Towards Accurate PAH IR Spectra Prediction: Handling Charge Effects with Classical and Deep Learning Models

Babken G. Beglaryan, Aleksandr S. Zakuskin, Viktor A. Nemchenko, Timur A. Labutin *

Lomonosov Moscow State University, 119234 Moscow, Russia

E-mail: timurla@laser.chem.msu.ru

ABSTRACT: Polycyclic aromatic hydrocarbons (PAHs) play a crucial role in astrochemistry, environmental studies, and combustion chemistry, yet interpreting their infrared (IR) spectra remains challenging due to similarity of spectral features of many molecules. Presumable presence of both neutral and charged PAHs in mixtures complicates spectra interpretation too. While firstprinciples calculations provide accurate spectral predictions, their high computational cost limits scalability. This study employs machine learning (ML) to predict PAH IR spectra, emphasizing applicability of the developed models simultaneously for neutral and ionized molecules. Two models are introduced: a XGBoost model trained on Morgan fingerprints and a graph neural network (GNN) that employs molecular graph representations. Charged molecules are treated by incorporating one-hot or learnable NN encoding to molecular representations. Both models demonstrate excellent predictive capabilities, for the first time enabling fast and accurate prediction of charged PAHs IR spectra. While the XGBoost model demonstrates the highest accuracy achieved up to date, the GNN shows significant promise for future advancements due to the inherent capabilities of molecular graph representations. Remaining challenges, such as scarcity of data on heteroatomic PAHs, and potential approaches of addressing them are also discussed in the manuscript.

1. INTRODUCTION

Polycyclic aromatic hydrocarbons (PAHs) and their physicochemical and spectral properties are of significant interest in the fields of astronomy,^{1, 2} environmental science,^{3, 4} and related fields, including sensor development⁵ and studies on optimization of combustion process to reduce PAHs emission.^{6, 7} Environmental studies have focused on PAHs due to their classification as hazardous pollutants, given their toxicity, mutagenicity, and carcinogenicity, which pose significant health risks.⁸ PAHs are widespread in the space, from circumstellar environments to planetary atmospheres and meteorites,⁹⁻¹² making their investigation one of the main objectives in astrochemistry.^{13,14}

Mid-infrared (IR) spectroscopy is one of the most widely used techniques for PAH analysis, enabling the characterization of their molecular structure during combustion processes.^{6, 7} It is also employed for monitoring the concentration profiles of specific hydrocarbons in air and water.^{5, 15} The presence of PAHs in various astronomical objects, including emission, reflection and planetary nebulae,¹⁶⁻²⁰ young stellar objects,^{21,22} ultraluminous infrared galaxies²³ has also been extensively investigated in the mid-IR range. These molecules play a dominant role in heating neutral gas, maintaining the ionization balance within molecular clouds, influencing star formation,¹³ and contributing to the production of carbon particles and fullerenes.^{24, 25}

Despite the importance and presumed abundance of PAHs, detailed studies remain limited due to the complexity of interpreting observed IR spectra from astronomical objects, polluted air, or combustion products. These spectra result from the cumulative contributions of tens or even hundreds of PAHs species.^{26, 27} Distinguishing signals from different PAHs is particularly challenging due to their spectral similarities. Moreover, the vast number of theoretically possible PAH structures and their various charged states further complicate the IR spectra interpretation, especially in case of astronomical data.¹³ Consequently, the development and application of

methods for systematically obtaining IR spectra for as many individual PAH molecules as possible are essential for their identification in the environment, ISM, and combustion products.

IR spectra of individual PAHs can be obtained using both experimental and computational approaches. The experimental measurements of IR spectra for PAHs are limited due to challenges associated primarily with isolating and holding PAHs in the gaseous phase.¹³ In several studies, the infrared multiple photon dissociation (IRMPD) technique has been employed for recording the spectra of small PAH cations.²⁸⁻³¹ Computational methods are extensively used in combination with such experimental techniques.^{30, 32} In particular, density functional theory (DFT), is employed to calculate PAHs IR spectra.³³ However, as the molecular mass increases, the DFT calculations become more complicated and resource-intensive.

To accelerate the prediction of spectra, the application of machine learning (ML) models presents a promising approach. Machine learning has already found successful applications in two domains closely related to the prediction of PAH IR spectra: creating latent representations of molecular structure for further prediction of some properties and the prediction or interpretation of spectral data. Various ML approaches are already used to predict the properties of individual molecules, including solubility,^{34, 35} toxicity,^{36, 37} melting points,³⁸ as well as for screening molecules for antibacterial activity.³⁹ It is important to note that opinions differ on the relative effectiveness of traditional descriptor-based machine learning methods versus graph neural networks for molecular property prediction. One review suggests that classical models trained on molecular descriptors typically achieve higher accuracy with significantly lower computational cost,⁴⁰ while other highlight several successful applications of graph neural networks in this field.⁴¹ Given these contrasting perspectives, it is reasonable to explore both approaches to address our research objectives. ML has also been successfully applied for predicting the spectral characteristics of various sources,^{42, 43} including the prediction of IR-spectra of organic molecules.⁴⁴ Several studies have focused on solving the inverse problem of infrared spectrum

interpretation and identifying possible fragments and functional groups in organic molecules.^{45, 46}

McCarthy and Lee⁴⁷ proposed a method for molecular identification based on experimental rotational data using deep learning networks. Kovács et al.⁴⁸ previously demonstrated the potential of ML for predicting the IR spectra of neutral PAHs using multilayer perceptron (MLP) and random forest model. Despite the success of neural networks to bypass expensive DFT calculations and achieve high-quality predictions for numerous molecules, this approach has excluded ionized molecules from training. More importantly, it prevents the prediction of their spectra, even though they are crucial for identifying emitters and constraining the physical properties of the irradiating source in astrophysical studies.⁴⁹

In this study, we focus on leveraging the significantly expanded PAHs spectral data to enable the prediction of IR spectra for the widest possible range of molecules, with a particular emphasis on ionized species. To achieve this, we propose the implementation of two machine learning techniques combined with capacious representations of molecular structure and ionization state for predicting the IR spectra of PAHs. The first approach employs a classical ML model based on gradient boosting—XGBoost.⁵⁰ The second approach applies a graph neural network.⁵¹ Both methods are capable of providing state-of-the-art results at the moment, advancing the quality of predicted PAH IR spectra, but have different peculiarities that define possible directions of further development.

2. METHODOLOGY

2.1. Database and data preprocessing.

We gathered data from the publicly available NASA Ames PAH IR Spectroscopic Database (version 3.20).⁵² The chemical structure of the PAH molecules in the database is represented in the form of atomic coordinates (XYZ format). Most of the PAH IR spectra are calculated using DFT, with only a dozen of molecules having experimentally recorded spectra. The low representativeness of experimental data (only 84 unique spectra) prevents training models solely on them. Another challenge pertains to the noticeable discrepancies between DFT-calculated and experimentally recorded spectra (Figure S1), raising concerns about the validity of their integration

into a unified dataset. Given that DFT-calculated spectra constitute the majority of the database, totaling 4233 unique spectra, this study uses exclusively theoretical spectra for model training, excluding experimental data from consideration. Since the IR spectra of various molecules consist of a variable number of signals (wavenumbers), we conducted resampling of the spectra to a unified resolution of ≈ 21.33 cm⁻¹ across the range of 0.21–5376.69 cm⁻¹ through spectral binning. To determine the optimal number of bins, we calculated the average percentage of non-empty bins for each molecule. With 252 bins the average percentage of non-empty bins is approximately 50%. It means that each molecule is well-represented across the spectral range. The resulting spectral resolution of 21.33 cm⁻¹ is very close to that used by Kovács et al.⁴⁸ – 21.39 cm⁻¹. We observed that even after spectral binning, each spectrum still contains a large gap between 1500 and 3000 cm⁻¹. The low-frequency region of the spectrum contains most of the signals, while the highfrequency part includes only a few. The high-frequency region (from $\approx 3000 \text{ cm}^{-1}$) corresponds primarily to stretching vibrations along the C-H bonds or those associated with functional groups and radicals,^{13, 52, 53} such as -CH₃, -OH, -CH₂•. Meanwhile, signals in the low-frequency region, up to approximately 1500 cm⁻¹, are primarily attributed to deformation vibrations in aromatic rings. As the gap region contains zero-value points, the low-frequency region (105 bins from 0.21 to 2219.07 cm⁻¹) was selected as the target for machine learning predictions in this study, as it encompasses the majority of spectral signals.

The XYZ format provided in the database⁵² is not straightforward regarding the molecular structure, as it does not contain direct information on the bonds between atoms and other atom properties, but only their coordinates. Therefore, it lacks crucial information for spectra prediction. There are numerous molecular representations that reflect structural formula. One such method is the Simplified Molecular Input Line Entry System (SMILES) representation.⁵⁴ Molecular SMILES strings store information about atom types and properties, bond types, stereochemistry, and aromaticity. Conversion of molecules from XYZ format to SMILES can be performed using the functionality of the RDKit library⁵⁵ as well as with OpenBabel.⁵⁶ The initial database consists

primarily of hydrocarbons, but it also includes a few PAHs that are dehydrogenated or heterocyclic. While the presence of these molecules does not hinder the format conversion process in general, in some cases one or the other tool could not handle conversion correctly. Therefore, to transform as many molecules as possible from XYZ to SMILES, we had to use both RDKit and OpenBabel conversion algorithms. The database also contains samples that are complexes of PAHs with magnesium and iron ions, such as Mg⁺, Fe⁺, Mg⁺² and Fe⁺². Since the interaction between the metal ion and the neutral PAH molecule is primarily electrostatic,^{57, 58} it is not possible to include such information in SMILES. Therefore, we decided to include such molecules by designating the interaction in these complexes as an "unspecified" bond between the iron or magnesium ion and a carbon atom in a specified ring, depending on the ion's position in the molecule according to the database⁵² (Figure S2). However, the database still included molecules for which valid SMILES corresponded to different molecular structures. These molecules were excluded from the dataset, resulting in a final processed set consisting of 4137 unique molecules.

2.2. Train test split.

An important aspect of data in NASA Ames PAH IR Spectroscopic Database is the presence of a notable number of molecular ions, comprising nearly a third of the dataset (Figure 1). But the delocalized molecular charge cannot be described by topological descriptors, including SMILES, so the search for an alternative approach for charge encoding was a separate task, that will be discussed in the following sections. Figure 1 also shows a significant decrease in the number of molecules as the number of constituent heavy atoms increases above the values of \approx 50-70.



Figure 1. Distribution of molecules by charge and size. The number at each violin plot represents the total number of molecules of the given charge.

Both distributions across charges and the number of atoms must be considered when partitioning the data into training, validation, and test sets to ensure the preservation of the original distributions across all three subsets. To achieve this, we propose a custom train-test split method. First, the dataset is divided into quarters based on the number of atoms in molecules. Subsequently, within each quarter, the molecules of each charge are distributed into the training, validation, and test subsets in 70:15:15 ratio (number of molecules: 2898, 620, 619 respectively) for GNN and into the training and test subsets in 85:15 ratio (number of molecules: 3518, 619 respectively) for XGBoost (test subset is the same for GNN and XGBoost models), ensuring that the original charge state distribution is preserved in each subset.

2.3. Molecular Descriptors and models

2.3.1. Classical machine learning approach. Prior to any learning step, SMILES representations of molecules should be converted into a numerical format suitable for input into a model. Usually, molecular descriptors are employed as numerical representations of a molecule's structural and chemical characteristics in tasks of similarity search or ML prediction of properties. These descriptors encode comprehensive information on the types and properties of constituent

atoms, their local environments, and bonding patterns. Among these methods are molecular fingerprints, which result from fragmenting molecules and encoding these fragments into bit vectors. The position of a number in a molecular fingerprint corresponds to a structural fragment, while the value indicates the number of such fragments or their absence. In this work, we use the Morgan Count Bit Vector (Figure 2), where bits represent the count of each fragment in molecule formed by Morgan's algorithm.⁵⁹ The number of bits is a variable parameter that can be set directly, enabling precise control over the fingerprint's resolution and sensitivity in representing molecular features. This approach also avoids the challenges observed earlier⁴⁸ with the molecular fingerprint generation, when the number of bits for the Morgan Count Bit Vector equal to 2048, as this number of fragments provides detailed enough representation of the structure of each molecule in the dataset (Figure S3).

To enable the distinction between molecular representations with varying charges, we used a one-hot encoding method. The charge of each molecule was converted into a 5-bit vector that was concatenated with the corresponding molecular fingerprint. Thus, each molecule is represented by a unique feature vector of length 2053.



Figure 2. Molecular fragmentation into a bit vector: a) the structure of the molecule; b) molecular fingerprints; c) Morgan count bit vector

After data transformation, the molecular fingerprints were fed into the XGBoost model to predict the IR spectra. Hyperparameters optimization of the XGBoost model was done with the openaccess Optuna library⁶⁰ with TPESampler and 5-fold cross-validation to ensure resource efficiency, model robustness and to prevent overfitting. The ranges of hyperparameter optimization, as well as the optimal values, are given in (Table S1).

2.3.2 Graph Neural Network (GNN). Every molecule can be represented as a graph, defined as a set of nodes corresponding to the atoms in the molecule and edges indicating the connections between nodes, thus reflecting the chemical bonds (Figure 3).



Figure 3. A molecular graph representation. The different bond types are marked as follows: yellow square – aromatic; black diamonds – single; green triangle – double.

Graph neural networks (GNNs) are employed for interacting with molecular graphs and extracting valuable information. These networks rely on such operations as graph convolution and message passing. GNNs receive vector representations of nodes and edges, as well as an adjacency matrix or pairwise connection tensors, which are encoded from atom and bond properties, such as atom type, bond type, charge, and hybridization, etc. The message-passing mechanism involves the exchange of information between adjacent nodes. Each atom is assigned an embedding vector that incorporates both its own features and the embeddings of the atoms to which it is chemically bonded. Through the convolution operation, messages are aggregated and averaged, leading to an update of atom embeddings at each convolutional layer. Additionally, bond embeddings, which facilitate message transmission between atoms, may also be included.^{51, 61}

The architecture of the graph neural network developed in this study is comprised of two primary components: a graph block and a fully connected block. The graph block consists of graph convolution layers enhanced with an attention mechanism,⁶² which computes weight coefficients for each neighboring atom. The Graph Attention Network (GAT) layer convolutions allow for varying contributions from each atom during the update of feature embeddings, thereby enabling the model to capture the differential influence of atoms within the molecular structure more effectively. For designing the GNN architecture, we employed the PyTorch Geometric library,⁶³ which provides a comprehensive set of tools for working with graph-structured data. At the output of the graph convolution block, the obtained node feature embeddings of each molecule are aggregated and pooled to form molecular (graph) fingerprints. Subsequently, the embeddings are processed through fully connected layers, resulting in the predicted IR spectrum at the output. The described architecture of the graph neural network is illustrated in Figure 4.



Figure 4. Graph Neural Network Architecture

As well as in the case of classical ML models described above, the process of learning molecular embeddings does not account for molecular charge as graph representations are purely topological mappings of molecules. To encode the charges, we employed the learnable embeddings from the PyTorch library,⁶⁴ which generate vector representations that are updated during the model's training process. These embeddings are concatenated with the molecular embeddings following the pooling stage. Optimization of GNN parameters, including the number of graph and linear layers, channel dimensions, and the number of heads in the GAT cell, was also performed using Optuna.

2.4. Loss function and metrics

Considering the complex structure of the targets (IR spectra), where both peak intensity and position are crucial and values within a peak are inherently correlated, the choice of the loss function is particularly important. It must ensure that the optimization process produces physically meaningful results while effectively minimizing the loss function. When choosing a loss function for the prediction of IR spectra of PAH molecules, it is essential to consider two major features: the accuracy of predicted signal position and the spectral shape, determined by the relative signal intensities and signal widths. Additionally, given the relatively small dataset size (4137 molecules), it is advisable to select loss functions that are robust to outliers and provide strong generalization capabilities. Classical loss functions, such as Mean Absolute Error (MAE) and Mean Squared Error (MSE), while easy to interpret, are not well-suited under these conditions.

MSE is sensitive to outliers, whereas MAE may impose insufficient penalties on local deviations. Therefore, we prefer the Huber Loss function. It serves as a compromise between MAE and MSE: it behaves quadratically for small deviations and linearly for larger ones. This balance allows it to maintain robustness to outliers while simultaneously focusing on significant errors. The Huber Loss function is defined as follows:

$$L_{\delta}(y,\hat{y}) = \begin{cases} \frac{1}{2} * (y - \hat{y})^2, & \text{if } |y - \hat{y}| \le \delta, \\ \delta * |y - \hat{y}| - \frac{1}{2} * \delta, & \text{otherwise } |y - \hat{y}| > \delta, \end{cases}$$

where:

y-target

 \hat{y} – predicted value

 δ - a hyperparameter that sets the threshold at which the loss function transitions from quadratic to linear. Another function that combines quadratic and linear behavior is the Pseudo-Huber Loss:

$$L_{\delta}(y,\hat{y}) = \delta^{2}(\sqrt{1 + (\frac{y - \hat{y}}{\delta})^{2} - 1})$$

Unlike Huber Loss, Pseudo-Huber Loss is a smooth function, providing a more gradual transition between MSE and MAE.

Another important aspect is the selection of a metric that most accurately captures the difference between predicted and true values. If we consider the IR spectra represented as intensity vectors, the cosine similarity metric can be used to calculate the difference between them. It is computed as $\cos(A, B) = \frac{A \cdot B}{\|A\| \|B\|}$, where $A \cdot B$ is the dot product of vectors and $\|A\|$, $\|B\|$ are the norms of vectors A and B, respectively. Cosine similarity measures the angular convergence of vectors based on their orientation. For spectra represented by vectors with non-negative intensity values, cosine similarity ranges from 0 to 1, where 0 indicates that the vectors are orthogonal (the least similarity), and 1 means the vectors have the same direction (identical spectra). The advantage of cosine similarity lies in its sensitivity to signal position and relative intensities within spectra. It does not account for differences in absolute intensity values between two spectra, which is

particularly useful when spectra are not normalized to a single scale. Implementation of cosine similarity as a loss function in PyTorch⁶⁵ is called Cosine Embedding Loss.

Another metric for ML spectra prediction applied by Kovács et al.,⁴⁸ is the Earth Mover's Distance (EMD), or alternatively the Wasserstein metric, evaluates the difference between two distributions (spectra) by defining "distance" as a measure of their dissimilarity. These distributions are thought of as histograms and EMD is interpreted as the minimum amount of work needed to redistribute values across bins until the two histograms are fully aligned. A lower EMD value indicates a higher similarity between the distributions. Mathematically, it is expressed as follows:

 $EMD(a,b) = \sum_i |\sum_{j \le i} (a_j - b_j)|,$

where:

 a_i , b_i – are the *i-th* elements of distributions a and b, respectively. Applied to spectra, EMD ensures that peak positions, including their relative alignment, are considered. This aspect can support more accurate reconstruction of the predicted spectral shapes. However, the absolute values of spectra intensity also contribute to EMD, making it essential to normalize the spectra beforehand to ensure optimal algorithm performance.

3. RESULTS

3.1. XGBoost and GNN predictions

All the spectra predicted by XGBoost discussed in this section were obtained using the optimized model trained with the built-in Pseudo-Huber loss function. For a fair comparison of classic ML and GNN models, as well as for GNN optimization, we required the same metric. To determine the most suitable metric, we trained two instances of GNNs, optimized with EMD loss and Cosine Embedding Loss. Rather than focusing on their absolute predictive accuracy, we thoroughly analyzed patterns in their predictions to identify the most appropriate metric.



Figure 5. Predictions of PAH IR spectra using GNN trained with: a) Earth Mover's Distance (EMD); b) Cosine Embedding Loss as the loss function.

Without delving into the numerical or statistical analysis of cosine similarity and EMD values, we observed that using EMD loss (Figure 5a) resulted in broader and less aligned predicted signals (red curves) compared to the true signals (green curves), particularly the most intense peaks. In contrast, training the model with Cosine Embedding Loss (Figure 5b) led to predictions that exhibited a closer alignment in both the shape and intensities of the target spectra. Absolute values of cosine similarity and EMD for cases a and b (Figure 5) show that cosine similarity metrics correlate better with the general ideas on the quality of predictions (match in peak position and signal shape with the target spectrum). For example, for the top-row molecule, cosine similarity indicates that model b prediction is more accurate (0.96 vs 0.74), while the EMD metric leads to opposite: EMD of 1.6 and 2.4 in case of a and b, respectively. But the discrepancies between the predicted and the target spectra in case of model a are more obvious: prediction exhibits worse

resolution between major peaks (800-1000cm⁻¹), non-zero background from \approx 1100 to 1500 cm⁻¹ and broader signals in general. The latter is of major importance since the overall spectral resolution (21.33 cm⁻¹) is already far from ideal due to lack of data. Therefore, we chose cosine similarity as the metric function for all further studies, and we consider that it correctly indicates the quality of predicted spectra.

Table 1. Predictive performance of XGBoost and GNN models.

Model (Loss function)	Cosine similarity
XGBoost (Pseudo-Huber)	0.789
GNN (Cosine Embedding Loss)	0.764
GNN (Huber loss)	0.740
GNN (EMD)	0.717

The average values of cosine similarity on a common test set for all models are summarized in Table 1. The mean metric values reveal an expected trend for a group of GNNs, where the highest average cosine similarity is observed in the predictions made by the model trained using the Cosine Embedding Loss function. However, while it shows the best average prediction of all GNNs, the model trained with the Huber Loss excels in individual prediction accuracy, achieving the highest cosine similarity score for a single molecule (0.995 vs 0.992 for the one with Cosine Embedding Loss) and also mitigates the worst-case error giving the worst value of cosine similarity of 0.115 opposed to 0.049 for the model with Cosine Embedding Loss, it is evident that Cosine Embedding Loss is more suitable for predicting the IR spectra of PAHs due to its overall quality and consistency of results. The closeness of the results achieved by the XGBoost and GNN models (Table 1) demonstrates their ability to predict spectra with high accuracy. However, deeper insight into the nature of errors of both classes is required for recommendation of the best model.

 Table 2. Predictive performance of models based solely on structural representations of PAH molecules.

Molecular charge	Cosine similarity	
	XGBoost	GNN
-1	0.742	0.705
0	0.707	0.716
+1	0.693	0.654
+2	0.634	0.511
+3	0.673	0.705
Average	0.694	0.704

To demonstrate the impact of charge on the spectra and to enable a comparison between the developed models as well as with the results of Kovács et al. study,⁴⁸ we trained the models without inclusion of charge representation. This led to moderate prediction accuracy (Table 2), with average cosine similarity of 0.694 and 0.704 for XGBoost and GNN, respectively. These findings highlight the necessity of explicitly incorporating molecular charges into molecular representations.

As a next step, we carefully considered the variability of molecules across their size and charge while defining encoding and splitting strategies, so the assessment of their efficacy is needed. We analyze IR spectra of structurally identical PAHs in different charge states from the test set to assess the models' ability to differentiate between charged and neutral molecules, identify their characteristic spectral features, and uncover patterns in the generated output. Figure 6 illustrates examples of such molecules and corresponding IR target (green) and predicted (red) by XGBoost and GNN spectra. These molecules are present in the test set in neutral, negatively, and positively charged forms. As distinct structural units, the three of them differ solely in size (basically, the number of aromatic rings) and share identical structural properties, consisting only of carbon and hydrogen atoms without any functional groups.



Figure 6. Prediction accuracy of IR-spectra for molecules with different charges - green line represents the target spectra, red line indicates the predicted spectra: a) XGBoost predictions; b) GNN predictions.

The target IR spectra of molecules with varying charges show notable differences. The presented spectra of PAHs with a charge of (-1) contain two prominent signals near 1200 cm⁻¹ and 1500 cm⁻¹ ¹, along with a smaller signal around 700 cm⁻¹. The one at 1200 cm⁻¹ has notable "splitting". For positively charged PAHs (+1), the spectral features resemble those observed in the case of a negative charge. However, the signals are slightly shifted. Notably, the signal near 1200 cm⁻¹ is less split compared to (-1) charge state. Another trend is observed with changes in PAH size: as the molecule size increases, the signals in the 1200 cm⁻¹ region for cationic and anionic PAHs broaden and merge into a single peak. In the case of neutral molecules, the spectral pattern differs significantly. A strong, singular signal is present near 700 cm⁻¹, with all other peaks being comparatively weaker. A much deeper discussion of assignment of spectral features across the IRrange can be found elsewhere,⁵³ but we focus primarily on the models' capabilities to follow major trends. We can state that both the XGBoost and GNN models, trained on the corresponding representations, successfully recognize key features that allow them to distinguish between molecules charges, as clearly demonstrated in Figure 6. The high values of cosine similarity confirm the effectiveness of both charge encoding approaches. The GNN performs slightly better in predicting the shapes of signals of charged molecules, for instance, more accurately depicting the "splitting" of the 1200 cm⁻¹ signal. This fact may be attributed to the enhanced representational capacity of graph embeddings, characterized by their higher dimensionality and learnable embeddings for molecular charge. At the same time, XGBoost provides more accurate predictions in the region of high wavenumbers (around 2000 cm⁻¹), where no significant signals are present for these example molecules. In this region, the predictions of the GNN for charged molecules are noisy and exhibit small fluctuations. The models' predictions also reflect differences in molecular size. The predicted spectral shapes and relative intensities differ distinctly with the change of molecular size and follow the same descriptive trends that are given for the target spectra.

Molecular charge	Cosine similarity	
	XGBoost	GNN
-1	0.800	0.801
0	0.794	0.766
+1	0.763	0.739
+2	0.736	0.541
+3	0.824	0.847
Average	0.789	0.764

Table 3. Mean metric value for each charge state

The mean cosine similarity on the test set across different charge states (Table 3) demonstrates a decline as the molecular charge increases from (0) to (+2). This trend aligns with the representativity of molecules in the dataset, as illustrated in Figure 1. The decrease in the total number of molecules in the train set with increasing charge is expected to diminish the models' learning capacity. However, the highest mean cosine similarity values are observed for molecules with charges of (-1) and (+3). Based on comparison of Table 2 and Table 3, we can conclude that the suggested approaches, based on one-hot encoding and learnable embeddings, demonstrate their effectiveness in capturing charge-related properties.

A thorough consideration of the results revealed one possible reason for low prediction accuracy cases. The accuracy is influenced by the presence of heteroatomic molecules. It gets worse when heteroatomic PAHs are included, although they remain underrepresented, and the metric shows higher values when all molecules in the subset (corresponding to a single charge state) exhibit greater structural uniformity. The IR spectra of heteroatomic PAHs (e.g. those with O or N in aromatic rings) can vary significantly from those of ordinary PAHs, and since their representativity in the dataset is extremely low, the prediction results tend to deteriorate considerably when such

molecules are present in the test set. The number of neutral molecules containing nitrogen or oxygen atoms within aromatic rings or in functional groups in the test set is 9, compared to 17 for positively charged molecules (+1) and only 1 for negatively charged molecules (-1). Notably, the highest metric value across these three charge states is observed for the negatively charged molecules (-1), while the singly positively charged molecules (+1) exhibit the lowest metric value among these three charge states. The high cosine similarity observed for the (+3) charge state can be attributed to its limited representation in the dataset, with only one molecule in the test set and six in the training set, alongside the uniformity of molecules, as indicated by the absence of heteroatomic molecules among the (+3) PAHs. More discussion of predicting spectra of heteroatomic molecules will follow in section 3.2.

3.2. The past, present, and future of PAH IR spectra prediction

Firstly, and most importantly, we introduce a harmonious expansion of the concept of ML-based prediction of PAH IR spectra by addressing, for the first time, the problem of predicting spectra across different charge states of PAHs. Secondly, our approach to molecular representations – using Morgan Count Bit Vector for XGBoost – offers a more universal and efficient alternative to the previously used Morgan Fingerprints for PAH IR-spectra prediction.⁴⁸ This representation provides a reduced dimensionality that can be controlled directly, enhancing its versatility. Finally, to ensure a statistically based comparison of the prediction quality we trained optimal instances of our XGBoost and GNN models, as well as the fully connected network proposed by Kovács et al.,⁴⁸ and tested their performance on a unified test set composed exclusively of neutral PAH molecules. In the latter case, charged PAHs were excluded from the test set, as the model was not designed to account for molecular charge. The training set described in Section 2.2 was used for our models. The distributions of cosine similarity values on the test set are illustrated in Figure 7.



Figure 7. PAH IR spectra prediction accuracy by models: a) XGBoost; b) GNN; c) Fully Connected Neural Network proposed by Kovács et al.⁴⁸.

The higher mean cosine similarity values for both the XGBoost and GNN models than for fully connected neural network indicates the obvious benefits of suggested molecular and charge encodings as well as the model architecture in terms of average precision of neutral PAH IR spectra prediction. Analysis of quartiles ranges of these distributions provides some additional insights. The maximum achievable cosine similarity for a single molecule from the test set by the XGBoost and GNN models (0.992 and 0.989, respectively) exceeds that of the fully connected neural network (0.971). The maxima of probability density function (PDF) values are also located at higher metric values in the case of both of our models. A similar trend is observed at the third quartile and the median value ranges. However, the shape of PDF distribution remains nearly the

same across all three models. The situation differs in the lowest quarter, where most molecules from the underrepresented class of heteroatomic molecules are found. This range exhibits a relatively flat distribution with the highest Q1 (25% boundary) value for the XGBoost model, while the worst-case cosine similarity is comparable across all models with a slight advantage for the fully connected neural network. Indeed, upon examining the molecules with the lowest accuracy, a significant proportion of PAHs containing heteroatoms, either within the aromatic ring or as functional groups, is observed. Examples of structures and spectra of such molecules are presented in Figure 8.



Figure 8. Predicted IR spectra by GNN for heteroatomic PAH molecules.

Moreover, the group of PAHs with poorest prediction accuracy includes dehydrogenated molecules composed solely of carbon atoms. Since there is a limited number of heteroatomic PAHs in the dataset, to tackle the challenge of low prediction accuracy of the IR spectra of heteroatomic PAH molecules, we assumed that a transfer learning approach can be employed in future study. It would involve implementing a method commonly used in neural networks domain, where a pretrained on a large dataset model is further fine-tuned on a smaller but more specific. If one can select molecules structurally similar to PAHs, including those containing nitrogen and oxygen atoms, and apply a transfer learning technique, it might be possible to enhance IR spectrum predictions not only for heteroatomic PAH molecules but for all PAHs in general. The primary task is to establish a criterion for evaluating a molecule's suitability, based on its structural similarity to PAHs, for the proposed approach. Additionally, the molecule's IR spectrum should be accessible. As mentioned earlier, the high-frequency region of the spectrum features signals influenced by stretching vibrations within functional groups (e.g., NH₂, OH). Including this spectral region may improve the model's predictive accuracy for heteroatomic PAHs. We have identified the primary limitations of our models and outlined the possible future direction for enhancing PAHs IR-spectra prediction.

4. CONCLUSIONS

We have demonstrated the predictive power of both classical ML approaches and graph neural networks trained on the largest dataset of DFT-calculated IR spectra of PAH molecules available to date. The XGBoost model currently represents the most effective approach for spectral prediction tasks, as evidenced by its consistently superior performance on average, and in extreme cases, compared to prior methodologies and the GNN model.

Despite this, the GNN model shows significant promise for future advancements. Its potential lies in its inherently flexible and expressive molecular graph encoding framework, which facilitates detailed and nuanced representations of molecular structures. The GNN's ability to incorporate transfer learning provides an opportunity to further improve its predictive performance and broaden its applicability.

This study marks a step forward by enabling the prediction of PAH ion spectra through comprehensive encoding of molecular charge states. Accurate and detailed encoding has been shown to enhance the versatility and applicability of predictive models considerably. The achieved accuracy—both averaged across the independent test set and within specific charge states and molecular size categories—enables reliable prediction of spectra for a wide range of PAH structures. These advancements pave the way for precise modeling of PAH mixture spectra, addressing the critical challenge of investigating the compositional properties of astronomical objects.

ASSOCIATED CONTENT

Data Availability

The dataset, models and code used to generate the results in this paper are available on Zenodo⁶⁶. The dataset files are also provided as Supporting Information.

Supporting Information.

The Supporting Information is available free of charge at

- Supplementary figures and tables referenced in the text. (Supplementary.doc)
- A database of PAH molecules containing UIDs, SMILES, and additional data. (database.csv)
- The test set used to evaluate the prediction accuracy of PAH IR spectra by the XGBoost and GNN models. (database_test.csv)

AUTHOR INFORMATION

Corresponding Author

Timur A. Labutin - Lomonosov Moscow State University, 119234 Moscow, Russia; Email: timurla@laser.chem.msu.ru

Author Contributions

The manuscript was prepared through the contributions of all authors. All authors have given approval to the final version of the manuscript.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENT

The work was supported by the Non-commercial Foundation for the Advancement of Science

and Education INTELLECT. T.A.L.'s and A.S.Z.'s work was conducted under the state

assignment of Lomonosov Moscow State University (Project No. 121031300173-2).

REFERENCES

1. Allamandola, L. J.; Hudgins, D. M.; Sandford, S. A., Modeling the Unidentified Infrared Emission with Combinations of Polycyclic Aromatic Hydrocarbons. *Astrophys. J.* **1999**, *511* (2), L115.

2. Snow, T. P.; Le Page, V.; Keheyan, Y.; Bierbaum, V. M., The interstellar chemistry of PAH cations. *Nature* **1998**, *391* (6664), 259-260.

3. Moorthy, B.; Chu, C.; Carlin, D. J., Polycyclic Aromatic Hydrocarbons: From Metabolism to Lung Cancer. *Toxicol. Sci.* **2015**, *145* (1), 5-15.

4. Ball, A.; Truskewycz, A., Polyaromatic hydrocarbon exposure: an ecological impact ambiguity. *Environ. Sci. Pollut. Res.* **2013**, *20* (7), 4311-4326.

5. Kumar, S.; Negi, S.; Maiti, P., Biological and analytical techniques used for detection of polyaromatic hydrocarbons. *Environ. Sci. Pollut. Res.* **2017**, *24* (33), 25810-25827.

6. Jin, H.; Yuan, W.; Li, W.; Yang, J.; Zhou, Z.; Zhao, L.; Li, Y.; Qi, F., Combustion chemistry of aromatic hydrocarbons. *Prog. Energy Combust. Sci.* **2023**, *96*, 101076.

7. Michela, A.; Barbara, A.; Antonio, T.; Anna, C., Identification of large polycyclic aromatic hydrocarbons in carbon particulates formed in a fuel-rich premixed ethylene flame. *Carbon* **2008**, *46* (15), 2059-2066.

8. Boström, C.-E.; Gerde, P.; Hanberg, A.; Jernström, B.; Johansson, C.; Kyrklund, T.; Rannug, A.; Törnqvist, M.; Victorin, K.; Westerholm, R., Cancer risk assessment, indicators, and guidelines for polycyclic aromatic hydrocarbons in the ambient air. *Environ. Health Perspect.* **2002**, *110* (suppl 3), 451-488.

9. Wenzel, G.; Cooke, I. R.; Changala, P. B.; Bergin, E. A.; Zhang, S.; Burkhardt, A. M.; Byrne, A. N.; Charnley, S. B.; Cordiner, M. A.; Duffy, M.; Fried, Z. T. P.; Gupta, H.; Holdren, M. S.; Lipnicky, A.; Loomis, R. A.; Shay, H. T.; Shingledecker, C. N.; Siebert, M. A.; Stewart, D. A.; Willis, R. H. J.; Xue, C.; Remijan, A. J.; Wendlandt, A. E.; McCarthy, M. C.; McGuire, B. A., Detection of interstellar 1-cyanopyrene: A four-ring polycyclic aromatic hydrocarbon. *Science* **2024**, *386* (6723), 810-813.

10. Grübel, F.; Molaverdikhani, K.; Ercolano, B.; Rab, C.; Trapp, O.; Dubey, D.; Arenales-Lope, R., Detectability of polycyclic aromatic hydrocarbons in the atmosphere of WASP-6 b with JWST NIRSpec PRISM. *Mon. Not. R. Astron. Soc.* **2025**, *536*, 324-339.

11. Clemett, S. J.; Maechling, C. R.; Zare, R. N.; Swan, P. D.; Walker, R. M., Identification of Complex Aromatic Molecules in Individual Interplanetary Dust Particles. *Science* **1993**, *262* (5134), 721-725.

12. Li, A. In *PAHs in Comets: An Overview*, Deep Impact as a World Observatory Event: Synergies in Space, Time, and Wavelength, Berlin, Heidelberg, 2009//; Käufl, H. U.; Sterken, C., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2009; pp 161-175.

13. Tielens, A. G. G. M., Interstellar Polycyclic Aromatic Hydrocarbon Molecules. *Annu. Rev. Astron. Astrophys.* **2008**, *46* (Volume 46, 2008), 289-337.

14. Li, A., Spitzer's perspective of polycyclic aromatic hydrocarbons in galaxies. *Nature* Astronomy **2020**, *4* (4), 339-351.

15. Pejcic, B.; Boyd, L.; Myers, M.; Ross, A.; Raichlin, Y.; Katzir, A.; Lu, R.; Mizaikoff, B., Direct quantification of aromatic hydrocarbons in geochemical fluids with a mid-infrared attenuated total reflection sensor. *Org. Geochem.* **2013**, *55*, 63-71.

16. Moutou, C.; Sellgren, K.; Verstraete, L.; Léger, A., Upper limit on C60 and C60+ features in the ISO-SWS spectrum of the reflection nebula NGC 7023. *Astron Astrophys* **1999**.

17. Croiset, B. A.; Candian, A.; Berné, O.; Tielens, A. G. G. M., Mapping PAH sizes in NGC 7023 with SOFIA. *Astron Astrophys* **2016**, *590*, A26.

18. Guzman-Ramirez, L.; Lagadec, E.; Jones, D.; Zijlstra, A. A.; Gesicki, K., PAH formation in O-rich planetary nebulae. *Mon. Not. R. Astron. Soc.* **2014**, *441* (1), 364-377.

19. Peeters, E.; Hony, S.; Van Kerckhoven, C.; Tielens, A. G. G. M.; Allamandola, L. J.; Hudgins, D. M.; Bauschlicher, C. W., The rich 6 to 9 μm spectrum of interstellar PAHs*. *Astron Astrophys* **2002**, *390* (3), 1089-1113.

20. Murga, M. S.; Kirsanova, M. S.; Wiebe, D. S.; Boley, P. A., Orion Bar as a window to the evolution of PAHs. *Mon. Not. R. Astron. Soc.* **2021**, *509* (1), 800-817.

21. Acke, B.; Bouwman, J.; Juhász, A.; Henning, T.; van den Ancker, M. E.; Meeus, G.; Tielens, A. G. G. M.; Waters, L. B. F. M., Spitzer's view on aromatic and aliphatic hydrocarbon emission in Herbig Ae stars. *Astrophys. J.* **2010**, *718* (1), 558.

22. Seok, J. Y.; Li, A., Polycyclic Aromatic Hydrocarbons in Protoplanetary Disks around Herbig Ae/Be and T Tauri Stars. *Astrophys. J.* **2017**, *835* (2), 291.

23. Genzel, R.; Lutz, D.; Sturm, E.; Egami, E.; Kunze, D.; Moorwood, A. F. M.; Rigopoulou, D.; Spoon, H. W. W.; Sternberg, A.; Tacconi-Garman, L. E.; Tacconi, L.; Thatte, N., What Powers Ultraluminous IRAS Galaxies? *Astrophys. J.* **1998**, *498* (2), 579.

24. Homann, K.-H., Fullerenes and Soot Formation— New Pathways to Large Particles in Flames. *Angew. Chem.* **1998**, *37* (18), 2434-2451.

25. Murga, M. S., Evolution of carbon particles from the stage of asymptotic giant branch stars to planetary nebulae: observations, experiments, and theory. *Phys.-Uspekhi* **2023**, *67* (10), 961-987.

26. Tommasini, M.; Lucotti, A.; Alfè, M.; Ciajolo, A.; Zerbi, G., Fingerprints of polycyclic aromatic hydrocarbons (PAHs) in infrared absorption spectroscopy. *Spectrochim. Acta A* **2016**, *152*, 134-148.

27. Pentsak, E. O.; Murga, M. S.; Ananikov, V. P., Role of Acetylene in the Chemical Evolution of Carbon Complexity. *ACS Earth Space Chem.* **2024**, *8* (5), 798-856.

28. Oomens, J.; Tielens, A. G. G. M.; Sartakov, B. G.; von Helden, G.; Meijer, G., Laboratory Infrared Spectroscopy of Cationic Polycyclic Aromatic Hydrocarbon Molecules. *Astrophys. J.* **2003**, *591* (2), 968.

29. Lorenz, U. J.; Solcà, N.; Lemaire, J.; Maître, P.; Dopfer, O., Infrared Spectra of Isolated Protonated Polycyclic Aromatic Hydrocarbons: Protonated Naphthalene. *Angew. Chem.* **2007**, *46* (35), 6714-6716.

30. Kamer, J.; Schleier, D.; Jiao, A.; Schneider, G.; Martens, J.; Berden, G.; Oomens, J.; Bouwman, J., IR spectra of cationic 1,5,9-triazacoronene and two of its cationic derivatives. *Phys. Chem. Chem. Phys.* **2024**, *26* (44), 27912-27921.

31. Martens, J.; Berden, G.; Gebhardt, C. R.; Oomens, J., Infrared ion spectroscopy in a modified quadrupole ion trap mass spectrometer at the FELIX free electron laser laboratory. *Rev. Sci. Instrum.* **2016**, *87* (10).

32. Palotás, J.; Martens, J.; Berden, G.; Oomens, J., Laboratory IR spectroscopy of protonated hexa-peri-hexabenzocoronene and dicoronylene. *J. Mol. Spectrosc.* **2021**, *378*, 111474.

33. Esposito, V. J.; Fortenberry, R. C.; Boersma, C.; Allamandola, L. J., High-Resolution Farto Near-Infrared Anharmonic Absorption Spectra of Cyano-Substituted Polycyclic Aromatic Hydrocarbons from 300 to 6200 cm-1. *ACS Earth Space Chem.* **2024**, *8* (9), 1890-1900.

34. Lusci, A.; Pollastri, G.; Baldi, P., Deep Architectures and Deep Learning in Chemoinformatics: The Prediction of Aqueous Solubility for Drug-Like Molecules. *J. Chem. Inf. Model.* **2013**, *53* (7), 1563-1575.

35. Boobier, S.; Hose, D. R. J.; Blacker, A. J.; Nguyen, B. N., Machine learning with physicochemical relationships: solubility prediction in organic solvents and water. *Nat. Commun.* **2020**, *11* (1), 5753.

36. Maltarollo, V. G.; Gertrudes, J. C.; Oliveira, P. R.; Honorio, K. M., Applying machine learning techniques for ADME-Tox prediction: a review. *Expert Opin. Drug Metab. Toxicol.* **2015**, *11* (2), 259-271.

37. Zhang, J.; Mucs, D.; Norinder, U.; Svensson, F., LightGBM: An Effective and Scalable Algorithm for Prediction of Chemical Toxicity–Application to the Tox21 and Mutagenicity Data Sets. *J. Chem. Inf. Model.* **2019**, *59* (10), 4150-4158.

38. Galeazzo, T.; Shiraiwa, M., Predicting glass transition temperature and melting point of organic compounds via machine learning and molecular embeddings. *Environ Sci Atmos.* **2022**, *2* (3), 362-374.

39. Stokes, J. M.; Yang, K.; Swanson, K.; Jin, W.; Cubillos-Ruiz, A.; Donghia, N. M.; MacNair, C. R.; French, S.; Carfrae, L. A.; Bloom-Ackermann, Z.; Tran, V. M.; Chiappino-Pepe, A.; Badran, A. H.; Andrews, I. W.; Chory, E. J.; Church, G. M.; Brown, E. D.; Jaakkola, T. S.; Barzilay, R.; Collins, J. J., A Deep Learning Approach to Antibiotic Discovery. *Cell* **2020**, *180* (4), 688-702.e13.

40. Jiang, D.; Wu, Z.; Hsieh, C.-Y.; Chen, G.; Liao, B.; Wang, Z.; Shen, C.; Cao, D.; Wu, J.; Hou, T., Could graph neural networks learn better molecular representation for drug discovery? A comparison study of descriptor-based and graph-based models. *J. Cheminform.* **2021**, *13* (1), 12. 41. Wieder, O.; Kohlbacher, S.; Kuenemann, M.; Garon, A.; Ducrot, P.; Seidel, T.; Langer,

T., A compact review of molecular property prediction with graph neural networks. *Drug Discov. Today* **2020,** *37*, 1-12.

42. Zakuskin, A. S.; Labutin, T. A., StarkML: application of machine learning to overcome lack of data on electron-impact broadening parameters. *Mon. Not. R. Astron. Soc.* **2023**, *527* (2), 3139-3145.

43. 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* (9), 1801367.

44. McGill, C.; Forsuelo, M.; Guan, Y.; Green, W. H., Predicting Infrared Spectra with Message Passing Neural Networks. *J. Chem. Inf. Model.* **2021**, *61* (6), 2594-2609.

45. Enders, A. A.; North, N. M.; Fensore, C. M.; Velez-Alvarez, J.; Allen, H. C., Functional Group Identification for FTIR Spectra Using Image-Based Machine Learning Models. *Anal Chem* **2021**, *93* (28), 9711-9718.

46. Koshelev, D. S., Expert System for Fourier Transform Infrared Spectra Recognition Based on a Convolutional Neural Network With Multiclass Classification. *Appl Spectrosc* **2024**, *78* (4), 387-397.

47. McCarthy, M.; Lee, K. L. K., Molecule Identification with Rotational Spectroscopy and Probabilistic Deep Learning. *J. Phys. Chem. A* **2020**, *124* (15), 3002-3017.

48. 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* (2), 100.

49. Maragkoudakis, A.; Peeters, E.; Ricca, A., Probing the size and charge of polycyclic aromatic hydrocarbons. *Mon. Not. R. Astron. Soc.* **2020**, *494* (1), 642-664.

50. Chen, T.; Guestrin, C., XGBoost: A Scalable Tree Boosting System. In *Proceedings of the* 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, Association for Computing Machinery: San Francisco, California, USA, 2016; pp 785–794.

51. Kipf, T. N.; Welling, M., Semi-supervised classification with graph convolutional networks. In *International Conference on Learning Representations*, Toulon, France, 2016.

52. Boersma, C.; Bauschlicher, C. W.; Ricca, A.; Mattioda, A. L.; Cami, J.; Peeters, E.; de Armas, F. S.; Saborido, G. P.; Hudgins, D. M.; Allamandola, L. J., The NASA Ames PAH IR

Spectroscopic Database version 2.00: Updated Content, Web Site, and On(Off)Line Tools. *Astrophys. J. Suppl. Ser.* 2014, 211 (1), 8.

53. Ricca, A.; Roser, J. E.; Peeters, E.; Boersma, C., Polycyclic Aromatic Hydrocarbons with Armchair Edges: Potential Emitters in Class B Sources. *Astrophys. J.* **2019**, *882* (1), 56.

54. Weininger, D., SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Model.* **1988**, *28* (1), 31-36.

55. Landrum, G. The RDKit documentation. https://www.rdkit.org/docs/ (accessed December6).

56. O'Boyle, N. M.; Banck, M.; James, C. A.; Morley, C.; Vandermeersch, T.; Hutchison, G. R., Open Babel: An open chemical toolbox. *J. Cheminform.* **2011**, *3* (1), 33.

57. Bauschlicher, C. W.; Ricca, A., The Far-Infrared Emission from the Mg⁺–PAH Species. *Astrophys. J.* **2009**, *698* (1), 275.

58. Bauschlicher Jr, C. W., Fe+- and Mg+-polycyclic aromatic hydrocarbon complexes. *Mol. Phys.* **2009**, *107* (8-12), 809-818.

59. Morgan, H. L., The Generation of a Unique Machine Description for Chemical Structures-A Technique Developed at Chemical Abstracts Service. *J. Chem. Inf. Model.* **1965**, *5* (2), 107-113.

60. Akiba, T.; Sano, S.; Yanase, T.; Ohta, T.; Koyama, M., Optuna: A Next-generation Hyperparameter Optimization Framework. In *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, Association for Computing Machinery: Anchorage, AK, USA, 2019; pp 2623–2631.

61. Gilmer, J.; Schoenholz, S. S.; Riley, P. F.; Vinyals, O.; Dahl, G. E., Neural Message Passing for Quantum Chemistry. In *Proceedings of the 34th International Conference on Machine Learning*, Doina, P.; Yee Whye, T., Eds. PMLR: Proceedings of Machine Learning Research, 2017; Vol. 70, pp 1263--1272.

62. Velickovic, P.; Cucurull, G.; Casanova, A.; Romero, A.; Lio, P.; Bengio, Y., Graph attention networks. In *International Conference on Learning Representations*, Vancouver, Canada, 2018; Vol. 1050, pp 10-48550.

63. PyTorch Geometric. https://pytorch-geometric.readthedocs.io/en/latest/# (accessed December 6).

64. PyTorch Embedding. https://pytorch.org/docs/stable/generated/torch.nn.Embedding.html (accessed December 6).

65. PyTorch

CosineEmbeddingLoss.

https://pytorch.org/docs/stable/generated/torch.nn.CosineEmbeddingLoss.html (accessed December 6).

66. Beglaryan, B. G.; Zakuskin, A. S.; Nemchenko, V. A.; Labutin, T. A. Towards Accurate PAH IR Spectra Prediction: Handling Charge Effects with Classical and Deep Learning Models. DOI: <u>10.5281/zenodo.14894779</u>.