Research Square

Preprints are preliminary reports that have not undergone peer review.
They should not be considered conclusive, used to inform clinical practice,
or referenced by the media as validated information.

GeoLinker: An Equivariant Diffusion Model for 3D Molecular Linker Design

Srajal Tiwari
Shiv Nadar Institution of Eminence Deemed to be University
Dolly Sharma
dolly.sharma@snu.edu.in

Shiv Nadar Institution of Eminence Deemed to be University
Ranjit Kumar
Shiv Nadar Institution of Eminence Deemed to be University

Research Article

Keywords: fragment-based drug discovery, molecular linker design, 3D molecular generation, equivariant diffusion, SE(3)-equivariance, cheminformatics

Posted Date: September 25th, 2026

DOI: https://doi.org/10.21203/rs.3.rs-10622385/v1

License: This work is licensed under a Creative Commons Attribution 4.0 International License.
Read Full License

Additional Declarations: No competing interests reported.

GeoLinker: An Equivariant Diffusion Model for 3D Molecular Linker Design

Srajal Tiwari[1], Dolly Sharma[1*], Ranjit Kumar[2]

[1]Department of Computer Science and Engineering, Shiv Nadar Institution of Eminence Deemed to be University, NH-91, Delhi-NCR, Uttar Pradesh, 201314, India.
[2]Department of Chemical Engineering, Shiv Nadar Institution of Eminence Deemed to be University, NH-91, Delhi-NCR, Uttar Pradesh, 201314, India.

*Corresponding author(s). E-mail(s): dolly.sharma@snu.edu.in;

Abstract

Fragment-based drug discovery combines weakly binding molecular fragments to produce more potent candidates, but designing three-dimensional linkers that connect fragments while preserving their geometry and chemical validity remains challenging. Existing generative methods typically optimize for a specific conditioning setting, limiting flexibility across heterogeneous fragment-design scenarios. We present GeoLinker, a unified SE(3)-equivariant diffusion framework supporting multiple fragment-linking scenarios within a single training framework: unconstrained, anchor-guided, size-conditioned, and jointly anchor- and size-conditioned generation. Anchor-aware cross-attention and learned size embeddings incorporate user-defined geometric priors within a shared Equivariant Graph Neural Network backbone. Auxiliary endpoint, steric-clash, and non-anchor repulsion losses promote connected and physically plausible structures. On the ZINC benchmark, GeoLinker achieves over 80% novelty and consistently higher uniqueness than the evaluated baselines, while experiments on CASF-2016 and GEOM-drugs assess transfer to bioactive and structurally diverse molecular settings. Case studies demonstrate the practical applicability of the proposed framework in representative fragment-linking tasks.

Scientific Contribution: GeoLinker provides a single SE(3)-equivariant diffusion framework supporting four linker-design constraint modes within one training framework. Anchor-aware attention, size conditioning, and topology-oriented auxiliary losses provide flexible 3D linker generation across ZINC, CASF-2016, and GEOM-drugs evaluation settings.

Keywords: fragment-based drug discovery, molecular linker design, 3D molecular generation, equivariant diffusion, SE(3)-equivariance, cheminformatics

1 Introduction

The space of pharmacologically significant compounds, estimated to have more than 10^{60} structures, poses a big challenge for drug design [1]. Fragment-Based Drug Discovery (FBDD) provides a systematic strategy for assembling high-affinity leads from low-molecular-weight fragments whose binding modes are individually characterised by high-resolution structural data [2, 3]. The clinical success of this strategy depends critically on Structure-Based Drug Design (SBDD). Because fragments exhibit weak initial binding affinities, high-resolution structural data is essential to provide a geometric blueprint for their subsequent optimization. This process frequently involves combining multiple fragments into a single high-affinity molecule, this task presents a design challenge that the chemical linker must be synthetically feasible while perfectly preserving the original binding site of the fragments, governed by 3D geometric constraints [3, 4].

Previous methods for molecular linker design were based on database searches or relied on 1D and 2D molecular representations, such as SMILES strings [6]. Such techniques are limited in their ability to represent the three-dimensional geometry of the problem [7]. Graph-based approaches such as DeLinker condition molecular graph generation on inter-fragment 3D distances, yet their sequential atom-by-atom construction limits holistic reasoning over the final molecule's global geometry [8]. The field has increasingly adopted SE(3)-equivariant diffusion probabilistic models operating directly on 3D atomic point clouds [9, 10], enabling geometrically consistent linker generation without molecular graph pre-specification. Existing methods have demonstrated excellent performance under specific conditioning settings; however, practical fragment-design workflows frequently require switching between different constraint configurations within the same project. A unified framework capable of consistently supporting multiple conditioning scenarios remains practically valuable for heterogeneous fragment-design workflows. In structure-based drug design, recent equivariant diffusion probabilistic models have shown strong performance. Existing methods differ in the types of conditioning information they utilize and the design assumptions under which they are evaluated. DiffLinker [11] conditions linker generation on fragment geometry and optional design information, whereas DiffSBDD [12] incorporates protein structural context. LinkerNet [15] and DAD-PROTAC [16] extend this area through joint fragment-pose and linker co-design and domain adaptation, respectively. Applications such as PROteolysis Targeting Chimaeras (PROTACs), in which linkers influence three-dimensional ternary-complex formation, place additional geometric and chemical demands on generative models [13, 14].

Here, we present GeoLinker, an SE(3)-equivariant diffusion model that addresses a practical limitation in flexible fragment-linker generation through a unified, multi-constraint framework. The modular architecture supports four design modes, ranging from unconstrained generation to jointly anchor-guided and size-conditioned generation. Anchor-aware cross-attention and learned size embeddings fuse geometric priors with

the diffusion dynamics, while endpoint-distance, non-anchor-repulsion, and steric-clash losses encourage connected and physically plausible structures. This design provides a flexible framework for fragment-linker generation under multiple conditioning scenarios.

Figure 1 summarizes the generation task, conditioning variants, diffusion process, network architecture, and molecular-graph inference pipeline.

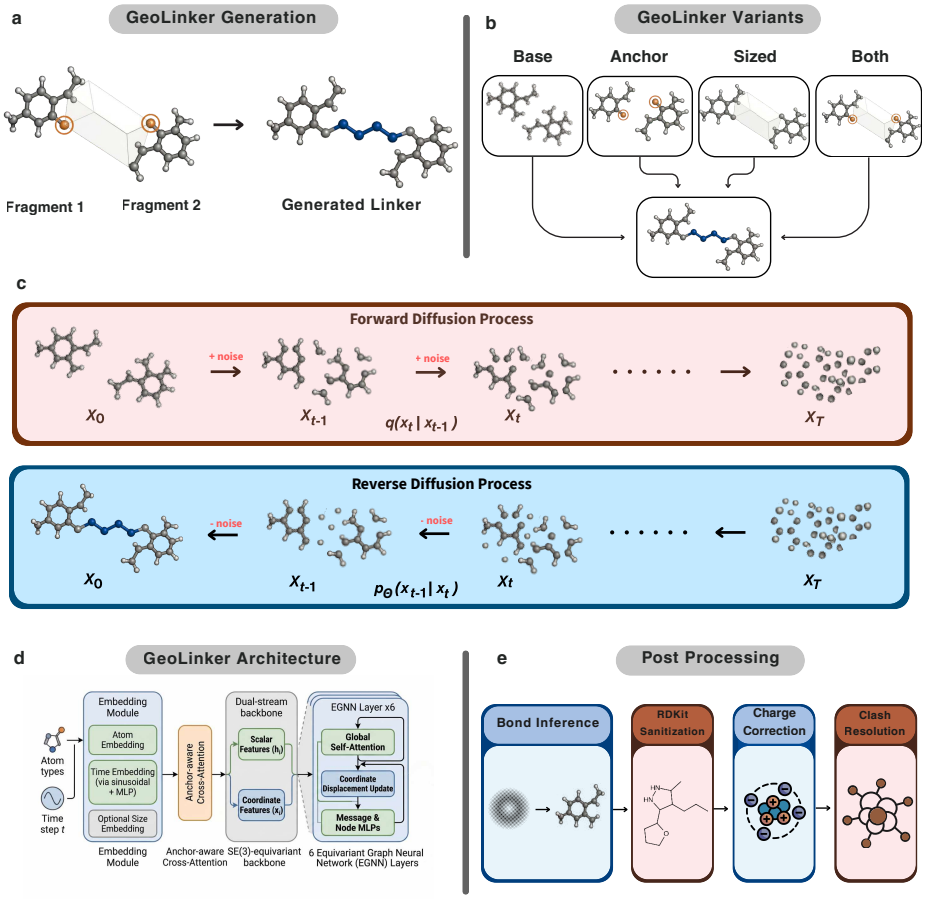


Fig. 1 Overview of the proposed GeoLinker workflow. a, Fragment-to-linker generation. b, Four design variants spanning unconstrained to fully constrained generation. c, Forward (top) and reverse (bottom) equivariant diffusion process. d, GeoLinker network architecture comprising the embedding module, anchor-aware cross-attention, SE(3)-equivariant dual-stream backbone, and six Equivariant Graph Neural Network (EGNN) layers. e, Post-processing pipeline used to infer molecular graphs from generated atomic point clouds.

2 Results

Table 1 Performance comparison on ZINC, CASF-2016, and GEOM-drugs benchmarks.

Methods	QED↑	SA↓	Rings↑	Valid (%)	Unique (%)	Novel (%)
ZINC						
DeLinker+ConfVAE+MMFF†	0.64	3.11	0.21	98.3	44.2	47.1
3DLinker†	0.65	3.14	0.24	71.5	29.2	41.9
3DLinker (given anchors)†	0.65	3.11	0.23	99.3	29.0	41.2
DiffLinker†	0.68	3.01	0.25	93.8	24.0	30.3
DiffLinker (given anchors)†	0.68	3.03	0.26	97.6	22.7	32.4
DiffLinker (sampled size)†	0.65	3.19	0.32	90.6	51.4	42.9
DiffLinker (anchors+size)†	0.65	3.24	0.36	94.8	50.9	47.7
GeoLinker (base)	0.50	5.14	0.34	45.5	**86.4**	**78.7**
GeoLinker (anchors)	0.49	5.04	**0.57**	44.1	**86.4**	**82.1**
GeoLinker (sized)	0.49	5.05	**0.63**	45.3	**88.7**	**81.9**
GeoLinker (both)	0.49	5.02	**0.54**	44.8	**87.5**	**80.8**
CASF						
DiffLinker†	0.41	4.00	0.34	85.3	40.5	41.8
DiffLinker (given anchors)†	0.40	4.03	0.38	90.2	37.3	48.4
DiffLinker (sampled size)†	0.40	4.06	0.30	63.7	60.0	49.3
DiffLinker (anchors+size)†	0.40	4.10	0.38	68.4	57.1	56.9
GeoLinker (base)‡	0.36	5.22	**0.88**	57.6	**86.2**	**71.4**
GeoLinker (anchors)‡	0.37	5.22	**0.89**	57.7	**86.4**	**72.6**
GeoLinker (sized)‡	0.37	5.19	**0.89**	59.0	**85.8**	**71.8**
GeoLinker (both)‡	0.37	5.31	**1.00**	58.5	**85.9**	**70.6**
GEOM						
DiffLinker†	0.48	2.99	0.75	93.4	31.7	68.6
DiffLinker (given anchors)†	0.49	3.02	0.79	93.5	32.1	68.4
DiffLinker (sampled size)†	0.45	3.27	0.76	87.1	57.3	76.1
DiffLinker (anchors+size)†	0.46	3.33	0.84	88.6	58.2	76.1
GeoLinker (base)	0.17	5.72	**2.67**	**98.0**	**98.5**	**98.8**
GeoLinker (anchors)	0.17	5.89	**3.04**	**98.0**	**99.2**	**98.0**
GeoLinker (sized)	0.17	5.87	**2.95**	**98.0**	**98.3**	**98.3**
GeoLinker (both)	**0.17**	**6.09**	**3.36**	**98.0**	**99.0**	**98.5**

†Baseline metrics retrieved from the official DiffLinker benchmarking protocols [11].
‡GeoLinker models evaluated zero-shot on CASF-2016 with no target-pose or pocket-conditioning fine-tuning.

GeoLinker achieves robust generalization across three benchmark datasets. These datasets are in the order of increasing structural and biological complexity, from in-distribution fragment pairs of ZINC to experimentally determined protein–ligand co-crystal structures CASF-2016 complexes and structurally diverse conformational ensembles GEOM-drugs. For each dataset, the proposed model variants (Base, Anchor, Sized, and Both) were evaluated, with 250 linkers generated per input fragment pair. As detailed in Table 1, GeoLinker achieves 44.1%–45.5% chemical validity on the highly constrained ZINC benchmark, lower than graph-based methods whose bond topology

is explicitly modeled. However, GeoLinker consistently achieves higher structural diversity, attaining 86.4%–88.7% uniqueness and 78.7%–82.1% novelty, compared with DiffLinker's ranges of 22.7%–51.4% and 30.3%–47.7%, respectively. GeoLinker also generates linkers with greater topological complexity, reaching 0.63 rings per linker compared with at most 0.36 for DiffLinker. These results suggest a trade-off between strict sanitization success and exploration of diverse bridging scaffolds for early-stage fragment-based drug discovery.

GeoLinker prioritizes structural diversity over strict chemical validity, highlighting an explicit diversity–validity trade-off.

2.1 Adaptation to Bioactive Poses (CASF-2016)

Using the CASF-2016 benchmark of high-quality protein-ligand co-crystal structures, we assessed zero-shot generalisation to bioactive chemical space. This evaluation is particularly stringent because no protein-pocket features, target-pose supervision, or CASF-specific fine-tuning are used, and the model must generate linkers conditioned solely on fragment geometry and optional design constraints. GeoLinker achieves validity of 57.6%–59.0%, uniqueness of 85.8%–86.4%, and novelty of 70.6%–72.6% across its four inference modes. Ring content increases to 0.88–1.00 compared with 0.30–0.38 for the DiffLinker baselines, while SA scores of 5.19–5.31 indicate that this added structural diversity is accompanied by a synthetic-accessibility trade-off.

2.2 Generalisation to Structurally Diverse Molecules (GEOM)

Performance on the GEOM-drugs dataset assesses GeoLinker's behavior on structurally diverse, conformationally flexible compounds. All four modes achieve 98.0% validity, 98.3%–99.2% uniqueness, and 98.0%–98.8% novelty while producing higher ring counts (2.67–3.36 rings per linker) than the DiffLinker baselines (0.75–0.84). The Sized mode is consistent with the other GeoLinker variants, achieving QED 0.17, SA 5.87, 2.95 rings per linker, 98.3% uniqueness, and 98.3% novelty. These results suggest that GeoLinker retains high generative diversity across all four conditioning modes on structurally heterogeneous molecules.

2.3 Case Studies

To assess GeoLinker's applicability in a representative fragment-linking task, we applied it to the Hsp90 ATPase pocket case study originally reported by Woodhead et al. [42] and subsequently used as a benchmark by DiffLinker. Two fragments experimentally identified in the ATPase pocket of Hsp90 (PDB: 3HZ1), an aminopyrimidine and a resorcinol derivative, were used as input to GeoLinker without pocket conditioning. After removing the methyl ester group following the same procedure as DiffLinker, the fragments were provided in their experimentally determined 3D coordinates. GeoLinker-Both generated sanitized connected molecules linking the two fragments through a three-carbon chain and recovered the linker topology of the reference inhibitor (PDB: 3HZ5; Tanimoto similarity = 0.32). The representative pose in Fig. 2 suggests that the framework can support this fragment-linking configuration without explicit

pocket conditioning. The limited diversity (two unique molecules) reflects the out-of-distribution nature of the Hsp90 fragments relative to the ZINC training set and suggests that training on protein-fragment datasets could improve performance in structure-based applications.

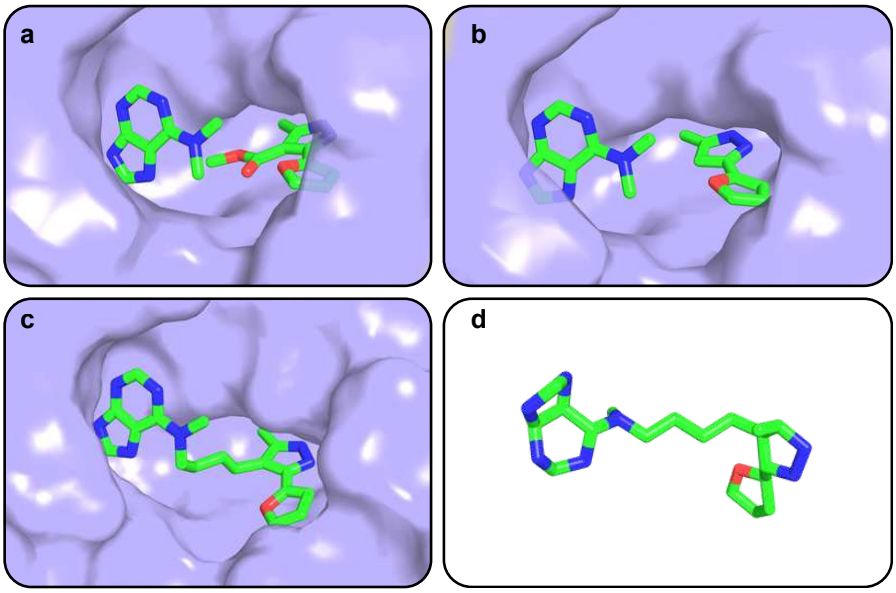


Fig. 2 GeoLinker applied to Hsp90 fragment linking. a, Experimentally identified fragment hits in the ATPase pocket of Hsp90 (PDB: 3HZ1) [42]. b. Input fragments after removal of the methyl ester group. c. Representative GeoLinker-generated molecule shown docked in the Hsp90 binding pocket. d, Three-dimensional structure of the generated molecule recovering the three-carbon linker topology of the reference inhibitor AT13387 (PDB: 3HZ5) (Tanimoto similarity = 0.32).

The IMPDH case study presents two structurally complex fragments, a methylimidazolidine-cyclohexyl unit and a bromocyclohexyl-imidazolidine unit, connected through a short six-atom linker bridging anchor carbons at positions 3 and 18 of the full molecular graph. Figure 3 illustrates the generation pipeline: panel (a) shows the input fragments and attachment points; panel (b) depicts the fragments in three-dimensional space with the linker region initialized as Gaussian noise between the anchors; panel (c) presents the reference molecule; and panel (d) shows a representative molecule generated by GeoLinker-Both using anchor and linker-size conditioning. The generated molecule achieves a Tanimoto similarity of 0.643 to the reference while preserving both input fragments, suggesting that the model can propose heteroatom-containing bridging motifs consistent with the supplied fragment geometry.

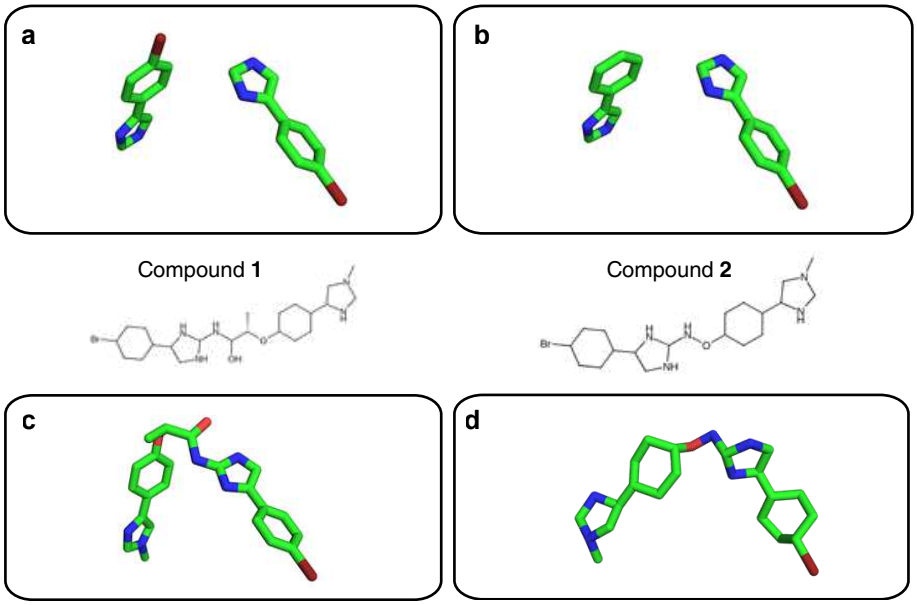


Fig. 3 Case study of fragment-based IMPDH inhibitor design. a, Experimentally identified fragment hits [41] bound to IMPDH (PDB: 5OU2), with the protein shown in blue. b, Starting fragments obtained after bromine removal. c, Crystal structure of the most potent inhibitor (compound 31) [41] (PDB: 5OU3). d, A representative molecule generated by GeoLinker that reconstructs compound 31 with a Tanimoto similarity of 0.643.

GeoLinker exhibits an explicit diversity–validity trade-off. While graph-based approaches optimize discrete bond topology during generation and therefore achieve higher chemical validity, the proposed coordinate-first framework explores a broader structural space, leading to substantially higher uniqueness and novelty. Future work will focus on integrating bond prediction directly into the diffusion process to improve validity while preserving diversity.

3 Discussion

The proposed model, GeoLinker, presents a unified SE(3)-equivariant diffusion framework for three-dimensional molecular linker design across multiple conditioning regimes. Practical fragment-design campaigns may operate with different levels of information at different stages; supporting unconstrained, anchor-guided, size-conditioned, and jointly constrained generation within one architecture provides a practical solution for these heterogeneous scenarios. On ZINC, GeoLinker achieves higher uniqueness, novelty, and ring content than the evaluated DiffLinker settings, but lower chemical validity. The results therefore indicate competitive diversity characteristics despite lower validity, while maintaining high diversity across multiple conditioning scenarios. Future work will investigate joint bond-coordinate generation to improve chemical validity while preserving structural diversity.

The current framework has several limitations. First, bond topology is inferred post-hoc from generated atomic coordinates using a greedy valence-constrained perception procedure followed by MMFF94 relaxation. This separation between coordinate generation and bond inference can produce locally implausible bonding environments, particularly at ring junctions and heteroatom-dense linker segments. Second, fragment poses remain fixed throughout generation, although fragment reorientation may occur during linker optimization. Third, the model does not use protein-pocket conditioning, so the case studies evaluate geometric fragment linking rather than prospective protein-aware molecular design. Joint atom-type, coordinate, and bond generation, dynamic fragment-pose optimization [15], and explicit pocket conditioning [12] are therefore important directions for further development.

Synthetic accessibility provides an additional limitation. Although GeoLinker samples structurally diverse molecules, synthesis feasibility is inferred only indirectly from the training distribution. Incorporating reaction-aware constraints, retrosynthetic scores, or synthesis-based filtering into training and sampling may improve the practical prioritization of generated candidates.

GeoLinker is trained exclusively on ZINC fragment pairs, whose linker-length distribution is compact, with a modal length of three to four atoms. Applications requiring substantially longer or more conformationally flexible linkers, including PROTAC ternary-complex design, are likely to be outside the current training distribution. Extending training to more diverse linker lengths and protein-associated fragment pairs will be necessary to assess these applications.

4 Conclusions

GeoLinker supports unconstrained, anchor-guided, size-conditioned, and jointly constrained 3D molecular linker generation within a single SE(3)-equivariant diffusion architecture. Across ZINC, CASF-2016, and GEOM-drugs, the proposed framework maintains high structural diversity across multiple evaluation settings while exhibiting dataset-dependent chemical validity. These findings support flexible conditioning as a practical direction for fragment-based design and motivate future joint modelling of coordinates, atom types, bonds, synthesis constraints, fragment poses, and protein-pocket context.

5 Methods

5.1 Problem Formulation

Molecular linker design is formulated as a conditional three-dimensional generative problem. Given two molecular fragments $\mathcal{F}_1$ and $\mathcal{F}_2$ represented as atomic point clouds, the objective is to generate a chemically valid linker $\mathcal{L}$ whose geometry bridges both fragments into a topologically connected molecule. The model learns the conditional distribution:

$$p_\theta(\boldsymbol{X}_{\text{link}}, \boldsymbol{H}_{\text{link}} \mid \boldsymbol{X}_{\text{frag}}, \boldsymbol{H}_{\text{frag}}, \mathcal{C}) \tag{1}$$

where $\boldsymbol{X}_{\text{link}} \in \mathbb{R}^{N_{\text{link}} \times 3}$ and $\boldsymbol{H}_{\text{link}} \in \{0,1\}^{N_{\text{link}} \times 8}$ are the positions and atom types of the N_{link} linker atoms, and $\mathcal{C}$ denotes optional design constraints (anchor atoms

and/or target linker size). This treats generation as a geometric inpainting problem over the fragment point cloud (Supplementary Note 1).

5.2 Model Architecture

GeoLinker employs an SE(3)-equivariant backbone comprising $L = 6$ E(n) Equivariant Graph Neural Network (EGNN) layers [19]. The network processes invariant scalar features $\boldsymbol{h}_i$ and equivariant coordinate features $\boldsymbol{x}_i$. Coordinate updates are computed by summing the relative displacement vectors between neighboring atoms, weighted by an invariant message MLP, ensuring SE(3)-equivariance by construction. To stabilize high-noise diffusion trajectories and prevent gradient explosions, the coordinate updates are bounded using a Tanh activation. Full architectural specification and hyperparameters are provided in Supplementary Note 1 and Supplementary Table S3.

Four design variants are implemented within this shared backbone. **GeoLinker-Base** generates linkers conditioned exclusively on fragment geometry, with an auxiliary anchor prediction head providing implicit attachment guidance. **GeoLinker-Anchor** incorporates known attachment atoms via anchor-aware cross-attention. Here linker atom representations attend over anchor atom representations, and anchor positions are held fixed throughout reverse diffusion. **GeoLinker-Sized** conditions on target linker atom count N_{link} via a learned size embedding table $\mathbf{E} \in \mathbb{R}^{21\times 64}$ projected into the node feature space. **GeoLinker-Both** composes anchor cross-attention and size conditioning within a single forward pass, providing the most constrained and controlled design regime. Variant-specific conditioning equations are detailed in Supplementary Note 2.

5.3 SE(3)-Equivariance Verification

Each EGNN layer constructs edge messages from scalar node features and squared pairwise distances:

$$\boldsymbol{m}_{ij} = \phi_e\left(\boldsymbol{h}_i^{(\ell)}, \boldsymbol{h}_j^{(\ell)}, \left\|\boldsymbol{x}_i^{(\ell)} - \boldsymbol{x}_j^{(\ell)}\right\|^2\right). \tag{2}$$

Because squared distances are invariant to global translation and rotation, the scalar edge messages are also invariant. Coordinates are updated using weighted relative displacements:

$$\boldsymbol{x}_i^{(\ell+1)} = \boldsymbol{x}_i^{(\ell)} + \frac{1}{|\mathcal{N}(i)|}\sum_{j\in\mathcal{N}(i)}\left(\boldsymbol{x}_i^{(\ell)} - \boldsymbol{x}_j^{(\ell)}\right)\tanh(\phi_x(\boldsymbol{m}_{ij}))\,. \tag{3}$$

Under a rotation $\boldsymbol{R} \in \mathrm{SO}(3)$, each relative displacement transforms as $\boldsymbol{x}_i - \boldsymbol{x}_j \mapsto \boldsymbol{R}(\boldsymbol{x}_i - \boldsymbol{x}_j)$, while its scalar weight is unchanged. Consequently,

$$\boldsymbol{x}_i^{(\ell+1)} \mapsto \boldsymbol{R}\boldsymbol{x}_i^{(\ell+1)}, \tag{4}$$

which establishes rotational equivariance. Translation equivariance follows because the messages depend on relative distances and the coordinate update uses only relative displacements. Atom-type predictions remain SE(3)-invariant because they are computed from invariant scalar features.

5.4 Equivariant Diffusion Process

GeoLinker employs a DDPM [9] formulated as a conditional inpainting process. Fragment coordinates are held fixed. The forward process adds noise to the linker positions over $T = 1000$ timesteps under a cosine noise schedule [32]:

$$q(\boldsymbol{x}_t \mid \boldsymbol{x}_{t-1}) = \mathcal{N}\Big(\boldsymbol{x}_t;\ \sqrt{1-\beta_t}\,\boldsymbol{x}_{t-1},\ \beta_t\,\mathbf{I}\Big) \tag{5}$$

The network is parameterised as an $\boldsymbol{x}_0$-predictor, directly outputting clean coordinates $\hat{\boldsymbol{x}}_0 = f_\theta(\boldsymbol{x}_t, t)$. Fragment consistency is enforced at every denoising step by replacing generated fragment positions with their clean ground-truth coordinates (geometric inpainting). Generation uses 500 DDPM reverse steps with centroid-interpolated initialisation. Sampling details and post-processing pipeline are described in Supplementary Note 3.

5.5 Training Objective

The combined training loss is formulated as a weighted sum of denoising, chemical identity, and geometric auxiliary objectives:

$$\begin{aligned}\mathcal{L} = \lambda_{\text{denoise}}\mathcal{L}_{\text{denoise}} + \lambda_{\text{type}}\mathcal{L}_{\text{type}} + \lambda_{\text{endpt}}\mathcal{L}_{\text{endpt}} \\ + \lambda_{\text{clash}}\mathcal{L}_{\text{clash}} + \lambda_{\text{repel}}\mathcal{L}_{\text{repel}}.\end{aligned} \tag{6}$$

The coordinate reconstruction loss $\mathcal{L}_{\text{denoise}}$ is the mean squared error between predicted and ground-truth linker atom coordinates, computed exclusively over linker atoms:

$$\mathcal{L}_{\text{denoise}} = \mathbb{E}_{t,\boldsymbol{x}_0}\left[\frac{1}{N_{\text{link}}}\sum_{i:\,m_i=1}\left\|\boldsymbol{x}_0^{(i)} - f_\theta(\boldsymbol{x}_t, t)^{(i)}\right\|^2\right]. \tag{7}$$

The atom type loss $\mathcal{L}_{\text{type}}$ is a standard cross-entropy over all generated atoms with inverse-frequency class weights to account for rare heteroatoms. The weights for the unified objective were selected based on validation performance: $\lambda_{\text{denoise}} = 1.0$, $\lambda_{\text{type}} = 0.5$, $\lambda_{\text{endpt}} = 0.4$, $\lambda_{\text{clash}} = 0.2$, and $\lambda_{\text{repel}} = 0.08$.

5.5.1 Auxiliary Geometric Losses

While $\mathcal{L}_{\text{denoise}}$ drives global geometric reconstruction, it provides no explicit guarantee of topological connectivity between the linker and fragments. The **Endpoint Distance Loss** explicitly drives the generated linker to physically connect with the designated fragment attachment points. It applies a Smooth-L1 objective penalizing the distance from each fragment anchor atom ($i \in \mathcal{A}$) to its nearest generated linker atom ($j \in \text{link}$), targeting an ideal single-bond distance of 1.5 Å:

$$\mathcal{L}_{\text{endpt}} = \frac{1}{|\mathcal{A}|}\sum_{i\in\mathcal{A}}\text{SmoothL1}\Big(\min_{j\in\text{link}}\|\boldsymbol{x}_i - \boldsymbol{x}_j\|_2,\ 1.5\Big) \tag{8}$$

Smooth-L1 is used in place of MSE to bound gradient magnitude during the high-noise early training phases where linker atoms may be several angstroms from fragments. To prevent the generation of chemically unviable bonds to stable, non-anchor fragment atoms, we introduce a **Non-Anchor Repulsion Loss**. This penalises generated linker atoms that fall within 1.8 Å of any fragment atom not designated as an anchor, excluding the immediate geometric neighbors of anchors to allow for natural bond angles. Furthermore, a **Strict Steric Clash Loss** applies a steep quadratic penalty to any inter-atomic distance falling below 1.2 Å, ensuring that the generated coordinates do not collapse into unphysical steric overlaps. These auxiliary geometric constraints are active during the unified training run to explicitly guide the model toward topologically valid configurations without relying solely on the coordinate reconstruction loss.

5.6 Training Setup

Rather than training separate networks for each constraint regime, GeoLinker utilizes a unified training paradigm. A single model was trained on the ZINC fragment-linker dataset [11, 38] (438,610 molecules; maximum 50 heavy atoms, maximum linker size 20 atoms; 400-molecule held-out validation set) for 150 epochs. Constraint modalities (anchor atoms and linker sizes) were dynamically masked during training using a Bernoulli distribution, allowing the single network to generalize across unconstrained, size-conditioned, and anchor-guided generation tasks. Optimisation used AdamW [33] with a learning rate of 2×10^{-4}, weight decay 10^{-5}, and cosine annealing. To optimize GPU memory and prevent fragmentation across highly variable graph sizes, dynamic atom-budget batching was employed with a maximum limit of 650 atoms per batch. Gradient clipping was applied at $\ell_2 = 1.0$. All training was conducted on a single NVIDIA GPU.

5.7 Datasets and Evaluation

GeoLinker was evaluated on three benchmarks of increasing structural diversity. The *in-distribution* test set consists of 400 ZINC hold-out molecules. The *CASF-2016* [39, 48] provides 285 high-quality protein-ligand co-crystal structures for out-of-distribution evaluation on bioactive conformations. The *GEOM-drugs* dataset[30, 40] tests generalisation to structurally diverse conformers of drug-like molecules. For each input fragment pair, 250 candidate linkers were generated, matching the DiffLinker evaluation protocol [11]. All evaluation protocols, preprocessing procedures, and baseline settings follow the published DiffLinker benchmark wherever applicable to ensure fair comparison.

5.8 Metrics

The generated linkers have been evaluated using six metrics. The Quantitative Estimate of Drug-likeness (QED) [34], synthetic accessibility (SA score) [20], and mean ring count per produced linker, calculated on linker-only canonical SMILES, are used to measure drug-likeness. The generated structures' significance to medicinal chemistry can be seen by these three properties considered together. In accordance with the DiffLinker evaluation protocol [11], novelty (fraction of valid molecules whose canonical SMILES is missing from the ZINC training set), validity (fraction of generated molecules passing

RDKit sanitisation), and uniqueness (fraction of valid molecules with distinct canonical SMILES, averaged per input fragment pair) has also been reported. Supplementary Note 4 contains complete metric definitions and the precise evaluation process.

Abbreviations. FBDD, fragment-based drug discovery; SBDD, structure-based drug design; SE(3), three-dimensional special Euclidean group; EGNN, Equivariant Graph Neural Network; DDPM, denoising diffusion probabilistic model; QED, quantitative estimate of drug-likeness; SA, synthetic accessibility; PROTAC, proteolysis-targeting chimera; IMPDH, inosine-5′-monophosphate dehydrogenase.

Declarations

Supplementary information. Additional file 1 (.pdf): Supplementary Information for GeoLinker, containing detailed model architecture, multi-constraint conditioning, sampling and post-processing procedures, the evaluation protocol, auxiliary-loss justification, dataset preparation, Supplementary Figures S1–S2, and Supplementary Tables S1–S3.

Availability of data and materials. All experiments were conducted in Python (v. 3.10.19). Core deep learning was implemented using PyTorch (v. 2.5.1, CUDA 12.1) [45]. Molecular processing relied on RDKit (v. 2025.09.6) [31], including BRICS-based fragment–linker decomposition and post-processing of generated structures. Bond topology was inferred from the generated 3D atomic point clouds using a greedy valence-constrained perception algorithm, scoring candidate bonds against atom-pair-specific covalent radii. To resolve minor coordinate strains, generated linker atoms were physically relaxed using the Merck Molecular Force Field (MMFF94) implemented in RDKit, with input fragment coordinates frozen to preserve the original binding pose. Protein structure handling used BioPython (v. 1.86) [44].

Additional packages included NumPy (v. 2.0.1), SciPy (v. 1.15.3), pandas (v. 2.3.3), NetworkX (v. 3.4.2), matplotlib (v. 3.10.8), seaborn (v. 0.13.2), tqdm (v. 4.67.3), and Weights & Biases (WandB; v. 0.25.0) for experiment tracking. Baseline comparisons used publicly available implementations of DiffLinker (https://github.com/igashov/DiffLinker) [11] and 3DLinker (https://github.com/YinanHuang/3DLinker).

The processed fragment–linker datasets used in this work were derived from the publicly available DiffLinker datasets [11]. The original datasets include ZINC [38], CASF-2016 [39], and GEOM-drugs [40]. The original ZINC compound library is available through the ZINC database [17], while the CASF-2016 evaluation data were derived from the PDBbind database [48]. The original GEOM dataset is available at https://github.com/learningmatter-mit/geom [30].

The Protein Data Bank provides the crystal structures of the linked inhibitor (PDB: `5OU3`) and experimentally identified fragment hits (PDB: `5OU2`) used in the IMPDH case study, originally reported by Trapero *et al.* [41]. The Hsp90 case study used the fragment hits (PDB: `3HZ1`) and reference inhibitor AT13387 (PDB: `3HZ5`), originally reported by Woodhead *et al.* [42].

The complete source code for GeoLinker, including all four variant implementations, training scripts, preprocessing procedures, and post-processing pipeline, is publicly

available at https://github.com/srajalcodes/GeoLinker. The datasets, model configurations, trained model weights, and associated reproducibility materials used in this study are archived at Zenodo: https://zenodo.org/records/21806622.

The authors declare that they have no competing interests..

Authors' contributions. S.T. conceived the GeoLinker framework, designed and implemented all model variants, conducted all experiments, and drafted the manuscript. D.S. supervised the research, provided scientific guidance, and critically revised the manuscript. R.K. provided chemical domain expertise and contributed to interpretation of drug-likeness results. All authors reviewed and approved the final manuscript.

Acknowledgements. The authors thank the Department of Computer Science and Engineering and the Department of Chemical Engineering at Shiv Nadar Institution of Eminence Deemed to be University, Delhi-NCR, for supporting the research and providing resources.

References

[1] Polishchuk P G, Madzhidov T I, Varnek A (2013) Estimation of the Size of Drug-Like Chemical Space Based on GDB-17 Data. *J. Comput.-Aided Mol. Des.* 27(8):675–679.

[2] Erlanson D A, Fesik S W, Hubbard R E, Jahnke W, Jhoti H (2016) Twenty Years On: The Impact of Fragments on Drug Discovery. *Nat. Rev. Drug Discovery* 15(9):605–619.

[3] Hajduk P J, Greer J (2007) A Decade of Fragment-Based Drug Design: Strategic Advances and Lessons Learned. *Nat. Rev. Drug Discovery* 6(3):211–219.

[4] Congreve M, Carr R, Murray C, Jhoti H (2003) A 'Rule of Three' for Fragment-Based Lead Discovery?. *Drug Discovery Today* 8(19):876–877.

[5] Murray C W, Rees D C (2009) The Rise of Fragment-Based Drug Discovery. *Nat. Chem.* 1(3):187–192.

[6] Yang Y, Zheng S, Su S, Zhao C, Xu J, Chen H (2020) SyntaLinker: Automatic Fragment Linking with Deep Conditional Transformer Neural Networks. *Chem. Sci.* 11(31):8312–8322.

[7] Meyers J, Fabian B, Brown N (2021) De Novo Molecular Design and Generative Models. *Drug Discovery Today* 26(11):2707–2715.

[8] Imrie F, Bradley A R, van der Schaar M, Deane C M (2020) Deep Generative Models for 3D Linker Design. *J. Chem. Inf. Model.* 60(4):1983–1995.

[9] Ho J, Jain A, Abbeel P (2020) Denoising Diffusion Probabilistic Models. *Adv. Neural Inf. Process. Syst.* 33:6840–6851.

[10] Hoogeboom E, Satorras V G, Vignac C, Welling M (2022) Equivariant Diffusion for Molecule Generation in 3D. *Proc. Int. Conf. Mach. Learn.* 162:8867–8887.

[11] Igashov I, Stärk H, Vignac C, Schneuing A, Satorras V G, Frossard P, Welling M, Bronstein M, Correia B (2024) Equivariant 3D-Conditional Diffusion Model for Molecular Linker Design. *Nat. Mach. Intell.* 6(4):417–427.

[12] Schneuing A, Harris C, Du Y, Didi K, Jamasb A, Igashov I, Du W, Blundell T L, Liò P, Gomes C, et al (2024) Structure-Based Drug Design with Equivariant Diffusion Models. *Nat. Comput. Sci.* 4(3):224–235.

[13] Crews C M (2017) Targeted Protein Degradation: From Chemical Biology to Drug Discovery. *Cell Chem. Biol.* 24(9):1197–1207.

[14] Edmondson S D, Yang B, Fallan C (2019) PROTACs: A New Approach to Targeted Protein Degradation. *Bioorg. Med. Chem. Lett.* 29(13):1555–1564.

[15] Guan J, Qian W, Liu Q, Ma W-Y, Ma J, Peng J (2023) LinkerNet: Fragment Poses and Linker Co-Design with 3D Equivariant Diffusion. *Adv. Neural Inf. Process. Syst.* 36:77503–77519.

[16] Luo S, Guan J, Ma J, Peng J (2023) Domain-Adapted Diffusion Models for PROTAC Linker Design. *Adv. Neural Inf. Process. Syst.* 36.

[17] Irwin J J, Shoichet B K (2005) ZINC—A Free Database of Commercially Available Compounds for Virtual Screening. *J. Chem. Inf. Model.* 45(1):177–182.

[18] Kingma D P, Ba J (2015) Adam: A Method for Stochastic Optimization. *Proc. Int. Conf. Learn. Represent.*.

[19] Satorras V G, Hoogeboom E, Welling M (2021) E(n) Equivariant Graph Neural Networks. *Proc. Int. Conf. Mach. Learn.* 139:9323–9332.

[20] Ertl P, Schuffenhauer A (2009) Estimation of Synthetic Accessibility Score of Drug-Like Molecules Based on Molecular Complexity and Fragment Contributions. *J. Cheminf.* 1:8.

[21] Levy D, Rector-Brooks J (2023) Molecular Fragment-Based Diffusion Model for Drug Discovery. *Proc. ICLR Workshop Mach. Learn. Drug Discovery.*

[22] Ghorbani M, Gendelev L, Beroza P, Keiser M J (2024) Autoregressive Fragment-Based Diffusion for Pocket-Aware Ligand Design *arXiv* preprint arXiv:2401.05370.

[23] Polykovskiy D, Zhebrak A, Sanchez-Lengeling B, Golovanov S, Tatanov O, Belyaev S, Kurbanov R, Artamonov A, Aladinskiy V, Veselov M, et al (2020) Molecular Sets (MOSES): A Benchmarking Platform for Molecular Generation Models. *Front. Pharmacol.* 11:565644.

[24] Brown N, Fiscato M, Segler M H S, Vaucher A C (2019) GuacaMol: Benchmarking Models for De Novo Molecular Design. *J. Chem. Inf. Model.* 59(3):1096–1108.

[25] Sohl-Dickstein J, Weiss E, Maheswaranathan N, Ganguli S (2015) Deep Unsupervised Learning Using Nonequilibrium Thermodynamics. *Proc. Int. Conf. Mach. Learn.* 37:2256–2265.

[26] Lin Z, Akin H, Rao R, Hie B, Zhu Z, Lu W, Smetanin N, Verkuil R, Kabeli O, Shmueli Y, et al (2023) Evolutionary-Scale Prediction of Atomic-Level Protein Structure with a Language Model. *Science* 379(6637):1123–1130.

[27] Fuchs F B, Worrall D E, Fischer V, Welling M (2020) SE(3)-Transformers: 3D Roto-Translation Equivariant Attention Networks. *Adv. Neural Inf. Process. Syst.* 33:1970–1981.

[28] Geiger M, Smidt T (2022) e3nn: A Modular Library for Equivariant Neural Networks *arXiv* preprint arXiv:2207.09453.

[29] Irwin J J, Sterling T, Mysinger M M, Bolstad E S, Coleman R G (2012) ZINC: A Free Tool to Discover Chemistry for Biology. *J. Chem. Inf. Model.* 52(7):1757–1768.

[30] Axelrod S, Gomez-Bombarelli R (2022) GEOM: Energy-Annotated Molecular Conformations for Property Prediction and Force Field Development. *Sci. Data* 9:185.

[31] Landrum G (2025) RDKit: open-source cheminformatics. https://www.rdkit.org. Accessed 6 Aug 2026.

[32] Nichol A Q, Dhariwal P (2021) Improved Denoising Diffusion Probabilistic Models. *Proc. Int. Conf. Mach. Learn.* 139:8162–8171.

[33] Loshchilov I, Hutter F (2019) Decoupled Weight Decay Regularization. *Proc. Int. Conf. Learn. Represent.*.

[34] Bickerton G R, Paolini G V, Besnard J, Muresan S, Hopkins A L (2012) Quantifying the Chemical Beauty of Drugs. *Nat. Chem.* 4(2):90–98.

[35] Wishart D S, Feunang Y D, Guo A C, Lo E J, Marcu A, Grant J R, Sajed T, Johnson D, Li C, Sayeeda Z, et al (2018) DrugBank 5.0: A Major Update to the DrugBank Database for 2018. *Nucleic Acids Res.* 46(D1):D1074–D1082.

[36] Lipinski C A, Lombardo F, Dominy B W, Feeney P J (1997) Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Delivery Rev.* 23(1–3):3–25.

[37] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez A N, Kaiser L, Polosukhin I (2017) Attention Is All You Need. *Adv. Neural Inf. Process. Syst.* 30.

[38] Igashov I, Stärk H, Vignac C, Schneuing A, Satorras V G, Frossard P, Welling M, Bronstein M, Correia B (2022) DiffLinker ZINC Dataset (Processed). *Zenodo* . https://doi.org/10.5281/zenodo.7121271.

[39] Igashov I, Stärk H, Vignac C, Schneuing A, Satorras V G, Frossard P, Welling M, Bronstein M, Correia B (2022) DiffLinker CASF Dataset (Processed). *Zenodo* . https://doi.org/10.5281/zenodo.7121264.

[40] Igashov I, Stärk H, Vignac C, Schneuing A, Satorras V G, Frossard P, Welling M, Bronstein M, Correia B (2022) DiffLinker GEOM Dataset (Processed). *Zenodo* . https://doi.org/10.5281/zenodo.7121278.

[41] Trapero A, Muhammad A, Fletcher I W, Bhatt J S, Bhatt S, O'Neill P M, Berry N G, Coyne A G, Abell C (2018) Fragment-Based Approach to Targeting Inosine-5′-Monophosphate Dehydrogenase (IMPDH) from *Mycobacterium tuberculosis*. *J. Med. Chem.* 61(7):2806–2822.

[42] Woodhead A J, Angove H, Carr M G, Chessari G, Congreve M, Coyle J E, Cosme J, Graham B, Day P J, Downham R, et al (2010) Discovery of (2,4-Dihydroxy-5-isopropylphenyl)-[5-(4-methylpiperazin-1-ylmethyl)-1,3-dihydroisoindol-2-yl]methanone (AT13387), a Novel Inhibitor of the Molecular Chaperone Hsp90 by Fragment Based Drug Design. *J. Med. Chem.* 53(15):5956–5969.

[43] O'Boyle N M, Banck M, James C A, Morley C, Vandermeersch T, Hutchison G R (2011) Open Babel: An Open Chemical Toolbox. *J. Cheminf.* 3:33.

[44] Cock P J A, Antao T, Chang J T, Chapman B A, Cox C J, Dalke A, Friedberg I, Hamelryck T, Kauff F, Wilczynski B, et al (2009) Biopython: Freely Available Python Tools for Computational Molecular Biology and Bioinformatics. *Bioinformatics* 25(11):1422–1423.

[45] Paszke A, Gross S, Massa F, Lerer A, Bradbury J, Chanan G, Killeen T, Lin Z, Gimelshein N, Antiga L, et al (2019) PyTorch: An Imperative Style, High-Performance Deep Learning Library. *Adv. Neural Inf. Process. Syst.* 32.

[46] Jumper J, Evans R, Pritzel A, Green T, Figurnov M, Ronneberger O, Tunyasuvunakool K, Bates R, Žídek A, Potapenko A, et al (2021) Highly Accurate Protein Structure Prediction with AlphaFold. *Nature* 596(7873):583–589.

[47] Askr H, Elgeldawi E, Aboul Ella H, Elshaier Y A M M, Gomaa M M, Hassanien A E (2023) Deep Learning in Drug Discovery: An Integrative Review and Future Challenges. *Artif. Intell. Rev.* 56(7):5975–6037.

[48] Su M, Yang Q, Du Y, Feng G, Liu Z, Li Y, Wang R (2019) Comparative Assessment of Scoring Functions: The CASF-2016 Update. *J. Chem. Inf. Model.* 59(2):895–913.

Supplementary Files

This is a list of supplementary files associated with this preprint. Click to download.

- GeoLinkerSupp.pdf