



## **Évaluation de méthodes d'apprentissage profond pour les données spectroscopiques et la chimiométrie : un cas d'étude pour les applications agro-environnementales**

Florent Abdelghafour

Maxime Metz



# Contexte : LabCom Aioly



## Étudier les architectures de réseaux de neurones existantes

→ Identifier leurs forces et limites pour les données spectrales.

## Ouvrir les outils à la communauté scientifique et industrielle

→ Favoriser la collaboration, la reproductibilité et l'innovation.



## Développer des architectures dédiées aux données spectrales

→ Prendre en compte leurs spécificités (structure, bruit, redondance).

## Questionner et améliorer les systèmes de mesure

→ Optimiser la qualité des données à la source pour de meilleures performances.



Artificial Intelligence  
& Optics LaboratorY  
(AIOLY)



## Questions opérationnelles :

**Peut-on entraîner des NN's «from scratch» ?**

Actuellement les réseaux sont réutilisés et affinés pour une nouvelle tâche

**Quelle architecture semble être la plus prometteuse ?**

Transformers, RNN, LSTM, CNN, State-space model

**Peut-on avoir un seul jeu de paramètres pour plusieurs bases de données ?**

Pour permettre un transfert de connaissances simple

**Quel rapport entre gain de performance et complexité de réglage des modèles ?**

Peut-on augmenter la complexité même sur une petite base de données ?

# Un benchmark des architectures classiques en AP



## L'étude

**Comparaison de 4 typologies d'architectures :** CNN, ResNet, Inception, Transformer

**Etude réalisée sur trois bases de données agroenvironnementales :** Mangues, Sols, Blé

**Deux tâches distinctes :** classification et régression

**Des tailles de bases de données variées :** de 10 000 à 150 000 mesures

# Un benchmark des architectures classiques en AP



	Mangues [1]	OSSL [2]	WHEAT [3]
<b>Nb échantillons</b>	11 691	23 447	147 096
<b>Plage spectrale (nm)</b>	684-990	400-2500	975-1645
<b>Nb variables spectrales</b>	103	1075	200
<b>Y</b>	% DM	oC g.kg <sup>-1</sup>	30 variétés
<b>Spécificité</b>	Prédiction de l'année suivante	Hétérogénéité des données	Grand nombre de classes

[1] N.T. Anderson et al., 2020 « Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content » Postharvest Biology and Technology, 168,111202

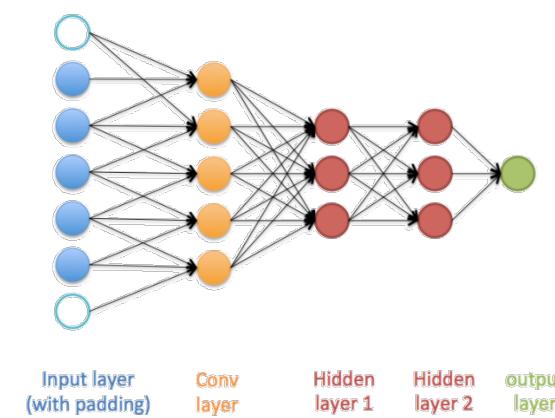
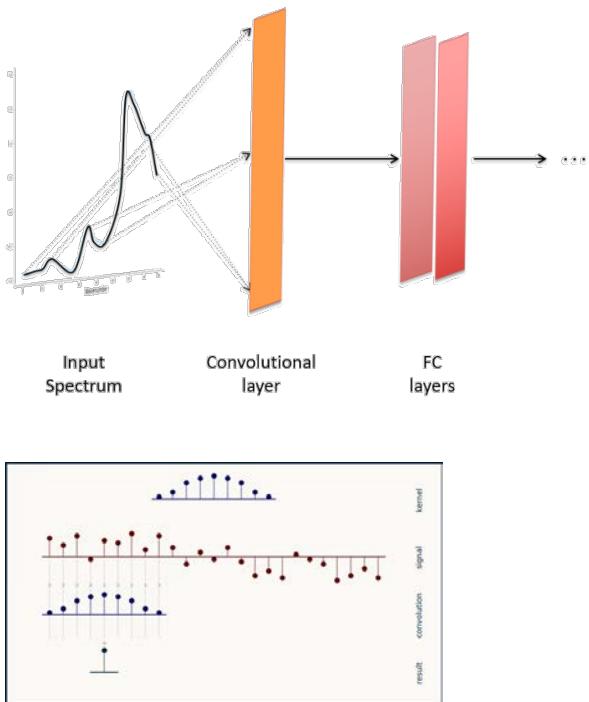
[2] J. L. Stefanelli et al., 2023 . Open Soil Spectral Library soil (OSSL): Building reproducible soil calibration models through open development and community engagement. bioRxiv 2023.12.16.57201

[3] Zhou, L., Zhang, C., Taha, M. F., Wei, X., He, Y., Qiu, Z., & Liu, Y. (2020). Wheat kernel variety identification based on a large near-infrared spectral dataset and a novel deep learning-based feature selection method. Frontiers in Plant Science, 11, 575810.

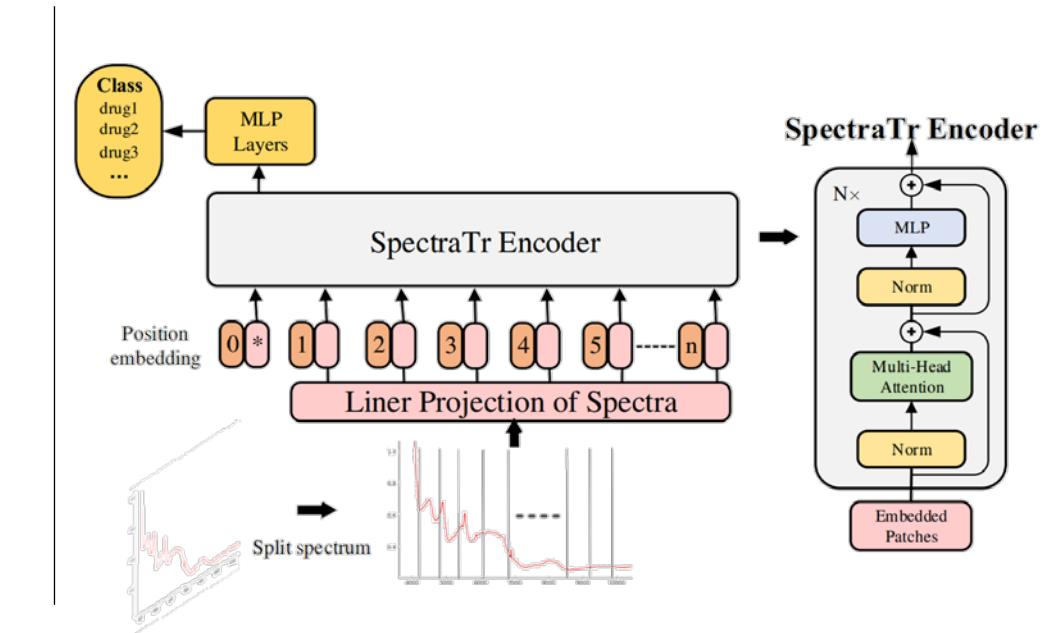
# Un benchmark des architectures classiques en AP

## Les architectures :

**CNN 1-D: Cuinet [4]**



**Transformer : SpectraTr [5]**



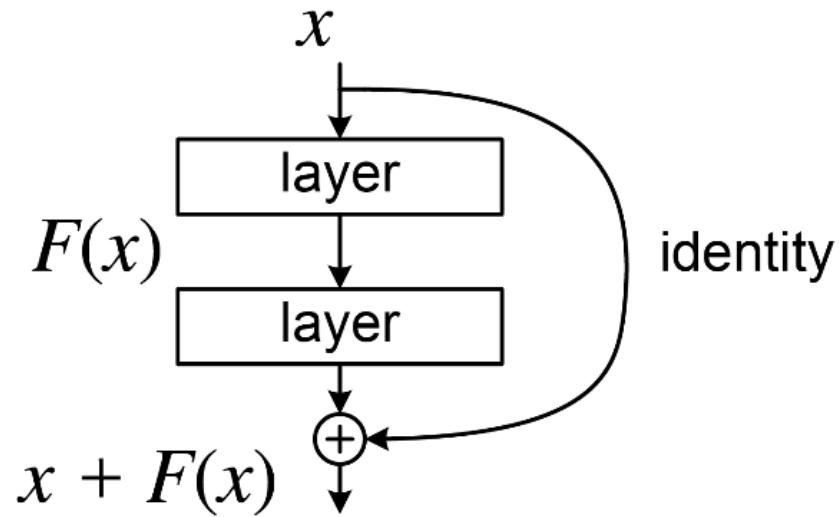
[4] Cui, C., Fearn, T., 2018. Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration. *Chemometrics and Intelligent Laboratory Systems* 182, 9–20.

[5] Fu P, et al. SpectraTr: A novel deep learning model for qualitative analysis of drug spectroscopy based on transformer structure[J]. *Journal of Innovative Optical Health Sciences*, 2022, 15(03): 2250021

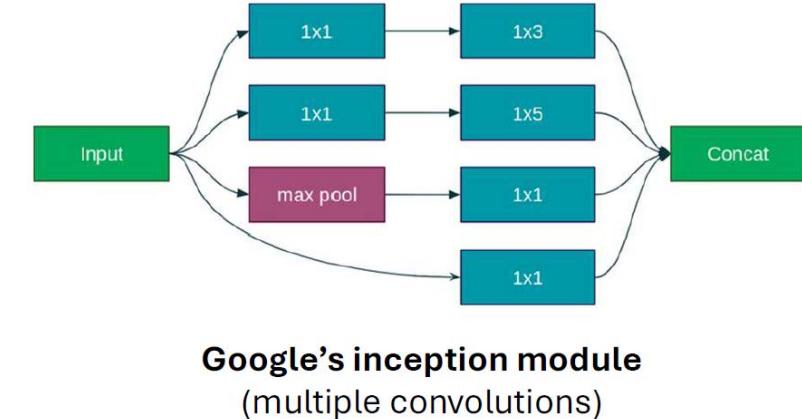
# Un benchmark des architectures classiques en AP

## Les architectures :

Resnet : Resnet-1D [6]



Inception : DeepSpectra [7]



**Google's inception module**  
(multiple convolutions)

[6] Martins, J., Guerra, R., Pires, R., Antunes, M., Panagopoulos, T., Brázio, A., Afonso, A., Silva, L., Lucas, & Cavaco, A. (2022). SpectraNet-53 : A deep residual learning architecture for predicting soluble solids content with VIS–NIR spectroscopy. *Computers And Electronics In Agriculture*, 197, 106945.

[7] Zhang, X., Lin, T., Xu, J., Luo, X., & Ying, Y. (2019). DeepSpectra : An end-to-end deep learning approach for quantitative spectral analysis. *Analytica Chimica Acta*, 1058, 48-57.

## Paramètres benchmark :

- **Paramètres d'apprentissage :**
  - Taux d'apprentissage (LR)
  - Taux de Dropout (DP)
- **Paramètres dimensions des modèles :**
  - Taille des filtres
  - Nombres de couches
  - Dimensions des « blocs »  
*e.g. (nb encodeurs « transformers », nb mécanismes d'attention)*

Au total : 3-60 Combinaisons de paramètres testés selon les combinaisons architecture -BDD

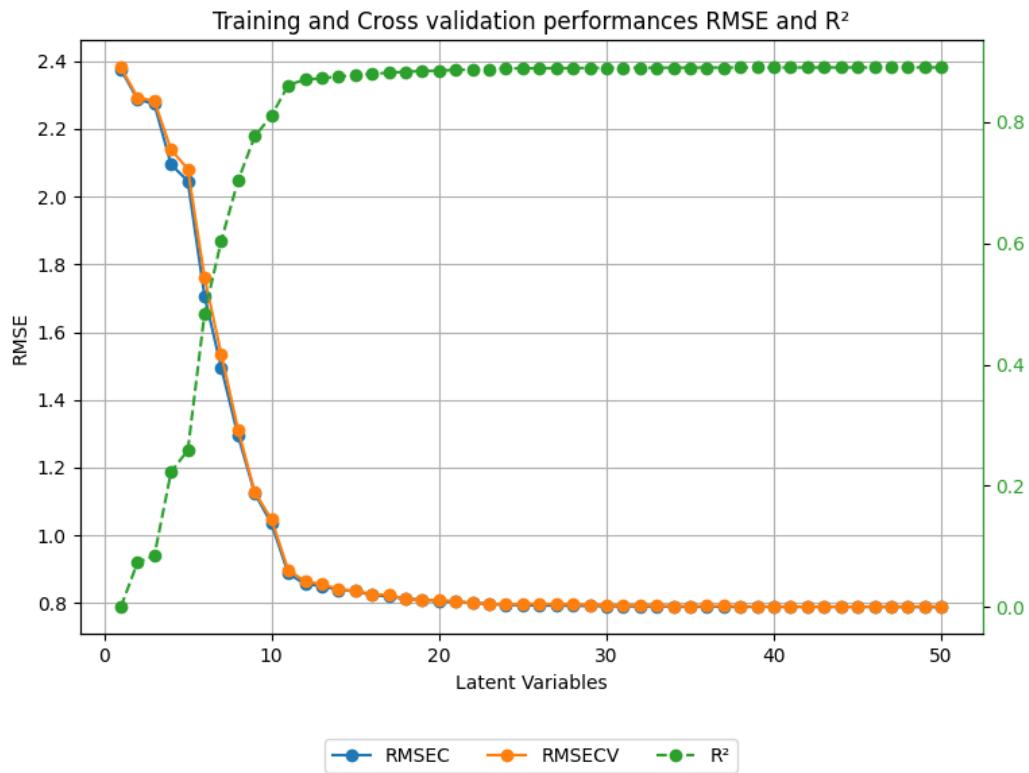
# Un benchmark des architectures classiques en AP



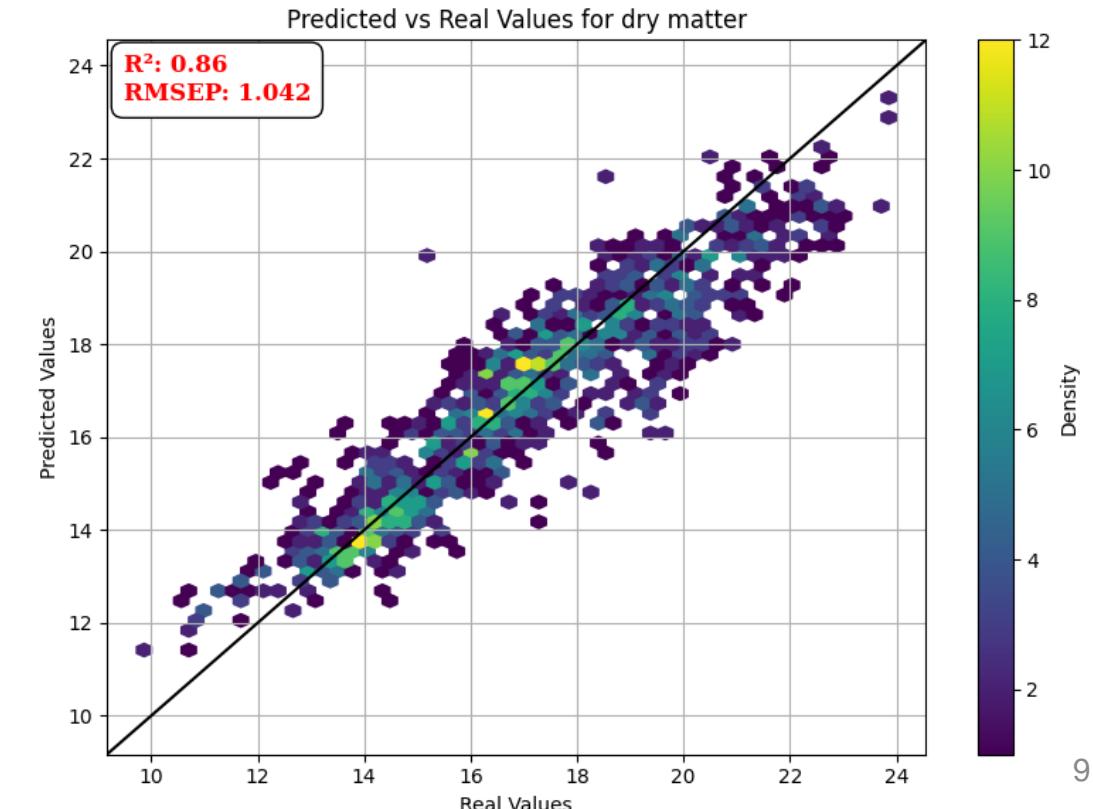
## Résultats Mangues



Modèle PLS (SVG d1)



Test 12 LV



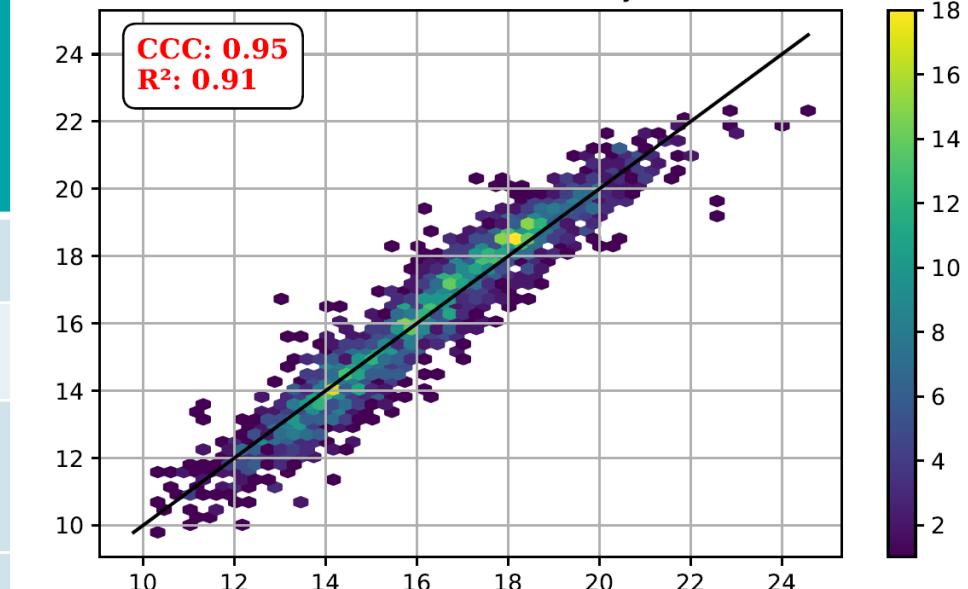
# Un benchmark des architectures classiques en AP



## Résultats Mangues

Model	RMSEP	R <sup>2</sup>	N paramètres	Best epoch
CuiNet	0.892	0.88	4 513	721
DarioNet	0.888	0.89	22 637	563
DeepSpectra	0.989	0.86	36 975	210
ResNet18	1.020	0.85	61 431	648
ResNet50	0.995	0.86	255 607	272
ViT-1D	<b>0.746</b>	<b>0.91</b>	<b>814 907</b>	<b>1952</b>

Test ViT



State of the art perf : R<sup>2</sup> = 0,898 RMSE = 0,853

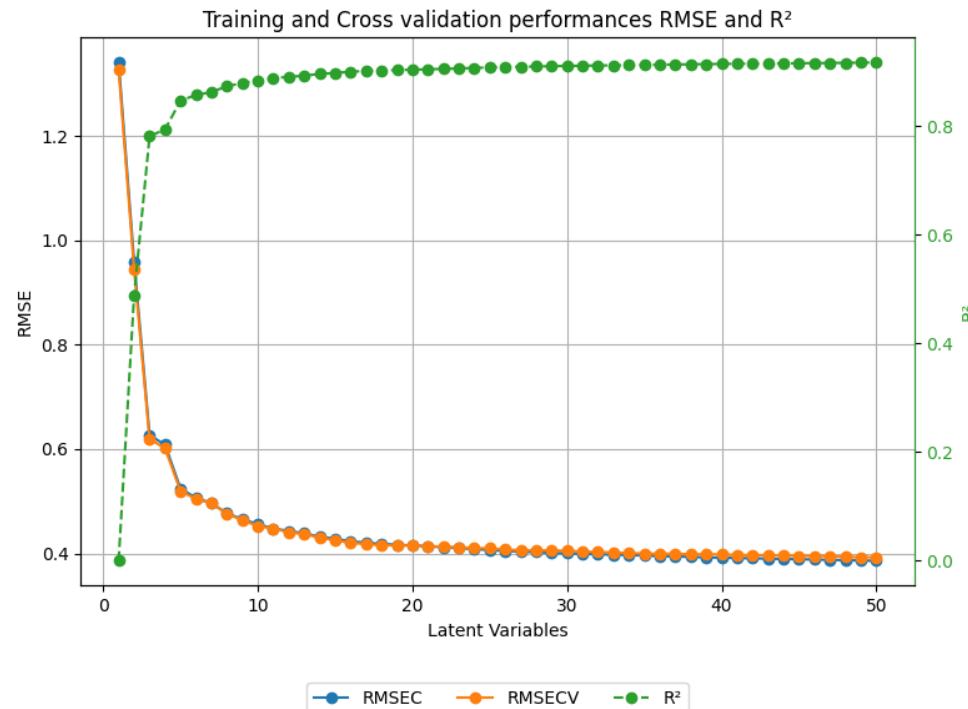
Passos and Mishra 2021 :  
<https://doi.org/10.1016/j.chemolab.2021.104354>

# Un benchmark des architectures classiques en AP

## Résultats OSSL



## Modèle PLS

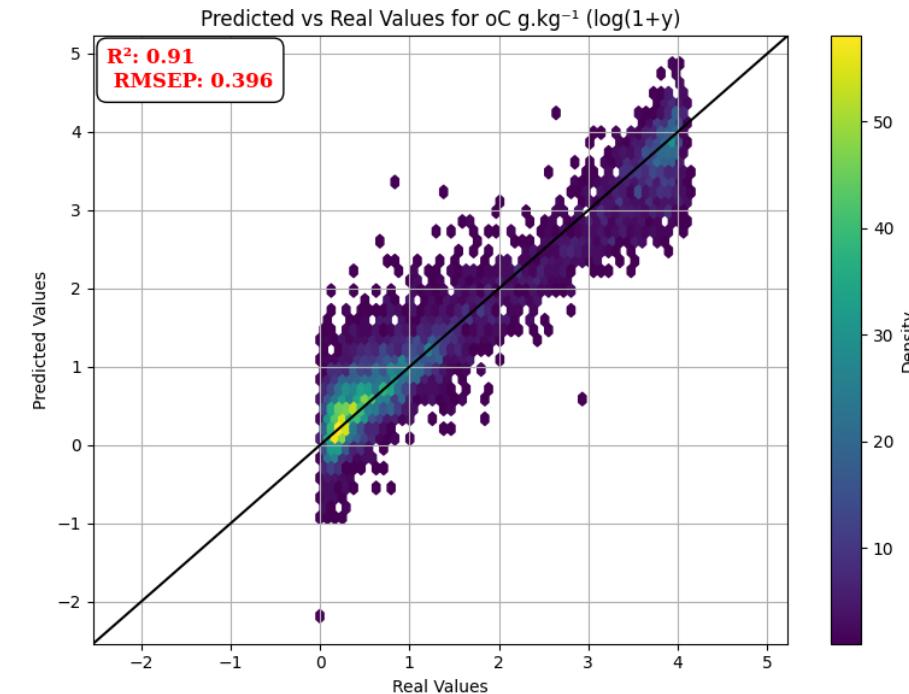


Valeurs  $\log(1 + y) \rightarrow$  Echelle originale  $\exp(y-1)$

**R<sup>2</sup> = 0.81**

**RMSEP = 7.449 g.kg<sup>-1</sup>**

**Test 36 LV**

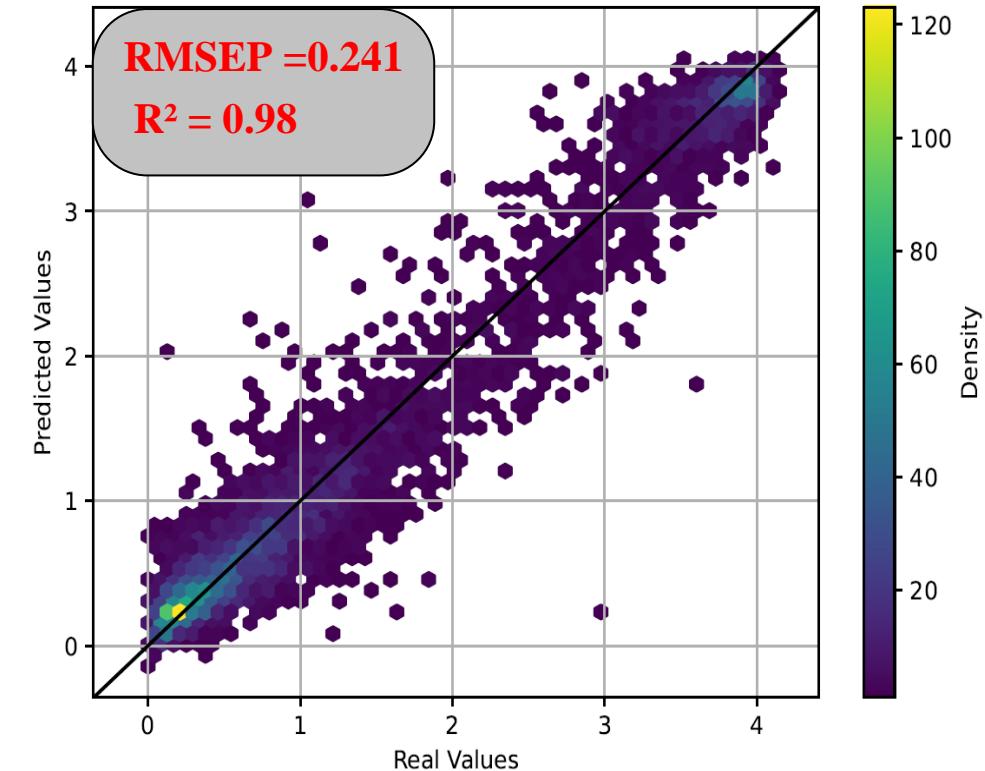


# Un benchmark des architectures classiques en AP

## Résultats OSSL

Model	RMSEP	R <sup>2</sup>	N parameters	Best epoch
CuiNet	0.272	0.95	39 541	318
DarioNet	<b>0.234</b>	<b>0.96</b>	147 191	154
DeepSpectra	<b>0.229</b>	<b>0.97</b>	<b>370 697</b>	<b>57</b>
ResNet18	0.264	0.96	63 377	586
ResNet50	0.289	0.95	257 553	1000
ViT-1D	<b>0.241</b>	<b>0.98</b>	547 249	1673

Test ViT

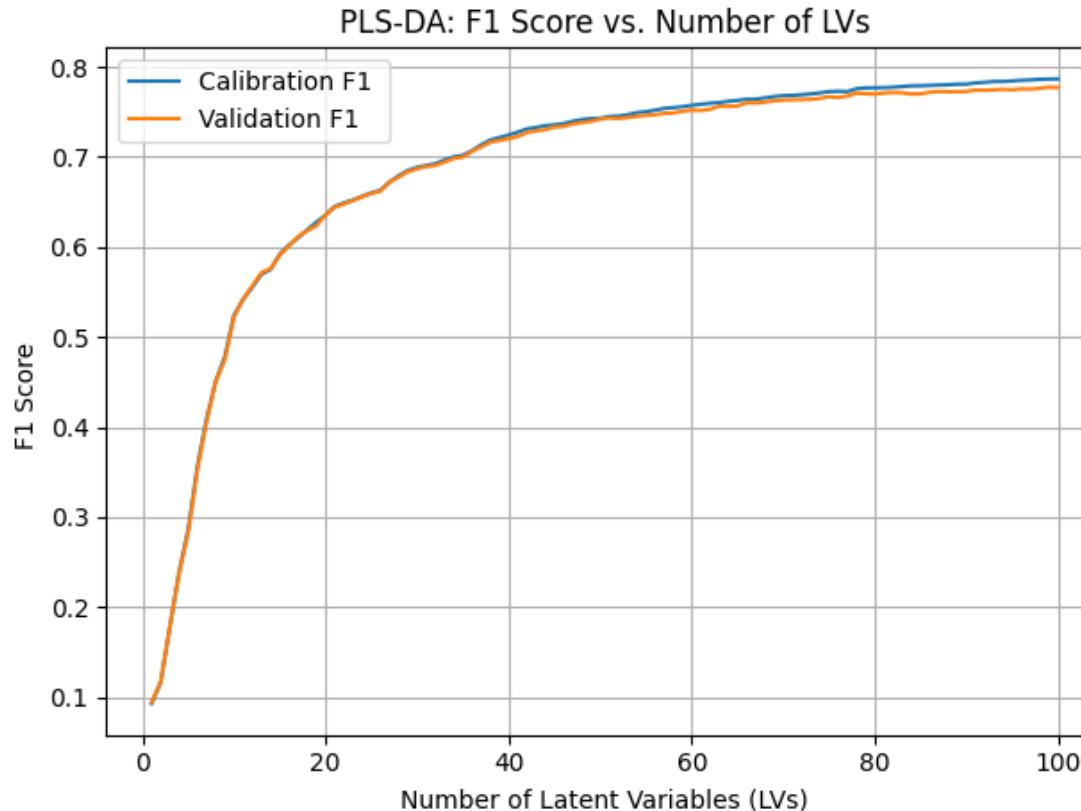


# Un benchmark des architectures classiques en AP

## Résultats Wheat

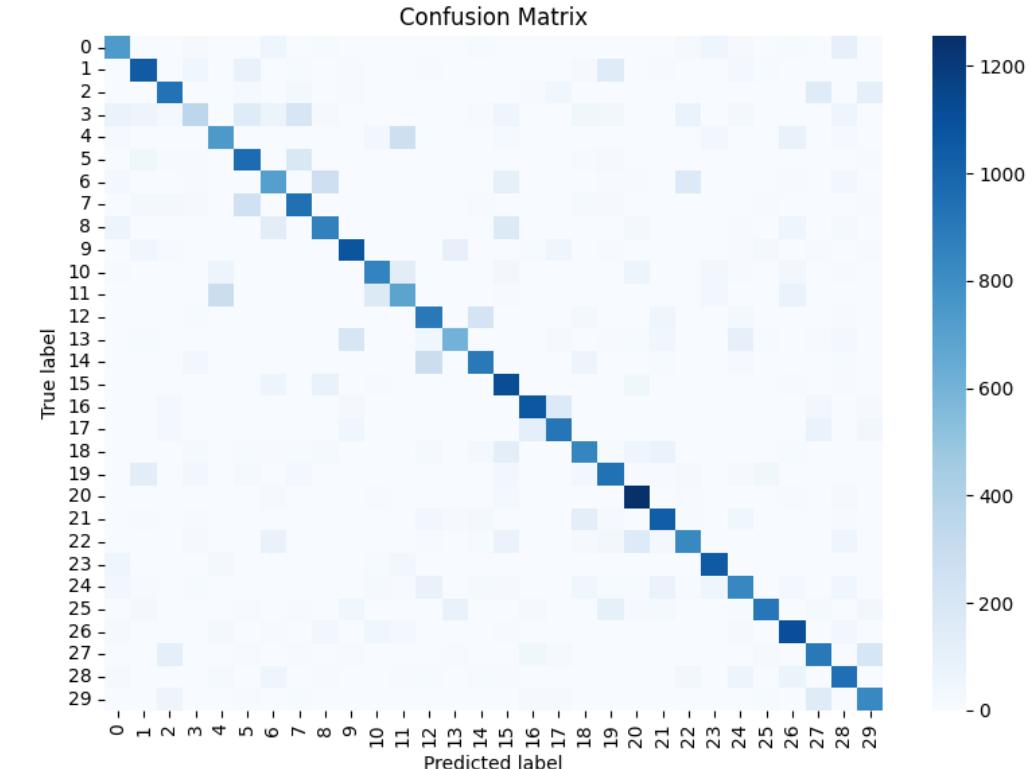


### Modèle PLS - LDA



### Test 60 LV

Accuracy = 0.71    F1 = 0.71  
Precision = 0.69    Recall = 0.74



## Résultats Wheat

Model	F1	Precision	Recall	N parameters	Best epoch
CuiNet	0.63	0.64	0.65	44 382	236
DarioNet	0.71	0.72	0.72	163 556	309
DeepSpectra	<b>0.85</b>	<b>0.85</b>	<b>0.85</b>	<b>414 814</b>	<b>207</b>
ResNet18	0.68	0.72	0.70	65510	1000
ResNet50					
ViT-1D	0.75	0.72	0.79	4 193 702	1587

## Conclusion :

**Peut-on entraîner des NN's «from scratch» ?**

Capable d'apprendre « from scratch » sur des bases chimiométriques sans augmentation de données

**Quelle architecture semble être la plus prometteuse ?**

Actuellement les architectures type DeepSpectra ou Transformer semblent de distinguer

**Peut-on avoir un seul jeu de paramètres pour plusieurs bases de données ?**

Les architectures actuelles en chimiométrie ne permettent pas de s'adapter à la taille des données d'entrée

**Quel rapport entre gain de performance et complexité de réglage des modèles ?**

Permet de reproduire ou dépasser les résultats de l'art sur les grandes bases de données même avec une complexité élevée

Sensibles aux hyperparamètres (défaut de convergence / instabilités)

Certains paramètres intuitifs selon les propriétés des spectres et de l'application

## Perspectives :

- Comparer avec des approches non linéaires en chimiométrie (ex : LW-PLSR)
- Etudier le comportement de la performance en fonction des paramètres
- Trouver des compromis entre performances et architectures génériques

## Futurs travaux/volonté dans le LabCom :

- Développer avec les chimiométriciens des stratégies de DL adaptées, des premiers résultats montre déjà l'intérêt des approches sans modifications
- Nourrir le DL avec les connaissances métiers de la chimiométrie
- Manque d'outils et de protocoles standard  
⇒ Mise à disposition d'outils didactiques open source pour le benchmark dans le cadre du Labcom AIOLY

INRAe

どうもありがとう

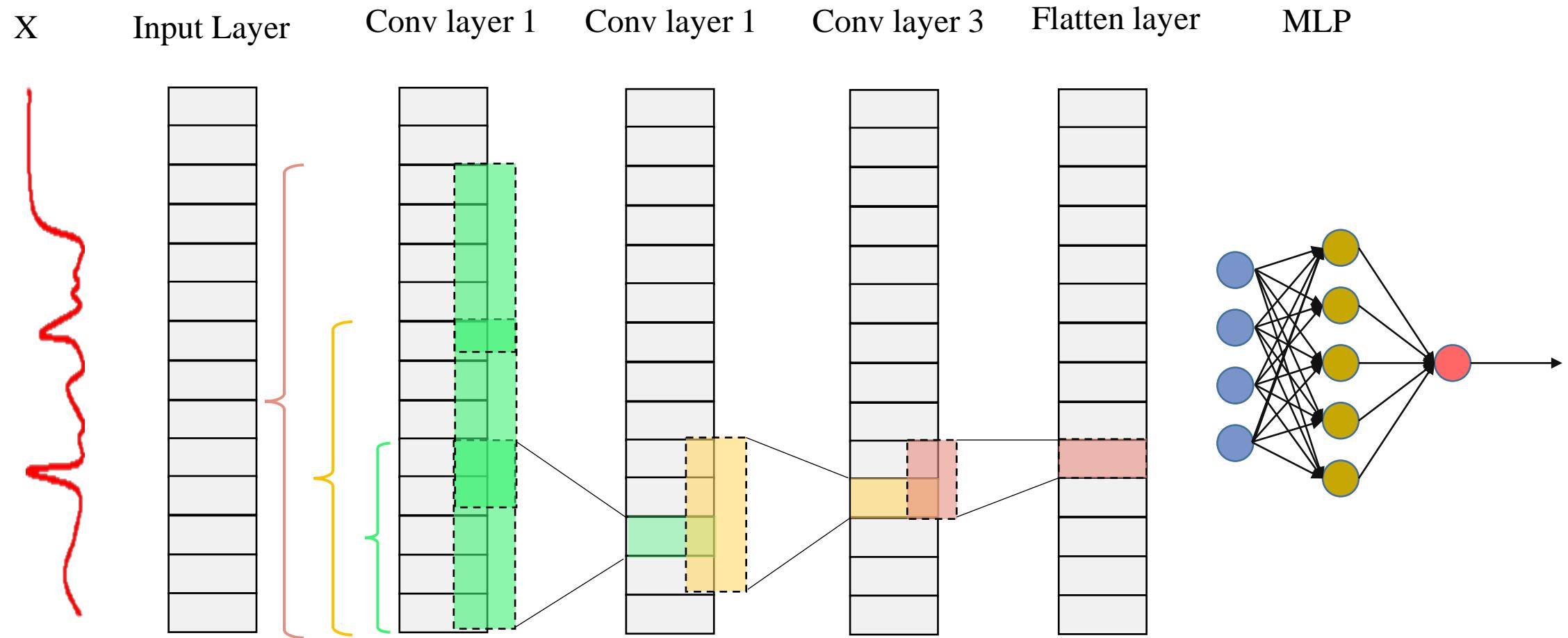
Merci شكرأ

Gracias

Thank You



# CNN-1D



# Base de données Mangues : prédition %DM



NIR [684 – 990]nm,  $\lambda = 103$  variables (résolution 3 nm)  
11 691 spectres (4675 fruits) répliquats  
 $Y = \% DM \in [9,25]$   
4 saisons de récoltes (2015-2018)  
⇒ problématique transfert d'étalonnage / écarts de distribution

Jeu nettoyé (DOI: 10.1016/j.chemolab.2021.104287 D. Passos)  
Outliers ⇒  $T^2$  Hotelling and  $Q$  résidus (PLS)  
**Données brutes filtrées** + prétraitées  
Prétraitements : SNV + dérivées 1<sup>ère</sup>, 2<sup>nd</sup>

**N.T. Anderson et al., 2020** « Achieving robustness across season, location and cultivar for a NIRS model for intact mango fruit dry matter content » Postharvest Biology and Technology, 168,111202 DOI: 10.1016/j.postharvbio.2020.111202

# Base OSSL : prédition taux oC dans les sols



VisNIR [350 – 2500]nm,  $\lambda = 1075$  variables (resolution 2 nm)

N spectres = 23447

$Y = oC \in [0,150] g.kg^{-1}$

Hétérogène : nature des sols, sites de mesures, années, instruments, standards de mesures

J. L. Stefanelli et al., 2023 . Open Soil Spectral Library soil (OSSL): Building reproducible soil calibration models through open development and community engagement. bioRxiv 2023.12.16.57201  
doi: <https://doi.org/10.1101/2023.12.16.572011>

# Base Wheat: classification variétés de grain de blé



Images hyperspectrales SWIR

147 096 spectres , moyenne pour chaque grain de blé

$[975 - 1645]nm$   $\lambda = 1075$  variables spectrales

Caméra ImSpector V10E; Spectral Imaging Ltd., Oulu, Finland

30 variétés de blé

Zhou, L., Zhang, C., Taha, M. F., Wei, X., He, Y., Qiu, Z., & Liu, Y. (2020). Wheat kernel variety identification based on a large near-infrared spectral dataset and a novel deep learning-based feature selection method.

*Frontiers in Plant Science*, 11, 575810. doi:10.3389/fpls.2020.575810