

Utilisation de réseaux de neurones pour l'étude phyldynamique de modèles épidémiologiques

Inès Vati

Encadrant : Patrick Hoscheit¹

Tuteur Ecole : Romain Loiseau²

¹MaIAGE INRAE
Laboratoire de recherche Jouy-en-Josas

²Ecole Des Ponts ParisTech
Département Ingénierie Mathématique et Informatique



Plan de l'exposé

- 1 Introduction
 - Modèle épidémiologique
 - Méthodes de Deep Learning
- 2 Simulation des données
 - Simulation des arbres phylogénétiques
 - Vectorisation des arbres - *CBLV representation*
- 3 Apprentissage et Évaluation des Réseaux de Neurones
 - Architectures
 - *Loss* et *Optimizer*
 - Courbes d'apprentissage
 - Sélection de modèle
 - Comparaison au modèle nul
 - Relation entre erreur et taille des arbres
- 4 Conclusion

Modèle épidémiologique

Observations

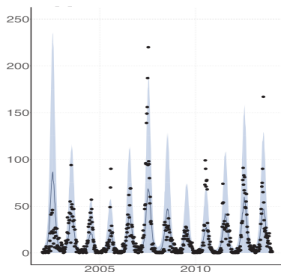


Figure 1 – Nombre de cas recensés par le NDSS au Cambodge (CHAMPAGNE et al., 2019)

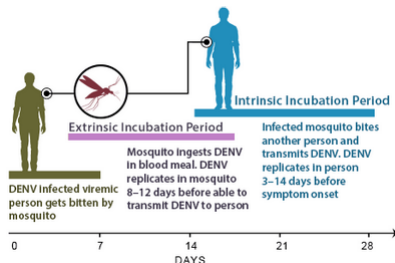


Figure 2 – Dengue : maladie virale transmise par les moustiques du genre *Aedes* (DISEASE CONTROL et PREVENTION, 2023)

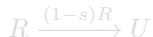
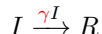
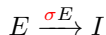
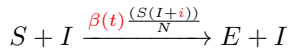
Modèle épidémiologique

Paramètres

- N : nombre total d'individus dans la population
($N = S + E + I + R$)
- S : individu sensible ($S(t = 0) = N - 1$)
- E : individu infecté mais non contagieux
- I : individu infecté et contagieux
- R : individu rétabli
- U : individu rétabli dont la séquence génétique virale n'a pas été analysée

Modèle épidémiologique

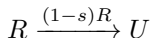
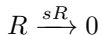
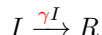
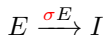
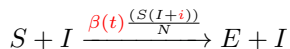
Réactions



- $\beta(t)$: taux de transmission
- σ : taux d'incubation
- γ : taux de guérison
- i : "paramètre d'import"
- s : taux d'échantillonnage

Modèle épidémiologique

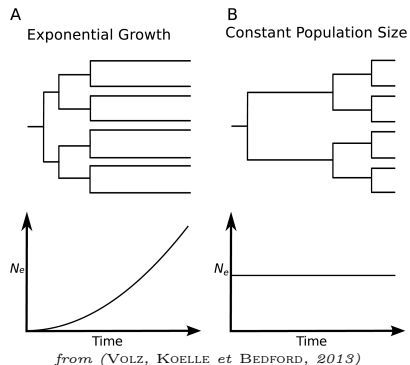
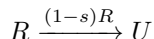
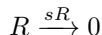
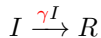
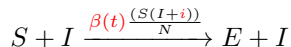
Réactions



- $\beta(t)$: taux de transmission
- σ : taux d'incubation
- γ : taux de guérison
- i : "paramètre d'import"
- s : taux d'échantillonnage

Modèle épidémiologique

Réactions



Modèle épidémiologique

Saisonnalité

(CHAMPAGNE et al., 2019)

$$\beta(t) = B \left[1 + b \sin \left(2\pi \left(\frac{t}{T} + p \right) \right) \right]$$

avec

- B : moyenne temporelle du taux de transmission
- b : amplitude du taux de transmission
- p : phase
- $T=1$ an : période du signal

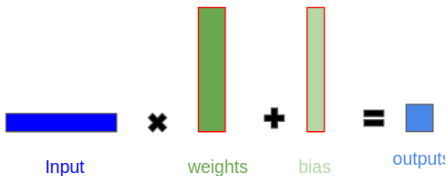
Deep Learning - Intérêts

Pourquoi étudier des méthodes deep learning? (VOZNICA et al., 2022)

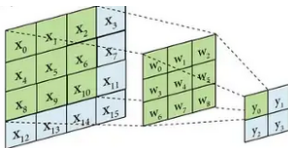
- Les modèles épidémiologiques complexes reposent sur des ensembles d'EDO qui ne peuvent être résolues analytiquement
- Imprécision et instabilité numérique pour les approches Bayésiennes (dû aux approximations à chaque noeud de l'arbre)
- D'après (VOZNICA et al., 2022), estimations des paramètres plus précises que par les méthodes standards
- Les réseaux de neurones sont des méthodes d'inférence rapides
- Meilleure efficacité sur les modèles épidémiologiques complexes

Deep Learning - Principes

- Perceptron



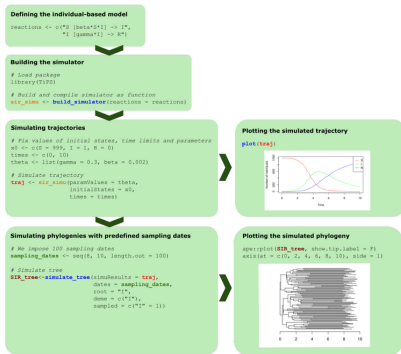
- Convolution



from A.Vidhya

Simulation des arbres

Simulateur TiPS (DANESH et al., 2022)



- Méthode "approximate" (τ -leaping algorithm (CAO et SAMUELS, 2009))
- Filtration des séries temporelles périodiques en calculant la transformée de Fourier

Figure 3 – Pipeline de *TiPS*

Simulation des arbres

Paramètres des simulations

- $N=100000$, simulation sur $5T = 500$
- Période d'incubation $\frac{1}{\sigma} \sim \mathcal{N}(5.9, 0.5)$ dans $[3, 15]$ (CHAMPAGNE et al., 2019)
- Période infectieuse $\frac{1}{\gamma} \sim \mathcal{N}(7, 0.5)$ dans $[3, 10]$ (CHAMPAGNE et al., 2019)
- Temps de simulation (1 arbre) :
~ 0.15s pour un nombre de feuilles entre 60 et 600
- Taux de succès : 20%

Simulation des arbres

Paramètres des simulations

- $N=100000$, simulation sur $5T = 500$
- Période d'incubation $\frac{1}{\sigma} \sim \mathcal{N}(5.9, 0.5)$ dans $[3, 15]$ (CHAMPAGNE et al., 2019)
- Période infectieuse $\frac{1}{\gamma} \sim \mathcal{N}(7, 0.5)$ dans $[3, 10]$ (CHAMPAGNE et al., 2019)
- Temps de simulation (1 arbre) :
~ 0.15s pour un nombre de feuilles entre 60 et 600
- Taux de succès : 20%

Simulation des arbres

Exemple de trajectoires obtenues

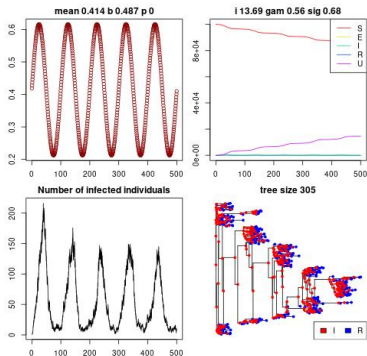


Figure 4 – Dynamique simulée et filtrée

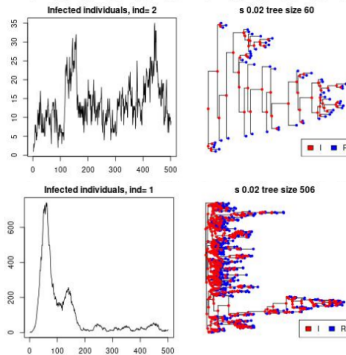
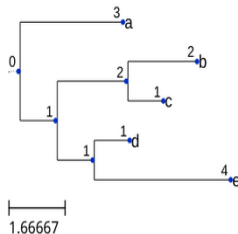


Figure 5 – Dynamiques rejetées par l'analyse du spectre de Fourier (McMASTER, 2010)

Vectorisation des arbres - *CBLV representation*

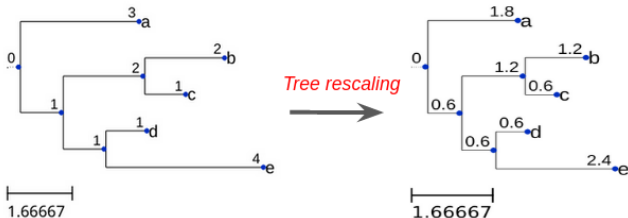
Étapes de l'encodage des arbres (VOZNICA et al., 2022)



- Python, package *ete3*

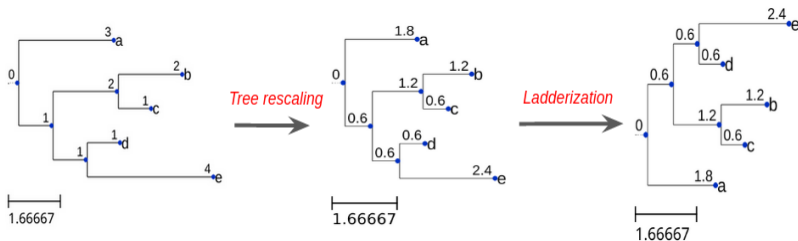
Vectorisation des arbres - *CBLV representation*

Étapes de l'encodage des arbres (VOZNICA et al., 2022)



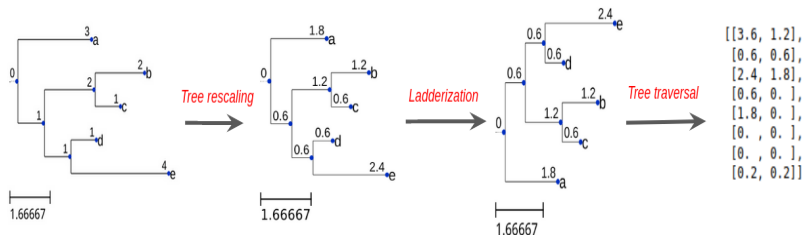
Vectorisation des arbres - *CBLV representation*

Étapes de l'encodage des arbres (VOZNICA et al., 2022)



Vectorisation des arbres - *CBLV representation*

Étapes de l'encodage des arbres (VOZNICA et al., 2022)



Vectorisation des arbres - *CBLV representation*

Algorithme 1 *traversal*(noeud courant, noeud précédent)

si noeud courant est la racine **alors**

profondeur précédente \leftarrow 0

sinon

profondeur précédente \leftarrow profondeur du parent du noeud courant

fin si

si noeud courant est une feuille **alors**

stocker distance entre noeud courant et noeud précédent

sinon

traversal(enfant1 du noeud courant, noeud précédent)

stocker profondeur du noeud courant

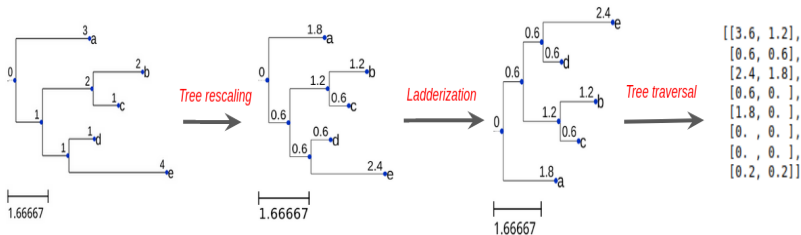
noeud précédent \leftarrow noeud courant

traversal(enfant2 du noeud courant, noeud précédent)

fin si

Vectorisation des arbres - *CBLV representation*

Étapes de l'encodage des arbres (VOZNICA et al., 2022)



Temps d'exécution de vectorisation : 12s pour 1400 arbres

Architectures

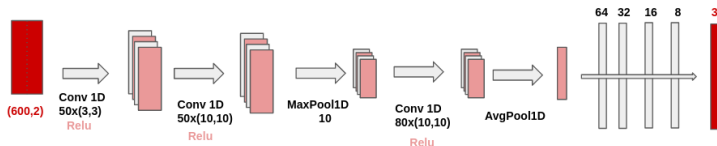


Figure 6 – Différents modèles de réseaux de neurones

Architectures

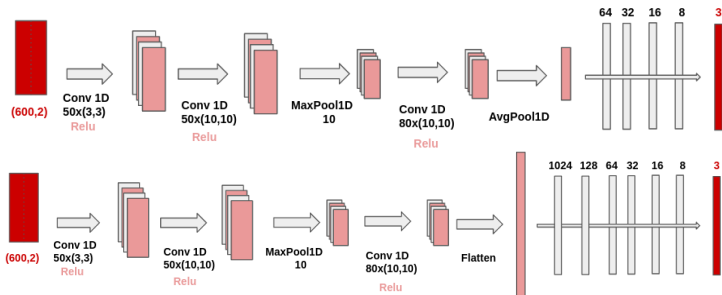


Figure 6 – Différents modèles de réseaux de neurones

Architectures

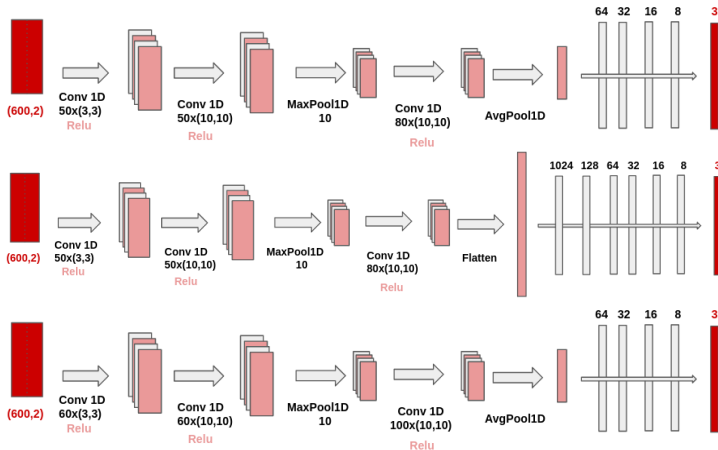


Figure 6 – Différents modèles de réseaux de neurones

Fonction de Loss

- MAPE : *Mean Absolute Percentage Error* (VOZNICA et al., 2022)

$$\text{MAPE} = 100 \frac{|y_{true} - y_{pred}|}{y_{true}}$$

- MSE : *Mean Squared Error*

$$\text{MSE} = |y_{true} - y_{pred}|^2$$

Optimizer (DOZAT, 2016)

Algorithme 2 Adam (*Adaptive moment estimation*)

$$\begin{aligned} \mathbf{g}_t &\leftarrow \nabla_{\theta_{t-1}} l(\theta_{t-1}) \\ \mathbf{m}_t &\leftarrow \mu \mathbf{m}_{t-1} + (1 - \mu) \mathbf{g}_t \\ \hat{\mathbf{m}}_t &\leftarrow \frac{\mathbf{m}_t}{1 - \mu^t} \\ \mathbf{n}_t &\leftarrow \nu \mathbf{n}_{t-1} + (1 - \nu) \mathbf{g}_t^2 \\ \hat{\mathbf{n}}_t &\leftarrow \frac{\mathbf{n}_t}{1 - \nu^t} \\ \theta_t &\leftarrow \theta_{t-1} - l_r \frac{\hat{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{n}}_t + \epsilon}} \end{aligned}$$

Algorithme 3 NAdam (*with Nesterov's accelerated gradient*)

$$\begin{aligned} \mathbf{g}_t &\leftarrow \nabla_{\theta_{t-1}} l(\theta_{t-1}) \\ \mathbf{m}_t &\leftarrow \mu \mathbf{m}_{t-1} + (1 - \mu) \mathbf{g}_t \\ \hat{\mathbf{m}}_t &\leftarrow \frac{\mathbf{m}_t}{1 - \mu^t} \\ \mathbf{n}_t &\leftarrow \nu \mathbf{n}_{t-1} + (1 - \nu) \mathbf{g}_t^2 \\ \hat{\mathbf{n}}_t &\leftarrow \frac{\mathbf{n}_t}{1 - \nu^t} \\ \bar{\mathbf{m}}_t &\leftarrow (1 - \mu_t) \hat{\mathbf{g}}_t + \mu_{t+1} \hat{\mathbf{m}}_t \\ \theta_t &\leftarrow \theta_{t-1} - l_r \frac{\bar{\mathbf{m}}_t}{\sqrt{\hat{\mathbf{n}}_t + \epsilon}} \end{aligned}$$

Courbes d'apprentissage

Epoch 225/1000
156/156 [=====] - 15s 95ms/step - loss: 0.0105 - val_loss: 0.0089
Training with 318586 training data took 55.55382755994797 minutes

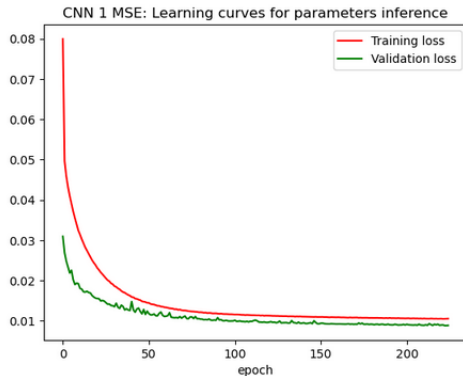


Figure 7 – *Learning Curves* obtenue pour le modèle 1 $l_r = 1.10^{-4}$

Sélection de modèle

Comparaison des erreurs sur les données de test

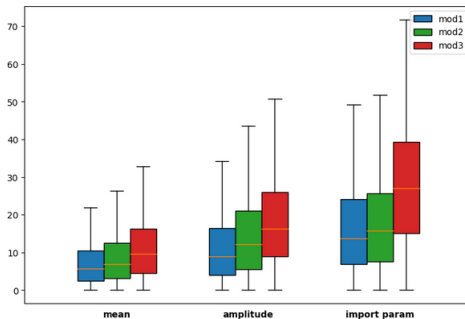


Figure 8 – Boxplot des erreurs relatives absolues par paramètre et par réseau (en %)

Sélection de modèle

Erreur sur les données de test du réseau 1

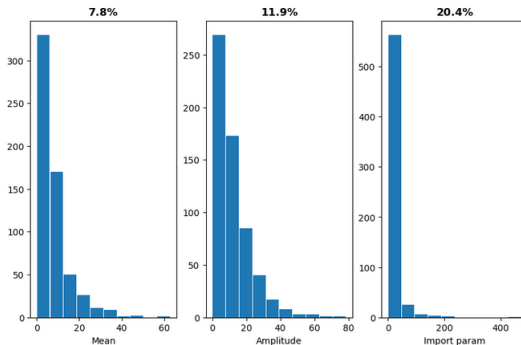


Figure 9 – Histogramme des erreurs absolues relatives pour le modèle 1 (en **gras** est indiquée la moyenne des erreurs relatives absolues)

Comparaison au modèle nul

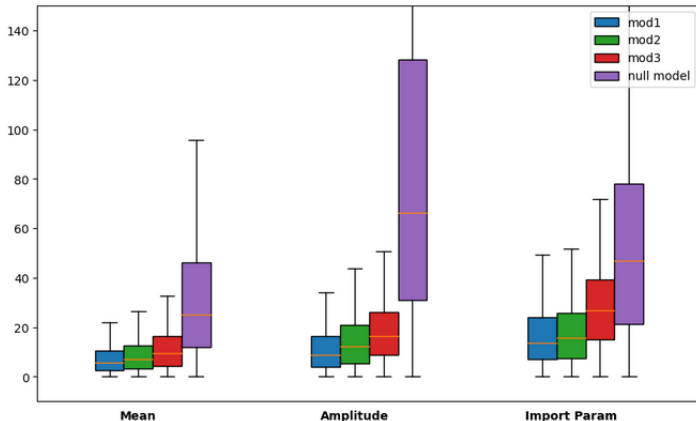


Figure 10 – Boxplots des erreurs sur les données de test (en %)
Le modèle nul est un prédicteur aléatoire

Relation entre erreur et taille des arbres

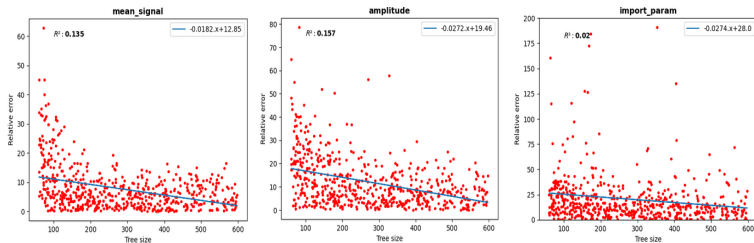
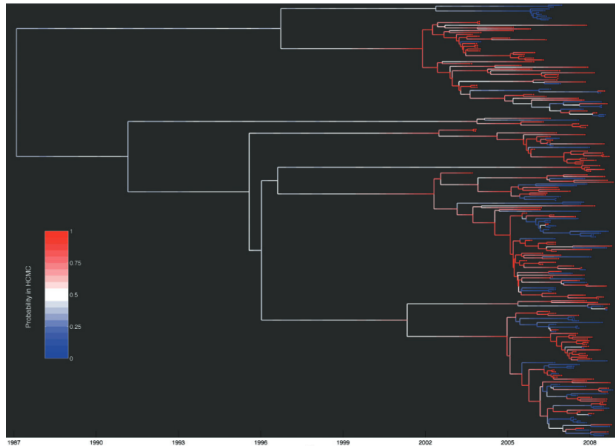


Figure 11 – L'erreur des prédictions du réseaux diminue avec la taille des arbres. (regression linéaire sur 600 arbres)

Conclusion

Perspectives

- Appliquer à des données réelles



from (RASMUSSEN, BONI et KOELLE, 2014)

Conclusion

Perspectives

- Comparer à d'autres méthodes d'inférence de paramètres (*BEAST*)
- Inférer la probabilité d'échantillonnage
- Comparer avec d'autres représentations d'arbres
- Comparer avec d'autres modèles épidémiologiques (CHAMPAGNE et al., 2019) de la dengue
- Approches ensemblistes de deep learning

References I






CAO, Yang et David C. SAMUELS (2009). “Discrete Stochastic Simulation Methods for Chemically Reacting Systems”. In : *Methods in enzymology* 454, p. 115-140. ISSN : 0076-6879. DOI : 10.1016/S0076-6879(08)03805-6. URL : <https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3492891/> (visité le 11/01/2023).



CHAMPAGNE, Clara et al. (mars 2019). “Dengue modeling in rural Cambodia: Statistical performance versus epidemiological relevance”. In : *Epidemics* 26, p. 43-57. ISSN : 17554365. DOI : 10.1016/j.epidem.2018.08.004. URL : <https://linkinghub.elsevier.com/retrieve/pii/S1755436517301706> (visité le 10/01/2023).

References II

-  DANESH, Gonché et al. (18 mai 2022). *TiPS: rapidly simulating trajectories and phylogenies from compartmental models*. Pages: 2020.11.09.373795 Section: New Results. DOI : 10.1101/2020.11.09.373795. URL : <https://www.biorxiv.org/content/10.1101/2020.11.09.373795v2> (visité le 10/01/2023).
-  DISEASE CONTROL, Centers for et PREVENTION (2023). *Mosquito-borne Transmission*. URL : <https://www.cdc.gov/dengue/training/cme/ccm/page45915.html> (visité le 11/01/2023).
-  DOZAT, Timothy (2016). “Incorporating Nesterov Momentum into Adam”. In.

References III



McMASTER (2010). *Spectral Analysis in R - McMaster Universitybolker/eeid/2010/Ecology/Spectral.pdf · Spectral Analysis in R Helen J. Wearing June 8, 2010 Contents 1 Motivation 1 2 What is spectral - [PDF Document].*

fddocuments.net. URL :

<https://fddocuments.net/document/spectral-analysis-in-r-mcmaster-university-bolkereeid2010ecologyspectralpdf.html> (visité le 11/01/2023).

References IV



RASMUSSEN, David A., Maciej F. BONI et Katia KOELLE (fév. 2014). “Reconciling Phylodynamics with Epidemiology: The Case of Dengue Virus in Southern Vietnam”. In : *Molecular Biology and Evolution* 31.2, p. 258-271. ISSN : 0737-4038. DOI : 10.1093/molbev/mst203. URL : <https://doi.org/10.1093/molbev/mst203> (visité le 14/01/2023).



VOLZ, Erik M., Katia KOELLE et Trevor BEDFORD (mars 2013). “Viral Phylodynamics”. en. In : *PLOS Computational Biology* 9.3. Publisher: Public Library of Science, e1002947. ISSN : 1553-7358. DOI : 10.1371/journal.pcbi.1002947. URL : <https://journals.plos.org/ploscompbiol/article?id=10.1371/journal.pcbi.1002947> (visité le 14/01/2023).

References V



VOZNICA, J. et al. (6 juill. 2022). “Deep learning from phylogenies to uncover the epidemiological dynamics of outbreaks”. In : *Nature Communications* 13.1. Number: 1 Publisher: Nature Publishing Group, p. 3896. URL : <https://www.nature.com/articles/s41467-022-31511-0> (visité le 10/01/2023).

Merci de votre attention !



<https://github.com/InesVATI/phylodeepINRAE>