

UNIVERSITÉ NATIONALE DU VIETNAM, HANOI

LA ROCHELLE UNIVERSITÉ, FRANCE

INTÉGRATION D'UNE APPROCHE
D'INTELLIGENCE ARTIFICIELLE POUR LA
DéTECTION DE MODULES ACTIFS DANS LES
RÉSEAUX D'INTERACTIONS BIOLOGIQUES À
TRAVERS DES DONNÉES MULTIVUES

MÉMOIRE DE FIN D'ÉTUDES DE MASTER EN
INFORMATIQUE

REDIGÉ PAR: GIRESSE TCHOTANEU NGATCHA

SOUS LA DIRECTION DE : CLAUDE PASQUIER

*HDR en Informatique et Chercheur en Biologie
Computationnelle*

Institut pour la Francophonie et l'Innovation

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis

Institut pour la Francophonie et l'Innovation

Code : 8480201.02v

Systèmes Intelligents & Multimédia

HANOI, FÉVRIER 2024

UNIVERSITÉ NATIONALE DU VIETNAM, HANOI

INTÉGRATION D'UNE APPROCHE D'INTELLIGENCE
ARTIFICIELLE POUR LA DÉTECTION DE MODULES ACTIFS
DANS LES RÉSEAUX D'INTERACTIONS BIOLOGIQUES À
TRAVERS DES DONNÉES MULTIVUES

MÉMOIRE DE FIN D'ÉTUDES DE MASTER EN INFORMATIQUE

REDIGÉ PAR: GIRESSE TCHOTANEU NGATCHA

Student No. 2230455

SOUS LA DIRECTION DE : CLAUDE PASQUIER

HDR en Informatique et Chercheur en Biologie Computationnelle

Lu et approuvé

HANOI, FÉVRIER 2024

DÉCLARATION D'AUTHENTICITÉ

Nous attestons sur l'honneur que le travail présenté dans ce Memoire de Recherche, intitulée "Intégration d'une approche d'intelligence artificielle pour la détection de modules actifs dans les réseaux d'interactions biologiques à travers des données multivues," est original et a été réalisé par **Giresse TCHOTANEU NGATCHA** Sous la direction de **Claude PASQUIER** (*HDR en Informatique et Chercheur en Biologie Computationnelle*)

Lien du dépôt GitHub : https://github.com/tchotaneu/Intership_I3S

Hanoi, février 2024

Redigé par: Giresse TCHOTANEU
NGATCHA

Sous la direction de : Claude PASQUIER

*HDR en Informatique et Chercheur en Biologie
Computationnelle*

REMERCIEMENTS

Je tiens à exprimer ma profonde gratitude envers les personnes et les institutions suivantes pour leur soutien précieux tout au long de la réalisation de ce mémoire :

- Claude PASQUIER, dont l’encadrement attentif, les conseils avisés et le soutien moral ont été d’une importance capitale tout au long de ce stage.
- Sabine Barrere, pour son accueil chaleureux et son aide précieuse dans les démarches administratives lors de mon intégration au laboratoire.
- NGUYEN TRAN Minh Anh, pour sa présence constante, ses encouragements et son écoute attentive pendant ma formation et mon stage.
- Serge SONFACK pour ces encouragements
- Je souhaite exprimer ma gratitude envers notre institution de formation, l’Institut Francophone International (IFI) de l’Université Nationale du Vietnam à Hanoi. Je tiens à remercier chaleureusement le corps enseignant pour la qualité de la formation dispensée et le personnel administratif pour leur collaboration efficace.
- Mes remerciements vont également à la famille CERISE et la famille KAMANDA pour leur aide précieuse dans les moments difficiles.
- Je souhaite également exprimer ma gratitude envers mes parents, mes frères et sœurs pour leur soutien moral et financier, qu’ils m’ont apporté durant tout ce parcours.

Cette année de Master Recherche a été un défi, et je suis reconnaissant envers tous ceux qui ont contribué à mon succès.

RESUMÉ

Dans les expériences biologiques, les chercheurs cherchent à mesurer l'activité des gènes pour identifier ceux qui sont activés et potentiellement liés aux phénotypes observés. Ces analyses visent à comprendre les relations entre les caractéristiques observées (phénotypes) et à identifier des schémas qui pourraient devenir des cibles thérapeutiques. Cependant, la sélection des gènes à étudier reste un défi crucial. La méthode conventionnelle consiste à rechercher des modules de gènes dont l'action combinée réalise une fonction spécifique, souvent à travers l'approche des "top k-gènes" les plus variables. Cependant, cette méthode présente des limites car la variabilité ne garantit pas toujours la pertinence biologique.

L'identification d'ensembles de gènes spécifiques à une condition à partir d'expériences transcriptomiques représente un défi majeur dans la recherche biologique. Malgré les diverses approches proposées pour surmonter ces limites, elles présentent souvent des limitations qui les rendent peu utiles aux biologistes. Dans ce contexte, notre recherche, intitulée "Intégration d'une approche d'intelligence artificielle pour la détection de modules actifs dans les réseaux d'interactions biologiques à travers des données multivues", explore une méthode novatrice pour identifier ces modules génétiques actifs. Notre approche combine plusieurs vues de données issues de données génomiques. Les résultats obtenus sur des jeux de données réels démontrent que notre méthode permet d'identifier de nouveaux groupes de gènes d'intérêt.

Notre recherche se concentre sur la construction des critères de vue de données, l'intégration des vues au sein d'un espace unifié et le regroupement des modules en clusters. L'intégration des vues de données génomiques met en évidence la flexibilité et la capacité à gérer plusieurs vues issues de plusieurs expériences transcriptomiques. Pour y parvenir, nous avons développé des critères essentiels pour la construction des graphes représentant les vues. En utilisant des réseaux de neurones avec des mécanismes d'attention, nous avons unifié les différentes vues de données par des vecteurs d'incorporation denses (embedding vectors) dans un espace vectoriel de faible dimension, en nous appuyant sur les propriétés d'intégration dans un espace vectoriel unifié des réseaux collaboratifs multi-vues (*ATA et al., 2020*). Nous avons ensuite appliqué différentes techniques de clustering pour détecter les modules actifs, dont la plus efficace est celle basée sur une approche gloutonne (*PASQUIER et al., 2022*) mais associée à la similarité de Pearson. Cette technique de clustering s'est avérée meilleure que les méthodes DBSCAN, OPTICS et Spectral Clustering. Notre modèle, que nous avons surnommé Amine_multiview, surpasse certains modèles de détection de modules actifs en bioinformatique tels que Bionet, Cosine et Diamond. Ces premiers résultats doivent toutefois être approfondis.

Mots Clés: Données génomiques, Transcriptomique, Réseaux d'interactions biologiques, Données multivues, Clustering, Vecteurs d'incorporation (Embedding vectors), Réseaux de neurones, Mécanisme d'attention, Similarité de Pearson, Modules actifs, Approche gloutonne

ABSTRACT

In biological experiments, researchers aim to measure gene activity to identify those that are activated and potentially linked to observed phenotypes. These analyses seek to understand the relationships between observed characteristics (phenotypes) and to identify patterns that could become therapeutic targets. However, the selection of genes to study remains a crucial challenge. The conventional method involves searching for sets of genes, called modules, whose combined action achieves a specific function, often through the "top k-genes" approach that selects the most variable genes. However, this method has limitations as variability does not always guarantee biological relevance.

Identifying gene sets specific to a condition from transcriptomic experiments is a major challenge in biological research. Despite various approaches proposed to overcome these limitations, they often have constraints that make them less useful to biologists. In this context, our research, entitled "Integration of an artificial intelligence approach for the detection of active modules in biological interaction networks through multiview data," explores an innovative method to identify these active genetic modules. Our approach combines multiple data views from genomic data. Results on real datasets demonstrate that our method identifies new groups of genes of interest.

Our research focuses on constructing data view criteria, integrating views within a unified space, and clustering modules into clusters. The integration of genomic data views highlights flexibility and the ability to handle multiple views from multiple transcriptomic experiments. To achieve this, we have developed essential criteria for constructing graphs representing views. Additionally, by using neural networks with attention mechanisms, we have unified the different data views through embedding vectors in a low-dimensional vector space, relying on the properties of integration into a unified vector space of multi-view collaborative networks *ATA et al., 2020*. We then applied various clustering techniques to detect active modules, with the most effective being one based on a greedy approach *PASQUIER et al., 2022* associated with Pearson similarity. This clustering technique outperformed methods such as DBSCAN, OPTICS, and Spectral Clustering. Our model, which we dubbed Amine_multiview, surpasses some active module detection models in bioinformatics such as Bionet, Cosine, and Diamond. However, these initial results need further exploration.

Keywords: Genomic data, Transcriptomics, Biological interaction networks, Multiview data, Clustering, Embedding vectors, Neural networks, Attention mechanism, Pearson similarity, Active modules, Greedy approach

TABLE DES MATIÈRES

Table des matières	vi
Table des figures	ix
Liste des tableaux	1
1 Introduction	2
1.1 Contexte	2
1.2 Problématique	3
1.3 Motivation du projet	4
1.4 Objectifs de la recherche	4
1.5 Questions de recherche	5
2 Etat de l'Art	7
2.1 Introduction	7
2.2 Théories et modèles existants	7
2.2.1 Notion de Multivues	7
2.2.2 Factorisation Matricielle par des Composants Liés pour l'Intégration Unifié	8
2.2.3 La Théorie Basée sur les cadre de probabiliste	8
2.2.4 La Théorie Basée sur les Méthodes de Conflation de distribution de probabilité :	9
2.2.5 La Théorie Basée sur les Méthodes de Collaboration de données :	10
2.3 Présentation des principaux modèles generant les embbedding multivues	12
2.3.1 Le modele MvNE (Multi-view Neighbourhood Embedding)	12
2.3.2 Le modele MANE (Multi-View Collaborative Network Embedding) . . .	13
3 Méthodologie	16
3.1 Cadre conceptuel	16
3.2 Conception du modèle	17
3.2.1 Modèles de Construction de la Première Vue	17
3.2.2 Modèles de Construction de la Deuxième Vue	17
3.2.3 choix du Modèles de theorie d'unifictaion d'embbedding	18
3.2.4 Choix des Mesures de Similarité dans l'Espace Unifié	19
3.2.5 Critères d'Évaluation	19
3.3 Collecte de données :	20

3.3.1	Description de l'Algorithme de Génération de Données	20
3.3.2	Génération de la structure topologique du graphe	21
3.3.3	Génération des Modules dans le Graphe	22
3.3.4	Attribution des Poids (p_value) aux Nœuds	23
4	Validation du Modèle	24
4.1	Environnement d'Implémentation : Choix Technologiques	24
4.1.1	Langages de Programmation	24
4.1.2	Outils de Traitement de Données	24
4.1.3	Librairies de Machine Learning et Statistiques	24
4.1.4	Technologies Spécifiques pour l'Analyse de Graphes	25
4.1.5	Gestion des Fichiers et Utilitaires	25
4.1.6	Autres Librairies et Fonctions	25
4.2	Approche de simulation	25
4.3	Analyse des Résultats MANE pour les Modèles de Construction	26
4.3.1	Contexte	26
4.3.2	Objectif	26
4.3.3	Evaluation avec le Modèle de Construction 1	27
4.3.4	Evaluation avec le Modèle de Construction 2	29
4.3.5	Evaluation avec le Modèle de Construction 3	30
4.3.6	Evaluation avec le Modèle de Construction 4	31
4.4	Analyse des Résultats avec MvME pour les Modèles de Construction	32
4.4.1	Contexte et Méthodologie	32
4.4.2	Résultats et Problématiques	32
4.4.3	Décision et Implications	32
4.5	Choix du Modèle d'intégration d'embedding	32
4.6	Comparaison avec les méthodes de clustering classiques	33
5	Application du modèle à des données réelles	35
5.1	Test du modèle sur les données réelles	35
5.1.1	Description du Jeu de Données :	35
5.1.2	Préparation des Données	35
5.1.3	Résultats du Test :	36
5.2	Interprétation des résultats	36
5.2.1	Interpretation de l'enrichissement du module 1	36
5.2.2	Enrichement des autres modules	38
5.3	Visualisation des interactions entre les modules	41
6	Conclusion	45
6.1	Contributions de l'étude	45
6.2	Résumé des résultats	45
6.3	Réponses aux questions de recherche et Limites	45
6.4	Suggestions pour des travaux futurs	46

Anexos

M	illustration du déroulement du Projet	54
N	Courbe d'apprentissage de la fonction de Perte	55
O	Procédure de travail avec le Cadre MANE	56
P	Procédure de travail avec Cadre MvME	57
Q	Presentation du Laboratoire	58
R	Presentation de l'équipe de travail SPARKS	59

TABLE DES FIGURES

3.1	Illustration de notre cadre conceptuel via l'approche multivues	17
3.2	Illustration de notre modele conceptuel	21
3.3	Illustration en digaramme classe de la description de l'agorithme qui genere les données artificielles	22
3.4	Illustration la gereration des données artificielles	23
4.1	tableau de F1 score des modeles bioinformatique	33
5.1	Enrichement du module 1 via la plateforme GSEA	37
5.2	Enrichement du module 2 via la plateforme GSEA	38
5.3	Enrichement du module 4 via la plateforme GSEA	39
5.4	Enrichement du module 20 via la plateforme GSEA	39
5.5	Enrichement du module 85 via la plateforme GSEA	40
5.6	Enrichement du module 90 via la plateforme GSEA	41
5.7	La visualisation des interactions entre les modules (avec la voie hyposie)	42
5.8	La visualisation des interactions entre les modules (avec la voie EMT	43
5.9	La visualisation des interactions entre les modules (avec la voie uv reponse) . . .	43
5.10	La visualisation des interactions entre les modules (avec la voie MTORC1 SIGNALING)	44
5.11	La visualisation des interactions entre les modules (avec la voie OXIDATIVE PHOSPHORYLATION)	44
M.1	deroulement du projet	54
N.1	La visualisation de progression de la fonction de perte avec le Framework MANE sur des Données Réelles	55
O.1	différents types de relation entre les vues avec le cadre MANE	56
O.2	Framework MANE	56
P.1	Procédure de travail avec MvME	57
P.2	Framework MvME	57
Q.1	Presentation du Laboratoire	58
R.1	Presentation de l'équipe de travail SPARKS	59

LISTE DES TABLEAUX

4.1	<i>Analyse comparative des F1 scores des modèles avec un true hit de taille 10</i>	27
4.2	<i>Analyse comparative des scores F1 des modèles avec un true hit de taille 20.</i>	28
4.3	<i>Analyse comparative des scores F1 des modèles avec un true hit de 10 pour la contruction 2</i>	29
4.4	<i>Analyse comparative des scores F1 des modèles avec un true hit de 20 pour la contruction 2.</i>	29
4.5	<i>Analyse comparative des modèles avec un true hit de 10 gènes avec le modele de contruction 4.</i>	31
4.6	<i>Analyse comparative des scores F1 modèles avec un true hit de 20 gènes pour la construction 4.</i>	31
4.7	<i>Analyse comparative des scores F1 de Amine_Multiview et des modèles classiques de clustering, avec un true hit de taille 10.</i>	34

INTRODUCTION

Les dernières décennies ont été marquées par une évolution significative dans la manière dont les sciences sont abordées. Alors qu'auparavant, les disciplines scientifiques étaient souvent traitées de manière indépendante, la tendance actuelle est à la transdisciplinarité. Un exemple de cette convergence entre différentes disciplines est la bioinformatique, qui réunit la biologie, l'informatique et les statistiques mathématiques.

1.1 Contexte

La bioinformatique représente un mariage harmonieux entre la biologie et l'informatique, exploitant les outils statistiques et la modélisation informatique au service de la recherche biologique. Elle se définit par le développement et l'application de méthodes informatiques et statistiques visant à comprendre et interpréter les informations génétiques et moléculaires. Cette approche transcende les frontières traditionnelles des disciplines, offrant ainsi un moyen puissant d'aborder les complexités des données biologiques à grande échelle.

Au cœur de la bioinformatique se trouve son rôle essentiel dans l'analyse des données génomiques. Elle va au-delà de la simple manipulation de séquences génétiques pour devenir un moteur de création de cadres expérimentaux pour les biologistes. En fournissant des outils, des méthodes et des ressources indispensables, la bioinformatique facilite la planification, l'exécution et l'analyse d'expériences biologiques.

Dans ces cadres expérimentaux, les biologistes s'emploient à mesurer l'activité des gènes pour identifier des gènes ou groupes de gènes activés, potentiellement liés aux phénotypes observés. Ces analyses, basées sur la mesure de l'activité génique, visent à comprendre les relations entre les caractéristiques observées (phénotypes) et à identifier des schémas susceptibles de devenir des cibles thérapeutiques.

Notre projet, en collaboration étroite avec des biologistes, illustre cette convergence. L'objectif premier de cette collaboration est de déterminer le rôle de la protéine SigmaR1 dans le contexte spécifique du cancer du pancréas. Bien que le rôle précis de cette protéine reste à élucider, son étude revêt un intérêt particulier dans la compréhension des mécanismes génétiques liés à cette forme de cancer pour le développement de thérapies plus ciblées et efficaces.

1.2 Problématique

L'évolution vers la transdisciplinarité dans le domaine scientifique, incarnée par la bioinformatique, suscite des questionnements essentiels sur la manière dont nous abordons la recherche et la compréhension des phénomènes biologiques. Face à la complexité croissante des données génomiques, la bioinformatique doit relever plusieurs défis, appelant à une réévaluation des méthodes et des approches utilisées.

La complexité inhérente aux données génomiques à grande échelle constitue l'un des principaux défis pour la bioinformatique. Malgré ses outils avancés pour manipuler et interpréter ces données, la sélection des gènes à étudier par les biologistes demeure un enjeu crucial. La méthode conventionnelle consiste à rechercher des modules de gènes dont l'action combinée réalise une fonction spécifique, et cela se fait souvent à travers l'approche des **"top k-gènes"** les plus variables (RAPAPORT *et al.*, 2007). Cependant, cette méthode montre ses limites, car la variabilité ne garantit pas toujours la pertinence biologique. Cette limite peut parfois être observée dans le cas des gènes inflammatoires, où les gènes les plus variables représentent souvent des causes plutôt que des facteurs directement liés aux observations. Ce défi devient particulièrement pressant lorsque l'on cherche à démêler les mécanismes d'une fonction biologique. Comment facilement identifier ces modules de gènes qui varient dans l'expérience mais qui interagissent également entre eux pour une fonction biologique associée ?

La recherche de modules de gènes actifs, similaire à la détection de communautés dans les réseaux sociaux, s'avère cruciale pour comprendre les interactions génétiques. Toutefois, contrairement aux réseaux sociaux où les attributs des individus raffinent les communautés, en bioinformatique, la valeur associée aux gènes est aussi cruciale que la topologie du réseau. La recherche de modules nécessite des méthodes innovantes intégrant à la fois la topologie du graphe et la valeur des nœuds. Comment parvenir à une intégration efficace de ces deux aspects pour dévoiler des modules d'activité génique pertinents ? ou Comment réussir à fusionner de manière efficace la topologie des réseaux génétiques avec la valeur associée aux gènes pour dévoiler des modules actifs significatifs ?

L'embedding de réseau, à travers une représentation vectorielle dense en dimension réduite, émerge comme une solution prometteuse dans cette quête. Un autre aspect crucial réside dans la dynamique des valeurs des nœuds dans les réseaux biologiques. Alors que la topologie des réseaux reste stable, les valeurs des nœuds peuvent évoluer à chaque nouvelle expérience. Comment adapter les méthodes d'embedding de réseau pour garantir une représentation fidèle aux données, prenant en compte cette dynamique propre à la biologie expérimentale et à la variabilité des valeurs des nœuds ?

Notre parcours de recherche s'anime autour de ces questionnements complexes, avec pour objectif d'explorer de nouvelles perspectives dans l'utilisation de l'embedding multivues en bioinformatique. En répondant à ces défis, nous espérons contribuer à une compréhension approfondie de la recherche des modules actifs des réseaux génétiques, ouvrant la voie à des avancées significatives dans le domaine des thérapies ciblées.

1.3 Motivation du projet

Notre quête pour une compréhension approfondie des mécanismes génétiques est motivée par la nécessité de développer une méthode générique pour la détection de modules actifs. Dans le cadre de ce projet, cette méthode sera appliquée au contexte complexe du cancer du pancréas, une maladie dévastatrice souvent détectée en phase avancée mais elle pourra être également utilisée pour traiter de nombreux autres jeux de données produits par les biologistes. Les méthodes conventionnelles pour identifier les modèles de gènes qui varient dans l'expérience mais qui interagissent également entre eux pour une fonction peuvent parfois se révéler insuffisantes face à la complexité des activités biologiques.

Notre aspiration est de dépasser les limites des méthodes conventionnelles en explorant des approches plus adaptées pour faire des choix initiaux plus judicieux. Nous aspirons à contribuer à la recherche de méthodologies avancées permettant de discerner avec précision les gènes actifs impliqués dans des activités biologiques complexes. Notre projet vise à devenir un élément clé dans la recherche en bioinformatique, en utilisant l'embedding multivue. Cette technique, qui exploite la flexibilité de l'utilisation de plusieurs vues de données, promet une meilleure compréhension des mécanismes génétiques.

Dans notre projet, nous adoptons une approche multi-vue où un réseau d'interaction est considéré comme une vue et un ensemble de mesures effectuées comme une autre. Cette stratégie permet l'intégration de multiples perspectives de données, offrant une flexibilité accrue. En cas de succès avec deux vues - un graphe d'interaction et des valeurs d'expression - nous envisageons d'étendre la méthode en intégrant d'autres réseaux ou mesures. Cette stratégie est d'autant plus pertinente dans le contexte biologique où la topologie du réseau reste stable, tandis que les valeurs des nœuds varient avec chaque nouvelle expérience.

Notre intérêt est de réaliser un embedding multivues, où chaque série d'expériences pourrait être représentée par des graphes basés sur les valeurs des nœuds, formant ainsi plusieurs réseaux. Cette approche pourrait améliorer significativement notre compréhension et notre analyse des données dans des environnements biologiques complexes, ouvrant la voie à de nouvelles perspectives dans l'étude du multivue.

En somme, notre motivation repose sur la volonté de transcender les limitations actuelles et d'explorer des voies novatrices en bioinformatique. En utilisant des techniques avancées comme l'embedding multivue et le deep learning, nous visons à repousser les frontières de la recherche, en allant au-delà de la simple sélection des "top k-gènes" par leur variation. Notre but ultime est de mieux comprendre les mécanismes sous-jacents complexes entre les gènes et ainsi orienter la recherche vers des cibles thérapeutiques plus pertinentes.

1.4 Objectifs de la recherche

Notre recherche a pour objectif principal d'explorer et d'évaluer l'efficacité de la détection de modules actifs en utilisant des données multivue, en exploitant les techniques d'apprentissage profond offertes par le deep learning pour former des embeddings en dimension réduite. Cette approche novatrice vise à surmonter les limitations de l'approche traditionnelle, qui combine la structure du graphe génétique avec la valeur des nœuds, à l'instar de la méthode AMINE

(PASQUIER *et al.*, 2022). Nous envisageons également de comparer ces résultats à ceux obtenus par la méthode AMINE, reconnue pour ses performances supérieures dans la détection de modules actifs.

La flexibilité de l'approche multivue constitue le pivot central de notre investigation. Comme souligné dans notre motivation, cette méthode ouvre la porte à l'utilisation de plusieurs perspectives de données, telles que différents réseaux génétiques ou les résultats d'expériences réalisées à différents moments. En exploitant cette diversité de points de vue, la séparation des vues pourrait offrir des avantages significatifs en termes d'efficacité de calcul. Avec cette approche, nous avons l'opportunité de combiner plusieurs vues représentant des expériences distinctes. En cas de résultats supérieurs à ceux d'AMINE, notre deuxième objectif consistera à démontrer que cette approche permettra une représentation plus précise et adaptable des mécanismes génétiques, offrant ainsi une perspective plus complète pour la détection de modules actifs, compte tenu de l'interdépendance des activités biologiques.

En résumé, nos objectifs de recherche sont les suivants :

- Explorer l'efficacité de la détection de modules actifs en utilisant des données multivue : Nous souhaitons évaluer la capacité de l'embedding multivue à identifier de manière précise et complète les modules actifs dans les réseaux génétiques.
- Comparer les résultats avec la méthode AMINE : En confrontant les performances de notre approche à celles d'AMINE, reconnue pour ses succès dans la détection de modules actifs, nous cherchons à évaluer le potentiel de notre méthode à surpasser les approches existantes.
- Démontrer l'efficacité et l'adaptabilité de l'approche multivue : Si nos résultats confirment la supériorité de l'embedding multivue, notre objectif sera de démontrer comment cette approche offre une représentation plus précise et adaptable des mécanismes génétiques, fournissant ainsi une perspective plus complète pour la détection de modules actifs dans des conditions biologiques variées.

En poursuivant ces objectifs, notre ambition est de contribuer significativement à la recherche en bioinformatique et d'ouvrir de nouvelles voies pour une compréhension approfondie des mécanismes génétiques, particulièrement dans le contexte complexe du cancer du pancréas.

1.5 Questions de recherche

Notre parcours de recherche est guidé par des questionnements complexes visant à explorer les possibilités novatrices de l'embedding multivue en bioinformatique, en vue de générer une méthode générique applicable au cas particulier du cancer du pancréas. Ces interrogations émergent des lacunes identifiées dans la recherche actuelle et cherchent à éclairer les défis spécifiques liés à l'analyse des réseaux génétiques à grande échelle.

- Comment l'embedding multivue peut-il optimiser la détection de modules actifs dans les réseaux génétiques par rapport à l'approche combinant la topologie du graphe et la valeur des nœuds?
- Quels critères de construction adopter pour élaborer le graphe de la vue dépendant des valeurs de poids des nœuds, afin de mieux conserver l'information dans l'embedding multivue?

- Quels critères de collaboration entre les vues devrions-nous mettre en exergue pour implémenter l’embedding multivue de manière optimale ?
- Quels sont les avantages et les limitations de l’embedding multivue par rapport à la méthode AMINE, largement reconnue pour ses performances dans la détection de modules actifs ?

En répondant à ces questions de recherche, notre ambition est d’apporter des contributions significatives à la compréhension des mécanismes génétiques, de développer des approches innovantes en bioinformatique, et de jeter les bases pour des avancées substantielles dans le domaine des thérapies ciblées. Notre objectif ultime est de mettre en place une méthode générique pour la détection des modules actifs dans les réseaux génétiques, que nous appliquerons particulièrement dans le contexte complexe du cancer du pancréas. En combinant l’embedding multivue avec des critères de construction et de collaboration judicieux, nous espérons ouvrir de nouvelles perspectives pour une meilleure compréhension des processus biologiques et contribuer ainsi à l’élaboration de thérapies plus précises et efficaces.

ÉTAT DE L'ART

2.1 Introduction

Dans l'analyse et le traitement des données, il est crucial de comprendre les interactions et les corrélations entre différents ensembles de données. L'approche des embeddings multi-vues se distingue par son efficacité et sa flexibilité, jouant un rôle essentiel dans l'intégration d'ensembles de données hétérogènes. Cette méthode offre une représentation unifiée et globale qui saisit avec précision la complexité des données. Cette capacité est particulièrement utile pour extraire des informations pertinentes lors du traitement de données, que ce soit pour le regroupement (clustering), la classification ou la prédiction de liens. Compte tenu de l'importance d'intégrer les données dans un espace unifié, de nombreuses théories ont été élaborées, donnant naissance à des modèles à la fois robustes et efficaces.

2.2 Théories et modèles existants

2.2.1 Notion de Multivues

La notion de multivues fait référence à une approche analytique dans laquelle plusieurs "vues" ou représentations d'un même ensemble de données sont examinées et intégrées pour obtenir une compréhension plus profonde et détaillée. Cette méthode est extrêmement utile dans les domaines caractérisés par des données complexes et multidimensionnelles. Prenons l'exemple de la bioinformatique : une vue peut représenter des données génomiques, tandis qu'une autre pourrait se concentrer sur les données protéomiques ou transcriptomiques du même échantillon biologique. L'approche multivues permet d'analyser ces différentes perspectives simultanément, offrant ainsi une compréhension plus complète et nuancée du sujet étudié. Cette méthode diffère fondamentalement de l'approche univue, qui se limite à une seule perspective des données. En intégrant diverses vues, les chercheurs peuvent identifier des corrélations et des interactions qui ne seraient pas évidentes en examinant les vues séparément. Cela conduit à des insights plus riches et ouvre la voie à de nouvelles découvertes dans leur domaine de recherche.

L'importance d'intégrer diverses sources d'information dans un cadre unifié a conduit au développement de nombreuses théories, parfois basées sur des principes statistiques et probabilistes. Parmi cet éventail de concepts, quatre ont particulièrement retenu notre attention, chacun apportant une perspective unique et enrichissante à notre compréhension du multivues.

Ces théories nous permettent de naviguer et d'exploiter efficacement la complexité inhérente aux données multidimensionnelles, ouvrant ainsi de nouvelles voies dans la recherche et l'analyse de données.

2.2.2 Factorisation Matricielle par des Composants Liés pour l'Intégration Unifié

Dans leur approche de la "Factorisation Matricielle par des Composants Liés pour l'Intégration Unifiée", (IRINA GAYNANOVA, 2017) apportent une contribution notable avec leur modèle SLIDE (Structural Learning and Integrative DEcomposition). Ce modèle se distingue par l'intégration de structures partiellement partagées dans la factorisation matricielle des données multivues, une avancée par rapport au modèle JIVE (Joint and Individual Variation Explained) (LOCK et al., 2013). En fait dans la plupart des ensembles multivues réels, il existe des données dont les instances ne sont pas présents dans toutes les vues.

SLIDE offre une représentation efficace des données multivues à travers des composants liés, utilisée pour la réduction dimensionnelle exploratoire et l'analyse d'association entre les vues. Cette intégration de composants partiellement partagés aborde un défi important dans la factorisation structurale des données multivues. Dans les études empiriques, notamment avec des données sur le cancer issues du répertoire "The Cancer Genome Atlas", Le modèle SLIDE a démontré d'excellentes performances en termes d'estimation du signal et de sélection des composants.

Cependant, SLIDE présente des limitations. La méthode de détermination du nombre de composants pour chaque type (partagés, individuels, partiellement partagés) pour chaque vue, bien qu'innovante, peut rencontrer des difficultés en termes de complexité computationnelle et de précision dans des contextes de données variés. De plus, l'utilisation d'un cadre de factorisation matricielle pénalisée pour réduire la complexité peut limiter la flexibilité et l'adaptabilité du modèle dans certaines applications.

Ces limitations ouvrent la voie à l'exploration de nouvelles théories pour unifier les données multivues dans un cadre unifié, suggérant la nécessité de modèles plus flexibles et adaptatifs.

2.2.3 La Théorie Basée sur le cadre de probabiliste

La transformation des données issues de différentes vues dans un espace unifié, appelé espace partagé, est un processus fondamental dans le domaine de l'apprentissage automatique. Le modèle **Probabilistic Multi-view Graph Embedding (PMvGE)** (OKUNO et al., 2018) offre une solution innovante à ce défi, en combinant des techniques avancées d'incorporation de graphes (graph embedding) avec des approches probabilistes pour unifier les données multivues dans un espace commun.

Dans le cadre de PMvGE, les données de chaque vue sont initialement transformées en vecteurs de caractéristiques au sein d'un espace partagé. Cette étape est essentielle pour aligner les données provenant de sources diverses. Elle est effectuée à l'aide de réseaux neuronaux, où les données d'entrée $x^{(v)}$ sont converties en vecteurs de caractéristiques $y^{(v)}$ dans l'espace partagé selon la fonction

$$y^{(v)} = f^{(v)}(x^{(v)}; \theta^{(v)}) \quad (2.1)$$

Le cœur de PMvGE réside dans sa capacité à modéliser la probabilité d'association entre des paires de vecteurs de caractéristiques issus de différentes vues. Cette probabilité est estimée par le produit scalaire des vecteurs de caractéristiques, exprimé par

$$P(y^{(v)}, y^{(u)}) = \sigma(y^{(v)T} y^{(u)}) \quad (2.2)$$

Cette méthode probabiliste dans le cas de l'implémentation pourrait être une sigmoïde, cela permettrait non seulement d'identifier les associations entre les données, mais aussi d'en quantifier la force.

L'approche probabiliste pour modéliser les associations entre les vecteurs de caractéristiques est un élément clé de PMvGE. Elle permet de détecter et de mesurer la force des associations entre les données de différentes vues, offrant ainsi une compréhension plus détaillée des relations entre les données. L'objectif principal de PMvGE est de maximiser la vraisemblance des associations observées dans les données multivues. La fonction de vraisemblance $L(\theta)$ est optimisée pour ajuster les paramètres du modèle, suivant l'équation :

$$L(\theta) = \sum_{v,u} \sum_{i,j} w_{ij}^{(vu)} \log P(y_i^{(v)}, y_j^{(u)}; \theta) + (1 - w_{ij}^{(vu)}) \log(1 - P(y_i^{(v)}, y_j^{(u)}; \theta)) \quad (2.3)$$

Cette optimisation est cruciale pour assurer que le modèle reflète fidèlement les relations complexes entre les données multivues pour ce distinguer des autres modèles. **PMvGE** (Probabilistic Multi-view Graph Embedding) se distingue de CDMCA par l'introduction de transformations non linéaires. Les réseaux neuronaux employés par PMvGE transforment les données de chaque vue en vecteurs de caractéristiques dans un espace partagé. Cette approche non linéaire permet à PMvGE de capturer des associations plus complexes et subtiles entre les vues, surpassant ainsi les capacités de CDMCA. En exploitant des concepts d'incorporation de graphes et de modélisation probabiliste, PMvGE facilite une analyse approfondie et une compréhension des données multivues. Toutefois, la transformation des données via des réseaux neuronaux peut être complexe et nécessite un ajustement précis des paramètres. De plus, la performance du modèle peut être limitée par la taille et la qualité des données disponibles.

Ces limitations soulignent l'importance de développer de nouvelles méthodes pour l'intégration de données multivues, incluant des techniques d'apprentissage plus simples

2.2.4 La Théorie Basée sur les Méthodes de Conflation de distribution de probabilité :

Le concept de conflation de distributions de probabilité joue un rôle essentiel dans la création d'embeddings unifiés. Cette approche (MITRA *et al.*, 2020), axée sur l'intégration et la fusion de données multivues en une représentation unifiée, repose sur deux principes fondamentaux : la symétrie probabiliste et la réduction de la divergence de Kullback-Leibler. Dans ce contexte, chaque vue possède sa propre distribution de probabilité. Ces distributions sont calculées avant la fusion des données. Après cette fusion, une nouvelle distribution est obtenue dans l'espace unifié. Cela contraste avec les modèles probabilistes qui calculent la probabilité directement dans l'espace unifié.

1. Probabilité Symétrique

- Chaque point dans un ensemble de données, relatif à une vue spécifique, calcule une probabilité symétrique pour chaque autre point potentiellement voisin.

- La probabilité, notée p_{ij}^v , est déterminée par une formule basée sur l'exponentielle de la dissimilarité au carré entre les points, divisée par la somme de ces exponentielles pour tous les voisins potentiels.
- La formule spécifique est :

$$p_{ij}^v = \frac{\exp(-d_{ij}^v)^2}{\sum_k \exp(-d_{ik}^v)^2} \quad (2.4)$$

où d_{ij}^v désigne la dissimilarité entre les échantillons i et j dans la vue v .

2. Réduction de la Divergence de Kullback-Leibler

- L'objectif est de minimiser la divergence de Kullback-Leibler entre la distribution de probabilité dans l'espace de haute dimension et sa contrepartie dans l'espace de basse dimension (embedding).
- Cette minimisation est cruciale pour optimiser l'embedding afin qu'il représente fidèlement les relations de probabilité des points dans l'espace de haute dimension.
- La divergence est exprimée par :

$$C = KL(P||Q) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}}$$

Ce processus utilise des principes de probabilité et d'optimisation pour transformer efficacement des données multi-vues de haute dimension en un espace unifié de basse dimension, préservant ainsi la structure essentielle des données originelles.

2.2.5 La Théorie Basée sur les Méthodes de Collaboration de données :

Cette approche se concentre sur l'analyse collaborative des données (ATA et al., 2020) afin de construire un embedding unifié. Elle repose sur trois propriétés essentielles pour explorer les interactions et relations entre les données dans différentes vues. Ces propriétés sont : la diversité, la collaboration de premier ordre et la collaboration de second ordre

1. Diversité

Cette phase vise à capturer l'unicité de chaque vue en produisant des paires d'échantillons ou individus qui sont réellement connectés au sein de chaque vue. Ces paires illustrent la similarité entre les échantillons dans une vue donnée. Pour une vue v , un ensemble de paires intra-vue $\Omega(v)$ est constitué, chaque paire $(u(v), w(v)) \in \Omega(v)$ comprenant un échantillon central $u(v)$ et un échantillon contextuel $w(v)$. L'objectif est d'optimiser la probabilité de prédire l'échantillon contextuel à partir de l'échantillon central, en réduisant la perte $Div(\Theta)$, définie comme :

$$L_{Div}(\Theta) = - \sum_{v \in V} \sum_{(i(v), j(v)) \in \Omega(v)} \log P(j(v)|i(v); \Theta) \quad (2.5)$$

2. Collaboration de Premier Ordre

Bien que les différentes vues d'un réseau multi-vues présentent de la diversité, elles convergent finalement vers un ensemble commun d'individu ou échantillon. Les instances d'un même individu à travers différentes vues décrivent fondamentalement la même entité.

Cette collaboration de premier ordre vise à aligner les représentations spécifiques d'un même échantillon à travers différentes vues. Pour ce faire, des paires intra-échantillon sont formées pour toute les vues dans laquelle l'instance de l'échantillon existe. Cette relation peut être vue comme une relation identitaire car chaque paire représente le même individu observé dans différentes vues. Ainsi, nous avons les paires $(u(v), u(v'))$, où

- $u(v)$ est l'échantillon dans la vue v
- $u(v')$ est l'échantillon dans la vue v' .

Comme il s'agit du même individu observé sous différentes perspectives, la perte $C1(\Theta)$ est minimiser pour maximiser la similarité des représentations vectorielles de l'échantillon dans les différentes vues. Cette perte $C1(\Theta)$ est exprimée de la manière suivante :

$$L_{C1}(\Theta) = - \sum_{v \in V} \sum_{(i(v), \cdot) \in \Omega(v)} \sum_{v' \neq v} \log P(i(v')|i(v); \Theta) \quad (2.6)$$

ou

- $P(i(v')|i(v); \Theta)$ exprime la probabilité de prédire correctement la représentation d'un nœud dans une vue v' , en fonction de sa représentation dans une vue v , sous les paramètres Θ .
- $i(v)$ et $i(v')$ indiquent respectivement le nœud central dans la vue v et sa représentation dans une autre vue v' .

L'objectif de cette fonction de perte $L_{C1}(\Theta)$ est de garantir que les embeddings d'un même nœud soient similaires à travers les différentes vues. En optimisant cette fonction, le modèle aligne efficacement les représentations de chaque nœud à travers les vues, assurant ainsi que les caractéristiques fondamentales du nœud sont cohérentes et fidèlement représentées dans l'ensemble du réseau multi-vues.

3. Collaboration de Second Ordre

Cette collaboration utilise les associations entre les échantillons d'une vue pour améliorer la collaboration entre différentes vues. Des paires d'échantillons croisées sont établies selon les associations entre les échantillons de chaque vue, dans le but de mettre à jour les représentations d'un échantillon pour qu'elles ressemblent à celles des échantillons associés dans une autre vue. La perte $C2(\Theta)$ est minimisée et formulée comme suit :

$$L_{C2}(\Theta) = - \sum_{v \in V} \sum_{(i(v), j(v)) \in \Omega(v)} \sum_{v' \neq v} \log P(j(v')|i(v); \Theta) \quad (2.7)$$

En combinant ces différentes propriétés de collaboration de données, la méthode développe un embedding unifié qui intègre la diversité intrinsèque à chaque vue et la collaboration entre les vues, tout en prenant en compte les relations de second ordre entre les échantillons.

2.3 Présentation des principaux modèles générant les embeddings multivues

2.3.1 Le modèle MvNE (Multi-view Neighbourhood Embedding)

Le modèle Multi-view Neighbourhood Embedding (MvNE) représente une approche sophistiquée pour l'intégration unifiée de données multi-vues. Cette méthode se décompose en plusieurs étapes clés.

1. Génération de l'ensemble de données unifié :

La première étape implique la fusion des différentes vues de l'ensemble de données en une seule représentation unifiée. Chaque vue capte un aspect distinct des données. En cas d'absence de certaines caractéristiques ou échantillons dans des vues, ils seront remplacés par des valeurs nulles. Ainsi, l'ensemble de données unifié englobe l'intégralité des échantillons et caractéristiques issues des diverses vues.

2. Autoencodeur empilé (SAE) pour l'intégration initiale :

L'autoencodeur empilé (SAE) est un modèle d'apprentissage profond non supervisé. Il sera utilisé sur l'ensemble de données unifié pour générer l'embedding initiale. Le SAE se compose d'un encodeur et d'un décodeur. L'encodeur transforme les données d'entrée en un espace de dimension réduite, formant ainsi l'embedding initial. Et le Décodeur du SAE essayer de reconstitue l'échantillon originale à partir de sa représentation fournie par l'Encodeur. Le modèle SAE sera entraîné en minimisant l'erreur de reconstruction sur la différence entre les données d'entrée x et la sortie reconstruite \hat{x} , mesurée par l'erreur quadratique moyenne suivante :

$$\text{MSE} = \frac{1}{n} \sum_{i=1}^n (x_i - \hat{x}_i)^2 \quad (2.8)$$

3. Génération de distributions de probabilités unifiées de l'ensemble de données concaténé :

Pour chaque échantillon de l'ensemble de données unifié, des probabilités symétriques p_{ij} sont calculées afin d'estimer la probabilité de sélectionner un point voisin. La formule est la suivante :

$$p_{ij} = \frac{\prod_v p_{vij}}{\prod_v p_{vij} + \prod_v \sum_{k \neq j} p_{vik}} \quad (2.9)$$

- p_{ij} : Probabilité combinée que l'échantillon i choisisse l'échantillon j comme voisin, en tenant compte de toutes les vues.
- $\prod_v p_{vij}$: Produit des probabilités p_{vij} sur toutes les vues v . Chaque p_{vij} indique la probabilité que, dans la vue v , l'échantillon i choisisse j comme voisin.
- $\prod_v \sum_{k \neq j} p_{vik}$: Produit des sommes des probabilités que l'échantillon i choisisse un autre échantillon k (différent de j) comme voisin, calculé pour chaque vue v .

Les probabilités p_{vij} , basées sur une distribution gaussienne, sont préalablement calculées séparément pour chaque vue. La probabilité est donnée par

$$p_{ij}^v = \frac{\exp(-d_{ij}^v)^2}{\sum_k \exp(-d_{ik}^v)^2} \quad (2.10)$$

où d_{ij}^v représente la dissimilarité entre les échantillons i et j dans la vue v .

Cette approche est particulièrement pertinente dans les scénarios nécessitant l'intégration de données provenant de sources diverses pour obtenir une vue complète et unifiée, comme dans l'analyse de données multi-omiques ou la fusion de données issues de capteurs multiples.

4. Génération de distributions de probabilités dans l'espace intégré :

Dans cette étape, nous calculons la probabilité symétrique q_{ij} pour chaque échantillon dans l'espace latent de l'Autoencodeur empilé (SAE). Cette probabilité représente la chance que le point i sélectionne le point j comme voisin. Elle est basée sur une distribution de Student t. L'équation correspondante est

$$q_{ij} = \frac{1 + \|y_i - y_j\|^2^{-1}}{\sum_{l \neq k} (1 + \|y_l - y_k\|^2)^{-1}} \quad (2.11)$$

, où y_i et y_j désignent les représentations des points i et j dans l'espace d'embedding. La norme $\|y_i - y_j\|$ mesure la distance euclidienne entre ces deux points dans cet espace.

Cette formulation permet de capturer les relations de proximité entre les échantillons dans un espace de dimension réduite, facilitant ainsi la compréhension des structures intrinsèques des données multi-vues.

5. **Optimisation de l'intégration unifiée** L'objectif final est de trouver un embedding dans un espace à faible dimension qui reflète au mieux la distribution de probabilité unifiée. Cette optimisation est réalisée en minimisant la divergence de Kullback-Leibler (KL) entre la distribution de probabilité unifiée et celle de l'espace intégré. La formule de divergence KL est exprimée par :

$$C = KL(P||Q) = \sum_{i,j} p_{ij} \log \frac{p_{ij}}{q_{ij}} \quad (2.12)$$

La descente de gradient est utilisée pour ajuster itérativement la position des échantillons dans l'espace intégré.

2.3.2 Le modèle MANE (Multi-View Collaborative Network Embedding)

Le modèle MANE, développé pour les réseaux multi-vues, part du principe que ces réseaux sont formés de graphes non orientés. Ce modèle s'appuie sur la théorie des modèles de collaboration de données, en examinant de près les interactions et les relations entre les différentes vues. Il utilise la fonction exponentielle pour calculer les probabilités de prédire avec précision la représentation d'un nœud contextuel en fonction des nœuds central. les etapes de la contruction de l'embedding unifié des noeud est la suivantes :

1. Construction de l'Ensemble de Paires de Nœuds

Cette phase implique la création d'un ensemble de paires de nœuds, divisées en trois catégories :

(a) Paires de Nœuds Intra-Vue

Pour générer ce type de paires de nœuds, le modèle définit des marches aléatoires au sein de chaque vue afin de générer des séquences de nœuds. Ces séquences ont pour but de révéler la structure topologique de chaque vue. Ensuite, ces séquences sont tronquées en paires de nœuds

(b) Paires de Nœuds Inter-Vues Intra-Nœud

Ces paires sont composées d'instances d'un même nœud dans différentes vues.

(c) Paires de Nœuds Inter-Vues et Inter-Nœuds

Incluant des paires d'un nœud dans une vue et de différents nœuds dans une autre, ces associations aident à déchiffrer la collaboration de second ordre.

2. Définir la dimension de la representation vectorielle des noeuds dans l'espace unifie

L'approche MANE consiste à représenter un nœud dans chaque vue, puis à concaténer la représentation de ce nœud dans chaque vue pour former sa représentation finale. L'objectif est de capturer la diversité propre à chaque vue en traitant les opérations sur les paires de nœuds intra-vue de manière distincte. Pour s'assurer que chaque vue contribue de manière égale à la représentation globale du réseau, le cadre conceptuel de MANE divise l'espace d'embedding entre les différentes vues de manière équitable. Elle définit la fonction de représentation de l'embedding dans une vue de la manière suivante :

$$\begin{cases} f^v : U \rightarrow \mathbb{R}^{[D/|V|]} \\ i \mapsto f_{i(v)} \end{cases} \quad (2.13)$$

- Ici, U est l'ensemble des nœuds.
- $\mathbb{R}^{[D/|V|]}$ est l'ensemble de sortie, représentant l'espace vectoriel dans lequel les embeddings sont placés.
- D est la dimension totale de l'espace d'embedding,
- $|V|$ est le nombre de vues dans le réseau.
- $f_{i(v)}$ représente le vecteur dense du nœud i pour la vue v

3. Calcul de la fonction de Perte pour l'entrainement du modele

(a) definition de la fontion de perte :

pour prendre en compte la diversité intra-vue ainsi que les interactions inter-vues, le framework MANE utilise une combinaison linéaire de fonctions de pertes consue sur les trois typees d'emsembles de noeuds (les trois fonctions de pertes definir dans la theories des modeles de collaboration de données). elle est de finide la maniere suivante

$$Loss = L_{Div} + \alpha \cdot L_{C1} + \beta \cdot L_{C2} \quad (2.14)$$

où :

- L_{Div} : Représente la perte liée à la diversité intra-vue. Elle vise à capturer la diversité et les caractéristiques uniques de chaque vue individuelle dans le réseau multi-vues.
- L_{C1} : Correspond à la perte de collaboration de premier ordre. Cette composante de la perte aligne les représentations spécifiques d'un même nœud à travers différentes vues, assurant la cohérence et la similitude des représentations d'un nœud d'une vue à l'autre vue .
- L_{C2} : Représente la perte de collaboration de second ordre. Elle se concentre sur les relations entre les nœuds à travers différentes vues, exploitant les associations entre les nœuds d'une vue pour renforcer la collaboration entre les différentes vues.

- α et β : Sont des hyperparamètres qui déterminent l'importance relative des pertes de collaboration de premier et de second ordre par rapport à la perte de diversité. Ces hyperparamètres sont ajustés pour équilibrer le modèle en fonction des particularités du réseau et des objectifs de l'analyse.

(b) **definition de la fonction de probabilité**

La probabilité $P(j(v)|i(v); \Theta)$ est mise en pratique par le modèle MANE via une fonction softmax à travers l'implemetation de l'équation :

$$P(j(v)|i(v); \Theta) = \log \left(\frac{\exp(f_{i(v)} \cdot f_{j(v)})}{\sum_{k \in U} \exp(f_{i(v)} \cdot f_{k(v)})} \right) \quad (2.15)$$

Où :

- $f_{i(v)}$ est la representation vectoreille dense du noeud i dans la vue v
- $f_{i(v')}$ est la representation vectoreille dense du noeud i dans la vue v'
- La fonction exponentielle $\exp(f_{i(v)} \cdot f_{j(v)})$ transforme le score de similarité en probabilité.
- $P(j(v)|i(v); \Theta)$ est la probabilité de prédire correctement la représentation du nœud j dans une vue v' , en se basant sur sa connexion avec le nœud i dans la vue v , sous les paramètres du modèle Θ .

Ainsi , nous aurons expresiion des des differents

$$L_{Div}(\Theta) = - \sum_{v \in V} \sum_{(i(v), j(v)) \in \Omega(v)} \log \left(\frac{\exp(f_{i(v)} \cdot f_{j(v)})}{\sum_{k \in U} \exp(f_{i(v)} \cdot f_{k(v)})} \right) \quad (2.16)$$

$$L_{C1}(\Theta) = - \sum_{v \in V} \sum_{(i(v), \cdot) \in \Omega(v)} \sum_{v' \neq v} \log \left(\frac{\exp(f_{i(v)} \cdot f_{i(v')})}{\sum_{k \in U} \exp(f_{k(v)} \cdot f_{i(v')})} \right) \quad (2.17)$$

$$L_{C2}(\Theta) = - \sum_{v \in V} \sum_{(i(v), j(v)) \in \Omega(v)} \sum_{v' \neq v} \log \left(\frac{\exp(f_{i(v)} \cdot f_{j(v')})}{\sum_{k \in U} \exp(f_{i(v)} \cdot f_{k(v')})} \right) \quad (2.18)$$

Où :

- V représente l'ensemble des vues dans le réseau multi-vues.
- $\Omega(v)$ désigne les paires de nœuds connectés dans la vue v .
- $i(v)$ et $j(v)$ sont des nœuds connectés dans la vue v , et $j(v')$ est la représentation du nœud j dans une autre vue v' .
- U est l'ensemble total des nœuds dans le réseau.

MÉTHODOLOGIE

Notre recherche, axée sur l'efficacité de la détection de modules actifs dans des réseaux biologiques, adopte une approche innovante basée sur des données multivues. Afin de rester en continuité avec les recherches précédentes, nous envisageons d'intégrer cette approche multivue au sein du framework AMINE, reconnu pour ses performances supérieures dans l'analyse de données univues avec des graphes pondérés au niveau des nœuds. Pour cela, nous établirons un cadre conceptuel qui intègre la construction de différentes vues du graphe et d'autres règles sur le processus d'intégration d'embeddings multivues dans AMINE.

3.1 Cadre conceptuel

Pour atteindre nos objectifs, nous avons défini un cadre conceptuel articulé autour de plusieurs processus clés. Initialement, nous construisons deux vues distinctes à partir d'un graphe de données pondéré, adapté au framework AMINE. Cette approche vise à assurer une continuité avec les recherches précédentes utilisant AMINE, qui a démontré d'excellents résultats dans la détection de modules actifs. La première vue sera influencée par la topologie structurelle du graphe, tandis que la seconde se concentrera sur les poids des nœuds, reflétant les p-values des protéines dans le graphe.

L'étape suivante consiste à unifier ces deux vues dans un espace vectoriel commun. Cette fusion est cruciale pour créer une représentation complète et intégrée des données. Après avoir établi cet espace vectoriel unifié, nous appliquerons l'algorithme glouton d'AMINE sur l'embedding résultant. Cependant, avant d'appliquer cet algorithme, il est essentiel de définir des métriques de similarité adaptées à ce nouvel embedding. Ces métriques seront déterminantes pour évaluer la pertinence et l'efficacité de notre approche multivue dans la détection de modules actifs.

Notre démarche implique également une analyse approfondie des caractéristiques intrinsèques des données biologiques (au niveau de la validation des données), en tenant compte de la variabilité et de la complexité des interactions génétiques. En intégrant ces aspects, nous visons à améliorer la précision de la détection des modules actifs.

Nous illustrons notre cadre conceptuel avec la Figure 3.1.

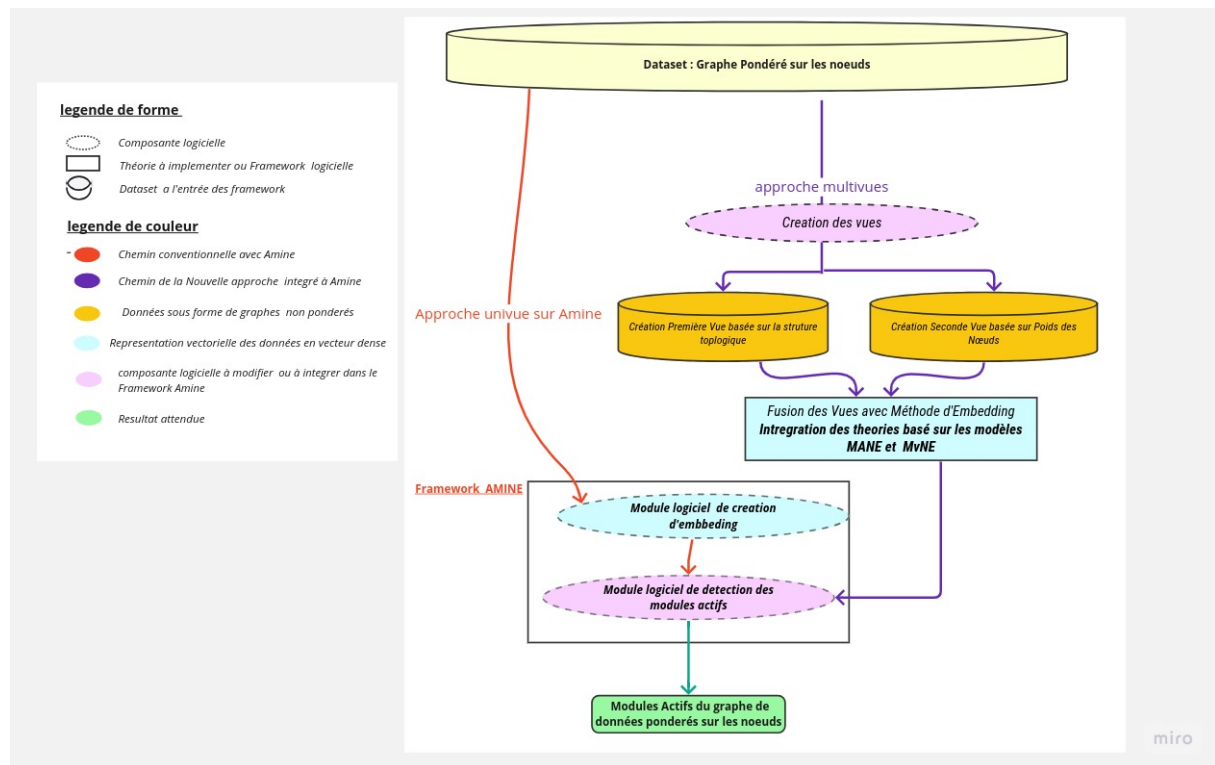


FIGURE 3.1 – Illustration de notre cadre conceptuel via l'approche multivues

3.2 Conception du modèle

3.2.1 Modèles de Construction de la Première Vue

La construction des vues est un élément central de notre recherche. La première vue, en particulier, joue un rôle crucial dans l'analyse des données. Cette vue est conçue comme une réplique fidèle du graphe d'origine, mais sans pondération au niveau des nœuds. En adoptant cette méthode, nous préservons intégralement la structure topologique du graphe, ce qui nous permet de capturer et d'analyser les relations et les connexions intrinsèques entre les différents nœuds. Cette préservation de la structure topologique est essentielle pour deux raisons. Premièrement, elle permet une interprétation plus directe et intuitive des relations entre les nœuds, car chaque lien ou connexion dans le graphe reflète une interaction ou une association réelle, non influencée par des poids. Deuxièmement, en conservant la structure originale, elle nous permet de fidéliser les propriétés de connectivité des protéines pour une activité biologique.

3.2.2 Modèles de Construction de la Deuxième Vue

Pour la deuxième vue, nous avons élaboré quatre modèles de construction distincts, chacun proposant deux sous-variantes basées sur des règles de filtrage spécifiques. Les nœuds qui ne respectent pas ces règles seront traités soit comme des singletons (première variante) soit retirés complètement (seconde variante)

1. **Modèle Construction 1** : Filtrage des composantes connexes du graphe où les relations sont établies uniquement avec les nœuds ayant une p-value inférieure ou égale à 0.05.

- Sous-Variante 1 : Les autres nœuds non connectés sont considérés comme des singletons.
 - Sous-Variante 2 : Ces nœuds sont retirés.
2. **Modèle Construction 2** : Filtrage de tous les nœuds avec une p-value inférieure à 0.05 pour créer une composante connexe complète avec cet ensemble de nœuds de p_value inférieur ou égale à 0.05 .
- Sous-Variante 1 : Les nœuds avec une p-value supérieure à 0.05 deviennent des singletons.
 - Sous-Variante 2 : Ces nœuds sont retirés.
3. **Modèle Construction 3** : La seconde vue a la même structure topologique que la première, mais avec des arêtes supplémentaires entre les nœuds de p-value inférieure à 0.05 pour former un sous-graphe complet.
- Sous-Variante 1 : Traitement des nœuds singletons existants.
 - Sous-Variante 2 : dans cette construction la sous variante n'existe pas car dans le graphe d'origine nous n'avons pas de nœud singletons, lors de la construction du graphe nous veillons à ce que toutes les composantes connexes soient liées par au moins une arête .
4. **Modèle Construction 4** : Construction d'un graphe en ajoutant des arêtes entre les nœuds dont la différence de "ZScores" est inférieure ou égale à 0.4.
- Sous-Variante 1 : Les nœuds non connectés sont considérés comme des singletons.
 - Sous-Variante 2 : Ces nœuds sont retirés

La sous-variante 1 de la vue 2 sera considérée comme une vue complète lors de l'intégration des vues, car même si les nœuds sont des singletons, ils représentent des instances existantes dans la vue. En revanche, la sous-variante 2 sera considérée comme une vue incomplète, car nous avons des instances de nœuds qui ne sont pas présentes dans la deuxième vue lors de l'intégration unifiée. .

3.2.3 choix du Modèles de theorie d'unification d'embedding

Le choix entre ces variantes influencera la méthode d'intégration unifiée, en tenant compte des capacités des frameworks comme MANE (Multi-View Collaborative Network Embedding), adapté aux vues complètes, et MvMe (Multi-view Neighbourhood Embedding), plus efficace pour les vues incomplètes. Des tests préliminaires seront réalisés pour évaluer l'efficacité de ces différentes approches afin d'adopter un modèle final.

La sélection du modèle optimal pour la deuxième vue est essentielle, car elle influence directement la qualité des informations utilisées dans notre analyse finale. Chaque modèle d'intégration unifiée sera évalué sous les deux variantes de la seconde vue. L'objectif est de déterminer la combinaison la plus efficace pour la détection de modules actifs, en se basant sur des critères tels que la précision, la robustesse et la pertinence biologique, ainsi que la métrique de similarité de validation sur l'embedding unifié. En appliquant ces critères d'évaluation rigoureux, nous nous assurons que le modèle choisi offre non seulement une représentation précise des données biologiques, mais est également robuste et adapté à divers contextes d'analyse.

3.2.4 Choix des Mesures de Similarité dans l'Espace Unifié

Nous avons défini trois métriques dans l'espace unifié pour évaluer la similarité des représentations vectorielles des nœuds. Ces métriques comprennent la similarité cosinus, la distance euclidienne, et la similarité de Pearson.

- **La Similarité de Pearson :** La similarité de Pearson (ou corrélation de Pearson) mesure la corrélation linéaire entre deux variables aléatoires. La formule pour calculer la corrélation de Pearson entre deux vecteurs X et Y est :

$$r = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^n (X_i - \bar{X})^2} \sqrt{\sum_{i=1}^n (Y_i - \bar{Y})^2}} \quad (3.1)$$

où \bar{X} et \bar{Y} sont les moyennes des vecteurs X et Y respectivement. La valeur de r se situe entre -1 et 1.

- **La Similarité Cosinus :**

La similarité cosinus mesure l'angle entre deux vecteurs dans un espace vectoriel. Elle est calculée comme le produit scalaire des vecteurs normalisé par le produit de leurs normes. La formule est :

$$\text{similarity_cosine} = \frac{X \cdot Y}{\|X\| \|Y\|} \quad (3.2)$$

où $X \cdot Y$ est le produit scalaire des vecteurs X et Y , et $\|X\|$ et $\|Y\|$ sont les normes de X et Y .

- **La Similarité Euclidienne :**

La similarité euclidienne (ou distance euclidienne normalisée) est basée sur la distance euclidienne entre deux points. Si D est la distance euclidienne entre deux vecteurs X et Y , la similarité S est calculée comme :

$$\text{Similarity_euclidienne} = \frac{1}{1 + D} \quad (3.3)$$

où D est la distance euclidienne, calculée comme :

$$D = \sqrt{\sum_{i=1}^n (X_i - Y_i)^2}$$

Chacune de ces métriques de similarité apporte une perspective différente sur la manière dont les nœuds ou les modules sont liés ou distants les uns des autres dans l'espace vectoriel.

3.2.5 Critères d'Évaluation

Pour évaluer et valider nos modèles dans le but d'assurer leur efficacité et leur fiabilité, nous avons élaboré deux critères principaux, chacun ciblant un aspect différent de la performance du modèle :

1. la Métrique Précision :

La capacité du modèle à identifier correctement les modules actifs est mesurée par le score F1. Ce critère évalue l'équilibre entre la précision (proportion de vrais positifs parmi les identifications) et le rappel (proportion de vrais positifs parmi les cas réels). La formule du score F1 est la suivante :

$$\text{F1 Score} = 2 \times \frac{\text{Précision} \times \text{Rappel}}{\text{Précision} + \text{Rappel}} \quad (3.4)$$

Où :

- La précision est calculée comme $\frac{TP}{TP+FP}$
- Le rappel est calculé comme $\frac{TP}{TP+FN}$
- TP représente le nombre de vrais positifs (Vrai positives).
- FP représente le nombre de faux positifs (Faux positives).
- FN représente le nombre de faux négatifs (Faux negatives).

2. Robustesse : (Test sur Ensemble de Données Artificiel de 1000 graphes pondérés)

Le critère de "Robustesse" constituera un test essentiel qui sera effectué sur notre modèle, visant à évaluer sa résilience et sa fiabilité face à des variations aléatoires. Cette évaluation spécifique s'effectuera à l'aide d'un ensemble de données artificielles comprenant pas moins de 1000 graphes.

L'objectif central de ce test résidera dans la capacité du modèle à maintenir des performances constantes malgré les fluctuations imprévisibles qui seront introduites délibérément par l'algorithme de génération des données artificielles. En d'autres termes, le modèle sera soumis à une série de scénarios où chaque graphe sera généré de manière aléatoire, simulant ainsi des conditions réalistes. La mesure de la robustesse du modèle reposera sur sa capacité à fournir des résultats cohérents et fiables dans des conditions dynamiques. Cette évaluation approfondie visera à assurer que notre modèle puisse maintenir des performances stables.

En somme, ce test de robustesse sera un indicateur crucial, puisque nous envisagerons de réaliser une étude statistique de chaque modèle de machine learning sur l'ensemble des 1000 valeurs de F1 scores collectées. L'objectif sera de déterminer notamment le modèle qui fournira les meilleures statistiques sur l'ensemble des 1000 graphes.

Nous pouvons illustrons notre modele conceptuel avec par la Figure 3.2.

3.3 Collecte de données :

Dans le cadre de notre étude, nous avons opté pour la création de données artificielles afin de simuler des réseaux biologiques complexes. Cette méthode nous permet de contrôler précisément les paramètres du réseau, ce qui est crucial pour tester l'efficacité de notre modèle de détection de modules actifs. Après avoir validé le modèle sur ces données artificielles, nous l'appliquerons sur des données réelles, celles générées dans l'article de (CHIOU *et al.*, 2017) . Le modèle AMINE a également été évalué sur ces données.

3.3.1 Description de l'Algorithme de Génération de Données

Cet algorithme génère des données en simulant des réseaux biologiques complexes à l'aide de graphes sans échelle. Il s'appuie sur une version améliorée de la méthode de Barabási-Albert, idéale pour modéliser des réseaux où la distribution des degrés des nœuds suit une loi de puissance. En plus d'intégrer la méthode de Barabási-Albert, l'algorithme comprend trois fonctions principales qui sont utiles dans la construction des clusters :

1. **neighbors** : son Rôle est de détermine le nombre de voisins d'un nœud spécifique à une distance d'orde k . Partant d'un nœud initial (*start*), elle explore et compte les voisins

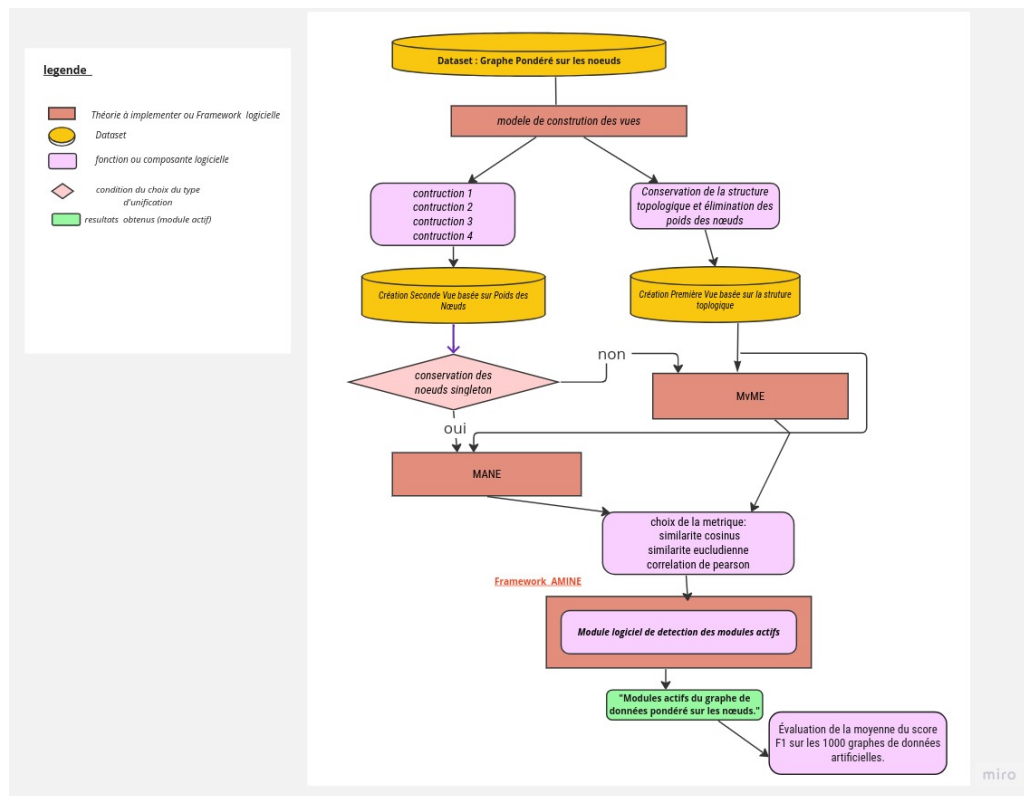


FIGURE 3.2 – Illustration de notre modèle conceptuel

jusqu'à atteindre le niveau k , offrant ainsi une vue sur la connectivité locale du nœud dans le graphe.

2. **knbrs** : Cette fonction identifie tous les voisins d'un nœud à un niveau k . Similaire à *neighbors_order*, elle retourne un ensemble des voisins jusqu'au niveau k , permettant de comprendre les interactions potentielles d'un nœud donné.
3. **get_seeds** : Cette fonction sélectionne des nœuds initiaux pour la création de modules dans le graphe. et veuille à ce que les nœuds soient suffisamment éloignés les uns des autres (*min_distance*), assurant ainsi une distribution équilibrée et non chevauchante des modules dans le graphe.

Chaque sous-fonction joue un rôle clé dans l'élaboration d'un graphe complexe et structuré, reflétant les propriétés des réseaux biologiques. *neighbors_order* et *knbrs* sont cruciales pour analyser la structure locale des nœuds, tandis que *get_seeds* est essentielle pour initier la formation de modules distincts au sein du graphe.

3.3.2 Génération de la structure topologique du graphe

la structure topologique du reseaux de données est basé sur le modèle de Barabási-Albert étendu, permettant la création de graphe invariant d'échelle (*scale-free graph*) avec la propriété fondamentale de l'attachement préférentiel (Les nouveaux noeuds ont tendance à se connecter à des nœuds déjà bien connecté). Les paramètres clés de ce modèle sont :

1. Nœuds Initiaux (*nb_initial_nodes* = 3) : Les nœuds initiaux forment le noyau de départ du graphe. Ils sont essentiels pour commencer le processus de croissance du réseau selon la

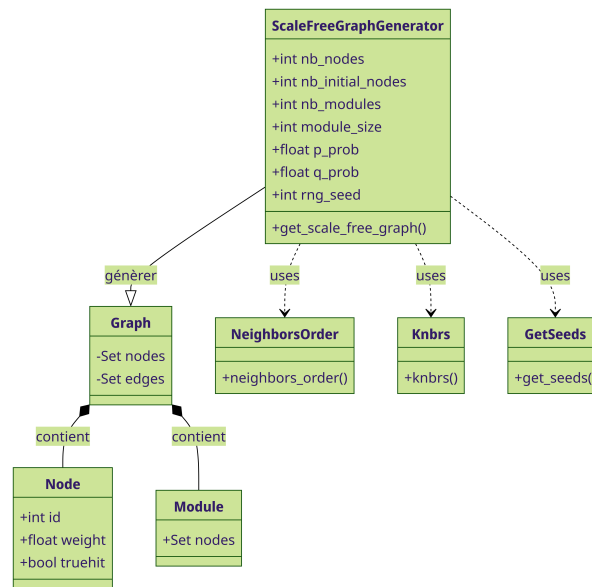


FIGURE 3.3 – Illustration en digramme classe de la description de l'algorithme qui genere les données artificielles

méthode de Barabási-Albert. Le nombre de nœuds initiaux influence la structure initiale du réseau. Un petit nombre de nœuds initiaux peut conduire à un réseau plus centralisé autour de ces nœuds, tandis qu'un plus grand nombre peut favoriser une structure plus distribuée. Le choix du nombre de nœuds initiaux doit refléter l'objectif de la simulation. Pour un réseau biologique, il est souvent souhaitable de commencer avec un petit nombre de nœuds initiaux pour simuler le développement naturel d'un réseau biologique à partir de quelques éléments clés.

2. Probabilités (p_prob) et (q_prob) : la probabilité ($p_prob=0.09$) : Contrôle l'ajout de nouvelles arêtes entre les nœuds existants. Un (p_prob) élevé favorise la création de nouvelles connexions. et la probabilité ($q_prob=0.7$) : Gère la réorganisation des arêtes existantes. Un (q_prob) élevé permet une plus grande dynamique dans la structure du réseau. pour assure un équilibre entre la croissance et la reorganistaion du réseau la somme de probabilité doit soumise a la contrainte suivante $(p_prob) + (q_prob) < 1$,

Après avoir construit la structure topologique du graphe basée sur le modèle de Barabási-Albert, nous veillerons à ne pas laisser de composantes connexes disjointes. Le principe est simple, il s'agira de créer un lien aléatoire entre les composantes connexes.

3.3.3 Génération des Modules dans le Graphe

La formation des modules dans le réseau est une étape cruciale, simulant la création de groupes de gènes ou de protéines fonctionnellement liés. Des nœuds "graines" sont sélectionnés en fonction de leur degré de connectivité et de leur distance relative, assurant une répartition équilibrée des modules dans le graphe. Autour de chaque graine, un module est formé en intégrant des nœuds voisins, choisis selon un processus aléatoire pondéré par la distance dans le graphe. La taille de chaque module est contrôlée par le paramètre 'module_size'. Par exemple, si une graine est sélectionnée, les nœuds à une distance de 1 ou 2 pas sont progressivement inclus dans le module, en fonction de leur probabilité de connexion.

3.3.4 Attribution des Poids (p_value) aux Nœuds

Enfin, des poids sont attribués à chaque nœud du graphe pour simuler des caractéristiques biologiques spécifiques. Les nœuds hors modules reçoivent des poids selon une distribution uniforme $[0,1]$, tandis que ceux au sein des modules suivent une distribution normale tronquée. Cette distribution est choisie pour refléter une concentration élevée de caractéristiques biologiquement significatives dans les modules, comme on pourrait s'y attendre dans des groupes de gènes ou de protéines actifs. Les poids des nœuds dans les modules sont donc générés selon la formule :

$$P(p_value) = \text{TruncNorm}(\mu, \sigma, a, b) \quad (3.5)$$

où $\mu = 0$, $\sigma = 0.05$, et les bornes a et b sont ajustées pour maintenir les poids entre 0 et 1.

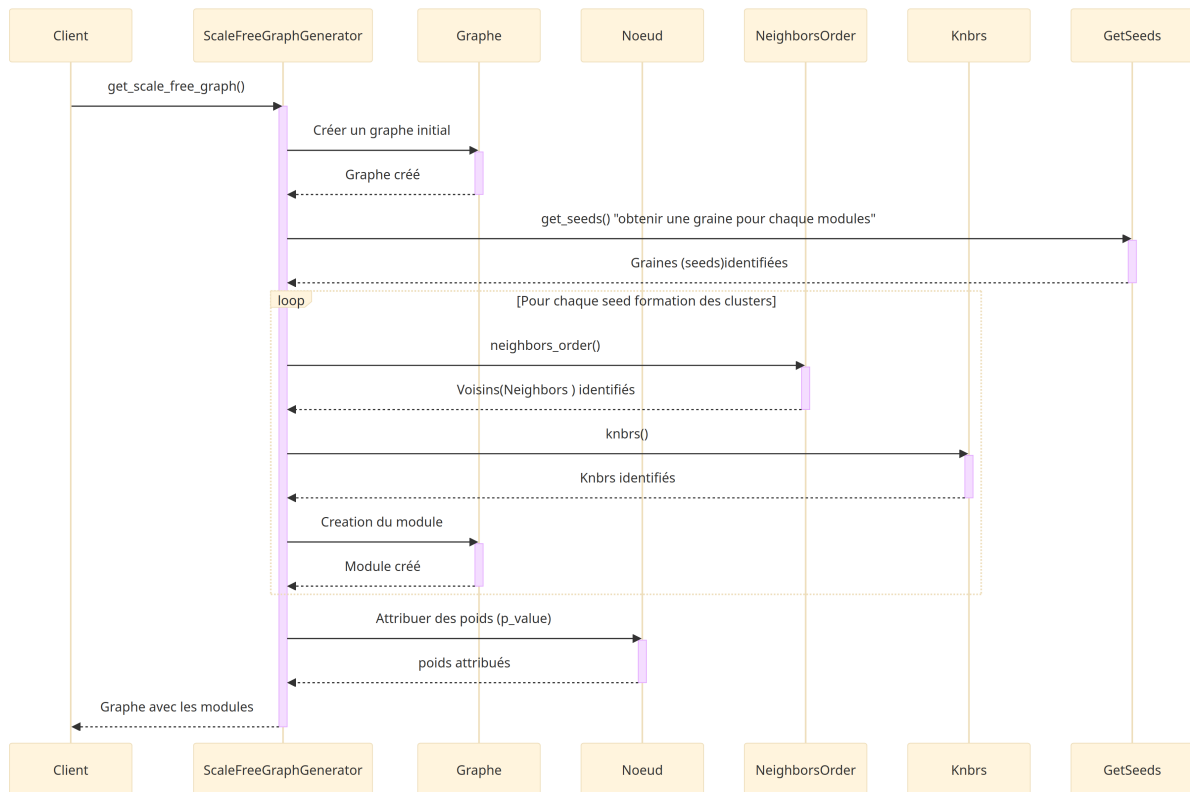


FIGURE 3.4 – Illustration la generation des données artificielles

La méthodologie proposée offre une nouvelle perspective pour l'analyse des interactions complexes au sein des réseaux biologiques en intégrant le concept de données multivues et en appliquant des critères d'évaluation rigoureux, nous visons à améliorer la précision et la robustesse de la détection des modules actifs, un aspect crucial pour la compréhension des mécanismes biologiques sous-jacents

VALIDATION DU MODÈLE

Dans ce chapitre, nous présentons en détail l'implémentation et les expérimentations réalisées afin de valider notre modèle Multivue . Ce modèle vise à détecter des modules actifs au sein de graphes artificiels. Cette démarche s'appuie sur une méthodologie multivues que nous avons spécifiquement élaborée pour notre projet de recherche.

4.1 Environnement d'Implémentation : Choix Technologiques

4.1.1 Langages de Programmation

Le langage de programmation choisi pour cette recherche est Python, reconnu pour sa polyvalence, sa compatibilité étendue avec diverses bibliothèques, et son efficacité dans le traitement de données complexes. Python se distingue par sa facilité de lecture, sa syntaxe claire et son vaste écosystème de bibliothèques scientifiques, ce qui en fait un choix idéal pour les projets de recherche en sciences des données et en intelligence artificielle.

4.1.2 Outils de Traitement de Données

Pour la gestion et la manipulation des données, nous avons intégré plusieurs outils spécialisés :

- Pandas : Utilisé pour sa capacité à manipuler et analyser efficacement de grands ensembles de données.
- NetworkX : Employé pour créer et analyser des graphes complexes, un élément central dans notre étude des réseaux biologiques.
- Numpy : Indispensable pour les calculs numériques et la manipulation de tableaux de données.
- Xlrd et Openpyxl : Pour la lecture et l'écriture de fichiers Excel, facilitant l'intégration de données hétérogènes.

4.1.3 Bibliothèques de Machine Learning et Statistiques

- Torch.nn et Torch.optim : Ces composants de PyTorch sont utilisés pour le développement et l'optimisation de réseaux de neurones, ainsi que pour la mise en œuvre de techniques d'embedding avancées

- Scikit-learn(Sklearn.metrics) : Offrant des outils statistiques et des mesures de performance avancées.
- Scipy : Incluant 'scipy.spatial.distance', 'scipy.stats', 'scipy.stats.norm', pour des mesures de distance précises, des calculs statistiques et des analyses de distributions normales.

4.1.4 Technologies Spécifiques pour l'Analyse de Graphes

Nous avons choisi la librairie **Node2Vec** qui est fondamentale pour l'embedding de graphes. Elle offre une approche innovante pour transformer les nœuds en vecteurs efficaces, ce qui facilite grandement l'analyse avancée des graphes. Son rôle est crucial dans la création d'un vocabulaire vectoriel à partir des nœuds, permettant une représentation vectorielle plus efficace et une meilleure interprétation des structures de graphe.

4.1.5 Gestion des Fichiers et Utilitaires

- Shutil : Pour la gestion efficace de fichiers et de répertoires.
- ZipFile : Essentiel pour la manipulation de fichiers compressés, utile pour gérer de grands ensembles de données.
- CSV : Pour la lecture et l'écriture standard de fichiers CSV.
- Gensim : Utilisé pour le traitement avancé du langage naturel et l'analyse de texte.

4.1.6 Autres Librairies et Fonctions

- Math : Pour les fonctions mathématiques de base.
- Statistics : Pour des analyses statistiques simples.
- Random : Crucial pour la génération de nombres aléatoires.
- Typing : Pour une définition et une gestion claires des types de données.
- Argparse : Facilite la gestion des arguments de ligne de commande.
- Time : Pour mesurer les durées d'exécution.
- Matplotlib.pyplot : Pour la visualisation des données.
- Numba : Pour l'optimisation des performances via la compilation JIT.
- Progressbar2 : Pour afficher des barres de progression dans les processus longs.
- Python-Levenshtein : Pour des calculs rapides de distance de Levenshtein.
- PyYAML : Pour le parsing et la production de fichiers YAML.
- Powerlaw : Pour analyser les distributions en loi de puissance.

Cet ensemble diversifié de technologies nous a permis de mettre en place notre framework que nous avons élaboré pour la recherche sur les réseaux biologiques, englobant des aspects essentiels comme la manipulation de données, l'analyse de réseaux, le machine learning et la visualisation

4.2 Approche de simulation

Dans notre démarche de recherche, nous adoptons une approche de simulation visant à générer des embeddings à partir de données multivues en utilisant deux cadres conceptuels

distincts : **MANE** (Multi-View Collaborative Network Embedding) et **MvME** (Multi-view Neighbourhood Embedding) que nous allons intégrer au sein du Framework **AMINE**. L'objectif est d'appliquer une métrique de similarité adéquate dans l'espace d'embedding pour identifier efficacement le groupe de "true hits" dans les réseaux biologiques. Cette méthode nous permettra de comparer les deux cadres en termes de leur capacité à fournir une représentation vectorielle précise et pertinente des données issues de multiples vues.

4.3 Analyse des Résultats MANE pour les Modèles de Construction

4.3.1 Contexte

Dans cette section, nous explorons l'utilisation du cadre conceptuel MANE (*Multi-View Collaborative Network Embedding*) pour l'intégration de différentes vues dans les réseaux biologiques. Notre objectif est d'évaluer l'efficacité de diverses métriques de similarité vectorielle dans l'embedding résultant de l'intégration des vues. La théorie développée par le cadre **MANE** est particulièrement adaptée à notre étude, car elle nécessite des vues sans données manquantes, c'est-à-dire que chaque nœud doit avoir une instanciation dans toutes les vues.

Dans notre implémentation avec **MANE**, les deux vues utilisées seront les suivantes :

- La première vue est celle d'un graphe non pondéré qui conserve la topologie du graphe de départ, comme mentionné dans la méthodologie.
- La deuxième vue est celle d'un graphe construit sur les valeurs des p-values suivant la **variante 1** de chaque modèle de construction, laquelle conserve les nœuds singletons.

Ces deux vues ont le même nombre de nœuds, chaque nœud ayant des instances dans les deux vues. Cette intégralité des vues nous permet d'appliquer efficacement MANE pour obtenir une représentation vectorielle cohérente et significative.

4.3.2 Objectif

L'objectif principal de cette analyse est d'identifier la configuration de construction la plus efficace avec le cadre Conceptuel **MANE** pour détecter les "**True hits**" dans des réseaux biologiques. Notre démarche vise à déterminer la représentation vectorielle la plus adaptée, où les nœuds correspondant aux "**True hits**" seront bien regroupés à l'aide d'une métrique de similarité pertinente, à définir avec précision.

Nous nous focaliserons particulièrement sur les avantages de l'approche multivues dans notre domaine de recherche, en évaluant comment cette intégration renforce la capacité du modèle à identifier de manière fiable et précise les éléments clés dans des réseaux biologiques. Cette évaluation se fera sur un jeu de données de 1000 graphes artificiels et se reposera sur une définition claire et rigoureuse de la mesure de similarité, élément crucial pour sélectionner la représentation vectorielle la plus pertinente, et par extension, le modèle le plus adapté à nos données.

4.3.3 Evaluation avec le Modèle de Construction 1

Analyse des résultats pour des graphes avec un True Hit de taille 10

Notre objectif était d'évaluer l'efficacité de ce modèle dans la détection d'un module actif de taille 10 au sein de graphes artificiels comprenant 1000 nœuds.

1. Tableau des Résultats :

Modèle	Variance	Q25	Q50 (Médiane)	Q75	Moyenne sur 1000 graphes
Amine	0.2751	0.4000	0.63397	0.7368	0.5972
MANE Pearson	0.1270	0.2727	0.35294	0.4324	0.3359
MANE Euclidien	0.1315	0.2152	0.30303	0.4166	0.3055
MANE Cosinus	0.1305	0.2553	0.3333	0.4210	0.3325

TABLE 4.1 – Analyse comparative des F1 scores des modèles avec un true hit de taille 10

2. Interprétation : Nous évaluons trois critères principaux à savoir performance générale, consistance et fiabilité des résultats.

— Performance Générale :

- Le modèle Amine se distingue par une performance supérieure, affichant une moyenne de 0.5972, ce qui illustre sa capacité élevée à détecter les "True hits".
- Les modèles MANE (Pearson, Euclidien, et Cosinus) affichent des performances moindres, avec des moyennes allant de 0.3055 à 0.3359.
- Le modèle MANE Pearson se distingue parmi les variantes de MANE, avec une moyenne de 0.3359 accompagnée d'une faible variance, reflétant une capacité significative à détecter les "true hits" par rapport aux autres variantes de MANE.

— Consistance et Variabilité : La variance est un indicateur clé de la consistance

- Les modèles MANE présentent une variabilité réduite (variance entre 0.1270 et 0.1315), suggérant une plus grande consistance malgré une performance globale inférieure.
- Bien que le modèle Amine affiche une performance supérieure avec une moyenne de 0.5972, sa variance élevée de 0.2751 révèle une variabilité dans ses résultats. En contraste, la faible variance de 0.1270 du modèle MANE Pearson, la plus basse parmi les variantes de MANE, souligne sa consistance et sa fiabilité, faisant de lui un choix potentiellement meilleur pour des résultats consistants..

— Fiabilité des Résultats : La médiane (Q50) et les quartiles (Q25 et Q75) offrent des perspectives supplémentaires sur la distribution des performances.

- En se focalisant d'abord sur le modèle MANE Pearson, nous observons une médiane de 0.35294, indiquant une meilleure distribution de performances sur les variantes de MANE.
- Le modèle Amine affiche une médiane plus élevée (0.63397) qui est environ le double de l'autre variante de MANE (cosinus, euclidienne).

Analyse des Résultats avec des Graphes d'un True Hit de Taille 20

Lors de l'évaluation de graphes avec un **true hit** de taille 10, nous avons observé que l'implémentation de **MANE** démontrait une excellente capacité de **Rappel** (*la proportion des cas positifs réels de true hits qui ont été correctement identifiés par le modèle*) dans le module détecté. Plus précisément, dans le module identifié, nous retrouvons environ 60 à 80% des nœuds du **true hit**, notamment avec **la première variante de la Construction 1** mais cela était accompagné d'une faible **précision** (*matérialisée par une grande quantité de faux positifs*). Cette observation nous a incités à explorer l'évaluation de l'impact de l'augmentation de la taille du **true hit** à 20 nœuds. Les résultats de cette analyse sont présentés dans le tableau suivant :

1. Tableau des Résultats :

Modèle	Variance	Q25	Q50 (Médiane)	Q75	Moyenne sur 1000 graphes
Amine	0.1861	0.4666	0.5806	0.6666	0.5486
MANE Pearson	0.1083	0.4084	0.4846	0.5614	0.48562
MANE Euclidien	0.0940	0.4106	0.4404	0.4908	0.45463
MANE Cosinus	0.1080	0.4154	0.4807	0.5696	0.48010

TABLE 4.2 – Analyse comparative des scores F1 des modèles avec un true hit de taille 20.

2. Interprétation des Résultats pour un True Hit de Taille 20 :

— Performance face à une Complexité Accrue :

- Avec un "true hit" de taille 20, le modèle Amine conserve une performance comparable (moyenne de 0.5486), démontrant ainsi une bonne adaptabilité à des réseaux plus complexes.
- Les modèles MANE montrent une amélioration ou une stabilisation de leurs performances .par exemple la variante Pearson passe de 0.3309 de moyenne à 0.48562 en moyenne, ce qui peut indiquer une meilleure capacité à gérer des situations plus complexes avec des **True hit** plus important .

— Considérations Statistiques :

- La variance pour tous les modèles diminue avec un "True hit" de taille 20, suggérant une plus grande cohérence dans les résultats.
- Les quartiles montrent une amélioration notable des performances des modèles MANE par rapport aux résultats avec des true hit de taille 10 , bien qu'une part significative de leurs résultats demeure légèrement en retrait par rapport au modèle **AMINE**

Ces analyses révèlent que le modèle Amine tend à surclasser les modèles MANE basés sur la Construction 1 dans les scénarios testés, malgré une variabilité plus marquée. L'augmentation de la taille des "true hits" semble bénéfique pour les modèles MANE dans le cadre de la Construction 1, laissant supposer leur meilleure adaptabilité à des réseaux de grande envergure ou plus complexes. Cependant, il est crucial de souligner que la performance supérieure du modèle Amine s'accompagne d'une variabilité élevée, ce qui pourrait influencer sa fiabilité dans différents contextes.

4.3.4 Evaluation avec le Modèle de Construction 2

Analyse des résultats avec des graphes d'un True Hit de taille 10

Les résultats suivants ont été obtenus en utilisant le modèle MANE avec la première sous-variante 1 du Modèle de Construction 2, c'est-à-dire que nous avons conservé des nœuds singletons.

— **Tableau des Résultats :**

Modèle	Variance	Q25	Q50 Médiane	Q75	Moyenne sur 1000 graphes
Amine	0.27516	0.40000	0.63397	0.73684	0.56932
MANE Pearson	0.04592	0.18182	0.21538	0.24615	0.22391
MANE Euclidien	0.12091	0.1912	0.2191	0.21391	0.21381
MANE Cosinus	0.13914	0.1791	0.2170	0.21359	0.21940

TABLE 4.3 – Analyse comparative des scores F1 des modèles avec un true hit de 10 pour la construction 2

— **interprétation** Les modèles MANE, bien qu'affichant une plus grande consistance (faible variance), ne parviennent pas à atteindre les niveaux de performance du modèle Amine. Ils sont moins efficaces dans la détection de "true hits" de taille 10. La différence de performance entre Amine et les modèles MANE souligne l'importance d'optimiser davantage les approches de construction des vues pour améliorer leur efficacité.

Analyse des résultats avec des graphes d'un True Hit de taille 20

1. Tableau des Résultats :

Modèle	Variance	Q25	Q50 Médiane	Q75	Moyenne sur 367 graphes
Amine	0.1861	0.4666	0.5806	0.6666	0.52640
Mane Pearson	0.0531	0.30380	0.33766	0.37209	0.34055
Mane Euclidien	0.0559	0.30000	0.33735	0.37333	0.35133
Mane Cosinus	0.0530	0.29971	0.33333	0.37397	0.34081

TABLE 4.4 – Analyse comparative des scores F1 des modèles avec un true hit de 20 pour la construction 2.

2. Interprétation des Résultats

- Tous les modèles MANE (Pearson, Euclidien, et Cosinus) affichent des performances inférieures comparées au modèle Amine, mais elles se rapprochent de celles obtenues avec la construction 1 sur les graphes de true hit de taille 10. Ce qui pourrait confirmer que la construction 2 est moins efficace que celle de la construction 1.
- Bien que les modèles MANE sont moins performants nous pouvons remarquer que la variance des modèles MANE est relativement faible (autour de 0.053), indiquant une certaine consistance dans leurs performances sur les données multivues.
- Pour défaut de mémoire, le processus computationnel de cette session s'est arrêté au niveau de l'évaluation du 368ème graphe de données artificielle. Vu les résultats mitigés, nous avons jugé qu'il n'était pas nécessaire de relancer la session de travail le

temps pour nous d'évaluer d'autres méthodes en exploitant au mieux les ressources pour atteindre nos objectifs.

Bien que nous ayons une faible variance avec cette construction 2, les performances du modèle avec cette construction ne parviennent pas à atteindre le niveau de précision et de fiabilité du modèle Amine. Cette observation souligne le besoin de continuer à explorer et à affiner les modèles de construction pour améliorer leur capacité à détecter des **true hits** dans des réseaux biologiques.

4.3.5 Evaluation avec le Modèle de Construction 3

Pour évaluer l'efficacité de la détection des **true hits** au sein de réseaux biologiques, nous avons intégré le Modèle de Construction 3 dans notre étude. Ce modèle se distingue par l'ajout d'arêtes supplémentaires entre les nœuds dont les p-values sont inférieures à 0,05, une modification qui se fait tout en conservant la structure topologique d'origine. Un point notable de cette construction est l'absence de nœuds singletons, ce qui pourrait influencer significativement les résultats obtenus.

Problématiques Rencontrées

- Durée de Traitement Élevée : Durant les phases initiales de notre expérimentation avec le Modèle de Construction 3, nous avons constaté que le processus de traitement était extrêmement long. Cette durée prolongée a posé des défis significatifs en termes de temps et de ressources, rendant le processus impraticable pour une analyse complète dans un délai raisonnable.
- Faible Performance dans les Résultats Préliminaires : Les résultats préliminaires obtenus avant l'arrêt du processus indiquaient des performances très faibles.

Considérations et Décisions

- Arrêt de l'Expérimentation : En raison de la durée excessive de traitement et des faibles performances préliminaires, il a été décidé d'arrêter l'expérimentation avec le Modèle de Construction 3. Cette décision nous a permis de concentrer nos efforts et nos ressources sur des approches plus prometteuses et pratiques.
- Réévaluation du Modèle : Il est impératif de procéder à une réévaluation approfondie de ce modèle afin de démêler les causes profondes des difficultés rencontrées. Des ajustements stratégiques pourraient être envisagés dans le but de diminuer le temps de traitement et d'augmenter l'efficacité globale. Parmi les modifications potentielles figurent la révision des critères d'ajout d'arêtes, l'optimisation des sous-algorithmes en place, ou encore l'ajustement des valeurs des hyperparamètres de node2vec. Ces modifications ciblées pourraient apporter des améliorations significatives aux performances du modèle.

Le Modèle de Construction 3, bien qu'intéressant dans son approche théorique, s'est avéré peu pratique et inefficace dans notre contexte de recherche. Les résultats préliminaires et la durée de traitement ont souligné les limitations de ce modèle pour l'analyse de réseaux biologiques complexes. L'arrêt de l'expérimentation avec ce modèle spécifique reflète notre engagement à poursuivre les méthodes les plus efficaces et réalisables pour notre recherche.

4.3.6 Evaluation avec le Modèle de Construction 4

Analyse des resultats avec des graphes d'un True Hit de taille 10

Dans cette section, nous nous penchons sur l'évaluation des performances du Modèle de Construction 4, appliqué à des 'true hits' de taille 10 nœuds. L'objectif principal de cette analyse est d'examiner en détail la capacité des différents modèles à identifier de manière précise ces 'true hits' dans une configuration de réseau bien définie.

— Tableau des Résultats

Modèle	Variance	Q25	Q50 Médiane	Q75	Moyenne sur 289 graphes
Amine	0.27516	0.40000	0.63397	0.73684	0.60763
Mane Pearson	0.04592	0.18182	0.21538	0.24615	0.21391
Mane Euclidien	<i>Arret</i>	<i>de</i>	<i>evaluation</i>	<i>dans l'espace</i>	<i>intégré</i>
Mane Cosinus	0.09355	0.00000	0.00000	0.11111	0.04920

TABLE 4.5 – Analyse comparative des modèles avec un true hit de 10 gènes avec le modele de contruction 4.

— Interprétation des Résultats

- Vu les performances mitigées, nous avons décidé de ne pas poursuivre la session de traitement. En effet, presque un quart des données évaluées indiquent que la médiane de Mane Cosinus est toujours à zéro, et celle de Pearson n'est qu'à 0.24, tandis que pour Amine, elle est au-delà de 0.6.
- Les résultats montrent une performance significativement plus basse de Mane par rapport au modèle Amine, ce qui suggère des limitations dans cette configuration lié à la construction 4.

Les résultats actuels suggèrent que le modèle de construction Graphe_4 offre des performances variables selon le modèle MANE appliqué. En particulier, les modèles Pearson et Cosinus de MANE montrent des limitations notables dans la détection des "true hits".

Analyse des resultats avec des graphes d'un True Hit de taille 20

Avec un 'true hit' de taille 20, les résultats se sont progressivement dégradés. Face à l'interruption due au manque de ressources de la machine, nous avons jugé qu'il n'était pas pertinent de relancer le traitement

— Tableau des Résultats :

Modèle	Variance	Q25	Q50 Médiane	Q75	Moyenne sur 362 graphes
Amine	0.1861	0.4666	0.5806	0.6666	0.53487
Mane pearson	0.07950	0.00000	0.00000	0.09091	0.05604

TABLE 4.6 – Analyse comparative des scores F1 modèles avec un true hit de 20 gènes pour la construction 4.

- Interprétation des Résultats Lors de l'évaluation, nous avons constaté une dégradation des performances du modèle MANE, ce qui nous a conduit à décider de ne pas poursuivre le traitement.

4.4 Analyse des Résultats avec MvME pour les Modèles de Construction

4.4.1 Contexte et Méthodologie

L'utilisation de MvME (Multi-view Neighbourhood Embedding) pour les différents modèles de construction a constitué une partie importante de notre recherche. Cette technique, adaptée aux situations où certaines vues peuvent présenter des données manquantes, semblait prometteuse pour notre analyse multivue. Dans le cas de notre étude, nous avons spécifiquement appliqué MvME à la deuxième vue construite selon la seconde variante de notre méthodologie, qui suggérerait la suppression des nœuds singletons pour donner plus d'information aux nœuds connectés.

4.4.2 Résultats et Problématiques

- **Performances Faibles** : Les résultats obtenus avec MvME ont été relativement faibles. Cela a soulevé des questions quant à l'efficacité de cette approche dans le contexte spécifique de nos réseaux biologiques et dans la détection des "true hits". La faible performance pourrait être attribuée à la suppression des nœuds singletons, ce qui pourrait avoir réduit la richesse des informations disponibles dans l'embedding et impacté négativement l'analyse.
- **Processus de Calcul Long** : Le temps nécessaire pour mener à bien le processus de calcul avec MvME a été particulièrement long. Cette durée prolongée rend le processus inutilisable en situation réelle.

4.4.3 Décision et Implications

En raison des performances faibles et du temps de traitement élevé, nous avons pris la décision de suspendre les évaluations supplémentaires avec MvME. Cette décision a été guidée par la nécessité d'optimiser l'efficacité de notre recherche et de concentrer nos ressources sur des méthodes plus prometteuses.

L'expérience avec MvME, bien que n'ayant pas produit les résultats escomptés, offre des enseignements précieux. Elle souligne l'importance de l'intégrité des données et de l'efficacité du processus dans l'analyse des réseaux biologiques. La suspension de cette méthode nous incite à explorer d'autres voies qui pourraient être plus fructueuses pour notre objectif de détection précise des "true hits" dans des réseaux biologiques.

Cette étape de notre recherche illustre l'importance d'être flexible et réactif face aux défis méthodologiques. En fin de compte, chaque étape, même celles menant à des résultats non concluants, est cruciale pour affiner notre compréhension et guider nos efforts futurs dans l'étude complexe des réseaux biologiques.

4.5 Choix du Modèle d'intégration d'embedding

Après une analyse approfondie des résultats obtenus de nos diverses expérimentations, nous avons décidé d'adopter le modèle MANE en combinaison avec la Construction 1. Notre choix se

fonde sur l'objectif de tester une approche multivue, ce qui nous a conduit à privilégier MANE avec la construction 1 et la métrique de similarité de Pearson. Ce modèle, que nous nommerons Amine_Multiview, bien qu'affichant une performance légèrement inférieure à celle d'Amine, se distingue néanmoins par sa capacité à surclasser d'autres modèles standards (Expr value, Bionet, COSINE, DIAMOND, DOMINO) lors des tests sur données artificielles avec un **True hit** de taille 10 voir figure 4.1. Amine_Multiview obtient une Moyenne de 0.3309, une médiane de 0.35294 et une variance de 0.1270 (cf Table 4.1) sur un ensemble de 1000 valeurs de F1_Scores.

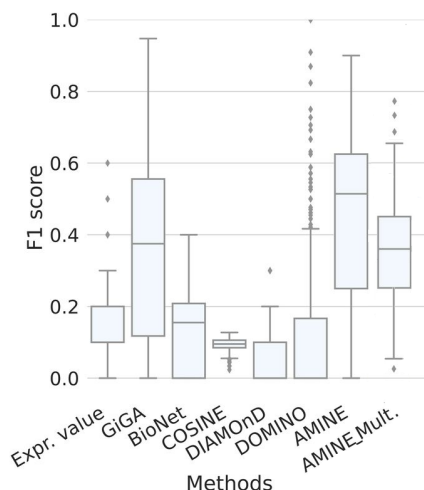


FIGURE 4.1 – tableau de F1 score des modeles bioinformatique

Nous prévoyons donc de poursuivre en validant et en confirmant ces résultats sur des ensembles de données réels. En approfondissant notre exploration avec le modèle AMINE_Multiview sur des données réelles, nous aspirons non seulement à valider nos résultats préliminaires mais également à acquérir une compréhension plus profonde des mécanismes biologiques sous-jacents. Cette démarche vise à assurer que notre modèle, Amine_Multiview, ne se contente pas de performances théoriques mais s'avère également efficace et pertinent dans des applications réelles. Avant de procéder à l'application de notre modèle AMINE_Multiview sur des données réelles, Nous devons initier une comparaison rigoureuse avec les modèles de clustering classiques.

4.6 Comparaison avec les méthodes de clustering classiques

Après avoir réalisé une comparaison entre Amine_Multiview et divers modèles spécialisés en bioinformatique, notre objectif suivant consiste à évaluer si les méthodes de clustering traditionnelles peuvent offrir de meilleures performances sur embedding par rapport à Amine_Multiview. Nous avons mené notre analyse sur une série de 40 embeddings de graphes, chaque graphe contenant 1000 nœuds et un groupe cible "**True hit**" de 10 nœuds spécifiques à identifier. Les algorithmes de clustering classiques testés sont les suivants :

- **DBSCAN** : Pour cet algorithme, les valeurs des hyperparamètres qui ont produit les meilleurs résultats sont : $\text{eps} = 0.26$, $\text{min_samples} = 3$ et $\text{metric} = \text{'cosine'}$.
- **OPTICS** : Une extension de DBSCAN, conçue pour gérer les variations de densité. Les paramètres sélectionnés sont : $\text{min_samples} = 3$ et $\text{metric} = \text{'cosine'}$.

- **Spectral Clustering** : Utilisé pour regrouper des données qui ne sont pas linéairement séparables, ce clustering est particulièrement efficace pour identifier des structures complexes dans les données. Les paramètres utilisés sont : `n_clusters = 10` et `affinity = 'nearest_neighbors'`.

Modèle	Variance	Q25	Q50 Médiane	Q75	Moyenne sur 40 graphes
AMINE_Multiview	0.0271	0.2340	0.4273	0.4419	0.3947
DBSCAN	0.0883	0.2974	0.3053	0.3739	0.3141
OPTICS	0.0477	0.1146	0.1946	0.2712	0.2727
Spectral Clustering	0.0203	0.2768	0.4316	0.4539	0.4061

TABLE 4.7 – Analyse comparative des scores F1 de Amine_Multiview et des modèles classiques de clustering, avec un *true hit* de taille 10.

L'analyse des différents points que nous pouvons déduire de notre tableau comparatif sur des performances de différents modèles de clustering classiques avec Amine_Multiview sont les suivantes

- Le modèle **AMINE_Multiview** et le **Spectral Clustering** présentent la variance la plus basse, indiquant une consistance dans les performances sur les différents graphes testés.
- **AMINE_Multiview** montre une médiane (Q50) relativement élevée (0.4273), ce qui indique que la moitié de ses scores F1 sont au-dessus de cette valeur. Cela suggère que AMINE_Multiview produit une performance globalement supérieure pour une partie des graphes. Cependant, OPTICS, malgré sa bonne variance, montre une médiane plus basse (0.1946), ce qui pourrait refléter une performance plus faible sur certains graphes.
- **Spectral Clustering** présente la moyenne la plus élevée (0.4061), ce qui suggère que, globalement, ce modèle a le mieux performé en moyenne sur l'ensemble des graphes.

Bien que le modèle Spectral Clustering semble offrir les meilleures performances en moyenne associées à une faible variance, nous pouvons penser qu'elle ne sera pas efficace sur des données réelles de grande taille car pour une bonne performance d'un modèle basé sur Spectral Clustering, il faut déjà une idée du nombre de clusters à identifier.

APPLICATION DU MODÈLE À DES DONNÉES RÉELLES

5.1 Test du modèle sur les données réelles

5.1.1 Description du Jeu de Données :

Dans notre étude visant à tester le modèle *Amine_multivue*, nous avons utilisé un ensemble de données d'expression génique de PDAC téléchargée depuis Gene Expression Omnibus. Ces données comprennent des mesures d'expression génique de populations de cellules PDAC, distinguant les cellules exprimant HMGA2 (HMGA2+) de celles ne l'exprimant pas (HMGA2-), dérivées de tumeurs PDAC primaires dans un modèle de souris génétiquement modifiée. L'analyse des variations génétiques et d'expression est axée sur les processus liés aux métastases du PDAC, notamment :

- l'hypoxie,
- la transition épithélio-mésenchymateuse (EMT) ,
- et d'autres caractéristiques métastatiques (la capacité des cellules cancéreuses à se propager à partir de la tumeur primaire vers d'autres parties du corps)

HMGA2 (High Mobility Group AT-Hook 2) est un gène qui code pour une protéine jouant un rôle important dans la régulation de l'expression des gènes. Cette protéine est impliquée dans plusieurs processus cellulaires tels que la croissance, la prolifération et la différenciation. Dans le contexte du cancer, HMGA2 est souvent associé à un potentiel métastatique élevé et à un mauvais pronostic. Une expression accrue de HMGA2 peut contribuer à la progression et aux métastases de divers types de cancers, notamment le PDAC (adénocarcinome canalaire pancréatique). Ces données, analysées dans l'étude de Claude Pasquier et al, permettent de valider l'efficacité d'*Amine_multivue* dans l'identification de modules génétiques clés associés à la propagation métastatique dans le PDAC.

5.1.2 Préparation des Données

Dans le processus de préparation des données pour l'étude avec le modèle *AMINE_Multiview*, l'ensemble de données extrait du Gene Expression Omnibus a été soumis aux procédures suivantes :

1. *Nettoyage des données* : Les premières étapes ont consisté à éliminer les erreurs de formatage, à retirer les gènes présentant une faible expression, et à effectuer une vérification

- de la qualité des données pour éliminer les incohérences potentielles
2. **Normalisation** : Les données d'expression génique ont été normalisées afin de réduire les variations techniques qui ne reflètent pas les différences biologiques entre les échantillons, ce qui est essentiel pour les comparaisons ultérieures.
 3. **Transformation des données** : Les valeurs d'expression génique ont subi une transformation logarithmique pour atténuer les disparités dues à des niveaux d'expression élevés, rendant les données plus adaptées pour l'analyse
 4. **Intégration avec les réseaux d'interactions protéiques** : Les données d'expression génique ont été combinées avec des informations issues de bases de données d'interactions protéiques (comme STRING, BioGRID, et IntAct), facilitant l'étude des interactions génétiques au sein des réseaux cellulaires.
 5. **Application de filtres** : Des critères de sélection ont été appliqués aux interactions protéiques, utilisant par exemple un score de confiance minimal pour les données issues de STRING, afin de ne retenir que les interactions les plus pertinentes.
 6. **Paramétrage pour l'analyse** : Des paramètres spécifiques, incluant la sélection de l'espèce modèle (la souris "*Mus musculus*") et le réseau d'interactions protéiques de référence (par exemple, STRING), ont été définis pour l'analyse.
 7. **Analyse différentielle** : Utilisation d'algorithmes tels que DESeq2 pour identifier les gènes différentiellement exprimés entre les populations de cellules HMGA2+ et HMGA2-.

Ces étapes ont préparé les données pour une analyse approfondie par Amine_multivue, dans le but d'identifier les modules génétiques clés liés à la propagation métastatique dans le PDAC.

5.1.3 Résultats du Test :

Après avoir appliqué le modèle AMINE_Multiview à des ensembles de données réelles, un ensemble de données d'expression génique provenant de tumeurs PDAC primaires de souris génétiquement modifiées. Cet ensemble incluait 21 317 nœuds et 7 248 179 arêtes. L'analyse a permis d'identifier 90 modules distincts, parmi lesquels 6 modules ont été considérés comme particulièrement significatifs après une analyse d'enrichissement fonctionnel effectuée via la plateforme GSEA. En illustration, la figure 5.1 montre l'enrichissement du module 1. L'examen de la colonne FDR (taux de fausse découverte) a révélé des valeurs extrêmement significatives ; en effet, une valeur inférieure à $10e-5$ est considérée comme très significative pour les voies biologiques identifiées.

5.2 Interprétation des résultats

5.2.1 Interprétation de l'enrichissement du module 1

Dans le contexte du cancer du pancréas (PDAC), les modules génétiques mentionnés sont importants pour comprendre les mécanismes sous-jacents de la maladie :

- **Hypoxie** : Les tumeurs PDAC peuvent développer des zones hypoxiques, hypoxie est un facteur connu pour influencer la progression tumorale dans de nombreux types de cancer. Elle peut favoriser l'adaptation des cellules cancéreuses à un environnement pauvre en oxygène, contribuant ainsi à la résistance aux thérapies.

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value ?	FDRq-value ?
HALLMARK_HYPOXIA [199]	Mouse genes annotated to HALLMARK_HYPOXIA based on orthology mappings provided by the Alliance Genome Consortium	23		2.77 e-27	1.39 e-25
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION [194]	Mouse genes annotated to HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION based on orthology mappings provided by the Alliance Genome Consortium	20		7.97 e-23	1.99 e-21
HALLMARK_GLYCOLYSIS [200]	Mouse genes annotated to HALLMARK_GLYCOLYSIS based on orthology mappings provided by the Alliance Genome Consortium	15		2.62 e-15	4.37 e-14
HALLMARK_ANGIOGENESIS [36]	Mouse genes annotated to HALLMARK_ANGIOGENESIS based on orthology mappings provided by the Alliance Genome Consortium	5		2.77 e-7	3.46 e-6
HALLMARK_ADIPOGENESIS [200]	Mouse genes annotated to HALLMARK_ADIPOGENESIS based on orthology mappings provided by the Alliance Genome Consortium	8		1.23 e-6	1.23 e-5
HALLMARK_UV_RESPONSE_DN [144]	Mouse genes annotated to HALLMARK_UV_RESPONSE_DN based on orthology mappings provided by the Alliance Genome Consortium	7		1.59 e-6	1.32 e-5
HALLMARK_APOPTOSIS [161]	Mouse genes annotated to HALLMARK_APOPTOSIS based on orthology mappings provided by the Alliance Genome Consortium	7		3.34 e-6	2.38 e-5
HALLMARK_MTORC1_SIGNALING [199]	Mouse genes annotated to HALLMARK_MTORC1_SIGNALING based on orthology mappings provided by the Alliance Genome Consortium	7		1.34 e-5	8.34 e-5
HALLMARK_FATTY_ACID_METABOLISM [155]	Mouse genes annotated to HALLMARK_FATTY_ACID_METABOLISM based on orthology mappings provided by the Alliance Genome Consortium	6		3.29 e-5	1.83 e-4
HALLMARK_OXIDATIVE_PHOSPHORYLATION [195]	Mouse genes annotated to HALLMARK_OXIDATIVE_PHOSPHORYLATION based on orthology mappings provided by the Alliance Genome Consortium	6		1.17 e-4	5.86 e-4

FIGURE 5.1 – Enrichement du module 1 via la plateforme GSEA

- **Transition Épithélio-Mésenchymateuse (EMT)** : L'EMT est un processus par lequel les cellules épithéliales acquièrent des caractéristiques mésenchymateuses, ce qui leur permet de migrer et d'envahir d'autres tissus. Ce processus est crucial pour la métastase dans de nombreux cancers, y compris le PDAC.
- **Glycolyse** : Les cellules PDAC peuvent dépendre de la glycolyse pour l'énergie, un phénomène connu sous le nom d'effet Warburg. L'effet Warburg décrit la préférence des cellules cancéreuses pour la glycolyse pour produire de l'énergie même en présence d'oxygène, ce qui est un trait caractéristique de nombreux cancers, y compris le PDAC.
- **Angiogenèse** : Essentielle pour la croissance tumorale en fournissant des nutriments et de l'oxygène
- **Adipogenèse** : Peut être liée à la progression du PDAC en modifiant le microenvironnement tumoral
- **Réponse aux UV** : Moins directement liée au PDAC, mais peut influencer les mutations génétiques.
- **Apoptose** : Les modifications dans les voies apoptotiques peuvent effectivement favoriser la survie des cellules cancéreuses, contribuant à la progression du cancer et à la résistance

aux traitements

- **Signalisation de mTORC1** : Importante dans la croissance et la survie des cellules PDAC.
- **Métabolisme des Acides Gras** : Peut jouer un rôle dans la bioénergétique des cellules cancéreuses.
- **Phosphorylation Oxydative** : Influence la production d'énergie cellulaire, affectant la croissance tumorale

Ces processus sont fondamentaux pour comprendre le comportement des cellules souches de PDAC et pour le développement de thérapies ciblées. En effet, Chaque processus apporte un éclairage sur les mécanismes sous-jacents du PDAC et offre des pistes potentielles pour le développement de nouvelles thérapies ciblées.

5.2.2 Enrichement des autres modules

Parmi les 90 modules détectés, cinq ont également révélé un enrichissement positif, contribuant ainsi à l'identification de potentielles cibles thérapeutiques

Module 2 : HALLMARK_MYOGENESIS

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value ?	FDRq-value ?
HALLMARK_MYOGENESIS [199]	Mouse genes annotated to HALLMARK_MYOGENESIS based on orthology mappings provided by the Alliance Genome Consortium	2	<div style="width: 2.2%; height: 10px; background-color: green;"></div>	6.46 e ⁻⁵	3.23 e ⁻³

FIGURE 5.2 – Enrichement du module 2 via la plateforme GSEA

La signature "HALLMARK_MYOGENESIS" représente un ensemble de gènes impliqués dans le processus de myogenèse, c'est-à-dire le développement et la différenciation des cellules musculaires. La régénération musculaire est un processus contrôlé et bénéfique, visant à réparer le tissu endommagé.

Module 4 : HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION

Après l'enrichissement, nous avons observé que le module 4 est associé à la transition épithélio-mésenchymateuse (EMT), similairement au module 1. Cependant, une différence notable réside dans le nombre de gènes et la p-value associées à chaque module. Cette variation peut être interprétée comme reflétant l'intensité et la consistance de l'expression génique liée à l'EMT au sein de chaque module. En générale, un module présentant une meilleure p-valeur indique une corrélation plus forte et uniforme avec la voie associée. Cela suggère que les gènes inclus dans le module 4 pourraient être moins directement engagés ou moins systématiquement régulés dans le cadre du processus d'EMT par rapport à ceux analysés dans les échantillons. En d'autres termes, bien que les deux modules soient liés à l'EMT, le module 4 montre une association moins significative et moins cohérente avec l'EMT par rapport au module 1. Cela pourrait indiquer que les gènes du module 1 jouent un rôle plus central ou sont plus étroitement

régulés dans le processus d'EMT observé, comparativement au module 4, où l'implication peut être moins directe ou la régulation moins uniforme.


Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value ?	FDRq-value ?
HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION [194]	Mouse genes annotated to HALLMARK_EPITHELIAL_MESENCHYMAL_TRANSITION based on orthology mappings provided by the Alliance Genome Consortium	3		9.22 e ⁻⁸	4.61 e ⁻⁶

FIGURE 5.3 – Enrichement du module 4 via la plateforme GSEA

Module 20 : HALLMARK_APOPTOSIS et HALLMARK_P53_PATHWAY

Nous avons identifié la signature HALLMARK_APOPTOSIS dans le module 1. Cependant, nous allons plutôt explorer la seconde voie, HALLMARK_P53_PATHWAY, dont les fonctions peuvent contribuer à des cibles thérapeutiques contre le cancer. En effet, certaines de ces fonctions sont :

- **Arrêt du Cycle Cellulaire :** HALLMARK_P53_PATHWAY peut induire un arrêt temporaire du cycle cellulaire pour permettre la réparation de l'ADN endommagé, empêchant ainsi la propagation des mutations.
- **Apoptose :** Si les dommages à l'ADN sont trop importants pour être réparés, p53 peut activer des voies menant à l'apoptose, éliminant ainsi les cellules potentiellement cancéreuses.
- **Sénescence Cellulaire :** HALLMARK_P53_PATHWAY peut également induire un état de sénescence, dans lequel les cellules cessent de se diviser mais restent métaboliquement actives, comme un mécanisme de protection contre la transformation maligne.
- **Réparation de l'ADN :** Par l'activation de gènes impliqués dans la réparation de l'ADN, p53 aide à maintenir l'intégrité génomique.

Les fonctions de la voie HALLMARK_P53_PATHWAY dans l'arrêt du cycle cellulaire, l'apoptose, la sénescence cellulaire, et la réparation de l'ADN reflètent son rôle central dans la protection contre le cancer. Souvent appelé "gardien du génome", p53 surveille l'intégrité de l'ADN et orchestre des réponses cellulaires pour prévenir la propagation de cellules avec des dommages génétiques. Cela pourrait déterminer une cible thérapeutique d'importance dans la recherche sur le cancer et le développement de thérapies.



Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value ?	FDRq-value ?
HALLMARK_APOPTOSIS [161]	Mouse genes annotated to HALLMARK_APOPTOSIS based on orthology mappings provided by the Alliance Genome Consortium	2		8.42 e ⁻⁵	3.25 e ⁻³
HALLMARK_P53_PATHWAY [200]	Mouse genes annotated to HALLMARK_P53_PATHWAY based on orthology mappings provided by the Alliance Genome Consortium	2		1.3 e ⁻⁴	3.25 e ⁻³

FIGURE 5.4 – Enrichement du module 20 via la plateforme GSEA

Module 85 : HALLMARK_OXIDATIVE_PHOSPHORYLATION

Dans ce module, la p-value associée à HALLMARK_OXIDATIVE_PHOSPHORYLATION est plus significative que celle obtenue dans le module 1. Ceci suggère que les gènes de ce module pourraient jouer un rôle plus important dans l'activation de cette fonction. La signature "HALLMARK_OXIDATIVE_PHOSPHORYLATION" fait référence à un ensemble de gènes impliqués dans le processus de phosphorylation oxydative, qui est une voie métabolique essentielle qui se déroule dans les mitochondries et joue un rôle central dans la production d'énergie cellulaire sous forme d'ATP (adénosine triphosphate). Voici les principales fonctions de la phosphorylation oxydative :

- **Production d'ATP** : la production d'ATP, la principale source d'énergie pour de nombreux processus cellulaires, y compris la contraction musculaire, la synthèse de protéines, et le transport actif à travers les membranes cellulaires.
- **Régulation du Métabolisme** : La phosphorylation oxydative est étroitement régulée et s'adapte aux besoins énergétiques de la cellule. Elle est influencée par la disponibilité des substrats (comme le glucose et les acides gras) et par les signaux hormonaux et métaboliques.
- **Rôle dans l'Apoptose** : Les mitochondries et la phosphorylation oxydative jouent également un rôle dans l'induction de l'apoptose, ou mort cellulaire programmée, en réponse à certains signaux de stress ou de dommages cellulaires.
- **Utilisation de l'Oxygène** : Ce processus utilise l'oxygène comme accepteur final d'électrons dans la chaîne de transport d'électrons mitochondriale. L'oxygène est réduit pour former de l'eau, un processus qui contribue à la demande en oxygène des cellules et des tissus.

La signature "HALLMARK_OXIDATIVE_PHOSPHORYLATION" reflète donc l'importance de la phosphorylation oxydative dans la production d'énergie cellulaire, la régulation du métabolisme, et la participation à des processus cellulaires critiques comme l'apoptose. Des dysfonctionnements dans cette voie peuvent être impliqués dans le développement de maladies, y compris les troubles métaboliques, les maladies neurodégénératives, et le cancer.

Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value ?	FDRq-value ?
HALLMARK_OXIDATIVE_PHOSPHORYLATION [195]	Mouse genes annotated to HALLMARK_OXIDATIVE_PHOSPHORYLATION based on orthology mappings provided by the Alliance Genome Consortium	3	<div style="width: 100%; height: 10px; background-color: green;"></div>	1.85 e-6	9.27 e-5

FIGURE 5.5 – Enrichement du module 85 via la plateforme GSEA

Module 90 : HALLMARK_G2M_CHECKPOINT

la signature HALLMARK_G2M_CHECKPOINT met en lumière l'importance du point de contrôle G2M_Checkpoint dans le maintien de l'intégrité génomique et dans la prévention de la prolifération de cellules anormales dont l'une de ces fonctions essentielles sont les suivantes

- **Surveillance de l'ADN** : il surveille l'intégrité de l'ADN et s'assure que la réplication de

l'ADN est complète avant que la cellule n'entre en mitose. Cela empêche la ségrégation de chromosomes incomplets ou endommagés, qui pourrait conduire à l'instabilité génomique.

- **Réparation de l'ADN** : Si des dommages à l'ADN sont détectés, il active des voies de réparation de l'ADN et retarde temporairement la progression du cycle cellulaire, donnant à la cellule le temps nécessaire pour réparer

La compréhension des mécanismes régulant ce point de contrôle est essentielle pour le développement de stratégies thérapeutiques ciblées contre le cancer et d'autres maladies associées à l'instabilité génomique.


Gene Set Name [# Genes (K)]	Description	# Genes in Overlap (k)	k/K	p-value ?	FDRq-value ?
HALLMARK_G2M_CHECKPOINT [195]	Mouse genes annotated to HALLMARK_G2M_CHECKPOINT based on orthology mappings provided by the Alliance Genome Consortium	2		3.07 e ⁻⁴	1.54 e ⁻²

FIGURE 5.6 – Enrichement du module 90 via la plateforme GSEA

5.3 Visualisation des interactions entre les modules

La visualisation des interactions entre les modules revêt une importance cruciale dans notre démarche visant à identifier des cibles thérapeutiques potentielles. Cette étape permet de mieux comprendre les relations entre les différents processus biologiques impliqués dans notre étude. Nous avons choisi d'explorer ces relations en utilisant la plateforme en ligne "String Database" pour une analyse ontologique approfondie.

Notre approche commence par l'enrichissement de chaque module sur la plateforme GEAS, où nous rassemblons une diversité de données moléculaires. Ensuite, nous sélectionnons avec soin les éléments les plus significatifs associés à chaque voie spécifique dans chaque module. Cette démarche vise à affiner l'identification des gènes clés et des processus biologiques pertinents au sein de chaque voie.

Une fois ces éléments sélectionnés, nous utilisons la base de données String pour visualiser graphiquement les interactions entre eux. Cette analyse permet de mettre en lumière les interactions directes entre les gènes et les protéines, ainsi que les voies biologiques et les processus fonctionnels dans lesquels ils sont impliqués. Cette visualisation offre ainsi une perspective détaillée des relations complexes qui se tissent à la fois à l'intérieur de chaque module et entre les différents modules.

En résumé, le processus de visualisation des interactions entre les modules se divise en trois étapes principales :

1. Enrichissement des modules sur la plateforme GEAS : Rassemblement des données moléculaires pertinentes pour chaque module.
2. Sélection des éléments significatifs associés à chaque voie spécifique : Identification des gènes clés et des processus biologiques pertinents dans chaque module.

- Utilisation de la base de données String pour la visualisation des interactions : Représentation graphique des interactions entre les éléments sélectionnés dans chaque module, facilitant ainsi une compréhension approfondie des processus biologiques étudiés.

Cette approche nous permet de mieux appréhender les mécanismes moléculaires et d'ouvrir de nouvelles perspectives pour l'identification de cibles thérapeutiques potentielles.

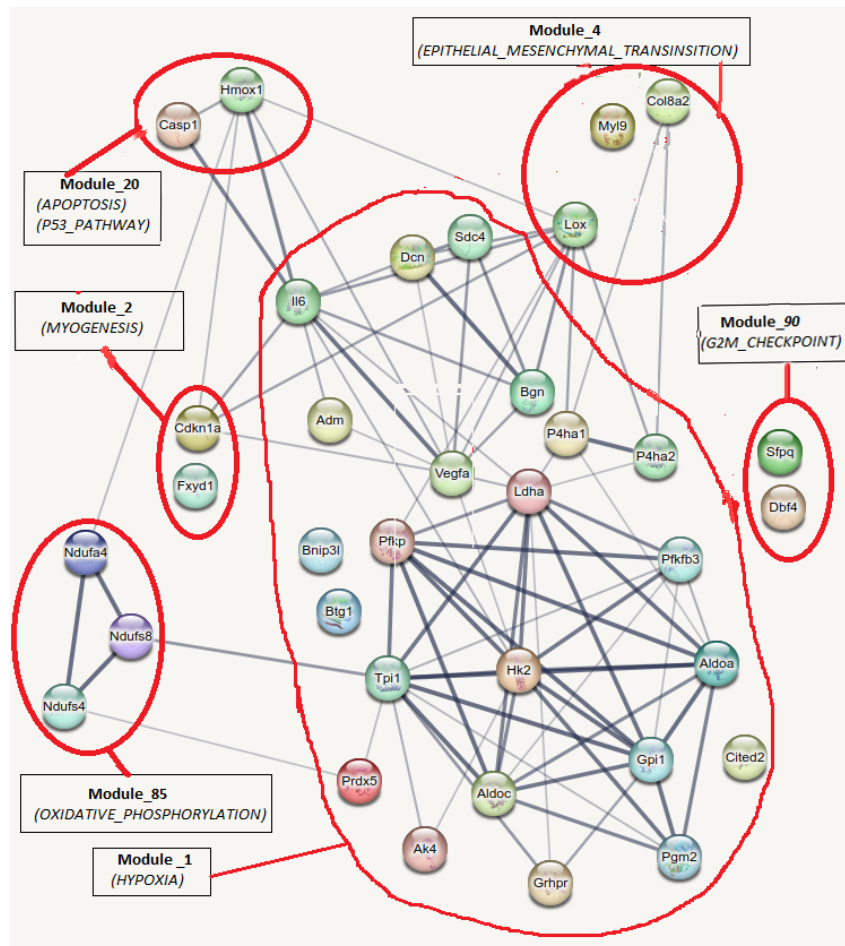


FIGURE 5.7 – La visualisation des interactions entre les modules (avec la voie hypoxie)

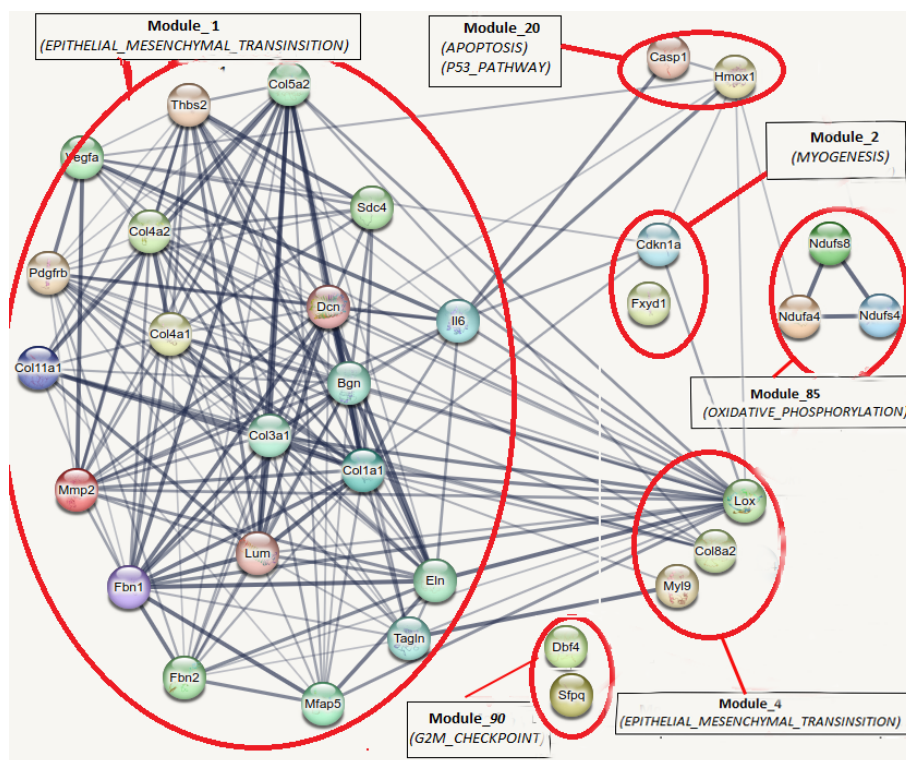


FIGURE 5.8 – La visualisation des interactions entre les modules (avec la voie EMT

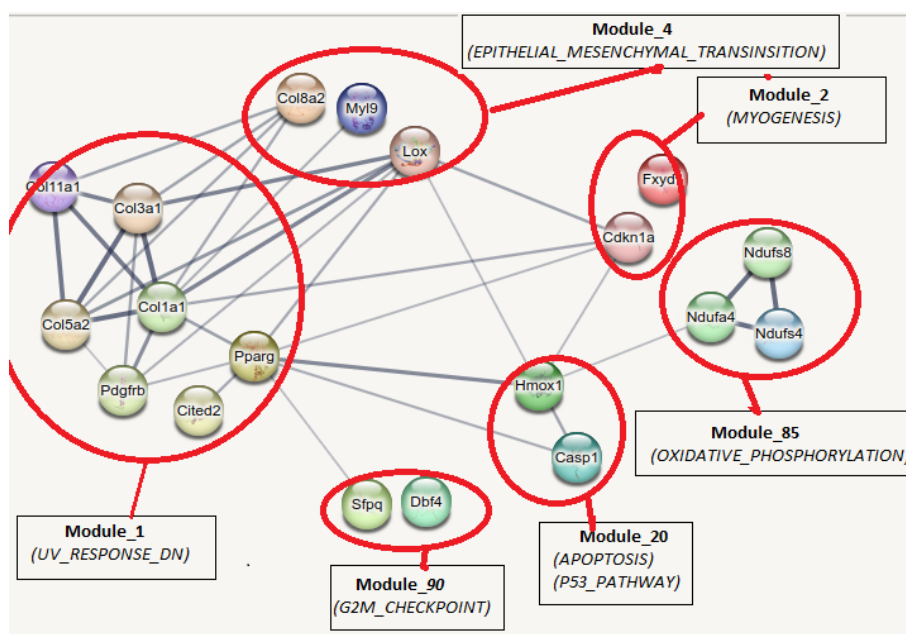


FIGURE 5.9 – La visualisation des interactions entre les modules (avec la voie uv reponse)

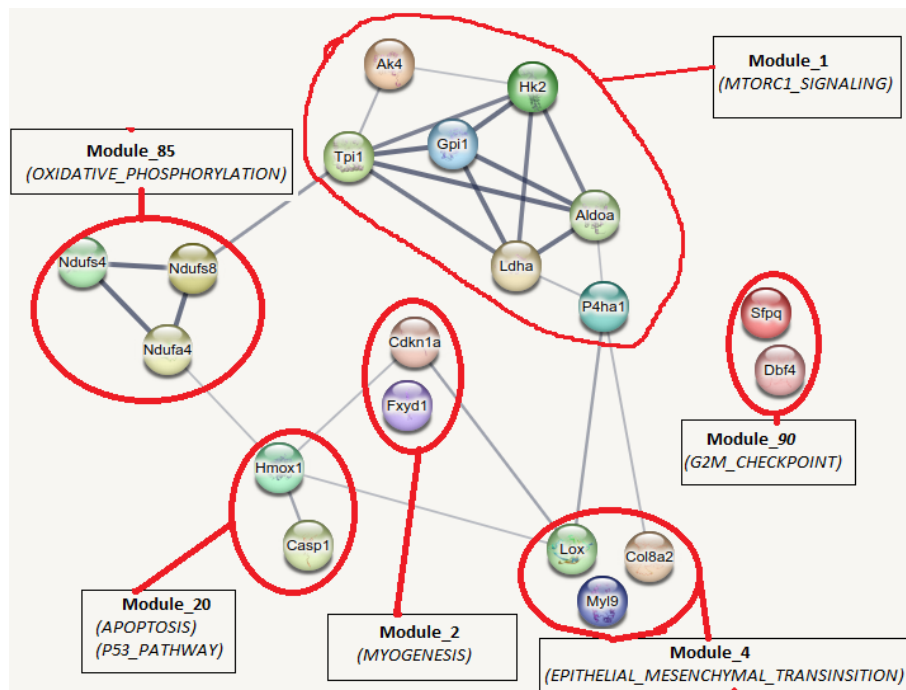


FIGURE 5.10 – La visualisation des interactions entre les modules (avec la voie MTORC1 SIGNALING)

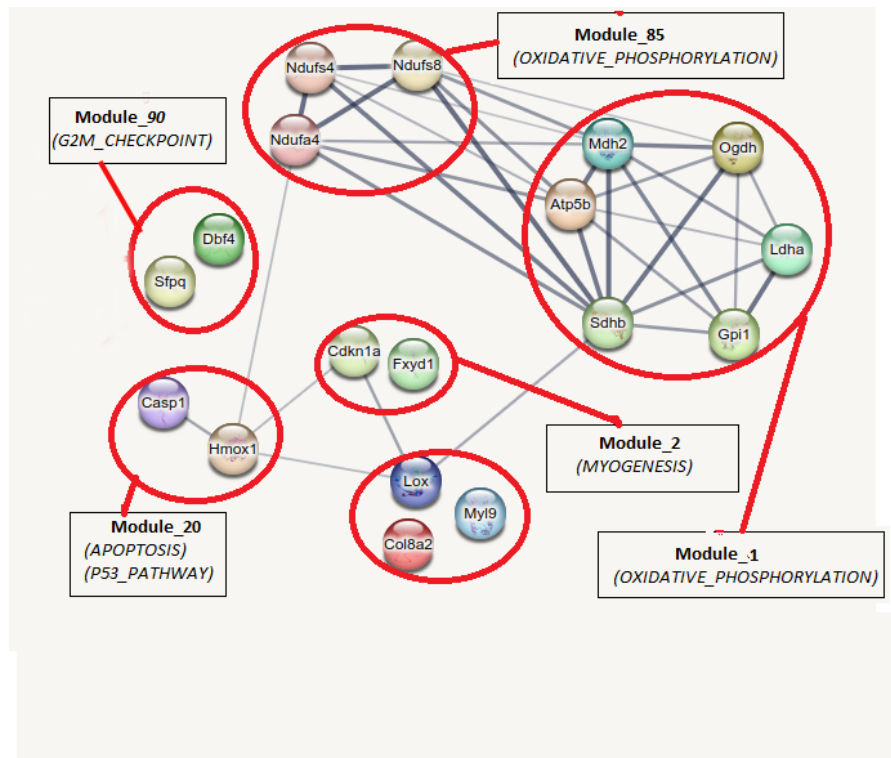


FIGURE 5.11 – La visualisation des interactions entre les modules (avec la voie OXIDATIVE PHOSPHORYLATION)

CONCLUSION

6.1 Contributions de l'étude

Notre recherche a exploré le potentiel des méthodes d'apprentissage profond pour l'identification de modules actifs, en exploitant des données multi-vues. Nous avons suivi une approche méthodique pour relever les défis de la création de vues cohérentes à partir de données biologiques complexes, et avons appliqué diverses techniques d'embedding afin d'intégrer les données de chaque vue dans un espace unifié. Notre principal objectif était de dépasser les performances de l'approche AMINE, qui représente le standard de référence dans la sélection de groupes de gènes essentiels à l'activité génétique. La sélection de ces groupes de gènes sont déterminants dans l'élaboration de thérapies ciblées, notamment pour le traitement du cancer du pancréas. Bien que nous n'ayons pas totalement atteint cet objectif, notre étude a démontré que notre méthode AMINE_multivue surpasse les autres modèles de sélection de gènes et se rapproche des performances d'AMINE.

6.2 Résumé des résultats

Nos résultats confirment l'efficacité de l'apprentissage profond, particulièrement lorsqu'il est appliqué à des données multi-vues, pour affiner la sélection de groupes de gènes et élucider les interactions génétiques complexes. Cette méthode a permis une analyse plus détaillée que celles offertes par les techniques conventionnelles, se rapprochant des performances de l'approche AMINE. Ainsi, elle jette les bases solides nécessaires au développement futur de la recherche en bioinformatique. En outre, cette avancée met en évidence le rôle crucial de la collaboration interdisciplinaire dans l'exploitation optimale des données omiques pour la construction de vues intégrées, ouvrant la voie à la découverte de nouvelles approches thérapeutiques ciblées.

6.3 Réponses aux questions de recherche et Limites

Notre étude a approfondi la compréhension de l'embedding multivue dans l'optimisation de la détection de modules actifs dans les réseaux génétiques, démontrant son potentiel à surpasser les approches traditionnelles. Nous avons identifié des critères essentiels pour la construction du graphe en nous concentrant sur les valeurs de poids des nœuds, ce qui permet de mieux

conserver l'information pertinente pour l'embedding multivue. Nous pensons que d'autres critères pourraient se greffer à la valeur de la p-value pour mettre en lumière l'importance d'une intégration soignée afin de maximiser les bénéfices de l'embedding multivue. Par rapport à la méthode AMINE, reconnue pour son efficacité dans la détection de modules actifs, l'embedding multivue offre des avantages significatifs en termes de flexibilité et de capacité à gérer des données complexes, avec des performances se rapprochant de celles d'AMINE. Cependant, malgré son caractère innovant, il convient également de reconnaître les limites de notre étude, notamment en termes de complexité computationnelle, de la nécessité d'une calibration minutieuse des paramètres et d'une bonne collaboration avec les biologistes pour élaborer des meilleurs critères de construction des vues.

6.4 Suggestions pour des travaux futurs

Il est crucial de poursuivre l'exploration de l'apprentissage profond appliqué aux données multi-vues en bioinformatique. Les recherches futures devraient se concentrer sur le développement de nouveaux critères pour la construction de vues et l'amélioration des hyperparamètres pour l'intégration des données multi-vues. Cela nécessite une collaboration étroite entre informaticiens, bioinformaticiens et biologistes pour valider efficacement les différentes vues et enrichir les analyses génomiques. Cette démarche favorisera une meilleure compréhension des processus biologiques et le développement de stratégies thérapeutiques innovantes.

BIBLIOGRAPHIE

- ATA, Sezin Kircali et al. (jan. 2020). « Multi-View Collaborative Network Embedding ». In : *ACM Transactions on Knowledge Discovery from Data (TKDD)* 1.1, Article 1. DOI : 10.1145/3441450. URL : <https://doi.org/10.1145/3441450>.
- CHIOU, S-H et al. (2017). « BLIMP1 induces transient metastatic heterogeneity in pancreatic cancer ». In : *Cancer Discov.* DOI : 10.1158/2159-8290.cd-17-0250. URL : <https://aacrjournals.org/cancerdiscovery/article/7/10/1184/5714/BLIMP1-Induces-Transient-Metastatic-Heterogeneity>.
- CROTÈS, David (2014). « Rôle du récepteur Sigma-1 sur la régulation des canaux ioniques impliqués dans la carcinogénèse ». Thèse de doctorat, École Doctorale des Sciences de la Vie et de la Santé, Interactions moléculaires et cellulaires, Institut de Biologie Valrose (CNRS UMR7277, Inserm U1091, UNS). Thèse de doct. Université de Nice Sophia-Antipolis.
- CUI, Peng et al. (2017). « A Survey on Network Embedding ». In : *IEEE Transactions on Knowledge and Data Engineering* 31.5, p. 833-852. URL : <https://arxiv.org/pdf/1711.08752.pdf>.
- GOYAL, P. et E. FERRARA (juill. 2018). « Graph embedding techniques, applications, and performance : A survey ». In : *Knowledge-Based Systems*. DOI : 10.1016/j.knosys.2018.03.022. URL : <https://doi.org/10.1016/j.knosys.2018.03.022>.
- IRINA GAYNANOVA, Gen Li (2017). « Structural learning and integrative decomposition of multi-view data ». In : *Journal of Statistical Software*. URL : <https://arxiv.org/pdf/1707.06573.pdf>.
- LESTER, Brian et al. (2020). « Multiple Word Embeddings for Increased Diversity of Representation ». In : *Interactions*.
- LOCK, Eric F. et al. (2013). « Discovering Structure in High-Dimensional Data Through Methodology and Application of the Joint and Individual Variation Explained (JIVE) Approach ». In : *Journal of Multivariate Analysis*. URL : <https://arxiv.org/pdf/1707.06573.pdf>.
- MITRA, Sayantan, Sriparna SAHA et Mohammed HASANUZZAMAN (2020). « Multi-view clustering for multi-omics data using unified embedding ». In : *Scientific Reports*. DOI : 10.1038/s41598-020-71487-3. URL : <https://doi.org/10.1038/s41598-020-71487-3>.
- OKUNO, Akifumi et Tetsuya Hada Hidetoshi SHIMODAIRA (fév. 2018). « A probabilistic framework for multi-view feature learning with many-to-many associations via neural networks ». In : DOI : 10.48550/arXiv.1802.04630. URL : <https://arxiv.org/abs/1802.04630>.
- PASQUIER, Claude et al. (juin 2022). « A network embedding approach to identify active modules in biological interaction networks ». In : *Life Science Alliance*. DOI : 10.26508/lsa.202201550. URL : <https://www.life-science-alliance.org/content/6/9/e202201550>.

- PENKE, Botond et al. (jan. 2018). « The Role of Sigma-1 Receptor, an Intracellular Chaperone in Neurodegenerative Diseases ». In : *Current Neuropharmacology* 16.1, p. 97-116. DOI : 10.2174/1570159X15666170529104323. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC5771390/>.
- RAPAPORT, P. et al. (2007). « Classification of microarray data using gene networks ». In : *BMC Bioinformatics* 8. DOI : 10.1186/1471-2105-8-35. URL : <https://arxiv.org/pdf/1707.06573.pdf>.
- XU, Xinxing, Ivor W. TSANG et Dong XU (2013). « Soft Margin Multiple Kernel Learning ». In : *Journal of Multivariate Analysis*. DOI : 10.1109/TNNLS.2012.2237183.
- ZHANG, D. et al. (juill. 2018). « Network representation learning : A survey ». In : *IEEE Transactions on Big Data*. DOI : 10.1016/j.knosys.2018.03.022. URL : <https://ieeexplore.ieee.org/document/8395024>.

ANNEXES

ILLUSTRATION DU DEROULEMENT DU PROJET

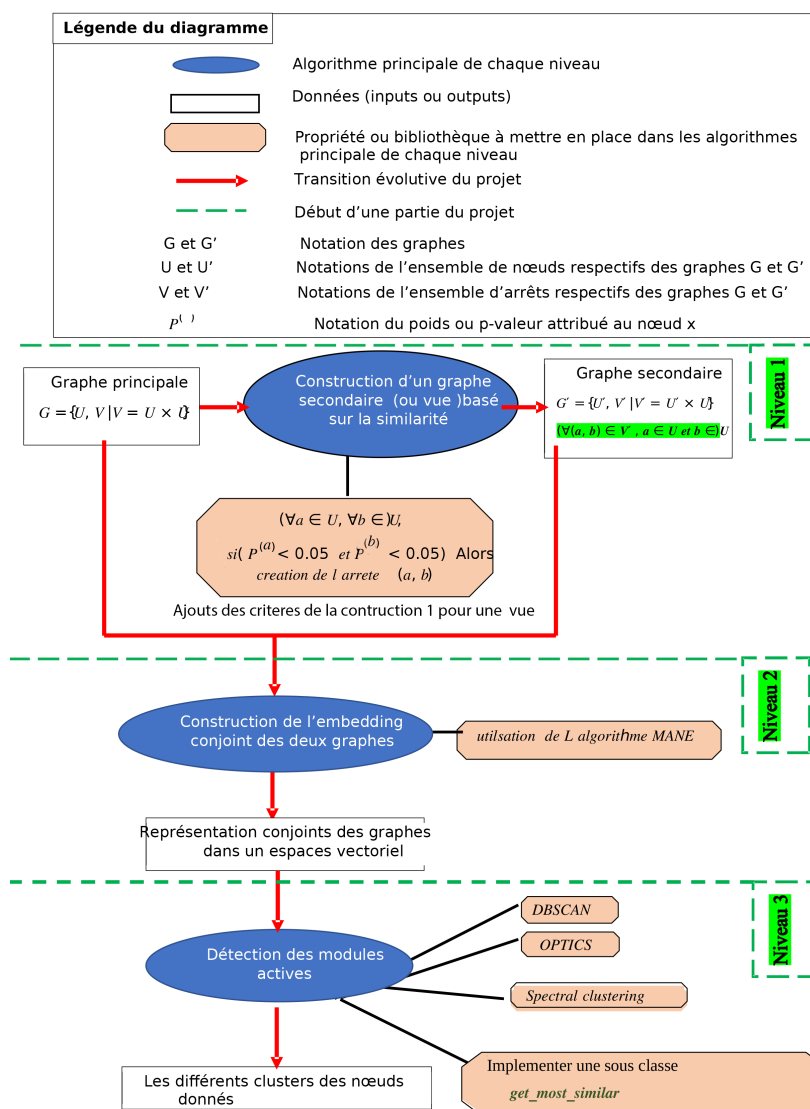


FIGURE M.1 – déroulement du projet

COURBE D'APPRENTISSAGE DE LA FONCTION DE PERTE

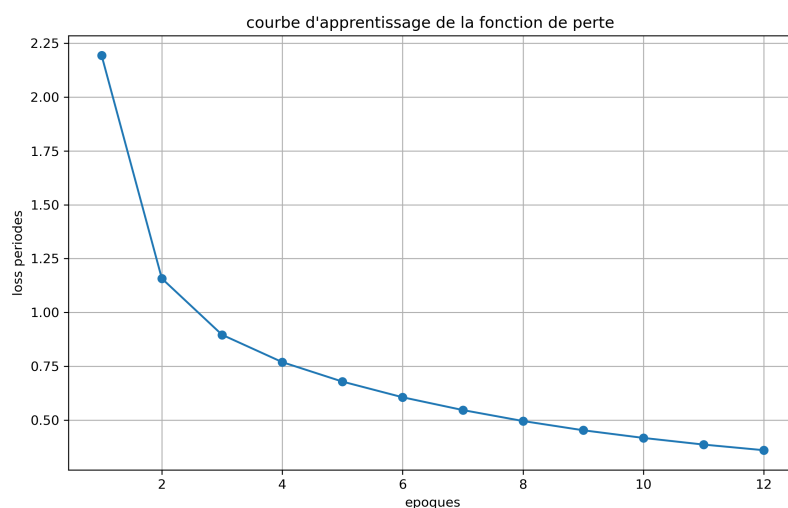


FIGURE N.1 – *La visualisation de progression de la fonction de perte avec le Framework MANE sur des Données Réelles*

PROCÉDURE DE TRAVAIL AVEC LE CADRE MANE

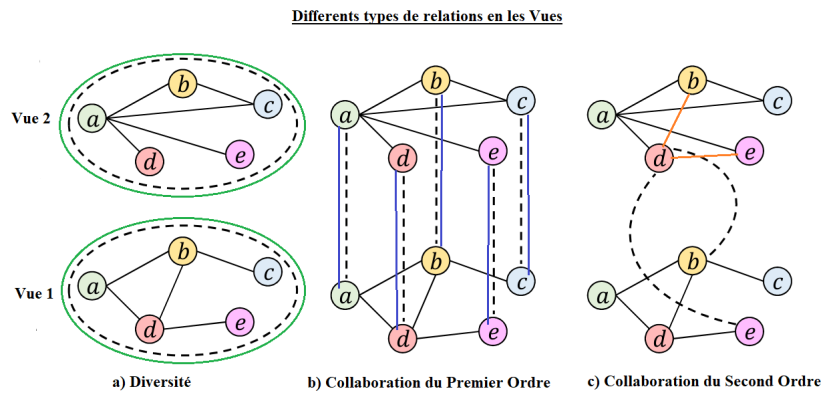


FIGURE O.1 – différents types de relation entre les vues avec le cadre MANE

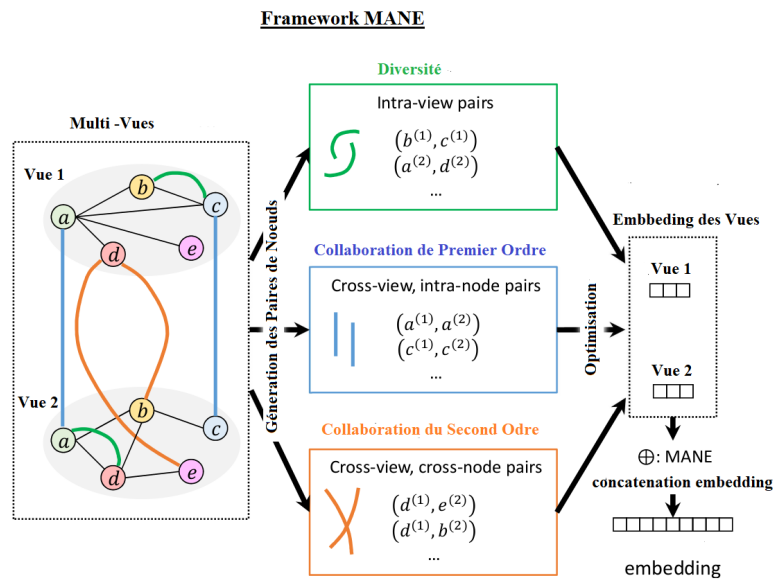


FIGURE O.2 – Framework MANE

PROCÉDURE DE TRAVAIL AVEC CADRE MvME

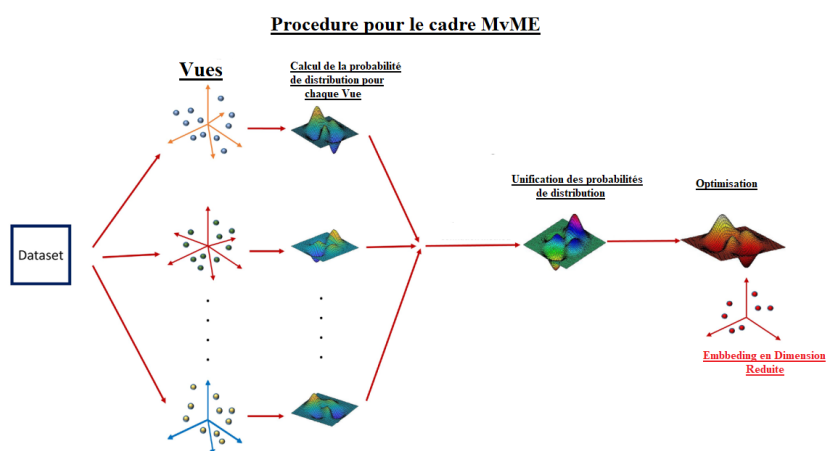


FIGURE P.1 – Procédure de travail avec MvME

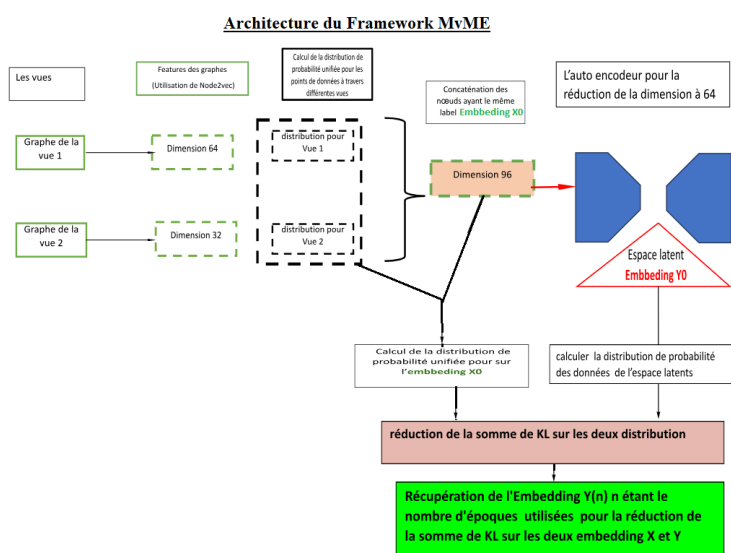


FIGURE P.2 – Framework MvME

PRESENTATION DU LABORATOIRE

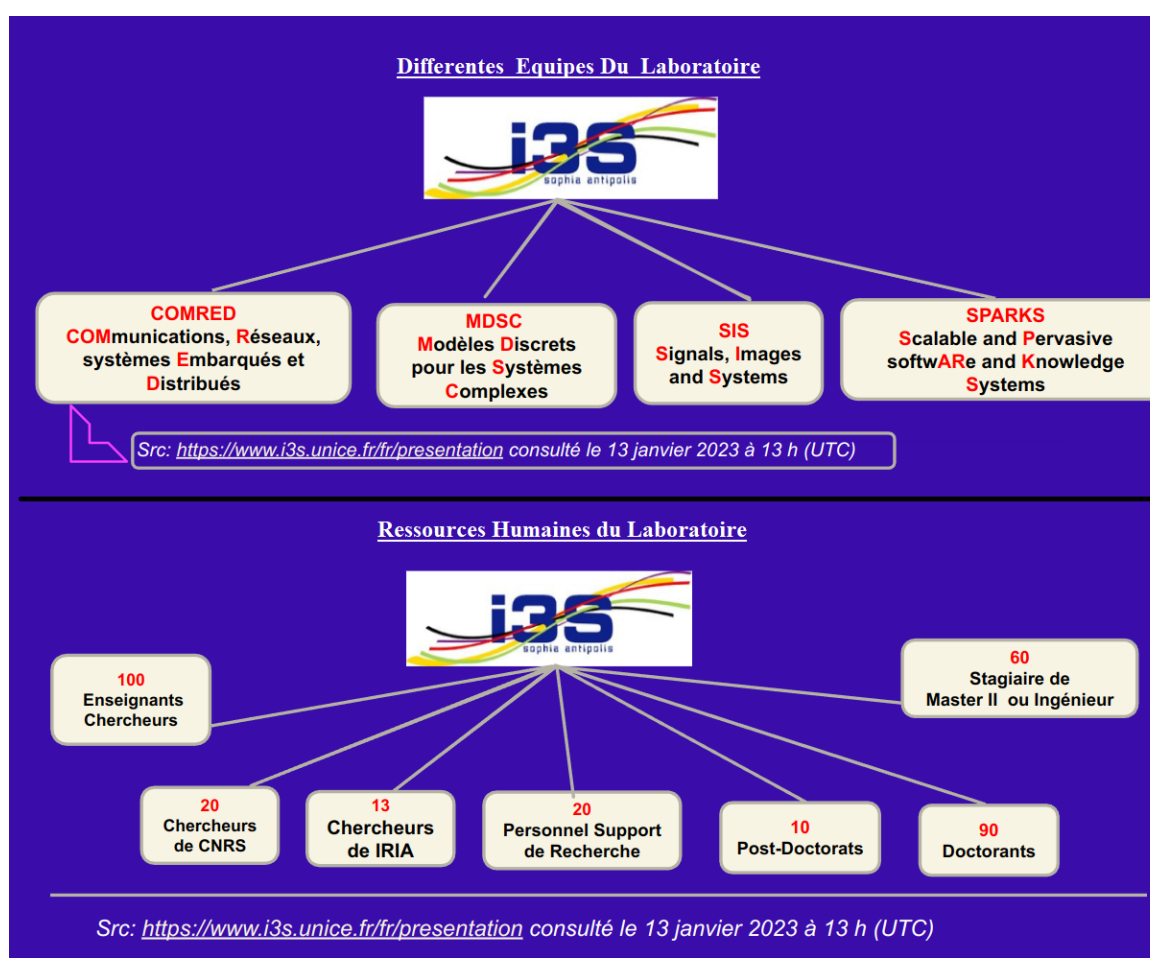


FIGURE Q.1 – Presentation du Laboratoire

PRESENTATION DE L'ÉQUIPE DE TRAVAIL SPARKS

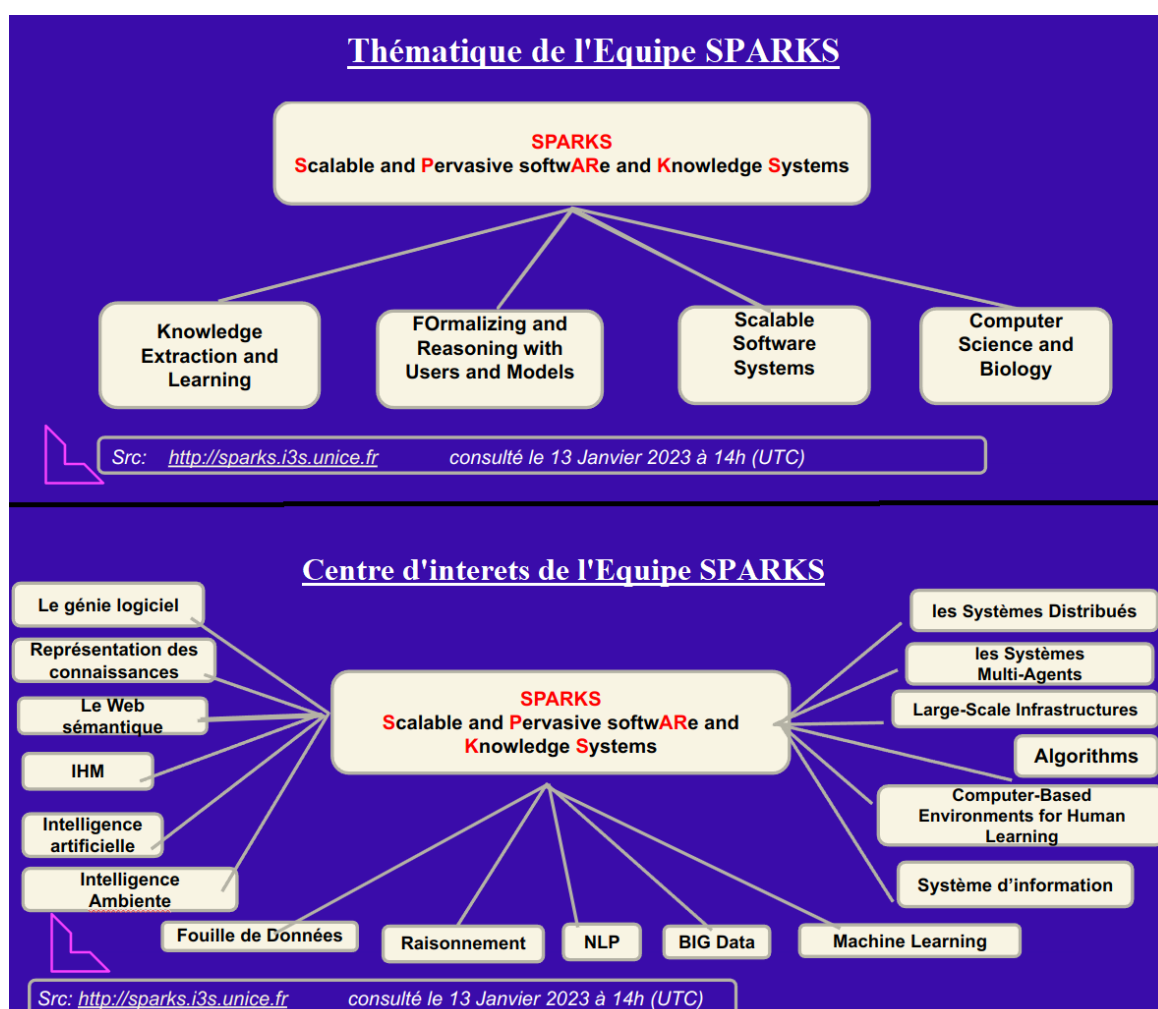


FIGURE R.1 – Presentation de l'équipe de travail SPARKS

UNIVERSITÉ NATIONALE DU
VIETNAM, HANOI

INTÉGRATION D'UNE APPROCHE
D'INTELLIGENCE ARTIFICIELLE POUR LA
DÉTECTION DE MODULES ACTIFS DANS
LES RÉSEAUX D'INTERACTIONS
BIOLOGIQUES À TRAVERS DES DONNÉES
MULTIVUES

REDIGÉ PAR: GIRESSE TCHOTANEU
NGATCHA
Student No. 2230455

SOUS LA DIRECTION DE : CLAUDE
PASQUIER

*HDR en Informatique et Chercheur en
Biologie Computationnelle*

HANOI, FÉVRIER 2024