

UE 2: Apprentissage, Intelligence Artificielle et Optimisation

# Prédiction de la structure secondaire des ARNm

AGSOUS Salim, FAUCHOIS Antoine, HOLO Donovan  
et YOUJIL ABADI Souad

Master 2 Biologie-Informatique

Vendredi 27 octobre 2023

# Sommaire

## 1. Introduction

- Vaccin contre la COVID-19, une approche innovante qui ouvre la voie vers une nouvelle génération de thérapie
- Biochimie structurale des ARN, quelques rappels
- Détermination expérimentale des structures secondaires des ARN

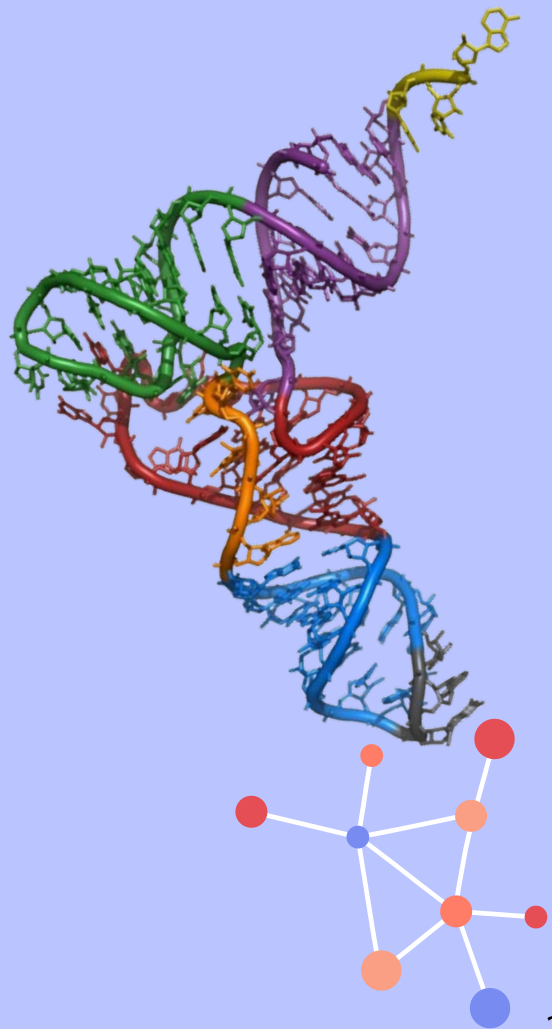
## 2. Matériels et méthodes

- Présentation des données
- Méthodologie d'analyses exploratoires
- Les séquences d'ARN sont-elles suffisamment informatives
- Méthodologie du features engineering
- Structure des réseaux élaborés

## 4. Résultats et discussions

- Résultats des analyses exploratoires
- Performances des modèles sur un sous échantillon
- Apprentissage partiel sur le jeu complet

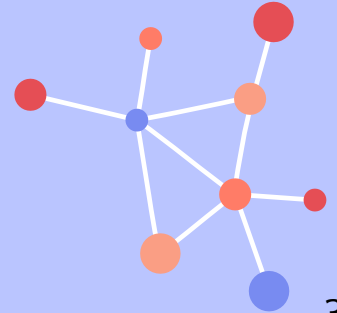
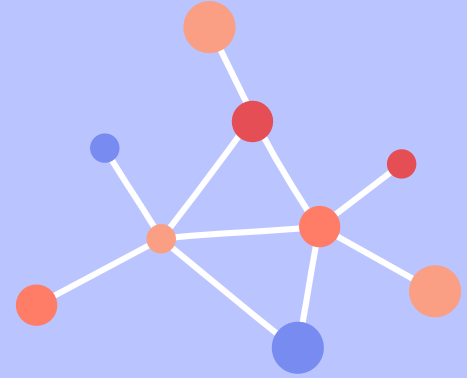
## 5. Conclusion





01

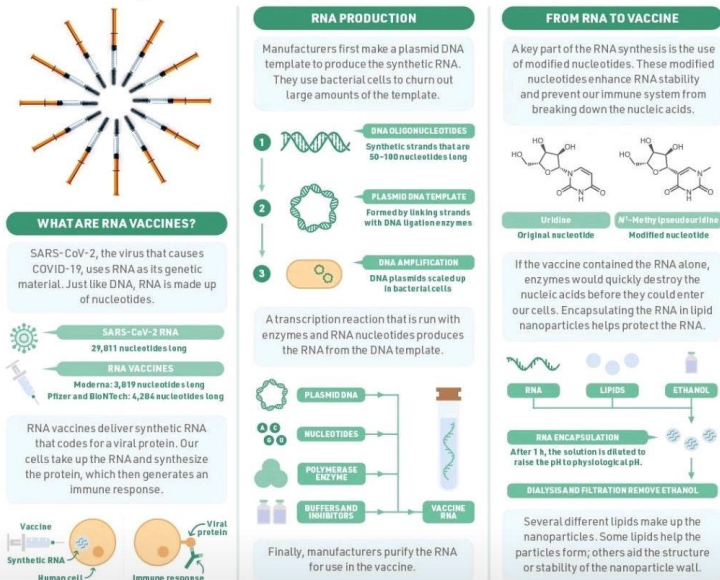
# Introduction





## HOW ARE RNA VACCINES MADE?

RNA vaccines produced by Pfizer and BioNTech and Moderna have become the first COVID-19 vaccines approved for emergency use in the US. How are these vaccines made?



PERIODIC  
GRAPHICS

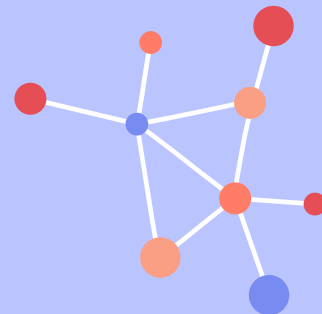


© C&EN 2021 Created by Andy Brunning for Chemical & Engineering News

- Concept pensé par Katalin Karikó (prix Nobel de la médecine 2023)
- Injection d'un ARNm codant pour la synthèse d'un antigène spécifique d'un agent pathogène
- Induit ensuite une réponse immunitaire adaptative
- Vaccin COVID-19 (Pfizer/BioNtech et Moderna): 96 millions de doses injectées en France (*source: covidtracker.fr*)
- De nouveaux projets de vaccins à ARNm: Zika, Cancer ou encore le VIH !

Une technologie innovante, avec des inconvénients:

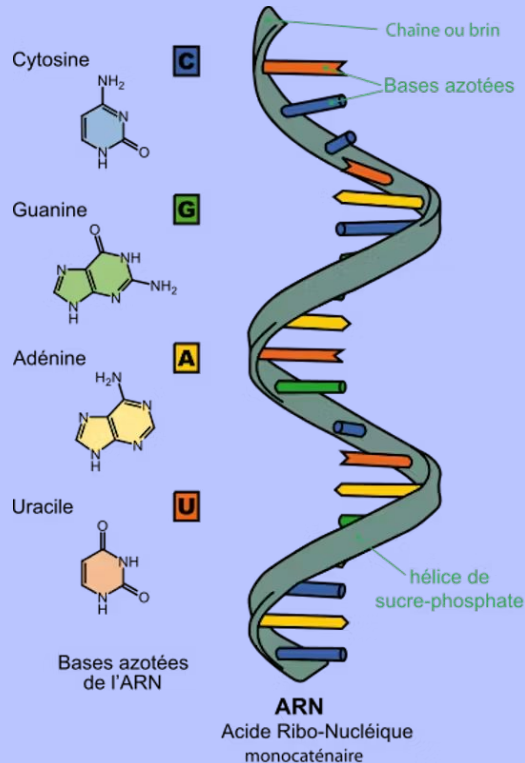
- Encapsulation lipidique nécessaire
- Stabilité des ARN



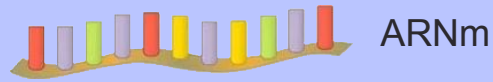
## 1.2 Biochimie structurale des ARN, quelques rappels

### Composition des ARN

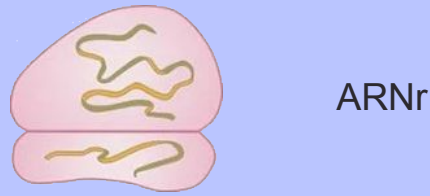
ARN = **A**cide **R**ibo**N**ucléique



### Différents types d'ARN



- Transcription des ADN codants
- Traduits en protéine



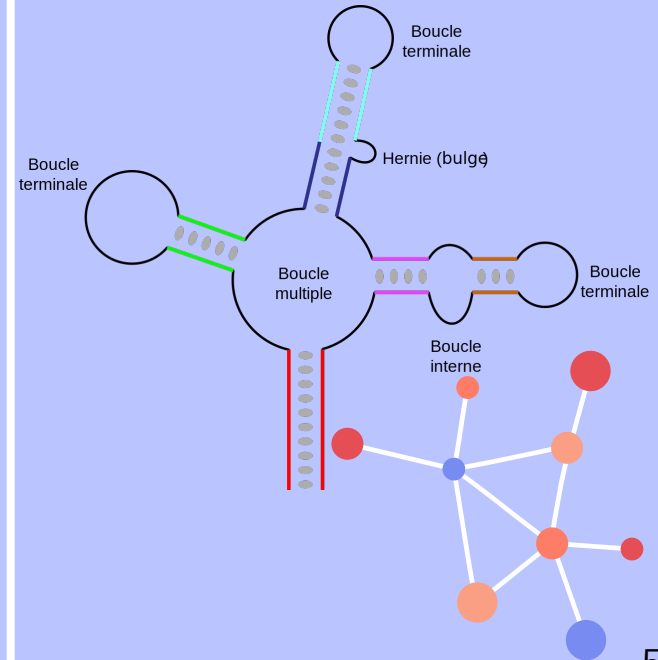
- Eléments structurants des ribosomes



- Comportent un anti-codon avec un acide aminés spécifiques

### Éléments de structure secondaire

- Essentiellement dirigée par des motifs répétés et inversés



## 1.3

## Détermination expérimentale des structures secondaires des ARN

## SHAPE seq

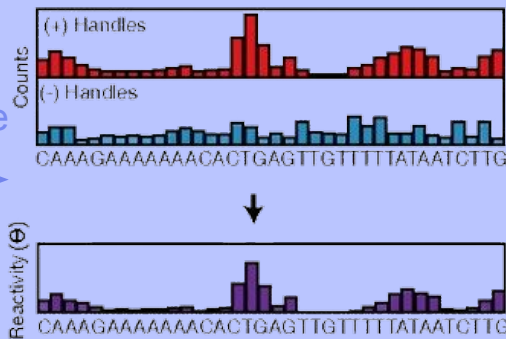
Contrôle positive

Shape / DMS

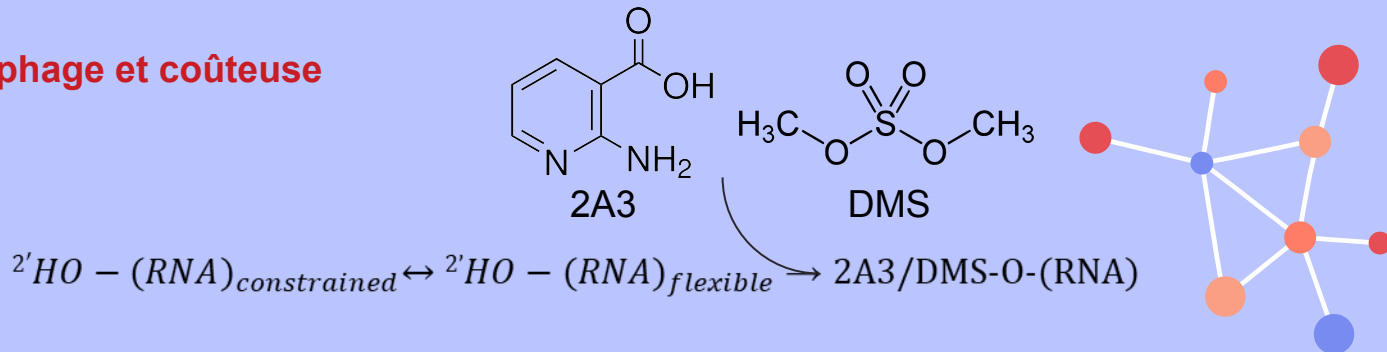
Reverse  
transcription

Séquençage

Contrôle négative



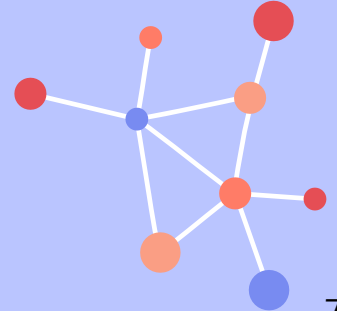
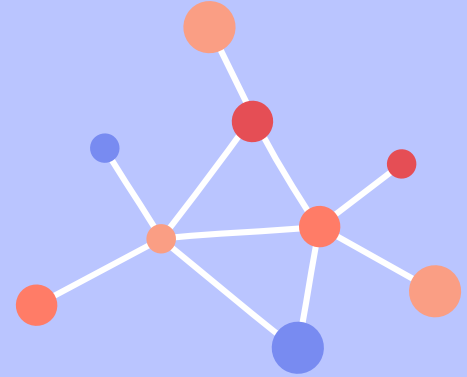
□ Une méthode chronophage et coûteuse





02

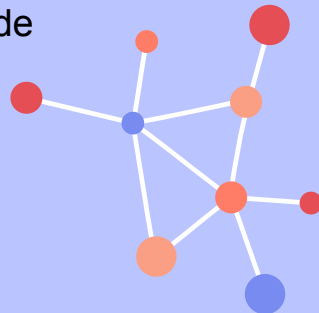
# Matériels et Méthodes



# kaggle Stanford Ribonanza RNA Folding

Train\_data.csv (2,37 Gb)

- **Sequence\_id**: ID des séquences d'ARN
- **Sequence**: séquence d'ARN
- **Experiment\_type**: réactif de sondage utilisé (DMS ou 2A3)
- **Dataset\_name**: nom de la base de données
- **Reads**: nombre de reads séquencés
- **Signal\_to\_noise**: rapport signal sur bruit de fond
- **SN\_filter**: booléen sur la qualité du séquençage (reads et rapport signal sur bruit de fond)
- **Reactivity\_0001 à reactivity\_0256**: réactivités pour les positions n°1 à n°256 (longueur maximale)
- **Reactivity\_error\_0001 à reactivity\_error\_0256**: erreurs des réactivités pour les positions n°1 à n°256





## 2.2 Méthodologie d'analyses exploratoires des données

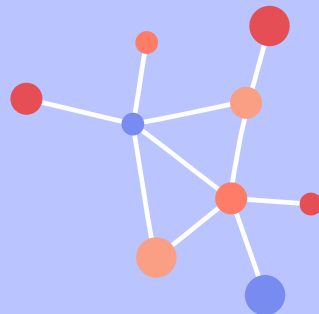
### ➤ Analyses statistiques

- paramètres de tendances centrales et de dispersion sur les réactivités,  $\frac{signal}{bruit\ de\ fond}$  et nombre de reads
- tableau de contingence des variables SN filter et experiment type
- Comptages des séquences par base de données

### ➤ Recherche des valeurs manquantes

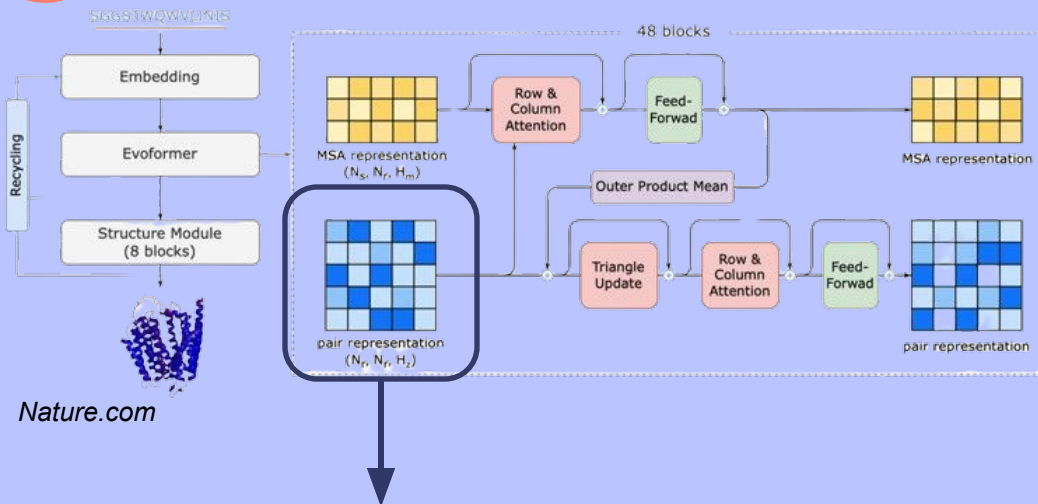
### ➤ Visualisation des séquences

- tSNE appliqué sur les distances entre les séquences réencodés (analyse par base de données)

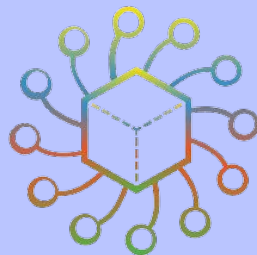
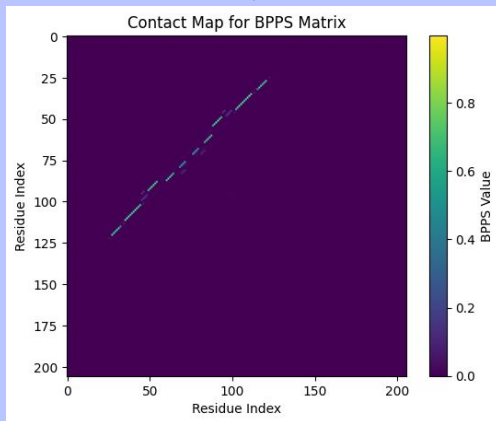


## 2.3

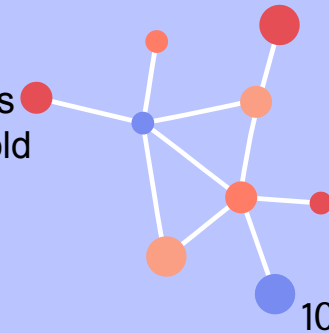
## Les séquences d'ARN sont-elles suffisamment informatives ?



Nature.com



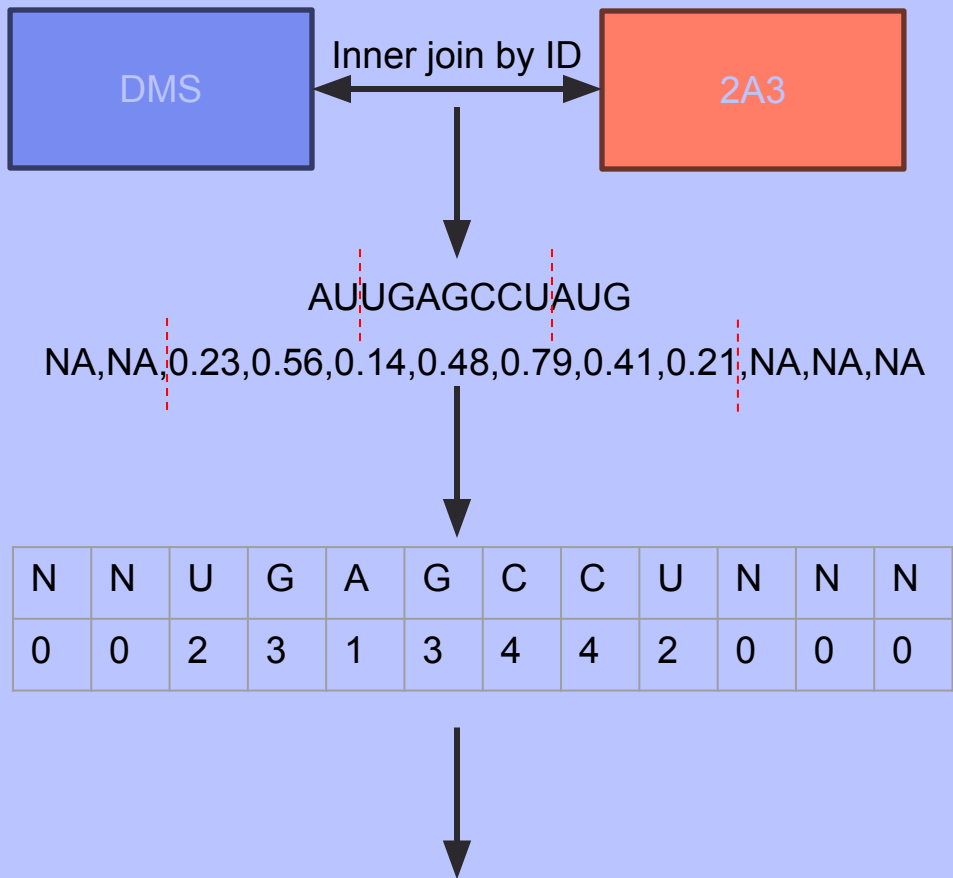
□ Probabilités d'appariement des bases estimées avec EternaFold utilisable via la librairie Arnie



## 2.4

## Méthodologie du features engineering

## Analyse par database

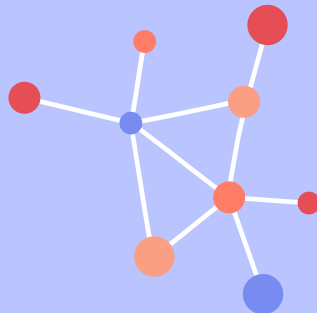


1) Jointure complète pour conserver les séquences avec des réactivités en DMS et 2A3

2) Trimming des séquences en fonction des réactivités

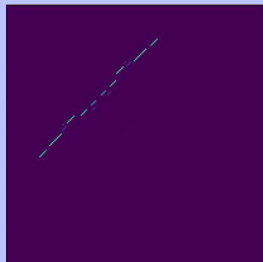
3) Sequence encoding and padding

{N:0, A:1, U:2, G:3, C:4}



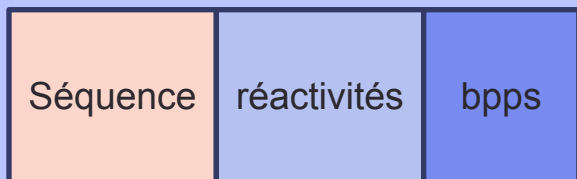
## 2.4

## Méthodologie du features engineering



0	0,53	0,14
0,08	0	0,17
0,02	0,49	0

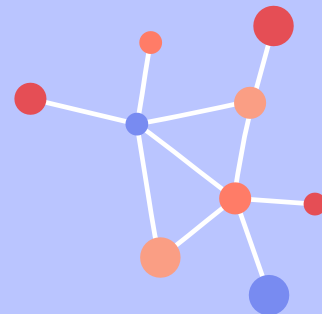
0	0,53	0,14	0,08	0	0,17	0,02	0,49	0
---	------	------	------	---	------	------	------	---



4) Calcule des cartes de contact ou matrice des probabilités d'appariement (Eternafold)

5) Flat des cartes de contacts

6) Assemblage du dataset final



## 2.5

## Structure des réseaux élaborés

## a) Vue d'ensemble des structures de réseaux envisagées

- Il existe un lien de dépendance des réactivités entre les bases
- L'utilisation d'un **Recurrent Neural Network** (RNN) est privilégiée

## Réseaux se basant sur les séquences d'ARN

- RNN à LSTM
- RNN à GRU

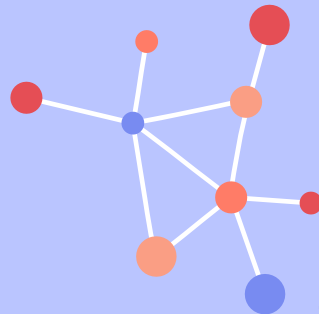
## Réseaux se basant sur les séquences et les cartes de contact

- RNN à LSTMs combiné à un CNN
- RNN à GRUs combiné à un CNN

## Réseaux inspirés du concours OpenVaccine

- RNN à LSTM combiné à un CNN avec un multihead attention
- RNN à GRU combiné à un CNN avec un multihead attention

CNN: Convolutional Neural Network  
LSTM: Long Short Term Memory  
GRU: Gated Recurrent Unit



## 2.5

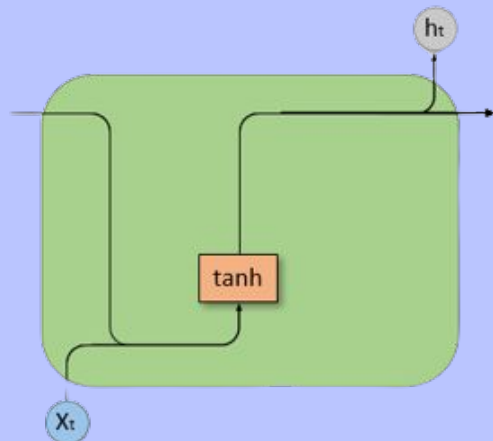
## Structure des réseaux élaborés

## b) GRU et LSTM, de quoi parle-t-on ?

- Le nombre de paramètres explosent assez vite

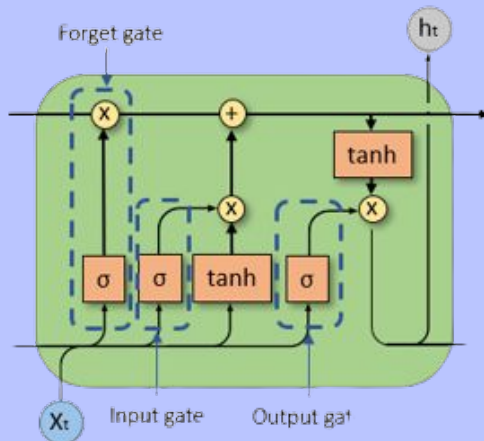
LSTM: Long Short Term Memory

GRU: Gated Recurrent Unit

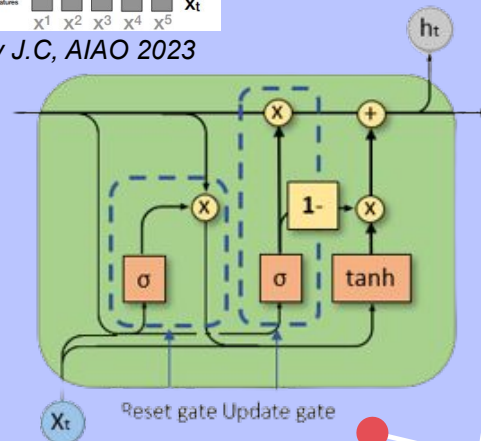


RNN

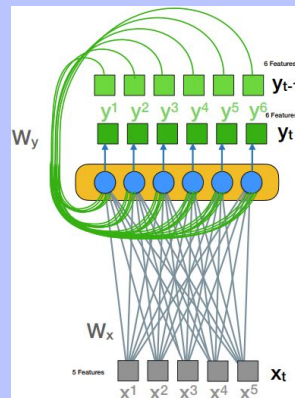
*towardsdatascience.com*



LSTM

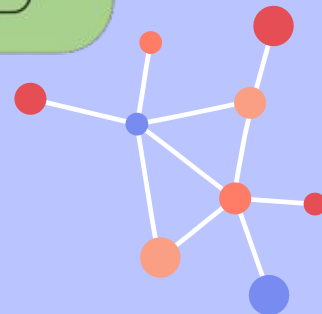


GRU



Gelly J.C, AIAO 2023

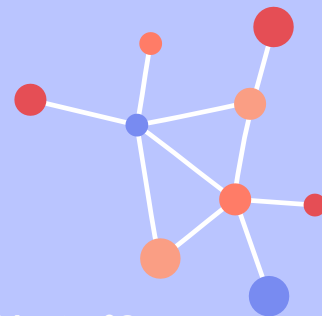
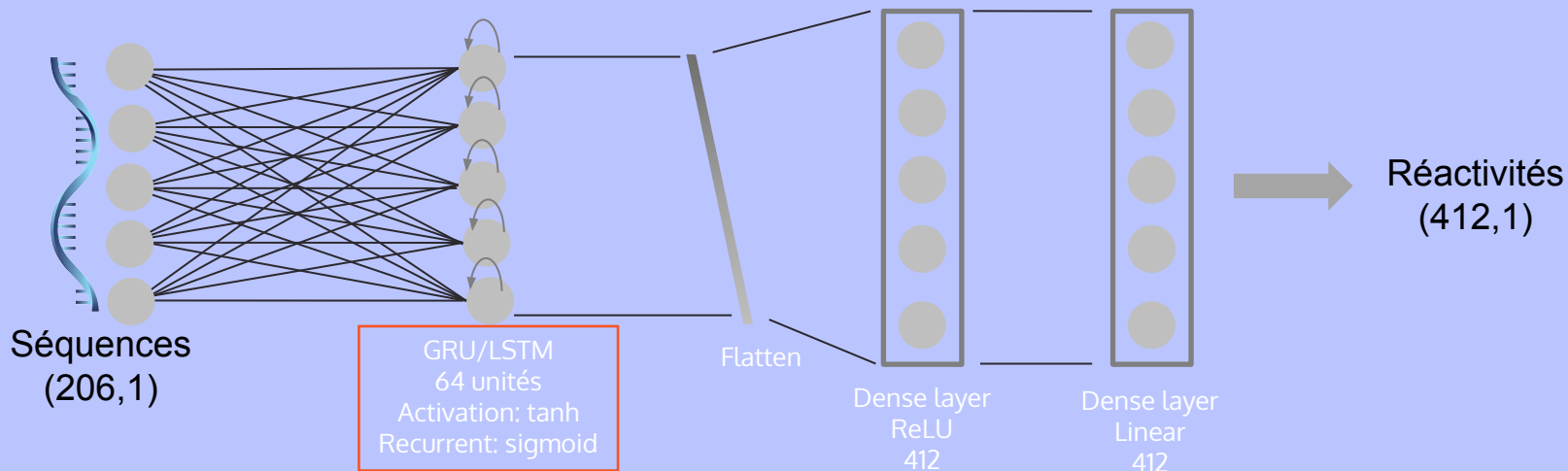
- Utilisation d'un système de portes
- Gestion dynamique de la mémoire lors de l'apprentissage



## 2.5

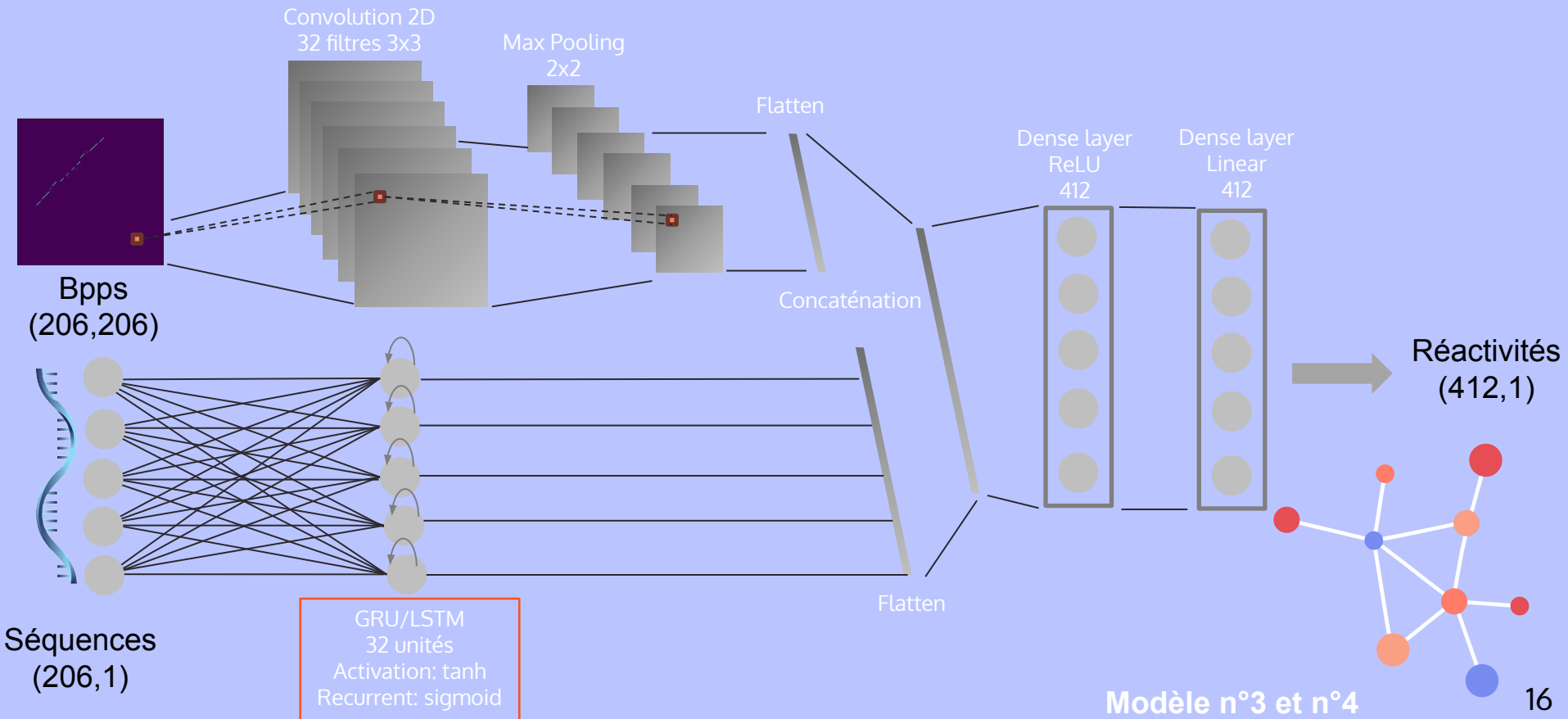
## Structures détaillées des réseaux

## c) Réseaux se basant uniquement sur les séquences



## 2.5 Structures des réseaux élaborés

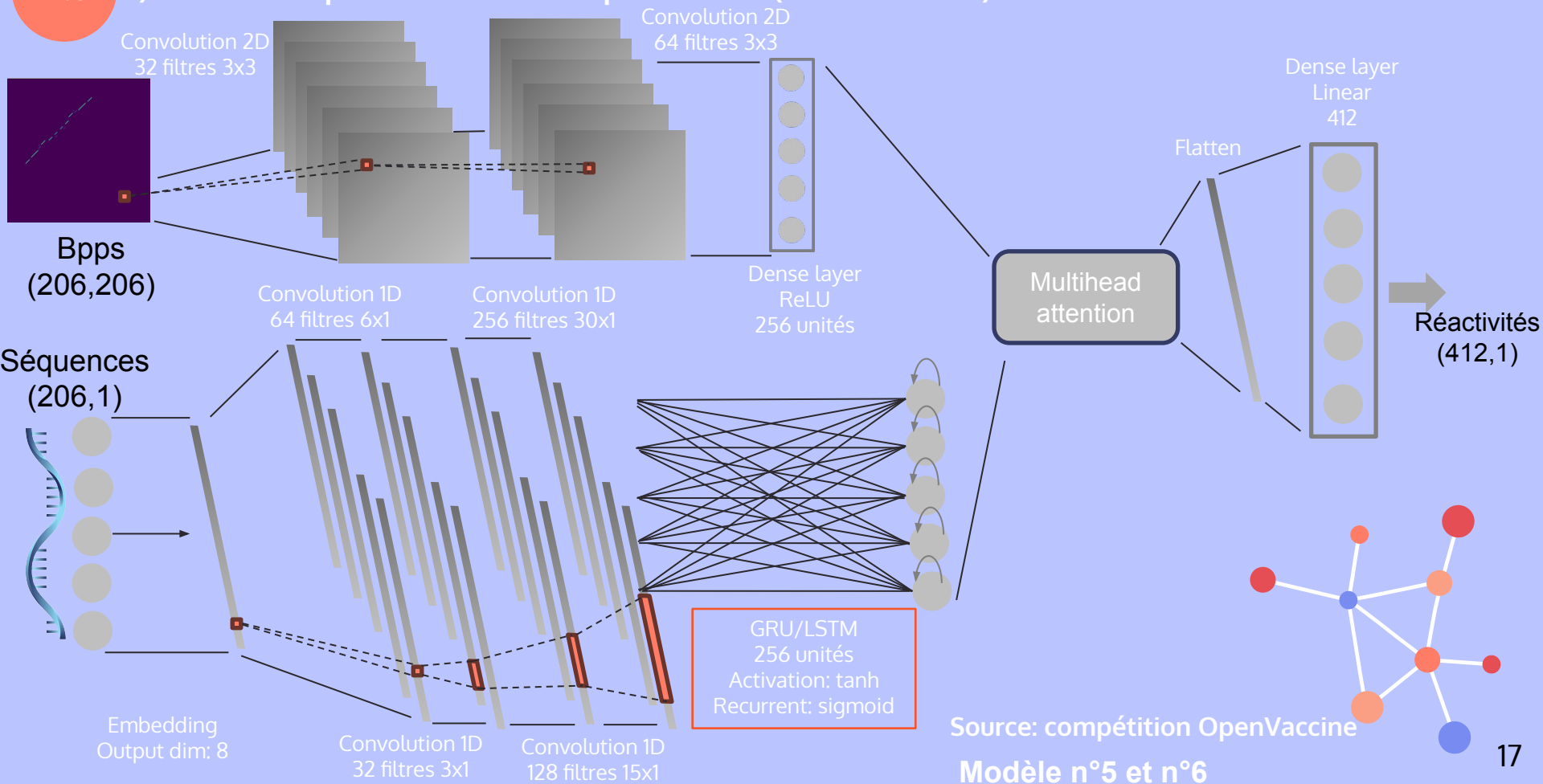
### d) Réseaux se basant sur les séquences et les cartes de contact



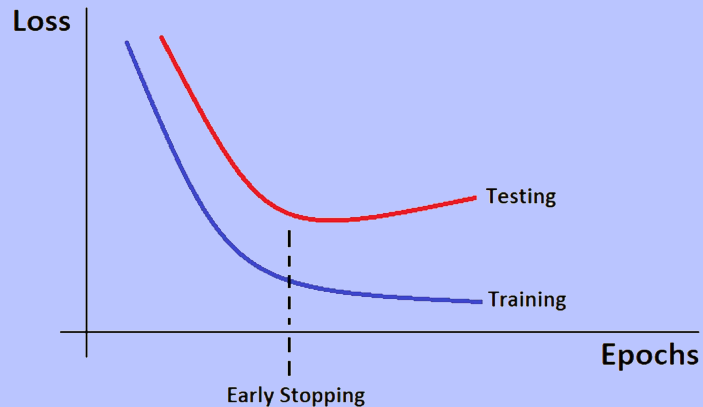


## 2.5 Structures des réseaux élaborés

### e) Réseaux inspirés du concours OpenVaccine (NullRecurrent)

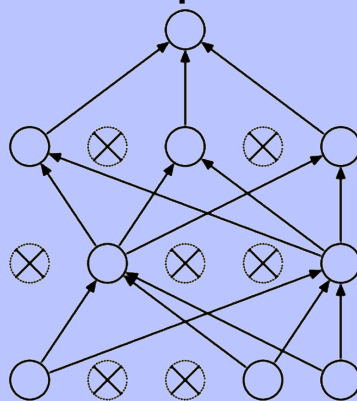


## Early stopping



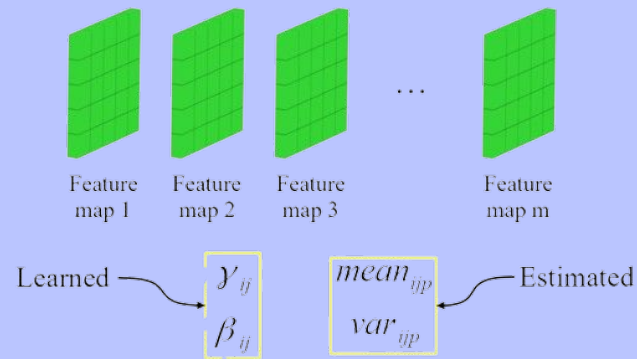
- Interruption de l'apprentissage dès le début du sur-apprentissage
- Contrôle glissant sur 5 epochs

## Dropout

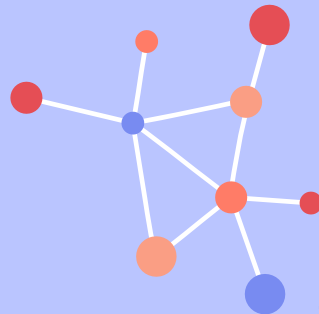


- Fixation aléatoire de paramètres aux valeurs nulles
- Fixé à 20%

## Batch normalization



- Normalisation batches



## 2.7 Choix de l'optimiseur et des hyperparamètres

- Optimiseur: Adam
- Batch size: 100
- Learning rate: 0,001
- Loss function utilisée:

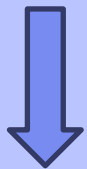
$$loss = \sum_{i=1}^n \sum_{j=1}^p (reactivity_{i,j} - N_w(sequence)_{i,j})^2$$

L2

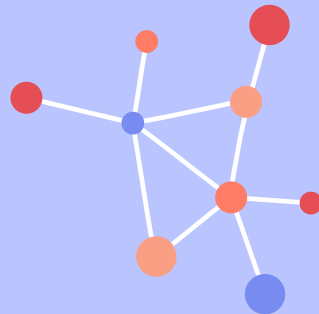
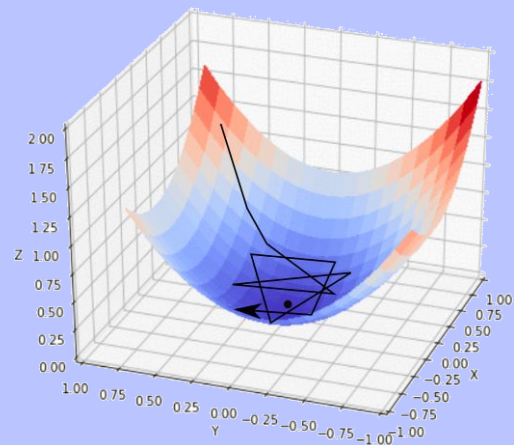
- Autre métrique utilisée:

$$mean\ absolute\ error = \sum_{i=1}^n \sum_{j=1}^p |reactivity_{i,j} - N_w(sequence)_{i,j}|$$

L1



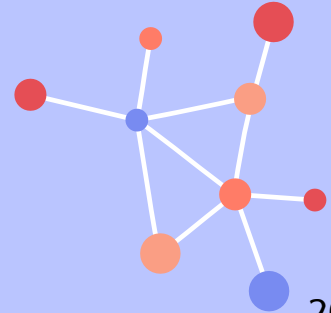
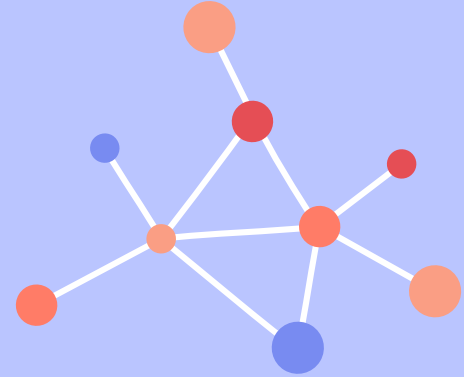
Métrique d'évaluation de Kaggle



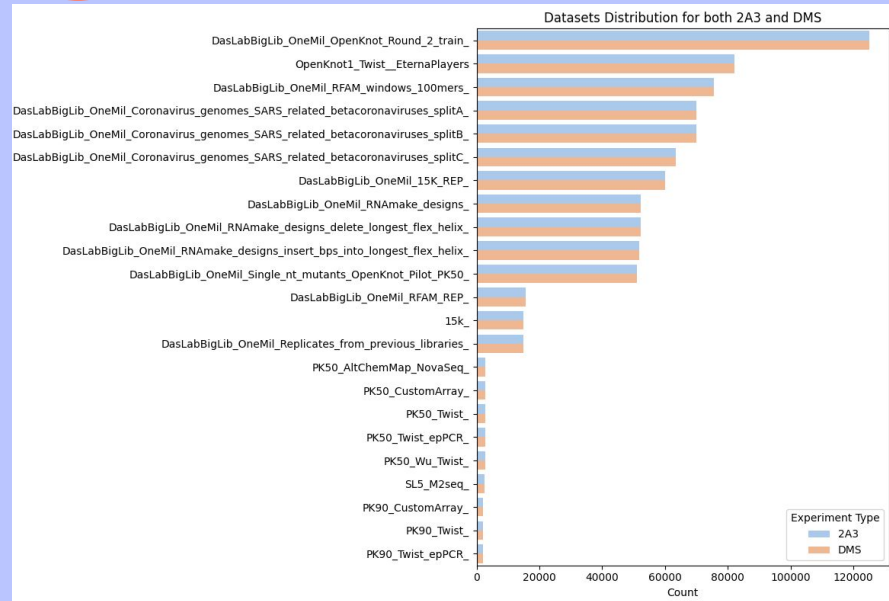


03

# Résultats et discussions

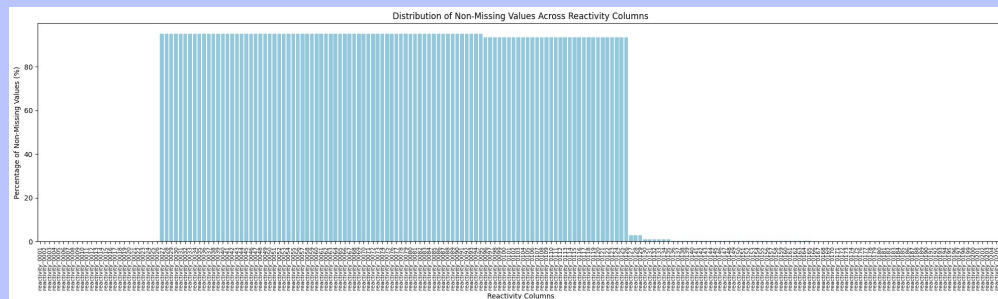


## 3.1 Résultats des analyses exploratoires des données



Répartition des séquences par base de données et par expérimentation

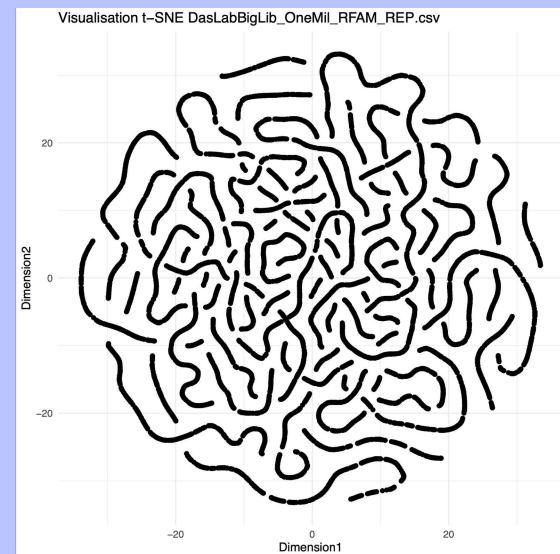
□ 764 117 séquences après nettoyage du train



Valeurs de réactivités manquantes

177 - 95.42% (1568354)  
170 - 1.83% (30000)  
115 - 1.66% (27290)  
155 - 0.79% (13038)  
206 - 0.30% (4998)

Longueurs des séquences



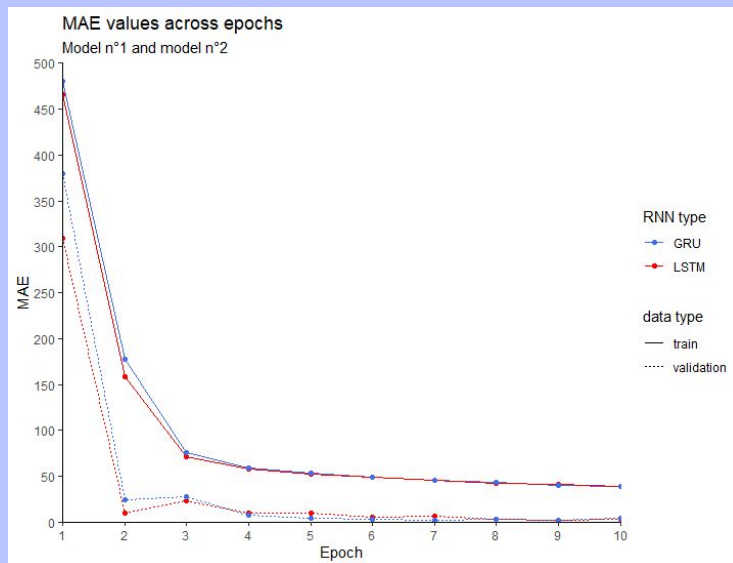
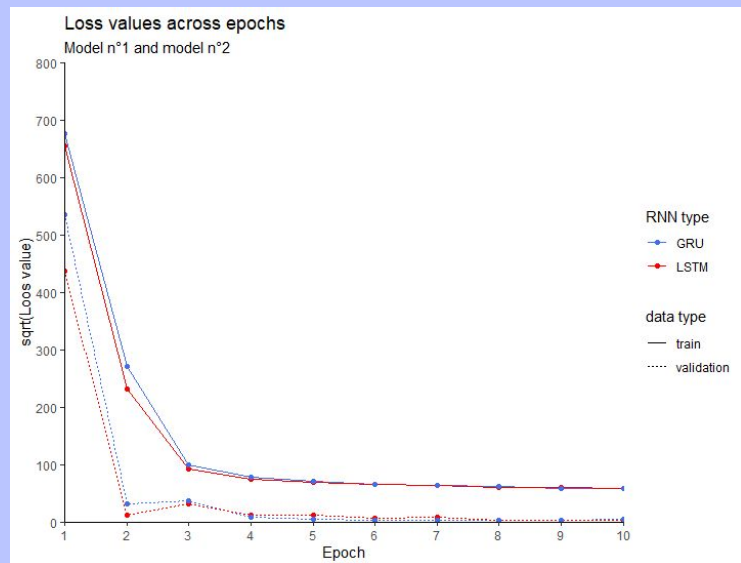
tSNE sur une base de données

## 3.2

## Performances des modèles

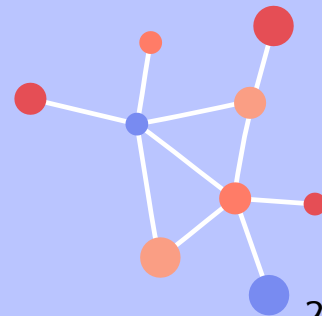
### a) Modèles se basant uniquement sur les séquences

#### Performances sur un sous échantillon



- Pas de sur/sous apprentissage
- Pas de différence flagrante de performances
- Modèle GRU plus « léger »
- En revanche, jeu de validation peut-être pas suffisamment représentatif

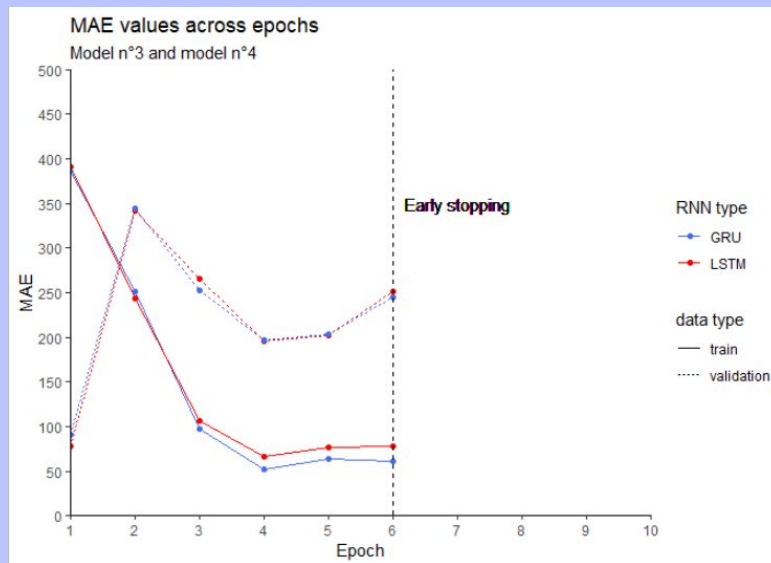
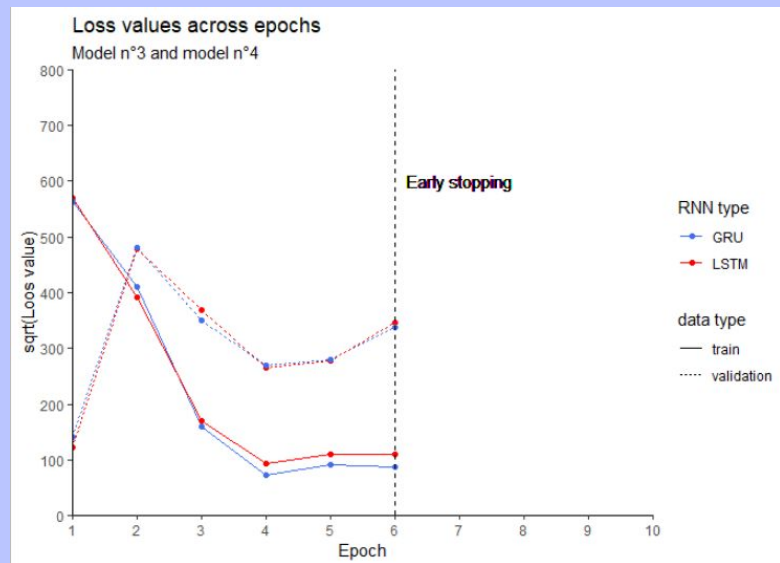
Modèle	Nombre de paramètres
Modèle n°1: LSTM	2 818 290
Modèle n°2: GRU	2 814 258



## 3.2

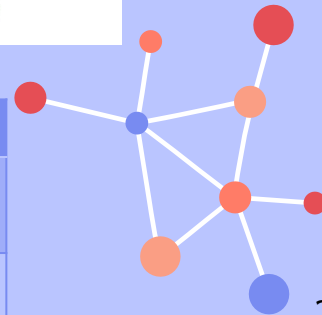
## Performances des modèles

## b) Modèles se basant sur les séquences et les cartes de contacts



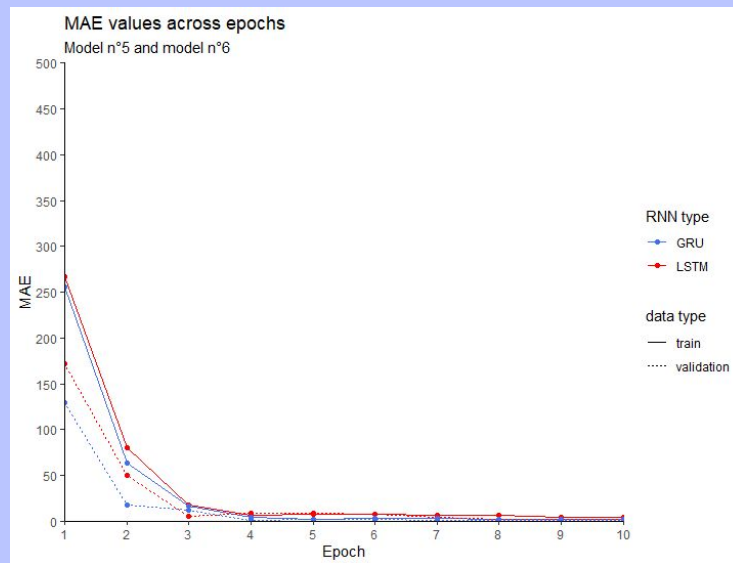
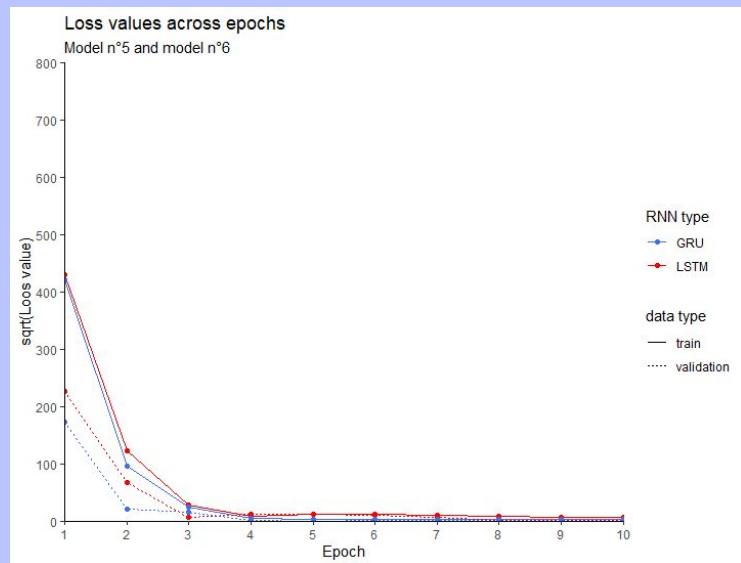
- Modèles présentant beaucoup de paramètres
- Surapprentissage (early stopping à 6 epoch)
- Une nouvelle fois, problème de représentativité de l'échantillon de validation

Modèle	Nombre de paramètres
Modèle n°3: LSTM + CNN	142 760 328
Modèle n°4: GRU + CNN	142 759 336



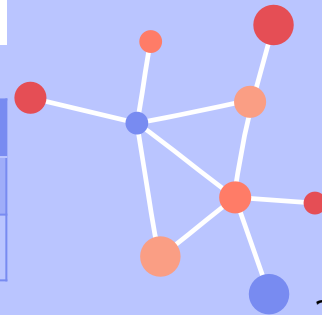
## 3.2

### Performances des modèles c) Modèles inspirés du concours OpenVaccine



- Même remarques que pour les modèles n°1 et n°2
- Modèles plus légers que les n°3 et n°5
- Meilleure apprentissage que les autres modèles (MAE = 1,20 à l'époch 10 sur le le NR2)

Modèle	Nombre de paramètres
Modèle n°5: NR 1	27 292 708
Modèle n°6: NR2	27 162 148





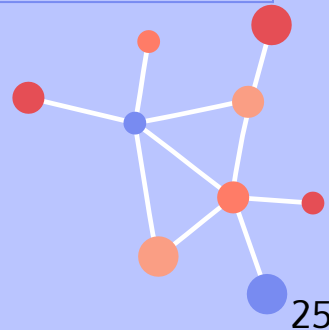
## 3.3

## Apprentissage partiel sur le jeu complet

Modèles	Epoch	Batch	Loss	Mean absolute error
Modèle n°1 LSTM	1	2902	5516,54	29,59
Modèle n°2 GRU	1	3114	5314.89	24.12
Modèle n°3 LSTM + CNN	1	17	364153,84	424,64
Modèle n°4 GRU + CNN	1	110	76344,67	128,43
Modèle n°5 NR1	1	2904	1268,62	4,20
Modèle n°6 NR2	1	1262	3211,80	8,34

*Snapshot des apprentissages des différents modèles*

- Un apprentissage chronophage sur l'ensemble des données





# CONCLUSIONS



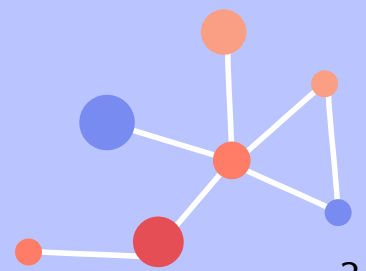
## ❑ Apprentissage sur un sous échantillon

- Manque de représentativité du jeu de validation
- Modèles GRU/LSTM simple manquent de performances
- Modèles GRU/LSTM combinés à un CNN comporte beaucoup de paramètres => sur-apprentissage
- Modèles NR1/NR2 possèdent à priori les meilleures performances et sont plus légers que les modèles GRU/LSTM combinés à un CNN

## ❑ Apprentissage sur le jeu complet

- Manque de ressources computationnelles

## ❑ Pour aller plus loin:

- Apprentissage sur jeu de données complet
  - Utilisation d'une validation croisée
- 

# CONCLUSIONS

## ▣ Compétences acquises:

- Prise en main de Tensorflow et Keras
- Connaissances plus approfondies sur le Machine Learning
- Utilisation d'un cluster de calcul
- Manipulation de fichiers volumineux avec les générateurs

## ▣ Difficultés rencontrées:

- Entraînement des modèles chronophages
- Manipulation des générateurs pour alimenter les modèles

# Bibliographie

1. *Thérapies à ARN* · Inserm, *La science pour la santé*. (s. d.). Inserm. <https://www.inserm.fr/dossier/therapies-a-arn/>
2. Contributeurs aux projets Wikimedia. (2023, 8 janvier). *Structure de l'ARN*.  
[https://fr.wikipedia.org/wiki/Structure\\_de\\_l%27ARN#:~:text=La%20structure%20secondaire%20d'un,r%C3%A9gions%20non%20appari%C3%A9es%20\(boucles\)](https://fr.wikipedia.org/wiki/Structure_de_l%27ARN#:~:text=La%20structure%20secondaire%20d'un,r%C3%A9gions%20non%20appari%C3%A9es%20(boucles))