Predicting Infrared Spectra with Message Passing Neural Networks

Message Passing Neural Networks for Infrared Spectra Predictions

Charles McGill,[†,‡] Michael Forsuelo,[†,‡] Yanfei Guan,[†] and William H. Green[*,†]

†*Department of Chemical Engineering, Massachusetts Institute of Technology, 77 Massachusetts Avenue, Cambridge, MA 02139, USA*

‡*Contributed equally to this work*

E-mail: whgreen@mit.edu

Abstract

Infrared spectroscopy remains an important tool for chemical characterization and identification. Chemprop-IR has been developed as a software package for the prediction of IR spectra through the use of machine learning. This work serves the dual purpose of providing a trained general-purpose model for the prediction of IR spectra with ease and providing the Chemprop-IR software framework for the training of new models. In Chemprop-IR, molecules are encoded using a directed message passing neural network, allowing for molecule latent representations to be learned and optimized for the task of spectra prediction. Model training incorporates spectra metrics and normalization techniques that offer better performance with spectra predictions than standard practice in regression models. The model makes use of pretraining using quantum chemistry calculations and ensembling of multiple sub-models to improve generalizability and performance. The spectra predictions that result are high quality, showing capability to capture the extreme diversity of spectra forms over chemical space and represent complex peak structures.

Introduction

Advances in deep neural networks (DNNs) have improved the state of the art in molecular property prediction. Improved computational accuracy and speed have been seen in a variety of property prediction tasks, ranging from atomization energies to protein-ligand binding affinities.[1–4] In this work, we investigate the application of DNNs to infrared spectroscopy predictions.

Infrared (IR) spectroscopy remains an important analytical tool for molecular identification. This technique has found applications in a diverse array of chemical industries, including both petrochemicals and pharmaceuticals.[5–7] However, spectra predictions are complex, especially in the fingerprint region (500-1500 cm^{-1}). Developing a high-fidelity model, whether by *ab initio* or machine learning (ML) techniques, is crucial not only for molecular identification but for more complex tasks such as molecular discovery and design.[8]

Traditionally, *ab initio* quantum mechanics has served as a preferred tool for IR spectra predictions. Density Functional Theory (DFT) and Møller-Plesset Theory (MP2) have become common for gas-phase predictions.[9,10] Commonly used density functionals include B3LYP and M05.[11] Even so, these methods are not always in quantitative agreement with experiment. Factors such as anharmonicity and finite temperature effects can introduce errors.[12] Scaling factors also play a crucial role in the agreement or disagreement between static DFT calculations and experiment. Variation between prototypical frequency factors can result in a discrepancy of as much as 30 cm^{-1}.[13] Various approaches have included non-uniform scaling factors that depend on local frequency and molecular identity.[14,15] A higher level of theory, such as CCSD(T) or QCISD, can improve some aspects of quantitative agreement. Sophisticated computational software, such as Gaussian or SPECTRO, may be used to capture anharmonicity and Fermi resonances.[16–19] The fidelity of these first principles methodologies, still, remains limited to the completeness and accuracy of the Hamiltonian description. Further, more sophisticated methods require significant computational expense, limiting these methods to small molecule applications.[20] Similar claims follow

for condensed-phase IR predictions. *Ab initio* molecular dynamics (AIMD) has become increasingly common for first-principles predictions of liquid-phase IR spectra.[21,22] However, AIMD suffers from computational expense. The accuracy of the spectra in turn depends on the accuracy of available force fields.[23] Additionally, application of these tools to both gas-phase and condensed-phase spectra requires significant user expertise and involvement.

The shortcomings of accurate first-principles techniques motivate the use of machine learning methodology. Recently, ML models have been applied to IR prediction tasks with promising results. Gastegger et al. have employed high-dimensional neural network potentials (HDNNPs) to accelerate AIMD simulations for IR predictions.[24] The neural networks n2p2 and FieldSchNet facilitate the prediction of infrared spectra by means of predicting molecular potential energies and dipole moments. In particular, the n2p2 architecture utilizes symmetry functions at the input layer, while FieldSchNet makes use of learned radial interaction functions when updating atomic representations.[25–27] Kovács et al. predict infrared spectra with a neural network utilizing Morgan-based fingerprints in combination with the Earth Mover's Distance as a metric and loss function.[28] The above works are a marked improvement from the seminal works of Clerc et al., which employed a feedforward neural network and an engineered molecular representation.[29,30] However, ML techniques have more often been used for the reverse prediction tasks: IR spectra are analyzed as inputs for a classification or regression task.[31–35] Thus, an expanded and rigorous investigation for mapping molecular structure to IR spectra with ML remains a significant innovation.

Deep neural network architectures, such as Message Passing Neural Networks (MPNNs), have shown promise in surpassing traditional techniques from quantum chemistry in both chemical accuracy and computational expense.[36,37] Typical MPNN architectures represent molecular inputs as 2D graphs consisting of node and edge features. After several graph convolutions, the MPNN returns a learned molecular fingerprint which can be mapped to target properties by a feedforward neural network. These learned fingerprints have been shown to outperform engineered representations such as Morgan fingerprints.[38]

In this paper, we investigate the application of Chemprop,[38] a state-of-the-art MPNN, towards the prediction of IR spectra. We propose a novel family of spectra metrics which accelerate the prediction of high-fidelity IR spectra. We find that our spectra metrics can offer much stronger peak resolution and baseline drive than traditional metrics offer. The trained prediction models, spectra metrics, and supporting features are packaged into a software extension named Chemprop-IR.

Methods

The Chemprop-IR prediction model training process uses sampled molecules from Pubchem to generate 85,232 computed spectra to pretrain a model, then refines the parameters in the feedforward neural network using 56,955 experimental spectra. The prediction model is an ensemble of 10 individually trained sub-models. Spectra are represented as vectors with 1801 normalized absorbances (one each 2 cm^{-1} from 400 cm^{-1} to 4000 cm^{-1}).

Datasets

External Data Sources

The IR prediction model is pretrained with computed spectra and subsequently refined by training with experimental spectra. The experimental spectra were obtained from four external data sources: the National Institute of Standards and Technology (NIST),[39] Pacific Northwest National Laboratory (PNNL),[40–42] the National Institute of Advanced Industrial Science and Technology (AIST),[43] and the Coblentz Society.[44] See the Supporting Information for a detailed data access statement. The datasets from NIST and PNNL contain gas-phase spectra. The AIST dataset contains spectra collected in four condensed phases: pure component liquid film, pressed KBr pellet, mineral oil suspension (nujol mull), and CCl_4 solution. The dataset from the Coblentz Society contains each of the five previously noted phases. After data pruning, these sources provide the model with 56,955 spectra and

31,439 unique molecule SMILES. The number of spectra from each source is provided in Table 1.

Table 1: Sources of spectra used in model.

Source	Phase	Spectra
NIST	Gas	8754
PNNL	Gas	503
Coblentz		
	Gas	254
	CCl_4 Solution	43
	Liquid Film	1271
	KBr Pellet	1769
	Nujol Mull	915
AIST		
	CCl_4 Solution	1409
	Liquid Film	7365
	KBr Pellet	17630
	Nujol Mull	17042

Select spectra from the three data sources were removed from consideration for a variety of reasons. Spectra from any source that could not be linked to valid SMILES strings were not included. Included species were limited to those that existed as a single molecule (e.g., excluding mixtures and salts). Species were also limited to contain only the following atoms: C, H, O, N, Si, P, S, F, Cl, Br, and I. The dataset from PNNL contained some species with repeated spectra collected at multiple different temperatures, of which only the spectrum collected closest to 25°C was used in model construction.

Spectra from different sources were processed to conform to a uniform data format. The prediction model uses normalized absorbance coefficients (referred to simply as absorbance later) at 2 cm^{-1} intervals over the range 400 - 4000 cm^{-1}. Source data originally had various levels of measurement resolution: NIST data had resolution of 2 cm^{-1}; PNNL data had resolution of between 2^{-1} and 0.25 cm^{-1}; AIST data had resolution of 4 cm^{-1} in the low wavenumber range and 8 cm^{-1} in the high wavenumber range; and Coblentz Society data had a wide variety of resolutions. AIST and Coblentz spectra were converted from a transmittance basis to an absorbance basis. Absorbance values were interpolated at 2 cm^{-1}

intervals from cubic splines of the raw spectra to provide a uniform data spacing. To be compatible with logarithmic loss functions, each individual value that is negative or zero is replaced with a small positive value (10^{-10}). No additional baseline corrections were made to spectra beyond what may have been performed by the data providers.

Normalization of spectra is necessary to present spectra from different sources on the same scale. Experimental spectra and model predictions of absorbance around 3000 cm^{-1} were observed to have large magnitudes and significant differences between phases, resulting in inconsistent scaling when used to calculate normalization factors. To promote consistent scaling across data sources and phases in model inputs and model predictions, normalization was carried out so that values in the fingerprint region (500 - 1500 cm^{-1}) sum to unity. The normalization considering the fingerprint region is distinct from the whole-spectrum normalization that is required for loss function calculations as discussed later. Computed spectra were constructed with the same bounds, data spacing, and scaling.

Input spectra may have gaps in usable data and still be used by the model. In these experimental spectra, gaps occur when the raw experimental data were not measured over the full range 400 - 4000 cm^{-1} or when the solvent background is heavily absorbing. Values from spectra collected in CCl_4 solution were excluded for the range obstructed by heavy solvent absorbance, 696 - 850 cm^{-1} and 1500 - 1600 cm^{-1}. Values from spectra collected in mineral oil suspension (nujol mull) were excluded in the range 2750 - 3000 cm^{-1}. In training, these gaps are not considered by the loss function or model optimization steps and do not affect the learned model. Spectrum predictions made by the model will not supply values in the phase exclusion regions.

Computed Spectra

In this work, experimental spectra were supplemented with additional computed spectra. These computed spectra were used for pretraining, to give the learned fingerprint encodings the benefit of exposure to a greater diversity of molecule types than are available from the

experimental dataset alone.

Molecules used for generating computed spectrum were curated from Pubchem database, which contains 103 million compounds at the time of writing.[45] We sampled 60,000 molecules from Pubchem. Entries containing only one neutral-charge molecule with molecular weight < 500 were selected. Molecules for which we had experimental spectra were added to this set of 60,000, resulting in 85,232 total molecules chosen for the generation of computed spectra.

We then computed a spectrum for each of the selected molecules. A high-throughput computing workflow was developed using Python to automatically generate spectrum from SMILES strings. The workflow starts from 3D conformer sampling via distance geometry, using RDKit.[46] Conformer searching was then performed under Merck Molecular Force Field (MMFF94s), considering up to 500 potential conformers each.[47] Accessible low-lying conformers with relative energy below 2.5 kcal/mol were selected for further calculations. See the Supporting Information for a distribution of the number of conformers that ultimately contributed to each each spectrum. In order to generate the IR-spectrum, we then optimize the structure and compute the harmonic frequencies for all selected conformers under GFN2-xTB level of theory of Grimme and co-workers,[48] which supports elements through Radon with emphasis on yielding reasonable structures and vibrational frequencies.

The GFN2-xTB spectra are computed making the harmonic oscillator approximation in the gas phase, assuming the dipole moment is a linear function of the atomic displacements and only considering the 3N-6 vibrations. While the calculations do include multiple conformers, they do not include external or internal rotations, anharmonicities, hot bands, or any transitions that violate the $\Delta v = 1$ selection rule. The calculated absorbances are initially for infinitely-sharp stick spectra. To make them more similar to the actual gas-phase spectra we apply a function to broaden the peaks based on the peak position, fitted to the peak shapes observed in the gas-phase experimental spectra. An example of peak broadening is provided in Figure 1. Details on the peak broadening procedure are provided in the Supporting Information. Spectra from different conformers were combined using a Boltzmann

weighting at 25°C.

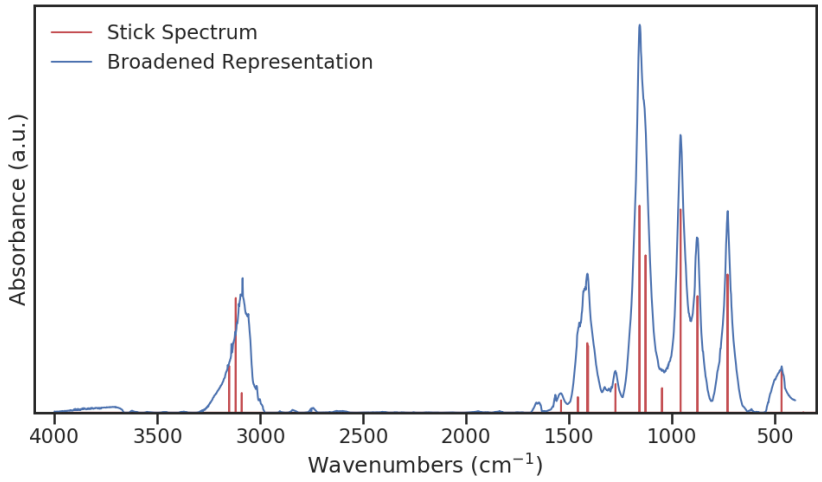

Figure 1: An example of a spectrum formed by using fitted peak distributions to broaden peak locations into a continuous representation for a single conformer of 1,1,2-trifluoroethane.

The computed spectra on their own should not be relied on to be accurate predictions of experimental spectra. They are subject to the shortcomings of the GFN2-xTB method as applied and the fitted peak distributions method. When compared to corresponding experimental spectra using Spectral Information Similarity (SIS, defined in a later section), these computed spectra have an average similarity score of only 0.38. The computed spectra are sufficiently dissimilar to the experimental spectra that they should not be considered as initial estimates of spectra to be fine-tuned. The computed spectrum is better understood as a distinct property that can inform the learning of molecular representations for IR spectra by analogy. These computed spectra fulfill an important purpose by incorporating great complexity. They sample from a diverse chemical space, account for the contribution of many conformers, and represent peaks with realistic peak shapes. These spectra contain enough complexity to be used to learn a method for encoding molecules that captures the relevant characteristics for vibrational spectra.

Network Architecture

We implement an end-to-end learning architecture to predict the IR spectrum from a 2D molecular graph. The end-to-end learning architecture is based on an encoder that can capture the most relevant properties of a given molecule corresponding to the IR spectrum and convert the molecular structure into a vector representation. Relevant feature learning methods, such as graph neural networks (GNN) and language models, have achieved state-of-the-art accuracy in molecular property predictions.[37,38,49–51] Our method builds upon the Chemprop architecture developed by Yang et al.[38] Chemprop consists of a directed MPNN and a Feedforward Neural Network (FFNN) readout. The directed MPNN constructs a learned vector representation, or molecular fingerprint, from a given molecular graph by message passing. Atom and bond features are combined through the message passing steps to generate the learned molecular fingerprint. In the application of directed message passing, the messages and hidden states of the network are associated with directed edges (bonds) instead of nodes (atoms). Directed message passing is thought to minimize the introduction of noise into the graph representation, since the architecture prevents messages being received by the same object in the following iteration. Details of the directed message passing structure may be found in the work of Yang et al.[38]

Chemprop-IR uses an extension of the Chemprop network structure that specializes in the prediction of IR spectra, in the form of 1,801 output elements. Each element corresponds to the absorbance at a fixed wavenumber position in the spectrum. Although we represent the spectra as a series of distinct absorbances, it is important to keep in mind that the absorbances are not independent scalars but the discretized representation of a smooth curve. Usually several absorbances represent a single physical peak, and also a single functional group in the molecule creates several peaks with characteristic relative intensities. To be satisfactory, the predicted spectra must be smooth and respect these correlations. Any scaling or normalization of the experimental data must be smooth across all the absorbances in a single spectrum.

Hyperparameter optimization for the Chemprop-IR architecture was performed initially using the software Hyperopt, followed by manual tuning.[52] The MPNN is operated with a depth of 6 and a hidden size with 2,200 elements. The FFNN is operated with a depth of 3 and a hidden size of 2,200 elements. A dropout ratio of 0.05 is applied to the model network. Adam is used as the optimizer.[53] The loss function is the Spectral Information Divergence, defined in the spectra loss function section. In total, a model trained for the prediction of experimental spectra contains 23,957,601 trainable parameters.

As in Chemprop, the structure of Chemprop-IR first encodes molecules with a MPNN before further processing the output in a FFNN. The initial input to the model is a SMILES string, which is interpreted as a 2D molecular graph. Each bond and heavy atom in the molecule is used to generate a vector of atom or bond features. The considered atom features are atomic number, atomic mass, formal charge, the number of bonded hydrogens, whether the atom is included in an aromatic ring, the bonding degree, the chirality, and the hybridization. The bond features are the type of bond, the stereochemistry of the bond, whether the bond is included in a ring, and whether the bond is conjugated with other bonds. The bond and atom features are used to generate hidden states of size 2,200 at each of the directed edges. In each message passing step, the hidden states are updated according to the directed message passing procedure described by Yang et al.[38] Following the message passing steps, atom vectors are generated from the associated directed edge hidden states. The atom vectors are averaged to form a molecular fingerprint vector of size 2,200.

The phase of the molecule is introduced to the molecular fingerprint vector following message passing. This process is carried out for experimental spectra models where each entry is specified as one of the five supported phases. The phase features are represented as a one-hot vector of size five. This vector is concatenated to the end of the molecular fingerprint vector. For models trained on computed spectra, there is no differentiation between collection phases (all are calculated in gas phase), so the concatenation of phase features is omitted. The molecular vector following this concatenation is size 2,205 for experimental spectra

models and 2,200 for computed spectra models.

After phase feature concatenation, the resulting vector is fed to a FFNN for readout of the model results. For each intermediate layer in the FFNN, the produced vector size is 2,200. The final layer of the FFNN provides an output size of 1,801. An exponential activation function is applied to the FFNN output to enforce value positivity. Output absorbances predicted by the model are normalized to sum the fingerprint region to unity. A separate normalization to sum the absorbances within the whole spectrum to unity is carried out as part of the loss functions and does not change the scaling of the model predictions. The loss calculated at the model output is fully differentiable through the FFNN and the MPNN, allowing weights from both portions of the model to be optimized simultaneously.

Model Training

Models for spectra prediction used an ensemble approach, where ten sub-models were constructed independently and the model output is the average of the outputs of the sub-models. Model training made use of a transfer learning strategy where the model was initially pre-trained using computed spectra before the final training using experimental spectra (Figure 2).

In the initial training, the model was trained using the 85,232 computed spectra. The computed spectra were divided along a 80-10-10 train-test-validation split randomly. Each of the ten sub-models was run independently, with a different set of initialization weights. MPNN weights from the initial model were carried over to the final model training and frozen, holding constant the encoding model learned in the initial model.

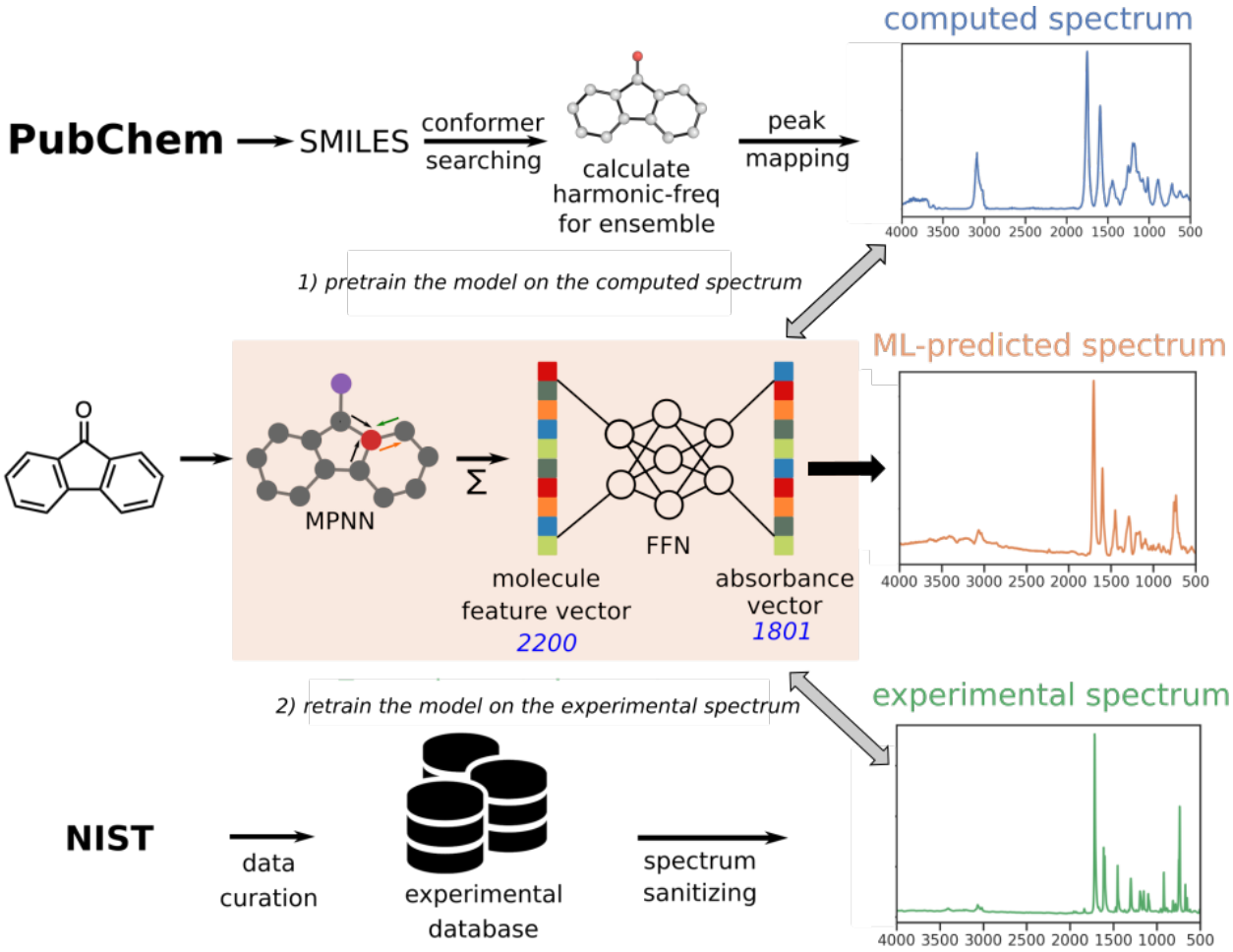

Figure 2: Chemprop-IR model transfer-learning training scheme, using sampled molecules from Pubchem to generate 85,232 computed spectra to pretrain a model, then refining the parameters in the FFNN using 56,955 experimental spectra. Spectra are represented as vectors with 1801 normalized absorbances (one each 2 cm^{-1} from 400 cm^{-1} to 4000 cm^{-1}).

In the final training, the model was trained using the 56,955 experimental spectra. Phase information associated with each of the experimental spectra were also incorporated in model training. The experimental spectra were divided along a 80-10-10 train-test-validation split randomly but with the constraint that all spectra sharing a SMILES representation had to occupy the same split. The MPNN weights from the initial sub-models were frozen and applied to the ten final sub-models. Each of the ten final sub-models were initialized with different FFNN weights. All ten of the final sub-models shared the same random data split.

Spectra Loss Functions

Several metrics have been proposed for use in the study of spectra predictions.[54–56] These metrics can serve as high-performance loss functions for the prediction of machine-learned spectra or be used to assess similarity between predictions and targets via similarity scores.

Spectra metrics include the Spectral Information Divergence (SID) as developed by Chang[54] and the Target-weighted Mean Squared Error (TMSE). Though previously developed as a divergence measure, to our knowledge SID has not been used as a loss function in a machine learning application before. The Spectral Information Divergence may be defined as

$$\mathrm{SID}(\mathbf{y}_{pred}, \mathbf{y}_{target}) = \sum_i y_{pred,i} \ln \frac{y_{pred,i}}{y_{target,i}} + y_{target,i} \ln \frac{y_{target,i}}{y_{pred,i}} \tag{1}$$

where $\mathbf{y}_{pred}$ and $\mathbf{y}_{target}$ are vectors corresponding to the predicted and target spectra. Note that SID can be used to compare any two spectra generally, not just predictions and targets. Further, SID can be interpreted as a symmetric sum of the Kullback-Liebler Divergence.[54] In this work, we restrict the domain of SID such that the components of $\mathbf{y}_{pred}$ and $\mathbf{y}_{target}$ are positive. Note that positivity is enforced by our architecture with output exponential activation as described in the network architecture section. As with the Kullback-Liebler Divergence, the SID operation is defined such that such that the sum of the absorbances is normalized to unity for each spectrum. For consistent treatment, exponential activation and normalization of spectra to sum all absorbances to unity may be carried out with any loss function.

The Target-weighted Mean Squared Error is defined as

$$\mathrm{TMSE}(\mathbf{y}_{pred}, \mathbf{y}_{target}) = \frac{1}{N} \sum_i \frac{(y_{pred,i} - y_{target,i})^2}{y_{target,i}} \tag{2}$$

for N properties.

Note that SID converges to TMSE within a scalar for sufficiently small residuals $\Delta\mathbf{y} =$

$\mathbf{y}_{pred} - \mathbf{y}_{target}$. Observe

$$\begin{aligned}
\mathrm{SID}(\mathbf{y}_{pred}, \mathbf{y}_{target}) &= \sum_i \Delta y_i \ln\left(1 + \frac{\Delta y_i}{y_{target,i}}\right) \\
&= \sum_i \Delta y_i \sum_{j=1}^{\infty} \frac{(-1)^{j-1}}{j} \left(\frac{\Delta y_i}{y_{target,i}}\right)^j \\
&= \sum_i \frac{\Delta y_i^2}{y_{target,i}} + O\left(\frac{\Delta y_i^3}{y_{target,i}^2}\right)
\end{aligned}$$

Spectral Information Similarity Metric

Generation and testing of spectral representations requires the use of a consistent method for judging the degree of similarity between two spectra. In particular, developing such a measure is necessary to assess the success of a spectrum prediction relative to the target spectrum. In the course of this study, we have used a scalar score that we call spectral information similarity (SIS).

The calculation of SIS between two spectra involves the following series of operations:

1. Each spectrum is broadened by applying a Gaussian convolution. In this work a standard deviation of 10 cm^{-1} was used.

$$\breve{\mathbf{y}} = G_\sigma * \mathbf{y}$$

2. Each spectrum is normalized to sum all the spectrum absorbances to unity.

3. The divergence between these two spectra is determined using the SID calculation, as provided in Equation 1.

4. The divergence value is rescaled to a similarity value on the interval (0,1].

$$\mathrm{SIS}(\breve{\mathbf{y}}_{pred}, \breve{\mathbf{y}}_{target}) = \frac{1}{1 + \mathrm{SID}(\breve{\mathbf{y}}_{pred}, \breve{\mathbf{y}}_{target})} \tag{3}$$

The calculated SIS is a single scalar value expressing the similarity between two spectra. On its own, SID is a measure of divergence and peak overlap. Extending beyond a pure measure of spectrum overlap, the Gaussian convolution operation in SIS is used to provide for a level of detected similarity for imperfect spectrum matches. When a predicted peak is shifted or distorted in shape relative to the reference, the SIS score still reflects some amount of similarity in prediction. The Gaussian convolution allows the degree of deviation for peak shifts or distortions to be continuously represented, with the SIS score diminishing as the deviation becomes more significant. The standard deviation for the Gaussian convolution is a variable value that can be adjusted to provide for looser or stricter penalties for deviations. In this case the standard deviation of 10 cm^{-1} was chosen to be on the same order of magnitude of peak shifts reported by other prediction methods while being smaller than the spacing between most peaks. The Gaussian convolution is valuable in assessing the similarity between spectra, but it makes SIS unsuitable for use as a training loss function in the model as it results in aphysical prediction artifacts (see the Supporting Information for an example).

In practice, the SIS provides an easily accessible measure of prediction quality and typically follow trends in prediction behavior. Predictions with SIS in the range 0.40-0.70 are loosely predictive, matching the general regions of absorbance peaks but usually only predicting the locations of major peaks. In the range 0.70 - 0.85 the magnitude and peak shapes of major peaks in the spectrum prediction are improved, with reasonable peak location predictions for some minor peaks. In the range 0.85 - 0.95, the predictions of peak magnitudes and locations are good for most major and many minor peaks. In this SIS range, errors often manifest as predictions of broad peaks in place of a series of sharp minor peaks. In the range above 0.95, the predicted spectra have increasing levels of detailed reproduction of the reference peak shapes. Specific examples of SIS scores are provided in the model results section.

Results and Discussion

Model Performance

Two separate models were trained using the Chemprop-IR framework. The first is a model for the prediction of the computed spectra, as derived from GFN2-xTB calculations. This initial model was trained using a 80-10-10 random splitting of the 85,232 computed spectra generated. The final model is for the prediction of experimental spectra. This model was trained using a 80-10-10 random splitting of the 56,955 spectra obtained from external sources, constrained such that spectra with the same SMILES had to occupy the same split. The final model used MPNN weights that were extracted from the initial model, trained using the computed spectra. Both models are composed of ensembles of ten individually trained sub-models. The success of the models in predicting a spectrum accurately was assessed using SIS. The resulting performance distribution for both models is provided in Figure 3.

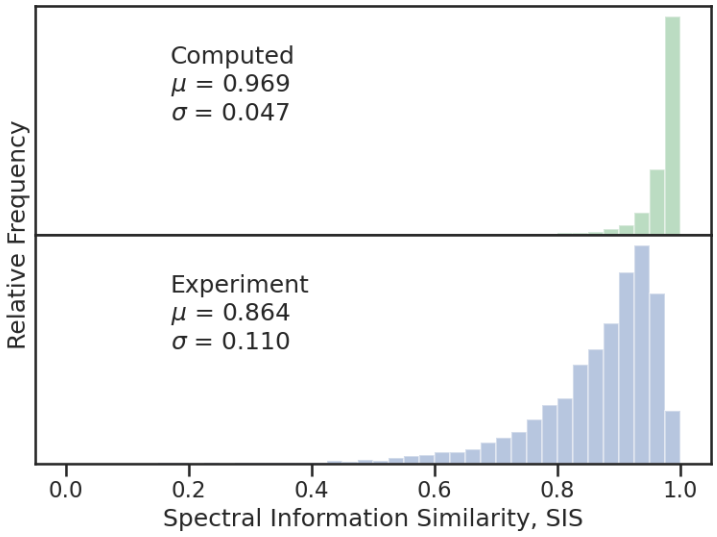

Figure 3: Distribution of SIS score for spectral predictions in the test split for (top) computed spectra and (bottom) experimental spectra. The average value and standard deviation for SIS are provided as μ and σ, respectively.

With an average SIS of 0.969 and a median of 0.983, the initial model for computed

spectra is able to reproduce spectra with its predictions to an exceptional degree. Examples of the model predictions at different percentile locations in the performance distribution are provided in Figure 4. The predictions of computed spectra do benefit from several simplifying aspects of the underlying calculations. Several factors contribute to make the problem of fitting the computed spectra a tractable one for the model: the method GFN2-xTB is a semiempirical method rather than higher levels of theory, only simple harmonic oscillator modes were considered, and the peak distributions were applied using a function that only varies with peak position. The modeled system remains far from simple though. These fits encompass a large number of spectra over a diverse set of molecule types. Each spectrum includes contributions from many different conformers, combined through Boltzmann weighting. The high level of accuracy in the reproduction of the computed spectra is a demonstration of the flexibility and capability of the Chemprop-IR framework.

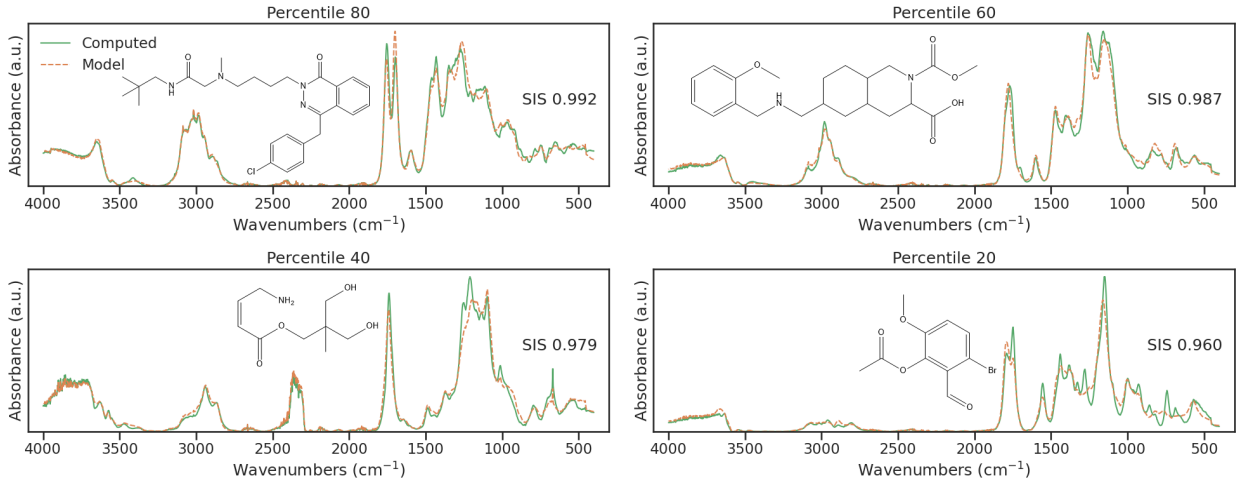

Figure 4: Examples of spectrum predictions in the initial model for computed spectra from different percentile locations in the model performance distribution. The computed spectrum (blue), the model-predicted spectrum (orange), and the SIS score are provided.

The final model for experimental spectra performs very well, with an average SIS score in the test set of 0.864 and a median of 0.896. In these ranges, the majority of predictions are carried out with good accuracy. Examples of spectra at different percentile locations in the performance distribution are provided in Figure 5. These examples show increasing levels

of qualitative agreement through the performance distribution. Roughly two thirds of the predicted spectra exist in the SIS range above 0.85 where the major peak shapes of the spectra are well captured. Half of the predicted spectra have SIS scores above 0.9, corresponding with appropriate recreation of many or most of the minor peaks as well. An alternative version of the final model was also tested where the computed spectra used for pretraining only considered the lowest energy conformer. This alternative model performed just as well as the ordinary model, with an average SIS of 0.863 and a median SIS of 0.895, indicating that this instance of the model is not sensitive to the inclusion of multiple conformers.

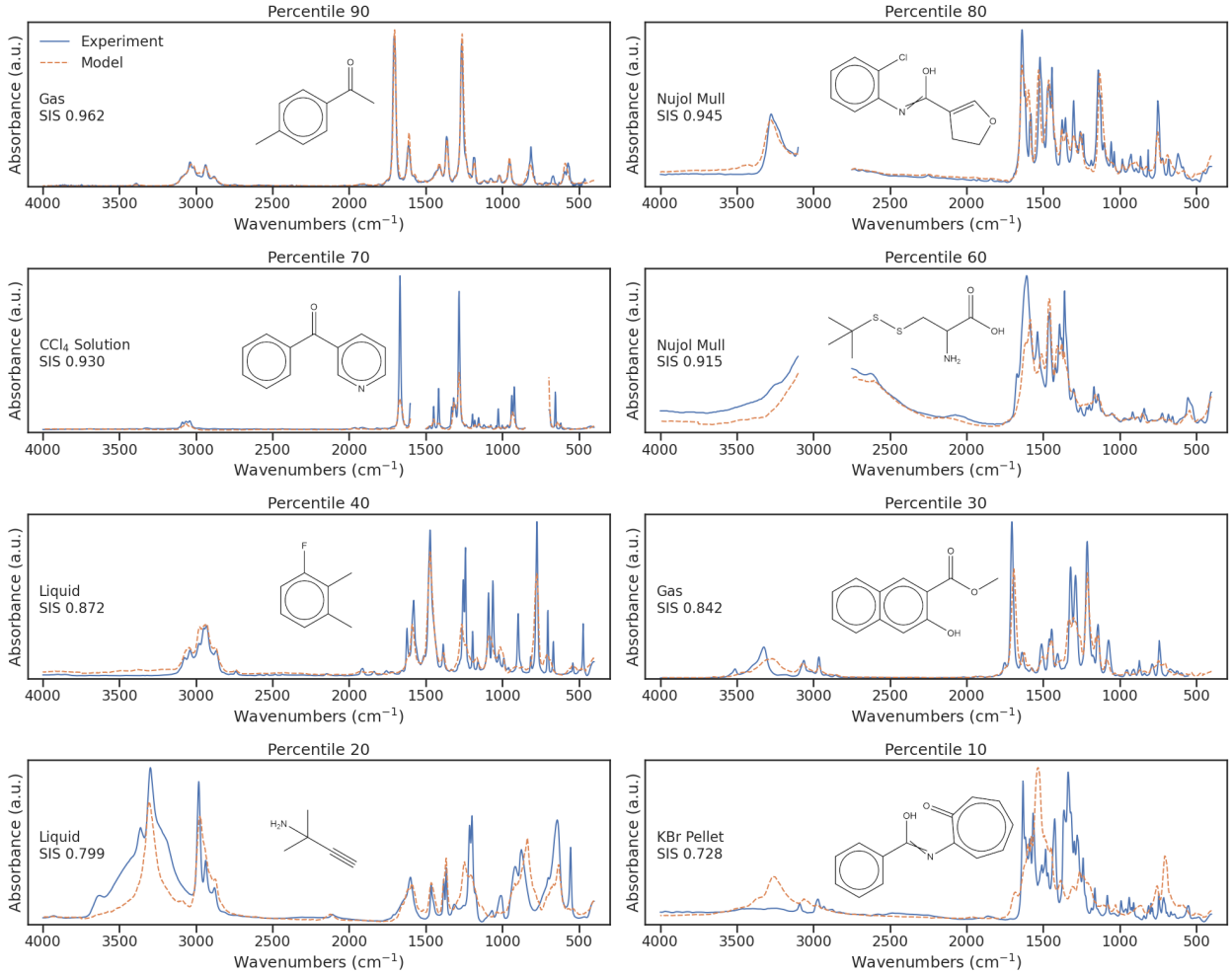

Figure 5: Examples of spectrum predictions for the final model for experimental spectra from different percentile locations in the model performance distribution. The experimental spectrum (blue), model-predicted spectrum (orange), phase of collection, and SIS score are provided.

In model training, the iteration of the model was used that produced the lowest error in predictions of the validation set. The performance on the validation set resulting from this procedure is nearly the same as that of the test set, with an average SIS score of 0.866 and a median of 0.896. The performance on the training set shows evidence of overfitting, with an average SIS of 0.957 and a median of 0.967. Predictions made on molecules in the test set that are similar to molecules in the training set tend to perform better on average. As can be observed in Figure 6, the performance of specific predictions in the test set is positively correlated to the similarity between the prediction molecule and the molecules in the training set. Similarity between molecules was calculated using a Tanimoto similarity between the molecular Morgan fingerprints, using a radius of 6 and a hashed bit length of 2048. The datapoints in the figure are clustered to emphasize the correlation.

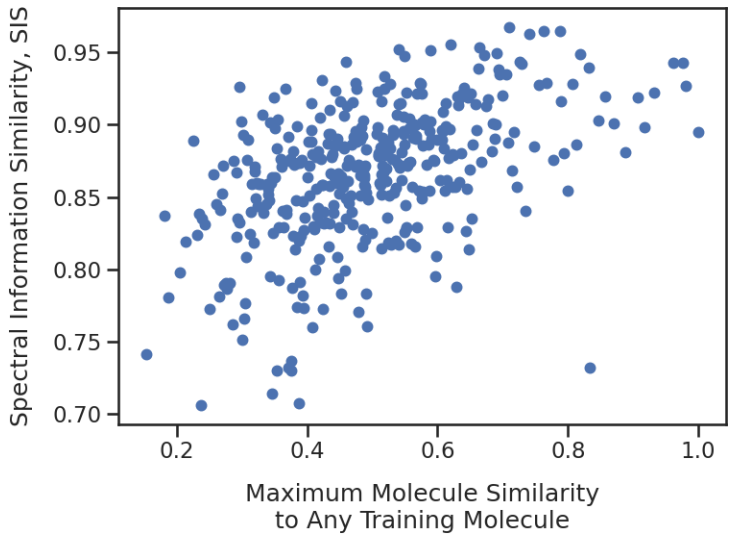

Figure 6: Test set prediction SIS scores related to the maximum molecular similarity between the prediction molecule and the molecules of the training set. Datapoints are clustered and averaged in groups of 10.

The final model for experimental spectra was trained using the five sample phases. Though other sample phases are present in the experimental source data, the other phases beyond the chosen five did not have enough datapoints to support training. The performance distribution for each supported phase is presented in Figure 7, with specific example

spectra at different locations in the distributions provided in the Supporting Information. Additional analysis on model performance across phases may also be found in the Supporting Information.

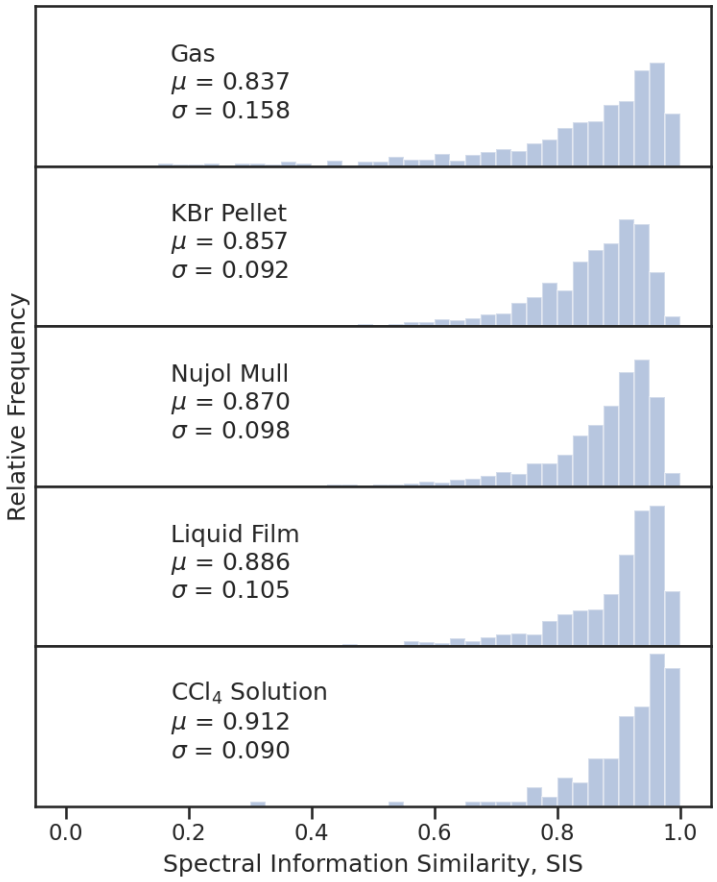

Figure 7: Distribution of SIS score for spectral predictions in the test split, separated according to the sample phase in which the spectrum was collected. The average value and standard deviation for SIS are provided as μ and σ, respectively.

The predictions of gas-phase spectra are worse than those in the other four phases. The average gas-phase SIS appears relatively low at 0.837 as a result of tail and outlier performance. The median SIS score for a gas-phase prediction is good at 0.889. Gas-phase spectra tend to have fewer and more narrow peaks, compared to condensed-phase spectra. The baseline absorption away from peaks also tends to be lower in gas-phase spectra. These features all contribute to a scoring disparity where the overlap measurement in SIS will score

peak shifts more harshly in the gas phase than in other phases. The tail performance of the gas-phase dataset may also reflect a selection bias where the gas-phase dataset contains smaller molecules on average than the other datasets. A prediction for a small molecule containing a difficult-to-predict mode will be scored more harshly than the same mode in a large molecule because the associated peak makes up a larger fraction of the spectrum.

The SIS scores for predictions in KBr pellets and in nujol mull are notably sensitive to how well the phase effects are predicted. KBr spectra sometimes will exhibit a low broad peak above 3000 cm^{-1} due to moisture absorption. The significant absorbance in nujol mull around 3000 cm^{-1} has been excluded from training and scoring of spectra, but broader absorption by the phase in the surrounding region remains to some extent. The magnitude of those phase features relative to the target sample spectrum is not easily predicted and can sometimes affect scores significantly. The relative intensities of the phase features are a function of sample concentration, a variable not available from the data source and therefore not accounted for in spectral predictions.

The prediction of spectra in both liquid film and CCl_4 solution phases is good. The liquid film spectra would be less likely than other condensed phases to have significant phase absorption and interference since it is a pure sample collection. In both cases, there may be a selection bias effect where a higher proportion of the species measured in these phases are types where we may expect the model to perform better, such as in the distribution of molecule sizes or polarity.

For further consideration of the peak shapes and positions generated in predicted spectra, consider the example provided in Figure 8. This example exhibits good prediction of peak locations, showing only minor peak shift and recognition of most peaks. We also see in this example a shortcoming of the model, a tendency to predict peaks lower and broader than they appear in experiment spectra. This tendency leads to the reduced representation of fine details, such as the diminished prominence of the doublet above 1400 cm^{-1} or the loss of the shoulder ridges on the peak at 1280 cm^{-1}. In this respect, the model is challenged

when there are a series of sharp peaks clustered closely together in a spectrum, as this will often result in a low rolling peak that does not separate the different modes present.

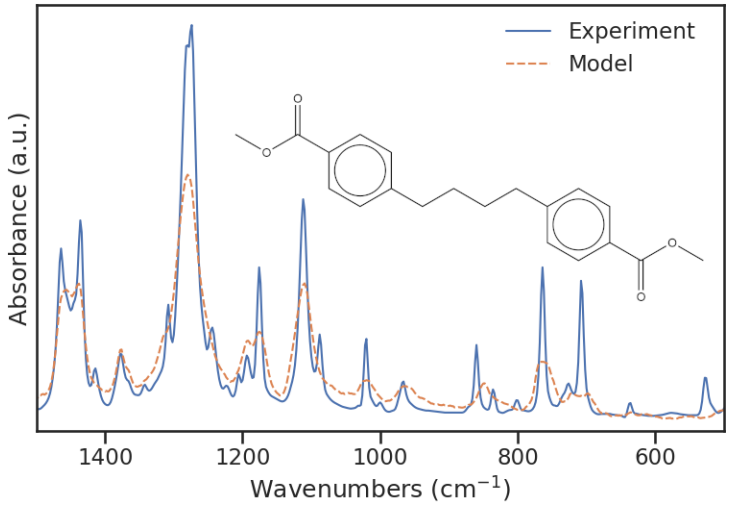

Figure 8: Example of a spectrum fingerprint region in a predicted spectrum.

Benefits of Ensembling

The prediction models being used are ensembles of ten individual sub-models. Each sub-model is trained using a different weight initialization. In the case of the experimental spectra predicting models, the pretrained MPNN weights fitted from training on computed spectra are also different for each member of the ensemble. These different initializations, paired with some amount of randomness in optimizer performance, result in different training trajectories for each model. The complex optimization landscape of the training sets presents a multitude of local minima where the optimized models will terminate. The final form of each model in the end generally results in similar but distinct results. By averaging the results of the ten sub-models into a single ensemble average prediction, the faults of a single sub-model can be outweighed by the central tendency of the population. The approach is analogous to sampling a distribution many times in order to better estimate the population mean. The result is general improvement across the performance distribution of predictions

using an ensemble compared to those using only a single model, as shown in Figure 9. The improvement from ensembling is observed in both the initial model and the final model.

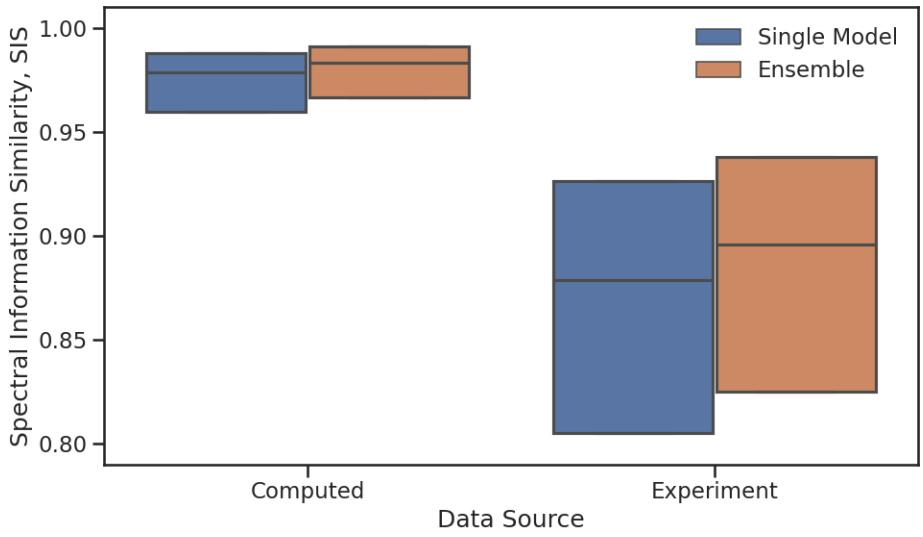

Figure 9: Distribution of SIS score for spectral predictions in the test split, comparing results for an ensemble of ten models and a single trained model. The box plot shows the SIS scores for the inner quartiles of the distribution. The graphical indication of the full range of data is omitted due to the long distribution tail.

In addition to the improved accuracy resulting from ensemble average predictions, the variability observed within an ensemble is a measurement of model and dataset characteristics. As studied in the work of Kendall and Gal, neural network model predictions are affected by epistemic uncertainty, uncertainties inherent to the model.[57] The sources of epistemic uncertainty include such factors as sparsity of training data, rigidity of the model structure, and aggressiveness of regularization. Epistemic uncertainty is distinct from aleatoric uncertainty, which results from uncertainties in the data itself. As demonstrated in the study by Lakshminarayan, the variability observed within an ensemble of predictions can be used as an estimate of the epistemic uncertainty affecting each prediction.[58] Quantifying the ensemble variability for a scalar prediction is a simple matter of calculating the sample variance among the sub-model outputs. For the present case using an ensemble of spectra, the uncertainty was quantified by calculating the SIS between each of the individual model

predictions in a pairwise fashion (45 combinations for a set of ten sub-models) and averaging them together. As with the previously discussed SIS measure, the pairwise SIS calculations include a Gaussian convolution with standard deviation 10 cm^{-1}. A high ensemble SIS indicates that the sub-models return predictions with low variability among them. Likewise, a low ensemble SIS indicates a high degree of variability among them. We interpret this measure of ensemble output stability to be inversely related to the epistemic uncertainty for the predictions.

Calculated uncertainty often can be correlated with the error of the prediction.[59,60] As shown in Figure 10, there is a clear positive correlation between ensemble SIS and the SIS between the model prediction and the reference spectrum. The spread of the data in this correlation is significant and the deviation from the regression line for any individual measurement can be large. The ensemble SIS measurement does not include the aleatoric component of total uncertainty and is therefore an incomplete estimator of prediction accuracy.

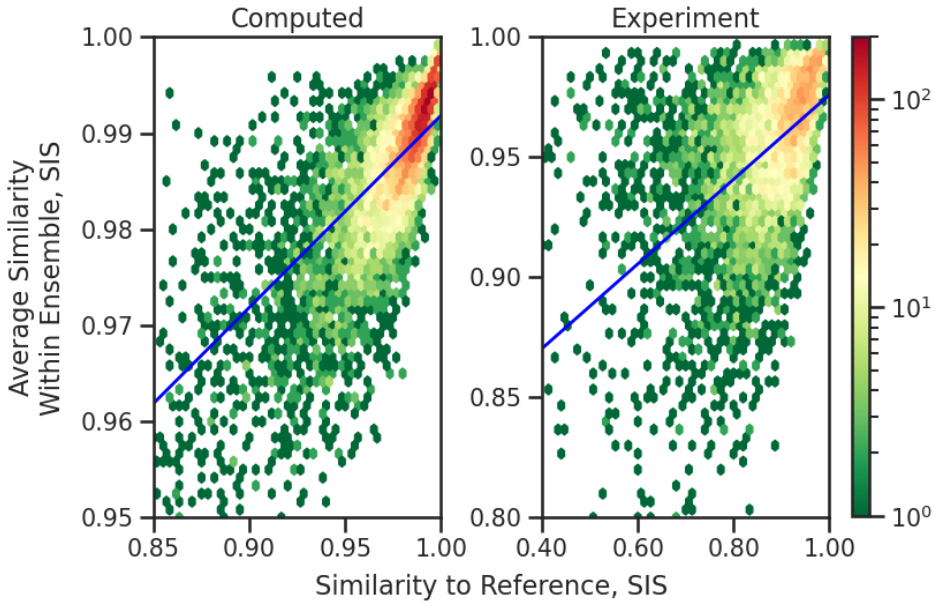

Figure 10: Intensity map of the similarity of results within the ensemble correlated with the similarity of the ensemble prediction to the reference in predictions of (left) computed spectra using the initial model and (right) experimental spectra using the final model. The least squares regression (blue) is also provided. The color indicates the number of test set predictions to fall in each area unit.

Though not quantitatively precise in estimating prediction accuracy, the ensemble SIS function is useful as a classification and ordering function in selecting which predicted spectra in a set are least likely to be accurate. Such a function can be used to prioritize targets for data acquisition. A demonstration of ensemble SIS as a prioritization function is shown in Figure 11. This case uses the final model test set as a hypothetical pool of spectra with unknown performance with the objective to identify the molecules for which the model would perform poorly. Please see exercise details in Supporting Information. By random selection it would take over 500 samples out of the pool of 5,698 to find 10 spectra with SIS scores averaging under 0.4. The search using ensemble SIS opens with an average SIS of 0.450 in its first 10 selections. Ensemble SIS finds 10 spectra with SIS scores averaging below 0.4 and then below 0.3 within the first 50 samples.Ensemble SIS performs drastically better than random selection in identifying the spectra for which predictions are likely inaccurate. We believe this demonstrates ensemble SIS may be used as part of a strategy for prioritizing data collection and as a classification tool for spectra likely to be poorly predicted.

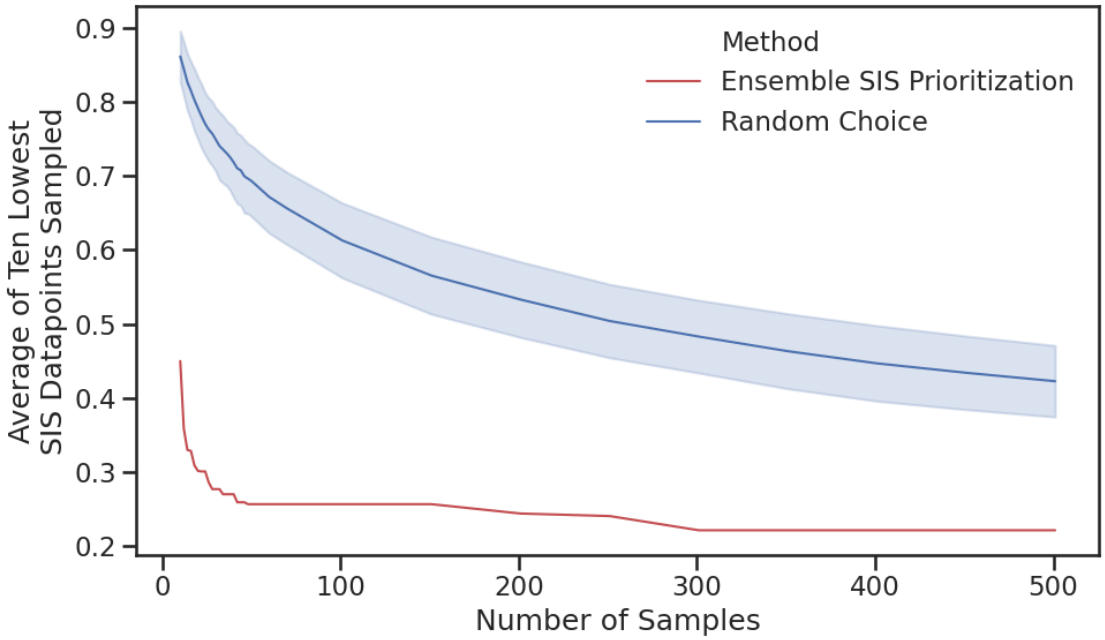

Figure 11: A demonstration of two different sampling methods for finding low scoring data, as represented by the average of the lowest ten scores sampled. These data are taken from the final model test set. The shaded blue region represents the one-standard-deviation range for outcomes of the random selection.

Model Pretraining for Generalizability

The IR spectra data available for use in the model is inherently limited. Though the datasets available from our data sources comprise tens of thousands of molecules, that is many orders of magnitude lower than the number of molecules that have been studied or could exist. Further, even with access to some of these molecules, many structures would be so unstable that they would not be practically measurable. The practical and theoretical limits of our data constrain the molecular space available to us for direct training. There are significantly fewer limits on the scope of molecules that can be studied with computed spectra, and the cost of data generation in the computational space using the methods we have outlined is dramatically lower. By pretraining with computed spectra, the MPNN weights used for molecule fingerprint generation are exposed to a more diverse set of molecules and, as a result, can generate appropriate fingerprints for a wider array of molecules.

In our application, the computed dataset includes all the unique SMILES from the experimental dataset, guaranteeing at least the same level of molecular diversity. In terms of dataset size, the 85,232 unique SMILES in the computed dataset available offer significantly more opportunity for coverage compared to the 31,439 unique SMILES available in the experimental data.

Transfer learning approaches with pretraining are often used to compensate for data scarcity. In those cases, a large training set of lower quality data is used to set some portion of the learned weights before a smaller amount of higher quality data is used to refine the model. When used to compensate for dataset size, the pretraining improves the model accuracy above the level of direct training. This model appears to have sufficient experiment spectra available for training that it does not fall into the small-data regime. However, when the training dataset is artificially reduced (Figure 12), the models show clear benefits from using pretrained MPNN weights.

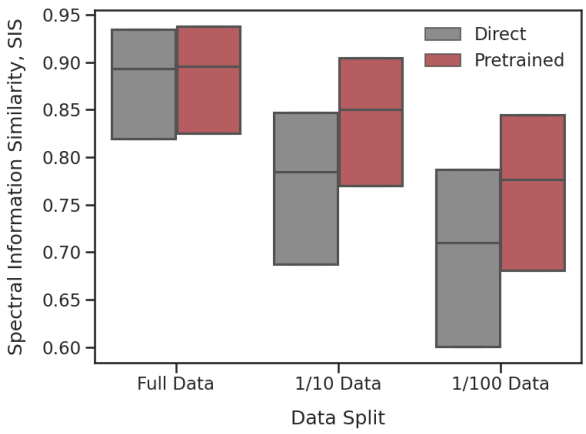

Figure 12: Distribution of SIS score for spectra predictions using different relative sizes of training dataset, comparing direct training to pretraining. The box plot shows the SIS scores for the inner quartiles of the distributions. The graphical indication of the full range of data is omitted due to the long distribution tail.

Beyond the pretraining benefits related to dataset size, there are other benefits based on dataset composition. It is believed that with the more diverse set of molecules used to pretrain the MPNN weights for fingerprint generation, that the model will function better

for molecules outside of the scope of molecules included in experimental training spectra. To test generalizability of the model, alternative methods for splitting the training data from test data were used to compare the pretrained model to a direct trained model, Figure 13. In these cases, the indicated class of molecules was excluded from the experimental training set and were used to make up the entirety of the test set. In all cases, the compared models used the same model parameters and used ensembles of ten sub-models.

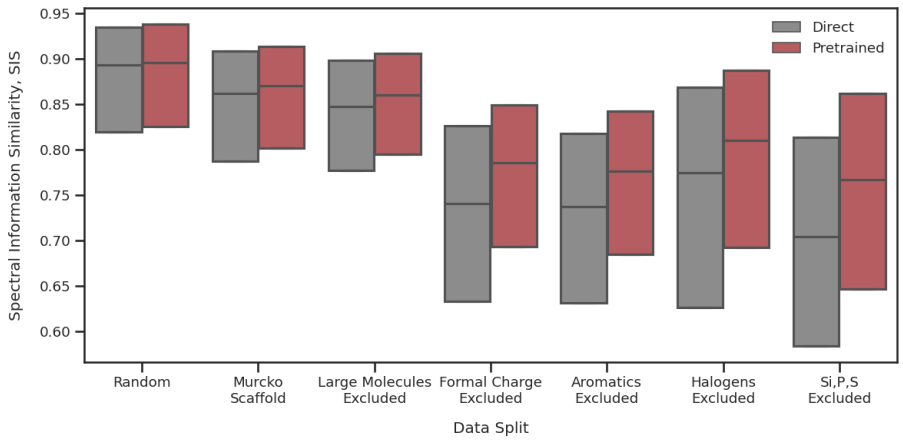

Figure 13: Distribution of SIS score for spectra predictions in the test split, comparing direct training to pretraining for cases using different splitting criteria. The Murcko scaffold comparison constrains the splits such that all molecules of a given scaffold are contained within the same split. The indicated molecule type exclusions are absent from the training set and make up the entirety of the test set. The box plot shows the SIS scores for the inner quartiles of the distributions. The graphical indication of the full range of data is omitted due to the long distribution tail.

As shown in Figure 13, there are clear benefits to pretraining on quantum data if the experimental training set omits all molecules of a certain type. The separation on the basis of Murcko scaffold constrained the test set to share no scaffolds with the test set, as a representation of predictions on novel molecule classes. The Murcko scaffold separation shows a minor reduction in performance relative to the ordinary random split. Both random and scaffold splits show minor improvement with pretraining. The exclusion of large molecules is defined here as having greater than 14 heavy atoms (the median for the dataset), with a small associated benefit from pretraining. Formal charge separation refers to molecules that

are neutral on net but have some individual atoms with non-neutral formal charge, such as would be found in the common SMILES representation of the nitro group. The halogens-excluded and Si-P-S-excluded test cases address the particularly aggressive test case in which any molecule containing those types of atoms would be excluded. In all these cases, the lack of relevant training data hinders the appropriate training of the FFNN weights in the model but is compensated for by the consideration of the excluded types in the pretrained MPNN weights. These tests indicate that generalizability for these and other classes of molecules can be improved using pretraining. Further, predictions for a desired but underrepresented class of molecules could likely be improved by adding focused examples of the class to the computed spectra for pretraining models.

Comparing Loss Function Performance

In the course of this study, we perform a preliminary investigation into alternative loss functions. In particular, we would like to compare the performance of the Spectral Information Divergence (SID) with conventional loss functions such as the Mean Squared Error (MSE) and the Root Mean Squared Error (RMSE). We also include comparisons with the Target-weighted Mean Squared Error (TMSE).

For the purposes of comparison, several models are trained, each with different loss functions. All else is held constant broadly speaking. Specifically, each model utilizes the same neural architecture described in the network architecture section. The exponential activation enforcing positivity and the whole-spectrum normalization to sum absorbances to unity is applied for each of the loss functions. All models are trained using the same 80-10-10 random splitting of the 56,955 spectra obtained from external sources, constrained such that spectra with the same SMILES had to occupy the same split. The same hyperparameters used in the model training section are enforced for all models. The maximum number of training epochs is fixed to 250 for each model. Each model uses its corresponding loss function as the metric for the evaluation of validation scores. The models with the best

validation scores are compared. Models used in this comparison are single models rather than ensembles of sub-models. To enable side-by-side comparisons of model predictions, a consistently applied similarity measure must be chosen. The SIS measure is applied for such comparisons.

Controls on the model structure are necessary for fair comparisons between loss functions. Alternative versions of these comparisons with different controls are presented in the Supporting Information. These variations are the use of scaffold splitting for training splits, the use of a MSE-based similarity in place of SIS, and the use of different random seeds for the model. Nevertheless, the results below illustrate loss function performance where all controls are enforced.

Figure 14 illustrates the performance of each loss function on the experimental dataset. For the level of control specified, MSE is outperformed by SID, RMSE, and TMSE. SID and RMSE appear fairly competitive with one another. Given that TMSE is a first-order approximation of SID, the difference in performance between the two loss functions is a result of higher-order terms. MSE and RMSE appear to offer differing performance although the losses are related monotonically. Suggested hypotheses into the differing performance may be attributed to differences in loss function and gradient scaling.

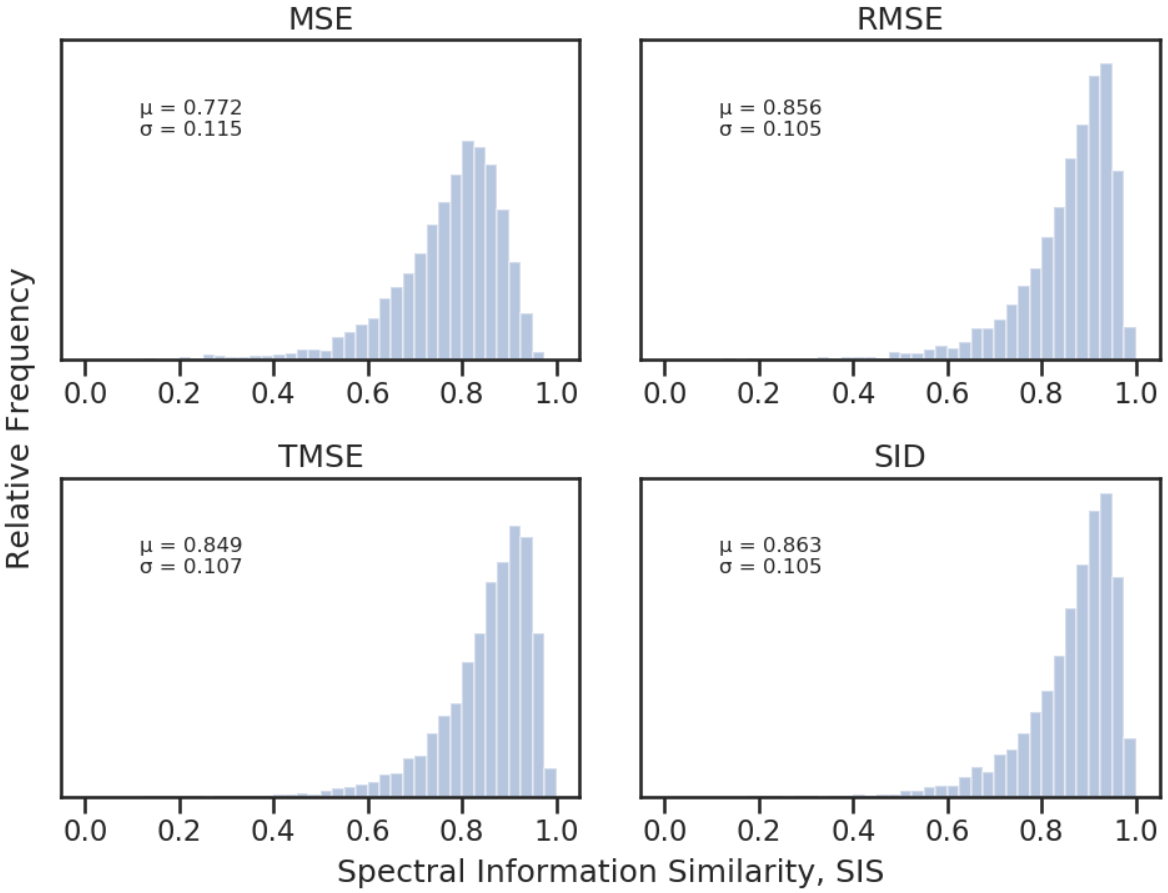

Figure 14: Distribution of SIS score for spectra predictions in the test split, comparing different loss functions.

Representative spectra are reported for the distributions in Figure 14. The best predictions for each model are illustrated in Figure 15. Each loss function appears to offer quantitative predictions in their best cases. This is consistent with the notion of neural networks serving as universal function approximators. Figure 16 illustrates the performance of average predictions. MSE appears to offer weaker peak resolution, especially in the fingerprint region. In contrast, TMSE, RMSE, and SID offer quantitative performance.

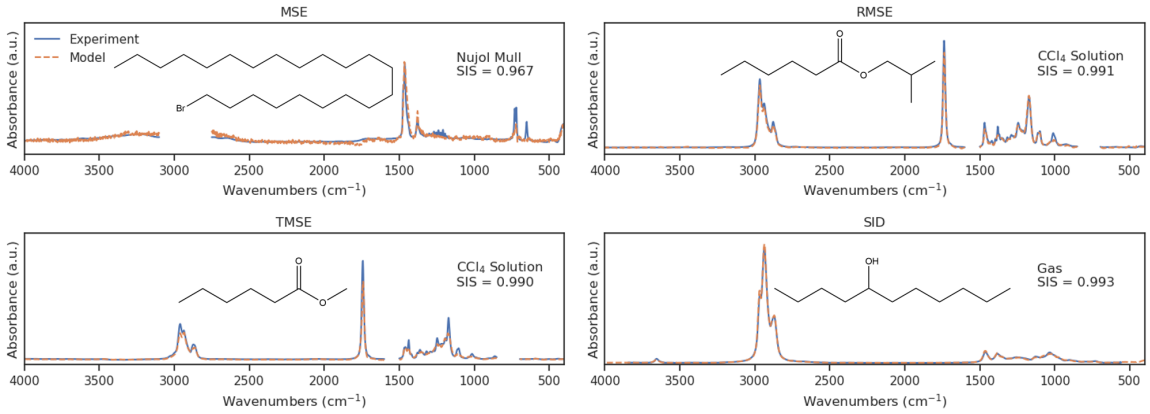

Figure 15: Cases with best model vs experimental agreement for models trained with various loss functions.

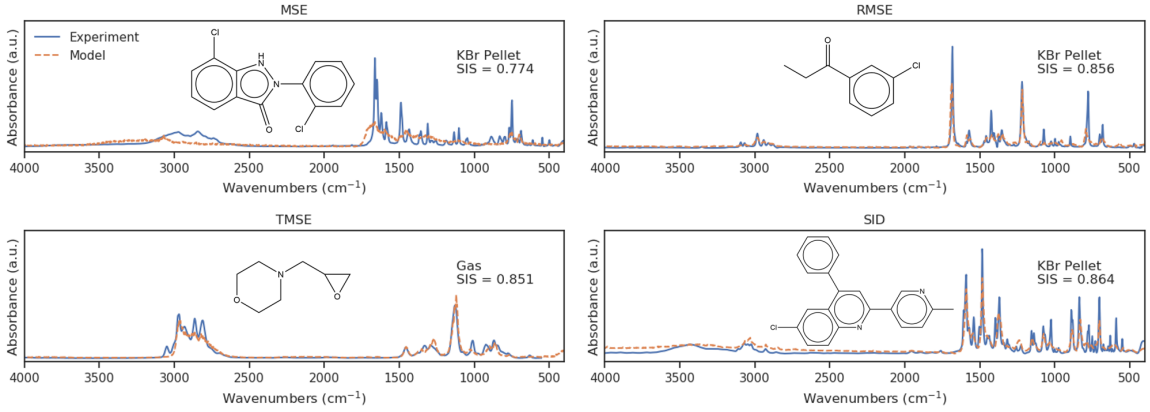

Figure 16: Average predictions vs experiment for models trained with various loss functions.

One hypothesis into the success of SID as a loss function is due to its target-weighting. SID may be re-expressed as a polynomial of weighted residuals with reciprocal target-weighting as discussed in the spectra loss function section. In fact, it can be proved that SID may be expressed as a polynomial of reciprocally-weighted residuals across its entire domain. Please see the Supporting Information for proof. This target-weighting emphasizes a strong baseline drive. A strong baseline drive in turn facilitates peak identification.

Other spectral metrics could show promise as competitive loss functions. In particular, the spectral correlation measure, the spectral angle measure, and the Euclidean distance measure could be promising alternatives. However, van der Meer observed that the spectral

information divergence exhibits better spectral discrimination than the prior alternatives,[56] which may be an important property for training machine learning architectures.

Comparing Fingerprint Performance

Expert-crafted molecular representations have been used traditionally for molecular applications of machine learning. Extended-Connectivity Fingerprints (ECFP) have been a popular class in Quantitative Structure-Activity Relationship (QSAR) models. The ECFP algorithm identifies chemical substructures in circular layers of fixed radius and hashes these into a representation vector.[61] Note that ECFP representations offer a direct analog to learned MPNN fingerprints, albeit with the use of manual feature engineering. The Morgan fingerprint, a common ECFP, will serve as a baseline for our machine learning architecture.

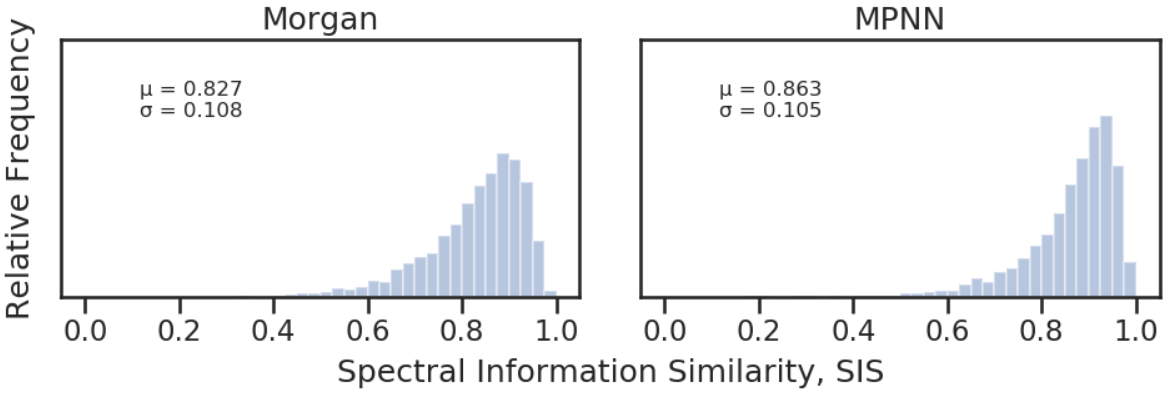

Figure 17: Distribution of SIS score for spectra predictions in the test split, comparing different fingerprint.

This fingerprint baseline helps assess the performance of the learned MPNN representation. For an analogous comparison, the Morgan fingerprint radius is set to 6 with hashed bit length of 2048, a conventional size.[37,62,63] SIS histograms are generated for the comparison. Similar controls are enforced as described in the comparison of loss functions. The SID loss function is enforced while the fingerprint is varied between a Morgan and MPNN fingerprint.

Figure 17 illustrates the performance of two fingerprints. The trial suggest that utilizing a MPNN improves prediction performance over use of a Morgan fingerprint. One hypothesis into the apparent success of the MPNN over the Morgan fingerprint could be attributed to the flexibility of the deep learning representation. Further, the MPNN representation is trained specifically for the IR prediction task, whereas the Morgan fingerprint is engineered as a fixed representation. Additional comparisons varying similarity measure, random seed, and split type can be found in the Supporting Information.

Comparison to Other Methods

The Chemprop-IR architecture offers an accurate and competitive alternative to first-principles techniques. For comparison, *ab initio* quantum mechanics yield prediction errors typically on the order of 10 cm^{-1} for both gas and condensed-phase.[9,23] Chemprop-IR can offer minimal peak shifts in predictive performance. This can be most readily seen for predictions with SIS>0.85, covering the majority of the experimental spectra performance distribution. Examples of the low level of blue or red shift are available in Figures 4 and 5. The promising predictive accuracy could be attributed to the SID loss function and the model's network structure. The SID loss function has an aggressive drive to baseline, giving the model good performance in distinguishing peak locations from the the gaps between, especially in cases of fully separated peaks. With the network structure of the Chemprop-IR architecture, the model asymptotically enjoys the benefits of universal function approximators when operating with a sufficiently large and flexible set of parameters. Provided with a sufficient amount of well-chosen training spectra, a large network has the potential to predict peak position well. These two factors give Chemprop-IR the potential to be competitive with *ab initio* methods with respect to predicting peak locations.

A similar argument holds for the absorbance intensities. Current *ab initio* methods have an estimated prediction error on the order of 10% based on a relative ordering of peaks, though a range of more and less severe peak errors are reported in recent literature.[23,24]

Calculations of this accuracy are attainable through calculations of the dipole moment and polarizability derivative in quantum chemistry calculations.[20] Chemprop-IR is able to incorporate such considerations implicitly through its learned recognition of functional groups and their interactions. In this way, the model is capable of predicting peak intensities to an accuracy often comparable with first principles approaches. For much of the performance distribution for the experimental spectrum model, the intensities of major spectra peaks are well predicted. For the high end of performance, SIS>0.95, the predicted intensities reproduce experimental peak magnitudes well for major and most minor peaks.

Predictions made using a previously trained model in Chemprop-IR offer benefits complementary to spectral predictions derived from *ab initio* quantum calculations. Traditionally, *ab initio* methods excel in providing a high degree of interpretability. The precise modes and conformers contributing to a particular set of peaks can be more easily identified with DFT. Further, DFT methods and higher levels of theory can offer significant predictive accuracy given sufficient computational resources and expert knowledge. Chemprop-IR may complement *ab initio* approaches by offering high-throughput predictions which can recreate target spectra with little need for expert knowledge. Using Chemprop-IR for predictions bypasses the large CPU-time requirements of DFT for large molecules. With minimal computational resources, absorbance spectra for thousands of molecules can be calculated in minutes. Predictions through Chemprop-IR include phase and conformer effects implicitly rather than requiring multiple calculations to represent an array of conformers or the solvent environment. Predictions are made with minimal input data, requiring only SMILES strings and no initial geometries for optimization.

Computational Time of Calculations

Computational cost for machine-learning models, such as those using Chemprop-IR, is predominantly a factor during model training. Prediction tasks are relatively fast and inexpensive by comparison. Users may invest the significant time for training a model up front and

then make use of fast predictions later. Interested readers may download the trained models discussed in this work and use them for prediction tasks without needing to undergo the costly training process.

For reference, some example computational times are provided (using 4 cores of an Intel Xeon Gold 6248 and a Nvidia Volta V100 gpu). Predictions made for the 5,695 spectra of the experiment test set took 3 minutes to calculate from the ensemble of 10 submodels and 1.5 minutes when using a single submodel. Prediction times were unaffected by whether a direct training model or a pretrained model was used. Training a single submodel took 6 hours and 41 minutes when trained directly and 16 hours and 58 minutes when trained with pretraining for the model discussed in this work. Training an ensemble of 10 submodels took 66 hours when trained directly and 165 hours when trained with pretraining. Ensemble submodels can be trained independently and in parallel if resources are available.

Potential Future Improvements

The trained prediction model and the Chemprop-IR software function well to provide high-fidelity predictions of IR spectra. Even so, some relevant aspects of the results indicate areas where the prediction performance could improve: rolling peaks shapes, minor direct improvement from pretraining, overperformance of the training set relative to the test set, and long tails on performance distributions. There are many avenues to attempt improvement that appear promising.

Higher levels of theory than GFN2-xTB can and should be used to generate computed spectra for the model. Testing alternative methods for broadening peaks in the computed spectra also would be appropriate, in the pursuit of making the pretraining step higher impact. Improvements in the quality of the computed spectra could advance them to the point where they are more closely related to the experimental spectra targets. In so doing, the learned molecule encoder would likely improve in its identification of molecular features relevant to prediction of vibrational spectra.

Increases in the amount of data would likely have a positive effect on the model performance. As shown in Figure 6, performance in the test set does correlate with having nearby molecules in the training set. A targeted effort to enlarge the dataset, focusing on areas of chemical space with sparse coverage, could be effective. Uncertainty metrics, such as ensemble SIS, would be useful in such targeting.

A further opportunity for improvement must be balanced against the ease of usability of the model. At present, only a SMILES string is needed for prediction. Models could alternatively be constructed that use calculated features from *ab initio* methods as input features to the model alongside the SMILES string. Such additional features would be more informative to a model regarding the 3D structure of a molecule and could lead to significantly better performance, but would come at the cost of requiring such a calculation before any molecule prediction.

Data and Software Availability

The Chemprop-IR software is available through a public repository with GitHub.[64] An archived version of the software and the trained models associated with this publication are shared through Zenodo.[65] In this repository, the trained models for the computed spectra and for experimental spectra are provided, allowing predictions of new spectra in any of the supported phases. The full set of computed spectra used for pretraining as well as code for generating additional computed spectra are available. The software provides tools for calculating ensemble uncertainty for predicted spectra and for calculating SIS between predictions and reference spectra. The software enables the training of new models from input spectra, using such features as the SID loss function and spectrum normalization by specified absorbance ranges. The tuned settings used in the prediction of IR spectra in this work are provided as recommendations for future applications. These software tools are designed to be generalizable to spectra types beyond IR.

The experimental data used in the training of these models are restricted by copyright held by the respective organizations and cannot be shared directly. Please see the data access statement in the Supporting Information for details on the datasets used and available channels for requesting these data.

Conclusion

The Chemprop-IR architecture and associated trained models offer high-fidelity infrared predictions in gas phase and condensed phases. The spectral loss metrics, namely SID, may provide an accelerated drive to baseline and peak resolution compared to traditional metrics. These predictions are competitive with *ab initio* methods, avoiding peak shift pathologies due to an imperfect or incomplete Hamiltonian. The software provides easy, accessible predictions of infrared spectra from only a molecular SMILES representation, making spectral predictions for large batches of prospective molecules possible for a wealth of user applications.

Acknowledgement

This work was supported by the DARPA Accelerated Molecular Discovery (AMD) program under contract # DARPA HR00111920025. We acknowledge NIST, PNNL, AIST, and Coblentz Society for making available the collected spectra that made this work possible. We acknowledge the MIT SuperCloud and Lincoln Laboratory Supercomputing Center for providing high performance computing resources that have contributed to the research results reported within this paper.[66] We would like to thank Ellen Miseo and Mary Carrabba of the Coblentz Society for discussion of the Coblentz Society Spectra. We would like to thank Ye Li, MIT Chemistry and Chemical Engineering Librarian, for helpful input as we gathered the dataset. We would also like to thank our colleagues Florence Vermeire, Lagnajit Pattanaik, and Kevin Greenman for the helpful discussion and insight.

Supporting Information Available

The Supporting Information contains the data access statement as well as additional analysis on the computational dataset, the peak broadening procedure, the SIS loss function, phase performance, ensemble SIS, loss function and fingerprint baseline comparisons, the polynomial expansion of SID, and a listing of SMILES strings associated with molecule diagrams in figures.

References

(1) Hansen, K.; Montavon, G.; Biegler, F.; Fazli, S.; Rupp, M.; Scheffler, M.; von Lilienfeld, O. A.; Tkatchenko, A.; Müller, K.-R. Assessment and Validation of Machine Learning Methods for Predicting Molecular Atomization Energies. *J. Chem. Theory Comput.* **2013**, *9*, 3404–3419.

(2) Ghosh, K.; Stuke, A.; Todorović, M.; Jørgensen, P. B.; Schmidt, M. N.; Vehtari, A.; Rinke, P. Deep Learning Spectroscopy: Neural Networks for Molecular Excitation Spectra. *Adv. Sci.* **2019**, *6*, 1801367.

(3) Wang, X.; Ye, S.; Hu, W.; Sharman, E.; Liu, R.; Liu, Y.; Luo, Y.; Jiang, J. Electric Dipole Descriptor for Machine Learning Prediction of Catalyst Surface–Molecular Adsorbate Interactions. *J. Am. Chem. Soc.* **2020**, *142*, 7737–7743.

(4) Feinberg, E. N.; Sur, D.; Wu, Z.; Husic, B. E.; Mai, H.; Li, Y.; Sun, S.; Yang, J.; Ramsundar, B.; Pande, V. S. PotentialNet for Molecular Property Prediction. *ACS Cent. Sci.* **2018**, *4*, 1520–1530.

(5) Chung, H. Applications of Near-Infrared Spectroscopy in Refineries and Important Issues to Address. *Appl. Spectrosc. Rev.* **2007**, *42*, 251–285.

(6) Barth, A.; Haris, P. *Biological and Biomedical Infrared Spectroscopy*; IOS Press BV, 2009; Vol. 2.

(7) Wu, H.; Saltzberg, D. J.; Kratochvil, H. T.; Jo, H.; Sali, A.; DeGrado, W. F. Glutamine Side Chain 13C18O as a Nonperturbative IR Probe of Amyloid Fibril Hydration and Assembly. *J. Am. Chem. Soc.* **2019**, *141*, 7320–7326.

(8) Kim, K.; Kang, S.; Yoo, J.; Kwon, Y.; Nam, Y.; Lee, D.; Kim, I.; Choi, Y.-S.; Jung, Y.; Kim, S.; Son, W.-J.; Son, J.; Kim, S.; Shin, J.; Hwang, S. Deep-learning-based Inverse Design Model for Intelligent Discovery of Organic Molecules. *npj Comput. Mater.* **2018**, *4*, 67.

(9) Bouteiller, Y.; Gillet, J.-C.; Grégoire, G.; Schermann, J. P. Transferable Specific Scaling Factors for Interpretation of Infrared Spectra of Biomolecules from Density Functional Theory. *J. Phys. Chem. A* **2008**, *112*, 11656–11660.

(10) Bouteiller, Y.; Poully, J. C.; Desfrançois, C.; Grégoire, G. Evaluation of MP2, DFT, and DFT-D Methods for the Prediction of Infrared Spectra of Peptides. *J. Phys. Chem. A* **2009**, *113*, 6301–6307.

(11) Malik, M.; Wysokiński, R.; Zierkiewicz, W.; Helios, K.; Michalska, D. Raman and Infrared Spectroscopy, DFT Calculations, and Vibrational Assignment of the Anticancer Agent Picoplatin: Performance of Long-Range Corrected/Hybrid Functionals for a Platinum(II) Complex. *J. Phys. Chem. A* **2014**, *118*, 6922–6934.

(12) DeBlase, A. F.; Bloom, S.; Lectka, T.; Jordan, K. D.; McCoy, A. B.; Johnson, M. A. Origin of the Diffuse Vibrational Signature of a Cyclic Intramolecular Proton Bond: Anharmonic Analysis of Protonated 1,8-Disubstituted Naphthalene Ions. *J. Chem. Phys.* **2013**, *139*, 024301.

(13) Katari, M.; Nicol, E.; Steinmetz, V.; vanderRest, G.; Carmichael, D.; Frison, G. Im-

proved Infrared Spectra Prediction by DFT from a New Experimental Database. *Chem. - Eur. J.* **2017**, *23*, 8414–8423.

(14) Halls, M. D.; Velkovski, J.; Schlegel, H. B. Harmonic Frequency Scaling Factors for Hartree-Fock, S-VWN, B-LYP, B3-LYP, B3-PW91 and MP2 with the Sadlej pVTZ Electric Property Basis Set. *Theor. Chem. Acc.* **2001**, *105*, 413–421.

(15) Alcolea Palafox, M. Scaling Factors for the Prediction of Vibrational Spectra. I. Benzene Molecule. *Int. J. Quantum Chem.* **2000**, *77*, 661–684.

(16) Maltseva, E.; Petrignani, A.; Candian, A.; Mackie, C. J.; Huang, X.; Lee, T. J.; Tielens, A. G. G. M.; Oomens, J.; Buma, W. J. High-Resolution IR Absorption Spectroscopy of Polycyclic Aromatic Hydrocarbons: The Realm of Anharmonicity. *Astrophys. J.* **2015**, *814*, 23.

(17) Mackie, C. J.; Candian, A.; Huang, X.; Maltseva, E.; Petrignani, A.; Oomens, J.; Buma, W. J.; Lee, T. J.; Tielens, A. G. G. M. The Anharmonic Quartic Force Field Infrared Spectra of Three Polycyclic Aromatic Hydrocarbons: Naphthalene, Anthracene, and Tetracene. *J. Chem. Phys.* **2015**, *143*, 224314.

(18) Mackie, C. J.; Chen, T.; Candian, A.; Lee, T. J.; Tielens, A. G. G. M. Fully Anharmonic Infrared Cascade Spectra of Polycyclic Aromatic Hydrocarbons. *J. Chem. Phys.* **2018**, *149*, 134302.

(19) Gaw, J. F.; Willetts, A.; Green, W. H.; Handy, N. C. In *Advances in Molecular Vibrations and Collision Dynamics*; Bowman, J. M., Ratner, M. A., Eds.; 1991; Vol. 1B; p 169.

(20) Zvereva, E. E.; Shagidullin, A. R.; Katsyuba, S. A. Ab Initio and DFT Predictions of Infrared Intensities and Raman Activities. *J. Phys. Chem. A* **2011**, *115*, 63–69.

(21) Thomas, M.; Brehm, M.; Fligg, R.; Vöhringer, P.; Kirchner, B. Computing Vibrational Spectra from Ab Initio Molecular Dynamics. *Phys. Chem. Chem. Phys.* **2013**, *15*, 6608–6622.

(22) Gaigeot, M.-P.; Sprik, M. Ab Initio Molecular Dynamics Computation of the Infrared Spectrum of Aqueous Uracil. *J. Phys. Chem. B* **2003**, *107*, 10344–10358.

(23) Katsyuba, S. A.; Spicher, S.; Gerasimova, T. P.; Grimme, S. Fast and Accurate Quantum Chemical Modeling of Infrared Spectra of Condensed-Phase Systems. *The Journal of Physical Chemistry B* **2020**, *124*, 6664–6670, PMID: 32633534.

(24) Gastegger, M.; Behler, J.; Marquetand, P. Machine Learning Molecular Dynamics for the Simulation of Infrared Spectra. *Chem. Sci.* **2017**, *8*, 6924–6935.

(25) Laurens, G.; Rabary, M.; Lam, J.; Peláez, D.; Allouche, A.-R. Infrared Spectra of Neutral Polycyclic Aromatic Hydrocarbons by Machine Learning. *arXiv e-prints* **2020**, arXiv:2010.13686.

(26) Lam, J.; Abdul-Al, S.; Allouche, A.-R. Combining Quantum Mechanics and Machine-Learning Calculations for Anharmonic Corrections to Vibrational Frequencies. *J. Chem. Theory Comput.* **2020**, *16*, 1681–1689.

(27) Gastegger, M.; Schütt, K. T.; Müller, K.-R. Machine Learning of Solvent Effects on Molecular Spectra and Reactions. *arXiv e-prints* **2020**, arXiv:2010.14942.

(28) Kovács, 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*, 100.

(29) Clerc, J.-T.; Terkovics, A. L. Versatile Topological Structure Descriptor for Quantitative Structure/Property Studies. *Anal. Chim. Acta* **1990**, *235*, 93 – 102.

(30) Affolter, C.; Clerc, J. Prediction of Infrared Spectra from Chemical Structures of Organic Compounds Using Neural Networks. *Chemom. Intell. Lab. Syst.* **1993**, *21*, 151 – 157.

(31) Barbon, S.; Costa Barbon, A. P. A. d.; Mantovani, R. G.; Barbin, D. F. Machine Learning Applied to Near-Infrared Spectra for Chicken Meat Classification. *J. Spectrosc.* **2018**, *2018*, 8949741.

(32) Cadet, X. F.; Lo-Thong, O.; Bureau, S.; Dehak, R.; Bessafi, M. Use of Machine Learning and Infrared Spectra for Rheological Characterization and Application to the Apricot. *Sci. Rep.* **2019**, *9*, 19197.

(33) Mwanga, E. P.; Mapua, S. A.; Siria, D. J.; Ngowo, H. S.; Nangacha, F.; Mgando, J.; Baldini, F.; González Jiménez, M.; Ferguson, H. M.; Wynne, K.; Selvaraj, P.; Babayan, S. A.; Okumu, F. O. Using Mid-Infrared Spectroscopy and Supervised Machine-Learning to Identify Vertebrate Blood Meals in the Malaria Vector, Anopheles Arabiensis. *Malar. J.* **2019**, *18*, 187.

(34) Lansford, J. L.; Vlachos, D. G. Infrared Spectroscopy Data- and Physics-Driven Machine Learning for Characterizing Surface Microstructure of Complex Materials. *Nat. Commun.* **2020**, *11*, 1513.

(35) Le, B. T. Application of Deep Learning and Near Infrared Spectroscopy in Cereal Analysis. *Vib. Spectrosc.* **2020**, *106*, 103009.

(36) Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E. Neural Message Passing for Quantum Chemistry. *arXiv e-prints* **2017**, arXiv:1704.01212.

(37) Duvenaud, D. K.; Maclaurin, D.; Iparraguirre, J.; Bombarell, R.; Hirzel, T.; Aspuru-Guzik, A.; Adams, R. P. In *Advances in Neural Information Processing Systems 28*; Cortes, C., Lawrence, N. D., Lee, D. D., Sugiyama, M., Garnett, R., Eds.; Curran Associates, Inc., 2015; pp 2224–2232.

(38) Yang, K.; Swanson, K.; Jin, W.; Coley, C.; Eiden, P.; Gao, H.; Guzman-Perez, A.; Hopper, T.; Kelley, B.; Mathea, M.; Palmer, A.; Settels, V.; Jaakkola, T.; Jensen, K.; Barzilay, R. Analyzing Learned Molecular Representations for Property Prediction. *J. Chem. Inf. Model.* **2019**, *59*, 3370–3388.

(39) NIST Mass Spectrometry Data Center, In *"Infrared Spectra" in NIST Chemistry WebBook, NIST Standard Reference Database Number 69*; Lindstrom, P. J., Mallard, W. G., Eds.; National Institute of Standards and Testing: Gaithersburg, MD, `http://webbook.nist.gov` (Accessed 2019-10-3).

(40) Johnson, T. J.; Profeta, L. T.; Sams, R. L.; Griffith, D. W.; Yokelson, R. L. An Infrared Spectral Database for Detection of Gases Emitted by Biomass Burning. *Vibrational Spectroscopy* **2010**, *53*, 97–102.

(41) Sharpe, S. W.; Sams, R. L.; Johnson, T. J.; Chu, P. M.; Rhoderick, G. C.; Guenther, F. R. Creation of 0.10-cm-1 resolution quantitative infrared spectral libraries for gas samples. Vibrational Spectroscopy-based Sensor Systems. 2002; pp 12 – 24.

(42) Sharpe, S. W.; Johnson, T. J.; Sams, R. L.; Chu, P. M.; Rhoderick, G. C.; Johnson, P. A. Gas-phase databases for quantitative infrared spectroscopy. *Appl. Spectrosc.* **2004**, *58*, 1452–1461.

(43) National Institute of Advanced Science and Technology, SDBS Web. `https://sdbs.db.aist.go.jp` (Accessed 2019-10-1).

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

(45) Kim, S.; Chen, J.; Cheng, T.; Gindulyte, A.; He, J.; He, S.; Li, Q.; Shoemaker, B. A.; Thiessen, P. A.; Yu, B.; Zaslavsky, L.; Zhang, J.; Bolton, E. E. PubChem 2019 Update: Improved Access to Chemical Data. *Nucleic Acids Res.* **2018**, *47*, D1102–D1109.

(46) Riniker, S.; Landrum, G. A. Better Informed Distance Geometry: Using What We Know To Improve Conformation Generation. *J. Chem. Inf. Model.* **2015**, *55*, 2562–2574, PMID: 26575315.

(47) Halgren, T. A. Merck Molecular Force Field. I. Basis, Form, Scope, Parameterization, and Performance of MMFF94. *J. Comput. Chem.* **1996**, *17*, 490–519.

(48) Bannwarth, C.; Ehlert, S.; Grimme, S. GFN2-xTB—An Accurate and Broadly Parametrized Self-Consistent Tight-Binding Quantum Chemical Method with Multipole Electrostatics and Density-Dependent Dispersion Contributions. *J. Chem. Theory Comput.* **2019**, *15*, 1652–1671, PMID: 30741547.

(49) Wu, Z.; Ramsundar, B.; Feinberg, E. N.; Gomes, J.; Geniesse, C.; Pappu, A. S.; Leswing, K.; Pande, V. MoleculeNet: a Benchmark for Molecular Machine Learning. *Chem. Sci.* **2017**, *9*, 513–530, 29629118[pmid].

(50) Schütt, K. T.; Sauceda, H. E.; Kindermans, P.-J.; Tkatchenko, A.; Müller, K.-R. SchNet – A Deep Learning Architecture for Molecules and Materials. *J. Chem. Phys.* **2018**, *148*, 241722.

(51) St. John, P. C.; Guan, Y.; Kim, Y.; Kim, S.; Paton, R. S. Prediction of Organic Homolytic Bond Dissociation Enthalpies at Near Chemical Accuracy with Sub-Second Computational Cost. *Nat. Commun.* **2020**, *11*, 2328.

(52) Bergstra, J.; Yamins, D.; Cox, D. D. Making a Science of Model Search: Hyperparameter Optimization in Hundreds of Dimensions for Vision Architectures. Proceedings of the 30th International Conference on International Conference on Machine Learning - Volume 28. 2013; p I–115–I–123.

(53) Kingma, D. P.; Ba, J. Adam: A Method for Stochastic Optimization. *arXiv e-prints* **2014**, arXiv:1412.6980.

(54) Chang, C.-I. An Information-Theoretic Approach to Spectral Variability, Similarity, and Discrimination for Hyperspectral Image Analysis. *IEEE Trans. Inf. Theory* **2000**, *46*, 1927–1932.

(55) Kumar, M. N.; Seshasai, M. V. R.; Prasad, K. S. V.; Kamala, V.; Ramana, K. V.; Dwivedi, R. S.; Roy, P. S. A New Hybrid Spectral Similarity Measure for Discrimination among Vigna Species. *Int. J. Remote Sens.* **2011**, *32*, 4041–4053.

(56) van der Meer, F. The Effectiveness of Spectral Similarity Measures for the Analysis of Hyperspectral Imagery. *Int. J. Appl. Earth Obs. Geoinf.* **2006**, *8*, 3 – 17.

(57) Kendall, A.; Gal, Y. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA, 2017; p 5580–5590.

(58) Lakshminarayanan, B.; Pritzel, A.; Blundell, C. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook, NY, USA, 2017; p 6405–6416.

(59) Guo, C.; Pleiss, G.; Sun, Y.; Weinberger, K. Q. On Calibration of Modern Neural Networks. Proceedings of the 34th International Conference on Machine Learning. International Convention Centre, Sydney, Australia, 2017; pp 1321–1330.

(60) Scalia, G.; Grambow, C. A.; Pernici, B.; Li, Y.-P.; Green, W. H. Evaluating Scalable Uncertainty Estimation Methods for Deep Learning-Based Molecular Property Prediction. *J. Chem. Inf. Model.* **2020**, *60*, 2697–2717.

(61) Probst, D.; Reymond, J.-L. A Probabilistic Molecular Fingerprint for Big Data Settings. *J. Cheminf.* **2018**, *10*, 66.

(62) Kensert, A.; Alvarsson, J.; Norinder, U.; Spjuth, O. Evaluating Parameters for Ligand-Based Modeling with Random Forest on Sparse Data Sets. *J. Cheminf.* **2018**, *10*, 49.

(63) Gütlein, M.; Kramer, S. Filtered Circular Fingerprints Improve Either Prediction or Runtime Performance While Retaining Interpretability. *J. Cheminf.* **2016**, *8*, 60–60.

(64) McGill, C.; Forsuelo, M.; Guan, Y.; Green, W. Chemprop-IR. `https://github.com/gfm-collab/chemprop-IR` (Accessed 2021-4-17).

(65) McGill, C.; Forsuelo, M.; Guan, Y.; Green, W. Chemprop-IR. `https://zenodo.org/record/4698943` (Accessed 2021-4-17).

(66) Reuther, A.; Kepner, J.; Byun, C.; Samsi, S.; Arcand, W.; Bestor, D.; Bergeron, B.; Gadepally, V.; Houle, M.; Hubbell, M.; Jones, M.; Klein, A.; Milechin, L.; Mullen, J.; Prout, A.; Rosa, A.; Yee, C.; Michaleas, P. Interactive Supercomputing on 40,000 Cores for Machine Learning and Data Analysis. 2018 IEEE High Performance extreme Computing Conference (HPEC). 2018; pp 1–6.

Graphical TOC Entry