Uni-Bond: Learning Chemical Bonds from Atomic Coordinates

Stepan Pavlenko [*1] **Pavel Maslov** [*2] **Ivan Burov** [*3] **Artem Tsypin** [4] **Alexander Telepov** [4]
Konstantin Ushenin [4 5 6] **Kuzma Khrabrov** [4] **Denis Potapov** [4] **Artur Kadurin** [4]

Abstract

Recovering a molecular graph from 3D atomic coordinates is a fundamental step in computational chemistry, linking continuous geometry to the discrete graphs on which chemoinformatics and property prediction depend. Rule-based bond perception is brittle to geometric distortion and strained conformations, motivating learned alternatives. We introduce Uni-Bond, which predicts typed bonds directly from atom types and Cartesian coordinates by combining a pretrained Uni-Mol encoder with a lightweight pairwise classification head. In our final architecture, the two atomic embeddings associated with each candidate bond are combined using a symmetric pooling operation and concatenated with the Uni-Mol pair representation before classification. On the GEOM scaffold split, Uni-Bond reduces the exact-match error rate by roughly 20× relative to the strongest baseline, and transfers zero-shot to water clusters, peptide-like SPICE molecules, and molecular dimers, reliably distinguishing covalent bonds from close non-covalent contacts. We further study direct geometry-to-SMILES generation with a coordinate-tokenized language model: it performs well on GEOM but collapses under chemical domain shift, indicating that structured geometric foundation models offer a more reliable route from 3D coordinates to molecular graphs.

1. Introduction

The molecular graph is the central abstraction for representing small molecules in machine learning: atoms define nodes, chemical bonds define typed edges, and this discrete structure can be equivalently written as a line notation such as SMILES (Weininger, 1988; Wu et al., 2018). This graph-based representation underlies a broad range of molecular learning pipelines, including property prediction from SMILES strings (Goh et al., 2017; Francoeur & Koes, 2021), large-scale chemical language modeling and molecular representation learning (Ross et al., 2022), and template-free reaction prediction from molecular sequences (Schwaller et al., 2019). Molecules, however, are not fully described by their graph alone: their physical, chemical, and biological behavior is shaped by three-dimensional conformation and spatial arrangement (Zhou et al., 2023). This creates a fundamental representational gap between continuous 3D geometry and discrete graph structure, and motivates a central question: how can a chemically meaningful molecular graph be recovered directly from atomic coordinates?

This geometry-to-graph mapping problem is nontrivial because molecular topology is difficult to recover using heuristic rules based on coordinate information. In reactive molecular dynamics, thermal fluctuations and transient bond stretching can make static distance-based edge definitions unreliable (Hutchings et al., 2020); in structural biology, noise and resolution limits can complicate ligand connectivity assignment in Cryo-EM maps (Muenks et al., 2023); and in quantum chemical datasets, tautomerism and context-dependent interpretation can affect downstream property prediction (Vazquez-Salazar et al., 2022). This problem is therefore better viewed as a global structure reconstruction task that requires integrating information across the entire molecular geometry, rather than relying solely on local heuristic rules or pairwise distance-based methods.

The need for robust geometry-to-graph reconstruction has become especially pressing with the rise of structure-based *de novo* molecular design. Recent 3D molecular generative models operate directly on atomic coordinates: E(3)-equivariant diffusion processes generate atom types and Cartesian coordinates jointly (Hoogeboom et al., 2022; Le et al., 2023), some methods couple discrete molecular topology with continuous 3D geometry (Vignac et al., 2023; Hua et al., 2024), and others represent molecules through voxelized 3D density grids (Ragoza et al.; Pinheiro et al., 2023). A central application is structure-based drug design, where generative models build 3D ligand structures directly inside protein binding pockets conditioned on re-

[1]HSE University [2]Moscow Center for Advanced Studies [3]Lomonosov Moscow State University [4]AXXX [5]Tomsk State University [6]Ural Federal University. Correspondence to: Stepan Pavlenko <spavlenko@hse.ru>.

The Workshop on Graph Foundation Models at the 43rd International Conference on Machine Learning (ICML 2026), Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

ceptor geometry (Ragoza et al., 2022; Guan et al., 2023; Schneuing et al., 2024; Zholus et al., 2024). Yet the downstream ecosystem operates almost exclusively on molecular graphs and SMILES: chemoinformatics software, property predictors, reaction prediction systems, and molecular databases all assume an explicit graph representation as input (O'Boyle et al., 2011; Zhang et al., 2012; Kadukova & Grudinin, 2016). Converting generated or simulated 3D structures into typed molecular graphs is therefore a critical interface between geometric generative models and graph-based molecular learning workflows. The atom–bond inconsistency problem identified in recent 3D diffusion work makes this gap concrete: distance-based bond perception algorithms (e.g., RDKit's `rdDetermineBonds` and OpenBabel's bond perception routines) compare interatomic distances against covalent-radius thresholds, and small coordinate errors can push them into incorrect bond orders or sanitization failures even for chemically plausible structures (Peng et al., 2023). This issue already appears in geometry-complete 3D diffusion models such as GCDM (Morehead & Cheng, 2024), where atom types and coordinates are generated directly but the molecular graph is subsequently reconstructed by heuristic bond-perception rules. This motivates learned bond reconstruction methods that infer connectivity from data-driven geometric representations rather than hand-crafted distance thresholds.

To address this, we build on Uni-Mol (Zhou et al., 2023), a graph foundation model for 3D molecules pretrained on roughly 209M conformations with explicit hydrogens. Architecturally, Uni-Mol is a graph Transformer that operates on a fully connected molecular graph and maintains both per-node atomic representations and per-edge pairwise representations, with spatial self-attention exchanging information through SE(3)-invariant distance features. The Uni-Mol backbone has been successfully transferred to a wide range of downstream problems, such as molecular property prediction (Potapov et al., 2026; Wang et al., 2024; Gao et al., 2023), protein–ligand binding pose prediction (Fan et al., 2024; Alcaide et al., 2025), and reaction modeling (Wu et al., 2025). Bond reconstruction provides a particularly natural application of this model, since the dual atomic/pairwise representations align directly with the per-atom and per-edge predictions required by the task.

Building on this, we propose Uni-Bond, a bond reconstruction model that predicts bond orders for every atom pair from contextualized Uni-Mol embeddings. Uni-Bond outperforms both rule-based and GNN-based baselines on in-distribution GEOM reconstruction, and remains accurate when transferred zero-shot to out-of-distribution chemical systems including biomolecules, water clusters, and molecular dimers — illustrating that strong molecular foundation models can transfer effectively to graph-reconstruction tasks not seen during pre-training. As a complementary study, we examine direct geometry-to-SMILES generation with a coordinate-tokenized language model. This formulation is appealing in principle, since it bypasses graph sanitization and can represent annotations difficult to recover from a bond-order matrix alone, such as charges and stereochemistry. In practice, however, we find that it generalizes poorly under chemical domain shift, despite strong in-distribution performance. Together, these results position transfer from a structured graph foundation model as the more robust route from 3D geometry to molecular graphs, and identify direct coordinate-tokenized generation as expressive but currently less reliable.

2. Related Work

2.1. Molecular Graph Reconstruction from 3D Coordinates

While sequence modeling has advanced significantly, interpreting raw 3D geometries presents a distinct challenge. Historically, this task has relied on rule-based chemoinformatics toolkits. Algorithms such as `xyz2mol` (Kim & Kim, 2015), widely deployed in toolkits like RDKit, infer connectivity by comparing interatomic distances against a look-up table of covalent radii. While computationally efficient, these deterministic rules are inherently brittle; they lack the capacity to handle geometric noise, steric strain, or dynamic fluctuations common in generated or experimental data.

Deep learning has been utilized to address these limitations. The recent state-of-the-art framework, **YuelBond** (Wang & Dokholyan, 2026), treats bond reconstruction as a supervised edge classification task. YuelBond employs a Graph Neural Network (GNN) to encode spatial relationships, significantly outperforming rule-based methods on distorted geometries. However, YuelBond, like most standard GNNs, relies on iterative message passing between neighboring atoms. This architectural choice imposes a local receptive field, effectively limiting the model's ability to capture long-range dependencies. Consequently, resolving ambiguous bond orders in complex conjugated systems where the global chemical context (e.g., aromaticity or tautomeric state) defines the local bond type remains a challenge for purely local GNNs. Moreover, YuelBond is trained and evaluated only on the lowest-energy GEOM conformers and their artificially perturbed versions. Therefore, it may not fully capture the variability of molecular geometry across different conformations of the same molecule. This can limit its robustness when the input structure differs substantially from the training distribution.

2.2. Large Language Models in Chemistry

The remarkable success of Large Language Models (LLMs) in natural language processing has motivated their rapid

adoption in chemistry, where molecules can be treated as a form of language through linear representations such as SMILES (Weininger, 1988), SELFIES (Krenn et al., 2020), group SELFIES (Cheng et al., 2023), and SAFE (Noutahi et al., 2023). This "molecules as language" perspective enables the direct application of sequence modeling architectures to molecular tasks. Early chemical language models, including ChemBERTa, X-MOL, MolFormer and Chemformer, showed that large-scale pre-training can yield transferable molecular representations for property prediction, reaction modeling, and molecular optimization (Chithrananda et al., 2020; Xue et al., 2022; Ross et al., 2022; Irwin et al., 2022). In parallel, generative models such as MolGPT and LigGPT showed that language models can be used to design new molecules, including those tailored to specific structures or properties (Bagal et al., 2022; 2021).

More recent models have extended this paradigm beyond single-modality SMILES modeling by aligning molecular strings with graphs, 2D and 3D structures, images, and natural-language descriptions, enabling tasks such as molecule captioning, text-guided generation, and cross-modal retrieval (Edwards et al., 2022; Liu et al., 2023; 2024; Tang et al., 2024; Fang et al., 2024). Overall, these developments position chemical LLMs as a complementary alternative to handcrafted fingerprints and message-passing neural networks, combining scalable pre-training with controllable molecular generation and optimization (Rogers & Hahn, 2010; Gilmer et al., 2017).

2.3. Integrating 3D Structures with LLMs

Bridging the gap between 3D Euclidean space and the tokenized latent space of LLMs is an active frontier in molecular machine learning. BindGPT (Zholus et al., 2024) demonstrated that a single autoregressive language model can jointly handle structural SMILES and tokenized XYZ coordinates, supporting tasks such as 3D molecule generation and conformer generation. Nach0-pc (Kuznetsov et al., 2024) introduced a dedicated molecular point-cloud encoder coupled to a text-based language model to support a range of multimodal generative tasks involving 3D structures. Building on this line of work, we adopt a coordinate-tokenization strategy similar in spirit to BindGPT's and apply it to a different question: whether a fine-tuned language model can directly recover SMILES from atom identities and 3D coordinates alone, providing a complementary route to structured bond-order classification.

3. Methods

3.1. Problem Setup

We formulate the task as mapping a set of 3D atomic coordinates to a discrete molecular graph. The input is a molecule defined by its atomic numbers $\mathcal{Z} = \{z_1, \ldots, z_N\}$ and 3D coordinates $\mathcal{R} \in \mathbb{R}^{N\times 3}$. The objective is to predict the connectivity matrix $\mathcal{C}$, whose entry c_{ij} represents the bond order between atoms i and j. We treat connectivity as undirected, so $\mathcal{C}$ is symmetric: $c_{ij} = c_{ji}$.

We explore two distinct conceptual approaches to this problem:

1. **Local, Pairwise Prediction:** This approach decomposes the problem into a set of independent classification tasks, where the bond type for each atom pair (i, j) is predicted from local geometric features. The final molecular graph is then assembled from these individual predictions. Once the bond network is recovered, a SMILES string can be obtained in a standard post-processing step using RDKit (Landrum, 2013) (e.g., by building an `rdkit.Chem.RWMol` from the predicted bonds and calling `rdkit.Chem.MolToSmiles` after sanitization).

2. **Global, Sequence-Level Inference:** In this setting, the molecule is represented only by atom types and tokenized 3D coordinates, without an explicit molecular graph or bond labels. Instead of predicting a bond label for each candidate pair (i, j) and assembling a graph, we condition a pretrained language model on this coordinate-based input and directly generate the SMILES string as a sequence.

3.2. Dataset Construction

For our experiments, we utilized the **Geometric Ensemble Of Molecules (GEOM)** dataset (Axelrod & Gómez-Bombarelli, 2022), which contains conformers for 133,000 distinct molecule samples from QM9 and 317,000 samples with experimental data related to biophysics, physiology, and physical chemistry. We developed an automated pipeline to construct a dataset suitable for training. All molecules were represented with explicit hydrogen atoms, so the model was trained and evaluated on full molecular geometries rather than heavy-atom-only graphs. For each unique molecule identified by its SMILES, we extracted its lowest-energy conformer along with five additional conformers randomly sampled from the conformational ensembles provided within the GEOM dataset. This process yielded approximately 2.2 million conformers from ∼437,000 unique molecules.

The reference bond labels were inherited from the molecular graph encoded by the source SMILES associated with each GEOM molecule, rather than being re-determined from every optimized geometry using rule-based bond perception. This defines the task as recovering the intended molecular topology from a 3D conformer, rather

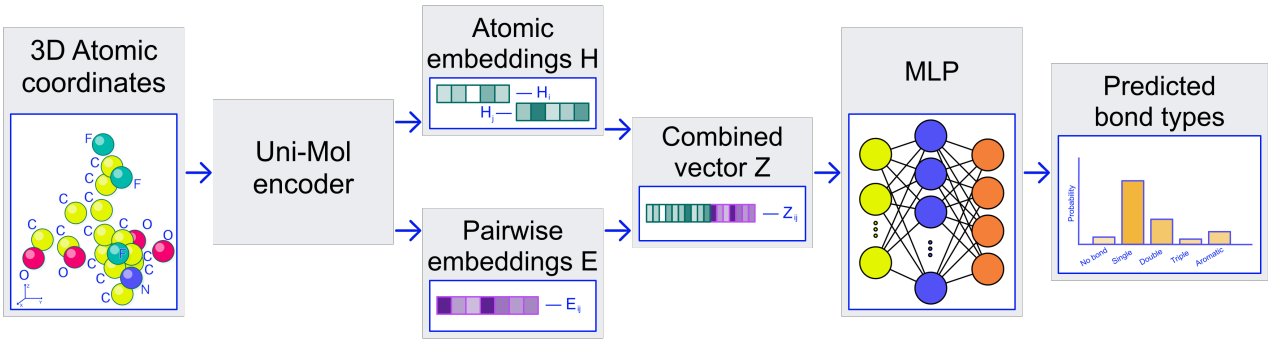

Figure 1. Architecture of Uni-Bond. The model extracts atomic and pairwise embeddings from the Uni-Mol geometric encoder. For each candidate atom pair, the two atomic embeddings are first combined using symmetric pooling, and the resulting pooled atomic representation is concatenated with the corresponding Uni-Mol pair embedding before being processed by an MLP classification head.

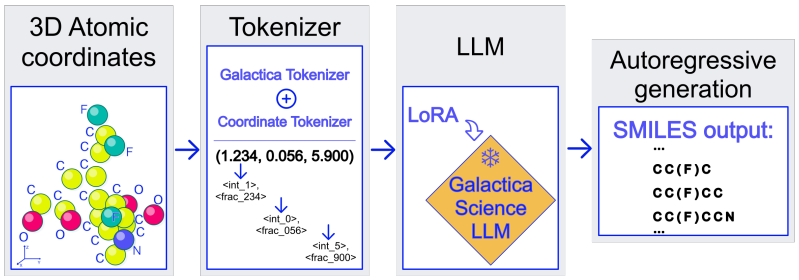

Figure 2. Direct SMILES generation pipeline with the BindGPT-style coordinate tokenizer. An XYZ-style molecular structure is converted into coordinate-aware tokens by splitting each Cartesian coordinate into signed integer and three-digit fractional parts, then processed by a LoRA-adapted Galactica language model to autoregressively generate the target SMILES string.

than reproducing a geometry-dependent rule-based assignment. This choice is supported on two grounds. First, the GEOM conformers were generated using CREST (Pracht et al., 2020), which samples a conformer–rotamer ensemble through metadynamics-driven exploration at the GFN2-xTB level (Grimme, 2019) followed by geometry optimization. The procedure is explicitly designed to discover low-energy conformers and rotamers of a fixed molecular graph, not to explore distinct bonding patterns or constitutional isomers, so a single reference graph is the appropriate shared label across conformers of the same molecule. Second, using source-SMILES labels avoids circularity when comparing against rule-based baselines such as RDKit: if the labels were themselves derived from RDKit-like rules, the learned model would partly be evaluated on its ability to reproduce those rules rather than the true molecular graph. Under this setup, a prediction is considered correct if it matches the source-SMILES graph for the corresponding conformer, and rare cases in which relaxation or graph re-identification changes the apparent topology are treated as label noise.

Since our models predict explicit bond types, an additional step is required to reconstruct valid SMILES strings from the resulting molecular graphs for verification. The source SMILES provided in GEOM use aromatic bond notation rather than the kekulized form, so aromatic systems are represented by a dedicated aromatic bond type instead of alternating single and double bonds. Consequently, our reference bond labels include aromatic as one of five classes — no-bond, single, double, triple, and aromatic — and our training targets correspond to molecular graphs in the same aromatic representation. Given a labeled molecular graph (atom types and typed bonds), we need to verify that it can be round-tripped into a canonical SMILES string using RDKit. We do this by constructing an `RWMol` from the atoms and bonds, applying `Chem.SanitizeMol` for valence and aromaticity perception, and serializing with `Chem.MolToSmiles`. For most molecules this reconstruction is straightforward, but two well-known limitations make it unreliable for a subset of cases. First, formal charges and bond orders are co-dependent, and ambiguous formal-charge assignments are possible for a given connectivity, with no guarantee that perceived charges match the true molecular charge (Welsh & Allison, 2019). Second, the underlying procedure is formulated for closed-shell systems and does not handle unpaired electrons in a principled way. We therefore removed molecules carrying a non-zero formal charge or an open-shell electronic structure from our dataset,

amounting to approximately 10% of the total. Restricting training data to neutral closed-shell molecules is a common choice in the field of AI for chemistry (Kovács et al., 2025; Smith et al., 2017). The final dataset consists of molecules for which this reconstruction step succeeded, yielding a valid SMILES string from each ground-truth graph.

To ensure a comprehensive evaluation of the model's generalization capabilities, we used an 80/10/10 scaffold split based on Bemis–Murcko scaffolds. In this setting, molecules are grouped according to their core chemical scaffolds, and the training, validation, and test sets contain different scaffold groups. This provides a stringent evaluation protocol, since the model must make predictions for molecules with previously unseen chemical core structures rather than merely interpolating between closely related molecules.

3.3. Baselines

To evaluate the effectiveness of our approach, we compare three distinct methods for bond reconstruction, ranging from rule-based heuristics to advanced deep learning frameworks.

First, as a classical heuristic baseline, we employ a **Rule-based predictor**, using RDKit's `rdDetermineBonds` module (Kim & Kim, 2015). This rule-based approach infers molecular connectivity from 3D coordinates using heuristic chemical rules. The resulting bond graph is then used for evaluation and SMILES generation.

Second, we compare against **YuelBond** (Wang & Dokholyan, 2026), a supervised GNN-based bond reconstruction model. In comparison with the rule-based methods, YuelBond learns bond perception directly from molecular examples. Given atomic types and 3D coordinates, it first constructs a candidate graph of atom pairs and represents each pair by geometric edge features, such as interatomic distance. A message-passing GNN then iteratively updates atom and edge representations, allowing each candidate edge to incorporate information from its local chemical environment. Finally, an edge classifier predicts the bond type for each candidate atom pair, including the no-bond class. The predicted bond orders are then used to reconstruct the molecular graph for evaluation and SMILES generation. Because the original YuelBond setup does not use explicit hydrogen atoms, we retrain the YuelBond baseline from scratch on our explicit-hydrogen GEOM split to ensure a fair comparison with Uni-Bond; additional details are provided in Appendix D.1.

3.4. Uni-Bond: Uni-Mol-Based Bond Predictor

As introduced, our primary geometric pipeline, **Uni-Bond**, relies on the pretrained Uni-Mol encoder to process raw 3D coordinates. For each molecule containing N atoms, the encoder generates two distinct types of embeddings that are critical for our bond reconstruction task:

1. **Atomic Representations** ($\mathbf{H} \in \mathbb{R}^{N \times D_{\text{atomic}}}$): These vectors encode the local chemical environment and geometric context of each individual atom.

2. **Pair Representations** ($\mathbf{E} \in \mathbb{R}^{N \times N \times D_{\text{pair}}}$): These tensors explicitly capture the spatial and chemical interactions between every pair of atoms (i, j), providing the detailed relational information necessary for determining bond existence and order.

To improve computational efficiency and ensure the physical validity of the reconstructed graphs, we apply two critical constraints to the pair representations before classification. First, we implement a geometric cutoff, retaining only candidate pairs (i, j) with an interatomic distance d_{ij} such that $d_{ij} \leq r_c$, where $r_c = 3.0$ Å in our experiments. This threshold safely encompasses all plausible covalent interactions while significantly reducing the computational complexity.

Second, to inherently enforce the undirected nature of chemical bonds, we restrict our candidate set exclusively to the upper triangular indices of the distance matrix (i.e., $i < j$). Because a standard concatenation-based classifier could potentially output asymmetric predictions for (i, j) and (j, i) due to the ordering of the input vectors, querying the model only once per unique atom pair mathematically guarantees a symmetric connectivity matrix. This avoids any contradictory predictions without requiring post-hoc probability averaging.

For the final bond classification, a lightweight MLP head is integrated on top of the Uni-Mol encoder (Figure 1). For each candidate atom pair (i, j), we combine the corresponding atomic embeddings Hi and Hj using a symmetric pooling operation and concatenate the pooled atomic representation with the pairwise interaction embedding Eij. This pooled pair representation is then passed to an MLP head to predict the bond type. We train Uni-Bond in two configurations: a frozen-encoder variant where only the MLP head is updated, and a full fine-tuning variant where the Uni-Mol backbone is updated jointly with the head.

We formulate bond reconstruction as a multi-class edge-classification problem trained with a standard cross-entropy objective. During inference, bond types are assigned greedily using the maximum predicted logit, and the resulting predictions are assembled into an undirected connectivity matrix. The reconstructed topology is subsequently converted into a molecular graph and sanitized using RDKit. Additional implementation details, including the MLP architecture, candidate-pair selection, loss formulation, and decoding constraints, are provided in Appendix A.

Table 1. Overall bond reconstruction on GEOM. Macro F1 is computed over covalent bond classes, excluding the no-bond class.

Model	Macro F1	Exact Match
RDKit	0.973	0.899
YuelBond	0.975	0.832
Uni-Bond (frozen)	0.9948	0.9718
Uni-Bond	**0.9997**	**0.9953**
LLM	–	0.951

3.5. Direct SMILES Generation with LLMs

In addition to graph-level bond prediction, we consider an autoregressive language-model formulation that maps molecular 3D geometry directly to a SMILES string. This formulation targets a richer molecular representation than the bond-order matrix predicted by Uni-Bond. As discussed in Section 3.2, charged and open-shell molecules create difficulties for graph-to-SMILES reconstruction because bond-order assignment, formal-charge assignment, and radical-electron assignment are coupled during sanitization. Direct SMILES generation avoids this intermediate graph-sanitization step by learning to predict the source molecular string directly from atom identities and coordinates.

Stereochemistry presents a separate limitation of the graph-reconstruction formulation. Uni-Bond predicts only bond classes and therefore does not directly output tetrahedral chirality or directional double-bond stereochemistry, which are represented explicitly in isomeric SMILES. Such information would require an additional stereochemical perception step from the reconstructed graph and 3D geometry. We therefore evaluate the LLM formulation as a complementary approach that can, in principle, generate stereochemical and charged SMILES annotations directly. In Appendix C.3, we also report direct-SMILES experiments on both filtered and unfiltered GEOM data (see Appendix C.1 for the definition of filtered/unfiltered GEOM data).

LLM input representation and fine-tuning. For each input, the 3D molecular geometry is serialized in an XYZ-style format, where each atom is represented by its atomic symbol followed by three Cartesian coordinates. This sequence is then tokenized into a flat token stream that conditions the language model on the molecular structure (Figure 2; see Appendix C.2 for the exact prompt schema). We consider two coordinate tokenization strategies. The first follows the BindGPT-style (Zholus et al., 2024) representation, where each Cartesian coordinate is represented by two tokens: one token for the signed integer part and one token for the three-digit fractional part. This gives six coordinate tokens per atom. The second uses a compact binned representation, where each Cartesian coordinate is discretized into one of 1024 bins over the interval $[-12, 12]$ Å, resulting in three coordinate tokens per atom.

For both representations, we extend the Galactica tokenizer with the corresponding coordinate vocabulary. The model is fine-tuned using LoRA adapters (Hu et al., 2022) applied to all linear layers, while newly introduced coordinate-token embeddings are kept trainable together with the LoRA parameters. We use random conformer selection during training and apply target-side SMILES augmentation with probability 0.5 using kekulized, rooted, randomized, and cycle-renumbered SMILES variants (Ganeeva et al., 2024). At inference time, decoding is performed greedily with a maximum of 256 generated tokens. See Appendix C.1 for training hyperparameters.

Training objective. Let $\mathbf{x}$ denote the prompt tokens constructed from atom identities and tokenized 3D coordinates, and let $\mathbf{s} = (s_1, \ldots, s_T)$ denote the target answer sequence, including the `[START_SMILES]` and `[END_SMILES]` tokens. The LLM is trained by maximizing the conditional likelihood of the SMILES sequence given the coordinate prompt:

$$p_\theta(\mathbf{s} \mid \mathbf{x}) = \prod_{t=1}^{T} p_\theta(s_t \mid \mathbf{x}, s_{<t}). \tag{1}$$

Equivalently, we minimize the autoregressive cross-entropy loss

$$\mathcal{L}_{\text{LLM}} = -\frac{1}{T} \sum_{t=1}^{T} \log p_\theta(s_t \mid \mathbf{x}, s_{<t}). \tag{2}$$

The loss is computed only over the answer tokens; all coordinate-prompt tokens are masked out. During training, one conformer is sampled from the stored conformer tensor for each molecule, and the target SMILES may be replaced by an augmented equivalent representation. This objective therefore trains the model to recover a molecular string from atom types and 3D coordinates without requiring intermediate bond labels or graph sanitization.

4. Experiments

We evaluated the performance of several architectures on the geometry-to-graph task, using the scaffold split of the GEOM dataset (Section 3.2) as the primary benchmark to assess generalization to novel chemical structures. The models compared include: (1) a rule-based baseline – RDKit (Landrum, 2013), (2) a GNN baseline – YuelBond (Wang & Dokholyan, 2026), (3) Uni-Bond, (4) an LLM-based approach – Galactica (Taylor et al., 2022).

4.1. In-Domain Evaluation on GEOM

Table 1 reports the overall bond-reconstruction performance on the GEOM test set, where F1 is computed as a macro

Table 2. Out-of-Domain (OOD) generalization comparison between the Uni-Bond predictor and RDKit. For Uni-Bond, F1 is macro-averaged over covalent bond classes excluding the no-bond class; for Water Clusters, only the single-bond class is present. Uni-Bond was trained exclusively on the GEOM dataset and evaluated zero-shot on structurally disparate chemical systems. **Bold** indicates the best result within each dataset and metric; ties are marked consistently.

OOD Dataset	Model	Macro F1	Exact Match
Water Clusters	RDKit	**1.0000**	**1.0000**
	Uni-Bond	**1.0000**	0.9880
SPICE test	RDKit	**0.9964**	**0.9723**
	Uni-Bond	0.9896	0.9307
Dimers	RDKit	0.8755	0.8606
	Uni-Bond	**0.9183**	**0.8949**

average over covalent bond classes, excluding the no-bond class. Table 3 further decomposes these results by bond order, showing separate performance for single, double, triple, aromatic and no-bond classes.

Exact Match measures the fraction of molecules for which the entire predicted bond topology matches the reference graph, i.e., all atom-pair bond assignments are correct. Uni-Bond achieves a near-perfect Exact Match of 99.53% after full fine-tuning of the Uni-Mol backbone with symmetric pooling of atomic embeddings. This approach substantially outperforms the GNN baseline (YuelBond) and the traditional rule-based method (RDKit), which achieve 83.2% and 89.9% Exact Match, respectively. These results indicate that learned Uni-Mol representations provide highly accurate and consistent features for geometry-to-graph translation, while symmetric pooling preserves the permutation symmetry of atom-pair representations without degrading reconstruction accuracy.

In addition to graph-level bond prediction, we evaluate direct SMILES generation on the GEOM test set. This setting maps atom types and 3D coordinates directly to a molecular SMILES string using an autoregressive LLM, without explicitly constructing a molecular graph before decoding. We evaluate the BindGPT-style coordinate representation and the compact binned representation described in Section 3.5. Detailed per-task LLM results are reported in Appendix Table 6.

On GEOM, our best-performing model with BindGPT-style tokenization achieves strong in-domain performance: exact match is approximately 95%, validity exceeds 99.7%, and mean Morgan-fingerprint Tanimoto similarity is close to 0.99. We provide performance metrics of LLM models with different tokenization strategies trained on both filtered and unfiltered GEOM data in Appendix C.3.

Table 3. Per-bond-type metrics on the GEOM dataset. **Bold** indicates the best result; *italic* indicates the second-best. Ties are marked consistently.

Bond Type	Model	F1	Precision	Recall
No-bond	RDKit	0.994	0.988	**1.000**
	YuelBond	**1.000**	**1.000**	**1.000**
	Uni-Bond	**1.000**	**1.000**	**1.000**
Single	RDKit	0.990	*0.994*	0.985
	YuelBond	*0.994*	*0.994*	*0.995*
	Uni-Bond	**1.000**	**1.000**	**1.000**
Double	RDKit	0.932	0.924	0.939
	YuelBond	*0.977*	*0.976*	*0.977*
	Uni-Bond	**0.997**	**0.997**	**0.997**
Triple	RDKit	*0.976*	*0.988*	*0.964*
	YuelBond	0.940	*0.988*	0.896
	Uni-Bond	**0.994**	**0.995**	**0.993**
Aromatic	RDKit	*0.994*	**1.000**	*0.989*
	YuelBond	0.987	0.988	0.986
	Uni-Bond	**1.000**	*0.999*	**1.000**

4.2. Out-of-Domain Generalization of Bond Prediction

To rigorously assess the physical validity of our learned representations, we evaluated the fine-tuned Uni-Bond model on three out-of-domain (OOD) subsets from the well-established SPICE dataset (Eastman et al., 2023): SPICE test, Water Clusters, and Dimers. Because the model was exclusively trained on small, drug-like organic molecules from the GEOM dataset, these subsets allow us to test two distinct axes of zero-shot generalization: scaling to larger, unseen biomolecular classes, and handling disconnected, non-covalently bound systems. The evaluation results are summarized in Table 2.

First, we evaluated the model on the **SPICE test** subset to test generalization to novel, larger structural domains. The SPICE test set includes peptide-like systems that are substantially different from the typical GEOM molecules and introduce larger structures, biomolecular fragments, and repetitive backbone motifs. Despite this distinct domain shift and the increased structural complexity, the sum-pooling Uni-Bond model maintained a robust Exact Match accuracy of 93.07% and a macro F1-score of 0.9896. This demonstrates that the model's structural reasoning transfers effectively to larger molecules without requiring domain-specific fine-tuning.

Second, we assessed the model's performance on non-covalently bound systems using the **Water Clusters** and **Dimers** subsets. Crucially, the GEOM training dataset consisted exclusively of single, fully connected molecular graphs. Therefore, evaluating multi-component systems

is a severe test of whether the model can distinguish close non-covalent interactions (e.g., hydrogen bonds or van der Waals contacts) from actual covalent bonds based solely on geometry. On the Water Clusters dataset, the sum-pooling Uni-Bond model achieved perfect covalent-class Macro F1 and 98.80% Exact Match accuracy, successfully reconstructing intramolecular O-H bonds while largely ignoring the dense intermolecular hydrogen-bond networks. For the Dimers dataset, Uni-Bond achieved 89.49% Exact Match and a macro F1-score of 0.9183, reflecting the inherent geometric difficulty of distinguishing highly compressed van der Waals complexes from covalent bonds. Overall, these results suggest that the learned representations capture the geometric signatures of covalent bonding well enough to generalize beyond the training distribution, distinguishing covalent bonds from close non-covalent contacts in chemical regimes the model has not seen during training.

4.3. Out-of-Domain Direct SMILES Generation

We further evaluate whether coordinate-tokenized LLMs generalize out of distribution. Unlike bond prediction, direct SMILES generation must recover the full molecular string autoregressively, making it sensitive to chemical domain shift.

On SPICE OOD subsets, canonical match drops to near zero despite nonzero validity, indicating that the models often generate parseable but incorrect molecules. We provide a detailed breakdown of failures, as well as the comparison between BindGPT and binned tokenization, in Appendix C.4.

5. Conclusion

We presented Uni-Bond, a learned bond-reconstruction model that maps 3D atomic coordinates to discrete molecular graphs by combining symmetric sum-pooled atomic representations and pair representations from a pretrained Uni-Mol encoder with a lightweight pairwise classification head. On the GEOM scaffold split, Uni-Bond achieves near-perfect per-bond accuracy and 99.53% exact molecular-graph reconstruction, substantially outperforming both rule-based RDKit and the retrained GNN-based YuelBond baseline. Qualitative examples in Appendix F further show that errors are localized to specific ambiguous motifs, where Uni-Bond often resolves cases that distance-based and local message-passing methods misclassify.

The OOD experiments expose a sharper contrast between approaches. RDKit remains a strong rule-based baseline on several SPICE-derived subsets, particularly water clusters and peptide-like systems, while Uni-Bond achieves stronger edge-level performance on dimers and remains robust in disconnected, non-covalently bound systems where YuelBond produces spurious intermolecular bonds (Appendix D.2). We attribute this robustness to the combination of contextualized Uni-Mol pair representations with a factorized output space: each atom-pair prediction can incorporate global geometric context, while the final decision remains a local bond-type classification. This structure appears central to distinguishing covalent bonds from close non-covalent contacts in regimes the model has not seen during training.

The same pattern does not hold for the direct LLM pipeline, which is substantially less robust under distribution shift. Although the coordinate-tokenized Galactica model matches Uni-Bond on the in-domain GEOM test distribution, its exact-match accuracy collapses on SPICE test, water-cluster, and dimer subsets, and the diagnostic analysis on peptides indicates a stereochemical generalization failure: chiral annotations are recovered through scaffold-conditioned priors learned on GEOM rather than from the 3D input itself, and these priors do not transfer to peptide backbones. Unlike pairwise bond prediction, SMILES generation must jointly recover connectivity, bond orders, branches, ring closures, charges, and stereochemistry in a single autoregressive sequence, where early token errors propagate through the entire output. Coordinate tokenization alone, without an explicit geometric inductive bias, does not appear sufficient for reliable OOD geometry-to-SMILES generation.

Our study has several limitations that point to natural extensions. Most notably, the graph-reconstruction pipeline restricts training and evaluation to neutral closed-shell molecules, since charged species and open-shell radicals introduce ambiguities in RDKit-based graph sanitization that are difficult to resolve from bond orders alone. As a result, roughly 10% of the original GEOM data was excluded, and the model has not been exposed to the geometric signatures of charged or radical systems. Extending Uni-Bond to richer chemical regimes, for example by jointly predicting bond orders together with formal charges and radical electrons, or by training on more diverse datasets that span charged, open-shell, and organometallic systems, is an important direction for capturing a broader portion of chemical space. The direct LLM pipeline is in principle better suited to such annotations, but its current OOD fragility suggests that combining the two routes may be the most promising path forward. As a further future direction, we plan to evaluate Uni-Bond as a geometry-to-graph post-processing module for 3D molecular generative models, testing whether learned reconstruction from generated atom types and coordinates can reduce the atom–bond inconsistencies introduced by heuristic bond-perception pipelines.

Taken together, these results position transfer from a structured geometric foundation model as the more robust route from 3D coordinates to molecular graphs, while identifying direct coordinate-tokenized SMILES generation as an expressive but currently less reliable alternative.

Acknowledgmenents

The work of S.Pavlenko was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement **000000C313925P4E0002** and the agreement with HSE University № 139-15-2025-009. This research was supported in part through computational resources of HPC facilities at HSE University (Kostenetskiy et al., 2021).

Impact Statement

This paper presents work whose primary goal is to advance machine learning methods for molecular structure interpretation, specifically the reconstruction of molecular graphs from 3D atomic coordinates. The proposed approach is most likely to find use in computational chemistry, materials science, and structure-based drug discovery, where reliable geometry-to-graph translation is a recurring bottleneck between 3D generative models and downstream chemoinformatics pipelines.

We do not foresee specific harms beyond those generally associated with machine learning for chemistry. The contribution here is upstream of any particular molecular design objective and does not by itself enable new hazardous capabilities.

References

Alcaide, E., Gao, Z., Ke, G., Li, Y., Zhang, L., Zheng, H., and Zhou, G. Uni-mol docking v2: Towards realistic and accurate binding pose prediction. In *International Conference on Artificial Neural Networks*, pp. 34–41. Springer, 2025.

Axelrod, S. and Gómez-Bombarelli, R. GEOM: Energy-annotated molecular conformations for property prediction and molecular generation. *Scientific Data*, 9(1):185, 2022. doi: 10.1038/s41597-022-01288-4.

Bagal, V., Aggarwal, R., Vinod, P. K., and Priyakumar, U. D. Liggpt: Molecular generation using a transformer-decoder model, 2021. Preprint.

Bagal, V., Aggarwal, R., Vinod, P. K., and Priyakumar, U. D. Molgpt: Molecular generation using a transformer-decoder model. *Journal of Chemical Information and Modeling*, 62(9):2064–2076, 2022. doi: 10.1021/acs.jcim.1c00600.

Cheng, A. H., Cai, A., Miret, S., Malkomes, G., Phielipp, M., and Aspuru-Guzik, A. Group selfies: a robust fragment-based molecular string representation. *Digital Discovery*, 2:748–758, 2023. doi: 10.1039/D3DD00012E. URL http://dx.doi.org/10.1039/D3DD00012E.

Chithrananda, S., Grand, G., and Ramsundar, B. Chemberta: Large-scale self-supervised pretraining for molecular property prediction, 2020. URL https://arxiv.org/abs/2010.09885.

Eastman, P., Behara, P. K., Dotson, D. L., Galvelis, R., Herr, J. E., Horton, J. T., Mao, Y., Chodera, J. D., Pritchard, B. P., Wang, Y., De Fabritiis, G., and Markland, T. E. Spice, a dataset of drug-like molecules and peptides for training machine learning potentials. *Scientific Data*, 10 (1):11, 2023. doi: 10.1038/s41597-022-01882-6.

Edwards, C., Lai, T., Ros, K., Honke, G., Cho, K., and Ji, H. Translation between molecules and natural language. In Goldberg, Y., Kozareva, Z., and Zhang, Y. (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.26. URL https://aclanthology.org/2022.emnlp-main.26/.

Fan, J., Li, Z., Alcaide, E., Ke, G., Huang, H., and E, W. Accurate conformation sampling via protein structural diffusion. *Journal of Chemical Information and Modeling*, 64(22):8414–8426, 2024.

Fang, J., Zhang, S., Wu, C., Yang, Z., Liu, Z., Li, S., Wang, K., Du, W., and Wang, X. MolTC: Towards molecular relational modeling in language models. In Ku, L.-W., Martins, A., and Srikumar, V. (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1943–1958, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.116. URL https://aclanthology.org/2024.findings-acl.116/.

Francoeur, P. G. and Koes, D. R. Soltrannet—a machine learning tool for fast aqueous solubility prediction. *Journal of Chemical Information and Modeling*, 61(6):2530–2536, 2021. doi: 10.1021/acs.jcim.1c00331. URL https://doi.org/10.1021/acs.jcim.1c00331.

Ganeeva, V., Khrabrov, K., Kadurin, A., Savchenko, A., and Tutubalina, E. Chemical language models have problems with chemistry: A case study on molecule captioning task. In *The Second Tiny Papers Track at ICLR 2024*, 2024.

Gao, Z., Ji, X., Zhao, G., Wang, H., Zheng, H., Ke, G., and Zhang, L. Uni-qsar: an auto-ml tool for molecular property prediction. *arXiv preprint arXiv:2304.12239*, 2023.

Gilmer, J., Schoenholz, S. S., Riley, P. F., Vinyals, O., and Dahl, G. E. Neural message passing for quantum chemistry. In Precup, D. and Teh, Y. W. (eds.), *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pp. 1263–1272. PMLR, 06–11 Aug 2017. URL https://proceedings.mlr.press/v70/gilmer17a.html.

Goh, G. B., Hodas, N. O., Siegel, C., and Vishnu, A. Smiles2vec: An interpretable general-purpose deep neural network for predicting chemical properties. *arXiv preprint arXiv:1712.02034*, 2017. doi: 10.48550/arXiv.1712.02034. URL https://arxiv.org/abs/1712.02034.

Grimme, S. Exploration of chemical compound, conformer, and reaction space with meta-dynamics simulations based on tight-binding quantum chemical calculations. *Journal of Chemical Theory and Computation*, 15:2847–2862, 2019. doi: 10.1021/acs.jctc.9b00143.

Guan, J., Qian, W. W., Peng, X., Su, Y., Peng, J., and Ma, J. 3d equivariant diffusion for target-aware molecule generation and affinity prediction. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=kJqXEPXMsE0.

Hoogeboom, E., Garcia Satorras, V., Vignac, C., and Welling, M. Equivariant diffusion for molecule generation in 3d. *arXiv preprint arXiv:2203.17003*, 2022. doi: 10.48550/arXiv.2203.17003. URL https://arxiv.org/abs/2203.17003.

Hu, E. J., yelong shen, Wallis, P., Allen-Zhu, Z., Li, Y., Wang, S., Wang, L., and Chen, W. LoRA: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=nZeVKeeFYf9.

Hua, C., Luan, S., Xu, M., Ying, R., Fu, J., Ermon, S., and Precup, D. Mudiff: Unified diffusion for complete molecule generation. In *Proceedings of the Second Learning on Graphs Conference*, volume 231 of *Proceedings of Machine Learning Research*, 2024. URL https://proceedings.mlr.press/v231/hua24a.html.

Hutchings, M. E., Liu, J., Qiu, Y., Song, C., and Wang, L.-P. Bond-order time series analysis for detecting reaction events in ab initio molecular dynamics simulations. *Journal of Chemical Theory and Computation*, 16(3): 1606–1617, 2020. doi: 10.1021/acs.jctc.9b01039.

Irwin, R., Dimitriadis, S., He, J., and Bjerrum, E. J. Chemformer: a pre-trained transformer for computational chemistry. *Machine Learning: Science and Technology*, 3(1):015022, 2022. doi: 10.1088/2632-2153/ac3ffb.

Kadukova, M. and Grudinin, S. Knodle: A support vector machines-based automatic perception of organic molecules from 3d coordinates. *Journal of Chemical Information and Modeling*, 56(8):1410–1419, 2016. doi: 10.1021/acs.jcim.5b00512. URL https://doi.org/10.1021/acs.jcim.5b00512.

Kim, Y. and Kim, W. Universal structure conversion method for organic molecules: From atomic connectivity to three-dimensional geometry. *Bulletin of the Korean Chemical Society*, 36(7):1769–1777, 2015. doi: 10.1002/bkcs.10334.

Kostenetskiy, P., Chulkevich, R., and Kozyrev, V. Hpc resources of the higher school of economics. In *Journal of Physics: Conference Series*, volume 1740, pp. 012050. IOP Publishing, 2021.

Kovács, D. P., Moore, J. H., Browning, N. J., Batatia, I., Horton, J. T., Pu, Y., Kapil, V., Witt, W. C., Magdau, I.-B., Cole, D. J., et al. Mace-off: Short-range transferable machine learning force fields for organic molecules. *Journal of the American Chemical Society*, 147(21):17598–17611, 2025. doi: 10.1021/jacs.4c07099.

Krenn, M., Hase, F., Nigam, A., Friederich, P., and Aspuru-Guzik, A. Self-referencing embedded strings (selfies): A 100% robust molecular string representation. *Machine Learning: Science and Technology*, 1(4):045024, 2020. doi: 10.1088/2632-2153/aba947.

Kuznetsov, M., Valiev, A., Aliper, A., Polykovskiy, D., Tutubalina, E., Shayakhmetov, R., and Miftahutdinov, Z. nach0-pc: Multi-task language model with molecular point cloud encoder, 2024. URL https://arxiv.org/abs/2410.09240.

Landrum, G. RDKit: Open-source cheminformatics, 2013. URL http://www.rdkit.org. Accessed: 2023.

Le, T., Cremer, J., Noé, F., Clevert, D.-A., and Schütt, K. Navigating the design space of equivariant diffusion-based generative models for de novo 3d molecule generation, 2023. URL https://arxiv.org/abs/2309.17296.

Liu, P., Ren, Y., Tao, J., and Ren, Z. Git-mol: A multi-modal large language model for molecular science with graph, image, and text. *Computers in Biology and Medicine*, 171: 108073, 2024. doi: 10.1016/j.compbiomed.2024.108073.

Liu, Z., Zhang, W., Xia, Y., Wu, L., Xie, S., Qin, T., Zhang, M., and Liu, T.-Y. Molxpt: Wrapping molecules with text for generative pre-training. In *Proceedings of the 61st Annual Meeting of the Association for Computational*

Linguistics (Volume 2: Short Papers), pp. 1606–1616. Association for Computational Linguistics, 2023. doi: 10.18653/v1/2023.acl-short.138.

Morehead, A. and Cheng, J. Geometry-complete diffusion for 3d molecule generation and optimization, 2024. URL https://arxiv.org/abs/2302.04313.

Muenks, A., Zepeda, S., Zhou, G., Veesler, D., and DiMaio, F. Automatic and accurate ligand structure determination guided by cryo-electron microscopy maps. *Nature Communications*, 14(1):1164, 2023. doi: 10.1038/s41467-023-36732-5.

Noutahi, E., Gabellini, C., Craig, M., Lim, J. S. C., and Tossou, P. Gotta be safe: A new framework for molecular design, 2023. URL https://arxiv.org/abs/2310.10773.

O'Boyle, N. M., Banck, M., James, C. A., Morley, C., Vandermeersch, T., and Hutchison, G. R. Open babel: An open chemical toolbox. *Journal of Cheminformatics*, 3(1): 33, 2011. doi: 10.1186/1758-2946-3-33. URL https://doi.org/10.1186/1758-2946-3-33.

Peng, X., Guan, J., Liu, Q., and Ma, J. Moldiff: Addressing the atom-bond inconsistency problem in 3d molecule diffusion generation, 2023. URL https://arxiv.org/abs/2305.07508.

Pinheiro, P. O., Kirchmeyer, M., and Saremi, S. 3d molecule generation by denoising voxel grids. *arXiv preprint arXiv:2306.07473*, 2023. doi: 10.48550/arXiv.2306.07473. URL https://arxiv.org/abs/2306.07473.

Potapov, D., Rogovoi, S., Khrabrov, K., Ushenin, K., Korovin, A., Ber, A., Kadurin, A., and Tsypin, A. A conformational benchmark for optical property prediction with solvent-aware graph neural networks. *Communications Chemistry*, 2026.

Pracht, P., Bohle, F., and Grimme, S. Automated exploration of the low-energy chemical space with fast quantum chemical methods. *Physical Chemistry Chemical Physics*, 22:7169–7192, 2020. doi: 10.1039/C9CP06869D.

Ragoza, M., Masuda, T., and Koes, D. R. Learning a continuous representation of 3d molecular structures with deep generative models.

Ragoza, M., Masuda, T., and Koes, D. R. Generating 3d molecules conditional on receptor binding sites with deep generative models. *Chemical science*, 13(9):2701–2713, 2022.

Rogers, D. and Hahn, M. Extended-connectivity fingerprints. *Journal of Chemical Information and Modeling*, 50(5):742–754, 2010. doi: 10.1021/ci100050t.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., and Das, P. Large-scale chemical language representations capture molecular structure and properties. *Nature Machine Intelligence*, 4(12):1256–1264, 2022. doi: 10.1038/s42256-022-00580-7.

Schneuing, A., Harris, C., Du, Y., Didi, K., Jamasb, A., Igashov, I., Du, W., Gomes, C., Blundell, T., Lio, P., Welling, M., Bronstein, M., and Correia, B. Structure-based drug design with equivariant diffusion models, 2024. URL https://arxiv.org/abs/2210.13695.

Schwaller, P., Laino, T., Gaudin, T., Bolgar, P., Hunter, C. A., Bekas, C., and Lee, A. A. Molecular transformer: A model for uncertainty-calibrated chemical reaction prediction. *ACS Central Science*, 5(9):1572–1583, 2019. doi: 10.1021/acscentsci.9b00576. URL https://doi.org/10.1021/acscentsci.9b00576.

Smith, J. S., Isayev, O., and Roitberg, A. E. Ani-1: an extensible neural network potential with dft accuracy at force field computational cost. *Chemical science*, 8(4):3192–3203, 2017. doi: https://doi.org/10.1039/C6SC05720A.

Tang, X., Tran, A., Tan, J., and Gerstein, M. B. Mollm: a unified language model for integrating biomedical text with 2d and 3d molecular representations. *Bioinformatics*, 40(Supplement 1):i357–i368, 2024. doi: 10.1093/bioinformatics/btae260.

Taylor, R., Kardas, M., Cucurull, G., Scialom, T., Hartshorn, A. S., Saravia, E., Poulton, A., Kerkez, V., and Stojnic, R. Galactica: A large language model for science. *ArXiv*, abs/2211.09085, 2022. URL https://api.semanticscholar.org/CorpusID:253553203.

Vazquez-Salazar, L. I., Boittier, E. D., and Meuwly, M. Uncertainty quantification for predictions of atomistic neural networks, 2022. URL https://arxiv.org/abs/2207.06916.

Vignac, C., Osman, N., Toni, L., and Frossard, P. Midi: Mixed graph and 3d denoising diffusion for molecule generation. In *Machine Learning and Knowledge Discovery in Databases: Research Track*, pp. 560–576, 2023. doi: 10.1007/978-3-031-43415-0_33.

Wang, J. and Dokholyan, N. V. Multimodal bond reconstruction toward generative molecular design. *Journal of chemical information and modeling*, 66(2):1003–1012, 2026.

Wang, J., Liu, J., Wang, H., Zhou, M., Ke, G., Zhang, L., Wu, J., Gao, Z., and Lu, D. A comprehensive transformer-based approach for high-accuracy gas adsorption predic-

tions in metal-organic frameworks. *Nature Communications*, 15(1):1904, 2024.

Weininger, D. Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules. *Journal of Chemical Information and Computer Sciences*, 28(1):31–36, 1988. doi: 10.1021/ci00057a005.

Welsh, I. D. and Allison, J. R. Automated simultaneous assignment of bond orders and formal charges. *Journal of Cheminformatics*, 11(1):18, 2019. doi: https://doi.org/10.1186/s13321-019-0340-0.

Wu, L., Wang, J., Gao, Z., Ji, X., Zhu, R., Li, X., Zhang, L., Ke, G., et al. Uni-mol3: A multi-molecular foundation model for advancing organic reaction modeling. *arXiv preprint arXiv:2508.00920*, 2025.

Wu, Z., Ramsundar, B., Feinberg, E. N., Gomes, J., Geniesse, C., Pappu, A. S., Leswing, K., and Pande, V. Moleculenet: A benchmark for molecular machine learning, 2018. URL https://arxiv.org/abs/1703.00564.

Xue, D., Zhang, H., Chen, X., Xiao, D., Gong, Y., Chuai, G., Sun, Y., Tian, H., Wu, H., Li, Y., and Liu, Q. X-mol: large-scale pre-training for molecular understanding and diverse molecular analysis. *Science Bulletin*, 67(9): 899–902, 2022. doi: 10.1016/j.scib.2022.01.029.

Zhang, Q., Zhang, W., Li, Y., Wang, J., Zhang, L., and Hou, T. A rule-based algorithm for automatic bond type perception. *Journal of Cheminformatics*, 4(1):26, 2012. doi: 10.1186/1758-2946-4-26. URL https://doi.org/10.1186/1758-2946-4-26.

Zholus, A., Kuznetsov, M., Schutski, R., Shayakhmetov, R., Polykovskiy, D., Chandar, S., and Zhavoronkov, A. Bindgpt: A scalable framework for 3d molecular design via language modeling and reinforcement learning, 2024. URL https://arxiv.org/abs/2406.03686.

Zhou, G., Gao, Z., Ding, Q., Zheng, H., Xu, H., Wei, Z., Zhang, L., and Ke, G. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=6K2RM6wVqKu.

A. Uni-Mol Implementation Details

For the geometric encoding stage, we utilized the pretrained Uni-Mol base architecture, which contains approximately 47 million parameters. The encoder produces atomic embeddings of dimensionality $d = 512$ and pairwise embeddings of dimensionality $c = 64$.

A.1. MLP Classification Head.

The bond-classification head consists of a two-layer Multi-Layer Perceptron (MLP) with approximately 74K trainable parameters. For a candidate atom pair (i, j), we first apply symmetric sum pooling to the two atomic embeddings, $H_i + H_j$, where $H_i, H_j \in \mathbb{R}^{512}$. We then concatenate this pooled atomic representation with the pairwise interaction embedding $E_{ij} \in \mathbb{R}^{64}$, forming $x_{ij} = [(H_i + H_j) \parallel E_{ij}]$. The resulting 576-dimensional vector is projected to a hidden dimension of 128, followed by a GELU activation and Layer Normalization, and then projected to 5 output logits corresponding to the bond classes.

A.2. Candidate Selection and Loss Formulation.

Candidate atom pairs were restricted to atoms within a spatial cutoff radius of $r_c = 3.0$ Å. To enforce the undirected nature of chemical bonds, only upper-triangular atom pairs $(i < j)$ were considered. The model predicts a categorical distribution

$$\hat{p}_{ij} \in \mathbb{R}^5$$

over the five bond classes for every candidate pair. Training was performed using a standard multi-class cross-entropy loss averaged over all valid candidate pairs in the candidate set $\mathcal{E}$.

A.3. Training and Decoding.

We employed full fine-tuning of the complete Uni-Mol encoder jointly with the MLP head. The model was trained for 5 epochs using the AdamW optimizer with learning rate 5×10^{-5} and batch size 8.

At inference time, bond types were decoded greedily according to

$$\hat{c}_{ij} = \arg\max_c \hat{p}_{ij}^{(c)}.$$

Because only pairs satisfying $(i < j)$ are evaluated, the resulting predictions directly define a symmetric connectivity matrix used for graph reconstruction and RDKit sanitization.

B. Preliminary Extension to Formal Charges and Radical Electrons

The main Uni-Bond experiments restrict graph reconstruction to neutral closed-shell molecules, since formal charges and radical electrons introduce additional atom-level annotations that are not determined by bond orders alone. As a preliminary extension, we evaluated whether the same Uni-Mol representation can support these annotations by adding two atom-level prediction heads on top of the shared encoder: one for formal charge and one for radical-electron count. The bond head is unchanged. The model is initialized from the Uni-Bond bond-reconstruction checkpoint, the Uni-Mol backbone is kept frozen, and only the task heads are trained.

For each atom, the charge head predicts formal charge values in $\{-1, 0, +1\}$, while the radical head predicts values in $\{0, 1\}$. The total loss is

$$\mathcal{L} = \mathcal{L}_{\text{bond}} + \lambda_{\text{charge}}\mathcal{L}_{\text{charge}} + \lambda_{\text{radical}}\mathcal{L}_{\text{radical}},$$

with $\lambda_{\text{charge}} = 0.25$ and $\lambda_{\text{radical}} = 0.1$. Because the atom-level targets are highly imbalanced, we report not only accuracy and macro F1, but also nonzero-class detection metrics and exact-match accuracy restricted to molecules that contain at least one nonzero charge or radical atom.

The results suggest that the Uni-Mol representations used by Uni-Bond contain substantial information about charge and open-shell structure. Radical prediction is particularly strong, reaching a nonzero atom F1 of 0.9042 and a true-radical molecule exact-match rate of 76.12%. Formal-charge prediction is more challenging: the model detects charged molecules reliably, with molecule-level has-charge F1 of 0.9954, but exact atom-level charge assignment within charged molecules

Table 4. Preliminary multitask Uni-Bond results for joint bond, formal-charge, and radical-electron prediction. Atom-level charge and radical accuracies are dominated by the neutral/no-radical classes; nonzero detection and true-nonzero molecule exact match are therefore more diagnostic.

Metric	**Formal charge**	**Radical electrons**
Atom accuracy	0.9988	0.9999
Macro F1	0.8725	0.9521
Nonzero atom precision	0.8022	0.9672
Nonzero atom recall	0.8739	0.8489
Nonzero atom F1	0.8365	0.9042
Molecule has-nonzero F1	0.9954	0.9498
Molecule exact match, all molecules	96.85%	99.64%
Molecule exact match, true-nonzero only	59.35%	76.12%

remains harder. The weakest class is formal charge -1, whose atom-level F1 is 0.6277, compared with 0.9903 for formal charge $+1$.

These findings support joint prediction of bonds, charges, and radical electrons as a promising route toward extending Uni-Bond beyond neutral closed-shell molecules. At the same time, the gap between high aggregate accuracy and lower true-nonzero exact match highlights the need for imbalance-aware objectives and evaluation protocols when modeling rare chemical annotations.

C. LLM Implementation Details and Additional Experiments

C.1. LLM Dataset Construction and Training Hyperparameters

For the direct-SMILES generation experiments, we constructed LMDB datasets at the molecule level. Each record contains the source SMILES string, the atomic numbers, and a tensor of 3D conformations for the same molecule. The atomic identities are shared across conformers, while the coordinate tensor provides multiple geometric realizations of the same molecular graph and SMILES target.

During training, one conformer is sampled at random from the stored conformer tensor each time a molecule is loaded. This exposes the model to conformational variation while keeping the target SMILES fixed. During evaluation on GEOM, we use all stored conformers, so each molecule contributes multiple coordinate prompts with the same reference SMILES. This evaluates whether the model can recover the same molecular identity consistently from different 3D conformations.

We trained models on two dataset variants. The filtered variant follows the graph-reconstruction setting and excludes molecules that are difficult to sanitize or reconstruct unambiguously through RDKit, such as charged molecules. The unfiltered variant keeps these molecules because the LLM directly predicts the source SMILES string from the XYZ prompt and does not require an intermediate RDKit-sanitized bond graph. This allows us to test whether direct sequence prediction can use a broader training distribution than graph-based bond reconstruction.

All direct-SMILES generation models were initialized from a local Galactica-1.3B checkpoint and fine-tuned with supervised learning on 8 H100 80GB GPUs. We used LoRA adaptation rather than full-model fine-tuning, applying adapters to all linear layers and training the newly introduced coordinate-token embeddings. Both coordinate encodings used the same optimizer settings, scheduler, precision, and generation limit; the main training difference was the per-device batch size, which was smaller for the longer BindGPT-style sequences. The training hyperparameters are listed in Table 5.

C.2. LLM Prompt Format

For direct SMILES generation, each molecule is serialized into a causal-language-model sequence consisting of a coordinate prompt followed by the target SMILES. The coordinate prompt is not stored as ordinary decimal XYZ text. Instead, it is assembled directly as token IDs. The sequence begins with the special token `<XYZ>`, followed by a newline, the number of atoms, and another newline. The dataset then appends, for each atom in order, the tokenizer IDs for the atomic symbol followed immediately by the coordinate-token IDs for that atom.

Table 5. Hyperparameters for Galactica-1.3B supervised fine-tuning. Shared values apply to both BindGPT and binned coordinate encodings.

Hyperparameter	**BindGPT**	**Binned**
Base model	Galactica-1.3B	
Data regimes	filtered, unfiltered	
Stored conformers per molecule	10: lowest-energy conformer + 9 random conformers	
Coordinate preprocessing	centered, principal-axis aligned	
Train/eval conformer selection	random / random	
GEOM test conformer selection	all stored conformers	
Target SMILES augmentation probability	0.5	
Target SMILES augmentations	kekule, rooted, random, cycle renumber	
Coordinate encoding	2 tokens / coordinate	1 token / coordinate
Coordinate vocabulary	integer + 3-digit fraction	1024 bins
Coordinate range	integer range ± 15	$[-12, 12]$ Å
Tokens per atom for coordinates	6	3
LoRA rank r	32	
LoRA α	64	
LoRA dropout	0.05	
LoRA target modules	all linear layers	
Per-device batch size	48	64
Number of GPUs	8× H100 80GB	
Gradient accumulation	1	
Effective batch size	384	512
Epochs	50	
Learning rate	3×10^{-4}	
LR schedule	cosine	
Warmup ratio	0.05	
Weight decay	0.01	
Max gradient norm	1.0	
Precision	bfloat16	
Gradient checkpointing	enabled	
Max generation tokens	256	

```
<XYZ>
N
A_1 τ(x_1) τ(y_1) τ(z_1) A_2 τ(x_2) τ(y_2) τ(z_2) ··· A_N τ(x_N) τ(y_N) τ(z_N)
Provide the SMILES formula for the molecule.  If the molecule is corrupted or
invalid, respond with ``no smiles possible''.
[START_SMILES] target_smiles [END_SMILES]
```

Here A_i denotes the tokenized atomic symbol and $\tau(\cdot)$ denotes the coordinate-tokenization function. The apparent spaces in the schematic are for readability only: in the implementation, the atom-coordinate block is assembled directly as token IDs, and no explicit separator token is inserted between successive atoms beyond the atom and coordinate tokens themselves. The supervised loss is applied only to the final answer segment, while all prompt tokens are masked out.

We evaluated two coordinate tokenization schemes. In the BindGPT-style encoding, each scalar coordinate is represented by two special tokens: one for the signed integer part and one for the three-digit fractional part. For example, a coordinate such as -0.640 is encoded as `<INT_-0>` followed by `<FRAC_640>`. Thus, each atom contributes six coordinate tokens. In the binned encoding, each scalar coordinate is discretized into one of 1024 bins over $[-12, 12]$ Å, so each atom contributes three coordinate tokens. Coordinates are centered and aligned to principal axes before tokenization.

The answer is appended as `[START_SMILES] target_smiles [END_SMILES]`. During supervised fine-tuning, the model is optimized with the standard causal language-modeling objective on this answer segment only. In the experiments reported here, the corruption probability was set to zero, so training targets were valid source SMILES strings rather than the `no smiles possible` label.

Table 6. Direct SMILES generation results on the GEOM test set. Each cell reports exact match, valid SMILES rate, and mean Morgan fingerprint Tanimoto similarity.

Coordinate Encoding	Filtered GEOM	Unfiltered GEOM
BindGPT	Exact: 95.10% Valid: 99.78% Tanimoto: 0.9913	Exact: 95.27% Valid: 99.79% Tanimoto: 0.9911
Binned	Exact: 94.81% Valid: 99.78% Tanimoto: 0.9899	Exact: 95.04% Valid: 99.77% Tanimoto: 0.9903

At evaluation time, generation starts from the prompt and proceeds greedily until `[END_SMILES]` is produced or the maximum generation length of 256 tokens is reached. The generated SMILES is extracted from the text between `[START_SMILES]` and `[END_SMILES]` and evaluated using exact match, RDKit-canonicalized match, valid SMILES rate, and Morgan fingerprint Tanimoto similarity.

C.3. In-Domain GEOM Evaluation of LLM Models

We evaluated the direct-SMILES generation models on the GEOM test set using the procedure described in Section C.2. We report results separately for the filtered and unfiltered GEOM datasets, where the unfiltered setting retains charged and otherwise graph-sanitization-challenging molecules.

We use three metrics. **Exact match** canonicalizes both generated and reference SMILES with RDKit before comparison, making the metric invariant to many equivalent SMILES representations. **Validity** is the fraction of generated outputs that RDKit can parse as valid SMILES. Finally, **Tanimoto** is the mean Morgan fingerprint Tanimoto similarity between the generated and reference molecules.

Table 6 shows that the LLM achieves strong in-distribution performance. Exact match is approximately 95% across settings, validity exceeds 99.7%, and mean Tanimoto similarity is close to 0.99. The similar performance of filtered and unfiltered runs suggests that direct SMILES generation can make use of the broader unfiltered dataset without requiring the charge-related filtering needed by graph-reconstruction pipelines.

C.4. OOD Evaluation of LLM Models

We evaluated the direct-SMILES LLM models on two out-of-domain subsets derived from SPICE (Eastman et al., 2023): SPICE test and SPICE dipeptides. These subsets expose different failure modes. SPICE test primarily probes constitutional generalization to larger and chemically different molecules; its reference SMILES do not contain explicit stereochemical annotations. SPICE dipeptides, in contrast, introduces a strong stereochemical shift, since nearly all reference molecules contain explicit chiral annotations.

For SPICE test, we report the standard direct-generation metrics: exact match, validity, and mean Morgan fingerprint Tanimoto similarity. For SPICE dipeptides, standard exact match becomes a strict isomeric metric because a prediction must recover both the molecular constitution and the correct stereochemistry. We therefore additionally report **constitutional match**, computed by removing stereochemical distinctions before exact comparison. This separates failures in recovering the molecular graph from failures in assigning the correct stereochemical form.

Table 7 shows that all LLM variants degrade sharply on SPICE test relative to the in-domain GEOM evaluation. Although validity remains relatively high, around 73–76%, exact match is below 1% for every model. Mean Tanimoto similarity is around 0.21–0.23, indicating that the models often generate valid molecules sharing some local fragments with the reference, but rarely recover the full molecular constitution. This suggests that SPICE test primarily exposes constitutional OOD failure rather than stereochemical failure.

Table 8 shows a different pattern for SPICE dipeptides. The gap between constitutional and exact recovery is large. For example, the filtered BindGPT model reaches 56.0% constitutional match and a mean Tanimoto similarity of 0.7323, but its exact match is only 0.25%. This indicates that the model can often recover much of the peptide constitution, but almost never assigns the exact stereochemical form. The same pattern appears across the other models: validity and Tanimoto

Table 7. OOD direct-SMILES generation results on SPICE test. Exact match and validity are reported as percentages; Tanimoto is the mean Morgan fingerprint Tanimoto similarity.

Training Set	Encoding	Exact Match	Tanimoto	Validity
Unfiltered	BindGPT	0.80%	0.2264	76.20%
Unfiltered	Binned	0.50%	0.2123	74.03%
Filtered	BindGPT	0.75%	0.2289	73.55%
Filtered	Binned	0.66%	0.2086	73.16%

Table 8. OOD direct-SMILES generation results on SPICE dipeptides. Exact match accounts for stereochemistry. Constitutional match ignores stereochemical distinctions before canonical comparison. Validity is reported as a percentage; Tanimoto is the mean Morgan fingerprint Tanimoto similarity.

Training Set	Encoding	Exact Match	Constitutional Match	Tanimoto	Validity
Unfiltered	BindGPT	0.18%	27.9%	0.5975	97.75%
Unfiltered	Binned	0.15%	11.8%	0.4173	94.80%
Filtered	BindGPT	0.25%	56.0%	0.7323	96.41%
Filtered	Binned	0.30%	15.5%	0.4343	94.81%

similarity is reasonable, while exact match remains below 0.3%.

These results suggest that coordinate-tokenized LLMs do not reliably infer chirality from 3D geometry under distribution shift. Instead, they appear to learn scaffold-conditioned priors for stereochemical tokens from GEOM, where chiral molecules are relatively rare and appear in different chemical contexts. Those priors do not transfer to peptide backbones, where stereochemistry is nearly ubiquitous. Thus, SPICE test exposes poor constitutional OOD generalization, while SPICE dipeptides additionally exposes a stereochemical generalization failure that is hidden by aggregate GEOM in-domain metrics.

This stereochemical failure is consistent with the structure of the training distribution. Chiral annotations are rare in GEOM, whereas they are nearly ubiquitous in SPICE dipeptides. As a result, the language model can obtain low training loss by learning scaffold-conditioned priors for when to emit stereochemical tokens, rather than learning a coordinate-driven parity rule. This shortcut works in distribution, where chiral scaffolds resemble those seen during training, but it does not transfer to peptide backbones. A diagnostic analysis supports this interpretation: on SPICE dipeptides, the model may recover the correct constitution but still omit or misassign `@`/`@@` annotations; conversely, on SPICE test, where the references contain no stereochemical markers, the model still frequently emits them. Thus, stereochemistry is not controlled reliably by the 3D input in the current coordinate-tokenized LLM.

BindGPT-style tokenization consistently outperforms the binned representation, especially on SPICE dipeptides. The likely reason is that BindGPT preserves coordinate values at much higher effective resolution: each coordinate is represented by signed integer and three-digit fractional tokens, whereas the binned representation quantizes the full $[-12, 12]$ Å range into 1024 bins. The binned representation is shorter, but adjacent bins are independent discrete tokens and the model must learn their ordering from data. By contrast, the BindGPT-style representation exposes sign, integer part, and fractional structure explicitly, which may make local geometric patterns easier to transfer. This appears particularly important for peptide geometries, where small local coordinate differences determine backbone and side-chain arrangements.

D. YuelBond Baseline Details

D.1. Retraining on Explicit-Hydrogen GEOM

The original YuelBond model (Wang & Dokholyan, 2026) was trained in a molecular representation without explicit hydrogen atoms, whereas our dataset retains all hydrogens and evaluates full molecular graphs, including X–H bonds. Directly using the original checkpoint would therefore introduce a representation mismatch: the baseline would be evaluated on atom types, candidate edges, and graph topologies that differ from those used during its training.

To ensure a fair comparison with Uni-Bond, we trained a YuelBond baseline from scratch on our processed GEOM dataset

Table 9. Out-of-domain evaluation of YuelBond on SPICE-derived subsets. Macro F1 is computed excluding the no-bond class.

OOD Dataset	Macro F1	Exact Match Accuracy
Dipeptides	0.6959	0.5022
Dimers	0.7943	0.5894
Water Clusters	0.8742	0.0000
SPICE test	0.7257	0.2453

using the same explicit-hydrogen molecular representation, the same scaffold split, and the same bond-label convention. Unless stated otherwise, all YuelBond results reported in this work correspond to this retrained explicit-hydrogen checkpoint rather than the original pretrained model.

This retraining is important because explicit hydrogens substantially change the bond-reconstruction problem. They increase the number of atoms and candidate atom pairs, introduce many additional X–H covalent bonds, and modify the class imbalance between covalent and no-bond edges. As a result, the baseline must be adapted to the same input and output space used by Uni-Bond.

D.2. OOD Evaluation of YuelBond

To provide an additional comparison for out-of-domain generalization, we evaluated our retrained YuelBond baseline on SPICE-derived OOD subsets. This checkpoint was trained from scratch on our explicit-hydrogen GEOM training set, using the same molecular representation as Uni-Bond. Since the candidate-edge distribution is strongly dominated by the no-bond class, we report the macro F1-score after excluding the no-bond class. This metric better reflects performance on chemically meaningful covalent bond assignments than a score dominated by the abundant no-bond class.

The results are shown in Table 9. YuelBond preserves strong edge-level performance on several OOD subsets, but Exact Match accuracy decreases substantially under distribution shift. This indicates that relatively localized bond-assignment errors can be sufficient to break complete molecular-graph reconstruction.

A characteristic failure mode appears for disconnected molecular systems, especially water clusters. In this subset, YuelBond recovers all true intramolecular O–H bonds, giving perfect recall for the single-bond class. However, it also predicts additional false-positive single bonds between nearby molecules. This behavior is visible from the decrease in the predicted no-bond fraction relative to the true no-bond fraction: for water clusters, the true no-bond fraction is 0.8128, whereas the predicted no-bond fraction is only 0.7585. As a result, the model constructs spurious covalent links across non-covalently bound components. Even though most individual edge labels remain correct, a single extra intermolecular bond is enough to make the entire predicted graph differ from the reference graph, leading to zero exact-match accuracy on this subset.

This failure mode cannot be attributed solely to the absence of disconnected molecular systems in the GEOM training set, since Uni-Bond is trained under the same data regime but remains substantially more robust on water clusters and dimers. Rather, the result suggests a limitation of the YuelBond inductive bias in this setting. YuelBond performs bond classification through local message passing on a candidate graph constructed from spatially close atom pairs. As a result, short non-covalent contacts in disconnected systems can be difficult to distinguish from genuine covalent interactions when the decision is dominated by local geometry and neighborhood-level features. In water clusters, this manifests as false-positive intermolecular single bonds: the model recovers the true intramolecular O–H bonds, but also connects nearby molecules that should remain separate. By contrast, Uni-Bond uses contextual pair representations from the Uni-Mol encoder, which appear to provide a more robust distinction between covalent bonds and close non-covalent contacts.

E. Confusion Matrix Analysis

To further characterize the error structure of Uni-Bond, we report row-normalized confusion matrices for the in-domain GEOM scaffold test set and the three OOD datasets in Figure 3. Each row corresponds to the reference bond class and each column to the predicted class; therefore, off-diagonal entries directly indicate systematic bond-type confusions. The matrices show that most probability mass is concentrated on the diagonal across datasets, confirming that Uni-Bond errors are sparse and class-specific. The main degradation under OOD shift is observed for the Dimers subset, where double and triple bonds are more frequently confused with single or lower-order bonds.

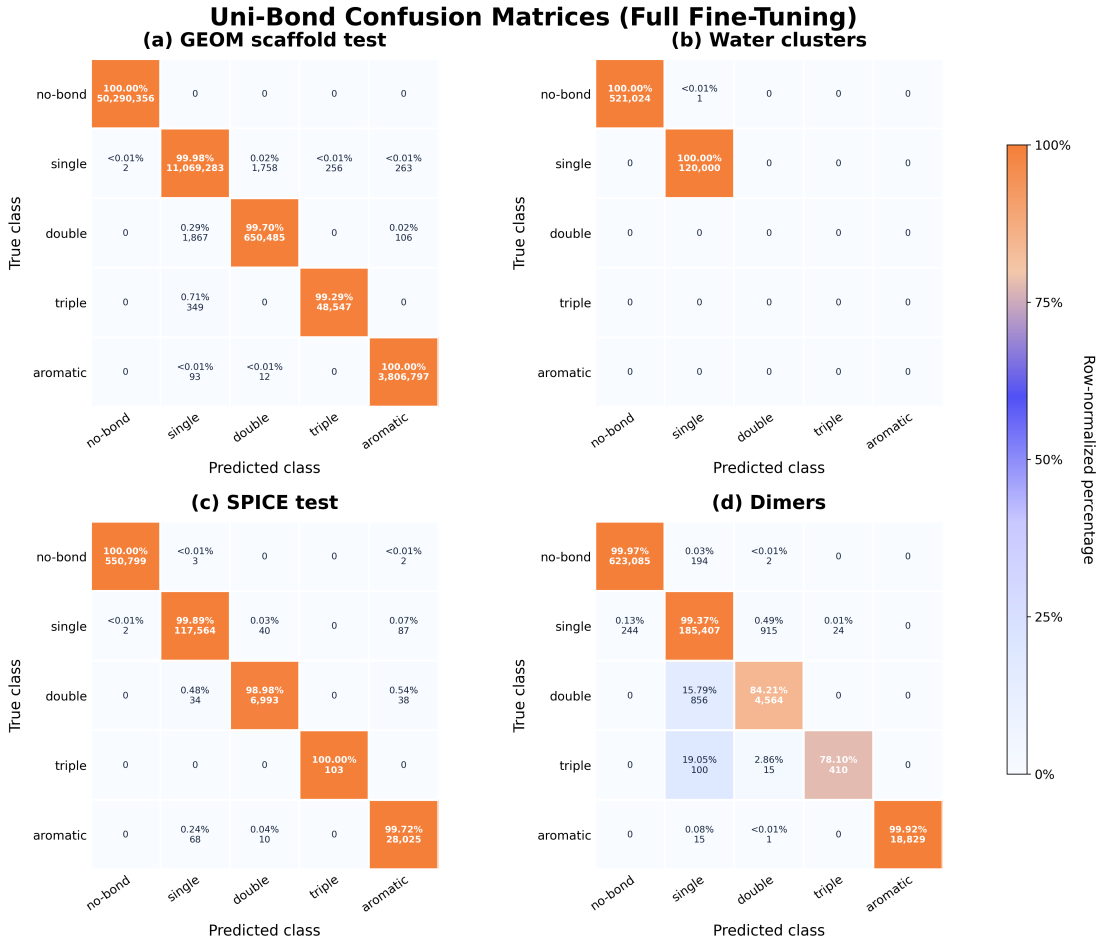

Figure 3. Row-normalized confusion matrices for Uni-Bond with full Uni-Mol fine-tuning. Panels show (a) the GEOM scaffold test set, (b) Water Clusters, (c) SPICE test, and (d) Dimers. Rows correspond to reference bond classes and columns to predicted bond classes. Each cell reports the row-normalized percentage and the corresponding number of atom-pair predictions.

F. Qualitative Examples of Bond Reconstruction

To complement the quantitative evaluation, we provide representative qualitative examples of bond reconstruction errors and corrections in Figure 4. Each row corresponds to one molecular conformer, and the three columns show the same case from different perspectives: the input 3D conformer, the RDKit prediction, and the Uni-Bond prediction. The highlighted regions indicate atom pairs or local substructures where the predicted connectivity differs from the reference molecular graph.

These examples illustrate that reconstruction errors are often highly localized, even when the rest of the molecular graph is predicted correctly. RDKit failures typically arise in geometrically ambiguous regions, where distance-based and valence-based heuristics can assign an incorrect bond order or miss a chemically intended connection. Uni-Bond corrects some of these cases by using contextual geometric representations from the Uni-Mol encoder rather than relying only on local distance thresholds. At the same time, the examples also show that learned bond reconstruction is not error-free: some local motifs remain challenging when the geometry is ambiguous or when the relevant chemical context is underrepresented in the training data.

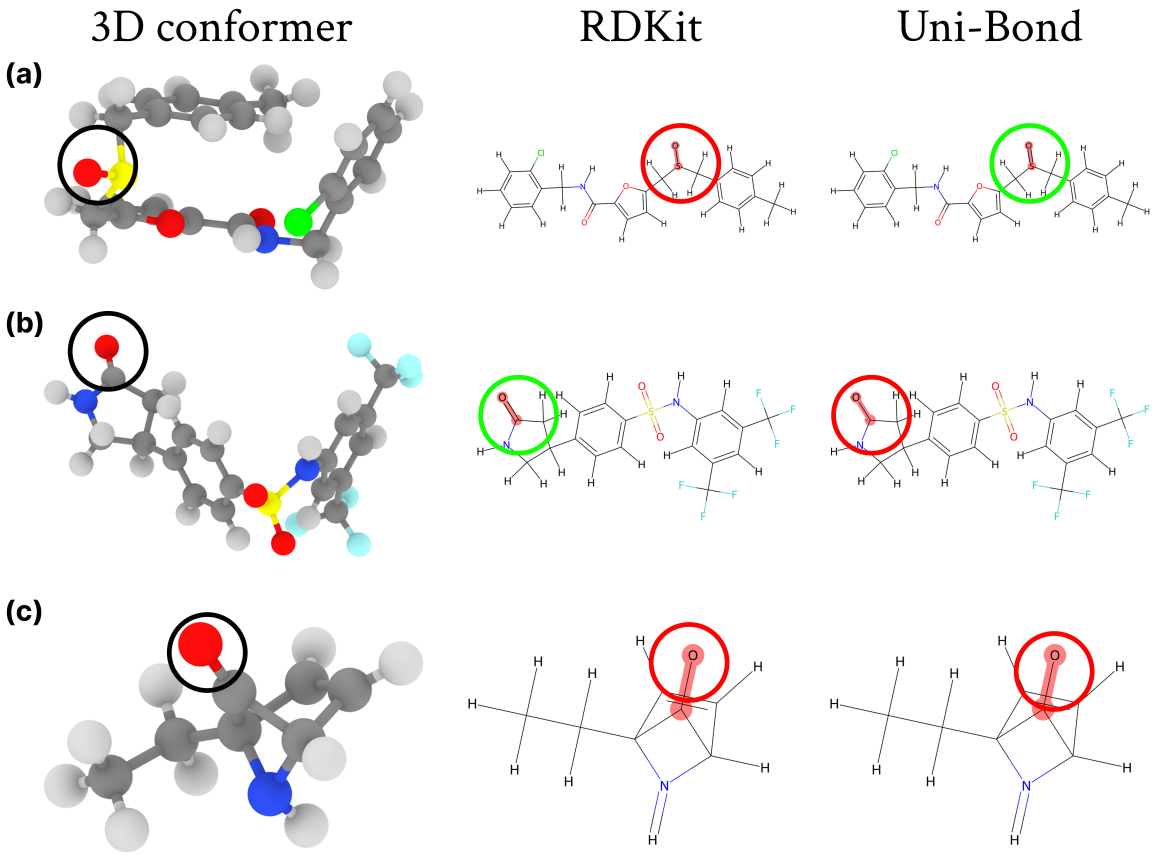

Figure 4. Representative qualitative examples of bond reconstruction on molecular conformers. Rows (a)–(c) correspond to different molecules. Columns show, from left to right, the input 3D conformer, the RDKit prediction, and the Uni-Bond prediction. Black circles mark the highlighted regions in the 3D conformers. Red circles indicate incorrect bond assignments, while green circles indicate correct assignments in the corresponding 2D molecular graphs.