Soutenance du projet de fin d'études

BIOINFORMATIQUE ET MODELISATION DES SYSTEMES COMPLEXES LIEE A LA SANTE

Construction d'un graphe de connaissances biomédicales: Cas de COVID-19

Préparé par :
SNOUSSI Youssef

Encadré par :
Pr. ABIK Mounia

Soutenu le 29 septembre 2021 devant le jury :

- Président *Pr. TABII Youness* ENSIAS
- Examinateur *Pr. EZZAHOUT Abderrahmane* FSR
- Encadrant *Pr. ABIK Mounia* ENSIAS
- Co-Encadrant *Mr. HAJHOUJ Mohammed* ENSIAS

Plan

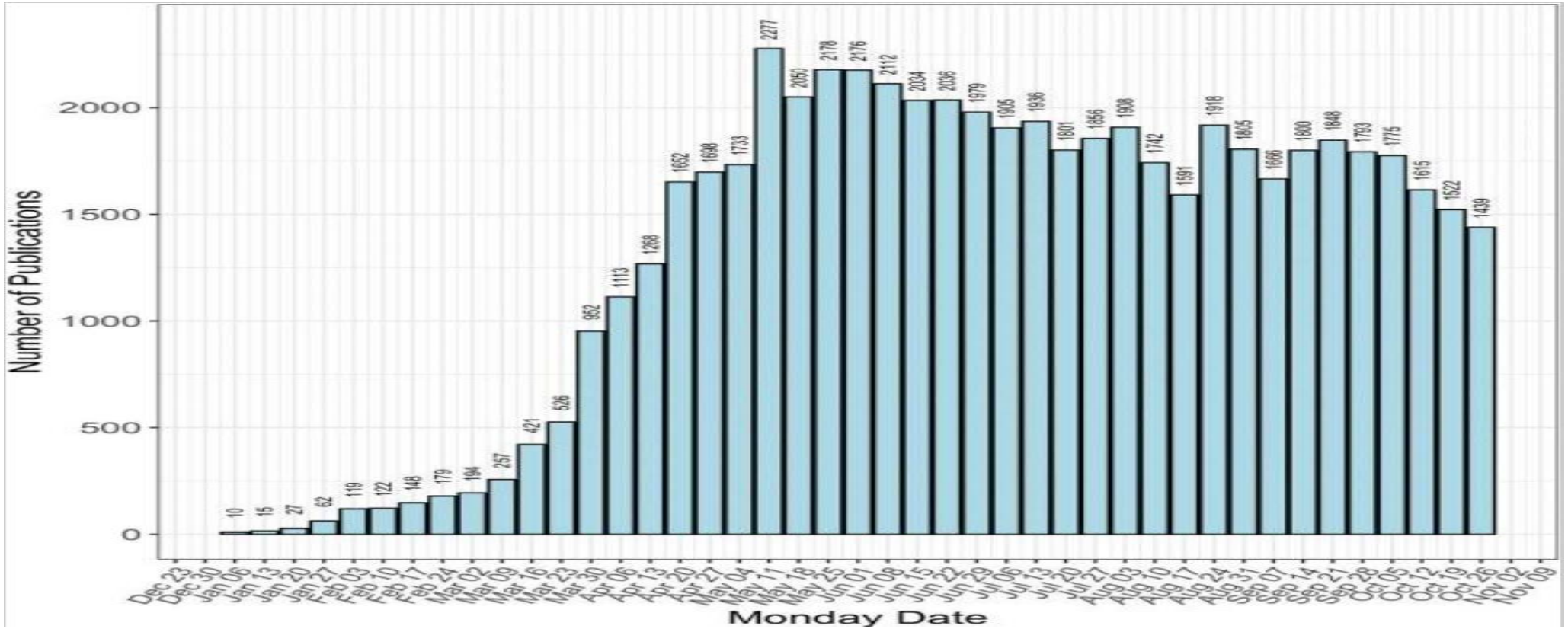
I. INTRODUCTION

II. ÉTAT DE L'ART

III. RÉALISATION ET RÉSULTATS

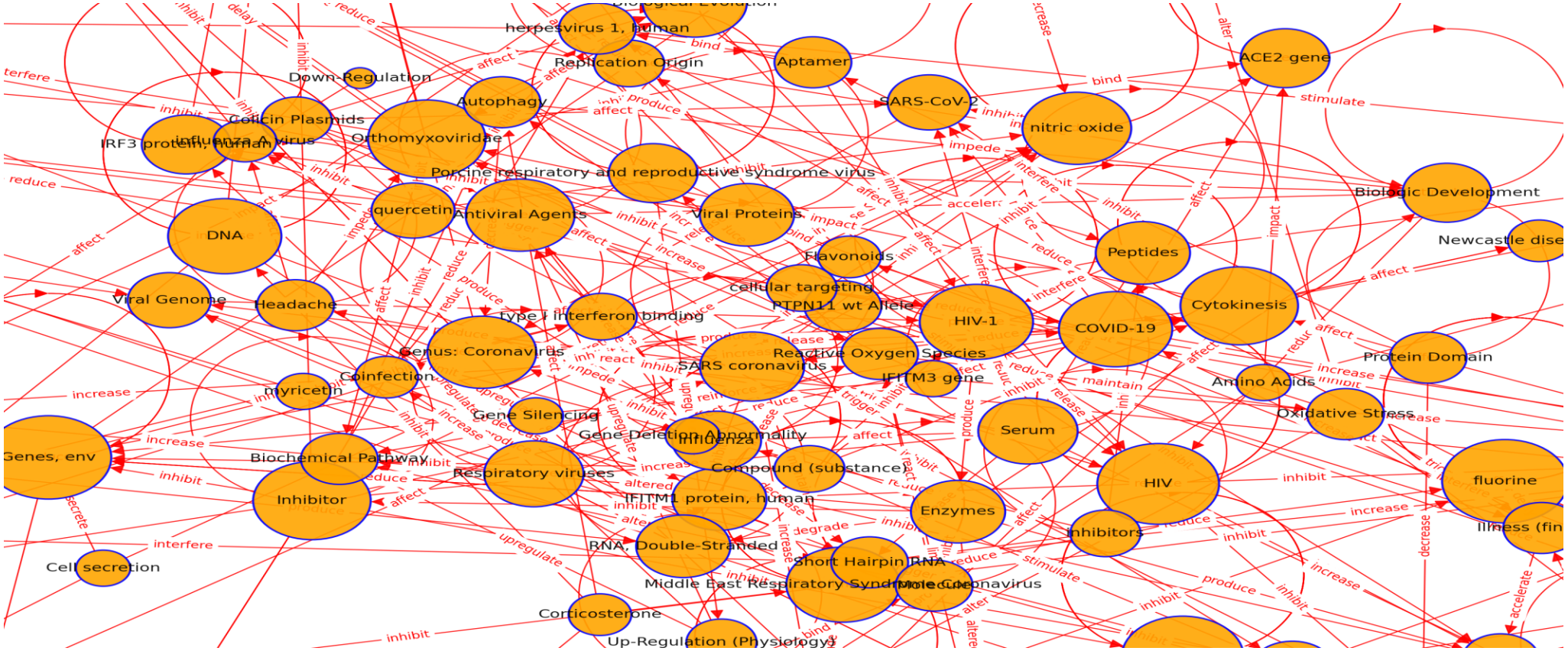
IV. CONCLUSION ET PERSPECTIVES

INTRODUCTION



Nombre de publications par semaine dans PubMed

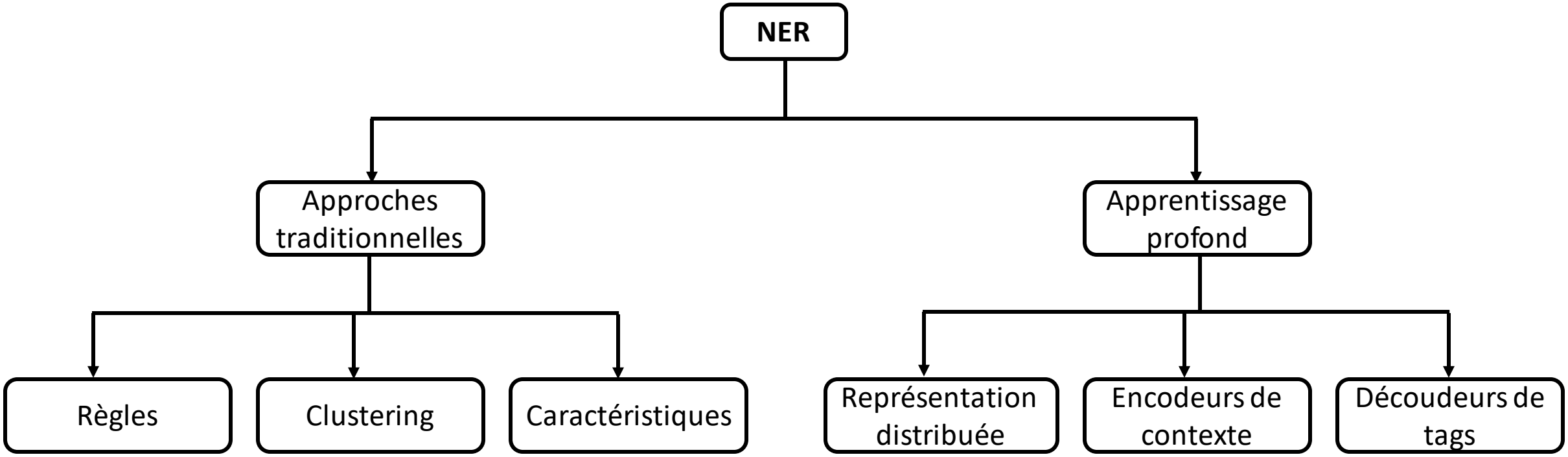
INTRODUCTION



Exemple d'un graphe de connaissances biomédicales

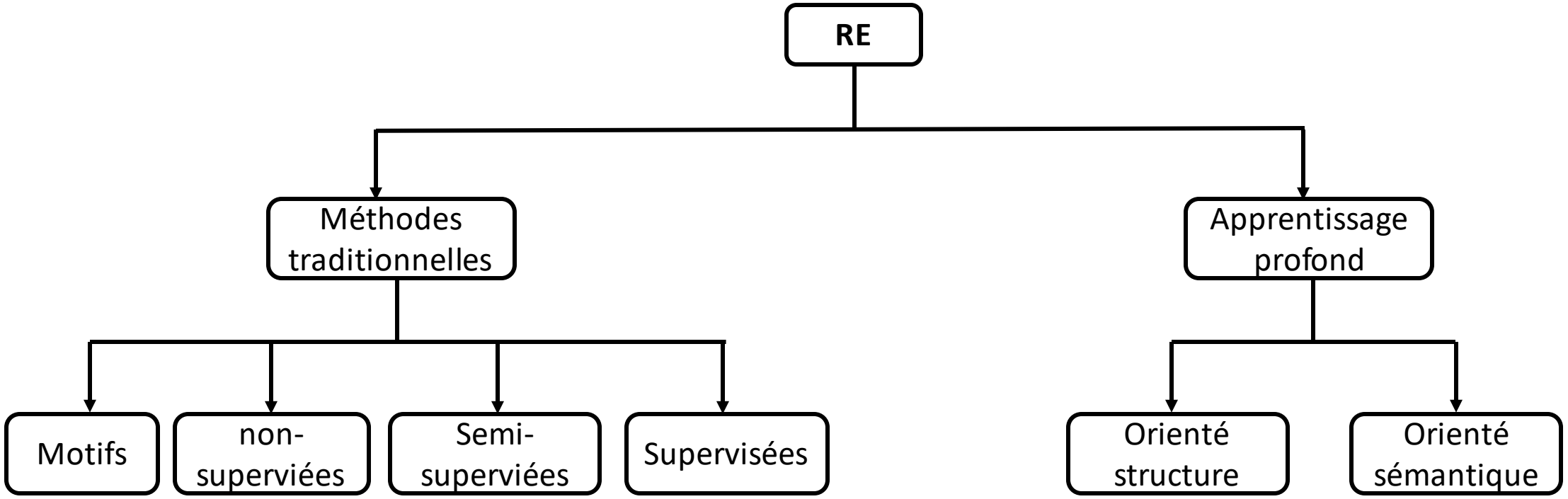
ÉTAT DE L'ART

RECONNAISSANCE D'ENTITÉS NOMMÉES (NER)



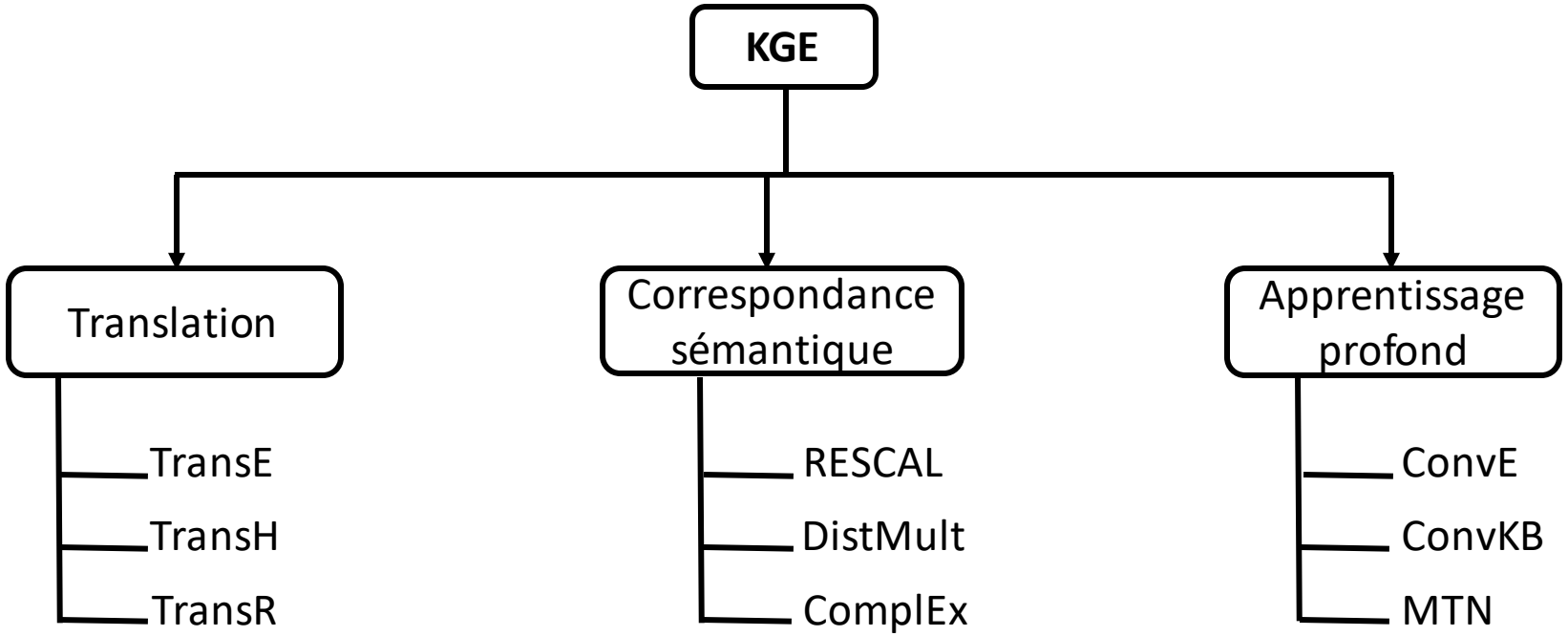
ÉTAT DE L'ART

EXTRACTION DES RELATIONS (RE)



ÉTAT DE L'ART

INTÉGRATION DES GRAPHES DE CONNAISSANCES (KGE)



RÉALISATION ET RÉSULTATS

PRÉPARATION DES DONNÉES

- Nous avons utilisé la collection des publications relatives à COVID-19 à partir de la base de données CORD19 qui disponible sur Kaggle.
- Nous avons pris 400000 publications en texte intégrale, et nous n'avons considérer que les publications anglaises.
- Nous avons pu extraire 6.5 millions de phrases
- Vu à la nature des modèles utilisés, nous n'avons pas modifié le texte pour améliorer l'analyse sémantique.

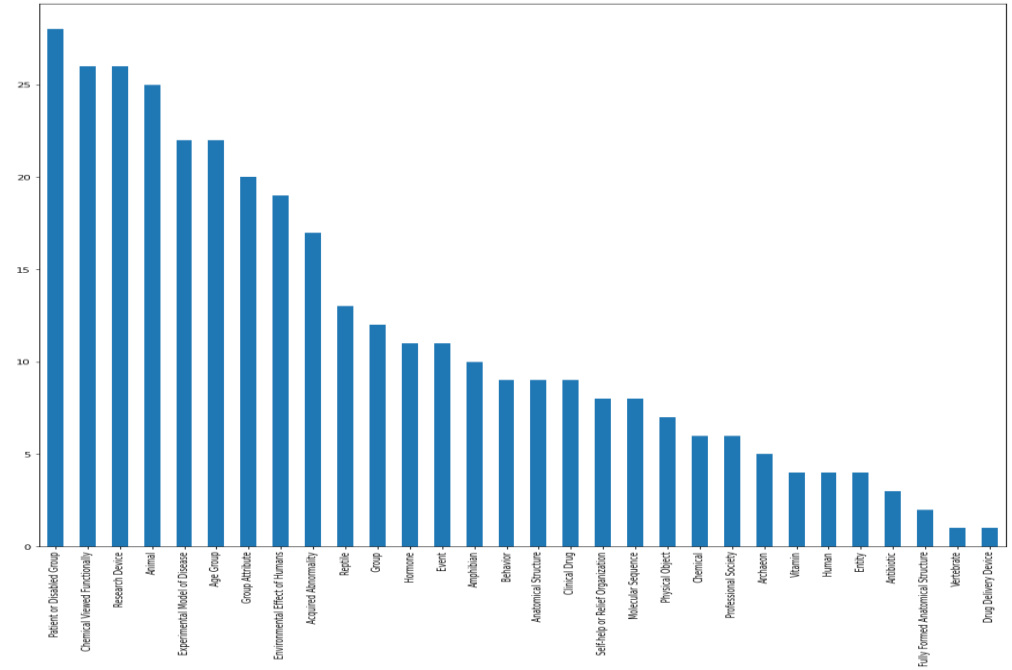
RÉALISATION ET RÉSULTATS

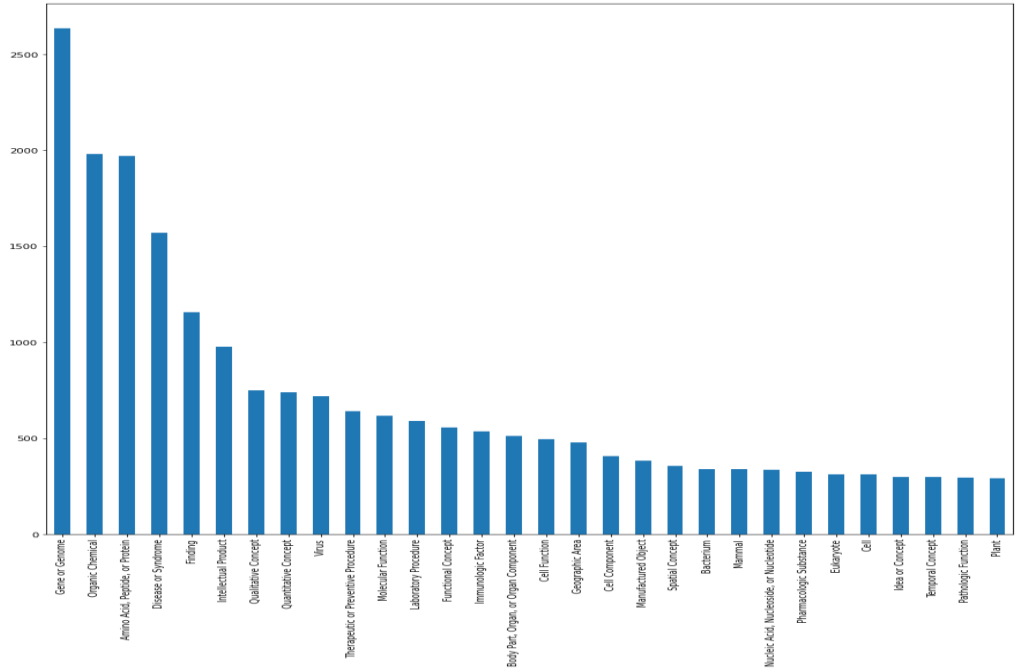
RECONNAISSANCE D'ENTITÉS NOMMÉES

- Nous avons utilisé le modèle pré-entraîné SCI-BERT qui est disponible dans la bibliothèque scispaCy.
- Nous avons pris seulement les entités ayant un score de liaison avec l'UMLS supérieur à 0.85.
- Nous avons pu extraire 29609 entités nommées tout en liant chaque entité avec l'UMLS par un identifiant unique de concept.
- En utilisant le REST API de UMLS, nous avons récupéré le type sémantique de chaque entité.

RÉALISATION ET RÉSULTATS

RECONNAISSANCE D'ENTITÉS NOMMÉES





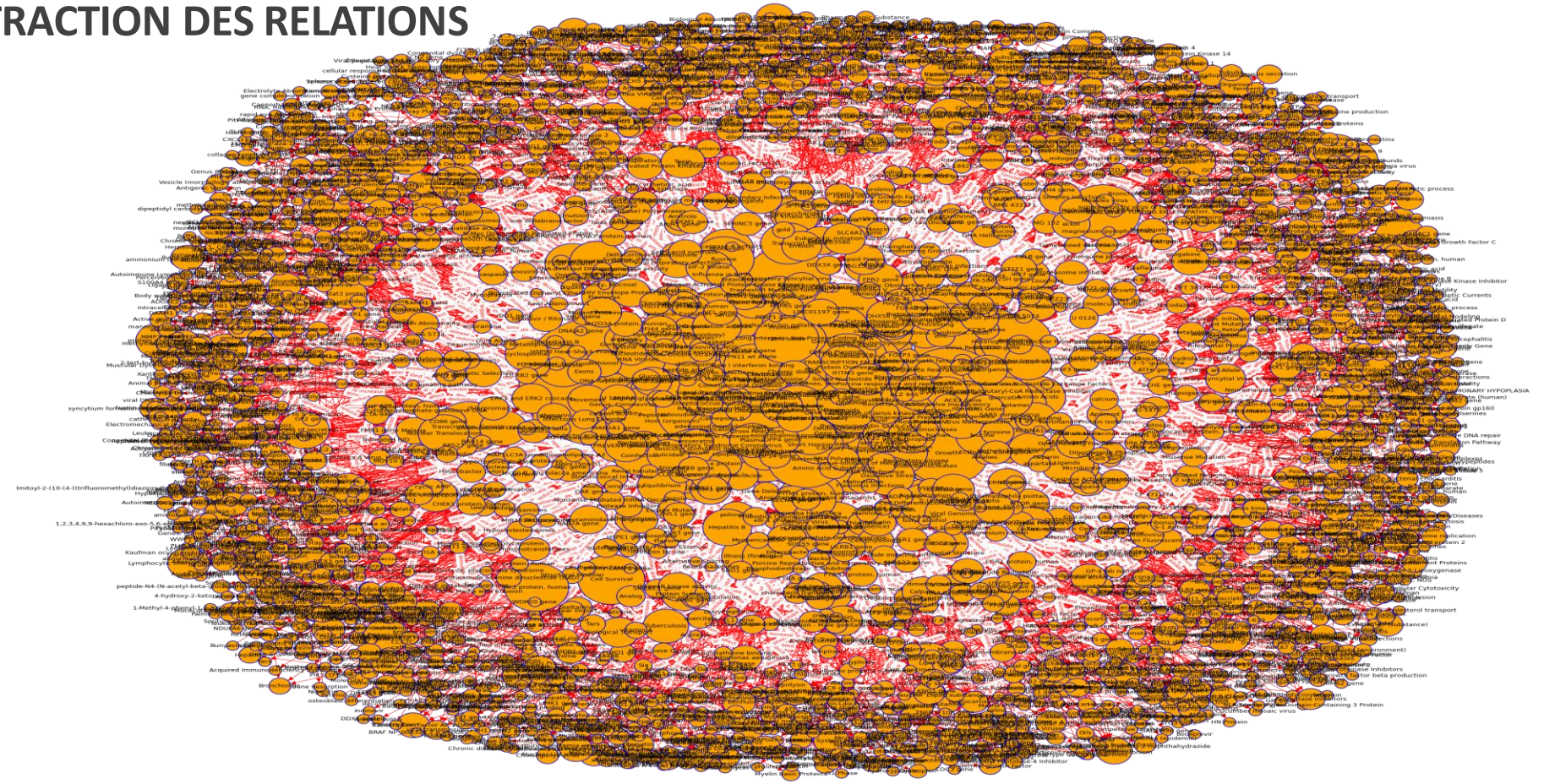
RÉALISATION ET RÉSULTATS

EXTRACTION DES RELATIONS

- Nous avons utilisé le modèle de langage BERT pour étiquetage des rôles sémantiques, et par la suite extraire les relations sous forme de verbes entre entités.
- A cause des contraintes de puissance de calcul requise, nous avons pu seulement traiter 25% des phrases obtenues dans la phase de préparation de données.
- Nous avons pu extraire 413064 triplets.
- Les relations obtenues nécessitent encore du travail puisque nous avons obtenu plus de 4000 relations différentes.

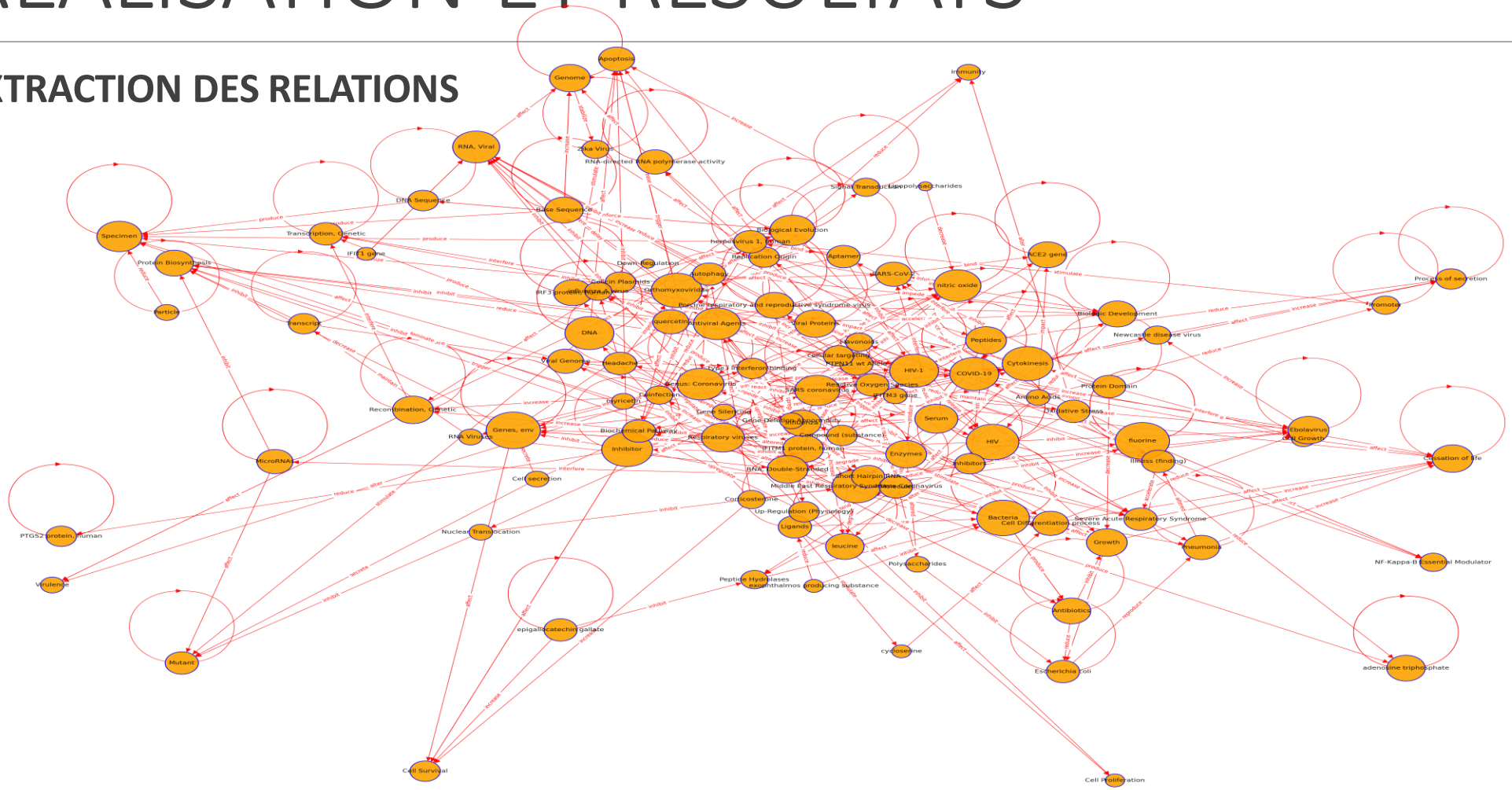
RÉALISATION ET RÉSULTATS

EXTRACTION DES RELATIONS



RÉALISATION ET RÉSULTATS

EXTRACTION DES RELATIONS



RÉALISATION ET RÉSULTATS

INTÉGRATION DES GRAPHES DE CONNAISSANCES

- Nous avons essayé plusieurs modèles, TransE, TransH, TransR, ComplEx et ConvE.
- A cause des contraintes de puissance de calcul requise, nous avons seulement considéré un sous-graphe contenant les maladies et les substances chimiques (3000 entités et 40000 triplets).
- Nous avons conduit l'entraînement sur 80%, 10% pour le test et 10% pour la validation.
- Les métriques d'évaluations utilisés sont basées sur les ranks.

RÉALISATION ET RÉSULTATS

INTÉGRATION DES GRAPHES DE CONNAISSANCES

	MR	MRR	AMRI	HITS@10
TransE	381.40	0.1851	0.7464	0.2440
TransH	607.84	0.0707	0.5955	0.1538
TransR	1063.33	0.0728	0.2920	0.1239
ComplEx	1301.14	0.0322	0.1335	0.0476
ConvE	**178.59**	**0.2542**	**0.8816**	**0.4265**

RÉALISATION ET RÉSULTATS

INTÉGRATION DES GRAPHES DE CONNAISSANCES

Sujet	relation	objet
nicotine	Inconnue	Acute gastroenteritis
Hydroxychloroquine	Inconnue	Influenza A Virus, H1N1
Fatty Liver	Host	Daclatasvir
Asthma	require	Aarskog syndrome

CONCLUSION ET PERSPECTIVES

- Le but de ce travail était de construire un graphe de connaissances biomédicales à partir de la littérature de COVID-19 et l'utiliser par la suite pour la pévision de vouveaux liens.
- La construction de notre graphe a été conduite en utilisant les variantions de BERT pour la reconnaissance d'entités nommées et l'extraction de relations.
- Pour la privision des nouveaux liens, nous avons utilisé plusieurs modèles d'incorporation des graphes de connaissances.
- Les résultats on été prometteurs puisque nous avons obtenu des nouveaux liens signifiants, et peut être amélioré par l'améliorations des relations.

RÉFÉRENCES

Kang M, Gurbani SS, Kempker JA. The Published Scientific Literature on COVID-19: An Analysis of PubMed Abstracts. *J Med Syst*. 2020;45(1):3. Published 2020 Nov 25. doi:10.1007/s10916-020-01678-4

Jing Li, Aixin Sun, Jianglei Han, Chenliang Li,A Survey on Deep Learning for Named Entity Recognition

Pawar, S., Palshikar, G.K., & Bhattacharyya, P. (2017). Relation Extraction : A Survey. *ArXiv, abs/1712.05191*.

Shivani Choudhary, Tarun Luthra, Ashima Mittal, Rajat Singh, A Survey of Knowledge Graph Embedding and Their Applications