Graphormer-IR: Graph Transformers Can Predict Experimental IR Spectra Using Highly Specialized Attention

Cailum M. K. Stienstra\textsuperscript{a}, Liam Hebert\textsuperscript{b}, Patrick Thomas\textsuperscript{a}, Alexander Haack\textsuperscript{a}, Jason Guo\textsuperscript{a}, W. Scott Hopkins\textsuperscript{a,c,d}

\textsuperscript{a} Department of Chemistry, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada,
\textsuperscript{b} Cheriton School of Computer Science, University of Waterloo, Waterloo, Ontario, N2L 3G1, Canada
\textsuperscript{c} Waterline Innovation, Waterloo, Ontario, NOB 270, Canada
\textsuperscript{d} Centre for Eye and Vision Research, Hong Kong Science Park, New Territories, 999077, Hong Kong.

**ABSTRACT:** Infrared (IR) spectroscopy is crucial in various chemical and forensic domains, but faster in silico methods for predicting experimental spectra are needed due to the time and accuracy limitations of ab initio methods. We employ Graphormer, a graph neural network (GNN) transformer, to predict IR spectra using only Simplified Molecular-Input Line-Entry System (SMILES) strings. Our dataset includes 53,528 high-quality spectra with elements H, C, N, O, F, Si, S, P, Cl, Br, and I in five solvent phases. When using only atomic numbers for node encodings, Graphormer-IR achieved $\text{SI}_\mu$ test scores of $0.8494 \pm 0.0012$ (n=5), surpassing the state-of-the-art Chemprop-IR ($\text{SI}_\mu = 0.8409 \pm 0.0014$, n=5), with only 36% of the encoded information. Augmenting node embeddings with additional node-level descriptors in learned embeddings generated through a multi-layer perceptron improves scores to $\text{SI}_\mu = 0.8523 \pm 0.0006$, a total improvement of 19.7%. These improved scores show how Graphormer-IR excels in capturing long-range interactions like hydrogen bonding, anharmonic peak positions in experimental spectra, and stretching frequencies of uncommon functional groups. Scaling our architecture to 210 attention heads demonstrates specialist-like behavior for distinct IR frequencies that improves model performance. Our model utilizes novel architectures, including a global node for solvent phase encoding, learned node feature embeddings, and a 1D smoothing CNN. Graphormer-IR’s innovations underscore its potency over traditional message-passing neural networks (MPNNs) due to its expressive embeddings and ability to capture long-range intra-molecular relationships.

The study of infrared (IR) radiation was pioneered by William Herschel in 1800, unlocking a new dimension of molecular analysis by harnessing the interaction between molecules and infrared radiation. Herschel discovered that different molecules absorb and emit distinct infrared frequencies, allowing for their identification and characterization.\textsuperscript{1} Infrared spectra were not measured until 1881, when Abney and Festing were able to capture vibrational frequencies used photographic plates.\textsuperscript{2} The first breakthrough towards “modern” IR spectroscopy came in 1905 when William Coblentz published a large library of IR spectra for molecules between 1,000 and 16,000 nm that allowed researchers to fingerprint specific molecules using frequencies that could be related to pieces of molecular structure.\textsuperscript{1,3} IR spectroscopy provides an invaluable tool for identifying compounds and monitoring changes in an analyte matrix because vibrational modes are tightly linked to molecular substructure. IR spectroscopy assumes a vital role in ensuring pharmaceuticals quality control. It enables the identification and quantification of impurities or degradation products in drugs to determine product quality and to meet stringent regulatory requirements.\textsuperscript{4,5} In biochemistry, protein structure (alpha helices, beta sheets, and random coil regions), conformational changes, and interactions (e.g., hydrogen bonding) can all be monitored with IR spectroscopy.\textsuperscript{6,7} Analyzing infrared spectra aids in understanding the chemical composition, chain length, and degree of polymerization.\textsuperscript{8,9} In inorganic chemistry, this technique aids in identifying metal complexes, coordination compounds, and minerals. Analyzing vibrational modes (and their relative shifts) and absorption frequencies provides insights into their structures, bonding, and reactivity.\textsuperscript{10,11}

Forensic science finds use from IR spectroscopy in evidence analysis, characterizing substances like drugs, fibers, and paints. By scrutinizing the infrared spectra of unknown substances found at crime scenes, forensic chemists can determine their composition and origin, providing valuable evidence for criminal investigations.\textsuperscript{12,13} In the domain of cultural heritage preservation, IR spectroscopy plays a significant role in analyzing pigments, binders, and coatings on artworks and artifacts to support conservation and restoration efforts.\textsuperscript{14,15}

Given such wide practical application, a means to consistently predict accurate experimental infrared spectra for experimental benchmarking or pre-screening is highly desirable. Current ab initio methods have inherent limitations when it comes to accurately calculating infrared (IR) spectra. One notable challenge is accurately capturing the complex interplay of molecular vibrations that contribute to IR spectra. Harmonic calculations, based on the assumption of small atomic displacement from their equilibrium positions, are computationally feasible in a reasonable amount of time but may not adequately account for anharmonic effects in highly complex systems. Calculation of anharmonic IR spectra, which consider larger displacements and higher-order vibrational couplings, are difficult to predict accurately due to the computational demands associated with high-level quantum mechanical calculations.\textsuperscript{16} Calculating typical anharmonic infrared spectra of small organic molecules using ab initio methods can take days to weeks, scaling rapidly with the number of heavy atoms. Frequencies in both harmonic and anharmonic calculations are also typically adjusted by a method-specific scaling factor. Previous work has shown that variation in these scaling factors can lead to discrepancy in peak positions of as high as 30 cm\textsuperscript{-1}.\textsuperscript{17} Additionally, environmental factors like solvent interactions or temperature effects complicate spectra are also not easily captured by fundamental calculations.\textsuperscript{18}
Given these difficulties in generalizing calculations ab initio anharmonic IR spectra, deep learning methods provide an attractive solution because of the speed of evaluation and the ad hoc evaluation of environment (temperature, solvent, etc.) from the dataset. As such, infrared spectra have been utilized in machine learning models in diverse ways. Medium and near-IR spectra have been used for classification and prediction of various “real” matrices. This is included but not limited to classification of plastic waste,\textsuperscript{19} prediction of beer quality (pH, alcohol content, foam content),\textsuperscript{20} prediction of neonatal respiratory distress biomarker concentrations,\textsuperscript{21} among other systems.\textsuperscript{22–28} Deep learning attempts to predict IR spectra include the use of multilayer perceptrons (MLPs)\textsuperscript{30,31}, prediction of protein IR spectra,\textsuperscript{32} and use of Morgan fingerprints to predict interstellar polycyclic aromatic hydrocarbon (PAH) infrared spectra.\textsuperscript{33} MLPs have also been used to enable calculation of infrared spectra by predicting potential energies and dipole moments using symmetry\textsuperscript{34} and radial interaction functions.\textsuperscript{35,36} The most relevant example of deep learning is Chemprop-IR,\textsuperscript{37} a study predicting IR spectra from chemical structure using a derivation of the Directed Message Passing Neural Network (D-MPNN) package Chemprop.\textsuperscript{38}

Originally conceived in 2019 to predict the output of quantum chemical calculations,\textsuperscript{39} MPNNs are a type of graph neural networks (GNNs) that have proven to be a valuable tool for modeling chemical systems that can be expressed as mathematical graphs. Notably, MPNNs have yielded state-of-the-art scores in a variety of deep learning tasks in chemistry including prediction of retention times,\textsuperscript{40} pharmacological properties,\textsuperscript{41,42} and various physicochemical properties.\textsuperscript{43–45} Like any architecture, MPNNs also possess certain limitations. One prominent flaw lies in their ability to capture complex higher-order relationships within chemical systems. MPNNs typically rely on a fixed number of message-passing steps, which can restrict their capacity to capture intricate long-range interactions and dependencies. Chemical systems often exhibit long-range interactions that play a vital role in determining properties and behaviors. In the prediction of infrared spectra, this could be relevant in calculating the “true” anharmonicity of a particular vibrational frequency, which is influenced by global molecule electronics, capturing solvent effects, or considering long-range interactions like hydrogen bonding.

We investigate the implementation of Graphormer\textsuperscript{46} to improve the prediction of infrared spectra using deep learning. Graphormer is a graph neural network that adapts the transformer architecture\textsuperscript{47} and was made by Microsoft that was the winner of the 2021 Open Catalyst challenge.\textsuperscript{48} By utilizing a global receptive field (i.e., all nodes in the graph exchange information with all other nodes) and the attention mechanism that allows for contextual weighting, Graphormer has achieved state of the art predictions in molecular HOMO-LUMO gaps,\textsuperscript{49} inhibition of HIV viral reproduction,\textsuperscript{50} biological activities,\textsuperscript{50} and molecular docking.\textsuperscript{51} At the time of writing, Graphormer has only been applied to spectroscopic predictions to tandem mass spectra (MS\textsuperscript{2}).\textsuperscript{52} In this work, we show how Graphormer can be used to improve on the groundwork laid by McGill,\textsuperscript{37} demonstrating the immense expressiveness achieved by graph transformer network in the prediction of IR spectra.

Methods

Dataset

Large collections of infrared spectra are owned by private organizations across a variety of domains, so no unified dataset is available. Due to the local copyrights of the external data sources, no complete “machine learning ready” dataset is publicly available (spectra stored on separate websites, not standardized, undigitized, etc.). As such, it was necessary to obtain, clean, and process a library of IR spectra from different domains. IR spectra were taken from three online sources: the National Institute of Advanced Industrial Science and Technology (AIST),\textsuperscript{53} the National Institute of Standards and Technology (NIST),\textsuperscript{54} and the Coblenz Society.\textsuperscript{55} Complete data access statements can be found in the supplementary information. In total, these datasets contain a mixture of gas-phase and condensed-phase IR spectra. The condensed phase spectra included spectra obtained in media including pressed KBr pellets, nujol mull (suspension in mineral oil), carbon tetrachloride (CCl\textsubscript{4}) solution, and pure liquid film. A complete breakdown of the composition of these spectra can be found in Figure S3.

To improve the quality of the dataset for deep learning, the initial collection of spectra (~60,000) was pruned such that it:

i) Only included molecules that could be expressed as molecular graphs (no solvent clusters, adducts);

ii) Included only features that were well-represented within the dataset (e.g., atom type, solvent phase);

iii) Included well-resolved, high-quality spectra to limit the effects of noise on learning.

To these ends, only molecules containing main block elements (H, C, N, O, F, Si, S, P, Cl, Br, I) were considered. An analysis of the representation of these elements and the distribution of molecular weights is available in figure S1. Any species containing adducts, dimers, multimers or ionic complexes (i.e., salts) were excluded from the dataset. Low quality spectra (i.e., broad peaks, noisy spectra, contamination from water at 3,400 cm\textsuperscript{-1}) were excluded from the dataset at the discretion of the authors. After pruning, the final dataset contained 53,528 spectra representing a total of 30,951 unique Simplified Molecular-Input Line-Entry System (SMILES) strings. While there is likely large overlap between our dataset and that produced by McGill,\textsuperscript{37} there are likely meaningful differences due to time between data collection and the stringency of data pruning. As such results obtained using our dataset and reported in reference 37 must be considered in isolation from one another. To account for this, we reproduce the MPNN models found and test performance on our dataset to provide a real comparison, which is discussed in more detail below.

Data Processing:

The raw spectra obtained from external sources had variable resolutions ranging from 0.1 cm\textsuperscript{-1} to 10 cm\textsuperscript{-1}. AIST spectra were obtained as images and subsequently digitized using
pixel mapping to obtain numeric spectra. As such, different graph and axes sizes resulted in variable spectral resolutions due to different numbers of pixels in the x-dimension. After digitization, all spectra were processed using simple linear interpolation to produce arrays of 1,800 elements with a step size of 2 cm\(^{-1}\), representing spectra ranging from 400 cm\(^{-1}\) to 4,000 cm\(^{-1}\) to ensure homogenous data spacing. For spectra that did not cover this entire range (i.e., termination at 3,800 cm\(^{-1}\)), the spectra were interpolated to their endpoints and then padded with the “Not a Number” (NaN) value so that they were consistently sized. In all subsequent loss, normalization, or accuracy computations During model training and evaluation, NaN values were excluded using masking operations.

Solvent absorbance arising from condensed phase have well defined absorbances arising from solvent infrared frequencies. Spectra measured in CCl\(_4\) have large absorbances from 696-850 cm\(^{-1}\) and 1,500-1,600 cm\(^{-1}\). Spectra measured in nujol mull contain broad absorbance from 2,750 to 3,000 cm\(^{-1}\). While these peaks are consistent and theoretically learnable by a machine learning model, the relative concentration of the solvent/solute is not known across all spectra, leading to inconsistency in relative peak sizes and intensities. As such, solvent absorbance frequencies were replaced with NaNs and excluded from normalization steps and accuracy measurements.

All spectra were corrected using a baseline correction procedure (discussed in more detail below). Then, all spectra were normalized to sum to unity. While the goal of this baseline correction procedure is to standardize spectra and reduce variance due to baseline noise, it does not eliminate solvent effects from spectra. As such, a description of solvent is still needed for model learning, which is implemented as a global solvent node. The implementation of this method will be discussed in the Feature Engineering section.

Model Architecture:

Transformers:

Graphormer was chosen as our model architecture to address the known limitation of the MPNN architecture and because of its recent success across a wide variety of domains.\(^{46,56-59}\) Transformers are type of encoder that are successful because they learn contextual relationships between input embeddings which are aggregated to a learned description of the input. This is achieved using a self-attention mechanism, where the model assigns varying degrees of importance to between different items in the input sequence. These importance values can then be utilized to generate contextual hidden representations which incorporate long-range contextual dependencies between items.\(^{46,47,60}\) Transformers are most well known for their applications in Natural Language processing (NLP), where the attention mechanisms and grasp of context allows models to parse the recombinant meaning of sentences and paragraphs.\(^{60}\) Before the advent of Graphormer, transformers were thought to perform poorly on graph structured data, despite their broad success in NLP tasks as embodied by the

---

**Figure 1.** Schematic illustration of the Graphormer-IR architecture in which a molecular SMILES string for picolinic acid is mapped into a mathematical graph object by the Graph Encoder (1) with node and edge encodings. In this example, atomic number is the only node feature encoded and edge encodings are the combinatoric representation described in the text. Molecular graphs are then passed to Graphormer (2), whose centrality, spatial, and node encodings are illustrated alongside the model architecture. The depiction of Graphormer is inspired by reference 46. The learned molecular representation generated by Graphormer is then passed to out MLP (3), from which the final IR spectrum is predicted.

---

https://doi.org/10.26434/chemrxiv-2023-f38b5 ORCID: https://orcid.org/0000-0001-7365-1558 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0
BERT (Bidirectional Encoder Representations from Transformers)\textsuperscript{60} and GPT (Generative Pre-trained Transformer)\textsuperscript{61,62} models.

Graphormer extends the success of Transformers to graph applications, addressing the unique challenges of complex relational data. By building upon the success of transformers in sequential data processing, Graphormer presents an exciting advancement in deep learning architectures tailored for graph-structured data, addressing the unique challenges posed by complex relational data. Many examples of graph-structured data, such as social networks, molecular structures, or recommendation systems (e.g., Youtube, Spotify),\textsuperscript{63} necessitate the modeling of intricate structural relationships between entities. Graphormer accomplishes this by integrating multi-headed self-attention mechanisms, where several attention heads allow the model to selectively attend to each node and its relationship to its neighbors.

**Dot Product and Multi Headed Attention:**

Attention mechanisms are at the heart of the success of transformers, where providing a device to evaluate the relative importance of nodes in a graph. Traditional dot product attention consists of a procedure where the layer inputs are projected into a set of queries (Q), keys (K), and values (V) representations. Using these projections, new attention-weighted representations can be calculated using equation 1:

\[
\text{Attention}(Q, K, V) = \text{softmax} \left( \frac{QK^T}{\sqrt{d_k}} \right) V
\]

Where \( d_k \) is the keys dimension, \( T \) is the transpose operation and the softmax function serves as a scoring function to map the logits (\( QK^T \)) into probabilities that serve as the attention values. Multiplying by the values (V) serves to produce the scaled embedding values.\textsuperscript{68}

Graphormer utilizes a multi-headed attention mechanism akin to the original transformer architecture, where \( h \) attention heads calculate parallelized attention scores. First, input data is decomposed into the query, keys, and values which consist of \( h \) sets of learned linear projections. Attention scores are then calculated in parallel for each head and then concatenated and projected back into embedding space to produce the final unified score (shown in the centrality encoding in Figure 1).\textsuperscript{46,47} Multi-headed attention is thought to improve on dot product attention because of the model’s ability to simultaneously different types of relationships in the input that govern relative importance.\textsuperscript{68}

**Talking Heads (TH) Attention:**

It is important to note that in multi-headed attention, each head acts as an independent entity in the evaluation of relationships in the data. If each attention head can be thought of as an “expert” developing a theory for causal relationships between features and prediction, each expert is locked in a separate room independently processing the same data. The final attention-weighted representations are averaged versions of all the experts’ conclusions.

In 2020, Shazeer et al. of Google implemented an improvement to multi-headed attention known as Talking Heads (TH) Attention. The modification is simple, where a learned linear projection across the attention heads dimension is performed on the attention- logits vector (\( QK^T \) in equation 1)) before evaluating the softmax function.\textsuperscript{68} As the name suggests, this allows the attention heads to “talk” prior to generating the attention scores. Instead of making decisions as completely independent actors, there is discourse between the experts and the attention scores become a composite of the intermediate opinions of each individual head, instead of just the average of all final decisions. In their paper, Shazeer found that their modification provided meaningful improvements on transformers studying language comprehension and question answering tasks in the BERT\textsuperscript{69} and ALBERT\textsuperscript{70} models.\textsuperscript{68}

Talking heads attention has been notably implemented by Mao et al., who used to great success in GraphNovo, a tandem mass spectral (MS\textsuperscript{2}) peptide sequencing algorithm built on the original Graphormer architecture.\textsuperscript{59} We test the talking heads attention using the method of Shazeer et al.\textsuperscript{68} and a modified version of the GraphoNovo code found in reference 60.

**Structural Encoding:**

Graphormer extends the Transformer architecture to graphs. However, due to the linear sequence limitations of Transformers, it is crucial to inject graph structural information into the model operators. To encode structure of the graph, Graphormer proposes two structural embeddings: Edge encodings \( c_{ij} \) and Spatial encodings \( b_{\phi(v_i,v_j)} \). Both encodings are injected during the Graphormer self-attention mechanism, defined as:

\[
A_{ij} = \frac{(h_iW_Q)(h_jW_K)^T}{\sqrt{d}} + b_{\phi(v_i,v_j)} + c_{ij}
\]

Where \( W_Q \) and \( W_K \) are projection matrices representing query and keys projections, \( h \) and \( h \) are the hidden representations of node \( I \) and \( j \), \( d \) is the hidden dimension size. Spatial encodings, represented as \( b_{\phi(v_i,v_j)} \), is a set of learnable scalars indexed by the function \( \phi(v_i,v_j) \). In Graphormer-IR, this function is chosen to be the shortest path between two nodes in the graph \( (v_i,v_j) \), assuming that they are connected.

Since Graphormer’s global receptive field means that all nodes attend to all other nodes in graph,\textsuperscript{46} Graphormer can adapt \( b_{\phi(v_i,v_j)} \) to incorporate the relationships and weightings which depend on graph structure (e.g., nodes that are closer may provide more meaning than those that are far away). This is in stark contrast to the MPNN architecture, where node/edge level information is only passed a finite number of steps away from its starting position.\textsuperscript{46}

Next, Graphormer also utilizes a novel edge encoding scheme \( c_{ij} \) where edges are considered in the correlation of node features. For each ordered node pair \( (v_i,v_j) \), Graphormer determines the shortest path between them and compute an average of the dot products of all the edge features and a learnable embedding along that path:

4
\[
c_{ij} = \frac{1}{N} \sum_{n=1}^{N} x_{n}(w_{n}^{E})^{T}  \tag{3}
\]

Where \(x_{n}\) is feature of the \(n^{th}\) edge and \(w_{n}^{E}\) is the \(n^{th}\) weight embedding. \(^{46}\) This amalgamation of self-attention and positional encodings empowers Graphormer to excel in a variety of graph-related tasks due to its ability to capture the complexities of graphs and extract highly expressive representations from graph-structured data. \(^{46}\)

**Spectral Loss Function:**

In their study, McGill et al. showed that the spectral information divergence (SID) outperformed other loss functions in the prediction of IR spectra including mean squared error (MSE), root mean squared error (RMSE), and target-weighted mean squared error (TMSE). \(^{37}\) As such, we choose to implement this loss function into the Graphormer-IR framework as well. SID is a logarithmic measure that was develop by Chang\(^{41}\) for use a spectral divergence measure and is defined as:

\[
SID(y_{\text{pred}}, y_{\text{target}}) = \sum_{i} y_{\text{pred},i} \ln \left( \frac{y_{\text{pred},i}}{y_{\text{target},i}} \right) + \frac{y_{\text{target},i}}{y_{\text{pred},i}} \tag{4}
\]

where \(y_{\text{pred}}\) and \(y_{\text{target}}\) are tensors containing intensities in the predicted and target spectra after normalization. Since logarithms require that all values be greater than zero, all spectra in our model are clipped such that the minimum intensity value in a spectrum is \(10^{-4}\). Prior to calculating the loss, model predictions are normalized to sum to unity, and all NaN values are excluded from the calculation of the loss. The SID metric is defined such that the sum of the absorbances must be normalized to unity in each individual spectrum. \(^{37}\) In training, the SID value is scaled by a factor of 1,000 to help manage gradient overflow in the first few training epochs.

As noted in the baseline correction section above, the multiplication factors of \(y_{\text{pred},i}\) and \(y_{\text{target},i}\) in the loss function provide innate scaling such that incorrect prediction of large absorbances will be more severely punished than incorrect prediction of values near zero. While this effect is desirable when examining peaks, an unintended consequence is that for broad regions of moderate absorbance, like from solvent baselines, the sum of correct large number of moderate intensities will contribute meaningfully to the composite SID value. As such, without baseline correction, models would tend to learn solvent baselines, which would inflate the prediction scores of “messier” spectra.

**Spectral Information Similarity Score:**

To consistently evaluate the success of a model in measuring the extent of similarity between two spectra across a set of predictions, a consistent scalar metric is needed. Once again, we borrow a method known as Spectral Information Similarity (SIS) from the work of McGill to maximize the comparability of Graphormer-IR with respect to Chemprop-IR. \(^{37}\)

Test predictions in our model are first normalized such that the fingerprint region (600-1,500 cm\(^{-1}\)) sums to unity. This procedure was performed in training the Chemprop study, but we found that it decreased performance and interfered with model gradients. The goal of this normalization is to emphasize the weight of the fingerprint region in evaluation by increasing the relative intensity of peaks within this region. To reproduce this scaling effect, during evaluation of fully trained models, the fingerprint regions of all spectra are first normalized to sum to unity. Then, both predicted and target spectra are smoothed with a gaussian convolution with a standard deviation of 15 cm\(^{-1}\), which is comparable to typical peak-to-peak distances in IR spectra. The goal of the gaussian convolution is to provide some tolerance with respect to minor deviations or noise in both the model predictions and the ground truth spectra themselves. \(^{37}\) While this step is not necessarily needed due to the inclusion of the 1D-CNN, it was done to ensure internal consistency between Graphormer-IR and Chemprop-IR. Spectra were then normalized to globally sum to unity and SID scores are calculated using equation (2). In both models, SID was calculated using wavenumber values from 500 cm\(^{-1}\) to 4,000 cm\(^{-1}\), since the lowest possible vibrational frequency in the dataset, C-I stretching, is known to occur between 500-600 cm\(^{-1}\). \(^{37}\) SID is then converted to SIS by rescaling the score to a range of 0 to 1, as shown in equation 5:

\[
SIS(\bar{y}_{\text{pred}}, \bar{y}_{\text{target}}) = \frac{1}{1 + SID(\bar{y}_{\text{pred}}, \bar{y}_{\text{target}})} \tag{5}
\]

Where \(\bar{y}_{\text{pred}}, \bar{y}_{\text{target}}\) represent the predicted and target tensor after the Gaussian convolution and re-normalization procedure.

SIS provides an excellent metric for qualitatively assessing the quality of model predictions. A SIS value less than 0.4 indicates a spectrum that has little to no predictive value. An SIS value from 0.40 - 0.70 will have loosely predictive qualities where regions of absorbance are relatively consistent for major peaks. \(^{37}\) For predictions with SIS of 0.70 - 0.85, the peak predictions of major peaks have improved, and smaller features are inconsistently predicted by the model. For SIS = 0.85 - 0.95, the predictions of both major and minor peaks are good, with some small inconsistencies in the intensity profiles. For the range >0.95, the predicted spectra are almost perfectly predicted in a detailed manner, with only slight deviations from the ground truths. \(^{37}\)

**Baseline Correction**

While peak exclusion of solvent matrices addresses the acute peaks introduced in condensed-phase spectra, it fails to address the systematic change to the baseline absorbance profile from variable solvents, instruments, and impurities. As shown in figure 2, sometimes spectra measured in the same condensed phase can have wildly different baseline noise profiles, meaning that baselines may not be consistently learnable.
While originally designed for Raman spectroscopy, this method minimizes the baseline overfitting that traditional asymmetric least squares fitting suffers from by only considering the smoothness constraint at the second derivative and is equally suitable for IR spectroscopy because of the rising baseline noise suffered by both spectroscopies. The KBr disk IR spectra of 2-((pyridin-2-ylamino)methylene)malononitrile in figure 2B has a large rising baseline from 2,000-4,000 cm\(^{-1}\) likely arising from water absorption in the salt disk, and is corrected appropriately by the algorithm. Figure 2A shows a spectrum with minimal baseline noise, which is relatively unaffected by the correction procedure.

We believe that the baseline correction procedure results in a systematic deflation of scores relative to uncorrected spectra because of the reduction of baseline as discussed above. While lowering scores seems counterintuitive, by reducing spectra to only their constituent peaks, we provide a more direct measure of the “success” of a model at predicting infrared spectra instead of a convolution of predicting variable noise.

**Feature Engineering:**

For large models, a traditional deep learning approach involves providing models the general information needed to understand the problem instead of providing human-understood heuristics and quantities, allowing the model to infer relevant characteristics from context, relative node importance, and data structure. In doing so, models can develop learned representations without being bound to the same human-understood descriptors, allowing for more flexible machine understanding and improved expressiveness. Atomic number is the prototypical “minimal” descriptor for molecular graphs because it provides sufficient information for a model to infer other characteristics when considered in the context of graph structure and edge encodings.

This is best exemplified by hybridization, a commonly used node encoding in graph neural network descriptors that could prove useful in the prediction of infrared spectroscopy. An atom’s hybridization can give vital insight into bonding angles and distances, which are known to directly correlate with the energy of a molecular fragment’s bending and stretching frequencies. But could a description of hybridization prove redundant by a sufficiently expressive neural network? For instance, a node labelled as being carbon and is connected to 4 other nodes in a graph, that node will always be SP\(^3\) hybridized and will have corresponding bonding angles and lengths. By explicitly describing atomic hybridization, this may provide a label that may unnecessarily shoehorn the model into an association with the human-label that may not be flexible enough for the complexities of real molecular environments. In the set of all possible SP\(^3\) carbon environments, the constituent bonding partners could drastically change the bond lengths and angles. This variance may be better suited to a learned understanding of hybridization, wherein a model can derive its own understanding of the bonding environment of the 2D chemical graph.

---

Figure 2. (A) Baseline corrected infrared spectra for ((5-Isopropyl-2-methylcyclohexyl)sulfonylmethyl)benzene measured using a pressed KBr disk. The red curve illustrates the original spectrum, the dotted red line is the baseline calculated using the method described in the work of He et al. \(\lambda = 3.0 \times 10^5, \lambda_1 = 1.0 \times 10^{-3}\) and \(p = 0.04\), and the black spectrum is the corrected spectrum. (B) The analogous baseline corrected IR spectra of 2-((pyridin-2-ylamino)methylene)malononitrile. These spectra have not yet been normalized to sum to unity.

Qualitatively, it was also observed in model training that noisier data (e.g., a higher solvent baseline) would often yield higher accuracy/lower loss values. This is because models learned to propagate noisy baselines into predictions because the loss function would reward higher intensity values (seen in the 2,000-4,000 cm\(^{-1}\) range in figure 1B). While the intended effect of the loss function is to reward correctly predicted high-intensity peaks, if there are enough noisy spectra in the dataset, the model will conversely learn to reward large noise baselines. We believe that this effect was observed in the MPNN study (no baseline correction), where predictions of condensed phase spectra that had higher average baseline noise performed better on average than gas phase spectra.\(^{32}\) To account for this loss function bias, spectral baseline correction was applied universally to correct for different experimental conditions and baseline noise. This correction was performed using the improved asymmetric least squares method described by He et al.,\(^{73}\) and implemented in Python in the work of Haack et al.\(^{74}\) with baseline parameters \(\lambda = 3.0 \times 10^5, \lambda_1 = 1.0 \times 10^{-3}\) and \(p = 0.04\).
Conversely, if a human-understood descriptor is accessible to a model’s understanding in the embedding space, it may provide the model with a more knowledgeable embedding initialization. In terms of hybridization, consider the case where hydrogens are implicit in a graph, described as a node level feature from 0 to 4 that corresponds to the number of bonded hydrogens to that atom. If a carbon is connected to three carbons and has one implicit hydrogen, it is readily apparent that that node will be as equally SP^3 hybridized as the previous system and will have comparable vibrations for any C-C bonds. However, the relative equivalence of these systems may not be evident to the model, and it may take several epochs to learn the similarities in these two graphs, which could be perhaps better spent discerning more subtle chemical interactions. This time spent learning could be eliminated by providing that initial description of hybridization, which has the potential to improve model scores by starting from a more sophisticated initial description. Explicit labels of hybridization may also prove helpful in discerning more complicated cases that break conventional bonding rules, like in the expanded octets of sulfur and phosphorus that are SP^3D and SP^3D^2 hybridized and bonding motifs may be less intuitive from graph structure.

Given this complex interplay between flexibility and necessary knowledge, to perform feature engineering in a deep learning model is to tow a narrow line between knee-capping a model’s expressiveness and leaving it blind to the realities of the molecular graph. Several different approaches to node and edge encodings were tested to maximize model performance and determine the best balance between these extrema. In feature engineering, the DGL-LifeSci^5 and RDKit^5 packages were used for featurization functions.

**Learned Feature Representations**

In this study the best performing encoding scheme was like the one implemented in reference 77, where graph node features are generated as learned feature embeddings. The graph feature encoder (1) in figure 2 is replaced with a MLP that projects node features into a fixed-size latent space that is the same size as the Graphormer’s embedding dimension. The goal of this encoder is to provide a learned representation of the input features in the graph nodes themselves, such that the model has the maximum flexibility possible for engineering descriptions of local atomic environments while still utilizing human-engineered information. Masters et al.^77 found that the addition or removal of specific features for molecular encoding still provided meaningful changes in model performance. This is likely due to redundant or irrelevant information contained in encodings.

This feature encoder used in this study consisted of two linear layers (projecting from # features → 1,050 → 2,100), with ReLU activation and two dropout layers (p = 0.05). Adding layer normalization was tested but did not improve model performance. To the best of the author’s knowledge at the time of writing, this is only the second time that such an encoding scheme has been used, and the first time learned feature embeddings have been used with Graphormer as well as the first time for multi-class prediction tasks.

**Initial Bond Encodings:**

Chemprop^38 incorporates four bond features from RDKit, represented as one-hot encodings: bond type (single, double, triple, aromatic), aromaticity (aromatic, not aromatic), cyclicity (in a ring, not in a ring), and stereochemistry (None, Cis, Trans, E, Z, Other). These features are concatenated together. In Graphormer-IR, the same bond embeddings are used across all models, except for the aromaticity encoding, which is omitted as it becomes redundant due to the bond type feature. Additionally, a custom descriptor for connections to the global solvent node (connected, not connected) is introduced, which will be elaborated on in the Solvent Phase section. The bond encodings are separate one-hot encodings, and the total bond encoding is expressed combinatorially. This means that every edge in the graph is described by an integer from 0 to 95, where each number corresponds to a specific combination of the bond encodings.

**Initial Atom Encodings:**

Chemprop utilizes a variety of other node features to improve learning including aromaticity (aromatic, not aromatic), number of hydrogens (0 to 4), hybridization (S, SP, SP^2, SP^3, SP^3D, SP^3D^2), explicit valence (0 to 7), atom chirality (none, other, R, or S), total bonds (0 to 7), formal charge (−3 to 3), atomic mass (isotope-weighted mass in amu, scaled by a factor of 100), partial charge (as calculated by the method of Gastegier),^38,76 All of these encodings were included in our work, with the exception of atom chirality. This was in part previous work into deep learning using transformer/MPNN hybrids^77 found that inclusion of a stereochemical description decreased model performance in predictions of HOMO-LUMO gaps due to redundancy. We believe that there is no meaningful reason that chirality will influence the vibrational frequencies of a molecule in enough cases to be represented in the dataset such that it could have a meaningful impact on model performance. Enantiomers are expected to have functionally identical IR spectra, where diastereomers are expected to have slight differences in peak positions. In all molecular graphs, hydrogens are implicit, to reduce the average graph size since transformers are known to perform better on smaller input due to hardware limitations.46

As discussed previously, the simple “atomic number only” approach for node encodings was first tested to determine the extent to which our models could discern infrared spectra with the minimum information possible, using only learned chemical relationships. To incorporate additional node-level features into Graphormer’s encodings, two different methods were tested. First, all integer features (all features except atomic mass and partial charge) were mapped combinatorially to integer classes representing each unique combination of all possible node features, analogous to the method used for edge features. This method provides a naïve approach to describing node features, wherein all possible node environments as defined by the feature set are unique classes which the model can use to distinguish vibrational frequencies. The second method was by way of the learned feature embeddings, where node features are passed through an
MLP in a variety of different combinations to obtain flexible initial node descriptions.

**Solvent Phase**

One way that the Chemprop architecture seeks to overcome the locality limitations of D-MPNNs is in the introduction of global features. After the learned molecular fingerprint is generated, global molecular properties are appended and then processed by the MLP amid spectral prediction. In doing so, the spectra produced by the network can consider the total molecular environment instead of just the sum of local descriptions generated by the MLP, which Yang showed measurably improved model performance across a variety of regression and classification tasks. The most notable global feature using in the Chemprop-IR study was a description of the solvent phase, with a one-hot encoding describing whether an IR spectrum was measured in CCl₄, the gas phase, nujol mull, liquid film, or KBr disks. With this inclusion, the MLP can theoretically learn the approximate baseline profile of each phase and adjust spectra accordingly.

The location of this descriptor begs the question: which neural network in our workflow is best suited to use the solvent phase information? Are the GNNs generating a customized generalizable learned molecular fingerprint whose complexities unlocked by the MLP? Or is the MLP simply providing a projection with minor logical inference of the richly expressive embedding generated by Graphormer? Preliminary experiments of Graphormer-IR found that in this context, relatively high model performance could be retained with only a single-layer MLP, strongly supporting the latter conclusion in this use case.

Given this understanding, we chose to include the solvent phase in the graph structure itself, with the assumption that the encodings generated by Graphormer will more robustly describe the solvent environment. To this end, we introduce a global solvent node to all graphs that provides a distinct encoding for each of the solvent phases without any other node information. This node was connected to all other nodes in the graph to propagate solvent information. An additional edge feature was included that would encode whether an edge was connected to the global node.

At the time of writing, the authors believe that this is a novel use of a global node has been used to describe molecule-wide properties in GNNs. By including this feature in the innate graph, we believe that solvent effects can be considered in the highly expressive embedding space generated by Graphormer. Solvent effects can be considered with context within the graph itself instead of just uniformly applied to the MLP. This will allow the model to more robustly consider if specific functional groups interact more with the solvent phase, if these solvent absorbances disproportionately impact certain wavenumber regions or if certain solvent phases commonly introduce impurities into IR spectra.

The idea of this global node should not be limited to only expressing the solvent phase. Any global feature that could be expressed as an integer class can very readily be expressed as a binary node level feature that is only expressed in this global node. Float features could theoretically also be expressed in a global node using the learned feature embedding technique discussed above, where all non-global node values of that float feature are consistent (zero, negative, or a large positive value depending on the domain of the global feature). In doing so, global features can benefit from the context and relative importance provided by transformers, instead of being limited by the interpretation of MLPs.

**Comparison to Message Passing Neural Networks:**

The only other large-scale attempt to predict infrared spectra from molecular structure using deep learning has been performed in the work of McGill et al. As such, it is desirable to have a true comparison of these model scores to determine if the results achieved in this work are state-of-the-art. However, due to the fundamental differences between the dataset used in the original work and this work (constituent spectra, baseline correction, etc.), model scores reported in the work of McGill et al. are apples with respect to the oranges reported in this work. To overcome this hurdle and to have some measure of relative model quality, the Chemprop-IR source code included in reference 37 was installed. In the original study, DFT-computed infrared spectra were predicted using the same architecture as a test of transfer learning, and those spectra were made available in the published work. With advice from the original author (specific thanks to Dr. James McGill), identical train/test splits of the computed IR spectra were generated for Chemprop-IR. Using these splits, MPNN models were trained, and the test scores reported in reference 37 were successfully reproduced.

Our experimental dataset was then restructured such that the format was compatible with the MPNN infrastructure. Five training, validation, and test splits were generated that were identical to those used in the training of Graphormer-IR. Models were trained using the hyperparameters of the most successful models reported in reference 37 and are available in table S2. Any deviation from the original evaluation procedure made in Graphormer-IR (evaluation from 500-4,000 cm⁻¹, 15 cm⁻¹ Gaussian convolution) was also propagated back into the Chemprop-IR architecture. Only single model performances were considered for both Graphormer and Chemprop. MPNN models were tested using two sets of node-level features: only atomic number, and the complete suite of node features described in reference 38.

**Model Training:**

In this work, we implement an end-to-end network based on the original implementation of Graphormer that was built using the PyTorch Python package with a variety of modifications (see figure 2). Our model consists of a novel graph feature encoder (1), that creates molecular graphs with node- (atom) and edge- (bond) level encodings using the DGL-LifeSci package. These graphs are passed to Graphormer (2), that constructs a learned, dense molecular representation. More details on this architecture can be
found in the work of Ying et al. 66 This dense molecular representation is then passed to the MLP (3), which consists of intermediate linear layers and activation functions. We also tested the inclusion of a one-dimensional convolutional neural network (1D-CNN) as the last layer of the MLP. Pure model outputs suffered from point-to-point noise, since the model does not innately understand that adjacent points are related. By incorporating convolution, the authors hoped to smooth peak predictions at each epoch, theoretically allowing the model to learn more realistic peak profiles through backpropagation.

The best performing model was operated with four Graphormer encoder layers, each with an embedding dimension of 2100. A multheaded attention mechanism was used with 210 attention heads and an attention dropout of 0.10. Adam was used as the optimizer with $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$, 56,250 warmup updates, and 375,000 total updates. The best performing model used a learning rate of $3.0 \times 10^{-8}$, though models with less dense encodings (i.e., only atomic number as a node feature) performed better with learning rates of $6.0-7.0 \times 10^{-5}$. The MLP consisted of 3 linear layers with an output of size 1800, interleaved with Rectified Linear Units (ReLU)67 as the activation function. In cases where a 1D-CNN was tested, the last linear layer produced an output of size 1840, followed by a 1D-CNN with a kernel size of 40. The loss function used in training was the spectral information divergence (SID), which was used in the work of McGill57 and will be further described in the Spectral Loss Function section. In total, the best performing model had 138,700,619 tunable parameters, which represents a very wide but short model architecture. All hyperparameters for the best performing models can be found in tables S2 of the supplementary information.

All models were trained for 250 epochs and with a batch size of 32 and an 80-10-10 (%) training-validation-test split. Some molecules had multiple spectra measured in different solvent environments, but since these spectra could be expressed as unique graphs, all measurements were included. These repeat measurements were restricted to the same split (i.e., test, training, or validation) to mitigate possible data leakage between the training and test sets. All models were trained on a single NVIDIA GeForce RTX 3080 graphics card with 10 GB of video random access memory (VRAM). Training a typical model would take roughly 10.2 hours to complete. One of the major benefits of the machine learning approach to infrared spectral prediction is that once a model is trained, calculation of infrared spectra is extremely rapid. A typical test split evaluation using a trained model consisting of 5338 spectra took approximately 2.8 minutes, which is approximately 31 ms, much shorter than the typical experimental measurement time.

Results and Discussion:

Model Performance:

Using five identical training, validation and test splits each consisting of only unique SMILES strings, Chemprop-IR and Graphormer-IR models were trained and tested. As discussed above, models were trained using only atomic number as a node-level feature, as well as the full suite of node features. Unless otherwise mentioned, all models contained the full suite of edge features discussed above. Test score distributions for the best performing Graphormer-IR split are shown in figure 3.

Figure 3. Violin plot showing the test split of the best performing Graphormer-IR model (full node features) with swarm plot overlain. Each individual point represents the score of a single test prediction and is color coded according to the prediction’s solvent phase. The median of the distribution is drawn as a gray dashed line.

In figure 3, we observe that half of the spectra have an SIS greater than 0.87, indicating accurate prediction of most major and minor peaks, with some discrepancies in the intensity profile. The distribution’s long tail results from lower quality spectra and unrepresented functional groups, leading to poorly learned spectral features. Figure 4 presents example spectral predictions, showcasing representative spectra from each percentile range (10th to 100th), featuring a variety of chemical structures and sizes. Generally, Graphormer-IR shows optimal performance with molecules having large carbon chains and functional groups with conjugated systems, oxygen (e.g., ethers, esters, carboxylic acids, alcohols), and nitrogen (amines, amides, nitriles). This is particularly evident in the spectra from the 60th to 100th percentiles.

Graphormer-IR sometimes was able to succeed when considering molecules with unconventional functional groups where atomic charge distributions are not evident from the molecular structure. For example, as depicted in figure 4, percentile 90, the N=C=S stretching frequencies (ca. 2,050 cm$^{-1}$) that are unique to the relatively uncommon isothiocyanate functional group are successfully predicted with high accuracy. In figure 4, percentile 100, Graphormer-
Figure 4. Sample spectra showing ground truth spectra (black) alongside model predictions (teal) from the top performing test split. Each spectrum is depicted with its calculated SIS, phase environment, and percentile location in the total test split (i.e., percentile 50 indicates that the spectrum is better than the bottom 40% of spectra, but worse than the top 50%).

IR successfully predicts the entire spectrum and peak intensities of a molecule containing a 1,3 dipole, a highly delocalized structure with large charge spread across the \( \text{O} = \text{N} = \text{N} \) moiety. Spectra with poorly represented molecular features like substituted cycloheptatriene (figure 4, percentile 10), bicyclic rings (percentile 20), and highly halogenated species (percentile 30) create situations where subtle molecular interactions are missed, and peaks are not correctly predicted. For example, in the case of 1,2-dibromo-1-chloro-1,2,2-trifluoroethane, the bromination at adjacent carbons that have slightly different chemical environments. This manifests in the ground truth spectrum as a doublet of C-Br peaks (ca. 650-700 cm\(^{-1}\)) that are chemically shifted from each other due to the asymmetry of the chlorine and fluorine substitution at those moieties. The predicted spectrum only contains one C-Br stretching frequency, drastically impacting the total SIS. This missed peak is likely due to underrepresentation of halogenated compounds in the dataset (see figure S2), where the model does not have a large enough sample size to discern the differences in local chemical environment that result from different halogen substitutions.

The IR spectrum of 1,2-dibromo-1-chloro-1,2,2-trifluoroethane illustrates a systematic effect where spectra with fewer vibrational frequencies, such as highly symmetric systems or small molecules, typically have lower SIS values than larger molecules. Each SID value is adjusted based on its intensity, as shown in equation 4, meaning baseline values with lower intensity have a lesser impact on the overall loss. A spectrum with a small number of significant peaks is more adversely affected by incorrect predictions of even a single peak or region. This trend is reflected in the distribution of gas phase spectra performance (see figure 5), because small molecules are more readily measured in the gas phase.
It's important to note that IR spectra measurement techniques inherently possess unique sampling biases in each solvent phase, influenced by the compounds best measured in those phases. Molecules in the gas phase usually have lower polarity, resulting in weaker and fewer IR intensities than those measured in other phases. This leads to spectra with fewer peaks on average, where inaccurate peak predictions cause a more pronounced drop in average SIS values, as seen in the extended tail of the gas phase spectra in figure 5.

Graphormer-IR performed best on liquid film spectra ($SIS_\mu = 0.8868$, see figure 4), which is likely because it is the phase that is least likely to exhibit complex phase absorbance. Spectra measured in nujol mull ($SIS_\mu = 0.8437$) and KBr disks ($SIS_\mu = 0.8338$) had intermediate absorbances likely due to inconsistencies in solvent phases that may not have completed eliminated by baseline correction. Spectra in nujol mull often exhibited solvent absorbance outside of the excluded range discussed above, and KBr disk spectra often had broad regions of absorbance from moisture. Spectra measured in CCl$_4$ had the lowest average $SIS_\mu$ of 0.8138 due to their relative underrepresentation in the dataset (see Figure 5).

![Figure 5](https://example.com/figure5)

**Figure 5.** Breakdown of model performance on test set for each spectral phase of the best performing model ($SIS_\mu = 0.8531$) with associated means ($\mu$) and standard deviations ($\sigma$).

Figure 5 also demonstrates the intended effect of the baseline corrections. Gas-phase spectra represent the “truest” form of the vibrational modes in an IR spectrum. In an ideal world, the performance of the solvent-phase IR spectra after inclusion of the global node would on average roughly correspond to the gas-phase spectra because it best represents the prediction task. The solvent doesn’t fundamentally change the underlying chemistry of an IR spectrum (assuming that solvent peaks are excluded) and the “goal” of predicting vibrational peaks is best represented by the successful prediction of gas phase IR spectra. The solvent phase and its associated baseline should introduce local variance (positive or negative) to the ensemble of test scores. In the case of the original Chemprop-IR dataset where no baseline correction was applied, this variance was purely positive, resulting in SIS scores where the gas phase predictions perform poorest and $SIS_{CCL4} > SIS_{LiquidFilm} > SIS_{Nujol} > SIS_{KBr} > SIS_{Gas}$.

As discussed previously, this strongly suggests that the extent to which the model is learning the solvent baseline introduces systematic positive bias and deviates from the goal of predicting vibrational frequencies. As shown in figure 5, the predictions made by Graphormer-IR on the baseline-corrected dataset do not exhibit this systematic bias, where $SIS_{LiquidFilm} > SIS_{Gas} > SIS_{Nujol} > SIS_{KBr} > SIS_{CCL4}$.

The gas phase performance is near the center of the phase distribution and most closely resembles the overall average ($SIS_{gas} = 0.8567, SIS_\mu = 0.8531$), which was not true for Chemprop-IR dataset ($SIS_{gas} = 0.837, SIS_\mu = 0.864$). As such, we believe that training on the baseline-corrected dataset provides a much better measure representation of the prediction of the vibrational frequencies without prediction of the solvent noise.

**Feature Representation**

The goal of testing different schemes for feature representation is to determine the most effective balance between expressiveness and insight from human knowledge. Single model performance was evaluated on an identical test split for a variety of different encoding schemes that were described in the Feature Engineering Section. Combinatoric edge mappings were used for all models. All test scores are available in table 2.

The first node encoding scheme tested was atomic number only, to provide the maximum flexibility to Graphormer and to provide a baseline relative to models with additional features. Use of atomic node encodings yielded $SIS_\mu = 0.8464$ (n=1) and $SIS_\mu = 0.8449 \pm 0.0012$ (n=5) which is 3.3$\sigma$ better than the average result of Chemprop-IR with the full suite of features ($SIS_\mu = 0.8405 \pm 0.0013$). This result is even more meaningful when put in context of the Chemprop-IR models that were trained on only atomistic encodings, achieving an average score of $SIS_\mu = 0.8326 \pm 0.0015$ (See Table 1). From this, we can conclude that graph transformers are much more expressive and are able to derive sufficient understanding of bond frequencies from atomic number, bond types, and graph structure, using only 36% of the information found in the Chemprop-IR “full features” model.

**Table 1.** Five-fold random cross validation results for the best performing models produced by Graphormer-IR and Chemprop-IR with only atomistic node encodings and the full suite of node features discussed in the Methods section. Both sets of models were tested on identical training, test, and validation splits.

[ORCID: https://orcid.org/0000-0001-7365-1558] Content not peer-reviewed by ChemRev. License: CC BY 4.0
In general, a limitation of both Chemprop-IR and Graphormer-IR predictions was that models were not able to consistently predict peak intensity profiles. This effect can be prominently seen in the lower percentiles displayed in figure 4, where peak locations are often correct, but predicted intensities are often much lower or higher than the ground truths. McGill posited that this effect was due in part due to unknown parameters in measured spectra like analyte and solvent concentration that make relative intensities uncertain. We believe that this effect is further compounded by the Gaussian smoothing procedure, wherein sharp intensities are reduced across the spectrum. While this step is necessary, a less intrusive means to generate realistic peak intensity profiles is desirable.

To attempt to improve these effects, we decided to include partial charge as an additional feature into node encodings. Local electron density plays an important role in the relative intensity of peaks in IR spectra, so having even an approximate value could theoretically help models better judge peak intensities. Since partial charge is a float feature, it cannot be readily mapped as a one-hot class like integer features. In this work, we tested adding the pure partial charge to atomic number, as well as combining the atomic number encoding with a learned representation of partial charge to produce the final node embeddings. The latter of these options was found to perform better, but only produced an $SIS_\mu$ of 0.8399. This performance is worse than the pure atomistic encoding, suggesting that whatever derived understanding of partial charge learned by the model is better for predicting peak profiles and positions. This is likely due to the way in which partial charges was calculated, using the method of Gastegier, which is known to perform well for organic molecules containing C, H, N, and O. However, there are a variety of assumptions baked into these approximate calculations (e.g., assumed partial charge “starting” values) that do not lend themselves to understanding systems with unconventional dipoles (e.g., molecules with thiocyanate, ketene, or nitroso groups) or highly delocalized charges. We believe that the decrease in performance associated with the inclusion of partial charges is because of these assumptions, where Gastegier partial charge does not do a good job at describing unconventional electron densities, shoehorning model predictions into poorer predictions of intensities. The inclusion of more robust partial charges (i.e., calculated using more sophisticated electronic structure methods), may be able to circumvent this limitation.

The second encoding method involved mapping all integer node features combinatorically into numeric class values as discussed in the Methods section. It was observed that the combinatoric mapping of node features ($SIS_\mu = 0.8440$) performed worse than the state-of-the-art model ($SIS_\mu = 0.8531$), as shown in table 2. To understand why this model was limited in scope, one must consider the representation of this embedding space. Naïvely, the possible encoding space as described by the chemical restrictions in the Methods section (e.g., only eleven possible elements) would suggest that the total possible number of possible atomic classes using this method would be 194,040 (the product of the number of classes per features). When this mapping was performed on the 53,528 relatively diverse, small, organic molecules in our dataset, only 98 of these classes are represented, with the largest plurality being associated with carbon environments. This fact reveals two things about this dataset and chemical feature engineering more broadly. First, something that chemists have been claiming for centuries: molecules bond predictably.

### Table 2. Model performance (n=1) as a function of different node encoding schemes.

<table>
<thead>
<tr>
<th>Encoding Description</th>
<th>$SIS_\mu$ (n=5)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Atomic Number</td>
<td>0.8464</td>
</tr>
<tr>
<td>Atomic Number + MLP(Partial Charge)</td>
<td>0.8399</td>
</tr>
<tr>
<td>Combinatoric Node Feature</td>
<td>0.8440</td>
</tr>
<tr>
<td>MLP(All Features)</td>
<td>0.8531</td>
</tr>
<tr>
<td>MLP(All Features + period, row)</td>
<td>0.8523</td>
</tr>
</tbody>
</table>

The second and more critical insight is the set of commonly used integer descriptors for deep learning with GNNs contains immense redundancy. Providing a complete description of all the possible one-hot encodings for a given atom is providing too specific and rigid of a description for the model to learn from. It’s likely that within those 98 classes, there is meaningful overlap between IR frequencies produced from each node. Both SP² and SP³ carbon atoms connected to hydrogens by a single bond will produce very similar C=H bending frequencies that may not be readily differentiable. By defining these cases as different classes, the model will learn to discriminate between these two classes in the output and may not reconcile the common frequencies generated by these contexts. In other words, each node is over-described by providing a hard classification that doesn’t necessarily encode all the subtleties of anharmonicity experimental IR peaks.

The most successful node mapping scheme determined by this study was with complete learned feature generation (MLP(All Features)). By passing all node and float features into an MLP that generates learned encodings for each node, Graphormer-IR can best tow that line between encoding flexibility and application of human engineered information. This scheme achieved a best $SIS_\mu = 0.8531$ in single model performance, and $SIS_\mu = 0.8523±0.0006$ in the n=5 case.

In figure 6A, the measured spectra (black) of 10-bromoanthracen-9(10H)-one is depicted alongside predictions from...
Graphormer-IR (atomic number) in teal, and predictions from the state-of-the-art model using learned feature generation (red) where there exists a difference of \( \Delta SIS_\mu = +0.16 \) between the model predictions. In both predicted spectra, the correct peaks are generally present, but the large discrepancy in \( SIS_\mu \) is caused by differences in the peak locations in both spectra. For predictions made using the full set of node features, Graphormer-IR is much better at predicting the wavenumber of the peaks, drastically raising the SIS. Notable examples are highlighted in gray, including successful prediction of C-Br stretching (ca. 695 cm\(^{-1}\)), C-H bending (ca. 800 cm\(^{-1}\)) and C=O stretching (ca. 1,600-1,700 cm\(^{-1}\)).

Figure 6A shows the difference in wavenumber prediction for C=O stretching, with Graphormer-IR predicting a peak near 1,760 cm\(^{-1}\) while Chemprop-IR predicts a peak near 1,650 cm\(^{-1}\). This difference is not because of peak position, but the successful prediction of peaks not predicted by the atomistic model. At 1,600 cm\(^{-1}\), the atomistic model does not discern the presence of the C=O stretching frequency. By including additional node level information relevant for describing double bonds (e.g., hybridization), the “full features” model can better parse this stretching frequency.

Masters et al. found that the inclusion of period and group number as node features in their learned embedding scheme improved prediction of HOMO-LUMO gaps,\(^{27}\) presumably because they communicate physical trends embedded in the period table (atomic radius, electronegativity, etc.) that are not apparent from only atomic number. We tested the inclusion of these features in our encoding scheme and found that they did not improve predictions of infrared spectra relative to our state-of-the-art model, producing an \( SIS_\mu = 0.8523 \) (n=1) (See table 2).

The best performing Graphormer-IR model (\( SIS_\mu = 0.8523 \pm 0.0006 \)) improvement is roughly 19.7\( \sigma \) better than best performing Chemprop-IR model (\( SIS_\mu = 0.8405 \pm 0.0013 \)). Figure 7 shows several examples highlighting instances in which Graphormer-IR meaningfully improves the predictions made from Chemprop-IR on identical test splits after training on identical training/validation splits. Notable improvements to spectral features are highlighted in gray. In Figure 7A, the predicted gas-phase spectra of (isocyanatomethanetriyl)tribenzene are displayed alongside the ground truth (black) and demonstrated an impressive \( \Delta SIS_\mu = +0.42 \) for Graphormer-IR predictions (red) as compared to Chemprop-IR predictions (green). This large improvement is largely in part due to the successful prediction of the N=C=O stretching frequency at 2,250 cm\(^{-1}\). While Chemprop-IR does predict a minor peak at 2,000, it fails to capture the real peak intensity and location. Graphormer-IR can capture this frequency and reasonably approximate the intensity profile, suggesting that the embeddings generated from Graphormer can more effectively describe three body stretching frequencies than those generated by MPNN frameworks.

In Figure 7B, the spectral predictions for 4,4’-(butane-1,1-diyl)bis(2,6-di-tert-butylphenol) are depicted showing an improvement of \( \Delta SIS_\mu = +0.25 \), largely explained by Chemprop-IR missing the broad absorbance from 3,200-3,500 cm\(^{-1}\). In IR spectra, O-H stretching frequencies manifest as sharp peaks ca. 3,500-3,700 cm\(^{-1}\). When alcohols experience inter- or intra- molecular hydrogen bonding (from the presence of solvents containing -OH or from high analyte concentrations), this peak becomes a broad absorbance from 3,200-3,500 cm\(^{-1}\).4,4’-(butane-1,1-diyl)bis(2,6-di-tert-butylphenol) has two hydroxy groups and due to intermolecular hydrogen bonding (intramolecular hydrogen bonding is also possible but unlikely due to the bulky tert-butyl groups adjacent to the hydroxides), we, the -OH stretching frequency is broad and red-shifted. We believe that Graphormer-IR can discern this interaction in a way that Chemprop-IR cannot because of its global receptive field. Because the embedding from each node considers all other nodes in the graph, the embedding generated from each hydroxide will encode information from each other and will “know” that there is more than one hydroxide groups in the molecule and that hydrogen bonding is possible. Conversely, the D-MPNN in Chemprop-IR uses a depth of 6, meaning that node-level information can only be propagated a maximum of 6 hops away from the initial node. This means that the embeddings for the hydroxide groups in the molecule will not contain information from the other group (see Figure 7B).

As such, the broad hydrogen-bonding absorbance is not predicted. This difference in prediction accuracy exemplifies the richness of the embeddings generated by Graphormer-IR.
Figure 7. Select spectral predictions from Graphormer-IR (red) and Chemprop-IR (green) models trained and tested on identical splits including spectra for (isocyanatomethanetriyl)tribenzene (A), 4, 4'- (butane-1,1-diyl)bis(2,6-di-tert-butylphenol) (B), and 1,2-bis(tetrahydrothiophen-2-yl)disulfane (C), and 1,1,1,3,5,5,5-heptamethyl-3λ3-trisiloxane (D). All predictions are depicted alongside the respective $SIS_\mu$ values, solvent phase, and molecule structure. Graphormer-IR spectral improvements that contribute meaningfully to the change in $SIS_\mu$ are highlighted in gray.

and highlights how architecture novelties allow GNN transformer hybrids to discern much more subtle chemical interactions than MPNNs. Molecules in Figure 7C and 7D show two more examples of how Graphormer-IR can more effectively parse experimental peak positions as well as correctly identify frequencies for more uncommon elements (Si, S) spectra of highly symmetric molecules.

Ablation Studies:

In this work, we performed an ablation study of the features utilized in the final state-of-the-art model on a single test split. The results from this study are available in Table 3.

Table 3. Ablation Study (n=1) for “Full Features” models with novel features implemented in the state-of-the-art Graphormer-IR model.

<table>
<thead>
<tr>
<th>Node Feature</th>
<th>Comb. Edges</th>
<th>Single Edge</th>
<th>Global Node</th>
<th>TH Attn.</th>
<th>1D-CNN</th>
<th>Test $SIS_\mu$</th>
</tr>
</thead>
<tbody>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.8531</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.8503</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.8521</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.8109</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.7006</td>
<td></td>
</tr>
<tr>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>✓</td>
<td>0.8512</td>
<td></td>
</tr>
</tbody>
</table>

The 1D-CNN was not found to improve model performance in the case of learned feature embeddings ($SIS_\mu = 0.8503$, see Table 3). We found that inclusions of the 1D-CNN would slow down the rate of learning in early epochs, eventually resulting in overall lower model accuracy. The ablation study also showed that the inclusion of additional edge features by using the combinatoric method discussed above provided only minor improvements to model performance ($SIS_\mu = 0.8531$) as compared to a model that only included bond type ($SIS_\mu = 0.8521$). The global node was found to provide a very large benefit to model performance, when ablated to the $SIS$ dropped massively to a value of 0.8109. While solvent absorbances were excluded, the surrounding regions to the peak would often have increased absorbance, and as discussed earlier, each solvent would introduce a consistent change to the baseline. While our correction procedure would standardize spectra, it does not eliminate the entirety of the solvent baseline (as can be seen in Figures 4, 6, and 7). As such, the inclusion of a description of the solvent phase in the global node allows models to learn the unique absorbances associated with each solvent phase and increases model performance. This also shows that inclusion of features via a global node is not only functional in a GNN environment but can provide meaningful benefit to model performance when describing global features like the solvent phase.
Graphormer-IR was also trained where all node features were eliminated by setting them to a uniform, constant value. Despite having no encoding for atomic number, Graphormer-IR still achieved a score of $\text{SIS} = 0.7006$. While this performance is greatly lower than all other models, it shows how Graphormer-IR learns greatly from the graph structure and implicit bonding patterns such that it can still discern partial vibrational frequencies from each molecule. Lastly, ablation showed that the Talking Heads attention mechanism had a minor negative effect on single model performance as compared to the best model ($\text{SIS}_\mu = 0.8512$). We believe that this decrease in performance is related to a broader phenomenon related to the number of attention heads used in this work. As described in the Methods section and shown in Table 4, hyperparameter optimization showed that Graphormer-IR models performed best with 210 attention heads. This number is shocking because each attention head represents its own MLP, introducing a very large number of parameters to the model at a high computational cost. Typical transformers tend to use 16-64 attention heads, where the top performing models used by Graphormer in the 2021 Open Catalyst Challenge used only 8-32.\(^{46}\) So why does Graphormer-IR perform so much better with 210 attention heads ($\text{SIS}_\mu = 0.8531$) instead of 50 ($\text{SIS}_\mu = 0.8386$)\

The authors of the text believe the answer to this question lies in the idea of specialization. In a professional kitchen, each sous-chef is assigned a specific task (i.e., soups, fish, dessert) for which they exclusively perform those duties. The kitchen is successful because the compartmentalization of each role allows the sous-chefs to become a local expert and not become overwhelmed by too broad a role. We believe that an analogous idea might apply in Graphormer-IR. At a coarse grain, predicting IR spectra is effectively a classification problem. Unique functional group vibrations result in distinctive peaks that occur at predictable locations in the spectrum, where absolute peak position and width depend on the local chemical environment. As chemists have done for almost a century thanks to the work of Dr. Coblentz, these peak positions can be used to fingerprint a molecule and determine the chemical structure because of the innate predictability of these peaks to within 50-100 cm\(^{-1}\).

With the classification mindset in mind, how many classes would one expect in the dataset constructed in this work (containing only eleven elements)? Using IR peak tables from reference 74, the authors of this text counted and found that there were between approximately 150-200 unique frequencies associated with unique functional groups that could be distinguished in graph structure by Graphormer-IR. The variance in this number comes from the extent to what one considers to be a unique “graph structure”: is the C=O stretching frequency in a $\beta$-lactone part of a meaningfully different class than C=O stretch of a $\gamma$-lactone if they both occur ca. 1,750±25 cm\(^{-1}\)? The important part of this exercise is that the number of attention heads is close to the number of possible unique frequency/graph structure pairs. It is also important to note that attention heads are a mathematical tool used to understand relationships in the input data that aren’t strictly limited to structure to frequency. The number of frequencies is smaller than the total number of heads because there are likely attention heads attending to other relationships related to global graph structure or long-range interactions that aren’t unique to a functional group like hydrogen bonding.

We propose that the reason the model performs better with 210 attentions heads is because attention heads used in the state-of-the-art Graphormer-IR becomes a specialist in that frequency and is responsible for casting attention such that the model can successfully predict that peak alone. When adding together all the attention scores from each frequency specialist, the final prediction made is the sum of all the individual peaks developed by each attention head.

To support this claim with empirical evidence, we developed a subset of our complete dataset in which we only include molecules containing the elements C, H, N, and O. An 80-10-10 train-validation-test split was performed randomly on this subset of the data, once again obeying the restriction that spectra with the same SMILES string can be in the same split. This subset contained 34,382 spectra representing 19,557 unique SMILES strings. Model performance was studied as a function of the number of attention heads, keeping all other hyperparameters constant. Graphormer’s embedding dimension must be divisible by the number of attention heads, so slight modifications to this value was made to accommodate different numbers of heads. The results of this study are depicted in Table 4:

### Table 4. Graphormer-IR performance (n=1) as a function of the number of attention heads. All models here are variations of the state-of-the-art model with learned feature embeddings. Results depicted in the “Only C, N, H, O” column were trained on a subset of the dataset that contained only those four elements. Note that embedding dimension must be divisible by the number of attention heads.

<table>
<thead>
<tr>
<th># Attention Heads</th>
<th>Embedding Dimension</th>
<th>$\text{SIS}_\mu$ (Full Dataset)</th>
<th>$\text{SIS}_\mu$ (Only C, N, H, O)</th>
</tr>
</thead>
<tbody>
<tr>
<td>50</td>
<td>2100</td>
<td>0.8386</td>
<td>0.8504</td>
</tr>
<tr>
<td>100</td>
<td>2100</td>
<td>0.8403</td>
<td>0.8543</td>
</tr>
<tr>
<td>150</td>
<td>2100</td>
<td>0.8501</td>
<td>0.8534</td>
</tr>
<tr>
<td>175</td>
<td>2100</td>
<td>0.8397</td>
<td>0.8626</td>
</tr>
<tr>
<td>210</td>
<td>2100</td>
<td>0.8531</td>
<td>0.8551</td>
</tr>
<tr>
<td>232</td>
<td>2112</td>
<td>0.8529</td>
<td>0.8535</td>
</tr>
<tr>
<td>250</td>
<td>2000</td>
<td>0.8529</td>
<td>0.8549</td>
</tr>
</tbody>
</table>

The first thing to note about these results which is to be expected, is that model performance on the C, H, N, O dataset is higher than the full dataset because those elements are the best represented elements in organic chemistry and our dataset. In the world where the frequency specialist hypothesis is true, the maximum model performance as a function of number of attention heads should be lower for the C, H, N, O split than the full dataset because fewer specialists will be needed for fewer distinct functional groups. As shown in Table 4, that is exactly the outcome that is observed, where the
model trained on the C, H, N, O dataset reaches peak performance with 175 attention heads, almost 40 heads fewer than the complete dataset. We propose that fewer specialists are needed because there are simply fewer unique frequencies and functional groups that need to be attended to in the C, H, N, O data. This change in maximum performance supports the idea of attention heads acting as specialists for specific frequencies for specific functional groups. We believe that at the time of writing that this implementation of transformers is novel, no other work has found that including this number of attention heads provide meaningful model performance via specialization. Previous works have examined how in BERT models, some attention heads can track some individual syntactic dependency types \(^7^8\), but to the best of our knowledge, there has never been a wide scale example of use of specialist attention heads for classification like in this work.

We believe that the lower performance of the talking heads attention mechanism further supports the specialization hypothesis. For the same reason that a kitchen works better with highly specialized cooks that won’t get overwhelmed by sharing tasks that they are not experts at, each attention head is best suited to parsing exactly one relationship. In TH attention, the experts “talk” to each other at intermediate stages and may distract one another from performing their specialized task. An expert on fish won’t have a lot of useful advice for the pastry chef; an expert in N=N=N stretching won’t be useful to an attention head trying to understand C=Cl bending frequencies. By having a variety of specialists working individually on spectral prediction in a multi-headed attention mechanism, the sum of the highly specialized set of opinions performs better than one in which all work together at intermediate stages.

Conclusions:

In this work we have explored how GNN transformers can be applied to the prediction of infrared spectra using only SMILES strings as input. Infrared spectra were obtained from a variety of different online sources that were measured on different instruments and in five unique solvent environments (nujol mull, gas phase, KBr disk, CCl\(_4\), liquid film). Spectra were pruned to only high-quality spectra of molecules containing 11 elements (C, H, N, O, F, Si, P, S, Cl, Br, I), then standardized using baseline correction algorithms, exclusion of solvent absorbance peaks, interpolation, and padding to produce learnable data. Molecular graphs consist of edge features (e.g., bond type, stereoisometry, etc.) and node features (e.g., atomic number, hybridization, partial charge, etc.). Solvent phase was encoded through use of a global node that was connected to all other nodes by a custom edge type. Graphs were then passed to Graphormer, which generates a learned molecular fingerprint that is analyzed by a multi-layer perceptron to produce the final IR spectrum. Using only atomic number as a node feature, Graphormer-IR obtained final test scores of \(SIS_\mu = 0.8449 \pm 0.0012\) (\(n=5\)). To validate these scores, the previous MPNN model, Chemprop-IR, was installed from reference 37 and tested using identical test, validation, and training splits. Chemprop-IR achieved test \(SIS_\mu = 0.8326 \pm 0.0015\) with only atomic number as a node encoding and \(0.8405 \pm 0.0013\) with the full suite of node level features. This proved that with statistical significance (3.3\(\sigma\)), Graphormer-IR (atomic number only) is the state-of-art-model for prediction of infrared spectra.

Additional node features were included using a variety of different schemes to improve model performance. This included combinatoric mapping of integer features and the inclusion of a graph encoder MLP to produce learned node embeddings at the graph level, to provide the most human engineered information possible, while giving the graph encoder the most flexibility in the embedding space. At the time of writing, this is only the second time that such a technique has been implemented, and the first time it has been incorporated in any context outside of single value regression (the previous use case was in prediction of HOMOLUMO gaps). \(^7^7\) A variety of different node features and encoding schemes were tested, but the best performing encoder involved a two-layer MLP that received atomic number, hybridization, atomic mass, number of bonds, explicit valence, number of hydrogens, aromaticity, and Gasetgier partial charge. This version of Graphormer-IR obtained final test \(SIS_\mu = 0.8523 \pm 0.0006\), which is on average 19.7\(\sigma\) greater than the previous state-of-the-art results.

Analysis of these test results demonstrated a variety of ways in which Graphormer-IR outperformed MPNN frameworks. First, the baseline-corrected data allowed Graphormer-IR to obtain scores where the inclusion of solvent data did not inflate results, and that the gas-phase IR spectra provided the most representative test scores as compared to the MPNN study where phase effects resulted in inflation of test scores. Comparison of test predictions showed the Graphormer-IR could more accurately discern anharmonic peak locations, predict uncommon frequencies from more than two atoms (i.e., N=C=O, N=C=S, stretching), and prediction of hydrogen-bonding interactions than the MPNN model could not parse. Ablation studies showed that combinatoric edge encodings and the global node provided meaningful benefits to predictions, whereas the 1D-CNN and talking heads attention did not. At the time of writing, the authors of this text believe that this a use global node to encode global graph features to improve learning is novel. By improving this information at the graph level, the graph encoder can apply the global information with context and relative importance in a way that was not possible in previous approaches and could theoretically be applied to any global information.

It was also determined that by using many attention heads (210), each head likely becomes a “specialist” for each different frequency that could be produced from a chemical structure. By allowing each attention head to focus on only one relationship (and without the discourse of talking heads attention), the sum performance of the model with many highly specialized actors outperforms the approach with fewer heads that try to broadly encode more different types

https://doi.org/10.26434/chemrxiv-2023-f38b5 ORCID: https://orcid.org/0000-0001-7365-1558 Content not peer-reviewed by ChemRxiv. License: CC BY 4.0
of logic. We believe that this approach highlights a new possibility for how to use transformer models in regression problems that are built on coarse-grained classification problems. To the best of the author's knowledge, there has been no other application of large-scale implementation of attention heads to regression problems that involve coarse grained classification tasks that improved model performance.

In essence, this study makes significant progress in precisely predicting infrared spectra through deep learning approaches. Our discoveries underscore that transformer graph neural networks achieve cutting-edge performance. We've highlighted numerous valuable improvements in model architecture and innovative training techniques, including the integration of global feature nodes, specialized attention heads, and learned node feature embeddings. These advancements have notably enhanced the impromptu forecasting of complex chemical interactions in infrared spectroscopy, providing an attractive alternative to laborious ab initio methods. While a step forward, there still exists large room for improvement to model predictions and other areas to explore in prediction of IR spectra. Emergent architectures like GPS++77 which amalgamate the local focus of MPNNs with the global attention of transformers, have already exhibited enhanced performance compared to pure graph transformers. Challenges remain in precisely predicting outcomes for intensity profiles and atypical functional groups. Furthermore, the potential application of graph transformers to related IR techniques like Near-IR (NIR) spectroscopy, Raman spectroscopy, or Infrared Multi-Photon Dissociation (IRMPD) remains underexplored. In those regards, our work is ongoing.

ASSOCIATED CONTENT

Data access statements are available in the supplementary information. Code for this project will be made available online at https://github.com/HopkinsLaboratory/Graphformer-IR

AUTHOR INFORMATION

Corresponding Author
*W. Scott Hopkins, shopkins@uwaterloo.ca

Present Addresses

Author Contributions

CMKS performed data collection, data processing, built the model architecture, performed all experiments, and wrote the manuscript. LH helped conceptualize the project, helped build and debug the model, and aided in manuscript preparation. PT aided in data collection and processing. AH provided experimental guidance and aided in manuscript preparation. JG helped collect data. WSH was responsible for the concept and funding, provided experimental guidance, and aided in manuscript preparation.

Funding Sources

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. 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 Canadian Graduate Scholarship. PT acknowledges financial support from NSERC in the form of an undergraduate research award.

ACKNOWLEDGMENT

The authors would like to acknowledge Dr. James McGill for their assistance in obtaining spectra and reconstructing the MPNNs model used as a benchmark in this study. The authors would also like to thank Zeping Mao for their advice on implementing the Talking Heads attention scheme.

ABBREVIATIONS

REFERENCES


