Supplementary Information (SI) for Digital Discovery.

SUPPLEMENTARY INFORMATION

Mol2Raman: a graph neural network model for predicting Raman spectra from SMILES representations

Salvatore Sorrentino,*[a] Alessandro Gussoni,[b] Francesco Calcagno,[c,d] Gioele Pasotti,[a] Davide Avagliano,[e] Ivan Rivalta,[c,d] Marco Garavelli,[c] Dario Polli *[a,f]

a. Department of Physics, Politecnico di Milano, Milan, 20133, Italy
b. Citizen Scientist
c. Department of Industrial Chemistry "Toso Montanari", Universita' degli Studi di Bologna, Via Piero Gobetti, 85, I-40129 Bologna, Italy
d. Center for Chemical Catalysis - C3, Alma Mater Studiorum University of Bologna, via Piero Gobetti 85, 40129 Bologna, Italy
e. Chimie ParisTech, PSL University, CNRS, Institute of Chemistry for Life and Health Sciences (iCLeHS UMR 8060), 75005 Paris, France
f. CNR-Institute for Photonics and Nanotechnologies (CNR-IFN), Milan, Italy

** Correspondence: salvatore.sorrentino@polimi.it (S.S), dario.polli@polimi.it (D.P.)*

Supplementary Note 1 Mol2Raman Computational Time of Calculations. All models presented in this paper are trained on an NVIDIA GeForce RTX 3090 GPU, enabling efficient training with competitive runtimes. The two networks predicting the number of Raman-active frequencies requires approximately 6 hours each for training, while the two networks predicting Raman activity completes training in about 9 hours each. Once trained, the inference time for all 3,168 molecules in the test dataset is approximately 70 seconds, which corresponds to an average of 22 milliseconds per molecule. This represents a substantial improvement over DFT calculations, which require approximately 3 hours per molecule on a standard PC.

These results demonstrate that the *Mol2Raman* architecture effectively captures complex molecular features, enabling accurate Raman spectral predictions while maintaining computational efficiency. The combination of precise predictions, even though not perfect with respect to DFT, and the extremely fast inference time (<< 1s) compared to the hours required for DFT calculations, is the key strength of *Mol2Raman*.

Supplementary Note 2 Analysis of model performances for C-H and fingerprint region in the prediction of the number of Raman-active frequencies. The higher R^2 value observed for the fingerprint region (see Table 1) is primarily due to the greater asymmetry in the distribution of Raman-active frequencies, as indicated by the skewness values reported in Supplementary Figures S2 and S3. This skewness results in a more populated upper tail in the fingerprint region, meaning that molecules with a larger number of peaks are more frequent. Since R^2 is sensitive to larger absolute deviations, the model's ability to accurately predict these high-count cases boosts the R^2 metric in the fingerprint region compared to the C–H region.

Conversely, the better performance observed in the C–H stretching region in terms of accuracy (0.458 vs 0.349) and RMSE (1.089 vs 1.282) is mainly attributed to the narrower distribution of the number of Raman-active frequencies, as shown in Supplementary Figures S2 and S3. A tighter distribution reduces the variance in the output, making the prediction task statistically easier and minimizing the spread of errors.

Supplementary Note 3 Peak-conditioned attributions for atom and bond relevance

To visualize which parts of a molecule most influence a specific predicted Raman peak, we computed peak-conditioned Integrated Gradients (IG)[1] on the trained graph network. For a target wavenumber bin j(selected as the largest peak either in the fingerprint region ~500–1800 cm^{-1} or in the C–H stretching region ~2800–3050 cm^{-1}), we treated the scalar model output $\hat{y}_j$ as the function of interest and integrated input gradients along a straight path from a zero baseline to the molecule's input. Atom-level scores were obtained by summing the absolute IG values across all node-feature channels for each atom; bond scores were defined as the mean of the scores of the two incident atoms. Scores for every atom were normalized, and we report to them as "relative attribution (0–1)." These attributions quantify model sensitivity to atoms and bonds for that peak; they are explanations of the predictor, not normal-mode vectors.

For the peak with the largest intensity in the fingerprint window, attributions are distributed across the full molecular scaffold, as can be see both in Supplementary Figures 15 and 16, with strong weights on multiple ring atoms and the connecting bonds. This pattern is consistent with the fact that fingerprint bands often arise from concerted, molecule-wide vibrations (e.g., coupled ring deformations and C–C/C–O bends), so the network draws information from many atoms to set the intensity at that bin. By contrast, for the largest peak in the C–H region, the highest attributions concentrate on aliphatic carbon atoms and their adjacent bonds, with comparatively low weights on heteroatoms, matching the expected CH_2/CH_3 symmetric/asymmetric stretching origin of this band, as can be seen in Figure 5 and Supplementary Figures 15. Together, these qualitative maps indicate that the model's decisions for representative peaks align with chemically plausible assignments: broad, delocalized contributions in the fingerprint and localized C–H stretching contributions in the high-wavenumber region. We note that attribution magnitudes depend on the chosen baseline and within-molecule normalization; consequently, colors should be interpreted relatively inside the molecule, not as absolute contributions.

Supplementary Table 1 - Molecular composition for the entire dataset and for the test dataset. Summary of molecular composition for the full dataset and the test subset. The table reports the number of molecules containing at least one fluorine, oxygen, or nitrogen atom. Additionally, it includes counts of molecules exhibiting specific structural features—namely the presence or absence of ring systems, chirality, and conjugation.

Property	Full Dataset	Test Dataset
Contains Fluorine	484	44
Contains Oxygen	27129	2703
Contains Nitrogen	19416	1940
Has Ring	28456	2837
No Ring	3320	331
Chiral	22420	2199
Non-Chiral	9356	969
Conjugated	13995	1399
Non-Conjugated	17781	1769

Supplementary Table 2 - Ablation study on global molecular descriptors.

We evaluated the contribution of global molecular descriptors to Mol2Raman by comparing four variants of the activity-prediction network in the entire spectral region: using only the predicted number of Raman-active peaks; Number of peaks + Daylight fingerprint ; Number of peaks + Morgan fingerpring; and Number of peaks + Daylight fingerprint + Morgan fingerprint (i.e. the actual architecture used in this work). Performance was assessed by F1 with ±15 cm^{-1} tolerance—our peak-localization metric—and by Spectral Information Similarity (SIS) and Cosine Similarity on Lorentzian-convolved spectra. The results are reported in the Table below.

Adding Daylight fingerprint to the Number of Raman active peaks improved spectral-shape similarity (SIS from 0.646 → 0.665; Cosine Similarity from 0.716 → 0.732) and slightly increased F1 with ±15 cm^{-1} tolerance (0.633 → 0.640). Morgan fingerprint alone did not surpass the Number of peaks-only baseline (F1 with ±15 cm^{-1} tolerance = 0.629; SIS = 0.637; Cosine Similarity = 0.708). Crucially, combining Daylight and Morgan fingerprints with Number of Raman active peaks yielded the best results across all metrics— F1 with ±15 cm^{-1} tolerance = 0.642, SIS = 0.669, Cosine Similarity = 0.735—indicating complementarity: Daylight encodes path-based/global motifs, while Morgan captures circular/local neighborhoods; together they enrich the global embedding beyond what either provides in isolation. These trends align with the model's design, which integrates global descriptors with graph-derived local features and the predicted number of Raman-active modes to jointly guide both peak placement and intensity distribution.

The inclusion of Daylight and Morgan fingerprints does not increase inference throughput time in practice, so we keep the combination Number of predicted Raman peaks + Daylight fingerprint + Morgan fingerprint as the default configuration given its consistent improvements in F1 with ±15 cm^{-1} tolerance, SIS, and Cosine Similarity.

Global molecular description employed	F1 Tol. 15 cm^{-1}	SIS	Cosine Similarity
Number of Raman peaks	0.633	0.646	0.716
Number of Raman peaks + Daylight fingerprint	0.640	0.665	0.732
Number of Raman peaks + Morgan fingerprint	0.629	0.637	0.708
Number of Raman peaks + Daylight fingerprint + Morgan fingerprint	0.642	0.669	0.735

Supplementary Table 3 — Ablation study on a 3D global descriptor

To probe whether explicitly 3D-aware features help Raman prediction, we augmented the global molecular description with a Smooth Overlap of Atomic Positions (SOAP) vector provided in the *DScribe* library[2]. SOAP encodes each atom's local 3D environment by expanding the neighbor density into radial basis functions and spherical harmonics, then forming rotation-invariant power spectra from the overlaps of these expansions. When pooled across atoms, SOAP yields a fixed-length, species-aware global descriptor that reflects molecular shape and symmetry while remaining invariant to rigid translations and rotations. This descriptors adds 1,261 components to the global feature block used by the network predicting the Raman activity. We then evaluated two variants over the entire spectral range with the same splits and metrics used elsewhere (F1 with ±15 cm^{-1} tolerance, SIS and Cosine Similarity on Lorentzian-convolved spectra):

1) Mol2Raman + SOAP (global) — i.e., our standard architecture (GINE + global features) with SOAP appended to the global vector — yields F1 with 15cm^{-1} tolerance = 0.6363, SIS = 0.6514, Cosine Similarity = 0.7230. This is *slightly below* the default configuration without SOAP (F1 with 15cm^{-1} tolerance = 0.642; SIS = 0.669; Cosine Similarity = 0.735), suggesting modest degradation when the much larger global input increases parameter count in the first dense layers and introduces additional redundancy/noise leading to a small overfitting which reduces the global performances.
2) SOAP-only MLP (global-only) — i.e., a small fully connected network composed of 3 fully connected layer over SOAP description + the other global features, without GINE — performs F1 with 15cm^{-1} tolerance = 0.5269, SIS = 0.5761, Cosine Similarity = 0.6828, confirming that removing atom/bond-level message passing markedly hurts accuracy because purely global summaries cannot capture local, collective vibrational patterns that our graph encoder learns.

Mechanistically, these outcomes align with the design of Mol2Raman, which couples local graph features (GINE) with compact global descriptors (Daylight + Morgan + predicted number of peaks). Appending a long SOAP vector inflates the global block (and thus the subsequent dense layers) without providing commensurate new information for our task, leading to mild overfitting/optimization confusion; conversely, discarding the graph pathway removes essential local chemistry.

Moreover, we believe that further gains in the fingerprint region, where geometry is most critical, may require a model natively operating on 3D coordinates rather than adding a pooled 3D descriptor. However this is beyond the scope of our work, because it would forgo the SMILES-only, user-friendly workflow and would require in input a per-molecule geometry optimization to know the exact coordinates of atoms in a molecule.

Model (global features)	F1 Tol. 15 cm^{-1}	SIS	Cosine Similarity
Mol2Raman + SOAP (with GINE)	0.6363	0.6514	0.7230
SOAP-only MLP (no GINE)	0.5269	0.5761	0.6828

Supplementary Table 4 — Subgroup-Stratified Split Analysis

To verify that known skews in the dataset (fluorination and chirality) do not unduly influence evaluation, we constructed a subgroup-stratified 80/10/10 split preserving the joint distribution of (i) fluorination (presence of ≥1 F atom) and (ii) chirality (presence of ≥1 chiral center) across train/validation/test. We then retrained the model with identical hyperparameters and evaluated on the stratified test set. The overall dataset proportions for the four strata is the following:

No Fluorine × Achiral (F–, non-chiral)	= 28.19
Fluorine × Achiral (F+, non-chiral)	= 1.35
No Fluorine × Chiral (F–, chiral)	= 70.45
Fluorine × Chiral (F+, chiral)	= 0.10

Test performance on the stratified split is reported in Table. While the aggregate F1 scores are slightly lower than those obtained under a single random split, they are comparable in magnitude. We attribute the modest reduction to the constraints imposed by stratification across all four partitions, which (especially for rare subgroups) can marginally reduce the diversity/quantity of similar training examples available for those chemistries—an effect known to yield small dips in single-split performance. Importantly, qualitative findings and relative model comparisons remain unchanged: peak-placement trends across baselines are consistent.

Splitting strategy	F1 Tol. 10 cm^{-1}	F1 Tol. 15 cm^{-1}	F1 Tol. 20 cm^{-1}
Stratify Splitting	0.50	0.58	0.65
Random Splitting	0.56	0.64	0.71

Supplementary Table 5 — Ablation study using GATv2Conv Layers

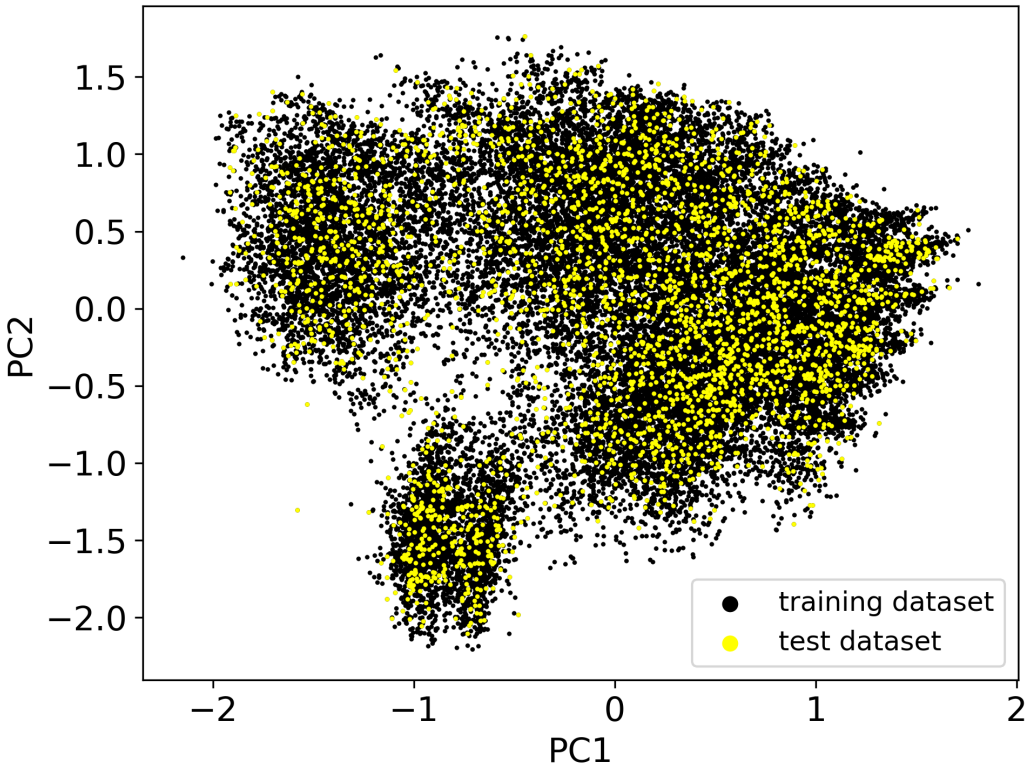
To assess whether attention improves performance for our task, we ran a controlled ablation in which we took the same backbone used for Mol2Raman and only replaced the four GINEConv layers with four GATv2Conv. Edge (bond) features were provided to both models as edge attributes. For GATv2 we used one attention head, no attention dropout, and shared weights to match capacity.

We evaluated peak-matching performance using the F1 score, computed by matching predicted and reference spectral peaks within a wavenumber tolerance, as described in the main part of this work. The GATv2 variant achieved the following resuts, compared to the results obtained using GINEConv layers:

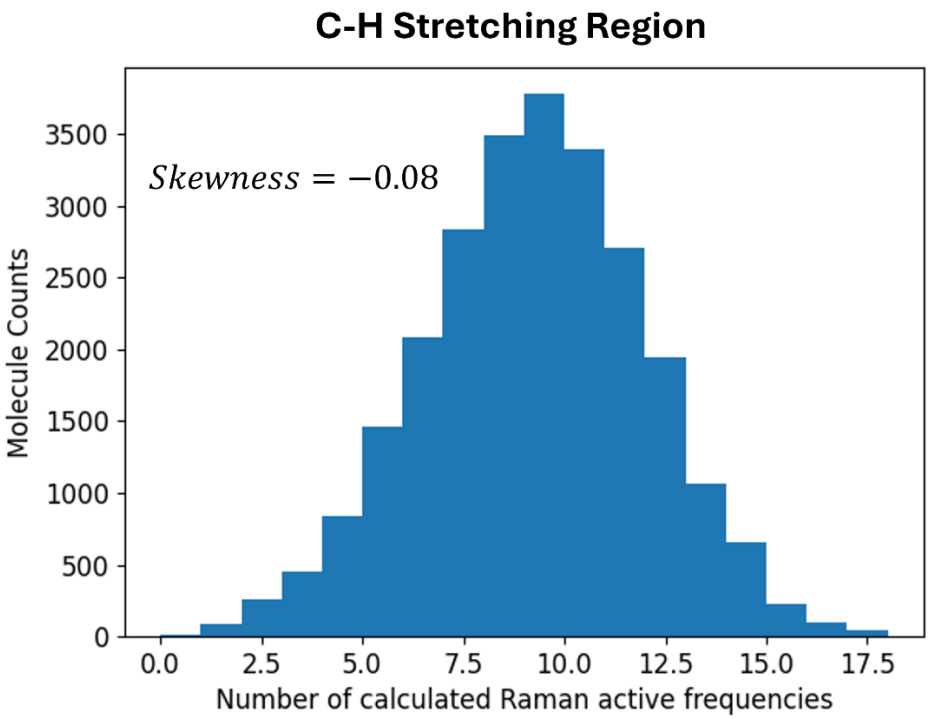
Layers type	F1 Tol. 10 cm^{-1}	F1 Tol. 15 cm^{-1}	F1 Tol. 20 cm^{-1}	Cosine Similarity
GATv2Conv	0.27	0.31	0.34	0.41
GINEConv	0.56	0.64	0.71	0.73

These values are substantially lower than those obtained with the original GINEConv backbone on the same split. We also observed markedly slower training per epoch with GATv2.

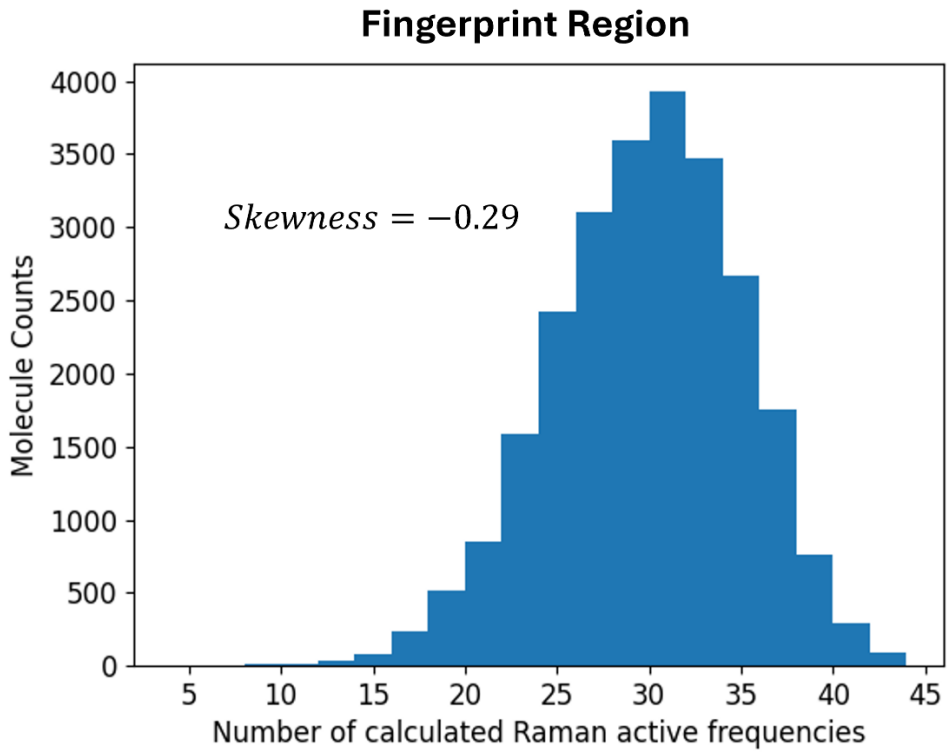
A possible explanation is that, for molecular graphs with informative bond features, edge-conditioned message passing (GINEConv) leverages bond attributes by transforming the message content directly via an edge multilayer perceptron, whereas GATv2 mainly uses edge attributes to re-weight messages through attention coefficients. In our data regime, this extra flexibility appears to increase optimization variance without yielding gains, and the local, chemically regular connectivity reduces the benefit of learning highly selective attention patterns. Given the lower peak-matching accuracy and higher computational cost, we keep GINEConv as the primary architecture and report GATv2 as an ablation.



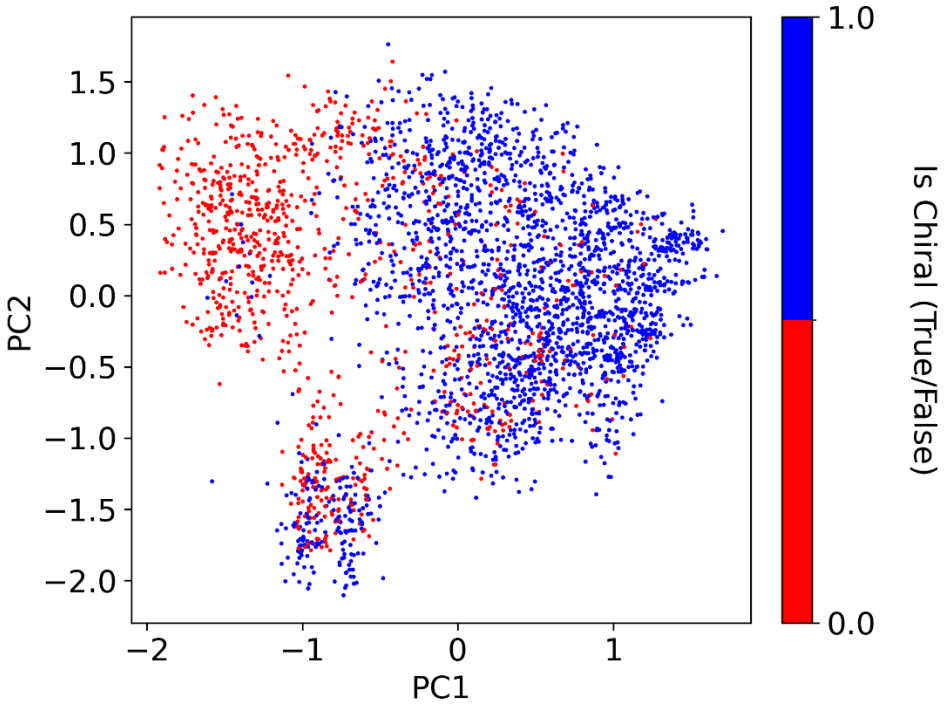
Supplementary Figure 1 (S1). Principal Component Analysis (PCA) of the molecular space based on Morgan fingerprint representations. Each point corresponds to a molecule, projected onto the first two principal components (PC1 and PC2). Points are colored according to being molecule in the training or test dataset. To verify the random sampling strategy used to construct the test dataset, we show the overlap between training and test molecules in the PC1/PC2 space. This random sampling ensures that the performance discussed for the Mol2Raman model does not depend on a specific portion of chemical space, but instead reflects its ability to generalize across a diverse and representative set of molecular structures.



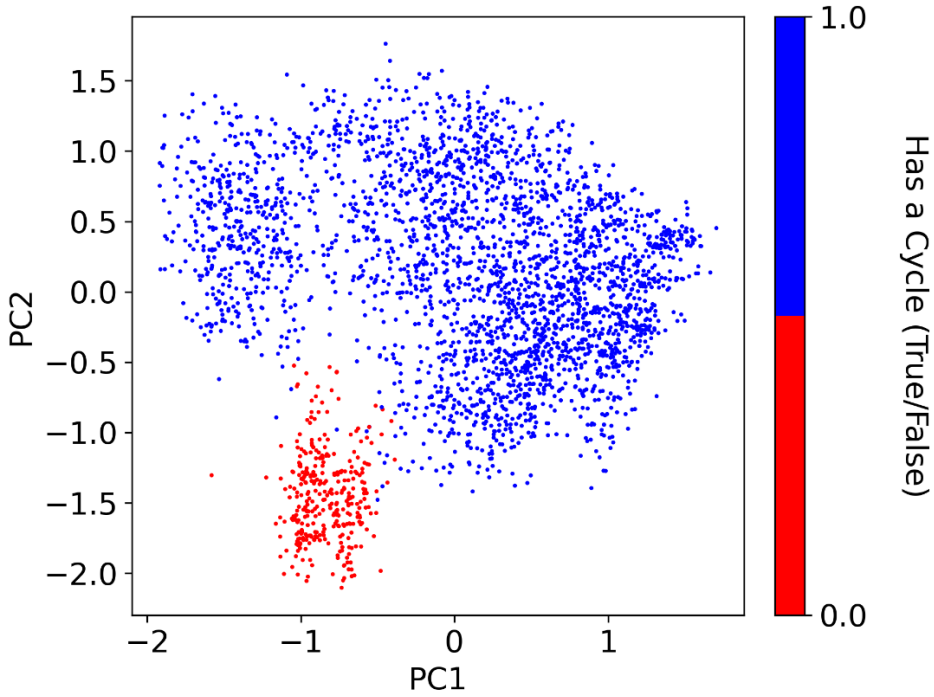
Supplementary Figure 2 (S2). Distribution of the number of calculated Raman-active frequencies in the C-H stretching region for the molecules in the training dataset. The skewness value of -0.08 indicates a nearly symmetric distribution.



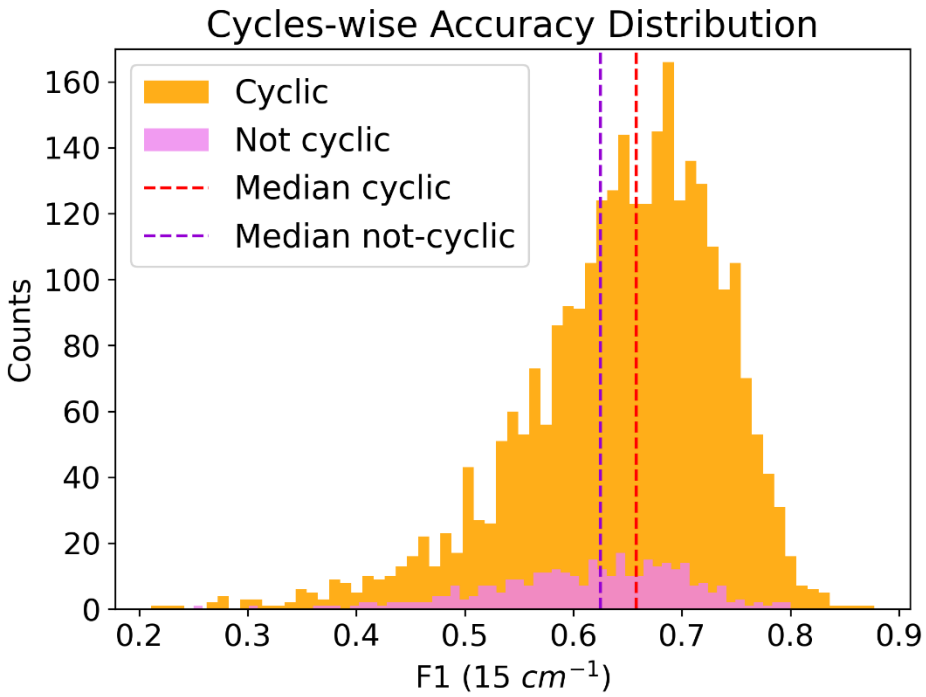
Supplementary Figure 3 (S3). Distribution of the number of calculated Raman-active frequencies in the fingerprint region for the molecules in the training dataset. The skewness value of -0.29 indicates a slight asymmetry towards higher values, suggesting that the a higher fraction of molecules have a number of Raman peaks concentrated near the upper end of the distribution.



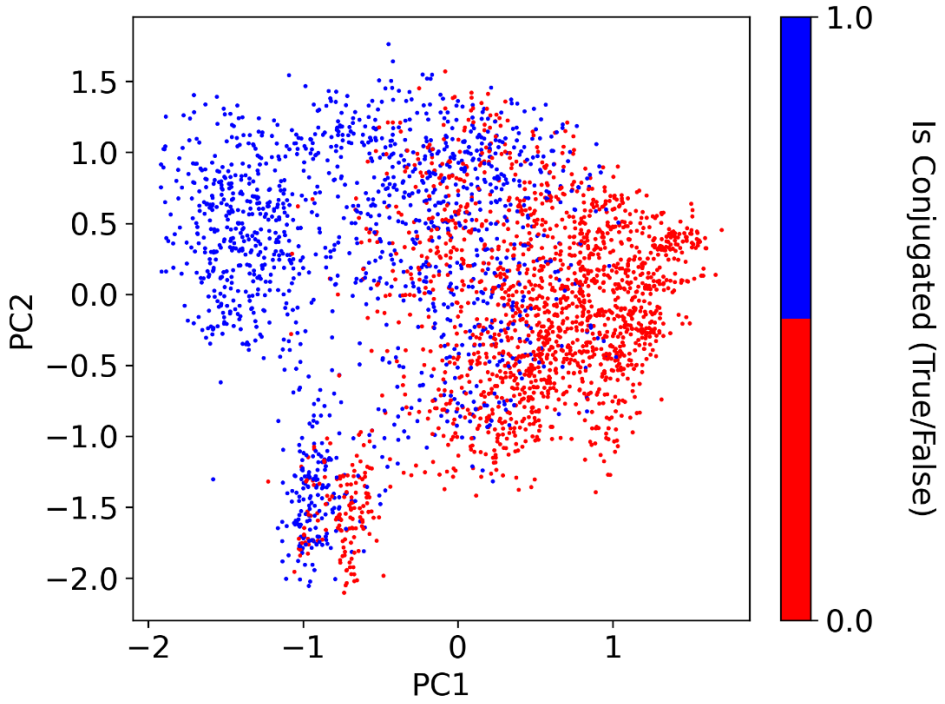
Supplementary Figure 4 (S4). Distribution of test molecules in the space of PC1 and PC2 calculated over the Morgar Fingerprint representation of each SMILES. Points are colored according to chirality: red for not-chiral molecules and blue for chiral molecules. Although chirality was not explicitly encoded in the Morgan fingerprint calculation, a partial clustering of chiral and not-chiral molecules is observed in the PCA space. This can be attributed to the fact that chirality often correlates with other structural features, such as molecular complexity, asymmetry, or presence of specific functional groups, that can be captured in the fingerprint. As a result, the unsupervised projection can still reflect underlying stereochemical diversity. However, this distribution demonstrates that the chiral and achiral molecules span comparable regions of chemical space, supporting balanced model evaluation across stereochemical diversity.



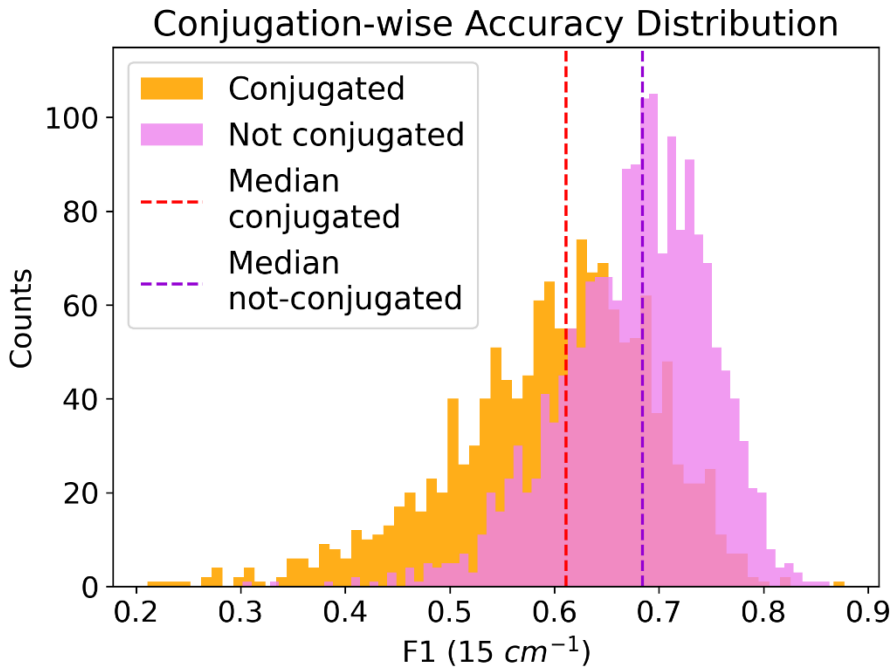
Supplementary Figure 5 (S5). Distribution of test molecules in the space of PC1 and PC2 calculated over the Morgar Fingerprint representation of each SMILES. Molecules are colored by the presence (blue) or absence (red) of a ring structure. The clear separation along PC2 indicates that cyclic and acyclic molecules activate distinct substructures in the fingerprint space, confirming that ring-related topological information is a dominant feature captured by the encoding scheme.



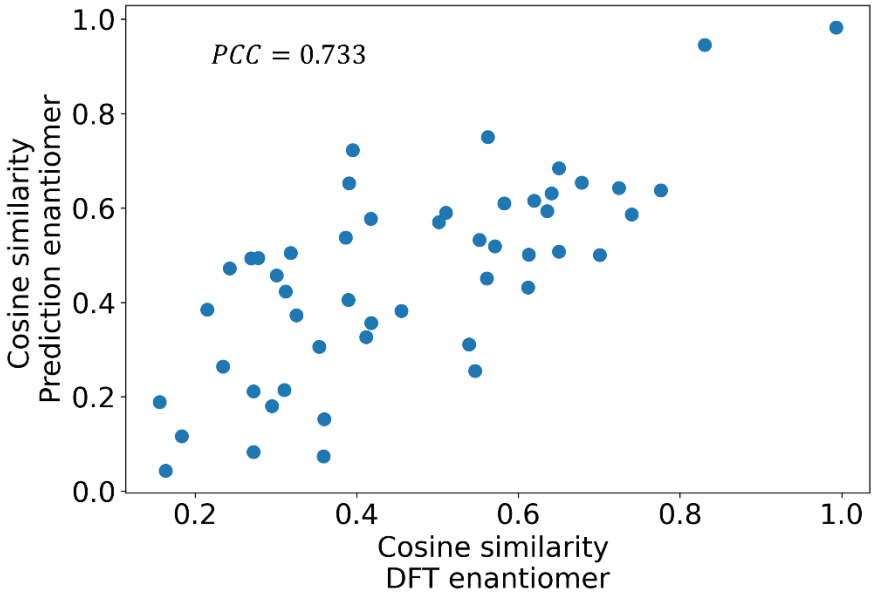
Supplementary Figure 6 (S6). Distribution of F1 scores (with a ±15 cm^{-1} tolerance) for molecules with and without cyclic structures (rings) in the test dataset. Despite the imbalance in representation (cyclic molecules significantly outnumber acyclic ones as expressed in Supplementary Table 1) the performance difference between the two groups is modest. The mean F1 score is 0.65 for cyclic and 0.61 for acyclic molecules, with a standard deviation of 0.09 for both, while the median are respectivly 0.66 and 0.62 (dashed red and purple lines, respectively). This result indicates that the presence or absence of a ring system, while structurally significant, does not substantially affect prediction accuracy, suggesting the model's robustness to this topological feature.



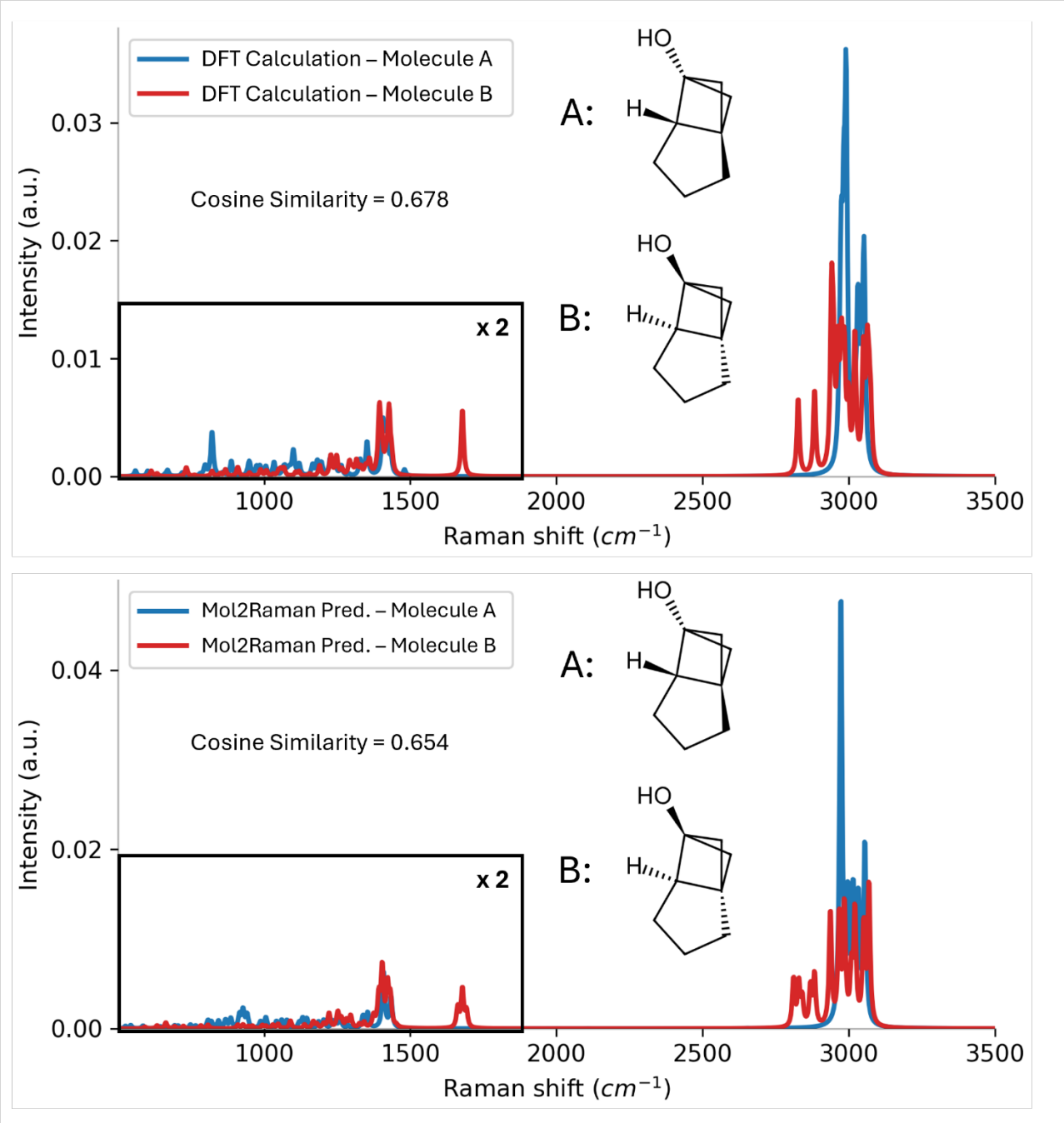
Supplementary Figure 7 (S7). Distribution of test molecules in the space of PC1 and PC2 calculated over the Morgar Fingerprint representation of each SMILES, colored according to molecular conjugation (red = conjugated, blue = non-conjugated). Despite conjugation not being explicitly encoded as a feature, molecules with conjugated systems occupy a distinct region in the fingerprint-based chemical space. This separation reflects the impact of extended π-systems and delocalized bonding patterns on the local atomic environments captured by the fingerprinting algorithm.



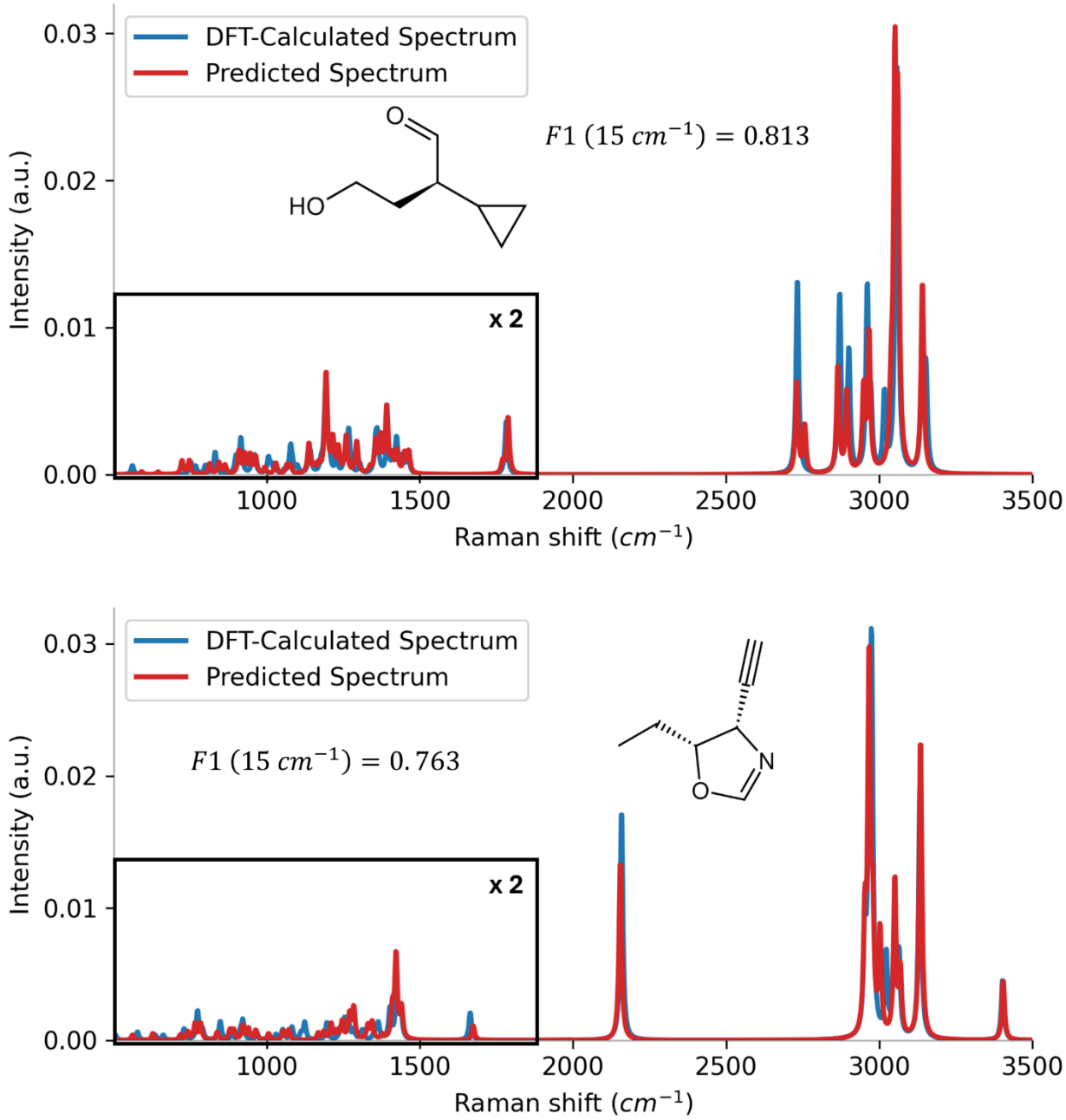
Supplementary Figure 8 (S8). Distribution of F1 scores (with a ±15 cm^{-1} tolerance) for conjugated and non-conjugated molecules in the test dataset. Although the number of molecules in both classes is balanced across the training and test sets, the model demonstrates markedly better performance on non-conjugated molecules. The median F1 score is 0.68 for non-conjugated species and 0.61 for conjugated ones, as indicated by the dashed purple and red lines, respectively. This discrepancy likely arises from the spectroscopic complexity of conjugated systems, where π-electron delocalization gives rise to Raman-active modes that are more spatially extended and less localized to specific bonds or functional groups. As a result, these vibrational modes are more challenging for Mol2Raman model, trained on local atom- and bond-centered features, to predict the Raman spectrum with the same precision of not-conjugated ones, highlighting a limitation in capturing long-range electronic effects inherent in conjugated frameworks.



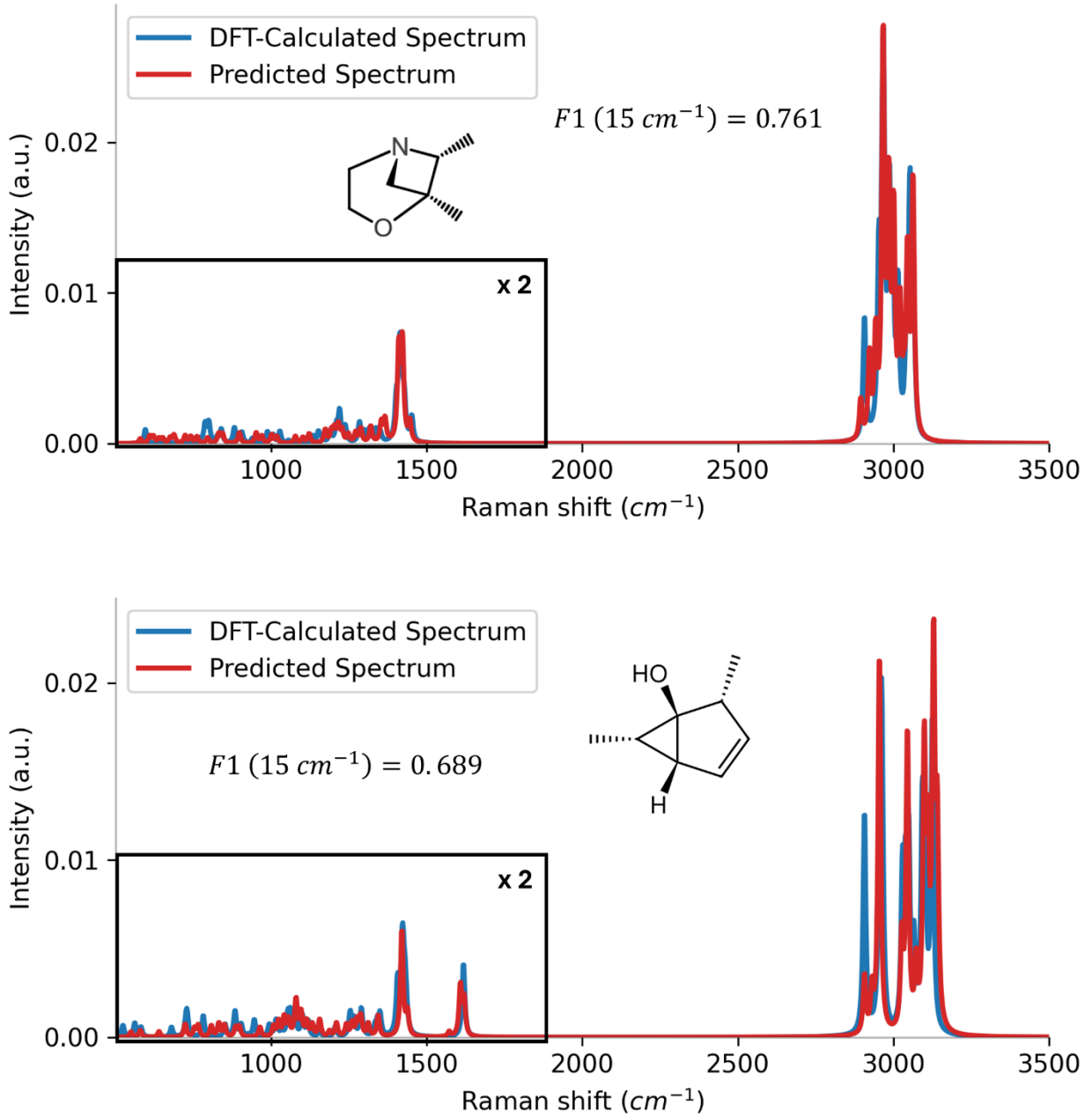
Supplementary Figure 9 (S9). Evaluation of enantiomeric similarity in Raman spectra using both DFT calculations and Mol2Raman predictions. To assess the model's ability to reproduce the spectral relationships between enantiomers, we select 50 chiral molecules from the test dataset and computed their Raman spectra via density functional theory (DFT). For each enantiomeric pair, we calculate the cosine similarity between their DFT-generated Raman spectra (x-axis) and compared it to the cosine similarity of their Mol2Raman-predicted spectra (y-axis). Ideally, if the model accurately captures enantiomeric invariance in Raman spectra, all points should align along the bisector (indicating identical similarity trends between predicted and calculated spectra). The observed high Pearson correlation coefficient (PCC = 0.733) confirms that the predicted Raman spectra preserve the enantiomeric spectral relationships found in DFT calculations. This result highlights Mol2Raman's ability to maintain subtle spectral differences inherent to chiral molecules, reinforcing its reliability for molecular vibrational analysis across stereochemical variants.



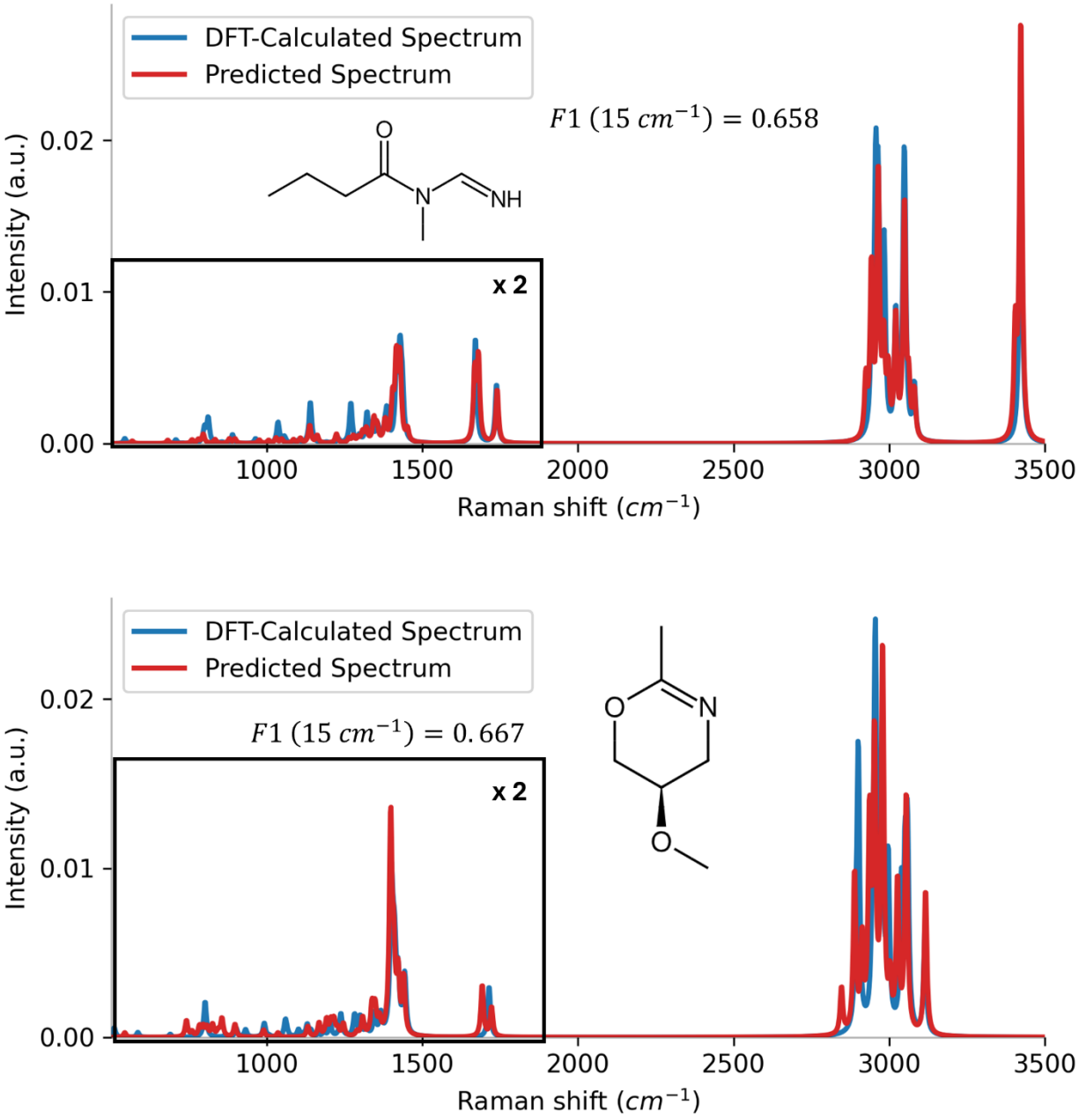
Supplementary Figure 10 (S10). Comparison of Raman spectra for a representative enantiomeric pair, computed via DFT (top panel) and predicted by Mol2Raman (bottom panel). The molecular structures of the two enantiomers (A and B) are shown in the inset. The spectral differences between enantiomers, quantified using cosine similarity (0.678 for DFT, 0.654 for Mol2Raman), are preserved between calculated and predicted spectra, demonstrating that Mol2Raman successfully captures the subtle spectral variations associated with enantiomeric inversion. The fingerprint region is scaled by a factor of two for better visualization. This result strengthens the conclusions drawn from the scatter plot analysis, confirming that the model reproduces enantiomeric spectral relationships with high fidelity.



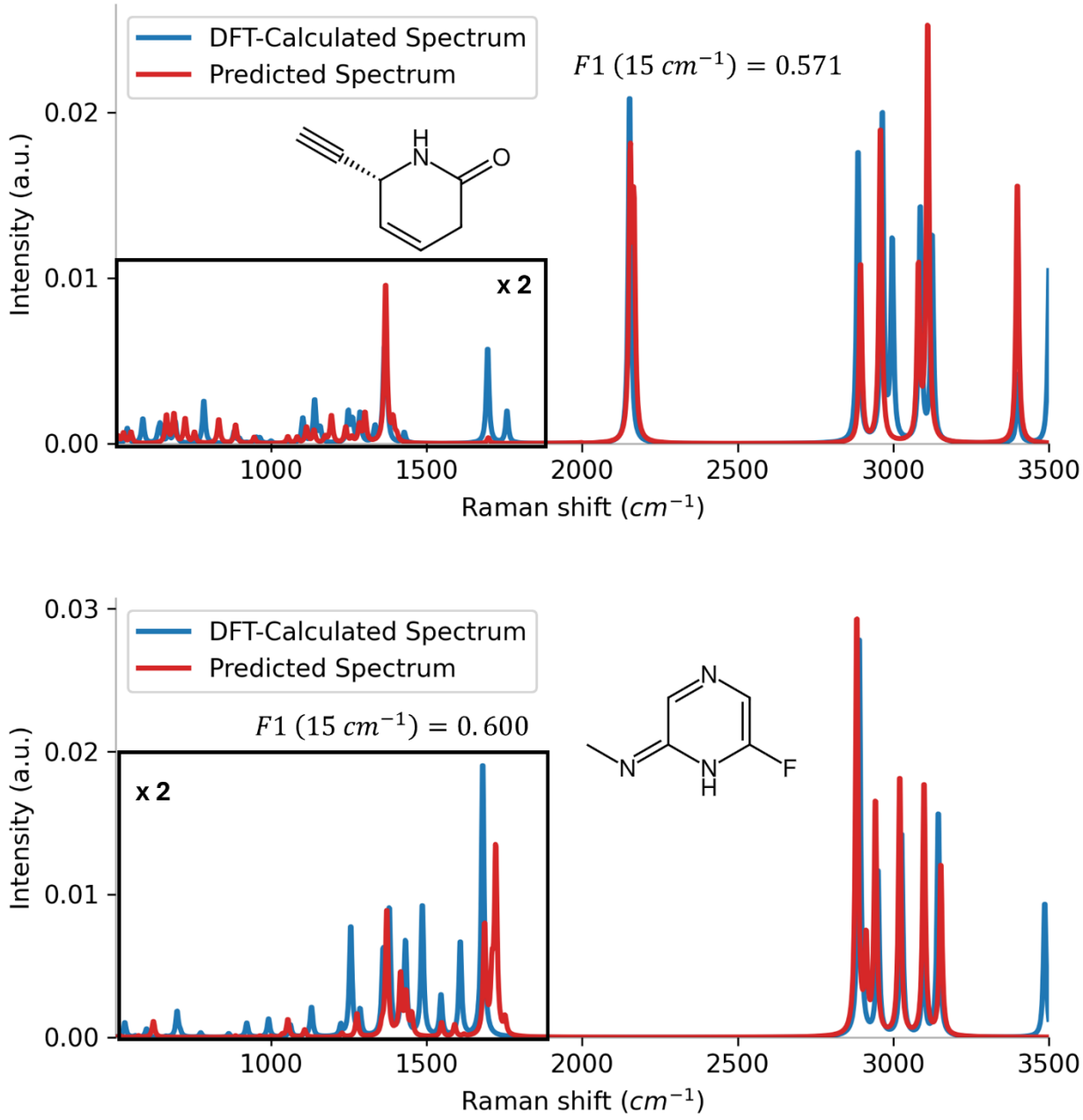
Supplementary Figure 11 (S11). Additional examples of the comparison between Mol2Raman predictions and DFT-calculated spectrum for the percentile 80.



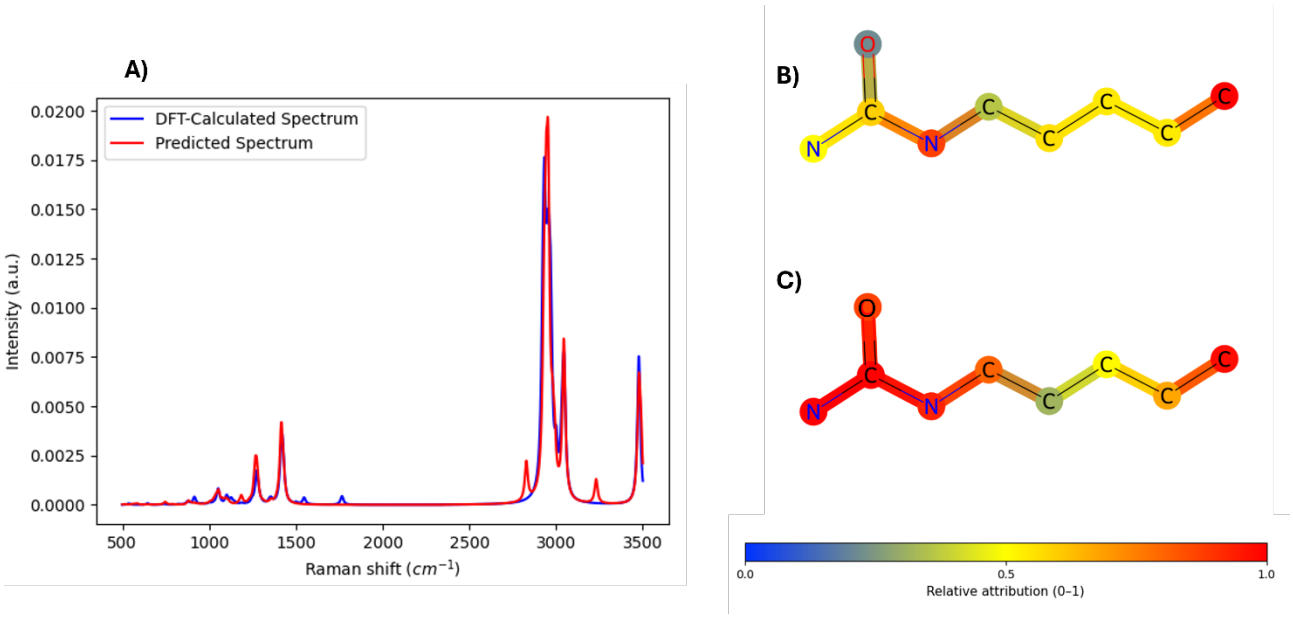
Supplementary Figure 12 (S12). Additional examples of the comparison between Mol2Raman predictions and DFT-calculated spectrum for the percentile 60.



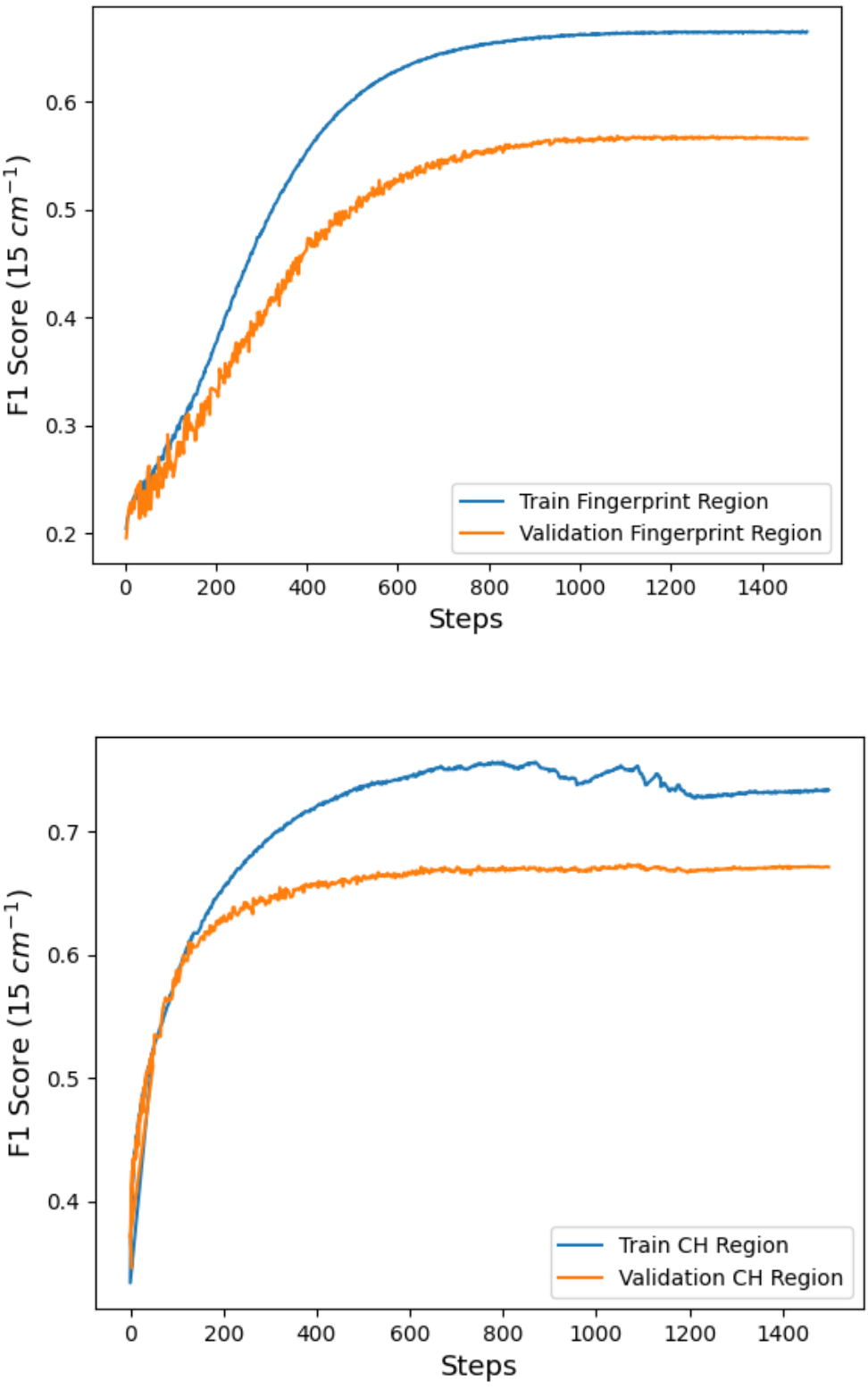
Supplementary Figure 13 (S13). Additional examples of the comparison between Mol2Raman predictions and DFT-calculated spectrum for the percentile 40.



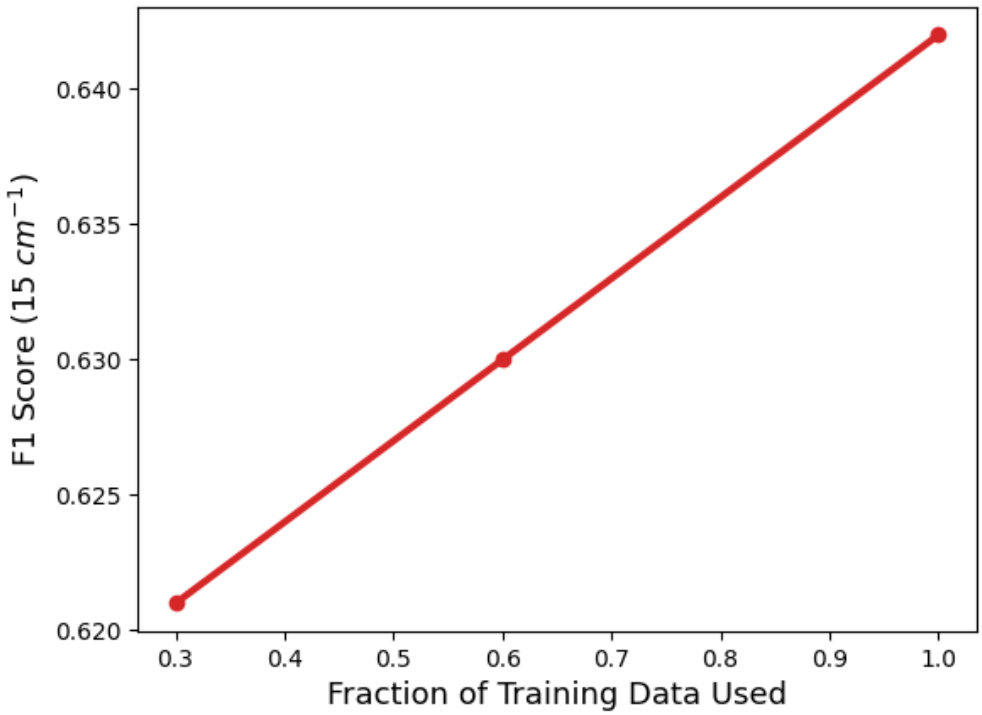
Supplementary Figure 14 (S14). Additional examples of the comparison between Mol2Raman predictions and DFT-calculated spectrum for the percentile 20.



Supplementary Figure 15 (S15). Peak-conditioned IG attribution calculation for the most intense CH and fingerprint peaks. A) shows the Mol2Raman and DFT-calculate spectra for the molecule reported in B) (CH region) and C) (fingerprint region), in which atom and bond colors denote relative attribution (0–1) in the peak-conditioned IG calculation. Similarly to Figure 5 in the manuscript, we observe a localized attribution over CH and NH bonds in B) and a widespread high attribution across the entire molecule shown in C), which are consistent with te analogous results reported in Figure S15[3, 4].



Supplementary Figure 16 (S16). Training and validation F1 with 15 cm^{-1} tolerance vs training steps for the fingerprint (upper panel) and C–H stretching region (bottom panel). The validation curves follows the training curves closely with a small, stable gap, indicating good generalization in the training process.



Supplementary Figure 17 (S17). Validation F1 with 15 cm^{-1} tolerance as a function of the fraction of the training set used (30%, 60%, Full Dataset). Performance increases at each step. The modest absolute change suggests that most of the predictive signal exploited by the model is already present in the first ~30% of the training molecules; however, adding more data provides still global improvements under the same model and training protocol.

References

1. McCloskey, K., Taly, A., Monti, F., Brenner, M. P. & Colwell, L. J. Using attribution to decode binding mechanism in neural network models for chemistry. *Proceedings of the National Academy of Sciences* **116**, 11624–11629 (2019).
2. Himanen, L. *et al.* DScribe: Library of descriptors for machine learning in materials science. *Computer Physics Communications* **247**, 106949 (2020).
3. Howell, N. K., Arteaga, G., Nakai, S. & Li-Chan, E. C. Y. Raman Spectral Analysis in the C−H Stretching Region of Proteins and Amino Acids for Investigation of Hydrophobic Interactions. *J. Agric. Food Chem.* **47**, 924–933 (1999).
4. Pezzotti, G. Raman spectroscopy in cell biology and microbiology. *Journal of Raman Spectroscopy* **52**, 2348–2443 (2021).