

**CT IA, Licence 2
Portail SV
Chapitre 4 : L'IA pour les sciences de la vie**

Pr. Vincent Vandewalle

Professeur des Universités en Mathématique Appliquée, UniCA

Directeur scientifique adjoint d'EFELIA Côte d'Azur

Image by Alan Warburton / © BBC / Better Images of AI / Nature / CC BY 4.0

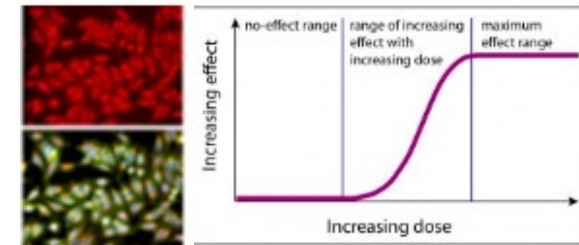
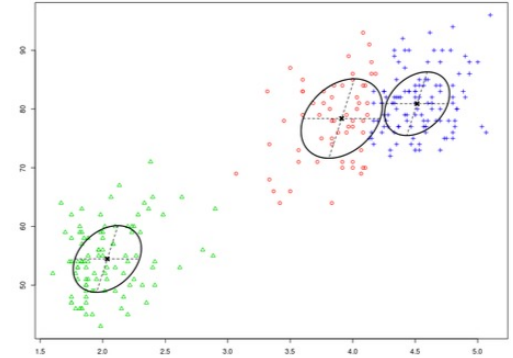
Parcours initial en sciences de la vie

- Classe prépa BCPST
- Ecole d'ingénieur AgroParisTech :
 - Cours en agronomie et biotechnologie
 - Suivi de modules en statistique, bio-informatique : Modélisation du hasard en biologie, neurones et modèles, bio-puces et bio-informatique
 - M2 en probabilités et statistiques
- 2009 : Thèse en statistique « Estimation et sélection en classification semi-supervisée »
- 2010-2022 : MCF à l'Université de Lille avec développement de méthodes statistiques pour différents types de données, et application à divers domaines dont la biologie
- 2022-maintenant : PR à l'Université Côte d'Azur, même thématiques de recherche, directeur scientifique adjoint du projet EFELIA Côte d'Azur



Ma recherche

- Développement de modèles statistiques pour le *machine learning* : modèles pour le clustering (classification non supervisée)
- Développement de modèles pour des applications spécifiques :
 - Développement de méthodes pour la détection d'effets de substances chimiques sur la morphologie cellulaire, thèse 3IA Côte d'Azur, Mariam Grigoryan (2023-aujourd'hui), en collaboration avec Bayer
 - Clustering pour la détection de famille d'astéroïdes, thèse 3IA Côte d'Azur, Maya Guy (2025-aujourd'hui), co-encadrée avec l'OCA



Objectifs du chapitre

A la fin de ce chapitre :

- Vous connaîtrez les principales méthodes d'IA (statistique/machine-learning/deep-learning) utilisées dans les sciences de la vie
- Vous serez capable de citer des exemples d'application en biologie moléculaire, développement de thérapies, études et protection de la biodiversité, agriculture
- Vous pourrez discuter de la pertinence de mettre en place une approche d'IA selon la problématique rencontrée
- Vous pourrez discuter des limites et enjeux éthiques, légaux et sociétaux de la mises en place de telles approches



Vocabulaire et rappels fondamentaux

- Dans ce chapitre nous focaliseront essentiellement sur les **approches statistiques de l'IA** aussi appelées **machine learning/apprentissage automatique**
- Ces approches font usage de **données** pour déterminer les **paramètres** d'un **modèle mathématique**
- Les paramètres sont déterminés de manière à minimiser un écart entre les valeurs observées dans les données et les prédictions faites par le modèle
- Cette phase de **détermination des paramètres** du modèle est appelée phase d'**apprentissage** ou d'**entraînement**
- Il existe différents modèles mathématiques selon la tâche qu'on souhaite résoudre :
 - S'il s'agit de prédire une réponse y à partir de valeurs x en entrée sur la base d'exemples fournis sous la forme $\{(x_1, y_1), \dots, (x_n, y_n)\}$ on parle d'**apprentissage supervisé** :
 - S'il s'agit de prédire un **nombre réel** on parle de **régression**
 - S'il s'agit de prédire la modalité d'une **variable qualitative** on parle de **classification supervisée**
 - S'il s'agit d'organiser les données en **groupes homogènes** sur la base de données de la forme $\{x_1, \dots, x_n\}$ on parle de **classification non-supervisée** ou **clustering**
 - S'il s'agit de **générer une nouvelle** valeur x_{n+1} sur la base d'exemples antérieurs $\{x_1, \dots, x_n\}$ on parle de **modèles génératifs**
- Pour résoudre chacune de ces tâches différentes fonctions mathématiques peuvent être envisagées, le type de fonctions qui a connu le plus de succès au cours des 15 dernières années sont les **réseaux de neurones profonds**, il s'agit de modèles comportant un très grand nombre de paramètres dont l'apprentissage a été rendu possible à l'aide de la disponibilité de grands volumes de données et de l'augmentation des capacités de calcul.



IA et statistique

- La plupart des systèmes d'IA actuellement utilisés s'appuient sur des données pour déterminer des paramètres en ce sens ce sont des approches statistiques
- A ce titre elle permettent de détecter des corrélations dans les données, mais elle ne permettent en général pas de déterminer explicitement une relation de cause à effet
- Par exemple si à partir de données d'expression de gène on voit que certains gène sont particulièrement associés à l'apparition d'une maladie cela ne permet pas de conclure que ces gènes sont responsables de la maladie et seules des investigations biologiques plus avancées permettent éventuellement de conclure, l'approche statistique servant uniquement à pointer des candidats potentiels pour une investigation approfondie



Remarques sur l'utilisation des système d'IA en science du vivant

- Les sciences du vivant n'échappent pas à l'augmentation des volumes de données disponibles, notamment en biologie moléculaire depuis le séquençage du génome à haut débit, mais aussi dans domaines comme la protection de la biodiversité ou l'agriculture où les approches telles que la reconnaissance d'image jouent un rôle important
- Comme dans tout les domaines la mise en place de système d'IA en science de la vie pose un certains nombre de questions, par exemple si vous souhaitez automatiser une tâche à l'aide d'un approche de *machine learning* vous devrez vous poser les questions :
 - Intérêt et caractère éthique de la tâche je vous souhaitez automatiser
 - Représentativité des données utilisées pour l'apprentissage par rapport à la population sur laquelle la tâches doit être effectuée
 - Aspects légaux relatifs aux données recueillies
 - ...



Plan du chapitre

1. IA et biologie moléculaire
2. IA dans les thérapies
3. IA et biodiversité
4. IA et agriculture
5. Synthèse



Plan du chapitre

- 
1. IA et biologie moléculaire
 2. IA dans les thérapies
 3. IA et biodiversité
 4. IA et agriculture
 5. Synthèse



Quelques exemples d'application l'IA en biologie moléculaire

- Prédiction de structures protéiques (ex. AlphaFold)
- Analyse de séquences génomiques et annotation fonctionnelle
- Découverte de médicaments assistée par IA
- Diagnostic biomoléculaire basé sur le deep learning
- Modélisation des interactions protéine–protéine

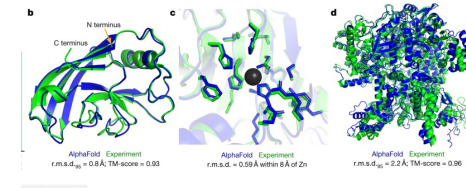


Focus sur l'Algorithme AlphaFold¹

Prédiction de la structure 3D des protéines à partir de leur séquences d'acides aminés :

- Données d'entraînement : les structures protéiques connues dans la Protein Data Bank (PDB), de vastes banques de séquences pour construire des alignements multiples de séquences (MSA)
- Dualité MSA <-> Matrice de distances entre paires de résidus :
 - Représentation MSA : Capture l'information évolutive (co-variations entre résidus), modélisée via une architecture de réseau de neurone spécialisée (transformer).
 - Matrice de distances entre paires (2D) : Matrice décrivant les relations géométriques entre paires de résidus. Encode distances, orientations, angles potentiels.
 - Les deux représentations s'alimentent mutuellement à travers un module central appelé Evoformer.
- Evoformer applique des blocs de transformers adaptés :
 - Attentions sur les MSA,
 - Attentions sur les paires de résidus,
 - Echanges bidirectionnels d'information MSA ↔ distances entre paires
 - Permet d'apprendre à déduire les interactions entre résidus de la chaînes d'acides aminés et de prédire comment la chaîne doit se plier.
- Prédiction de la structure 3D :
 - Les coordonnées 3D des atomes du squelette protéique,
 - En respectant des contraintes géométriques différentiables
- Optimisation end-to-end :
 - Les paramètres sont optimisés pour minimiser un écart géométrique entre la structure prédite et la structure réelle.
 - Tout le pipeline est différentiable, du MSA jusqu'aux coordonnées 3D.
 - Cette approche permet d'apprendre directement la physique implicite du repliement sans l'encoder explicitement.
 - Approche itérative : raffine progressivement les positions atomiques.

Attention les auteurs mentionnent que des écarts peuvent subsister avec la réalité et que la détermination expérimentale de la structure 3D reste importante.



Jumper, J. et al. (2021). *Highly accurate protein structure prediction with AlphaFold*. **Nature**, 596, 583–589.

Développement de modèle fondation pour la biologie qui permettent la prédiction et la génération du niveau moléculaire jusqu'à l'échelle du génome avec le modèle Evo¹

Jeux de données OpenGenome avec les génomes de 80,000 bactéries et archées et des millions de phages procaryotiques et de séquences de plasmides.

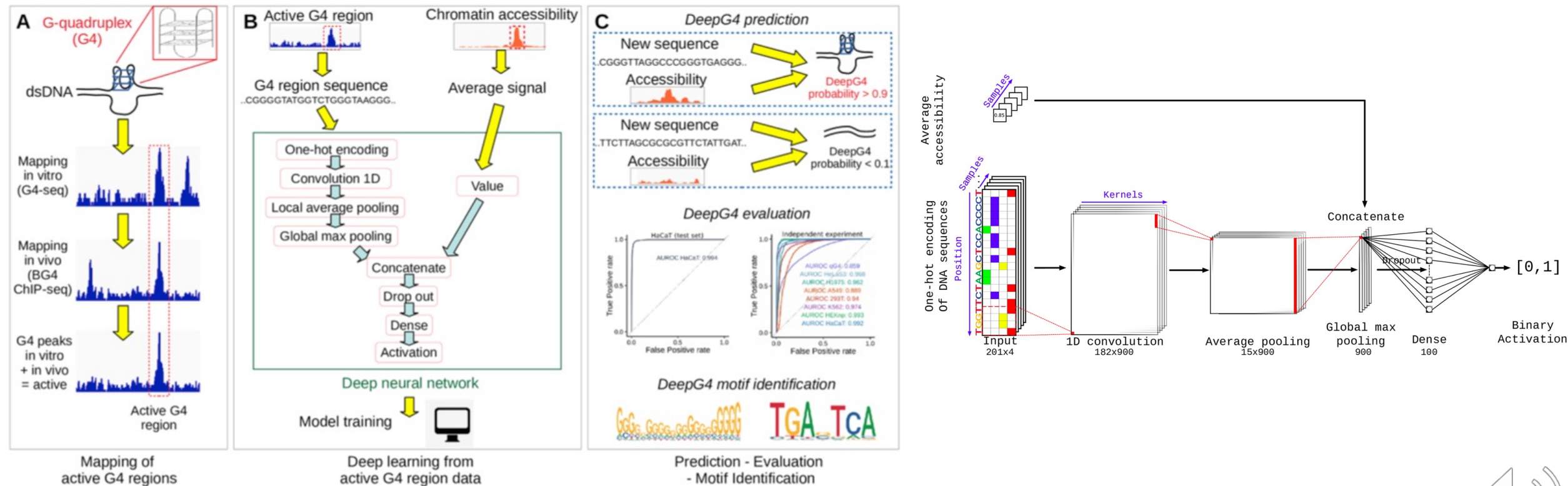
Modèle qui permet à l'aide d'un réseau de neurone commun aux différentes tâches de :

- Prédire les conséquences d'une mutation sur la fonction d'une protéine
- Prédiction des effets des mutations sur les ARN non codants
- Prédiction de l'expression des gènes à partir de l'ADN régulateur



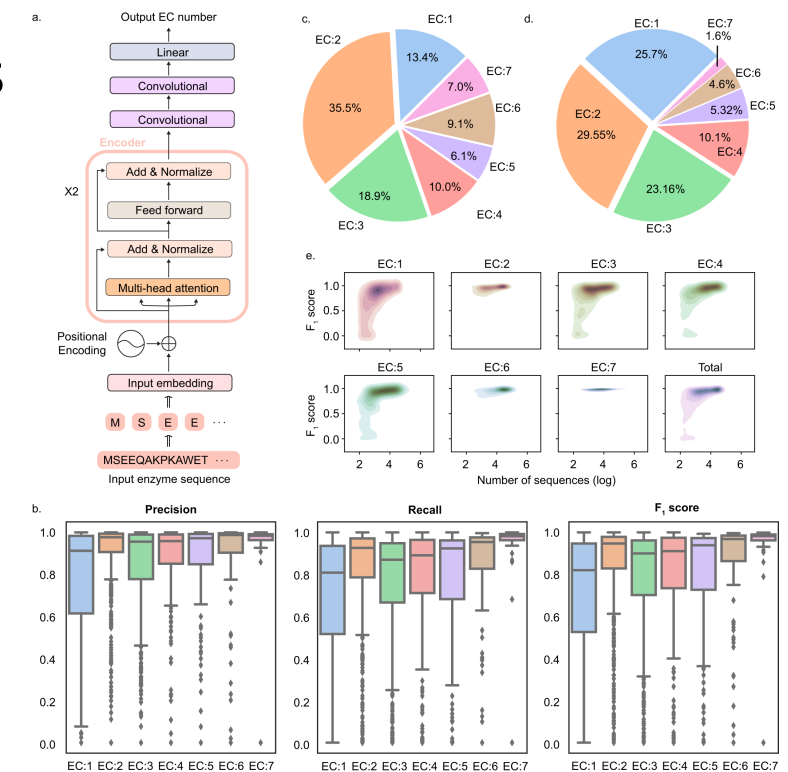
Détection de structures secondaires dans l'ADN

- La formation de structures secondaires type G-quadruplexes (ou G4) peut entraîner l'instabilité du génome en créant des mutations
- Les auteurs de ¹ démontrent qu'il est possible de prédire avec précision les régions du génome où se forment les G4 dans les lignées cellulaires, tissus et cancers.



Autres exemples d'applications

1. Prédiction de l'épissage à partir de la séquence primaire grâce à l'apprentissage profond¹, prédit avec précision les variants génétiques non codants qui provoquent un épissage cryptique
2. Prédiction de l'influence de mutations de la bactérie *Mycobacterium tuberculosis* (*M.tb*) pour la résistance à des médicaments en fonction de mutations sur certains gènes², application d'approches de machine learning (algorithmes classiques + algorithmes de deep learning)
3. Prédiction de la classe fonctionnelle des enzymes à partir d'une architecture comprenant des couches transformers³



Kim et al., "Functional Annotation of Enzyme-Encoding Genes Using Deep Learning with Transformer Layers," Nat Commun, 2023, 14(1):7370.

1. Jaganathan et al., "Predicting Splicing from Primary Sequence with Deep Learning," Cell, 2019, 176(3):535–548.e24.
2. Jamal et al., "Artificial Intelligence and Machine Learning Based Prediction of Resistant and Susceptible Mutations in Mycobacterium tuberculosis," Sci Rep, 2020, 10(1):5487.
3. Kim et al., "Functional Annotation of Enzyme-Encoding Genes Using Deep Learning with Transformer Layers," Nat Commun, 2023, 14(1):7370.

IA et biologie moléculaire : opportunités et enjeux éthiques

- Avancées majeures
 - Prédiction de structures protéiques
 - Analyse de mutations et diagnostics améliorés
 - Modélisation de réseaux géniques complexes
- Enjeux éthiques
 - Données génétiques : confidentialité & risque de ré-identification
 - Biais dans les modèles : inégalités en santé
 - Capacité à modifier ou générer du matériel génétique
 - Risques d'usage détourné (biosécurité)
- Besoin de gouvernance
 - Transparence des modèles
 - Cadres réglementaires adaptés
 - Surveillance et responsabilité partagée
 - Usage responsable au bénéfice de tous



Plan du chapitre

1. IA et biologie moléculaire
- ➔ 2. IA dans les thérapies
3. IA et biodiversité
4. IA et agriculture
5. Synthèse



Systèmes d'IA en thérapie : opportunités et enjeux éthiques

Apports majeurs

- Analyse massive des données cliniques et biologiques
- Prédiction de l'efficacité thérapeutique
- Personnalisation des traitements
- Accélération de la découverte de médicaments
- Gains d'efficacité en imagerie médicale

Risques et défis

- Confidentialité et sécurité des données
- Consentement éclairé difficile à garantir
- Biais algorithmiques → inégalités de santé
- Opacité des modèles & manque d'explicabilité
- Déploiement clinique encore hétérogène

Santé mentale : potentiel & prudence

- Agents conversationnels : soutien, dépistage, monitoring
- Réductions observées des symptômes anxieux et dépressifs
- Usage complémentaire, non substitutif de la relation thérapeutique

Conditions pour un usage responsable

- Gouvernance et supervision humaine
- Validation clinique rigoureuse
- Transparence des modèles & gestion éthique des données
- Formation des professionnels de santé
- Approche centrée patient & équité

1. Younis HA, et al. A Systematic Review and Meta-Analysis of Artificial Intelligence Tools in Medicine and Healthcare: Applications, Considerations, Limitations, Motivation and Challenges. *Diagnostics* (Basel). 2024 Jan 4;14(1):109
2. Wenderott K, et al. Effects of artificial intelligence implementation on efficiency in medical imaging-a systematic literature review and meta-analysis. *NPJ Digit Med*. 2024 Sep 30;7(1):265
3. Li H, et al. Systematic review and meta-analysis of AI-based conversational agents for promoting mental health and well-being. *NPJ Digit Med*. 2023 Dec 19;6(1):236.
4. Berkhout WEM, et al. Operationalization of Artificial Intelligence Applications in the Intensive Care Unit: A Systematic Review. *JAMA Netw Open*. 2025 Jul 1;8(7):e2522866.
5. Morone G, De Angelis L, Martino Cinnera A, Carbonetti R, Bisirri A, Ciancarelli I, Iosa M, Negrini S, Kiekens C, Negrini F. Artificial intelligence in clinical medicine: a state-of-the-art overview of systematic reviews with methodological recommendations for improved reporting. *Front Digit Health*. 2025 Mar 5;7:1550731.



Intelligence Artificielle en hématologie¹

1. Explosion des données en hématologie moléculaire

- Les technologies haut débit (NGS, transcriptomique, génomique) génèrent un volume massif de données difficile à analyser manuellement.

2. L'intelligence artificielle comme outil indispensable

- IA et machine learning deviennent essentiels pour exploiter ces données, notamment pour la prédiction de pathogénicité, le diagnostic, le pronostic et l'identification de cibles thérapeutiques.

3. Applications majeures du ML

- Pathogénicité des variants : modèles comme EVE ou BoostDM identifient mutations driver vs passager.
- Pronostic : modèles intégrant données multi-omiques (ex. LymForest-25, signatures génomiques dans LAM).
- Thérapeutique : algorithmes d'assignation numérique des médicaments (DDA) qui prédisent la réponse aux thérapies ciblées.

4. Diagnostic et classification

- Random Forest et réseaux neuronaux permettent de classer les hémopathies (lymphomes, LAM...) via l'expression génique, avec une précision proche de celle des experts.

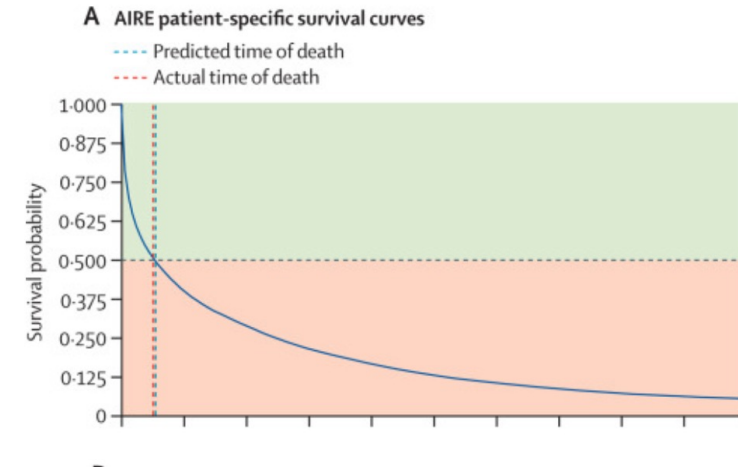
5. Limites et risques

- Boîte noire : difficulté à interpréter les inférences des modèles.
- Risque de dépendre excessivement des algorithmes.
- Les système d'IA doivent rester un outil d'aide, non un substitut au jugement clinique.



Modèle de prédiction de la survie à partir de données d'ECG¹

- Prédiction de la survie à partir de données d'ECG : $P(\text{Survie} > t \mid \text{ECG})$
- 1 163 401 ECGs provenant de 189 539 patients + date éventuelle d'occurrence d'un événement pour les patients
- Utilisation d'un modèle de deep-learning pour ajuster cette probabilité
- Prédiction de différents type d'événements :
 - Mort toutes causes confondues
 - Arthrosclérose cardiovasculaire
 - Crise cardiaque
- Permet d'identifier niveau de risque et d'adapter la prise en charge au niveau de risque estimé



1. Sau et al., "Artificial Intelligence-Enabled Electrocardiogram for Mortality and Cardiovascular Risk Estimation: A Model Development and Validation Study," Lancet Digit Health, 2024, 6(11):e791–e802.

Plan du chapitre

1. IA et biologie moléculaire
2. IA dans les thérapies
- ➔ 3. IA et biodiversité
4. IA et agriculture
5. Synthèse



Exemples d'applications¹

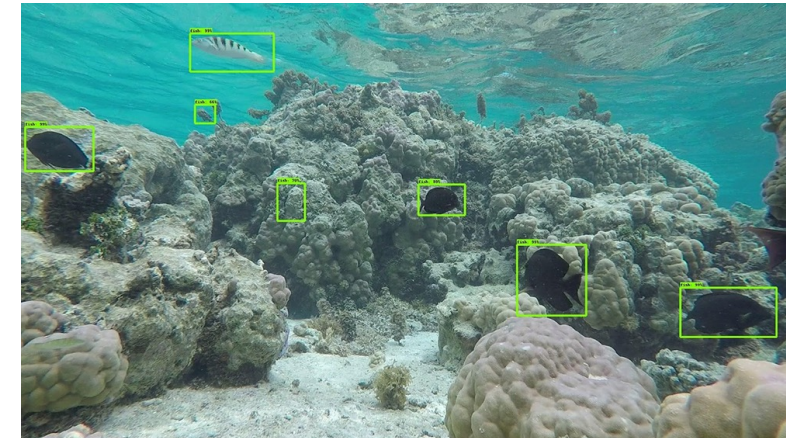
- **Suivi d'espèces en temps réel** grâce à l'analyse automatisée de données massives.
- **Prévention de la déforestation** via des capteurs mesurant humidité, carbone ou stress hydrique permettant d'anticiper sécheresses et feux de forêt.
- **Lutte contre le braconnage**, par la détection automatique de présences humaines dans des zones protégées et l'envoi d'alertes aux gardes.
- **Préservation de la biodiversité marine**, au cœur de plusieurs initiatives comme le *Challenge IA-BIODIV*.
- **Protection des éléphants**, grâce à des systèmes de surveillance aérienne capables de cartographier en quelques heures plus de 5 000 km² de savanes, prairies ou forêts et d'identifier automatiquement les populations.

1. <https://idverde.fr/actualites/proteger-la-biodiversite-grace-a-lintelligence-artificielle/>



Projet FISH-PREDICT

- Améliorer les modèles de distribution des espèces et suivi d'indicateurs de biodiversité à l'aide d'approches d'IA (projet FISH-PREDICT¹) :
 - Analyser la variation spatiale des paramètres environnementaux
 - Meilleures explication de la distribution des espèces
 - Capturer des interactions complexes (influence des bateaux de plaisance, l'impact de l'aménagement de la côte, ...)
- Protéger la biodiversité grâce à l'IA



Légende : Exemple de détection automatique de spécimens de poissons dans des vidéos sous-marines.

© Valentine Fleuré

1. <https://lejournel.cnrs.fr/nos-blogs/un-ocean-de-decouvertes/des-modeles-dapprentissage-profond-pour-comprendre-et-preserver-la>



Surveillance automatisée des aires marines protégées¹

Fonctionnement

- Véhicules autonomes de surface pour la surveillance en continu
- Détection et reconnaissance automatique des navires

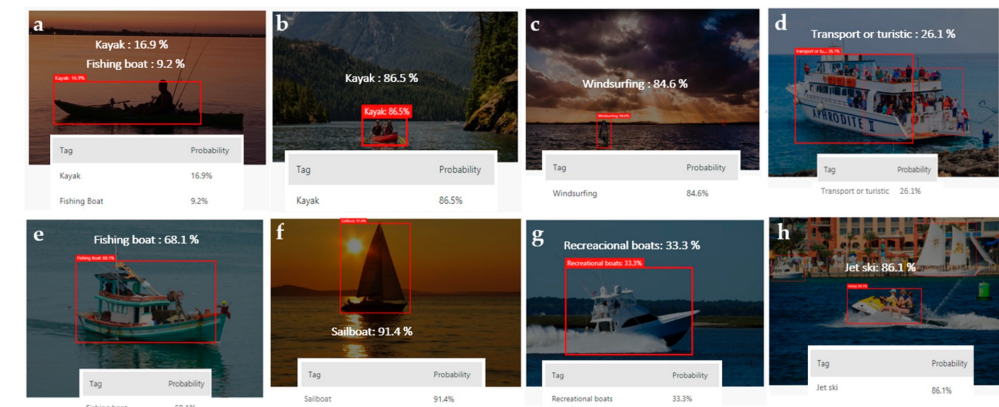
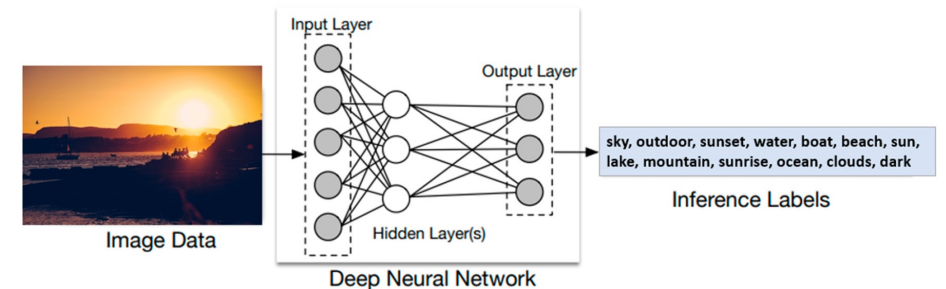
Défis techniques

- Capacités de calcul embarquées limitées
- Besoin fréquent de recourir à des services en ligne pour exécuter les modèles d'IA
- Sélection automatique du modèle d'IA selon le scénario :
 - **Modèles légers** → exécution embarquée
 - **Modèles complexes** → exécution via réseau

Mise en place de l'approche

- Données d'apprentissage étiquetées avec un cadre indiquant l'objet
- Entraînement d'un réseau de neurone profond
- Prédiction sur de nouvelles données

Dataset	Recreational boat	Fishing boat	Jet ski	Kayak	Transport and tourism	Sailboat	Windsurfing
							
Trained Photos	65	104	54	156	50	121	50

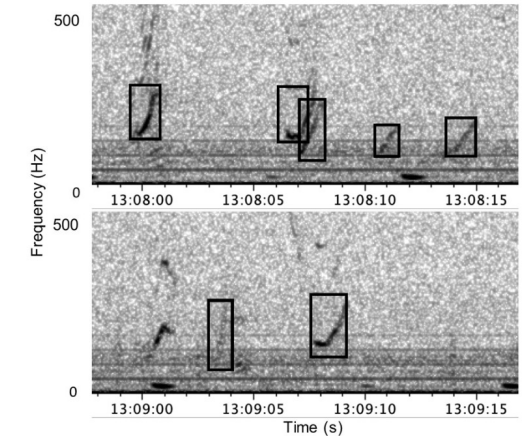


1. Autonomous Marine Robot Based on AI Recognition for Permanent Surveillance in Marine Protected Areas. Molina-Molina JC, Salhaoui M, Guerrero-González A, Arioua M. Sensors (Basel). 2021 Apr 10;21(8):2664

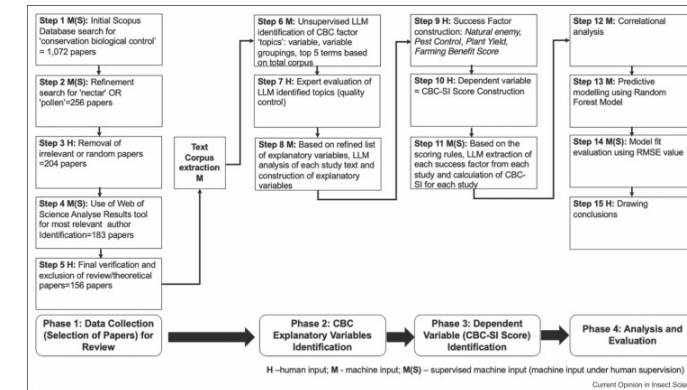


Autres exemples d'applications

- Détection des vocalises de baleines blanches en risque d'extinction¹
- Analyse de la pollution aux microplastiques à l'aide de techniques d'imagerie²
- Utilisation pour déterminer le succès d'une approche de lutte biologique³ :
 - Objectif déterminer les facteurs qui déterminent le succès d'une lutte biologique à l'aide d'information extraites dans 147 articles à l'aide d'un LLM
 - Informations extraites utilisées comme variables explicatives dans un modèle statistique
 - Attention les auteurs soulignent que l'extraction d'information peut être sujette à des erreurs et donc qu'il faut interpréter les résultats avec prudence



1. Deep neural networks for automated detection of marine mammal species. Shiu Y, Palmer KJ, Roch MA, Fleishman E, Liu X, Nosal EM, Helble T, Cholewiak D, Gillespie D, Klinck H. Sci Rep. 2020 Jan 17;10(1):607



Harnessing artificial intelligence for analysing the impacts of nectar and pollen feeding in conservation biological control. Gurr GM, Liu J, Pogrebna G. Curr Opin Insect Sci. 2024 Apr;62:101176.

1. Deep neural networks for automated detection of marine mammal species. Shiu Y, Palmer KJ, Roch MA, Fleishman E, Liu X, Nosal EM, Helble T, Cholewiak D, Gillespie D, Klinck H. Sci Rep. 2020 Jan 17;10(1):607
2. A Critical Review on Artificial Intelligence-Based Microplastics Imaging Technology: Recent Advances, Hot-Spots and Challenges. Zhang Y, Zhang D, Zhang Z. Int J Environ Res Public Health. 2023 Jan 9;20(2):1150.
3. Harnessing artificial intelligence for analysing the impacts of nectar and pollen feeding in conservation biological control. Gurr GM, Liu J, Pogrebna G. Curr Opin Insect Sci. 2024 Apr;62:101176.



Plan du chapitre

1. IA et biologie moléculaire
2. IA dans les thérapies
3. IA et biodiversité
-  4. IA et agriculture
5. Synthèse



Quelques exemples d'utilisation de systèmes d'IA dans le domaine de l'agriculture

Surveillance des cultures et des élevages :

- Analyse des données satellitaires
- Détection de maladie ou de parasite
- Surveillance intelligente

Optimisation de la production :

- Prédiction des rendements
- Gestion de l'irrigation

Automatisation et robotisation :

- Cueillette robotisée des fruits
- Désherbage automatisé

Gestion de la chaîne logistique :

- Optimisation de stock



Exemples issus d'articles récents

- L'intelligence artificielle (IA) est-elle devenue indispensable en agriculture ?¹ : applications à bases d'images, applications axées sur le traitement de données massives
- Détection de maladies et de nuisibles dans les cultures²
- Optimisation des récoltes à partir de modèles prédictifs de la maturité des fruits^{3,4}



1. <https://hal.science/hal-05036786/document#>, Académie d'agriculture de France
2. Intelligent agriculture: deep learning in UAV-based remote sensing imagery for crop diseases and pests detection [Academic Journal] Hongyan Zhu | Chengzhi Lin | Gengqi Liu | Dani Wang | Shuai Qin Frontiers in Plant Science, Vol 15 (2024)
3. AI-driven time series analysis for predicting strawberry weekly yields integrating fruit monitoring and weather data for optimized harvest planning [Academic Journal] Liu, Shiyu | Ampatzidis, Yiannis | Zhou, Congliang | Lee, Won Suk In Computers and Electronics in Agriculture June 2025 233
4. Zero-shot transfer learned generic AI models for prediction of optimally ripe climacteric fruits. [Academic Journal] Dutta, Jayita | Patwardhan, Manasi | Deshpande, Parijat | Karande, Shirish | Rai, Beena Scientific Reports 27 5/5/2023, Vol. 13 Issue 1, p1-11. 11p. DOI : 10.1038/s41598-023-34527-8

Plan du chapitre

1. IA et biologie moléculaire
2. IA dans les thérapies
3. IA et biodiversité
4. IA et agriculture
-  5. Synthèse



Synthèse

- Nombreuses applications des approches de machine learning dans le domaine des sciences de la vie
- Essentielles pour le domaine de la biologie moléculaire où elles viennent enrichir les méthodes de statistique et de bioinformatiques déjà existante
- Méthodes liées à l'analyse d'images qui a fait le succès des méthodes de deep learning très présentes
- Place croissante dans ces différents domaines
- Importance d'un usage éthique et légal associé aux différents usages des système d'IA

