Nucleotide Transformer

Building and evaluating robust foundation models for human genomics

Hugo Dalla-Torre[1], Liam Gonzalez[2], Javier Mendoza-Revilla[3], Nicolas Lopez Carranza[1], Adam Henryk Grzywaczewski[2], Francesco Oteri[1], Christian Dallago[2], Evan Trop[1], Bernardo P. de Almeida[1], Hassan Sirelkhatim[1], Guillaume Richard[1], Marcin Skwark[1], Karim Beguir[1], Marie Lopez[1] & Thomas Pierrot[1]

[1] *InstaDeep*
[2] *Nvidia*
[3] *Technical University of Munich*

28/11/2024

https://www.nature.com/articles/s41592-024-02523-z

Contexte

- **Problèmes :**
  - **Données Annotées Rares :** Les données annotées en génomique sont coûteuses et difficiles à produire.
  - **Modèles Spécialisés :** Les modèles existants sont souvent conçus pour des tâches spécifiques et ne généralisent pas bien à d'autres tâches.
  - **Dépendance aux Données :** Les modèles actuels nécessitent de grandes quantités de données annotées pour être performants.

Objectifs :

Modèle Généraliste : Proposer un modèle de fondation capable de s'adapter à de nouvelles tâches .

Pré-entraînement sur Données Diversifiées : Utiliser des données génomiques humaines et multispecies pour améliorer la généralisation.

Fine-Tuning Efficace : Adapter rapidement le modèle à de nouvelles tâches en réentraînant seulement 0.1% des paramètres.

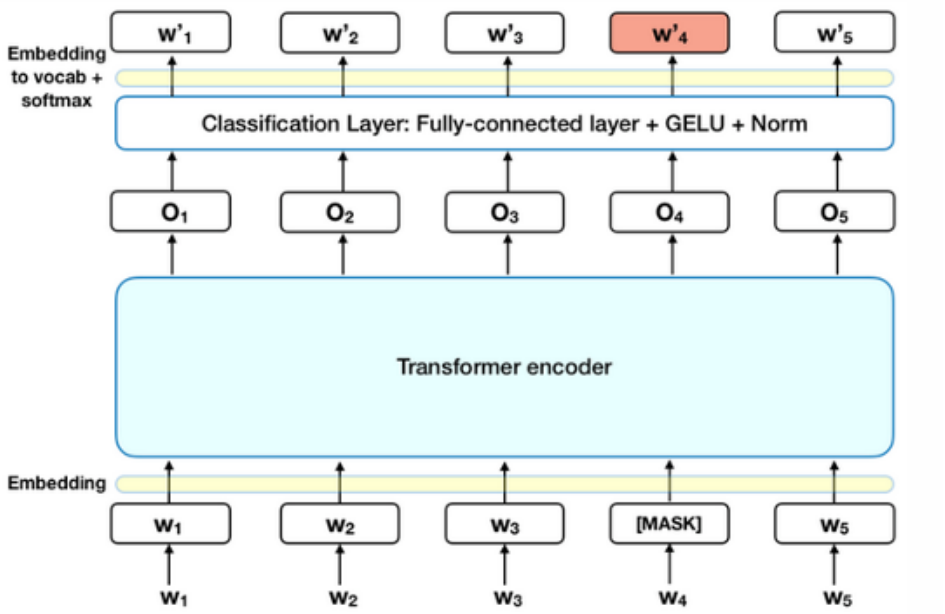
Zero-Shot Learning : Permettre au modèle de prédire l'impact des variants génétiques sans données annotées spécifiques.

Sommaire

- **NLP et biologie**
- **Entrainement**
- **Méthodes d’évaluation**
- **Interprétation de l’évaluation**
- **Conclusion**

NLP et biologie

- **Révolution du NLP**: Modèles de fondation
  - Masked Language Modeling (MLM)



Application en biologie:

1. Prédiction Protéine: Séquence d'acides aminés
   - Structure: AlphaFold
   - Fonction biologique : ProtBERT, ESM

2. Extension aux séquences nucléotidiques (ADN):
   - Comprendre la régulation génétique
   - Identifier des mutations, des régions régulatrices (enhancers, promoteurs)

Limitations des modèles actuels

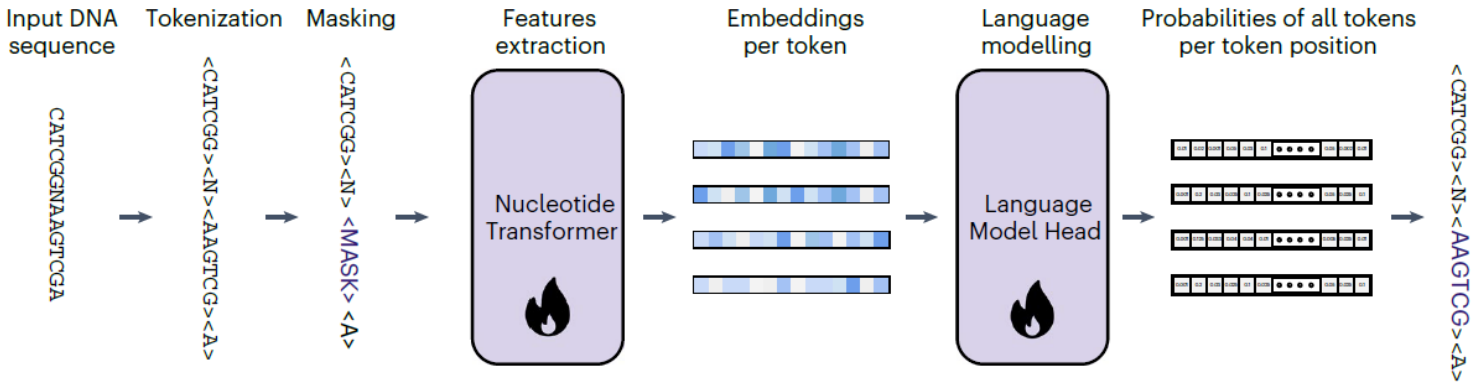
- **Dégradation performances** sur **tâches spécifiques**
- Manque de **généralisation inter-espèces**
  - **Multispecies 2.5B** ➔ 850 génomes d'espèces + 3202 génomes humains
- **Dépendance aux annotations** ➔ manque de données annotées en génomique
- **Longueur des séquences**
  - ***DNABERT*** : limité à 512 bp, insuffisant pour longues régions génomiques
  - ***Enformer*** *(Transformer & Convolutions)*
  - ***HyenaDNA*** *(32 – 200 kb)* ➔ coût computationnel important

Entrainement

- 6 kb de données génomiques non annotées
- 4 modèles

Human ref 500M	1000G 500M	1000G 2.5B	Multispecies 2.5B
1 génome de référence	3202 génomes humains	3202 génomes humains	850 espèces diverses

- Masked Language Modelling
- Séquences de nucléotides ➔ phrases
- 6-mers ➔ mot

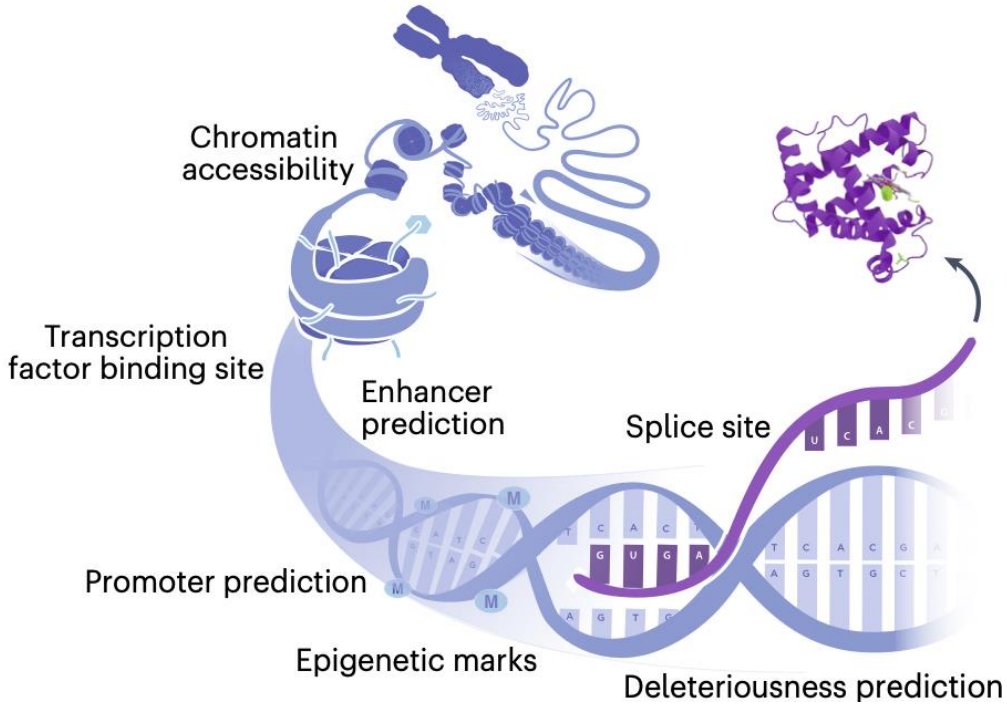


Entrainement

- **Méthode BERT**
- **Taille de batch**
  - 500M : 14 séquences
  - 2.5B : 2 séquences
- **MLM :** 15% des tokens sont sélectionnés pour être modifiés :
  - 80% des tokens ➔ ‘MASK’
  - 10% ➔ token aléatoire (sauf CLS, PAD, MASK & 1000G)
- **Optimisation**
  - Cross-entropy sur les positions masquées
  - Accumulation de gradients (batchsize ~ 1M de tokens)
  - Adam : $\beta_1 = 0.9$, $\beta_2 = 0.999$, $\epsilon = 10^{-8}$

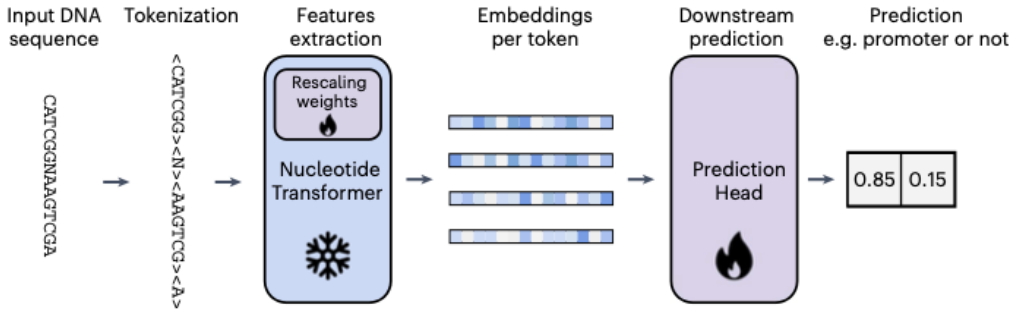
Limitations des modèles actuels

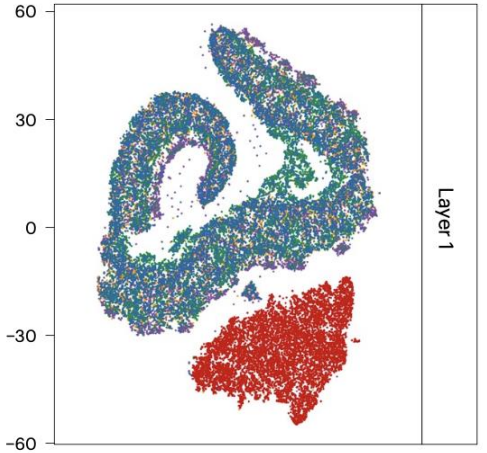
- Évaluer la capacité des NT à prédire divers phénomènes génomiques
- **18 tâches** de prédiction génomique :
  - Sites d'épissage (splice sites) → GENCODE
  - Promoteurs → Eukaryotic Promoter Database
  - Modifications d'histones & enhancers → ENCODE
- **Validation croisée** en 10 plis
- **Probing** vs **Fine-tuning** pour tester les performances des modèles

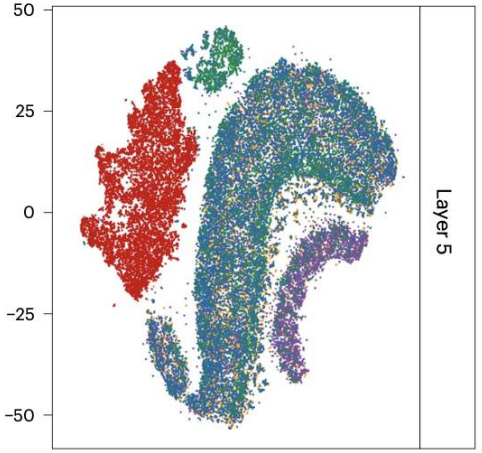


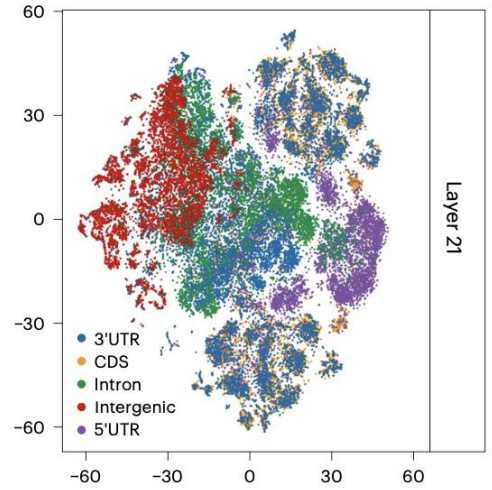
Probing et Fine-tuning

- Tester **la qualité des embeddings** du modèle NT sans changer ses poids:
  - Extraction des embeddings des **couches intermédiaires**.
  - On utilise ces embeddings comme entrées pour un modèle classique de Machine Learning
  - On entraîne ce modèle simple pour prédire une tâche spécifique.
- **Adapter NT à une tâche spécifique** en modifiant ses poids





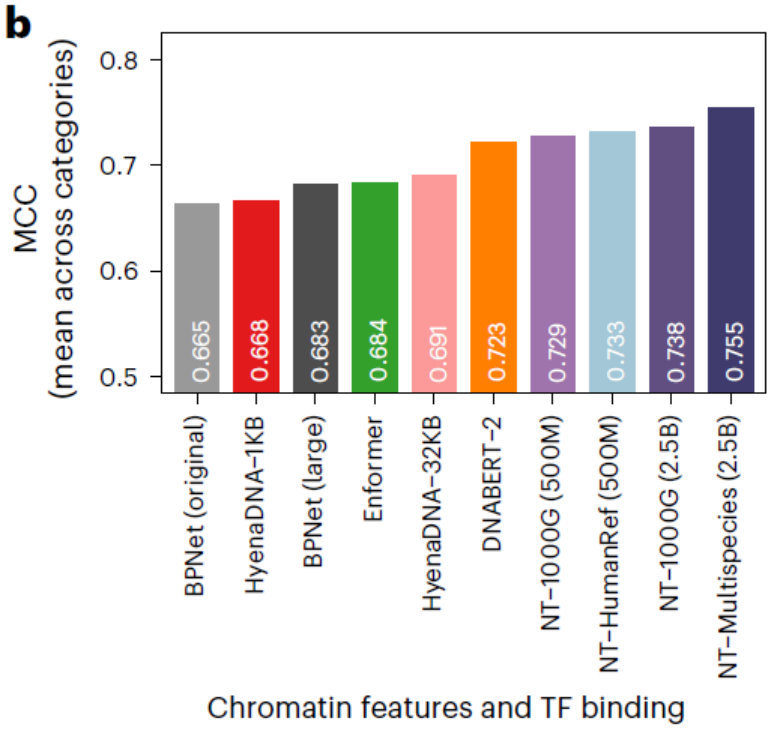




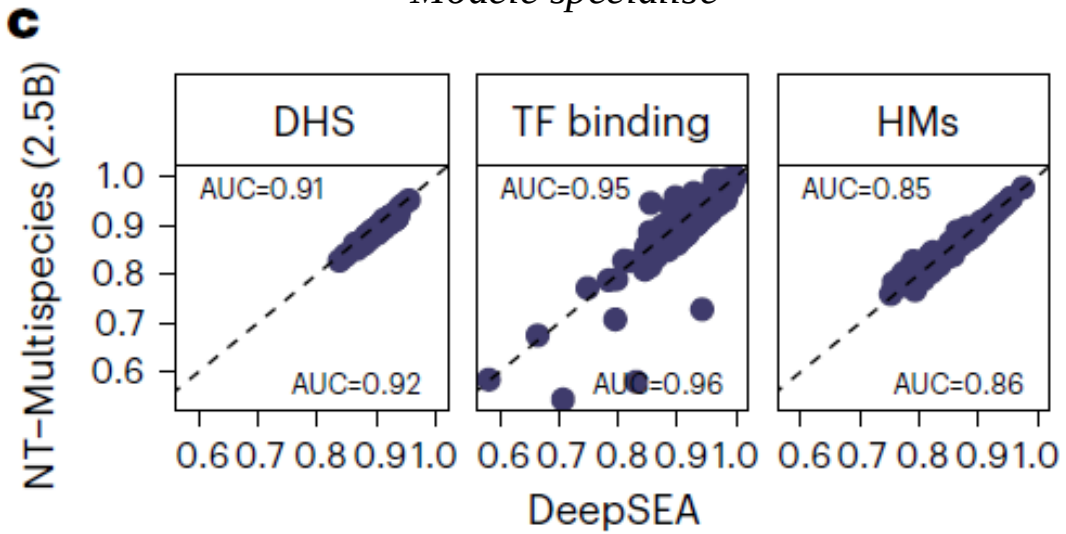
Résultats : Comparaison des approches

Par rapport à la prédiction génomique, son efficacité computationnelle et son potentiel d'application meow

Performance Générale



Modèle spécialisé

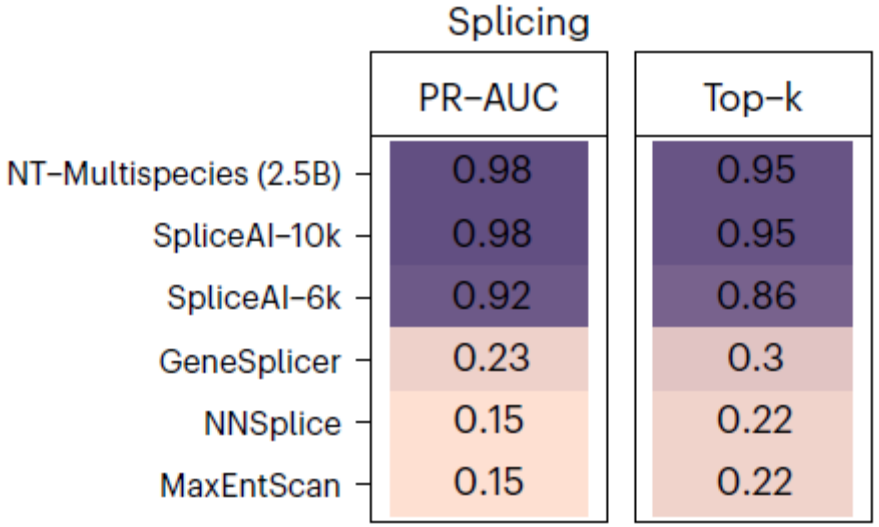


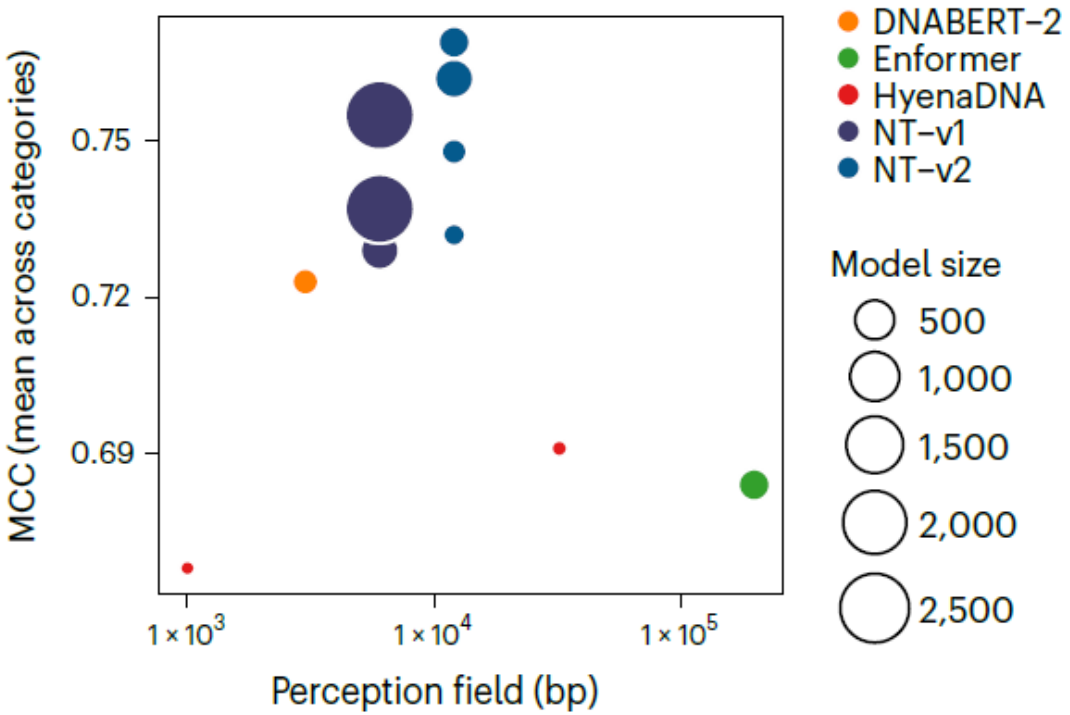
Sans supervision directe sur ces tâches, est capable de **capturer des motifs génomiques essentiels** et de rivaliser avec des modèles explicitement optimisés pour la prédiction de ces éléments.

Résultats : Comparaison des approches

- Égalant **SpliceAI-10k** malgré un **contexte d'entrée plus court**
- Surpasse **SpliceAI** lorsque les **séquences d'entrée sont limitées à 6 kb**

Modèle spécialisé de splicing





Utilisations possibles et applications futures

- **Modèle généraliste** puissant qui **surpasse** ou **égale** les modèles supervisés spécialisés
- Représentations apprises par NT sont **riches** et **transférables** à des tâches complexes en génomique.
- **Modèle fondationnel** évite d'avoir à entraîner un modèle spécifique pour chaque type de tâche, réduisant ainsi les coûts de calcul et améliorant la flexibilité.

Prédiction des mutations pathogènes

Attribuer un score d'impact aux mutations

Amélioration des modèles de médecine personnalisée

Affiner les diagnostics

Applications en biotechnologie et en génomique évolutive

Comprendre l'émergence de nouvelles mutations

Merci