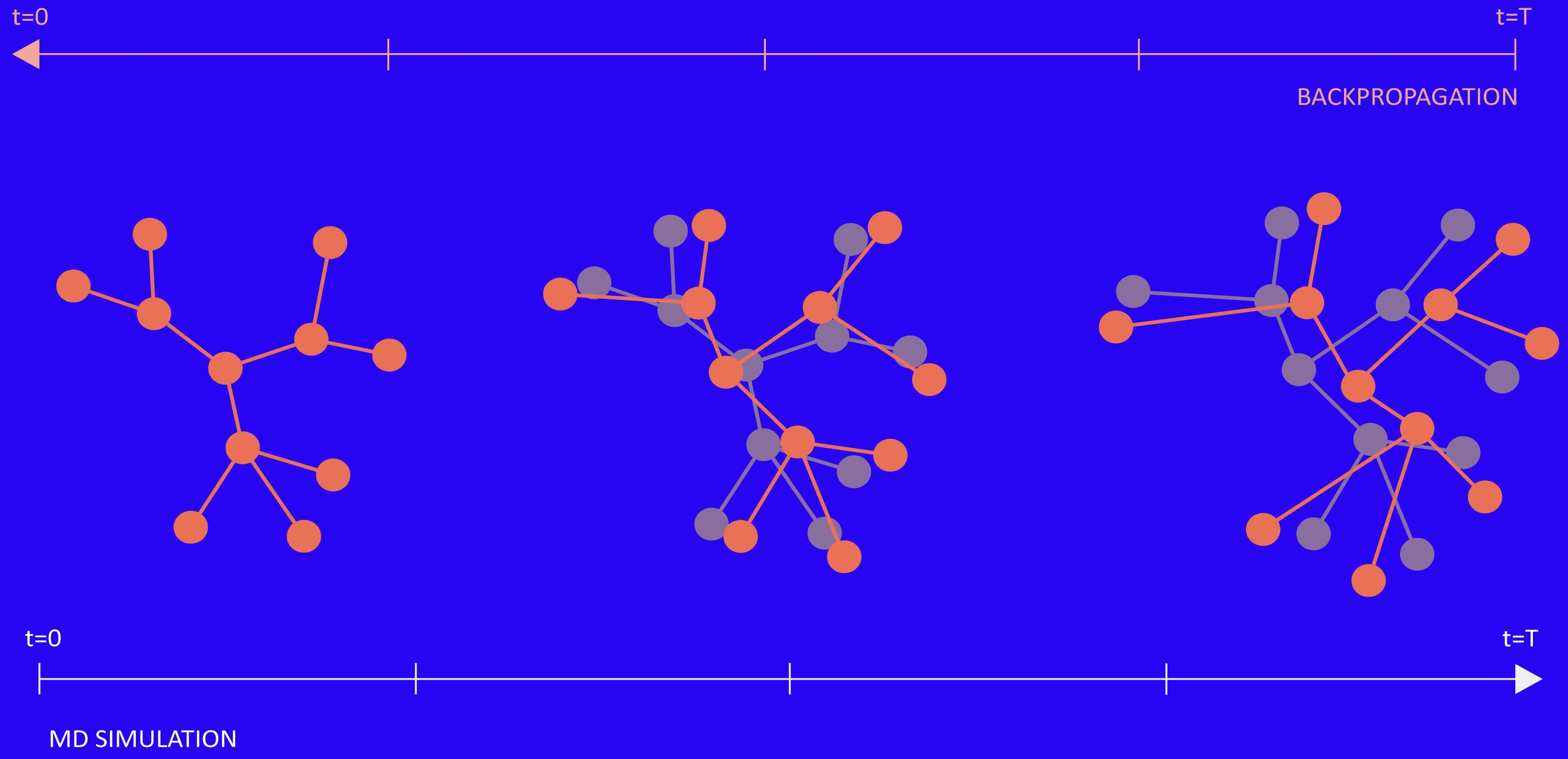
AI-driven structure-based drug discovery using generative equivariant diffusion

AIDD TALK| 02-2023

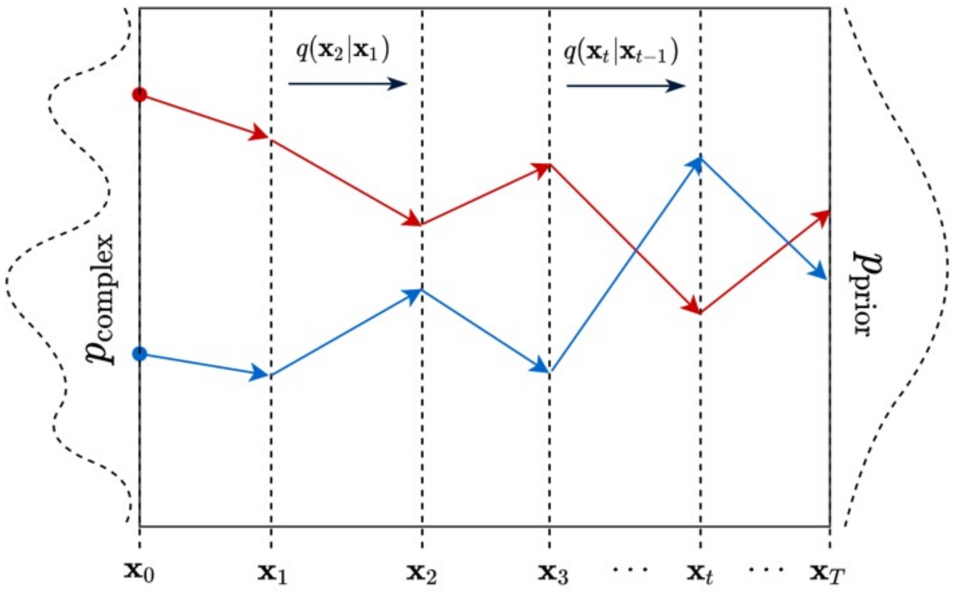
Pfizer

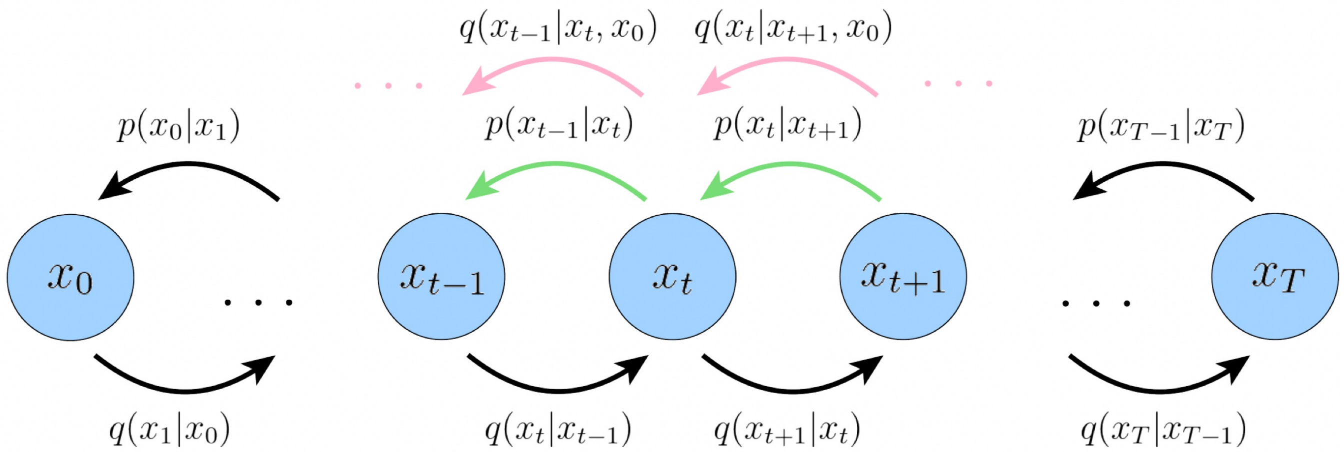
Universitat Pompeu Fabra Barcelona

Diffusion?
Equivariance?


t=0
t=T
BACKPROPAGATION
t=0
t=T
MD SIMULATION

Overview: Diffusion



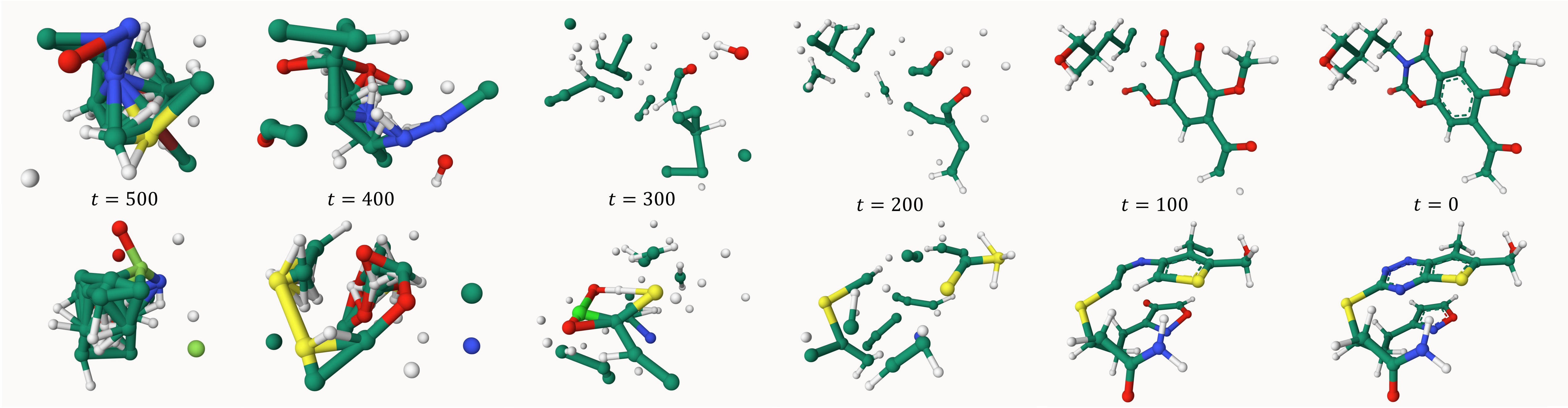


$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t|\sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \quad \text{and} \quad q(\mathbf{c}_t|\mathbf{c}_0) = \mathcal{C}(\mathbf{c}_t|\bar{\alpha}_t\mathbf{c}_0 + (1-\bar{\alpha}_t)\tilde{\mathbf{c}})$$

$$L_{t,\epsilon} = w(t)||\epsilon_t - \hat{\epsilon}_\theta(x_t, t)||^2 \quad \text{and} \quad L_{t,x_0} = w(t) \cdot l_d(x_0, \hat{x}_\theta(x_t, t); \lambda_m)$$

$$L_{t-1} = w_s(t)\Big(\lambda_x||\mathbf{X}_0 - \hat{\mathbf{X}}_0||^2 + \lambda_h \mathrm{CE}(\mathbf{H}_0, \hat{\mathbf{H}}_0) + \lambda_e \mathrm{CE}(\mathbf{E}_0, \hat{\mathbf{E}}_0)\Big)$$

Overview: Diffusion



EQGAT-diff

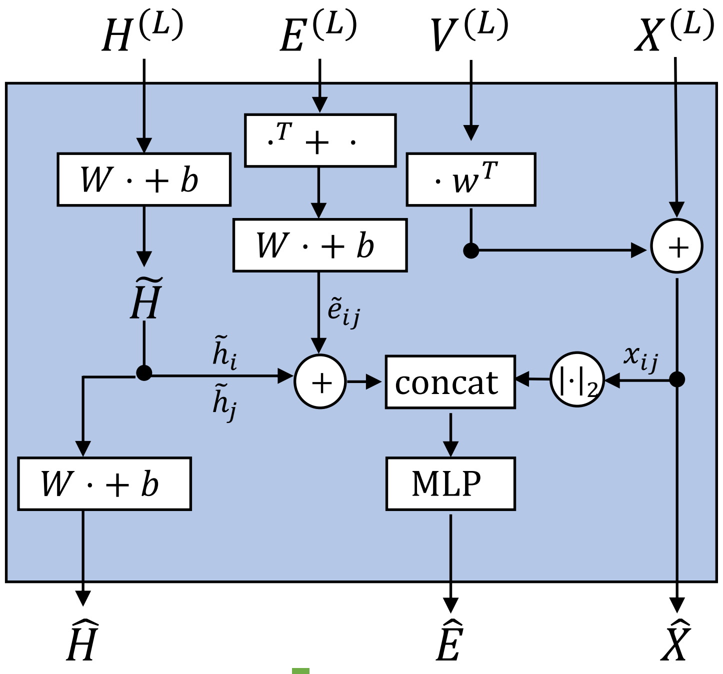
Model architecture: EQGAT-diff

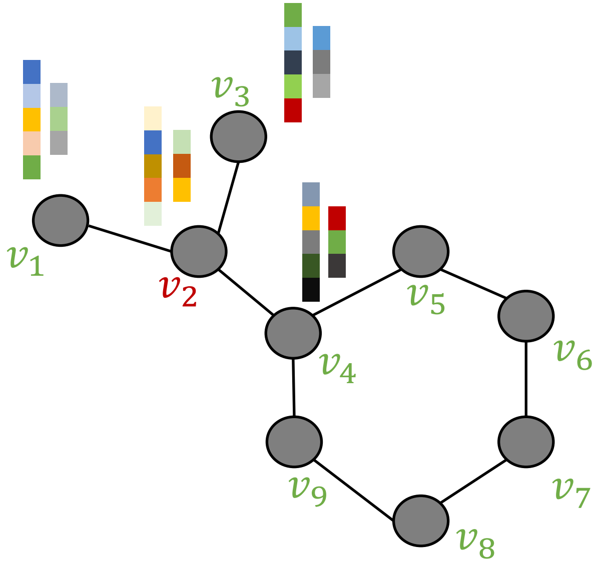
$$\mathbf{m}_{ji}^{(l)} = \mathrm{MLP}([\mathbf{h}_j^{(l)}; \mathbf{h}_i^{(l)}; \mathbf{W}_{e_0}^{(l)}\mathbf{e}_{ji}^{(l)}; d_{ji}^{(l)}; d_j^{(l)}; d_i^{(l)}; \mathbf{p}_j^{(l)} \cdot \mathbf{p}_i^{(l)}]),$$

$$\mathbf{h}_i^{(l+1)} = \mathbf{h}_i^{(l)} + \sum_j \frac{\exp(\mathbf{a}_{ji}^{(l)})}{\sum_{j'} \exp(\mathbf{a}_{j'i}^{(l)})} \mathbf{W}_h^{(l)} \mathbf{h}_j^{(l)} \quad \text{and} \quad \mathbf{e}_{ji}^{(l+1)} = \mathbf{W}_{e_1}^{(l)} \sigma(\mathbf{e}_{ji}^{(l)} + \mathbf{d}_{ji}^{(l)}),$$

$$\mathbf{v}_i^{(l+1)} = \mathbf{v}_i^{(l)} + \frac{1}{N} \sum_j \mathbf{x}_{ji,n} \otimes \mathbf{b}_{ji}^{(l)} + (\mathbf{1} \otimes \mathbf{c}_{ji}^{(l)}) \odot \mathbf{v}_j^{(l+1)} \mathbf{W}_v^{(l)},$$

$$\mathbf{x}_i^{(l+1)} = \mathbf{x}_i^{(l)} + \frac{1}{N} \sum_j s_{ji}^{(l)} \mathbf{x}_{ji,n}^{(l)},$$





Loss weighting

$$w_s(t) = \min(0.05,\ \max(1.5,\ \mathrm{SNR}(t)))$$

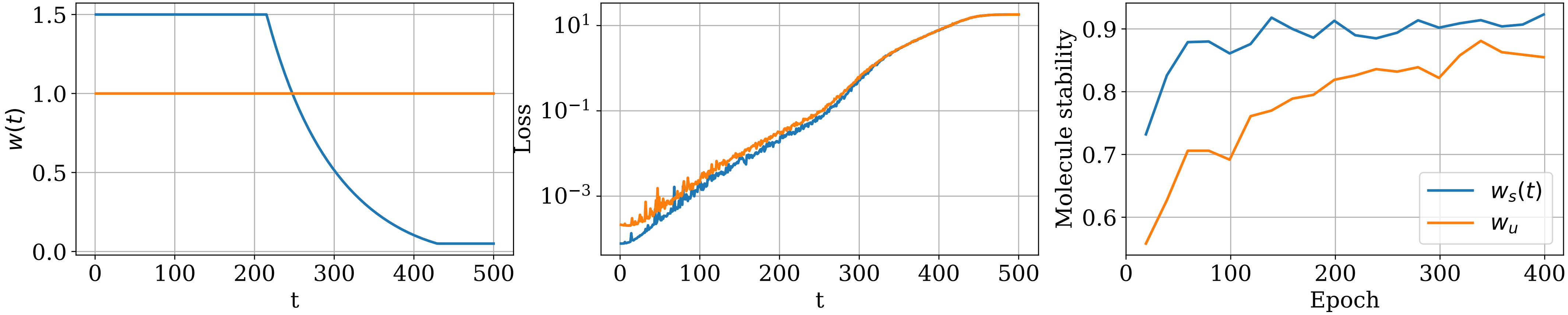


Table 1: Comparison of EQGAT-diff on QM9 and GEOM-Drugs trained with w_u or $w_s(t)$ loss-weighting. We report the mean values over five runs of selected evaluation metrics with the margin of error for the 95% confidence level given as subscripts. The best results are in bold.

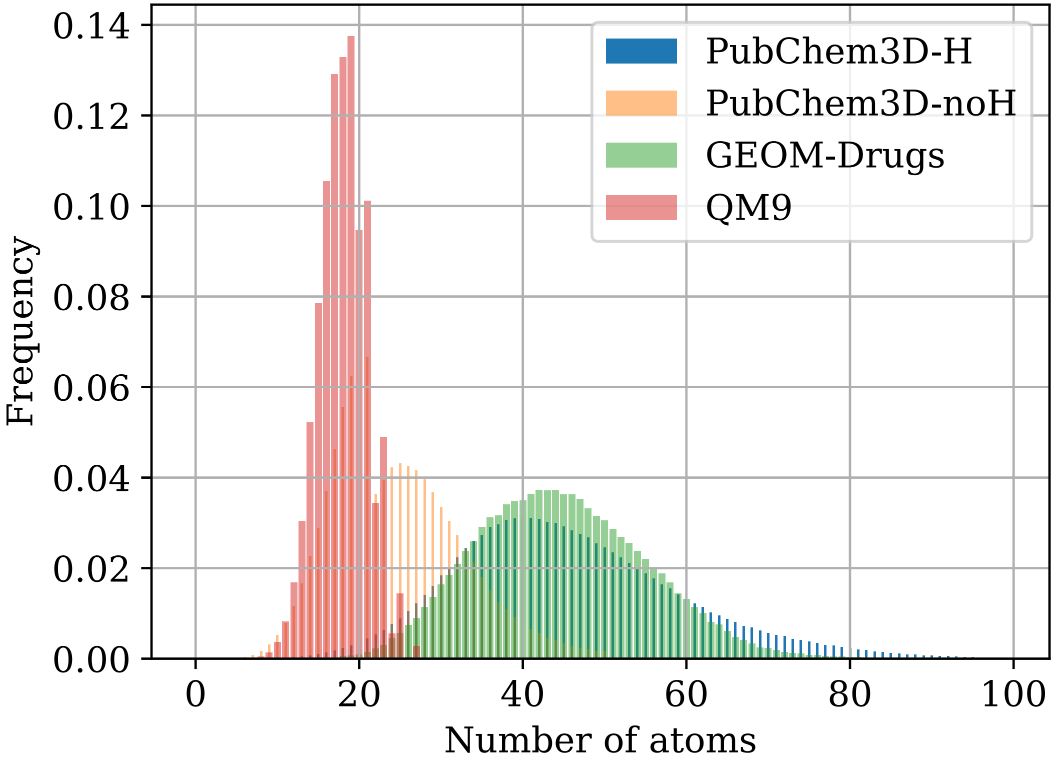
	QM9			GEOM-Drugs		
Weighting	Mol. Stability ↑	Validity ↑	Connect. Comp. ↑	Mol. Stability ↑	Validity ↑	Connect. Comp. ↑
w_u	$97.39_{\pm 0.23}$	$97.99_{\pm 0.20}$	$99.70_{\pm 0.03}$	$87.59_{\pm 0.19}$	$71.44_{\pm 0.22}$	$86.57_{\pm 0.33}$
$w_s(t)$	$\mathbf{98.68}_{\pm 0.11}$	$\mathbf{98.96}_{\pm 0.07}$	$\mathbf{99.94}_{\pm 0.03}$	$\mathbf{91.60}_{\pm 0.14}$	$\mathbf{84.02}_{\pm 0.19}$	$\mathbf{95.08}_{\pm 0.12}$

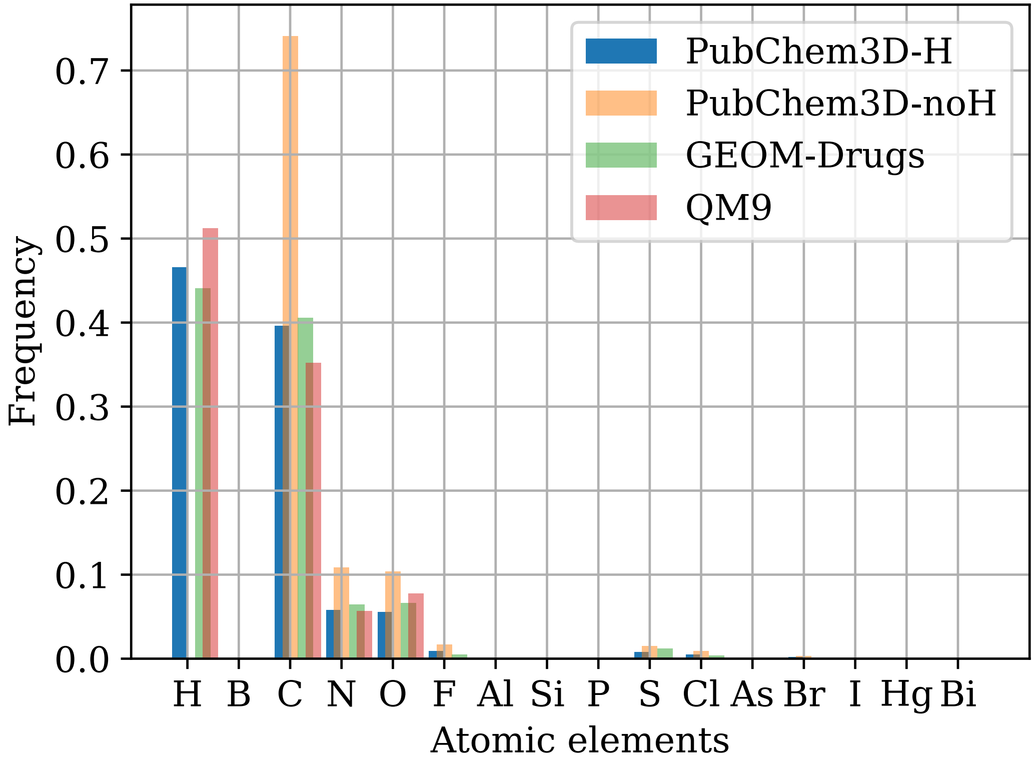
Design Space

Table 2: Overall performance of EQGAT-diff on QM9 and GEOM-Drugs for discrete and continuous diffusion as well as noise (ϵ) and data learning (x_0). Discrete or continuous diffusion is denoted as 'disc' and 'cont', respectively, given as subscripts, ϵ- and x_0-parameterization as superscripts. We report mean values over five sampling runs with 95% confidence intervals as subscripts. The best results are in bold.

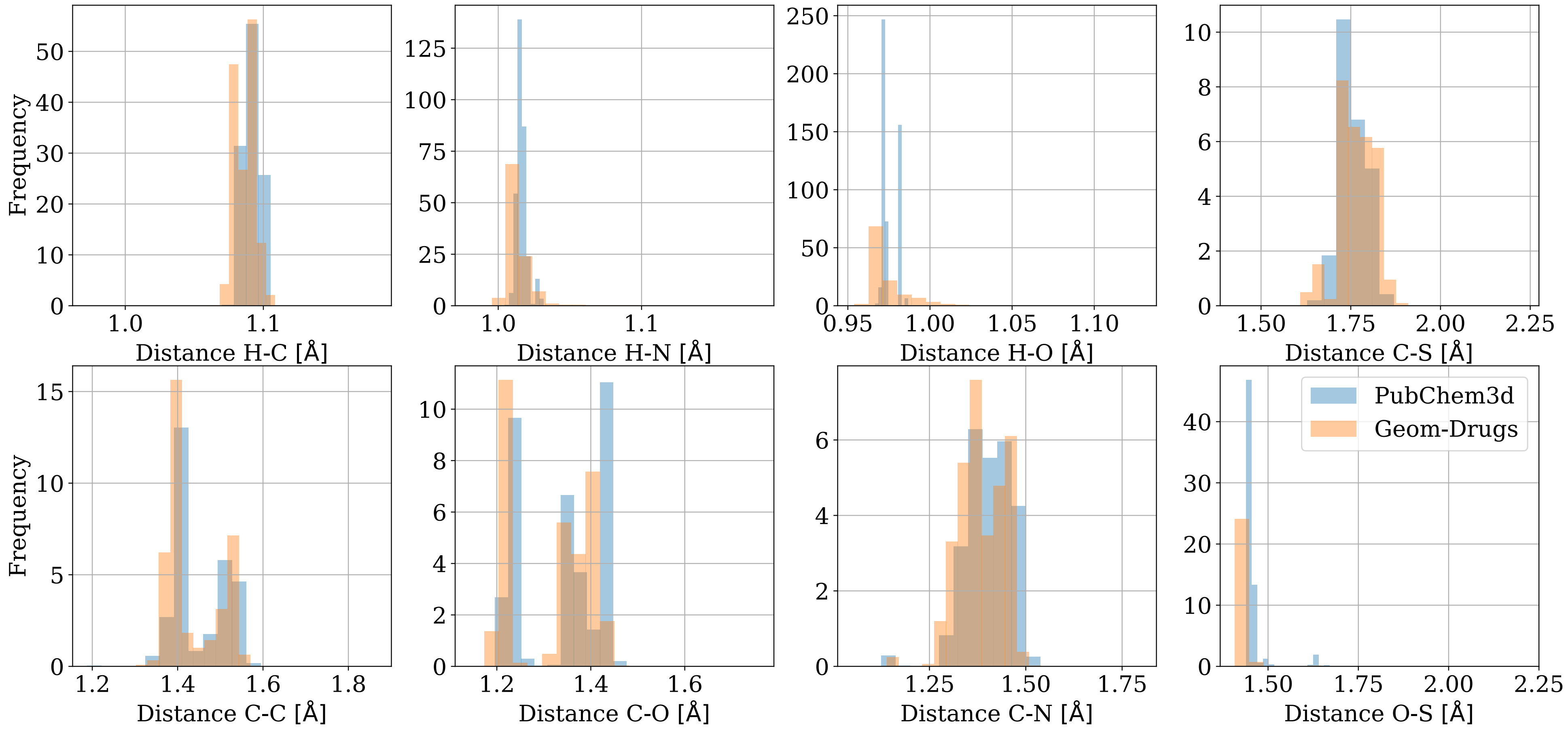
Dataset	**QM9**			**GEOM-Drugs**		
Model	EQGAT^{x0}_{disc}	EQGAT^{x0}_{cont}	$\text{EQGAT}^{\epsilon}_{cont}$	EQGAT^{x0}_{disc}	EQGAT^{x0}_{cont}	$\text{EQGAT}^{\epsilon}_{cont}$
Mol. Stab. ↑	**98.68**$_{\pm 0.11}$	96.45$_{\pm 0.17}$	96.18$_{\pm 0.16}$	**91.60**$_{\pm 0.14}$	90.46$_{\pm 0.09}$	85.19$_{\pm 0.72}$
Atom. Stab ↑	**99.92**$_{\pm 0.00}$	99.79$_{\pm 0.01}$	99.68$_{\pm 0.02}$	**99.72**$_{\pm 0.01}$	**99.73**$_{\pm 0.01}$	99.32$_{\pm 0.04}$
Validity ↑	**98.96**$_{\pm 0.07}$	96.79$_{\pm 0.15}$	97.04$_{\pm 0.17}$	**84.02**$_{\pm 0.19}$	80.96$_{\pm 0.38}$	79.13$_{\pm 0.58}$
Connect. Comp. ↑	**99.94**$_{\pm 0.03}$	99.82$_{\pm 0.05}$	99.71$_{\pm 0.03}$	**95.08**$_{\pm 0.12}$	93.30$_{\pm 0.21}$	94.10$_{\pm 0.48}$
Novelty ↑	64.03$_{\pm 0.24}$	60.96$_{\pm 0.54}$	**73.40**$_{\pm 0.32}$	**99.87**$_{\pm 0.04}$	**99.83**$_{\pm 0.04}$	99.82$_{\pm 0.0}$
Uniqueness ↑	**100.00**$_{\pm 0.00}$	**100.0**$_{\pm 0.00}$	**100.00**$_{\pm 0.00}$	**100.00**$_{\pm 0.00}$	**100.00**$_{\pm 0.00}$	**100.00**$_{\pm 0.00}$
Diversity ↑	91.72$_{\pm 0.02}$	91.51$_{\pm 0.03}$	**91.89**$_{\pm 0.03}$	**89.00**$_{\pm 0.03}$	88.87$_{\pm 0.04}$	88.97$_{\pm 0.05}$
KL Divergence ↑	**91.36**$_{\pm 0.29}$	**91.41**$_{\pm 0.54}$	88.97$_{\pm 0.31}$	87.17$_{\pm 0.34}$	87.35$_{\pm 0.35}$	**87.70**$_{\pm 0.58}$
Train Similarity ↓	0.076$_{\pm 0.00}$	0.076$_{\pm 0.00}$	**0.075**$_{\pm 0.00}$	**0.113**$_{\pm 0.00}$	0.114$_{\pm 0.00}$	0.114$_{\pm 0.00}$
AtomsTV $[10^{-2}]$ ↓	1.0$_{\pm 0.00}$	2.0$_{\pm 0.00}$	2.7$_{\pm 0.00}$	3.4$_{\pm 0.10}$	3.6$_{\pm 0.10}$	**2.9**$_{\pm 0.20}$
BondsTV $[10^{-2}]$ ↓	1.2$_{\pm 0.00}$	1.8$_{\pm 0.00}$	1.2$_{\pm 0.00}$	**2.4**$_{\pm 0.00}$	**2.4**$_{\pm 0.00}$	**2.4**$_{\pm 0.00}$
ValencyW$_1$ $[10^{-2}]$ ↓	0.6$_{\pm 0.10}$	1.9$_{\pm 0.00}$	0.9$_{\pm 0.00}$	1.2$_{\pm 0.10}$	1.9$_{\pm 0.10}$	1.6$_{\pm 0.00}$
BondLenghtsW$_1$ $[10^{-2}]$ ↓	**0.2**$_{\pm 0.10}$	0.5$_{\pm 0.00}$	**0.2**$_{\pm 0.10}$	**0.2**$_{\pm 0.10}$	0.3$_{\pm 0.00}$	0.7$_{\pm 0.40}$
BondAnglesW$_1$ ↓	**0.42**$_{\pm 0.03}$	1.86$_{\pm 0.06}$	0.52$_{\pm 0.03}$	**0.92**$_{\pm 0.02}$	0.95$_{\pm 0.02}$	1.07$_{\pm 0.06}$

Large-Scale Pre-Training: PubChem3D

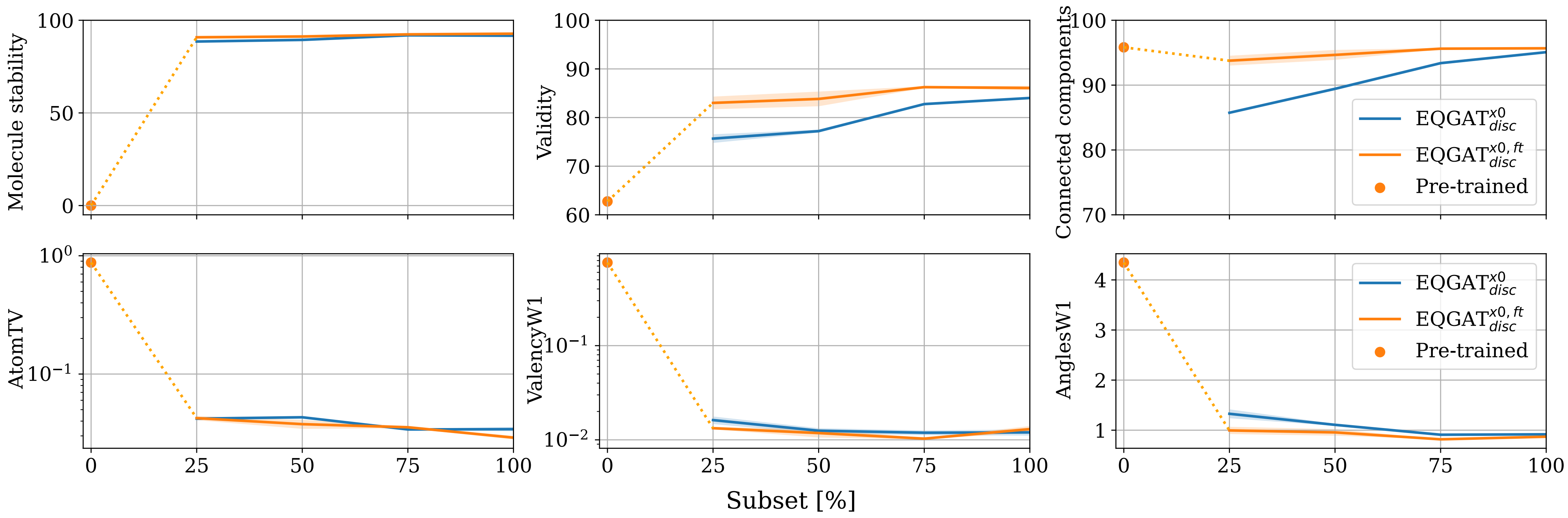




Large-Scale Pre-Training: PubChem3D



Large-Scale Pre-Training: PubChem3D

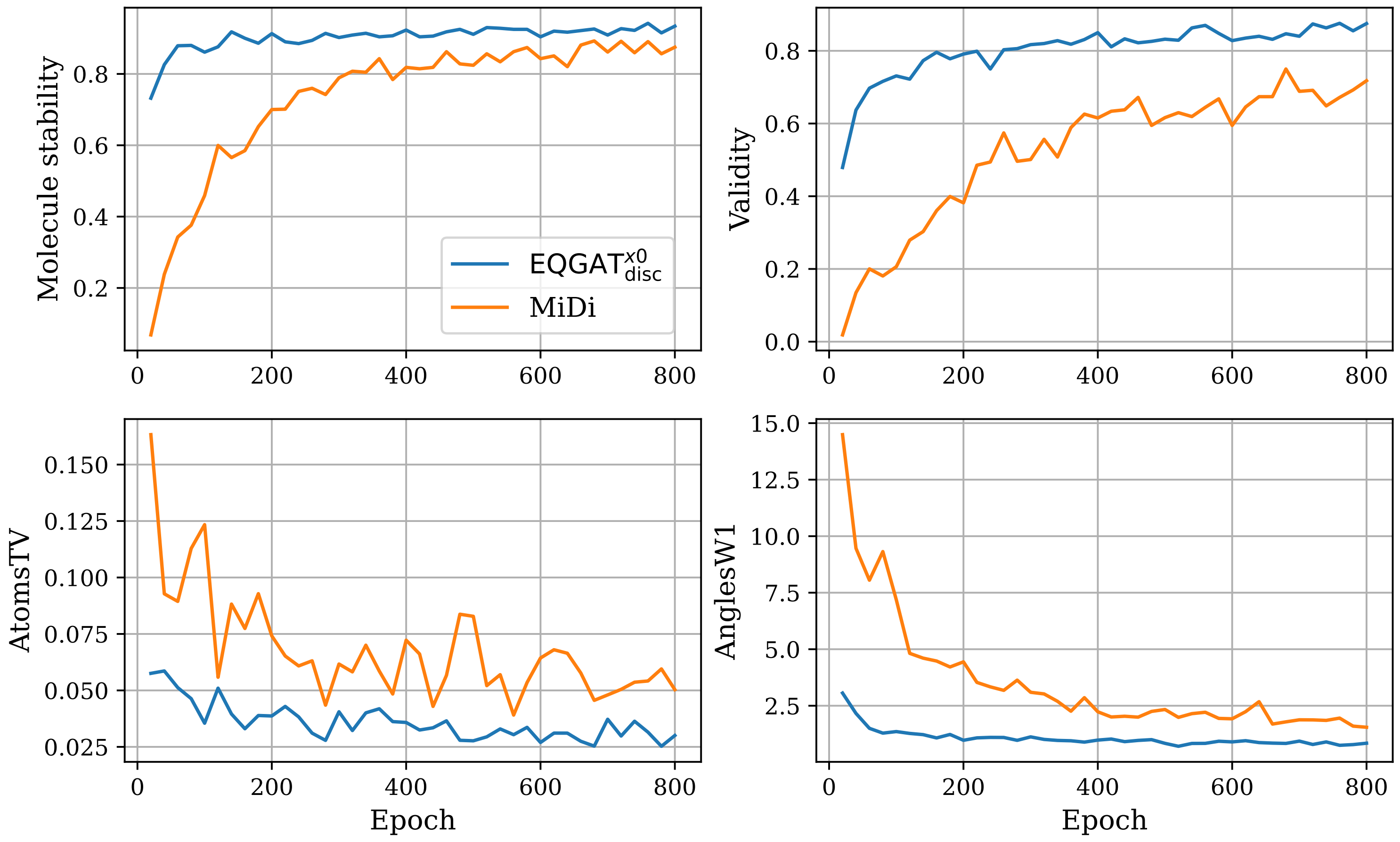


State-of-the-Art

Table 3: Comparison of $\text{EQGAT}_{\text{disc}}$ models trained for 800 epochs on GEOM-Drugs. The superscripts 'ft' and 'af' abbreviate *fine-tuned* and *additional-features*. The margin of error for the 95% confidence level is given as subscripts. We also compare EDM and the current SOTA, MiDi. Training details for MiDi are given in Appendix A.6. The best results are in bold.

Dataset	**GEOM-Drugs**					
Model	$\text{EQGAT}^{x0}_{\text{disc}}$	$\text{EQGAT}^{x0,ft}_{\text{disc}}$	$\text{EQGAT}^{x0,af}_{\text{disc}}$	$\text{EQGAT}^{x0,af,ft}_{\text{disc}}$	EDM	MiDi
Mol. Stab. ↑	$93.11_{\pm 0.31}$	$93.92_{\pm 0.13}$	$\mathbf{94.51}_{\pm 0.18}$	$\mathbf{95.01}_{\pm 0.37}$	40.3	$89.7_{\pm 0.60}$
Atom. Stab ↑	$99.79_{\pm 0.01}$	$\mathbf{99.81}_{\pm 0.01}$	$\mathbf{99.83}_{\pm 0.01}$	$\mathbf{99.84}_{\pm 0.00}$	97.8	$99.7_{\pm 0.01}$
Validity ↑	$85.86_{\pm 0.33}$	$\mathbf{88.04}_{\pm 0.17}$	$87.89_{\pm 0.31}$	$\mathbf{88.42}_{\pm 0.26}$	87.8	$70.5_{\pm 0.41}$
Connect. Comp. ↑	$\mathbf{96.32}_{\pm 0.25}$	$\mathbf{96.57}_{\pm 0.18}$	$96.36_{\pm 0.25}$	$\mathbf{96.71}_{\pm 0.20}$	41.4	$88.76_{\pm 0.55}$
Novelty ↑	$99.82_{\pm 0.05}$	$99.84_{\pm 0.02}$	$99.82_{\pm 0.05}$	$99.82_{\pm 0.03}$	**100.00**	$\mathbf{100.00}_{\pm 0.00}$
Diversity ↑	$\mathbf{89.03}_{\pm 0.03}$	$\mathbf{89.05}_{\pm 0.05}$	$\mathbf{88.98}_{\pm 0.02}$	$\mathbf{88.96}_{\pm 0.01}$	-	-
KL Divergence ↑	$87.66_{\pm 0.31}$	$87.58_{\pm 0.56}$	$\mathbf{88.38}_{\pm 0.25}$	$87.62_{\pm 0.19}$	-	-
Train Similarity ↓	$0.114_{\pm 0.0}$	$\mathbf{0.113}_{\pm 0.0}$	$0.114_{\pm 0.0}$	$0.114_{\pm 0.0}$	-	-
AtomsTV $[10^{-2}]$ ↓	$3.02_{\pm 0.08}$	$3.02_{\pm 0.10}$	$\mathbf{2.88}_{\pm 0.10}$	$\mathbf{2.91}_{\pm 0.10}$	21.2	$5.11_{\pm 0.19}$
BondsTV $[10^{-2}]$ ↓	$2.44_{\pm 0.01}$	$\mathbf{2.40}_{\pm 0.00}$	$2.42_{\pm 0.00}$	$\mathbf{2.40}_{\pm 0.00}$	4.8	$2.44_{\pm 0.00}$
ValencyW$_1$ $[10^{-2}]$ ↓	$1.18_{\pm 0.09}$	$1.20_{\pm 0.00}$	$\mathbf{0.85}_{\pm 0.12}$	$\mathbf{0.90}_{\pm 0.10}$	28.5	$2.48_{\pm 0.52}$
BondLenghtsW$_1$ $[10^{-2}]$ ↓	$0.56_{\pm 0.38}$	$\mathbf{0.10}_{\pm 0.00}$	$0.50_{\pm 0.51}$	$0.20_{\pm 0.10}$	0.2	$0.2_{\pm 0.10}$
BondAnglesW$_1$ ↓	$0.83_{\pm 0.03}$	$0.79_{\pm 0.02}$	$0.65_{\pm 0.01}$	$\mathbf{0.62}_{\pm 0.01}$	6.23	$1.73_{\pm 0.32}$

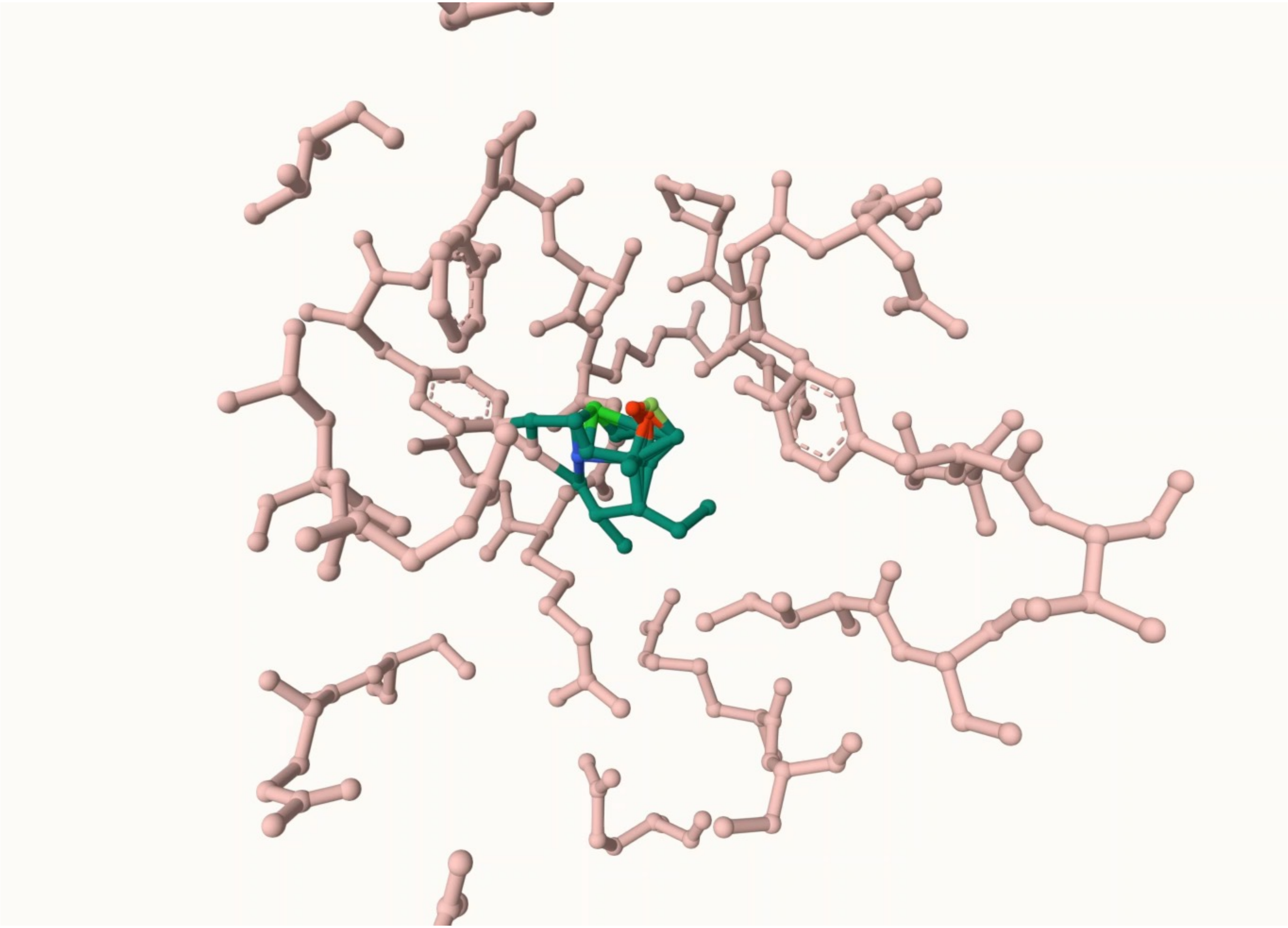
State-of-the-Art



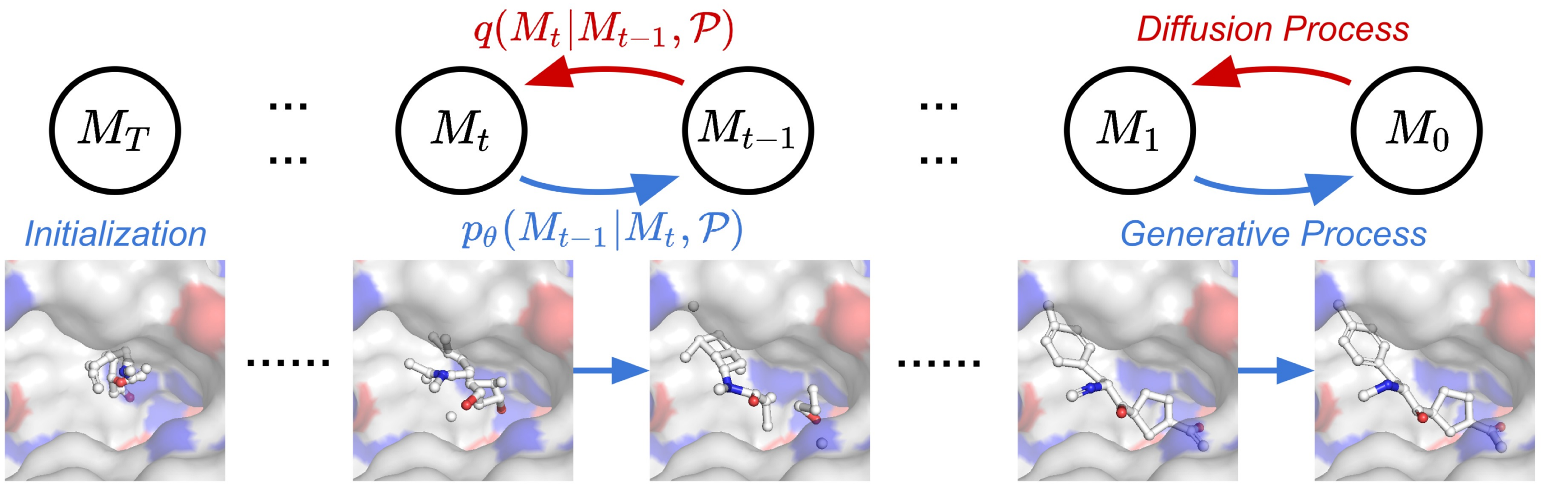
Why 3D-based modelling?

Table 9: Classifier-guidance on EQGAT-diff to shift the reverse sampling towards low or high polarizability values. We report the mean polarizability values of sampled molecules with standard deviations as subscripts.

Guidance	Polarizability
Minimization	$195.19_{\pm 4.9}$
Maximization	$400.21_{\pm 8.3}$



Target-aware de novo generation



Model	Validity ↑	Connect. Comp. ↑	BondLengths W1 $[10^{-2}]$ ↓	BondAngles W1 ↓
$\text{EQGAT}^{x0}_{disc}(w_u)$	$85.51_{\pm 0.09}$	$95.15_{\pm 0.14}$	$0.20_{\pm 0.0}$	$4.37_{\pm 0.20}$
$\text{EQGAT}^{x0}_{disc}(w_s(t))$	$89.62_{\pm 0.08}$	$97.65_{\pm 0.11}$	$0.12_{\pm 0.0}$	$2.12_{\pm 0.26}$
$\text{EQGAT}^{x0,ft}_{disc}(w_s(t))$	$\mathbf{95.65}_{\pm 0.12}$	$\mathbf{99.66}_{\pm 0.10}$	$\mathbf{0.11}_{\pm 0.0}$	$\mathbf{1.55}_{\pm 0.21}$

Model	Vina (All) ↓	Vina (Top-10%) ↓	QED ↑	SA ↑	Lipinski ↑	Diversity ↑
$\text{EQGAT}^{x0,ft}_{disc}(w_s(t))$	$\mathbf{-7.423}_{\pm 2.33}$	$-9.571_{\pm 2.14}$	$\mathbf{0.522}_{\pm 0.18}$	$\mathbf{0.697}_{\pm 0.20}$	$4.66_{\pm 0.72}$	$\mathbf{0.742}_{\pm 0.07}$
TargetDiff	$-7.318_{\pm 2.47}$	$\mathbf{-9.669}_{\pm 2.55}$	$0.483_{\pm 0.20}$	$0.584_{\pm 0.13}$	$4.594_{\pm 0.83}$	$0.718_{\pm 0.09}$
DiffSBDD-cond	$-6.950_{\pm 2.06}$	$-9.120_{\pm 2.16}$	$0.469_{\pm 0.21}$	$0.578_{\pm 0.13}$	$4.562_{\pm 0.89}$	$0.728_{\pm 0.07}$

Thanks!

Questions?

AIDD TALK| 02-2023

Pfizer

upf. Universitat Pompeu Fabra Barcelona