## UE: "Méthodologies de la recherche"

Master 2: Machine Learning for Data Science
University of Paris

---

# ChemNLP: A Natural Language Processing based Library for Materials Chemistry Text Data

---

**Encadré par :** Mr. Nadif Mohamed


**Membres du groupe :**

- Belkaid Meryem
- Oumghar Abir
- Boanani Hafsa


**Class of :** 2023-2024

# Table des matières

# Introduction :

Au croisement de la chimie des matériaux et de l'intelligence artificielle, le Traitement du Langage Naturel (NLP) se révèle être un outil précieux pour aborder les défis inhérents à la gestion et à l'analyse des vastes corpus de textes scientifiques. L'importance de ***l'application du NLP dans le domaine de la chimie des matériaux*** réside dans sa capacité à transcender les limites traditionnelles de l'exploration textuelle, offrant ainsi des perspectives novatrices pour la recherche et le développement de matériaux avancés.

Les implications du NLP dans l'analyse de textes scientifiques dépassent largement la simple automatisation des processus de recherche. En effet, le NLP ouvre la voie à des applications variées, allant de la *classification et du regroupement de textes* à la *génération de résumés abstraits*, en passant par la reconnaissance d'entités nommées à grande échelle. Dans le domaine spécifique de la chimie des matériaux, où les articles regorgent de terminologies techniques et de méthodologies complexes, le NLP offre la possibilité d'extraire des informations pertinentes, de classer les textes en fonction de leurs caractéristiques spécifiques, et même de générer des résumés concis facilitant l'accès à la connaissance.

Cependant, cette alliance entre le NLP et la chimie des matériaux n'est pas sans défis. Les spécificités du langage scientifique, la présence fréquente de termes techniques complexes, et la nécessité de résoudre des dépendances entre les concepts au sein de textes spécialisés posent des défis uniques pour les outils traditionnels de NLP. L'accès à des ensembles de données complets, la standardisation des méthodes d'évaluation, et la résolution des problèmes de dépendance croisée entre différents domaines de la chimie des matériaux constituent des obstacles qui nécessitent une attention particulière.

Dans cette perspective, notre étude explore **le potentiel du NLP dans le domaine de la chimie des matériaux**, en mettant en lumière ses applications novatrices tout en abordant de manière critique les défis spécifiques qui jalonnent cette voie prometteuse de recherche.

# État de l’art en NLP pour la chimie des matériaux:

L'évolution de l'application du Traitement du Langage Naturel (NLP) dans le domaine de la chimie des matériaux a été marquée par des percées notables, avec des outils tels que ChemDataExtractor et Chemical Tagger  occupant une place prépondérante dans cette transition.
Ces deux outils, chacun dans son domaine d'expertise, ont ainsi joué un rôle crucial dans le progrès de la recherche et de l'exploration scientifique en chimie des matériaux.

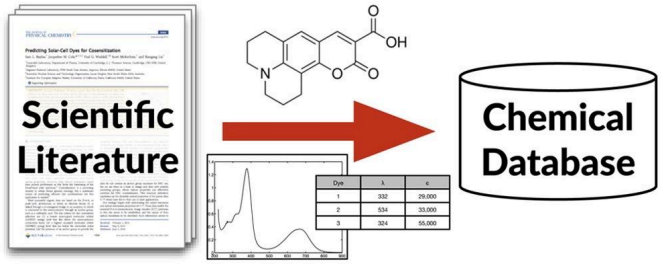



**Figure 1: Extraction de données de la littérature scientifique pour l'alimentation d'une base de données chimique"**

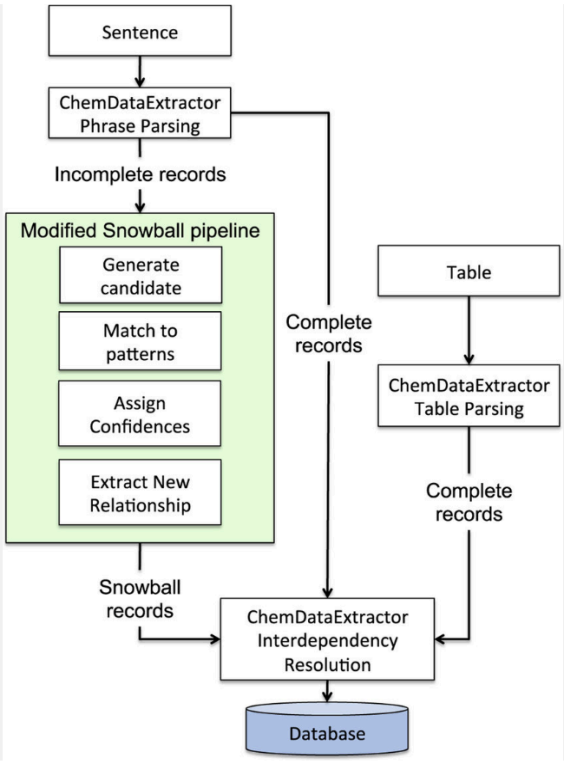
➔ **ChemDataExtractor** (CDE) s'est imposé comme un pionnier en démontrant sa capacité à extraire des informations spécifiques à partir de textes scientifiques complexes. Spécialisé dans la *catégorisation automatique*, cet outil a été particulièrement efficace dans l'organisation structurée des données provenant de documents liés aux matériaux magnétiques et aux batteries. Il a contribué de manière significative à la création de bases de données structurées, facilitant ainsi l'accès rapide et précis à des informations cruciales pour les chercheurs en chimie des matériaux.

Avec l'accent croissant sur les initiatives de Big Data, le besoin d'outils automatisés pour extraire des informations chimiques précieuses à partir de vastes quantités de données non structurées est devenu impératif.
Le ChemDataExtractor répond à ce besoin en fournissant un pipeline NLP conscient de:

- la chimie qui couvre la tokenisation;
- l'étiquetage morphosyntaxique;
- la reconnaissance d'entités nommées;
- l'analyse de phrases.

Le système est conçu pour interpréter simultanément divers domaines de documents, y compris *les paragraphes textuels*, *les légendes* et *les tableaux*, démontrant ainsi sa polyvalence dans le traitement de sources diverses d'informations chimiques.

Les performances de cette library ont été évaluées, démontrant une grande précision dans l'extraction de divers types de données, notamment les identifiants chimiques, les attributs spectroscopiques et les attributs de propriétés chimiques.




***-Sentence:*** Le texte brut, qui est probablement une phrase ou un paragraphe contenant des données chimiques.

***-Table:*** Ceci représente les données chimiques qui sont présentées sous forme de tableaux, souvent trouvés dans des articles de recherche et des rapports.

➢ **ChemDataExtractor Phrase Parsing:** L'outil ChemDataExtractor analyse la phrase pour identifier et extraire des informations chimiques. Le parsing de phrase est le processus de décomposition de la phrase en éléments plus petits pour comprendre sa structure et extraire des informations.

➢ **ChemDataExtractor Table Parsing:** Un processus par lequel ChemDataExtractor analyse et extrait des informations à partir de tableaux.

**Figure 2: Flux de travail de l'extraction de données chimiques pour la création de base de données**

1. **Incomplete records:** Ces informations extraites peuvent être incomplètes et nécessitent donc un traitement supplémentaire pour former des enregistrements complets.

➔ **Modified Snowball pipeline:** C'est une méthode itérative d'extraction d'informations qui utilise des données partiellement structurées pour extraire de nouvelles relations et informations.

- **Generate candidate:** Générer des candidats pour les relations ou les données possibles à partir de l'entrée.
- **Match to patterns:** Faire correspondre ces candidats à des modèles connus pour identifier des données valables.
- **Assign Confidences:** Attribuer un niveau de confiance à chaque relation ou donnée extraite pour indiquer la probabilité que l'information soit correcte.
- **Extract New Relationship:** Extraire de nouvelles relations entre les entités chimiques identifiées.

2. **Snowball records:** Les enregistrements résultants du processus Snowball, qui peuvent être partiels ou entiers, en fonction de la qualité et de la complétude des données extraites.

3. **Complete records:** Les données extraites à la fois du texte (via le parsing de phrase) et des tableaux (via le parsing de table) sont complétées et structurées en enregistrements complets.

➢ **ChemDataExtractor Interdependency Resolution:** Une fois que les enregistrements complets sont obtenus, il peut y avoir des dépendances et des relations entre les données extraites du texte et des tableaux. Cette étape résout ces interdépendances pour intégrer les informations de manière cohérente.

➢ **Database:** Enfin, toutes les informations structurées et résolues sont stockées dans une base de données pour un accès et une analyse ultérieurs.

---

**Figure: Processus d'Extraction et d'Intégration de Données Chimiques pour la Constitution d'une Base de Données (ChemDataExtractor)**

Le processus d'extraction comprend la tokenisation, l'étiquetage morphosyntaxique, la détection des entités nommées chimiques via l'apprentissage automatique, l'identification des relations chimiques à partir du texte et des tableaux avec des règles imbriquées, la résolution des interdépendances, aboutissant ainsi à un ensemble d'enregistrements chimiques cohérents. Ceci permet à ChemDataExtractor d'extraire des informations chimiques de manière indépendante du domaine.

➔ D'autre part, **ChemicalTagger** a joué un rôle clé dans l'évolution de l'efficacité globale du NLP dans le traitement des textes scientifiques. Cet outil met l'accent sur la *reconnaissance* et *l'étiquetage de termes chimiques* complexes, contribuant ainsi à la compréhension fine des composés spécifiques dans le domaine de la chimie des matériaux. Grâce à son expertise dans la détection de termes chimiques et de concepts associés, ChemicalTagger a élargi la portée de l'analyse NLP, permettant une interprétation plus approfondie des informations contenues dans les articles scientifiques.

1. **Outil NLP Open Source pour le traitement du texte chimique :**c'est un outil de traitement du langage naturel (NLP) en code source ouvert conçu pour traiter le texte chimique.
2. **Combinaison des Reconnaissances d'Entités Chimiques (OSCAR) avec des Techniques de NLP :** cet outil combine les reconnaissances d'entités chimiques (OSCAR) avec diverses techniques de traitement du langage naturel (NLP). Cette intégration améliore probablement sa capacité à extraire des informations pertinentes à partir de textes chimiques.

3. **Taggers et Parsers Extensibles et Reconfigurables générés avec ANTLR :** il utilise ANTLR (Another Tool for Language Recognition) pour générer des taggers et des parsers extensibles et reconfigurables. ANTLR est un outil puissant pour la création de langages, de parsers et d'interprètes, et dans ce contexte, il semble être utilisé pour créer des taggers et des parsers qui peuvent être étendus et reconfigurés en fonction des besoins spécifiques.

# Contextualisation du projet:

## 1. Objectifs de la bibliothèque ChemNLP:

- Développement de ChemNLP pour le traitement du langage naturel en chimie des matériaux.
- Curatelle de jeux de données arXiv et PubChem.
- Comparaison et développement de modèles d'apprentissage automatique.
- Résolution de défis spécifiques du domaine.
- Génération de résumés abstraits.
- Intégration avec des ensembles de données, comme la théorie fonctionnelle de la densité.
- Infrastructure robuste pour l'avancement de la recherche.

## 2. Fonctionnalités clés de la bibliothèque:

La bibliothèque ChemNLP présente un ensemble varié de fonctionnalités clés qui renforcent sa polyvalence et son utilité dans le domaine du traitement du langage naturel appliqué à la chimie des matériaux.

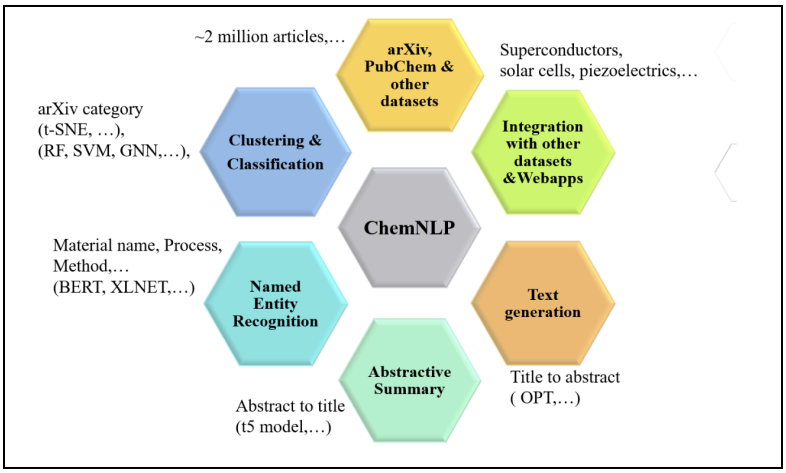



**Figure 3 : Composants et flux de travail de l'analyse de données en ChemNLP**

Voici une expansion sur ces fonctionnalités principales :

- ➢ **Curatelle de Données :**ChemNLP assure la qualité des données à partir de sources comme arXiv et PubChem pour le traitement ultérieur du langage naturel.
- ➢ **Classification de Texte :** Facilite la catégorisation automatique des textes pour une recherche et une organisation efficaces.
- ➢ **Extraction d'Entités Nommées : I**dentifie et extrait des informations spécifiques dans les textes, comme les noms de produits chimiques.
- ➢ **Génération de Résumés :** Crée des résumés significatifs à partir de textes longs, facilitant la compréhension rapide du contenu.

- ➢ **Génération de Texte :** Suggère des résumés à partir de titres d'articles pour une création automatisée de contenu.
- ➢ **Intégration avec des Ensembles de Données Spécifiques :** Permet une analyse ciblée des données pertinentes pour la chimie des matériaux.
- ➢ **Développement d'Interfaces Web :** Offre des interfaces web conviviales pour accéder aux fonctionnalités sans expertise technique approfondie.

En somme, les fonctionnalités de ChemNLP sont conçues pour répondre aux besoins spécifiques du traitement du langage naturel dans le contexte de la recherche en chimie des matériaux, offrant ainsi une suite complète d'outils pour diverses applications.

## 3. Utilisation principale des ensembles de données arXiv et PubChem:

Les ensembles de données arXiv et PubChem occupent une place centrale dans le projet ChemNLP, fournissant une base de données riche et diversifiée pour le développement, l'entraînement et l'évaluation des modèles de traitement du langage naturel. Voici une expansion sur l'utilisation principale de ces ensembles de données :

- **arXiv:**

La base de données **arXiv** est une vaste collection de publications prépubliées couvrant une multitude de domaines scientifiques. Avec un corpus de plus d'un million d'articles, il est particulièrement riche dans les domaines de la *physique*, des mathématiques et de l'informatique, où les chercheurs partagent leurs découvertes avant la révision par les pairs. La physique de la matière condensée représente une part significative de ce dataset, subdivisée en plusieurs sous-catégories allant de la science des matériaux à la supraconductivité.

Ci-dessous, le Tableau 1 présente la répartition par Catégories Principales dans le Dataset arXiv.
Ce tableau offre un aperçu de la distribution des articles par catégories principales, soulignant l'importance de la physique de la matière condensée dans la base de données arXiv.

| Catégories | Nombre d'Articles |
|---|---|
| Physique | 1,042,227 |
| Mathématiques | 425,745 |
| Informatique | 209,068 |
| Statistiques | 72,058 |
| Biologie Quantitative | 24,720 |
| Génie Électrique | 13,064 |
| Finance Quantitative | 8,920 |
| Économie | 1,109 |

**Tableau 1: Répartition par Catégories Principales dans le Dataset arXiv**

Le fichier **"cond-mat.csv**" représente un extrait de données tirées de la catégorie "physique de la matière condensée" de la collection arXiv. Ce fichier contient vraisemblablement des détails sur les publications scientifiques répertoriées selon les sous-catégories définies en physique de la matière condensée. Parmi celles-ci, on compte les sciences des matériaux, la physique à l'échelle mésoscopique et nanoscopique, les électrons fortement corrélés, la supraconductivité, ainsi que la matière condensée douce, les gaz quantiques et d'autres sujets pertinents. L'analyse de ces données s'avère essentielle pour la compréhension et la visualisation des tendances prédominantes en matière de recherche dans ce domaine scientifique.

Le dataset "cond-mat" d'arXiv se caractérise par trois colonnes principales : 'id_1', 'title', et 'abstract'. La colonne **'id_1'** classe chaque article dans une sous-catégorie spécifique, **'title'** en indique le titre, et **'abstract'** en résume le contenu.

| Information | Valeur |
|---|---|
| Nombre de catégories uniques | 9 |
| Nombre de titres uniques | 86,849 |
| Nombre d'abstracts uniques | 86,820 |
| Nombre de lignes | 86,849 |
| Nombre de colonnes | 3 |

**Tableau 2 : Synthèse Générale du Dataset "cond-mat"**

| Catégories | Nombre |
|---|---|
| cond-mat.mtrl-sci | 18,943 |
| cond-mat.mes-hall | 18,761 |
| cond-mat.str-el | 14,174 |
| cond-mat.stat-mech | 11,014 |
| cond-mat.supr-con | 9,157 |
| cond-mat.soft | 6,849 |
| cond-mat.quant-gas | 3,182 |
| cond-mat.other | 2,470 |
| cond-mat.dis-nn | 2,299 |

**Tableau 3: Distribution des Articles par Catégories dans le Dataset "cond-mat"**

- **PubChem:**

Le dataset PubChem, géré par le National Center for Biotechnology Information (NCBI), est une ressource de référence pour la chimie et la biologie. Il propose des informations détaillées sur une variété de molécules, y compris leur structure, leurs propriétés chimiques et physiques, leur activité biologique et leur sécurité. Récemment, ce dataset a été enrichi par des études et des articles liés à des sujets d'actualité tels que le deep-learning, la réalité virtuelle, et des problématiques sanitaires globales telles que la COVID-19. Cette base de données est essentielle pour les professionnels de la santé et les chercheurs qui s'intéressent aux interactions moléculaires et à leur potentiel thérapeutique.

| Catégories | Nombre d'Articles |
|---|---|
| Deep-Learning | 13,18Okii7 |
| Réalité Virtuelle | 11,245 |
| COVID-19 | 8,694 |
| Humane-Connectcome | 4,646 |
| Interface Cerveau-Machine | 4,052 |
| Polymère Électroactif | 896 |
| Électrodes PEDOT | 666 |
| Neuroprothèses | 534 |

**Tableau 4: Répartition par Catégories dans le Dataset PubChem**

Les articles provenant de PubChem, tout comme ceux d'arXiv, sont convertis en format JARVIS-Tools. Cette conversion standardise la représentation des données et facilite leur manipulation au sein de la bibliothèque ChemNLP.
En utilisant ces ensembles de données diversifiés, ChemNLP bénéficie d'une richesse d'informations permettant un apprentissage automatique plus robuste et des performances améliorées sur des tâches spécifiques liées à la chimie des matériaux.

Somme toute, PubChem et arXiv offrent une variété de tâches NLP, telles que la classification de texte, l'extraction d'entités nommées, la génération de résumés, et bien plus encore. Cette variété permet de développer et d'évaluer des modèles de ChemNLP sur un éventail complet de contextes scientifiques.

# Étude de l’existant:

Nous avons choisi de décrire d'abord l'approche de l'article, de la critiquer et de voir où elle échoue, puis de passer à expliquer l'approche que nous avons adoptée pour pallier les lacunes de l'approche de l'article.

## 1. Approche d’un point de vue théorique:

***Clustering Analysis:*** Le workflow décrit ici comprend plusieurs étapes clés pour appliquer l'algorithme t-distributed stochastic neighbor analysis (t-SNE) à la visualisation des catégories dans les ensembles de données arXiv et PubChem. Après le prétraitement initial des titres d'articles, incluant le stemming, le calcul du TF-IDF est effectué pour évaluer l'importance des termes dans les documents. La réduction de dimension avec Truncated Singular Value Decomposition (TruncatedSVD) est ensuite appliquée pour obtenir un espace d'embedding de taille 128. Enfin, l'algorithme t-SNE réduit cet espace d'embedding à une visualisation en deux dimensions, permettant ainsi d'observer la structure locale des catégories dans les ensembles de données.

***Classification:*** L'analyse de clustering est étendue vers une approche plus quantitative à travers la classification supervisée. Cette méthodologie vise à attribuer automatiquement des articles à des catégories prédéfinies dans les ensembles de données arXiv:cond-mat et PubChem. Pour ce faire, le texte subit une transformation en vecteurs numériques à l'aide du modèle de sac de mots (bag of words) et de la représentation Term Frequency-Inverse Document Frequency (TF-IDF), mis en œuvre avec Scikit-learn. Trois approches distinctes sont examinées pour la classification, impliquant l'utilisation exclusive des titres, des résumés (abstracts), ou une combinaison des deux pour une représentation textuelle exhaustive. Divers algorithmes de machine learning, tels que Random Forest (RF), Support Vector Machine linéaire (SVM), Régression logistique (LR), et Réseaux neuronaux graphiques (GNN), sont appliqués aux données transformées.

***Reconnaissance d'Entités Nommées (NER) pour Applications Chimiques :*** L'objectif principal est l'extraction d'informations cruciales, également appelées entités, à partir de textes pertinents en chimie. Pour atteindre cet objectif, l'entraînement d'un modèle de transformation XLNet a été réalisé en utilisant le jeu de données MatScholar, qui comprend des abstracts soigneusement annotés.

***Modèles Texte-vers-Texte pour la Génération Abstraite et de Texte :*** Cette approche novatrice implique l’utilisation un modèle pré-entraîné T5, permettant ainsi la génération de résumés abstraits, la création de titres à partir des abstracts, et la production d'abstracts à partir des titres d'articles.

***Integration with DFT Database:*** L'intégration de ChemNLP avec les bases de données DFT, dont JARVIS-DFT, marque une avancée significative dans l'exploitation synergique des données. Cette collaboration permet d'explorer en détail les propriétés matérielles telles que les structures atomiques et les énergies de formation. La capacité de ChemNLP à générer des descriptions formatées, notamment en texte JSON, offre une grande flexibilité pour l'entraînement de modèles de langage. Cette intégration facilite une gestion efficiente des données et ouvre des perspectives prometteuses pour l'entraînement de modèles prédictifs sur des propriétés matérielles telles que les énergies de formation et les bandgaps.

***WebApp Development:*** Pour le développement de l'application web, l'outil ChemDataExtractor 22 a été employé en conjonction avec JARVIS-Tools et des modèles de reconnaissance d'entités nommées. Cet outil a permis l'extraction automatisée d'entités chimiques, de leurs propriétés associées, de mesures et de relations à partir de documents scientifiques. En utilisant les informations chimiques extraites, une application web a été créée à l'aide de JARVIS-Tools et du Configurable Data Curation System (CDCS), basé principalement sur Django-Python et des bibliothèques JavaScript. Cette application web offre la possibilité de rechercher des articles en fonction de systèmes chimiques spécifiques.

## 2. Approche d'un point de vue technique:

### *1) Installation de ChemNLP:*

La bibliothèque ChemNLP et ses dépendances sont installées avec pip, spécifiquement conçue pour le traitement des données textuelles en chimie des matériaux. Les commandes d'installation assurent la disponibilité de tous les outils nécessaires dans l'environnement Python.

### *2) Analyse des formules chimiques dans le texte:*

Extraction des formules chimiques dans le texte Pour comprendre le contenu des documents en chimie des matériaux, nous avons identifié et extrait les formules chimiques dans le texte en utilisant les outils d'analyse de ChemNLP.

### *3) Téléchargement et visualisation des données:*

Une fois le jeu de données obtenu, généralement à partir d'un référentiel public ou via une API, une analyse exploratoire initiale des données commence.
Cette étape est essentielle pour obtenir un aperçu de la composition du jeu de données, telle que la distribution des catégories de documents ou la fréquence de certains termes. Des outils de visualisation de ***Seaborn*** et ***Matplotlib*** sont utilisés pour créer des représentations graphiques des données.

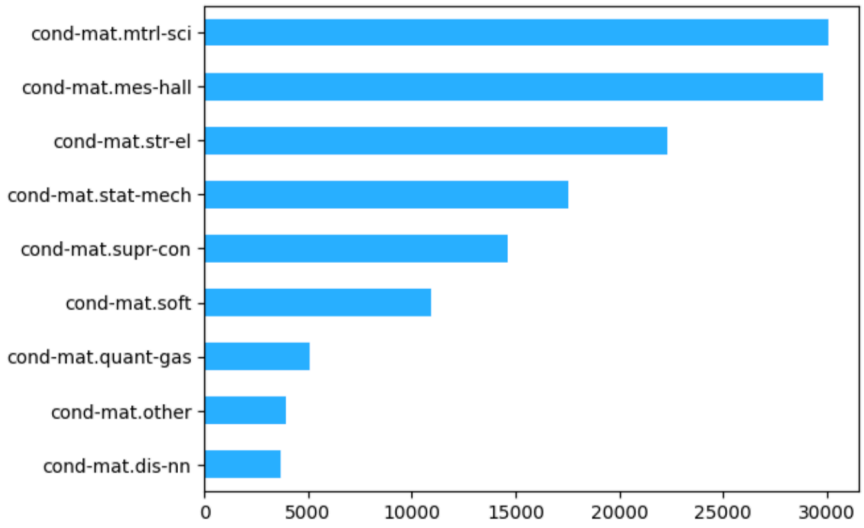



**Figure 4 : Distribution des publications par sous-domaines**

*4) Clustering à l'aide de t-SNE:*

Le clustering est une tâche d'apprentissage non supervisée qui découvre des groupements inhérents dans les données sans étiquettes prédéfinies. En exploitant **t-SNE**, une technique puissante de réduction de la dimensionnalité, les données textuelles de haute dimension sont transformées **en un espace 2D** où les modèles et les relations peuvent être inspectés visuellement. Cette représentation graphique permet la reconnaissance intuitive des graphes, éclairant la catégorisation naturelle au sein du corpus de documents.

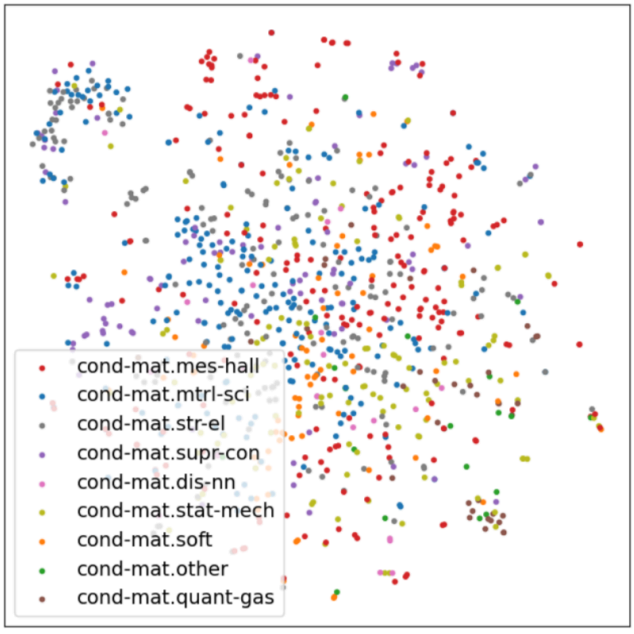



**Figure 5 : Analyse de clustering des publications en physique de la matière condensée**

*5) Classification à l'aide de la régression logistique:*

La classification consiste à attribuer des étiquettes prédéfinies aux documents en fonction de leur contenu. Cette étape tire parti de la régression logistique, un algorithme d'apprentissage automatique fondamental adapté aux tâches de classification binaire et multinomiale. Le modèle est formé sur un sous-ensemble du jeu de données et validé sur des données de test séparées. Les performances du modèle sont évaluées de manière critique à l'aide de métriques telles que la précision et les matrices de confusion, qui fournissent des mesures quantitatives des capacités prédictives du modèle.

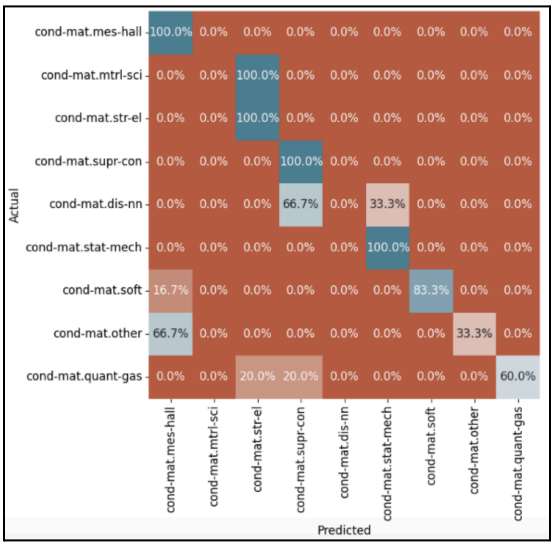



**Figure 6 : Matrice de confusion pour la classification des publications en matière condensée**

*6) Génération de texte à l'aide de l'interface HuggingFace:*

La bibliothèque transformers de HuggingFace offre une riche collection de modèles pré-entraînés qui peuvent être utilisés pour des tâches de génération de texte. En employant ces modèles, l'algorithme peut synthétiser du texte qui reflète le style et le contenu des données d'entraînement. Cela peut servir à diverses fins, de la génération de résumés pour des articles scientifiques à la prédiction de la continuation de données textuelles incomplètes, démontrant la puissance des modèles NLP modernes.

*7) Entrée au classement JARVIS:*

Pour comparer les performances des modèles développés dans le cadre du projet ChemNLP, les résultats peuvent être soumis au **classement JARVIS**, une plateforme de comparaison des algorithmes d'informations sur les matériaux. Cette étape implique de préparer le modèle et ses résultats selon les directives de soumission du classement, ce qui inclut généralement la spécification de la catégorie sous laquelle le modèle concourt et la fourniture de toute documentation ou métadonnée nécessaire.

3. **Limitations de l'approche:**

**-**La visualisation avec T-SNE révèle des **clusters visuellement distincts** mais avec des chevauchements significatifs, soulevant des questions sur leur interprétation et la fiabilité des paramètres de T-SNE. Bien que T-SNE soit perspicace, il ne constitue pas une méthode de clustering à part entière et nécessite une validation avec des techniques explicites telles que *K-means* ou des approches basées sur la densité. De plus, il est important de noter que les **données étaient distribuées mais non regroupées**.

**-**La classification effectuée à l'aide de la **régression logistique a révélé des degrés de précision variables**, comme le montrent les matrices de confusion. Tandis que certaines catégories ont été prédites avec une grande précision, d'autres ont montré des confusions, signalant des déséquilibres potentiels de classe ou des insuffisances dans la représentativité des données d'entraînement.

En conclusion, l'analyse des textes en chimie des matériaux tire parti des capacités de ChemNLP, de t-SNE et des modèles de classification. Toutefois, une évaluation continue est nécessaire pour affiner ces outils. L'utilisation de NER et de modèles pré-entraînés comme BERT complète cette approche, offrant une puissante capacité d'analyse pour les données textuelles complexes et ouvrant la voie à des découvertes significatives et c'est ce que nous prévoyons d'intégrer dans notre approche afin d'affiner nos outils d'analyse des textes en chimie des matériaux.

# Modification et optimisation de l'approche:

1. **Contexte générale:**

Dans cette étude, nous proposons une approche novatrice pour dépasser les limitations des méthodes précédentes en analyse textuelle en chimie des matériaux. Nous avons combiné les capacités de ChemNLP, de t-SNE et des modèles de classification avec des ajustements méticuleux et des améliorations essentielles pour affiner notre méthode.

Notre schéma global de procédure, *représenté ci-dessous*, illustre comment nous avons intégré ces technologies préexistantes avec les nouvelles méthodes pour surmonter les défis identifiés dans l'étude précédente. En utilisant une approche holistique, nous visons à fournir une analyse textuelle plus précise et plus fiable, ouvrant ainsi de nouvelles perspectives dans le domaine de la chimie des matériaux.

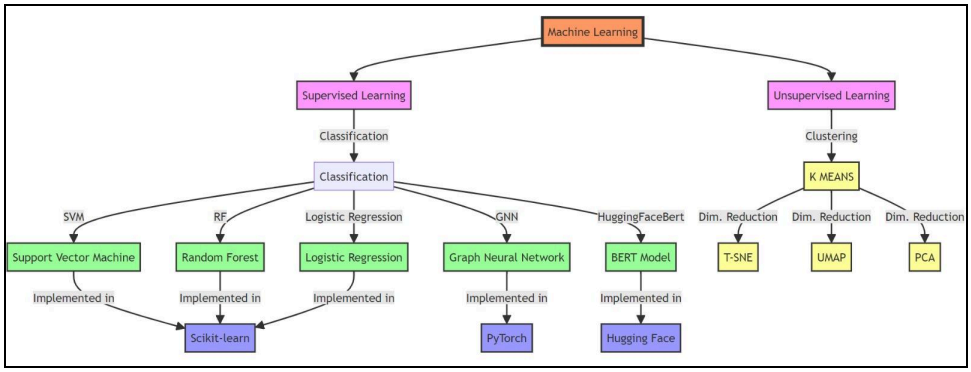



**Figure 7: Schéma golabale de procédures**

## 2. Approche suivie et analyse des résultats:

### ➔ Phase 1: *Préparation et Nettoyage des Données*

- ***Nettoyage et Formatage des Données:*** La première partie de notre exploration avant de plonger dans les complexités des analyses de clustering a été dédiée au nettoyage et au formatage des données, assurant ainsi la qualité et la précision des informations à analyser.
- ***Extraction des caractéristiques avec BERT:*** Nous avons combiné l'extraction des caractéristiques en utilisant le puissant modèle BERT avec la transformation des textes en vecteurs numériques, capturant ainsi à la fois l'essence sémantique et linguistique des documents scientifiques.
- ***Segmentation des Données:*** Division du jeu de données en segments pour le traitement. Pour faciliter le traitement, les données ont été segmentées, permettant ainsi des calculs plus gérables et une meilleure organisation des informations. Nous avons combiné l'extraction des caractéristiques en utilisant le puissant modèle BERT avec la transformation des textes en vecteurs numériques, capturant ainsi à la fois l'essence sémantique et linguistique des documents scientifiques.
- ***Génération et sauvegarde des embeddings par segment:*** Chaque segment a été traité pour générer des embeddings, qui ont ensuite été sauvegardés dans des fichiers pickle pour une utilisation ultérieure, garantissant l'intégrité des données pour les étapes suivantes.

### ➔ Phase 2*: Réduction de Dimension et Clustering*

- ***Réduction de Dimension avec PCA, T-SNE, et UMAP:*** La deuxième phase a impliqué l'application de méthodes PCA, T-SNE et UMAP de réduction de dimensionnalité pour réduire la dimensionnalité des embeddings et simplifier la visualisation des complexes données textuelles.

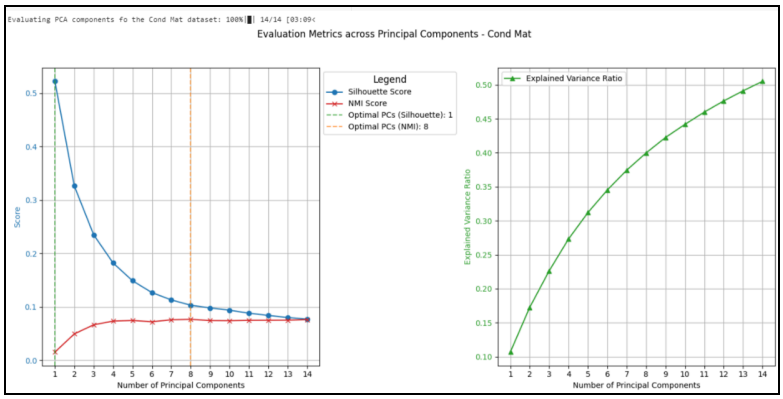



**Figure 8: Évaluation des métriques de PCA avec le calcul des scores de NMI et de la silhouette**

-Le score Silhouette diminue rapidement avec le nombre croissant de composants, suggérant que moins de dimensions sont nécessaires pour capturer les clusters.
-La courbe NMI indique qu'un nombre réduit de composants peut suffire à préserver les structures importantes.
-Le ratio de variance expliquée augmente de manière monotone, mais il est important de considérer la clarté des clusters pour choisir le nombre optimal de composants.

- ***Visualisation des embeddings réduits et identification des clusters:*** Nous avons combiné la visualisation des embeddings réduits pour identifier les clusters, révélant ainsi des groupements distincts au sein des données, avec des clusters reflétant les sous-catégories de la chimie.
- ***Clustering et Évaluation des Résultats:*** Mise en œuvre du clustering K-means sur les embeddings réduits.Le clustering K-means a été utilisé pour regrouper les embeddings, avec une évaluation attentive des résultats pour assurer la validité des groupes formés.
- ***Évaluation des clusters à l'aide de métriques telles que le score de silhouette et l'information mutuelle normalisée:*** Les métriques de score de silhouette et d'information mutuelle normalisée ont servi à évaluer la qualité des clusters, confirmant l'alignement entre les groupements obtenus et les catégories de sujets sous-jacents.
- ***Analyse des clusters par rapport aux sous-disciplines chimiques:*** L'analyse des clusters a offert des aperçus précieux sur les relations entre les documents et les sous-disciplines chimiques, offrant des pistes pour de futures recherches.
- ***Évaluation des Clusters selon la Réduction de Dimension:*** Le graphique des métriques d'évaluation illustre l'impact de la réduction de dimension sur la qualité des clusters, soulignant l'importance de choisir le nombre optimal de composants principaux pour une segmentation claire et significative.
- ***Dissection Visuelle des Groupes Chimiques par t-SNE, UMAP et PCA:*** Les visualisations obtenues par t-SNE,UMAP et PCA ont permis de déceler des structures latentes, mettant en lumière des clusters distincts qui reflètent les catégories scientifiques présumées, offrant une nouvelle perspective sur les données chimiques.

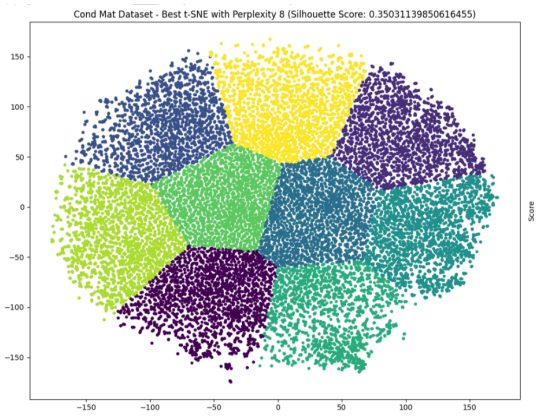



**Figure 9: t-SNE avec Perplexité 8: Cartographie des Clusters en Chimie de la Matière Condensée**

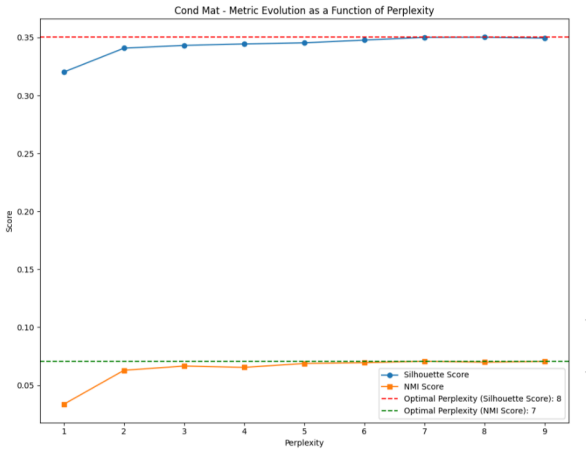



**Figure 10: Évolution des Métriques de Clustering: Impact de la Perplexité sur t-SNE**

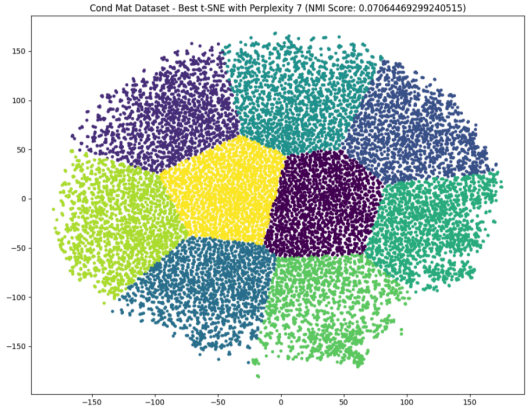



**Figure 11: t-SNE avec Perplexité 7: Cartographie des Clusters en Chimie de la Matière Condensée**

-Ces graphiques t-SNE fournissent une représentation visuelle des structures cachées au sein du jeu de données en matière condensée, avec une mise en lumière particulière des variations dans la perplexité.

-La perplexité, un hyperparamètre clé dans t-SNE, influence la façon dont les clusters sont formés et perçus. Une perplexité de 8 semble offrir une séparation légèrement plus claire des clusters, comme en témoigne le score de silhouette supérieur, indiquant une meilleure définition des limites entre les clusters.

-En revanche, une perplexité de 7 révèle des clusters plus compacts avec des distinctions moins marquées, ce qui est corroboré par un score NMI légèrement inférieur. Ces différences subtiles dans les scores mettent en évidence l'importance de choisir judicieusement la perplexité pour capturer la structure des données de manière fidèle.

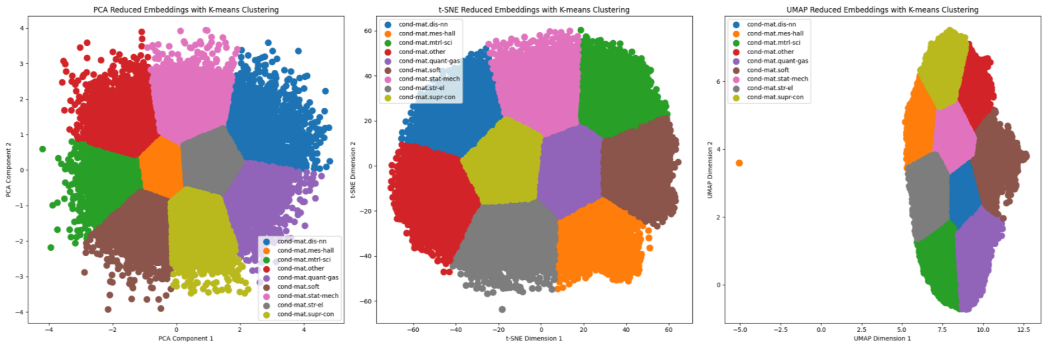



**Figure 12: Clustering avec T-SNE, UMAP et PCA**

- ***Comparaison des Topologies de Clusters 2D vs 3D:*** La comparaison entre les visualisations 2D et 3D a démontré que les représentations en deux dimensions offrent une clarté supérieure, rendant l'interprétation des clusters plus intuitive et moins sujette à confusion due à la superposition des points dans l'espace tridimensionnel.

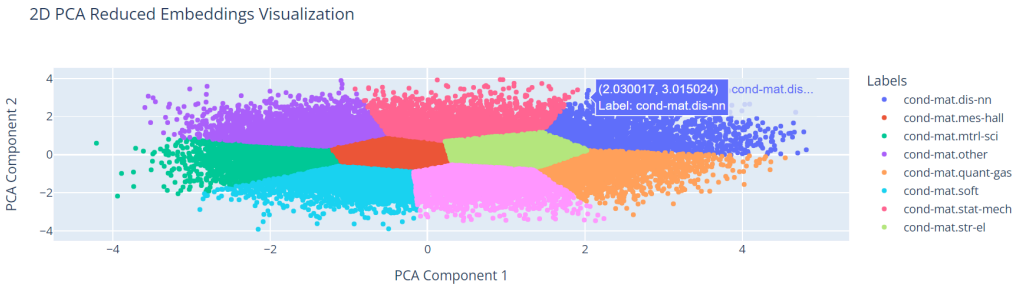



**Figure 13: Visualisation 2D PCA**

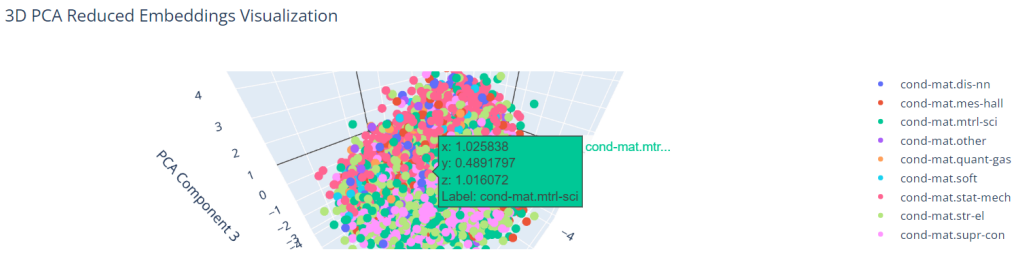



**Figure 14: Visualisation 3D PCA**

- La visualisation en 2D des embeddings réduits par PCA montre clairement les différents clusters, avec des couleurs distinctes pour chaque groupe. Cela suggère que la représentation en 2D est suffisamment informative pour distinguer les groupes de données.
-En comparaison, la visualisation en 3D, bien que fournissant une perspective plus profonde de la disposition des clusters dans l'espace des caractéristiques, peut sembler moins claire en raison de la superposition des points et de la complexité ajoutée par la troisième dimension.

## ➔ Phase 3: Affinement et Validation des Clusters

- ***Sélection des Paramètres pour K-means et ajustement fin des hyperparamètres:*** Cette phase initiale a été consacrée à la sélection minutieuse des paramètres pour l'algorithme de clustering K-means. Un ensemble optimal d'hyperparamètres a été déterminé pour assurer que les clusters générés soient aussi précis et significatifs que possible.
- ***Application du Clustering K-means et Classification des données en clusters:*** Avec les paramètres choisis, l'algorithme K-means a été appliqué aux embeddings réduits. Cette étape cruciale a permis de classer les documents du jeu de données en différents clusters basés sur leur proximité dans l'espace des caractéristiques.
- ***Analyse approfondie des clusters formés:*** Chaque cluster formé par l'algorithme K-means a été évalué pour mesurer la cohérence interne et la démarcation par rapport aux autres clusters. L'analyse des clusters a permis de mettre en évidence leur alignement avec les catégories scientifiques attendues et d'identifier les domaines de la chimie représentés par chaque groupe.
- ***Interprétation tirée de la classification des Résultats de la Matrice de Confusion:*** L'analyse du Scatter Plot PCA et de la matrice de confusion a révélé la structure complexe du jeu de données en matière condensée. La superposition entre les clusters a souligné des ambiguïtés et des similitudes subtiles entre les catégories scientifiques, suggérant des défis dans la distinction nette des groupes par l'algorithme K-means.

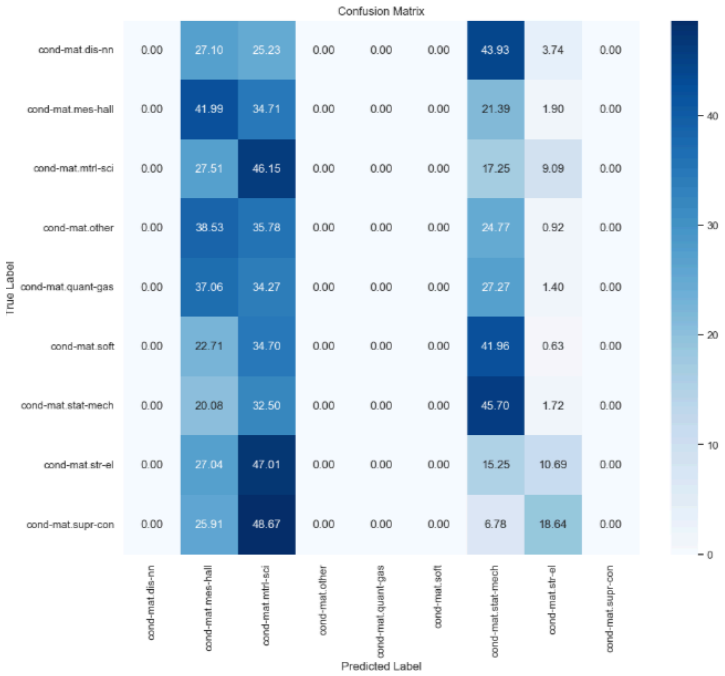



**Figure 15: Matrice de Confusion de  Cond Mat**

La matrice de confusion dévoile la performance de notre modèle de classification. Les valeurs diagonales montrent le pourcentage de prédictions correctes pour chaque label réel, tandis que les valeurs hors diagonale indiquent des erreurs de classification, révélant des confusions entre les catégories. Des valeurs nulles indiquent une absence de confusion pour certaines paires de catégories, ce qui est un bon signe. Cependant, l'existence de chevauchements signale des défis pour l'algorithme K-means dans la distinction des groupes.


**Figure 16:  Analyse des Composantes PCA des Étiquettes de Classe Réelles**

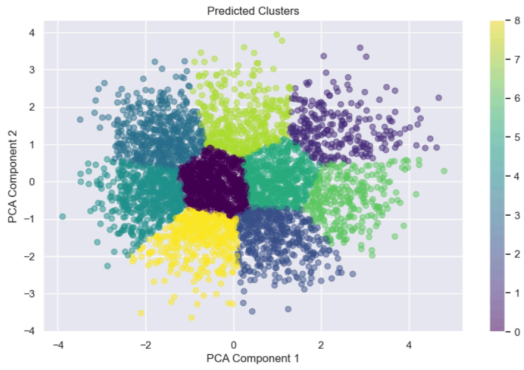



**Figure 17: Visualisation PCA des Clusters Prédits dans les Donnée**

-*Le premier graphique* montre la répartition des points de données du jeu de test dans un espace à deux dimensions obtenu après une réduction de dimensionnalité par PCA. Les différentes couleurs représentent les clusters prédits par l'algorithme de K-means. On observe une formation de groupes distincts, ce qui suggère que l'algorithme a identifié des structures de données naturellement regroupées. Toutefois, certains chevauchements de clusters indiquent que toutes les caractéristiques des données ne sont pas complètement séparables dans cet espace réduit.

-*Le deuxième graphique* présente la même projection PCA que le premier, mais cette fois-ci, les couleurs indiquent les véritables étiquettes des points de données, et non les clusters prédits par le K-means. Le mélange des couleurs illustre le degré de mélange des classes réelles dans l'espace de caractéristiques réduit. La comparaison entre les clusters prédits et les étiquettes réelles peut révéler la précision de la méthode de clustering utilisée. Si l'on compare les deux graphiques, on peut évaluer la performance du modèle de clustering en termes de séparation des différentes classes.

# Conclusion :

À travers les trois phases de ce projet, nous avons navigué à travers les défis de l'analyse de données textuelles en chimie, en utilisant des outils de pointe en TAL et en apprentissage automatique. De la préparation initiale des données à l'application sophistiquée de techniques de clustering, chaque étape a contribué à une meilleure compréhension de la littérature scientifique dans ce domaine.

Les résultats obtenus illustrent l'efficacité des embeddings BERT chemical-bert-uncased et des méthodes de réduction de dimension dans la distinction des documents, malgré les chevauchements observés. Cela indique que des techniques supplémentaires pourraient être envisagées pour améliorer la séparation des clusters.

En résumé, ce projet a franchi des étapes cruciales dans l'analyse des données textuelles en chimie des matériaux, en exploitant la synergie entre ChemNLP, les visualisations t-SNE et les modèles de classification robustes. Chaque phase a contribué à façonner une compréhension plus profonde de la littérature scientifique, démontrant l'efficacité des embeddings BERT et des méthodes de réduction de dimension pour distinguer les documents malgré les chevauchements inhérents.

La nécessité d'une approche itérative est mise en évidence, suggérant que l'adoption de techniques d'embedding avancées et l'exploration de méthodes de clustering alternatives pourraient améliorer la séparation des clusters. Ce projet n'a pas seulement enrichi notre compréhension des modèles présents dans les données chimiques mais a également posé les jalons pour des recherches futures, incitant à de nouvelles avancées dans l'analyse des données scientifiques et la découverte dans le vaste domaine de la chimie.

# Bibliographie:


Hawizy, L., Jessop, D. M., Adams, N., Murray-Rust, P., & Robertson, S. P. (2011). "ChemicalTagger: A tool for semantic text-mining in chemistry." Journal of Cheminformatics, 3(1), 17. DOI: 10.1186/1758-2946-3-17.

Swain, M. C., & Cole, J. M. (2016). "ChemDataExtractor: A Toolkit for Automated Extraction of Chemical Information from Scientific Literature." Journal of Chemical Information and Modeling, 56(10), 1894–1904. DOI: 10.1021/acs.jcim.6b00207.

Olivetti, E. A.; Cole, J. M.; Kim, E.; Kononova, O.; Ceder, G.; Han, T. Y.-J.; Hiszpanski, A. M. Data-driven materials research enabled by natural language processing and information extraction. Applied Physics Reviews 2020, 7, 041317.

Choudhary, K.; DeCost, B.; Chen, C.; Jain, A.; Tavazza, F.; Cohn, R.; Park, C. W.; Choudhary, A.; Agrawal, A.; Billinge, S. J., et al. Recent advances and applications of deep learning methods in materials science. npj Computational Materials 2022, 8, 1–26.

Kononova, O.; He, T.; Huo, H.; Trewartha, A.; Olivetti, E. A.; Ceder, G. Opportunities and challenges of text mining in materials research. Iscience 2021, 24, 102155.

"BERT for Chemical Industry." Hugging Face Model Hub, Recobo. [En ligne]. Disponible sur : https://huggingface.co/recobo/chemical-bert-uncased

Choudhary, K., Kelley, M. L. (2022, Dec 22). "ChemNLP: A Natural Language Processing based Library for Materials Chemistry Text Data.". [En ligne]. Disponible sur: https://ar5iv.labs.arxiv.org/html/2209.08203

Chithrananda, S., Grand, G., Ramsundar, B. (2020, October 23). "ChemBERTa: Large-Scale Self-Supervised Pre Training for Molecular Property Prediction." [En ligne]. Disponible sur : https://arxiv.org/pdf/2010.09885.pdf.

Belkin, M., & Niyogi, P. (2003). Laplacian eigenmaps and spectral techniques for embedding and clustering. In Advances in neural information processing systems (pp. 585-591).

Lloyd, S. (1982). Least squares quantization in PCM. IEEE transactions on information theory, 28(2), 129-137.

Milligan, G. W., & Cooper, M. C. (1985). An examination of procedures for determining the number of clusters in a data set. Psychometrika, 50(2), 159-179.

Rousseeuw, P. J. (1987). Silhouettes: A graphical aid to the interpretation and validation of cluster analysis. Journal of Computational and Applied Mathematics, 20, 53-65.

Vinh, N. X., Epps, J., & Bailey, J. (2010). Information theoretic measures for clusterings comparison: Variants, properties, normalization and correction for chance. Journal of Machine Learning Research, 11(Oct), 2837-2854.

Gasteiger, J., & Engel, T. (2003). Chemoinformatics: A Textbook. Wiley-VCH.

Hastie, T., Tibshirani, R., & Friedman, J. (2009). The Elements of Statistical Learning: Data Mining, Inference, and Prediction. Springer.

Maaten, L. v. d., & Hinton, G. (2008). Visualizing data using t-SNE. Journal of Machine Learning Research, 9(Nov), 2579-2605.

McInnes, L., Healy, J., & Melville, J. (2018). UMAP: Uniform Manifold Approximation and Projection for Dimension Reduction. arXiv preprint arXiv:1802.03426.

Jolliffe, I. T., Principal Component Analysis, 2nd Edition, Springer, 2002.