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

JOURNAL OF INTERDISCIPLINARY METHODOLOGIES AND ISSUES IN SCIENCE

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

Corentin VAN DEN BROEK D'OBRENAN[1,2], Frédéric GALLIANO[1], Jeremy MINTON[2], Viktor BOTEV[2], Ronin WU[*2]

[1]Université Paris-Saclay, Université Paris Cité, CEA, CNRS, AIM, 91191, Gif-sur-Yvette, France
[2]Iris AI, Bekkestua, Norway

*Corresponding author: ronin@iris.ai

DOI: 10.46298/jimis.9388
Submitted: 22/04/2023 - Accepted: 01/04/2023
Volume: **11** - Year: **2023**
Issue: **Penser l'interdisciplinarité en pratique**
Editors: *Deborah Nourrit, Guillaume Alevêque, Anne Laurent, Thérèse Libourel*

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

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

II DIFFUSE INTERSTELLAR BANDS

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

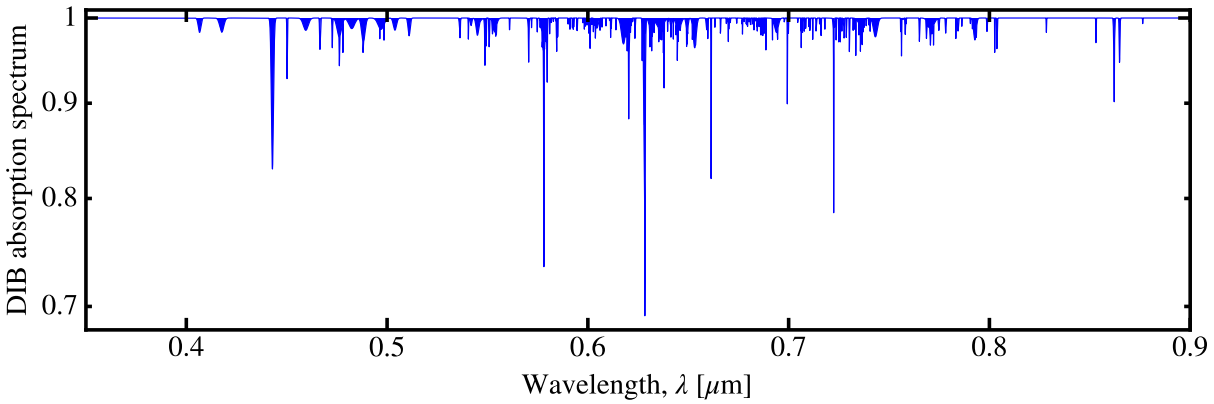


Figure 1: Synthetic absorption spectrum of the most prominent DIBs. The parameters of each feature (position, width and depth) come from the observational compilation by Jenniskens and Desert (1994).

2.1 The Interstellar Medium

Since DIBs are interstellar features, we first need to review the general properties of the ISM. Broadly speaking, the ISM is constituted of all the matter filling the volume of a galaxy between the stars. This matter is essentially gaseous, but about half a percent of its mass is made of small solid particles, the dust grains (Tielens, 2005; Draine, 2011, for textbooks). The elemental composition of the ISM is 74 % hydrogen, 25 % helium, and the remaining 1 % contains all the heavier elements (Asplund *et al.*, 2009). Figure 2 shows the abundances, in the Solar system, of the first elements in the periodic table. We see that besides hydrogen and helium, the two 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.

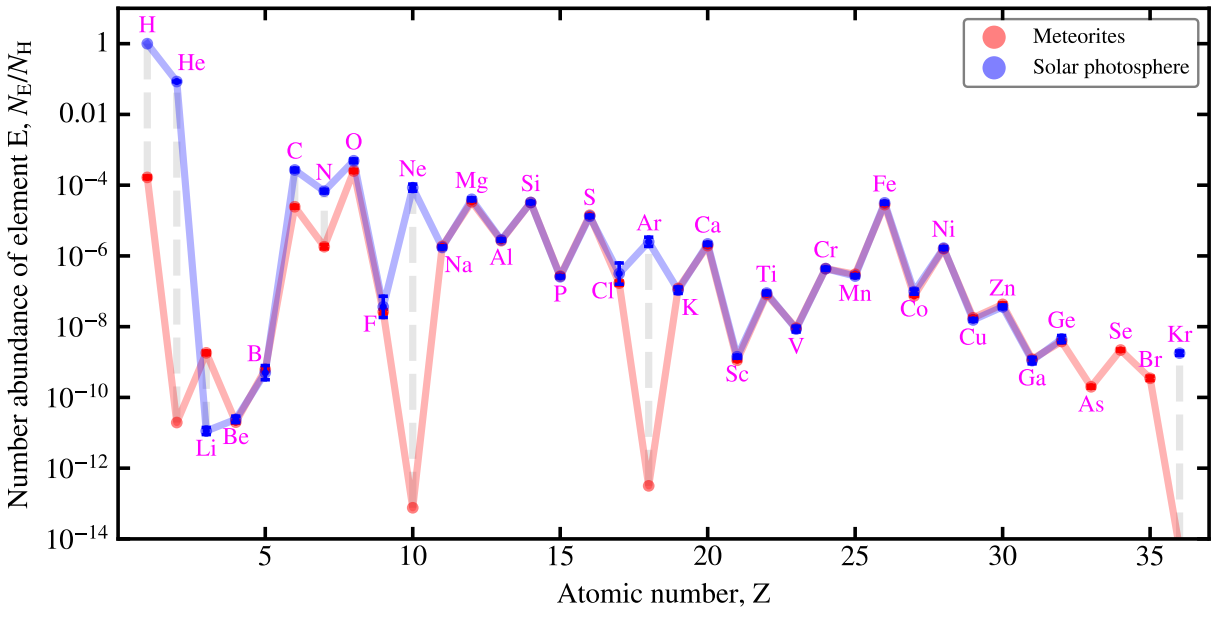


Figure 2: Elemental abundances in the Solar system, from the data in Asplund *et al.* (2009). Those are *number* abundances (*i.e.* number of atoms relative to hydrogen). The blue dots represent measures made in the Solar photosphere, through absorption spectroscopy, and the red dots correspond to the 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.

2.1.1 *The Phases of the ISM*

The ISM is a highly heterogeneous medium (*e.g.* Chap. III.3 of Galliano, 2022). Half of the volume of our galaxy is filled with a permeating hot ionized gas with a very low density (temperature, $T = 10^6$ K; density, $n \simeq 3 \times 10^{-3}$ cm^{-3}). This phase is heated by the shock waves from supernova explosions. The rest of the volume covers a wide range of density, temperature and atomic state. The coldest and densest interstellar regions are called molecular clouds. As their name indicates, the elements in these clouds have combined to form molecules, majoritarily H_2 and CO. Their high density (up to $n \simeq 10^6$ cm^{-3}) allow them to be shielded from the stellar radiation and to reach low temperatures ($T \simeq 10$ K). Between these two extremes, there are nine orders of magnitude in density and five in temperature, unlike anything we can find on Earth. DIBs are found preferentially in the diffuse ISM, and disappear in dense regions (*e.g.* Lan *et al.*, 2015), although some specific DIBs are found in diffuse molecular clouds (density $n \simeq 10^2$ cm^{-3}; Thorburn *et al.*, 2003).

2.1.2 *Interstellar Molecules*

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

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

2.1.3 *Interstellar Dust Grains*

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

[1] Rotational lines arise from transitions between different values of the quantum angular momentum of the molecule.

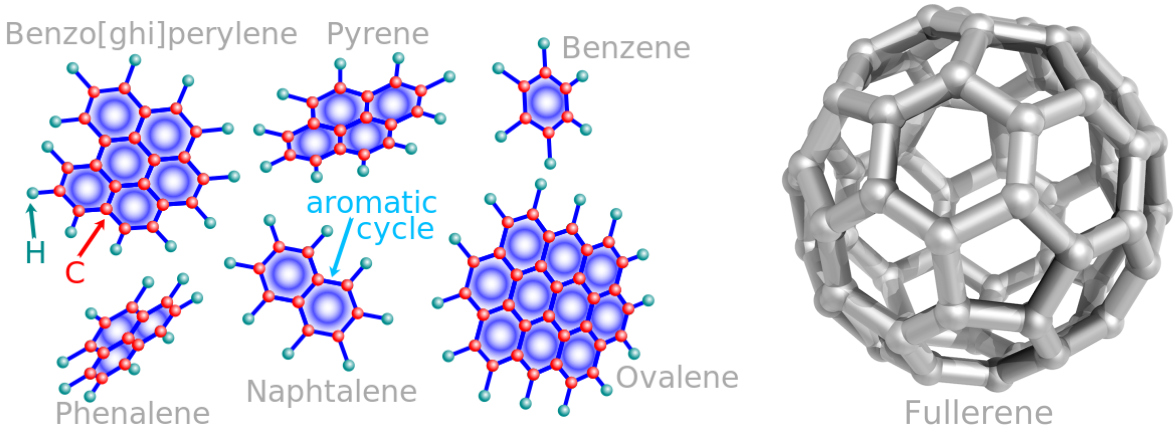


Figure 3: The PAH molecular family. On the left side, we show six different PAHs. Hydrogen atoms are represented in cyan, and carbon atoms, in red. On the right side, we show the buckminsterfullerene, which is a spherical molecule, composed of 60 carbon atoms. Credit: the left figure is adapted from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

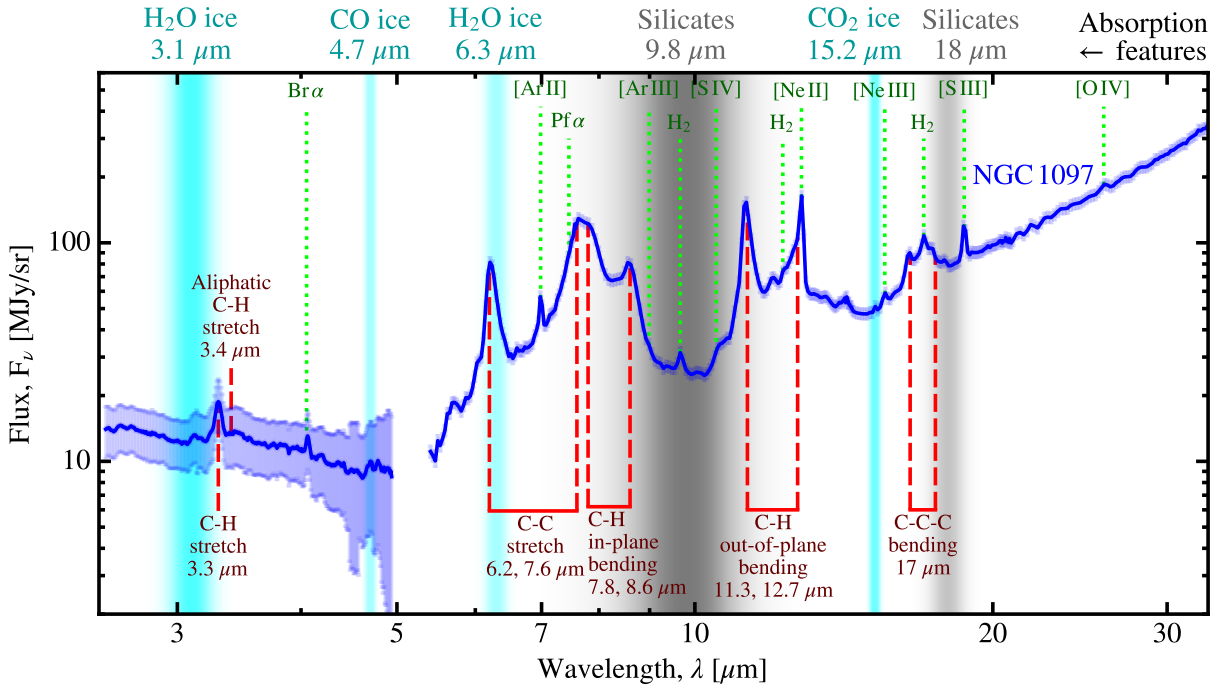


Figure 4: Mid-infrared spectrum of the galaxy NGC 1097. The blue line, with error bars, is the observed spectrum of the central region of this object. We have highlighted the main spectral features: on top, we have indicated the main silicate and ice interstellar absorption bands; in green, we have shown 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).

Interstellar dust grains are small solid particles with radii ranging from $\simeq$ 3 Å to 300 nm (*e.g.* Draine, 2003, for a review). They account for about half of the mass of heavy elements in the ISM, thus about half a percent of its total mass. Yet, these grains, which are predominantly silicate and carbonaceous compounds, have a very important role. In a quiescent galaxy, such as the Milky Way, they absorb about 25 % of the stellar power, in the ultraviolet (UV) and visible range, and re-emit it thermally in the infrared (*e.g.* Bianchi *et al.*, 2018). In regions of massive star formation, this fraction can go up to 99 %. These regions are thus totally opaque to visible photons and can be studied only through their infrared radiation, emitted by the dust. Grains are also the catalysts of several chemical reactions, including the formation of H_2, the most abundant molecule in the Universe (*e.g.* Bron, 2014).

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

2.1.4 The Lifecycle of a Galaxy

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

2.2 The Uncertain Nature of DIBs

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

2.2.1 Constraints on the Nature of the DIB Carriers

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

2.2.2 *Identification of Buckminsterfullerene*

The only molecule to date, unambiguously identified as a DIB carrier is the *buckminsterfulerene*[2] (Campbell *et al.*, 2015; Walker *et al.*, 2015). The cation of this molecule, C_{60}^{+}, can account for two, possibly four DIBs, at near-infrared wavelengths. Notice that this molecule satisfies all the constraints we have listed in Sect. 2.2.1. The fact that this molecule had been detected beforehand, in a planetary nebula, *via* its mid-IR features (Cami *et al.*, 2010), makes this identification trustworthy. It is possible that other COMs, and probably other PAHs, are DIB carriers. There are however variations of the relative strengths of DIBs, across sightlines (Herbig, 1995). This indicates that DIB carriers likely come from a diversity of molecules whose relative abundance varies with the environment.

2.2.3 *The Significance of DIB Identification*

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

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

III NATURAL LANGUAGE PROCESSING

NLP is a sub-discipline of machine-learning, that combines linguistics and artificial intelligence to allow computers to interface with human language. In this application specifically, it is used to extract data from natural language documents: academic papers.

3.1 Machine learning and neural networks

Machine-Learning (ML) is a methodology for producing computer programs whose main decision making is determined by data, not directly coded by a computer scientist. This methodology enables us to perform automated decision making too detailed for a human to design by using a set of examples too numerous for a human to fully perceive. Done correctly, the decision making should remain valid for data not explicitly included in the set of training examples (Goodfellow *et al.*, 2016, for a text book). *Neural Networks* (NNs) are a popular algorithm for many modern ML applications. Modelled on an organic brain, NNs are comprised of artificial neurons that perform a non-linear operation on a signal and propagate the result as a signal to the next neurons. When that signal has passed through the entire network of neurons, it can be interpreted as a meaningful prediction. Each neuron's behaviour is parameterized.

[2] Fullerenes constitute a family of compact closed-mesh carbon compounds. Buckminsterfullerene is the variety with formula C_{60}.

Corrections are made to iteratively update the neurons' parameters during a dedicated training process by propagating a prediction error backwards through the network and calculating the error gradient with respect to each parameter. Neurons are typically arranged in layers because the corresponding vectors of parameters allow efficient matrix arithmetic to be used to perform the necessary operations.

3.2 Word-embeddings

Word-embeddings are numerical representations of words, useful for computation, and NNs are a common tool for producing and using word-embeddings.

In recent years, development of word-embeddings models has taken big steps forward from NNs with few layers, such as the Word2Vec (Mikolov *et al.*, 2013), Glove (Pennington *et al.*, 2014), and FastText (Bojanowski *et al.*, 2017) models, to complex models that require enormous volumes of data to train, such as ELMo (Peters *et al.*, 2018) and BERT (Devlin *et al.*, 2019).

3.2.1 Concepts

Word-embeddings are underpinned by the distributional hypothesis: that words with similar co-occurrence distributions have similar meanings or "a word is characterized by the company it keeps" (Firth, 1957). Techniques leveraging this hypothesis are able to represent semantic properties of words and capture meaningful syntactic and semantic regularities. Often, regularities are observed as constant vector offsets between pairs of words sharing a particular relationship. For example, it has been demonstrated that the Word2Vec model in Mikolov *et al.* (2013) can learn the male/female relationship, which is expressed as the vector expression, $\overrightarrow{king} - \overrightarrow{man} + \overrightarrow{woman}$ producing a vector very close to $\overrightarrow{queen}$. Such behaviour makes word-embedding models an effective tool for a number of NLP applications such as identifying contextual synonyms, ranking keywords, and computing similarities between millions of documents (Botev *et al.*, 2017).

3.2.2 The Word2Vec Model

Given the large volume of publications to be processed in this study, we opt to use the `gensim` library (Řehůřek and Sojka, 2010) implementation of the Word2Vec Continuous Bag-of-Words (Word2Vec-CBOW) model for it's computational speed. The Word2Vec model is a simple architecture of two sets of embeddings where one encodes input words that should approximate output words encoded by the other. This will draw similar words together and push dissimilar words apart within the embedding space. The two variants, CBOW and skip-gram, swap the target word and context words as input and output. The CBOW model uses context words that appear in a fixed window around the target word as input to predict the target word. See Mikolov *et al.* (2013) for details about context word sampling, negative-word sampling and the loss-function used.

In order to obtain a model trained on the cross-domain scientific literature, we compile a generic corpus of 1.5 million open access English articles across all domains from the CORE service (Knoth and Zdrahal, 2012). To ensure the articles cover a diverse range of domains, we randomly select these articles out of a pool of 20 million English articles hosted by the CORE service with any of 24 keywords. The keywords used in our selection and the resultant numbers of articles are listed in Table 1.

Table 1: Genre keywords used to select articles in July 2021

Keyword	# of articles
”starburst” ∧ ”galaxy”	55k
”quantum” ∧ ”field” ∧ ”theory”	50k
”artificial” ∧ ”intelligence”	80k
”nanotechnology”	75k
”surface” ∧ ”science”	25k
”microbiome”	75k
”computer” ∧ ”architecture”	75k
”game” ∧ ”theory”	25k
”bioinformatics”	25k
”quantum” ∧ ”computation”	75k
”cognitive” ∧ ”behavioral” ∧ ”therapy”	40k
”obsessive” ∧ ”compulsive” ∧ ”disorder”	40k
”renaissance”	20k
”crystallography”	75k
”dielectric”	75k
”Ising”	25k
”Markov” ∧ ”chain”	75k
”thin” ∧ ”film”	75k
”cancer”	80k
”corrosion”	75k
”neural” ∧ ”network”	75k
”geochemistry”	15k
”HIV”	75k
”renewable” ∧ ”energy”	75k

Due to the resource limitation in this pilot study, we focus on the abstracts, which are usually concise but still contain essential information of the article. We then train a generic Word2Vec-CBOW model of 100 dimensions on this corpus for ten epochs with a window size of five: where the dimensionality is the size of the vector embedding for each word, one for input and one for output, and the epochs are the number of times the entire corpus is repeated during training, and the window size specifies the region around the target word from which context words can be selected. While hyperparameter optimization was not performed on this specific dataset, the choice of dimension was based on previous experiments (Yin, 2018) and the `gensim` library's recommended value for CBOW models. We observe that training over the corpus for ten epochs showed thorough convergence of validation loss. In the following section, we explain how this model is used in our analysis.

3.3 Extracting physical quantities from scientific documents

As the main objective of this study is to discover possible carriers of DIBs in the literature outside of the astrophysics domain, recognition of physical quantities is an essential step in the corpus processing. A physical quantity in the corpus is defined as a numeric value that appears with a physical unit, such as “1500 Å”. However, recognizing physical quantities from existing documents is not a trivial task. For example, MWC 349A, is a member of the double star system,

MWC 349 (Gvaramadze, V. V. and Menten, K. M., 2012), and should not be recognized as a physical quantity of 349 Å. Furthermore, there are numerous ways to denote one physical quantity. For example, 1500 Å and 0.15 μm are equivalent and should be recognized as such by the model.

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

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

We detail the unit-disambiguation algorithm in the following.

With the RegEx and unit-parsing algorithms listed above, we mask the identified physical quantities by numbers and units in the articles of the generic corpus. Physical quantities with ambiguous units are masked by all candidate units, for example, 1500 Å is masked as "NUM–Angstrom–Ampere".

Using this model, we compute a context vector, which is the average vector from a window of 5 words around the physical quantity excluding itself, for each physical quantity with an ambiguous unit. The choice of window size, 5, is to be consistent with the hyperparameters used in the Word2Vec-CBOW model training discussed previously. For each candidate unit assigned to the ambiguous unit, we then calculate the cosine similarity, which quantifies how parallel are two vectors, or in our case, how contextually-correlated are two words, between the calculated context vector to other units of the same dimension as the candidate unit. The highest cosine similarity is then assigned to the candidate unit as its score to represent the ambiguous unit.

IV TACKLING DIBS WITH NLP

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

4.1 Compiling the DIB corpus

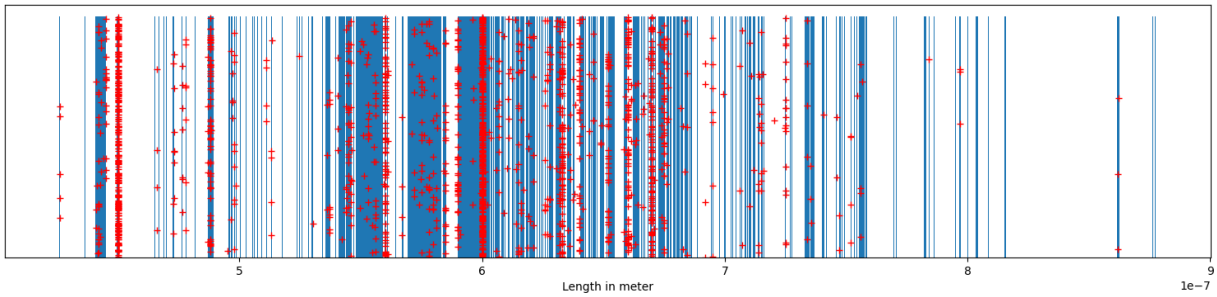


Figure 5: Indication of identified known DIBs (blue bands) and identified physical quantities from the generic corpus that overlap with DIBs (red crosses). The y-axis is in arbitrary unit to guide the eyes on how identified physical quantities distribute through the DIBs bands.

[3] https://pint.readthedocs.io/en/stable/

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

We first search for DIBs-related papers on NASA's Astrophysics Data System[4] and look for their relevant papers with Iris.ai's Explore Tool[5]. We then examine all articles appearing in the search and hand pick the articles that are directly DIBs-related. The final selected DIB corpus then includes 939 articles.We then search through these 939 articles to annotate all valid physical quantities in the texts that are directly associated with the DIBs catalog (Jenniskens and Desert, 1994)[6]. In total, there are 243 annotated wavelengths identified out of these articles. The physical quantity recognition algorithm described in Section 3.3 correctly identified 203 and missed 40 of them. The algorithm also wrongly identified 41 candidates that are not valid physical quantities. The precision and recall values are summarized in Table 2. These 243 annotated wavelengths, along with the full-width-half-maximum (FWHM), are marked as blue bands in Figure 5.

Table 2: Precision and recall for the physical quantity recognition algorithm.

precision	recall
203/244	203/243
83.2%	83.5%

As many tokens that are specific to the DIB corpus are infrequent in the generic corpus, in order to well represent these tokens, it is important to fine-tune the generic Word2Vec-CBOW model. The DIB corpus, with the disambiguated units, is then used to fine-tune the generic model. During the fine-tuning process, the vectors of tokens that appear in the DIB corpus are updated and thus displaced with the new context words. We observe that after training the generic model on the DIB corpus for ten epochs, the displacement of vectors stabilized. As demonstrated in Figure 6, in the fine-tuning process, vectors of tokens that are frequent in the DIB corpus and infrequent in the generic corpus are displaced the most while vectors of tokens that are infrequent in the DIB corpus yet frequent in the generic corpus are generally still.

4.2 Using the unit quantity recognition

We then apply the physical quantity recognition algorithm on the generic corpus. Out of the 1.5 million articles, we identify ~ 2000 physical quantities from $\sim 20,000$ articles, that overlap with DIBs and mark their locations as red crosses in Figure 5. Some of these physical quantities describe molecular sizes, which are unrelated to DIBs, but others describe the wavelength of the energy transitions and are thus of direct interest to us.

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

[4] https://ui.adsabs.harvard.edu/
[5] https://the.iris.ai/
[6] https://leonid.arc.nasa.gov/DIBcatalog.html

e-ISSN: 2430-3038, © JIMIS, CC BY
Volume: 11 - Year: 2023, DOI: 10.46298/jimis.9388

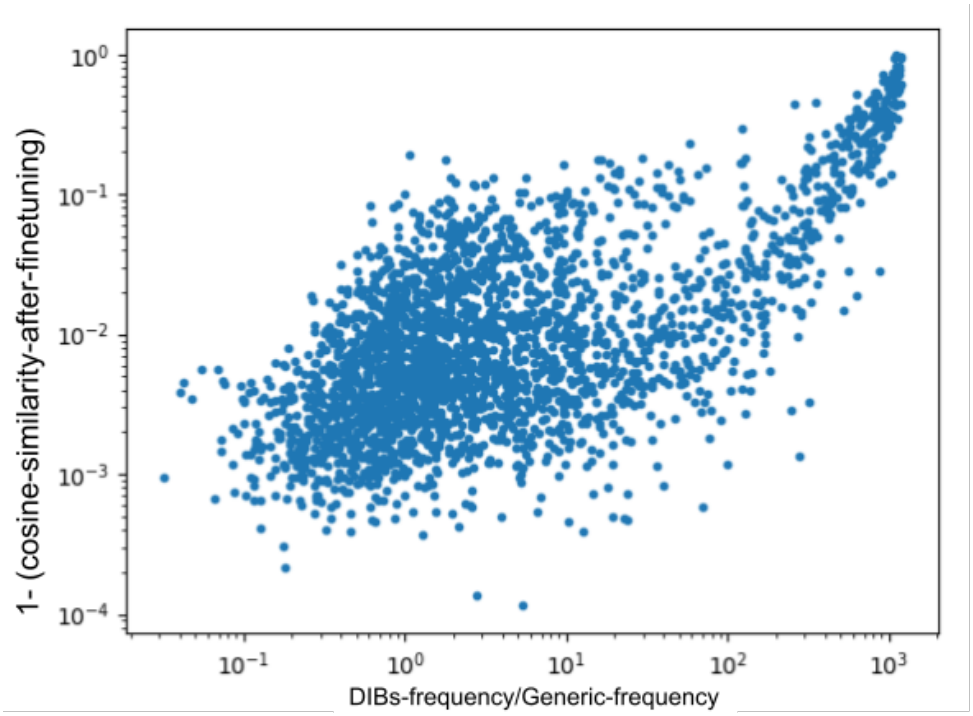


Figure 6: This graph shows how tokens specific to the DIB corpus vary before and after the fine-tuning. One can see that tokens that are frequent in the DIB corpus yet infrequent in the genric corpus displace the most, while tokens that are infrequent in the DIB corpus yet frequent in the generic corpus do not displace much.

1. In the DIBs wavelength range, $0.1 - 1\ \mu m$, physical quantities that are described in the unit of μm in the generic corpus are often found to be associated to diameters or distances, such as the DNA lengths. On the contrary, physical quantities used to describe wavelengths in this range often appear in units of angstrom or nanometer. As a consequence, we discard quantities that are identified within this range that appear in μm.
2. Known laser wavelengths are irrelevant for identifying DIB carriers so we discard any identified physical quantities that co-occur with the token "laser" or "light" within a window of 5 words.
3. Using the DIBs-enhanced Word2Vec-CBOW model in the sentence window as the context where a physical quantity is identified, we compute a cosine similarity between the context vector and the physical-quantity vector. Articles of low (< 0.5) cosine similarities are filtered out to remove false positives given by the unit-recognition algorithm.

After applying filters, we manually screen through all 1932 candidate articles and identify twelve papers of high relevance, which are discussed in the following section.

V ASSESSMENT OF THE RESULTS

Our model points us toward twelve articles presenting spectroscopic measurements of molecules having transitions corresponding to some DIBs. These findings are summarized in Table 3. They now need to be scrutinized.

5.1 The Results

Article	Transitions	Closest DIB	Molecule
Wakakuwa *et al.* (2010)	425 nm	425.90 ± 0.01 nm	11-cis retinal

Article	Transitions	Closest DIB	Molecule
Butterfly eye pigment	453 nm	450.18 ± 0.02 nm	chromophore
	563 nm	563.50 ± 0.01 nm	within opsin
	620 nm	619.90 ± 0.01 nm	(H, C, O)
	640 nm	640.05 ± 0.01 nm	
Dove *et al.* (1995)	560 nm	560.09 ± 0.01 nm	Chromophores
Coral pigment	580 nm	580.66 ± 0.01 nm	within
	590 nm	590.06 ± 0.01 nm	pocilloporin
Davies *et al.* (2009)	441.9 ± 1.0 nm	442.82 ± 0.17 nm	Chromophores
Elephant shark eye pigment	493.7 ± 2.6 nm	494.74 ± 0.01 nm	within
	496.3 ± 0.1 nm	496.39 ± 0.01 nm	proteins
	498.3 ± 0.3 nm	498.21 ± 0.01 nm	(H, C, N, O)
	498.7 ± 0.3 nm	498.74 ± 0.01 nm	
	504.1 ± 1.0 nm	505.48 ± 0.01 nm	
	509.5 ± 0.5 nm	509.21 ± 0.01 nm	
	510.1 ± 0.2 nm	510.10 ± 0.01 nm	
	520.9 ± 2.0 nm	521.79 ± 0.01 nm	
	534.2 ± 1.0 nm	534.25 ± 0.01 nm	
	547.8 ± 2.2 nm	548.08 ± 0.01 nm	
Spady *et al.* (2006)	423 nm	425.90 ± 0.01 nm	Retinal
Cichlid eye pigment	456 nm	450.18 ± 0.15 nm	chromophores
	472 nm	472.68 ± 0.02 nm	within
	518 nm	517.81 ± 0.01 nm	opsin
	528 nm	529.8 ± 0.01 nm	
	561 nm	560.98 ± 0.01 nm	
Wolfbeis *et al.* (2001)	488 nm	488.00 ± 0.01 nm	Valinomycin
Valinomycin			(H, C, N, O)
Filosa (2001)	695 nm	694.46 ± 0.01 nm	Amide II
Blood proteins			(H, C, N, O)
Davies *et al.* (2009)	501.0 ± 0.1 nm	500.36 ± 0.01 nm	Chromophores
Agnathan eye pigments	535.5 ± 3.3 nm	535.88 ± 0.01 nm	within
	544.1 ± 5.0 nm	543.35 ± 0.01 nm	opsin
	554.3 ± 2.0 nm	554.51 ± 0.01 nm	
	562.8 ± 0.4 nm	563.50 ± 0.01 nm	
Schoot Uiterkamp *et al.* (1976)	653 nm	653.65 ± 0.01 nm	Tyrosinase
Mushroom absorption	755 nm	755.94 ± 0.01 nm	(H, C, O, Cu)
Davies *et al.* (2007)	439 nm	436.39 ± 1.10 nm	Chromophores
Lamprey eye pigment	492 nm	494.74 ± 0.01 nm	
	497 nm	496.91 ± 0.01 nm	
Maréchal *et al.* (2007)	445 nm	442.82 ± 1.69 nm	Heme intermediate
Nitric-oxide synthase			(H, C, N, O, Fe)
Rémigy *et al.* (2003)	420 nm	425.90 ± 0.01 nm	Heme-type
Bacterium cytochrome	525.2 nm	525.18 ± 0.01 nm	molecule
	545.4 nm	545.06 ± 0.83 nm	(H, C, O, N, Fe)
Fasick *et al.* (1998)	488 nm	488.00 ± 0.12 nm	11-cis retinal
Dolphin eye pigment	545 nm	545.06 ± 0.83 nm	chromophore

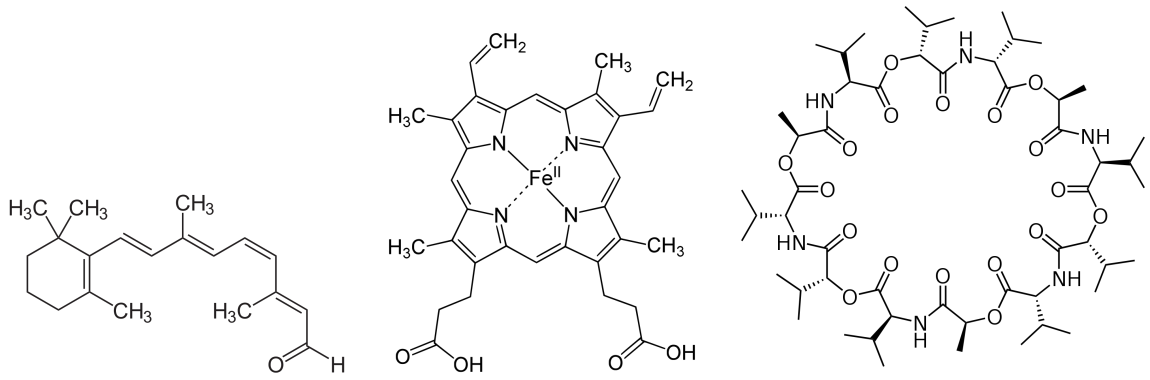
Article	Transitions	Closest DIB	Molecule

Table 3: *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, which is the median uncertainty of published transition quoted in Table 3. 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.

The twelve interdisciplinary articles listed in Table 3 are all biochemistry studies. They all report experimental spectroscopic measurements of organic molecules. Most of these molecules are constituted of abundant interstellar atoms, mainly, H, C, N and O. We can divide them into the two following categories.

5.1.1 Chromophores

A majority of the papers in Table 3 deal with animal retinal eye pigments (Fasick *et al.*, 1998; Spady *et al.*, 2006; Davies *et al.*, 2007, 2009; Wakakuwa *et al.*, 2010). These studies indeed measure the absorption bands of organic molecules in the visible range. Their potential relevance to DIBs is thus obvious. All these molecules are proteins (such as opsin) containing *chromophores*. Chromophores are molecular groupings, often part of a larger molecule, that are responsible for the color of an organism (*e.g.* Shukla *et al.*, 2017, for a review). It is thus reasonable to assume that they are responsible for the reported bands. A recurring chromophore in these studies is 11-cis retinal (Figure 7). Not all these papers discuss the actual chromophores present in their molecules, probably because they are not always known.



11-cis retinal Heme b Valinomycin

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

5.1.2 Heme and other molecules

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

5.2 Discussion

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

5.2.1 Accuracy of the reported transitions

The second column of Table 3 lists the central wavelengths of the bands reported by the articles found by our model. Unfortunately, only a few of these papers quote measurement errors on the centroid. This adds some uncertainty to our assessment. We assume that these errors are the last significant digit, that is 1 nm in most cases. Yet, we note that the median of the uncertainties in Table 3, when they are reported, is exactly 1 nm. Our assumption is thus realistic, assuming errors are comparable across all these studies. The third column of Table 3 gives the closest DIB from the measured band, using the Hobbs *et al.* (2009) compilation. Out of the 43 centroids, only 8 (19 %) do not have a DIB within $\pm 1\sigma$ (grey values).

The first question one can ask is how probable is it to draw a centroid wavelength at $\pm 1\sigma$ of a DIB. The compilation of Hobbs *et al.* (2009) contains 380 DIBs in the wavelength range $\lambda = 320 - 810$ nm. It gives a probability of 0.78 DIB per nm. Assuming that DIBs are randomly, uniformly distributed with this density, the Poissonian probability to find at least one DIB within $\pm 1\sigma$ (± 1 nm in our case) is 79 %. Matching bands is thus very likely. However, several studies list a large number of bands. The most spectacular is the paper by Davies *et al.* (2009), listing 11 bands with their uncertainties. It happens that all of them coincide with a DIB within $\pm 1\sigma$. Using the uncertainties quoted by Davies *et al.* (2009) and assuming they are normally distributed and independent, the probability to randomly obtain such a result is 0.2 %. These results are thus non trivial.

We have mentioned in Sect. II that DIBs were a few Angstrom wide. The bands reported in these articles are however wider (up to several tens of nm). One can thus wonder if they are relevant to our problem. The answer to this question is not straightforward. Molecular band width depends on the molecular size. Our candidate molecules contain a few tens to a hundred atoms, which is the expected size of DIB carriers (MacIsaac *et al.*, 2022). If these molecules are however bonded in a larger matrix, such as a protein, the line widths will surely be broadened with respect to an isolated molecule. This will be because, unless the material is very pure and well structured, the local environment of the chromophore will be varied thus giving to slight spread in line positions due to different (steric or other) interactions for each molecule.

5.2.2 Insights for astrochemistry

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

Our highest-score molecules, chromophores, have been proposed as DIB carriers (*e.g.* Johnson, 2006; Adams and Oka, 2019), and this is probably why our NLP model has selected them. Chromophores are probably too brittle molecules to survive in the diffuse ISM. For instance, the long chain of 11-cis retinal (Figure 7) will likely be photolyzed by UV photons. However, chromophores could form moieties in larger molecules or, possibly, in nanometer-size dust grains (Jones, 2014, 2016b). This would allow them to be present in the ISM and carry some of

the DIBs. We note this is also the case in the biochemistry studies. Several of the found papers mention that the chromophores are covalently bonded with their proteins.

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

5.3 Limitation of the methodology and prospects for interdisciplinarity

This pilot study aims to demonstrate a proof of concept on using NLP as a tool for users to identify cross-domain knowledge. We have shown that it could bring down the borders between different disciplines and perform a rather unbiased search for relevant information. Our uncovering of meaningful classes of molecules, studied in biochemistry, that could provide insights into an astrophysical question shows the practical relevance of NLP for interdisciplinary studies. Could this method, applied to the particular problem of the origin of DIBs, be applied to other open questions? There are no reason why this could not be the case. The method is however not completely straightforward and will need to be adapted to each new problem and its specificity: what is the size of the specialized corpus, are numerical quantities instrumental, and so on? Fully automatizing the search in order to release an open source software is thus a challenge for future studies.

As the nature of our research question is closely associated to the DIBs wavelengths, these physical quantities are thus the ideal lampposts guiding us in the literature outside of the astrophysics domain. To reproduce such process for research questions that are associated to some semantic concepts, an embeddings space of higher precision is crucial. It is thus a natural next step to explore how one can disambiguate word senses in the embeddings space. It is also worthwhile to explore other domain-adaptation techniques, such as the graph-based latent semantic imputation (Yao *et al.*, 2019), to obtain an embeddings space with higher precision in the domain of interest.

As mentioned in Section 3.2.2, the study is based on a generic Word2Vec-CBOW model that is trained on 1.5 million English articles selected with some specific keywords. These manually selected keywords inevitably create some domain biases in the generic corpus. This can be further improved by randomizing our search across the entire CORE database, which contains more than 500 million articles. Furthermore, we include only abstracts in our analysis in this study due to the limitation of resources. A lot of important contextual information written in the full text is therefore absent from our analysis. To thoroughly cover all information in the generic and DIB corpus, it would be essential to include the full text information in the future study.

Another limitation of the demonstrated method in this work is the language. As the curated DIB corpus includes only English articles, for research questions that are in low-resource languages, it would be difficult to adopt such methodology in cross-domain literature search.

VI SUMMARY AND CONCLUSION

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

Our main objectives, the feasibility to use NLP to address open scientific questions, is reached. We have shown that NLP is able to surface candidate DIBs molecules from an interdisciplinary corpus of scientific literature. This confirms that NLP-methodologies can generate plausible and non-trivial hypotheses for future investigation. First, we have shown that some associations between the reported transitions and the DIBs would be unlikely if they were purely random. Second, our NLP model has pointed us toward molecules relevant to the *InterStellar Medium* (ISM). These molecules are constituted of abundant interstellar atoms. In addition, several of these molecules have been proposed in the past to be DIB carriers. Their presence in the ISM has not been proven, yet, but it is conjectured that they could form moieties in interstellar grains. In light of our study, NLP thus appears as a practical tool for interdisciplinarity. We have shown that this computational linguistic technique could uncover information in the biochemistry literature that happen to be relevant to astrophysics. We have also discussed how this technique could be applied to other open questions in an interdisciplinary fashion. In the future, extending our analysis to the full text, and expanding the generic corpus to thoroughly-cover cross-domain literature, can potentially give us a more complete result. From an astrophysical point of view, our work adds credibility to the possibility that some DIBs could be carried by chromophores and to the role of heme and porphyrins in interstellar chemistry. These findings need now to be further investigated in the laboratory, in interstellar conditions. As a further outlook, our methodology should be applicable to other open scientific questions which require interdisciplinary knowledge. More generally, our study shows it is possible to retrieve universal concepts from multiple disciplines. An interesting prospective would be to look for hidden interdisciplinary concepts from related universal concepts used in multiple domains.

Acknowledgements

We thank Anthony Jones for a useful discussion about the molecules found by our model and about DIBs in general, and Jason Hoelscher-Obermaier for constructive comments on the prospects of this study. We also thank the two anonymous referees for their comments which improved the quality and clarity of this article. This paper was supported by funding from NRC, Project ID: 309594, the AI Chemist under the collaboration of Iris.ai AS with Dr. Frédéric GALLIANO covering the contributions of Viktor BOTEV, Jeremy MINTON, Ronin WU, and partly Corentin VAN DEN BROEK D'OBRENAN. This work was supported by the Programme National "Physique et Chimie du Milieu Interstellaire" (PCMI) of the CNRS/INSU with INC/INP co-funded by CEA and CNES.

References

Adams K., Oka T. (2019, December). Relating the Carriers of λ5797.1 Diffuse Interstellar Band and λ5800 Red Rectangle Band. *ApJ 886*(2), 138. doi:10.3847/1538-4357/ab4c49.

Asplund M., Grevesse N., Sauval A. J., Scott P. (2009, September). The Chemical Composition of the Sun. *ARA&A 47*, 481–522. arXiv:0909.0948, doi:10.1146/annurev.astro.46.060407.145222.

Belloche A., Garrod R. T., Müller H. S. P., Menten K. M. (2014, September). Detection of a branched alkyl molecule in the interstellar medium: iso-propyl cyanide. *Science 345*(6204), 1584–1587. arXiv:1410.2607, doi:10.1126/science.1256678.

Bianchi S., De Vis P., Viaene S., Nersesian A., Mosenkov A. V., Xilouris E. M., Baes M., Casasola V., Cassarà L. P., Clark C. J. R., Davies J. I., De Looze I., Dobbels W., Galametz M., Galliano F., Jones A. P., Lianou S., Madden S. C., Trčka A. (2018, December). Fraction of bolometric luminosity absorbed by dust in DustPedia galaxies. *A&A 620*, A112. arXiv:1810.01208, doi:10.1051/0004-6361/201833699.

Bojanowski P., Grave E., Joulin A., Mikolov T. (2017, 06). Enriching Word Vectors with Subword Information. *Transactions of the Association for Computational Linguistics 5*, 135–146. URL: https://doi.org/10.1162/tacl_a_00051, arXiv:https://direct.mit.edu/tacl/article-pdf/doi/10.1162/tacl_a_00051/1567442/tacl_a_00051.pdf, doi:10.1162/tacl_a_00051.

Botev V., Marinov K., Schäfer F. (2017). Word importance-based similarity of documents metric (wisdm): Fast and scalable document similarity metric for analysis of scientific documents. In *Proceedings of the 6th International Workshop on Mining Scientific Publications*, WOSP 2017, pp. 17. ACM. URL: http://doi.acm.org/10.1145/3127526.3127530, doi:10.1145/3127526.3127530.

Bron E. (2014, November). *Stochastic processes in the interstellar medium*. Ph. D. thesis, LERMA, Observatoire de Paris, PSL Research University, CNRS, Sorbonne Universités, UPMC Univ. Paris 06, F-92190, Meudon, France.

Cami J., Bernard-Salas J., Peeters E., Malek S. E. (2010, September). Detection of C_{60} and C_{70} in a Young Planetary Nebula. *Science 329*(5996), 1180. doi:10.1126/science.1192035.

Campbell E. K., Holz M., Gerlich D., Maier J. P. (2015, July). Laboratory confirmation of C_{60}^+ as the carrier of two diffuse interstellar bands. *Nature 523*, 322–323. doi:10.1038/nature14566.

Davies W. L., Carvalho L. S., Tay B. H., Brenner S., Hunt D. M., Venkatesh B. (2009, April). Into the blue: gene duplication and loss underlie color vision adaptations in a deep-sea chimaera, the elephant shark callorhinchus milii. *Genome research 19(3)*, 415–426. doi:10.1101/gr.084509.108.

Davies W. L., Collin S. P., Hunt D. M. (2009, August). Adaptive gene loss reflects differences in the visual ecology of basal vertebrates. *Molecular biology and evolution 26(8)*, 1803–1809. doi:10.1093/molbev/msp089.

Davies W. L., Cowing J. A., Carvalho L. S., Potter I. C., Trezise A. E., Hunt D. M., Collin S. P. (2007). Functional characterization, tuning, and regulation of visual pigment gene expression in an anadromous lamprey. *FASEB journal : official publication of the Federation of American Societies for Experimental Biology 21(11)*, 2713–2724. doi:10.1096/fj.06-8057com.

Densen P. (2011). Challenges and opportunities facing medical education. *Transactions of the American Clinical and Climatological Association*.

Devlin J., Chang M.-W., Lee K., Toutanova K. (2019, June). BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, Minneapolis, Minnesota, pp. 4171–4186. Association for Computational Linguistics. URL: https://aclanthology.org/N19-1423, doi:10.18653/v1/N19-1423.

Dove S. G., Takabayashi M., Hoegh-Guldberg O. (1995). Isolation and partial characterization of the pink and blue pigments of pocilloporid and acroporid corals. *The Biological bulletin 189(3)*, 288—-297. doi:10.2307/1542146.

Draine B. T. (2003). Interstellar Dust Grains. *ARA&A 41*, 241–289. arXiv:arXiv:astro-ph/0304489, doi:10.1146/annurev.astro.41.011802.094840.

Draine B. T. (2011). *Physics of the Interstellar and Intergalactic Medium*. Princeton University Press.

Fan H., Hobbs L. M., Dahlstrom J. A., Welty D. E., York D. G., Rachford B., Snow T. P., Sonnentrucker P., Baskes N., Zhao G. (2019, June). The Apache Point Observatory Catalog of Optical Diffuse Interstellar Bands. *ApJ 878*(2), 151. arXiv:1905.05962, doi:10.3847/1538-4357/ab1b74.

Fasick J. I., Cronin T. W., Hunt D. M., Robinson P. R. (1998). The visual pigments of the bottlenose dolphin (tursiops truncatus). *Visual neuroscience 15(4)*, 643–651. doi:10.1017/s0952523898154056.

Filosa A. (2001). *Study of locally unfolded forms of cytochrome c by Fourier transform infrared and of H2O2-mediated oxidation of ferricytochrome c and metmyoglobin by mass spectrometry*. Ph. D. thesis, Concordia University. Unpublished. URL: https://spectrum.library.concordia.ca/id/eprint/1417/.

Firth J. R. (1957). A synopsis of linguistic theory, 1930-1955.

Galliano F. (2022, February). A Nearby Galaxy Perspective on Interstellar Dust Properties and their Evolution. *Habilitation Thesis*, 1. arXiv:2202.01868.

Galliano F., Galametz M., Jones A. P. (2018, September). The Interstellar Dust Properties of Nearby Galaxies. *ARA&A 56*, 673–713. doi:10.1146/annurev-astro-081817-051900.

Goodfellow I., Bengio Y., Courville A. (2016). *Deep Learning*. MIT Press. http://www.deeplearningbook.org.

Grezes F., Blanco-Cuaresma S., Accomazzi A., Kurtz M. J., Shapurian G., Henneken E., Grant C. S., Thompson D. M., Chyla R., McDonald S., Hostetler T. W., Templeton M. R., Lockhart K. E., Martinovic N., Chen S., Tanner C., Protopapas P. (2021, December). Building astroBERT, a language model for Astronomy & Astrophysics. *arXiv e-prints*, arXiv:2112.00590. arXiv:2112.00590.

Gvaramadze, V. V., Menten, K. M. (2012). Discovery of a parsec-scale bipolar nebula around mwca. *A&A 541*, A7. URL: https://doi.org/10.1051/0004-6361/201218841, doi:10.1051/0004-6361/201218841.

Hastings J., Magka D., Batchelor C., Duan L., Stevens R., Ennis M., Steinbeck C. (2012, April). Structure-based classification and ontology in chemistry. *Journal of Cheminformatics 4*, 8. doi:10.1186/1758-2946-4-8.

Hearnshaw J. B. (2014). *The interpretation of stellar spectra and the birth of astrophysics* (2 ed.)., pp. 127–151. Cambridge University Press. doi:10.1017/CBO9781139382779.009.

Heger M. L. (1922). Further study of the sodium lines in class B stars . *Lick Observatory Bulletin 10*, 141–145. doi:10.5479/ADS/bib/1922LicOB.10.141H.

Herbig G. H. (1995, January). The Diffuse Interstellar Bands. *ARA&A 33*, 19–74. doi:10.1146/annurev.aa.33.090195.000315.

Herbst E., van Dishoeck E. F. (2009, September). Complex Organic Interstellar Molecules. *ARA&A 47*(1), 427–480. doi:10.1146/annurev-astro-082708-101654.

Herzberg G. (1988, June). Historical Remarks on the Discovery of Interstellar Molecules. *JRASC 82*, 115.

Hobbs L. M., York D. G., Thorburn J. A., Snow T. P., Bishof M., Friedman S. D., McCall B. J., Oka T., Rachford B., Sonnentrucker P., Welty D. E. (2009, November). Studies of the Diffuse Interstellar Bands. III. HD 183143. *ApJ 705*(1), 32–45. arXiv:0910.2983, doi:10.1088/0004-637X/705/1/32.

Jenniskens P., Desert F.-X. (1994, July). A survey of diffuse interstellar bands (3800-8680 A). *A&AS 106*.

Johnson F. M. (1994, May). Porphyrins in the interstellar medium (in grains). In A. G. G. M. Tielens (Ed.), *The Diffuse Interstellar Bands*, pp. 47–52.

Johnson F. M. (2006, December). Diffuse interstellar bands: A comprehensive laboratory study. *Spectrochimica Acta Part A: Molecular Spectroscopy 65*, 1154–1179. arXiv:1706.04273, doi:10.1016/j.saa.2006.03.004.

Jones A. P. (2014, October). A framework for resolving the origin, nature and evolution of the diffuse interstellar band carriers? *Planetary and Space Science 100*, 26–31. arXiv:1411.5854, doi:10.1016/j.pss.2013.11.011.

Jones A. P. (2016a, December). Dust evolution, a global view I. Nanoparticles, nascence, nitrogen and natural selection ... joining the dots. *Royal Society Open Science 3*, 160221. doi:10.1098/rsos.160221.

Jones A. P. (2016b, December). Dust evolution, a global view: II. Top-down branching, nanoparticle fragmentation and the mystery of the diffuse interstellar band carriers. *Royal Society Open Science 3*, 160223. doi:10.1098/rsos.160223.

Kerzendorf W. E. (2019, June). Knowledge discovery through text-based similarity searches for astronomy literature. *Journal of Astrophysics and Astronomy 40*(3), 23. arXiv:1705.05840, doi:10.1007/s12036-019-9590-5.

Knoth P., Zdrahal Z. (2012). Core: three access levels to underpin open access. *D-Lib Magazine 18*(11/12). URL: http://oro.open.ac.uk/35755/.

Lamurias A., Sousa D., Clarke L. A., Couto F. M. (2019, January). BO-LSTM: classifying relations via long 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.

Lan T.-W., Ménard B., Zhu G. (2015, October). Exploring the diffuse interstellar bands with the Sloan Digital Sky Survey. *MNRAS 452*(4), 3629–3649. arXiv:1406.7284, doi:10.1093/mnras/stv1519.

MacIsaac H., Cami J., Cox N. L. J., Farhang A., Smoker J., Elyajouri M., Lallement R., Sarre P. J., Cordiner M. A., Fan H., Kulik K., Linnartz H., Foing B. H., van Loon J. T., Mulas G., Smith K. T. (2022, March). The EDIBLES survey V: Line profile variations in the $\lambda\lambda$5797, 6379, and 6614 diffuse interstellar bands as a tool to constrain carrier sizes. *arXiv e-prints*, arXiv:2203.01803. arXiv:2203.01803.

Manning C. D., Schütze H. (1999). *Foundations of Statistical Natural Language Processing*. Cambridge, Massachusetts: The MIT Press. URL: http://nlp.stanford.edu/fsnlp/.

Maréchal A., Mattioli T. A., Stuehr D. J., Santolini J. (2007). Activation of peroxynitrite by inducible nitric-oxide synthase: a direct source of nitrative stress. *The Journal of biological chemistry 282(19)*, 14101–14112. doi:10.1074/jbc.M609237200.

McGuire B. A. (2018, December). 2018 Census of Interstellar, Circumstellar, Extragalactic, Protoplanetary Disk, and Exoplanetary Molecules. *ApJS 239*(2), 17. arXiv:1809.09132, doi:10.3847/1538-4365/aae5d2.

McGuire B. A., Loomis R. A., Burkhardt A. M., Lee K. L. K., Shingledecker C. N., Charnley S. B., Cooke I. R., Cordiner M. A., Herbst E., Kalenskii S., Siebert M. A., Willis E. R., Xue C., Remijan A. J., McCarthy M. C. (2021, March). Detection of two interstellar polycyclic aromatic hydrocarbons via spectral matched filtering. *Science 371*(6535), 1265–1269. arXiv:2103.09984, doi:10.1126/science.abb7535.

Merrill P. W. (1934, August). Unidentified Interstellar Lines. *PASP 46*(272), 206–207. doi:10.1086/124460.

Mikolov T., Chen K., Corrado G., Dean J. (2013). Efficient estimation of word representations in vector space. *CoRR abs/1301.3781*. URL: http://arxiv.org/abs/1301.3781.

Mikolov T., Yih W.-t., Zweig G. (2013, June). Linguistic regularities in continuous space word representations. In *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Atlanta, Georgia, pp. 746–751. Association for Computational Linguistics. URL: https://aclanthology.org/N13-1090.

Pennington J., Socher R., Manning C. (2014, October). GloVe: Global vectors for word representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, Doha, Qatar, pp. 1532–1543. Association for Computational Linguistics. URL: https://aclanthology.org/D14-1162, doi:10.3115/v1/D14-1162.

Peters M. E., Neumann M., Iyyer M., Gardner M., Clark C., Lee K., Zettlemoyer L. (2018). Deep contextualized word representations. *CoRR abs/1802.05365*. URL: http://arxiv.org/abs/1802.05365, arXiv:1802.05365.

Řehůřek R., Sojka P. (2010, May). Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, Valletta, Malta, pp. 45–50. ELRA. http://is.muni.cz/publication/884893/en.

Rémigy H. W., Aivaliotis M., Ioannidis N., Jenö P., Mini T., Engel A., Jaquinod M., Tsiotis G. (2003). Characterization by mass spectroscopy of a 10 kda c-554 cytochrome from the green sulfur bacterium chlorobium tepidum. *Photosynthesis research 78(2)*, 153–160. doi:10.1023/B:PRES.0000004347.34228.2c.

Schoot Uiterkamp A. J., Evans L. H., Jolley R. L., Mason H. S. (1976). Absorption and circular dichroism spectra of different forms of mushroom tyrosinase. *Biochimica et biophysica acta 453(1)*, 200–204. doi:10.1016/0005-2795(76)90264-6.

Shukla R., Dubey A., Pandey V., Golhani D., Jain A. (2017, 07). Chromophore- an utility in uv spectrophotometer. *Inventi Rapid: Pharm Ana & Qual Assur*.

Spady T. C., Parry J. W., Robinson P. R., Hunt D. M., Bowmaker J. K., Carleton K. L. (2006). Evolution of the cichlid visual palette through ontogenetic subfunctionalization of the opsin gene arrays. *Molecular biology and evolution 23(8)*, 1538–1547. doi:10.1093/molbev/msl014.

Thomas B., Thronson H., Buonomo A., Barbier L. (2022, January). Determining Research Priorities for Astronomy Using Machine Learning. *Research Notes of the American Astronomical Society 6*(1), 11. arXiv:2203.00713, doi:10.3847/2515-5172/ac4990.

Thorburn J. A., Hobbs L. M., McCall B. J., Oka T., Welty D. E., Friedman S. D., Snow T. P., Sonnentrucker P.,

York D. G. (2003, February). Some Diffuse Interstellar Bands Related to Interstellar C_2 Molecules. *ApJ 584*(1), 339–356. doi:10.1086/345665.

Tielens A. G. G. M. (2005, September). *The Physics and Chemistry of the Interstellar Medium*. Cambridge University Press.

Tielens A. G. G. M. (2008, September). Interstellar Polycyclic Aromatic Hydrocarbon Molecules. *ARA&A 46*, 289–337. doi:10.1146/annurev.astro.46.060407.145211.

Wakakuwa M., Terakita A., Koyanagi M., Stavenga D., Shichida Y., Arikawa K. (2010, nov). Evolution and mechanism of spectral tuning of blue-absorbing visual pigments in butterflies. *PLoS One 5(11):e15015*. doi:10.1371/journal.pone.0015015.

Walker G. A. H., Bohlender D. A., Maier J. P., Campbell E. K. (2015, October). Identification of More Interstellar C_{60}^+ Bands. *ApJL 812*(1), L8. arXiv:1509.06818, doi:10.1088/2041-8205/812/1/L8.

Wolfbeis O. S., Opitz D., Werner T., Ouart A. (2001). Chiroptic recognition of potassium ion. *Journal of molecular recognition 14(1)*, 13–17. doi:10.1002/1099-1352(200101/02)14:1¡13::AID-JMR514¿3.0.CO;2-K.

Yao S., Yu D., Xiao K. (2019, July). Enhancing Domain Word Embedding via Latent Semantic Imputation. *Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 557–565. arXiv: 1905.08900. URL: http://arxiv.org/abs/1905.08900, doi:10.1145/3292500.3330926.

Yin Z. (2018, May). Understand Functionality and Dimensionality of Vector Embeddings: the Distributional Hypothesis, the Pairwise Inner Product Loss and Its Bias-Variance Trade-off. *arXiv:1803.00502 [cs, stat]*. arXiv: 1803.00502. URL: http://arxiv.org/abs/1803.00502.