A Machine-Learned "Chemical Intuition" to Overcome Spectroscopic Data Scarcity

Cailum M. K. Stienstra,[a] Teun van Wieringen,[b] Liam Hebert,[c] Patrick Thomas,[a] Kas J. Houthuijs[b], Giel Berden[b], Jos Oomens[b], Jonathan Martens,[b*] and W. Scott Hopkins[a,d,e*]

[a] *Department of Chemistry, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada,*

[b] *Radboud University, Institute for Molecules and Materials, FELIX Laboratory, 6525 ED, Nijmegen, The Netherlands;*

[c] *Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada*

[d] *WaterFEL Free Electron Laser Laboratory, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada*

[e] *Centre for Eye and Vision Research, Hong Kong Science Park, New Territories, 999077, Hong Kong.*

**Corresponding Authors*

ABSTRACT: Machine learning models for predicting IR spectra of molecular ions (infrared ion spectroscopy, IRIS) have yet to be reported owing to the relatively sparse experimental datasets available. To overcome this limitation, we employ the Graphormer-IR model for neutral molecules as a knowledgeable starting point, then employ transfer learning to refine the model to predict the spectra of gaseous ions. A library of 10,336 computed spectra and a small dataset of 312 experimental IRIS spectra is used for model fine-tuning. Nonspecific global graph encodings that describe the molecular charge state (*i.e.*, (de)protonation, sodiation), combined with an additional transfer learning step that considers computed spectra for ions, improved model performance. The resulting Graphormer-IRIS model yields spectra that are 21% more accurate than those produced by commonly employed DFT quantum chemical models, while capturing subtle phenomena such as spectral red-shifts due to sodiation. Dimensionality reduction of model embeddings demonstrate derived "chemical intuition" of functional groups, trends in molecular electron density, and the location of charge sites. Our approach will enable fast IRIS predictions for determining the structures of unknown small molecule analytes (*e.g.*, metabolites, lipids) present in biological samples.

Introduction

Mass spectrometry (MS) is an invaluable tool in chemical analysis with application in areas such as metabolomics, environmental monitoring, and forensics.[1–3] Traditionally, targeted MS analyses have been employed for analyte detection and quantification, but in recent years the development of non-targeted approaches has become a priority for the community.[4,5] In a typical non-targeted MS workflow, signals are labelled with, *e.g.*, exact ion mass, isotope patterns, tandem MS fragmentation patterns (*i.e.*, MS^n), and liquid chromatographic retention times (or a subset of these parameters), and analytes are annotated based on matches to a library or agreement with the predictions of a model.[6,7] For unambiguous assignment, the measured parameters must match those of an internal standard.[8,9] This approach, of course, is limited by the availability of chemical standards and the chemical coverage of MS libraries.[10,11] Consequently, novel compounds such as previously uncharacterized natural products and "dark metabolites" typically cannot be identified using these routine approaches.[12,13] Further to this, the mechanisms of fragmentation (*e.g.*, cyclization, neutral losses) observed via MS^n methods are often unclear and it can be challenging to distinguish closely related isomeric species from one another.[14] As a result, traditional mass spectrometric measurements (*viz.* MS^n methods) are often insufficient for complete experimental characterization of molecular structure. This limitation is reflected in the performance of top *de novo* methods, which identify the correct molecular structure as the best candidate less than 30% of the time.[13,15]

Experimentalists have turned to orthogonal techniques to improve structural elucidation of small molecules. A key challenge in this regard is the separation and isolation of analytes from complex mixtures. For example, traditional benchtop Fourier transform infrared (FT-IR) spectroscopy measures simultaneously the IR spectrum for all components in a mixture, which results in spectra that are the convolution of the components in the mixture, often leading to non-detection of low-abundance species.[16] By coupling MS with infrared ion spectroscopy (IRIS), one can overcome this hurdle by mass-selecting analytes of interest prior to spectroscopic interrogation.[16,17] Selected ions that absorb IR photons can be internally heated to induce photofragmentation, thereby providing a measurable change in the MS signal.[17] Owing to the sensitivity of the MS detector, this form of action spectroscopy overcomes the limitation of absorption measurements for which the number density of the absorbing analyte must be several orders of magnitude higher than is common in a mass spectrometer.[18] By monitoring ion fragmentation efficiency as a function of IR excitation wavelength, one can generate a vibrational spectrum wherein the vibrational frequencies may be directly related to the molecular structure .

IRIS has been employed successfully to characterize numerous interesting chemical systems,[16,19–22] and researchers continue to actively advance its application in new areas. For example, experimental and computed IR spectra have shown promise as discriminators in self-driving labs for probing chemical reactivity; this could be extended to incorporating IRIS in high-throughput MS screening workflows.[23,24] As another example, Martens *et al.* demonstrated the use of IRIS to characterize individual components of complex samples as they are separated using hybrid liquid chromatographic (LC) MS workflows.[22,25–27] By comparing the measured "fingerprint" spectra to calculated reference spectra (typically computed using electronic structure theory methods),[28] researchers can precisely differentiate chemically similar moieties, including isomeric species, without the need for chemical

standards.[19,22] IRIS workflows have been applied in lipidomics,[1,29,30] proteomics,[31–33] metabolomics,[1,21,28,34,35] and small molecule applications;[19,20,22] for example, IRIS has shown utility in characterizing sugars,[36,37] polymers,[38] and environmental pollutants.[39–42] Maitre *et al.* have also described the use of IRIS for the structural characterization of post-translational modifications of proteins (*e.g.*, phosphorylation, glycosylation), which often initiate signaling processes for downstream physiological function.[33]

The success of IRIS as a tool for improved structural annotation of unknown analytes depends strongly on the quality and speed of the computational tool used to generate reference spectra. The conventional approach to computing reference spectra involves quantum chemical calculation of harmonic vibrational frequencies and integrated intensities followed by empirical scaling to correct for anharmonicity of vibrational potentials.[28] One can improve harmonic frequency predictions by calculating anharmonic corrections, but the cost of these calculations are much higher than those using the harmonic approximation and they scale rapidly with molecular size, requiring substantial time and computational resources for even moderately sized molecules. Moreover, quantum chemical methods compute *absorption* spectra, which differ subtly from IRIS spectra, where intensities are associated with IR absorption cross section, as well as the efficiency of anharmonic coupling between normal modes (leading to intramolecular vibrational energy redistribution; IVR) and coupling to dissociation thresholds.[17]

By enabling a learned understanding of peak position and mode-coupling, deep learning might provide a solution to the challenges of quickly and accurately predicting IRIS spectra. Further, trained machine learning (ML) models can produce predicted IRIS spectra for thousands of molecules in seconds – much faster than can be achieved by quantum chemical methods for all but the smallest molecules. To date, ML models of IR absorption spectra have been developed using Message Passing Neural Networks (MPNNs),[43] Graph Attention Networks (GATs),[44] graph transformers,[45] and other frameworks.[46–50] The success of these models is in part due to the large training datasets of absorption IR spectra (tens of thousands) available in online repositories.[51–53] However, no large compendium of IRIS spectra for model training exists. Consequently, to the best of our knowledge, a predictive model for IRIS spectroscopy is yet to be reported.

In this work, we describe a ML model for predicting IRIS spectra from chemical structure. This model is based on the Graphormer architecture, which extends graph neural networks (GNNs) via transformers.[54,55] Graphormer has already been applied successfully to chemical systems, having won the 2021 Open Catalyst challenge and achieving state-of-the-art predictions across a variety of (bio)chemical domains.[45,56–59] Using a multi-staged transfer learning scheme inspired by the psychology of human learning and natural language processing (NLP), we show how different spectral libraries offer distinct contributions (*e.g.*, anharmonic scaling, denoising) to the final model's understanding of experimental IRIS measurements. Further, we utilize a flexible description of molecular charge (*i.e.*, (de)protonation, sodiation) and explainability techniques to show that Graphormer-IRIS successfully derives "chemical intuition" from transfer learning and a highly contextual understanding of molecular charge and bulk thermodynamic behavior.

Results and Discussion

Model Performance. Using the datasets described in reference 28 (see Figures S1-S3), models (see Figure S4) were trained using ten-fold cross validation to determine the optimal transfer learning strategy (see Figure 1 and Table S1), encodings (see Tables S1 and S2), and hyperparameters (see Table S3). The best performing model obtained a spectral similarity score of $SIS_{\mu} = 0.6823 \pm 0.0343$ (ten-fold cross validation, $c = 10$, Table 1, Ablation #10).[43] Histograms showing performance for all test splits ($c = 10$) are depicted in Figure 2. Detailed results showing all training strategies are shown in Table 1 and are discussed in more detail below. For 195 molecular ions, DFT-computed and experimental IRIS spectra were available; we calculated the SIS score for the DFT-computed spectra using the same pre-processing method as was used for our ML model. This DFT/IRIS comparison yielded $SIS_{\mu} = 0.5616$. Although the DFT/IRIS comparison is not perfect due to differences in the composition (*e.g.* molecular, charge state) and the number of spectra in the two evaluation sets, this benchmark suggests that Graphormer-IRIS (using only a small training dataset) is approximately 21% more accurate than conventional quantum-chemical methods for predicting IRIS spectra. Graphormer-IRIS predictions are also completed in seconds on GPUs, in contrast to the hours-to-days timescale for common DFT workflows (see Supporting Information*:Model Training*).[28]

The best performing transfer learning strategy started with a pretrained Graphormer-IR model,[45] followed by finetuning on DFT-computed spectra, and finally training on the experimental IRIS spectra. For DFT and IRIS finetuning, the first layer of the graph encoder was frozen. This model utilized the novel phase encodings for the global charge node described in the *Supporting Information: Charge Encoding* section for the IRIS and DFT spectra. As shown in Figure 2, the best performing model yielded $SIS_{\mu} = 0.6823 \pm 0.0343$ ($c = 10$) for spectra of all species, $SIS_{\mu} = 0.6683$ ($n = 164$) for spectra of protonated species, $SIS_{\mu} = 0.6782$ ($n = 80$) for spectra of deprotonated species, and $SIS_{\mu} = 0.7209$ ($n = 68$) for spectra of sodiated species.

Figure 3 compares the predictions for nine molecules in the test split with their respective experimental spectra – three for each charge state. Figures 3A-C show predictions that are well below the average SIS, Figures 3D-F show predictions of average accuracy, and Figures 3G-I show predictions with above average SIS scores. In general, Graphormer-IRIS performs best for organic molecules that have common organic functional groups/structural motifs (*e.g.*, carboxylic acids, ethers).

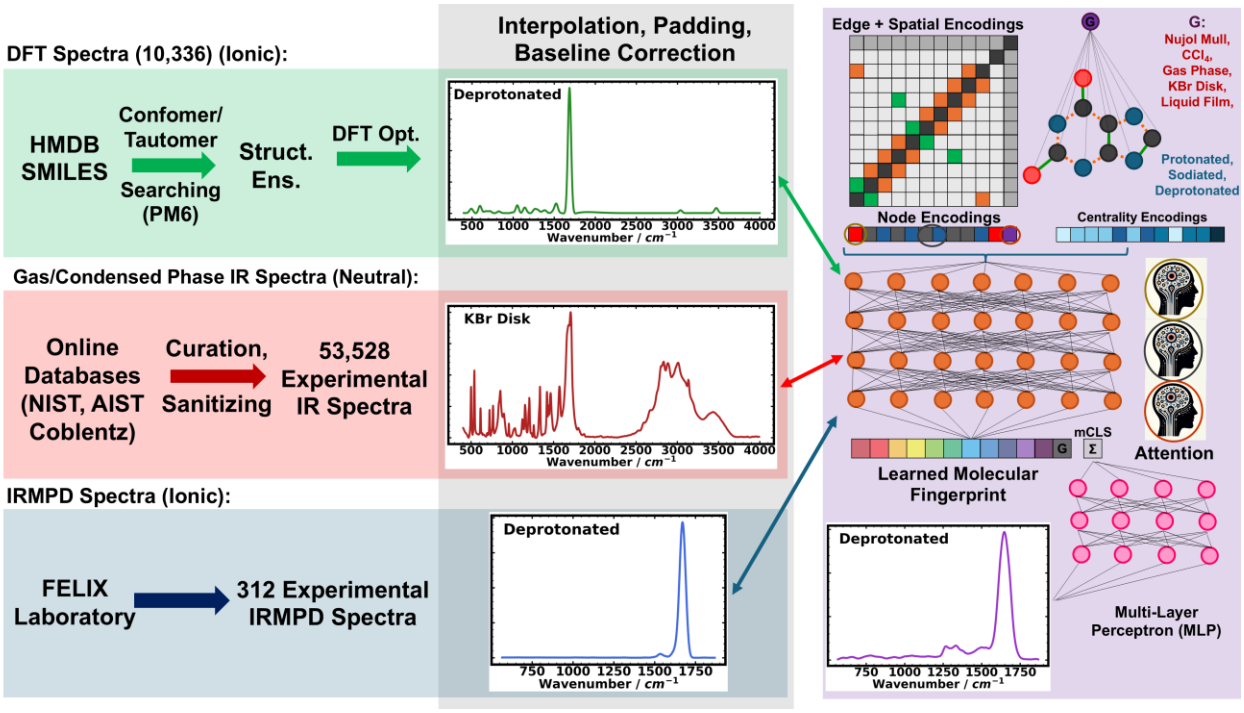

Figure 1. Schematic showing the datasets, transfer learning workflow, and model architecture utilized in this study. The plotted spectra are associated with *xanthine*, which is depicted as a graph structure in the top right.

However, given the relatively sparse chemical diversity of the IRIS dataset, we cannot make any comprehensive claims regarding broad molecular generalizability. We also acknowledge that a limitation of the small experimental dataset is that some species contain similar scaffolds (*e.g.*, the agrochemical depicted in Figure 3H). In future work, splitting via a Murcko scaffold may be useful to assess the extent that this structural similarity impacts model performance.[60] The below average IRIS spectral predictions shown in Figure 3A-C highlight that the underrepresentation of functional groups in the IRIS dataset impacts model predictions even if similar features are well represented in pretraining datasets (*e.g.*, IR, DFT). This effect is most obvious in the IRIS spectra of 4-bromo aniline(Figure 3C) and (8-chloro-6-(2-fluorophenyl)-4H-benzo[f]imidazo[1,5-a][1,4]diazepin-1-yl)methanol (Figure 3D), where Graphormer-IRIS predicts poorly the C-Br, C-Cl, and C-F stretching frequencies at *ca.* 600–800 cm^{-1}. Although halogenated species with similar stretching frequencies are represented in the experimental IR absorption database that was used for pretraining, these frequencies are poorly represented and predicted in the IRIS dataset. These prediction errors indicate that the IRIS dataset would benefit from expansion to improve the representation and diversity of chemical moieties.

Charge State Encodings. We introduce a global charge node to all molecular graphs for IRIS spectra (see *Supporting Information: Charge Encodings*) so that Graphormer-IRIS could obtain a contextual understanding of molecular charge. While training on IRIS spectra without a description of charge, Graphormer-IRIS yields SIS_μ = 0.4749 ± 0.0228 for predictions of IRIS spectra (Table 1; row #2). Upon transfer learning using the IRIS dataset but signifying that the molecules are in the gas phase using the global node (and the encoding from the Graphormer-IR study),[45] scores improve to SIS_μ = 0.5491 ± 0.0329 (Table 1; row #3). Introducing an explicit global node description of molecular charge state yielded SIS_μ = 0.6259 ± 0.0400 (Table 1; row #4), which is ~50% improved over the naïve Graphormer-IR evaluation.

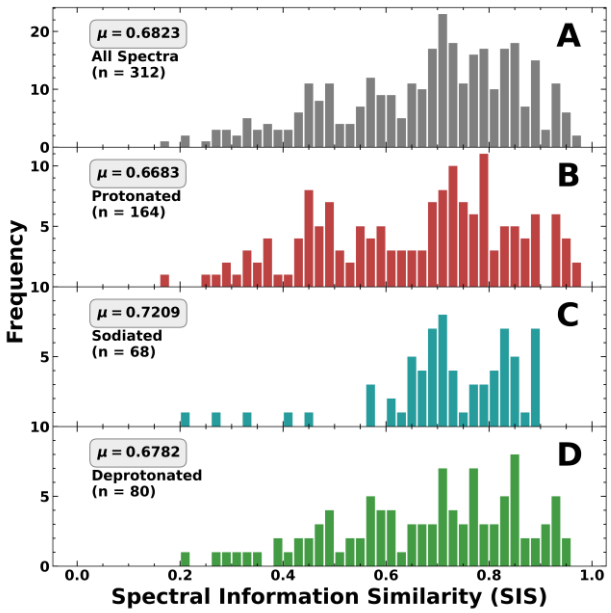

Figure 2. Graphormer-IRIS performance for ten test folds (c = 10) using the best performing model (SIS_μ = 0.6823 ± 0.0343). SIS distribution for (**A**) all IRIS spectra, (**B**) protonated molecules, (**C**) sodiated molecules, and (**D**) deprotonated molecules. Associated mean SIS scores (μ) and standard deviations (σ) for the distributions are reported in each panel. The variable *n* indicates the number of spectra of that type included in the IRIS spectra library.

Table 1. Results of the ablation study for the Graphormer-IRIS models. A checkmark indicates that the feature is present in the trained model. Uncertainty is described by the standard deviation of the 10-fold cross validated results (*i.e.*, $c = 10$).

#	Train on IRIS Spectra	Train on IR Spectra	Charge Graph Encodings	Train on DFT Spectra	Freeze Graph Encoder	Freeze Feature Encoder	Test SIS_{μ} ($c = 10$)
1.		✓					0.4219 ± 0.0337
2.	✓						0.4749 ± 0.0228
3.	✓	✓					0.5491 ± 0.0329
4.	✓	✓	✓				0.6259 ± 0.0400
5.	✓	✓	✓		✓		0.6336 ± 0.0396
6.	✓	✓	✓		✓	✓	0.6285 ± 0.0485
7.			✓	✓			0.4777 ± 0.0329
8.	✓		✓	✓			0.5080 ± 0.0434
9.	✓	✓	✓	✓			0.6811 ± 0.0339
10.	✓	✓	✓	✓	✓		**0.6823 ± 0.0343**
DFT-Computed Spectra							0.5616*****

*Note that DFT spectra do not overlap with the entire training/test dataset and scores cannot be cross validated.

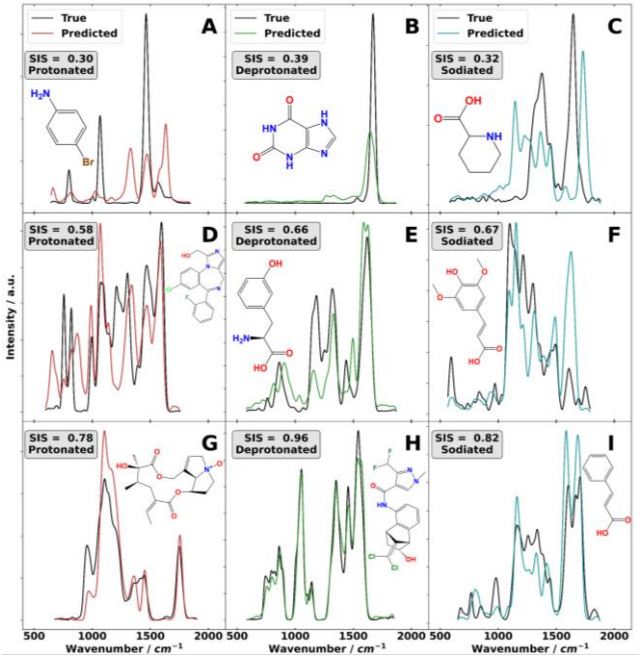

Figure 3. Graphormer-IRIS test predictions (red/green/blue) overlaid on experimental spectra (black). Predictions include below-average (A-C), average (D-F), and above-average (G-I) errors for protonated (A,D,G, red), deprotonated (B,E,H, green), and sodiated (C,F,I, blue) analyte spectra.

Figure 4 further emphasizes the importance of an explicit, contextual, global description of charge for predicting IRIS spectra generated from our ML architecture. Figure 4 panels A, B, and C show the IRIS spectra for phenylacetylglycine as measured and predicted for the deprotonated, protonated, and sodiated charge states, respectively. Similarly, Figure 4 panels D, E, and F show the measured and Graphormer-IRIS spectra for L-aspartic acid in the same charge states. By inspection, Graphormer-IRIS clearly captures the dramatic differences in IRIS spectra as a function of charge state, yielding high-quality predictions. In the case of phenylacetylglycine, several vibrational bands are at similar wavenumbers from one charge state to another, such as those associated with the C=C aromatic stretching frequency at ~1,650 cm^{-1}, but the IRIS spectrum of the deprotonated species (Figure 4A) exhibits fewer peaks than the protonated and sodiated species. Furthermore, the IRIS spectrum of protonated phenylacetylglycine (Figure 4B) exhibits relatively intense bands below 750 cm^{-1}, which are very weak or absent from the spectra of the sodiated and deprotonated charge states. The IRIS spectrum for the deprotonated charge state of L-aspartic acid (Figure 4D) demonstrates another interesting phenomenon that is captured by Graphormer-IRIS – molecules that contain multiple hydrogen bonding moieties can exhibit significant broadening of some spectral features. Interestingly, this band broadening is not observed experimentally for the protonated or sodiated forms of L-aspartic acid (Figures 4E and 4F), and Graphormer-IRIS also predicts these spectra reasonably well. For L-aspartic acid (Figure 4E), protonation occurs on the amine moiety and an intramolecular hydrogen bond can form to either of the carboxyl groups. Due to the asymmetry of the amine group along the carbon backbone, the C=O vibrational frequencies of the different hydrogen-bonded conformers are not equivalent, and one observes a pair of peaks at *ca.* 1,750 cm^{-1}. Graphormer-IRIS successfully captures this subtle interaction, likely owing to the robust pretraining steps. Moreover, Graphormer-IRIS can capture the variable red shifting of the carboxylic acid (Figure 4B: *ca.* 1,780 cm^{-1}, Figure 4C: *ca.* 1,736 cm^{-1}; $\Delta\nu = 44$ cm^{-1}) and amide (Figure 4B: *ca.* 1,682 cm^{-1}, Figure 4C: *ca.* 1,650 cm^{-1}; $\Delta\nu = 24$ cm^{-1}) stretching frequencies upon sodiation. The ability to capture this subtle effect is a testament to the model's contextual understanding of charge provided by the global node.

Transfer Learning Strategies. Given the relative paucity of the IRIS spectral library ($n = 312$), the success of Graphormer-IRIS hinges on effectively transferring knowledge from models trained on IR absorption and DFT-

computed spectra. As discussed in the *Supporting Information:Transfer Learning*, transfer learning can be seen through the lens of human learning, where analogies (*e.g.*, learning to eat yogurt compared to learning to eat soup) and easier versions of tasks (*e.g.*, walking compared to running) can improve model performance. In the context of IRIS, the task of predicting IR spectra for neutral species (*i.e.*, Graphormer-IR's large training set) can be thought to teach the model structure-to-vibration analogies. The DFT spectra data set, on the other hand, represents an "easier" version of gas phase ion spectra (*i.e.*, harmonic modes, no noise) that allows learning on a gradient. Finetuning the best-performing Graphormer-IR model on IRIS spectra (*i.e.*, no intermediate DFT pretraining step) resulted in a

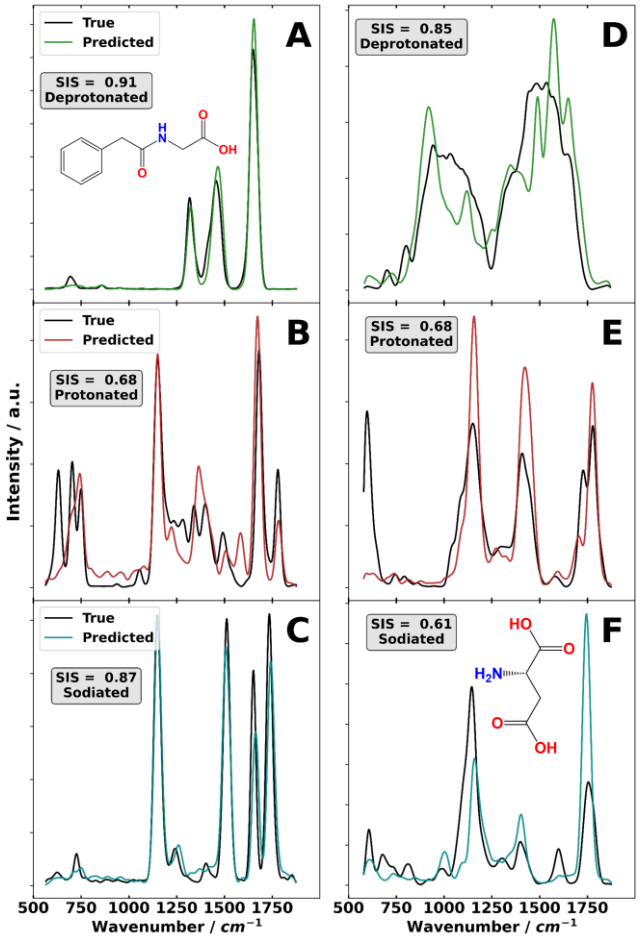

Figure 4. IRIS spectra of phenylacetylglycine in its (**A**) deprotonated (**B**) protonated, and (**C**) sodiated state, and of L-aspartic acid in its (**D**) deprotonated, (**E**) protonated, and (**F**) sodiated state. Graphormer-IRIS predictions are plotted in colour, experimental data are plotted in black. SIS scores for the various predicted spectra are reported in the appropriate panels.

performance increase of nearly 50%; SIS improved from $SIS_{\mu} = 0.4219 \pm 0.0337$ (Table 1; row #1) to $SIS_{\mu} = 0.6259 \pm 0.0400$ (Table 1; row #4). We also found that freezing the first layer of the graph feature encoder offered a small (but not statistically significant) improvement in model performance $SIS_{\mu} = 0.6336 \pm 0.0396$ (Table 1; row #5). Moreover, freezing the graph feature encoder offered no further improvement to model performance, yielding $SIS_{\mu} = 0.6285 \pm 0.0485$ (Table 1; row #6).

DFT Spectra Pretraining. The DFT spectral library consists of spectra generated from ionic structure ensembles that have been explicitly (de)protonated/sodiated and then optimized using DFT methods.[28] More details are available in the *Supporting Information: Computed DFT Spectra* section. Because the original Graphormer-IR models have minimal awareness of charge state (See *Supporting Information: Charge Encodings)*, we used the DFT-computed spectral library developed in reference 28 to improve the model's understanding of the charge state encodings. Since the DFT spectra dataset contains ionic conformers, we hypothesized that models pretrained on this dataset might gain a sense of (de)protonation/sodiation. Of course, the potential downside of using DFT-computed spectra in training purposes is that peak positions, which are computed within the harmonic oscillator approximation and corrected with a uniform scaling factor, are possibly different from those observed experimentally and so might be detrimental to model accuracy. When pretraining on the DFT spectra (after already pretraining on the experimental IR library for neutral species), then finetuning on experimental IRIS spectra, the model similarity score improves to $SIS_{\mu} = 0.6811 \pm 0.0339$ (Table 1; row #9). This accuracy is a substantial improvement over the DFT-prediction performance benchmark of $SIS_{\mu} = 0.5616$. The improvement in performance as a function of DFT pretraining suggests that the difference between experimental and DFT-computed peak positions might not be especially detrimental to model learning. When transferring knowledge of vibrational frequencies from the model for predicting experimental IR spectra to the DFT-refined model and, ultimately, to the experimental IRIS model, the transformer architecture might be better able to focus on robust anharmonic frequency correction and IRIS band intensity correction in the final step, having already gained some understanding for the impact of charge state via the intermediate DFT spectra pretraining step. To explore this notion, we examined model predictions made at intermediate stages of transfer learning. Figure S5C highlights an additional explanation for the positive effect of the DFT spectra finetuning procedure, which might act as a "cleaning" step in our workflow. Experimental IR spectra can contain noise and impurities, leading to extemporaneous signals that the model can learn during the training phase and subsequently introduce into predictions. By pretraining on DFT spectra before training on the IRIS spectra, the model learns to eliminate these "noisy" peaks from predictions and consider only "real" vibrational features. We also trained an uninitialized model (*i.e.*, one not pretrained on absorption IR spectra of neutral molecules) on only the DFT library using the global charge encoding. Doing so yields $SIS_{\mu} = 0.4777 \pm 0.0329$ (Table 1; row #7). Further finetuning of this model (as opposed to evaluating) on the experimental IRIS spectra improves the similarity score to $SIS_{\mu} = 0.5080 \pm 0.0434$ (Table 1; row #8). This result indicates that training solely on the DFT spectra is not

sufficient to yield an accurate IRIS model. Instead, it is necessary to first pretrain on experimental IR spectra for neutral species.

UMAP: Uniform Manifold Approximation and Projection for Dimensionality Reduction. To explore why pretraining steps are necessary to achieve optimal model performance (*i.e.*, what knowledge is passed to Graphormer-IRIS from transfer learning), we perform dimensionality reduction on model embeddings. Figure 5A shows the UMAP projection (See *Supporting Information: Universal Manifold Approximation and Projection)* of node embeddings generated by Graphormer-IR, illustrating some of the "chemical first principles" that might be passed to Graphormer-IRIS. This figure exhibits well-defined clusters based on atom type and hybridization, demonstrating a sophisticated understanding of local chemical environments and allowing us to draw three conclusions. First, the clustering of specific node level embeddings demonstrates that the Graphormer-IR model (*i.e.*, the first pretraining step) derives an understanding of the emergent "concept" of functional groups without explicit annotation of these geometric motifs and their properties. For instance, in Figure 5A, the embeddings of both the sp^2 oxygen (light red) and sp^2 nitrogen (light blue) moieties are clustered in the vicinity of one another on the UMAP plot. Secondly, the embeddings for moieties with dissimilar encodings but similar vibrational frequencies tend to be clustered. For example, the halogens (R-F R-Cl, R-Br, R-I) which produce similar vibrational frequencies (*ca.*550 – 850 cm^{-1}) have similar embeddings in the UMAP space (see Figure 5A). Likewise, 1,3 dipolar species including azides (R-N_3), isocyanates (R-N=C=O), and thiocyanates (R-N=C=S) have similar embeddings, which is likely due to the model's understanding of the similar stretching frequencies (*ca.* 2,250 cm^{-1}). Lastly, we observe that the UMAP x-dimension acts as a proxy for the extent of electron delocalization in each molecular ion. At the coarsest grain, the embeddings in the most negative x-direction consist largely of sp^2-hybridized carbon moieties and highly conjugated systems like pyridines, nitro groups, and aromatic species. Molecules in the most positive x-direction largely consisting of alkanes, silicon derivatives, and halogenated species. At a finer grain, this trend continues where conjugated R-N=C=O moieties (*e.g.*, R = benzyl groups) have embeddings shifted to the "delocalized" x-direction, relative to those that are not conjugated (*e.g.* R = alkyl groups). At their core, these phenomena speak to the derived "chemical intuition" of the foundational Graphormer-IR model and demonstrate the analogous chemical principles and structure-to-vibrational frequency relationships that are being passed to Graphormer-IRIS in the transfer learning process.

Figure 5B shows UMAP performed for embeddings generated by the best performing IRIS model (Table 1, Ablation #10). These embeddings cluster similarly to Graphormer-IR, but with lower diversity due to the smaller IRIS dataset. Figure 5B highlights specific node embeddings made for predictions of (E)-hex-2-enoylglycine and exemplifies Graphormer-IRIS's high contextualization of charge. The embedding for the carboxylic acid hydroxyl group oxygen atom for a deprotonated analyte (circled in green, Figure 5B) exhibits a much larger relative shift compared to the same hydroxyl moiety in sodiated and protonated charge states.

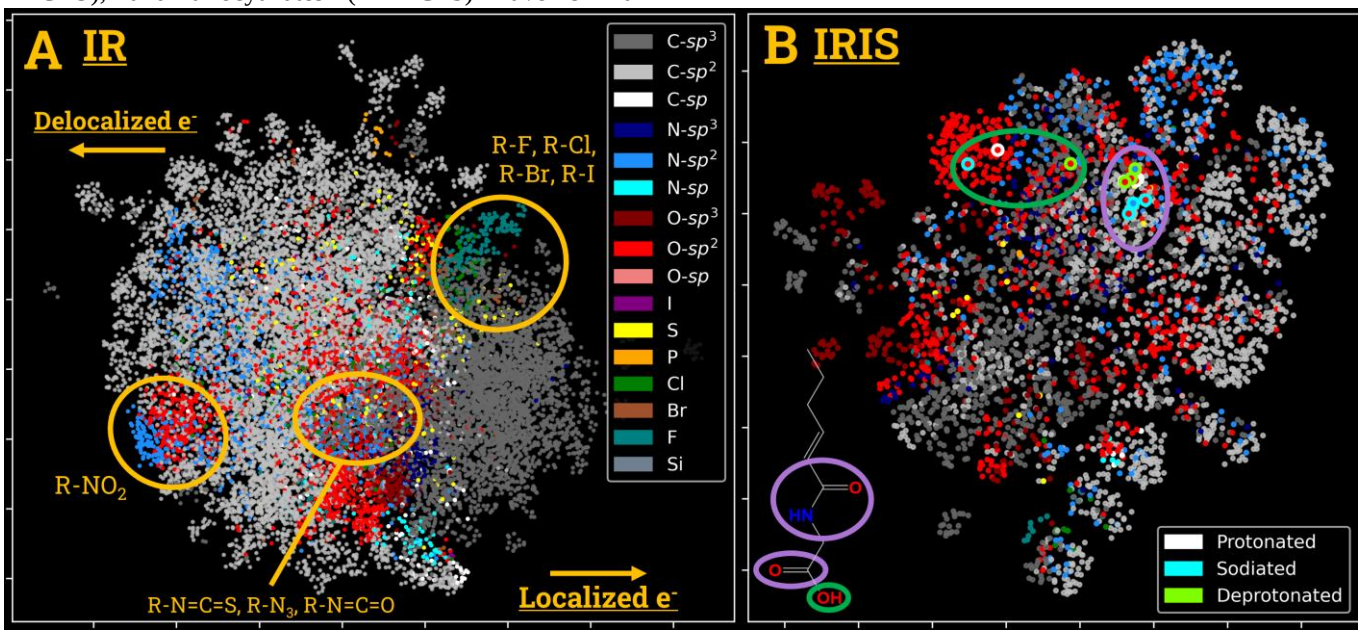

Figure 5. UMAP projections for (**A**) the best performing Graphormer-IR model and (**B**) the best performing Graphormer-IRIS model. Embeddings for panel **A** (1,200 molecules, 17,530 embeddings) were taken from a random sample of the gas- and condensed-phase test set from reference 45, and embeddings for panel **B** are the complete set of IRIS spectra (312 molecules, 4,166 embeddings). The dimensions for all plots are the arbitrary UMAP dimensions. Each point is labelled using the atom type and hybridization as encoded at the graph level using the DGL and RDkit featurization functions. Note that carboxylic acid hydroxyl moieties are labelled as sp^2 hybridized by the DGL featurization functions. Functional group and atom labels were determined by plotting the results of individual molecules/atom embeddings after UMAP projection

This relative shift is also not observed for the deprotonated versions of the other heteroatoms in the same molecule (circled in purple, Figure 5B). Given this embedding shift, we can conclude that Graphormer-IRIS has likely identified the site of deprotonation in (E)-hex-2-enoylglycine without an explicit annotation of deprotonation (other than the nonspecific global charge node). This identification of charge moiety is impressive given that this information has been learned indirectly from only the IRIS and DFT spectra prediction tasks. We can thus conclude that the global charge embedding succeeds at communicating molecule-wide shifts as a function of charge state (see Figure 4) resulting from highly specific and localized embedding shifts associated with the most probable sites of charge.

Conclusions

This study demonstrates the application of GNN transformers with multi-staged transfer learning to create a predictive model of IRIS spectroscopy using only SMILES codes as input. To create the Graphormer-IRIS model, transfer learning was used to refine the Graphormer-IR model via two separate steps: (*i*) pretraining on DFT-computed spectra and (*ii*) refining on experimental IRIS spectra. DFT-computed spectra for 10,336 ionic molecules were accessed from reference 28, and experimental IRIS spectra were obtained for 312 molecular ions. A key addition to the molecular graph structures that we used as inputs was the introduction of a flexible global node connected to all other (atom) nodes by a special edge type. This global node was used to describe the molecular charge state (*i.e.*, (de)protonated, sodiated). Our best-performing IRIS model achieved a test SIS_{μ}= 0.6823 ± 0.0343, which is 9.1σ (t = 15.9) or ~ 44% more accurate than the original Graphormer-IR model (SIS_{μ}= 0.4749 ± 0.0228; employing no transfer learning) and ~ 21% more accurate than the benchmark quantum-chemical methods (SIS_{μ}= 0.5616).[28] Moreover, our ML model predictions require seconds on a GPU, rather than hours-to-days for the calculation of a single DFT spectrum on a high performance cluster. [28]

Since the IRIS dataset employed in this study is relatively small, transfer learning is necessary to improve model generalizability. Borrowing from the field of NLP, we pretrain models on simplified versions of a task and with synthetic/approximate data, or by analogy, to provide the models with a more knowledgeable starting point. This approach ultimately improves prediction accuracy. Here, we explored using experimental IR spectra for neutral species and DFT-computed harmonic vibrational spectra for molecular ions for pretraining purposes.[28] The optimal scheme was first pretrained on the experimental IR spectra for neutrals, followed by an intermediate refinement step that used DFT spectra, then final refinement and transfer of information to the IRIS spectral predictions. Fine-tuning models on DFT spectra provided Graphormer-IRIS with an improved understanding of charge and additional examples of metabolite IRIS spectra. This approach might also effectively act as an intermediate "cleaning" step, where spectral noise found in the gas/condensed IR spectra are eliminated.

The ability of Graphormer-IRIS to accurately predict the (often dissimilar) spectra associated with different charge states of the same molecule arises from the richness of the description generated by the global charge node. For example, Graphormer-IRIS captures subtle effects such as shifting band positions due to hydrogen bonding interactions and red-shifting of carbonyl frequencies upon sodium adduction. The use of UMAP visualization provides additional explainability, demonstrating that the initial embeddings of the Graphormer-IR model provide a rich description of chemical structures, leading to an emergent understanding of functional groups and generalized trends in how electron density impacts IR frequencies. These rich embeddings provide the foundation of understanding for the successful predictions made by Graphormer-IRIS via transfer learning. Similar dimensionality reduction of the final Graphormer-IRIS models shows that its embeddings retain the knowledge of functional groups from Graphormer-IR, and that the charge encoding communicates a contextual understanding of likely sites of charge state derived from the global charge node. Given the richness of the "learned" chemical intuition that we observe in our model, we speculate that experimental IR(IS) data are an avenue to train "generalist" foundational deep learning frameworks by describing the patterns in molecular structure and bonding that are not obvious from a simple molecule graph.[61,62]

To employ IRIS as a tool for annotation of unknown features detected by mass spectrometry or for self-driving process optimization, it is necessary to overcome the current issues of data scarcity and the complex nuances of molecular charge. Here, we achieved this using multi-staged transfer learning to pass a learned "chemical intuition" from one model to another. Although a meaningful step forward, there remains a great deal to accomplish *en route* to improved MS feature annotation. For example, Graphormer-IRIS can provide accurate spectral predictions for small molecule metabolites, but our experimental training set requires significant expansion to create a generally applicable model. Such a model could be used as a discriminator for non-targeted MS feature annotation. ML-predicted IR and IRIS spectra could also prove valuable in a closed-loop self-driving framework for process optimization.

DATA AND SOFTWARE AVAILABILTY

Data access statements are available in the Supporting Information. Code for this project is available online alongside Graphormer-IR code at https://github.com/HopkinsLaboratory/Graphormer-IR

AUTHOR INFORMATION

Corresponding Authors

*W. Scott Hopkins, shopkins@uwaterloo.ca

*Jonathan Martens, jonathan.martens@ru.nl

AUTHOR CONTRIBUTIONS

CMKS built the model architecture, performed all machine learning experiments, and wrote the manuscript. LH helped

build and debug the model and aided in manuscript preparation. PT aided in data collection. TvW, KJH, GB, JO and JM helped conceive the project, provided the data, and aided in manuscript preparation. WSH was responsible for the concept and funding, provided experimental guidance, and aided in manuscript preparation.

FUNDING SOURCES

CMKS acknowledges financial support from NSERC in the form of a Canadian Graduate Scholarship and an Ontario Graduate Scholarship. LH acknowledges financial support from NSERC in the form of a Vanier Canada Graduate Scholarship. PT acknowledges financial support from NSERC in the form of a graduate scholarship. TvW, KJH, GB, JO and JM gratefully acknowledge the support of the FELIX technical staff. This project received funding from the Dutch Research Council (NWO) under grant number GWI Roadmap 184.034.022. Computations were performed at the national supercomputer Snellius at SurfSara in Amsterdam with the compute budget kindly provided through NWO Rekentijd grant 2021.055. WSH acknowledges funding from the Canadian Foundation for Innovation (CFI), Ontario Research Fund (ORF), and Natural Sciences and Engineering Research Council (NSERC) of Canada in the form of a Discovery Grant.

REFERENCES

(1) Heiles, S. Advanced Tandem Mass Spectrometry in Metabolomics and Lipidomics—Methods and Applications. *Analytical and Bioanalytical Chemistry 2021 413:24* **2021**, *413* (24), 5927–5948. https://doi.org/10.1007/S00216-021-03425-1.

(2) Hoffmann, W. D.; Jackson, G. P. Forensic Mass Spectrometry. *Annual Review of Analytical Chemistry* **2015**, *8* (Volume 8, 2015), 419–440. https://doi.org/10.1146/ANNUREV-ANCHEM-071114-040335/CITE/REFWORKS.

(3) Strynar, M.; Dagnino, S.; McMahen, R.; Liang, S.; Lindstrom, A.; Andersen, E.; McMillan, L.; Thurman, M.; Ferrer, I.; Ball, C. Identification of Novel Perfluoroalkyl Ether Carboxylic Acids (PFECAs) and Sulfonic Acids (PFESAs) in Natural Waters Using Accurate Mass Time-of-Flight Mass Spectrometry (TOFMS). *Environ Sci Technol* **2015**, *49* (19), 11622–11630. https://doi.org/10.1021/ACS.EST.5B01215/.

(4) Koelmel, J. P.; Stelben, P.; McDonough, C. A.; Dukes, D. A.; Aristizabal-Henao, J. J.; Nason, S. L.; Li, Y.; Sternberg, S.; Lin, E.; Beckmann, M.; Williams, A. J.; Draper, J.; Finch, J. P.; Munk, J. K.; Deigl, C.; Rennie, E. E.; Bowden, J. A.; Godri Pollitt, K. J. FluoroMatch 2.0—Making Automated and Comprehensive Non-Targeted PFAS Annotation a Reality. *Anal Bioanal Chem* **2022**, *414* (3), 1201–1215. https://doi.org/10.1007/S00216-021-03392-7/.

(5) Bach, E.; Schymanski, E. L.; Rousu, J. Joint Structural Annotation of Small Molecules Using Liquid Chromatography Retention Order and Tandem Mass Spectrometry Data. *Nature Machine Intelligence 2022 4:12* **2022**, *4* (12), 1224–1237. https://doi.org/10.1038/s42256-022-00577-2.

(6) Wang, M.; Jarmusch, A. K.; Vargas, F.; Aksenov, A. A.; Gauglitz, J. M.; Weldon, K.; Petras, D.; da Silva, R.; Quinn, R.; Melnik, A. V.; van der Hooft, J. J. J.; Caraballo-Rodríguez, A. M.; Nothias, L. F.; Aceves, C. M.; Panitchpakdi, M.; Brown, E.; Di Ottavio, F.; Sikora, N.; Elijah, E. O.; Labarta-Bajo, L.; Gentry, E. C.; Shalapour, S.; Kyle, K. E.; Puckett, S. P.; Watrous, J. D.; Carpenter, C. S.; Bouslimani, A.; Ernst, M.; Swafford, A. D.; Zúñiga, E. I.; Balunas, M. J.; Klassen, J. L.; Loomba, R.; Knight, R.; Bandeira, N.; Dorrestein, P. C. Mass Spectrometry Searches Using MASST. *Nature Biotechnology 2020 38:1* **2020**, *38* (1), 23–26. https://doi.org/10.1038/s41587-019-0375-9.

(7) Petras, D.; Phelan, V. V.; Acharya, D.; Allen, A. E.; Aron, A. T.; Bandeira, N.; Bowen, B. P.; Belle-Oudry, D.; Boecker, S.; Cummings, D. A.; Deutsch, J. M.; Fahy, E.; Garg, N.; Gregor, R.; Handelsman, J.; Navarro-Hoyos, M.; Jarmusch, A. K.; Jarmusch, S. A.; Louie, K.; Maloney, K. N.; Marty, M. T.; Meijler, M. M.; Mizrahi, I.; Neve, R. L.; Northen, T. R.; Molina-Santiago, C.; Panitchpakdi, M.; Pullman, B.; Puri, A. W.; Schmid, R.; Subramaniam, S.; Thukral, M.; Vasquez-Castro, F.; Dorrestein, P. C.; Wang, M. GNPS Dashboard: Collaborative Exploration of Mass Spectrometry Data in the Web Browser. *Nature Methods 2021 19:2* **2021**, *19* (2), 134–136. https://doi.org/10.1038/s41592-021-01339-5.

(8) Schymanski, E. L.; Jeon, J.; Gulde, R.; Fenner, K.; Ruff, M.; Singer, H. P.; Hollender, J. Identifying Small Molecules via High Resolution Mass Spectrometry: Communicating Confidence. *Environ Sci Technol* **2014**, *48* (4), 2097–2098. https://doi.org/10.1021/ES5002105/ASSET/IMAGES/LARGE/ES-2014-002105_0001.JPEG.

(9) Schrimpe-Rutledge, A. C.; Codreanu, S. G.; Sherrod, S. D.; McLean, J. A. Untargeted Metabolomics Strategies—Challenges and Emerging Directions. *J Am Soc Mass Spectrom* **2016**, *27* (12), 1897–1905. https://doi.org/10.1007/S13361-016-1469-Y/ASSET/IMAGES/MEDIUM/JS8B05178_0005.GIF.

(10) Stein, S. E.; Scott, D. R. Optimization and Testing of Mass Spectral Library Search Algorithms for Compound Identification. *J Am Soc Mass Spectrom* **1994**, *5* (9), 859–866. https://doi.org/10.1016/1044-0305(94)87009-8.

(11) Kind, T.; Tsugawa, H.; Cajka, T.; Ma, Y.; Lai, Z.; Mehta, S. S.; Wohlgemuth, G.; Barupal, D. K.; Showalter, M. R.; Arita, M.; Fiehn, O. Identification of Small Molecules Using Accurate Mass MS/MS Search. *Mass Spectrom Rev* **2018**, *37* (4), 513–532. https://doi.org/10.1002/MAS.21535.

(12) Da Silva, R. R.; Dorrestein, P. C.; Quinn, R. A. Illuminating the Dark Matter in Metabolomics. *Proc Natl Acad Sci U S A* **2015**, *112* (41), 12549–12550. https://doi.org/10.1073/PNAS.1516878112/ASSET/FCB96DFB-4497-4C31-97E2-265AF95CF1E8/ASSETS/GRAPHIC/PNAS.1516878112FIG01.JPEG.

(13) Stravs, M. A.; Dührkop, K.; Böcker, S.; Zamboni, N. MSNovelist: De Novo Structure Generation from Mass Spectra. *Nature Methods 2022 19:7* **2022**, *19* (7), 865–870. https://doi.org/10.1038/s41592-022-01486-3.

(14) Song, Y.; Song, Q.; Liu, W.; Li, J.; Tu, P. High-Confidence Structural Identification of Metabolites Relying on Tandem Mass Spectrometry through Isomeric Identification: A Tutorial. *TrAC Trends in Analytical Chemistry* **2023**, *160*, 116982. https://doi.org/10.1016/J.TRAC.2023.116982.

(15) Liu, Y.; Zhang, X.; Zhao, W.; Zhu, D.; Cui, X. De Novo Molecular Structure Generation from Mass Spectra. *Proceedings - 2023 2023 IEEE International Conference on Bioinformatics and Biomedicine, BIBM 2023* **2023**, 373–378. https://doi.org/10.1109/BIBM58861.2023.10385903.

(16) Martens, J.; van Outersterp, R. E.; Vreeken, R. J.; Cuyckens, F.; Coene, K. L. M.; Engelke, U. F.; Kluijtmans, L. A. J.; Wevers, R. A.; Buydens, L. M. C.; Redlich, B.; Berden, G.; Oomens, J. Infrared Ion Spectroscopy: New Opportunities for Small-Molecule Identification in Mass Spectrometry - A Tutorial Perspective. *Anal Chim Acta* **2020**, *1093*, 1–15. https://doi.org/10.1016/J.ACA.2019.10.043.

(17) Polfer, N. C. Infrared Multiple Photon Dissociation Spectroscopy of Trapped Ions. *Chem Soc Rev* **2011**, *40* (5), 2211–2221. https://doi.org/10.1039/C0CS00171F.

(18) Martens, J.; Berden, G.; Gebhardt, C. R.; Oomens, J. Infrared Ion Spectroscopy in a Modified Quadrupole Ion Trap Mass Spectrometer at the FELIX Free Electron Laser Laboratory. *Review of Scientific Instruments* **2016**, *87* (10). https://doi.org/10.1063/1.4964703/368272.

(19) Van Geenen, F. A. M. G.; Kranenburg, R. F.; Van Asten, A. C.; Martens, J.; Oomens, J.; Berden, G. Isomer-Specific Two-Color Double-Resonance IR2MS3 Ion Spectroscopy Using a Single Laser: Application in the Identification of Novel Psychoactive Substances. *Anal Chem* **2021**, *93* (4), 2687–2693. https://doi.org/10.1021/ACS.ANALCHEM.0C05042/ASSET/IMAGES/LARGE/AC0C05042_0006.JPEG.

(20) Vink, M. J. A.; Van Geenen, F. A. M. G.; Berden, G.; O'Riordan, T. J. C.; Howe, P. W. A.; Oomens, J.; Perry, S. J.; Martens, J. Structural Elucidation of Agrochemicals and Related Derivatives Using Infrared Ion Spectroscopy. *Environ Sci Technol* **2022**, *56* (22), 15563–15572. https://doi.org/10.1021/ACS.EST.2C03210.

(21) Martens, J.; Berden, G.; Van Outersterp, R. E.; Kluijtmans, L. A. J.; Engelke, U. F.; Van Karnebeek, C. D. M.; Wevers, R. A.; Oomens, J. Molecular Identification in Metabolomics Using Infrared Ion Spectroscopy. *Scientific Reports 2017 7:1* **2017**, *7* (1), 1–5. https://doi.org/10.1038/s41598-017-03387-4.

(22) Martens, J.; Koppen, V.; Berden, G.; Cuyckens, F.; Oomens, J. Combined Liquid Chromatography-Infrared Ion Spectroscopy for Identification of Regioisomeric Drug Metabolites. *Anal Chem* **2017**, *89* (8), 4359–4362. https://doi.org/10.1021/ACS.ANALCHEM.7B00577.

(23) Granda, J. M.; Donina, L.; Dragone, V.; Long, D. L.; Cronin, L. Controlling an Organic Synthesis Robot with Machine Learning to Search for New Reactivity. *Nature 2018 559:7714* **2018**, *559* (7714), 377–381. https://doi.org/10.1038/s41586-018-0307-8.

(24) Abolhasani, M.; Kumacheva, E. The Rise of Self-Driving Labs in Chemical and Materials Sciences. *Nature Synthesis 2023 2:6* **2023**, *2* (6), 483–492. https://doi.org/10.1038/s44160-022-00231-0.

(25) Abikhodr, A. H.; Warnke, S.; Ben Faleh, A.; Rizzo, T. R. Combining Liquid Chromatography and Cryogenic IR Spectroscopy in Real Time for the Analysis of Oligosaccharides. *Anal Chem* **2024**, *96* (4), 1462–1467. https://doi.org/10.1021/ACS.ANALCHEM.3C03578/ASSET/IMAGES/LARGE/AC3C03578_0005.JPEG.

(26) Schindler, B.; Laloy-Borgna, G.; Barnes, L.; Allouche, A. R.; Bouju, E.; Dugas, V.; Demesmay, C.; Compagnon, I. Online Separation and Identification of Isomers Using Infrared Multiple Photon Dissociation Ion Spectroscopy Coupled to Liquid Chromatography: Application to the Analysis of Disaccharides Regio-Isomers and Monosaccharide Anomers. *Anal Chem* **2018**, *90* (20), 11741–11745. https://doi.org/10.1021/ACS.ANALCHEM.8B02801/ASSET/IMAGES/LARGE/AC-2018-028012_0002.JPEG.

(27) van Outersterp, R. E.; Oosterhout, J.; Gebhardt, C. R.; Berden, G.; Engelke, U. F. H.; Wevers, R. A.; Cuyckens, F.; Oomens, J.; Martens, J. Targeted Small-Molecule Identification Using Heartcutting Liquid Chromatography-Infrared Ion Spectroscopy. *Anal Chem* **2023**, *95* (6), 3406–3413. https://doi.org/10.1021/ACS.ANALCHEM.2C04904/ASSET/IMAGES/LARGE/AC2C04904_0005.JPEG.

(28) Houthuijs, K. J.; Berden, G.; Engelke, U. F. H.; Gautam, V.; Wishart, D. S.; Wevers, R. A.; Martens, J.; Oomens, J. An In Silico Infrared Spectral Library of Molecular Ions for Metabolite Identification. *Anal Chem* **2023**, *95* (23), 8998–9005. https://doi.org/10.1021/ACS.ANALCHEM.3C01078/.

(29) Bonney, J. R.; Kang, W. Y.; Specker, J. T.; Liang, Z.; Scoggins, T. R.; Prentice, B. M. Relative Quantification of Lipid Isomers in Imaging Mass Spectrometry Using Gas-Phase Charge Inversion Ion/Ion Reactions and Infrared Multiphoton Dissociation. *Anal Chem* **2023**, *95* (48), 17766–17775. https://doi.org/10.1021/ACS.ANALCHEM.3C03804/ASSET/IMAGES/LARGE/AC3C03804_0007.JPEG.

(30) Becher, S.; Berden, G.; Martens, J.; Oomens, J.; Heiles, S. IRMPD Spectroscopy of [PC (4:0/4:0) + M]+ (M = H, Na, K) and Corresponding CID Fragment Ions. *J Am Soc Mass Spectrom* **2021**, *32* (12), 2874–2884. https://doi.org/10.1021/JASMS.1C00277.

(31) Ledvina, A. R.; Lee, M. V.; McAlister, G. C.; Westphall, M. S.; Coon, J. J. Infrared Multiphoton Dissociation for Quantitative Shotgun Proteomics. *Anal Chem* **2012**, *84* (10), 4513–4519. https://doi.org/10.1021/AC300367P/SUPPL_FILE/AC300367P_SI_001.PDF.

(32) Smyrnakis, A.; Levin, N.; Kosmopoulou, M.; Jha, A.; Fort, K.; Makarov, A.; Papanastasiou, D.; Mohammed, S. Characterisation of an Omnitrap-Orbitrap Platform Equipped with IRMPD, UVPD and ExD for the Analysis of Peptides and Proteins. *bioRxiv* **2023**, 2023.05.15.540788. https://doi.org/10.1101/2023.05.15.540788.

(33) Maitre, P.; Scuderi, D.; Corinti, D.; Chiavarino, B.; Crestoni, M. E.; Fornarini, S. Applications of Infrared Multiple Photon Dissociation (IRMPD) to the Detection of Posttranslational Modifications. *Chem Rev* **2020**, *120* (7), 3261–3295. https://doi.org/10.1021/ACS.CHEMREV.9B00395/ASSET/IMAGES/LARGE/CR9B00395_0035.JPEG.

(34) Cismesia, A. P.; Bell, M. R.; Tesler, L. F.; Alves, M.; Polfer, N. C. Infrared Ion Spectroscopy: An Analytical Tool for the Study of Metabolites. *Analyst* **2018**, *143* (7), 1615–1623. https://doi.org/10.1039/C8AN00087E.

(35) Vink, M. J. A.; Alarcan, J.; Martens, J.; Buma, W. J.; Braeuning, A.; Berden, G.; Oomens, J. Structural Elucidation of Agrochemical Metabolic Transformation Products Based on Infrared Ion Spectroscopy to Improve in Silico Toxicity Assessment. *Chem Res Toxicol* **2024**, *37* (1), 81–97. https://doi.org/10.1021/ACS.CHEMRESTOX.3C00316/ASSET/IMAGES/LARGE/TX3C00316_0011.JPEG.

(36) Polfer, N. C.; Valle, J. J.; Moore, D. T.; Oomens, J.; Eyler, J. R.; Bendiak, B. Differentiation of Isomers by Wavelength-Tunable Infrared Multiple-Photon Dissociation-Mass Spectrometry: Application to Glucose-Containing Disaccharides. *Anal Chem* **2006**, *78* (3), 670–679. https://doi.org/10.1021/AC0519458/ASSET/IMAGES/LARGE/AC0519458F00008.JPEG.

(37) Moge, B.; Yeni, O.; Infantino, A.; Compagnon, I. CO2 Laser Enhanced Rapid IRMPD Spectroscopy for Glycan Analysis. *Int J Mass Spectrom* **2023**, *490*, 117071. https://doi.org/10.1016/J.IJMS.2023.117071.

(38) Floris, F.; Vallotto, C.; Chiron, L.; Lynch, A. M.; Barrow, M. P.; Delsuc, M. A.; O'Connor, P. B. Polymer Analysis in the Second Dimension: Preliminary Studies for the Characterization of Polymers with 2D MS. *Anal Chem* **2017**, *89* (18), 9892–9899. https://doi.org/10.1021/ACS.ANALCHEM.7B02086/ASSET/IMAGES/LARGE/AC-2017-020868_0002.JPEG.

(39) J.A. Vink, M.; A.M.G. van Geenen, F.; Berden, G.; J. C. O'Riordan, T.; W.A. Howe, P.; Oomens, J.; J. Perry, S.; Martens, J. Structural Elucidation of Agrochemicals and Related Derivatives Using Infrared Ion Spectroscopy. *Environmental Science & Technology* **2022**, *56* (22), 15563–15572. https://doi.org/10.1021/acs.est.2c03210.

(40) Pascale, R.; Acquavia, M. A.; Onzo, A.; Cataldi, T. R. I.; Calvano, C. D.; Bianco, G. Analysis of Surfactants by Mass Spectrometry: Coming to Grips with Their Diversity. *Mass Spectrom Rev* **2023**, *42* (5), 1557–1588. https://doi.org/10.1002/MAS.21735.

(41) E. Lee, A.; Featherstone, J.; Martens, J.; B. McMahon, T.; Scott Hopkins, W. Fluorinated Propionic Acids Unmasked: Puzzling Fragmentation Phenomena of the Deprotonated Species. *J Phys Chem Lett* **2024**, *15* (11), 3029–3036. https://doi.org/10.1021/acs.jpclett.3c03400.

(42) Houthuijs, K. J.; Horn, M.; Vughs, D.; Martens, J.; Brunner, A. M.; Oomens, J.; Berden, G. Identification of Organic Micro-Pollutants in Surface Water Using MS-Based Infrared Ion Spectroscopy. *Chemosphere* **2023**, *341*, 140046. https://doi.org/10.1016/J.CHEMOSPHERE.2023.140046.

(43) McGill, C.; Forsuelo, M.; Guan, Y.; Green, W. H. Predicting Infrared Spectra with Message Passing Neural Networks. *Journal of Chemical Information and Modeling*. American Chemical Society June 28, 2021, pp 2594–2609. https://doi.org/10.1021/acs.jcim.1c00055.

(44) Saquer, N.; Iqbal, R.; Ellis, J. D.; Yoshimatsu, K. Infrared Spectra Prediction Using Attention-Based Graph Neural Networks. *Digital Discovery* **2024**. https://doi.org/10.1039/D3DD00254C.

(45) Stienstra, C. M. K.; Hebert, L.; Thomas, P.; Haack, A.; Guo, J.; Hopkins, W. S. Graphormer-IR: Graph Transformers Predict Experimental IR Spectra Using Highly Specialized Attention. *J Chem Inf Model* **2024**, *64* (12), 4613–4629. https://doi.org/10.1021/ACS.JCIM.4C00378.

(46) Ye, S.; Zhong, K.; Zhang, J.; Hu, W.; Hirst, J. D.; Zhang, G.; Mukamel, S.; Jiang, J. A Machine Learning Protocol for Predicting Protein Infrared Spectra. *J Am Chem Soc* **2020**, *142* (45), 19071–19077. https://doi.org/10.1021/JACS.0C06530/ASSET/IMAGES/LARGE/JA0C06530_0005.JPEG.

(47) Laurens, G.; Rabary, M.; Lam, J.; Peláez, D.; Allouche, A.-R. Infrared Spectra of Neutral Polycyclic Aromatic Hydrocarbons by Machine Learning. *Theor Chem Acc* **2020**, *140* (6). https://doi.org/10.1007/s00214-021-02773-6.

(48) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem Sci* **2017**, *8* (10), 6924–6935. https://doi.org/10.1039/C7SC02267K.

(49) Kovacs, P.; Zhu, X.; Carrete, J.; Madsen, G. K. H.; Wang, Z. Machine-Learning Prediction of Infrared Spectra of Interstellar Polycyclic Aromatic Hydrocarbons. *Astrophys J* **2020**, *902* (2), 100. https://doi.org/10.3847/1538-4357/abb5b6.

(50) Zou, Z.; Zhang, Y.; Liang, L.; Wei, M.; Leng, J.; Jiang, J.; Luo, Y.; Hu, W. A Deep Learning Model for Predicting Selected Organic Molecular Spectra. *Nature Computational Science 2023 3:11* **2023**, *3* (11), 957–964. https://doi.org/10.1038/s43588-023-00550-y.

(51) P.J. Linstrom and W.G. Mallard. "Infrared Spectra." In *NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; National Institute of Standards and Technology: Gaithersburg MD, 2022; Vol. 69.

(52) *National Institute of Advanced Science and Technology, SDBS, Web.*

(53) Craver, C. *The Coblentz Society Desk Book of Infrared Spectra*, 2nd ed.; Kirkwood, MO, 1982.

(54) Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; Polosukhin, I. Attention Is All You Need. *Adv Neural Inf Process Syst* **2017**, *2017-December*, 5999–6009.

(55) Ying, C.; Cai, T.; Luo, S.; Zheng, S.; Ke, G.; He, D.; Shen, Y.; Liu, T.-Y. Do Transformers Really Perform Bad for Graph Representation? *NIPS'21: Proceedings of the 35th International Conference on Neural Information Processing Systems* **2021**, 28877–28888. https://doi.org/https://arxiv.org/abs/2106.05234.

(56) Hu, W.; Fey, M.; Ren, H.; Nakata, M.; Dong, Y.; Leskovec, J. OGB-LSC: A Large-Scale Challenge for Machine Learning on Graphs. **2021**.

(57) Shi, Y.; Zheng, S.; Ke, G.; Shen, Y.; You, J.; He, J.; Luo, S.; Liu, C.; He, D.; Liu, T.-Y. Benchmarking Graphormer on Large-Scale Molecular Modeling Datasets. **2022**.

(58) *OGB-LSC: A Large-Scale Challenge for Machine Learning on Graphs | Papers With Code*. https://paperswithcode.com/paper/ogb-lsc-a-large-scale-challenge-for-machine (accessed 2023-07-20).

(59) Das, A.; Shuaibi, M.; Palizhati, A.; Goyal, S.; 1→3, A. G.; Kolluru, A.; Lan, J.; Rizvi, A.; Sriram, A.; Wood, B.; Parikh, D.; Ulissi, Z.; Zitnick, C. L.; Ke, G.; Zheng, S.; Shi, Y.; He, D.; Liu, T.-Y.; Ying, C.; You, J.; He, Y.; Grigoriev, R.; Lukin, R.; Yarullin, A.; Faleev, M.; Kiela, D.; Ciccone, M.; Caputo, B. The Open Catalyst Challenge 2021: Competition Report. *Proceedings of Machine Learning Research*. PMLR July 20, 2022, pp 29–40. https://proceedings.mlr.press/v176/das22a.html (accessed 2023-07-20).

(60) Langdon, S. R.; Brown, N.; Blagg, J. Scaffold Diversity of Exemplified Medicinal Chemistry Space. *J Chem Inf Model* **2011**, *51* (9), 2174–2185. https://doi.org/10.1021/CI2001428/SUPPL_FILE/CI2001428_SI_001.PDF.

(61) Ahmad, W.; Simon, E.; Chithrananda, S.; Ramsundar, B. ChemBERTa-2: Towards Chemical Foundation Models. **2022**.

(62) Clarté, L.; Loureiro, B.; Krzakala, F.; -, al; Hinz, F. B.; Mahmoud, A. H.; Lill -, M. A.; Emami, P.; Perreault, A.; Irwin, R.; Dimitriadis, S.; He, J.; Jannik Bjerrum, E. Chemformer: A Pre-Trained Transformer for Computational Chemistry. *Mach Learn Sci Technol* **2022**, *3* (1), 015022. https://doi.org/10.1088/2632-2153/AC3FFB.