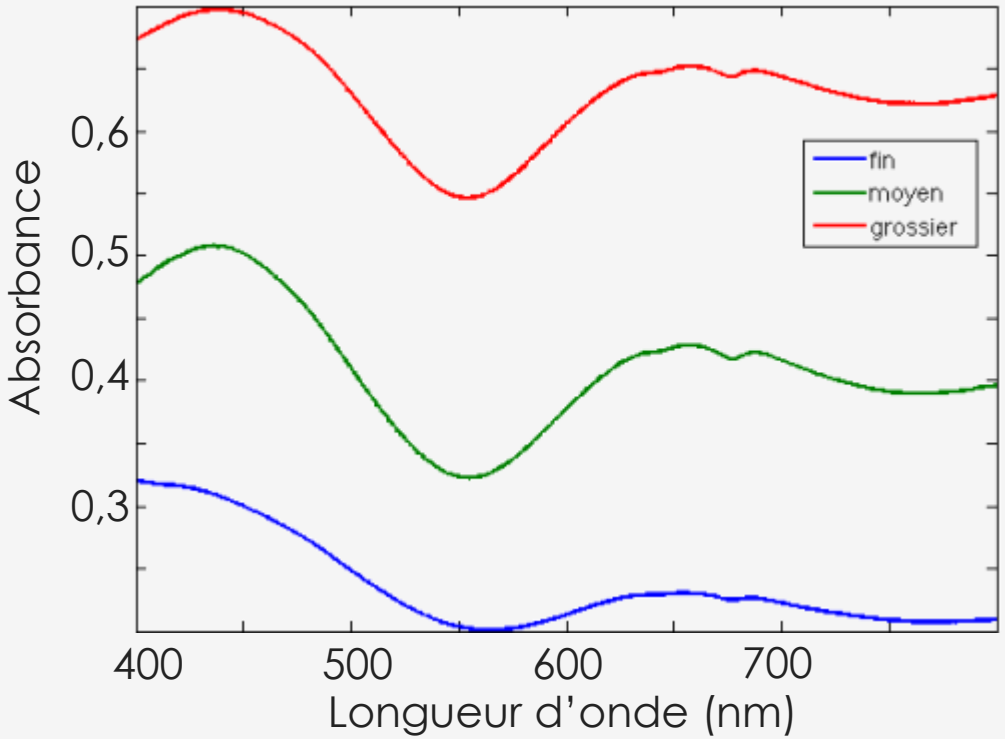
Comparaison de l’augmentation de données et du prétraitement pour l’apprentissage profond.

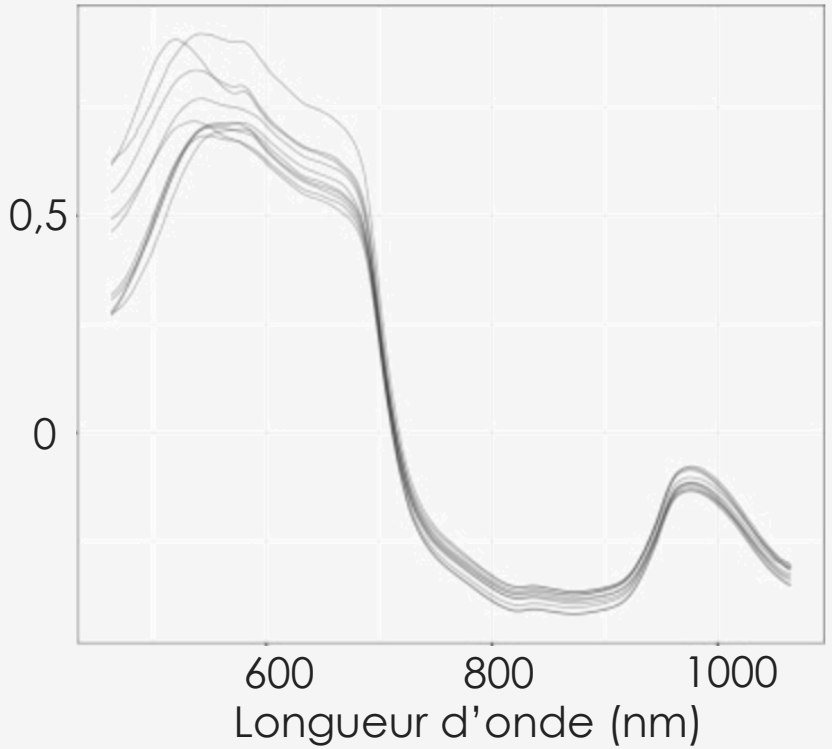
Ivy Tumoine, Florent Abdelghafour, Maxime Metz, Nicolas Grotus, Ryad Bendoula, Jean-Michel Roger
i.tumoine@pellencst.com

Déformation des spectres PIR

Les spectres peuvent être entachés d'**effets non-liés à la variable d'intérêt**.
➔ **Difficultés** pour les **modèles**



Effet de diffusion de la lumière : le même verre broyé à différentes granulométries [1]



Effet instrumental : le même fruit (kiwi) mesuré sur différents spectromètres [2]

[1] Phil Williams, Karl Norris, et al. « Near-infrared technology in the agricultural and food industries ». American Association of Cereal Chemists, Inc., 1987.
[2] M. Wohlers, et al. « Augmenting NIR Spectra in deep regression to improve calibration », *Chemometrics and Intelligent Laboratory Systems*, 2023, lien.

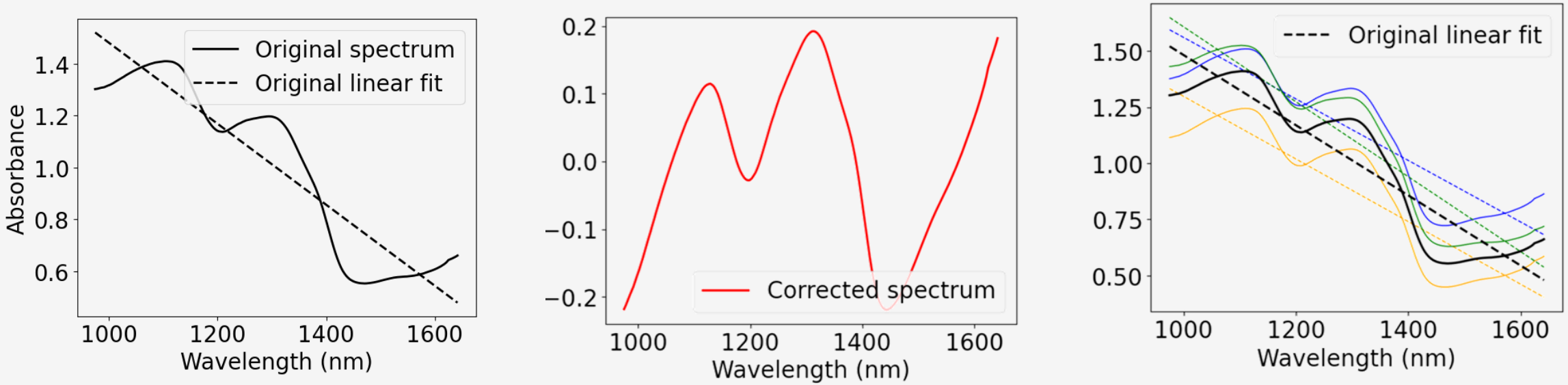
Encourager l'invariance aux effets

On voudrait que les **modèles** soient **invariants** à ces **effets** :

$$f(x) = f\left(g_{z\sim Pz}(x)\right) = y\ , f(.)\ le\ modèle\ et\ g_{z\sim Pz}(.)\ un\ générateur\ d'effets.$$

Approches possibles :

- **Prétraitements** : corriger par g^{-1}, chimiométrie.
- **Augmentation de données** : estimer un voisinage de x par échantillonnage dans g [3], apprentissage profond.



Pré-traitement VS augmentation de données pour la pente,
gauche : spectre pur, milieu : prétraitement, droite : augmentations (blé [4])

[3] O. Chapelle, J. Weston, L. Bottou, et V. Vapnik, « Vicinal Risk Minimization », in *Advances in Neural Information Processing Systems*, MIT Press, 2000, lien
[4] L. Zhou *et al*,«Wheat Kernel Variety Identification Based on a Large Near-Infrared Spectral Dataset and a Novel Deep Learning-Based Feature Selection Method», 2020, lien

Augmentations spectrales : questions de recherche

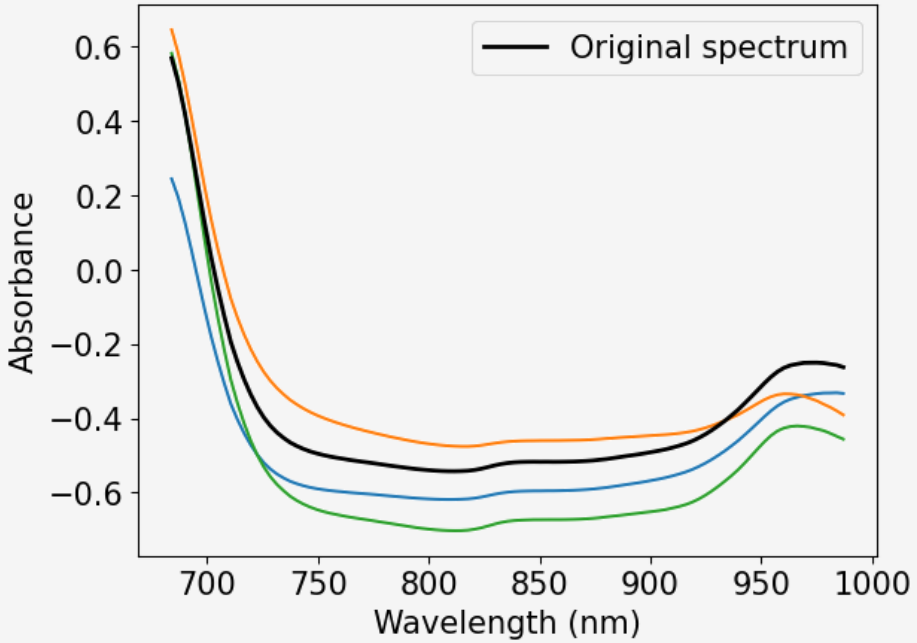
1. Pré-traitement VS augmentation de données, quel outil pour l'apprentissage profond ?

2. Comment faire de bonnes augmentations de données spectrales ?

Augmentation de données spectrales : existant

EMSA, Blazhko et al [5] :

- Utilisent le prétraitement EMSC pour identifier des effets de pentes et effets multiplicatifs.
- Modifient légèrement chaque effet, de façon indépendante.



EMSA sur un spectre de mangues [6]

Avantages :

- Estimation des perturbations à partir des données.
- Possibilité d'ajout de loadings de perturbation.

Limite :

- Hypothèse d'indépendance des effets.

Matrice de corrélation des coefficients EMSC, mangues [6]

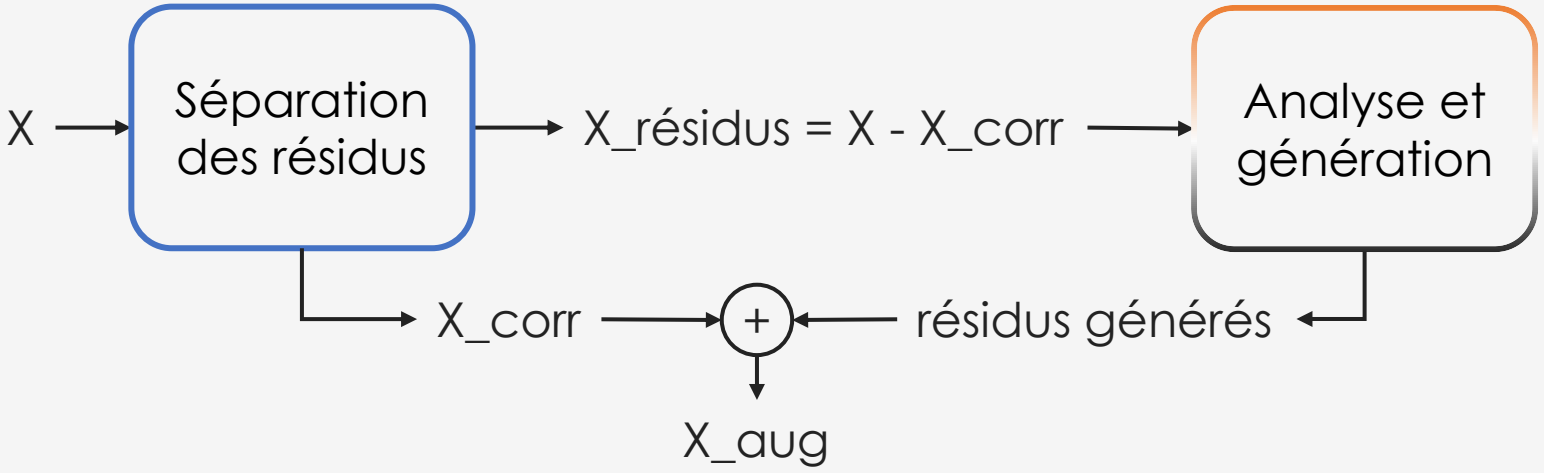
[5] U. Blazhko et al, « Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra », 2021, lien.
[6] P. Mishra et D. Passos, « A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit », 2021, doi: lien.

Proposition, hypothèse et postulat

Hypothèse :
Il existe une structure dans les effets qu'il faut respecter, sinon les augmentations dégradent les performances.

Postulat :
Les pré-traitements peuvent être utilisés pour créer des augmentations ➔ comparaison

Proposition, méthode



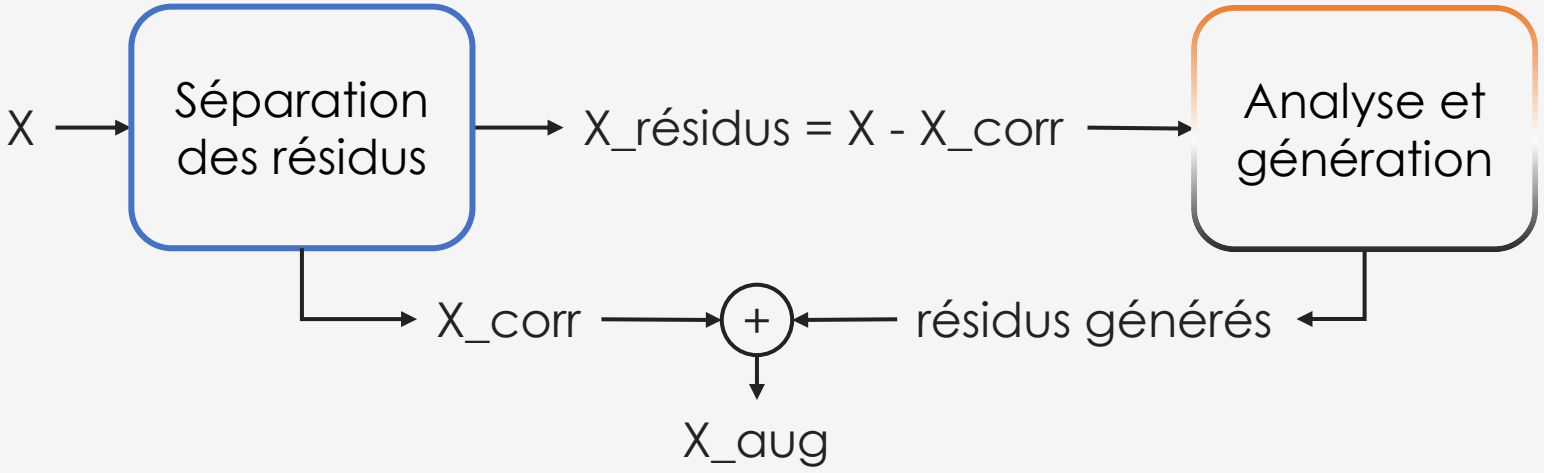
= Différents prétraitements : EMSC, SavGol et SNV

= Méthode d'analyse, ici ACP, $X\_résidus = TP'$

= Génération de nouveaux résidus :
pour chaque composante de l'ACP, $new\_T_{Ci} = old\_T_{Ci} + \epsilon$ avec $\epsilon \sim N(0, \frac{\sigma_{Ci}}{10})$

[5] U. Blazhko et al, « Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra », 2021, lien.

Proposition, méthode



<u>Évaluation des entraînements via la RMSEP :</u>

Entraînements avec…

- Données brutes
- Données prétraitées
- Données augmentées (méthode proposée et EMSA[5])

[5] U. Blazhko et al, « Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra », 2021, lien.

Augmentations spectrales : proposition, matériels

Données : [6]
- Mangues, ≈ 12000 spectres pour 4675 fruits
- Problème : régression de la teneur en matière sèche (%), comprise entre 9 et 25%
- Différents lieux de culture, différentes saisons : 3 saisons pour l'entraînement, 1 saison de test

Modèle : [7]
- 1 convolution avec 1 filtre et 4 couches denses
- Activation ELU

Entraînement :
- Loss RMSE : $\frac{1}{n}\sum_{i=1}^{n}(f(x_i) - y_i)^2$
- Un GPU A100

[6] P. Mishra et D. Passos, « A synergistic use of chemometrics and deep learning improved the predictive performance of near-infrared spectroscopy models for dry matter prediction in mango fruit », 2021, doi: lien.
[7] C. Cui and T. Fearn, « Modern practical convolutional neural networks for multivariate regression: Applications to NIR calibration », Chemometrics and Intelligent Laboratory Systems, 2018, doi: 10.1016/j.chemolab.2018.07.008

Proposition, résultats

Prétraitement	Correction / Augmentation	RMSEP
Données brutes		1,00
EMSC3	Correction	1,10
	Augmentation	**0,93**
	EMSA (σ/10) [5]	*1,00*
SavGolw17p2d2	Correction	1,00
	Augmentation	0,97
SNV	Correction	1,14
	Augmentation	*1,23*

RMSEP moyenne sur 10 entraînements.

→ Les prétraitements classiques (correction) ne semblent pas être bénéfiques pour le modèle.

* signifie que le résultat est significativement différent de celui obtenu pour les données brutes.

[5] U. Blazhko et al, « Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra », 2021, lien.

Proposition, résultats

Prétraitement	Correction / Augmentation	RMSEP
Données brutes		1,00
EMSC3	Correction	1,10*
	Augmentation	0,93
	EMSA (σ/10) [5]	*1,00*
SavGolw17p2d2	Correction	1,00
	Augmentation	0,97
SNV	Correction	1,14*
	Augmentation	*1,23*

RMSEP moyenne sur 10 entraînements.

→ Les prétraitements classiques (correction) ne semblent pas être bénéfiques pour le modèle.

* signifie que le résultat est significativement différent de celui obtenu pour les données brutes.

[5] U. Blazhko et al, « Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra », 2021, lien.

Proposition, résultats

Prétraitement	Correction / Augmentation	RMSEP
Données brutes		1,00
EMSC3	Correction	1,10
	Augmentation	**0,93**
	EMSA (σ/10) [5]	*1,00*
SavGolw17p2d2	Correction	1,00
	Augmentation	0,97
SNV	Correction	1,14
	Augmentation	*1,23*

RMSEP moyenne sur 10 entraînements.

→ Les augmentations avec le prétraitement approprié semblent pouvoir bénéficier au modèle.

* signifie que le résultat est significativement différent de celui obtenu pour les données brutes.

[5] U. Blazhko et al, « Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra », 2021, lien.

Proposition, résultats

Prétraitement	Correction / Augmentation	RMSEP
Données brutes		1,00
EMSC3	Correction	1,10
	Augmentation	**0,93***
	EMSA (σ/10) [5]	*1,00*
SavGolw17p2d2	Correction	1,00
	Augmentation	0,97
SNV	Correction	1,14
	Augmentation	*1,23**

RMSEP moyenne sur 10 entraînements.

→ Les augmentations avec le prétraitement approprié semblent pouvoir bénéficier au modèle.

* signifie que le résultat est significativement différent de celui obtenu pour les données brutes.

[5] U. Blazhko et al, « Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra », 2021, lien.

Proposition, résultats

Prétraitement	Correction / Augmentation	RMSEP
Données brutes		1,00
EMSC3	Correction	1,10
	Augmentation	**0,93**
	EMSA (σ/10) [5]	*1,00*
SavGolw17p2d2	Correction	1,00
	Augmentation	0,97
SNV	Correction	1,14
	Augmentation	*1,23*

RMSEP moyenne sur 10 entraînements.

→ L'hypothèse de structure semble vérifiée.

* signifie que le résultat est significativement différent de celui obtenu avec l'EMSA.

[5] U. Blazhko et al, « Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra », 2021, lien.

Proposition, résultats

Prétraitement	Correction / Augmentation	RMSEP
Données brutes		1,00
EMSC3	Correction	1,10
	Augmentation	**0,93***
	EMSA (σ/10) [5]	*1,00*
SavGolw17p2d2	Correction	1,00
	Augmentation	0,97
SNV	Correction	1,14
	Augmentation	*1,23*

RMSEP moyenne sur 10 entraînements.

→ L'hypothèse de structure semble vérifiée.

* signifie que le résultat est significativement différent de celui obtenu avec l'EMSA.

[5] U. Blazhko et al, « Comparison of augmentation and pre-processing for deep learning and chemometric classification of infrared spectra », 2021, lien.

Limites de la proposition et conclusions

Limites de la proposition :

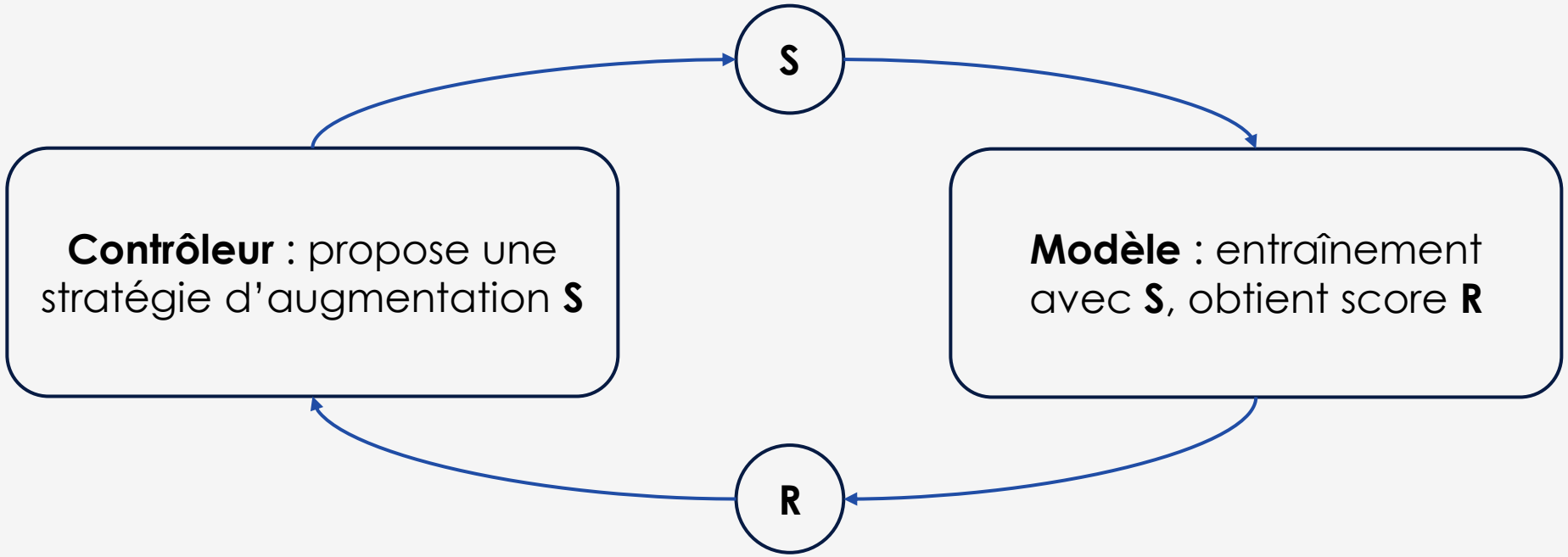
- On ne simule que des effets présents dans les données.
- Pour avoir une bonne estimation des effets, on a besoin de toutes les composantes et données.

Conclusions :

- La structure des effets doit être prise en compte pour l'augmentation de données.
- L'augmentation de données semble être un meilleur paradigme pour l'apprentissage profond.
- Les méthodes actuelles restent limitées.

Travaux futurs

- Travailler avec des données simulées → meilleure vue de ce qui peut être fait.
- Utiliser une stratégie d'apprentissage par renforcement :



Principe de AutoAugment [8].

[8] E. D. Cubuk et al, « AutoAugment: Learning Augmentation Policies from Data », 2019, *arXiv*: arXiv:1805.09501. doi: 10.48550/arXiv.1805.09501.