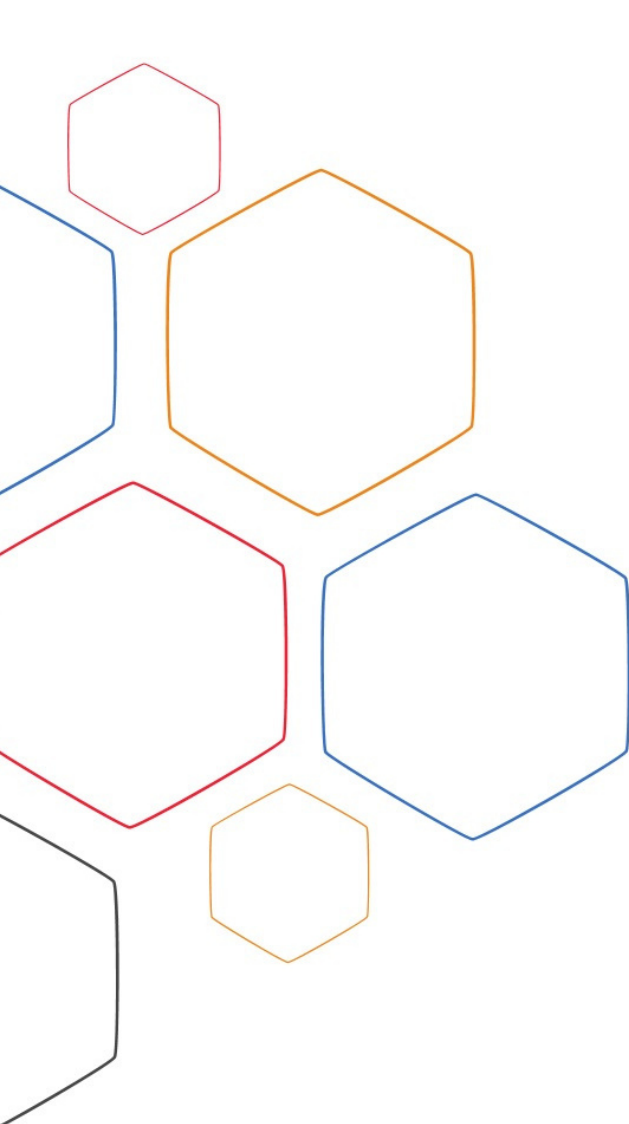




Journées NetBio · 16 Septembre 2026

Exploitation de graphes hétérogènes par Graph Neural Networks

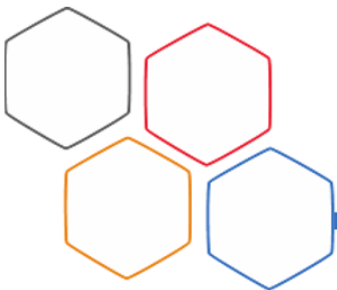
Application à l'annotation fonctionnelle des protéines

Antoine Toffano

Superviseurs : **Pierre Larmande & Jérôme Azé**

Laboratoire d'Informatique, de Robotique et de Microélectronique de
Montpellier (LIRMM)





Le défi de l'annotation fonctionnelle

Séquençage massif vs annotation expérimentale

> 150M

Séquences UniProt

Séquençage haut débit

< 0.5%

Annotées en pailasse

Validation expérimentale

La prédiction comme filtre de priorisation

- Tester chaque protéine pour chaque fonction est impossible.
- La prédiction computationnelle guide et accélère la découverte biologique (santé, agronomie, ..).

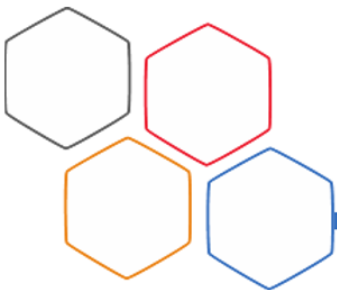
Objectifs :

Précision Identifier les fonctions les plus spécifiques.

Généralisation Généralisable à tout type de protéine.

Interprétabilité Expliquer les prédictions.

Zhou et al., « The CAFA challenge... », *Genome Biology*, 2019.



Formalisation des annotations par la Gene Ontology (GO)

Structure et sous-ontologies

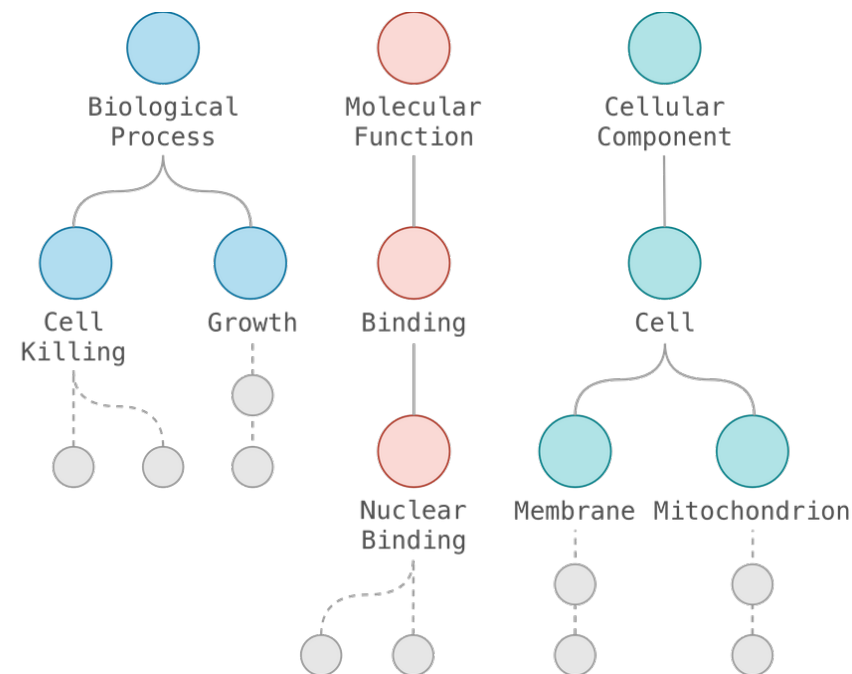
Vocabulaire structuré :

- Nœuds : termes fonctionnels biologiques précis.
- Arêtes : relations hiérarchiques.

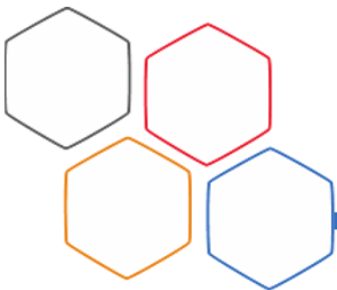
3 domaines fonctionnels :

- **BP Biological Process** : processus multi-étapes (ex: *croissance cellulaire, voies de signalisation*).
- **MF Molecular Function** : activité au niveau moléculaire (ex: *kinase, liaison d'ions*).
- **CC Cellular Component** : compartiment ou complexe (ex: *membrane, mitochondrie*).

Une protéine annotée à un nœud hérite obligatoirement de toutes les annotations de ses ancêtres.



Structure hiérarchique de la Gene Ontology (Ashburner et al., 2000).



L'objectif de la prédiction fonctionnelle

Évaluation multi-label & métriques

Classification hiérarchique multi-label :

- Pour chaque protéine, prédiction d'un ensemble de termes face aux annotations expérimentales.

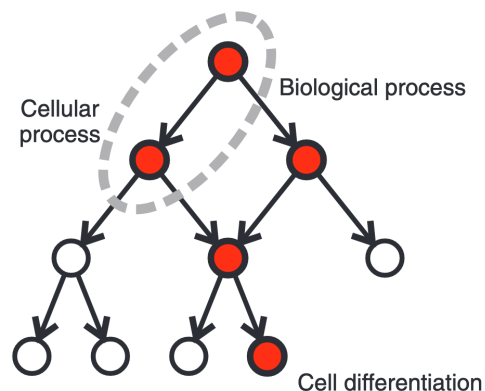
Métrique indicative: F max

$$\text{Pr}(\tau) = \frac{\text{TP}}{\text{TP} + \text{FP}} \quad \text{Rc}(\tau) = \frac{\text{TP}}{\text{TP} + \text{FN}}$$

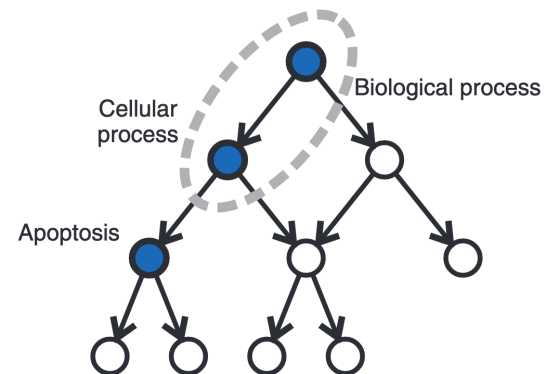
$$F_{\max} = \max_{\tau} \frac{2 \cdot \text{Pr}(\tau) \cdot \text{Rc}(\tau)}{\text{Pr}(\tau) + \text{Rc}(\tau)}$$

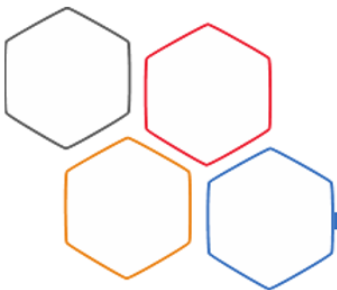
Compromis optimal entre précision et rappel sur les seuils τ .

a Predicted function



b True function





Du transfert d'alignement à l'apprentissage machine

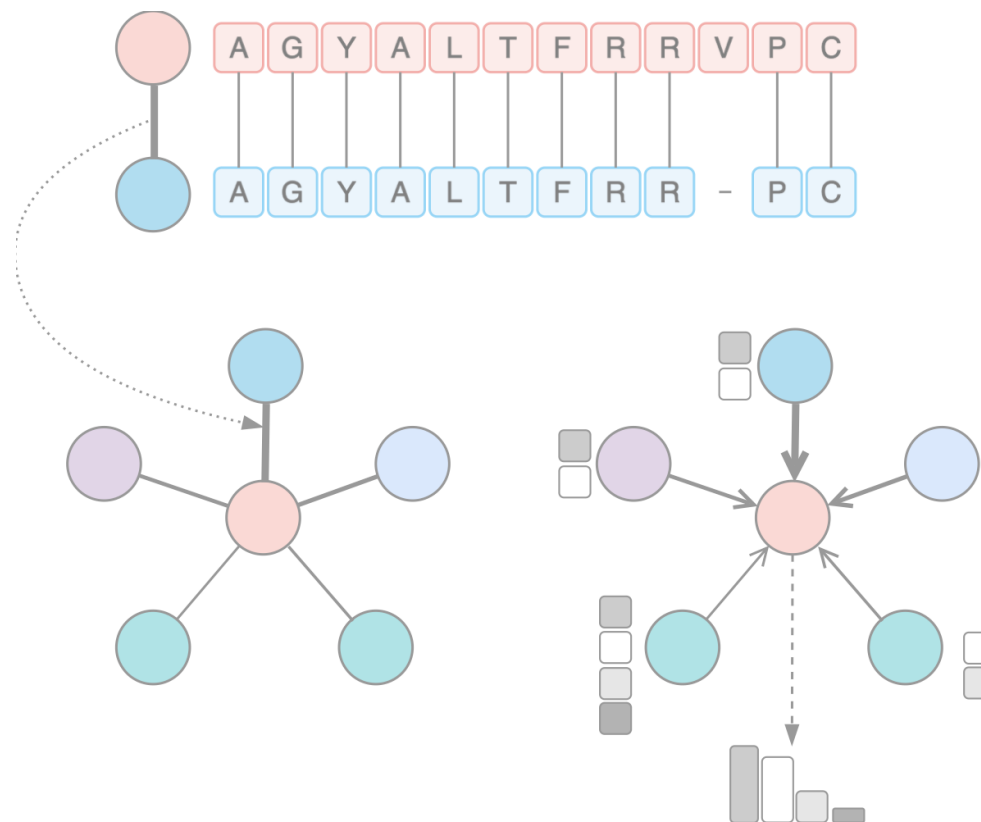
Limites de l'homologie

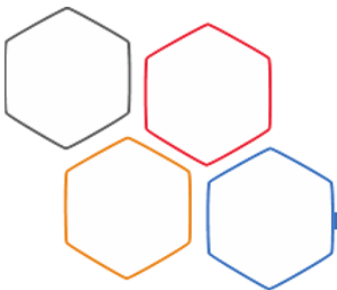
Le transfert classique par alignement :

- Recherche d'homologues (BLAST, Diamond).
- Transfert direct des annotations du plus proche voisin.

Pourquoi l'apprentissage machine s'est imposé :

- Perte de performance sur les protéines à faible homologie (< 30% d'identité)
- Échec sur les protéines isolées.
- **Ne prend en compte que la similarité de séquence !**



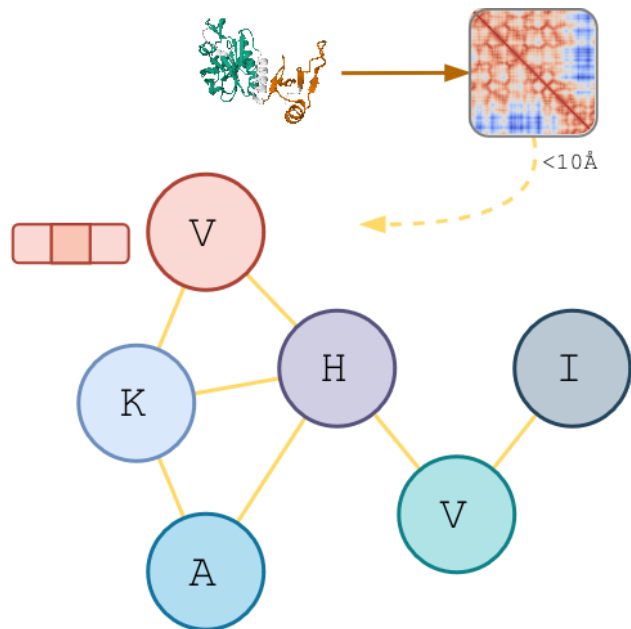


Modélisation par les graphes : Deux approches

Approche structurale vs Réseau systémique

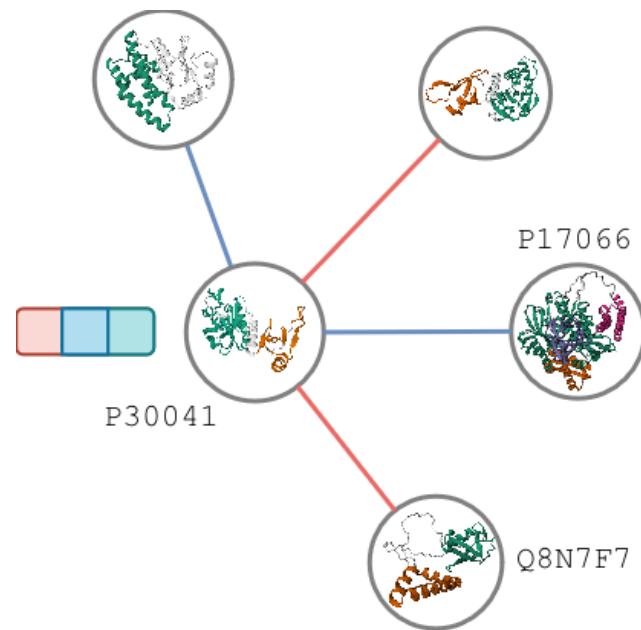
1. Protéine comme un graphe (Structurale)

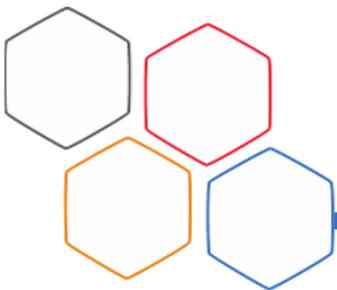
- Modélise la protéine elle-même au niveau résidu / atomique.



2. Représentation en réseau (Systémique)

- Modélise la protéine comme un nœud dans un réseau d'interactions.





Graph Neural Networks

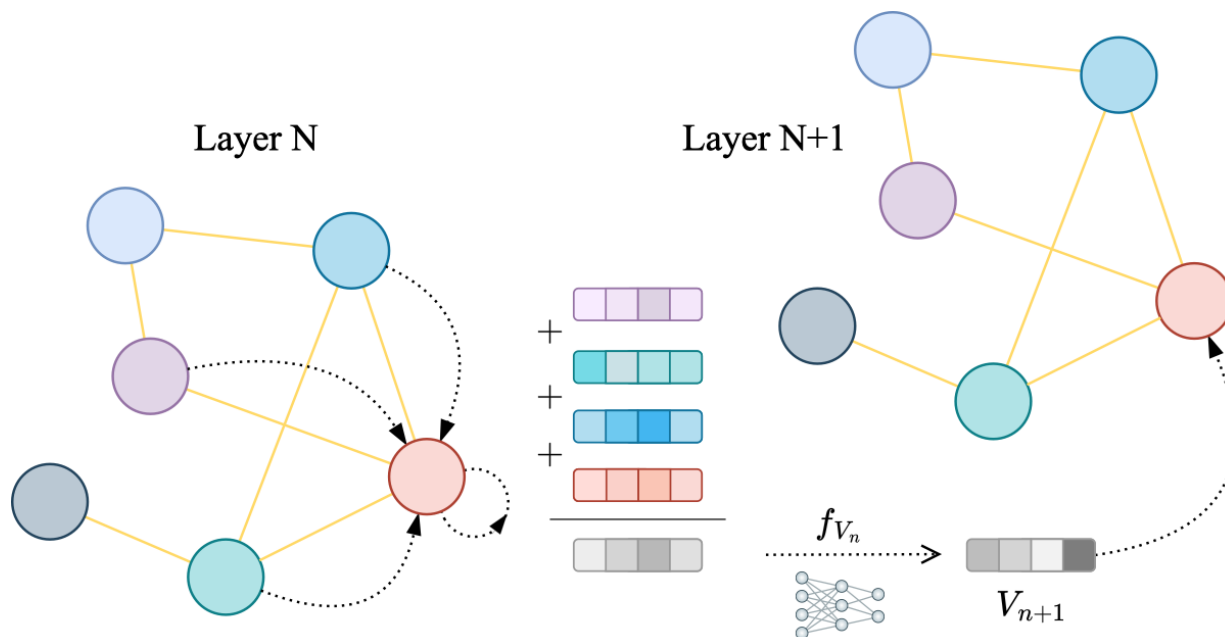
Message Passing: Agrégation & mise à jour des représentations

Message Passing

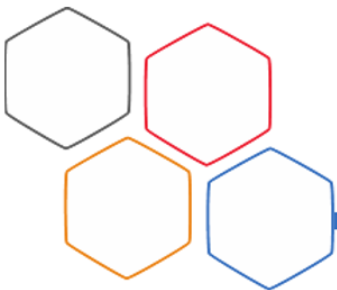
- **Agrégation** : Réception des “messages” (i.e. embeddings des voisins.)
- **Pooling** : Intégration des messages en une représentation unique
- **Mise à jour** : Transformation de la représentation du noeud par une couche linéaire.

Répéter k fois pour propager l'information aussi loin que voulu

Champ récepteur : L'empilement de k couches intègre les informations à une distance de k .



$$V_n \xrightarrow{p_{V_n \rightarrow V_n}} \text{Node} \xrightarrow{f_{V_n}} V_{n+1}$$



Mécanisme d'attention sur graph

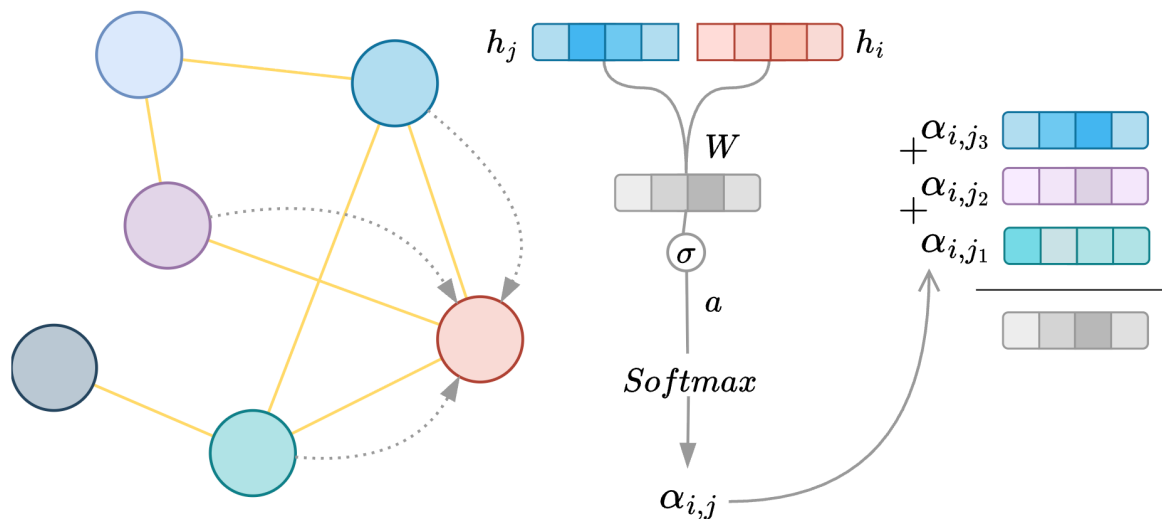
Et intégration des podérations des arêtes

Le problème du Message Passing standard :

- Un voisin a la même importance dans la somme, quelle que soit ses caractéristiques.

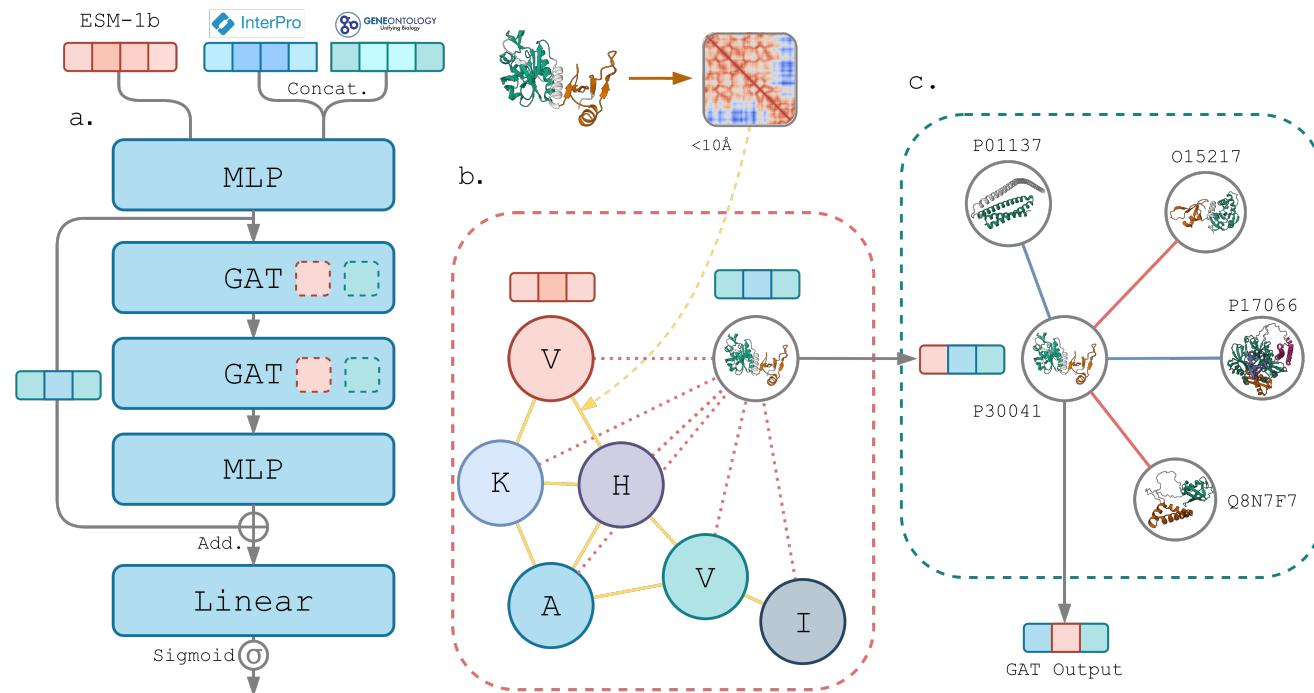
Intégration explicite des attributs d'arêtes :

- Distances spatiales C α -C α (AlphaFold), scores de confiance STRING, bitscores d'homologie (DIAMOND).
- Le modèle apprend **quelles interactions filtrer** pour chaque fonction.



Architecture GNN globale

Intégration multi-échelles résidu-protéine-réseau



1. Échelle locale des résidus

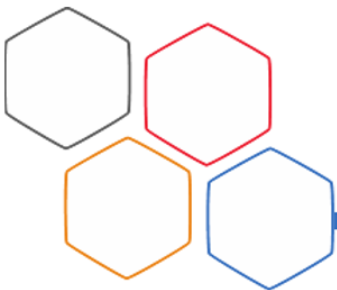
- Nœuds résidus initialisés par embeddings **ESM-1b (1280-d)**.
- Arêtes spatiales intra-protéine issues d'AlphaFold.

2. Couplage via nœud global

- Nœuds protéines initialisés par domaines InterPro et termes GO connus (voisins).

3. Échelle réseau & Inférence

- Arêtes inter-protéines : homologie (DIAMOND) et interactions (STRING).



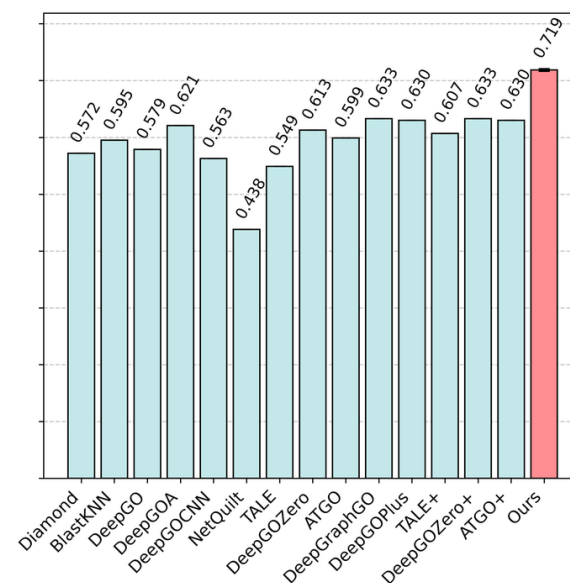
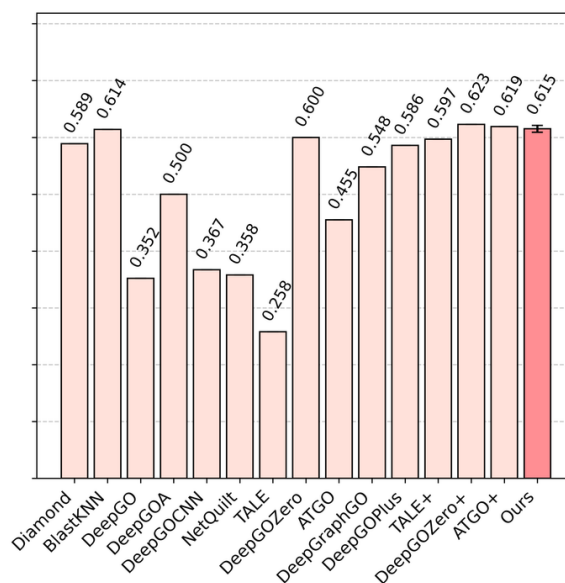
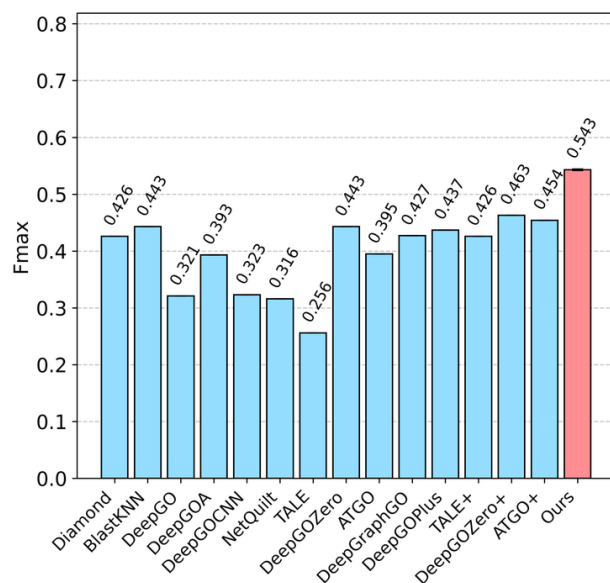
Évaluation sur benchmark

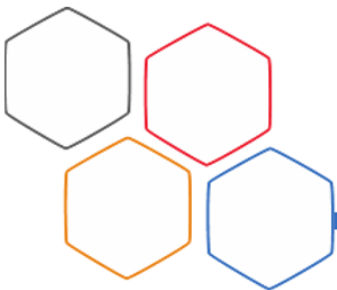
Comparaison face à l'état de l'art

+17.3% vs SoTA · Biological Process

-0.6% vs SoTA · Molecular Function

+13.6% vs SoTA · Cellular Component





Explicabilité réseau

Sélection de l'information des orthologues

Filtrage sélectif du voisinage :

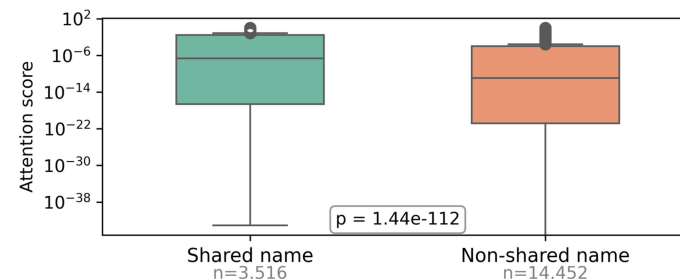
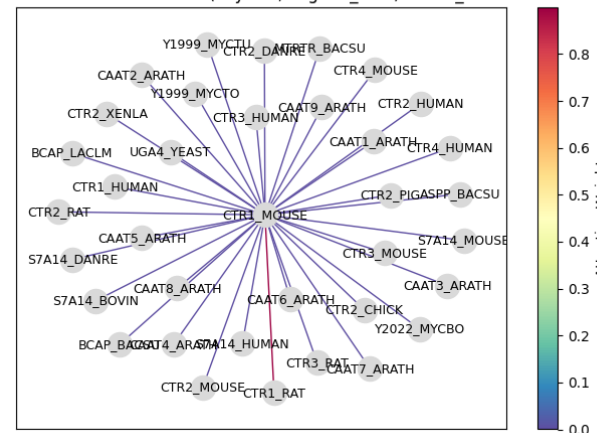
- Concentration du flux d'information sur l'orthologue direct.

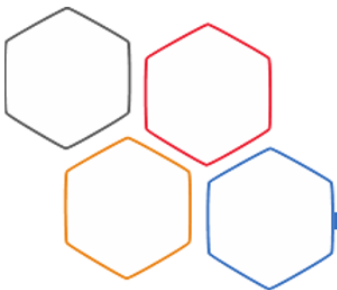
Expérience d'ablation :

- Supprimer les voisins homologues dégrade énormément les prédictions.

Les orthologues reçoivent une attention massivement supérieure face aux autres voisins.

Protein-Protein Attention (Layer 2, aligned_with): CTR1_MOUSE





Cartographie d'attention : Famille des transporteurs

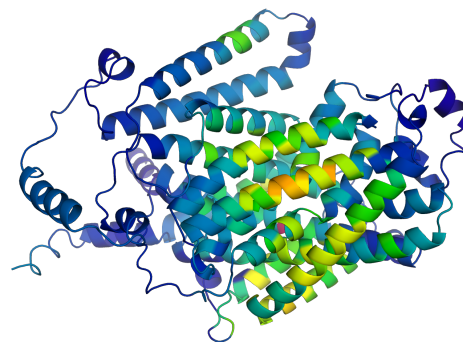
Haute attribution dans les zones transmembranaires

Consensus multi-modèles :

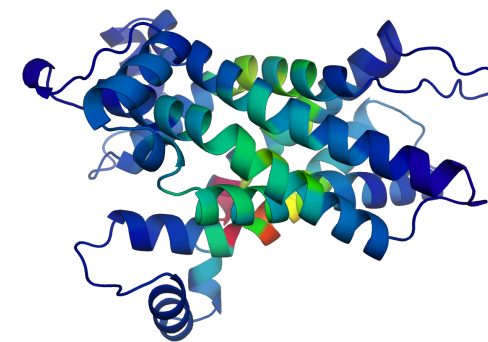
- Agrégation pondérée des contributions selon l'accord mutuel entre les modèles.

Propriété émergente :

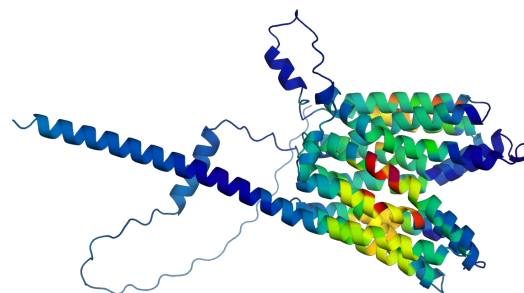
- L'attention résiduelle se concentre dans le domaine transmembranaire.
- Les boucles cytosoliques et extracellulaires sont ignorées.



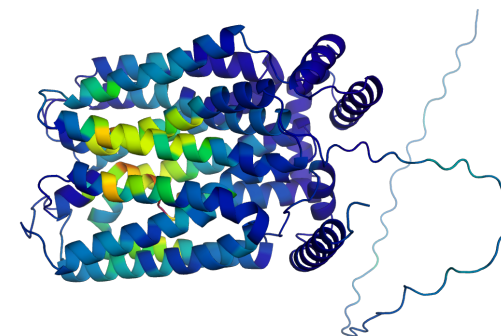
CTR1_HUMAN (Cuivre)



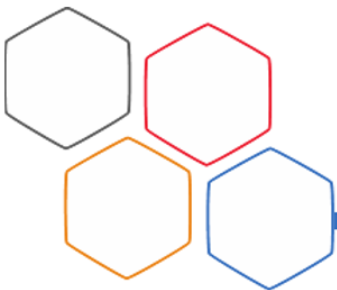
DIC_RAT (Dicarboxylate)



MOT10_RAT (Tryptophane)



ITR1_CANAL (Myo-inositol)



Étude de cas : Bêta-1,4-galactosyltransférase 1 humaine

Focalisation de l'attention sur les régions fonctionnelles

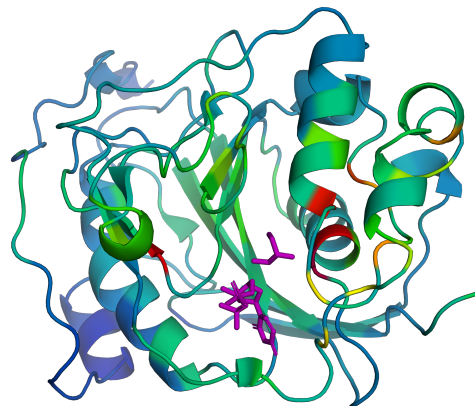
Activité catalytique (synthèse du lactose) :



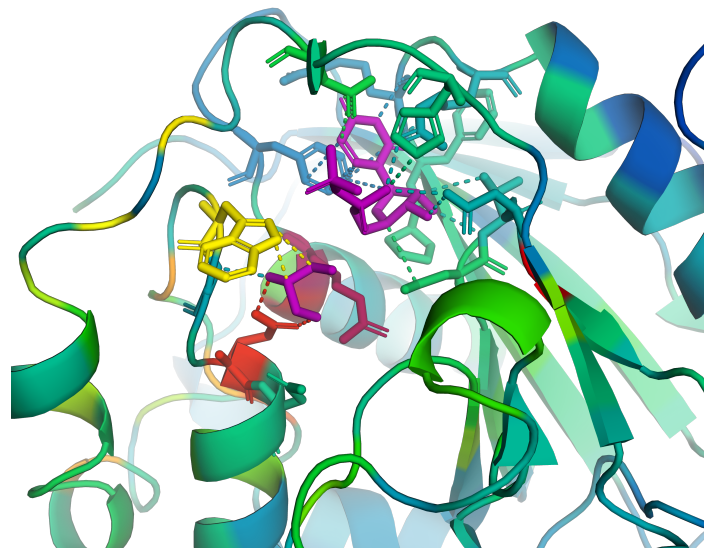
Concordance structurale :

- Les résidus à forte attention délimitent la cavité catalytique.
- Superposition directe avec les substrat (en violet).

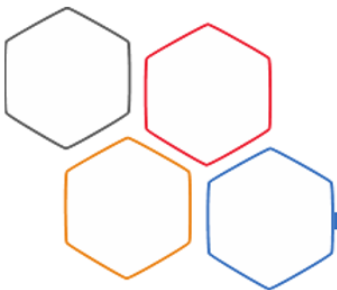
Détection non supervisée : les résidus en contact direct avec les substrats émergent spontanément.



Vue globale



Site actif & Substrat



Étude de cas : Aldéhyde déshydrogénase murine

Détection de deux poches de liaison (Site catalytique NAD⁺ & poche secondaire)

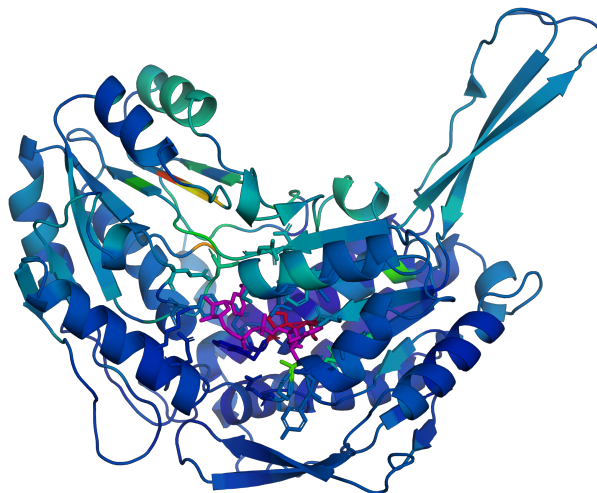
Validation par modèle spécialisé (AlphaFill) :

- Prédiction indépendante du cofacteur NAD⁺ (en violet).

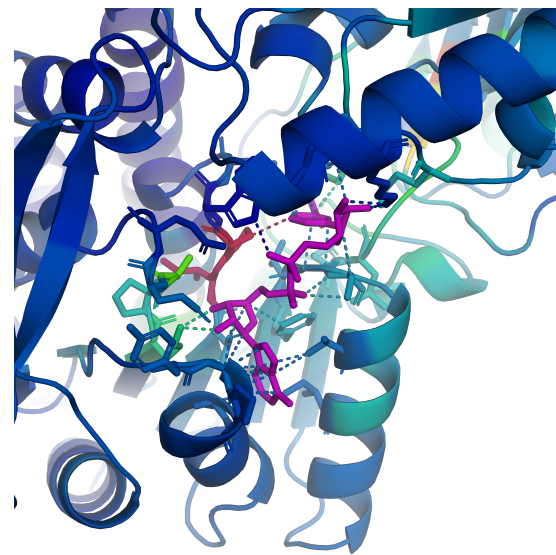
Détection de deux poches distinctes :

- **Poche 1 (Site catalytique)** : superposition directe avec le cofacteur NAD⁺.
- **Poche 2 (Secondaire)** : zone de haute attention sans rôle catalytique documenté.

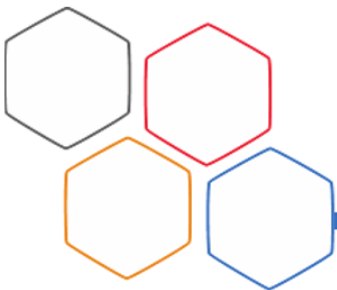
Isolation de régions fonctionnelles potentielles.



Vue globale (2 poches)



Poche NAD⁺ (AlphaFill)



Limitations

Sur-lissage et contrainte de taille de voisinage

Le phénomène de lissage (Over-Smoothing) :

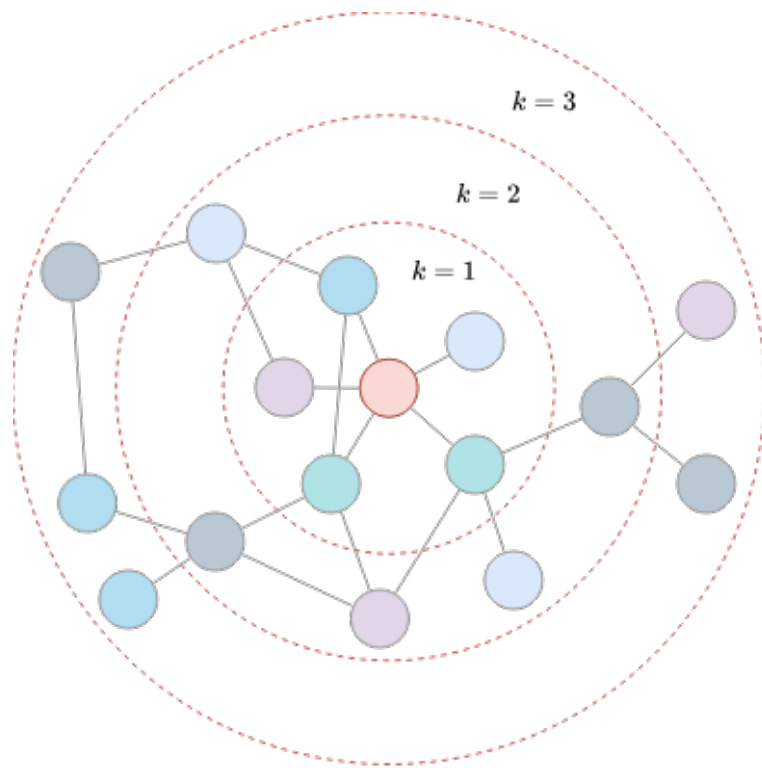
- L'agrégation répétée du voisinage entraîne une homogénéisation des représentations : les vecteurs convergent vers des profils indiscernables.

Contrainte de profondeur ($K = 2$)

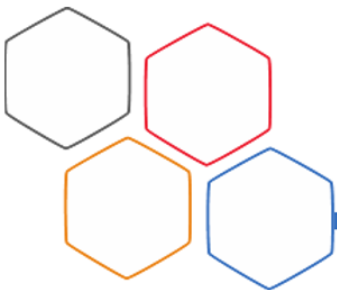
Nécessité de limiter le nombre de couches du réseau pour éviter la dilution complète du signal fonctionnel.

Explosion du nombre de voisins:

- La profondeur des GNNs est également limitée computationnellement via l'augmentation de K .



Expansion du champ récepteur.



Conclusion et perspectives

Graph Neural Networks comme outils d'intégration multimodaux

Dualité structurale & systémique

- Couplage synergique de la **structure 3D** (contacts AlphaFold) et du **réseau d'interactions** (PPI, homologie).

Attention & Explicabilité

- Émergence spontanée des régions d'importance sans supervision.

Performance & Anti-sur-lissage

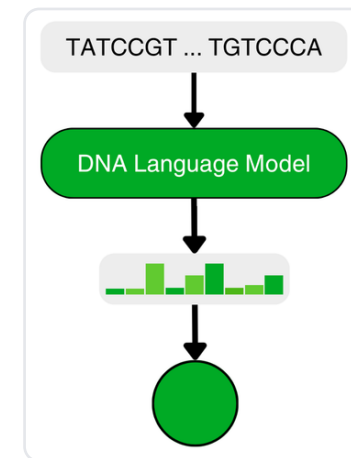
- **+17.3% (BPO)** et **+13.6% (CCO)** face à l'état de l'art.

Modèles de langage génomiques (gLM)

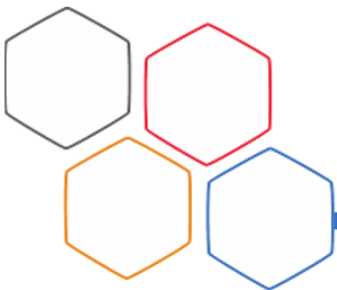
- Encodage de la séquence codante
- ..et des séquences régulatrices (promoteurs, enhancers, non-codant).

Réseaux de régulation (GRN)

- Au-delà du réseau PPI physique : modéliser la dynamique transcriptionnelle (TF → cibles).
- But: tendre vers un modèle intégrant génome, transcriptome et protéome 3D.



Vectorisation d'ADN par modèle de langage génomique.



Journées NetBio · 16 Septembre 2026

Merci pour votre attention !

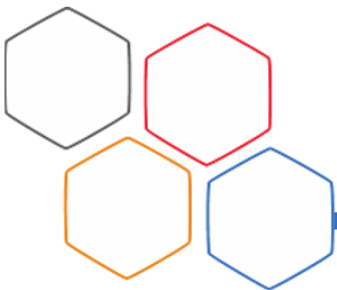
Antoine Toffano · antoine.toffano@lirmm.fr

Superviseurs : **Pierre Larmande** (IRD) & **Jérôme Azé** (Univ. Montpellier / LIRMM)



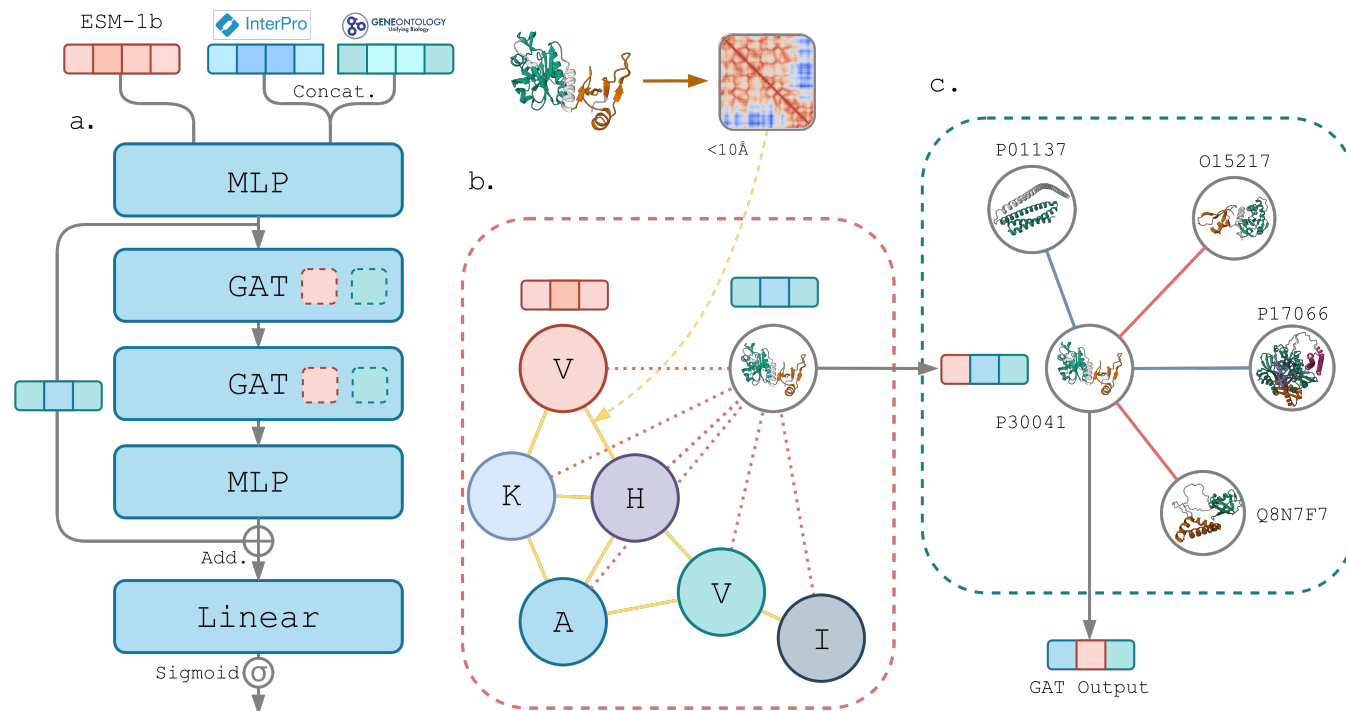
UNIVERSITÉ DE
MONTPELLIER



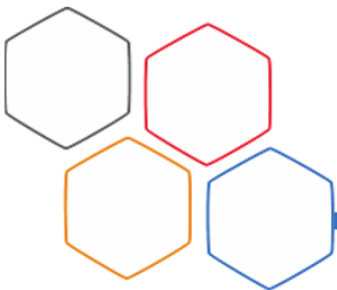


Annexe · Architecture MS-GNN (0/2)

Vue d'ensemble de l'architecture multi-échelles résidu-protéine-réseau



Architecture MS-GNN : intégration conjointe des niveaux résidus (AlphaFold/ESM-1b) et réseau (STRING/DIAMOND).



Annexe · Architecture MS-GNN (1/2)

Graphe hétérogène $\mathcal{G} = (\mathcal{V}, \mathcal{E})$

- **Nœuds résidus** $r \in \mathcal{V}_r$: plongements ESM-1b $\mathbf{x}_r \in \mathbb{R}^{1280}$.
- **Nœuds protéines** $p \in \mathcal{V}_p$: concaténation binaire InterPro et GO ($\mathbf{x}_p$). Masquage des termes GO pour cible et sets val/test.
- **Arêtes relationnelles** $e \in \mathcal{E}$:
 - Résidu $\rightarrow$ Résidu : contact AlphaFold si $d(C_\alpha) < 10 \text{ \AA}$, distance normalisée au carré en attribut ($\mathbf{e}_{r,r'}$).
 - Protéine $\rightarrow$ Résidu : bridge entre représentation globale $\rightarrow$ locale.
 - Protéine $\rightarrow$ Protéine : homologie (DIAMOND normalisé) et interactions physiques/fonctionnelles (STRING normalisé).

Bloc MLP & Projections initiales

Bloc élémentaire avec LayerNorm et activation PReLU :

$$\text{MLP}(\mathbf{x}) = \text{LayerNorm}(\text{PReLU}(\mathbf{W}\mathbf{x} + \mathbf{b}))$$

Projections vers l'espace latent partagé (d -dimensionnel) :

$$\mathbf{x}_r^0 = \text{MLP}_r(\mathbf{x}_r), \quad \mathbf{x}_p^0 = \text{MLP}_p(\mathbf{x}_p)$$

Propagation de messages par GAT hétérogène

Deux couches successives adaptées pour intégrer les attributs d'arêtes $\mathbf{e}_{i,j}$. Pour un nœud cible i à la couche l , sommation sur les relations $r \in \mathcal{E}$:

$$\mathbf{x}_i^l = \text{LayerNorm} \left(\text{PReLU} \left(\sum_{r \in \mathcal{E}} \text{GAT}_r^l(\mathcal{N}_r(i) \cup \{i\}) \right) \right)$$

Agrégation relationnelle pondérée par attention (projection cible Θ_t) :

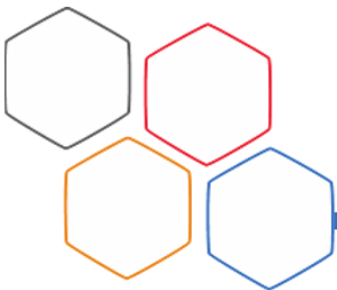
$$\mathbf{x}_{i,r}^l = \sum_{j \in \mathcal{N}_r(i) \cup \{i\}} \alpha_{i,j} \Theta_t \mathbf{x}_j^{l-1}$$

Fonction de score incluant explicitement les attributs d'arêtes $\mathbf{e}_{i,j}$:

$$S_{i,j} = \Theta_s \mathbf{x}_i^{l-1} + \Theta_t \mathbf{x}_j^{l-1} + \Theta_e \mathbf{e}_{i,j}$$

Coefficients d'attention normalisés par Softmax :

$$\alpha_{i,j} = \frac{\exp(\mathbf{a}^\top \text{LeakyReLU}(S_{i,j}))}{\sum_{j' \in \mathcal{N}_r(i) \cup \{i\}} \exp(\mathbf{a}^\top \text{LeakyReLU}(S_{i,j'}))}$$



Annexe · Architecture MS-GNN (2/2)

Formalisation mathématique : tête de classification résiduelle, perte BCE et consensus d'attention

Tête de prédiction & Connexion résiduelle

- **Préservation du signal & Anti-sur-lissage** : Concaténation de la projection initiale x_p^0 avec la sortie de la seconde couche d'attention x_p^2 :

$$\hat{y} = \sigma(W_{\text{out}} \text{MLP}_{\text{post}(x_p^0 \parallel x_p^2)} + b_{\text{out}})$$

- σ : activation sigmoïde pour une prédiction multi-label indépendante par terme GO.

Fonction de perte (BCE unweighted)

Optimisation par entropie croisée binaire non pondérée :

$$\mathcal{L} = -\frac{1}{N} \sum_{n=1}^N [y_n \log(\hat{y}_n) + (1 - y_n) \log(1 - \hat{y}_n)]$$

- Deux régimes d'entraînement évalués : **EXP** (labels expérimentaux) vs **CUR** (labels curés Swiss-Prot étendus). Évaluation toujours sur vérité terrain expérimentale.

Accord inter-modèles : Tau de Kendall (τ)

Mesure de l'accord sur l'ordre relatif d'importance des résidus/voisins entre deux modèles A et B :

$$\tau = \frac{\# \text{ paires concordantes} - \# \text{ paires discordantes}}{\text{nombre total de paires}}$$

- $\tau = 1$: classements identiques ; $\tau = -1$: classements opposés.
- Accord fort sur les partenaires homologues ($\tau \approx 0.75 - 0.80$), plus variable sur les résidus ($\tau \approx 0.0 - 0.40$).

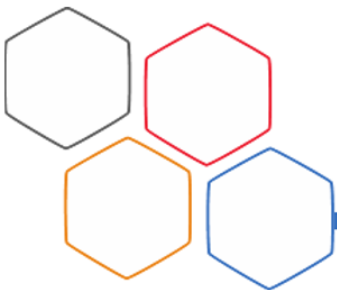
Pondération consensuelle multi-modèles (S)

Poids moyen d'accord du modèle i avec les $M - 1$ autres modèles :

$$\tilde{w}_i = \frac{1}{M-1} \sum_{j \neq i} \tau_{i,j}, \quad w_i = \frac{\tilde{w}_i}{\sum_{k=1}^M \tilde{w}_k}$$

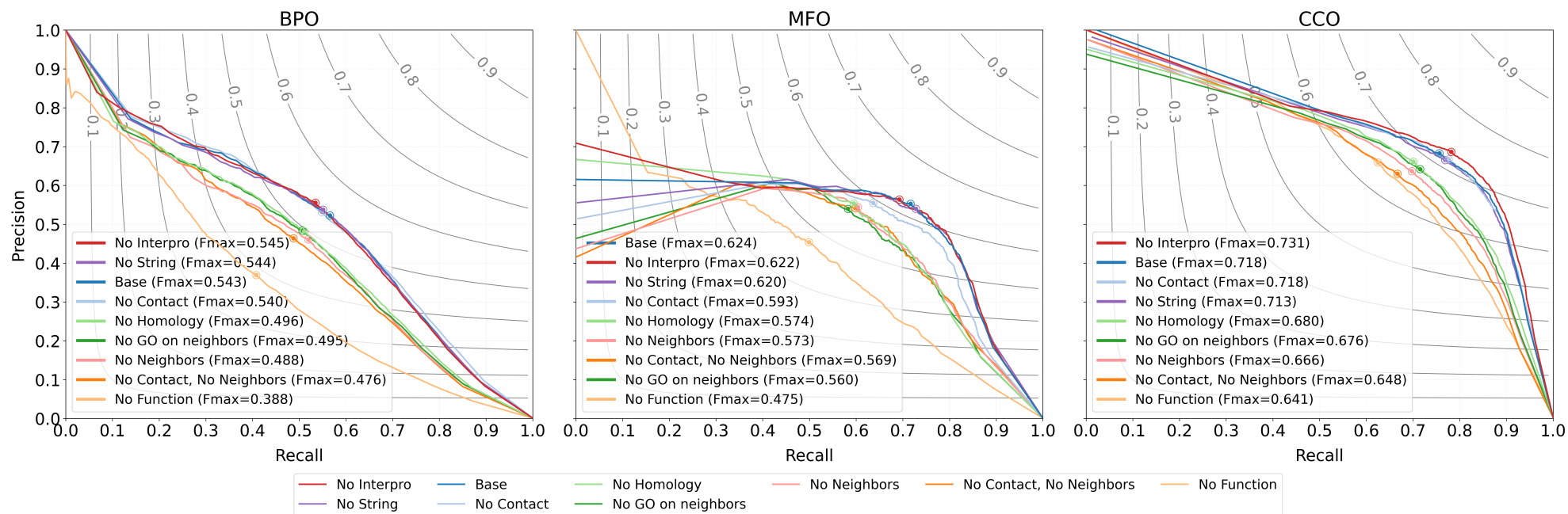
Score d'attention consensuel final :

$$S = \sum_{i=1}^M w_i \alpha_i$$

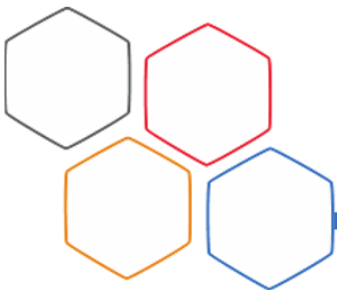


Annexe · Étude d'ablation des caractéristiques

Contribution des modalités biologiques et dynamique de compensation (MS-GNN CUR)



Courbes Précision-Rappel pour chaque condition d'ablation sur le jeu de test D1 (BPO, MFO, CCO). Points : $F_{\max}$.



Annexe · Évaluation CAFA (1/2)

Formalisation mathématique : métriques dépendantes du seuil et compromis $F_{\max}$

Cadre d'évaluation multi-label au seuil τ

Pour chaque protéine cible i et seuil de décision $\tau \in [0, 100]$:

- $P_{i(\tau)} = \{f \mid s_{i,f} \geq \tau\}$: ensemble des termes GO prédits.
- T_i : ensemble des termes GO vrais (annotations expérimentales).
- N : nombre total de protéines évaluées.
- $m(\tau)$: nombre de protéines avec au moins une prédiction au seuil τ ($|P_{i(\tau)}| > 0$).
- $I(\cdot)$: fonction indicatrice valant 1 si la condition est vérifiée, 0 sinon.

Aire sous la courbe Précision-Rappel (AUPR)

- Évalue la capacité globale d'ordonnement sans sélection de seuil.
- Préférée au ROC en génomique fonctionnelle car très sensible aux faux positifs face à la rareté extrême des annotations positives.

Précision et Rappel moyens au seuil τ

- **Précision moyenne** calculée sur les protéines avec au moins une prédiction :

$$\text{pr}(\tau) = \frac{1}{m(\tau)} \sum_{i=1}^{m(\tau)} \frac{\sum_f I(f \in P_{i(\tau)} \wedge f \in T_i)}{\sum_f I(f \in P_{i(\tau)})}$$

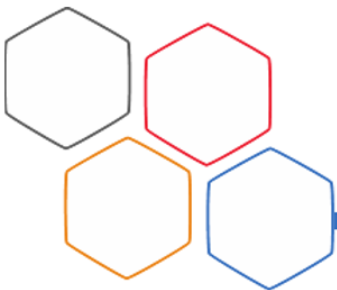
- **Rappel moyen** calculé sur l'ensemble des N protéines cibles :

$$\text{rc}(\tau) = \frac{1}{N} \sum_{i=1}^N \frac{\sum_f I(f \in P_{i(\tau)} \wedge f \in T_i)}{\sum_f I(f \in T_i)}$$

Mesure F maximale ($F_{\max}$)

Métrique principale CAFA ; représente la moyenne harmonique optimale entre précision et rappel sur l'ensemble des seuils τ :

$$F_{\max} = \max_{\tau} \left(\frac{2 \cdot \text{pr}(\tau) \cdot \text{rc}(\tau)}{\text{pr}(\tau) + \text{rc}(\tau)} \right)$$



Annexe · Évaluation CAFA (2/2)

Formalisation mathématique : contenu d'information (IC) et distance sémantique minimale ($S_{\min}$)

Contenu d'Information (IC - Information Content)

- **Limite des métriques classiques** : précision et rappel traitent tous les termes de manière uniforme. Prédire un terme racine (e.g., *binding*) est trivial, alors qu'un terme spécifique (e.g., *sequence-specific DNA binding*) est hautement informatif.

- **Rareté conditionnelle** dans le DAG de la Gene Ontology :

$$IC(f) = -\log(Pr(f \mid \text{Parents}(f)))$$

- Plus un terme est rare par rapport à ses parents, plus son IC est élevé.

Incertitude restante (ru) et Désinformation (mi)

- **Incertitude restante** : somme d' IC des termes vrais manqués (faux négatifs) :

$$ru(\tau) = \frac{1}{N} \sum_{i=1}^N \sum_f IC(f) \cdot I(f \notin P_{i(\tau)} \wedge f \in T_i)$$

- **Désinformation** : somme d' IC des termes prédits à tort (faux positifs) :

$$mi(\tau) = \frac{1}{N} \sum_{i=1}^N \sum_f IC(f) \cdot I(f \in P_{i(\tau)} \wedge f \notin T_i)$$

Distance sémantique minimale ($S_{\min}$)

Distance euclidienne minimale à l'origine dans l'espace (ru, mi) :

$$S_{\min} = \min_{\tau} \sqrt{ru(\tau)^2 + mi(\tau)^2}$$

- Récompense la prédiction de termes précis et pénalise sévèrement les prédictions confiantes de termes rares erronés.