluation of existing datasets, I realized that much of the existing MD trajectory data wasn't at a high enough resolution and sampled initial configurations based on prior knowledge about the known binding site. In order to develop an unbiased RL model to surrogate atomistic MD, I decided to generate my own data from scratch. This took significant effort and, thus, I was only able to experiment over policy expressivity and show proof-of-concept of the surrogate.

## References

Robert Abel, Lingle Wang, Edward D. Harder, B. J. Berne, and Richard A. Friesner. Advancing drug discovery through enhanced free energy calculations. *Accounts of Chemical Research*, 50(7): 1625–1632, 2017. doi: 10.1021/acs.accounts.7b00083.

Jonas Abramson, Jonas Adler, Jack Dunger, and et al. Accurate structure prediction of biomolecular interactions with alphafold 3. *Nature*, 630:493–500, 2024. doi: 10.1038/s41586-024-07487-w.

Dariush W. Borhani and David E. Shaw. The future of molecular dynamics simulations in drug discovery. *Journal of Computer-Aided Molecular Design*, 26:15–26, 2012.

Gabriele Corso, Hannes Stärk, Bowen Jing, Regina Barzilay, and Tommi Jaakkola. Diffdock: Diffusion steps, twists, and turns for molecular docking, 2022. International Conference on Learning Representations (ICLR 2023).

B. L. de Groot and H. Grubmüller. Water permeation across biological membranes: mechanism and dynamics of aquaporin-1 and glpf. *Science*, 294:2353–2357, 2001.

C. A. F. de Oliveira, D. Hamelberg, and J. A. McCammon. On the application of accelerated molecular dynamics to liquid water simulations. *Journal of Physical Chemistry B*, 110:22695–22701, 2006.

Ron O. Dror, Albert C. Pan, Daniel H. Arlow, Dariush W. Borhani, Paul Maragakis, Yibing Shan, Huafeng Xu, and David E. Shaw. Pathway and mechanism of drug binding to g-protein-coupled receptors. *Proceedings of the National Academy of Sciences of the United States of America*, 108 (32):13118–13123, August 2011. doi: 10.1073/pnas.1104614108.

Jie Fan, Aiping Fu, and Lin Zhang. Progress in molecular docking. *Quantitative Biology*, 7:83–89, 2019. doi: 10.1007/s40484-019-0172-y.

Scott A. Hollingsworth and Ron O. Dror. Molecular dynamics simulation for all. *Neuron*, 99(6): 1129–1143, 2018. doi: 10.1016/j.neuron.2018.08.011.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland, Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with alphafold. *Nature*, 596:583–589, 2021. doi: 10.1038/s41586-021-03819-2.

Martin Karplus and J. Andrew McCammon. Molecular dynamics simulations of biomolecules. *Nature Structural Biology*, 9(9):646–652, September 2002. doi: 10.1038/nsb0902-646. Erratum in: Nat Struct Biol. 2002 Oct;9(10):788. PMID: 12198485.

Saro Passaro, Gabriele Corso, Jeremy Wohlwend, Mateo Reveiz, Stephan Thaler, Vignesh Ram Somnath, Noah Getz, Tally Portnoi, Julien Roy, Hannes Stark, David Kwabi-Addo, Dominique Beaini, Tommi Jaakkola, and Regina Barzilay. Boltz-2: Towards accurate and efficient binding affinity prediction, 2025. URL `https://github.com/jwohlwend/boltz`. Preprint. URL: `https://github.com/jwohlwend/boltz`.

Romelia Salomon-Ferrer, Andreas W. Götz, Duncan Poole, Scott Le Grand, and Ross C. Walker. Routine microsecond molecular dynamics simulations with amber on gpus. 2. explicit solvent particle mesh ewald. *Journal of Chemical Theory and Computation*, 9(9):3878–3888, 2013. doi: 10.1021/ct400314y.

Chao Wang, Yuhua Chen, Yuwei Zhang, and et al. A reinforcement learning approach for protein–ligand binding pose prediction. *BMC Bioinformatics*, 23:368, 2022. doi: 10.1186/s12859-022-04912-7.

Jeremy Wohlwend, Gabriele Corso, Mateo Reveiz, Saro Passaro, Stephan Thaler, Tommi Jaakkola, and Regina Barzilay. Boltz: Diffusion-based generative models for structure-based drug design, 2024. URL `https://github.com/jwohlwend/boltz`. Preprint. URL: `https://github.com/jwohlwend/boltz`.

Marriam Yamin, Zafar Khan Ghouri, Nashiour Rohman, Junaid Ali Syed, Adam Skelton, and Khalid Ahmed. Unravelling ph/pka influence on ph-responsive drug carriers: Insights from ibuprofen-silica interactions and comparative analysis with carbon nanotubes, sulfasalazine, and alendronate. *Journal of Molecular Graphics and Modelling*, 128:108720, 2024. doi: 10.1016/j.jmgm.2024.108720.

Ting Yang, Yanyan Wang, Fang Ma, and et al. Build the virtual cell with artificial intelligence: a perspective for cancer research. *Military Medical Research*, 12:4, 2025. doi: 10.1186/s40779-025-00591-6.

## A Additional Experiments

Experiments were run over many different HSA-ibuprofen configurations as well as HSA with warfarin (another known interacting small molecule). Models were trained over multiple configuration trajectories, but after TA feedback from the poster session, I ended up training the models over a single trajectory to show proof-of-concept of RL as a new paradigm to model protein-ligand interactions.

## B Implementation Details

Please see the attached code for full implementation details of (1) MD pipeline, (2) data post-processing, and (3) imitation learning with the MLP and diffusion-based policies.