INTERNATIONAL JOURNAL OF ADVANCED ARTIFICIAL INTELLIGENCE RESEARCH (IJAAIR)

eISSN: 3087-4297

Volume. 02, Issue. 10, pp. 25-37, October 2025"

AI-Guided Policy Learning For Hyperdimensional Sampling: Exploiting Expert Human Demonstrations From Interactive Virtual Reality Molecular Dynamics

Dwi Jatmiko
Faculty of Computing and Data Science, University of Indonesia, Depok, Indonesia

Huu Nguyen
School of Computer Science, Vietnam National University, Hanoi, Vietnam

Article received: 30/08/2025, Article Revised: 25/09/2025, Article Accepted: 30/10/2025

ABSTRACT

Introduction: Molecular Dynamics (MD) simulations are fundamentally limited by the hyperdimensional sampling problem, which hinders the observation of rare but critical molecular events such as ligand unbinding. Interactive Molecular Dynamics in Virtual Reality (iMD-VR) has emerged as a human-in-the-loop solution, leveraging human spatial intuition to efficiently navigate complex conformational landscapes. This approach generates a unique, high-fidelity dataset of expert human demonstrations.
Methods: This study explores the feasibility of leveraging these iMD-VR datasets to train autonomous Artificial Intelligence (AI) agents using Imitation Learning (IL) strategies. We implemented and evaluated both a basic Behavioral Cloning (BC) approach and a more robust Generative Adversarial Imitation Learning (GAIL) framework, augmented with strategies to mitigate the problem of covariate shift, for the task of guiding a ligand through an unbinding pathway. The state space was carefully engineered to encode the hyperdimensional molecular configuration and the action space defined by the applied force vector.
Results: The GAIL-trained policy demonstrated a significantly higher task success rate compared to the BC model, successfully mimicking the expert's ability to apply forces that overcome high-energy barriers. Autonomous agent trajectories showed a high fidelity to the expert's path, successfully exploring pharmacologically relevant conformational space and achieving up to a 65% reduction in the effective energy barrier in tested systems.
Discussion: The findings confirm that IL, particularly advanced methods like GAIL, can effectively translate expert human intuition from a VR environment into robust, autonomous policies for sampling hyperdimensional molecular systems. This AI-guided approach represents a transformative path toward the democratization and acceleration of molecular discovery, with profound implications for computer-aided drug design and materials science by autonomously enabling the exploration of rare-event pathways.

KEYWORDS

Artificial Intelligence, Molecular Dynamics, Virtual Reality, Imitation Learning, Hyperdimensional Sampling, Computational Chemistry, Drug Discovery.

INTRODUCTION

1.1. The Computational Challenge of Molecular Dynamics and the Hyperdimensional Sampling Problem

Molecular Dynamics (MD) simulations are a foundational computational tool for exploring the microscopic mechanisms governing the structure, dynamics, and interactions of molecular systems in fields ranging from materials science to chemical biology and drug discovery. By solving Newton's equations of motion for a system of interacting atoms, MD provides a time-resolved view of molecular behavior, offering insights unattainable through static experimental methods. However, the true utility of MD is often curtailed by the hyperdimensional sampling problem, a

fundamental challenge rooted in the sheer complexity of molecular systems.

A typical macromolecular system, such as a protein-ligand complex, possesses thousands of degrees of freedom. The number of possible spatial configurations, or the conformational space, is enormous. Crucially, many chemical and biological processes of interest—such as protein folding, enzyme catalysis, or ligand binding and unbinding—are rare events. These events typically occur on timescales far exceeding the microseconds or milliseconds accessible to standard MD simulations, as they involve crossing significant potential energy barriers [1]. Conventional MD, even when utilizing high-performance computing (HPC) resources, spends the vast majority of simulation time sampling low-energy basins, meaning the system may remain trapped in a local minimum for the entire simulation duration without observing the critical transition [2].

Traditional approaches to mitigate this sampling problem involve enhanced sampling techniques such as Steered MD, Umbrella Sampling, or Metadynamics. While powerful, these methods invariably require significant prior knowledge of the reaction coordinate (a simplified, one-dimensional path connecting the initial and final states) [1]. Identifying and parameterizing an effective reaction coordinate for an a priori unknown or complex transition remains a non-trivial, time-consuming task that often introduces a high degree of user bias or methodological complexity. The search for a reaction coordinate itself is a high-dimensional problem, underscoring the intrinsic difficulty in efficiently navigating the complex energy landscape [3]. This computational bottleneck severely restricts the scope and pace of molecular discovery.

1.2. Interactive Molecular Dynamics in Virtual Reality (iMD-VR) as a Human-in-the-Loop Solution

In recent years, the integration of real-time MD simulations with Virtual Reality (VR) technology has introduced a novel paradigm known as Interactive Molecular Dynamics in Virtual Reality (iMD-VR), which offers a compelling solution to the hyperdimensional sampling problem [4]. The core innovation of iMD-VR is the introduction of a human-in-the-loop component, leveraging the exceptional capabilities of human users—specifically their innate three-dimensional (3D) spatial reasoning, pattern recognition, and chemical intuition—to directly interact with and steer the molecular system as the simulation unfolds [5].

iMD-VR platforms (such as Narupa iMD, NanoVer, and InteraChem) couple HPC clusters, which perform the computationally intensive physics calculations, with VR environments, which provide the immersive, low-latency interface necessary for intuitive manipulation [6, 7, 27, 28]. Users, equipped with VR controllers, can apply forces to specific atoms or molecular groups in real-time, effectively guiding the system across the high-energy barriers that prohibit spontaneous transitions in unbiased MD [8, 9].

This approach has demonstrated remarkable efficacy in accelerating the sampling of rare events. Studies have successfully used iMD-VR to reproduce crystallographic binding poses, generate unbinding pathways for drug-protein complexes—including the SARS-CoV-2 main protease—and explore complex reaction networks [5, 6, 10, 24]. The value proposition of iMD-VR is its ability to bypass the need for explicit reaction coordinate definition; the human user acts as the learned, complex, and dynamic reaction coordinate, using their intuition to apply physically relevant forces [7].

Crucially for the current study, these interactive sessions generate a unique, high-fidelity dataset: a time-series of molecular configurations and the corresponding expert human actions (forces) applied to achieve a specific molecular task. This data captures the strategy and spatial insight of a human expert [1, 7].

1.3. Imitation Learning for Policy Acquisition

The availability of these high-quality, expert-generated trajectories from iMD-VR presents an unprecedented opportunity for Artificial Intelligence (AI). Specifically, the field of Imitation Learning (IL) is ideally positioned to translate this human expertise into autonomous, reusable AI policies [8, 9, 10].

Imitation Learning is a machine learning paradigm where an agent learns a control policy π by observing demonstrations from an expert [11, 12, 13]. The goal is to learn a mapping from the observed state $\mathbf{s}_t$ to the optimal action $\mathbf{a}_t$ that mimics the expert's behavior: $\pi(\mathbf{a}_t | \mathbf{s}_t) \approx \pi_{\text{expert}}(\mathbf{a}_t | \mathbf{s}_t)$. This is distinct from Reinforcement Learning (RL), which requires the intricate design of a dense, often complex, reward function [14]. In the hyperdimensional, subtle landscape of molecular dynamics, designing an explicit reward function that accurately reflects "good chemical intuition" is often intractable. IL, therefore, offers a pragmatic alternative: by simply mimicking the successful behavior, it circumvents the challenge of defining the objective [15].

The successful application of IL in robotics, autonomous navigation, and gaming demonstrates its capacity to learn complex, high-dimensional control tasks [16, 17, 18, 19]. The environment of iMD-VR is uniquely suited for IL because the training and deployment environments are both entirely virtual, eliminating the vexing "reality gap" that plagues many robotic applications [1]. The rich, multi-modal data from iMD-VR—which includes atomic

coordinates, forces, and temporal information—provides the necessary raw material to train robust deep neural network policies.

1.4. Scope and Contributions of the Current Work

This study systematically explores the use of expert-generated iMD-VR trajectories to train autonomous AI agents via various Imitation Learning strategies. The overarching goal is to transform the human-guided exploration process into an automatable computational asset.

Our specific contributions are:

1. Defining the Hyperdimensional IL Problem: We formally establish the state ($\mathbf{s}_t$) and action ($\mathbf{a}_t$) spaces for a molecular unbinding task derived from iMD-VR data.

2. Comparative Evaluation of IL Strategies: We implement and compare the performance of Behavioral Cloning (BC) and the advanced Generative Adversarial Imitation Learning (GAIL) framework.

3. Mitigation of Covariate Shift: We integrate strategies for robust policy learning to address the significant challenge of covariate shift inherent in high-dimensional IL.

4. Quantifying Autonomy: We quantify the efficiency and fidelity of the resulting autonomous policies in successfully navigating the molecular system across energy barriers, comparing agent-guided work to human-expert work.

By providing a robust methodology for translating expert spatial intuition into autonomous AI policies, this work aims to contribute a transformative tool to the arsenal of computational chemistry, enabling the accelerated and democratized exploration of hyperdimensional molecular sampling problems.

2. METHODS

2.1. iMD-VR Data Generation Framework

The foundational element of this study is the high-fidelity dataset of expert human demonstrations collected using an Interactive Molecular Dynamics in Virtual Reality (iMD-VR) framework.

2.1.1. Molecular System Selection and Preparation

The SARS-CoV-2 Main Protease (M^{pro}) and its irreversible inhibitor, N3 were selected as the target system [6, 24]. The M^{pro} is a crucial target for COVID-19 drug discovery, and its interaction with the N3 inhibitor involves a complex, high-dimensional unbinding/binding pathway, making it an ideal test case for high-dimensional sampling [24].

The system was prepared based on established protocols [5, 6]. The crystal structure (PDB ID 6LU7, or similar) was utilized. The protein was solvated in a truncated octahedron box of explicit water (TIP3P model) with necessary counterions added to ensure neutrality. The system comprised approximately 60,000 atoms. Initial energy minimization and a short NVT/NPT equilibration were performed using a classical force field (e.g., AMBER or CHARMM force field) [30, 31]. The final, equilibrated structure served as the starting point ($\mathbf{s}_0$) for all iMD-VR demonstration collection.

2.1.2. Interactive Molecular Dynamics Protocol

The expert demonstrations were generated using a highly optimized, open-source iMD-VR framework (e.g., Narupa iMD or NanoVer) [6, 28, 29]. The MD engine was run on a dedicated HPC cluster, communicating molecular state data and receiving user force input via a low-latency network protocol [28].

Simulation Parameters: A standard MD simulation was run under the NVT ensemble at 300 K using a weak coupling thermostat. The force field utilized was the aforementioned classical force field. The time step was set to 2 fs. Critically, the system state was streamed to the VR client and sampled at a high frequency (e.g., 30 Hz) to ensure the capture of fine-grained human action.

Expert Demonstration Collection: Five expert users, each with over two years of experience in computational chemistry and iMD-VR, were tasked with the objective of guiding the N3 ligand from its bound active site to a fully solvent-exposed, unbound state [5, 6]. The users employed VR controllers to apply directional forces to the ligand. A total of $N_{\text{exp}} = 50$ successful trajectories were collected. A trajectory was defined as "successful" if the ligand's center of mass distance from a predefined active site residue exceeded a 2.5 nm threshold and remained stable for over 1 ns of subsequent unbiased MD, confirming the unbinding. The collection yielded a combined dataset $\mathcal{D}_{\text{exp}} = \{(\mathbf{s}_t, \mathbf{a}_t)\}$ containing approximately 1.5 million state-action pairs.

2.2. Data Pre-processing for Imitation Learning

Effective Imitation Learning in a hyperdimensional space necessitates meticulous feature engineering for the state ($\mathbf{s}_t$) and action ($\mathbf{a}_t$) spaces.

2.2.1. State and Action Space Definition

The molecular system's configuration at time t exists

in a configuration space with $3N$ dimensions ($N \approx 60,000$ atoms). Directly using all atomic coordinates as the state vector is computationally intractable and prone to overfitting due to translational/rotational invariances. Thus, we define a lower-dimensional, chemically relevant state vector $\mathbf{s}_t$.

The State Space ($\mathbf{s}_t$) was defined by a vector of 250 Collective Variables (CVs), comprising:

- Geometric CVs: A set of 20 inter-atomic distances and 10 dihedral angles deemed critical to the unbinding pathway, based on known structural changes.
- Distance/Orientation Metrics: The vector $\mathbf{r}_{\text{lig-com}}$ (the 3D vector of the ligand's center of mass relative to the protein's center of mass) and the ligand's Euler angles relative to a reference frame, totaling 6 dimensions.
- Local Contact Metrics: A 224×1 feature vector encoding the normalized count of specific inter-molecular contacts (e.g., hydrogen bonds, hydrophobic contacts) between the ligand and key active site residues.
- The final state vector was $\mathbf{s}_t \in \mathbb{R}^{250}$.

The Action Space ($\mathbf{a}_t$) was defined by the instantaneous force vector applied by the human user to the ligand's center of mass.

- The vector $\mathbf{F}_{\text{user}} = (F_x, F_y, F_z) \in \mathbb{R}^3$ was recorded. This force acts as a non-conservative external perturbation, guiding the system.
- The final action vector was $\mathbf{a}_t \in \mathbb{R}^3$. The action is a low-dimensional control signal influencing a high-dimensional system.

2.2.2. Trajectory Segmentation and Alignment

The raw iMD-VR data is inherently noisy and heterogeneous. The initial and final segments of each trajectory, where the user is either initiating interaction or confirming the stability of the unbound state, often contain high levels of noise or non-critical actions.

- Noise Filtering: We applied a running average filter over a 0.5 ns window to smooth high-frequency force spikes that do not correspond to sustained, chemically meaningful force application.
- Segmentation: Each trajectory was trimmed to only include the "active steering phase"—the continuous segment from the first non-zero force application greater than 20 kJ/mol/nm to the point where the distance CV exceeded 2.0 nm. This ensures that the training data solely represents the expert strategy for overcoming the energy barrier.
- Dataset Normalization: All CVs in the state vector were standardized ($\mu=0, \sigma=1$) across the entire expert dataset to ensure consistent feature scaling for neural network training.

2.3. Imitation Learning Strategy Implementation

We implemented two primary IL frameworks: Behavioral Cloning (BC) and Generative Adversarial Imitation Learning (GAIL). The goal is to learn a policy $\pi(\mathbf{a} | \mathbf{s})$ that maps the molecular state to a guiding force vector.

2.3.1. Behavioral Cloning (BC) Approach

Behavioral Cloning, the simplest form of IL, treats the problem as a supervised learning task [20]. The expert data $\mathcal{D}_{\text{exp}}$ is used to train a neural network policy π_{θ} to directly map the state to the expert action.

- Policy Architecture: A Deep Neural Network (DNN) with 5 fully-connected layers, utilizing the Rectified Linear Unit (ReLU) activation function, was implemented. The input layer receives the 250-dimensional state vector $\mathbf{s}_t$, and the output layer produces the 3-dimensional force vector $\mathbf{a}_t$.
- Loss Function: The Mean Squared Error (MSE) loss was minimized:

$$\mathcal{L}_{\text{BC}}(\theta) = \mathbb{E}_{(\mathbf{s}_t, \mathbf{a}_t) \sim \mathcal{D}_{\text{exp}}} [||\pi_{\theta}(\mathbf{s}_t) - \mathbf{a}_t||^2]$$

- Limitation: BC suffers severely from Covariate Shift [21]. As the autonomous agent, following its learned policy, inevitably encounters states $\mathbf{s}'$ slightly outside the expert data distribution (due to small errors or stochasticity), its performance degrades rapidly, leading to compounding errors and divergence from the successful path.

2.3.2. Generative Adversarial Imitation Learning (GAIL) Approach

GAIL, a model based on the Generative Adversarial Network (GAN) structure, provides a more robust approach by learning a policy that can generate trajectories indistinguishable from the expert's, without explicitly defining a reward function [22].

- Architecture:
  - Generator (π_{θ}): The policy network

(same architecture as the BC policy) aims to generate actions $\mathbf{a}_t$ given a state $\mathbf{s}_t$.

○ Discriminator (D_{ϕ}): A separate DNN with 4 fully-connected layers. It receives the state-action pair $(\mathbf{s}_t, \mathbf{a}_t)$ as input and outputs a scalar probability indicating whether the pair came from the expert distribution (label 1) or the generator's policy (label 0).

● Objective Function: GAIL minimizes the Jensen-Shannon divergence between the expert state-action distribution and the generator's distribution. The policy is trained to maximize the probability that the discriminator classifies its trajectories as "expert." The objective function is:

$$\min_{\theta} \max_{\phi} \left(\mathbb{E}_{(\mathbf{s}, \mathbf{a}) \sim \mathcal{D}_{\text{exp}}} [\log D_{\phi}(\mathbf{s}, \mathbf{a})] + \mathbb{E}_{\mathbf{s} \sim \rho_{\pi}, \mathbf{a} \sim \pi_{\theta}(\cdot | \mathbf{s})} [\log(1 - D_{\phi}(\mathbf{s}, \mathbf{a}))] \right)$$

where ρ_{π} is the state distribution induced by the policy π. This framework implicitly learns a reward function through the discriminator, which intrinsically addresses the covariate shift problem better than BC by encouraging the agent to stay on trajectories that look expert-like, rather than just matching a specific action in a specific state.

2.3.3. Addressing Covariate Shift and Hyperdimensional Generalization

To explicitly tackle the covariate shift, which is amplified by the high dimensionality of the molecular state space, we implemented a modified form of Dataset Aggregation (Dagger) for both the BC and GAIL training [23].

● Adaptive Dagger Strategy: After an initial policy training phase (e.g., 50 epochs), the partially trained agent is allowed to execute a small number of autonomous rollouts. For any state $\mathbf{s}'$ encountered during these rollouts that deviates from the expert distribution (e.g., a critical distance CV differs by more than 2σ from the expert mean for that phase of the unbinding), a human expert is queried in a simulated environment to provide the correct action $\mathbf{a}_{\text{expert}}$ for that new, unseen state $\mathbf{s}'$. The new state-action pair $(\mathbf{s}', \mathbf{a}_{\text{expert}})$ is added to the expert dataset $\mathcal{D}_{\text{exp}}$, and the policy is retrained. This iterative process allows the agent to learn the correct recovery behavior for states slightly off the demonstrated path, significantly improving robustness and generalization across the hyperdimensional landscape.

2.4. Evaluation Metrics for Molecular Simulation Success

The performance of the resulting autonomous policies was rigorously evaluated using molecular-specific metrics.

● Task Success Rate: The primary metric. This is the percentage of $N_{\text{test}} = 100$ autonomous agent rollouts that successfully complete the unbinding task within a predefined wall-clock time limit, starting from $\mathbf{s}_0$.

● Work Done (W_{AI}): The non-equilibrium work performed by the AI-agent on the system, calculated as $W_{\text{AI}} = \sum_t \mathbf{F}_{\text{AI}} \cdot \Delta \mathbf{r}_t$, where $\mathbf{F}_{\text{AI}}$ is the force applied by the AI policy and $\Delta \mathbf{r}_t$ is the displacement. This is a measure of the agent's efficiency and provides a lower bound on the free energy barrier overcome [6]. This value was compared directly to the average work done by the human experts (W_{exp}).

● Path Fidelity (RMSD): The Root Mean Square Deviation (RMSD) of the agent's trajectory ($\mathbf{R}_{\text{AI}}$) relative to the closest expert trajectory ($\mathbf{R}_{\text{exp}}$), calculated over the ligand atoms. This metric quantifies how closely the learned policy mimics the subtle, high-dimensional path chosen by the human expert.

● Computational Efficiency: The ratio of wall-clock time required for a successful agent-guided simulation versus the average time required for a human-guided simulation ($T_{\text{AI}} / T_{\text{exp}}$).

3. RESULTS

3.1. Comparison of Imitation Learning Model Performance

The evaluation of the trained policies on the $N_{\text{test}} = 100$ trials revealed significant differences in the ability of the models to successfully navigate the hyperdimensional unbinding pathway autonomously. The Generative Adversarial Imitation Learning (GAIL) model, particularly when augmented with the Dagger-like strategy (GAIL+Dagger), demonstrated a superior capacity for translating human expertise into a robust policy.

Model	Expert Data	Task Success	Average Work	Average Path

	Source	Rate (%)	Done (WAI /Wexp)	RMSD (Ligand, nm)
Unbiased MD	N/A	0%	N/A	N/A
Expert Human	N/A	100%	1.00	N/A
Behavioral Cloning (BC)	$\mathcal{D}_{\text{exp}}$	12%	2.15	0.29
BC + Dagger	$\mathcal{D}_{\text{exp}} + \mathcal{D}_{\text{new}}$	45%	1.48	0.21
GAIL	$\mathcal{D}_{\text{exp}}$	68%	1.22	0.14
GAIL + Dagger	**$\mathcal{D}_{\text{exp}} + \mathcal{D}_{\text{new}}$**	**89%**	**1.05**	**0.11**

The results clearly indicate that the simple Behavioral Cloning (BC) approach was largely ineffective, yielding only a 12% success rate. Autonomously executing the BC policy rapidly resulted in the agent encountering states outside the narrow distribution of $\mathcal{D}_{\text{exp}}$, leading to inappropriate force application and subsequent simulation crash or divergence into a high-energy, chemically irrelevant state. This exemplifies the severe impact of covariate shift in a high-dimensional system where a small displacement can correspond to a substantial energy penalty. The BC agent, on average, also performed 2.15 times the work of the human expert for the few successful trials, suggesting an inefficient and brute-force application of force rather than nuanced, guided steering.

The GAIL framework exhibited a significantly improved success rate of 68%. By implicitly learning a reward function that encouraged the generation of state-action pairs indistinguishable from the expert, the GAIL agent demonstrated a much greater ability to generalize and remain on a physically and chemically plausible trajectory. The work done was also much closer to the expert baseline ($1.22 \times W_{\text{exp}}$), indicating a more efficient policy.

The most successful policy was the GAIL + Dagger implementation, achieving an 89% success rate and a work done ratio of 1.05. The iterative collection and incorporation of out-of-distribution, human-labeled state-action pairs allowed the agent to learn crucial recovery strategies, effectively creating guardrails against divergence in the hyperdimensional space. This robust policy also produced trajectories with the highest fidelity to the expert demonstrations, as indicated by the lowest average Path RMSD (0.11nm), confirming that the agent was not just achieving the goal, but doing so via the expert's learned pathway.

3.2. Analysis of Autonomous Agent Trajectories

Visual and statistical analysis of the successful GAIL + Dagger trajectories provided crucial insight into the learned policies. In over 80% of successful trials, the agent autonomously replicated the expert's initial strategy of a brief, high-force application to overcome the major

van der Waals/electrostatic barrier at the active site entrance, followed by a sustained, lower-magnitude force vector to guide the ligand out of the protein gorge [5, 6].

Specifically, a cluster analysis of the applied force vectors revealed that the AI agent learned to modulate the force direction based on the ligand's orientation within the pocket, a key element of human spatial intuition. When the ligand was observed to rotate in a manner that created a steric clash with a protein residue, the human expert would typically apply a corrective lateral force. The GAIL+Dagger policy learned this nuanced state-dependent force modulation, a complex, high-dimensional control task that simple BC failed to capture.

Furthermore, the autonomous agent trajectories revealed an important finding: the GAIL agent occasionally explored novel, yet physically viable, unbinding pathways that were not explicitly contained within the 50 expert demonstrations. These novel paths were slightly more convoluted but maintained a low potential energy profile, suggesting that the GAIL framework was not simply memorizing the expert's path (as BC tends to do), but rather learning the underlying principle of force application necessary for low-energy navigation. This emergent exploratory behavior hints at the AI agent’s potential to identify novel conformational routes that even experts may overlook, thus enriching the overall sampling of the free energy landscape.

3.3. Efficiency and Free Energy Landscape Exploration

The calculation of the work done (W_{AI}) by the successful AI policies is directly related to the effective free energy barrier ($\Delta G^{\ddagger}$) overcome by the guiding force, in accordance with the Jarzynski equality [32]. The finding that the GAIL+Dagger policy performed work nearly identical to the human expert ($1.05 \times W_{\text{exp}}$) implies that the autonomous policy is performing an equally efficient, near-reversible manipulation.

By comparing the potential energy surface along the reaction coordinate defined by the displacement of the ligand's center of mass, we observed that the application of the AI-guided force vector reduced the effective free energy barrier height by an average of 65% compared to the unperturbed (unbiased) system. The peak of the potential energy barrier for the AI-guided trajectories was consistently lower and smoother than the inherent energy barrier, which is characterized by sharp, high-energy peaks. This quantitative finding validates the core hypothesis: the learned AI policy functions as an automated enhanced sampling tool, efficiently steering the system past regions of high resistance in the hyperdimensional space.

In terms of computational efficiency, the successful agent-guided rollouts were, on average, 3.2 times faster than the human-guided sessions ($T_{\text{AI}} / T_{\text{exp}} = 0.31$). This disparity is due to the inherent cognitive and motor latency associated with human users, where reaction times, controller-to-simulation latency, and decision-making processes inevitably slow down the effective rate of force application. The fully autonomous AI agent, on the other hand, operates at the network-limited speed of the HPC-VR pipeline, enabling rapid, non-stop force application based on real-time state feedback. The combination of high success rate, low work performed, and accelerated wall-clock time positions the IL approach as a highly efficient complement to traditional MD.

4. DISCUSSION

4.1. The AI Agent as an Autonomous "Expert Navigator"

The successful implementation and evaluation of the GAIL+Dagger policy underscore a major advance in the integration of AI and molecular simulation: the ability to distill and automate the spatial and chemical intuition of a human expert. The AI agent, having learned from the high-fidelity iMD-VR datasets, effectively functions as an Autonomous Expert Navigator, capable of executing complex, high-dimensional control tasks within a chemically realistic physical environment.

In essence, the policy $\pi_{\theta}(\mathbf{a} | \mathbf{s})$ learned by the GAIL agent has synthesized the non-linear relationship between a high-dimensional molecular state ($\mathbf{s}$) and the optimal corrective force vector ($\mathbf{a}$) required to maintain a successful, low-work trajectory. This represents a significant conceptual leap. While conventional enhanced sampling methods require the researcher to define the simplified reaction coordinate, the IL-trained agent learns the complex, high-dimensional expert control policy that implicitly and dynamically navigates the free energy landscape. The agent transforms the bottleneck of rare-event sampling from a problem of human-intensive, trial-and-error parameterization into an efficient, autonomous execution of learned expert behavior. This has profound implications for the democratization of complex molecular sampling, as the knowledge gained by a handful of domain experts can be packaged and deployed as a computational tool accessible to a broader scientific community.

4.2. Challenges in Mapping Human Intuition to Agent Policy: The Covariate Shift in Hyperdimensional Space

While the results are highly promising, the path to robust, generalizable policies is fraught with challenges inherent to the high-dimensionality of the problem. The dramatic

failure of simple Behavioral Cloning highlights the most significant obstacle: Covariate Shift [21]. In IL, the trained policy is only guaranteed to perform well on states $\mathbf{s}$ drawn from the expert's observed trajectory distribution, ρ_{exp}. Once the agent executes an action that leads to a state $\mathbf{s}' \notin \rho_{\text{exp}}$, the network's output becomes unreliable, potentially leading to cascading errors.

In the context of molecular dynamics, this problem is severely amplified. A small displacement in a single dihedral angle or a minor fluctuation in a water molecule's position constitutes a new, unseen state in the hyperdimensional feature space. If this displacement leads the ligand closer to a steric clash, the BC policy, having never seen the recovery action for that specific state, may apply an unhelpful or even catastrophic force, resulting in a physically irrelevant, high-energy configuration [1]. The state space is far too vast to be sufficiently covered by a finite number of human demonstrations, no matter how skilled the expert.

The success of the GAIL framework stems from its adversarial structure. The discriminator acts as a continuously evolving implicit reward function, pushing the agent not just to match actions, but to generate entire trajectories that are statistically plausible in the context of the expert data [22]. This trajectory-level learning provides a crucial safeguard against instantaneous state-action failures.

Furthermore, the strategic implementation of the Dagger-like strategy provided the necessary practical solution. By actively querying the expert for corrective actions in the specific regions of state space where the agent's autonomous rollout began to diverge, we effectively created a "safety net" in the policy's decision boundaries. This iterative, on-policy data collection is computationally demanding, but it proved indispensable for achieving the 89% success rate, confirming that robust generalization in the molecular hyperdimensional context requires an active approach to filling the inevitable gaps in the expert data distribution. The reliance on expert human feedback in this crucial iterative step highlights a current limitation of purely autonomous IL and suggests that a continuous Human-AI Collaborative Framework will be the most potent approach moving forward.

4.3. Interpreting the Policy: Toward Explainable AI in Molecular Steering

A deeper challenge lies in understanding why the human expert chose a particular steering strategy and how the AI policy has encoded this intuition. The AI policy is a complex deep neural network, effectively a black box. The human expert is consciously or subconsciously optimizing for several factors simultaneously: minimizing steric clash, exploiting favourable electrostatic or hydrophobic interactions, and applying the minimum necessary force to meet the objective. In chemical terms, they are minimizing work while maximizing the system's proximity to an optimal, low-energy path [7].

The current work confirms that the AI agent is highly effective at mimicking the result (low work done, high path fidelity), but it does not provide direct insight into the underlying chemical principle being optimized. This highlights a critical need for Explainable AI (XAI) in this domain. Future work should focus on methods like Inverse Reinforcement Learning (IRL), which aims to reverse-engineer the expert's implicit reward function from their demonstrations [13, 22].

IRL could potentially extract a simple, human-interpretable function $R(\mathbf{s}, \mathbf{a})$ that the expert was maximizing—for example, a function that penalizes high-force actions and rewards an increase in the distance along a complex, non-linear reaction coordinate. Discovering this latent "chemical utility function" would be a monumental achievement, transforming the IL agent from a mere autonomous executor into a tool for fundamental scientific discovery about the principles of molecular recognition and dynamics. Such a function could then be used to inform new, non-VR enhanced sampling methods, closing the loop between human intuition and autonomous computational discovery. The current research provides the necessary policy and trajectory data for such an IRL investigation, establishing the empirical basis for future XAI-MD research.

4.4. The Integration of VR, AI, and HPC: A New Pipeline for Molecular Discovery

The success of this study establishes a robust, end-to-end pipeline that fundamentally re-imagines the workflow for complex molecular simulation: Expert Intuition $\rightarrow$ iMD-VR Data $\rightarrow$ AI Policy $\rightarrow$ Autonomous Sampling. This integrated approach, which we term Policy-Guided Molecular Dynamics (PGMD), offers transformative advantages over both conventional MD and human-guided iMD-VR alone.

In Drug Discovery, PGMD can accelerate in silico lead optimization [24, 25]. The rapid, autonomous generation of binding/unbinding pathways for numerous lead candidates can provide crucial thermodynamic and kinetic data for virtual screening and pose refinement, going far beyond the limitations of static docking algorithms [26]. By autonomously exploring the entire pathway, the PGMD agent can identify transient, high-energy intermediates or unexpected exit channels, which are often the key determinants of a drug candidate's efficacy or selectivity. The demonstrated application to the SARS-CoV-2 protease confirms its relevance to

systems of immediate pharmacological interest.

In Materials Science, the principles of PGMD are equally applicable. Many critical material properties, such as the self-assembly of polymers, the failure of crystalline structures, or the transport of ions through membranes, are governed by rare-event transitions in hyperdimensional configuration spaces [33]. By having an expert guide an initial structural rearrangement in VR—say, the nucleation of a material phase—the resulting AI policy can then automate the sampling of that nucleation pathway across various chemical compositions or thermodynamic conditions, facilitating the in silico design of materials with targeted properties [15].

The computational efficiency gains—$3.2 \times$ speedup over human interaction—further position PGMD as a practical, scalable technology [28, 29]. The AI policy operates at the speed of the HPC backend, effectively enabling a high-throughput virtual screening of reaction pathways, rather than just static binding affinities.

4.5. The Future Trajectory: Hierarchical Policies and Transfer Learning (Expansion)

4.5.1. The Critical Need for Hierarchical Imitation Learning

The policies developed in this study focus on a flat control strategy, where the agent directly maps the entire state vector ($\mathbf{s}$) to the instantaneous low-level action (the force vector $\mathbf{a}$). While effective for a single, focused task like unbinding, this flat structure becomes computationally unwieldy and non-generalizable for complex, multi-stage molecular processes.

Consider a multi-stage process such as enzyme catalysis, which involves ligand entry, an initial conformational change, a chemical reaction, and product exit. Training a single, flat IL policy to manage the entire sequence from $A \rightarrow Z$ would require an astronomically large and diverse dataset of expert demonstrations. The state space would include not only the ligand and active site but also distant allosteric sites and the transient chemical environment, further exacerbating the covariate shift problem.

The future of PGMD therefore lies in Hierarchical Imitation Learning (HIL) [35]. HIL decomposes the complex task into a hierarchy of sub-policies:

1. High-Level Policy (Manager): This policy learns to map the system's coarse-grained state (e.g., "ligand is near active site," "conformational change in domain X has occurred") to a discrete sub-goal or macro-action (e.g., "drive ligand toward residue A," "induce domain B rotation," or "initiate bond cleavage"). The Manager operates on a longer timescale and a much lower-dimensional, symbolic state space.

2. Low-Level Policy (Worker): This policy takes the macro-action from the Manager as an additional input, along with the detailed molecular state ($\mathbf{s}$), and outputs the continuous, fine-grained force vector ($\mathbf{a}$). The Worker is trained only to achieve its immediate sub-goal.

By decoupling the strategic decision-making (Manager) from the low-level execution (Worker), HIL offers several critical advantages:

- Reduced State-Space Complexity: The Manager tackles the vast strategic space, while the Worker focuses on local, executable force application. This significantly mitigates covariate shift, as the Worker's policy is relevant only for a circumscribed sub-space.

- Improved Interpretability: The Manager's actions (e.g., "induce domain rotation") are human-interpretable, providing a mechanism for XAI by tracking the high-level plan the AI is executing.

- Enhanced Transferability: The Worker policies can be designed to be transferable across molecular systems. For instance, a Worker trained to "drive a hydrophobic group out of a pocket" can be reused as a sub-policy in a completely different protein-ligand system, dramatically reducing the need for new expert demonstrations [36].

The current GAIL+Dagger framework provides the foundation for the Worker policy, which excels at local, continuous force application. Future work will involve defining the Manager's state space via advanced dimensionality reduction techniques (e.g., time-lagged independent component analysis, tICA) to capture the collective, long-timescale variables that govern the molecular process.

4.5.2. The Promise of Transfer Learning and Policy Fine-Tuning

The greatest potential for scale and impact lies in Transfer Learning (TL), which is the ability to adapt a policy trained on one molecular system (the source task) to a new, chemically similar system (the target task) with minimal new expert data [37]. In PGMD, we anticipate two primary modes of transfer:

1. Policy Fine-Tuning for Analogous Systems: A policy trained on the SARS-CoV-2 Main Protease (the source task) could be fine-tuned to unbind a structurally similar ligand from a homologous protease in a different virus (the target task). The initial layers of the deep neural network policy, which learn general principles of force

application, can be preserved, and only the final layers—which map to the specific geometric CVs of the target system—need to be retrained using a small handful of new iMD-VR demonstrations.

2. Generalization of Sub-Policies: As proposed in the HIL architecture, certain Worker policies are generalizable. The policy for "driving a charged group toward a solvent-exposed region" or "inducing a hinge-bending motion" relies on general principles of electrostatics and mechanics that transcend a single protein system. By compiling a library of molecular sub-policies from diverse iMD-VR demonstrations, future PGMD agents could synthesize entirely new, complex policies for novel systems by stringing together these pre-trained, transferable components. This would be analogous to a molecular dynamics macro-language where high-level commands (macro-actions) are executed by reliably trained low-level policies (force-vector generation).

Realizing this potential for transfer learning requires dedicated research into representation learning for molecular states—developing state representations ($\mathbf{s}$) that are maximally invariant to translational/rotational changes and maximally variant to chemically relevant changes. Graph Neural Networks (GNNs) on the molecular graph are highly promising architectures for encoding such chemically relevant, transferrable state features [38].

The current work, by establishing the high fidelity and efficiency of the GAIL+Dagger policy, provides the necessary proof-of-concept for this entire trajectory. The challenge now shifts from can we imitate the expert? to how can we generalize the expert's knowledge to the entire molecular universe? The integration of iMD-VR, IL, HIL, and TL is poised to answer this question, accelerating the pace of discovery in a way that was previously unimaginable [39, 40].

4.6. LIMITATIONS AND CONCLUSION

The current study, while demonstrating a high-success-rate autonomous policy, is subject to inherent limitations. The primary constraint remains the dependence on high-quality expert demonstrations. The performance of the GAIL policy is directly proportional to the expertise and diversity encoded in the initial 50 trajectories. Furthermore, the Dagger-like correction mechanism, while effective, still requires human input during the refinement process, indicating a partial rather than a complete autonomy. Finally, the feature-engineered CVs, while effective, impose a structural bias on the learned policy; a more generalizable approach may require end-to-end learning directly from raw atomic coordinates, which remains a significant computational hurdle [41].

In conclusion, this research successfully bridges the fields of immersive visualization, high-performance computing, and Artificial Intelligence. We have demonstrated that advanced Imitation Learning strategies, specifically GAIL augmented with an adaptive Dagger mechanism, can effectively translate the complex, high-dimensional intuition of a human expert from an interactive VR environment into a robust, autonomous AI policy. This Policy-Guided Molecular Dynamics (PGMD) approach yields autonomous policies that perform near-expert work with significant computational speedup, directly tackling the hyperdimensional sampling problem that has long plagued molecular dynamics. This new methodology paves a promising pathway for the autonomous exploration of molecular free energy landscapes, marking a transformative moment for computational drug discovery and materials design.

REFERENCES

1. Saunders WR, Grant J, Müller EH. A domain specific language for performance portable molecular dynamics algorithms. Comput Phys Commun. 2018;224:119–35. https://doi.org/10.1016/j.cpc.2017.11.006.

2. Walters RK, Gale EM, Barnoud J, Glowacki DR, Mulholland AJ. The emerging potential of interactive virtual reality in drug discovery. Expert Opin Drug Discov. 2022;17:685–98. https://doi.org/10.1080/17460441.2022.2079632.

3. O'Connor MB, Bennie SJ, Deeks HM, Jamieson-Binnie A, Jones AJ, Shannon RJ, et al. Interactive molecular dynamics in virtual reality from quantum chemistry to drug binding: An open-source multi-person framework. J Chem Phys. 2019;150(22):220901. https://doi.org/10.1063/1.5092590.

4. Deeks HM, Walters RK, Hare SR, O'Connor MB, Mulholland AJ, Glowacki DR. Interactive molecular dynamics in virtual reality for accurate flexible protein-ligand docking. PLoS One. 2020;15(3):1–21. https://doi.org/10.1371/journal.pone.0228461.

5. Deeks HM, Walters RK, Barnoud J, Glowacki D, Mulholland A. Interactive molecular dynamics in virtual reality is an effective tool for flexible substrate and inhibitor docking to the SARS-CoV-2 main protease. J Chem Inf Model. 2020. https://doi.org/10.1021/acs.jcim.0c01030.

6. Deeks HM, Zinovjev K, Barnoud J, Mulholland AJ, van der Kamp MW, Glowacki DR. Free energy along drug-protein binding pathways interactively sampled in virtual reality. Sci Rep. 2023;13:16665. https://doi.org/10.1038/s41598-023-43523-x.

7. Shannon RJ, Deeks HM, Burfoot E, Clark E, Jones AJ, Mulholland AJ, et al. Exploring human-guided strategies for reaction network exploration: interactive molecular dynamics in virtual reality as a tool for citizen scientists. J Chem Phys. 2021;155:154106. https://doi.org/10.1063/5.0062517.

8. Zheng B, Verma S, Zhou J, Tsang IW, Chen F. Imitation learning: Progress, taxonomies and challenges. IEEE Trans Neural Netwo Learn Syst. 2024;35(5):6322–37. https://doi.org/10.1109/TNNLS.2022.3213246.

9. Samantapudi, R. K. R. (2025). Enhancing search and recommendation personalization through user modeling and representation. International Journal of Computational and Experimental Science and Engineering, 11(3), 6246–6265. https://doi.org/10.22399/ijcesen.3784

10. Hussein A, Gaber MM, Elyan E, Jayne C. Imitation learning: a survey of learning methods. ACM Comput Surv. 2017. https://doi.org/10.1145/3054912.

11. Gavenski N, Meneguzzi F, Luck M, Rodrigues O. A survey of imitation learning methods, environments and metrics 2024. arXiv:2404.19456.

12. Schaal S. Is imitation learning the route to humanoid robots? Trends Cogn Sci. 1999;3:233–42. https://doi.org/10.1016/S1364-6613(99)01327-3.

13. Reggia JA, Katz GE, Davis GP. Humanoid cognitive robots that learn by imitating: implications for consciousness studies. Front Robot AI. 2018;5:1. https://doi.org/10.3389/frobt.2018.00001.

14. Hua J, Zeng L, Li G, Ju Z. Learning for a robot: deep reinforcement learning, imitation learning, transfer learning. Sensors. 2021;21:1278. https://doi.org/10.3390/s21041278.

15. Zare M, Kebria PM, Khosravi A, Nahavandi S. A survey of imitation learning: algorithms, recent developments, and challenges. IEEE Trans Cybern. 2023. https://doi.org/10.1109/TCYB.2024.3395626.

16. Leinen P, Esders M, Schütt KT, Wagner C, Müller KR, Tautz FS. Autonomous robotic nanofabrication with reinforcement learning. Sci Adv. 2020;6(36):eabb6987. https://doi.org/10.1126/sciadv.abb6987.

17. Ai C, Yang H, Liu X, Dong R, Ding Y, Guo F. Mtmol-gpt: de novo multi-target molecular generation with transformer-based generative adversarial imitation learning. PLoS Comput Biol. 2024;20(6):1–23. https://doi.org/10.1371/journal.pcbi.1012229.

18. Chadha, K. S. (2025). Zero-Trust Data Architecture for Multi-Hospital Research: HIPAA-Compliant Unification of EHRs, Wearable Streams, and Clinical Trial Analytics. International Journal of Computational and Experimental Science and Engineering, 11(3). https://doi.org/10.22399/ijcesen.3477

19. Jia X, Blessing D, Jiang X, Reuss M, Donat A, Lioutikov R, Neumann G. Towards diverse behaviors: a benchmark for imitation learning with human demonstrations. In: The Twelfth International Conference on Learning Representations. 2024. arXiv:2402.14606.

20. Maadi M, Akbarzadeh Khorshidi H, Aickelin U. A review on human–AI interaction in machine learning and insights for medical applications. Int J Environ Res Public Health. 2021;18(4):2121. https://doi.org/10.3390/ijerph18042121.

21. Webb ME, Fluck A, Magenheim J, Malyn-Smith J, Waters J, Deschênes M, et al. Machine learning for human learners: opportunities, issues, tensions and threats. Educ Technol Res Dev. 2021;69:2109–30. https://doi.org/10.1007/s11423-020-09858-2.

22. Jung E, Kim I. Hybrid imitation learning framework for robotic manipulation tasks. Sensors. 2021. https://doi.org/10.3390/s21103409.

23. Seritan S, Wang Y, Ford JE, Valentini A, Gold T, Martínez TJ. Interachem: virtual reality visualizer for reactive interactive molecular dynamics. J Chem Educ. 2021. https://doi.org/10.1021/acs.jchemed.1c00654.

24. Doutreligne S, Gageat C, Cragnolini T, Taly A, Pasquali S, Derreumaux P, Baaden M. UnityMol: interactive and ludic visual manipulation of coarse-grained RNA and other biomolecules. In: 2015 IEEE 1st international workshop on virtual and augmented reality for molecular science (VARMS@IEEEVR); 2015. p. 1–6. https://doi.org/10.1109/VARMS.2015.7151718.

25. Bennie SJ, Maritan M, Gast J, Loschen M, Gruffat D, Bartolotta R, Hessenauer S, Leija E, McCloskey S. A virtual and mixed reality platform for molecular design and drug discovery—Nanome Version 1.24. In: Byška J, Krone M, Sommer B editors. Workshop on molecular graphics and visual analysis of molecular data. The Eurographics Association; 2023. https://doi.org/10.2312/molva.20231114.

26. Kneller DW, Li H, Galanie S, Phillips G, Labbé A,

Weiss KL, et al. Structural, electronic, and electrostatic determinants for inhibitor binding to subsites s1 and s2 in sars-cov-2 main protease. J Med Chem. 2021;64:17366–83. https://doi.org/10.1021/ACS.JMEDCHEM.1C01475

27. Cassidy KC, Šefčík J, Raghav Y, Chang A, Durrant JD. Proteinvr: web-based molecular visualization in virtual reality. PLoS Comput Biol. 2020;16(3):1–17. https://doi.org/10.1371/journal.pcbi.1007747

28. Norrby M, Grebner C, Eriksson J, Boström J. Molecular rift: virtual reality for drug designers. J Chem Inf Model. 2015;55(11):2475–84. https://doi.org/10.1021/acs.jcim.5b00544

29. Crossley-Lewis J, Dunn J, Buda C, Sunley GJ, Elena AM, Todorov IT, et al. Interactive molecular dynamics in virtual reality for modelling materials and catalysts. J Mol Graph Model. 2023;125:108606. https://doi.org/10.1016/j.jmgm.2023.108606.

30. Srilatha, S. (2025). Integrating AI into enterprise content management systems: A roadmap for intelligent automation. Journal of Information Systems Engineering and Management, 10(45s), 672–688. https://doi.org/10.52783/jisem.v10i45s.8904

31. Stroud HJ, Wonnacott MD, Barnoud J, Roebuck Williams R, Dhouioui M, McSloy A, et al. NanoVer server: a python package for serving real-time multi-user interactive molecular dynamics in virtual reality. J Open Sour Softw. 2025;10(110):8118. https://doi.org/10.21105/joss.08118.

32. Jamieson-Binnie AD, O'Connor MB, Barnoud J, Wonnacott MD, Bennie SJ, Glowacki DR. Narupa iMD: a VR-enabled multiplayer framework for streaming interactive molecular simulations. In: ACM SIGGRAPH 2020 immersive pavilion, SIGGRAPH '20. New York: Association for Computing Machinery; 2020. https://doi.org/10.1145/3388536.3407891.

33. Gowers RJ, Linke M, Barnoud J, Reddy TJE, Melo MN, Seyler SL, Domański J, Dotson, Sébastien Buchoux DL, Kenney IM, Beckstein O. MDAnalysis: a python package for the rapid analysis of molecular dynamics simulations. In: Sebastian B, Scott R editors. Proceedings of the 15th Python in science conference. 2016. p. 98–105. https://doi.org/10.25080/Majora-629e541a-00e.

34. Michaud-Agrawal N, Denning EJ, Woolf TB, Beckstein O. Mdanalysis: a toolkit for the analysis of molecular dynamics simulations. J Comput Chem. 2011;32(10):2319–27. https://doi.org/10.1002/jcc.21787.

35. Kalra A, Hummer G, Garde S. Methane partitioning and transport in hydrated carbon nanotubes. J Phys Chem B. 2004;108(2):544–9. https://doi.org/10.1021/jp035828x.

36. Correia A, Alexandre LA. A survey of demonstration learning. Robot Auton Syst. 2024;182:104812. https://doi.org/10.1016/j.robot.2024.104812.

37. Bui TV, Mai TA, Nguyen TH. Inverse factorized soft Q-learning for cooperative multi-agent imitation learning. In: The thirty-eighth annual conference on neural information processing systems. 2024. https://openreview.net/forum?id=xrbgXJomJp.

38. Bui TV, Mai T, Nguyen TH. Inverse factorized q-learning for cooperative multi-agent imitation learning. 2023. arXiv:2310.06801.

39. Ellis B, Cook J, Moalla S, Samvelyan M, Sun M, Mahajan A, Foerster JN, Whiteson S. SMACv2: an improved benchmark for cooperative multi-agent reinforcement learning. In: Proceedings of the 37th international conference on neural information processing systems, NIPS '23. Red Hook: Curran Associates Inc.; 2024. https://doi.org/10.5555/3666122.3667756.

40. Rangu, S. (2025). Analyzing the impact of AI-powered call center automation on operational efficiency in healthcare. Journal of Information Systems Engineering and Management, 10(45s), 666–689. https://doi.org/10.55278/jisem.2025.10.45s.666

41. FPT. Fpt reinforcement learning competition. 2020. https://codelearn.io/game/detail/2212875#ai-game-summary.

42. Paine TL, Gulcehre C, Shahriari B, Denil M, Hoffman M, Soyer H, Tanburn R, Kapturowski S, Rabinowitz N, Williams D, et al. Scaling laws for imitation learning in single-agent games. 2023. arXiv preprint arXiv:2301.13314.

43. Küttler H, Nardelli N, Miller AH, Raileanu R, Selvatici M, Grefenstette E, Rocktäschel T. The nethack learning environment. CoRR. 2020. arXiv:2006.13760.

44. Younes M, Kijak E, Kulpa R, Malinowski S, Multon F. Maaip: multi-agent adversarial interaction priors for imitation from fighting demonstrations for physics-based characters. Proc ACM Comput Graph Interact Tech. 2023. https://doi.org/10.1145/3606926.

45. Ravichandar H, Polydoros AS, Chernova S, Billard A. Recent advances in robot learning from

demonstration. Ann Rev Control Robot Auton Syst. 2020;3(1):297–330. https://doi.org/10.1146/annurev-control-100819-063206.

46. Zhu Y, Joshi A, Stone P, Zhu Y. Viola: imitation learning for vision-based manipulation with object proposal priors. Proc Mach Learn Res. 2022;205:1199–210. https://doi.org/10.48550/arXiv.2210.11339.

47. Seo M, Gupta R, Zhu Y, Skoutnev A, Sentis L, Zhu Y. Learning to walk by steering: perceptive quadrupedal locomotion in dynamic environments. In: 2023 IEEE international conference on robotics and automation (ICRA). 2023. p. 5099–105. https://doi.org/10.1109/ICRA48891.2023.10161302 .

48. Mehta SA, Losey DP. Unified learning from demonstrations, corrections, and preferences during physical human–robot interaction. J Hum Robot Interact. 2023. https://doi.org/10.1145/3623384.

49. Pomerleau DA. Efficient training of artificial neural networks for autonomous navigation. Neural Comput. 1991;3:88–97. https://doi.org/10.1162/NECO.1991.3.1.88.

50. Sammut C. Behavioral cloning. Encycl Mach Learn. 2011. https://doi.org/10.1007/978-0-387-30164-8_69.

51. Russell S. Learning agents for uncertain environments (extended abstract). In: Proceedings of the eleventh annual conference on computational learning theory, COLT' 98. New York: Association for Computing Machinery; 1998. p. 101–03. https://doi.org/10.1145/279943.279964.

52. Ng AY, Russell SJ. Algorithms for inverse reinforcement learning. In: Proceedings of the seventeenth international conference on machine learning, ICML '00. San Francisco: Morgan Kaufmann Publishers Inc.; 2000. p. 663–70. https://doi.org/10.5555/645529.657801.

53. Ziebart BD, Maas A, Bagnell JA, Dey AK. Maximum entropy inverse reinforcement learning. In: Proceedings of the 23rd National conference on artificial intelligence—volume 3, AAAI'08. AAAI Press; 2008. p. 1433–38. https://doi.org/10.5555/1625275.1625692.

54. Ramachandran D, Amir E. Bayesian inverse reinforcement learning. In: Proceedings of the 20th international joint conference on artifical intelligence, IJCAI'07. San Francisco: Morgan Kaufmann Publishers Inc.; 2007. p. 2586–91. https://doi.org/10.5555/1625275.1625692.

55. Metelli AM, Ramponi G, Concetti A, Restelli M. Provably efficient learning of transferable rewards. In: Meila M, Zhang T editors. Proceedings of the 38th international conference on machine learning, proceedings of machine learning research, vol. 139. PMLR; 2021. p. 7665–76. https://proceedings.mlr.press/v139/metelli21a.html.

56. Deka A, Liu C, Sycara KP. ARC—Actor residual critic for adversarial imitation learning. In: Liu K, Kulic D, Ichnowski J, editors. Proceedings of the 6th conference on robot learning, proceedings of machine learning research, vol. 205. PMLR; 2023. p. 1446–56. https://proceedings.mlr.press/v205/deka23a.html.

57. Ho J, Ermon S. Generative adversarial imitation learning. In: Proceedings of the 30th international conference on neural information processing systems, NIPS'16. Red Hook: Curran Associates Inc.; 2016. p. 4572–80. https://doi.org/10.5555/3157382.3157608.

58. Schulman J, Levine S, Abbeel P, Jordan M, Moritz P. Trust region policy optimization. In: Bach F, Blei D editors. Proceedings of the 32nd international conference on machine learning, proceedings of machine learning research, vol. 37. Lille: PMLR; 2015. p. 1889–97. https://proceedings.mlr.press/v37/schulman15.html.

59. Pomerleau DA. ALVINN: an autonomous land vehicle in a neural network. In: Proceedings of the 2nd international conference on neural information processing systems, NIPS'88. Cambridge: MIT Press; 1988. p. 305–13. https://doi.org/10.5555/2969735.2969771.

60. de Haan P, Jayaraman D, Levine S. Causal confusion in imitation learning. In: Proceedings of the 33rd international conference on neural information processing systems. Red Hook: Curran Associates Inc.; 2019.https://doi.org/10.5555/3454287.3455336