Key Mathematical Formulations for 3D Ligand–Pharmacophore Mapping (DiffPhore)

Problem Definition

$$\hat{G}_l = \text{Model}(G_p,\, G_l)\,, \tag{1}$$

where G_p denotes the pharmacophore model (graphical 3D representation), G_l is the input ligand conformation, and $\hat{G}_l$ is the generated ligand conformation (same chemistry as G_l, different 3D pose).

The learning goal is to approximate the conditional density

$$P\Big(\hat{G}_l \mid G_p,\, G_l\Big)\,, \tag{2}$$

whose score (gradient of log-density) guides denoising/generation.

Score-based Generative Modeling

Langevin dynamics (generic). Given a stepsize $\epsilon > 0$ and Gaussian noise $z_t \sim \mathcal{N}(0, I)$, iterative denoising is

$$\hat{G}_{l,t} \;=\; \hat{G}_{l,t-1} \;+\; \epsilon\, \nabla_{\hat{G}_{l,t-1}} \log P\Big(\hat{G}_{l,t-1} \mid G_p, G_l\Big) \;+\; \sqrt{2\epsilon}\, z_t, \quad 1 \le t \le T. \tag{3}$$

Gaussian perturbation at noise level σ. At step t with noise scale σ_t, the perturbed sample is modeled as

$$P_\sigma(G_{l,t} \mid G_l^*,\, G_p) \;=\; \mathcal{N}\big(G_{l,t} \,\big|\, G_l^*,\, \sigma^2 I\big), \tag{4}$$

with a decreasing schedule $\sigma_1 > \sigma_2 > \cdots > \sigma_T$.

Score network training loss. Let $\text{CFGenerator}(G_{l,t}, G_p, \sigma_t)$ estimate the score $\nabla_{G_{l,t}} \log P_{\sigma_t}(G_{l,t} \mid G_l^*)$. The denoising score matching loss is

$$\mathcal{L} = \frac{1}{T} \sum_{t=1}^{T} \sigma_t^2\, \mathbb{E}_{G_l^* \sim P_{\text{data}}}\, \mathbb{E}_{G_{l,t} \sim P_{\sigma_t}(\cdot \mid G_l^*)} \Big\| \text{CFGenerator}(G_{l,t}, G_p, \sigma_t) - \nabla_{G_{l,t}} \log P_{\sigma_t}(G_{l,t} \mid G_l^*) \Big\|^2 . \tag{5}$$

Generation-time Langevin (with learned score). At inference, replace the true score with the network output:

$$\hat{G}_{l,t} = \hat{G}_{l,t-1} + \epsilon_{t-1}\, \text{CFGenerator}\big(\hat{G}_{l,t-1},\, G_p,\, \sigma_{t-1}\big) + \sqrt{2\epsilon_{t-1}}\, z_t, \quad 1 \le t \le T. \tag{6}$$

Knowledge-guided LPM Representation

Heterogeneous geometric graph. At step t, the LPM encoder builds

$$G_t \;=\; \text{LPMEncoder}(G_{l,t},\, G_p) \;=\; \big\{\, G_{l,t},\; G_p,\; G_{lp} \,\big\}, \tag{7}$$

where $G_{l,t}$ is the ligand graph (atoms V_l, coordinates x_t, edges E_l), G_p is the pharmacophore graph (feature points V_p, coordinates x_p, edges E_p plus connections from exclusion spheres), and G_{lp} is a bipartite graph linking ligand atoms to pharmacophore points. The bipartite features include type-matching vectors V_{lp} and direction-matching vectors N_{lp}.

Geometry: Translations, Rotations, and Torsions

Conformation manifold. Let m be the number of rotatable bonds. A ligand conformation lies on an $(m+6)$-dimensional product space

$$g = (\, r,\, R,\, \theta\,) \in \mathcal{P}, \tag{8}$$

$$\mathcal{P} = \mathbb{T}^3 \times \mathrm{SO}(3) \times \mathrm{SO}(2)^m, \tag{9}$$

where $r \in \mathbb{T}^3$ (translation), $R \in \mathrm{SO}(3)$ (rotation), and $\theta \in \mathrm{SO}(2)^m$ (torsion angles).

Applying a pose update. Given coordinates x_t at step t, a pose update $g = (r, R, \theta)$ yields

$$x_{t-\Delta t} = \mathcal{A}(r, R, \theta;\, x_t) = \mathcal{A}_{\mathrm{tr}}(r)\; \mathcal{A}_{\mathrm{rot}}(R)\; \mathcal{A}_{\mathrm{tor}}(\theta)\; x_t. \tag{10}$$

Predicting change directions (scores). Rather than predicting absolute poses, the generator outputs score-like directions:

$$(\alpha,\, \beta,\, \gamma) = \text{CFGenerator}(G_t,\, t), \tag{11}$$

interpretable as directions for translation Δr, rotation ΔR, and torsions $\Delta\theta$, respectively.