Machine-learning Interpretation of the Correlation between Infrared Emission Features of Interstellar Polycyclic Aromatic Hydrocarbons

Zhisen Meng1, Xiaosi Zhu1, Péter Kovács2, Enwei Liang1, and Zhao Wang1

1 Laboratory for Relativistic Astrophysics, Department of Physics, Guangxi University, 530004 Nanning, People’s Republic of China; zw@gxu.edu.cn
2 Institute of Materials Chemistry, TU Wien, A-1060 Vienna, Austria

Received 2021 July 26; revised 2021 September 17; accepted 2021 September 30; published 2021 November 24

Abstract

Supervised machine-learning models are trained with various molecular descriptors to predict infrared (IR) emission spectra of interstellar polycyclic aromatic hydrocarbons. We demonstrate that a feature importance analysis based on the random forest algorithm can be utilized to explore the physical correlation between emission features. Astronomical correlations between IR bands are analyzed as examples of demonstration by finding the common molecular fragments responsible for different bands, which improves the current understanding of the long-observed correlations. We propose a way to quantify the band correlation by measuring the similarity of the feature importance arrays of different bands, by which a correlation map is obtained for emissions in the out-of-plane bending region. Moreover, a comparison between the predictions using different combinations of descriptors underscores the strong prediction power of the extended-connectivity molecular fingerprint, and shows that the combinations of multiple descriptors of other types in general lead to improved predictivity.

Unified Astronomy Thesaurus concepts: Astronomy data modeling (1859); Random Forests (1935); Infrared astronomy (786); Interstellar molecules (849); Polycyclic aromatic hydrocarbons (1280)

1. Introduction

Polycyclic aromatic hydrocarbons (PAHs) are among the most widely studied compounds in astronomy (Herbst & van Dishoeck 2009), chemistry (Zhang et al. 2015), biology (Moorothy et al. 2015), and environmental science (Ravindra et al. 2008). PAHs are thought to exhibit emission features dominating the infrared (IR) spectra of a large variety of galactic and extragalactic sources, commonly known as the “unidentified” infrared emission (UIE) features (Tielens 2008). The assessment of interstellar molecular composition by IR spectroscopy crucially relies on knowledge of the bands ascribed to characteristic vibrational modes of specific bonds or functional groups in the molecule (Allamandola et al. 1989; Hudgins & Allamandola 1999; Peeters 2011). Specifically, the 3.3 μm band is due to C–H stretching, the 6.2 μm band to C–C stretching, the 7.7 μm band to coupled C–C stretching and C–H in-plane bending, the 8.6 μm band to the C–H in-plane bending, and the 10–15 μm band to the CH out-of-plane bending, etc. Regardless, identifying PAHs remains anything but straightforward due to the vast variety of PAH molecular structures (Qi et al. 2018; Hanine et al. 2020; Li 2020). In particular, it was often observed in interstellar spectra that the IR intensities strongly correlate with each other, the underlying cause of such cross-correlations among the PAH bands are poorly understood.

For instance, strong correlations were found between the bands at 3.3 and 11.2 μm (Hony et al. 2001), and between those at 11.3 and 12.7 μm (Acke et al. 2010), in IR spectra from the regions of H II, intermediate mass star-forming, planetary nebulae and Herbig AeBe stars. These emission features are known to be associated with C–H stretching or out-of-plane bending vibration of PAHs, but their specific common structural origin is unknown. It was also reported that the 6.2, 7.7, 7.8, and 16.4 μm bands exhibit strong correlations in spectra from the Milky Way, Magellanic Clouds, and nearby galaxies (Peeters 2011). These bands were extensively ascribed to the C–C stretching or bending mode. Likewise, cross-correlations between the 6.2, 7.8, 16.4, and 8.6 μm (Vermeij et al. 2002; Acke et al. 2010; Peeters 2011), the 6.2, 7.7, 16.4, and 12.7 μm (Keller et al. 2008; Acke et al. 2010; Peeters et al. 2012), the 11.0 and 16.4 μm (Peeters 2011), and the 11.3 and 17.0 μm bands (Smith et al. 2007) were reported for spectra from different interstellar regions, while the molecular structures responsible for these correlations are unclear. More importantly, there is a lack of quantitative measurement of the band correlation for analyzing these phenomena on the ground of molecular structure or energetics, while there is a clear interest in the identification of the underlying connection among the UIE bands for assessing the possible chemical composition of interstellar medium.

A promising approach to this end is the density functional theory (DFT) calculations, which are based on a solution of the Schrödinger equation in the Kohn–Sham formulation. DFT is able to accurately describe the changes in the molecular dipole moment with a good characterization of both the dynamics of the atomic nuclei and of the electronic charge distribution. However, a brute-force application of DFT calculations is unlikely to be successful in dealing with interstellar PAH IR spectra, due to its high computational cost versus the vast possible PAH structures existing in the emitting source. More critically for the aforementioned problem on the understanding of band correlation, the complex formulations of DFT bring great challenges for finding the molecular substructure, which gives rise to a specific IR band.

The development of machine-learning (ML) methods has opened new and reliable ways of investigating molecular structure-spectrum relationships (Butler et al. 2018; Ghosh et al. 2019). ML was used for analyzing molecular IR spectra for various applications (Gastegger et al. 2017; Fu & Hopkins 2018; Márquez-Neila et al. 2018; Ye et al. 2020; Calvo et al. 2021; Joung et al. 2021; Laurens et al. 2021; Zapata Trujillo et al. 2021). In particular, for the analysis of interstellar IR spectra, Kovács et al. (2020) proposed a neutral network (NN) model for predicting IR spectrum of interstellar PAHs. Their results...
demonstrated excellent predictive skill of the NN model based on molecular fingerprints for out-of-sample inputs. Recently, McGill et al. (2021) developed a software package for the prediction of IR spectra through the use of message passing NN based on spectra from various sources. Their method is in quantitative agreement with experiment.

In this work, we demonstrate that the ML model can be used to interpret the correlation between IR bands by measuring the similarity between their feature importance arrays. We take several astronomically observed correlations as examples, and show the capacity of ML in finding the molecular fragment responsible for a specific emission band, and in calculating the correlation map between bands in a spectra region. We also aim at comparing the performance of different common molecular descriptors in the ML prediction of PAH IR spectra, the choice of which is crucial for the ML prediction of molecular properties. The results of comparison are discussed in terms of the physical correlation between the IR bands and the molecular features represented by different descriptors.

### 2. Methods

#### 2.1. Molecular Descriptors for ML

Chemistry databases commonly use the atomic coordinate (AC) for representing a molecular compound. However, for ML, AC is an ineffective descriptor because it is not invariant to translation, rotation of the molecule, or the permutation of atoms. This problem can be solved by converting AC to molecular invariants, such as the extended-connectivity fingerprint (ECFP; Rogers & Mathew 2010), the sorted eigenvalues (SEVs) of the distance matrix (DM; Ivanciuc 2000) or coulomb matrix (CM; Rupp et al. 2012; Schrier 2020). We here use six types of common molecular invariants as the descriptors, including the distance matrix eigenvalues (DMEs), the coulomb matrix eigenvalues (CMEs), ECFP, the number of H adjacency classes (NHAC; Hony et al. 2001), the zero-point vibrational energy (ZPVE) and the atomization energy (AE), as listed in Table 1. These six descriptors are compared with that using a combination of molecular structural information (CMSI), which comprises the topological index meant to quantify the molecular complexity (BertzCT; Bertz 1981), etc. (obtained via RDKit.). Among the descriptors,

| Abbreviation | Descriptor Name                  | Feature Number |
|--------------|----------------------------------|----------------|
| DMEs         | Distance matrix eigenvalues       | 432            |
| CMEs         | Coulomb matrix eigenvalues        | 432            |
| ECFP         | Extended-connectivity fingerprint | 19794 (r = 4)  |
| NHAC         | Number of H adjacency classes     | 7              |
| ZPVE         | Zero-point vibrational energy     | 1              |
| AE           | Atomization energy                | 1              |
| CMSI         | Combined molecular structural information | 19          |

![Figure 1](image-url) Normalized feature importance (histogram) and root mean square error (RMSE, curves) for different descriptors. Error bars show the mean ±1 standard error. The shadowed zone includes OCFs that lead to the optimal performance of the model with high efficiency. In panel (c), OCFs of CMSI include BertzCT, the number of radical electrons (NumRadicalElectrons), the number of valence electrons (NumValenceElectrons), the number of rings (RingCount), the number of aromatic rings (NumAromaticRings), the average molecular weight (MolWt), the number of aliphatic carbocycles (NumAliphaticCarbocycles), the average molecular weight ignoring hydrogens (HeavyAtomMolWt), and the number of heavy atoms (HeavyAtomCount).
CMEs, DMEs, ZPVE, and AE crucially rely on the DFT or force field method computation of bond lengths or energetics, and are thus less interesting for astronomical applications than ECFP, NHAC, and CMSI that only depend upon the topology of molecule. A voting regressor and a bootstrapping data set treatment are also used to further improve the model predictivity, as discussed at the end of Section 2.2.

2.2. Selection of Important Features

Using all the features available in a descriptor data set may not result in the most effective model. A selection test is therefore performed to determine the optimal combination of features (OCFs). For each molecular descriptor, we first train a testing ML model using all features, which are subsequently sorted by their importance as shown by the histogram in Figure 1. A series of models are then trained with different numbers of sorted important features (SIFs) to determine OCFs as indicated by the shadowed zone. For example, first 30, 65, 9, and 1000 SIFs are chosen for DMEs, CMEs, CMSI, and ECFP (cutoff radius \( r = 4 \)), respectively. All features are used for NHAC, ZPVE, and AE, which have \( \leq 7 \) features.

2.3. Algorithm and Data Set Construction

The random forest (RF) algorithm is used here to predict PAH IR spectra as implemented in the open-source Scikit-Learn library (Pedregosa et al. 2011). It is chosen because of its intrinsic metric for the importance of the feature that is very practical for the task of this study. RF is an estimator that comprises a root node, leaf nodes, and branches (Breiman 2001). When the decision trees are trained, each node will randomly select a
number of specific features in which the best splitting condition is evaluated. RF uses random subsets of the features to include a number of decision networks in the learning process, which makes it efficient and naturally resistant to overfitting (Ho 2002). The workflow of our ML approach is schematically illustrated in Figure 2, taking DMEs as an example. The PAH AC are first converted to DMs ((a)–(b)), based on which SEVs are computed (c). The effective features determined by the aforementioned selection test (Figure 1(a)) are used to train an RF predictor (d). The predictor is used to simulate IR spectra of the molecules in the testing set (e), and to further perform a feature importance analysis for interpreting the band correlation (f). We use the hyperparameters adjusted in the previous work of Kovács et al. (2020) for predicting the IR spectra of PAHs using a earlier version of the same data set. This parameters can be found in a Python program via doi:10.5281/zenodo.5513073.

2925 spectra are chosen from the DFT-computed spectra in the 3.20 version of the NASA AMES PAH IR spectroscopic database (PAHdb; Boersma et al. 2014; Bauschlicher et al. 2018; Mattioda et al. 2020). The spectra of all charged PAHs are discarded, since we use topological descriptors such as ECFP and NHAC, which could generate exactly the same input for a neutral molecule and for its charged counterpart. The discrete IR intensity from PAHdb is converted to a histogram with a bin width of 16.65 cm\(^{-1}\), which is determined using Knuth’s Bayesian rule as implemented in the Astropy package (Knuth 2006). Each spectrum consists of 225 bins covering the wavelength range from 2.67 to 1172.33 \(\mu\)m (8.53 to 3746.25 cm\(^{-1}\)). The histograms are separated into low-frequency (\(v_{low}\)) and high-frequency (\(v_{high}\)) regions by a cutoff of 4.44 \(\mu\)m (2252.25 cm\(^{-1}\)), in order to distinguish the signals from localized vibrations of hydrogen bonds and those from other vibrations.

A 0.8/0.2 partition is determined by a best-predictivity test for constructing the training and the testing sets. The prediction of each model is repeated five times using different sample combinations to ensure that all samples are used in the testing and the training sets. The root mean square error (RMSE) is computed by averaging over all samples in the testing set to assess the model predictivity (namely mean RMSE). The feature importance is also evaluated to help understand the correlation between molecular substructures and IR bands. Note that the Earth mover’s distance (Monge 1781) is not used since this work needs a merit for the prediction of individual bands.

3. Results and Discussion

Figure 3 compares the performance of different descriptors in predicting PAH IR spectra. They are ordered by accuracy in \(v_{low}\) and \(v_{high}\) as follows: ECFP > CMEs > DMEs > NHAC > CMSI > AE > ZPVE, and ECFP > DMEs > CMEs > NHAC > CMSI > AE > ZPVE, respectively. The prediction in \(v_{high}\) is in general more accurate than that in \(v_{low}\). IR bands are correlated with the vibration characteristics of specific substructures of the compound. It is hence not surprising that ECFP leads to the best predictivity since it contains the most comprehensive information about the molecular substructures among the compared descriptors. Besides ECFP, the atomic distance (represented by DMEs) is proven to be the most important feature in \(v_{high}\), where IR emission is simply dominated by C–H stretching. It could be expected that CMEs outperform DMEs, since CMEs include not only a distance measure of the atomic configuration like DMEs does, but also additional information about the element species. This is, however, not the case in \(v_{high}\). The species information in CMEs turns out not to be helpful, but instead likely interferes.
with distance information in SEV and eventually leads to a decreased predictivity.

To give an idea about how “good” or “poor” our prediction could be, we plot in Figure 4 the distribution of RMSE and two examples of the predicted spectra, taking the models using ECFP and CMSI as examples. It is seen in panel (a) that RMSE is in a gamma-type distribution with a larger width for $\nu_{low}$. In the two example spectra, the bands of C$_{33}$H$_{19}$ are well predicted by ECFP in terms of both position and intensity (panel (b)). Even the relatively poor prediction for C$_{33}$ (panel (c)) is informative about the peak position. This is crucial for the subsequent analysis of the feature importance, in which the features are compared relative to each other for a specific band, despite the fact that the model may fail to match its absolute intensity in the cases of large RMSE.

For each band in a predicted spectrum, the importance of a feature can be computed by simply reverting all the choices based on that feature. This feature importance can be used to search for the common origins of two IR emission bands. For example, it has long been observed that the 3.3 and 11.2 $\mu$m bands exhibit strong correlation in the spectra from H$_{II}$ regions, intermediate mass star-forming regions and planetary nebulae (Hony et al. 2001). These bands were extensively ascribed to C–H stretching and C–H in-plane bending modes (Allamandola et al. 1989). Our ML analysis using ECFP first confirms this correlation for the spectra collected in PAHdb (Figure 5(a)), and further reveals common but slightly different structural origins of these two bands as shown in Figures 5(b)–(c). It is seen that the 3.3 $\mu$m band is singularly dominated by the emission feature of individual C–H bonds, while the 11.2 $\mu$m one depends much more on the number of adjacent CH groups on the peripheral aromatic rings, as discussed by Hony et al. (2001).

Another example is the strong correlation between 6.2, 7.7, and 8.6 $\mu$m bands previously reported by Galliano et al. (2008) for the ISO- and Spitzer-observed mid-IR spectra from various interstellar regions. These bands were extensively ascribed to C–C stretching and C–H in-plane bending modes (Allamandola et al. 1989; Draine 2003). The ML analysis confirms this correlation (Figures 5(d)–(e)), and reveals the common molecular fragments that are responsible for these bands as those
contoured by the coils in Figures 5(f)–(h). The 8.6 μm band appears to exhibit a stronger correlation with the 7.7 μm one than the 6.2 μm band does. However, this is not true if we compare more common features than the five ones shown in Figure 5(d), given that a band can have many (up to $n = 1000$ as indicated in Figure 1(d)) elements in its ECFP feature importance array and each one could represent a different molecular fragment. To give a comprehensive description, we calculate the cosine similarity (Singhal 2001) between the feature importance arrays as an index of the band correlation,

$$S(A, B) = \frac{\sum_{i=1}^{n} l_i(A) l_i(B)}{\sqrt{\sum_{i=1}^{n} l_i^2(A)} \sqrt{\sum_{i=1}^{n} l_i^2(B)}}.$$  

where $l_i(A)$ is the importance of the $i$th feature of the band $A$. For the 6.2, 7.7, and 8.6 μm bands, $S(6.2, 7.7)$, $S(7.7, 8.6)$, and $S(6.2, 8.6)$ are computed to be 0.79, 0.75, and 0.68, respectively. In contrast to the comparison using only five features, the 6.2 μm band is actually more strongly correlated with the 7.7 μm band than the 8.6 μm band if all 1000 features are taken into account.

The spectral region between 10 and 15 μm exhibits rich emission due to the out-of-plane (OOP) bending vibrations of C–H bonds (Allamandola et al. 1989). These bands are in general difficult to predict by ML (with high RMSE, as shown in Figure 6(a)). The emission in this region has long been recognized to rely on the configuration of adjacent aromatic rings carrying different numbers of C–H bonds (Bellamy 1958; Hudgins & Allamandola 1999), which can be characterized by NHAC (Hony et al. 2001). Figure 5(c) showed an example for the 11.2 μm band. Figure 6(a) confirms that NHAC outperforms other descriptors except for ECFP in the OOP spectral region. However, the significance of NHAC could have been underestimated; it also gives a fair prediction for the spectrum in other IR regions.

We measure the correlation between the bands in the OOP spectral region by computing the similarity between their feature importance arrays defined by Equation (1). The result is plotted in the correlation map in Figure 6(c). It is seen that the 12.64, 12.91, 13.20, 11.02, and 11.22 μm bands are loosely correlated with the other bands, while the 14.47 and 10.44 μm, the 14.82 and 13.80 μm bands, and the 14.82 and 10.63 μm bands exhibit the strongest correlations.

We perform a predictivity comparison among the models trained with different combinations of descriptors. These combinations are sorted by their predictivity from high (top) to low (bottom) in Figure 7. It is seen that the use of the combination of different types of descriptors (blue bars) in general improves the model predictivity with respect to the model using a single type of descriptor (gray bars). The exception of ECFP in this trend confirms the completeness of its molecular information, and underscores its importance in ML prediction of IR spectra. Besides ECFP, the combination of all descriptors (orange bars) does not lead to the best performance. Two combinations, NHAC+AE and NHAC+DMEs, show good predictivity in $\nu_{\text{low}}$ and $\nu_{\text{high}}$, respectively. This indicates that the combination of two descriptors with information of complementary types could be a reliable route toward optimized predictivity. Nevertheless, a combination with AE or with DMEs is not so interesting in the goal of improving the resolution of IR spectroscopy, since these descriptors would require time-consuming calculations (e.g., via DFT). In contrast, ECFP, NHAC, and CMSI depend only on the topology of a molecule. In particular, the combination of ECFP and NHAC can be expected to be the best candidate for ML prediction of molecular IR spectra.

The methods of the plurality voting (Lin et al. 2003) and the bootstrapping data set (Kohavi et al. 1995) are also used in order to further improve the model predictivity. The former method is a voting regressor that combines every model built with a different descriptor; it performs individual prediction with the most votes to form a final prediction. The latter method is an ensemble meta-estimator that fits base regressors each on random subsets of the original data set; it summarizes the individual predictions in a voting fashion to form a final prediction. Figure 7 shows that these two methods lead to
further improved predictivity, as indicated by the red and green bars. Moreover, we compute the degrees of mutual information between the different descriptor families and those between the features within the families, using the maximal information coefficient algorithm (Reshef et al. 2011). It is found that ECFP and NHAC are loosely correlated with the other descriptors, as shown in doi:10.5281/zenodo.5513073. The features of ECFP and NHAC exhibit relatively weak cross-correlation within the family, indicating less repeated information. This could be part of the reason why ECFP outperforms other descriptors, and why the descriptor pairs with NHAC exhibit improved predictivity.

4. Conclusions

Different ML models are trained to predict IR spectra of PAHs using several different types of representation of the compounds. We demonstrate that ML can help interpret the correlation between IR emission features. The astronomically observed correlations between the 3.3 and 11.2 μm bands, and between the 6.2, 7.7, and 8.6 μm ones are used as examples for demonstration. The RF feature importance is computed to find the common responsible molecular fragments, from which the correlation can be quantified by measuring the cosine similarity of their feature importance arrays (Equation (1)). It is found that the 3.3 μm band is singularly dominated by the emission feature of individual C–H bonds, while the 11.2 μm one depends much more on the local atomic environment (i.e., the number of adjacent CH groups on the peripheral aromatic rings). It is also shown that the 6.2 μm band is more strongly correlated with the 7.7 μm one than the 8.6 μm band is. The correlation map of the OOP spectral region on the ground of NHAC suggests that the 12.64, 12.91, 13.20, 11.02, and 11.22 μm bands have weak correlations, and that the 14.47 and 10.44 μm bands, the 14.82 and 13.80 μm bands, the 14.82 and 10.63 μm bands exhibit strong correlations.

We compare the performance of the descriptors, the choice of which is crucial for the model predictivity. ECFP has the best performance among the individual descriptors in both the low and high-frequency regions, underscoring the completeness of its contained molecular information. The second place of DMEs in υ_high indicates a strong correlation between the atomic distance and the high-frequency spectra. NHAC outperforms other descriptors (except for ECFP) for the bands in the 10–15 μm region. Moreover, besides ECFP, the combination of two descriptors with information of complementary types leads to improved predictivity. The voting regressor and the bootstrapping data set are shown to be reliable methods to further improve the model predictivity. Our results underscore the potential of ECFP and NHAC for accurately analyzing astronomical IR spectra, since they encode the topology of the molecule without reference to the exact coordinates of the atoms or the energy of molecule that need to be obtained by specific quantum chemical calculations.

Partial financial support from the National Natural Science Foundation of China (11964002, 12133003), the Guangxi Science Foundation (2020GXNSFAA159119, 2018GXNSFAA138179), the Special Funding for Guangxi Distinguished Professors
(2017AD22006), and the Scientific Research Foundation of Guangxi University (XTZ160532) are acknowledged. J. Carrete and G. Madsen are acknowledged for fruitful discussions.

**ORCID iDs**

Zhisen Meng [https://orcid.org/0000-0002-0761-7658](https://orcid.org/0000-0002-0761-7658)
Xiaosi Zhu [https://orcid.org/0000-0002-5110-0701](https://orcid.org/0000-0002-5110-0701)
Enwei Liang [https://orcid.org/0000-0002-7044-733X](https://orcid.org/0000-0002-7044-733X)
Zhao Wang [https://orcid.org/0000-0003-1887-223X](https://orcid.org/0000-0003-1887-223X)

**References**

Acke, B., Bouwman, J., Juhász, A., et al. 2010, *ApJ*, 718, 558
Allamandola, L. J., Tielens, A. G. G. M., & Barker, J. R. 1989, *ApJS*, 71, 733
Bauschlicher, C. W., Jr., Ricca, A., Boersma, C., & Allamandola, L. 2018, *ApJS*, 234, 32
Bellamy, L. J. 1958, The Infrared Spectra of Complex Organic Molecules (New York: Wiley)
Bertz, S. H. 1981, *JChS*, 103, 3599
Boersma, C., Bauschlicher, C. Jr., Ricca, A., et al. 2014, *ApJS*, 211, 8
Breiman, L. 2001, *Mach. Learn.*, 45, 5
Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. 2018, *Natur*, 559, 547
Calvo, F., Simon, A., Parneix, P., Falvo, C., & Dubosq, C. 2021, *JPCA*, 125, 5509
Draine, B. T. 2003, *ARA&A*, 41, 241
Fu, W., & Hopkins, W. S. 2018, *JPCA*, 122, 167
Galliano, F., Madden, S. C., Tielens, A. G., Peeters, E., & Jones, A. P. 2008, *ApJ*, 679, 310
Gasteiger, M., Behler, J., & Marquetand, P. 2017, *ChSci*, 8, 6924
Ghosh, K., Stuke, A., Todorović, M., et al. 2019, *AdSci*, 6, 1801367
Hanine, M., Meng, Z., Lu, S., et al. 2020, *ApJ*, 900, 188
Herbst, E., & van Dishoeck, E. F. 2009, *ARA&A*, 47, 427
Ho, T. K. 2002, *PAA*, 5, 102
Hony, S., Van Kerckhoven, C., Peeters, E., et al. 2001, *A&A*, 370, 1030
Hudgins, D. M., & Allamandola, L. 1999, *ApJL*, 516, L41
Ivanciuc, O. 2000, *JChICS*, 40, 1412
Joung, J. F., Han, M., Hwang, J., et al. 2021, *JACS*, 1, 427
Keller, L. D., Sloan, G., Forrest, W., et al. 2008, *ApJ*, 684, 411
Knuth, K. H. 2006, arXiv:physics/0605197
Kohavi, R. 1995, *IJCAI*, 14, 1137
Kovács, P., Zhu, X., Carrete, J., Madsen, G. K., & Wang, Z. 2020, *ApJ*, 902, 100
Laurens, G., Rabary, M., Lam, J., Peláez, D., & Allouche, A.-R. 2021, *TChAc*, 140, 1
Li, A. 2020, *NatAs*, 4, 339
Lin, X., Yacoub, S., Burns, J., & Simske, S. 2003, *PaRel.*, 24, 1959
Márquez-Neila, P., Fisher, C., Sznitzman, R., & Heng, K. 2018, *NatAs*, 2, 719
Mattioda, A., Hudgins, D., Boersma, C., et al. 2020, *ApJS*, 251, 22
McGill, C., Forsuelo, M., Guan, Y., & Green, W. H. 2021, *JChIM*, 61, 2594
Monge, G. 1781, *HARSB*, 666, 1781
Mooorthy, B., Chu, C., & Carlin, D. J. 2015, *ToxSci*, 145, 5
Pedregosa, F., Varoquaux, G., Gramfort, A., et al. 2011, *J. Mach. Learn. Res.*, 12, 2825
Peeters, E. 2011, IAU Symp. 280, The Molecular Universe, 149
Peeters, E., Tielens, A. G., Allamandola, L. J., & Wolfire, M. G. 2012, *ApJ*, 747, 44
Qi, H., Picaud, S., Devel, M., Liang, E., & Wang, Z. 2018, *ApJ*, 867, 133
Ravindra, K., Sokhi, R., & van Grieken, R. 2008, *AtmEs*, 42, 2895
RDKit Open-source cheminformatics, [http://www.rdkit.org](http://www.rdkit.org)
Reshef, D. N., Reshef, Y. A., Finucane, H. K., et al. 2011, *Sci*, 334, 1518
Rogers, D., & Mathew, H. 2010, *JChIM*, 50, 742
Rupp, M., Tkatchenko, A., Müller, K.-R., & Von Lilienfeld, O. A. 2012, *PRL*, 108, 058301
Schirer, J. 2020, *JChIM*, 60, 3804
Singhal, A. 2001, IDEB, 24, 35
Smith, J., Draine, B., Dale, D., et al. 2007, *ApJ*, 656, 770
Tielens, A. G. G. M. 2008, *ARA&A*, 46, 289
Vermeij, R., Peeters, E., Tielens, A., & Van der Hulst, J. 2002, *A&A*, 382, 1042
Ye, S., Zhong, K., Zhang, J., et al. 2020, *JChS*, 142, 19071
Zapata Trujillo, J. C., Syme, A.-M., Rowell, K. N., et al. 2021, *FrASS*, 8, 43
Zhang, L., Cao, Y., Colella, N. S., et al. 2015, *AcChR*, 48, 500