

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

Srajal Tiwari¹, Dolly Sharma^{1*}, Ranjit Kumar²

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

²Department 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;

This Supplementary Information includes:

- Supplementary Note 1—Model Architecture and Coordinate Updates;
- Supplementary Note 2—Multi-Constraint Variant Conditioning;
- Supplementary Note 3—Sampling Procedure and Post-processing Pipeline;
- Supplementary Note 4—Evaluation Protocol;
- Supplementary Note 5—Auxiliary Loss Design Justification;
- Supplementary Note 6—Dataset Preparation;
- Supplementary Figs. S1–S2; and
- Supplementary Tables S1–S3.

Supplementary Note 1: Model Architecture and Coordinate Updates

Atomic Representation and Graph Construction

Each heavy atom i is represented by its three-dimensional Cartesian coordinates $\mathbf{x}_i \in \mathbb{R}^3$ and an eight-dimensional one-hot vector $\mathbf{e}_i \in \{0, 1\}^8$ encoding chemical identity over the alphabet {C, O, N, F, S, Cl, Br, I}, covering all heavy-atom types in the ZINC fragment-linker dataset. A binary membership indicator $m_i \in \{0, 1\}$

distinguishes fragment atoms ($m_i = 0$) from linker atoms ($m_i = 1$). Atom features are projected to a $d_h = 128$ -dimensional hidden space by a learned linear embedding layer. The diffusion timestep is represented by a sinusoidal positional embedding [2] and fused into the node scalar features before the first EGNN layer.

At each denoising step, a fully connected geometric graph is constructed over all $N = N_{\text{frag}} + N_{\text{link}}$ atoms without a distance cutoff, ensuring that linker atoms receive information from both fragments simultaneously. Pairwise squared Euclidean distances, $\|\mathbf{x}_i - \mathbf{x}_j\|^2$, are computed directly and supplied as invariant edge features to the message MLP in each EGNN layer. No radial-basis-function expansion or neighbor-pooled distance embedding is applied.

EGNN Backbone and Equivariant Coordinate Update

The backbone comprises $L = 6$ stacked $E(n)$ Equivariant Graph Neural Network (EGNN) layers [6]. Each layer computes message passing over node scalar features $\mathbf{h}_i^{(\ell)} \in \mathbb{R}^{128}$ and updates coordinate features $\mathbf{x}_i^{(\ell)} \in \mathbb{R}^3$. Edge messages are formulated using invariant pairwise distances:

$$\mathbf{m}_{ij} = \phi_e\left(\mathbf{h}_i^{(\ell)}, \mathbf{h}_j^{(\ell)}, \left\|\mathbf{x}_i^{(\ell)} - \mathbf{x}_j^{(\ell)}\right\|^2\right). \quad (\text{S1})$$

Scalar features are updated using standard multi-head attention [3] ($H = 8$ heads), followed by a standard multilayer perceptron ($128 \rightarrow 128$, GELU activation) and layer normalization. Equivariant coordinate features are updated by a relative spatial displacement sum over same-molecule neighbors:

$$\begin{aligned} \mathbf{x}_i^{(\ell+1)} = & \mathbf{x}_i^{(\ell)} + \frac{1}{|\mathcal{N}(i)|} \sum_{j \in \mathcal{N}(i)} \left(\mathbf{x}_i^{(\ell)} - \mathbf{x}_j^{(\ell)}\right) \\ & \cdot \tanh(\phi_x(\mathbf{m}_{ij})), \end{aligned} \quad (\text{S2})$$

where $\tanh$ smoothly bounds coordinate step sizes, and normalization by the active same-molecule neighbor count ensures consistency across variable batch graphs. After the final layer, the updated coordinates provide the clean-coordinate prediction $\hat{\mathbf{x}}_0$, while an MLP maps invariant scalar features to eight atom-type logits. An auxiliary anchor head predicts binary attachment probabilities for unconstrained inference.

The SE(3)-equivariance verification is provided in the Methods section of the main manuscript.

Supplementary Note 2: Multi-Constraint Variant Conditioning

All four GeoLinker inference modes share a single EGNN backbone trained under dynamic constraint masking. Variant-specific conditioning signals are fused into the node representations before or after the backbone.

GeoLinker-Base generates linkers conditioned exclusively on fragment geometry. An auxiliary anchor-prediction head trained alongside the backbone outputs a binary attachment probability for each linker atom through $\sigma(\text{MLP}(\mathbf{h}_i^{(L)}))$ ($128 \rightarrow 32 \rightarrow 1$), providing implicit geometric guidance without requiring explicit anchor information at inference time.

GeoLinker-Anchor incorporates experimentally identified attachment atoms using `nn.MultiheadAttention` ($d = 128$, $H = 8$ heads), in which linker-atom representations attend to anchor-atom representations through query-key-value attention. The attended output is added to the linker scalar features through a residual connection. Anchor-atom positions are also held fixed throughout reverse diffusion as a hard constraint by replacing generated anchor coordinates with their ground-truth values at every denoising step.

GeoLinker-Sized conditions on the target linker atom count $N_{\text{link}} \in \{1, \dots, 20\}$ through a learnable embedding table $\mathbf{E} \in \mathbb{R}^{21 \times 64}$. The embedding is projected to the node-feature space and broadcast to all atoms before the first EGNN layer. The learnable table permits the model to acquire size-specific geometric priors directly from the training-data distribution.

GeoLinker-Both composes anchor cross-attention and size-embedding conditioning within a single forward pass. The size embedding is fused into the node features before the backbone, and anchor cross-attention is applied after the backbone output. These two conditioning signals are complementary: the size embedding constrains global linker topology, whereas anchor cross-attention constrains local attachment geometry, jointly reducing the effective search space to linkers of the correct length that connect to the specified anchor atoms.

Table S3 Complete hyperparameter specification.

Architecture	
Atom alphabet	C, N, O, F, S, Cl, Br, I
Hidden dimension d_h	128
EGNN layers	6
Attention heads	8
Dropout	0.1
Edge features	Squared pairwise distances
Size embedding	$\mathbf{E} \in \mathbb{R}^{21 \times 64}$
Parameters	2,386,184
Diffusion and atom-type weights	
Timesteps/reverse steps	1,000/500 (DDPM)
Schedule/clipping	Cosine, $s = 0.008$ / $[-30, 30]$ Å
C/O/N; F/S; Cl/Br/I	1.0/2.6/3.0; 50; 100
Training	
Optimizer/weight decay	AdamW/ 10^{-5}
Learning rate/epochs	2×10^{-4} /150
Gradient clipping	1.0
LR schedule	Cosine annealing
Batch size/GPU	Variable (max 650 atoms/batch)/one NVIDIA GPU
Loss weights and software	
$\lambda_{\text{denoise}}/\lambda_{\text{type}}$	1.0/0.5
$\lambda_{\text{endpt}}/\lambda_{\text{clash}}/\lambda_{\text{repel}}$	0.4/0.2/0.08
Framework/chemistry	PyTorch 2.5.1/RDKit

Supplementary Note 3: Sampling Procedure and Post-processing Pipeline

Sampling Initialization and Reverse Diffusion

GeoLinker uses the cosine variance schedule [4] with offset $s = 0.008$, satisfying $\bar{\alpha}_0 \approx 1$ and $\bar{\alpha}_{1000} \approx 0$. The forward process, reverse transition, posterior variance, and posterior mean follow the standard DDPM formulation [2]. Fragment coordinates are held fixed and are never corrupted; at every denoising step, generated fragment positions are replaced by their ground-truth values (a geometric inpainting constraint). Generation uses 500 DDPM reverse steps.

For the Base and Sized variants, linker-atom positions are initialized by centroid interpolation along the inter-fragment axis connecting fragment centroids $\mathbf{c}_1$ and $\mathbf{c}_2$:

$$\begin{aligned} \mathbf{x}_T^{(k)} &= (1 - t_k)\mathbf{c}_1 + t_k\mathbf{c}_2 + \boldsymbol{\varepsilon}_k, \\ t_k &= \frac{k}{N_{\text{link}} + 1}, \quad \boldsymbol{\varepsilon}_k \sim \mathcal{N}(\mathbf{0}, 0.1^2 \mathbf{I}), \end{aligned} \tag{S3}$$

for $k = 1, \dots, N_{\text{link}}$. This physically motivated initialization places linker atoms along the geometric axis connecting the two fragments, substantially reducing the number of denoising steps required for convergence relative to isotropic Gaussian initialization.

For the Anchor and Both variants, non-anchor linker atoms are initialized by forward-diffusing ground-truth positions to timestep $T_{\text{init}} = 500$.

During the final third of the reverse steps ($t < T_{\text{init}}/3$), a fragment-attraction heuristic (strength 0.05) nudges each linker atom toward its nearest fragment centroid when the minimum linker-fragment distance exceeds 3.0 Å. When protein-pocket coordinates are available (CASF-2016 evaluation), a post-hoc steric-repulsion heuristic displaces linker atoms within 1.5 Å of any pocket heavy atom by $\delta = 0.1$ Å along the repulsion vector. Both heuristics are applied exclusively at inference time, use no learned parameters, and were not applied to any baseline method.

Post-processing Pipeline

The raw typed atomic point cloud is converted to a molecular graph using a seven-stage pipeline. Stage-level statistics computed on 100,000 molecules (250 samples $\times$ 400 ZINC test inputs, GeoLinker-Base) are reported in Supplementary Table S2.

Stage 1—Connectivity perception. A custom greedy valence-constrained perception algorithm paired with RDKit’s `DetermineConnectivity` (`covFactor=1.3`) assigns single bonds between atom pairs using tailored covalent radii. RDKit’s `DetermineBondOrders` subsequently assigns double, triple, and aromatic bonds. Isolated atoms arise in 2.7% of generated molecules.

Stage 2—Chemical sanitization. `Chem.SanitizeMol` verifies valency, perceives aromaticity using the Hückel model, and assigns ring membership. This is the primary validity filter, at which 3.25% of molecules are rejected and excluded from the reported metrics.

Stage 3—Standardization. `MolStandardize.Cleanup` normalizes functional-group representations and resolves tautomeric ambiguities. In our experiments, 8.1% of molecules undergo nontrivial correction.

Stage 4—Charge neutralization. `rdMolStandardize.Uncharger` removes spurious formal charges arising from bond-order perception ambiguities, affecting 12.4% of nitrogen-containing molecules.

Stage 5—Fragment retention. `Chem.GetMolFragments` decomposes the graph, after which the largest connected component is retained. Disconnected noise atoms arise in 1.8% of molecules.

Stage 6—Radical resolution. `atom.SetNumRadicalElectrons(0)` resolves attachment-point radicals at the linker-fragment boundary (4.3% of molecules).

Stage 7—Steric-clash resolution. Atom pairs with $d_{ij} < 1.2$ Å are symmetrically displaced:

$$\begin{aligned} \mathbf{x}_i &\leftarrow \mathbf{x}_i + \frac{\mathbf{x}_i - \mathbf{x}_j}{d_{ij}} \frac{1.2 - d_{ij}}{2}, \\ \mathbf{x}_j &\leftarrow \mathbf{x}_j - \frac{\mathbf{x}_i - \mathbf{x}_j}{d_{ij}} \frac{1.2 - d_{ij}}{2}, \end{aligned} \tag{S4}$$

iterated for 20 steps. After Stage 7, all pairwise distances satisfy $d_{ij} \geq 1.2$ Å in 99.8% of cases; the remaining 0.2% retain near-bond distances (1.0–1.2 Å) that are resolved using RDKit’s valency-clean property cache.

Table S2 Post-processing stage statistics for 100,000 GeoLinker-Base samples. Only Stage 2 rejects molecules; percentages are not mutually exclusive.

Stage	Operation	Affected (%)	Action
1	Bond inference	2.7	Flagged
2	RDKit sanitization	3.25	Rejected
3	Standardization	8.1	Corrected
4	Charge neutralization	12.4	Corrected
5	Fragment retention	1.8	Discarded
6	Radical resolution	4.3	Corrected
7	Steric-clash correction	6.7	Adjusted

Supplementary Note 4: Evaluation Protocol

For each input fragment pair, GeoLinker generates 250 candidate linkers, matching the DiffLinker evaluation protocol [1]. All 250 samples pass through the seven-stage post-processing pipeline before evaluation. Generated molecules are converted to canonical SMILES using `Chem.MolToSmiles(mol, canonical=True)`; tautomers are canonicalized using `MolStandardize.TautomerEnumerator` before deduplication. Chirality is not considered, consistent with the DiffLinker protocol and reflecting the fact that GeoLinker does not currently model stereocenters.

Validity is the fraction of all generated molecules that pass `Chem.SanitizeMol` without a valency exception. **Uniqueness** is computed for each input fragment pair as the fraction of valid molecules with distinct canonical SMILES and is then averaged over all test inputs; this per-input definition prevents cross-input artifacts from inflating the score. Near-100% uniqueness reflects genuine stochastic diversity among the 250 samples generated for each input by the diffusion process. **Novelty** is the fraction of valid molecules whose canonical SMILES is absent from the ZINC training set $\mathcal{D}_{\text{train}}$, precomputed as a hash set for $O(1)$ lookup. High novelty across all benchmarks—78.7%–82.1% on ZINC, 70.6%–72.6% on CASF-2016, and 98.0%–98.8% on GEOM-drugs—indicates that GeoLinker explores structures beyond the training set rather than merely reproducing training molecules. All baseline results in Table 1 of the main manuscript are reproduced directly from the published DiffLinker supplementary information [1], using identical test sets and $N = 250$ samples per input.

Drug-likeness (QED) [10] integrates eight molecular properties into a single desirability score from 0 to 1. Approved drugs typically score between 0.4 and 0.9. **Synthetic accessibility (SA score)** [11] ranges from 1 (trivially synthesizable) to 10 (practically inaccessible) and is estimated from fragment contributions trained on commercial-compound databases. **Ring content** is the mean number of rings per generated molecule computed from linker-only canonical SMILES using `rdMolDescriptors.CalcNumRings`. Ring counts are evaluated on linker-only SMILES to ensure a like-for-like comparison with the published DiffLinker values.

Supplementary Note 5: Auxiliary Loss Design Justification

Motivation

The primary denoising objective $\mathcal{L}_{\text{denoise}}$ drives global geometric reconstruction but provides no explicit topological guarantee. A geometrically plausible linker that is spatially disconnected from its input fragments can incur a low loss yet remain chemically inert. GeoLinker therefore combines endpoint, repulsion, and steric-clash objectives with coordinate and atom-type supervision.

Endpoint Distance Loss

The Endpoint Distance Loss explicitly drives the generated linker to 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 ($j \in \mathcal{V}_{\text{link}}$), targeting an ideal single-bond distance of 1.5 Å:

$$\mathcal{L}_{\text{endpt}} = \frac{1}{|\mathcal{A}|} \sum_{i \in \mathcal{A}} \text{SmoothL1} \left(\min_{j \in \mathcal{V}_{\text{link}}} \|\mathbf{x}_i - \mathbf{x}_j\|_2, 1.5 \right), \quad (\text{S5})$$

where $\text{SmoothL1}(a, b) = \frac{1}{2}(a - b)^2$ if $|a - b| < 1$, and $|a - b| - \frac{1}{2}$ otherwise. Smooth-L1 bounds gradient magnitude during early high-noise training, when predicted linker atoms may be several angstroms from the fragment anchors.

Repulsion and Steric-Clash Losses

The Non-Anchor Repulsion Loss penalizes generated linker atoms that fall within 1.8 Å of fragment atoms not designated as anchors. Immediate geometric neighbors of anchor atoms are excluded to preserve natural bond angles at the attachment interface. The Strict Steric Clash Loss additionally applies a steep quadratic penalty whenever an inter-atomic distance falls below 1.2 Å, preventing coordinate collapse and unphysical steric overlap.

Unified Objective

All geometric constraints are active during unified training. The complete objective combines denoising, atom-type, endpoint, clash, and repulsion terms with weights 1.0, 0.5, 0.4, 0.2, and 0.08, respectively, as summarized in Supplementary Table S1.

Table S1 Weights used in the unified GeoLinker training objective.

Objective	Weight	Role
$\mathcal{L}_{\text{denoise}}$	1.00	Linker coordinate reconstruction
$\mathcal{L}_{\text{type}}$	0.50	Atom identity classification
$\mathcal{L}_{\text{endpt}}$	0.40	Anchor-to-linker endpoint connection
$\mathcal{L}_{\text{clash}}$	0.20	Strict short-range steric exclusion
$\mathcal{L}_{\text{repel}}$	0.08	Repulsion from non-anchor fragment atoms

Supplementary Note 6: Dataset Preparation

ZINC Fragment–Linker Dataset

The ZINC dataset was prepared using the DiffLinker decomposition protocol [1], adopted without modification to ensure a fair comparison. SMILES strings from the ZINC15 database were filtered to molecules with 10–50 heavy atoms and molecular weight ≤ 500 Da. Each molecule was decomposed into exactly two fragments and one linker using the BRICS algorithm implemented in RDKit; molecules that could not be decomposed into exactly two fragments were discarded. Three-dimensional coordinates were generated using the MMFF94 force field (`RDKit.EmbedMolecule`, random seed 42). The resulting 438,610 molecules were partitioned into training (438,210), validation (400), and test (400) sets. Dataset distributions are summarized in Supplementary Fig. S1, and representative GeoLinker outputs are shown in Supplementary Fig. S2.

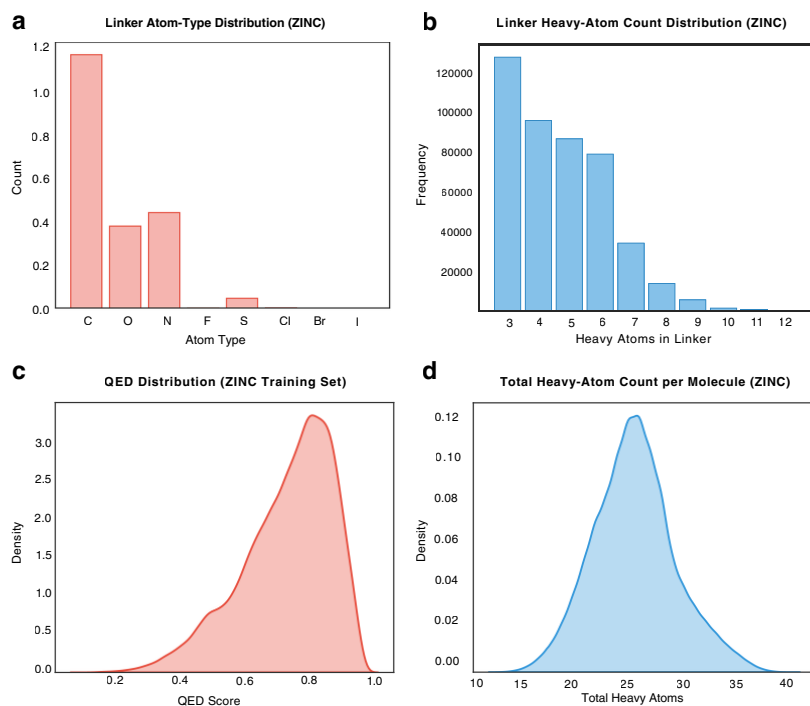


Fig. S1 ZINC fragment-linker dataset statistics. Training set: 438,610 molecules (438,210 training, 400 validation, and 400 test), prepared using the BRICS decomposition protocol and filtered to ≤ 50 heavy atoms and ≤ 20 linker atoms. **(a)** Atom-type frequency of ground-truth linker atoms; the strong carbon bias motivates inverse-frequency class weights in $\mathcal{L}_{\text{type}}$ (counts in millions). **(b)** Linker heavy-atom count distribution; the modal length is 3–4 atoms, with a right-skewed tail to a maximum of 12 atoms. **(c)** QED distribution of the training set, peaking near QED ≈ 0.8 , consistent with the drug-like character of ZINC compounds. **(d)** Total heavy-atom count per molecule, reflecting the ≤ 50 -atom filtering criterion.

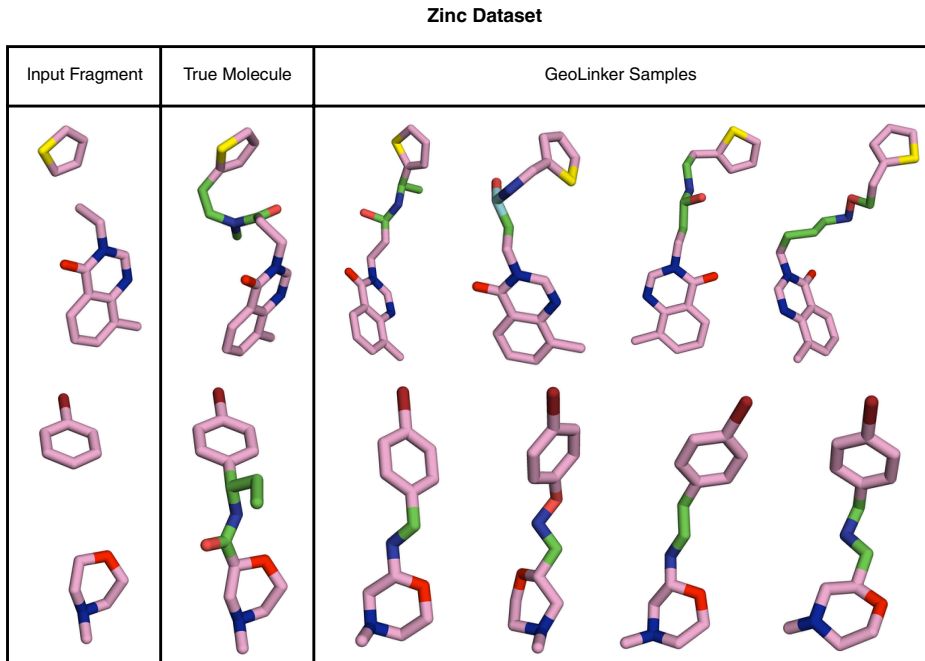


Fig. S2 GeoLinker-generated linkers on the ZINC test set. Each row shows one fragment pair from the ZINC test set. From left to right, the columns show the two input fragments, the ground-truth molecule, and representative valid samples generated by GeoLinker. Valid molecules shown passed the seven-stage post-processing pipeline and RDKit sanitization.

CASF-2016 Benchmark

The CASF-2016 benchmark [7] provides 285 high-quality protein–ligand co-crystal structures from the PDBbind database, each resolved to ≤ 2.5 Å. Ligands were extracted and subjected to the same BRICS decomposition. Protein-pocket atoms (all heavy atoms within 10 Å of any ligand heavy atom) were retained exclusively for the optional post-hoc steric-repulsion step at sampling time; no pocket features were used during training or model inference.

GEOM-Drugs Dataset

The GEOM-Drugs dataset [8] was processed using the DiffLinker protocol [9]. Molecules were filtered to the eight-element atom-type alphabet, and entries with linker phosphorus atoms were excluded. Multi-fragment entries were retained to probe linker generation under structurally diverse and conformationally flexible conditions.

References

- [1] 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.
- [2] Ho J, Jain A, Abbeel P (2020) Denoising Diffusion Probabilistic Models. *Adv. Neural Inf. Process. Syst.* 33:6840–6851.
 - [3] 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.
 - [4] Nichol A Q, Dhariwal P (2021) Improved Denoising Diffusion Probabilistic Models. *Proc. Int. Conf. Mach. Learn.* 139:8162–8171.
 - [5] 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.
 - [6] Satorras V G, Hoogeboom E, Welling M (2021) E(n) Equivariant Graph Neural Networks. *Proc. Int. Conf. Mach. Learn.* 139:9323–9332.
 - [7] 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.
 - [8] Axelrod S, Gomez-Bombarelli R (2022) GEOM: Energy-Annotated Molecular Conformations for Property Prediction and Force Field Development. *Sci. Data* 9:185.
 - [9] 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>.
 - [10] 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.
 - [11] 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.
 - [12] Loshchilov I, Hutter F (2019) Decoupled Weight Decay Regularization. *Proc. Int. Conf. Learn. Represent.*.