3D Equivariant Diffusion For Target-Aware Molecule Generation and Affinity Prediction

ICLR 2023

Jiaqi Guan[1], Wesley Wei Qian[1], Xingang Peng[2], Yufeng Su[1], Jian Peng[1], Jianzhu Ma[3]

1. Department of Computer Science, University of Illinois Urbana Champaign
2. School of Intelligence Science and Technology, Peking University
3. Institute for AI industry Research, Tsinghua University

Code: https://github.com/guanjq/targetdiff

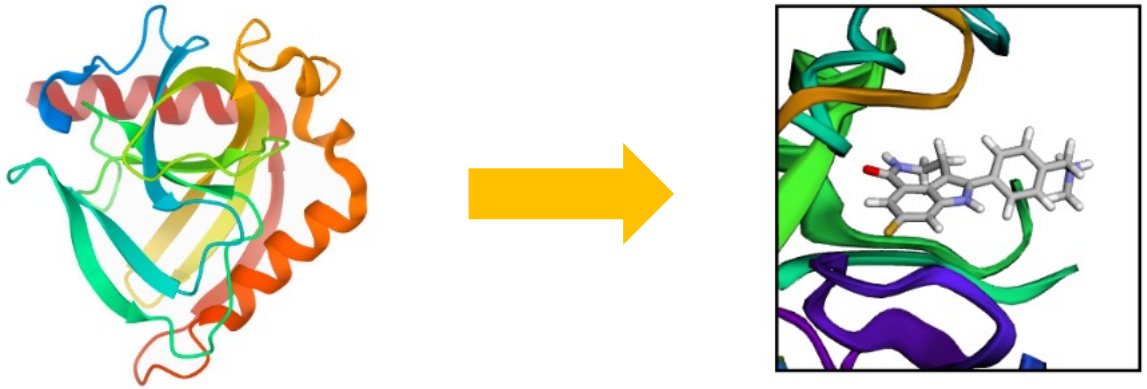
Problem Definition

Given protein binding site: $\mathcal{P} = \{(\boldsymbol{x}_P^{(i)}, \boldsymbol{v}_P^{(i)})\}_{i=1}^{N_P}$

$\boldsymbol{x}_P \in \mathbb{R}^3$ 3D atom coordinates

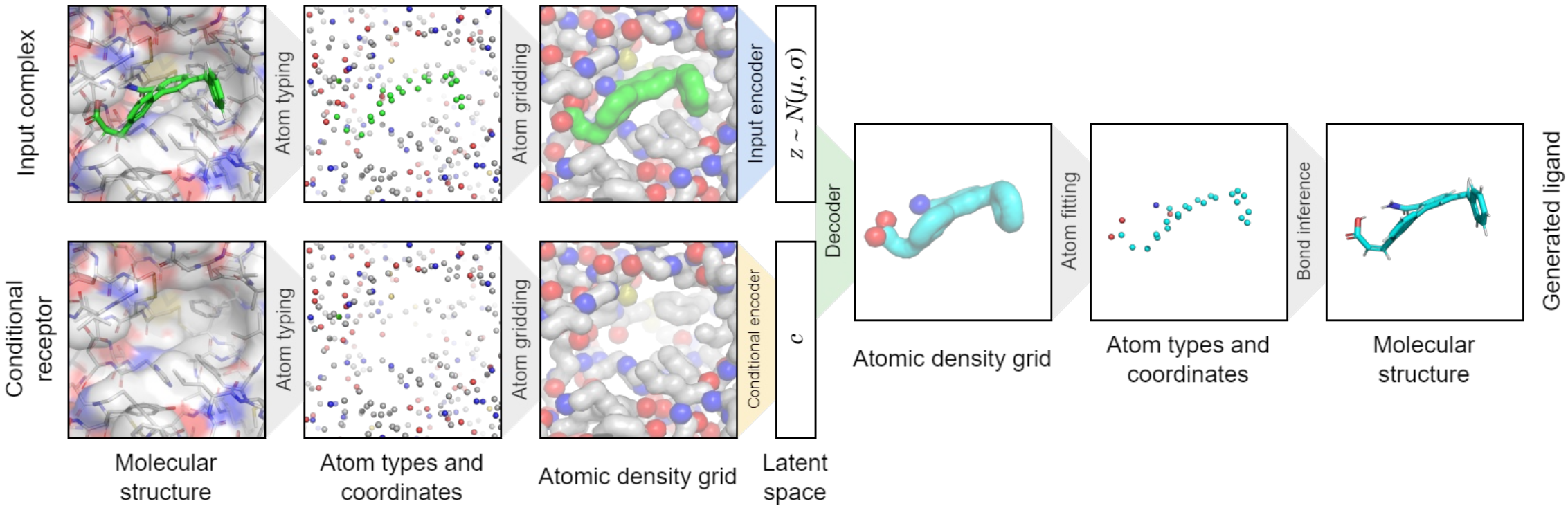
$\boldsymbol{v}_P \in \mathbb{R}^{N_f}$ Protein atom features, such as element types, amino acid types, etc.

Goal: Generate binding molecules $\mathcal{M} = \{(\boldsymbol{x}_L^{(i)}, \boldsymbol{v}_L^{(i)})\}_{i=1}^{N_M}$



Related Work

liGAN (Ragoza et al. 2022) : a conditional VAE model



Problem:

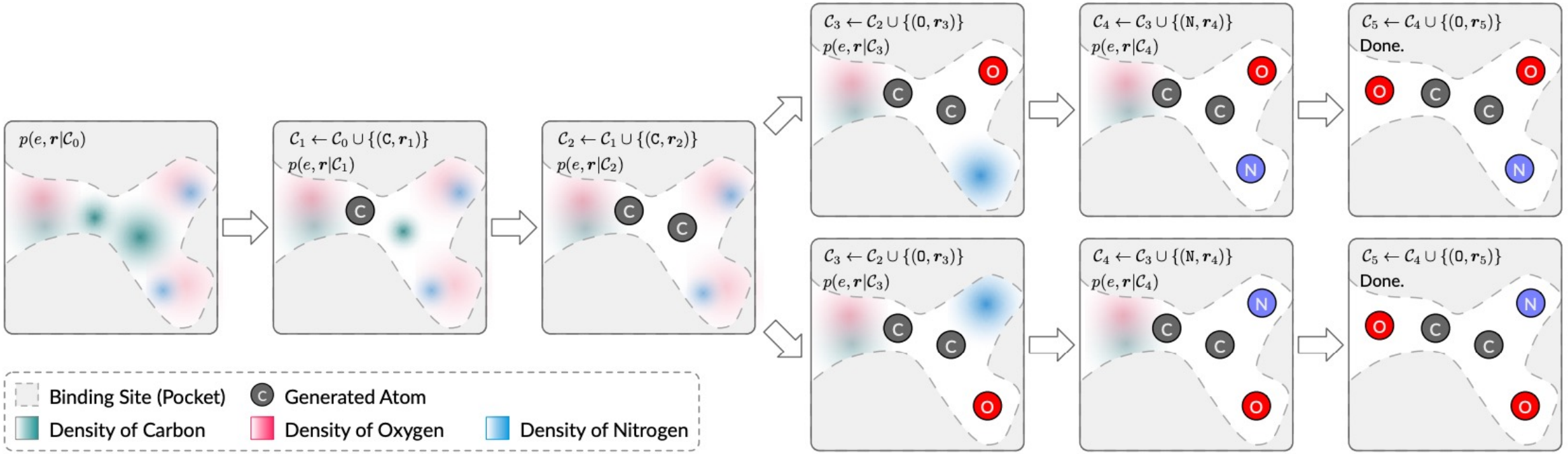
- Not rotational equivariant
- Voxelization operation will lead to poor scalability

Matthew Ragoza, Tomohide Masuda, and David Ryan Koes. Generating 3D molecules conditional on receptor binding sites with deep generative models. *Chem Sci*, 13:2701–2713, Feb 2022.

Related Work

Autoregressive Model (Luo et al., 2021): Learns $p(\boldsymbol{x}, v|\mathcal{P})$

3D Masked Language Modeling training + Autoregressive Sampling

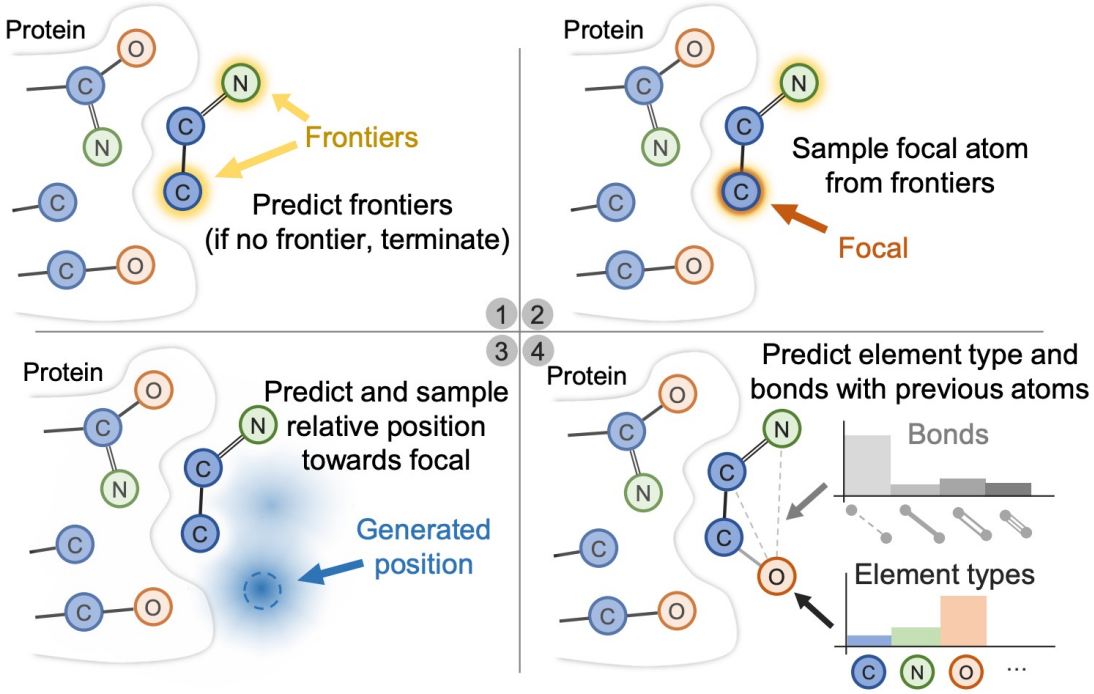


Shitong Luo, Jiaqi Guan, Jianzhu Ma, and Jian Peng. A 3d generative model for structure-based drug design. Advances in Neural Information Processing Systems, 34, 2021.

Related Work

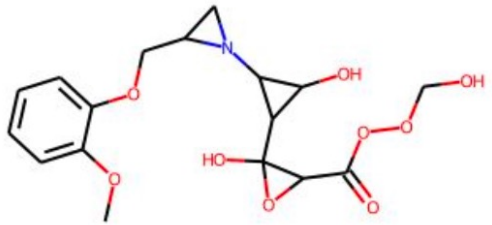
Pocket2Mol (Peng et al., 2022)
Improvement compared to AR (Luo et al., 2021):

- Relative Position Prediction
- Bond Prediction
- Apply Vector-Based Neural Network



Problem of autoregressive models

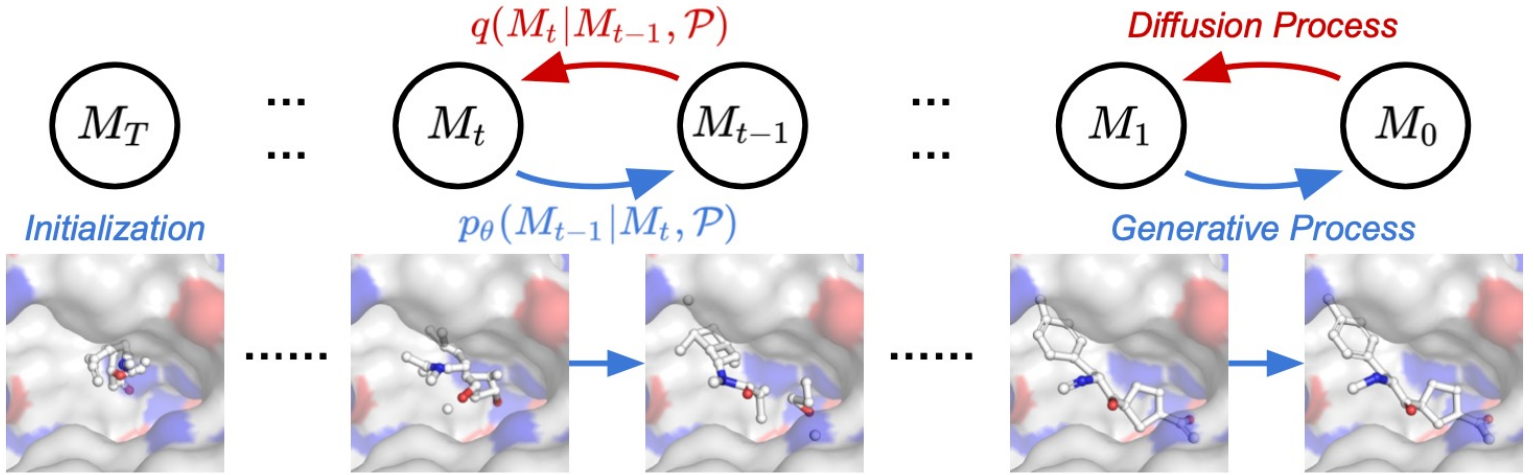
- Training and sampling doesn't align well (exposure bias)
- Doesn't consider the probability of the entire 3D structure → unrealistic fragments
- Doesn't scale well when generating large binding molecules is necessary



Xingang Peng, Shitong Luo, Jiaqi Guan, Qi Xie, Jian Peng, and Jianzhu Ma. Pocket2mol: Efficient molecular sampling based on 3d protein pockets. ICML, 2022

3D Equivariant Diffusion

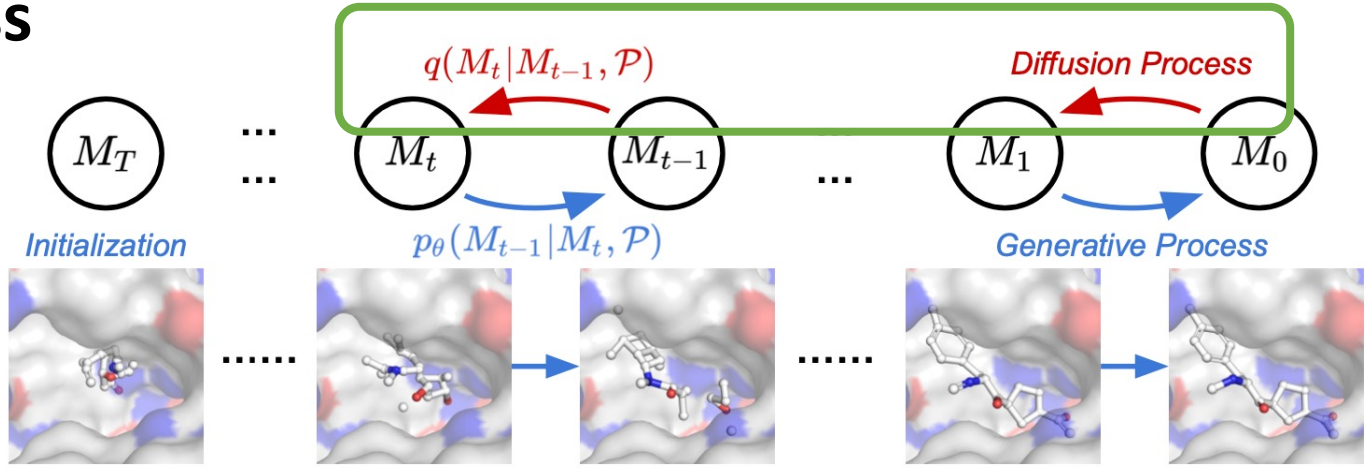
Overview



- Training and sampling are aligned
- Capture the global structure information
- Avoid voxelization and scale well for large molecules

3D Equivariant Diffusion

Diffusion Process



- Joint distribution of continuous atom coordinates and discrete atom types

$$q(M_t|M_{t-1},\mathcal{P}) = \mathcal{N}(\mathbf{x}_t; \sqrt{1-\beta_t}\mathbf{x}_{t-1}, \beta_t\mathbf{I}) \cdot \mathcal{C}(\mathbf{v}_t|(1-\beta_t)\mathbf{v}_{t-1} + \beta_t/K).$$

- Calculate the noisy distribution of any time step in closed-form:

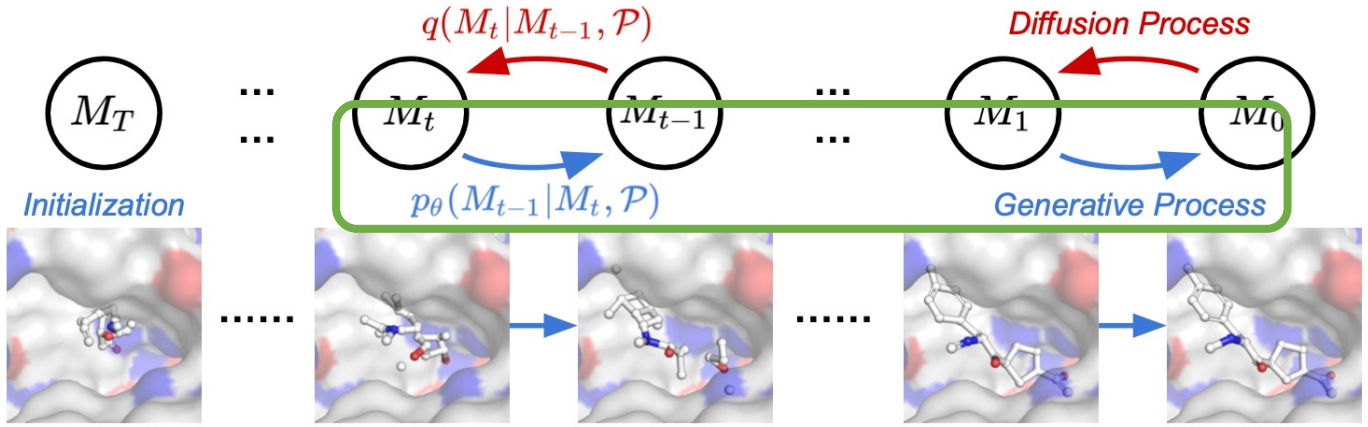
$$q(\mathbf{x}_t|\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_t; \sqrt{\bar{\alpha}_t}\mathbf{x}_0, (1-\bar{\alpha}_t)\mathbf{I}) \qquad q(\mathbf{v}_t|\mathbf{v}_0) = \mathcal{C}(\mathbf{v}_t|\bar{\alpha}_t\mathbf{v}_0 + (1-\bar{\alpha}_t)/K)$$

- Using Bayes theorem, we can also compute the posterior of x and v in closed-form:

$$q(\mathbf{x}_{t-1}|\mathbf{x}_t,\mathbf{x}_0) = \mathcal{N}(\mathbf{x}_{t-1}; \tilde{\boldsymbol{\mu}}_t(\mathbf{x}_t,\mathbf{x}_0), \tilde{\beta}_t\mathbf{I}) \qquad q(\mathbf{v}_{t-1}|\mathbf{v}_t,\mathbf{v}_0) = \mathcal{C}(\mathbf{v}_{t-1}|\tilde{\boldsymbol{c}}_t(\mathbf{v}_t,\mathbf{v}_0)).$$

3D Equivariant Diffusion

Equivariant Generative Process



$$p_\theta(M_{t-1}|M_t,\mathcal{P}) = \mathcal{N}(\mathbf{x}_{t-1};\boldsymbol{\mu}_\theta([\mathbf{x}_t,\mathbf{v}_t],t,\mathcal{P}),\sigma_t^2 I)\cdot\mathcal{C}(\mathbf{v}_{t-1}|\boldsymbol{c}_\theta([\mathbf{x}_t,\mathbf{v}_t],t,\mathcal{P})).$$

A denoising equivariant neural network:

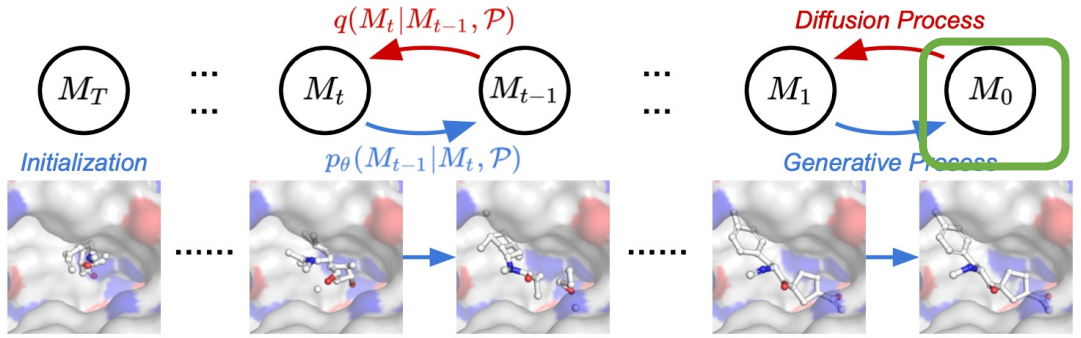
$$[\hat{\mathbf{x}}_0,\hat{\mathbf{v}}_0] = \phi_\theta(M_t,t,\mathcal{P}) = \phi_\theta([\mathbf{x}_t,\mathbf{v}_t],t,\mathcal{P}).$$

$$\mathbf{h}_i^{l+1} = \mathbf{h}_i^l + \sum_{j\in\mathcal{V},i\neq j} f_h(d_{ij}^l,\mathbf{h}_i^l,\mathbf{h}_j^l,\mathbf{e}_{ij};\theta_h)$$

$$\mathbf{x}_i^{l+1} = \mathbf{x}_i^l + \sum_{j\in\mathcal{V},i\neq j} (\mathbf{x}_i^l-\mathbf{x}_j^l) f_x(d_{ij}^l,\mathbf{h}_i^{l+1},\mathbf{h}_j^{l+1},\mathbf{e}_{ij};\theta_x)\cdot\mathbb{1}_{\text{mol}}$$

3D Equivariant Diffusion

Affinity Ranking and Prediction



A denoising equivariant neural network: $[\hat{\mathbf{x}}_0, \hat{\mathbf{v}}_0] = \phi_\theta(M_t, t, \mathcal{P}) = \phi_\theta([\mathbf{x}_t, \mathbf{v}_t], t, \mathcal{P}).$

$$\mathbf{h}_i^{l+1} = \mathbf{h}_i^l + \sum_{j \in \mathcal{V}, i \neq j} f_h(d_{ij}^l, \mathbf{h}_i^l, \mathbf{h}_j^l, \mathbf{e}_{ij}; \theta_h)_{l=1\ldots L-1} \qquad \hat{\mathbf{v}}_0 = \texttt{softmax}(\mathbf{MLP}(\mathbf{h}^L)).$$

Assumption: if the ligand molecule has a good binding affinity to protein

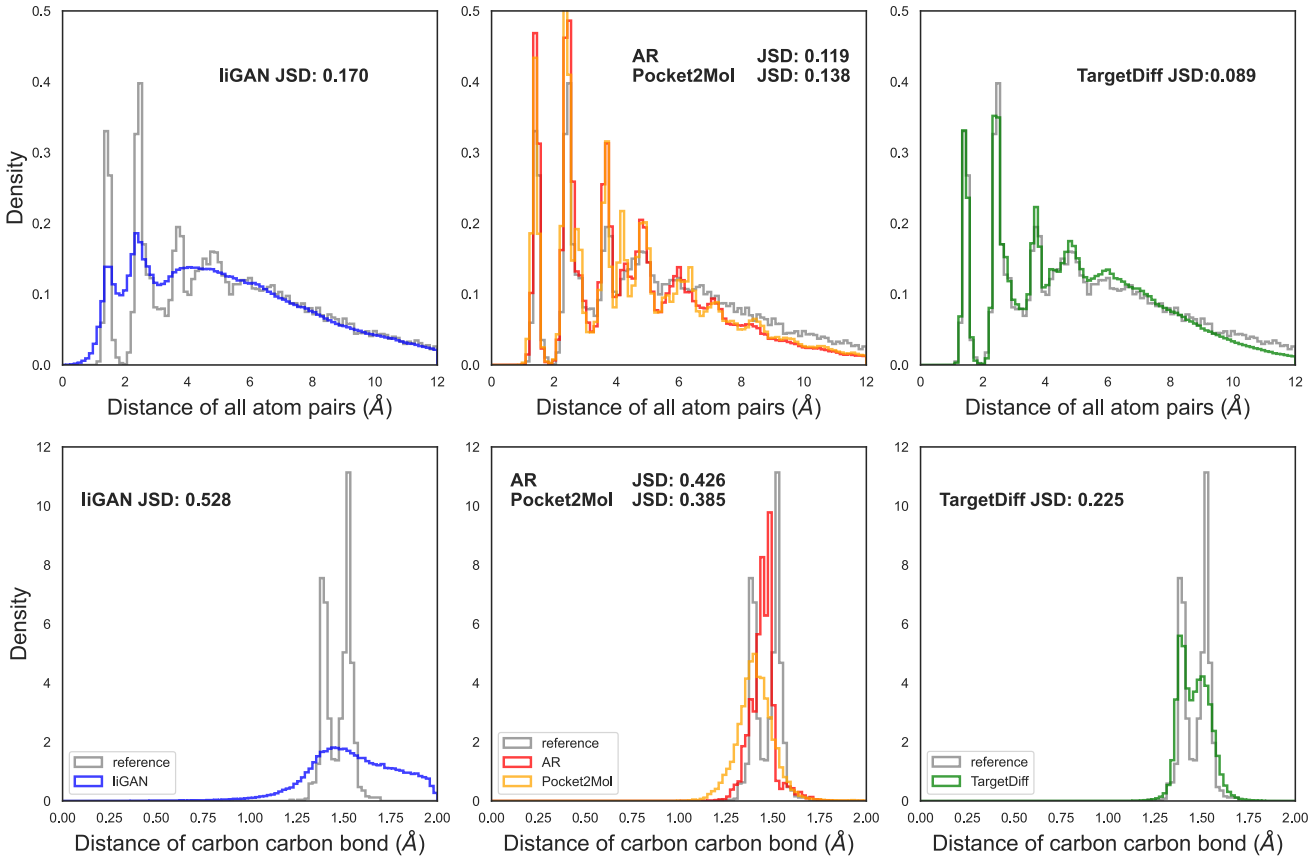
→ the flexibility of atom types should be low

→ entropy of $\hat{v}_0$ is low

Our model can serve as a scoring function to perform affinity ranking and prediction

Experiments

Molecular Structure Analysis

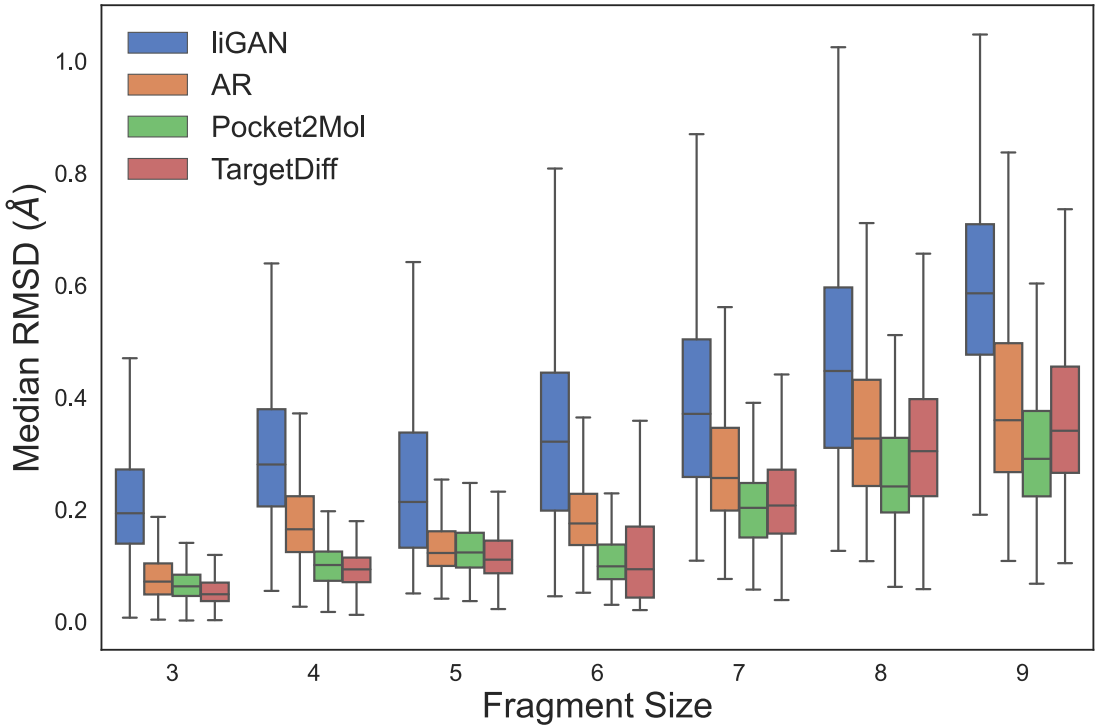


Bond	liGAN	AR	Pocket2Mol	TargetDiff
C−C	0.601	0.609	0.496	**0.369**
C=C	0.665	0.620	0.561	**0.505**
C−N	0.634	0.474	0.416	**0.363**
C=N	0.749	0.635	0.629	**0.550**
C−O	0.656	0.492	0.454	**0.421**
C=O	0.661	0.558	0.516	**0.461**
C:C	0.497	0.451	0.416	**0.263**
C:N	0.638	0.552	0.487	**0.235**

JS-div between distributions of bond distances

Experiments

Molecular Structure Analysis



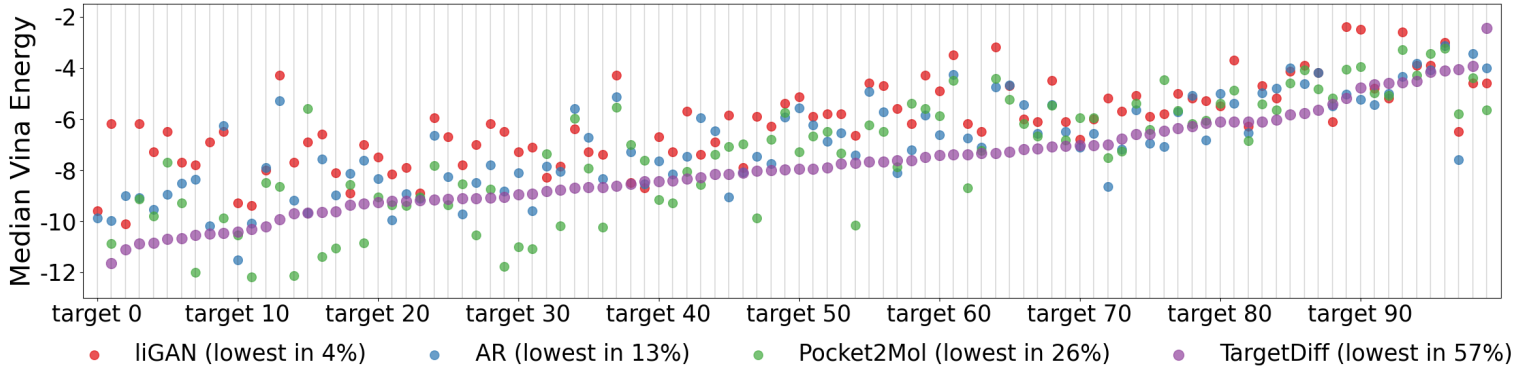
Median RMSD for rigid fragment 3D structure before and after the force-field optimization

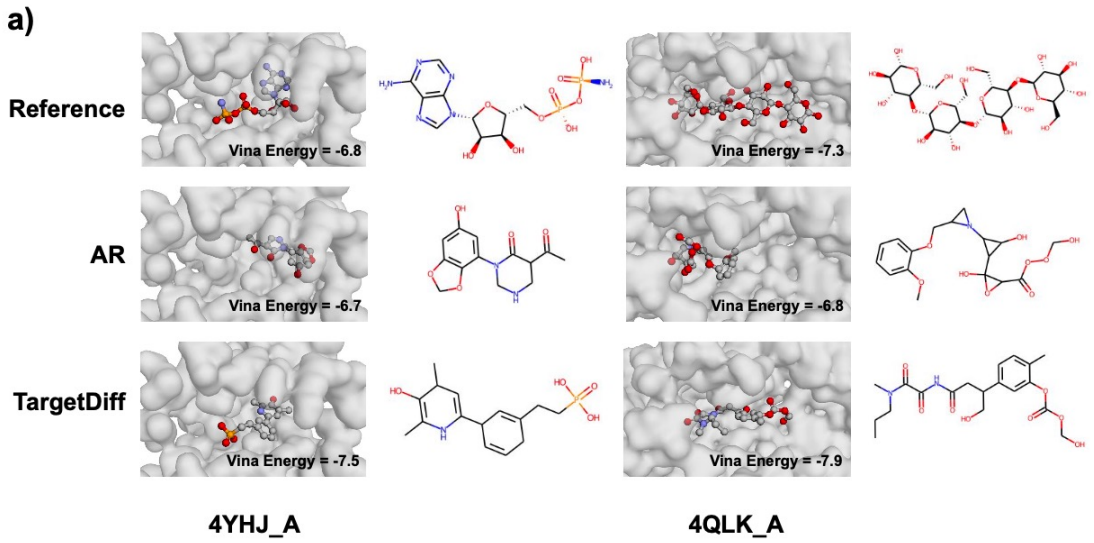
Ring Size	Ref.	liGAN	AR	Pocket2Mol	TargetDiff
3	1.7%	28.1%	29.9%	0.1%	0.0%
4	0.0%	15.7%	0.0%	0.0%	2.8%
5	30.2%	29.8%	16.0%	16.4%	30.8%
6	67.4%	22.7%	51.2%	80.4%	50.7%
7	0.7%	2.6%	1.7%	2.6%	12.1%
8	0.0%	0.8%	0.7%	0.3%	2.7%
9	0.0%	0.3%	0.5%	0.1%	0.9%

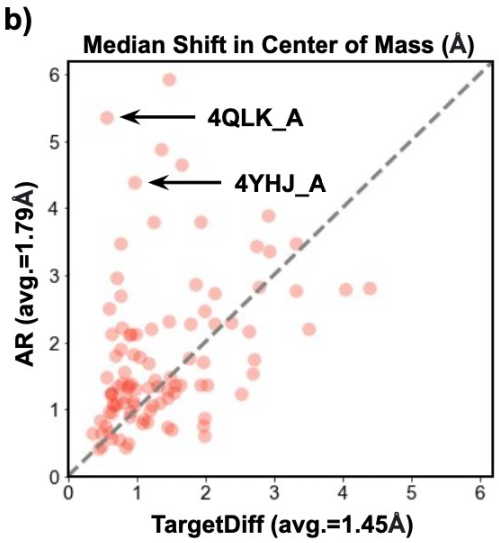
Percentage of different ring sizes for reference and model generated molecules

Experiments

Target Binding Affinity

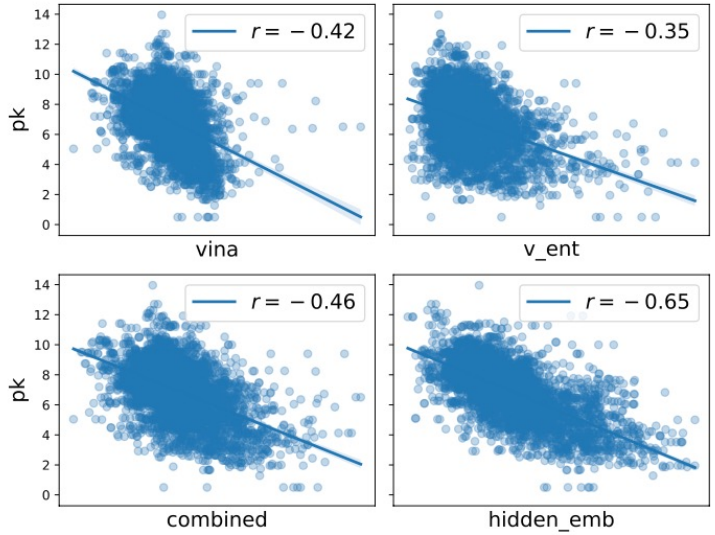






Experiments

Binding Affinity Ranking and Prediction



Spearman's rank correlation between different indicators and pK

Model \ Metric	RMSE ↓	Pearson ↑	Spearman ↑	MAE ↓
TransCPI	1.741	0.576	0.540	1.404
MONN	1.438	0.624	0.589	1.143
IGN	1.433	0.698	0.641	1.169
HOLOPROT	1.546	0.602	0.571	1.208
STAMP-DPI	1.658	0.545	0.411	1.325
EGNN	1.445	0.648	0.598	1.141
EGNN + ours	1.374	0.680	0.637	1.118

Binding affinity prediction on PDBBind v2020

- Unsupervised learning can provide useful information for binding affinity ranking.
- The entropy score provides some complementary information to traditional chemical / physical-based score function like Vina
- When provided with labeled data, the final hidden embedding h_L (*i.e.* hidden emb) with a simple linear transformation could improve the correlation to a large extent.

Thanks for watching!

3D Equivariant Diffusion For Target-Aware Molecule Generation and Affinity Prediction

ICLR 2023

- Jiaqi Guan[1], Wesley Wei Qian[1], Xingang Peng[2], Yufeng Su[1], Jian Peng[1], Jianzhu Ma[3]

1. Department of Computer Science, University of Illinois Urbana Champaign
2. School of Intelligence Science and Technology, Peking University
3. Institute for AI industry Research, Tsinghua University