CHEM4DLLM: 4D MULTIMODAL LLMS FOR CHEMICAL DYNAMICS UNDERSTANDING

Anonymous authors
Paper under double-blind review

ABSTRACT

Existing chemical understanding tasks primarily rely on static molecular representations, limiting their ability to model inherently dynamic phenomena such as bond breaking or conformational changes, which are essential for a chemist to understand chemical reactions. To address this gap, we introduce **Chem**ical **D**ynamics **U**nderstanding (ChemDU), a new task that translates 4D molecular trajectories into interpretable natural-language explanations. ChemDU focuses on fundamental dynamic scenarios, including simple molecular motion, gas-phase reactions and catalytic reactions, and requires models to reason about key events along molecular trajectories, such as bond formation and dissociation, and to generate coherent, mechanistically grounded narratives. To benchmark this capability, we construct **Chem4DBench**, the first dataset pairing 4D molecular trajectories with explanations across these settings. We further develop a suite of 4D molecular language models, including 4D Text-based, **4D-MolT5** and **Chem4DLLM**. Notably, Chem4DLLM is a trajectory-aware model that injects per-atom, per-frame geometry directly into the token stream of a pretrained large language model. This model suite enables a systematic evaluation of how different 4D representations and architectures support dynamic chemical understanding. We hope that ChemDU, together with Chem4DBench and our 4D model suite, will stimulate further research in dynamic chemical understanding and multimodal scientific reasoning. Our code is publicly available at https://anonymous.4open.science/r/Chem4DLLM-617E.

1 INTRODUCTION

The advancement of large language models (LLMs) (Achiam et al., 2023; Chen et al., 2024) has triggered a wave of innovation, moving these models from processors of human language to powerful reasoning engines for complex scientific data (AI4Science & Quantum, 2023; Wei et al., 2025; Xin et al., 2025). In chemistry and biology, this has sparked a new frontier of molecule-text modeling. Early efforts successfully unified textual knowledge with 1D molecular sequences (like SMILES (Weininger, 1988) and SELFIES (Krenn et al., 2022; Christofidellis et al., 2023) and/or 2D molecular graphs (Luo et al., 2023) enabling powerful models for tasks like molecule captioning and text-based retrieval (Edwards et al., 2021). More recently, research has pushed into incorporating static 3D structures (Li et al., 2024; Pei et al., 2025), recognizing that a molecule's geometry is critical to its function. However, these static snapshots fail to capture the true nature of biomolecular processes, which are fundamentally dynamic. Indeed, all chemical phenomena are dynamic since a molecule can be truly static only at absolute zero temperature, which does not exist in any known environment in the universe.

Computational chemistry simulations such as Molecular Dynamics (MD) (Herschbach, 1992) simulates such dynamic processes, generating massive spatio-temporal datasets that describe the fundamental mechanisms of biology and materials science. These trajectories, which can be conceptualized as dynamic 4D point clouds—where atomic coordinates (3D) evolve with time steps (1D)—hold the key to understanding processes like drug binding, protein folding, and catalytic reactions. Currently, such data is primarily interpreted through low-dimensional, quantitative metrics such as root-mean-square deviation or radial distribution functions. While useful, these metrics are insufficient; they fail to capture the qualitative, event-driven "story" of the chemical process. The true scientific insight lies in a high-level narrative: for example, "the inhibitor first docks into the

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

active site, inducing a subtle conformational change in a nearby loop, which in turn allosterically closes the binding pocket." Automatically generating such qualitative descriptions is not only fundamental for human scientific insight but also crucial for agentic scientific discovery, where an agent is expected to plan, execute, and refine in-silico simulations based on the trajectory evolution. However, translating these vast, high-dimensional, and unstructured spatio-temporal data into coherent narratives presents a formidable challenge for the machine learning and computational chemistry communities.

The challenge comes from two sides. On the data side, while using AI and machine learning to accelerate quantum chemistry simulations is a community focus and has generated plentiful molecular, protein, and materials 3D geometries or 4D trajectories (Ramakrishnan et al., 2014; Jain et al., 2013; Chanussot et al., 2021; Tran et al., 2023), corresponding text descriptions are rare. ChEBI-20 (Edwards et al., 2021) and 3D-MoIT (Li et al., 2024) are two datasets that do contain 3D geometry and a text description, but both of them are focused on describing static properties, therefore lacking descriptions of dynamic events. Another issue is that both datasets are designed for non-Periodic-Boundary-Condition (non-PBC) molecular systems, thereby lacking the capability to model PBC systems such as crystals and catalytic systems.

Beyond the data-side challenges, there are also unique challenges on the model side. First, the input is a high-dimensional, unstructured sequence of point clouds, how to encode such discrete graph data into LLM. Second, the sheer scale of the data is prohibitive for standard sequence models. A naive texturalization of Cartesian coordinates for hundreds of atoms over one hundred steps would generate sequences that vastly exceed the context window of most of LLM. Third, similar to video LLMs, a 4D molecular LLM must track state evolution over time, identify actions and events, model long-range temporal dependencies, and reason about causal relationships between events, such as recognizing that a conformational change is triggered by an earlier binding event. These requirements go far beyond the static state description typical of 3D molecular models.

To address these challenges, we make the following contributions:

- We introduce **ChemDU**, a chemical dynamics understanding task that translates 4D molecular trajectories into scientifically grounded text by identifying dynamic events and their temporal progression.
- We construct **Chem4DBench**, the first benchmark for 4D molecular trajectory understanding, covering gas-phase reactions, heterogeneous catalysis, and molecular motion to probe spatio-temporal reasoning, transition states, PBC systems, and temporal grounding.
- We develop a suite of **4D molecular language models**, including 4D Text-based, 4D-MolT5, and **Chem4DLLM**, where Chem4DLLM injects per-atom, per-frame graph embeddings into the token stream of a pretrained LLM, with coordinates expressed in a center-of-mass frame.

Together, ChemDU, Chem4DBench, and our 4D model suite provide a foundation for interpretable, temporally grounded molecular reasoning.

2 Related Work

Chemistry Large Language Models The application of LLMs (Achiam et al., 2023; Yang et al., 2025b; Ren et al., 2025) to chemistry has been progressed using difference level of molecular representation, from 1D strings to 2D graphs and 3D geometries. Early works represented molecules as SMILES (Weininger, 1988) or SELFIES (Krenn et al., 2022), adapting NLP architectures, especially T5-based models (Raffel et al., 2020), for property prediction and molecule–text generation (Edwards et al., 2021; 2022). Representative models include MolT5 (Edwards et al., 2022), ChemT5 (Christofidellis et al., 2023), BioT5 (Pei et al., 2023), BioT5$^+$ (Pei et al., 2024), and ChemLLM (Zhang et al., 2024), with further instruction-tuning efforts such as Mol-Instructions (Fang et al., 2024) and LlaSMol (Yu et al., 2024). However, 1D representations suffer from intrinsic limitations, including many-to-one SMILES mappings, sequence ambiguity, and invalid molecular generations.

To address these issues, 2D graph-based molecular LLMs incorporate GNNs or graph encoders to inject topological information. Cross-modal methods such as MoleculeSTM (Liu et al., 2023a),

MolFM (Luo et al., 2023), MolCA (Liu et al., 2023b), and MolX (Le et al., 2024) align molecular graphs with text, while UniMoT (Guo et al., 2024) tokenizes 2D graphs through vector quantization. ChemVLM (Li et al., 2025) instead converts 2D molecular structures into images for pretrained VLMs. More recently, 3D-aware models have been developed. MolBind (Xiao et al., 2024) aligns 2D graph, 3D structure, and language encoders; 3D-MoLM (Li et al., 2024) equips an LLM with an external Uni-Mol encoder (Zhou et al., 2023); and 3D-MolT5 uses E3FP (Axen et al., 2017) as 3D features. Beyond encoder-based designs, Jiang et al. (2025) introduces SDF2Text to directly serialize 3D coordinates as text, similar in spirit to CrystalLLM (Gruver et al., 2024), which generates crystal structures in CIF format. Despite recent advancements, most existing frameworks remain tailored for static structural analysis and lack a unified mechanism to encode the temporal evolution in 4D molecular trajectories.

Spatio-temporal Language Benchmarks The task of translating continuous spatio-temporal data into natural language has been studied beyond chemistry, most notably in 3D human motion captioning. Early resources such as KIT Motion-Language (Plappert et al., 2016) paired 3D pose sequences with simple textual descriptions, while larger benchmarks such as HumanML3D (Guo et al., 2022), built from motion repositories including AMASS (Mahmood et al., 2019), provide extensive 3D human motions represented by body models or joint coordinates with language annotations. However, these datasets mainly describe holistic actions of a single agent. In the broader context, video-language understanding is a mature field supported by large-scale benchmarks, including MSR-VTT (Xu et al., 2016) and VATEX (Wang et al., 2019). These datasets, however, pair 2D video clips with textual annotations. Consequently, they operate on 2D projections of the 3D world, lacking the precise, high-dimensional coordinate data inherent to our domain.

Representation Learning for 4D and 3D Data Machine learning for molecular modeling has transitioned from feature engineering, such as Coulomb Matrices (Montavon et al., 2012), toward the end-to-end GNN models. This field progressed from invariant models (e.g., SchNet (Schütt et al., 2017; Schütt et al., 2018) and DimeNet (Klicpera et al., 2020)), to equivariant models (e.g., NequIP (Batzner et al., 2022) and MACE (Batatia et al., 2022)). These equivariant architectures, which respect the necessary physical symmetry of rotation, have achieved state-of-the-art results in predicting quantum-chemical properties from static 3D geometries. More recently, universal atomic models—including M3GNet (Chen & Ong, 2022), MACE-MP (Batatia et al., 2023), and UMA (Wood et al., 2025)—have emerged. These models aim to be trained on massive, diverse datasets to provide a general-purpose potential that does not require task-specific fine-tuning. However, the primary goal of these GNNs is simulation rather than interpretation; they are trained to predict atomic forces to generate a trajectory at a much faster speed than quantum chemistry methods.

3 ChemDU Task and Chem4DBench

3.1 Problem Definition for ChemDU Task

We define Chemical Dynamics Understanding (ChemDU) as the task of translating a 4D molecular trajectory into a concise, scientifically grounded natural-language description of the underlying chemical processes. Formally, a trajectory is represented as

$$\mathcal{T} = \{\mathcal{X}_t\}_{t=1}^{T}, \tag{1}$$

where each frame $\mathcal{X}_t = (\mathbf{A}, \mathbf{C}_t)$ consists of a fixed set of atoms $\mathbf{A}$ and their 3D Cartesian coordinates $\mathbf{C}_t \in \mathbb{R}^{N\times 3}$ at time step t. The ChemDU task learns a mapping

$$f : \mathcal{T}, \mathcal{P} \to \mathcal{Y}, \tag{2}$$

where $\mathcal{P}$ denotes the task prompt and $\mathcal{Y}$ is the target natural-language explanation. Unlike static molecule captioning, ChemDU requires temporal grounding and event-level abstraction: models must identify salient events such as motion, bond breaking/formation, adsorption, or conformational change, determine when they occur, and describe their mechanistic progression.

3.2 Chem4DBench

Chem4DBench is a benchmark for evaluating LLMs on 4D dynamic trajectories, focusing on reaction trajectories with explicit transition states, temporal evolution, and periodic boundary conditions.

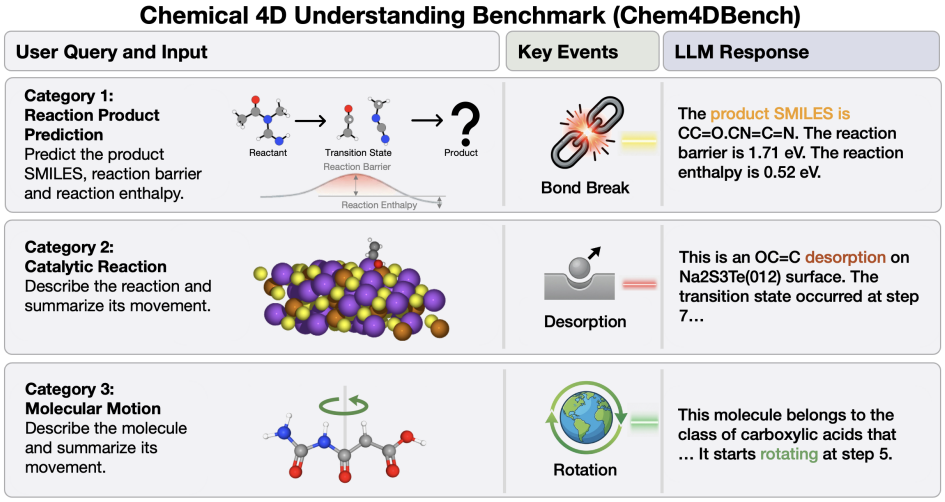


Figure 1: Overview of the Chem4D benchmark. This benchmark encompasses three distinct categories: (1) Reaction Product Prediction, involving the analysis of bond breaking/forming events and reaction barriers (derived from Transition1x and RGD1); (2) Catalytic Reaction, covering complex surface interactions such as desorption (derived from OC20). (3) Molecular Motion, augmenting 3D-MolT dataset with deterministic motion. For each category, the figure illustrates the workflow from the user query and 4D input, through the identification of key physical events, to the generation of the final scientific narrative.

It consists of three complementary categories: gas-phase reaction products prediction, catalytic reactions and molecular motion detection—designed to probe different aspects of spatio-temporal and chemical reasoning.

Reaction Products Prediction Chemical reactions are inherently 4D processes, involving atomic rearrangements from reactants through high-energy transition states (TS) to products. Conventional reaction benchmarks such as USPTO-50K (Seidl et al., 2022; Lowe, 2012), largely derived from patent data (Lowe, 2012), model reactions as reactant-to-product mappings using 2D graphs or SMILES. However, identical reactants can yield different products depending on the reaction pathway and conditions. Such endpoint-based representations overlook essential 3D information, including steric effects and the spatial orientation of functional groups.

To address this limitation, we formulate Reaction Product Prediction as a 4D task: given reactant and TS geometries, the model predicts the product and reaction energetics, including the reaction barrier (ΔE) and enthalpy (ΔH). Incorporating TS geometry encourages the model to reason about atomic migration and bond rearrangement along the reaction pathway rather than relying solely on statistical correlations. We construct this task using two complementary datasets: Transition1x (Schreiner et al., 2022) ($\sim$10k reactions) and RGD1 (Zhao et al., 2023) ($\sim$170k reactions), both providing reactant, TS, and product geometries alongside reaction energetics. Evaluation includes an in-distribution (ID) test set for Transition1x and both ID and three out-of-distribution (OOD) splits for RGD1: OOD-Reactants, OOD-Products, and OOD-Both. These OOD splits are constructed using scaffold splitting (Bemis & Murcko, 1996) to assess generalization to unseen chemical structures and reaction pathways.

Catalytic Reactions A problem of previous 3D-based LLMs is that they are designed for molecules therefore can not be used for periodic boundary systems such as crystals and catalysts. To addrss this limitation, we also designed the catalytic reactions category, which aims to evaluate a model's ability to understand complex 4D molecular dynamics in heterogeneous catalysis, where reactions occur at solid surfaces (with PBC constraint) and involve strong coupling between adsorbates and the catalyst lattice. This subset is constructed based on the OC20-NEB dataset (Wander et al., 2025), an extension of the Open Catalyst 2020 dataset (Chanussot et al., 2021) that provides NEB trajectories for surface reactions. OC20-NEB explicitly models transition pathways for three elementary surface reactions—transfer, dissociation, and desorption. To overcome the limited scale of OC20-NEB ($\sim$ 700 samples), we expand this subset with approximately 6,000 additional surface-reaction trajectories. This expansion is started by sampling catalyst–adsorbate reactant states from the OC20

dataset (Chanussot et al., 2021). Then we randomly assign a reaction from transfer, dissociaiton and desorption to each sample and re-simulate it using UMA model (Wood et al., 2025). In this category, the model is required to identify the reaction event, i.e., transfer, desorption, or dissociation, as well as the adsorbate and catalyst involved in the trajectory. Beyond event recognition and trajectory narration, the model is also required to identify reaction-critical frames and predict key physical properties, including the transition-state frame, reaction barrier, and reaction enthalpy.

Molecular Motion An important application of 4D LLMs is agentic simulation, where an LLM-based agent is expected to interpret simulated trajectories and take appropriate actions based on dynamic molecular states. To support this capability, we introduce a Molecular Motion category, which evaluates whether LLMs can understand elementary molecular motions such as translation and rotation. We reuse the 3D-MolT dataset introduced in 3D-MoLM (Li et al., 2024), which contains approximately 15,000 3D structure–text caption pairs, and extend it into 4D by applying deterministic geometric transformations. The actions include global translation, partial movement/bond breaking, global rotation, partial rotation, bending, vibration, and explosion. Except for vibration, these trajectories are not generated by molecular dynamics simulations and do not rely on force fields or potential energy surfaces. This subset provides a controlled test of whether a model can recognize explicit geometric actions and identify their starting time, while retaining the original 3D captioning task.

Evaluation We evaluate models on three Chem4DBench categories using metrics that jointly assess language quality, structural correctness, and physical fidelity. Tab. 1 summarizes the predicted targets and evaluation metrics for each category, with further details provided in Appx. B. For gas-phase reactions, we ask the model to predict the final product SMILES, and we measure product SMILES accuracy as well as the regression error for the reaction barrier and enthalpy. For catalytic reactions, predicting the final product is less meaningful (e.g., transfer and desorption can yield the same product); therefore, we ask the model to identify the transition-state step, as well as the reaction barrier and enthalpy. We report reaction type accuracy, adsorbate and product SMILES accuracy, and energetic MAE. For Molecular Motion, we evaluate the original 3D captioning quatiliy, as well as motion recognition accuracy and temporal grounding.

Category	Predicted targets	Metrics
Reaction Product Prediction	Product SMILES, ΔE, ΔH	BLEU, exact match, Levenshtein distance, validity, MACCS/RDK/Morgan similarity, and MAE for $\Delta E/\Delta H$.
Catalytic Reaction Understanding	Reaction type, adsorbate, product, TS step, ΔE, ΔH	Type accuracy, adsorbate/product exact match and Morgan similarity, TS Step MAE, and MAE for $\Delta E/\Delta H$.
Molecular Motion Understanding	Molecule caption, motion type, motion starting step	BLEU-2/4, ROUGE-1/2/L, METEOR, motion accuracy, and Step MAE.

Table 1: Overview of evaluation tasks and metrics in Chem4DBench.

4 Model Architecture

Our model, built upon the Qwen3-8B architecture (Yang et al., 2025a), is a multimodal LLM capable of processing and reasoning about 4D molecular trajectories. A key design choice is to use equivariant information: in 4D trajectories, rigid-body rotation and reorientation are meaningful dynamics, and invariant features discard this information, making distinct rotational states indistinguishable. Our architecture consists of three key components: (1) an graph encoder that translates each 3D frame into embeddings, (2) a graph projector, and (3) the Qwen3-8B LLM backbone.

Graph Encoder and Atom Representation. The input to our model is a 4D trajectory $\mathcal{T}$ represented as a sequence of atomic systems (frames). Each frame contains N atoms. For each frame, we use UMA (Wood et al., 2025), a pretrained equivariant GNN, to extract invariant $l = 0$ features that describe the local chemical environment of each atom. To explicitly retain equivariant geometric information, we additionally provide the Cartesian XYZ coordinates of each atom, centered by the system center of mass. Unlike learned $l = 1$ features, which primarily capture directional information within local atomic environments, Cartesian coordinates directly encode each atom's position within the entire system while transforming equivariantly under rotation. For example, when a molecule rotates, an atom may move from the right side

of the system to the left. Although learned $l = 1$ features capture the rotation of its local environment, they do not explicitly encode the atom's position within the system. In contrast, Cartesian XYZ coordinates directly capture such positional changes, enabling the model to track atomic motion across frames. We further append 12 auxiliary dimensions describing system-level information, including the atomic numbers, 3×3 lattice cell, structural flags such as periodicity and whether atoms are fixed. This yields a total feature dimensionality of **143** per atom. For a trajectory with K frames, we represent the input as a sequence of node-embedding tensors $\mathbf{H} = [\mathbf{H}_1, \ldots, \mathbf{H}_K]$, where each $\mathbf{H}_k \in \mathbb{R}^{N \times 143}$.

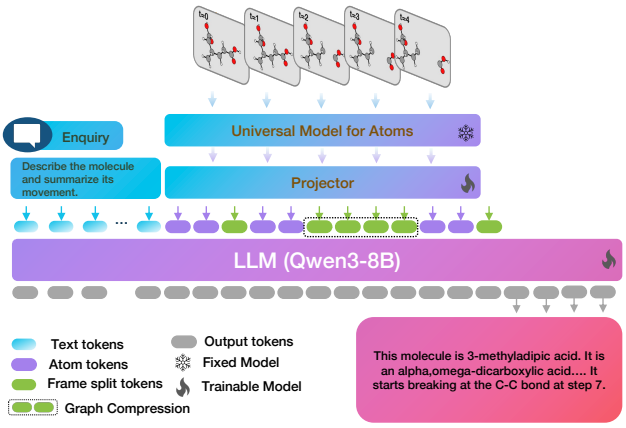


Figure 2: The Chem4DLLM model architecture. (1) A 4D graph encoder (UMA) processes each 3D frame $\mathcal{X}_t$ into a graph embedding, concatenated with raw channels; (2) A projector transforms the graph embeddings into vectors that are additively fused with the embeddings of special <node> tokens; (3) The language model (Qwen3-8B) takes the resulting embedding sequence $\mathbf{E}$ as a prefix and autoregressively generates the output.

Center-of-Mass Reparameterization. Center-of-mass reparameterization prevents the model from encountering extreme numerical out-of-distribution values during inference, which can occur if test molecules exhibit large global translations. Instead of using raw coordinates, each atom's position is expressed relative to the center of mass of the first frame,

$$\tilde{\mathbf{x}}_{k,i} = \mathbf{x}_{k,i} - \mathbf{c}_0, \qquad \mathbf{c}_0 = \frac{\sum_i m_i \, \mathbf{x}_{0,i}}{\sum_i m_i}, \tag{3}$$

and $\mathbf{c}_0$ is supplied to the model as a single line of text in the prompt. This design provides numerically stable inputs while preserving two critical properties. First, fixing a single reference point across the entire trajectory retains the system's global motion, a specific phenomenon evaluated in Chem4DBench. Second, providing $\mathbf{c}_0$ as text makes the transformation lossless. Absolute coordinates remain recoverable without redundantly encoding large global offsets across every atom token in every frame.

Coarse-to-Fine Trajectory Tokenization and Graph Projector. We introduce a coarse-to-fine trajectory tokenization scheme, analogous to video modeling, that preserves the global temporal structure while allocating fine-grained atom-level tokens only to informative frames. Specifically, we use three special tokens: <graph_start>, <node>, and <graph_end>. Starting from frame 0, a frame is retained if its maximum per-atom displacement from the last retained frame exceeds a threshold δ (default 0.005 Å). Retained frames are serialized as <graph_start>, followed by N <node> tokens and <graph_end>. A learnable linear projector $W \in \mathbb{R}^{143 \times 4096}$ maps atomic features into the LLM's hidden space and adds them to the corresponding node embeddings. Compressed frames contain only <graph_start><graph_end>, serving as lightweight temporal anchors. This design preserves trajectory ordering and duration while reducing token consumption, enabling efficient modeling of long trajectories without discarding key structural changes.

LLM Backbone and Fine-tuning The core of Chem4DLLM is the Qwen3-8B language model, implemented using a standard Transformer architecture with FlashAttention 2 (Dao, 2023). The model processes a mixed sequence of text tokens and projected graph tokens. The full input sequence Z is formed by concatenating the system prompt, the sequence of graph spans, and the user query or instruction. This formulation treats physical states as a foreign language that the LLM learns to interpret in context. We perform full-parameter fine-tuning on the combined graph–text inputs. The model is trained using the standard causal language modeling objective:

$$\mathcal{L} = -\sum_t \log p(y_t \mid y_{<t}, \mathbf{H}, \mathcal{P}), \tag{4}$$

where $\tilde{\mathcal{T}}$ denotes the compressed and projected trajectory-token sequence produced by the coarse-to-fine graph tokenization and graph projector, $\mathcal{P}$ is the task prompt, and y_t denotes the target text tokens, such as scientific descriptions or reasoning traces. Optimization is performed using AdamW with a cosine annealing learning rate schedule.

5 RESULTS

5.1 BASELINES

3D Baselines We compare Chem4DLLM with static 3D molecular LLMs and controlled 4D baselines constructed in this work. For static 3D models, we include **3D-MoLM** (Li et al., 2024) and **3D-MolT5** (Pei et al., 2025), which are evaluated on Reaction Product Prediction because they are designed for isolated molecular structures. Inspired by prior works that textualize molecular or crystalline geometries (Jiang et al., 2025; Gruver et al., 2024), we also develop a **3D Text-based** baseline that serializes atomic numbers, Cartesian coordinates, and, when applicable, lattice information into text. This representation is applicable to PBC systems, allowing evaluation across all three tasks.

Controlled 4D Model Variants Since no existing molecular LLM is designed for 4D trajectories, we construct two controlled temporal extensions. **4D-MolT5** extends 3D-MolT5 by serializing a trajectory as an ordered sequence of 3D molecular frames, allowing the pretrained backbone to attend to multiple conformations. **4D Text-based** extends the 3D Text-based baseline by serializing time-indexed coordinates and sparse geometric updates. Although simple and applicable to PBC systems, this text-only representation suffers from severe input-length limitations; for example, a trajectory with 20 frames and 500 atoms can easily exceed 10,000 tokens. These baselines are intended as controlled comparisons rather than independently proposed 4D methods, testing whether simple temporal serialization is sufficient for 4D molecular understanding.

5.2 CHEMICAL DYNAMICS UNDERSTANDING

Reaction Products Prediction We begin by evaluating 4D models on reaction product prediction against the baselines above. As shown in Tab. 2, Chem4DLLM shows a clear advantage in modeling atomistic rearrangements. On Transition1x, it achieves a BLEU score of 0.795 and an EXACT match rate of 0.593, substantially outperforming 4D-MolT5 (BLEU 0.480, EXACT 0.131). This gain also appears in structural metrics, where Chem4DLLM reaches a Morgan (Morgan, 1965) similarity of 0.687, compared with 0.308 for 4D-MolT5. On the larger and more diverse RGD1 dataset, Chem4DLLM remains strongest, achieving an ID EXACT match rate of 0.530 and a BLEU score of 0.484 even in the most challenging OOD-Both setting with unseen reactants and products. Beyond structural prediction, Chem4DLLM excels at estimating related physical properties, which is crucial for downstream chemical applications. It achieves a MAE for the reaction barrier of only 0.127 eV on Transition1x, compared to 0.900 eV for 4D-MolT5. Similarly, the MAE for enthalpy is 0.513 eV, demonstrating that the model effectively captures the energetic landscape of the reaction.

Catalytic Reaction Understanding We further evaluate catalytic reaction understanding on the OC20-based catalytic subset, which involves PBC, surface–adsorbate interactions, and heterogeneous reaction pathways. As shown in Tab. 3, the 4D models improve over the 3D Text-based baseline on most trajectory-dependent metrics, highlighting the value of temporal modeling for surface reactions. Chem4DLLM achieves a reaction-type accuracy of 0.828, compared with 0.517 for 3D Text-based and 0.535 for the 4D Text-based model, indicating stronger recognition of transfer, dissociation, and desorption events. It also improves adsorbate prediction, reaching an Adsorbate Exact match of 0.783 and an Adsorbate Morgan similarity of 0.795. Product prediction remains more challenging due to subtle surface-mediated bond rearrangements, but Chem4DLLM still improves Product Exact match to 0.519 and Product Morgan similarity to 0.549. For physical targets, Chem4DLLM reduces TS Step MAE to 1.138 and achieves the lowest reaction-barrier and enthalpy MAEs, with ΔE MAE of 0.791 and ΔH MAE of 0.931. These energetic results should, however, be interpreted with caution: a subset of the catalytic trajectories is simulated using UMA, while UMA also provides the per-atom representations used by Chem4DLLM, introducing a potential coupling between trajectory generation and model representation. Overall, the results demonstrate the utility of structured trajectory modeling for interpreting catalytic events and transition states under PBC.

Model	BLEU↑	EXACT↑	LEV.↓	MACCS↑	RDK↑	MOR.↑	VALID.↑	ΔE MAE↓	ΔH MAE↓
Transition1x In Distribution									
3D-MoLM	0.216	0.075	10.385	0.296	0.258	0.168	0.514	3.157	3.618
3D-MolT5	0.431	0.070	4.566	0.514	0.400	0.240	0.992	0.902	1.392
3D Text-based	0.059	0.000	8.347	0.206	0.124	0.053	0.954	1.262	2.064
4D-MolT5	0.480	0.131	4.381	0.568	0.477	0.308	0.990	0.900	1.265
4D Text-based	0.051	0.000	8.318	0.200	0.119	0.049	0.957	1.552	2.204
Chem4DLLM	**0.795**	**0.593**	**1.584**	**0.813**	**0.771**	**0.687**	0.992	**0.127**	0.513
RGD1 In Distribution									
3D-MoLM	0.182	0.042	12.472	0.344	0.213	0.222	0.501	1.295	1.205
3D-MolT5	0.640	0.280	4.416	0.722	0.476	0.484	**0.996**	0.687	0.639
3D Text-based	0.273	0.003	8.106	0.214	0.087	0.083	0.634	0.903	0.782
4D-MolT5	0.706	0.405	3.760	0.770	0.566	0.572	0.995	0.443	0.532
4D Text-based	0.258	0.002	8.418	0.211	0.084	0.080	0.648	0.823	0.784
Chem4DLLM	**0.792**	**0.530**	**3.085**	**0.870**	**0.739**	**0.746**	0.995	0.212	0.401
RGD1 Out of Distribution Reactants									
3D-MoLM	0.162	0.012	14.018	0.274	0.124	0.134	0.481	3.400	3.668
3D-MolT5	0.531	0.203	**5.564**	0.600	0.343	0.353	**0.980**	0.879	1.204
3D Text-based	0.281	0.003	9.592	0.333	0.125	0.119	0.856	1.205	1.080
4D-MolT5	**0.535**	**0.227**	5.648	0.604	0.360	0.367	0.954	0.797	1.253
4D Text-based	0.275	0.003	9.742	0.328	0.123	0.116	0.857	0.998	**1.042**
Chem4DLLM	0.467	0.195	7.796	**0.631**	**0.383**	**0.382**	0.954	**0.707**	1.314
RGD1 Out of Distribution Products									
3D-MoLM	0.149	0.002	13.377	0.284	0.102	0.091	0.598	1.906	2.051
3D-MolT5	0.346	0.019	**7.274**	0.483	0.180	0.169	**0.984**	0.906	1.254
3D Text-based	0.180	0.000	13.048	0.247	0.086	0.058	0.844	0.998	1.354
4D-MolT5	0.366	0.055	7.413	0.489	0.204	0.199	0.954	0.761	1.231
4D Text-based	0.179	0.000	13.111	0.243	0.084	0.057	0.847	0.994	1.417
Chem4DLLM	**0.485**	**0.124**	8.324	**0.602**	**0.374**	**0.363**	0.935	**0.419**	**1.046**
RGD1 Out of Distribution Both									
3D-MoLM	0.132	0.000	14.419	0.218	0.086	0.082	0.437	4.384	4.692
3D-MolT5	0.356	0.021	**7.702**	0.479	0.188	0.177	**0.959**	0.994	1.454
3D Text-based	0.214	0.000	12.545	0.267	0.099	0.060	0.915	1.132	1.433
4D-MolT5	0.338	0.046	8.200	0.460	0.198	0.193	0.860	1.004	1.560
4D Text-based	0.209	0.000	12.653	0.260	0.095	0.058	0.903	1.068	1.450
Chem4DLLM	**0.484**	**0.096**	8.754	**0.580**	**0.330**	**0.316**	0.895	**0.678**	**1.362**

Table 2: Reaction product prediction results on Transition1x and RGD1. We evaluate models on the task of predicting product SMILES from reaction trajectories. Results are reported for in-distribution (ID) and multiple out-of-distribution (OOD) settings with unseen reactants, products, or both. Rows shaded in light gray denote models proposed or constructed in this work: 4D-MolT5 and 4D Text-based are our controlled 4D variants of existing 3D molecular LLMs, while Chem4DLLM is our proposed 4D molecular LLM. Overall, Chem4DLLM achieves strong performance across string, structural, and energetic metrics, particularly in challenging OOD regimes. Details about the evaluation metrices can be found in Appx. B.

Model	Type Acc ↑	AExact ↑	AMOR. ↑	PExact ↑	PMOR. ↑	TSMAE ↓	ΔE MAE* ↓	ΔH MAE* ↓
3D Text-based	0.517	0.355	0.392	0.054	0.064	2.437	0.985	1.069
4D Text-based	0.535	0.373	0.410	0.068	0.084	1.947	1.019	1.102
Chem4DLLM	**0.828**	**0.783**	**0.795**	**0.519**	**0.549**	**1.138**	**0.791**	**0.931**

Table 3: Evaluation results on the catalytic reaction subset. Chem4DLLM achieves the highest accuracy in reaction-type classification and adsorbate/product identification. * Energetic MAEs should be interpreted with caution, as a subset of trajectories was generated using UMA, which also provides Chem4DLLM's atom-level representations.

Molecular Motion Understanding The Molecular Motion task is designed primarily to evaluate 4D spatio-temporal geometric understanding, rather than to compete directly with specialized 3D molecule captioning models on static caption-generation. As shown in Tab. 4, 3D-MolT5 remains the strongest model on conventional 3D captioning metrics, reflecting its strong pretraining and tokenization design for static molecule–text modeling. Among our 4D models, 4D-MolT5 and Chem4DLLM are now comparable on captioning: 4D-MolT5 retains an edge on BLEU-2, ROUGE-1 and METEOR, while Chem4DLLM is stronger on BLEU-4, ROUGE-2 and ROUGE-L.

On the trajectory-specific metrics, however, models with 4D inputs show clear advantages. The 3D Text-based baseline, which observes only static geometry, achieves a Motion Accuracy of 19.42 and a Step MAE of 2.95. In contrast, the 4D Text-based model improves Motion Accuracy to 72.18 and reduces Step MAE to 0.47, while Chem4DLLM reaches 81.15 with a Step MAE of 0.30 — the best of the three on both trajectory metrics. These results indicate that trajectory inputs provide essential information for identifying kinematic actions, as well as determining when the motion begins.

Model	BLEU-2↑	BLEU-4↑	ROUGE-1↑	ROUGE-2↑	ROUGE-L↑	MET.↑	MAcc.↑	Step MAE↓
Static 3D Molecule Captioning Models								
MolT5-Large†	25.87	17.28	34.07	16.42	23.41	28.04	-	-
MoMu-Large†	26.34	18.01	34.75	16.86	24.76	28.73	-	-
3D-MoLM†	29.82	22.39	37.23	22.49	31.07	32.69	-	-
3D-MoLM†	30.32	22.52	36.84	22.32	31.23	33.06	-	-
UniMoT†	31.30	23.80	37.50	23.70	33.60	34.80	-	-
MolX†	31.40	24.25	44.20	28.96	38.76	39.55	-	-
3D-MolT5†	42.05	34.16	48.13	33.20	42.33	44.69	-	-
3D Text-based	20.91	8.93	36.79	11.75	20.73	28.38	19.42	2.95
Ours								
4D Text-based	23.36	10.96	38.23	13.03	22.06	30.64	72.18	0.47
4D-MolT5	**33.19**	17.79	**45.99**	17.51	26.40	**36.43**	29.14	2.53
Chem4DLLM	32.65	**18.63**	45.84	**18.87**	**26.74**	34.13	**81.15**	**0.30**

Table 4: Results on the 3D-MoIT-based molecular motion subset. Standard captioning metrics evaluate the original 3D molecule-description task, while Motion Acc. (MAcc.) and Step MAE evaluate 4D trajectory understanding. † denotes captioning results taken from previous publications as reported in (Pei et al., 2025); these static 3D models were not evaluated on our motion-recognition task and therefore do not report MAcc. or Step MAE. The 3D Text-based baseline is evaluated with static geometric input as a reference point, but it does not observe the trajectory. Rows shaded in light gray denote 4D models constructed in this work. Underlining indicates the best result among static 3D captioning models, while bold indicates the best result among the 4D models constructed in this work.

5.3 Ablation Study

We conduct an ablation study on the Molecular Motion subset to investigate the contribution of different geometric representations to Chem4DLLM. Specifically, we examine Cartesian coordinates (xyz), rotation-invariant ℓ=0 features, and rotation-equivariant ℓ=1 features extracted from UMA. As shown in Tab. 5, explicit Cartesian coordinates play a central role in molecular motion understanding. The reference model, combining xyz coordinates with invariant ℓ=0 features, achieves a Motion Accuracy of 81.15. Incorporating additional equivariant ℓ=1 features reduces accuracy to 78.28, suggesting that higher-order geometric features do not necessarily improve performance when explicit coordinates are available. Removing xyz leads to substantially lower accuracy (66.52–69.98), supporting our hypothesis that xyz provide necessary equivariant and global information. These results support the use of explicit coordinates alongside invariant atomic representations in Chem4DLLM, while highlighting the challenge of effectively integrating equivariant features into the language model.

Variant	Mode	MAcc ↑	Step MAE ↓
xyz + ℓ=0	train	81.15	0.296
xyz + ℓ=0 + ℓ=1	train	78.28	0.312
ℓ=0 + ℓ=1, no xyz	train	66.52	0.293
ℓ=0 only, no xyz	train	69.98	0.340

Table 5: Ablation study of geometric representations on the Molecular Motion category. All variants are evaluated on the same test set as Tab. 4.

6 Conclusion

We introduced Chemical Dynamics Understanding (ChemDU), a task that translates 4D molecular trajectories into scientifically grounded descriptions, and Chem4DBench, a benchmark spanning gas-phase reactions, catalytic reactions, and molecular motion. We further developed three 4D molecular language models: 4D Text-based, 4D-MolT5, and Chem4DLLM. By integrating atom-level geometric features and temporal information into a pretrained LLM, Chem4DLLM demonstrates the importance of explicit 4D representations for trajectory-dependent reasoning. Our ablations reveal that Cartesian coordinates provide substantial benefits over pretrained geometric embeddings, while additional equivariant features offer limited gains. These findings also highlight the importance of coordinate normalization to avoid spurious dependence on absolute positions. Chem4DBench establishes a foundation for future research on scalable 4D tokenization, long-horizon reasoning, and mechanistic interpretation, supporting progress toward LLM-driven agentic scientific discovery.

AI use statement

We used ChatGPT, Claude and Gemini to assist with writing and language polishing, retrieve and discover relevant literature, support research ideation and execution, and draft portions of the

manuscript. These uses included discussing methodological and experimental designs, assisting with code revision, and refining the organization and presentation of the paper. The authors take responsibility for the final content of this work, including its text, claims, code, and results.

Ethics Statement

This work uses publicly available molecular simulation and reaction datasets and does not involve human participants or personal data. We acknowledge the ICLR Code of Ethics and provide dataset sources, experimental configurations, and evaluation protocols to support transparency and reproducibility. While our framework aims to facilitate chemical understanding, generated descriptions and predicted reaction properties may contain inaccuracies and should not replace independent quantum chemical or experimental validation. Furthermore, chemical modeling has potential dual-use implications; we encourage responsible application and appropriate scientific oversight.

Reproducibility Statement

All code is available at an anonymous repository: `https://anonymous.4open.science/r/Chem4DLLM-617E/README.md`, including model implementation, data processing, training, evaluation, and scoring scripts. Complete experimental configurations are specified in `scripts/run_eval.sh` through environment variables. To preserve anonymity, we cannot release the full preprocessed datasets (tens of gigabytes) or trained checkpoints (16 GB each). Instead, the repository provides deterministic test subsets for all three tasks (500, 100, and 100 trajectories), corresponding model predictions, reference metrics, and scoring scripts that verify exact matches without GPUs or model weights. Training one checkpoint takes approximately 11 hours on four H100 GPUs.

References

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report, 2023. URL https://arxiv.org/abs/2303.08774.

Microsoft Research AI4Science and Microsoft Azure Quantum. The impact of large language models on scientific discovery: a preliminary study using gpt-4, 2023. URL https://arxiv.org/abs/2311.07361.

Seth D. Axen, Xi-Ping Huang, Elena L. Cáceres, Leo Gendelev, Bryan L. Roth, and Michael J. Keiser. A simple representation of three-dimensional molecular structure. *Journal of Medicinal Chemistry*, 60(17):7393–7409, 2017. ISSN 1520-4804. doi: 10.1021/acs.jmedchem.7b00696. URL http://dx.doi.org/10.1021/acs.jmedchem.7b00696.

Ilyes Batatia, Dávid Péter Kovács, Gregor N. C. Simm, Christoph Ortner, and Gábor Csányi. MACE: higher order equivariant message passing neural networks for fast and accurate force fields. In Sanmi Koyejo, S. Mohamed, A. Agarwal, Danielle Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*, 2022. URL http://papers.nips.cc/paper_files/paper/2022/hash/4a36c3c51af11ed9f34615b81edb5bbc-Abstract-Conference.html.

Ilyes Batatia, Philipp Benner, Yuan Chiang, Alin M. Elena, Dávid P. Kovács, Janosh Riebesell, Xavier R. Advincula, Mark Asta, Matthew Avaylon, William J. Baldwin, Fabian Berger, Noam Bernstein, Arghya Bhowmik, Filippo Bigi, Samuel M. Blau, Vlad Cărare, Michele Ceriotti, Sanggyu Chong, James P. Darby, Sandip De, Flaviano Della Pia, Volker L. Deringer, Rokas Elijošius, Zakariya El-Machachi, Fabio Falcioni, Edvin Fako, Andrea C. Ferrari, John L. A. Gardner, Mikolaj J. Gawkowski, Annalena Genreith-Schriever, Janine George, Rhys E. A. Goodall, Jonas Grandel, Clare P. Grey, Petr Grigorev, Shuang Han, Will Handley, Hendrik H. Heenen, Kersti Hermansson, Christian Holm, Cheuk Hin Ho, Stephan Hofmann, Jad Jaafar, Konstantin S. Jakob, Hyunwook Jung, Venkat Kapil, Aaron D. Kaplan, Nima Karimitari, James R. Kermode, Panagiotis Kourtis, Namu Kroupa, Jolla Kullgren, Matthew C. Kuner, Domantas Kuryla, Guoda Liepuoniute, Chen Lin, Johannes T. Margraf, Ioan-Bogdan Magdău, Angelos Michaelides, J. Harry

Moore, Aakash A. Naik, Samuel P. Niblett, Sam Walton Norwood, Niamh O'Neill, Christoph Ortner, Kristin A. Persson, Karsten Reuter, Andrew S. Rosen, Louise A. M. Rosset, Lars L. Schaaf, Christoph Schran, Benjamin X. Shi, Eric Sivonxay, Tamás K. Stenczel, Viktor Svahn, Christopher Sutton, Thomas D. Swinburne, Jules Tilly, Cas van der Oord, Santiago Vargas, Eszter Varga-Umbrich, Tejs Vegge, Martin Vondrák, Yangshuai Wang, William C. Witt, Thomas Wolf, Fabian Zills, and Gábor Csányi. A foundation model for atomistic materials chemistry, 2023. URL https://arxiv.org/abs/2401.00096.

Simon Batzner, Albert Musaelian, Lixin Sun, Mario Geiger, Jonathan P. Mailoa, Mordechai Kornbluth, Nicola Molinari, Tess E. Smidt, and Boris Kozinsky. E(3)-equivariant graph neural networks for data-efficient and accurate interatomic potentials. *Nature Communications*, 13(1), 2022. ISSN 2041-1723. doi: 10.1038/s41467-022-29939-5. URL http://dx.doi.org/10.1038/s41467-022-29939-5.

Guy W. Bemis and Mark A. Murcko. The properties of known drugs. 1. molecular frameworks. *Journal of Medicinal Chemistry*, 39(15):2887–2893, 1996. ISSN 1520-4804. doi: 10.1021/jm9602928. URL http://dx.doi.org/10.1021/jm9602928.

Lowik Chanussot, Abhishek Das, Siddharth Goyal, Thibaut Lavril, Muhammed Shuaibi, Morgane Riviere, Kevin Tran, Javier Heras-Domingo, Caleb Ho, Weihua Hu, Aini Palizhati, Anuroop Sriram, Brandon Wood, Junwoong Yoon, Devi Parikh, C. Lawrence Zitnick, and Zachary Ulissi. Open catalyst 2020 (oc20) dataset and community challenges. *ACS Catalysis*, 11(10):6059–6072, 2021. ISSN 2155-5435. doi: 10.1021/acscatal.0c04525. URL http://dx.doi.org/10.1021/acscatal.0c04525.

Chi Chen and Shyue Ping Ong. A universal graph deep learning interatomic potential for the periodic table. *Nature Computational Science*, 2(11):718–728, 2022. ISSN 2662-8457. doi: 10.1038/s43588-022-00349-3. URL http://dx.doi.org/10.1038/s43588-022-00349-3.

Qi Chen, Bowen Zhang, Gang Wang, and Qi Wu. Weak-eval-strong: Evaluating and eliciting lateral thinking of llms with situation puzzles. In Amir Globersons, Lester Mackey, Danielle Belgrave, Angela Fan, Ulrich Paquet, Jakub M. Tomczak, and Cheng Zhang (eds.), *Advances in Neural Information Processing Systems 37: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*, 2024. URL http://papers.nips.cc/paper_files/paper/2024/hash/914b372356d58d9e9357b29332cb8fdc-Abstract-Conference.html.

Dimitrios Christofidellis, Giorgio Giannone, Jannis Born, Ole Winther, Teodoro Laino, and Matteo Manica. Unifying molecular and textual representations via multi-task language modelling. In Andreas Krause, Emma Brunskill, Kyunghyun Cho, Barbara Engelhardt, Sivan Sabato, and Jonathan Scarlett (eds.), *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pp. 6140–6157. PMLR, 2023. URL https://proceedings.mlr.press/v202/christofidellis23a.html.

Tri Dao. Flashattention-2: Faster attention with better parallelism and work partitioning, 2023. URL https://arxiv.org/abs/2307.08691.

Carl Edwards, ChengXiang Zhai, and Heng Ji. Text2Mol: Cross-modal molecule retrieval with natural language queries. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pp. 595–607, Online and Punta Cana, Dominican Republic, 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.emnlp-main.47. URL https://aclanthology.org/2021.emnlp-main.47/.

Carl Edwards, Tuan Lai, Kevin Ros, Garrett Honke, Kyunghyun Cho, and Heng Ji. Translation between molecules and natural language. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang (eds.), *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pp. 375–413, Abu Dhabi, United Arab Emirates, 2022. Association for Computational Linguistics. doi: 10.18653/v1/2022.emnlp-main.26. URL https://aclanthology.org/2022.emnlp-main.26/.

Yin Fang, Xiaozhuan Liang, Ningyu Zhang, Kangwei Liu, Rui Huang, Zhuo Chen, Xiaohui Fan, and Huajun Chen. Mol-instructions: A large-scale biomolecular instruction dataset for large language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=Tlsdsb6l9n.

Nate Gruver, Anuroop Sriram, Andrea Madotto, Andrew Gordon Wilson, C. Lawrence Zitnick, and Zachary W. Ulissi. Fine-tuned language models generate stable inorganic materials as text. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=vN9fpfqoP1.

Chuan Guo, Shihao Zou, Xinxin Zuo, Sen Wang, Wei Ji, Xingyu Li, and Li Cheng. Generating diverse and natural 3d human motions from text. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2022, New Orleans, LA, USA, June 18-24, 2022*, pp. 5142–5151. IEEE, 2022. doi: 10.1109/CVPR52688.2022.00509. URL https://doi.org/10.1109/CVPR52688.2022.00509.

Shuhan Guo, Yatao Bian, Ruibing Wang, Nan Yin, Zhen Wang, and Quanming Yao. Unimot: Unified molecule-text language model with discrete token representation, 2024. URL https://arxiv.org/abs/2408.00863.

Dudley R Herschbach. Molecular dynamics of elementary chemical reactions. *Nobel Lectures in Chemistry 1981-1990*, pp. 265–314, 1992.

Anubhav Jain, Shyue Ping Ong, Geoffroy Hautier, Wei Chen, William Davidson Richards, Stephen Dacek, Shreyas Cholia, Dan Gunter, David Skinner, Gerbrand Ceder, and Kristin A. Persson. Commentary: The materials project: A materials genome approach to accelerating materials innovation. *APL Materials*, 1(1), 2013. ISSN 2166-532X. doi: 10.1063/1.4812323. URL http://dx.doi.org/10.1063/1.4812323.

Lei Jiang, Shuzhou Sun, Biqing Qi, Yuchen Fu, Xiaohua Xu, Yuqiang Li, Dongzhan Zhou, and Tianfan Fu. Chem3dllm: 3d multimodal large language models for chemistry, 2025. URL https://arxiv.org/abs/2508.10696.

Johannes Klicpera, Janek Groß, and Stephan Günnemann. Directional message passing for molecular graphs. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020. URL https://openreview.net/forum?id=B1eWbxStPH.

Mario Krenn, Qianxiang Ai, Senja Barthel, Nessa Carson, Angelo Frei, Nathan C. Frey, Pascal Friederich, Théophile Gaudin, Alberto Alexander Gayle, Kevin Maik Jablonka, Rafael F. Lameiro, Dominik Lemm, Alston Lo, Seyed Mohamad Moosavi, José Manuel Nápoles-Duarte, AkshatKumar Nigam, Robert Pollice, Kohulan Rajan, Ulrich Schatzschneider, Philippe Schwaller, Marta Skreta, Berend Smit, Felix Strieth-Kalthoff, Chong Sun, Gary Tom, Guido Falk von Rudorff, Andrew Wang, Andrew D. White, Adamo Young, Rose Yu, and Alán Aspuru-Guzik. Selfies and the future of molecular string representations. *Patterns*, 3(10):100588, 2022. ISSN 2666-3899. doi: https://doi.org/10.1016/j.patter.2022.100588. URL https://www.sciencedirect.com/science/article/pii/S2666389922002069.

Khiem Le, Zhichun Guo, Kaiwen Dong, Xiaobao Huang, Bozhao Nan, Roshni Iyer, Xiangliang Zhang, Olaf Wiest, Wei Wang, Ting Hua, and Nitesh V. Chawla. Molx: Enhancing large language models for molecular understanding with a multi-modal extension, 2024. URL https://arxiv.org/abs/2406.06777.

Junxian Li, Di Zhang, Xunzhi Wang, Zeying Hao, Jingdi Lei, Qian Tan, Cai Zhou, Wei Liu, Yaotian Yang, Xinrui Xiong, Weiyun Wang, Zhe Chen, Wenhai Wang, Wei Li, Mao Su, Shufei Zhang, Wanli Ouyang, Yuqiang Li, and Dongzhan Zhou. Chemvlm: Exploring the power of multimodal large language models in chemistry area. In Toby Walsh, Julie Shah, and Zico Kolter (eds.), *AAAI-25, Sponsored by the Association for the Advancement of Artificial Intelligence, February 25 - March 4, 2025, Philadelphia, PA, USA*, pp. 415–423. AAAI Press, 2025. doi: 10.1609/AAAI.V39I1.32020. URL https://doi.org/10.1609/aaai.v39i1.32020.

Sihang Li, Zhiyuan Liu, Yanchen Luo, Xiang Wang, Xiangnan He, Kenji Kawaguchi, Tat-Seng Chua, and Qi Tian. Towards 3d molecule-text interpretation in language models. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net, 2024. URL https://openreview.net/forum?id=xI4yNlkaqh.

Shengchao Liu, Weili Nie, Chengpeng Wang, Jiarui Lu, Zhuoran Qiao, Ling Liu, Jian Tang, Chaowei Xiao, and Animashree Anandkumar. Multi-modal molecule structure–text model for text-based retrieval and editing. *Nature Machine Intelligence*, 5(12):1447–1457, 2023a. ISSN 2522-5839. doi: 10.1038/s42256-023-00759-6. URL http://dx.doi.org/10.1038/s42256-023-00759-6.

Zhiyuan Liu, Sihang Li, Yanchen Luo, Hao Fei, Yixin Cao, Kenji Kawaguchi, Xiang Wang, and Tat-Seng Chua. MolCA: Molecular graph-language modeling with cross-modal projector and uni-modal adapter. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 15623–15638, Singapore, 2023b. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.966. URL https://aclanthology.org/2023.emnlp-main.966/.

Daniel Mark Lowe. *Extraction of chemical structures and reactions from the literature*. PhD thesis, Apollo - University of Cambridge Repository, 2012. URL https://www.repository.cam.ac.uk/handle/1810/244727.

Yizhen Luo, Kai Yang, Massimo Hong, Xing Yi Liu, and Zaiqing Nie. Molfm: A multimodal molecular foundation model, 2023. URL https://arxiv.org/abs/2307.09484.

Naureen Mahmood, Nima Ghorbani, Nikolaus F. Troje, Gerard Pons-Moll, and Michael J. Black. AMASS: archive of motion capture as surface shapes. In *2019 IEEE/CVF International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019*, pp. 5441–5450. IEEE, 2019. doi: 10.1109/ICCV.2019.00554. URL https://doi.org/10.1109/ICCV.2019.00554.

Grégoire Montavon, Katja Hansen, Siamac Fazli, Matthias Rupp, Franziska Biegler, Andreas Ziehe, Alexandre Tkatchenko, Anatole von Lilienfeld, and Klaus-Robert Müller. Learning invariant representations of molecules for atomization energy prediction. In Peter L. Bartlett, Fernando C. N. Pereira, Christopher J. C. Burges, Léon Bottou, and Kilian Q. Weinberger (eds.), *Advances in Neural Information Processing Systems 25: 26th Annual Conference on Neural Information Processing Systems 2012. Proceedings of a meeting held December 3-6, 2012, Lake Tahoe, Nevada, United States*, pp. 449–457, 2012. URL https://proceedings.neurips.cc/paper/2012/hash/115f89503138416a242f40fb7d7f338e-Abstract.html.

H. L. Morgan. The generation of a unique machine description for chemical structures-a technique developed at chemical abstracts service. *Journal of Chemical Documentation*, 5(2):107–113, 1965. ISSN 1541-5732. doi: 10.1021/c160017a018. URL http://dx.doi.org/10.1021/c160017a018.

Qizhi Pei, Wei Zhang, Jinhua Zhu, Kehan Wu, Kaiyuan Gao, Lijun Wu, Yingce Xia, and Rui Yan. BioT5: Enriching cross-modal integration in biology with chemical knowledge and natural language associations. In Houda Bouamor, Juan Pino, and Kalika Bali (eds.), *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 1102–1123, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.70. URL https://aclanthology.org/2023.emnlp-main.70/.

Qizhi Pei, Lijun Wu, Kaiyuan Gao, Xiaozhuan Liang, Yin Fang, Jinhua Zhu, Shufang Xie, Tao Qin, and Rui Yan. BioT5+: Towards generalized biological understanding with IUPAC integration and multi-task tuning. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 1216–1240, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-acl.71. URL https://aclanthology.org/2024.findings-acl.71/.

Qizhi Pei, Rui Yan, Kaiyuan Gao, Jinhua Zhu, and Lijun Wu. 3d-molt5: Leveraging discrete structural information for molecule-text modeling. In *The Thirteenth International Conference on*

Learning Representations, ICLR 2025, Singapore, April 24-28, 2025. OpenReview.net, 2025. URL https://openreview.net/forum?id=eGqQyTAbXC.

Matthias Plappert, Christian Mandery, and Tamim Asfour. The KIT motion-language dataset. *Big Data*, 4(4):236–252, 2016. doi: 10.1089/big.2016.0028. URL http://dx.doi.org/10.1089/big.2016.0028.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020. URL http://jmlr.org/papers/v21/20-074.html.

Raghunathan Ramakrishnan, Pavlo O. Dral, Matthias Rupp, and O. Anatole von Lilienfeld. Quantum chemistry structures and properties of 134 kilo molecules. *Scientific Data*, 1(1), 2014. ISSN 2052-4463. doi: 10.1038/sdata.2014.22. URL http://dx.doi.org/10.1038/sdata.2014.22.

Xuan Ren, Qi Chen, and Lingqiao Liu. Efficiently selecting response generation strategies for synthetic data construction by self-aligned perplexity. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (eds.), *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 11584–11605, Suzhou, China, 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.621. URL https://aclanthology.org/2025.findings-emnlp.621/.

Mathias Schreiner, Arghya Bhowmik, Tejs Vegge, Jonas Busk, and Ole Winther. Transition1x - a dataset for building generalizable reactive machine learning potentials. *Scientific Data*, 9(1), 2022. ISSN 2052-4463. doi: 10.1038/s41597-022-01870-w. URL http://dx.doi.org/10.1038/s41597-022-01870-w.

K. T. Schütt, H. E. Sauceda, P. -J. Kindermans, A. Tkatchenko, and K. -R. Müller. SchNet – A deep learning architecture for molecules and materials. *The Journal of Chemical Physics*, 148 (24):241722, 2018. doi: 10.1063/1.5019779.

Kristof Schütt, Pieter-Jan Kindermans, Huziel Enoc Sauceda Felix, Stefan Chmiela, Alexandre Tkatchenko, and Klaus-Robert Müller. Schnet: A continuous-filter convolutional neural network for modeling quantum interactions. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett (eds.), *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pp. 991–1001, 2017. URL https://proceedings.neurips.cc/paper/2017/hash/303ed4c69846ab36c2904d3ba8573050-Abstract.html.

Philipp Seidl, Philipp Renz, Natalia Dyubankova, Paulo Neves, Jonas Verhoeven, Jörg K. Wegner, Marwin Segler, Sepp Hochreiter, and Günter Klambauer. Improving few- and zero-shot reaction template prediction using modern hopfield networks. *Journal of Chemical Information and Modeling*, 62(9):2111–2120, 2022. ISSN 1549-960X. doi: 10.1021/acs.jcim.1c01065. URL http://dx.doi.org/10.1021/acs.jcim.1c01065.

Richard Tran, Janice Lan, Muhammed Shuaibi, Brandon M. Wood, Siddharth Goyal, Abhishek Das, Javier Heras-Domingo, Adeesh Kolluru, Ammar Rizvi, Nima Shoghi, Anuroop Sriram, Félix Therrien, Jehad Abed, Oleksandr Voznyy, Edward H. Sargent, Zachary Ulissi, and C. Lawrence Zitnick. The open catalyst 2022 (oc22) dataset and challenges for oxide electrocatalysts. *ACS Catalysis*, 13(5):3066–3084, 2023. ISSN 2155-5435. doi: 10.1021/acscatal.2c05426. URL http://dx.doi.org/10.1021/acscatal.2c05426.

Brook Wander, Muhammed Shuaibi, John R. Kitchin, Zachary W. Ulissi, and C. Lawrence Zitnick. Cattsunami: Accelerating transition state energy calculations with pretrained graph neural networks. *ACS Catalysis*, 15(7):5283–5294, 2025. ISSN 2155-5435. doi: 10.1021/acscatal.4c04272. URL http://dx.doi.org/10.1021/acscatal.4c04272.

Xin Wang, Jiawei Wu, Junkun Chen, Lei Li, Yuan-Fang Wang, and William Yang Wang. Vatex: A large-scale, high-quality multilingual dataset for video-and-language research. In *2019 IEEE/CVF*

International Conference on Computer Vision, ICCV 2019, Seoul, Korea (South), October 27 - November 2, 2019, pp. 4580–4590. IEEE, 2019. doi: 10.1109/ICCV.2019.00468. URL https://doi.org/10.1109/ICCV.2019.00468.

Jiaqi Wei, Yuejin Yang, Xiang Zhang, Yuhan Chen, Xiang Zhuang, Zhangyang Gao, Dongzhan Zhou, Guangshuai Wang, Zhiqiang Gao, Juntai Cao, Zijie Qiu, Ming Hu, Chenglong Ma, Shixiang Tang, Junjun He, Chunfeng Song, Xuming He, Qiang Zhang, Chenyu You, Shuangjia Zheng, Ning Ding, Wanli Ouyang, Nanqing Dong, Yu Cheng, Siqi Sun, Lei Bai, and Bowen Zhou. From ai for science to agentic science: A survey on autonomous scientific discovery, 2025. URL https://arxiv.org/abs/2508.14111.

David Weininger. 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. ISSN 1520-5142. doi: 10.1021/ci00057a005. URL http://dx.doi.org/10.1021/ci00057a005.

Brandon M. Wood, Misko Dzamba, Xiang Fu, Meng Gao, Muhammed Shuaibi, Luis Barroso-Luque, Kareem Abdelmaqsoud, Vahe Gharakhanyan, John R. Kitchin, Daniel S. Levine, Kyle Michel, Anuroop Sriram, Taco S. Cohen, Abhishek Das, Sushree Jagriti Sahoo, Ammar Rizvi, Zachary W. Ulissi, and Larry Zitnick. UMA: A family of universal models for atoms. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/bbf23e81b0ad7637fe9a731d0b676ca6-Abstract-Conference.html.

Teng Xiao, Chao Cui, Huaisheng Zhu, and Vasant G. Honavar. Molbind: Multimodal alignment of language, molecules, and proteins, 2024. URL https://arxiv.org/abs/2403.08167.

Hongliang Xin, John R. Kitchin, and Heather J. Kulik. Towards agentic science for advancing scientific discovery. *Nature Machine Intelligence*, 7(9):1373–1375, 2025. ISSN 2522-5839. doi: 10.1038/s42256-025-01110-x. URL http://dx.doi.org/10.1038/s42256-025-01110-x.

Jun Xu, Tao Mei, Ting Yao, and Yong Rui. MSR-VTT: A large video description dataset for bridging video and language. In *2016 IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2016, Las Vegas, NV, USA, June 27-30, 2016*, pp. 5288–5296. IEEE Computer Society, 2016. doi: 10.1109/CVPR.2016.571. URL https://doi.org/10.1109/CVPR.2016.571.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. Qwen3 technical report, 2025a. URL https://arxiv.org/abs/2505.09388.

Zhen Yang, Zhuo Tao, Qi Chen, Liang Li, Yuankai Qi, Anton van den Hengel, and Qingming Huang. Separation of powers: On segregating knowledge from observation in llm-enabled knowledge-based visual question answering. In *IEEE/CVF Conference on Computer Vision and Pattern Recognition, CVPR 2025, Nashville, TN, USA, June 11-15, 2025*, pp. 24753–24762. Computer Vision Foundation / IEEE, 2025b. doi: 10.1109/CVPR52734.2025.02305. URL https://openaccess.thecvf.com/content/CVPR2025/html/Yang_Separation_of_Powers_On_Segregating_Knowledge_from_Observation_in_LLM-enabled_CVPR_2025_paper.html.

Botao Yu, Frazier N. Baker, Ziqi Chen, Xia Ning, and Huan Sun. LlaSMol: Advancing large language models for chemistry with a large-scale, comprehensive, high-quality instruction tuning dataset. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=lY6XTF9tPv.

Di Zhang, Wei Liu, Qian Tan, Jingdan Chen, Hang Yan, Yuliang Yan, Jiatong Li, Weiran Huang, Xiangyu Yue, Wanli Ouyang, Dongzhan Zhou, Shufei Zhang, Mao Su, Han-Sen Zhong, and Yuqiang Li. Chemllm: A chemical large language model, 2024. URL https://arxiv.org/abs/2402.06852.

Qiyuan Zhao, Sai Mahit Vaddadi, Michael Woulfe, Lawal A. Ogunfowora, Sanjay S. Garimella, Olexandr Isayev, and Brett M. Savoie. Comprehensive exploration of graphically defined reaction spaces. *Scientific Data*, 10(1), 2023. ISSN 2052-4463. doi: 10.1038/s41597-023-02043-z. URL http://dx.doi.org/10.1038/s41597-023-02043-z.

Gengmo Zhou, Zhifeng Gao, Qiankun Ding, Hang Zheng, Hongteng Xu, Zhewei Wei, Linfeng Zhang, and Guolin Ke. Uni-mol: A universal 3d molecular representation learning framework. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net, 2023. URL https://openreview.net/pdf?id=6K2RM6wVqKu.

A CHEM4DBENCH

We provide further details in constructing the Chem4D benchmark

A.1 REACTION PRODUCT PREDICTION

Chemical reactions represent a fundamentally 4D scenario, where atoms evolve from initial reactant positions through high-energy transition states to final product geometries. To decouple the understanding of reaction mechanisms from simple 2D pattern matching, we constructed the Reaction Product Prediction category using two complementary quantum-chemical datasets: Transition1x (Schreiner et al., 2022) and RGD1 (Zhao et al., 2023). For both datasets, we constructed the task by providing the LLM with two 3D frames—the initial reactant geometries and the TS geometries—and asking it to predict the product SMILES, reaction barrier, and reaction enthalpy. To evaluate the model's performance in various OOD cases, we constructed four distinct evaluation splits: an in-distribution (ID) split and three out-of-distribution (OOD) splits—OOD-Reactants, OOD-Products, and OOD-Both. These OOD splits are determined via scaffold partitioning, ensuring that the evaluation probes the model's ability to extrapolate to molecular frameworks that are structurally distinct from those seen during training.

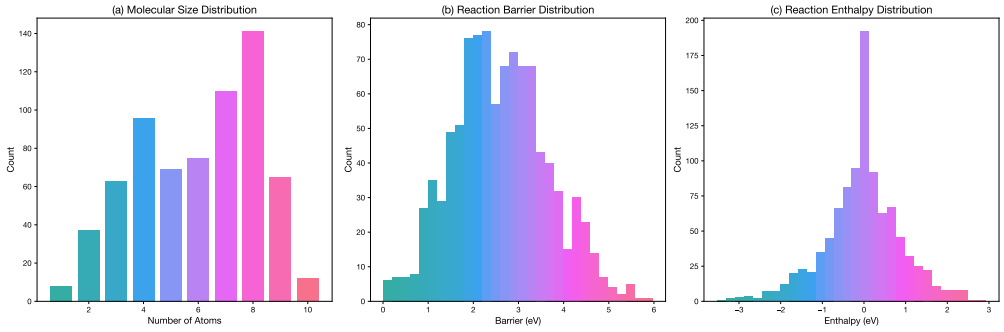


Figure 3: Statistical distribution of Reaction Product Prediction in the Chem4D benchmark. (a) The distribution of the number of atoms. (b) The distribution of the reaction barrier (eV). (c) The distribution of the reaction enthalpy (eV).

A.2 CATALYTIC REACTION

Heterogeneous catalysis involves complex spatio-temporal processes where reactants adsorb onto extended surfaces, undergo bond rearrangements mediated by the catalyst, and subsequently desorb as products. Compared to gas-phase or isolated molecular reactions, catalytic systems introduce additional challenges, including periodic boundary conditions (PBC), surface–adsorbate interactions, and collective atomic rearrangements of the catalyst lattice. To capture these characteristics, we construct the Catalytic Reactions category, which focuses on surface-mediated reaction dynamics in periodic environments.

We build this subset based on the OC20-NEB dataset (Wander et al., 2025), an extension of Open Catalyst 2020 (Chanussot et al., 2021) that provides nudged elastic band (NEB) trajectories for surface reactions. OC20-NEB explicitly resolves transition pathways for three key classes of elementary surface reactions—transfer, dissociation, and desorption—by tracing the atomic-scale evolution from reactant states, through transition-state (TS) regions, to product configurations.

To overcome the limited scale of OC20-NEB ($\sim$ 700 samples), we further augment this category with approximately 6,000 additional surface-reaction trajectories. We start by sampling catalyst–adsorbate reactant states from OC20 (Chanussot et al., 2021). For each sampled structure, we randomly assign a reaction class (transfer, dissociation, or desorption) and re-simulate the corresponding pathway with NEB optimization using the pretrained UMA model as the underlying surrogate potential. We record energies along the reaction coordinate to identify the TS frame, reaction barrier (ΔE), and reaction enthalpy (ΔH). This augmentation yields a larger and more diverse set

of kinetically meaningful surface reaction pathways, substantially strengthening evaluation of 4D understanding under PBC.

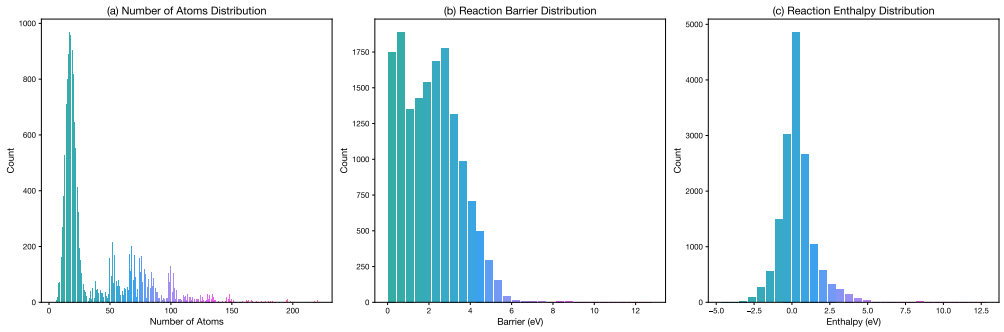


Figure 4: Statistical distribution of Catalytic Reaction Understanding in the Chem4D benchmark. (a) The distribution of the number of atoms. (b) The distribution of the reaction barrier (eV). (c) The distribution of the reaction enthalpy (eV).

A.3 Molecular Motion

A foundational capability of a 4D LLM model is the ability to recognize and describe explicit geometric evolutions in space and time. To this end, we constructed a dataset of algorithmic, kinematic trajectories derived from static structures. The static starting structures are sourced from the PubChem-based corpus used in 3D-MoLM (Li et al., 2024) and 3D-MolT5 (Pei et al., 2025), containing approximately 15,000 molecule-caption pairs. We generate 4D trajectories via the following procedure. First, each molecule is assigned one of seven distinct actions: moving, exploding, partially rotating, globally rotating, vibrating, bending, and bond breaking. Second, we generate a trajectory by applying the corresponding action to the initial geometry. A fundamental capability required for a 4D LLM is temporal grounding—identifying when and what event occurs. To evaluate this capability, we design a random delay $t_{start} \in [2, 10]$ that is prepended to the trajectory, during which the molecule remains in its initial configuration. The dynamic action executes immediately following this prefix. The textual annotations are programmatically generated to capture this temporal structure. Unlike standard captions, these descriptions explicitly state the nature of the geometric event and its precise onset time (e.g., *"This molecule starts exploding at step 3."*). This structure enforces a strict requirement for the model to attend to specific frames within the 4D point cloud sequence. Fig. 5 shows the statistical distributions of action types, start steps, and trajectory lengths.

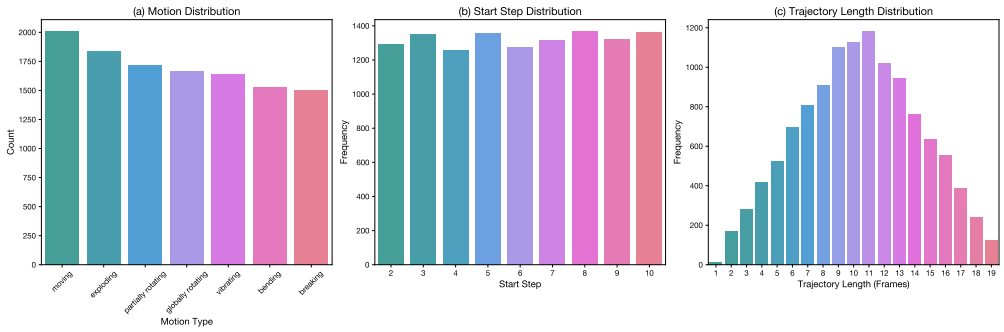


Figure 5: Statistical distribution of the Molecular Motion in the Chem4D benchmark. (a) The frequency distribution of the seven distinct motion types, demonstrating a relatively balanced dataset. (b) The distribution of the start step (t_{start}), confirming uniform sampling of the motion onset delay across the interval $[2, 10]$. (c) The distribution of total trajectory lengths (in frames).

B EVALUATION PROTOCOL

Evaluation is performed separately for each Chem4D category, with metrics designed to reflect linguistic correctness, structural correctness, temporal grounding, and physical validity. Because several tasks require models to generate SMILES strings and numerical physical quantities in free-form text, we first parse each model output into structured fields before computing metrics. Unless otherwise stated, all reported metrics are computed after this parsing step. Higher values indicate better performance for accuracy- and similarity-based metrics, while lower values indicate better performance for error-based metrics such as MAE. Because evaluation plays a central role in Chem4DBench, we include the evaluation scripts used to produce all reported results in the supplementary materials.

Output parsing and normalization. For all generation-based evaluations, we extract the answer after the reasoning delimiter when present, e.g., text following `</think>`. We then use task-specific regular expressions to extract the predicted fields, such as product SMILES, adsorbate SMILES, reaction type, transition-state step, reaction barrier, reaction enthalpy, motion type, and motion starting step. Numerical quantities are parsed as floating-point values and compared to the corresponding ground-truth values. If a required numerical field cannot be extracted, the prediction is treated as invalid for that field and assigned a fixed penalty in the MAE computation.

SMILES canonicalization and molecular comparison. SMILES strings are non-unique: the same molecular graph can be represented by many syntactically different SMILES. Therefore, we do not rely on raw string equality as the primary notion of chemical correctness. Before computing molecule-level metrics, both predicted and reference SMILES are parsed using RDKit. A prediction is considered valid if it can be parsed into a non-empty molecular graph. For exact molecular matching, we compare the canonical molecular identity after RDKit parsing, implemented through canonical molecular representations such as InChI/canonical SMILES rather than raw string equality. Thus, two syntactically different SMILES strings are counted as an exact match if they correspond to the same parsed molecular graph.

String-level metrics, including Levenshtein distance and BLEU, are used only as auxiliary sequence-generation metrics. To reduce the effect of SMILES non-uniqueness, these metrics are computed on normalized SMILES when both prediction and reference can be parsed by RDKit. If a predicted SMILES cannot be parsed, it is counted as invalid, receives zero fingerprint similarity, and its sequence-level distance is computed against the reference using the extracted raw string. Structural metrics are therefore the more chemically meaningful indicators, while sequence-level metrics mainly quantify the difficulty of generating syntactically close molecular strings.

Fingerprint similarity. For valid predicted and reference molecules, we compute Tanimoto similarity between molecular fingerprints. We report three complementary fingerprint metrics: MACCS keys, RDKit topological fingerprints (RDK), and Morgan fingerprints. MACCS captures predefined substructure keys, RDK captures path-based topological features, and Morgan fingerprints capture circular local environments. If either the predicted or reference SMILES cannot be parsed, the fingerprint similarity for that example is set to zero. The reported MACCS, RDK, and Morgan scores are averaged over all examples in the corresponding split.

Reaction Product Prediction. For gas-phase reaction product prediction, each input consists of reaction geometries along the reaction coordinate, including the reactant and transition-state geometries. The model is asked to predict the final product SMILES, the reaction barrier ΔE, and the reaction enthalpy ΔH. We evaluate the predicted product using validity, exact molecular match, Levenshtein distance, BLEU, and MACCS/RDK/Morgan fingerprint similarities. Validity measures the fraction of predicted SMILES that RDKit can parse into chemically valid molecular graphs. Exact match evaluates graph-level identity after canonicalization, rather than raw SMILES equality. Levenshtein distance and BLEU provide auxiliary sequence-level measures of similarity between the generated and reference SMILES. Reaction barriers and enthalpies are evaluated using mean absolute error (MAE) in eV:

$$\mathrm{MAE}(q) = \frac{1}{N}\sum_{i=1}^{N} |\hat{q}_i - q_i|,$$

where q denotes either ΔE or ΔH. Performance is reported on the Transition1x in-domain split and on the RGD1 in-domain and out-of-distribution splits, including OOD-Reactants, OOD-Products, and OOD-Both.

Catalytic Reactions. For heterogeneous catalysis under periodic boundary conditions, evaluation focuses on joint reasoning over surface structure, reaction type, transition-state localization, and energetics. The model is asked to identify the reaction type, the adsorbate, the product when applicable, the transition-state step, the reaction barrier ΔE, and the reaction enthalpy ΔH. Reaction type is evaluated by classification accuracy over the three elementary surface reaction classes: transfer, dissociation, and desorption. Adsorbate prediction is evaluated using exact molecular match and Morgan similarity after RDKit parsing and canonicalization. Product prediction is evaluated similarly, but product SMILES metrics are computed for dissociation cases where product fragments are explicitly defined. For transfer and desorption, final products can be ambiguous or shared across reaction classes, so these cases are primarily evaluated through reaction-type recognition, adsorbate identification, transition-state localization, and energetic prediction.

Transition-state localization is evaluated using TS Step MAE:

$$\text{TS Step MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{s}_i - s_i|,$$

where $\hat{s}_i$ and s_i are the predicted and ground-truth transition-state frame indices. Reaction barriers and enthalpies are evaluated using MAE in eV. Together, these metrics assess whether the model can recognize the catalytic event, identify the relevant surface-bound species, localize the reaction-critical frame, and capture the associated energetic landscape.

Molecular Motion. The Molecular Motion task isolates spatio-temporal geometric reasoning from chemical energetics. Each example contains a molecular trajectory generated by applying a controlled geometric action to an initial 3D molecular structure. The model is asked to generate a description of the molecule and identify the dynamic event and its onset time. We evaluate the original molecule-captioning component using standard text-generation metrics, including BLEU-2, BLEU-4, ROUGE-1, ROUGE-2, ROUGE-L, and METEOR. These metrics quantify the similarity between the generated and reference textual descriptions.

For the trajectory-specific component, we evaluate Motion Accuracy and Step MAE. Motion Accuracy measures whether the predicted action type matches the reference action type, including moving, exploding, partial rotation, global rotation, vibrating, bending, and bond breaking. Step MAE measures the absolute error between the predicted and reference motion starting step:

$$\text{Step MAE} = \frac{1}{N}\sum_{i=1}^{N}|\hat{t}_i^{\text{start}} - t_i^{\text{start}}|.$$

This metric directly evaluates temporal grounding, i.e., whether the model can determine not only what event occurs but also when it begins.

Interpretation of metrics. The metrics are complementary. Sequence-level metrics such as BLEU and Levenshtein distance measure the textual form of the generated SMILES, but they are not sufficient to establish chemical correctness because SMILES are non-unique. For this reason, our main chemically meaningful metrics are validity, canonical exact molecular match, and fingerprint similarity. Similarly, reaction-type accuracy and motion accuracy evaluate discrete event recognition, whereas TS Step MAE and motion Step MAE evaluate temporal grounding. Energetic MAEs evaluate physical fidelity. Strong performance on Chem4DBench therefore requires not only fluent language generation, but also chemically valid structure prediction, temporally grounded event recognition, and accurate physical-property estimation.

C Details of baselines

C.1 3D Text-based

Inspired by Gruver et al. (2024)'s and Jiang et al. (2025)'s work, we develop a 3D Text-based baseline that converts molecular or periodic geometries into structured natural-language descriptions.

This baseline allows a standard LLM to consume 3D structural information without using a dedicated geometric encoder, learned 3D embeddings, or equivariant neural networks. It is intentionally simple and serves as a controlled variant for assessing how much 3D and periodic structural information can be captured through text serialization alone.

Difference from SDF-to-text and CIF-to-text The 3D Text-based baseline differs from the SDF-to-text encoding or text-to-CIF used in Gruver et al. (2024) and Jiang et al. (2025). SDF-to-text is designed to serialize molecular structures in a relatively complete and reconstructable form, including detailed bond and coordinate information. In contrast, our 3D Text-based baseline is descriptive rather than generative: it does not aim to reconstruct the original structure from text, nor does it preserve complete bonding information. Instead, it provides a compact, human-readable geometric summary, including atomic identities, Cartesian coordinates, and a bounded set of local interatomic distances.

Handling periodic boundary conditions. For periodic systems, which are common in heterogeneous catalysis, the 3D Text-based baseline explicitly exposes lattice information rather than attempting to encode periodicity implicitly. When PBC are present, the 3×3 lattice matrix is serialized as three basis vectors along the x, y, and z directions. Atomic positions are reported in Cartesian coordinates within the reference cell.

Neighbor distances under PBC are computed using a radius-based search that accounts for periodic images. In practice, the unit cell is tiled sufficiently to include all neighbors within the cutoff radius. For slab geometries, periodicity is applied only in the surface plane, while the surface-normal direction is treated as non-periodic. After candidate neighbors are generated from periodic images, we enforce a strict cap on the maximum number of neighbors per atom to maintain a bounded and consistent textual representation.

C.2 4D TEXT-BASED

We modify 3D Text-based to 4D Text-based by representing a molecular trajectory as an ordered sequence of frames, where $\mathcal{X}_0$ denotes the initial state and $\mathcal{X}_t$ denotes the configuration at time step t. To control prompt length and ensure scalability to long trajectories, we adopt a compact text-based encoding that combines numeric truncation with sparse temporal updates.

Initial State Encoding ($\mathcal{X}_0$). The first frame provides a complete description of the system. For each atom, we record its chemical symbol and absolute Cartesian coordinates (x, y, z), rounded to two decimal places. For periodic systems, the 3×3 lattice matrix is explicitly included as three basis vectors.

To expose local geometric context, we construct a radius-based neighborhood graph and, to bound the textual length, cap the number of neighbors per atom at k, retaining only the nearest ones. The two settings differ slightly because the systems do: Reaction Product Prediction uses a 2.5 Å cutoff with $k = 5$, while Molecular Motion and Catalytic Reaction use a 3.0 Å cutoff with $k = 4$, the larger radius accommodating the longer contacts found in surface–adsorbate geometries. For systems under PBC, distances are computed using the minimum image convention. The resulting neighborhood information is serialized as tuples of the form `(src_idx, tgt_idx, distance)`, where all distances are also rounded to two decimal places.

Sparse Differential Encoding ($\mathcal{X}_{t>0}$). For subsequent frames, we apply a sparse, event-driven encoding strategy. Instead of re-listing the full structure, we include only atoms whose displacement exceeds a predefined threshold. Specifically, an atom i is reported at frame t only if

$$\|\mathbf{r}_i^{(t)} - \mathbf{r}_i^{(t-1)}\| > \delta, \quad \delta = 0.1\,\text{Å}.$$

For each such atom, we report: 1st, Geometry updates: the updated Cartesian coordinates, rounded to two decimal places. 2nd, Topology updates: newly computed neighbor distances, but only for edges incident to atoms that satisfy the displacement criterion.

Atoms that do not exceed the displacement threshold are omitted entirely from the frame description. This design significantly reduces token consumption for rigid or inactive regions, while preserving high-resolution information around dynamically evolving reaction centers.

Table 6: Example inputs for Chem3D-style (3D) and 4D-text-based models. Both settings aim to predict the product SMILES, reaction barrier, and reaction enthalpy, but differ in how geometric information is provided.

3D-Text-based-style (3D Input)

```
Given the reactant and transition state
geometries,  predict the product SMILES,
reaction barrier and reaction enthalpy.
The geometry is as follows:
The system is non-periodic. The atom and
its  position are as follows:
C1 -1.6169 0.5542 0.9918
C2 -2.2864 -0.7707 1.3265
C3 -1.4256 0.7574 -0.4801
...
H21 1.0702 -0.4440 -2.0669
The distance between atoms is as follows:
atom1_index atom2_index distance(A)
1 7 1.0930
1 6 1.0964
...
21 19 2.3363
```

4D-Text-based (Trajectory Input)

```
Given the reactant and transition state
geometries, predict the product SMILES,
reaction barrier and  reaction enthalpy.
The geometry is as follows:
Frame 1:
The system is non-periodic. The atom and
its position are as follows:
C1 -1.62 0.55 0.99
C2 -2.29 -0.77 1.33
...
The distance between atoms is as follows:
atom1_index atom2_index distance(A)
1 7 1.09
1 6 1.10
...
21 19 2.34

Frame 2:
The atom with big movements are listed as
follows:
C1 0.77 1.22 -0.69
C2 2.26 1.20 -0.37
...
H21 -1.48 -1.01 1.37
New edge distances:
atom1_index atom2_index distance(A)
1 7 1.09
1 6 1.09
...
21 19 2.25
```

The final input to the language model is formed by concatenating the textual descriptions of $\mathcal{X}_0$ and all subsequent sparse updates $\{\mathcal{X}_t\}_{t>0}$. This representation allows the LLM to reason over temporal evolution while remaining within a practical context window.

Tab. 6 provides the examples of 3D Text-based and 4D Text-based methods.

4D-MolT5. We further develop **4D-MolT5** as a temporal extension of 3D-MolT5, designed to interpret molecular dynamics and reaction pathways by lifting static spatial encoding into the temporal domain. In 3D-MolT5, a 1D SELFIES sequence is aligned with a single static 3D molecular conformation, which limits its ability to represent conformational ensembles, transition states, or trajectory-dependent phenomena. In contrast, 4D-MolT5 processes an entire molecular trajectory.

Formally, a 4D trajectory is represented as a sequence of T frames,

$$\mathcal{D} = \{D^{(1)}, D^{(2)}, \ldots, D^{(T)}\},$$

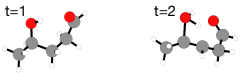
where each $D^{(t)}$ corresponds to the 3D token sequence of the molecule at time step t. To capture spatiotemporal dependencies without modifying the underlying T5 encoder architecture, we flatten the temporal dimension into the token sequence dimension. This results in an extended input sequence of length $T \times N$, where N is the number of atoms.

To preserve atom-level semantic alignment across time, the corresponding 1D SELFIES-based semantic tokens are broadcasted T times and paired with each frame. This serialization strategy allows the global self-attention mechanism of T5 to jointly model intra-frame spatial geometry and inter-frame temporal evolution, effectively enabling the model to *observe* molecular dynamics and reaction trajectories as a unified 4D sequence.

While this approach allows 4D-MolT5 to ingest time-resolved structural information using an unmodified language backbone, it relies on sequence flattening and lacks explicit mechanisms to preserve atomic identity or enforce physical inductive biases across frames, which can limit scalability and fine-grained causal reasoning for long trajectories or large systems.

D OUTPUT EXAMPLE

Query: Given the reactant and transition state geometries, predict the product SMILES, reaction barrier and reaction enthalpy.



Label: The product SMILES is CCCCCO.O, the reaction barrier is 3.225 eV, reaction enthalpy is 0.945 eV.

3D-MoLM: The product SMILES is CC(O)CCCO.[H][H], the reaction barrier is 5.060 eV, reaction enthalpy is 1.780 eV.

3D-MolT5: The product SMILES is CC(C=O)CCO, the reaction barrier is 4.610 eV, reaction enthalpy is 0.000 eV.

3D Text-based : The product SMILES is CC(C(=O)C)CO, the reaction barrier is 2.420 eV, reaction enthalpy is -1.220 eV.

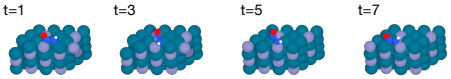
4D-MolT5: The product SMILES is CC(O)CCCO, the reaction barrier is 2.780 eV, reaction enthalpy is 0.390 eV.

4D Text-based: The product SMILES is CC(=O)CC(=C)O, the reaction barrier is 1.920 eV, reaction enthalpy is -0.600 eV.

Chem4DLLM: The product SMILES is CCCCCO.O, the reaction barrier is 3.280 eV, reaction enthalpy is 0.890 eV.

Figure 6: Sample of Reaction Product Prediction

Query: Given the provided reaction trajectory on a catalyst surface, identify the reaction, the catalyst involved, the transition state along the pathway, and determine the reaction energy barrier and reaction enthalpy.



Label: This is a [H][N-][N-][O-] transfer on Pd3Zn(111) surface. The Transition State occurred at step 1. The reaction barrier is 0.56 eV and the reaction enthalpy is 0.52 eV.

3D Text-based : This is a [H]O[N-][N+]([H])[H] transfer on Pd2Zn(111) surface. The Transition State occurred at step 3. The reaction barrier is 0.71 eV and the reaction enthalpy is 0.67 eV.

4D Text-based: This is a [H]ON=O desorption on Pd3Zn(101) surface. The Transition State occurred at step 3. The reaction barrier is 0.82 eV and the reaction enthalpy is 0.82 eV.

Chem4DLLM: This is a [H][N-][N-][O-] transfer on Pd3Zn(001) surface. The Transition State occurred at step 2. The reaction barrier is 0.01 eV and the reaction enthalpy is -0.02 eV.

Figure 7: Sample of Catalytic Reaction Understanding