BIOMOLECULAR AGENT DIFFERENTIAL GAMES FOR MODELING PROTEIN DYNAMICS

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein dynamics characterize structural fluctuations and transitions among conformational states relevant to biomolecular function, complementing static structural descriptions. Such dynamics involve heterogeneous residue motions and inter-residue couplings within the evolving protein structure. However, many learning-based approaches model a protein as a single dynamical object, leaving these residue-wise dynamics implicit. We propose *Biomolecular Agent Differential Games*, a centralized differential game that coordinates biomolecular agents through local feedback policies. We characterize the protein residue coalition by a cooperative Nash equilibrium and recast the resulting game as a centralized stochastic control problem. Its optimality is characterized by the Hamilton–Jacobi–Bellman (HJB) equation and an equivalent forward–backward stochastic differential equation (FBSDE). We optimize the policy profile through fictitious play, where each agent assimilates the central planner proposal according to its algorithmic Nash gap. We extend the framework to protein–ligand dynamics by formulating the joint dynamics of the residue coalition and a distinct ligand agent as a non-cooperative game. Across diverse biomolecular systems, our framework achieves improvements over baselines in protein dynamics modeling and ligand–pocket interaction fidelity. These results indicate that biomolecular agent coordination provides a principled framework for accurate and faithful biomolecular dynamics modeling.

1 INTRODUCTION

Biomolecular dynamics describes how atomic-scale fluctuations give rise to conformational changes over time, linking microscopic motions to biomolecular function. Because these fluctuations can propagate through the molecular structure, understanding protein function requires a dynamical perspective that goes beyond static conformations (Henzler-Wildman et al., 2007; Popovych et al., 2006). A standard physical framework for this evolution is *underdamped Langevin dynamics* (Leimkuhler & Matthews, 2015; Pavliotis, 2014), in which deterministic molecular forces are coupled with frictional dissipation and stochastic thermal fluctuations from the heat bath (Bussi & Parrinello, 2007). This formulation reflects that biomolecules continuously exchange energy with their surrounding environment, making stochastic trajectories a natural representation of protein motion. Recent learning-based methods have been proposed for modeling protein conformational variation and dynamics (Jing et al., 2024a;b; Yu et al., 2025; Sengar et al., 2026; Yu et al., 2026). These approaches typically treat the protein as a single and globally dynamical object, either in the original molecular state space or in a learned latent representation. Although this holistic view can recover overall conformational evolution, the temporal development of residue-level fluctuations and inter-residue couplings remains unresolved, even though these localized interactions are often central to physically faithful biomolecular dynamics.

We address this limitation by introducing *biomolecular agents* formulation for biomolecular dynamics. Rather than modeling the protein with a single monolithic transition rule, we represent it as a collection of interacting residue-level biomolecular agents, each associated with a local feedback policy. This view provides a structured parameterization of high-dimensional stochastic dynamics, in which heterogeneous local motions and inter-residue couplings are modeled explicitly rather than absorbed into a single global evolution map. A central planner coordinates these biomolecular agents through a shared protein-level objective (Nisan et al., 2007), ensuring that local updates remain aligned with the dynamics of the full molecular system. The resulting framework combines modular local control

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

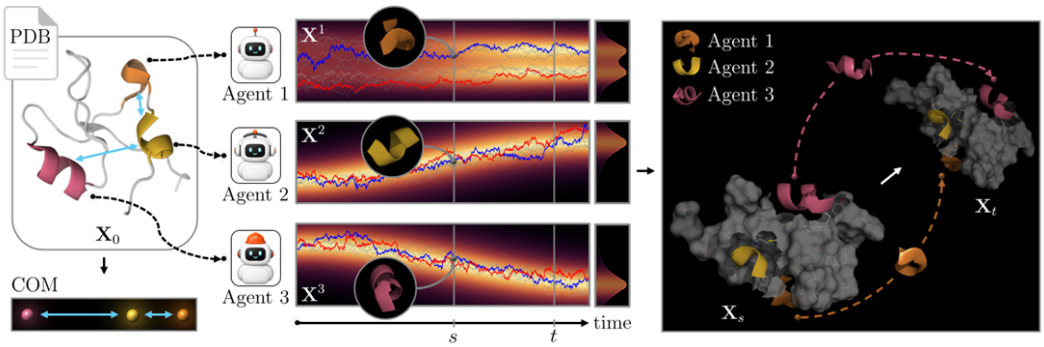

Figure 1: **Overall pipeline.** *(Left)* We partition the protein complex into molecular blocks, assign blocks to agents, and model pairwise couplings using inter-block COM distances. Initial state $\mathbf{X}_0$ comes from a PDB structure. *(Middle)* Over time, each agent uses a feedback policy to control its block dynamics within the coupled system, with reference (red) and controlled dynamics (blue). *(Right)* A central planner coordinates the agent coalition under a shared objective to integrate block predictions and model protein dynamics from $\mathbf{X}_s$ to $\mathbf{X}_t$.

with global dynamical consistency and provides a *divide-and-conquer* formulation for learning biomolecular dynamics under underdamped Langevin dynamics. From this perspective, the model is naturally formulated as a centralized **differential game** (Isaacs, 1999), in which biomolecular agents form a coalition under a common objective. Overall, the main contribution of this work is two-fold.

1. **Residue Agent Decomposition for Molecular Dynamics.** We propose *Biomolecular Agents* as a new modeling principle for molecular dynamics, replacing the conventional single-system formulation with a centralized cooperative differential game over residue-level agents. Each Biomolecular Agent represents local residue dynamics, while a shared protein-level objective couples the agents into a coherent joint system that recovers full molecular evolution.
2. **Agent Coalition Nash Equilibrium.** We develop a centralized cooperative learning framework that coordinates residue-level feedback policies under a shared protein-level objective. The resulting residue coalition captures residue-wise motions and couplings while preserving global protein evolution. We optimize this residue-agent system via fictitious play toward a cooperative Nash equilibrium and analyze its learning dynamics.

2 Game-theoretic Formulation for Protein Complex Dynamics

Problem Formulation. In this paper, we formulate biomolecular dynamics modeling as a centralized differential game over an agnet-level coalition. Let $\mathbf{x}_t$ denote the protein-level reference generated by underdamped Langevin dynamics (Leimkuhler & Matthews, 2015; Pavliotis, 2014):

$$\textbf{(Red)} \qquad d\mathbf{x}_t = \mathbf{M}^{-1}\mathbf{P}_t dt, \quad d\mathbf{P}_t = -\nabla\mathcal{E}(\mathbf{x}_t)dt - \gamma\mathbf{P}_t dt + \sqrt{2\gamma k_B \mathfrak{T}\mathbf{M}}d\mathbf{W}_t, \tag{1}$$

where $\mathbf{P}_t$ denotes the conjugate momenta and V denotes the force field, and quadruplet $(\gamma, k_B, \mathfrak{T}, \mathbf{M})$ is thermodynamic environmental variables. As mentioned above, our goal is to represent it through coordinated residue-level dynamics rather than modeling it as a single global dynamical object.

Let the protein complex contain M atoms, and let $\{\mathcal{I}_i\}_{i=1}^N$ be a residue-level or fragment-level partition of the atom indices. Each block $\mathcal{I}_i$ is assigned to a *biomolecular agent* $i \in \mathcal{C} := \{1, \dots, N\}$, whose local state $\mathbf{X}_t^i$ contains the molecular variables of that block, including positions and momenta of a single residue. The agent states are assembled as $\mathbf{X}_t = \bigoplus_{i=1}^N \mathbf{X}_t^i \in \mathbb{R}^d$, where $d = \sum_{i=1}^N d_i$. The objective is to learn the local agent dynamics such that their joint trajectory reproduces the Langevin reference trajectory $\mathbf{x}_t$ at the full protein-complex level.

Agent Coalition Biomolecular Dynamics. We associate the agent coalition with a joint control profile $\mathbf{u} = (\mathbf{u}^1, \dots, \mathbf{u}^N) \in \mathrm{U}_1 \times \cdots \times \mathrm{U}_N := \mathrm{U}$, where $\mathrm{U}_i \subset \mathcal{H}_i$ is an admissible subset of a complete Hilbert space $\mathcal{H}_i$ of biomolecular-level control functions. The i-th component is generated by a local agent through a parameterized Markovian feedback policy $\mathbf{u}^i = \alpha_i(\cdot, \cdot; \theta^i) \in \mathrm{U}_i$, where $\theta^i \in \Theta_i \subseteq \mathbb{R}^p$ denotes the trainable neural network parameter of agent i. At time t, the realized control is evaluated as $\mathbf{u}_t^i = \alpha_i(t, \mathbf{X}_t; \theta^i)$. The biomolecular-block dynamics and the full protein

complex dynamics are written compactly through the block-assembly operator as

$$\textbf{(Blue)} \qquad d\mathbf{X}_t := \bigoplus_{i=1}^{N} d\mathbf{X}_t^i = \mathrm{B}(t, \mathbf{X}_t, \mathbf{u}_t)dt + \sqrt{2\gamma k_B \mathfrak{T}\mathbf{M}}d\mathbf{W}_t, \quad \mathbf{X}_0 \sim \mathbb{P}_{\text{init}} \propto e^{-\mathcal{E}_{\text{init}}}, \quad (2)$$

where $\mathrm{B} = \bigoplus_{i=1}^{N} \mathrm{B}_i$ is the assembled drift, $\mathbf{W}_t = \bigoplus_{i=1}^{N} \mathbf{W}_t^i$ denotes the concatenated Brownian motion, and $\mathbb{P}_{\text{init}}$ is a probability measure over initial configuration. To mimic residue-level interactions within the protein complex, we parameterize the residue-block drift as

$$\mathrm{B}_i(t, \mathbf{X}_t, \mathbf{u}_t) = \Phi_i(t, \mathbf{X}_t^i, \mathbf{u}_t^i) + \sum_{j \neq i} \exp\left(-\gamma d^2_{\mathbb{R}^3}(\mathbb{E}_{\nu_i}[\mathbf{X}_t^i], \mathbb{E}_{\nu_j}[\mathbf{X}_t^j])\right) \Psi_{ij}(t, \mathbf{X}_t^i, \mathbf{X}_t^j, \mathbf{u}_t^i, \mathbf{u}_t^j), \quad (3)$$

where $\nu_i \sim \mathrm{Law}(\mathbf{X}^i)$ denotes the mass probability measure on agent block $\mathcal{I}_i$, thus $\mathbb{E}_{\nu_i}[\mathbf{X}_t^i]$ represents the agent-level *center of mass* (COM) at time t. The exponential factor therefore defines a distance-decay interaction kernel on $\mathbb{R}^3$, assigning larger coupling weights to agent pairs with nearby mass centers and exponentially suppressing long-range interactions. The parameter $\gamma > 0$ controls the rate of this spatial decay. The term Φ_i is a learned local component for unit block $\mathcal{I}_i$, while Ψ_{ij} provides a learned interaction component through which agent j modulates the evolution of agent i. This decomposition serves as a biomolecular inductive bias for constructing full biomolecular dynamics from coupled residue-level stochastic dynamics. In this setting, all agents are optimized under the same global molecular information filtration $\mathcal{F}_t = \sigma(\mathbf{X}_s, \mathbf{u}_s^1, \ldots, \mathbf{u}_s^N : 0 \leq s \leq t)$. Thus, each admissible policy is a centralized Markovian feedback policy $\mathbf{u}_t^i = \alpha_i(t, \mathbf{X}_t; \theta^i)$, with $\alpha_i \in \mathcal{A}_i$.

Objective Function. For a policy profile of N biomolecular agents $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_N)$, let $\mathbf{u}_s^{\boldsymbol{\alpha}} = (\alpha_1(s, \mathbf{X}_s^{\boldsymbol{\alpha}}; \theta^1), \ldots, \alpha_N(s, \mathbf{X}_s^{\boldsymbol{\alpha}}; \theta^N))$ denote the induced control profile, and let $\mathbf{X}_s^{\boldsymbol{\alpha}}$ be the corresponding controlled agent coalition trajectory. Then, we estimate a trajectory-matching running loss (*i.e.*, ℓ) and a terminal negative log-likelihood (*i.e.*, g_T) as a state-time conditioned *cost functional* given as follows:

$$\mathcal{J}_x(\boldsymbol{\alpha}) := \mathbb{E}\left[\int_t^T \underbrace{\left(\|\mathbf{X}_s^{\boldsymbol{\alpha}} - \mathbf{x}_s\|_2^2 + \lambda_{\mathrm{u}}\|\mathbf{u}_s^{\boldsymbol{\alpha}}\|_2^2\right)}_{\ell(s, \mathbf{X}_s^{\boldsymbol{\alpha}}, \mathbf{u}_s^{\boldsymbol{\alpha}}; \mathbf{x}_s)} ds + \underbrace{\left(-\lambda \log \pi(\mathbf{X}_T^{\boldsymbol{\alpha}})\right)}_{g_T(\mathbf{X}_T^{\boldsymbol{\alpha}})} \,\middle|\, \mathbf{X}_t^{\boldsymbol{\alpha}} = x\right], \quad (4)$$

where the optimal behavior of the agent coalition is characterized by the *value function* $\mathcal{V}(t, x) = \inf_{\boldsymbol{\alpha} \in \mathcal{A}} \mathcal{J}_x(\boldsymbol{\alpha})$, with $\mathcal{A} = \mathcal{A}_1 \times \cdots \times \mathcal{A}_N$, and $\pi(x) = \mathscr{N}_{\mathcal{E},\mathfrak{T}}^{-1} \exp\left(-\mathcal{E}(x)/(k_B\mathfrak{T})\right)$ denotes the *Gibbs density* associated with the molecular energy $\mathcal{E}$, so the terminal term favors thermodynamically plausible low-energy states. Equivalently, one may rewrite $g_T(\mathbf{X}_T) = \lambda\mathcal{E}(\mathbf{X}_T)/(k_B\mathfrak{T}) + \lambda \log \mathscr{N}_{\mathcal{E},\mathfrak{T}}$, where the additive constant $\lambda \log \mathscr{N}_{\mathcal{E},\mathfrak{T}}$ can be ignored during optimization.

The agent-wise construction assigns a local dynamical module to each agent, but the purpose of these local modules is to jointly reproduce the physically consistent evolution of the protein complex. Since the motion of one agent can influence the global conformation, independently optimized local updates may not compose into a reliable molecular trajectory. We therefore model the agent coalition as ***differential games*** (Basar & Olsder, 1999), in which a central planner coordinates agents through a single objective and seeks a joint policy that minimizes the protein-level cost. The resulting solution is characterized by a *Nash equilibrium*:

Definition 2.1. *We say that a policy profile $\boldsymbol{\alpha}^\star = (\alpha_1^\star, \ldots, \alpha_N^\star) \in \mathcal{A}$ is a **Nash equilibrium** if no agent can lower the shared cost $\mathcal{J}_x$ by deviating from its own policy alone:*

$$(\mathbf{NE}) \qquad \mathcal{J}_x(\alpha_i^\star, \boldsymbol{\alpha}_{-i}^\star) \leq \mathcal{J}_x(\alpha_i, \boldsymbol{\alpha}_{-i}^\star), \quad \forall \alpha_i \in \mathcal{A}_i, \quad i = 1, \ldots, N, \quad (5)$$

where $\boldsymbol{\alpha}_{-i}^\star = (\alpha_j^\star)_{j \neq i}$ denotes the policies of all agents other than i.

Definition 2.1 states a protein condition for the residue agent coalition. At a Nash equilibrium, each residue policy is aligned with the surrounding interaction pattern induced by $\alpha_{-i}^\star$. However, directly searching for this joint policy across the infinite-dimensional space of joint policies $\alpha \in \mathcal{A}$ is computationally intractable since the coupling of molecular dynamics across agents intertwines the cost evaluation under any anomaly (Laine et al., 2023). In Section 3, we resolve this difficulty by reformulating the residue coalition game as a centralized stochastic optimal-control problem, allowing us to identify the team optimal profile through a computable equilibrium condition.

3 AGENT COALITION EQUILIBRIUM FOR PROTEIN DYNAMICS

In previous section, we formulated protein dynamics modeling as a centralized differential game and introduced the corresponding Nash equilibrium. The remaining challenge is to characterize this equilibrium in a form that is analytically tractable and computationally implementable. To this end, we recast the agent coalition game as a stochastic optimal-control problem under the shared protein-level objective, so that the value function $\mathcal{V}$ can be characterized through a computable equilibrium condition rather than by a direct search over the infinite-dimensional joint policy space of joint policies $\boldsymbol{\alpha} \in \mathcal{A}$. The associated Hamiltonian $\mathscr{H}$ is then given by:

$$\mathscr{H}(t, x, \mathbf{u}, p, q) = \ell(t, x, \mathbf{u}; \mathbf{x}_t) + p^\top \mathrm{B}(t, x, \mathbf{u}) + \frac{1}{2} \operatorname{Tr}\left(\Sigma\Sigma^\top q\right), \;\; \Sigma = \mathbf{I}_d \sqrt{2\gamma k_B T \mathbf{M}}. \quad (6)$$

Under the controlled residue dynamics, value function $\mathcal{V}$ formally satisfies the following *Hamilton-Jacobi-Bellman* (HJB) equation:

$$\partial_t \mathcal{V}(t, x) + \inf_{\mathbf{u} \in \mathrm{U}} \mathscr{H}\left(t, x, \mathbf{u}, \nabla_x \mathcal{V}(t, x), \nabla_x^2 \mathcal{V}(t, x)\right) = 0, \quad \mathcal{V}(T, x) = g_T(x). \quad (7)$$

Although the HJB characterization in Eq. 7 provides a formal optimality condition, directly solving this PDE is impractical for high-dimensional protein states. In particular, the grid-based representation of $\mathcal{V}(t, x)$ and its derivatives suffers from the *curse of dimensionality* and is not well aligned with GPU-parallel tensor computation. We therefore turn to the equivalent stochastic representation of the centralized control problem, in which the value function is characterized along sampled agent coalition trajectories through *forward–backward stochastic differential equations* (FBSDEs):

Definition 3.1. *Let $\mathbf{u}(t, x)$ denote the centralized control of residue agents. Then, the corresponding centralized controlled FBSDE is given by*

$$\begin{aligned} &d\mathbf{X}_t = \mathrm{B}(t, \mathbf{X}_t, \mathbf{u}_t)dt + \Sigma d\mathbf{W}_t, \quad d\mathbf{Y}_t^{\mathbf{u}} = -\ell(t, \mathbf{X}_t, \mathbf{u}_t; \mathbf{x}_t)dt + \mathbf{Z}_t^\top d\mathbf{W}_t, \\ &\mathbf{Y}_t^{\mathbf{u}} = \mathcal{J}(t, \mathbf{X}_t), \quad \mathbf{Z}_t^{\mathbf{u}} = \Sigma^\top \nabla_x \mathcal{J}(t, \mathbf{X}_t), \end{aligned} \quad (8)$$

where $(\mathbf{Y}_t^{\mathbf{u}}, \mathbf{Z}_t^{\mathbf{u}})$ denotes the adapted backward pair associated with the agent coalition control $\mathbf{u}$. In particular, the value function is derived as $\inf_{\alpha \in \mathcal{A}} \mathcal{J}_{0,x}(\alpha) = \mathcal{V}(0, x) = \mathbf{Y}_0^{\mathbf{u}^\star}$.

The FBSDE representation avoids explicit state-space discretization by evaluating the cost-to-go along sampled trajectories. However, it fails to eliminate the coupled policy search required for equilibrium. Since computing exact best responses over the full feedback policy classes remains intractable, practical approaches commonly rely on approximate iterative schemes such as *fictitious play* (Laine et al., 2023; Daskalakis et al., 2009). Accordingly, we use the backward processes as policy improvement signals within a *two-stage fictitious play* scheme described in the following:

Two-Stage Fictitious Play. To translate the FBSDE characterization into a practical learning procedure, we develop a two-stage fictitious play that progressively refines the joint policy profile. At each iteration, the first stage proposes a coordinated update for the entire agent coalition, while the second determines how strongly each agent incorporates this proposal. Together, the two stages balance coalition-level coordination and agent-level adjustment and provide a practical route to an approximate equilibrium in high-dimensional protein dynamics.

Let $\bar{\boldsymbol{\theta}}^k = (\bar{\theta}_1^k, \ldots, \bar{\theta}_N^k)$ be the current running parameter profile at iteration k. Starting from an initial profile $\bar{\boldsymbol{\theta}}^0$, the algorithm alternates between the following two stages for $k \leq K$, so as to progressively refine the agent coalition policy toward the equilibrium. We first observe that the backward process[1] provides this gradient direction signal, since it represents the continuation value of the cost functional $(i.e., \mathcal{J})$ described as follows:

$$\nabla_\theta \mathcal{J}_{\mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}}(\bar{\boldsymbol{\alpha}}^k) = \nabla_\theta \mathbf{Y}_t^{k,\bar{\boldsymbol{\alpha}}^k}, \quad \bar{\boldsymbol{\alpha}}^k := \left(\alpha_1(\cdot, \cdot; \bar{\theta}_1^k), \ldots, \alpha_N(\cdot, \cdot; \bar{\theta}_N^k)\right), \quad \widetilde{\alpha}_i^{k+1} := \alpha_i(\cdot, \cdot; \widetilde{\theta}_i^{k+1}).$$

Based on this value-gradient identity, the first stage performs a centralized policy update by moving the current joint parameter profile $\bar{\theta}^k$ along the aggregated FBSDE descent direction. This produces the planner proposal $\widetilde{\theta}^{k+1}$ and the corresponding proposed policy profile as follows:

[1] Please refer to Sec B.3.1 for detailed information

Stage I: Central Planner Coordination

At this stage, the central planner updates the residue-agent policy profile jointly by applying a coordinated descent step on the shared team objective. The resulting policy proposal is given by

$$\boldsymbol{\alpha}\left(t, \mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}; \widetilde{\boldsymbol{\theta}}^{k+1} := \bar{\boldsymbol{\theta}}^k - \mathbb{E}_{Y,t\leq T}\left[\eta_k \nabla_{\boldsymbol{\theta}} \mathbf{Y}_t^{k,\bar{\boldsymbol{\alpha}}^k}\right]\right) = \boldsymbol{\alpha}\left(t, \mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}; \widetilde{\boldsymbol{\theta}}^{k+1}\right) := \tilde{\boldsymbol{\alpha}}^{k+1}. \quad (9)$$

To connect the fictitious play with the Nash equilibrium criterion, we introduce a computable certificate for the theoretical *Nash gap* (*i.e.*, Gap). While the theoretical gap is defined through the best response over $\mathcal{A}_i$, computing this best response at every iteration is infeasible. We therefore use the planner proposal $\widetilde{\alpha}_i^{k+1}$ to measure an algorithmic cost reduction:

$$\underbrace{\widehat{\mathsf{Gap}}_i^k := \left[\mathcal{J}_x(\bar{\alpha}^k) - \mathcal{J}_x\left(\bar{\boldsymbol{\alpha}}_{-i}^k, \widetilde{\alpha}_i^{k+1}\right)\right]_+}_{\text{Algorithmic Nash Gap}} \leq \underbrace{\mathsf{Gap}_i^k := \mathcal{J}_x(\bar{\alpha}^k) - \inf_{\beta_i} \mathcal{J}_x(\bar{\boldsymbol{\alpha}}_{-i}^k, \beta_i)}_{\text{Approximated Nash Gap, Sec. B.3.2}}. \quad (10)$$

Using the algorithmic Nash gap computed above, the second stage assigns an agent-wise assimilation weight for incorporating the planner proposal into the running coalition profile:

Stage II: Agent-coalition Fictitious Play

In the second stage, each agent uses the algorithmic Nash gap in Eq. 10 to determine how strongly it assimilates the planner proposal into the running profile through a fictitious play:

$$\bar{\theta}_i^{k+1} = (1-\lambda_i^k)\bar{\theta}_i^k + \lambda_i^k \widetilde{\theta}_i^{k+1}, \quad \lambda_i^k := \mathrm{Prob}\left(i \,\middle|\, \{\widehat{\mathrm{Gap}}_j^k\}_{j=1}^N\right) \propto \exp\left(\beta \widehat{\mathrm{Gap}}_i^k\right). \quad (11)$$

Here, $\beta > 0$ controls the sensitivity of the assimilation probability to the algorithmic Nash gaps. A larger β favors agents with greater certified cost reductions, while a smaller β yields more uniform assimilation. The two-stage procedure progressively drives the coalition profile toward an approximate equilibrium satisfying the following ε-Nash condition:

Proposition 3.2. ***(informal).*** *For a sufficiently large neural network (i.e., $\mathbf{u}$), there exist constants $\mathsf{C}_D, \mathsf{C}_{i,T}$ and $\mathsf{D}_{i,K}$ such that the learned policy profile $\bar{\boldsymbol{\alpha}}^K$ at iteration $k = K$ satisfies*

$$\mathbb{E}_{\mathbf{X}_0 \sim \mathbb{P}_{init}}\left[\mathsf{Gap}_i^K\left(\bar{\boldsymbol{\alpha}}^K\right)\right] \leq \varepsilon_K := \mathsf{C}_D\left(\sqrt{\log 2} + \frac{\sqrt{\pi}}{2}\right) \max_{1\leq i\leq N}\left\{\mathsf{C}_{i,T}\mathsf{D}_{i,K}\sqrt{P_i}\right\}. \quad (12)$$

Consequently, for every non-optimal policy γ_i of the i-th agent, the ε-Nash equilibrium is satisfied:

$$\mathbb{E}_{\mathbf{X}_0 \sim \mathbb{P}_{init}}\left[\mathcal{J}_{\mathbf{X}_0}\left(\bar{\boldsymbol{\alpha}}^K\right)\right] \leq \mathbb{E}_{\mathbf{X}_0 \sim \mathbb{P}_{init}}\left[\mathcal{J}_{\mathbf{X}_0}\left(\bar{\boldsymbol{\alpha}}_{-i}^K, \gamma^i\right)\right] + \varepsilon_K, \quad (13)$$

where P_i is the number of trainable parameters used to represent the neural policy u_i ($\varepsilon_K \propto \sqrt{P_i}$), $\mathsf{D}_{i,K}$ measures the policy variation generated by the gradient updates ($\varepsilon_K \propto \mathsf{D}_{i,K} \propto \nabla_{\boldsymbol{\theta}}\mathbf{Y}$), and $\mathsf{C}_{i,T}$ summarizes the system regularity ($\varepsilon_K \propto \mathsf{C}_{i,T}$). Once the neural network is sufficiently expressive, the bound tightens as fictitious play stabilizes the gradient updates and reduces the resulting policy variation. Hence, gradient convergence certifies that the learned agent coalition approaches a Nash equilibrium. Since this certificate is obtained by averaging policy deviations over the initial state distribution, rather than controlling them uniformly over all states and policies, the result is stated as an expected weak Nash equilibrium.

4 Related Work

Game-Theoretic Formulations for Static Molecular Systems. Several studies have sought to understand molecular systems through interactions among multiple agents. GARN (Boudard et al., 2015) applies regret minimization updates to sample diverse candidate RNA structures broadly. To predict how multiple chains assemble into a stable structure, DockGame (Somnath et al., 2024) models protein chains as cooperative interacting players. Combining game theory with reinforcement learning, another study (Bougueroua et al., 2023) infers peptide conformations from graph-based representations of molecular structure. Although these methods seek stable molecular structures, their focus remains mainly on static molecular configurations rather than how different parts of a protein

Table 1: **Quantitative results on ATLAS and PED.** We evaluate biomolecular dynamics modeling quality across coordinate accuracy, residue-level structural fidelity, and coupled residue dynamics. For each method, the upper and lower entries correspond to ATLAS and PED, respectively. The best and second-best results within each dataset are shown in **bold** and **blue**, respectively.

Method	Coordinate fidelity		Residue-level fidelity			Coupled dynamics
	MSE↓	RMSD↓	RMSF Err.↓	Contact RMSE↓	Residue-lDDT↑	DCCM Err.↓
AlphaFlow	3.96×10^{-1} 6.54×10^{-1}	5.68×10^{-1} 9.79×10^{-1}	2.44×10^{-1} 2.05×10^{-1}	3.44×10^{-1} 2.62×10^{-1}	3.73×10^{-1} 2.54×10^{-1}	2.88×10^{-1} 2.96×10^{-1}
MDGen	6.86×10^{-1} 2.29×10^{0}	1.09×10^{0} 1.75×10^{0}	1.34×10^{-1} 3.09×10^{-1}	6.61×10^{-1} 7.89×10^{-1}	1.31×10^{-1} 1.49×10^{-1}	2.79×10^{-1} 4.54×10^{-1}
UniSim	3.39×10^{-2} 3.74×10^{-2}	1.77×10^{-1} 2.72×10^{-1}	9.65×10^{-2} 1.05×10^{-1}	5.22×10^{-2} 3.96×10^{-2}	7.95×10^{-1} 7.80×10^{-1}	1.16×10^{-1} 6.15×10^{-2}
AlphaFolding	1.24×10^{-2} 4.82×10^{-2}	1.32×10^{-1} 3.01×10^{-1}	5.01×10^{-2} 2.64×10^{-2}	7.69×10^{-2} 8.58×10^{-2}	8.13×10^{-1} 6.33×10^{-1}	2.24×10^{-1} 2.45×10^{-1}
PVB	2.93×10^{-2} 2.35×10^{-2}	1.65×10^{-1} 2.31×10^{-1}	9.06×10^{-2} 6.05×10^{-2}	5.78×10^{-2} 4.03×10^{-2}	8.09×10^{-1} 7.74×10^{-1}	**7.23×10^{-2}** 5.83×10^{-2}
GLDP	**4.32×10^{-3}** **1.54×10^{-2}**	**1.13×10^{-1}** **2.05×10^{-1}**	**3.40×10^{-2}** **1.88×10^{-2}**	**4.09×10^{-2}** **3.56×10^{-2}**	**9.35×10^{-1}** **7.84×10^{-1}**	1.11×10^{-1} **2.94×10^{-2}**
Ours	**1.63×10^{-3}** **1.44×10^{-3}**	**6.02×10^{-2}** **6.22×10^{-2}**	**1.93×10^{-3}** **2.19×10^{-3}**	**1.31×10^{-2}** **4.75×10^{-3}**	**9.36×10^{-1}** **9.33×10^{-1}**	**6.10×10^{-3}** **3.97×10^{-3}**

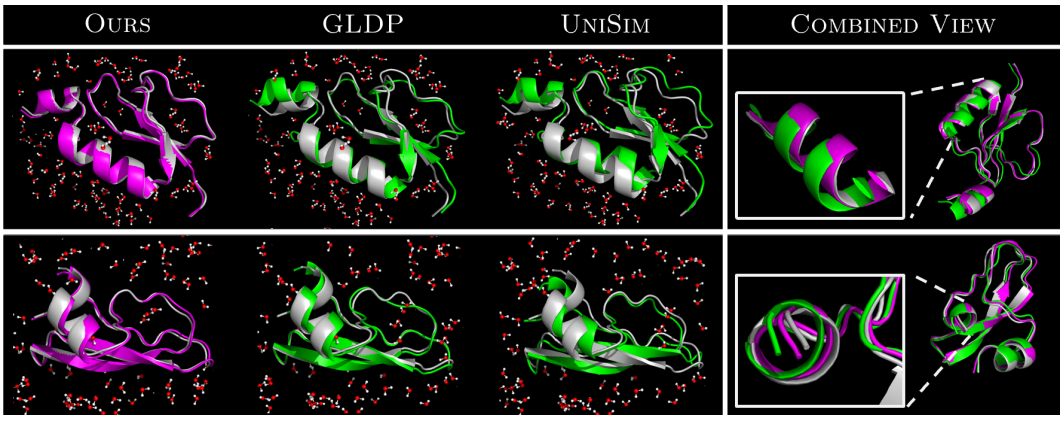

Figure 2: **Qualitative comparison of protein structures for ATLAS.** We visualize protein structures from the ATLAS dataset, with 2PST on the top and 1KNT on the bottom. *(Left)* Our method (magenta) and prior works (green) are separately overlaid on the ground truth (gray). *(Right)* We show the ground truth, our method, and GLDP together, accompanied by magnified insets corresponding to the boxed regions shown for each protein.

move together over time. We instead formulate protein conformational evolution as a differential game, where residue-level policies interact to capture coordinated motions over time.

Generative Protein Ensembles and Trajectories. Beyond static protein structure modeling, generative frameworks model conformational ensembles and molecular trajectories from molecular dynamics simulations. Using flow matching, AlphaFlow (Jing et al., 2024a) models sequence-conditioned conformational ensembles. AlphaFolding (Cheng et al., 2025) uses 4D diffusion to model dynamic protein structures across time steps. At the trajectory level, MDGen (Jing et al., 2024b) models molecular trajectories through a conditional generative framework. In contrast, our method formulates biomolecular dynamics as a centralized differential game, in which interacting residue-level feedback policies model coupled residue motions under a shared objective.

Learning Molecular State Transitions. Recent methods learn transition models to describe how molecular structures change over time. UniSim (Yu et al., 2025) models time-coarsened biomolecular dynamics by learning transitions between molecular states at larger temporal intervals. To capture long-horizon motion, latent-space approaches such as GLDP (Sengar et al., 2026) evolve molecular structures in learned representation spaces. For transitions toward intermediate target states, PVB (Yu et al., 2026) moves an initial structure through a pretrained variational bridge. Whereas these methods model dynamics through transitions in global or latent state spaces, our method represents biomolecular dynamics through contact-aware interactions among agents under a shared objective.

Table 2: **Quantitative results on MISATO biomolecular dynamics.** We evaluate biomolecular dynamics modeling quality on solvated protein systems across coordinate accuracy, residue-level structural fidelity, and coupled residue dynamics. The best and second-best available results are shown in **bold** and **blue**, respectively.

Method	Coordinate fidelity		Residue-level fidelity			Coupled dynamics
	MSE↓	RMSD↓	RMSF Err.↓	Contact RMSE↓	Residue-lDDT↑	DCCM Err.↓
AlphaFlow	4.62×10^{-1}	5.59×10^{-1}	2.50×10^{-1}	3.41×10^{-1}	3.48×10^{-1}	2.53×10^{-1}
MDGen	6.43×10^{-1}	1.03×10^{0}	1.61×10^{-1}	7.52×10^{-1}	1.09×10^{-1}	3.09×10^{-1}
UniSim	5.01×10^{-2}	2.16×10^{-1}	1.21×10^{-1}	5.57×10^{-2}	7.74×10^{-1}	8.96×10^{-2}
AlphaFolding	1.51×10^{-2}	1.44×10^{-1}	5.41×10^{-2}	6.51×10^{-2}	8.12×10^{-1}	2.16×10^{-1}
PVB	3.25×10^{-2}	2.16×10^{-1}	7.80×10^{-2}	5.10×10^{-2}	7.71×10^{-1}	**6.46×10^{-2}**
GLDP	**5.78×10^{-3}**	**1.30×10^{-1}**	**3.25×10^{-2}**	**2.99×10^{-2}**	**9.29×10^{-1}**	1.21×10^{-1}
Ours	**1.81×10^{-3}**	**6.38×10^{-2}**	**2.24×10^{-3}**	**1.36×10^{-2}**	**9.30×10^{-1}**	**6.36×10^{-3}**

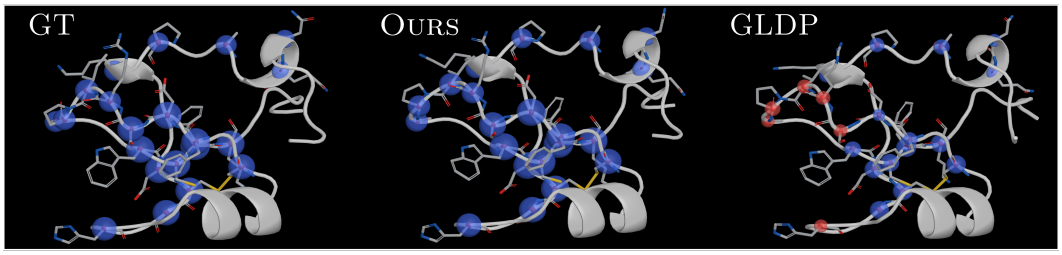

Figure 3: **Qualitative comparison of residue motion correlations on MISATO.** We compare DCCM correlations in the 6V1C system across the ground truth (GT), our method, and GLDP. We visualize spheres of positive (blue) and negative (red) correlations and scale the sphere radii by its absolute magnitude. Larger spheres therefore represent residue pairs with stronger correlated or anticorrelated motions.

5 Experiments

We evaluate the proposed centralized differential game across two settings that capture different biomolecular dynamics. First, we study protein dynamics under a cooperative formulation, where residue agents coordinate to model protein conformational evolution. Second, we turn to protein–ligand dynamics by introducing the ligand as a distinct agent and formulating its interaction with the residue coalition as a non-cooperative game. Across multiple biomolecular systems, we compare the proposed formulation with existing methods under a common evaluation protocol.

Datasets. We construct our evaluation with protein systems from ATLAS (Vander Meersche et al., 2024), PED (Ghafouri et al., 2024), and MISATO (Siebenmorgen et al., 2024). While ATLAS covers standardized molecular dynamics simulations of representative folded proteins, PED broadens the evaluation to intrinsically disordered proteins and flexible regions. To examine ligand-bound dynamics, we further include protein–ligand complexes from MISATO. We evaluate cooperative protein dynamics across all three datasets and use MISATO for the non-cooperative protein–ligand setting. We provide details of the selected systems and data generation protocol in Appendix B.1.

Evaluation Metrics. We evaluate biomolecular dynamics using three groups of metrics. For coordinate fidelity, MSE and RMSD assess how accurately the generated trajectories reproduce the reference dynamics. At the residue level, we use RMSF Error, Residue-lDDT (Mariani et al., 2013), and Contact RMSE to assess flexibility, local geometry, and persistent interaction patterns, thereby evaluating how well the generated dynamics reproduce residue-wise behavior over time beyond simple Cartesian deviation. Through RMSF Error, we compare residue fluctuation profiles and examine the reproduction of rigid and flexible regions. Using local distance patterns, Residue-lDDT provides an alignment-insensitive measure of residue-level structural consistency. By comparing contact occupancies accumulated across the trajectory, Contact RMSE evaluates persistent contact patterns between residues. For coupled dynamics, we further report DCCM Error (Ichiye & Karplus, 1991), which measures errors in correlated and anticorrelated residue motions. With these metrics, we evaluate whether generated biomolecular dynamics preserve residue-level behavior while maintaining coupled motions across the protein. Details of the evaluation metrics are provided in Appendix B.1.

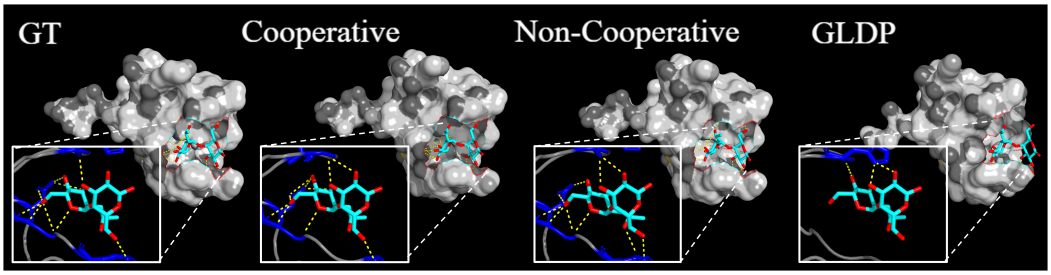

Figure 4: **Visualization of ligand pocket contact geometry on MISATO.** We visualize the bound protein–ligand complex of 6V1C for ground truth (GT), cooperative and non-cooperative modeling of our framework and GLDP from left to right. The ligand is shown in sticks (cyan), and the binding pocket is shown as an atom colored molecular surface. We identify ligand pocket contacts (yellow) from potential polar contacts.

Scenario I: Cooperative Games for Protein Dynamics. We formulate protein dynamics as a cooperative game in which residues act as biomolecular agents. These agents are coupled through a shared objective, which enables a compact representation of conformational flexibility while preserving protein topology. We evaluate this cooperative formulation on protein systems from ATLAS and PED, with results reported in Table 1. On ATLAS, our method reduces MSE and RMSD by $\mathbf{2.7}\times$ and $\mathbf{1.9}\times$, respectively, compared with the strongest baseline. The gains are more pronounced for residue-level dynamics, with $\mathbf{17.6}\times$ lower RMSF Error and $\mathbf{11.9}\times$lower DCCM Error. As illustrated in Figure 2, our predictions closely follow the ground truth conformations in both the overall structure and the magnified regions. These advantages persist on PED, despite its broader and more flexible conformational ensembles with the lower Contact RMSE further confirming that the predicted dynamics preserve persistent interaction patterns across conformations.

We also observe consistent improvements in ligand bound systems from MISATO, where protein motions are additionally influenced by molecular binding. As reported in Table 2, our method achieves $\mathbf{14.5}\times$ lower RMSF Error and $\mathbf{10.2}\times$ lower DCCM Error, indicating more accurate residue fluctuations and correlated motions in this setting. In Figure 3, the DCCM visualization further shows that our prediction more closely matches the ground truth distribution of positive and negative correlations than the baseline. Overall, these findings demonstrate that cooperative biomolecular agents effectively model protein dynamics across diverse protein structures and ligand influenced molecular systems, while preserving residue-level fluctuations and long-range correlated motions.

Scenario II: Non-Cooperative Games for Protein-Ligand Interaction. Protein dynamics often arise through interactions with other biomolecular entities, among which protein–ligand binding represents an important problem in biomolecular modeling. While the first scenario focuses on protein motion, the second extends the framework to model the coupled evolution of protein and ligand pose. Given the distinct physical characteristics and dynamical roles of proteins and ligands (Alberts et al., 1989), we assign them different objectives and formulate their interaction as a *non-cooperative game* (Basar & Olsder, 1999). Within this formulation, the objectives from the first scenario remain with the residue coalition, whereas the ligand optimizes a separate objective for binding-pocket compatibility. This objective includes a contact term that encourages ligand poses to match the predicted ligand–pocket contact pattern. Further details of the objective are provided in Appendix B.1.

Table 3: **Protein–ligand dynamics on MISATO.** We evaluate ligand–pocket interaction fidelity, coordinate accuracy, and correlated biomolecular dynamics.

Metric ↓	SCENARIO 1 (coop.)	SCENARIO 2 (non–coop.)	GLDP
Contact Occ.	$\mathbf{8.11\times10^{-2}}$	$\mathbf{7.98\times10^{-2}}$	1.35×10^{-1}
Pocket RMSF	$\mathbf{3.75\times10^{-3}}$	$\mathbf{2.21\times10^{-3}}$	1.44×10^{-2}
COM Fluct.	$\mathbf{1.42\times10^{-3}}$	$\mathbf{5.49\times10^{-4}}$	2.44×10^{-3}
MSE	$\mathbf{1.08\times10^{-3}}$	$\mathbf{1.10\times10^{-3}}$	2.85×10^{-3}
DCCM Error	$\mathbf{4.97\times10^{-1}}$	$\mathbf{5.01\times10^{-1}}$	9.45×10^{-1}

We compare the two scenarios and GLDP on MISATO to evaluate the non-cooperative formulation for protein–ligand dynamics. Table 3 reports coordinate accuracy, ligand–pocket fidelity, and correlated residue dynamics. Contact occupancy RMSE evaluates contact frequency, while pocket-aligned ligand RMSF MAE and ligand COM fluctuation MAE measure atom-wise flexibility and pocket-ligand dynamics, respectively. The first scenario achieves the lowest MSE and DCCM Error, whereas the second scenario achieves the best ligand–pocket fidelity. Relative to GLDP, the second scenario reduces pocket-aligned ligand RMSF MAE and ligand COM fluctuation MAE by $\mathbf{6.52}\times$ and $\mathbf{4.44}\times$, respectively. These results suggest that the separate objective for binding-pocket compatibility improves overall ligand–pocket interaction fidelity while retaining competitive coordinate accuracy.

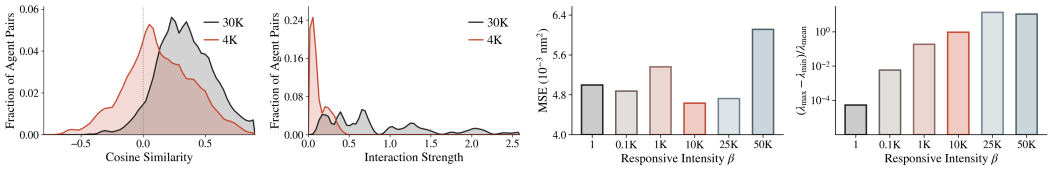

Figure 5: **Ablation study.** *(Left)* We compare distributions of pairwise cosine similarities between agent control profiles $\mathbf{u}^i$ and $\mathbf{u}^j$ and the norms of the interaction components Ψ_{ij} at 4K and 30K learning steps. *(Right)* We assess how response intensity β affects MSE and relative range of belief update weight λ at 4K learning steps.

As shown in Figure 4, the improvements in ligand–pocket interaction fidelity are also reflected in the predicted contact patterns at the binding interface. The ligand (cyan) is shown within the binding-pocket surface, while the dashed lines (yellow) indicate potential polar contacts between the ligand and surrounding pocket residues associated with hydrogen-bond formation. Despite its relatively low MSE, GLDP shows noticeable discrepancies in these contacts, whereas the proposed formulations more closely reproduce the binding geometry and the polar-contact pattern around the ligand. Taken together, these observations support the second scenario as a faithful description of ligand motion and ligand–pocket interactions while retaining accurate modeling of protein dynamics.

Ablation Study. We conduct an ablation study on the 6V1C protein–ligand complex from MISATO to examine two aspects of the centralized differential game. First, we test whether the central planner coordinates biomolecular agents while preserving distinct control directions. Second, we examine how the response intensity β affects coalition updates through the algorithmic Nash gap.

Centralized Coordination and Agent Control Diversity. The two-stage fictitious play is designed to coordinate the residue coalition while maintaining diversity among agent controls. We examine this behavior at 4K and 30K steps using pairwise control similarity $\langle \mathbf{u}^i, \mathbf{u}^j \rangle / (\|\mathbf{u}^i\| \|\mathbf{u}^j\|)$ for directional agreement and interaction strength $\|\Psi_{ij}\|$ for pairwise coupling. Figure 5 *(Left)* shows that the similarities are centered around zero and most interactions remain weak at 4K steps. At 30K steps, the similarity distribution shifts toward positive values, while the interaction strengths spread over a broader range. Agent controls therefore become more aligned overall, but their coordination relies on differentiated coupling across residues rather than a uniform control profile. These patterns show that fictitious play forms coalition behavior while preserving variation among agent controls.

Response Intensity and Belief Update Selectivity. We study how the response intensity β governs the influence of algorithmic Nash-gap differences on coalition updates. The ϵ-Nash analysis characterizes β as the sensitivity of the belief update to agent-wise algorithmic Nash gaps. A small β approaches uniform fictitious play, whereas a larger β assigns greater influence to agents with larger certified cost reductions. We examine how this theoretical role appears in practice by varying β and measuring the relative range of the belief update weights, $(\lambda_{\max} - \lambda_{\min})/\lambda_{\text{mean}}$, together with MSE at 4K steps. Figure 5 *(Right)* shows that small β values produce nearly uniform belief updates and largely suppress the agent-wise differences identified by the Nash gaps. An intermediate β retains these differences while distributing the update across the coalition, which yields the lowest MSE. Larger values over-amplify the same differences and concentrate the update on fewer agents, accompanied by degraded MSE. These results support the theoretical role of β and show that effective coalition refinement requires selective belief updates without excessive concentration.

6 Conclusion

In this work, we propose a centralized differential game for biomolecular dynamics, in which protein residues act as cooperative agents under a shared objective. We formulate the cooperative game as a stochastic optimal-control problem, characterize its optimality through the HJB equation and an FBSDE, and develop a two–stage fictitious play algorithm for policy optimization. Across protein dynamics benchmarks, our approach improves structural fidelity and more accurately captures residue-level fluctuations and coupled motions than existing baselines. We further extend the framework to protein–ligand systems through a non-cooperative game, assigning the ligand an objective that encourages compatibility with the binding pocket. This extension strengthens ligand–pocket interactions while preserving protein dynamics. Overall, biomolecular agent games provide a stochastic control framework for modeling both protein dynamics and protein–ligand interactions.

AI USE STATEMENT

In this work, we used generative AI tools to improve the clarity of the manuscript and assist in mathematical proofs. We have reviewed all AI-assisted content and verified the mathematical arguments, and we take full responsibility for the final manuscript.

REFERENCES

Bruce Alberts, Dennis Bray, Julian Lewis, Martin Raff, Keith Roberts, and James D. Watson. *Molecular Biology of the Cell*. Garland Publishing, 2nd edition, 1989. ISBN 0824036956.

T. Basar and G.J. Olsder. *Dynamic Noncooperative Game Theory*. Classics in Applied Mathematics. Society for Industrial and Applied Mathematics, 1999. ISBN 9780898714296.

Mélanie Boudard, Julie Bernauer, Dominique Barth, Johanne Cohen, and Alain Denise. Garn: Sampling rna 3d structure space with game theory and knowledge-based scoring strategies. *PLOS ONE*, 10(8):1–21, 08 2015. doi: 10.1371/journal.pone.0136444. URL https://doi.org/10.1371/journal.pone.0136444.

Sana Bougueroua, Marie Bricage, Ylène Aboulfath, Dominique Barth, and Marie-Pierre Gaigeot. Algorithmic graph theory, reinforcement learning and game theory in md simulations: From 3d structures to topological 2d-molecular graphs (2d-molgraphs) and vice versa. *Molecules*, 28(7), 2023. ISSN 1420-3049. doi: 10.3390/molecules28072892. URL https://www.mdpi.com/1420-3049/28/7/2892.

Giovanni Bussi and Michele Parrinello. Accurate sampling using langevin dynamics. *Phys. Rev. E*, 75:056707, May 2007. doi: 10.1103/PhysRevE.75.056707. URL https://link.aps.org/doi/10.1103/PhysRevE.75.056707.

Kaihui Cheng, Ce Liu, Qingkun Su, Jun Wang, Liwei Zhang, Yining Tang, Yao Yao, Siyu Zhu, and Yuan Qi. 4d diffusion for dynamic protein structure prediction with reference and motion guidance. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 39, pp. 93–101, 2025.

Constantinos Daskalakis, Paul W. Goldberg, and Christos H. Papadimitriou. The complexity of computing a nash equilibrium. *Communications of the ACM*, 52(2):89–97, 2009. doi: 10.1145/1461928.1461951.

Peter Eastman, Jason Swails, John D. Chodera, Robert T. McGibbon, Yutong Zhao, Kyle A. Beauchamp, Lee-Ping Wang, Andrew C. Simmonett, Matthew P. Harrigan, Chaya D. Stern, Rafal P. Wiewiora, Bernard R. Brooks, and Vijay S. Pande. Openmm 7: Rapid development of high performance algorithms for molecular dynamics. *PLoS computational biology*, 13(7):1–17, 2017. URL https://doi.org/10.1371/journal.pcbi.1005659.

Hamidreza Ghafouri, Tamas Lazar, Alessio Del Conte, Luiggi G Tenorio Ku, PED Consortium, Peter Tompa, Silvio C E Tosatto, and Alexander Miguel Monzon. Ped in 2024: improving the community deposition of structural ensembles for intrinsically disordered proteins. *Nucleic acids research*, 52(D1):D536–D544, 2024. doi: 10.1093/nar/gkad947.

Katherine A Henzler-Wildman, Vu Thai, Ming Lei, Maria Ott, Magnus Wolf-Watz, Tim Fenn, Ed Pozharski, Mark A Wilson, Gregory A Petsko, Martin Karplus, et al. Intrinsic motions along an enzymatic reaction trajectory. *Nature*, 450(7171):838–844, 2007.

P.H. Hünenberger, A.E. Mark, and W.F. van Gunsteren. Fluctuation and cross-correlation analysis of protein motions observed in nanosecond molecular dynamics simulations. *Journal of Molecular Biology*, 252(4):492–503, 1995. ISSN 0022-2836. doi: https://doi.org/10.1006/jmbi.1995.0514. URL https://www.sciencedirect.com/science/article/pii/S0022283685705141.

Toshiko Ichiye and Martin Karplus. Collective motions in proteins: a covariance analysis of atomic fluctuations in molecular dynamics and normal mode simulations. *Proteins: Structure, Function, and Bioinformatics*, 11(3):205–217, 1991. doi: https://doi.org/10.1002/prot.340110305.

Rufus Isaacs. *Differential games: a mathematical theory with applications to warfare and pursuit, control and optimization*. Dover Publications, 1999. ISBN 9780486406824.

Bowen Jing, Bonnie Berger, and Tommi Jaakkola. AlphaFold meets flow matching for generating protein ensembles. In Ruslan Salakhutdinov, Zico Kolter, Katherine Heller, Adrian Weller, Nuria Oliver, Jonathan Scarlett, and Felix Berkenkamp (eds.), *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 22277–22303. PMLR, 21–27 Jul 2024a. URL https://proceedings.mlr.press/v235/jing24a.html.

Bowen Jing, Hannes Stärk, Tommi Jaakkola, and Bonnie Berger. Generative modeling of molecular dynamics trajectories. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 40534–40564. Curran Associates, Inc., 2024b. doi: 10.52202/079017-1282.

Wolfgang Kabsch. A solution for the best rotation to relate two sets of vectors. *Acta Crystallographica Section A*, 32(5):922–923, Sep 1976. doi: 10.1107/S0567739476001873. URL https://doi.org/10.1107/S0567739476001873.

Forrest Laine, David Fridovich-Keil, Chih-Yuan Chiu, and Claire Tomlin. The computation of approximate generalized feedback nash equilibria. *SIAM Journal on Optimization*, 33(1):294–318, 2023. doi: 10.1137/21M142530X.

Ben Leimkuhler and Charles Matthews. *Molecular Dynamics: With Deterministic and Stochastic Numerical Methods*, volume 39 of *Interdisciplinary Applied Mathematics*. Springer, Cham, 2015. ISBN 978-3-319-16374-1. doi: 10.1007/978-3-319-16375-8.

Valerio Mariani, Marco Biasini, Alessandro Barbato, and Torsten Schwede. lddt: a local superposition-free score for comparing protein structures and models using distance difference tests. *Bioinformatics*, 29(21):2722–2728, 2013. doi: 10.1093/bioinformatics/btt473.

Noam Nisan, Tim Roughgarden, Eva Tardos, and Vijay V Vazirani. *Algorithmic game theory*. Cambridge university press, 2007. ISBN 9781139466547.

Grigorios A. Pavliotis. *Stochastic Processes and Applications: Diffusion Processes, the Fokker–Planck and Langevin Equations*, volume 60 of *Texts in Applied Mathematics*. Springer, 2014. ISBN 978-1-4939-1323-7. doi: 10.1007/978-1-4939-1323-7.

Nataliya Popovych, Shangjin Sun, Richard H Ebright, and Charalampos G Kalodimos. Dynamically driven protein allostery. *Nature structural & molecular biology*, 13(9):831–838, 2006.

Aditya Sengar, Jiying Zhang, Pierre Vandergheynst, and PATRICK BARTH. Beyond ensembles: Simulating all-atom protein dynamics in a learned latent space. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=AwowReRWXI.

Till Siebenmorgen, Filipe Menezes, Sabrina Benassou, Erinc Merdivan, Kieran Didi, André Santos Dias Mourão, Radosław Kitel, Pietro Liò, Stefan Kesselheim, Marie Piraud, et al. Misato: machine learning dataset of protein–ligand complexes for structure-based drug discovery. *Nature computational science*, 4(5):367–378, 2024.

Vignesh Ram Somnath, Pier Giuseppe Sessa, Maria Rodriguez Martinez, and Andreas Krause. Dockgame: Cooperative games for multimeric rigid protein docking, 2024. URL https://openreview.net/forum?id=wmq67R2PIu.

Yann Vander Meersche, Gabriel Cretin, Aria Gheeraert, Jean-Christophe Gelly, and Tatiana Galochkina. Atlas: protein flexibility description from atomistic molecular dynamics simulations. *Nucleic Acids Research*, 52(D1):D384–D392, 01 2024. ISSN 0305-1048. doi: 10.1093/nar/gkad1084.

Ziyang Yu, Wenbing Huang, and Yang Liu. Unisim: A unified simulator for time-coarsened dynamics of biomolecules. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=W598uW37zt.

Ziyang Yu, Wenbing Huang, and Yang Liu. Unified biomolecular trajectory generation via pretrained variational bridge. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=8HH9dBOxwu.

A NOTATION TABLE

We summarize the notation used throughout the paper. Symbols introduced only for a specific proof are defined locally within the corresponding proof.

Notation	Description
N	Number of biomolecular agents.
i, j	Indices of biomolecular agents.
$\mathcal{I}_i$	Set of atoms assigned to agent i.
$\mathbf{x}_t$	Protein-level reference state at time t.
$\mathbf{P}_t$	Conjugate momentum of the reference molecular system.
$\mathbf{X}_t^i$	Local molecular state associated with agent i.
$\mathbf{X}_t$	Assembled state of the full agent coalition.
$\mathbf{u}_t^i$	Realized control of agent i at time t.
$\mathbf{u}_t$	Joint control profile of all biomolecular agents.
B_i	Drift function governing the dynamics of agent i.
B	Assembled drift of the full biomolecular system.
Φ_i	Learned local dynamics component of agent i.
Ψ_{ij}	Learned interaction component between agents i and j.
ν_i	Mass probability measure associated with agent i.
E	Molecular potential-energy function.
$\mathbf{M}$	Molecular mass matrix.
γ	Friction or spatial-decay coefficient, as specified by context.
k_B	Boltzmann constant.
T	Terminal time horizon or bath temperature, as specified by context.
$\mathbf{W}_t$	Brownian motion driving the stochastic molecular dynamics.
Σ	Diffusion matrix of the controlled molecular dynamics.
$\mathbb{P}_{init}$	Probability distribution of the initial molecular state.
$\mathcal{F}_t$	Global molecular-information filtration up to time t.
θ_i	Trainable neural-network parameter of agent i.
Θ_i	Parameter space of agent i.
α_i	Parameterized feedback policy of agent i.
$\boldsymbol{\alpha}$	Joint feedback-policy profile of all agents.
$\boldsymbol{\alpha}_{-i}$	Policy profile of all agents other than agent i.
$\mathcal{U}_i$	Admissible control set of agent i.
$\mathcal{A}_i$	Admissible feedback-policy class of agent i.
$\mathcal{A}$	Joint admissible policy class of the agent coalition.
ℓ	Running cost for trajectory matching and control regularization.
g_T	Terminal molecular cost.
π	Gibbs density associated with the molecular energy.
Z_β	Partition function of the Gibbs density.
$\mathcal{J}_x$	State-time conditioned objective functional.
$\mathcal{V}$	Value function of the centralized control problem.
$\mathscr{H}$	Hamiltonian associated with the centralized control problem.
$\mathbf{Y}_t$	Backward value process in the FBSDE representation.
$\mathbf{Z}_t$	Backward martingale process in the FBSDE representation.
$\mathsf{Gap}_{i,t,x}$	Nash gap associated with agent i.
$\mathsf{Gap}_{t,x}$	Maximum Nash gap over all biomolecular agents.
$\mathsf{Gap}_{i,0,x}^{\mathcal{C}}$	Pointwise Nash gap restricted to the record policy class.
k, K	Current fictitious-play iteration and total number of iterations.
$\bar{\boldsymbol{\theta}}^k$	Running parameter profile at iteration k.
$\widetilde{\boldsymbol{\theta}}^{k+1}$	Central-planner parameter proposal produced in Stage I.
$\bar{\boldsymbol{\alpha}}^k$	Running policy profile at iteration k.
$\widetilde{\boldsymbol{\alpha}}^{k+1}$	Central-planner policy proposal produced in Stage I.
η_k	Gradient step size at iteration k.
$G_{i,k}$	Aggregated FBSDE gradient update for agent i.

Notation	Description
$\widehat{\mathsf{Gap}}_i^{\,k}$	Algorithmic Nash gap of agent i.
λ_i^k	Agent-wise assimilation coefficient used in Stage II.
β	Sensitivity parameter controlling the assimilation weights.
ρ_k	Time-sampling measure used to aggregate FBSDE gradients.
$\Phi_{i,k}$	Stage-I operator generating the local proposal class.
$\Psi_{i,k}$	Stage-II operator assimilating the proposal into the running policy.
$\widetilde{\mathcal{A}}_i^{k+1}$	Local policy class generated in Stage I.
$\mathcal{A}_i^{k+1}$	Local policy class generated after Stage-II assimilation.
$\mathcal{C}_i$	Retained record policy class of agent i.
$\mathcal{H}_i$	Hilbert space of feedback policies for agent i.
$\mu_{i,K}$	Policy-evaluation measure generated up to iteration K.
ρ_i	Metric on the record policy class of agent i.
d_i	Canonical pseudometric of the value-deviation process.
$N(\varepsilon, \mathcal{C}_i, \rho_i)$	Covering number of the record policy class.
$\mathrm{diam}_{\rho_i}(\mathcal{C}_i)$	Diameter of the record policy class.
$\|\cdot\|_{\psi_2}$	Sub-Gaussian Orlicz norm.
m_0	Mean of the initial molecular state.
K_0	Sub-Gaussian scale of the initial-state distribution.
P_i	Number of trainable parameters in the neural policy of agent i.
$R_{i,K}$	Input-radius term associated with the policy-evaluation measure.
$\mathsf{J}_{i,K}^{\mathrm{NN}}$	Neural-policy Jacobian bound.
$\mathsf{D}_{i,K}$	Gradient-controlled diameter bound for the record policy class.
$\mathsf{C}_{i,K}^{\mathrm{NN}}$	Neural-network complexity term in the covering-number bound.
C_D	Universal constant in Dudley's entropy inequality.
$\mathsf{C}_{i,T}$	Agent-dependent sensitivity constant over the time horizon.
$\varepsilon_{i,K}^{\star}$	Optimized splitting scale for the entropy bound.
ε_K	Expected equilibrium tolerance after K iterations.

B APPENDIX

B.1 EXPERIMENTAL DETAILS

Model Architecture. We represent each molecular system as a collection of biomolecular agents. In the protein-only setting, each residue serves as an individual agent. For the protein–ligand setting, we retain the residue-level agents and introduce the complete ligand as one additional agent. To accommodate differences in atom count and composition, a shared encoder maps every molecular block to a 64-dimensional latent representation. Interactions among agents are modeled through a two-layer distance-aware graph constructed from their spatial relationships. With a hidden dimension of 128, the graph produces pairwise interaction features that are aggregated into a 64-dimensional context representation for each agent. A context-conditioned latent update network with a hidden dimension of 1,024 then updates each agent representation based on this context. Finally, a corresponding decoder reconstructs the molecular coordinates of the original block from the updated latent representation. This design allows both the latent state and the reconstructed coordinates to account for the surrounding molecular environment.

We optimize the model through a two-stage game-optimization procedure. At every optimization step, the first stage performs one inner update. Every ten optimization steps, the second stage evaluates all agents in the system. In the cooperative protein-only setting, this evaluation uses the Algorithmic Nash gap. In the non–cooperative protein–ligand setting, the Nash gap is applied to the residue agents, whereas a separate ligand loss is used for the ligand agent. For each molecular system, we train the model for 30,000 optimization iterations with the Adam optimizer and a learning rate of 1×10^{-4}.

Stochastic Protein–Ligand Objective. In the protein–ligand setting, we introduce the ligand as the $(N+1)$-st biomolecular agent. The residue agents retain the shared coalition objective $\mathcal{J}_x$, whereas the ligand agent is evaluated through the state-conditioned stochastic objective

$$\mathcal{J}_x^{N+1}(\boldsymbol{\alpha}) := \mathbb{E}\left[\int_t^T \left(\left\|\mathbf{X}_s^{N+1,\boldsymbol{\alpha}} - \mathbf{x}_s^{N+1}\right\|_2^2 + \mathcal{L}\left(\mathbf{X}_s^{\boldsymbol{\alpha}}, \mathbf{x}_s\right)\right) ds \,\middle|\, \mathbf{X}_t^{\boldsymbol{\alpha}} = x\right]. \tag{14}$$

The specialized term in Eq. 14 is defined as follows:

$$\mathcal{L}(\mathbf{X}, \mathbf{x}) = \mathbb{E}_{(a,b)\sim \mathrm{Unif}(\mathcal{I}_{\mathrm{lig}} \times \mathcal{I}_{\mathrm{prot}})}\left[\left|\kappa_{ab}(\mathbf{X}) - \kappa_{ab}(\mathbf{x})\right|^2\right]. \tag{15}$$

Here, κ_{ab} denotes the smooth contact score between ligand atom a and protein atom b:

$$\kappa_{ab}(z) = \mathrm{sigmoid}\left(\frac{r_{\mathrm{c}} - \|z_a - z_b\|_2}{\tau}\right). \tag{16}$$

The contact cutoff is denoted by r_{c}, while $\tau > 0$ controls the smoothness of the contact transition. Using the same notation as in Eq. 10, algorithmic Nash gap for the ligand agent can be described as

$$\widehat{\mathsf{Gap}}_{N+1}^{\,k} := \left[\mathcal{J}_x^{N+1}\left(\bar{\boldsymbol{\alpha}}^k\right) - \mathcal{J}_x^{N+1}\left(\bar{\boldsymbol{\alpha}}_{-(N+1)}^k, \widetilde{\alpha}_{N+1}^{k+1}\right)\right]_+. \tag{17}$$

Protein Systems and Reference Trajectories. We use 2FMA, 2PST, and 1KNT from ATLAS, three conformations each of 2EC3 and 6UCK from PED, and the 6V1C, 5YJ8, and 5EPK protein–ligand complexes from MISATO. Table 5 summarizes the protein, ligand, and total heavy-atom counts for all evaluated systems. We generate all reference trajectories with OpenMM (Eastman et al., 2017). We use AMBER19 for proteins, TIP3P-FB for water, and separately parameterized ligand topologies for the MISATO complexes. Long-range electrostatic interactions are treated with particle-mesh Ewald, and the nonbonded cutoff is set to $1.0\mathrm{nm}$. All simulations use a Langevin integrator at $350\mathrm{K}$ with an integration timestep of $2\mathrm{fs}$.

For ATLAS, we generate trajectories of $1\mathrm{ns}$, $2\mathrm{ns}$, and $4\mathrm{ns}$, respectively. We use the same simulation horizons for MISATO, while each selected PED conformation is simulated for $1\mathrm{ns}$. Each reference trajectory is uniformly subsampled into one conditioning frame and 1,024 target frames over the full simulation horizon. The conditioning frame is excluded from the reported metrics, and all coordinates are represented in nanometers.

Evaluation Protocol. We evaluate all methods over the same physical prediction horizon and exclude the conditioning frame at $t = 0$. Each metric is computed for every paired generated and

Table 5: **Atom composition of the evaluated systems.** For each system, we report the protein and ligand atom counts together with their total, using dashes when no ligand is present. Across the selected conformations of each PED protein, the atom count remains unchanged.

Dataset	System	Atom counts		
		Protein	Ligand	Total
ATLAS	1KNT	440	—	440
	2FMA	475	—	475
	2PST	513	—	513
PED	2EC3	509	—	509
	6UCK	520	—	520
MISATO	5YJ8	504	14	518
	5EPK	484	33	517
	6V1C	404	26	430

reference trajectory, and we report the average across the evaluated trajectory pairs. The same metric definitions are applied to all compared methods. All methods except ALPHAFOLDING produce $1{,}024$ future frames, whereas ALPHAFOLDING produces 16 frames. For ALPHAFOLDING, we sample the reference trajectory at the corresponding output times and evaluate the matched reference–prediction pairs. The remaining methods are evaluated over the full $1{,}024$ frame trajectories.

Evaluation Metrics. We compare the coordinate component of the generated dynamics $\mathbf{X}_t$ with the reference coordinates $\mathbf{x}_t$. All metrics are implemented within the same evaluation pipeline and adapted to paired trajectories, consistent atom selections, and the alignment procedures described below.

Coordinate Fidelity. We measure coordinate accuracy with Cartesian MSE and RMSD. MSE is computed directly in the original coordinate frame, whereas RMSD is computed after frame-wise Cα alignment using the Kabsch algorithm (Kabsch, 1976). Our RMSD implementation follows MDTraj[2]. We report MSE in nm^2 and RMSD in nm.

Residue-Level Fidelity. We compare the Cα RMSF profiles of the generated and reference dynamics to evaluate residue-wise structural fluctuations (Hünenberger et al., 1995). RMSF Error reports the mean absolute difference between the two profiles, following the implementation in MDAnalysis[3].

We further evaluate structural agreement with residue-lDDT based on the local Distance Difference Test (Mariani et al., 2013) and its public implementation in Biotite[4]. We use a local cutoff of $1.5\,\mathrm{nm}$ and exclude residue pairs separated by fewer than three sequence positions.

Contact RMSE compares the occupancies of nonlocal Cα contacts defined with a $0.8\,\mathrm{nm}$ distance cutoff. Contact occupancy is the fraction of evaluated frames in which each pair lies below this cutoff. Pairwise distances follow the contact routines in MDTraj, while the trajectory-level occupancy error is computed within our evaluation pipeline.

Coupled Residue Dynamics. We evaluate correlated Cα motions with the dynamic cross-correlation matrix (Ichiye & Karplus, 1991). DCCM Error reports the mean absolute difference between the generated and reference matrices over their strict upper triangles. We implement this metric directly from its definition.

Ligand–Pocket Fidelity. For protein–ligand systems, Ligand–protein Contact RMSE compares heavy-atom contact occupancies over reference-native contacts using a $0.45\,\mathrm{nm}$ distance cutoff. We further evaluate ligand flexibility with pocket-aligned ligand RMSF Error.

The reference pocket is defined at $t = 0$ using protein heavy atoms within $0.6\,\mathrm{nm}$ of any ligand heavy atom. Ligand motion within the pocket is measured by the absolute difference between the standard

[2]MDTraj GitHub repository
[3]MDAnalysis GitHub repository
[4]Biotite GitHub repository

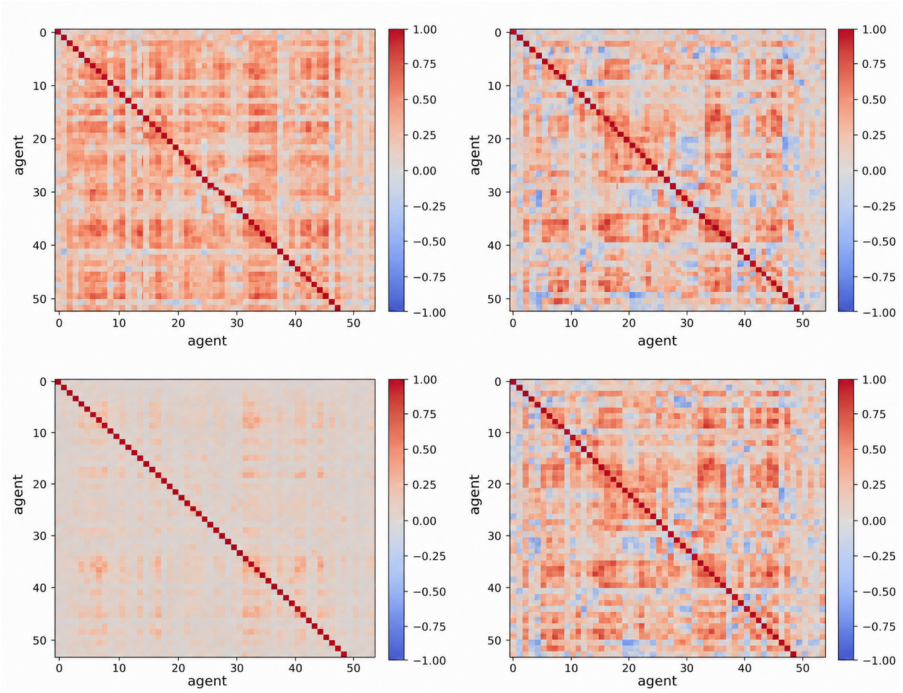

Figure 6: **Evolution of latent-state and residual similarities during cooperative training.** The top and bottom rows show pairwise centered cosine similarities of the latent states z_t and their residual updates. The left and right columns correspond to the 4K and 30K steps.

deviations of the radial fluctuations of the generated and reference ligand geometric centers. The specific ligand metrics are implemented directly within our evaluation pipeline.

Baseline Implementations. We compare our method with ALPHAFLOW, MDGEN, ALPHAFOLDING, UNISIM, PVB, and GLDP using their official public implementations. We preserve their original model architectures and inference procedures and modify only the input and output processing required for evaluation on the selected molecular systems.

- ALPHAFLOW: `https://github.com/bjing2016/alphaflow`
- MDGEN: `https://github.com/bjing2016/mdgen`
- ALPHAFOLDING: `https://github.com/fudan-generative-vision/dynamicPDB/tree/main/applications/4d_diffusion`
- UNISIM: `https://github.com/yaledeus/UniSim`
- PVB: `https://github.com/yaledeus/PVB`
- GLDP: `https://github.com/adityasengar/GLDP`

B.2 ADDITIONAL EXPERIMENTS

Evolution of Agent Representations. Figure 6 compares pairwise centered cosine similarities of the residue agent latent states and their residual updates at the 4K and 30K learning steps. For each representation h_i, we subtract the coalition mean $\bar{h} = N^{-1}\sum_{\ell=1}^{N} h_\ell$ and compute $(h_i - \bar{h})^\top (h_j - \bar{h})/(\|h_i - \bar{h}\|_2 \|h_j - \bar{h}\|_2)$. At 4K steps, the latent states contain both positive and negative pairwise similarities, while the residual updates form several block-structured regions with similar directions. At 30K steps, the latent state similarities shift toward positive values, whereas the off-diagonal similarities between residual updates become weaker. These results show that common update directions are more pronounced across subsets of agents during early training, while later residual updates retain greater agent-wise variation.

Learned Residue Interactions. Figure 7 visualizes the learned pairwise interaction weights A_{ij} for 6V1C, 5EPK, and 5YJ8 from MISATO. The weights vary across residue pairs and exhibit distinct

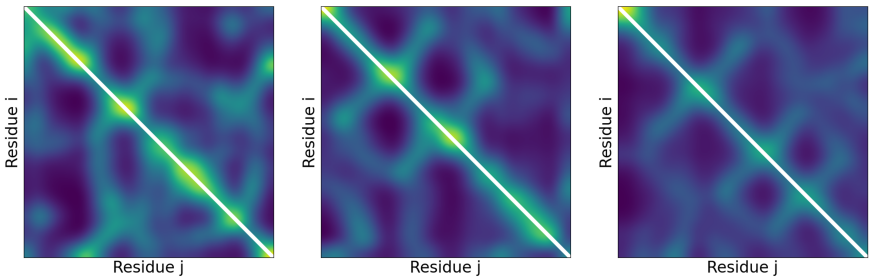

Figure 7: **Learned residue-wise interaction maps on MISATO.** We visualize the pairwise attention weights A_{ij} for 6V1C, 5EPK, and 5YJ8 from left to right. Brighter off-diagonal regions indicate stronger learned interactions between residue agents, while the diagonal is masked to exclude self interactions.

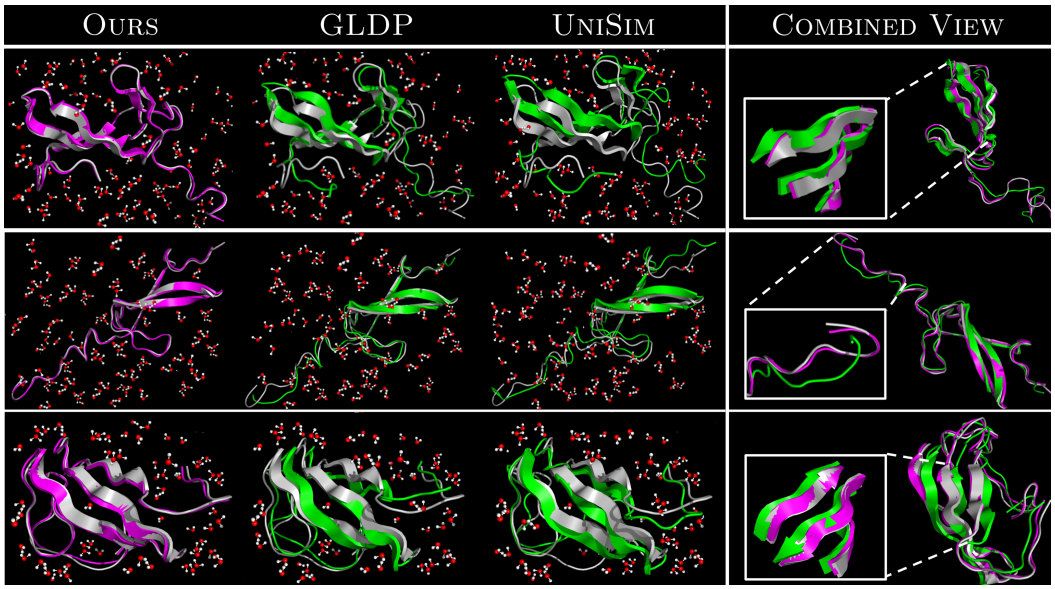

Figure 8: **Qualitative comparison of protein structures for PED.** We visualize three distinct conformations of 2EC3 from the PED dataset, one in each row. *(Left)* Our method (magenta) and prior works (green) are separately overlaid on the ground truth (gray). *(Right)* We show the ground truth, our method, and GLDP together, accompanied by magnified insets corresponding to the boxed regions shown for each protein.

patterns for the three molecular systems. Localized off-diagonal regions correspond to residue pairs assigned larger interaction weights, whereas weaker regions receive smaller pairwise weights. These differences show that the learned residue agent coupling is system dependent rather than fixed across molecular structures. As A_{ij} is an internal model quantity, the maps characterize learned interaction patterns without directly identifying physical residue contacts.

Visualization of PED. We visualize protein dynamics on PED, whose protein systems contain more heterogeneous and flexible conformational variations. Figure 8 compares the dynamics of our framework with GLDP and UNISIM, on diverse conformations of 2EC3. Our method closely predicts the ground truth dynamics accurately capturing the overall protein structure and fluctuation of residues. The combined visualization further shows that our framework better preserves residue motions throughout the trajectory.

B.3 Mathematical Preliminaries

Before presenting the proofs, we introduce the mathematical ingredients required for the subsequent theoretical analysis. We first formulate the gradient system of forward-backward stochastic differential equations underlying the two-stage fictitious play and then establish the policy-space framework used to certify an approximate response equilibrium over the generated local policy classes.

B.3.1 Gradient System of Forward-Backward Stochastic Differential Equations

The two-stage fictitious play introduced in the main text is driven by the value information obtained from the FBSDE representation. For the subsequent approximate-equilibrium analysis, it is therefore useful to describe the controlled stochastic dynamics and the resulting policy updates within a unified system. The following definition formalizes this construction.

Definition B.1. ***(Gradient System of FBSDEs).*** *For the fixed parameter profile $\bar{\boldsymbol{\theta}}^k = (\bar{\theta}_1^k, \dots, \bar{\theta}_N^k)$ at gradient iteration k, we consider a family of processes $\left(\mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}, \mathbf{Y}_t^{k,\bar{\boldsymbol{\alpha}}^k}, \mathbf{Z}_t^{k,\bar{\boldsymbol{\alpha}}^k}\right)$ that solves the following forward-backward stochastic differential system:*

$$
\begin{aligned}
d\mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k} &= \mathrm{B}\left(t, \mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}, \mathbf{u}_t^k\right) dt + \Sigma\, d\mathbf{W}_t, \\
d\mathbf{Y}_t^{k,\bar{\boldsymbol{\alpha}}^k} &= -\ell\left(t, \mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}, \mathbf{u}_t^k; \mathbf{x}_t\right) dt + \left(\mathbf{Z}_t^{k,\bar{\boldsymbol{\alpha}}^k}\right)^\top d\mathbf{W}_t, \\
\widetilde{\boldsymbol{\alpha}}^{k+1} &:= \boldsymbol{\alpha}\left(t, \mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}; \bar{\boldsymbol{\theta}}^k - \mathbf{G}_k\right) = \boldsymbol{\alpha}\left(t, \mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}; \widetilde{\boldsymbol{\theta}}^{k+1}\right), \\
\bar{\theta}_i^{k+1} &:= (1-\lambda_i^k)\bar{\theta}_i^k + \lambda_i^k \widetilde{\theta}_i^{k+1}, \quad \bar{\alpha}_i^{k+1} := \alpha_i\left(t, \mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}; \bar{\theta}_i^{k+1}\right), \quad i = 1, \dots, N. \qquad (18)
\end{aligned}
$$

where we denote $\mathbf{u}_t^k = \bar{\boldsymbol{\alpha}}^k\left(t, \mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}\right)$, $\eta_k > 0$ is the gradient step size, and λ_i^k is the agent-wise assimilation coefficient. The aggregated FBSDE direction $\mathbf{G}_k = (\mathbf{G}_{1,k}, \dots, \mathbf{G}_{N,k})$ is defined in Eq. 22. Then,

$$
\left(\mathbf{Y}_t^{k,\bar{\boldsymbol{\alpha}}^k}, \mathbf{Y}_T^{k,\bar{\boldsymbol{\alpha}}^k}, \mathbf{Z}_t^{k,\bar{\boldsymbol{\alpha}}^k}\right) = \left(\mathcal{J}_{t,\mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}}\left(\bar{\boldsymbol{\alpha}}^k\right), g_T\left(\mathbf{X}_T^{k,\bar{\boldsymbol{\alpha}}^k}\right), \Sigma^\top \nabla_x \mathcal{J}_x\left(\bar{\boldsymbol{\alpha}}^k\right)\big|_{x=\mathbf{X}_t^{k,\bar{\boldsymbol{\alpha}}^k}}\right).
$$

Definition B.1 summarizes one complete iteration of the two-stage fictitious play:

$$
\bar{\boldsymbol{\alpha}}^k \quad \longmapsto \quad \left(\mathbf{X}^{k,\bar{\boldsymbol{\alpha}}^k}, \mathbf{Y}^{k,\bar{\boldsymbol{\alpha}}^k}, \mathbf{Z}^{k,\bar{\boldsymbol{\alpha}}^k}\right) \quad \longmapsto \quad \widetilde{\boldsymbol{\alpha}}^{k+1} \quad \longmapsto \quad \bar{\boldsymbol{\alpha}}^{k+1}.
$$

The forward process evolves the protein state under the current running coalition policy $\bar{\boldsymbol{\alpha}}^k$, whereas the backward pair evaluates the corresponding continuation cost along this trajectory. In particular, $\mathbf{Y}_t^{k,\bar{\boldsymbol{\alpha}}^k}$ coincides with the cost-to-go of the current policy, and $\mathbf{Z}_t^{k,\bar{\boldsymbol{\alpha}}^k}$ represents its state sensitivity through the diffusion-weighted gradient of $\mathcal{J}$.

The gradient of the backward value process is then used to construct the joint planner proposal $\widetilde{\boldsymbol{\alpha}}^{k+1}$. This proposal is not adopted directly by every residue agent. Instead, each parameter is incorporated into the running profile according to the agent-specific coefficient λ_i^k, producing $\bar{\boldsymbol{\alpha}}^{k+1}$. Thus, the system couples coalition-level policy improvement with agent-wise adjustment and provides the iterative stochastic dynamics on which the subsequent approximate-equilibrium analysis is based.

B.3.2 Restricted Approximate-Response Certificate

This subsection connects the two-stage fictitious-play dynamics to an approximate-response guarantee. Although the algorithm updates neural parameters, the certificate must be assessed through the induced feedback policies and their agent-wise deviations. We therefore formulate the cost functional and response gap in policy space and describe the two stages through separate operators. The first operator constructs a local response class from the FBSDE gradient, and the second incorporates this response into the running policy profile. The resulting guarantee is restricted to the generated response classes and does not assert a global Nash equilibrium over the full admissible policy spaces.

Accordingly, the response condition is stated directly for parameterized feedback policies

$$\alpha_i : [0,T] \times \mathbb{R}^d \times \Theta \to \mathrm{U}_i, \qquad u_t^i = \alpha_i(t, X_t; \theta_i),$$

where the realized control depends on both the trainable parameter θ_i and the controlled state X_t. Let $\mathcal{H}_i := L^2\left([0,T] \times \mathbb{R}^d \times \Theta, \mu_i; \mathbb{R}^{m_i}\right)$ be the Hilbert space of square-integrable parameterized feedback policies for agent i, equipped with the inner product

$$\langle f, g \rangle_{\mathcal{H}_i} := \int_{[0,T]\times\mathbb{R}^d\times\Theta} f(t,x;\theta)^\top g(t,x;\theta) d\mu_i(t,x,\theta), \quad \|f\|_{\mathcal{H}_i} = \sqrt{\langle f, f\rangle_{\mathcal{H}_i}}.$$

The admissible policy class $\mathcal{A}_i$ is then a U_i-valued subset of this ambient Hilbert space. The joint parameterized feedback profile and its associated parameter profile are, respectively,

$$\alpha = (\alpha_1, \ldots, \alpha_N) \in \mathcal{A} := \mathcal{A}_1 \times \cdots \times \mathcal{A}_N, \quad \theta = (\theta_1, \ldots, \theta_N) \in \Theta^N,$$

Then, we denote $X^{\alpha;t,x}$ as the controlled agent coalition trajectory starting from $X_t = x$ and evolving according to Eq. 2:

$$dX_s^{\alpha;t,x} = B\left(s, X_s^{\alpha;t,x}, u_s^\alpha\right) ds + \Sigma dW_s, \;\; u_s^\alpha = \left(\alpha_1(s, X_s^{\alpha;t,x}; \theta_1), \ldots, \alpha_N(s, X_s^{\alpha;t,x}; \theta_N)\right).$$

The coalition objective can then be written as the policy functional

$$\mathcal{J}_x(\alpha) := \mathbb{E}\left[\int_t^T \ell\left(s, X_s^{\alpha;t,x}, u_s^\alpha; y_s\right) ds + g_T\left(X_T^{\alpha;t,x}\right) \middle| X_t^{\alpha;t,x} = x\right]. \tag{19}$$

For a deviation $\beta_i : [0,T] \times \mathbb{R}^d \times \Theta \to \mathrm{U}_i$ in $\mathcal{A}_i$, write

$$(\alpha^{-i}, \beta_i) := (\alpha_1, \ldots, \alpha_{i-1}, \beta_i, \alpha_{i+1}, \ldots, \alpha_N).$$

Accordingly, $\mathcal{J}_x(\alpha^{-i}, \beta_i)$ denotes the team cost when agent i deviates to β_i, evaluated through $u_s^i = \beta_i(s, X_s; \theta_i)$, while all other agents keep α^{-i} fixed. For a prescribed response class $\mathcal{R}_i \subseteq \mathcal{A}_i$, define the response gap and its agent-wise maximum by

$$\mathsf{Gap}_{i,t,x}(\alpha) := \mathcal{J}_x(\alpha) - \inf_{\beta_i \in \mathcal{R}_i} \mathcal{J}_x(\alpha^{-i}, \beta_i), \quad \mathsf{Gap}_{t,x}(\alpha) := \max_{1\le i\le N} \mathsf{Gap}_{i,t,x}(\alpha). \tag{20}$$

We say that α is an ε-approximate response equilibrium relative to $(\mathcal{R}_i)_{i=1}^N$ if

$$\mathsf{Gap}_{t,x}(\alpha) \le \varepsilon. \iff \mathcal{J}_x(\alpha) \le \mathcal{J}_x(\alpha^{-i}, \beta_i) + \varepsilon, \quad \forall \beta_i \in \mathcal{R}_i. \tag{21}$$

The full Nash gap is recovered only in the special case $\mathcal{R}_i = \mathcal{A}_i$. In the results below, $\mathcal{R}_i = \mathcal{C}_{i,K}$ is the retained gradient-generated response class; hence no conclusion is claimed for deviations outside $\mathcal{C}_{i,K}$.

Gradient System and Fictitious Play. We now describe the two operators that generate and assimilate the local action sets during fictitious play. At iteration k, let the current feedback profile and its associated parameter profile be

$$\bar{\alpha}^k = (\bar{\alpha}_1^k, \ldots, \bar{\alpha}_N^k), \qquad \bar{\theta}^k = (\bar{\theta}_1^k, \ldots, \bar{\theta}_N^k).$$

The FBSDE system in Definition B.1 produces the backward process, whose parameter-gradient signal defines the first-stage proposal

$$\widetilde{\theta}_i^{k+1} = \bar{\theta}_i^k - \mathbf{G}_{i,k}, \qquad i = 1, \ldots, N.$$

Let $\rho_k \in \mathcal{P}([0,T])$ be the time-sampling measure used to aggregate the backward FBSDE gradient, for example $\rho_k(dt) = \frac{1}{T}dt$ or the empirical measure on the discretization grid. Assume that the following quantity is finite

$$\int_0^T \mathbb{E}\left[\left\|\nabla_{\theta_i} \mathbf{Y}_t^{k,\bar{\alpha}^k}\right\|_2\right] \rho_k(dt) < \infty.$$

We define the aggregated first-stage FBSDE gradient for agent i by

$$\mathbf{G}_{i,k} := \eta_k \int_0^T \mathbb{E}\left[\nabla_{\theta_i} \mathbf{Y}_t^{k,\bar{\alpha}^k}\right] \rho_k(dt) \in \mathbb{R}^{p_i}. \tag{22}$$

The induced Hilbert-space radius is then defined as

$$r_{i,k} := \mathrm{Lip}_{\alpha,i} \|\mathbf{G}_{i,k}\|_2, \quad \|\alpha_i(\cdot,\cdot;\theta_i) - \alpha_i(\cdot,\cdot;\theta_i')\|_{\mathcal{H}_i} \le \mathrm{Lip}_{\alpha,i} \|\theta_i - \theta_i'\|_2,$$

where $\mathrm{Lip}_{\alpha,i}$ is a local parameter-to-policy Lipschitz constant for θ_i, θ_i' in the local parameter neighborhood considered at stage k. Then, we define the *first-stage operator* (*i.e.*, Φ) as the set-valued map

$$\boxed{\Phi_{i,k} : \mathcal{A} \times \Theta^N \rightrightarrows \mathcal{A}_i, \quad \Phi_{i,k}\left(\bar{\alpha}^k, \bar{\theta}^k\right) := \left\{\beta_i \in \mathcal{A}_i : \left\|\beta_i - \bar{\alpha}_i^k\right\|_{\mathcal{H}_i} \le r_{i,k}\right\}.} \tag{23}$$

By definition, the first stage restricts the response search to the local FBSDE-gradient ball generated by backward signals. Specifically, we write

$$\widetilde{\mathcal{A}}_i^{k+1} := \Phi_{i,k}\left(\bar{\alpha}^k, \bar{\theta}^k\right) \subseteq \mathcal{A}_i, \quad \widetilde{\alpha}_i^{k+1} = \alpha_i\left(t, \mathbf{X}_t^{k,\bar{\alpha}^k}; \widetilde{\theta}_i^{k+1}\right) \in \widetilde{\mathcal{A}}_i^{k+1}. \tag{24}$$

The *second-stage operator* (*i.e.*, Ψ) assimilates the first-stage response into the fictitious profile. It is defined as

$$\boxed{\Psi_{i,k} : \mathcal{A}_i \times 2^{\mathcal{A}_i} \rightrightarrows \mathcal{A}_i, \ \Psi_{i,k}\left(\bar{\alpha}_i^k, \widetilde{\mathcal{A}}_i^{k+1}\right) := \left\{\Pi_{\mathcal{A}_i}^{\mathcal{H}}\left((1-\lambda_i^k)\bar{\alpha}_i^k + \lambda_i^k \beta_i\right) : \beta_i \in \widetilde{\mathcal{A}}_i^{k+1}\right\}.} \tag{25}$$

Here $\lambda_i^k \in [0,1]$ is the assimilation coefficient and $\Pi_{\mathcal{A}_i}^{\mathcal{H}}$ denotes the Hilbert-space projection onto $\mathcal{A}_i$. If $\mathcal{A}_i$ is convex and the interpolation remains inside $\mathcal{A}_i$, this projection is the identity. The second-stage action set can therefore be written in the equivalent forms

$$\mathcal{A}_i^{k+1} := \Psi_{i,k}\left(\bar{\alpha}_i^k, \widetilde{\mathcal{A}}_i^{k+1}\right) \longleftrightarrow \mathcal{A}_i^{k+1} = \Psi_{i,k}\left(\bar{\alpha}_i^k, \Phi_{i,k}\left(\bar{\alpha}^k, \bar{\theta}^k\right)\right). \tag{26}$$

The realized second-stage fictitious policy satisfies $\bar{\alpha}_i^{k+1} \in \mathcal{A}_i^{k+1}$. In the parameterized representation, this is exactly

$$\bar{\theta}_i^{k+1} = (1-\lambda_i^k)\bar{\theta}_i^k + \lambda_i^k \widetilde{\theta}_i^{k+1}, \qquad \bar{\alpha}_i^{k+1} = \alpha_i\left(t, \mathbf{X}_t^{k,\bar{\alpha}^k}; \bar{\theta}_i^{k+1}\right).$$

B.3.3 Standing Assumptions

Before turning to the proofs, we state the regularity conditions used in the analysis. These conditions are imposed only on the parameter regions visited by the algorithm and on the retained response classes. We do not require global convexity of the neural parameter space, and we do not control deviations over the full admissible classes $\mathcal{A}_i$.

Local regularity of the policy update. For every agent i and iteration k, let

$$\mathcal{F}_i^k(\theta_i) := \mathcal{J}_x\left(\bar{\alpha}_{-i}^k, \alpha_i(\cdot,\cdot;\theta_i)\right), \quad \mathsf{S}_{i,k} := \left\{(1-s)\bar{\theta}_i^k + s\widetilde{\theta}_i^{k+1} : s \in [0,1]\right\}.$$

We assume that $\mathsf{S}_{i,k}$ lies in a convex local region $\Theta_{i,k}^{\mathrm{loc}}$. On this region, $\mathcal{F}_i^k$ is continuously differentiable and convex, with an L_i Lipschitz gradient:

$$\left\|\nabla_{\theta_i}\mathcal{F}_i^k(\theta_i) - \nabla_{\theta_i}\mathcal{F}_i^k(\theta_i')\right\|_2 \le L_i \|\theta_i - \theta_i'\|_2, \quad \theta_i, \theta_i' \in \Theta_{i,k}^{\mathrm{loc}}. \tag{27}$$

The aggregated FBSDE direction must also provide sufficient descent at the current iterate:

$$\left\langle \nabla_{\theta_i}\mathcal{F}_i^k(\bar{\theta}_i^k), \mathbf{G}_{i,k}\right\rangle \ge \frac{1}{\eta_k}\|\mathbf{G}_{i,k}\|_2^2, \quad 0 < \eta_k \le \frac{2}{L_i}. \tag{28}$$

We further assume that the local parameter to policy Lipschitz condition used to define $r_{i,k}$ holds on the same neighborhood. The inequality above does not identify every FBSDE signal with the exact objective gradient. It becomes an equality when the time sampling measure is concentrated at $t = 0$ and the FBSDE signal equals the exact parameter gradient of $\mathcal{J}_x$.

Initial distribution and FBSDE regularity. Let $\mathbf{X}_0 \sim \mathbb{P}_{\text{init}}$ have mean $\mathbf{m}_0$. We assume that the centered initial distribution is sub-Gaussian with parameter K_0:

$$\sup_{\|v\|_2=1} \|\langle v, \mathbf{X}_0 - \mathbf{m}_0\rangle\|_{\psi_2(\mathbb{P}_{\text{init}})} \le \mathsf{K}_0. \tag{29}$$

For each retained class $\mathcal{C}_{i,K}$, let $\mathcal{Z}_{i,K} \subset [0,T] \times \mathbb{R}^d$ be a compact domain on which the policies are evaluated. We use the uniform metric

$$\rho_i(\gamma, \gamma') := \sup_{z \in \mathcal{Z}_{i,K}} \|\gamma(z) - \gamma'(z)\|_2.$$

We assume that $(\mathcal{C}_{i,K}, \rho_i)$ is separable. We also assume that the local interpolation hull of $\mathcal{C}_{i,K}$ is admissible. The diffusion matrix does not depend on the policy, and the state and adjoint equations are well posed and differentiable with respect to the interpolation parameter λ. For $\gamma^\lambda = (1-\lambda)\gamma + \lambda\gamma'$, let X^λ denote the corresponding state. Along this interpolation, define

$$A_s^\lambda := \nabla_x \left[B(s,x,u^\lambda(s,x))\right]_{x=X_s^\lambda}, \quad c_s^\lambda := \nabla_x \left[\ell(s,x,u^\lambda(s,x);y_s)\right]_{x=X_s^\lambda}.$$

We assume that there are finite constants $L_B^x, L_B^{u_i}, L_\ell^x, L_\ell^{u_i}, L_g$, and C_X such that the following bounds hold uniformly over $\lambda \in [0,1]$:

$$\begin{aligned} &\|A_s^\lambda\|_{\text{op}} \le L_B^x, \quad \|B_{u_i}(s, X_s^\lambda, u^\lambda)\|_{\text{op}} \le L_B^{u_i}, \quad \|c_s^\lambda\| \le L_\ell^x(1 + \|X_s^\lambda\|), \\ &\|\ell_{u_i}(s, X_s^\lambda, u^\lambda; y_s)\| \le L_\ell^{u_i}(1 + \|X_s^\lambda\|), \quad \|\nabla g_T(X_T^\lambda)\| \le L_g(1 + \|X_T^\lambda\|), \\ &\sup_{0 \le s \le T} \mathbb{E}_x[1 + \|X_s^\lambda\|] \le C_X(1 + \|x\|). \end{aligned} \tag{30}$$

Neural policy class and covering numbers. Each retained policy is represented by a feedforward network F_{θ_i} with L_i layers and P_i trainable parameters. The activation functions satisfy $\sigma_{i,\ell}(0) = 0$ and are Lipschitz with constants $\varrho_{i,\ell}$. The weights satisfy

$$\|W_{i,\ell}\|_{\text{op}} \le s_{i,\ell}, \quad \|W_{i,\ell}^\top\|_{2,1} \le b_{i,\ell}, \quad w_{i,\ell} \le W_i, \qquad 1 \le \ell \le L_i. \tag{31}$$

We assume that $\mathcal{C}_{i,K}$, together with every parameter segment joining a retained parameter to $\bar{\theta}_i^K$, remains in this constrained neural parameter space. Under the same metric ρ_i, we use the following spectral covering estimate for some finite $\mathsf{C}_{i,K}^{\text{NN}}$:

$$\log N(\varepsilon, \mathcal{C}_{i,K}, \rho_i) \le \frac{\left(\mathsf{C}_{i,K}^{\text{NN}}\right)^2}{\varepsilon^2}, \qquad 0 < \varepsilon \le \mathsf{D}_{i,K}. \tag{32}$$

Here $\mathsf{D}_{i,K}$ is the policy diameter controlled by the gradient updates, as derived in the proof. The neural network conditions above provide both parameter and spectral covering estimates. They do not identify the policy metric with the Euclidean parameter norm.

Localization as the iteration count increases. The finite K certificate does not require an asymptotic condition. To obtain a vanishing bound as K increases, we assume that the retention scheme satisfies

$$\lim_{K \to \infty} \max_{1 \le i \le N} \left\{ \mathsf{C}_{i,T} \mathsf{D}_{i,K} \sqrt{P_i} \right\} = 0. \tag{33}$$

The regularity and covering conditions above give the finite K certificate. The final localization condition is used only to conclude that $\varepsilon_K \to 0$. Because this condition concerns $\mathcal{C}_{i,K}$, the resulting limit does not imply convergence of the global Nash gap over $\mathcal{A}_i$.

B.4 PROOFS

The purpose of this section is to derive a probabilistic certificate of an approximate response equilibrium over the gradient-generated local policy classes introduced in Proposition 3.2. The result is restricted to these generated response classes and does not assert a global Nash equilibrium over the full admissible policy spaces.

Proof Strategy. We first establish the descent property of the two-stage update and identify a retained response class formed by certified gradient segments in Proposition B.2. We then use the FBSDE sensitivity system to represent agent-wise cost deviations as a sub-Gaussian process indexed by this class and apply Dudley's chaining argument to bound their expected supremum by a policy-space entropy integral in Proposition B.4. Next, Proposition B.5 controls this entropy integral through neural-network covering-number estimates, and Proposition B.6 optimizes the resulting bound over the entropy splitting scale. Finally, combining these results yields the explicit expected restricted response-gap certificate under the initial-state distribution stated in Proposition B.7.

Proposition B.2 (Two-Stage Descent and Retained Response Class)**.** *Under the local update conditions specified in the proof, the first- and second-stage policies satisfy*

$$\mathcal{J}_x\big(\bar{\alpha}_{-i}^k, \widetilde{\alpha}_i^{k+1}\big) \le \mathcal{J}_x\big(\bar{\alpha}_{-i}^k, \bar{\alpha}_i^k\big) - \left(\frac{1}{\eta_k} - \frac{L_i}{2}\right) \|\mathbf{G}_{i,k}\|_2^2, \tag{34}$$

$$\mathcal{J}_x\big(\bar{\alpha}_{-i}^k, \bar{\alpha}_i^{k+1}\big) \le \mathcal{J}_x\big(\bar{\alpha}_{-i}^k, \bar{\alpha}_i^k\big) - \lambda_i^k \left(\frac{1}{\eta_k} - \frac{L_i}{2}\right) \|\mathbf{G}_{i,k}\|_2^2 \le \mathcal{J}_x\big(\bar{\alpha}_{-i}^k, \bar{\alpha}_i^k\big). \tag{35}$$

Moreover, the certified gradient segment obeys

$$\mathcal{S}_{i,k} := \big\{\alpha_i\left(\cdot, \cdot; \bar{\theta}_i^k - a\mathbf{G}_{i,k}\right) : a \in [0,1]\big\} \subseteq \widetilde{\mathcal{A}}_i^{k+1}. \tag{36}$$

Every policy in the retained class $\mathcal{C}_{i,K}$ constructed in the proof therefore has the parameter representation

$$\theta_i^\gamma = \bar{\theta}_i^0 - \sum_{k=0}^{K-1} a_{i,k}(\gamma)\mathbf{G}_{i,k}, \qquad 0 \le a_{i,k}(\gamma) \le 1. \tag{37}$$

Proof. Fix an agent i and an iteration k. For the current opponent profile $\bar{\alpha}_{-i}^k$, define

$$\bar{\alpha}^k := (\bar{\alpha}_i^k, \bar{\alpha}_{-i}^k), \;\; \widetilde{\mathcal{A}}_i^{k+1} := \Phi_{i,k}(\bar{\alpha}^k, \bar{\theta}^k), \;\; \mathcal{A}_i^{k+1} := \Psi_{i,k}\big(\bar{\alpha}_i^k, \widetilde{\mathcal{A}}_i^{k+1}\big), \;\; \mathcal{A}_i^0 := \{\bar{\alpha}_i^0\}. \tag{38}$$

The realized first- and second-stage policies are

$$\widetilde{\alpha}_i^{k+1} := \alpha_i(\cdot, \cdot; \widetilde{\theta}_i^{k+1}) \in \widetilde{\mathcal{A}}_i^{k+1}, \quad \bar{\alpha}_i^{k+1} := \alpha_i(\cdot, \cdot; \bar{\theta}_i^{k+1}) \in \mathcal{A}_i^{k+1}, \tag{39}$$

where $\widetilde{\theta}_i^{k+1} := \bar{\theta}_i^k - \mathbf{G}_{i,k}$, $\bar{\theta}_i^{k+1} := (1-\lambda_i^k)\bar{\theta}_i^k + \lambda_i^k \widetilde{\theta}_i^{k+1}$, and $\lambda_i^k \in [0,1]$. For notational brevity, define the parameterized objective

$$\mathcal{F}_i^k(\theta_i) := \mathcal{J}_x\big(\bar{\alpha}_{-i}^k, \alpha_i(\cdot, \cdot; \theta_i)\big). \tag{40}$$

Let the following denote the parameter segment generated by the first-stage update:

$$\mathsf{S}_{i,k} := \Big\{(1-s)\bar{\theta}_i^k + s\widetilde{\theta}_i^{k+1} : s \in [0,1]\Big\} \tag{41}$$

We assume that $\mathsf{S}_{i,k}$ is contained in a local convex parameter region $\Theta_{i,k}^{\mathrm{loc}}$ on which $\mathcal{F}_i^k$ is locally convex and continuously differentiable, and its gradient satisfies

$$\left\|\nabla_{\theta_i}\mathcal{F}_i^k(\theta_i) - \nabla_{\theta_i}\mathcal{F}_i^k(\theta_i')\right\|_2 \le L_i \left\|\theta_i - \theta_i'\right\|_2, \qquad \theta_i, \theta_i' \in \Theta_{i,k}^{\mathrm{loc}}. \tag{42}$$

Assume that the aggregated FBSDE direction satisfies

$$\left\langle \nabla_{\theta_i}\mathcal{F}_i^k(\bar{\theta}_i^k), \mathbf{G}_{i,k}\right\rangle \ge \frac{1}{\eta_k}\|\mathbf{G}_{i,k}\|_2^2, \qquad 0 < \eta_k \le \frac{2}{L_i}. \tag{43}$$

Thus, neither global convexity nor global smoothness over the entire parameter space is required. It is sufficient that these properties hold on the local region containing the parameter updates considered at iteration k. Recall that the first-stage parameter update is

$$\widetilde{\theta}_i^{k+1} = \bar{\theta}_i^k - \mathbf{G}_{i,k}, \tag{44}$$

where $\mathbf{G}_{i,k}$ is the aggregated FBSDE direction and satisfies Eq. 43. This condition holds with equality when the time-sampling measure is concentrated at $t = 0$ and the FBSDE signal equals the exact parameter gradient of $\mathcal{J}_x$. By the fundamental theorem of calculus along the segment $\mathsf{S}_{i,k}$, for $\mathsf{D}_{i,k} := \widetilde{\theta}_i^{k+1} - \bar{\theta}_i^k$, we have

$$\mathcal{F}_i^k(\bar{\theta}_i^k + \mathsf{D}_{i,k}) - \mathcal{F}_i^k(\bar{\theta}_i^k) = \int_0^1 \left\langle \nabla_{\theta_i} \mathcal{F}_i^k(\bar{\theta}_i^k + s\mathsf{D}_{i,k}), \mathsf{D}_{i,k} \right\rangle ds. \tag{45}$$

Adding and subtracting $\nabla_{\theta_i} \mathcal{F}_i^k(\bar{\theta}_i^k)$ inside the integral and using Eq. 42 gives the local descent lemma

$$\mathcal{F}_i^k(\widetilde{\theta}_i^{k+1}) \le \mathcal{F}_i^k(\bar{\theta}_i^k) + \left\langle \nabla_{\theta_i} \mathcal{F}_i^k(\bar{\theta}_i^k), \widetilde{\theta}_i^{k+1} - \bar{\theta}_i^k \right\rangle + \frac{L_i}{2} \left\| \widetilde{\theta}_i^{k+1} - \bar{\theta}_i^k \right\|_2^2. \tag{46}$$

Consequently,

$$\begin{aligned}
\mathcal{F}_i^k(\widetilde{\theta}_i^{k+1}) &\le \mathcal{F}_i^k(\bar{\theta}_i^k) + \left\langle \nabla_{\theta_i} \mathcal{F}_i^k(\bar{\theta}_i^k), \widetilde{\theta}_i^{k+1} - \bar{\theta}_i^k \right\rangle + \frac{L_i}{2} \|\widetilde{\theta}_i^{k+1} - \bar{\theta}_i^k\|_2^2 \\
&= \mathcal{F}_i^k(\bar{\theta}_i^k) + \left\langle \nabla_{\theta_i} \mathcal{F}_i^k(\bar{\theta}_i^k), -\mathbf{G}_{i,k} \right\rangle + \frac{L_i}{2} \|\mathbf{G}_{i,k}\|_2^2 \\
&\le \mathcal{F}_i^k(\bar{\theta}_i^k) - \frac{1}{\eta_k} \|\mathbf{G}_{i,k}\|_2^2 + \frac{L_i}{2} \|\mathbf{G}_{i,k}\|_2^2 \\
&= \mathcal{F}_i^k(\bar{\theta}_i^k) - \left(\frac{1}{\eta_k} - \frac{L_i}{2} \right) \|\mathbf{G}_{i,k}\|_2^2.
\end{aligned} \tag{47}$$

Substituting the definition of $\mathcal{F}_i^k$, together with

$$\widetilde{\alpha}_i^{k+1} = \alpha_i(\cdot, \cdot; \widetilde{\theta}_i^{k+1}), \qquad \bar{\alpha}_i^k = \alpha_i(\cdot, \cdot; \bar{\theta}_i^k), \tag{48}$$

gives

$$\begin{aligned}
\mathcal{J}_x\big(\bar{\alpha}_{-i}^k, \widetilde{\alpha}_i^{k+1}\big) &= \mathcal{F}_i^k(\widetilde{\theta}_i^{k+1}) \\
&\le \mathcal{F}_i^k(\bar{\theta}_i^k) - \left(\frac{1}{\eta_k} - \frac{L_i}{2} \right) \|\mathbf{G}_{i,k}\|_2^2 \\
&= \mathcal{J}_x\big(\bar{\alpha}_{-i}^k, \bar{\alpha}_i^k\big) - \left(\frac{1}{\eta_k} - \frac{L_i}{2} \right) \|\mathbf{G}_{i,k}\|_2^2.
\end{aligned} \tag{49}$$

This proves Eq. 34. Next, the second-stage parameter satisfies

$$\bar{\theta}_i^{k+1} = (1 - \lambda_i^k)\bar{\theta}_i^k + \lambda_i^k \widetilde{\theta}_i^{k+1}, \qquad \lambda_i^k \in [0, 1]. \tag{50}$$

In particular, $\bar{\theta}_i^{k+1} \in \mathsf{S}_{i,k} \subset \Theta_{i,k}^{\mathrm{loc}}$. Therefore, the local convexity assumption is sufficient to apply the convex interpolation inequality along this particular segment:

$$\begin{aligned}
\mathcal{F}_i^k(\bar{\theta}_i^{k+1}) &\le (1 - \lambda_i^k)\mathcal{F}_i^k(\bar{\theta}_i^k) + \lambda_i^k \mathcal{F}_i^k(\widetilde{\theta}_i^{k+1}) \\
&\le (1 - \lambda_i^k)\mathcal{F}_i^k(\bar{\theta}_i^k) + \lambda_i^k \left[\mathcal{F}_i^k(\bar{\theta}_i^k) - \left(\frac{1}{\eta_k} - \frac{L_i}{2} \right) \|\mathbf{G}_{i,k}\|_2^2 \right] \\
&= \mathcal{F}_i^k(\bar{\theta}_i^k) - \lambda_i^k \left(\frac{1}{\eta_k} - \frac{L_i}{2} \right) \|\mathbf{G}_{i,k}\|_2^2 \\
&\le \mathcal{F}_i^k(\bar{\theta}_i^k),
\end{aligned} \tag{51}$$

where the last inequality follows from the step-size constraint and $\lambda_i^k \ge 0$. Using

$$\bar{\alpha}_i^{k+1} = \alpha_i(\cdot, \cdot; \bar{\theta}_i^{k+1}), \qquad \bar{\alpha}_i^k = \alpha_i(\cdot, \cdot; \bar{\theta}_i^k), \tag{52}$$

we obtain

$$\begin{aligned}\mathcal{J}_x\big(\bar{\alpha}_{-i}^k, \bar{\alpha}_i^{k+1}\big) &= \mathcal{F}_i^k(\bar{\theta}_i^{k+1}) \\ &\le \mathcal{F}_i^k(\bar{\theta}_i^k) - \lambda_i^k \left(\frac{1}{\eta_k} - \frac{L_i}{2}\right) \|\mathbf{G}_{i,k}\|_2^2 \\ &= \mathcal{J}_x\big(\bar{\alpha}_{-i}^k, \bar{\alpha}_i^k\big) - \lambda_i^k \left(\frac{1}{\eta_k} - \frac{L_i}{2}\right) \|\mathbf{G}_{i,k}\|_2^2 \\ &\le \mathcal{J}_x\big(\bar{\alpha}_{-i}^k, \bar{\alpha}_i^k\big). \end{aligned} \tag{53}$$

This proves Eq. 35. We next verify that the certified segment belongs to the first-stage policy ball. Assume that the parameter-to-policy map is locally Lipschitz with constant $\mathrm{Lip}_{\alpha,i}$ and that $r_{i,k} = \mathrm{Lip}_{\alpha,i}\|\mathbf{G}_{i,k}\|_2$. For any $a \in [0,1]$, the local parameter-to-policy Lipschitz condition gives

$$\begin{aligned}&\left\|\alpha_i\left(\cdot, \cdot; \bar{\theta}_i^k - a\mathbf{G}_{i,k}\right) - \bar{\alpha}_i^k\right\|_{\mathcal{H}_i} \\ &\le \mathrm{Lip}_{\alpha,i}\left\|a\mathbf{G}_{i,k}\right\|_2 \le \mathrm{Lip}_{\alpha,i}\|\mathbf{G}_{i,k}\|_2 = r_{i,k}. \end{aligned} \tag{54}$$

Hence Eq. 36 follows directly from the definition of $\Phi_{i,k}$. For any retained index set $\mathcal{I}_{i,K} \subseteq \{0, \dots, K-1\}$, define

$$\mathcal{C}_{i,K} := \left\{\bar{\alpha}_i^K\right\} \cup \bigcup_{k \in \mathcal{I}_{i,K}} \mathcal{S}_{i,k}. \tag{55}$$

The recursive second-stage update can be written as

$$\bar{\theta}_i^{k+1} = \bar{\theta}_i^k - \lambda_i^k \mathbf{G}_{i,k}. \tag{56}$$

Iterating this identity yields

$$\bar{\theta}_i^k = \bar{\theta}_i^0 - \sum_{j=0}^{k-1} \lambda_i^j \mathbf{G}_{i,j}. \tag{57}$$

If $\gamma \in \mathcal{S}_{i,k}$, then for some $a \in [0,1]$,

$$\theta_i^\gamma = \bar{\theta}_i^k - a\mathbf{G}_{i,k} = \bar{\theta}_i^0 - \sum_{j=0}^{k-1} \lambda_i^j \mathbf{G}_{i,j} - a\mathbf{G}_{i,k}. \tag{58}$$

Thus Eq. 37 holds with $a_{i,j}(\gamma) = \lambda_i^j$ for $j < k$, $a_{i,k}(\gamma) = a$, and $a_{i,j}(\gamma) = 0$ for $j > k$. For $\gamma = \bar{\alpha}_i^K$, it holds with $a_{i,j}(\gamma) = \lambda_i^j$ for $j < K$. Since all these coefficients lie in $[0,1]$, the claimed representation follows for every $\gamma \in \mathcal{C}_{i,K}$. □

Corollary B.3 (Sub-Gaussian norm of the initial state)**.** *Under the initial-state condition specified in the proof,*

$$\|1 + \|\mathbf{X}_0\|\|_{\psi_2(\mathbb{P}_{\text{init}})} \leq \frac{1 + \|\mathbf{m}_0\|}{\sqrt{\log 2}} + 2\mathsf{K}_0\sqrt{\frac{\log(2 \cdot 5^{d_x})}{\log 2}}. \tag{59}$$

Proof. Let $\mathbf{X}_0 \in \mathbb{R}^{d_x}$ have distribution $\mathbb{P}_{\text{init}}$, and let $\mathbf{m}_0 := \mathbb{E}_{\mathbb{P}_{\text{init}}}[\mathbf{X}_0]$ be its mean. Assume that $\mathbb{P}_{\text{init}}$ is K_0-sub-Gaussian in the sense that

$$\sup_{\|v\|=1} \|\langle v, \mathbf{X}_0 - \mathbf{m}_0\rangle\|_{\psi_2(\mathbb{P}_{\text{init}})} \leq \mathsf{K}_0.$$

As a first step, let us define the centered initial state under $\mathbb{P}_{\text{init}}$ as $\widetilde{\mathbf{X}}_0 := \mathbf{X}_0 - \mathbf{m}_0$. Then, we choose a $1/2$-net $\mathcal{N}_{1/2}$ of the Euclidean unit sphere $\mathbb{S}^{d_x-1}$ satisfying $M := |\mathcal{N}_{1/2}| \leq 5^{d_x}$. Next, for every $x \in \mathbb{R}^{d_x}$, if $x \neq 0$, choose $v_x \in \mathcal{N}_{1/2}$ such that

$$\left\| v_x - \frac{x}{\|x\|} \right\| \leq \frac{1}{2}. \tag{60}$$

Then, one can find the direct upper bound of vector norm as

$$\begin{aligned} \|x\| &= \left\langle \frac{x}{\|x\|}, x \right\rangle = \langle v_x, x\rangle + \left\langle \frac{x}{\|x\|} - v_x, x \right\rangle \\ &\leq |\langle v_x, x\rangle| + \left\| \frac{x}{\|x\|} - v_x \right\| \|x\| \leq \max_{v \in \mathcal{N}_{1/2}} |\langle v, x\rangle| + \frac{1}{2}\|x\|. \end{aligned} \tag{61}$$

Rearranging these terms gives

$$\|x\| \leq 2 \max_{v \in \mathcal{N}_{1/2}} |\langle v, x\rangle|. \tag{62}$$

Applying Eq. 62 to $x = \widetilde{\mathbf{X}}_0$, define $\mathsf{M}_0 := \max_{v \in \mathcal{N}_{1/2}} \left|\left\langle v, \widetilde{\mathbf{X}}_0 \right\rangle\right|$. We then have $\left\|\widetilde{\mathbf{X}}_0\right\| \leq 2\mathsf{M}_0$. Thus, by the definition, for every $v \in \mathcal{N}_{1/2}$, one obtains

$$\mathbb{E}_{\mathbb{P}_{\text{init}}}\left[\exp\left(\frac{\left|\langle v, \widetilde{\mathbf{X}}_0\rangle\right|^2}{\mathsf{K}_0^2}\right)\right] \leq 2. \tag{63}$$

Consequently,

$$\begin{aligned} \mathbb{E}_{\mathbb{P}_{\text{init}}}\left[\exp\left(\frac{\mathsf{M}_0^2}{\mathsf{K}_0^2}\right)\right] &= \mathbb{E}_{\mathbb{P}_{\text{init}}}\left[\max_{v \in \mathcal{N}_{1/2}} \exp\left(\frac{\left|\langle v, \widetilde{\mathbf{X}}_0\rangle\right|^2}{\mathsf{K}_0^2}\right)\right] \\ &\leq \sum_{v \in \mathcal{N}_{1/2}} \mathbb{E}_{\mathbb{P}_{\text{init}}}\left[\exp\left(\frac{\left|\langle v, \widetilde{\mathbf{X}}_0\rangle\right|^2}{\mathsf{K}_0^2}\right)\right] \\ &\leq 2M. \end{aligned} \tag{64}$$

Let us define the M-dependent constant $a_M := \frac{\log(2M)}{\log 2}$. Since $a_M \geq 1$, the map $u \mapsto u^{1/a_M}$ is concave. Therefore, Jensen's inequality and Eq. 64 give

$$\begin{aligned}
\mathbb{E}_{\mathbb{P}_{\text{init}}}\left[\exp\left(\frac{\mathsf{M}_0^2}{a_M \mathsf{K}_0^2}\right)\right] &= \mathbb{E}_{\mathbb{P}_{\text{init}}}\left[\left\{\exp\left(\frac{\mathsf{M}_0^2}{\mathsf{K}_0^2}\right)\right\}^{1/a_M}\right] \\
&\leq \left\{\mathbb{E}_{\mathbb{P}_{\text{init}}}\left[\exp\left(\frac{\mathsf{M}_0^2}{\mathsf{K}_0^2}\right)\right]\right\}^{1/a_M} \\
&\leq (2M)^{1/a_M} \\
&= \exp\left(\frac{\log(2M)}{a_M}\right) \\
&= \exp(\log 2) \\
&= 2. \qquad (65)
\end{aligned}$$

Using the upper bound above, it follows that

$$\mathbb{E}_{\mathbb{P}_{\text{init}}}\left[\exp\left(\frac{\|\widetilde{\mathbf{X}}_0\|^2}{4a_M \mathsf{K}_0^2}\right)\right] \leq \mathbb{E}_{\mathbb{P}_{\text{init}}}\left[\exp\left(\frac{\mathsf{M}_0^2}{a_M \mathsf{K}_0^2}\right)\right] \leq 2. \qquad (66)$$

Hence, by the definition of the ψ_2-norm,

$$\left\|\|\widetilde{\mathbf{X}}_0\|\right\|_{\psi_2(\mathbb{P}_{\text{init}})} \leq 2\mathsf{K}_0\sqrt{a_M} = 2\mathsf{K}_0\sqrt{\frac{\log(2M)}{\log 2}} \leq 2\mathsf{K}_0\sqrt{\frac{\log(2\cdot 5^{d_x})}{\log 2}}. \qquad (67)$$

Moreover,

$$1 + \|\mathbf{X}_0\| = 1 + \left\|\mathbf{m}_0 + \widetilde{\mathbf{X}}_0\right\| \leq 1 + \|\mathbf{m}_0\| + \|\widetilde{\mathbf{X}}_0\|. \qquad (68)$$

By monotonicity and the triangle inequality of the ψ_2-norm,

$$\|1 + \|\mathbf{X}_0\|\|_{\psi_2(\mathbb{P}_{\text{init}})} \leq \|1 + \|\mathbf{m}_0\|\|_{\psi_2(\mathbb{P}_{\text{init}})} + \left\|\|\widetilde{\mathbf{X}}_0\|\right\|_{\psi_2(\mathbb{P}_{\text{init}})}. \qquad (69)$$

For the deterministic first term,

$$\begin{aligned}
&\mathbb{E}_{\mathbb{P}_{\text{init}}}\left[\exp\left(\frac{(1+\|\mathbf{m}_0\|)^2}{t^2}\right)\right] \leq 2 \\
&\iff \exp\left(\frac{(1+\|\mathbf{m}_0\|)^2}{t^2}\right) \leq 2 \\
&\iff \frac{(1+\|\mathbf{m}_0\|)^2}{t^2} \leq \log 2 \\
&\iff t \geq \frac{1+\|\mathbf{m}_0\|}{\sqrt{\log 2}}. \qquad (70)
\end{aligned}$$

$$\|1 + \|\mathbf{m}_0\|\|_{\psi_2(\mathbb{P}_{\text{init}})} = \frac{1+\|\mathbf{m}_0\|}{\sqrt{\log 2}}. \qquad (71)$$

Substituting this identity and Eq. 67 into Eq. 69 proves Eq. 59. □

Proposition B.4 (Sub-Gaussian Pontryagin–Dudley Certificate)**.** *There exist a universal constant $\mathsf{C}_D > 0$ and a deterministic constant $\mathsf{C}_{i,T} > 0$ such that*

$$\mathbb{E}_{\mathbf{X}_0\sim\mathbb{P}_{\text{init}}}\left[\sup_{\gamma\in\mathcal{C}_{i,K}}\left|\mathbf{Y}_0^{i,\bar{\alpha}}(\mathbf{X}_0) - \mathbf{Y}_0^{i,(\alpha_{-i},\gamma)}(\mathbf{X}_0)\right|\right] \leq \mathsf{C}_D\mathsf{C}_{i,T}\int_0^{\operatorname{diam}_{\rho_i}(\mathcal{C}_{i,K})}\sqrt{\log N\left(\varepsilon, \mathcal{C}_{i,K}, \rho_i\right)}d\varepsilon. \qquad (72)$$

Thus, its expected uniform FBSDE value deviation is controlled by the policy entropy of $\mathcal{C}_{i,K}$, rather than by a global response class.

Proof. Fix $\bar{\alpha} = (\alpha_{-i}, \bar{\alpha}_i)$ with $\bar{\alpha}_i \in \mathcal{C}_{i,K}$. Assume that $(\mathcal{C}_{i,K}, \rho_i)$ is separable, its local interpolation hull is admissible, $\mathbb{P}_{\text{init}}$ is sub-Gaussian, the diffusion matrix is policy independent, and the state and adjoint equations are well posed and differentiable with respect to λ for every interpolated policy between elements of $\mathcal{C}_{i,K}$. Assume also that the interpolated trajectories are evaluated on a compact computational domain $\mathcal{Z}_{i,K}$. Let $\mathcal{Z}_{i,K} \subset [0,T] \times \mathbb{R}^d$ denote the bounded time-state domain on which the generated policies are evaluated. For policies $\gamma, \gamma' \in \mathcal{C}_{i,K}$, define the policy-space metric

$$\rho_i(\gamma, \gamma') := \sup_{0 \le s \le T} \sup_{x:(s,x) \in \mathcal{Z}_{i,K}} \|\gamma(s,x) - \gamma'(s,x)\|. \tag{73}$$

The interpolation used below need not belong to the record class $\mathcal{C}_{i,K}$. It is sufficient that it belongs to the admissible local policy hull containing $\mathcal{C}_{i,K}$. Let us define

$$\gamma^\lambda := (1-\lambda)\gamma + \lambda\gamma', \qquad \lambda \in [0,1], \tag{74}$$

$$\delta\gamma := \gamma' - \gamma, \qquad \frac{d}{d\lambda}\gamma^\lambda = \delta\gamma \tag{75}$$

Then, the arbitrary deviated action policy driven by γ^λ can be described as

$$u^\lambda(s,x) := \left(\alpha_{-i}(s,x), \gamma^\lambda(s,x)\right), \tag{76}$$

and its corresponding protein dynamics X^λ solve

$$dX_s^\lambda = B\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda)\right) ds + \Sigma dW_s, \quad \mathbf{X}_0^\lambda = x. \tag{77}$$

For notational convenience, define

$$A_s^\lambda := \nabla_x \left[B\left(s, x, u^\lambda(s,x)\right)\right]_{x=X_s^\lambda}, \quad c_s^\lambda := \nabla_x \left[\ell\left(s, x, u^\lambda(s,x); y_s\right)\right]_{x=X_s^\lambda}. \tag{78}$$

Assume that, uniformly in $\lambda \in [0,1]$,

$$\begin{aligned} &\|A_s^\lambda\|_{\text{op}} \le L_B^x, \quad \|B_{u_i}(s, X_s^\lambda, u^\lambda)\|_{\text{op}} \le L_B^{u_i}, \quad \|c_s^\lambda\| \le L_\ell^x(1 + \|X_s^\lambda\|), \\ &\|\ell_{u_i}(s, X_s^\lambda, u^\lambda; y_s)\| \le L_\ell^{u_i}(1 + \|X_s^\lambda\|), \quad \|\nabla g_T(X_T^\lambda)\| \le L_g(1 + \|X_T^\lambda\|), \\ &\sup_{0 \le s \le T} \mathbb{E}_x[1 + \|X_s^\lambda\|] \le C_X(1 + \|x\|). \end{aligned} \tag{79}$$

The corresponding adjoint process satisfies

$$dP_s^\lambda = -\left[(A_s^\lambda)^\top P_s^\lambda + c_s^\lambda\right] ds + Q_s^\lambda dW_s, \quad P_T^\lambda = \nabla g_T(X_T^\lambda). \tag{80}$$

Let us further define the variation of the proteins state: $\xi_s^\lambda := \frac{d}{d\lambda} X_s^\lambda$. Since the diffusion matrix is independent of the policy, differentiating variation with respect to λ yields

$$d\xi_s^\lambda = \left[A_s^\lambda \xi_s^\lambda + B_{u_i}\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda)\right) \delta\gamma(s, X_s^\lambda)\right] ds, \quad \xi_0^\lambda = 0. \tag{81}$$

While the backward dynamics at time $t = 0$ is associated with the value function, we specify the value function with respect to the variation at arbitrary λ:

$$\begin{aligned} \mathcal{V}_i^\lambda(x) &:= \mathbf{Y}_0^{i,(\alpha_{-i}, \gamma^\lambda)}(x) \\ &= \mathbb{E}_x\left[\int_0^T \ell\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda); y_s\right) ds + g_T(X_T^\lambda)\right]. \end{aligned} \tag{82}$$

Differentiating the running cost gives

$$\begin{aligned} &\frac{d}{d\lambda}\ell\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda); y_s\right) \\ &= \left\langle c_s^\lambda, \xi_s^\lambda \right\rangle + \left\langle \ell_{u_i}\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda); y_s\right), \delta\gamma(s, X_s^\lambda) \right\rangle. \end{aligned} \tag{83}$$

Similarly,

$$\frac{d}{d\lambda} g_T(X_T^\lambda) = \left\langle \nabla g_T(X_T^\lambda), \xi_T^\lambda \right\rangle. \tag{84}$$

Consequently,

$$\frac{d}{d\lambda}\mathcal{V}_i^\lambda(x) = \mathbb{E}_x[\int_0^T \left\langle c_s^\lambda, \xi_s^\lambda \right\rangle ds] + \mathbb{E}_x\left[\int_0^T \left\langle \ell_{u_i}\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda); y_s\right), \delta\gamma(s, X_s^\lambda)\right\rangle ds\right] + \mathbb{E}_x\left[\left\langle \nabla g_T(X_T^\lambda), \xi_T^\lambda \right\rangle\right]. \tag{85}$$

We next eliminate the terms containing ξ^λ. Applying Itô's product rule to $\langle P_s^\lambda, \xi_s^\lambda \rangle$ gives

$$d\left\langle P_s^\lambda, \xi_s^\lambda \right\rangle = \left\langle dP_s^\lambda, \xi_s^\lambda \right\rangle + \left\langle P_s^\lambda, d\xi_s^\lambda \right\rangle + d[P^\lambda, \xi^\lambda]_s. \tag{86}$$

Because ξ^λ has no diffusion term, one can observe that $d[P^\lambda, \xi^\lambda]_s = 0$. Substituting Eq. 81, we obtain

$$\begin{aligned} d\left\langle P_s^\lambda, \xi_s^\lambda \right\rangle = &- \left\langle (A_s^\lambda)^\top P_s^\lambda, \xi_s^\lambda \right\rangle ds - \left\langle c_s^\lambda, \xi_s^\lambda \right\rangle ds \\ &+ \left\langle P_s^\lambda, A_s^\lambda \xi_s^\lambda \right\rangle ds \\ &+ \left\langle P_s^\lambda, B_{u_i}\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda)\right) \delta\gamma(s, X_s^\lambda) \right\rangle ds \\ &+ \left\langle Q_s^\lambda dW_s, \xi_s^\lambda \right\rangle. \end{aligned} \tag{87}$$

Since

$$\left\langle (A_s^\lambda)^\top P_s^\lambda, \xi_s^\lambda \right\rangle = \left\langle P_s^\lambda, A_s^\lambda \xi_s^\lambda \right\rangle, \tag{88}$$

the two drift-linearization terms cancel. Hence,

$$\begin{aligned} d\left\langle P_s^\lambda, \xi_s^\lambda \right\rangle = &- \left\langle c_s^\lambda, \xi_s^\lambda \right\rangle ds \\ &+ \left\langle B_{u_i}\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda)\right)^\top P_s^\lambda, \delta\gamma(s, X_s^\lambda) \right\rangle ds \\ &+ \left\langle Q_s^\lambda dW_s, \xi_s^\lambda \right\rangle. \end{aligned} \tag{89}$$

Integrating from 0 to T, using $\xi_0^\lambda = 0$, and taking expectations yields

$$\begin{aligned} &\mathbb{E}_x\left[\left\langle \nabla g_T(X_T^\lambda), \xi_T^\lambda \right\rangle\right] \\ &= -\mathbb{E}_x\left[\int_0^T \left\langle c_s^\lambda, \xi_s^\lambda \right\rangle ds\right] \\ &\quad + \mathbb{E}_x\left[\int_0^T \left\langle B_{u_i}\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda)\right)^\top P_s^\lambda, \delta\gamma(s, X_s^\lambda) \right\rangle ds\right]. \end{aligned} \tag{90}$$

Substituting this identity into Eq. 85 cancels the state-variation terms:

$$\begin{aligned} \frac{d}{d\lambda}\mathcal{V}_i^\lambda(x) &= \mathbb{E}_x[\int_0^T \langle \ell_{u_i}\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda); y_s\right) \\ &\qquad + B_{u_i}\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda)\right)^\top P_s^\lambda, \\ &\qquad \delta\gamma(s, X_s^\lambda)\rangle ds] \\ &= \mathbb{E}_x\left[\int_0^T \left\langle \nabla_{u_i}\mathfrak{H}\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda), P_s^\lambda\right), \delta\gamma(s, X_s^\lambda)\right\rangle ds\right]. \end{aligned} \tag{91}$$

This is the required stochastic maximum-principle representation of the policy derivative. We next estimate the adjoint process. Let $\Phi_{r,s}^\lambda$ be the state-transition matrix satisfying

$$\frac{\partial}{\partial r}\Phi_{r,s}^\lambda = A_r^\lambda \Phi_{r,s}^\lambda, \qquad \Phi_{s,s}^\lambda = \mathbf{I}. \tag{92}$$

For any $z \in \mathbb{R}^d$, we define $\eta^\lambda_{r,s}(z) := \Phi^\lambda_{r,s} z$ for $s \le r \le T$. Then, the vector $\eta^\lambda_{r,s}(z)$ satisfies

$$\frac{\partial}{\partial r}\eta^\lambda_{r,s}(z) = A^\lambda_r \eta^\lambda_{r,s}(z), \qquad \eta^\lambda_{s,s}(z) = z. \tag{93}$$

Therefore, the chain rule with respect to r gives

$$\begin{aligned}
\frac{\partial}{\partial r}\left\|\eta^\lambda_{r,s}(z)\right\|^2 &= \frac{\partial}{\partial r}\left\langle \eta^\lambda_{r,s}(z), \eta^\lambda_{r,s}(z)\right\rangle \\
&= \left\langle \frac{\partial}{\partial r}\eta^\lambda_{r,s}(z), \eta^\lambda_{r,s}(z)\right\rangle + \left\langle \eta^\lambda_{r,s}(z), \frac{\partial}{\partial r}\eta^\lambda_{r,s}(z)\right\rangle \\
&= \left\langle A^\lambda_r \eta^\lambda_{r,s}(z), \eta^\lambda_{r,s}(z)\right\rangle + \left\langle \eta^\lambda_{r,s}(z), A^\lambda_r \eta^\lambda_{r,s}(z)\right\rangle \\
&= 2\left\langle \eta^\lambda_{r,s}(z), A^\lambda_r \eta^\lambda_{r,s}(z)\right\rangle \\
&\le 2\left\|\eta^\lambda_{r,s}(z)\right\| \left\|A^\lambda_r \eta^\lambda_{r,s}(z)\right\| \\
&\le 2\left\|A^\lambda_r\right\|_{\mathrm{op}} \left\|\eta^\lambda_{r,s}(z)\right\|^2 \\
&\le 2L^x_B \left\|\eta^\lambda_{r,s}(z)\right\|^2.
\end{aligned} \tag{94}$$

Multiplying by the integrating factor $e^{-2L^x_B(r-s)}$, we obtain

$$\begin{aligned}
&\frac{\partial}{\partial r}\left[e^{-2L^x_B(r-s)}\left\|\eta^\lambda_{r,s}(z)\right\|^2\right] \\
&= -2L^x_B e^{-2L^x_B(r-s)}\left\|\eta^\lambda_{r,s}(z)\right\|^2 + e^{-2L^x_B(r-s)}\frac{\partial}{\partial r}\left\|\eta^\lambda_{r,s}(z)\right\|^2 \\
&\le -2L^x_B e^{-2L^x_B(r-s)}\left\|\eta^\lambda_{r,s}(z)\right\|^2 + 2L^x_B e^{-2L^x_B(r-s)}\left\|\eta^\lambda_{r,s}(z)\right\|^2 \\
&= 0.
\end{aligned} \tag{95}$$

Integrating Eq. 95 from s to r and using $\eta^\lambda_{s,s}(z) = z$ gives

$$e^{-2L^x_B(r-s)}\left\|\eta^\lambda_{r,s}(z)\right\|^2 - \|z\|^2 \le 0, \tag{96}$$

and hence

$$\left\|\eta^\lambda_{r,s}(z)\right\|^2 \le e^{2L^x_B(r-s)}\|z\|^2, \qquad \left\|\Phi^\lambda_{r,s} z\right\| \le e^{L^x_B(r-s)}\|z\|. \tag{97}$$

Finally, taking the supremum over all z with $\|z\| = 1$ yields

$$\left\|\Phi^\lambda_{r,s}\right\|_{\mathrm{op}} = \sup_{\|z\|=1}\left\|\Phi^\lambda_{r,s} z\right\| \le e^{L^x_B(r-s)}. \tag{98}$$

The adjoint representation is

$$P^\lambda_s = \mathbb{E}\left[(\Phi^\lambda_{T,s})^\top \nabla g_T(X^\lambda_T) + \int_s^T (\Phi^\lambda_{r,s})^\top c^\lambda_r dr \middle| \mathcal{F}_s\right]. \tag{99}$$

Consequently,

$$\begin{aligned}
\mathbb{E}_x\left[\|P^\lambda_s\|\right] &\le e^{L^x_B(T-s)}\mathbb{E}_x\left[\|\nabla g_T(X^\lambda_T)\|\right] + \int_s^T e^{L^x_B(r-s)}\mathbb{E}_x\left[\|c^\lambda_r\|\right] dr \\
&\le e^{L^x_B(T-s)} L_g \mathbb{E}_x\left[1 + \|X^\lambda_T\|\right] + L^x_\ell \int_s^T e^{L^x_B(r-s)}\mathbb{E}_x\left[1 + \|X^\lambda_r\|\right] dr \\
&\le C_X\left[e^{L^x_B(T-s)} L_g + L^x_\ell \int_s^T e^{L^x_B(r-s)} dr\right](1 + \|x\|) \\
&\le C_X\left[e^{L^x_B T} L_g + L^x_\ell \int_0^T e^{L^x_B r} dr\right](1 + \|x\|).
\end{aligned} \tag{100}$$

Using Eq. 91, the Hamiltonian-gradient growth conditions, and Eq. 100, we obtain

$$\begin{aligned}\left|\frac{d}{d\lambda}\mathcal{V}_i^\lambda(x)\right| &\le \mathbb{E}_x\left[\int_0^T \left\|\ell_{u_i}\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda); y_s\right)\right\| \left\|\delta\gamma(s, X_s^\lambda)\right\| ds\right] \\ &\quad + \mathbb{E}_x\left[\int_0^T \left\|B_{u_i}\left(s, X_s^\lambda, u^\lambda(s, X_s^\lambda)\right)\right\|_{\mathrm{op}} \|P_s^\lambda\| \left\|\delta\gamma(s, X_s^\lambda)\right\| ds\right] \\ &\le \rho_i(\gamma,\gamma')L_\ell^{u_i}\int_0^T \mathbb{E}_x\left[1+\|X_s^\lambda\|\right] ds + \rho_i(\gamma,\gamma')L_B^{u_i}\int_0^T \mathbb{E}_x\left[\|P_s^\lambda\|\right] ds \\ &\le TC_X\left[L_\ell^{u_i} + L_B^{u_i}\left(e^{L_B^x T}L_g + L_\ell^x\int_0^T e^{L_B^x r}dr\right)\right](1+\|x\|)\rho_i(\gamma,\gamma'). \end{aligned} \tag{101}$$

Integrating over $\lambda \in [0,1]$ gives

$$\begin{aligned}\left|\mathbf{Y}_0^{i,(\alpha_{-i},\gamma)}(x) - \mathbf{Y}_0^{i,(\alpha_{-i},\gamma')}(x)\right| &= \left|\mathcal{V}_i^0(x) - \mathcal{V}_i^1(x)\right| \le \int_0^1 \left|\frac{d}{d\lambda}\mathcal{V}_i^\lambda(x)\right| d\lambda \\ &\le TC_X\left[L_\ell^{u_i} + L_B^{u_i}\left(e^{L_B^x T}L_g + L_\ell^x\int_0^T e^{L_B^x r}dr\right)\right](1+\|x\|)\rho_i(\gamma,\gamma'). \end{aligned} \tag{102}$$

Define the value-deviation process indexed directly by the record policy class:

$$\mathsf{Z}_\gamma^0 := \left|\mathbf{Y}_0^{i,\bar{\alpha}}(\mathbf{X}_0) - \mathbf{Y}_0^{i,(\alpha_{-i},\gamma)}(\mathbf{X}_0)\right|, \qquad \gamma \in \mathcal{C}_{i,K}. \tag{103}$$

For any $\gamma, \gamma' \in \mathcal{C}_{i,K}$, the reverse triangle inequality and Eq. 102 imply

$$\begin{aligned}|\mathsf{Z}_\gamma^0 - \mathsf{Z}_{\gamma'}^0| &\le \left|\mathbf{Y}_0^{i,(\alpha_{-i},\gamma)}(\mathbf{X}_0) - \mathbf{Y}_0^{i,(\alpha_{-i},\gamma')}(\mathbf{X}_0)\right| \\ &\le TC_X\left[L_\ell^{u_i} + L_B^{u_i}\left(e^{L_B^x T}L_g + L_\ell^x\int_0^T e^{L_B^x r}dr\right)\right] \\ &\quad \times (1+\|\mathbf{X}_0\|)\rho_i(\gamma,\gamma'). \end{aligned} \tag{104}$$

By Corollary B.3, the standing sub-Gaussian initial-law assumption ensures that $\|1+\|\mathbf{X}_0\|\|_{\psi_2(\mathbb{P}_{\mathrm{init}})} < \infty$. Therefore,

$$\begin{aligned}&\|\mathsf{Z}_\gamma^0 - \mathsf{Z}_{\gamma'}^0\|_{\psi_2(\mathbb{P}_{\mathrm{init}})} \\ &\le TC_X\left[L_\ell^{u_i} + L_B^{u_i}\left(e^{L_B^x T}L_g + L_\ell^x\int_0^T e^{L_B^x r}dr\right)\right]\|1+\|\mathbf{X}_0\|\|_{\psi_2(\mathbb{P}_{\mathrm{init}})}\,\rho_i(\gamma,\gamma'). \end{aligned} \tag{105}$$

Define the canonical Dudley pseudometric

$$\begin{aligned}d_i(\gamma,\gamma') &:= TC_X\left[L_\ell^{u_i} + L_B^{u_i}\left(e^{L_B^x T}L_g + L_\ell^x\int_0^T e^{L_B^x r}dr\right)\right] \\ &\quad \times \|1+\|\mathbf{X}_0\|\|_{\psi_2(\mathbb{P}_{\mathrm{init}})}\,\rho_i(\gamma,\gamma'). \end{aligned} \tag{106}$$

Then, one finds the natural inequality $\|\mathsf{Z}_\gamma^0 - \mathsf{Z}_{\gamma'}^0\|_{\psi_2(\mathbb{P}_{\mathrm{init}})} \le d_i(\gamma,\gamma')$. Moreover, since $\bar{\alpha}_i \in \mathcal{C}_{i,K}$, the deviation process

$$\mathsf{Z}_{\bar{\alpha}_i}^0 = \left|\mathbf{Y}_0^{i,\bar{\alpha}}(\mathbf{X}_0) - \mathbf{Y}_0^{i,\bar{\alpha}}(\mathbf{X}_0)\right| = 0.$$

Aggregating the results above, we apply Dudley's entropy inequality to have

$$\mathbb{E}\left[\sup_{\gamma\in\mathcal{C}_{i,K}} \mathsf{Z}_\gamma^0\right] = \mathbb{E}\left[\sup_{\gamma\in\mathcal{C}_{i,K}} |\mathsf{Z}_\gamma^0 - \mathsf{Z}_{\bar{\alpha}_i}^0|\right] \le \mathsf{C}_D\int_0^{\mathrm{diam}_{d_i}(\mathcal{C}_{i,K})} \sqrt{\log N\left(\varepsilon, \mathcal{C}_{i,K}, d_i\right)}d\varepsilon, \tag{107}$$

where $\mathsf{C}_D > 0$ is some numerical constant. Next, we introduce the constant $\mathsf{C}_{i,T}$ defined by

$$\mathsf{C}_{i,T} := TC_X \left[L_\ell^{u_i} + L_B^{u_i} \left(e^{L_B^x T} L_g + L_\ell^x \int_0^T e^{L_B^x r} dr \right) \right] \|1 + \|\mathbf{X}_0\|\|_{\psi_2(\mathbb{P}_{\text{init}})} . \tag{108}$$

$$\leq TC_X \left[L_\ell^{u_i} + L_B^{u_i} \left(e^{L_B^x T} L_g + L_\ell^x \int_0^T e^{L_B^x r} dr \right) \right] \left[\frac{1 + \|\mathbf{m}_0\|}{\sqrt{\log 2}} + 2\mathsf{K}_0 \sqrt{\frac{\log(2 \cdot 5^{d_x})}{\log 2}} \right] \tag{109}$$

With this constant, the metric is further improved by $d_i(\gamma, \gamma') = \mathsf{C}_{i,T} \rho_i(\gamma, \gamma')$, and hence one obtains $\operatorname{diam}_{d_i}(\mathcal{C}_{i,K}) = \mathsf{C}_{i,T} \operatorname{diam}_{\rho_i}(\mathcal{C}_{i,K})$. This leads to the following equality comparing two covering numbers:

$$N(\varepsilon, \mathcal{C}_{i,K}, d_i) = N\left(\frac{\varepsilon}{\mathsf{C}_{i,T}}, \mathcal{C}_{i,K}, \rho_i \right). \tag{110}$$

Using the result, the change of variables with $r := \frac{\varepsilon}{\mathsf{C}_{i,T}}$, and $d\varepsilon = \mathsf{C}_{i,T} dr$ gives the following:

$$\int_0^{\operatorname{diam}_{d_i}(\mathcal{C}_{i,K})} \sqrt{\log N(\varepsilon, \mathcal{C}_{i,K}, d_i)} d\varepsilon = \mathsf{C}_{i,T} \int_0^{\operatorname{diam}_{\rho_i}(\mathcal{C}_{i,K})} \sqrt{\log N(r, \mathcal{C}_{i,K}, \rho_i)} dr. \tag{111}$$

Substituting the definition of $\mathsf{C}_{i,T}$ into Eq. 107 proves the main statement. □

Proposition B.5 (Neural-Policy Entropy of the Retained Class)**.** *The Dudley entropy integral of $\mathcal{C}_{i,K}$ satisfies*

$$\int_0^{\mathrm{diam}_{\rho_i}(\mathcal{C}_{i,K})} \sqrt{\log N\left(\varepsilon, \mathcal{C}_{i,K}, \rho_i\right)} d\varepsilon$$

$$\leq \inf_{0<\varepsilon_0\leq \mathsf{D}_{i,K}} \left\{ \varepsilon_0 \sqrt{\sum_{\ell=1}^{L_i} w_{i,\ell} w_{i,\ell-1}} \left[\sqrt{\log\left(1+\frac{\mathsf{D}_{i,K}}{\varepsilon_0}\right)} + \frac{\sqrt{\pi}}{2} \right] \right. \tag{112}$$

$$\left. + \mathsf{C}_{\mathrm{cov}} \left[\overline{R}_{i,K}^2 \log(2W_i^2) \left(\prod_{\ell=1}^{L_i} s_{i,\ell}^2 \varrho_{i,\ell}^2 \right) \left(\sum_{\ell=1}^{L_i} \left(\frac{b_{i,\ell}}{s_{i,\ell}} \right)^{2/3} \right)^3 \right]^{1/2} \log\left(\frac{\mathsf{D}_{i,K}}{\varepsilon_0}\right) \right\}. \tag{113}$$

The two terms quantify, respectively, the small-scale parameter cover and the large-scale spectral cover of the same restricted policy class.

Proof. Assume the neural-network regularity, parameter-hull, and uniform spectral covering conditions specified below. We retain the uniform policy-space metric introduced in Eq. 73 on the compact computational domain $\mathcal{Z}_{i,K}$ from Proposition B.4. Accordingly,

$$\rho_i(\gamma, \gamma') := \sup_{z\in\mathcal{Z}_{i,K}} \|\gamma(z) - \gamma'(z)\|_2 . \tag{114}$$

Thus, the same metric ρ_i is used in both the Pontryagin–Dudley bound and the neural-policy entropy estimate. It measures a function-space discrepancy and is not identified with the Euclidean parameter metric $\|\theta_i - \theta_i'\|_2$. We now introduce a neural policy class analogous to the score-network class used in the accompanying analysis. To be more specific, let us define the generic neural network as

$$F_{\theta_i}(z) := W_{i,L_i}\sigma_{i,L_i-1}\left(W_{i,L_i-1}\sigma_{i,L_i-2}\left(\cdots\sigma_{i,1}\left(W_{i,1}z\right)\cdots\right)\right) \tag{115}$$

with an L_i-layer feedforward neural policy. The hidden widths and all the layers satisfy

$$\|W_{i,\ell}\|_{\mathrm{op}} \leq s_{i,\ell}, \quad \|W_{i,\ell}^\top\|_{2,1} \leq b_{i,\ell}, \quad w_{i,\ell} \leq W_i, \quad 1 \leq \ell < L_i. \tag{116}$$

Next, let the activation satisfy $\sigma_{i,\ell}(0) = 0$ with $\|\sigma_{i,\ell}(v) - \sigma_{i,\ell}(v')\|_2 \leq \varrho_{i,\ell}\|v - v'\|_2$. For ReLU activations, for instance, one finds $\varrho_{i,\ell} = 1$, suggesting $\varrho_{i,L_i} := 1$. The record class $\mathcal{C}_{i,K}$ defined earlier is assumed to satisfy

$$\mathcal{C}_{i,K} \subseteq \left\{F_{\theta_i} : \|W_{i,\ell}\|_{\mathrm{op}} \leq s_{i,\ell},\ \|W_{i,\ell}^\top\|_{2,1} \leq b_{i,\ell},\ 1 \leq \ell \leq L_i\right\}, \tag{117}$$

where the number of trainable weight parameters is defined as $P_i := \sum_{\ell=1}^{L_i} w_{i,\ell}w_{i,\ell-1}$ for simplicity. Let us define the uniform input radius as follows:

$$\overline{R}_{i,K} := \sup_{z\in\mathcal{Z}_{i,K}} \|z\|_2. \tag{118}$$

We first derive how a parameter perturbation changes a neural policy under the uniform policy metric ρ_i. For $1 \leq \ell < L_i$, we then define the hidden representations of the neural network recursively by

$$h_{i,0}(z) := z, \quad h_{i,\ell}(z) := \sigma_{i,\ell}\left(W_{i,\ell}h_{i,\ell-1}(z)\right), \quad F_{\theta_i}(z) := W_{i,L_i}h_{i,L_i-1}(z), \tag{119}$$

Repeated application of spectral norm and Lipschitz constants, one is able to have

$$\|h_{i,1}(z)\|_2 = \|\sigma_{i,1}(W_{i,1}z)\|_2 \leq \varrho_{i,1}\|W_{i,1}z\|_2 \leq \varrho_{i,1}s_{i,1}\|z\|_2, \tag{120}$$

and

$$\|h_{i,2}(z)\|_2 = \|\sigma_{i,2}\left(W_{i,2}h_{i,1}(z)\right)\|_2 \leq \varrho_{i,2}s_{i,2}\|h_{i,1}(z)\|_2 \leq \varrho_{i,2}s_{i,2}\varrho_{i,1}s_{i,1}\|z\|_2. \tag{121}$$

We proceed this process inductively until reach the end of the layer to have

$$\|h_{i,\ell}(z)\|_2 \leq \left(\prod_{r=1}^{\ell} \varrho_{i,r}s_{i,r}\right)\|z\|_2. \tag{122}$$

Next, let us consider a parameter perturbation $\Delta\theta_i := (\Delta W_{i,1}, \ldots, \Delta W_{i,L_i})$ with its squared norm $\|\Delta\theta_i\|_2^2 := \sum_{\ell=1}^{L_i} \|\Delta W_{i,\ell}\|_F^2$. The directional derivative of the network can be decomposed layer by layer:

$$D_{\theta_i} F_{\theta_i}(z)[\Delta\theta_i] = \sum_{\ell=1}^{L_i} D_{W_{i,\ell}} F_{\theta_i}(z)[\Delta W_{i,\ell}]. \tag{123}$$

Within this interpretation, we compute the contribution ofthe ℓ-th layer in the followings:

$$\begin{aligned}
&\left\| D_{W_{i,\ell}} F_{\theta_i}(z)[\Delta W_{i,\ell}] \right\|_2 \\
&\le \left(\prod_{r=\ell+1}^{L_i} s_{i,r}\varrho_{i,r} \right) \|\Delta W_{i,\ell}\|_{\mathrm{op}} \|h_{i,\ell-1}(z)\|_2 \\
&\le \left(\prod_{r=\ell+1}^{L_i} s_{i,r}\varrho_{i,r} \right) \|\Delta W_{i,\ell}\|_F \left(\prod_{r=1}^{\ell-1} s_{i,r}\varrho_{i,r} \right) \|z\|_2 \\
&= \left(\prod_{r=1}^{L_i} s_{i,r}\varrho_{i,r} \right) \frac{\|\Delta W_{i,\ell}\|_F}{s_{i,\ell}\varrho_{i,\ell}} \|z\|_2.
\end{aligned} \tag{124}$$

Summing over the layers gives

$$\begin{aligned}
&\|D_{\theta_i} F_{\theta_i}(z)[\Delta\theta_i]\|_2 \\
&\le \left(\prod_{r=1}^{L_i} s_{i,r}\varrho_{i,r} \right) \|z\|_2 \sum_{\ell=1}^{L_i} \frac{\|\Delta W_{i,\ell}\|_F}{s_{i,\ell}\varrho_{i,\ell}}.
\end{aligned} \tag{125}$$

By the Cauchy–Schwarz inequality,

$$\begin{aligned}
\sum_{\ell=1}^{L_i} \frac{\|\Delta W_{i,\ell}\|_F}{s_{i,\ell}\varrho_{i,\ell}} &\le \left(\sum_{\ell=1}^{L_i} \frac{1}{s_{i,\ell}^2 \varrho_{i,\ell}^2} \right)^{1/2} \left(\sum_{\ell=1}^{L_i} \|\Delta W_{i,\ell}\|_F^2 \right)^{1/2} \\
&= \left(\sum_{\ell=1}^{L_i} \frac{1}{s_{i,\ell}^2 \varrho_{i,\ell}^2} \right)^{1/2} \|\Delta\theta_i\|_2.
\end{aligned} \tag{126}$$

Combining Eq. 125 and Eq. 126,

$$\|D_{\theta_i} F_{\theta_i}(z)[\Delta\theta_i]\|_2 \le \left(\prod_{r=1}^{L_i} s_{i,r}\varrho_{i,r} \right) \left(\sum_{\ell=1}^{L_i} \frac{1}{s_{i,\ell}^2 \varrho_{i,\ell}^2} \right)^{1/2} \|z\|_2 \|\Delta\theta_i\|_2. \tag{127}$$

Taking the supremum over $z \in \mathcal{Z}_{i,K}$ and using Eq. 118, we obtain

$$\sup_{z \in \mathcal{Z}_{i,K}} \|D_{\theta_i} F_{\theta_i}(z)[\Delta\theta_i]\|_2 \le \overline{R}_{i,K} \left(\prod_{r=1}^{L_i} s_{i,r}\varrho_{i,r} \right) \left(\sum_{\ell=1}^{L_i} \frac{1}{s_{i,\ell}^2 \varrho_{i,\ell}^2} \right)^{1/2} \|\Delta\theta_i\|_2. \tag{128}$$

Therefore,

$$\sup_{z \in \mathcal{Z}_{i,K}} \|D_{\theta_i} F_{\theta_i}(z)[\Delta\theta_i]\|_2 \le J_{i,K}^{\mathrm{NN}} \|\Delta\theta_i\|_2, \;\; J_{i,K}^{\mathrm{NN}} := \overline{R}_{i,K} \left(\prod_{r=1}^{L_i} s_{i,r}\varrho_{i,r} \right) \left(\sum_{\ell=1}^{L_i} \frac{1}{s_{i,\ell}^2 \varrho_{i,\ell}^2} \right)^{1/2}. \tag{129}$$

Note that this value is a derived neural-network Jacobian estimate under ρ_i, not an identification of ρ_i with the Euclidean parameter norm. By Proposition B.2, every retained policy has the gradient expansion in Eq. 37. Since

$$\bar{\theta}_i^K = \bar{\theta}_i^0 - \sum_{k=0}^{K-1} \lambda_i^k \mathbf{G}_{i,k}, \tag{130}$$

we define the terminal-centered parameter radius

$$R^{\theta}_{i,K} := \sup_{\gamma \in \mathcal{C}_{i,K}} \left\| \theta^{\gamma}_i - \bar{\theta}^K_i \right\|_2 . \tag{131}$$

The gradient representation provides the explicit finite-K bound

$$R^{\theta}_{i,K} \le \sup_{\gamma \in \mathcal{C}_{i,K}} \sum_{k=0}^{K-1} \left| a_{i,k}(\gamma) - \lambda^k_i \right| \|\mathbf{G}_{i,k}\|_2 \le \sum_{k=0}^{K-1} \|\mathbf{G}_{i,k}\|_2. \tag{132}$$

Assume that every parameter segment joining $\bar{\theta}^K_i$ to a retained parameter remains in the constrained neural parameter hull in Eq. 117. The fundamental theorem of calculus and Eq. 129 then give

$$\rho_i \left(F_{\theta^{\gamma}_i}, F_{\bar{\theta}^K_i} \right) \le J^{\mathrm{NN}}_{i,K} \left\| \theta^{\gamma}_i - \bar{\theta}^K_i \right\|_2 \le J^{\mathrm{NN}}_{i,K} R^{\theta}_{i,K}. \tag{133}$$

For arbitrary $\gamma, \gamma' \in \mathcal{C}_{i,K}$, the triangle inequality therefore yields

$$\rho_i(\gamma, \gamma') \le 2 J^{\mathrm{NN}}_{i,K} R^{\theta}_{i,K}. \tag{134}$$

We define

$$\mathsf{D}_{i,K} := 2 J^{\mathrm{NN}}_{i,K} R^{\theta}_{i,K}, \tag{135}$$

so that $\mathrm{diam}_{\rho_i}(\mathcal{C}_{i,K}) \le \mathsf{D}_{i,K}$. The record parameters lie in $B_2(\bar{\theta}^K_i, R^{\theta}_{i,K})$. A Euclidean parameter cover of radius $\delta := \varepsilon / J^{\mathrm{NN}}_{i,K}$ induces an ε-cover under ρ_i by Eq. 129. Therefore,

$$\begin{aligned} N(\varepsilon, \mathcal{C}_{i,K}, \rho_i) &\le \left(1 + \frac{2R^{\theta}_{i,K}}{\delta} \right)^{P_i} \\ &= \left(1 + \frac{2 J^{\mathrm{NN}}_{i,K} R^{\theta}_{i,K}}{\varepsilon} \right)^{P_i} \\ &= \left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon} \right)^{P_i}. \end{aligned} \tag{136}$$

Consequently,

$$\log N(\varepsilon, \mathcal{C}_{i,K}, \rho_i) \le P_i \log \left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon} \right). \tag{137}$$

For the large scales, we impose the following spectral $(2,1)$-norm covering condition for the same neural policy class under the uniform metric ρ_i:

$$\log N(\varepsilon, \mathcal{C}_{i,K}, \rho_i) \le \frac{\left(\mathsf{C}^{\mathrm{NN}}_{i,K} \right)^2}{\varepsilon^2}, \tag{138}$$

where

$$\mathsf{C}^{\mathrm{NN}}_{i,K} := \mathsf{C}_{\mathrm{cov}} \left[\overline{R}^2_{i,K} \log(2W^2_i) \left(\prod_{\ell=1}^{L_i} s^2_{i,\ell} \varrho^2_{i,\ell} \right) \left(\sum_{\ell=1}^{L_i} \left(\frac{b_{i,\ell}}{s_{i,\ell}} \right)^{2/3} \right)^3 \right]^{1/2}. \tag{139}$$

Here $\mathsf{C}_{\mathrm{cov}} > 0$ is the universal constant in the neural-network covering inequality. The form of Eq. 139 is the same spectral $(2,1)$-norm complexity used for the neural class. Combining the small- and large-scale bounds,

$$\log N(\varepsilon, \mathcal{C}_{i,K}, \rho_i) \le \min \left\{ P_i \log \left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon} \right), \frac{\left(\mathsf{C}^{\mathrm{NN}}_{i,K} \right)^2}{\varepsilon^2} \right\}. \tag{140}$$

We now estimate the Dudley entropy integral. Let us fix an arbitrary splitting scale $0 < \varepsilon_0 \le \mathsf{D}_{i,K}$. Using the diameter of gradient, we split the entropy integral:

$$\int_0^{\mathrm{diam}_{\rho_i}(\mathcal{C}_{i,K})} \sqrt{\log N(\varepsilon, \mathcal{C}_{i,K}, \rho_i)} d\varepsilon \le \int_0^{\varepsilon_0} \sqrt{\log N(\varepsilon, \mathcal{C}_{i,K}, \rho_i)} d\varepsilon + \int_{\varepsilon_0}^{\mathsf{D}_{i,K}} \sqrt{\log N(\varepsilon, \mathcal{C}_{i,K}, \rho_i)} d\varepsilon.$$

For the small-scale part, we use Eq. 137:

$$\int_0^{\varepsilon_0} \sqrt{\log N\left(\varepsilon, \mathcal{C}_{i,K}, \rho_i\right)} d\varepsilon \leq \sqrt{P_i} \int_0^{\varepsilon_0} \sqrt{\log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon}\right)} d\varepsilon. \tag{141}$$

$$= \varepsilon_0 \sqrt{P_i} \int_0^1 \sqrt{\log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_0 u}\right)} du, \tag{142}$$

where we set $u := \frac{\varepsilon}{\varepsilon_0}$, $\varepsilon = \varepsilon_0 u$, and $d\varepsilon = \varepsilon_0 du$ for applying the change of variables. Note that for $u \in (0, 1]$, one obtains the following inequality:

$$1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_0 u} \leq \frac{1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_0}}{u}. \tag{143}$$

Therefore,

$$\log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_0 u}\right) \leq \log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_0}\right) + \log\left(\frac{1}{u}\right). \tag{144}$$

Using $\sqrt{a+b} \leq \sqrt{a} + \sqrt{b}$,

$$\sqrt{\log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_0 u}\right)}$$

$$\leq \sqrt{\log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_0}\right)} + \sqrt{\log\left(\frac{1}{u}\right)}. \tag{145}$$

It follows that

$$\int_0^{\varepsilon_0} \sqrt{\log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon}\right)} d\varepsilon$$

$$\leq \varepsilon_0 \sqrt{\log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_0}\right)} + \varepsilon_0 \int_0^1 \sqrt{\log\left(\frac{1}{u}\right)} du$$

$$= \varepsilon_0 \left[\sqrt{\log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_0}\right)} + \frac{\sqrt{\pi}}{2}\right]. \tag{146}$$

Hence,

$$\int_0^{\varepsilon_0} \sqrt{\log N\left(\varepsilon, \mathcal{C}_{i,K}, \rho_i\right)} d\varepsilon \leq \varepsilon_0 \sqrt{P_i} \left[\sqrt{\log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_0}\right)} + \frac{\sqrt{\pi}}{2}\right]. \tag{147}$$

For the large-scale part, use Eq. 138:

$$\int_{\varepsilon_0}^{\mathsf{D}_{i,K}} \sqrt{\log N\left(\varepsilon, \mathcal{C}_{i,K}, \rho_i\right)} d\varepsilon \leq \int_{\varepsilon_0}^{\mathsf{D}_{i,K}} \frac{\mathsf{C}_{i,K}^{\mathrm{NN}}}{\varepsilon} d\varepsilon = \mathsf{C}_{i,K}^{\mathrm{NN}} \left[\log \varepsilon\right]_{\varepsilon=\varepsilon_0}^{\varepsilon=\mathsf{D}_{i,K}}$$

$$= \mathsf{C}_{i,K}^{\mathrm{NN}} \left(\log \mathsf{D}_{i,K} - \log \varepsilon_0\right)$$

$$= \mathsf{C}_{i,K}^{\mathrm{NN}} \log\left(\frac{\mathsf{D}_{i,K}}{\varepsilon_0}\right). \tag{148}$$

Combining results in Eq. 147 and Eq. 148, we obtain, for every $0 < \varepsilon_0 \leq \mathsf{D}_{i,K}$,

$$\int_0^{\mathrm{diam}_{\rho_i}(\mathcal{C}_{i,K})} \sqrt{\log N\left(\varepsilon, \mathcal{C}_{i,K}, \rho_i\right)} d\varepsilon$$

$$\leq \inf_{0 < \varepsilon_0 \leq \mathsf{D}_{i,K}} \left\{\varepsilon_0 \sqrt{P_i} \left[\sqrt{\log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_0}\right)} + \frac{\sqrt{\pi}}{2}\right] + \mathsf{C}_{i,K}^{\mathrm{NN}} \log\left(\frac{\mathsf{D}_{i,K}}{\varepsilon_0}\right)\right\}. \tag{149}$$

Taking the infimum over $\varepsilon_0 \in (0, \mathsf{D}_{i,K}]$ proves the main statement by showing Eq. 113. □

Proposition B.6 (Optimized Neural-Policy Entropy). *If $\mathsf{D}_{i,K} = 0$, the entropy integral vanishes. If $\mathsf{D}_{i,K} > 0$ and $\mathsf{C}_{i,K}^{\mathrm{NN}} > 0$, then at the minimizing splitting scale $\varepsilon_{i,K}^{\star}$ defined in the proof,*

$$\int_0^{\mathrm{diam}_{\rho_i}(\mathcal{C}_{i,K})} \sqrt{\log N\left(\varepsilon, \mathcal{C}_{i,K}, \rho_i\right)} d\varepsilon \leq \varepsilon_{i,K}^{\star} \sqrt{P_i} \left[\sqrt{\log\left(1 + \frac{\mathsf{D}_{i,K}}{\varepsilon_{i,K}^{\star}}\right)} + \frac{\sqrt{\pi}}{2} \right] + \mathsf{C}_{i,K}^{\mathrm{NN}} \log\left(\frac{\mathsf{D}_{i,K}}{\varepsilon_{i,K}^{\star}}\right). \tag{150}$$

If $\mathsf{C}_{i,K}^{\mathrm{NN}} \geq \mathsf{D}_{i,K}\sqrt{P_i}\left(\sqrt{\log 2} + \frac{\sqrt{\pi}}{2} - \frac{1}{4\sqrt{\log 2}}\right)$, then $\varepsilon_{i,K}^{\star} = \mathsf{D}_{i,K}$. Thus, the optimized estimate selects the boundary scale whenever the large-scale entropy coefficient dominates the stated threshold.

Proof. If $\mathsf{D}_{i,K} = 0$, the entropy integral vanishes and the claim is immediate. Hence, assume $\mathsf{D}_{i,K} > 0$. For brevity, define $D := \mathsf{D}_{i,K}$, $C := \mathsf{C}_{i,K}^{\mathrm{NN}}$, $P := P_i$, and $a := \frac{\sqrt{\pi}}{2}$. By the preceding covering-number estimates, for every $0 < \varepsilon_0 \leq D$,

$$\int_0^{\mathrm{diam}_{\rho_i}(\mathcal{C}_{i,K})} \sqrt{\log N\left(\varepsilon, \mathcal{C}_{i,K}, \rho_i\right)} d\varepsilon \leq \varepsilon_0 \sqrt{P} \left[\sqrt{\log\left(1 + \frac{D}{\varepsilon_0}\right)} + a \right] + C \log\left(\frac{D}{\varepsilon_0}\right). \tag{151}$$

It therefore remains to minimize the right-hand side over $\varepsilon_0 \in (0, D]$. For further simplicity, let us denote

$$F(\varepsilon_0) := \varepsilon_0 \sqrt{P} \left[\sqrt{\log\left(1 + \frac{D}{\varepsilon_0}\right)} + a \right] + C \log\left(\frac{D}{\varepsilon_0}\right). \tag{152}$$

While $C > 0$, one obtains the $\lim_{\varepsilon_0 \downarrow 0} F(\varepsilon_0) = +\infty$. Since F is continuous on $(0, D]$, its minimum is attained either at an interior critical point or at the boundary $\varepsilon_0 = D$. Let $\varepsilon_{i,K}^{\star} \in (0, \mathsf{D}_{i,K}]$ denote such a minimizer. Evaluating Eq. 151 at $\varepsilon_{i,K}^{\star}$ proves Eq. 150. Upon these preliminary setting, we first compute its derivative explicitly. Let us define another function h and its derivative with respect to variable ε_0:

$$h(\varepsilon_0) := \sqrt{\log\left(1 + \frac{D}{\varepsilon_0}\right)}, \quad \frac{d}{d\varepsilon_0} \log\left(1 + \frac{D}{\varepsilon_0}\right) = \frac{-D/\varepsilon_0^2}{1 + D/\varepsilon_0} = -\frac{D}{\varepsilon_0(D + \varepsilon_0)}. \tag{153}$$

One can easily find out the critical point of h as

$$h'(\varepsilon_0) = \frac{1}{2h(\varepsilon_0)} \frac{d}{d\varepsilon_0} \log\left(1 + \frac{D}{\varepsilon_0}\right) = -\frac{D}{2\varepsilon_0(D + \varepsilon_0)h(\varepsilon_0)}. \tag{154}$$

Therefore,

$$\begin{aligned} F'(\varepsilon_0) &= \sqrt{P}\left[h(\varepsilon_0) + a + \varepsilon_0 h'(\varepsilon_0)\right] - \frac{C}{\varepsilon_0} \\ &= \sqrt{P}\left[h(\varepsilon_0) + a - \frac{D}{2(D + \varepsilon_0)h(\varepsilon_0)}\right] - \frac{C}{\varepsilon_0}. \end{aligned} \tag{155}$$

Consequently, every interior critical point satisfies

$$C = \varepsilon_0 \sqrt{P} \left[\sqrt{\log\left(1 + \frac{D}{\varepsilon_0}\right)} + a - \frac{D}{2(D + \varepsilon_0)\sqrt{\log\left(1 + \frac{D}{\varepsilon_0}\right)}} \right]. \tag{156}$$

To determine the existence and uniqueness of such a point, set $r := \frac{D}{\varepsilon_0}$. The constraint $0 < \varepsilon_0 \leq D$ is equivalent to $r \in [1, \infty)$, and $\varepsilon_0 = \frac{D}{r}$. Substituting this identity into Eq. 152 gives

$$f(r) := F\left(\frac{D}{r}\right) = \frac{D\sqrt{P}}{r} \left[\sqrt{\log(1 + r)} + a\right] + C \log r. \tag{157}$$

Let us define another function $g(r) := \sqrt{\log(1+r)}$, with constant $\lambda := \frac{C}{D\sqrt{P}}$. Since its derivative is simply derived as $g'(r) = \frac{1}{2(1+r)g(r)}$, one can obtain

$$\begin{aligned} f'(r) &= D\sqrt{P}\left[-\frac{g(r)+a}{r^2} + \frac{g'(r)}{r}\right] + \frac{C}{r} \\ &= \frac{D\sqrt{P}}{r^2}\left[-g(r) - a + \frac{r}{2(1+r)g(r)} + \lambda r\right]. \end{aligned} \tag{158}$$

For the following function ϕ,

$$\phi(r) := \frac{1}{r}\left[g(r) + a - \frac{r}{2(1+r)g(r)}\right] = \frac{g(r)+a}{r} - \frac{1}{2(1+r)g(r)}, \tag{159}$$

The calculated value in Eq. 158 becomes

$$f'(r) = \frac{D\sqrt{P}}{r}\left[\lambda - \phi(r)\right]. \tag{160}$$

We next show that ϕ is strictly decreasing on $[1, \infty)$ by differentiating Eq. 159 to have

$$\phi'(r) = -\frac{g(r)+a}{r^2} + \frac{1}{2r(1+r)g(r)} + \frac{1}{2(1+r)^2 g(r)} + \frac{1}{4(1+r)^2 g(r)^3}. \tag{161}$$

Multiplying by r^2, we have

$$r^2\phi'(r) = -g(r) - a + \frac{r}{2(1+r)g(r)} + \frac{r^2}{2(1+r)^2 g(r)} + \frac{r^2}{4(1+r)^2 g(r)^3}. \tag{162}$$

Since $\frac{r}{1+r} \leq 1$ and $\frac{r^2}{(1+r)^2} \leq 1$, it follows that the following inequality holds:

$$r^2\phi'(r) \leq -g(r) - a + \frac{1}{g(r)} + \frac{1}{4g(r)^3}, \quad g(r) \geq g(1) = \sqrt{\log 2}. \tag{163}$$

For further analysis, let us define q and its derivative q' as follows:

$$q(v) := -v - a + \frac{1}{v} + \frac{1}{4v^3}, \quad q'(v) = -1 - \frac{1}{v^2} - \frac{3}{4v^4} < 0 \tag{164}$$

Following the definition, we have

$$\begin{aligned} r^2\phi'(r) &\leq q\left(\sqrt{\log 2}\right) \\ &= -\sqrt{\log 2} - \frac{\sqrt{\pi}}{2} + \frac{1}{\sqrt{\log 2}} + \frac{1}{4(\log 2)^{3/2}} \\ &< 0. \end{aligned} \tag{165}$$

Therefore, ϕ is strictly decreasing. At the value $r = 1$, we observe the following two facts:

$$\phi(1) = \sqrt{\log 2} + \frac{\sqrt{\pi}}{2} - \frac{1}{4\sqrt{\log 2}}, \lim_{r\to\infty} \phi(r) = 0 \tag{166}$$

Suppose that the neural network is large enough to satisfy $C \geq D\sqrt{P}\left(\sqrt{\log 2} + \frac{\sqrt{\pi}}{2} - \frac{1}{4\sqrt{\log 2}}\right)$. Equivalently, this condition can be described as $\lambda \geq \phi(1)$. Since $\phi(r) \leq \phi(1)$ for all $r \geq 1$, Eq. 160 implies $f'(r) \geq 0$ and $r \geq 1$. Hence, f is minimized at $r = 1$, which is equivalent to

$$\varepsilon^\star_{i,K} = D = \mathsf{D}_{i,K}. \tag{167}$$

□

Proposition B.7 (Expected Restricted Response-Gap Certificate)**.** *The expected restricted response gap satisfies*

$$\max_{1\le i\le N}\mathbb{E}_{\mathbf{X}_0\sim\mathbb{P}_{\text{init}}}\left[\mathit{Gap}_{i,0,\mathbf{X}_0}(\bar\alpha^K)\right]\le\varepsilon_K:=\mathsf{C}_D\left(\sqrt{\log 2}+\frac{\sqrt{\pi}}{2}\right)\max_{1\le i\le N}\left\{\mathsf{C}_{i,T}\mathsf{D}_{i,K}\sqrt{P_i}\right\} \tag{168}$$

Under the localization condition specified in the proof, $\varepsilon_K\to 0$. This concerns only the restricted response classes $\mathcal{C}_{i,K}$, not the global Nash gap over the full admissible policy classes.

Proof. Assume the conditions of Propositions B.4 and B.5. Define the restricted unilateral response gap over $\mathcal{C}_{i,K}$ by

$$\mathsf{Gap}_{i,0,x}(\bar\alpha^K):=\mathcal{J}_{0,x}(\bar\alpha^K)-\inf_{\gamma\in\mathcal{C}_{i,K}}\mathcal{J}_{0,x}\left(\bar\alpha^K_{-i},\gamma\right). \tag{169}$$

Fix an arbitrary agent $i\in\{1,\dots,N\}$. By the FBSDE value representation,

$$\mathcal{J}_{0,x}(\bar\alpha^K)=\mathbf{Y}_0^{i,\bar\alpha^K}(x),\quad \mathcal{J}_{0,x}\big(\bar\alpha^K_{-i},\gamma\big)=\mathbf{Y}_0^{i,(\bar\alpha^K_{-i},\gamma)}(x),\quad \gamma\in\mathcal{C}_{i,K}. \tag{170}$$

It follows from the definition of the pointwise restricted response gap that

$$\begin{aligned}\mathsf{Gap}_{i,0,x}(\bar\alpha^K)&=\mathcal{J}_{0,x}(\bar\alpha^K)-\inf_{\gamma\in\mathcal{C}_{i,K}}\mathcal{J}_{0,x}\big(\bar\alpha^K_{-i},\gamma\big)\\&=\sup_{\gamma\in\mathcal{C}_{i,K}}\left\{\mathcal{J}_{0,x}(\bar\alpha^K)-\mathcal{J}_{0,x}\big(\bar\alpha^K_{-i},\gamma\big)\right\}\\&=\sup_{\gamma\in\mathcal{C}_{i,K}}\left\{\mathbf{Y}_0^{i,\bar\alpha^K}(x)-\mathbf{Y}_0^{i,(\bar\alpha^K_{-i},\gamma)}(x)\right\}\\&\le\sup_{\gamma\in\mathcal{C}_{i,K}}\left|\mathbf{Y}_0^{i,\bar\alpha^K}(x)-\mathbf{Y}_0^{i,(\bar\alpha^K_{-i},\gamma)}(x)\right|.\end{aligned} \tag{171}$$

Evaluating equation 171 at the random initial state $\mathbf{X}_0\sim\mathbb{P}_{\text{init}}$ and taking expectations gives

$$\mathbb{E}_{\mathbf{X}_0\sim\mathbb{P}_{\text{init}}}\left[\mathsf{Gap}_{i,0,\mathbf{X}_0}(\bar\alpha^K)\right]\le\mathbb{E}_{\mathbf{X}_0\sim\mathbb{P}_{\text{init}}}\left[\sup_{\gamma\in\mathcal{C}_{i,K}}\left|\mathbf{Y}_0^{i,\bar\alpha^K}(\mathbf{X}_0)-\mathbf{Y}_0^{i,(\bar\alpha^K_{-i},\gamma)}(\mathbf{X}_0)\right|\right]. \tag{172}$$

Applying the Sub-Gaussian Pontryagin–Dudley Certificate in previous result to the right-hand side yields

$$\mathbb{E}_{\mathbf{X}_0\sim\mathbb{P}_{\text{init}}}\left[\mathsf{Gap}_{i,0,\mathbf{X}_0}(\bar\alpha^K)\right]\le\mathsf{C}_D\mathsf{C}_{i,T}\int_0^{\operatorname{diam}_{\rho_i}(\mathcal{C}_{i,K})}\sqrt{\log N\left(\varepsilon,\mathcal{C}_{i,K},\rho_i\right)}\,d\varepsilon. \tag{173}$$

Since $\operatorname{diam}_{\rho_i}(\mathcal{C}_{i,K})\le\mathsf{D}_{i,K}$, the small-scale covering bound Eq. 137 and Eq. 147 with $\varepsilon_0=\mathsf{D}_{i,K}$ give

$$\begin{aligned}&\int_0^{\operatorname{diam}_{\rho_i}(\mathcal{C}_{i,K})}\sqrt{\log N\left(\varepsilon,\mathcal{C}_{i,K},\rho_i\right)}\,d\varepsilon\\&\le\mathsf{D}_{i,K}\sqrt{P_i}\left(\sqrt{\log 2}+\frac{\sqrt{\pi}}{2}\right).\end{aligned} \tag{174}$$

Combining equation 173 and equation 174 yields

$$\mathbb{E}_{\mathbf{X}_0\sim\mathbb{P}_{\text{init}}}\left[\mathsf{Gap}_{i,0,\mathbf{X}_0}(\bar\alpha^K)\right]\le\mathsf{C}_D\mathsf{C}_{i,T}\mathsf{D}_{i,K}\sqrt{P_i}\left(\sqrt{\log 2}+\frac{\sqrt{\pi}}{2}\right). \tag{175}$$

Taking the maximum over $i=1,\dots,N$ proves Eq. 168.

Finally, fix any $\gamma\in\mathcal{C}_{i,K}$. By the definition of the pointwise gap, for every x,

$$\mathcal{J}_{0,x}(\bar\alpha^K)-\mathcal{J}_{0,x}\big(\bar\alpha^K_{-i},\gamma\big)\le\mathsf{Gap}_{i,0,x}(\bar\alpha^K). \tag{176}$$

Taking expectations with respect to $\mathbf{X}_0 \sim \mathbb{P}_{\text{init}}$ and using linearity of expectation gives

$$\begin{aligned}
&\mathbb{E}_{\mathbf{X}_0 \sim \mathbb{P}_{\text{init}}}\left[\mathcal{J}_{0,\mathbf{X}_0}(\bar{\alpha}^K)\right] - \mathbb{E}_{\mathbf{X}_0 \sim \mathbb{P}_{\text{init}}}\left[\mathcal{J}_{0,\mathbf{X}_0}\left(\bar{\alpha}^K_{-i}, \gamma\right)\right] \\
&= \mathbb{E}_{\mathbf{X}_0 \sim \mathbb{P}_{\text{init}}}\left[\mathcal{J}_{0,\mathbf{X}_0}(\bar{\alpha}^K) - \mathcal{J}_{0,\mathbf{X}_0}\left(\bar{\alpha}^K_{-i}, \gamma\right)\right] \\
&\leq \mathbb{E}_{\mathbf{X}_0 \sim \mathbb{P}_{\text{init}}}\left[\mathsf{Gap}_{i,0,\mathbf{X}_0}(\bar{\alpha}^K)\right] \leq \varepsilon_K. \qquad (177)
\end{aligned}$$

Rearranging gives, for every $\gamma \in \mathcal{C}_{i,K}$ and $i = 1, \ldots, N$,

$$\mathbb{E}_{\mathbf{X}_0 \sim \mathbb{P}_{\text{init}}}\left[\mathcal{J}_{0,\mathbf{X}_0}(\bar{\alpha}^K)\right] \leq \mathbb{E}_{\mathbf{X}_0 \sim \mathbb{P}_{\text{init}}}\left[\mathcal{J}_{0,\mathbf{X}_0}\left(\bar{\alpha}^K_{-i}, \gamma\right)\right] + \varepsilon_K. \qquad (178)$$

This is the policywise form of Eq. 168. Finally, assume that the retention scheme localizes the generated response classes so that

$$\lim_{K \to \infty} \max_{1 \leq i \leq N} \left\{ \mathsf{C}_{i,T} \mathsf{D}_{i,K} \sqrt{P_i} \right\} = 0.$$

The definition of ε_K in Eq. 168 then gives $\varepsilon_K \to 0$. □