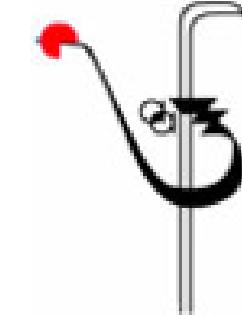
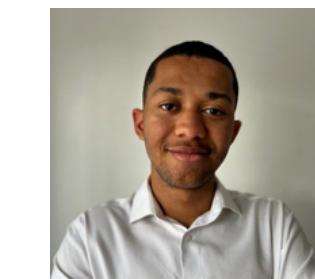
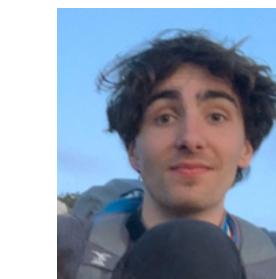


Analyse de données pour la **prédiction des valeurs nutritionnelles des fourrages** pour ruminants par **apprentissage automatique** et **Grands Modèles de Langue (LLM)**

Un projet de l'**Association Française de Zootechnie (AFZ)** encadré par Valérie Heuzé et Gilles Tran



Réalisé par
Aristide Lauront, Matéo Petitet,
Raphaël Genin et Raphaël Rubrice



Plan de la présentation

I/Des tables INRAe aux valeurs nutritionnelles ? 3/Résultats

- enjeux - problématique
- présentation des données à disposition
- performance des modèles
- cas d'utilisation concret

2/D'une approche non neuronale aux grands modèles de langue (LLM)

- exploitation des valeurs chimiques
- couplage avec des valeurs textuelles structurées
- passage au LLM

4/Discussion

- impact des valeurs numériques
- impact des valeurs textuelles
- analyse des poids

II. Des tables INRAe aux valeurs nutritionnelles ?

Enjeux : la prédiction des valeurs nutritionnelles

Composition des rations :

- impacte le rendement
- impacte le bien-être animal

Mais coûts variables et pouvant très vite augmenter --> intérêt de maximiser le rapport qualité/prix

Pour aider à l'élaboration de rations, l'INRAE a développé un système de caractérisation des rations selon leur part d'énergie ou de nutriment effectivement disponible pour l'animal

Pourquoi ce besoin ?

- calcul exact coûteux
- description du fourrage complexe
- fortes variation

Pour quel usage ?

- adaptation à la volée des rations
- suivi précis des apports

But : utiliser un **Grand Modèle de Langue (LLM)** afin de prédire plus précisément des valeurs nutritionnelles de fourrages

Objectif

Associer une description qualitative des fourrages à des caractéristiques chimiques mesurées sur place par infrarouge pour en déterminer les valeurs nutritionnelles

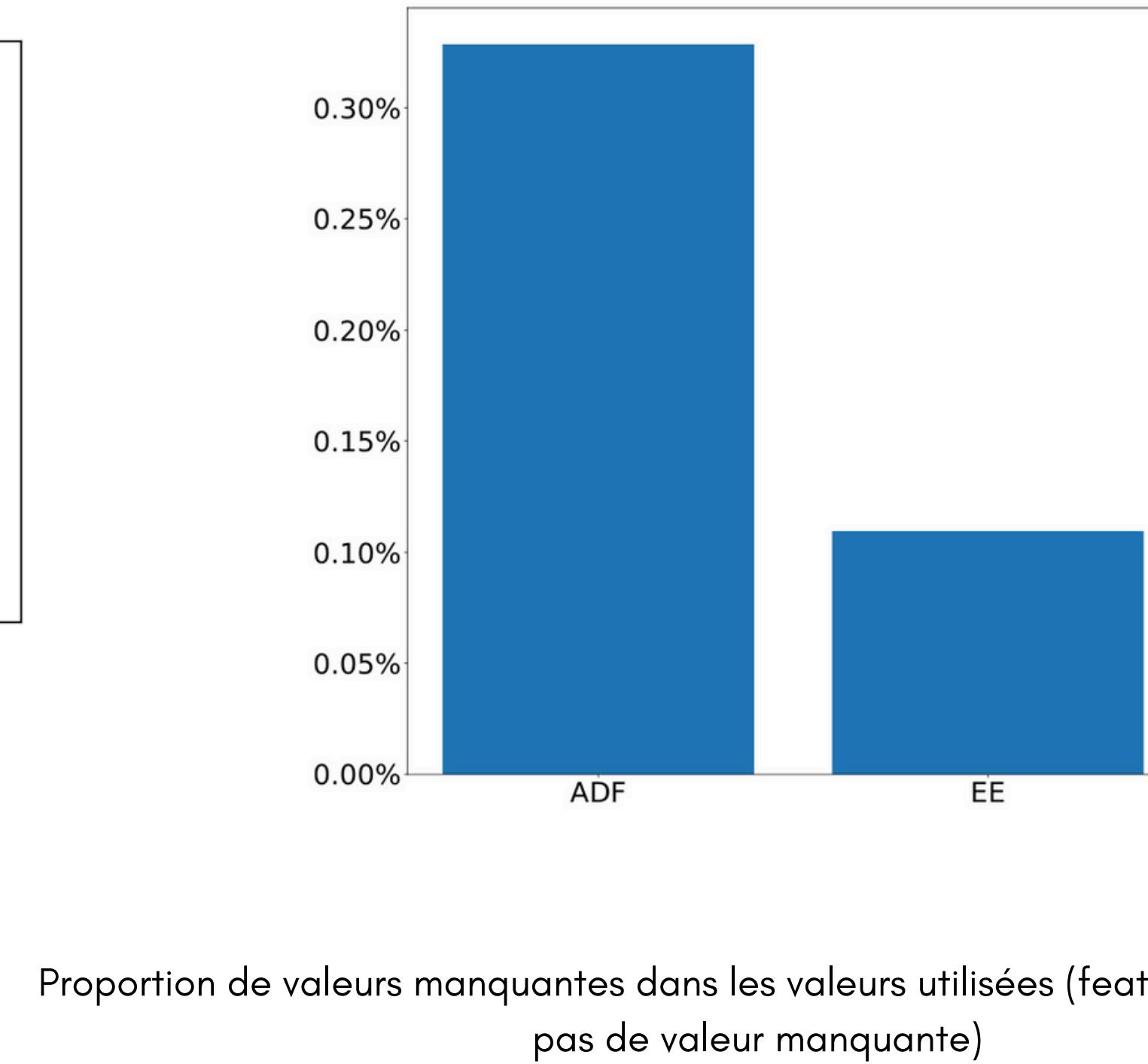
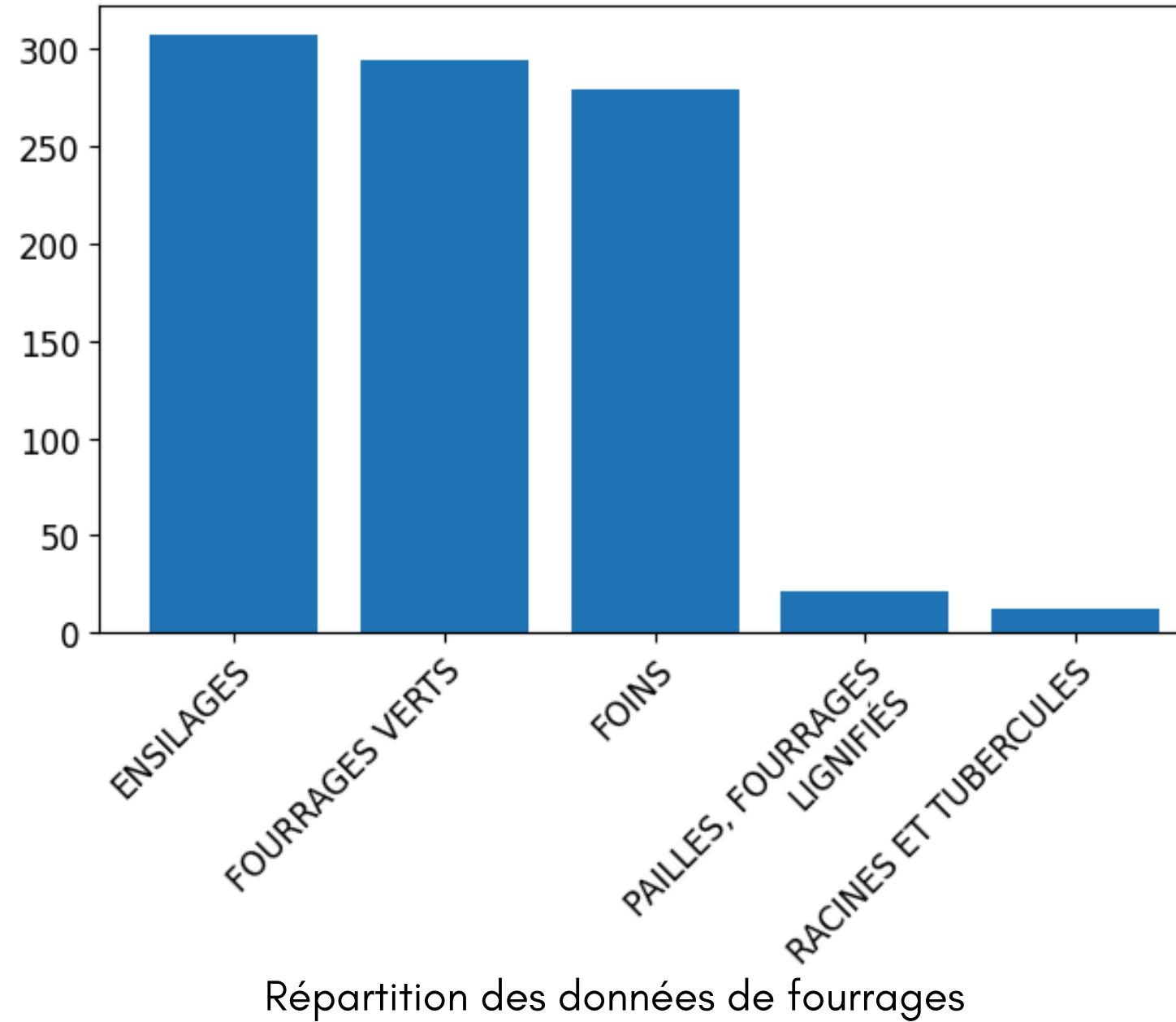
Le **LLM** permet ici l'utilisation d'une description en langage naturel

Nature des **données** utilisées

Des données issus des tables INRAE décrivant des caractéristiques de **fourrages** et de **concentrés** ont été fournis. Chaque aliment est caractérisé par :

- un numéro de ligne
- son **identifiant** INRAe unique
- **5 libellés** de plus en plus précis (du libellé 0, systématique, au libellé 4, facultatif) ; ces libellés décrivent qualitativement le fourrage
- **92 valeurs chimiques** mesurées en laboratoire ; nous cherchons à prédire 5 d'entre elles

Typologie des données



Focus sur les valeurs à prédire

Valeurs définies par l'INRAE

Valeurs énergétiques	Valeurs protéiques
UFL : Unité Fourragère Lait, quantité d'énergie nette absorbable pendant la lactation ou l'entretien du ruminant (1 UFL = 1700 kcal)	PDI : Protéines Digestibles dans l'Intestin, valeurs nutritives en azote (protéines métabolisables) chez les ruminants
UFV : Unité Fourragère Viande, quantité d'énergie nette absorbable lors de l'engraissement d'un ruminant (1 UFL = 1820 kcal)	PDIA : PDI d'origine Alimentaire, non dégradées dans le rumen BPR : Bilan Protéique du Rumen, différence entre les protéines ingérées et celles passant au duodénum

Focus sur les **variables explicatives** - Valeurs chimiques (1)

Valeurs chimiques mesurables facilement et à faible coût

- **MS** : Matière Sèche, le restant après retrait de toute l'eau du produit
- **MM** : Matière Minérale, portion non-organique du produit
- **MAT** : Matière Azotée Totale, protéines brutes
- **CB** : Cellulose Brute
- **NDF** : Neutral Detergent Fiber, fibres totales présentes dans un aliment
- **ADF** : Acid Detergent Fibers, fibres totales présentes dans un aliment sauf hemicellulose
- **EE** : Ether Extract, lipides brutes

Focus sur les **variables explicatives** - Libellés (2)

5 niveaux de précision sont donnés par les libellés

Général

Niveau 0 : catégorie générale de l'aliment considéré (ex : FOURRAGES VERT) ; 5 modalités

Niveau 1 : sous-catégorie de l'aliment (ex : PRAIRIES PERMANENTES, PLAINE (NORMANDIE)) ; 53 modalités

Niveau 2 : précisions sur les conditions de culture et/ou récolte de l'aliments (ex : 1er cycle) ; 42 modalités, présence non-systématique

Niveau 3 : précisions supplémentaires sur les conditions de culture et/ou récolte de l'aliments (ex : 15-25 avril, déprimage, ST = 172°C) ; 113 modalités, présence non-systématique, informations parfois de même nature que le niveau 2

Niveau 4 : informations complémentaires sur l'aliment (ex : Épiaison du dactyle) ; 55 modalités, présence non-systématique

Spécifique

Les libellés sont standards du niveau 0 à 2, et deviennent ensuite plus imprécis et inconsistants.

Données **additionnelle**

Fichiers excel d'entraînement des modèles insuffisamment dotés en vocabulaire dans les libellés pour entraîner un LLM ; ajout de la base **Feedipedia** (© INRAE CIRAD AFZ FAO) intégrale en français et anglais, fournissant un corpus détaillé relatif à l'alimentation animale.

Corpus se sous forme d'un fichier excel contenant des **noms et descriptions d'aliments** en français et en anglais --> **16 186** lignes de données présentées

II. D'une approche non neuronale aux grands modèles de langue (LLM)

Campagne d'optimisation des **modèles non-neuronaux**

- Validation croisée en 5 plis
- Test de toutes les combinaisons (produit cartésien prétraitements x régresseurs)

Pré-traitements

- Sélection d'un ou plusieurs libellés
- Encodage one-hot des libellés
- Avec ou sans reduction de dimension (ACM, AFDM)
- Avec ou sans valeurs chimiques



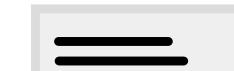
Régresseurs

- > Régression Ridge, Lasso Linéaire
- > Séparateur à Vaste Marge (SVR)
- > SVR non linéaire
- > Random Forests, Gradient Boosting Trees

Non linéaire

Approche d'apprentissage automatique non neuronale

Description du fourrage



Libellé 0 → 4



Libellé 1

GRAMINÉES FOURRAGÈRES, RAY-GRASS D'ITALIE, ALTERNATIF

GRAMINÉES FOURRAGÈRES

RAY-GRASS D'ITALIE

ALTERNATIF

Splitting

GRAMINÉES FOURRAGÈRES

(0, 1, ..., 0,0)

PRAIRIES PERMANENTES

(0, 1, ..., 0,0)

Encodage OneHot

Représentation
Vectorielle

Réduction de dimension
FAMD/ACM

Valeurs chimiques

- MS
- MM
- MAT
- CB
- NDF
- ADF
- EE



Standard Scaler

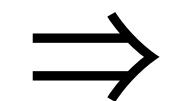
KNN Imputer

Modèle de machine learning

UFL

Limites des approches **non-neuronales**

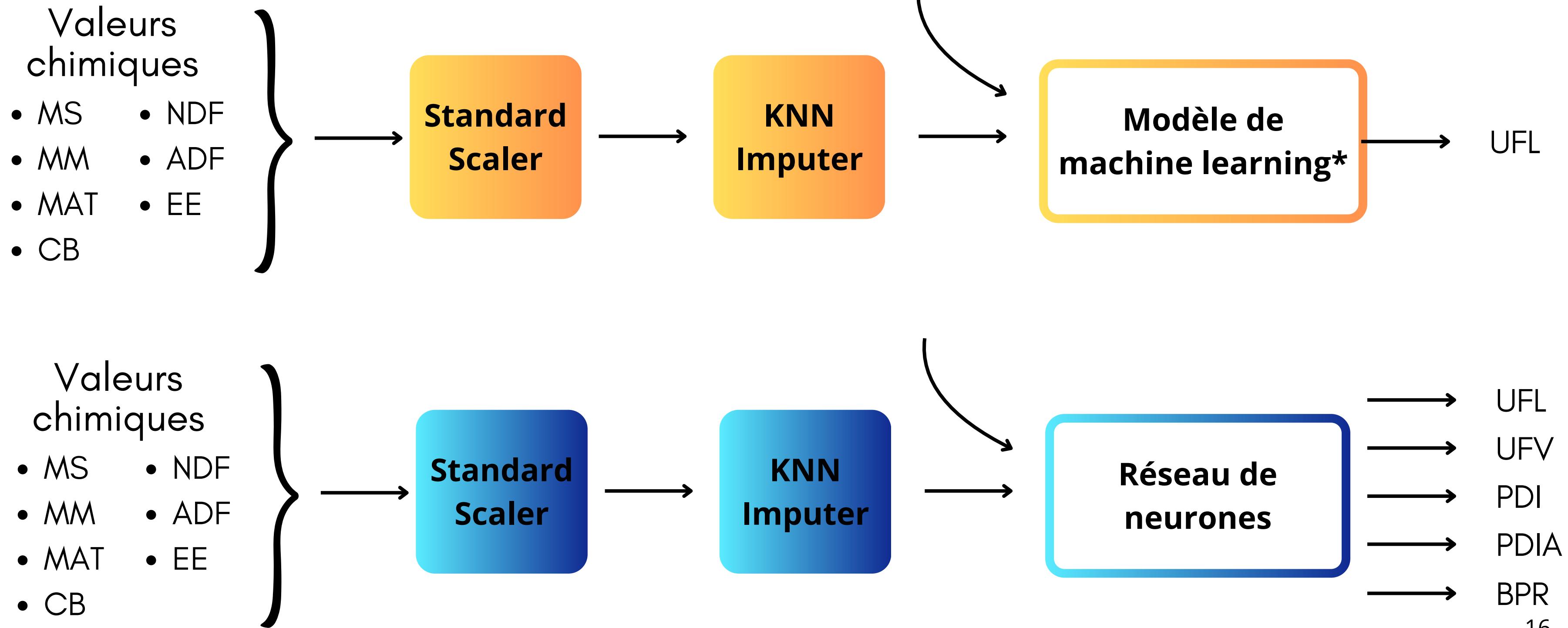
- Difficulté à capturer les dépendances complexes du langage
- Représentation limitée des mots et du contexte
- Généralisation limitée



Deep Learning

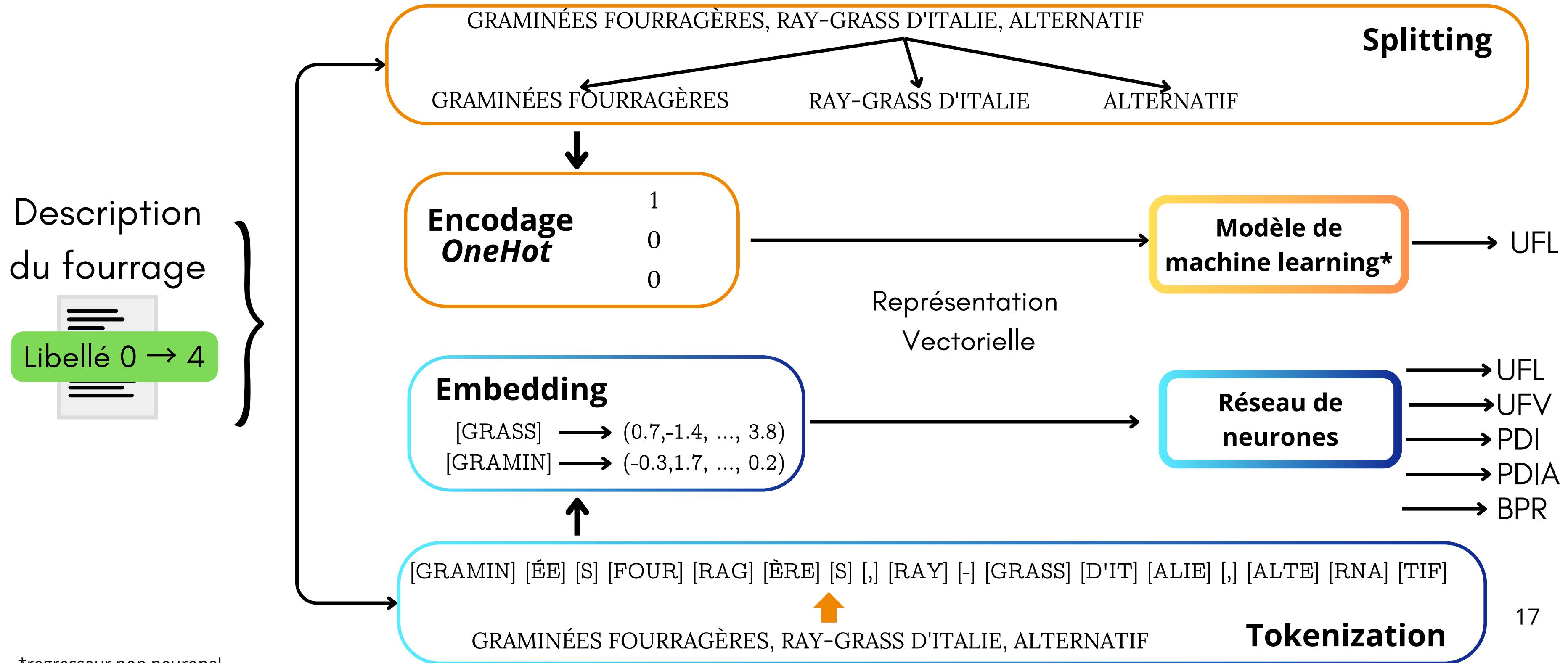
= Approches neuronales

Comparaison DeepLearning/Non neuronal (quantitatif)



*regresseur non neuronal

Comparaison DeepLearning/Non neuronal (textuel)

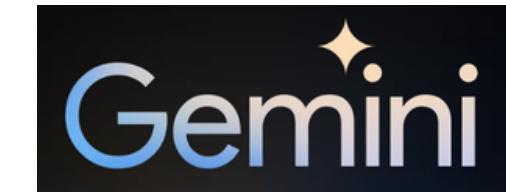


Revolution du traitement automatique de la langue naturelle : les *large language model (LLM)*

- Une nouvelle architecture : “Attention Is All You Need” 2017
- Architecture encodeurs / décodeur /encoder-décodeur
- Accès aux modèles facilité et frugalité des modèles
- Apprentissage par transfert

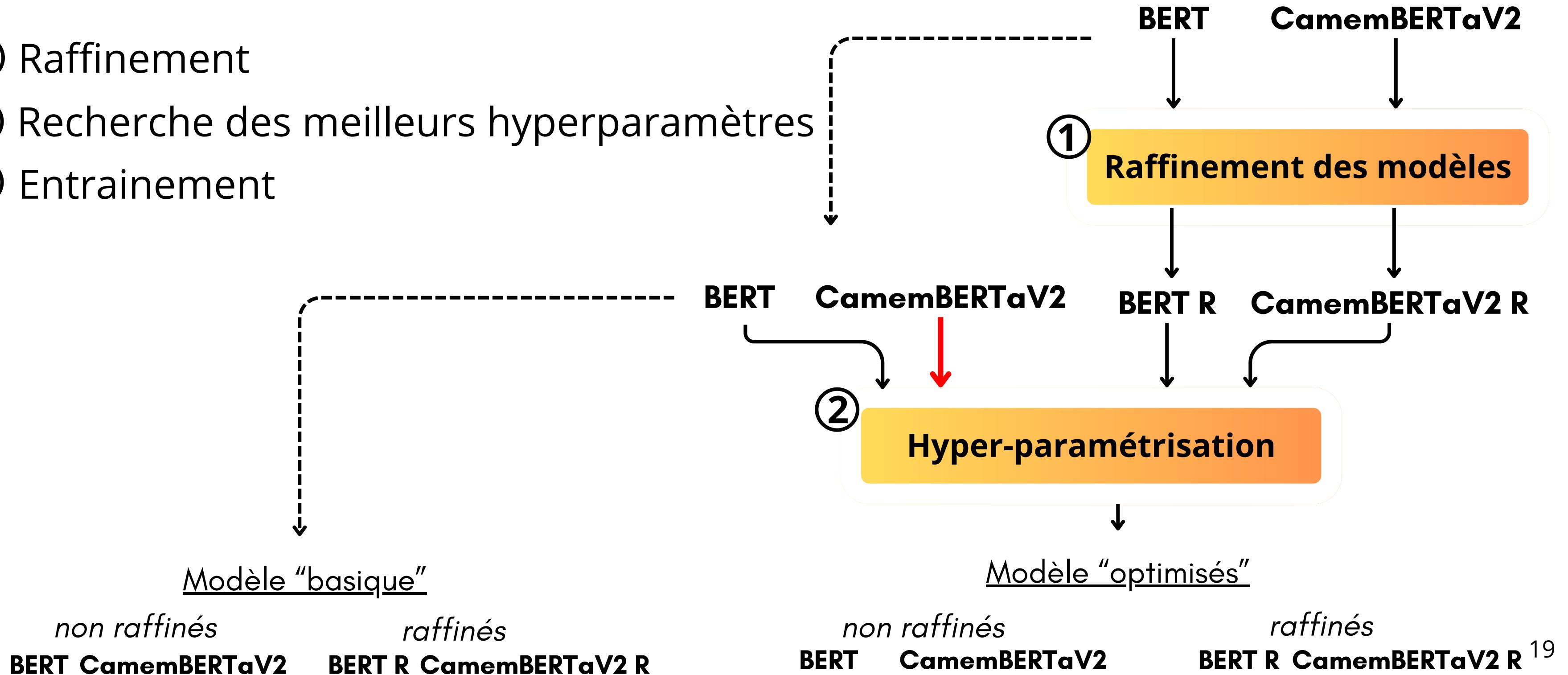


Hugging Face

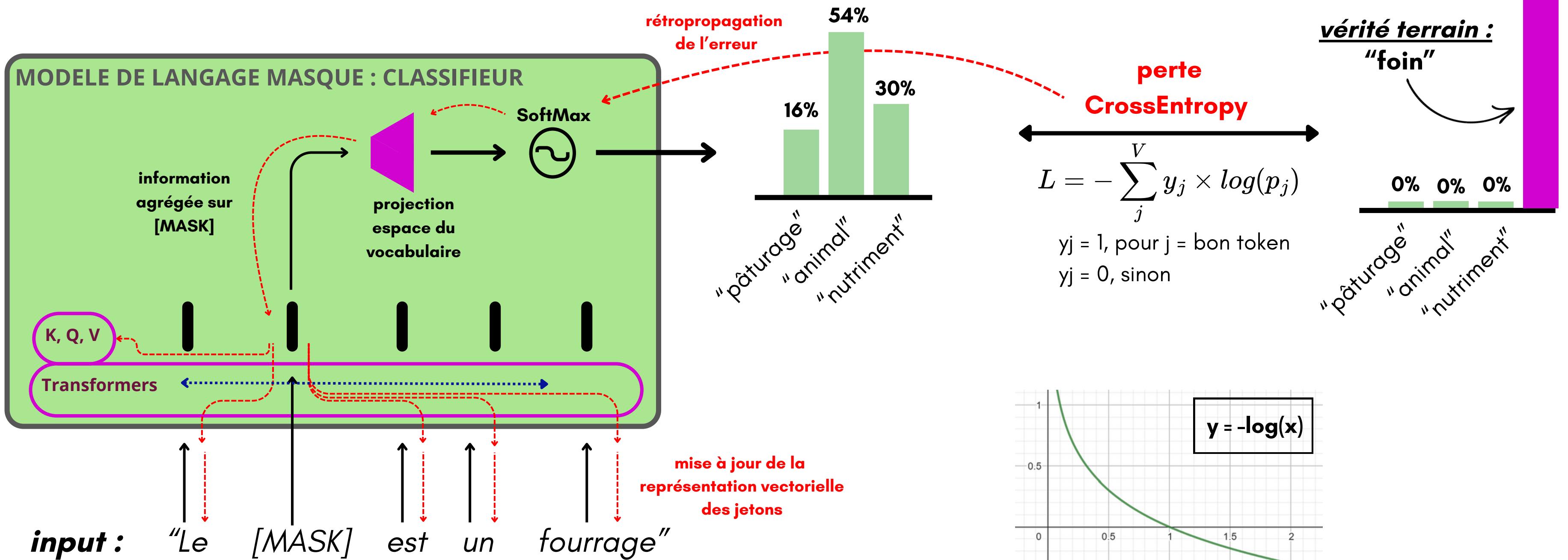


Construction des modèles profonds : **pipeline**

- ① Raffinement
- ② Recherche des meilleurs hyperparamètres
- ③ Entrainement

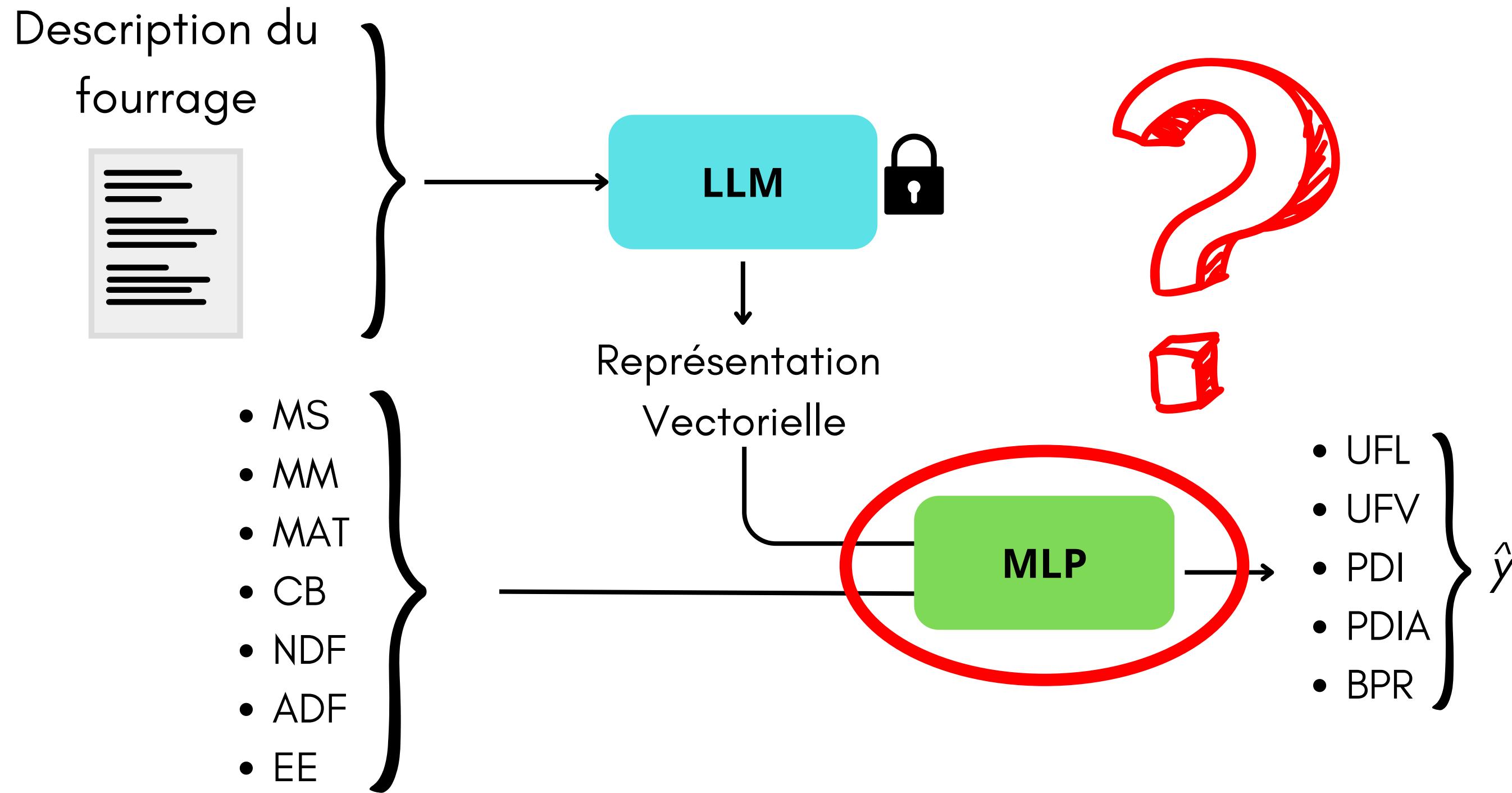


Construction des modèles profonds : raffinement des modèles de langues



100%

Hyperparamétrisation de la tête de régression



Hyperparamétrisation de la tête de régression

- Choix non arbitraires des hyperparamètres
 - > Définitions d'un espace de recherche
 - > Algorithmes divers pour explorer cet espace



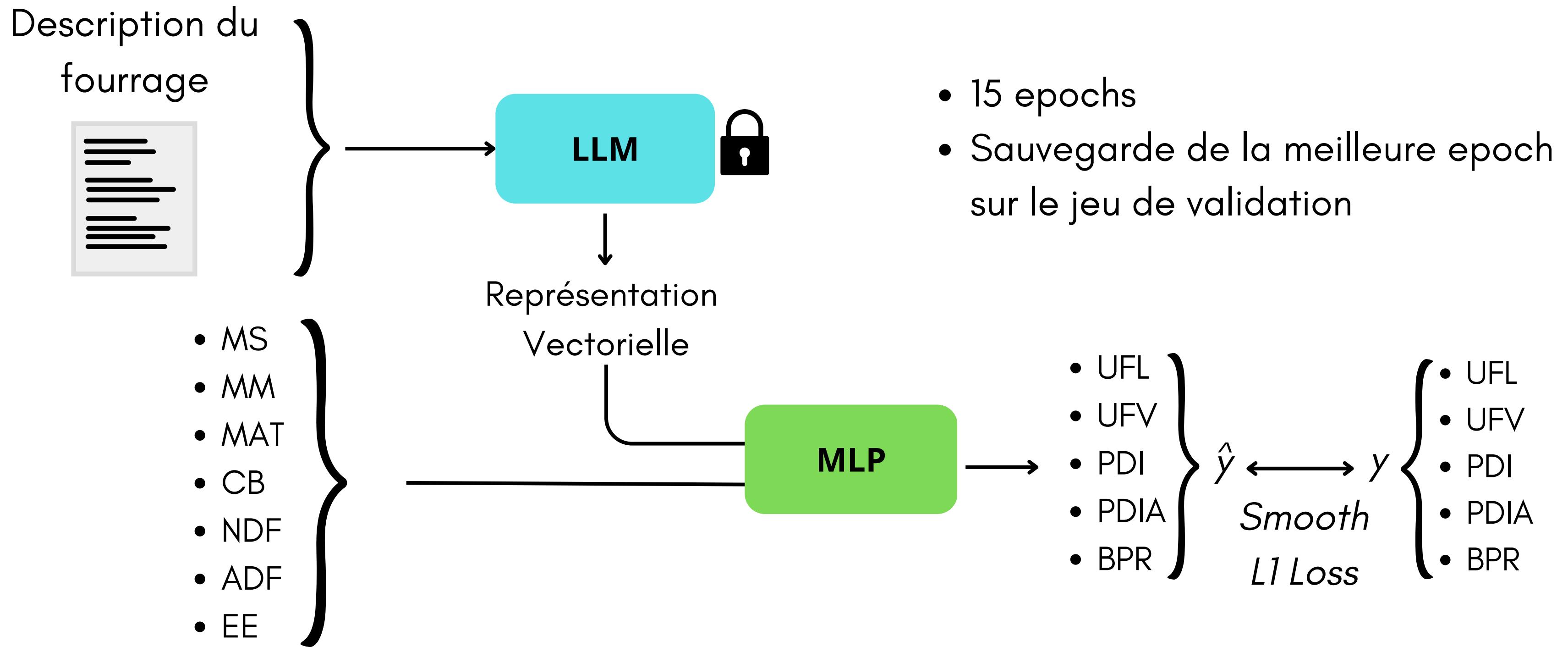
Idéal

- Exploration exhaustive de l'espace de recherche ($N \rightarrow +\infty$)
- Entrainement identique à celui prévu en aval (15 epochs)

Réalisable

- Exploration superficielle (100 essais)
- Entrainement limité pour chaque essai (1 epoch)

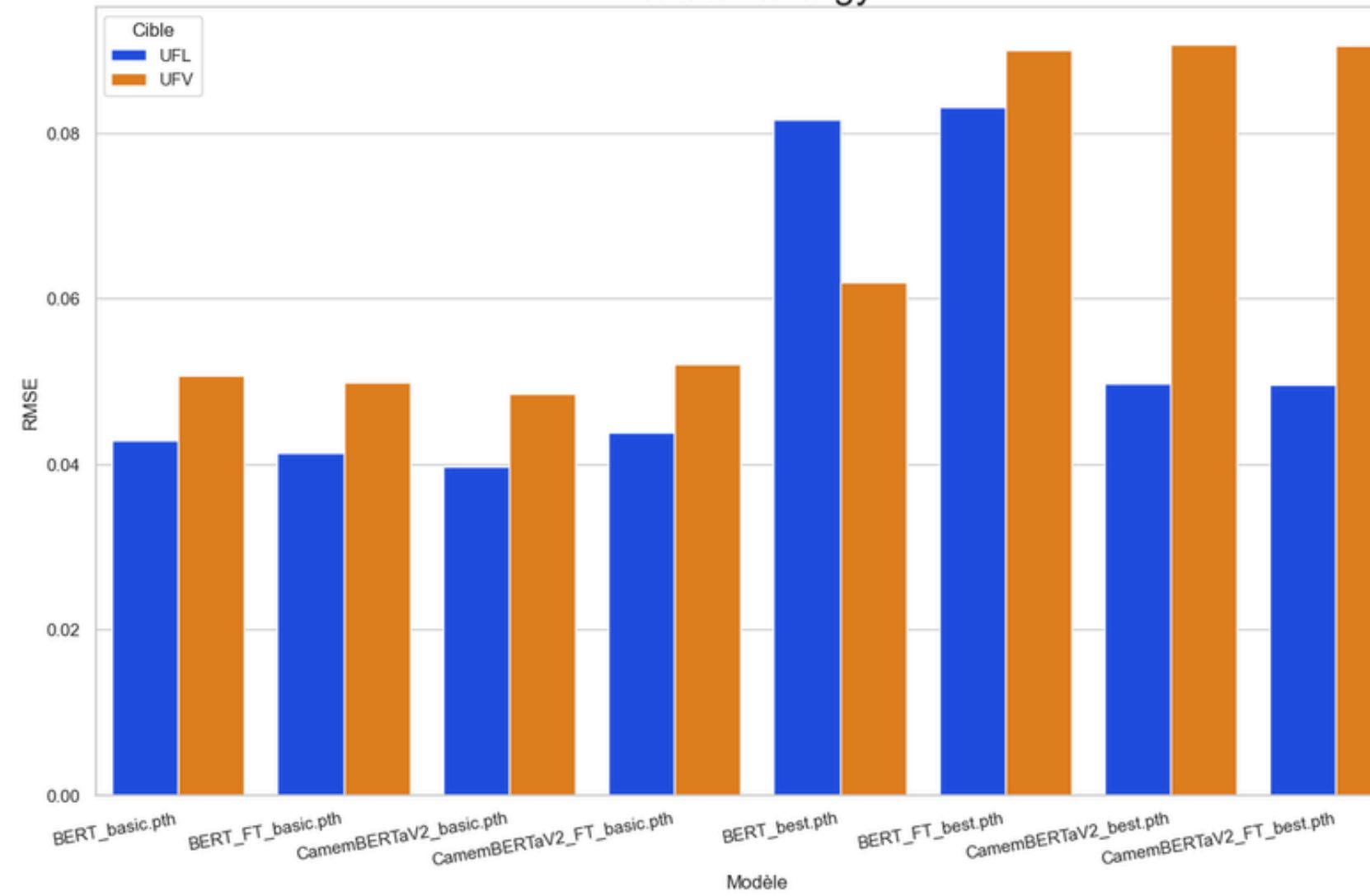
Apprentissage de la tâche de régression



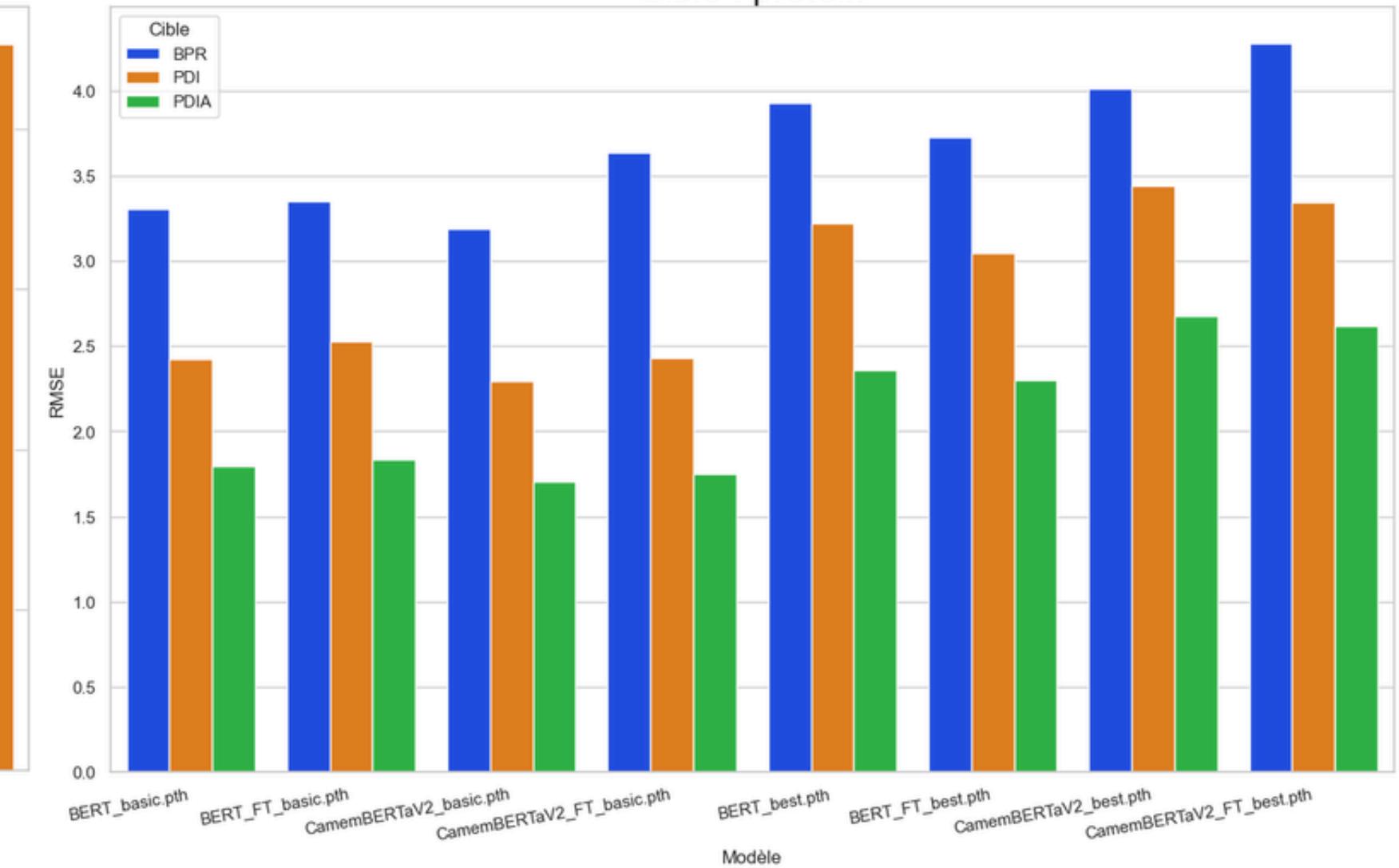
III. Résultats

Apprentissage profond :Performances globales

Comparaison des modèles (basic & best & ML)
métrique : RMSE
Cible : energy

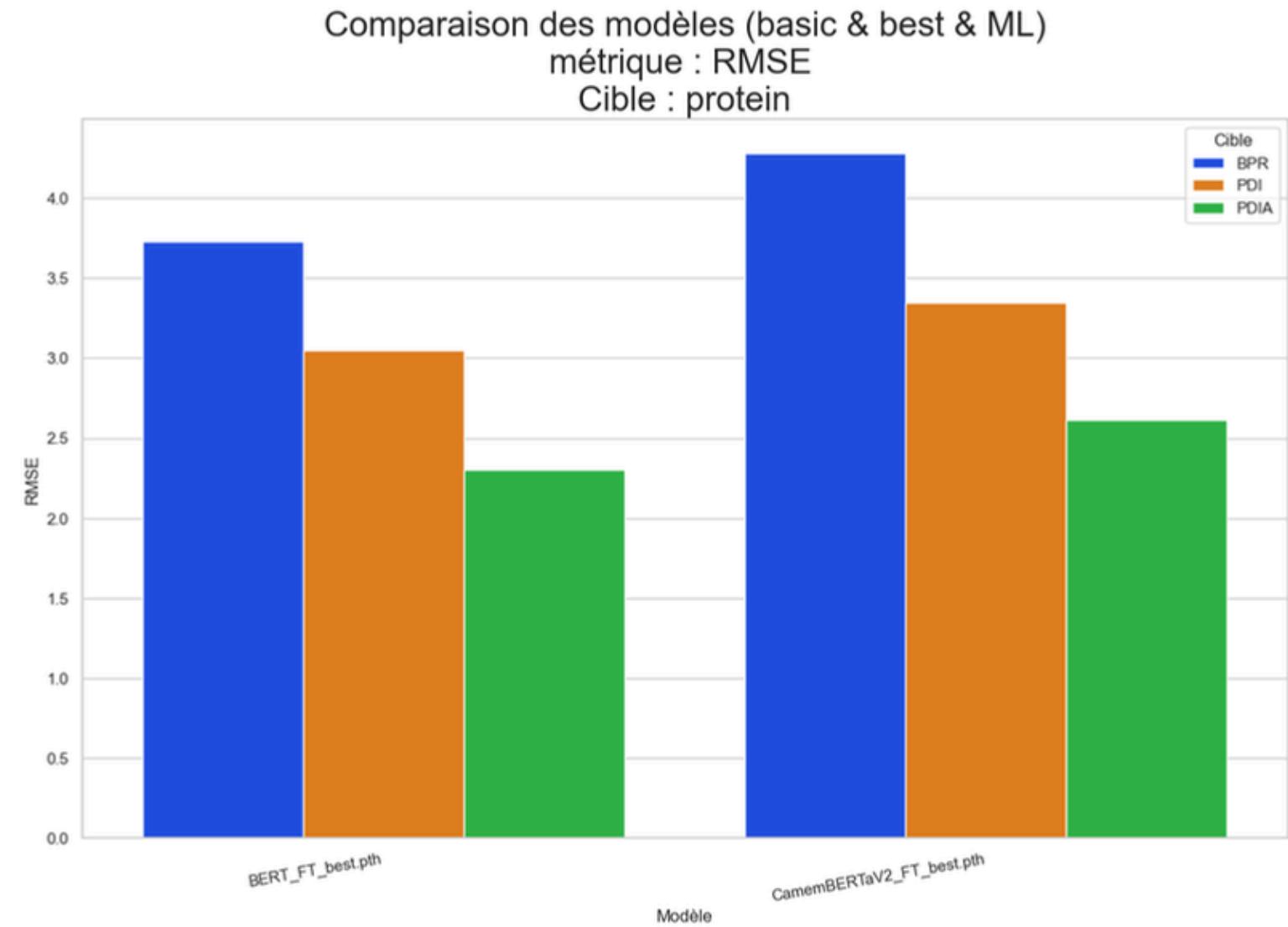
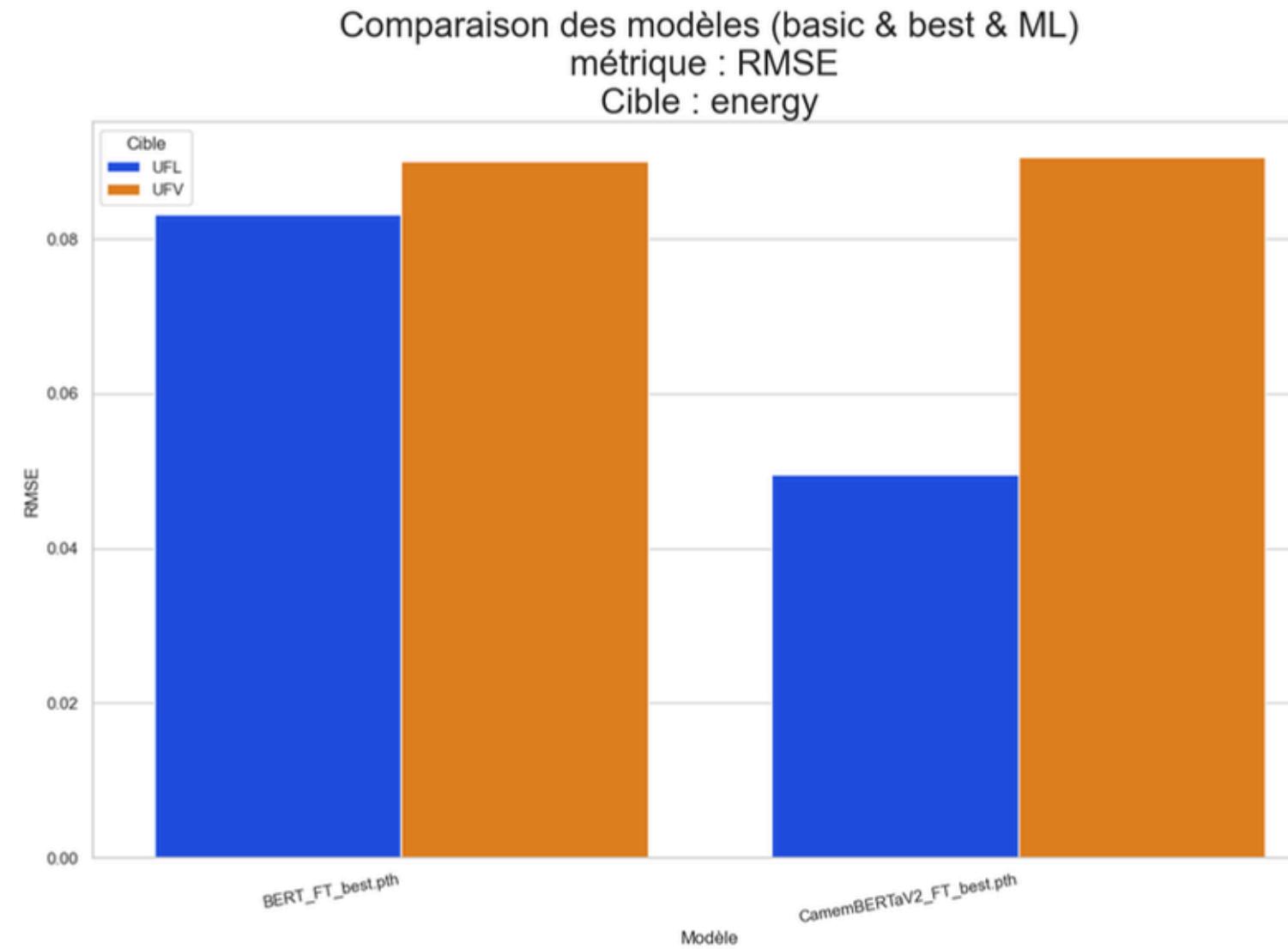


Comparaison des modèles (basic & best & ML)
métrique : RMSE
Cible : protein



Des performances globales pertinentes pour l'ensemble des cibles

Apprentissage profond : effet des modèles

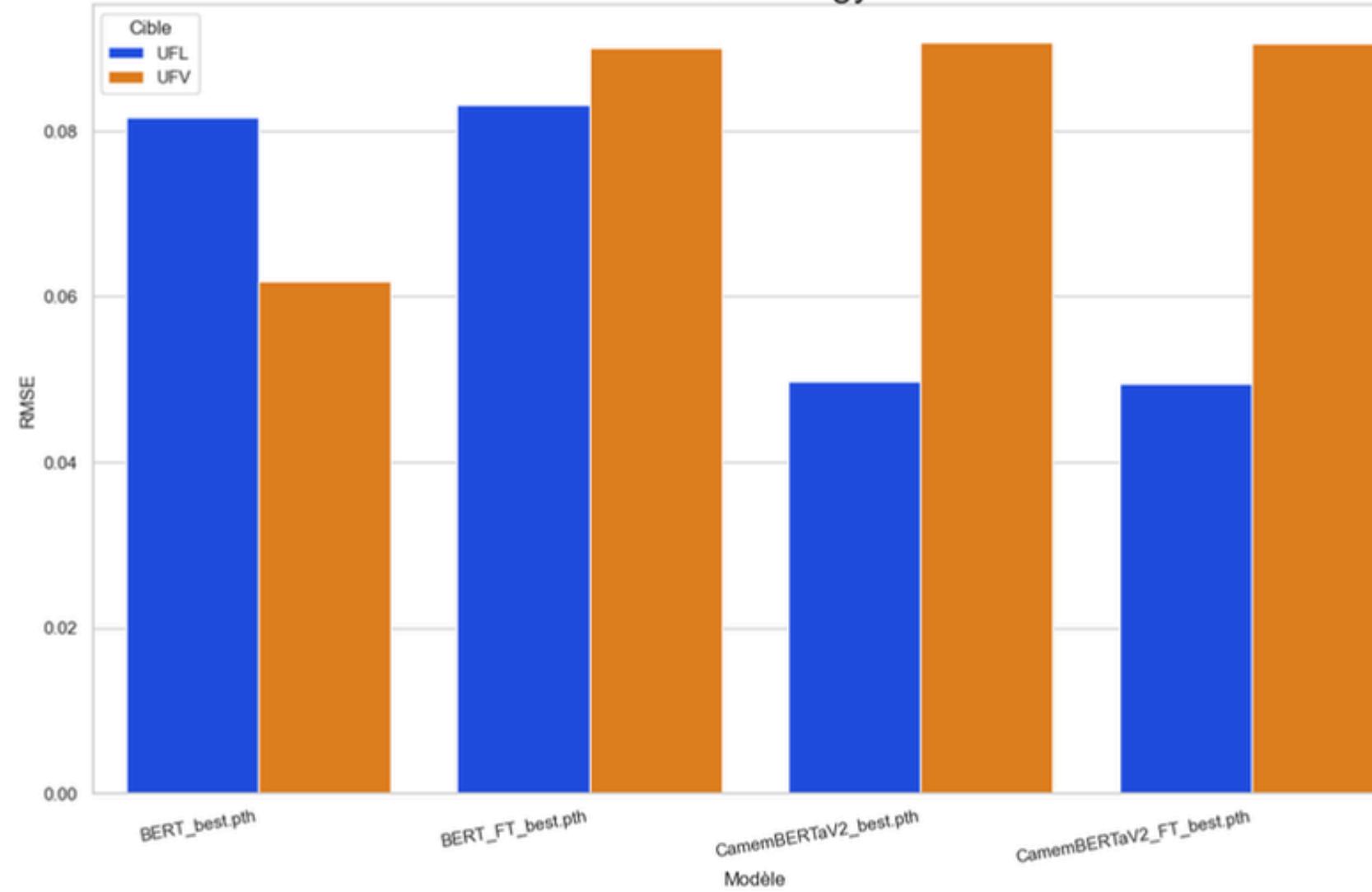


Performances hétérogènes sur les cibles selon les modèles :

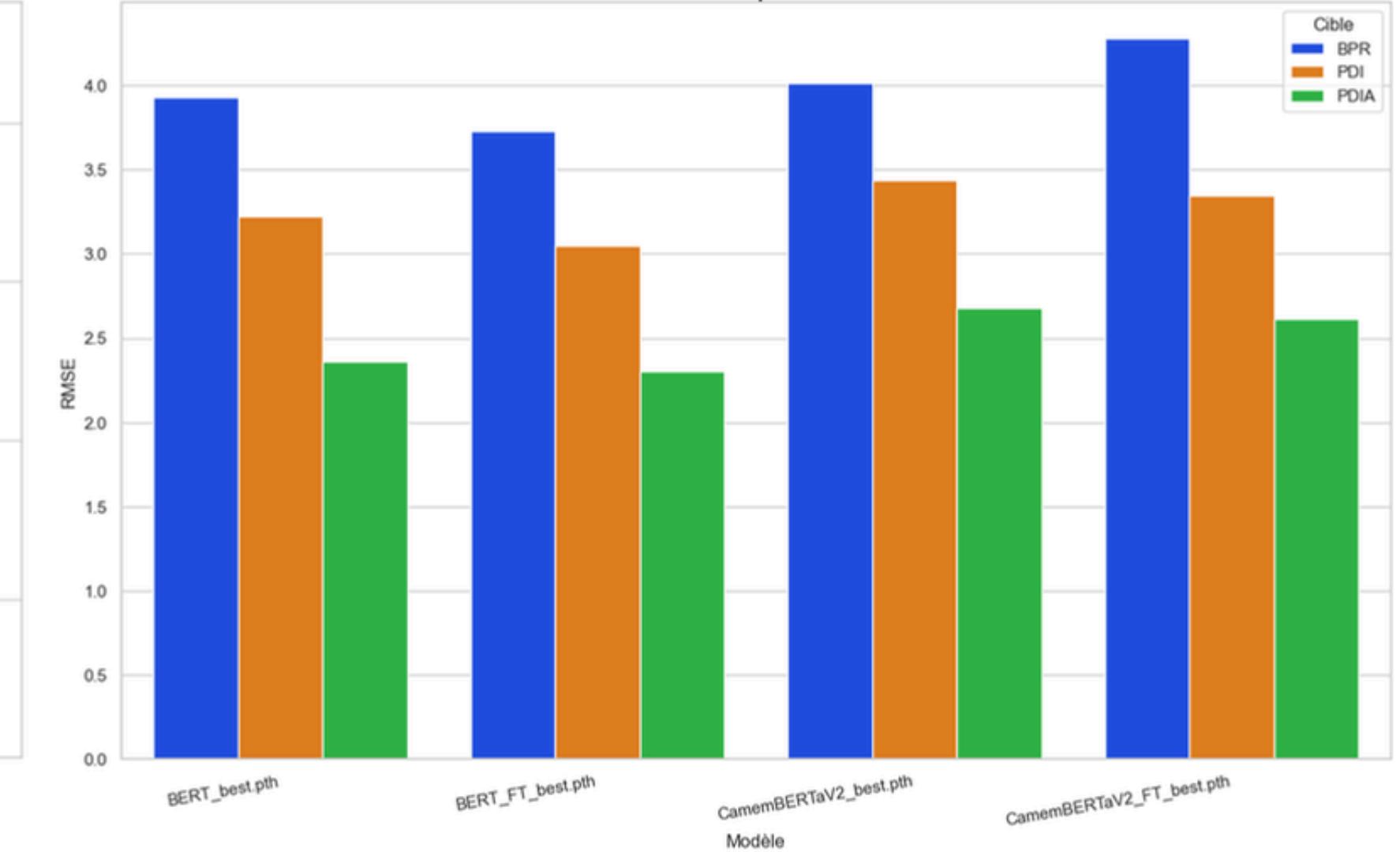
- **BERT meilleur sur valeurs énergétiques**
- **CamemBERTaV2 meilleure sur UFL,**
- **Performances similaires sur RMSE**

Apprentissage profond : effet du raffinage

Comparaison des modèles (basic & best & ML)
métrique : RMSE
Cible : energy



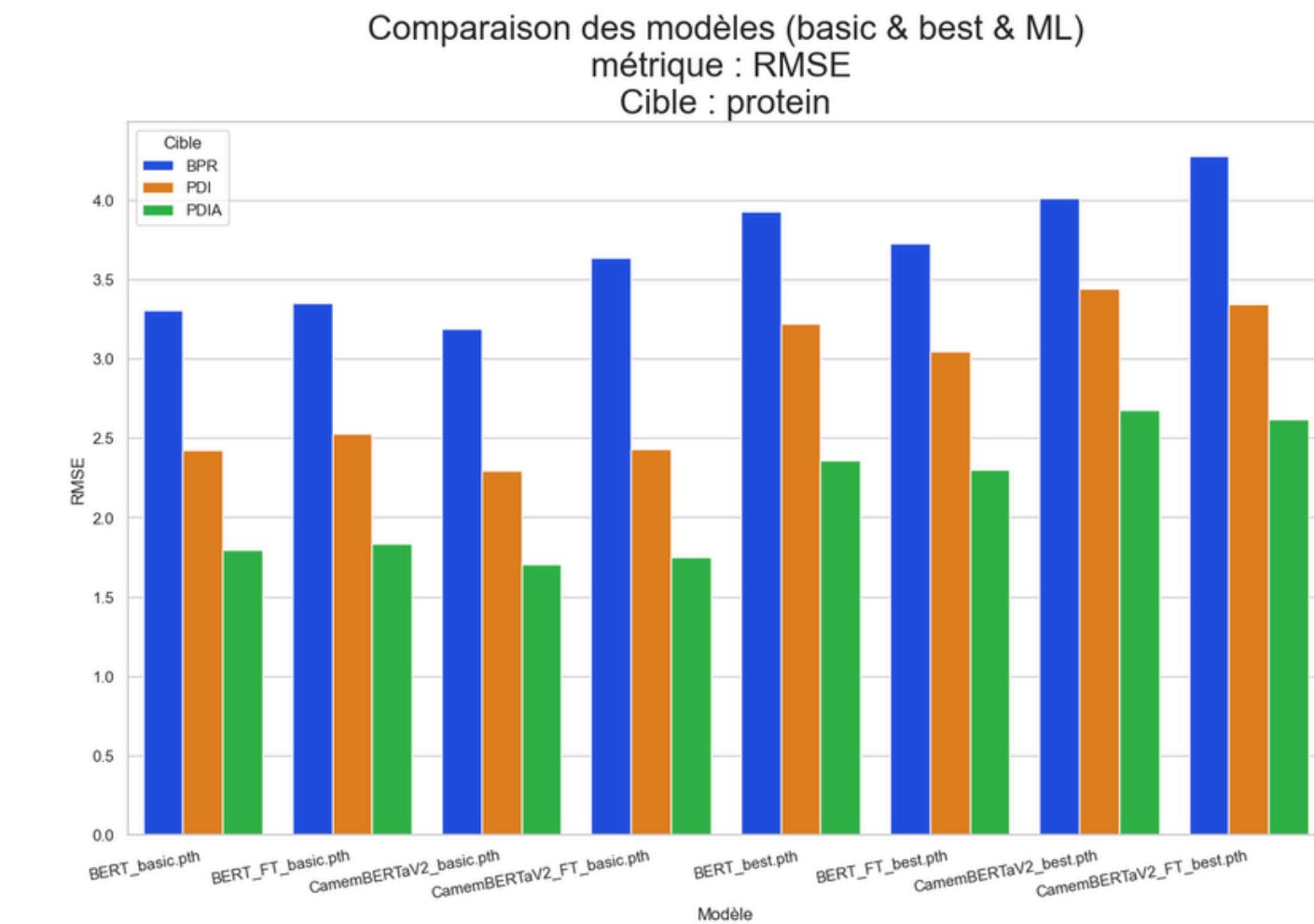
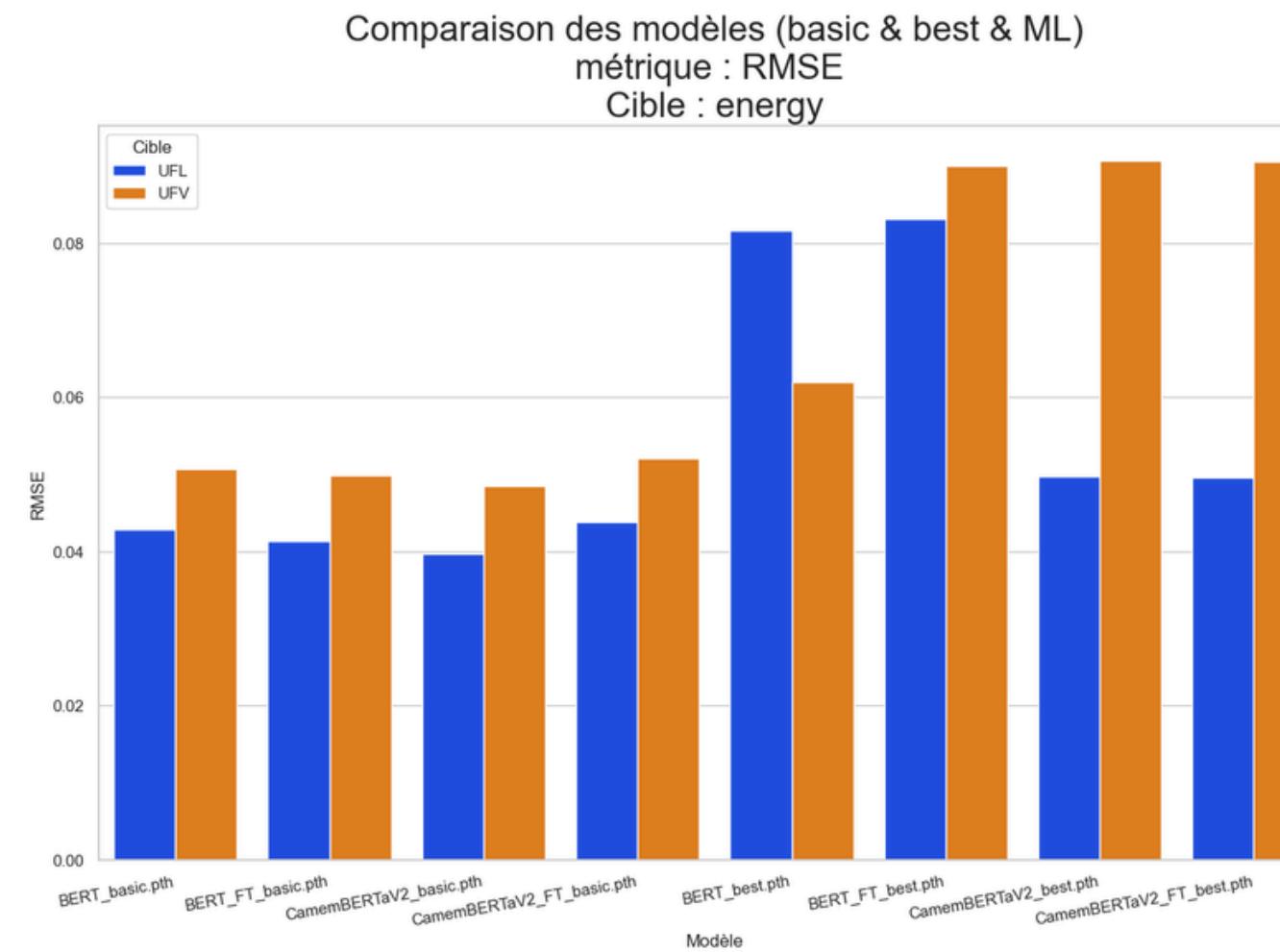
Comparaison des modèles (basic & best & ML)
métrique : RMSE
Cible : protein



Effet du raffinage est contrasté :

- **tendance à l'amélioration pour les valeurs protéiques**
- **tendance plus hétérogène pour les valeurs énergétique**

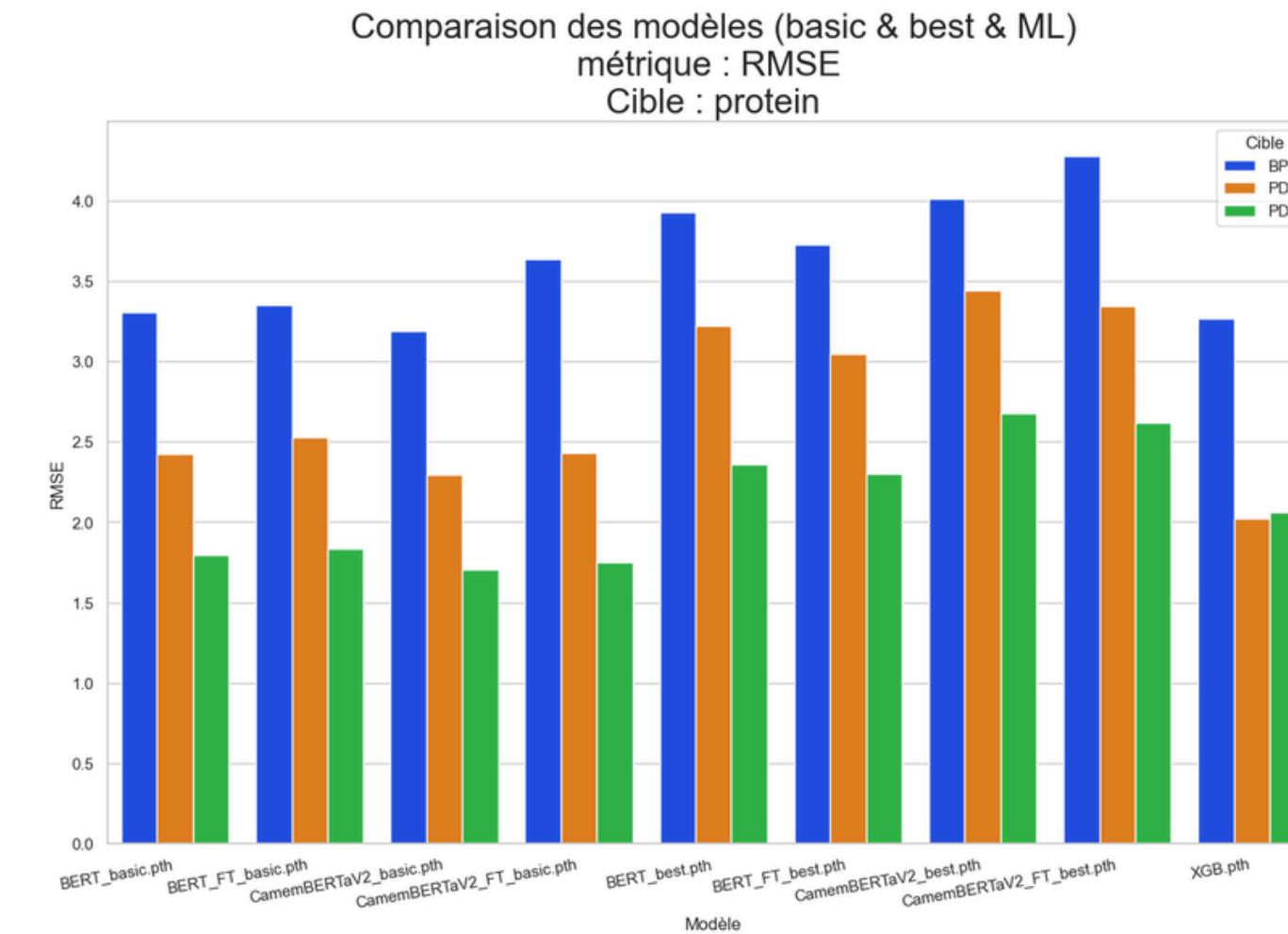
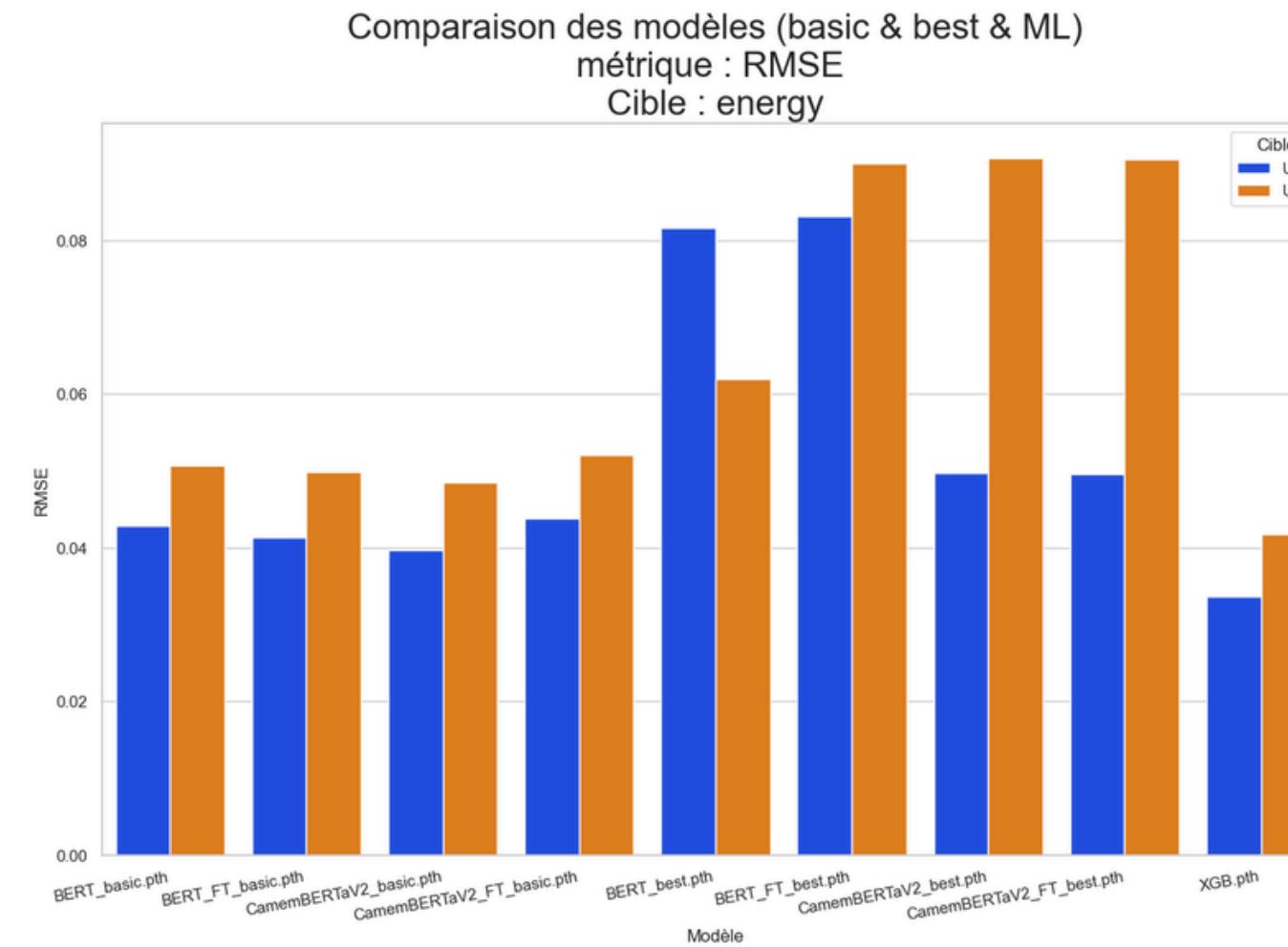
Apprentissage profond : effet de l'optimisation de la tête de régression



Effet de l'optimisation de la tête de régression optimisée :

- homogénéité de la qualité des prédictions des valeurs énergétiques : meilleures prédictions
- Hétérogénéité chez les valeurs protéiques :
 - BPR : faible diminution
 - PDI, PDIA : diminution plus marqué

Approche non neuronale :



Valeurs énergétiques : gain sur UFV et UFL

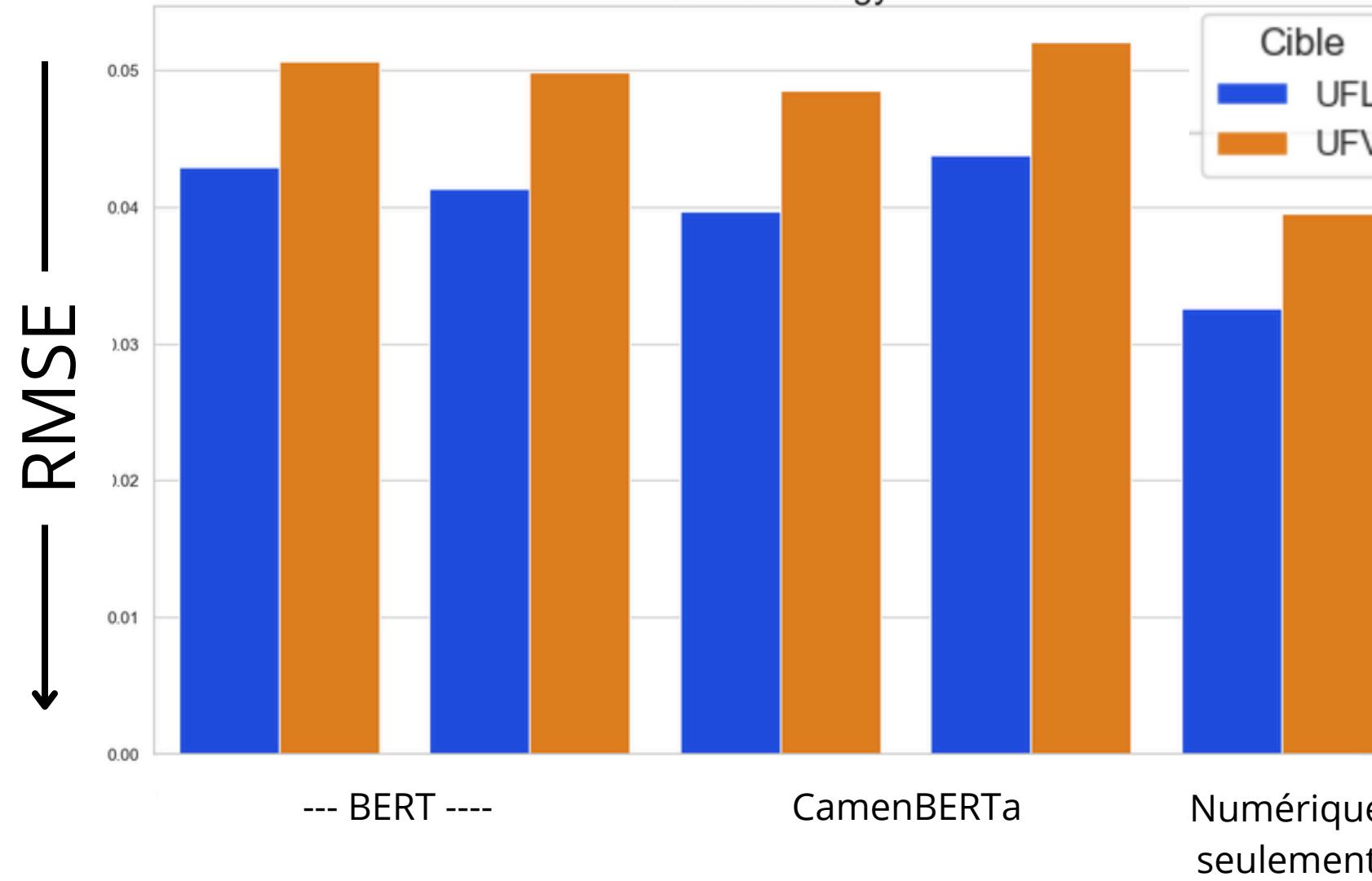
Performances plus contrastés :

- similaire en performances BPR, PDIA
- meilleur sur PDI

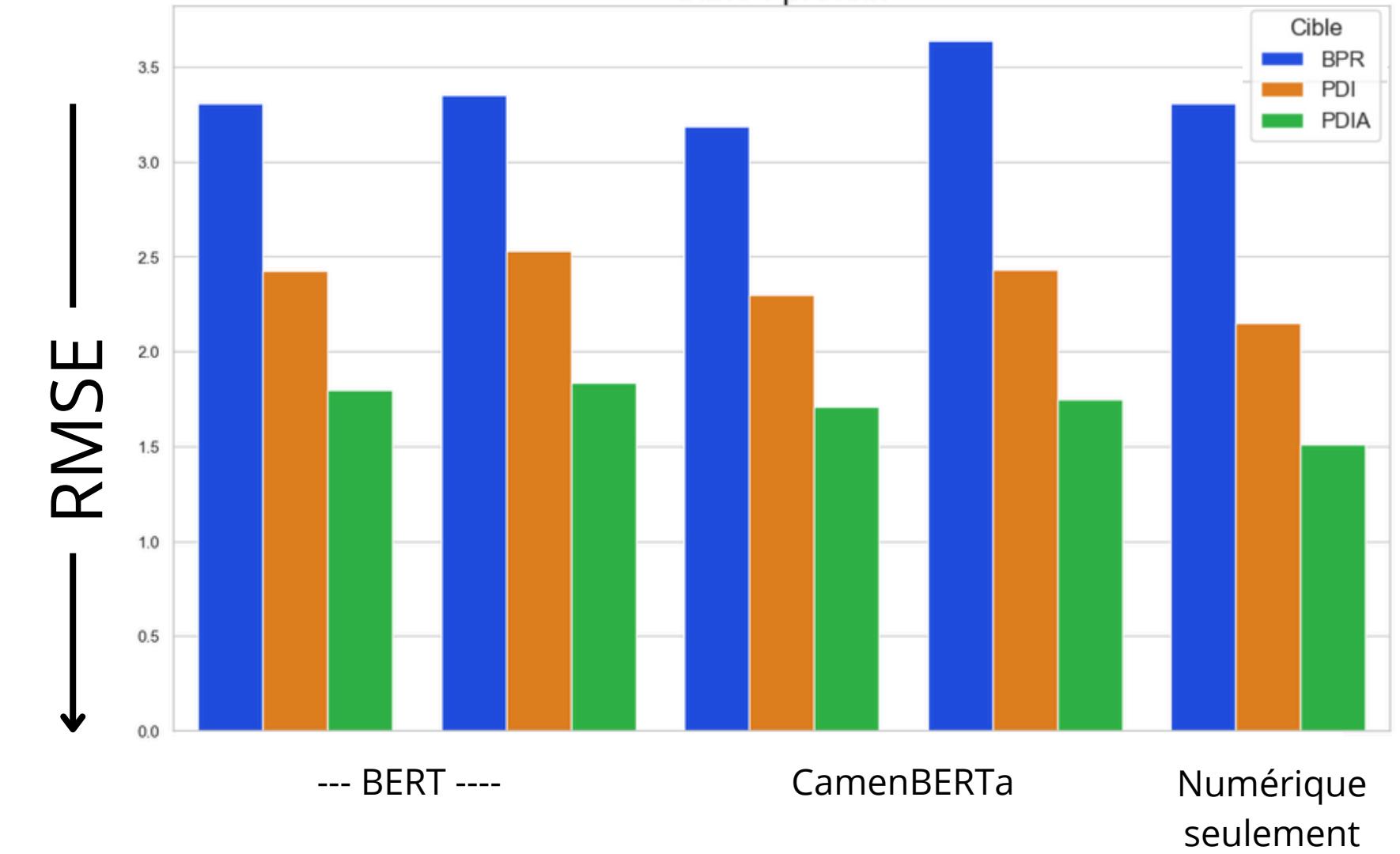
IV. Discussion

Apprentissage profond : effet des valeurs numériques

Comparaison des modèles (basic & best & ML)
métrique : RMSE
Cible : energy

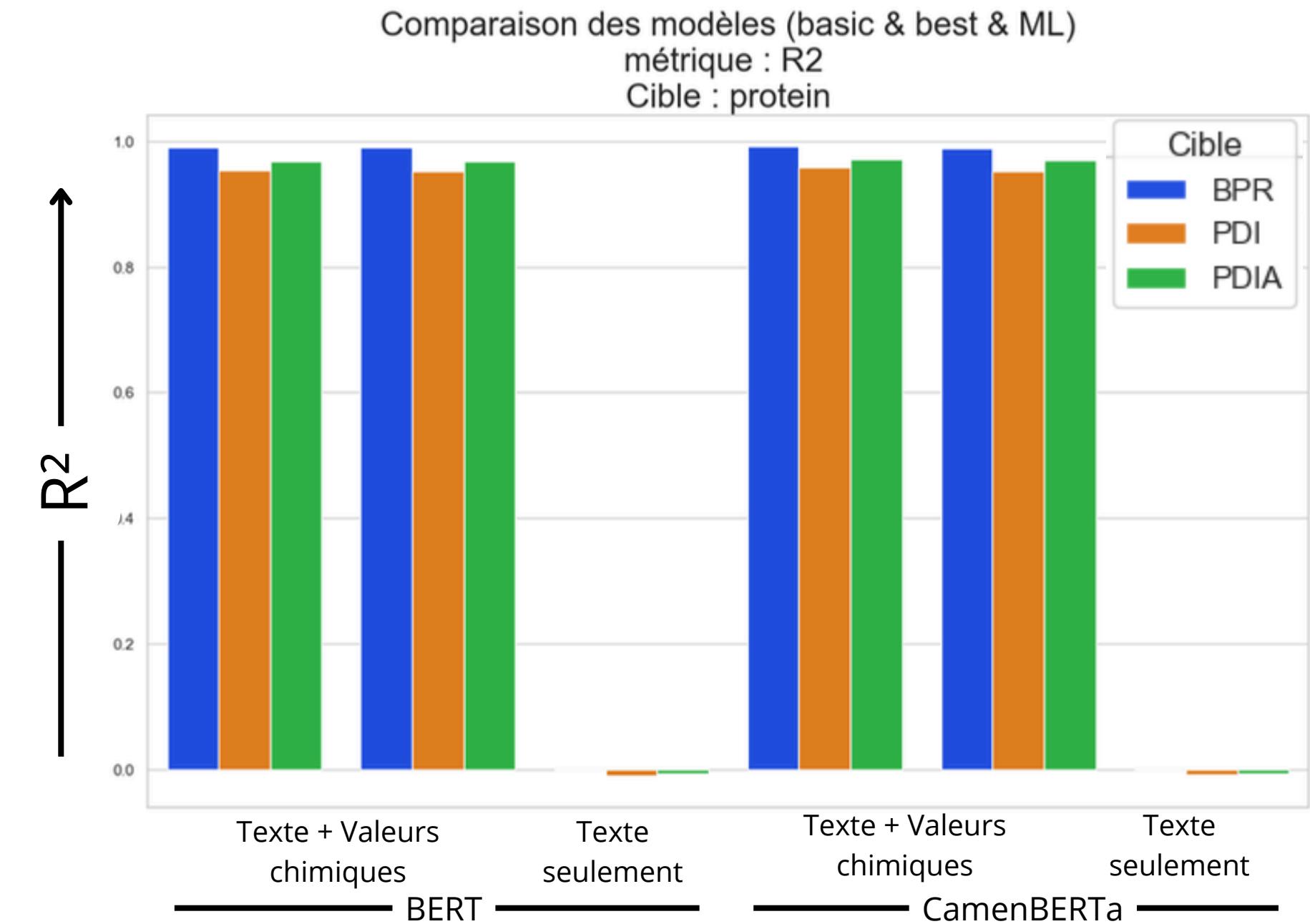
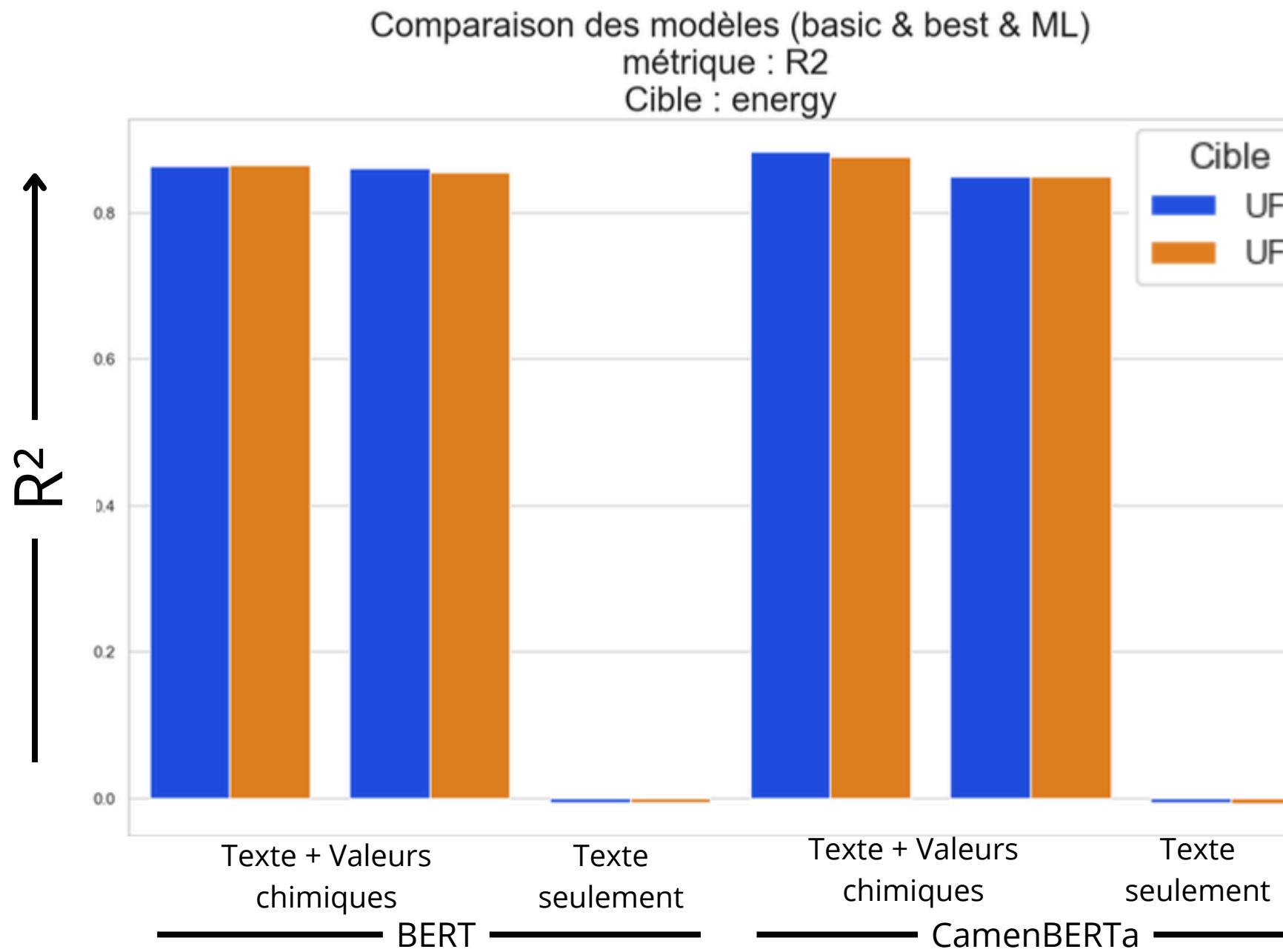


Comparaison des modèles (basic & best & ML)
métrique : RMSE
Cible : protein



Effet du texte : meilleures performances avec seulement les numériques

Apprentissage profond : effet des valeurs textuelles



Effet du texte : diminution des performances, aucune information ne semble être capturé par le modèle

Discussion : effet de la dilution des informations des valeurs chimiques par les valeurs textuelles

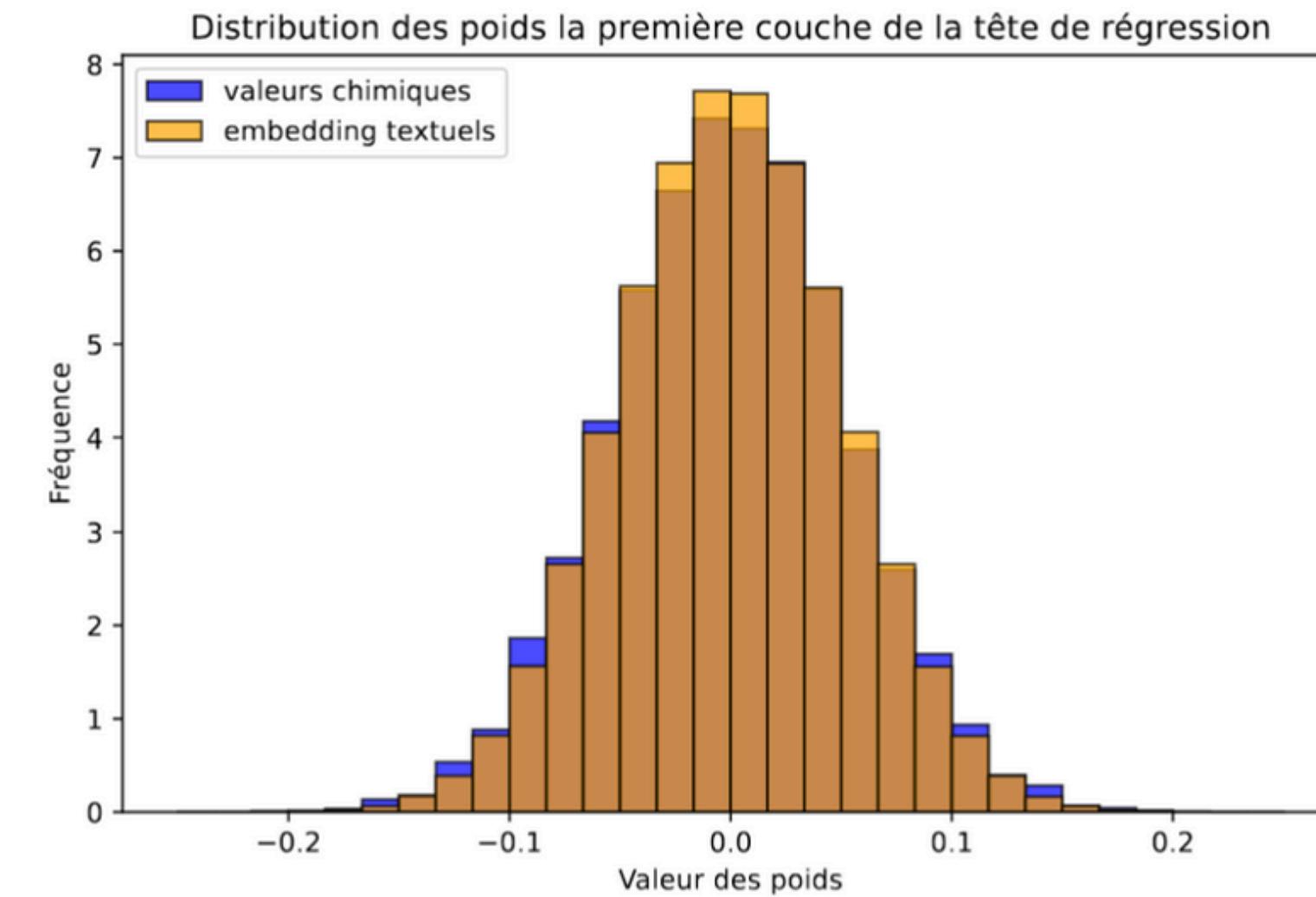
Configuration de l'encodeur

Description du fourrage



Valeurs chimiques

- MS
- MM
- MAT
- CB
- NDF
- ADF
- EE



Conclusion

- Des résultats encourageants :
 - toutes les cibles
- Meilleur modèle en apprentissage statistique non neuronale :
XGBoost
- Intérêt des modèles d'apprentissage profond :
 - palier l'absence d'exhaustivité des données de calibrage des modèles mécanistiques
 - capacité de généralisation inter-régionales
 - gratuité et facilité

ANNUXES

Annexe - Sommaire

- Reproductibilité : 37
- Figures supplémentaires résultats approche non neuronale: 38-46
- Explication de principes en approche neuronale (Qu'est-ce qu'... ?) : 47-49
- Figures supplémentaires résultats approche neuronale : 50-54

Dépôt de nos travaux : GitHub et HugginFace



- scripts brutes
- https://github.com/lauronta/projet_fil_rouge_afz
- <https://github.com/lauronta/pfr-demo>



• Modèles

Raphaël GENIN
RaphyThePingouin

Models 2

- RaphyThePingouin/CamemBERTaV2_FT_10_epoch
- RaphyThePingouin/BERT_FT_2_epoch

Datasets

- None yet

Recent Activity

- Updated a model 10 days ago
- Published a model 10 days ago

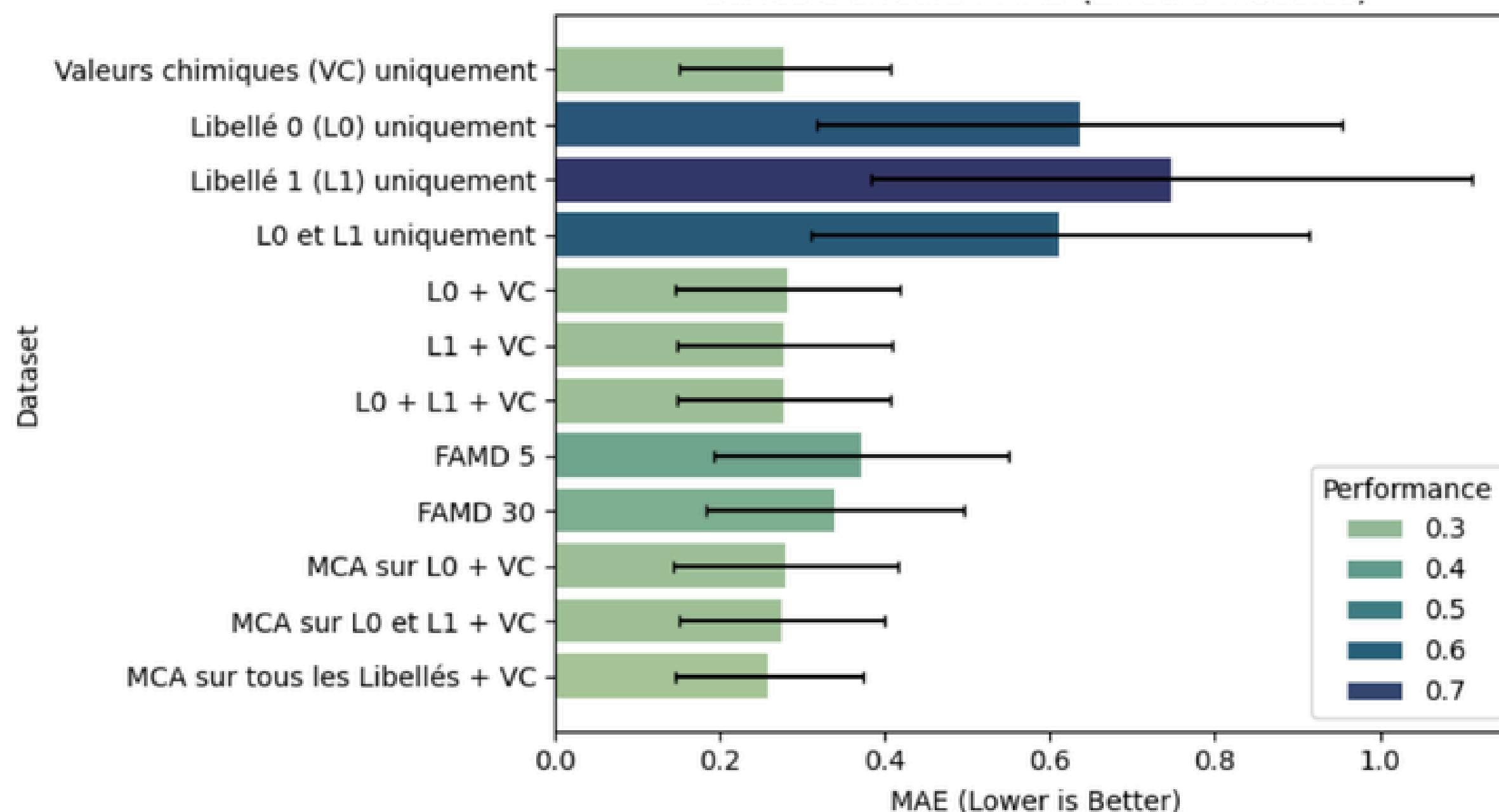
• Démo

- <https://huggingface.co/RaphyThePingouin>
- Grands modèles de langage rafinés

Approche non neuronale : Résultats

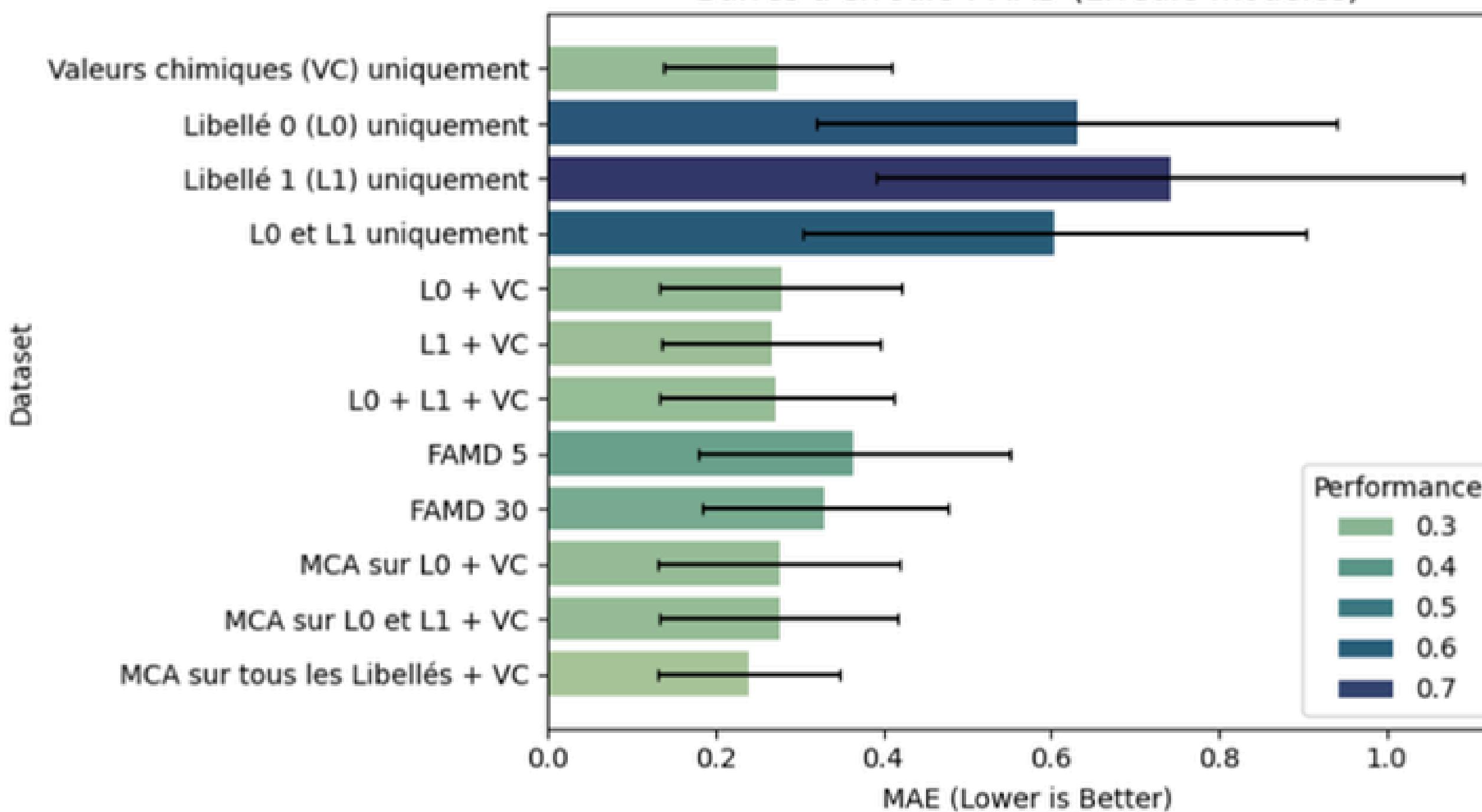
Quel pré-traitement apporte les meilleurs résultats ?

Exemple pour les valeurs UFL :

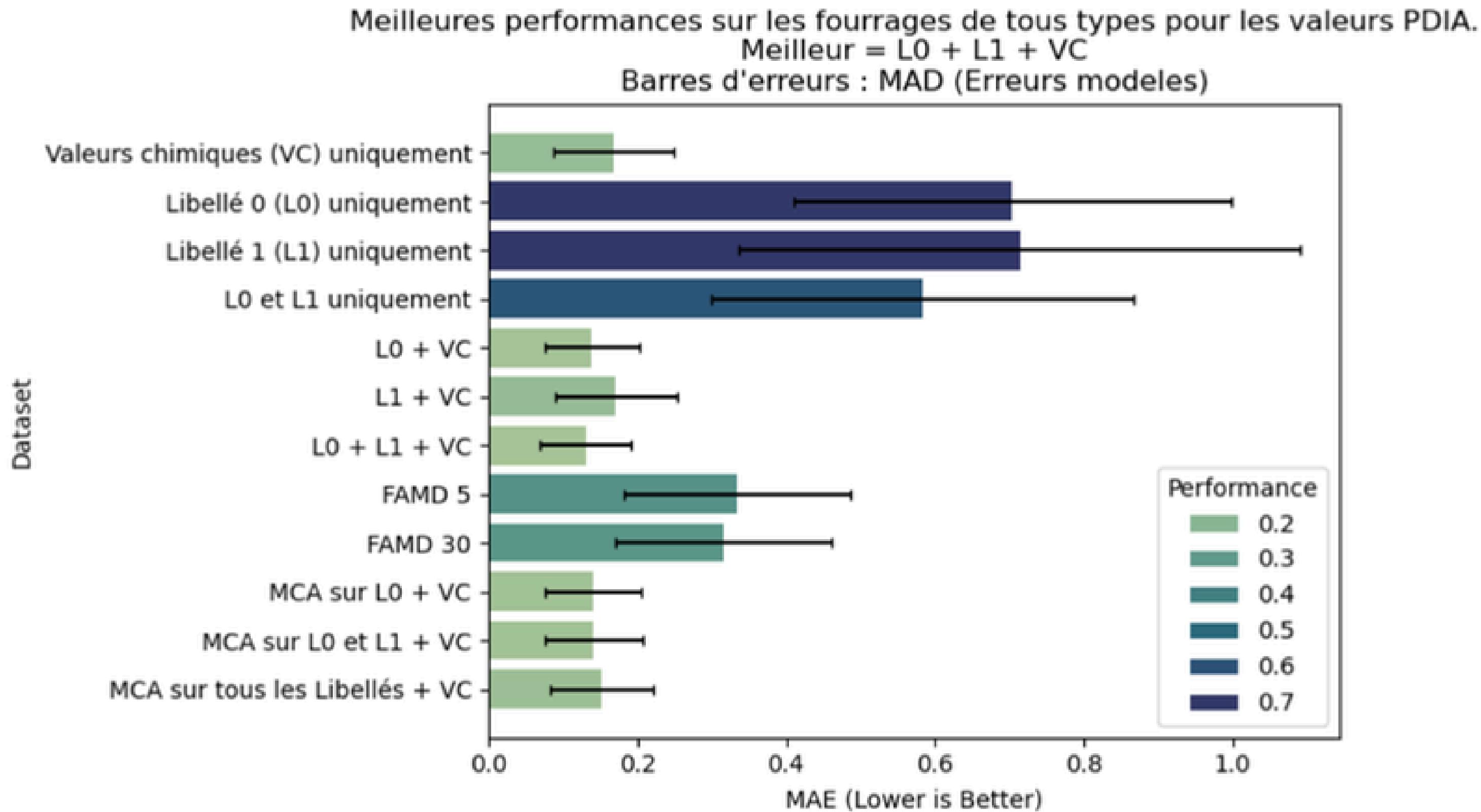


Meilleurs pré-traitements UFV

Meilleures performances sur les fourrages de tous types pour les valeurs UFV.
Meilleur = MCA sur tous les Libellés + VC
Barres d'erreurs : MAD (Erreurs modeles)



Meilleurs pré-traitements PDIA

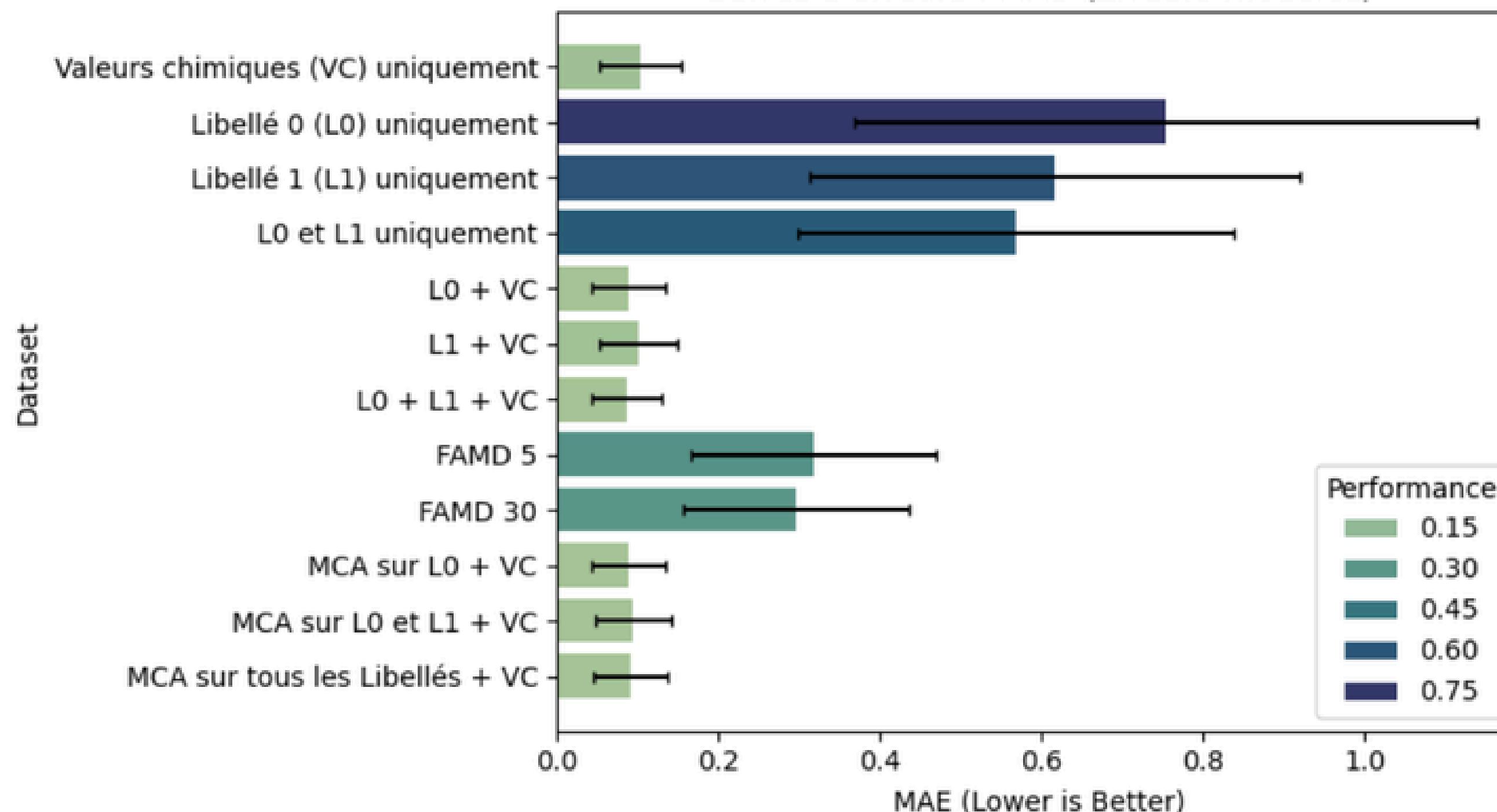


Meilleurs pré-traitements BPR

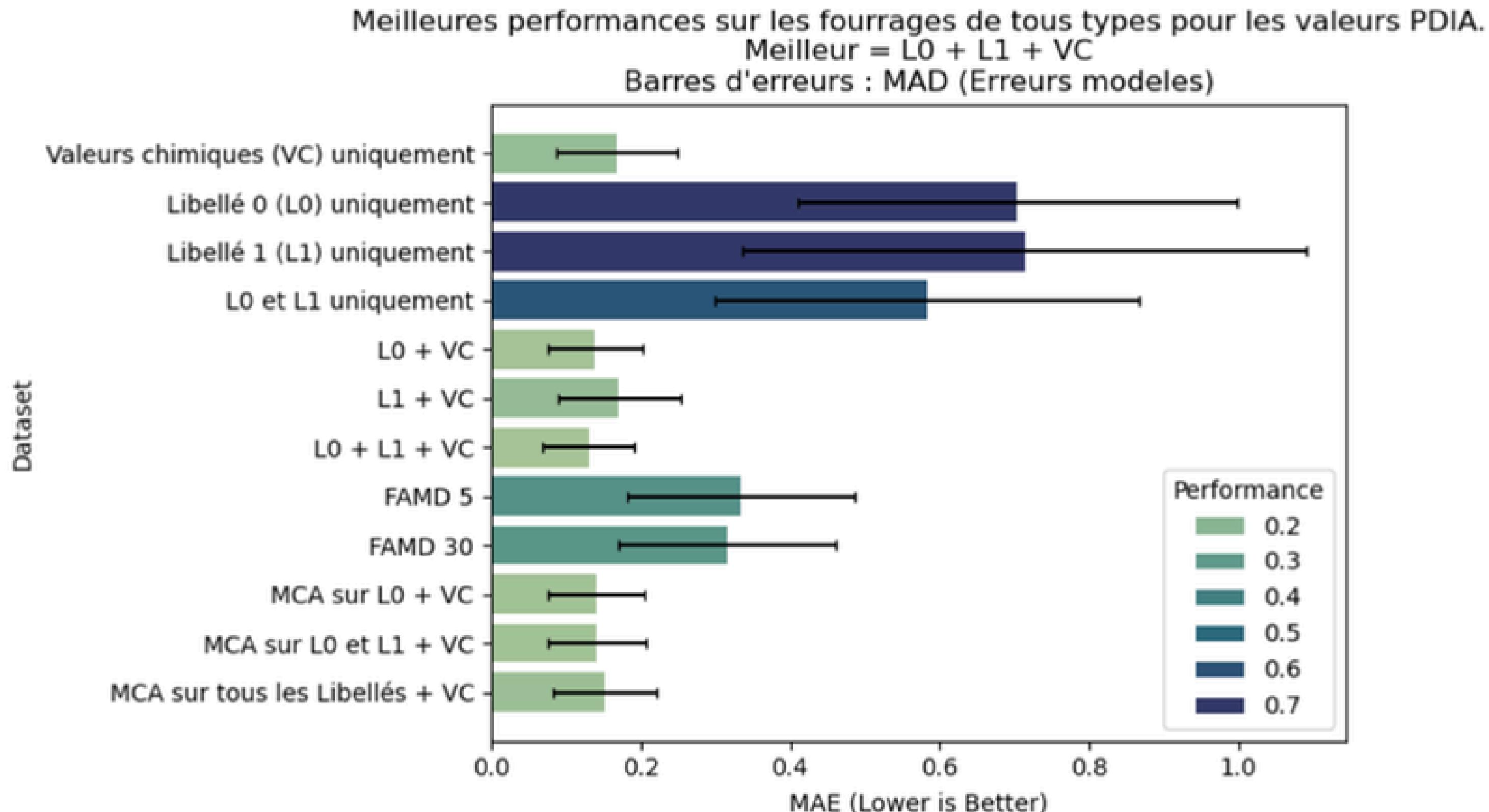
Meilleures performances sur les fourrages de tous types pour les valeurs BPR.

$$\text{Meilleur} = L0 + L1 + VC$$

Barres d'erreurs : MAD (Erreurs modeles)



Meilleurs pré-traitements PDIA

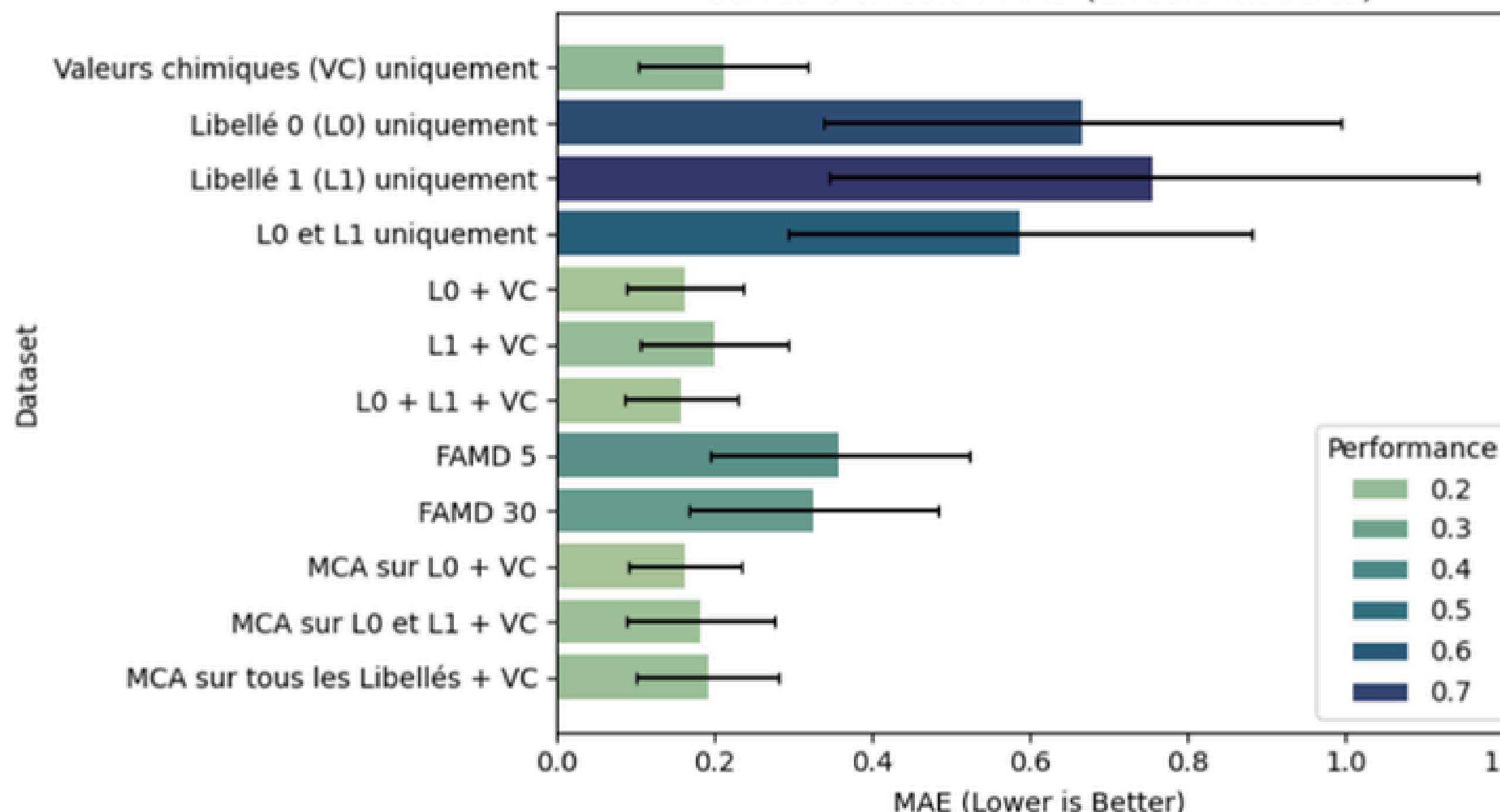


Meilleurs pré-traitements PDI

Meilleures performances sur les fourrages de tous types pour les valeurs PDI.

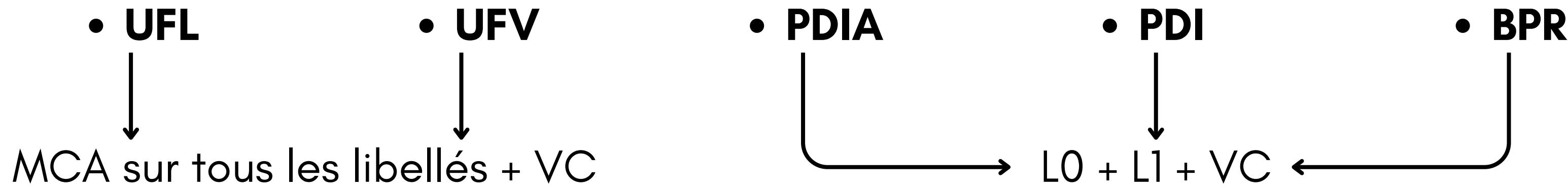
Meilleur = L0 + L1 + VC

Barres d'erreurs : MAD (Erreurs modeles)



Approche non neuronale : **Résultats**

Quel pré-traitement apporte les meilleurs résultats ?

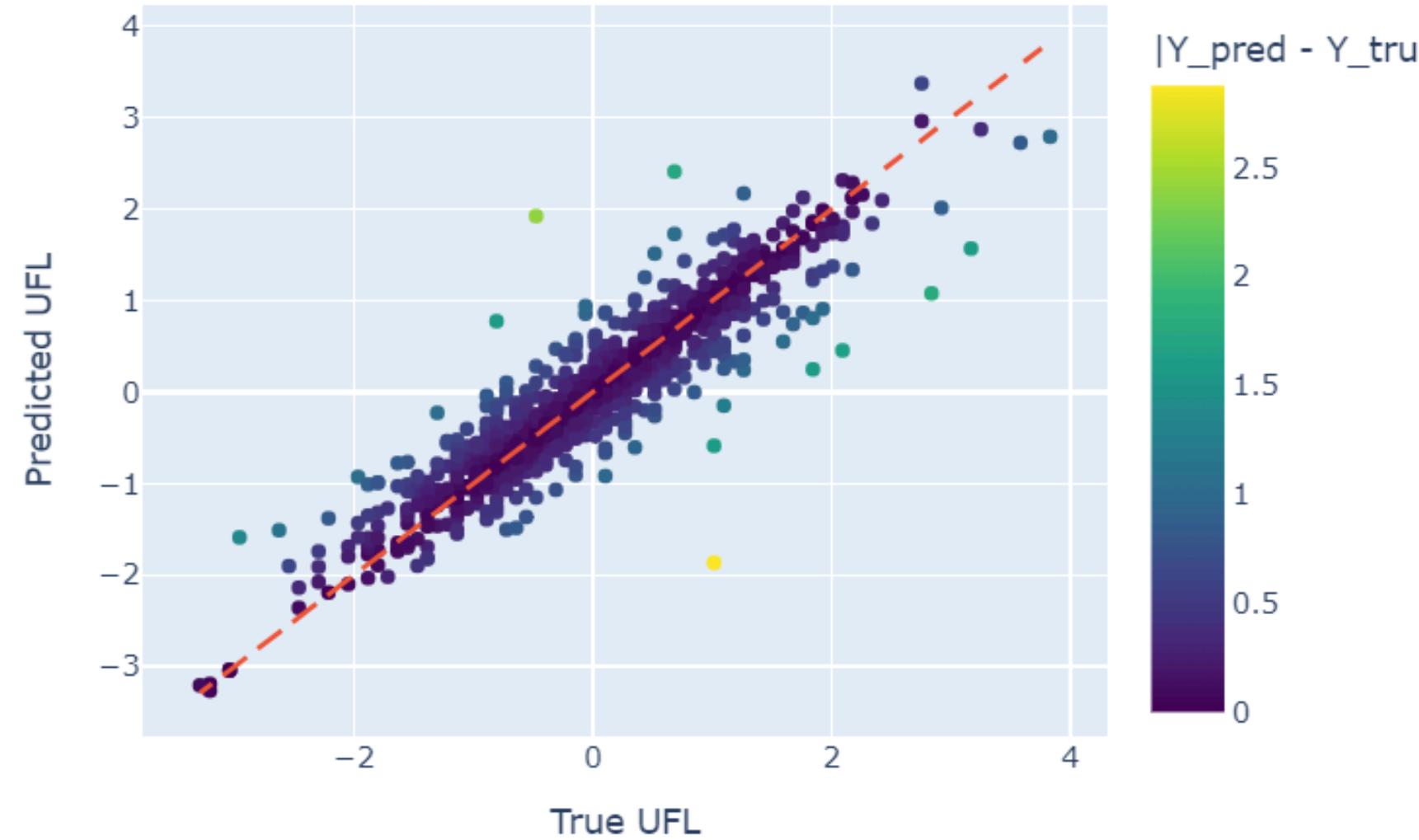


2 algorithmes performent mieux que les autres :

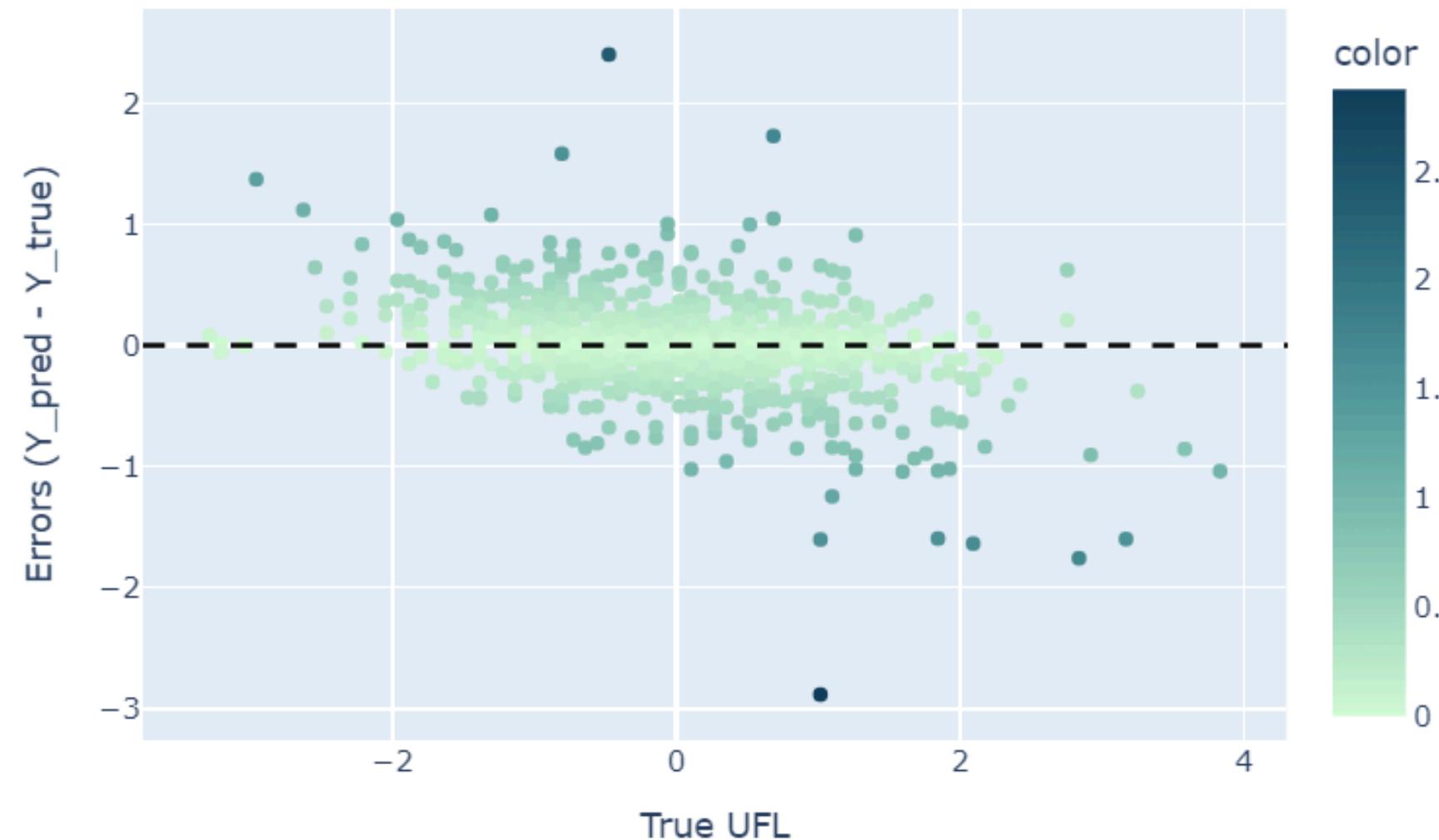
- Séparateur à Vaste Marge à noyau Non Linéaire (rbf)
- Ensemble d'Arbre de régression avec Boosting de Gradient

Meilleur modèle non neuronal pour UFL

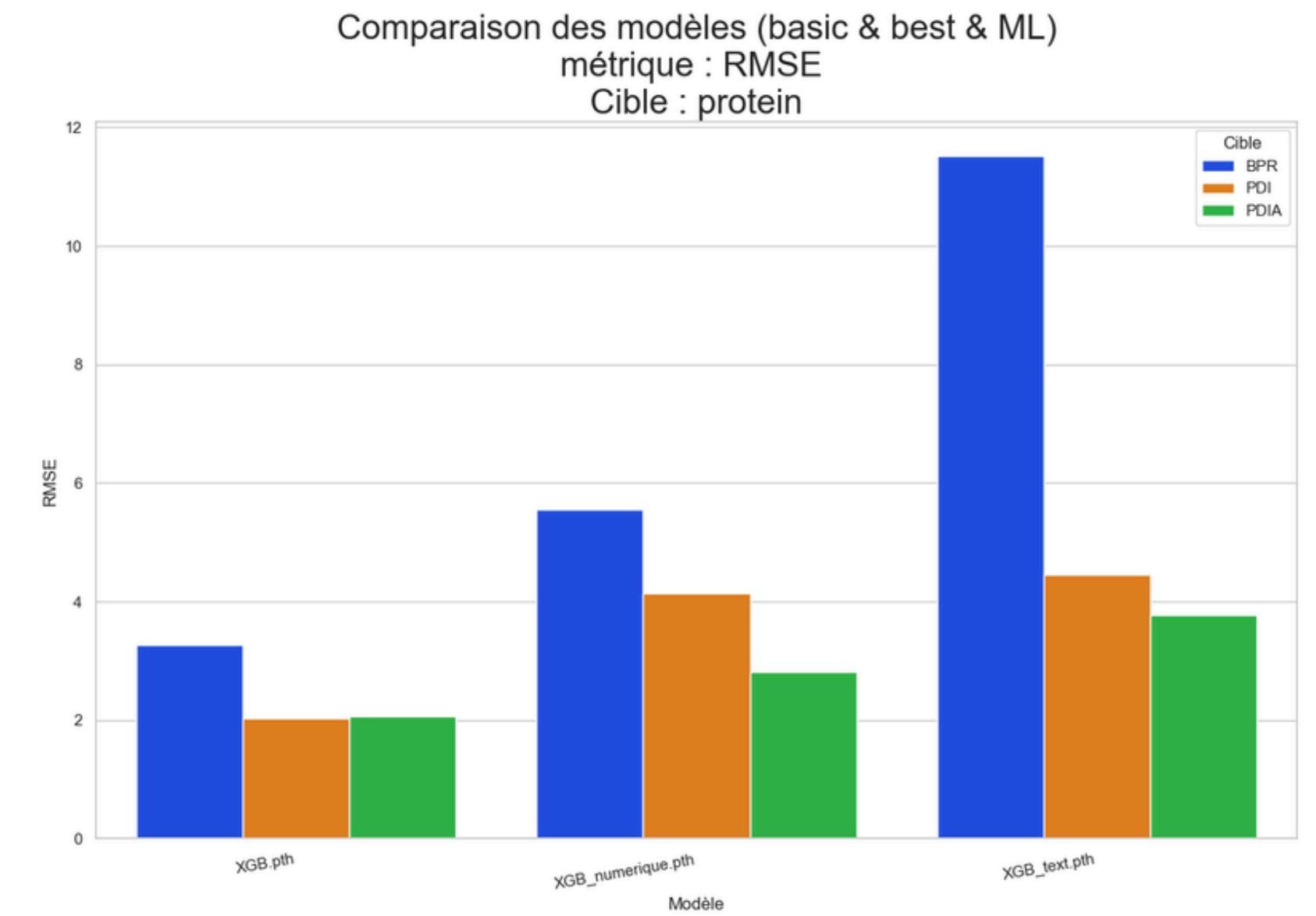
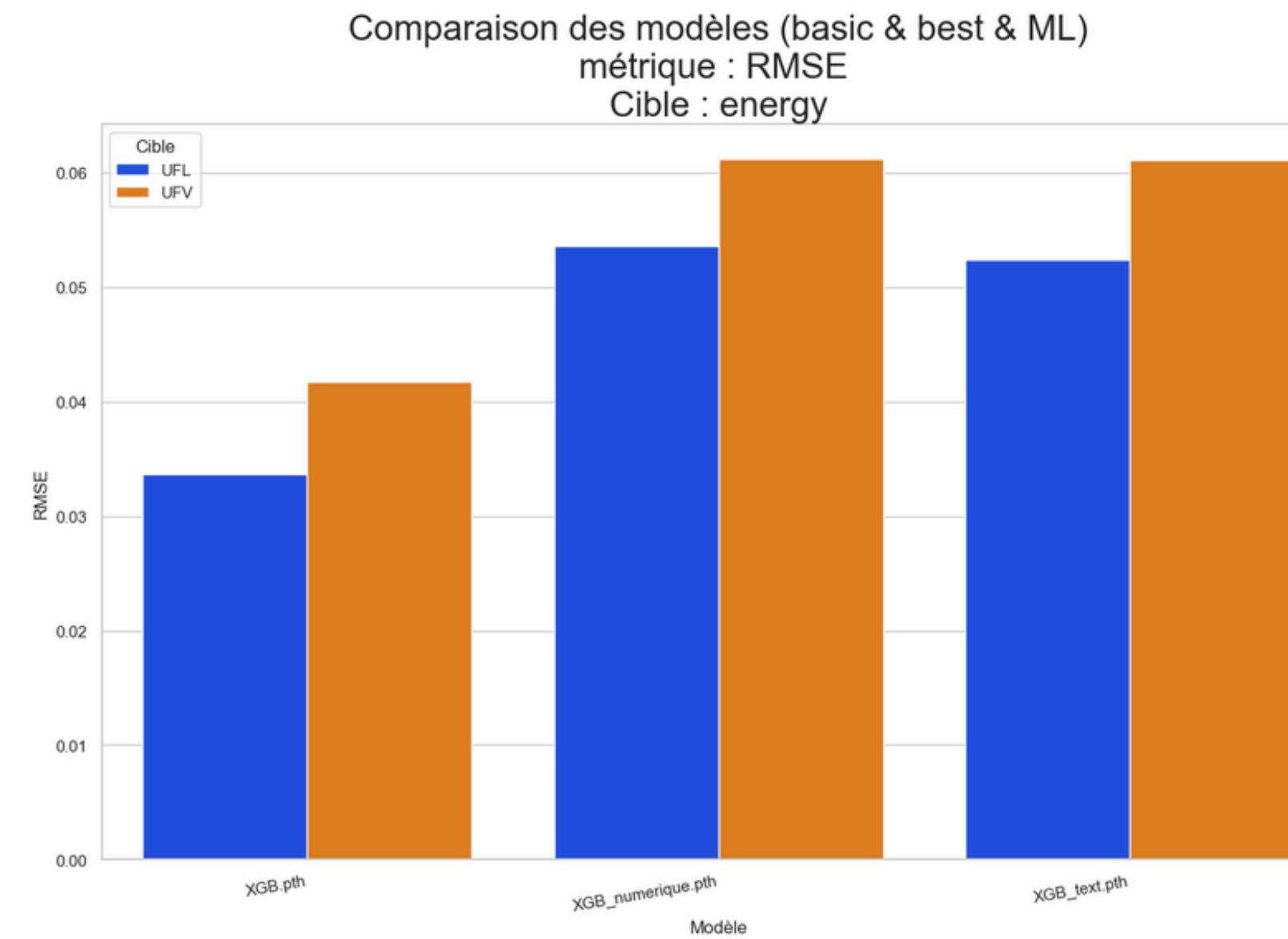
Meilleur Modèle pour prédire UFL: XGBoostRegressor
Dataset: MCA sur tous les Libellés + VC
MAE score: 0.2591, R² score: 0.8521



Plot des Résidus.
Ecart-type = 0.2842, Erreur Absolue Moyenne = 0.2591



Effets des valeurs numériques et textuelles sur XGBoost

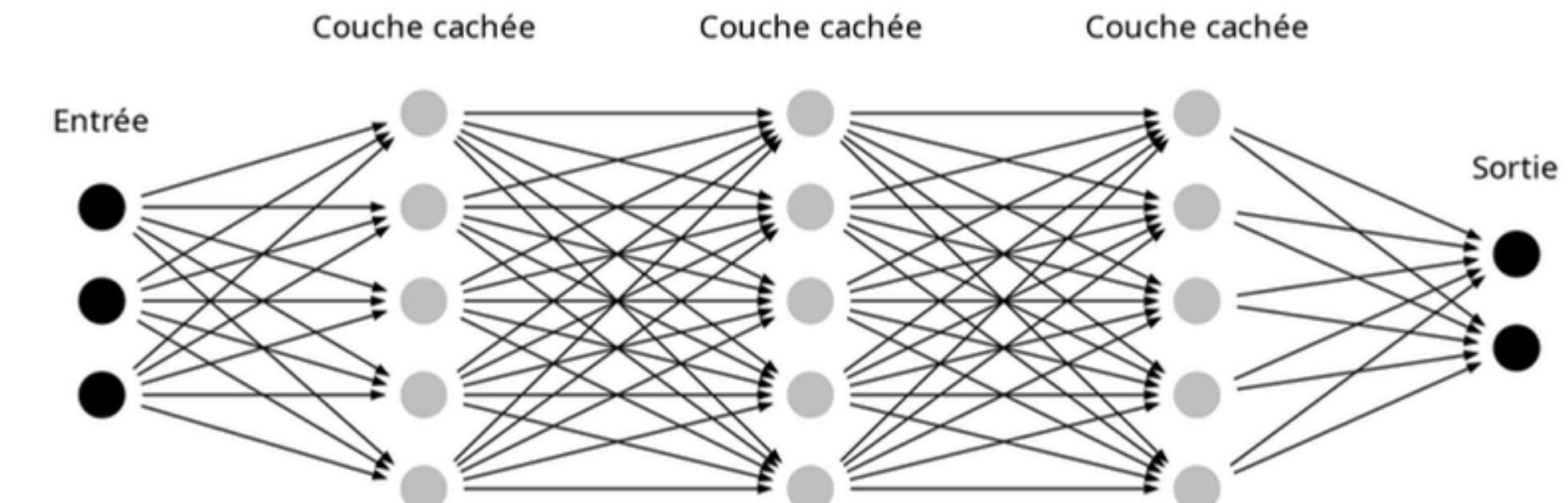


Qu'est-ce qu'un **LLM** ?

Le chat mange des croquettes

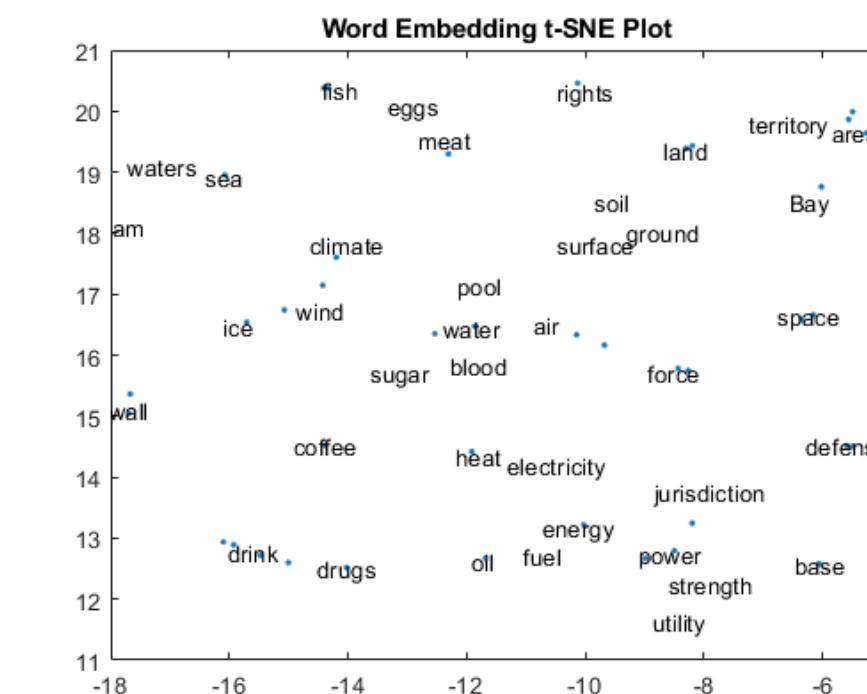
[Le] [chat] [mange] [des] [croqu] [ette] [s] [.]

Tokenization



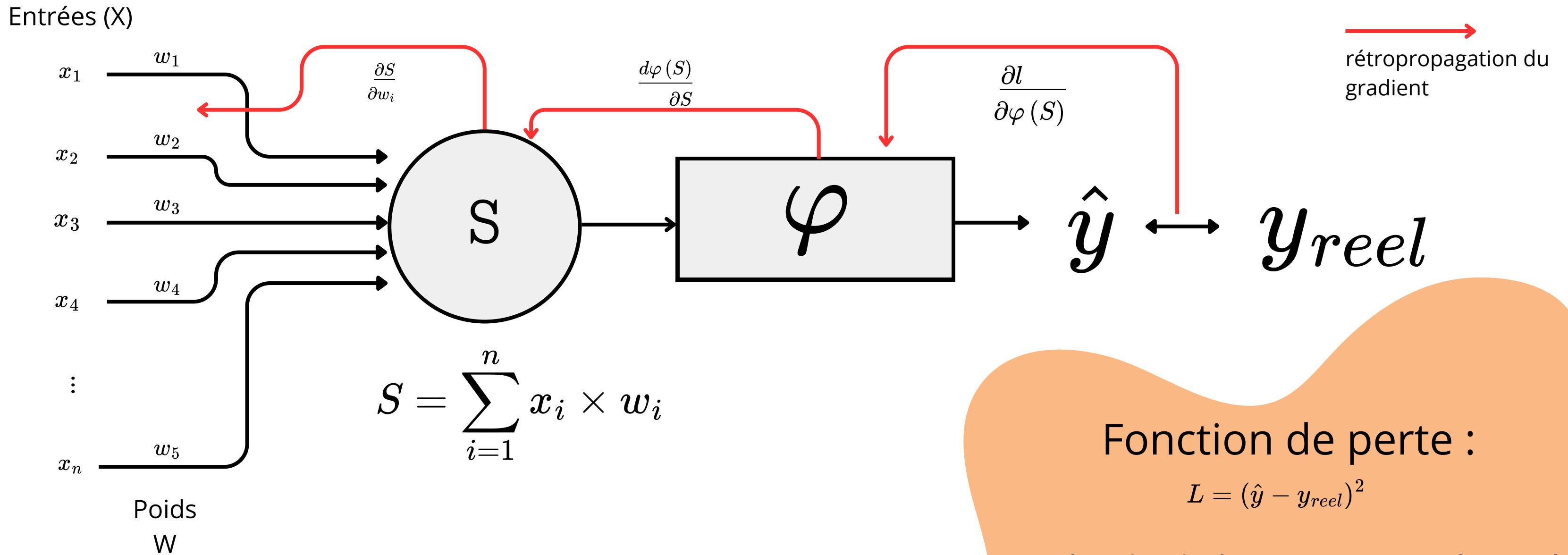
Structure d'un réseau de neurone (NN)

Les grands modèles de langue (ou *lLM*) sont la combinaison d'un **tokenizer**, d'un **embedder** et d'un **réseau de neurones** (avec une architecture particulière appelée transformer)

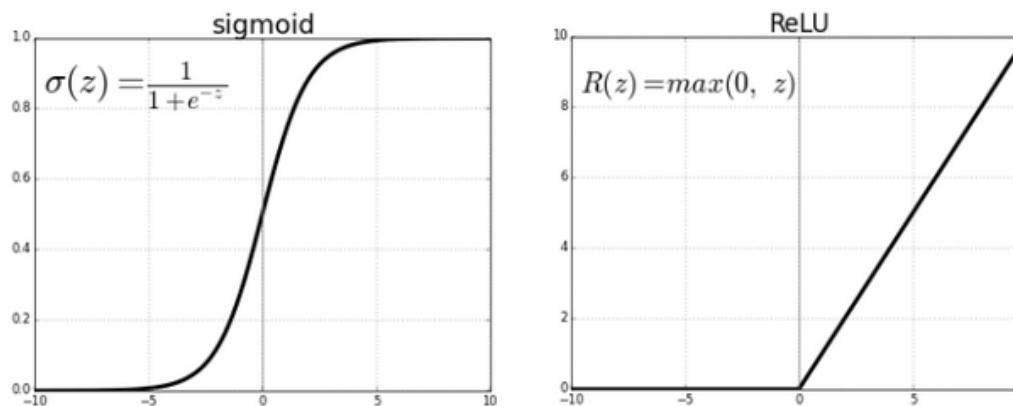


Embedding

Qu'est-ce qu'un **réseau de neurones (RN)** ?



φ : fonction d'activation
ReLU, tanh,...



Fonction de perte :

$$L = (\hat{y} - y_{reel})^2$$

que l'on cherche à minimiser. Pour cela on utilise la méthode dite de **descente du gradient**, consistant à se déplacer dans la direction de plus grande pente à chaque étape

Revolution du traitement automatique de la langue naturelle : les *large language model (LLM)*

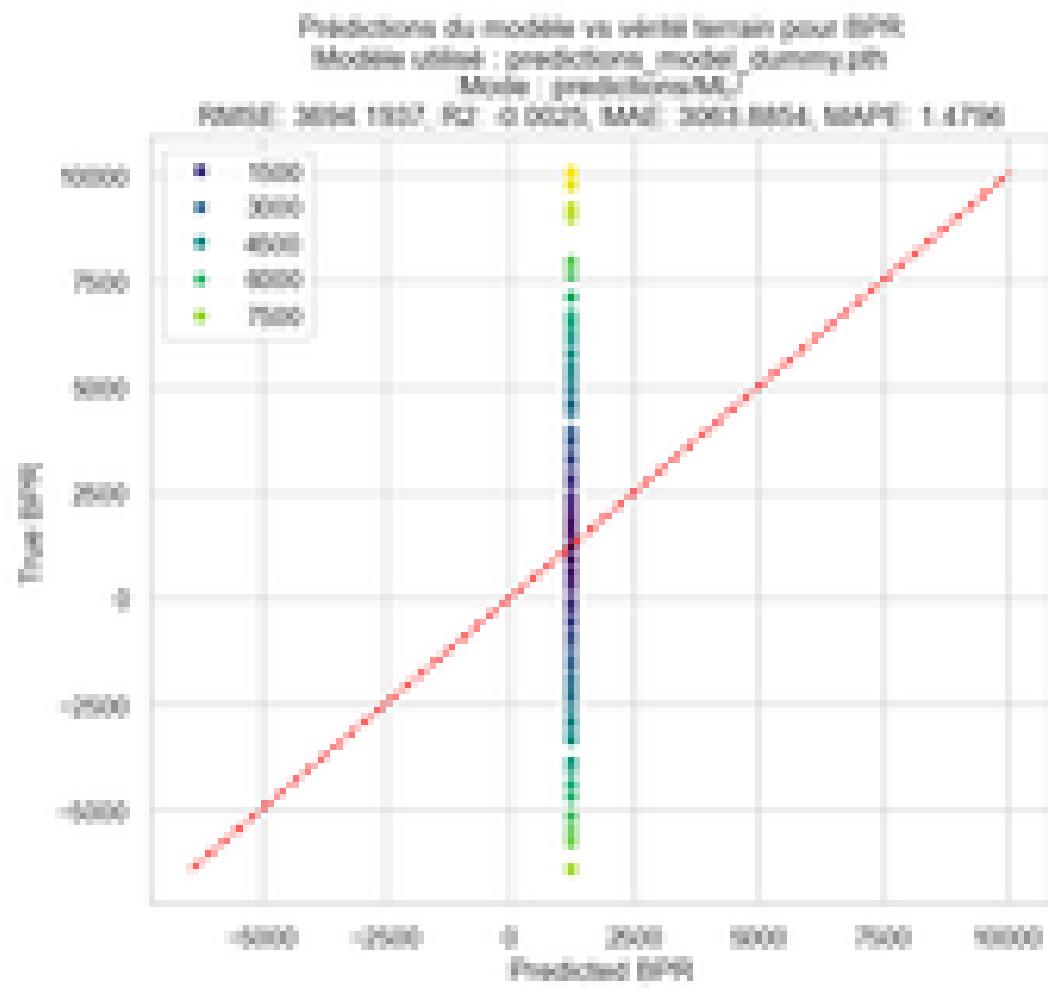
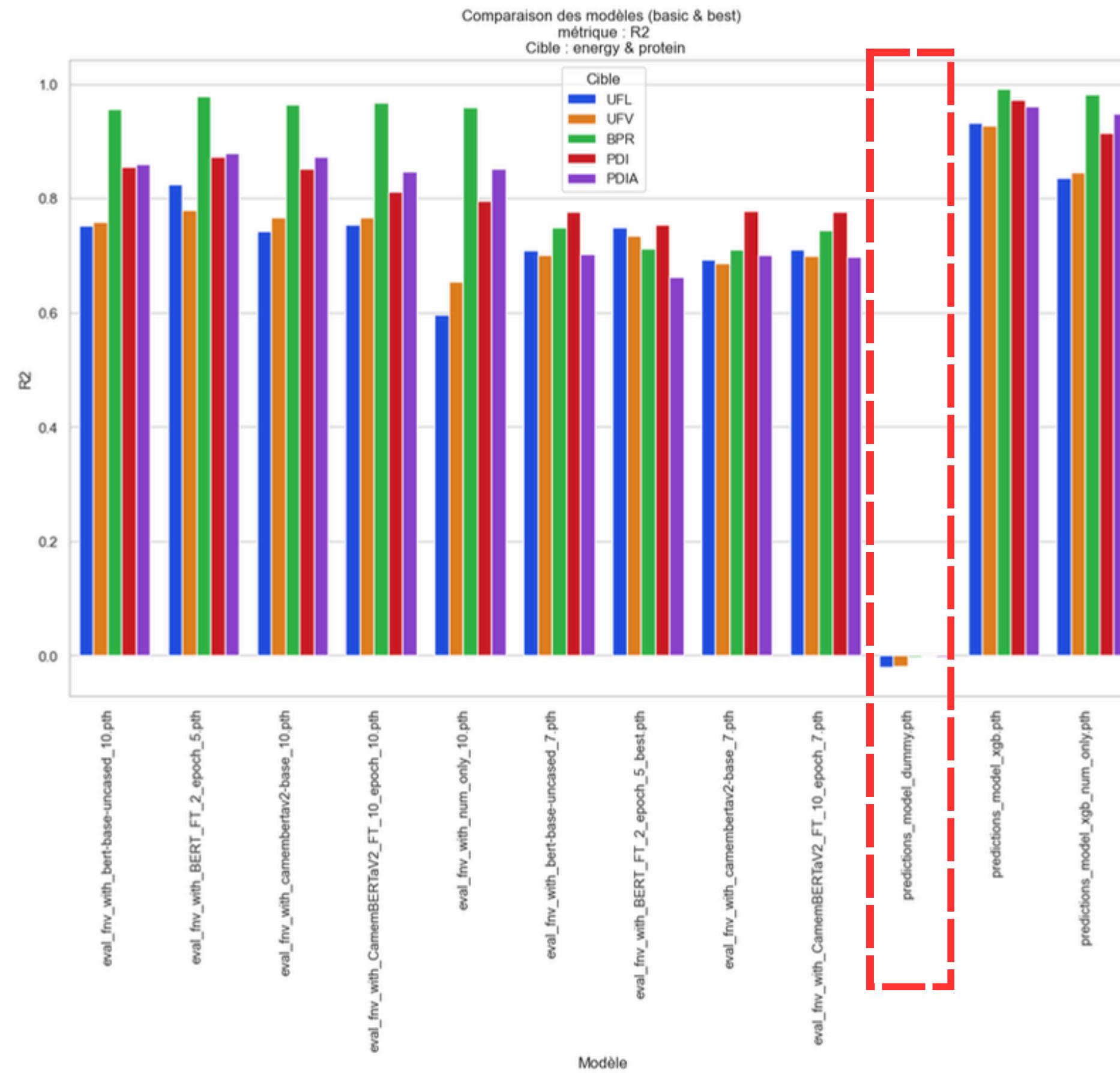
- Une nouvelle architecture : “Attention Is All You Need” 2017
- Les premiers LLMs : BERT GPT, GPT-2, RoBERTa 2018-2019
- De nos jours => Paysage des LLMs très diversifié

=> Innovation architecturale + Passage à l'échelles

=> Performances et Transférabilité



Résultats réseaux de neurones : comparaison à un régresseur “dummy”



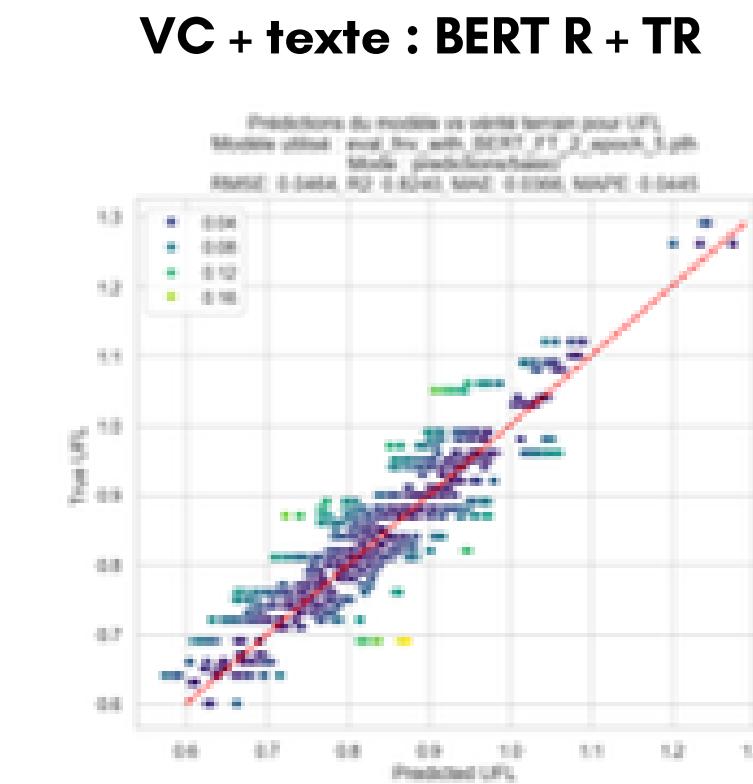
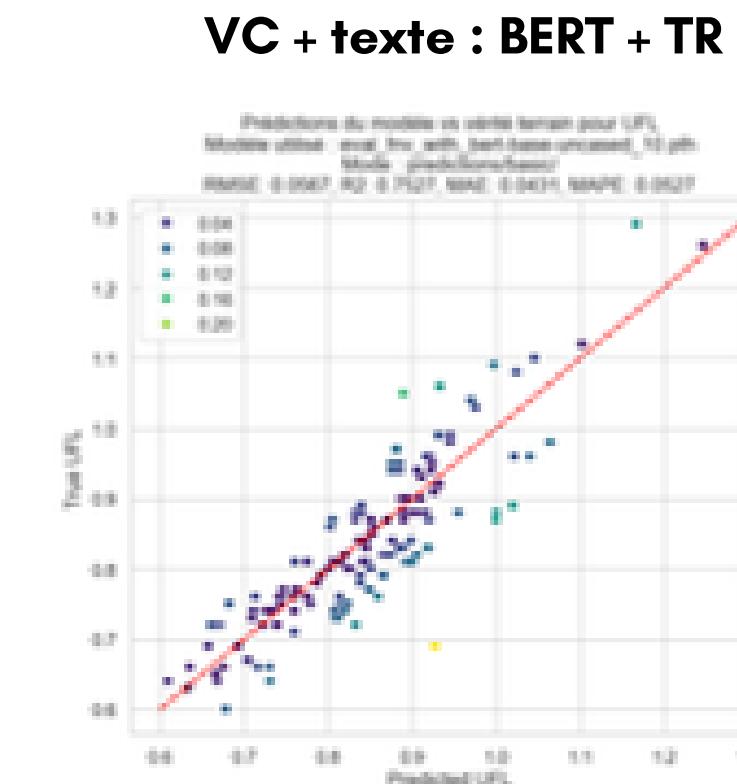
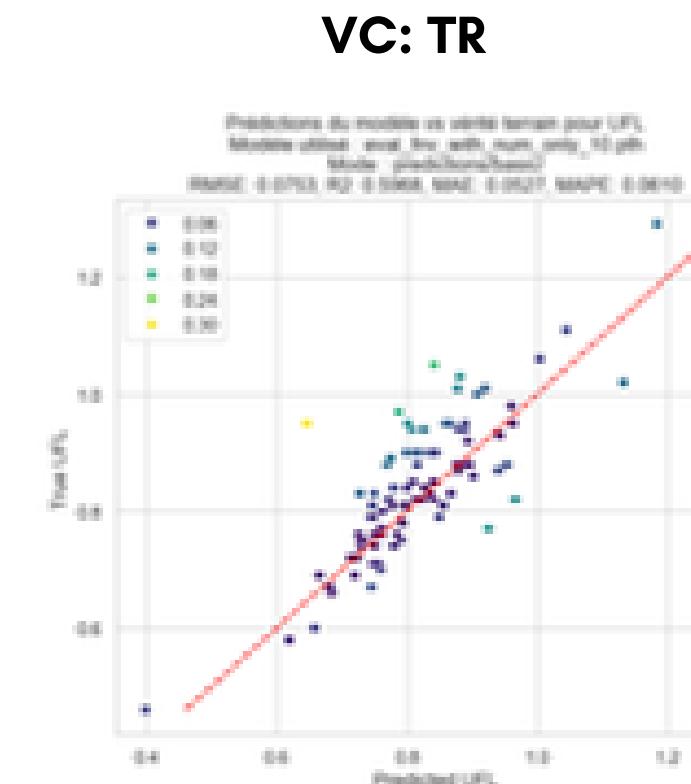
- Meilleure performances qu'un régresseur dummy

Résultats : des performances différentes en fonction des approches

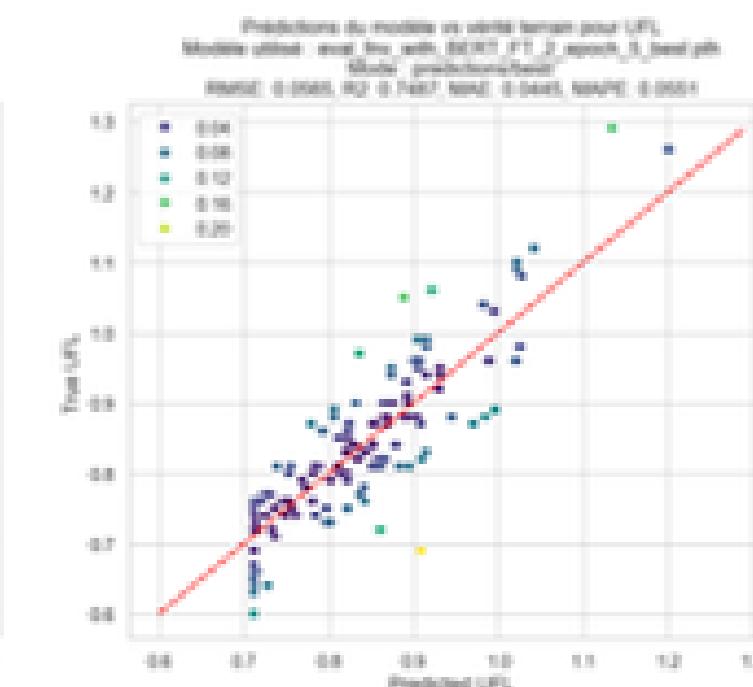
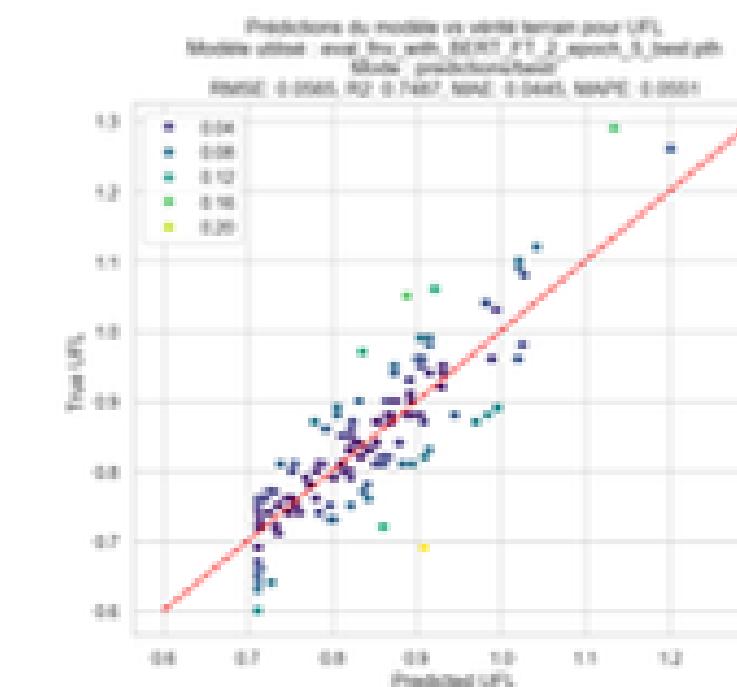
Tête de régression

Données et modèle profond

Basic

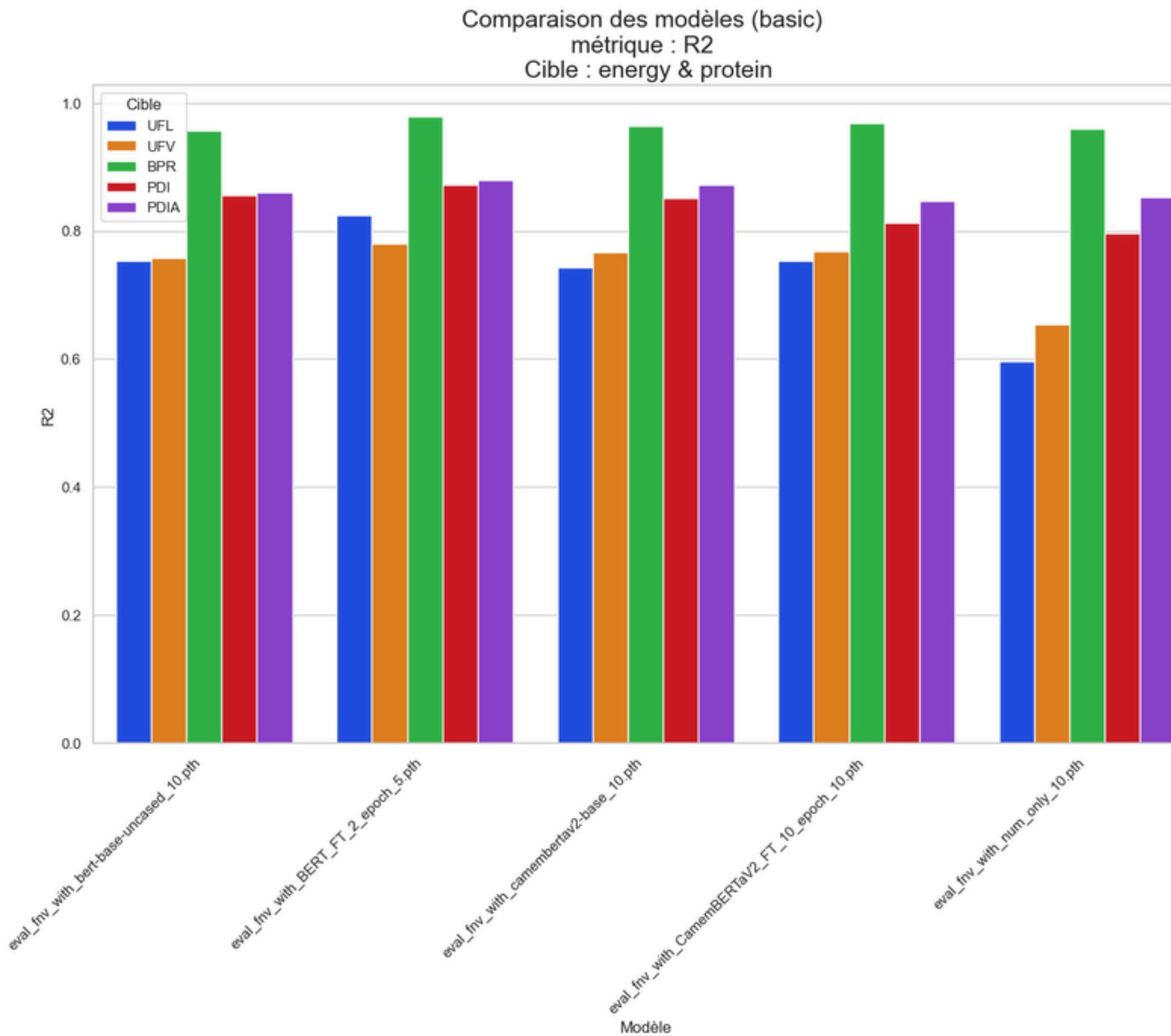


Best



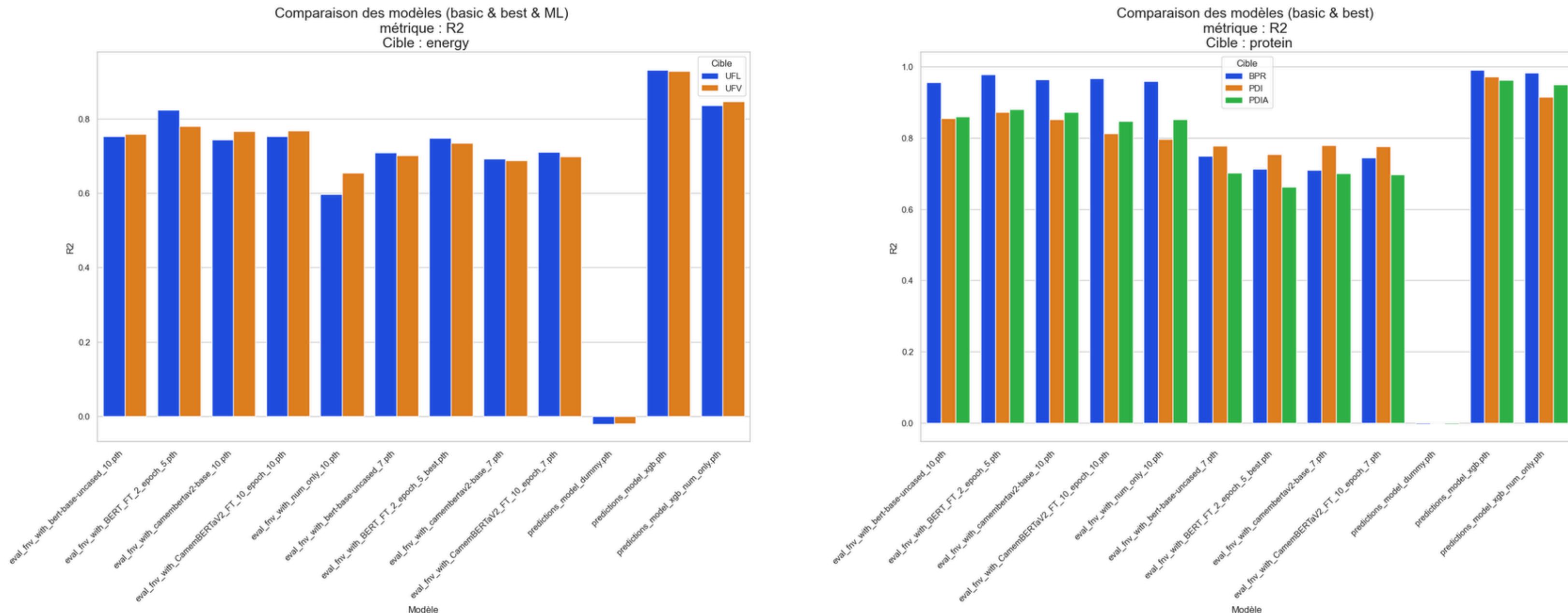
- Une tête de régression neuronale appliquées aux valeurs chimiques permet d'extraire des résultats
- Le texte contient en effet de l'information
- Le raffinage du modèle de language permet d'améliorer les performances
- Correction nécessaire des têtes des têtes de régressions optimisées

Résultats : performances diffèrent également sur la nature des cibles



- Grandes disparités au sein des cibles à prédire :
 - homogénéité des valeurs énergétiques
 - hétérogénéité des valeurs protéiques
 - des performances sur le BPR particulièrement élevées

Résultats : performances diffèrent également sur la nature des cibles



- Meilleure performances sur les protéines
- Homogénéité des approches "basic" et "best" sur l'énergie
- Davantage d'hétérogénéité sur les performances associées aux protéines



Résultats : des performances en approches non neuronales qui dépassant celles de l'apprentissages profond neuronnale

