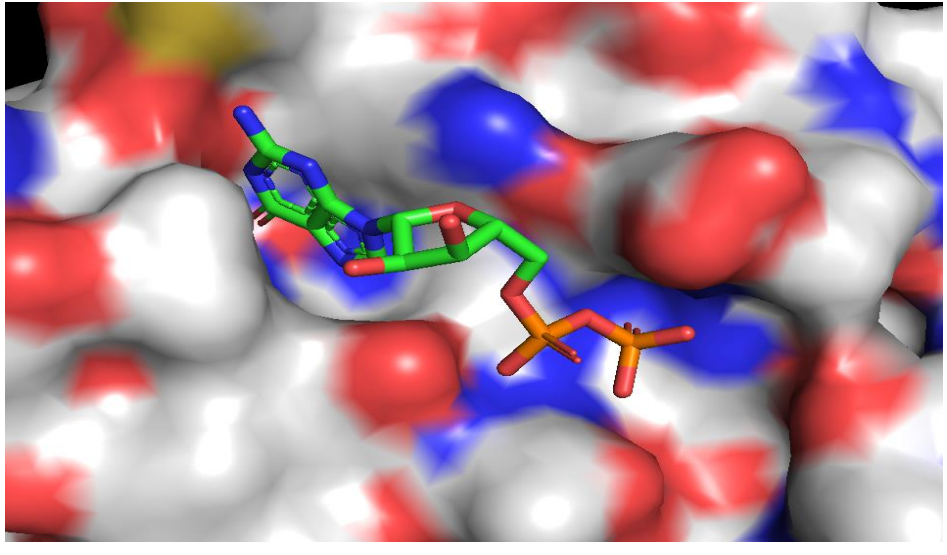
Pocket2Mol: Efficient Molecular Sampling Based on 3D Protein Pockets

Xingang Peng [1] Shitong Luo [2] Jiaqi Guan [3] Qi Xie [4] Jian Peng [2 3 5] Jianzhu Ma [2 6 7]

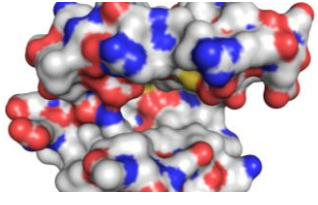


Content

Background

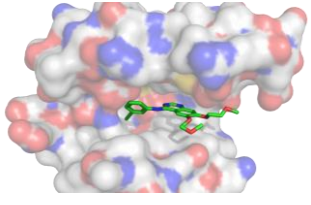
1. Molecular generation is important but hard
2. Most previous work:
   - Pocket-free generation
   - Molecular representation: 1D SMILES or 2D graph

Generating 3D molecules that bind to protein pockets



Protein Pocket

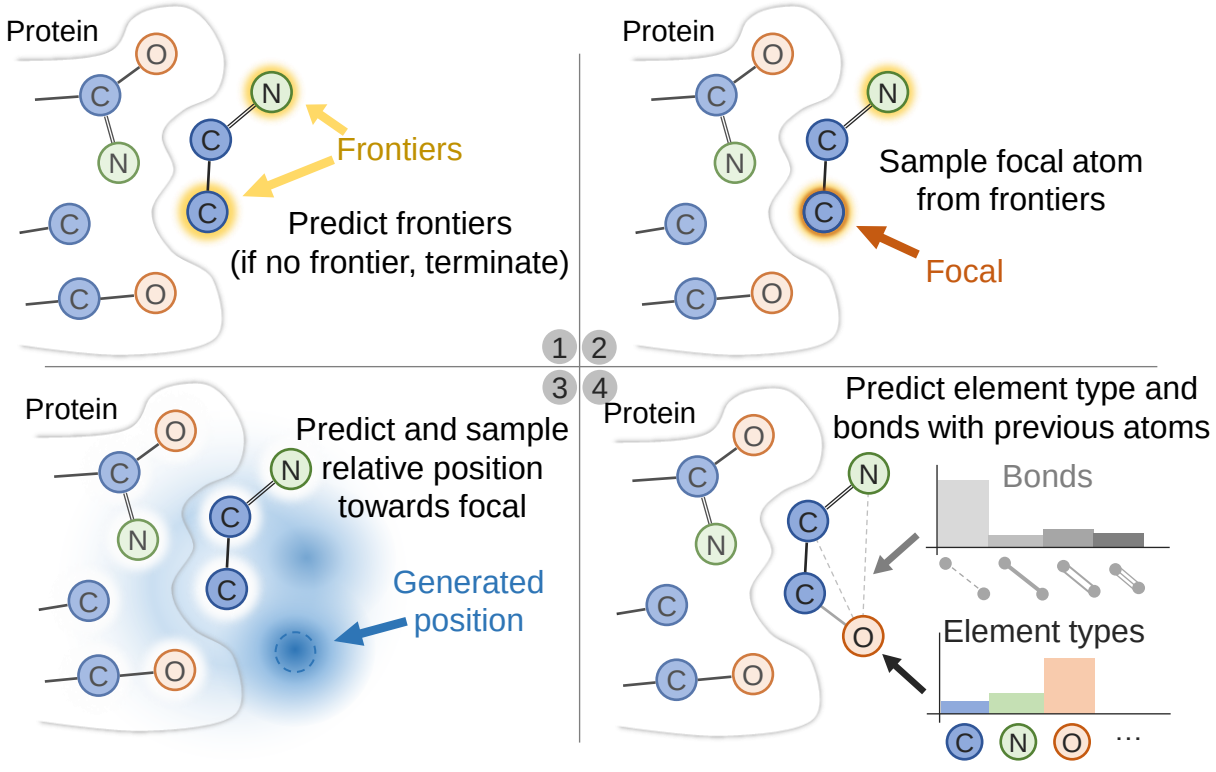
Molecular Generation



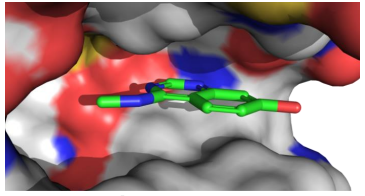
3D Molecule

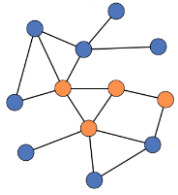
Methods

The generation is **auto-regressive** and includes **four** steps



Methods



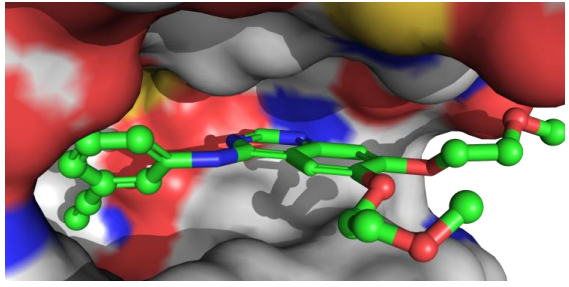


Encoder

Frontier Predictor

Position Predictor

Element-and-Bond Predictor



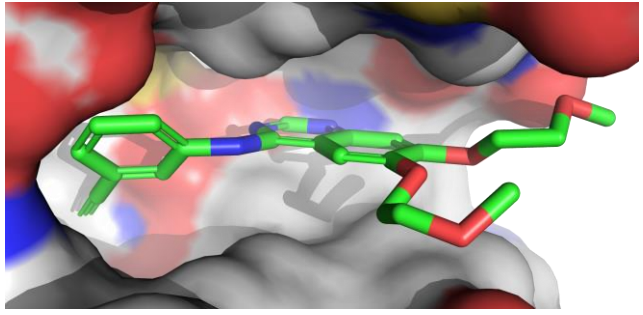
Methods

1. E(3)-equivariant network
   - Atoms and edges are represented by **scalar and vector** features
   - Utilize **Geometric Vector Perceptron** and **Vector Neuron** to design equivariant building blocks

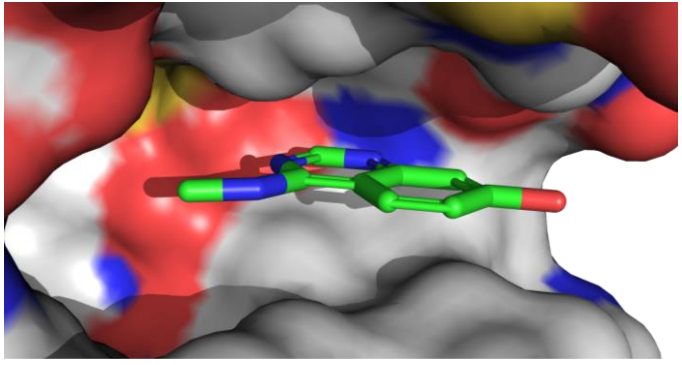
2. Direct generation of atom positions
   - Use **Vector features** and **Gaussian Mixture Model** to predict the atom positions

3. Joint Prediction of atom element and bonds

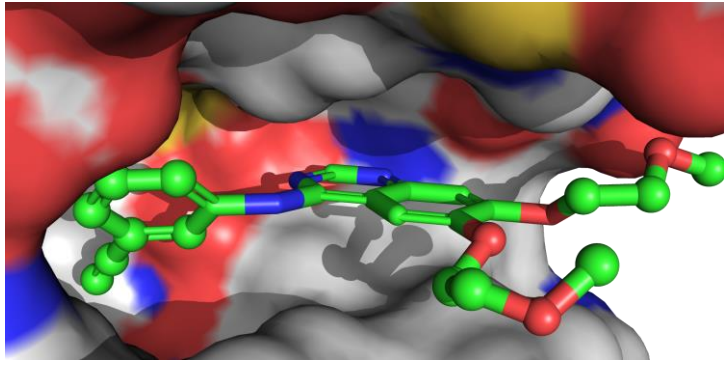
Methods



Randomly **mask** molecular atoms



Model Prediction



Train the model to recover atoms

Results

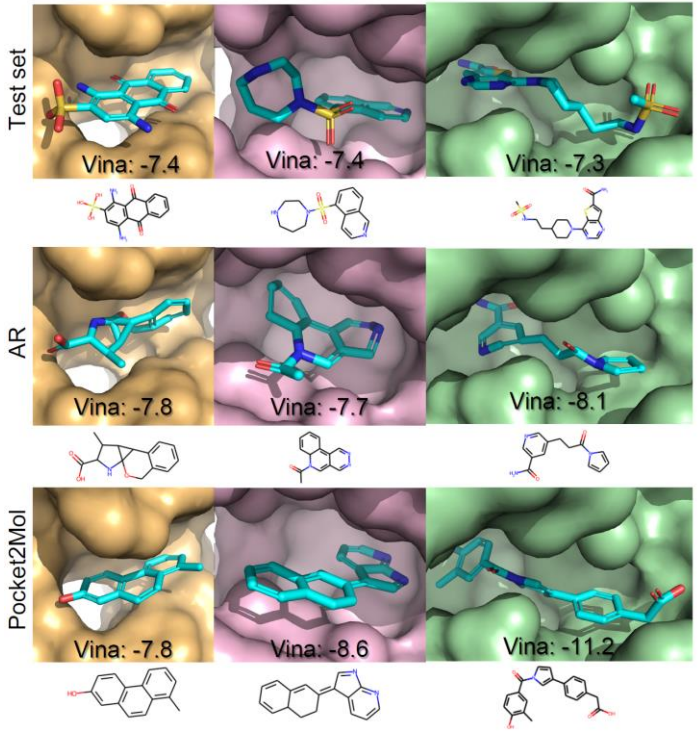
Baselines:

CVAE (Masuda et al., 2020): 3D CNN + CVAE

AR (Luo et al., 2021): auto-regressive model

	Test Set	CVAE	AR	Pocket2 Mol
Vina Score (kcal/mol, ↓)	-7.158 ± 2.10	-6.144 ± 1.57	-6.215 ± 1.54	**-7.288** ± 2.53
High Affinity (↑)	-	0.238 ± 0.28	0.267 ± 0.31	**0.542** ± 0.32
QED (↑)	0.484 ±0.21	0.369 ±0.22	0.502 ±0.17	**0.563** ±0.16
SA (↑)	0.732 ±0.14	0.590 ±0.15	0.675 ±0.14	**0.765** ±0.13
LogP	0.947 ±2.65	-0.140 ±2.73	0.257 ±2.01	**1.586** ±1.82
Lipinski (↑)	4.367 ±1.14	4.027 ±1.38	4.787 ±0.50	**4.902** ±0.42
Sim. Train (↓)	-	0.460 ±0.18	0.409 ±0.19	**0.376** ±0.22
Diversity (↑)	-	0.654 ±0.12	**0.742** ±0.09	0.688 ±0.14
Time (s, ↓)	-	-	19658.56 ±14704	**2503.51** ±2207

Better **drug-likeness** and **vina score**



Results

Sub-structure Analysis

Table 2. The ratio of the molecules containing different rings in the datasets and those generated by different methods.

Ring Size	Train Set	Test Set	CVAE	AR	Pocket2 Mol
3	0.034	0.033	0.361	0.484	**0.002**
4	0.005	0.000	0.248	0.005	**0.000**
5	0.572	0.475	0.397	0.276	**0.415**
6	0.903	0.833	0.300	0.693	**0.885**
7	0.028	0.017	0.044	**0.033**	0.076
8	0.001	0.000	0.014	**0.007**	**0.007**
9	0.000	0.000	0.006	0.006	**0.002**

More **realistic** sub-structures

Table 3. The KL divergence of the bond angles and dihedral angles with the test set. The lower letters represent the atoms in the aromatic rings.

	Val. Set	Test. Set	CVAE	AR	Pocket2 Mol
CCC	0.12	0.00	7.08	1.80	**0.97**
CCO	0.10	0.00	7.58	2.02	**0.95**
CNC	0.11	0.00	7.74	2.86	**0.49**
OPO	0.10	0.00	4.72	2.06	**0.23**
NCC	0.09	0.00	7.86	2.55	**0.95**
CC=O	0.07	0.00	7.41	2.90	**0.76**
COC	0.12	0.00	6.32	3.88	**0.24**
CCCC	0.14	0.00	**0.59**	0.78	0.71
cccc	0.08	0.00	7.91	10.64	**4.49**
CCCO	0.55	0.00	0.94	1.23	**0.56**
OCCO	1.01	0.00	1.92	1.85	**1.56**
Cccc	0.28	0.00	5.78	7.91	**2.85**
CC=CC	0.68	0.00	4.96	7.07	**4.09**

Better **local** conformation

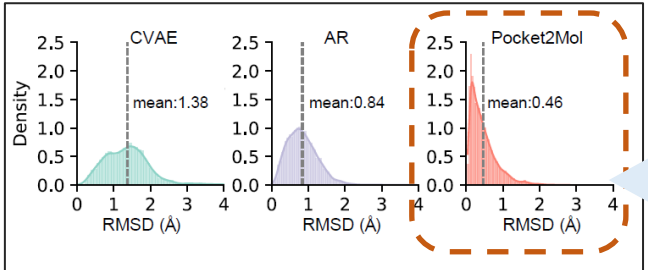


Figure 4. The distributions of RMSD of the generated 3D molecular structures.

Better **global** conformation

Conclusion

1. We proposed Pocket2Mol to efficiently generate **3D molecules** directly binding to the given **3D protein pockets**.

2. The generated molecules show better **chemical properties** and more reasonable **structures / conformations**.

Thank You!