Monthly Notices of the ROYAL ASTRONOMICAL SOCIETY
MNRAS **510,** 3546–3560 (2022) https://doi.org/10.1093/mnras/stab3651
Advance Access publication 2021 December 16

Families and clusters of diffuse interstellar bands: a data-driven correlation analysis

Haoyu Fan [1,2]★ Madison Schwartz,[1,2] Amin Farhang,[3,4] Nick L. J. Cox,[5] Pascale Ehrenfreund,[6] Ana Monreal-Ibero [6] Bernard H. Foing,[6] Farid Salama [7] Klay Kulik,[1,2] Heather MacIsaac,[1,2] Jacco Th. van Loon[8] and Jan Cami[1,2,9]

[1] *Department of Physics and Astronomy, The University of Western Ontario, London, ON N6A 3K7, Canada*
[2] *The Institute for Earth and Space Exploration, The University of Western Ontario, London, ON N6A 3K7, Canada*
[3] *Department of Physics, University of Tehran, North Karegar Ave, 14395-547 Tehran, Iran*
[4] *School of Astronomy, Institute for Research in Fundamental Sciences, 19395-5531 Tehran, Iran*
[5] *ACRI-ST, 260 Route du Pin Montard, F-06904 Sophia-Antipolis, France*
[6] *Leiden Observatory, Leiden University, Niels Bohrweg 2, NL-2333 CA Leiden, the Netherlands*
[7] *NASA Ames Research Centre, Space Science & Astrobiology Division, Moffett Field, CA 94035, USA*
[8] *Lennard-Jones Laboratories, Keele University, Keele, Newcastle ST5 5BG, UK*
[9] *SETI Institute, 189 Bernardo Ave, Suite 100, Mountain View, CA 94043, USA*

Accepted 2021 December 4. Received 2021 December 1; in original form 2021 September 24

ABSTRACT

More than 500 diffuse interstellar bands (DIBs) have been observed in astronomical spectra, and their signatures and correlations in different environments have been studied over the past decades to reveal clues about the nature of the carriers. We compare the equivalent widths of the DIBs, normalized to the amount of reddening, $E_{\text{B-V}}$, to search for anticorrelated DIB pairs using a data sample containing 54 DIBs measured in 25 sightlines. This data sample covers most of the strong and commonly detected DIBs in the optical region, and the sightlines probe a variety of interstellar medium conditions. We find that 12.9 per cent of the DIB pairs are anticorrelated, and the lowest Pearson correlation coefficient is $r_{\text{norm}} \sim -0.7$. We revisit correlation-based DIB families and are able to reproduce the assignments of such families for the well-studied DIBs by applying hierarchical agglomerative and k-means clustering algorithms. We visualize the dissimilarities between DIBs, represented by $1 - r_{\text{norm}}$, using multidimensional scaling (MDS). With this representation, we find that the DIBs form a rather continuous sequence, which implies that some properties of the DIB carriers are changing gradually following this sequence. We also find that at that least two factors are needed to properly explain the dissimilarities between DIBs. While the first factor may be interpreted as related to the ionization properties of the DIB carriers, a physical interpretation of the second factor is less clear and may be related to how DIB carriers interact with surrounding interstellar material.

Key words: dust, extinction – ISM: lines and bands – ISM: molecules.

1 INTRODUCTION

The diffuse interstellar bands (DIBs) are a set of absorption features that represent a century-long mystery regarding the interstellar medium (ISM). The first DIBs, $\lambda\lambda$5780 and 5797,[1] were noted towards ζ Per by Heger (1922), and now more than 500 such absorption features have been catalogued in the optical region (Hobbs et al. 2008, 2009; Fan et al. 2019), and tens more in the near-infrared (e.g. Joblin et al. 1990; Cox et al. 2014; Hamano et al. 2015, 2016). The substructures within several DIB profiles strongly suggest a molecular origin (e.g. Sarre et al. 1995; Ehrenfreund & Foing 1996; Kerr et al. 1998; Walker et al. 2001; Cami et al. 2004), yet their specific carriers remain unknown, maybe except the two near-infrared DIBs at $\lambda\lambda$9577 and 9633 and three weaker DIBs that have been assigned to C_{60}^{+} – a finding that is supported by an impressive array of observational and experimental studies (Foing & Ehrenfreund 1994, 1997; Campbell et al. 2015; Walker et al. 2015, 2016, 2017; Cordiner et al. 2017; Spieler et al. 2017; Lallement et al. 2018; Cordiner et al. 2019; see Linnartz et al. 2020 for a review). We note that this assignment was recently challenged by Galazutdinov et al. (2021), who reported that the two DIBs $\lambda\lambda$9577 and 9633 are poorly correlated – they report a Pearson correlation coefficient $r = 0.32$. A thorough review of this claim, however, finds the opposite, that the $\lambda\lambda$9577 and 9633 DIBs do in fact correlate very well ($r \sim 0.9$; Schlarmann et al. 2021), thus further supporting this identification.

To guide and support laboratory efforts to identify more DIB carriers, astronomers perform analyses of astronomical observations to provide constraints on the properties of DIB carriers and thus narrowing down the candidates to be examined. Such efforts often include correlation studies, where the DIB strengths, most often

★ E-mail: hfan58@uwo.ca

[1] We follow the convention that the DIBs are referred to by their approximate central wavelength expressed in Å.

represented by their equivalent widths [EWs, or denoted as W(DIBs)], are compared to ISM parameters such as $E_{\text{B-V}}$, column densities of various species (N(X)), or to other DIBs (e.g. Herbig 1993; Cami et al. 1997; Friedman et al. 2011; Vos et al. 2011; Fan et al. 2017). It has been recognized long ago that many DIBs respond to changing physical conditions (termed the 'DIB behaviour', typically observed by changes in the EW) but not necessarily in the same way (Snow & Cohen 1974; Adamson, Whittet & Duley 1991; Jenniskens, Ehrenfreund & Foing 1994; Cami et al. 1997; Vos et al. 2011; Sonnentrucker 2014). The classical example in this context is illustrated by comparing the $\lambda\lambda$5780 and 5797 DIBs in the sightlines towards ζ Oph (HD 149757; $E_{\text{B-V}} = 0.28$) and σ Sco (HD 147165, $E_{\text{B-V}} = 0.34$). The sightline of ζ Oph crosses a cloud interior and thus probes an environment that is shielded from UV irradiation. On the other hand, the sightline of σ Sco represents a more exposed environment. In both sightlines, the EW of the λ5797 DIB is similar ($\sim$35 mÅ), while W(5780) increases from $\sim$73 mÅ in ζ Oph to $\sim$240 mÅ in σ Sco. Clearly, the λ5797 DIB carrier is somewhat indifferent to the changing UV exposure, whereas the λ5780 is very sensitive to it. This could indicate different photo-chemical properties of the DIB carriers themselves but could also be due to more indirect effects where the UV radiation dissociates molecular hydrogen, which in turn then affects the chemical network of DIB carriers (Webster 1993).

The changing response to environment then also leads to the idea of DIB 'families'. Members of the same family show a similar response to environmental factors and thus also exhibit good mutual correlations, whereas members of different families show a much poorer correlation (Krelowski & Walker 1987; Westerlund & Krelowski 1989; Cami et al. 1997; Wszołek & Godłowski 2003). Strong correlations between two DIBs from the same family could also indicate the same or related carriers. While there is no general agreement about precisely which DIBs belong to a specific family, some connections are well established. For instance, the $\lambda\lambda$5797, 6379 and 6613 DIBs show similar behaviour, favouring less exposed (ζ-type) environments, and have been grouped into a family by various authors. In this paper, we will refer to this family of DIBs as the ζ-DIBs. The DIBs $\lambda\lambda$5780 and 6284 (along with several others) on the other hand thrive in exposed (σ-type) regions (Lan, Ménard & Zhu 2015; Ensor et al. 2017; Krełowski 2018; Omont & Bettinger 2020). They can also be considered as members of a family that we will call σ-DIBs for what follows. In this context, another group of the so-called C_2-DIBs can be seen as a third family, whose members trace dense and molecular regions of the ISM cloud (Thorburn et al. 2003).

While many studies have focused on correlations, anticorrelations between DIBs could also be of particular interest. Such DIB pairs could indicate that their carriers are the start and end products of the same physical or chemical processes. For example, if one DIB would be carried by a neutral species and another DIB by its cation, one would expect the strength of one DIB to decrease as the strength of the other increases. Similar arguments of course hold for other processes such as hydrogenation. A key issue that plagues correlation studies, however, is that different sightlines typically represent different amounts of interstellar material. Since more interstellar material in general also implies higher column densities for individual species, the EWs of DIBs thus always have some positive correlation with each other (e.g. Bailey et al. 2015, 2016). This effect tends to hide anticorrelations. But by comparing the normalized EWs of DIBs (W(DIB)/$E_{\text{B-V}}$), Cami et al. (1997) were among the first to identify several anticorrelated DIB pairs within a small sightline sample.

Motivated by more robustly confirming the existence of anticorrelated DIB pairs, we revisit the topic of DIB (anti-)correlations as well as DIB families. We include more DIBs than most previous analyses to obtain a more general picture of DIB correlations, rather than focusing on a handful of well-studied DIBs. This paper is organized as follows. Section 2 describes the data we use and the selection criteria of our target DIBs. We search for anticorrelated DIB pairs in Section 3 and sort the DIBs into groups in Section 4 according to their mutual correlations. Section 5 contains our efforts to visualize the DIB correlations, and the implications of our findings are discussed in Section 6. Finally, we summarize this work in Section 7.

2 DATA AND DIB SELECTION

We base our work on the DIB measurements from Fan et al. (2019), which is part of the DIB survey project carried out at the Apache Point Observatory and the University of Chicago. We refer the reader to the paper series 'Studies of the Diffuse Interstellar Bands' and related publications (i.e. McCall et al. 2001; Thorburn et al. 2003; Hobbs et al. 2008, 2009; McCall et al. 2010; Friedman et al. 2011; Dahlstrom et al. 2013; Welty et al. 2014; Fan et al. 2017, 2019) for more details on the data and measurement techniques than outlined here.

The DIB measurements are made with $R \sim 38\,000$ spectra towards 25 medium to highly reddened targets ($E_{\text{B-V}}$ between 0.31 and 3.31 mag). The spectral types of the background stars are between O6 and A5, and the sightlines cover a great variety of ISM conditions as characterized by their f_{H2} values. To identify the presence of stellar lines that may compromise the DIB measurements, each of these target stars is paired with a standard star with similar spectral type but very low reddening. Table A1 summarizes the information on the target stars and their corresponding standard star, and a full version of the same table can be found in Fan et al. (2019). Telluric lines are removed from the raw data by fitting a template spectrum based on airmass, and a telluric reference spectrum is displayed during the DIB-measuring process to provide guidance on possible residuals from the correction.

Direct integration is used to measure the EWs of DIBs without assuming any specific profile, and the uncertainties are estimated based on the signal-to-noise ratio and the width of the profile. Since the spectra data are normalized by the data reduction pipeline, only a local continuum is needed around the target DIB. Consistent measuring techniques, especially regarding the selections of continuum regions and integration limits, are kept for each DIB to all sightlines. Such effort ensures a uniformly made data set with great self-consistency. We flag defects such as contamination from adjacent stellar/telluric lines or large uncertainty in the continuum level and exclude such measurements in the analysis. We also include $E_{\text{B-V}}$ and column densities of some molecular species in our study. The sources of these data are described in Fan et al. (2017).

The selection of target DIBs in this work is based on the number of measurements available among the 25 sightlines, and we require each of the target DIBs to have no more than five excluded measurements or upper limits (non-detection). This is to ensure that many environments can be considered in the DIB correlations and that the resulting correlations are robust. However, we made exceptions for a few DIBs of particular interest, such as the broad DIB λ4429 and the strong DIB λ5780. In total, we include 54 DIBs in our correlation analysis as summarized in Table 1. These DIBs cover most of the strong and/or well-studied optical DIBs in the literature (e.g. Cami et al. 1997; Cox et al. 2006; Friedman et al. 2011; Vos et al. 2011; Kos & Zwitter 2013).

Table 1. Properties of the 54 target DIBs in this work.

Label Å	Wave. Å	FWHM Å	No. of Mea.	Avg. Nor. EW mÅ/mag	Comments	Index Å	Wavelength Å	FWHM Å	No. of Mea.	Avg. Nor. EW mÅ/mag	Comments
4429	4429.33	24.13	10	2005.95	Broad DIB[a]	6270[b]	6269.88	1.48	24	104.64	
4501	4501.51	2.53	17	73.84		6284	6284.05	4.49	24	958.09	
4726	4726.98	2.76	21	168.87	C_2 DIB	6324	6324.91	0.74	20	10.00	
4762	4762.44	1.94	20	48.81		6330	6330.03	0.73	21	10.11	
4963	4963.92	0.68	25	29.20	C_2 DIB	6353	6353.31	1.66	20	26.77	
4984	4984.78	0.51	18	16.11	C_2 DIB	6362	6362.26	1.61	20	18.76	
5418	5418.87	0.75	20	22.01	C_2 DIB	6367	6367.30	0.52	21	11.50	
5512	5512.68	0.54	20	14.85	C_2 DIB	6376	6376.14	0.76	23	32.59	
5545	5545.08	0.84	20	25.02		6377	6377.07	0.57	22	11.34	
5546	5546.46	0.68	21	12.38	C_2 DIB	6379	6379.25	0.64	23	79.12	
5705	5705.12	2.68	20	90.65		6397	6397.04	1.27	20	23.85	
5766	5766.16	0.76	22	16.31		6439	6439.51	0.82	23	18.31	
5780	5780.67	2.09	17	398.84		6445	6445.30	0.60	20	23.56	
5793	5793.24	0.96	20	14.32	C_2 DIB	6449	6449.27	0.94	21	19.77	
5797	5797.18	0.89	23	140.63		6520	6520.74	1.00	21	25.07	
5828	5828.50	0.78	20	10.96		6553	6553.88	0.51	21	11.04	
5849	5849.82	0.83	24	50.88		6613	6613.74	1.05	25	185.14	
5923	5923.51	0.74	21	16.34		6622	6622.84	0.58	21	9.67	
6065	6065.32	0.64	21	12.19		6660	6660.67	0.63	23	39.22	
6089	6089.85	0.58	22	18.74		6699	6699.28	0.67	22	24.66	
6108	6108.06	0.49	20	8.34		6702	6702.07	0.74	22	10.05	
6116	6116.80	0.87	20	10.56		6729	6729.22	0.29	21	9.16	C_2 DIB
6185	6185.79	0.48	22	6.38		6993	6993.12	0.77	21	64.72	
6196	6195.99	0.51	24	47.06		7224	7224.16	1.12	19	162.17	
6203	6203.58	1.63	25	157.31		7367	7367.08	0.64	20	11.50	
6212	6211.69	0.62	20	8.95		7559	7559.43	0.91	20	14.17	
6234	6234.01	0.65	20	16.93		7562	7562.16	1.55	20	51.96	

[a]See Sonnentrucker et al. (2018).
[b]This DIB was divided into three separate DIBs in Fan et al. (2019) to reflect the structures within its profile. We re-measured the entire feature as one DIB in this work to follow the convention of most DIB studies.

3 SEARCHING FOR ANTICORRELATED DIBS

As is the case for any interstellar material, DIB carriers generally would accumulate over distance. Hence, their EWs always have a positive correlation with each other to a certain degree, especially when some of the most heavily reddened sightlines are included in the data sample (Friedman et al. 2011). To reveal possible anticorrelations among DIBs, we choose to follow the approach outlined in Cami et al. (1997) and work with the EWs of DIBs normalized by the reddening (i.e. using $W(\mathrm{DIB})/E_{\mathrm{B\text{-}V}}$). When discussing Pearson correlation coefficients using this normalized EW, we will denote them with r_{norm} to differentiate them from the 'regular' correlations using W(DIBs) that we will denote with r_{reg}. The total extinction A_{V} also scales with the gas column densities along the sightline and has been favoured in some works to normalize W(DIBs) (e.g. Ramírez-Tannus et al. 2018). While identifying the best normalizer of DIB strengths would be a worthy future project, it is beyond the scope of the current paper, and we found similar results as those reported in the following sections when analysing the correlations of $W(\mathrm{DIB})/A_{\mathrm{V}}$.

Fig. 1 provides some comparisons between r_{norm} and r_{reg}. The r_{norm} value is in most cases smaller than r_{reg} value of the same DIB pair. The highest r_{norm} value is between λλ6089 and 6379 with $r_{\mathrm{norm}} = 0.964$, and there are 22 DIB pairs with $r_{\mathrm{norm}} \geq 0.9$. Many of these DIB pairs are known to be well correlated when the comparison is between their EWs (e.g. McCall et al. 2010; Smith et al. 2021), such as λ6196 versus λ6613 ($r_{\mathrm{reg}} = 0.980$, $r_{\mathrm{norm}} = 0.952$) and λ6203 versus λ6284 ($r_{\mathrm{reg}} = 0.991$, $r_{\mathrm{norm}} = 0.936$).

While all r_{reg} values are positive, we find that 184 DIB pairs have negative r_{norm} values. They take up 12.9 per cent of the total possible combinations among the 54 target DIBs. The most anticorrelated DIB pairs, λ4984 versus λ7559, and λ5418 versus λ7562, have $r_{\mathrm{norm}} \sim -0.7$, and their scatter plots are presented in the middle and lower panels of Fig. 1. As indicated by the plotting axes, $W(\mathrm{DIBs})/E_{\mathrm{B\text{-}V}}$ still vary over a factor of 5–10 among the target sightlines. This reflects the impact of the environmental factors on DIBs, and how DIBs may be used to trace such differences (Vos et al. 2011; Kos & Zwitter 2013; van Loon et al. 2013; Bailey et al. 2015; Fan et al. 2017).

While the r values offer a good estimation on the similarity between two DIBs, the exact values depend on the composition of the sightline sample and the uncertainties in the measurements. There are 40 DIB pairs with $r_{\mathrm{norm}} \leq -0.5$ and 17 unique DIBs are involved in these pairs. We show their correlation coefficients in the heat map of Fig. 2, and two DIB groups emerge. The first group consists of DIBs λλ5705, 5780, 6203, 6270, 6284, 6324, 6353, 6362, 6993, 7224, 7559, and 7562, and many of them are known to be DIBs that remain prominent under strong radiation (i.e. in σ-type environments). The second group contains λλ4963, 4984, 5418, 5512, and 5546. They are all C_2-DIBs that trace denser regions of interstellar clouds and display different dependencies on environmental factors like radiation and density (Thorburn et al. 2003; Fan et al. 2017). This result suggests that DIB clusters or families can be identified according to their mutual correlations (e.g. Cami et al. 1997; Wszołek & Godłowski 2003; Omont & Bettinger 2020). That is, members of the same group share more similarities and thus have better correlations, while DIBs from different groups have reduced r values due to their different preferences on the environments.

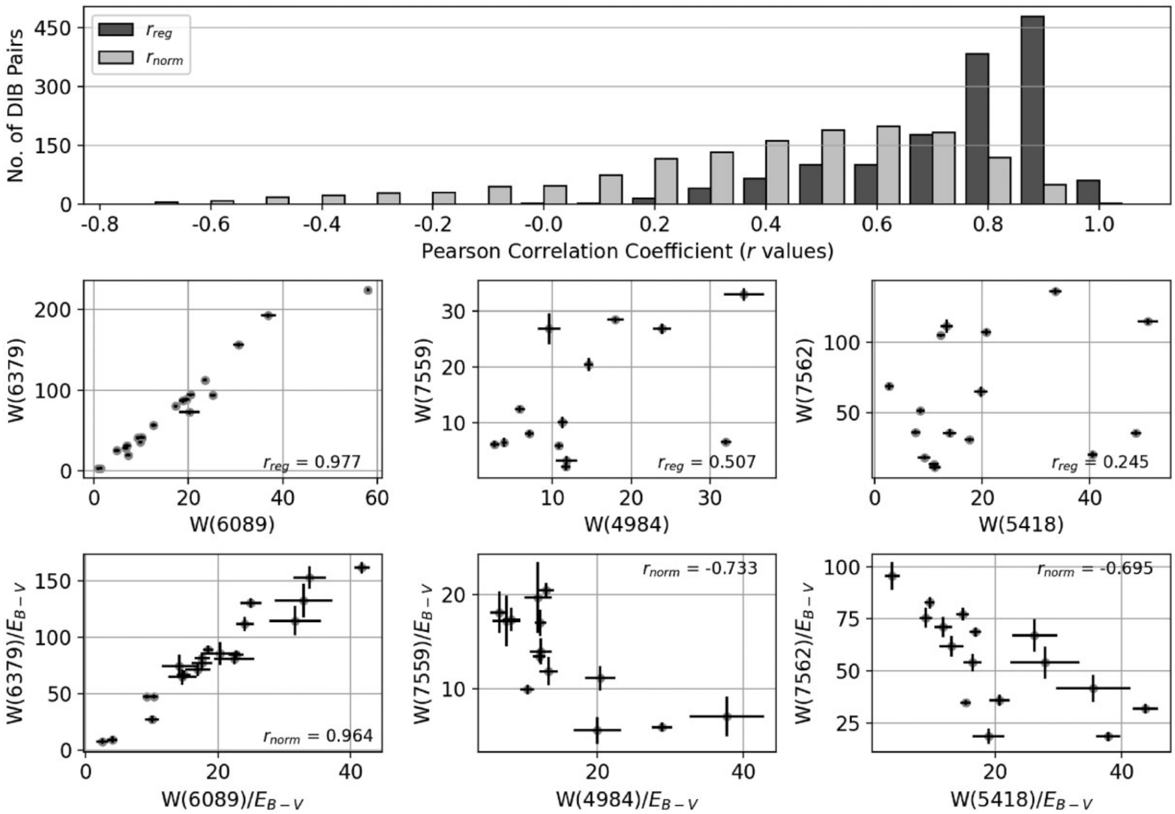

Figure 1. Comparison between $r_{\rm reg}$ and $r_{\rm norm}$. Upper panel: histogram of $r_{\rm reg}$ and $r_{\rm norm}$ values. We find that ∼1/7 of the DIB pairs have negative $r_{\rm norm}$ values, and the most anticorrelated DIB pairs have $r_{\rm norm} \sim -0.7$. Middle and bottom panels: scatter plots of W(DIB) and W(DIB)/$E_{\rm B\text{-}V}$ correlations between DIB pairs λ6089 versus λ6379, λ4984 versus λ7559, and λ5418 versus λ7562. The first pair has the highest $r_{\rm norm}$ value, while the second and third are the most anticorrelated DIB pairs in our data sample. The units for W(DIB) and W(DIB)/$E_{\rm B\text{-}V}$ are mÅ and mÅ· mag^{-1}, respectively.

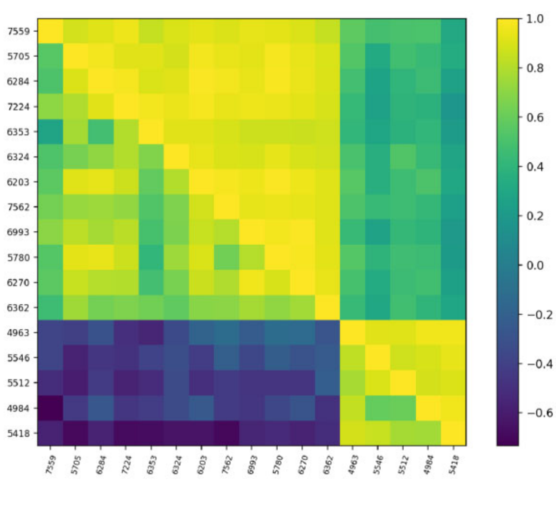

Figure 2. Heat map for the Pearson correlation coefficients among the 17 DIBs involved in the 40 DIB pairs with $r_{\rm norm} < -0.5$. Note that the figure is asymmetrical, where the lower triangle is for the $r_{\rm norm}$ values and the upper triangle is for the $r_{\rm reg}$ values. The DIBs are sorted by the sequence in Section 5 and two groups can be identified. They are, respectively, parts of the σ-type and C_2 DIB groups (Section 4), and all cross-group comparisons yield negative $r_{\rm norm}$ values. A full-scale heat map for all target DIBs of this work is presented in the Appendix B.

4 CLUSTERING OF DIBS

The apparent clustering of the C_2-DIBs in the previous section encouraged us to study clustering of the DIBs more closely based on their correlation coefficients. We first constructed a 54 × 54 matrix of $r_{\rm norm}$ values.[2] Each row of this matrix is an array of 54 $r_{\rm norm}$ values between 1.0 and −1.0 and can be interpreted as a coordinate of a parameter space. DIB families can be identified when a group of DIBs are located closely in this parametric space, i.e. when they are mutually well correlated and have similar $r_{\rm norm}$ values with other DIBs outside the group.

As a first attempt to cluster the target DIBs, we applied hierarchical agglomerative clustering (HAC[3]) to the $r_{\rm norm}$ matrix. With this method, the algorithm initially takes each DIB as a singleton cluster. During each iteration, two clusters are merged in such a way that the sum of within-cluster variance is kept minimal after they are merged [Ward's minimum variance method (Ward 1963)]. This process is repeated until all DIBs are merged into a single cluster and is similar to the approach by Baron et al. (2015) except we use a different 'linkage' function that determines which clusters to merge.

[2]We use the $r_{\rm norm}$ values since this work is originally motivated by the search for anticorrelated DIB pairs. We will show in Appendix C that the general picture of DIB correlations does not change if the analysis is based on $r_{\rm reg}$ values.

[3]Not to be confused with hydrogenated amorphous carbon that has been proposed as a DIB carrier.

This process is graphically represented in Fig. 3 as a dendrogram or 'tree diagram'. In the figure, the DIBs are sorted along the x-axis and gradually merged into nodes and finally the root on the top, and DIBs connected by a lower node share more similarities. The criterion to decide where precisely to cut off the tree (vertically) will then decide on the number of branches (clusters) to be kept. This criterion is subjective. However, as will be discussed in the next section, the number of groups is not very important since the DIBs in fact exhibit a rather continuous sequence. We thus choose to keep four clusters, indicated by different colours in Fig. 3). This clustering accounts for the three known DIB families discussed above (the λ5797 family, the λ5780 family, and the C_2-DIBs) plus a possible unidentified group. This results in the following four clusters:

(i) The σ −DIB group (green in Fig. 3), containing λλ4429, 5705, 5780, 6065, 6108, 6196, 6203, 6270, 6284, 6324, 6330, 6353, 6362, 6445, 6520, 6613, 6993, 7224, 7559, and 7562.

(ii) The intermediate DIB group (red), containing λλ4501, 4762, 5923, 6185, 6212, 6367, 6376, 6377, 6397, 6553, 6622, 6660, 6699, 6702, and 7367.

(iii) The ζ −DIB group (purple), containing λλ5545, 5766, 5793, 5797, 5849, 6089, 6116, 6234, 6379, 6439, 6449, and 6729.

(iv) The C_2 DIB group (orange), containing λλ4726, 4963, 4984, 5418, 5512, 5546, and 5828.

Unlike the classical approach where several DIBs are grouped solely for having good correlations, the HAC algorithm also considers whether they are less correlated with other DIBs to the same degree. This difference should be subtle since two perfectly correlated DIBs would always have the same r values with a third DIB. Our results are fully consistent with the literature for the well-known DIBs, such as for the typical σ-type DIBs λλ5705, 5780, and 6284, as well as ζ-type DIB λλ5797 and 6379 (Lan et al. 2015; Ensor et al. 2017; Krełowski 2018; Omont & Bettinger 2020; Galazutdinov et al. 2021). We also have eight C_2 DIBs from the original reference (i.e. Thorburn et al. 2003) and find six of them in the C_2 DIB group, while the two exceptions are assigned to the ζ-type group whose members also prefer shielded environments. Our clustering efforts also expand the knowledge to some less-often targeted DIBs. For example, DIBs λλ5849 and 6379 are a factor of 2 stronger in the shielded sightline of BD + 63° 1964 compared to HD 183143 (Ehrenfreund et al. 1997), and our analysis confirms them as ζ-type DIBs like λ5797. Lastly, a new 'intermediate' DIB group is introduced. We will show in Section 5 that members of this group have properties between the σ-type and ζ-type DIBs, and together they form a rather continuous spectrum of DIB behaviour.

We also adopt k-means clustering, another widely used clustering algorithm, to test the robustness of the grouping result. The k-means clustering algorithm sorts all data points into k groups so that: (a) the centre of each group is given by the average coordinate of its members, and (b) each data point is closer to its own group centre than to other group centres. The k-means clustering results are overall quite similar to our HAC results. Ten DIBs, however, are assigned to a different 'adjacent' group compared to the HAC results. Indeed, the seven DIBs in the λλ6196 and 6613 sub-branch of Fig. 3 are assigned to the intermediate group rather than σ group; the λ6397 DIB is assigned to the σ group rather than intermediate group; the λ6553 DIB is assigned to the ζ group rather than intermediate group; and the λ6729 DIB, recognized as a C_2 DIB in Thorburn et al. (2003), to C_2 group rather than ζ group. As will be shown in the next section, these DIBs are mostly located around the 'junction regions' between clusters. Their membership of specific DIB families is thus somewhat ambiguous, and the uncertainty in their membership would not change the overall picture of DIB behaviour that emerges from this clustering.

5 MULTIDIMENSIONAL SCALING ANALYSIS

While clustering algorithms sort the DIBs into groups, they provide limited information on any possible linkages between these groups. In this section, we use a Multi-Dimensional Scaling (MDS) analysis to visualize the r_{norm} matrix and provide a general picture on the similarities/dissimilarities among our target DIBs. We also tested other dimensional reduction and data visualization algorithms such as t-Distributed Stochastic Neighbor Embedding (tSNE, see Van der Maaten & Hinton 2008) and UMAP (McInnes, Healy & Melville 2018), and both methods produce very similar results.

The MDS algorithm maps a set of N data points on to an abstract M-dimensional Cartesian space (with $M < N$). The only input that MDS requires is an N-by-N *dissimilarity matrix* that contains a measure for pairwise distances among the N observations. Somewhat akin to a principal component analysis (PCA), the algorithm then maps these N points on to the new M-dimensional space in such a way that these pairwise distances are preserved as much as possible. This is done by minimizing the *stress* function:

$$\mathrm{Stress} = \sqrt{\frac{\sum_{i,j}(d_{ij} - \hat{d}_{ij})^2}{\sum_{i,j} d_{ij}^2}}. \quad (1)$$

Here, d_{ij} is the observed distance from the input matrix, and $\hat{d}_{ij}$ is the distance between points in the new M-dimensional space mapped by the algorithm. The details of how the algorithm optimizes this mapping can be found in Mead (1992). An classical example of MDS is to feed its pairwise distances between cities and let MDS reconstruct a map from those distances.

Since we are using Pearson correlation coefficients that measure the similarity between two DIBs, we use $1 - r_{\mathrm{norm}}$ as input values for the dissimilarity matrix. In this way, the dissimilarity between two DIBs is minimized to zero when they are perfectly correlated and maximized to 2.0 when they are perfectly anticorrelated. We also include correlations with the column densities of H_2, CH, C_2, and CN in the analysis – i.e. we expanded our dissimilarity matrix to the size of 58 × 58 that includes correlations with these column densities. Their r_{norm} values are calculated normalized to the $E(B-V)$ as is the case for the DIBs, and the resulting $1 - r_{\mathrm{norm}}$ is added to the dissimilarity matrix.

Including more dimensions (i.e. the higher M is) would always reduce the stress (equation 1) and bring better agreement between the input dissimilarities and the distances in the new M-dimensional space. However, it is equally important to use a small number of dimensions so that the results are easier to be interpreted. For the visualization purposes, it is most common for MDS to map the dissimilarity matrix to a one-dimensional (1D), two-dimensional (2D), or three-dimensional (3D) space so that the result can be demonstrated as scatter plots (Borg & Groenen 2005). In Fig. 4, we examine the results of 1D-, 2D-, and 3D-MDS by comparing the pairwise input dissimilarity $[d_{ij}] = [1 - r_{ij}]$ to the pairwise distances $[\hat{d}_{ij}]$ in the new M-dimensional space. If the MDS algorithm would have preserved all distances, the comparison should yield $Y = X$, but we find considerable scatter in the top panel that shows the output of the 1D-MDS. Mapping the DIBs along a single axis is thus not sufficient to fully explain their dissimilarities, and an additional mapping axis is required. The scatter is greatly reduced after adding a second projection axis, and including a third axis provides little

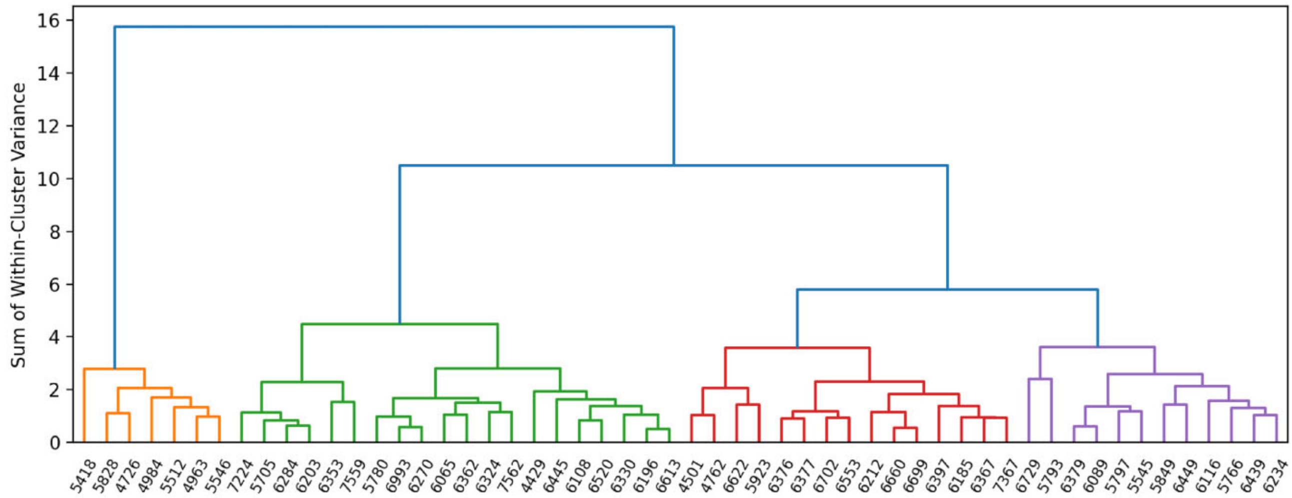

Figure 3. Dendrogram for the hierarchical agglomerative clustering analysis. The 54 DIBs are arranged along the *x*-axis and gradually merged into the root on top of the plot. The horizontal bars indicate which clusters/DIBs are being merged during each iteration and the sum of within-cluster variance after the merge. We choose to keep four clusters that correspond to the σ-type (green), ζ-type (purple), C_2 DIBs (orange), plus an 'intermediate' (red) group whose properties is between the σ- and ζ-types of DIBs.

improvement (Fig. 4 middle and bottom panels). We will thus adopt the result of the 2D-MDS analysis on our dissimilarity matrix for the remainder of this paper.

Fig. 5 presents the 2D-MDS result, i.e. the locations of the normalized parameters (W(DIBs)/$E_{\text{B-V}}$ or N(Xs)/$E_{\text{B-V}}$) in the new 2D space, using the same colours for different DIB groups as in Fig. 3. The 10 DIBs assigned to different groups by the HAC and *k*-means clustering algorithms are plotted with different edge and face colours. Note that the MDS analysis focuses on distances between points, and that distances would not change if the projection coordinate is rotated, flipped, or translated. Thus, the plotting axes of Fig. 5 do not necessarily have a physical meaning. We focus on the general layout and trend of the projected points.

The clustering result in Section 4 remains valid in the MDS analysis. From the lower left to the upper right of Fig. 5, we find in a roughly linear manner of the σ-type, intermediate, ζ-type, and finally the C_2 DIBs and the molecular species. This trend agrees with the general knowledge on how DIBs react to environmental factors especially regarding the radiation field: σ-DIBs like $\lambda\lambda$5780 and 6284 are much more prominent in radiative environments than in shielded environments, whereas ζ-DIBs like λ5797 are strong in shielded environments as well, and the C_2 DIBs trace denser, more shielded regions than other DIBs. Such diversity in DIB behaviour is often associated with the ionization potentials of their carriers (Cami et al. 1997; Sonnentrucker et al. 1997), although other mechanism like hydrogenation, dehydrogenation, and depletion may be influencing DIBs and other ISM species alike (e.g. Cardelli 1994; Jensen & Snow 2007a, b; Welty & Crowther 2010; Zhen et al. 2014; Fan et al. 2017). We also find DIBs assigned to different groups by the HAC and *k*-means clustering algorithms to be mostly located in the inter-cluster regions. These DIBs may not be the 'typical members' of any of the DIB groups described earlier, making their assignments more difficult.

Fan et al. (2017) proposed a sequence of eight DIBs based on how their strength ratios change with the f_{H2} value of the sightline. In order of favouring decreasing radiation and increasing shielding, the sequence goes $\lambda\lambda$6284 and 5780, then $\lambda\lambda$6196 and 6613, then λ5797, and finally the C_2 DIBs $\lambda\lambda$4727, 4963, and 4984. This is fully consistent with the more expanded trend observed in Fig. 5. To extract this sequence in our data, we perform a linear fit to the DIBs [i.e. excluding all N(Xs)] and project all points to this best-fitting line. The observed sequence is as follows: $\lambda\lambda$7559, 5705, 6284, 7224, 6353, 6324, 6203, 7562, 6993, 5780, 6270, 6362, 6065, 6520, 6445, 6108, 6196, 6613, 6330, 5923, 4429, 6367, 6376, 6212, 6622, 7367, 6377, 4501, 6702, 6397, 6699, 6660, 6553, 6185, 4762, 6379, 5797, 5793, 6234, 6089, 6439, 5766, 5545, 6116, 6449, 5849, 5828, 6729, 4726, 4963, 5546, 5512, N(CH), N(H_2), 4984, 5418, N(C_2), and N(CN).

6 DISCUSSION

6.1 DIB families

DIB families have been discussed in various publications as the result of correlation analysis and the observed changes in their band strength ratios (e.g. Krelowski & Walker 1987; Cami et al. 1997; Moutou et al. 1999; Lan et al. 2015). By definition, DIBs from the same family demonstrate similar behaviour under varying ISM conditions. They are thus mutually well correlated and have relatively constant strength ratios, and their carriers are expected to share certain properties such as ionization potentials.

The effort of DIB classification has been carried out for strong DIBs since their measurements are more accessible. Some well-defined DIB families, like the σ-type, ζ-type, and the C_2 DIBs, have been described in Section 4, and such classification echos studies on DIB profiles (e.g. Josafatsson & Snow 1987; Galazutdinov et al. 2003; Wszołek & Godłowski 2003). For example, some σ-type DIBs like $\lambda\lambda$5780 and 6284 have broad and smooth profiles, while ζ-type DIB λ5797 (along with many other DIBs) have narrow profile and clear substructures, indicating gas-phase molecules as their carriers (Foing & Ehrenfreund 1994; Sarre et al. 1995; Ehrenfreund & Foing 1996).

On the other hand, the classification of certain DIBs can be ambiguous and the result varies among analyses. This issue first results from the uncertainties in correlation analysis, especially when weak DIBs are involved (e.g. Cami et al. 1997; Krełowski 2018). The *r* values observed are dependent on factors such as the sample of sightlines, data quality, and the selected measuring method. It is

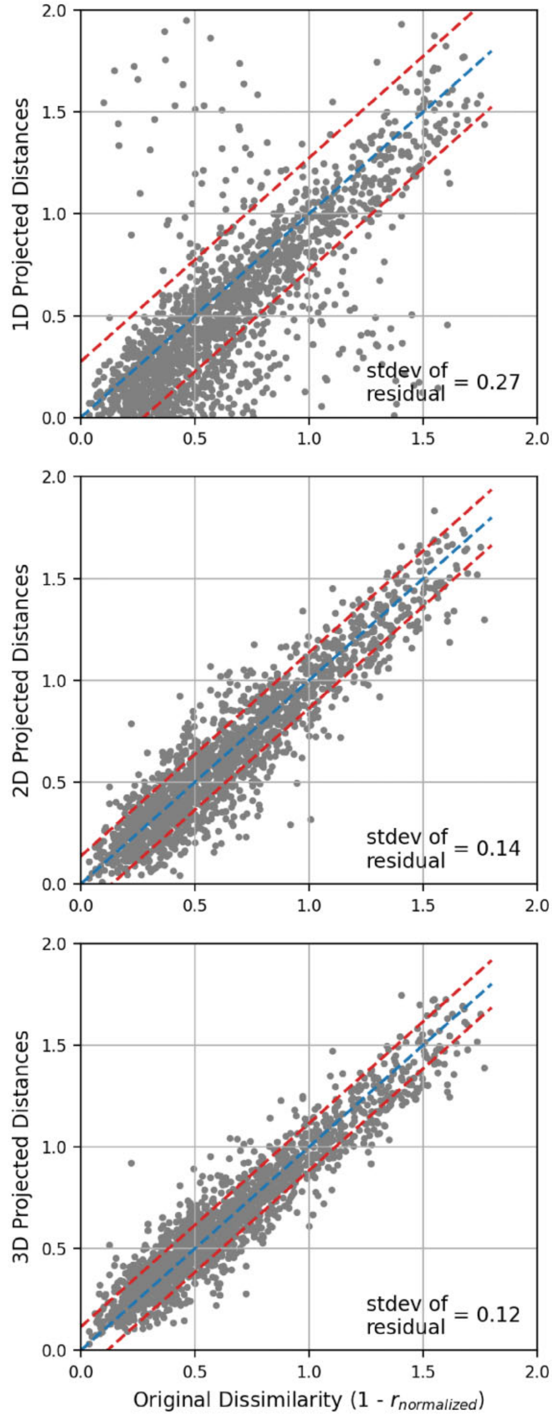

Figure 4. Comparing the results of the 1D-, 2D-, and 3D-MDS analyses. Each panel compares the input $d = 1 - r_{\mathrm{norm}}$ values (abscissa) to the MDS-mapped distances (ordinate). The blue dashed lines represent the ideal case $Y = X$ and the red dashed lines illustrate the standard deviation of the residuals.

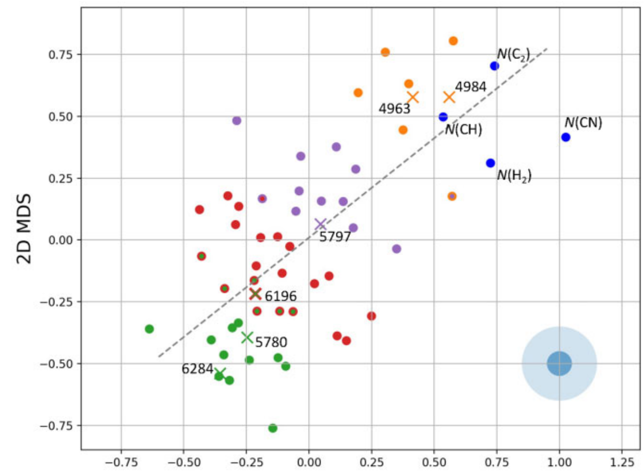

Figure 5. Scatter plot for the 2D-MDS results, where each point represents a DIB (or other tracer) and the distance between them represents their $1 - r_{\mathrm{normalized}}$ dissimilarity. The plotting axes are abstract coordinates in the new 2D space and do not necessarily represent a physical quantity – hence, they are not labelled. We follow the same colour code as in Fig. 3 for DIB groups, except for the 10 DIBs assigned to different groups by the HAC and *k*-means clustering algorithms. For these DIBs, the face (inner) colour represents the HAC result and the edge (outer) colour represents the *k*-means result, and they are located around the transition regions between clusters. DIBs λλ6284, 5780, 6196, 5797, 4963, and 4984 are highlighted, and we also include $N(H_2)$, N(CH), $N(C_2)$, and N(CN) for comparison (blue dots). The dashed line is a least-squares fit representing a straight line through all DIB points. The transparent blue circles in the lower right corner have radii of 0.05 (dark blue) and 0.15 units (light blue) and thus correspond to r_{norm} values of 0.95 and 0.85, respectively. Thus, points that are separated by the radius of the light blue circle have a mutual correlation coefficient of 0.85. We find a rather smooth transition from one DIB group to the next especially among the non-C_2 DIBs. The molecular species are all located around the C_2 DIBs and they seem to form a somewhat separate cluster from the non-C_2 DIBs. We also provide an enlarged version of this plot in Fig. B2, where all data points have been labelled.

thus very hard to compare r values across different analyses in a qualitative manner. The assignments can still be difficult for some strong DIBs with good measurements and well-defined correlation coefficients. For example, DIBs λλ6196 and 6613 are known for their close-to-perfect correlation (Cami et al. 1997; Moutou et al. 1999; Galazutdinov et al. 2002; McCall et al. 2010; Bailey et al. 2016). But since they are equally well correlated with both typical σ-type DIB λ6284 and ζ-type DIB λ5797 (e.g. Friedman et al. 2011; Fan et al. 2017), which DIB family should they be assigned to?

6.1.1 σ-type, intermediate, and ζ-type DIBs

By including most of the strong DIBs in the optical region, our analysis finds rather continuously distributed data points in Fig. 5, especially among the non-C_2 DIBs (i.e. the σ-type, intermediate, and ζ-type groups). For most of these DIBs, several other DIBs can be found within an ∼0.15 radius and sometimes from a different DIB family.

This continuous trend among the non-C_2 DIBs goes against sorting them into several distinguishable clusters. While the behaviour of typical σ- and ζ-type DIBs can be very different under different ISM conditions, there are many DIBs between them and the transition is gradual and without clear boundaries. The membership of certain DIBs to specific clusters can be thus ambiguous, such as the 10 DIBs assigned to different groups by the HAC and *k*-means clustering algorithms (Section 4). In Fig. 5, all these DIBs are located at the junctions of the neighbouring groups, whereas the terms σ- and ζ-type DIBs may be applied to only some of the most representative DIBs like λλ5780, 5797, and 6284.

DIB correlations reflect the similarities between their behaviour under different ISM conditions and thus potentially the underlying properties of their carriers. The continuous trend among non-C_2 DIBs suggests progressive changes in the response to their environments, due to e.g. gradually changing ionization potentials or maybe molecular sizes of their carriers.

6.1.2 The C_2 DIBs

The C_2 DIBs are first introduced in Thorburn et al. (2003) and described as 'a class of weak, narrow bands whose normalized EWs W(X)/W(6196) are well correlated specifically with $N(C_2)/E_{B\text{-}V}$ via power laws'. However, despite seemingly suggested by the name, many C_2 DIBs are in fact better correlated with $E_{B\text{-}V}$ than with $N(C_2)$ (e.g. Galazutdinov et al. 2006; Elyajouri et al. 2018). We note that some of the strongest C_2 DIBs are observed towards HD 37061 and HD 37903 (albeit at greatly reduced strengths; see Fan et al. 2019). However, these sightlines do not show any evidence for C_2 absorption (Fan et al., in preparation). We thus emphasize that the detection of the C_2 DIBs does not depend on the prior existence of the C_2 molecules.

Despite a list of C_2 DIBs, Thorburn et al. (2003) do not provide a quantitative examination to check if a new DIB belongs to the family. The membership of the C_2 DIB family thus varies among publications when different definitions are adopted (e.g. Galazutdinov et al. 2006; Elyajouri et al. 2018). In this work, six of the eight C_2 DIBs identified as such by Thorburn et al. (2003) are grouped together when using the HAC method and seven of them when performing k-means clustering. The remaining one or two C_2-DIBs in both cases are assigned to the ζ-type group. This seems acceptable, given the loose definition of C_2 DIBs, the uncertainties in the correlation coefficients, and the fact that the ζ-type DIBs also trace denser regions than other non-C_2 DIBs. At the same time, we also identify the λ5828 DIB [not targeted in Thorburn et al. (2003)] as a promising new member of the C_2 DIB family, since both our clustering methods and the MDS analysis put it squarely in the same group as the other C_2-DIBs.

The C_2 DIBs are known to differ from non-C_2 DIBs especially regarding the 'skin effect'. This phenomenon refers to the reduced EWs of certain DIBs (relative to $E_{B\text{-}V}$) when the sightline passes through denser regions of the ISM cloud (Wampler 1966; Strom et al. 1975; Meyer & Ulrich 1984; Herbig 1995). It is best explained if those DIB carriers are more abundant in the outer layers ('skin') of interstellar clouds, whereas the denser internal regions contribute little to the column density of DIB material. A survey of C_2 and C_3 (Fan et al., in preparation) in the EDIBLES (Cox et al. 2017) data set finds that the EWs of the C_2 DIBs are in fact indifferent to C_2 and C_3 detection in the sightlines. Thus, the C_2 DIBs are even less sensitive to the skin effect and may thus trace denser regions than the non-C_2 DIBs (Thorburn et al. 2003; Fan et al. 2017). But since the C_2 DIBs are neither enhanced in sightlines with C_2, it is likely that the C_2 DIBs are tracing less dense regions in the ISM clouds than the C_2 molecules.

The current analysis suggests that the C_2 DIBs are really separated from the non-C_2 DIBs. In Fig. 3, they are the two top branches of the DIB tree. In Fig. 5, there seems to be a boundary between the C_2 and ζ-type DIBs, despite a continuous trend among the non-C_2 DIBs and that our analysis includes most of the strong optical DIBs. And finally in Fig. B1 that presents the r values between all target DIBs, the non-C_2 DIBs appear to form an extension of the continuous progression, while there is a noticeable gap between the C_2 and the non-C_2 DIBs, especially for the r_{reg} values. This gap is even more pronounced when carrying out the MDS analysis using the non-normalized correlation coefficients (i.e. based on r_{reg}; see Fig. C2). The C_2 DIBs may thus arise from a very different family of molecules than the non-C_2 DIBs.

6.2 Anticorrelations

As discussed in Section 3, the most robust anticorrelations we identify are all between a σ-type DIB and a C_2 DIB (see also Fig. 2). The most anticorrelated DIB pairs have $r_{norm} \sim -0.7$ and are far away from a perfect anticorrelation. This lack of perfect anticorrelation is less likely to be the result of the uncertainties introduced in the normalization process, since we are able to identify plenty of DIB pairs at r_{norm} ~0.9 level.

This lack of a perfect anticorrelation makes sense if we consider the implications from the point of view of DIB environmental behaviour. For several DIBs, it has been documented that there is a 'rise and fall' of their strengths when comparing it to indicators for the exposure to radiation (see e.g. Cami et al. 1997; Sonnentrucker et al. 1997). Fan et al. (2017) demonstrate that the strengths of the σ-type DIBs such as $\lambda\lambda$5780 and 6284 demonstrate a clear Λ-shaped behaviour when compared to the mass fraction of molecular hydrogen f_{H2}: their W(DIBs)/$E_{B\text{-}V}$ peaks at a 'sweet spot' of $f_{H2} \sim 0.2$ and decreases towards the low f_{H2} end due to radiation and towards the high f_{H2} end due to the skin effect. Given the tight mutual correlations among the σ-type DIBs, we can expect similar Λ-shaped behaviours for the other members. Thus, in order for a DIB to have a perfect anticorrelation with a σ-type DIB, it should exhibit a 'V-shaped' behaviour with f_{H2}. In that case, the hypothesized DIB carrier must thrive under the most exposed and shielded environment at the same time, which seems not plausible.

The above discussion applies only to ideal condition with infinite sensitivities. In reality, the strengths of the C_2 DIBs decrease dramatically towards the low f_{H2} end since they are more sensitive to the presence of radiation field. By dropping below the detection limit in low f_{H2} sightlines, a C_2 DIB may demonstrate a portion of the required 'V-shaped' behaviour, but the degree of anticorrelation is dependent on the composition of the sightline sample. For example, the sightlines involved in the most anticorrelated DIB pairs λ4984 versus λ7559 and λ5418 versus λ7562 all have $f_{H2} > 0.2$. In this f_{H2} region, the strengths of the σ-type DIBs start to decrease while $W(C_2$ DIBs)/$E_{B\text{-}V}$ remain roughly constant but with large scatter (Fan et al. 2017). The anticorrelation between the σ-type and C_2 DIBs thus reflects how they demonstrate different behaviours in sightlines with medium to large f_{H2} values.

Since we have targeted most of the strong and commonly detected DIBs in the optical region, it seems safe to conclude that we cannot expect perfect anticorrelations from these DIBs. On the other hand, the near-infrared DIBs may demonstrate quite different behaviour than the optical DIBs (Cox et al. 2014; Hamano et al. 2015, 2016), but we are not able to target them in our spectral data. Future projects would also benefit from including more sightlines and targeting weaker DIBs. This would also help to identify more close-to-perfect positive correlations, since molecules are expected to a few strong features along with more well-correlated weaker spectral signatures.

6.3 Factors governing DIB behaviour

In Section 5 and Fig. 4, we find that at least two projection axes are required to properly reproduce the $1 - r_{norm}$ dissimilarities among DIBs. These two projection axes may correspond to two underlying physical factors that set the degree of correlations between DIBs.

Using a PCA, Ensor et al. (2017) find that four principal components together determine ~ 93 per cent of the variations in the observed DIB strengths and sightline parameters in a sample of single-cloud sightlines. The first and most dominant factor is well traced by $W(5797)$ and is interpreted as the total amount of DIB-producing material along the sightline. To first order, this factor would then determine the approximate strengths of all DIBs and be the main cause for all DIBs showing some degree of mutual correlation. The actual r values are then determined by how much the data points deviate from this basic linear relationship in response to the physical conditions.

The second of the four factors is best traced by the $W(5780)/W(5797)$ ratio, which is often used as a measure of UV exposure of the sightline. The radiation field can influence the behaviour of interstellar species via photo-ionization and photo-dissociation, and the presence of a strong radiation field is often associated with lower abundances of molecular species and reduced W(DIBs) along the sightline (see e.g. Savage et al. 1977; Herbig 1995; Welty & Hobbs 2001; Cox et al. 2006; Friedman et al. 2011; Vos et al. 2011). However, it is important to realize that the $W(5780)/W(5797)$ ratio does not trace UV exposure directly but rather the ratio of UV exposure to the density. This is perhaps best illustrated by the sightline towards star number 46 of IC 62 that penetrates a high-density photodissociation region (PDR) with $n > 10^4$ cm^{-3} (Lai et al. 2020). The UV radiation in this PDR is 150 times stronger than in the typical diffuse environments (Andrews et al. 2018), leaving hydrogen in almost purely atomic form. Given the very intense radiation field, one would expect most of the DIBs to have greatly reduced strengths or vanish as seen in the sightlines towards the Orion Trapezium stars (Fan et al. 2017). However, Lai et al. (2020) find that the normalized EWs and strength ratios of the DIBs are comparable to a regular interstellar low-f_{H2} sightline. This observation thus suggests that different DIB behaviour [and thus the $W(5780)/W(5797)$ ratio] is related to the H I/H_2 ratio (which in itself is determined by the radiation field and the density) rather than to the radiation field directly. This should be kept in mind for the discussion that follows.

It is common for DIB strength ratios to vary among sightlines harbouring different environmental factors especially regarding the intensity of UV radiation. Fan et al. (2017) propose a sequence of eight DIBs to account for such variations (see also van Loon et al. 2013), and we find a more extended sequence in Fig. 5. This sequence follows the transitions between DIB families and acts as the first projection axis of our MDS analysis. In the classical radiation-shielding picture of DIB behaviours, this sequence could reflect the ionization potentials of DIB carriers. In this picture, carriers of the C_2 and ζ-type DIBs have lower ionization potential and require more shielding, while σ-type DIBs arise from molecules of higher ionization potentials that are able to survive in more exposed environments (Cami et al. 1997; Sonnentrucker et al. 1997; Farhang et al. 2019). The DIB families are formed among DIB carriers with approximate ionization potentials that would trace similar environment and thus develop good correlations.

This, of course, is a simplified picture and in reality the photo-ionization process depends on the ionization parameter G_0/n and thus the density (which has been assumed be constant in many cases), and other factors like ionization cross-sections and temperature may also contribute. One should also consider the possible role of hydrogenation and dehydrogenation that depends on the density of hydrogen and the intensity of UV radiation field (e.g. Vuong & Foing 2000). Large organic molecules, especially those with more than 50 carbon atoms, are expected to be stable and remain well hydrogenated in typical interstellar radiation fields (e.g. Allain, Leach & Sedlmayr 1996; Omont & Bettinger 2021).

Additional factors influencing the observed $W(\mathrm{DIBs})/E_{\mathrm{B\text{-}V}}$ and thus their correlations have been proposed in the literature. For example, anomalously weak DIBs are detected towards bright stars in the Large and Small Magellanic Clouds (LMC and SMC) and have been attributed to lower metallicities (Cox & Spaans 2006; Welty et al. 2006; Bailey et al. 2015). In this case, the imperfect DIB correlations may be partly due to the variable metallicity of the local ISM (De Cia et al. 2021; Zuo, Li & Zhao 2021). By comparing DIBs to known interstellar atomic and molecular species, Fan et al. (2017) find that mechanisms such as depletion on to dust grains or chemical reactions of various sorts may be required to explain the decreased $W(\mathrm{DIBs})/E_{\mathrm{B\text{-}V}}$ in denser regions. The PCA analysis by Ensor et al. (2017) suggests dust-related factors, such as dust-to-gas ratio or grain size along the sightline, are affecting the strength of DIBs. Many of these hypotheses involve the interaction between DIB carriers or their precursors and other species or particles but should play a minor role compared to the radiation field. The second axis would then reflect related properties of DIB carriers that lead to different behaviour in the same environment, such as the reaction rate of the key process(es).

On the other hand, it is also possible that the second axis reflects the non-linear effect in correlation coefficients and their uncertainties (Cami et al. 1997). But this would not change our findings that DIB carriers respond to the radiation field in a relatively continuous manner. These DIBs may trace certain environments, especially the diffuse atomic gas regions (van Loon et al. 2013; Bailey et al. 2015). In principle, the strength ratios between any two DIBs that are sufficiently apart in the sequence can be used to characterize the radiation field, similar to the $W(5780)/W(5780)$ ratio and the $\sigma - \zeta$ effect (Krelowski & Walker 1987; Vos et al. 2011; Kos & Zwitter 2013).

7 SUMMARY AND CONCLUSIONS

In this work, we set off from a uniform data set sampling 54 strong and commonly detected DIBs in 25 sightlines representing various ISM environments. We investigate their pairwise correlations and provide further analysis using data science tools like clustering and MDS. The major conclusions we have reached are as follows:

(i) Using normalized EWs $W(\mathrm{DIBs})/E_{\mathrm{B\text{-}V}}$, we confirm the common presence of anticorrelated DIB pairs, most notably between the C_2 DIBs and the σ-type DIBs like λλ5780 and 6284. We find several DIB pairs with $r_{\mathrm{norm}} \geq 0.9$, but the most negative r_{norm} values are ~ -0.7, far from a perfect anticorrelation. We do not see convincing evidence that any of the anticorrelated DIB pairs is from successive ionization states of a single carrier.

(ii) We use multidimensional scaling (MDS) to visualize the $1 - r_{\mathrm{norm}}$ dissimilarities as distances between data points. At least two projection axes are required to properly explain the dissimilarities among DIBs. The first factor is associated with radiation. The second factor is not fully understood and might be related to the interaction of DIB carriers with other particles, species, or dust grains.

(iii) Hierarchical agglomerative clustering and k-means clustering reproduce previous divisions of DIB families, including the σ-, ζ-, and C_2 DIBs. However, the MDS analysis shows a rather continuous sequence especially among the non-C_2 DIBs, and the σ- and ζ-type DIBs appear to be the two extreme ends of this continuous sequence. The continuous sequence suggests that the DIB carriers differ by a property that is continuously variable, such as ionization potential and maybe molecular size. There is a gap between the non-C_2 DIBs and the C_2 DIBs, which suggests more fundamental differences between the carriers of the two groups of DIBs.

ACKNOWLEDGEMENTS

HF, MS and JC acknowledge support from an NSERC Discovery Grant, a Western SERB Accelerator Award, and the USRI program. HF would like to thank Prof. Gang Zhao from NAOC and Prof. Donald G. York from the University of Chicago for their supervision and kind support during his PhD study. The hierarchical agglomerative clustering (HAC), *k*-means clustering, and multidimensional scaling (MDS) analyses in this work are based on the scikit-learn package (Pedregosa et al. 2011). This research has made use of NASA's Astrophysics Data System Bibliographic Services and the SIMBAD data base operated at CDS, Strasbourg, France (Wenger et al. 2000).

DATA AVAILABILITY

DIB measurements used in this work are taken from Fan et al. (2019). These data are open-access at https://cdsarc.unistra.fr/viz-bin/cat/J/ApJ/878/151. The code used to generate the figures is available at https://github.com/HaoyuFan-DIB/FamilyAndClusterOfDIBs.

REFERENCES

Adamson A. J., Whittet D. C. B., Duley W. W., 1991, MNRAS, 252, 234
Allain T., Leach S., Sedlmayr E., 1996, A&A, 305, 616
Andrews H., Peeters E., Tielens A. G. G. M., Okada Y., 2018, A&A, 619, A170
Bailey M., van Loon J. T., Farhang A., Javadi A., Khosroshahi H. G., Sarre P. J., Smith K. T., 2016, A&A, 585, A12
Bailey M., van Loon J. T., Sarre P. J., Beckman J. E., 2015, MNRAS, 454, 4013
Baron D., Poznanski D., Watson D., Yao Y., Cox N. L. J., Prochaska J. X., 2015, MNRAS, 451, 332
Borg I., Groenen P., 2005, Modern Multidimensional Scaling: Theory and Applications. Springer Series in Statistics. Springer New York, https://books.google.ca/books?id = duTODldZzRcC
Cami J., Salama F., Jiménez-Vicente J., Galazutdinov G. A., Krełowski J., 2004, ApJ, 611, L113
Cami J., Sonnentrucker P., Ehrenfreund P., Foing B. H., 1997, A&A, 326, 822
Campbell E. K., Holz M., Gerlich D., Maier J. P., 2015, Nature, 523, 322
Cardelli J. A., 1994, Science, 265, 209
Cordiner M. A. et al., 2017, ApJ, 843, L2
Cordiner M. A. et al., 2019, ApJ, 875, L28
Cox N. L. J. et al., 2017, A&A, 606, A76
Cox N. L. J., Cami J., Kaper L., Ehrenfreund P., Foing B. H., Ochsendorf B. B., van Hooff S. H. M., Salama F., 2014, A&A, 569, A117
Cox N. L. J., Cordiner M. A., Cami J., Foing B. H., Sarre P. J., Kaper L., Ehrenfreund P., 2006, A&A, 447, 991
Cox N. L. J., Ehrenfreund P., Foing B. H., D'Hendecourt L., Salama F., Sarre P. J., 2011, A&A, 531, A25
Cox N. L. J., Spaans M., 2006, A&A, 451, 973
Dahlstrom J. et al., 2013, ApJ, 773, 41
De Cia A., Jenkins E. B., Fox A. J., Ledoux C., Ramburth-Hurt T., Konstantopoulou C., Petitjean P., Krogager J.-K., 2021, Nature, 597, 206
Ehrenfreund P., Cami J., Dartois E., Foing B. H., 1997, A&A, 318, L28
Ehrenfreund P., Foing B. H., 1996, A&A, 307, L25
Elyajouri M. et al., 2018, A&A, 616, A143
Ensor T., Cami J., Bhatt N. H., Soddu A., 2017, ApJ, 836, 162
Fan H. et al., 2017, ApJ, 850, 194
Fan H. et al., 2019, ApJ, 878, 151
Farhang A., van Loon J. T., Khosroshahi H. G., Javadi A., Bailey M., 2019, Nature Astron., 3, 922
Foing B. H., Ehrenfreund P., 1994, Nature, 369, 296
Foing B. H., Ehrenfreund P., 1997, A&A, 317, L59
Friedman S. D. et al., 2011, ApJ, 727, 33
Galazutdinov G. A., Gnaciński P., Han I., Lee B.-C., Kim K.-M., Krełowski J., 2006, A&A, 447, 589
Galazutdinov G. A., Musaev F. A., Bondar A. V., Krełowski J., 2003, MNRAS, 345, 365
Galazutdinov G. A., Valyavin G., Ikhsanov N. R., Krełowski J., 2021, AJ, 161, 127
Galazutdinov G., Moutou C., Musaev F., Krełowski J., 2002, A&A, 384, 215
González Hernández J. I., Iglesias-Groth S., Rebolo R., García-Hernández D. A., Manchado A., Lambert D. L., 2009, ApJ, 706, 866
Hamano S. et al., 2015, ApJ, 800, 137
Hamano S. et al., 2016, ApJ, 821, 42
Heger M. L., 1922, Lick Obs. Bull., 10, 146
Herbig G. H., 1993, ApJ, 407, 142
Herbig G. H., 1995, ARA&A, 33, 19
Hobbs L. M. et al., 2008, ApJ, 680, 1256
Hobbs L. M. et al., 2009, ApJ, 705, 32
Jenniskens P., Ehrenfreund P., Foing B., 1994, A&A, 281, 517
Jensen A. G., Snow T. P., 2007a, ApJ, 669, 378
Jensen A. G., Snow T. P., 2007b, ApJ, 669, 401
Joblin C., Maillard J. P., D'Hendecourt L., Léger A., 1990, Nature, 346, 729
Josafatsson K., Snow T. P., 1987, ApJ, 319, 436
Kerr T. H., Hibbins R. E., Fossey S. J., Miles J. R., Sarre P. J., 1998, ApJ, 495, 941
Kos J., Zwitter T., 2013, ApJ, 774, 72
Krełowski J., 2018, PASP, 130, 071001
Krelowski J., Walker G. A. H., 1987, ApJ, 312, 860
Lai T. S. Y., Witt A. N., Alvarez C., Cami J., 2020, MNRAS, 492, 5853
Lallement R. et al., 2018, A&A, 614, A28
Lan T.-W., Ménard B., Zhu G., 2015, MNRAS, 452, 3629
Linnartz H., Cami J., Cordiner M., Cox N., Ehrenfreund P., Foing B., Gatchell M., Scheier P., 2020, J. Mol. Spectrosc., 367, 111243
McCall B. J. et al., 2010, ApJ, 708, 1628
McCall B. J., Thorburn J., Hobbs L. M., Oka T., York D. G., 2001, ApJ, 559, L49
McInnes L., Healy J., Melville J., 2018, preprint (arXiv:1802.03426)
Mead A., 1992, J. R. Stat. Soc. D Stat., 41, 27
Meyer D. M., Ulrich R. K., 1984, ApJ, 283, 98
Moutou C., Krełowski J., D'Hendecourt L., Jamroszczak J., 1999, A&A, 351, 680
Omont A., Bettinger H. F., 2020, A&A, 637, A74
Omont A., Bettinger H. F., 2021, A&A, 650, A193
Pedregosa F. et al., 2011, J. Mach. Learn. Res., 12, 2825
Ramírez-Tannus M. C., Cox N. L. J., Kaper L., de Koter A., 2018, A&A, 620, A52
Rawlings M. G., Juvela M., Lehtinen K., Mattila K., Lemke D., 2013, MNRAS, 428, 2617
Sarre P. J., Miles J. R., Kerr T. H., Hibbins R. E., Fossey S. J., Somerville W. B., 1995, MNRAS, 277, L41
Savage B. D., Bohlin R. C., Drake J. F., Budich W., 1977, ApJ, 216, 291
Schlarmann L., Foing B., Cami J., Fan H., 2021, A&A, 656, 17
Schlarmann L., Foing B., Cami J., Fan H. , 2021, A&A, 656L, 17S
Smith F. M., Harriott T. A., Majaess D., Massa L., Matta C. F., 2021, MNRAS, 507, 5236
Snow T. P. J., Cohen J. G., 1974, ApJ, 194, 313
Sonnentrucker P., 2014, in Cami J., Cox N. L. J., eds, Proc. IAU Symp. 297, The Diffuse Interstellar Bands. p.13
Sonnentrucker P., Cami J., Ehrenfreund P., Foing B. H., 1997, A&A, 327, 1215
Sonnentrucker P., York B., Hobbs L. M., Welty D. E., Friedman S. D., Dahlstrom J., Snow T. P., York D. G., 2018, ApJS, 237, 40
Spieler S. et al., 2017, ApJ, 846, 168

Strom K. M., Strom S. E., Carrasco L., Vrba F. J., 1975, ApJ, 196, 489
Thorburn J. A. et al., 2003, ApJ, 584, 339
Van der Maaten L., Hinton G., 2008, J. Mach. Learn. Res., 9, 11
van Loon J. T. et al., 2013, A&A, 550, A108
Vos D. A. I., Cox N. L. J., Kaper L., Spaans M., Ehrenfreund P., 2011, A&A, 533, A129
Vuong M. H., Foing B. H., 2000, A&A, 363, L5
Walker G. A. H., Bohlender D. A., Maier J. P., Campbell E. K., 2015, ApJ, 812, L8
Walker G. A. H., Campbell E. K., Maier J. P., Bohlender D., 2017, ApJ, 843, 56
Walker G. A. H., Campbell E. K., Maier J. P., Bohlender D., Malo L., 2016, ApJ, 831, 130
Walker G. A. H., Webster A. S., Bohlender D. A., Krełowski J., 2001, ApJ, 561, 272
Wampler E. J., 1966, ApJ, 144, 921
Ward J. H. J., 1963, J. Am. Stat. Assoc., 58, 236
Webster A., 1993, MNRAS, 262, 831
Welty D. E., Crowther P. A., 2010, MNRAS, 404, 1321
Welty D. E., Federman S. R., Gredel R., Thorburn J. A., Lambert D. L., 2006, ApJS, 165, 138
Welty D. E., Hobbs L. M., 2001, ApJS, 133, 345
Welty D. E., Ritchey A. M., Dahlstrom J. A., York D. G., 2014, ApJ, 792, 106
Wenger M. et al., 2000, A&AS, 143, 9
Westerlund B. E., Krelowski J., 1989, A&A, 218, 216
Wszołek B., Godłowski W., 2003, MNRAS, 338, 990
Zhen J., Castellanos P., Paardekooper D. M., Linnartz H., Tielens A. G. G. M., 2014, ApJ, 797, L30
Zuo W., Li A., Zhao G., 2021, ApJS, 252, 22

APPENDIX A: SIGHTLINE INFORMATION

The following table contains basic information about the target stars and sightlines, along with their low-reddening standard stars used in the DIB measuring process. A full version of this table can be found in Fan et al. (2019).

Table A1. Information on the 25 target sightlines.

Star name (HD/BD)	Identifier	Spectral type	$E_{\rm B\text{-}V}$ (mag)	f_{H2} [a]	Std. star (HD)	Std. star Spectral Type	Std. star $E_{\rm B\text{-}V}$ (mag)	Remarks
20041		A0Ia	0.72	0.42*	46300	A0Ib	0.01	
BD $+40^\circ4220$	Cernis 52	A3V	0.90	>0.78*	107966	A3V	0.00	Report of PAH (González Hernández et al. 2009)
23180	omi Per	B1III + B2V	0.31	0.55	44743	B1II-III	0.02	Steep ext. curve; Broad 2175 Åbump.
281159		B5V	0.85	0.50*	16219	B5V	0.04	
23512		A0V	0.36	0.62*	31647	A1V	0.01	Steep ext. curve; Broad 2175 Åbump.
24534	X Per	O9.5pe	0.59	0.76	214680	O9V	0.11	Translucent cloud; Steep ext.curve; Broad 2175 Åbump.
24912	xi Per	O7e	0.33	0.38	47839	O7Ve	0.07	
28482		B8III	0.52	0.66*	4382	B8III	0.01	Steep ext. curve; Broad 2175 Åbump.
37061	NU Ori	B1V	0.52	0.02*	36959	B1V	0.03	Intense radiation field; Flat ext. curve; Weak 2175 Åbump.
37903		B1.5V	0.35	0.53	37018	B1V	0.07	Anamalous 5780/5797 ratio; Flat ext. curve; Weak 2175 Åbump.
43384	9 Gem	B3Ib	0.58	0.44*	52089	B2II	0.01	
147084	omi Sco	A5II	0.73	0.59*	186377	A5III	0.04	
147889		B2V	1.07	0.45	42690	B2V	0.04	Embedded and ionizing nearby cloud (Rawlings et al. 2013); Steep ext. curve.
148579		B9V	0.34	0.45*	201433	B9V	0.00	Flat ext. curve; Weak 2175 Åbump.
166734		O8e	1.39	0.39*	47839	O7Ve	0.07	
168625		B8Ia	1.48	0.33*	34085	B8Iae	0.00	
175156		B5II	0.31	0.31*	34503	B5III	0.05	Steep ext. curve; Weak 2175 Åbump.
183143		B7Iae	1.27	0.31*	63975	B8II	0.00	
190603		B1.5Iae	0.72	0.16	52089	B2II	0.01	Flat ext. curve; Weak 2175 Åbump.
194279		B2Iae	1.20	0.30*	53138	B3Iab	0.05	Multiple components but average condition (Cox et al. 2011)
BD $+40^\circ4220$	VI Cyg 5	O7f	1.99	0.47*	47839	O7Ve	0.07	
	VI Cyg 12	B5Ie	1.11	0.67*	36959	B1V	0.03	Schulte's Star.
204827		O9.5V + B0.5III	1.11	0.67*	36959	B1V	0.03	Steep ext. Curve; Weak 2175 Åbump.
206267		O6f	0.53	0.42	47839	O7Ve	0.07	
223385	6Cas	A3Iae	0.67	0.12*	197345	A2Ia	0.09	

[a]Values marked by asterisks are surrogate result, i.e. N(H) estimated from W(5780) and/or $N(H_2)$ estimated from N(CH). See Fan et al. (2017) for details.

APPENDIX B: PAGEWIDE PLOTS

This appendix contains pagewide figures to include more details and information than their counterparts in the main text. Fig. B1 demonstrates $r_{\rm reg}$ (upper triangle) and $r_{\rm norm}$ (lower triangle) values among the 54 target DIBs. The DIBs are sorted by the sequence proposed in Section 5. As in the MDS result, the transitions in the r values (colours) are smooth across the DIBs, except around the grey lines separating the C_2 DIBs from the non-C_2 DIBs. This gap is more obvious in the $r_{\rm reg}$ values. Fig. B2 is the enlarged version of Fig. 5, except all data points are labelled. Note that the colours represent the clustering results (Section 4) and DIBs assigned to different groups by HAC and k-means clustering algorithms are plotted in two colours. Detailed discussions on this plot can be found in Sections 5 and 6.3.

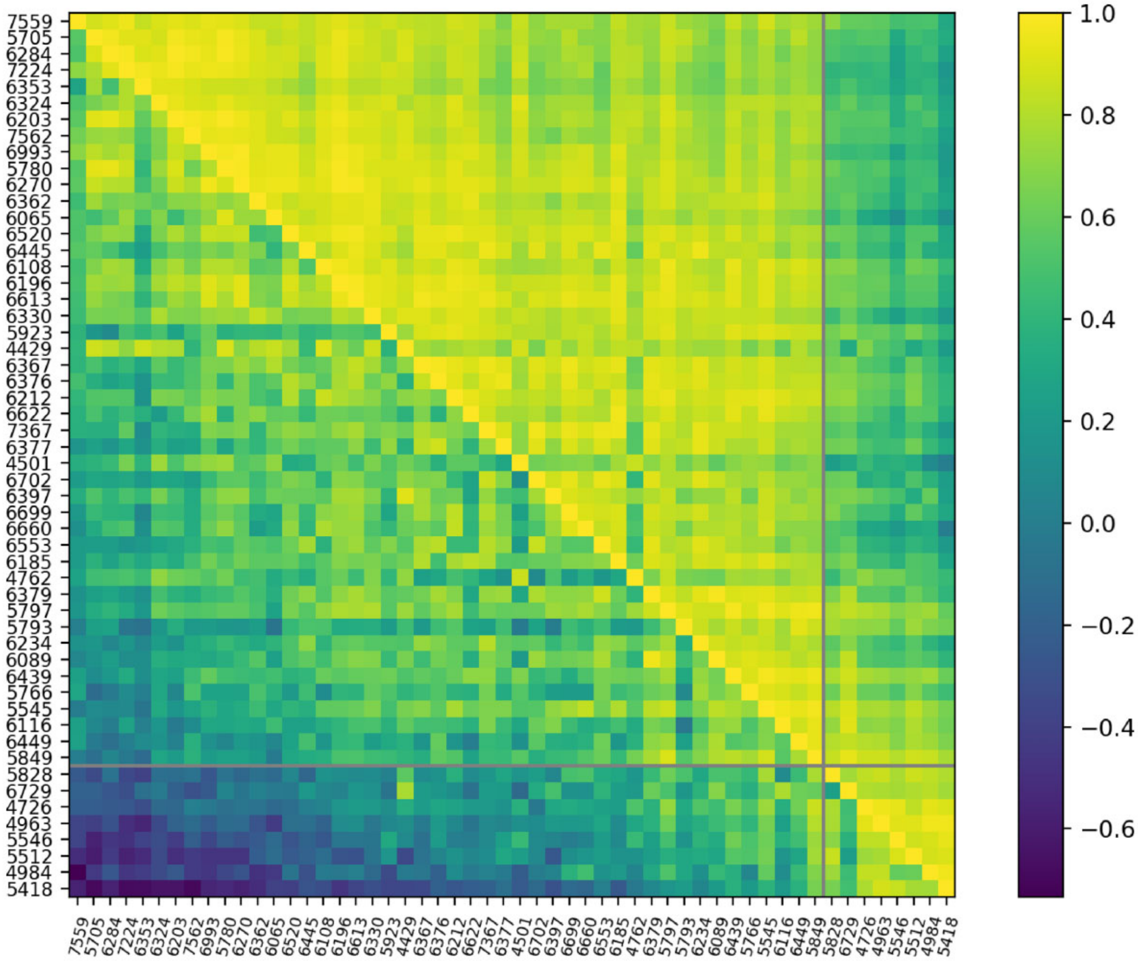

Figure B1. Heat map for the Pearson correlation coefficients (r values) among all 54 target DIBs. The figure is asymmetrical, where the lower triangle is for the $r_{\rm norm}$ values and the upper triangle is for the $r_{\rm reg}$ values. The DIBs are sorted by the sequence in Section 5, and the grey lines separate the C_2 DIBs from the non-C_2 DIBs. The transition in the colours and hence r values are rather smooth, except around the grey lines, which is more obvious in the $r_{\rm reg}$ section.

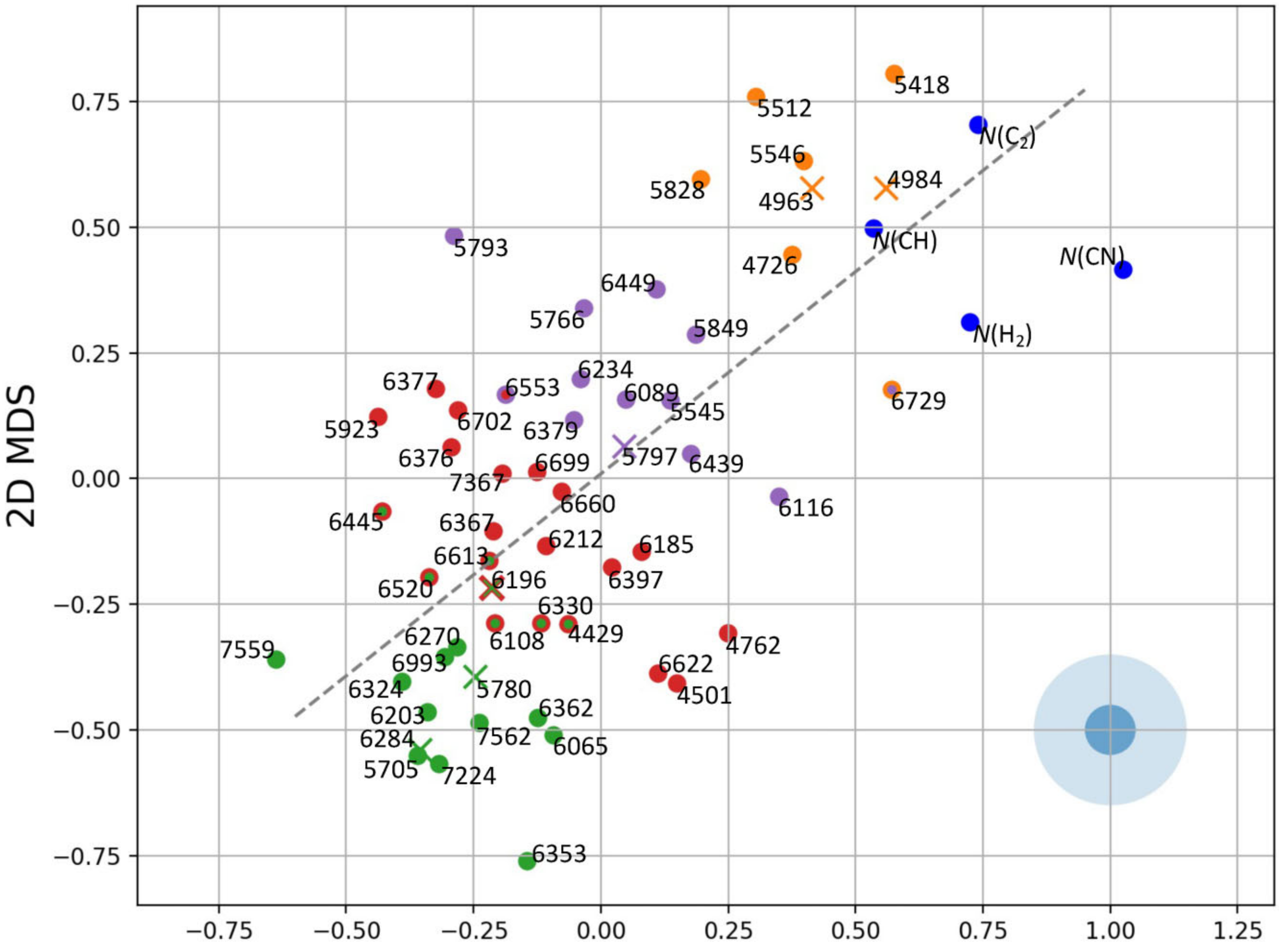

Figure B2. Result of 2D-MDS where all data points are labelled. Colours reflect the clustering result as in Fig. 3, where green is for the σ-type DIBs, red is for the intermediate DIBs, purple is for the ζ-type DIBs, orange is for the C_2 DIBs, and blue for the column densities of molecular species included for comparison. The 10 DIBs assigned to different groups by HAC and k-means algorithms are plotted with different colours, where the face (inner) colour represents the HAC result and the edge (outer) colour represents the k-means result. DIBs λλ6284, 5780, 6196, 5797, 4963, and 4984 are highlighted as 'X' for their known sequence. The transparent blue circles indicate 0.05 and 0.15 radius on the plot, and the dashed line is the best-fitting line of all DIB points. It follows the sequence of DIBs and acts as the first projection axis. We find a smooth and continuous distribution among the non-C_2 DIBs, whereas the C_2 DIBs seem to form a separate cluster with other small molecules.

APPENDIX C: ANALYSIS USING r_{REG} VALUES

We focus on the r_{norm} values in the main text since this work is originally motivated by the search of anticorrelated DIBs. But in principle, the clustering and MDS analyses can be applied to the r_{reg} matrix as well. Here, we follow the same routine as in Sections 4 and 5 to show that using r_{reg} would not change most of our findings.

We start with sorting the target DIBs into four clusters using the r_{reg} value matrix. Fig. C1 shows the dendrogram from HAC, and the difference between C_2 and non-C_2 DIBs is more highlighted. The assignments among the three non-C_2 groups are somewhat shuffled. For example, the intermediate group (red) now contains about half of the non-C_2 DIBs and includes λ5780, which is traditionally recognized as a σ-type DIB. In Table C1, we summarize the clustering results from different methods and inputs (i.e. the HAC or k-means clustering algorithms, r_{norm} or r_{reg} value matrix). The DIBs are ordered by the sequence from Section 5. For all combinations of method and input data, the assignments of DIBs generally follow the flow as σ-type, intermediate, ζ-type, and C_2 DIBs, and the boundary between C_2 and non-C_2 DIBs is quite solid. We find 34 out of 54 target DIBs being assigned to the same group throughout different methods and inputs, and the rest are assigned to two groups. There is no DIB assigned to three different groups in our analysis.

Fig. C2 presents the 2D-MDS results from the $1 - r_{\mathrm{reg}}$ matrix. Compared to Fig. 5, the C_2 DIBs here are still clustered with the molecular species and more separated from the non-C_2 DIBs. The continuous distribution among non-C_2 DIBs remains valid, and there is no clear boundary to further divide them into smaller clusters. On the other hand, the overall trend bends into an arch. It highlights the necessity of including a second projection axis, and this arched distribution might be related to the non-linear effect in correlation coefficients.

To sum up, analyses based on the r_{reg} values would provide a similar picture on DIB correlations and strengthens the discussions we made in the text, especially regarding the following:

(i) The C_2 DIBs are a unique DIB family that are more clustered with molecular species and are well separated from the non-C_2 DIBs.

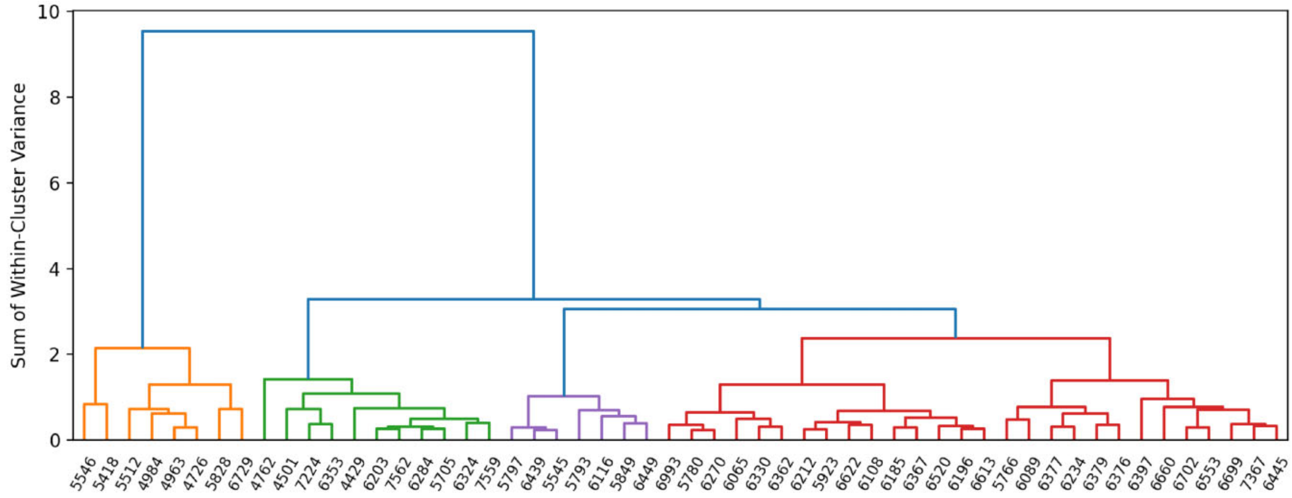

Figure C1. Dendrogram using the $r_{\rm reg}$ values. DIBs connected by a lower horizontal bar (node) share more similarities. We note that the difference between C_2 and non-C_2 DIBs is more obvious, and the assignments among the non-C_2 DIBs can be different compared to Fig. 3.

Table C1. Results of different clustering algorithms and inputs.

Idx[a]	DIB	$r_{\rm norm}$		$r_{\rm reg}$		Idx[a]	DIB	$r_{\rm norm}$		$r_{\rm reg}$	
		HAC	k-means	HAC	k-means			HAC	k-means	HAC	k-means
1	7559	σ	σ	σ	σ	28	4501	Inter.	Inter.	σ	σ
2	5705	σ	σ	σ	σ	29	6702	Inter.	Inter.	Inter.	Inter.
3	6284	σ	σ	σ	σ	30	6397	Inter.	Inter.	Inter.	Inter.
4	7224	σ	σ	σ	σ	31	6699	Inter.	Inter.	Inter.	Inter.
5	6353	σ	σ	σ	σ	32	6660	Inter.	Inter.	Inter.	Inter.
6	6324	σ	σ	σ	σ	33	6553	Inter.	ζ	Inter.	Inter.
7	6203	σ	σ	σ	σ	34	6185	Inter.	Inter.	Inter.	Inter.
8	7562	σ	σ	σ	σ	35	4762	Inter.	Inter.	σ	σ
9	6993	σ	σ	Inter.	σ	36	6379	ζ	ζ	Inter.	Inter.
10	5780	σ	σ	Inter.	σ	37	5797	ζ	ζ	ζ	ζ
11	6270	σ	σ	Inter.	Inter.	38	5793	ζ	ζ	ζ	ζ
12	6362	σ	σ	Inter.	Inter.	39	6234	ζ	ζ	Inter.	Inter.
13	6065	σ	σ	Inter.	Inter.	40	6089	ζ	ζ	Inter.	ζ
14	6520	σ	Inter.	Inter.	Inter.	41	6439	ζ	ζ	ζ	ζ
15	6445	σ	Inter.	Inter.	Inter.	42	5766	ζ	ζ	Inter.	ζ
16	6108	σ	Inter.	Inter.	Inter.	43	5545	ζ	ζ	ζ	ζ
17	6196	σ	Inter.	Inter.	Inter.	44	6116	ζ	ζ	ζ	ζ
18	6613	σ	Inter.	Inter.	Inter.	45	6449	ζ	ζ	ζ	ζ
19	6330	σ	Inter.	Inter.	Inter.	46	5849	ζ	ζ	ζ	ζ
20	5923	Inter.	Inter.	Inter.	Inter.	47	5828	C_2	C_2	C_2	C_2
21	4429	σ	Inter.	σ	σ	48	6729	ζ	C_2	C_2	C_2
22	6367	Inter.	Inter.	Inter.	Inter.	49	4726	C_2	C_2	C_2	C_2
23	6376	Inter.	Inter.	Inter.	Inter.	50	4963	C_2	C_2	C_2	C_2
24	6212	Inter.	Inter.	Inter.	Inter.	51	5546	C_2	C_2	C_2	C_2
25	6622	Inter.	Inter.	Inter.	Inter.	52	5512	C_2	C_2	C_2	C_2
26	7367	Inter.	Inter.	Inter.	Inter.	53	4984	C_2	C_2	C_2	C_2
27	6377	Inter.	Inter.	Inter.	Inter.	54	5418	C_2	C_2	C_2	C_2

[a]DIBs are sorted according to the sequence in Section 5.

(ii) There is a rather continuous transition among the non-C_2 DIBs, although some of the members may demonstrate quite different behaviour. This continuous trend is against further sorting all non-C_2 DIBs into clearly distinguishable groups, and terms like σ- or ζ-type DIBs may be applied only to the most representing members.

(iii) At least two factors are needed to properly reproduce the DIB correlation.

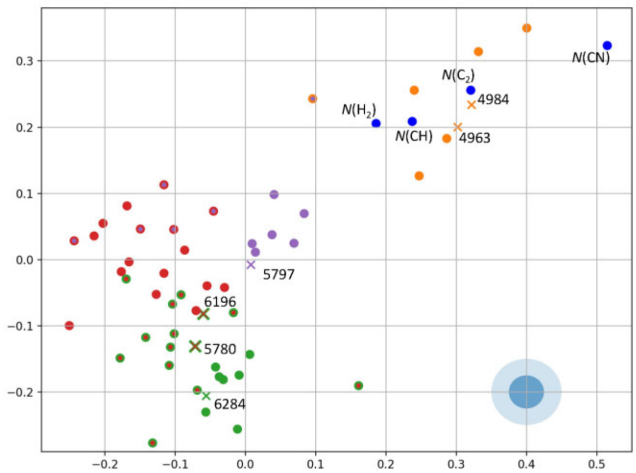

Figure C2. 2D-MDS result where distance between points represents $1 - r_{\rm reg}$ dissimilarity. Colours represent memberships of group, i.e. green for the σ-type, red for intermediate, purple for the ζ-type, orange for the C_2 DIBs, and blue for known molecular species. DIBs assigned to two groups in Table C1 have distinct face and edge colours. DIBs $\lambda\lambda$6284, 5780, 6196, 5797, 4963, and 4984 are highlighted as 'X' for their known sequence. The transparent blue circles at the lower right corner indicate 0.05 and 0.1 radius in the plot. The C_2 DIBs remain clustered with molecular species and are more separated from the non-C_2 DIBs compared to Fig. 5. We find that the overall trend bends into an arch, which might be due to the non-linear effects in correlation coefficient.

This paper has been typeset from a TEX/LATEX file prepared by the author.