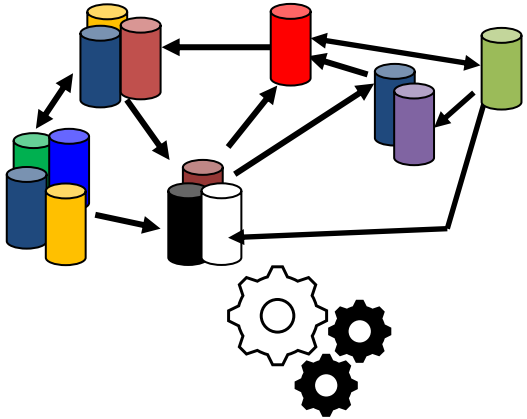
# Réseau Recherche Reproductible

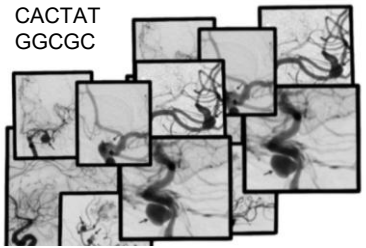
## *Journée Préservation de l'Intelligence Artificielle*

Sarah Cohen-Boulakia
Université Paris-Saclay

# Analyse de données en Bioinformatique



TGACTT
TGACTT TTGGGC
TTGGGC CACTAT
TGACTT GGCGC
TTGGGC
CACTAT
GGCGC

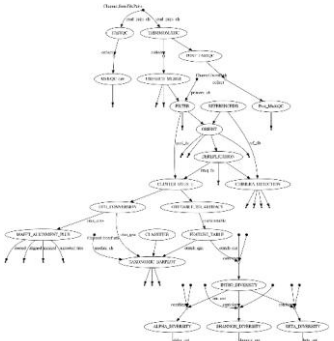


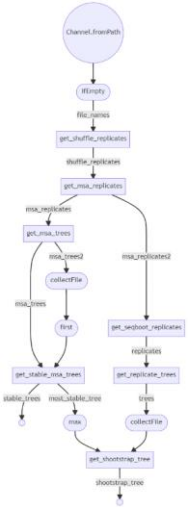
**Sources de données**
Distribuées
Réseau hétérogène
> 1,000 (NAR)

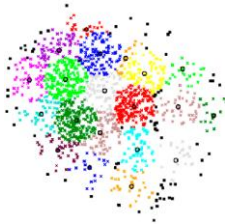
**Outils – Scripts**
Distribués
Hétérogènes
> 30,000
(bio.tools)

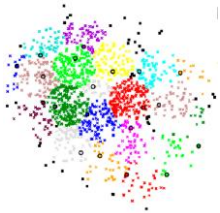
*Quel jeu de données d'entrée ai-je utilisé ? Quel paramétrage? Quelles versions d'outils ?*

**Pipelines d'analyse**
Combinaison d'outils
Environnements & platformes variées
> 1,000 workflows
(GitLab, ….)







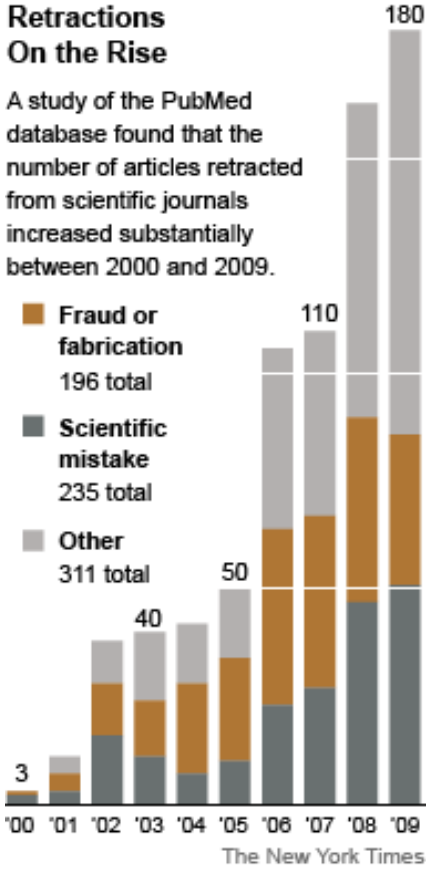


# La crise de la reproductibilité

Chaque communauté a ses papiers de reference

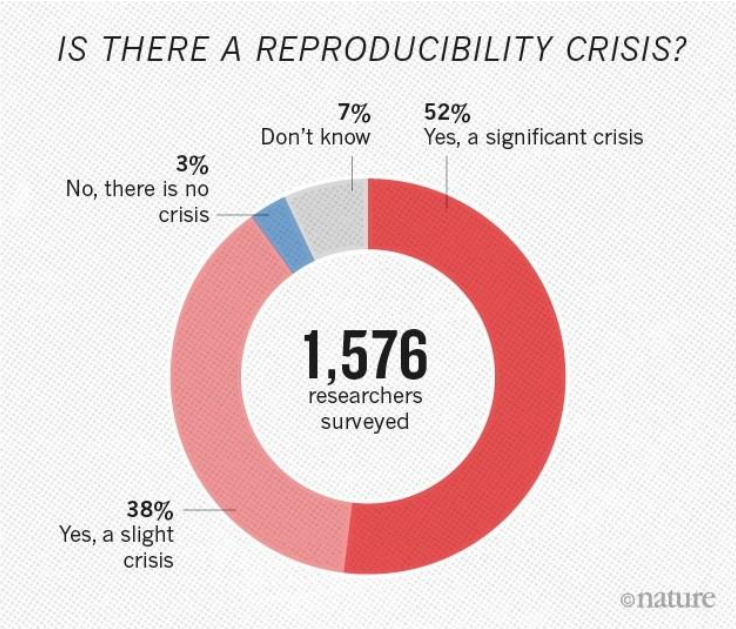
**Nekrutenko & Taylor - Nature Genetics (2012) [1]**
50 articles 2011 utilisant l'aligneur Burrows-Wheeler
31/50 (62%) ne fournissent aucune information
pas de version des outils, pas de paramètres
pas de séquence génomique de réference
7/50 (14%) fournissent toutes les informations necessaires

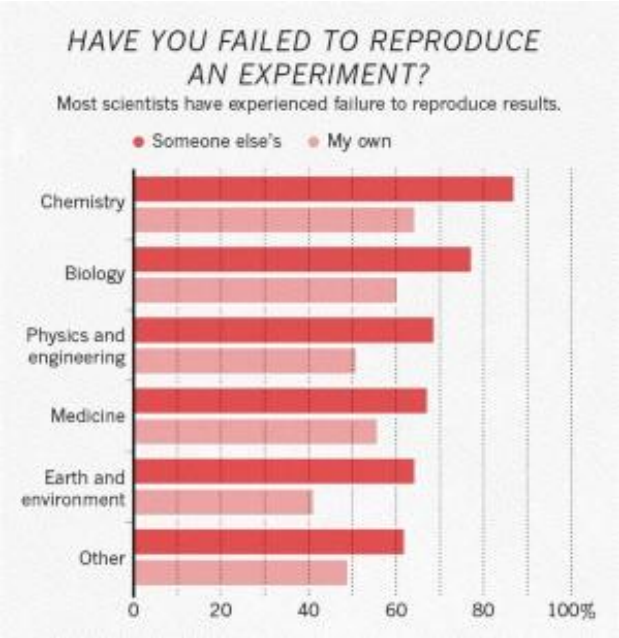
**Alsheikh-Ali et al, PLoS one (2011) [2]**
10 articles du top-50 des revues fort IF → 500 articles
149 (30%) n'ont pas de politique de mise à disposition des données (0% données mises à disposition)
Sur les 351 articles restants
208 articles (59%) ne respectent pas les instructions de mise à disposition des données des éditeurs
143 indique un souhait de partager "*willingness* to share"
Seuls 47 des articles (9%) ont deposé les données en ligne


New York Times 2011

# Survey Nature - *1,500 scientists lift the lid on reproducibility (2016)*





nature

Explore content · About the journal · Publish with us

nature › news feature › article

News Feature | Published: 25 May 2016

**1,500 scientists lift the lid on reproducibility**

Monya Baker

*Nature* **533**, 452–454 (2016) | Cite this article

**209k** Accesses | **2372** Citations | **5151** Altmetric | Metrics

La crise a touché toutes les disciplines [3]

# Types de Reproductibilité

## Reproductibilité Empirique

Informations détaillées sur les expériences
NB : Le chercheur contrôle le cadre de l'expérience
Son savoir faire peut entrer en jeu
Mise à disposition des données, méthodes de collecte des données

## Reproductibilité Observationnelle

Informations détaillées sur les observations
NB : Le chercheur ne contrôle pas le cadre – il observe
Mise à disposition des données, méthodes de collecte des données


Le Monde

SCIENCES • PHYSIQUE

### Des meringues pour faire goûter la démarche scientifique

**Afin d'expliquer au grand public le défi que représente la reproductibilité des résultats en sciences, une équipe rennaise a choisi une approche originale : utiliser la confection de pâtisseries comme terrain d'expérimentation.**

Par David Larousserie
Publié le 23 décembre 2024 à 06h00, modifié le 23 décembre 2024 à 15h17 · Lecture 3 min.

# Types de Reproductibilité

**Reproductibilité Empirique**
Informations détaillées sur les expériences
NB : Le chercheur contrôle le cadre de l'expérience
Son savoir faire peut entrer en jeu
Mise à disposition des données, méthodes de collecte des données

**Reproductibilité Observationnelle**
Informations détaillées sur les observations
NB : Le chercheur ne contrôle pas le cadre – il observe
Mise à disposition des données, méthodes de collecte des données

**Reproductibilité Statistique**
Informations détaillées sur le choix des tests statistiques, les paramètres des modèles, les seuils de décision...
Pré-enregistrement du design de l'étude pour prévenir la manipulation des p-valeurs et autres manipulations

**Reproductibilité Computationnelle**
Informations détaillées sur le code, le logiciel, le matériel et les détails de l'implémentation
Documenter *comment* les données ont été produites

# Remarques

Un même résultat scientifique peut avoir trait à **plusieurs types de reproductibilité**

E.g., Épidémiologie : reproductibilité observationnelle, statistique, computationnelle

La **reproductibilité statistique** ne concerne pas que les mathématiciens et la reproductibilité computationnelle pas que les informaticiens

**Science des données touche un nombre toujours croissant de disciplines**

La **reproductibilité computationnelle** a longtemps été considérée comme facile à obtenir (**à tort**)

Il existe des *niveaux de* reproductibilité : redo –replicate – reproduce – reuse…

Attention : d'une communauté à l'autre les termes changent !

# Niveaux de reproductibilité : cadre (proposition)

## Niveau 1

**Repeat**

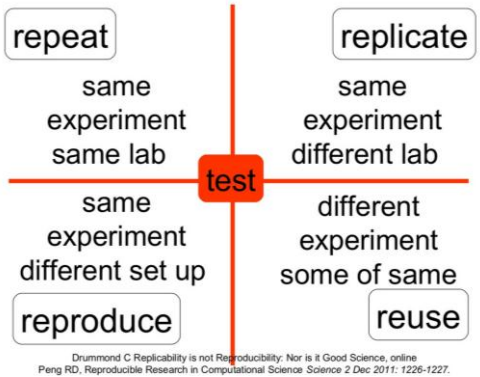
La traçabilité complète permet de *refaire* à l'identique

Mêmes données

*Redo* – Réexecute

But : capter le maximum d'*informations*

*permettant d'expliquer un résultat*




Drummond C Replicability is not Reproducibility: Nor is it Good Science, online
Peng RD, Reproducible Research in Computational Science *Science 2 Dec 2011: 1226-1227.*

## Niveau 2

**Replicate**

On s'autorise quelques variations

Mêmes résultats – données similaires

But : on teste les limites d'une approche

# Niveaux de reproductibilité : cadre (proposition)

## Niveau 1

**Repeat**

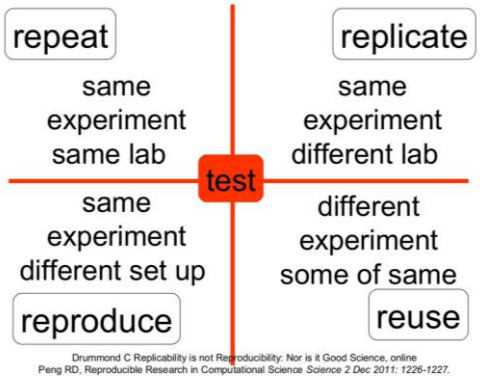
La traçabilité complète permet de *refaire* à l'identique

Mêmes données

*Redo* – Réexecute

But : capter le maximum d'*informations*

*permettant d'expliquer un résultat*




## Niveau 2

**Replicate**

On s'autorise quelques variations

Mêmes résultats – données similaires

But : on teste les limites d'une approche

## Niveau 3

**Reproduce**

Même résultat – même *inférence*

*Mais les moyens/procédures/méthodes/données peuvent avoir changé*

## Reuse

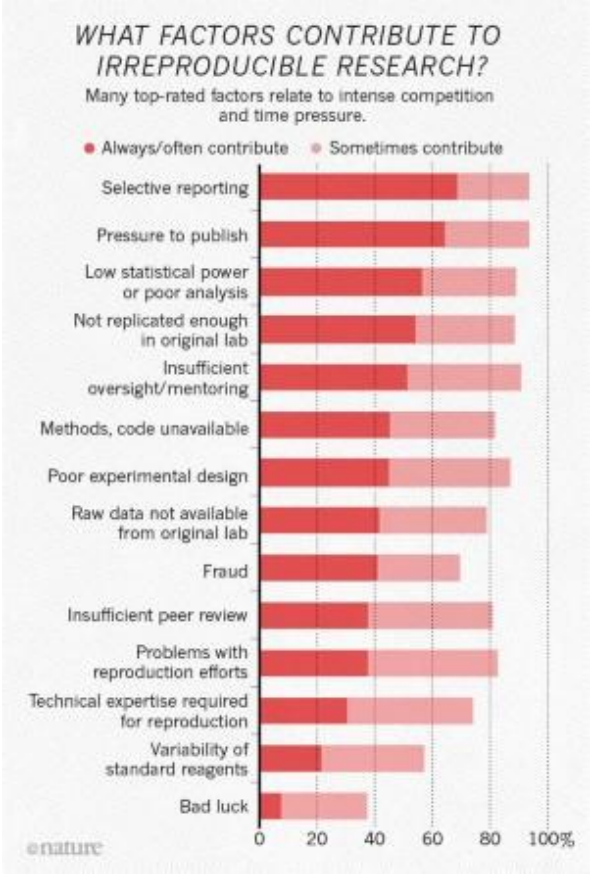
On s'adapte à de nouveaux besoins

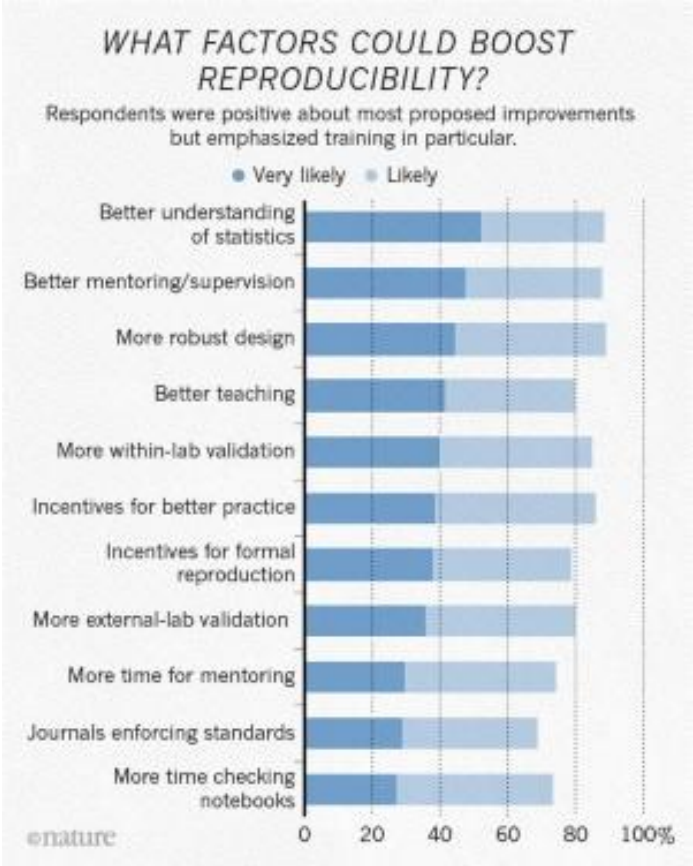
On réutilise en partie – dans un autre contexte

On peut obtenir un résultat différent

→ Science Cumulative

# Survey Nature – *Les raisons et les pistes de solutions*




**Ralentir**

Moins de pression à publier

Plus de temps pour des résultats plus robustes

**Être formé**

Compétences à acquérir

Bonnes pratiques

Echanges

**Valoriser (*incentives*)**

Journaux

Institutions

# IA et reproductibilité : l'IA, un objet difficile à reproduire (1/2)

## Les freins à la reproductibilité

### Les données
Datasets propriétaires, non versionnés, biaisés…
Problèmes d'accès et de licence

### Le code et les modèles
Code mal documenté ou non publié
Poids de modèles non partagés
Enfer des dépendances logicielles

### Le hasard et le matériel
Variabilité selon le hardware (GPU/TPU)
Seeds aléatoires non fixées
Coût prohibitif de ré-entraînement

### Le prompt et le contexte
Sensibilité au prompt : impact formulation
Absence de standard : prompts rarement publiés
Le contexte de la conversation : historique…
La température et les paramètres de génération
Dérive des modèles dans le temps

### Les pratiques de publication
Hyperparamètres non documentés
Cherry-picking des métriques

# IA et reproductibilité : l'IA, un objet difficile à reproduire (2/2)

## De bonnes pratiques existent (un bon début…)

### Les données
Publier les datasets sur plateformes ouvertes
Versionner les données
Documenter…

### Le code et les modèles
Publier le code sur Git avec licence claire
Partager les (poids des) modèles ☺
Conteneuriser pour figer les dépendances
Utiliser des gestionnaires d'environnement

### Le hasard et le matériel
Fixer les **seeds aléatoires** et les documenter
Décrire le matériel utilisé (GPU, mémoire)
Rapporter une **moyenne sur plusieurs runs**

### Le prompt et le contexte
Publier les prompts exacts utilisés, mot pour mot
Documenter tous les paramètres de génération (température…)
Utiliser des **versions figées de modèles**
Préférer des modèles **open source**

### Les pratiques de publication
Suivre les **Checklist** imposée par les conférences
Soumettre le code + données avec le papier
Encourager les **reproduction papers** : des publications dont le but est de reproduire un résultat existant
Éviter le cherry-picking en enregistrant **toutes** les expériences

# IA et reproductibilité : l'IA comme un outil pour la reproductibilité

## Documentation automatique

extraction des étapes d'un pipeline (code)
génération de descriptions textuelles d'un workflow
identification des entrées, sorties, paramètres et dépendances…

## Détection des informations manquantes

version des logiciels
paramètres
prétraitements
provenance des données

## Génération de métadonnées FAIR

suggestion des métadonnées
Vers lien auto données, logiciels et publications

## Assistance à la reproduction ?

convertir un workflow d'un système à un autre
générer des scripts d'installation
détecter les incompatibilités d'environnement

# LES RESEAUX

## Global Networks

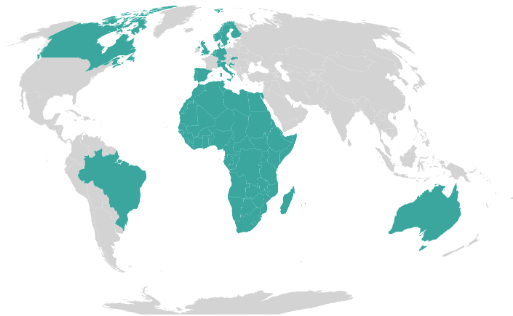
Outside the UK? Find a Reproducibility Network in your area

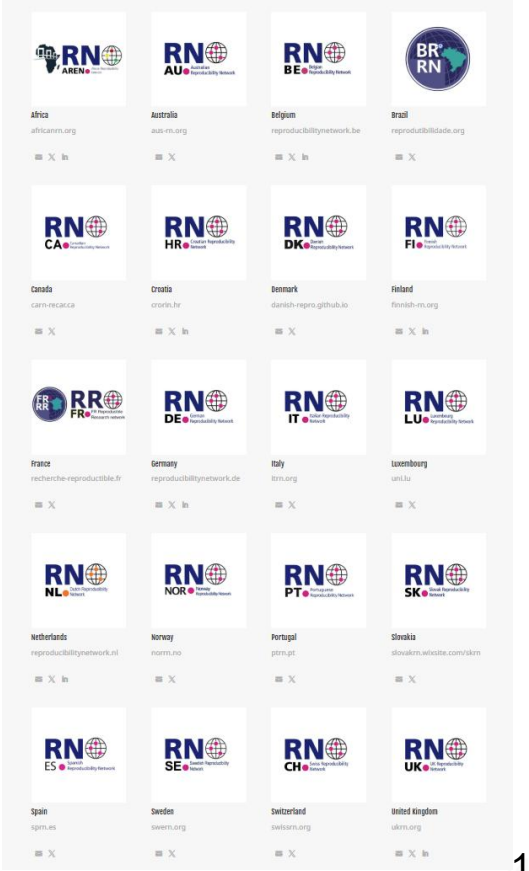
See full Global Networks Statement

### Global Reproducibility Networks

A Reproducibility Network (RN) is a national, peer-led consortium of researchers that aims to promote and ensure rigorous research practices by establishing appropriate training activities, designing and evaluating research improvement efforts, disseminating best practice and working with stakeholders to coordinate efforts across the sector. RNs aim for broad disciplinary representation and an intensive interdisciplinary dialogue (e.g., with funding agencies, publishers, learned societies and other sectoral organisations, as well as researchers from all disciplines and across all career stages).

To reach as many researchers as possible, and to operate as efficiently as possible, we are keen to support other countries interested in creating similar networks. If you are interested in setting up a national RN, or finding out who in your country is working towards this, please email: contact@ukrn.org.

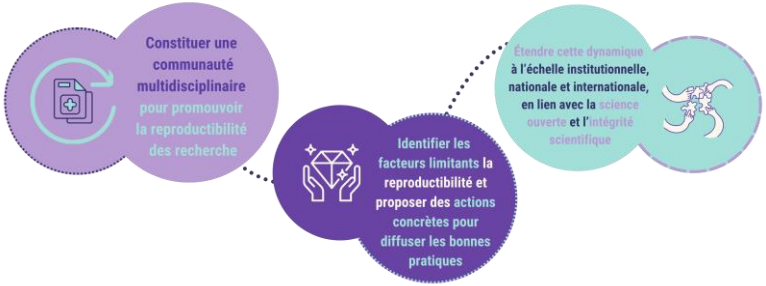




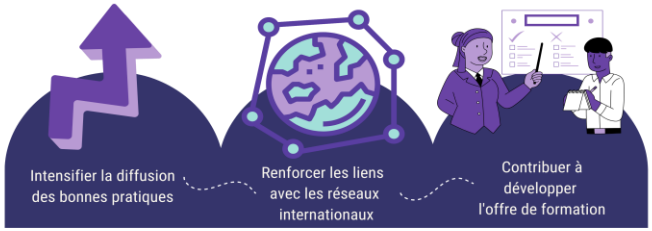
# Le Réseau Français Recherche Reproductible

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
*Liberté*
*Égalité*
*Fraternité*

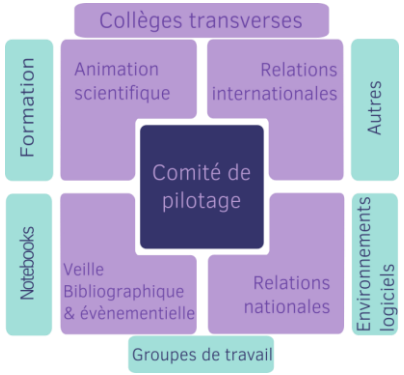
## Objectifs




## Actions



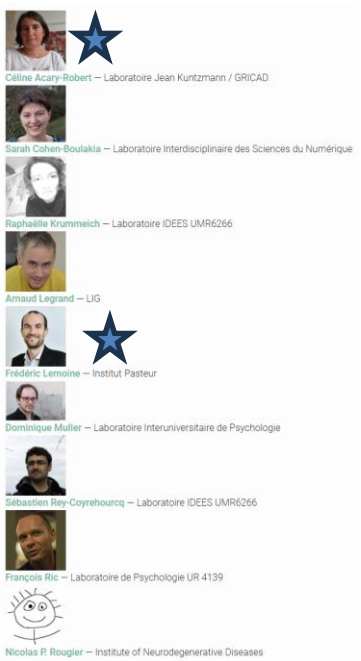

## Structuration




Notre réseau c'est **300+ membres - 30 disciplines** : Informatique 20% - Bioinfo 10% - Physique 8% - Neurosciences 8%....

FR RR FR Reproducible Research network

## Comité de pilotage




https://groupes.renater.fr/sympa/info/recherche-reproductible

# La vie du réseau – Actions en cours et à venir

## https://www.recherche-reproductible.fr

### Wébinaires

Dec 12, 2025 The African Reproducibility Network: Building Grassroots Capacity in Africa

Oct 10, 2025 The Brazilian Reproducibility Network: past, present, and future.

### Conférences

Oct 3, 2025 Replication Games - Paris

SCIENCES • RECHERCHE SCIENTIFIQUE

**Les « jeux de la réplication », ou quand la science s'amuse à se reproduire**

Une vingtaine de chercheurs ont traqué, le 3 octobre, les failles de plusieurs articles de sciences sociales parus dans des revues de renom.

Le Monde

## Nombreux évènements sur le site web

**ANF "Workflows et reproductibilité en bioinformatique", Paris, 25 au 27 novembre 2025.**

Le réseau MétiER en bioinformatique **MERIT** et l'Institut Français de Bioinformatique (**IFB**) organisent une Action Nationale de Formation (ANF) sur les principes FAIR appliqués aux workflows bioinformatiques. Cette formation, portée par le CNRS (INS2I), vise à former des bioinformaticiens sur les bonnes pratiques de reproductibilité. Les participants acquièrent des compétences avancées en gestion de workflows (Nextflow et Snakemake), et maîtrisent les outils de dépôt (SWH, HAL) et d'intégration continue pour garantir la qualité de leur code. Le réseau français de la recherche reproductible est ainsi impliqué, renforçant l'adoption de méthodes rigoureuses en bioinformatique.

MINISTÈRE
DE L'ENSEIGNEMENT
SUPÉRIEUR
ET DE LA RECHERCHE
*Liberté*
*Égalité*
*Fraternité*

# Rejoignez-nous !

Collez des **affiches** de promotion du réseau autour de vous !

Impliquez-vous dans le **collège Animation ou Relations Européennes ou Formation** ☺