**IRFU : Institut de recherche sur les lois fondamentales de l'univers**

**Saclay**

DAp/LFEMI

# Traitement automatique du langage naturel pour explorer la nature des énigmatiques bandes diffuses interstellaires

**Spécialité** Astrophysique

**Niveau d'étude** Bac+4/5

**Formation** Ingenieur/Master

**Unité d'accueil** DAp/LFEMI

**Candidature avant le** 12/07/2021

**Durée** 6 mois

**Poursuite possible en thèse** oui

**Contact** GALLIANO Frederic
+33 1 69 08 18 21
frederic.galliano@cea.fr

**Autre lien** https://iris.ai/

## Résumé

Les DIBs sont des bandes en absorption, omniprésentes dans le milieu interstellaires, dont l'origine reste un mystère, près d'un siècle après leur découverte. Le but du stage est d'utiliser la technique du traitement automatique du langage, afin de trouver dans la littérature scientifique des composés moléculaires permettant d'expliquer certaines de ces DIBs.

## Sujet détaillé

Le Milieu InterStellaire (MIS) est une intrication complexe de phases remplissant le volume d'une galaxie entre les étoiles. Il est constitué de: (i) gaz, principalement d'hydrogène (75%) et d'hélium (23%), mais aussi d'éléments plus lourds (C, N, O, etc.; 1%) qui peuvent se trouver sous forme moléculaire; et (ii) de grains de poussières (le 1% restant de la masse), qui sont de petites particules solides de tailles inférieures au micron. Le MIS est un constituant fondamental de l'Univers, puisque les étoiles naissent de l'effondrement de nuages interstellaires denses, et retournent une partie de leur masse, enrichie en éléments fraîchement synthétisés, à la fin de leur vie. Le MIS est aussi le milieu ou la complexité moléculaire croît, fournissant les briques élémentaires pour l'émergence de la vie. Le contenu moléculaire du MIS est actuellement mal connu. Les premières molécules complexes n'ont commencé à être détectées que depuis une décennie.

Les Bandes Diffuses Interstellaires (DIBs) sont des bandes en absorption omniprésentes, dans le domaine spectrale entre 0.4 et 2 microns. Leurs propriétés spectrales correspondent à celles de grosses molécules ou de petits solides, mais leur nature exacte reste un mystère, près d'un siècle après leur découverte par Heger (1922). Plus de 400 DIBs distinctes ont été détectées dans le MIS et un plus grand nombre est en train d'être observé par le télescope spatial GAIA (2013-présent). À l'heure actuelle, seulement 2 DIBs ont été identifiées de manière claire et ont été attribuées au buckminsterfullerene ionisé (C_{60}^{+}), un hydrocarbure aromatique polycyclique en forme de ballon de football, il y a 5 ans (Campbell et al. 2015). Dévoiler la nature des DIBs est une question ouverte importante, car cela

représenterait une clef pour accéder au contenu moléculaire et poussiéreux du MIS de manière plus précise. Cela fournirait également des diagnostics des conditions physiques locales.

Les approches astrophysiques et astrochimiques traditionnelles ont échoué à aborder cette question ouverte. La raison probable est que ce problème touche la complexité chimique. Il est par conséquent crucial d'explorer des approches non conformistes qui permettraient une synergie entre notre connaissance des DIBs et la recherche dans d'autres domaines, tels que la science des matériaux, la chimie ou la biochimie.

Afin d'explorer de telles synergies, nous tirerons profit des techniques d'apprentissage automatique du langage (NLP). Le projet est de lancer des recherches systématiques, à grande échelle, dans les publications mondiales de chimie. Le NLP est une application de l'apprentissage automatique dont le but est de comprendre du texte écrit. Notre espoir est de trouver des publications présentant des transitions moléculaires pouvant expliquer certaines DIBs.

La première étape de ce projet est d'injecter la connaissance des DIBs, depuis un corpus hautement spécialisé, dans une intégration sémantique entraînée sur un grand corpus de littérature scientifique. La seconde étape sera d'explorer les relations sémantiques entre les entités étiquetées sur un prototype ontologique des DIBs, et comprendre comment enrichir l'ontologie en retour. Finalement, nous construirons un graphe de connaissance des DIBs qui nous aidera à trouver des propriétés chimiques potentiellement non explorées, pertinentes aux DIBs.

Le but de ce stage est de réaliser une étude de faisabilité sur la pertinence du NLP pour s'attaquer au problème de l'origine des DIBs. Le stagiaire travaillera sous la supervision du Dr. Frédéric GALLIANO du Département d'Astrophysique du CEA-Saclay, et du Dr. Ronin WU de la compagnie IRIS.AI, qui développe un modèle NLP unique, afin d'explorer la littérature chimique.

**Mots clés**

Apprentissage automatique, chimie

**Compétences**

- Astrophysique élémentaire - Physique moléculaire - Apprentissage automatique

**Logiciels**

Python

## Using Natural Language Processing to Unveil the Nature of the Enigmatic Diffuse Interstellar Bands

### Summary

DIBs are ubiquitous absorption bands in the interstellar medium, whose origin is still a mystery, one century after their discovery. The goal of this internship is to use the technics of natural language processing, in order to find molecular compounds, in the scientific literature, that could explain some of these DIBs.

### Full description

The InterStellar Medium (ISM) is the complex intertwining of phases filling the volume of a galaxy between the stars. It is constituted of: (i) gas, principally hydrogen (75%) and helium (23%), but also heavier elements (C, N, O, etc.; 1%) that can be found in molecular forms; and (ii) dust grains (the remaining 1% of the ISM mass), which are small solid particles of sub-micronic sizes. The ISM is a fundamental constituent of the Universe, as stars are born out of the collapse of dense interstellar clouds, and return some of their mass, enriched in freshly synthesized heavy elements, at the end of their lifetime. The ISM is also the medium where the molecular complexity grows, providing the basic building blocks for the emergence of life. The molecular content of the ISM is currently poorly known, complex interstellar molecules have been detected for the first time, a decade ago.

Diffuse Interstellar Bands (DIBs) are ubiquitous absorption bands found preferentially in the 0.4 to 2 microns wavelegnth range. Their spectral properties correspond to those of large molecules or small solids, but their exact nature is still a mystery, a century after their discovery by Heger (1922). More than 400 distinct DIBs have been detected in the ISM and many more are currently observed by the GAIA space telescope (2013-present). Only 2 of them have been unambiguously identified and attributed to ionized buckminsterfullerene ($C_{60}^{+}$), a football-shaped polycyclic aromatic carbon compound, 5 years ago (Campbell et al. 2015). Unveiling the nature of DIBs is an important open question as it could help us unlock the ISM molecular and dusty content. It could also provide diagnostics of the local physical conditions where they are found.

Traditional astrophysical and astrochemical approaches have failed at tackling this open question. The reason is probably that it involves chemical complexity. It is therefore crucial to explore out-of-the-box methods that could synergize our current knowledge on DIBs with research in other domains, such as material science, biochemistry, or chemistry.

To explore such synergies, we will take advantage of the techniques of Natural Language Processing (NLP), in order to perform systematic, large-scale researches in the world-wide chemistry literature. NLP is a machine-learning application aimed at automatically understanding written text. Our hope is to find publications reporting molecular transitions that could explain some of the DIBs.

The first step of this project aims to inject the knowledge of DIBs from a highly specialized corpus into a general word embedding model trained on a large diverse corpus of scientific literature. The second step would be to explore the semantic relationships between entities labeled on a prototype ontology of DIBs, and how they can in turn enrich the ontology. Finally, we aim to build a knowledge graph of DIBs which can help indicate a potentially unexplored chemical properties of DIBs.

The goal of this internship is to perform a feasability study of the relevance of NLP to tackle the origin of DIBs. The intern will work under the supervision of both Dr. Frédéric GALLIANO of the Department of Astrophysics at CEA-Saclay, and Dr. Ronin WU of the IRIS.AI company developing a unique NLP model to explore the chemistry literature.

### Keywords

Machine learning, chemistry

**Skills**

- Elementary astrophysics - Molecular physics - Machine learning

**Softwares**

Python