INFORMATIONS GENERALES

Intitulé du projet : Apprentissage géométrique : vers des réseaux de neurones sur graphes à haute expressivité

English version : Graph Learning : toward **Hi**gh e**X**pressivity of GNNs)

Acronyme du projet : GrapHiX

Etablissement porteur : Université de Rouen Normandie

Localisation du projet (nom du laboratoire et adresse) : Laboratoire d'Informatique, du Traitement de l'Information et des Systèmes, LITIS UR 4108

Discipline du projet : Informatique

Coordinateur du projet dans le laboratoire d'accueil :

Nom : ADAM

Prénom : Sébastien

Courriel : Sebastien.Adam@univ-rouen.fr

Équipe de recherche (si existante) : Équipe Apprentissage

Adresse : Avenue de l'Université

N° - Libellé de la voie :

Code postal : 76800

Commune : Saint Etienne du Rouvray

École doctorale de rattachement du directeur de thèse : *MIIS*

Période d'exécution du projet : Du *01/10/2025* au *30/09/2028* , soit *36* mois de projet.

RÉSUMÉ DU PROJET

Résumé vulgarisé du projet :

Version Française

Les objets qui nous entourent se caractérisent non seulement par leurs propriétés intrinsèques, mais également par les relations qu'ils entretiennent avec d'autres objets. Si ces relations sont souvent séquentielles (par exemple, un texte est une séquence de mots, une image est une grille 2D de pixels), elles peuvent également adopter des formes plus complexes, ne suivant pas une structure régulière. Parmi les exemples notables figurent les interactions sociales dans un réseau, les liaisons chimiques entre atomes dans une molécule ou les interactions entre acides aminés au sein d'une protéine.

Dans de tels cas, les graphes constituent un cadre mathématique et informatique pertinent pour représenter ces relations. Dès lors que les données à étudier sont représentées par des graphes, la prédiction des propriétés associées à ces données et la génération de nouvelles données relèvent du domaine de l'apprentissage sur graphes, aussi appelé apprentissage géométrique et tire partie de la géométrie non-euclidienne sous-jacente. Cette thèse se situe dans ce champ de recherche de l'intelligence artificielle utile pour de très nombreuses applications interdisciplinaires, comme en témoignent les récents prix Nobel de physique et de chimie.

L'objectif principal du projet GrapHiX est la proposition de nouveaux modèles d'apprentissage sur graphes dépassant les capacités des modèles existants, du point de vue des performances, tout en gardant des coûts calculatoires maîtrisés, dans une optique de frugalité et de soutenabilité, les algorithmes d'analyse de graphes étant habituellement connus pour leur consommation importante de ressources. Les travaux développés s'appuieront sur les résultats récents du LITIS dans ce domaine, qui adoptent une vision grammaticale de la conception de réseaux de neurones sur graphes. Les résultats théoriques et méthodologiques seront appliqués à des données issues du domaine de la chimie, dans le cadre du projet ANR Octopussy.

English Version

The objects around us are characterized not only by their intrinsic properties but also by the relationships they maintain with other objects. While these relationships are often sequential (for example, a text is a sequence of words, an image is a 2D grid of pixels), they can also take more complex forms that do not follow a regular structure. Notable examples include social interactions within a network, chemical bonds between atoms in a molecule or interactions between amino acids within a protein.

In such cases, graphs provide a relevant mathematical and computational framework for representing these relationships. When the data to be studied is represented as graphs, predicting the properties associated with this data and generating new data fall within the field of graph learning, also known as geometric learning and leveraging the underlying non-euclidean geometry. This thesis lies within this area of artificial intelligence research, which is valuable for numerous interdisciplinary applications, as evidenced by recent Nobel Prizes in Physics and Chemistry.

The primary objective of the GrapHiX project is to propose new graph learning models that surpass the capabilities of existing models in terms of performance, while limiting the

computational costs, having in mind frugality and sustainability, as graph analysis algorithms are usually known for their high resource consumption. The work to be developed will build upon recent results from LITIS in this domain, which adopt a grammatical perspective on the design of graph neural networks. The theoretical and methodological results will be applied to data from the field of chemistry as part of the ANR Octopussy project.

Mots clés liés au projet (5 mots maximum) :

Apprentissage sur graphes

Expressivité des réseaux de neurones sur graphes

Langages algébriques

Décomposition de tenseurs

Chemo-informatique

PRÉSENTATION DU PROJET :

Contexte et Objectifs :

Au cours des dernières années, les réseaux de neurones sur graphes (Graph Neural Networks, GNNs) ont émergé comme une méthode incontournable pour résoudre des tâches d'apprentissage sur des données structurées sous forme de graphes, en tirant parti des relations entre les entités. Pour évaluer le pouvoir expressif de ces GNNs (i.e. leur capacité à distinguer des graphes non isomorphes et donc leur adaptabilité à des données diverses), la hiérarchie de Weisfeiler-Lehman (WL), fondée sur le test d'isomorphisme éponyme [1], s'est imposée comme la méthode de référence [2,3,4,5] en complément des analyses centrées sur leurs capacités spectrales [6].

Un résultat clé dans ce domaine réside dans la démonstration que les réseaux de type *Message Passing Neural Networks (MPNNs)* utilisés dans la plupart des applications réelles des GNNs, sont, au mieux, aussi puissants que le test WL de premier ordre (1-WL) [7][8]. Sur cette base, de nombreuses contributions se sont focalisées sur le dépassement de cette limite 1-WL pour concevoir des GNNs plus expressifs (et donc mieux adaptés à des données plus difficiles). Parmi ces approches, les GNNs basés sur des sous-graphes ont démontré une expressivité supérieure à 1-WL, tout en restant limités par la puissance de 3-WL [9].

En adoptant un point de vue algébrique, l'équipe de F. Geerts a proposé dans [10] de reformuler les tests 1-WL et 3-WL grâce à des langages basés sur des sous-ensembles spécifiques d'opérations algébriques appliquées à la matrice d'adjacence. Ces fragments du langage matriciel MATLANG, nommés ML(L1) et ML(L3), ont été prouvés comme aussi expressifs que 1-WL et 3-WL [10].

S'appuyant sur ces avancées, nous avons proposé au sein de l'équipe Apprentissage du LITIS le modèle GNNML1 [11]. Ce modèle, prouvé comme équivalent à 1-WL, est capable de générer l'ensemble des phrases de ML(L1). Un modèle plus expressif, GNNML3, a également été développé dans cette étude. Bien que ce dernier ait démontré des performances expérimentales comparables à celles des modèles 3-WL existants [12], son expressivité théorique au niveau 3-WL n'a pas été établie dans [11].

Dans le cadre de la thèse de Jason Piquenot qui sera soutenue en 2025, cette recherche a été approfondie et étendue pour développer une procédure de génération de GNNs à partir de la définition d'un langage, en s'appuyant sur des grammaires non-contextuelles. Ce travail, publié à ICLR'24 [13], a conduit à la conception du modèle G2N2, démontré comme atteignant une expressivité équivalente à 3-WL. Ce résultat positionne G2N2 parmi les deux seuls modèles de réseaux de neurones sur graphes actuels à posséder cette expressivité, tout en affichant des performances expérimentales supérieures à l'autre modèle 3-WL existant : PPGN. Les développements théoriques issus de ces travaux ont également permis la proposition du modèle GPN (Grammatical Path Network) [14] , un GNN spécialement conçu pour capturer des cycles de longueur (l+1) au niveau des arêtes en pré-calculant des chemins de longueur (l).

Ces résultats marquent une avancée significative, positionnant l'équipe Apprentissage du LITIS comme un acteur majeur de la communauté scientifique internationale travaillant à la conception de modèles à l'état de l'art en apprentissage sur graphes. Ils ont également mis en lumière certaines limitations dont en particulier l'enjeu central de définir de nouveaux langages permettant de dépasser l'expressivité de 3-WL.

Cette thèse propose d'explorer précisément cet axe, ainsi que ses applications.

[1] AA Lehman and B. Weisfeiler. A reduction of a graph to a canonical form and an algebra arising during this reduction. Nauchno-Technicheskaya Informatsiya, 2(9):12–16, 1968.
[2] C. Morris, M. Ritzert, M. Fey, W. L Hamilton, JE Lenssen, G. Rattan, and M. Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 4602–4609, 2019.
[3] C. Bodnar, F. Frasca, Y. Wang, N. Otter, G. F Montufar, P. Lio, and M. Bronstein. Weisfeiler and lehman go topological: Message passing simplicial networks. In International Conference on Machine Learning, pages 1026–1037. PMLR, 2021.
[4] C. Bodnar, F. Frasca, N. Otter, Y. Wang, P. Lio, G. F Montufar, and M. Bronstein. Weisfeiler and lehman go cellular: Cw networks. Advances in Neural Information Processing Systems, 34:2625–2640, 2021.
[5] B. Zhang, G. Feng, Y. Du, D. He, and L. Wang. A complete expressiveness hierarchy for subgraph gnns via subgraph weisfeilerlehman tests. In International Conference on Machine Learning, 2023.
[6] M. Balcilar, G. Renton, P. Heroux, B. Gauzere, S. Adam, and P. Honeine. Analyzing the expressive power of graph neural networks in a spectral perspective. In International Conference on Learning Representations ICLR, 2021.
[7] C. Morris, M. Ritzert, M. Fey, W. L Hamilton, JE Lenssen, G. Rattan, and M. Grohe. Weisfeiler and leman go neural: Higher-order graph neural networks. In Proceedings of the AAAI conference on artificial intelligence, volume 33, pages 4602–4609, 2019.
[8] K. Xu, W. Hu, J. Leskovec, and S. Jegelka. How powerful are graph neural networks? In International Conference on Learning Representations, 2019.
[9] F. Frasca, B. Bevilacqua, M. M Bronstein, and H. Maron. Understanding and extending subgraph gnns by rethinking their symmetries. In Advances in Neural Information Processing Systems, 2022.
[10] F. Geerts. On the expressive power of linear algebra on graphs. Theory of Computing Systems, Oct 2020.
[11] M. Balcilar, P. Heroux, B. Gauzere, P. Vasseur, S. Adam, and P. Honeine. Breaking the limits of message passing graph neural networks. In Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event, 2021.
[12] H. Maron, H. Ben-Hamu, H. Serviansky, and Y. Lipman. Provably powerful graph networks. Advances in neural information processing systems, 32, 2019.
[13] J. Piquenot, A. Moscatelli, M. Berar, P. Heroux, R. Raveaux, J-Y Ramel, and S. Adam. G2n2: Weisfeiler and lehman go grammatical. In The Twelfth International Conference on Learning Representations ICLR, 2024.
[14] J. Piquenot, L. Bouzidi, M. Berar, P. Heroux, R. Raveaux, J-Y Ramel, and S. Adam. Grammatical path network : You want cycles, paths is all you need. In The third Learning on Graphs Conference LOG, 2024.

Projet détaillé

Pour atteindre une expressivité de GNN allant au-delà de 3-WL, les fragments du langage MATLANG manipulant des matrices [15] ne sont pas suffisants. Une expressivité supérieure nécessite la manipulation de tuples de noeuds et donc d'entrées tensorielles décrivant les graphes de façon plus riche que la simple adjacence entre noeuds. Dans le cadre grammatical décrit ci-dessus, cette contrainte impose naturellement l'introduction d'opérations entre tenseurs au sein du langage.

La première question de recherche qui sera abordée dans le cadre de la thèse concerne la conception de nouveaux fragments de langage tensoriel capables d'atteindre théoriquement une expressivité de 4-WL, et des grammaires associées. Ce travail, extension naturelle des travaux de la thèse de Jason Piquenot, s'appuiera sur les contributions théoriques récentes [12] et [16].

La seconde question de recherche qui sera abordée dans la thèse concerne l'optimisation des architectures découlant des travaux théoriques décrits ci-dessus. En effet, la complexité calculatoire inhérente aux calculs tensoriels requiert des approximations d'opérations tensorielles et/ou des représentations en mémoire compressées pour pouvoir être appliquées à des graphes à large échelle. Leur développement devra donc faire appel aux résultats de la communauté scientifique de la décomposition de tenseurs [17], en les adaptant au contexte d'apprentissage géométrique. Les liens entre de telles stratégies et les opérations de pooling, qui sont encore assez peu explorées dans la communauté scientifique de l'apprentissage sur graphes, seront étudiées dans ce cadre [18]. Ces travaux s'appuieront sur les compétences de l'équipe dans ce domaine [19].

Pour valider les résultats de ces travaux, les expérimentations seront dans un premier temps menées sur le dataset BREC, spécifiquement conçu pour évaluer des architectures expressives au sens de WL. Des expérimentations seront également réalisées sur des données de chimie, dont nous bénéficions dans le cadre du projet ANR OCTOPUSSY qui vise à concevoir et synthétiser de nouveaux polymères issus de matières premières biologiques renouvelables, avec des applications potentielles dans les emballages flexibles et durables. Ce projet est mené avec des collègues du COBRA (https://www.lab-cobra.fr/), de PBS (www.pbs.cnrs.fr/) et du PIMM (https://pimm.artsetmetiers.fr/).

La troisième question de recherche qui sera abordée dans la thèse est celle de la généralisation des contributions proposées dans [20]. L'objectif sera de proposer un modèle de machine learning s'appuyant sur les grammaires proposées dans le premier axe de recherche pour générer des formules tensorielles capables d'extraire des sous-structures telles que des cycles ou des cliques au sein du graphe. À notre connaissance, de telles formules n'existent pas pour des cycles d'une taille supérieure à 8 alors qu'une expressivité 4-WL doit théoriquement permettre d'atteindre cet objectif. De telles formules pourront être exploitées dans des architectures telles que GPN [14].

Finalement, pour mesurer l'apport des architectures expressives proposées dans les deux premières questions de recherche, nous les évaluerons dans un contexte d'apprentissage de métriques entre graphes, à des fins d'explicabilité des modèles. Ce travail s'appuiera sur les résultats récents obtenus dans le cadre de la thèse d'Aldo Moscatelli [21], au travers d'architecture siamoises produisant des embeddings des

graphes à comparer. La production par l'architecture de tenseurs d'ordres supérieurs nécessitera une adaptation des composants de calcul de métriques qui reposent sur le transport optimal. Là encore, les travaux bénéficieront des compétences de l'équipe dans ce domaine [22,23,24].

Les différentes questions scientifiques mentionnées ci-dessous ont toutes en commun de reposer sur des problèmes d'optimisation non-triviaux dans des espaces non-euclidiens (que ce soit des espaces de graphes ou des espaces de matrices/tenseurs devant respecter des contraintes ad-hoc). Les outils d'optimisation sur des espaces non-euclidiens font partie de l'arsenal scientifique de l'équipe qui a pu les mettre en œuvre dans divers contextes [25, 26, 27, 28].

[15] R. Brijder, F. Geerts, JV den Bussche, and T. Weerwag. On the expressive power of query languages for matrices. ACM Trans. Database Syst., 44(4):15:1–15:31, 2019.
[16] F. Geerts and JL. Reutter. Expressiveness and approximation properties of graph neural networks. In The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022. OpenReview.net, 2022.
[17] M. Wang, Y. Pan, Z. Xu, X. Yang, G. Li, and A. Cichocki. Tensor networks meet neural networks: A survey and future perspectives. 2023.
[18] C. Hua, G. Rabusseau, and J. Tang. High-order pooling for graph neural networks with tensor decomposition. Advances in Neural Information Processing Systems, 35:6021–6033, 2022.
[19] A. Traore, M. Berar, and A. Rakotomamonjy. Singleshot : a scalable tucker tensor decomposition. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019.
[20] J. Piquenot, M. Berar, P. Heroux, R. Raveaux, JY Ramel, and S. Adam. Grammar reinforcement learning: path and cycle counting in graphs with a context-free grammar and transformer approach. In The thirteenth International Conference on Learning Representations ICLR, 2025 (submission : https://openreview.net/forum?id=yEox25xAED, decision 22/01/2025).
[21] A. Moscatelli, J. Piquenot, M. Berar, P. Heroux, and S. Adam. Graph node matching for edit distance. Pattern Recognition Letters, 184:14–20, 2024.
[22] M. Z. Alaya, M. Berar, G. Gasso, and A. Rakotomamonjy. Gaussian-smoothed sliced probability divergences. Transactions on Machine Learning Research, 2024.
[23] M. Z. Alaya, M. Berar, G. Gasso, and A. Rakotomamonjy. Screening sinkhorn algorithm for regularized optimal transport. In Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada, 2019.
[24] A. Rakotomamonjy, R. Flamary, G. Gasso, M. Z. Alaya, M. Berar, and N. Courty. Optimal transport for conditional domain matching and label shift. Machine Learning, 2022.
[25] F. Yger, M. Berar, G. Gasso and A. Rakotomamonjy. Adaptive canonical correlation analysis based on matrix manifolds. ICML 2012
[26] I. Horev, F. Yger, and M. Sugiyama. "Geometry-aware principal component analysis for symmetric positive definite matrices." ACML, 2016.
[27] F. Yger, S. Chevallier, Q. Barthélemy and S. Sra. Geodesically-convex optimization for averaging partially observed covariance matrices. ACML 2022
[28] T. De Surrel, S. Chevallier, F. Lotte, F. Yger. Geometry-Aware visualization of high dimensional Symmetric Positive Definite matrices. TMRL (accepted for publication) 2025

Principales actions et calendrier détaillés de mise en œuvre :

T0 -> T0+9

Après une analyse approfondie de la littérature liée à la fois à l'expressivité des réseaux de neurones sur graphes, aux grammaires algébriques et aux décompositions tensorielles, la première étape de la thèse consistera à proposer une première contribution qui réponde à la première question de recherche mentionnée ci-avant. Le travail consistera à proposer

une grammaire tensorielle permettant d'atteindre une expressivité de 4WL. Il s'agit ici d'une contribution théorique à faible risque car le travail a déjà été largement abordé dans le cadre de la thèse de Jason Piquenot. Le doctorant pourra bénéficier de ces travaux. Le travail d'analyse de la littérature sera également facilité par l'existence des manuscrits de thèse de Jason Piquenot et d'Aldo Moscatelli, tous les deux en fin de thèse sur des sujets très proches.

T0+9 -> T0+18

Dans un second temps, nous nous appuierons sur les travaux menés lors de la première étape pour proposer l'implémentation d'un GNN 4WL opérationnel, inexistant à ce jour. Un travail conséquent et minutieux d'implémentation devra être réalisé, avec un focus particulier sur la complexité calculatoire des modèles développés. Ce travail s'appuiera sur l'utilisation de techniques de décomposition de tenseur pour lesquelles des compétences sont présentes au sein de l'équipe d'encadrement. Un travail important d'expérimentation devra être mené, sur des bases de données de la littérature, mais aussi sur des données issues du projet ANR OCTOPUSSY.

T0+18 -> T0+24

À l'issue de ces travaux, une troisième étape consistera à proposer un modèle de machine learning de type LLM à base de transformers et d'apprentissage par renforcement pour générer des formules tensorielles capables d'extraire des sous-structures telles que des cycles ou des cliques au sein du graphe. Cette étape s'appuiera sur les grammaires proposées dans la première étape de recherche et sur les travaux décrits dans [20].

T0+24 -> T0+30

Enfin, la quatrième étape avant la rédaction de la thèse consistera à combiner les travaux développés dans l'étape 2 avec les résultats de la thèse d'Aldo Moscatelli [21] pour proposer un modèle de calcul de métriques entre graphes exploitant les représentations expressives produites par le nouveau modèle.

T0+30 -> T0+36 : rédaction du manuscrit

Moyens humains, matériels, financiers … mis en œuvre et demandés pour atteindre les objectifs :

Pour atteindre les objectifs mentionnés, le doctorant pourra s'appuyer sur les compétences de l'équipe d'encadrement de la thèse, composée de Sébastien Adam (PR LITIS-APP), Maxime Bérar (MCF LITIS-APP) et Florian Yger (MCF LITIS-APP) qui bénéficient d'une expérience importante dans les domaines des réseaux de neurones sur graphes, de la décomposition tensorielle et de l'optimisation [6,11,13,14,19,20,21,22,23,24,25,26,27,28].

Les travaux seront appliqués sur des bases de données de référence à des fins de comparaison avec les modèles de la littérature, et sur des bases liées à des problématiques d'analyse de molécules chimiques, dans le cadre du projet ANR OCTOPUSSY. Cette partie applicative du travail bénéficiera en outre des interactions et des échanges avec un autre doctorant débutant sa thèse en septembre 2025 sur les réseaux de neurones sur graphes dynamiques et du renfort d'un post-doc recruté sur le

projet ANR.

Pour ses expériences numériques, le doctorant bénéficiera des infrastructures de calcul disponibles au LITIS pour tester ses développements. Il aura également, pour les tests à plus large échelle, accès aux super-calculateurs du CRIANN et à l'infrastructure Jean Zay.

Procédure de candidature :

Les candidats intéressés par ce sujet doivent envoyer à Sebastien.Adam@univ-rouen.fr, Maxime.Berar@univ-rouen.fr, Florian.Yger@insa-rouen.fr :

- CV
- Lettre de motivations
- Relevés de notes du Master (M1+M2) ou des deux dernières années d'école d'ingénieur.

La date limite de candidature est fixée au 30/04/2025