

# Méthodes Big Data pour l'Analyse de Très Grands Graphes de Protéines

Projet PLDAC

17 mai 2024

---

**Anyes TAFOUGHALT**  
**Racha Nadine DJEGHALI**

- Encadré par : Hubert NAACKE

En collaboration avec le MNHN

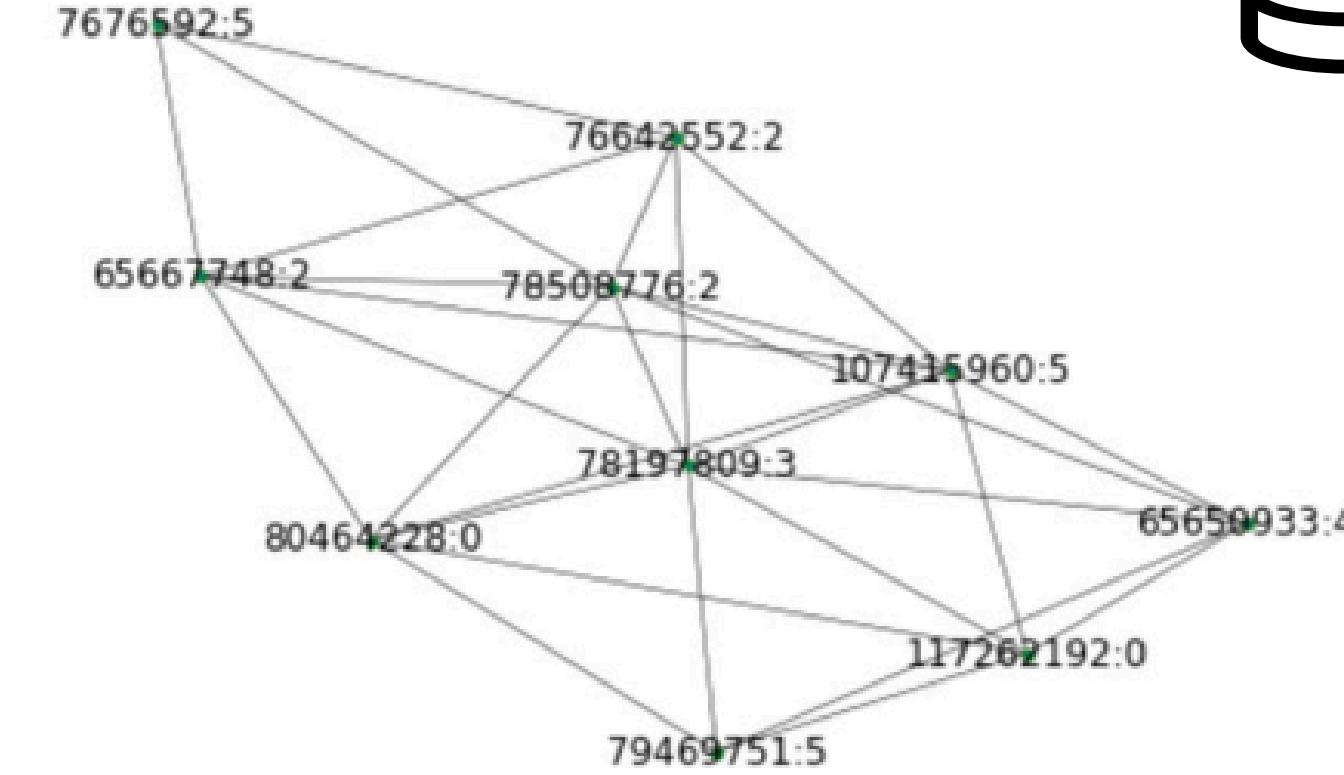
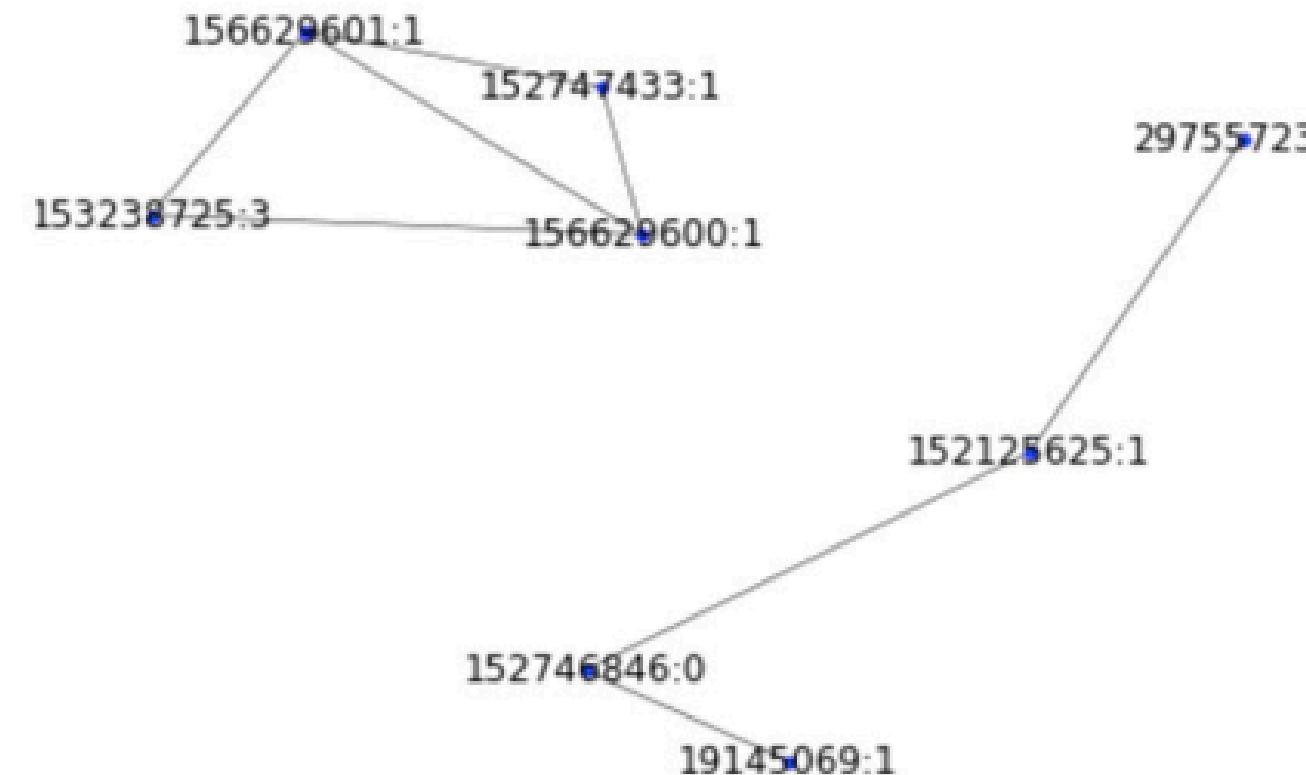
# Plan

---

- 1 Contexte du Problème
- 2 Détection de composantes
- 3 Prédiction d'annotations

# Données manipulées

- Graphe de similarité protéique :
  - Un nœud = identifiant de protéine.
  - Un arc = poids de similarité (80% à 100%).
- Annotations fonctionnelles pour certaines protéines.
- Échelle massive : plus de 20 milliards d'arcs.



# Objectifs

---

1

Déetecter les composantes d'un graphe volumineux

2

Prédire les labels manquants

# Détection de composantes

---

# Solutions existantes

---

## 1) Petite échelle:

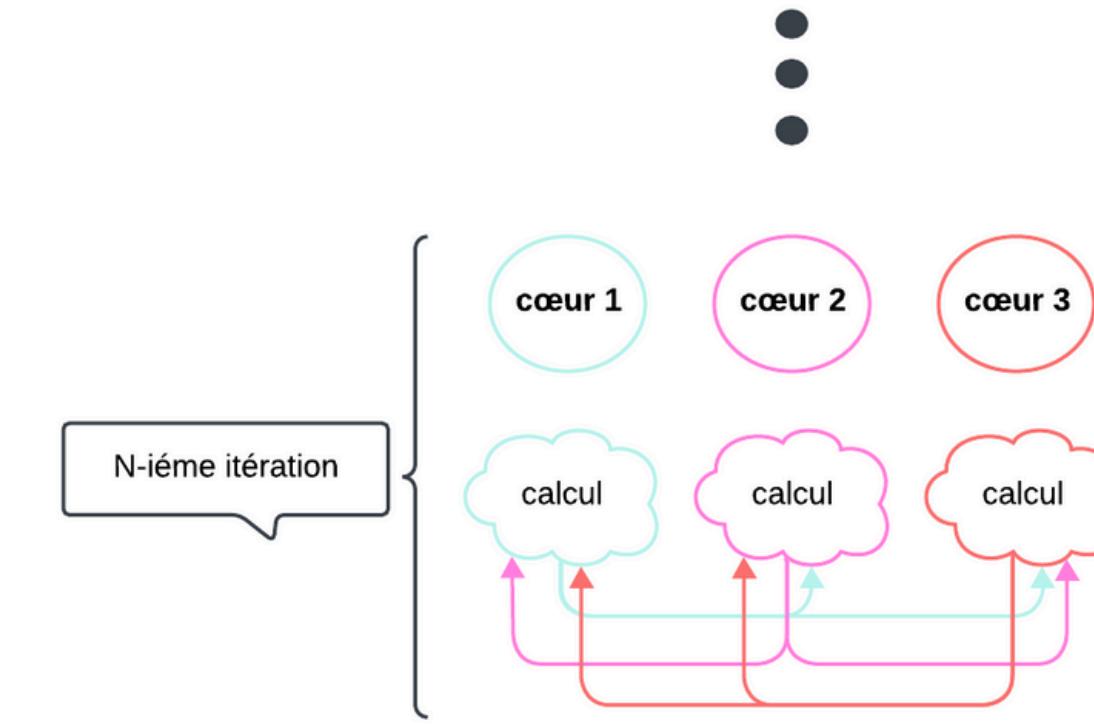
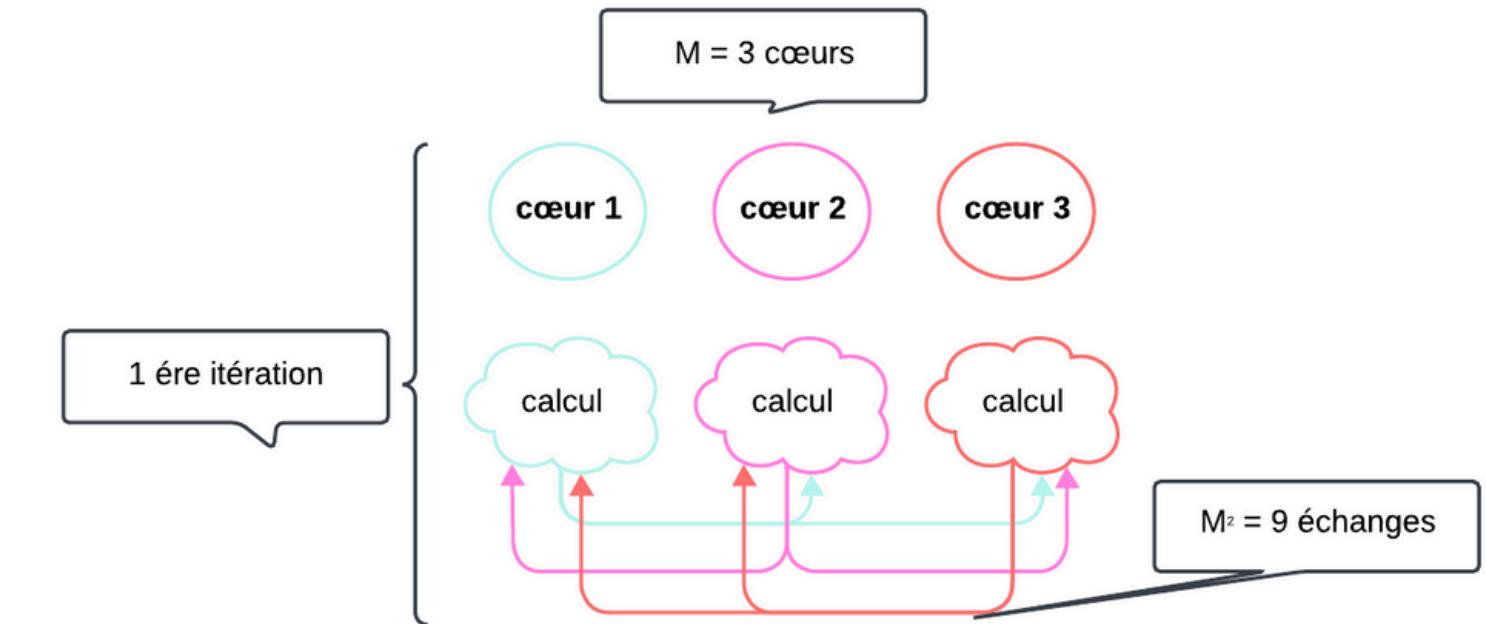
- Solution centralisée.
- Calcul des composantes sur un seul cœur.
- Limitée aux graphes qui tiennent en mémoire.



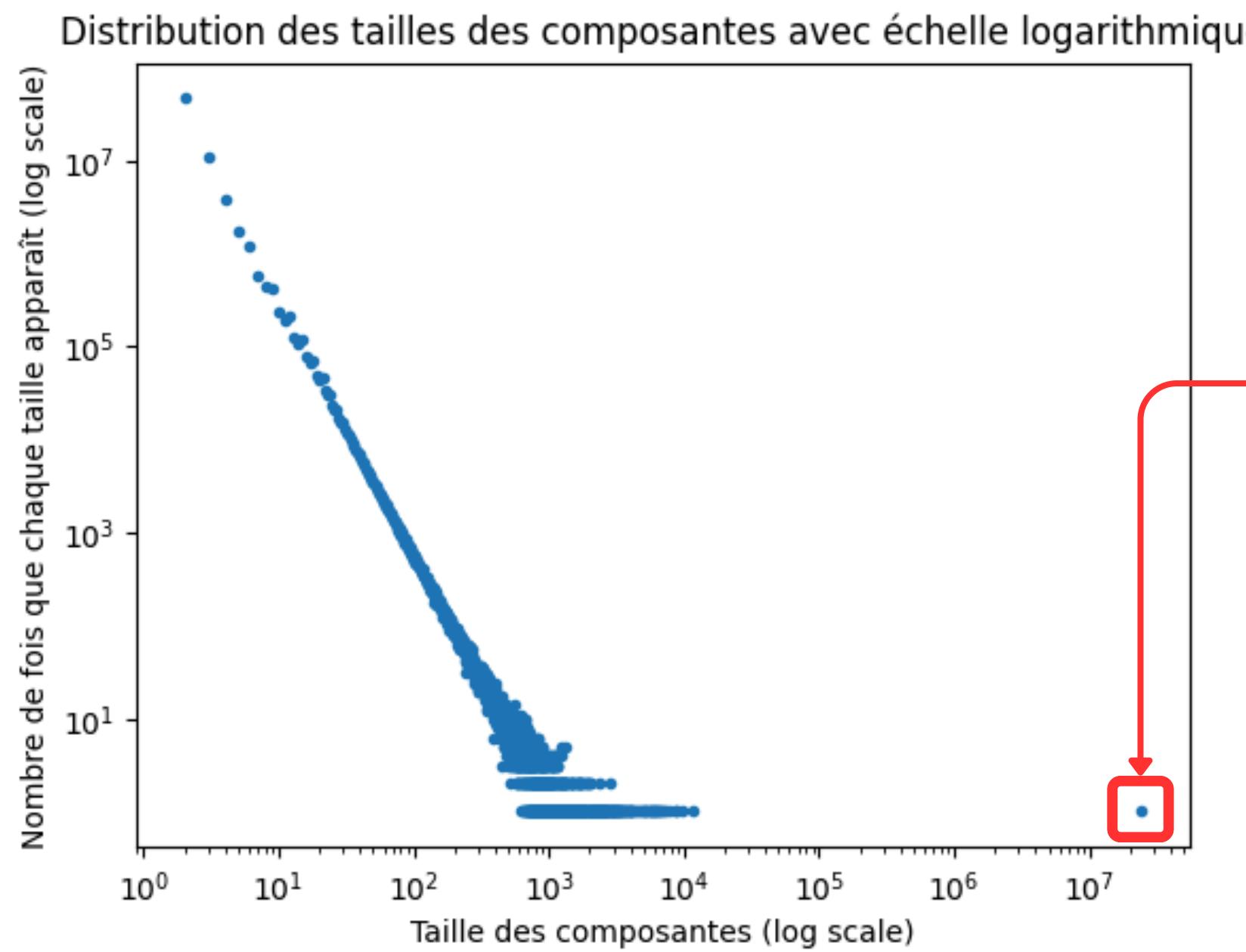
# Solutions existantes

## 2) Grande échelle:

- Propagation de labels itérative.
- Échange de données entre les cœurs à chaque itération.



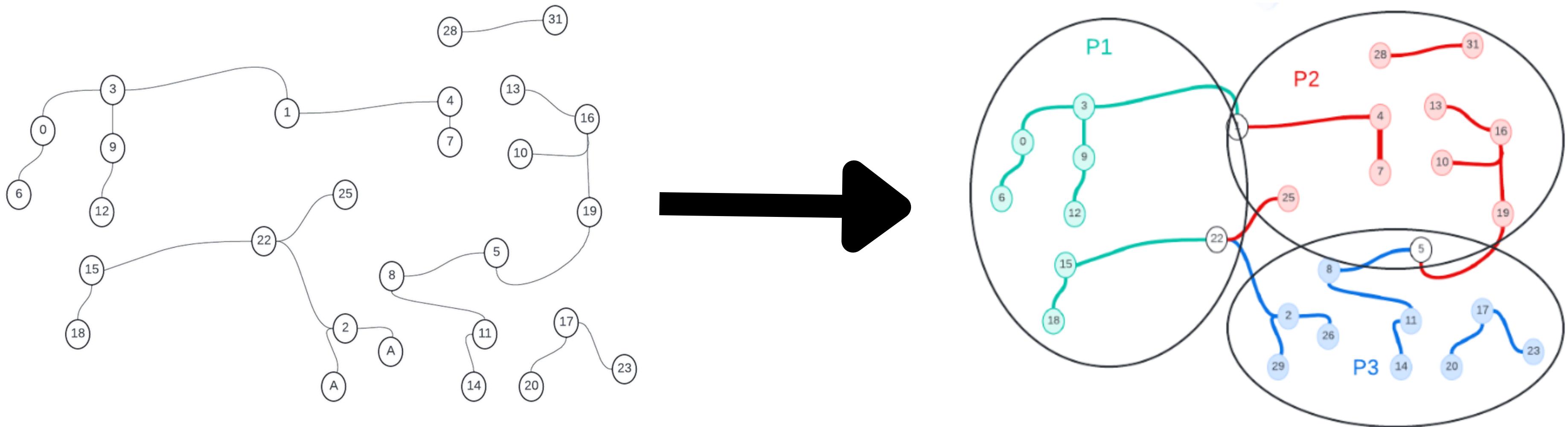
# Défi des composantes volumineuse



La plus grande  
composante a une taille  
d'environ 24 millions  
(24 479 543).

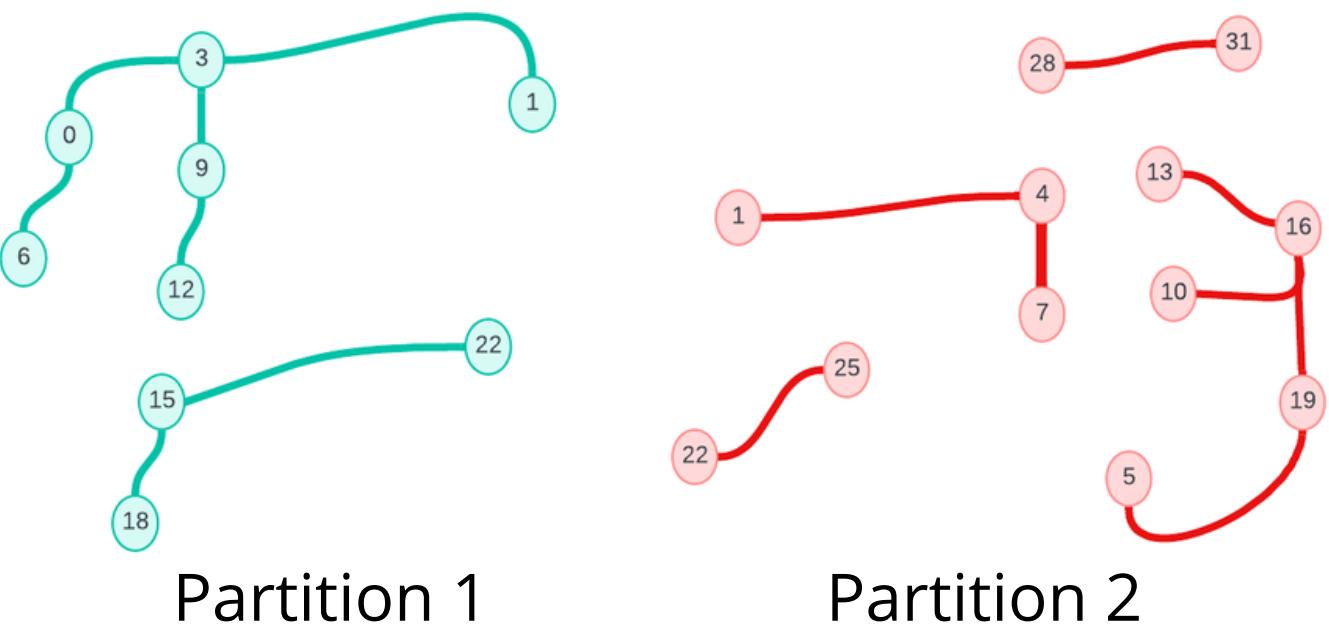
# Méthode proposée

1) Partitionnement du graphe G.

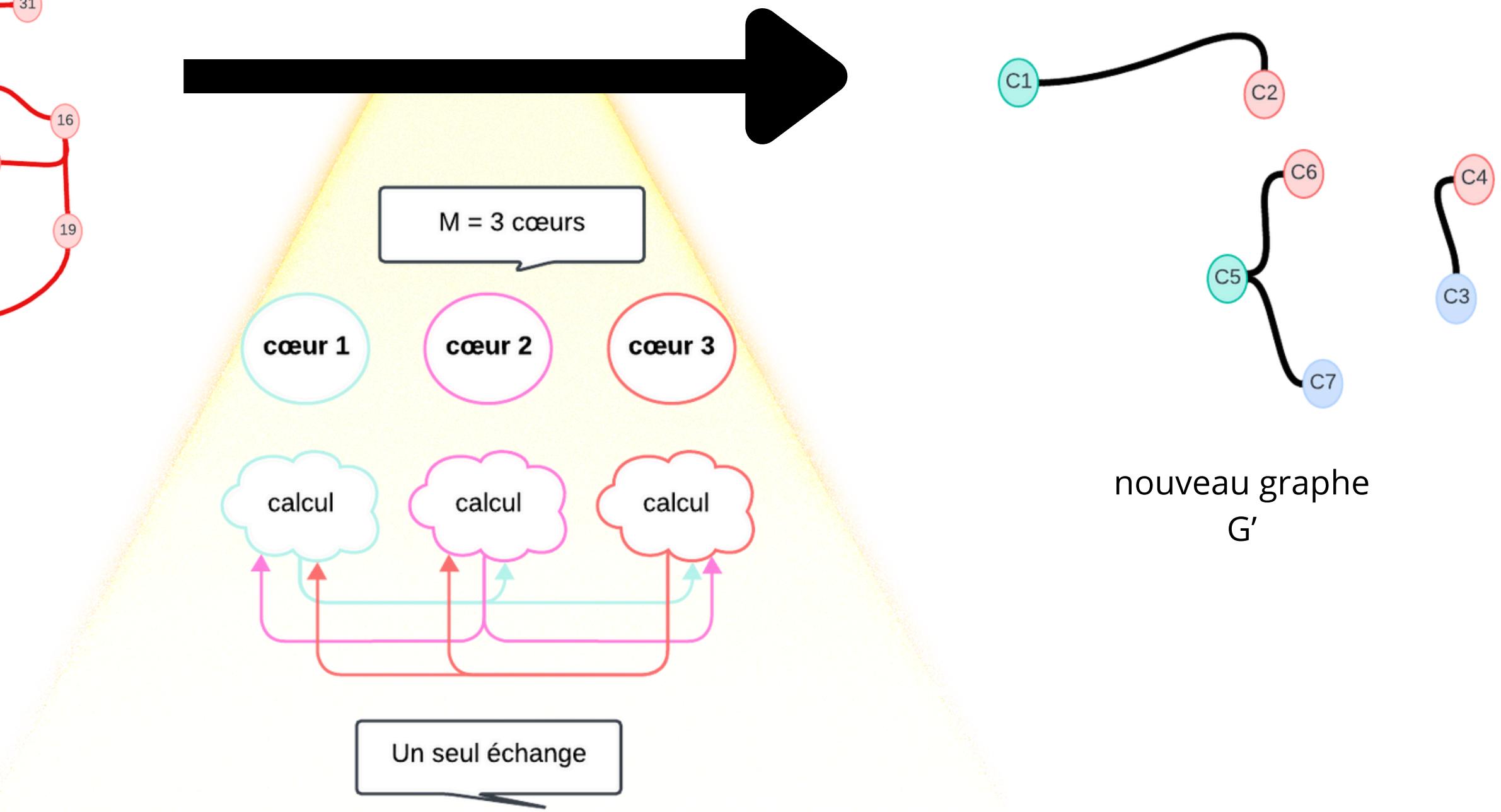


# Méthode proposée

2) Calcul des composantes partielles CP.

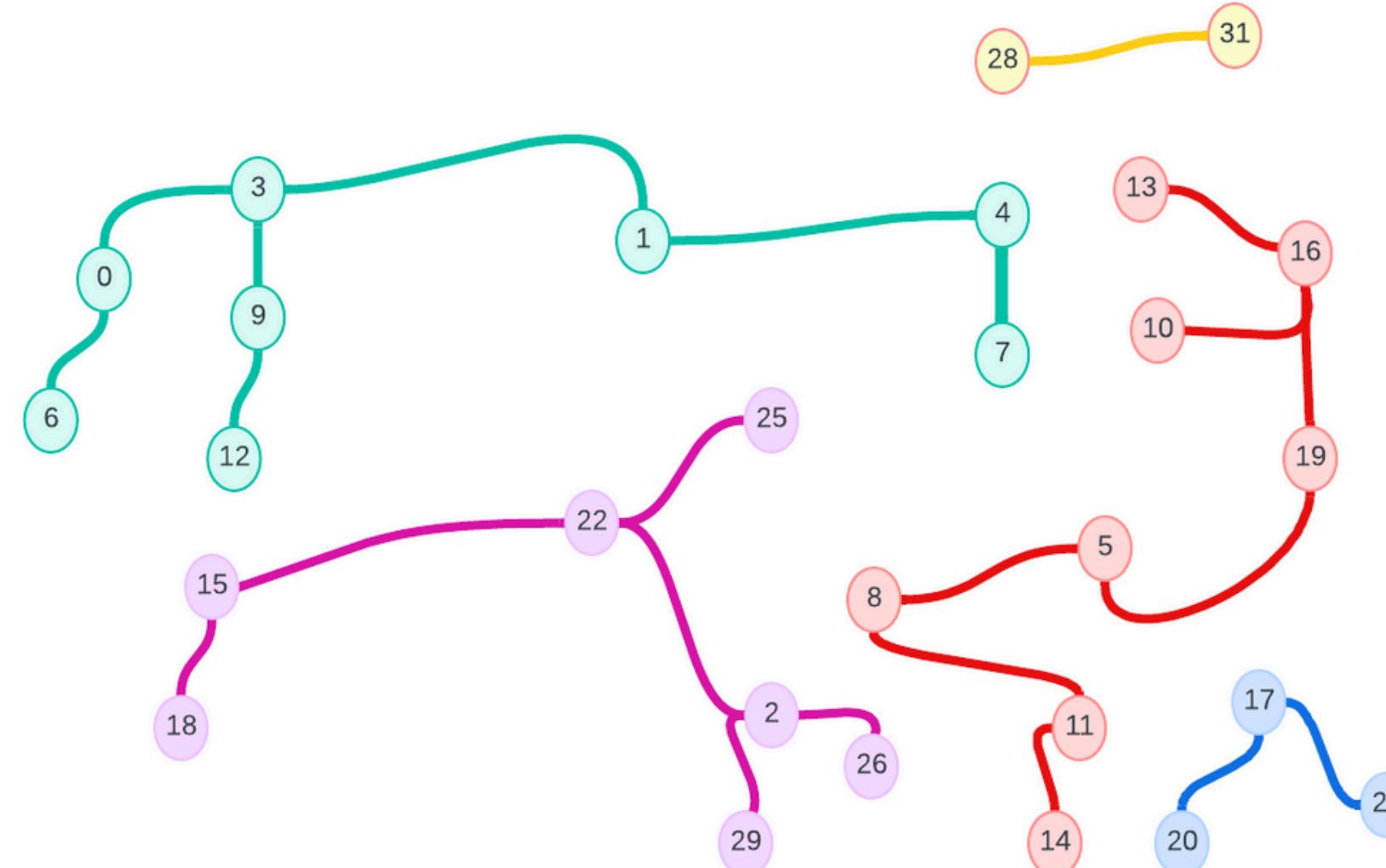


3) Crédit du nouveau graphe  $G'$ .



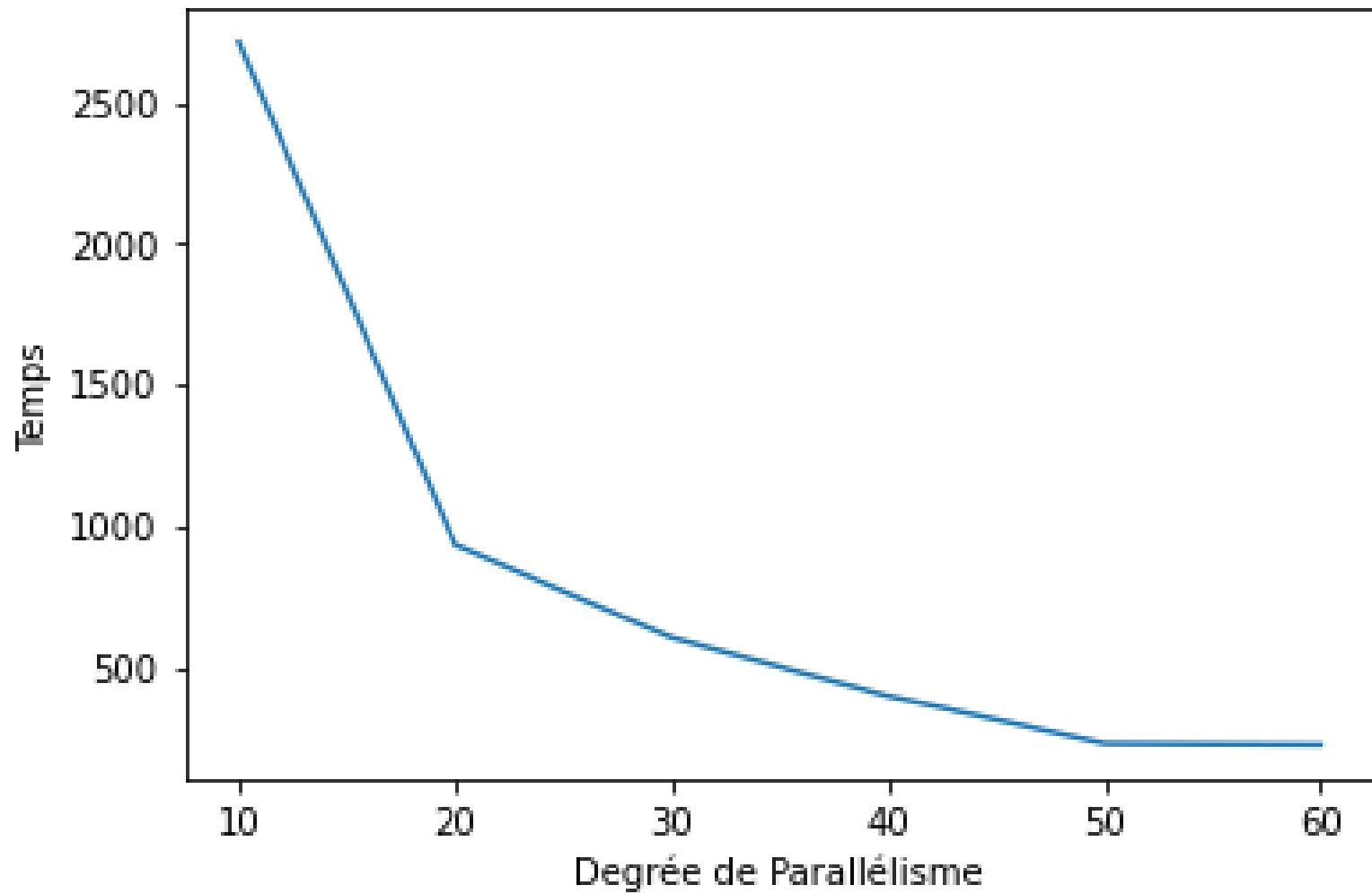
# Méthode proposée

- 4) Calcul des composantes C du graphe  $G'$ .
- 5) Associer les nœuds N du graphe G à leurs composantes réelles.

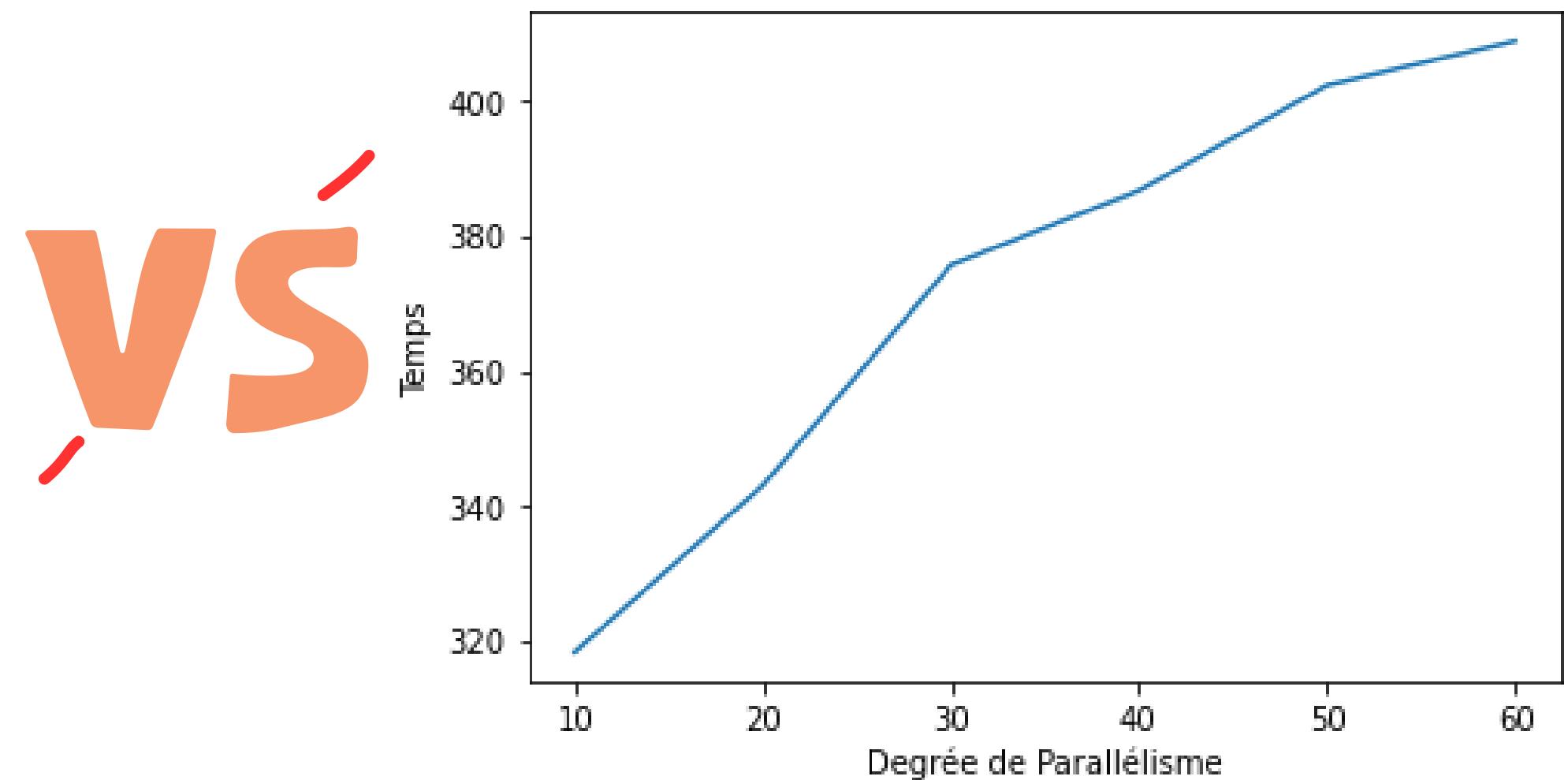


# Quel degré de parallélisme choisir?

- Temps de calcul des composantes partielles



- Temps de calcul des composantes de  $G'$



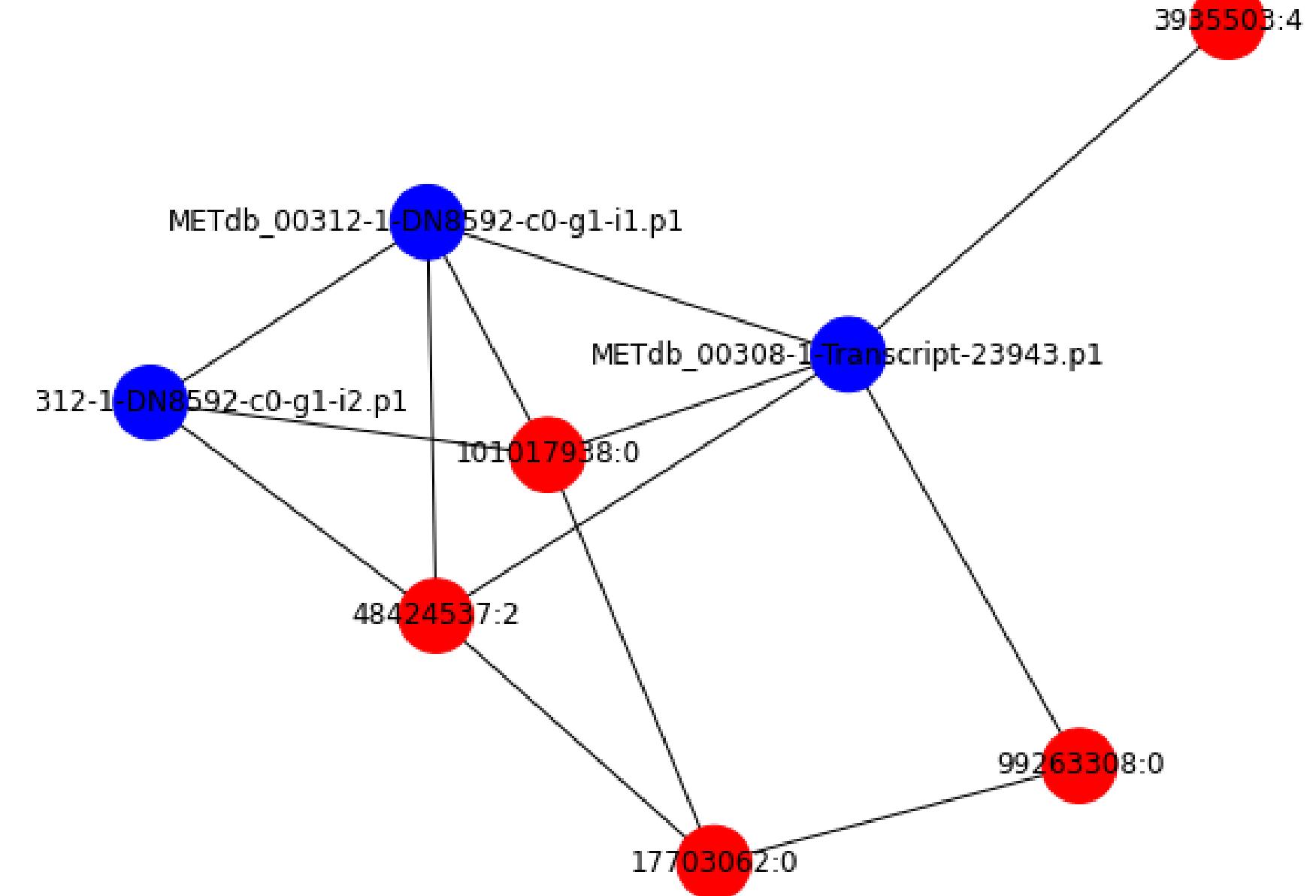
VS

# Prédiction d'annotations

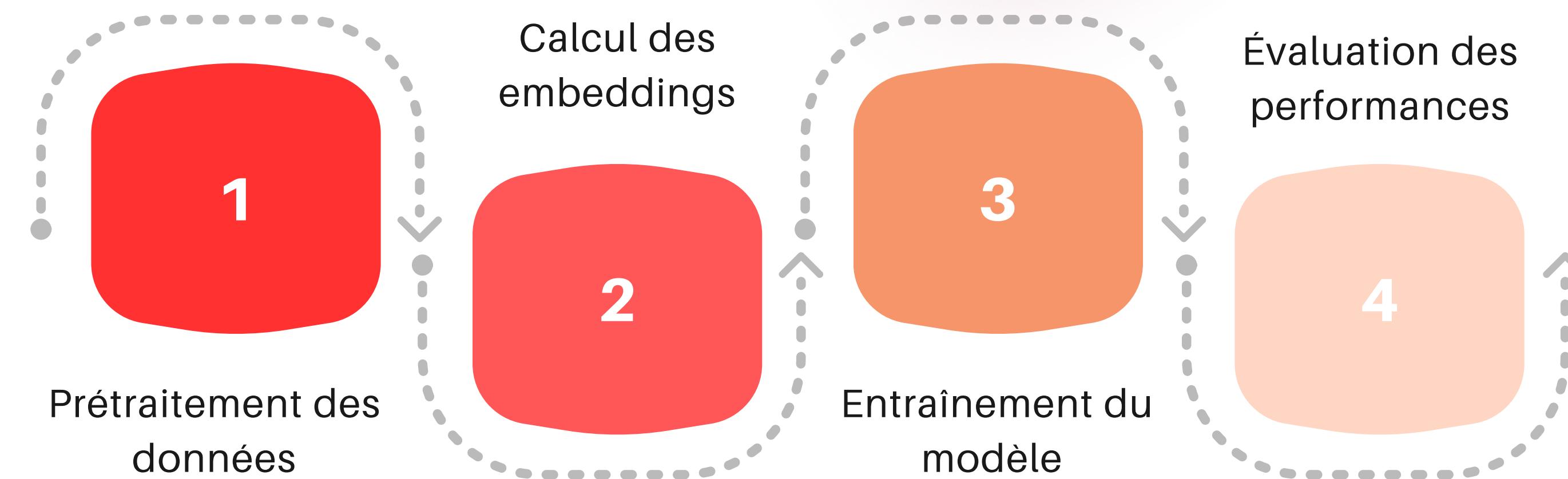
---

# Exemple d'annotation sur une composante

- Chaque nœud dans le graphe représente une protéine.
- Les nœuds **bleus** correspondent aux protéines **non annotées**.
- Les nœuds **rouges** correspondent aux protéines **annotées**.
- Un nœud peut avoir plusieurs labels.



# Prédiction des labels dans une composante



# Prétraitement des données

---

- Sélection d'une composante homogène.

Caractéristiques de la composante choisie	Valeur
Nombre d'arêtes	600 863
Nombre de Nœuds	110 442
Nombre de Nœuds annotés	7 444
Nombre d'annotations différentes	4

- Proposition et implémentation d'un algorithme d'attribution de label unique à chaque nœud annoté.

# Pipeline du machine learning

## Topologie en embedding

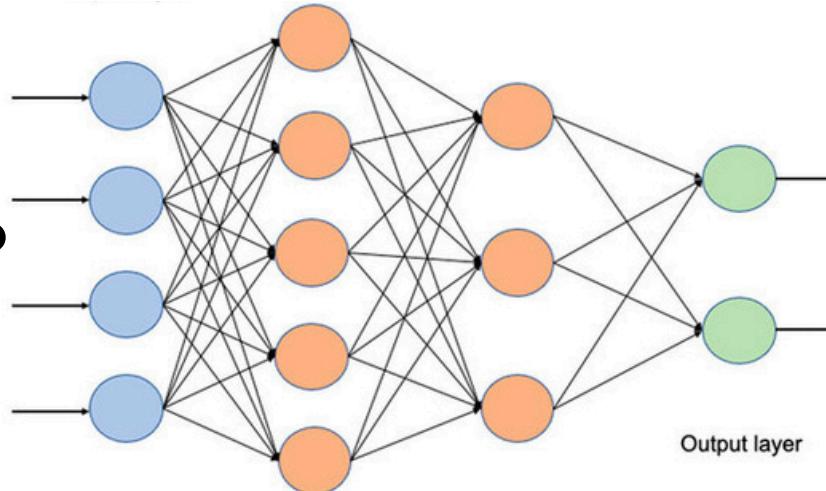
N1	(x <sub>11</sub> , x <sub>12</sub> , ..., x <sub>1d</sub> )
N2	(x <sub>21</sub> , x <sub>22</sub> , ..., x <sub>2d</sub> )
N3	(x <sub>31</sub> , x <sub>32</sub> , ..., x <sub>3d</sub> )
⋮	⋮
N <sub>k</sub>	(x <sub>k1</sub> , x <sub>k2</sub> , ..., x <sub>kd</sub> )



Nœuds annotés		Labels
N2	(x <sub>21</sub> , x <sub>22</sub> , ..., x <sub>2d</sub> )	Y <sub>2</sub>
⋮	⋮	⋮
N <sub>k</sub>	(x <sub>k1</sub> , x <sub>k2</sub> , ..., x <sub>kd</sub> )	Y <sub>k</sub>

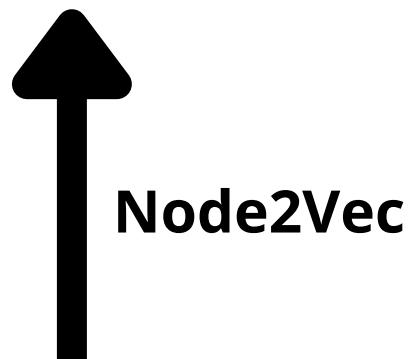


## MLP Classifier

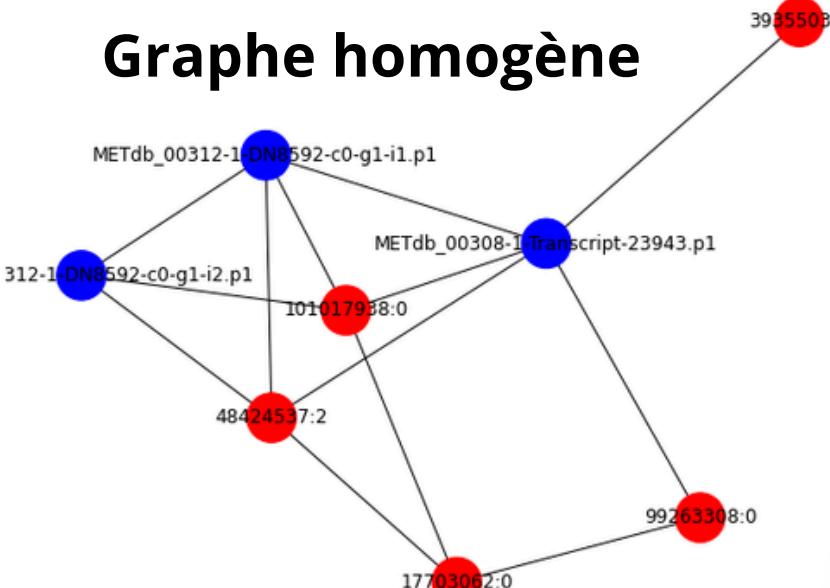


## Labels prédicts

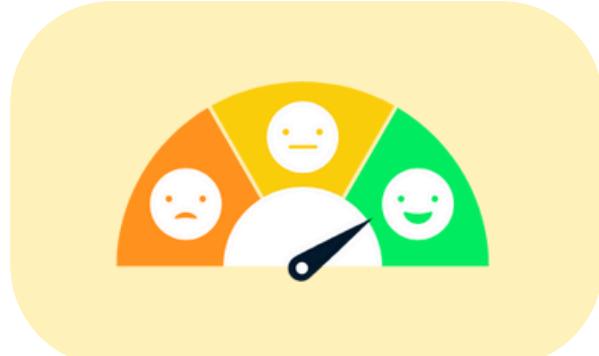
Y' <sub>1</sub>
Y' <sub>2</sub>
⋮
Y' <sub>k</sub>



## Graphe homogène



## Évaluation du modèle



# Conclusion

- Optimisation de l'espace mémoire et maximisation de l'utilisation des ressources locales tout en minimisant les échanges.
- Méthode permettant une analyse rapide et efficace de graphes massifs.
- Prétraitement optimal attribuant des labels uniques à chaque nœud.
- Modèle prédictif pour annotations manquantes, enrichissant la compréhension biologique.

# Perspectives

- Variation du degré de parallélisme entre le calcul des composantes partielles et celui de  $G'$  pour améliorer les performances.
- Amélioration des prédictions d'annotations par la division des composantes en communautés.
- Utilisation du clustering et propagation d'annotations majoritaires pour des composantes hétérogènes.

# Merci de votre attention !

---