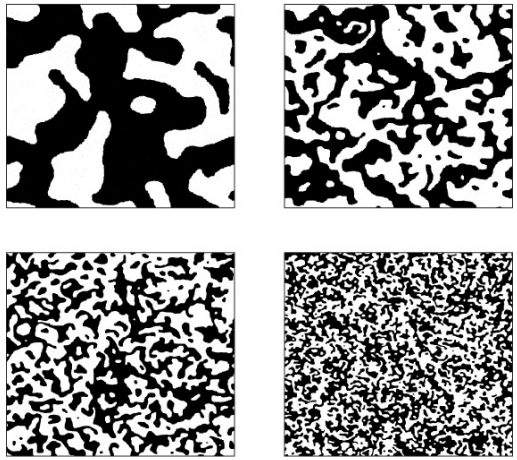
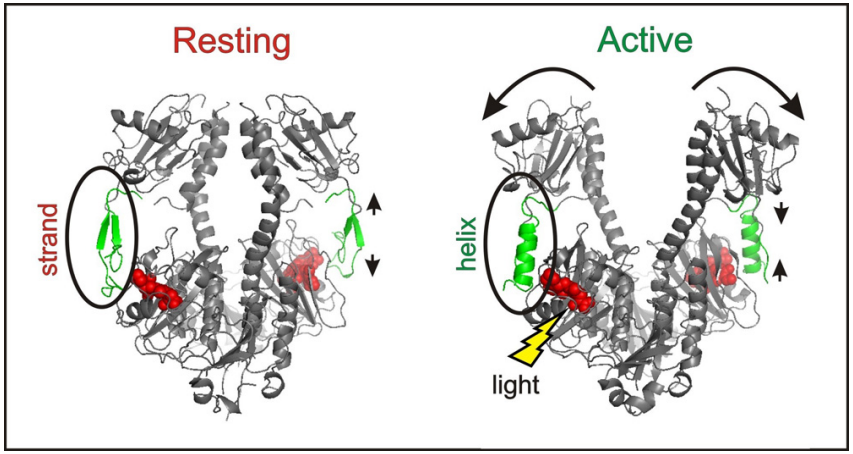
Boltzmann Generators

F. Noé[1]

Group Seminar, Nov 6, 2018

Limitations of Monte Carlo Sampling



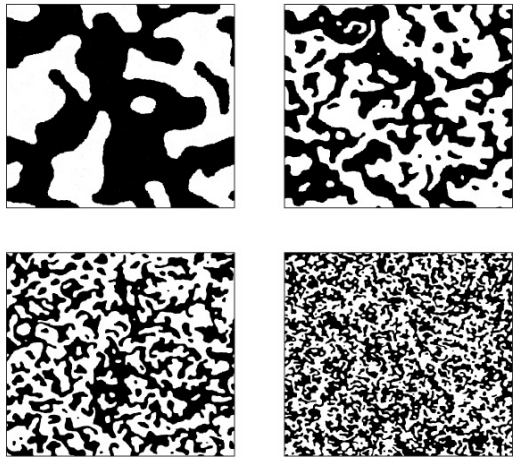


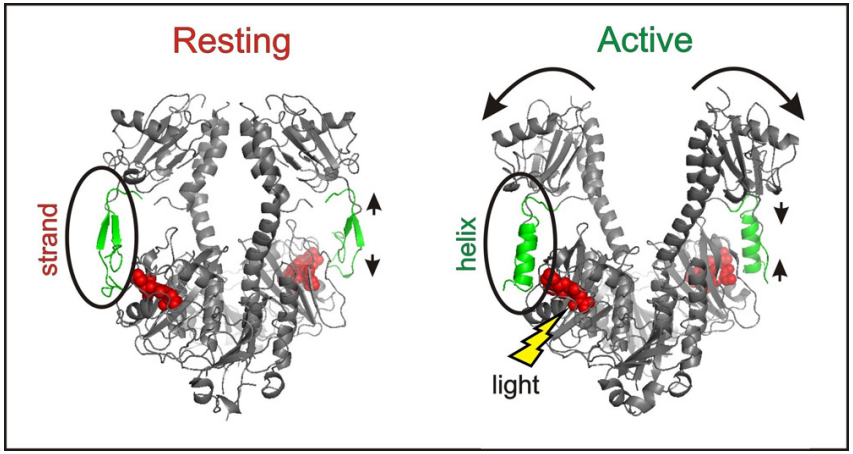
- **Input**: Reduced Potential Energy $u(\mathbf{x})$ in coordinates $\mathbf{x} \in \mathbb{R}^n$.
- **Aim**: Sample Equilibrium Distribution.

$$\mu(\mathbf{x}) \propto \mathrm{e}^{-u(\mathbf{x})}$$

- **Problem 1**: For increasing n, the subvolume of low-energy configurations is vanishingly small compared to $\mathbb{R}^n$ and has a complex shape.
  - Direct MD sampling (with rejection or reweighting) in configuration space is hopeless .
- **Problem 2**: Metastable states or phases
  - MCMC or MD methods with small steps converge very slowly
  - Guessing large MCMC proposal steps is hard and problem-specific.

Limitations of Monte Carlo Sampling



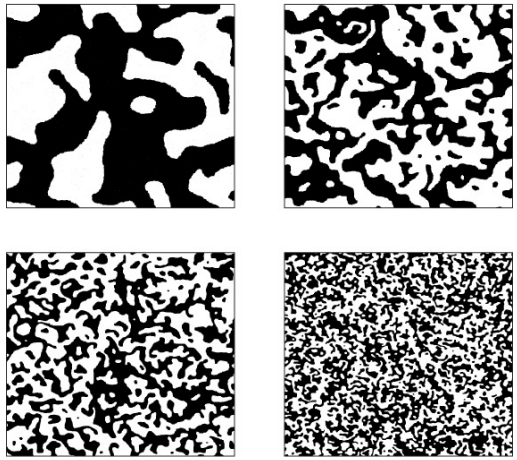


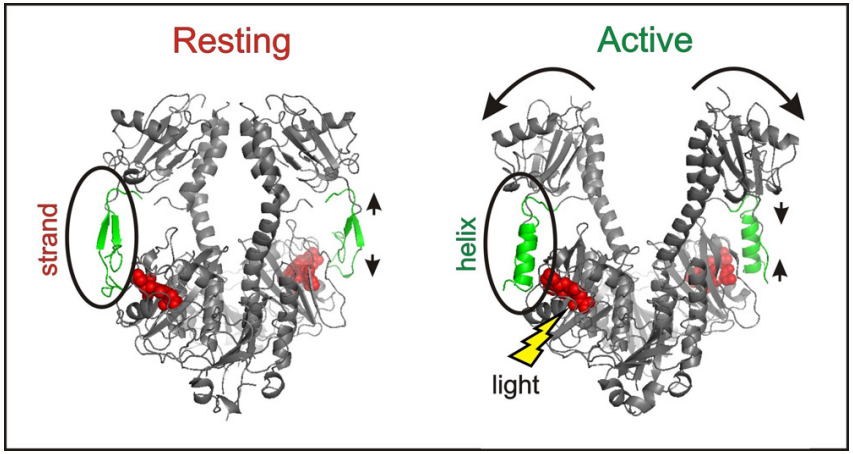
- **Input**: Reduced Potential Energy $u(\mathbf{x})$ in coordinates $\mathbf{x} \in \mathbb{R}^n$.
- **Aim**: Sample Equilibrium Distribution.

$$\mu(\mathbf{x}) \propto \mathrm{e}^{-u(\mathbf{x})}$$

- **Problem 1**: For increasing n, the subvolume of low-energy configurations is vanishingly small compared to $\mathbb{R}^n$ and has a complex shape.
  - Direct MD sampling (with rejection or reweighting) in configuration space is hopeless .
- **Problem 2**: Metastable states or phases
  - MCMC or MD methods with small steps converge very slowly
  - Guessing large MCMC proposal steps is hard and problem-specific.

Limitations of Monte Carlo Sampling



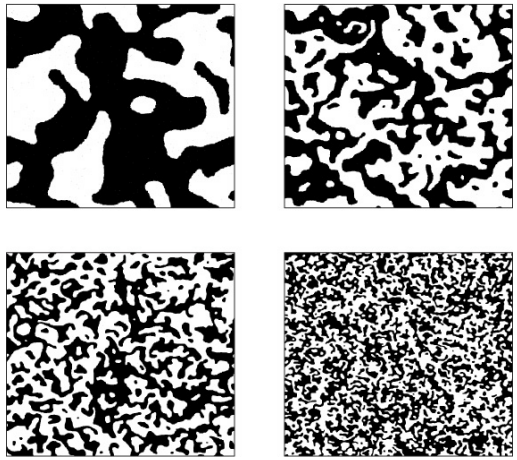


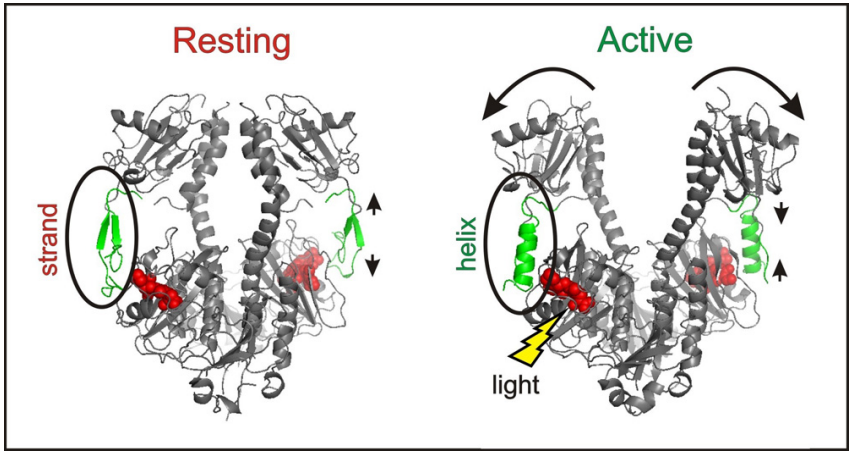
- **Input**: Reduced Potential Energy $u(\mathbf{x})$ in coordinates $\mathbf{x} \in \mathbb{R}^n$.
- **Aim**: Sample Equilibrium Distribution.

$$\mu(\mathbf{x}) \propto \mathrm{e}^{-u(\mathbf{x})}$$

- **Problem 1**: For increasing n, the subvolume of low-energy configurations is vanishingly small compared to $\mathbb{R}^n$ and has a complex shape.
  - Direct MD sampling (with rejection or reweighting) in configuration space is hopeless .
- **Problem 2**: Metastable states or phases
  - MCMC or MD methods with small steps converge very slowly
  - Guessing large MCMC proposal steps is hard and problem-specific.

Limitations of Monte Carlo Sampling



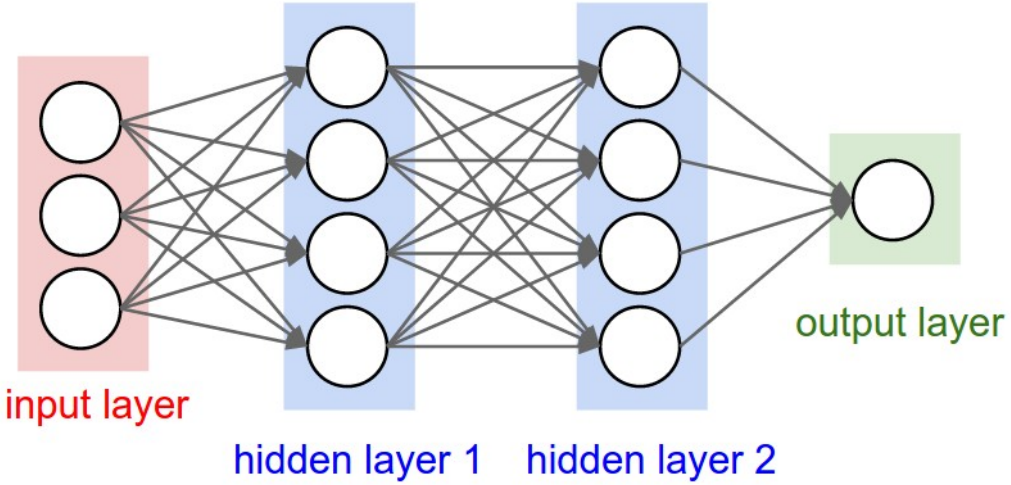


- **Input**: Reduced Potential Energy $u(\mathbf{x})$ in coordinates $\mathbf{x} \in \mathbb{R}^n$.
- **Aim**: Sample Equilibrium Distribution.

$$\mu(\mathbf{x}) \propto \mathrm{e}^{-u(\mathbf{x})}$$

- **Problem 1**: For increasing n, the subvolume of low-energy configurations is vanishingly small compared to $\mathbb{R}^n$ and has a complex shape.
  - Direct MD sampling (with rejection or reweighting) in configuration space is hopeless .
- **Problem 2**: Metastable states or phases
  - MCMC or MD methods with small steps converge very slowly
  - Guessing large MCMC proposal steps is hard and problem-specific.

Reminder: Multilayer Neural Networks

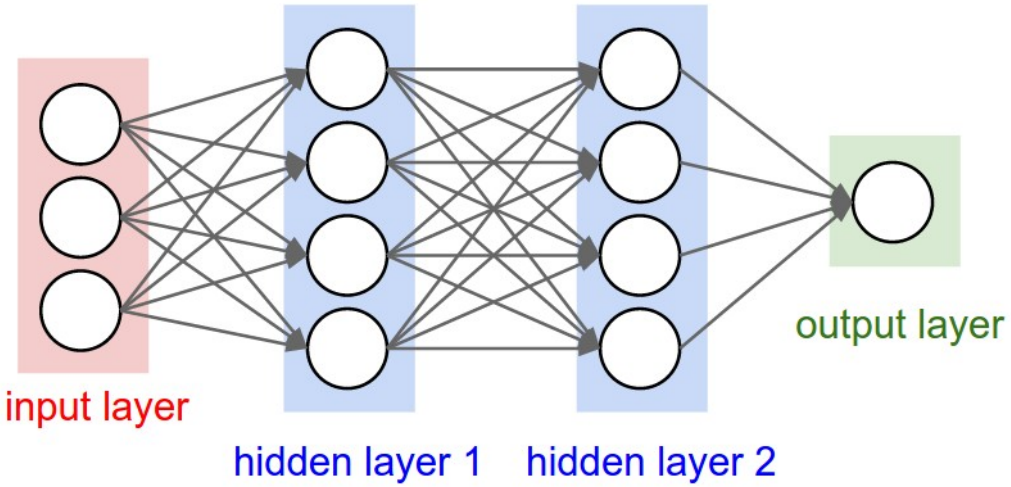


Sequence of linear and nonlinear transforms defined by the recursion:

$$\mathbf{x}^{l+1} = \sigma\left(\mathbf{W}^l \mathbf{x}^l + \mathbf{b}^l\right)$$

- L layers indexed by $l = 1, \ldots, L$. Input vector $\mathbf{x}^{(0)}$ does not count as a layer.
- $\mathbf{x}^l \in \mathbb{R}^{n_l}$: Activations of n_l neurons at layer l
- Trainable weights $\mathbf{W}^l \in \mathbb{R}^{n_{l-1} \times n_l}$ and biases $\mathbf{b}^l \in \mathbb{R}^{n_l}$ at each layer. W^l_{ij} is connecting neuron j of layer $l-1$ with neuron i of layer l.
- Nonlinear function $\sigma : \mathbb{R} \to \mathbb{R}$.
- Output vector: $\hat{\mathbf{y}} = \mathbf{x}^L$ ($\hat{\mathbf{y}}$: network predictions. $\mathbf{y}$: training values)

Reminder: Multilayer Neural Networks

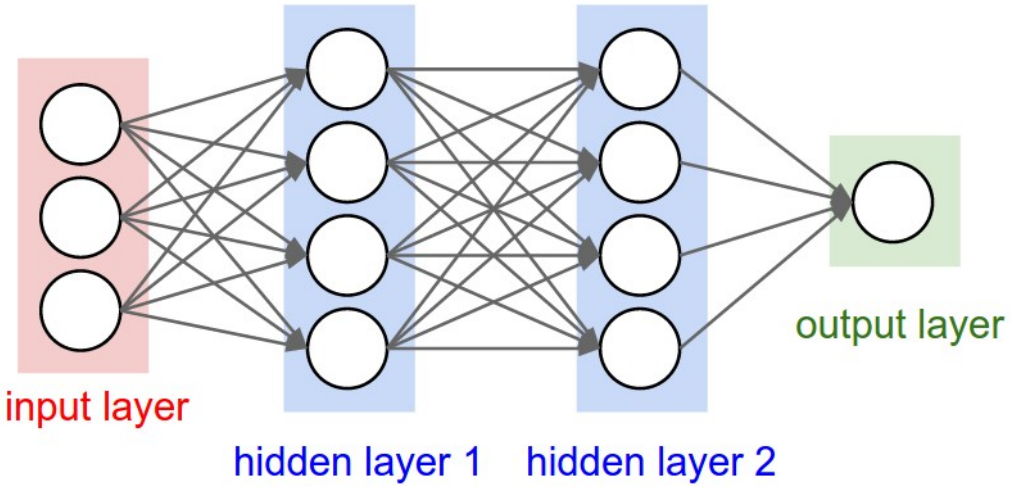


Sequence of linear and nonlinear transforms defined by the recursion:

$$\mathbf{x}^{l+1} = \sigma\left(\mathbf{W}^l \mathbf{x}^l + \mathbf{b}^l\right)$$

- L layers indexed by $l = 1, \ldots, L$. Input vector $\mathbf{x}^{(0)}$ does not count as a layer.
- $\mathbf{x}^l \in \mathbb{R}^{n_l}$: Activations of n_l neurons at layer l
- Trainable weights $\mathbf{W}^l \in \mathbb{R}^{n_{l-1} \times n_l}$ and biases $\mathbf{b}^l \in \mathbb{R}^{n_l}$ at each layer. W^l_{ij} is connecting neuron j of layer $l-1$ with neuron i of layer l.
- Nonlinear function $\sigma : \mathbb{R} \to \mathbb{R}$.
- Output vector: $\hat{\mathbf{y}} = \mathbf{x}^l$ ($\hat{\mathbf{y}}$: network predictions. $\mathbf{y}$: training values)

Reminder: Multilayer Neural Networks

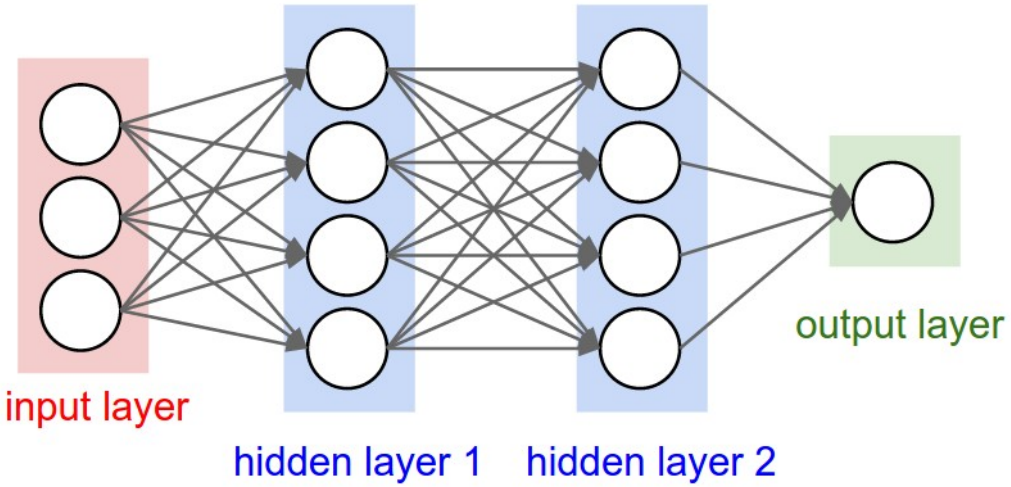


Sequence of linear and nonlinear transforms defined by the recursion:

$$\mathbf{x}^{l+1} = \sigma\left(\mathbf{W}^l \mathbf{x}^l + \mathbf{b}^l\right)$$

- L layers indexed by $l = 1, \ldots, L$. Input vector $\mathbf{x}^{(0)}$ does not count as a layer.
- $\mathbf{x}^l \in \mathbb{R}^{n_l}$: Activations of n_l neurons at layer l
- Trainable weights $\mathbf{W}^l \in \mathbb{R}^{n_{l-1} \times n_l}$ and biases $\mathbf{b}^l \in \mathbb{R}^{n_l}$ at each layer. W^l_{ij} is connecting neuron j of layer $l-1$ with neuron i of layer l.
- Nonlinear function $\sigma : \mathbb{R} \to \mathbb{R}$.
- Output vector: $\hat{\mathbf{y}} = \mathbf{x}^l$ ($\hat{\mathbf{y}}$: network predictions. $\mathbf{y}$: training values)

Reminder: Multilayer Neural Networks

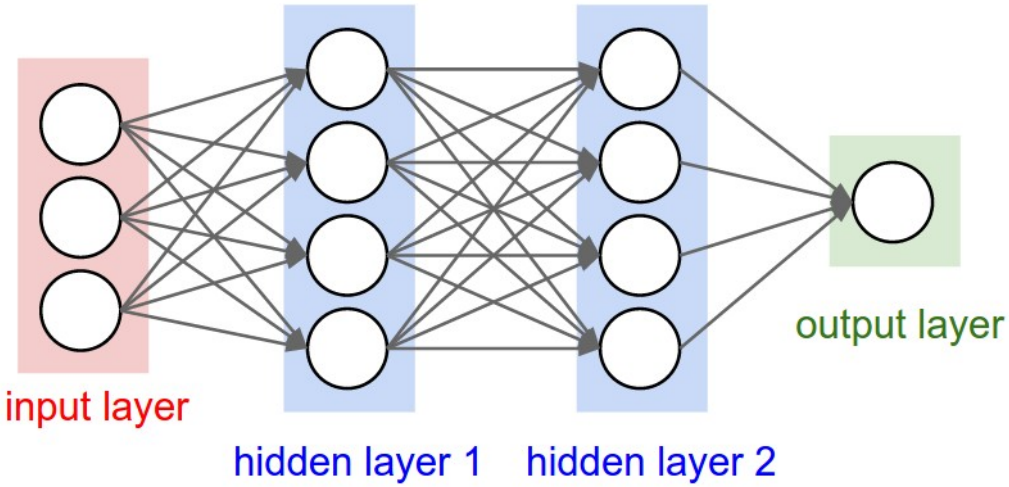


Sequence of linear and nonlinear transforms defined by the recursion:

$$\mathbf{x}^{l+1} = \sigma\left(\mathbf{W}^l \mathbf{x}^l + \mathbf{b}^l\right)$$

- L layers indexed by $l = 1, \ldots, L$. Input vector $\mathbf{x}^{(0)}$ does not count as a layer.
- $\mathbf{x}^l \in \mathbb{R}^{n_l}$: Activations of n_l neurons at layer l
- Trainable weights $\mathbf{W}^l \in \mathbb{R}^{n_{l-1} \times n_l}$ and biases $\mathbf{b}^l \in \mathbb{R}^{n_l}$ at each layer. W^l_{ij} is connecting neuron j of layer $l-1$ with neuron i of layer l.
- Nonlinear function $\sigma : \mathbb{R} \to \mathbb{R}$.
- Output vector: $\hat{\mathbf{y}} = \mathbf{x}^l$ ($\hat{\mathbf{y}}$: network predictions. $\mathbf{y}$: training values)

Reminder: Multilayer Neural Networks



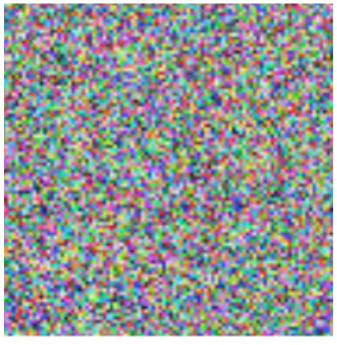
Sequence of linear and nonlinear transforms defined by the recursion:

$$\mathbf{x}^{l+1} = \sigma\left(\mathbf{W}^l \mathbf{x}^l + \mathbf{b}^l\right)$$

- L layers indexed by $l = 1, \ldots, L$. Input vector $\mathbf{x}^{(0)}$ does not count as a layer.
- $\mathbf{x}^l \in \mathbb{R}^{n_l}$: Activations of n_l neurons at layer l
- Trainable weights $\mathbf{W}^l \in \mathbb{R}^{n_{l-1} \times n_l}$ and biases $\mathbf{b}^l \in \mathbb{R}^{n_l}$ at each layer. W^l_{ij} is connecting neuron j of layer $l-1$ with neuron i of layer l.
- Nonlinear function $\sigma : \mathbb{R} \to \mathbb{R}$.
- Output vector: $\hat{\mathbf{y}} = \mathbf{x}^l$ ($\hat{\mathbf{y}}$: network predictions. $\mathbf{y}$: training values)

Reminder: Generative Neural Networks

Noise ~ N(0,1)



Generative Model

Reminder: Generative Neural Networks

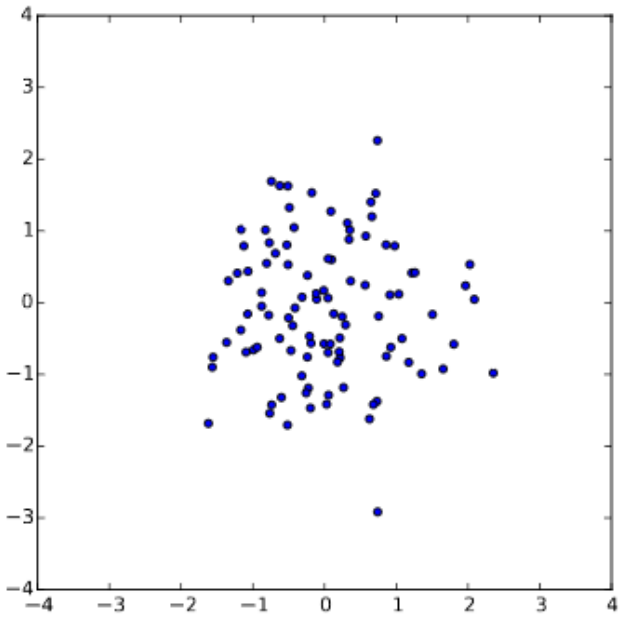
- **Idea**: Learn to sample intractable $p(\mathbf{x})$ by sampling tractable latent distribution

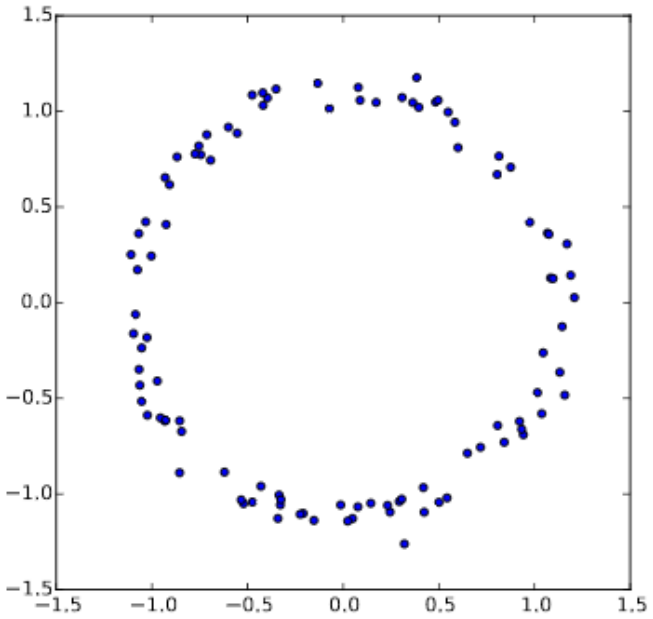
$$\mathbf{z} \sim p(\mathbf{z})$$

and perform a linear transformation to a desired distribution:

$$\mathbf{x} = G(\mathbf{z}, \theta) \sim p(\mathbf{x}).$$

- **Example**:
  - **Left**: Samples from normal distribution, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
  - **Right**: Samples mapped through $G(\mathbf{z}) = \frac{\mathbf{z}}{10} + \frac{\mathbf{z}}{\|\mathbf{z}\|}$ to form a ring.





Reminder: Generative Neural Networks

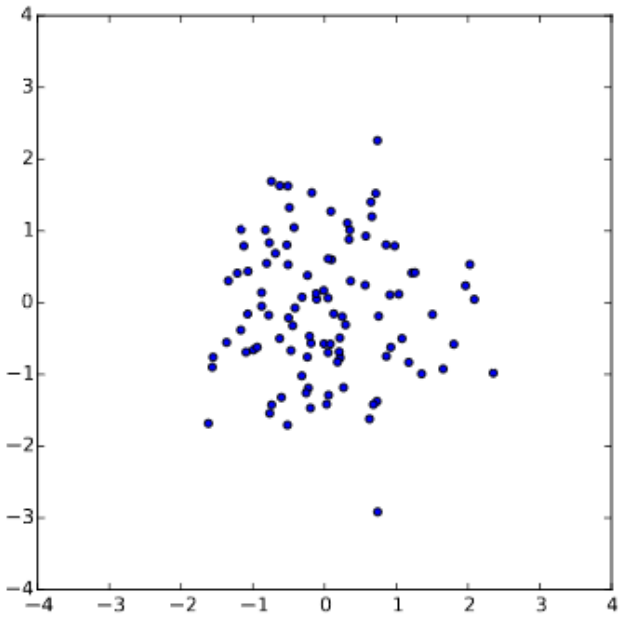
- **Idea**: Learn to sample intractable $p(\mathbf{x})$ by sampling tractable latent distribution

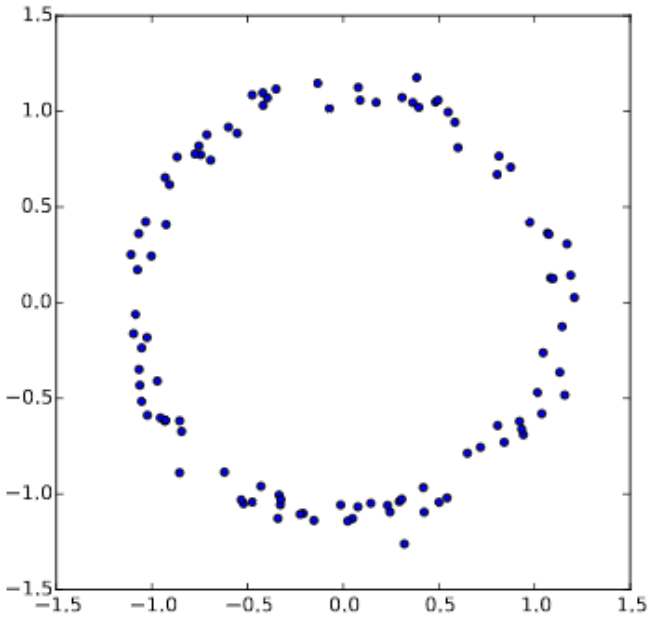
$$\mathbf{z} \sim p(\mathbf{z})$$

and perform a linear transformation to a desired distribution:

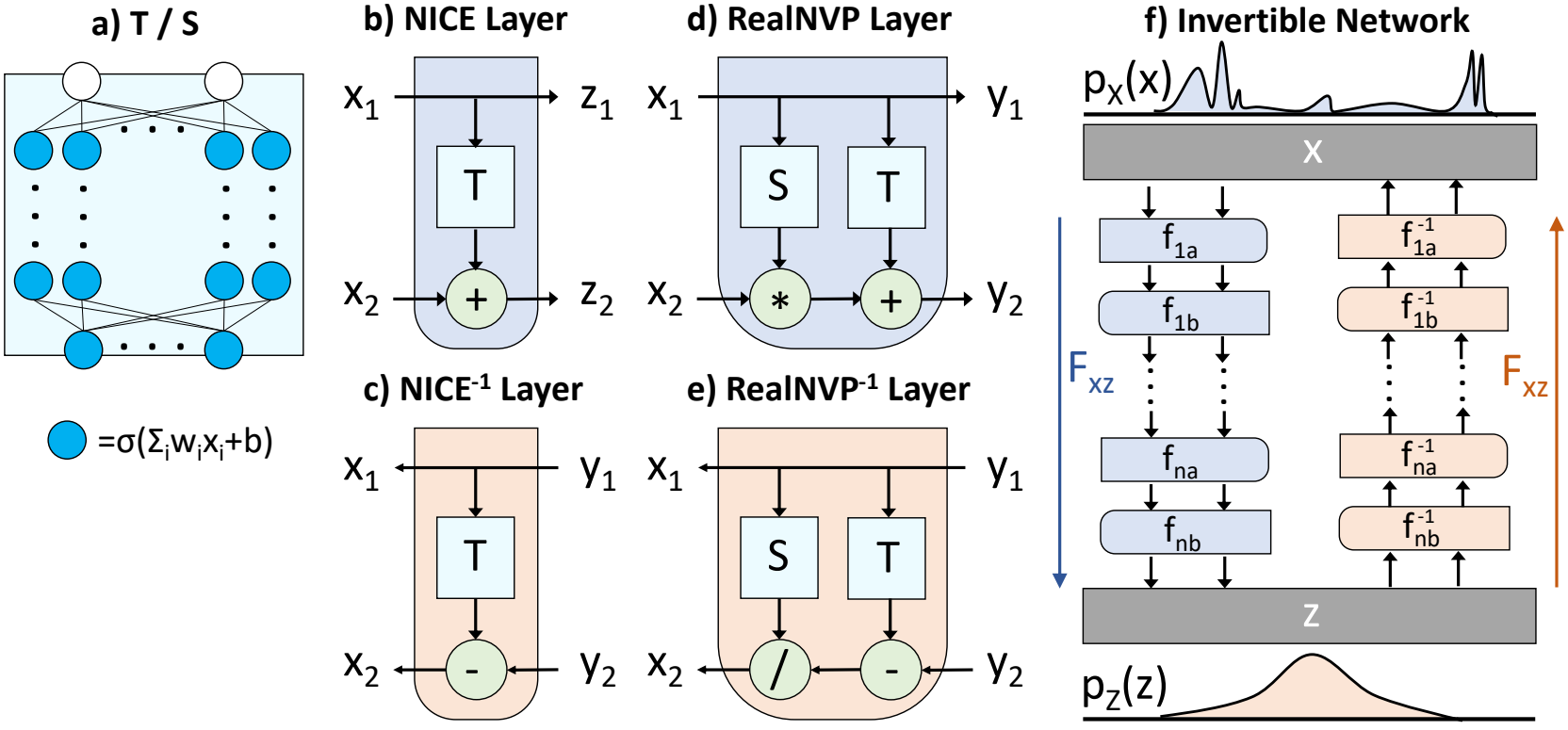
$$\mathbf{x} = G(\mathbf{z}, \theta) \sim p(\mathbf{x}).$$

- **Example**:
  - **Left**: Samples from normal distribution, $\mathbf{z} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$.
  - **Right**: Samples mapped through $G(\mathbf{z}) = \frac{\mathbf{z}}{10} + \frac{\mathbf{z}}{\|\mathbf{z}\|}$ to form a ring.





Boltzmann Generators



- **NICE**: Dinh, Krueger, Y. Bengio, ICLR 2015
- **RealNVP**: Dinh, Sohl-Dickstein, S. Bengio, ICLR 2017

Transformation of random variables

- Invertible transformation:

$$\mathbf{z} = F_{xz}(\mathbf{x}; \theta)$$
$$\mathbf{x} = F_{zx}(\mathbf{z}; \theta).$$

- Jacobians:

$$\mathbf{J}_{zx}(\mathbf{z}; \theta) = \left[\frac{\partial F_{zx}(\mathbf{z}; \theta)}{\partial z_1}, \ldots, \frac{\partial F_{zx}(\mathbf{z}; \theta)}{\partial z_n} \right]$$
$$\mathbf{J}_{xz}(\mathbf{x}; \theta) = \left[\frac{dF_{xz}(\mathbf{x}; \theta)}{dx_1}, \ldots, \frac{dF_{xz}(\mathbf{x}; \theta)}{dx_n} \right]$$

- Transformation of random variables:

$$\begin{aligned} p_X(\mathbf{x}) &= p_Z(\mathbf{z}) \left| \det \mathbf{J}_{zx}(\mathbf{z}) \right|^{-1} \\ &= p_Z(T_{xz}(\mathbf{x})) \left| \det \mathbf{J}_{xz}(\mathbf{x}) \right| \qquad (1) \\ p_Z(\mathbf{z}) &= p_X(\mathbf{x}) \left| \det \mathbf{J}_{xz}(\mathbf{x}) \right|^{-1} \\ &= p_X(T_{zx}(\mathbf{z})) \left| \det \mathbf{J}_{zx}(\mathbf{z}) \right| \qquad (2) \end{aligned}$$

Transformation of random variables

- Invertible transformation:

$$\mathbf{z} = F_{xz}(\mathbf{x}; \theta)$$
$$\mathbf{x} = F_{zx}(\mathbf{z}; \theta).$$

- Jacobians:

$$\mathbf{J}_{zx}(\mathbf{z}; \theta) = \left[\frac{\partial F_{zx}(\mathbf{z}; \theta)}{\partial z_1}, \ldots, \frac{\partial F_{zx}(\mathbf{z}; \theta)}{\partial z_n} \right]$$
$$\mathbf{J}_{xz}(\mathbf{x}; \theta) = \left[\frac{dF_{xz}(\mathbf{x}; \theta)}{dx_1}, \ldots, \frac{dF_{xz}(\mathbf{x}; \theta)}{dx_n} \right]$$

- Transformation of random variables:

$$\begin{aligned} p_X(\mathbf{x}) &= p_Z(\mathbf{z}) \left| \det \mathbf{J}_{zx}(\mathbf{z}) \right|^{-1} \\ &= p_Z(T_{xz}(\mathbf{x})) \left| \det \mathbf{J}_{xz}(\mathbf{x}) \right| \end{aligned} \tag{1}$$

$$\begin{aligned} p_Z(\mathbf{z}) &= p_X(\mathbf{x}) \left| \det \mathbf{J}_{xz}(\mathbf{x}) \right|^{-1} \\ &= p_X(T_{zx}(\mathbf{z})) \left| \det \mathbf{J}_{zx}(\mathbf{z}) \right| \end{aligned} \tag{2}$$

Transformation of random variables

- Invertible transformation:

$$\mathbf{z} = F_{xz}(\mathbf{x}; \theta)$$
$$\mathbf{x} = F_{zx}(\mathbf{z}; \theta).$$

- Jacobians:

$$\mathbf{J}_{zx}(\mathbf{z}; \theta) = \left[\frac{\partial F_{zx}(\mathbf{z}; \theta)}{\partial z_1}, \ldots, \frac{\partial F_{zx}(\mathbf{z}; \theta)}{\partial z_n} \right]$$
$$\mathbf{J}_{xz}(\mathbf{x}; \theta) = \left[\frac{dF_{xz}(\mathbf{x}; \theta)}{dx_1}, \ldots, \frac{dF_{xz}(\mathbf{x}; \theta)}{dx_n} \right]$$

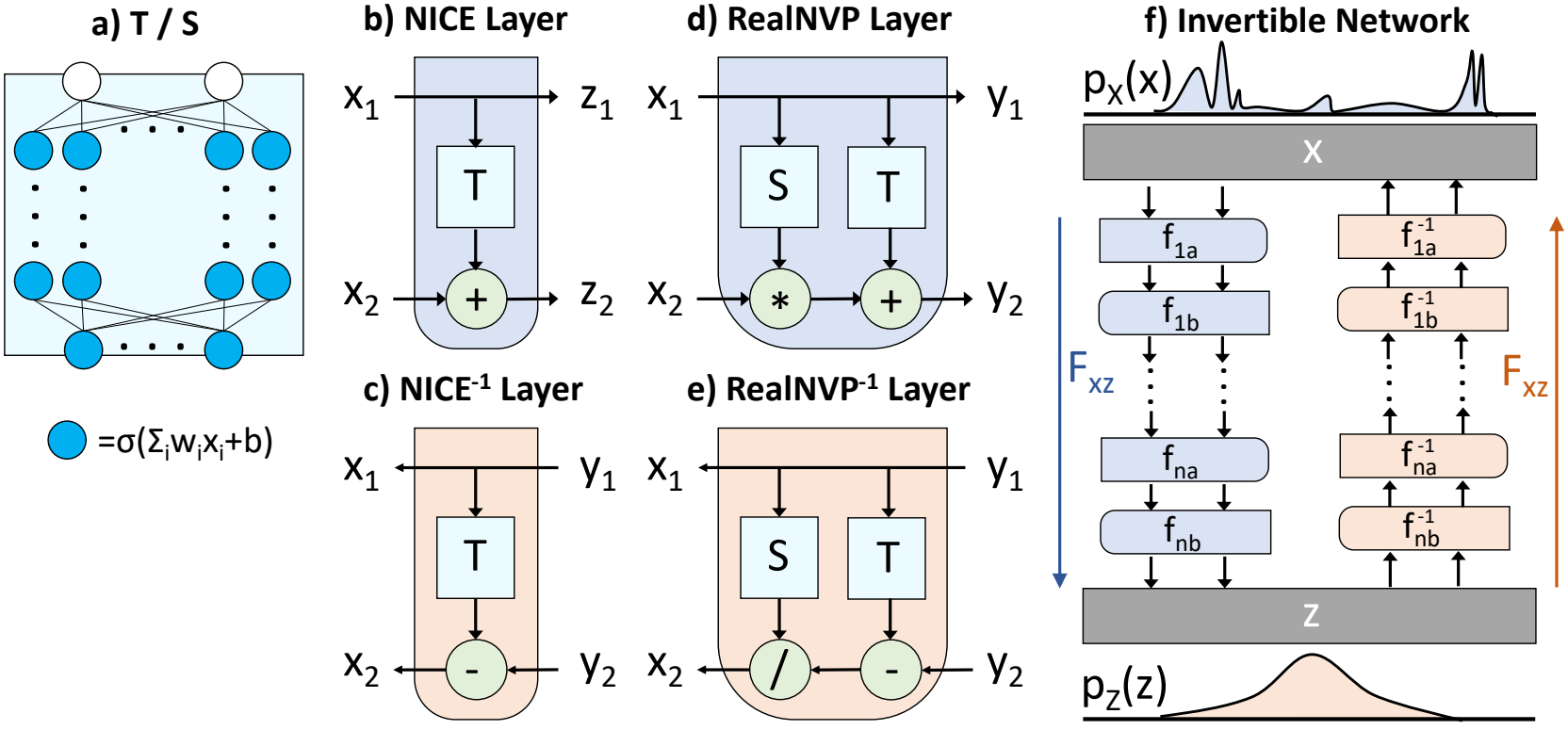
- Transformation of random variables:

$$\begin{aligned} p_X(\mathbf{x}) &= p_Z(\mathbf{z}) \left| \det \mathbf{J}_{zx}(\mathbf{z}) \right|^{-1} \\ &= p_Z(T_{xz}(\mathbf{x})) \left| \det \mathbf{J}_{xz}(\mathbf{x}) \right| \qquad (1) \\ p_Z(\mathbf{z}) &= p_X(\mathbf{x}) \left| \det \mathbf{J}_{xz}(\mathbf{x}) \right|^{-1} \\ &= p_X(T_{zx}(\mathbf{z})) \left| \det \mathbf{J}_{zx}(\mathbf{z}) \right| \qquad (2) \end{aligned}$$

Invertible network components

Layer	f_{xz}	$\lvert\det \mathbf{J}_{xz}\rvert$	f_{zx}	$\lvert\det \mathbf{J}_{zx}\rvert$
NICE	$\mathbf{z}_1 = \mathbf{x}_1$ $\mathbf{z}_2 = \mathbf{x}_2 + T(\mathbf{x}_1;\theta)$	1	$\mathbf{x}_1 = \mathbf{z}_1$ $\mathbf{x}_2 = \mathbf{z}_2 - T(\mathbf{y}_1;\theta)$	1
Scaling, Exp	$\mathbf{z} = \mathrm{e}^{\mathbf{k}} \circ \mathbf{x}$	$\mathrm{e}^{\sum_i k_i}$	$\mathbf{x} = \mathrm{e}^{-\mathbf{k}} \circ \mathbf{z}$	$\mathrm{e}^{-\sum_i k_i}$
RealNVP	$\mathbf{z}_1 = \mathbf{x}_1$ $\mathbf{z}_2 = \mathbf{x}_2 \odot \exp(S(\mathbf{x}_1;\theta)) + T(\mathbf{x}_1;\theta)$	$\mathrm{e}^{\sum_i S_i(\mathbf{x}_1;\theta)}$	$\mathbf{x}_1 = \mathbf{z}_1$ $\mathbf{x}_2 = (\mathbf{z}_2 - T(\mathbf{x}_1;\theta)) \odot \exp(-S(\mathbf{z}_1;\theta))$	$\mathrm{e}^{-\sum_i S_i(\mathbf{z}_1;\theta)}$

Boltzmann Generators



- **NICE**: Dinh, Krueger, Y. Bengio, ICLR 2015
- **RealNVP**: Dinh, Sohl-Dickstein, S. Bengio, ICLR 2017

Transformation of random variables

- Distributions:

$$\text{Prior} \quad q_Z(\mathbf{z}) \xrightarrow{F_{zx}} p_X(\mathbf{x}) \quad \text{generated}$$

$$\text{Boltzmann} \quad \mu_X(\mathbf{x}) \xrightarrow{F_{xz}} p_Z(\mathbf{z}) \quad \text{generated}$$

- Aim: sample configurations $\mathbf{x}$ from **Boltzmann distribution**

$$\mu_X(\mathbf{x}) = Z_X^{-1} \mathrm{e}^{-u(\mathbf{x})} \tag{3}$$

- Reduced energy at temperature T:

$$u(\mathbf{x}) = \frac{U(\mathbf{x})}{k_B T}$$

- **Prior distribution**: Sample input in $\mathbf{z}$ from isotropic Gaussian:

$$q_Z^k(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}) = Z_Z^{-1} \mathrm{e}^{-\frac{1}{2}\|\mathbf{z}\|^2/\sigma_k^2}, \tag{4}$$

  Prior energy:

$$u_Z^k(\mathbf{z}) = -\log q_Z^k(\mathbf{z}) = \frac{1}{2\sigma_k^2} \|\mathbf{z}\|^2 + \text{const}$$

Transformation of random variables

- Distributions:

$$\text{Prior} \quad q_Z(\mathbf{z}) \xrightarrow{F_{zx}} p_X(\mathbf{x}) \quad \text{generated}$$

$$\text{Boltzmann} \quad \mu_X(\mathbf{x}) \xrightarrow{F_{xz}} p_Z(\mathbf{z}) \quad \text{generated}$$

- Aim: sample configurations $\mathbf{x}$ from **Boltzmann distribution**

$$\mu_X(\mathbf{x}) = Z_X^{-1} \mathrm{e}^{-u(\mathbf{x})} \tag{3}$$

- Reduced energy at temperature T:

$$u(\mathbf{x}) = \frac{U(\mathbf{x})}{k_B T}$$

- **Prior distribution**: Sample input in $\mathbf{z}$ from isotropic Gaussian:

$$q_Z^k(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}) = Z_Z^{-1} \mathrm{e}^{-\frac{1}{2}\|\mathbf{z}\|^2/\sigma_k^2}, \tag{4}$$

  Prior energy:

$$u_Z^k(\mathbf{z}) = -\log q_Z^k(\mathbf{z}) = \frac{1}{2\sigma_k^2} \|\mathbf{z}\|^2 + \text{const}$$

Transformation of random variables

- Distributions:

$$\text{Prior} \quad q_Z(\mathbf{z}) \xrightarrow{F_{zx}} p_X(\mathbf{x}) \quad \text{generated}$$

$$\text{Boltzmann} \quad \mu_X(\mathbf{x}) \xrightarrow{F_{xz}} p_Z(\mathbf{z}) \quad \text{generated}$$

- Aim: sample configurations $\mathbf{x}$ from **Boltzmann distribution**

$$\mu_X(\mathbf{x}) = Z_X^{-1} \mathrm{e}^{-u(\mathbf{x})} \tag{3}$$

- Reduced energy at temperature T:

$$u(\mathbf{x}) = \frac{U(\mathbf{x})}{k_B T}$$

- **Prior distribution**: Sample input in $\mathbf{z}$ from isotropic Gaussian:

$$q_Z^k(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}) = Z_Z^{-1} \mathrm{e}^{-\frac{1}{2}\|\mathbf{z}\|^2/\sigma_k^2}, \tag{4}$$

  Prior energy:

$$u_Z^k(\mathbf{z}) = -\log q_Z^k(\mathbf{z}) = \frac{1}{2\sigma_k^2} \|\mathbf{z}\|^2 + \text{const}$$

Transformation of random variables

- Distributions:

$$\text{Prior} \quad q_Z(\mathbf{z}) \xrightarrow{F_{zx}} p_X(\mathbf{x}) \quad \text{generated}$$

$$\text{Boltzmann} \quad \mu_X(\mathbf{x}) \xrightarrow{F_{xz}} p_Z(\mathbf{z}) \quad \text{generated}$$

- Aim: sample configurations $\mathbf{x}$ from **Boltzmann distribution**

$$\mu_X(\mathbf{x}) = Z_X^{-1} \mathrm{e}^{-u(\mathbf{x})} \tag{3}$$

- Reduced energy at temperature T:

$$u(\mathbf{x}) = \frac{U(\mathbf{x})}{k_B T}$$

- **Prior distribution**: Sample input in $\mathbf{z}$ from isotropic Gaussian:

$$q_Z^k(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}) = Z_Z^{-1} \mathrm{e}^{-\frac{1}{2}\|\mathbf{z}\|^2/\sigma_k^2}, \tag{4}$$

  Prior energy:

$$u_Z^k(\mathbf{z}) = -\log q_Z^k(\mathbf{z}) = \frac{1}{2\sigma_k^2} \|\mathbf{z}\|^2 + \text{const}$$

Latent KL divergence

- Loss function:

$$J = \underbrace{w_{ML} J_{ML}}_{\text{max likelihood}} + \underbrace{w_{KL} J_{KL}}_{\text{Kullback}-\text{Leibler}} + \underbrace{w_{RC} J_{RC}}_{\text{reaction coordinate}}.$$

- KL divergence between q and p, entropy H_q:

$$\begin{aligned}\mathrm{KL}(q \parallel p) &= \int q(\mathbf{x}) \left[\log q(\mathbf{x}) - \log p(\mathbf{x})\right] \mathrm{d}\mathbf{x}, \\ &= H_q - \int q(\mathbf{x}) \log p(\mathbf{x}) \mathrm{d}\mathbf{x},\end{aligned}$$

- KL divergence between generated $p_X(\mathbf{x})$ and Boltzmann distribution:

$$\begin{aligned}\mathrm{KL}_\theta \left[q_Z \parallel p_Z\right] &= H_Z - \int q_Z(\mathbf{z}) \log p_Z(\mathbf{z}; \theta) \mathrm{d}\mathbf{z}, \\ &= H_Z - \int q_Z(\mathbf{z}) \left[\log \mu_X(F_{zx}(\mathbf{z}; \theta)) + \log \left|\mathbf{J}_{zx}(\mathbf{z}; \theta)\right|\right] \mathrm{d}\mathbf{z}, \\ &= \underbrace{H_Z + \log Z_X}_{\text{const}} + \mathbb{E}_{\mathbf{z} \sim q_Z(\mathbf{z})} \left[u(F_{zx}(\mathbf{z}; \theta)) - \log \left|\mathbf{J}_{zx}(\mathbf{z}; \theta)\right|\right]\end{aligned}$$

- KL loss:

$$J_{KL} = \mathbb{E}_{\mathbf{z} \sim q_Z(\mathbf{z})} \left[u(F_{zx}(\mathbf{z}; \theta)) - \log \left|\mathbf{J}_{zx}(\mathbf{z}; \theta)\right|\right]. \quad (5)$$

Latent KL divergence

- Loss function:

$$J = \underbrace{w_{ML} J_{ML}}_{\text{max likelihood}} + \underbrace{w_{KL} J_{KL}}_{\text{Kullback}-\text{Leibler}} + \underbrace{w_{RC} J_{RC}}_{\text{reaction coordinate}}.$$

- KL divergence between q and p, entropy H_q:

$$\begin{aligned}\mathrm{KL}(q \parallel p) &= \int q(\mathbf{x}) \left[\log q(\mathbf{x}) - \log p(\mathbf{x})\right] \mathrm{d}\mathbf{x}, \\ &= H_q - \int q(\mathbf{x}) \log p(\mathbf{x}) \mathrm{d}\mathbf{x},\end{aligned}$$

- KL divergence between generated $p_X(\mathbf{x})$ and Boltzmann distribution:

$$\begin{aligned}\mathrm{KL}_\theta\left[q_Z \parallel p_Z\right] &= H_Z - \int q_Z(\mathbf{z}) \log p_Z(\mathbf{z};\theta) \mathrm{d}\mathbf{z}, \\ &= H_Z - \int q_Z(\mathbf{z}) \left[\log \mu_X(F_{zx}(\mathbf{z};\theta)) + \log\left|\mathbf{J}_{zx}(\mathbf{z};\theta)\right|\right] \mathrm{d}\mathbf{z}, \\ &= \underbrace{H_Z + \log Z_X}_{\text{const}} + \mathbb{E}_{\mathbf{z}\sim q_Z(\mathbf{z})}\left[u(F_{zx}(\mathbf{z};\theta)) - \log\left|\mathbf{J}_{zx}(\mathbf{z};\theta)\right|\right]\end{aligned}$$

- KL loss:

$$J_{KL} = \mathbb{E}_{\mathbf{z}\sim q_Z(\mathbf{z})}\left[u(F_{zx}(\mathbf{z};\theta)) - \log\left|\mathbf{J}_{zx}(\mathbf{z};\theta)\right|\right]. \quad (5)$$

Latent KL divergence

- Loss function:

$$J = \underbrace{w_{ML} J_{ML}}_{\text{max likelihood}} + \underbrace{w_{KL} J_{KL}}_{\text{Kullback}-\text{Leibler}} + \underbrace{w_{RC} J_{RC}}_{\text{reaction coordinate}}.$$

- KL divergence between q and p, entropy H_q:

$$\begin{aligned}\mathrm{KL}(q \parallel p) &= \int q(\mathbf{x}) \left[\log q(\mathbf{x}) - \log p(\mathbf{x})\right] \mathrm{d}\mathbf{x}, \\ &= H_q - \int q(\mathbf{x}) \log p(\mathbf{x}) \mathrm{d}\mathbf{x},\end{aligned}$$

- KL divergence between generated $p_X(\mathbf{x})$ and Boltzmann distribution:

$$\begin{aligned}\mathrm{KL}_\theta \left[q_Z \parallel p_Z\right] &= H_Z - \int q_Z(\mathbf{z}) \log p_Z(\mathbf{z}; \theta) \mathrm{d}\mathbf{z}, \\ &= H_Z - \int q_Z(\mathbf{z}) \left[\log \mu_X(F_{zx}(\mathbf{z}; \theta)) + \log \left|\mathbf{J}_{zx}(\mathbf{z}; \theta)\right|\right] \mathrm{d}\mathbf{z}, \\ &= \underbrace{H_Z + \log Z_X}_{\text{const}} + \mathbb{E}_{\mathbf{z} \sim q_Z(\mathbf{z})} \left[u(F_{zx}(\mathbf{z}; \theta)) - \log \left|\mathbf{J}_{zx}(\mathbf{z}; \theta)\right|\right]\end{aligned}$$

- KL loss:

$$J_{KL} = \mathbb{E}_{\mathbf{z} \sim q_Z(\mathbf{z})} \left[u(F_{zx}(\mathbf{z}; \theta)) - \log \left|\mathbf{J}_{zx}(\mathbf{z}; \theta)\right|\right]. \quad (5)$$

Latent KL divergence

- Loss function:

$$J = \underbrace{w_{ML} J_{ML}}_{\text{max likelihood}} + \underbrace{w_{KL} J_{KL}}_{\text{Kullback}-\text{Leibler}} + \underbrace{w_{RC} J_{RC}}_{\text{reaction coordinate}}.$$

- KL divergence between q and p, entropy H_q:

$$\begin{aligned}\mathrm{KL}(q \parallel p) &= \int q(\mathbf{x}) \left[\log q(\mathbf{x}) - \log p(\mathbf{x})\right] \mathrm{d}\mathbf{x}, \\ &= H_q - \int q(\mathbf{x}) \log p(\mathbf{x}) \mathrm{d}\mathbf{x},\end{aligned}$$

- KL divergence between generated $p_X(\mathbf{x})$ and Boltzmann distribution:

$$\begin{aligned}\mathrm{KL}_\theta \left[q_Z \parallel p_Z\right] &= H_Z - \int q_Z(\mathbf{z}) \log p_Z(\mathbf{z}; \theta) \mathrm{d}\mathbf{z}, \\ &= H_Z - \int q_Z(\mathbf{z}) \left[\log \mu_X(F_{zx}(\mathbf{z}; \theta)) + \log \left|\mathbf{J}_{zx}(\mathbf{z}; \theta)\right|\right] \mathrm{d}\mathbf{z}, \\ &= \underbrace{H_Z + \log Z_X}_{\text{const}} + \mathbb{E}_{\mathbf{z} \sim q_Z(\mathbf{z})} \left[u(F_{zx}(\mathbf{z}; \theta)) - \log \left|\mathbf{J}_{zx}(\mathbf{z}; \theta)\right|\right]\end{aligned}$$

- KL loss:

$$J_{KL} = \mathbb{E}_{\mathbf{z} \sim q_Z(\mathbf{z})} \left[u(F_{zx}(\mathbf{z}; \theta)) - \log \left|\mathbf{J}_{zx}(\mathbf{z}; \theta)\right|\right]. \quad (5)$$

Multi-T latent KL divergence

- **Boltzmann distribution** $\mathrm{e}^{-u^k(\mathbf{x})}$ at **multiple temperatures** $(T^1, ..., T^K)$. Reference temperature T^0, relative temperature τ_k:

$$u^k(\mathbf{x}) = \frac{T^0}{T^k} u^0(\mathbf{x}) = \frac{u^0(\mathbf{x})}{\tau_k}$$

- **Prior distribution** $q_Z^k(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}) = Z_Z^{-1} \mathrm{e}^{-\frac{1}{2}\|\mathbf{z}\|^2/\sigma_k^2}$, **Prior energy**:

$$u_Z^k(\mathbf{z}) = -\log q_Z^k(\mathbf{z}) = \frac{1}{2\sigma_k^2} \|\mathbf{z}\|^2 + \text{const}$$

$\rightarrow$ variance takes same role as relative temperature. Setting variance$= 1$ at standard temperature, we obtain:

$$\sigma_k^2 = \tau_k.$$

- **Multi-temperature KL loss**:

$$J_{KL}^{T^1,...,T^K} = \sum_{k=1}^{K} \mathbb{E}_{\mathbf{z}\sim q_Z^k(\mathbf{z})} \left[u^k(F_{zx}(\mathbf{z};\theta)) - \log |\mathbf{J}_{zx}(\mathbf{z};\theta)| \right].$$

Multi-T latent KL divergence

- **Boltzmann distribution** $\mathrm{e}^{-u^k(\mathbf{x})}$ at **multiple temperatures** $(T^1, ..., T^K)$. Reference temperature T^0, relative temperature τ_k:

$$u^k(\mathbf{x}) = \frac{T^0}{T^k} u^0(\mathbf{x}) = \frac{u^0(\mathbf{x})}{\tau_k}$$

- **Prior distribution** $q_Z^k(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}) = Z_Z^{-1} \mathrm{e}^{-\frac{1}{2}\|\mathbf{z}\|^2/\sigma_k^2}$, **Prior energy**:

$$u_Z^k(\mathbf{z}) = -\log q_Z^k(\mathbf{z}) = \frac{1}{2\sigma_k^2} \|\mathbf{z}\|^2 + \text{const}$$

$\rightarrow$ variance takes same role as relative temperature. Setting variance$= 1$ at standard temperature, we obtain:

$$\sigma_k^2 = \tau_k.$$

- **Multi-temperature KL loss**:

$$J_{KL}^{T^1,...,T^K} = \sum_{k=1}^{K} \mathbb{E}_{\mathbf{z}\sim q_Z^k(\mathbf{z})} \left[u^k(F_{zx}(\mathbf{z};\theta)) - \log |\mathbf{J}_{zx}(\mathbf{z};\theta)| \right].$$

Multi-T latent KL divergence

- **Boltzmann distribution** $\mathrm{e}^{-u^k(\mathbf{x})}$ at **multiple temperatures** $(T^1, ..., T^K)$. Reference temperature T^0, relative temperature τ_k:

$$u^k(\mathbf{x}) = \frac{T^0}{T^k} u^0(\mathbf{x}) = \frac{u^0(\mathbf{x})}{\tau_k}$$

- **Prior distribution** $q_Z^k(\mathbf{z}) = \mathcal{N}(\mathbf{0}, \sigma_k^2 \mathbf{I}) = Z_Z^{-1} \mathrm{e}^{-\frac{1}{2}\|\mathbf{z}\|^2/\sigma_k^2}$, **Prior energy**:

$$u_Z^k(\mathbf{z}) = -\log q_Z^k(\mathbf{z}) = \frac{1}{2\sigma_k^2} \|\mathbf{z}\|^2 + \text{const}$$

$\rightarrow$ variance takes same role as relative temperature. Setting variance$= 1$ at standard temperature, we obtain:

$$\sigma_k^2 = \tau_k.$$

- **Multi-temperature KL loss**:

$$J_{KL}^{T^1,...,T^K} = \sum_{k=1}^{K} \mathbb{E}_{\mathbf{z}\sim q_Z^k(\mathbf{z})} \left[u^k(F_{zx}(\mathbf{z};\theta)) - \log |\mathbf{J}_{zx}(\mathbf{z};\theta)| \right].$$

Configuration KL divergence

- Loss function:

$$J = \underbrace{w_{ML} J_{ML}}_{\text{max likelihood}} + \underbrace{w_{KL} J_{KL}}_{\text{Kullback}-\text{Leibler}} + \underbrace{w_{RC} J_{RC}}_{\text{reaction coordinate}}.$$

- KL divergence between generated distribution $q_Z(\mathbf{z})$ and Gaussian:

$$\begin{aligned}
\mathrm{KL}_\theta(\mu_X \parallel p_X) &= H_X - \int \mu_X(\mathbf{x}) \log p_X(\mathbf{x};\theta) \mathrm{d}\mathbf{x} \\
&= H_X - \int \mu_X(\mathbf{x}) \left[\log q_Z(F_{xz}(\mathbf{x};\theta)) + \log |\mathbf{J}_{xz}(\mathbf{z};\theta)|\right] \mathrm{d}\mathbf{x}. \\
&= \underbrace{H_X + \log Z_Z}_{\text{const}} + \mathbb{E}_{\mathbf{x}\sim\mu(\mathbf{x})} \left[\frac{1}{\sigma^2} \|F_{xz}(\mathbf{x};\theta)\|^2 - \log |\mathbf{J}_{xz}(\mathbf{x};\theta)|\right].
\end{aligned}$$

  Problem: sampling $\mathbf{x} \sim \mu(\mathbf{x})$ is difficult and our goal!

- **Maximum Likelihood** loss:

$$\begin{aligned}
J_{ML} &= -\mathbb{E}_{\mathbf{x}\sim\rho(\mathbf{x})} \left[\log p_X(\mathbf{x};\theta)\right] \\
&= \mathbb{E}_{\mathbf{x}\sim\rho(\mathbf{x})} \left[\frac{1}{\sigma^2} \|F_{xz}(\mathbf{x};\theta)\|^2 - \log |\mathbf{J}_{xz}(\mathbf{x};\theta)|\right]
\end{aligned}$$

Configuration KL divergence

- Loss function:

$$J = \underbrace{w_{ML} J_{ML}}_{\text{max likelihood}} + \underbrace{w_{KL} J_{KL}}_{\text{Kullback}-\text{Leibler}} + \underbrace{w_{RC} J_{RC}}_{\text{reaction coordinate}}.$$

- KL divergence between generated distribution $q_Z(\mathbf{z})$ and Gaussian:

$$\begin{aligned}
\mathrm{KL}_\theta(\mu_X \parallel p_X) &= H_X - \int \mu_X(\mathbf{x}) \log p_X(\mathbf{x};\theta) \mathrm{d}\mathbf{x} \\
&= H_X - \int \mu_X(\mathbf{x}) \left[\log q_Z(F_{xz}(\mathbf{x};\theta)) + \log |\mathbf{J}_{xz}(\mathbf{z};\theta)|\right] \mathrm{d}\mathbf{x}. \\
&= \underbrace{H_X + \log Z_Z}_{\text{const}} + \mathbb{E}_{\mathbf{x}\sim\mu(\mathbf{x})} \left[\frac{1}{\sigma^2} \left\| F_{xz}(\mathbf{x};\theta) \right\|^2 - \log |\mathbf{J}_{xz}(\mathbf{x};\theta)| \right].
\end{aligned}$$

  Problem: sampling $\mathbf{x} \sim \mu(\mathbf{x})$ is difficult and our goal!

- **Maximum Likelihood** loss:

$$\begin{aligned}
J_{ML} &= -\mathbb{E}_{\mathbf{x}\sim\rho(\mathbf{x})} \left[\log p_X(\mathbf{x};\theta)\right] \\
&= \mathbb{E}_{\mathbf{x}\sim\rho(\mathbf{x})} \left[\frac{1}{\sigma^2} \left\| F_{xz}(\mathbf{x};\theta) \right\|^2 - \log |\mathbf{J}_{xz}(\mathbf{x};\theta)| \right]
\end{aligned}$$

Configuration KL divergence

- Loss function:

$$J = \underbrace{w_{ML} J_{ML}}_{\text{max likelihood}} + \underbrace{w_{KL} J_{KL}}_{\text{Kullback}-\text{Leibler}} + \underbrace{w_{RC} J_{RC}}_{\text{reaction coordinate}}.$$

- KL divergence between generated distribution $q_Z(\mathbf{z})$ and Gaussian:

$$\begin{aligned}
\mathrm{KL}_\theta(\mu_X \parallel p_X) &= H_X - \int \mu_X(\mathbf{x}) \log p_X(\mathbf{x};\theta) \mathrm{d}\mathbf{x} \\
&= H_X - \int \mu_X(\mathbf{x}) \left[\log q_Z(F_{xz}(\mathbf{x};\theta)) + \log |\mathbf{J}_{xz}(\mathbf{z};\theta)|\right] \mathrm{d}\mathbf{x}. \\
&= \underbrace{H_X + \log Z_Z}_{\text{const}} + \mathbb{E}_{\mathbf{x}\sim\mu(\mathbf{x})} \left[\frac{1}{\sigma^2} \left\| F_{xz}(\mathbf{x};\theta) \right\|^2 - \log |\mathbf{J}_{xz}(\mathbf{x};\theta)| \right].
\end{aligned}$$

  Problem: sampling $\mathbf{x} \sim \mu(\mathbf{x})$ is difficult and our goal!

- **Maximum Likelihood** loss:

$$\begin{aligned}
J_{ML} &= -\mathbb{E}_{\mathbf{x}\sim\rho(\mathbf{x})} \left[\log p_X(\mathbf{x};\theta)\right] \\
&= \mathbb{E}_{\mathbf{x}\sim\rho(\mathbf{x})} \left[\frac{1}{\sigma^2} \left\| F_{xz}(\mathbf{x};\theta) \right\|^2 - \log |\mathbf{J}_{xz}(\mathbf{x};\theta)| \right]
\end{aligned}$$

Reaction coordinate loss

- Loss function:

$$J = \underbrace{w_{ML} J_{ML}}_{\text{max likelihood}} + \underbrace{w_{KL} J_{KL}}_{\text{Kullback}-\text{Leibler}} + \underbrace{w_{RC} J_{RC}}_{\text{reaction coordinate}}.$$

- Reaction coordinate loss:

$$\begin{aligned} J_{RC} &= \int p(R(\mathbf{x})) \log p(R(\mathbf{x})) \, \mathrm{d}R(\mathbf{x}) \\ &= \mathbb{E}_{\mathbf{x} \sim p_X(\mathbf{x})} \log p(R(\mathbf{x})) . \end{aligned}$$

- Implementation:
  - Reaction coordinate function R is user input
  - Min and max bounds are given
  - $p(R(\mathbf{x}))$ is computed as a batchwise kernel density estimate.

Reaction coordinate loss

- Loss function:

$$J = \underbrace{w_{ML} J_{ML}}_{\text{max likelihood}} + \underbrace{w_{KL} J_{KL}}_{\text{Kullback}-\text{Leibler}} + \underbrace{w_{RC} J_{RC}}_{\text{reaction coordinate}}.$$

- Reaction coordinate loss:

$$\begin{aligned} J_{RC} &= \int p(R(\mathbf{x})) \log p(R(\mathbf{x})) \, \mathrm{d}R(\mathbf{x}) \\ &= \mathbb{E}_{\mathbf{x} \sim p_X(\mathbf{x})} \log p(R(\mathbf{x})). \end{aligned}$$

- Implementation:
  - Reaction coordinate function R is user input
  - Min and max bounds are given
  - $p(R(\mathbf{x}))$ is computed as a batchwise kernel density estimate.

Reaction coordinate loss

- Loss function:

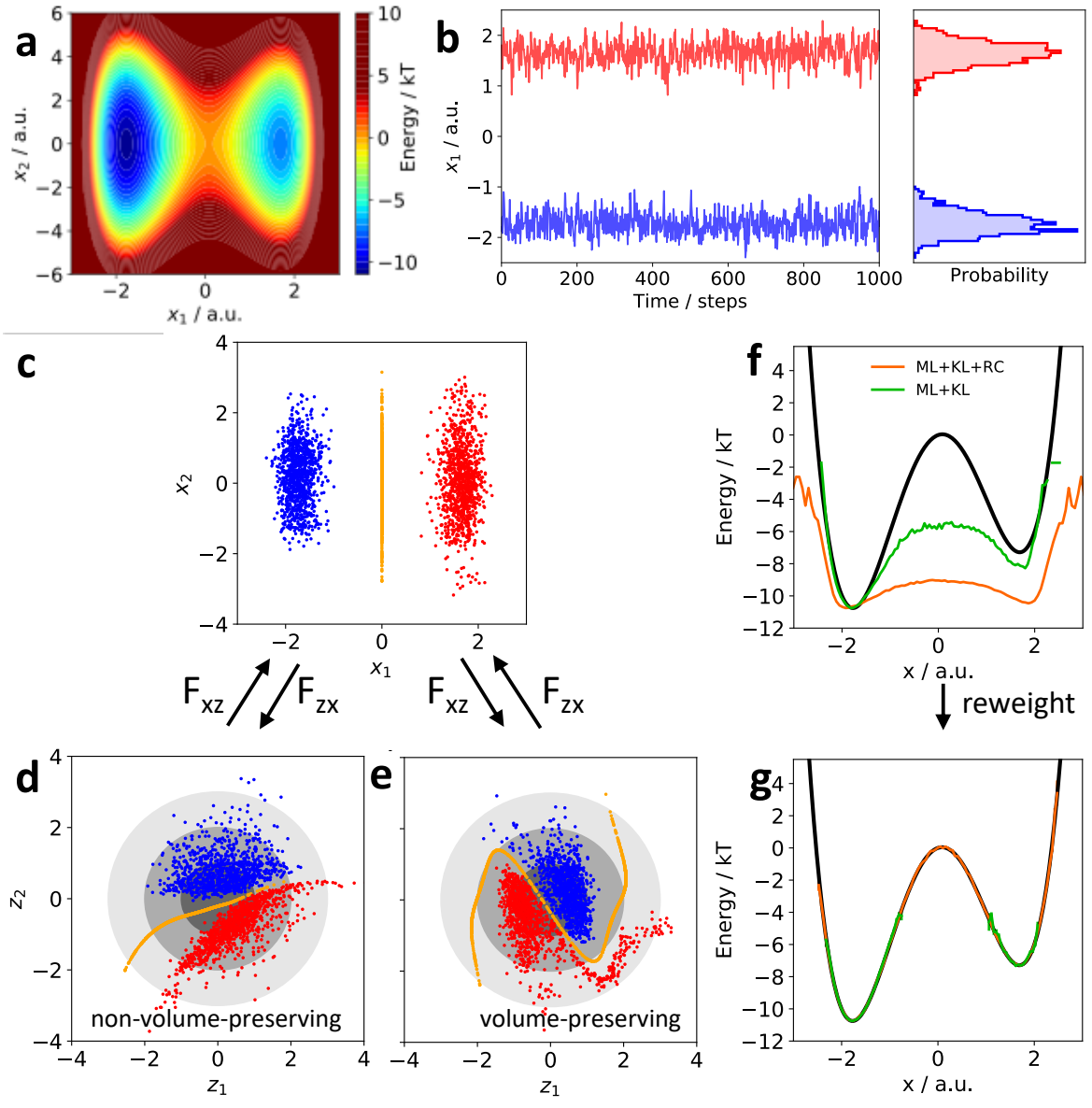
$$J = \underbrace{w_{ML} J_{ML}}_{\text{max likelihood}} + \underbrace{w_{KL} J_{KL}}_{\text{Kullback}-\text{Leibler}} + \underbrace{w_{RC} J_{RC}}_{\text{reaction coordinate}}.$$

- Reaction coordinate loss:

$$\begin{aligned} J_{RC} &= \int p(R(\mathbf{x})) \log p(R(\mathbf{x})) \, \mathrm{d}R(\mathbf{x}) \\ &= \mathbb{E}_{\mathbf{x} \sim p_X(\mathbf{x})} \log p(R(\mathbf{x})). \end{aligned}$$

- Implementation:
  - Reaction coordinate function R is user input
  - Min and max bounds are given
  - $p(R(\mathbf{x}))$ is computed as a batchwise kernel density estimate.

Boltzmann Generator - 1D Example



Reweighting

- Probability reweighting: assign to each generated configuration $\mathbf{x}$ the statistical weight:

$$w_X(\mathbf{x} \mid \mathbf{z}) = \frac{\mu_X(\mathbf{x})}{p_X(\mathbf{x})} = \frac{p_Z(\mathbf{z})}{q_Z(\mathbf{z})}. \tag{6}$$
$$\propto \mathrm{e}^{-u_X(T_{zx}(\mathbf{z}))+u_Z(\mathbf{z})+\log|\det(\mathbf{J}_{zx}(\mathbf{z}))|}$$

- Weighted expectation values

$$\mathbb{E}[O] \approx \frac{\sum_{i=1}^{N} w_X(\mathbf{x}) O(\mathbf{x})}{\sum_{i=1}^{N} w_X(\mathbf{x})}.$$

Reweighting

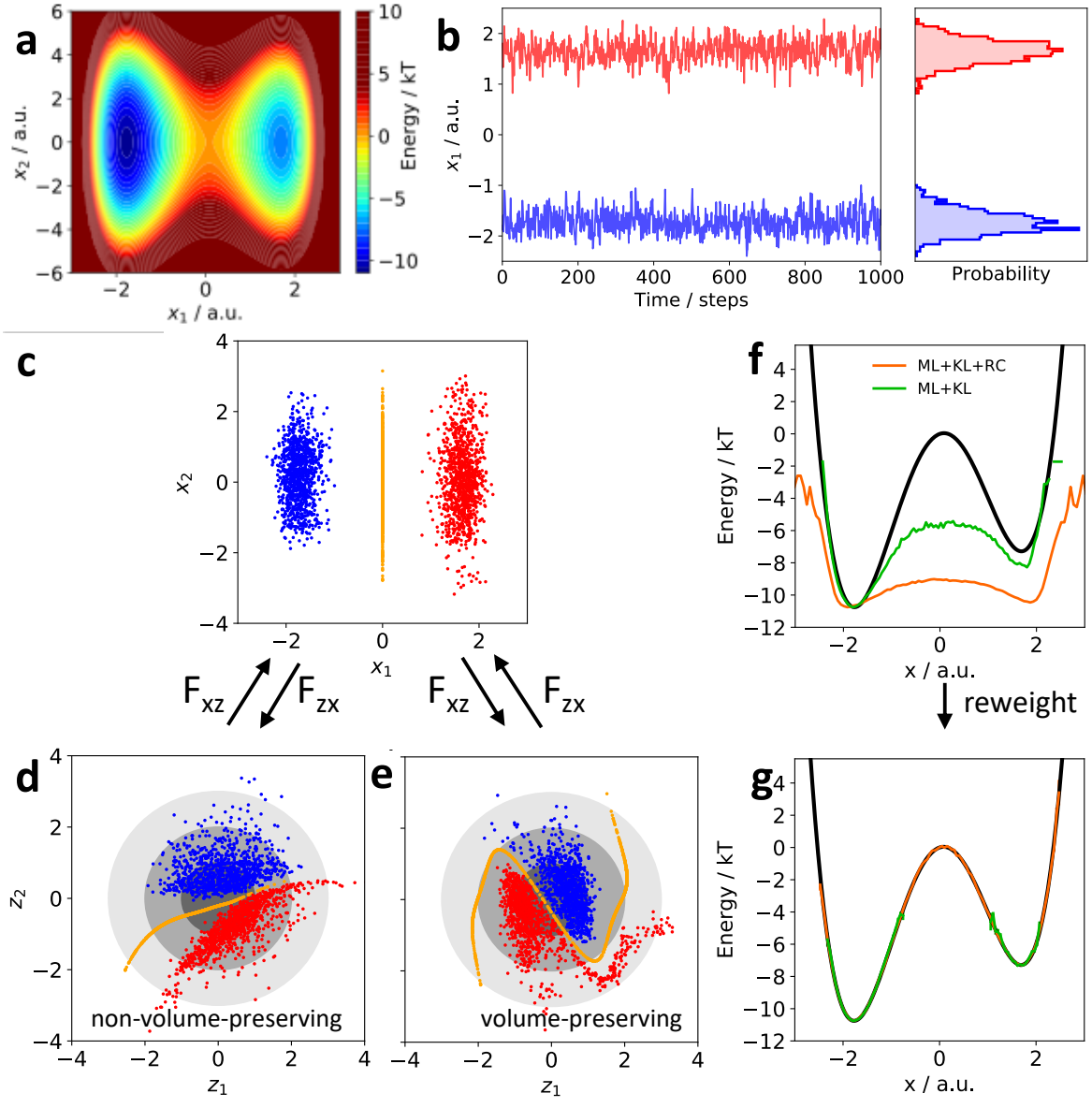
- Probability reweighting: assign to each generated configuration $\mathbf{x}$ the statistical weight:

$$w_X(\mathbf{x} \mid \mathbf{z}) = \frac{\mu_X(\mathbf{x})}{p_X(\mathbf{x})} = \frac{p_Z(\mathbf{z})}{q_Z(\mathbf{z})}. \tag{6}$$
$$\propto \mathrm{e}^{-u_X(T_{zx}(\mathbf{z}))+u_Z(\mathbf{z})+\log|\det(\mathbf{J}_{zx}(\mathbf{z}))|}$$

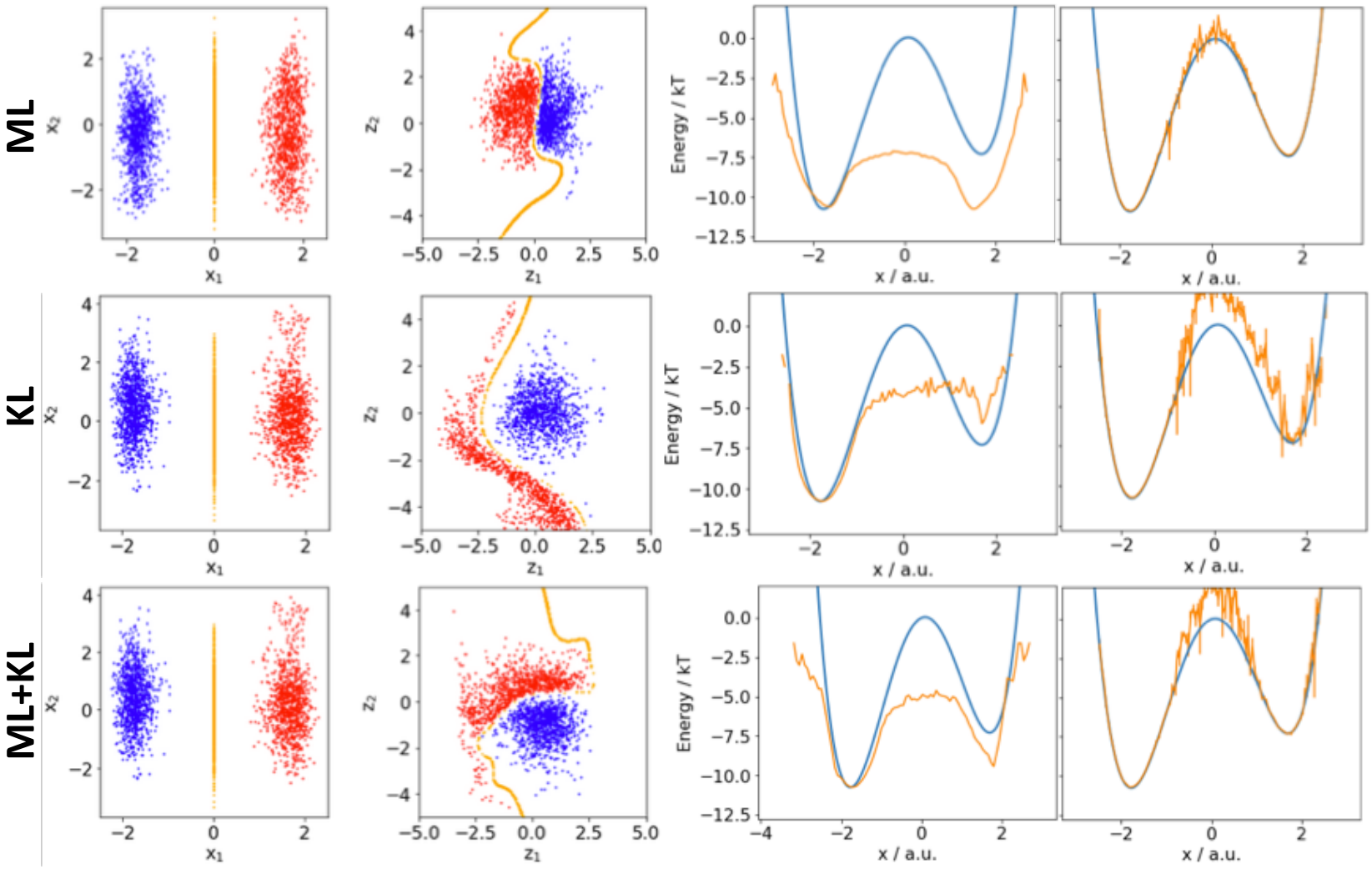
- Weighted expectation values

$$\mathbb{E}[O] \approx \frac{\sum_{i=1}^{N} w_X(\mathbf{x}) O(\mathbf{x})}{\sum_{i=1}^{N} w_X(\mathbf{x})}.$$

Boltzmann Generator - 1D Example

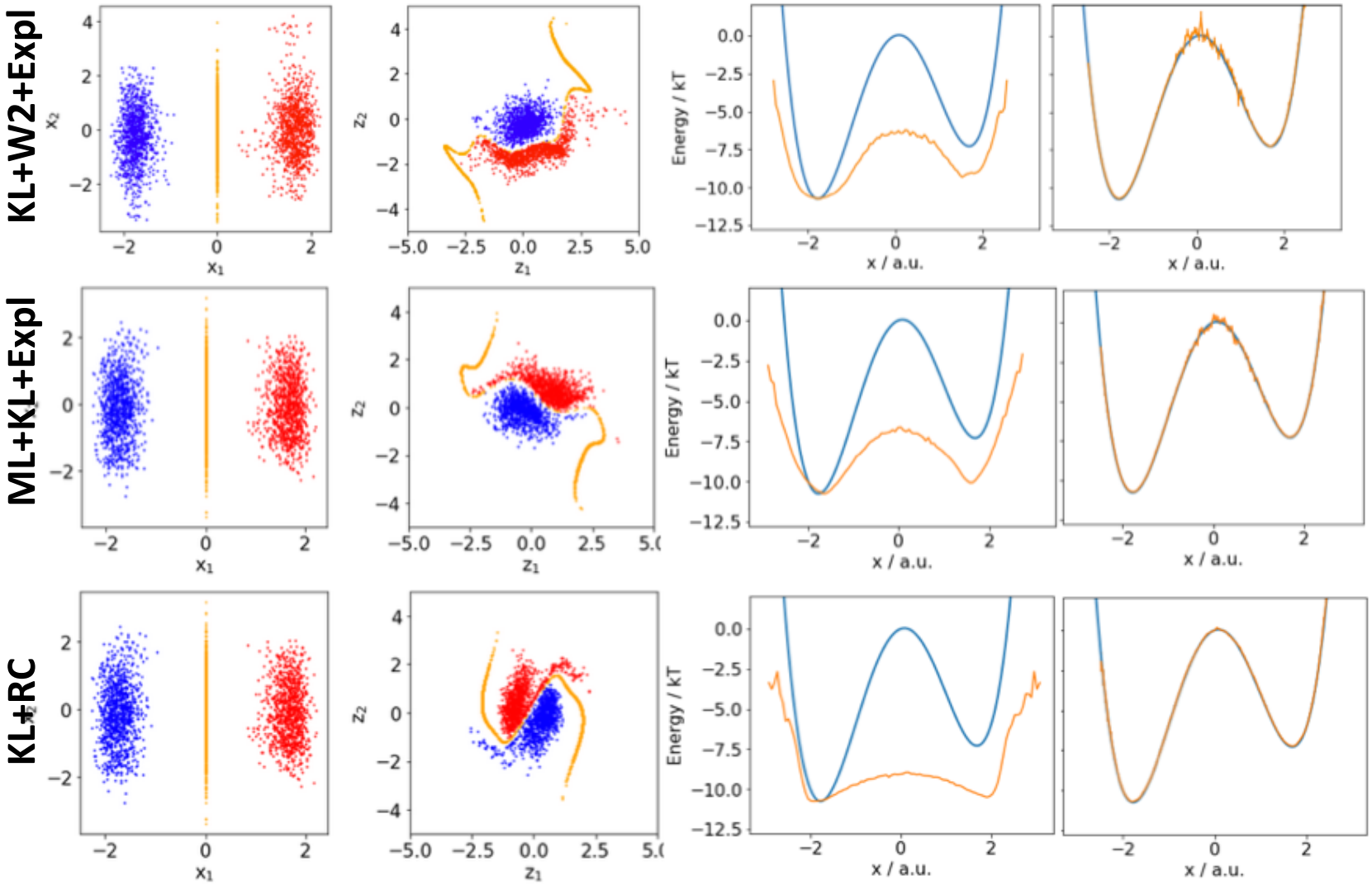


Different training methods



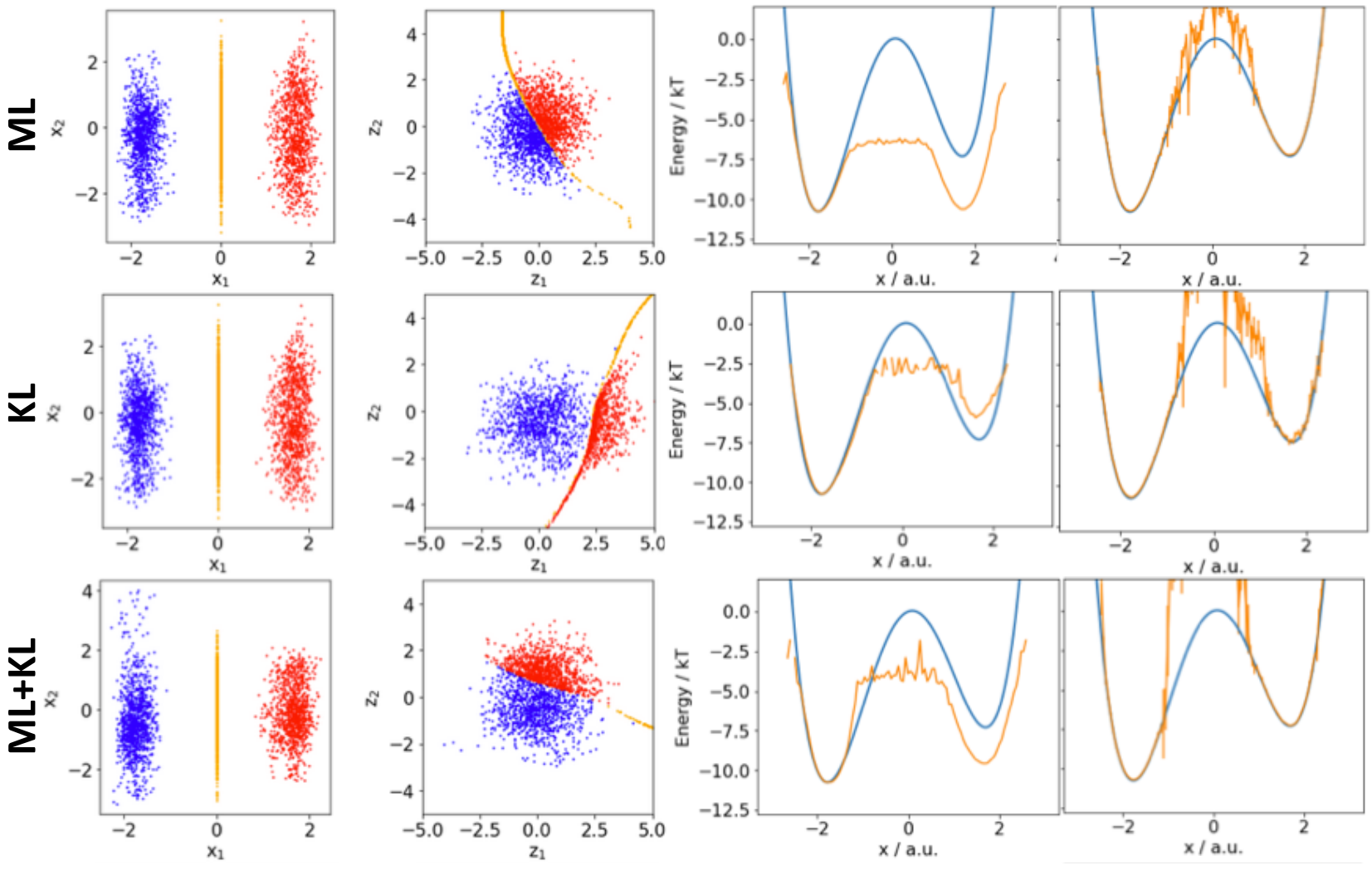
Architecture $\mathrm{N}_{10}\mathrm{S}$, $nl_{layers} = 2$, $nl_{hidden} = 100$, $\sigma = \tanh$

Different training methods



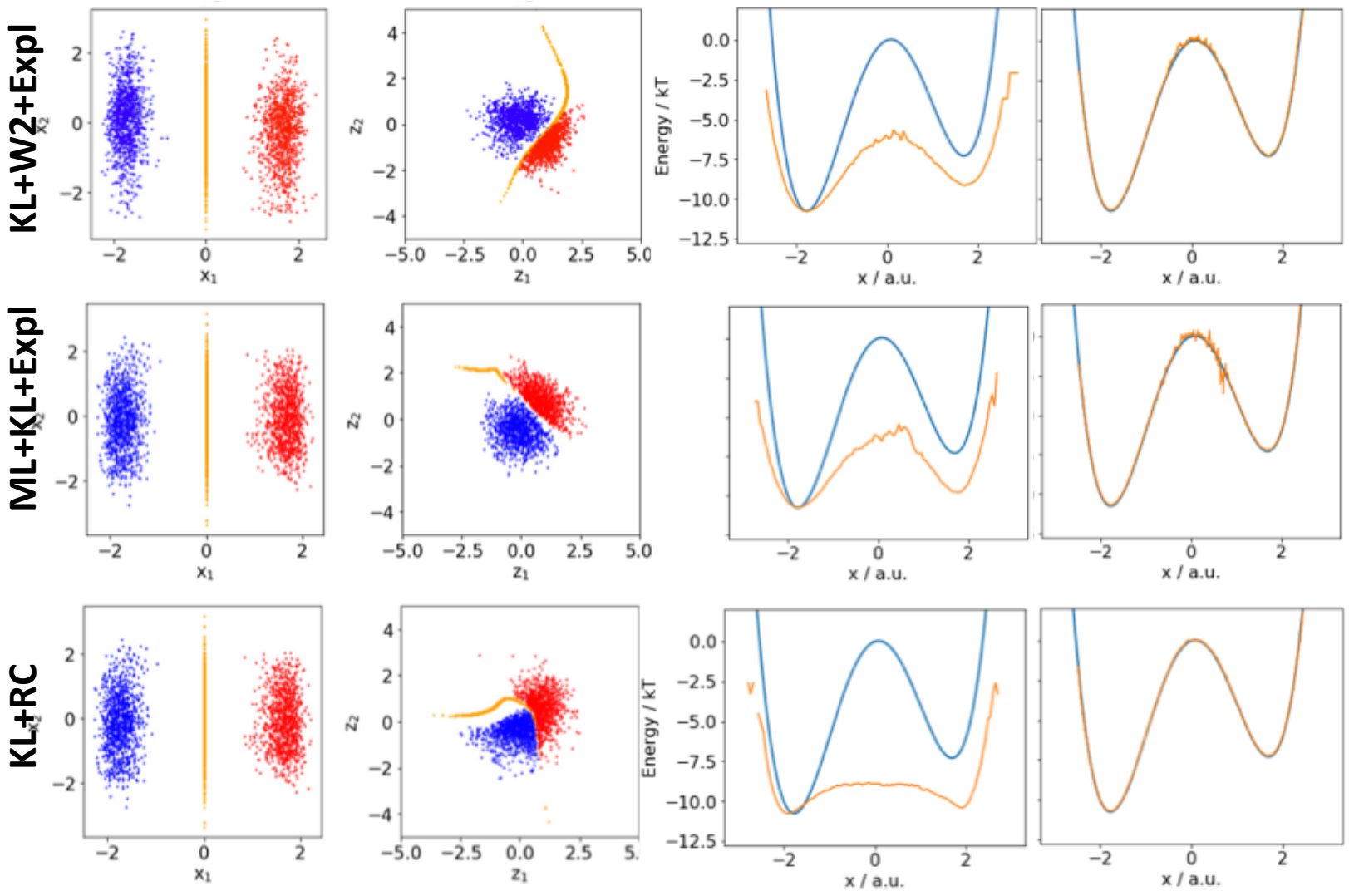
Architecture $\mathrm{N}_{10}\mathrm{S}$, $nl_{layers} = 2$, $nl_{hidden} = 100$, $\sigma = \tanh$

Different training methods



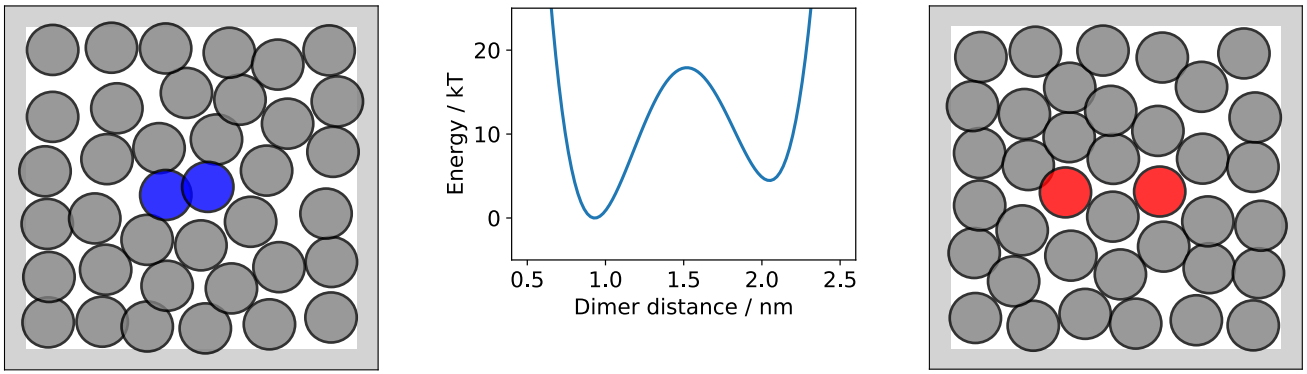
Architecture R_{10}, $nl_{layers} = 2$, $nl_{hidden} = 100$, $\sigma = \tanh$

Different training methods



Architecture R_{10}, $nl_{layers} = 2$, $nl_{hidden} = 100$, $\sigma = \tanh$

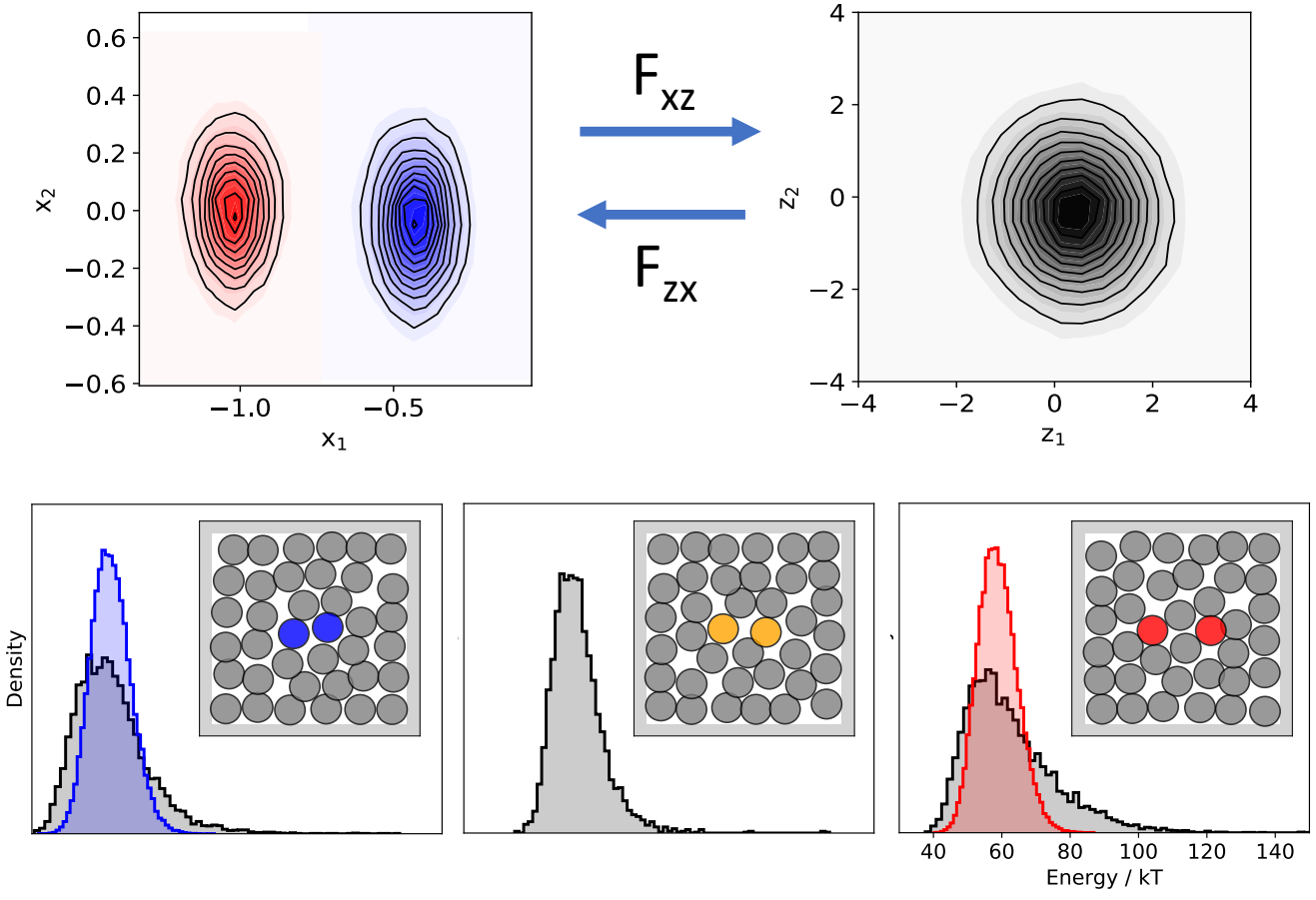
Boltzmann Generator - Particle dimer



- Configuration: $\mathbf{x} = \left[\mathbf{x}_{1x}, \mathbf{x}_{1y}, \mathbf{x}_{2x}, \mathbf{x}_{2y}, ..., \mathbf{x}_{(n_s+2)x}, \mathbf{x}_{(n_s+2)y}\right]$
- Dimer distance $d = \|\mathbf{x}_1 - \mathbf{x}_2\|$, Heavyside function h, **potential energy**:

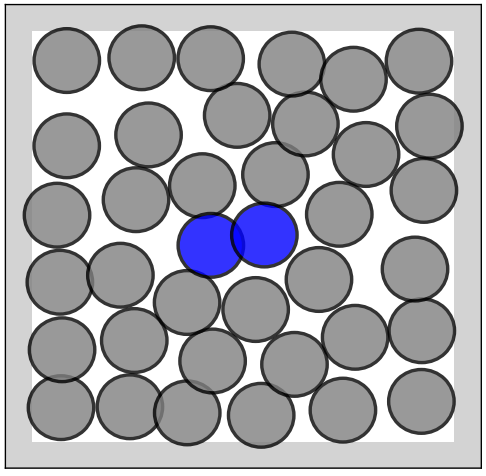
$$
\begin{aligned}
U(\mathbf{x}) &= k_d(\mathbf{x}_{1x} + \mathbf{x}_{2x})^2 + k_d\mathbf{x}_{1y}^2 + k_d\mathbf{x}_{2y}^2 \\
&+ \frac{1}{4}a(d-d_0)^4 - \frac{1}{2}b(d-d_0)^2 + c(d-d_0)^4 \\
&+ \sum_{i=1}^{n+2} h(-\mathbf{x}_{1x} - l_{\text{box}})k_{\text{box}}(-\mathbf{x}_{1x} - l_{\text{box}})^2 + \sum_{i=1}^{n+2} h(\mathbf{x}_{1x} - l_{\text{box}})k_{\text{box}}(\mathbf{x}_{1x} - l_{\text{box}})^2 \\
&+ \sum_{i=1}^{n+2} h(-\mathbf{x}_{1y} - l_{\text{box}})k_{\text{box}}(-\mathbf{x}_{1y} - l_{\text{box}})^2 + \sum_{i=1}^{n+2} h(\mathbf{x}_{1y} - l_{\text{box}})k_{\text{box}}(\mathbf{x}_{1y} - l_{\text{box}})^2 \\
&+ \varepsilon \sum_{i=1}^{n+1} \sum_{j=i+1, j\neq 2}^{n+2} \left(\frac{\sigma}{\|\mathbf{x}_i - \mathbf{x}_j\|}\right)^{12}
\end{aligned}
$$

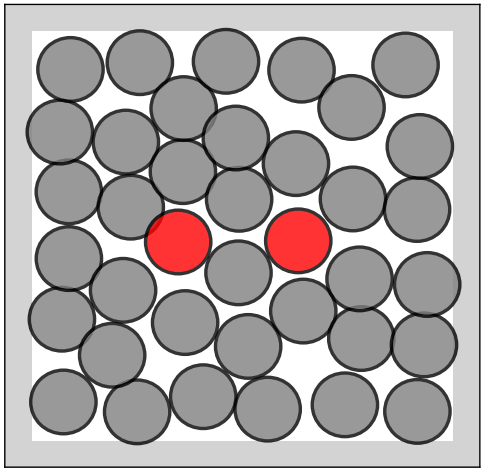
Boltzmann Generator - Particle dimer

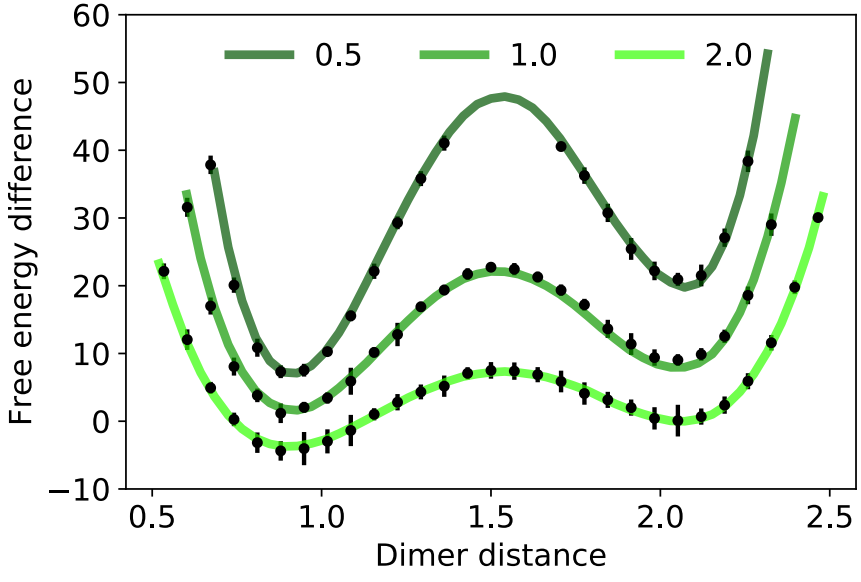


Architecture R_8, $nl_{layers} = 3$, $nl_{hidden} = 200$, $\sigma = \tanh$, $w_{KL} = 1$, $w_{ML} = 0.01$, $w_{RC} = 10$

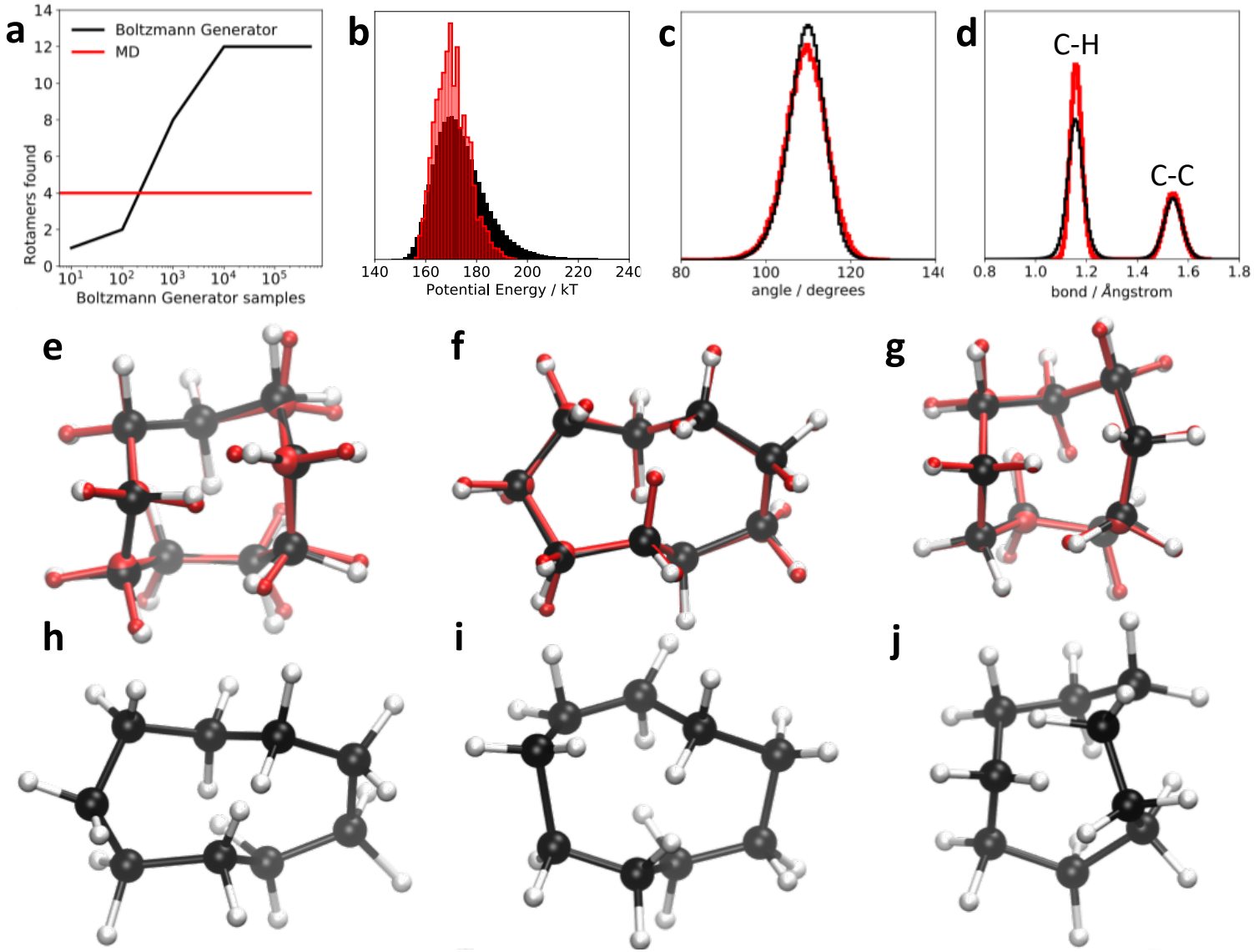
Boltzmann Generator - Particle dimer





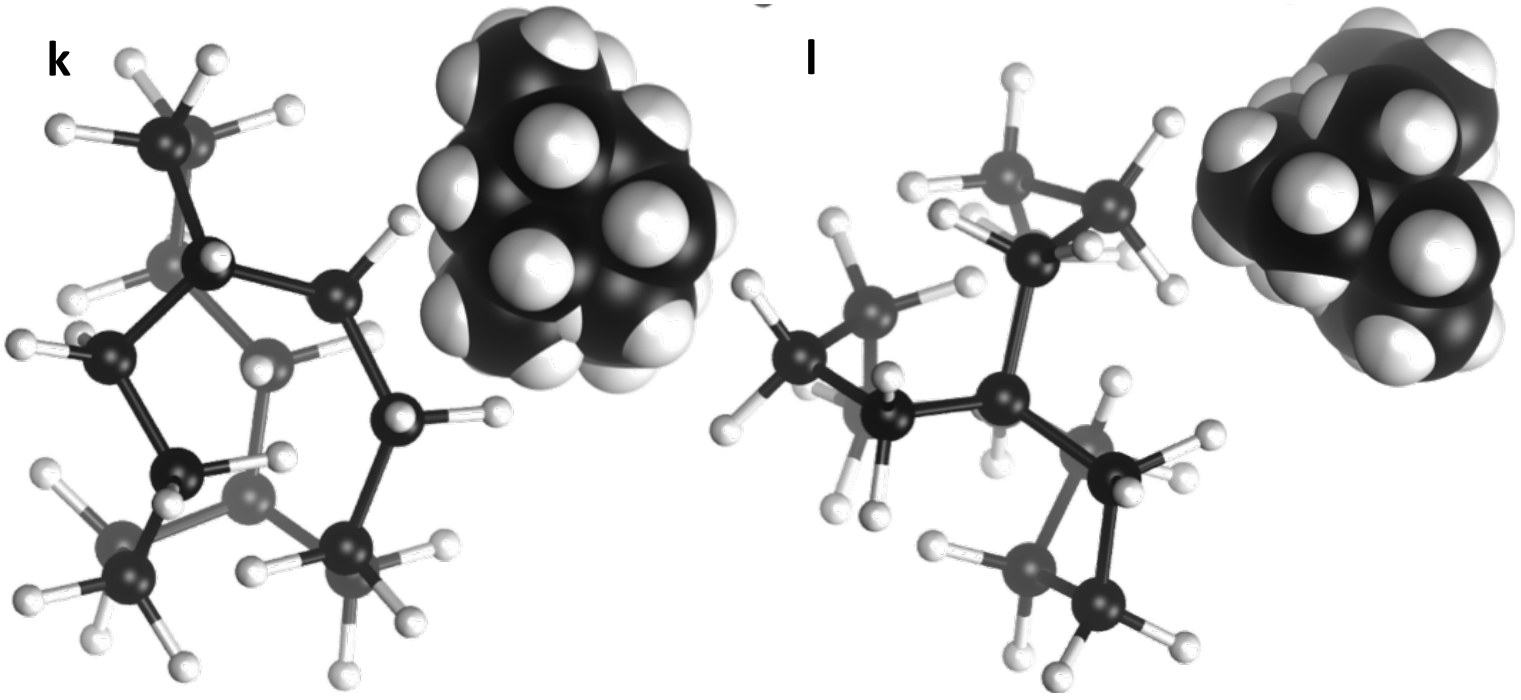


Boltzmann Generator - Hydrocarbons



Architecture R_4, $nl_{layers} = 2$, $nl_{hidden} = 100$, $\sigma = \tanh$, $w_{KL} = 0.01$, $w_{ML} = 1$

Boltzmann Generator - Hydrocarbons



Architecture R_4, $nl_{layers} = 2$, $nl_{hidden} = 100$, $\sigma = \tanh$, $w_{KL} = 0.01$, $w_{ML} = 1$