

UNIVERSITÉ NATIONALE DU VIETNAM, HANOI

Institut de la Francophonie pour l'Innovation

Laboratoire d'Informatique, Signaux et Systèmes de Sophia Antipolis

LA ROCHELLE UNIVERSITÉ

Détermination du rôle de la protéine SigmaR1 dans le cancer du pancréas par une approche d'intelligence artificielle

MÉMOIRE DE FIN D'ÉTUDES DE MASTER EN INFORMATIQUE

Spécialité :

Systèmes Intelligents et Multimédias

Redigé par : Sous la direction de :

2023



1	Introduction	1
1.1	Contexte	1
1.2	Problématique	2
1.3	Motivation	3
1.4	Objectifs de la recherche	4
	Objectifs principaux	4
	Objectifs secondaires	5
1.5	Lacunes de recherche	5
1.6	Questions de recherche	7
2	Chapter 1: Revue de la littérature	9
2.1	Introduction	9
2.2	Théories et modèles existants	9
2.3	Présentation des principaux modèles generant les embbedding multivues	9
2.4	Défis computationnels	9
2.5	Lacunes de recherche existantes	9
3	Chapter 3:Méthodologie	10
3.1	Cadre conceptuel	10
3.2	Conception du modèle	10
3.3	Validation du modèle	10
3.4	Approche de simulation	10
3.5	Collecte de données	10
3.6	Analyse de données	10
3.7	Justification des méthodes choisies	10

4	Chapter 4: Résultats	11
4.1	Statistiques descriptives	11
4.2	les metriques utilisées	11
4.3	Statistiques inférentielles	11
4.4	Réponses aux questions de recherche	11
4.5	Visualisation	11
4.6	Analyse de sensibilité des cluster	11
4.7	Comparaison avec d'autres modèles clusters	11
5	Discussion	12
5.1	Comparaison avec Amine	12
5.2	Interprétation des résultats	12
5.3	Implications pratiques	12
5.4	Limites de l'étude	12
5.5	Contributions de l'étude	12
5.6	Suggestions pour des travaux futurs	12
6	Conclusion	13
6.1	Résumé des résultats	13
6.2	Réponses aux questions de recherche	13
6.3	Contributions de l'étude	13
6.4	Limites de l'étude	13
6.5	Suggestions pour des travaux futurs	13
6.6	Réflexion personnelle	13

Depuis les débuts de la biologie moléculaire, les chercheurs ont aspiré à décrypter les interactions complexes qui se manifestent au niveau moléculaire. Historiquement, les techniques en biologie moléculaire se sont concentrées sur l'étude d'interactions spécifiques, souvent à une échelle limitée. Des méthodes telles que le Western blot, l'immunohistochimie et l'analyse des polymorphismes de l'ADN ont constitué des outils inestimables pour les biologistes. Toutefois, bien que ces méthodes soient puissantes, elles étaient généralement restreintes à l'étude d'interactions binaires ou à petite échelle [ZK19].

1.1 Contexte

L'émergence des technologies actuelles à haut débit sont désormais capables de quantifier de manière fiable, à l'échelle d'un organisme entier, les changements moléculaires qui surviennent en réponse à des maladies ou à des perturbations environnementales [Pas+23]. Ces innovations ont permis aux chercheurs de générer d'immenses volumes de données, offrant une vue sans précédent des processus biologiques à une échelle bien plus large. Auparavant, les chercheurs se concentraient sur des études ciblées, examinant des gènes ou des protéines individuels. Aujourd'hui, ils peuvent étudier l'ensemble du génome, de l'épigénome ou du protéome d'un organisme, ouvrant ainsi la voie à des découvertes qui étaient auparavant inimaginables [Mer+14].

Ces avancées technologiques ont conduit à l'émergence de la biologie des systèmes, une approche qui vise à étudier les systèmes biologiques dans leur ensemble plutôt que de se concentrer sur des composants individuels. Cette approche reconnaît que les processus biologiques sont le résultat d'interactions complexes entre de nombreux composants, et que comprendre ces interactions est essentiel pour comprendre la biologie à un niveau plus profond.

Toutefois, cette abondance de données a posé un défi majeur : comment analyser, interpréter et tirer des conclusions pertinentes de ces vastes ensembles de données ? Les données génomiques, transcriptomiques, protéomiques et métabolomiques sont souvent volumineuses et complexes, avec des milliers, voire des millions, de points de données à analyser. De plus, ces données sont souvent bruyantes, incomplètes et peuvent être affectées par divers artefacts techniques.

Étant donné que certaines activités cellulaires sont parfois très interdépendantes, la question se pose de savoir comment fusionner les données de différentes activités pour une analyse efficace. Par exemple, une mutation génétique peut affecter l'expression de nombreux gènes, qui à leur tour peuvent affecter la fonction de plusieurs protéines, conduisant à des changements phénotypiques observables. Comprendre ces chaînes causales nécessite une analyse intégrée des données à plusieurs niveaux. La réponse à ces défis se trouve à la croisée de la biologie, des mathématiques et de l'informatique.

Les méthodes de bio-informatique et d'analyse systémique sont devenues essentielles pour naviguer dans ce labyrinthe complexe de données omiques. Des algorithmes sophistiqués sont développés pour traiter, analyser et visualiser ces données, permettant aux chercheurs de découvrir des motifs, des associations et des causalités qui étaient auparavant cachés [Wu+14].

En outre, l'importance de la collaboration interdisciplinaire est devenue évidente. Les biologistes collaborent avec des mathématiciens et des informaticiens pour développer des outils et des méthodes pour traiter ces données multivues, parfois représentées sous forme de graphes où les nœuds sont des protéines et les arêtes sont des relations entre ces protéines. Ces collaborations ont conduit à des avancées majeures dans notre compréhension de la biologie et ont ouvert la voie à de nouvelles approches thérapeutiques pour traiter les maladies. Face à ces défis, de nouvelles approches et méthodes ont été nécessaires pour permettre aux chercheurs de naviguer efficacement dans le paysage complexe des données omiques et d'en tirer des informations significatives. La difficulté réside dans le fait que ces méthodes doivent être capables de gérer la complexité et la dimensionnalité élevée des données, tout en étant suffisamment flexibles pour s'adapter aux changements et aux nouvelles découvertes.

La représentation, l'analyse et l'interprétation de ces graphes posent un défi majeur : comment obtenir une représentation vectorielle (embedding) pertinente des activités cellulaires à partir de ces graphes ?

1.2 Problématique

La biologie moléculaire a connu une évolution rapide avec l'émergence des technologies de séquençage à haut débit et des plateformes omiques, permettant aux chercheurs de générer d'immenses volumes de données. Ces données, offrant une vue sans précédent sur les processus biologiques et les activités cellulaires avec leurs interactions complexes et leurs réseaux de régulation, sont souvent représentées sous forme de graphes. Ces graphes peuvent illustrer diverses interactions, telles que les interactions protéine-protéine, les voies métaboliques ou les réseaux de régulation génique. Bien que les graphes soient des outils puissants pour visualiser et comprendre ces interactions, leur analyse à grande échelle, en particulier dans le contexte des données omiques, peut s'avérer complexe. L'un des défis majeurs en bio-informatique est la représentation vectorielle des données.

L'embedding de graphes, ou la transformation de ces graphes en représentations vectorielles, est une technique qui vise à transformer les nœuds, les arêtes ou les structures entières des graphes en vecteurs

de faible dimension tout en préservant certaines propriétés du graphe. Un embedding réussi permettrait de capturer l'essence des interactions et des motifs dans le graphe sous une forme qui peut être facilement analysée à l'aide de techniques d'apprentissage automatique telles que la classification des cellules, la prédiction des fonctions géniques ou la découverte de nouvelles interactions. Cette transformation est cruciale pour permettre l'application de techniques d'apprentissage automatique et d'analyse de données sur des graphes. Dans le contexte de la biologie moléculaire, l'embedding de graphes pourrait permettre de mieux comprendre les interactions moléculaires complexes, comme celles observées dans les interactions médicamenteuses [SN22] ou l'organisation spatiale de l'environnement tumoral [Din+21].

Plusieurs méthodes ont été proposées pour obtenir des embeddings de graphes, allant des techniques basées sur la factorisation de matrices aux approches basées sur l'apprentissage profond comme les Graph Neural Networks (GNN). Cependant, chaque méthode a ses avantages et ses inconvénients, et le choix de la meilleure méthode dépend souvent de la nature spécifique des données et du problème à résoudre. Malgré son potentiel, l'embedding de graphes en biologie moléculaire reste un domaine émergent avec de nombreux défis à relever. Par exemple, comment garantir que l'embedding préserve les propriétés biologiques essentielles des graphes ? Comment intégrer efficacement différentes sources de données omiques dans un embedding unifié ? De plus, comment utiliser ces embeddings pour faire des découvertes biologiques significatives, comme la prédiction des modules actifs (truehits) [Ash+20] ?

Dans ce travail, nous explorons différentes techniques d'embedding de graphes dans le contexte des activités cellulaires. Nous cherchons à répondre à des questions clés telles que : Comment peut-on obtenir une représentation vectorielle qui capture fidèlement les informations contenues dans le graphe ? Quelle est la meilleure méthode pour intégrer ces embeddings dans l'analyse des données omiques multi-vues ? Et comment ces embeddings peuvent-ils être utilisés pour faire avancer notre compréhension du cancer du pancréas ? Face à ces défis, il est essentiel de développer de nouvelles méthodes et approches pour l'embedding de graphes en biologie moléculaire. Ces méthodes doivent non seulement être mathématiquement rigoureuses, mais aussi biologiquement pertinentes pour garantir leur applicabilité dans la recherche biomédicale.

1.3 Motivation

Le cancer du pancréas est l'un des cancers ayant le taux de mortalité le plus élevé. Une étude menée par Xiaoyue Jia et al. a révélé une augmentation significative du taux de mortalité associé à ce cancer en Chine au cours de la dernière décennie, soulignant ainsi l'urgence de la situation [Jia+18]. Malgré les avancées notables dans la compréhension de la biologie du cancer, trouver un traitement efficace pour le cancer du pancréas demeure un défi majeur. Dans ce contexte, la protéine SigmaR1 est apparue comme une cible thérapeutique prometteuse.

La complexité du cancer du pancréas, avec ses multiples voies de signalisation et ses mécanismes

d'évasion immunitaire, nécessite une approche intégrée pour comprendre et traiter efficacement la maladie. Les avancées récentes dans les technologies omiques ont ouvert la voie à une compréhension plus profonde des mécanismes moléculaires sous-jacents, offrant ainsi de nouvelles opportunités pour le développement de thérapies ciblées.

Selon une étude récente, Sigma1 (également connu sous le nom de récepteur sigma-R1) est une protéine chaperonne régulée pharmacologiquement qui joue un rôle varié dans la survie cellulaire. Plusieurs publications ont suggéré un rôle pour SigmaR1 dans le cancer, et des composés ayant une affinité pour Sigma1 ont été identifiés comme inhibant la prolifération et la survie des cellules cancéreuses, l'adhésion et la migration cellulaire, ainsi que la croissance tumorale [KM17]. De plus, une autre étude a montré que Sigma1 régule le métabolisme des gouttelettes lipidiques dans les cellules du cancer de la prostate, suggérant un rôle essentiel dans la régulation de l'homéostasie redox nécessaire aux programmes métaboliques oncogéniques favorisant la prolifération du cancer [OS+23].

Notre motivation pour ce sujet est double :

- D'une part, il y a le défi scientifique : comment fusionner les différentes perspectives représentées par des graphes pour extraire un maximum d'informations sur les activités des composants moléculaires et leurs interactions ? Les techniques traditionnelles d'analyse des données, bien qu'efficaces pour certains ensembles de données, peuvent ne pas être adaptées à la fusion des différentes perspectives des données.
- D'autre part, il y a l'impératif clinique : chaque découverte offre l'espoir de nouvelles thérapies pouvant améliorer la vie des patients atteints de cancer du pancréas. Dans notre cas, il s'agit de la possibilité de traduire les découvertes en laboratoire en recherches plus efficaces sur de véritables échantillons biologiques.

Face à l'augmentation constante du nombre de cas de cancer du pancréas et à la gravité de la maladie, il est impératif de poursuivre les recherches pour développer des traitements plus efficaces. En combinant ces motivations, notre recherche vise à combler le fossé entre la génération de données et leur fusion en multi-vues. En développant de nouvelles méthodes pour l'analyse des données omiques et en explorant le rôle de SigmaR1 dans le cancer du pancréas, nous espérons contribuer à la recherche sur le cancer et, à terme, améliorer les résultats pour les échantillons biologiques.

1.4 Objectifs de la recherche

Objectifs principaux

Le cancer du pancréas est l'une des maladies les plus mortelles avec un taux de survie relativement faible. Dans ce contexte, l'objectif principal de cette recherche est de développer une méthode intégrée pour fusionner et analyser les données omiques multi-vues. Cette approche vise à améliorer la compréhension des

mécanismes sous-jacents de la maladie et à mettre en évidence de nouvelles cibles thérapeutiques potentielles. Plus précisément, nous visons à :

- Développer une méthode bio-informatique innovante pour intégrer efficacement les données omiques multi-vues. Cette méthode combinera des techniques d'analyse de graphes et d'apprentissage automatique pour offrir une meilleure représentation des données.
- Identifier les sous-groupes actifs dans l'embedding obtenu des vues fusionnées. Ces sous-groupes pourraient représenter des phénotypes ou des états de la maladie qui n'ont pas encore été découverts.
- Valider les sous-groupes identifiés en utilisant des données artificielles. Cela permettra de s'assurer que les groupes identifiés ne sont pas le résultat de bruits ou d'artefacts dans les données. Nous nous appuierons sur la continuité de l'algorithme Amine développé par [Pas+23] pour cette validation.
- Explorer le rôle potentiellement crucial de la protéine SigmaR1 dans le cancer du pancréas. Cette protéine, en interagissant avec d'autres protéines, pourrait jouer un rôle clé dans la progression de la maladie.

Objectifs secondaires

Outre les objectifs principaux, cette recherche vise également à :

- Évaluer la performance de notre méthode. Il est essentiel de s'assurer que notre approche est non seulement innovante mais aussi efficace. Pour ce faire, nous la comparerons à d'autres méthodes existantes dans la littérature, notamment AMINE [Pas+23].
- Identifier d'autres protéines ou voies moléculaires potentiellement impliquées dans le cancer du pancréas. En utilisant différentes méthodes de clustering sur l'embedding résultant des multi-vues, nous pourrions découvrir de nouvelles cibles thérapeutiques.
- Enfin, nous explorerons les implications cliniques de nos découvertes. Si nos résultats sont prometteurs, ils pourraient ouvrir la voie à de nouvelles stratégies thérapeutiques et à des expérimentations sur des échantillons biologiques pour confirmer nos hypothèses.

1.5 Lacunes de recherche

L'intégration des données omiques multi-vues est un domaine en pleine évolution. Bien que de nombreuses avancées aient été réalisées, plusieurs lacunes demeurent dans la littérature :

- Intégration limitée des données omiques : La plupart des méthodes existantes se concentrent sur l'intégration de deux types de données omiques, laissant de côté la complexité inhérente à l'intégration

de multiples types de données. Cette limitation empêche une compréhension holistique des interactions moléculaires complexes. Les données génomiques, transcriptomiques, protéomiques et métabolomiques, bien que complémentaires, sont souvent traitées séparément, ce qui peut entraîner une perte d'informations cruciales [HCG17].

- Manque de prise en compte de la nature hiérarchique : Les données omiques sont souvent interdépendantes et hiérarchiques. Cependant, peu de méthodes actuelles tiennent compte de cette nature lors de l'intégration. Cette lacune peut conduire à des représentations inexactes ou incomplètes, car les interactions entre différentes échelles omiques peuvent être omises ou mal interprétées [VM22].
- Rôle mal compris de certaines protéines : Bien que certaines protéines, comme SigmaR1, aient été identifiées comme des cibles thérapeutiques potentielles, leur rôle précis et leurs interactions avec d'autres protéines restent mal compris. Cette lacune souligne la nécessité d'une approche intégrative pour étudier les protéines et leurs interactions dans le contexte des maladies [Dua+21].
- Défis de l'embedding de graphes : Transformer les données omiques, souvent représentées sous forme de graphes, en représentations vectorielles est un défi majeur. Les méthodes actuelles, bien que prometteuses, sont confrontées à des défis tels que la préservation des propriétés biologiques essentielles des graphes, la gestion de la grande dimensionnalité des données, et l'intégration efficace de différentes sources de données omiques dans un embedding unifié [btad353].
- Analyse à grande échelle : Avec l'augmentation exponentielle des données omiques disponibles, les méthodes actuelles d'analyse et d'intégration peuvent ne pas être adaptées à une analyse à grande échelle. Les outils et les infrastructures doivent être adaptés pour gérer ces grands volumes de données, tout en garantissant une analyse précise et pertinente [CG22].
- Méthodes d'intégration non optimisées : Bien que plusieurs méthodes d'intégration des données omiques aient été proposées, il n'existe pas de consensus sur la meilleure méthode. Chaque méthode a ses avantages et ses inconvénients, et le choix de la meilleure méthode dépend souvent de la nature spécifique des données et du problème à résoudre. De plus, la validation de ces méthodes est souvent limitée à des ensembles de données spécifiques, ce qui rend difficile leur généralisation à d'autres contextes [WK23].
- Manque de standards : L'absence de standards pour l'intégration des données omiques multi-vues rend difficile la comparaison des résultats entre différentes études. Cela limite également la reproductibilité des recherches, un élément clé pour valider et généraliser les découvertes.
- Complexité computationnelle : L'intégration des données omiques multi-vues est souvent associée à une complexité computationnelle élevée. Les méthodes actuelles nécessitent souvent des ressources

computationnelles importantes, ce qui peut limiter leur applicabilité à des ensembles de données de grande taille.

Ces lacunes soulignent la nécessité de poursuivre les recherches dans ce domaine pour développer des méthodes plus robustes et précises d'intégration des données omiques multi-vues.

1.6 Questions de recherche

Avec l'explosion des données omiques, la nécessité d'outils informatiques robustes pour traiter, intégrer et interpréter ces données est devenue primordiale. Les données omiques, en raison de leur complexité et de leur volume, nécessitent des techniques avancées pour être transformées en informations exploitables. L'une des approches prometteuses dans ce domaine est la construction d'embeddings, qui vise à représenter ces données sous une forme vectorielle, facilitant ainsi leur analyse et leur interprétation. Une fois ces embeddings créés, la détection de clusters ou de sous-groupes au sein de ces données peut révéler des motifs ou des structures cachés, offrant des insights précieux sur les interactions moléculaires et les phénomènes biologiques.

Dans ce contexte, plusieurs questions cruciales se posent, notamment sur la manière de construire efficacement ces embeddings, sur les techniques optimales pour détecter des clusters pertinents et sur la manière d'évaluer et de valider les résultats obtenus. Ces questions sont au cœur de la fusion des domaines de la bio-informatique et de l'analyse des données, et y répondre pourrait ouvrir la voie à de nouvelles découvertes et innovations dans le domaine de la biologie moléculaire. Ainsi, nous formulons les questions de recherche suivantes :

1. Construction d'Embeddings pour les Données Omiques :

- Quelles sont les meilleures techniques d'embedding pour transformer efficacement les données omiques multi-vues en représentations vectorielles?
- Comment les propriétés et les caractéristiques des données omiques influencent-elles la qualité et la pertinence des embeddings générés?
- Quels sont les défis computationnels associés à la création d'embeddings pour de grands ensembles de données omiques, et comment peuvent-ils être surmontés?

2. Détection de Clusters à partir des Embeddings :

- Quelles méthodes de clustering sont les plus efficaces pour identifier des sous-groupes pertinents à partir des embeddings des données omiques?
- Comment la dimensionnalité et la structure des embeddings influencent-elles la performance des algorithmes de clustering ?

- Existe-t-il des techniques spécifiques d'optimisation qui peuvent améliorer la précision et la robustesse de la détection de clusters dans le contexte des données omiques ?

3. Évaluation et Validation :

- Comment évaluer objectivement la qualité des embeddings générés en termes de préservation de l'information biologique?
- Quels critères ou métriques peuvent être utilisés pour évaluer la pertinence et la cohérence des clusters identifiés?
- Comment les résultats obtenus à partir des embeddings et du clustering peuvent-ils être validés en utilisant des données biologiques réelles ou des ensembles de données de référence ?

4. Comparaison avec les Approches Existantes :

- Comment la méthode d'embedding proposée dans cette recherche se compare-t-elle aux techniques existantes en termes d'efficacité, de rapidité et de précision?
- Quels sont les avantages et les inconvénients des différentes méthodes de clustering lorsqu'elles sont appliquées aux embeddings des données omiques?
- Dans quelle mesure les techniques proposées dans cette recherche améliorent-elles la capacité à extraire des informations significatives des données omiques par rapport aux méthodes existantes ?

2.1 Introduction**2.2 Théories et modèles existants****2.3 Présentation des principaux modèles generant les embbedding multivues****2.4 Défis computationnels****2.5 Lacunes de recherche existantes**

3.1 Cadre conceptuel**3.2 Conception du modèle****3.3 Validation du modèle****3.4 Approche de simulation****3.5 Collecte de données****3.6 Analyse de données****3.7 Justification des méthodes choisies**

- 4.1 Statistiques descriptives**
- 4.2 les metriques utilisées**
- 4.3 Statistiques inférentielles**
- 4.4 Réponses aux questions de recherche**
- 4.5 Visualisation**
- 4.6 Analyse de sensibilité des cluster**
- 4.7 Comparaison avec d'autres modèles clusters**

- 5.1 Comparaison avec Amine**
- 5.2 Interprétation des résultats**
- 5.3 Implications pratiques**
- 5.4 Limites de l'étude**
- 5.5 Contributions de l'étude**
- 5.6 Suggestions pour des travaux futurs**

6.1 Résumé des résultats**6.2 Réponses aux questions de recherche****6.3 Contributions de l'étude****6.4 Limites de l'étude****6.5 Suggestions pour des travaux futurs****6.6 Réflexion personnelle**

Bibliographie

- [Ash+20] Ashoor, Haitham et al. (2020). “Graph embedding and unsupervised learning predict genomic sub-compartments from HiC chromatin interaction data”. In: *Nature Communications*. URL: <https://dx.doi.org/10.1038/s41467-020-14974-x>.
- [CG22] Cao, Zhi-Jie and Ge Gao (2022). “Multi-omics single-cell data integration and regulatory inference with graph-linked embedding”. In: *Nature Biotechnology*. URL: <https://www.nature.com/articles/s41587-022-01284-4>.
- [Din+21] Ding, Liang et al. (2021). “Abstract 237: Inferring spatial organization of tumor microenvironment from single-cell RNA sequencing data using graph embedding”. In: *Cancer Research*. URL: <https://dx.doi.org/10.1158/1538-7445.AM2021-237>.
- [Dua+21] Duan, R. et al. (2021). “Evaluation and comparison of multi-omics data integration methods for cancer subtyping”. In: *PLOS Computational Biology*. URL: <https://journals.plos.org/ploscompbiol/article/file?id=10.1371/journal.pcbi.1009224&type=printable>.
- [HCG17] Huang, Sijia, Kumardeep Chaudhary, and Lana Garmire (2017). “More Is Better: Recent Progress in Multi-Omics Data Integration Methods”. In: *Frontiers in Genetics*. URL: <https://www.frontiersin.org/articles/10.3389/fgene.2017.00084/pdf>.
- [Jia+18] Jia, Xiaoyue et al. (2018). “Pancreatic Cancer Mortality in China: Characteristics and Prediction”. In: *BioMed Research International*. URL: <https://dx.doi.org/10.1097/MPA.0000000000000976>.
- [KM17] Kim, E and Christina M. Maher (2017). “Signal Pharmacology in the Context of Cancer”. In: *Handbook of Experimental Pharmacology*. URL: https://dx.doi.org/10.1007/164_2017_38.
- [Mer+14] Merelli, I. et al. (2014). “High-Performance Computing and Big Data in Omics-Based Medicine”. In: *BioMed Research International*. URL: <https://downloads.hindawi.com/journals/bmri/2014/825649.pdf>.

- [OS+23] Oyer, Halley M., Alexandra R. Steck, et al. (2023). “Abstract 284: Sigma1 regulates lipid droplet mediated redox homeostasis required for prostate cancer proliferation”. In: *Cancer Research*. URL: <https://dx.doi.org/10.1158/1538-7445.am2023-284>.
- [Pas+23] Pasquier, Claude et al. (June 2023). “A network embedding approach to identify active modules in biological interaction networks”. In: *Life Science Alliance*. URL: <https://www.life-science-alliance.org/content/6/9/e202201550>.
- [SN22] Sadeghi, Seyedeh Shaghayegh and A. Ngom (2022). “DDIPred: Graph Convolutional Network-Based Drug-drug Interactions Prediction Using Drug Chemical Structure Embedding”. In: *IEEE International Conference on Computational Intelligence in Bioinformatics and Computational Biology (CIBCB)*. URL: <https://dx.doi.org/10.1109/CIBCB55180.2022.9863018>.
- [VM22] Vahabi, Nasim and G. Michailidis (2022). “Unsupervised Multi-Omics Data Integration Methods: A Comprehensive Review”. In: *Frontiers in Genetics*. URL: <https://www.frontiersin.org/articles/10.3389/fgene.2022.854752/pdf>.
- [WK23] Wekesa, Jael Sanyanda and Michael Kimwele (2023). “A review of multi-omics data integration through deep learning approaches for disease diagnosis, prognosis, and treatment”. In: *Frontiers in Genetics*. URL: <https://doi.org/10.3389/fgene.2023.1199087>.
- [Wu+14] Wu, F. et al. (2014). “Biomolecular Networks and Human Diseases”. In: *BioMed Research International*, URL: <https://downloads.hindawi.com/journals/bmri/2014/363717.pdf>.
- [ZK19] Zhang, B. and B. Kuster (2019). “Proteomics Is Not an Island: Multi-omics Integration Is the Key to Understanding Biological Systems”. In: *Molecular & Cellular Proteomics*. URL: [https://www.mcponline.org/article/S1535-9476\(20\)32756-0/pdf](https://www.mcponline.org/article/S1535-9476(20)32756-0/pdf).