A 3D Generative Model for Structure-Based Drug Design

Shitong Luo, Jianzhu Ma, Jiaqi Guan, Jian Peng
University of Illinois Urbana-Champaign
2021

Problem

- The existing models are mostly **string-based or graph-base**, they are limited by the **lack of spatial information**.
- Authors propose a 3D generative model that **generates molecules given a designated 3D protein binding site** by estimating **the probability density of atom**'s occurrences in 3D space

Methods

- Step 1: present a 3D generative model that predicts the **probability of atom occurrence in 3D space** of the binding site.
- Step 2: the auto-regressive **sampling algorithm** for generating valid and multi-modal molecules from the model.
- Step 3: derive the training objective, by which the **model learns to predict** where should be placed and atoms and what type of atom should be placed.

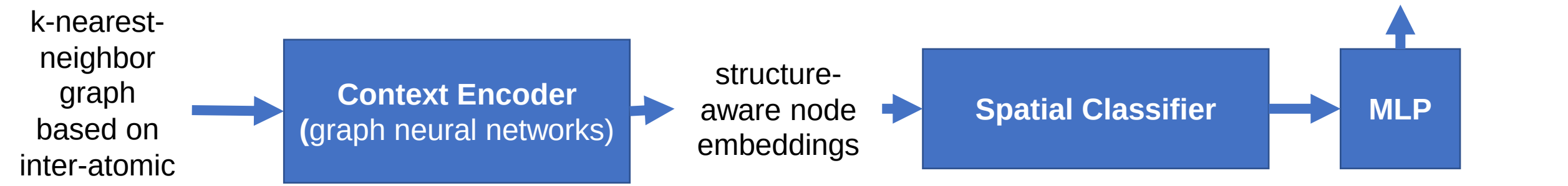
Step 1: 3D Generative Model Design

set of atoms $\mathcal{C} = \{(\boldsymbol{a}_i, \boldsymbol{r}_i)\}_{i=1}^{N_b}$, $e \in \mathcal{E} = \{\mathrm{H}, \mathrm{C}, \mathrm{O}, \ldots\}$

probability of atom occurring at some position $\boldsymbol{r}$ in the site. $p(e|\boldsymbol{r}, \mathcal{C})$

- **Context Encoder** learns the representation of each atom in the context C via graph neural networks
- **Spatial Classifier** takes as input a query position r, then aggregates the representation of contextual atoms nearby it, and finally predicts **p(e|r,C)**

$$p(e|\boldsymbol{r}, \mathcal{C}) = \frac{\exp(\boldsymbol{c}[e])}{1 + \sum_{e' \in \mathcal{E}} \exp(\boldsymbol{c}[e'])},$$



$$\mathcal{G} = \langle \mathcal{C}, \boldsymbol{A} \rangle \qquad \boldsymbol{h}_i^{(\ell+1)} = \sigma\left(\boldsymbol{W}_0^\ell \boldsymbol{h}_i^{(\ell)} + \sum_{j \in N_k(\boldsymbol{r}_i)} \boldsymbol{W}_1^\ell \boldsymbol{w}(d_{ij}) \odot \boldsymbol{W}_2^\ell \boldsymbol{h}_j^{(\ell)}\right)$$

$$\boldsymbol{v} = \sum_{j \in N_k(\boldsymbol{r})} \boldsymbol{W}_0 \boldsymbol{w}_{\text{aggr}}(\|\boldsymbol{r} - \boldsymbol{r}_j\|) \odot \boldsymbol{W}_1 \boldsymbol{h}_j^{(L)},$$

Step 2: Sampling

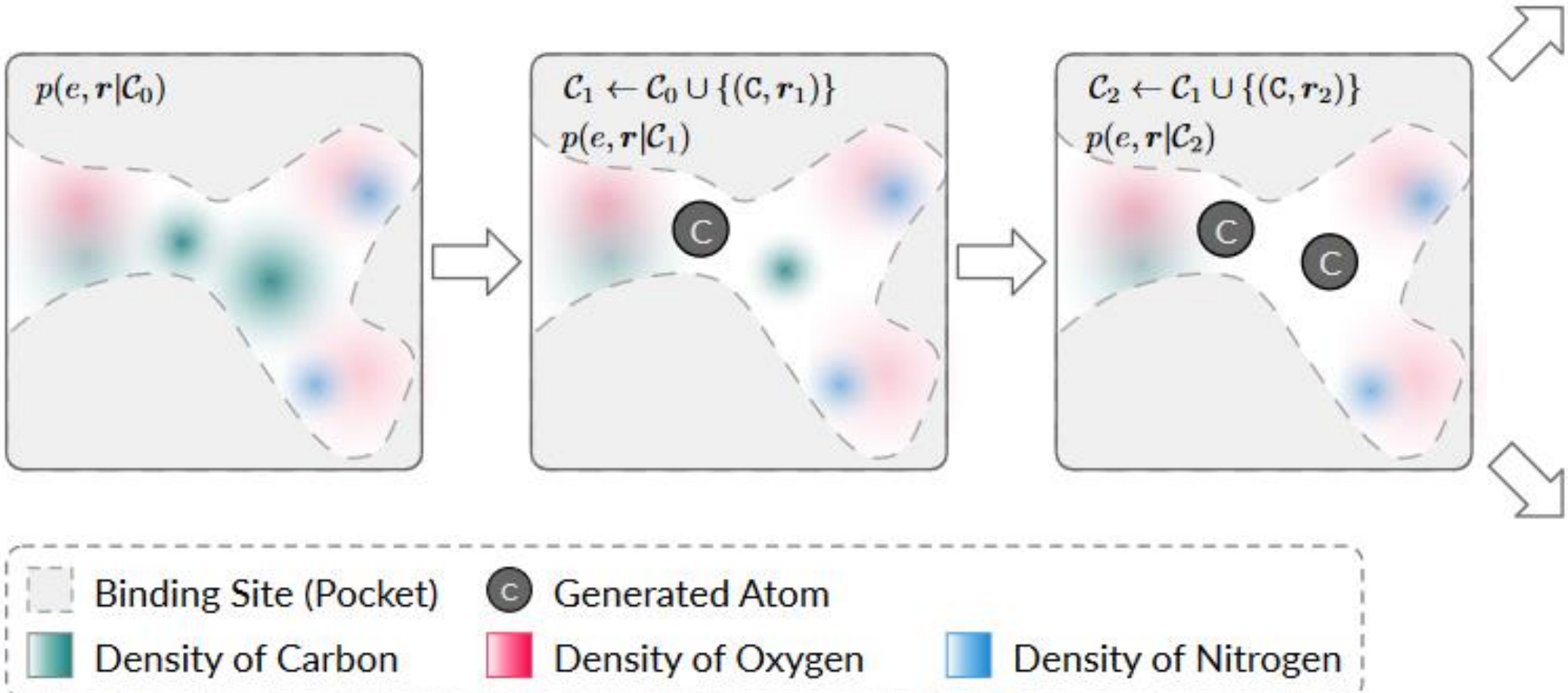
Sampling a molecule amounts to generating a set of atoms $\{(e_i, \boldsymbol{r}_i)\}_{i=1}^{N_a}$

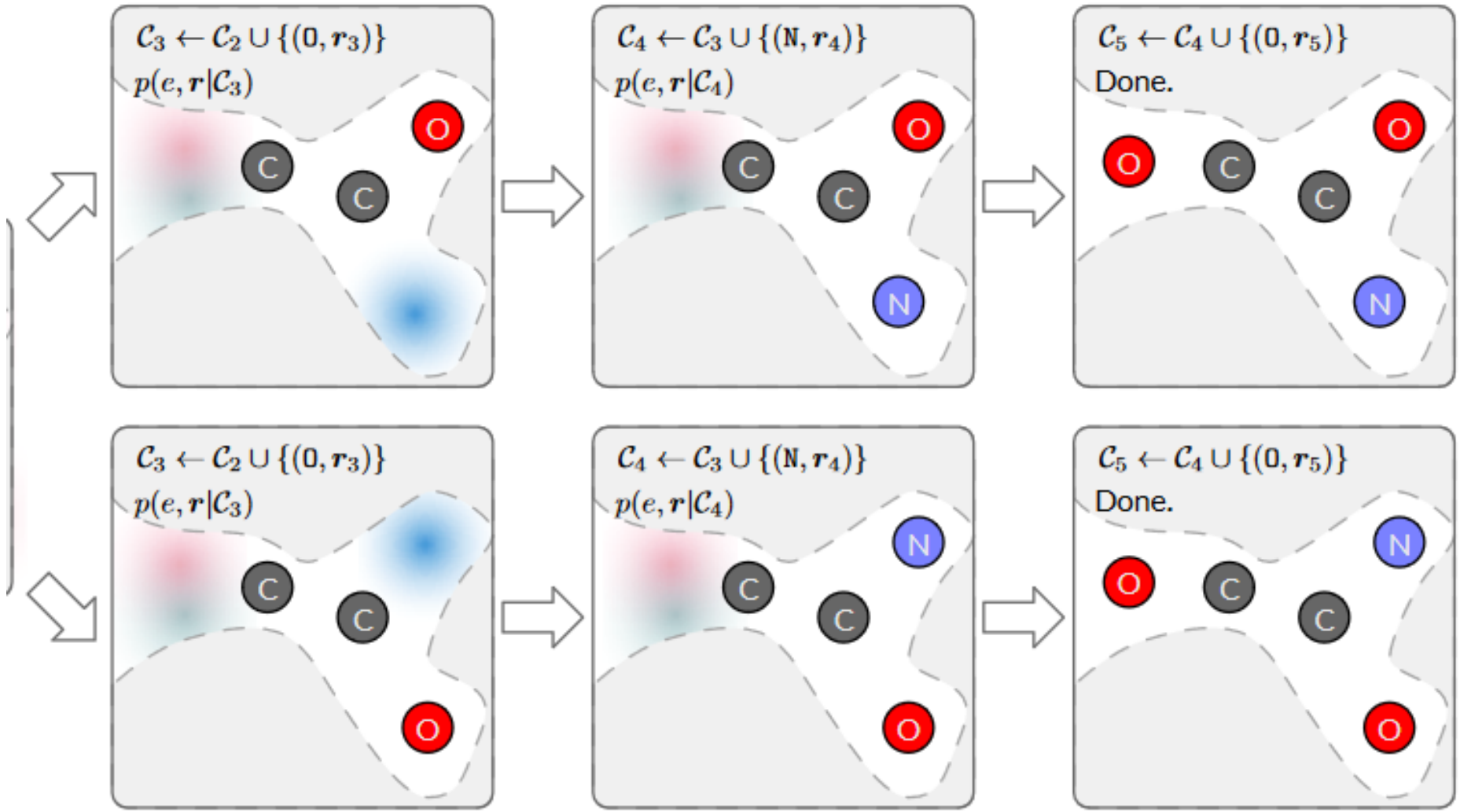
Joint Distribution We define the joint distribution of coordinate $\boldsymbol{r}$ and atom type e

$$p(e, \boldsymbol{r}|\mathcal{C}) = \frac{\exp\left(\boldsymbol{c}[e]\right)}{Z},$$

Auto-Regressive Sampling We sample a molecule by progressively sampling one atom at each step. In specific, at step t, the context $\mathcal{C}_t$ contains not only protein atoms but also t atoms sampled beforehand. Sampled atoms in $\mathcal{C}_t$ are treated equally as protein atoms in the model, but they have different attributes in order to differentiate themselves from protein atoms. Then, the $(t+1)$-th atom will be sampled from $p(e, \boldsymbol{r}|\mathcal{C}_t)$ and will be added to $\mathcal{C}_t$, leading to the context for next step $\mathcal{C}_{t+1}$.

$$\begin{aligned} (e_{t+1}, \boldsymbol{r}_{t+1}) &\sim p(e, \boldsymbol{r}|\mathcal{C}_t), \\ \mathcal{C}_{t+1} &\leftarrow \mathcal{C}_t \cup \{(e_{t+1}, \boldsymbol{r}_{t+1})\}. \end{aligned} \tag{7}$$


$p(e, \boldsymbol{r}|\mathcal{C}_0)$
$\mathcal{C}_1 \leftarrow \mathcal{C}_0 \cup \{(\mathrm{C}, \boldsymbol{r}_1)\}$
$p(e, \boldsymbol{r}|\mathcal{C}_1)$
$\mathcal{C}_2 \leftarrow \mathcal{C}_1 \cup \{(\mathrm{C}, \boldsymbol{r}_2)\}$
$p(e, \boldsymbol{r}|\mathcal{C}_2)$
C
C
C
Binding Site (Pocket)
Generated Atom
Density of Carbon
Density of Oxygen
Density of Nitrogen


$\mathcal{C}_3 \leftarrow \mathcal{C}_2 \cup \{(\mathtt{O}, \boldsymbol{r}_3)\}$
$p(e, \boldsymbol{r}|\mathcal{C}_3)$
$\mathcal{C}_4 \leftarrow \mathcal{C}_3 \cup \{(\mathtt{N}, \boldsymbol{r}_4)\}$
$p(e, \boldsymbol{r}|\mathcal{C}_4)$
$\mathcal{C}_5 \leftarrow \mathcal{C}_4 \cup \{(\mathtt{O}, \boldsymbol{r}_5)\}$
Done.
C
C
O
N
O
C
C
O
N
O
$\mathcal{C}_3 \leftarrow \mathcal{C}_2 \cup \{(\mathtt{O}, \boldsymbol{r}_3)\}$
$p(e, \boldsymbol{r}|\mathcal{C}_3)$
$\mathcal{C}_4 \leftarrow \mathcal{C}_3 \cup \{(\mathtt{N}, \boldsymbol{r}_4)\}$
$p(e, \boldsymbol{r}|\mathcal{C}_4)$
$\mathcal{C}_5 \leftarrow \mathcal{C}_4 \cup \{(\mathtt{O}, \boldsymbol{r}_5)\}$
Done.
C
C
O
N
O
C
C
O
N
O

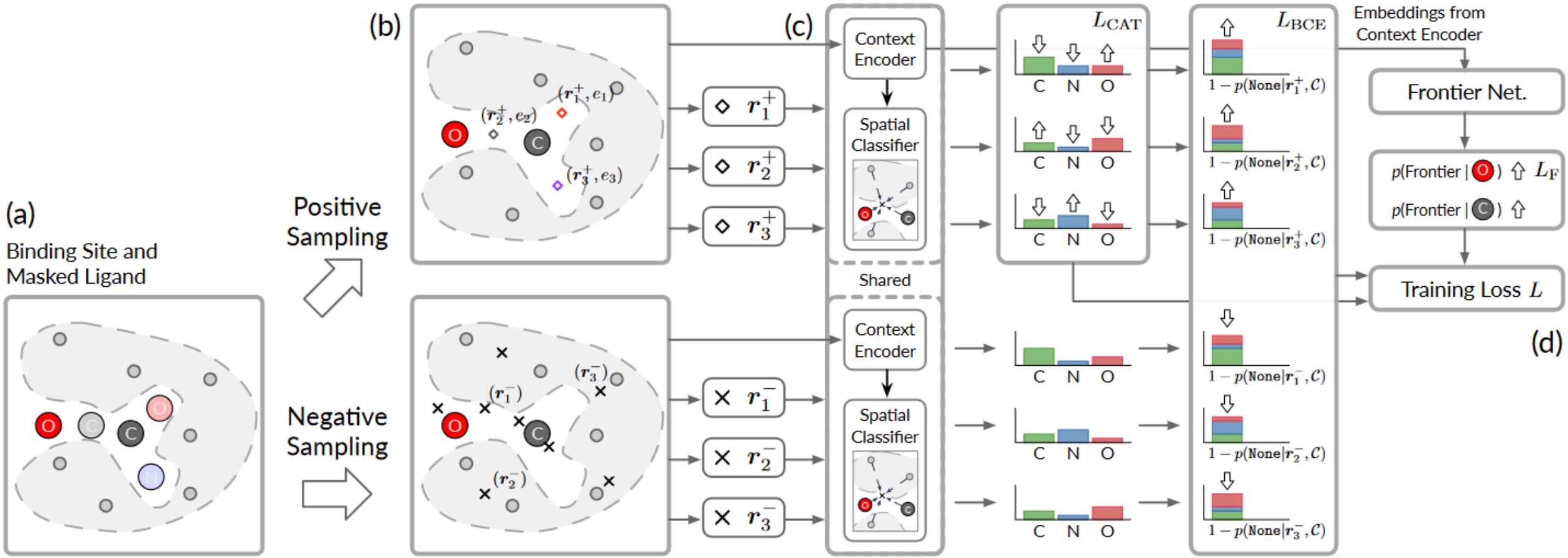
Step 3: Training

$$L_{\text{BCE}} = -\mathbb{E}_{\boldsymbol{r}\sim p_+}\left[\log\left(1 - p(\texttt{Nothing}|\boldsymbol{r}, \mathcal{C})\right)\right] - \mathbb{E}_{\boldsymbol{r}\sim p_-}\left[\log p(\texttt{Nothing}|\boldsymbol{r}, \mathcal{C})\right].$$

$$L_{\text{CAT}} = -\mathbb{E}_{(e,\boldsymbol{r})\sim p_+}\left[\log p(e|\boldsymbol{r}, \mathcal{C})\right].$$

$$L_{\text{F}} = \sum_{i\in\mathcal{F}\subseteq\mathcal{C}} \log\sigma(F(\boldsymbol{h}_i)) + \sum_{i\notin\mathcal{F}\subseteq\mathcal{C}} \log(1 - \sigma(F(\boldsymbol{h}_i))),$$

$$L = L_{\text{BCE}} + L_{\text{CAT}} + L_{\text{F}}.$$


(a)
Binding Site and
Masked Ligand
Positive
Sampling
Negative
Sampling
(b)
(c)
Context
Encoder
Spatial
Classifier
Shared
Context
Encoder
Spatial
Classifier
L_{CAT}
L_{BCE}
C N O
$1 - p(\mathbf{None}|\boldsymbol{r}_1^+, \mathcal{C})$
$1 - p(\mathbf{None}|\boldsymbol{r}_2^+, \mathcal{C})$
$1 - p(\mathbf{None}|\boldsymbol{r}_3^+, \mathcal{C})$
$1 - p(\mathbf{None}|\boldsymbol{r}_1^-, \mathcal{C})$
$1 - p(\mathbf{None}|\boldsymbol{r}_2^-, \mathcal{C})$
$1 - p(\mathbf{None}|\boldsymbol{r}_3^-, \mathcal{C})$
Embeddings from
Context Encoder
Frontier Net.
p(Frontier | O) ⇧ L_{F}
p(Frontier | C) ⇧
Training Loss L
(d)

Molecule Design Data:

- CrossDocked: 184,057 docked protein-ligand pairs.
- mmseqs2 to cluster data at 30% sequence identity, =>100,000 protein-ligand pairs for training and 100 proteins from remaining clusters for testing

Metric:

- Quality of generated molecules
  - **Binding affinity** measures how well the generated molecules fit the binding site. (VinaScore)
  - **Drug likeness** reflects how much a molecule is like a drug. (QED score)
  - **Synthesizability** assesses the ease of synthesis of generated molecules. (SA score)
- Generation quality and diversity:
  - **Percentage of Samples with High Affinity**, which measures the percentage of a binding site's generated molecules whose binding affinity is higher than or equal to the reference ligand.
  - **Diversity** measures the diversity of generated molecules for a binding site.

Evaluation:

Metric		liGAN	**Ours**	Ref
Vina Score (kcal/mol, ↓)	Avg.	-6.144	**-6.344**	-7.158
	Med.	-6.100	**-6.200**	-6.950
QED (↑)	Avg.	0.371	**0.525**	0.484
	Med.	0.369	**0.519**	0.469
SA (↑)	Avg.	0.591	**0.657**	0.733
	Med.	0.570	**0.650**	0.745
High Affinity (%, ↑)	Avg.	23.77	**29.09**	-
	Med.	11.00	**18.50**	-
Diversity (↑)	Avg.	0.655	**0.720**	-
	Med.	0.676	**0.736**	-

Table 1: Mean and median values of the four metrics on generation quality. (↑) indicates higher is better. (↓) indicates lower is better.

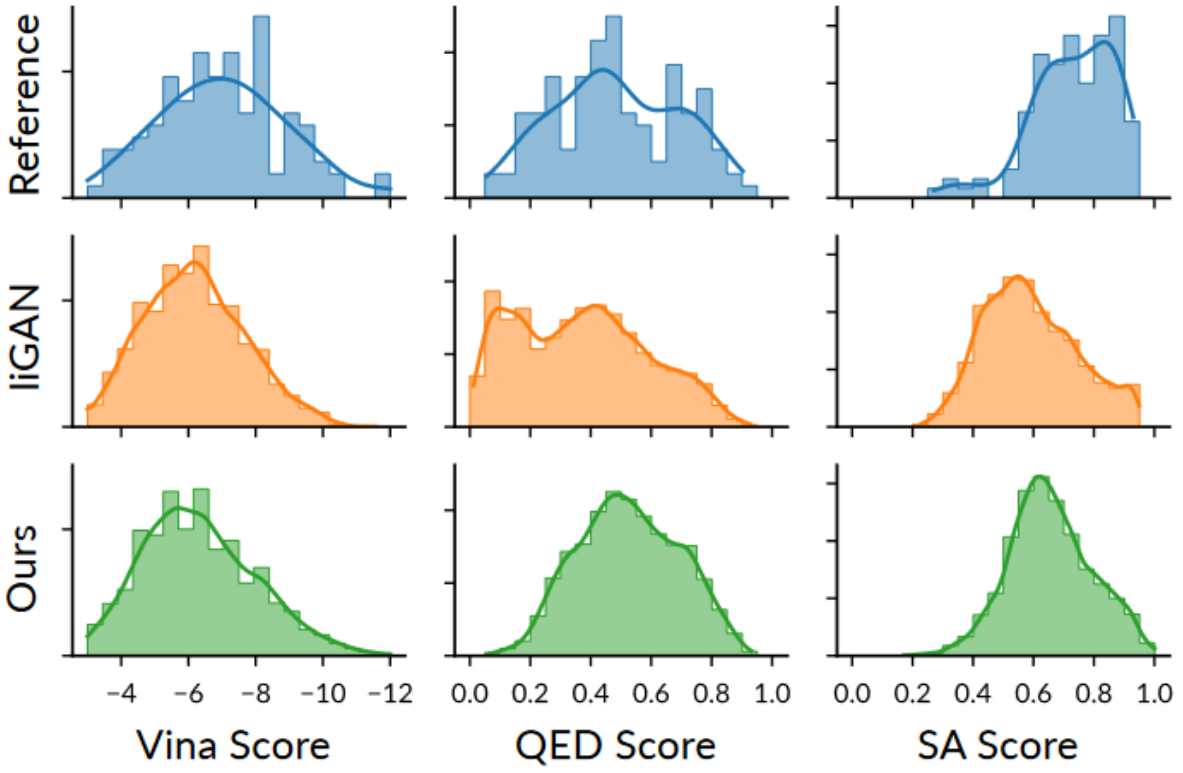
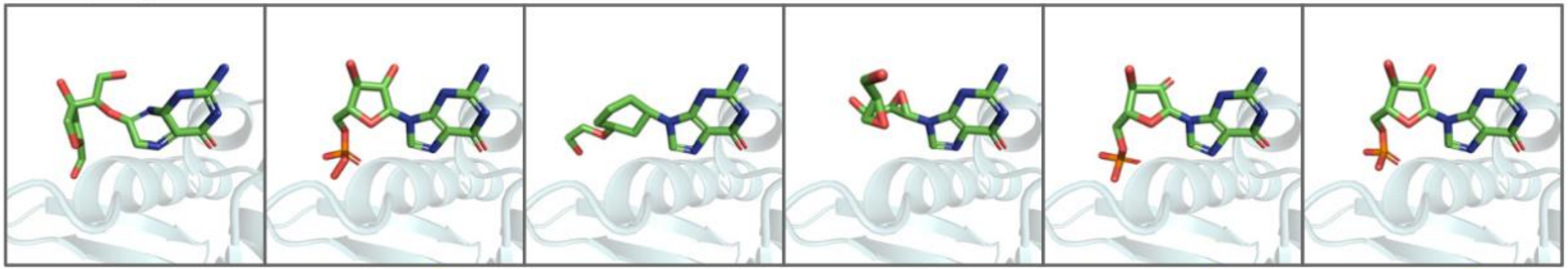
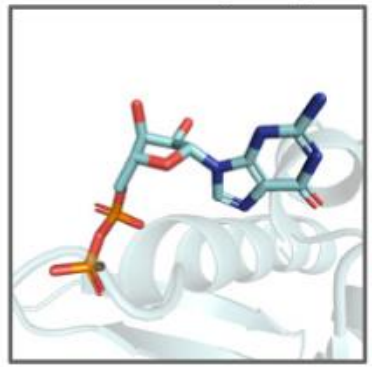


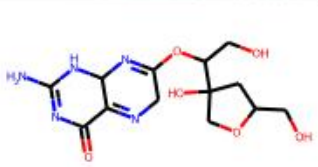
Figure 3: Distributions of Vina, QED, and SA scores over all the generated molecules.

Ours (2hcj)



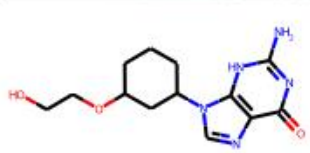
Reference (2hcj)





Vina: -7.3
QED: 0.35 SA: 0.50

Vina: -7.1
QED: 0.19 SA: 0.63

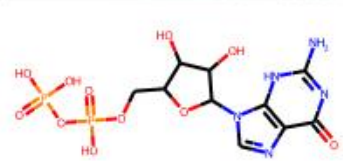


Vina: -6.8
QED: 0.74 SA: 0.72

Vina: -6.7
QED: 0.46 SA: 0.61

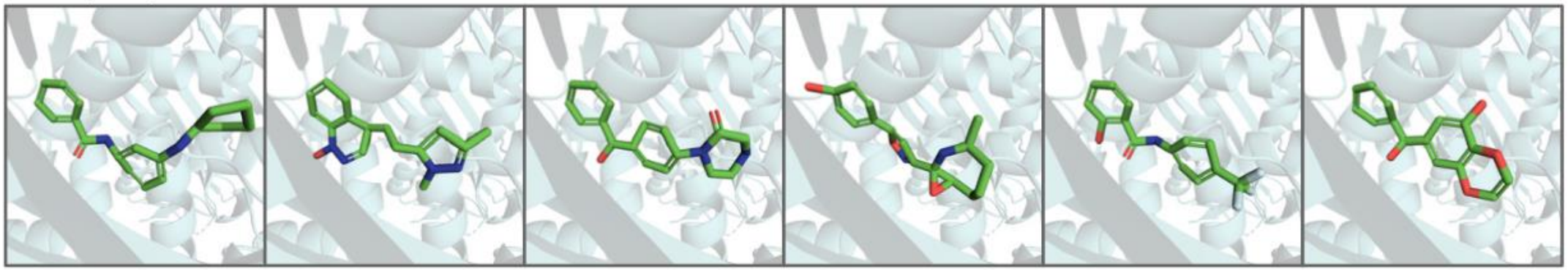
Vina: -6.7
QED: 0.36 SA: 0.64

Vina: -6.6
QED: 0.31 SA: 0.67

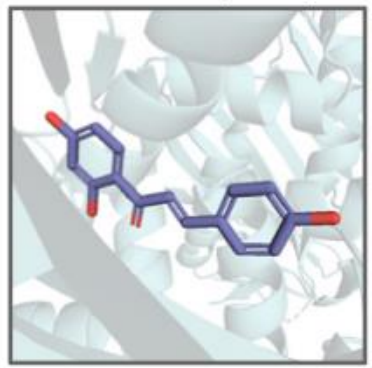


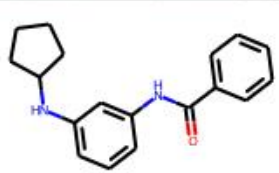
Vina: -6.5
QED: 0.23 SA: 0.64

Ours (4rlu)

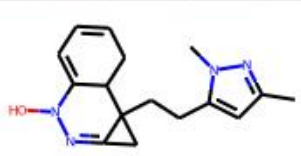


Reference (4rlu)

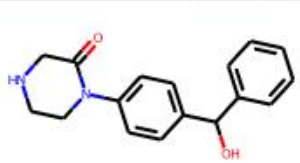




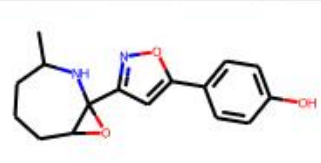
Vina: -9.4
QED: 0.88 SA: 0.93



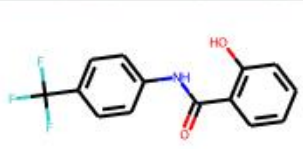
Vina: -9.3
QED: 0.93 SA: 0.60



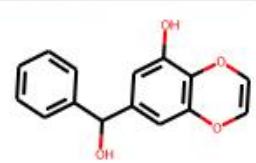
Vina: -9.1
QED: 0.90 SA: 0.83



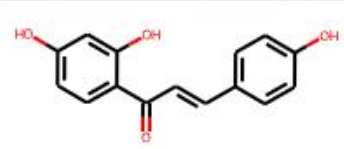
Vina: -9.1
QED: 0.83 SA: 0.66



Vina: -9.0
QED: 0.88 SA: 0.93



Vina: -9.0
QED: 0.87 SA: 0.75



Vina: -8.4
QED: 0.58 SA: 0.88

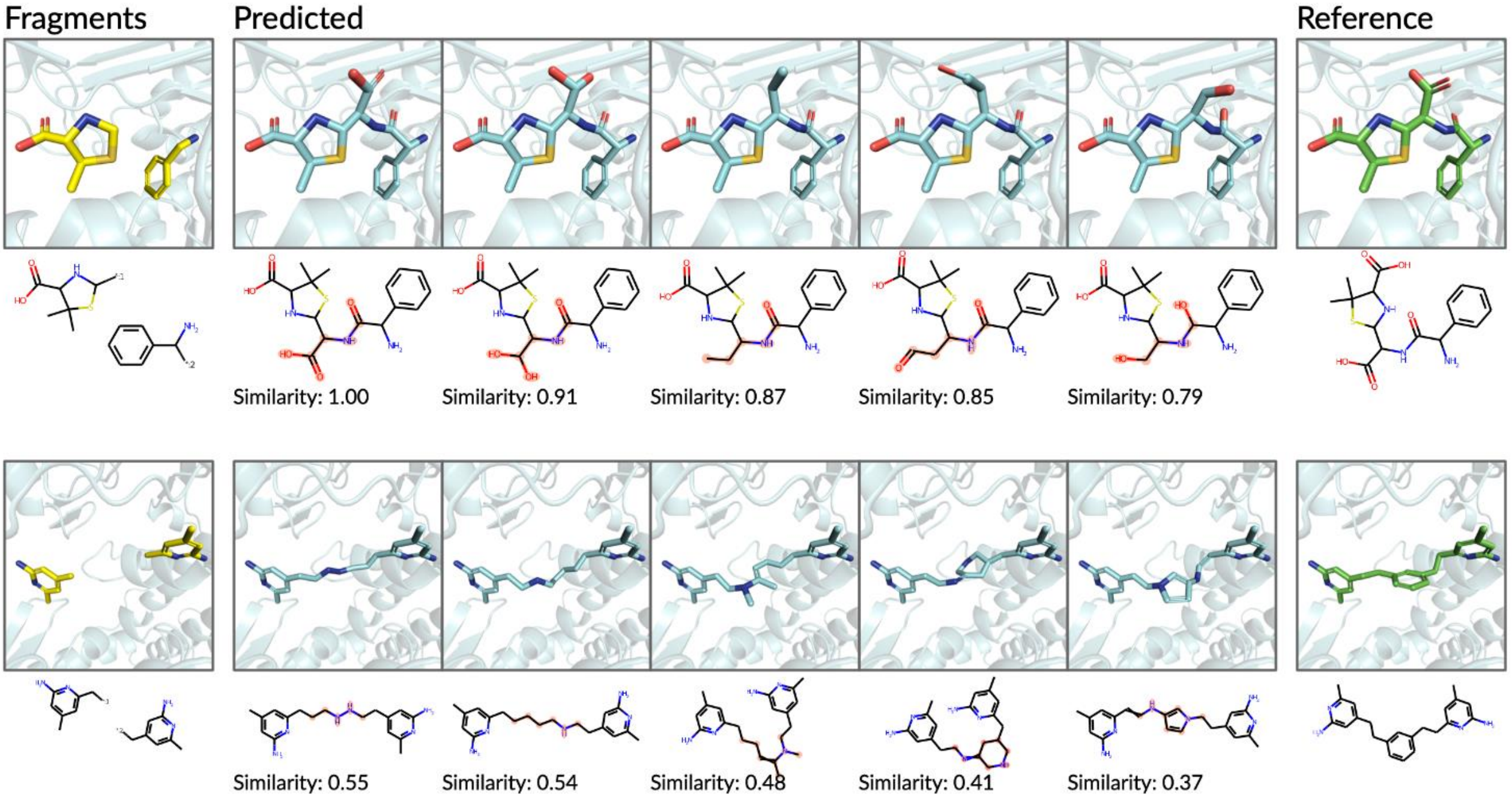
Linker prediction

- Data: 120 data points in total. Each of them consists of two disconnected molecule fragments
- Model is compared with DeLinker
- Metrics:
  - **Similarity**: Tanimoto Similarity over Morgan fingerprints
  - **Percentage of recovered molecules**: We calculate the percentage of test molecules that are recovered by the model
  - **Binding Affinity**: VinaScore

Table 2: Performance of linker prediction.

Metric		DeLinker	**Ours**
Similarity (↑)	Avg.	0.612	**0.701**
	Med.	0.600	**0.722**
Recovered (%, ↑)		40.00	**48.33**
Vina Score (kcal/mol, ↓)	Avg.	-8.512	-8.603
	Med.	-8.576	-8.575

ete.


Fragments
Predicted
Reference
Similarity: 1.00
Similarity: 0.91
Similarity: 0.87
Similarity: 0.85
Similarity: 0.79
Similarity: 0.55
Similarity: 0.54
Similarity: 0.48
Similarity: 0.41
Similarity: 0.37