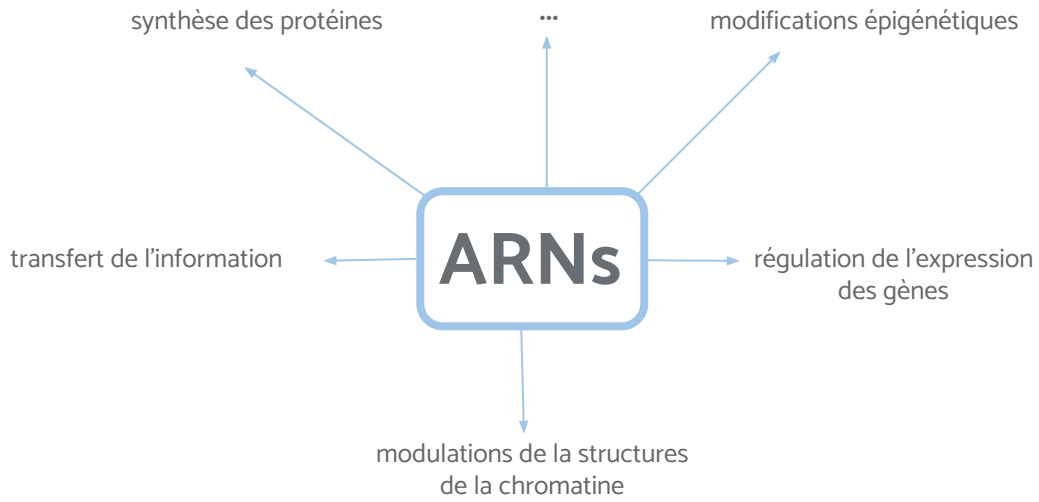
Stanford Ribonanza RNA Folding

Projet Kaggle : Création d'un modèle de prédiction de la structure de molécules d'ARN

UE Apprentissage, Intelligence artificielle et Optimisation 1 - M2BI

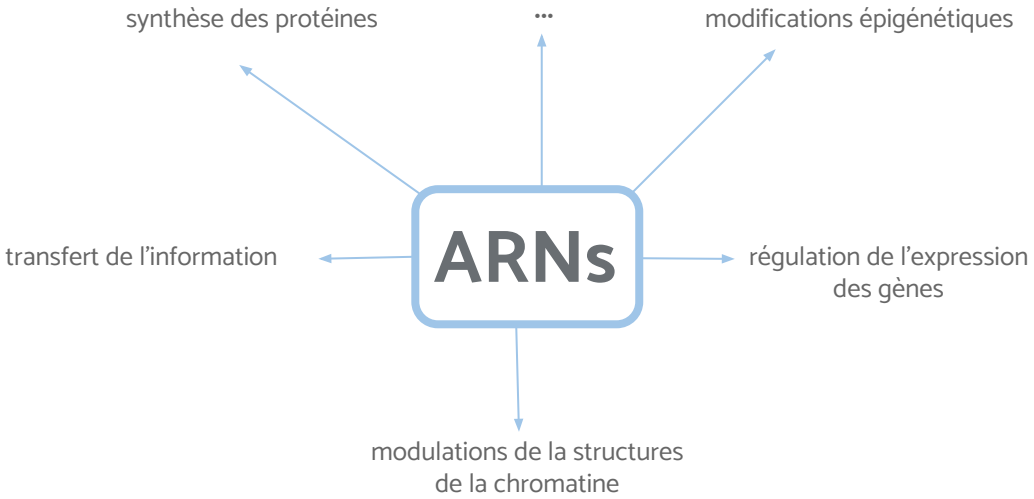
OUADAH Lilia - MERABET Anis - BAILLIF Marine - JEAN-MARIE ROUDE

Introduction

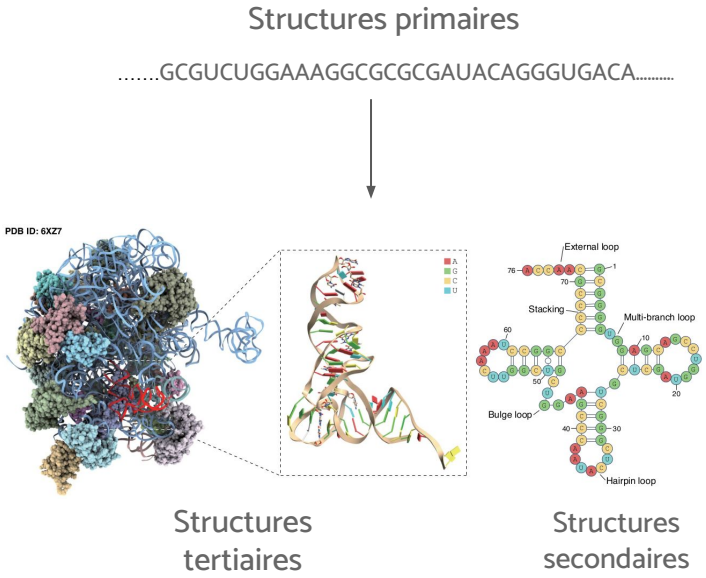


fonctions des ARNs Dogini et al. 2014

Introduction

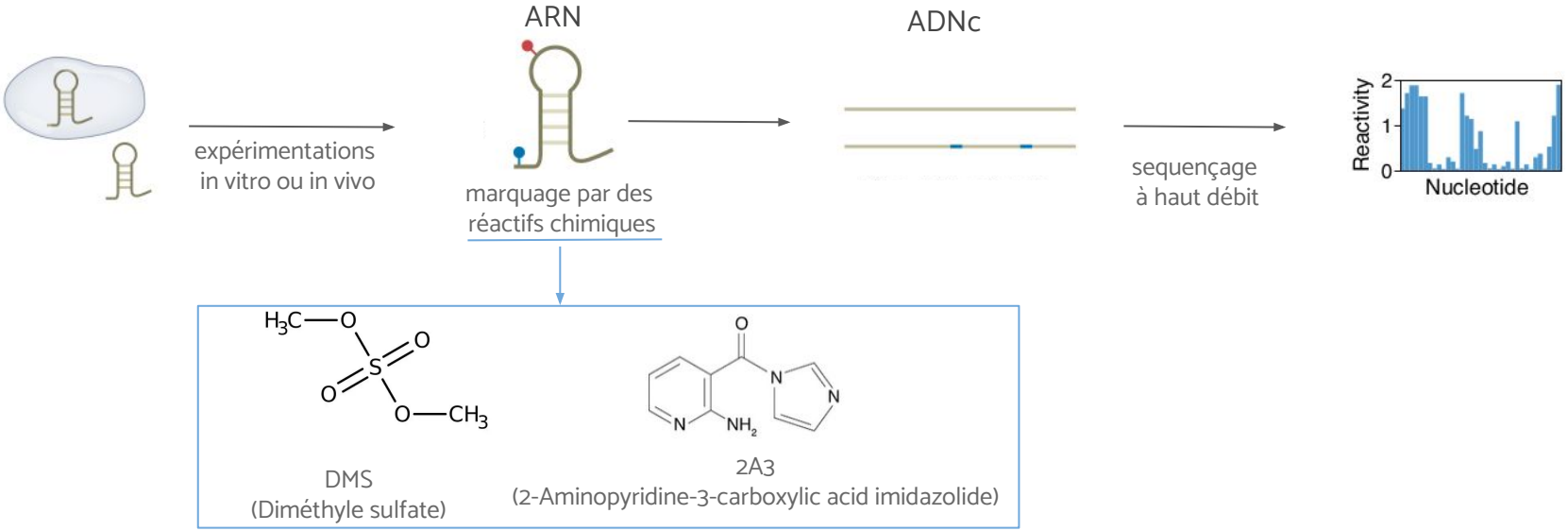


fonctions des ARNs Dogini et al. 2014

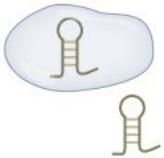


Structures de l'ARN Yesselman et al. 2019

Introduction



Introduction



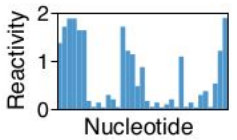
expérimentations
in vitro ou in vivo

ARN

marquage par des
réactifs chimiques

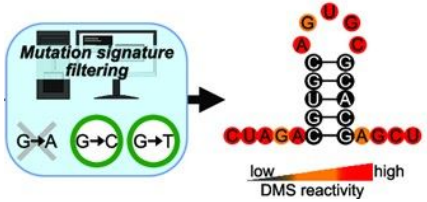
ADNc

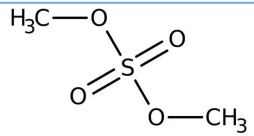
séquençage
à haut débit



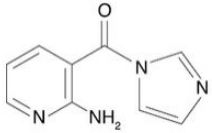
algorithmes

prédiction du repliement





DMS
(Diméthyle sulfate)



2A3
(2-Aminopyridine-3-carboxylic acid imidazolide)

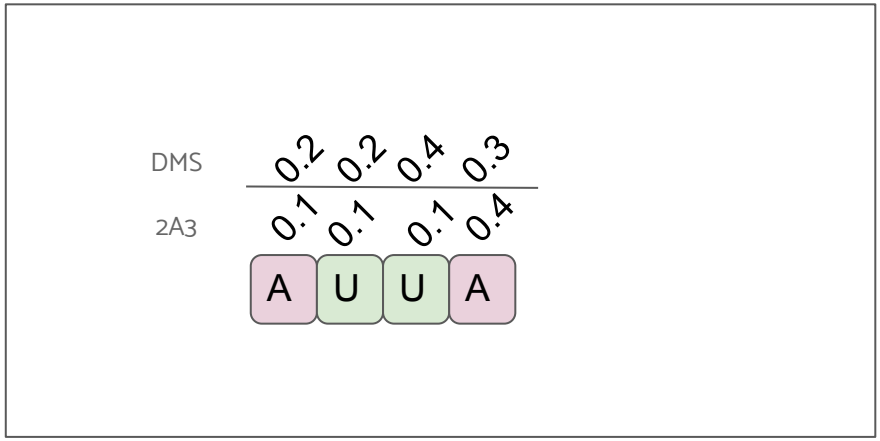
Problématique

Peut on prédire computationnellement le profil de réactivité d'une séquence d'ARN ?

Projet Kaggle

Objectif

Prédire pour une séquence d'ARN la réactivité de chacun de ces nucléotides pour les expériences DMS et 2A3



Projet Kaggle

Présentation des données

1643800 observations
806578 séquences uniques

Sequence	Dataset name	Experiment type	Signal to noise	SN Filter	Reactivity 1	…	Reactivity n	Reactivity error 1	…	Reactivity error n
AU…A	…	2A3/DMS	0.02	0/1	0.012	…	0.003	0.02	…	0.001

Preprocessing

Filtrage

1643680 séquences

SN FILTER = 1

437919 observations
242729 séquences uniques

Retrait des séquences avec des valeurs pour un seul type d'expérience

410708 observations
242729 séquences uniques

Retrait des séquences avec des duplicatas de valeurs pour un seul type d'expérience

197258 observations pour 2A3
213450 observations pour DMS

JEU DE DONNÉES 2

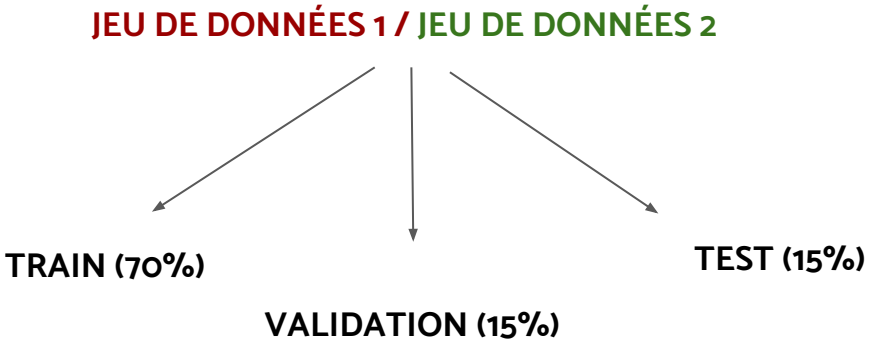
Retrait des séquences avec des duplicatas de valeurs pour un seul type d'expérience

335958 observations
167979 séquences uniques

JEU DE DONNÉES 1

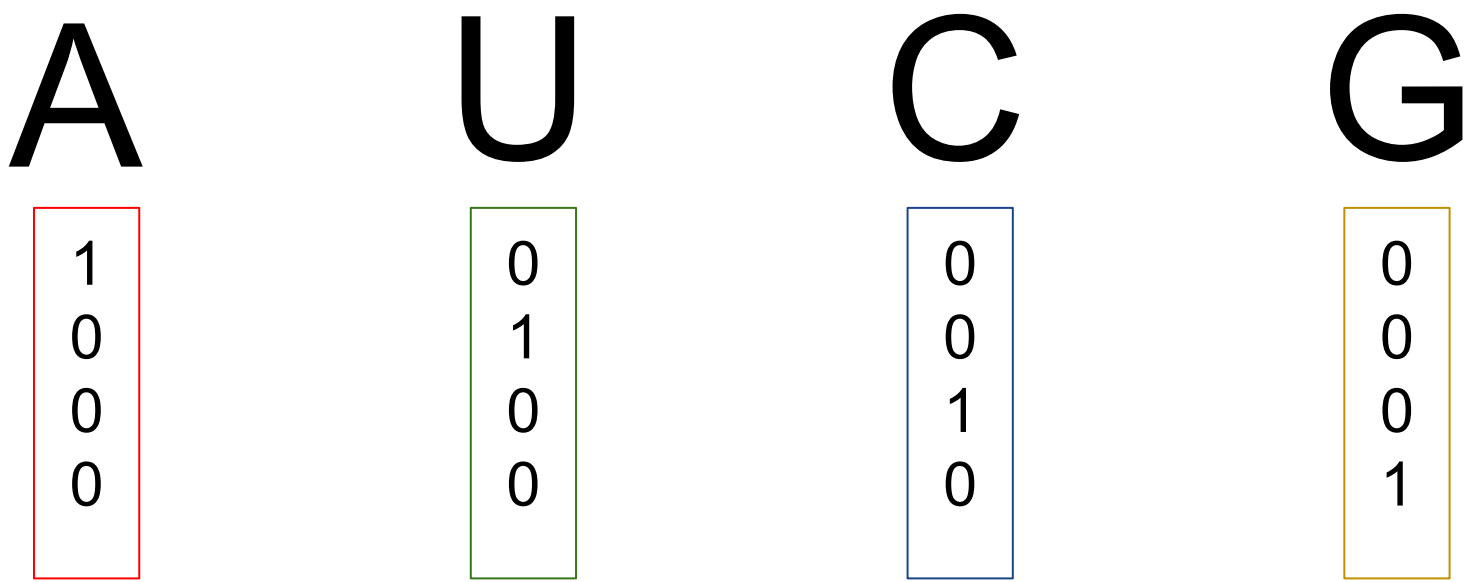
Preprocessing

Filtrage



Preprocessing

Encoding

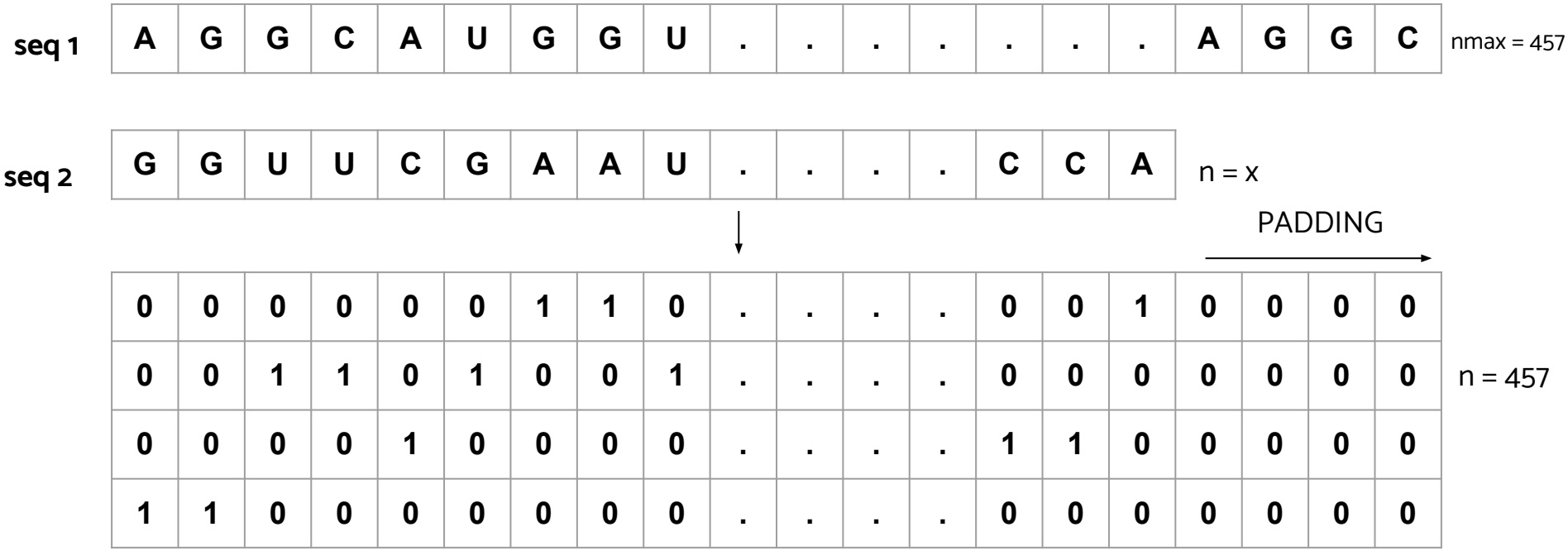


Preprocessing

Encoding

Preprocessing

Padding



Preprocessing

Padding

Preprocessing

Padding

Encoded sequence

0	0	0	0	0	0	1	1	0	.	.	.	.	0	0	1	0	0	0	0
0	0	1	1	0	1	0	0	1	.	.	.	.	0	0	0	0	0	0	0
0	0	0	0	1	0	0	0	0	.	.	.	.	1	1	0	0	0	0	0
1	1	0	0	0	0	0	0	0	.	.	.	.	0	0	0	0	0	0	0

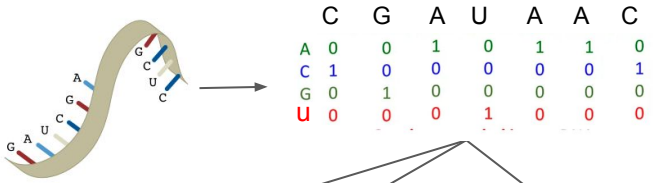
Reactivities

0.25	0.11	NA	NA	0.3	0.41	0.29	NA	0.31	.	.	.	.	0.2	0.1	NA	0	0	0	0

Masking

0.25	0.11			0.3	0.41	0.29		0.31	.	.	.	.	0.2	0.1					

Modèles simples :



	C	G	A	U	A	A	C
A	0	0	1	0	1	1	0
C	1	0	0	0	0	0	1
G	0	1	0	0	0	0	0
U	0	0	0	1	0	0	0

simple LSTM

Input	shape (457,4)
LSTM	noeuds = 32
Dense	noeuds = 2

valeurs de réactivités
DMS et 2A3
pour chaque nucléotides

bi - LSTM

Input	shape (457,4)	
Bi-LSTM	noeuds = 32	
LSTM	noeuds = 32	
Dropout	0.1	
Dense	ReLU	noeuds =16
Dense	ReLU	noeuds = 8
Dense	noeuds = 2	

valeurs de réactivités
DMS et 2A3
pour chaque nucléotides

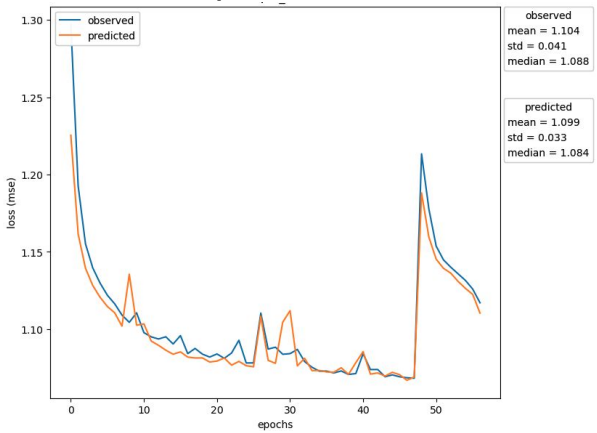
LSTM + Conv

Input	shape (457,4)
Bi-LSTM	noeuds = 32
LSTM	noeuds = 32
Dropout	0.1
ConV1D	filtres = 64, kernel size = 3, dilatation rate = 1 , ReLU
	filtres = 128, kernel size = 3, dilatation rate = 2 , ReLU
	filtres = 256, kernel size = 3 dilatation rate = 4 , ReLU
Dense ReLU	noeuds =16
Dense ReLU	noeuds = 8
Dense	noeuds = 2

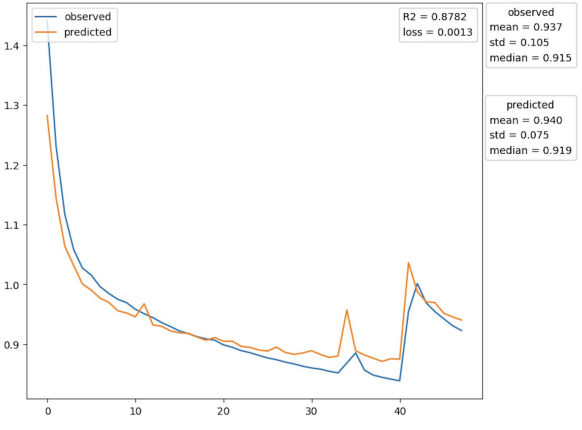
valeurs de réactivités
DMS et 2A3
pour chaque nucléotides

Résultats et Discussions :

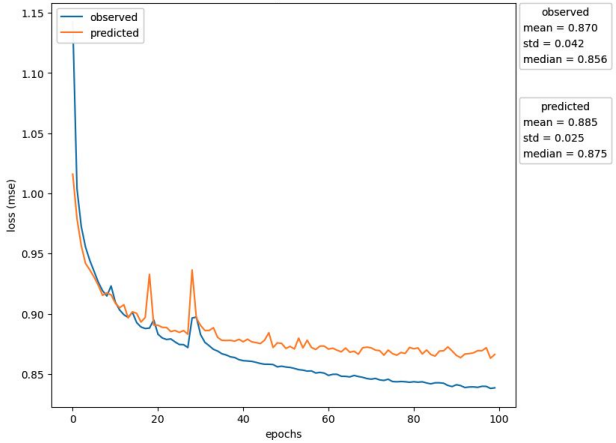
simple LSTM



Bi - LSTM

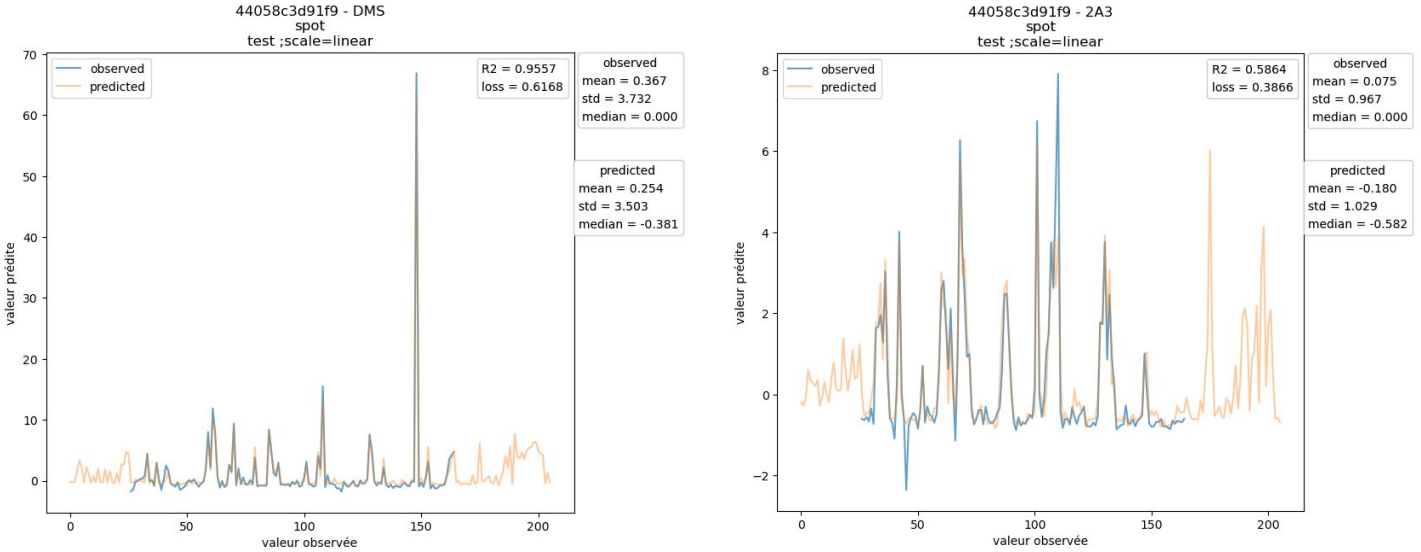


LSTM + ConV



Courbes Loss des modèles simples

Résultats et Discussions :

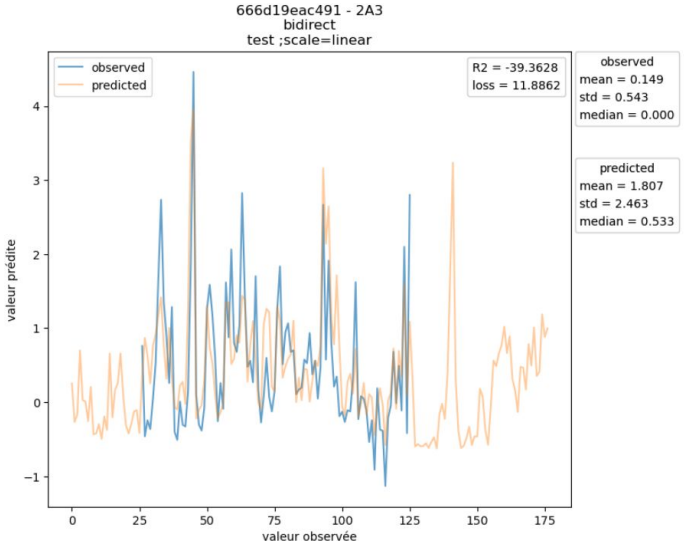


Comparaison des valeurs observées et prédites de la réactivité par le modèle LSTM Convolutionnel pour une séquence (id: 44058c3d91f9)

meilleures prédictions des modèles simples pour le jeu test

Résultats et Discussions :

Bi - LSTM



Comparaison des valeurs observées et prédites de la réactivité par le modèle Bi- LSTM pour une séquence (id: 666d19eac491)

pires prédictions des modèles simples pour le jeu test

Prédiction du jeu de données test

UMAP : nombre local de voisins = 100 , distance minimal entre les voisins = 0.4

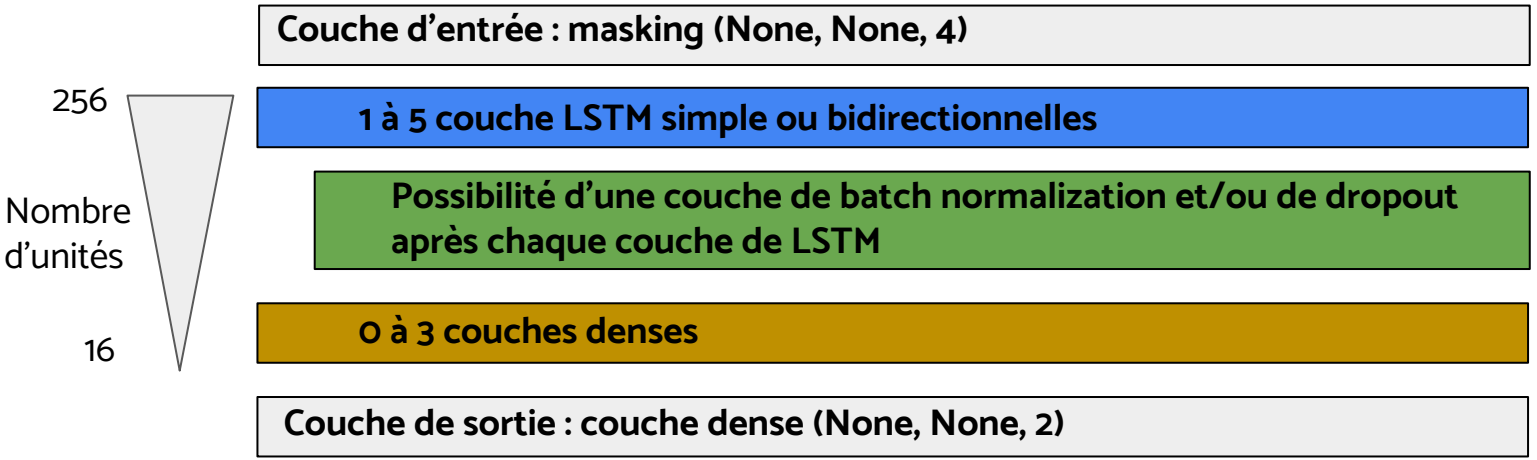
Modèles	$R^2 > 0.5$
Simple LSTM 2A3	3

Modèles	$R^2 \geq 0.5$
Simple LSTM 2A3	3
Simple LSTM DMS	42
Bi -LSTM 2A3	19
Bi -LSTM DMS	610
LSTM + Conv 2A3	385
LSTM + Conv DMS	1312

Nombre de prédictions associé à un R ≥ 0.5 pour chaque modèles

Optimisation des hyperparamètres du modèle

Approche bayésienne visant réduire le nombre d'essais nécessaires pour identifier les hyperparamètres optimaux.



Paramètres fixes : taille des batchs : 64. Nombre d'epochs : 10. Optimiseur : Adam. fonction d'activation : "relu" pour les couches dense et "linear" pour la couche de sortie.

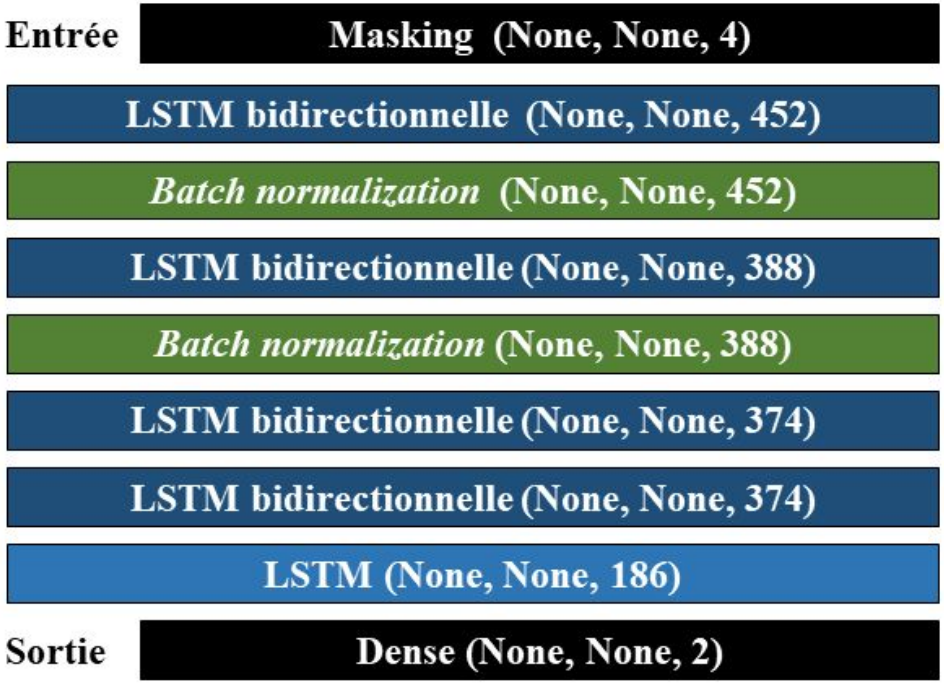
Optimisation des hyperparamètres du modèle

Un total de 35 itérations.

Le meilleure modèle a été enregistré à l'**itération 20**

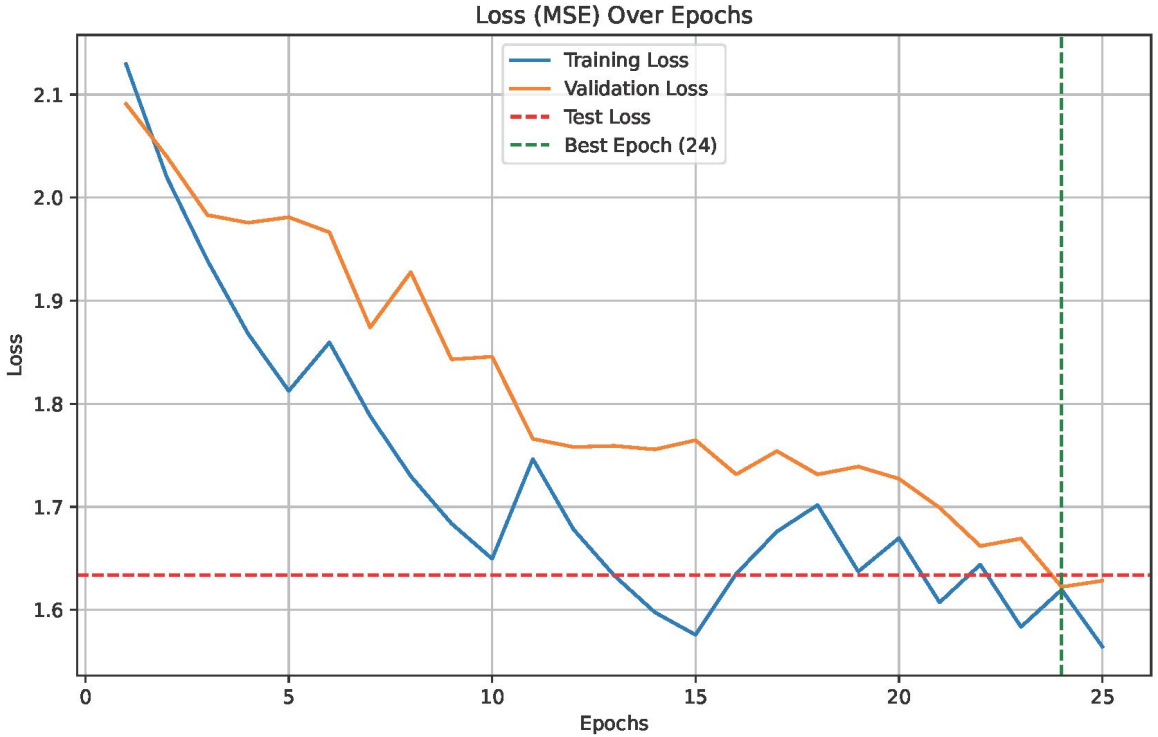
Optimisation des hyperparamètres du modèle

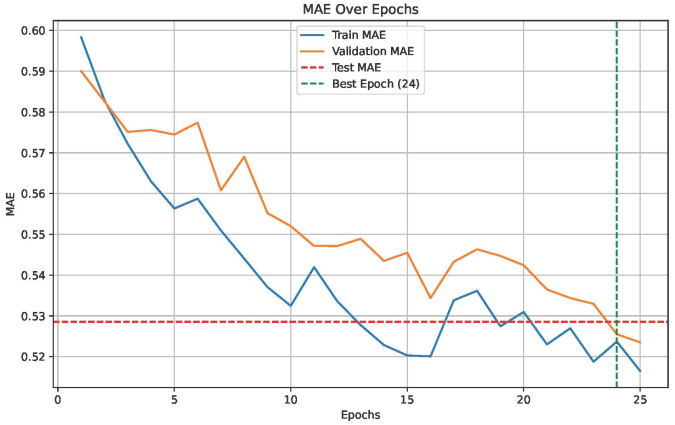
Approche bayésienne visant réduire le nombre d'essais nécessaires pour identifier les hyperparamètres optimaux.

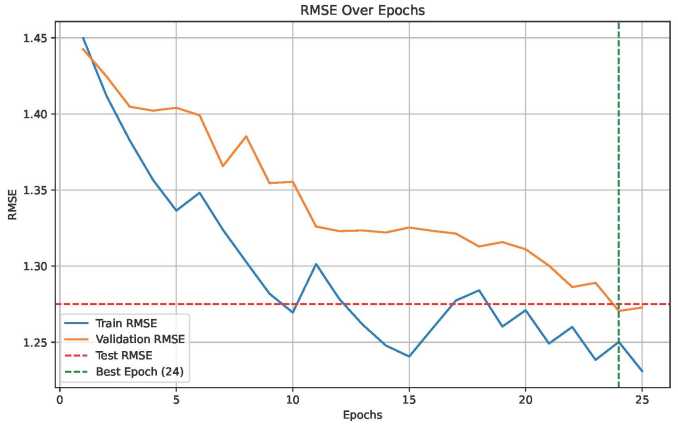


Nombre de paramètres : **3 537 886**

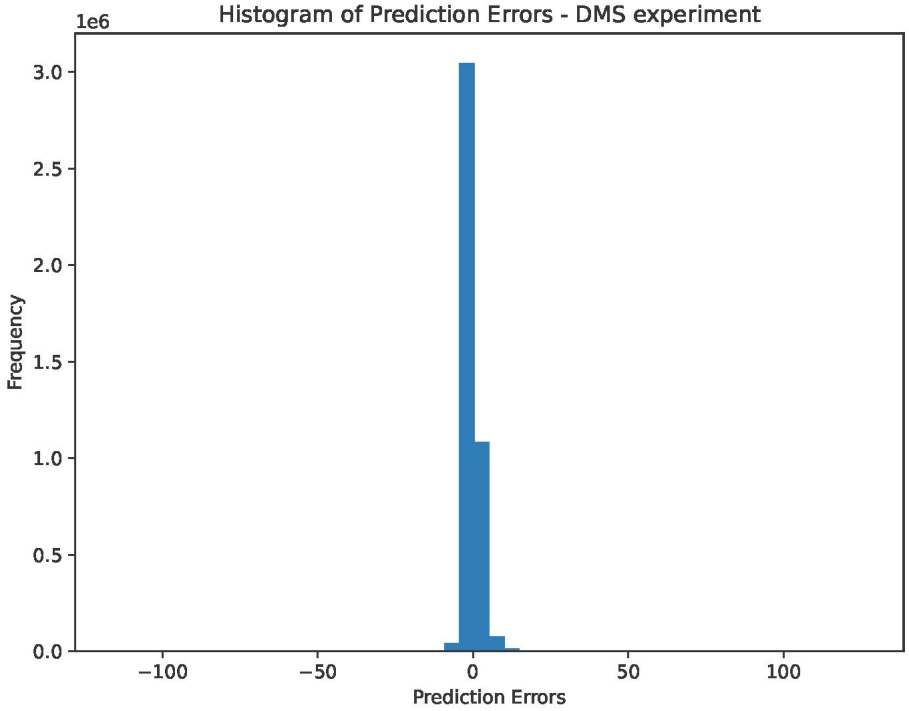
Entrainement du meilleur modèle

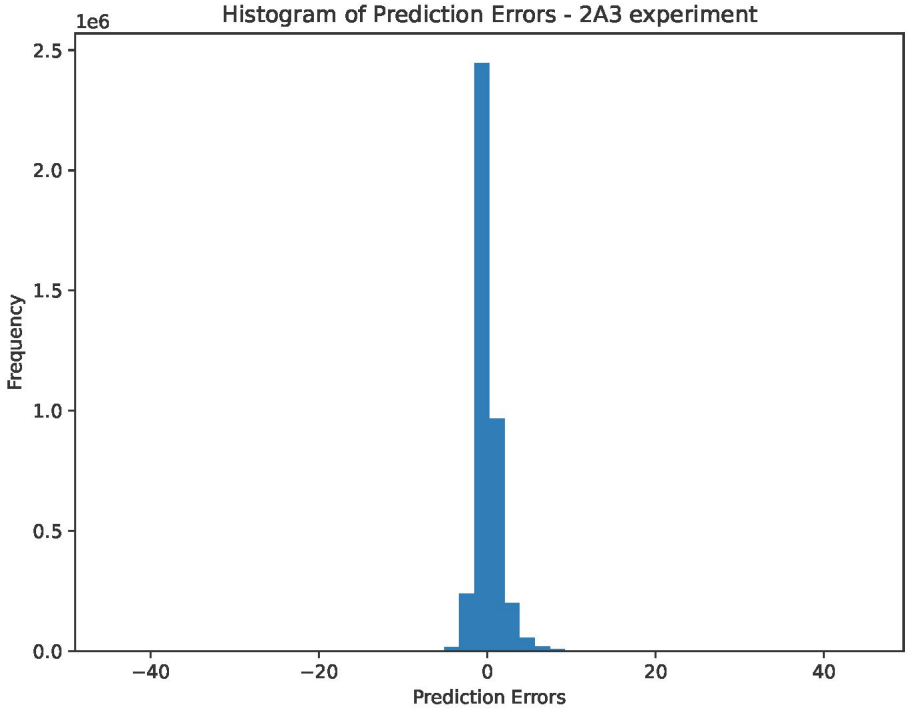






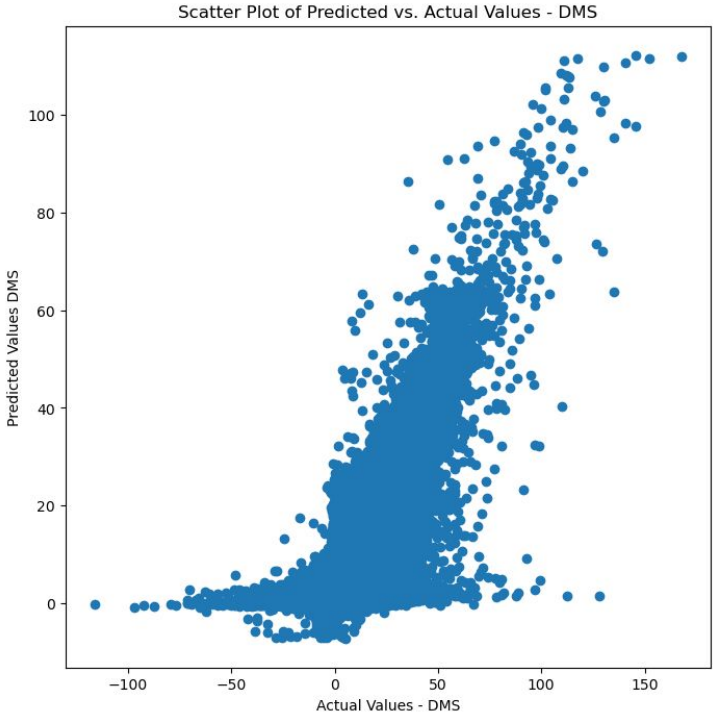
Erreurs de prédiction du meilleur modèle

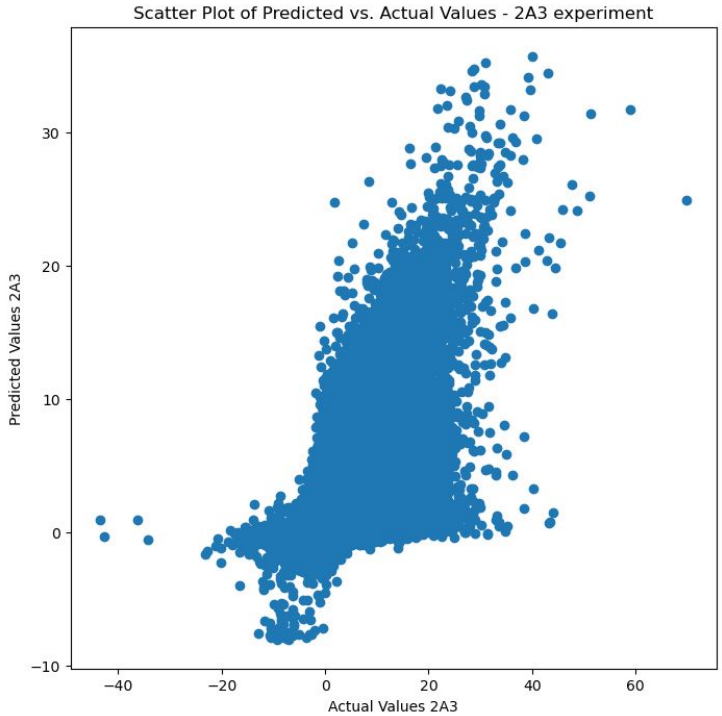




Les valeurs de réactivité dans l'expérience de **DMS** sont mieux prédites que celles dans l'expérience de **2A3**

Homogénéité des prédictions des valeurs de réactivité

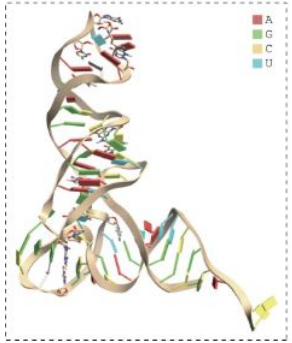




Le modèle arrive à prédire aussi bien les valeurs proches de zéro que les valeurs extrêmes

Conclusion :

- Performances différentielles entre les différents modèles
- Meilleures prédictions des réactivités pour le DMS
- Bonne généralisation aux données du test pour le modèle obtenu par optimisation des hyperparamètres
- Bonne capacité de prédiction des valeurs extrêmes



Notre modèle optimisé pourrait être un bon modèle de prédiction des structures d'ARN

Perspectives :

- Modèle simples : modèle combinant couches de convolution et couches LSTM à explorer davantage *(cf SPOTRNA1, SPOTRNA2)*
- Utiliser des modèles basés sur les transformers afin de concevoir un embedding qui sera utilisé pour mieux prédire les valeurs de réactivités
- Utiliser des données redondantes en appliquant un poids en fonction de la redondance des séquences
- Appliquer un poids selon la valeur de signal_to_noise

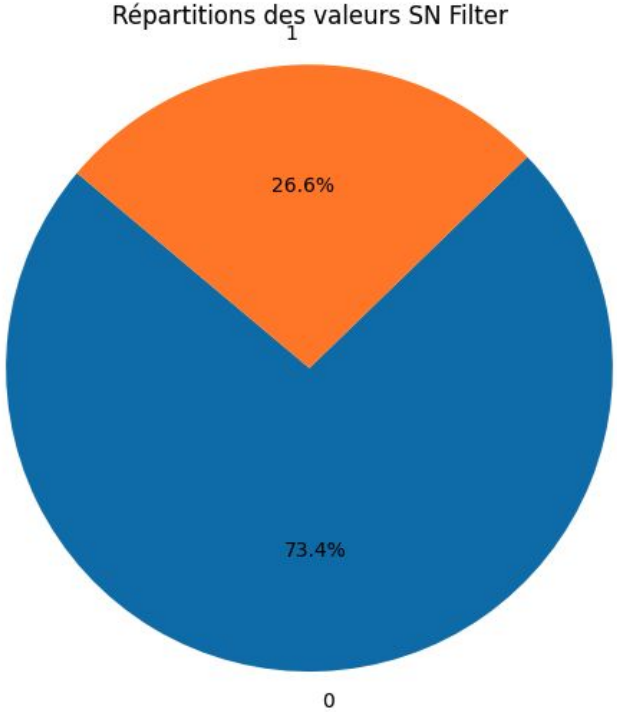
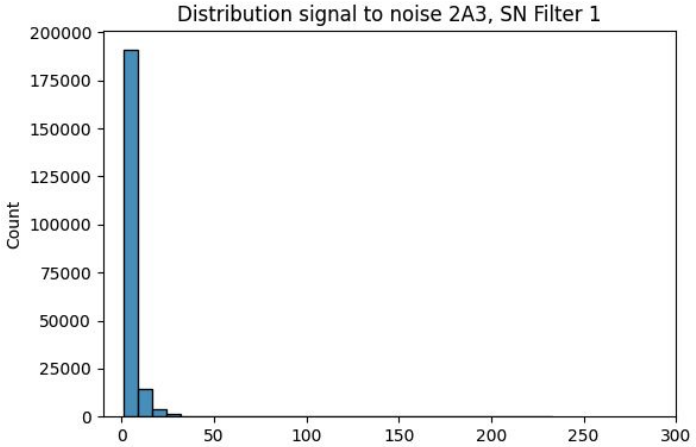
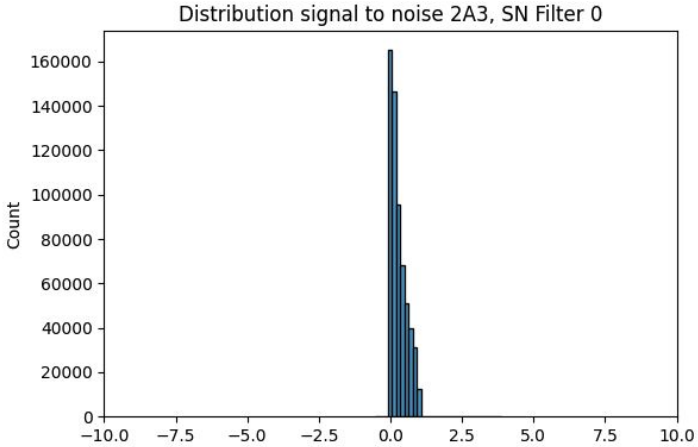
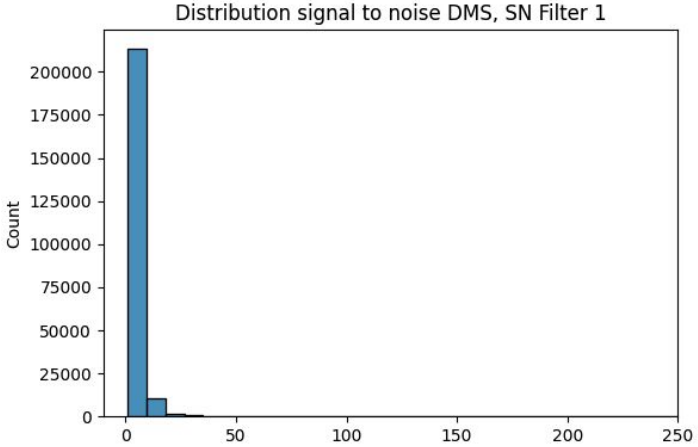


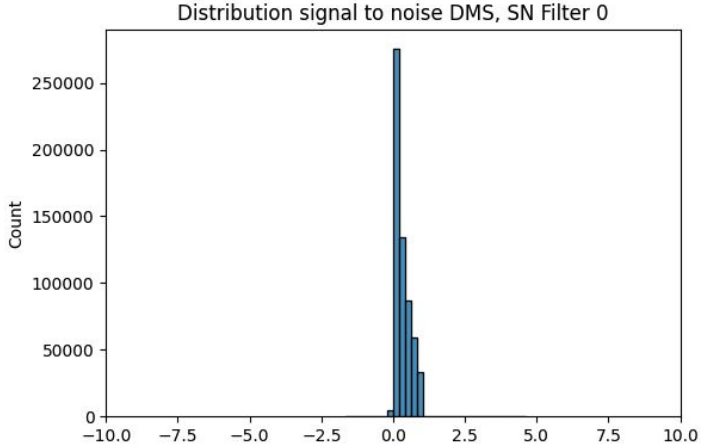
Figure 1 : Répartition des valeurs de SN Filter (0 ou 1) dans le jeu de données.

Annexes

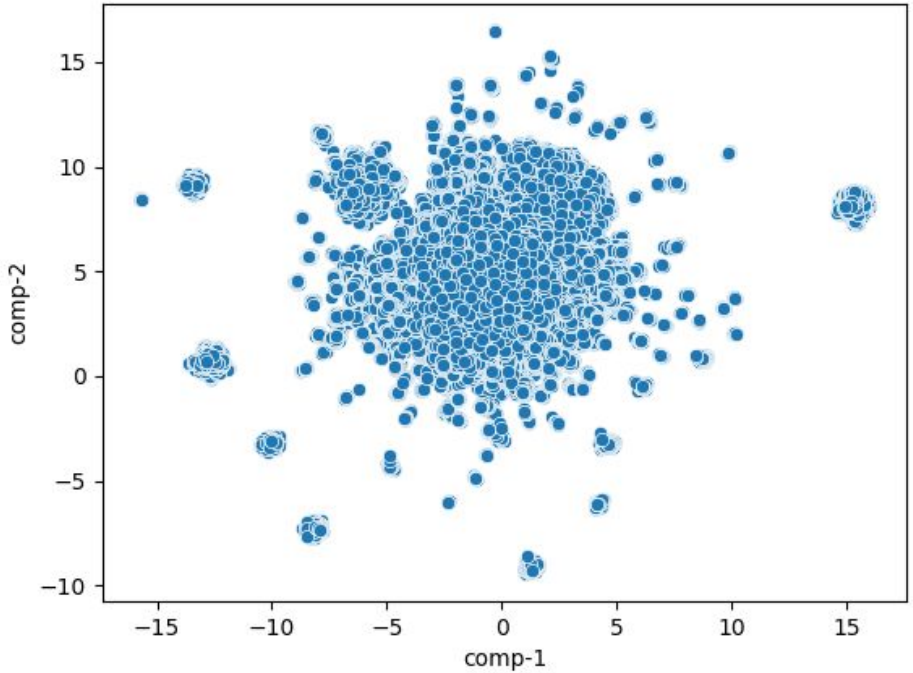

Distribution signal to noise 2A3, SN Filter 1
Count
200000
175000
150000
125000
100000
75000
50000
25000
0
0
50
100
150
200
250
300


Distribution signal to noise 2A3, SN Filter 0
Count
160000
140000
120000
100000
80000
60000
40000
20000
0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

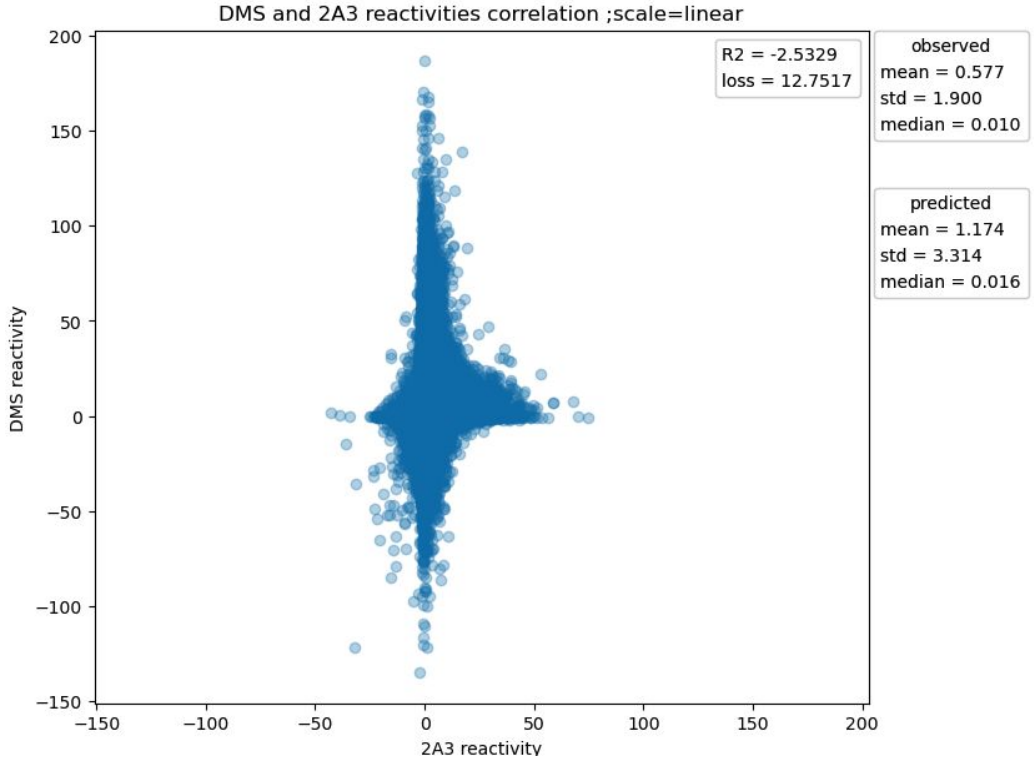

Distribution signal to noise DMS, SN Filter 1
Count
200000
175000
150000
125000
100000
75000
50000
25000
0
0
50
100
150
200
250


Distribution signal to noise DMS, SN Filter 0
Count
250000
200000
150000
100000
50000
0
−10.0
−7.5
−5.0
−2.5
0.0
2.5
5.0
7.5
10.0

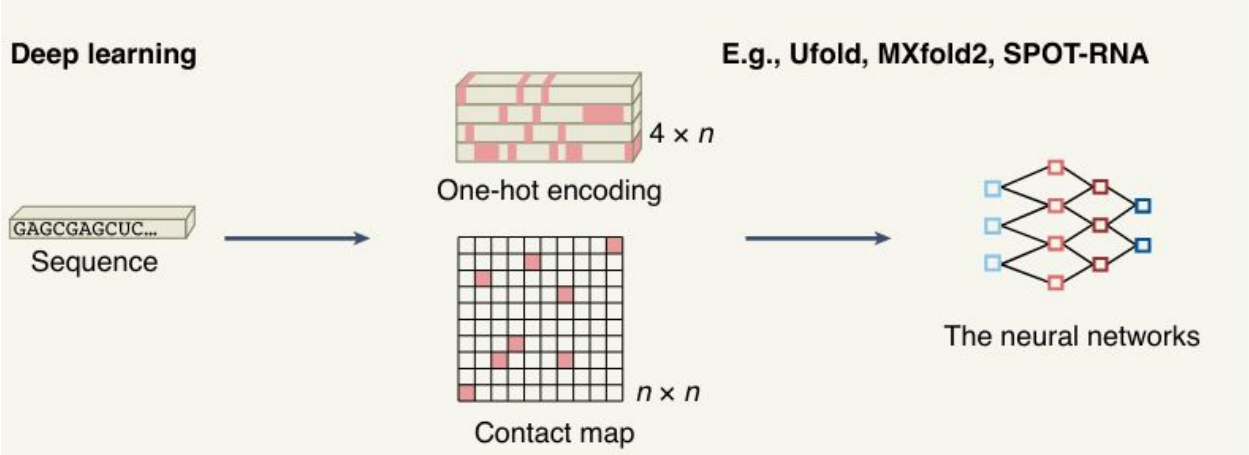
Annexes



Annexes


DMS and 2A3 reactivities correlation ;scale=linear
R2 = -2.5329
loss = 12.7517
observed
mean = 0.577
std = 1.900
median = 0.010
predicted
mean = 1.174
std = 3.314
median = 0.016
DMS reactivity
2A3 reactivity
200
150
100
50
0
−50
−100
−150
−150
−100
−50
0
50
100
150
200

Annexes



approches dans la littérature