Equivariant Diffusion for Molecule Generation in 3D

Emiel Hoogeboom *, Victor Garcia Satorras *, Clément Vignac, Max Welling

Presenter: Will Hua (McGill & Mila)

Nov 2022

Content

Mila

01

Contribution

Finding

1) An E(3) Equivariant Diffusion Model (EDM)

Jointly operates on both continuous (atom coordinates) and categorical features (atom types).

2) Interesting results

Can directly generate molecules in 3D space.

Can outperform previous methods regarding the quality of generated samples and efficiency at training time.

Limitation of Existing Generative Models

1) Diffusion Models

Problem: Cannot directly generate molecules. Atoms must be given.
EDM Advantage: Does not require atom types.

2) Autoregressive Models

Problem: A particular atom ordering must be given.
EDM Advantage: Does not require atom orderings.

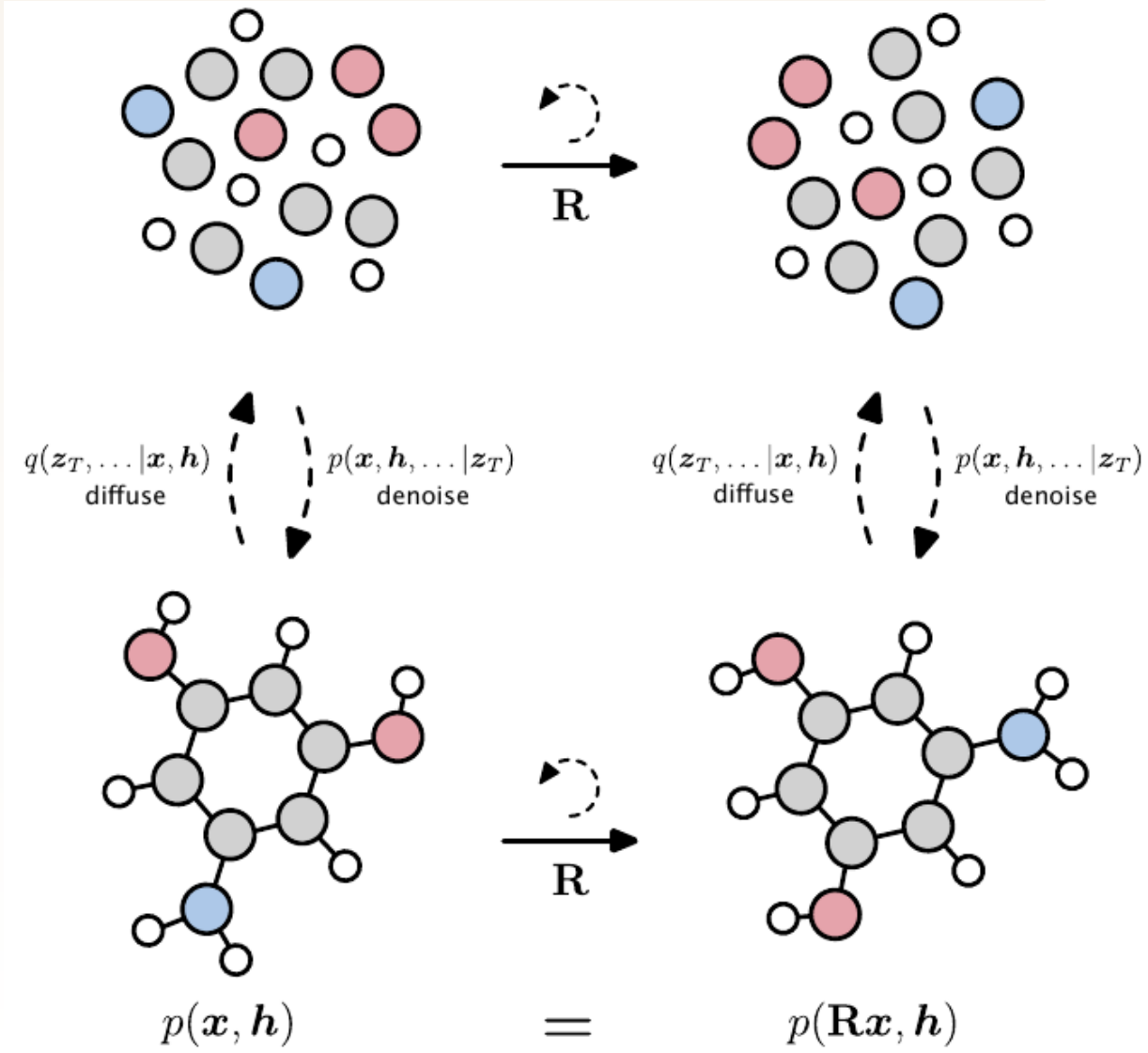
3) Flow Models

Problem: Slow and inefficient. Expensive.
EDM Advantage: Up to 16 times more stable molecules with half of the training time.

Equivariant Diffusion Model Overview

A set of points are denoised into a molecule consisting of atom coordinates x and atom types h.

EDM is rotation equivariant.



Mila

02

Background

Noising Process

Given a data point x, a diffusion process that adds noise to x for $t = 0, \ldots, T$ is defined by the multivariate normal distribution:

$$q(z_t \vee x) = N(z_t \vee \alpha_t x_t, \sigma_t^2 I)$$

α controls the signal, σ controls the noise.

This diffusion process is Markov and:

$$q(z_t \vee z_{t-1}) = N(z_t \vee \alpha_{t \vee t-1} z_{t-1}, \sigma_{t \vee t-1}^2 I), \alpha_{t \vee t-1} = \frac{a_t}{a_{t-1}}, \sigma_{t \vee t-1}^2 = \sigma_t^2 - \alpha_{t \vee t-1}^2 \sigma_{t-1}^2$$

The entire noising process is then written as:

$$q(z_0, z_1, \ldots, z_T \vee x) = q(z_0 \vee x) \prod_{t=1}^{T} q\,(z_t \vee z_{t-1})$$

Denoising Process

Follow the inverse of diffusion process:

$$q(z_s \vee x, z_t) = N(z_s \vee \mu_{t\to s}(x, z_t), \sigma^2_{t\to s} I)$$

the distribution parameters can be obtained:

$$\mu_{t\to s}(x, z_t) = \frac{\alpha_{t\vee s}\sigma_s^2}{\sigma_t^2} z_t + \frac{\alpha_s \sigma^2_{t\vee s}}{\sigma_t^2} x \qquad \sigma_{t\to s} = \frac{\sigma_{t\vee s}\sigma_s}{\sigma_t}$$

The generative transition distribution is chosen to be,

$$p(z_s \vee z_t) = q(z_s \vee \hat{x}, z_t) = N(z_s \vee \mu_{t\to s}(\hat{x}, z_t), \sigma^2_{t\to s} I), \hat{x} = \phi(z_t, t)$$

Equivariance

A function f is said to be **equivariant** to the action of a group G if

$$T_g\big(f(x)\big) = f\left(S_g(x)\right), \forall g \in G$$

where T_g, S_g are linear representations related to the group element g.

EDM consider the E(3) generated by **translations, rotations and reflections**

T_g, S_g can be represented by a **translation** t and an orthogonal matrix R that **rotates or reflects coordinates**.

$$R\big(f(x)\big) = f\big(R(x)\big)$$

Mila

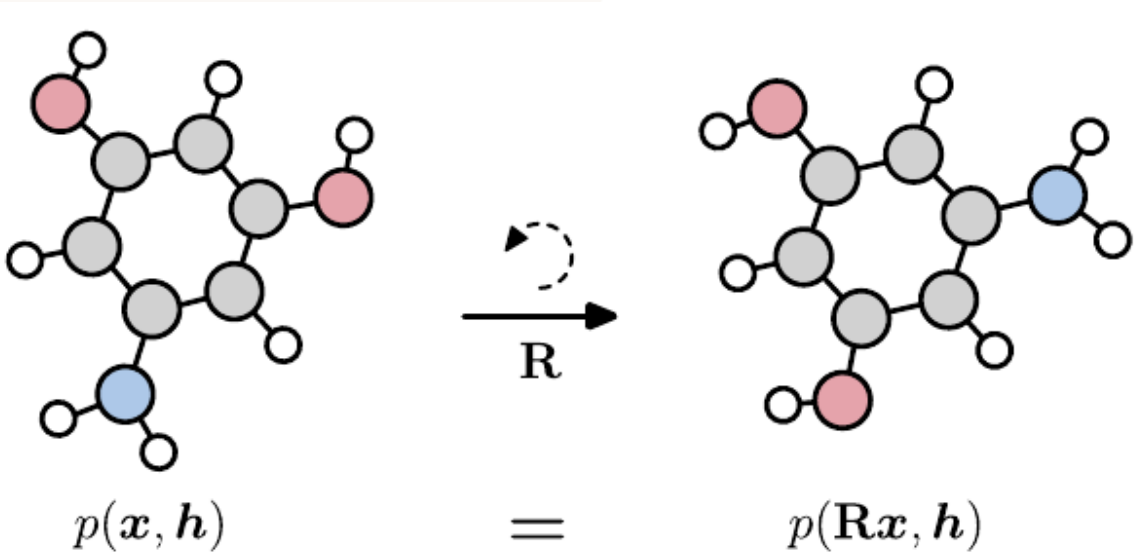
Equivariance

For a set of positions $x = (x_1, \ldots, x_n) \in \mathbb{R}^{M \times 3}$

and features $h = (h_1, \ldots, h_n) \in \mathbb{R}^{M \times nf}$

$$Rz_x + t, z_h = f(Rx + t, h)$$

Positions are equivariant, features are invariant



Mila

Equivariant Graph Convolutional Layer

For every update,

$$x^{l+1}, h^{l+1} = EGCL(x^l, h^l)$$

For every vertex , v_i

$$h_i^{l+1} = \phi_h\left(h_i^l, \sum_{j \neq i} \phi_{inf}\left(m_{ij}\right) m_{ij}\right), m_{ij} = \phi_e\left(h_i^l, h_j^l, d_{ij}^2, a_{ij}\right)$$

$$x_i^{l+1} = x_i^l + \frac{\sum_{j \neq i} x_i^l - x_j^l}{d_{ij} + 1} \phi_x\left(h_i^l, h_j^l, d_{ij}^2, a_{ij}\right)$$

$d_{ij} = \left\|x_i^l - x_j^l\right\|_2, a_{ij}$ **are edge attributes.**

Mila

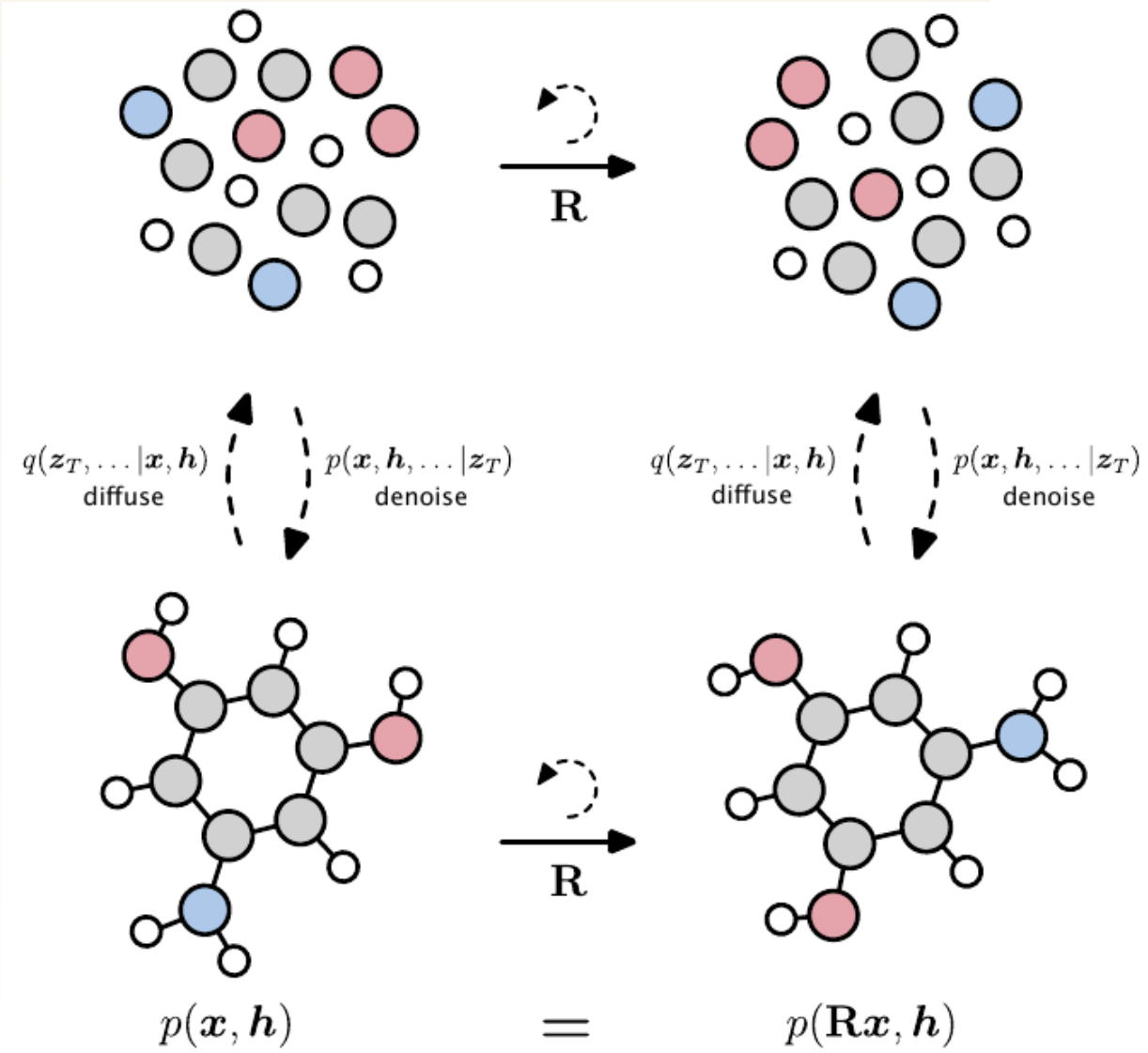
03

Model

EDM Overview

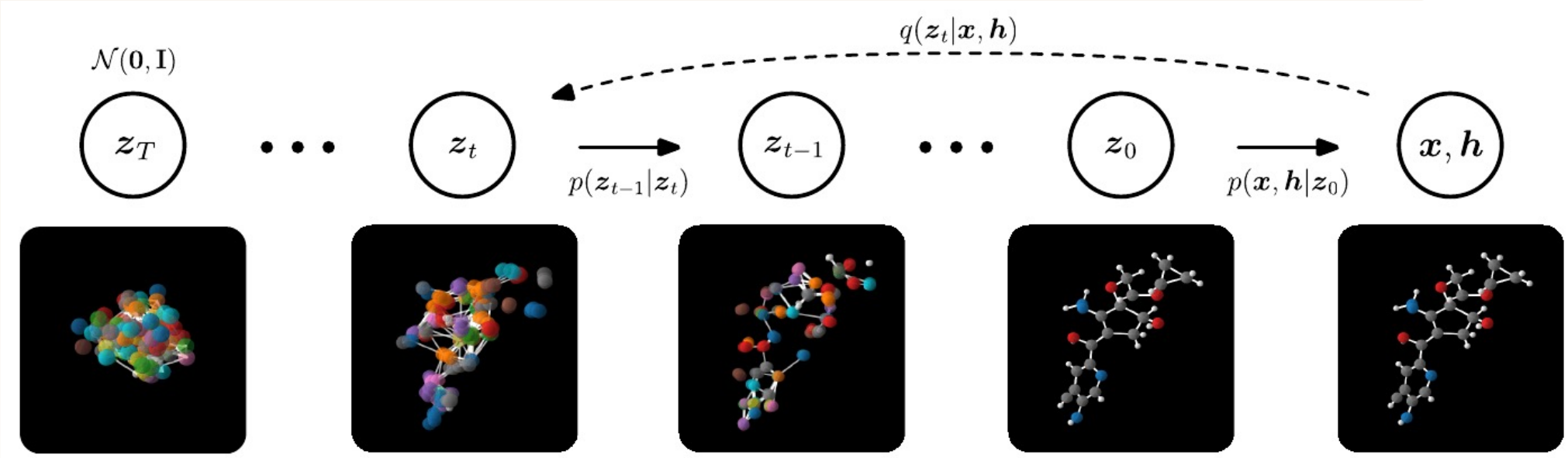
EDM defines a noising process on both node positions and features.

EDM learns the generative denoising process using an equivariant neural network.



Mila

EDM Noising Process



Noising operates on atom coordinates x_i with atom features h_i.

$$q(z_t \vee x, h) = N_{xh}(z_t \vee \alpha_t[x,h], \sigma_t^2 I)$$

$$N_x(z_t^x \vee \alpha_t x, \sigma_t^2 I) \cdot N_h\left(z_t^h \vee \alpha_t h, \sigma_t^2 I\right)$$

Mila

EDM Denoising Process

Variables x, h obtained by neural network approximation $\hat{x}, \hat{h}$.

$$p(z_s \vee z_t) = q(z_s \vee \hat{x}, \hat{h}, z_t) = N_{xh}(z_s \vee \mu_{t\to s}([\hat{x}, \hat{h}], z_t), \sigma^2_{t\to s} I)$$

$$\hat{x} = \phi(z_t^x, t), \hat{h} = \phi(z_t^h, t)$$

EDM use the noise parametrization to obtain $\hat{x}, \hat{h}$.

$$\hat{\epsilon}_t = [\hat{\epsilon}_t^x, \hat{\epsilon}_t^h] = \phi(z_t^x, z_t^h, t)$$

$$[\hat{x}, \hat{h}] = f(\hat{\epsilon}_t) = \frac{z_t}{\alpha_t} - \epsilon_t \cdot \frac{\sigma_t}{\alpha_t}$$

Mila

EDM Denoising Process

$\hat{\epsilon}_t$ are computed by equivariant neural networks , ϕ

$$R\hat{\epsilon}_t = \phi(Rz_t, t)$$

$$\hat{\epsilon}_t^x = EGNN\left(z_t^x, \left[z_t^h, \frac{t}{T}\right]\right) - [z_t^x, 0]$$

Note: the outputs have to lie on a zero center of gravity subspace.
The mean of the denoising equation rotates,

$$R\hat{x} = \frac{Rz_t^x}{\alpha_t} - \frac{R\hat{\epsilon}_t^x \sigma_t}{\alpha_t}$$

Mila

EDM Denoising Process

Algorithm 2 Sampling from EDM

Sample $z_T \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
for t in $T, T-1, \ldots, 1$ where $s = t-1$ **do**
  Sample $\epsilon \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$
  Subtract center of gravity from $\epsilon^{(x)}$ in $\epsilon = [\epsilon^{(x)}, \epsilon^{(h)}]$
  $z_s = \frac{1}{\alpha_{t|s}} z_t - \frac{\sigma^2_{t|s}}{\alpha_{t|s}\sigma_t} \cdot \phi(z_t, t) + \sigma_{t \to s} \cdot \epsilon$
end for
Sample $x, h \sim p(x, h | z_0)$

To sample from the model, one first samples $z_T \sim N_{xh}(0, I)$

and then iteratively samples $z_{t-1} \sim p(z_{t-1} \vee z_t)$

and then finally samples $x, h \sim p(x, h \vee z_0)$

Mila

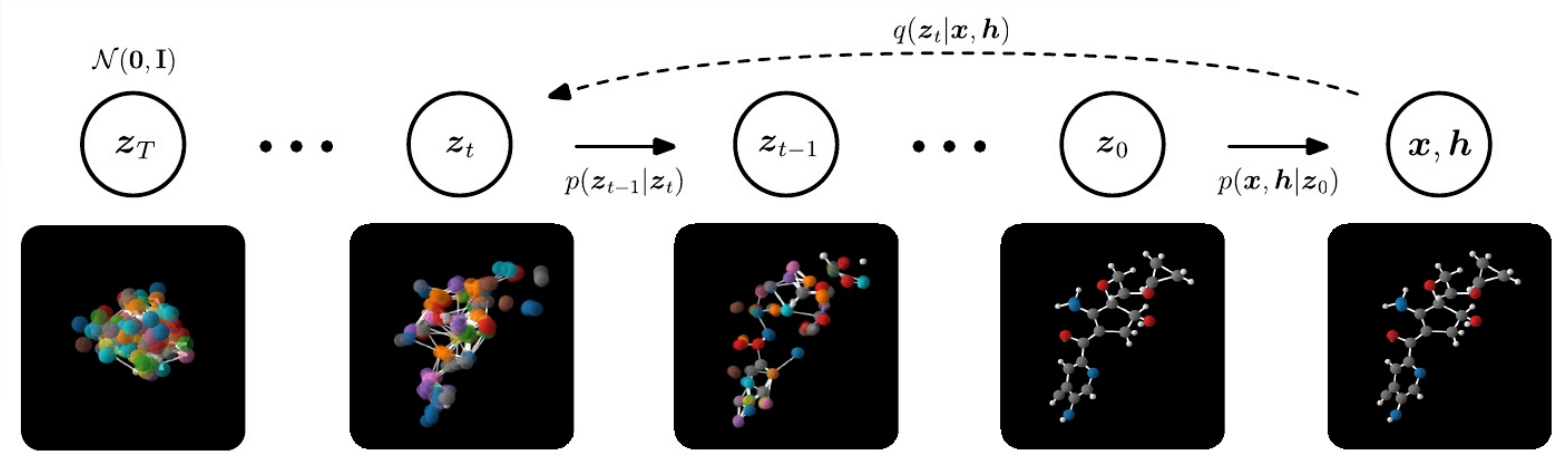
Categorical Atom Feature

Integer atom type representation is unnatural and introduces bias, use one-hot representation,

$$h^{atom} \Rightarrow h^{onehot}, h_{ij}^{onehot} = 1_{h_i = atom_j}$$

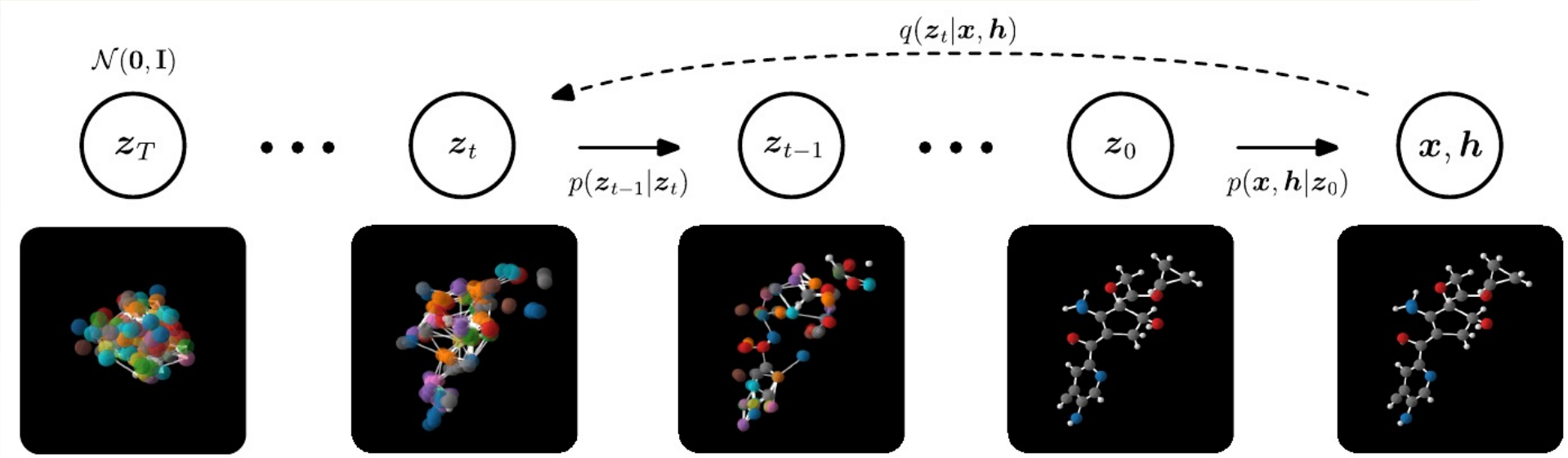
The noising process follows,

$$q\left(z_t \vee h^{onehot}\right) = N\left(z_t \vee \alpha_t h^{onehot}, \sigma_t^2 I\right)$$



Mila

Scaling Feature



Input to EDM model,

$$[x, 0.25h^{oenhot}, 0.1h^{charge}]$$

Mila

Number of Atoms

To work with different sizes, EDM compute the categorical distribution of molecule sizes. $p(M)$

First sample molecule size , M

$$M \sim P(M)$$

Then sample x, h

$$x, h \sim P(x, h \vee M)$$

04

Result

QM9

130k small molecules with up to 9 heavy atoms (hydrogen not included).

They train EDM to unconditionally generate molecules with 3-dimensional coordinates, atom types (H, C, N, O, F) and integer-valued atom charges.

100K/18K/13K samples for train/val/test partitions.

QM9

Table 1. Neg. log-likelihood $-\log p(\mathbf{x}, \mathbf{h}, M)$, atom stability and molecule stability with standard deviations across 3 runs on QM9, each drawing 10000 samples from the model.

# Metrics	NLL	Atom stable (%)	Mol stable (%)
E-NF	-59.7	85.0	4.9
G-Schnet	N.A	95.7	68.1
GDM	-94.7	97.0	63.2
GDM-aug	-92.5	97.6	71.6
EDM (ours)	**-110.7**$\pm$1.5	**98.7**$\pm$0.1	**82.0**$\pm$0.4
Data		99.0	95.2

atom stability (the proportion of atoms that have the right valency) and
molecule stability (the proportion of generated molecules for which all atoms are stable).

Mila

QM9

Table 2. Validity and uniqueness over 10000 molecules with standard deviation across 3 runs. Results marked (*) are not directly comparable, as they do not use 3D coordinates to derive bonds. H: model hydrogens explicitly

Method	H	Valid (%)	Valid and Unique (%)
Graph VAE (*)		55.7	42.3
GTVAE (*)		74.6	16.8
Set2GraphVAE (*)		59.9±1.7	56.2±1.4
EDM (ours)		**97.5**±0.2	**94.3**±0.2
E-NF	✓	40.2	39.4
G-Schnet	✓	85.5	80.3
GDM-aug	✓	90.4	89.5
EDM (ours)	✓	**91.9**±0.5	**90.7**±0.6
Data	✓	97.7	97.7

validity and
uniqueness of the generated compounds.

Mila

QM9

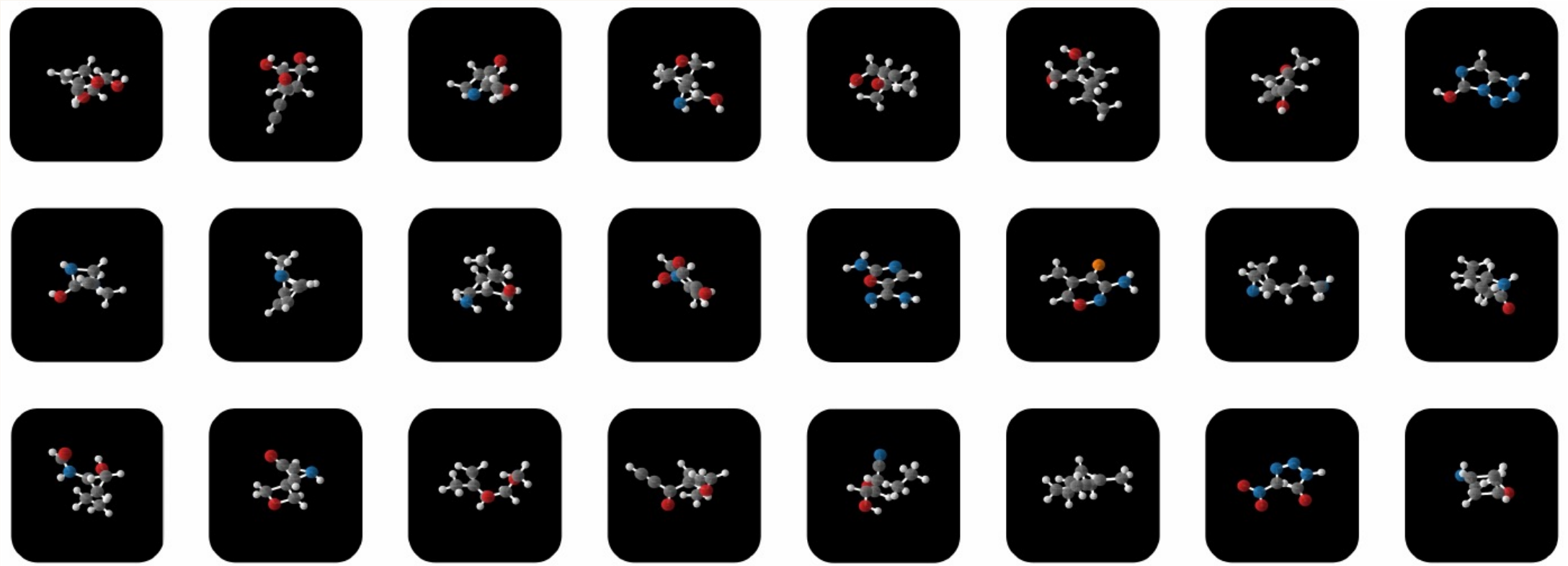


Figure 6. Random samples taken from the EDM trained on QM9.

Samples from EDM trained on QM9

Mila

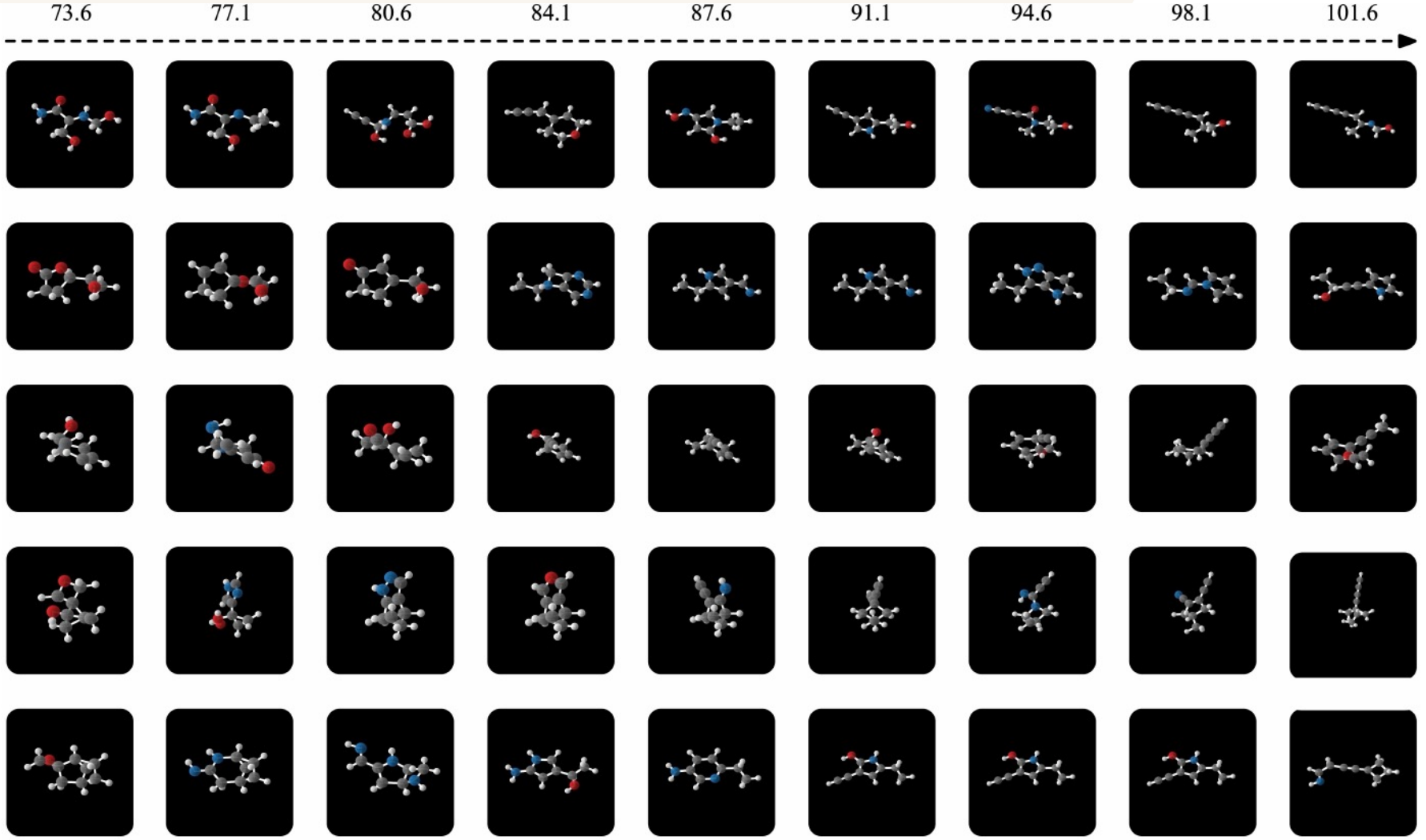
QM9 Conditional Molecule Generation

Table 3. Mean Absolute Error for molecular property prediction by a EGNN classifier ϕ_c on a QM9 subset, EDM generated samples and two different baselines "Naive (U-bounds)" and "# Atoms".

Task	α	$\Delta\varepsilon$	$\varepsilon_{\text{HOMO}}$	$\varepsilon_{\text{LUMO}}$	μ	C_v
Units	Bohr^3	meV	meV	meV	D	$\frac{\text{cal}}{\text{mol}}\text{K}$
Naive (U-bound)	9.01	1470	645	1457	1.616	6.857
#Atoms	3.86	866	426	813	1.053	1.971
EDM	2.76	655	356	584	1.111	1.101
QM9 (L-bound)	0.10	64	39	36	0.043	0.040

QM9 Conditional Molecule Generation

Molecules generated by Conditional EDM when interpolating among different polarizability values (from left to right).



Mila

DRUGS

430k small molecules with up to 181 atoms and 44.4 atoms on average.

For each molecule, many conformers are given along with their energy.

Retain the 30 lowest energy conformations for each molecule

DRUGS

Table 4. Neg. log-likelihood, atom stability and Wasserstein distance between generated and training set energy distributions.

# Metrics	NLL	Atom stability (%)	$\mathcal{W}$
GDM	− 14.2	75.0	3.32
GDM-aug	− 58.3	77.7	4.26
EDM	**−137.1**	**81.3**	**1.41**
Data		86.5	0.0

DRUGS

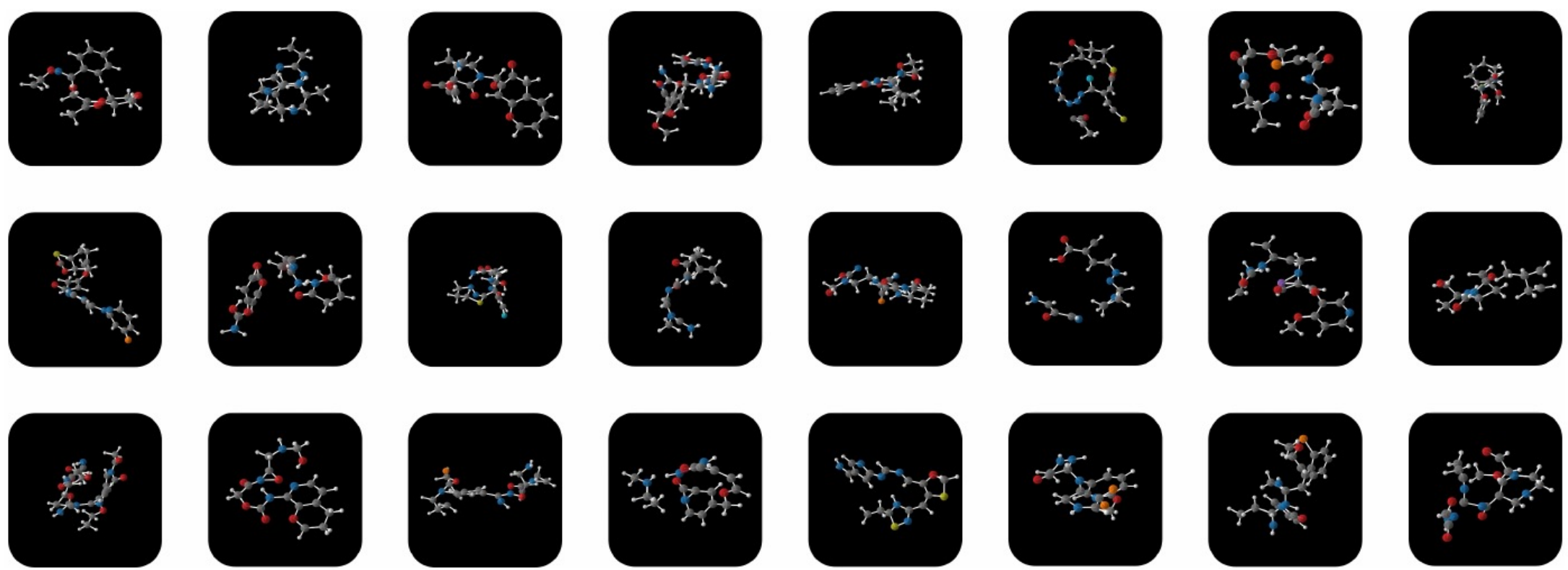


Figure 7. Random samples taken from the EDM trained on geom drugs. While most samples are very realistic, we observe two main failure cases: some molecules that are disconnected, and some that contain long rings. We note that the model does not feature any regularization to prevent these phenomena.

Samples from EDM trained on DRUGS

Mila

DRUGS

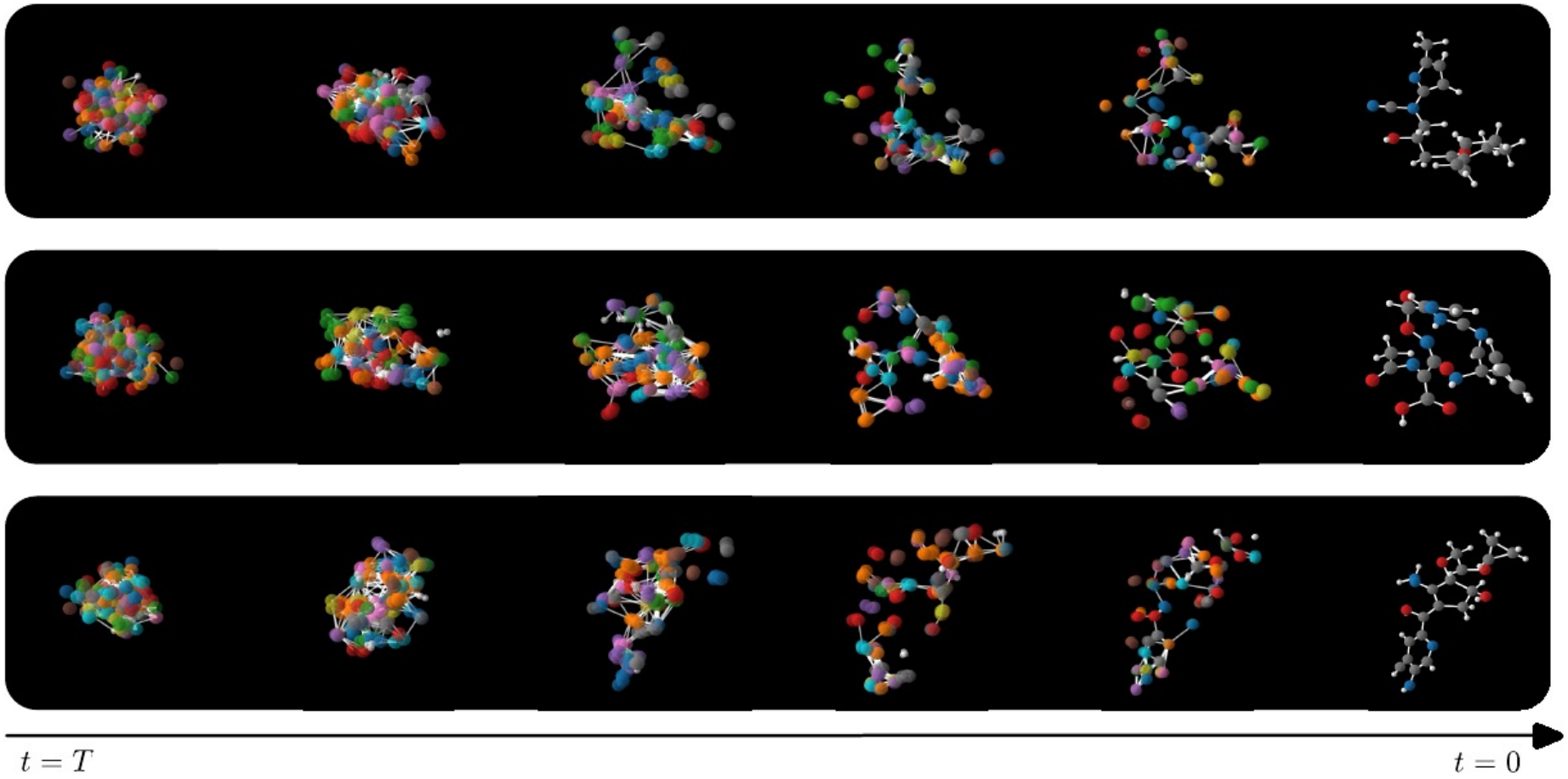


Figure 8. Selection of sampling chains at different steps from a model trained on GEOM-Drugs. The final column shows the resulting sample from the model.

Mila

Ablation on Scaling Feature

Table 10. Ablation study on the scaling of features of the EDM. Comparing our proposed scaling to no scaling.

# Metrics	Scaling	NLL	Atom stable (%)	Mol stable (%)
EDM (ours)	$[\boldsymbol{x}, 1.00\ \boldsymbol{h}^{\text{onehot}}, 1.0\ \boldsymbol{h}^{\text{atom charge}}]$	-103.4	95.7	46.9
EDM (ours)	$[\boldsymbol{x}, 0.25\ \boldsymbol{h}^{\text{onehot}}, 0.1\ \boldsymbol{h}^{\text{atom charge}}]$	**-110.7**± 1.5	**98.7**± 0.1	**82.0**± 0.4
Data			99.0	95.2

05

Conclusion

Conclusion

Advantage:

Does not require atom types.

Does not require atom orderings.

Efficient and effective.

Disadvantage:

No significant improvment.