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

- 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 $p(e|r, C)$

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

$k$-nearest-neighbor graph based on inter-atomic distances

$$\mathcal{G} = \langle \tilde{C}, \tilde{A} \rangle$$

$$h_i^{(\ell+1)} = \sigma \left( W_0 h_i^{(\ell)} + \sum_{j \in N_k(r_i)} W_1 w(d_{ij}) \odot W_2 h_j^{(\ell)} \right)$$

$$v\sum_{j \in N_k(r)} W_0 w_{aggr}(\|r - r_j\|) \odot W_1 h_j^{(L)}$$
Step 2: Sampling

Sampling a molecule amounts to generating a set of atoms \[\{(e_i, r_i)\}_{i=1}^{N_a}\].

**Joint Distribution** We define the joint distribution of coordinate \( r \) and atom type \( e \)

\[
p(e, r|C) = \frac{\exp(c[e])}{Z},
\]

**Auto-Regressive Sampling** We sample a molecule by progressively sampling one atom at each step. In specific, at step \( t \), the context \( C_t \) contains not only protein atoms but also \( t \) atoms sampled beforehand. Sampled atoms in \( C_t \) 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 \( p(e, r|C_t) \) and will be added to \( C_t \), leading to the context for next step \( C_{t+1} \).

\[
(e_{t+1}, r_{t+1}) \sim p(e, r|C_t), \quad C_{t+1} \leftarrow C_t \cup \{(e_{t+1}, r_{t+1})\}. \tag{7}
\]
$C_3 \leftarrow C_2 \cup \{(0, r_3)\}$
$p(e, r|C_3)$

$C_4 \leftarrow C_3 \cup \{(N, r_4)\}$
$p(e, r|C_4)$

Done.

$C_5 \leftarrow C_4 \cup \{(0, r_5)\}$

Done.
Step 3: Training

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

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

\[ L_F = \sum_{i \in \mathcal{F} \subseteq C} \log \sigma(F(h_i)) + \sum_{i \notin \mathcal{F} \subseteq C} \log(1 - \sigma(F(h_i))), \]

\[ L = L_{\text{BCE}} + L_{\text{CAT}} + L_F. \]
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:**

<table>
<thead>
<tr>
<th>Metric</th>
<th>liGAN</th>
<th>Ours</th>
<th>Ref</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Vina Score</strong></td>
<td>Avg.</td>
<td>-6.144</td>
<td>-6.344</td>
</tr>
<tr>
<td>(kcal/mol, ↓)</td>
<td>Med.</td>
<td>-6.100</td>
<td>-6.200</td>
</tr>
<tr>
<td><strong>QED (↑)</strong></td>
<td>Avg.</td>
<td>0.371</td>
<td>0.525</td>
</tr>
<tr>
<td></td>
<td>Med.</td>
<td>0.369</td>
<td>0.519</td>
</tr>
<tr>
<td><strong>SA (↑)</strong></td>
<td>Avg.</td>
<td>0.591</td>
<td>0.657</td>
</tr>
<tr>
<td></td>
<td>Med.</td>
<td>0.570</td>
<td>0.650</td>
</tr>
<tr>
<td><strong>High Affinity</strong></td>
<td>Avg.</td>
<td>23.77</td>
<td>29.09</td>
</tr>
<tr>
<td>(%, ↑)</td>
<td>Med.</td>
<td>11.00</td>
<td>18.50</td>
</tr>
<tr>
<td><strong>Diversity (↑)</strong></td>
<td>Avg.</td>
<td>0.655</td>
<td>0.720</td>
</tr>
<tr>
<td></td>
<td>Med.</td>
<td>0.676</td>
<td>0.736</td>
</tr>
</tbody>
</table>

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.
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.

<table>
<thead>
<tr>
<th>Metric</th>
<th>DeLinker</th>
<th>Ours</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Similarity (↑)</strong></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Avg.</td>
<td>0.612</td>
<td>0.701</td>
</tr>
<tr>
<td>Med.</td>
<td>0.600</td>
<td>0.722</td>
</tr>
<tr>
<td><strong>Recovered (%) (↑)</strong></td>
<td>40.00</td>
<td>48.33</td>
</tr>
<tr>
<td><strong>Vina Score (kcal/mol, ↓)</strong></td>
<td>Avg.</td>
<td></td>
</tr>
<tr>
<td></td>
<td>-8.512</td>
<td>-8.603</td>
</tr>
<tr>
<td></td>
<td>Med.</td>
<td>-8.576</td>
</tr>
</tbody>
</table>