

GraphChem3D: Bond-Centric Prediction of Molecular 3D Structures

Matthew Osvaldo

Nanyang Technological University

Joshua Adrian Cahyono

Nanyang Technological University

Leonard Ng Wei Tat

`leonard.ngwt@ntu.edu.sg`

Nanyang Technological University

Research Article

Keywords:

Posted Date: September 24th, 2026

DOI: <https://doi.org/10.21203/rs.3.rs-10379050/v1>

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

[Read Full License](#)

Additional Declarations: No competing interests reported.

GraphChem3D: Bond-Centric Prediction of Molecular 3D Structures

Matthew Osvaldo^{1*} Joshua Adrian Cahyono^{2†} Leonard Wei Tat Ng^{1‡}

¹School of Materials Science and Engineering, Nanyang Technological University, Singapore

²College of Computing and Data Science, Nanyang Technological University, Singapore
May 2026

Abstract

Predicting the three-dimensional conformation of molecules from their bonding graph is a fundamental problem in computational chemistry with applications in drug discovery, materials science, and molecular simulation. We present a bond-centric graph neural network framework for predicting local molecular geometry directly from graph connectivity. Instead of generating full Cartesian coordinates in a single step, the method decomposes the problem into local predictions around individual bonds, using a graph attention network (GNN) to encode chemical environments and a rotationally invariant spherical representation to model bond distances and angular structure. To support this formulation, we construct an independently generated benchmark dataset of 9,691 DFT-optimized molecular structures using a large-scale Quantum ESPRESSO workflow initialized from PubChemQC molecular candidates. The resulting dataset provides rotationally consistent local geometric supervision for bond-centric molecular reconstruction while remaining independent from the original molecular optimization pipeline. On the benchmark set, the proposed framework achieves a heavy-atom bond-distance mean absolute error (MAE) of 0.103 Å and a root mean squared error (RMSE) of 0.152 Å, corresponding to an average relative bond-length error of approximately 7.37%. Local angular reconstruction achieves a polar-angle (θ) MAE of 4.56° and RMSE of 7.49°, while azimuthal-angle ($\Delta\phi$) prediction remains substantially more challenging, with an MAE of 18.19° and RMSE of 31.34°. Error analysis shows that bond lengths and local bond-angle magnitudes are reconstructed reliably across diverse chemical environments, whereas torsional and rotational degrees of freedom contribute the dominant source of geometric uncertainty, particularly in flexible single-bond conformations. Qualitative reconstruction results demonstrate strong preservation of molecular topology and local bonding environments, with successful extrapolation to polymeric systems significantly larger than those observed during training. From a computational perspective, the method exhibits predictable scaling with system size, following a power-law relationship of approximately $T \propto B^{1.98}$ with respect to bond count. In direct comparison, RDKit ETKDG-only and ETKDG + MMFF workflows exhibit steeper empirical scaling trends of approximately $T \propto B^{3.04}$ and $T \propto B^{2.99}$, respectively. These results indicate that the proposed bond-centric reconstruction strategy becomes increasingly advantageous for large polymer-scale molecular graphs. These results demonstrate that bond-centric local geometry prediction provides a scalable and physically consistent approach for rapid three-dimensional molecular structure generation, offering a practical alternative for high-throughput molecular modeling.

Scientific Contribution. GraphChem3D reformulates molecular three-dimensional reconstruction as a set of local, bond-centered geometric predictions rather than global coordinate generation. On an independently generated DFT benchmark, bond lengths and polar bond angles are recovered accurately (0.103 Å and 4.56° MAE), while azimuthal orientation remains substantially less accurate, isolating torsional degrees of freedom as the specific bottleneck limiting purely local geometry formulations. Because reconstruction cost is governed by bond count rather than global embedding, runtime scales near quadratically with bond number against near cubic empirical scaling for ETKDG based workflows, and reconstructions of polymer chains well beyond the training size range remain chemically plausible under qualitative inspection. The work therefore positions bond-centric prediction as a scalable initialization

*Co-first author. Email: MATTHEWO001@e.ntu.edu.sg

†Co-first author. Email: JCAHYONO001@e.ntu.edu.sg

‡Email: leonard.ngwt@ntu.edu.sg

step for geometry dependent simulation of large molecular graphs, rather than a replacement for conformer search or force field relaxation.

1 Introduction

Accurate prediction of molecular geometry is a foundational problem in computational chemistry because the three-dimensional arrangement of atoms governs the electronic structure, chemical reactivity, intermolecular interactions, and thermodynamic stability relevant to pharmaceuticals, catalysis, organic electronics, and materials science. Since the development of quantum mechanics, first-principles electronic structure methods have provided rigorous frameworks for predicting molecular properties directly from fundamental physical laws [2, 4]. Among these approaches, density functional theory (DFT) has become one of the most widely adopted methods due to its balance between predictive accuracy and computational efficiency through an electron-density formulation of the many-body problem [7, 11]. However, despite its success, DFT geometry optimization remains computationally expensive when applied to large molecular libraries or atomistically large systems, creating a major bottleneck for high-throughput molecular screening, polymer simulation, and large-scale dataset generation.

This bottleneck is increasingly important as materials discovery moves toward automated, data-rich, and self-driving workflows. In high-throughput experimental domains such as printed electronics and organic photovoltaics, advances in roll-to-roll processing, functional inks, and in-line characterization have enabled rapid generation of large experimental datasets, but have also exposed the difficulty of navigating high-dimensional materials and processing spaces [8, 17, 24, 18]. Similar challenges arise in computational molecular discovery, where the ability to rapidly generate physically reasonable three-dimensional structures can strongly affect the throughput of downstream property prediction, simulation, and optimization. Recent work on self-driving laboratories and digital twins for high-throughput photovoltaic manufacturing further illustrates the growing need for computational tools that can provide fast, reliable structure–property inputs within closed-loop optimization workflows [13, 6, 18, 12, 3]. In this context, molecular geometry generation is not only a classical computational chemistry problem, but also an enabling component for scalable AI-driven materials discovery.

Recent advances in machine learning have created new opportunities for rapid molecular geometry prediction. In particular, graph neural networks (GNNs) provide a natural framework for molecular learning because molecules can be represented as graphs where atoms form nodes and chemical bonds form edges. Existing approaches, including SchNet, directional message passing networks, autoregressive coordinate generators, learned gradient-field models, and diffusion-based geometric models, have demonstrated strong performance for molecular property prediction and small-molecule conformation generation [19, 10, 14, 21, 25]. Nevertheless, direct generation of full molecular geometry remains fundamentally challenging. Global Cartesian coordinate prediction introduces rotational and translational ambiguities, while the high dimensionality of molecular conformational space becomes increasingly difficult to model for large or flexible systems. These challenges are particularly pronounced for polymeric, organic-semiconductor, and atomistically large molecular structures, where existing generative approaches often exhibit substantial computational cost and reduced structural reliability. The main practical objective of this work is therefore not to replace high-accuracy quantum chemical relaxation, but to provide a fast geometry initialization and reconstruction method for large molecular graphs where conventional conformer generation and force-field refinement become increasingly expensive.

In this work, we propose GraphChem3D, a bond-centric framework for molecular geometry prediction that reformulates global three-dimensional reconstruction into a sequence of local geometric prediction problems. Rather than directly generating an entire molecular conformation in a single step, the proposed approach predicts physically meaningful local geometric structure around individual bonds and progressively reconstructs the full molecular geometry through graph-based propagation. By decomposing molecular structure into localized bond-centered environments, the framework reduces the complexity of the reconstruction problem while preserving chemically relevant structural information. This locality-driven formulation enables improved computational scalability and facilitates efficient reconstruction of larger molecular systems that are challenging for conventional global coordinate generation methods.

To encode the surrounding chemical environment of each target bond, the proposed framework employs a

graph attention network, allowing atom representations to be iteratively updated through attention-weighted message passing over neighboring molecular connectivity while incorporating bond-type information as edge features. The target geometry is represented using a rotationally invariant spherical formulation consisting of bond distances and local angular relationships. This representation improves physical interpretability while avoiding dependence on arbitrary Cartesian coordinate frames, thereby improving geometric consistency and learning stability during reconstruction.

As illustrated in Figure 1, the proposed workflow combines graph-based local environment encoding with rotationally invariant geometric prediction to reconstruct large-scale molecular structure from local representations centered on bonds. To enable supervised learning, the model is trained on an independently generated dataset of DFT-optimized molecular structures. Molecular candidates are initialized from the PubChemQC dataset and subsequently relaxed using a separate DFT workflow, ensuring that training and evaluation are performed on geometries generated from an independent computational pipeline. This strategy provides high-quality structural supervision while avoiding direct dependence on the geometries of the source data set.

The proposed formulation provides several important advantages. First, the bond-centric local representation captures chemically meaningful structure while avoiding direct prediction of high-dimensional global coordinate distributions. Second, the rotationally invariant geometric formulation improves physical consistency and reduces sensitivity to arbitrary coordinate frame transformations. Third, the locality-driven reconstruction strategy exhibits favorable computational scaling behavior, enabling efficient reconstruction of larger molecular and polymeric systems beyond the practical size range typically accessible to first-principles geometry optimization methods. This is particularly relevant for high-throughput and self-driving materials workflows, where geometric initialization may need to be repeated across large candidate libraries, parameter sweeps, or closed-loop optimization cycles [?, 18, 12, 3]. As demonstrated in the experimental results, the proposed framework achieves scalable molecular reconstruction with near power-law computational behavior while maintaining accurate local geometric prediction.

The main contributions of this work are as follows. First, we introduce a bond-centric graph neural network framework that reformulates molecular geometry generation as a localized reconstruction problem, enabling progressive molecular assembly from local geometric predictions. Second, we propose a rotationally invariant spherical geometric representation for predicting bond distances and local angular structure while avoiding direct dependence on global Cartesian coordinate frames. Third, we construct a large-scale independently generated DFT molecular geometry benchmark using over one million CPU core-hours of Quantum ESPRESSO simulations, together with a graph-to-coordinate alignment pipeline enabling localized geometric supervision for bond-centric molecular reconstruction. Finally, we demonstrate scalable molecular reconstruction behavior across large molecular and polymeric systems and quantitatively compare the runtime scaling against an ETKDG + MMFF reconstruction workflow, showing a more favorable empirical scaling trend for the proposed method.

The remainder of this paper is organized as follows. Section 2 surveys related work on quantum chemical methods, machine learning approaches for molecular geometry prediction, and high-throughput materials discovery contexts. Section 3 describes the dataset construction, preprocessing pipeline, molecular geometry representation, model architecture, and loss functions. Section 4 presents the training setup, evaluation metrics, and quantitative results. Section 5 discusses the findings and limitations of the proposed approach. Section 6 outlines directions for future work. Section 7 summarizes the contributions of this work.

2 Related Work

Between empirical force-field methods and first-principles quantum chemistry lies an important intermediate class of approaches known as semi-empirical methods. These methods simplify the electronic Hamiltonian through parameterized approximations while retaining an explicit quantum mechanical treatment of valence electrons. As a result, methods such as PM6 can provide substantially faster geometry optimization than conventional DFT while maintaining reasonable structural accuracy for many molecular systems [23]. Their favorable balance between computational cost and accuracy has enabled large-scale molecular data generation efforts, most notably the PubChemQC dataset, which reports B3LYP/6-31G* calculations initialized from PM6 geometries for approximately 86 million molecules from the PubChem database [15]. Nevertheless,

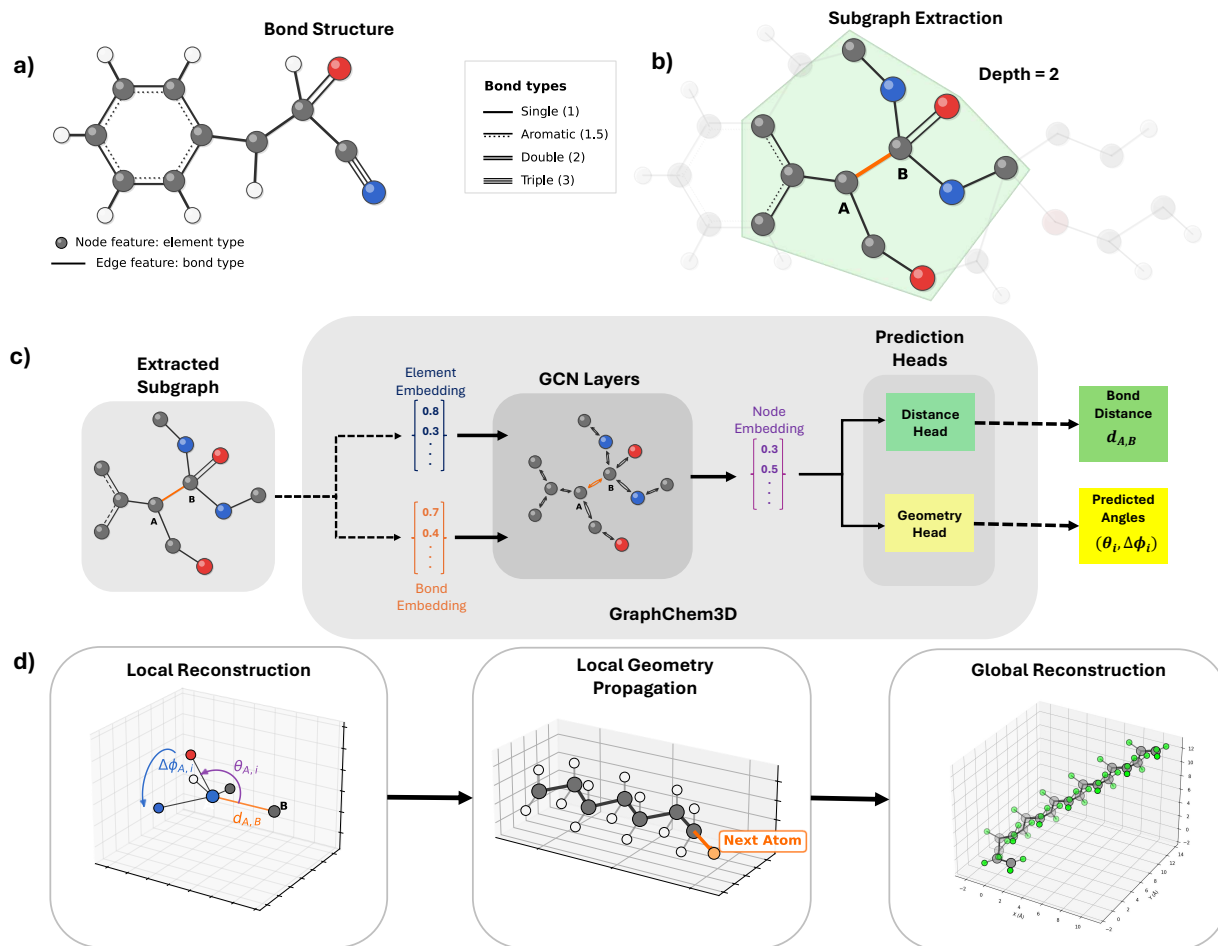


Figure 1: Overview of the proposed GraphChem3D framework for bond-centric molecular geometry prediction. (a) Molecules are represented as attributed graphs, where atoms are encoded as node features and bond types are encoded as edge features. (b) For each target bond, a bond-centered k -hop neighborhood is extracted to capture the surrounding local chemical environment. (c) The extracted subgraph is processed using edge-aware graph convolution layers to generate latent node embeddings, which are passed to separate prediction heads for bond-distance and local angular-geometry prediction. (d) Predicted bond lengths and local spherical geometric components are first assembled into local three-dimensional structures, which are sequentially propagated through the molecular graph to construct a globally reconstructed molecular conformation.

despite their reduced computational prefactor, semi-empirical approaches still rely on iterative self-consistent field procedures whose computational cost increases rapidly with system size. As a result, routine simulation of polymeric systems, organic semiconductor molecules, and atomistically large molecular structures remains computationally challenging.

The need for faster molecular and materials modeling is closely connected to broader developments in high-throughput materials discovery. In organic and printed electronic systems, high-throughput fabrication platforms have enabled large-scale exploration of ink formulations, deposition conditions, annealing parameters, and device architectures [8, 16, 9, 17]. Roll-to-roll manufacturing studies of organic and perovskite photovoltaics further demonstrate how material composition and processing conditions jointly determine device performance, producing optimization spaces that are too large for exhaustive experimental exploration [24, 18]. These developments have motivated the use of Bayesian optimization, digital twins, and self-driving laboratories to guide high-throughput experiments with fewer trials [20, 22, 5, 13, 6, 12]. However, such closed-loop workflows depend on the availability of fast computational representations and predictive models. For molecular systems, rapid three-dimensional reconstruction can therefore serve as an important bridge between graph-level molecular design and geometry-dependent simulation or property prediction.

Recent advances in machine learning have introduced alternative approaches for molecular structure prediction that bypass explicit quantum mechanical optimization during inference. Graph neural networks are particularly well suited to chemistry because molecules can naturally be represented as graphs where atoms form nodes and bonds form edges. Architectures such as SchNet and directional message passing networks have demonstrated strong performance for molecular property prediction and interatomic interaction modeling through learned graph-based representations [19, 10]. Building upon these developments, generative molecular geometry models including autoregressive coordinate generators, learned gradient fields, and diffusion-based frameworks have extended graph learning toward direct conformation generation [14, 21, 25].

Despite substantial progress, direct molecular geometry generation remains fundamentally difficult due to rotational and translational symmetries, the existence of multiple valid conformations, and the need to simultaneously model both local chemical structure and long-range geometric constraints. Many existing approaches operate directly on global Cartesian coordinate distributions, introducing additional complexity during both learning and generation. These limitations become increasingly pronounced for large or highly flexible molecular systems, where diffusion and autoregressive methods often exhibit significant computational overhead and reduced structural consistency. In particular, efficient reconstruction of polymeric, organic electronic, and atomistically large systems remains an open challenge for many current generative molecular geometry frameworks.

The present work addresses these limitations through a fundamentally different formulation of molecular geometry prediction. Rather than directly generating entire molecular conformations globally, the proposed framework decomposes reconstruction into localized bond-centric geometric prediction tasks. This locality-driven approach enables molecular structure to be reconstructed progressively from physically meaningful local environments while avoiding direct modeling of full high-dimensional coordinate distributions. By combining graph-based local environment encoding with rotationally invariant spherical geometry prediction, the proposed framework aims to provide scalable and computationally efficient molecular reconstruction suitable for larger molecular systems, polymeric structures, and high-throughput computational workflows.

3 Methodology

3.1 Dataset and Preprocessing

3.1.1 Data Source

The primary source of molecular candidates used in this work is the PubChemQC dataset reported by Nakata and Maeda, which provides large-scale quantum chemically optimized molecular structures for approximately 86 million compounds using B3LYP/6-31G* calculations initialized from PM6 geometries [15]. Each entry contains Cartesian atomic coordinates, elemental identities, and associated electronic structure information, making it a valuable source of initial molecular geometries and chemical metadata.

In this study, the PubChemQC structures were not used directly as training targets. Instead, they were used as initialization geometries for an independently generated benchmark dataset produced through our

own density functional theory (DFT) calculations. This allows model training and evaluation on structures obtained from a separate computational workflow while retaining access to large-scale curated molecular candidates.

3.1.2 Independent DFT Geometry Generation

To construct the benchmark dataset, we performed a large-scale independent DFT simulation campaign using Quantum ESPRESSO on the NSCC Aspire 2A high-performance computing cluster, consuming over one million cumulative CPU core-hours. Simulations employed the Perdew–Burke–Ernzerhof (PBE) exchange–correlation functional together with ultrasoft pseudopotentials (USPP family). These calculations provide an independently generated geometry dataset separate from the original PubChemQC workflow.

Geometry optimizations were performed using a standardized parameter set for all calculations to ensure consistency across the benchmark dataset. The initial atomic configurations used for each relaxation were taken from the PM6 structures provided through the source database. The job identifier for each calculation corresponds to its atomic CID. Detailed implementation of the automated workflow, pseudopotential library, and full simulation input parameters are provided in the public code repository described in the Code and Data Availability sections.

The independently simulated dataset contains 9,691 molecules that successfully converged during geometry optimization, with molecular sizes reaching up to 180 atoms. The composition spans a broad range of molecular sizes, bond counts, and elemental compositions. Summary statistics of the simulated molecules are presented in Figure 2.

3.1.3 Connectivity Mapping and Atom Correspondence

The independently relaxed geometries consist primarily of atomic identities and Cartesian coordinates, while supervised graph learning additionally requires explicit bond connectivity and consistent atom indexing. Therefore, bond topology information was obtained from the corresponding PubChem records and aligned with the optimized coordinate structures.

Because coordinate files do not explicitly encode bonds, connectivity was inferred through a hybrid procedure combining interatomic distance criteria with expected atomic degree information obtained from the PubChem molecular graph. Graph isomorphism matching was then performed between the reference PubChem connectivity graph and the inferred coordinate-based graph to establish one-to-one atom correspondence.

This procedure ensures consistent mapping between graph nodes and optimized atomic coordinates, enabling reliable supervision for node-level and bond-level geometric prediction tasks. Additional implementation details and edge-case handling procedures are provided in the accompanying GitHub repository.

3.1.4 Streaming Data Pipeline

A streaming dataset pipeline was implemented to load molecular data incrementally in chunks rather than storing the full corpus in memory. Each CSV file is treated as one training epoch, after which files are cycled sequentially. An LRU-based chunk cache with configurable RAM limits is used to minimize redundant disk access and improve data throughput.

Only molecules with valid geometries, successful relaxations, and verified graph-to-structure correspondence are retained for training. To ensure reproducibility, dataset partitioning is performed using deterministic random seeds, and 5-fold cross-validation is applied at the file level.

3.1.5 Local Subgraph Extraction

For each bond (A, B) in a molecule, we extract a local subgraph capturing the chemical context around both endpoints. Specifically, we perform a BFS/DFS from each of A and B up to a configurable depth (default: 1 hop) and take the union of the two neighborhoods. This dual-extraction strategy correctly handles both:

- **Cyclic structures** (e.g., rings): The union captures the full local cycle.
- **Linear chains**: The union captures two hops on both sides of the bond.

All edges within the subgraph are preserved, along with their bond type attributes.

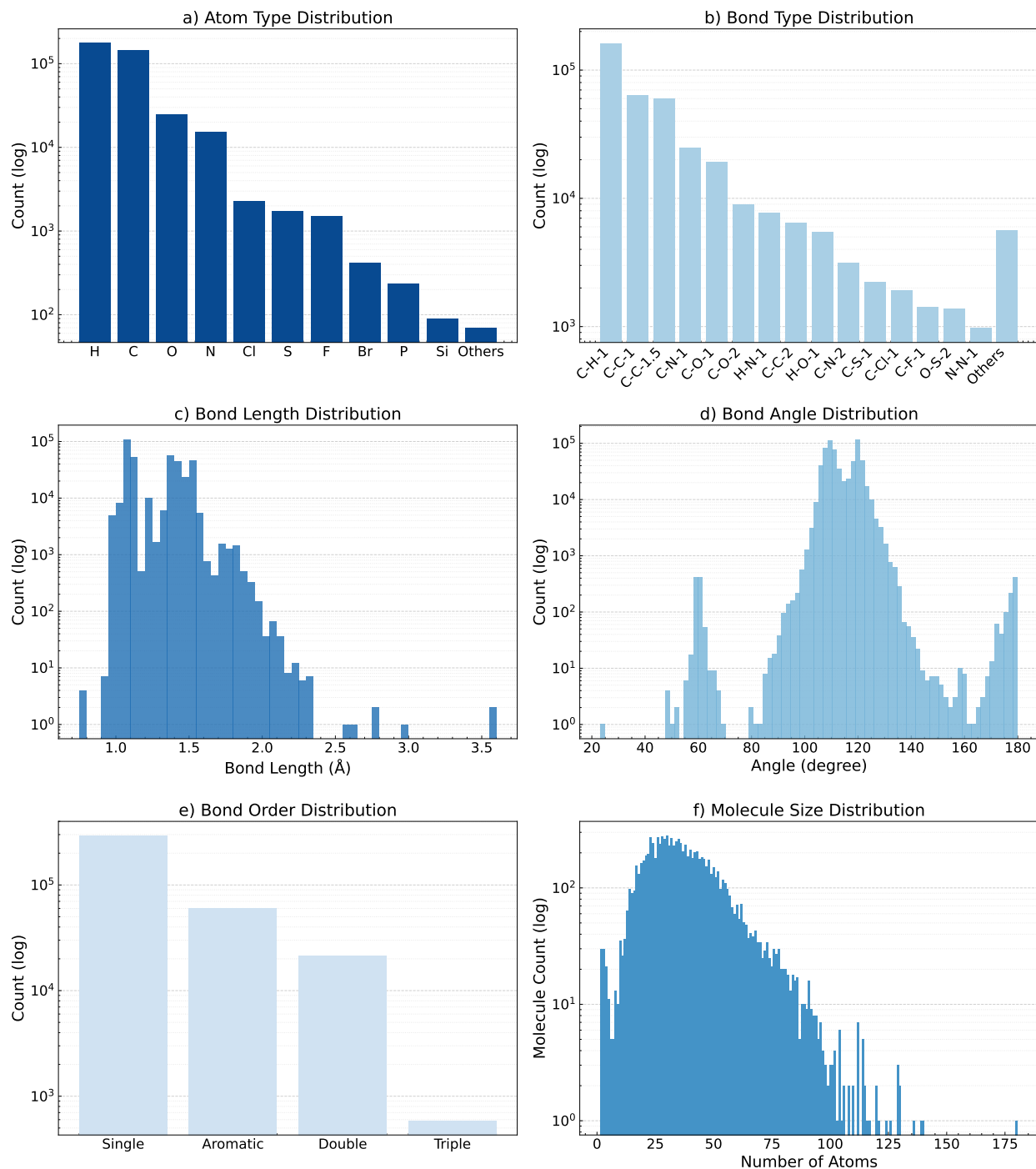


Figure 2: Statistical overview of the independently generated DFT molecular geometry benchmark used for bond-centric geometry learning and reconstruction. The benchmark was constructed from independently relaxed Quantum ESPRESSO simulations initialized from PubChemQC molecular candidates and spans a broad range of molecular sizes, bonding environments, bond orders, and angular configurations. Shown are the distributions of (a) atom type frequency, (b) bond type frequency, (c) bond length distribution, (d) bond angle distribution, (e) bond order distribution including aromatic bonds, and (f) molecule size distribution measured by total atom count per molecule. Together, these statistics illustrate the chemical and structural diversity of the benchmark dataset and its suitability for localized graph-based molecular geometry supervision.

3.1.6 Geometry Representation

Rather than predicting raw Cartesian coordinates (which are not rotationally invariant), we represent local geometry in a spherical coordinate frame anchored at each bond endpoint. For atom A with neighbors $\{n_1, n_2, \dots\}$, we place $n_1 = B$ (the bond partner) as the polar axis and compute for each neighbor n_i :

- θ_i : Polar angle between the $A \rightarrow B$ axis and the $A \rightarrow n_i$ direction, $\theta_i \in [0^\circ, 180^\circ]$.
- $\Delta\phi_i$: Relative azimuthal difference from the previous neighbor, $\Delta\phi_i \in [0^\circ, 360^\circ]$. The first neighbor (by a deterministic element/bond-order sort) has $\Delta\phi_1 = 0$ by definition.

The same representation is computed independently for atom B , using A as the polar reference. This yields a rotationally and translationally invariant description of local geometry. During training, a random global rotation is applied to each molecule for augmentation, testing the invariance of the representation.

Neighbors are sorted deterministically by element symbol and bond order to ensure a consistent canonical ordering across random rotations. The target distance d_{AB} is simply the Euclidean distance between atoms A and B in the DFT geometry.

3.2 Model Architecture

3.2.1 Input Encoding

Each atom in the subgraph is represented by a learned element embedding of dimension x , indexed by the atom’s element symbol (with a special **OTHER** token for rare elements). Bond types are encoded using a separate learned embedding of dimension x . The element vocabulary is constructed by scanning the dataset to identify all unique element symbols.

3.2.2 Graph Attention Network Encoder

The core of the model is a stack of GATv2 [1] layers, which use dynamic attention that conditions on both source and target node features, fixing the static-attention limitation of the original GAT formulation. Each layer uses 4 attention heads with averaged (non-concatenated) outputs, producing hidden representations of dimension 128. Bond type embeddings are passed as native edge features to each GATv2 layer via the `edge_dim` parameter, so that attention scores between atom i and atom j condition directly on the bond type of edge (i, j) . This allows the model to distinguish, for example, a stiff double bond from a flexible single bond when computing neighborhood attention.

Residual connections are added at each layer to prevent over-smoothing: the input to layer l is added to the output of the GATv2 operation. A linear projection aligns dimensions for the first-layer residual; subsequent layers use identity residuals. Batch normalization and dropout (rate 0.2) are applied after each layer for regularization.

The encoder is configurable in depth (number of layers), allowing the receptive field to be extended beyond the immediate neighbors when needed.

3.2.3 Prediction Heads

Given the node embeddings $\mathbf{h}_A$ and $\mathbf{h}_B$ for the bond endpoints after GNN encoding:

Distance head. The concatenated embedding $[\mathbf{h}_A \parallel \mathbf{h}_B]$ is passed through a two-layer MLP with a Softplus output activation to predict the positive bond distance $\hat{d}_{AB}$:

$$\hat{d}_{AB} = \text{Softplus}(\text{MLP}_{\text{dist}}([\mathbf{h}_A \parallel \mathbf{h}_B])).$$

Geometry head. Each bond endpoint is decoded through a two-stage pipeline. First, a two-layer MLP encoder takes $[\mathbf{h}_{\text{self}} \parallel \mathbf{h}_{\text{partner}}]$ as input—incorporating cross-atom context so that geometry predictions around A are conditioned on the nature of B and vice versa—and produces a slot representation $\mathbf{s} \in \mathbb{R}^{d_{\text{slot}}}$. Separate encoder networks are used for atoms A and B to capture their distinct roles in the bond. Second, a learnable positional embedding $\mathbf{e}_i$ for each bond slot $i \in \{0, \dots, \text{max_bonds} - 1\}$ is added to the slot representation:

$$\mathbf{r}_{X,i} = \mathbf{s}_X + \mathbf{e}_i,$$

allowing the model to learn slot-specific geometric biases (e.g., slot 0 has $\Delta\phi = 0$ by definition; sp^2 vs. sp^3 patterns differ across slots). A shared single-layer decoder then maps each slot representation to a $(\theta, \Delta\phi)$ pair:

$$\hat{\theta}_{X,i} = 180 \cdot \sigma(\mathbf{W}\mathbf{r}_{X,i} + b)_\theta, \quad \hat{\Delta\phi}_{X,i} = 360 \cdot \sigma(\mathbf{W}\mathbf{r}_{X,i} + b)_\phi.$$

Sigmoid is preferred over clamping because it maintains nonzero gradients across the full output range. The same decoder weights $(\mathbf{W}, b)$ are shared between atoms A and B , since the role differences are already captured by the separate encoders.

3.3 Loss Functions

3.3.1 Distance Loss

We use a scale-invariant percentage error loss for the bond distance:

$$\mathcal{L}_{\text{dist}} = \frac{1}{N} \sum_{n=1}^N \frac{|\hat{d}_n - d_n|}{|d_n| + \epsilon},$$

where $\epsilon = 10^{-3}$ prevents division by zero. This formulation treats prediction errors proportionally, so a 10% error on a 1 Å bond contributes the same as a 10% error on a 2 Å bond.

3.3.2 Angular Geometry Loss

For polar angles $\theta \in [0^\circ, 180^\circ]$:

$$\mathcal{L}_\theta = \frac{1}{|\mathcal{M}|} \sum_{(n,i) \in \mathcal{M}} \frac{|\hat{\theta}_{n,i} - \theta_{n,i}|}{180},$$

where $\mathcal{M}$ indexes valid (non-padded) neighbor slots. For relative azimuthal differences $\Delta\phi \in [0^\circ, 360^\circ]$:

$$\mathcal{L}_{\Delta\phi} = \frac{1}{|\mathcal{M}|} \sum_{(n,i) \in \mathcal{M}} \frac{|\hat{\Delta\phi}_{n,i} - \Delta\phi_{n,i}|}{360}.$$

The normalization factors (180 and 360) place both angle losses in the range $[0, 1]$, making them commensurable with each other and with the distance loss.

3.3.3 Total Loss

The per-node geometry loss combines the polar and azimuthal components for all valid neighbors $\mathcal{M}_X$ of node $X \in \{A, B\}$ using configurable component weights w_θ and w_ϕ :

$$\mathcal{L}_X = \frac{1}{w_\theta + w_\phi} \left(\frac{w_\theta}{|\mathcal{M}_X|} \sum_{i \in \mathcal{M}_X} \frac{|\hat{\theta}_{X,i} - \theta_{X,i}|}{180} + \frac{w_\phi}{|\mathcal{M}_X|} \sum_{i \in \mathcal{M}_X} \frac{|\hat{\Delta\phi}_{X,i} - \Delta\phi_{X,i}|}{360} \right).$$

The division by $w_\theta + w_\phi$ averages (rather than sums) the two normalized components, keeping $\mathcal{L}_X$ in the range $[0, 1]$ regardless of the chosen weights. By default $w_\theta = w_\phi = 1$, giving equal contribution from polar and azimuthal errors. The total loss then averages the distance and geometry components with configurable weights (w_d, w_g) :

$$\mathcal{L}_{\text{total}} = \frac{w_d \cdot \mathcal{L}_{\text{dist}} + w_g \cdot \frac{1}{2}(\mathcal{L}_A + \mathcal{L}_B)}{w_d + w_g},$$

where $\mathcal{L}_A$ and $\mathcal{L}_B$ are the per-node geometry losses for bond endpoints A and B respectively, averaged to treat both endpoints symmetrically. By default we set $w_d = 1, w_g = 2$.

3.4 Molecular Reconstruction

Given a trained model, full three-dimensional molecular coordinates are recovered from predicted local geometry through a three-phase pipeline: BFS-based atom placement, angle-conditioned position prediction, and local constraint-based position refinement.

3.4.1 Phase 1: BFS-Based Atom Placement

Reconstruction begins by selecting a *center atom*: the atom with the highest degree in the molecular graph (ties broken by index). The center atom is placed at the coordinate origin and its first neighbor is placed along the x -axis at the model-predicted bond distance, establishing a fixed coordinate frame for the rest of the molecule.

The remaining atoms are then placed in breadth-first order. At each step, a *center atom* C with at least one already-placed neighbor is dequeued. The model is queried for bond (C, R) , where R is the first placed neighbor of C , producing predicted angles $(\hat{\theta}_i, \hat{\Delta}\phi_i)$ for each neighbor slot i of C . A local orthonormal coordinate frame $(\mathbf{e}_x, \mathbf{e}_y, \mathbf{e}_z)$ is constructed with $\mathbf{e}_z$ aligned to the $C \rightarrow R$ bond vector.

Neighbors of C are sorted deterministically by element symbol and bond order to produce a canonical slot ordering consistent with training. For each unplaced neighbor n_i at slot i , the cumulative azimuthal angle is accumulated as:

$$\Phi_i = \sum_{k=1}^i \hat{\Delta}\phi_k,$$

where $\hat{\Delta}\phi_1 = 0$ by definition (the first neighbor anchors the azimuthal reference). The predicted spherical coordinates $(r_i, \hat{\theta}_i, \Phi_i)$ are then converted to a Cartesian displacement via:

$$\mathbf{p}_{n_i} = \mathbf{p}_C + r_i \left(\sin \hat{\theta}_i \cos \Phi_i \mathbf{e}_x + \sin \hat{\theta}_i \sin \Phi_i \mathbf{e}_y + \cos \hat{\theta}_i \mathbf{e}_z \right),$$

where r_i is obtained from a separate model query for bond (C, n_i) . The newly placed atom n_i is then appended to the BFS queue.

To improve robustness, a *multi-start* strategy is used: the BFS placement is repeated from two different starting atoms (the two highest-degree atoms), producing two independent coordinate sets. The second reconstruction is rigidly aligned to the first using the Kabsch algorithm, and the two sets are averaged atom-wise to produce a single consensus initial geometry.

3.4.2 Phase 2: Angle Prediction for Refinement

Before global refinement, pairwise bond-angle targets are collected from the model. For each atom C with at least two neighbors, and for each distinct ordered pair of neighbors (n_1, n_2) , the model is queried for bond (C, n_1) and the predicted polar angle $\hat{\theta}_{n_2}$ of n_2 in C 's local frame is recorded as an approximate bond-angle target:

$$\alpha_{n_1, C, n_2} \approx \hat{\theta}_{n_2}.$$

These targets supplement the bond-distance constraints during the subsequent optimization step.

3.4.3 Phase 3: Local Constraint-Based Position Refinement

BFS placement is inherently greedy: each atom is positioned using only the local predictions available at the moment it is dequeued, so small errors accumulate along the chain. Phase 3 corrects this by treating all atom positions simultaneously as a single optimization problem — nudging every atom at once until the full set of predicted bond lengths and bond angles is jointly satisfied as closely as possible.

The consensus geometry from Phase 1 is refined by minimizing a constraint objective over the positions of all free atoms using L-BFGS-B. Three anchor atoms (chosen to be non-collinear and highly connected) are held fixed to remove the global translation and rotation degrees of freedom. The objective function combines relative distance errors and angle errors:

$$\mathcal{E}(\mathbf{P}) = \sum_{(i,j) \in \mathcal{B}} \left(\frac{\|\mathbf{p}_i - \mathbf{p}_j\| - \hat{r}_{ij}}{\hat{r}_{ij} + \epsilon} \right)^2 + \lambda \sum_{(n_1, C, n_2) \in \mathcal{A}} (\alpha(\mathbf{P})_{n_1, C, n_2} - \alpha_{n_1, C, n_2})^2,$$

where $\mathcal{B}$ is the set of all bonds, $\mathcal{A}$ is the set of collected angle triples, $\alpha(\mathbf{P})_{n_1,C,n_2}$ is the bond angle computed from current positions, and $\lambda = 0.3$ is the angle constraint weight. Analytical gradients of $\mathcal{E}$ with respect to free atom positions are provided to the optimizer. The final refined positions are returned as the reconstructed molecular geometry.

4 Experiment Results

4.1 Training Setup

The model was trained using mini-batch gradient-based optimization within the streaming dataset framework described above. Validation performance was monitored throughout training, and the checkpoint achieving the best validation loss was retained for final evaluation. Five-fold cross-validation and deterministic random seeds were used to ensure robust and reproducible assessment. Detailed training hyperparameters, optimizer settings, learning-rate schedules, early-stopping criteria, and implementation-specific parameters are provided in the public code repository described in the Code and Data Availability section.

4.1.1 Hardware Configuration

All model training and evaluation experiments were conducted on a desktop system running Microsoft Windows 11 Pro (Version 10.0.26100). The system is equipped with an Intel(R) Core(TM) Ultra 9 285K processor (24 cores, 24 logical processors) operating at 3.70 GHz, 64 GB of installed system memory, and an NVIDIA GeForce RTX 5090 GPU. The system is built on a Gigabyte Z890 AORUS ELITE WIFI7 platform. Model training was accelerated using the GPU, while data preprocessing and auxiliary operations were performed on the CPU. The available memory and compute resources were sufficient to support the streaming dataset pipeline and mini-batch processing without requiring distributed training.

4.2 Evaluation Metrics

We report the following metrics on the held-out validation set:

- Mean Absolute Error (MAE) and Root Mean Square Error (RMSE) for bond distances (in Å).
- MAE and RMSE for polar angles θ (in degrees).
- MAE and RMSE for relative azimuthal differences $\Delta\phi$ (in degrees).

4.3 Results

The proposed graph-based geometry prediction framework demonstrates stable convergence during training and produces physically consistent molecular reconstructions across a wide range of chemical environments.

4.3.1 Training Convergence

Model optimization remained stable throughout training, with both training and validation losses decreasing substantially before gradually plateauing, as shown in Figure 3. The best validation performance was achieved at epoch 49, where the validation loss reached 1.893×10^{-2} . At this checkpoint, the validation distance loss reached 1.139×10^{-2} , while the validation geometry loss reached 2.270×10^{-2} . The corresponding training losses at this stage were 2.403×10^{-2} for the total loss, 2.115×10^{-2} for the distance component, and 2.547×10^{-2} for the geometry component.

The consistently decreasing validation losses indicate strong generalization on held-out molecular structures without evidence of severe overfitting. Bond-distance prediction converged more rapidly and to lower loss values than angular reconstruction, suggesting that local distance relationships are easier to learn than rotational geometric features. This trend is consistent with the final evaluation metrics, where bond lengths achieved substantially lower prediction error compared to azimuthal angular components. In particular, the geometry-related losses plateaued at higher values due to the increased complexity of predicting torsional and rotational degrees of freedom, especially for flexible single-bond conformations. Overall, the training dynamics demonstrate stable optimization behavior and effective learning of both bond-length and local structural information required for molecular geometry reconstruction.

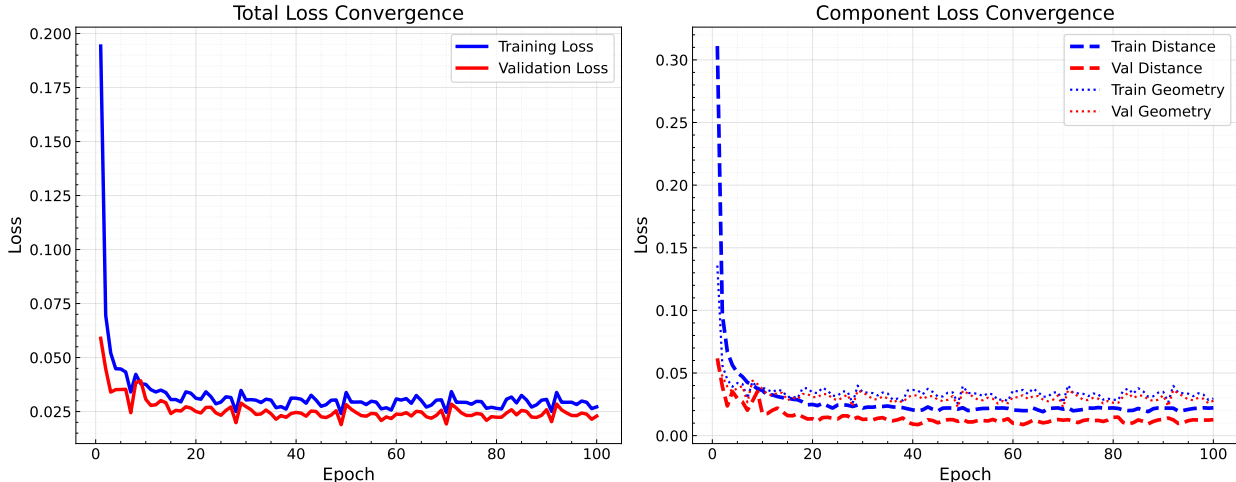


Figure 3: Training and validation convergence behavior over 100 epochs. Left: total training and validation loss. Right: distance and geometry loss components for both training and validation sets.

4.3.2 Bond Length Accuracy

Bond Type	Count	MAE ($\text{\AA}$)	RMSE ($\text{\AA}$)
O-C (2.0)	9,006	0.3395	0.3459
N-C (2.0)	3,136	0.2892	0.3018
N-C (1.0)	24,902	0.2134	0.2305
O-C (1.0)	19,168	0.2097	0.2262
C-C (1.0)	62,165	0.0654	0.0736
C-C (2.0)	5,595	0.0513	0.0652
C-C (1.5)	62,407	0.0119	0.0155

Table 1: Bond length prediction errors for representative high-frequency heavy-atom bond types.

A detailed evaluation over 199,505 heavy-atom bonds yields a mean absolute error (MAE) of $0.1033 \text{ \AA}$ and a root mean squared error (RMSE) of $0.1518 \text{ \AA}$. The average relative error is 7.37%, indicating generally strong predictive accuracy for bond distances across diverse chemical environments. Aromatic C-C bonds exhibit particularly low error (MAE: $0.0119 \text{ \AA}$ over 62,407 samples), followed by C=C double bonds (MAE: $0.0513 \text{ \AA}$), suggesting that rigid conjugated bonding environments are reconstructed especially well. In contrast, larger deviations are observed for polar and chemically asymmetric bonds such as O=C and N=C double bonds, with MAEs of $0.3395 \text{ \AA}$ and $0.2892 \text{ \AA}$, respectively. Single bonds involving heteroatoms, including O-C and N-C, also exhibit elevated error relative to carbon-only bonds. These results suggest that the model reliably captures common carbon bonding environments while struggling more with electronically polarized or chemically heterogeneous bonds, as summarized in Table 1.

4.3.3 Angular Geometry Accuracy

Local angular structure was evaluated over 547,710 heavy-atom angles. The polar angle (θ) prediction achieves an MAE of 4.56° and an RMSE of 7.49° , indicating accurate reconstruction of local bond-angle geometry. In contrast, the azimuthal component ($\Delta\phi$) exhibits substantially higher error, with an MAE of 18.19° and an RMSE of 31.34° , reflecting the increased difficulty of learning rotational degrees of freedom. Unlike bond angles, which are strongly constrained by local hybridization and steric geometry, azimuthal orientations are highly sensitive to torsional flexibility and conformational variability, particularly around rotatable single bonds where multiple low-energy configurations may exist. In addition, the relative angu-

Angle Type	Count	θ MAE ($^{\circ}$)	θ RMSE ($^{\circ}$)
O-C-N (1.0-2.0)	4,521	10.96	11.27
N-C-O (1.0-2.0)	4,521	10.62	10.96
O-C-O (1.0-2.0)	7,654	10.13	10.48
N-C-N (1.0-2.0)	3,518	9.83	11.14
O-C-C (1.0-2.0)	10,068	8.86	10.24
C-C-O (1.0-2.0)	10,068	7.58	8.87
N-C-C (1.0-2.0)	5,640	6.87	8.11
C-C-N (1.0-2.0)	5,640	6.39	7.61
C-N-C (1.0-1.0)	30,516	6.24	8.51
C-C-C (1.0-1.0)	95,260	5.67	8.32

Table 2: Polar angle (θ) prediction errors for representative high-frequency heavy-atom local angle types.

Angle Type	Count	$\Delta\phi$ MAE ($^{\circ}$)	$\Delta\phi$ RMSE ($^{\circ}$)
C-C-C (1.0-1.0)	95,260	47.99	53.16
C-C-O (1.0-1.0)	16,334	46.29	52.65
C-C-N (1.0-1.0)	17,786	43.04	49.89
N-C-C (1.0-1.0)	17,786	42.57	50.50
O-C-C (1.0-1.0)	16,334	42.33	50.11
C-N-C (1.0-1.0)	30,516	16.91	27.09
O-C-O (1.0-2.0)	7,654	7.95	8.39
O-C-N (1.0-2.0)	4,521	7.43	8.40
N-C-O (1.0-2.0)	4,521	5.35	6.96
N-C-N (1.0-2.0)	3,518	5.03	5.90

Table 3: Azimuthal angle ($\Delta\phi$) prediction errors for representative high-frequency heavy-atom local angle types.

lar representation introduces sensitivity to neighbor ordering and angular periodicity, which can amplify prediction error even for geometrically similar configurations.

Error analysis reveals strong dependence on local chemical environment, as summarized in Tables 2 and 3. For θ prediction, the largest errors are primarily associated with environments adjacent to double bonds, such as O-C-N (1.0–2.0), O-C-O (1.0–2.0), and N-C-N (1.0–2.0), where MAEs exceed 9° . These motifs are typically associated with planar conjugated geometries, where small deviations in local arrangement can produce larger angular discrepancies.

In contrast, $\Delta\phi$ prediction errors are dominated by flexible single-bond carbon backbone environments. Common motifs such as C-C-C (1.0–1.0), C-C-O (1.0–1.0), and C-C-N (1.0–1.0) exhibit very large azimuthal deviations, with MAEs approaching 48° . This behavior is consistent with the large torsional flexibility associated with rotatable single bonds, where multiple conformations may be energetically accessible. Meanwhile, environments adjacent to double bonds exhibit substantially lower $\Delta\phi$ error, reflecting the stronger rotational constraints imposed by conjugation and partial bond-order rigidity. Overall, the results indicate that the model captures local bond-angle magnitudes reasonably well, while rotational alignment and torsional degrees of freedom remain the primary challenge in angular reconstruction.

4.3.4 Small Molecule Reconstruction Study

In addition to quantitative evaluation metrics, qualitative reconstruction tests were performed on representative small molecular systems to assess the model’s ability to recover chemically meaningful local geometry and structural motifs. Figure 4 presents reconstructed geometries for (a) acetylsalicylic acid (aspirin) and (b) cyclohexane. These examples were selected to probe distinct structural characteristics, including aromatic conjugation, heteroatom-containing functional groups, and non-planar cyclic conformations.

The reconstructed acetylsalicylic acid structure preserves the aromatic benzene ring topology, ester linkage, and carboxylic acid functional group geometry with good qualitative consistency. In particular, the planar conjugated ring environment and relative arrangement of oxygen-containing substituents remain chemically plausible, indicating that the model successfully captures both local bonding structure and broader geometric organization in heteroatom-rich molecular systems. Meanwhile, the cyclohexane reconstruction correctly reproduces the characteristic non-planar six-membered ring conformation rather than collapsing into an unrealistic planar geometry. The preservation of ring closure and approximate chair-like spatial arrangement suggests that the bond-centric representation effectively encodes local geometric constraints necessary for cyclic molecular reconstruction.

Together, these examples demonstrate that the proposed framework is capable of reconstructing chemically diverse small molecules while maintaining realistic local geometry, aromatic connectivity, and ring structure organization. The results further support the ability of the learned local geometric representation to generalize across both rigid conjugated systems and flexible non-planar molecular motifs.

4.3.5 Polymer Extrapolation Study

Although the model was trained primarily on relatively small molecular systems, qualitative extrapolation tests were also performed on substantially larger polymeric structures to assess generalization behavior beyond the training distribution. Figure 5 presents representative reconstructions for (a) polytetrafluoroethylene (PTFE), (b) polyethylene glycol (PEG), (c) polyester, and (d) polystyrene. The reconstructed polymer structures preserve chemically plausible local geometry, continuous backbone connectivity, and characteristic structural motifs associated with each polymer class. In particular, aromatic ring topology in polystyrene and fluorinated backbone geometry in PTFE remain well preserved despite the significantly larger system sizes compared to the training molecules.

The results suggest that the bond-centric formulation generalizes effectively to extended molecular systems by leveraging local chemical environments rather than relying on fixed global molecular templates. Since reconstruction is performed through local bond-centered predictions, the framework remains applicable to polymer-scale systems substantially larger than those encountered during training. While these extrapolation results are qualitative, they demonstrate that the learned local geometric representation captures transferable chemical information sufficient to maintain realistic large-scale molecular structure organization.

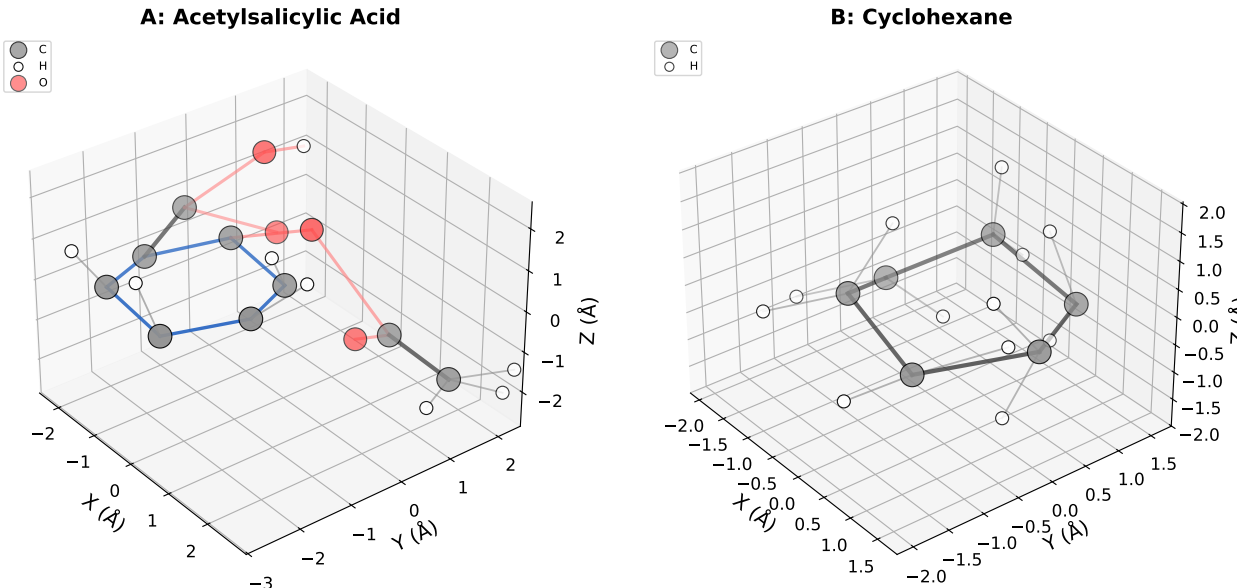


Figure 4: Qualitative reconstruction results for representative small molecular systems: (a) acetylsalicylic acid and (b) cyclohexane. The reconstructed geometries preserve aromatic ring topology, functional group organization, and non-planar cyclic conformations, demonstrating accurate local structural reconstruction across chemically distinct molecular motifs.

4.3.6 Computational Efficiency and Baseline Scaling Comparison

In addition to geometric accuracy, the computational efficiency of the proposed method was evaluated by measuring reconstruction time as a function of molecular size. Figure 6 compares the scaling behavior of GraphChem3D against two RDKit-based baseline workflows: ETKDG-only conformer generation and ETKDG followed by MMFF force-field optimization. For the RDKit baselines, ETKDGv3 was used to generate an initial conformer from the polymer molecular graph. In the ETKDG + MMFF workflow, this conformer was subsequently refined using MMFF force-field optimization with a fixed maximum iteration limit. ETKDG embedding success, MMFF convergence status, and wall-clock runtime were recorded separately. Molecular size is represented by the number of bonds, which provides a direct graph-based measure of reconstruction complexity.

For GraphChem3D, reconstruction time increases smoothly with system size and follows a near-quadratic trend. A linear fit in log-log space yields a scaling relationship of approximately $T \propto B^{1.98}$, indicating that the computational cost grows close to quadratically with the number of bonds. This behavior is consistent with the bond-centric formulation of the method, where reconstruction is governed primarily by repeated local bond-centered predictions rather than global conformer embedding or iterative force-field relaxation. A formal worst-case complexity analysis showing why the algorithm scales as $\Theta(N^2)$ is provided in Supplementary Section S1.

In contrast, both RDKit-based workflows exhibit substantially steeper empirical scaling. ETKDG-only conformer generation follows an approximate relationship of $T \propto B^{3.04}$, while the combined ETKDG + MMFF workflow follows a similar near-cubic trend of approximately $T \propto B^{2.99}$. The similarity between these two exponents indicates that the initial ETKDG embedding step already contributes a strong size-dependent computational cost, while MMFF refinement adds additional runtime overhead. In several larger flexible polymer cases, MMFF optimization did not fully converge within the fixed iteration limit, further emphasizing the practical difficulty of force-field refinement for large molecular systems.

This comparison clarifies the main computational advantage of GraphChem3D. The proposed method is not intended to replace high-accuracy quantum chemical optimization or force-field refinement when fully relaxed equilibrium structures are required. Instead, its strength lies in rapidly generating chemically

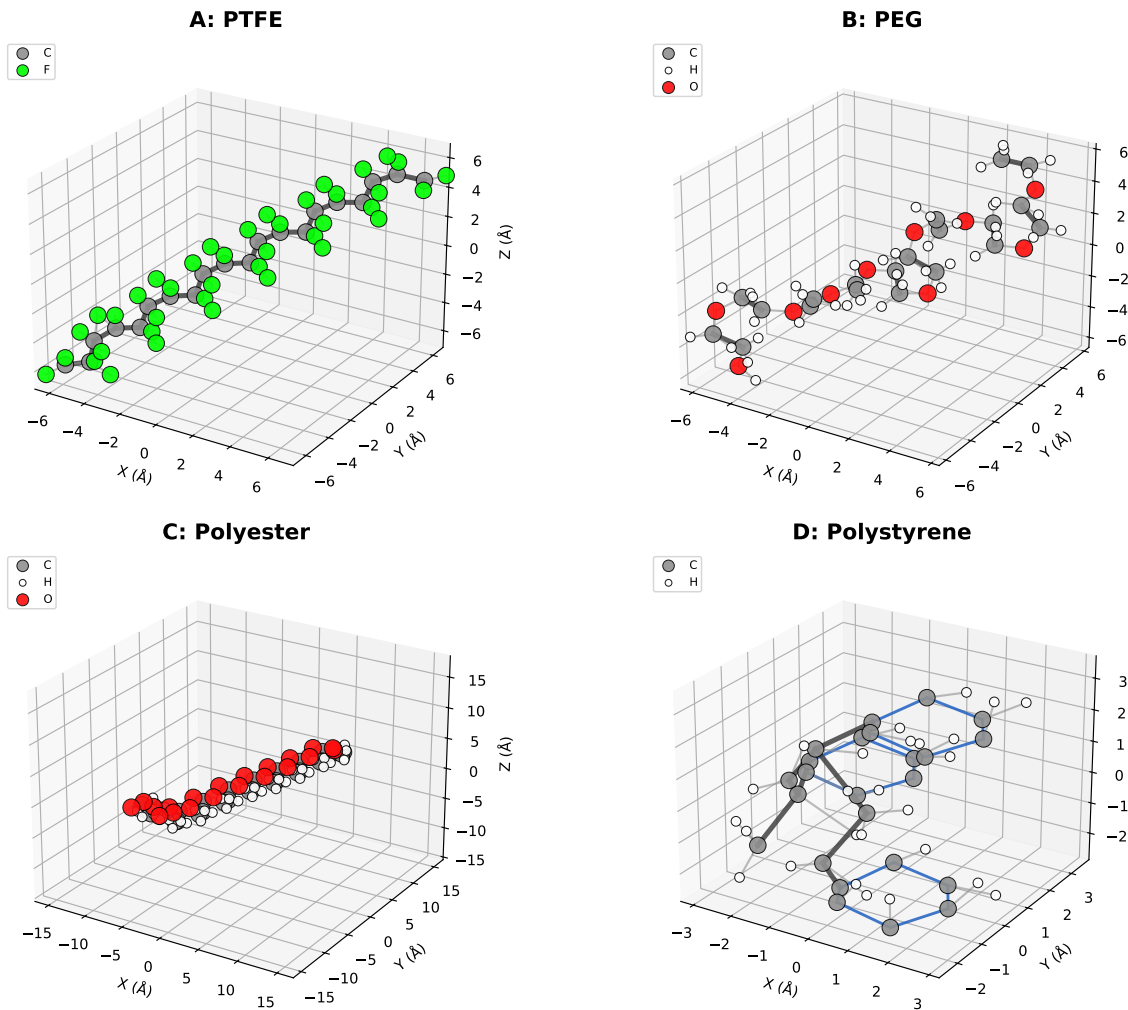


Figure 5: Qualitative extrapolation results on larger polymer systems: (a) polytetrafluoroethylene (PTFE), (b) polyethylene glycol (PEG), (c) polyester, and (d) polystyrene. The reconstructed structures preserve chemically consistent local geometry, continuous backbone organization, and characteristic polymer-specific structural motifs.

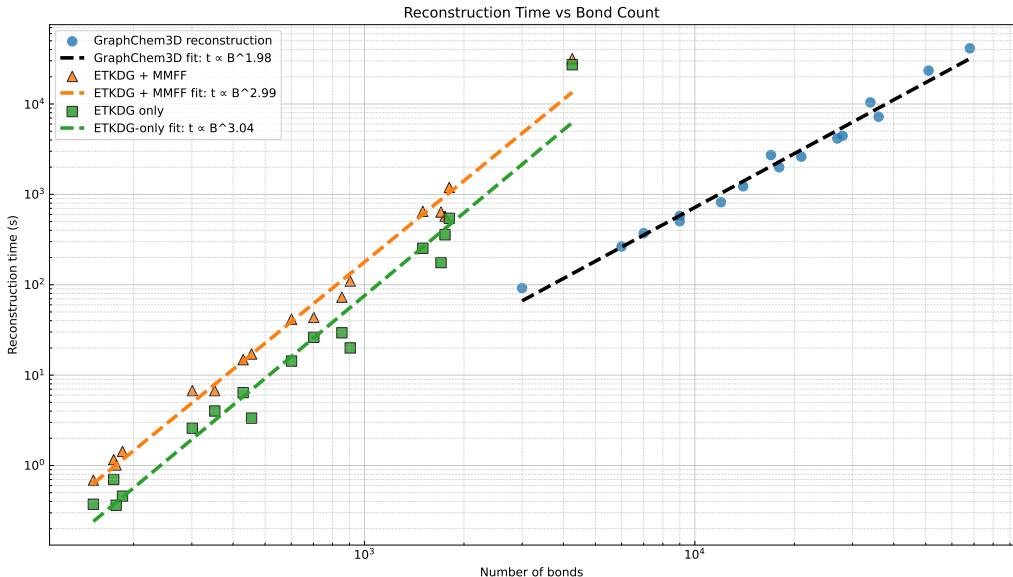


Figure 6: Reconstruction time as a function of molecular size, measured by the number of bonds, for GraphChem3D, RDKit ETKDG-only conformer generation, and ETKDG followed by MMFF force-field optimization. GraphChem3D exhibits near-quadratic scaling with a fitted relationship of approximately $T \propto B^{1.98}$, while ETKDG-only and ETKDG + MMFF show steeper empirical trends of approximately $T \propto B^{3.04}$ and $T \propto B^{2.99}$, respectively. The comparison shows that the proposed reconstruction method has more favorable scaling behavior for large polymer systems.

plausible initial geometries for large molecular graphs. While all tested methods become slower as molecular size increases, the lower scaling exponent of GraphChem3D makes it more tractable for high-throughput molecular generation and polymer-scale structural reconstruction, particularly in regimes where conventional ETKDG-based and force-field-based workflows become increasingly expensive.

5 Discussion and Limitations

5.1 Discussion

The results demonstrate that bond-centric graph-based modeling provides an effective framework for predicting local molecular geometry while maintaining computational efficiency. The model achieves strong accuracy in bond-length and polar angle prediction, indicating that local chemical environments are well captured through graph attention mechanisms and edge-conditioned message passing. However, the larger errors observed in azimuthal components highlight a fundamental challenge associated with rotational degrees of freedom. Unlike bond angles, which are strongly constrained by local hybridization, azimuthal orientations are sensitive to torsional flexibility, conformational variability, and the choice of local reference frame. These factors introduce ambiguity in the target representation and make precise prediction inherently more challenging. Despite this, qualitative reconstruction results show that the predicted structures preserve overall molecular topology and chemically plausible geometry, and the model successfully generalizes to polymeric systems significantly larger than those seen during training. This suggests that the bond-centric decomposition effectively separates local geometric prediction from global structural complexity.

From a computational perspective, the method exhibits predictable scaling with system size, following a power-law relationship of approximately $T \propto B^{1.98}$ with respect to bond count. In direct comparison, a conventional ETKDG + MMFF reconstruction workflow exhibits a steeper empirical scaling of approximately $T \propto B^{2.99}$, indicating that the proposed bond-centric reconstruction strategy becomes increasingly advantageous for large polymer-scale molecular graphs.

The comparison with ETKDG + MMFF clarifies the main practical advantage of the proposed approach. GraphChem3D is not positioned as a replacement for quantum chemical optimization or force-field refinement when high-accuracy equilibrium structures are required. Instead, its advantage lies in rapid generation of chemically plausible initial geometries for large molecular graphs. Because the learned model predicts local bond-centered geometry directly, reconstruction avoids the increasingly expensive distance-geometry embedding and iterative force-field refinement steps that dominate conventional workflows for large flexible polymers. This makes the method most useful as a scalable initialization or high-throughput reconstruction tool for polymer-scale molecular systems.

5.2 Limitations

Several methodological and scope limitations should be noted. First, the reconstruction procedure is inherently sequential and greedy: each bond geometry is predicted independently and then propagated through the molecular graph. Consequently, errors in early bond predictions may accumulate and propagate to downstream atoms, potentially degrading global reconstruction quality for longer molecular chains or highly branched structures. This limitation is expected to be particularly significant for highly rigid systems, where local geometric inconsistencies cannot be easily relaxed through conformational flexibility.

Second, the model predicts a single deterministic conformation for each molecule rather than an ensemble. For flexible molecules with multiple low-energy conformers - particularly those containing rotatable single bonds - the predicted structure represents only one possible geometry, and the current framework does not provide a mechanism for sampling or quantifying conformational diversity. This remains a fundamental limitation of the present formulation.

Third, the training dataset remains limited in chemical scope. The 9,691 successfully converged molecules contain at most 180 atoms and are dominated by a relatively narrow subset of elements (C, H, O, N, F, S, and Cl). Predictions involving heavier or less common elements, unusual coordination environments, or substantially larger molecular systems should therefore be interpreted cautiously. This limitation is expected to be particularly pronounced for highly specific or underrepresented functional groups that are insufficiently covered within the training dataset.

Finally, future studies should extend the present RDKit ETKDG and ETKDG + MMFF runtime-scaling comparison into a broader conformer-quality benchmark, including standardized local-geometry validity, steric clash, and ensemble-diversity metrics.

6 Future Work

Future work should focus on improving rotational consistency and reducing error accumulation during sequential reconstruction, particularly for highly rigid molecular systems where local geometric inaccuracies may propagate more strongly through the molecular graph. Additional directions include incorporating longer-range structural interactions through deeper or global message-passing mechanisms, integrating global geometric constraints to improve overall conformational fidelity, and extending the framework toward probabilistic or ensemble-based conformation generation for flexible molecules with multiple low-energy states. Expanding the training dataset to include broader chemical diversity and more specialized functional groups would further improve generalization across underrepresented chemistries. Finally, future studies should include direct benchmarking against established conformation-generation approaches such as RDKit ETKDG and diffusion-based molecular geometry models to better characterize the accuracy, scalability, and computational efficiency trade-offs of the proposed framework. These developments could further establish bond-centric local geometry prediction as a scalable alternative to conventional molecular structure generation methods.

7 Conclusion

We have presented a bond-centric, rotationally invariant framework for predicting local three-dimensional molecular geometry using graph neural networks. By decomposing the conformation problem into local predictions around individual bonds, the model achieves strong accuracy in bond-length and bond-angle

prediction while maintaining computational scalability. The results demonstrate that the proposed approach can reliably reconstruct molecular structures and generalize to larger polymeric systems beyond the training distribution. In particular, the observed near power-law scaling with system size highlights the practical advantage of the method for large-scale molecular modeling, where conventional quantum chemistry methods become computationally prohibitive. At the same time, the work identifies key challenges in modeling rotational degrees of freedom, particularly in the azimuthal component, where conformational variability and representation sensitivity limit predictive accuracy. Addressing these limitations will be critical for further improving geometric fidelity.

Code and Data Availability

The code, processed dataset, configuration files, and trained model checkpoint associated with this work have been uploaded to [Zenodo](#). Details of the deposited files, reproducibility workflow, and benchmark regeneration procedure are provided in Supplementary Section S2.

Author Contributions

M.O. and J.A.C. contributed equally to this work. J.A.C. developed the model architecture and implemented the training pipeline. M.O. performed data extraction, conducted DFT simulations, and carried out data processing and analysis of results. L.W.T.N. supervised the project and contributed to the interpretation of results. All authors contributed to writing and reviewing the manuscript.

Funding information

L.W.T.N. acknowledges funding from Singapore MOE Tier 2 grant (MOE-T2EP50125-0020).

References

- [1] Shaked Brody, Uri Alon, and Eran Yahav. How attentive are graph attention networks? In *International Conference on Learning Representations*, 2022.
- [2] P. A. M. Dirac. Quantum mechanics of many-electron systems. *Proceedings of the Royal Society A*, 123:714–733, 1929.
- [3] Sayan Doloi, Maloy Das, Yujia Li, Zen Han Cho, Xingchi Xiao, John V. Hanna, Matthew Osvaldo, and Leonard Ng Wei Tat. Democratizing self-driving labs: Advances in low-cost 3d printing for laboratory automation. *Digital Discovery*, 4(7):1685–1721, 2025.
- [4] V. Fock. Näherungsmethode zur lösung des quantenmechanischen mehrkörperproblems. *Zeitschrift für Physik*, 61:126–148, 1930.
- [5] Peter I Frazier. A tutorial on bayesian optimization. *ArXiv (Cornell University)*, 07 2018.
- [6] Kedar Hippalgaonkar, Qianxiao Li, Xiaonan Wang, John W. Fisher, James Kirkpatrick, and Tonio Buonassisi. Knowledge-integrated machine learning for materials: Lessons from gameplaying and robotics. *Nature Reviews Materials*, 8:241–260, 04 2023.
- [7] P. Hohenberg and W. Kohn. Inhomogeneous electron gas. *Physical Review*, 136:B864–B871, 1964.
- [8] Guohua Hu, Joohoon Kang, Leonard W. T. Ng, Xiaoxi Zhu, Richard C. T. Howe, Christopher G. Jones, Mark C. Hersam, and Tawfique Hasan. Functional inks and printing of two-dimensional materials. *Chem. Soc. Rev.*, 47:3265–3300, 2018.

- [9] Guohua Hu, Lisong Yang, Zongyin Yang, Yubo Wang, Xinxin Jin, Jie Dai, Qing Wu, Shouhu Liu, Xiaoxi Zhu, Xiaoshan Wang, and et al. A general ink formulation of 2d crystals for wafer-scale inkjet printing. *Science Advances*, 6(33), Aug 2020.
- [10] Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *International Conference on Learning Representations*, 2020.
- [11] W. Kohn and L. J. Sham. Self-consistent equations including exchange and correlation effects. *Physical Review*, 140:A1133–A1138, 1965.
- [12] Andre K. Low, Jayce J. Cheng, Kedar Hippalgaonkar, and Leonard W. Ng. Self-driving laboratories: Translating materials science from laboratory to factory. *ACS Omega*, 10(28):29902–29908, Jun 2025.
- [13] B. P. MacLeod, F. G. L. Parlane, T. D. Morrissey, F. Häse, L. M. Roch, K. E. Dettelbach, R. Moreira, L. P. E. Yunker, M. B. Rooney, J. R. Deeth, V. Lai, G. J. Ng, H. Situ, R. H. Zhang, M. S. Elliott, T. H. Haley, D. J. Dvorak, A. Aspuru-Guzik, J. E. Hein, and C. P. Berlinguette. Self-driving laboratory for accelerated discovery of thin-film materials. *Science Advances*, 6, 05 2020.
- [14] Elman Mansimov, Omar Mahmood, Seokho Kang, and Kyunghyun Cho. Molecular geometry prediction using a deep generative graph neural network. *Scientific Reports*, 9(1):1–13, 2019.
- [15] Maho Nakata and Toshiyuki Maeda. Pubchemqc b3lyp/6-31g*//pm6 data set: The electronic structures of 86 million molecules using b3lyp/6-31g* calculations. *Journal of Chemical Information and Modeling*, 63(18):5734–5754, 2023. PMID: 37677147.
- [16] Leonard W. Ng, Guohua Hu, Richard C. Howe, Xiaoxi Zhu, Zongyin Yang, Christopher G. Jones, and Tawfique Hasan. Printing technologies. *Printing of Graphene and Related 2D Materials*, page 135–178, Jul 2018.
- [17] Leonard W. T. NG, Seok Woo Lee, Dong Wook Chang, Justin M. Hodgkiss, and Doojin Vak. Organic photovoltaics’ new renaissance: Advances toward roll-to-roll manufacturing of non-fullerene acceptor organic photovoltaics. *Advanced Materials Technologies*, 7(10):2101556, 2022.
- [18] Leonard Wei Ng, Na Gyeong An, Liu Yang, Yinhua Zhou, Dong Wook Chang, Jueng-Eun Kim, Luke J. Sutherland, Tawfique Hasan, Mei Gao, and Doojin Vak. A printing-inspired digital twin for the self-driving, high-throughput, closed-loop optimization of roll-to-roll printed photovoltaics. *Cell Reports Physical Science*, 5(6):102038, Jun 2024.
- [19] Kristof T. Schütt, Pieter-Jan Kindermans, Huziel E. Sauceda, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. SchNet: A continuous-filter convolutional neural network for modeling quantum interactions. In *Advances in Neural Information Processing Systems*, volume 30, 2017.
- [20] Bobak Shahriari, Kevin Swersky, Ziyu Wang, Ryan P. Adams, and Nando de Freitas. Taking the human out of the loop: A review of bayesian optimization. *Proceedings of the IEEE*, 104:148–175, 01 2016.
- [21] Chence Shi, Shitong Luo, Minkai Xu, and Jian Tang. Learning gradient fields for molecular conformation generation. In *International Conference on Machine Learning*, 2021.
- [22] Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. Practical bayesian optimization of machine learning algorithms. *ArXiv (Cornell University)*, 01 2012.
- [23] James J. P. Stewart. Optimization of parameters for semiempirical methods v: Modification of nddo approximations and application to 70 elements. *Journal of Molecular Modeling*, 13(12):1173–1213, 2007.
- [24] Hasitha Weerasinghe, Nasiruddin Macadam, Jueng-Eun Kim, Luke Sutherland, Dechan Angmo, Leonard Ng, Andrew Scully, Fiona Glenn, Regine Chantler, Nathan Chang, and et al. The first demonstration of entirely roll-to-roll fabricated perovskite solar cell modules under ambient room conditions. *Nature Communications*, 15(1), Mar 2024.

- [25] Minkai Xu, Lantao Yu, Yang Song, Chence Shi, Stefano Ermon, and Jian Tang. GeoDiff: A geometric diffusion model for molecular conformation generation. In *International Conference on Learning Representations*, 2022.