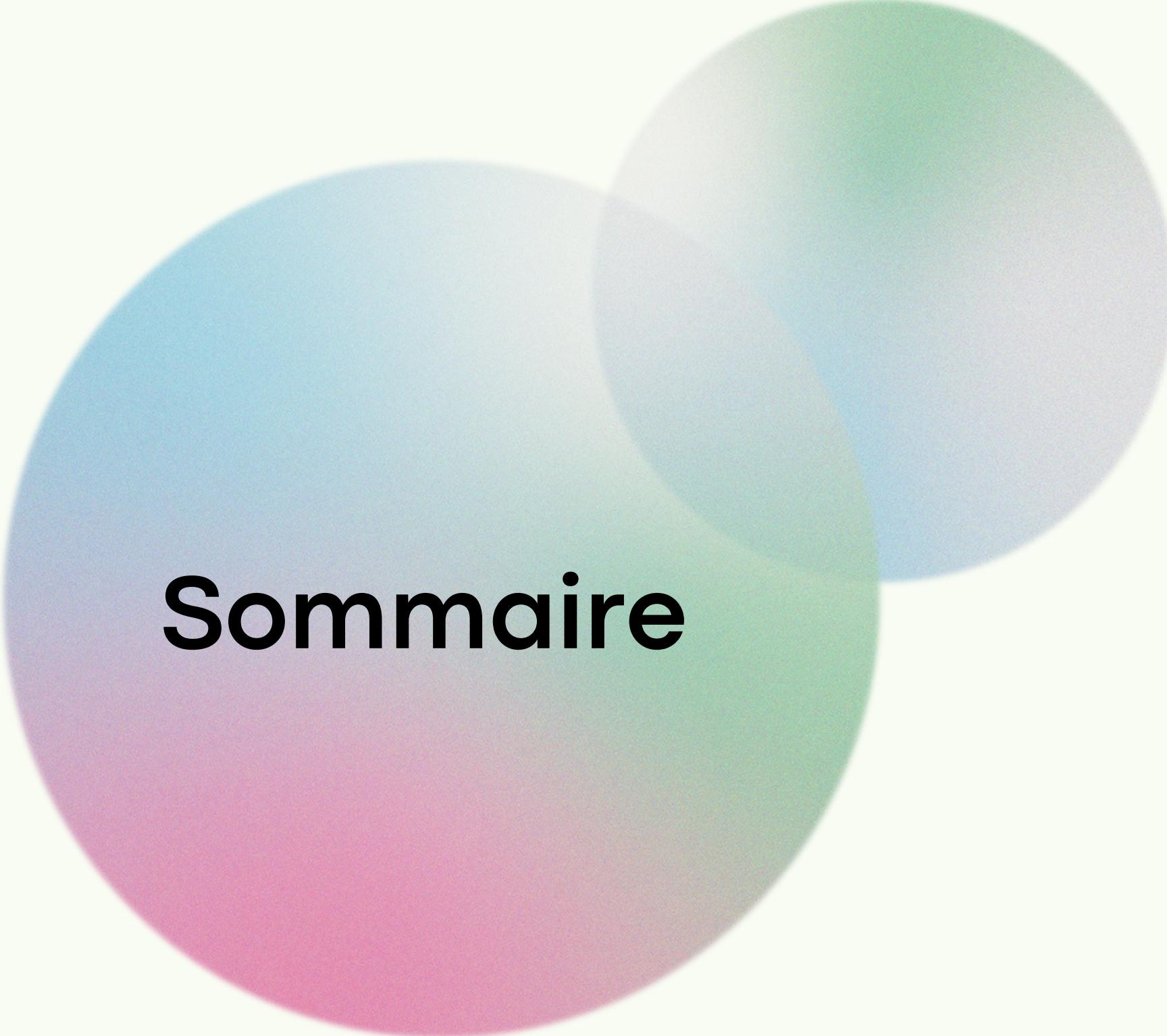


Prédiction de la structure secondaire des protéines

**Beaufils Constance
Boulet Faustine
Placier Moïse**

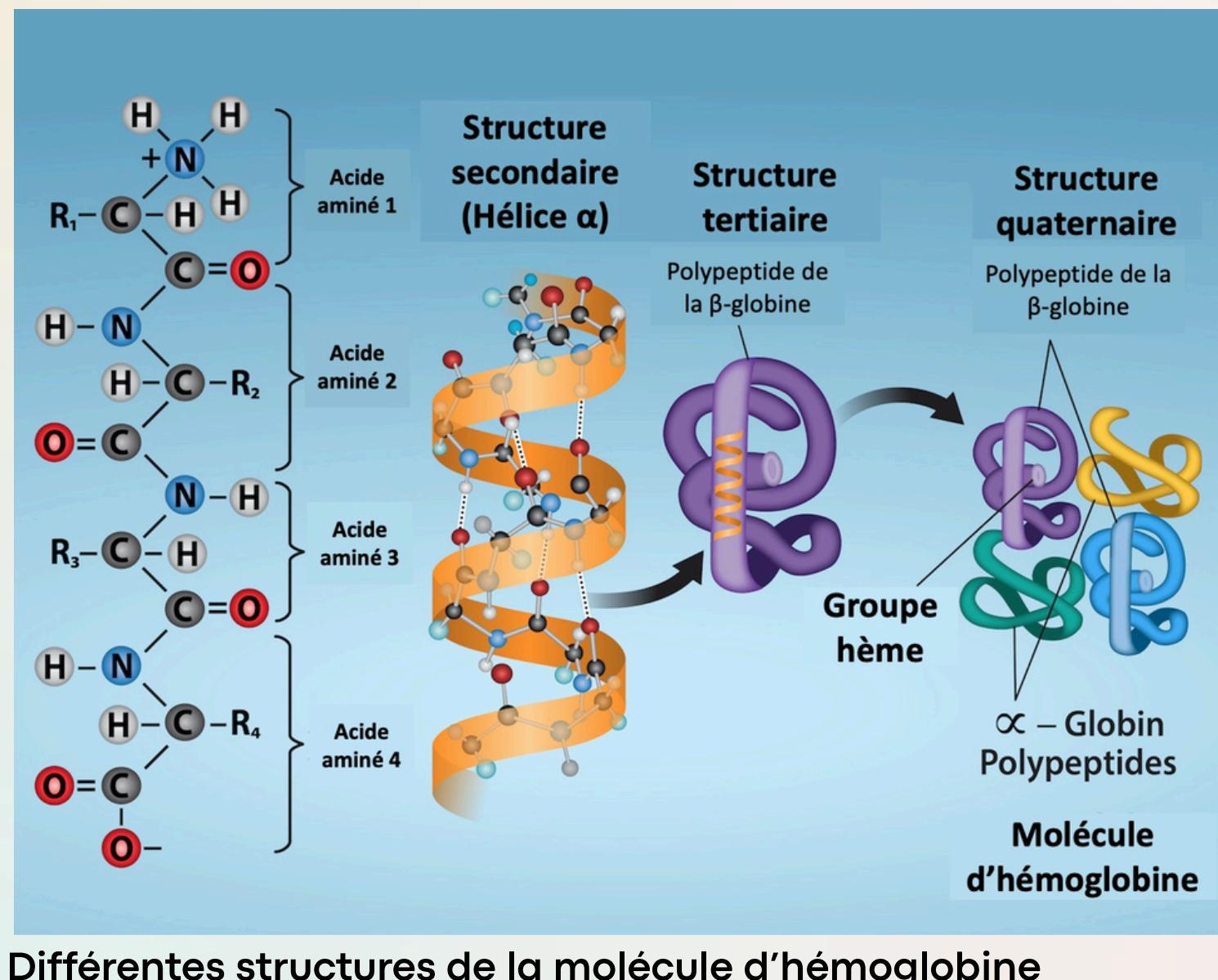
**École d'automne sur l'apprentissage
automatique en sciences du vivant**



Sommaire

- Définitions
- Enjeux
- Dataset
- Approches
 - Windows + Random Forest
 - PSSM + CNN
 - ProtBERT + Transformer
- Conclusion

Définitions : les différentes structures des protéines



Nom de la Structure	Définition	Stabilisée par...
Primaire	Séquence linéaire d'acides aminés	Liaisons peptidiques (covalentes)
Secondaire	Repliement local régulier (Hélices alpha et Feuilles beta)	Liaisons hydrogène dans le squelette
Tertiaire	Forme tridimensionnelle globale unique de la chaîne	Interactions des chaînes latérales (R) (ponts disulfure, ioniques, hydrophobes)
Quaternaire	Assemblage de plusieurs chaînes (sous-unités)	Toutes les interactions ci-dessus, entre les sous-unités

Pourquoi de telles prédictions ?

- La séquence primaire d'acides aminés détermine la structure 3D des protéines.
- La détermination expérimentale des structures 3D est coûteuse et longue.

Structure secondaire ?

- Simplification de la tâche : output à 1 dimension & metrics d'évaluation facile
- Les grands principes pour la prédiction des structures secondaires sont transposables pour les prédictions de structures tertiaires.



Difficulté de prédire la structure tertiaire

Anfinsen's dogma (1973)

La séquence d'acides aminés (structure primaire) détermine entièrement la structure 3D (structure tertiaire) de la protéine.

MAIS

Levinthal paradox (1968)

Pour un peptide de 101 acides aminés (un résidu = 3 états stables)
=> 3^{100} conformations possibles

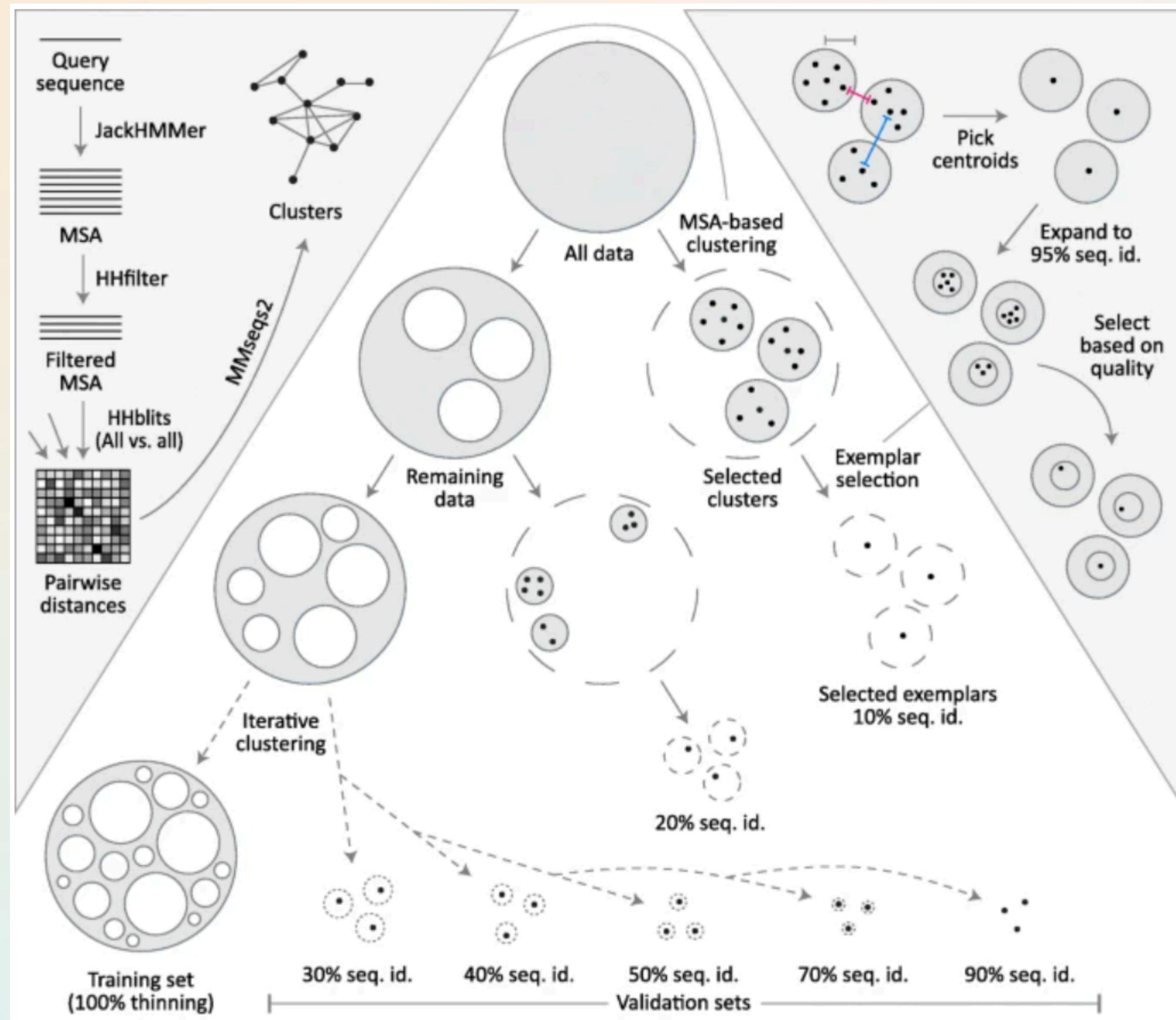
Dataset : ProteinNet Casp8

Caractéristiques des Datasets de Protéines :

- **Interdépendance Évolutive (Non-I.I.D.)** : les protéines partagent des relations évolutives
- **Identité Structurelle et Séquentielle Élevée** : deux protéines liées peuvent être plus de 90 % identiques au niveau de leur séquence d'acides aminés VS l'imagerie où deux classes similaires sont distinctes au niveau des pixels

Risque :

- **Overfitting** : le modèle ne fait que mémoriser des séquences très similaires sans apprendre la fonction de repliement



Test set ?

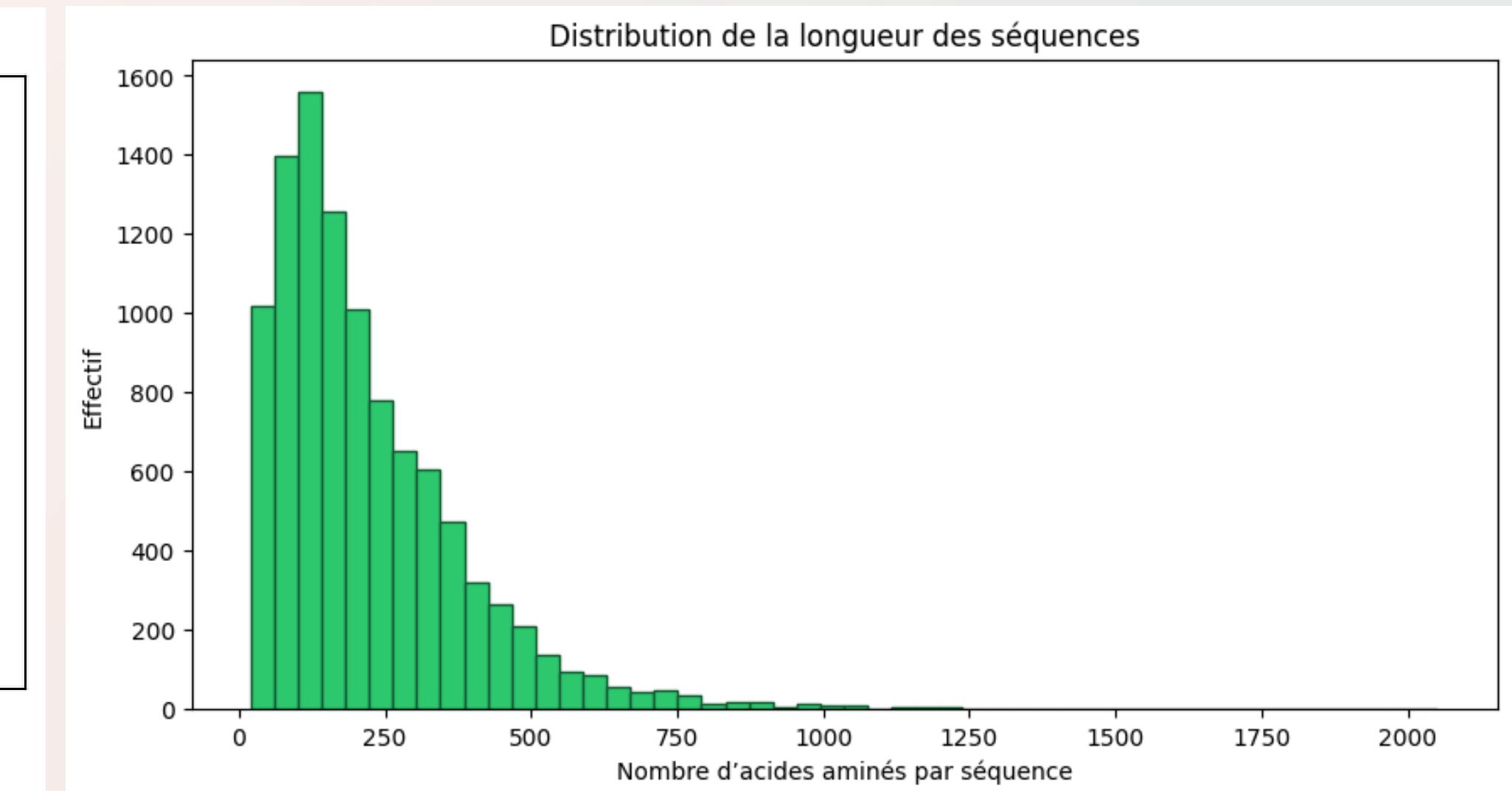
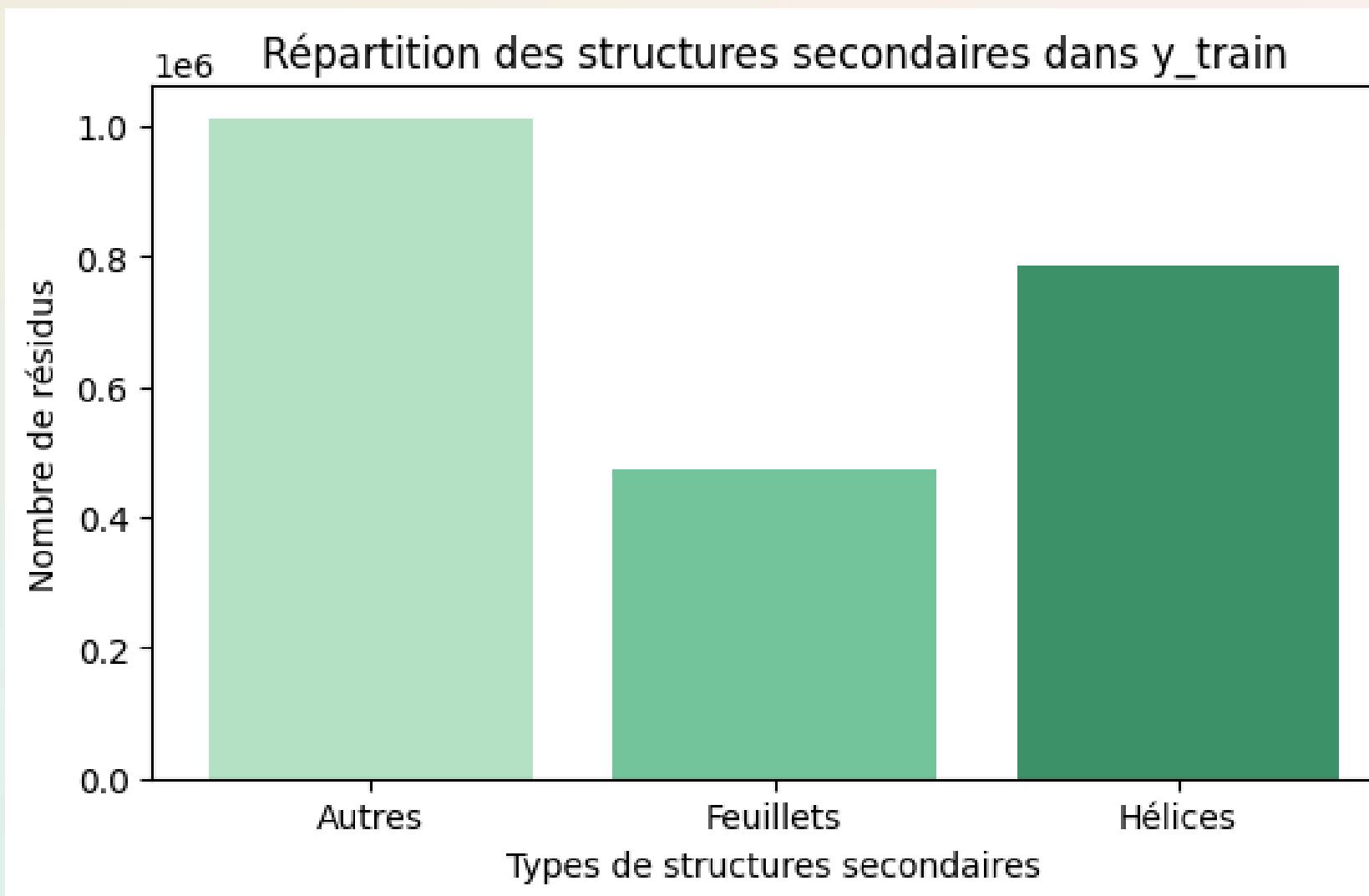
Compétition CASP !
(Critical Assessment of Structure Prediction)

= structures de protéines qui ne sont pas encore connues

ProteinNet : a standardized data set for machine learning of protein structure

Caractéristiques du dataset

10 174 Séquences
2 270 581 Acides Aminés



Longueur moyenne : 223, 17 AA

Simplification du nombre de classes

8 class symbols	8 class names	3 class symbols	3 class names
H	α -helix	H	Helix
L	Loop/Irregular	C	Coil/Loop
T	B-Turn	C	Coil/Loop
S	Bend	C	Coil/Loop
G	310-helix	H	Helix
B	B-bridge	E	Sheet
I	π -helix	H	Helix

8 classes → 3 classes

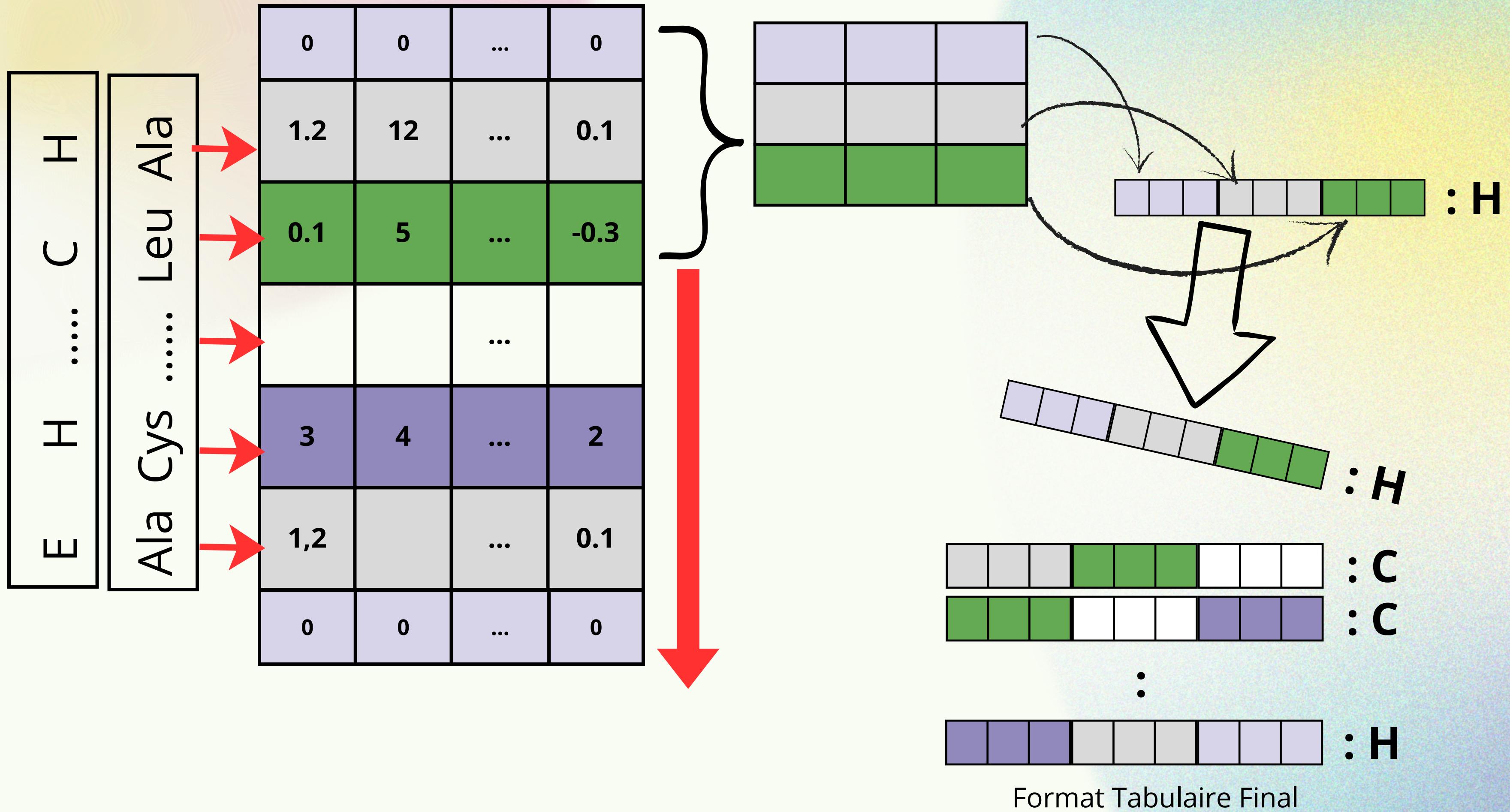


Approche locale : Random Forest sur Features Physico- Chimiques

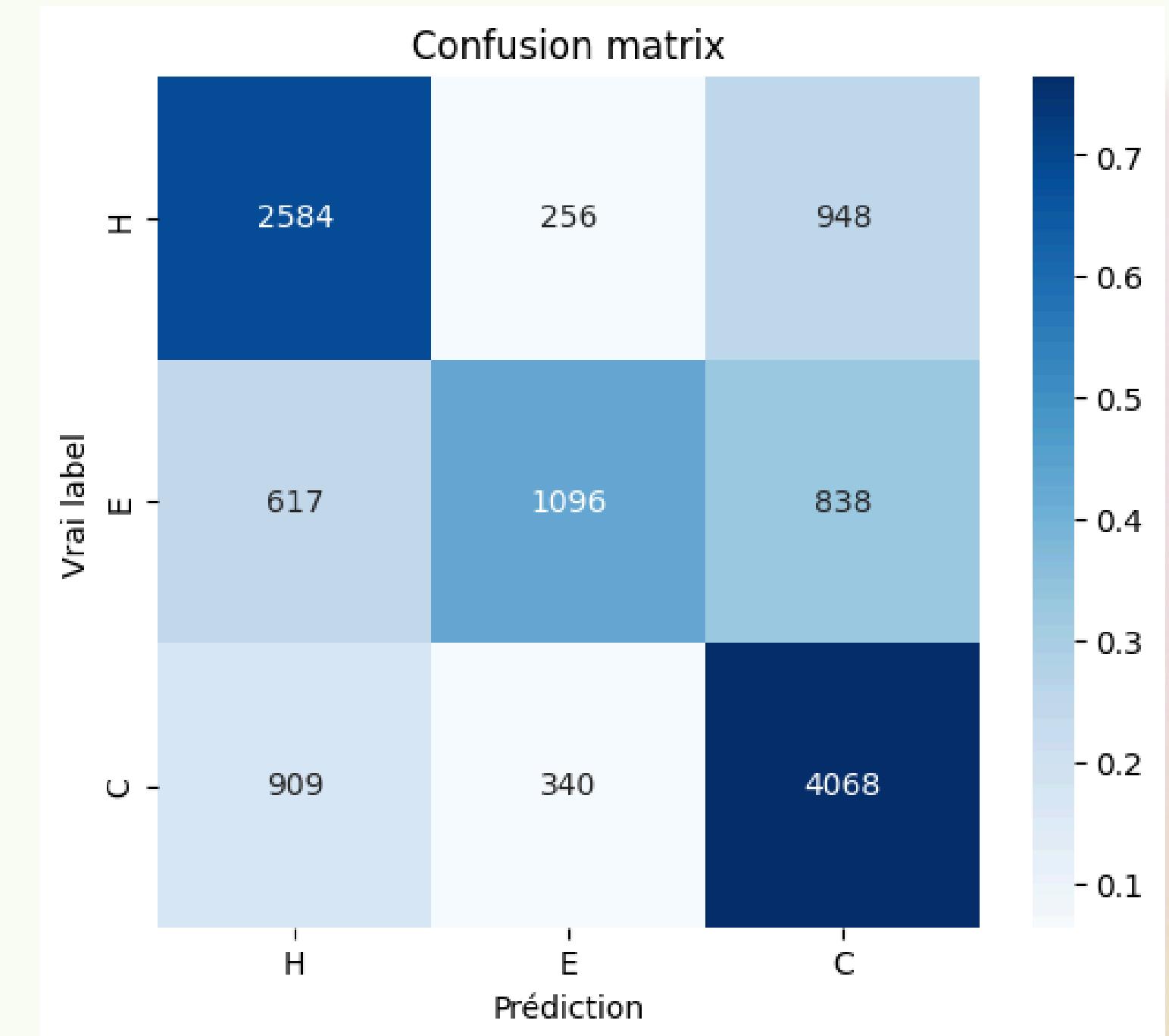
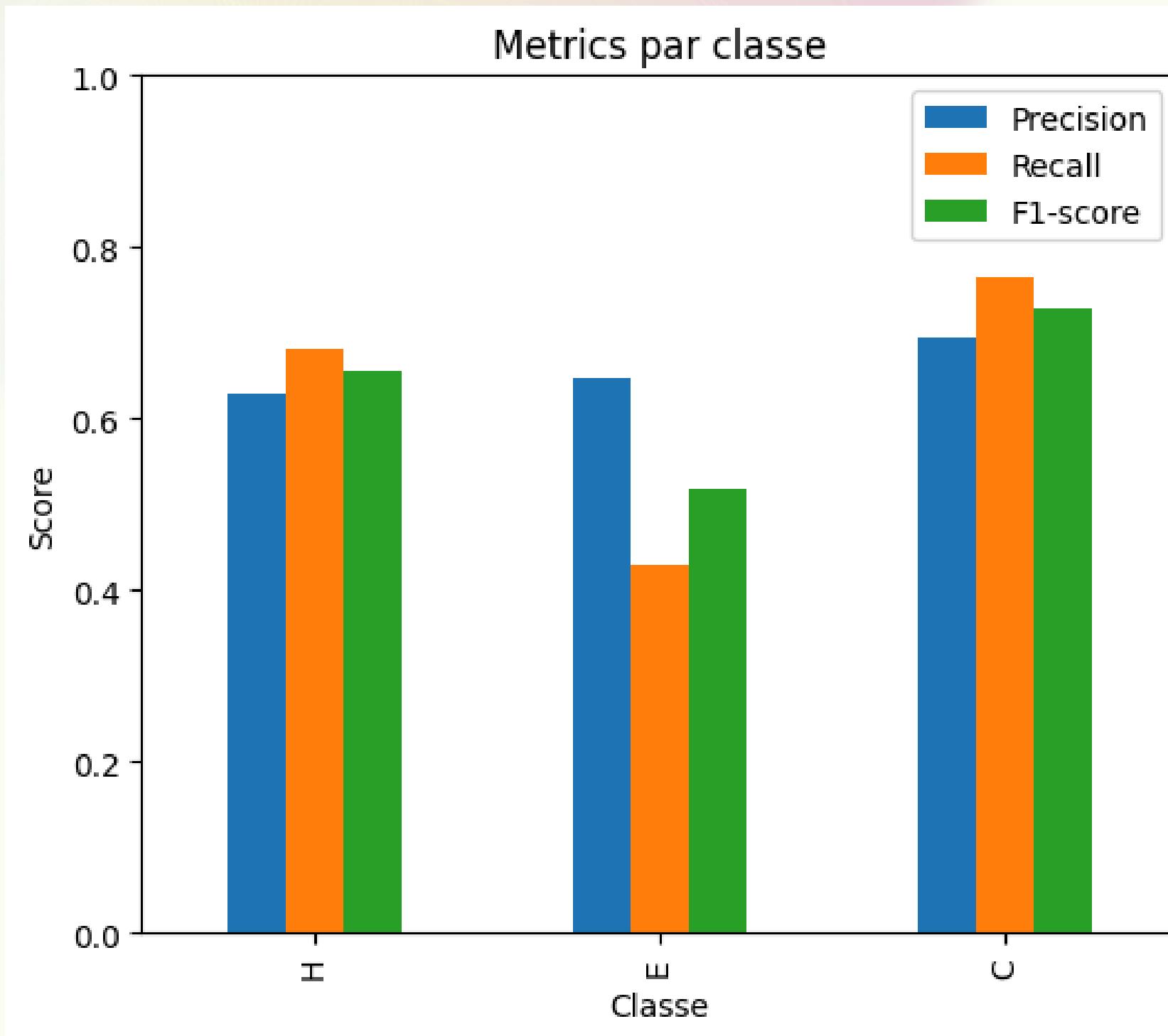
Embedding des acides aminés



Fenêtre glissante



Résultats et discussion



Limites de la méthode :

- RF invariant par permutation des features
- Prédiction à l'échelle du résidu

séquence cible ['C' 'H' 'H' 'H' 'H' 'H' 'H' 'C' 'C' 'C' 'C' 'E' 'E' 'E' 'E']

séquence prédite ['C' '**H**' '**H**' '**H**' 'C' 'C' 'E' 'C' 'C' 'E' 'E' 'E' 'E' 'E' 'E']

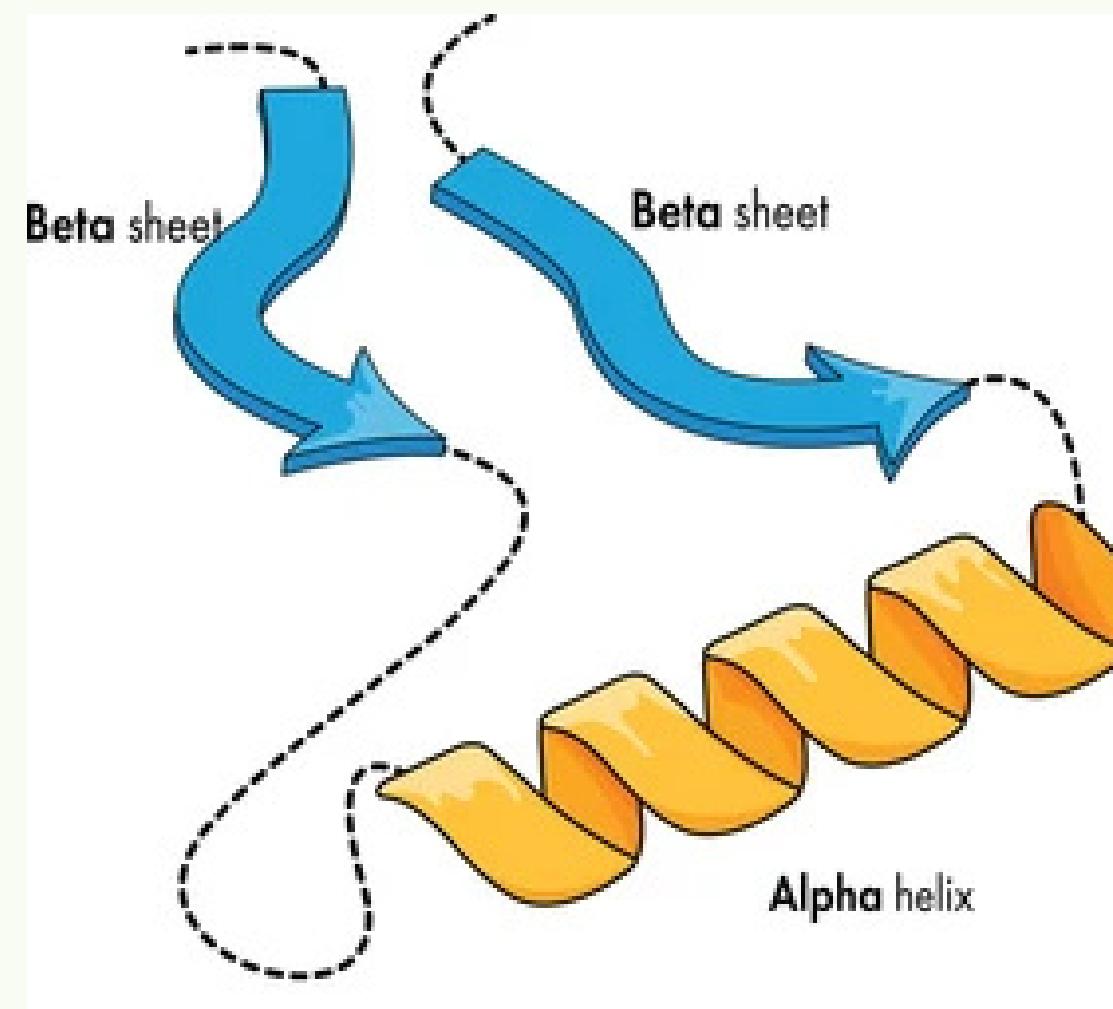
=> ne reflète pas les motifs réels

Limites de l'input biologique :

- Pas spécifique à la position
- Information redondante

Limites de l'approche locale :

- les feuillets beta ne sont pas des structures locales
- 35 % des structures secondaires sont déterminées par des interactions à longue distance (Robert L. Baldwin et al)





**Approche séquentielle :
CNN + features évolutives**

L'input biologique : one-hot encoding + PSSM

One-hot encoding

Vecteur : chaque acide aminé représenté par un vecteur binaire de longueur 20 (car 20 acides aminés)

Encodage : dans ce vecteur, une seule position correspondante à l'AA est mise à 1, tandis que toutes les autres positions sont à 0



PSSM (Position-Specific Scoring Matrix)

= information évolutive positionnelle spécifique (pour chaque position i et acide aminé a, une probabilité ou score de substitution $P(a/i)$ dérivé d'un MSA = Multiple Sequence Alignment)

Protein sequence	Sequence one-hot encoding
M	0 0
S	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0
T	0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0
Q	0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Y	0 1 0 0
T	0 1 0 0
Y	0 1 0
E	0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
Q	0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
I	0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
:
A	1 0



PSSM generation

MSA for a protein segment of length 5

Sequence	1	2	3	4	5
S1	A	L	K	A	V
S2	A	L	R	A	V
S3	A	I	K	A	I
S4	A	L	K	A	V

AA ↓ / Pos →	1	2	3	4	5
A	$\log(4/4 / 0.05) = \log(20) = 3.00$	0	0	3.00	0
L	0	$\log(3/4 / 0.05) = 2.30$	0	0	0
I	0	$\log(1/4 / 0.05) = 1.61$	0	0	1.61
K	0	0	2.71	0	0
R	0	0	1.61	0	0
V	0	0	0	0	2.30

Position	A	L	I	K	R	V
1	4	0	0	0	0	0
2	0	3	1	0	0	0
3	0	0	0	3	1	0
4	4	0	0	0	0	0
5	0	0	1	0	0	3

$$PSSM(a, i) = \log \left(\frac{P(a|i)}{P(a)} \right)$$

Architecture du CNN 1D

Préparation et encodage des données

Vecteur de caractéristiques : concaténation
encodage one-hot et profil PSSM

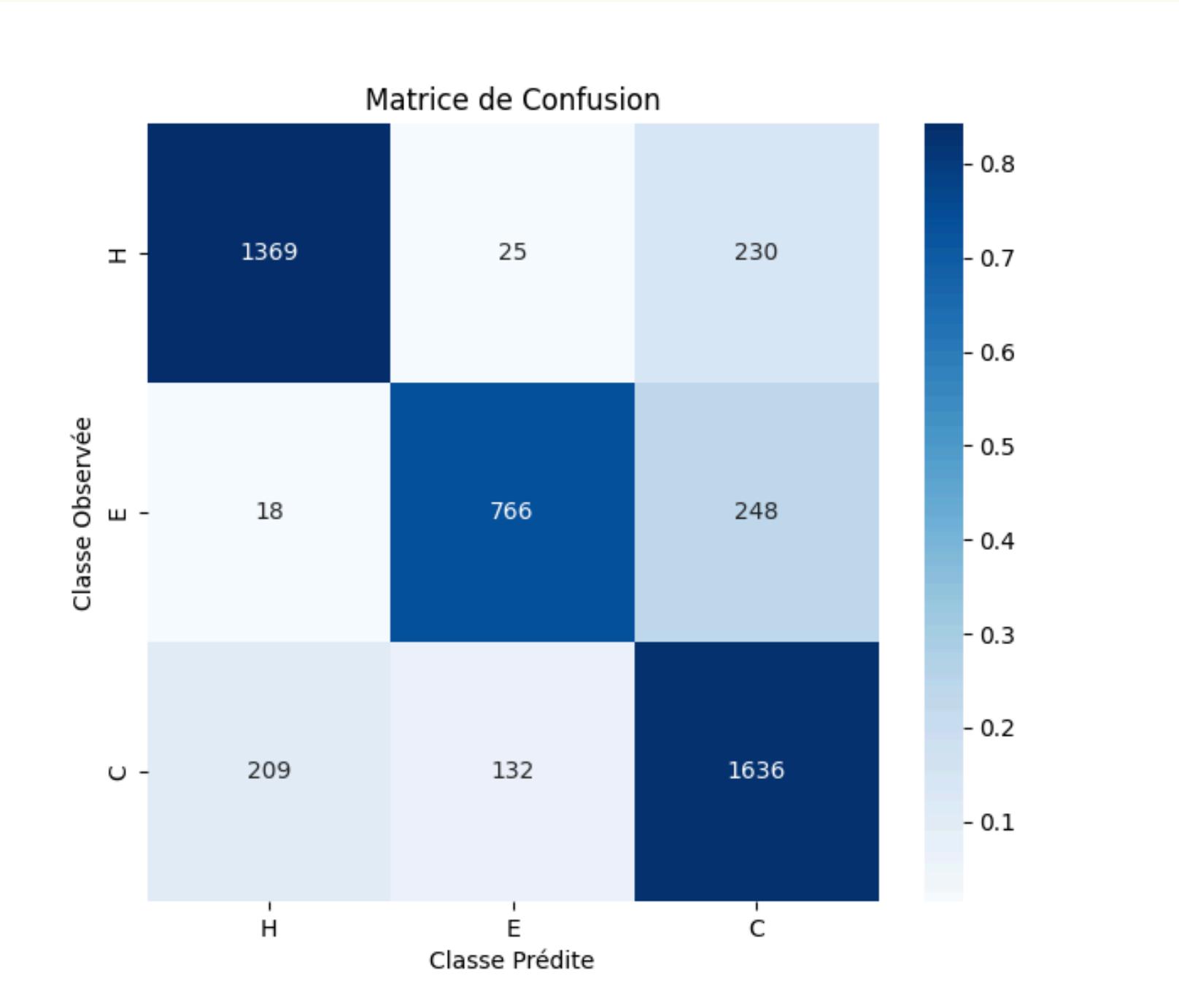
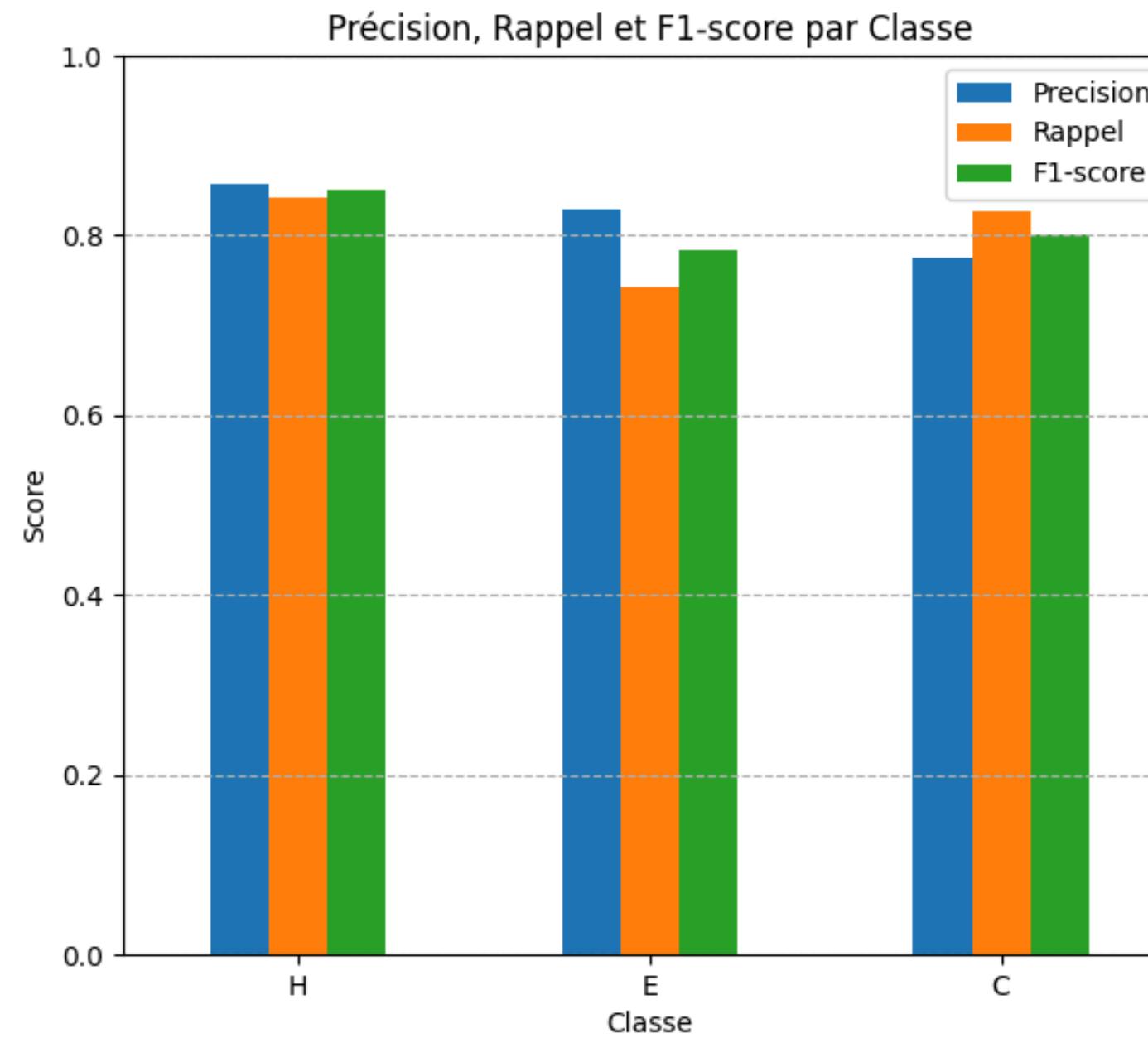
Padding à la longueur maximale du jeu de
données



Architecture du réseau

- Couches convolutionnelles successives : 3 couches de convolution 1D
- Extraction multi-échelle : tailles de noyaux multiples (3, 7 et 11) pour capturer différentes longueurs de motifs structurels locaux
- Couche de classification finale

Résultats et discussion



Q3 Accuracy = 82.56 %

Avantage de la méthode :

- Extraction de Motifs Locaux Supérieure : utilisation de filtres qui balayent la séquence, ce qui permet d'identifier des motifs locaux et des patterns conservés avec plus de puissance que les méthodes classiques comme Random Forest

Limites de l'input biologique :

- Problème de Localité Persistant : les CNN 1D restent intrinsèquement des modèles locaux, malgré les noyaux larges, ils peinent à capturer l'effet des interactions réelles entre résidus très éloignés dans la séquence primaire
- Signal PSSM Non Contextuel : les profils PSSM fournissent une information position-spécifique riche, mais cette information est non-contextuelle par nature (elle ne tient pas compte des résidus environnants) → l'effet des résidus éloignés n'est donc pas explicitement représenté
- La PSSM n'encode pas explicitement les co-évolutions qui renseignent sur des interactions structurales distantes entre AA.



Exploitation d'un Modèle de Langage Protéique Profond : ProtBERT

ProtBERT

Origine :

- Adaptation de BERT (NLP) au domaine des séquences protéiques.

Principe :

- Apprentissage auto-supervisé par Masked Language Modeling (prédirer des acides aminés masqués).

Entrée :

- Séquence d'acides aminés (tokens).

Sortie :

- Embedding contextuel (vecteur de 1024 dimensions par résidu).

Forces :

- Capture des dépendances locales et globales dans les séquences.
- Encode motifs structuraux et fonctionnels.
- Réutilisable pour diverses tâches (classification, prédiction de structure, etc.).

Utilisation typique :

- Extraire les embeddings puis ajouter une tête de classification adaptée à la tâche.

Transformer

Mécanisme Clé = Multi-Head Self-Attention :

Chaque token pondère l'influence de TOUS les autres tokens de la séquence, sans dépendre de leur proximité physique.

Avantages Majeurs :

- Contexte global :

Capture les dépendances à longue portée dans les données séquentielles, garantissant que la représentation de chaque token est informée par l'intégralité du contexte.

- Vitesse :

Le calcul de l'attention est parallélisable (non-séquentiel), ce qui permet un entraînement beaucoup plus rapide sur les GPU par rapport aux architectures récurrentes.

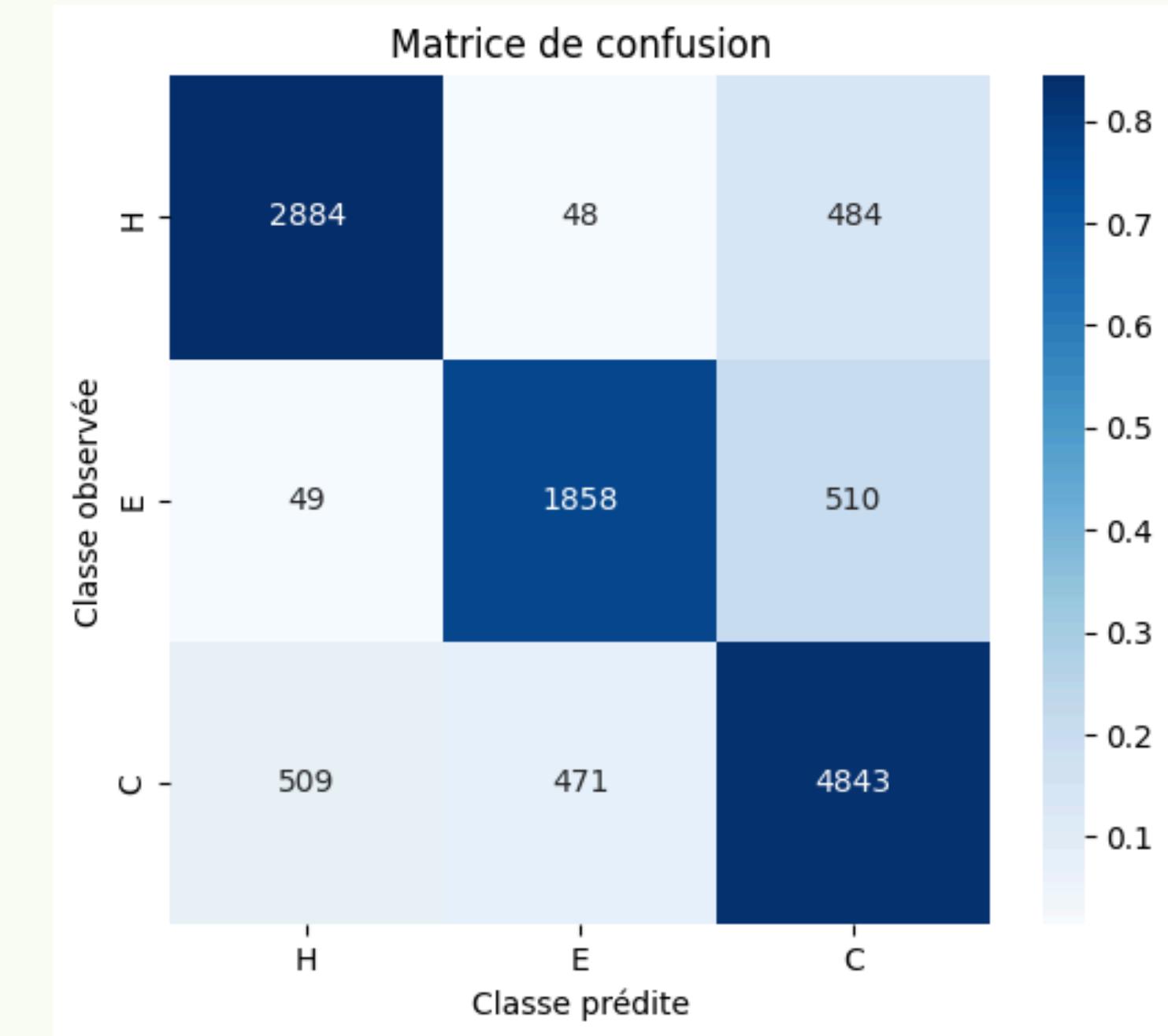
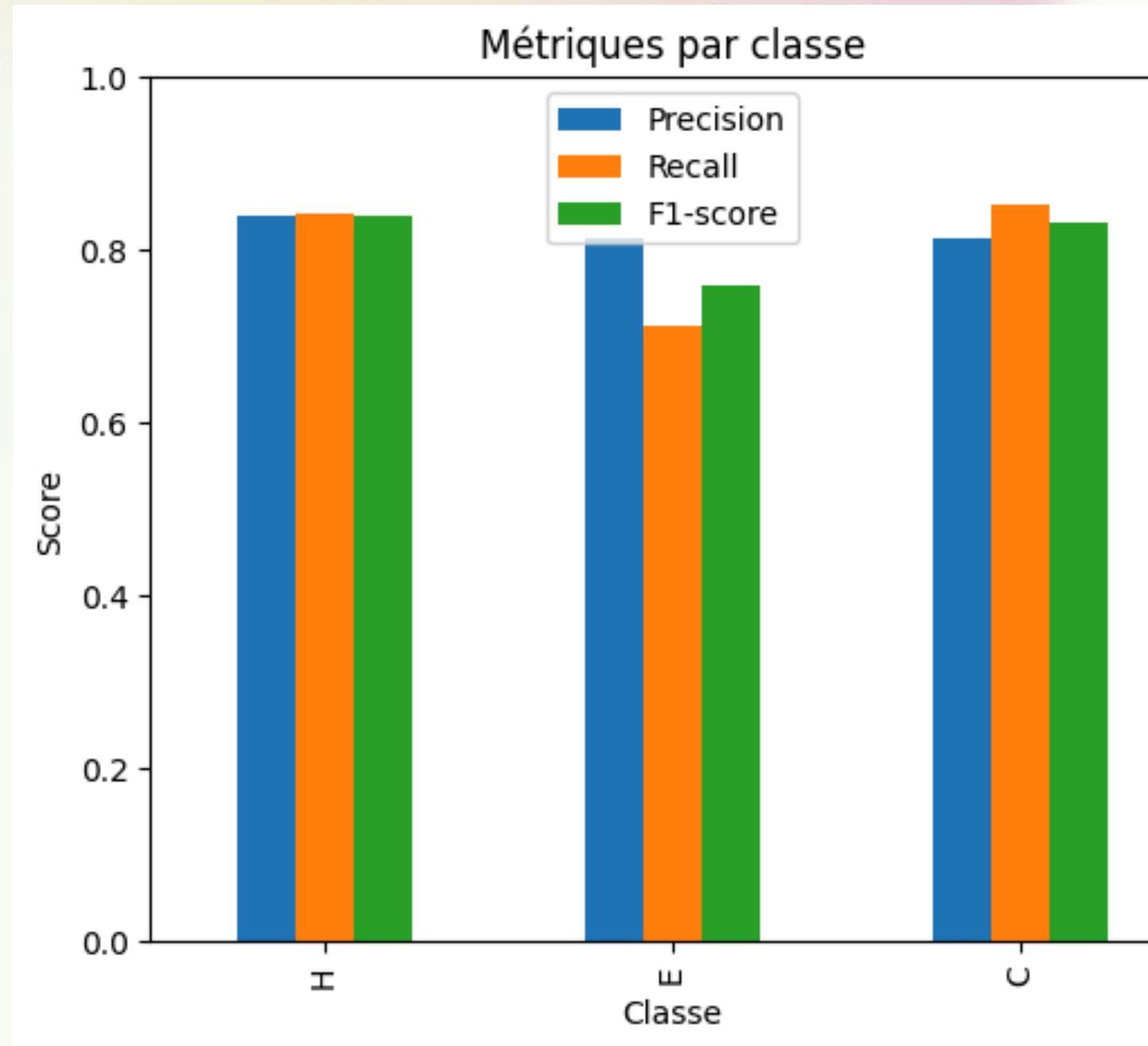
- Richesse :

Les "multi-têtes" apprennent différents types de relations ou de motifs simultanément.

Pipeline

Étape	Composant	Action Clé	Format des Données
Entrée	Séquence Protéique	Fournir la donnée brute	Séquence d'Acides Aminés (e.g., ATGEK...)
Encodage	ProtBERT	Générer des représentations contextuelles (Embeddings)	Vecteurs 1024-dimensions
Modélisation	Transformer	Apprendre les dépendances résidu-résidu et les motifs spécifiques H/E/C	Vecteurs 256-dimensions
Sortie	Classifieur Linéaire	Prédire l'état (H/E/C) pour chaque résidu	Prédiction H/E/C

Résultats et discussion



ProtBERT-Transformer : performance de prédiction = Q3 de 82.23 % et F1-Macro de 0.82%..

Conclusion

Modèle	Entrée principale	Q3 Accuracy
Random Forest	Descripteurs Physico-chimiques (Locaux)	66.5
PSSM - CNN	PSSM	82.56
ProtBERT-Transformer	Séquence Brute (Embeddings 1024-D)	82.23

Les modèles locaux/physico-chimiques (Random Forest, Q3=66.5) sont limités par l'absence d'information non-locale, tandis que l'incorporation d'information biologique supplémentaire (via PSSM pour le CNN, Q3 =82.5, ou via embeddings pour ProtBERT, Q3 = 82.2) est le déterminant critique pour une bonne prédiction de structure secondaire.

Merci !

Avez-vous des questions ?

