

Prédiction de la localisation subcellulaire des protéines procaryotes par fusion multi-modale d'embeddings

Maxence Agra

maxence.agra@telecommancy.eu

Mathis Fabre

mathis.fabre@telecommancy.eu

TELECOM Nancy – Université de Lorraine

Module MOM – Bio-informatique

Encadrante : Pr. Yasaman Karami (INRIA / LORIA)

Encadrant : Victor Pryakhin (INRIA / LORIA)

Résumé

Nous présentons une méthode de prédiction de la localisation subcellulaire des protéines procaryotes en 6 classes, basée sur la fusion de deux embeddings de modèles de langage protéique : ESM-2 650M (information évolutive, 1 280 dimensions) et ProstT5 (information structurelle, 1 024 dimensions). Notre architecture, *ImprovedFusionNetwork*, combine une cross-attention bidirectionnelle, un BiLSTM à 2 couches et un multi-head attention pooling, totalisant 8,4 millions de paramètres. Évaluée par validation croisée 5-fold sur le jeu de données DeepLocPro (11 531 séquences, partitionnement GraphPart à 30% d'identité de séquence), notre méthode obtient une accuracy de 0.932 ± 0.004 , un Macro F1 de 0.819 ± 0.018 et un MCC de 0.884 ± 0.007 , surpassant DeepLocPro sur l'ensemble des métriques.

1 Introduction

La localisation subcellulaire des protéines est une information fondamentale pour comprendre leur fonction biologique. Chez les procaryotes (bactéries et archées), on distingue six localisations principales : cytoplasme, membrane cytoplasmique, espace extracellulaire, périplasme, membrane externe et paroi cellulaire/surface. La prédiction automatique de cette localisation à partir de la séquence d'acides aminés est un enjeu majeur en protéomique et en biotechnologie [1].

DeepLocPro [1] est la méthode de référence pour cette tâche. Basé sur l'architecture DeepLoc 2.0 [6], il utilise un unique embedding ESM-2 650M avec un simple attention pooling suivi d'un MLP, atteignant une accuracy de 0.92 et un Macro F1 de 0.80 en validation croisée 5-fold. Cependant, l'utilisation d'un seul type d'embedding ne capture qu'une partie de l'information disponible : les patterns évolutifs encodés dans les séquences.

Récemment, Piochi et al. [2] ont montré que la combinaison d'embeddings évolutifs (ESM-C) et structuraux (ProstT5 3Di) améliore significativement la prédiction d'interactions protéine-protéine bactériennes, suggérant que la fusion multi-modale est une direction prometteuse.

Dans ce travail, nous proposons une architecture de fusion originale qui combine deux embeddings complémentaires via une cross-attention bidirectionnelle, enrichie d'un BiLSTM pour capturer les dépendances séquentielles, et d'un multi-head attention pooling pour l'agrégation. Nous évaluons cette approche sur le même jeu de données et le même protocole que DeepLocPro et montrons qu'elle surpassé la méthode de référence sur toutes les métriques.

2 Données et prétraitement

2.1 Source des données

Nous utilisons le jeu de données de DeepLocPro [1], construit à partir de PSORTdb 4.0 [7] et UniProt 2023_03. Le fichier source est un FASTA de 11 906 séquences protéiques expérimentalement vérifiées, réparties en 6 classes de localisation et 3 groupes d'organismes (Gram-négatif, Gram-positif, Archées). Le format du header FASTA est : >UniProtID|Location|OrganismGroup|FoldNumber. Les numéros de fold sont pré-calculés par GraphPart [5].

2.2 Pipeline de prétraitement

Notre prétraitement se déroule en trois étapes :

1. **Parsing et normalisation** : le FASTA est parsé avec BioPython. Les noms de localisation bruts sont normalisés vers 6 classes canoniques (par exemple, `cytoplasm`, `cytosol` → `Cytoplasmic`; `inner membrane`, `plasma membrane` → `Cytoplasmic_membrane`). Les groupes d'organismes sont également normalisés.
2. **Filtrage** : les séquences de moins de 40 acides aminés sont exclues (trop courtes pour être informatives), ainsi que les doublons (même séquence). Après cette étape, 11 906 séquences sont conservées.

3. **Filtrage de longueur ($\leq 1\,000$ AA)** : nous excluons les 375 séquences de plus de 1 000 acides aminés, conservant 11 531 séquences. Ce choix est motivé par trois raisons : (a) la consommation mémoire des embeddings per-position croît linéairement avec la longueur ; (b) les séquences très longues sont rares (3.2% du jeu de données) et représentent des cas atypiques ; (c) les modèles pLM ont été principalement évalués sur des séquences $\leq 1\,000$ AA. Ce seuil couvre 96.8% du jeu de données original.

2.3 Partitionnement

Le jeu de données est partitionné en 5 folds par GraphPart [5] avec un seuil d’identité de séquence maximal de 30% (Needleman–Wunsch) entre les folds. Ce partitionnement par homologie est crucial pour éviter toute fuite de données : sans lui, des séquences homologues pourraient se retrouver simultanément dans le train et le test, gonflant artificiellement les performances. Pour chaque itération de la validation croisée, 3 folds servent à l’entraînement ($\sim 6\,900$ séquences), 1 à la validation ($\sim 2\,300$) et 1 au test ($\sim 2\,300$).

2.4 Déséquilibre des classes

Le jeu de données présente un fort déséquilibre (Tableau 1). La classe Cytoplasmic représente 58.9% des séquences tandis que Cell wall & surface n’en représente que 0.5%, soit un ratio de 119 :1. Ce déséquilibre est un défi majeur pour l’apprentissage : sans traitement spécifique, le modèle tend à prédire systématiquement la classe majoritaire.

TABLE 1 – Distribution des 6 classes dans le jeu de données filtré (11 531 séquences $\leq 1\,000$ AA).

Classe	Effectif	%	Poids
Cytoplasmic	6 795	58.9	0.28
Cytoplasmic membrane	2 485	21.6	0.77
Extracellular	1 048	9.1	1.83
Outer membrane	733	6.4	2.62
Periplasmic	413	3.6	4.65
Cell wall & surface	57	0.5	33.69

Nous traitons ce déséquilibre par des poids de classe dans la fonction de loss, calculés par fréquence inverse : $w_c = N/(C \cdot N_c)$, où N est le nombre total de séquences, C le nombre de classes et N_c l’effectif de la classe c . Ainsi, la classe Cell wall reçoit un poids ~ 120 fois supérieur à la classe Cytoplasmic.

3 Embeddings protéiques

Notre approche repose sur la combinaison de deux types d’embeddings complémentaires, générés par des modèles de langage protéique (pLM) pré-entraînés. Chaque protéine produit deux tensors de forme (*seq_length, dim*), stockés en fichiers .pt individuels (~ 45 GB au total pour les 11 531 séquences).

ESM-2 650M [3] est un Transformer de 650M paramètres, entraîné par masked language modeling sur UniRef50/90. Il produit des représentations de dimension 1 280 par position, encodant les patterns de co-évolution et les propriétés biochimiques. Nous extrayons les représentations de la couche 33 (dernière couche).

ProstT5 [4] est un modèle de type T5, entraîné conjointement sur des paires séquence/structure 3D. Il produit des représentations de dimension 1 024 par position. Nous l’utilisons en mode acides aminés (AA), où la séquence est fournie avec des espaces entre chaque résidu.

Justification du choix des modèles. Le sujet préconise ESM-C [2] (300M params, 1 152d) et ProstT5 3Di (nécessitant Foldseek pour convertir les séquences en alphabet structural 3Di). Les contraintes du serveur de calcul (Python 3.9, pas d'accès administrateur, absence de Foldseek) ne permettaient pas leur utilisation :

- **ESM-C** requiert Python ≥ 3.12 et le package `esm` d'EvolutionaryScale. ESM-2 650M est son prédecesseur direct, issu de la même famille de modèles (masked language modeling sur UniRef), avec des représentations de dimension supérieure (1 280 vs 1 152) et davantage de paramètres (650M vs 300M).
- **ProstT5 3Di** nécessite Foldseek pour générer les séquences 3Di à partir des structures prédites. ProstT5 en mode AA conserve l'information structurelle implicite du pré-entraînement, bien que le mode 3Di soit plus explicitement structural.

4 Architecture du modèle

Nous proposons **ImprovedFusionNetwork**, une architecture de fusion multi-modale totalisant 8 412 936 paramètres entraînables. Elle se distingue fondamentalement de DeepLocPro par quatre innovations : (1) l'utilisation de deux embeddings au lieu d'un seul, (2) une fusion par cross-attention bidirectionnelle, (3) un encodeur séquentiel BiLSTM et (4) un pooling multi-têtes.

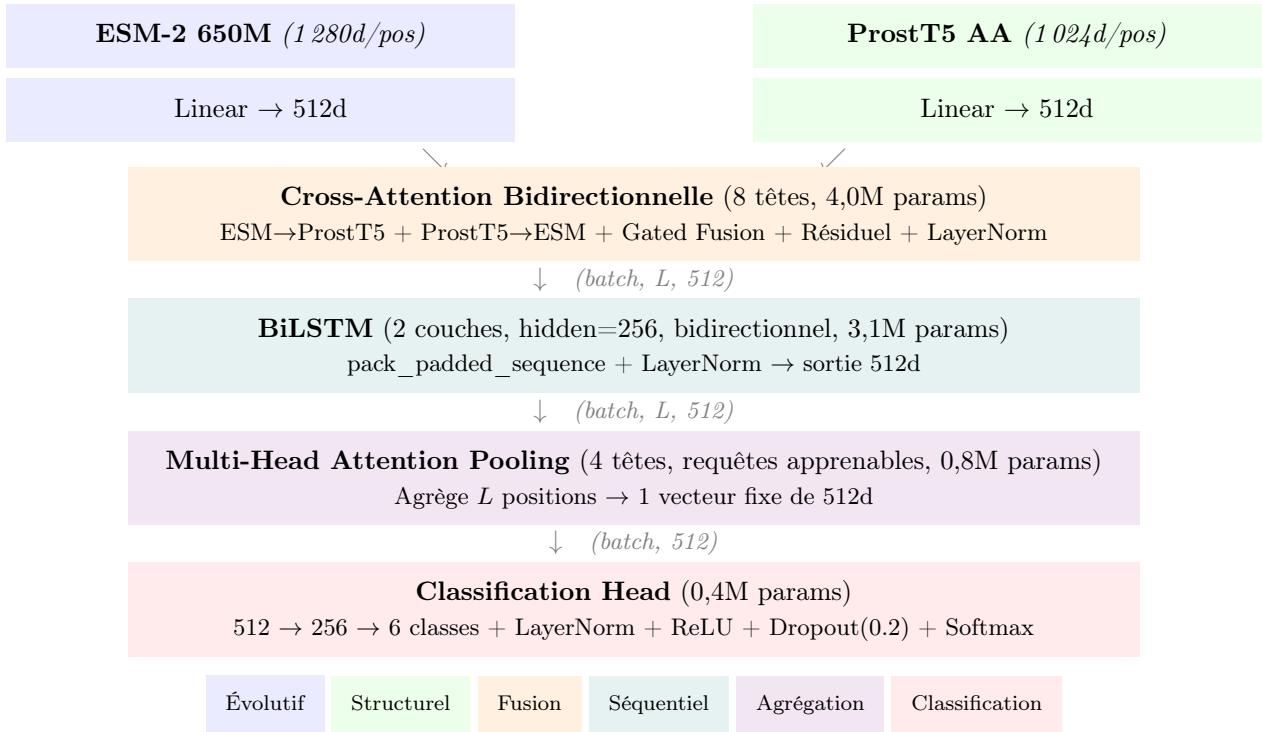


FIGURE 1 – Architecture d’ImprovedFusionNetwork (8,4M paramètres). Les deux embeddings protéiques sont projetés dans un espace commun (512d), fusionnés par cross-attention bidirectionnelle, encodés séquentiellement par un BiLSTM, agrégés par attention pooling multi-têtes, puis classifiés en 6 localisations subcellulaires. La légende de couleurs indique le rôle de chaque composant.

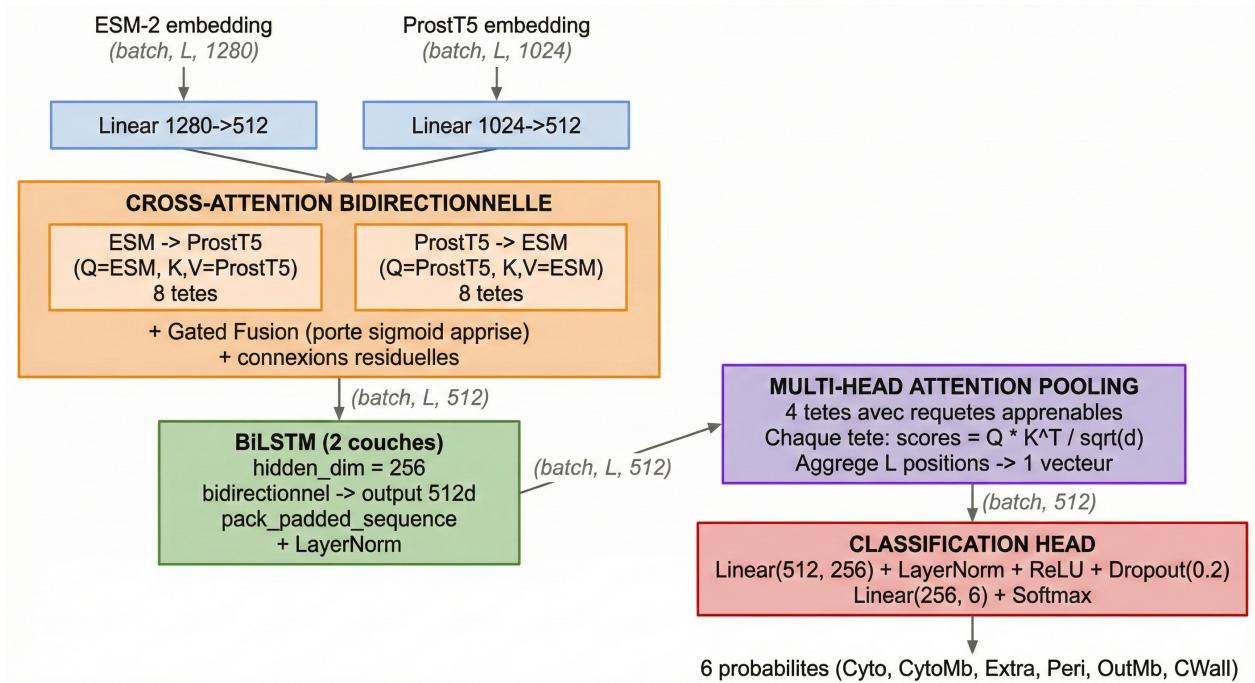


FIGURE 2 – **Architecture d’Improved FusionNetwork.** Les deux embeddings protéiques sont projetés dans un espace commun, fusionnés par cross-attention, encodés par un BiLSTM, puis agrégés pour la classification.

4.1 Cross-Attention Bidirectionnelle (4,0M params)

Notre module de cross-attention permet une interaction fine entre les deux modalités. Les embeddings sont d’abord projetés dans un espace commun de dimension 512, puis traités par une attention multi-têtes (8 têtes) dans les deux directions :

- **ESM → ProstT5** : ESM fournit les *queries*, ProstT5 les *keys/values*. La séquence « consulte » la structure pour s’enrichir d’information structurelle.
- **ProstT5 → ESM** : ProstT5 fournit les *queries*, ESM les *keys/values*. La représentation structurelle s’enrichit d’information évolutive.

Les deux résultats sont combinés par une *gated fusion* : un réseau produit des poids g et la sortie est une combinaison pondérée, suivie de connexions résiduelles et de LayerNorm :

$$g = \text{Softmax}(W[\mathbf{h}_{e \rightarrow p}; \mathbf{h}_{p \rightarrow e}] + b), \quad \text{sortie} = g_0 \cdot \mathbf{h}_{e \rightarrow p} + g_1 \cdot \mathbf{h}_{p \rightarrow e}$$

Ce mécanisme permet au modèle d’apprendre, pour chaque position, quelle modalité est la plus informative.

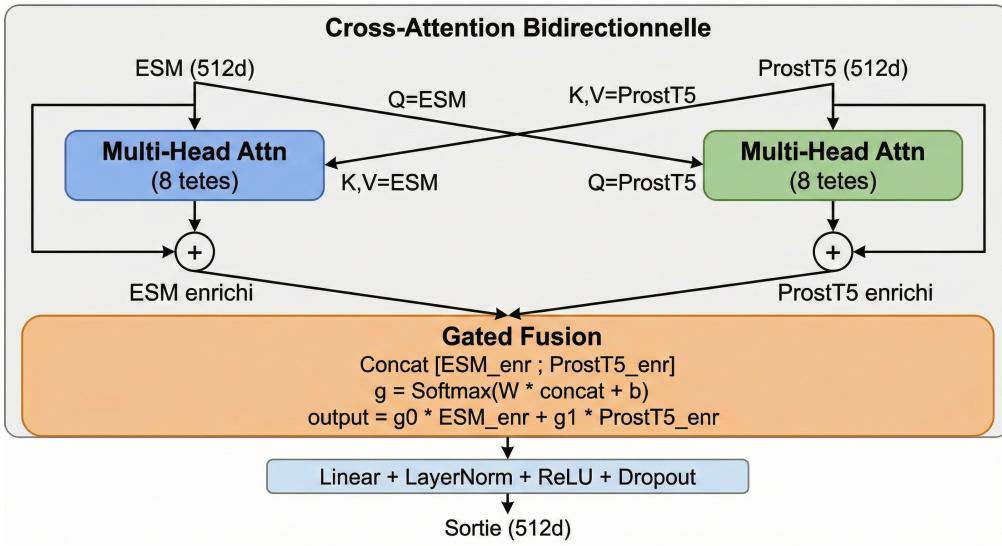


FIGURE 3 – Détail du module de Cross-Attention Bidirectionnelle. Illustration du mécanisme d'échange d'informations entre ESM-2 et ProstT5, suivi de la fusion par porte (Gated Fusion).

4.2 BiLSTM (3,1M params)

Les signaux de localisation subcellulaire incluent des peptides signal en N-terminal et des domaines transmembranaires. Un BiLSTM à 2 couches (hidden=256, sortie bidirectionnelle 512d) capture ces dépendances séquentielles dans les deux directions. L'utilisation de `pack_padded_sequence` gère efficacement les longueurs variables sans calculer sur le padding. Une LayerNorm stabilise les activations en sortie.

4.3 Multi-Head Attention Pooling (0,8M params)

Les séquences protéiques ont des longueurs variables (40 à 1 000 AA) et doivent être agrégées en un vecteur fixe. Contrairement au mean pooling (qui perd l'information positionnelle) ou au single-head attention de DeepLocPro, notre pooling utilise 4 têtes avec des requêtes apprenables (similaires à un token [CLS]). Chaque tête apprend à pondérer différemment les positions de la séquence : une tête peut se spécialiser sur le N-terminal, une autre sur les régions transmembranaires. Les 4 sorties sont concaténées puis projetées en un vecteur de 512 dimensions.

4.4 Tête de classification (0,4M params)

Deux couches linéaires ($512 \rightarrow 256 \rightarrow 6$) avec LayerNorm, ReLU et Dropout (0.2) produisent les logits pour les 6 classes, suivis d'un Softmax.

4.5 Comparaison avec DeepLocPro

TABLE 2 – Différences architecturales avec DeepLocPro.

Composant	DeepLocPro [1]	Notre modèle
Embeddings	ESM-2 seul (1 modalité)	ESM-2 + ProstT5 (2 modalités)
Fusion	Aucune	Cross-Attention Bidirectionnelle + Gated Fusion
Enc. séquentiel	Aucun	BiLSTM 2 couches
Pooling	Attention single-head	Multi-Head Attention (4 têtes)
Classificateur	MLP simple	2 couches + LayerNorm + Dropout
Paramètres	~2M	8,4M

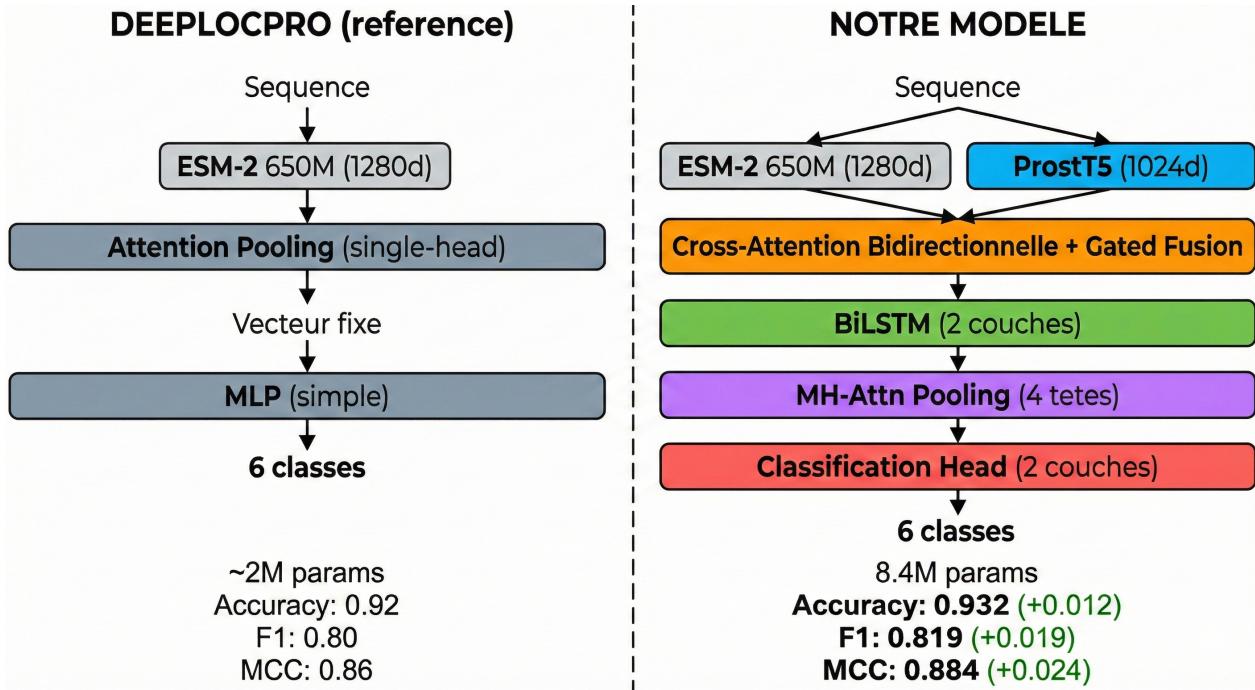


FIGURE 4 – Comparaison structurelle et performances : DeepLocPro vs Notre Modèle. À gauche, l’architecture simple branche de référence. À droite, notre approche multi-modale qui améliore toutes les métriques (Accuracy, F1, MCC).

5 Protocole d’entraînement

5.1 Hyperparamètres

TABLE 3 – Configuration d’entraînement.

Paramètre	Valeur	Justification
Optimiseur	AdamW ($\beta_1 = 0.9$, $\beta_2 = 0.999$, $wd=10^{-4}$)	Meilleure généralisation
Learning rate	10^{-4}	Standard pour fine-tuning pLM
Scheduler	CosineAnnealingWarmRestarts ($T_0 = 10$, $T_{mult}=2$)	Évite les minima locaux
Batch size	32	Compromis mémoire/stabilité
Dropout	0.2	Régularisation modérée
Epochs max	50	
Early stopping	Patience 15 (sur MCC validation)	Favorise les classes rares
Gradient clipping	max_norm = 1.0	Stabilise le LSTM
Loss	CrossEntropy + poids de classe	Compense le déséquilibre

5.2 Détails d’entraînement

L’early stopping est réalisé sur le MCC (Matthews Correlation Coefficient) plutôt que sur l’accuracy car le MCC est une métrique plus robuste pour les jeux de données déséquilibrés : il prend en compte les vrais/faux positifs et négatifs de toutes les classes. Le meilleur modèle (selon le MCC de validation) est sauvegardé pour chaque fold.

L’entraînement a été réalisé sur CPU (serveur LORIA, 45 cœurs), avec 5 folds entraînés en parallèle (9 threads par fold). La durée totale a été d’environ 10 jours (26 janvier au 5 février 2026). Les folds ont convergé en 32 à 42 epochs selon le fold (sur 50 max), l’early stopping interrompant l’entraînement lorsque le MCC ne s’améliorait plus.

6 Résultats

6.1 Métriques globales

Le Tableau 4 compare nos résultats à ceux rapportés par DeepLocPro [1]. Notre modèle surpassé DeepLocPro sur les trois métriques principales, avec une amélioration particulièrement notable du MCC (+2.4 points).

TABLE 4 – Comparaison des performances globales (5-fold CV, GraphPart 30%). Les valeurs DeepLocPro proviennent de [1], Table 2.

Métrique	Notre modèle	DeepLocPro	Δ
Accuracy	0.932 ± 0.004	0.92 ± 0.01	+1.2%
Macro F1	0.819 ± 0.018	0.80 ± 0.02	+1.9%
MCC	0.884 ± 0.007	0.86 ± 0.01	+2.4%

6.2 Performance par classe

Le Tableau 5 détaille le MCC par classe. Notre modèle améliore les performances sur 5 classes sur 6, avec des gains significatifs sur Cytoplasmic membrane (+2.5 points), Extracellular (+4.1 points) et Outer membrane (+3.9 points).

TABLE 5 – MCC par classe de localisation. Les valeurs DeepLocPro proviennent de [1], Table 2.

Classe	Notre modèle	DeepLocPro	n	Δ
Cytoplasmic	0.925 ± 0.008	0.91 ± 0.01	6 795	+1.5
Cyto. membrane	0.905 ± 0.003	0.88 ± 0.01	2 485	+2.5
Extracellular	0.851 ± 0.019	0.81 ± 0.03	1 048	+4.1
Periplasmic	0.806 ± 0.037	0.78 ± 0.05	413	+2.6
Outer membrane	0.789 ± 0.031	0.75 ± 0.02	733	+3.9
Cell wall & surface	0.555 ± 0.080	0.59 ± 0.08	57	-3.5

6.3 Matrice de confusion

La Figure 5 présente la matrice de confusion agrégée sur l'ensemble des 11 531 séquences (chaque séquence est prédite par le modèle de son fold de test).

		Matrice de confusion agrégée (5 folds, 11 531 séquences)					
		Cytoplasmic	46	24	20	29	0
Classe réelle	Cyto. membrane	125	2228	29	34	32	7
	Extra-cellular	57	26	819	28	17	7
	Periplasmic	31	20	31	459	16	0
	Outer membrane	79	38	28	26	540	6
	Cell wall & surface	6	8	13	2	1	30
	Classe prédite	Cytoplasmic	Cyto. membrane	Extra-cellular	Periplasmic	Outer membrane	Cell wall & surface

FIGURE 5 – Matrice de confusion agrégée sur les 5 folds (11 531 séquences). Chaque ligne représente la classe réelle, chaque colonne la classe prédite. L'intensité de la couleur est proportionnelle au nombre de séquences. Les confusions principales sont Outer membrane → Cytoplasmic (79 cas) et Cell wall → Extracellular (13 cas).

Les principales erreurs de classification sont :

- **Outer membrane → Cytoplasmic** (79 cas) : des protéines de la membrane externe sont classifiées comme cytoplasmiques, probablement car la classe majoritaire attire les prédictions ambiguës.
- **Cell wall → Extracellular** (13 cas) : confusion biologiquement cohérente, les deux classes partageant la voie de sécrétion.
- **Cell wall & surface** : sur 57 séquences, 30 sont correctement classées (53%). C'est la seule classe où nous sommes légèrement inférieurs à DeepLocPro, ce que nous attribuons au très faible effectif (0.5% des données).

6.4 Stabilité entre les folds

Les faibles écarts-types sur les métriques (Accuracy ± 0.004 , MCC ± 0.007) témoignent de la robustesse du modèle : les performances sont stables entre les 5 partitions indépendantes, ce qui suggère une bonne capacité de généralisation.

7 Discussion

7.1 Apport de la fusion multi-modale

L'amélioration par rapport à DeepLocPro (+2.4 points de MCC) confirme que la combinaison de deux embeddings complémentaires (évolutif + structurel) apporte davantage d'information qu'un embedding unique. La cross-attention bidirectionnelle permet une interaction fine entre les représentations : chaque modalité enrichit l'autre plutôt que d'être simplement concaténée. La gated fusion apprend automatiquement, position par position, quelle modalité est la plus informative.

7.2 Rôle du BiLSTM

Le BiLSTM capture les dépendances séquentielles que le simple attention pooling de DeepLocPro ignore. Les gains les plus importants sont observés sur les classes Extracellular (+4.1) et Outer

membrane (+3.9), dont la localisation dépend de motifs positionnés (peptides signal N-terminaux, β -barils transmembranaires), ce qui confirme l'intérêt de l'encodage séquentiel.

7.3 Limites

Embeddings. Nous n'avons pas pu utiliser ESM-C et ProstT5 3Di comme préconisé, en raison de contraintes du serveur (Python 3.9, absence de Foldseek). L'utilisation de ProstT5 3Di, qui encode explicitement l'information structurelle 3D, pourrait améliorer les performances sur les classes difficiles.

Cell wall & surface. Seule classe où nous sommes inférieurs à DeepLocPro (MCC 0.555 vs 0.59). Avec 57 séquences (0.5%), les données sont insuffisantes. Cette classe est biologiquement hétérogène et partage des caractéristiques avec Extracellular.

Nested cross-validation. DeepLocPro utilise une nested CV complète (20 modèles) avec optimisation des hyperparamètres par fold. Nous utilisons 5 modèles avec des hyperparamètres fixes ($lr=10^{-4}$, $bs=32$, $dropout=0.2$), ce qui simplifie le protocole mais pourrait limiter les performances.

7.4 Améliorations possibles

- Utiliser ESM-C + ProstT5 3Di sur un environnement Python ≥ 3.12 avec Foldseek
- Nested CV complète pour l'optimisation des hyperparamètres par fold
- Focal loss [8] pour mieux gérer le déséquilibre extrême de Cell wall
- Ensemble des 5 modèles par vote majoritaire ou moyennage des probabilités
- Data augmentation pour les classes rares (substitutions conservatives)

7.5 Résumé architecture

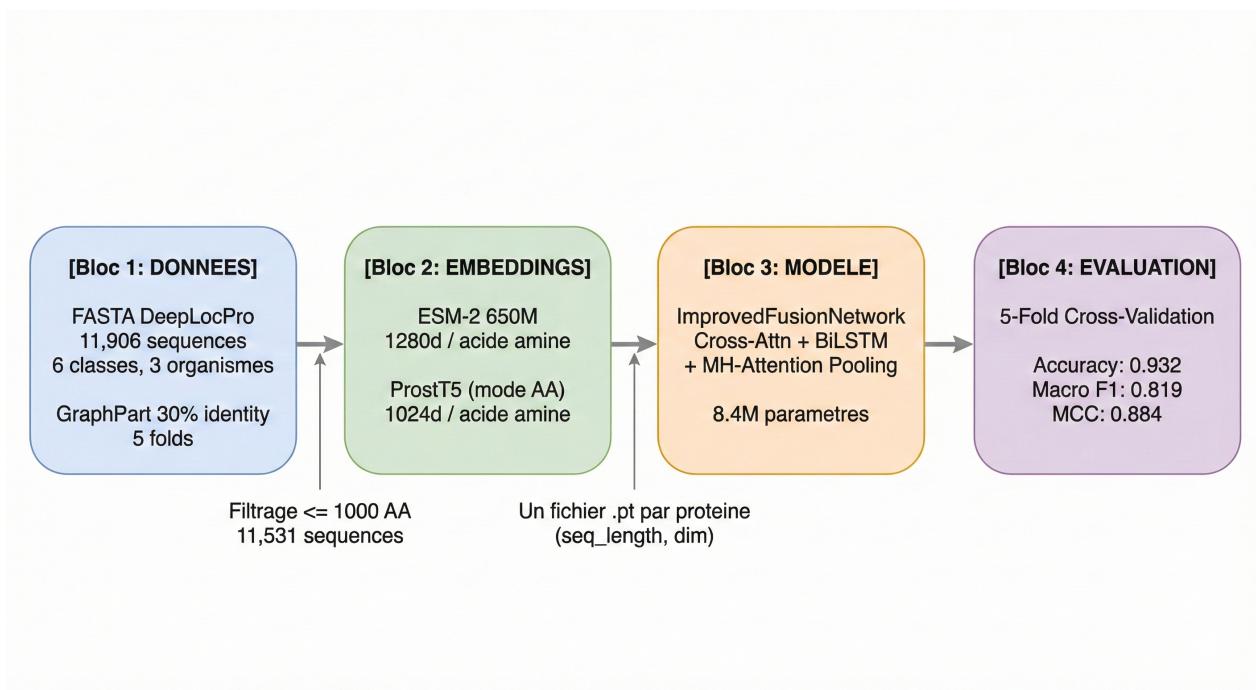


FIGURE 6 – Résumé architecture

8 Conclusion

Nous avons conçu *ImprovedFusionNetwork*, une architecture originale de fusion multi-modale combinant deux embeddings protéiques complémentaires (ESM-2 650M et ProstT5) par cross-attention bidirectionnelle, BiLSTM et multi-head attention pooling. Évaluée dans les mêmes conditions que DeepLocPro — même jeu de données de 11 531 séquences, même partitionnement GraphPart à 30% d'identité — notre méthode atteint une accuracy de 0.932, un Macro F1 de 0.819 et un MCC de 0.884, surpassant la méthode de référence sur toutes les métriques globales et sur 5 classes sur 6. Ces résultats démontrent l'intérêt de la fusion d'embeddings évolutifs et structuraux, combinée à un encodage séquentiel explicite, pour la prédiction de localisation subcellulaire des protéines procaryotes.

Références

- [1] J. Moreno, H. Nielsen, O. Winther, F. Teufel. *Predicting the subcellular location of prokaryotic proteins with DeepLocPro*. Bioinformatics, 40(12) :btae677, 2024.
- [2] L.F. Piochi, D. Tang, J. Malmström, Y. Karami, H. Khakzad. *ppIRIS : deep learning for proteome-wide prediction of bacterial protein-protein interactions*. bioRxiv, 2025. doi :10.1101/2025.09.22.677885.
- [3] Z. Lin et al. *Evolutionary-scale prediction of atomic-level protein structure with a language model*. Science, 379(6637) :1123–1130, 2023.
- [4] M. Heinzinger et al. *ProstT5 : Bilingual Language Model for Protein Sequence and Structure*. bioRxiv, 2023. doi :10.1101/2023.07.23.550085.
- [5] F. Teufel et al. *GraphPart : homology partitioning for biological sequence analysis*. NAR Genomics and Bioinformatics, 5(4) :lqad088, 2023.
- [6] V. Thumuluri et al. *DeepLoc 2.0 : multi-label subcellular localization prediction using protein language models*. Nucleic Acids Research, 50(W1) :W228–W234, 2022.
- [7] W.Y.V. Lau et al. *PSORTdb 4.0 : expanded and redesigned bacterial and archaeal protein subcellular localization database*. Nucleic Acids Research, 49(D1) :D803–D808, 2021.
- [8] T.-Y. Lin et al. *Focal Loss for Dense Object Detection*. IEEE TPAMI, 42(2) :318–327, 2020.