A&A, 710, L17 (2026)
https://doi.org/10.1051/0004-6361/202659999

Astronomy & Astrophysics

LETTER TO THE EDITOR

Full-spectrum infrared fingerprinting: A transformative AI paradigm for interstellar polycyclic aromatic hydrocarbons

Zhao Wang*

Laboratory for Relativistic Astrophysics, Department of Physics, Guangxi University, 530004 Nanning, China

Received 23 March 2026 / Accepted 25 May 2026

ABSTRACT

Context. In the era of high-sensitivity infrared (IR) astronomy, traditional manual diagnostics are no longer sufficient to harvest the complex physical insights hidden within interstellar spectra.
Aims. We introduce a machine learning paradigm that bypasses the limitations of empirical band ratios by treating the complete IR spectrum of polycyclic aromatic hydrocarbons (PAHs) as a high-dimensional fingerprint.
Methods. Using a random forest classifier trained on ∼23 000 spectra, we achieved a robust F_1 score of ∼0.963 across 12 size and charge categories, maintaining high performance on unseen molecular mixtures.
Results. Interrogating the model's decision-making process reveals that PAH size diagnostics are charge-dependent. Neutral PAHs are traced by C−H modes, while ionized species rely on 6−8 μm C−C morphology; however, the 12.5 μm feature remains a versatile tracer across multiple charge states.
Conclusions. This AI-driven paradigm offers a new route to interpret IR signatures and probe the chemical complexity of the interstellar medium.

Key words. ISM: molecules – infrared: ISM

1. Introduction

Interstellar polycyclic aromatic hydrocarbons (PAHs) are primary subjects in astrophysical exploration; they are studied via their infrared (IR) emission (Leger & Puget 1984; Allamandola et al. 1985). Their characteristic mid-IR (MIR) features (notably at 3.3, 6.2, 7.7, 8.6, 11.2, and 12.7 μm) are ubiquitous, appearing in diverse environments ranging from individual stellar sources to entire galaxies (Peeters et al. 2002; Smith et al. 2007). Since PAH sizes and ionization states are sensitive to the local ultraviolet field and electron density, these molecules serve as vital diagnostics for energetic regions such as photodissociation regions and active galactic nuclei (Tielens 2008; Li 2020).

Historically, inferring PAH properties relied on empirical band ratios calibrated against limited laboratory or theoretical datasets (Allamandola et al. 1989; Draine & Li 2007). These ratios are rooted in vibrational physics, and the $I_{11.2}/I_{3.3}$ ratio serves as a primary proxy for PAH size. This is because smaller grains reach higher peak temperatures during stochastic heating, preferentially exciting the 3.3 μm C−H stretch over the 11.2 μm mode (Draine & Li 2001). Similarly, ratios such as $I_{6.2}/I_{11.2}$ (or $I_{7.7}/I_{11.2}$) and $I_{11.2}/I_{12.7}$ are standard tools for tracing ionization states and molecular edge structures, respectively (Hony et al. 2001; Galliano et al. 2008; Boersma et al. 2018). Despite the evolution from the "blind" mathematical decomposition to a template-based fitting approach using the NASA Ames PAH IR spectroscopic database (PAHdb; Boersma et al. 2015; Mattioda et al. 2020), band-ratio analysis remains the conventional (if limited) standard in the *James Webb* Space Telescope (JWST) era (Maragkoudakis et al. 2022; Rigopoulou et al. 2024; Gregg et al. 2026).

However, Fig. 1 demonstrates a critical shortcoming of the ratio-based paradigm: its perceived accuracy is often a byproduct of sample selection, rather than physical universality (Maragkoudakis et al. 2020). For instance, the $I_{11.2}/I_{3.3}$ size trend performs well for a specific subset of 81 molecules (R^2 = 0.82), yet it collapses and becomes unreliable (R^2 = 0.23) when applied to a comprehensive library of 15 022 neutral PAHs. This discrepancy highlights a fundamental mismatch between classical diagnostic methods and modern observations. With JWST's unprecedented spectral resolution and sensitivity, reducing complex spectral profiles to a few discrete ratios is no longer a necessary simplification, but it comes at the cost of discarding valuable diagnostic information. To fully exploit the capabilities of current and future observatories, it is essential to transition from empirical ratios toward full-spectrum inference methods.

The advent of machine learning (ML) in astrophysics, coupled with expansive spectral libraries and JWST observations, presents a timely opportunity to move beyond traditional band-ratio diagnostics. In this work, we transition from discrete ratios to full-spectrum morphology, treating the complete IR spectrum as a high-dimensional fingerprint of PAH size and charge. Specifically, we trained an ML classifier on a large ensemble of spectra and evaluated its performance on observation-like mixtures of previously unseen species. This approach aims to provide more accurate and probabilistic constraints on molecular properties than conventional empirical methods. Instead of relying on degenerate band ratios, our method directly maps the full spectrum to radiation field hardness, PAH size, and charge, overcoming degeneracies, such as that seen between ionization and radiation fields (Rigopoulou et al. 2021).

* Corresponding author: zw@gxu.edu.cn

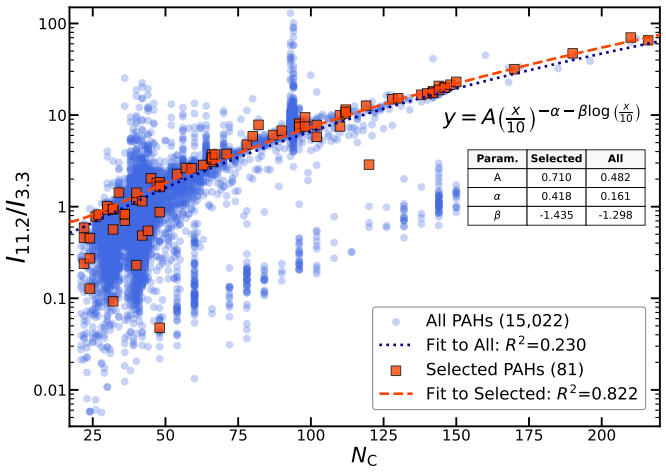

Fig. 1. Emission intensity ratio ($I_{11.2}/I_{3.3}$) vs. the number of carbon atoms (N_C). Circles represent the full dataset of 15 022 neutral PAHs. Squares denote the subset of 81 species selected by Maragkoudakis et al. (2020), characterized by $N_C > 20$, the presence of solo C–H bonds, and the absence of heteroatoms. All spectra were re-computed using a 6 eV cascade model. The lines depict the fit where R^2 values highlight the fitting quality.

2. Methodology

To streamline our analysis, we built a data-driven pipeline to infer PAH size and charged directly from their IR emission spectra. It integrates a large dataset of emission spectra with a ML classifier. While the model is trained on spectra of individual molecules, its performance is validated on complex, observation-like spectral mixtures drawn from an independent "unseen" molecular pool.

The dataset comprises 23 653 unique PAH structures compiled from PAHdb (10 404 theoretical spectra, $6 \leq N_C \leq 384$, charges –1, 0, +1, +2; Ricca et al. 2026) and first-principles DFT calculations (13 986 spectra, $8 \leq N_C \leq 160$, charges –1, 0, +1; He et al. 2026; Meng et al. 2023). After de-duplication, the raw dataset spans 23 653 PAH structures with harmonic IR spectra from 6.95 to 3751 cm^{-1}, including 15 911 neutrals (15 022 with $N_C > 20$), 2047 anions, 2972 cations, and 2723 dications.

The molecules were categorized into a 12-class framework based on size and charge state. For size: small ($N_C < 50$), medium ($50 \leq N_C \leq 99$), and large ($N_C \geq 100$). For charge state: anion (–1), neutral (0), cation (+1), and dication (+2).

These spectra represent ground-state absorption. To simulate astrophysically relevant conditions, we converted them to emission spectra using the thermal-cascade approximation within the AmesPAHdbIDLSuite tool (Boersma et al. 2013; Boersma & Bauschlicher 2014), assuming a representative excitation energy of 6 eV.

To maintain consistency with traditional theory, we focused on the 2.76–20 μm window (500–3620 cm^{-1}), which encompasses the characteristic PAH IR bands. Each discrete line spectrum was converted into a fixed-length feature vector by binning onto a common histogram grid with a fixed bin width and normalizing to unit area, so the model learns spectral shape rather than absolute intensity. Following sensitivity testing (see Appendix A), we used a bin width of 20 cm^{-1} and excluded any features (bins) containing contributions from fewer than ten molecules. While this bin width optimizes the model's signal-to-noise ratio (S/N) and generalization potential, we acknowledge that this resolution may smooth out fine-grained spectral substructures potentially resolvable by JWST. This choice prioritizes robust morphological patterns over noise-sensitive narrow-band variations.

Table 1. Performance metrics of PAH mixture classification across 12 size and charge categories.

Category	Training size	Precision	Recall	F_1 score
Small	13 626	0.985	0.983	0.985
Medium	4638	0.910	0.938	0.923
Large	663	0.975	0.992	0.983
−1	1639	0.970	0.983	0.977
0	12730	0.893	0.973	0.930
+1	2379	0.977	0.980	0.977
+2	2179	0.987	0.947	0.963
Average	–	0.957	0.971	0.963

Notes. Results are grouped by molecular size ($N_C < 50$, 50–99, and ≥100) and charge state (−1, 0, +1, and +2).

To ensure the model's astrophysical relevance, we partitioned the data by randomly selecting 20% of the molecules from each of the twelve classes to form an "unseen" pool, while the remaining 80% constituted the training set. We simulated realistic astronomical observations by constructing synthetic mixtures, averaging the spectra of molecules drawn at random from the unseen pool. This uniform weighting ensures the model remains independent of specific astrophysical priors. These mixtures were generated across varying population sizes, defined as $N_{mol} \in \{1, 5, 10, 20, 50, 100, 200\}$. For each N_{mol} value, we produced 100 mixed spectra, resulting in 700 synthetic spectra per class. All mixtures derived from this unseen subset were reserved strictly for final model evaluation to ensure unbiased performance metrics.

We adopted a random forest (RF) classifier, an ensemble of decision trees that combines numerous threshold-based rules to produce robust predictions (Breiman 2001). The model was trained on normalized spectral features using 500 trees with a maximum depth of 25, utilizing out-of-bag scoring for internal validation. To mitigate class imbalance, we implemented the synthetic minority oversampling technique (SMOTE) in conjunction with adjusted class weighting. The trained classifier was evaluated on synthetic mixtures derived from the "unseen" molecular pool. Performance was quantified using standard metrics: precision, recall, and F_1 score. To ensure reproducibility, the source code, datasets, and the trained model are openly accessible on Git repository: AstroPAH-MLDiag. Comprehensive hyperparameter configurations can be found in this code.

3. Results and discussion

The trained RF model shows high-fidelity performance on the mixed PAH spectra, achieving a macro-averaged F_1 score of approximately 0.963 (Table 1). This robust classification indicates that full-profile spectral features contain sufficient information to simultaneously constrain molecular size and ionization state, even within complex molecular mixtures.

The performance scales nonlinearly with molecular size. Overall, small PAHs ($N_C < 50$) achieve the highest F_1 score (0.985), supported by a large training set of 13 626 spectra. Large PAHs ($N_C \geq 100$) perform nearly as well ($F_1 = 0.983$) despite having only 663 training samples, which is the smallest among all size classes. This suggests that spectral features of large PAHs are sufficiently distinctive that the model requires relatively

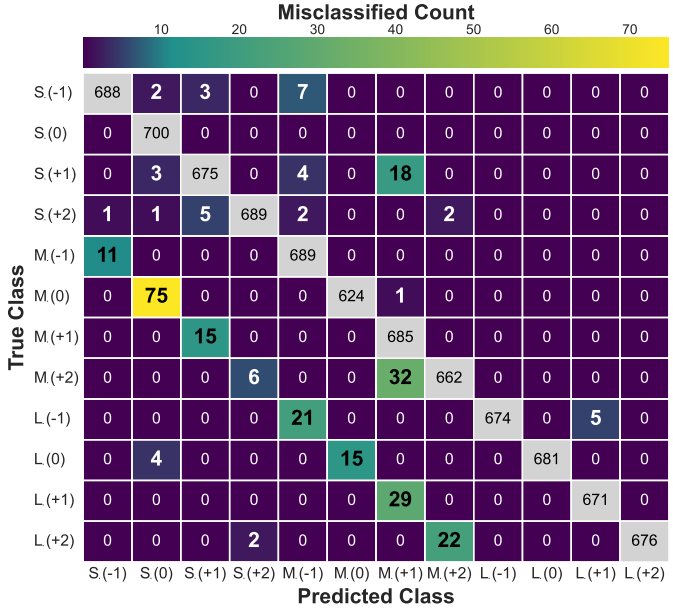

Fig. 2. Confusion matrix for the 12-class PAH mixture classification. Diagonal elements indicate correct classifications, while off-diagonal values reveal specific misclassification patterns among the 700 samples per class. Class indices are categorized by size: S (small), M (medium), and L (large), followed by the charge state in parentheses.

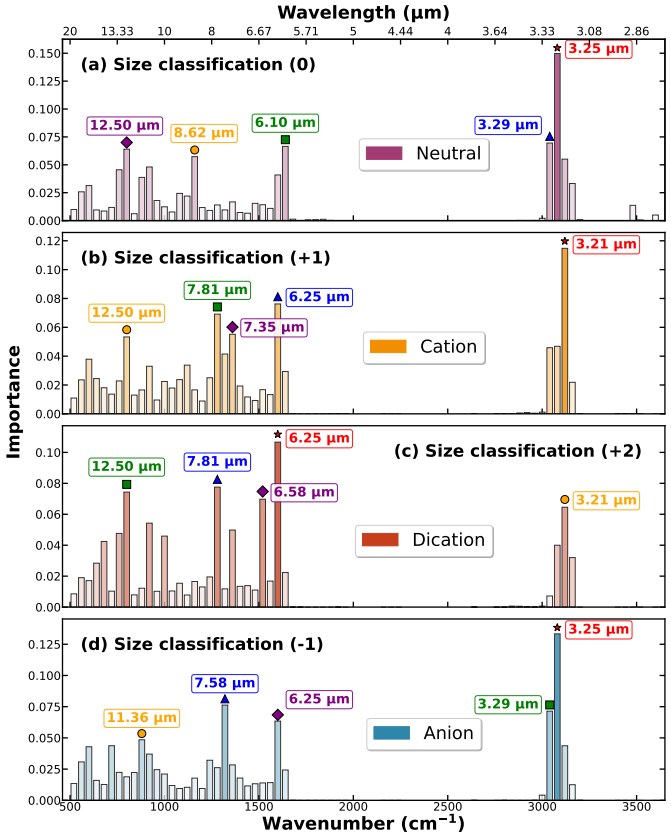

Fig. 3. Feature importance for size classification across different charge states: (a) neutral (0), (b) cation (+1), (c) dication (+2), and (d) anion (−1). Highlighted markers indicate the top five influential features.

few examples to generalize accurately (Bauschlicher et al. 2008; Ricca et al. 2012). In contrast, medium-sized PAHs ($50 \leq N_C \leq 99$) show a clear performance drop ($F_1 = 0.923$) even though they are reasonably well represented with 4638 training samples.

Charge classification performance is consistently high, with both cations (+1) and anions (−1) achieving F_1 scores of 0.977, and dications (+2) following at 0.963. The neutral class (0) exhibits the highest recall (0.973) but the lowest precision (0.893), indicating that while the model successfully identifies nearly all neutral species, it misassigns ionized or differently sized molecules to the neutral category.

To understand these performance dips, we can take a look at the confusion matrix (Fig. 2), which reveals specific interclass leakage. Within the medium-sized PAHs (M), misclassifications are dominated by neutral species [M (0)], which are frequently mistaken for small neutral [S (0)] PAHs, rather than ionized species. This confusion suggests that the spectral signatures of medium-sized neutrals begin to converge with those of smaller counterparts. As molecular size increases, rising vibrational mode density and spectral broadening are likely to diminish the distinctiveness between size classes (Bauschlicher et al. 2009; Knight et al. 2021; Draine & Li 2007). Given the high sampling rate for the medium category, this ambiguity appears to be a physical limitation of the spectral features rather than a lack of training data.

This trend extends across all charge states. Systematic misclassifications in Fig. 2 appear along the lines parallel to the diagonal, offset by exactly four classes, the interval corresponding to the same charge state in an adjacent size family. This pattern confirms that while the model readily distinguishes charge states, molecular size remains the more challenging attribute to constrain (Draine et al. 2021). Furthermore, charge-state residuals are confined to adjacent ionization levels; for example, 32 (out of 700) instances occurred where medium-sized dications [M (+2)] were misidentified as cations [M (+1)].

To evaluate the influence of the training set size range on our results, we performed two sensitivity analyses. First, acknowledging that small PAHs ($N_C < 20$) are susceptible to photodissociation in the diffuse interstellar medium (Li & Draine 2001), we retrained the model with a minimum threshold of $N_C = 20$; the resulting mean F_1 score remained stable at 0.947, confirming that the inclusion of very small species does not bias performance for more resilient populations. Second, while computational costs currently cap our training data at $N_C = 384$, we tested the model's so-called out-of-distribution generalizability by training on subsets with $N_{C,max} \leq 200$. The consistently high performance ($F_1 > 0.95$) indicates that the model can effectively learn fundamental spectral motifs that remain characteristic in larger molecular structures ($N_C \sim 1000$), as modeled in traditional frameworks (Draine & Li 2007).

The inherent interpretability of the RF model allows for an examination of its "reasoning" through a feature importance analysis. Using the Gini importance metric (Breiman 2001), we identified the spectral regions most diagnostic of PAH properties. Figure 3 shows that the size information is distributed across multiple spectral regions in a charge-dependent manner. For neutral PAHs (panel a), the model relies heavily on the 3.25 and 3.29 μm bins, consistent with the classical proxy (Lemmens et al. 2023). The analysis also identifies the 6.1 (C−C stretching), 8.6 (C−H in-plane bending), and 12.5 μm (C−H out-of-plane (OOP) bending) bins as critical features for sizing the neutrals.

For cations (panel b), and especially dications (panel c), size diagnostic importance shifts toward the 6−8 μm C−C stretching modes at 6.25 and 7.81 μm. The ML model extracts

size information within these complexes, likely tracing charge-specific modifications or edge geometries, but still supports caution against using the 6.2/7.7 ratio alone for size determination (Peeters et al. 2002; Bauschlicher et al. 2008). The analysis further reveals a significant departure from canonical diagnostics for anionic PAHs (panel d), where size is primarily traced by the relative strengths of the 3.3 μm complex and the 7.58 μm C−C stretch (Allamandola et al. 1989; Schutte et al. 1993). However, the high diagnostic importance of the 7.58 μm C−C stretch in anions should be interpreted with caution. Given the relatively small training pool for large anions (only 145 samples), the model's reliance on this specific bin might partially reflect the limited structural diversity within this subset, rather than a universal physical law.

Figure 3 suggests the heightened importance of the 12.5 μm bin across multiple charge states (0, +1, and +2), identifying it as a robust complementary size tracer. This importance stems from the C−H OOP bending modes of "duo" and "trio" hydrogen atoms, which are spatially stable across varying radiation fields (Shannon et al. 2016). The intensity of the 12.5−12.7 μm complex relative to the 11.2 μm solo-H mode tracks the evolution of edge structure as PAHs grow in size (Shannon et al. 2015; Maragkoudakis et al. 2023). This aligns with longer-wavelength signatures proposed as alternatives for observations lacking 3.3 μm coverage (Draine et al. 2021; Berné et al. 2022).

We further assessed the model's applicability by testing it against JWST MIRI observations of NGC 7027 (see Appendix B). The model identified small-to-medium neutral PAHs as the dominant species, aligning with the observed spectral profile. However, the moderate confidence levels that we obtained in this work point to an important limitation: astronomical observations represent an integrated line-of-sight view of mixed PAH populations. Although our current model is designed as a classifier that identifies a single dominant species or state, real spectra are inherently mixtures. This mismatch naturally reduces the confidence of any single-label prediction. Additionally, the model remains limited by its training on discrete energy levels rather than on continuous interstellar radiation fields (Li et al. 2024a,b). As a preliminary test, this work also shows that future high-confidence diagnostics will require integrating multi-instrument data (e.g., NIRSpec and MIRI) to capture the full range of diagnostic vibrational modes. Moreover, we evaluated the impact of excitation conditions on classification performance by applying our workflow to emission spectra at 3 and 9 eV. The classification accuracy remained consistent (see Appendix C), demonstrating a robustness across diverse astrophysical regimes.

4. Conclusions

We present a ML framework designed to infer the size and charge of interstellar PAHs directly from full-spectrum IR morphology, achieving a macro F_1 score of 0.963 across 12 categories. Moving beyond "black-box" AI, our feature-importance analysis reveals that PAH size diagnostics are fundamentally charge-dependent. Specifically, while traditional empirical proxies such as $I_{11.2}/I_{3.3}$ show diminished reliability across large, diverse datasets, our model identifies the 3.21–3.29 and 11–14 μm spectral regions as the most informative feature clusters for size inference. Notably, the 12.5 μm feature emerges as a physically grounded, versatile tracer that remains robust across multiple charge states, effectively breaking the degeneracies that limit single-band-ratio diagnostics.

The model currently relies on synthetic mixtures as proxies for unknown astronomical ground truths. Further development of this work will prioritize addressing class imbalances, particularly the under-representation of large molecules, which currently limits the statistical confidence in their identified diagnostic footprints (Kovács et al. 2020; Mai et al. 2025; Tang et al. 2026). From a broader perspective, applications to JWST observations of NGC 7027 show that future efforts will require regression-based mixed-population modeling, continuous excitation frameworks, and multi-instrument data.

Data availability

Source code and datasets are available on Git repository: AstroPAH-MLDiag.

Acknowledgements. The authors acknowledge financial support from: National Natural Science Foundation of China (Grant No. 12463005), Guangxi Natural Science Foundation under (Grant No. 2026GXNSFHA00640301), and Guangxi Talent Programme (Highland of Innovation Talents).

References

Allamandola, L. J., Tielens, A. G. G. M., & Barker, J. R. 1985, ApJ, 290, L25
Allamandola, L. J., Tielens, A. G. G. M., & Barker, J. R. 1989, ApJS, 71, 733
Bauschlicher, C. W., Jr., Peeters, E., & Allamandola, L. J. 2008, ApJ, 678, 316
Bauschlicher, C. W., Jr., Peeters, E., & Allamandola, L. J. 2009, ApJ, 697, 311
Berné, O., Habart, É., Peeters, E., et al. 2022, PASP, 134, 054301
Boersma, C., & Bauschlicher, C. W. Jr. 2014, ApJS, 211, 8
Boersma, C., Bregman, J., & Allamandola, L. J. 2013, ApJ, 769, 117
Boersma, C., Bregman, J., & Allamandola, L. J. 2015, ApJ, 806, 121
Boersma, C., Bregman, J., & Allamandola, L. J. 2018, ApJ, 858, 67
Breiman, L. 2001, Mach. Learn., 45, 5
Donnan, F. R., García-Bernete, I., Rigopoulou, D., et al. 2024, MNRAS, 529, 1386
Draine, B. T., & Li, A. 2001, ApJ, 551, 807
Draine, B. T., & Li, A. 2007, ApJ, 657, 810
Draine, B. T., Li, A., Hensley, B. S., et al. 2021, ApJ, 917, 3
Galliano, F., Madden, S. C., Tielens, A. G. G. M., Peeters, E., & Jones, A. P. 2008, ApJ, 679, 310
Gregg, B., Calzetti, D., Adamo, A., et al. 2026, ApJ, 997, 20
He, J., Mai, X., & Wang, Z. 2026, A&A, 708, A335
Hony, S., van Kerckhoven, C., Peeters, E., et al. 2001, A&A, 370, 1030
Knight, C., Peeters, E., Stock, D. J., & Tielens, A. G. G. M. 2021, ApJ, 918, 8
Kovács, P., Zhu, X., Carrete, J., Madsen, G., & Wang, Z. 2020, ApJ, 902, 100
Leger, A., & Puget, J. L. 1984, A&A, 137, L5
Lemmens, A. K., Mackie, C. J., Candian, A., et al. 2023, Faraday Discuss., 245, 380
Li, A. 2020, Nat. Astron., 4, 339
Li, A., & Draine, B. T. 2001, ApJ, 554, 778
Li, K. J., Li, A., & Gang, Z. 2024a, ApJ, 961, 107
Li, K. J., Li, A., & Gang, Z. 2024b, MNRAS, 529, 4425
Mai, X., Wang, Z., Pan, L., et al. 2025, MNRAS, 541, 3073
Maragkoudakis, A., Peeters, E., & Ricca, A. 2020, MNRAS, 494, 642
Maragkoudakis, A., Boersma, C., Temi, P., Bregman, J. D., & Allamandola, L. J. 2022, ApJ, 931, 38
Maragkoudakis, A., Peeters, E., Ricca, A., & Boersma, C. 2023, MNRAS, 524, 3429
Mattioda, A. L., Hudgins, D. M., Boersma, C., et al. 2020, ApJS, 251, 22
Meng, Z., Zhang, Y., Liang, E., & Wang, Z. 2023, MNRAS, 525, L29
Peeters, E., Hony, S., van Kerckhoven, C., et al. 2002, A&A, 390, 1089
Ricca, A., Bauschlicher, C. W., Jr, Boersma, C., Tielens, A. G. G. M., & Allamandola, L. J. 2012, ApJ, 754, 75
Ricca, A., Boersma, C., Maragkoudakis, A., et al. 2026, ApJS, 282, 7
Rigopoulou, D., Barale, M., Clary, D. C., et al. 2021, MNRAS, 504, 5287
Rigopoulou, D., Donnan, F. R., García-Bernete, I., et al. 2024, MNRAS, 532, 1598
Schutte, W. A., Tielens, A. G. G. M., & Allamandola, L. J. 1993, ApJ, 415, 397
Shannon, M. J., Stock, D. J., & Peeters, E. 2015, ApJ, 811, 153
Shannon, M. J., Stock, D. J., & Peeters, E. 2016, ApJ, 824, 111
Smith, J. D. T., Draine, B. T., Dale, D. A., et al. 2007, ApJ, 656, 770
Tang, G., He, J., Wang, Z., Qiu, D., et al. 2026, MNRAS, 546, stag283
Tielens, A. G. G. M. 2008, ARA&A, 46, 289

Appendix A: Sensitivity test for varying bin widths

The spectral preprocessing requires balancing resolution against model robustness. We tested bin widths of 8.4 (determined by the Knuth Bayesian rule), 12, 20, 30, and 40 cm^{-1}, with results summarized in Table A.1.

Table A.1. Comparison of model performance (average F_1) across different bin widths (6 eV dataset).

Bin Width (cm^{-1})	Macro Average F_1
8.4	0.90
12	0.93
20	0.96
30	0.90
40	0.92

Although finer binning (e.g., 8.4 or 12 cm^{-1}) provides higher nominal resolution, the resulting high-dimensional feature space (more bins) makes the classifier sensitive to non-diagnostic spectral noise and minor vibrational shifts.

Appendix B: Test on JWST MIRI observations

We selected ten high signal-to-noise ratio (S/N > 30.0) spectra from the planetary nebula NGC 7027 (JWST Program 1523, PI: D. R. Law), as shown in the inset of Fig. B.1. To isolate the PAH emission, we applied the differential extinction method of Donnan et al. (2024), which extracts intrinsic spectral features. The resulting spectra are presented in Fig. B.1.

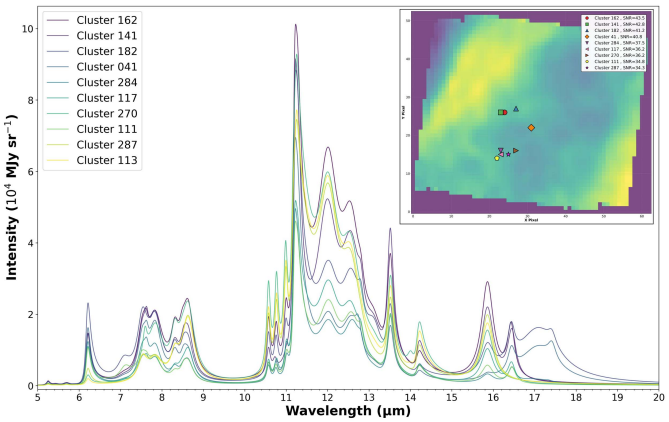

Fig. B.1. JWST spectra of ten distinct PAH clusters in NGC 7027 over the 5–20 μm wavelength range. Inset: spatial distribution of the 10 PAH spectra in NGC 7027, where the colors of the background correspond to the mean flux.

We trained a separated RF classifier to predict the excitation energy across three classes: 3, 6, or 9 eV. When evaluated on synthetic mixed spectra, the model achieved F_1 scores of about 0.989, 0.947, and 0.955 for the 3, 6, and 9 eV classes, respectively. Using this classifier, we estimated the excitation environment for the observed spectra to be approximately 3.0 eV.

We then applied our 12-class ML framework (re-trained with 3-eV dataset) to determine PAH size and charge states. The model identifies the carriers as small-to-medium neutral PAHs, which is qualitatively consistent with the observed spectral profiles (e.g., the prominent 11.3 μm feature relative to the 6.2, 7.7, and 8.6 μm bands, relatively well-resolved discrete features in 11–15 μm region, and weak feature at 17 μm). However, we noted that the classification confidence remains low. For the ten selected PAH clusters in NGC 7027, the model predicts eight clusters (IDs: 111, 141, 117, 113, 270, 284, 162, and 287) as small neutral PAHs (Small (0)), with confidence scores of 0.696, 0.654, 0.650, 0.598, 0.574, 0.572, 0.568, and 0.564, respectively; along with two clusters (IDs: 041 and 182) as medium neutral PAHs (Medium (0)), with confidence scores of about 0.578.

Appendix C: Impact of excitation energy

We tested model performance across three excitation energies (3, 6, and 9 eV) by simulating PAH emission spectra from absorption data via a thermal cascade model (Fig. C.1).

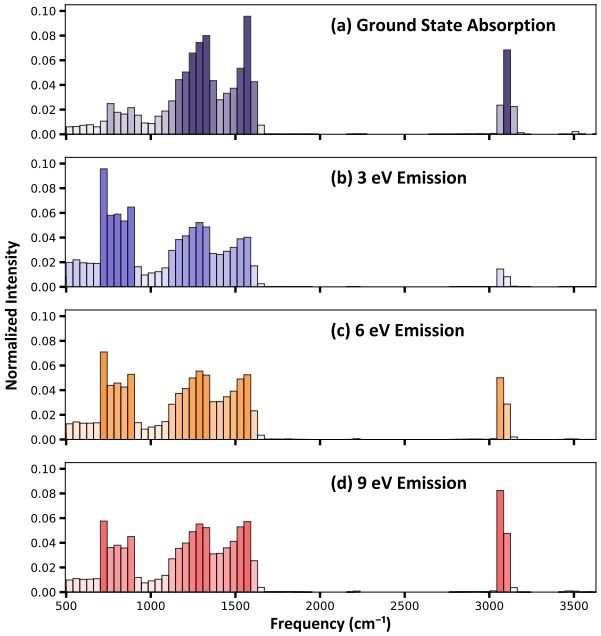

Fig. C.1. Normalized summed emission spectra of all 23 653 PAHs at excitation energies of 3 (b), 6 (c), and 9 eV (d), shown alongside their ground-state absorption spectrum (a). Each spectrum was normalized with the total area scaled to unity.

As shown in Table C.1, classification performance peaks at 6 eV. Performance remains high but shows a slight decrease at both 3 eV and 9 eV which yield Macro F_1 scores of 0.95 and 0.93.

Table C.1. Model performance metrics across different emission excitation energies.

Spectral Type	Precision	Recall	F_1 score
3 eV	0.95	0.95	0.95
6 eV	0.96	0.97	0.96
9 eV	0.94	0.93	0.93