Simulation-to-real transfer learning for infrared spectroscopic chemical sensing and analysis from molecules to complex samples

Yusen Tan[1], Yixuan Chen[2], Zheng Fang[1], Pan Liu[1], Yifan Li[1], Qinyu Guo[3], Zhedong Lin[3], Yuqiang Li[4], Xiangxiang Zeng[†5], Tong Wang[†6], Jun Xia[†1, 7]

[1]Information Hub, The Hong Kong University of Science and Technology (Guangzhou), Guangzhou 511453, China
[2]College of Computer Science and Technology, Jilin University, Changchun 130012, China
[3]School of Computer Science, University of Auckland, Auckland 1142, New Zealand
[4]AI for Chemistry Center, Shanghai Artificial Intelligence Laboratory, Shanghai 200232, China
[5]College of Computer Science and Electronic Engineering, Hunan University, Changsha 410082, China
[6]State Key Laboratory of Chemo and Biosensing, College of Chemistry and Chemical Engineering, Hunan University, Changsha 410082, China
[7]Department of Computer Science and Engineering, The Hong Kong University of Science and Technology, Hong Kong SAR, China

Abstract

Infrared (IR) spectroscopy is widely used for chemical sensing, from molecular characterization to the analysis of complex samples, but extracting reliable chemical information from spectra remains challenging. Conventional interpretation based on expert peak assignment, rule-based analysis, and library matching is labor-intensive and difficult to scale across analytical objectives and datasets while maintaining consistently high accuracy, and its reliance on prior knowledge and reference spectra makes it less reliable for unfamiliar compounds and complex samples. Machine-learning methods have enabled more automated and scalable chemical inference from IR spectra, but most approaches remain tailored to individual tasks or datasets, require large training sets of labeled spectra, and show limited transferability across analytical objectives and experimental datasets. Together, these limitations and the scarcity of labeled experimental IR spectra motivate simulation-to-real transfer learning to enable data-efficient chemical inference under limited supervision. Here we introduce UltraIR, a foundation model for IR spectroscopy with more than 100 million parameters that enables simulation-to-real transfer learning for chemical sensing and analysis from molecules to complex samples. UltraIR learns a shared spectral representation by pretraining on approximately 60 million simulated IR spectra using three complementary pretraining objectives for spectral reconstruction, molecular fingerprint similarity alignment, and functional-group prediction. The pretrained encoder is then adapted to each downstream objective using task-specific labels or targets paired with the corresponding IR spectra, typically with only a limited number of labeled spectra available. Across benchmark evaluations of functional-group prediction, molecular structure elucidation, physicochemical property prediction, and mixture-component identification and quantification, as well as real-world chemical sensing applications involving bacterial classification, medicinal-herb geographic origin traceability and constituent quantification using two newly generated in-house datasets, microplastics classification, and soil property prediction, UltraIR outperforms conventional machine-learning and task-specific deep-learning baselines. UltraIR demonstrates the practical value of simulation-to-real transfer learning through strong performance in two challenging real-world chemical sensing scenarios: downstream adaptation using limited labeled experimental spectra and zero-shot inference for the same analytical task across Fourier-transform infrared spectrometers and laboratories. More broadly, this approach provides a route to adaptable, data-efficient chemical sensing systems for extracting reliable chemical information from IR spectra of complex real-world samples.

Correspondence: Xiangxiang Zeng (xzeng@foxmail.com); Tong Wang (wangtong@hnu.edu.cn); Jun Xia (junxia@hkust-gz.edu.cn)

1 Introduction

Infrared (IR) spectroscopy is an analytical technique that probes molecular vibrations through their interaction with infrared radiation to generate chemically informative spectral fingerprints [1–3]. Its rapid, label-free, and often non-destructive measurements support widespread use in chemical sensing and analysis, ranging from molecular characterization to complex-sample analysis in environmental monitoring, clinical analysis, and materials characterization [4–7]. However, the ability to acquire IR spectra at scale has not been matched by an equally scalable ability to convert them into reliable chemical information [8–10]. Conventional workflows still rely heavily on expert peak assignment, rule-based interpretation, and library matching. These approaches depend on prior expert knowledge, predefined rules, or reference spectra, which can limit reliable interpretation when suitable reference information is unavailable, as may occur for unfamiliar compounds and complex samples [8, 11]. Such dependencies also make conventional workflows difficult to scale across analytical objectives and datasets while maintaining consistent accuracy [1, 12, 13]. A central challenge for next-generation IR chemical sensing is therefore to translate spectra into reliable chemical information at scale across analytical objectives and datasets while maintaining high accuracy.

Machine learning has begun to address the challenge of automated IR analysis at scale by learning predictive relationships directly from spectral data [14–16]. Recent studies have shown that IR spectra support functional-group recognition, molecular structure elucidation, library-based molecular retrieval, and mixture-component identification [17–22]. Together, these advances demonstrate that IR spectra contain molecular information that can support diverse analytical tasks. However, most approaches are tailored to individual tasks or datasets and require large task-specific collections of labeled spectra, limiting the development of shared spectral representations that transfer across analytical objectives and experimental datasets. Despite enabling more rapid and scalable analysis than expert peak assignment, rule-based interpretation, and library matching, these models often achieve only modest accuracy gains over strong conventional machine-learning baselines, and these gains may disappear when few labeled spectra are available for training. These limitations motivate a general representation-learning approach that enables scalable, reliable, transferable, and data-efficient IR analysis.

Foundation models provide such a general representation-learning approach, as demonstrated by large language models (LLMs), whose large-scale pretraining supports adaptation across diverse downstream tasks [23, 24]. Similar large-scale representation-learning paradigms have emerged across sensing domains, including medical imaging [25, 26], continuous glucose monitoring [27], neural-activity modeling [28], remote sensing [29], and robotic tactile sensing [30, 31]. However, foundation-model approaches remain underexplored in IR spectroscopy, with their development constrained by the scarcity of large, chemically diverse collections of labeled experimental spectra [18, 32]. Simulated IR spectra offer a scalable, complementary source of pretraining data that can broaden molecular coverage and support shared representation learning [33–36]. To date, however, simulated IR spectra have mainly been used within individual chemical sensing and analysis tasks, with reported performance gains demonstrating their value for data-driven IR analysis [19]. Together, these observations motivate a simulation-to-real representation-learning paradigm that combines large-scale pretraining on simulated spectra with supervised adaptation using task-specific labeled experimental spectra, offering a promising route toward transferable and data-efficient IR foundation models. Specifically, pretraining can capture recurring relationships among IR spectral patterns, molecular structures and chemical properties across a broad chemical space, while downstream supervision adapts the shared representation to each chemical sensing and analysis objective.

Here we introduce UltraIR, a foundation model for IR spectroscopy with more than 100 million parameters that enables simulation-to-real transfer learning for chemical sensing and analysis from molecules to complex samples. UltraIR uses a two-stage learning framework in which a shared spectral encoder is first pretrained on approximately 60 million simulated IR spectra. For each downstream objective, the pretrained encoder is paired with a newly initialized task-specific module, and all components are jointly optimized using labeled experimental IR spectra. The encoder combines a derivative-aware multi-channel input, hierarchical convolutional modules, and a patch-based Transformer. These components capture local line shapes and peak shifts, extract multiscale vibrational signatures, and model longer-range spectral dependencies, respectively. Pretraining integrates three complementary objectives. Wavelet-domain spectral reconstruction encourages preservation of broad spectral envelopes and fine absorption features. Molecular fingerprint similarity alignment organizes the latent space according to graded structural relationships among molecules, while multi-label functional-group prediction anchors the representation to interpretable chemical motifs. Together, this design couples broad molecular exposure from simulated spectra with supervised adaptation using task-specific labeled experimental IR spectra, often available only in limited numbers, to provide a shared representation backbone for molecular interpretation, mixture analysis, and chemical sensing in complex samples.

We evaluate UltraIR across molecular and mixture benchmarks, as well as real-world chemical sensing applications. The benchmarks cover functional-group prediction, molecular structure elucidation, physicochemical property prediction, and mixture-component identification and quantification. The real-world applications include bacterial classification, medicinal-herb geographic origin traceability and constituent quantification using two newly generated in-house datasets, microplastics classification, and soil property prediction. Across these evaluations, UltraIR outperforms conventional machine-learning and task-specific deep-learning baselines. The practical value of UltraIR's simulation-to-real transfer learning is demonstrated under two demanding conditions encountered in real-world chemical sensing: downstream adaptation with limited labeled experimental spectra and zero-shot inference for the same analytical task across Fourier-transform infrared (FTIR) spectrometers and laboratories. More broadly, this learning framework provides a basis for adaptable and data-efficient chemical sensing systems that extract reliable chemical information from IR spectra of complex real-world samples.

2 Results

UltraIR overview

UltraIR is a foundation model for IR spectroscopy with more than 100 million parameters that enables simulation-to-real transfer learning for chemical sensing and analysis from molecules to complex samples. The framework integrates two complementary resources: approximately 60 million simulated IR spectra for large-scale pretraining and a diverse collection of downstream datasets for supervised adaptation and evaluation across molecular-level benchmarks, mixture analysis, and real-world chemical sensing applications (Fig. 1a).

The standardized pretraining dataset comprises molecule-associated IR spectra from three sources: approximately 1.5 million publicly released simulated spectra from IRtoMol, the multimodal spectroscopy dataset, and QM9S [19, 34, 35]; approximately 7.5 million newly generated spectra from molecular-dynamics simulations; and approximately 51 million spectra generated using a machine-learning-based spectral predictor [33]. Before pretraining, simulated spectra corresponding to molecules present in the downstream molecular datasets were excluded from the pretraining pool to prevent molecule-level overlap. Together, these sources provide the scale and molecular diversity required for large-scale pretraining.

For supervised adaptation and evaluation, we assembled approximately 120,000 experimental IR spectra together with the simulated USPTO molecular benchmark. Molecular-level benchmarks use spectra from the National Institute of Standards and Technology (NIST) Chemistry WebBook [37], the Spectral Database for Organic Compounds (SDBS) [38], and USPTO [36] for functional-group prediction, molecular structure elucidation, and physicochemical property prediction. Mixture analysis uses NIST, SDBS, and USPTO spectra for targeted component detection and targeted fractional contribution estimation, the DeepMIR liquid-mixture dataset for additional evaluation of targeted component detection [22], and FTIR mixture spectra for mixture-level component quantification [39]. Real-world chemical sensing applications include genus-level classification of bacterial isolates from green snow [40], medicinal-herb geographic origin traceability and constituent quantification, microplastics classification using environmentally sourced spectra [7], and soil property prediction using the Open Soil Spectral Library [6]. Together, these datasets span chemical sensing and analysis from molecules and mixtures to biological and botanical samples and complex environmental samples.

UltraIR follows a two-stage simulation-to-real learning framework (Fig. 1a). During pretraining, a shared encoder learns spectral representations from simulated IR spectra through three complementary objectives. Wavelet-domain spectral reconstruction encourages the encoder to retain coarse spectral envelopes and fine absorption features; molecular fingerprint similarity alignment encourages the latent space to reflect graded structural similarity among molecules; and multi-label functional-group prediction links the learned representations to interpretable chemical motifs. Together, these objectives integrate spectral morphology, molecular structural relationships, and functional-group information into a shared representation. For downstream adaptation, the pretrained encoder initializes the shared representation backbone, while a newly initialized task-specific head is added for each downstream objective. The encoder and head are then jointly optimized using the corresponding supervised data. This framework combines broad exposure to chemical diversity through pretraining on simulated spectra with task-specific supervision, enabling transfer of the shared representation across analytical objectives.

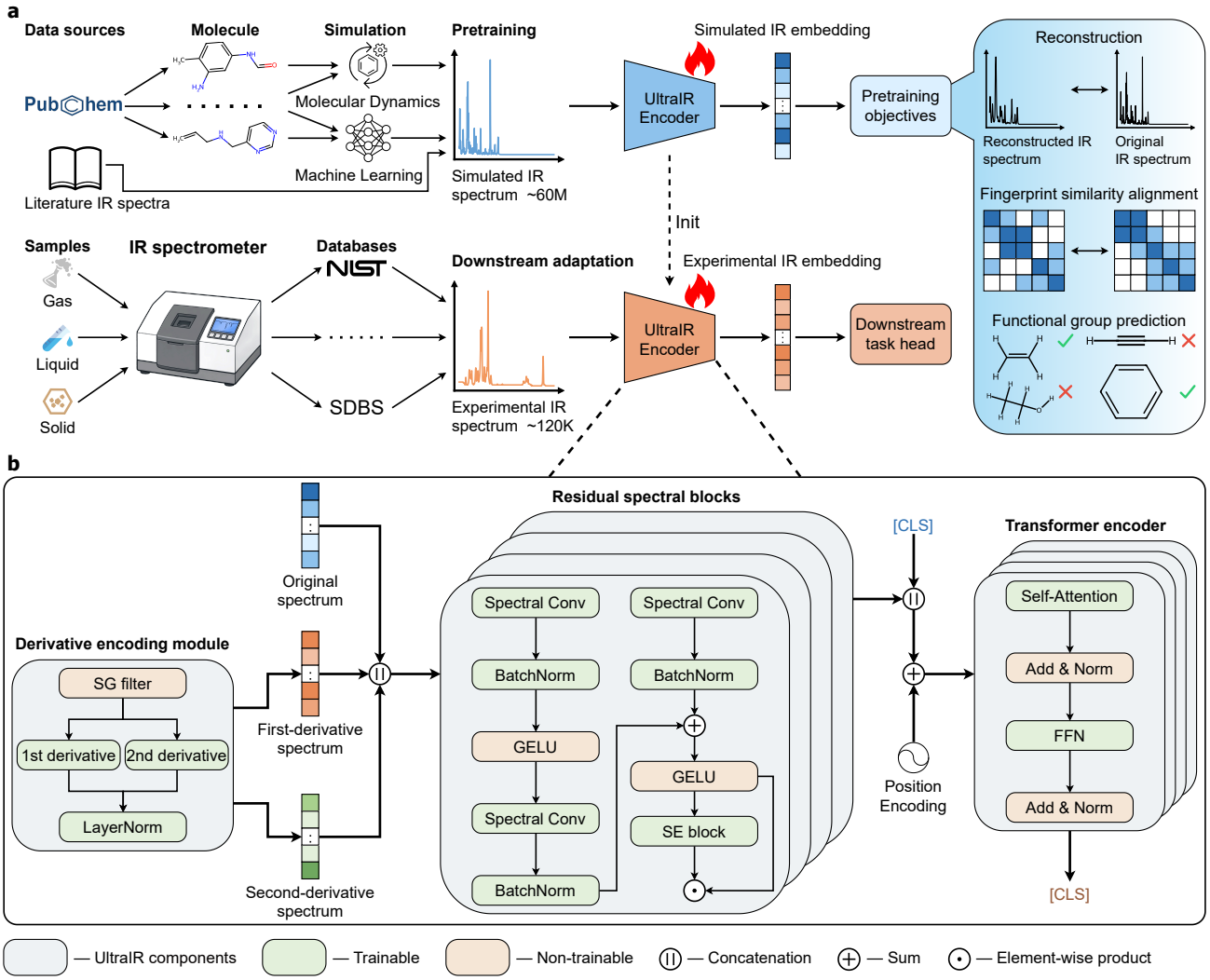
a
Data sources
Molecule
Simulation
Pretraining
Simulated IR embedding
PubChem
Molecular Dynamics
Machine Learning
Literature IR spectra
Simulated IR spectrum ~60M
UltraIR Encoder
Pretraining objectives
Reconstruction
Reconstructed IR spectrum
Original IR spectrum
Fingerprint similarity alignment
Functional group prediction
Init
Samples
IR spectrometer
Databases
Gas
Liquid
Solid
SDBS
Downstream adaptation
Experimental IR embedding
Experimental IR spectrum ~120K
UltraIR Encoder
Downstream task head
b
Residual spectral blocks
Original spectrum
Derivative encoding module
SG filter
1st derivative
2nd derivative
LayerNorm
First-derivative spectrum
Second-derivative spectrum
Spectral Conv
BatchNorm
GELU
Spectral Conv
BatchNorm
GELU
SE block
BatchNorm
[CLS]
Position Encoding
Transformer encoder
Self-Attention
Add & Norm
FFN
Add & Norm
[CLS]
— UltraIR components
— Trainable
— Non-trainable
— Concatenation
— Sum
— Element-wise product
c
Functional group prediction
Molecular structure elucidation
Physicochemical property prediction
TPSA: 53
LogP: 3.15
H acceptors: 4
QED: 0.70
H donors: 1
Rotatable bonds: 8
SA score: 2.48
BertzCT: 687
Aliphatic rings: 0
Fraction Csp3: 0.62
Aromatic rings: 2
Mixture component identification and quantification
IR spectrometer
18.5%
23.5%
IR spectrometer
18.5%
21.0%
23.5%
37.0%

Figure 1 | Overview of the UltraIR simulation-to-real framework, encoder architecture, and molecular- and mixture-level analytical tasks. a, UltraIR data resources and the two-stage simulation-to-real learning framework. The pretraining dataset contains approximately 60 million simulated IR spectra assembled from publicly released spectrum–molecule pairs and spectra predicted or simulated for PubChem-derived molecules using machine-learning models and molecular dynamics. The shared encoder is pretrained with three chemically motivated objectives: wavelet-domain spectral reconstruction, molecular fingerprint similarity alignment and multi-label functional-group prediction. For downstream adaptation, the pretrained encoder initializes the shared representation backbone and is jointly optimized with a newly initialized task-specific head using supervised downstream data, including approximately 120,000 experimental IR spectra from gas-, liquid-, and solid-phase samples. **b,** UltraIR encoder architecture. The original spectrum is concatenated with first- and second-derivative channels generated by the derivative encoding module and processed through hierarchical residual spectral blocks. The resulting features are concatenated with a `[CLS]` token, combined with positional encodings, and processed by a Transformer encoder to produce the final `[CLS]` representation. Colors distinguish UltraIR components and separate trainable from non-trainable operations; symbols denote concatenation, summation, and element-wise multiplication. **c,** Molecular- and mixture-level analytical tasks used to evaluate UltraIR, comprising functional-group prediction, molecular structure elucidation, physicochemical property prediction, and mixture-component identification and quantification. Mixture analysis includes targeted component detection, targeted fractional contribution estimation, and mixture-level component quantification.

Downstream task landscape

To evaluate whether the shared spectral representation transfers across analytical objectives and datasets, we assessed UltraIR across a downstream task landscape spanning molecular- and mixture-level benchmarks and real-world chemical sensing applications. The molecular- and mixture-level benchmarks link IR spectra to functional-group annotations, molecular structures, physicochemical properties, and mixture composition (Fig. 1c), whereas the real-world applications evaluate chemical inference from complex biological, botanical, and environmental samples.

Functional-group prediction evaluates whether UltraIR can identify functional groups from their characteristic vibrational features. Molecular structure elucidation evaluates whether UltraIR can identify the correct molecular structure from spectral information under a specified molecular-formula constraint. Physicochemical property prediction examines whether the shared spectral representation encodes continuous molecular attributes beyond discrete structural annotations. Mixture analysis assesses component detection and quantification through three tasks. In targeted component detection, the model receives a single-compound reference spectrum and a mixture spectrum and predicts whether the corresponding compound is present in the mixture. In targeted fractional contribution estimation, the same pair of spectra is used to estimate the target compound's fractional contribution. In mixture-level component quantification, the model receives only a mixture spectrum and predicts the proportions of multiple components.

Real-world chemical sensing applications span four sample domains and multiple analytical objectives. Genus-level bacterial classification evaluates whether biological spectral fingerprints can distinguish fast-growing bacterial isolates collected from green snow. Medicinal-herb characterization combines geographic origin traceability with chemical constituent quantification, requiring both categorical and continuous inference from botanical spectra. Microplastics classification uses environmental IR spectra to distinguish polymer types. Soil property prediction evaluates whether IR spectra support the prediction of continuous soil physicochemical properties.

The following sections benchmark UltraIR against established conventional machine-learning and task-specific deep-learning methods appropriate to each task.

UltraIR enables molecular structural interpretation from infrared spectra

We first evaluated UltraIR on functional-group prediction and molecular structure elucidation using the public experimental NIST and SDBS datasets and the simulated USPTO benchmark [36–38].

Functional-group prediction.

For functional-group prediction, UltraIR was benchmarked against task-specific deep-learning models, including FCG-Former and IRAnalysis [17, 18], and conventional machine-learning baselines, including XGBoost, random forest (RF), k-nearest neighbors (KNN), and a logistic classifier implemented using scikit-learn [41–43]. Performance was evaluated using three standard and complementary metrics: Micro-F1, Macro-F1, and exact match ratio (EMR). Micro-F1 and Macro-F1 are commonly used label-level metrics for multi-label classification. Micro-F1 aggregates true positives, false positives, and false negatives across all functional-group labels, providing an overall measure of label-wise prediction performance, whereas Macro-F1 first computes the F1 score for each functional group and then averages across groups, giving equal weight to frequent and infrequent functional groups. In contrast, EMR provides a molecule-level assess-

ment of complete functional-group assignment: a prediction is counted as correct only when the full set of functional groups is recovered without missing labels or false assignments. We assessed model behavior from three perspectives: overall performance on each dataset, performance variation as a function of labeled training-data size, and performance for individual functional groups.

Across the three datasets and all three metrics, UltraIR consistently and substantially outperformed both task-specific deep-learning models and conventional machine-learning baselines (Fig. 2a). On SDBS and USPTO, FCGFormer and IRAnalysis either only marginally outperformed XGBoost or performed comparably to it. On NIST, both models slightly underperformed XGBoost across all three metrics. These results demonstrate that task-specific deep learning alone does not confer a decisive advantage over strong conventional baselines for functional-group prediction. By contrast, UltraIR established a clear and consistent performance advantage over both classes of methods, with the most pronounced gains in EMR on NIST and USPTO. These gains are chemically meaningful rather than merely statistical. Functional-group prediction converts an IR spectrum into an interpretable molecular annotation that chemists can use to assess molecular composition, validate spectral interpretations, and evaluate plausible structural hypotheses. A partially correct label-wise prediction may still miss a key group or introduce a false assignment, thereby altering the chemical interpretation of the molecule even when Micro-F1 or Macro-F1 remains high. Higher EMR therefore shows that UltraIR more frequently recovers the complete functional-group set, rather than merely improving isolated label detections.

We next examined whether UltraIR's performance advantage persisted as the amount of labeled training data varied. Using EMR as the evaluation metric, UltraIR consistently outperformed all competing methods at every training proportion on NIST and USPTO (Fig. 2b). On SDBS, UltraIR performed comparably to the strongest competing method at the two lowest training proportions, 0.1% and 1%, and clearly outperformed it at all higher proportions. By contrast, the two task-specific deep-learning models, particularly FCGFormer, underperformed the strongest conventional baseline at the two lowest training proportions and were markedly inferior to UltraIR on NIST and USPTO. These results support the central simulation-to-real premise of UltraIR: large-scale simulated spectra provide transferable chemical information that enables reliable inference across a broad range of labeled-data availability, particularly when downstream supervision is scarce.

Finally, we examined whether the aggregate performance advantage of UltraIR extended across individual functional groups rather than being driven by a small number of frequent labels. Per-group F1-score rankings placed UltraIR consistently among the top-ranked methods across NIST, SDBS, and USPTO (Fig. 2c). This pattern covered 17 functional groups spanning hydrocarbons and unsaturated motifs, oxygen- and nitrogen-containing groups, halogenated groups, and carbonyl-containing derivatives. The aggregate gains therefore reflect broad performance across chemically diverse functional groups rather than gains driven by a narrow subset of common labels. This breadth supports the use of UltraIR for functional interpretation across diverse molecular chemistry.

Molecular structure elucidation.

For molecular structure elucidation, UltraIR was benchmarked against representative IR structure-elucidation models, including IRtoMol, AISE, PBSA, and DLIR [19, 20, 32, 44]. Given an IR spectrum and its molecular formula, each model generates a ranked list of molecular candidates represented as Simplified Molecular Input Line Entry System (SMILES) strings. Performance was evaluated using top-1, top-5, and top-10 accuracy, together with fingerprint-based Tanimoto similarity. The top-k metrics assess exact structure recovery by determining whether the reference molecule appears among the k highest-ranked candidates. Fingerprint-based Tanimoto similarity provides a complementary assessment of structural agreement between the top-ranked candidate and the reference molecule, with higher values indicating closer structural similarity. We examined performance from three perspectives: overall exact-recovery accuracy across datasets, structural similarity of top-ranked candidates, and the overlap and characteristics of molecules correctly solved by different methods.

Across NIST, SDBS, and USPTO, UltraIR achieved the strongest overall performance among the compared methods for molecular structure elucidation (Fig. 2d). Its top-1, top-5, and top-10 accuracies show that the learned spectral representation can prioritize the reference structure within a molecular-formula-constrained candidate set. In contrast, the other methods exhibited greater dataset-to-dataset variation in both overall performance and relative ranking.

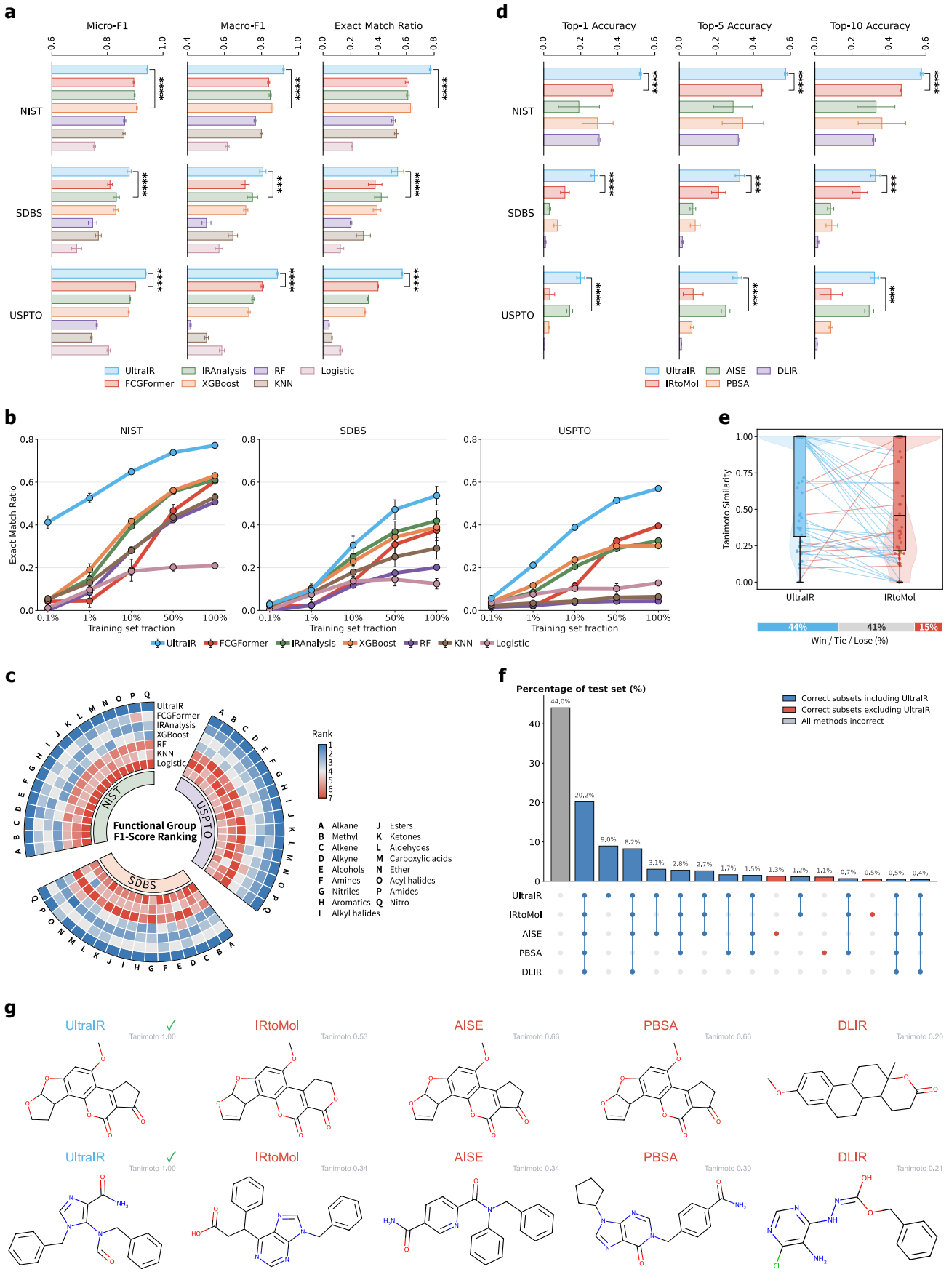
a
Micro-F1
Macro-F1
Exact Match Ratio
NIST
SDBS
USPTO
UltraIR
FCGFormer
IRAnalysis
XGBoost
RF
KNN
Logistic
d
Top-1 Accuracy
Top-5 Accuracy
Top-10 Accuracy
IRtoMol
AISE
DLIR
PBSA
b
Exact Match Ratio
Training set fraction
e
Tanimoto Similarity
Win / Tie / Lose (%)
44%
41%
15%
c
Functional Group F1-Score Ranking
Rank
A Alkane
B Methyl
C Alkene
D Alkyne
E Alcohols
F Amines
G Nitriles
H Aromatics
I Alkyl halides
J Esters
K Ketones
L Aldehydes
M Carboxylic acids
N Ether
O Acyl halides
P Amides
Q Nitro
f
Percentage of test set (%)
Correct subsets including UltraIR
Correct subsets excluding UltraIR
All methods incorrect
g
Tanimoto 1.00
Tanimoto 0.53
Tanimoto 0.66
Tanimoto 0.66
Tanimoto 0.20
Tanimoto 1.00
Tanimoto 0.34
Tanimoto 0.34
Tanimoto 0.30
Tanimoto 0.21

Figure 2 | UltraIR enables molecular structural interpretation from infrared spectra. **a,** Overall functional-group prediction performance on NIST, SDBS, and USPTO, evaluated using Micro-F1, Macro-F1, and exact match ratio (EMR). EMR counts a prediction as correct only when the complete functional-group set is recovered. **b,** Training-data scaling analysis for functional-group prediction, showing EMR as the fraction of labeled training spectra varies from 0.1% to 100%. **c,** Functional-group-wise F1-score rankings across 17 chemically diverse labels on NIST, SDBS, and USPTO, with lower ranks indicating stronger relative performance for each functional group. **d,** Molecular structure elucidation on NIST, SDBS, and USPTO, evaluated using top-1, top-5, and top-10 accuracy. A prediction is counted as correct when at least one of the top k candidates matches the ground-truth molecular structure after canonicalization. **e,** Paired comparison of fingerprint-based Tanimoto similarity between each ground-truth molecule and the corresponding top-1 candidates generated by UltraIR and IRtoMol on the NIST test set. Lines connect predictions for the same molecule. **f,** Correct-case overlap among UltraIR, IRtoMol, AISE, PBSA, and DLIR on the NIST test set. The 15 most frequent correct-case subsets are shown together with the fraction of molecules for which all methods were incorrect. Blue and red denote subsets including and excluding UltraIR, respectively, whereas gray denotes cases missed by all methods. **g,** Two representative top-1 structure-elucidation examples from the NIST test set, comparing molecular structures predicted by UltraIR and competing methods.

We further assessed the structural similarity of the top-1 candidates on NIST using fingerprint-based Tanimoto similarity. For each spectrum, we compared the Tanimoto similarity between the ground-truth molecule and the top-1 candidates generated by UltraIR and IRtoMol, the strongest competing method on NIST in the exact-recovery evaluation. UltraIR produced candidates with higher, equal, and lower similarity than IRtoMol for 44%, 41%, and 15% of spectra, respectively (Fig. 2e). Thus, UltraIR matched or exceeded IRtoMol in structural similarity for 85% of the evaluated spectra, and higher-similarity candidates outnumbered lower-similarity candidates by 29 percentage points. Together with the exact-recovery results, this comparison shows that UltraIR not only more often ranks the ground-truth molecule among its leading candidates, but also more often produces a top-ranked candidate that is structurally closer to the ground-truth molecule.

We next analyzed the overlap of correctly solved molecules across UltraIR, IRtoMol, AISE, PBSA, and DLIR on the NIST dataset, focusing on the 15 most frequent correct-case subsets (Fig. 2f). The largest subset comprised molecules that were not solved by any of the five methods, accounting for 44.0% of the test set. Among the displayed correct-case subsets, those including UltraIR accounted for 52.0% of test samples, whereas subsets excluding UltraIR accounted for only 2.9%. Notably, all five methods correctly solved 20.2% of the test set, while UltraIR alone correctly solved a further 9.0%. This pattern shows that UltraIR both recovers a substantial set of molecules shared with existing methods and contributes a sizable set of correct predictions not obtained by the competing structure-elucidation approaches.

Representative NIST test cases further illustrate the chemical basis of this advantage (Fig. 2g). In the first case, the ground-truth molecule comprised an oxygen-rich fused polycyclic scaffold containing a methoxy substituent, cyclic ethers, and two ring-associated carbonyl groups. UltraIR recovered the exact top-1 structure with a Tanimoto similarity of 1.00. AISE and PBSA retained most of the fused scaffold but introduced an incorrect double bond within an oxygen-containing ring, each yielding a similarity of 0.66. IRtoMol introduced the same bond-order error and replaced a ring carbon with an oxygen atom, resulting in a similarity of 0.53, whereas DLIR generated a substantially different and less oxygenated polycyclic framework with a similarity of 0.20. In the second case, the ground-truth molecule contained a nitrogen-rich heteroaromatic core bearing carboxamide and formamide functionalities and two benzyl substituents. UltraIR again recovered the exact structure with a similarity of 1.00. By contrast, the competing candidates altered the heteroaromatic core and the connectivity of the carbonyl-containing and aromatic substituents, with DLIR additionally introducing a chlorine atom absent from the ground truth. IRtoMol, AISE, PBSA, and DLIR achieved similarities of only 0.34, 0.34, 0.30, and 0.21, respectively.

Collectively, these results establish UltraIR as a strong approach for functional and structural interpretation of IR spectra, spanning functional-group prediction and molecular structure elucidation.

UltraIR enables physicochemical property prediction from infrared spectra

Beyond molecular structural interpretation, we evaluated whether UltraIR could predict continuous molecular physicochemical properties from IR spectra. Physicochemical property prediction was evaluated on NIST, SDBS, and USPTO for 11 structure-derived molecular descriptors and scores: synthetic accessibility (SA) score, $\log P$, topological polar surface area (TPSA), hydrogen-bond donors, hydrogen-bond acceptors, rotatable bonds, the fraction of sp^3-hybridized carbon atoms (Fraction Csp3), quantitative estimate of drug-likeness (QED), aromatic rings, aliphatic rings, and Bertz complexity (BertzCT). UltraIR was compared with conventional regression baselines, including XGBoost regression, support-vector regression (SVR), KNN regression, and partial least-squares regression (PLSR) [41, 43, 45]. Perfor-

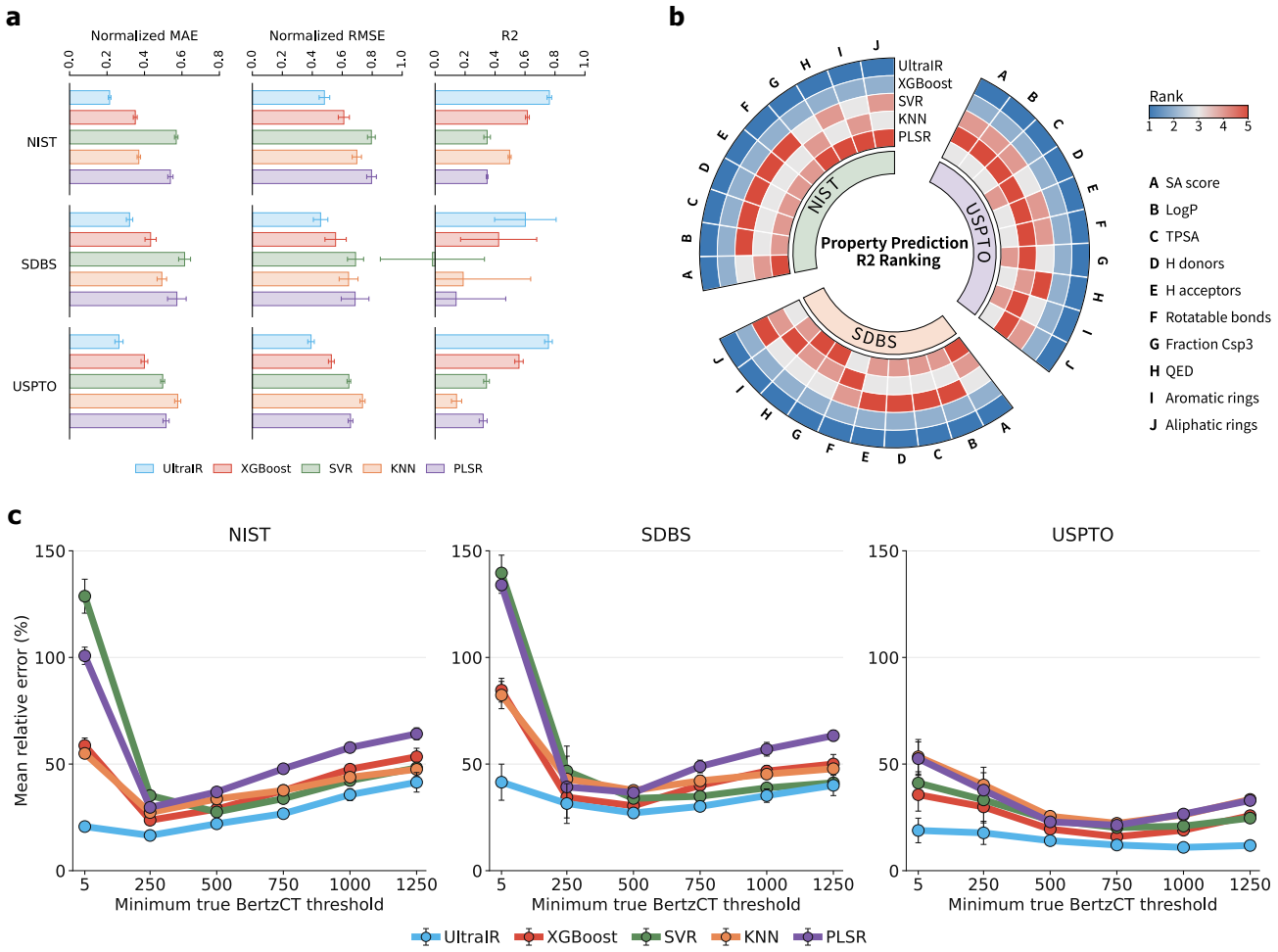

Figure 3 | UltraIR enables physicochemical property prediction from infrared spectra. a, Overall physicochemical property prediction on NIST, SDBS, and USPTO, evaluated using normalized MAE, normalized RMSE, and R^2. **b,** Property-wise R^2 rankings for 10 of the 11 physicochemical properties across NIST, SDBS, and USPTO. The remaining property, BertzCT, is examined separately in c. Lower ranks indicate stronger relative performance for each property. **c,** Mean relative error for BertzCT prediction across cumulative molecular-complexity thresholds on NIST, SDBS, and USPTO. At each threshold, performance is evaluated for molecules with true BertzCT values greater than or equal to that threshold.

mance was evaluated using normalized mean absolute error (normalized MAE), normalized root mean squared error (normalized RMSE), and the coefficient of determination R^2. Normalized MAE and normalized RMSE quantify scale-adjusted prediction error, whereas R^2 measures the proportion of variance explained by the model. We further analyzed property-wise R^2 rankings and BertzCT complexity-thresholded prediction error.

Across all three datasets, UltraIR achieved the lowest normalized MAE and normalized RMSE, together with the highest R^2 among the evaluated methods (Fig. 3a). XGBoost was the strongest conventional baseline, whereas SVR, KNN, and PLSR showed larger normalized errors or lower R^2 values in several settings. Property-wise R^2 rankings for the ten properties included in the ranking analysis placed UltraIR first in all 30 property–dataset combinations (Fig. 3b). This consistent ranking shows that UltraIR's advantage extends across the ten ranked properties rather than being driven by a single target.

We next examined BertzCT prediction across cumulative molecular-complexity thresholds. At each threshold, mean relative error was calculated for molecules whose true BertzCT value met or exceeded that threshold. UltraIR achieved the lowest mean relative error at every threshold across NIST, SDBS, and USPTO, whereas conventional regression baselines showed larger errors and more pronounced threshold-dependent variation, particularly for high-complexity subsets (Fig. 3c). These results indicate that UltraIR maintains accurate continuous-property prediction across a broad range of molecular complexity.

Together, these results extend the molecular-level interpretation capability of UltraIR beyond discrete functional-group labels and molecular structure elucidation. The learned IR representations capture continuous physicochemical attributes across diverse molecular properties and complexity regimes.

UltraIR enables mixture-component identification and quantification from infrared spectra

We evaluated targeted component detection and fractional contribution estimation across NIST, SDBS, and USPTO. Targeted component detection under increasing mixture complexity was further evaluated using the DeepMIR liquid-mixture dataset [22], whereas mixture-level component quantification was assessed using FTIR mixture datasets [39].

Targeted component detection and fractional contribution estimation.

For pairwise mixture analysis, we evaluated UltraIR on targeted component detection and targeted fractional contribution estimation across NIST, SDBS, and USPTO. In both tasks, the input consisted of a reference single-compound spectrum and a mixture spectrum. For targeted component detection, UltraIR was benchmarked against DeepMIR [22], reverse match, and the hit quality index (HQI). Performance was evaluated using accuracy, Macro-F1, and the area under the receiver operating characteristic curve (ROC-AUC). For targeted fractional contribution estimation, UltraIR was compared with a DeepMIR-derived regression baseline and conventional regression baselines, including XGBoost regression, support-vector regression (SVR), KNN regression, and partial least-squares regression (PLSR). Performance was evaluated using mean absolute error (MAE), root mean squared error (RMSE), and R^2.

UltraIR achieved strong targeted component-detection performance across all three datasets (Fig. 4a). On NIST, UltraIR attained higher accuracy and Macro-F1 than DeepMIR while achieving comparable ROC-AUC. On SDBS and USPTO, UltraIR and DeepMIR performed comparably across all three metrics. Across all datasets, UltraIR consistently outperformed the direct spectral-matching baselines: reverse match and HQI.

We next examined whether this performance persisted as mixture complexity increased. UltraIR and DeepMIR performed comparably for binary and ternary mixtures, whereas UltraIR established a clearer advantage in both accuracy and Macro-F1 for quaternary mixtures (Fig. 4b). This pattern indicates that UltraIR retains stronger target-component detection as additional components increase spectral overlap and make the target contribution more difficult to isolate.

UltraIR also achieved the strongest performance in targeted fractional contribution estimation across all three datasets, substantially and significantly outperforming the strong DeepMIR-derived regression baseline. It attained the lowest MAE and RMSE, together with the highest R^2 among all evaluated methods (Fig. 4c). Its R^2 values reached 0.956 on NIST, 0.986 on SDBS, and 0.996 on USPTO (Fig. 4d). The parity plots showed that UltraIR predictions closely followed the identity line, whereas competing methods exhibited larger deviations and weaker calibration (Fig. 4d). These results establish that UltraIR supports both qualitative identification of a target component and quantitative estimation of its contribution to a mixture spectrum.

Mixture-level component quantification.

We finally evaluated mixture-level component quantification, in which the model receives only a mixture spectrum and predicts the proportions of multiple components. This task followed a simulation-to-experiment transfer protocol: models were first trained on a simulated FTIR mixture dataset and then fine-tuned and evaluated only on experimentally acquired FTIR mixture spectra [39]. UltraIR was compared with the mixture-analysis models AIMWSP and ML-FTIR [39, 46], as well as conventional regression baselines including XGBoost regression, SVR, KNN regression, and PLSR. Performance was assessed using normalized residual distributions and component-wise R^2.

UltraIR produced residuals most tightly concentrated around zero, indicating the most accurate and stable mixture-level concentration estimates among the evaluated methods (Fig. 4e). Component-wise analysis further confirmed this advantage. UltraIR achieved the highest R^2 for each of the four quantified components: acrylonitrile, adiponitrile, propionitrile, and glycerol (Fig. 4f). Several competing methods showed pronounced component-dependent degradation. PLSR yielded negative R^2 values across all four components, and KNN also produced a negative R^2 for propionitrile, indicating performance below that of a mean-value predictor in these settings.

Together, these results show that UltraIR extends molecular-level IR analysis from individual compounds to chemical mixtures. It enables reference-guided component detection and fractional contribution estimation, as well as multi-component quantification from a single mixture spectrum. UltraIR thus supports functional-group prediction, molecular structure elucidation, physicochemical property prediction, and mixture analysis.

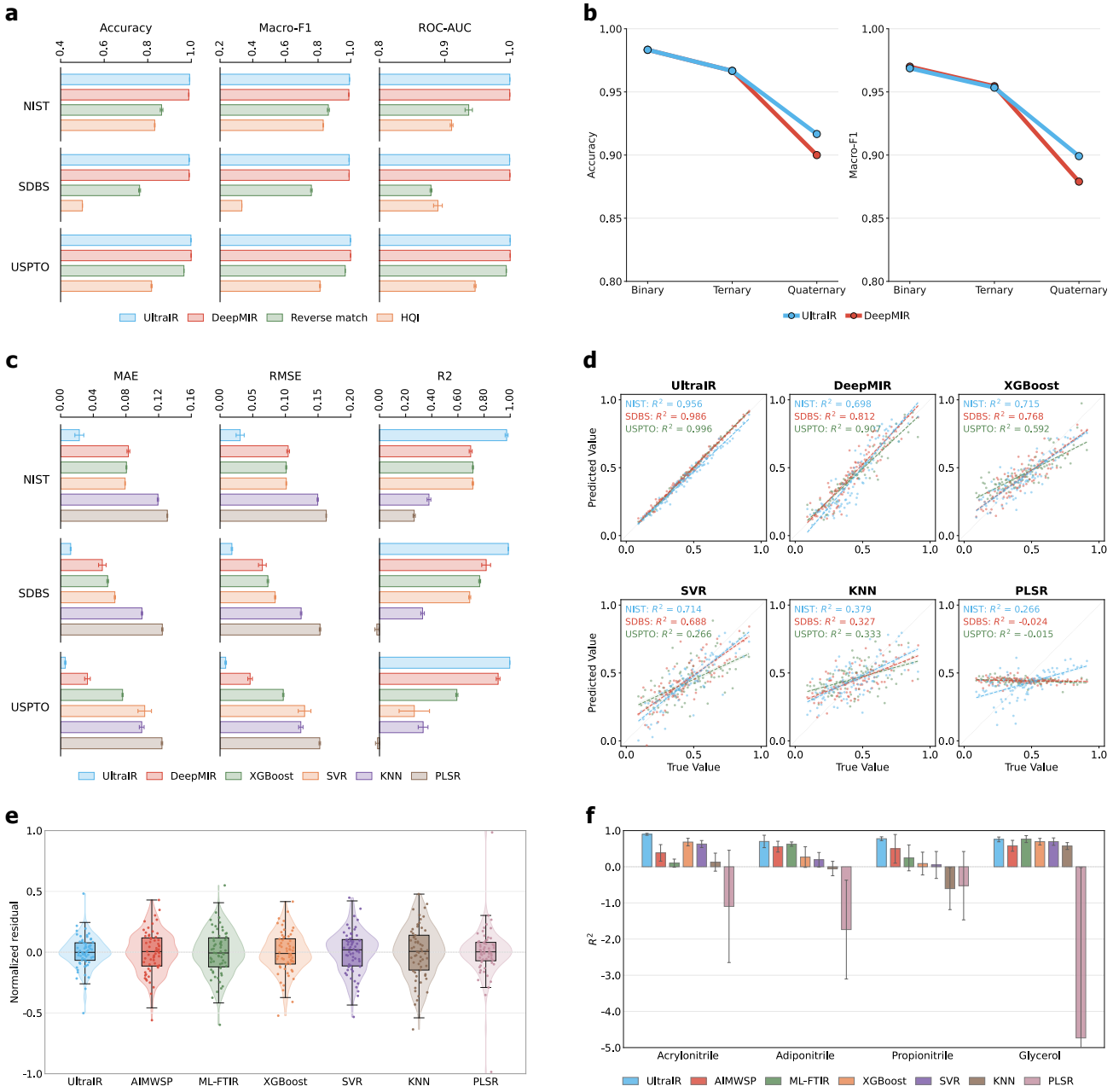

Figure 4 | UltraIR enables mixture-component identification and quantification from infrared spectra. a, Targeted component detection on NIST, SDBS, and USPTO, evaluated using accuracy, Macro-F1, and ROC-AUC. UltraIR is compared with DeepMIR, reverse match, and hit quality index (HQI). **b,** Targeted component detection in binary, ternary, and quaternary mixtures, comparing UltraIR and DeepMIR using accuracy and Macro-F1. **c,** Targeted fractional contribution estimation on NIST, SDBS, and USPTO, evaluated using mean absolute error (MAE), root mean squared error (RMSE), and R^2. **d,** Parity plots for targeted fractional contribution estimation, comparing predicted and true target-component contributions for UltraIR and competing methods across NIST, SDBS, and USPTO. The identity line denotes ideal agreement, and dashed lines show fitted relationships. **e,** Normalized residual distributions for mixture-level multi-component quantification from mixture spectra. **f,** Component-wise R^2 values for mixture-level quantification of acrylonitrile, adiponitrile, propionitrile, and glycerol.

UltraIR supports biological and botanical infrared sensing

Having established UltraIR across molecular-level tasks, we next examined whether it could transfer to biological and botanical IR sensing tasks involving experimentally acquired spectra from complex real-world samples. We considered two applications with distinct sample types and analytical objectives: genus-level bacterial classification and medicinal-

herb characterization, encompassing geographic origin traceability and chemical constituent quantification.

Bacterial classification.

For bacterial classification (Fig. 5a), UltraIR was evaluated on FTIR spectra of fast-growing bacterial isolates from green snow [40]. The task was formulated as genus-level classification and benchmarked against conventional machine-learning baselines, including XGBoost, RF, KNN, and a logistic classifier. Performance was evaluated using accuracy, Macro-F1, and the Matthews correlation coefficient (MCC), which provides a robust measure under class imbalance.

UltraIR achieved the highest accuracy, Macro-F1, and MCC among all compared methods (Fig. 5c). Genus-wise F1 analysis further showed that UltraIR consistently outperformed the conventional baselines across all nine evaluated genera, whereas baseline performance varied substantially among genera (Fig. 5d). Thus, UltraIR's aggregate advantage reflected broad genus-level discrimination rather than performance gains confined to a small subset of genera.

Medicinal herb characterization.

We next evaluated UltraIR on two newly generated in-house experimental datasets of Jinyinhua (*Lonicerae Japonicae Flos*) and Shanyinhua (*Lonicerae Flos*), comprising newly acquired IR spectra for both geographic origin traceability and chemical constituent quantification (Fig. 5b). The origin-traceability datasets comprised 120 Jinyinhua spectra from Shandong, Henan, Hebei, and Sichuan and 150 Shanyinhua spectra from Hunan, Hubei, Sichuan, Henan, and Guangdong, with 30 spectra per origin. Liquid chromatography–mass spectrometry (LC-MS)-derived relative abundances of the target constituents, expressed in arbitrary units (a.u.), served as regression labels for 60 Jinyinhua and 75 Shanyinhua samples. Details of the in-house datasets, including dataset composition, LC-MS sample preparation, and data acquisition, are provided in the Appendix. We retained a matched UltraIR model without simulated pretraining as an ablation for these tasks. For geographic origin traceability, UltraIR and the no-pretraining ablation were compared with XGBoost, RF, KNN, and a logistic classifier using accuracy, Macro-F1, and MCC. For chemical constituent quantification, they were compared with XGBoost regression, SVR, KNN regression, and PLSR using normalized residual distributions and constituent-wise R^2.

For geographic origin traceability, UltraIR achieved the highest mean accuracy, Macro-F1 and MCC on both Jinyinhua and Shanyinhua, outperforming all conventional baselines (Fig. 5e). By contrast, the no-pretraining ablation underperformed UltraIR across all three metrics on both Jinyinhua and Shanyinhua and also fell below XGBoost, the strongest conventional baseline, in every comparison. A Uniform Manifold Approximation and Projection (UMAP) of embeddings extracted by UltraIR provided a complementary qualitative view of the learned spectral organization (Fig. 5f). Spectra from the same geographic origin largely co-localized within each herb species, while Jinyinhua and Shanyinhua occupied largely distinct regions of the embedding space rather than being broadly intermixed. This organization emerged even though herb species identity was not used as a supervised label in the origin-traceability task and is consistent with UltraIR encoding spectral variation associated with both medicinal-herb identity and geographic origin.

We next moved from categorical origin assignment to quantitative estimation of chemical constituents. For both Jinyinhua and Shanyinhua, UltraIR produced normalized residual distributions that were more tightly concentrated around zero than those of the no-pretraining ablation and the conventional regression baselines, with fewer large residuals (Fig. 5g,i). The no-pretraining ablation nevertheless produced a compact interquartile residual distribution. Its interquartile range was narrower than that of KNN regression, the strongest conventional baseline by this criterion, on Jinyinhua and comparable to that of KNN regression on Shanyinhua. However, the ablation also produced more large-residual outliers.

Because residual centering and dispersion do not establish how much constituent-specific variation a model explains, we next examined constituent-wise R^2. UltraIR achieved the highest R^2 for every evaluated constituent in both Jinyinhua and Shanyinhua, with particularly large gains over the no-pretraining ablation (Fig. 5h,j). Several conventional regression baselines showed marked constituent-dependent degradation and, in some cases, negative R^2 values. The no-pretraining ablation also showed weak constituent-level generalization despite its residual distributions being centered close to zero. On Jinyinhua, its R^2 values were negative for all six constituents and exceeded only those of PLSR, the weakest method on this dataset. On Shanyinhua, it produced the lowest R^2 among all methods for three of the four constituents. The consistent contrast between UltraIR and the matched ablation provides direct evidence that simulated pretraining was critical under limited labeled supervision.

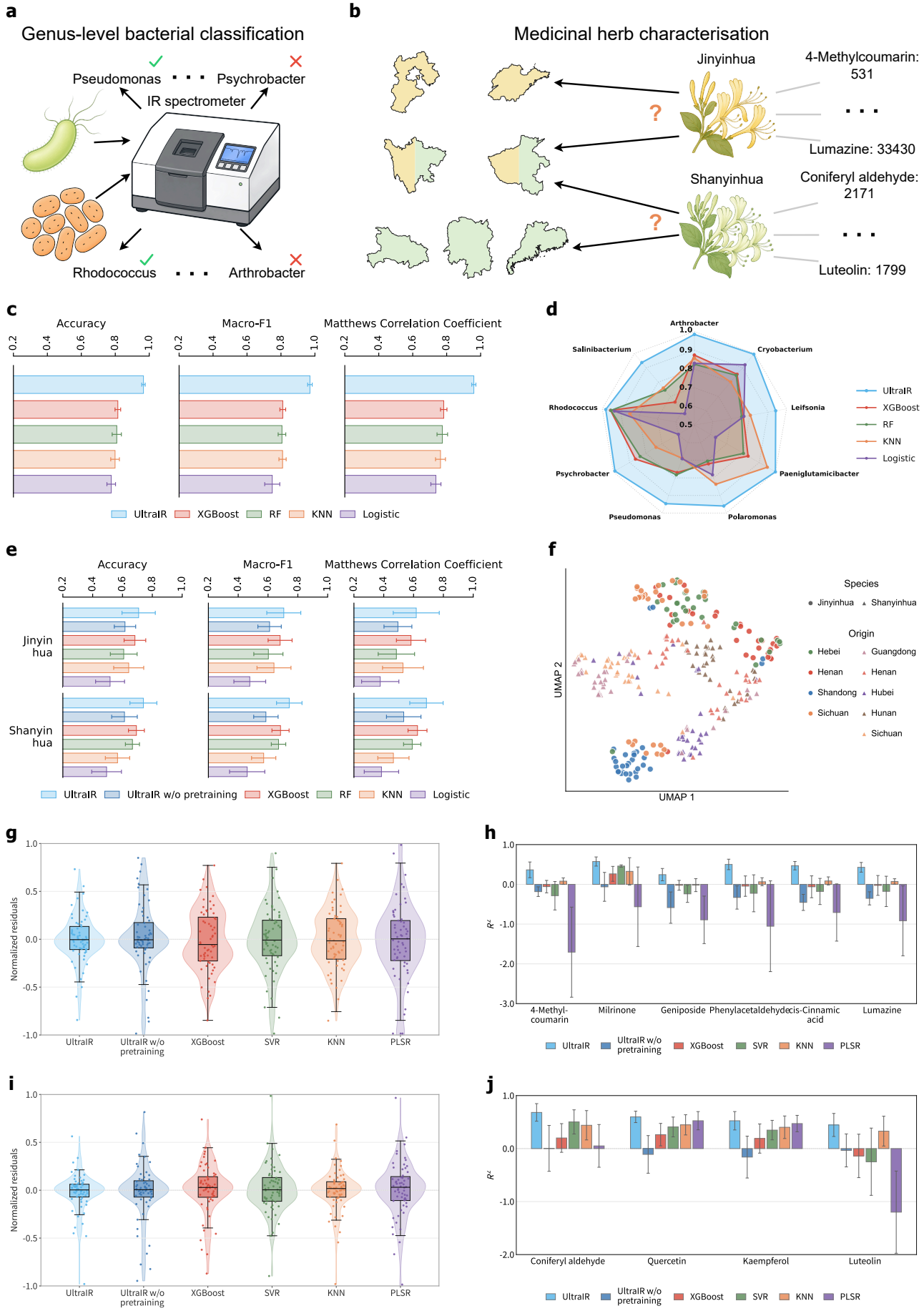
a
Genus-level bacterial classification
Pseudomonas
Psychrobacter
IR spectrometer
Rhodococcus
Arthrobacter
b
Medicinal herb characterisation
Jinyinhua
4-Methylcoumarin: 531
Lumazine: 33430
Shanyinhua
Coniferyl aldehyde: 2171
Luteolin: 1799
c
Accuracy
Macro-F1
Matthews Correlation Coefficient
UltraIR
XGBoost
RF
KNN
Logistic
d
Arthrobacter
Cryobacterium
Leifsonia
Paeniglutamicibacter
Polaromonas
Pseudomonas
Psychrobacter
Rhodococcus
Salinibacterium
e
Jinyinhua
Shanyinhua
UltraIR w/o pretraining
f
Species
Origin
UMAP 1
UMAP 2
g
Normalized residuals
SVR
PLSR
h
4-Methyl-coumarin
Milrinone
Geniposide
Phenylacetaldehyde
cis-Cinnamic acid
Lumazine
i
j
Coniferyl aldehyde
Quercetin
Kaempferol
Luteolin

Figure 5 | UltraIR supports biological and botanical infrared sensing. a, Schematic of genus-level bacterial classification from infrared spectra. **b,** Schematic of medicinal-herb characterization using in-house experimental datasets, comprising geographic origin traceability and chemical constituent quantification for Jinyinhua and Shanyinhua. **c,** Overall genus-level bacterial classification, evaluated using accuracy, Macro-F1, and the Matthews correlation coefficient (MCC). **d,** Genus-wise F1 scores across the nine evaluated bacterial genera. **e,** Geographic origin traceability of Jinyinhua and Shanyinhua, evaluated using accuracy, Macro-F1, and MCC. UltraIR is compared with its matched no-pretraining ablation and conventional classifiers. **f,** Uniform Manifold Approximation and Projection (UMAP) of medicinal-herb spectral embeddings extracted by UltraIR. Points are colored by geographic origin, and marker shapes indicate herb species. **g,** Normalized residual distributions for chemical constituent quantification in Jinyinhua, comparing UltraIR, its matched no-pretraining ablation, and conventional regression baselines. **h,** Constituent-wise R^2 values for the six quantified Jinyinhua constituents across the same methods. **i,** Normalized residual distributions for chemical constituent quantification in Shanyinhua, comparing UltraIR, its matched no-pretraining ablation, and conventional regression baselines. **j,** Constituent-wise R^2 values for the four quantified Shanyinhua constituents across the same methods.

Together with the bacterial classification results, these findings show that UltraIR supports IR analysis of complex biological and botanical samples across genus-level identification, medicinal-herb geographic origin traceability, and chemical constituent quantification.

UltraIR supports environmental infrared sensing

Having established UltraIR's performance in biological and botanical IR sensing, we finally evaluated the model for environmental IR sensing using experimentally acquired spectra from complex samples. We considered two representative environmental applications: microplastics classification and soil physicochemical property prediction.

Microplastics classification.

For microplastics classification (Fig. 6a), UltraIR was evaluated on IR spectra of environmentally sourced microplastics [7]. UltraIR was compared with task-specific deep-learning baselines, including Softmax and DB-CNN-CBAM [7, 47], and conventional machine-learning baselines, including XGBoost, RF, KNN, and a logistic classifier. Performance was evaluated using accuracy, Macro-F1, and MCC. We further analyzed the true-class margin, defined as the correct-class logit minus the highest logit among the non-true classes. Larger positive values indicate stronger separation of the correct class from its nearest competitor.

Across accuracy, Macro-F1, and MCC, UltraIR achieved the strongest performance among all evaluated methods (Fig. 6b). Its advantage over both task-specific deep-learning and conventional machine-learning baselines demonstrates effective transfer from molecular benchmarks to polymer identification in complex environmental spectra.

We next examined whether this aggregate performance advantage was accompanied by larger true-class margins. In the pairwise margin plots, many spectra lay above the diagonal, indicating that UltraIR assigned a larger true-class margin than Softmax or DB-CNN-CBAM, including numerous cases in which both methods correctly predicted the polymer class (Fig. 6c). UltraIR also corrected substantially more baseline errors than it introduced. Relative to Softmax, UltraIR corrected 472 baseline errors while introducing 112 errors. Relative to DB-CNN-CBAM, it corrected 657 baseline errors while introducing 97 errors. Thus, UltraIR not only corrected more misclassified spectra, but also achieved stronger separation between the correct polymer class and its nearest alternatives.

Finally, class-wise F1 analysis showed that UltraIR maintained consistently high performance across all 18 polymer categories, including common commodity polymers and more specialized materials (Fig. 6d). This broad class-wise advantage shows that the overall improvement was not driven by a narrow subset of readily distinguishable polymers, but extended across the microplastics classification label space.

Soil property prediction.

For soil property prediction (Fig. 6e), we evaluated whether UltraIR could estimate continuous physicochemical properties from soil IR spectra using the Open Soil Spectral Library (OSSL) [6]. UltraIR was compared with soil spectral-modeling baselines, including GADF-Swin and LSTM-CNN [48, 49], as well as conventional regression baselines, including XGBoost regression, SVR, KNN regression, and PLSR. Performance was evaluated using normalized MAE and RMSE, together with property-wise R^2, across ten soil properties: total carbon, total nitrogen, total sulfur, clay content, pH (H_2O), cation-exchange capacity (CEC), and exchangeable Ca, Mg, K, and Na.

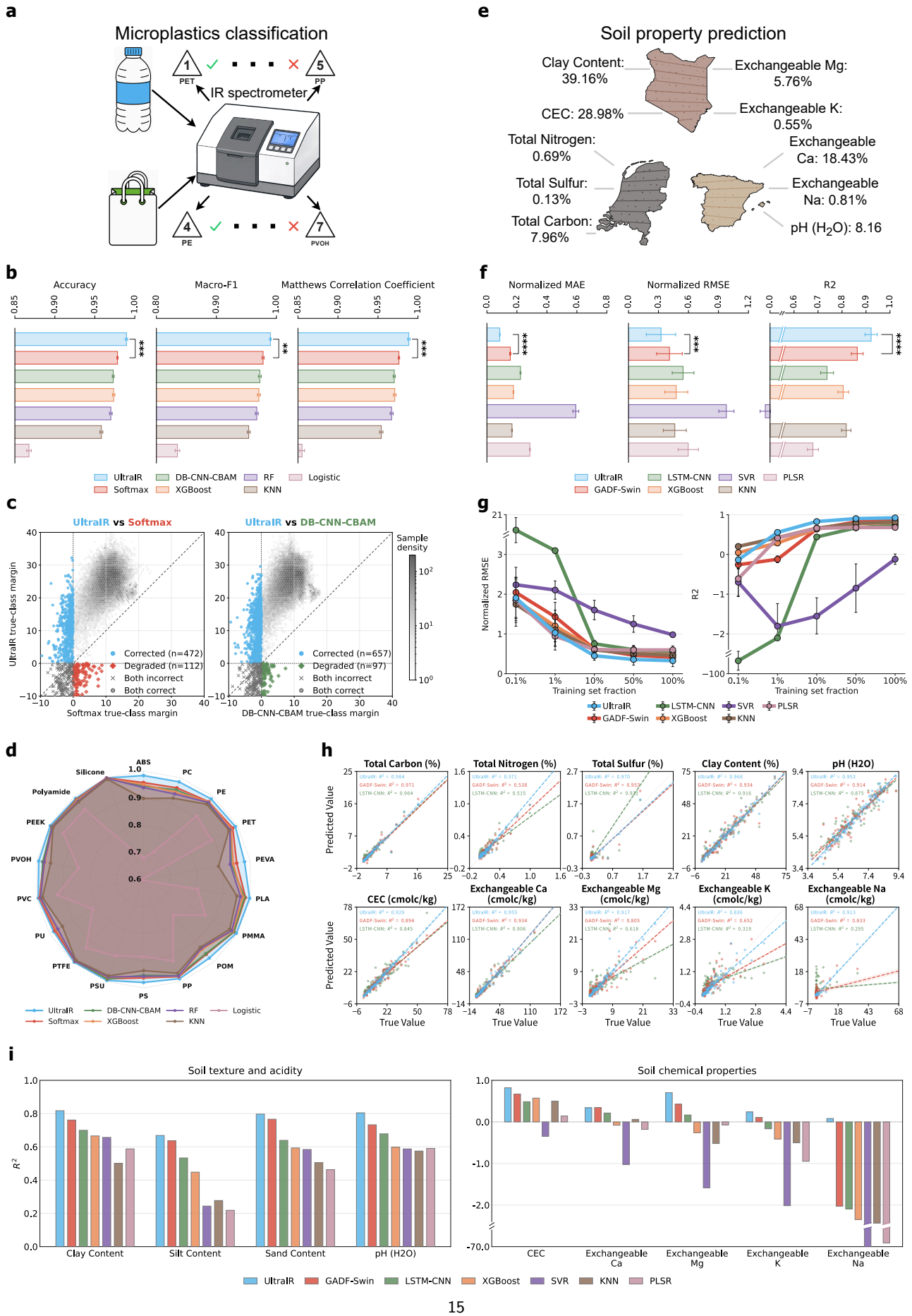
a
Microplastics classification
1 PET
5 PP
IR spectrometer
4 PE
7 PVOH
e
Soil property prediction
Clay Content: 39.16%
Exchangeable Mg: 5.76%
CEC: 28.98%
Exchangeable K: 0.55%
Total Nitrogen: 0.69%
Exchangeable Ca: 18.43%
Total Sulfur: 0.13%
Exchangeable Na: 0.81%
Total Carbon: 7.96%
pH (H2O): 8.16
b
Accuracy
Macro-F1
Matthews Correlation Coefficient
UltraIR
Softmax
DB-CNN-CBAM
XGBoost
RF
KNN
Logistic
f
Normalized MAE
Normalized RMSE
R2
UltraIR
GADF-Swin
LSTM-CNN
XGBoost
SVR
KNN
PLSR
c
UltraIR vs Softmax
UltraIR vs DB-CNN-CBAM
Sample density
UltraIR true-class margin
Softmax true-class margin
DB-CNN-CBAM true-class margin
Corrected (n=472)
Degraded (n=112)
Both incorrect
Both correct
Corrected (n=657)
Degraded (n=97)
g
Normalized RMSE
R2
Training set fraction
d
ABS
PC
PE
PET
PEVA
PLA
PMMA
POM
PP
PS
PSU
PTFE
PU
PVC
PVOH
PEEK
Polyamide
Silicone
h
Total Carbon (%)
Total Nitrogen (%)
Total Sulfur (%)
Clay Content (%)
pH (H2O)
CEC (cmolc/kg)
Exchangeable Ca (cmolc/kg)
Exchangeable Mg (cmolc/kg)
Exchangeable K (cmolc/kg)
Exchangeable Na (cmolc/kg)
Predicted Value
True Value
i
Soil texture and acidity
Soil chemical properties
Clay Content
Silt Content
Sand Content
pH (H2O)
CEC
Exchangeable Ca
Exchangeable Mg
Exchangeable K
Exchangeable Na

Figure 6 | UltraIR supports environmental infrared sensing across microplastics and soil analysis. a, Schematic of microplastics classification from infrared spectra of environmentally sourced microplastics. **b,** Overall microplastics classification, evaluated using accuracy, Macro-F1, and the Matthews correlation coefficient (MCC). **c,** Pairwise true-class margin comparisons for UltraIR versus Softmax and UltraIR versus DB-CNN-CBAM. The true-class margin is defined as the correct-class logit minus the highest logit among the non-true classes, with points above the diagonal indicating a larger margin for UltraIR. Corrected and degraded samples denote cases classified correctly by UltraIR but incorrectly by the comparator, and vice versa, respectively. **d,** Class-wise F1 scores across 18 polymer categories. **e,** Schematic of soil physicochemical property prediction from infrared spectra. **f,** Overall soil physicochemical property prediction, evaluated using normalized mean absolute error (normalized MAE), normalized root mean squared error (normalized RMSE), and R^2. **g,** Training-data scaling analysis for soil physicochemical property prediction, reporting normalized RMSE and R^2 as the labeled training fraction varies from 0.1% to 100%. **h,** Parity plots for soil property prediction, comparing predicted and measured values for UltraIR, GADF-Swin, and LSTM-CNN across ten properties: total carbon, total nitrogen, total sulfur, clay content, pH (H_2O), cation-exchange capacity (CEC), and exchangeable Ca, Mg, K, and Na. The identity line denotes ideal agreement. **i,** Cross-instrument and cross-laboratory soil property prediction. Bars show property-wise R^2 for soil texture and acidity and for soil chemical properties.

Across the three aggregate regression metrics, UltraIR achieved the strongest overall performance in soil property prediction (Fig. 6f). It attained the lowest normalized MAE and normalized RMSE and the highest R^2 among all evaluated methods.

We further assessed performance under reduced labeled-data settings. As the training-set fraction decreased from 100% to 0.1%, normalized RMSE increased and R^2 decreased across all methods (Fig. 6g). At the 1% training fraction, UltraIR achieved lower normalized RMSE than XGBoost regression and KNN regression and attained the highest R^2 among all evaluated methods, although its normalized RMSE was slightly higher than that of PLSR. At this fraction, UltraIR substantially outperformed the task-specific neural baselines GADF-Swin and LSTM-CNN, with LSTM-CNN showing particularly severe performance degradation. As the labeled training fraction increased to 10%, 50%, and 100%, both task-specific models achieved progressively lower normalized RMSE and higher R^2, but remained inferior to UltraIR. This strong performance with only 1% of the labeled training data demonstrates the data efficiency of UltraIR for soil spectroscopy, which is practically important because soil reference measurements are expensive to acquire and are often unevenly available across soil properties and sampling regions.

Property-wise analysis further confirmed the consistency of this advantage. UltraIR achieved the highest R^2 for all ten soil properties (Fig. 6h). Its predictions closely followed the identity line for total carbon, total sulfur, clay content, exchangeable calcium, and pH, for which it achieved particularly strong performance. Total nitrogen remained the most challenging property across the evaluated methods, yet UltraIR still achieved the highest R^2. These results demonstrate broad predictive capability across soil attributes with distinct chemical origins, concentration ranges, and spectral signatures.

Together with the microplastics results, these findings show that UltraIR supports environmental IR sensing across both discrete material classification and continuous quantitative estimation. Its accuracy, robustness under limited labeled data, and consistent property-wise performance support reliable IR-based analysis of complex environmental samples.

Soil property prediction as a case study of cross-instrument and cross-laboratory generalization

To examine UltraIR's capacity to generalize across FTIR instruments and laboratories, we used soil property prediction as a case study. UltraIR was compared with the same soil spectral-modeling baselines: GADF-Swin, LSTM-CNN, XGBoost regression, SVR, KNN regression, and PLSR. Relative to the preceding experiment, this comparison used different subsets of soil IR spectra from OSSL [6] and a partially different set of soil properties. Performance was evaluated using property-wise R^2 for clay content, silt content, sand content, pH (H_2O), CEC, and exchangeable Ca, Mg, K, and Na.

All models were trained using spectra from the Kellogg Soil Survey Laboratory (KSSL) subset of OSSL and evaluated using zero-shot target-domain inference on the ICRAF–ISRIC subset of OSSL. The source and target subsets differed in laboratory, FTIR instrument, and sample preparation procedures. Detailed metadata and label distributions for both subsets are provided in Appendix Tables D and E.

UltraIR achieved the highest R^2 for all soil texture and acidity properties and for most soil chemical properties (Fig. 6i). UltraIR outperformed all competing methods for CEC and exchangeable Mg, K, and Na, while performing comparably to GADF-Swin and better than the remaining methods for exchangeable Ca. UltraIR was also the only method to maintain positive R^2 across all evaluated properties. Performance differences were most pronounced for

the chemical properties, for which several baselines yielded negative R^2, particularly for exchangeable Na. This case study demonstrates that UltraIR achieved more stable and generally stronger performance during zero-shot inference across instruments and laboratories, which represents an important practical challenge in real-world chemical sensing.

3 Discussion

IR spectroscopy is widely used for chemical sensing, but extracting reliable chemical information from complex spectra at scale remains difficult across analytical objectives and datasets. UltraIR addresses this challenge by enabling simulation-to-real transfer learning through a foundation model for IR spectroscopy with more than 100 million parameters. The model learns a shared spectral representation by pretraining on approximately 60 million simulated IR spectra using complementary objectives for spectral reconstruction, molecular fingerprint similarity alignment, and functional-group prediction. The pretrained encoder is then adapted to each downstream analytical objective using task-specific supervision. This design separates large-scale spectral representation learning from downstream adaptation, allowing the same pretrained encoder to support diverse forms of IR chemical sensing and analysis.

A central advance of UltraIR lies in the breadth of chemical inference supported by its shared pretrained encoder. The evaluated tasks span molecular structural interpretation, physicochemical property prediction, mixture-component identification and quantification, and real-world chemical sensing in biological and botanical samples and complex environmental samples, with two newly generated in-house experimental datasets used for botanical sensing. These tasks differ in their inputs, output spaces, and analytical objectives, encompassing multi-label classification, molecular generation, regression, reference-guided mixture analysis, and multi-component quantification. Across these evaluations, UltraIR outperformed conventional machine-learning and task-specific deep-learning baselines. This consistent performance across tasks and datasets indicates that the learned representation can transfer after supervised adaptation rather than remaining tied to a single label space or dataset. UltraIR therefore provides a practical basis for reusing a shared pretrained encoder with a task-specific output module for each analytical objective.

The practical value of UltraIR's simulation-to-real transfer learning was particularly evident in two challenging real-world chemical sensing scenarios: downstream adaptation with limited labeled experimental spectra and zero-shot target-domain inference for the same analytical task across FTIR instruments and laboratories. In training-data scaling experiments, UltraIR retained stronger performance than task-specific neural baselines as labeled supervision decreased. The medicinal-herb ablation further showed that simulated pretraining benefited complex sample analysis under limited labeled supervision. A complementary soil property prediction case study showed that UltraIR maintained more stable and generally stronger performance than competing methods during zero-shot target-domain inference across instruments and laboratories. Simulated pretraining provides broad chemical exposure, whereas task-specific labels align the shared representation with individual analytical objectives, together supporting data-efficient learning and generalization across measurement conditions.

Although UltraIR demonstrated strong performance across diverse infrared chemical analysis tasks, several limitations remain and point to directions for further improvement. UltraIR still requires supervised adaptation and a task-specific output module for each analytical objective. The present study therefore demonstrates broad transferability across analytical objectives after adaptation, rather than zero-shot or universal inference across tasks. A simulation-to-real domain gap also remains because simulated IR spectra cannot fully reproduce experimental variation. This variation can arise from instrument response, spectral resolution, measurement geometry, sample state, preparation procedures, temperature, concentration, scattering, complex matrices, and baseline artifacts. Improving simulation fidelity will require explicit modeling of these factors. Future adaptation strategies could exploit unlabeled experimental IR spectra to reduce the need for labeled experimental spectra.

Overall, UltraIR establishes a scalable framework for simulation-to-real transfer learning in foundation modeling for IR spectroscopy. By coupling pretraining on chemically diverse simulated spectra with supervised downstream adaptation, UltraIR supports chemical sensing and analysis across analytical objectives and datasets, spanning molecules and mixtures, biological and botanical samples, and complex environmental samples. This framework provides a route beyond collections of independently trained task-specific models toward adaptable and data-efficient IR-based chemical sensing systems built on reusable spectral representations.

4 Methods

Overview of UltraIR

UltraIR is a foundation model for IR spectroscopy with more than 100 million parameters that enables simulation-to-real transfer learning for chemical sensing and analysis from molecules to complex samples. The framework comprises a shared spectral encoder and task-specific output modules. During pretraining, the encoder is optimized on approximately 60 million simulated IR spectra to learn a shared, chemically informative spectral representation. For downstream adaptation, the pretrained encoder is paired with a newly initialized task-specific output module, and both components are jointly optimized using labeled spectra for the corresponding task. The encoder architecture and pretrained parameters provide a common initialization across downstream objectives, whereas the output module, output dimensionality, and supervised training objective are defined separately for each task.

In UltraIR, the shared encoder is initialized from the pretrained checkpoint. In designated ablation analyses, we additionally evaluated a matched no-pretraining control in which the same encoder architecture and task-specific output modules were initialized randomly and trained using the same downstream data and optimization procedure. Formally, given a preprocessed IR spectrum $\mathbf{x}$, the shared encoder produces a latent representation

$$\mathbf{z} = f_\theta(\mathbf{x}), \tag{1}$$

where f_θ denotes the shared encoder with parameters θ. For UltraIR, θ is initialized from the pretrained parameters θ_{pre}. For the no-pretraining control, θ is instead initialized randomly. A task-specific output module $h_\phi^{(t)}$ then transforms the shared representation, together with any auxiliary input required by task t, into the corresponding analytical output:

$$\hat{\mathbf{y}}^{(t)} = h_\phi^{(t)}\left(\mathbf{z}, \mathbf{u}^{(t)}\right), \tag{2}$$

where $\mathbf{u}^{(t)}$ denotes optional task-specific inputs, such as a molecular formula or a reference single-compound spectrum. The downstream outputs include functional-group annotations, formula-conditioned molecular structure candidates, molecular physicochemical property estimates, mixture-component detection and quantification outputs, bacterial genus labels, medicinal-herb geographic origin labels and constituent relative abundances, polymer-type labels for microplastics, and soil physicochemical property estimates.

Spectral preprocessing and data augmentation

Before being supplied to the model, all IR spectra were converted to a common spectral range, sampling grid, and intensity scale. The same deterministic preprocessing pipeline was used for large-scale pretraining on simulated spectra and downstream adaptation on experimental spectra. Stochastic augmentation was stage-specific. A broader set of augmentation operations was used during pretraining, whereas only weak, validation-selected augmentations were considered during downstream adaptation.

Standard spectral preprocessing pipeline.

Each spectrum was first cropped to the mid-IR range of $400\,\mathrm{cm}^{-1}$ to $4000\,\mathrm{cm}^{-1}$. Spectra that did not cover the complete range were zero-padded at the missing boundaries. The resulting spectrum was resampled by piecewise-linear interpolation onto a canonical 3,600-point grid and then min–max normalized to the interval $[0, 1]$. This canonical representation was subsequently mapped to the fixed input length required by each model implementation.

Pretraining augmentation.

During pretraining, four stochastic augmentation operations were applied to each spectrum. Additive Gaussian noise was sampled at a random signal-to-noise ratio to represent electronic measurement noise. A random shift along the wavenumber axis represented calibration uncertainty. An additive intensity offset represented baseline drift. Finally, a random subset of spectral positions was masked by setting the corresponding intensities to zero. This masking objective encouraged the encoder to use information from the broader spectral context rather than relying exclusively on isolated local features.

Downstream adaptation augmentation.

For downstream adaptation, stochastic augmentation was not imposed as a universal part of the training pipeline. Instead, weak augmentations, including additive noise, small wavenumber shifts, and baseline offsets, were enabled only when selected using the validation split for the corresponding model–task setting. Random spectral masking was not used during downstream adaptation because most downstream objectives require access to the complete experimental spectral profile.

Hybrid convolutional–Transformer architecture

As illustrated in Fig. 1b, UltraIR uses a shared hybrid convolutional–Transformer encoder to map an IR spectrum to a latent spectral representation $\mathbf{z}$ for downstream chemical analysis. The encoder comprises three principal modules: a derivative-aware multi-channel input module, a hierarchical convolutional backbone with multi-scale feature fusion, and a patch-based Transformer encoder. Together, these modules combine derivative-aware spectral feature extraction, multi-scale local feature aggregation, and long-range spectral-context modeling.

Derivative-aware multi-channel input module.

Given the preprocessed IR spectrum $\mathbf{x} \in \mathbb{R}^{1\times L}$, where L denotes the number of spectral data points, this module augments the original absorbance signal with derivative-based spectral representations to enhance sensitivity to spectral line shape features. Specifically, $\mathbf{x}$ is first smoothed via a Savitzky–Golay (SG) filter to yield $\tilde{\mathbf{x}}$. Two learnable derivative convolutions then compute the first- and second-order spectral derivatives $\mathbf{d}_1$ and $\mathbf{d}_2$, where the corresponding kernels $\mathbf{w}_1$ and $\mathbf{w}_2$ are initialized as first- and second-order central-difference operators, respectively, and remain trainable throughout optimization. Each derivative signal is subsequently normalized using layer normalization, and the three signals are concatenated to form the three-channel representation $\mathbf{x}_{\text{cat}} = [\mathbf{x}, \mathbf{d}_1, \mathbf{d}_2] \in \mathbb{R}^{3\times L}$. To adaptively reweight each spectral channel, $\mathbf{x}_{\text{cat}}$ is passed through a gated input fusion module that computes a position-wise, input-dependent channel gate:

$$\mathbf{g} = \sigma(\mathbf{W}_2^{\text{g}}\,\text{GELU}(\mathbf{W}_1^{\text{g}}\,\mathbf{x}_{\text{cat}})) \in \mathbb{R}^{3\times L}, \tag{3}$$

where $\mathbf{W}_1^{\text{g}}$ and $\mathbf{W}_2^{\text{g}}$ are 1×1 convolutional projections, GELU denotes the Gaussian error linear unit activation, and $\sigma(\cdot)$ denotes the sigmoid function. The fused multi-channel representation is then obtained via element-wise modulation:

$$\mathbf{x}_{\text{multi}} = \mathbf{x}_{\text{cat}} \odot \mathbf{g} \in \mathbb{R}^{3\times L}, \tag{4}$$

where $\odot$ denotes element-wise multiplication.

Convolutional spectral encoder.

The fused representation $\mathbf{x}_{\text{multi}} \in \mathbb{R}^{3\times L}$ is fed into a convolutional encoder that hierarchically extracts local spectral features across multiple spectral scales. Prior to feature extraction, $\mathbf{x}_{\text{multi}}$ is rescaled by a learnable per-channel scale vector $\boldsymbol{\gamma} \in \mathbb{R}^3$, broadcast along the spectral dimension and initialized as $[1.0, 0.5, 0.5]^\top$ to encode the inductive bias that the original absorbance signal should initially dominate over the two derivative channels, while allowing the network to adjust this balance during training, yielding $\hat{\mathbf{x}}_{\text{multi}} \in \mathbb{R}^{3\times L}$. The encoder then begins with a convolutional stem that applies a wide-kernel convolution followed by batch normalization and a GELU activation to $\hat{\mathbf{x}}_{\text{multi}}$, mapping it to an initial feature map $\mathbf{H}_0 \in \mathbb{R}^{C_0\times L}$, where C_0 denotes the base channel width.

Four successive residual blocks then progressively encode the representation. The first three blocks each apply a strided convolution with stride 2, yielding intermediate feature maps $\mathbf{F}_1 \in \mathbb{R}^{C_1\times L/2}$, $\mathbf{F}_2 \in \mathbb{R}^{C\times L/4}$, and $\mathbf{F}_3 \in \mathbb{R}^{C\times L'}$, where $C_1 = 2C_0$, $C = 4C_0$, and $L' = L/8$. The fourth block operates at the same spectral resolution, producing $\mathbf{F}_4 \in \mathbb{R}^{C\times L'}$. Each residual block comprises two convolutional layers, each followed by batch normalization, with a GELU activation applied after the first convolution–batch normalization (Conv-BN) layer and again after the residual addition, together with a squeeze-and-excitation (SE) module for adaptive channel-wise recalibration. Formally, for block $i \in \{1, 2, 3, 4\}$ with input $\mathbf{F}_{i-1}$ (where $\mathbf{F}_0 = \mathbf{H}_0$), the residual branch is defined as:

$$f_i(\mathbf{F}) = \text{BN}\Big(\mathbf{W}_i^{(2)} * \text{GELU}\Big(\text{BN}\Big(\mathbf{W}_i^{(1)} * \mathbf{F}\Big)\Big)\Big), \tag{5}$$

where $\mathbf{W}_i^{(1)}$ and $\mathbf{W}_i^{(2)}$ are the convolutional filters of the two Conv-BN layers within block i. The block output is then obtained by adding the residual branch to the shortcut and applying a final activation:

$$\tilde{\mathbf{F}}_i = \mathrm{GELU}(f_i(\mathbf{F}_{i-1}) + \mathcal{P}_i(\mathbf{F}_{i-1}))\,, \tag{6}$$

where $\mathcal{P}_i(\cdot)$ is a strided pointwise projection for the first three blocks and the identity mapping for the fourth. The SE module then computes a channel-wise attention vector:

$$\mathbf{s}_i = \sigma\big(\mathbf{W}_{i,2}^{\mathrm{se}}\,\mathrm{ReLU}\big(\mathbf{W}_{i,1}^{\mathrm{se}}\,\mathrm{GAP}(\tilde{\mathbf{F}}_i)\big)\big)\,, \tag{7}$$

where $\mathrm{ReLU}(\cdot)$ denotes the rectified linear unit activation, $\mathrm{GAP}(\cdot)$ denotes global average pooling, and $\mathbf{W}_{i,1}^{\mathrm{se}}$ and $\mathbf{W}_{i,2}^{\mathrm{se}}$ are pointwise linear projections. The recalibrated block output is then $\mathbf{F}_i = \tilde{\mathbf{F}}_i \odot \mathbf{s}_i$, where $\mathbf{s}_i$ is broadcast along the spectral dimension.

To preserve fine-grained spectral details from shallower layers, multi-scale feature fusion is performed by resampling $\mathbf{F}_2$ and $\mathbf{F}_1$ to resolution L' via nearest-neighbor interpolation $\mathcal{R}(\cdot)$, with $\mathbf{F}_1$ additionally projected to C channels via a pointwise convolution $\phi(\cdot)$. The resampled features are concatenated with $\mathbf{F}_4$ along the channel dimension to form

$$\mathbf{Y} = \big[\mathbf{F}_4,\ \mathcal{R}(\mathbf{F}_2),\ \phi(\mathcal{R}(\mathbf{F}_1))\big] \in \mathbb{R}^{3C\times L'}. \tag{8}$$

A pointwise channel-mixing multilayer perceptron (MLP) $\psi(\cdot)$ is then applied independently at each spectral position to reduce the channel dimension from $3C$ to C, producing the unified feature map:

$$\mathbf{F} = \psi(\mathbf{Y}) \in \mathbb{R}^{C\times L'}. \tag{9}$$

Patch-based Transformer encoder.

The convolutional feature map $\mathbf{F} \in \mathbb{R}^{C\times L'}$ is tokenized by a strided convolutional patch embedding layer with kernel size P and stride $P/2$, which projects each overlapping local segment into a d-dimensional token embedding and produces a sequence of N patch tokens $\mathbf{E} \in \mathbb{R}^{N\times d}$. A learnable `[CLS]` token $\mathbf{e}_{\mathrm{cls}} \in \mathbb{R}^d$ is prepended, and a learnable positional embedding $\mathbf{E}_{\mathrm{pos}} \in \mathbb{R}^{(N+1)\times d}$ is added to encode sequential positional information, yielding the initial token sequence:

$$\mathbf{T}^{(0)} = \big[\mathbf{e}_{\mathrm{cls}};\ \mathbf{E}\big] + \mathbf{E}_{\mathrm{pos}} \in \mathbb{R}^{(N+1)\times d}. \tag{10}$$

The sequence is then processed by N_{layers} stacked Transformer encoder layers with pre-layer normalization, where each layer applies a multi-head self-attention sub-layer (H heads, per-head dimension $d_h = d/H$) followed by a two-layer feed-forward network with GELU activation and hidden dimension $4d$, each preceded by layer normalization and wrapped with a residual connection. The final hidden state corresponding to the `[CLS]` token is extracted from the last Transformer layer as the global latent spectral representation:

$$\mathbf{z} = \mathbf{T}_{[0]}^{(N_{\mathrm{layers}})} \in \mathbb{R}^d. \tag{11}$$

Task-specific downstream architectures

The hybrid convolutional–Transformer encoder is shared across downstream tasks. Most classification and regression tasks apply the shared spectral encoder to a single input spectrum and pass the resulting representation to a task-specific MLP head. Two task families require additional task-specific modules. Formula-conditioned molecular structure elucidation takes a molecular formula as an additional input and generates SMILES strings. Pairwise mixture analysis takes a reference single-compound spectrum together with a mixture spectrum to determine whether the reference compound is present and to estimate its fractional contribution. By contrast, mixture-level component quantification from a mixture spectrum alone follows the standard single-spectrum regression pipeline. The following sections describe the additional modules used for structure generation and pairwise mixture analysis.

Molecular structure elucidation.

Molecular structure elucidation is formulated as formula-conditioned autoregressive SMILES generation. Given an IR spectrum and its associated molecular formula, the model generates a ranked list of candidate SMILES strings by connecting the pretrained encoder to a molecular formula encoder, a Perceiver-style IR token resampler, a spectral vocabulary prompt module, and an autoregressive SMILES decoder.

The molecular formula string is tokenized at the character level and encoded by a formula encoder. The encoder consists of a token embedding layer with sinusoidal positional encodings and dropout, followed by N_f pre-norm Transformer encoder layers, each with multi-head self-attention and a GELU feed-forward sublayer, and a final layer normalization. It produces formula hidden states $\mathbf{H}_f \in \mathbb{R}^{M_f \times d}$, where M_f denotes the number of formula tokens.

On the spectral side, the backbone extracts both the global `[CLS]` feature $\mathbf{z} \in \mathbb{R}^d$ and the full patch token sequence $\mathbf{T}_{\text{patch}} \in \mathbb{R}^{N \times d}$ from its final Transformer layer. A projection head comprising layer normalization, a linear layer, and GELU activation maps $\mathbf{z}$ to a single IR memory token $\mathbf{m}_{\text{IR}} \in \mathbb{R}^d$. The patch tokens are further compressed by a Perceiver-style resampler into K latent vectors. The resampler maintains K learnable query embeddings and refines them over N_r cross-attention blocks. Unlike standard post-norm Perceivers, each block applies pre-norm layer normalization separately to the queries and the key–value source before cross-attention. A pre-norm GELU feed-forward sublayer follows, with residual connections throughout. Denoting the latent state at layer l as $\mathbf{X}^{(l)} \in \mathbb{R}^{K \times d}$, the update rule per block is:

$$\mathbf{X}^{(l)} = \mathbf{X}^{(l-1)} + \text{CrossAttn}\Big(\text{LN}_q\Big(\mathbf{X}^{(l-1)}\Big),\ \text{LN}_{kv}(\mathbf{T}_{\text{patch}})\Big), \tag{12}$$

$$\mathbf{X}^{(l)} \leftarrow \mathbf{X}^{(l)} + \text{FFN}\Big(\mathbf{X}^{(l)}\Big), \tag{13}$$

with $\mathbf{X}^{(0)} = \mathbf{Q}_{\text{lat}} \in \mathbb{R}^{K \times d}$ the learnable query matrix, yielding $\mathbf{M}_{\text{lat}} = \mathbf{X}^{(N_r)} \in \mathbb{R}^{K \times d}$.

A soft vocabulary prompt token is constructed from $\mathbf{z}$ to bias the decoder toward chemically plausible token sequences. A task-specific MLP head maps $\mathbf{z}$ to logits over the SMILES vocabulary of size V. A temperature-scaled softmax then converts these logits into a probability distribution, which is used to compute a soft embedding as a weighted sum over the decoder embedding matrix $\mathbf{E} \in \mathbb{R}^{V \times d}$. The soft embedding is then projected and scaled by a learnable scalar gate α to yield the prompt token:

$$\mathbf{p} = \text{softmax}\left(\frac{\text{MLP}_{\text{cls}}(\mathbf{z})}{\tau}\right) \in \mathbb{R}^V, \qquad \mathbf{m}_{\text{prompt}} = \alpha \cdot f_{\text{proj}}\big(\mathbf{p}^\top \mathbf{E}\big) \in \mathbb{R}^d, \tag{14}$$

where MLP_{cls} denotes the task-specific vocabulary-classification head, τ is a fixed temperature hyperparameter, and f_{proj} denotes a three-stage projection comprising layer normalization, a linear transformation, and GELU activation.

The complete decoder memory is formed by concatenating all spectral tokens and the formula hidden states:

$$\mathbf{M} = \big[\mathbf{m}_{\text{IR}};\ \mathbf{M}_{\text{lat}};\ \mathbf{m}_{\text{prompt}};\ \mathbf{H}_f\big] \in \mathbb{R}^{(K+2+M_f) \times d}. \tag{15}$$

The SMILES decoder consists of N_s pre-norm Transformer layers, each applying causal self-attention under a triangular mask, cross-attention to $\mathbf{M}$, and a GELU feed-forward sublayer, followed by a final layer normalization and a linear projection to vocabulary logits. During training, teacher forcing is applied, and the primary loss is token-level cross-entropy $\mathcal{L}_{\text{LM}}$. Two contrastive losses further shape the representation. An IR–formula contrastive loss $\mathcal{L}_{\text{IR-F}}$ aligns $\mathbf{z}$ with the masked mean-pooled formula encoding of $\mathbf{H}_f$, whereas an IR–target contrastive loss $\mathcal{L}_{\text{IR-T}}$ aligns $\mathbf{z}$ with the masked mean-pooled decoder hidden states at the final Transformer layer. Both use symmetric information noise-contrastive estimation (InfoNCE) with two-layer projection heads that map the inputs into a shared d_c-dimensional contrastive space. The training objective is:

$$\mathcal{L}_{\text{struct}} = \mathcal{L}_{\text{LM}} + \lambda_{\text{IF}}\,\mathcal{L}_{\text{IR-F}} + \lambda_{\text{IT}}\,\mathcal{L}_{\text{IR-T}}. \tag{16}$$

At inference, beam search with length-normalized scoring generates multiple ranked candidate SMILES sequences. A formula-consistency reranking step then promotes candidates whose SMILES match the exact atomic composition of the provided molecular formula and demotes chemically invalid structures.

Targeted component detection and contribution estimation.

The pairwise mixture-analysis tasks, targeted component detection and targeted fractional contribution estimation, are unified under a common spectral comparison framework. Given a reference single-compound spectrum and a mixture spectrum, the model predicts whether the reference compound is present in the mixture and estimates its fractional contribution. The two tasks share the same pairwise representation backbone and differ only in their final prediction head.

Each input pair $(\mathbf{x}_{\text{ref}}, \mathbf{x}_{\text{mix}}) \in \mathbb{R}^{1\times L} \times \mathbb{R}^{1\times L}$ is encoded by three parallel branches that extract complementary representations of the spectral pair.

The first branch uses a weight-shared encoder to process each spectrum independently, producing a global `[CLS]` feature and a patch token sequence for each spectrum. The patch tokens are projected to d_p dimensions and aggregated by attentive pooling, combining an attention-weighted sum with an element-wise maximum:

$$\mathbf{p}_{\text{pool}} = \tfrac{1}{2}\left(\sum_{n=1}^{N} \alpha_n \, \mathbf{t}_n + \max_n \mathbf{t}_n\right), \qquad \alpha_n = \frac{\exp(s_n)}{\sum_{n'} \exp(s_{n'})}, \tag{17}$$

where $\mathbf{t}_n$ are the projected patch tokens and s_n is an attention score produced by layer normalization followed by a linear projection.

The second branch operates directly on the raw input signals. The reference spectrum, mixture spectrum, and their absolute pointwise difference are stacked into a three-channel tensor $[\mathbf{x}_{\text{ref}}, \mathbf{x}_{\text{mix}}, |\mathbf{x}_{\text{mix}} - \mathbf{x}_{\text{ref}}|] \in \mathbb{R}^{3\times L}$. Successive Conv-BN-GELU blocks with progressively doubling channel widths process this tensor and reduce its spectral resolution by a fixed factor. The resulting feature sequence is augmented with sinusoidal positional encodings, processed by a shallow Transformer encoder, attentively pooled, and projected to yield the joint feature $\mathbf{h}_{\text{joint}} \in \mathbb{R}^{d_p}$.

The third branch provides complementary frequency-domain features. A real-input fast Fourier transform (FFT) is applied to each spectrum, yielding complex coefficients $\hat{r}_f$ and $\hat{m}_f$ for the reference and mixture spectra at frequency bin f. Eight frequency-domain channels are then constructed: the log-amplitude spectra $\log(1 + |\hat{r}_f|)$ and $\log(1 + |\hat{m}_f|)$, their sum and absolute difference, the normalized cross-magnitude term $\kappa_f = |\hat{r}_f \hat{m}_f^*| / (|\hat{r}_f|\,|\hat{m}_f| + \varepsilon)$, the normalized cross-spectral phase $\phi_f = \angle(\hat{r}_f \hat{m}_f^*)/\pi$, the log-amplitude ratio $\rho_f = \log(1 + (|\hat{m}_f| + \varepsilon)/(|\hat{r}_f| + \varepsilon))$, and the log-amplitude product $\pi_f = \log(1 + |\hat{r}_f|\,|\hat{m}_f|)$, where $(\cdot)^*$ denotes complex conjugation, $\angle(\cdot)$ denotes the complex argument, and ε is a small constant for numerical stability. These channels are concatenated to form the frequency-domain tensor $\mathbf{X}_{\text{freq}} \in \mathbb{R}^{8\times F}$, which is processed by three convolutional sub-branches with short, medium, and long receptive fields. The resulting feature maps are concatenated, compressed by strided convolutions, globally average-pooled, and projected to yield the frequency representation $\mathbf{h}_{\text{freq}} \in \mathbb{R}^{d_p}$.

The pair representation $\mathbf{r}_{\text{pair}}$ is assembled by concatenating five d_p-dimensional feature vectors with a ten-dimensional scalar statistics vector $\mathbf{s}_{\text{stats}}$. The five feature vectors are the projected `[CLS]` features $\mathbf{c}_{\text{ref}}$ and $\mathbf{c}_{\text{mix}}$, their absolute difference $|\mathbf{c}_{\text{ref}} - \mathbf{c}_{\text{mix}}|$, the absolute difference of the pooled token features $|\mathbf{p}_{\text{ref}} - \mathbf{p}_{\text{mix}}|$, and the frequency feature $\mathbf{h}_{\text{freq}}$. The statistics vector $\mathbf{s}_{\text{stats}}$ collects the cosine similarity between the `[CLS]` features, the cosine similarity between the pooled-token features, the cosine similarity between $\mathbf{h}_{\text{joint}}$ and $\mathbf{h}_{\text{freq}}$, the mean squared activation energy of the joint Transformer tokens, the cosine similarity of the log-amplitude spectra, the mean and maximum absolute log-amplitude difference, the mean and maximum normalized cross-magnitude term, and the mean absolute normalized cross-spectral phase:

$$\mathbf{r}_{\text{pair}} = \big[\mathbf{c}_{\text{ref}};\ \mathbf{c}_{\text{mix}};\ |\mathbf{c}_{\text{ref}} - \mathbf{c}_{\text{mix}}|;\ |\mathbf{p}_{\text{ref}} - \mathbf{p}_{\text{mix}}|;\ \mathbf{h}_{\text{freq}};\ \mathbf{s}_{\text{stats}}\big]. \tag{18}$$

The pair representation $\mathbf{r}_{\text{pair}}$ is then passed to a task-specific prediction head. For targeted component detection, a binary classification head maps $\mathbf{r}_{\text{pair}}$ to a scalar logit and is trained with binary cross-entropy loss. For targeted fractional contribution estimation, a regression head maps $\mathbf{r}_{\text{pair}}$ to a scalar contribution estimate $\hat{a} \in [0, 1]$ through a sigmoid output layer and is trained with mean squared error against the normalized target contribution.

Pretraining

In the large-scale simulation-based pretraining stage, each preprocessed IR spectrum $\mathbf{x}$ is first stochastically augmented to obtain an encoder input $\tilde{\mathbf{x}}$. The augmented spectrum is passed through the hybrid convolutional–Transformer encoder f_θ, producing the latent representation $\mathbf{z} = f_\theta(\tilde{\mathbf{x}})$. Three complementary objectives are then applied to outputs derived from $\mathbf{z}$: wavelet-domain spectral reconstruction, molecular fingerprint similarity alignment, and multi-label functional-group prediction. Together, these objectives encourage the learned representation to capture spectral morphology, molecular structural relationships, and interpretable functional-group signatures. The total pretraining loss is their weighted combination:

$$\mathcal{L}_{\text{pre}} = \lambda_{\text{recon}}\,\mathcal{L}_{\text{recon}} + \lambda_{\text{contrast}}\,\mathcal{L}_{\text{contrast}} + \lambda_{\text{fg}}\,\mathcal{L}_{\text{fg}}, \tag{19}$$

where λ_{recon}, $\lambda_{\text{contrast}}$, and λ_{fg} are scalar loss weights. The encoder and pretraining heads are optimized jointly by backpropagating $\mathcal{L}_{\text{pre}}$ using AdamW. Each objective is described below.

Wavelet-domain spectral reconstruction.

To encourage the encoder to preserve fine-grained spectral structure, wavelet-domain spectral reconstruction is used as one of the pretraining objectives. A dedicated wavelet reconstruction head maps $\mathbf{z}$ through a two-stage bottleneck, each stage comprising layer normalization, a linear transformation, a GELU activation, and dropout, yielding a compressed representation $\mathbf{b} \in \mathbb{R}^{d_b}$. This representation is decoded into the approximation coefficients and $J = 4$ levels of detail coefficients of a discrete wavelet transform (DWT) using a Daubechies-4 wavelet. Specifically, the predicted approximation coefficients $\hat{\mathbf{a}} \in \mathbb{R}^{L_0}$ and the j-th-level detail coefficients $\hat{\mathbf{d}}_j \in \mathbb{R}^{L_j}$ are given by:

$$\hat{\mathbf{a}} = e^{\alpha_0}\,\mathbf{W}_a\,\mathbf{b}, \qquad \hat{\mathbf{d}}_j = e^{\alpha_j}\,\mathbf{W}_{d,j}\,\mathbf{b}, \quad j = 1, \ldots, J, \tag{20}$$

where $\mathbf{W}_a$ and each $\mathbf{W}_{d,j}$ are linear projections to the respective band lengths, and $\alpha_0, \alpha_1, \ldots, \alpha_J$ are learnable log-scale parameters that explicitly accommodate the heterogeneous energy scales across wavelet subbands. The predicted coefficients are then passed through the inverse DWT (IDWT) to recover the full-resolution spectral estimate $\hat{\mathbf{x}} \in \mathbb{R}^L$.

The reconstruction objective combines a wavelet-coefficient-domain loss and a signal-domain loss. Denoting by $\mathbf{a}$ and $\{\mathbf{d}_j\}_{j=1}^{J}$ the target DWT coefficients of the unaugmented spectrum $\mathbf{x}$, the coefficient loss applies the smooth-ℓ_1 loss $\ell_{\text{H}}(\cdot,\cdot)$ independently to each wavelet band and averages uniformly across all $J+1$ bands:

$$\mathcal{L}_{\text{coeff}} = \frac{1}{J+1}\left[\ell_{\text{H}}(\hat{\mathbf{a}},\,\mathbf{a}) + \sum_{j=1}^{J} \ell_{\text{H}}\left(\hat{\mathbf{d}}_j,\,\mathbf{d}_j\right)\right]. \tag{21}$$

The signal-domain loss is computed between $\hat{\mathbf{x}}$ and the ground-truth $\mathbf{x}$ using a spatially weighted smooth-ℓ_1 loss. Leveraging the binary mask $\mathbf{m}$ introduced during data augmentation, masked positions are assigned an elevated weight $m_w > 1$ to concentrate reconstruction capacity on the corrupted spectral regions. The per-position weight is $w_l = 1 + (m_w - 1)\,m_l$, and the signal-domain loss is computed as the weighted average:

$$\mathcal{L}_{\text{signal}} = \frac{\sum_{l=1}^{L} w_l\;\ell_{\text{H}}(\hat{x}_l,\,x_l)}{\sum_{l=1}^{L} w_l}. \tag{22}$$

The two reconstruction terms are combined as a weighted sum:

$$\mathcal{L}_{\text{recon}} = \lambda_{\text{coeff}}\,\mathcal{L}_{\text{coeff}} + \lambda_{\text{signal}}\,\mathcal{L}_{\text{signal}}. \tag{23}$$

By jointly supervising coarse-scale spectral envelopes through the approximation band and fine-scale absorption line shapes through the detail bands, this pretraining objective encourages the encoder to learn representations that capture both global spectral morphology and local absorption features.

Molecular fingerprint similarity alignment.

To encourage the latent space to reflect graded structural relationships among molecules, UltraIR is additionally trained with a soft Tanimoto contrastive objective. This objective operates on a dedicated projection of the encoder output: a two-layer projection head maps $\mathbf{z}$ to a d_p-dimensional embedding $\mathbf{p} \in \mathbb{R}^{d_p}$ via layer normalization, a linear projection, a GELU activation, dropout, and a final linear projection.

Rather than defining hard positive and negative pairs, this objective constructs a continuous structural similarity target from precomputed binary molecular fingerprints. For a batch of B samples with binarized fingerprints $\mathbf{f}_i \in \{0,1\}^M$, the pairwise Tanimoto similarity is:

$$T_{ij} = \frac{\mathbf{f}_i^\top \mathbf{f}_j}{\|\mathbf{f}_i\|_1 + \|\mathbf{f}_j\|_1 - \mathbf{f}_i^\top \mathbf{f}_j + \varepsilon}. \tag{24}$$

A soft target distribution p_{ij} over off-diagonal batch members is obtained by applying a softmax with temperature τ_t to each row of T after masking the diagonal. Student similarity logits are computed as scaled inner products between ℓ_2-normalized embeddings $\bar{\mathbf{p}}_i = \mathbf{p}_i / \|\mathbf{p}_i\|_2$, giving $s_{ij} = \bar{\mathbf{p}}_i^\top \bar{\mathbf{p}}_j / \tau_s$. The contrastive loss is the soft cross-entropy between the student distribution and the Tanimoto-derived target:

$$\mathcal{L}_{\text{contrast}} = -\frac{1}{B} \sum_i \sum_{j \neq i} p_{ij} \log \frac{\exp(s_{ij})}{\sum_{k \neq i} \exp(s_{ik})}. \tag{25}$$

This objective encourages spectra of structurally related compounds to occupy nearby regions of the latent space in proportion to their fingerprint overlap.

Multi-label functional-group prediction.

A multilayer classification head maps the latent representation $\mathbf{z}$ to logits over the functional-group classes. It consists of layer normalization followed by three linear projections with progressively reducing dimensionality, interleaved with GELU activations and dropout. The model is trained to predict the presence or absence of each functional group as an independent binary classification task using binary cross-entropy with logits. This objective provides explicit chemical supervision that anchors the latent space to interpretable structural motifs.

Downstream adaptation

For downstream adaptation, the pretrained checkpoint was used to initialize only the shared spectral encoder components present in both the pretraining and downstream models. These components included the derivative-aware multichannel input module, the convolutional spectral encoder, and the patch-based Transformer encoder. Pretraining-specific heads, including the reconstruction head, contrastive projection head, and pretraining functional-group prediction head, were not transferred to downstream tasks. Consequently, even when a downstream task shared a prediction type or label space with a pretraining objective, its task-specific output module was newly initialized rather than inherited from the pretraining checkpoint. Other components introduced exclusively for downstream prediction were also initialized randomly.

Downstream task heads

Classification head.

All classification tasks use task-specific instances of the same MLP head architecture. The head consists of layer normalization followed by three linear projections with progressively reducing dimensionality, interleaved with GELU activations and a dropout layer. The final projection maps the hidden representation to C output logits, where C denotes the number of classes for each task. Single-spectrum classification tasks receive the global spectral representation $\mathbf{z}$ as input, whereas targeted component detection receives the dedicated pair representation $\mathbf{r}_{\text{pair}}$.

For functional-group prediction, each output logit corresponds to an independent binary classification target, and the head is trained with binary cross-entropy loss. Bacterial classification, medicinal-herb geographic origin traceability, and microplastics classification are formulated as single-label multiclass classification tasks and trained with cross-entropy loss, with label smoothing applied for medicinal-herb geographic origin traceability. Targeted component detection produces a scalar logit and is trained with binary cross-entropy loss.

Regression head.

For regression tasks, including physicochemical property prediction, targeted fractional contribution estimation, mixture-level component quantification, medicinal-herb constituent quantification, and soil property prediction, a task-specific three-layer MLP regression head processes the input representation. This head follows the same architecture as the classification head but uses task-specific output layers. Single-spectrum regression tasks receive the global spectral representation $\mathbf{z} \in \mathbb{R}^d$ as input, whereas targeted fractional contribution estimation receives the dedicated pair representation $\mathbf{r}_{\text{pair}}$.

To stabilize training across target dimensions with heterogeneous numerical scales, regression targets were transformed using target-wise scaling. For single-target regression tasks, including targeted fractional contribution estimation, the final output is passed through a sigmoid activation and optimized using mean squared error loss. For multi-target regression tasks, including physicochemical property prediction, mixture-level component quantification, medicinal-herb constituent quantification, and soil property prediction, the regression head adopts a linear output layer and is optimized using a smooth-ℓ_1 regression loss.

5 Data and code availability

The code is available at https://github.com/AIMS-Lab-HKUSTGZ/UltraIR. The checkpoints of UltraIR are available at https://huggingface.co/yusentan/UltraIR.

The newly generated UltraIR molecular-dynamics simulated infrared dataset used for UltraIR pretraining can be accessed at https://huggingface.co/yusentan/UltraIR. The simulated infrared dataset from IRtoMol [19] used for UltraIR pretraining can be accessed through https://zenodo.org/records/7928396. The simulated infrared dataset from the multimodal spectroscopy dataset [34] used for UltraIR pretraining can be accessed through https://zenodo.org/records/14770232. The simulated infrared dataset from QM9S [35] used for UltraIR pretraining can be accessed through https://figshare.com/articles/dataset/QM9S_dataset/24235333.

The real infrared dataset from the NIST Chemistry WebBook [37] used for molecular-level interpretation and mixture analysis benchmarks can be accessed through https://webbook.nist.gov/chemistry/. The real infrared dataset from SDBS [38] used for molecular-level interpretation and mixture analysis benchmarks can be accessed through https://sdbs.db.aist.go.jp/. The simulated infrared dataset from USPTO [36] used for molecular-level interpretation and mixture analysis benchmarks can be accessed through https://zenodo.org/records/16417648. The external liquid-mixture benchmark from DeepMIR [22] used for mixture analysis benchmarks can be accessed through https://github.com/LinTan-CSU/DeepMIR. The experimental FTIR mixture dataset [39] used for mixture analysis benchmarks can be accessed through https://doi.org/10.5281/zenodo.5498197.

The green-snow bacterial FTIR dataset [40] used for bacterial classification can be accessed through https://zenodo.org/records/4297950. The two newly generated in-house experimental datasets of Jinyinhua and Shanyinhua used for medicinal-herb characterization can be accessed at https://huggingface.co/yusentan/UltraIR. The environmentally sourced microplastics infrared dataset [7] used for microplastics classification can be accessed through https://drive.google.com/drive/folders/11MofhjEchgZelWPcHUvIMRPNEQPaLfUO?usp=sharing. The OSSL dataset [6] used for soil property prediction can be accessed through https://docs.soilspectroscopy.org/db-access.html.

6 Acknowledgments

This research was supported by the National Natural Science Foundation of China Project (No. 623B2086), CCF-GHFund (No. OF 2026005), the CIPS-SMP-Zhipu Large Model Fund, Ant Group, Shanghai Artificial Intelligence Laboratory, and TeleAI of China Telecom.

7 Author contributions

Conceptualization, Y.T. and J.X.; methodology, Y.T., Y.C., and J.X.; software, Y.T.; data curation, Y.T., Y.C., Z.F., P.L., and Y.F.L.; formal analysis, Y.T., Y.C., and Z.F.; investigation, Y.T., Y.C., Z.F., P.L., and Y.F.L.; validation, Y.T., Y.C., Z.F., P.L., and Y.F.L.; visualization, Y.T., Y.C., Q.G., and Z.L.; writing – original draft, Y.T.; writing – review & editing, all authors; resources, J.X., Y.Q.L., and T.W.; supervision, J.X., X.Z., and T.W.; project administration, J.X.; funding acquisition, J.X.

8 Competing interests

The authors declare no competing interests.

References

[1] Barbara H Stuart. Infrared spectroscopy: fundamentals and applications. John Wiley & Sons, 2004.

[2] Peter R Griffiths. Fourier transform infrared spectrometry. Science, 222(4621):297–302, 1983.

[3] Sergei G Kazarian and KL Andrew Chan. ATR-FTIR spectroscopic imaging: recent advances and applications to biological systems. Analyst, 138(7):1940–1951, 2013.

[4] Sander De Bruyne, Marijn M Speeckaert, and Joris R Delanghe. Applications of mid-infrared spectroscopy in the clinical laboratory setting. Critical Reviews in Clinical Laboratory Sciences, 55(1):1–20, 2018.

[5] Marinus Huber, Kosmas V Kepesidis, Liudmila Voronina, Maša Božić, Michael Trubetskov, Nadia Harbeck, Ferenc Krausz, and Mihaela Žigman. Stability of person-specific blood-based infrared molecular fingerprints opens up prospects for health monitoring. Nature Communications, 12(1):1511, 2021.

[6] José L Safanelli, Tomislav Hengl, Leandro L Parente, Robert Minarik, Dellena E Bloom, Katherine Todd-Brown, Asa Gholizadeh, Wanderson de Sousa Mendes, and Jonathan Sanderman. Open Soil Spectral Library (OSSL): Building reproducible soil calibration models through open development and community engagement. PLOS ONE, 20(1):e0296545, 2025.

[7] Junhao Xie, Aoife Gowen, and Jun-Li Xu. Open-set convolutional neural network for infrared spectral classification of environmentally sourced microplastics. npj Emerging Contaminants, 2(1):6, 2026.

[8] Tarek Eissa, Liudmila Voronina, Marinus Huber, Frank Fleischmann, and Mihaela Žigman. The perils of molecular interpretations from vibrational spectra of complex samples. Angewandte Chemie International Edition, 63(50):e202411596, 2024.

[9] Matthew J Baker, Júlio Trevisan, Paul Bassan, Rohit Bhargava, Holly J Butler, Konrad M Dorling, Peter R Fielden, Simon W Fogarty, Nigel J Fullwood, Kelly A Heys, et al. Using Fourier transform IR spectroscopy to analyze biological materials. Nature Protocols, 9(8):1771–1791, 2014.

[10] Rekha Gautam, Sandeep Vanga, Freek Ariese, and Siva Umapathy. Review of multidimensional data processing approaches for Raman and infrared spectroscopy. EPJ Techniques and Instrumentation, 2(1):8, 2015.

[11] Kas J Houthuijs, Giel Berden, Udo FH Engelke, Vasuk Gautam, David S Wishart, Ron A Wevers, Jonathan Martens, and Jos Oomens. An in silico infrared spectral library of molecular ions for metabolite identification. Analytical Chemistry, 95 (23):8998–9005, 2023.

[12] Brian C Smith. Infrared spectral interpretation: a systematic approach. CRC Press, 2018.

[13] Lisa M. Miller and John P. Coates. Interpretation of Infrared Spectra: A Practical and Systematic Approach. In Encyclopedia of Analytical Chemistry, pages 1–24. Wiley, 2025.

[14] Joshua L Lansford and Dionisios G Vlachos. Infrared spectroscopy data- and physics-driven machine learning for characterizing surface microstructure of complex materials. Nature Communications, 11(1):1513, 2020.

[15] Zhihao Ren, Zixuan Zhang, Jingxuan Wei, Bowei Dong, and Chengkuo Lee. Wavelength-multiplexed hook nanoantennas for machine learning enabled mid-infrared spectroscopy. Nature Communications, 13(1):3859, 2022.

[16] Jianxiong Zhu, Shanling Ji, Zhihao Ren, Wenyu Wu, Zhihao Zhang, Zhonghua Ni, Lei Liu, Zhisheng Zhang, Aiguo Song, and Chengkuo Lee. Triboelectric-induced ion mobility for artificial intelligence-enhanced mid-infrared gas spectroscopy. Nature Communications, 14(1):2524, 2023.

[17] Vu Hoang Minh Doan, Cao Duong Ly, Sudip Mondal, Thi Thuy Truong, Tan Dung Nguyen, Jaeyeop Choi, Byeongil Lee, and Junghwan Oh. Fcg-Former: identification of functional groups in FTIR spectra using enhanced transformer-based model. Analytical Chemistry, 96(30):12358–12369, 2024.

[18] Dev Punjabi, Yu-Chieh Huang, Laura Holzhauer, Pierre Tremouilhac, Pascal Friederich, Nicole Jung, and Stefan Bräse. Infrared spectrum analysis of organic molecules with neural networks using standard reference data sets in combination with real-world data. Journal of Cheminformatics, 17(1):24, 2025.

[19] Marvin Alberts, Teodoro Laino, and Alain C Vaucher. Leveraging infrared spectroscopy for automated structure elucidation. Communications Chemistry, 7(1):268, 2024.

[20] Wenjin Wu, Ales Leonardis, Jianbo Jiao, Jun Jiang, and Linjiang Chen. Transformer-based models for predicting molecular structures from infrared spectra using patch-based self-attention. The Journal of Physical Chemistry A, 129(8):2077–2085, 2025.

[21] Ganesh Chandan Kanakala, Bhuvanesh Sridharan, and U Deva Priyakumar. Spectra to structure: contrastive learning framework for library ranking and generating molecular structures for infrared spectra. Digital Discovery, 3(12):2417–2423, 2024.

[22] Lin Tan, Yue Wang, Hailiang Zhang, Jinyu Sun, Qiong Yang, Xiao Yang, Zhimin Zhang, and Hongmei Lu. DeepMIR: A Hybrid Convolutional Neural Network-Transformer Framework for Accurate Identification of Target Components from Mid-Infrared Spectra of Mixtures. Analytical Chemistry, 97(50):27706–27715, 2025.

[23] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. In Advances in Neural Information Processing Systems, volume 33, pages 1877–1901, 2020.

[24] Rishi Bommasani, Drew A Hudson, Ehsan Adeli, Russ Altman, Simran Arora, Sydney von Arx, Michael S Bernstein, Jeannette Bohg, Antoine Bosselut, Emma Brunskill, et al. On the opportunities and risks of foundation models, 2021.

[25] Yukun Zhou, Mark A Chia, Siegfried K Wagner, Murat S Ayhan, Dominic J Williamson, Robbert R Struyven, Timing Liu, Moucheng Xu, Mateo G Lozano, Peter Woodward-Court, et al. A foundation model for generalizable disease detection from retinal images. Nature, 622(7981):156–163, 2023.

[26] DongAo Ma, Jiaxuan Pang, Michael B Gotway, and Jianming Liang. A fully open AI foundation model applied to chest radiography. Nature, 643(8071):488–498, 2025.

[27] Guy Lutsker, Gal Sapir, Smadar Shilo, Jordi Merino, Anastasia Godneva, Jerry R Greenfield, Dorit Samocha-Bonet, Raja Dhir, Francisco Gude, Shie Mannor, et al. A foundation model for continuous glucose monitoring data. Nature, 650(8103): 978–986, 2026.

[28] Eric Y Wang, Paul G Fahey, Zhuokun Ding, Stelios Papadopoulos, Kayla Ponder, Marissa A Weis, Andersen Chang, Taliah Muhammad, Saumil Patel, Zhiwei Ding, et al. Foundation model of neural activity predicts response to new stimulus types. Nature, 640(8058):470–477, 2025.

[29] Kang Wu, Yingying Zhang, Lixiang Ru, Bo Dang, Jiangwei Lao, Lei Yu, Junwei Luo, Zifan Zhu, Yue Sun, Jiahao Zhang, et al. A semantic-enhanced multi-modal remote sensing foundation model for Earth observation. Nature Machine Intelligence, 7(8):1235–1249, 2025.

[30] Shoujie Li, Tong Wu, Jianle Xu, Yan Huang, Zongwen Zhang, Hongfa Zhao, Qinghao Xu, Zihan Wang, Linqi Ye, Yang Yang, et al. Biomimetic multimodal tactile sensing enables human-like robotic perception. Nature Sensors, 1(1):52–62, 2026.

[31] Hao Sun, Xin Gao, Wanhao Niu, Longwei Li, Yutong Song, Lei Liu, Huaidong Zhou, Huaping Liu, Zhong Lin Wang, Xiong Pu, et al. A spike–language dual framework bridges fast perception and deep reasoning in artificial tactile somatosensory systems. Nature Sensors, pages 1–12, 2026.

[32] Marvin Alberts, Federico Zipoli, and Teodoro Laino. Setting new benchmarks in AI-driven infrared structure elucidation. Digital Discovery, 4(7):1936–1943, 2025.

[33] Charles McGill, Michael Forsuelo, Yanfei Guan, and William H Green. Predicting infrared spectra with message passing neural networks. Journal of Chemical Information and Modeling, 61(6):2594–2609, 2021.

[34] Marvin Alberts, Oliver Schilter, Federico Zipoli, Nina Hartrampf, and Teodoro Laino. Unraveling molecular structure: A multimodal spectroscopic dataset for chemistry. In Advances in Neural Information Processing Systems, volume 37, pages 125780–125808, 2024.

[35] Zihan Zou, Yujin Zhang, Lijun Liang, Mingzhi Wei, Jiancai Leng, Jun Jiang, Yi Luo, and Wei Hu. A deep learning model for predicting selected organic molecular spectra. Nature Computational Science, 3(11):957–964, 2023.

[36] Federico Zipoli, Marvin Alberts, and Teodoro Laino. IR-NMR multimodal computational spectra dataset for 177K patent-extracted organic molecules. Scientific Data, 12(1):1375, 2025.

[37] Peter J Linstrom and William G Mallard. The NIST Chemistry WebBook: A chemical data resource on the internet. Journal of Chemical & Engineering Data, 46(5):1059–1063, 2001.

[38] National Institute of Advanced Industrial Science and Technology. SDBS: Spectral Database for Organic Compounds, 2026. URL https://sdbs.db.aist.go.jp/.

[39] Andrea Angulo, Lankun Yang, Eray S Aydil, and Miguel A Modestino. Machine learning enhanced spectroscopic analysis: towards autonomous chemical mixture characterization for rapid process optimization. Digital Discovery, 1(1):35–44, 2022.

[40] Margarita Smirnova, Achim Kohler, and Volha Shapaval. FTIR Dataset, 2020. URL https://doi.org/10.5281/zenodo.4297950.

[41] Tianqi Chen and Carlos Guestrin. XGBoost: A Scalable Tree Boosting System. In Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pages 785–794, 2016.

[42] Leo Breiman. Random forests. Machine Learning, 45(1):5–32, 2001.

[43] Fabian Pedregosa, Gaël Varoquaux, Alexandre Gramfort, Vincent Michel, Bertrand Thirion, Olivier Grisel, Mathieu Blondel, Peter Prettenhofer, Ron Weiss, Vincent Dubourg, et al. Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12:2825–2830, 2011.

[44] Colin Zhang and Yang Ha. Toward Complete Molecular Structure Prediction from Infrared Spectroscopy Using Deep Learning. Journal of Chemical Information and Modeling, 66(1):100–109, 2026.

[45] Svante Wold, Michael Sjöström, and Lennart Eriksson. PLS-regression: a basic tool of chemometrics. Chemometrics and Intelligent Laboratory Systems, 58(2):109–130, 2001.

[46] Jingkai Zhou, Zixuan Zhang, Bowei Dong, Zhihao Ren, Weixin Liu, and Chengkuo Lee. Midinfrared spectroscopic analysis of aqueous mixtures using artificial-intelligence-enhanced metamaterial waveguide sensing platform. ACS Nano, 17(1): 711–724, 2022.

[47] Min He, Jingjing Tong, Xiangxian Li, Xin Han, Yusheng Qin, Renjie Fang, Zidong Chen, and Minguang Gao. Research on hybrid microplastic recognition method based on dual-branch convolutional neural network combined with attention mechanism. Microchemical Journal, 218:115131, 2025.

[48] Wenqi Guo, Shichen Gao, Yaohui Ding, and Daming Dong. Simultaneous prediction of multiple soil components using Mid-Infrared Spectroscopy and the GADF-Swin Transformer model. Computers and Electronics in Agriculture, 237:110507, 2025.

[49] Yiqiang Liu, Luming Shen, Xinghui Zhu, Yangfan Xie, and Shaofang He. Spectral data-driven prediction of soil properties using LSTM-CNN-attention model. Applied Sciences, 14(24):11687, 2024.

[50] Yongqi Jin, Jun-Jie Wang, Fanjie Xu, Xiaohong Ji, Zhifeng Gao, Linfeng Zhang, Guolin Ke, Rong Zhu, and Weinan E. NMR-Solver: automated structure elucidation via large-scale spectral matching and physics-guided fragment optimization. Nature Communications, 17(1):4740, 2026.

[51] Peter Eastman, Jason Swails, John D Chodera, Robert T McGibbon, Yutong Zhao, Kyle A Beauchamp, Lee-Ping Wang, Andrew C Simmonett, Matthew P Harrigan, Chaya D Stern, et al. OpenMM 7: Rapid development of high performance algorithms for molecular dynamics. PLOS Computational Biology, 13(7):e1005659, 2017.

[52] Simon Boothroyd, Pavan Kumar Behara, Owen C. Madin, David F. Hahn, Hyesu Jang, Vytautas Gapsys, Jeffrey R. Wagner, Joshua T. Horton, David L. Dotson, Matthew W. Thompson, Jessica Maat, Trevor Gokey, Lee-Ping Wang, Daniel J. Cole, Michael K. Gilson, John D. Chodera, Christopher I. Bayly, Michael R. Shirts, and David L. Mobley. Development and Benchmarking of Open Force Field 2.0.0: The Sage Small Molecule Force Field. Journal of Chemical Theory and Computation, 19(11):3251–3275, 2023.

Appendix

A Supplementary experimental details

Pretraining hyperparameters

The principal hyperparameters used for large-scale UltraIR pretraining are summarized in Table A.

Table A | Pre-training hyperparameters of UltraIR.

Hyperparameter	Value
Optimization	
Learning rate	1×10^{-4}
Batch size	128
Epochs	5
Warmup ratio	5%
Transformer encoder	
Hidden dimension d	1024
Patch length P	16
Number of attention heads H	16
Number of Transformer layers N_{layers}	8
Dropout	0.05
Pre-training heads	
Fingerprint embedding dimension d_p	512
Number of wavelet decomposition levels J	4
Wavelet reconstruction bottleneck dimension d_b	256
Pre-training objectives	
Wavelet-domain spectral reconstruction loss weight λ_{recon}	50.0
Molecular fingerprint similarity alignment loss weight $\lambda_{\text{contrast}}$	0.2
Multi-label functional-group prediction loss weight λ_{fg}	1.25
Wavelet-coefficient-domain loss weight λ_{coeff}	0.2
Signal-domain loss weight λ_{signal}	0.8
Masked-position weighting factor m_w	3.0
Fingerprint student temperature τ_s	0.1
Fingerprint target temperature τ_t	0.05

Pretraining dataset construction

The UltraIR pretraining dataset comprised 60,000,000 simulated infrared (IR) spectra assembled from three complementary sources: publicly released simulated spectra, spectra generated using molecular dynamics, and spectra predicted using a machine-learning model. After removing molecules overlapping with the National Institute of Standards and Technology (NIST) Chemistry WebBook, Spectral Database for Organic Compounds (SDBS), and USPTO datasets, the publicly released sources contributed 630,395 spectra from IRtoMol, 743,955 from the multimodal spectroscopy dataset, and 127,898 from QM9S, yielding 1,502,248 spectra in total [19, 34, 35]. From structures in the PubChem-derived molecular collection distributed with the SimNMR-PubChem database [50] that were not represented in NIST, SDBS, or USPTO, we generated 7,534,293 spectra through molecular-dynamics simulation and predicted the remaining 50,963,459 spectra were predicted using Chemprop-IR [33].

For the molecular-dynamics component, three-dimensional molecular conformers were generated from Simplified Molecular Input Line Entry System (SMILES) representations and subjected to geometry relaxation before simulation. Each molecule was simulated without periodic boundary conditions at 300 K using a Langevin middle integrator in OpenMM, with the Open Force Field Sage 2.2.1 and fixed atom-centered partial charges [51, 52]. Molecular dipole trajectories were calculated from the fixed partial charges and instantaneous atomic positions. After temporal mean removal and Hann windowing, the three Cartesian dipole components were transformed independently using a fast Fourier transform. The resulting power spectra were summed, weighted by wavenumber, interpolated onto the common wavenumber grid, and normalized to a maximum intensity of one.

Table B | Functional-group classes and SMARTS definitions used for pretraining and downstream functional-group prediction. The listed order corresponds to the 17 dimensions of the multi-label target used in both stages. A class was assigned when the corresponding SMARTS pattern matched the molecular graph, and each molecule could receive multiple labels.

Index	Functional-group class	SMARTS definition
1	Alkane	`[CX4;H3,H2]`
2	Methyl	`[CH3]`
3	Alkene	`[CX3]=[CX3]`
4	Alkyne	`[CX2]#[CX2]`
5	Alcohols	`[#6][OX2H]`
6	Amines	`[NX3;H2,H1,H0;!$(N[CX3](=O))]`
7	Nitriles	`[NX1]#[CX2]`
8	Aromatics	`[$([cX3](:*):*),$([cX2+](:*):*)]`
9	Alkyl halides	`[#6][F,Cl,Br,I]`
10	Esters	`[#6][CX3](=O)[OX2H0][#6]`
11	Ketones	`[#6][CX3](=O)[#6]`
12	Aldehydes	`[CX3H1](=O)[#6]`
13	Carboxylic acids	`[CX3](=O)[OX2H1]`
14	Ether	`[OD2]([#6;!$(C=O)])([#6;!$(C=O)])`
15	Acyl halides	`[CX3](=[OX1])[F,Cl,Br,I]`
16	Amides	`[NX3][CX3](=[OX1])[#6]`
17	Nitro	`[$([N+](=O)[O-]),$([NX3](=O)=O)][#6]`

For the machine-learning-generated component, IR spectra were predicted using the two officially released pretrained Chemprop-IR checkpoints [33]. The two models were applied independently, and their predicted intensities were combined using an equally weighted arithmetic mean to obtain the final spectrum for each molecular structure.

Functional-group labels and molecular fingerprints used during pretraining were generated from molecular structures using RDKit. Functional-group labels were encoded as 17-dimensional multi-hot vectors, with each class assigned when the molecular graph matched its corresponding SMILES Arbitrary Target Specification (SMARTS) definition. Multiple classes could therefore be assigned to the same molecule. The functional-group classes, output order, and operational SMARTS definitions are provided in Table B. For molecular fingerprint similarity alignment, each structure was represented by a 2,048-bit Morgan fingerprint generated with a radius of 2.

Downstream datasets and benchmark summary

The eight downstream tasks were evaluated using ten datasets, with the Jinyinhua and Shanyinhua datasets each contributing separate subsets for geographic origin traceability and chemical constituent quantification. For the molecular-level benchmarks, we used NIST, SDBS, and USPTO. The NIST Chemistry WebBook, maintained as NIST Standard Reference Database 69, is a public resource that compiles chemical and spectroscopic data, including IR spectra [37]. We retained 23,286 NIST IR spectra after filtering. SDBS is a public database maintained by the National Institute of Advanced Industrial Science and Technology and contains experimentally measured Fourier-transform infrared (FTIR) spectra with corresponding molecular records for organic compounds [38]. We retained 18,253 SDBS IR spectra after filtering. The USPTO dataset is a computational multimodal spectroscopy resource constructed from patent-extracted organic molecules and contains 177,461 IR spectra generated using long-timescale molecular-dynamics simulations and machine-learning-assisted dipole-moment prediction [36].

For functional-group prediction, labels for NIST, SDBS, and USPTO were generated using the same 17 classes, class order, and SMARTS definitions as in pretraining (Table B). For physicochemical property prediction, all 11 targets were calculated from molecular structures using RDKit. The synthetic accessibility score (SAScore) was obtained using the RDKit SA_Score implementation, and the remaining targets were calculated using their corresponding RDKit descriptor functions.

Mixture-level and application-facing evaluations used several additional data sources. For external evaluation of targeted component detection, we used 360 spectra from the external liquid-mixture benchmark described in DeepMIR, comprising binary, ternary and quaternary mixtures [22]. Mixture-level component quantification was evaluated using 74 experimentally acquired spectra from a four-component FTIR mixture dataset comprising acrylonitrile, adiponi-

Table C | Summary of downstream benchmarks and comparison methods. The table summarizes the datasets and competing methods evaluated for each downstream task.

Task	Dataset	Compared methods
Functional-group prediction	NIST [37], SDBS [38], USPTO [36]	FCGFormer [17], IRAnalysis [18], XGBoost, Random Forest, KNN, Logistic classifier
Molecular structure elucidation	NIST [37], SDBS [38], USPTO [36]	IRtoMol [19], AISE [32], PBSA [20], DLIR [44]
Physicochemical property prediction	NIST [37], SDBS [38], USPTO [36]	XGBoost regression, SVR, KNN regression, PLSR
Targeted component detection	NIST [37], SDBS [38], USPTO [36] , DeepMIR liquid-mixture dataset [22]	DeepMIR [22], reverse match, HQI
Targeted fractional contribution estimation	NIST [37], SDBS [38], USPTO [36]	DeepMIR [22], XGBoost regression, SVR, KNN regression, PLSR
Mixture-level component quantification	Experimental FTIR mixture dataset [39]	AIMWSP [46], ML-FTIR [39], XGBoost regression, SVR, KNN regression, PLSR
Bacterial genus classification	Green-snow bacterial FTIR spectra [40]	XGBoost, Random Forest, KNN, Logistic classifier
Medicinal-herb origin traceability	Jinyinhua, Shanyinhua	XGBoost, Random Forest, KNN, Logistic classifier
Medicinal-herb constituent quantification	Jinyinhua, Shanyinhua	XGBoost regression, SVR, KNN regression, PLSR
Microplastics classification	Environmentally sourced microplastics IR dataset [7]	Softmax [7], DB-CNN-CBAM [47], XGBoost, Random Forest, KNN, Logistic classifier
Soil property prediction	OSSL [6]	GADF-Swin [48], LSTM-CNN [49], XGBoost regression, SVR, KNN regression, PLSR

trile, propionitrile and glycerol [39]. For this task, models were first trained on a simulated four-component mixture dataset containing 2,400 synthetic mixtures, and subsequently fine-tuned on the experimentally acquired spectra for evaluation [39]. Bacterial genus classification used 795 FTIR spectra from 45 fast-growing bacterial isolates collected from Antarctic green snow and spanning nine genera [40]. For medicinal-herb geographic origin traceability and constituent quantification, the dataset composition, LC-MS-derived relative-abundance regression targets, together with the corresponding analytical procedures, are detailed in the following subsection. The microplastics benchmark contained 32,965 IR spectra assembled from laboratory and open-source collections to represent diverse environmentally relevant polymer classes [7]. After filtering for complete wavenumber coverage and availability of the labels required for soil property prediction, we retained 46,099 mid-IR spectra from the Kellogg Soil Survey Laboratory (KSSL) collection within the Open Soil Spectral Library (OSSL) [6]. For cross-instrument and cross-laboratory evaluation, we independently filtered the datasets using the labels required for the cross-laboratory tasks and enforced complete wavenumber coverage. This procedure retained 55,974 KSSL spectra for model training and 3,753 ICRAF–ISRIC spectra for zero-shot target-domain inference. Table C summarizes the datasets and comparison methods used for each downstream task.

In-house medicinal-herb datasets and LC-MS-derived regression targets

In addition to the public and external benchmarks summarized above, we assembled two in-house medicinal-herb datasets: Jinyinhua (*Lonicerae Japonicae Flos*) and Shanyinhua (*Lonicerae Flos*). For geographic origin traceability, the Jinyinhua dataset contains 120 IR spectra from Shandong, Henan, Hebei, and Sichuan, and the Shanyinhua dataset contains 150 IR spectra from Hunan, Hubei, Sichuan, Henan, and Guangdong; each origin is represented by 30 spectra. For chemical constituent quantification, LC-MS-derived target relative abundances measured for 60 Jinyinhua and 75 Shanyinhua samples served as regression labels and were expressed in arbitrary units (a.u.). The Jinyinhua targets were 4-methylcoumarin, milrinone, geniposide, phenylacetaldehyde, cis-cinnamic acid, and lumazine; the Shanyinhua targets were coniferyl aldehyde, quercetin, kaempferol, and luteolin.

For LC-MS analysis, 30 mg of powdered material was extracted with 1.5 mL of 70% aqueous methanol. The extract was vortexed, ultrasonicated at 30 °C for 30 min, centrifuged at 12,000 r.p.m. for 15 min, and passed through a 0.22 μm membrane filter. Chromatographic separation was performed on a Kinetex F5 column (2.6 μm, 100 mm × 2.1 mm; Phenomenex) at 40 °C. The mobile phases were 0.1% formic acid in water (A) and 0.1% formic acid in acetonitrile (B), with a flow rate of 0.2 mL min^{-1} and an injection volume of 4 μL. The percentage of A was programmed as 100% at 0 min, 99% at 2 min, 95% at 3 min, 90% at 6 min, 85% at 14 min, 82% at 15 min, 80% at 18 min, 70% at 20 min, 60% at 23 min, 22% at 31 min, 10% at 33 min, 0% at 36–44 min, and 100% at 44.1–50 min. Full-scan time-of-flight mass spectrometry (TOF-MS) and information-dependent acquisition (IDA) tandem mass spectrometry (MS/MS) spectra were acquired in positive-ion mode over m/z 100–1,000 and m/z 50–1,000, respectively. The source temperature was 500 °C, the ion-spray voltage was +5,500 V, ion-source gases 1 and 2 were each 50 psi, curtain gas

Table D | Dataset metadata for cross-instrument and cross-laboratory soil-property prediction. Spectra from the Kellogg Soil Survey Laboratory (KSSL) and ICRAF–ISRIC, together with their associated metadata, were obtained from the Open Soil Spectral Library (OSSL) [6].

Characteristic	**Kellogg Soil Survey Laboratory database** *Source training domain*	**ICRAF–ISRIC Soil Spectral Library** *Target test domain*
Institution	USDA National Soil Survey Center	World Agroforestry Centre / ISRIC
Sample origin	United States	Not recorded in OSSL
FTIR system	Bruker Vertex 70 with HTS-XT	Bruker Tensor 27 with HTS-XT
Sample preparation	Ground to < 80 mesh	Ground to < 0.1 mm

Table E | Label distributions for cross-instrument and cross-laboratory soil-property prediction. The Kellogg Soil Survey Laboratory (KSSL) and ICRAF–ISRIC spectra and associated soil-property labels were obtained from the Open Soil Spectral Library (OSSL) [6]. Values are reported as median [minimum, maximum].

Property	Unit	**Kellogg Soil Survey Laboratory database** *Source training domain*	**ICRAF–ISRIC Soil Spectral Library** *Target test domain*
Clay Content	%	20.88 [0, 96.14]	30.10 [0, 96.80]
Silt Content	%	39.10 [0, 94.50]	24.90 [0.20, 100.00]
Sand Content	%	32.60 [0.10, 100.00]	33.05 [0, 99.50]
pH (H_2O)	–	6.37 [2.29, 10.70]	5.80 [3.00, 10.50]
CEC	$cmol_c\ kg^{-1}$	15.62 [0, 584.59]	11.50 [0, 189.60]
Exchangeable Ca	$cmol_c\ kg^{-1}$	11.92 [0, 410.41]	3.80 [0, 168.20]
Exchangeable Mg	$cmol_c\ kg^{-1}$	2.86 [0, 172.64]	1.10 [0, 68.00]
Exchangeable K	$cmol_c\ kg^{-1}$	0.364 [0, 32.33]	0.20 [0, 9.80]
Exchangeable Na	$cmol_c\ kg^{-1}$	0 [0, 868.36]	0.10 [0, 31.60]

was 30 psi, declustering potential was +80 V, and collision energy was 35 V.

Dataset metadata and label distributions for cross-instrument and cross-laboratory soil-property prediction

For cross-instrument and cross-laboratory soil-property prediction, the source- and target-domain dataset metadata and corresponding label distributions across the nine soil properties are summarized in Tables D and E, respectively.

Cross-validation and evaluation protocol

All downstream evaluations requiring supervised model training were conducted using five-fold cross-validation. For each such dataset, samples were partitioned into five non-overlapping folds, with each fold used once as the test set. In each cross-validation run, one fold, corresponding to approximately 20% of the dataset, was held out for testing. The remaining samples were divided into training and validation subsets to approximate an overall 70:10:20 train–validation–test ratio. When this ratio could not be achieved exactly because of discrete sample counts, the partition with the smallest deviation from the target ratio was used. For the molecule-level NIST, SDBS, and USPTO benchmarks, scaffold-based partitioning was used to minimize structural overlap among the training, validation, and test sets. Identical partitions were used for UltraIR, the matched no-pretraining ablation where included, and all comparison methods, ensuring directly comparable evaluations.

For the reduced-training-data experiments, the validation and test sets remained fixed within each cross-validation run. Only the original training subset was randomly subsampled to the specified fractions of labeled training data. These fractions therefore refer to the available training subset rather than the complete dataset. At each fraction, identical subsampled training sets were used for all compared methods.

Model configurations, training duration, early-stopping criteria, checkpoint selection, and the optional use of downstream spectral augmentation were determined exclusively using the training and validation subsets within each cross-validation run. Weak physically motivated augmentations, when enabled, were applied only to training samples. Except for the DeepMIR liquid-mixture dataset, downstream performance was aggregated across the five held-out

test folds. For DeepMIR, checkpoints trained on the corresponding NIST mixture-analysis tasks with matched target labels were applied directly for inference. No additional model training or five-fold cross-validation was performed on the DeepMIR dataset.

Evaluation metrics

Statistical significance testing.

All significance annotations shown in the figures were calculated independently for each dataset–metric combination. Methods were ranked according to their mean performance across the five held-out test folds, using the appropriate direction for each metric, and only the top-ranked and second-ranked methods were compared. Their fold-level results were paired by the corresponding test fold and analyzed using a two-sided paired t-test implemented with `scipy.stats.ttest_rel`. The null hypothesis was that the mean paired difference between the two methods across the five folds was zero. Significance was annotated as **** for $P < 0.0001$, *** for $0.0001 \leq P < 0.001$, ** for $0.001 \leq P < 0.01$, * for $0.01 \leq P < 0.05$, and n.s. for $P \geq 0.05$.

Classification metrics.

Single-label classification tasks, including bacterial classification, medicinal-herb geographic origin traceability, and microplastics classification, were evaluated using accuracy, Macro-F1, and the multiclass Matthews correlation coefficient (MCC). Targeted component detection was evaluated using accuracy, Macro-F1, and the area under the receiver operating characteristic curve (ROC-AUC), calculated from the continuous component-presence scores before thresholding. Macro-F1 was calculated as the unweighted mean of the one-versus-rest F1 scores across classes, thereby assigning equal weight to each class. For the class-wise radar analyses of bacterial and microplastics classification, the value reported for each bacterial genus or polymer class was its corresponding one-versus-rest F1 score. These class-specific F1 scores were calculated independently for each class and were distinct from the Macro-F1 values reported for overall task performance.

Functional-group prediction was formulated as multi-label classification and evaluated using Micro-F1, Macro-F1, and exact match ratio (EMR). Micro-F1 was calculated by aggregating true positives, false positives, and false negatives across all samples and functional-group labels, whereas Macro-F1 was the unweighted mean of the label-specific F1 scores. EMR required the complete predicted functional-group vector to match the ground-truth vector:

$$\mathrm{EMR} = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left(\mathbf{y}_i = \hat{\mathbf{y}}_i\right), \tag{26}$$

where N is the number of evaluated spectra, $\mathbf{y}_i$ and $\hat{\mathbf{y}}_i$ are the ground-truth and predicted binary label vectors, respectively, and $\mathbb{I}(\cdot)$ is the indicator function. A prediction containing either a missed functional group or an additional false-positive group was therefore counted as incorrect.

Molecular structure elucidation metrics.

Formula-conditioned molecular structure elucidation was evaluated using top-k accuracy for $k \in \{1, 5, 10\}$. Let s_i denote the ground-truth SMILES string for sample i, and let $\mathcal{P}_i^{(k)}$ denote the first k generated candidate SMILES strings in their original ranked order. Before structure matching, the ground-truth SMILES and every generated candidate were parsed using RDKit and converted to canonical isomeric SMILES. Denoting this canonicalization operation by $\mathcal{C}(\cdot)$, top-k accuracy was defined as

$$\text{Top-}k = \frac{1}{N}\sum_{i=1}^{N} \mathbb{I}\left[\exists\, \hat{s} \in \mathcal{P}_i^{(k)} \text{ such that } \mathcal{C}(\hat{s}) = \mathcal{C}(s_i)\right]. \tag{27}$$

A sample was counted as correct when at least one of the first k generated candidates represented the same canonical molecular structure as the ground truth. Candidates that could not be parsed into valid molecular structures remained in their generated rank positions but could not contribute a correct match. Stereochemical annotations were retained during canonicalization where present in the corresponding SMILES strings.

Structural similarity between a generated candidate and the ground-truth molecule was assessed using the Tanimoto

similarity of their binary molecular fingerprints. For fingerprint vectors $\mathbf{f}$ and $\mathbf{g}$, the Tanimoto similarity was defined as

$$T(\mathbf{f},\mathbf{g}) = \frac{\mathbf{f}^{\mathsf{T}}\mathbf{g}}{\|\mathbf{f}\|_1 + \|\mathbf{g}\|_1 - \mathbf{f}^{\mathsf{T}}\mathbf{g}}. \tag{28}$$

The score ranges from 0 to 1, with larger values indicating greater fingerprint overlap.

Regression metrics.

Regression tasks were evaluated using mean absolute error (MAE), root mean squared error (RMSE), and the coefficient of determination R^2. For single-target regression tasks, including targeted fractional contribution estimation, MAE, RMSE, and R^2 were calculated directly in the original target space.

For multi-target regression tasks, including physicochemical property prediction, mixture-level component quantification, medicinal-herb constituent quantification, and soil property prediction, normalized MAE and normalized RMSE were used to account for differences in numerical ranges among targets. These metrics were calculated after target-wise normalization and averaged across targets such that each target contributed equally. Overall R^2 was calculated as the arithmetic mean of target-specific R^2 values across targets.

For normalized residual analyses, residuals were calculated independently for each target and normalized by the corresponding target range. Let y_{ij} and $\hat{y}_{ij}$ denote the ground-truth and predicted values, respectively, for target j of sample i. The normalized residual was defined as

$$r'_{ij} = \frac{\hat{y}_{ij} - y_{ij}}{y_{j,\max} - y_{j,\min}}, \tag{29}$$

where $y_{j,\max}$ and $y_{j,\min}$ denote the maximum and minimum values of target j, respectively. To visualize multi-target regression residuals, the normalized residuals from all samples and targets were pooled such that each sample–target pair contributed equally to the resulting distribution.

BertzCT complexity-thresholded analysis.

Prediction errors across molecular complexity were assessed using cumulative BertzCT thresholds. For threshold τ, the evaluated subset was defined as

$$\mathcal{S}_\tau = \{i \colon B_i \geq \tau\}, \tag{30}$$

where B_i is the ground-truth BertzCT value of molecule i. The mean relative error (MRE) was then calculated as

$$\mathrm{MRE}(\tau) = \frac{1}{|\mathcal{S}_\tau|} \sum_{i \in \mathcal{S}_\tau} \frac{\left|\hat{B}_i - B_i\right|}{|B_i| + \epsilon}, \tag{31}$$

where ϵ is a small constant.

True-class margin analysis.

For microplastics classification, prediction confidence was further examined using the true-class margin. Given the logit s_{ic} assigned by a model to class c for sample i, with y_i denoting the ground-truth class label of sample i, the margin was defined as

$$m_i = s_{iy_i} - \max_{c \neq y_i} s_{ic}. \tag{32}$$

A positive margin indicates that the ground-truth class received the highest logit, whereas a negative margin indicates that at least one competing class received a higher score. Larger positive values indicate greater separation between the true class and its closest competing class.

B More results

Extended ablation analysis of UltraIR

Table F | Extended ablation analysis of simulated pretraining across representative downstream tasks. UltraIR is compared with the matched no-pretraining ablation across four downstream tasks. Values are means ± standard deviations across five test folds. Bold values indicate the better result for each metric. Arrows indicate the preferred direction.

a, Functional-group prediction *Results on the NIST dataset*			
Model	Macro-F1 ↑	Micro-F1 ↑	EMR ↑
UltraIR	$\mathbf{0.919}_{\pm 0.004}$	$\mathbf{0.943}_{\pm 0.002}$	$\mathbf{0.772}_{\pm 0.007}$
w/o pretraining	$0.901_{\pm 0.005}$	$0.932_{\pm 0.002}$	$0.719_{\pm 0.009}$
b, Molecular structure elucidation *Results on the NIST dataset*			
Model	Top-1 accuracy ↑	Top-5 accuracy ↑	Top-10 accuracy ↑
UltraIR	$\mathbf{0.522}_{\pm 0.005}$	$\mathbf{0.575}_{\pm 0.007}$	$\mathbf{0.576}_{\pm 0.007}$
w/o pretraining	$0.477_{\pm 0.003}$	$0.563_{\pm 0.006}$	$0.567_{\pm 0.007}$
c, Mixture-level component quantification			
Model	Normalized MAE ($\times 10^{-3}$) ↓	Normalized RMSE ($\times 10^{-3}$) ↓	R^2 ↑
UltraIR	$\mathbf{2.036}_{\pm 0.366}$	$\mathbf{2.621}_{\pm 0.502}$	$\mathbf{0.782}_{\pm 0.047}$
w/o pretraining	$2.288_{\pm 0.225}$	$2.919_{\pm 0.370}$	$0.714_{\pm 0.029}$
d, Soil property prediction			
Model	Normalized MAE ↓	Normalized RMSE ↓	R^2 ↑
UltraIR	$\mathbf{0.086}_{\pm 0.002}$	$\mathbf{0.325}_{\pm 0.147}$	$\mathbf{0.921}_{\pm 0.026}$
w/o pretraining	$0.104_{\pm 0.002}$	$0.360_{\pm 0.138}$	$0.901_{\pm 0.025}$

Additional functional-group prediction results

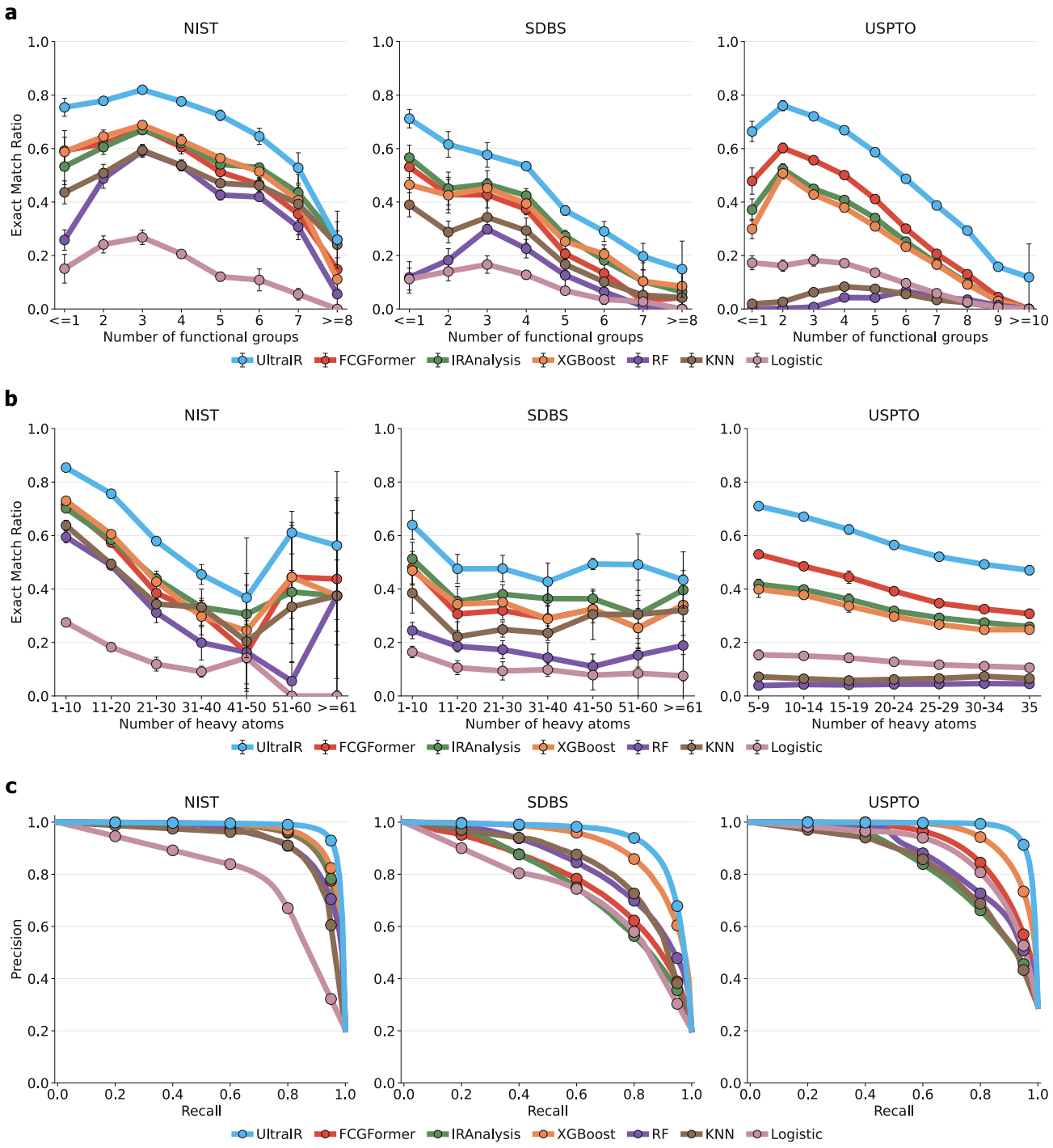

Figure A | Functional-group prediction across molecular-complexity strata and precision–recall trade-offs. a, Exact match ratio (EMR) for functional-group prediction stratified by the number of functional groups in each molecule on NIST, SDBS, and USPTO. **b,** EMR for functional-group prediction stratified by the number of heavy atoms in each molecule on NIST, SDBS, and USPTO. **c,** Precision–recall curves for functional-group prediction on NIST, SDBS, and USPTO.

Table G | Per-functional-group classification performance on the NIST benchmark. F1 scores are reported as means ± standard deviations across five folds. The best result for each functional group is highlighted in bold.

Functional group	Positive rate (%)	Logistic	KNN	RF	XGBoost	IRAnalysis	FCGFormer	UltraIR
Alkane	79.76	$0.919_{\pm 0.005}$	$0.931_{\pm 0.008}$	$0.952_{\pm 0.003}$	$0.966_{\pm 0.003}$	$0.960_{\pm 0.003}$	$0.962_{\pm 0.002}$	$\mathbf{0.978_{\pm 0.001}}$
Methyl	63.08	$0.785_{\pm 0.009}$	$0.865_{\pm 0.007}$	$0.877_{\pm 0.010}$	$0.920_{\pm 0.005}$	$0.910_{\pm 0.006}$	$0.911_{\pm 0.006}$	$\mathbf{0.950_{\pm 0.002}}$
Alkene	13.16	$0.449_{\pm 0.024}$	$0.684_{\pm 0.024}$	$0.689_{\pm 0.027}$	$0.783_{\pm 0.012}$	$0.761_{\pm 0.011}$	$0.731_{\pm 0.008}$	$\mathbf{0.869_{\pm 0.015}}$
Alkyne	2.04	$0.544_{\pm 0.047}$	$0.821_{\pm 0.031}$	$0.747_{\pm 0.057}$	$0.874_{\pm 0.023}$	$0.874_{\pm 0.015}$	$0.884_{\pm 0.035}$	$\mathbf{0.949_{\pm 0.022}}$
Alcohols	23.84	$0.682_{\pm 0.037}$	$0.827_{\pm 0.008}$	$0.850_{\pm 0.009}$	$0.887_{\pm 0.003}$	$0.889_{\pm 0.008}$	$0.889_{\pm 0.010}$	$\mathbf{0.928_{\pm 0.004}}$
Amines	22.74	$0.621_{\pm 0.044}$	$0.816_{\pm 0.005}$	$0.812_{\pm 0.017}$	$0.865_{\pm 0.011}$	$0.849_{\pm 0.005}$	$0.850_{\pm 0.005}$	$\mathbf{0.918_{\pm 0.005}}$
Nitriles	3.82	$0.291_{\pm 0.048}$	$0.528_{\pm 0.057}$	$0.466_{\pm 0.048}$	$0.795_{\pm 0.025}$	$0.659_{\pm 0.015}$	$0.586_{\pm 0.033}$	$\mathbf{0.931_{\pm 0.016}}$
Aromatics	58.84	$0.872_{\pm 0.016}$	$0.924_{\pm 0.008}$	$0.920_{\pm 0.011}$	$0.963_{\pm 0.004}$	$0.953_{\pm 0.009}$	$0.956_{\pm 0.007}$	$\mathbf{0.982_{\pm 0.005}}$
Alkyl halides	25.41	$0.542_{\pm 0.034}$	$0.764_{\pm 0.009}$	$0.750_{\pm 0.010}$	$0.810_{\pm 0.010}$	$0.800_{\pm 0.004}$	$0.791_{\pm 0.013}$	$\mathbf{0.889_{\pm 0.010}}$
Esters	11.78	$0.732_{\pm 0.021}$	$0.884_{\pm 0.014}$	$0.832_{\pm 0.029}$	$0.907_{\pm 0.013}$	$0.912_{\pm 0.013}$	$0.902_{\pm 0.019}$	$\mathbf{0.945_{\pm 0.010}}$
Ketones	8.96	$0.419_{\pm 0.045}$	$0.756_{\pm 0.023}$	$0.677_{\pm 0.026}$	$0.783_{\pm 0.029}$	$0.801_{\pm 0.011}$	$0.785_{\pm 0.022}$	$\mathbf{0.894_{\pm 0.014}}$
Aldehydes	1.97	$0.552_{\pm 0.050}$	$0.789_{\pm 0.034}$	$0.827_{\pm 0.025}$	$0.873_{\pm 0.021}$	$0.866_{\pm 0.035}$	$0.891_{\pm 0.022}$	$\mathbf{0.944_{\pm 0.016}}$
Carboxylic acids	6.85	$0.706_{\pm 0.017}$	$0.856_{\pm 0.014}$	$0.854_{\pm 0.017}$	$0.899_{\pm 0.016}$	$0.881_{\pm 0.009}$	$0.889_{\pm 0.012}$	$\mathbf{0.931_{\pm 0.015}}$
Ether	13.86	$0.580_{\pm 0.024}$	$0.784_{\pm 0.023}$	$0.717_{\pm 0.022}$	$0.816_{\pm 0.022}$	$0.818_{\pm 0.024}$	$0.818_{\pm 0.020}$	$\mathbf{0.905_{\pm 0.014}}$
Acyl halides	0.92	$0.559_{\pm 0.033}$	$0.827_{\pm 0.068}$	$0.781_{\pm 0.046}$	$0.843_{\pm 0.049}$	$0.864_{\pm 0.037}$	$0.861_{\pm 0.050}$	$\mathbf{0.897_{\pm 0.062}}$
Amides	6.19	$0.443_{\pm 0.040}$	$0.680_{\pm 0.026}$	$0.478_{\pm 0.052}$	$0.685_{\pm 0.021}$	$0.712_{\pm 0.010}$	$0.678_{\pm 0.016}$	$\mathbf{0.791_{\pm 0.014}}$
Nitro	5.63	$0.761_{\pm 0.018}$	$0.845_{\pm 0.026}$	$0.811_{\pm 0.031}$	$0.889_{\pm 0.027}$	$0.884_{\pm 0.017}$	$0.867_{\pm 0.019}$	$\mathbf{0.919_{\pm 0.006}}$

Table H | Per-functional-group classification performance on the SDBS benchmark. F1 scores are reported as means ± standard deviations across five folds. The best result for each functional group is highlighted in bold.

Functional group	Positive rate (%)	Logistic	KNN	RF	XGBoost	IRAnalysis	FCGFormer	UltraIR
Alkane	74.19	$0.858_{\pm 0.015}$	$0.875_{\pm 0.015}$	$0.866_{\pm 0.016}$	$0.915_{\pm 0.011}$	$0.723_{\pm 0.028}$	$0.633_{\pm 0.015}$	$\mathbf{0.941_{\pm 0.008}}$
Methyl	55.78	$0.695_{\pm 0.039}$	$0.767_{\pm 0.011}$	$0.776_{\pm 0.011}$	$0.852_{\pm 0.011}$	$0.053_{\pm 0.020}$	$0.052_{\pm 0.015}$	$\mathbf{0.881_{\pm 0.008}}$
Alkene	12.07	$0.401_{\pm 0.042}$	$0.470_{\pm 0.073}$	$0.301_{\pm 0.085}$	$0.565_{\pm 0.032}$	$0.107_{\pm 0.029}$	$0.187_{\pm 0.034}$	$\mathbf{0.699_{\pm 0.050}}$
Alkyne	1.07	$0.415_{\pm 0.089}$	$0.536_{\pm 0.104}$	$0.102_{\pm 0.071}$	$0.525_{\pm 0.067}$	$0.198_{\pm 0.087}$	$0.208_{\pm 0.060}$	$\mathbf{0.726_{\pm 0.061}}$
Alcohols	29.72	$0.631_{\pm 0.060}$	$0.800_{\pm 0.026}$	$0.773_{\pm 0.041}$	$0.844_{\pm 0.027}$	$0.226_{\pm 0.062}$	$0.254_{\pm 0.057}$	$\mathbf{0.893_{\pm 0.018}}$
Amines	27.35	$0.587_{\pm 0.105}$	$0.753_{\pm 0.008}$	$0.703_{\pm 0.026}$	$0.783_{\pm 0.015}$	$0.345_{\pm 0.061}$	$0.359_{\pm 0.069}$	$\mathbf{0.872_{\pm 0.017}}$
Nitriles	2.98	$0.586_{\pm 0.039}$	$0.278_{\pm 0.092}$	$0.152_{\pm 0.077}$	$0.800_{\pm 0.040}$	$0.297_{\pm 0.047}$	$0.338_{\pm 0.082}$	$\mathbf{0.876_{\pm 0.025}}$
Aromatics	60.47	$0.831_{\pm 0.017}$	$0.865_{\pm 0.009}$	$0.855_{\pm 0.050}$	$0.935_{\pm 0.014}$	$0.761_{\pm 0.049}$	$0.741_{\pm 0.056}$	$\mathbf{0.968_{\pm 0.004}}$
Alkyl halides	18.62	$0.363_{\pm 0.035}$	$0.429_{\pm 0.055}$	$0.296_{\pm 0.045}$	$0.497_{\pm 0.030}$	$0.187_{\pm 0.027}$	$0.195_{\pm 0.034}$	$\mathbf{0.607_{\pm 0.035}}$
Esters	11.95	$0.735_{\pm 0.009}$	$0.799_{\pm 0.017}$	$0.719_{\pm 0.024}$	$0.836_{\pm 0.017}$	$0.130_{\pm 0.053}$	$0.208_{\pm 0.060}$	$\mathbf{0.884_{\pm 0.018}}$
Ketones	9.24	$0.384_{\pm 0.051}$	$0.545_{\pm 0.040}$	$0.226_{\pm 0.106}$	$0.569_{\pm 0.036}$	$0.034_{\pm 0.010}$	$0.101_{\pm 0.027}$	$\mathbf{0.720_{\pm 0.031}}$
Aldehydes	2.01	$0.251_{\pm 0.060}$	$0.407_{\pm 0.075}$	$0.175_{\pm 0.074}$	$0.446_{\pm 0.082}$	$0.083_{\pm 0.097}$	$0.176_{\pm 0.113}$	$\mathbf{0.647_{\pm 0.073}}$
Carboxylic acids	12.17	$0.677_{\pm 0.017}$	$0.791_{\pm 0.043}$	$0.785_{\pm 0.039}$	$0.844_{\pm 0.036}$	$0.302_{\pm 0.065}$	$0.312_{\pm 0.058}$	$\mathbf{0.882_{\pm 0.030}}$
Ether	14.63	$0.518_{\pm 0.033}$	$0.615_{\pm 0.031}$	$0.368_{\pm 0.092}$	$0.673_{\pm 0.030}$	$0.044_{\pm 0.029}$	$0.129_{\pm 0.033}$	$\mathbf{0.784_{\pm 0.021}}$
Acyl halides	0.99	$0.606_{\pm 0.108}$	$\mathbf{0.728_{\pm 0.126}}$	$0.560_{\pm 0.157}$	$0.692_{\pm 0.119}$	$0.310_{\pm 0.125}$	$0.251_{\pm 0.146}$	$0.719_{\pm 0.120}$
Amides	7.95	$0.468_{\pm 0.045}$	$0.612_{\pm 0.036}$	$0.279_{\pm 0.111}$	$0.583_{\pm 0.055}$	$0.175_{\pm 0.050}$	$0.168_{\pm 0.047}$	$\mathbf{0.746_{\pm 0.030}}$
Nitro	6.17	$0.685_{\pm 0.076}$	$0.699_{\pm 0.066}$	$0.606_{\pm 0.067}$	$0.794_{\pm 0.031}$	$0.253_{\pm 0.097}$	$0.139_{\pm 0.104}$	$\mathbf{0.871_{\pm 0.040}}$

Table I | Per-functional-group classification performance on the USPTO benchmark. F1 scores are reported as means ± standard deviations across five folds. The best result for each functional group is highlighted in bold.

Functional group	Positive rate (%)	Logistic	KNN	RF	XGBoost	IRAnalysis	FCGFormer	UltraIR
Alkane	93.15	$0.989_{\pm 0.001}$	$0.964_{\pm 0.001}$	$0.965_{\pm 0.001}$	$0.994_{\pm 0.000}$	$0.856_{\pm 0.009}$	$0.848_{\pm 0.007}$	$\mathbf{0.999_{\pm 0.000}}$
Methyl	73.10	$0.914_{\pm 0.004}$	$0.839_{\pm 0.004}$	$0.851_{\pm 0.004}$	$0.964_{\pm 0.001}$	$0.017_{\pm 0.007}$	$0.013_{\pm 0.006}$	$\mathbf{0.981_{\pm 0.001}}$
Alkene	11.34	$0.160_{\pm 0.040}$	$0.204_{\pm 0.006}$	$0.000_{\pm 0.000}$	$0.240_{\pm 0.012}$	$0.000_{\pm 0.001}$	$0.001_{\pm 0.001}$	$\mathbf{0.603_{\pm 0.017}}$
Alkyne	1.91	$0.301_{\pm 0.022}$	$0.165_{\pm 0.014}$	$0.000_{\pm 0.000}$	$0.653_{\pm 0.035}$	$0.233_{\pm 0.040}$	$0.291_{\pm 0.019}$	$\mathbf{0.929_{\pm 0.013}}$
Alcohols	26.22	$0.636_{\pm 0.018}$	$0.467_{\pm 0.012}$	$0.634_{\pm 0.009}$	$0.840_{\pm 0.005}$	$0.032_{\pm 0.003}$	$0.137_{\pm 0.027}$	$\mathbf{0.953_{\pm 0.002}}$
Amines	44.48	$0.682_{\pm 0.022}$	$0.613_{\pm 0.003}$	$0.659_{\pm 0.008}$	$0.806_{\pm 0.004}$	$0.069_{\pm 0.002}$	$0.195_{\pm 0.020}$	$\mathbf{0.917_{\pm 0.003}}$
Nitriles	6.69	$0.426_{\pm 0.031}$	$0.321_{\pm 0.015}$	$0.001_{\pm 0.001}$	$0.813_{\pm 0.012}$	$0.158_{\pm 0.020}$	$0.135_{\pm 0.041}$	$\mathbf{0.945_{\pm 0.006}}$
Aromatics	91.42	$0.976_{\pm 0.004}$	$0.954_{\pm 0.007}$	$0.955_{\pm 0.013}$	$0.983_{\pm 0.005}$	$0.368_{\pm 0.034}$	$0.266_{\pm 0.012}$	$\mathbf{0.993_{\pm 0.001}}$
Alkyl halides	46.50	$0.653_{\pm 0.017}$	$0.600_{\pm 0.002}$	$0.661_{\pm 0.006}$	$0.731_{\pm 0.005}$	$0.051_{\pm 0.006}$	$0.021_{\pm 0.008}$	$\mathbf{0.809_{\pm 0.004}}$
Esters	18.23	$0.650_{\pm 0.025}$	$0.551_{\pm 0.024}$	$0.312_{\pm 0.012}$	$0.790_{\pm 0.016}$	$0.031_{\pm 0.007}$	$0.016_{\pm 0.010}$	$\mathbf{0.925_{\pm 0.006}}$
Ketones	8.22	$0.319_{\pm 0.041}$	$0.268_{\pm 0.012}$	$0.002_{\pm 0.001}$	$0.433_{\pm 0.043}$	$0.004_{\pm 0.002}$	$0.003_{\pm 0.002}$	$\mathbf{0.692_{\pm 0.027}}$
Aldehydes	2.67	$0.653_{\pm 0.037}$	$0.426_{\pm 0.025}$	$0.117_{\pm 0.020}$	$0.956_{\pm 0.009}$	$0.483_{\pm 0.013}$	$0.798_{\pm 0.035}$	$\mathbf{0.988_{\pm 0.003}}$
Carboxylic acids	9.73	$0.806_{\pm 0.014}$	$0.524_{\pm 0.019}$	$0.739_{\pm 0.021}$	$0.947_{\pm 0.005}$	$0.461_{\pm 0.024}$	$0.664_{\pm 0.017}$	$\mathbf{0.985_{\pm 0.001}}$
Ether	37.05	$0.656_{\pm 0.023}$	$0.571_{\pm 0.015}$	$0.627_{\pm 0.013}$	$0.785_{\pm 0.008}$	$0.019_{\pm 0.004}$	$0.014_{\pm 0.008}$	$\mathbf{0.905_{\pm 0.003}}$
Acyl halides	0.59	$0.132_{\pm 0.030}$	$0.277_{\pm 0.044}$	$0.000_{\pm 0.000}$	$0.205_{\pm 0.029}$	$0.054_{\pm 0.016}$	$0.092_{\pm 0.044}$	$\mathbf{0.719_{\pm 0.026}}$
Amides	25.61	$0.657_{\pm 0.009}$	$0.620_{\pm 0.003}$	$0.558_{\pm 0.015}$	$0.764_{\pm 0.004}$	$0.028_{\pm 0.001}$	$0.019_{\pm 0.006}$	$\mathbf{0.899_{\pm 0.003}}$
Nitro	5.27	$0.346_{\pm 0.036}$	$0.190_{\pm 0.017}$	$0.000_{\pm 0.000}$	$0.518_{\pm 0.023}$	$0.049_{\pm 0.010}$	$0.031_{\pm 0.009}$	$\mathbf{0.814_{\pm 0.014}}$

Additional molecular structure elucidation results

Table J | Ablation of IR and molecular-formula conditioning for molecular structure elucidation on the NIST benchmark. Tanimoto@k denotes the maximum Morgan Tanimoto similarity between the ground-truth structure and the first k candidates. Results are reported as means ± standard deviations across five folds.

Model	Top-1		Top-5		Top-10	
	Acc (%) ↑	Tanimoto ↑	Acc (%) ↑	Tanimoto ↑	Acc (%) ↑	Tanimoto ↑
UltraIR w/o IR	$9.71_{\pm 0.21}$	$0.293_{\pm 0.002}$	$25.54_{\pm 0.48}$	$0.476_{\pm 0.003}$	$33.17_{\pm 0.43}$	$0.548_{\pm 0.003}$
UltraIR w/o Formula	$38.94_{\pm 0.26}$	$0.594_{\pm 0.003}$	$46.66_{\pm 0.56}$	$0.676_{\pm 0.005}$	$49.45_{\pm 0.46}$	$0.702_{\pm 0.005}$
UltraIR	$\mathbf{52.20_{\pm 0.54}}$	$\mathbf{0.682_{\pm 0.004}}$	$\mathbf{57.48_{\pm 0.71}}$	$\mathbf{0.742_{\pm 0.003}}$	$\mathbf{57.57_{\pm 0.74}}$	$\mathbf{0.748_{\pm 0.003}}$

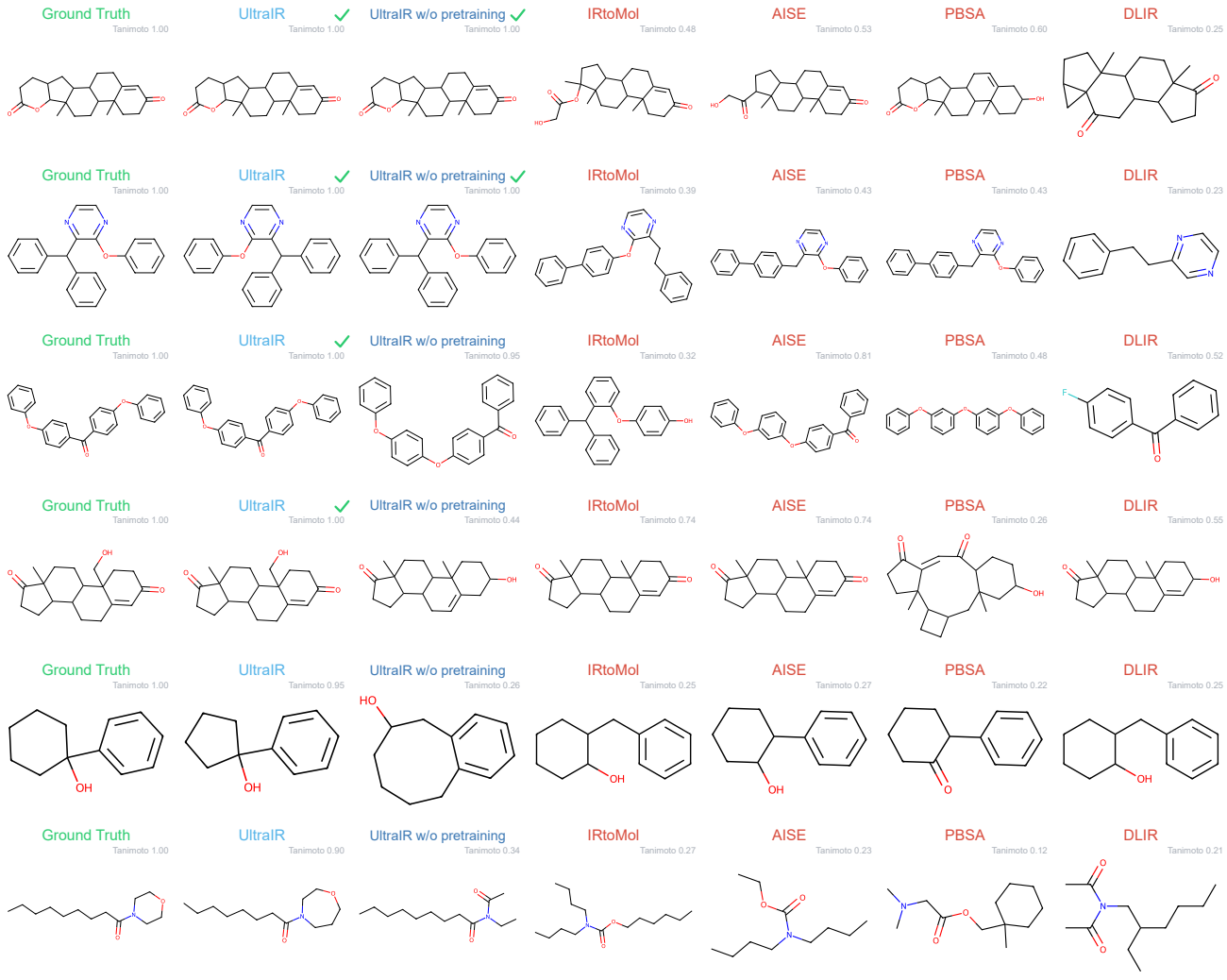

Figure B | Additional molecular structure elucidation examples. Six representative examples from the NIST test set comparing the ground-truth molecular structures with candidates generated by UltraIR, UltraIR without pretraining, and the competing structure-elucidation models IRtoMol, AISE, PBSA, and DLIR. Fingerprint-based Tanimoto similarities between each generated candidate and the corresponding ground-truth structure are reported. Green check marks indicate exact structure recovery.

Additional physicochemical property prediction results

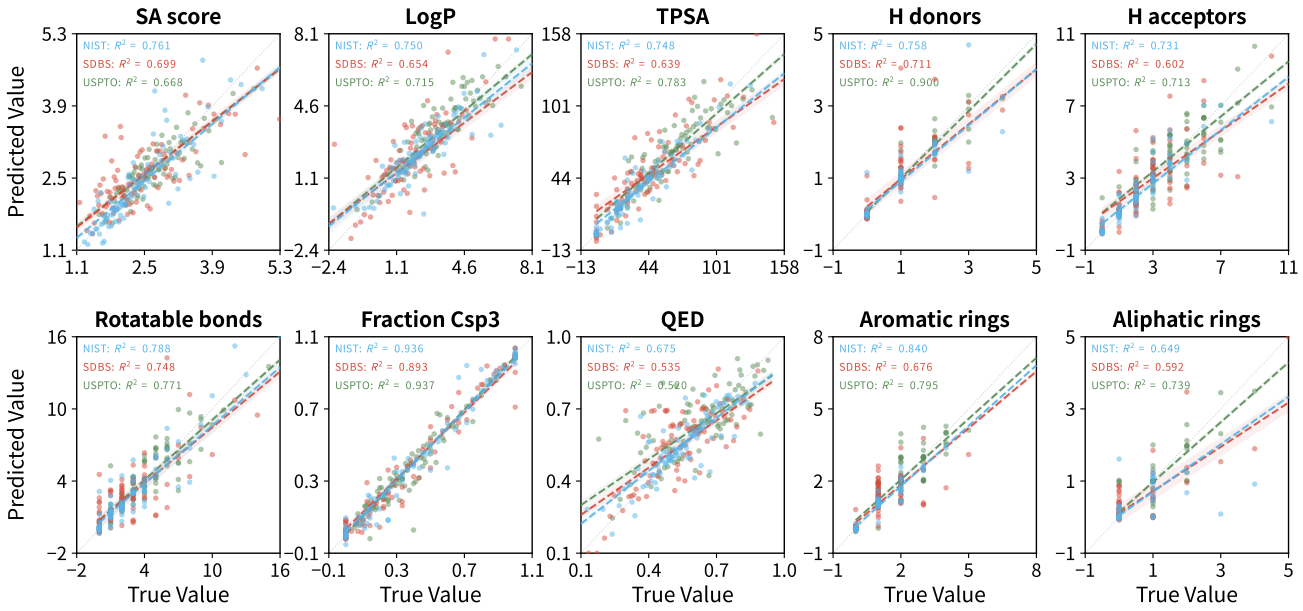

Figure C | Property-wise physicochemical prediction performance of UltraIR across NIST, SDBS, and USPTO. Predicted-versus-true parity plots are shown for synthetic accessibility (SA) score, log P, topological polar surface area (TPSA), hydrogen-bond donors, hydrogen-bond acceptors, rotatable bonds, the fraction of sp^3-hybridized carbon atoms (Fraction Csp3), quantitative estimate of drug-likeness (QED), aromatic rings, and aliphatic rings. The identity line denotes ideal agreement.

Additional bacterial classification results

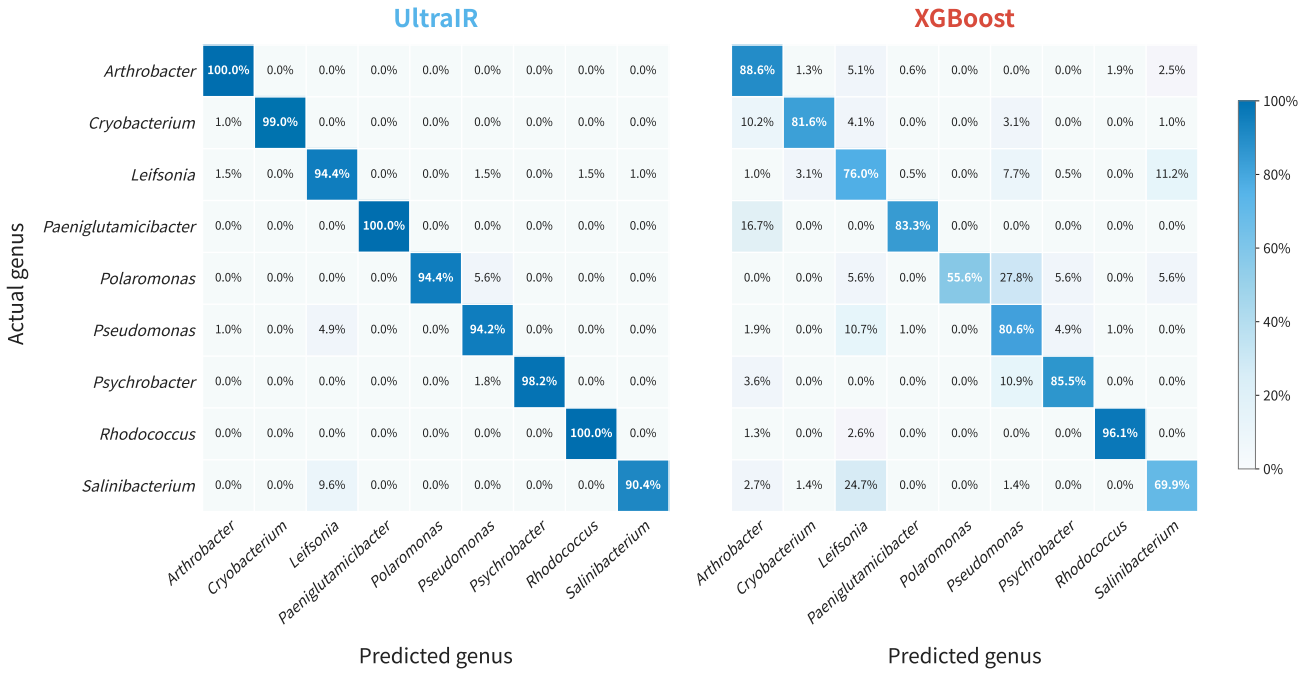

Figure D | Confusion matrices for genus-level bacterial classification. Confusion matrices are shown for UltraIR and XGBoost across the nine evaluated bacterial genera. Rows denote actual genera and columns denote predicted genera, with each cell indicating the percentage of samples assigned to the corresponding predicted genus.

Additional medicinal-herb geographic origin traceability results

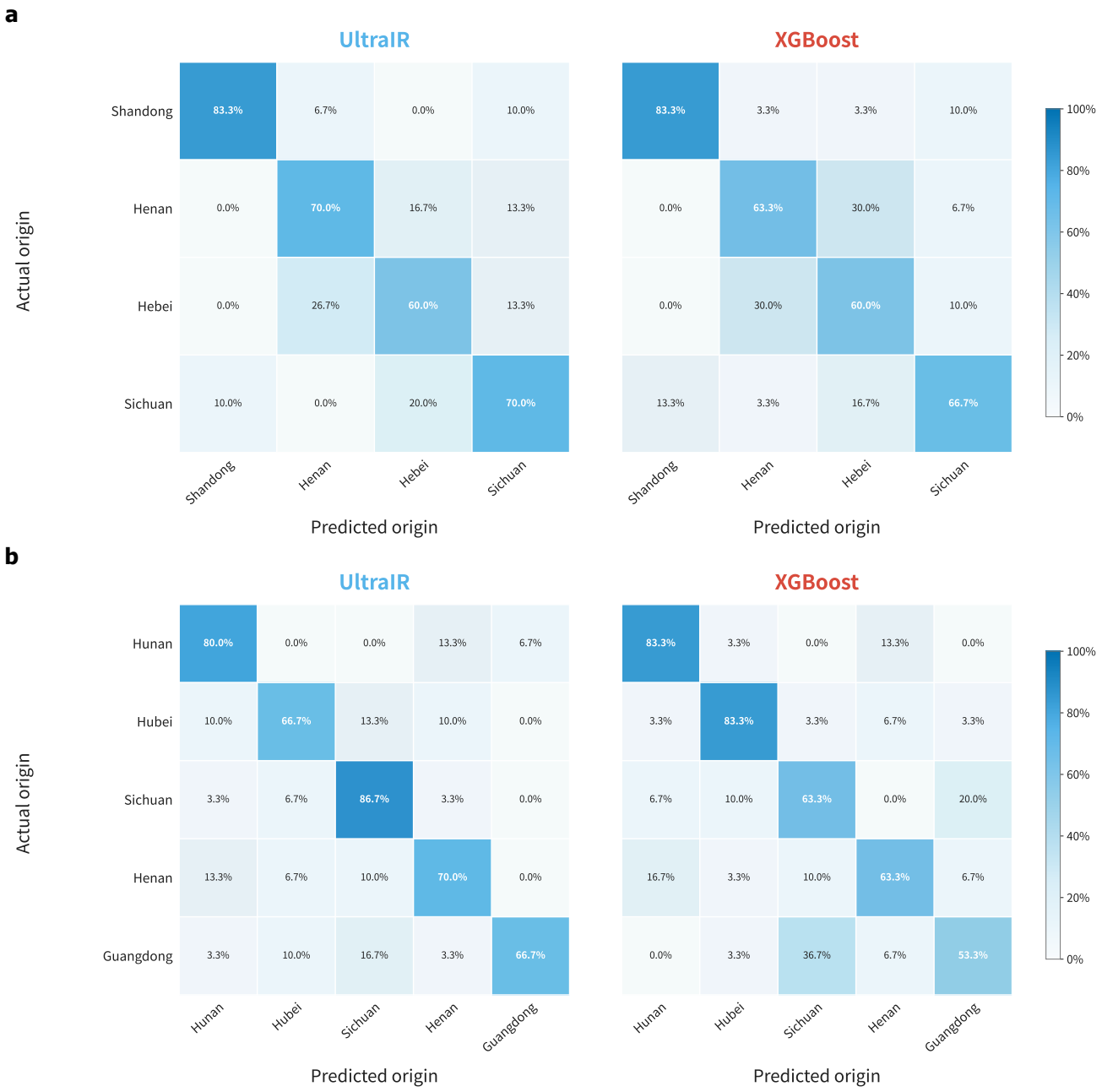

Figure E | Confusion matrices for medicinal-herb geographic origin traceability. a, Confusion matrices for Jinyinhua geographic origin traceability, comparing UltraIR and XGBoost across Shandong, Henan, Hebei, and Sichuan. **b,** Confusion matrices for Shanyinhua geographic origin traceability, comparing UltraIR and XGBoost across Hunan, Hubei, Sichuan, Henan, and Guangdong. Rows denote actual origins and columns denote predicted origins, with each cell indicating the percentage of samples assigned to the corresponding predicted origin.

Additional microplastics classification results

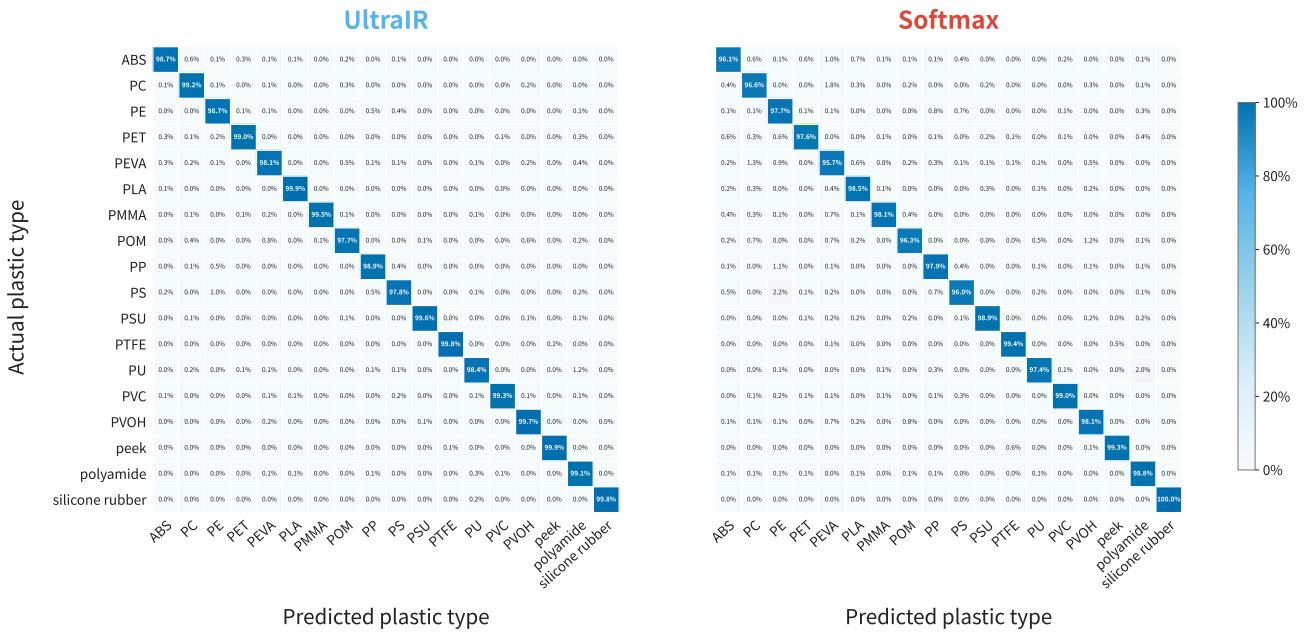

Figure F | Confusion matrices for microplastics classification. Confusion matrices are shown for UltraIR and Softmax across the 18 evaluated polymer classes. Rows denote actual plastic types and columns denote predicted plastic types, with each cell indicating the percentage of samples assigned to the corresponding predicted class.

Additional cross-instrument and cross-laboratory generalization results

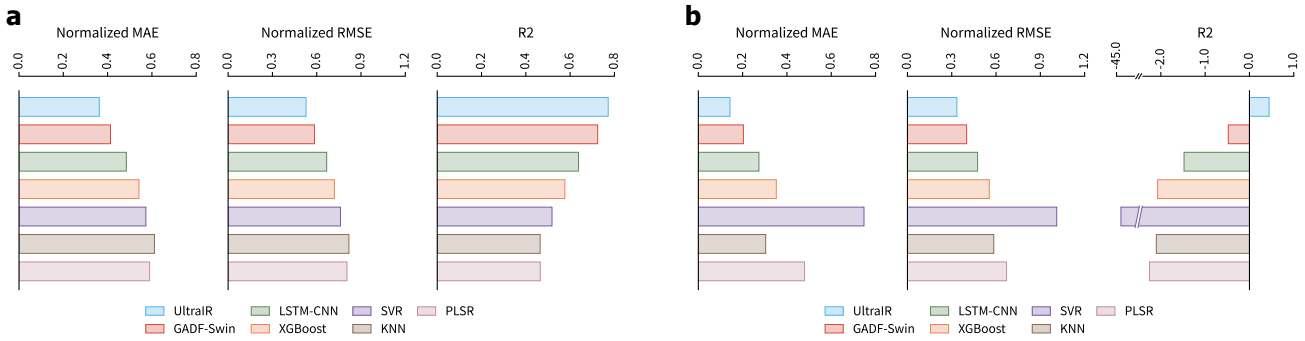

Figure G | Cross-instrument and cross-laboratory generalization for soil property prediction. All models were trained on the Kellogg Soil Survey Laboratory (KSSL) subset of OSSL and evaluated by zero-shot inference on the ICRAF–ISRIC subset. **a,** Prediction performance for soil texture and acidity properties, including clay content, silt content, sand content, and pH (H_2O), evaluated using Normalized MAE, Normalized RMSE, and R^2. **b,** Prediction performance for soil chemical properties, including cation exchange capacity (CEC) and exchangeable Ca, Mg, K, and Na, evaluated using the same metrics.