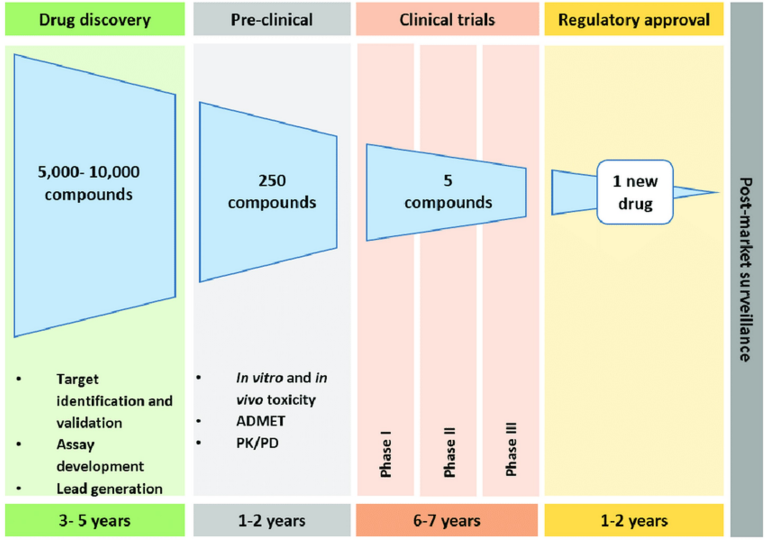
3DLinker: An E(3) Equivariant Variational Autoencoder for Molecular Linker Design

Yinan Huang[1], Xingang Peng[2], Jianzhu Ma[3,1], Muhan Zhang[3,1]

[1]Beijing Institute for General Artificial Intelligence
[2]Tsinghua University
[3]Institute for Artificial Intelligence, Peking University

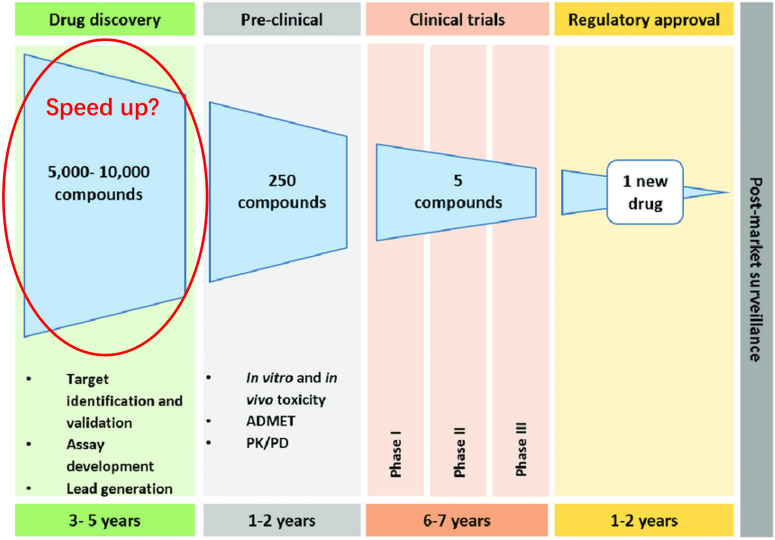
Drug Development Pipeline

It often takes more than a decade to develop even one new drug.



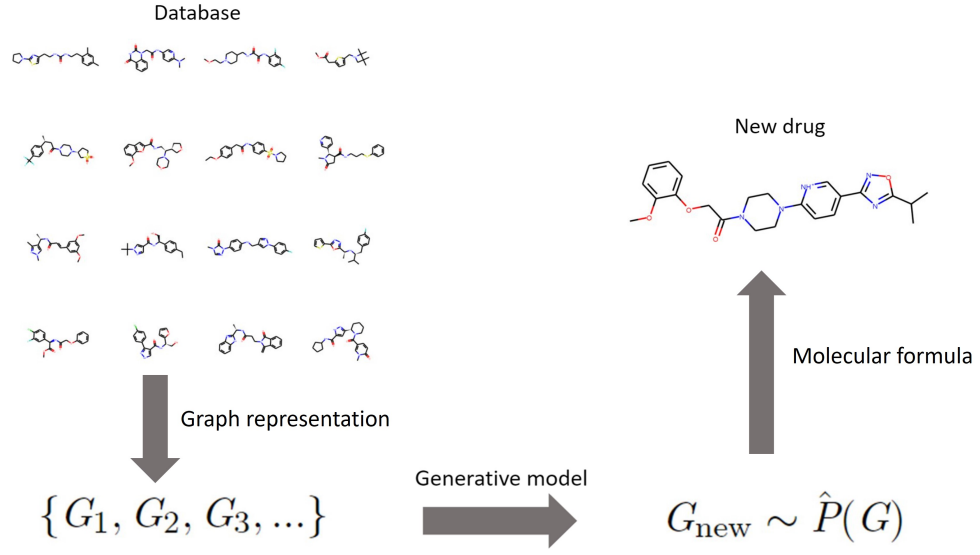
Drug Development Pipeline

Can we leverage machine learning to speed up drug design?



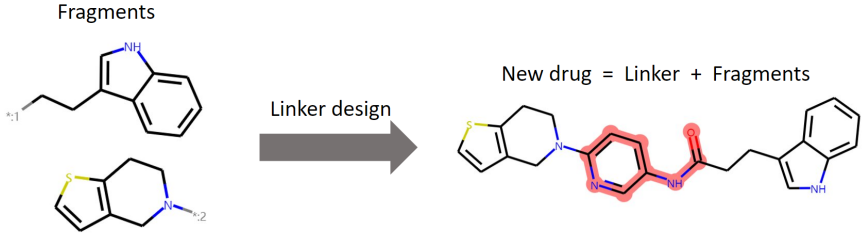
Drug Design using Machine Learning

A drug design algorithm can be described by a generative model.



Linker Design

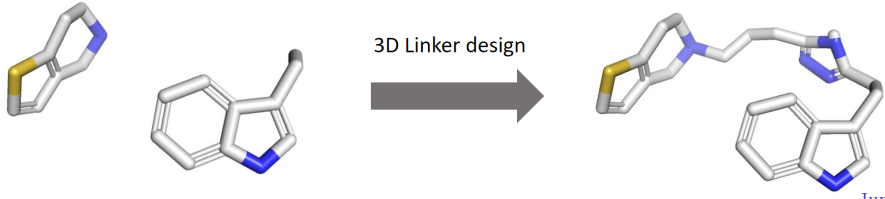
- However, generate the whole molecule from scratch is less than a practical approach.
- Linker design is a popular drug design method that finds a “linker” molecule attaching two given “fragment” molecules.
- The two given fragments are usually known to have specific chemical functions, thus making linker design more promising against designing the whole drug from scratch.
- Conditional graph generation, i.e. $P(G_{\text{link}}|G_{\text{frag}})$.



What We Propose: 3D Linker Design

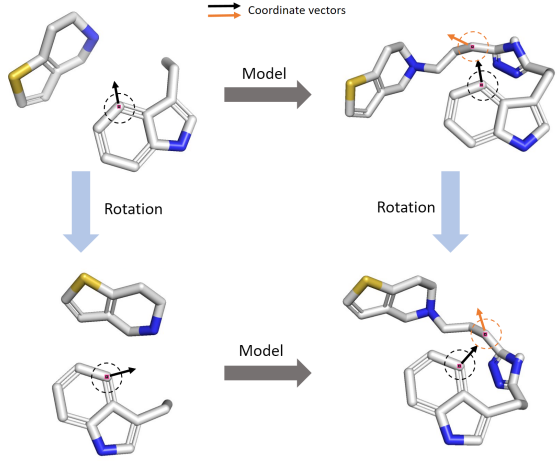
Key motivations:

- Previous works only generate the graphs and cannot generate the 3D coordinates of the linker. The lack of geometric information may lead to unrealistic generations that do not physically exist, or hard to adapted to downstream tasks that require 3D outputs (e.g. compute affinity to protein pocket).
- Therefore, we propose a new task named 3D linker design, that is to generate both the graph and the coordinates of the linker, given the graph and the coordinates of the fragments.
- Conditional 3D graph generation, i.e. $P(G_{\text{link}}, \boldsymbol{X}_{\text{link}} | G_{\text{frag}}, \boldsymbol{X}_{\text{frag}})$.
- Interestingly, this task can also be seen as a pre-training strategy that "masks" part of the 3D molecular graphs and learns to recover it.



Inductive Bias: Symmetry and Equivariance

The generated linker should be **insensitive** to rigid body transformations, i.e. E(3) group.

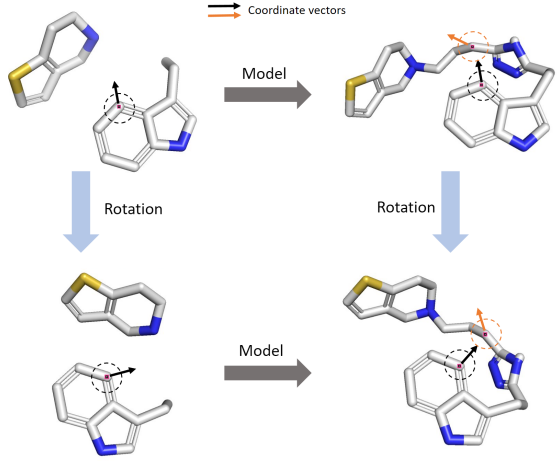


Inductive Bias: Symmetry and Equivariance

The generated linker should be **insensitive** to rigid body transformations, i.e. E(3) group.

- Graphs (atom types, edge types) are *invariant* to E(3).
- Coordinates are *equivariant* to E(3).

How to build a powerful model while satisfying the equivariance constraints?



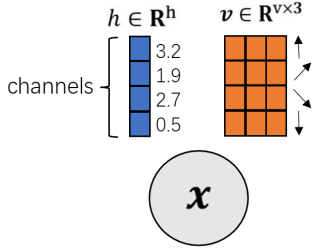
$$\forall g \in E(3), \quad P(G_{\text{link}}, \boldsymbol{X}_{\text{link}} | G_{\text{frag}}, \boldsymbol{X}_{\text{frag}}) = P(G_{\text{link}}, g \cdot \boldsymbol{X}_{\text{link}} | G_{\text{frag}}, g \cdot \boldsymbol{X}_{\text{frag}}).$$

Building Block: Mixed-Features Message Passing (MF-MP)

Main ideas:

- Each node has invariant and equivariant features.

  For example, a node at position $\boldsymbol{x}$ has invariant features $h \in \mathbf{R}^{\mathrm{h}}$ and equivariant features $\boldsymbol{v} \in \mathbf{R}^{\mathrm{v}\times 3}$. The row index represents different channels (so a node has v 3D "directional vectors/arrows" as its equivariant features).



Building Block: Mixed-Features Message Passing (MF-MP)

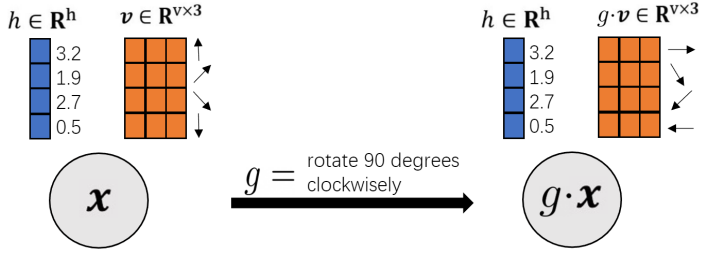
Main ideas:

- Each node has invariant and equivariant features.

  For example, a node at position $\boldsymbol{x}$ has invariant features $h \in \mathbf{R}^{\mathrm{h}}$ and equivariant features $\boldsymbol{v} \in \mathbf{R}^{\mathrm{v}\times 3}$. The row index represents different channels (so a node has v 3D "directional vectors/arrows" as its equivariant features).
  Transformation $g \in E(3)$ thus acts independently on every single row:

$$[g \cdot h]_i = h_i, \quad [g \cdot \boldsymbol{v}]_i = g \cdot \boldsymbol{v}_i.$$

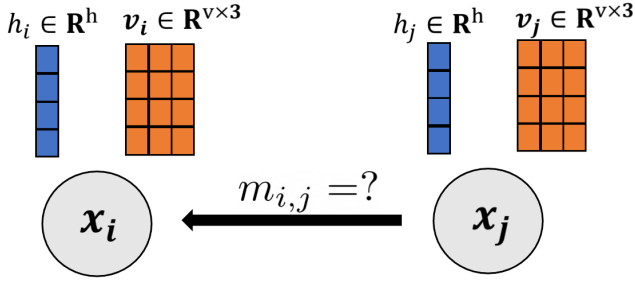
  As initialization we can let h be one-hot embedding of node types and $\boldsymbol{v}$ be all zeros.



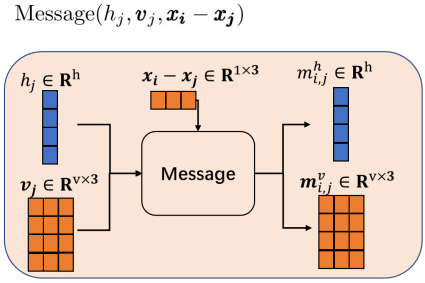
Building Block: Mixed-Features Message Passing (MF-MP)

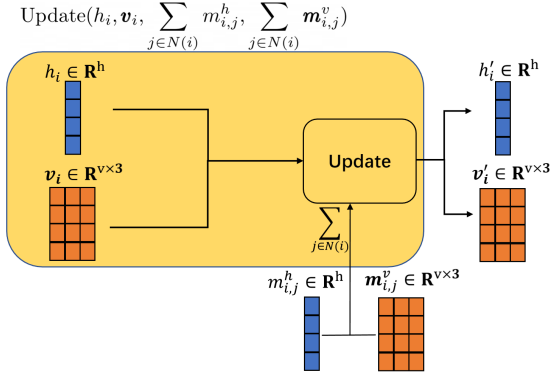
Main ideas:

- Each node has invariant and equivariant features.
- Message passing should preserve the equivariance.
  In message-passing scheme, one has to define *message function* and *update function*. They have to be designed properly in order to preserve equivariance.



Building Block: Mixed-Features Message Passing (MF-MP)





$$m^h_{i,j} = \mathrm{Ker}_1(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|) \odot h_j,$$
$$\boldsymbol{m}^v_{i,j} = \mathrm{diag}\left\{\mathrm{Ker}_2(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|)\right\} \cdot \boldsymbol{v}_j + \left(\mathrm{Ker}_3(\|\boldsymbol{x}_i - \boldsymbol{x}_j\|) \odot h_j\right) \cdot (\boldsymbol{x}_i - \boldsymbol{x}_j).$$

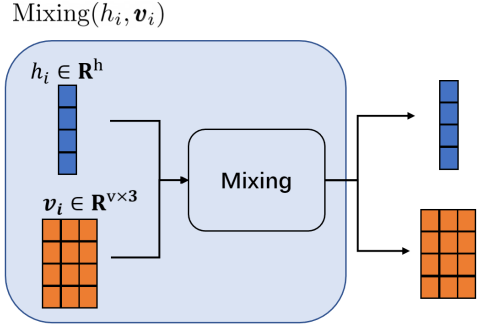
$$h'_i = \mathrm{GRU}(h_i, \sum_{j\in N(i)} m^h_{i,j}),$$
$$\boldsymbol{v}'_i = \mathrm{VN}(\boldsymbol{v}_i \oplus \sum_{j\in N(i)} \boldsymbol{m}^v_{i,j}).$$

- We have separate messages for invariant and equivariant features.

Here Ker is a radial basis kernel, and VN is an equivariant neuron mimicking normal neuron.

Building Block: Mixed-Features Message Passing (MF-MP)

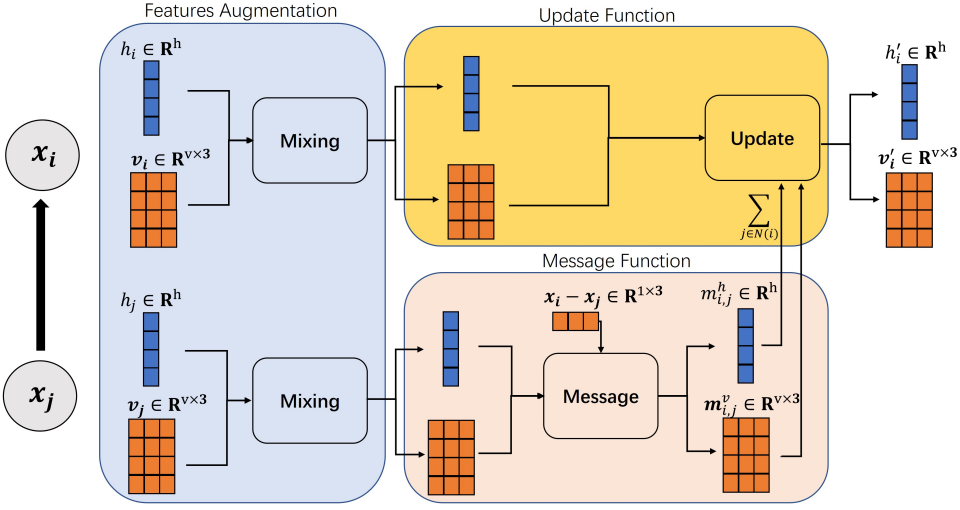
We also add a mixing unit to augment features before message passing:



$$h_i \to \mathrm{MLP}_1(h_i \oplus \|\mathrm{VN}_1(\boldsymbol{v}_i)\|),$$
$$\boldsymbol{v}_i \to \mathrm{diag}\{\mathrm{MLP}_2(h_i)\} \cdot \mathrm{VN}_2(\boldsymbol{v}_i).$$

Building Block: Mixed-Features Message Passing (MF-MP)

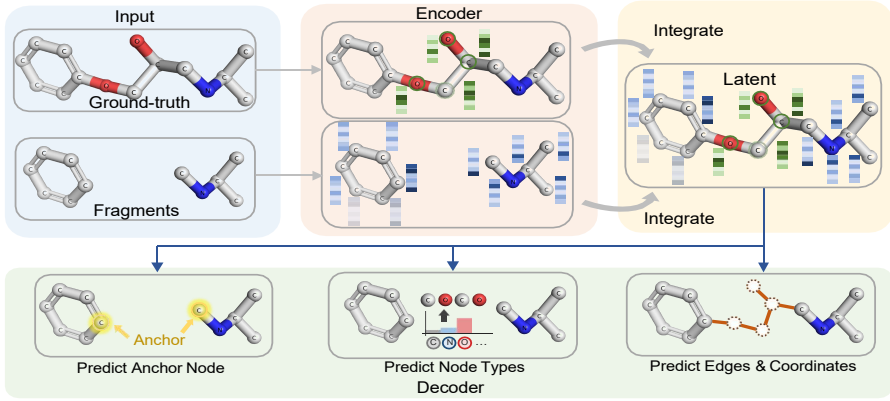
Let us get things together.



3DLinker: Architecture Overview

Using MF-MP as building blocks, 3DLinker follows an standard encoding-decoding process. Compared to a normal graph VAE, it has the following features:

- it not only encodes the ground truth, but also the fragments, reflecting it is a conditional generation;
- the encoding and decoding process preserve equivariance.



Experimental Results

- Since there are rarely 3D graph generative models, we integrate graph generative models (DeLinker, GraphAF, GraphVAE, SyntaLinker) and a graph-to-3D generative model (ConfVAE) as baselines.
- We conducted experiments on ZINC dataset with comparison to these baselines. Variants of metrics are estimated by samples generated from each models.

Metrics (%)	Valid	Recover	Pass 2D filters	S_{tani}	RMSD (Å)	Unique	Novel
3DLinker (ours)	**98.68**$_{\pm 0.03}$	**93.75**$_{\pm 0.36}$	90.32$_{\pm 0.18}$	**54.83**$_{\pm 0.08}$	**0.081**$_{\pm 0.004}$	29.34$_{\pm 0.17}$	32.53$_{\pm 0.12}$
DeLinker+ConfVAE	98.35$_{\pm 0.07}$	80.35$_{\pm 2.58}$	89.92$_{\pm 0.53}$	52.86$_{\pm 0.04}$	1.345$_{\pm 0.028}$	44.53$_{\pm 0.42}$	39.53$_{\pm 0.58}$
GraphAF+ConfVAE	34.25$_{\pm 0.17}$	21.23$_{\pm 1.82}$	82.00$_{\pm 2.21}$	38.00$_{\pm 0.14}$	1.263$_{\pm 0.124}$	84.06$_{\pm 0.31}$	78.33$_{\pm 0.12}$
GraphVAE+ConfVAE	63.07$_{\pm 0.59}$	1.40$_{\pm 1.37}$	86.17$_{\pm 0.26}$	51.31$_{\pm 0.07}$	1.523$_{\pm 0.478}$	43.59$_{\pm 0.35}$	91.59$_{\pm 1.47}$
SyntaLinker+ConfVAE	80.14$_{\pm 0.31}$	85.47$_{\pm 2.46}$	97.41$_{\pm 0.11}$	54.80$_{\pm 0.05}$	1.402$_{\pm 0.016}$	41.51$_{\pm 0.75}$	13.17$_{\pm 0.27}$
Gen3D	75.92$_{\pm 0.10}$	37.99$_{\pm 0.67}$	81.41$_{\pm 0.12}$	48.67$_{\pm 0.06}$	1.415$_{\pm 0.272}$	49.86$_{\pm 0.49}$	64.17$_{\pm 0.21}$

- Particularly, “Recover” represents the probability of generating the ground-truth linker, while RMSD measures the deviation of coordinates between ground-truth and generated ones. Our model shows a significant improvement over these two metrics.

Å$= 10^{-10}$m.

Experimental Results

Metrics (%)	Valid	Recover	Pass 2D filters	S_{tani}	RMSD (Å)	Unique	Novel
3DLinker (ours)	**98.68**$_{\pm 0.03}$	**93.75**$_{\pm 0.36}$	90.32$_{\pm 0.18}$	**54.83**$_{\pm 0.08}$	**0.081**$_{\pm 0.004}$	29.34$_{\pm 0.17}$	32.53$_{\pm 0.12}$
DeLinker+ConfVAE	98.35$_{\pm 0.07}$	80.35$_{\pm 2.58}$	89.92$_{\pm 0.53}$	52.86$_{\pm 0.04}$	1.345$_{\pm 0.028}$	44.53$_{\pm 0.42}$	39.53$_{\pm 0.58}$
GraphAF+ConfVAE	34.25$_{\pm 0.17}$	21.23$_{\pm 1.82}$	82.00$_{\pm 2.21}$	38.00$_{\pm 0.14}$	1.263$_{\pm 0.124}$	84.06$_{\pm 0.31}$	78.33$_{\pm 0.12}$
GraphVAE+ConfVAE	63.07$_{\pm 0.59}$	1.40$_{\pm 1.37}$	86.17$_{\pm 0.26}$	51.31$_{\pm 0.07}$	1.523$_{\pm 0.478}$	43.59$_{\pm 0.35}$	91.59$_{\pm 1.47}$
SyntaLinker+ConfVAE	80.14$_{\pm 0.31}$	85.47$_{\pm 2.46}$	97.41$_{\pm 0.11}$	54.80$_{\pm 0.05}$	1.402$_{\pm 0.016}$	41.51$_{\pm 0.75}$	13.17$_{\pm 0.27}$
Gen3D	75.92$_{\pm 0.10}$	37.99$_{\pm 0.67}$	81.41$_{\pm 0.12}$	48.67$_{\pm 0.06}$	1.415$_{\pm 0.272}$	49.86$_{\pm 0.49}$	64.17$_{\pm 0.21}$

We also conducted ablation study, showing the effectiveness of using equivariant features.

Metrics(%)	Valid	Recovered	Pass 2D filters	RMSD(Å)	Unique	novel
3DLinker	98.67	93.58	90.37	0.079	29.42	32.48
3DLinker (w/o equivariant)	99.42	86.59	92.68	1.352	34.58	27.02
3DLinker (w/o update)	98.85	39.94	62.81	0.399	55.93	72.25

Conclusions and Limitations

Conclusions

- Linker design is one of the most popular drug design techniques. However, many previous works either focus on generation from scratch or cannot generate 3D coordinates.
- we thus come up with a new task: generate both graph and coordinates of the linker given two 3D fragments. This task can also be seen as a pre-training process on 3D graphs.
- We propose mixed-features message passing (MF-MP) as building blocks and develop an equivariant VAE called 3DLinker to address the task above. Thanks to equivariance, 3DLinker is insensitive to choices of coordinate system and thus generalizes well in such 3D task. Finally it shows a significant boost of important metrics (recovery rate, RMSD, etc.) over multiple baselines in the experiments.

Limitations

- One concern is that in practice the relative position of two fragments may not be known, which is assumed to be given in our setting. In that case it needs an additional step to predict 6DOFs (displacement and rotation) between two fragments before applying 3DLinker.

Thank you!

Paper: https://arxiv.org/abs/2205.07309
Code: https://github.com/GraphPKU/3DLinker