



**HAL**  
open science

## Searching for carriers of the diffuse interstellar bands across disciplines, using Natural Language Processing

Corentin van den Broek d'Obrenan, Frédéric Galliano, Jeremy Minton, Viktor Botev, Ronin Wu

### ► To cite this version:

Corentin van den Broek d'Obrenan, Frédéric Galliano, Jeremy Minton, Viktor Botev, Ronin Wu. Searching for carriers of the diffuse interstellar bands across disciplines, using Natural Language Processing. *Journal of Interdisciplinary Methodologies and Issues in Science*, In press, 10.18713/JIMIS-ddmmyy-v-a . hal-03651314v1

**HAL Id: hal-03651314**

**<https://hal.science/hal-03651314v1>**

Submitted on 25 Apr 2022 (v1), last revised 1 Aug 2023 (v3)

**HAL** is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.



## Searching for Carriers of the Diffuse Interstellar Bands Across Disciplines, using Natural Language Processing

Corentin VAN DEN BROEK D'OBRENAN<sup>1,2,3</sup>, Frédéric GALLIANO<sup>1,2</sup>, Jeremy MINTON<sup>3</sup>,  
Viktor BOTEV<sup>3</sup>, Ronin WU<sup>\*3</sup>

<sup>1</sup>Université Paris-Saclay, Université Paris-Cité, CEA, CNRS, France

<sup>2</sup>Astrophysique, Instrumentation, Modélisation Paris-Saclay, 91191 Gif-sur-Yvette, France

<sup>3</sup>Iris AI, Bekkestua, Norway

\*Corresponding author: [ronin@iris.ai](mailto:ronin@iris.ai)

DOI: [10.18713/JIMIS-ddmmyy-v-a](https://doi.org/10.18713/JIMIS-ddmmyy-v-a)

Submitted: Jour Mois-en-lettres Année - Published: Jour Mois-en-lettres Année

Volume: N - Year: AAAA

Issue: thème interdisciplinaire

Editors: Prénom-1 Nom-1, Prénom-2 Nom-2, Prénom-k Nom-k...

### Abstract

The explosion of scientific publications overloads researchers with information. This is even more dramatic for interdisciplinary studies, where several fields need to be explored. A tool to help researchers overcome this is *Natural Language Processing* (NLP): a machine-learning (ML) technique that allows scientists to automatically synthesize information from many articles. As a practical example, we have used NLP to conduct an interdisciplinary search for compounds that could be carriers for *Diffuse Interstellar Bands* (DIBs), a long-standing open question in astrophysics. We have trained a NLP model on a corpus of 1.5 million cross-domain articles in open access, and fine-tuned this model with a corpus of astrophysical publications about DIBs. Our analysis points us toward several molecules, studied primarily in biology, having transitions at the wavelengths of several DIBs and composed of abundant interstellar atoms. Several of these molecules contain chromophores, small molecular groups responsible for the molecule's colour, could be promising candidate carriers. Identifying viable carriers demonstrates the value of using NLP to tackle open scientific questions, in an interdisciplinary manner.

### Keywords

Machine-learning; natural language processing; astrophysics; interstellar medium; diffuse interstellar bands

## I INTRODUCTION

A prerequisite to interdisciplinarity is the ability of researchers in a given field to explore the literature of other fields and easily extract relevant information. In particular, finding similar

25 concepts in two different fields, adapting methods from one field to another, or re-purposing  
26 data acquired in an unrelated field are all potentially fruitful approaches for scientists to make  
27 discoveries. It is not unlikely that clues to various open questions in the global scientific lit-  
28 erature currently exist, but looking for them is like searching for a needle in the proverbial  
29 haystack. However, as the scientific knowledge expands exponentially (Densen, 2011), human  
30 ability to keep up with both sorting and navigating diminishes quickly. It is humanly impossi-  
31 ble to read all published articles, and keyword searches that do not include contextual semantic  
32 meaning or conceptual reasoning are extremely limited. Fortunately, tremendous progress has  
33 recently been made in the automatic analysis of written documents. *Natural Language Process-*  
34 *ing* (NLP) is a branch of linguistics that leverages the statistical properties of a given corpus with  
35 machine learning (ML) methods to explore the semantic relationships between texts (Manning  
36 and Schütze, 1999). It is being used to extract information, to draw parallels between problems  
37 and to formulate new research directions. It is aimed at solving information overload. Recently,  
38 ML techniques have been used to generate scientific hypotheses in many scientific domains,  
39 such as the drug repositioning / discovery research (Hastings *et al.*, 2012; Lamurias *et al.*,  
40 2019). Going beyond the boundary of disciplines, we pioneer the use of ML techniques for  
41 hypothesis generation from cross-domain literature. It is a practical method, quickly becoming  
42 popular, that will have an important role in interdisciplinary studies in the coming years.

43 The present article discusses the results of a collaboration between astrophysicists and com-  
44 putational linguists. NLP techniques have already been applied to astrophysics to outline re-  
45 search priorities (Thomas *et al.*, 2022) and to thoroughly search the literature (Kerzendorf,  
46 2019; Grezes *et al.*, 2021). In these recent studies, NLP is used as an exploratory tool that can  
47 be used to help refine a project. Here, we take a step further and use NLP as a primary research  
48 tool, with the hope of making actual discoveries. We have applied it to the long-lasting question  
49 of the origin of the *Diffuse Interstellar Bands* (DIBs). DIBs are ubiquitous spectral absorption  
50 features observed at visible wavelengths, along the sightline to stars in the Milky Way (Hobbs  
51 *et al.*, 2009). They were discovered a century ago, but the chemical composition of their carriers  
52 is still unknown. Thus, we have trained a NLP model on a cross-domain corpus, which includes  
53 astrophysical publications about DIBs, then explored other fields, such as biochemistry, where  
54 relevant molecules could have been studied. We start this article by presenting the astrophysical  
55 context of our study, in Section II. In Section III, we review the NLP techniques used in this  
56 study. We then discuss how we have applied NLP to address the origin of DIBs, in Section IV.  
57 The astrophysical relevance of the molecules we have found is assessed in Section V and our  
58 results are summarized in Section VI.

## 59 II DIFFUSE INTERSTELLAR BANDS

60 The object of our study, the *Diffuse Interstellar Bands* (DIBs), are spectroscopic absorption  
61 features that are ubiquitously observed in the Milky Way, our own galaxy (Hobbs *et al.*, 2009;  
62 Jones, 2016b, for reviews). They appear as a forest of bands along the line of sight towards  
63 stars, in the visible electromagnetic spectrum (wavelengths,  $\lambda \simeq 0.4 - 0.8 \mu\text{m}$ ), sometimes  
64 extending up to the near-infrared domain (up to  $\lambda = 2 \mu\text{m}$ ). Figure 1 shows a synthetic DIB  
65 absorption spectrum. They originate from the *InterStellar Medium* (ISM).

### 66 2.1 The Interstellar Medium

67 Since DIBs are interstellar features, we first need to review the general properties of the ISM.  
68 Broadly speaking, the ISM is constituted of all the matter filling the volume of a galaxy between  
69 the stars. This matter is essentially gaseous, but about half a percent of its mass is made of small

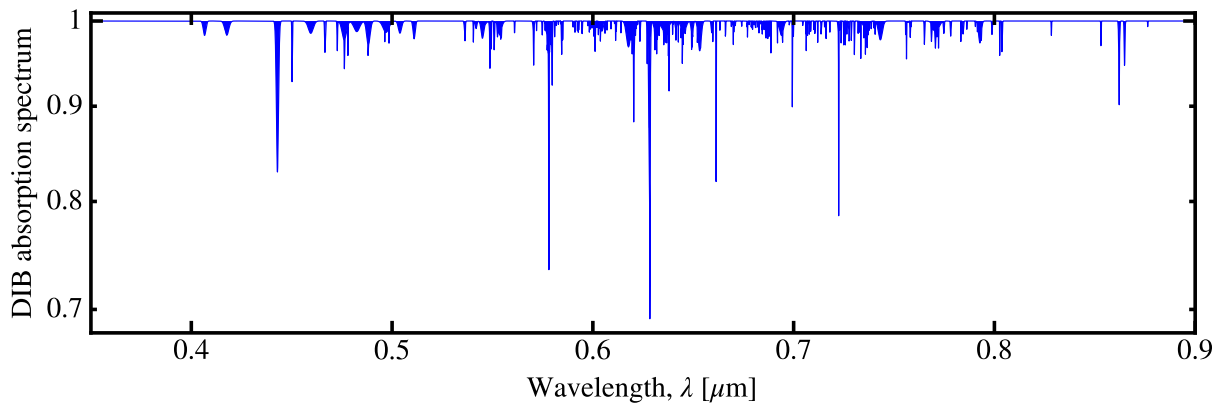


Figure 1: Synthetic absorption spectrum of the most prominent DIBs. The parameters of each feature (position, width and depth) come from the observational compilation by Jenniskens and Desert (1994).

70 solid particles, the dust grains (Tielens, 2005; Draine, 2011, for textbooks). The elemental  
 71 composition of the ISM is 74 % hydrogen, 25 % helium, and the remaining 1 % contains all the  
 72 heavier elements (Asplund *et al.*, 2009). Figure 2 shows the abundances, in the Solar system,  
 73 of the first elements in the periodic table. We see that besides hydrogen and helium, the two  
 74 most abundant species are carbon and oxygen. This puts some constraint on the most abundant  
 molecules that can be formed, and this will be important for our serendipitous search.

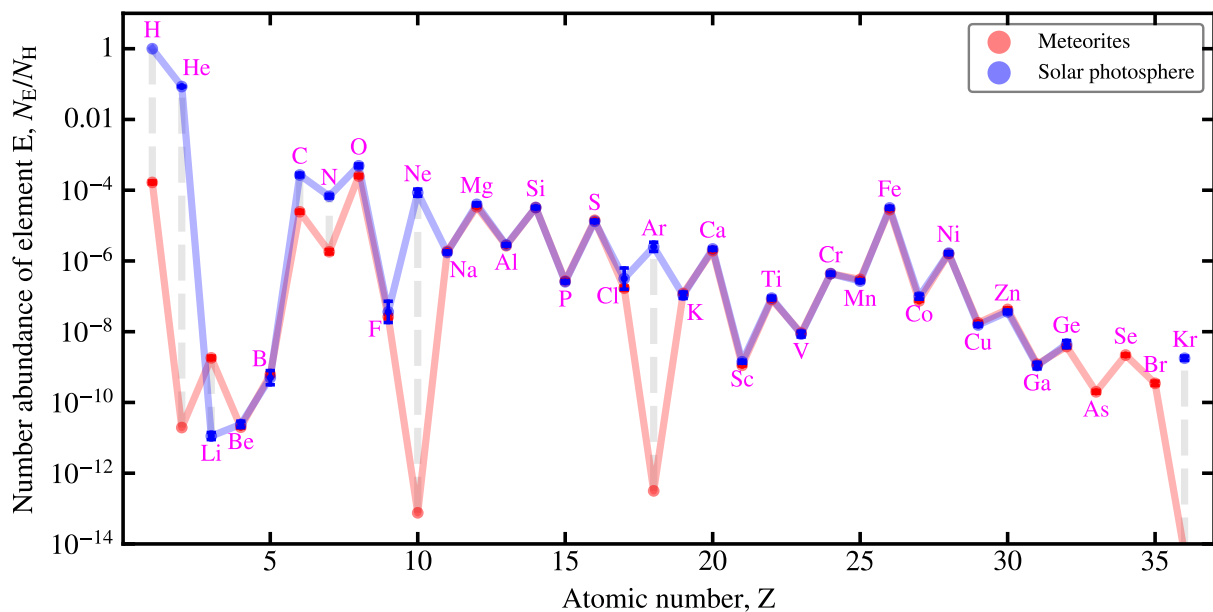


Figure 2: Elemental abundances in the Solar system, from the data in Asplund *et al.* (2009). Those are *number* abundances (*i.e.* number of atoms relative to hydrogen). The blue dots represent measures made in the Solar photosphere, through absorption spectroscopy, and the red dots correspond to the chemical composition of chondrite meteorites. Both are in good agreement, except for the lightest, most volatile elements, which do not remain trapped in meteorites. These abundances are representative of the ISM in our Solar neighborhood, and are even used as a reference when studying other galaxies.

75

### 76 2.1.1 The Phases of the ISM

77 The ISM is a highly heterogeneous medium (*e.g.* Chap. III.3 of Galliano, 2022). Half of the  
 78 volume of our galaxy is filled with a permeating hot ionized gas with a very low density (tem-  
 79 perature,  $T = 10^6$  K; density,  $n \simeq 3 \times 10^{-3} \text{ cm}^{-3}$ ). This phase is heated by the shock waves

80 from supernova explosions. The rest of the volume covers a wide range of density, temperature  
81 and atomic state. The coldest and densest interstellar regions are called molecular clouds. As  
82 their name indicates, the elements in these clouds have combined to form molecules, majoritar-  
83 ily H<sub>2</sub> and CO. Their high density (up to  $n \simeq 10^6 \text{ cm}^{-3}$ ) allow them to be shielded from the  
84 stellar radiation and to reach low temperatures ( $T \simeq 10 \text{ K}$ ). Between these two extremes, there  
85 are nine orders of magnitude in density and five in temperature, unlike anything we can find on  
86 Earth. DIBs are found preferentially in the diffuse ISM, and disappear in dense regions (*e.g.*  
87 [Lan et al., 2015](#)), although some specific DIBs are found in diffuse molecular clouds (density  
88  $n \simeq 10^2 \text{ cm}^{-3}$ ; [Thorburn et al., 2003](#)).

### 89 2.1.2 Interstellar Molecules

90 As we will see in Section 2.2, DIB carriers are likely large molecules. Until now, more than  
91 200 individual molecules have been identified in space ([McGuire, 2018](#)). The first one to be  
92 discovered was CH, in the 1930s ([Herzberg, 1988](#), for an historical review). Although this first  
93 observation was performed in the visible domain, the majority of the subsequent detections were  
94 achieved at radio wavelengths, through the various rotational lines<sup>1</sup> of the molecules. Several  
95 new molecules containing more than six atoms are now discovered each year. These large  
96 molecules all contain carbon atoms. They are thus called *Complex Organic Molecules* (COMs;  
97 [Herbst and van Dishoeck, 2009](#), for a review). Even branched molecules, which were believed  
98 to be too brittle to survive the harsh interstellar environment, have been detected ([Belloche](#)  
99 *et al.*, 2014).

100 An important class of organic molecules is *Polycyclic Aromatic Hydrocarbons* (PAHs). They are  
101 constituted of several *aromatic cycles*, which are hexagonal structures made of carbon atoms,  
102 such as in benzene, with peripheral hydrogen atoms (Figure 3). This species was introduced,  
103 in astrophysics, to provide an interpretation for the bright mid-infrared features (in the spectral  
104 range  $\lambda = 3 - 20 \mu\text{m}$ ) observed ubiquitously in the ISM and in galaxies (Figure 4). The vibra-  
105 tional modes of the C–C and C–H bonds in PAHs, which have similar resonance frequencies  
106 across the PAH family, indeed provide a good account for these mid-infrared bands ([Tielens,](#)  
107 [2008](#), for a review). This however forbids the identification of individual PAHs, as these broad  
108 mid-infrared bands arise from the mixture of several molecules. Only a few individual PAHs  
109 have unambiguously been detected, either because they have very peculiar features due to an  
110 atypical structure, such as fullerenes (Figure 3; [Cami et al., 2010](#)), or, recently, through their  
111 rotational lines ([McGuire et al., 2021](#)).

### 112 2.1.3 Interstellar Dust Grains

113 When the number of atoms in a molecule becomes large, its resonance features become broader  
114 and a noticeable continuum arises (*e.g.* Chap. 1 of [Galliano, 2022](#)). This is because, with a large  
115 number of atoms, we are entering the solid-state realm. Large molecules are at the interface with  
116 dust grains, and there is probably a continuity between both in the ISM, although there is no  
117 well-defined limit between these two categories.

118 Interstellar dust grains are small solid particles with radii ranging from  $\simeq 3 \text{ \AA}$  to 300 nm (*e.g.*  
119 [Draine, 2003](#), for a review). They account for about half of the mass of heavy elements in the  
120 ISM, thus about half a percent of its total mass. Yet, these grains, which are predominantly  
121 silicate and carbonaceous compounds, have a very important role. In a quiescent galaxy, such

---

<sup>1</sup>Rotational lines arise from transitions between different values of the quantum angular momentum of the molecule.

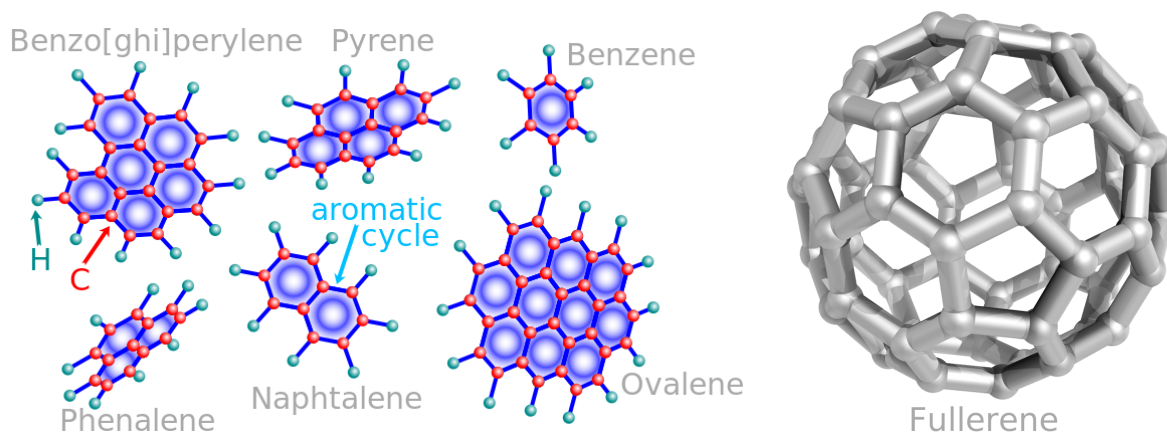


Figure 3: The PAH molecular family. On the left side, we show six different PAHs. Hydrogen atoms are represented in cyan, and carbon atoms, in red. On the right side, we show the buckminsterfullerene, which is a spherical molecule, composed of 60 carbon atoms. Credit: the left figure is adapted from Galliano (2022); the fullerene image is from Yassine Mrabet, licensed under CC BY 3.0.

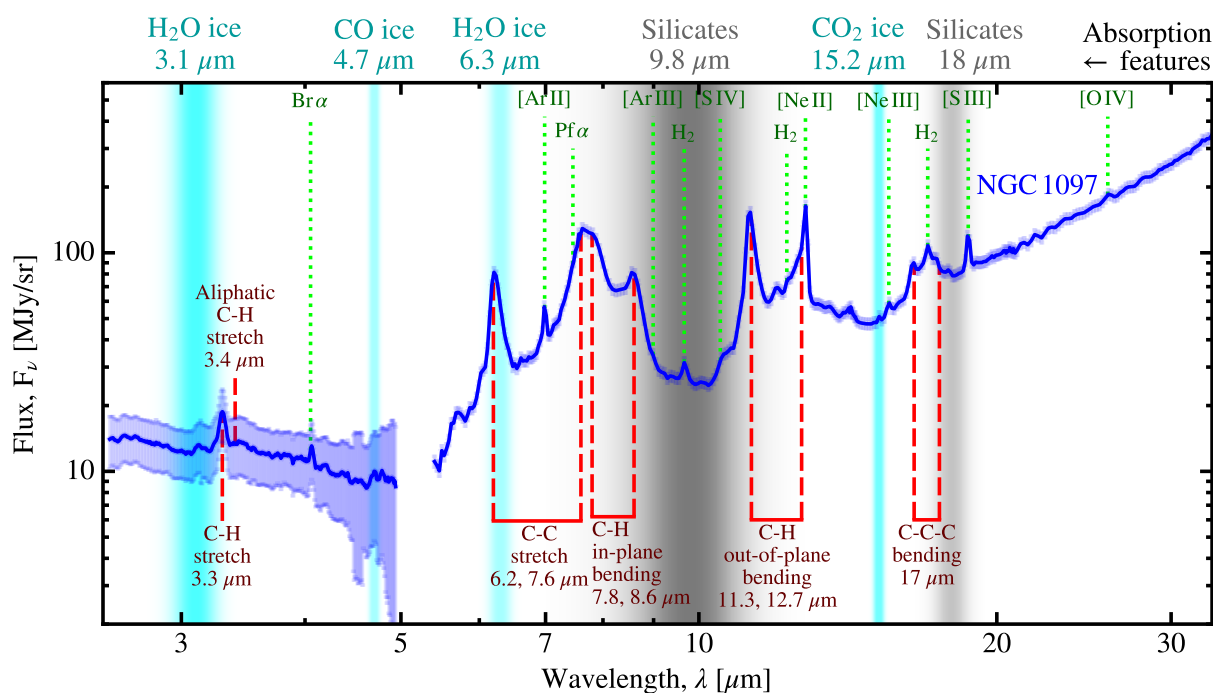


Figure 4: Mid-infrared spectrum of the galaxy NGC 1097. The blue line, with error bars, is the observed spectrum of the central region of this object. We have highlighted the main spectral features: on top, we have indicated the main silicate and ice interstellar absorption bands; in green, we have shown the position of the brightest gas lines; in red, we have pointed towards the brightest PAH features and have noted which vibrational C–C or C–H mode they correspond to. Credit: figure adapted from Galliano *et al.* (2018).

122 as the Milky Way, they absorb about 25 % of the stellar power, in the ultraviolet (UV) and  
123 visible range, and re-emit it thermally in the infrared (e.g. Bianchi *et al.*, 2018). In regions of  
124 massive star formation, this fraction can go up to 99 %. These regions are thus totally opaque  
125 to visible photons and can be studied only through their infrared radiation, emitted by the dust.  
126 Grains are also the catalysts of several chemical reactions, including the formation of H<sub>2</sub>, the  
127 most abundant molecule in the Universe (e.g. Bron, 2014).

128 When looking at a star in the Milky Way, the dust present along the line of sight extincts its  
129 radiation. It means that a fraction of the stellar light is either absorbed by grains or scattered  
130 in another direction. Dust grains are so well-mixed with the gas in the ISM that the extinction  
131 amplitude (usually quoted in the visual band, V, at  $\lambda = 0.55 \mu\text{m}$ ) is considered as a reliable  
132 tracer of interstellar matter.

#### 133 2.1.4 *The Lifecycle of a Galaxy*

134 Finally, it is important to understand that the ISM is a dynamical environment, constantly evol-  
135 ving. It is indeed the fuel of star formation. Stars form by the gravitational collapse of molecular  
136 clouds. Once ignited, the most massive stars, which are also the brightest, blow their interstellar  
137 cocoon away and ionize their surroundings. At the end of their lifetime, stars eject in the ISM  
138 fresh heavy elements that they have formed in their core by nucleosynthesis. The more a galaxy  
139 is evolved, the more tenuous and rich in heavy elements its ISM is.

### 140 2.2 The Uncertain Nature of DIBs

141 The first DIBs were discovered exactly one century ago by Heger (1922), but the physical nature  
142 of their carriers largely remains a mystery, today. Their interstellar origin was demonstrated  
143 by Merrill (1934). Their intensity is indeed correlated with the dust extinction amplitude and  
144 independent of the intrinsic properties of the background stars. Over 500 DIBs have been  
145 detected so far, in the ISM (Fan *et al.*, 2019). DIBs are also detected in external galaxies, such  
146 as the Magellanic clouds (Galliano *et al.*, 2018, for a review).

#### 147 2.2.1 *Constraints on the Nature of the DIB Carriers*

148 As absorption features, DIBs must originate from the transition of an atom or a molecule, be-  
149 tween two of its quantum energy levels. The intrinsic line width of DIBs, which is typically  
150  $\simeq 1 \text{ \AA}$ , excludes that they originate from free-flying atoms. They must come from molecules  
151 with a few to  $\simeq 100$  atoms (MacIsaac *et al.*, 2022). Their strength and their ubiquity also tells  
152 us that they have to be made with the most abundant atoms in the ISM, mainly H, O, C, N.  
153 A last constraint is that, due to their presence in the diffuse ISM, a medium permeated with  
154 UV photons, these molecules need to be rather compact, to be more resilient. The branched  
155 molecules we have mentioned in Sect. 2.1.2 are only found in dense molecular clouds, well  
156 protected from UV radiation.

#### 157 2.2.2 *Identification of Buckminsterfullerene*

158 The only molecule to date, unambiguously identified as a DIB carrier is the *buckminster-*  
159 *fullerene*<sup>2</sup> (Campbell *et al.*, 2015; Walker *et al.*, 2015). The cation of this molecule, C<sub>60</sub><sup>+</sup>, can  
160 account for two, possibly four DIBs, at near-infrared wavelengths. Notice that this molecule  
161 satisfies all the constraints we have listed in Sect. 2.2.1. The fact that this molecule had been  
162 detected beforehand, in a planetary nebula, *via* its mid-IR features (Cami *et al.*, 2010), makes

---

<sup>2</sup>Fullerenes constitute a family of compact closed-mesh carbon compounds. Buckminsterfullerene is the variety with formula C<sub>60</sub>.

163 this identification trustworthy. It is possible that other COMs, and probably other PAHs, are  
164 DIB carriers. There are however variations of the relative strengths of DIBs, across sightlines  
165 (Herbig, 1995). This indicates that DIB carriers likely come from a diversity of molecules  
166 whose relative abundance varies with the environment.

### 167 2.2.3 The Significance of DIB Identification

168 DIBs are the last large class of interstellar spectral features that are still unidentified. The fact  
169 that, a hundred years after their discovery, less than one percent of these features has been  
170 identified, shows that this is an arduous challenge. This is however an important question, as  
171 spectroscopy is historically what transformed astronomy into astrophysics (Hearnshaw, 2014).  
172 This is because spectroscopy allows us to identify atoms and molecules from a distance, mea-  
173 sure their charge state, temperature, density and abundance, that we could learn so much about  
174 the Universe. Identifying the carriers of the DIBs would therefore be a major breakthrough in  
175 our understanding of the ISM. It could help us unlock the complexity of interstellar chemistry  
176 and provide a wealth diagnostics of the physical conditions where these bands are observed.

177 Astrophysicists have been stuck by this problem, because its answer lies in the wide diversity  
178 of potential molecules. Only a handful of these molecules can be measured in the laboratory  
179 or computed theoretically. Teams working on these identifications have limited resources. Yet,  
180 since we have seen that DIB carriers are likely large organic molecules, it is possible that some  
181 of these molecules have been studied in other fields, such as biochemistry or biology. We have  
182 thus performed a serendipitous, interdisciplinary search for these molecules, using the technique  
183 of natural language processing.

## 184 III NATURAL LANGUAGE PROCESSING

185 NLP is a machine-learning technique applied to written documents so we will first review some  
186 relevant principles of machine-learning.

### 187 3.1 What is machine learning

188 *Machine-Learning* (ML) is a methodology that paves the way to artificial intelligence. It is a  
189 type of computer program, where the main algorithm is not directly coded by a computer sci-  
190 entist, but inferred from a large data set. This methodology enables us to extrapolate patterns  
191 from a large quantity of data to unseen data (Goodfellow *et al.*, 2016, for a text book). Gener-  
192 ally, modern ML algorithms are built on layered *Neural Network* (NN). A NN can be seen as  
193 an empirical decision tree where the different decisions are controlled by *parameters*. These  
194 parameters are represented as matrices that act on the encoded input data to make predictions  
195 about the data. These parameters are determined during a training process by correcting errors  
196 between the predictions and desired outputs. Once trained, the NN can be used on other unseen  
197 input data to predict new outputs.

### 198 3.2 Word-embeddings

199 To apply ML to texts, we need a way to numerically represent words and their meaning. This  
200 is achieved with *word-embeddings*. Word-embeddings are techniques to transform contex-  
201 tual representations of words into high-dimensional real-valued vectors (Bengio *et al.*, 2003;  
202 Mikolov *et al.*, 2013). In recent years, development of word-embeddings models has taken big  
203 steps forward from shallow NN architectures, such as the Word2Vec (Mikolov *et al.*, 2013),  
204 Glove (Pennington *et al.*, 2014), and FastText (Bojanowski *et al.*, 2017) models, to complex  
205 and data-hungry models, such as ELMo (Peters *et al.*, 2018) and BERT (Devlin *et al.*, 2019).



206 3.2.1 Concepts

207 Word-embeddings are underpinned by the distributional hypothesis: that words with similar distributions have similar meanings or "a word is characterized by the company it keeps" (Firth, 208 1957). Techniques leveraging this hypothesis numerically represent semantic properties of words and capture meaningful syntactic and semantic regularities in a given vector space. Regularities are often observed as constant vector offsets between pairs of words sharing a particular 210 relationship. For example, as demonstrated with the Word2Vec model in Mikolov *et al.* (2013), the male/female relationship is learned through such techniques, and with the induced vector 212 representations,  $\vec{king} - \vec{man} + \vec{woman}$  results in a vector very close to  $\vec{queen}$ . Such characteristics make word-embedding models an efficient and effective tool for identifying contextual synonyms, ranking keywords, and computing similarities between millions of documents (Botev 216 *et al.*, 2017).

218 3.2.2 The Word2Vec Model

219 In order to efficiently identify relevant publications in the cross-domain literature, we opt to use the Word2Vec Continuous Bag-of-Words (Word2Vec-CBOW) model, which is implemented 220 as part of the gensim library (Řehůřek and Sojka, 2010). The architecture of the Word2Vec model is a simple feedforward NN with a hidden projection layer. In the CBOW model, vectors 222 of words that appear in a fixed window around the target word are averaged into the projection layer as demonstrated in Figure 5.

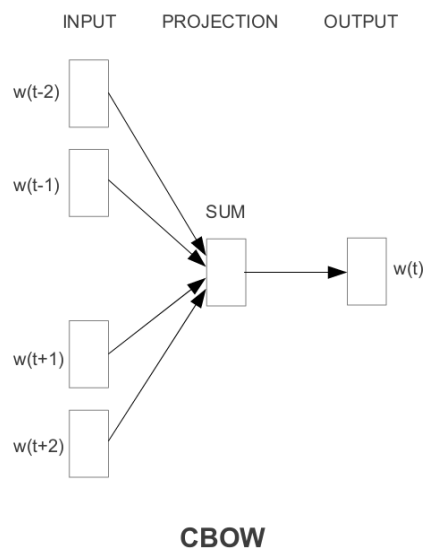


Figure 5: The architecture of the Word2Vec-CBOW model, where  $w(t)$  (on the right) represents the word in training by its co-occurred words,  $w(t - 2), w(t - 1), \dots$  (on the left), in the chosen window around it (Mikolov *et al.*, 2013).

224

225 In order to obtain a model trained on the cross-domain literature, we compile a generic corpus  
226 of 1.5 million open access English articles across all domains from the CORE service (Knoth  
227 and Zdrahal, 2012). Due to the resource limitation, in this pilot study, we focus on the abstracts,  
228 which are usually concisely written and already contains essential information of the articles.  
229 We then train a generic Word2Vec-CBOW model with 100 dimensions on the entire corpus ten  
230 times. In the following section, we explain how this model is used in our analysis.

### 231 **3.3 Extracting physical quantities from scientific documents**

232 As the main objective of this study is to discover possible carriers of DIBs in the literature  
233 outside of the astrophysics domain, recognition of physical quantities is an essential step in the  
234 corpus processing. A physical quantity in the corpus is defined as a numeric value that appears  
235 with a physical unit, such as “1500 Å” However, recognizing physical quantities from existing  
236 documents is not a trivial task. For example, MWC 349A, is a member of the double star system,  
237 MWC 349 (Gvaramadze, V. V. and Menten, K. M., 2012), and should not be recognized as a  
238 physical quantity of 349 Å. Furthermore, there are numerous ways to denote one physical  
239 quantity. For example, 1500 Å and 1.5 μm are equivalent and should be recognized as such by  
240 the model.

241 To identify all relevant wavelengths mentioned in the generic corpus, we implement a quantity  
242 recognition algorithm which includes three major components:

- 243 1. A regular expression (RegEx) algorithm that identifies all numerical values, such as  $28.4 \times$   
244  $10^3$ , 28,400 and  $(28.4 \pm 0.1) \times 10^3$ .
- 245 2. A unit-parsing algorithm that is based on the open source library Pint (version 0.19.1)<sup>3</sup>.
- 246 3. A unit-disambiguation algorithm that assigns probabilities to ambiguous units, such as  
247 angstrom (Å) and ampere (A), by comparing the unit predicted by the generic Word2Vec  
248 model.

249 We detail the unit-disambiguation algorithm in the following.

250 With the RegEx and unit-parsing algorithms listed above, we mask the identified physical quan-  
251 tities by numbers and units in the articles of the generic corpus. Physical quantities with am-  
252 biguous units are masked by all candidate units, for example, 1500 Å is masked as “NUM-  
253 Angstrom-Ampere”.

254 Using this model, we compute a context vector, which is the average vector from a window of 6  
255 words around the physical quantity excluding itself, for each physical quantity with an ambigu-  
256 ous unit. For each candidate unit assigned to the ambiguous unit, we then calculate the cosine  
257 similarity, which quantifies how parallel are two vectors, or in our case, how contextually-  
258 correlated are two words, between the calculated context vector to other units of the same  
259 dimension as the candidate unit. The highest cosine similarity is then assigned to the candidate  
260 unit as its score to represent the ambiguous unit.

## 261 **IV TACKLING DIBS WITH NLP**

262 We now discuss how we apply this NLP model to the question of DIBs. We start with the  
263 curation of the astrophysical literature corpus.

### 264 **4.1 Compiling the DIB corpus**

265 To improve accuracy of the semantic relationships between words directly associated with  
266 DIBs and common words used in other domains, a specialized corpus for DIBs is required.  
267 Our generic corpus of 1.5 million articles aims to cover cross-domain literature so will likely  
268 under-represent DIB articles making it inadequate for this purpose. To overcome this under-  
269 representation, we compile a corpus specialized in DIBs to enhance the representations of DIB  
270 words in the word-embeddings model.

---

<sup>3</sup><https://pint.readthedocs.io/en/stable/>

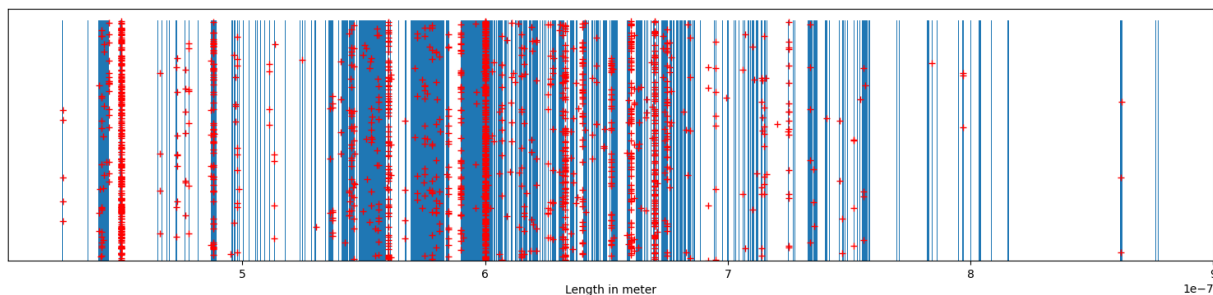


Figure 6: Indication of identified known DIBs (blue bands) and identified physical quantities from the generic corpus that overlap with DIBs (red crosses).

271 We first search for DIBs-related papers on NASA’s Astrophysics Data System<sup>4</sup> and look for  
 272 their relevant papers with Iris.ai’s Explore Tool<sup>5</sup>. We then examine all articles appearing in the  
 273 search and hand pick the articles that are directly DIBs-related. The final selected DIB corpus  
 274 then includes 939 articles.

275 Using our physical quantity recognition algorithm, we identify 284 wavelengths associated with  
 276 DIBs taken from the catalog of Jenniskens and Desert (1994)<sup>6</sup>. These wavelengths, along with  
 277 the full-width-half-maximum (FWHM), are marked as blue bands in Figure 6. The DIB corpus,  
 278 with the disambiguated units, is then used to fine-tune the generic Word2Vec-CBOW model  
 279 on this corpus for 10 times. This DIBs-enhanced Word2Vec-CBOW model is used for our  
 280 following analysis.

#### 281 4.2 Using the unit quantity recognition

282 We then apply the physical quantity recognition algorithm on the generic corpus. Out of the  
 283 1.5 million articles, we identify  $\sim 2000$  physical quantities from  $\sim 20,000$  articles, that overlap  
 284 with DIBs and mark their locations as red crosses in Figure 6. Some of these physical quantities  
 285 describe molecular sizes, which are unrelated to DIBs, but others describe the wavelength of  
 286 the energy transitions and are thus of direct interest to us.

287 In order to systematically filter the identified physical quantities down to the most relevant to  
 288 DIBs, we use three filters to select candidate articles.

- 289 1. Physical quantities in the range  $0.1 - 1 \mu m$  in the generic corpus are often found to de-  
 290 scribe diameters or distances. As a consequence, we discard quantities that are identified  
 291 within this range.
- 292 2. Known laser wavelengths are irrelevant for identifying DIB carriers so we discard any  
 293 identified physical quantities that co-occur with the token "laser" or "light" within a win-  
 294 dow of 3 words.
- 295 3. Using the DIBs-enhanced Word2Vec-CBOW model in the sentence window as the con-  
 296 text where a physical quantity is identified, we compute a cosine similarity between the  
 297 context vector and the physical-quantity vector. Articles of low ( $< 0.5$ ) cosine similarities  
 298 are filtered out to remove false positives given by the unit-recognition algorithm.

299 After applying filters, we screen through all candidate articles and identify twelve papers of  
 300 high relevance, which are discussed in the following section.

<sup>4</sup><https://ui.adsabs.harvard.edu/>

<sup>5</sup><https://the.iris.ai/>

<sup>6</sup><https://leonid.arc.nasa.gov/DIBcatalog.html>

301 **V ASSESSMENT OF THE RESULTS**

302 Our model points us toward twelve articles presenting spectroscopic measurements of molecules  
 303 having transitions corresponding to some DIBs. These findings are summarized in Table 1.  
 304 They now need to be scrutinized.

305 **5.1 The Results**

Article	Transitions	Closest DIB	Molecule
<i>Wakakuwa et al. (2010)</i> Butterfly eye pigment	425 nm 453 nm 563 nm 620 nm 640 nm	425.90 ± 0.01 nm 450.18 ± 0.02 nm 563.50 ± 0.01 nm 619.90 ± 0.01 nm 640.05 ± 0.01 nm	11-cis retinal chromophore within opsin (H, C, O)
<i>Dove et al. (1995)</i> Coral pigment	560 nm 580 nm 590 nm	560.09 ± 0.01 nm 580.66 ± 0.01 nm 590.06 ± 0.01 nm	Chromophores within pocilloporin
<i>Davies et al. (2009)</i> Elephant shark eye pigment	441.9 ± 1.0 nm 493.7 ± 2.6 nm 496.3 ± 0.1 nm 498.3 ± 0.3 nm 498.7 ± 0.3 nm 504.1 ± 1.0 nm 509.5 ± 0.5 nm 510.1 ± 0.2 nm 520.9 ± 2.0 nm 534.2 ± 1.0 nm 547.8 ± 2.2 nm	442.82 ± 0.17 nm 494.74 ± 0.01 nm 496.39 ± 0.01 nm 498.21 ± 0.01 nm 498.74 ± 0.01 nm 505.48 ± 0.01 nm 509.21 ± 0.01 nm 510.10 ± 0.01 nm 521.79 ± 0.01 nm 534.25 ± 0.01 nm 548.08 ± 0.01 nm	Chromophores within proteins (H, C, N, O)
<i>Spady et al. (2006)</i> Cichlid eye pigment	423 nm 456 nm 472 nm 518 nm 528 nm 561 nm	425.90 ± 0.01 nm 450.18 ± 0.15 nm 472.68 ± 0.02 nm 517.81 ± 0.01 nm 529.8 ± 0.01 nm 560.98 ± 0.01 nm	Retinal chromophores within opsin
<i>Wolfbeis et al. (2001)</i> Valinomycin	488 nm	488.00 ± 0.01 nm	Valinomycin (H, C, N, O)
<i>Filosa (2001)</i> Blood proteins	695 nm	694.46 ± 0.01 nm	Amide II (H, C, N, O)
<i>Davies et al. (2009)</i> Agnathan eye pigments	501.0 ± 0.1 nm 535.5 ± 3.3 nm 544.1 ± 5.0 nm 554.3 ± 2.0 nm 562.8 ± 0.4 nm	500.36 ± 0.01 nm 535.88 ± 0.01 nm 543.35 ± 0.01 nm 554.51 ± 0.01 nm 563.50 ± 0.01 nm	Chromophores within opsin
<i>Schoot Uiterkamp et al. (1976)</i> Mushroom absorption	653 nm 755 nm	653.65 ± 0.01 nm 755.94 ± 0.01 nm	Tyrosinase (H, C, O, Cu)
<i>Davies et al. (2007)</i> Lamprey eye pigment	439 nm 492 nm 497 nm	436.39 ± 1.10 nm 494.74 ± 0.01 nm 496.91 ± 0.01 nm	Chromophores
<i>Maréchal et al. (2007)</i>	445 nm	442.82 ± 1.69 nm	Heme intermediate

Article	Transitions	Closest DIB	Molecule
Nitric-oxide synthase			(H, C, N, O, Fe)
Rémigy <i>et al.</i> (2003) Bacterium cytochrome	420 nm 525.2 nm 545.4 nm	425.90 ± 0.01 nm 525.18 ± 0.01 nm 545.06 ± 0.83 nm	Heme-type molecule (H, C, O, N, Fe)
Fasick <i>et al.</i> (1998) Dolphin eye pigment	488 nm 545 nm	488.00 ± 0.12 nm 545.06 ± 0.83 nm	11-cis retinal chromophore

Table 1: *Summary of the results.* The first column lists the found articles, with their topics. The second column shows the transition wavelengths reported in each article. We quote the uncertainty when it is reported, otherwise we assume it is the last significant digit, that is 1 nm in most cases. The third column gives the closest DIB reported by Hobbs *et al.* (2009). This database contains 380 bands in the  $\lambda = 380 - 810$  nm range. We do not discuss bands reported by the interdisciplinary articles outside this spectral range. Values in grey correspond to the case where the DIB centroid is not consistent with the article measurement. The last column gives the studied molecule, with its constituting elements between parentheses, when they are provided.

306 The twelve interdisciplinary articles listed in Table 1 are all biochemistry studies. They all  
 307 report experimental spectroscopic measurements of organic molecules. Most of these molecules  
 308 are constituted of abundant interstellar atoms, mainly, H, C, N and O. We can divide them into  
 309 the two following categories.

### 310 5.1.1 Chromophores

311 A majority of the papers in Table 1 deal with animal retinal eye pigments (Fasick *et al.*, 1998;  
 312 Spady *et al.*, 2006; Davies *et al.*, 2007, 2009; Wakakuwa *et al.*, 2010). These studies indeed  
 313 measure the absorption bands of organic molecules in the visible range. Their potential rel-  
 314 evance to DIBs is thus obvious. All these molecules are proteins (such as opsin) containing  
 315 *chromophores*. Chromophores are molecular groupings, often part of a larger molecule, that  
 316 are responsible for the color of an organism. Their particular optical properties are due to the  
 317 presence of chemical double bonds. It is thus reasonable to assume that they are responsible for  
 318 the reported bands. A recurring chromophore in these studies is 11-cis retinal (Figure 7). Not  
 319 all these papers discuss the actual chromophores present in their molecules, probably because  
 they are not always known.

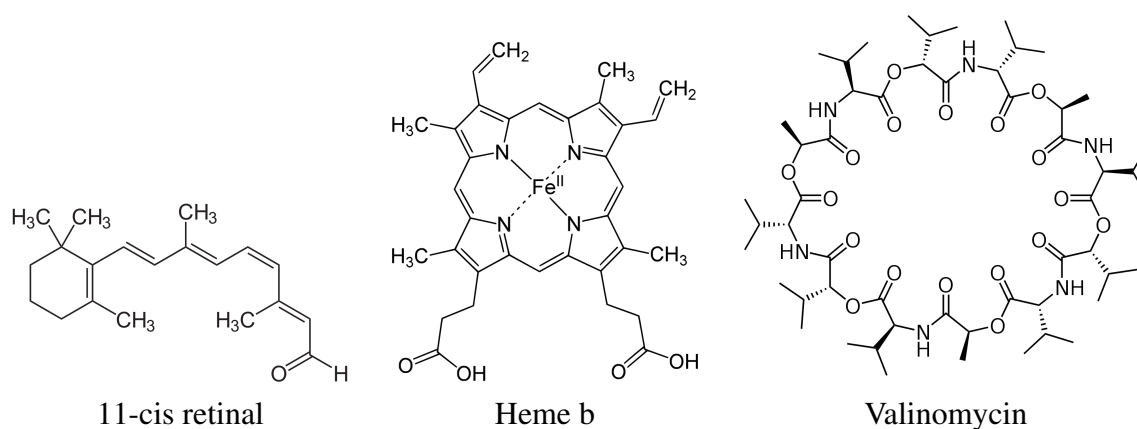


Figure 7: *Molecular structure of 11-cis retinal, heme b and valinomycin.*

320

### 321 5.1.2 Heme and other molecules

322 Apart from chromophores, our model points us toward several other organic molecules. In  
323 particular, two studies deal with molecules related to heme (Figure 7; [Rémigy et al., 2003](#);  
324 [Maréchal et al., 2007](#)). Heme is the molecule that allows hemoglobin to transport oxygen in  
325 the blood stream. Another potentially interesting molecule for interstellar chemistry is valino-  
326 mycin (Figure 7; [Wolfbeis et al., 2001](#)). We will discuss the likeliness of the existence of these  
327 molecules in the ISM in Sect. 5.2.2.

## 328 5.2 Discussion

329 We now attempt at assessing these results. We start by discussing the numerical values of the  
330 spectral band centroid. We then evaluate the likeliness of these molecules as carriers of some  
331 DIBs.

### 332 5.2.1 Accuracy of the reported transitions

333 The second column of Table 1 lists the central wavelengths of the bands reported by the articles  
334 found by our model. Unfortunately, only a few of these papers quote measurement errors on the  
335 centroid. This adds some uncertainty to our assessment. The third column of Table 1 gives the  
336 closest DIB from the measured band, using the [Hobbs et al. \(2009\)](#) compilation. Out of the 43  
337 centroids, only 8 (19 %) do not have a DIB within  $\pm 1\sigma$  (grey values).

338 The first question one can ask is how probable is it to draw a centroid wavelength at  $\pm 1\sigma$   
339 of a DIB. The compilation of [Hobbs et al. \(2009\)](#) contains 380 DIBs in the wavelength range  
340  $\lambda = 320\text{--}810$  nm. It gives a probability of 0.78 DIB per nm. Assuming that DIBs are randomly,  
341 uniformly distributed with this density, the Poissonian probability to find a DIB within  $\pm 1\sigma$   
342 ( $\pm 1$  nm in our case) is 33 %. Yet, our list matches 81 % of them. Moreover, several studies  
343 list a large number of bands. The most spectacular is the paper by [Davies et al. \(2009\)](#), listing  
344 11 bands with their uncertainties. It happens that all of them coincide with a DIB within  $\pm 1\sigma$ .  
345 Using the uncertainties quoted by [Davies et al. \(2009\)](#), the probability to randomly obtain such  
346 a result is 0.2 %. These results are thus non trivial.

347 We have mentioned in Sect. II that DIBs were a few Angstrom wide. The bands reported in these  
348 articles are however wider (up to several tens of nm). One can thus wonder if they are relevant to  
349 our problem. The answer to this question is not straightforward. Molecular band width depends  
350 on the molecular size as well as on the temperature. Concerning the first effect, our candidate  
351 molecules contain a few tens to a hundred atoms, which is the expected size of DIB carriers  
352 ([MacIsaac et al., 2022](#)). The second effect, the temperature, might explain the broadness of the  
353 features. All these molecules are indeed measured at room temperature ( $T \simeq 300$  K), whereas  
354 these molecules in the ISM would be close to  $T \simeq 20$  K. Low-temperature measurements  
355 should be made to confirm this, but we can reasonably assume that these molecules should have  
356 narrower bands in interstellar conditions.

### 357 5.2.2 Insights for astrochemistry

358 We have seen that most of the molecules our model points us toward are constituted of abundant  
359 interstellar atoms. This is a first requirement that our model has successfully accounted for. It  
360 is likely the result of the NLP training, associating DIBs with these elements and with organic  
361 molecules.

362 Our highest-score molecules, chromophores, have been proposed as DIB carriers (*e.g.* [Johnson,](#)  
363 [2006](#); [Adams and Oka, 2019](#)), and this is probably why our NLP model has selected them.

364 Chromophores are probably too brittle molecules to survive in the diffuse ISM. For instance,  
365 the long chain of 11-cis retinal (Figure 7) will likely be photolyzed by UV photons. However,  
366 chromophores could form moieties in larger molecules or, possibly, in nanometer-size dust  
367 grains (Jones, 2014, 2016b). This would allow them to be present in the ISM and carry some of  
368 the DIBs. We note this is also the case in the biochemistry studies. Several of the found papers  
369 mention that the chromophores are covalently bonded with their proteins.

370 Heme (Figure 7) has also been discussed as a possible interstellar molecule (Jones, 2016a).  
371 Together with porphyrin, which is also considered as a potential DIB carrier (Johnson, 1994),  
372 these molecules could result from the reaction of nitrogen atoms with hydrogenated amorphous  
373 carbon grains. The structure of valinomycin (Figure 7) is not unlike heme and porphyrins. The  
374 problem of valinomycin is that it is probably not going to be very stable in the ISM because of  
375 all its oxygen, which is going to make it very reactive with atomic C, N, H and S.

## 376 VI SUMMARY AND CONCLUSION

377 We have trained a *Natural Language Processing* (NLP) model on a corpus of astrophysical  
378 publications dealing with the open question of the origin of *Diffuse Interstellar Bands* (DIBs).  
379 We have then used this model to explore an interdisciplinary corpus of scientific literature,  
380 with the hope to find relevant molecules having transitions matching some DIBs. We have  
381 implemented a careful parsing of physical quantities and their units, in order to identify mea-  
382 sured wavelengths in the visible electromagnetic spectrum. Our model points us toward twelve  
383 biochemistry studies presenting spectroscopic measurements of molecules having transitions  
384 consistent with some DIBs. More than half of these molecules contain chromophores. Several  
385 other studies deal with molecules linked to heme and valinomycin.

386 Our main objectives, the feasibility to use NLP to address open scientific questions, is reached.  
387 We have shown that NLP is able to surface candidate DIBs molecules from an interdisciplinary  
388 corpus of scientific literature. This confirms that NLP-methodologies can generate plausible  
389 and non-trivial hypotheses for future investigation. First, the association between the reported  
390 transitions and the DIBs would be unlikely if they were purely random. Second, our NLP model  
391 has pointed us toward molecules relevant to the *InterStellar Medium* (ISM). These molecules  
392 are constituted of abundant interstellar atoms. In addition, several of these molecules have  
393 been proposed in the past to be DIB carriers. Their presence in the ISM has not been proven,  
394 yet, but it is conjectured that they could form moieties in interstellar grains. In light of our  
395 study, NLP thus appears as a practical tool for interdisciplinarity. In the future, extending our  
396 analysis to the full text, and expanding the generic corpus to thoroughly-cover cross-domain  
397 literature, can potentially give us a more complete result. From an astrophysical point of view,  
398 our work adds credibility to the possibility that some DIBs could be carried by chromophores  
399 and to the role of heme and porphyrins in interstellar chemistry. These findings need now to be  
400 further investigated in the laboratory, in interstellar conditions, at low temperature. As a further  
401 outlook, our methodology should be applicable to other open scientific questions which require  
402 interdisciplinary knowledge.

## 403 Acknowledgements

404 We thank Anthony Jones for a useful discussion about the molecules found by our model  
405 and about DIBs in general, and Jason Hoelscher-Obermaier for constructive comments on the  
406 prospects of this study. This paper was supported by funding from NRC, Project ID: 309594,  
407 the AI Chemist under the collaboration of Iris.ai AS with Dr. Frédéric GALLIANO covering  
408 the contributions of Viktor BOTEV, Jeremy MINTON, Ronin WU, and partly Corentin VAN

409 DEN BROEK D'OBRENAN. This work was supported by the Programme National "Physique  
410 et Chimie du Milieu Interstellaire" (PCMI) of the CNRS/INSU with INC/INP co-funded by  
411 CEA and CNES.

## 412 References

- 413 Adams K., Oka T. (2019, December). Relating the Carriers of  $\lambda$ 5797.1 Diffuse Interstellar Band and  $\lambda$ 5800 Red  
414 Rectangle Band. *ApJ* 886(2), 138. doi:10.3847/1538-4357/ab4c49.
- 415 Asplund M., Grevesse N., Sauval A. J., Scott P. (2009, September). The Chemical Composition of the Sun.  
416 *ARA&A* 47, 481–522. arXiv:0909.0948, doi:10.1146/annurev.astro.46.060407.145222.
- 417 Bellocche A., Garrod R. T., Müller H. S. P., Menten K. M. (2014, September). Detection of a branched alkyl  
418 molecule in the interstellar medium: iso-propyl cyanide. *Science* 345(6204), 1584–1587. arXiv:1410.2607,  
419 doi:10.1126/science.1256678.
- 420 Bengio Y., Ducharme R., Vincent P., Jauvin C. (2003). A neural probabilistic language model. *JOURNAL OF*  
421 *MACHINE LEARNING RESEARCH* 3, 1137–1155.
- 422 Bianchi S., De Vis P., Viaene S., Nersesian A., Mosenkov A. V., Xilouris E. M., Baes M., Casasola V., Cassarà  
423 L. P., Clark C. J. R., Davies J. I., De Looze I., Dobbels W., Galametz M., Galliano F., Jones A. P., Lianou S.,  
424 Madden S. C., Trčka A. (2018, December). Fraction of bolometric luminosity absorbed by dust in DustPedia  
425 galaxies. *A&A* 620, A112. arXiv:1810.01208, doi:10.1051/0004-6361/201833699.
- 426 Bojanowski P., Grave E., Joulin A., Mikolov T. (2017, 06). Enriching Word Vectors with  
427 Subword Information. *Transactions of the Association for Computational Linguistics* 5, 135–  
428 146. URL: [https://doi.org/10.1162/tacl\\_a\\_00051](https://doi.org/10.1162/tacl_a_00051), arXiv:[https://direct.mit.edu/tacl/article-  
429 pdf/doi/10.1162/tacl\\_a.00051/1567442/tacl\\_a.00051.pdf](https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a.00051/1567442/tacl_a.00051.pdf), doi:10.1162/tacl\_a.00051.
- 430 Botev V., Marinov K., Schäfer F. (2017). Word importance-based similarity of documents metric (wisdm): Fast and  
431 scalable document similarity metric for analysis of scientific documents. In *Proceedings of the 6th International*  
432 *Workshop on Mining Scientific Publications, WOSP 2017*, pp. 17. ACM. URL: [http://doi.acm.org/  
433 10.1145/3127526.3127530](http://doi.acm.org/10.1145/3127526.3127530), doi:10.1145/3127526.3127530.
- 434 Bron E. (2014, November). *Stochastic processes in the interstellar medium*. Ph. D. thesis, LERMA, Observatoire  
435 de Paris, PSL Research University, CNRS, Sorbonne Universités, UPMC Univ. Paris 06, F-92190, Meudon,  
436 France.
- 437 Cami J., Bernard-Salas J., Peeters E., Malek S. E. (2010, September). Detection of C<sub>60</sub> and C<sub>70</sub> in a Young  
438 Planetary Nebula. *Science* 329(5996), 1180. doi:10.1126/science.1192035.
- 439 Campbell E. K., Holz M., Gerlich D., Maier J. P. (2015, July). Laboratory confirmation of C<sub>60</sub><sup>+</sup> as the carrier of  
440 two diffuse interstellar bands. *Nature* 523, 322–323. doi:10.1038/nature14566.
- 441 Davies W. L., Carvalho L. S., Tay B. H., Brenner S., Hunt D. M., Venkatesh B. (2009, April). Into the blue: gene  
442 duplication and loss underlie color vision adaptations in a deep-sea chimaera, the elephant shark *Callorhynchus*  
443 *milii*. *Genome research* 19(3), 415–426. doi:10.1101/gr.084509.108.
- 444 Davies W. L., Collin S. P., Hunt D. M. (2009, August). Adaptive gene loss reflects differences in the visual ecology  
445 of basal vertebrates. *Molecular biology and evolution* 26(8), 1803–1809. doi:10.1093/molbev/msp089.
- 446 Davies W. L., Cowing J. A., Carvalho L. S., Potter I. C., Trezise A. E., Hunt D. M., Collin S. P. (2007). Func-  
447 tional characterization, tuning, and regulation of visual pigment gene expression in an anadromous lamprey.  
448 *FASEB journal : official publication of the Federation of American Societies for Experimental Biology* 21(11),  
449 2713–2724. doi:10.1096/fj.06-8057com.
- 450 Densen P. (2011). Challenges and opportunities facing medical education. *Transactions of the American Clinical*  
451 *and Climatological Association*.
- 452 Devlin J., Chang M.-W., Lee K., Toutanova K. (2019, June). BERT: Pre-training of deep bidirectional trans-  
453 formers for language understanding. In *Proceedings of the 2019 Conference of the North American Chap-  
454 ter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and*  
455 *Short Papers)*, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics. URL:  
456 <https://aclanthology.org/N19-1423>, doi:10.18653/v1/N19-1423.



- 457 Dove S. G., Takabayashi M., Hoegh-Guldberg O. (1995). Isolation and partial characterization of the pink  
458 and blue pigments of pocilloporid and acroporid corals. *The Biological bulletin* 189(3), 288—297.  
459 [doi:10.2307/1542146](https://doi.org/10.2307/1542146).
- 460 Draine B. T. (2003). Interstellar Dust Grains. *ARA&A* 41, 241–289. [arXiv:arXiv:astro-ph/0304489](https://arxiv.org/abs/astro-ph/0304489),  
461 [doi:10.1146/annurev.astro.41.011802.094840](https://doi.org/10.1146/annurev.astro.41.011802.094840).
- 462 Draine B. T. (2011). *Physics of the Interstellar and Intergalactic Medium*. Princeton University Press.
- 463 Fan H., Hobbs L. M., Dahlstrom J. A., Welty D. E., York D. G., Rachford B., Snow T. P., Sonnentrucker P.,  
464 Baskes N., Zhao G. (2019, June). The Apache Point Observatory Catalog of Optical Diffuse Interstellar Bands.  
465 *ApJ* 878(2), 151. [arXiv:1905.05962](https://arxiv.org/abs/1905.05962), [doi:10.3847/1538-4357/ab1b74](https://doi.org/10.3847/1538-4357/ab1b74).
- 466 Fasick J. I., Cronin T. W., Hunt D. M., Robinson P. R. (1998). The visual pigments of the bottlenose dolphin  
467 (*tursiops truncatus*). *Visual neuroscience* 15(4), 643–651. [doi:10.1017/s0952523898154056](https://doi.org/10.1017/s0952523898154056).
- 468 Filosa A. (2001). *Study of locally unfolded forms of cytochrome c by Fourier transform infrared and of H2O2-*  
469 *mediated oxidation of ferricytochrome c and metmyoglobin by mass spectrometry*. Ph. D. thesis, Concordia Uni-  
470 versity. Unpublished. URL: <https://spectrum.library.concordia.ca/id/eprint/1417/>.
- 471 Firth J. R. (1957). A synopsis of linguistic theory, 1930-1955.
- 472 Galliano F. (2022, February). A Nearby Galaxy Perspective on Interstellar Dust Properties and their Evolution.  
473 *Habilitation Thesis*, 1. [arXiv:2202.01868](https://arxiv.org/abs/2202.01868).
- 474 Galliano F., Galametz M., Jones A. P. (2018, September). The Interstellar Dust Properties of Nearby Galaxies.  
475 *ARA&A* 56, 673–713. [doi:10.1146/annurev-astro-081817-051900](https://doi.org/10.1146/annurev-astro-081817-051900).
- 476 Goodfellow I., Bengio Y., Courville A. (2016). *Deep Learning*. MIT Press. [http://www.  
477 deeplearningbook.org](http://www.deeplearningbook.org).
- 478 Grezes F., Blanco-Cuaresma S., Accomazzi A., Kurtz M. J., Shapurian G., Henneken E., Grant C. S., Thompson  
479 D. M., Chyla R., McDonald S., Hostetler T. W., Templeton M. R., Lockhart K. E., Martinovic N., Chen S., Tan-  
480 ner C., Protopapas P. (2021, December). Building astroBERT, a language model for Astronomy & Astrophysics.  
481 *arXiv e-prints*, arXiv:2112.00590. [arXiv:2112.00590](https://arxiv.org/abs/2112.00590).
- 482 Gvaramadze, V. V., Menten, K. M. (2012). Discovery of a parsec-scale bipolar nebula around mwca. *A&A* 541, A7.  
483 URL: <https://doi.org/10.1051/0004-6361/201218841>, [doi:10.1051/0004-6361/201218841](https://doi.org/10.1051/0004-6361/201218841).
- 484 Hastings J., Magka D., Batchelor C., Duan L., Stevens R., Ennis M., Steinbeck C. (2012, April). Structure-based  
485 classification and ontology in chemistry. *Journal of Cheminformatics* 4, 8. [doi:10.1186/1758-2946-4-8](https://doi.org/10.1186/1758-2946-4-8).
- 486 Hearnshaw J. B. (2014). *The interpretation of stellar spectra and the birth of astrophysics* (2 ed.), pp. 127–151.  
487 Cambridge University Press. [doi:10.1017/CBO97811139382779.009](https://doi.org/10.1017/CBO97811139382779.009).
- 488 Heger M. L. (1922). Further study of the sodium lines in class B stars. *Lick Observatory Bulletin* 10, 141–145.  
489 [doi:10.5479/ADS/bib/1922LicOB.10.141H](https://doi.org/10.5479/ADS/bib/1922LicOB.10.141H).
- 490 Herbig G. H. (1995, January). The Diffuse Interstellar Bands. *ARA&A* 33, 19–74.  
491 [doi:10.1146/annurev.aa.33.090195.000315](https://doi.org/10.1146/annurev.aa.33.090195.000315).
- 492 Herbst E., van Dishoeck E. F. (2009, September). Complex Organic Interstellar Molecules. *ARA&A* 47(1), 427–  
493 480. [doi:10.1146/annurev-astro-082708-101654](https://doi.org/10.1146/annurev-astro-082708-101654).
- 494 Herzberg G. (1988, June). Historical Remarks on the Discovery of Interstellar Molecules. *JRASC* 82, 115.
- 495 Hobbs L. M., York D. G., Thorburn J. A., Snow T. P., Bishof M., Friedman S. D., McCall B. J., Oka T., Rachford  
496 B., Sonnentrucker P., Welty D. E. (2009, November). Studies of the Diffuse Interstellar Bands. III. HD 183143.  
497 *ApJ* 705(1), 32–45. [arXiv:0910.2983](https://arxiv.org/abs/0910.2983), [doi:10.1088/0004-637X/705/1/32](https://doi.org/10.1088/0004-637X/705/1/32).
- 498 Jenniskens P., Desert F.-X. (1994, July). A survey of diffuse interstellar bands (3800-8680 Å). *A&AS* 106.
- 499 Johnson F. M. (1994, May). Porphyrins in the interstellar medium (in grains). In A. G. G. M. Tielens (Ed.), *The*  
500 *Diffuse Interstellar Bands*, pp. 47–52.
- 501 Johnson F. M. (2006, December). Diffuse interstellar bands: A comprehensive laboratory study. *Spectrochimica*  
502 *Acta Part A: Molecular Spectroscopy* 65, 1154–1179. [arXiv:1706.04273](https://arxiv.org/abs/1706.04273), [doi:10.1016/j.saa.2006.03.004](https://doi.org/10.1016/j.saa.2006.03.004).
- 503 Jones A. P. (2014, October). A framework for resolving the origin, nature and evolution of the diffuse interstellar  
504 band carriers? *Planetary and Space Science* 100, 26–31. [arXiv:1411.5854](https://arxiv.org/abs/1411.5854), [doi:10.1016/j.pss.2013.11.011](https://doi.org/10.1016/j.pss.2013.11.011).

- 505 Jones A. P. (2016a, December). Dust evolution, a global view I. Nanoparticles, nascence, nitrogen and natural  
506 selection ... joining the dots. *Royal Society Open Science* 3, 160221. doi:10.1098/rsos.160221.
- 507 Jones A. P. (2016b, December). Dust evolution, a global view: II. Top-down branching, nanoparticle frag-  
508 mentation and the mystery of the diffuse interstellar band carriers. *Royal Society Open Science* 3, 160223.  
509 doi:10.1098/rsos.160223.
- 510 Kerzendorf W. E. (2019, June). Knowledge discovery through text-based similarity searches for astronomy litera-  
511 ture. *Journal of Astrophysics and Astronomy* 40(3), 23. arXiv:1705.05840, doi:10.1007/s12036-019-9590-5.
- 512 Knoth P., Zdrahal Z. (2012). Core: three access levels to underpin open access. *D-Lib Magazine* 18(11/12). URL:  
513 <http://oro.open.ac.uk/35755/>.
- 514 Lamurias A., Sousa D., Clarke L. A., Couto F. M. (2019, January). BO-LSTM: classifying relations via long  
515 short-term memory networks along biomedical ontologies. *BMC Bioinformatics* 20(1), 10. URL: <https://doi.org/10.1186/s12859-018-2584-5>, doi:10.1186/s12859-018-2584-5.
- 517 Lan T.-W., Ménard B., Zhu G. (2015, October). Exploring the diffuse interstellar bands with the Sloan Digital Sky  
518 Survey. *MNRAS* 452(4), 3629–3649. arXiv:1406.7284, doi:10.1093/mnras/stv1519.
- 519 MacIsaac H., Cami J., Cox N. L. J., Farhang A., Smoker J., Elyajouri M., Lallement R., Sarre P. J., Cordiner M. A.,  
520 Fan H., Kulik K., Linnartz H., Foing B. H., van Loon J. T., Mulas G., Smith K. T. (2022, March). The EDIBLES  
521 survey V: Line profile variations in the  $\lambda\lambda 5797$ , 6379, and 6614 diffuse interstellar bands as a tool to constrain  
522 carrier sizes. *arXiv e-prints*, arXiv:2203.01803. arXiv:2203.01803.
- 523 Manning C. D., Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Mas-  
524 sachusetts: The MIT Press. URL: <http://nlp.stanford.edu/fsnlp/>.
- 525 Maréchal A., Mattioli T. A., Stuehr D. J., Santolini J. (2007). Activation of peroxynitrite by inducible nitric-  
526 oxide synthase: a direct source of nitrate stress. *The Journal of biological chemistry* 282(19), 14101–14112.  
527 doi:10.1074/jbc.M609237200.
- 528 McGuire B. A. (2018, December). 2018 Census of Interstellar, Circumstellar, Extragalactic, Protoplanetary Disk,  
529 and Exoplanetary Molecules. *ApJS* 239(2), 17. arXiv:1809.09132, doi:10.3847/1538-4365/aae5d2.
- 530 McGuire B. A., Loomis R. A., Burkhardt A. M., Lee K. L. K., Shingledecker C. N., Charnley S. B., Cooke I. R.,  
531 Cordiner M. A., Herbst E., Kalenskii S., Siebert M. A., Willis E. R., Xue C., Remijan A. J., McCarthy M. C.  
532 (2021, March). Detection of two interstellar polycyclic aromatic hydrocarbons via spectral matched filtering.  
533 *Science* 371(6535), 1265–1269. arXiv:2103.09984, doi:10.1126/science.abb7535.
- 534 Merrill P. W. (1934, August). Unidentified Interstellar Lines. *PASP* 46(272), 206–207. doi:10.1086/124460.
- 535 Mikolov T., Chen K., Corrado G., Dean J. (2013). Efficient estimation of word representations in vector space.  
536 *CoRR abs/1301.3781*. URL: <http://arxiv.org/abs/1301.3781>.
- 537 Mikolov T., Yih W.-t., Zweig G. (2013, June). Linguistic regularities in continuous space word representations.  
538 In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational*  
539 *Linguistics: Human Language Technologies*, Atlanta, Georgia, pp. 746–751. Association for Computational  
540 Linguistics. URL: <https://aclanthology.org/N13-1090>.
- 541 Pennington J., Socher R., Manning C. (2014, October). GloVe: Global vectors for word representation. In  
542 *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha,  
543 Qatar, pp. 1532–1543. Association for Computational Linguistics. URL: [https://aclanthology.org/](https://aclanthology.org/D14-1162)  
544 [D14-1162](https://aclanthology.org/D14-1162), doi:10.3115/v1/D14-1162.
- 545 Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. (2018). Deep contex-  
546 tualized word representations. *CoRR abs/1802.05365*. URL: <http://arxiv.org/abs/1802.05365>,  
547 arXiv:1802.05365.
- 548 Řehůřek R., Sojka P. (2010, May). Software Framework for Topic Modelling with Large Corpora. In *Proceedings*  
549 *of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, pp. 45–50. ELRA.  
550 <http://is.muni.cz/publication/884893/en>.
- 551 Rémy H. W., Aivaliotis M., Ioannidis N., Jenö P., Mini T., Engel A., Jaquinod M., Tsiotis G. (2003). Char-  
552 acterization by mass spectroscopy of a 10 kda c-554 cytochrome from the green sulfur bacterium chlorobium  
553 tepidum. *Photosynthesis research* 78(2), 153–160. doi:10.1023/B:PRES.0000004347.34228.2c.
- 554 Schoot Uiterkamp A. J., Evans L. H., Jolley R. L., Mason H. S. (1976). Absorption and circular dichroism spectra

- 555 of different forms of mushroom tyrosinase. *Biochimica et biophysica acta* 453(1), 200–204. doi:10.1016/0005-  
556 2795(76)90264-6.
- 557 Spady T. C., Parry J. W., Robinson P. R., Hunt D. M., Bowmaker J. K., Carleton K. L. (2006). Evolution of the  
558 cichlid visual palette through ontogenetic subfunctionalization of the opsin gene arrays. *Molecular biology and*  
559 *evolution* 23(8), 1538–1547. doi:10.1093/molbev/msl014.
- 560 Thomas B., Thronson H., Buonomo A., Barbier L. (2022, January). Determining Research Priorities for  
561 Astronomy Using Machine Learning. *Research Notes of the American Astronomical Society* 6(1), 11.  
562 arXiv:2203.00713, doi:10.3847/2515-5172/ac4990.
- 563 Thorburn J. A., Hobbs L. M., McCall B. J., Oka T., Welty D. E., Friedman S. D., Snow T. P., Sonnentrucker P.,  
564 York D. G. (2003, February). Some Diffuse Interstellar Bands Related to Interstellar C<sub>2</sub> Molecules. *ApJ* 584(1),  
565 339–356. doi:10.1086/345665.
- 566 Tielens A. G. G. M. (2005, September). *The Physics and Chemistry of the Interstellar Medium*. Cambridge  
567 University Press.
- 568 Tielens A. G. G. M. (2008, September). Interstellar Polycyclic Aromatic Hydrocarbon Molecules. *ARA&A* 46,  
569 289–337. doi:10.1146/annurev.astro.46.060407.145211.
- 570 Wakakuwa M., Terakita A., Koyanagi M., Stavenga D., Shichida Y., Arikawa K. (2010, nov). Evolution and  
571 mechanism of spectral tuning of blue-absorbing visual pigments in butterflies. *PLoS One* 5(11):e15015.  
572 doi:10.1371/journal.pone.0015015.
- 573 Walker G. A. H., Bohlender D. A., Maier J. P., Campbell E. K. (2015, October). Identification of More Interstellar  
574 C<sub>60</sub><sup>+</sup> Bands. *ApJL* 812(1), L8. arXiv:1509.06818, doi:10.1088/2041-8205/812/1/L8.
- 575 Wolfbeis O. S., Opitz D., Werner T., Ouart A. (2001). Chiroptic recognition of potassium ion. *Journal of molecular*  
576 *recognition* 14(1), 13–17. doi:10.1002/1099-1352(200101/02)14:1;13::AID-JMR514;3.0.CO;2-K.