SpecMol: A Unified Framework for Spectroscopy-Grounded Molecular Modeling and Evaluation

Shuaike Shen,[†,@] Jiaqing Xie,[‡,@] Zhuo Yang,[‡,¶,@] Antong Zhang,[§] Shuzhou Sun,[‡,‖] Ben Gao,[‡,⊥] Tianfan Fu,[#,‡] Biqing Qi,[*,‡] and Yuqiang Li[*,‡]

†*Carnegie Mellon University, Pittsburgh, PA, United States of America*

‡*Shanghai Artificial Intelligence Laboratory, Shanghai, China*

¶*Xidian University, Xi'an, Shaanxi,China*

§*Brown University, Rhode Island, RI, United States of America*

‖*University of Oulu, Oulu, Finland*

⊥*Wuhan University, Wuhan, Hubei, China*

#*Nanjing University, Nanjing, Jiangsu, China*

@*These authors contributed equally to this work.*

E-mail: qibiqing@pjlab.org.cn; liyuqiang@pjlab.org.cn

Datasets

PubChem.

PubChem is a large, publicly accessible chemical information database that integrates data from hundreds of sources. As of recent updates, it contains over 119 million unique compounds and aggregates information from more than 1000 data sources[1]. We leverage PubChem to obtain fundamental molecular identity records, including multiple representations

and textual descriptors for each compound. In practice, for each molecule we retrieve its SMILES strings, IUPAC names, molecular formulae, and known synonyms from PubChem, along with any brief descriptive annotations available. These rich, cross-referenced identifiers provide a foundation for tasks like molecular QA (querying chemical facts) and name-to-structure conversion, ensuring that the model can recognize and interconvert between different naming conventions and representations of the same compound. By using the extensive coverage of PubChem, which spans a broad chemical space and connects to many auxiliary data points, we ensure comprehensive molecular identity information is included for pretraining.

QM9S.

To incorporate high-quality quantum chemical references, we use the **QM9S** dataset [2]. QM9S is an augmented version of the popular QM9 dataset of small organic molecules (up to 9 non-hydrogen atoms). It consists of about 130,000 molecules (composed of C, H, N, O, F) derived from QM9, for which the geometries and properties have been recomputed at a higher level of theory. Specifically, Zou et al. optimized each molecule's 3D structure with DFT (B3LYP/def-TZVP) and then calculated a wide range of physico-chemical properties, including thermodynamic energies, partial charges, dipole moments, higher-order multipole moments, polarizabilities, and other tensorial properties. Importantly, they also simulated several types of spectra from first principles: frequency analysis and time-dependent DFT computations were used to generate **infrared (IR)** and **Raman** spectra, as well as **UV-Vis** absorption spectra for each molecule. This corpus thus offers chemically consistent 3D structures paired with theoretically calculated spectral data. In our training, we use QM9S both to teach the model about accurate molecular geometries and to enable *spectrum simulation tasks* under ideal conditions.

NMRBank.

For experimental spectroscopic data, we draw from **NMRBank**, a recently curated collection of nuclear magnetic resonance records built from the chemical literature[3]. Wang et al. constructed NMRBank by using a language-model-based text mining tool named NMRExtractor to process over 5.7 million scientific publications. The result is a database of about **225,809 entries** of compounds with their reported ^{1}H and ^{13}C NMR chemical shifts, along with metadata such as the experimental conditions (solvent, spectrometer frequency, etc.), confidence indicators, and reference citations. This offers an unprecedented scale of real-world NMR information, far surpassing older public NMR datasets in chemical diversity and size. We include NMRBank in our pretraining corpus to expose the model to genuine experimental spectra characteristics – for instance, the typical chemical shift ranges for various functional groups and the variability of NMR data across different molecules. By retaining the linkage between each NMR record and its compound, the model can learn to associate structural features with NMR signatures (and vice versa) in a realistic context.

Multimodal Spectroscopic Dataset.

In addition to NMRBank, we incorporate a broad **multimodal spectroscopic dataset** introduced by Alberts et al.. This dataset – one of the first of its kind – provides **simulated spectra across six different spectroscopic techniques** for approximately **790,000 organic molecules** extracted from reaction outcomes in patent databases. For each molecule, the dataset includes predicted spectra or spectral features: ^{1}H NMR, ^{13}C NMR, HSQC NMR (a 2D technique), infrared (IR) absorption, and tandem mass spectrometry (MS/MS) in both positive and negative ionization modes. All spectra are computationally simulated; for example, NMR peaks and shifts are predicted, IR intensities are generated over standard frequency ranges, and MS/MS data list fragment peaks with putative fragment formulas. Despite being synthetic, the dataset is designed to reflect realistic experimental outputs. By training on this multimodal dataset, our model learns to handle **multiple spectroscopic**

modalities in combination, mirroring how chemists use complementary techniques for structure elucidation.

Computed vs. Experimental Spectra.

It is important to note the differences between **computed** spectral data like QM9S and **experimental or realistic** spectral data such as NMRBank and the Multimodal Spectroscopic Dataset. QM9S provides high-quality, physics-based data generated under uniform theoretical conditions – highly consistent and reproducible, but lacking the variability of laboratory conditions such as solvent effects or instrument noise. In contrast, NMRBank entries and the patent-derived multimodal dataset embody the complexity of real-world chemistry. The multimodal dataset's spectra, although simulated, cover a broad range of molecular size and functional complexity, while NMRBank provides true experimental chemical shifts, inherently including condition-dependent variations. By combining these sources, we ensure that the model learns both idealized theoretical patterns and pragmatic, experimentally relevant spectra, improving robustness across both spectrum-to-structure and structure-to-spectrum tasks.

Data Processing

PubChem.

For molecular identity and descriptor information, we curated a large-scale dataset from the PubChem compound archive, which provides both 2D and 3D SDF files for millions of compounds. We processed these files using the RDKit cheminformatics toolkit to extract a comprehensive set of molecular features. Each molecule is indexed by its PubChem Compound ID (CID), and all parsed records are stored in both a dictionary (CID$\mapsto$features) and an indexed list for efficient retrieval.

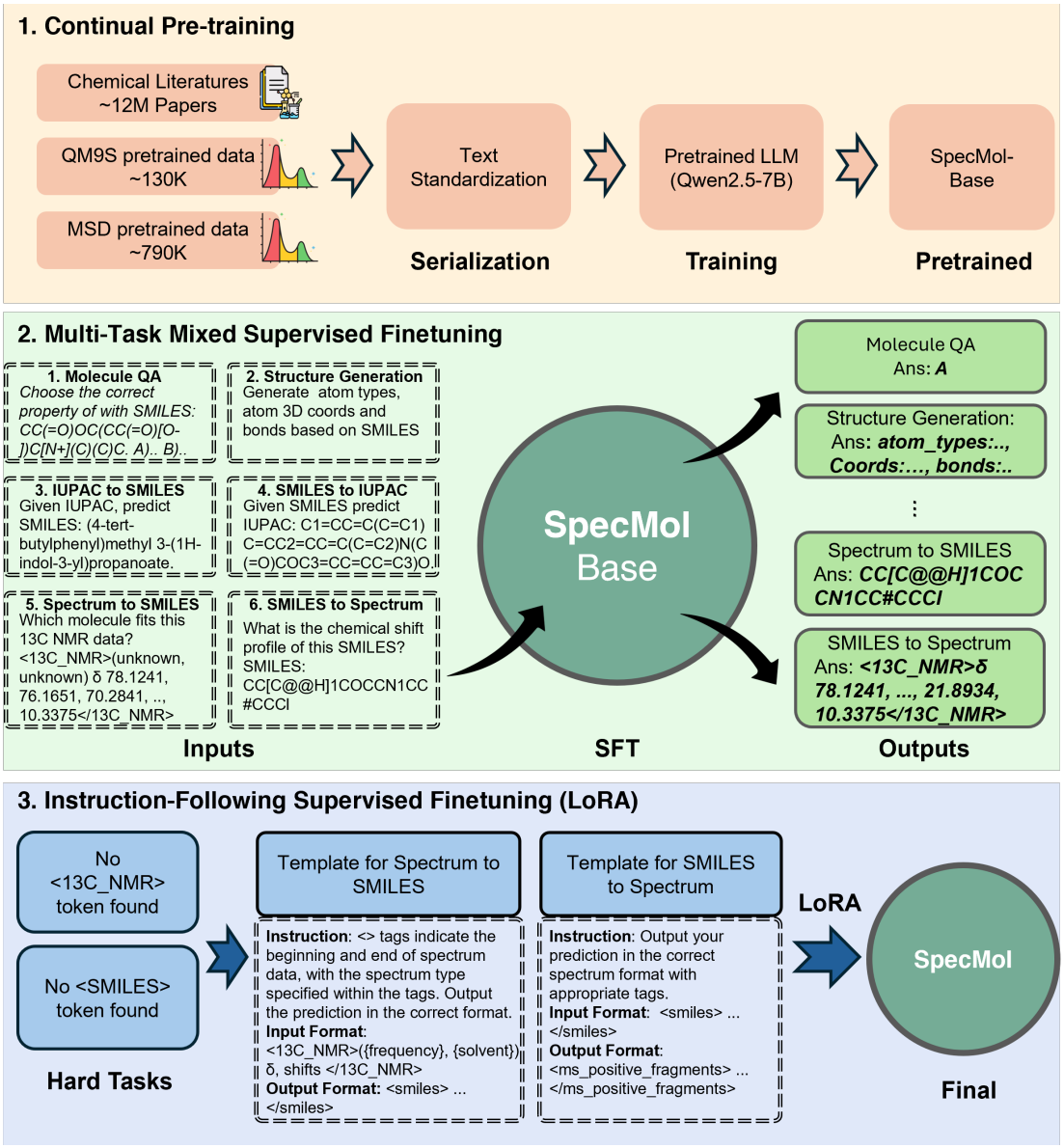

Figure 1: Overall training and alignment pipeline of SpecMol. We perform (1) continual pre-training with standardized spectral serialization and chemical corpora, (2) multi-task mixed supervised fine-tuning on spectroscopy- and structure-centric tasks, and (3) instruction-following LoRA with structured task templates to obtain the final SpecMolmodel.

2D information. From the 2D SDF files, we extracted the following fields explicitly provided by PubChem: canonical SMILES strings, molecular formulae, molecular weight, exact mass, heavy atom count, rotatable bond count, H-bond donors/acceptors, and associated identifiers. In addition, we recorded approximate 2D coordinates of atoms (when present in the file) for visualization or graph layout purposes. These descriptors cover basic chemical identity and structural properties.

3D information. From the 3D SDF files, we used RDKit to obtain a full set of atomic- and molecular-level descriptors:

- *Atomic-level features:* atom indices, atom types (element symbols), formal charges, aromaticity flags, chirality tags (whether an atom has a specified stereochemical label), ring membership (atoms in rings), 3D Cartesian coordinates (from conformers), and explicit bond connectivity (pairs of atom indices) with bond order (single/double/triple).
- *Ring structures:* list of the smallest set of smallest rings (SSSR) per molecule, allowing enumeration of aromatic and non-aromatic rings.
- *Electrostatic descriptors:* per-atom Gasteiger charges, which approximate atomic partial charges from electronegativity equalization.

Molecular fingerprints. Several widely used RDKit fingerprints were computed for each molecule:

- **MACCS keys:** a 166-bit structural key fingerprint capturing presence/absence of common substructures.
- **RDKit fingerprint:** a path-based hashed fingerprint enumerating atom-bond paths.
- **E-State fingerprint:** electrotopological state fragment counts.

Physicochemical descriptors. We also computed common molecular descriptors from RDKit's `Descriptors` module:

- Number of valence electrons (`NumValenceElectrons`);
- Topological polar surface area (TPSA);
- Octanol-water partition coefficient (MolLogP).

Feature annotations. Using RDKit's feature factory (`ChemicalFeatures`), we enumerated pharmacophore-like features such as hydrogen bond donors, acceptors, aromatic centers, and hydrophobic groups. These annotations provide higher-level semantic features for each molecule.

Data organization. The processed dataset thus contains, for each PubChem compound: (i) identifiers and basic properties from the raw SDF (CID, SMILES, formula, exact mass, etc.); (ii) 2D coordinates and counts of functional features (donors, acceptors, rotatable bonds); (iii) full 3D atomic and bonding information; (iv) aromaticity, chirality, and ring structures; (v) multiple types of fingerprints; (vi) physicochemical descriptors and atomic partial charges; and (vii) higher-level chemical features.

Finally, we selected some features suitable for use as language model input. These features form a unified molecular textual description combining identity, structural, electronic, and pharmacophoric information for millions of compounds, enabling downstream molecular QA, name conversion, and 3D generation tasks.

Spectrum Data

^{1}H NMR Spectroscopy. Proton nuclear magnetic resonance (^{1}H NMR) spectroscopy exploits the magnetic properties of hydrogen nuclei to probe molecular structure. Protons in different chemical environments resonate at characteristic frequencies (chemical shifts, δ in ppm), which reflect electron shielding effects[5]. Multiplicity arises from spin–spin coupling with neighboring hydrogens, quantified by coupling constants (J in Hz), while integration reveals the number of protons contributing to each signal[6]. These features provide detailed insights into functional groups and connectivity. We translate raw spectral vectors into structured textual descriptions capturing chemical shifts, multiplicities, couplings, and integrations, thereby embedding human-interpretable NMR cues in a form amenable to LLM-based reasoning.

Algorithm 1 Textual conversion for ^{1}H NMR

Require: Raw vector of ^{1}H NMR peaks: $\{\delta_i, n_i, m_i, J_i\}$
Ensure: Formatted textual representation
Initialize $rep \leftarrow$ "`<1H_NMR>(frequency, solvent)`"
for each peak i **do**
    Format shift δ_i (ppm)
    Extract multiplicity m_i and integration n_i
    if coupling constants J_i available **then**
        Append "J = ... Hz"
    end if
    Append to rep: "δ_i (m_i, n_iH)"
end for
Close tag: $rep \leftarrow rep+$ "`</1H_NMR>`"
return rep

^{13}C NMR Spectroscopy. Carbon-13 NMR (^{13}C NMR) provides a complementary view of molecular skeletons. ^{13}C chemical shifts span a wide range (0–220 ppm), diagnostic of hybridization and functional groups: sp^3 carbons at 0–50 ppm, sp^2 aromatic carbons around 110–160 ppm, and carbonyl carbons beyond 160 ppm[6]. Unlike ^{1}H NMR, broadband-decoupled ^{13}C spectra usually display single peaks per carbon environment without multiplicities, and intensities are not strictly quantitative[5]. Translating spectra into textual form involves listing chemical shift values and identifying characteristic regions (carbonyl, aromatic, aliphatic).

Algorithm 2 Textual conversion for ^{13}C NMR

Require: Raw vector of ^{13}C NMR shifts: $\{\delta_i\}$
Ensure: Formatted textual representation
Sort δ_i values descending
Initialize $rep \leftarrow$ "`<13C_NMR>(frequency, solvent)` δ"
for each shift δ_i **do**
    Append to rep: "δ_i"
end for
Close tag: $rep \leftarrow rep+$ "`</13C_NMR>`"
return rep

Infrared Spectroscopy. Infrared spectroscopy probes vibrational transitions of chemical bonds. Characteristic absorption bands correspond to functional groups: broad O–H stretches at 3200–3600 cm^{-1}, C=O carbonyl stretches at 1650–1800 cm^{-1}, C–H stretches near 2850–3000 cm^{-1}, and sharp nitrile bands at 2250 cm^{-1} [7,8]. By extracting peak positions and intensities, we generate textual summaries indicating functional group assignments.

Algorithm 3 Textual conversion for IR Spectrum

Require: Raw IR spectrum: frequency–intensity pairs $\{(\nu_i, I_i)\}$
Ensure: Formatted textual representation
1: Identify peaks above threshold
2: Initialize $rep \leftarrow$ "`<IR>(500~4000)`"
3: **for** each peak (ν_i, I_i) **do**
4: Append to rep: "ν_i`(`I_i`)`"
5: **end for**
6: Close tag: $rep \leftarrow rep+$ "`</IR>`"
7: **return** rep

Mass Spectrometry. Mass spectrometry (MS) measures mass-to-charge (m/z) ratios of ions, providing molecular weight and fragmentation patterns. The molecular ion (M^+) reveals molecular mass, while fragment ions (e.g. tropylium at m/z 91, phenyl cation at m/z 77) indicate structural motifs[9]. Textual conversion enumerates major peaks and their intensities, normalized to the base peak (100%).

Algorithm 4 Textual conversion for Mass Spectrum

Require: List of peaks $\{(m/z_i, I_i)\}$, normalized to base peak=100%
Ensure: Formatted textual representation
1: Sort peaks by m/z
2: Initialize $rep \leftarrow$ "`<ms_positive>`"
3: **for** each peak $(m/z_i, I_i)$ **do**
4: Append to rep: "$m/z_i : I_i$"
5: **end for**
6: Close tag: $rep \leftarrow rep+$ "`</ms_positive>`"
7: **return** rep

Instruction Data

Based on the processed features obtained from our datasets, we constructed a large collection of instruction-tuning data. As illustrated in Fig. **??**, these data cover a diverse set of tasks:

- **Molecule QA:** question–answer pairs targeting both local and global molecular features.
- **Structure Generation:** the model is required to generate 3D structural coordinates together with atom types and bond types.
- **IUPAC to SMILES:** the model is asked to convert a given IUPAC name into its corresponding SMILES string.
- **SMILES to IUPAC:** the model is asked to generate an IUPAC name from a given SMILES representation.
- **Spectrum to SMILES:** given one or multiple standard textual descriptions of spectra, the model is required to output the corresponding molecular SMILES.
- **SMILES to Spectrum:** given a molecular SMILES, the model is required to predict a specified spectrum in textual form.

Each task is instantiated in multiple QA formats, including free-form question answering, multiple-choice questions, and true/false judgments. Importantly, all QA templates used here are distinct from those employed in the Instruction-Following SFT stage, ensuring no overlap between training and evaluation templates and thus mitigating overfitting and data leakage.

Method Details

Implementation Details

All experiments are conducted on a single node with 8 NVIDIA A800 GPUs. During training, the sequence length is truncated to a maximum of 4096 tokens. The model is trained with a

per-device batch size of 4 and a gradient accumulation step of 8, yielding an effective batch size of 32. We employ a learning rate of 1.0×10^{-5} with a cosine learning rate scheduler and a warm-up ratio of 0.1.

Base Model

We primarily choose **Qwen2.5-7B**[10] as the base model architecture. Qwen2.5-7B is a 7-billion-parameter transformer-based language model, featuring a decoder-only architecture with multihead self-attention and rotary position embeddings. The model was pretrained on a large-scale mixed-domain corpus spanning web documents, code, and scientific texts, which endows it with strong general-purpose language understanding and generation capabilities. Compared to smaller variants, the 7B model strikes a balance between scalability and efficiency, offering sufficient parameter capacity to capture complex multimodal patterns while remaining feasible for fine-tuning on our spectroscopy–structure tasks.

Pretraining

As shown in Fig. 2, we pretrain SpecMol for one epoch on our unified molecular textual description dataset.

Figure 2: Training loss curve of pretraining.

Multi-task Mixed SFT

As shown in Fig. 3, we fine-tune the pretrained SpecMol on all kinds of instrcution data based on the unified molecular textual descriptions for three epochs.

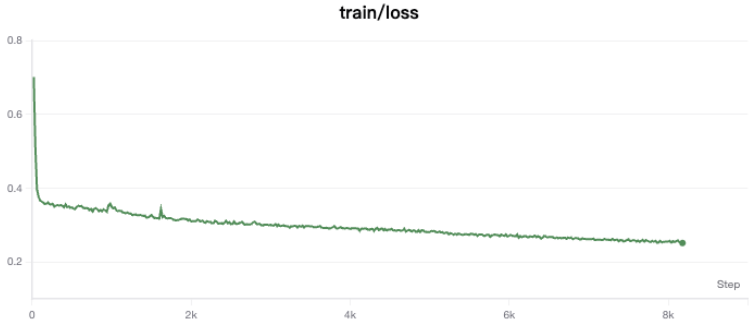

Figure 3: Training loss curve of Multi-task Mixed SFT.

Evaluation

Token-level and Sequence-level Accuracy. For sequence generation tasks (e.g., SMILES to IUPAC), let the test set be $\mathcal{D} = \{(T_i, \hat{T}_i)\}_{i=1}^{N}$, where $T_i = (t_{i,1}, \ldots, t_{i,m_i})$ is the ground-truth token sequence and $\hat{T}_i = (\hat{t}_{i,1}, \ldots, \hat{t}_{i,n_i})$ is the model output. We use a canonicalization map $\mathcal{C}(\cdot)$ (e.g., canonical SMILES) applied to whole sequences before exact comparison. The indicator $\mathbf{1}\{\cdot\}$ returns 1 if the condition holds and 0 otherwise.

Token Accuracy (per-sample).

$$\text{TokenAcc}(T_i, \hat{T}_i) = \frac{1}{|T_i|} \sum_{j=1}^{|T_i|} \mathbf{1}\big\{\hat{t}_{i,j} = t_{i,j}\big\}, \quad |T_i| = m_i. \tag{1}$$

$t_{i,j}$ is the j-th token of the ground truth for sample i; $\hat{t}_{i,j}$ is the j-th token of the prediction (if $j > n_i$, we treat $\hat{t}_{i,j}$ as missing and hence mismatched). The reported Token Accuracy is $\frac{1}{N}\sum_{i=1}^{N} \text{TokenAcc}(T_i, \hat{T}_i)$.

Sequence Accuracy (dataset-level).

$$\text{SeqAcc} \;=\; \frac{1}{N} \sum_{i=1}^{N} \mathbf{1}\big\{\mathcal{C}(\hat{T}_i) \equiv \mathcal{C}(T_i)\big\}. \tag{2}$$

$\equiv$ denotes exact string equality after canonicalization; N is the number of test samples. A sample contributes 1 iff the entire canonicalized ($\mathcal{C}(\cdot)$) prediction matches the canonicalized ground truth.

Algorithm 5 Compute Token & Sequence Accuracy

Require: Test set $\mathcal{D} = \{(T_i, \hat{T}_i)\}_{i=1}^{N}$; canonicalizer $\mathcal{C}(\cdot)$

$S \leftarrow 0$ ▷ exact sequence match counter
$A \leftarrow 0$ ▷ sum of per-sample token accuracies
for $i = 1$ to N **do**
  $T_i' \leftarrow \mathcal{C}(T_i)$; $\hat{T}_i' \leftarrow \mathcal{C}(\hat{T}_i)$
  if $\hat{T}_i' = T_i'$ **then**
    $S \leftarrow S + 1$
  end if
  $m \leftarrow |T_i|$; $n \leftarrow |\hat{T}_i|$; $c \leftarrow 0$
  for $j = 1$ to m **do**
    if $j \le n$ **and** $\hat{t}_{i,j} = t_{i,j}$ **then**
      $c \leftarrow c + 1$
    end if
  end for
  $A \leftarrow A + \frac{c}{m}$
end for
return SeqAcc $= \frac{S}{N}$, TokenAcc $= \frac{A}{N}$

Structure Validity and Geometry Quality. Let a predicted 3D structure be $M = (G, X)$ where $G = (V, E)$ is the molecular graph (V atoms with element types z_i, E bonds with orders o_{ij}) and $X \in \mathbb{R}^{|V|\times 3}$ are Cartesian coordinates (x_i for atom i).

SDF Validity.

$$\text{SDFValid} = \frac{N_{\text{valid}}}{N_{\text{total}}}. \tag{3}$$

N_{valid} is the count of predictions that can be parsed into chemically valid molecules by a toolkit (format OK, valence reasonable, non-empty graph), N_{total} is the number of generated files.

Atom Clash (steric overlap). Define the set of non-bonded pairs $\mathcal{P}_{\text{nb}}(G) = \{(i, j) : i < j,\ (i, j) \notin E\}$. Let $d_{ij} = \|x_i - x_j\|_2$ and r_i^{vdW} be the element-wise van der Waals radius.

A clash occurs if

$$d_{ij} < \alpha\,(r_i^{\mathrm{vdW}} + r_j^{\mathrm{vdW}}), \quad \alpha = 0.65. \tag{4}$$

The reported Atom Clash is the average number of clashing pairs per molecule.

Bond Violation (length out-of-range). For a bonded pair $(i, j) \in E$ with order o_{ij}, let the reference length be $\ell_{ij}^{(o_{ij})}$ from a lookup table conditioned on (z_i, z_j, o_{ij}). A violation occurs if

$$d_{ij} \notin \left[(1-\beta)\,\ell_{ij}^{(o_{ij})},\ (1+\beta)\,\ell_{ij}^{(o_{ij})}\right], \quad \beta = 0.20. \tag{5}$$

The reported Bond Violation is the average number of violated bonds per molecule.

Algorithm 6 Geometry Diagnostics: SDF Validity, Atom Clash, Bond Violation

Require: Predicted set $\{M_k = (G_k, X_k)\}_{k=1}^{N_{\mathrm{total}}}$; parser $\mathsf{Parse}(\cdot)$; radii $r^{\mathrm{vdW}}(z)$; reference lengths $\ell(z_i, z_j, o)$; α=0.65, β=0.20

$V \leftarrow 0$; $C \leftarrow 0$; $B \leftarrow 0$
for $k = 1$ to N_{total} **do**
  if $\mathsf{Parse}(M_k)$ succeeds **then**
    $V \leftarrow V + 1$
  end if
  $c \leftarrow 0$; $b \leftarrow 0$; $(G, X) \leftarrow M_k$
  for all $(i, j) \in \mathcal{P}_{\mathrm{nb}}(G)$ **do**
    $d \leftarrow \|x_i - x_j\|_2$
    if $d < \alpha\,[r^{\mathrm{vdW}}(z_i) + r^{\mathrm{vdW}}(z_j)]$ **then**
      $c \leftarrow c + 1$
    end if
  end for
  for all $(i, j, o) \in E$ **do**
    $d \leftarrow \|x_i - x_j\|_2$; $\quad L \leftarrow \ell(z_i, z_j, o)$
    if $d < (1-\beta)L$ **or** $d > (1+\beta)L$ **then**
      $b \leftarrow b + 1$
    end if
  end for
  $C \leftarrow C + c$; $\quad B \leftarrow B + b$
end for
return SDFValid $= \frac{V}{N_{\mathrm{total}}}$, AtomClash $= \frac{C}{N_{\mathrm{total}}}$, BondViolation $= \frac{B}{N_{\mathrm{total}}}$

Fingerprint Similarity. We compare predicted and reference structures via *Tanimoto similarity* on binary fingerprints. Let $b \in \{0, 1\}^K$ be a fingerprint bit vector and let $|b|_1$

denote its Hamming weight. For two fingerprints $b^{(\text{pred})}, b^{(\text{true})}$:

$$\text{Tanimoto}\big(b^{(\text{pred})}, b^{(\text{true})}\big) = \frac{\langle b^{(\text{pred})}, b^{(\text{true})} \rangle}{|b^{(\text{pred})}|_1 + |b^{(\text{true})}|_1 - \langle b^{(\text{pred})}, b^{(\text{true})} \rangle}. \tag{6}$$

$\langle \cdot, \cdot \rangle$ counts common set bits (intersection size); the denominator is the union size. Values lie in $[0, 1]$.

We report three RDKit[11] fingerprints:

Path-based (RDKFingerprint; "FP Sim"). Enumerate all simple paths $p = (v_1, \ldots, v_L)$ up to length $L \leq L_{\max}$ (typically 7 bonds). Encode a path feature $\phi_{\text{path}}(p)$ from atom types (z_{v_k}) and bond types along p; hash it to an index $h(\phi) \in \{1, \ldots, K\}$ and set $b_{h(\phi)} \leftarrow 1$.

Topological Torsion ("Torsion Sim"). Enumerate all sequences of four consecutively bonded atoms $q = (i, j, k, \ell)$ (paths of length 3). Form a torsion feature $\phi_{\text{tor}}(q) = (\tau(z_i), \tau(z_j), \tau(z_k), \tau(z_\ell),\ \text{bond}_{ij}, \text{bond}_{jk}, \text{bond}_{k\ell})$, where $\tau(\cdot)$ maps raw element/flags to an atom-type class (e.g., element + aromaticity). Hash ϕ_{tor} to set bits. This captures local 4-atom environments[12].

Atom-Pair ("Atom Pair Sim"). For every unordered atom pair (i, j), compute the topological distance δ_{ij} (shortest path length in G). Define an atom-pair feature $\phi_{\text{ap}}(i, j) = (\tau(z_i), \tau(z_j), \delta_{ij})$ and hash to set bits. This captures medium-range topology[13].

Algorithm 7 Fingerprint & Tanimoto Computation

Require: Molecules $M^{(\text{pred})}, M^{(\text{true})}$; hash $h(\cdot)$; path limit $L_{\max}$; bit length K

function PathFP(M)
  $b \leftarrow \mathbf{0}_K$
  for all simple paths p in M with length $\leq L_{\max}$ **do**
    $\phi \leftarrow \phi_{\text{path}}(p)$; $k \leftarrow h(\phi)$; $b_k \leftarrow 1$
  end for
  return b
end function
function TorsionFP(M)
  $b \leftarrow \mathbf{0}_K$
  for all bonded quadruples $q = (i, j, k, \ell)$ in M **do**
    $\phi \leftarrow \phi_{\text{tor}}(q)$; $k \leftarrow h(\phi)$; $b_k \leftarrow 1$
  end for
  return b
end function
function AtomPairFP(M)
  $b \leftarrow \mathbf{0}_K$
  for all unordered pairs (i, j) of atoms in M **do**
    $\delta \leftarrow$ shortest-path length between i and j in G
    $\phi \leftarrow \phi_{\text{ap}}(i, j)$; $k \leftarrow h(\phi)$; $b_k \leftarrow 1$
  end for
  return b
end function
function Tanimoto($b^{(1)}, b^{(2)}$)
  $c \leftarrow \langle b^{(1)}, b^{(2)} \rangle$; $a \leftarrow |b^{(1)}|_1$; $b \leftarrow |b^{(2)}|_1$
  if $a + b - c = 0$ **then**
    return 0
  else
    return $c/(a + b - c)$
  end if
end function
$b^{\text{path}}_{\text{pred}} \leftarrow$ PathFP($M^{(\text{pred})}$); $b^{\text{path}}_{\text{true}} \leftarrow$ PathFP($M^{(\text{true})}$)
$b^{\text{tor}}_{\text{pred}} \leftarrow$ TorsionFP($M^{(\text{pred})}$); $b^{\text{tor}}_{\text{true}} \leftarrow$ TorsionFP($M^{(\text{true})}$)
$b^{\text{ap}}_{\text{pred}} \leftarrow$ AtomPairFP($M^{(\text{pred})}$); $b^{\text{ap}}_{\text{true}} \leftarrow$ AtomPairFP($M^{(\text{true})}$)
return FP Sim = Tanimoto($b^{\text{path}}_{\text{pred}}, b^{\text{path}}_{\text{true}}$), Torsion Sim = Tanimoto($b^{\text{tor}}_{\text{pred}}, b^{\text{tor}}_{\text{true}}$), Atom Pair Sim = Tanimoto($b^{\text{ap}}_{\text{pred}}, b^{\text{ap}}_{\text{true}}$)

Metrics for Spectrum Generation

Details of Spectrum Evaluation Metrics

Here, we provide the formal definitions for the metrics used in the spectrum assessment section.

^{13}C NMR Metrics

Let the ground-truth carbon shifts be the multiset $C = \{\delta_i^{(C)}\}_{i=1}^{n_{\text{true}}}$ and the predictions be $\widehat{C} = \{\widehat{\delta}_j^{(C)}\}_{j=1}^{n_{\text{pred}}}$, where each δ denotes a chemical shift in ppm. We construct a one-to-one matching set M using greedy nearest-neighbor assignment subject to a tolerance $\tau_C = 0.5\,\text{ppm}$:

$$(j,i) \in M \iff \left|\widehat{\delta}_j^{(C)} - \delta_i^{(C)}\right| \le \tau_C$$

Matches are assigned iteratively, selecting the pair with the minimal absolute difference from the pool of unmatched peaks. Let $n_{\text{match}} = |M|$ be the number of matched pairs. The deviations are given by $d_{(j,i)} = \left|\widehat{\delta}_j^{(C)} - \delta_i^{(C)}\right|$. We calculate:

$$P = \frac{n_{\text{match}}}{n_{\text{pred}}}, \quad R = \frac{n_{\text{match}}}{n_{\text{true}}}, \quad \text{F1} = \frac{2PR}{P+R}, \quad \text{MAE} = \frac{1}{n_{\text{match}}} \sum_{(j,i)\in M} d_{(j,i)}.$$

^{1}H NMR Metrics

For ^{1}H NMR, each peak is a tuple (shift, integration). Let the ground truth be $H = \{(\delta_i^{(H)}, nH_i^{(\text{true})})\}_{i=1}^{N_{\text{true}}}$ and the predictions be $\widehat{H} = \{(\widehat{\delta}_j^{(H)}, nH_j^{(\text{pred})})\}_{j=1}^{N_{\text{pred}}}$. We employ a weighted matching protocol with tolerance $\tau_H = 0.12\,\text{ppm}$. For any potential match (j,i) satisfying $|\widehat{\delta}_j^{(H)} - \delta_i^{(H)}| \le \tau_H$, we define an overlap weight $w_{(j,i)}$:

$$w_{(j,i)} = \min\!\left(nH_j^{(\text{pred})}, nH_i^{(\text{true})}\right) \exp\left(-\frac{1}{2}\left(\frac{\widehat{\delta}_j^{(H)} - \delta_i^{(H)}}{\sigma}\right)^2\right), \quad \text{with } \sigma = 0.06 \text{ ppm}.$$

We perform greedy matching to maximize the total weight. Let $\widehat{M}$ be the set of matched indices. The total matched weight is $W_{\text{match}} = \sum_{(j,i)\in\widehat{M}} w_{(j,i)}$. The total proton counts are $W_{\text{pred}} = \sum nH_j^{(\text{pred})}$ and $W_{\text{true}} = \sum nH_i^{(\text{true})}$. The **Weighted Jaccard Similarity** is defined as:

$$\text{Jaccard} = \frac{W_{\text{match}}}{W_{\text{pred}} + W_{\text{true}} - W_{\text{match}}}.$$

Unweighted P, R, F1, and MAE are calculated similarly to the ^{13}C case, using the count of matched pairs $|\widehat{M}|$.

IR and MS Metrics

For Infrared (IR) and Mass Spectrometry (MS), spectra are converted into K-dimensional vectors $\mathbf{p}, \mathbf{q} \in \mathbb{R}^K$ via binning. The similarity is evaluated using the cosine similarity:

$$\text{CosSim}(\mathbf{p}, \mathbf{q}) = \frac{\mathbf{p}^\top \mathbf{q}}{\|\mathbf{p}\|_2 \|\mathbf{q}\|_2}.$$

References

(1) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; others PubChem 2023 update. *Nucleic acids research* **2023**, *51*, D1373–D1380.

(2) Zou, Z.; Zhang, Y.; Liang, L.; Wei, M.; Leng, J.; Jiang, J.; Luo, Y.; Hu, W. A deep learning model for predicting selected organic molecular spectra. *Nature Computational Science* **2023**, *3*, 957–964.

(3) Wang, Q.; Zhang, W.; Chen, M.; Li, X.; Xiong, Z.; Xiong, J.; Fu, Z.; Zheng, M. NMRExtractor: leveraging large language models to construct an experimental NMR

database from open-source scientific publications. *Chemical Science* **2025**, *16*, 11548–11558.

(4) Alberts, M.; Schilter, O.; Zipoli, F.; Hartrampf, N.; Laino, T. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. *Advances in Neural Information Processing Systems* **2024**, *37*, 125780–125808.

(5) Keeler, J. *Understanding NMR spectroscopy*; John Wiley & Sons, 2011.

(6) Claridge, T. D. *High-Resolution NMR Techniques in Organic Chemistry*; Elsevier, 2016.

(7) Colthup, N. *Introduction to infrared and Raman spectroscopy*; Elsevier, 2012.

(8) Smith, B. C. *Infrared spectral interpretation: a systematic approach*; CRC press, 2018.

(9) Gross, J. H. *Mass Spectrometry: A Textbook*; Springer, 2017.

(10) Qwen et al. Qwen2.5 Technical Report. 2025; `https://arxiv.org/abs/2412.15115`.

(11) Landrum, G.; others RDKit: Open-source cheminformatics. 2006.

(12) Nilakantan, R.; Rohrer, S.; Haraki, K.; Venkataraghavan, R. Topological torsion: a new molecular descriptor for SAR applications. *Journal of Chemical Information and Computer Sciences* **1987**, *27*, 82–85.

(13) Carhart, R. E.; Smith, D. H.; Venkataraghavan, R. Atom pairs as molecular features in structure-activity studies: definition and applications. *Journal of Chemical Information and Computer Sciences* **1985**, *25*, 64–73.