MOTIONPLAN: TRAINING-FREE LLM-GUIDED RESIDUE MOTION PLANNING FOR PROTEIN DYNAMICS GENERATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Protein dynamics generators can produce plausible states and temporally ordered rollouts, yet their objectives largely fit marginal distributions rather than the inter-residue coordination organizing transitions; coordination errors therefore accumulate over long horizons. Retraining for explicit residue-level control is costly and requires paired motion supervision rarely available in MD data. We introduce *MotionPlan*, a training-free framework in which a large language model (LLM) infers per-residue coordination roles and confidence from sequence and structural context. The plan becomes a localized, confidence-weighted intervention on a frozen generator's pair representations throughout reverse-SDE sampling, without parameter updates or coordinate editing. Because Pairformer-based generators share pair representations, this provides a reusable inference-time control interface. On the complete 82-protein ATLAS test split, MotionPlan reduces trajectory-level Cα MAE from 4.60 Å to 3.309 Å on the stride-256, $L = 9$ benchmark and improves frame/consecutive-frame Pearson correlation from 0.62/0.53 to 0.67/0.59. On the 100 ns benchmark, it lowers JSD from 0.52 to 0.44 and raises state recall from 0.36 to 0.59 while retaining comparable molecular quality. These results show that LLM reasoning can translate structural evidence into actionable motion priors for frozen protein dynamics generators.

1 INTRODUCTION

Proteins function through conformational ensembles and transitions rather than a single static structure (Henzler-Wildman & Kern, 2007). Molecular dynamics (MD) resolves these processes atomistically but is costly at long timescales. Diffusion- and flow-based generators amortize this cost by learning conformation distributions (Jing et al., 2023; Zheng et al., 2024; Jing et al., 2024a; Wang et al., 2024; Lewis et al., 2025; Daigavane et al., 2025; Lu et al., 2025a) or ordered rollouts (Klein et al., 2023; Jing et al., 2024b; Cheng et al., 2025; dos Santos Costa et al., 2024; Shen et al., 2025; dos Santos Costa et al., 2025; Shoghi et al., 2026; Liu et al., 2026). However, their objectives emphasize marginal or short-window statistics rather than the coordination organizing transitions. Rollouts may thus look plausible while coupling flexible loops too tightly to rigid cores or decoupling residues that should move coherently, with errors compounding over time. Figure 1 illustrates excessive and suppressed conformational change under a frozen baseline.

Retraining for relational control is unattractive: it requires trajectories annotated with desired rigid cores, independently varying regions, and their boundaries, which MD datasets do not provide, and re-optimizing large generators for each signal is expensive. From a world-model perspective, current generators learn molecular *states* and transitions but expose neither an inference-time *action* for residue coordination nor an *evaluator* that maps structural evidence to that action. Dynamic fidelity is thus inherited from training data and sampling noise.

A shared architectural pattern makes this training-free relational action possible. State-of-the-art systems such as ConfRover and STAR-MD follow a common *Pairformer features + diffusion generation* paradigm. ConfRover couples AlphaFold-derived Pairformer representations with an SE(3)-equivariant diffusion decoder (Shen et al., 2025); STAR-MD retains pretrained pairwise features and explicit pair updates while introducing scalable joint spatiotemporal attention for long-horizon

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: Failure cases of a frozen protein dynamics generator. Relative to reference MD, the baseline exaggerates motion in residues 229–241 of 6XB3-H and 191–203 of 7LAO-A, but suppresses motion in residues 64–76 of 7S86-A and 206–218 of 7LA6-A; local RMSDs are annotated in the figure. These cases expose the gap between plausible conformational marginals and the rigid–flexible coordination governing frame-to-frame dynamics.

generation (Shoghi et al., 2026). Across these temporal architectures, pair representations encode residue–residue relations before coordinates are generated. They therefore provide a native intervention surface that expresses coupling and decoupling biases while leaving geometric realization to the coordinate decoder.

We introduce **MotionPlan**, a training-free framework that follows the pipeline summarized in Figure 2. A causal LLM evaluator reasons jointly over the observed sequence, initial structure, secondary structure, contacts, exposure, and broader structural context to assign each residue a coordination role and confidence. The resulting static residue plan is lifted into localized pairwise coupling strengths and converted into a gated pair offset. This offset is reapplied to the frozen generator's base pair features at every denoising step, while the pretrained diffusion decoder remains responsible for backbone geometry and the resulting trajectory; optional replanning can refresh the residue plan when the conditioning context changes.

A trajectory-level coupling plan provides a coherent relational prior throughout the rollout. On the stride-256, $L = 9$ short-horizon benchmark, MotionPlan reduces Cα trajectory-change MAE by 28.1% relative to ConfRover and improves frame/consecutive-frame Pearson correlation by 8.1%/11.3%, respectively. At 100 ns, it reduces JSD by 15.4% and improves state recall by 63.9% relative to ConfRover. Relaxed-structure MolProbity remains comparable, differing by only 1.2% from ConfRover. Together, these results show that controlling inter-residue coordination improves short- and long-horizon dynamics while retaining the baseline level of molecular quality.

Our contributions are threefold:

- We identify a relational-control gap in protein dynamics generation: plausible conformational marginals alone leave the inference-time coordination or decoupling of residue pairs unspecified.
- We introduce MotionPlan, a training-free LLM-guided mechanism that jointly predicts per-residue coordination roles and five-level ordinal confidence scores, then converts them into a deterministic signed pair-latent offset. The frozen generator retains authority over coordinate-space motion directions and all model parameters remain fixed.
- We demonstrate improved short- and long-horizon dynamic fidelity on the complete 82-protein ATLAS test split while maintaining molecular quality at the ConfRover baseline level, establishing pair representations as a reusable control surface for inter-residue coordination.

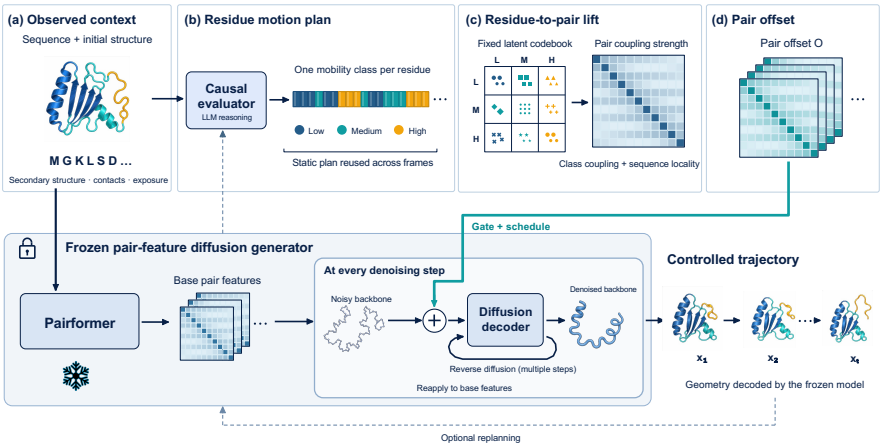


Figure 2: MotionPlan overview. (a) A causal evaluator observes sequence, initial structure, secondary structure, contacts, and exposure. (b) The LLM produces a static residue motion plan reused across frames; confidence weighting is omitted for clarity. (c) A fixed latent mapping and sequence locality lift the plan to pairwise coupling strengths. (d) A gate and denoising schedule form pair offset O, which is reapplied to the frozen generator's base pair features at every reverse step without accumulation. The frozen Pairformer–diffusion generator decodes the trajectory, with optional replanning when the conditioning context changes.

2 Related Works

2.1 Protein conformation and trajectory generation

Protein ensemble generators complement static predictors such as AlphaFold2/3 and ESMFold (Jumper et al., 2021; Abramson et al., 2024; Lin et al., 2023). AlphaFlow adapts structure predictors with flow matching (Jing et al., 2024a); diffusion and graph methods model equilibrium conformations (Jing et al., 2023; Lu et al., 2024; Zheng et al., 2024; Wang et al., 2024); and BioEmu, Energy-based Alignment, and JAMUN improve state coverage or physical weighting (Lewis et al., 2025; Lu et al., 2025a; Daigavane et al., 2025). De novo backbone generators (Yim et al., 2023; Lin & Alquraishi, 2023; Wu et al., 2024) likewise produce unordered structures. Trajectory models instead learn temporal transitions through finite-lag propagation (Klein et al., 2023; dos Santos Costa et al., 2024; 2025), joint frame modeling (Cheng et al., 2025; Jing et al., 2024b), autoregressive diffusion (Shen et al., 2025), spatiotemporal attention (Shoghi et al., 2026), or structure tokens (Liu et al., 2026). MotionPlan is orthogonal: it adds inference-time inter-residue control to a trained generator.

2.2 Inference-time diffusion guidance

Diffusion guidance steers a pretrained reverse process with a condition or external objective, building on denoising diffusion models (Ho et al., 2020). Classifier and classifier-free guidance modify the denoising prediction (Dhariwal & Nichol, 2021; Ho & Salimans, 2022), while Universal Guidance and diffusion posterior sampling incorporate general objectives or measurement consistency (Bansal et al., 2023; Chung et al., 2023). Training-free energy or differentiable-loss guidance further supports plug-and-play control (Yu et al., 2023; Song et al., 2023); decoupled annealing and twisted sequential Monte Carlo improve exploration or conditional sampling beyond local gradient corrections (Zhang et al., 2025; Wu et al., 2023). In protein generation, Energy-based Alignment uses physical feedback with additional optimization (Lu et al., 2025a). MotionPlan instead keeps all weights fixed and injects a signed coordination bias directly into the protein pair representation, avoiding class gradients, coordinate-space likelihoods, and training-stage alignment.

2.3 PROTEIN LANGUAGE MODELS

Protein language models span large-scale sequence representation (Elnaggar et al., 2022; Lin et al., 2023), structure-aware vocabulary learning (Su et al., 2024), sequence generation, and multimodal sequence–structure–function modeling (Madani et al., 2023; Hayes et al., 2025). Structure-token systems represent conformations discretely (Lu et al., 2025b); MD-LLM-1 generates tokenized conformational sequences (Murtada et al., 2025); and ProTDyn extends this paradigm to multi-timescale dynamics (Liu et al., 2026). MotionPlan uses an LLM differently: as a frozen evaluator that converts residue identity, secondary structure, contacts, and exposure into coordination roles for a continuous-coordinate generator.

3 METHOD

3.1 OVERVIEW AND PROBLEM FORMULATION

Let $a = (a_1, \ldots, a_N)$ denote a protein sequence and let $x_t \in \mathbb{R}^{N\times 3}$ denote the Cα coordinates of its state at physical time index t; the same construction applies when the backbone state contains residue frames or additional atoms. Given an observed state x_0, a pretrained dynamics generator samples $X_{1:T} = (x_1, \ldots, x_T)$ from $p_\theta(X_{1:T} \mid a, x_0)$, where all parameters θ are frozen. MotionPlan exposes an inference-time action over *inter-residue coordination*: it specifies which regions remain tightly coupled, which regions decouple, and which regions mediate their boundary, while the generator's learned geometric prior realizes the corresponding coordinate evolution. Figure 2 presents this pipeline from observed context and a reusable residue motion plan, through the residue-to-pair lift and gated pair offset, to repeated pair-feature control inside the frozen diffusion generator.

We consider the common pair-feature diffusion interface. At outer trajectory step t, the backbone constructs single features $s_t \in \mathbb{R}^{N\times d_s}$, pair features $z_t \in \mathbb{R}^{N\times N\times d_z}$, and an optional temporal state h_t. Starting from a noisy structural state $y_t^{(K)}$, its k-th reverse-SDE update has the abstract form

$$y_t^{(k-1)} = \Phi_\theta\Big(y_t^{(k)}, k; s_t, z_t, h_t\Big) + \sigma_k \epsilon_k, \qquad \epsilon_k \sim \mathcal{N}(0, I), \tag{1}$$

for $k = K, \ldots, 1$. Here Φ_θ includes the equivariant coordinate decoder and the base reverse update. The final $y_t^{(0)}$ is decoded as x_{t+1} and incorporated into h_{t+1} when the generator maintains temporal state. MotionPlan supplies a shape-preserving relational offset to z_t, while the frozen Φ_θ retains coordinate generation and all learned parameters.

3.2 LLM COORDINATION-ROLE EVALUATOR

MotionPlan separates *reasoning about coordination* from generating a concrete trajectory. An instruction-following LLM evaluator E_ψ maps the observable sequence and structural context o_t to a residue coordination plan,

$$M_t = E_\psi(o_t). \tag{2}$$

The observation contains only information available from the conditioning input: amino-acid identity, the initial structure, secondary structure, terminal proximity, contact count and density, graph centrality, burial depth, solvent exposure, and local packing. It never includes a future frame, reference displacement, RMSF, PCA/tICA coordinates, evaluation stride, or any score derived from the target trajectory.

The LLM jointly interprets these cues to assign contextual coordination roles and evidential confidence. Buried, well-connected residues tend to be rigid anchors, exposed loops and termini flexible executors, and their interfaces transitional linkers. Confidence is a five-level ordinal target whose order reflects increasing support. The structured residue-indexed plan is

$$M_t = \{m_{t,i}\}_{i=1}^{N}, \qquad m_{t,i} = (c_{t,i}, r_{t,i}), \tag{3}$$

where $c_{t,i} \in \{0, 1, 2\}$ denotes LOW, MEDIUM, or HIGH and $r_{t,i} \in \{1, 2, 3, 4, 5\}$ is the emitted ordinal confidence level. Level 1 means weak evidence for the assigned role and level 5 means strong evidence; the output schema therefore contains a discrete five-level score rather than an arbitrary real value. MotionPlan converts this score to an internal amplitude weight only after inference,

$$\rho_{t,i} = \frac{r_{t,i} - 1}{4} \in \left\{0, \frac{1}{4}, \frac{1}{2}, \frac{3}{4}, 1\right\}. \tag{4}$$

Semantically, LOW marks a rigid anchor, MEDIUM a transitional linker, and HIGH a flexible executor. The plan contains exactly one valid role and one integer confidence level for every residue. This joint role–confidence output preserves the ordinal regression semantics at the evaluator interface, while $\rho_{t,i}$ provides a calibrated weight for pair control.

3.3 FROM ORDINAL ROLE CONFIDENCE TO SIGNED PAIR CONTROL

Pair-feature generators naturally represent residue–residue relations, making them the appropriate control surface for the LLM output. MotionPlan lifts the joint role–confidence plan to an offset $O_t \in \mathbb{R}^{N\times N\times d_z}$ with the same shape as the pair representation. We assign the three coordination roles ordered scalar values

$$s_0 = -1, \qquad s_1 = 0, \qquad s_2 = +1, \tag{5}$$

corresponding to rigid coordination, transition, and flexible decoupling. The order supplies an explicit semantic axis: negative values favor rigid coordination, positive values favor flexible decoupling, and zero is neutral. We use a single fixed unit-norm basis $u \in \mathbb{R}^{d_z}$, selected once with a fixed seed and shared across all proteins and samples. Because every role combination acts along this common axis, the sign and magnitude are determined entirely by the ordered role values and ordinal confidence rather than by a learned or role-specific random codebook.

For a valid-residue mask $b_i \in \{0, 1\}$ and sequence-locality factor

$$\ell_{ij} = \exp\left(-\frac{|i-j|}{\max(4, N/4)}\right), \tag{6}$$

define the mean normalized confidence and signed pair coefficient as

$$\bar{\rho}_{t,ij} = \frac{\rho_{t,i} + \rho_{t,j}}{2}, \qquad \alpha_{t,ij} = \frac{s_{c_{t,i}} + s_{c_{t,j}}}{2}\,\bar{\rho}_{t,ij}. \tag{7}$$

The pair offset is

$$O_{t,ij} = m\, b_i b_j\, \ell_{ij}\, \alpha_{t,ij}\, u, \tag{8}$$

where m controls the global intervention strength. For a confident LOW–LOW pair, $\alpha_{t,ij} = -\bar{\rho}_{t,ij}$ produces the strongest negative coordination bias; for a confident HIGH–HIGH pair, $\alpha_{t,ij} = +\bar{\rho}_{t,ij}$ produces the strongest positive decoupling bias. A LOW–HIGH pair has $\alpha_{t,ij} = 0$, expressing a neutral boundary, while LOW–MEDIUM and MEDIUM–HIGH pairs receive half-strength signed biases. MEDIUM–MEDIUM is also neutral. Because both the role average and confidence average are symmetric, $O_{t,ij} = O_{t,ji}$ before any backbone-specific processing. The fixed basis determines the pair-space direction, the role scale determines its sign, and the ordinal confidence controls its magnitude. Sequence locality emphasizes nearby interactions while retaining long-range coordination. The operation costs $O(N^2 d_z)$, matching the asymptotic size of the pair tensor already present in the backbone.

3.4 REPEATED PAIR-LATENT INJECTION

At reverse step k, MotionPlan presents the controlled pair feature

$$\tilde{z}_t^{(k)} = z_t^{(k)} + \gamma\, q_k\, O_t, \tag{9}$$

where γ is an inference-time gate and q_k controls the injection strength at reverse step k. The controlled sampler is therefore

$$y_t^{(k-1)} = \Phi_\theta\left(y_t^{(k)}, k; s_t, \tilde{z}_t^{(k)}, h_t\right) + \sigma_k \epsilon_k. \tag{10}$$

The pair offset is applied at every reverse step. Equation 9 re-applies the LLM-inferred signed coupling field to the current base pair state, keeping the confidence-weighted relational bias consistent throughout denoising while avoiding offset accumulation across steps. The injected bias organizes coordination, decoupling, and neutral boundaries; the equivariant decoder produces every coordinate and autonomously realizes conformational change under this controlled pair context.

Table 1: Detailed multi-start evaluation at stride 256 and $L = 9$. Cα MAEs are in Å. Correlations are higher-is-better and MAEs are lower-is-better.

	Cα Pearson ↑			PCA Pearson ↑			Diversity	PCA MAE ↓			Cα MAE ↓		
Method	Traj.	Frame	Δ	Traj.	Frame	Δ	Pair RMSD	Traj.	Frame	Δ	Traj.	Frame	Δ
MDGen	0.57	0.46	0.41	0.18	0.13	0.11	1.34	7.66	1.54	1.06	6.00	1.14	0.85
MDGen-S256F9	0.56	0.45	0.38	0.21	0.19	0.11	1.59	6.90	1.47	1.03	5.33	1.13	0.82
ConfRover	**0.77**	0.62	0.53	**0.75**	0.50	0.44	1.78	3.91	**1.28**	0.87	4.60	0.91	0.75
MotionPlan	0.75	**0.67**	**0.59**	0.74	**0.61**	**0.48**	2.273	**3.65**	1.46	**0.81**	**3.309**	**0.872**	**0.675**

3.5 Training-free portability and implementation scope

MotionPlan has no optimization objective and updates no generator parameters. Integrating it requires (i) an LLM evaluator that emits the role and five-level ordinal confidence schema in Eq. 3, (ii) access to an $N \times N \times d_z$ pair representation, and (iii) a shape-preserving relational addition before coordinate decoding. The role scale $(s_0, s_1, s_2) = (-1, 0, +1)$, confidence normalization, and fixed basis u are deterministic inference rules shared across proteins. For a step-wise Pairformer–diffusion model, the signed offset is applied to the decoder pair latent independently at each outer frame. For a joint spatiotemporal generator, the same residue-to-pair operator can be broadcast or indexed over time before pair-biased temporal blocks. The sign semantics, ordinal confidence weighting, validity masking, and no-accumulation rule remain unchanged. MotionPlan defines its inference-time action entirely through the LLM-inferred role–confidence plan, with the pretrained generator checkpoint kept fixed.

4 Experiments

4.1 Experimental setup

Data and splits. The frozen generators retain the training data and checkpoints of their corresponding baseline methods; MotionPlan uses no trajectory training data and updates no model parameters. We evaluate on the public ATLAS benchmark (Vander Meersche et al., 2024) using the split and multi-start construction adopted by ConfRover (Shen et al., 2025). The test split contains 82 proteins, each with three 10,000-frame MD replicas saved every 10 ps. For short-horizon evaluation, each sample contains $L = 9$ frames. Start indices are drawn from 1,000–7,000 and temporal strides from $\{128, 256, 512, 1024\}$, yielding 2,706 start/stride conditions in the complete protocol. Short- and long-horizon dynamics are evaluated on the complete 82-protein test split. Molecular quality is evaluated separately on the 38 trajectories shared by the forward and interpolation settings, following the experimental setup of ConfRover (Shen et al., 2025); the stride-256 setting is used for controlled comparisons.

Long-horizon protocol. Following the public-ATLAS evaluation protocol of STAR-MD (Shoghi et al., 2026), we evaluate 100 ns trajectories on the 82-chain test split using 80 frames at 1.2 ns per frame. Every method starts from the same equilibrated structure. All trajectories are aligned to the starting frame of the reference MD trajectory by Cα superposition before computing coverage and kinetic statistics. Coverage is evaluated only on structurally valid conformations, and tICA is evaluated only on time-lagged pairs for which both frames are valid, using the joint Cα and all-atom validity criteria of STAR-MD.

Baselines and implementation. Short-horizon baselines include MDGen, MDGen-S256F9, and ConfRover (Jing et al., 2024b; Shen et al., 2025); the long-horizon comparison includes MD, MDGen, and ConfRover. In every matched controlled run, the checkpoint, conditioning structure, diffusion schedule, sampling horizon, and random seed are fixed. MotionPlan uses the LLM evaluator to infer residue coordination roles and five-level ordinal confidence scores, deterministically converts them into a signed coupling field, and changes only the relational pair representation supplied to the frozen generator. For each seed, MotionPlan and its corresponding baseline use the same initial structure and sampling seed.

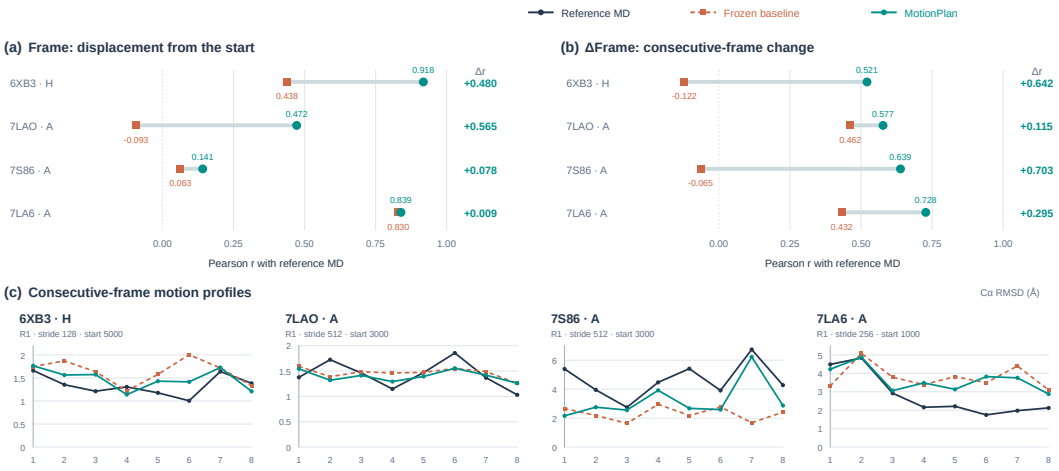


Figure 3: Effect of the LLM-inferred signed coupling field on short-horizon fidelity. MotionPlan assigns negative bias to confident rigid–rigid pairs, positive bias to confident flexible–flexible pairs, and neutral bias to rigid–flexible boundaries; five-level ordinal confidence controls the pair-bias magnitude. Panels (a) and (b) compare Pearson correlation for displacement from the initial structure and consecutive-frame change, respectively; the annotated Δr values show the gain from the relational intervention. Panel (c) plots consecutive-frame Cα RMSD profiles for four representative systems, where MotionPlan better follows the reference peaks and troughs.

Table 2: Long-horizon trajectory generation on the public ATLAS 100 ns benchmark with 82 proteins and 80 frames sampled at 1.2 ns intervals, following the evaluation protocol of STAR-MD (Shoghi et al., 2026). JSD and recall are computed on structurally valid conformations; tICA uses time-lagged pairs with both frames valid. Entries with $\pm$ report the mean and standard deviation across independent random seeds.

Method	JSD↓	Recall↑	tICA↑	RMSD Δ ↓	AutoCorr Δ ↓	VAMP-2 Δ ↓
MD oracle	0.31	0.67	0.17	0.00	0.00	0.02
MDGen	0.56 ± 0.01	0.28 ± 0.01	0.12 ± 0.00	0.38 ± 0.01	0.05 ± 0.00	0.38 ± 0.01
ConfRover	0.52 ± 0.01	0.36 ± 0.01	0.15 ± 0.01	0.20 ± 0.00	0.08 ± 0.00	0.47 ± 0.02
MotionPlan	$\mathbf{0.44\pm0.02}$	$\mathbf{0.59\pm0.01}$	$\mathbf{0.16\pm0.00}$	$\mathbf{0.18\pm0.02}$	$\mathbf{0.02\pm0.01}$	$\mathbf{0.27\pm0.01}$

4.2 Evaluation metrics

Short-horizon dynamics. Following ConfRover, we summarize conformational change at three temporal resolutions: *Trajectory*, the total change accumulated over a rollout; *Frame*, each frame's displacement from the initial structure after alignment; and Δ*Frame*, the displacement between consecutive aligned frames. Pearson correlation is computed across matched start/stride conditions between generated and reference changes in Cα coordinate and protein-specific 2D PCA spaces, measuring whether the model preserves their relative variation. MAE averages the corresponding absolute differences and therefore measures calibration in the original displacement scale. Pairwise RMSD is the average aligned RMSD across generated frame pairs and measures whether the rollout retains conformational spread.

Long-horizon dynamics and coverage. Following STAR-MD (Shoghi et al., 2026), Table 2 reports Jensen–Shannon distance (JSD) between generated and reference state distributions and reference-state recall in a PCA space fitted to reference MD conformations, with each principal component discretized into 10 evenly spaced bins. Both coverage metrics use only structurally valid generated conformations. The tICA metric measures agreement between per-residue contributions to the leading slow modes of generated and reference trajectories, using only time-lagged pairs whose two frames satisfy the joint Cα and all-atom validity criteria. RMSD, autocorrela-

Table 3: Molecular quality of relaxed conformations on the 38 trajectories shared by the forward and interpolation settings, following the experimental setup of ConfRover (Shen et al., 2025). All methods use the same relaxation and geometry-evaluation protocol. Geometric metrics are reported as mean±standard deviation across the 38 cases; MadraX energy is reported as the mean across these cases.

Method	Rama. outlier%↓	Rotamer outlier%↓	Clashscore↓	RMS bonds↓	RMS angles↓	MolProbity↓	MadraX↓
MD Reference	$\mathbf{0.38 \pm 0.49}$	$\mathbf{1.02 \pm 0.89}$	$\mathbf{0.04 \pm 0.16}$	$\mathbf{0.01 \pm 0.00}$	$\mathbf{1.88 \pm 0.05}$	$\mathbf{0.72 \pm 0.18}$	-519.3
MDGen	0.93 ± 0.86	2.86 ± 1.59	16.14 ± 20.05	0.02 ± 0.02	2.13 ± 0.30	2.24 ± 0.40	-314.7
ConfRover	0.58 ± 0.63	1.98 ± 1.48	7.81 ± 6.52	$\mathbf{0.01 \pm 0.01}$	$\mathbf{1.88 \pm 0.25}$	1.72 ± 0.38	$\mathbf{-522.2}$
MotionPlan	0.55 ± 0.61	2.06 ± 1.59	8.03 ± 6.34	$\mathbf{0.01 \pm 0.01}$	1.92 ± 0.26	1.70 ± 0.37	-518.7

tion, and VAMP-2 deviations are absolute discrepancies from the reference trajectory in structural spread, temporal persistence, and kinetic variational score, respectively (Wu & Noé, 2020).STAR-MD reports CA/AA/CA+AA validity thresholds but not the full Cα clash/break implementation, and its evaluation code is unavailable. We therefore omit these metrics to avoid implementation-dependent approximations, while assessing structural validity and quality separately under the ConfRover relaxed-structure protocol in Table 3.

Molecular quality. Following ConfRover's relaxed-structure protocol (Shen et al., 2025), all compared structures are refined with OpenFold (Ahdritz et al., 2024) and evaluated under one geometry pipeline. Ramachandran and rotamer outlier rates measure backbone and side-chain torsion violations, clashscore counts severe steric overlaps, and bond/angle RMS values measure covalent-geometry deviations. MolProbity aggregates stereochemical quality (Chen et al., 2010), while MadraX reports molecular-mechanics energy.

4.3 Short-horizon Results

Table 1 shows that MotionPlan provides the strongest improvement in local temporal organization while preserving trajectory-level coverage. Its signed pair bias gives the frozen generator an explicit relational prior: confident rigid–rigid pairs are encouraged to remain coordinated, flexible–flexible pairs are allowed to decouple, and mixed boundaries are left comparatively neutral. Ordinal confidence further prevents uncertain assignments from dominating the reverse process. The resulting trajectories therefore follow reference frame-to-frame changes more faithfully without collapsing conformational diversity, indicating that the gain comes from better organized motion rather than conservative sampling. Figure 3 shows how this mechanism appears at the system level.

4.4 Long-horizon Results

As shown in Table 2, MotionPlan achieves the most balanced long-horizon behavior, recovering broader reference-state support while maintaining the slow collective modes and temporal persistence of the underlying dynamics. Reapplying the signed coupling field throughout denoising prevents the relational prior from being washed out as errors accumulate across a rollout. Rigid regions remain coherent anchors, whereas flexible regions can explore alternative metastable states under the generator's learned coordinate-space dynamics. This separation explains why the method improves both conformational coverage and kinetic fidelity rather than trading one for the other.

4.5 Molecular Quality

Table 3 confirms that MotionPlan preserves the relaxed-structure quality of the frozen baseline while improving how conformational changes are organized over time. This behavior follows from applying control in pair-latent space rather than directly perturbing coordinates: the intervention changes which residues move together, while the pretrained equivariant decoder and the shared relaxation procedure continue to enforce local stereochemistry and covalent geometry. The molecular-quality results therefore rule out invalid or highly strained structures as the source of the dynamic gains.

Table 4: Factorized short-horizon ablation for MotionPlan at stride 256 and $L = 9$. The three panels separately vary the coupling-plan source, residue-to-pair lift, and injection schedule. All metrics are computed in Cα coordinate space; correlations are higher-is-better and MAEs are lower-is-better.

(a) Coupling-plan source

Variant	Frame $r \uparrow$	ΔFrame $r \uparrow$	Traj. MAE$\downarrow$	ΔFrame MAE$\downarrow$
Deterministic proxy evaluator	0.63	0.54	4.21	0.742
Shuffled role assignment	0.61	0.51	4.47	0.771
Random role assignment	0.59	0.49	4.93	0.801
Confidence-free scaling	0.65	0.57	3.61	0.699
MotionPlan	**0.67**	**0.59**	**3.309**	**0.675**

(b) Residue-to-pair lift

Variant	Frame $r \uparrow$	ΔFrame $r \uparrow$	Traj. MAE$\downarrow$	ΔFrame MAE$\downarrow$
Unsigned role scale	0.58	0.48	5.08	0.823
Negated basis ($-u$)	0.55	0.45	5.61	0.876
No locality decay	0.65	0.56	3.72	0.706
MotionPlan	**0.67**	**0.59**	**3.309**	**0.675**

(c) Injection schedule

Variant	Frame $r \uparrow$	ΔFrame $r \uparrow$	Traj. MAE$\downarrow$	ΔFrame MAE$\downarrow$
First-step injection only	0.62	0.53	4.57	0.754
MotionPlan	**0.67**	**0.59**	**3.309**	**0.675**

4.6 Ablation Study

Table 4 isolates each design choice. Panel (a) shows that plan quality matters: shuffling roles while preserving their aggregate distribution already degrades performance, confirming that the gain requires assigning the right coordination intent to the right regions, not just the right role counts; completely random assignment degrades further. Panel (b) shows that the lift must preserve direction and locality: inverting the basis to $-u$ produces the worst result in this panel, as it actively reverses coupling and decoupling throughout the field, while removing the role sign or locality decay each degrades results in smaller but consistent ways. Panel (c) confirms that repeated injection is essential, since a single offset is progressively absorbed by the frozen reverse process.

5 Conclusion

Protein dynamics generation requires more than plausible conformations: it must preserve the inter-residue relationships that organize frame-to-frame transitions. Existing generators learn structural statistics but cannot specify residue coordination at inference time, so local errors accumulate and training a separate controller would require paired motion supervision that MD data do not provide. MotionPlan fills this relational-control gap by translating sequence and structural evidence into residue-level intent and injecting it through the pair representations of a frozen generator, while leaving coordinate generation to the pretrained dynamics model. Improvements across short and long horizons show that this prior improves temporal organization rather than merely matching isolated states. Pair representations therefore offer a reusable control interface that can also support future learnable or physics-aware evaluators. We also discuss the current limitations of MotionPlan and promising directions for future work in Appendix C.

AI use statement. Generative AI tools were used to aid and polish writing and to assist with the retrieval and discovery of related work. All AI-assisted content was carefully reviewed, verified, and, where necessary, revised by the authors. In particular, retrieved literature and associated claims

were checked against the original sources. We take responsibility for the final content of this work, including text, claims, or artifacts produced with the aid of generative AI.

Reproducibility statement. MotionPlan requires no parameter updates; the main text and appendix specify the LLM setup, prompt, pair-latent intervention and evaluation protocol. Code will be released upon acceptance.

REFERENCES

Josh Abramson, Jonas Adler, Jack Dunger, Richard Evans, Tim Green, Alexander Pritzel, Olaf Ronneberger, Lindsay Willmore, Andrew J. Ballard, Joshua Bambrick, Sebastian W. Bodenstein, David A. Evans, Chia-Chun Hung, Michael O'Neill, David Reiman, Kathryn Tunyasuvunakool, Zachary Wu, Akvilė Žemgulytė, Eirini Arvaniti, Charles Beattie, Ottavia Bertolli, Alex Bridgland, Alexey Cherepanov, Miles Congreve, Alexander I. Cowen-Rivers, Andrew Cowie, Michael Figurnov, Fabian B. Fuchs, Hannah Gladman, Rishub Jain, Yousuf A. Khan, Caroline M. R. Low, Kuba Perlin, Anna Potapenko, Pascal Savy, Sukhdeep Singh, Adrian Stecula, Ashok Thillaisundaram, Catherine Tong, Sergei Yakneen, Ellen D. Zhong, Michal Zielinski, Augustin Žídek, Victor Bapst, Pushmeet Kohli, Max Jaderberg, Demis Hassabis, and John M. Jumper. Accurate structure prediction of biomolecular interactions with AlphaFold 3. *Nature*, 630(8016): 493–500, 2024. doi: 10.1038/s41586-024-07487-w. URL https://doi.org/10.1038/s41586-024-07487-w.

Gustaf Ahdritz, Nazim Bouatta, Christina Floristean, Sachin Kadyan, Qinghui Xia, William Gerecke, Timothy J. O'Donnell, Daniel Berenberg, Ian Fisk, Niccolò Zanichelli, Bo Zhang, Arkadiusz Nowaczynski, Bei Wang, Marta M. Stepniewska-Dziubinska, Shang Zhang, Adegoke Ojewole, Murat Efe Guney, Stella Biderman, Andrew M. Watkins, Stephen Ra, Pablo Ribalta Lorenzo, Lucas Nivon, Brian Weitzner, Yih-En Andrew Ban, Shiyang Chen, Minjia Zhang, Conglong Li, Shuaiwen Leon Song, Yuxiong He, Peter K. Sorger, Emad Mostaque, Zhao Zhang, Richard Bonneau, and Mohammed AlQuraishi. OpenFold: Retraining AlphaFold2 yields new insights into its learning mechanisms and capacity for generalization. *Nature Methods*, 21:1514–1524, 2024. doi: 10.1038/s41592-024-02272-z.

Arpit Bansal, Hong-Min Chu, Avi Schwarzschild, Soumyadip Sengupta, Micah Goldblum, Jonas Geiping, and Tom Goldstein. Universal guidance for diffusion models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*, pp. 843–852, June 2023.

Vincent B. Chen, III Arendall, W. Bryan, Jeffrey J. Headd, Daniel A. Keedy, Robert M. Immormino, Gary J. Kapral, Laura W. Murray, Jane S. Richardson, and David C. Richardson. MolProbity: All-atom structure validation for macromolecular crystallography. *Acta Crystallographica Section D: Biological Crystallography*, 66:12–21, 2010. doi: 10.1107/S0907444909042073.

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. doi: 10.1609/aaai.v39i1.31984. URL https://doi.org/10.1609/aaai.v39i1.31984.

Hyungjin Chung, Jeongsol Kim, Michael Thompson Mccann, Marc Louis Klasky, and Jong Chul Ye. Diffusion posterior sampling for general noisy inverse problems. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=OnD9zGAGT0k.

Ameya Daigavane, Bodhi Vani, Darcy Davidson, Saeed Saremi, Joshua Rackers, and Joseph Kleinhenz. Jamun: Bridging smoothed molecular dynamics and score-based learning for conformational ensemble generation. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, Main Conference, pp. 36769–36801. Curran Associates, Inc., 2025. doi: 10.52202/085713-1236. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/3476352269e5a76b91cb4670390f1b5c-Paper-Conference.pdf.

Prafulla Dhariwal and Alexander Nichol. Diffusion models beat gans on image synthesis. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan (eds.), *Advances in Neural Information Processing Systems*, volume 34, pp. 8780–8794. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/49ad23d1ec9fa4bd8d77d02681df5cfa-Paper.pdf.

Allan dos Santos Costa, Ilan Mitnikov, Franco Pellegrini, Ameya Daigavane, Mario Geiger, Zhonglin Cao, Karsten Kreis, Tess Smidt, Emine Kucukbenli, and Joseph Jacobson. Equijump: Protein dynamics simulation via so(3)-equivariant stochastic interpolants, 2024. URL https://arxiv.org/abs/2410.09667.

Allan dos Santos Costa, Manvitha Ponnapati, Dana Rubin, Tess Smidt, and Joseph Jacobson. Accelerating protein molecular dynamics simulation with deepjump, 2025. URL https://arxiv.org/abs/2509.13294.

Ahmed Elnaggar, Michael Heinzinger, Christian Dallago, Ghalia Rehawi, Yu Wang, Llion Jones, Tom Gibbs, Tamas Feher, Christoph Angerer, Martin Steinegger, Debsindhu Bhowmik, and Burkhard Rost. Prottrans: Toward understanding the language of life through self-supervised learning. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 44(10):7112–7127, 2022. doi: 10.1109/TPAMI.2021.3095381.

Thomas Hayes, Roshan Rao, Halil Akin, Nicholas J. Sofroniew, Deniz Oktay, Zeming Lin, Robert Verkuil, Vincent Q. Tran, Jonathan Deaton, Marius Wiggert, Rohil Badkundri, Irhum Shafkat, Jun Gong, Alexander Derry, Raul S. Molina, Neil Thomas, Yousuf A. Khan, Chetan Mishra, Carolyn Kim, Liam J. Bartie, Matthew Nemeth, Patrick D. Hsu, Tom Sercu, Salvatore Candido, and Alexander Rives. Simulating 500 million years of evolution with a language model. *Science*, 387(6736):850–858, 2025. doi: 10.1126/science.ads0018. URL https://www.science.org/doi/abs/10.1126/science.ads0018.

Katherine Henzler-Wildman and Dorothee Kern. Dynamic personalities of proteins. *Nature*, 450 (7172):964–972, 2007. doi: 10.1038/nature06522. URL https://doi.org/10.1038/nature06522.

Jonathan Ho and Tim Salimans. Classifier-free diffusion guidance, 2022. URL https://arxiv.org/abs/2207.12598.

Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 6840–6851. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/4c5bcfec8584af0d967f1ab10179ca4b-Paper.pdf.

Bowen Jing, Ezra Erives, Peter Pao-Huang, Gabriele Corso, Bonnie Berger, and Tommi Jaakkola. Eigenfold: Generative protein structure prediction with diffusion models, 2023. URL https://arxiv.org/abs/2304.02198.

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. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/478b06f60662d3cdc1d4f15d4587173a-Paper-Conference.pdf.

John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, Alex Bridgland,

Clemens Meyer, Simon A. A. Kohl, Andrew J. Ballard, Andrew Cowie, Bernardino Romera-Paredes, Stanislav Nikolov, Rishub Jain, Jonas Adler, Trevor Back, Stig Petersen, David Reiman, Ellen Clancy, Michal Zielinski, Martin Steinegger, Michalina Pacholska, Tamas Berghammer, Sebastian Bodenstein, David Silver, Oriol Vinyals, Andrew W. Senior, Koray Kavukcuoglu, Pushmeet Kohli, and Demis Hassabis. Highly accurate protein structure prediction with AlphaFold. *Nature*, 596(7873):583–589, 2021. doi: 10.1038/s41586-021-03819-2. URL https://doi.org/10.1038/s41586-021-03819-2.

Leon Klein, Andrew Foong, Tor Fjelde, Bruno Mlodozeniec, Marc Brockschmidt, Sebastian Nowozin, Frank Noe, and Ryota Tomioka. Timewarp: Transferable acceleration of molecular dynamics by learning time-coarsened dynamics. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 52863–52883. Curran Associates, Inc., 2023. doi: 10.52202/075280-2300. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/a598c367280f9054434fdcc227ce4d38-Paper-Conference.pdf.

Sarah Lewis, Tim Hempel, José Jiménez-Luna, Michael Gastegger, Yu Xie, Andrew Y. K. Foong, Victor García Satorras, Osama Abdin, Bastiaan S. Veeling, Iryna Zaporozhets, Yaoyi Chen, Soojung Yang, Adam E. Foster, Arne Schneuing, Jigyasa Nigam, Federico Barbero, Vincent Stimper, Andrew Campbell, Jason Yim, Marten Lienen, Yu Shi, Shuxin Zheng, Hannes Schulz, Usman Munir, Roberto Sordillo, Ryota Tomioka, Cecilia Clementi, and Frank Noé. Scalable emulation of protein equilibrium ensembles with generative deep learning. *Science*, 389(6761):eadv9817, 2025. doi: 10.1126/science.adv9817. URL https://www.science.org/doi/abs/10.1126/science.adv9817.

Yeqing Lin and Mohammed Alquraishi. Generating novel, designable, and diverse protein structures by equivariantly diffusing oriented residue clouds. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 20978–21002. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/lin23a.html.

Zeming Lin, Halil Akin, Roshan Rao, Brian Hie, Zhongkai Zhu, Wenting Lu, Nikita Smetanin, Robert Verkuil, Ori Kabeli, Yaniv Shmueli, Allan dos Santos Costa, Maryam Fazel-Zarandi, Tom Sercu, Salvatore Candido, and Alexander Rives. Evolutionary-scale prediction of atomic-level protein structure with a language model. *Science*, 379(6637):1123–1130, 2023. doi: 10.1126/science.ade2574. URL https://www.science.org/doi/abs/10.1126/science.ade2574.

Yikai Liu, Haoyang Zheng, Lining Mao, Yanbin Wang, Ming Chen, and Guang Lin. ProT-Dyn: A foundation protein language model for thermodynamics and dynamics generation. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=fvCHkWbdgX.

Jiarui Lu, Bozitao Zhong, Zuobai Zhang, and Jian Tang. Str2str: A score-based framework for zero-shot protein conformation sampling. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=C4BikKsgmK.

Jiarui Lu, Xiaoyin Chen, Stephen Zhewen Lu, Aurelie Lozano, Vijil Chenthamarakshan, Payel Das, and Jian Tang. Aligning protein conformation ensemble generation with physical feedback. In Aarti Singh, Maryam Fazel, Daniel Hsu, Simon Lacoste-Julien, Felix Berkenkamp, Tegan Maharaj, Kiri Wagstaff, and Jerry Zhu (eds.), *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 40436–40451. PMLR, 13–19 Jul 2025a. URL https://proceedings.mlr.press/v267/lu25b.html.

Jiarui Lu, Xiaoyin Chen, Stephen Zhewen Lu, Chence Shi, Hongyu Guo, Yoshua Bengio, and Jian Tang. Structure language models for protein conformation generation. In *The Thirteenth International Conference on Learning Representations*, 2025b. URL https://openreview.net/forum?id=OzUNDnpQyd.

Ali Madani, Ben Krause, Eric R. Greene, Subu Subramanian, Benjamin P. Mohr, James M. Holton, Jose Luis Olmos Jr., Caiming Xiong, Zachary Z. Sun, Richard Socher, James S. Fraser, and Nikhil Naik. Large language models generate functional protein sequences across diverse families. *Nature Biotechnology*, 41:1099–1106, 2023. doi: 10.1038/s41587-022-01618-2.

Mhd Hussein Murtada, Z. Faidon Brotzakis, and Michele Vendruscolo. Md-llm-1: A large language model for molecular dynamics, 2025. URL https://arxiv.org/abs/2508.03709.

Yuning Shen, Lihao Wang, Huizhuo Yuan, Yan Wang, Bangji Yang, and Quanquan Gu. Simultaneous modeling of protein conformation and dynamics via autoregression. In D. Belgrave, C. Zhang, H. Lin, R. Pascanu, P. Koniusz, M. Ghassemi, and N. Chen (eds.), *Advances in Neural Information Processing Systems*, volume 38, pp. 109782–109822. Curran Associates, Inc., 2025. doi: 10.52202/085713-3666. URL https://proceedings.neurips.cc/paper_files/paper/2025/file/9e1b570c9f8df9ea884b3307e86a7810-Paper-Conference.pdf.

Nima Shoghi, Yuxuan Liu, Yuning Shen, Rob Brekelmans, Pan Li, and Quanquan Gu. Scalable spatio-temporal SE(3) diffusion for long-horizon protein dynamics. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=Q1JpRZkR3S.

Jiaming Song, Qinsheng Zhang, Hongxu Yin, Morteza Mardani, Ming-Yu Liu, Jan Kautz, Yongxin Chen, and Arash Vahdat. Loss-guided diffusion models for plug-and-play controllable generation. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 32483–32498. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/song23k.html.

Jin Su, Chenchen Han, Yuyang Zhou, Junjie Shan, Xibin Zhou, and Fajie Yuan. Saprot: Protein language modeling with structure-aware vocabulary. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=6MRm3G4NiU.

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. URL https://doi.org/10.1093/nar/gkad1084.

Yan Wang, Lihao Wang, Yuning Shen, Yiqun Wang, Huizhuo Yuan, Yue Wu, and Quanquan Gu. Protein conformation generation via force-guided SE(3) diffusion models. 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. 56835–56859. PMLR, 21–27 Jul 2024. URL https://proceedings.mlr.press/v235/wang24cv.html.

Hao Wu and Frank Noé. Variational approach for learning markov processes from time series data. *Journal of Nonlinear Science*, 30(1):23–66, 2020. doi: 10.1007/s00332-019-09567-y.

Kevin E. Wu, Kevin K. Yang, Rianne van den Berg, Sarah Alamdari, James Y. Zou, Alex X. Lu, and Ava P. Amini. Protein structure generation via folding diffusion. *Nature Communications*, 15:1059, 2024. doi: 10.1038/s41467-024-45051-2. URL https://doi.org/10.1038/s41467-024-45051-2.

Luhuan Wu, Brian Trippe, Christian Naesseth, David Blei, and John Cunningham. Practical and asymptotically exact conditional sampling in diffusion models. In A. Oh, T. Naumann, A. Globerson, K. Saenko, M. Hardt, and S. Levine (eds.), *Advances in Neural Information Processing Systems*, volume 36, pp. 31372–31403. Curran Associates, Inc., 2023. doi: 10.52202/075280-1363. URL https://proceedings.neurips.cc/paper_files/paper/2023/file/63e8bc7bbf1cfea36d1d1b6538aecce5-Paper-Conference.pdf.

Jason Yim, Brian L. Trippe, Valentin De Bortoli, Emile Mathieu, Arnaud Doucet, Regina Barzilay, and Tommi Jaakkola. SE(3) diffusion model with application to protein backbone generation.

In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *Proceedings of the 40th International Conference on Machine Learning*, volume 202 of *Proceedings of Machine Learning Research*, pp. 40001–40039. PMLR, 23–29 Jul 2023. URL https://proceedings.mlr.press/v202/yim23a.html.

Jiwen Yu, Yinhuai Wang, Chen Zhao, Bernard Ghanem, and Jian Zhang. Freedom: Training-free energy-guided conditional diffusion model. In *2023 IEEE/CVF International Conference on Computer Vision (ICCV)*, pp. 23117–23127, 2023. doi: 10.1109/ICCV51070.2023.02118.

Bingliang Zhang, Wenda Chu, Julius Berner, Chenlin Meng, Anima Anandkumar, and Yang Song. Improving diffusion inverse problem solving with decoupled noise annealing. In *2025 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pp. 20895–20905, 2025. doi: 10.1109/CVPR52734.2025.01946.

Shuxin Zheng, Jiyan He, Chang Liu, Yu Shi, Ziheng Lu, Weitao Feng, Fusong Ju, Jiaxi Wang, Jianwei Zhu, Yaosen Min, He Zhang, Shidi Tang, Hongxia Hao, Peiran Jin, Chi Chen, Frank Noé, Haiguang Liu, and Tie-Yan Liu. Predicting equilibrium distributions for molecular systems with deep learning. *Nature Machine Intelligence*, 6(5):558–567, 2024. doi: 10.1038/s42256-024-00837-3. URL https://doi.org/10.1038/s42256-024-00837-3.

A LLM EVALUATOR SELECTION AND ANALYSIS

Selection and call frequency. We use GPT-5.6-sol as the default evaluator because the role assignment requires joint reasoning over sequence identity, secondary structure, contact context, and burial or exposure rather than independent residue classification. The evaluator is executed once for each protein-specific conditioning record, and the resulting ordered residue plan is cached. Each generated trajectory loads this cached plan once and reuses it for all rollout frames and reverse-SDE denoising steps, so the number of LLM calls does not grow with trajectory length, the number of trajectories drawn from that record, or the number of diffusion steps. GPT-5.6-sol provided the strongest short-horizon fidelity in our model-choice study while consistently satisfying the strict output contract.

Table 5: LLM evaluator choice under the short-horizon protocol at stride 256 and $L = 9$. Only the principal Cα fidelity metrics are reported. All evaluators use the same prompt, request wrapper, output-validation contract, residue-to-pair lift, and every-step injection schedule.

Evaluator	Frame $r \uparrow$	ΔFrame $r \uparrow$	Traj. MAE$\downarrow$
GPT-5.6-sol	**0.67**	**0.59**	**3.309**
Opus-4.8	0.66	0.58	3.42
GPT-5.6-luna	0.65	0.56	3.64
Seed-2.1-pro	0.64	0.55	3.78
DeepSeek-4.0-pro	0.63	0.54	4.02
Qwen-7B	0.60	0.50	4.72

Table 5 shows a monotonic degradation as evaluator capacity decreases. Opus-4.8 is the closest alternative, but GPT-5.6-sol remains best on all three metrics. Smaller evaluators most often confuse flexible loops with neutral boundary residues or assign uniformly high confidence, which weakens the signed pair field. Because the cached plan is reused throughout generation, the selected evaluator adds no per-frame or per-denoising-step LLM latency.

B PROMPT TEMPLATE AND OUTPUT SCHEMA

The production prompt is deliberately compact and excludes future coordinates, reference trajectories, force-field energies, and evaluation labels. It presents residues in sequence order and requests one role and one five-level ordinal confidence value per residue. The following pseudocode captures the model-facing contract; implementation-specific transport metadata is omitted.

```
SYSTEM:
  Infer a residue-level coordination plan for a frozen protein-dynamics
  generator. Use only the supplied sequence and structural context.

ROLE SEMANTICS:
  Low    = constrained/rigid residue; supports negative pair bias
  Medium = neutral or rigid-flexible boundary; supports near-zero bias
  High   = flexible/decoupled residue; supports positive pair bias

CONFIDENCE:
  Integer in {1,2,3,4,5}; 1 = weak evidence, 5 = strong evidence.
  Confidence changes bias magnitude, never the meaning of the role.

INPUT:
  sequence: [a_1, ..., a_N]
  residues: for i = 1..N
    {index: i,
     secondary_structure: ss_i,
     contact_context: contacts_i,
     burial_exposure: exposure_i}

OUTPUT (JSON only, sequence order, exactly N entries):
  {"residues": [
    {"index": 1, "role": "Low|Medium|High", "confidence": 1..5},
    ...
  ]}

CONSTRAINTS:
  Do not output coordinates, forces, energies, prose, or extra keys.
  Do not omit, duplicate, or reorder residues.
```

The response is accepted only if it is valid JSON, contains exactly N entries in the input order, uses a role in $\{\text{LOW}, \text{MEDIUM}, \text{HIGH}\}$, and assigns an integer confidence in $\{1, 2, 3, 4, 5\}$. Parsing or schema failure is recorded explicitly and the response is excluded rather than repaired with invented residue labels. After validation, confidence is normalized as $\rho_i = (r_i - 1)/4$ and the plan is lifted to pair space exactly as defined in Section 3.

C Limitations and Future Work

Limitations. The evaluator reasons from sequence and compact structural descriptors rather than an explicit physical force field. It therefore cannot directly test whether a proposed role assignment lowers energy, preserves solvent-mediated interactions, or respects the instantaneous force balance. The three roles and five confidence levels are intentionally coarse, and a plan produced once per protein-specific conditioning record cannot adapt when the generated structure enters a new metastable state. The current lift also projects every scalar coefficient through a fixed pair-space basis u; this preserves the frozen-generator interface but cannot learn channel-specific control directions. Finally, the global magnitude m, injection strength γ, and denoising schedule q_k are selected manually and may require recalibration across generators, proteins, or rollout horizons. Dependence on an external LLM also introduces latency and availability concerns; strict schema validation prevents malformed output from entering evaluation, but an unavailable or invalid response leaves that case without an LLM plan. At the time of writing, the training code, model checkpoints, and training data for the current state-of-the-art STAR-MD are not publicly available, which precludes evaluating whether our control mechanism transfers to STAR-MD.

Future work. A natural extension is to replace the LLM with a physics-aware learnable evaluator. Such a module could consume local geometry, forces, energy terms, solvent exposure, and temporal state, then predict either the residue plan or the pair control field directly. Joint training with the trajectory generator would turn MotionPlan from a training-free intervention into an end-to-end differentiable coordination controller, supervised by trajectory reconstruction together with energy, geometry, and kinetic-consistency objectives. The pair basis u, magnitude m, injection strength γ, and schedule q_k could then be learned rather than fixed. LLM plans may still provide initialization

or distillation targets, while the final controller grounds its decisions in physical state and adapts them throughout the rollout.

D Implementation and Reproducibility Details

Control hyperparameters and denoising schedule. The configuration used for the reported ConfRover experiments sets the offset magnitude to $m = 0.50$ and the inference-time gate to $\gamma = 14$. We use 100 reverse-SDE steps and a constant schedule,

$$q_k = 1, \qquad k = 1, \ldots, 100. \tag{11}$$

Thus, the same scaled offset is presented at every reverse step. The implementation forms each controlled pair state from the current unmodified pair latent plus the scaled offset; it does not add the offset to an already controlled state, thereby preventing accumulation across denoising steps. Invalid or padded residues are removed by the outer-product validity mask, and the locality factor is implemented exactly as $\exp[-|i-j|/\max(4, N/4)]$.

Backbone checkpoint and injection site. All controlled and uncontrolled ConfRover runs load the same frozen checkpoint, `confrover_base_20m_v1_0.pt`. MotionPlan modifies the pair latent only: the shape-preserving offset is added immediately before the reverse-step denoiser consumes that pair state, while the single-residue latent and all model parameters remain unchanged. The same base pair latent is used to reconstruct the controlled input at the next step.

LLM execution and validation. The closed-model plan-generation job invokes the model identifier `gpt-5.6-sol` once per protein-specific conditioning record through the evaluator CLI, with a 900-second per-call timeout and four concurrent workers. The wrapper does not set temperature, top-p, or an output-token cap, so these values are inherited from the serving backend rather than claimed as paper-controlled hyperparameters. It extracts the final JSON array, normalizes role strings, validates the role and confidence fields, and accepts a plan only when it contains exactly one entry for every residue in strictly increasing index order.