BINDRAE：Stage - wise Structural Prior + Conditional Bridge Flow for Ligand - Induced Apo→Holo Pathways

摘要（方法摘要式）

我们研究在已知蛋白序列、apo 构象与配体结合姿态的条件下，生成从 apo 到 holo 的**连续构象通路**并预测最终结合构象。我们提出一个两阶段框架：**Stage - 1** 学习一个高保真、几何一致的 **holo 先验解码器**，以冻结的蛋白语言模型（ESM）为语义底座，并在 SE(3) 等变结构模块中融合配体条件，输出 per - residue backbone 刚体帧与侧链 χ 扭转角，通过前向运动学（FK）重建全原子结构；**Stage - 2** 在乘积流形 $\mathcal{M} = \mathrm{SE}(3)^N \times (S^1)^{4N}$ 上学习一个**配体条件化、口袋—变构联合门控的**桥流向量场，使用带噪参考桥与 Conditional Flow Matching 训练，并将 Stage - 1 的几何先验（FK/FAPE/clash/contact）提升为**路径级约束**。为覆盖少数但关键的“大尺度变构”，Stage - 2 进一步引入基于弹性网络模型（ENM）的 **NMA 残基层移动性特征**，以物理可解释方式指导门控与损失权重，实现对铰链/结构域运动的自适应激活。

---

1. 问题定义与符号体系

1.1 输入与目标

给定：

- 蛋白序列 $S = (a_1, \dots, a_N)$，长度 N；
- apo 结构（观测）$\mathbf{X}^{\mathrm{apo}}$（至少包含 backbone N–Cα–C 坐标，可扩展至 atom14）；
- 配体 L 的重原子集合 $(\ell_m, \mathbf{y}_m)_{m=1}^{M}$，其中 ℓ_m 为类型/化学特征，$\mathbf{y}_m \in \mathbb{R}^3$ 为坐标；
- 关键约定：**配体坐标以 apo 坐标系表示**（或将 holo 中的配体刚体变换对齐到 apo）。

目标：

1. 预测 holo 终点结构 $\hat{\mathbf{X}}^{\mathrm{holo}}$；
2. 生成一条连续路径 $\hat{\mathbf{X}}(t)_{t\in[0,1]}$，满足 $\hat{\mathbf{X}}(0) \approx \mathbf{X}^{\mathrm{apo}}$，$\hat{\mathbf{X}}(1) \approx \hat{\mathbf{X}}^{\mathrm{holo}}$，且路径中间帧具有生物物理合理性（无严重 clash、口袋接触呈合理单调趋势、整体运动平滑）。

1.2 结构状态：骨架刚体帧与侧链 χ 扭转角

1.2.1 每残基 backbone 刚体帧（$\mathrm{SE}(3)$）

对每个残基 i，从其 backbone 三点 $(\mathbf{x}_i^N, \mathbf{x}_i^{C\alpha}, \mathbf{x}_i^C)$ 构造刚体帧

$$F_i = (R_i, t_i) \in \mathrm{SE}(3), \qquad R_i \in \mathrm{SO}(3),\ t_i \in \mathbb{R}^3.$$

一种标准构造是（给出可验证的确定性定义）：

$$\mathbf{e}_1 = \frac{\mathbf{x}_i^C - \mathbf{x}_i^{C\alpha}}{\|\mathbf{x}_i^C - \mathbf{x}_i^{C\alpha}\|},\ \mathbf{u} \quad = \mathbf{x}_i^N - \mathbf{x}_i^{C\alpha},\ \mathbf{e}_2 = \frac{\mathbf{u} - (\mathbf{u}^\top \mathbf{e}_1)\mathbf{e}_1}{\|\mathbf{u} - (\mathbf{u}^\top \mathbf{e}_1)\mathbf{e}_1\|},\ \mathbf{e}_3 \quad = \mathbf{e}_1 \times \mathbf{e}_2,\ R_i = [\mathbf{e}_1\ \mathbf{e}_2\ \mathbf{e}_3], \qquad t_i = \mathbf{x}_i^{C\alpha}.$$

则任意局部坐标 $\mathbf{p}$ 在全局坐标为 $R_i\mathbf{p} + t_i$。

全蛋白骨架帧 $F = (F_1, \ldots, F_N) \in \mathrm{SE}(3)^N$。

1.2.2 侧链 χ 扭转角（$(S^1)^{4N}$）

对每个残基 i，定义最多四个侧链扭转角

$$\chi_i = (\chi_{i,1}, \chi_{i,2}, \chi_{i,3}, \chi_{i,4}) \in (S^1)^4,$$

并用掩码 $m_{i,k} \in 0, 1$ 表示第 k 个 χ 是否存在（例如 Gly/Ala 全 0）。

我们将角变量视为流形 S^1，并使用 wrap 运算

$$\mathrm{wrap}(\theta) \in (-\pi, \pi]$$

确保差分与更新的周期一致性。网络内部对 χ 采用 $(\sin\chi, \cos\chi)$ 作为数值稳定表示，但数学对象仍是 S^1。

1.2.3 全状态空间

最终状态

$$x = (F, \chi) \in \mathcal{M}, \qquad \mathcal{M} = \mathrm{SE}(3)^N \times (S^1)^{4N}.$$

（这是无冗余的严格主线：backbone 自由度由 F 表示；侧链自由度由 χ 表示。将 backbone torsions 再显式纳入会与 F 产生不可辨识冗余，从而破坏动力学建模的良定性；见 § 4.2 的“良定性论证”。）

1.3 全原子解码：前向运动学 FK

令 FK 为确定性前向运动学映射：

$$\mathbf{X} = \mathrm{FK}(F, \chi),$$

输出 atom14 坐标 $\mathbf{X} = \{\mathbf{x}_{i,a}\}$（残基 i 的第 a 个原子）。FK 由标准化内坐标模板（键长/键角/扭转轴）定义，沿键旋转依次施加 χ，并用 F_i 将局部骨架嵌入全局。FK 的关键性质是：对全局刚体变换 $G \in \mathrm{SE}(3)$，有

$$\mathrm{FK}(G \cdot F, \chi) = G \cdot \mathrm{FK}(F, \chi),$$

即 FK 对全局 $\mathrm{SE}(3)$ 作用保持等变。

---

2. Stage - 1：静态 holo 先验解码器（Structural Prior Decoder）

Stage - 1 的定位不是“生成路径”，而是学习一个强先验：在给定 apo + ligand 的条件下，输出一个高可信的 holo 构象估计（或其参数化表示）。Stage - 2 的所有几何约束、端点引导与可解释路径都将以 Stage - 1 为锚。

2.1 Stage - 1 的条件化目标

给定输入条件 $c = (S, \mathbf{X}^{\text{apo}}, L)$，Stage - 1 学习映射

$$\mathcal{F}_\psi : c \mapsto (\tilde{F}^{\text{holo}}, \tilde{\chi}^{\text{holo}}),$$

并解码

$$\tilde{\mathbf{X}}^{\text{holo}} = \text{FK}(\tilde{F}^{\text{holo}}, \tilde{\chi}^{\text{holo}}).$$

其中 ψ 为 Stage - 1 参数。Stage - 1 输出将作为 Stage - 2 的**先验终态锚**与条件输入（见 § 5.1、§ 6.4）。

2.2 表征与条件输入（冻结语义 + 几何/配体注入）

2.2.1 冻结语言模型表征（ESM）

使用冻结蛋白语言模型 $E(\cdot)$：

$$e_i = E(S)_i \in \mathbb{R}^{d_{\text{esm}}}.$$

冻结意味着 Stage - 1/2 训练仅学习一个轻量 Adapter：

$$s_i = A(e_i) \in \mathbb{R}^d,$$

将语义表征投影到结构模块维度 d。

2.2.2 几何边特征与 SE(3) 等变约束

从 apo 的 backbone 帧 F^{apo} 构造残基图边特征。对任意 i, j，定义相对变换

$$\Delta F_{ij} = (F_i^{\text{apo}})^{-1} F_j^{\text{apo}} \in \text{SE}(3),$$

并从其 Lie 代数参数（例如 $\log(\Delta F_{ij}) \in \mathbb{R}^6$）与距离/邻域编码得到边特征 b_{ij}。所有几何特征均由相对量构造，从而对全局 SE(3) 平移/旋转天然不变。

2.2.3 配体 token 与蛋白 - 配体条件化

对配体每个原子 m，构造 token 嵌入

$$\ell_m = \text{Embed}(\text{atom type}, \text{chem features}) \in \mathbb{R}^{d_L},$$

并保留其位置 $\mathbf{y}_m \in \mathbb{R}^3$。通过蛋白 - 配体跨注意力/消息传递，得到对每个残基的配体条件化特征 $l_i \in \mathbb{R}^d$，并与 s_i 融合（例如 concat + MLP 或 FiLM）。

2.3 SE(3) 等变结构主干（FlashIPA/IPA - type）

Stage - 1 使用 SE(3) 等变的主干 Trunk_ψ 将 $(s_i, b_{ij}, l_i, F^{\mathrm{apo}})$ 映射到残基隐状态 h_i 与预测帧 $\tilde{F}_i$。抽象表示为：

$$(h, \tilde{F}) = \mathrm{Trunk}_\psi(s, b, l, F^{\mathrm{apo}}).$$

等变性要求（严格）：对任意全局变换 $G \in \mathrm{SE}(3)$（作用于所有坐标与帧），有

$$\mathrm{Trunk}_\psi(\cdot,\ G \cdot F^{\mathrm{apo}}) = (h,\ G \cdot \tilde{F}),$$

即隐状态 h 不随全局刚体变换改变（不变），预测帧 $\tilde{F}$ 随之等变。

2.4 输出头与解码

2.4.1 Frame head

预测帧用残差式更新（保证可控且数值稳定）：

$$\tilde{F}_i = F_i^{\mathrm{apo}} \exp(\widehat{\Delta\xi_i}), \quad \Delta\xi_i \in \mathbb{R}^6,$$

其中 $\widehat{\cdot}$ 是 $\mathfrak{se}(3)$ 的帽算子：

$$\widehat{\Delta\xi} = \begin{pmatrix} [\omega]_\times & v \\ 0 & 0 \end{pmatrix}.$$

2.4.2 χ head

预测侧链 χ：

$$\tilde{\chi}_{i,k} \in S^1, \quad k = 1, \ldots, 4,$$

网络输出 $(\sin\tilde{\chi}_{i,k}, \cos\tilde{\chi}_{i,k})$ 并归一化到单位圆，保证数学上一致。

2.5 Stage - 1 损失函数（几何一致、SE(3) 不变）

Stage - 1 在训练集 $\{(\mathbf{X}^{\mathrm{apo}}, \mathbf{X}^{\mathrm{holo}}, L, S)\}$ 上最小化以下损失：

2.5.1 FAPE（Frame - Aligned Point Error）

令预测结构 $\tilde{\mathbf{X}} = \mathrm{FK}(\tilde{F}, \tilde{\chi})$，真值 holo 为 $\mathbf{X}^{\mathrm{holo}}$。对每个残基 i 的局部帧 F_i^*（由真值 holo 构造）与预测帧 $\tilde{F}_i$，定义对点 $\mathbf{x}_{j,a}$ 的局部坐标：

$$\pi(F, \mathbf{x}) = R^\top(\mathbf{x} - t).$$

则 FAPE：

$$\mathcal{L}_{\mathrm{FAPE}} = \sum_i \sum_{j,a} w_{j,a} \left\| \pi(\tilde{F}_i, \tilde{\mathbf{x}}_{j,a}) - \pi(F_i^*, \mathbf{x}_{j,a}^{\mathrm{holo}}) \right\|_2.$$

性质（不变性）：对任意全局 $G \in \mathrm{SE}(3)$，同时变换预测与真值坐标，$\mathcal{L}_{\mathrm{FAPE}}$ 不变。证明：$\pi(GF, G\mathbf{x}) = (R_g R)^\top (G\mathbf{x} - (R_g t + t_g)) = R^\top R_g^\top (R_g \mathbf{x} + t_g - R_g t - t_g) = R^\top(\mathbf{x} - t) = \pi(F, \mathbf{x})$。因此各项差不变。

2.5.2 χ 的 wrap - aware 角度损失

$$\mathcal{L}_\chi = \sum_{i,k} m_{i,k}\, d_{S^1}(\tilde{\chi}_{i,k}, \chi^{\mathrm{holo}}_{i,k})^2, \quad d_{S^1}(a,b) = \mathrm{wrap}(a-b).$$

2.5.3 Clash（全原子碰撞）约束

对任意原子对 (p, q)，距离 d_{pq}，范德华半径和阈值 $r_p + r_q$。定义软碰撞罚项：

$$\mathcal{L}_{\mathrm{clash}} = \sum_{p<q} \mathbf{1}[\text{not covalent}]\ \mathrm{ReLU}(r_p + r_q - d_{pq})^2.$$

（用 atom14 邻接表排除共价键与 1–3/1–4 近邻。）

2.5.4 口袋加权（聚焦与配体相关自由度）

定义残基 - 配体最小距离

$$d_i = \min_{a \in \mathrm{atoms}(i), m \in L} \|\mathbf{x}^{\mathrm{apo}}_{i,a} - \mathbf{y}_m\|.$$

定义口袋权重

$$w_i^{\mathrm{pocket}} = \sigma\left(\frac{d_0 - d_i}{\tau}\right) \in (0,1),$$

其中 d_0 为接触阈值（例如 6Å），τ 为平滑尺度。Stage - 1 的残基相关损失采用加权：

$$\mathcal{L}_{\mathrm{Stage1}} = \mathcal{L}_{\mathrm{FAPE}} + \lambda_\chi \sum_i w_i^{\mathrm{pocket}}\, \mathcal{L}_{\chi,i} + \lambda_{\mathrm{clash}} \mathcal{L}_{\mathrm{clash}}.$$

2.6 Stage - 1 输出作为 Stage - 2 的强先验

Stage - 1 的输出定义为

$$(\tilde{F}^{\mathrm{holo}}, \tilde{\chi}^{\mathrm{holo}}) = \mathcal{F}_\psi(S, \mathbf{X}^{\mathrm{apo}}, L).$$

在 Stage - 2 中，它承担两种严格角色：

1. **条件输入（guidance condition）**：作为向量场输入的一部分，提供“目标构象锚”；
2. **路径后半段先验约束**：对 t 接近 1 的状态施加软约束（见 § 6.4），避免桥流在终段偏离可行 holo 流形。

---

3. Stage - 2：配体条件化 apo→holo 路径桥流（Conditional Bridge Flow on $\mathcal{M}$）

Stage - 2 在状态流形 $\mathcal{M} = \mathrm{SE}(3)^N \times (S^1)^{4N}$ 上学习时间连续向量场，使从 apo 起始状态出发积分得到一条路径，同时受路径级几何约束控制。

3.1 端点与路径的形式化

对训练样本，我们可提取：

- apo 状态 $x_0 = (F_0, \chi_0)$；
- holo 真值状态 $x_1 = (F_1, \chi_1)$；
- Stage - 1 先验终点 $\tilde{x}_1 = (\tilde{F}_1, \tilde{\chi}_1)$。

Stage - 2 的目标是学习向量场

$$v_\Theta(x, t \mid c) = (\dot{F}_\Theta(x, t), \dot{\chi}_\Theta(x, t)),$$

其中条件

$$c = \{S,\ L,\ \mathbf{X}^{\mathrm{apo}},\ \tilde{x}_1,\ w_i,\ M_i^{\mathrm{nma}}\}.$$

3.2 $\mathrm{SE}(3)$ 对称性与良定性（必须严格满足）

3.2.1 全局 $\mathrm{SE}(3)$ 作用

对任意 $G \in \mathrm{SE}(3)$，定义其对状态的作用

$$G \cdot (F, \chi) = (GF, \chi),$$

即仅作用于帧（左乘），χ 不变。

3.2.2 向量场的等变性要求

Stage - 2 要求对任意 $G \in \mathrm{SE}(3)$：

$$v_\Theta(G \cdot x, t \mid G \cdot c) = \mathcal{T}_G\, v_\Theta(x, t \mid c),$$

其中 $\mathcal{T}_G$ 是切空间中的自然变换。

我们采用 **右平凡化（body - frame）**表述，使该等变性变为一个更强、可实现的形式：向量场输出的 twist 用**每残基自身局部（body）坐标系**表示，从而对全局 G **不变**（见下）。

3.2.3 右平凡化 ODE 与等变性证明

在每个残基 i 上，Stage - 2 定义 ODE：

$$\dot{F}_i = F_i\,\widehat{\xi_i}, \qquad \xi_i = (\omega_i, v_i) \in \mathbb{R}^6,$$

即 ξ_i 为 **body twist**。离散更新：

$$F_i^{n+1} = F_i^n \exp(\Delta t,\, \widehat{\xi_i^n}).$$

命题（全局等变性）：若网络输出满足

$$\xi_i(G \cdot x, t \mid G \cdot c) = \xi_i(x, t \mid c),$$

则对任意初始状态 x_0，ODE 的解满足

$$F_i(t; G \cdot x_0) = G \cdot F_i(t; x_0),$$

且 χ 路径不变（χ 是不变量）。证明：令 $F_i'(t) = GF_i(t)$。则 $\dot{F}_i' = G\dot{F}_i = GF_i\widehat{\xi_i} = F_i'\widehat{\xi_i}$，且初值 $F_i'(0) = GF_i(0)$。由常微分方程解唯一性得 $F_i'(t)$ 即为以 $G \cdot x_0$ 为初值的解，故等变性成立。

> 因此：**只要网络的几何特征全用相对量（在局部坐标表达）构造，输出 body twist 就天然满足上述不变性，从而整体路径严格 $\mathrm{SE}(3)$ 等变。**

3.2.4 良定性：为何不把 backbone torsions 纳入主状态

backbone torsions（φ/ψ/ω）与相邻残基帧的相对旋转存在函数关系；在以每残基帧 F 作为 backbone 自由度的情况下，再显式引入 φ/ψ/ω 会造成：

- **冗余自由度**：同一几何构象可由不同 (F, ϕ, ψ, ω) 表示；
- **动力学不可辨识**：向量场可在冗余维度上"走捷径"，导致训练与积分不稳定；
- **约束冲突**：FK/几何损失对两套变量同时约束会产生不可控的梯度耦合。

因此主线严格采用 (F, χ) 作为状态，backbone torsions 仅作为可从 F 推导的派生量（若需要评估），不进入动力学状态。

---

4. 参考桥（Reference Bridge）在 $\mathcal{M}$ 上的严格构造

Stage - 2 训练使用带噪参考桥 X_t^{ref} 与其解析速度 u_t^{ref}，并通过 Conditional Flow Matching 学习真实向量场。

4.1 χ（圆周空间）上的 Brownian - bridge 参考路径

给定端点 $\chi_0, \chi_1 \in (S^1)^{4N}$：

1. 最短角差
$$\Delta\chi = \mathrm{wrap}(\chi_1 - \chi_0) \in (-\pi, \pi]^{4N}.$$
2. 选择平滑插值 $\gamma : [0,1] \to [0,1]$ 与噪声尺度 $\sigma : [0,1] \to \mathbb{R}_+$：
$$\gamma(t) = 3t^2 - 2t^3, \qquad \sigma(t) = \lambda_\chi \sqrt{t(1-t)}.$$
（$\sigma(0) = \sigma(1) = 0$ 保证端点一致性。）
3. 参考桥采样：
$$\chi_t^{\mathrm{ref}} = \mathrm{wrap}\big(\chi_0 + \gamma(t)\Delta\chi + \sigma(t), \varepsilon\big), \quad \varepsilon \sim \mathcal{N}(0, I).$$
4. 解析参考速度（在切空间的角速度）：
$$u_\chi^{\mathrm{ref}}(t) = \frac{d}{dt}(\chi_0 + \gamma(t)\Delta\chi) = \gamma'(t)\Delta\chi.$$
噪声项不影响 $\dot{\mu}(t)$ 的定义；CFM 的最优解为条件期望（见 § 5.3）。

4.2 $\mathrm{SE}(3)$ 上的 geodesic + right - noise 参考桥

对每个残基 i，给定端点 $F_{0,i}, F_{1,i} \in \mathrm{SE}(3)$。

1. 相对变换
$$\Delta_i = F_{0,i}^{-1} F_{1,i} \in \mathrm{SE}(3).$$
2. 参考均值路径（geodesic interpolation）
$$\mu_i(t) = F_{0,i} \exp\big(\gamma(t) \log(\Delta_i)\big).$$
3. right - multiplicative 噪声（Lie 代数高斯）
$$F_{i,t}^{\mathrm{ref}} = \mu_i(t) \exp(\sigma_F(t), \widehat{\eta_i}), \quad \eta_i \sim \mathcal{N}(0, I_6), \quad \sigma_F(t) = \lambda_F \sqrt{t(1-t)}.$$
（同样 $\sigma_F(0) = \sigma_F(1) = 0$。）
4. 参考速度（body twist）
$$u_{F,i}^{\mathrm{ref}}(t) = \gamma'(t) \log(\Delta_i) \in \mathbb{R}^6.$$

证明（速度表达）：$\mu_i(t) = F_{0,i} \exp(A(t))$ 且 $A(t) = \gamma(t) \log(\Delta_i)$。利用 $\frac{d}{dt} \exp(A(t)) = \exp(A(t)) \dot{A}(t)$（当 $[A, \dot{A}] = 0$，此处成立因为 $A(t)$ 与 $\dot{A}(t)$ 共线），得
$$\dot{\mu}_i(t) = F_{0,i} \exp(A(t)) \dot{A}(t) = \mu_i(t), \gamma'(t) \widehat{\log(\Delta_i)}.$$
对 right - trivialization，body twist 即 $\gamma'(t) \log(\Delta_i)$。

4.3 乘积参考桥

整体参考桥为

$$X_t^{\text{ref}} = (F_t^{\text{ref}}, \chi_t^{\text{ref}}), \qquad u_t^{\text{ref}} = (u_F^{\text{ref}}, u_\chi^{\text{ref}}).$$

端点一致性由 $\sigma(0) = \sigma(1) = 0$、$\gamma(0) = 0, \gamma(1) = 1$ 保证。

---

5. 向量场学习：Conditional Flow Matching（CFM）在乘积流形上的严格形式

5.1 向量场参数化（含 Stage - 1 先验条件）

Stage - 2 学习

$$v_\Theta(x, t \mid c) = (\xi_\Theta(x, t \mid c),\ \dot{\chi}_\Theta(x, t \mid c)),$$

其中 $\xi_\Theta \in \mathbb{R}^{N\times 6}$ 为每残基 body twist，$\dot{\chi}_\Theta \in \mathbb{R}^{N\times 4}$ 为 χ 角速度。

关键：Stage - 2 的条件必须包含 Stage - 1 预测终点 $\tilde{x}_1$，否则推理阶段（无真值 holo）无法提供终态锚。我们将 $\tilde{x}_1$ 以"相对量"形式注入：

- 相对帧误差：$\delta F_i = F_i^{-1}\tilde{F}_{1,i}$，编码为 $\log(\delta F_i) \in \mathbb{R}^6$；
- 相对 χ 误差：$\delta\chi_i = \text{wrap}(\tilde{\chi}_{1,i} - \chi_i) \in \mathbb{R}^4$。

它们与 $h_i(t)$、t、口袋/弹性权重一起作为门控与 head 的输入。

5.2 口袋—弹性联合权重（w_i^{eff}）与门控（$g_i(t)$）

Stage - 2 必须同时覆盖：

- 口袋局部诱导契合（多数样本）；
- 铰链/结构域大尺度变构（少数样本，但科学价值极高）。

因此我们定义两个来源的"应动性"信号：

1. **口袋权重** $w_i^{\text{pocket}} \in (0, 1)$，由 apo+ligand 计算（与 Stage - 1 一致）；
2. **NMA 移动性** $M_i^{\text{nma}} \in [0, 1]$，由 apo ENM - NMA 预计算（见 §7）。

将二者融合为 **有效权重**

$$w_i^{\text{eff}} = \max\left(w_i^{\text{pocket}},\ M_i^{\text{nma}}\right).$$

（该融合规则等价于"口袋应动 OR 物理易动"；是确定性且无兼容分支的严格定义。）

定义门控

$$g_i(t) = \sigma\left(\text{MLP}_g\left([h_i(t);\ e_t;\ w_i^{\text{pocket}};\ M_i^{\text{nma}};\ \log\delta F_i;\ \delta\chi_i]\right)\right) \in (0, 1).$$

并对输出速度强制门控：

$$\xi_i = g_i(t)\,\xi_i^{\text{raw}} \qquad \dot{\chi}_i = g_i(t)\,\dot{\chi}_i^{\text{raw}}.$$

这一步在数学上等价于将向量场限制在一个由网络学习的、随时间变化的“可动子空间”中；与纯 loss reweighting 不同，它直接改变动力学。

5.3 CFM 损失（带有效权重与掩码）

对参考桥样本 $(X_t^{\text{ref}}, u_t^{\text{ref}})$，定义 CFM 目标：

$$\mathcal{L}_{\text{CFM}} = \mathbb{E}_{t,X_0,X_1,\varepsilon}\left[\sum_i (w_i^{\text{eff}})^\alpha \left(\|\xi_{\Theta,i}(X_t^{\text{ref}},t) - u_{F,i}^{\text{ref}}(t)\|_2^2 + \sum_{k=1}^{4} m_{i,k}\,|\dot{\chi}_{\Theta,i,k}(X_t^{\text{ref}},t) - u_{\chi,i,k}^{\text{ref}}(t)|^2\right)\right].$$

其中 $\alpha > 0$ 固定（严格超参），用于增强“应动区域”的匹配强度。

定理（CFM 的最优性：条件期望投影）

令随机变量 $U = u_t^{\text{ref}}$，$Z = X_t^{\text{ref}}$。在平方损失下，最优向量场满足

$$v^*(z,t) = \mathbb{E}[U \mid Z = z, t].$$

证明：对每个 t 固定，$\arg\min_f \mathbb{E}[|f(Z) - U|^2]$ 的解为条件期望，这是 L^2 空间中的正交投影定理。

推论（边缘分布匹配的运输性质）

在满足常规正则条件（Lipschitz 连续、解存在唯一等）下，由 ODE

$$\dot{X} = v^*(X,t)$$

生成的连续流，其时间边缘分布与参考路径分布的边缘一致，从而 p_0 与 p_1 被正确连接。该性质是 CFM 在生成建模中作为“确定性桥”的理论基础。

---

6. 路径级几何与生物物理正则（严格主线）

CFM 仅对齐速度场的统计意义，不能保证中间帧物理可行。因此 Stage - 2 必须在学习到的路径上施加约束：

令通过数值积分得到路径 $x_\Theta(t)$（见 § 8 的算法），选取固定时间网格

$$0 = t_0 < t_1 < \cdots < t_T = 1.$$

在每个 t_j 解码得到 $\mathbf{X}(t_j) = \text{FK}(F(t_j), \chi(t_j))$。

6.1 状态级平滑（$\mathrm{SE}(3)+\chi$）

定义相邻时间步状态差：

$$\Delta F_i^{(j)} = F_i(t_j)^{-1} F_i(t_{j+1}), \qquad \Delta\chi_i^{(j)} = \mathrm{wrap}\big(\chi_i(t_{j+1}) - \chi_i(t_j)\big).$$

状态平滑损失：

$$\mathcal{L}_{\text{smooth}} = \sum_{j=0}^{T-1} \sum_i (w_i^{\text{eff}})^{\alpha} \left(\|\log(\Delta F_i^{(j)})\|_2^2 \sum_k m_{i,k} |\Delta\chi_{i,k}^{(j)}|^2 \right).$$

它在流形的切空间上定义，严格 $\mathrm{SE}(3)$ 一致。

6.2 原子级 clash 正则

对每个时间步的 atom14 坐标 $\mathrm{X}(t_j)$，计算 clash：

$$\mathcal{L}_{\text{clash}} = \sum_{j=0}^{T} \sum_{p<q} 1[\text{not covalent}]\ \mathrm{ReLU}(r_p + r_q - d_{pq}(t_j))^2.$$

6.3 口袋接触软单调性（方向性约束，路径级）

定义残基 i 与配体的软接触强度：

$$c_i(t) = \sum_{a \in \text{atoms}(i)} \sum_{m=1}^{M} \sigma\left(\frac{d_c - \|\mathbf{x}_{i,a}(t) - \mathbf{y}_m\|}{\tau_c} \right).$$

定义全局加权接触：

$$C(t) = \sum_i w_i^{\text{eff}}\, c_i(t).$$

要求结合过程接触总体不应显著回退（避免“先塞进再抽出”）。定义单调性损失：

$$\mathcal{L}_{\text{contact}} = \sum_{j=0}^{T-1} \mathrm{ReLU}\big(C(t_j) - C(t_{j+1}) + \delta\big)^2.$$

其中 $\delta \geq 0$ 为允许的微小回退容忍度（严格固定超参）。当 $\delta = 0$ 且步长趋于 0，该项约束 $C(t)$ 为弱单调非减函数。

6.4 Stage - 1 holo prior 对齐（路径后半段硬约束）

Stage - 1 给出 $\tilde{x}_1 = (\tilde{F}_1, \tilde{\chi}_1)$。Stage - 2 在后半段强制靠近该先验流形：

设 $t_{\text{mid}} \in (0,1)$ 固定（例如 0.5）。定义

$$\mathcal{L}_{\text{prior}} = \sum_{j:t_j \geq t_{\text{mid}}} \sum_i (w_i^{\text{eff}})^\alpha \left(\| \log(F_i(t_j)^{-1} \tilde{F}_{1,i}) \|_2^2 + \sum_k m_{i,k} d_{S^1}(\chi_{i,k}(t_j), \tilde{\chi}_{1,i,k})^2 \right).$$

这保证路径终段不会偏离 Stage - 1 已学习到的高可行区域，同时允许前半段自由探索过渡构象。

6.5 背景稳定性（非应动区域速度抑制，防止全局漂移）

定义非应动区域惩罚指数 $\beta > 0$ 固定。对任意时间 t，定义

$$\mathcal{L}_{\text{bg}}(t) = \sum_i (1 - w_i^{\text{eff}})^\beta \left(\|\xi_i(x,t)\|_2^2 + \sum_k m_{i,k} |\dot{\chi}_{i,k}(x,t)|^2 \right).$$

训练中对 $t \sim \mathcal{U}(0,1)$ 求期望得到 $\mathcal{L}_{\text{bg}}$。这项与门控 $g_i(t)$ 共同保证：非口袋且非铰链区域不会出现不受控漂移。

---

7. NMA - Guided Elastic Gating（ENM - NMA 的严格定义与不变性）

Stage - 2 的大尺度变构能力来自 M_i^{nma}。该特征在训练与推理中都由 apo 结构预计算，且作为必需输入。

7.1 弹性网络模型（ENM）

以 apo 的 Cα 节点 $\mathbf{r}_i \in \mathbb{R}^3$ 构建图：若 $\|\mathbf{r}_i - \mathbf{r}_j\| < r_c$ 则连接弹簧，势能

$$V = \frac{k}{2} \sum_{i<j} A_{ij} \left(\|\mathbf{r}_i - \mathbf{r}_j\| - d_{ij}^0 \right)^2,$$

其中 $A_{ij} \in 0,1$ 为邻接，$d_{ij}^0 = \|\mathbf{r}_i - \mathbf{r}_j\|$ 为平衡距离。

Hessian（3N×3N）：

$$H_{\alpha\beta}^{ij} = \frac{\partial^2 V}{\partial r_i^\alpha \partial r_j^\beta}.$$

对 H 做特征分解：

$$Hu_k = \lambda_k u_k.$$

去除 6 个刚体零模态（3 平移 + 3 旋转），取最低频的 K_{mode} 个非零模态。

7.2 残基层移动性（mobility amplitude）

令 $u_k^{(i)} \in \mathbb{R}^3$ 为模态 k 在残基 i 的 3 维位移向量。定义

$$M_i^{\text{nma}} = \sum_{k=1}^{K_{\text{mode}}} \omega_k \, \|u_k^{(i)}\|_2, \qquad \omega_k = \frac{1}{\lambda_k + \epsilon}.$$

并将 M_i^{nma} 线性归一化到 $[0, 1]$。

7.3 不变性验证（对全局坐标系无关）

命题：若对所有节点施加全局刚体变换 $\mathbf{r}_i \mapsto R\mathbf{r}_i + t$，则 M_i^{nma} 不变。证明要点：ENM 势能仅依赖成对距离 $\|\mathbf{r}_i - \mathbf{r}_j\|$，对刚体变换不变；因此 Hessian 的非零特征值谱不变，模态向量在旋转下整体左乘 R，从而 $\|u_k^{(i)}\|$ 不变，故 M_i^{nma} 不变。

---

8. 总损失、训练与推理算法（无兼容分支）

8.1 总损失

Stage - 2 的总目标为

$$\mathcal{L} = \mathcal{L}_{\text{CFM}} + \lambda_{\text{smooth}}\mathcal{L}_{\text{smooth}} + \lambda_{\text{clash}}\mathcal{L}_{\text{clash}} + \lambda_{\text{contact}}\mathcal{L}_{\text{contact}} + \lambda_{\text{prior}}\mathcal{L}_{\text{prior}} + \lambda_{\text{bg}}\mathcal{L}_{\text{bg}} + \lambda_{\text{end}}\mathcal{L}_{\text{end}},$$

其中 $\mathcal{L}_{\text{end}}$ 为训练时对真值 holo 的端点一致性（下述）。

端点一致性（训练时）

对从 x_0 积分得到的终点 $x_\Theta(1)$，定义

$$\mathcal{L}_{\text{end}} = \sum_i \| \log(F_{\Theta,i}(1)^{-1} F_{1,i})\|_2^2 + \sum_{i,k} m_{i,k} d_{S^1}(\chi_{\Theta,i,k}(1), \chi_{1,i,k})^2 + \lambda_{\text{FAPE}}\mathcal{L}_{\text{FAPE}}(\text{FK}(x_\Theta(1)), \mathbf{X}^{\text{holo}}).$$

该项保证“从 apo 出发的积分轨迹”确实落到真值 holo 端点附近，避免仅对齐参考桥速度却不达端点的退化。

8.2 数值积分（严格实现口径）

使用 Heun（二阶显式 Runge–Kutta）对 ODE 积分（固定步数 T）：

SE(3) 更新（右平凡化 right - multiply）：

$$F^{n+1} = F^n \exp\left(\Delta t, \widehat{\xi^n}\right), \quad \Delta t = \frac{1}{T}.$$

χ 更新（角度 wrap）：

$$\chi^{n+1} = \mathrm{wrap}(\chi^n + \Delta t, \dot{\chi}^n).$$

Heun：

$$k_1 = v_\Theta(x^n, t_n),\ \tilde{x} \ = \Phi(x^n, \Delta t, k_1),\ k_2 = v_\Theta(\tilde{x}, t_{n+1}),\ x^{n+1} \ = \Phi\left(x^n, \Delta t, \frac{k_1 + k_2}{2}\right),$$

其中 Φ 表示按上述 SE(3)+χ 规则做一步更新。

8.3 训练算法（Algorithm 1）

对每个 batch：

1. 从 apo/holo 构造 x_0, x_1；从 apo 构造 pocket 权重 w^{pocket}；预加载 M^{nma}；
2. 运行 Stage - 1 得到 $\tilde{x}_1$；
3. 采样 $t \sim \mathcal{U}(0,1)$，采样噪声，构造 $X_t^{\text{ref}}, u_t^{\text{ref}}$；
4. 计算 CFM：$v_\Theta(X_t^{\text{ref}}, t)$ 与 u_t^{ref} 的加权平方差；
5. 从 x_0 积分得到完整路径 $x_\Theta(t_j)_{j=0}^{T}$，并解码 $\mathrm{X}(t_j) = \mathrm{FK}(x_\Theta(t_j))$；
6. 计算路径级 $\mathcal{L}_{\text{smooth}}, \mathcal{L}_{\text{clash}}, \mathcal{L}_{\text{contact}}, \mathcal{L}_{\text{prior}}$；
7. 计算 $\mathcal{L}_{\text{bg}}$（可直接在积分过程中累加）；
8. 计算端点一致性 $\mathcal{L}_{\text{end}}$；
9. 合成总损失 $\mathcal{L}$ 并反向传播更新 Θ。

8.4 推理算法（Algorithm 2：apo+ligand → 路径 + holo）

给定 apo 与 ligand：

1. 构造 x_0；
2. 计算 w^{pocket}（apo+ligand），读取 M^{nma}；
3. 运行 Stage - 1 得到 $\tilde{x}_1$；
4. 以条件 c（包含 $\tilde{x}_1$）运行 Stage - 2 ODE 积分得到 $x_\Theta(t)$；
5. 对每个 t 解码 $\mathrm{X}(t) = \mathrm{FK}(x_\Theta(t))$ 得到全原子路径与终点 holo。

---

9. 关键数学自检（作为审稿人会看的"正确性证明块"）

这里把最容易被顶会/顶刊追问的"数学正确性"集中给出，确保整套方法在形式上无漏洞。

9.1 FK 等变性（已在 §1.3 给出）

$\mathrm{FK}(G \cdot F, \chi) = G \cdot \mathrm{FK}(F, \chi)$。

9.2 FAPE 全局不变性（§ 2.5.1 已证明）

确保 Stage - 1 训练不依赖坐标系选择。

9.3 Stage - 2 ODE 的 $\mathrm{SE}(3)$ 等变性（§ 3.2.3 已证明）

右平凡化 body - twist + 右乘更新，在“网络输出 body 坐标 twist 不变”的条件下严格等变。

9.4 参考桥端点一致性

因为 $\gamma(0)=0,\gamma(1)=1$ 且 $\sigma(0)=\sigma(1)=0$，直接得：

$$X_0^{\mathrm{ref}}=x_0,\qquad X_1^{\mathrm{ref}}=x_1.$$

9.5 CFM 的最优性与运输性质（§ 5.3）

平方损失最优解为条件期望，构成从参考路径边缘到目标边缘的确定性桥。

9.6 NMA 特征不变性（§ 7.3）

ENM 基于距离；模态幅值对刚体变换不变，保证特征可跨坐标系复用。

---

10. 实验与可复现要点

1. **数据构建必须可复现**：apo/holo 配对规则、链选择、缺失残基过滤、对齐方法、配体一致性/相似性阈值、去冗余 split（按蛋白序列与配体 scaffold）。
2. **严格的路径指标**：clash%、contact 单调违例率、状态平滑项、终点 pocket iRMSD/χ1 命中率。
3. **大变构分桶**：按 apo→holo backbone RMSD 或域运动指标分桶，报告 Stage - 2（含 NMA 门控）在“大变构桶”的收益。
4. **可解释性展示**：至少 2–3 个 case，展示 $C(t)$ 曲线、关键 χ 翻转时刻、铰链区域的 w^{eff} 与门控轨迹 $g_i(t)$。