*Option Santé & Vivant*

# Écueils de l'utilisation du ML dans le domaine biomédical

Chloé-Agathe Azencott

Center for Computational Biology (CBIO)
Mines ParisTech – Institut Curie – INSERM U1331
PSL Research University & PR[AI]RIE, Paris, France

Oct 14, 2025

http://cazencott.info chloe-agathe.azencott@minesparis.psl.eu @cazencott@lipn.info

# Objectif

- Le **machine learning** (deep learning, intelligence artificielle) est très à la mode
- Mais est-ce que **ça marche ?**
- Quelques **clés** permettant une **lecture critique**

# Moyens

- Des **exemples** tirés de ma propre **expérience**
- Quelques références :
  - S. Whalen et al (2021). **Navigating the pitfalls of applying machine learning in genomics**. Nat Rev Gen
  - Michael A. Lones (2024). **Avoiding common machine learning pitfalls**. Patterns

# Pourquoi c'est dur ?

## Succès du machine learning

- quand il y a **beaucoup de données**
  - ImageNet : 14 millions d'images
  - données d'entraînement de Llama4 : $30 \times 10^{12}$ tokens

# Pourquoi c'est dur ?

## Succès du machine learning

- quand il y a **beaucoup de données**
  - ImageNet : 14 millions d'images
  - données d'entraînement de Llama4 : $30 \times 10^{12}$ tokens
- quand on comprend bien la **nature des données**

  $\Rightarrow$ bonnes représentations & modèles

# Pourquoi c'est dur ?

## Succès du machine learning

- quand il y a **beaucoup de données**
  - ImageNet : 14 millions d'images
  - données d'entraînement de Llama4 : $30 \times 10^{12}$ tokens
- quand on comprend bien la **nature des données**

  $\Rightarrow$ bonnes représentations & modèles
- quand on comprend bien la **nature du problème**

  faisable par des êtres humains

# Pourquoi c'est dur ?

## Succès du machine learning

- quand il y a **beaucoup de données**
  - ImageNet : 14 millions d'images
  - données d'entraînement de Llama4 : $30 \times 10^{12}$ tokens
- quand on comprend bien la **nature des données**

  $\Rightarrow$ bonnes représentations & modèles
- quand on comprend bien la **nature du problème**

  faisable par des êtres humains
- pour **prédire** plutôt qu'**interpréter**

# À quoi faire attention ?

I. Avoir la bonne **métrique d'évaluation** pour l'application envisagée

II. Évaluer sur des **données de validation** appropriées

III. Comparer à de bonnes **baselines**

IV. Ce n'est pas parce que c'est du **deep learning** que c'est mieux

(V. Il y a loin du benchmarking à la **clinique**)

# I. Métriques d'évaluation

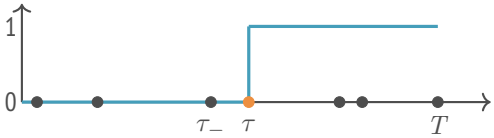
# Identification d'événements dans les parcours patient

- **But final** : utiliser le SNDS pour identifier des liens entre co-médications et **rechute**
- **Données** :
  - 200 000 patientes dans le SNDS (cancer du sein)
  - 5 892 patientes **annotées** : rechute/non, et si oui, date
- **But intermédiaire** : entraîner un modèle pour **annoter tous les parcours**

Dumas et al. (2022) **EDEN : An Event DEtection Network for the annotation of recurrences in administrative claims data**. ML4Health

# Identification d'événements dans les parcours patient

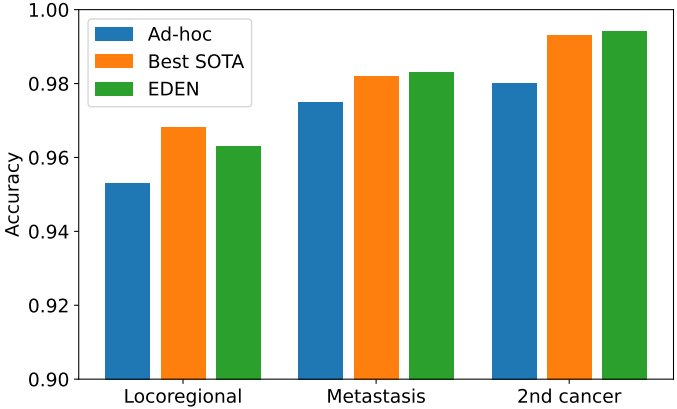
- **But intermédiaire** : entraîner un modèle pour **annoter tous les parcours**
- **Spéficités** : séquences **censurées** d'événements **irrégulièrement espacés dans le temps**
- **Idée** : modifier la fonction de coût d'un T-LSTM pour
  - modéliser le **taux de survie** $t \mapsto \mathbb{P}[t \geq \tau]$
  - adapter la **pénalisation** selon que l'observation est censurée ou non




Dumas et al. (2022) **EDEN : An Event DEtection Network for the annotation of recurrences in administrative claims data**. ML4Health

# Identification d'événements dans les parcours patient

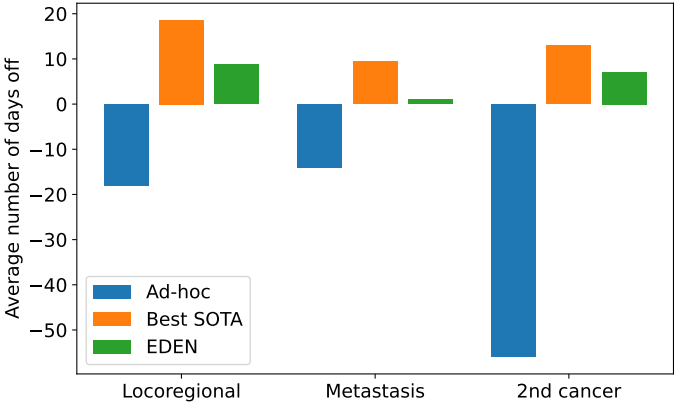
Évaluation sur un **jeu de données indépendant** de 800 parcours




Dumas et al. (2022) **EDEN : An Event DEtection Network for the annotation of recurrences in administrative claims data**. ML4Health

# Identification d'événements dans les parcours patient

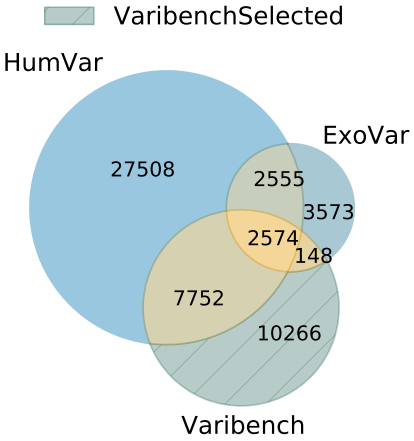
Évaluation sur un **jeu de données indépendant** de 800 parcours




Dumas et al. (2022) **EDEN : An Event DEtection Network for the annotation of recurrences in administrative claims data**. ML4Health

# II. Données d’évaluation

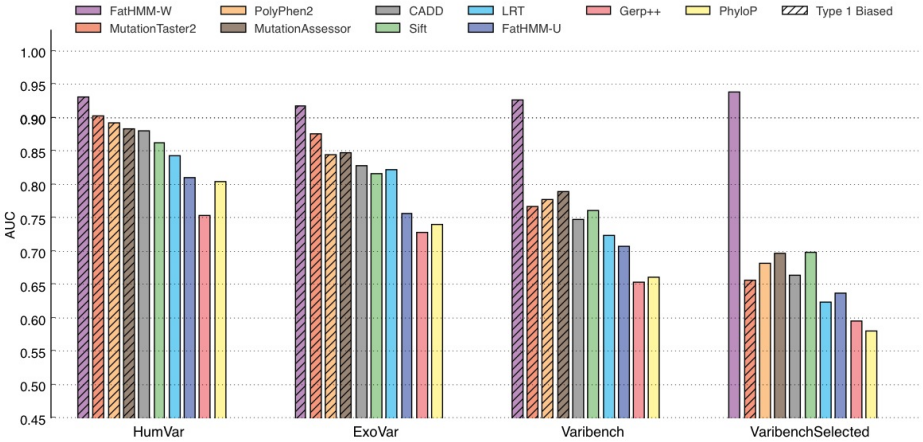
# 1. Prédiction de la pathogénicité de mutations




- De nombreux **outils** : PolyPhen2, GERP, CADD, PhyloP, MutationTaster, MutationAssessor, Sift, LRT, FatHMM
- Les **jeux de données** publics se **recouvrent** partiellement

Grimm et al. (2015). **The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity**. Hum Mut

# 1. Prédiction de la pathogénicité de mutations




Grimm et al. (2015). **The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity**. Hum Mut

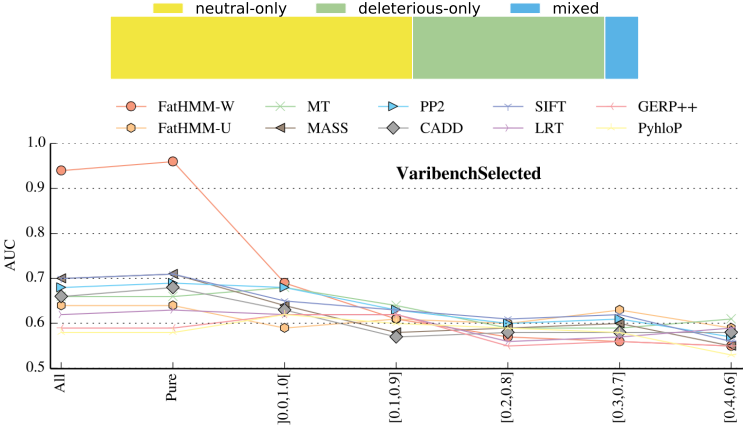
# 1. Prédiction de la pathogénicité de mutations

– Quasiment tous les variants **d'un même gène** ont la même annotation

Grimm et al. (2015). **The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity**. Hum Mut

# 1. Prédiction de la pathogénicité de mutations

– Quasiment tous les variants **d'un même gène** ont la même annotation

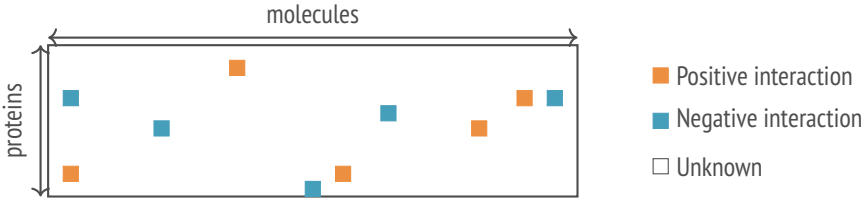



Grimm et al. (2015). **The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity**. Hum Mut

# 2. Prédiction d'interactions protéine-ligand

- **But**: prédire si une paire (ligand, protéine) interagit
  - pour une protéine et **toutes les molécules drug-like** : **découverte de médicaments, repositionnement**
  - pour une molécule et **toutes les protéines humaines**: **prédiction d'effets secondaires**, **méchanisme d'action**

- **LCIdb**: 271 180 molécules, 2 060 protéines, 0.07% paires étiquetées (98% positives)

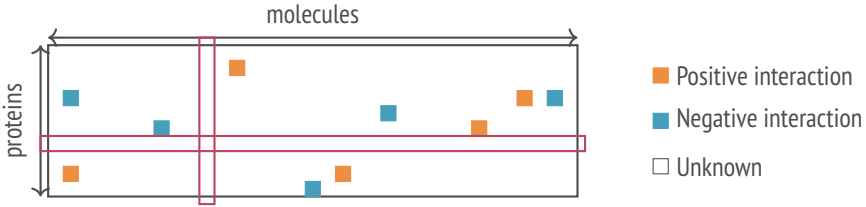



Guichaoua et al. (2024). **Drug-target interaction prediction at scale: the Komet algorithm with the LCIdb dataset**. J Chem Info Mod

# 2. Prédiction d'interactions protéine-ligand

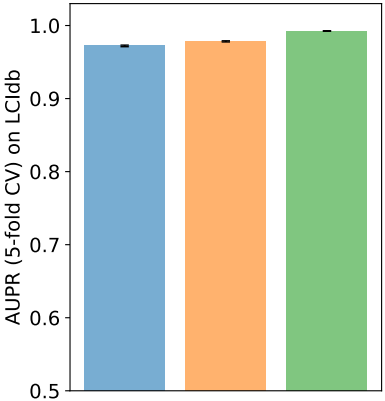
- **But**: prédire si une paire (ligand, protéine) interagit
  - pour une protéine et **toutes les molécules drug-like** : **découverte de médicaments, repositionnement**
  - pour une molécule et **toutes les protéines humaines**: **prédiction d'effets secondaires**, **méchanisme d'action**

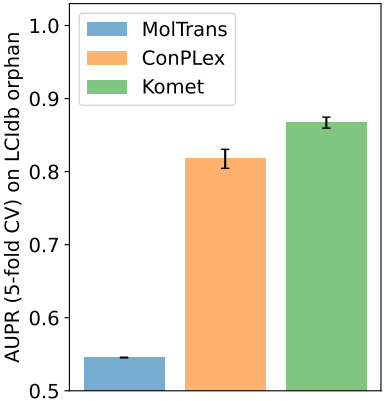
- **LCIdb**: 271 180 molécules, 2 060 protéines, 0.07% paires étiquetées (98% positives)




Guichaoua et al. (2024). **Drug-target interaction prediction at scale: the Komet algorithm with the LCIdb dataset**. J Chem Info Mod

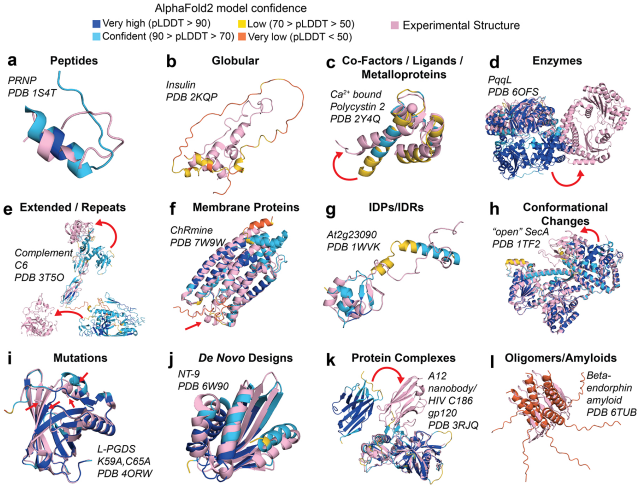
# 2. Prédiction d'interactions protéine-ligand




Guichaoua et al. (2024). **Drug-target interaction prediction at scale: the Komet algorithm with the LCIdb dataset**. J Chem Info Mod

# 3. AlphaFold2




Agrawal & McShan (2024). **The power and pitfalls of AlphaFold2 for structure prediction beyond rigid globular proteins.** Nat Chem Biol

# III. Points de comparaison

# 1. Prédiction de rechute à partir de parcours patient

- **But :** Prédire la rechute trois ans après une première chirurgie de cancer du sein
- **Données :** parcours **multi-modaux** pour 15 150 patientes
  - données tabulaires (biologie, données cliniques, traitement, service)
  - texte libre

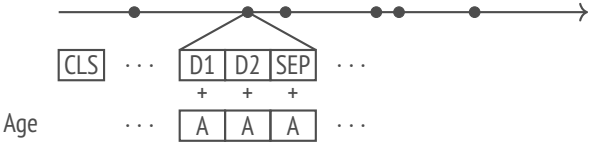
- **Nottingham Prognostic Index (NPI)** : $0.2T + G + N$

  $T$ = taille de la tumeur (cm); $G$ = grade de la tumeur ;
  $N$ = 1 if 0 ganglions lymphatiques envahis, 2 si 1 à 3 ganglions, 3 si +

Mbaye et al. (2025), **Multimodal BEHRT: transformers for multimodal Electronic Health Records**, Front Onc

# 1. Prédiction de rechute à partir de parcours patient

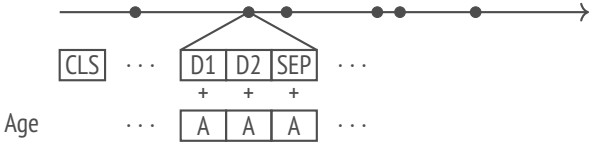
– **BEHRT** : classification de parcours patient **uni-modaux**



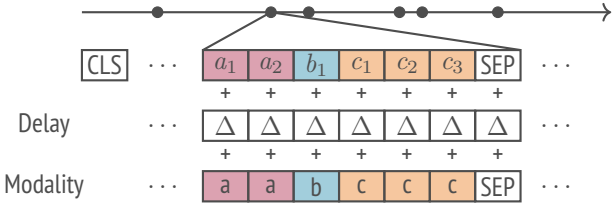

Mbaye et al. (2025), **Multimodal BEHRT: transformers for multimodal Electronic Health Records**, Front Onc

# 1. Prédiction de rechute à partir de parcours patient

– **BEHRT** : classification de parcours patient **uni-modaux**



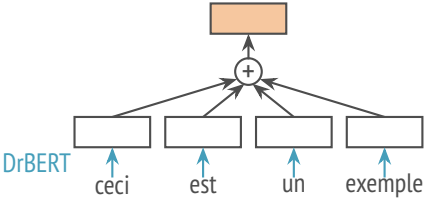

– **TabularBEHRT**




Mbaye et al. (2025), **Multimodal BEHRT: transformers for multimodal Electronic Health Records**, Front Onc

# 1. Prédiction de rechute à partir de parcours patient

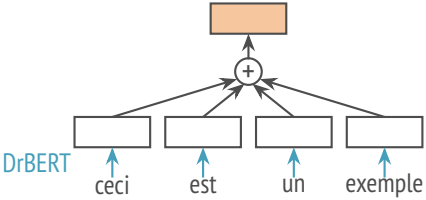
– **Représentation de rapports :** combiner les embeddings de mots par DrBERT



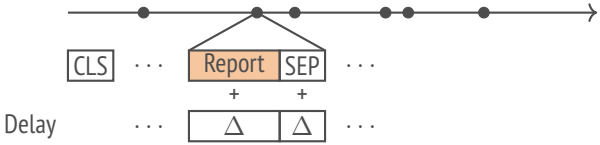

Mbaye et al. (2025), **Multimodal BEHRT: transformers for multimodal Electronic Health Records**, Front Onc

# 1. Prédiction de rechute à partir de parcours patient

– **Représentation de rapports :** combiner les embeddings de mots par DrBERT

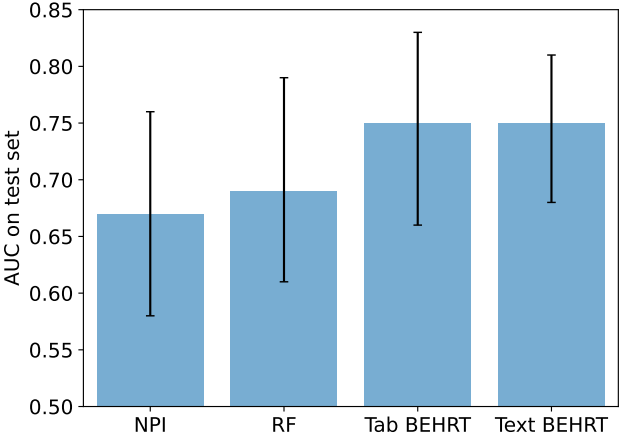



– **TextBEHRT:**




Mbaye et al. (2025), **Multimodal BEHRT: transformers for multimodal Electronic Health Records**, Front Onc

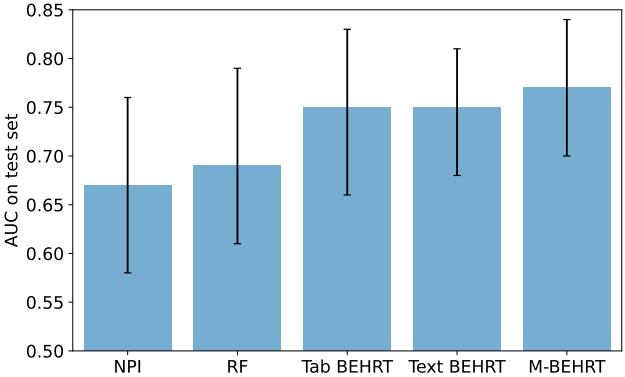
# 1. Prédiction de rechute à partir de parcours patient




Mbaye et al. (2025), **Multimodal BEHRT: transformers for multimodal Electronic Health Records**, Front Onc

# 1. Prédiction de rechute à partir de parcours patient

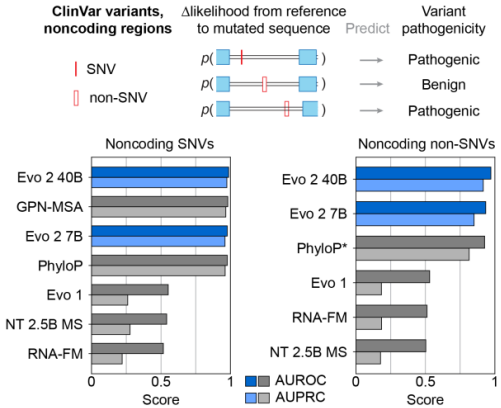
En **combinant** Tabular BEHRT and Text BEHRT par **cross-attention**




Mbaye et al. (2025), **Multimodal BEHRT: transformers for multimodal Electronic Health Records**, Front Onc

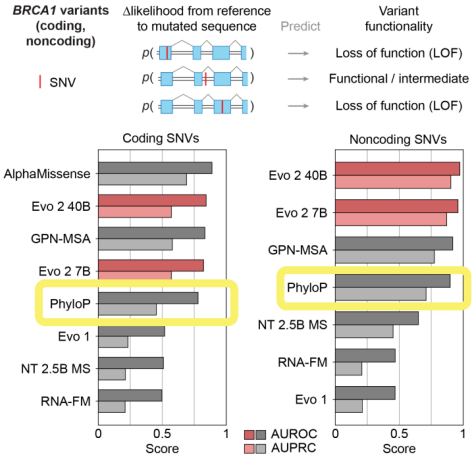
# 2. Prédiction de pathogénicité avec Evo2

- **Modèle de fondation** pour les séquences d'ADN
  - Entraîné sur presque 9 $\times 10^{12}$ paires de bases
  - Tâche prétexte : mascage
  - entre 7 et 40 $\times 10^{9}$ paramètres
  - 2.25 $\times 10^{24}$ FLOPS (similaire Llama 3.1)
- **Idée** : une séquence **improbable** est vraisemblablement **pathogène** $\rightarrow$**zero-shot** learning!




Brixi et al. (2025), Genome modeling and design across all domains of life with Evo 2, BioRxiv

# 2. Prédiction de pathogénicité avec Evo2




Brixi et al. (2025), Genome modeling and design across all domains of life with Evo 2, BioRxiv

# IV. Ce n’est pas parce que c’est du deep learning que c’est mieux

# Mieux ?

– Les algorithmes d'apprentissage profond n'ont **pas nécessairement de meilleures performances**

# Mieux ?

- Les algorithmes d'apprentissage profond n'ont **pas nécessairement de meilleures performances**

  Notamment : données **tabulaires** (ou **structurées**) : organisées en tableaux de valeurs

  Ex. données cliniques, biologie de laboratoire, signes vitaux, données omiques, etc.

- Malgré des efforts pour créer des **architectures adaptées** et des **modèles de fondation**, les **méthodes ensemblistes** (forêts aléatoires, CatBoost, XGBOOST) marchent souvent mieux.
  - Grinsztajn et al. Why do tree-based models still outperform deep learning on typical tabular data? NeurIPS 2022
  - McElfresh et al. When do neural nets outperform boosted trees on tabular data? NeurIPS 2023
  - Kim et al. CARTE: pretraining and transfer for tabular learning. ICML 2024
  - Hollman et al. Accurate predictions on small data with a tabular foundation model. Nature 2025
  - Ye et al. A closer look at deep learning methods on tabular datasets. arxiv 2025
  - Qu et al. TabICL: A tabular foundation model for in-context learning on large data ICML 2025

# Mieux ?

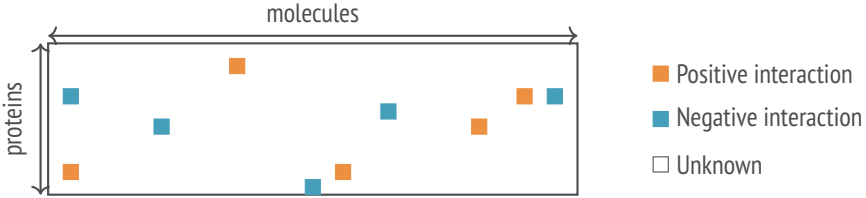
- Les algorithmes d'apprentissage profond n'ont **pas nécessairement de meilleures performances**

  Notamment : **données tabulaires** (mais pas seulement)

# Prédiction d'interactions protéine-ligand

- **But**: prédire si une paire (ligand, protéine) interagit
  - pour une protéine et **toutes les molécules drug-like** : **découverte de médicaments, repositionnement**
  - pour une molécule et **toutes les protéines humaines**: **prédiction d'effets secondaires**, **méchanisme d'action**

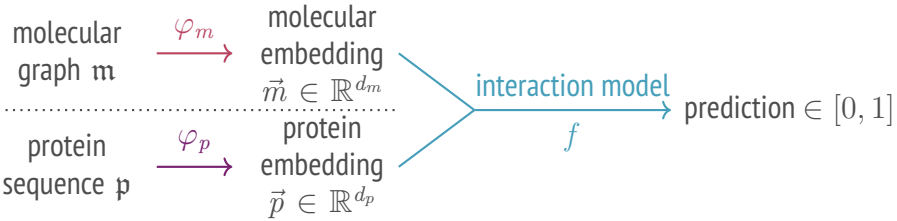
- **LCIdb**: 271 180 molécules, 2 060 protéines, 0.07% paires étiquetées (98% positives)




Guichaoua et al. (2024). **Drug-target interaction prediction at scale: the Komet algorithm with the LCIdb dataset**. J Chem Info Mod

# Prédiction d'interactions protéine-ligand

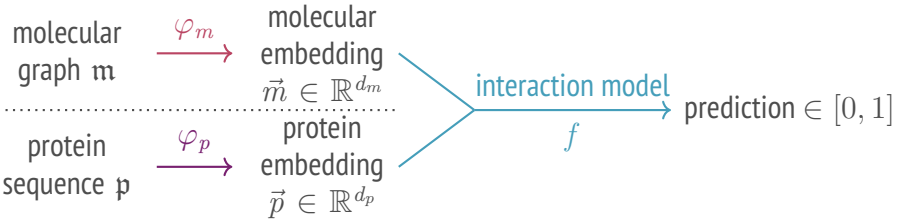
## Cadre méthodologique




Guichaoua et al. (2024). **Drug-target interaction prediction at scale: the Komet algorithm with the LCIdb dataset**. J Chem Info Mod

# Prédiction d'interactions protéine-ligand

## Cadre méthodologique



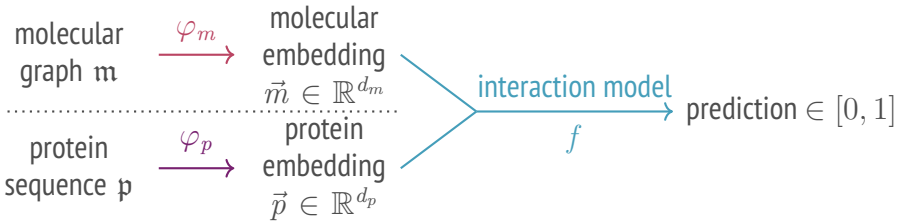

**État de l'art** (2023):

| | $\varphi_m$ | $\varphi_p$ | $f$ |
|---|---|---|---|
| MolTrans | Transformers | Transformers | neural network |
| ConPLex | ECFP | ProtBERT | neural network |

Guichaoua et al. (2024). **Drug-target interaction prediction at scale: the Komet algorithm with the LCIdb dataset**. J Chem Info Mod

# Prédiction d'interactions protéine-ligand

## Komet




**Komet** (Kronecker Optimized METhod):

- $f$ SVM avec noyau de Kronecker $k((\mathfrak{m}, \mathfrak{p}), (\mathfrak{m}', \mathfrak{p}')) = k_m(\mathfrak{m}, \mathfrak{m}')k_p(\mathfrak{p}, \mathfrak{p}')$
- $\varphi_m$ feature map de $k_m$: $\varphi_m$ s.t. $k_m(\mathfrak{m}, \mathfrak{m}') \cong \langle \varphi_m(\mathfrak{m}), \varphi_m(\mathfrak{m}') \rangle$
- $\varphi_p$ feature map de $k_p$

Guichaoua et al. (2024). **Drug-target interaction prediction at scale: the Komet algorithm with the LCIdb dataset**. J Chem Info Mod

# Prédiction d'interactions protéine-ligand

## Komet

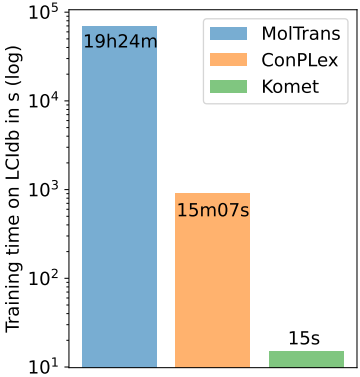
| | $\varphi_m$ | $\varphi_p$ | $f$ |
|---|---|---|---|
| MolTrans | Transformers | Transformers | neural network |
| ConPLex | ECFP | ProtBERT | neural network |
| Komet | feature map of Tanimoto kernel on ECFP | feature map of Local Alignment kernel | SVM with Kronecker kernel |

**Contribution méthodologique** : approximation de Nyström + réduction de dimension pour passer à l'échelle et faire tourner la SVM sur LCIdb

Guichaoua et al. (2024). **Drug-target interaction prediction at scale: the Komet algorithm with the LCIdb dataset**. J Chem Info Mod

# Prédiction d'interactions protéine-ligand

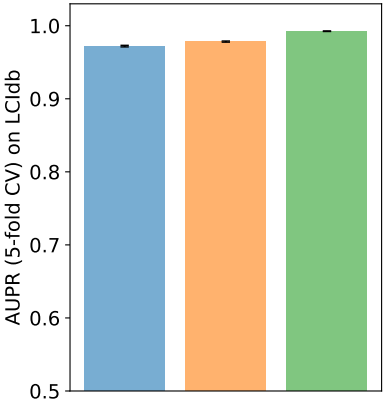
## Komet est bien plus rapide

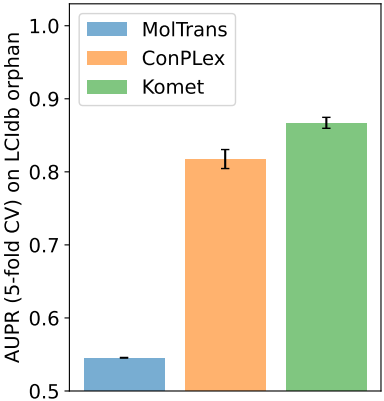



Guichaoua et al. (2024). **Drug-target interaction prediction at scale: the Komet algorithm with the LCIdb dataset**. J Chem Info Mod

# Prédiction d'interactions protéine-ligand

## Komet a de meilleures performances




Guichaoua et al. (2024). **Drug-target interaction prediction at scale: the Komet algorithm with the LCIdb dataset**. J Chem Info Mod

# Mieux ?

- Les algorithmes d'apprentissage profond n'ont **pas nécessairement de meilleures performances**

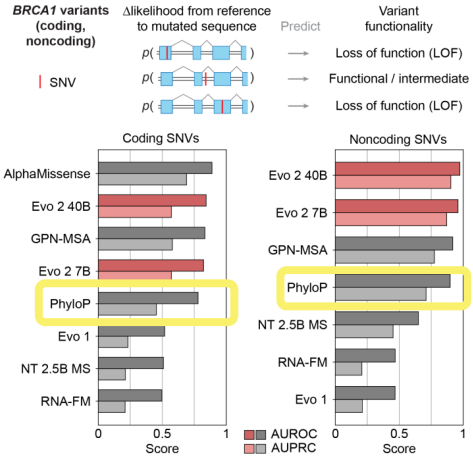
  Notamment : **données tabulaires** (mais pas seulement)

- Même si les performances sont meilleures, prendre en compte :
  - **coût** matériel/énergétique/écologique/temps
  - **efforts** d'optimisation des hyperparamètres
  - **opacité**
  - **biais** des modèles de fondation

    "Language models over-generate cases describing male patients, creating synthetic corpora that are not consistent with documented prevalence for these disorders"

    Ducel et al. (2025) **Women do not have heart attacks! Gender biases in automatically generated clinical cases in French**. Findings of the ACL

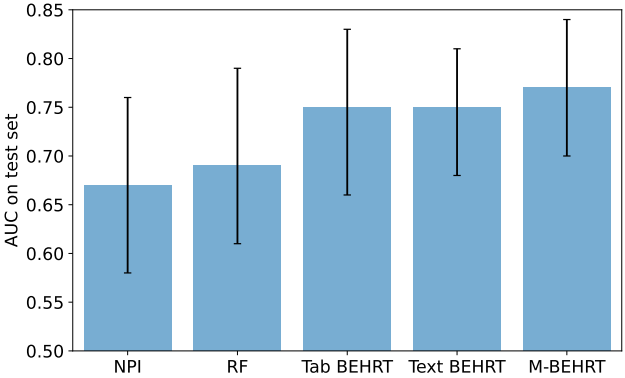
# 1. Prédiction de pathogénicité avec Evo2




Brixi et al. (2025), Genome modeling and design across all domains of life with Evo 2, BioRxiv

# 2. Prédiction de rechute à partir de parcours patient

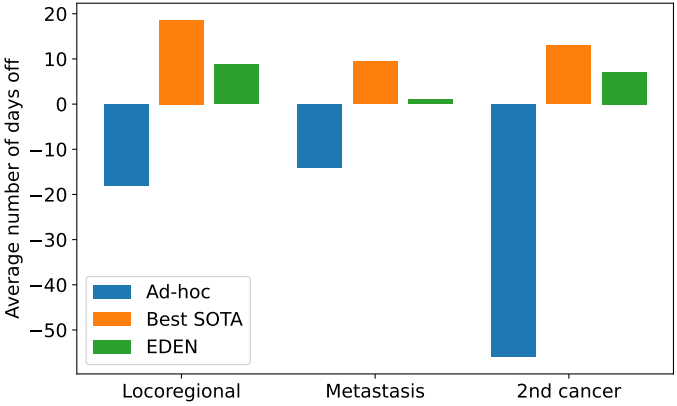
En **combinant** Tabular BEHRT and Text BEHRT par **cross-attention**




Mbaye et al. (2025), **Multimodal BEHRT: transformers for multimodal Electronic Health Records**, Front Onc

# 3. Identification d'événements dans les parcours patient

Évaluation sur un **jeu de données indépendant** de 800 parcours




Dumas et al. (2022) **EDEN : An Event DEtection Network for the annotation of recurrences in administrative claims data**. ML4Health

# Du benchmark à l'évaluation clinique

- Validation sur **plusieurs cohortes**
- **Essais cliniques**
- Intégration à la **pratique**
- **Modèle économique**

# Points clés

I. Avoir la bonne **métrique d'évaluation** pour l'application envisagée

II. Évaluer sur des **données de validation** appropriées

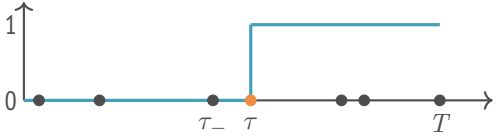
III. Comparer à de bonnes **baselines**

IV. Ce n'est pas parce que c'est du **deep learning** que c'est mieux

V. Il y a loin du **benchmark à la clinique**

# Event annotation in patient trajectories

- **EDEN:** Extends **T-LSTM** to **censored data**, using a new **loss function** to estimate the **event rate**
- **Patient event rate function:** $W(t) = \mathbb{P}[t \geq \tau]$ $\tau$ = time to event




- **Loss function:**

$$\underbrace{\alpha_1 \mathcal{L}_1}_{\text{BCE}} + \underbrace{\alpha_2 \mathcal{L}_2 + \alpha_3 \mathcal{L}_3}_{\text{uncensored}} + \underbrace{\alpha_4 \mathcal{L}_4}_{\text{censored}}$$

  - **uncensored data:** $\mathcal{L}_2$ penalizes low $\widetilde{W}(\tau)$ and $\mathcal{L}_3$ penalizes high $\widetilde{W}(\tau_-)$
  - **censored data:** $\mathcal{L}_4$ penalizes high $\widetilde{W}(T)$

# An example of gene-level confounding: FatHMM-W

$$\text{FatHMM-W} = \ln\left(\frac{(1 - P_{\text{wild-type}}).(F_{\textbf{neutral}} + 1)}{((1 - P_{\text{mutated}}).(F_{\textbf{pathogenic}} + 1))}\right)$$

FatHMM-W uses a conservation score, weighted by the relative proportions of pathogenic and benign mutations **in the same or homologous** protein or protein domains.