

Automated Structure Discovery for Tip-Enhanced Raman Spectroscopy

Harshit Sethi ¹, Markus Junttila ¹, Orlando J. Silveira ^{1,*} and Adam S. Foster ^{1,2,†}¹Department of Applied Physics, [Aalto University](#), Helsinki FI-02150, Finland²WPI Nano Life Science Institute (WPI-NanoLSI), [Kanazawa University](#), Kakuma-machi, Kanazawa 920-1192, Japan

(Received 25 February 2026; accepted 4 August 2026; published 24 September 2026)

Tip-Enhanced Raman Spectroscopy (TERS) provides nanoscale chemical fingerprints alongside high-resolution topographic mapping of molecules, offering a powerful tool for materials discovery. However, TERS image datasets are challenging to interpret and typically demand time-consuming, computationally intensive quantum-chemistry calculations. To overcome this problem, we present an encoder-decoder model trained and evaluated on simulated TERS images of planar molecules, enabling direct prediction of molecular structures from spectral simulated data with high accuracy. Our approach demonstrates the feasibility of automating molecular structure identification from TERS images, bypassing traditional manual analysis. These findings provide a foundation for extending machine learning methods to experimental TERS datasets, potentially accelerating molecular discovery by integrating nanoscale spectroscopy with automated computational analysis.

DOI: [10.1103/8k25-7jtd](https://doi.org/10.1103/8k25-7jtd)

I. INTRODUCTION

Achieving atomic resolution in optical imaging is a central goal for advancing nanoscale science and enabling the study and design of materials. In Scanning Tunneling Microscopy (STM), the functionalization of a tip apex with a single CO molecule enables unprecedented resolution of the internal structure of adsorbed molecules and their orbital densities [1–4]. Similar functionalization in noncontact Atomic Force Microscopy (AFM) provides stunning subatomic resolution of the skeletal structure of a molecule [2,4–6]. Both STM and AFM are among a family of techniques called scanning probe microscopy (SPM), and while it has been proposed that combining different modes of SPM allows for resolution not only the geometric arrangement of the atoms, but also their chemical composition, formidable challenges are still faced to achieve this routinely [7]. In contrast, conventional chemically specific nondestructive spectroscopies, such as infrared and Raman, readily offer chemical characterization, but are constrained by the Abbe diffraction limit and thus lack the type of spatial resolution that CO-tip functionalized SPM can give [8,9]. However, recent advances have shown that integrating an STM tip to the Raman spectrometer greatly enhances the Raman signal, giving the possibility to render both topographical and chemical information by imaging molecules at the angstrom scale [9–15]. In this integrated technique, namely,

tip-enhanced Raman spectroscopy (TERS), an incoming light triggers a collective response to the electrons at the STM tip apex, producing localized surface plasmon excitations and confining propagating light [16] (see Fig. 1). The STM tip acts then as a nanoantenna, transferring to the far-field information from the evanescent waves that originate from interactions of the light with sample sizes smaller than its wavelength, thus overcoming the diffraction limit and limiting the spatial resolution only to the apex radius. Despite these advances, the interpretation of TERS images remains challenging, owing to complex contrast mechanisms, tip dependence, and the large volume and dimensionality of the resulting datasets. Theoretical developments have successfully allowed individual comparisons of simulated TERS images and experiments through visual inspections [12,17–24], but they can be quite demanding tasks as TERS grows to belong to the mainstream group of experiments for molecular characterization [25,26]. All these challenges naturally motivate the use of data-driven approaches to assist in extracting physically and chemically meaningful information from such measurements.

Machine learning (ML) methods have been increasingly used in microscopy to assist with the analysis and interpretation of large, noisy, and high-dimensional datasets. Applications include image denoising and reconstruction, feature identification, and the extraction of underlying physical trends that are difficult to access with conventional analysis. In scanning transmission electron microscopy (STEM), for example, convolutional neural networks (CNNs) have been applied to STEM images to extract structural and compositional information at high throughput [27]. Domain-adaptation methods such as CycleGANs have been employed to render simulated STEM data more realistic for training, enabling robust identification of single-atom defects [28]. More recently, vision-language transformers have been used to predict atomic structure as captions from STEM images [29].

*Contact author: orlando.silveirajunior@aalto.fi†Contact author: adam.foster@aalto.fi

Published by the American Physical Society under the terms of the [Creative Commons Attribution 4.0 International](#) license. Further distribution of this work must maintain attribution to the author(s) and the published article's title, journal citation, and DOI.

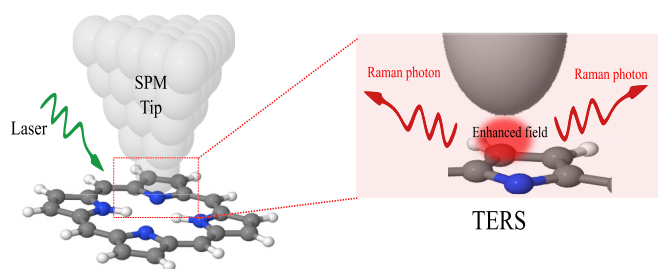


FIG. 1. Schematic illustration of the TERS setup, where laser illumination induces localized field enhancement at the SPM tip apex, enabling enhanced Raman signal intensity.

For SPM, in particular, recent advances at the intersection with ML have advanced prospects for autonomous materials discovery [30,31]. Early supervised ML studies established that SPM images can be mapped directly to atomic-scale structures. For example, by using CNNs to directly predict likely two-dimensional (2D) maps of structures from AFM images [32,33], or combined with graph neural networks [34] for both STM and AFM [35,36]. Complementary approaches used multimodal recurrent networks to infer International Union of Pure and Applied Chemists names from AFM data [37] and conditional generative adversarial networks (cGANs) to translate AFM images into ball-and-stick representations [38].

In this work, we present SMARTERS (Structural Machine Learning for High-Resolution Tip-Enhanced Raman Spectroscopy), an automated CNN-based ML model motivated by the growing use of ML in microscopy and the need for theoretical tools to decode TERS images. Inspired by the success of SPM models trained on simulations, here, the dataset of TERS images was simulated by considering the plasmonic resonance created at the SPM tip as the main source of enhancement of the Raman signal. Within this approach, the local plasmonic field is then treated as a three-dimensional Gaussian function with a specific full width at half maximum (FWHM) [18,21–23,39,40]. Despite the local field created on the tip, it is well known that the interactions of the molecules with the underlying substrate also play an important role in the enhancement of the Raman signal [20,24], and subtle variations of adsorption geometries may drastically affect the vibrations of the molecules [15]. In our simulations, we considered only isolated molecules, making the model more suitable for planar molecules lying flat on top of weakly interacting substrates, e.g., NaCl, where disturbances in the vibrational modes are expected to be less [11]. Additionally, since different molecules exhibit varying numbers of vibrational modes, we proposed a binning procedure along the frequency axes to uniformize the hyperspectral image with different numbers of spectral channels, making the data more suitable for the neural network training.

SMARTERS provides a strong methodological foundation for automated scanning probe microscopy platforms and represents a step toward realizing similar autonomous capabilities in TERS, possibly enabling in the future chemically specific discovery and characterization at the atomic and molecular scale. Automated molecular structure discovery from TERS hyperspectral images has not, to our knowledge,

been demonstrated prior to this work. Here, we address that gap by developing and validating ML methods for structural discovery from TERS hyperspectral datasets. Our model serves as a proof of concept for automating molecular structure identification from TERS images and contributes to a better understanding of light-matter interactions.

II. METHODS

We formulate the task of determining the geometry of molecules from TERS data as an image-to-image translation problem. To address this, the methodology used in SMARTERS is centered on training a deep encoder-decoder network on a large-scale, computationally simulated dataset. This process involves curating a cohort of suitable planar molecules, then for each one, generating a synthetic TERS spectral image paired with its ground-truth 2D atomic position map. The TERS images are generated using our Python reimplementation of the simulation method first implemented by Zhang *et al.* [19]. The network is then trained in a supervised manner to learn a direct mapping from the TERS image to the atomic map, where pixel intensity represents the presence of an atom. The entire end-to-end pipeline is depicted in Fig. 2.

A. Molecule dataset

For generating TERS simulated images, the geometries of sample molecules were taken from a well-known database of small organic molecules structurally optimized with density functional theory (DFT) [41]. The selected dataset was a subpart of the original dataset chosen for establishing automating structure discovery in AFM—we initially selected a subset of 28 570 molecules containing up to 40 atoms as representative of commonly studied systems in TERS. From this subset, we systematically identified planar molecules by applying principal component analysis (PCA) to the atomic coordinates of each molecule, retaining a molecule when its root-mean-square out-of-plane deviation was at most 0.05 Å. However, planarity alone was insufficient to ensure optimal TERS imaging conditions as planar molecules can be arbitrarily rotated, potentially obscuring parts of the structure from the tip’s view during a scan. We mitigated this by rotating each selected molecule to lie in the plane, perpendicular to the tip’s axis (see Appendix B for details). Following this preprocessing pipeline, the dataset was reduced to 1840 molecules, which were subsequently used to generate TERS hyperspectral images for model training. The statistical properties of this set, including distributions of molecule size and elemental composition, are presented in Fig. 3.

B. Simulation method

Our simulation approach to calculate the intensity of the Raman scattering is based on previous work [17,18,21–23,39,40], while the complete workflow of the simulation method is illustrated in Fig. 4. In SMARTERS, we ported the code developed by Zhang *et al.* [17] into Python to ensure compatibility within our ML framework. The simulation technique decouples the calculation of contribution of the tip and the molecule. In this approach, the induced dipole moment μ of a molecule polarized by an electric field E can be expanded

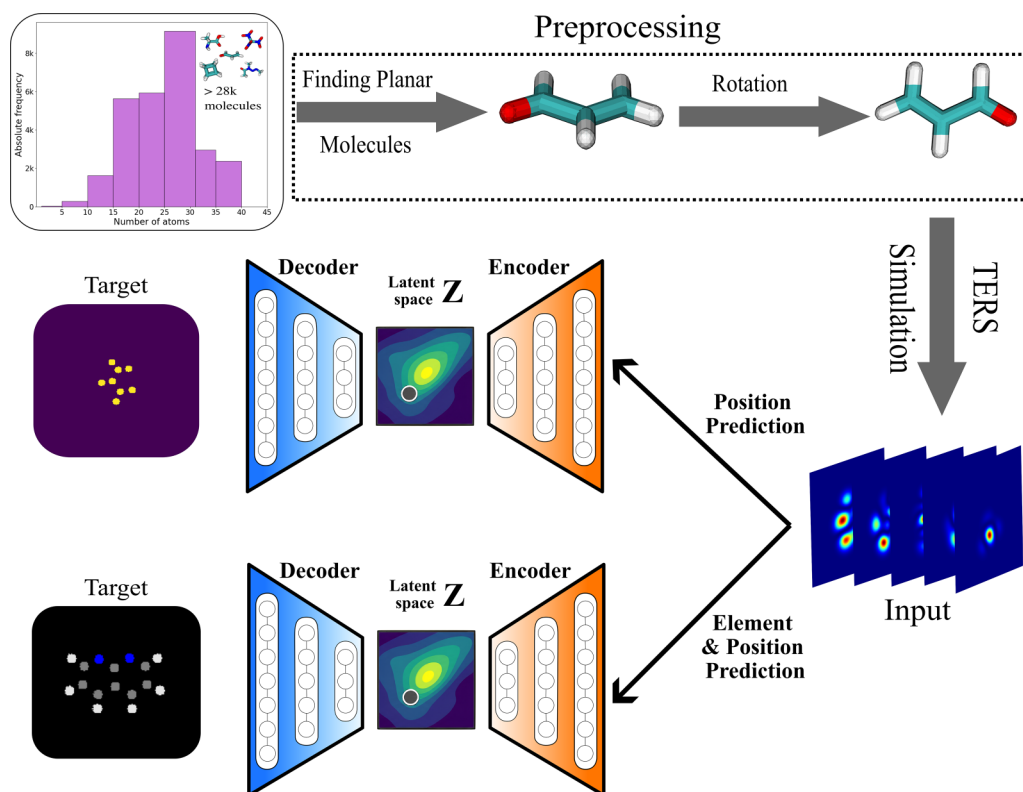


FIG. 2. Schematic of the deep learning workflow for structure prediction from TERS images.

in terms of the normal coordinates Q_k of the k th vibrational mode

$$\mu = \alpha E = \alpha_0 E + \sum_k \left(\frac{\partial \alpha}{\partial Q_k} \right)_0 Q_k E + \dots, \quad (1)$$

where α is the polarizability of the molecule. The Raman scattering process of the molecule is given by the second term of Eq. (1) in the form of a Raman dipole moment. By applying a transformation from normal coordinates Q_k to atomic coordinates $r^{(n)}$ in the derivative of the polarizability, one can

write

$$\frac{\partial \alpha}{\partial Q_k} = C \sum_{n=1}^M L_k^{(n)} \frac{\partial \alpha}{\partial r^{(n)}} = \sum_{n=1}^M \alpha_k^{(n)}, \quad (2)$$

where C is a constant, M is the number of atoms in the molecule, $L_k^{(n)}$ is the normalized displacement of the n th atom corresponding to the vibrational mode k , and $r^{(n)}$ is the coordinate of the n th atom. The $\alpha_k^{(n)}$ is defined as an atomistic polarizability of the atom n corresponding to vibrational mode k , allowing the understanding of the contribution of each atom

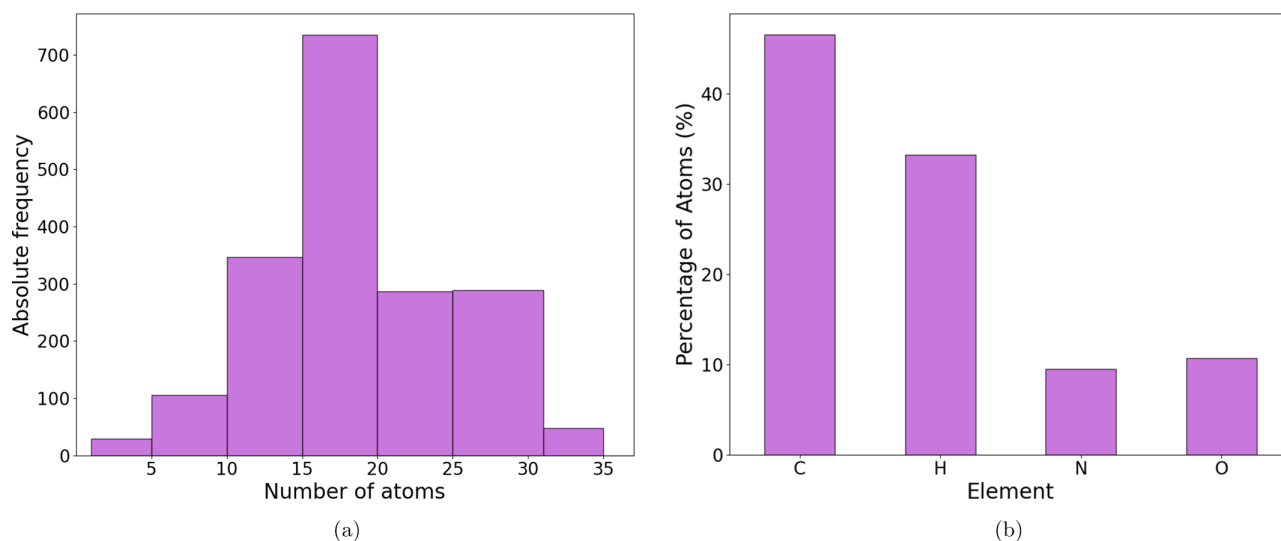


FIG. 3. Characteristics of the final dataset of 1840 preprocessed planar molecules used for simulated TERS image generation. (a) Distribution of molecule size (number of atoms). (b) Distribution of constituent elements.

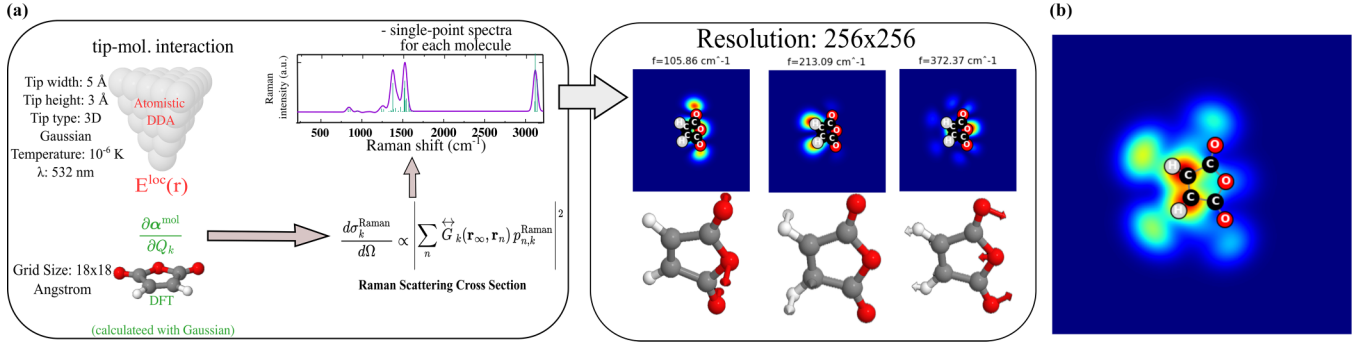


FIG. 4. TERS Simulation Pipeline. Our data generation process, adapted from Zhang *et al.* [17]. (a) The workflow takes a molecule’s geometry, calculates DFT properties, simulates interaction, and computes the Raman spectrum. (b) Sum of all TERS hyperspectral images for visualization.

to the Raman scattering. The atomistic dipole moment due to an incident inhomogeneous local field induced by the STM tip will then become

$$\mu_n^k = \alpha_k^{(n)} E^{\text{loc}}(\mathbf{r}_n), \quad (3)$$

and ultimately the Raman scattering cross section will be given by adding the contribution from each atom in the molecule [17]. The effect induced by the localized field is to enhance the atomistic Raman dipole moment, as well as the radiation of the molecular dipole. Finally, simulated TERS hyperspectral images corresponding to different vibrational modes are generated by combining spectral information from various spatial regions of the molecule. In our implementation, polarizabilities and their derivatives were calculated using DFT with the Gaussian software package [42].

One of the most important parameters for the TERS simulated images is the size of the spatially confined plasmon, which is implicitly defined by the incident inhomogeneous local field $E^{\text{loc}}(\mathbf{r}_n)$. The more localized the plasmon, the more subtle features of the spatial distribution of the vibrational modes can be identified, thus enabling the model to better distinguish between structurally similar molecules. In contrast, a larger plasmon is not able to distinguish between important spectral features, reducing the prediction reliability of our model with respect to those features. Our images were then generated by setting a plasmonic size of 5 Å [represented by a three-dimensional (3D) Gaussian with FWHM of 5 Å], which was enough to distinguish, for example, features arising from Jahn-Teller distortions [23]. The simulation field of view was set to $18 \times 18 \text{ Å}^2$ so that the full lateral extent of the largest molecules is contained within the scan window. Smaller grids would truncate peripheral regions of the structure and effectively crop the simulated images. The hyperspectral maps were rasterized at 256×256 pixels to preserve spatial detail. Coarser discretizations reduced performance by limiting the model’s ability to distinguish vibrational mode signals from noise.

C. Image dataset

Using the simulation method described above, we generated TERS hyperspectral image cubes for each of the 1840 planar molecules in our dataset. A key challenge arose from the fact that different molecules exhibit varying numbers of

vibrational modes, resulting in hyperspectral images with different numbers of spectral channels. To obtain a uniform representation suitable for neural network training, we developed a binning procedure along the frequency axis. The binning strategy divides the frequency range of interest (0–4000 cm^{-1}) into n uniform bins, where the i th bin covers the interval $[i \times \frac{4000}{n}, (i+1) \times \frac{4000}{n}]$, $i = 0, 1, \dots, n-1$. Each bin aggregates vibrational contributions within its frequency range, resulting in a standardized hyperspectral cube of size $H \times W \times n$, where H and W denote the spatial dimensions, and n represents the number of spectral bins (channels). This binning procedure provides two main advantages: (1) It produces hyperspectral cubes with a fixed number of channels across all molecules, enabling consistent model training; (2) it preserves coarse frequency localization of vibrational modes while reducing computational complexity.

We treated n as a hyperparameter and systematically compared different settings with $n = 400$ and $n = 100$. The region of interest for frequency was from 0 to 4000 cm^{-1} . Therefore, $n = 400$ corresponds to a spectral resolution of 10 cm^{-1} per bin, while $n = 100$ corresponds to 40 cm^{-1} per bin. Based on validation-set performance (Sec. III A), we determined that $n = 100$ provided better accuracy while simultaneously reducing training time due to the smaller input dimensionality.

D. Model training

The ML architecture of SMARTERS is similar to the encoder-decoder-type networks that have widely been used in computer vision problems, for example, image segmentation [43]. We selected the Attention U-Net architecture [44] as the backbone for both of our prediction tasks. The same architecture is used in two configurations that differ only in the number of output channels: an atomic structure prediction model that outputs a single-channel atomic map, where pixel intensity encodes the presence of an atom irrespective of its species; and a chemically resolved structure prediction model that outputs a four-channel map, one channel per element (H, C, N, O), encoding both atomic position and elemental identity. We chose this model motivated by its demonstrated prior success in automating structure discovery from scanning probe microscopy images [32,34,35]. The number of input channels of the model was treated as a hyperparameter to accommodate different spectral representations of the

simulated TERS hyperspectral images as input, with either 100 or 400 channels depending on the binning strategy described in Sec. II C. This design allowed us to assess whether higher spectral resolution provides additional predictive benefit, or whether a reduced representation obtained by summing nearby frequency bands is sufficient.

The ML model operates through two main phases: feature extraction and structure prediction. In the encoder phase, the network extracts spectroscopic features from each frequency channel as a function of their spectral character and spatial position. This process occurs simultaneously across all frequency channels in the hyperspectral data. Spectroscopic features that appear consistently across multiple frequency channels yield stronger aggregated activations in the convolutional layers. As the network processes data through its multiple layers, it filters these spectroscopic features according to learned weights and biases. This mathematically prioritizes persistent multichannel signals over uncorrelated background noise, ultimately generating critical feature maps that serve as statistical proxy for molecular structural elements. The decoder then reconstructs the 2D atomic map from these learned features, making predictions about atomic positions based on the spectroscopic input. The attention mechanism in the U-Net architecture is particularly valuable for TERS data analysis, as it allows the network to focus on the most informative spectral-spatial relationships while suppressing irrelevant background signals. This is crucial for TERS applications where signal-to-noise ratios can vary and specific vibrational modes may provide more structural information than others.

Our dataset of 1840 planar molecules was split into 80/10/10 partitions for training, validation, and testing, respectively. The network was trained for 50 epochs using the Adam optimizer [45]. Performance is quantified by the mean Dice Similarity Coefficient (DSC) [46], which measures the pixelwise overlap between the predicted and ground-truth maps. Although DSC is a region-overlap metric, in our sparse-map setting it is strongly sensitive to atom position, so misplaced atoms are penalized rather than tolerated (see Appendix C 1 for more details). For the atomic structure prediction model, the DSC is computed over the single foreground channel, whereas for the chemical species prediction model it is computed per element and averaged across the four channels, giving each species equal weight regardless of its abundance in the dataset. DSC loss is well suited to both prediction tasks, as it directly optimizes for spatial overlap and is robust to class imbalance. For the structure prediction model, this addresses the imbalance between sparse atomic pixels and the background. For the chemically resolved structure prediction model, the loss is computed per channel and averaged across the four elements, which additionally mitigates the imbalance arising from the underrepresentation of nitrogen and oxygen relative to carbon and hydrogen. Simple data augmentations were applied on-the-fly during training to encourage invariance to image orientation and robustness to additive noise, serving as a deliberately simplified proxy for common experimental variations. The augmentation pipeline included random rotations by multiples of 90° , horizontal and vertical flips, and additive Gaussian noise with standard deviation sampled uniformly from [0.01, 0.1]. Identical

geometric transformations were applied to the input images and corresponding masks. Validation and test sets were kept unchanged.

To optimize model performance, we conducted a systematic exploration of key hyperparameters including batch size, learning rate, encoder/decoder depth, and the number of attention channels. We employed Bayesian optimization using the Optuna library [47] to efficiently navigate the hyperparameter space. The explored ranges and final selected values are summarized in Table II in Appendix A. Our entire implementation is built on the PyTorch framework [48].

III. RESULTS

A. Atomic position prediction

We evaluate our proposed Attention U-Net model on the task of atomic position prediction for planar molecules. The model is trained to generate a labeled atom map from simulated images, where each atom is represented as a circle of fixed radius. Our model achieves a mean DSC of 0.862, 0.833, and 0.842 on the training, validation, and test set, respectively, demonstrating its efficacy in predicting the geometry of the molecule. The performance across the training, validation, and test splits remains consistent, indicating robust generalization.

To understand the factors influencing prediction accuracy, we investigated its relationship with molecular planarity. We hypothesized that performance would degrade for minor structural deviations, as the underlying TERS data acquisition is highly sensitive to nanoscopic changes in tip-molecule orientation. This sensitivity arises because deviations from ideal planarity alter the local electromagnetic field enhancement, causing a loss of key spectroscopic information. Our results, illustrated in Fig. 5, confirm this hypothesis. The plot of DSC against root-mean-square deviation (RMSD) shows a marked performance degradation for molecules with $\text{RMSD} > 0.01 \text{ \AA}$, characterized by a lower median DSC. This indicates that as molecules deviate from planarity, the model's predictions become less accurate and less reliable. To examine this effect further, we trained models on datasets including more nonplanar molecules, as will be shown in Sec. III B, confirming that higher RMSD correlates with lower DSC.

A qualitative analysis of the top and bottom 10% of test predictions, ordered by DSC, was performed to identify the factors influencing model performance. As shown in Fig. 6, predictions in the top decile exhibit high fidelity, accurately reconstructing atomic arrangements where the TERS signal is strong and uniform. Conversely, the primary causes of imperfect localization, illustrated in Fig. 7, can be attributed to two main factors: (1) weak Raman signals from atoms located further from the tip, leading to reduced positional precision, and (2) highly imbalanced Raman contribution of different bonds, which can bias the predicted geometry. Notably, even in these lower-ranked predictions, the model often generates chemically plausible structures, suggesting it has learned robust priors about molecular geometry even if the DSC is low. These limitations highlight clear directions for improving both future data acquisition protocols and model robustness.

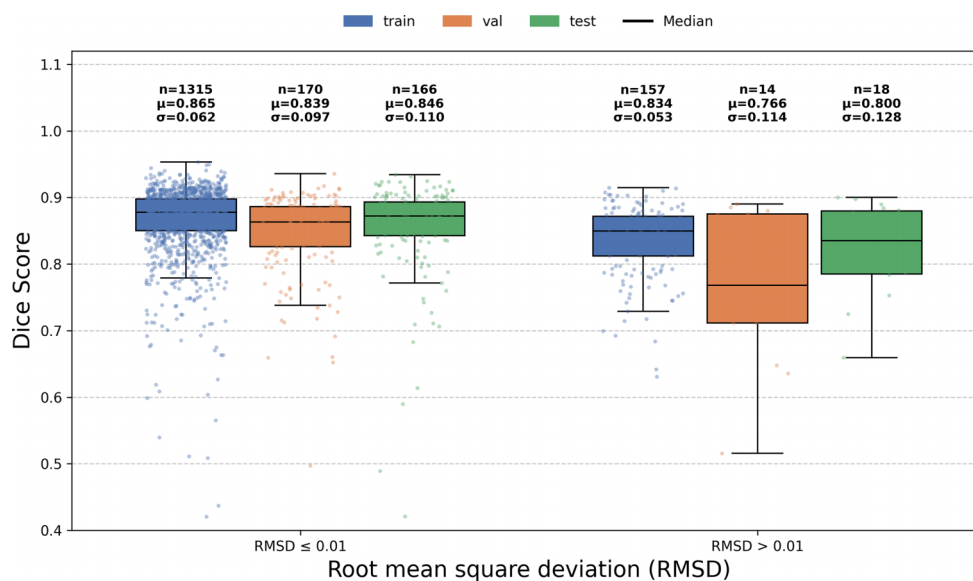


FIG. 5. Model performance as a function of planarity. Molecules that deviate more from being flat ($\text{RMSD} > 0.01 \text{ \AA}$) show a noticeable drop in DSC, indicating that the model has more difficulty with nonplanar structures.

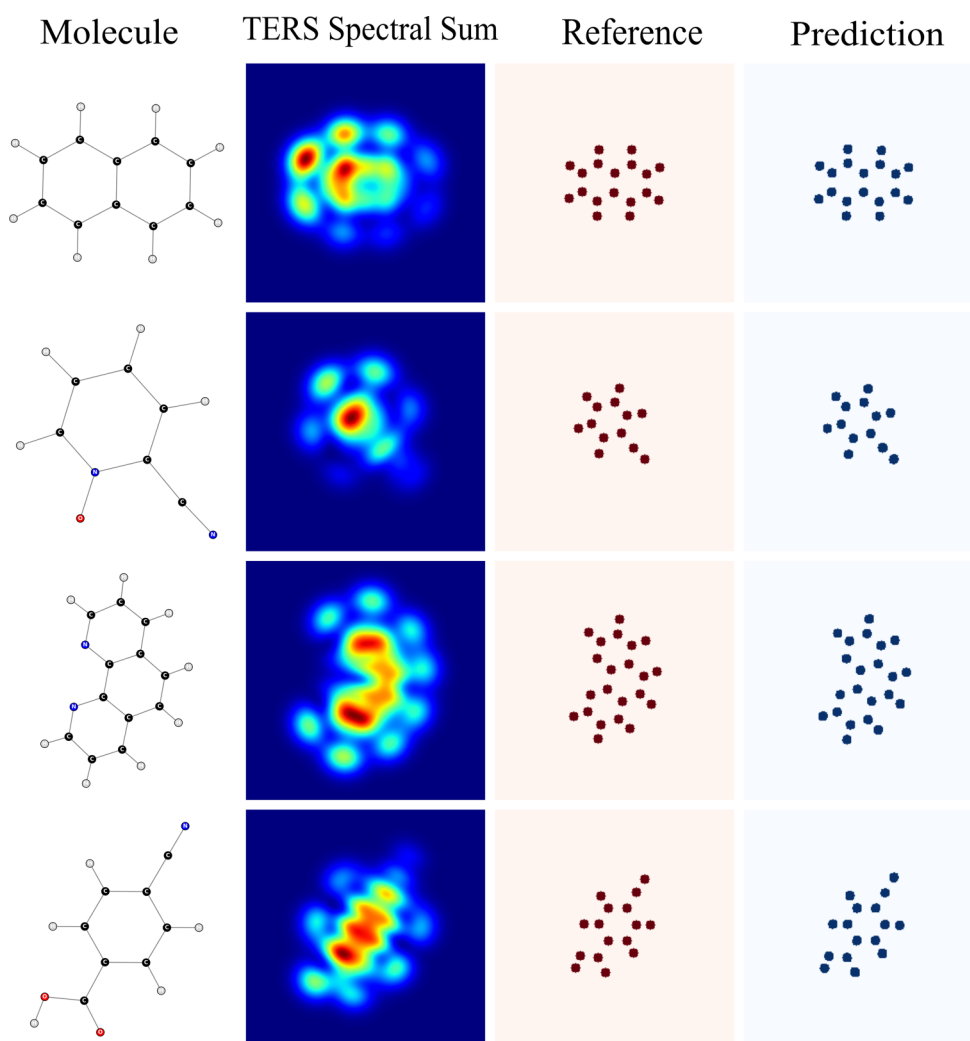


FIG. 6. Representative examples of high-fidelity predictions (top 10% by DSC). The model accurately reconstructs the atomic positions for molecules with strong, uniform TERS signals. Left: ground truth. Right: model prediction.

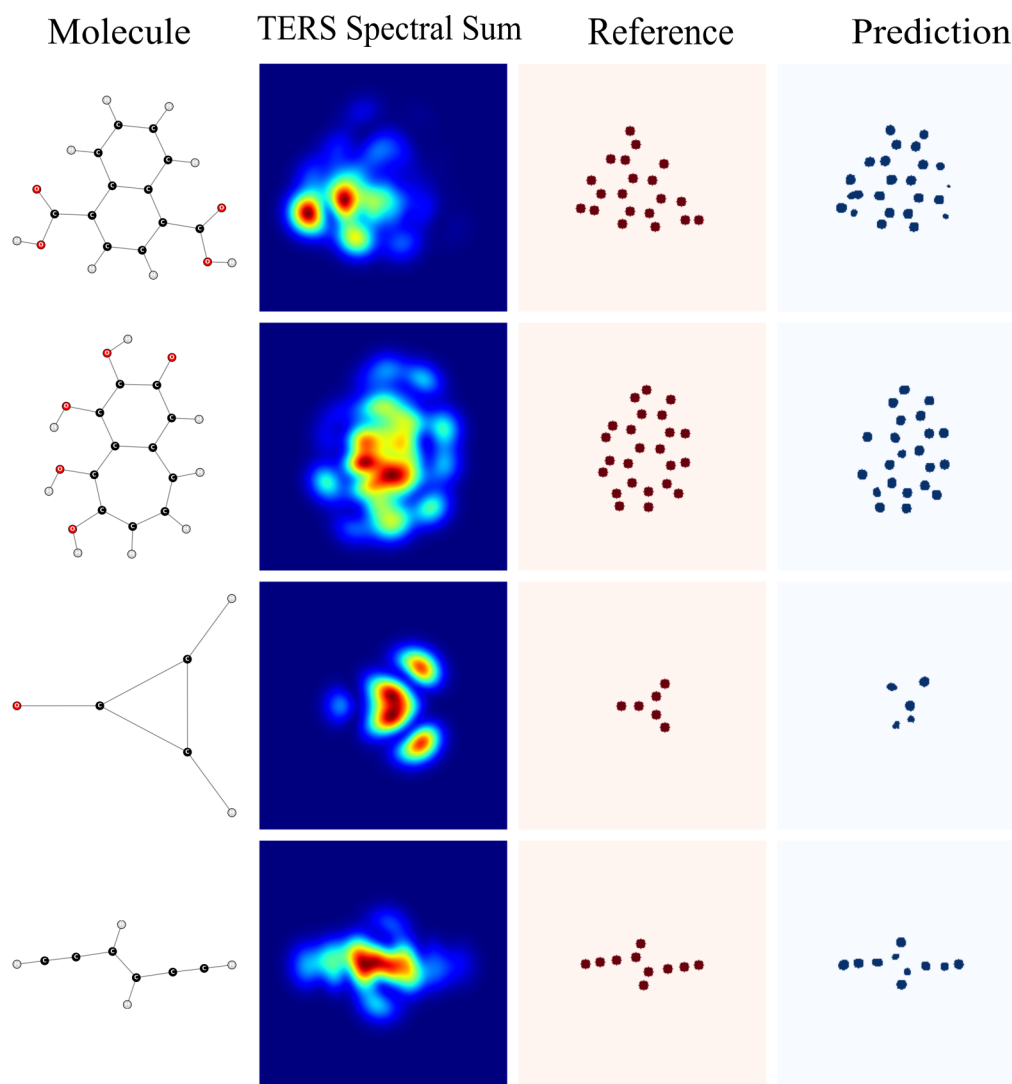


FIG. 7. Characteristic imperfect localization of the model (bottom 10% by DSC). Errors correlate with attenuated signals from distant atoms and geometric distortions from unequal bond signals.

B. Ablation study on the effect of planarity on the accuracy

Nonplanar molecular geometries present an inherent challenge for TERS imaging because the near-field enhancement decays rapidly with increasing distance from the surface. Consequently, atomic contributions originating from out-of-plane atoms are attenuated and less consistently resolved, leading to increased variability in the measured signal for otherwise structurally similar molecules. We refer to this effect throughout as *planarity-induced variability* (PIV).

To investigate the influence of PIV on model performance, datasets were constructed using progressively relaxed planarity constraints. Three models were independently trained, validated, and tested on datasets defined by RMSD thresholds of ≤ 0.05 , ≤ 0.1 , and ≤ 0.5 Å, enabling a controlled investigation of increasing deviations from molecular planarity. As illustrated in Fig. 8, model performance degrades systematically as the fraction of nonplanar molecules increases: Median DSC declines while the spread of the performance distribution grows. This pattern is consistent across training, validation, and test splits.

Although an increase in sample size would typically be expected to reduce variance under purely statistical assumptions, the observed increase in performance spread reflects the effect of PIV on the observed TERS signal intensity. Out-of-plane distortions alter the effective signal formation by weakening or obscuring near-field contributions. As a result, the mapping from molecular structure to measured signal becomes less deterministic. In other words, PIV increases the ambiguity of the inverse problem, which manifests as higher prediction variability for models trained on datasets containing greater structural planarity variation. This analysis highlights an important practical consideration for dataset curation in TERS-based machine learning. RMSD thresholds should be chosen to balance structural diversity against the physical limitations of near-field sensitivity: Including molecules with pronounced out-of-plane distortion increases PIV and degrades model performance for the present modeling approach.

Beyond PIV, TERS presents additional constraints arising from vibrational selection effects. Stretching modes are often weak or absent, and vibrational modes oriented predominantly parallel to the surface contribute less strongly to the measured

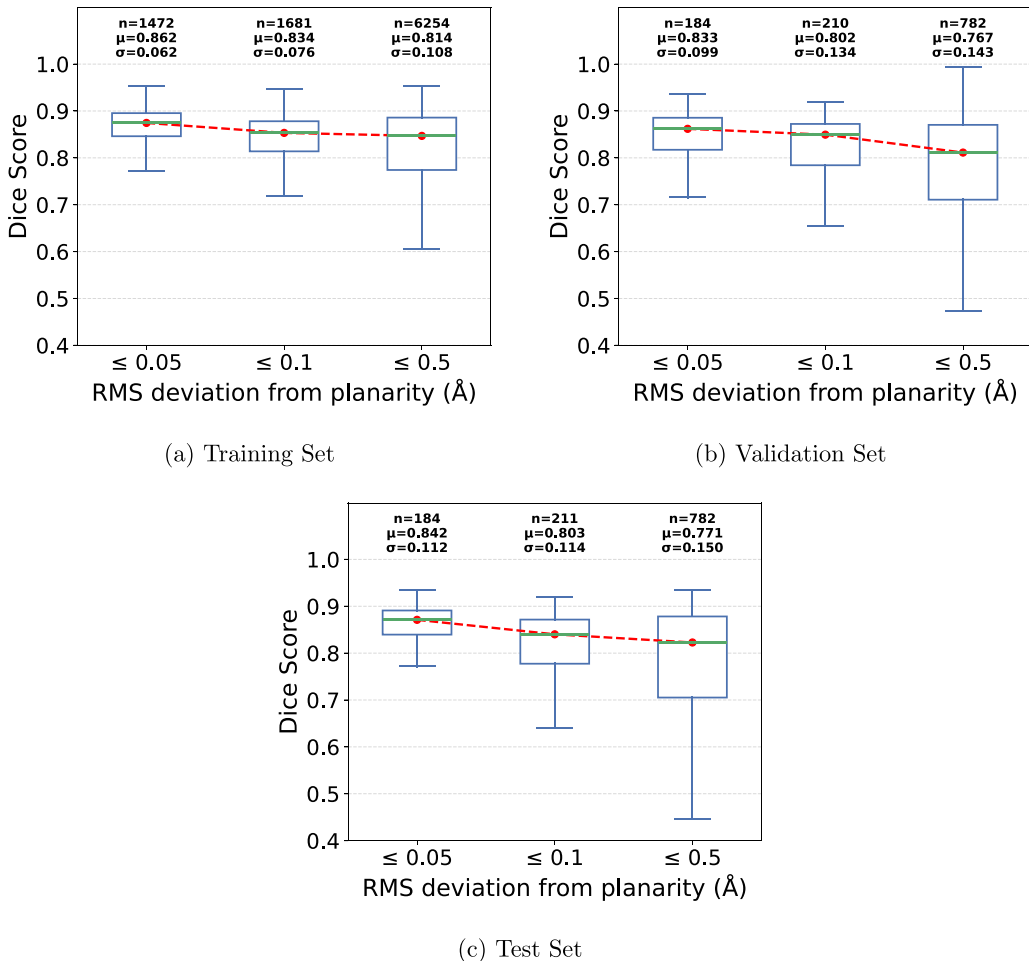


FIG. 8. Impact of molecular planarity on DSC. The x axis represents a cumulative threshold ($\text{RMSD} < x$), where each bin includes all molecules with a planarity deviation less than the specified value in angstroms. The median trend line shows the decrease in the accuracy with increase in nonplanarity. As the inclusion threshold increases (allowing less planar molecules), the variance in performance increases.

signal, whereas modes normal to the surface or aligned with the tip axis dominate the response. This introduces ambiguity in defining the ground-truth image, since atomic groups exhibiting primarily in-plane vibrations may remain weakly represented or undetected in the measurements. Consequently, certain structural regions may appear absent in the signal despite being physically present, introducing inconsistencies that can further complicate model training.

C. Chemically resolved structure prediction

We further evaluate our Attention U-Net model on the joint task of atomic position and chemical species prediction for planar molecules. Whereas the atomic position prediction model (Sec. III A) outputs a single foreground channel, here the network outputs one channel per element (H, C, N, O), so that each atom is represented as a circle of fixed radius assigned to its corresponding elemental channel, jointly encoding location and identity. The model achieves mean DSC values of 0.832, 0.802, and 0.810 on the training, validation, and test sets, respectively (Table I), demonstrating that the architecture can resolve elemental identity in addition to molecular geometry. These values are modestly lower than the values obtained for atomic position prediction alone, con-

sistent with the greater difficulty of the joint task, where the model must additionally discriminate between chemically distinct atoms rather than only localizing them. As in the atomic position prediction task, performance remains consistent across the three splits, indicating robust generalization.

The dataset is strongly imbalanced in chemical composition, with hydrogen and carbon far more abundant than nitrogen and oxygen. For evaluation, we report the mean of the per-element DSC values (Table I). As shown in Fig. 9, hydrogen and carbon are resolved most accurately, whereas nitrogen attains both the lowest median DSC and the highest variance, giving a spread of roughly 0.10 in DSC between the

TABLE I. Per-element DSC and overall mean DSC on the training, validation, and test sets.

	Train	Validation	Test
H	0.869	0.858	0.856
C	0.842	0.816	0.827
N	0.782	0.733	0.759
O	0.834	0.803	0.799
Mean	0.832	0.802	0.810

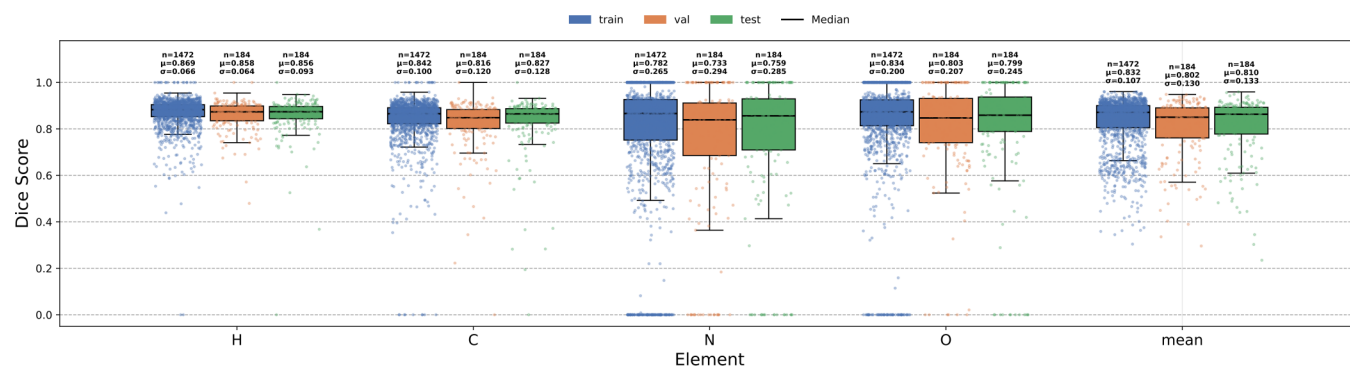


FIG. 9. Distribution of per-element DSC on the test set for hydrogen, carbon, nitrogen, and oxygen. Nitrogen shows both the lowest median DSC and the largest spread, while hydrogen and carbon are predicted most consistently. This trend mirrors the relative abundance of each element in the molecule dataset (Fig. 3).

best and worst element on the test set. This per-element trend in both mean and variance correlates directly with the relative abundance of each element in the dataset (Fig. 3).

To further characterize model behavior, we visualize the top and bottom 10% of test-set predictions, ranked by overall DSC. As shown in Fig. 10, predictions in the top decile

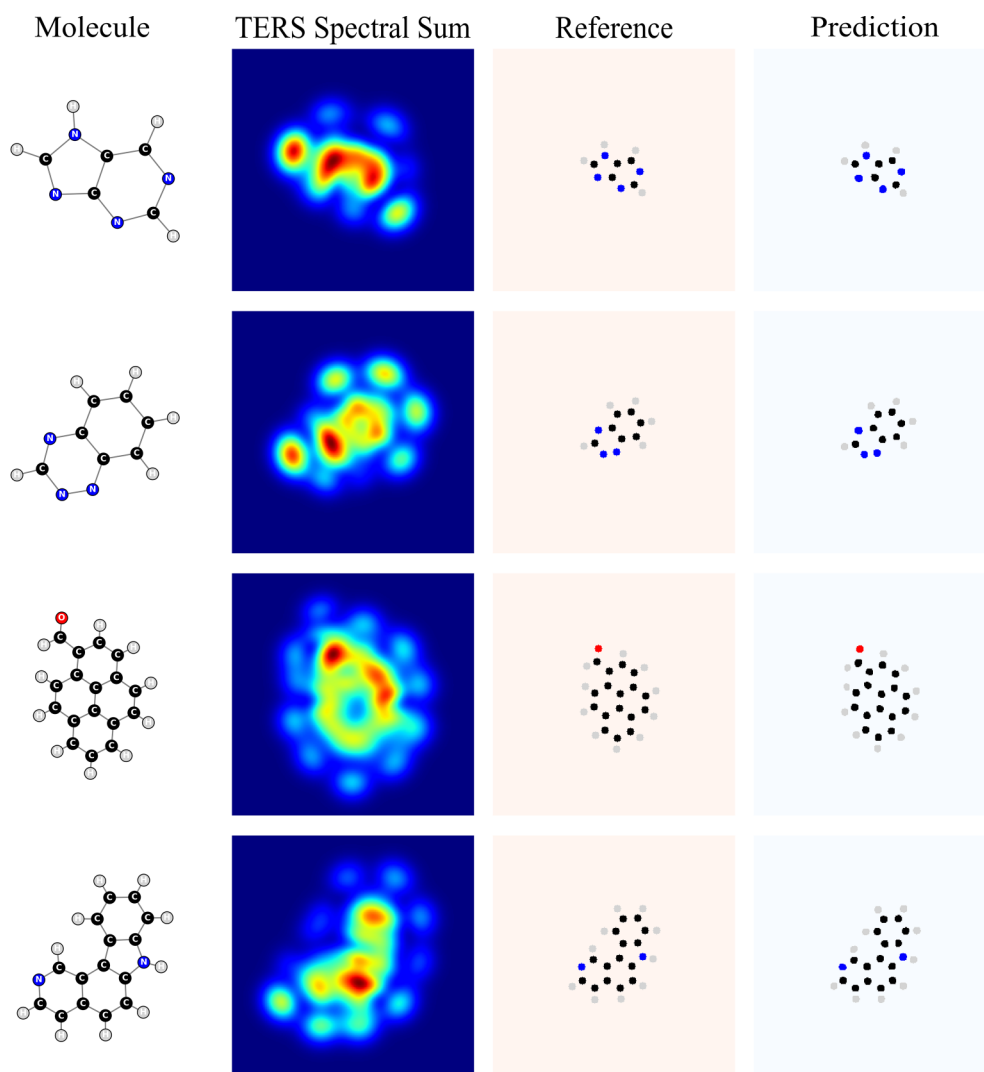


FIG. 10. Representative examples of high-fidelity chemical species predictions (top 10% by DSC). The model accurately reconstructs both the atomic positions and elemental identities (H, C, N, O) for molecules with strong, uniform TERS signals. Left: Ground Truth. Right: Model Prediction.

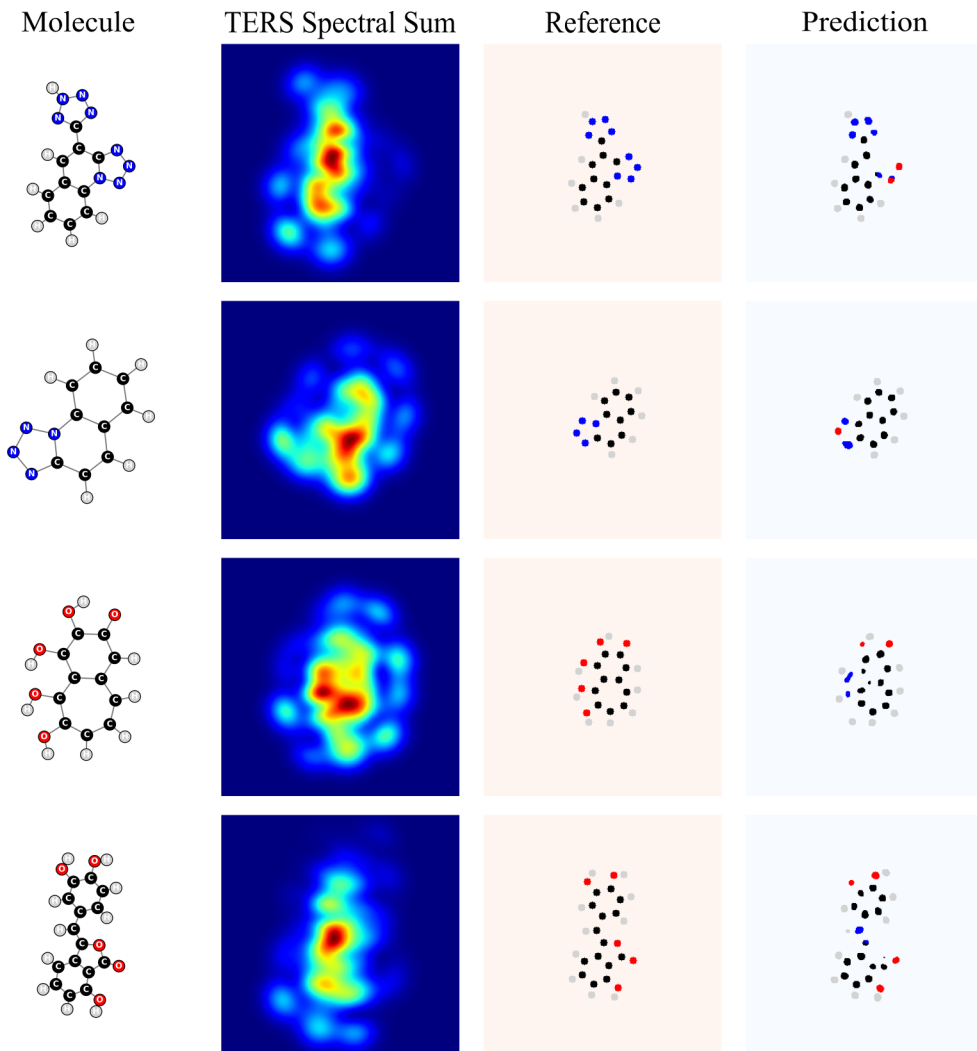


FIG. 11. Representative examples of low-fidelity chemical species predictions (bottom 10% by DSC). Carbon and hydrogen atoms remain well localized, but predictions for oxygen and nitrogen are frequently misplaced or misassigned, reflecting their underrepresentation in the training dataset. Left: Ground Truth. Right: Model Prediction.

closely reproduce the ground-truth atomic maps for all four elements, confirming the model’s ability to jointly resolve molecular geometry and chemical identity under favorable conditions. The bottom decile (Fig. 11), in contrast, highlights the main limitations of the current model. A closer inspection of these cases shows that, even when the overall TERS signal is strong, predictions for carbon and hydrogen remain accurate, whereas predictions for oxygen and nitrogen are markedly less reliable. This reinforces the need to enrich the training set with molecules containing a higher proportion of nitrogen and oxygen, exposing the model to a broader range of functional groups and thereby improving the accuracy of chemical species identification.

IV. CONCLUSIONS

In this work, we introduced a deep learning framework termed SMARTERS capable of predicting molecular structures from simulated TERS hyperspectral images. By training an encoder-decoder network to map vibrational spectra to atomic coordinates, we demonstrate high fidelity in recon-

structing planar molecular geometries from simulated images, going beyond simpler classification challenges [49]. This result establishes the viability of an ML-driven approach to bypass traditional, often ambiguous, manual interpretation, thereby opening an avenue for automated molecular discovery from TERS.

While this study serves as a critical proof of concept, we identify three key areas for future research. First, extending the approach to nonplanar systems remains an open challenge. The core difficulty is that atoms located beneath the surface do not contribute directly to the TERS signal but still influence the vibrational modes through their bonds. We posit that a transition from 2D representations to 3D molecular graphs, processed by graph neural networks, presents the most compelling path forward. Such a framework not only is essential for inferring the complete 3D geometry from the indirect vibrational contributions of subsurface atoms, but would also inherently provide robustness to molecular orientation and help disentangle the true signal from tip-positional artifacts. Second, while we have demonstrated chemically resolved prediction for H, C, N, and O, its reliability on the rarer

TABLE II. Hyperparameter search space and selected model configurations for the *atomic structure prediction* model, for datasets constructed using different RMSD thresholds. For each threshold, we report the configuration achieving the best validation performance for the model trained on the corresponding dataset. Filter configurations define both the number of filters per stage and the encoder/decoder depth: A = [16, 32, 64] (three levels), B = [16, 32, 64, 128] (four levels), C = [16, 32, 64, 128, 256] (five levels), and D = [32, 64, 128, 256, 512] (five levels).

Hyperparameter	Search space	Selected value		
		0.05 Å	0.10 Å	0.50 Å
Input channels	{100, 400}	100	100	400
Encoder/decoder filters	{A, B, C, D}	C	B	D
Attention channels	{16, 32, 64}	16	16	32
Batch size	{16, 32, 64, 128}	16	32	16
Learning rate	$[10^{-5}, 10^{-3}]$	7.35×10^{-4}	5.56×10^{-4}	1.29×10^{-4}

species (N, O) is currently limited by the chemical imbalance of our dataset. Extending this capability will require curating datasets with greater elemental and functional-group diversity, moving toward complete, chemically resolved structure determination. Finally, bridging the sim-to-real gap is paramount for practical application and the current model clearly is not yet capable of routinely handling experimental data (see Appendix D). Experience from ML models in SPM suggests that the model’s ability to generalize to diverse experimental TERS data is contingent on training with larger datasets, enriched with complex augmentations that realistically simulate a wide variety of experimental conditions, including thermal noise, tip variations, and substrate effects (see also Ref. [50]).

ACKNOWLEDGMENTS

The authors thank Krystof Brezina and Mariana Rossi for insightful discussions and helpful feedback. The authors also thank professor Fernando Aguilar-Galindo, professor Martin Švec, and professor Pablo Merino for providing original experimental data. The work was supported by the Finnish Ministry of Education and Culture through the Quantum Doctoral Education Pilot Program (QDOC VN/3137/2024-OKM-4) and the Research Council of Finland through the Finnish Quantum Flagship project (358877, Aalto University). O.J.S. acknowledges the financial support from the Research Council of Finland No. 371875. The authors acknowledge the computational resources provided by the Aalto Science-IT Project and CSC.

DATA AVAILABILITY

The data that support the findings of this study are openly available in Zenodo [51]. The scripts used for data analysis are publicly available on GitHub [52].

APPENDIX A: MODEL

The encoder-decoder structure used for structure discovery for TERS is Attention U-Net model originally introduced by Oktay *et al.* [44]. It uses U-Net architecture [53] with attention mechanism [54] and achieves higher accuracy than the U-Net baseline, as shown in Fig. 12. An overview of the network and the additive attention gate module is shown in Fig. 13. The

model operates on 2D TERS images and follows a symmetric encoder-decoder design [Fig. 13(a)]. Each encoder stage consists of two successive 3×3 convolutional layers with batch normalization and ReLU activation, followed by max pooling with a downsampling factor of 2.

In the encoder, the model learns hierarchical features by progressively reducing the spatial resolution while increasing the number of feature channels. The decoder mirrors this process using bilinear upsampling, concatenation with the corresponding encoder features, and convolutional refinement to recover spatial detail [Fig. 13(a)]. The encoder-decoder depth and channel widths used in each configuration are reported in Table II for the atomic structure prediction model and in Table III for the chemically resolved prediction model. To suppress irrelevant encoder activations, skip connections are filtered using additive attention gates [Fig. 13(b)]. Let $x_l \in \mathbb{R}^{C_l \times H_l \times W_l}$ denote the encoder feature map at level l , and let $g \in \mathbb{R}^{C_g \times H_l \times W_l}$ denote the decoder “gating” feature map from a coarser scale, resized to match the spatial resolution (H_l, W_l). Both are linearly projected using 1×1 convolutions into a common intermediate space of C_{att} channels and combined to compute attention coefficients

$$\alpha_l = \sigma(\psi^\top \text{ReLU}(W_x x_l + W_g g + b)), \quad (\text{A1})$$

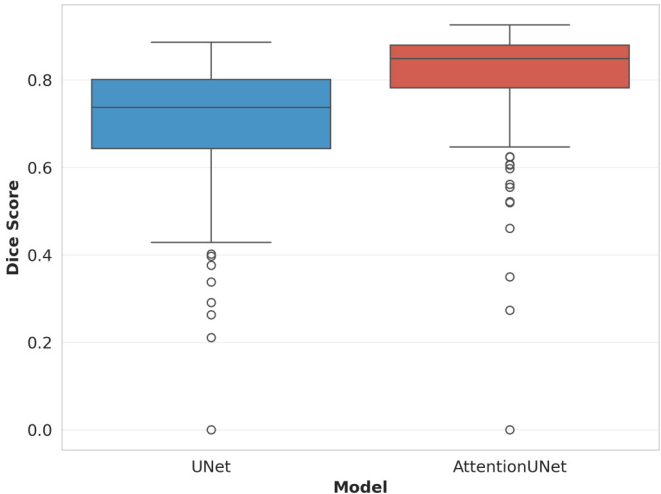


FIG. 12. Performance comparison between standard U-Net and Attention U-Net architectures, on TERS data.

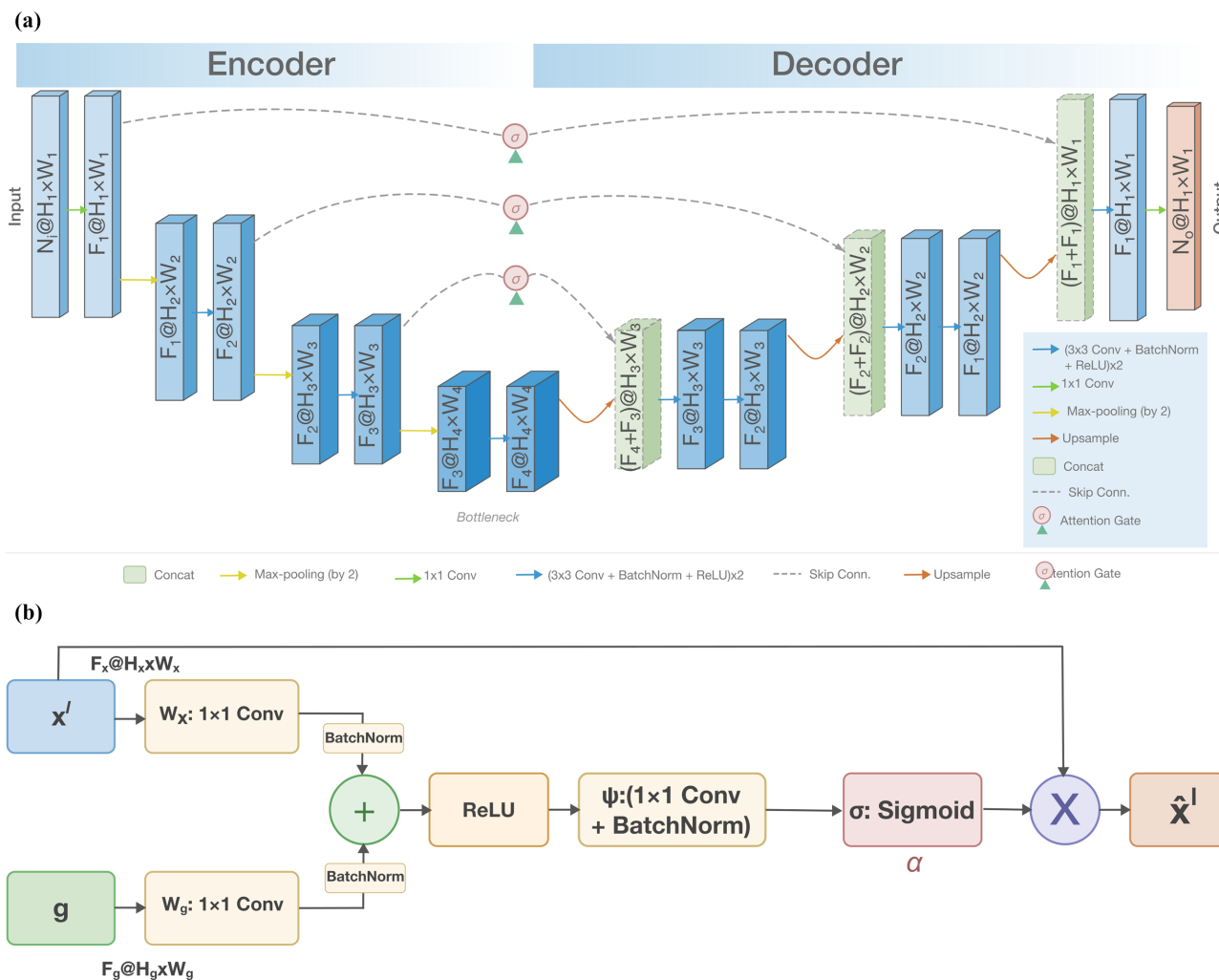


FIG. 13. (a) Attention U-Net architecture used for structure prediction from TERS images, illustrated here with a three-level encoder/decoder depth, together with the additive attention gate applied on the skip connections. F_k denotes the number of filters at stage k , while N_i and N_o denote the input and output channels, respectively; $N_o = 1, 4$ depending on the task (structure vs chemical species prediction). (b) Additive attention gate for skip-connection filtering. Decoder context g is used to compute attention coefficients α_l that scale encoder features x_l , suppressing irrelevant activations prior to concatenation in the decoder.

where W_x and W_g are learned 1×1 convolutional projections, b is a bias term, $\text{ReLU}(\cdot)$ is the rectified linear unit, and $\psi^\top$ denotes a learned 1×1 projection mapping the intermediate representation to a single-channel attention logit. The sigmoid activation produces an attention map $\alpha_l \in [0, 1]^{1 \times H_l \times W_l}$. This

TABLE III. Hyperparameter search space and selected configuration for the *chemically resolved prediction* model, trained on the 0.05 Å dataset. Only the encoder/decoder filters and the learning rate were searched; the remaining hyperparameters were fixed (100 input channels, 64 attention channels, and batch size 16, 50 epochs). Filter configurations define both the number of filters per stage and the encoder/decoder depth: A = [32, 64, 128, 256], B = [64, 128, 256], C = [64, 128, 256, 512], D = [32, 64, 128, 256, 512], E = [64, 128, 256, 512, 1024], and F = [64, 128, 256, 256].

Hyperparameter	Search space	Best value
Encoder/decoder filters	{A, B, C, D, E, F}	D
Learning rate	$[10^{-5}, 10^{-3}]$	2.33×10^{-4}

map is applied channelwise to the encoder features, yielding the gated representation $\tilde{x}_l = \alpha_l \odot x_l$, where $\odot$ denotes elementwise multiplication. The gated features $\tilde{x}_l$ are concatenated with decoder features before convolutional refinement. This mechanism allows decoder context to emphasize spatial regions consistent with the predicted molecular structure while suppressing background or noisy artefacts in simulated TERS images.

In contrast to the original medical-imaging implementation, our model processes 2D inputs with configuration-dependent depth and channel widths. A final 1×1 convolution maps the last decoder features to the output map, whose number of channels N_o depends on the task: $N_o = 1$ for the atomic structure prediction model (a single structure-vs-background mask) and $N_o = 4$ for the chemically resolved prediction model (one channel per element: H, C, N, O). The resulting architecture retains U-Net’s localization capability, while attention gating conditions skip-feature transfer on decoder context, enabling selective integration of spatial detail relevant to the predicted structure.

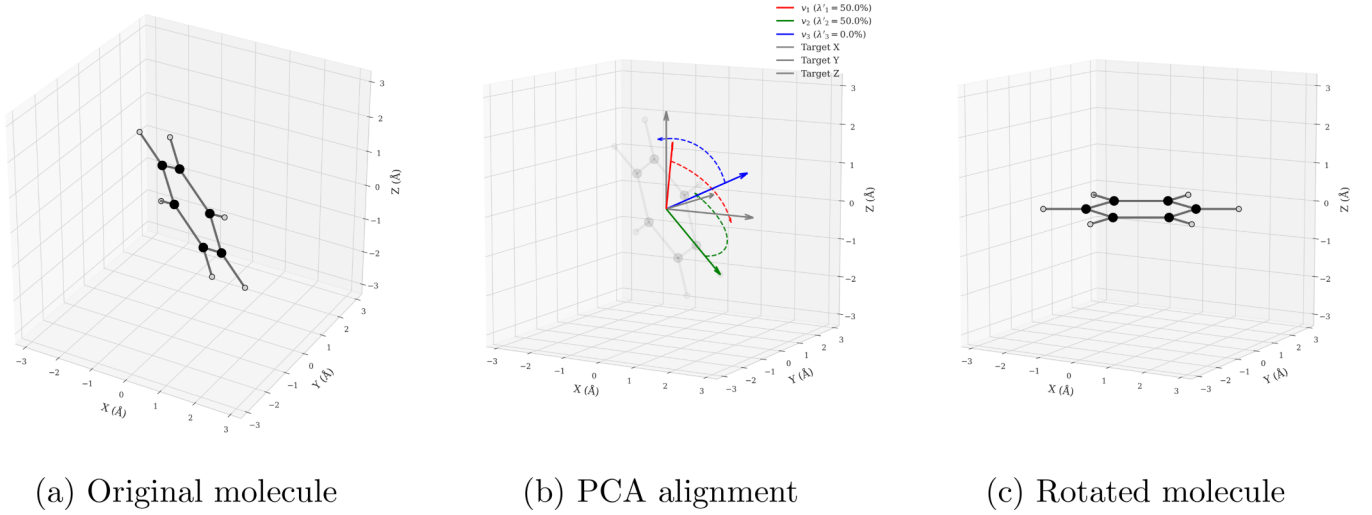


FIG. 14. Dataset preprocessing pipeline. (a) Original 3D coordinates of a molecule. (b) PCA identifies principal axes and eigenvalues; molecules with minimal out-of-plane variance ($\lambda_3 < \epsilon$) are selected as planar. (c) Selected planar molecules are rotated using rotation matrix $\mathbf{R}$ so the molecular plane lies in the XY plane and the tip is along the Z axis, ensuring consistent orientation for TERS simulations.

APPENDIX B: DATASET PREPROCESSING

Principal component analysis for planarity selection and tip alignment

To ensure reliable TERS simulations in the current model, only planar molecules were included in the dataset. Planarity was determined using PCA on the 3D atomic coordinates $\mathbf{X} \in \mathbb{R}^{N \times 3}$, where N is the number of atoms. PCA decomposes the molecular coordinates into orthogonal axes of maximum variance by diagonalizing the covariance matrix:

$$\mathbf{C} = \frac{1}{N}(\mathbf{X} - \bar{\mathbf{X}})^\top (\mathbf{X} - \bar{\mathbf{X}}) \quad (\text{B1})$$

where $\bar{\mathbf{X}}$ is the mean atomic coordinate vector. The eigen-decomposition of $\mathbf{C}$ yields three eigenvalues $\lambda_1 \geq \lambda_2 \geq \lambda_3 \geq 0$ and a corresponding orthonormal set of eigenvectors $\{\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3\}$. These eigenvalues quantify the variance of the atomic coordinates along the principal axes defined by the eigenvectors. The plane spanned by $\{\mathbf{v}_1, \mathbf{v}_2\}$ constitutes the best-fit plane through the molecule, while $\mathbf{v}_3$ is the normal vector to this plane.

Planar molecules were identified by checking the variance along the axis perpendicular to the molecular plane (see Fig. 14). Specifically, a molecule was classified as planar if

$$\lambda_3 < \epsilon, \quad (\text{B2})$$

where λ_3 is the smallest eigenvalue corresponding to the axis with minimal variance and ϵ is a small threshold indicating negligible out-of-plane deviation. This ensures that most atoms lie approximately within a single plane. Equivalently, $\sqrt{\lambda_3}$ corresponds to the root-mean-square out-of-plane deviation of the atoms from the best-fit plane, and the above criterion enforces negligible displacement normal to the plane.

After planar molecules were selected, each molecule was rotated to align its plane perpendicular to the TERS tip. Let $\mathbf{R}$ be the rotation matrix constructed from PCA eigenvectors.

The coordinates were transformed as

$$\mathbf{X}_{\text{aligned}} = \mathbf{R}(\mathbf{X} - \bar{\mathbf{X}}). \quad (\text{B3})$$

The rotation aligns the molecular plane with the XY plane and the tip along the Z axis, providing consistent orientation across all molecules and maximal structural visibility in TERS simulations.

APPENDIX C: EVALUATION METRICS

1. Dice similarity coefficient

The DSC is closely related to intersection over union (IoU) but gives more weight to the overlap between prediction and ground truth. It is defined as

$$\text{Dice}(P, G) = \frac{2|P \cap G|}{|P| + |G|}. \quad (\text{C1})$$

Like IoU, its value lies between 0 and 1, with higher scores indicating greater similarity. Both IoU and DSC provide complementary perspectives on segmentation accuracy, but the DSC emphasizes overlap accurately and is less sensitive to class imbalance.

In our setting, atoms are represented as small, sparse circles on a largely empty background. Although the DSC is a region-overlap measure, in this regime it responds sharply to

TABLE IV. Per-element detection and localization metrics for the chemically resolved prediction model, obtained by LoG blob detection at a binarization threshold of 0.5.

Element	Absolute count error	Precision	Recall	F1	RMSD (Å)
H	0.158	0.983	0.974	0.976	0.085
C	0.560	0.971	0.960	0.961	0.097
N	3.663	0.663	0.657	0.649	0.091
O	4.554	0.696	0.682	0.685	0.096
Mean	2.234	0.828	0.818	0.818	0.092

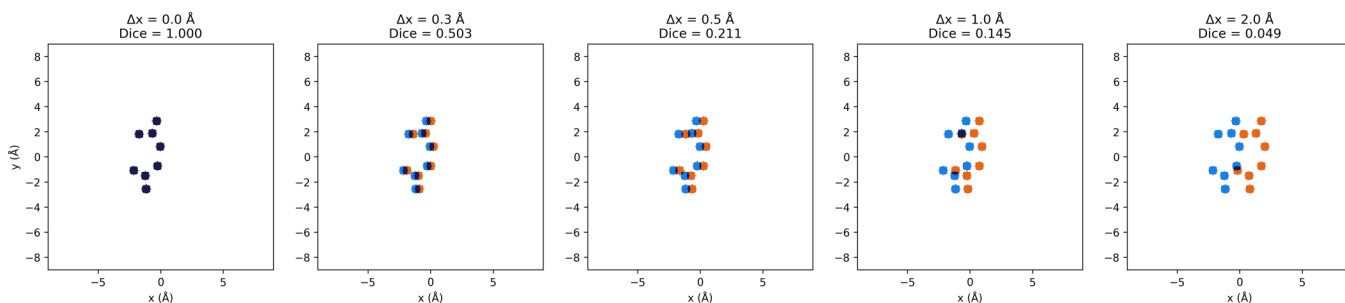


FIG. 15. Sensitivity of the DSC to translational shifts. The DSC between a predicted atomic map and its ground truth drops rapidly as the prediction is shifted, showing that subangstrom positional errors are heavily penalized.

positional error: Figure 15 shows that translating a predicted map relative to the ground truth by even a fraction of an angstrom produces a rapid drop in DSC. A misplaced atom therefore lowers the score substantially rather than being tolerated, so a high DSC reflects accurate localization and not merely a correct atom count.

2. Coordinate-based metrics

While the Dice similarity coefficient [Eq. (C1)] provides a single, threshold-free measure of pixelwise overlap, it does not on its own disentangle different aspects of structural accuracy, such as whether atoms are missed, spuriously detected, or merely displaced. To provide a more detailed picture of model performance, we additionally extract discrete atomic coordinates from the predicted probability maps using a Laplacian-of-Gaussian (LoG) blob-detection algorithm [55,56], allowing us to report atom-count error, precision, recall, and coordinate RMSD.

For the atomic structure prediction model, at a representative binarization threshold of 0.2, this yields a mean atom-count error of 0.38, precision and recall of 0.98, and an atomic coordinate rmsd of 0.097 Å over correctly detected atoms. The sub-0.1 Å localization error confirms that the model recovers atomic positions accurately.

Extracting discrete coordinates in this way, however, requires a binarization threshold whose optimal value is not guaranteed to generalize across molecules with differing signal strengths. As shown in Fig. 16, the number of detected atoms varies considerably with the threshold, and Fig. 17 shows that both the atom-count error and the coordinate RMSD become volatile and artificially dependent on the chosen value. Relying on such threshold-dependent metrics would compromise the fully automated nature of the pipeline

and obscure the network’s raw, threshold-independent predictive capability. We therefore retain threshold-agnostic metrics such as the DSC as the primary measure of performance, and report these coordinate-level metrics as a complementary, corroborating view.

The same analysis applied to the chemically resolved prediction model is reported per element in Table IV, with the corresponding threshold dependence shown in Fig. 18. For this model, we use a binarization threshold of 0.5, which gave the best results across the values we tested (Fig. 18). Hydrogen and carbon are detected and localized reliably (precision and recall ≈ 0.96 – 0.98 , atom-count error < 0.6), whereas nitrogen and oxygen show substantially higher atom-count errors and lower precision and recall. This mirrors the DSC trend reported in Sec. III C and reflects the same underlying cause, the underrepresentation of N and O in the training set, rather than a localization failure: When these atoms are detected they are placed accurately, as the coordinate RMSD remains uniformly low (~ 0.09 Å) across all four elements.

APPENDIX D: EXPERIMENTAL DATASET

We performed an ablation study of our geometry prediction model using both experimental and simulated data for the phthalocyanine molecule with an Fe center (FePc). The experimental dataset used in our geometry prediction has been already published in Ref. [15], corresponding to TERS images of the FePc adsorbed on a Ag(110) surface in an aligned fashion. The authors kindly provided us with .stp files of the TERS hyperspectral maps taken at high-intensity vibrations ranging from 300 to 2000 cm^{-1} , whereas the simulated data span a broad range of 0–3200 cm^{-1} , the full range of FePc vibrational modes. Figure 19 shows the geometry prediction on the simulated data, as well as two attempts on the

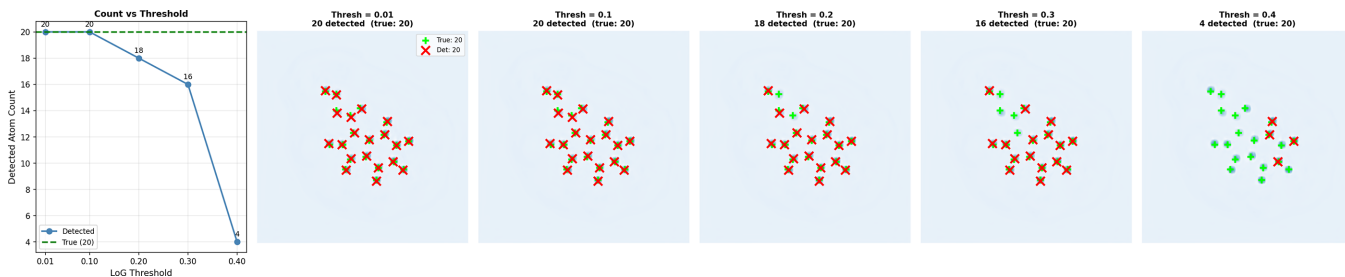


FIG. 16. Sensitivity of the predicted atom count to the binarization threshold used in LoG blob extraction.

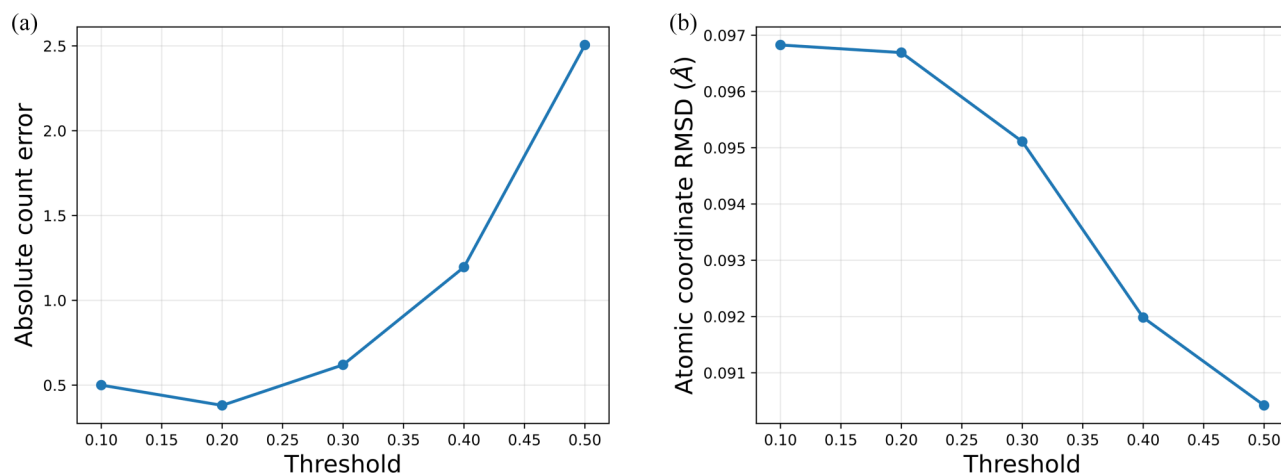


FIG. 17. Dependence of coordinate based evaluation metrics on the LoG binarization threshold. (a) Mean absolute atom-count error vs threshold. (b) Atomic coordinate RMSD vs threshold.

experimental data: on the raw experimental data and after applying a gaussian function to denoise the image.

Regarding the simulated data, our model successfully predicts the periphery of the phthalocyanine molecule when the broad range of vibrational modes is used, although the predictions at the center of the molecule deviate noticeably from the ground truth. This behavior is already expected since Fe atom is not present in our dataset, limiting the capability of our model. Restricting the amount of simulated TERS images to the same frequency range of the experiment drastically deviates the geometry prediction from the ground truth, which is also expected since it does not correspond to the range in which the model was trained on.

Regarding the experimental data, Fig. 19 already shows a large mismatch between the simulated and experimental TERS spectral sum. Such mismatch is attributed to the simplifications made to acquire the simulation data, as opposed to the complex experimental environment faced by the FePc molecule. For example, our model considers a relative sharp tip (5 Å FWHM gaussian function) that is more appropri-

ate for structural prediction at the angstrom level, whereas a blunter tip was most likely used in such experimental scenario. Additionally, it is also unknown how heavy is the influence of tip changes during experiment acquisition in the TERS images. Secondly, substrate effects also drastically change some features in the TERS images, which are not present in our current model. Because of those discrepancies, the geometry prediction on the raw experimental data consists of several incongruent atomic positions that do not resemble the structure of the phthalocyanine molecule, and applying a denoise function drastically reduces the number of incongruent atomic positions, placing only a few at random locations.

Given the remarks highlighted in the previous paragraphs, it is natural to expect that the model fails for the experimental images, especially given its limitations compared to the experiment. While our work still provides a proof of concept that TERS maps can be used for structural prediction, more robust models are necessary to achieve a better performance on experimental images.

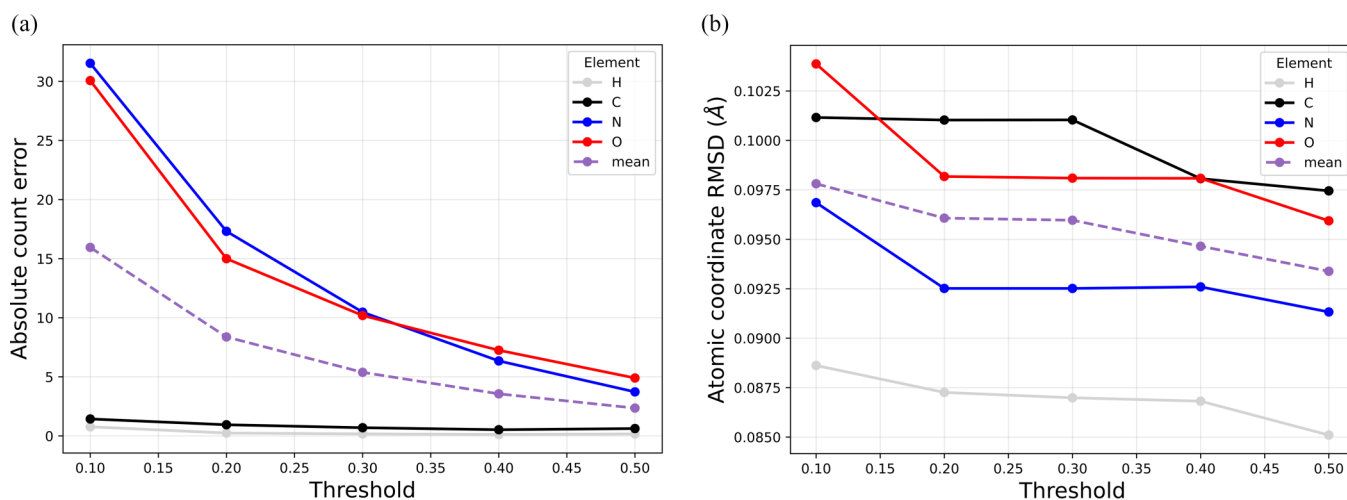


FIG. 18. Dependence of the per-element coordinate based metrics for the chemically resolved prediction model on the LoG binarization threshold. (a) Absolute atom-count error vs threshold. (b) Atomic coordinate RMSD vs threshold.

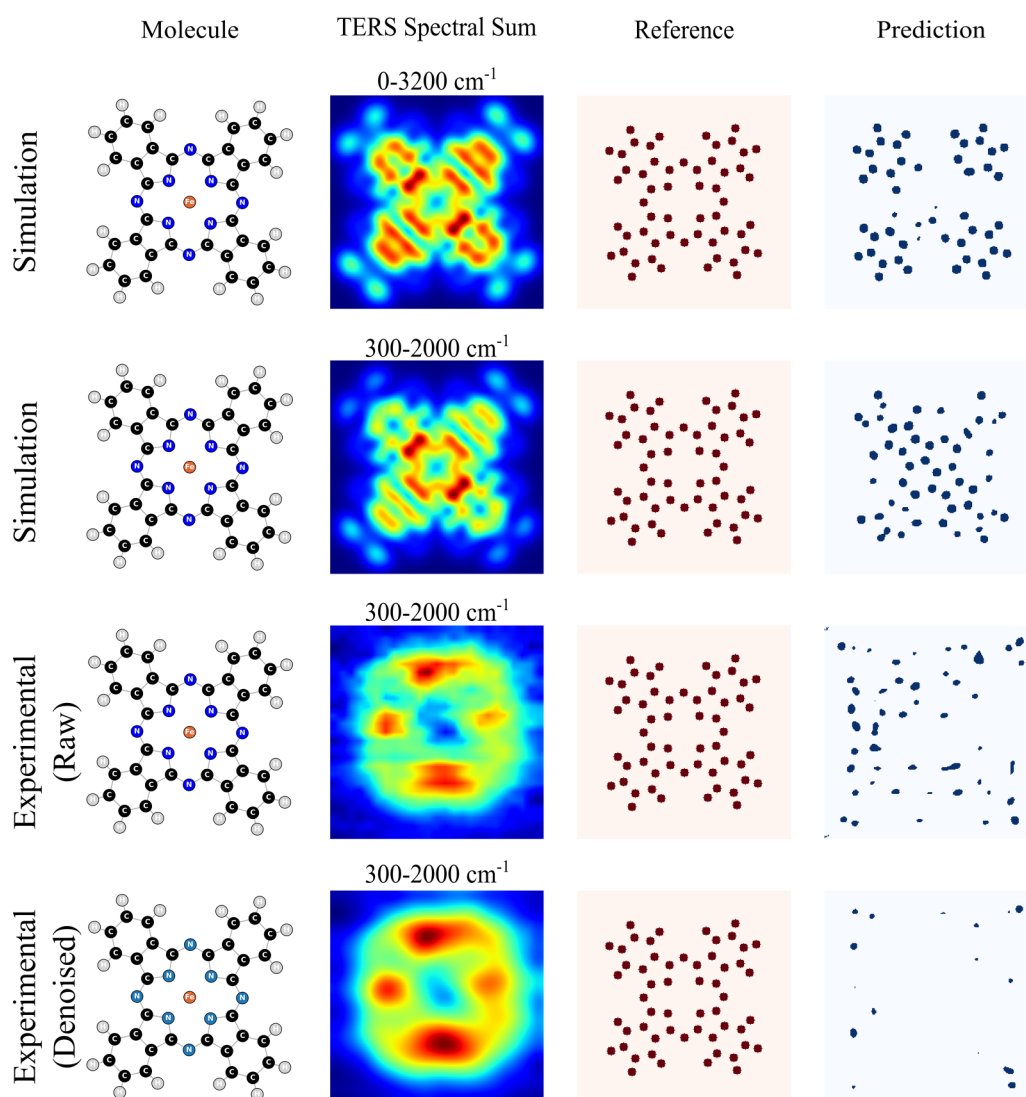


FIG. 19. Comparison of geometry predictions for FePc on Ag(110) using simulated and experimental TERS data.

- [1] F. Mohn, J. Repp, L. Gross, G. Meyer, M. S. Dyer, and M. Persson, Reversible bond formation in a gold-atom-organic-molecule complex as a molecular switch, *Phys. Rev. Lett.* **105**, 266102 (2010).
- [2] P. Hapala, G. Kichin, C. Wagner, F. S. Tautz, R. Temirov, and P. Jelínek, Mechanism of high-resolution STM/AFM imaging with functionalized tips, *Phys. Rev. B* **90**, 085421 (2014).
- [3] K. Sun, O. J. Silveira, Y. Ma, Y. Hasegawa, M. Matsumoto, S. Kera, O. Krejčí, A. S. Foster, and S. Kawai, On-surface synthesis of disilabenzene-bridged covalent organic frameworks, *Nat. Chem.* **15**, 136 (2023).
- [4] N. Oinonen, A. V. Yakutovich, A. Gallardo, M. Ondráček, P. Hapala, and O. Krejčí, Advancing scanning probe microscopy simulations: A decade of development in probe-particle models, *Comput. Phys. Commun.* **305**, 109341 (2024).
- [5] L. Gross, F. Mohn, N. Moll, P. Liljeroth, and G. Meyer, The chemical structure of a molecule resolved by atomic force microscopy, *Science* **325**, 1110 (2009).
- [6] L. Gross, F. Mohn, N. Moll, B. Schuler, A. Criado, E. Guitián, D. Peña, A. Gourdon, and G. Meyer, Bond-order discrimination by atomic force microscopy, *Science* **337**, 1326 (2012).
- [7] F. Schulz, J. Ritala, O. Krejčí, A. P. Seitsonen, A. S. Foster, and P. Liljeroth, Elemental identification by combining atomic force microscopy and Kelvin probe force microscopy, *ACS Nano* **12**, 5274 (2018).
- [8] E. Abbe, VII.—On the estimation of aperture in the microscope, *J. R. Microsc. Soc.* **1**, 388 (1881).
- [9] T. Itoh, M. Procházka, Z.-C. Dong, W. Ji, Y. S. Yamamoto, Y. Zhang, and Y. Ozaki, Toward a new era of SERS and TERS at the nanometer scale: From fundamentals to innovative applications, *Chem. Rev.* **123**, 1552 (2023).
- [10] A. B. Zrimsek, N. Chiang, M. Mattei, S. Zaleski, M. O. McAnally, C. T. Chapman, A.-I. Henry, G. C. Schatz, and R. P. Van Duyne, Single-molecule chemistry with surface- and tip-enhanced Raman spectroscopy, *Chem. Rev.* **117**, 7583 (2017).

- [11] B. Yang, G. Chen, A. Ghafoor, Y.-F. Zhang, X.-B. Zhang, H. Li, X.-R. Dong, R.-P. Wang, Y. Zhang, Y. Zhang, and Z.-C. Dong, Chemical enhancement and quenching in single-molecule tip-enhanced Raman spectroscopy, *Angew. Chem. Int. Ed.* **62**, e202218799 (2023).
- [12] J. Lee, K. T. Crampton, N. Tallarida, and V. A. Apkarian, Visualizing vibrational normal modes of a single molecule with atomically confined light, *Nature (London)* **568**, 78 (2019).
- [13] S. Jiang, X. Zhang, Y. Zhang, C. Hu, R. Zhang, Y. Zhang, Y. Liao, Z. J. Smith, Z. Dong, and J. G. Hou, Subnanometer-resolved chemical imaging via multivariate analysis of tip-enhanced Raman maps, *Light Sci. Appl.* **6**, e17098 (2017).
- [14] T. Kumagai, K. Miwa, and B. Cirera, Point-contact tip-enhanced Raman spectroscopy: Picoscale light-matter interactions within plasmonic cavities, *Nano Lett.* **25**, 15449 (2025).
- [15] R. C. de Campos Ferreira, B. Cirera, J. Doležal, Á. G. de Roa, A. Sagwal, P. Kahan, R. Canales, F. Aguilar-Galindo, M. Švec, and P. Merino, Adsorption-driven symmetry lowering in single molecules revealed by Ångström-scale tip-enhanced Raman imaging, *J. Am. Chem. Soc.* **148**, 9375 (2026).
- [16] C. Höppener, J. Aizpurua, H. Chen, S. Gräfe, A. Jorio, S. Kupfer, Z. Zhang, and V. Deckert, Tip-enhanced Raman scattering, *Nat. Rev. Methods Primers* **4**, 47 (2024).
- [17] Y. Zhang, Z.-C. Dong, and J. Aizpurua, Theoretical treatment of single-molecule scanning Raman picoscopy in strongly inhomogeneous near fields, *J. Raman Spectrosc.* **52**, 296 (2021).
- [18] S. Duan, G. Tian, Y. Ji, J. Shao, Z. Dong, and Y. Luo, Theoretical modeling of plasmon-enhanced Raman images of a single molecule with subnanometer resolution, *J. Am. Chem. Soc.* **137**, 9515 (2015).
- [19] Y. Zhang, R. Esteban, J. Cui, and J. Aizpurua, Software to implement simulations of Raman spectra of single-molecules in an inhomogeneous field, 2023, <https://doi.org/10.20350/digitalCSIC/15303>.
- [20] Y. Litman, F. P. Bonafé, A. Akkoush, H. Appel, M. Rossi, First-principles simulations of tip enhanced Raman scattering reveal active role of substrate on high-resolution images, *J. Phys. Chem. Lett.* **14**, 6850 (2023).
- [21] X. Chen, P. Liu, Z. Hu, and L. Jensen, High-resolution tip-enhanced Raman scattering probes sub-molecular density changes, *Nat. Commun.* **10**, 2567 (2019).
- [22] F. Qiu, Z.-Y. Gong, D. Cao, C. Song, G. Tian, S. Duan, and Y. Luo, Optical images of molecular vibronic couplings from tip-enhanced fluorescence excitation spectroscopy, *JACS Au* **2**, 150 (2022).
- [23] H.-Z. Yu, R.-L. Han, D. Chu, Y. Li, X.-R. Dong, Y. Zhang, L. Wang, Y. Song, C.-K. Wang, Z. Xie, and S. Duan, Theoretical prediction for monitoring Jahn-Teller vibrational evolution using real-space tip-enhanced Raman imaging, *Nat. Commun.* **17**, 1132 (2026).
- [24] K. Brezina, Y. Litman, and M. Rossi, Tip-enhanced Raman images of realistic systems through *ab initio* modeling, *ACS Nano* **20**, 5550 (2026).
- [25] X. Chen, S. Xu, S. Shabani, Y. Zhao, M. Fu, A. J. Millis, M. M. Fogler, A. N. Pasupathy, M. Liu, and D. N. Basov, Machine learning for optical scanning probe nanoscopy, *Adv. Mater.* **35**, 2109171 (2023).
- [26] N. Kumar, L.-Q. Zheng, A. J. Pollard, A. J. Wain, and R. Zenobi, Nanoscale chemical analysis of heterogeneous catalysts using tip-enhanced Raman spectroscopy, *Chem. Rev.* **126**, 2653 (2026).
- [27] M. Ziatdinov, A. Maksov, and S. V. Kalinin, Learning surface molecular structures via machine vision, *npj Comput. Mater.* **3**, 31 (2017).
- [28] A. Khan, C.-H. Lee, P. Y. Huang, and B. K. Clark, Leveraging generative adversarial networks to create realistic scanning transmission electron microscopy images, *npj Comput. Mater.* **9**, 85 (2023).
- [29] K. Choudhary, Microscopygpt: Generating atomic-structure captions from microscopy images of 2D materials with vision-language transformers, *J. Phys. Chem. Lett.* **16**, 7028 (2025).
- [30] S. V. Kalinin, E. Strelcov, A. Belianinov, S. Somnath, R. K. Vasudevan, E. J. Lingerfelt, R. K. Archibald, C. Chen, R. Proksch, N. Laanait, S. Jesse, Big, deep, and smart data in scanning probe microscopy, *ACS Nano* **10**, 9068 (2016).
- [31] Md A. R. Laskar and U. Celano, Scanning probe microscopy in the age of machine learning, *APL Mach. Learn.* **1**, 041501 (2023).
- [32] B. Alldritt, P. Hapala, N. Oinonen, F. Urtev, O. Krejci, F. F. Canova, J. Kannala, F. Schulz, P. Liljeroth, and A. S. Foster, Automated structure discovery in atomic force microscopy, *Sci. Adv.* **6**, eaay6913 (2020).
- [33] B. Tang, Y. Song, M. Qin, Y. Tian, Z. W. Wu, Y. Jiang, D. Cao, and L. Xu, Machine learning-aided atomic structure identification of interfacial ionic hydrates from AFM images, *Natl. Sci. Rev.* **10**, nwac282 (2023).
- [34] N. Oinonen, L. Kurki, A. Il'in, and A. S. Foster, Molecule graph reconstruction from atomic force microscope images with machine learning, *MRS Bull.* **47**, 895 (2022).
- [35] L. Kurki, N. Oinonen, and A. S. Foster, Automated structure discovery for scanning tunneling microscopy, *ACS Nano* **18**, 11130 (2024).
- [36] F. Priante, N. Oinonen, Y. Tian, D. Guan, C. Xu, S. Cai, P. Liljeroth, Y. Jiang, and A. S. Foster, Structure discovery in atomic force microscopy imaging of ice, *ACS Nano* **18**, 5546 (2024).
- [37] J. Carracedo-Cosme, C. Romero-Muñiz, P. Pou, and R. Pérez, Molecular identification from AFM images using the IUPAC nomenclature and attribute multimodal recurrent neural networks, *ACS Appl. Mater. Interfaces* **15**, 22692 (2023).
- [38] J. Carracedo-Cosme, R. Pérez, Molecular identification with atomic force microscopy and conditional generative adversarial networks, *npj Comput. Mater.* **10**, 19 (2024).
- [39] S. Duan, G. Tian, and Y. Luo, Visualization of vibrational modes in real space by tip-enhanced non-resonant Raman spectroscopy, *Angew. Chem. Int. Ed.* **55**, 1041 (2016).
- [40] Y. Li, D. Chu, W. Ma, H.-Z. Yu, Y. Song, C.-K. Wang, S. Duan, and Z. Xie, Visualization of hydrogen isotope exchange in single molecule by tip-enhanced Raman images, *J. Chem. Phys.* **163**, 224310 12 (2025).
- [41] N. Oinonen, C. Xu, B. Alldritt, F. F. Canova, F. Urtev, S. Cai, O. Krejčí, J. Kannala, P. Liljeroth, and A. S. Foster, Electrostatic discovery atomic force microscopy, *ACS Nano* **16**, 89 (2022).
- [42] M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, *et al.*, Gaussian 16 Revision C.01 (Gaussian Inc., Wallingford, CT, 2016).

- [43] V. Badrinarayanan, A. Kendall, and R. Cipolla, SegNet: A deep convolutional encoder-decoder architecture for image segmentation, *IEEE Trans. Pattern Anal. Mach. Intell.* **39**, 2481 (2017).
- [44] O. Oktay, J. Schlemper, L. L. Folgoc, M. Lee, M. Heinrich, K. Misawa, K. Mori, S. McDonagh, N. Y. Hammerla, B. Kainz, B. Glocker, and D. Rueckert, Attention U-Net: Learning where to look for the pancreas, in *Medical Imaging with Deep Learning* (OpenReview Foundation, Amherst, Massachusetts, 2018), <https://openreview.net/forum?id=Skft7cijM>.
- [45] D. P. Kingma and J. Ba, Adam: A method for stochastic optimization, in *Proceedings of the 3rd International Conference on Learning Representations* (ICLR, San Diego, 2015).
- [46] L. R. Dice, Measures of the amount of ecologic association between species, *Ecology* **26**, 297 (1945).
- [47] T. Akiba, S. Sano, T. Yanase, T. Ohta, and M. Koyama, Optuna: A next-generation hyperparameter optimization framework, in *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining* (ACM Press, New York, NY, 2019).
- [48] A. Paszke, S. Gross, F. Massa, A. Lerer, J. Bradbury, G. Chanan, T. Killeen, Z. Lin, N. Gimelshein, L. Antiga, *et al.*, PyTorch: An imperative style, high-performance deep learning library, in *Advances in Neural Information Processing Systems 32, Vancouver, BC* (Curran Associates, Inc., 2019), pp. 8024–8035.
- [49] Y. Tian, G. Cheng, G. Shen, X. Zhu, J. Xu, A.-I. Jia, Z. Liu, H. Du, S. Tan, and B. Wang, Autonomous single-molecule chemical identification with an AI-enabled STM-TERS platform, *Precis. Chem.* (2026), doi:10.1021/prechem.6c00036.
- [50] J. Cui, Y. Zhang, Y. Zhang, Y. Luo, and Z.-C. Dong, TERS-ABNet: A deep learning approach for automated single-molecule structure reconstruction with atomic precision from TERS mapping, [arXiv:2603.21579](https://arxiv.org/abs/2603.21579).
- [51] H. Sethi, M. K. Junttila, O. Silveira, A. Foster, Training dataset for SMARTERS: A machine learning model for automated structure discovery in tip-enhanced Raman spectroscopy [Data set], Zenodo, 2026, <https://doi.org/10.5281/zenodo.18768437>.
- [52] H. Sethi, M. K. Junttila, O. Silveira, A. Foster, SMARTERS (structural machine learning for high-resolution tip-enhanced Raman spectroscopy), Github, 2026, <https://github.com/SINGROUP/SMARTERS>.
- [53] O. Ronneberger, P. Fischer, and T. Brox, U-Net: Convolutional networks for biomedical image segmentation, in *Medical Image Computing and Computer-Assisted Intervention (MICCAI)* (Springer, Cham, Switzerland, 2015), pp. 234–241.
- [54] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, L. Kaiser, and I. Polosukhin, Attention is all you need, in *Advances in Neural Information Processing Systems 30, Long Beach, California* (Curran Associates, Inc., 2017), pp. 5998–6008.
- [55] T. Lindeberg, Feature detection with automatic scale selection, *Int. J. Comput. Vision* **30**, 79 (1998).
- [56] D. Marr and E. Hildreth, Theory of edge detection, *Proc. R. Soc. London, Ser. B Biol. Sci.* **207**, 187 (1980).