HBDRUG3D: A UNIFIED BENCHMARK DATASET OF HETEROBIFUNCTIONAL DRUGS FOR GEOMETRIC LINKER DESIGN

Anonymous authors
Paper under double-blind review

ABSTRACT

AI-driven drug design has demonstrated substantial potential in accelerating molecular discovery and optimization. However, for heterobifunctional drug modalities such as proteolysis-targeting chimeras (PROTACs), peptide-drug conjugates (PDCs), and antibody-drug conjugates (ADCs), the scarcity of publicly available 3D data and standardized evaluation protocols limits the development and comparison of machine learning methods. To address this gap, we introduce **HBDrug3D**, one of the largest publicly available 3D conformational datasets for heterobifunctional drugs, comprising 70,393 molecular conformations across PROTACs, PDCs, and ADCs. Each molecule is annotated at the component level, identifying its linker and modality-specific functional groups, including warheads, payloads, and E3 ligands. Based on HBDrug3D, we establish a standardized benchmark for geometric linker design, a central task in heterobifunctional drug development. The benchmark evaluates generated linkers in terms of validity, novelty, and physicochemical properties relevant to medicinal chemistry. By integrating diverse molecular modalities, component-level annotations, and reproducible evaluation protocols, HBDrug3D provides a common foundation for developing and systematically comparing machine learning approaches to heterobifunctional drug design.

1 INTRODUCTION

Heterobifunctional drugs, including PROTACs, Antibody-Drug Conjugates (ADCs), and Peptide-Drug Conjugates (PDCs), have emerged as a transformative modality in targeted cancer therapy (Bulldan et al., 2025; Matyskiela, 2025; Wang et al., 2025). These molecules leverage a dual-module mechanism to achieve highly specific therapeutic effects: PROTACs recruit E3 ligases to selectively degrade oncogenic proteins (Békés et al., 2022), while ADCs and PDCs deliver cytotoxic payloads directly to tumor cells via antibody or peptide targeting modules (Tsuchikama et al., 2024; Dai et al., 2024)(Figure 1). The linker, which connects the warhead and the payload or E3 ligand, serves as a pivotal element: it stabilizes the molecule's 3D conformation, fine-tunes pharmacokinetics, and ensures precise targeting, while simultaneously influencing its therapeutic efficacy and safety (Luo et al., 2025; Bouacha et al., 2025; Felber et al., 2025). Therefore, designing linkers that satisfy both chemical and structural constraints is a critical yet challenging task, given

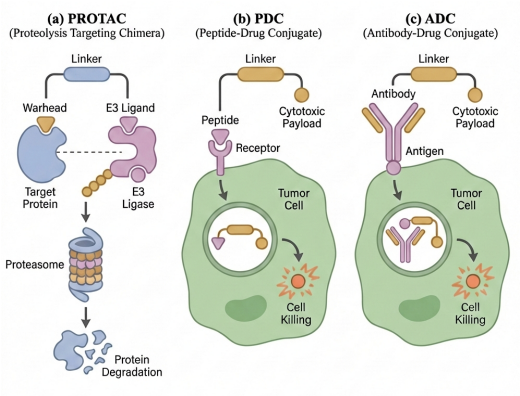


Figure 1: **Schematic illustration of heterobifunctional drugs.** PROTACs recruit E3 ligases to selectively degrade target proteins, while ADCs and PDCs deliver cytotoxic payloads to tumor cells via antibody or peptide targeting modules. In all cases, the linker connects the warhead and the effector module, playing a pivotal role in stabilizing 3D conformation, modulating pharmacokinetics, and ensuring targeted therapeutic activity.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

the complex 3D spatial arrangements and diverse
physicochemical properties required for effective heterobifunctional drug function.

Despite the critical role of linkers, computational modeling of heterobifunctional drug design is limited by the lack of appropriate datasets. Existing resources such as PROTAC-DB (Ge et al., 2025), PDC-DB (Sun et al., 2025), and ADC-DB Shen et al. (2024) are heterogeneous in format and do not include 3D structural information, which restricts their applicability for modeling linkers. General-purpose 3D molecular datasets, such as ZINC (Irwin & Shoichet, 2005) or GEOM (Axelrod & Gómez-Bombarelli, 2022), contain linker-like fragments derived from fragmenting small molecules, but these fragments differ significantly from heterobifunctional drug linkers in molecular weight, atom count, and linker length, introducing substantial domain mismatch and limiting their usefulness for rational linker design. A summary of these datasets and their characteristics is provided in Table 1.

Table 1: **Dataset comparison.** PROTAC, PDC, and ADC stand for proteolysis-targeting chimera, protein degrader conjugate, and antibody–drug conjugate, respectively. All three correspond to heterobifunctional drug modalities. We observe that existing datasets either focus on small molecules or lack 3D structural information for heterobifunctional drugs.

Dataset	Drug Modalities	#Samples	3D Confs.	Avg. Atoms
ZINC (Irwin & Shoichet, 2005)	Small Molecule	439,410	✓	25
GEOM (Axelrod & Gómez-Bombarelli, 2022)	Small Molecule	285,142	✓	32
Pockets (Schneuing et al., 2024)	Small Molecule	186,734	✓	31
PROTAC-DB (Ge et al., 2025)	PROTAC	6,111	✗	67
PDC-DB (Sun et al., 2025)	PDC	2,036	✗	143
ADC-DB Shen et al. (2024)	ADC	6,572	✗	95
HBDrug3D	PROTAC,PDC,ADC	70,393	✓	103

In parallel, generative models for 3D linker design have made notable progress in constructing chemically and geometrically valid linkers. Among them, **DeLinker** (Imrie et al., 2020) and **3DLinker** (Huang et al., 2022) are auto-regressive models that generate linker atoms and bonds sequentially, conditioned on fragment embeddings and coarse 3D spatial information. Building on these ideas, diffusion-based methods, such as **DiffLinker** (Igashov et al., 2024) and **LinkerNet** (Guan et al., 2023), simultaneously generate atom types and coordinates. LinkerNet further enhances flexibility by allowing rotation and translation of the fragments during generation, enabling the linker and fragments to adopt more favorable conformations. Despite these advances, all of these methods rely on 3D fragment conformations, which are not currently available for heterobifunctional drugs in existing datasets. Furthermore, their evaluations have so far been limited to general-purpose small molecule datasets, such as ZINC or GEOM, leaving their performance on heterobifunctional drug linker design unexplored.

To address these limitations, we present HBDrug3D, a comprehensive dataset of heterobifunctional drugs comprising 70,393 3D conformations across PROTACs, PDCs, and ADCs. Each molecule is annotated at the atomic level to indicate functional fragments, including warhead, linker, and payload/E3 ligand. In addition, we propose a unified benchmark for heterobifunctional linker design, which evaluates generated linkers not only in terms of validity and novelty but also across a range of medicinal chemistry–relevant properties, providing a comprehensive framework for assessing the performance of computational models in this domain.The main contributions of this paper are summarized as follows:

- **HBDrug3D dataset:** We construct HBDrug3D, a comprehensive 3D conformational dataset of heterobifunctional drugs, covering PROTACs, PDCs, and ADCs. Each atom is annotated according to its functional fragment (warhead, linker, or payload/E3 ligand), providing a standardized resource for data-driven linker design.
- **Benchmark for linker design:** We establish a unified benchmark for heterobifunctional drug linker design, including both conventional evaluation metrics such as validity and novelty, as well as a series of medicinal chemistry–relevant properties to comprehensively assess model performance.

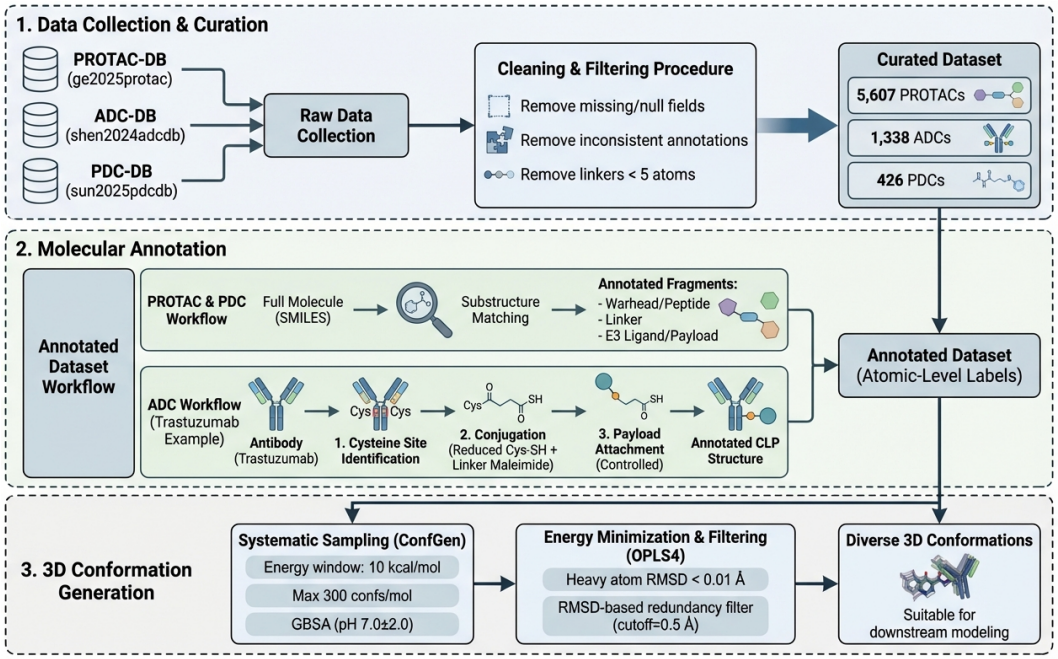


Figure 2: **Overview of the dataset construction workflow for heterobifunctional drugs.** The construction process includes three main steps: (1) data collection and curation from publicly available PROTAC, ADC, and PDC databases; (2) atomic-level annotation of functional modules, with ADC molecules annotated based on chemical conjugation reactions linking antibody residues, linker, and payload; and (3) 3D conformational generation using Schrödinger's ConfGen, producing diverse representative structures for downstream modeling and analysis.

- **Evaluation of existing models:** We systematically evaluate the performance of four representative structure-based generative linker design models on the HBDrug3D benchmark, providing insights into their strengths and limitations in heterobifunctional drug design.

2 HBDrug3D Dataset

2.1 Dataset Construction

Data Collection and Curation. We collected raw data from three publicly available heterobifunctional drug databases: PROTAC-DB (Ge et al., 2025), ADC-DB Shen et al. (2024), and PDC-DB (Sun et al., 2025). For PROTAC and PDC molecules, the datasets provide both the full molecular SMILES and separate SMILES for functional fragments, including the warhead/peptide, linker, and E3 ligand or payload. The ADC dataset additionally provides detailed information on the antibody–linker conjugate type.

To ensure data quality, we applied a uniform cleaning and filtering procedure across all datasets. Specifically, we removed: (1) entries with missing or null fields; (2) molecules with inconsistent or incorrect annotations for functional fragments or linkers; and (3) molecules with linkers shorter than 5 atoms, which lack physiological relevance. After this stringent curation, the resulting high-quality dataset comprises 5,607 PROTACs, 1,338 ADCs, and 426 PDC molecules.

Scaffold-based Split. We also performed a scaffold-based partitioning to evaluate model generalization to unseen scaffolds. First, conformations were clustered according to their molecular scaffolds. Training, validation, and test sets were then constructed such that scaffolds in validation and test sets did not overlap with those in the training set. To maintain comparable data sizes to the random split, we selected scaffolds so that the total number of conformations in validation and test sets matched: PROTAC: 400, PDC: 400, ADC: 400.

Detailed statistics on the molecular properties under two different data splitting strategies are provided in Appendix D.

Molecular Annotation. Building on the curated dataset, we performed atomic-level annotation to explicitly delineate functional modules. For PROTAC and PDC molecules, we annotated atoms according to the structure of warhead/peptide–linker–E3 ligand/payload. Each functional fragment was aligned to the full molecule via substructure matching, ensuring consistent atomic mapping.

Due to the involvement of antibodies, the annotation workflow for ADC molecules is relatively more complex. First, we identified the amino acid residues on the antibody that participate in linker conjugation based on the reported conjugate types. Across the entire dataset, we summarized 17 amino acids used for conjugation (see Appendix B). Subsequently, for each conjugate type, the linker was modified and conjugated to the reaction site, while the payload was attached to the linker in a controlled manner according to the corresponding reaction mechanism (see Appendix C for details). This process yielded well-defined conjugation–linker–payload (CLP) structures. Figure 3 illustrates this workflow using a trastuzumab-based ADC as a representative example, where interchain disulfide reduction generated reactive cysteine residues enabling selective conjugation. The workflow comprises three key stages: cysteine site identification, conjugation of the reduced cysteine -SH groups with the reactive maleimide groups on the linker, and controlled attachment of the payload to the linker via a dehydration condensation reaction. Finally, atomic-level annotations of the antibody conjugation sites, linker, and payload were completed using substructure matching, consistent with the approach applied to PROTAC molecules.

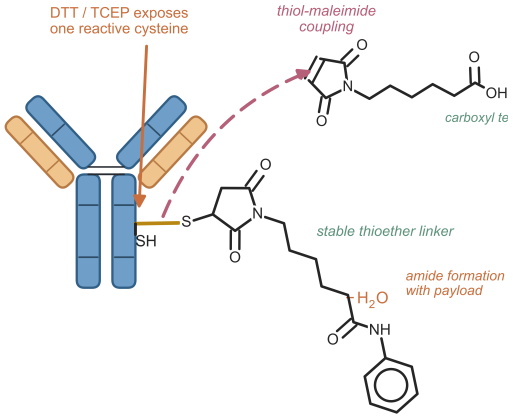


Figure 3: **Workflow of trastuzumab-based ADC conjugation.** Interchain disulfide reduction generates reactive cysteine residues, which selectively conjugate with maleimide groups on the linker. The payload is then attached via a dehydration condensation reaction.

3D Conformation Generation. All heterobifunctional molecules (including PROTACs, ADCs, and PDCs) were subjected to systematic 3D conformational sampling using Schrödinger's ConfGen module (Friesner et al., 2004). Sampling parameters were standardized: an energy window of 10 kcal/mol above the global minimum, a maximum of 300 conformations per molecule, and an implicit solvent model (GBSA) at physiological pH (7.0 ± 2.0).

Each generated conformation was energy-minimized using the OPLS4 force field (Lu et al., 2021) (heavy atom RMSD convergence threshold = 0.01 Å), followed by pairwise RMSD-based redundancy filtering (cutoff = 0.5 Å) to ensure conformational diversity. This protocol produced a diverse set of representative 3D structures suitable for downstream modeling and analysis.

2.2 Dataset Overview

Dataset Statistics. HBDrug3D covers three representative classes of linker-containing molecular systems, including PROTACs, PDCs, and ADCs. PROTACs and PDCs account for 41.15% and 37.89% of the generated conformations, respectively, while ADCs contribute 20.95%. These three subsets correspond to distinct linker design scenarios: PROTACs represent small-molecule-based bifunctional degraders, PDCs involve peptide-associated linkers, and ADCs contain antibody-conjugation-related linker structures. This dataset composition supports evaluation across chemically and structurally diverse linker regimes.

In terms of functional components, the dataset contains 808 warheads/conjugation sites, 2,379 linkers, and 184 E3 ligands for PROTACs. ADCs include 17 conjugation sites, 29 linkers, and 26 payloads, whereas PDCs have 426 peptides, 107 linkers, and 114 payloads. These numbers highlight the diversity of functional modules captured in the dataset, providing a rich resource for modeling both the chemical and structural variability of heterobifunctional molecules.

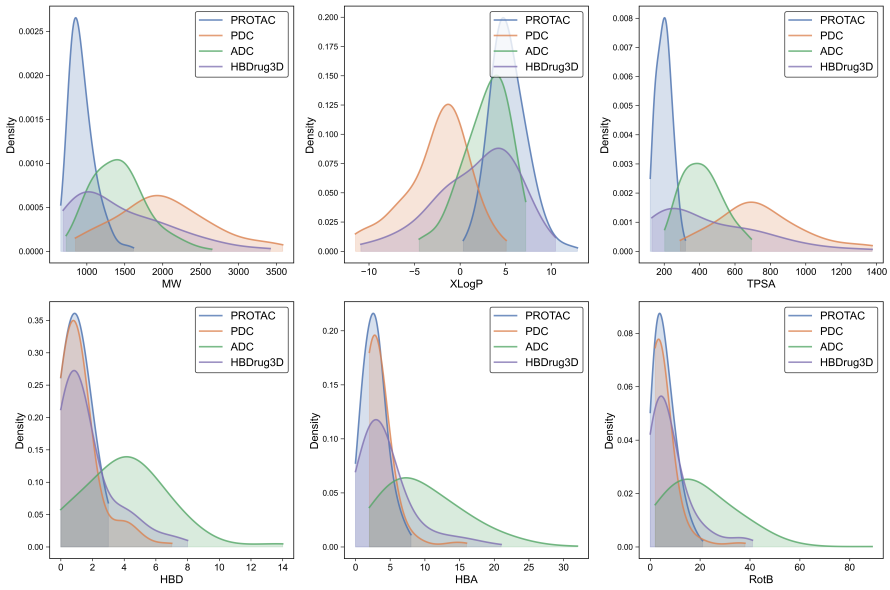


Figure 4: **Distribution of key physicochemical properties for PROTAC, PDC, ADC and HBDrug3D molecules.** The properties shown include molecular weight (MW), logP (XLogP), and topological polar surface area (TPSA), which are calculated for the entire molecule. In addition, hydrogen bond donor (HBD), hydrogen bond acceptor (HBA), and the number of rotatable bonds are calculated specifically for the linker region. The probability density function (KDE) curves, with shaded areas, represent the distribution of each property. Each compound type (PROTAC, PDC, ADC) is represented by a distinct color: PROTAC in blue, PDC in orange, and ADC in green.

Physicochemical Properties. Figure 4 illustrates the chemical properties of the three types of compounds. For ADC, we only report the conjugate-linker-payload portion rather than the entire antibody-drug conjugate. Regarding molecular weight (MW), PROTAC has the smallest MW. In terms of polarity, PDC has the lowest XlogP, indicating the highest hydrophilicity, followed by ADC and then PROTAC, which reflects PDC's better water solubility compared to the others. On the other hand, TPSA follows the opposite trend, with PROTAC having the highest TPSA, indicating its relatively higher hydrophilicity, while PDC has the lowest TPSA, indicating its greater hydrophobicity.

In the linker region, PRTOAC and PDC have similar hydrogen bond donor and acceptor counts, while ADC linker contains more hydrogen bond donors and acceptors. Moreover, the ADC linker exhibits significantly more rotatable bonds than PDC and PROTAC, indicating greater flexibility.

3 Benchmarking

3.1 Data Splits

We adopted two strategies to partition the dataset for benchmarking purposes.

Random Split. Each conformation was treated as an individual data point. The three subsets were split separately into training, validation, and test sets as follows: PROTAC: train/valid/test = 28,170 / 400 / 400, PDC: train/valid/test = 25,873 / 400 / 400, ADC: train/valid/test = 13,950 / 400 / 400.

For the combined HBDrug3D dataset, the splits were obtained by merging the three subsets, resulting in train/valid/test = 67,993 / 1200 / 1200.

Scaffold Split. For the scaffold split, we first extracted the fragment scaffold for each case and grouped cases sharing the same scaffold into the same cluster. The dataset was then partitioned at the cluster level, ensuring that all cases within a selected cluster were assigned to the same subset.

Clusters were sequentially assigned to the validation and test sets until each subset reached 400 cases, while the remaining clusters were used as the training set. This strategy prevents cases with the same fragment scaffold from appearing across different splits.

3.2 Benchmark Models

To systematically evaluate generative performance on heterobifunctional linker design, we select representative baseline models that cover the major modeling paradigms proposed in prior work. These benchmark models span conditional graph generation, equivariant variational autoencoders, and diffusion-based generative frameworks, providing a comprehensive comparison across different levels of 3D geometric modeling.

DeLinker (Imrie et al., 2020) is included as a representative conditional graph-based model. It generates linkers in an autoregressive manner conditioned on two fragments and predefined anchor atoms, while incorporating coarse geometric constraints such as inter-fragment distances and angles. DeLinker serves as a baseline that captures topological validity under limited 3D supervision.

3DLinker (Huang et al., 2022) extends this paradigm by explicitly modeling atom-level 3D geometry using an E(3)-equivariant graph variational autoencoder. It jointly predicts anchor atoms and generates both the linker graph and its 3D coordinates. As an autoregressive equivariant model, 3DLinker represents methods that enforce geometric consistency through symmetry-aware architectures.

Diffusion-based models are selected to represent a more recent class of generative approaches operating directly in continuous 3D space. **DiffLinker** (Igashov et al., 2024) performs diffusion over atom types and coordinates to generate linkers while keeping fragment geometries fixed, enabling flexible modeling of linker conformational distributions. **LinkerNet** (Guan et al., 2023) further generalizes this framework by allowing fragment translation and rotation during the diffusion process, thereby relaxing geometric constraints and improving adaptability to diverse linker conformations.

All benchmark models are retrained from scratch on each individual subset (PROTAC, ADC, and PDC) as well as on the combined HBDrug3D dataset, under both random and scaffold-based data splits, until convergence. For a fair comparison, the batch size is fixed to 64 across all models. The `max_num_atoms` parameter is adjusted to accommodate the size of molecules in our dataset, while all other hyperparameters are kept consistent with the original implementations.

3.3 Evaluation Metrics

General Metrics. We evaluate model performance using a set of commonly adopted metrics for 3D linker design. These include **validity**, which measures the percentage of chemically valid molecules; **recovery**, which assesses whether the ground-truth linker can be reproduced; **uniqueness**, indicating the diversity of generated molecules; and **novelty**, which quantifies the proportion of generated linkers not seen during training. In addition, we report the fraction of molecules that **pass standard 2D chemical filters**, ensuring basic chemical plausibility. To assess structural similarity between generated molecules and the ground truth, we further employ $\mathbf{SC_{RDKit}}$, which measures similarity based on 3D structural and conformational features.

3D Geometric Fidelity Metrics. Beyond molecular validity, we evaluate whether generated molecules preserve physically plausible 3D geometries. We report **Clash** to measure steric-clash-free structures and **Ring Flat.** to assess the planarity of ring systems. We further use **Min Bond**, **Max Bond**, and **Max Angle** to quantify local bond-length and bond-angle deviations, where values closer to 1 indicate more realistic geometry. Finally, **PB_Val.** denotes the overall PoseBusters validity pass rate, summarizing multiple 3D chemical and geometric checks.

Physicochemical and Linker-specific Properties. Beyond general generative metrics, we place particular emphasis on physicochemical properties that are critical for heterobifunctional drug design. These include molecular weight (**MW**), which is closely related to membrane permeability, topological polar surface area (**TPSA**), **XLogP** as a measure of molecular polarity, as well as the numbers of hydrogen bond acceptors (**HBA**) and donors (**HBD**). To specifically characterize linker behavior, we also analyze the **number of rotatable bonds** in the generated linkers, which serve as indicators of linker rigidity and flexibility. These properties are closely related to the ability of a linker to maintain favorable conformations and support stable ternary complex formation.

3.4 BENCHMARK RESULTS

Main Results. Table 2 summarizes the performance of four models evaluated on three subsets and on the full HBDrug3D dataset. Results are reported under the scaffold splitting strategy, while those obtained with random splitting are provided in Table 6.

Table 2: **Evaluation of generated molecules using general metrics under the scaffold splitting strategy.** For each case, each model generated 100 molecules, and the reported values are the mean ± standard deviation across three independent runs. SC_{RDKit} is reported with three decimal places; the standard deviation is not explicitly shown because rounding to two decimal places would result in zeros.

Dataset	Model	Valid (↑)	Unique (↑)	Novel (↑)	Recover (↑)	2D filters (↑)	SC_{RDKit} (↑)
PROTAC	DeLinker	87.83 ± 0.13	61.76 ± 2.46	79.86 ± 2.38	**40.00 ± 7.75**	87.25 ± 0.15	0.250
	3DLinker	70.35 ± 0.18	**94.48 ± 1.02**	**91.54 ± 1.25**	26.67 ± 6.83	69.72 ± 0.17	0.861
	DiffLinker	73.52± 0.08	85.19 ± 1.44	69.91 ± 1.56	22.73 ± 4.32	68.77± 0.12	**0.882**
	LinkerNet	**93.01 ± 0.06**	47.34 ± 2.58	55.13 ± 1.64	35.00 ± 4.97	**91.92 ± 0.06**	0.513
PDC	DeLinker	70.87 ± 0.13	45.14 ± 0.00	96.89 ± 0.00	60.00 ± 0.00	61.31 ± 0.12	0.097
	3DLinker	47.92 ± 0.15	**98.04 ± 0.00**	**99.40 ± 0.00**	0.00 ± 0.00	42.88 ± 0.19	0.916
	DiffLinker	46.67 ± 0.11	49.82 ± 0.09	39.77 ± 0.12	**75.15 ± 1.22**	41.28 ± 0.18	**0.964**
	LinkerNet	**86.38 ± 0.13**	36.38 ± 0.07	38.35 ± 0.14	51.17 ± 1.36	**72.03 ± 0.05**	0.626
ADC	DeLinker	11.36 ± 0.08	87.39 ± 0.05	100.0 ± 0.00	0.00 ± 0.00	**97.42 ± 0.20**	0.167
	3DLinker	29.28 ± 0.14	**100.0 ± 0.00**	100.0 ± 0.00	0.00 ± 0.00	38.34 ± 0.29	0.771
	DiffLinker	32.16 ± 0.21	95.47 ± 0.06	100.0 ± 0.00	0.00 ± 0.00	41.37 ± 0.28	**0.803**
	LinkerNet	**34.54 ± 0.34**	91.23 ± 0.06	100.0 ± 0.00	0.00 ± 0.00	45.73 ± 0.31	0.547
HBDrug3D	DeLinker	13.02 ± 0.05	47.74 ± 0.00	**100.0 ± 0.00**	5.88 ± 0.19	**81.13 ± 0.13**	0.220
	3DLinker	58.96 ± 0.09	**93.20 ± 0.02**	92.06 ± 0.07	17.58 ± 0.06	66.20 ± 0.10	0.884
	DiffLinker	59.22 ± 0.06	83.78 ± 0.66	68.74 ± 0.57	25.13 ± 2.16	64.22 ± 0.04	**0.886**
	LinkerNet	**73.42 ± 0.07**	41.47 ± 0.42	51.82 ± 0.64	**53.38 ± 3.41**	79.06 ± 0.03	0.548

Across all datasets, LinkerNet achieves the highest validity, while the second-best model varies by dataset. We observe a clear decline in validity from PROTAC to PDC and further to ADC, where validity ranges from about 11% to 35% under the scaffold split. This trend is likely attributable to the longer and more flexible linkers in the ADC subset, characterized by increased linker length and a higher number of rotatable bonds, which make it more challenging to generate chemically valid molecules. The lower validity may also reflect the additional constraints imposed by complex three-dimensional structures.

Table 3: **3D geometric fidelity under the scaffold splitting strategy.** For each case, 100 molecules were generated per model; values are mean ± standard deviation across three independent runs. Clash, Ring Flat., and PB_Val. are pass rates (higher is better); Min Bond, Max Bond, and Max Angle are relative ratios (closer to 1 is better). Bold indicates the best value within each dataset; columns in which all models tie are left unbolded.

Dataset	Model	Clash (↑)	Ring Flat. (↑)	Min Bond	Max Bond	Max Angle	PB_Val. (↑)
PROTAC	DeLinker	**98.33 ± 0.04**	100.0 ± 0.00	**0.97 ± 0.00**	**1.04 ± 0.00**	**1.06 ± 0.00**	**98.33 ± 0.04**
	3DLinker	14.67 ± 0.04	100.0 ± 0.00	0.26 ± 0.01	1.41 ± 0.02	1.76 ± 0.01	0.00 ± 0.00
	DiffLinker	95.85 ± 0.15	100.0 ± 0.00	0.95 ± 0.00	1.06 ± 0.00	1.09 ± 0.00	95.50 ± 0.15
	LinkerNet	80.87 ± 0.03	100.0 ± 0.00	0.82 ± 0.00	3.75 ± 0.04	2.69 ± 0.02	25.24 ± 0.48
PDC	DeLinker	62.32 ± 0.09	100.0 ± 0.00	**0.95 ± 0.00**	**1.04 ± 0.00**	**1.07 ± 0.00**	62.32 ± 0.09
	3DLinker	1.14 ± 0.03	100.0 ± 0.00	0.19 ± 0.01	3.36 ± 0.08	2.53 ± 0.04	0.00 ± 0.00
	DiffLinker	**98.77 ± 0.04**	100.0 ± 0.00	0.91 ± 0.00	1.06 ± 0.00	1.10 ± 0.02	**97.28 ± 0.03**
	LinkerNet	46.41 ± 0.02	100.0 ± 0.00	0.83 ± 0.00	6.54 ± 0.04	4.26 ± 0.02	3.9 ± 0.14
ADC	DeLinker	44.38 ± 0.13	100.0 ± 0.00	**0.96 ± 0.00**	**1.05 ± 0.00**	**1.08 ± 0.00**	**39.83 ± 0.22**
	3DLinker	0.00 ± 0.00	100.0 ± 0.00	0.03 ± 0.00	9.79 ± 0.66	6.58 ± 0.43	0.00 ± 0.00
	DiffLinker	**46.47 ± 0.00**	100.0 ± 0.00	0.89 ± 0.00	1.06 ± 0.00	1.12 ± 0.01	26.31 ± 0.24
	LinkerNet	24.75 ± 0.56	100.0 ± 0.00	0.78 ± 0.00	6.74 ± 0.77	3.25 ± 0.06	3.37 ± 0.55
HBDrug3D	DeLinker	66.73 ± 0.20	99.90 ± 0.01	**0.96 ± 0.00**	1.09 ± 0.01	**1.08 ± 0.00**	54.72 ± 0.10
	3DLinker	0.41 ± 0.01	99.65 ± 0.02	0.16 ± 0.02	4.99 ± 0.73	3.74 ± 0.52	0.00 ± 0.00
	DiffLinker	**74.56 ± 0.14**	**100.0 ± 0.00**	0.92 ± 0.00	**1.07 ± 0.00**	1.10 ± 0.00	**56.20 ± 0.14**
	LinkerNet	39.53 ± 0.09	99.95 ± 0.00	0.82 ± 0.00	5.58 ± 0.00	3.96 ± 0.00	7.46 ± 0.36

Table 4: **Evaluation of generated molecules using physicochemical and linker-specific properties under the scaffold splitting strategy.** For each case, each model generated 100 molecules, and the reported values are the mean ± standard deviation across three independent runs. Abbreviations: MW = molecular weight, XLogP = octanol–water partition coefficient, TPSA = topological polar surface area, HBD = hydrogen bond donor, HBA = hydrogen bond acceptor, RotB = number of rotatable bonds. Reference values correspond to the averages computed on the respective test sets.

Dataset	Model	MW	XLogP	TPSA	HBD	HBA	RotB
PROTAC	Reference	906	5.3	201.8	1.0	2.5	6.4
	DeLinker	924 ± 0.22	2.7 ± 0.01	191.3 ± 0.11	1.2 ± 0.00	2.5 ± 0.01	7.1 ± 0.01
	3DLinker	912 ± 0.12	2.4 ± 0.01	187.2 ± 0.11	1.1 ± 0.00	2.5 ± 0.01	5.7 ± 0.01
	DiffLinker	891 ± 0.13	4.3 ± 0.00	193.6 ± 0.03	1.1 ± 0.00	2.5 ± 0.00	6.7 ± 0.00
	LinkerNet	898 ± 0.06	4.1 ± 0.00	196.2 ± 0.06	1.0 ± 0.00	2.4 ± 0.00	6.9 ± 0.02
PDC	Reference	1817	0.73	623.7	1.3	3.2	5.1
	DeLinker	1597 ± 0.33	-1.1 ± 0.00	542.4 ± 0.15	1.1 ± 0.01	2.6 ± 0.01	3.5 ± 0.02
	3DLinker	1553 ± 1.87	-1.8 ± 0.02	537.3 ± 0.18	1.6 ± 0.00	3.5 ± 0.02	3.3 ± 0.02
	DiffLinker	1746 ± 0.16	-3.0 ± 0.01	606.5 ± 0.15	1.0 ± 0.00	2.8 ± 0.00	4.1 ± 0.00
	LinkerNet	1787 ± 0.45	-3.2 ±0.00	620.2 ± 0.16	1.0 ± 0.00	2.9 ± 0.00	4.5 ± 0.01
ADC	Reference	1611	1.7	495.7	4.2	11.9	28.1
	DeLinker	1491 ± 1.88	-4.4 ± 0.04	519.6 ± 0.78	8.1 ± 0.02	10.7 ± 0.02	11.7 ± 0.03
	3DLinker	1561 ± 2.07	-0.2 ± 0.02	528.2 ± 0.85	6.3 ± 0.02	12.9 ± 0.03	17.2 ± 0.05
	DiffLinker	1434 ± 0.12	-0.3 ± 0.03	456.7 ± 0.39	4.6 ± 0.01	6.7 ± 0.01	11.5 ± 0.05
	LinkerNet	1569 ± 0.48	-1.1 ± 0.02	501.3 ± 0.34	5.5 ± 0.02	9.4 ± 0.01	16.7 ± 0.14
HBDrug3D	Reference	1445	2.6	437.8	2.2	5.8	13.2
	DeLinker	1686 ± 2.40	3.5 ± 0.02	538.2 ± 0.90	1.8 ± 0.01	3.6 ± 0.02	3.2 ± 0.03
	3DLinker	1352 ± 0.79	2.9 ± 0.01	413.3 ± 0.30	2.1 ± 0.00	5.6 ± 0.01	7.6 ± 0.02
	DiffLinker	1369 ± 1.03	3.1 ± 0.01	442.6 ± 0.44	1.9 ± 0.00	5.2 ± 0.01	7.4 ± 0.01
	LinkerNet	1487 ± 0.32	1.7 ± 0.00	483.2 ± 0.14	1.9 ± 0.00	4.9 ± 0.00	8.3 ± 0.01

3D Geometric Fidelity of Generated Molecules.

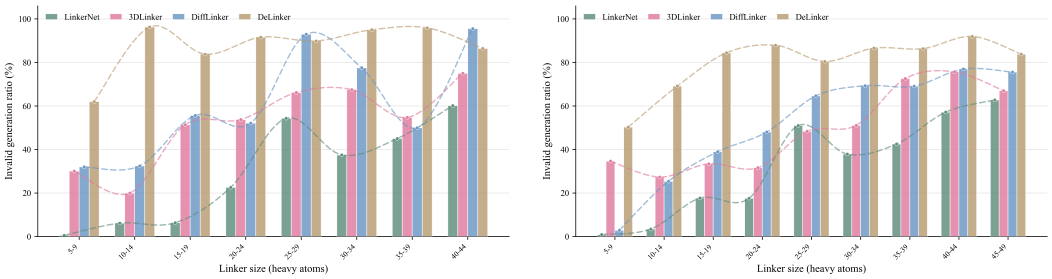
Table 3 reports the 3D geometric fidelity of generated molecules under the scaffold splitting strategy. Overall, DeLinker and DiffLinker consistently show stronger 3D validity than 3DLinker and LinkerNet across the four datasets. DeLinker maintains the most stable local geometry, with bond-length and bond-angle ratios generally closest to the ideal value of 1, while DiffLinker achieves the highest clash pass rate and PB validity on PDC and HBDrug3D. On PROTAC, DeLinker performs best overall, reaching 98.33% in both clash pass rate and PB validity.

The results also reveal clear failure patterns for the other two baselines. 3DLinker suffers from severe geometric degradation, with PB validity dropping to 0 across all datasets despite high ring-flatness scores. LinkerNet also shows large deviations in maximum bond length and bond angle, leading to low PB validity, especially on PDC, ADC, and HBDrug3D. These results suggest that preserving valid local bond and angle geometry remains a key challenge for 3D linker generation under scaffold-based evaluation. The corresponding results under the random splitting strategy are reported in Table 7.

Physicochemical Properties of Generated Molecules.

Table 4 reports the physicochemical and linker-specific properties of generated molecules under the scaffold splitting strategy. On PROTAC, all models produce molecules with molecular weight, TPSA, HBD, HBA, and RotB values close to the reference distribution, although XLogP is consistently underestimated. For PDC, DiffLinker and LinkerNet better preserve the reference MW and TPSA, while all models generate molecules with lower XLogP and fewer rotatable bonds than the reference.

On ADC, the generated molecules generally maintain comparable MW and TPSA, but show larger deviations in linker-specific properties, especially HBD, HBA, and RotB, reflecting the increased difficulty of matching the highly flexible ADC linker region. On the combined HBDrug3D benchmark, 3DLinker, DiffLinker, and LinkerNet produce overall physicochemical profiles closer to the reference averages than DeLinker, whereas DeLinker tends to generate molecules with higher MW and TPSA but substantially fewer rotatable bonds. Overall, these results show that different models capture global molecular properties and linker-specific properties with varying accuracy under scaffold-based generalization. Results on the random split are shown in Table 8.



(a) Scaffold split **(b) Random split**

Figure 5: Invalid generation ratio with respect to linker size on the HBDrug3D test set.

4 Failure Mode Analysis

To investigate the major factors underlying invalid generations, we analyzed the invalid candidate ratio as a function of linker size on the HBDrug3D test set. Fig. 5, the invalid ratio of all four models is strongly associated with linker size. DeLinker already maintains an invalid ratio above 80% once the linker size exceeds 15 heavy atoms. DiffLinker shows a clear increase in invalid generations when the linker size is larger than 25 heavy atoms, with the invalid ratio exceeding 50%. For 3DLinker, the invalid ratio exceeds 50% when the linker size reaches approximately 30–35 heavy atoms, whereas LinkerNet remains relatively more robust for shorter linkers but also exceeds 50% invalid generations when the linker size is larger than 40 heavy atoms. These results indicate that larger linkers substantially increase the design difficulty for all four generation models.

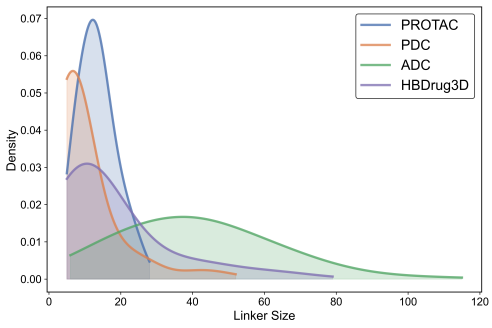


Figure 6: Linker size distributions across PROTAC, PDC, ADC, and HBDrug3D.

We further examined the linker-size distributions across PROTAC, PDC, ADC, and HBDrug3D, as shown in Fig. 6. Compared with PROTAC and PDC, ADC contains substantially larger linkers, shifting the data distribution toward the high-linker-size regime where all models exhibit higher invalid ratios. This distributional property provides a direct explanation for the consistently lower valid ratios of all four models on the ADC dataset reported in Table 2.

5 Conclusions

In this work, we introduce HBDrug3D, a large-scale dataset containing 70,393 conformational structures of heterobifunctional molecules. The dataset provides functional-group-level annotations for all atoms, together with low-energy three-dimensional conformations for each molecule, enabling detailed analysis and modeling of molecular geometry and chemical functionality. By explicitly associating atom-level functional roles with 3D conformations, HBDrug3D offers a unified representation that bridges molecular topology, chemistry, and spatial structure.

To demonstrate the value of HBDrug3D for structure-based molecular design, we focus on the representative linker design task in heterobifunctional drug discovery and conduct a systematic evaluation of four state-of-the-art models. In addition to commonly used linker design metrics, we further incorporate a diverse set of physicochemical properties and geometric metrics to enable a more comprehensive and fine-grained assessment of the generated molecules. This evaluation protocol establishes a solid foundation for future holistic and systematic benchmarking of linker generation models, and provides insights into model strengths, weaknesses, and characteristic tendencies across different molecular subsets and design scenarios.

AI USE STATEMENT

We used generative AI tools to edit the manuscript for readability, grammar, and clarity. We have reviewed all AI-assisted text and take responsibility for the final content of this work.

REFERENCES

Simon Axelrod and Rafael Gómez-Bombarelli. GEOM: Energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9:185, 2022. doi: 10.1038/s41597-022-01288-4.

Miklós Békés, David R Langley, and Craig M Crews. Protac targeted protein degraders: the past is prologue. *Nature reviews Drug discovery*, 21(3):181–200, 2022.

Célia Bouacha, Séverine Denoyelle, and Sonia Cantel. From design to outcome: The role of bi-and multifunctional ligands in modern therapeutics. *Advanced Therapeutics*, 8(10):e00146, 2025.

Ahmed Bulldan, Min Zheng, Christian Meyners, Patrick Purder, Johannes Krieger, Johannes Dreizler, Thomas M Geiger, Max Repity, Marte Høen Lein, Ingrid Quist-Løkken, et al. Heterobifunctional protein binders enable cell type-specific killing through in-cell enrichment. *bioRxiv*, pp. 2025–05, 2025.

Jingyuan Dai, Milad Ashrafizadeh, Amir Reza Aref, Gautam Sethi, and Yavuz Nuri Ertas. Peptide-functionalized,-assembled and-loaded nanoparticles in cancer therapy. *Drug Discovery Today*, pp. 103981, 2024.

Jan G Felber, Anil P Jagtap, Philipp Ochtrop, Saskia Schmitt, Isabelle Mai, Sarah Herterich, Paul Machui, Dietmar König, simon vogt, Dominik Schumacher, et al. Antibody-degrader conjugates unlock the therapeutic potential of heterobifunctional degraders through antibody-like pharmacokinetics and excellent in vivo efficacy. *Cancer Research*, 85(8_Supplement_1):7018–7018, 2025.

Richard A Friesner, Jay L Banks, Robert B Murphy, Thomas A Halgren, Jasna J Klicic, Daniel T Mainz, Matthew P Repasky, Eric H Knoll, Mee Shelley, Jason K Perry, et al. Glide: a new approach for rapid, accurate docking and scoring. 1. method and assessment of docking accuracy. *Journal of medicinal chemistry*, 47(7):1739–1749, 2004.

Jingxuan Ge, Shimeng Li, Gaoqi Weng, Huating Wang, Meijing Fang, Huiyong Sun, Yafeng Deng, Chang-Yu Hsieh, Dan Li, and Tingjun Hou. Protac-db 3.0: an updated database of protacs with extended pharmacokinetic parameters. *Nucleic Acids Research*, 53(D1):D1510–D1515, 2025.

Jiaqi Guan, Xingang Peng, PeiQi Jiang, Yunan Luo, Jian Peng, and Jianzhu Ma. Linkernet: Fragment poses and linker co-design with 3d equivariant diffusion. *Advances in Neural Information Processing Systems*, 36:77503–77519, 2023.

Yinan Huang, Xingang Peng, Jianzhu Ma, and Muhan Zhang. 3dlinker: an e (3) equivariant variational autoencoder for molecular linker design. *arXiv preprint arXiv:2205.07309*, 2022.

Robin C Humphreys, Jessica Kirtely, Amha Hewit, Sandra Biroc, Nick Knudsen, Lillian Skidmore, and Alan Wahl. Site specific conjugation of arx-788, an antibody drug conjugate (adc) targeting her2, generates a potent and stable targeted therapeutic for multiple cancers. *Cancer Research*, 75 (15_Supplement):639–639, 2015.

Ilia Igashov, Hannes Stärk, Clément Vignac, Arne Schneuing, Victor Garcia Satorras, Pascal Frossard, Max Welling, Michael Bronstein, and Bruno Correia. Equivariant 3d-conditional diffusion model for molecular linker design. *Nature Machine Intelligence*, pp. 1–11, 2024.

Fergus Imrie, Anthony R Bradley, Mihaela van der Schaar, and Charlotte M Deane. Deep generative models for 3d linker design. *Journal of chemical information and modeling*, 60(4):1983–1995, 2020.

John J. Irwin and Brian K. Shoichet. ZINC: A free database of commercially available compounds for virtual screening. *Journal of Chemical Information and Modeling*, 45(1):177–182, 2005. doi: 10.1021/ci049714+.

Sebastian Jäger, Stephan Dickgiesser, Jason Tonillo, Stefan Hecht, Harald Kolmar, and Christian Schröter. Egfr binding fc domain-drug conjugates: stable and highly potent cytotoxic molecules mediate selective cell killing. *Biological Chemistry*, 403(5-6):525–534, 2022.

Florence Lhospice, D Brégeon, C Belmant, Patrick Dennler, A Chiotellis, E Fischer, L Gauthier, A Boëdec, H Rispaud, S Savard-Chambard, et al. Site-specific conjugation of monomethyl auristatin e to anti-cd30 antibodies improves their pharmacokinetics and therapeutic index in rodent models. *Molecular pharmaceutics*, 12(6):1863–1871, 2015.

Chao Lu, Chuanjie Wu, Delaram Ghoreishi, Wei Chen, Lingle Wang, Wolfgang Damm, Gregory A Ross, Markus K Dahlgren, Ellery Russell, Christopher D Von Bargen, et al. Opls4: Improving force field accuracy on challenging regimes of chemical space. *Journal of chemical theory and computation*, 17(7):4291–4300, 2021.

Qunfeng Luo, Shuli Liu, Yaoguang Hua, Chunqiu Long, Juncheng Li, Yuzhi Zhang, et al. Heterobifunctional cross-linker with dinitroimidazole and azide modules for protein and oligonucleotide functionalization. *RSC advances*, 15(6):4526–4531, 2025.

Mary E Matyskiela. Targeted protein degradation comes of age. *nature chemical biology*, 21(1): 22–23, 2025.

Byeongkwi Min, Jonghwa Jin, Hyeree Kim, Nam-Gu Her, Changsik Park, Donggeon Kim, Jehoon Yang, Juhyeon Hwang, Eunmi Kim, Minji Choi, et al. circr201-dpbd, a novel pyrrolobenzodiazepine dimer-containing site-specific antibody–drug conjugate targeting c-met overexpression tumors. *ACS omega*, 5(40):25798–25809, 2020.

Suho Park, Sun Young Kim, Jongun Cho, Doohwan Jung, Donghoon Seo, Jaeho Lee, Sangkwang Lee, Sanghyeon Yun, Hyangsook Lee, Okku Park, et al. Aryl sulfate is a useful motif for conjugating and releasing phenolic molecules: sulfur fluorine exchange click chemistry enables discovery of ortho-hydroxy-protected aryl sulfate linker. *Bioconjugate Chemistry*, 30(7):1957–1968, 2019.

David Rabuka, Jason S Rush, Gregory W Dehart, Peng Wu, and Carolyn R Bertozzi. Site-specific chemical protein conjugation using genetically encoded aldehyde tags. *Nature protocols*, 7(6): 1052–1067, 2012.

Alexander Schneuing, Charles Harris, Yilun Du, et al. Structure-based drug design with equivariant diffusion models. *Nature Computational Science*, 4:899–909, 2024. doi: 10.1038/s43588-024-00737-x.

Liteng Shen, Xiuna Sun, Zhen Chen, Yu Guo, Zheyuan Shen, Yi Song, Wenxiu Xin, Haiying Ding, Xinyue Ma, Weiben Xu, et al. Adcdb: the database of antibody–drug conjugates. *Nucleic Acids Research*, 52(D1):D1097–D1109, 2024.

Wei Shi, Jianxin Zhang, Liya Liu, Wanzhen Li, Zhi Liu, Anni Ren, Jie Wang, Caihong Tang, Yang Yang, Dandan Xu, et al. Hiding payload inside the igg fc cavity significantly enhances the therapeutic index of antibody–drug conjugates. *Journal of Medicinal Chemistry*, 66(1):1011–1026, 2022.

Mayu Shiomi, Shinya Matsuzaki, Satoshi Serada, Koji Matsuo, Chihiro Mizuta-Odani, Mariko Jitsumori, Ruriko Nakae, Satoko Matsuzaki, Satoshi Nakagawa, Kosuke Hiramatsu, et al. Cd70 antibody-drug conjugate: A potential novel therapeutic agent for ovarian cancer. *Cancer Science*, 112(9):3655–3668, 2021.

Xiuna Sun, Hanyang Li, Zhen Chen, Yang Zhang, Zhangle Wei, Hangwei Xu, Yang Liao, Wanghao Jiang, Yichao Ge, Lingyan Zheng, et al. Pdcdb: the biological activity and pharmaceutical information of peptide–drug conjugate (pdc). *Nucleic Acids Research*, 53(D1):D1476–D1485, 2025.

Cheng-Yueh Ting, Paul T Kolbeck, Raffaele Colombo, Chacko Chakiath, Megan Rice, Marcello Marelli, and R James Christie. Cyclopentadiene as a multifunctional reagent for normal-and inverse-electron demand diels–alder bioconjugation. *Bioconjugate Chemistry*, 33(9):1609–1619, 2022.

K Tsuchikama and Z An. Antibody-drug conjugates: recent advances in conjugation and linker chemistries. protein cell 9: 33–46, 2018.

Kyoji Tsuchikama, Yasuaki Anami, Summer YY Ha, and Chisato M Yamazaki. Exploring the next generation of antibody–drug conjugates. *Nature Reviews Clinical Oncology*, 21(3):203–223, 2024.

Max M Wang, Mihai I Truica, Brayley S Gattis, Julia Oktawiec, Vinay Sagar, Ananya A Basu, Paul A Bertin, Xiaoyu Zhang, Sarki A Abdulkadir, and Nathan C Gianneschi. Heterobifunctional proteomimetic polymers for targeted protein degradation. *bioRxiv*, pp. 2025–03, 2025.

Xiao Zhang, Chong Ou, Huiying Liu, and Lai-Xi Wang. Synthesis and evaluation of three azide-modified disaccharide oxazolines as enzyme substrates for single-step fc glycan-mediated antibody-drug conjugation. *Bioconjugate chemistry*, 33(6):1179–1191, 2022.

A Related Work

A.1 Heterobifunctional Drugs and Linker Design

Heterobifunctional drugs have received increasing attention in recent years due to their ability to achieve targeted therapeutic effects via dual-module mechanisms. Several studies have explored their chemical design, structural characteristics, and pharmacological properties (Bulldan et al., 2025; Matyskiela, 2025; Wang et al., 2025). These molecules generally consist of a targeting warhead and an effector module connected by a linker, enabling precise control over protein degradation or cytotoxic payload delivery. Importantly, the linker plays a central role in stabilizing the 3D conformation, modulating pharmacokinetics, and ensuring targeting specificity, and it strongly influences overall efficacy and safety. Therefore, rational linker design is critical for the functional success of heterobifunctional drugs.

A.1.1 Proteolysis-Targeting Chimeras (PROTACs)

PROTACs recruit E3 ubiquitin ligases to mediate selective degradation of target proteins. Numerous studies have characterized PROTAC-mediated ternary complex formation, linker optimization, and degradation efficiency (Békés et al., 2022; Luo et al., 2025). Structural and biochemical analyses highlight that linker length and flexibility are key factors for effective ubiquitination, but computational models for rational linker design remain limited.

A.1.2 Peptide-Drug Conjugates (PDCs)

PDCs leverage peptide targeting modules to deliver cytotoxic payloads to tumor cells. Previous work has investigated peptide–receptor binding specificity, payload release mechanisms, and in vivo efficacy (Tsuchikama et al., 2024; Dai et al., 2024). While experimental studies have explored linker chemistry, systematic computational methods for PDC linker optimization are still scarce.

A.1.3 Antibody-Drug Conjugates (ADCs)

ADCs utilize antibodies as targeting modules to guide potent cytotoxins to cancer cells. Several reviews and experimental studies have addressed ADC pharmacokinetics, conjugation strategies, and linker stability (Felber et al., 2025; Bouacha et al., 2025). However, the diversity of antibody–payload combinations and the influence of linker properties pose challenges for computational modeling.

A.2 Datasets for Heterobifunctional Drug Design

A number of datasets have been developed to support molecular modeling, but they differ in focus and coverage. General-purpose small-molecule databases, such as ZINC (Irwin & Shoichet, 2005), GEOM (Axelrod & Gómez-Bombarelli, 2022), and Pockets (Schneuing et al., 2024), provide large collections of compounds with available 3D conformations. For instance, ZINC includes over 439,000 molecules with an average of 25 atoms, while GEOM offers around 285,000 molecules averaging 32 atoms. These datasets are highly valuable for general 3D molecular modeling; however, they are restricted to small molecules and do not capture the structural and functional characteristics of heterobifunctional drugs. As a result, although linker-like fragments can be derived from these compounds, they differ substantially from the linkers in PROTACs, PDCs, or ADCs in terms of molecular weight, atom count, and length, limiting their direct applicability for heterobifunctional drug design.

In contrast, specialized datasets have been assembled for heterobifunctional drugs. PROTAC-DB (Ge et al., 2025) contains 6,111 PROTACs, PDC-DB (Sun et al., 2025) includes 2,036 PDCs, and ADC-DB Shen et al. (2024) provides 6,572 ADC molecules. These resources offer useful information about sequences, activity, and conjugation sites, but they do not provide 3D structural data. Without conformational information, computational modeling of linkers remains challenging, and structure-based methods cannot be applied directly.

Taken together, these observations highlight a clear gap: existing datasets either provide 3D conformations but focus solely on small molecules, or cover heterobifunctional drugs but lack structural information. To bridge this gap, we constructed HBDrug3D, a large-scale dataset that includes over

70,393 3D conformations of PROTACs, PDCs, and ADCs, with an average of 103 atoms per molecule. By combining structural detail with functional diversity, HBDrug3D offers a standardized platform for systematic computational studies and rational linker design.

A.3 Generative Models for Linker Design

Recent years have seen a surge of interest in generative models for 3D linker design, aiming to construct linkers that connect two molecular fragments while satisfying both chemical and geometric constraints. **DeLinker** (Imrie et al., 2020) represents the first deep learning-based approach for this task. It formulates linker generation as a conditional graph generative problem, progressively constructing the linker graph starting from two given fragments and their anchor nodes. Spatial information, such as the distance and angle between fragments, is incorporated as side information to guide the generation. However, this approach only leverage coarse 3D information and do not explicitly model atom-level geometry, limiting their ability to fully capture fragment conformations.

Building on these limitations, **3DLinker** (Huang et al., 2022) introduces a method capable of jointly predicting anchor atoms and generating both the linker graph and its 3D structure. It is based on an E(3) equivariant graph variational autoencoder, which allows the model to respect 3D rotational and translational symmetries in molecular structures. Notably, both DeLinker and 3DLinker are auto-regressive models that progressively add linker atoms conditioned on fragment embeddings.

More recently, diffusion-based methods have emerged to model the linker generation process more flexibly. **DiffLinker** (Igashov et al., 2024) applies a diffusion process over atom types and coordinates to generate the linker while keeping the fragment coordinates fixed. **LinkerNet** (Guan et al., 2023) further extends this idea by allowing diffusion-based translation and rotation of fragments during linker generation, providing greater flexibility in modeling linker conformations and enabling the fragments to adjust their relative positions to better accommodate the generated linker.

Collectively, these models illustrate the progression from coarse 3D conditional graph generation to fully equivariant, diffusion-based approaches, highlighting the importance of precise 3D modeling in designing chemically and geometrically valid linkers for heterobifunctional drugs.

B Conjugation Sites

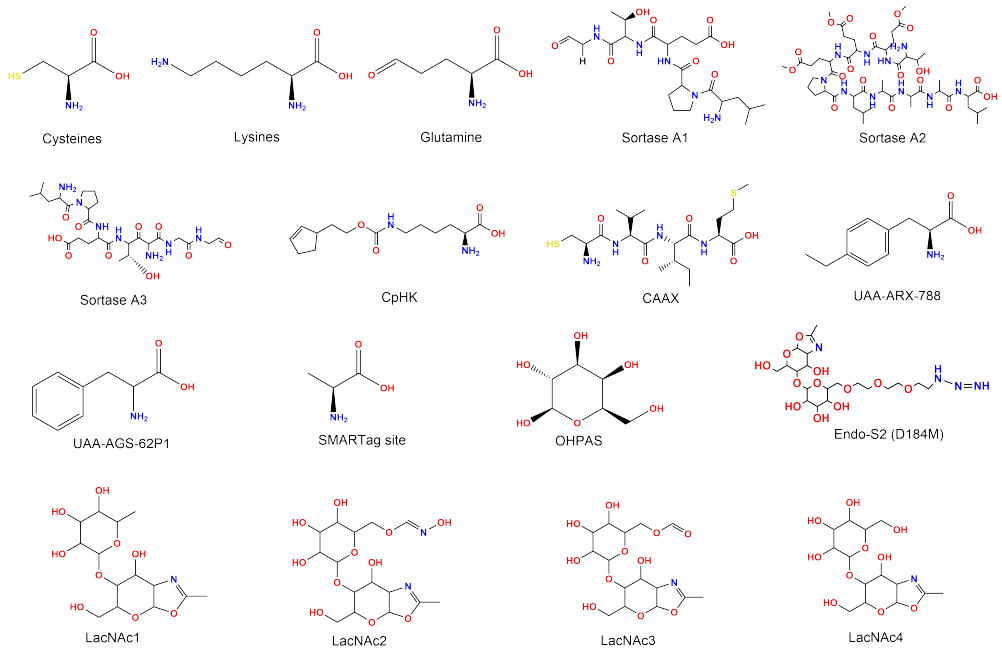


Figure 7: **Overview of antibody–drug conjugate (ADC) conjugation strategies included in this study.** The ADC subset encompasses 17 representative conjugation sites, spanning conventional chemical conjugation (e.g., cysteine thiol–maleimide and lysine NHS ester), enzymatic approaches (Sortase A transpeptidation and microbial transglutaminase–mediated amidation), glycoengineering-based strategies (LacNAc remodeling), and genetically encoded conjugation methods (unnatural amino acids and SMARTag formylglycine).

C Antibody–Drug Conjugate (ADC) Conjugation Strategies

Table 5: Representative antibody–drug conjugation sites and linkage strategies included in the ADC subset. The table summarizes commonly used chemical, enzymatic, glycoengineering, and genetically encoded approaches, together with their corresponding conjugation mechanisms and representative implementation strategies.

Conjugation Site	Chemical / Enzymatic Mechanism	Representative Strategies
Cysteine	Thiol-reactive chemistry (e.g., maleimide) targeting reduced cysteine -SH groups (Tsuchikama & An, 2018)	• Random conjugation via interchain disulfides • Engineered cysteine variants (e.g., THIOMAB™, H38C2) • Glycan-proximal cysteine targeting
Lysine	NHS ester-mediated amidation of ϵ-amino groups	• Random conjugation on surface-exposed lysines • Engineered lysine sites • Glycan-adjacent lysine modification

Continued on next page

Continued from previous page

Conjugation Site	Chemical / Enzymatic Mechanism	Representative Strategies
Glutamine (mTG)	Microbial transglutaminase (mTG)-mediated transamidation at native or engineered glutamine residues (Lhospice et al., 2015)	• Native Q295/Q311/Q438 sites • Engineered N→Q mutations • C-terminal glutamine-containing tags
Sortase A	Sortase A-mediated LPXTG motif cleavage followed by oligoglycine ligation (Jäger et al., 2022)	• C-terminal LPXTG-tagged antibodies • N-terminal oligoglycine-modified payloads
CpHK	Cyclopentadiene-based Diels–Alder cycloaddition chemistry (Ting et al., 2022)	• Rapid reaction kinetics ($> 10^4$ $M^{-1}s^{-1}$) • Compatibility with live-cell labeling
CAAX	Prenyltransferase-catalyzed installation of azido-isoprenyl groups at C-terminal CAAX motifs (Min et al., 2020)	• Light-chain C-terminal modification • Subsequent thioether click conjugation
Unnatural Amino Acids (UAA)	Genetic incorporation of bioorthogonal functional groups (e.g., azide or alkyne) (Shiomi et al., 2021; Humphreys et al., 2015)	• Direct encoding of pAcF or AzF residues • Copper-free click chemistry conjugation
SMARTag	Formylglycine-generating enzyme (FGE)-mediated oxidation of CxPxR motifs followed by HIPS ligation (Rabuka et al., 2012)	• Native CxPxR sequence tags • Engineered aldehyde-forming insertions
OHPAS	Self-immolative aryl sulfate linker with latent phenolic trigger (Park et al., 2019)	• Tumor-activated payload release • Enhanced plasma stability
Endo-S2	D184M Endo-S2 mutant-mediated transglycosylation of azido-GlcNAc to Fc Asn297 (Zhang et al., 2022)	• One-step glycan remodeling • Homogeneous DAR2 conjugates
LacNAc	Endo-S2-catalyzed incorporation of azido-LacNAc oxazoline followed by click chemistry (Shi et al., 2022)	• Fc-specific conjugation • Azido-GalNAc carrier installation

D Data Splits

In this section, we present the detailed statistics of molecular properties based on two data splitting strategies: random splitting and scaffold-based splitting, for all three subsets (train, validation, and test).

For the random splitting strategy, we observe that the molecular properties (including MW, XLogP, TPSA, HBD, HBA, and RotB) are almost identical across the training, validation, and test sets. This is because the random split does not impose any constraints on the distribution of properties between the subsets, resulting in very similar distributions for all three sets.

On the other hand, for the scaffold-based splitting strategy, which ensures non-overlapping scaffolds between subsets, the molecular properties exhibit some noticeable discrepancies. This is due to the inherent differences in physicochemical properties between the scaffolds, which leads to variations in the distributions across the training, validation, and test sets. These differences are more pronounced compared to the random split.

D.1 RANDOM SPLITS

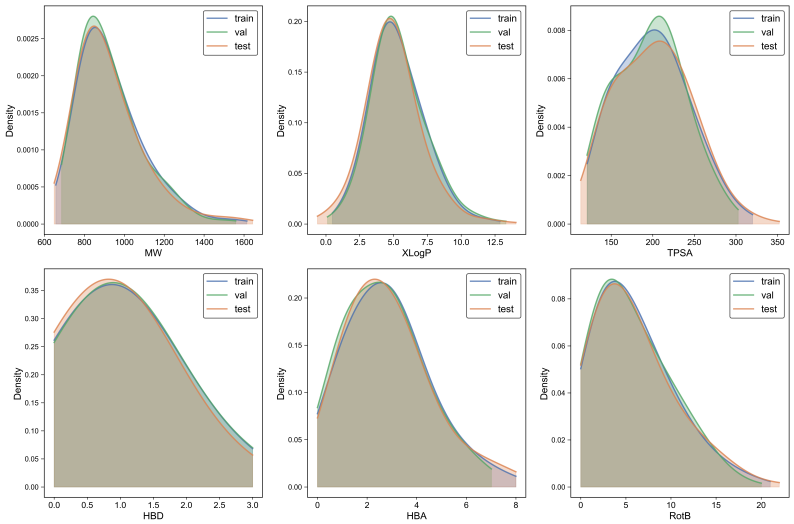


Figure 8: Distribution of physicochemical properties for the PROTAC subset based on the random splitting strategy into train, validation, and test sets. The properties are shown for each subset with distinct color coding, and the KDE plots with filled areas represent the distribution patterns across the subsets.

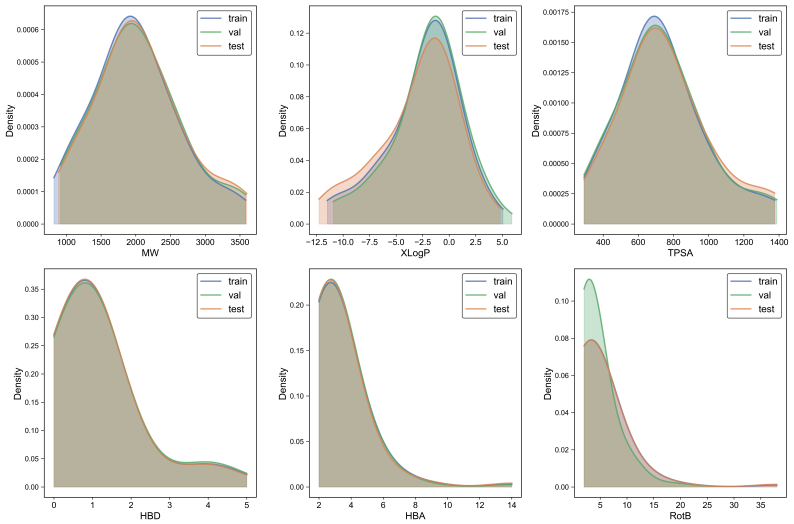


Figure 9: Distribution of physicochemical properties for the PDC subset based on the random splitting strategy into train, validation, and test sets. The properties are shown for each subset with distinct color coding, and the KDE plots with filled areas represent the distribution patterns across the subsets.

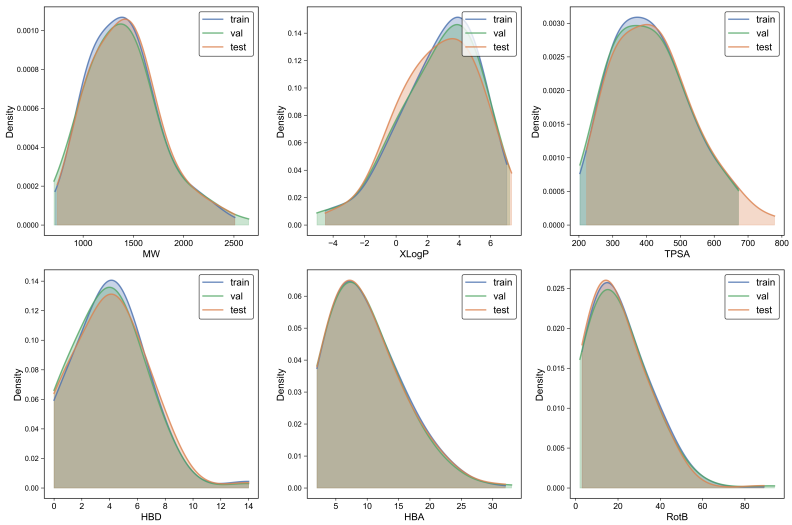


Figure 10: Distribution of physicochemical properties for the ADC subset based on the random splitting strategy into train, validation, and test sets. The properties are shown for each subset with distinct color coding, and the KDE plots with filled areas represent the distribution patterns across the subsets.

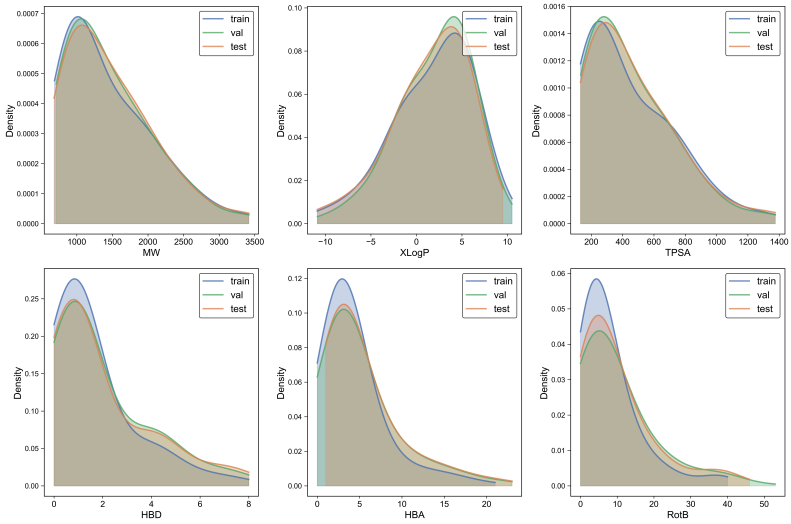


Figure 11: Distribution of physicochemical properties for the HBDrug3D dataset based on the random splitting strategy into train, validation, and test sets. The properties are shown for each subset with distinct color coding, and the KDE plots with filled areas represent the distribution patterns across the subsets.

D.2 SCAFFOLD SPLITS

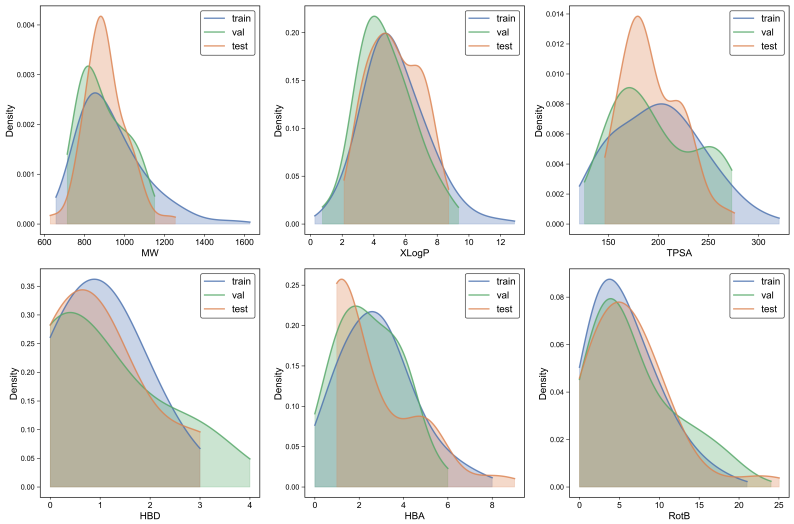


Figure 12: Distribution of physicochemical properties for the PROTAC subset based on the scaffold splitting strategy into train, validation, and test sets. The properties are shown for each subset with distinct color coding, and the KDE plots with filled areas represent the distribution patterns across the subsets.

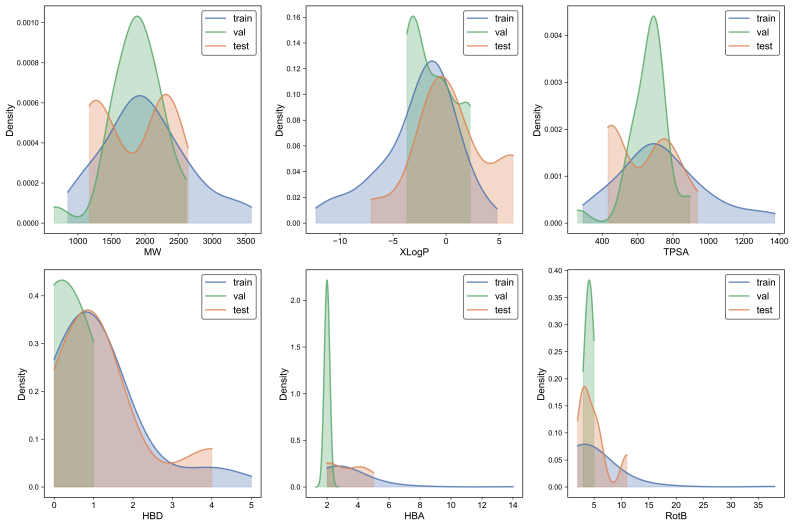


Figure 13: Distribution of physicochemical properties for the PDC subset based on the scaffold splitting strategy into train, validation, and test sets. The properties are shown for each subset with distinct color coding, and the KDE plots with filled areas represent the distribution patterns across the subsets.

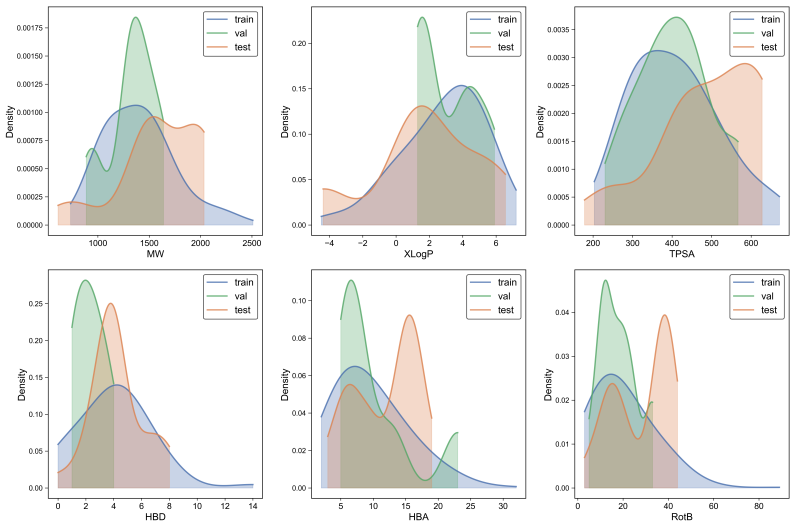


Figure 14: Distribution of physicochemical properties for the ADC subset based on the scaffold splitting strategy into train, validation, and test sets. The properties are shown for each subset with distinct color coding, and the KDE plots with filled areas represent the distribution patterns across the subsets.

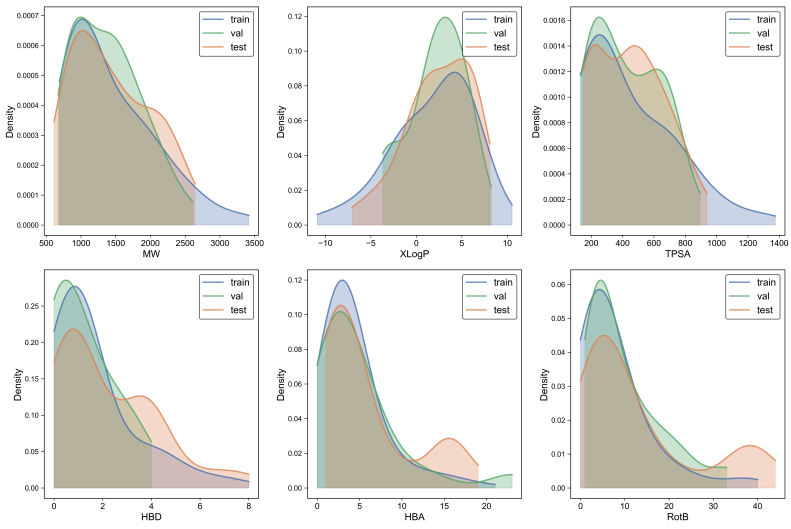


Figure 15: Distribution of physicochemical properties for the HBDrug3D dataset based on the scaffold splitting strategy into train, validation, and test sets. The properties are shown for each subset with distinct color coding, and the KDE plots with filled areas represent the distribution patterns across the subsets.

E Details of Benchmark Models

We summarize the implementation details of the four benchmark models. The code for each model was obtained from the respective open-source repositories: DeLinker (`https://github.`

com/oxpig/DeLinker), 3DLinker (https://github.com/YinanHuang/3DLinker), DiffLinker (https://github.com/igashov/DiffLinker), and LinkerNet (https://github.com/guanjq/LinkerNet). We only adjusted the max_num_atoms parameter to accommodate our datasets: PROTAC (160), PDC (256), ADC (163), and HBDrug3D (256). All models were trained with a batch size of 64. Regarding computational resources, DeLinker experiments were run on a single RTX 4090 GPU, DiffLinker required 4 RTX 4090 GPUs per experiment, and 3DLinker and LinkerNet each used one A100 GPU per experiment.

F Benchmark Results on Random Splits

Table 6: **Evaluation of generated molecules using general metrics under the random splitting strategy.** For each case, each model generated 100 molecules, and the reported values are the mean ± standard deviation across three independent runs. SC_{RDKit} is reported with three decimal places; the standard deviation is not explicitly shown because rounding to two decimal places would result in zeros.

Dataset	Model	Valid (↑)	Unique (↑)	Novel (↑)	Recover (↑)	2D filters (↑)	SC_{RDKit} (↑)
PROTAC	DeLinker	86.98 ± 0.02	40.50 ± 2.03	63.13 ± 2.96	**80.00 ± 4.47**	**86.60 ± 0.00**	0.261
	3DLinker	53.85 ± 0.23	**90.22 ± 0.45**	**85.35 ± 0.62**	13.33 ± 2.58	52.95 ± 0.25	0.873
	DiffLinker	56.94 ± 0.25	84.36 ± 1.33	70.65 ± 1.43	23.32 ± 1.46	44.82 ± 0.12	**0.893**
	LinkerNet	**91.69 ± 0.11**	33.26 ± 2.08	43.21 ± 1.58	76.42 ± 3.57	81.57 ± 0.11	0.502
PDC	DeLinker	51.94 ± 0.12	27.12 ± 0.46	64.54 ± 2.66	**100.0 ± 0.00**	50.63 ± 0.11	0.133
	3DLinker	72.16 ± 0.12	**61.89 ± 0.62**	**78.74 ± 0.14**	**100.0 ± 0.00**	**65.43 ± 0.17**	0.961
	DiffLinker	60.29 ± 0.16	29.54 ± 0.41	20.05 ± 0.24	78.48 ± 2.47	29.35 ± 0.08	**0.989**
	LinkerNet	**91.68 ± 0.07**	10.41 ± 0.55	9.51 ± 0.57	83.62 ± 1.23	49.31 ± 0.02	0.715
ADC	DeLinker	20.84 ± 0.13	90.32 ± 0.12	100.0 ± 0.00	**14.78 ± 0.25**	**79.74 ± 0.25**	0.163
	3DLinker	28.18 ± 0.14	**100.0 ± 0.00**	100.0 ± 0.00	0.00 ± 0.00	60.34 ± 0.31	**0.824**
	DiffLinker	29.54 ± 0.05	**100.0 ± 0.00**	100.0 ± 0.00	0.00 ± 0.00	62.37 ± 0.08	0.711
	LinkerNet	**31.42 ± 0.13**	64.71 ± 0.42	100.0 ± 0.00	0.00 ± 0.00	73.37 ± 0.13	0.554
HBDrug3D	DeLinker	29.20 ± 0.06	72.15 ± 0.10	**96.24 ± 0.03**	14.08 ± 0.20	**63.85 ± 0.14**	0.070
	3DLinker	59.03 ± 0.08	**97.91 ± 0.02**	92.36 ± 0.07	4.56 ± 0.06	58.59 ± 0.09	**0.893**
	DiffLinker	60.44 ± 0.08	92.55 ± 0.81	74.17 ± 1.32	14.70 ± 1.08	45.39 ± 0.13	0.817
	LinkerNet	**76.89 ± 0.07**	50.93 ± 0.16	51.55 ± 0.79	**33.42 ± 0.72**	61.48 ± 0.06	0.616

Table 7: **3D geometric fidelity under the random splitting strategy.** For each case, 100 molecules were generated per model; values are mean ± standard deviation across three independent runs. Clash, Ring Flat., and PB_Val. are pass rates (higher is better); Min Bond, Max Bond, and Max Angle are relative ratios (closer to 1 is better). Bold indicates the best value within each dataset; columns in which all models tie are left unbolded.

Dataset	Model	Clash (↑)	Ring Flat. (↑)	Min Bond	Max Bond	Max Angle	PB_Val. (↑)
PROTAC	DeLinker	**98.86 ± 0.02**	100.0 ± 0.00	**0.97 ± 0.00**	**1.04 ± 0.00**	**1.06 ± 0.00**	**98.87 ± 0.02**
	3DLinker	21.93 ± 0.16	100.0 ± 0.00	0.29 ± 0.01	1.58 ± 0.07	1.88 ± 0.03	0.00 ± 0.00
	DiffLinker	94.76 ± 0.23	100.0 ± 0.00	0.95 ± 0.00	1.06 ± 0.00	1.10 ± 0.00	93.45 ± 0.35
	LinkerNet	84.74 ± 0.02	100.0 ± 0.00	0.84 ± 0.00	2.76 ± 0.03	2.35 ± 0.02	28.37 ± 0.26
PDC	DeLinker	75.28 ± 0.10	100.0 ± 0.00	**0.95 ± 0.00**	**1.04 ± 0.00**	**1.07 ± 0.00**	75.42 ± 0.07
	3DLinker	5.26 ± 0.07	100.0 ± 0.00	0.27 ± 0.01	1.91 ± 0.04	1.89 ± 0.02	0.00 ± 0.00
	DiffLinker	**98.28 ± 0.12**	100.0 ± 0.00	0.91 ± 0.00	1.06 ± 0.00	1.08 ± 0.00	**98.28 ± 0.12**
	LinkerNet	47.51 ± 0.02	100.0 ± 0.00	0.84 ± 0.00	6.43 ± 0.03	4.06 ± 0.02	5.7 ± 0.11
ADC	DeLinker	74.86 ± 0.11	99.57 ± 0.00	**0.96 ± 0.00**	**1.05 ± 0.00**	1.09 ± 0.00	65.42 ± 0.25
	3DLinker	0.00 ± 0.00	99.82 ± 0.01	0.13 ± 0.01	7.76 ± 0.49	5.19 ± 0.33	0.00 ± 0.00
	DiffLinker	**83.33 ± 0.07**	**100.0 ± 0.00**	0.90 ± 0.00	**1.05 ± 0.00**	**1.08 ± 0.00**	**70.11 ± 0.48**
	LinkerNet	47.31 ± 0.14	99.21 ± 0.10	0.77 ± 0.00	4.89 ± 0.00	3.68 ± 0.00	6.98 ± 0.14
HBDrug3D	DeLinker	73.93 ± 0.09	**100.0 ± 0.00**	**0.96 ± 0.00**	**1.05 ± 0.00**	**1.08 ± 0.00**	55.20 ± 0.12
	3DLinker	0.38 ± 0.02	99.80 ± 0.00	0.16 ± 0.02	4.54 ± 0.62	3.32 ± 0.39	0.00 ± 0.00
	DiffLinker	**82.73 ± 0.15**	**100.0 ± 0.00**	0.94 ± 0.00	1.06 ± 0.00	1.12 ± 0.00	**63.59 ± 0.12**
	LinkerNet	49.36 ± 0.04	99.39 ± 0.00	0.84 ± 0.00	5.76 ± 0.00	3.86 ± 0.00	11.48 ± 0.36

Table 8: **Evaluation of generated molecules using physicochemical and linker-specific properties under the random splitting strategy.** For each case, each model generated 100 molecules, and the reported values are the mean ± standard deviation across three independent runs. Abbreviations: MW = molecular weight, XLogP = octanol–water partition coefficient, TPSA = topological polar surface area, HBD = hydrogen bond donor, HBA = hydrogen bond acceptor, RotB = number of rotatable bonds. Reference values correspond to the averages computed on the respective test sets.

Dataset	Model	MW	XLogP	TPSA	HBD	HBA	RotB
PROTAC	Reference	929	5.1	201.8	0.9	2.9	5.9
	DeLinker	952 ± 0.05	3.0 ± 0.00	198.1 ± 0.04	1.1 ± 0.00	2.6 ± 0.00	6.3 ± 0.01
	3DLinker	929 ± 1.15	2.9 ± 0.02	190.4 ± 0.04	1.3 ± 0.01	2.6 ± 0.01	5.2 ± 0.01
	DiffLinker	928 ± 0.39	4.1 ± 0.00	200.5 ± 0.07	1.0 ± 0.00	2.7 ± 0.00	6.1 ± 0.00
	LinkerNet	926 ± 0.18	4.2 ± 0.00	198.5 ± 0.02	1.0 ± 0.00	2.4 ± 0.00	6.1 ± 0.00
PDC	Reference	2046	-3.0	759.8	1.1	3.3	4.6
	DeLinker	1838 ± 1.36	-4.3 ± 0.01	653.6 ± 0.37	1.1 ± 0.00	2.7 ± 0.01	3.9 ± 0.02
	3DLinker	1862 ± 0.31	-4.8 ± 0.00	662.5 ± 0.22	1.2 ± 0.01	3.2 ± 0.01	3.1 ± 0.02
	DiffLinker	1928 ± 1.56	-7.0 ± 0.01	723.8 ± 0.08	0.9 ± 0.00	2.7 ± 0.00	3.0 ± 0.01
	LinkerNet	2035 ± 0.24	-7.4 ± 0.00	762.0 ± 0.16	0.9 ± 0.00	2.9 ± 0.01	3.9 ± 0.01
ADC	Reference	1419	2.6	416.4	4.0	9.8	19.7
	DeLinker	1283 ± 2.32	2.9 ± 0.02	365.4 ± 0.78	4.1 ± 0.02	7.4 ± 0.04	9.4 ± 0.06
	3DLinker	1269 ± 1.82	2.5 ± 0.02	368.5 ± 0.69	3.8 ± 0.01	8.7 ± 0.03	10.9 ± 0.04
	DiffLinker	1456 ± 2.34	2.6 ± 0.02	431.7 ± 0.23	4.3 ± 0.02	9.4 ± 0.09	14.3 ± 0.17
	LinkerNet	1402 ± 0.71	2.4 ± 0.03	397.2 ± 0.16	4.2 ± 0.01	7.9 ± 0.01	10.5 ± 0.01
HBDrug3D	Reference	1465	1.6	459.4	2.0	5.3	10.1
	DeLinker	1416 ± 1.23	2.5 ± 0.02	442.7 ± 0.61	2.1 ± 0.01	3.4 ± 0.01	3.8 ± 0.02
	3DLinker	1365 ± 0.96	2.4 ± 0.01	409.8 ± 0.45	1.4 ± 0.00	3.7 ± 0.01	4.4 ± 0.01
	DiffLinker	1387 ± 0.47	2.5 ± 0.01	408.6 ± 0.43	1.4 ± 0.00	3.3 ± 0.00	3.7 ± 0.00
	LinkerNet	1440 ± 0.52	2.3 ± 0.01	463.4 ± 0.29	1.9 ± 0.00	3.8 ± 0.01	5.9 ± 0.01