universität wien

Diffusion models for Molecular Dynamics

A First Look at MDGen

Bogdan Chuzhinov

Faculty of Mathematics, University of Vienna

Optimization Seminar, 13 January 2026

Proteins

Proteins are **folded chains of amino acids** that perform most functional tasks in the cell:

- catalysis,
- building,
- transport,
- signaling,
- defense.

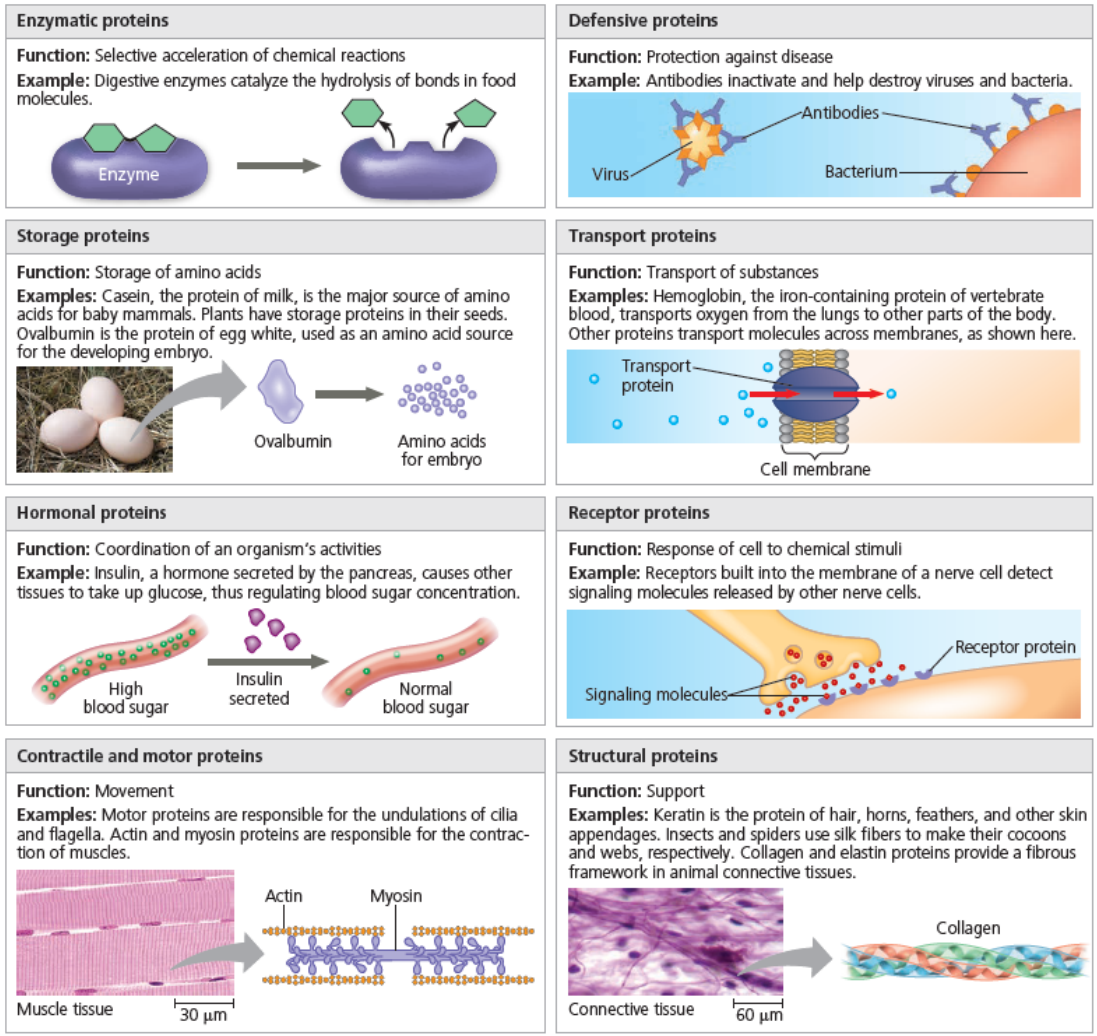


Image source: Fig. 5.13 from L. A. Urry *et al.*, *Campbell Biology*, 11th ed., Pearson, 2017.

Proteins

Proteins are **folded chains of amino acids** that perform most functional tasks in the cell:

- catalysis,
- building,
- transport,
- signaling,
- defense.

Function

=

Structure

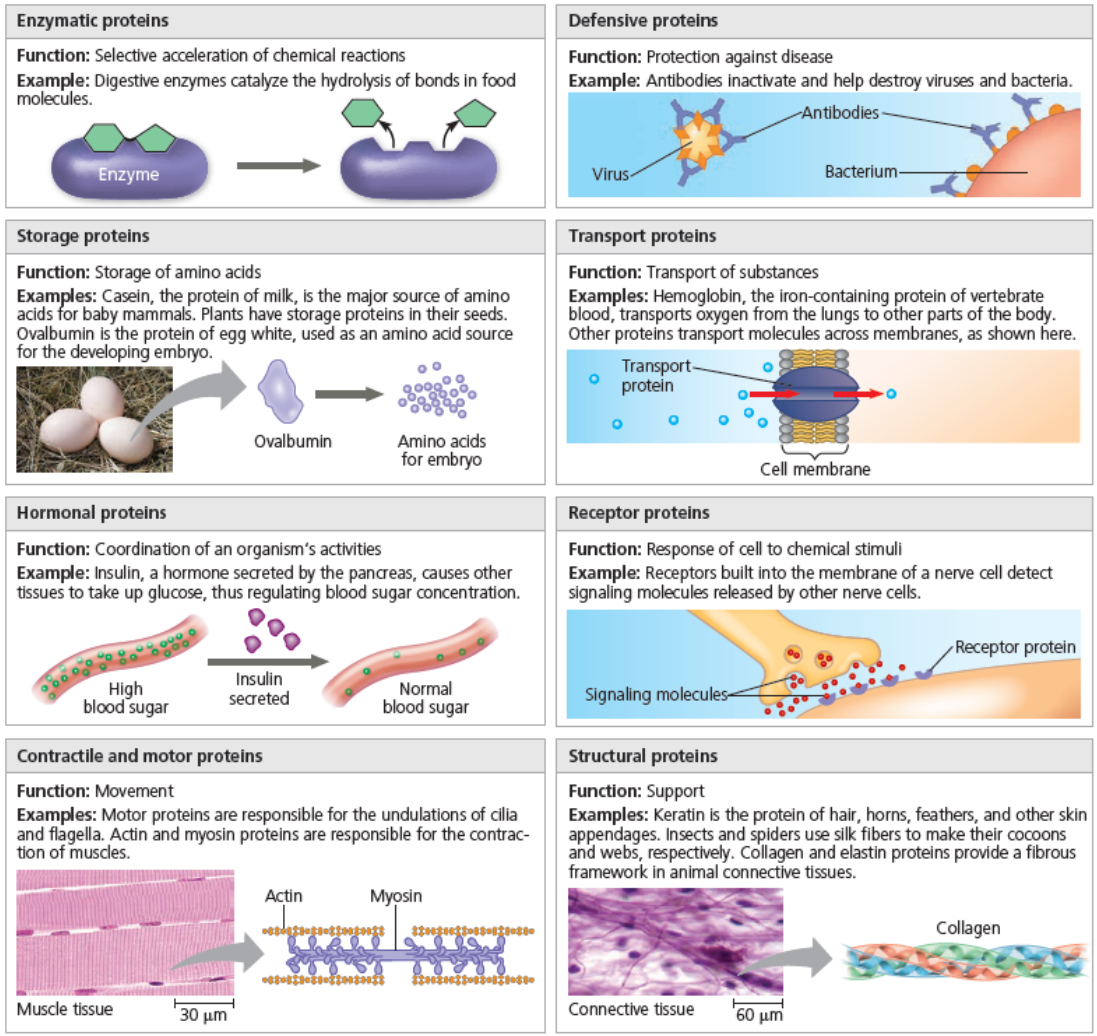


Image source: Fig. 5.13 from L. A. Urry *et al.*, *Campbell Biology*, 11th ed., Pearson, 2017.

Structure of Proteins

Protein structure is organized into **four hierarchical levels**:

Image: Jmarchn (after Thomas Shafee), *Protein structure (full)*, Wikimedia Commons, CC BY–SA 3.0.

Structure of Proteins

Protein structure is organized into **four hierarchical levels**:

- the amino acid sequence of the polypeptide chain,

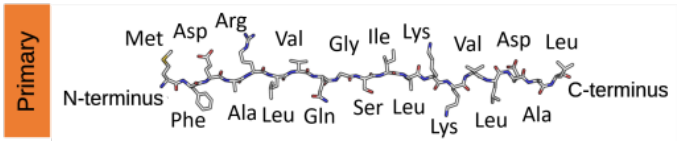


Image: Jmarchn (after Thomas Shafee), *Protein structure (full)*, Wikimedia Commons, CC BY–SA 3.0.

Structure of Proteins

Protein structure is organized into **four hierarchical levels**:

- the amino acid sequence of the polypeptide chain,
- local folding into α-helices and β-sheets stabilized by hydrogen bonds,

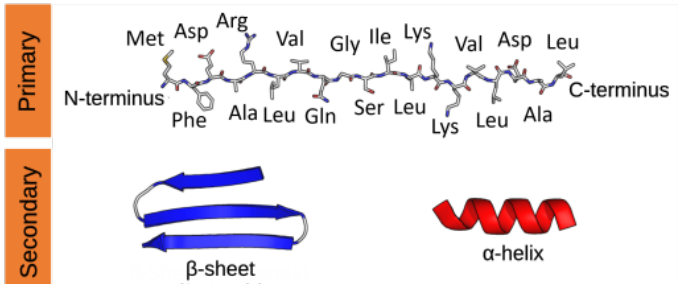


Image: Jmarchn (after Thomas Shafee), *Protein structure (full)*, Wikimedia Commons, CC BY–SA 3.0.

Structure of Proteins

Protein structure is organized into **four hierarchical levels**:

- the amino acid sequence of the polypeptide chain,
- local folding into α-helices and β-sheets stabilized by hydrogen bonds,
- the overall 3D shape of a single protein, determined by side-chain interactions,

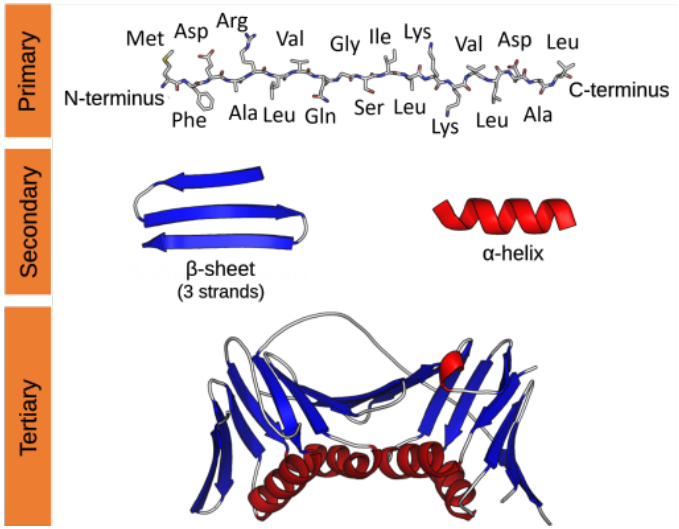
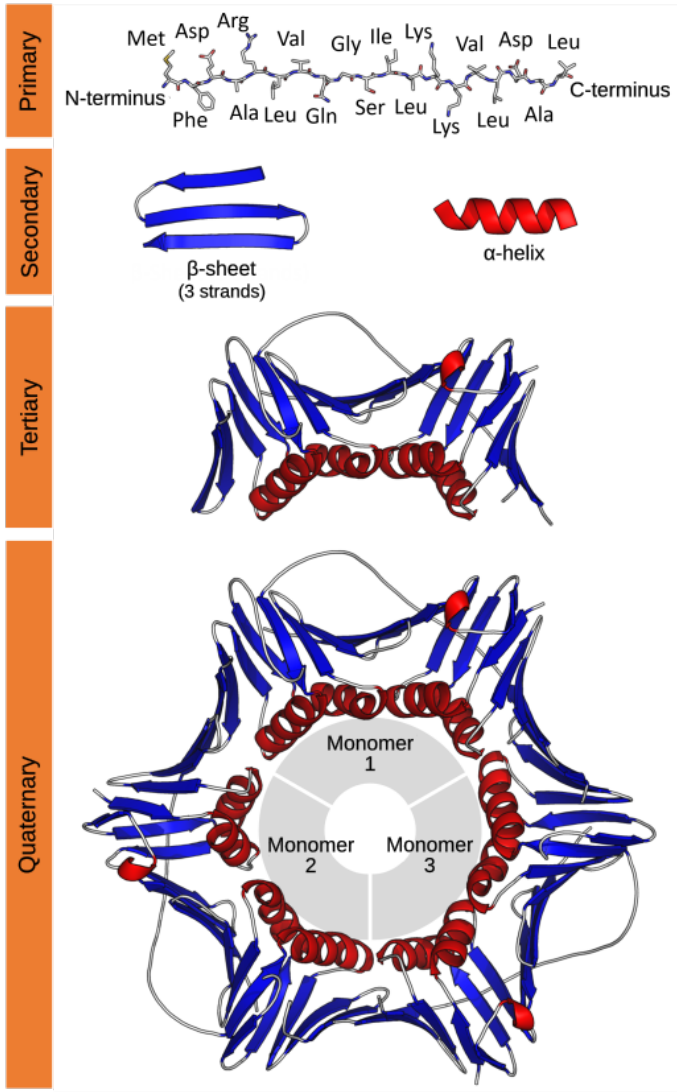


Image: Jmarchn (after Thomas Shafee), *Protein structure (full)*, Wikimedia Commons, CC BY–SA 3.0.

Structure of Proteins

Protein structure is organized into **four hierarchical levels**:

- the amino acid sequence of the polypeptide chain,
- local folding into α-helices and β-sheets stabilized by hydrogen bonds,
- the overall 3D shape of a single protein, determined by side-chain interactions,
- the assembly of multiple polypeptide chains into a functional complex.



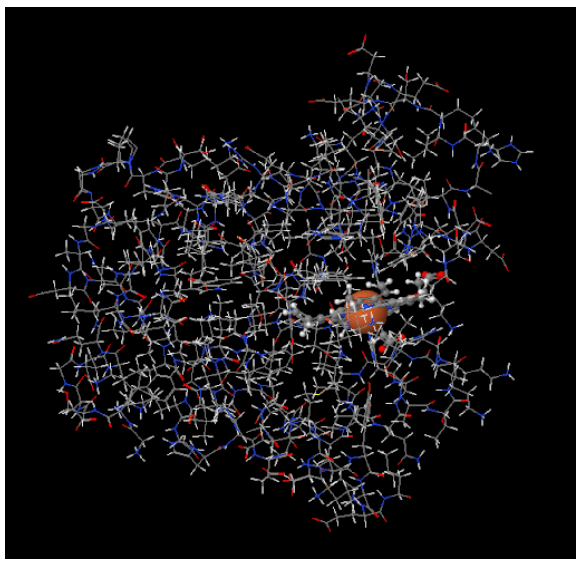
How proteins look like?

John Kendrew and Max Perutz with myoglobin model

How proteins look like?

1. Max Perutz (right) and John Kendrew with myoglobin model, ca. 1960s. Source: *Science History Institute*,
2. The original model of the myoglobin molecule, Science Museum Group Collection, ©The Board of Trustees of the Science Museum, CC BY–NC–SA 4.0,

How proteins look like?



1. Max Perutz (right) and John Kendrew with myoglobin model, ca. 1960s. Source: *Science History Institute*,
2. The original model of the myoglobin molecule, Science Museum Group Collection, ©The Board of Trustees of the Science Museum, CC BY–NC–SA 4.0,
3. Myoglobine (animated myoglobin structure), by Lesuperfétatoire, CC0 1.0 Universal Public Domain Dedication. Source: Wikimedia Commons.

Molecular Dynamics Force Field

$$m_i \ddot{x}_i = -\nabla_{x_i} U(x_1, \ldots, x_N)$$

Molecular Dynamics Force Field

$$m_i \ddot{x}_i = -\nabla_{x_i} U(x_1, \ldots, x_N)$$

$$U = \sum_{\text{bonds}} \tfrac{1}{2} k_r (r - r_0)^2$$

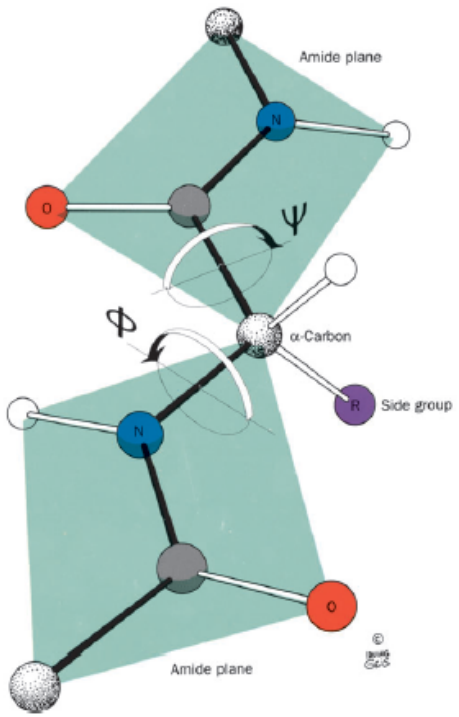


Image: Fig. 6.4 from D. Voet *et al.*, *Fundamentals of Biochemistry: Life at the Molecular Level*, 4th ed., John Wiley & Sons, Inc., 2013.

Molecular Dynamics Force Field

$$m_i \ddot{x}_i = -\nabla_{x_i} U(x_1, \ldots, x_N)$$

$$U = \sum_{\text{bonds}} \tfrac{1}{2} k_r (r - r_0)^2 \quad + \quad \sum_{\text{angles}} \tfrac{1}{2} k_\theta (\theta - \theta_0)^2 +$$

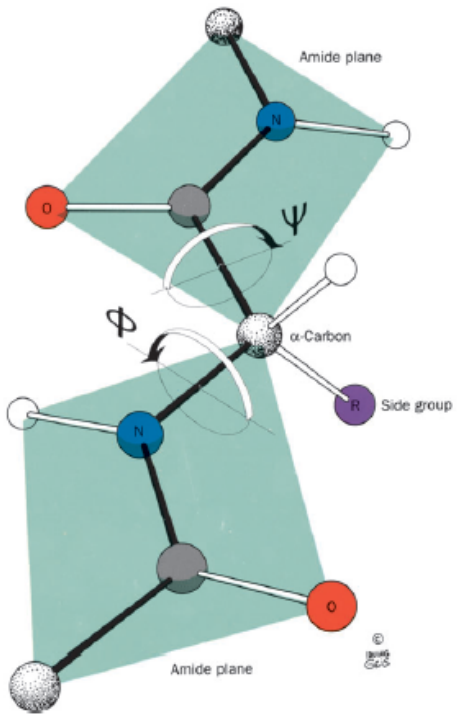


Image: Fig. 6.4 from D. Voet *et al.*, *Fundamentals of Biochemistry: Life at the Molecular Level*, 4th ed., John Wiley & Sons, Inc., 2013.

Molecular Dynamics Force Field

$$m_i \ddot{x}_i = -\nabla_{x_i} U(x_1, \ldots, x_N)$$

$$U = \sum_{\text{bonds}} \tfrac{1}{2} k_r (r - r_0)^2 + \sum_{\text{angles}} \tfrac{1}{2} k_\theta (\theta - \theta_0)^2 +$$

$$+ \sum_{\text{torsions}} \frac{V_n}{2} [1 + \cos(n\varphi - \delta)]$$

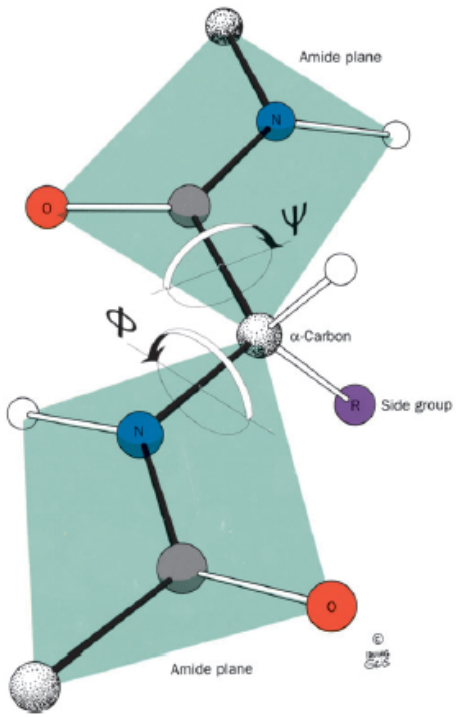


Image: Fig. 6.4 from D. Voet *et al.*, *Fundamentals of Biochemistry: Life at the Molecular Level*, 4th ed., John Wiley & Sons, Inc., 2013.

Molecular Dynamics Force Field

$$m_i \ddot{x}_i = -\nabla_{x_i} U(x_1, \ldots, x_N)$$

$$U = \sum_{\text{bonds}} \tfrac{1}{2} k_r (r - r_0)^2 + \sum_{\text{angles}} \tfrac{1}{2} k_\theta (\theta - \theta_0)^2 +$$

$$+ \sum_{\text{torsions}} \frac{V_n}{2} [1 + \cos(n\varphi - \delta)] + \sum_{\text{imp. tors.}} V(\omega) +$$

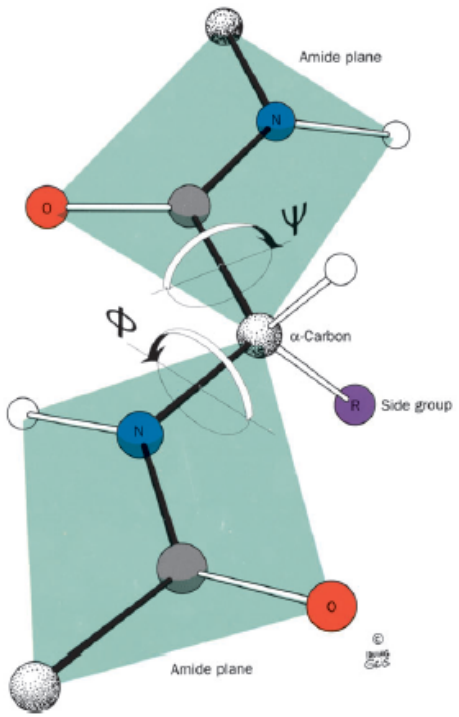


Image: Fig. 6.4 from D. Voet *et al.*, *Fundamentals of Biochemistry: Life at the Molecular Level*, 4th ed., John Wiley & Sons, Inc., 2013.

Molecular Dynamics Force Field

$$m_i \ddot{x}_i = -\nabla_{x_i} U(x_1, \ldots, x_N)$$

$$U = \sum_{\text{bonds}} \tfrac{1}{2} k_r (r - r_0)^2 + \sum_{\text{angles}} \tfrac{1}{2} k_\theta (\theta - \theta_0)^2 +$$

$$+ \sum_{\text{torsions}} \frac{V_n}{2} [1 + \cos(n\varphi - \delta)] + \sum_{\text{imp. tors.}} V(\omega) +$$

$$+ \sum_{\text{elec}} \frac{q_i q_j}{r_{ij}}$$

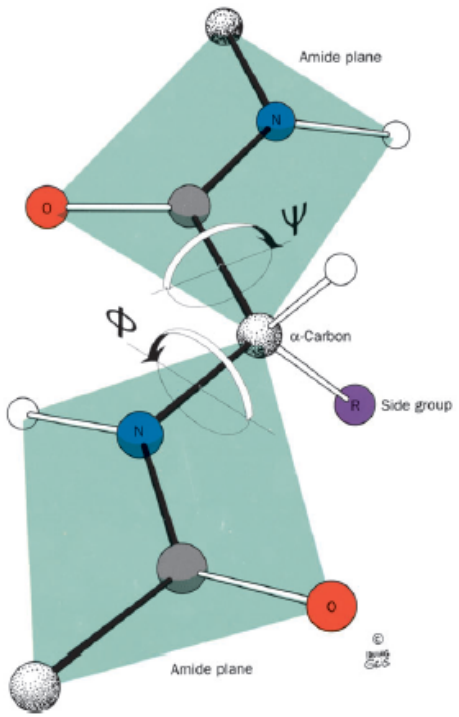


Image: Fig. 6.4 from D. Voet *et al.*, *Fundamentals of Biochemistry: Life at the Molecular Level*, 4th ed., John Wiley & Sons, Inc., 2013.

Molecular Dynamics Force Field

$$m_i \ddot{x}_i = -\nabla_{x_i} U(x_1, \ldots, x_N)$$

$$U = \sum_{\text{bonds}} \tfrac{1}{2} k_r (r - r_0)^2 + \sum_{\text{angles}} \tfrac{1}{2} k_\theta (\theta - \theta_0)^2 +$$

$$+ \sum_{\text{torsions}} \frac{V_n}{2} [1 + \cos(n\varphi - \delta)] + \sum_{\text{imp. tors.}} V(\omega) +$$

$$+ \sum_{\text{elec}} \frac{q_i q_j}{r_{ij}} + \sum_{\text{L-J}} \left(\frac{A_{ij}}{r_{ij}^{12}} - \frac{B_{ij}}{r_{ij}^{6}} \right)$$

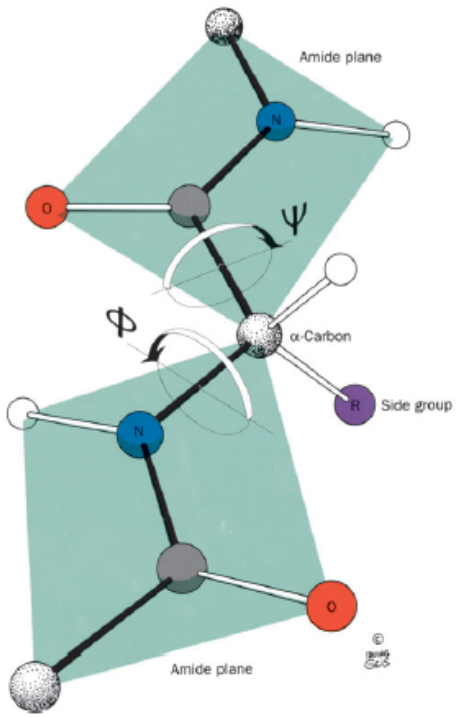
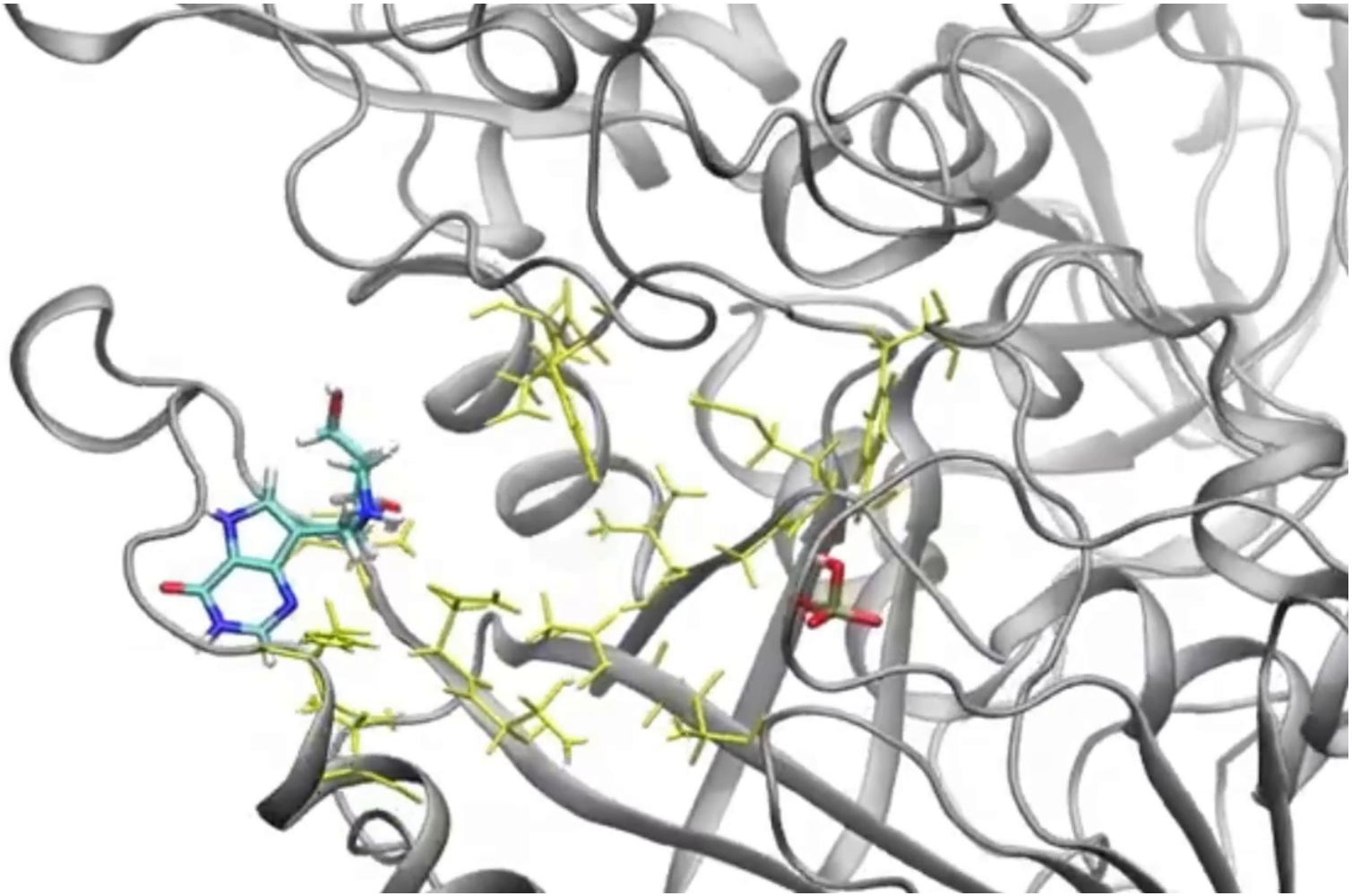


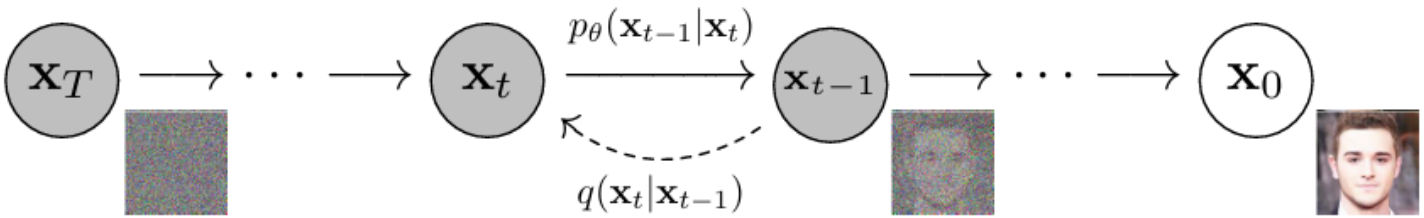
Image: Fig. 6.4 from D. Voet *et al.*, *Fundamentals of Biochemistry: Life at the Molecular Level*, 4th ed., John Wiley & Sons, Inc., 2013.



Video: Brain Science — Molecular dynamics simulation of a drug entering into a target protein. Source: YouTube channel *Istituto Italiano di Tecnologia*

Generative Models

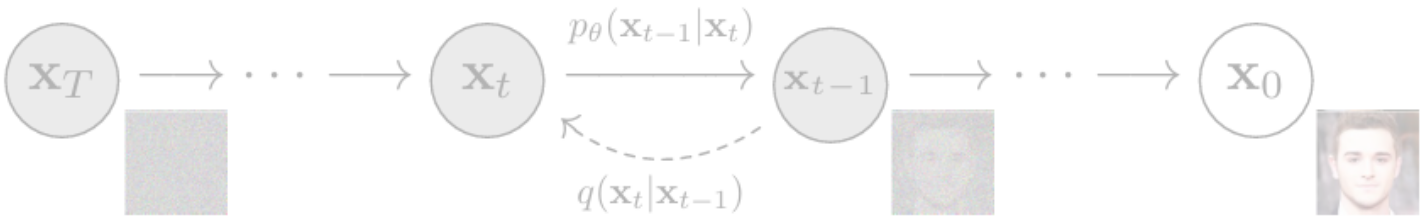
$$x_t = x_{t-1} + \varepsilon_t$$

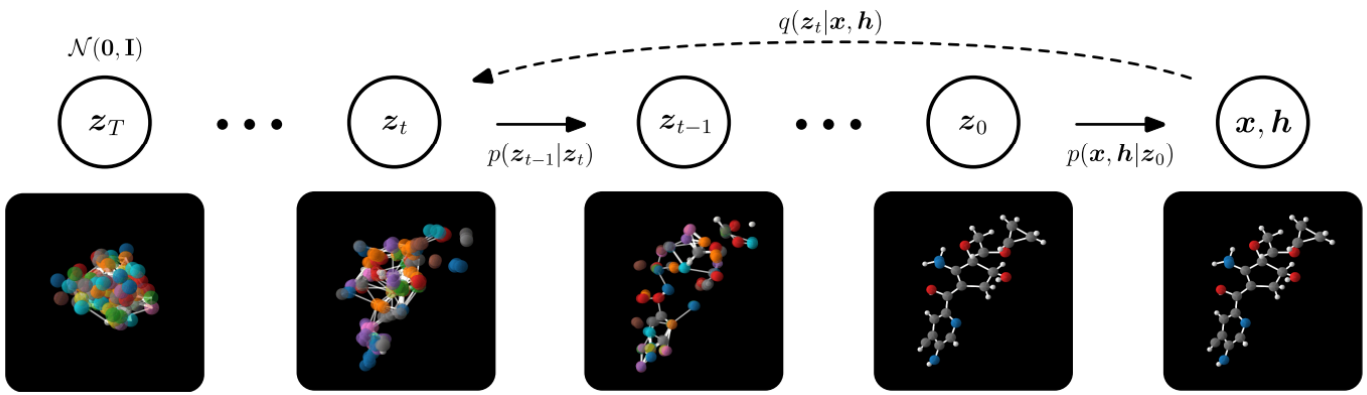


1. Image: Fig. 2 from J. Ho, A. Jain, P. Abbeel, *Denoising Diffusion Probabilistic Models*, arXiv:2006.11239 (2020).

Generative Models

$$x_t = x_{t-1} + \varepsilon_t$$





1. Image: Fig. 2 from J. Ho, A. Jain, P. Abbeel, *Denoising Diffusion Probabilistic Models*, arXiv:2006.11239 (2020).
2. Image: Fig. 2 from E. Hoogeboom, V. Garcia Satorras, C. Vignac, M. Welling, *Equivariant Diffusion for Molecule Generation in 3D*, arXiv:2203.17003 (2022).

Stochastic Interpolants: A Unifying Framework for Flows and Diffusions

Michael S. Albergo, Nicholas M. Boffi, Eric Vanden-Eijnden

Journal of Machine Learning Research, 2025

The section is based on this paper.

Definition (Stochastic Interpolant)

Given two probability density functions $\rho_0, \rho_1 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$, a *stochastic interpolant* between ρ_0 and ρ_1 is a stochastic process

$$x_t = I(t, x_0, x_1) + \gamma(t)\, z, \qquad t \in [0, 1],$$

where:

- $I \in C^2\big([0,1], C^2(\mathbb{R}^d \times \mathbb{R}^d; \mathbb{R}^d)\big)$ satisfies the boundary conditions $I(0, x_0, x_1) = x_0$, $I(1, x_0, x_1) = x_1$, as well as the Lipschitz condition

$$\exists C_1 < \infty : \quad |\partial_t I(t, x_0, x_1)| \leq C_1 |x_0 - x_1|, \qquad \forall (t, x_0, x_1) \in [0,1] \times \mathbb{R}^d \times \mathbb{R}^d,$$

- $\gamma^2 \in C^2([0,1])$ satisfies $\gamma(0) = \gamma(1) = 0$, $\gamma(t) > 0$ for all $t \in (0,1)$,
- the pair (x_0, x_1) is drawn from a probability measure ν whose marginals are ρ_0 and ρ_1, i.e.

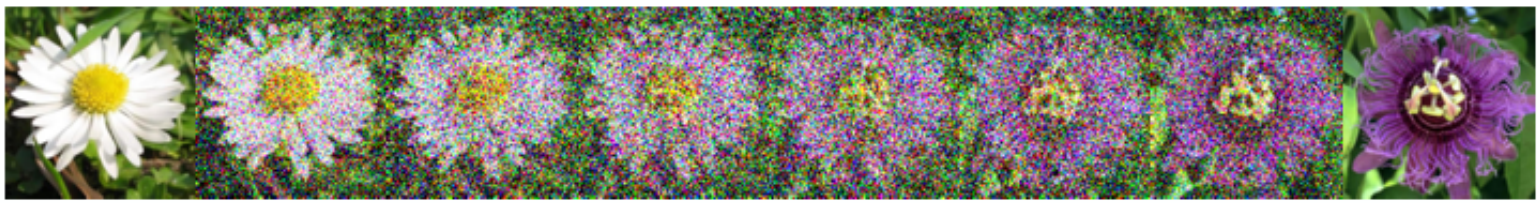
$$\nu(dx_0, \mathbb{R}^d) = \rho_0(x_0)\, dx_0, \qquad \nu(\mathbb{R}^d, dx_1) = \rho_1(x_1)\, dx_1,$$

- z is a standard Gaussian random variable independent of (x_0, x_1).

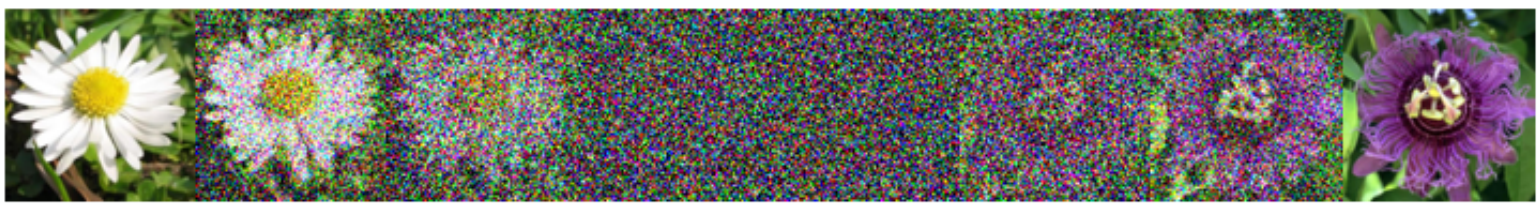
Example

Without latent variable $x_t = (1-t)x_0 + tx_1$

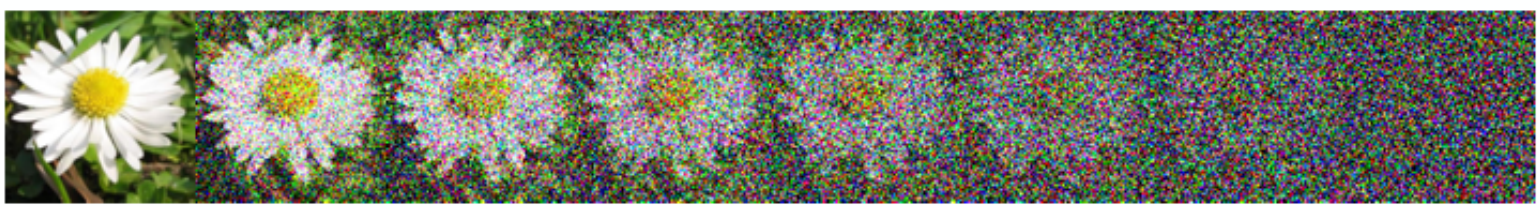
With latent variable $x_t = (1-t)x_0 + tx_1 + \sqrt{2t(1-t)}z$



Gaussian encoding-decoding $x_t = \cos^2(\pi t)(1_{[0,\frac{1}{2})}(t)x_0 + 1_{(\frac{1}{2},1]}(t)x_1) + \sqrt{2t(1-t)}z$



One-sided $x_t = (1-t)x_0 + tz$



Regularity assumptions

Assumption A. The endpoint densities $\rho_0, \rho_1 \in C^2(\mathbb{R}^d)$ are *strictly positive* and satisfy

$$\int_{\mathbb{R}^d} |\nabla \log \rho_0(x)|^2 \rho_0(x)\, dx < \infty, \qquad \int_{\mathbb{R}^d} |\nabla \log \rho_1(x)|^2 \rho_1(x)\, dx < \infty.$$

Assumption B. The coupling $\nu(x_0, x_1)$ and the interpolant $I(t, x_0, x_1)$ satisfy the moment bounds

$$\exists M_1, M_2 < \infty \text{ s.t. } \forall t \in [0,1]:$$

$$\mathbb{E}_{(x_0,x_1)\sim\nu}\left[\, |\partial_t I(t, x_0, x_1)|^4 \,\right] \le M_1, \qquad \mathbb{E}_{(x_0,x_1)\sim\nu}\left[\, |\partial_t^2 I(t, x_0, x_1)|^2 \,\right] \le M_2.$$

Transport Equation

Theorem

Let x_t be a stochastic interpolant. Then its law is absolutely continuous for all $t \in [0,1]$, with *strictly positive* density $\rho(t,x)$ satisfying $\rho(0,x) = \rho_0(x)$, $\rho(1,x) = \rho_1(x)$ and $\rho \in C^1\big([0,1]; C^p(\mathbb{R}^d)\big)$ for all $p \in \mathbb{N}$.

Moreover, ρ satisfies the *transport equation* (TE)

$$\partial_t \rho(t,x) + \nabla \cdot \big(b(t,x)\,\rho(t,x)\big) = 0,$$

where the *velocity field* is defined by

$$b(t,x) = \mathbb{E}[\,\dot{x}_t \mid x_t = x\,] = \mathbb{E}[\,\partial_t I(t,x_0,x_1) + \dot{\gamma}(t)\,z \mid x_t = x\,].$$

The velocity field satisfies $b \in C^0\big([0,1]; (C^p(\mathbb{R}^d))^d\big)$ for all $p \in \mathbb{N}$, and

$$\forall t \in [0,1]: \qquad \int_{\mathbb{R}^d} |b(t,x)|^2\,\rho(t,x)\,dx < \infty.$$

Objective

Theorem

The velocity field

$$b(t,x) = \mathbb{E}[\,\dot{x}_t \mid x_t = x\,] = \mathbb{E}[\,\partial_t I(t,x_0,x_1) + \dot{\gamma}(t)\,z \mid x_t = x]$$

is the unique minimizer in $C^0\big([0,1];(C^1(\mathbb{R}^d))^d\big)$ of the quadratic functional

$$\mathcal{L}_b[\hat{b}] = \int_0^1 \mathbb{E}\left[\frac{1}{2}\,|\hat{b}(t,x_t)|^2 - \big(\partial_t I(t,x_0,x_1) + \dot{\gamma}(t)\,z\big)\cdot\hat{b}(t,x_t)\right] dt,$$

where $x_t = I(t,x_0,x_1) + \gamma(t)z$ and the expectation is taken independently over $(x_0,x_1)\sim\nu$, $z\sim\mathcal{N}(0,I_d)$.

Score and Denoiser

Theorem

The *score* of the probability density ρ belongs to $C^1\big([0,1]; (C^p(\mathbb{R}^d))^d\big)$ for arbitrary $p \in \mathbb{N}$, and for all $(t,x) \in (0,1) \times \mathbb{R}^d$ is given by

$$s(t,x) = \nabla \log \rho(t,x) = -\gamma(t)^{-1}\, \mathbb{E}[\, z \mid x_t = x\,].$$

Moreover, it satisfies

$$\forall t \in [0,1]: \qquad \int_{\mathbb{R}^d} |s(t,x)|^2\, \rho(t,x)\, dx < \infty,$$

and it is the unique minimizer in $C^1\big([0,1]; (C^1(\mathbb{R}^d))^d\big)$ of the quadratic functional

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left[\frac{1}{2}\,|\hat{s}(t,x_t)|^2 + \gamma(t)^{-1}\, z \cdot \hat{s}(t,x_t)\right] dt.$$

Score and Denoiser

Theorem

The *score* of the probability density ρ belongs to $C^1\big([0,1]; (C^p(\mathbb{R}^d))^d\big)$ for arbitrary $p \in \mathbb{N}$, and for all $(t,x) \in (0,1) \times \mathbb{R}^d$ is given by

$$s(t,x) = \nabla \log \rho(t,x) = -\gamma(t)^{-1}\, \mathbb{E}[\, z \mid x_t = x\,].$$

Moreover, it satisfies

$$\forall t \in [0,1]: \qquad \int_{\mathbb{R}^d} |s(t,x)|^2\, \rho(t,x)\, dx < \infty,$$

and it is the unique minimizer in $C^1\big([0,1]; (C^1(\mathbb{R}^d))^d\big)$ of the quadratic functional

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left[\frac{1}{2}\,|\hat{s}(t,x_t)|^2 + \gamma(t)^{-1}\, z \cdot \hat{s}(t,x_t)\right] dt.$$

Score and Denoiser

Theorem

The *score* of the probability density ρ belongs to $C^1\big([0,1];(C^p(\mathbb{R}^d))^d\big)$ for arbitrary $p \in \mathbb{N}$, and for all $(t,x) \in (0,1)\times\mathbb{R}^d$ is given by

$$s(t,x) = \nabla \log \rho(t,x) = -\gamma(t)^{-1}\,\eta_z(t,x).$$

Moreover, it satisfies

$$\forall t \in [0,1]: \qquad \int_{\mathbb{R}^d} |s(t,x)|^2\,\rho(t,x)\,dx < \infty,$$

and it is the unique minimizer in $C^1\big([0,1];(C^1(\mathbb{R}^d))^d\big)$ of the quadratic functional

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left[\frac{1}{2}\,|\hat{s}(t,x_t)|^2 + \gamma(t)^{-1}\,z\cdot\hat{s}(t,x_t)\right] dt.$$

Score and Denoiser

Theorem

The *score* of the probability density ρ belongs to $C^1\big([0,1]; (C^p(\mathbb{R}^d))^d\big)$ for arbitrary $p \in \mathbb{N}$, and for all $(t,x) \in (0,1) \times \mathbb{R}^d$ is given by

$$s(t,x) = \nabla \log \rho(t,x) = -\gamma(t)^{-1}\, \eta_z(t,x).$$

Moreover, it satisfies

$$\forall t \in [0,1]: \qquad \int_{\mathbb{R}^d} |s(t,x)|^2\, \rho(t,x)\, dx < \infty,$$

and it is the unique minimizer in $C^1\big([0,1]; (C^1(\mathbb{R}^d))^d\big)$ of the quadratic functional

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \mathbb{E}\left[\frac{1}{2}\,|-\gamma(t)^{-1}\,\hat{\eta}_z(t,x)|^2 + \gamma(t)^{-1}\, z \cdot (-\gamma(t)^{-1}\,\hat{\eta}_z(t,x))\right] dt.$$

Score and Denoiser

Theorem

The *score* of the probability density ρ belongs to $C^1\big([0,1];(C^p(\mathbb{R}^d))^d\big)$ for arbitrary $p \in \mathbb{N}$, and for all $(t,x) \in (0,1)\times\mathbb{R}^d$ is given by

$$s(t,x) = \nabla \log \rho(t,x) = -\gamma(t)^{-1}\,\eta_z(t,x).$$

Moreover, it satisfies

$$\forall t \in [0,1]: \qquad \int_{\mathbb{R}^d} |s(t,x)|^2\,\rho(t,x)\,dx < \infty,$$

and it is the unique minimizer in $C^1\big([0,1];(C^1(\mathbb{R}^d))^d\big)$ of the quadratic functional

$$\mathcal{L}_s[\hat{s}] = \int_0^1 \gamma(t)^{-2}\,\mathbb{E}\left[\frac{1}{2}\,|\hat{\eta}_z(t,x)|^2 - z\cdot\hat{\eta}_z(t,x)\right] dt.$$

Score and Denoiser

Theorem

The *score* of the probability density ρ belongs to $C^1\big([0,1];(C^p(\mathbb{R}^d))^d\big)$ for arbitrary $p \in \mathbb{N}$, and for all $(t,x) \in (0,1)\times\mathbb{R}^d$ is given by

$$s(t,x) = \nabla \log \rho(t,x) = -\gamma(t)^{-1}\,\eta_z(t,x).$$

Moreover, it satisfies

$$\forall t \in [0,1]: \qquad \int_{\mathbb{R}^d} |s(t,x)|^2\,\rho(t,x)\,dx < \infty,$$

and it is the unique minimizer in $C^1\big([0,1];(C^1(\mathbb{R}^d))^d\big)$ of the quadratic functional

$$\mathcal{L}_{\eta_z}[\hat{\eta}_z] = \int_0^1 \mathbb{E}\left[\frac{1}{2}\,|\hat{\eta}_z(t,x)|^2 - z\cdot\hat{\eta}_z(t,x)\right] dt.$$

Fokker–Planck equations

Theorem

For any $\varepsilon \in C^0([0,1])$ with $\varepsilon \geq 0$, the probability density ρ satisfies:

1. Forward Fokker–Planck equation

$$\partial_t \rho + \nabla \cdot (b_F \rho) = \varepsilon(t)\, \Delta \rho, \qquad \rho(0) = \rho_0,$$

where the *forward drift* is

$$b_F(t,x) = b(t,x) + \varepsilon(t)\, s(t,x).$$

2. Backward Fokker–Planck equation

$$\partial_t \rho + \nabla \cdot (b_B \rho) = -\varepsilon(t)\, \Delta \rho, \qquad \rho(1) = \rho_1,$$

where the *backward drift* is

$$b_B(t,x) = b(t,x) - \varepsilon(t)\, s(t,x).$$

Generative models

Theorem

At any time $t \in [0,1]$, the law of the stochastic interpolant coincides with the law of the three processes X_t, X_t^F and X_t^B, defined as:

1. **Probability flow (transport ODE)**

$$\frac{d}{dt}X_t = b(t, X_t), \qquad \text{with} \quad X_{t=0} = x_0 \sim \rho_0 \quad \text{or} \quad X_{t=1} = x_1 \sim \rho_1.$$

2. **Forward SDE (associated with forward FPE)**

$$dX_t^F = b_F(t, X_t^F)\, dt + \sqrt{2\varepsilon(t)}\, dW_t,$$

solved forward in time from the initial data $X_{t=0}^F \sim \rho_0$, independent of W.

3. **Backward SDE (associated with backward FPE)**

$$dX_t^B = b_B(t, X_t^B)\, dt + \sqrt{2\varepsilon(t)}\, dW_t^B, \qquad W_t^B = -W_{1-t},$$

solved backward in time from $X_{t=1}^B \sim \rho_1$, independent of W^B.

KL divergence for TE

Theorem

Let $\rho_0 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be a fixed base probability density. Given two velocity fields $b, \hat{b} \in C^0([0,1], (C^1(\mathbb{R}^d))^d)$, let the time-dependent densities $\rho, \hat{\rho} : [0,1] \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ denote the solutions of the TEs

$$\partial_t \rho + \nabla \cdot (b\rho) = 0, \qquad \rho(0) = \rho_0,$$

$$\partial_t \hat{\rho} + \nabla \cdot (\hat{b}\,\hat{\rho}) = 0, \qquad \hat{\rho}(0) = \rho_0.$$

Then the Kullback–Leibler divergence of $\rho(1)$ from $\hat{\rho}(1)$ is

$$\mathrm{KL}\big(\rho(1) \,\|\, \hat{\rho}(1)\big) =$$

$$= \int_0^1 \int_{\mathbb{R}^d} \big(\nabla \log \hat{\rho}(t,x) - \nabla \log \rho(t,x)\big) \cdot \big(\hat{b}(t,x) - b(t,x)\big)\, \rho(t,x)\, dx\, dt.$$

KL divergence for Fokker–Planck equations

Theorem

Let $\rho_0 : \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ be a fixed base probability density. Given two velocity fields $b_F, \hat{b}_F \in C^0([0,1], (C^1(\mathbb{R}^d))^d)$, let the densities $\rho, \hat{\rho} : [0,1] \times \mathbb{R}^d \to \mathbb{R}_{\geq 0}$ denote the solutions to the Fokker–Planck equations

$$\partial_t \rho + \nabla \cdot (b_F \rho) = \varepsilon \, \Delta \rho, \qquad \rho(0) = \rho_0,$$

$$\partial_t \hat{\rho} + \nabla \cdot (\hat{b}_F \hat{\rho}) = \varepsilon \, \Delta \hat{\rho}, \qquad \hat{\rho}(0) = \rho_0.$$

Then the Kullback–Leibler divergence from $\rho(1)$ to $\hat{\rho}(1)$ is

$$\mathrm{KL}\big(\rho(1) \,\|\, \hat{\rho}(1)\big) = \int_0^1 \int_{\mathbb{R}^d} \big(\nabla \log \hat{\rho}(t,x) - \nabla \log \rho(t,x)\big) \cdot \big(\hat{b}_F(t,x) - b_F(t,x)\big) \, \rho(t,x) \, dx \, dt -$$

$$- \varepsilon \int_0^1 \int_{\mathbb{R}^d} \big|\nabla \log \rho(t,x) - \nabla \log \hat{\rho}(t,x)\big|^2 \, \rho(t,x) \, dx \, dt.$$

As a result,

$$\mathrm{KL}\big(\rho(1) \,\|\, \hat{\rho}(1)\big) \;\leq\; \frac{1}{4\varepsilon} \int_0^1 \int_{\mathbb{R}^d} |\hat{b}_F(t,x) - b_F(t,x)|^2 \, \rho(t,x) \, dx \, dt.$$

From training objectives to KL divergence

Theorem

Let ρ denote the solution of the FPE with $\varepsilon(t) = \varepsilon > 0$. Given two velocity fields $\hat{b}, \hat{s} \in C^0([0,1], (C^1(\mathbb{R}^d))^d)$, define

$$\hat{b}_F(t,x) = \hat{b}(t,x) + \varepsilon \hat{s}(t,x), \qquad \hat{v}(t,x) = \hat{b}(t,x) + \gamma(t)\dot{\gamma}(t)\,\hat{s}(t,x).$$

Let $\hat{\rho}$ denote the solution of the FPE: $\partial_t \hat{\rho} + \nabla \cdot (\hat{b}_F \hat{\rho}) = \varepsilon\, \Delta \hat{\rho}$ with $\hat{\rho}(0) = \rho_0$. Then

$$\mathrm{KL}\big(\rho_1 \,\|\, \hat{\rho}(1)\big) \;\le\; \frac{1}{2\varepsilon}\Big(\mathcal{L}_b[\hat{b}] - \min_{\tilde{b}} \mathcal{L}_b[\tilde{b}]\Big) + \frac{\varepsilon}{2}\Big(\mathcal{L}_s[\hat{s}] - \min_{\tilde{s}} \mathcal{L}_s[\tilde{s}]\Big),$$

where $\mathcal{L}_b[\hat{b}]$ and $\mathcal{L}_s[\hat{s}]$ are the objective functions defined previously. Moreover,

$$\mathrm{KL}\big(\rho_1 \,\|\, \hat{\rho}(1)\big) \;\le\; \frac{1}{2\varepsilon}\Big(\mathcal{L}_v[\hat{v}] - \min_{\tilde{v}} \mathcal{L}_v[\tilde{v}]\Big) + \frac{\sup_{t\in[0,1]} |\gamma(t)\dot{\gamma}(t) - \varepsilon|^2}{2\varepsilon}\Big(\mathcal{L}_s[\hat{s}] - \min_{\tilde{s}} \mathcal{L}_s[\tilde{s}]\Big),$$

where

$$\mathcal{L}_v[\hat{v}] = \int_0^1 \mathbb{E}\Big[\frac{1}{2}\,|\hat{v}(t,x_t)|^2 - \partial_t I(t,x_0,x_1)\cdot \hat{v}(t,x_t)\Big]\, dt.$$

Multilayer Perceptron

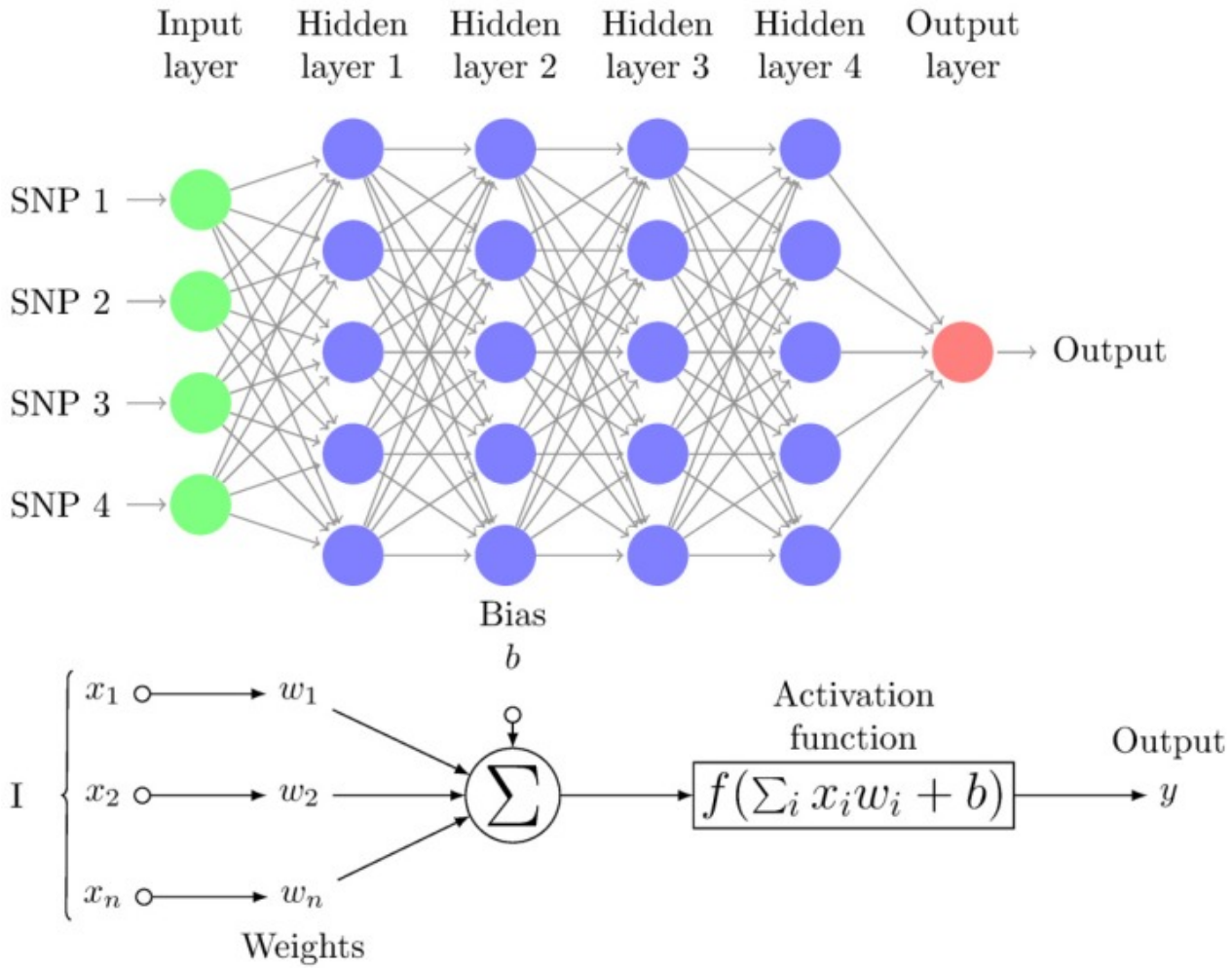


Image: Fig. 1 from M. Pérez-Enciso and L. M. Zingaretti, *A Guide for Using Deep Learning for Complex Trait Genomic Prediction*, Genes 10(7):553 (2019).

Convolutional Networks

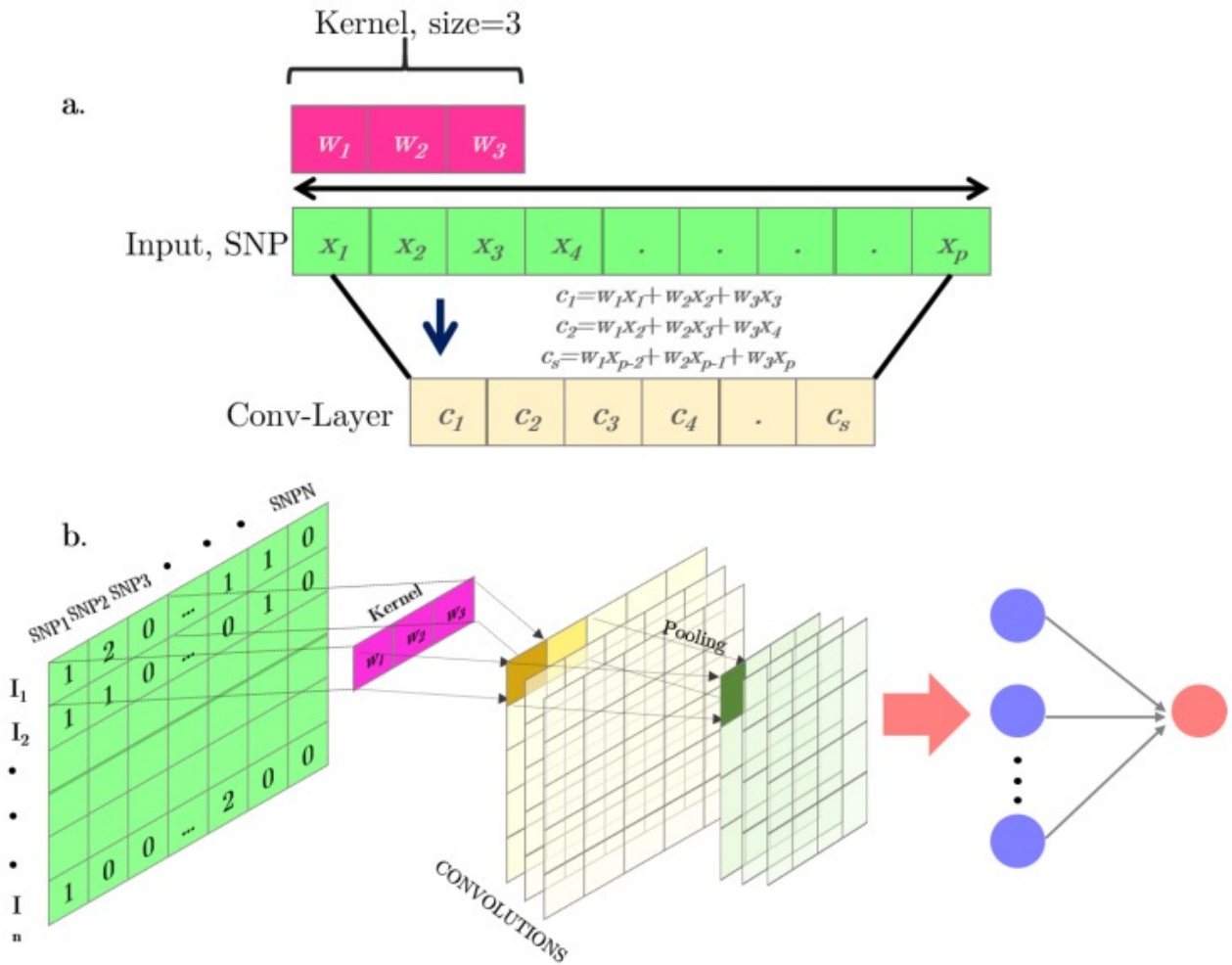


Image: Fig. 2 from M. Pérez-Enciso and L. M. Zingaretti, *A Guide for Using Deep Learning for Complex Trait Genomic Prediction*, Genes 10(7):553 (2019).

U-net

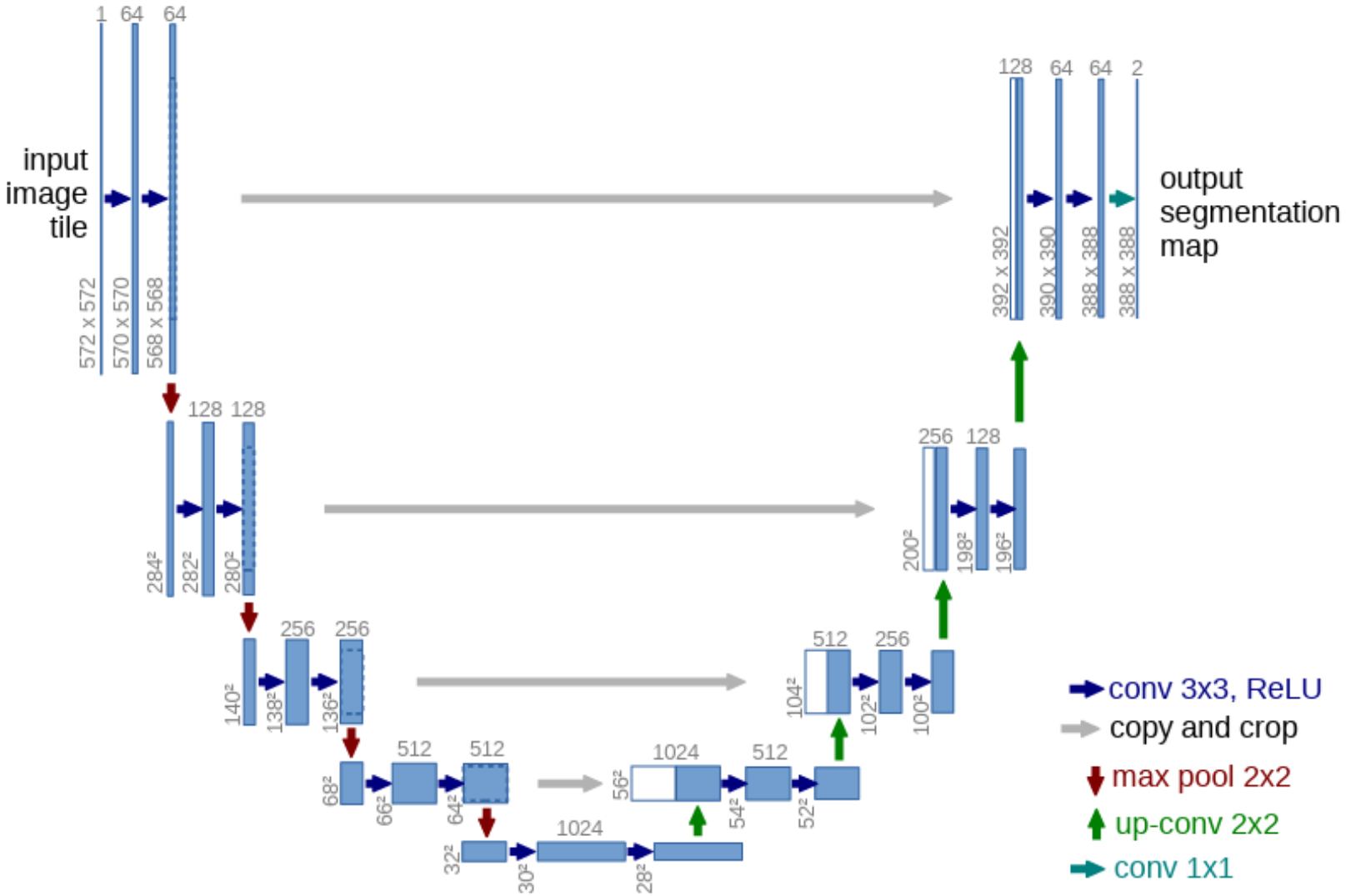


Image: Fig. 1 from O. Ronneberger, P. Fischer, and T. Brox, *U-Net: Convolutional Networks for Biomedical Image Segmentation*, arXiv:1505.04597, 2015.

Self-Attention

Each vector receives three representations ("roles")

Query: vector **from** which the attention is looking

"Hey there, do you have this information?"

Key: vector **at** which the query looks to compute weights

"Hi, I have this information – give me a large weight!"

Value: their weighted sum is attention output

"Here's the information I have!"

The formula for computing attention output is as follows:

$$Attention(q, k, v) = \overbrace{softmax\left(\frac{qk^T}{\sqrt{d_k}}\right)}^{\text{Attention weights}} v$$

q: from, k: to, v: vector dimensionality of K, V

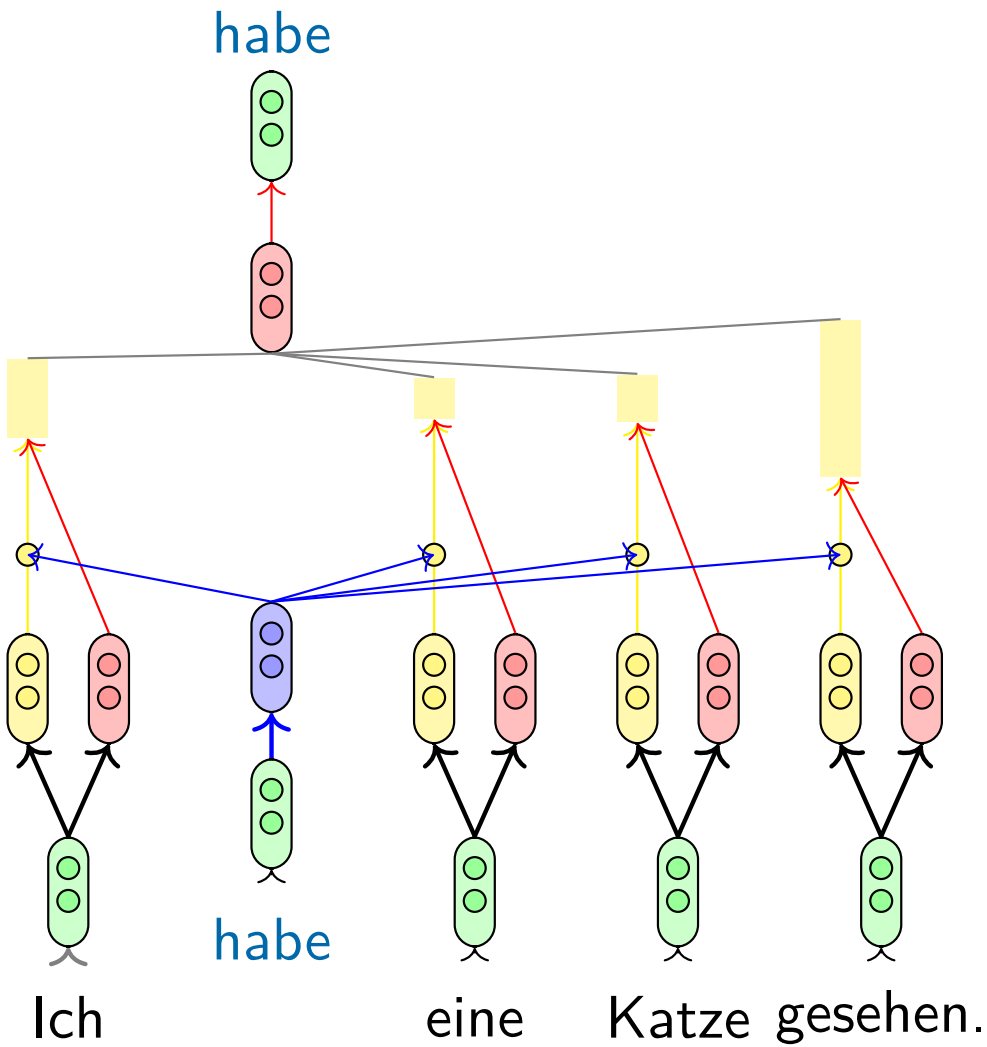


Image: after Lena Voita, *NLP Course For You*.

SE(3)-representation of a trajectory

Given a chemical specification of a molecular system with N atoms. We need to learn a generative model over time-series $[X_1, \ldots, X_T]$ of corresponding molecular structures $X_t \in \mathbb{R}^{3N}$ for a trajectory length T. Consider the following representation

$$\chi_t^j = \left((R_t^j,\ t_t^j),\ (\psi_t^j, \phi_t^j, \omega_t^j, \chi_{t,1}^j, \ldots, \chi_{t,4}^j)\right) \quad \text{with} \quad (R_t^j, t_t^j) \in SE(3), \quad (\psi, \phi, \omega, \chi) \in \mathbb{T}^7.$$

We assume that some frames $t_1, \ldots, t_K$ are known exactly. For every other frame t we model only *offsets* relative to them.

$$g_t^j \in SE(3) \qquad \Rightarrow \qquad [g_{t_i}^j]^{-1} g_t^j$$

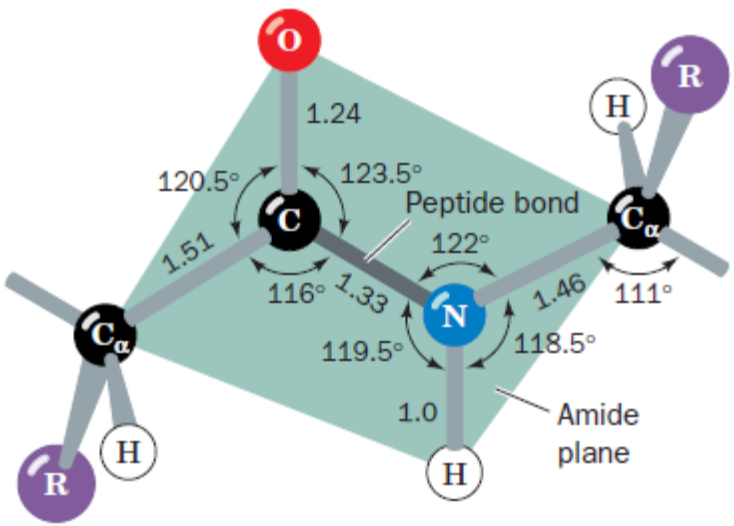


Image: Fig. 6.2 from D. Voet *et al.*, *Fundamentals of Biochemistry: Life at the Molecular Level*, 4th ed., John Wiley & Sons, Inc., 2013.

Numerical embedding

Each relative transform

$$[g_{t_i}^j]^{-1} g_t^j \in SE(3)$$

is encoded as:

$$(q_{t,i}^j,\ r_{t,i}^j) \in \mathbb{Q}^+ \times \mathbb{R}^3$$

- q — unit quaternion (rotation)
- r — translation

Torsion angles:

$$\theta \mapsto (\cos\theta, \sin\theta) \in S^1$$

Each key frame contributes:

$$4 \text{ (quaternion)} + 3 \text{ (translation)} = 7$$

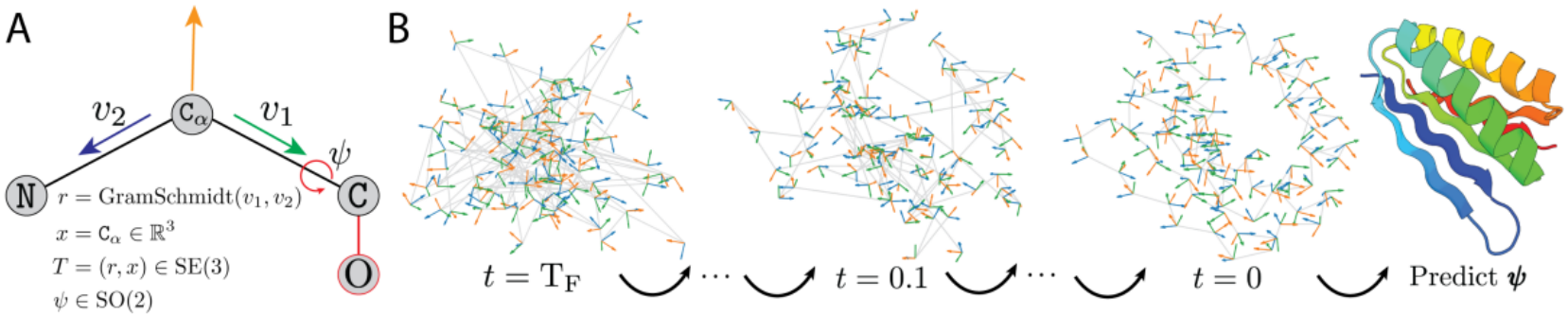
$$K \text{ key frames} \Rightarrow 7K$$

Torsions:

$$7 \text{ angles } \Rightarrow 14 \text{ numbers}$$

$$x_t^j \in \mathbb{R}^{7K+14}$$

Diffusion Model



Let consider the following process $x_t = \alpha_t x_* + \sigma_t \varepsilon$ with $\alpha_0 = \sigma_1 = 1$ and $\alpha_1 = \sigma_0 = 0$. Compute the velocity

$$v(x, t) = \mathbb{E}[\dot{x}_t \mid x_t = x] = \dot{\alpha}_t \, \mathbb{E}[x_* \mid x_t = x] + \dot{\sigma}_t \, \mathbb{E}[\varepsilon \mid x_t = x] \, .$$

Backward FPE gives us that time-dependent probability $\rho_t(x)$ of x_t coincides with the distribution of the reverse-time SDE

$$\mathrm{d}X_t = v(X_t, t) \, \mathrm{d}t - \frac{1}{2} \, w_t \, s(X_t, t) \, \mathrm{d}t + \sqrt{w_t} \, \mathrm{d}\bar{W}_t,$$

where $s(x, t) = -\sigma_t^{-1} \, \mathbb{E}[\varepsilon \mid X_t = x] \, .$
To learn Diffusion Model means to minimize the following loss-functions:

$$\mathcal{L}_s(\theta) = \int_0^T \mathbb{E}\Big[\big\| \sigma_t \, s_\theta(x_t, t) + \varepsilon \big\|^2\Big] \, \mathrm{d}t, \quad \mathcal{L}_v(\theta) = \int_0^T \mathbb{E}\Big[\big\| v_\theta(x_t, t) - \dot{\alpha}_t \, x_* - \dot{\sigma}_t \, \varepsilon \big\|^2\Big] \, \mathrm{d}t.$$

Using the fact that

$$s(x, t) = \sigma_t^{-1} \, \frac{\alpha_t \, v(x, t) - \dot{\alpha}_t \, x}{\dot{\alpha}_t \, \sigma_t - \alpha_t \, \dot{\sigma}_t}$$

we can learn only the velocity $v_\theta(x_t, t)$.

Image: Fig. 1 from J. Yim *et al.*, *SE(3)-diffusion model with application to protein backbone generation*, arXiv:2302.02277v3, 2023.

Euler–Maruyama Sampler

Require: Velocity model $v_\theta(x, t, y)$, time grid $\{t_i\}_{i=0}^{N}$, terminal time T, noise schedule $\{\omega_t\}$, $\{\alpha_t\}$, $\{\sigma_t\}$

Sample $x_0 \sim \mathcal{N}(0, I)$ ▷ initial noise
Convert s_θ from v_θ
Set $\Delta t \leftarrow t_1 - t_0$
for $i = 0, \ldots, N-1$ **do**
    Sample $\varepsilon_i \sim \mathcal{N}(0, I)$
    $d\varepsilon_i \leftarrow \varepsilon_i \sqrt{\Delta t}$
    $d_i \leftarrow v_\theta(x_i, t_i, y) + \frac{1}{2}\omega_{t_i}\, s_\theta(x_i, t_i, y)$ ▷ drift
    $\tilde{x}_{i+1} \leftarrow x_i + \Delta t\, d_i$
    $x_{i+1} \leftarrow \tilde{x}_{i+1} + \sqrt{\omega_{t_i}}\, d\varepsilon_i$ ▷ diffusion
end for
$h \leftarrow T - t_N$ ▷ final step to T, $x_T = x_*$
$d \leftarrow v_\theta(x_N, t_N, y) + \frac{1}{2}\omega_{t_N}\, s_\theta(x_N, t_N, y)$
$x \leftarrow x_N + h\, d$ ▷ noiseless final step
return x

Algorithm 2 from N. Ma *et al.*, *SiT: Exploring Flow and Diffusion-based Generative Models with Scalable Interpolant Transformers*, arXiv:2401.08740v2, 2024.

Invariant Point Attention

def $\text{InvariantPointAttention}(\{\mathbf{s}_i\}, \{\mathbf{z}_{ij}\}, \{T_i\}, N_{\text{head}} = 12, c = 16, N_{\text{query points}} = 4, N_{\text{point values}} = 8)$:

1: $\mathbf{q}_i^h, \mathbf{k}_i^h, \mathbf{v}_i^h = \text{LinearNoBias}(\mathbf{s}_i)$ $\qquad \mathbf{q}_i^h, \mathbf{k}_i^h, \mathbf{v}_i^h \in \mathbb{R}^c,\ h \in \{1, \ldots, N_{\text{head}}\}$

2: $\vec{\mathbf{q}}_i^{hp}, \vec{\mathbf{k}}_i^{hp} = \text{LinearNoBias}(\mathbf{s}_i)$ $\qquad \vec{\mathbf{q}}_i^{hp}, \vec{\mathbf{k}}_i^{hp}, \in \mathbb{R}^3,\ p \in \{1, \ldots, N_{\text{query points}}\},$ units: nanometres

3: $\vec{\mathbf{v}}_i^{hp} = \text{LinearNoBias}(\mathbf{s}_i)$ $\qquad \vec{\mathbf{v}}_i^{hp} \in \mathbb{R}^3,\ p \in \{1, \ldots, N_{\text{point values}}\},$ units: nanometres

4: $b_{ij}^h = \text{LinearNoBias}(\mathbf{z}_{ij})$

5: $w_C = \sqrt{\frac{2}{9N_{\text{query points}}}},$

6: $w_L = \sqrt{\frac{1}{3}}$

7: $a_{ij}^h = \text{softmax}_j\left(w_L\left(\frac{1}{\sqrt{c}}\,\mathbf{q}_i^{h\top}\mathbf{k}_j^h \;+ b_{ij}^h \;- \frac{\gamma^h w_C}{2}\sum_p \left\| T_i \circ \vec{\mathbf{q}}_i^{hp} - T_j \circ \vec{\mathbf{k}}_j^{hp} \right\|^2\right)\right)$

8: $\tilde{\mathbf{o}}_i^h = \sum_j a_{ij}^h \mathbf{z}_{ij}$

9: $\mathbf{o}_i^h = \sum_j a_{ij}^h \mathbf{v}_j^h$

10: $\vec{\mathbf{o}}_i^{hp} = T_i^{-1} \circ \sum_j a_{ij}^h \left(T_j \circ \vec{\mathbf{v}}_j^{hp}\right)$

11: $\tilde{\mathbf{s}}_i = \text{Linear}\left(\text{concat}_{h,p}(\tilde{\mathbf{o}}_i^h, \mathbf{o}_i^h, \vec{\mathbf{o}}_i^{hp}, \left\|\vec{\mathbf{o}}_i^{hp}\right\|)\right)$

12: **return** $\{\tilde{\mathbf{s}}_i\}$

Algorithm 22 from Suppl. Material for J. Jumper *et al.*, *Highly accurate protein structure prediction with AlphaFold*, Nature, 2021.

Architecture of MDGen

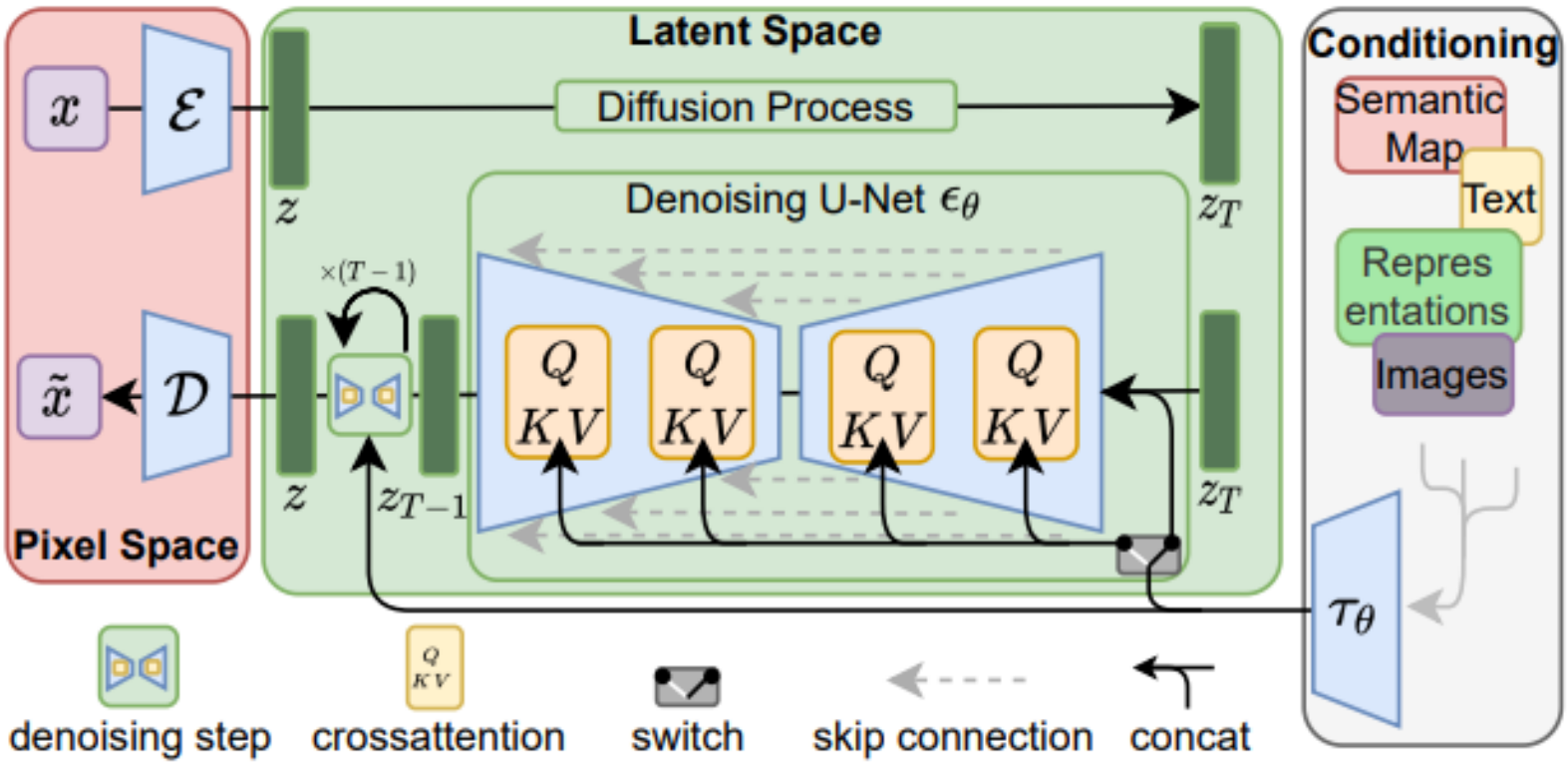
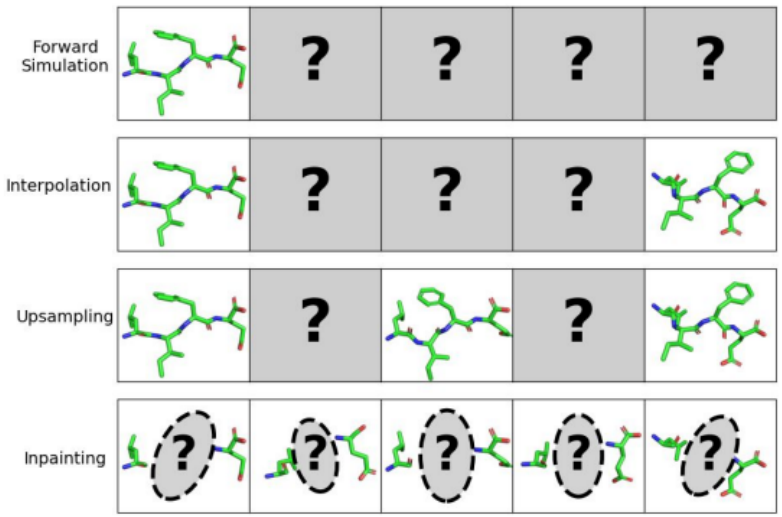
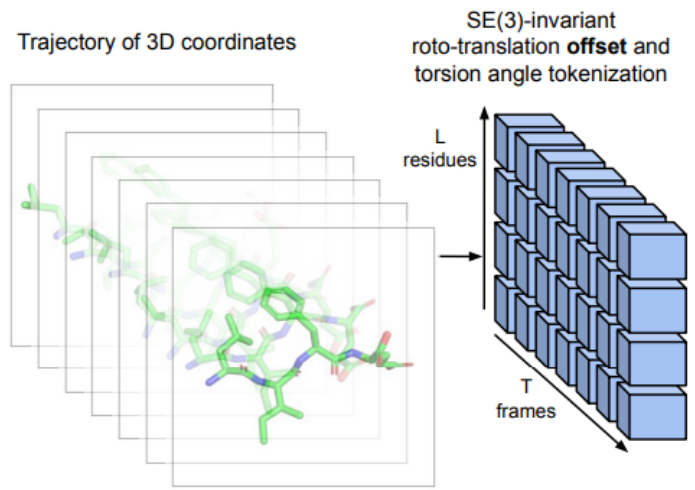


Image: Fig. 3 from R. Rombach, A. Blattmann, D. Lorenz, P. Esser, and B. Ommer, *High-Resolution Image Synthesis with Latent Diffusion Models*, arXiv:2112.10752, 2021.

What can it do?

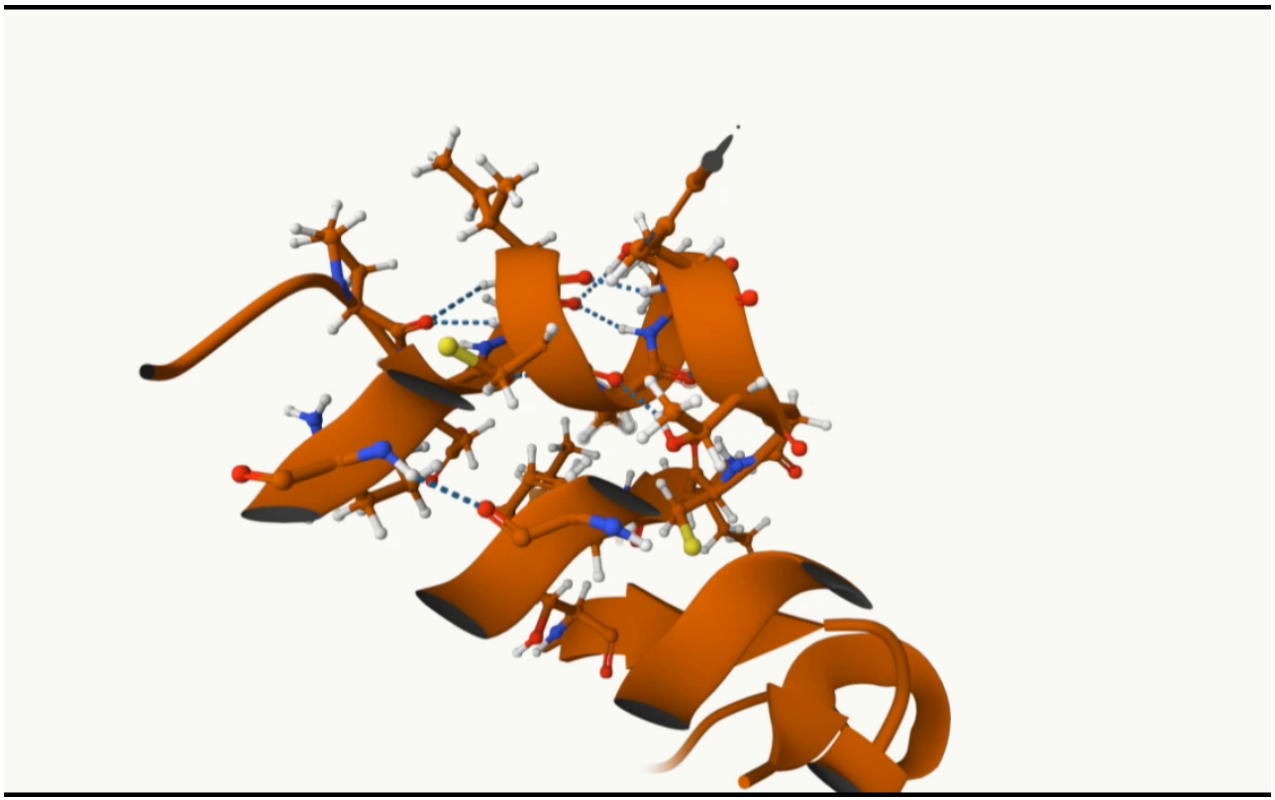




Setting	Key frames	Generate	Conditioned on	Token dim.
Forward simulation	g_1	$g_{1:T}$, $\tau_{1:T}$	g_1, τ_1, A	21
Interpolation	g_1, g_T	$g_{1:T}$, $\tau_{1:T}$	g_1, g_T, τ_1, τ_T, A	28
Upsampling	g_1	$g_{1:T}$, $\tau_{1:T}$	$g_{1+\{1,2,\dots\}M}$, $\tau_{1+\{1,2,\dots\}M}$, A	21
Inpainting	g_1, g_T	$g_{1:T}$, A	$g_{1:T}^{\text{known}}$	7 (+20)

Image: Fig. 1 from B. Jing, H. Stärk, T. Jaakkola, and B. Berger, *Generative Modeling of Molecular Dynamics Trajectories*, arXiv:2409.17808, 2024.

Forward Simulation for 1CRN



to be continued 😅

Thank you for your attention