Université Paris Cité

Projet long :

Predicting protein-carbohydrate binding sites using protein embeddings

Master 2 BI

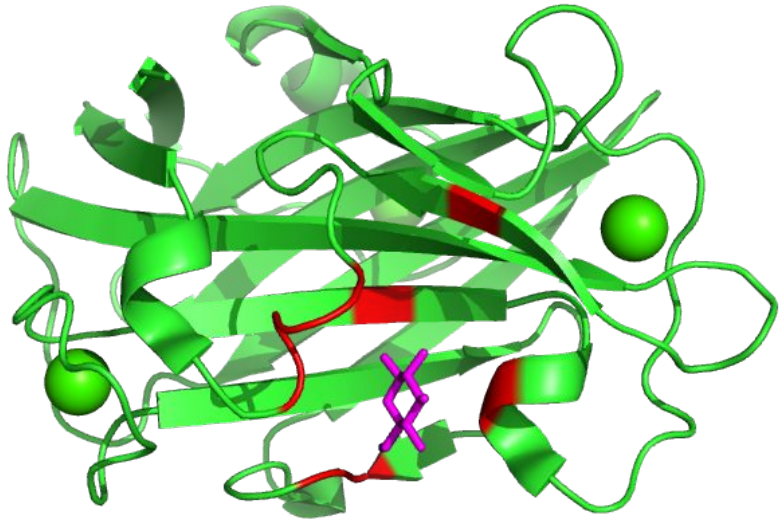
Auteur : Anas BELAKTIB

16 Janvier 2023

Interactions entre les protéines et les glucides

- Important dans différents processus biologiques
  - Embryogenèse [1]
  - Réponse immunitaire [2]
  - Absorption de toxines bactériennes [4]
- Regain d'intérêt depuis le SARS-CoV-2
  - Mise en place de vaccins
  - bouclier de glycanes [6]

- Événement RARE



Complexité de l’étude

Glucides forment des glycanes complexes ramifiés

Ligand très polyvalent

Mal annotées dans la PDB [10]

Flexibilité

Faible énergie des interactions protéine-glucide

Outils de prédiction actuels manquent de précision

Jeu de données

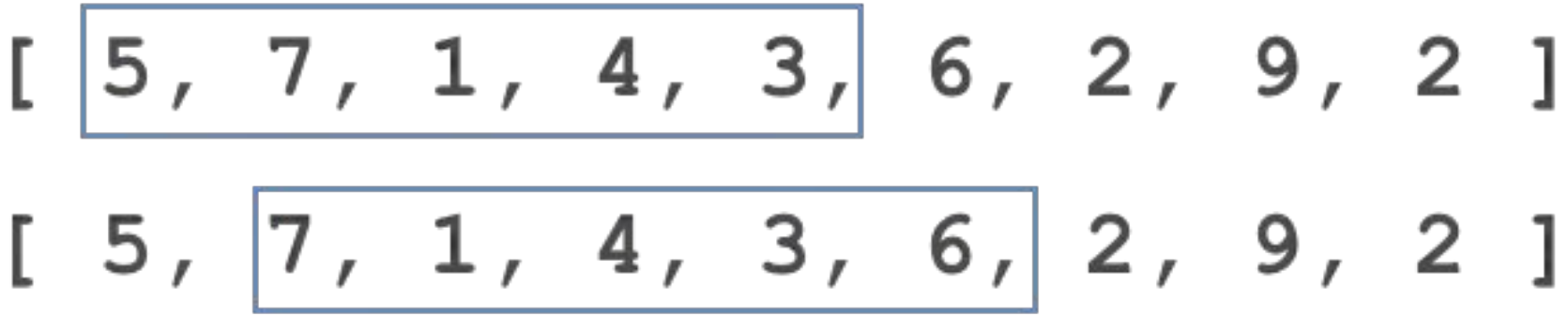
- 5172 protéines [12]
  - entre 14 et 1528 acides aminés
  - 321 features
  - 1 classes a prédire

- 3 jeux de données :
  - Learn, validation, test
  - Non redondante
  - Utilisation de fenêtre glissante

Fenêtre glissante

Permet de prendre en considération les résidus alentours

Taille de notre fenêtre = 13 (6 avant + 6 après)



https://itnext.io/sliding-window-algorithm-technique-6001d5fbe8b3

Déséquilibre des classes

Occurrence d'une des classes est très élevée

Les modèles ont tendance à attribuer la classe la plus fréquente

Fausses prédictions camouflées par la présence de nombreuses vraies prédictions

Fixation de sucre rare

Ajout de poids pour chaque classe : [0 : 0,537 1 : 7,233]

Transformers Evolutionary Scale Modeling

Facebook AI Research

Traitement du texte initial

Version utilisée Esm2_t6_8M_UR50D

320 features + 1 emplacement dans la fenêtre glissante

Modèles

Callbacks :

- Sauvegarde du meilleur modèle
- Création de matrices de confusion à la dernières époques
- EarlyStopping

Métriques :

- Accuracy = count / total
- Precision = (true_positives)/(true_positives + false_positives)
- Recall = (true_positives)/(true_positives + false_negatives)

	CNN	CNN simple
Nombre de paramètres	4.8M	93k
Durée d'une époque	14 mins	10 mins

Modèles

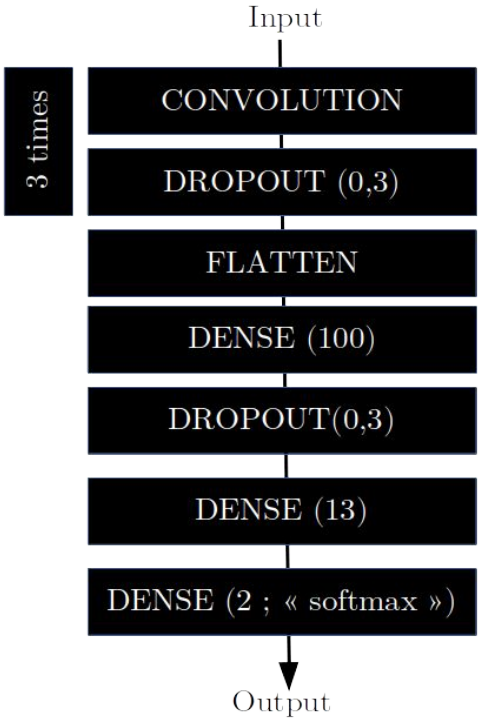


Schéma de conception du modèle CNN

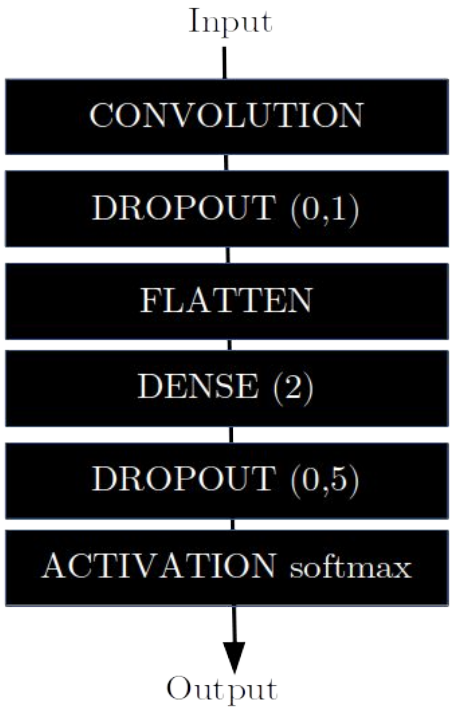
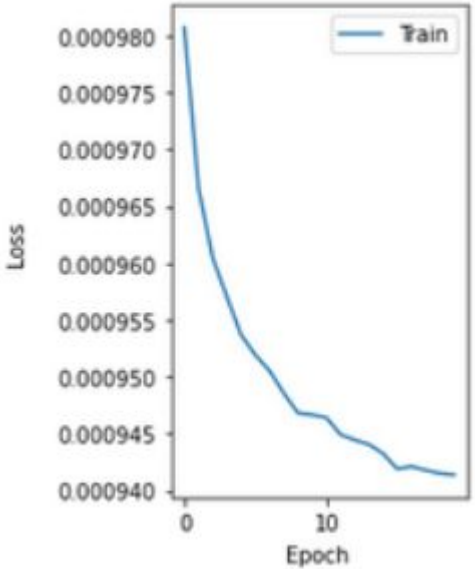
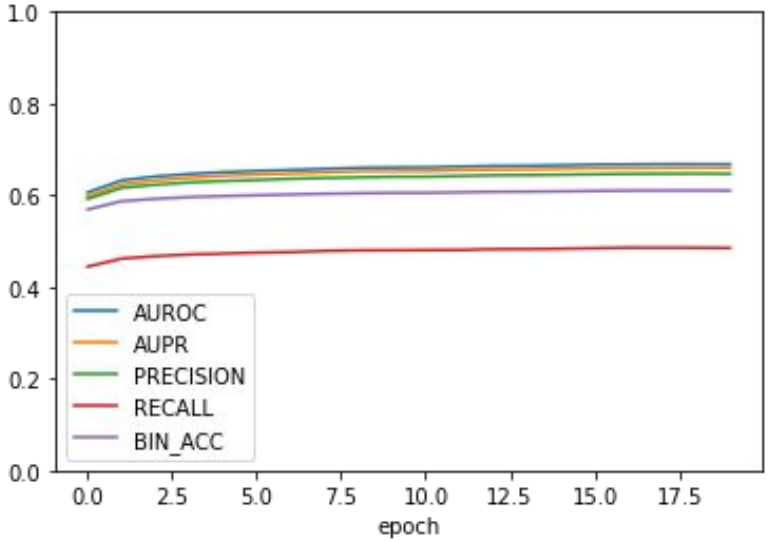


Schéma de conception du modèle CNN simple

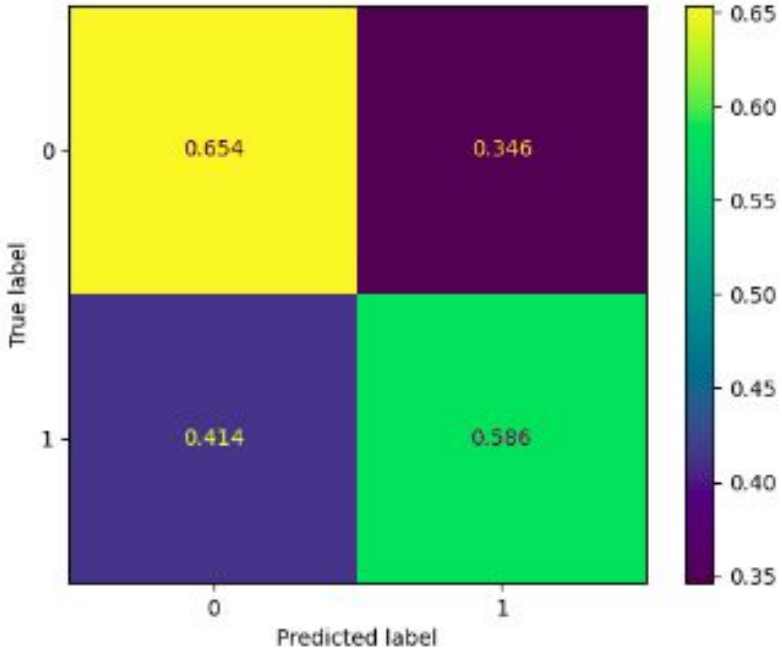
Résultat CNN simplifié Learn

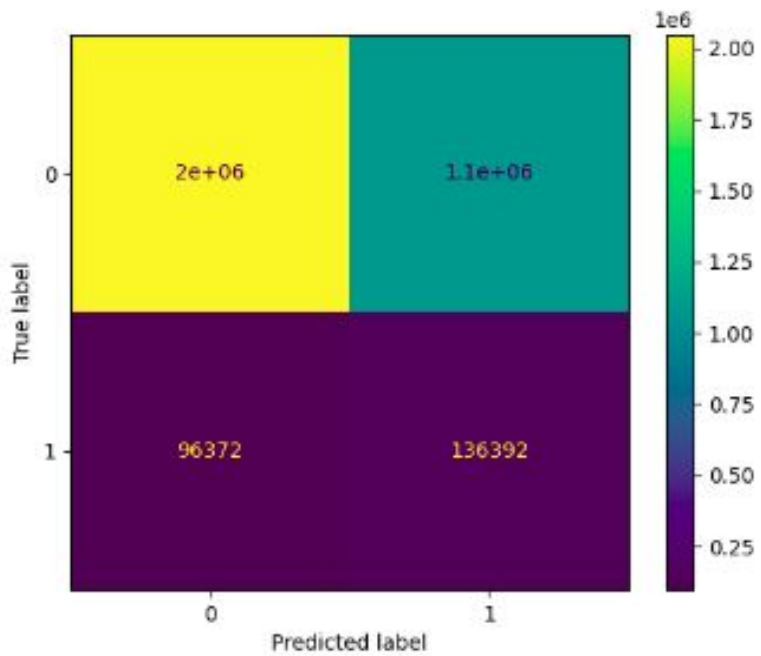




Graphiques des performances du modèle CNN simplifié en fonction des époques lors de la phase d'apprentissage

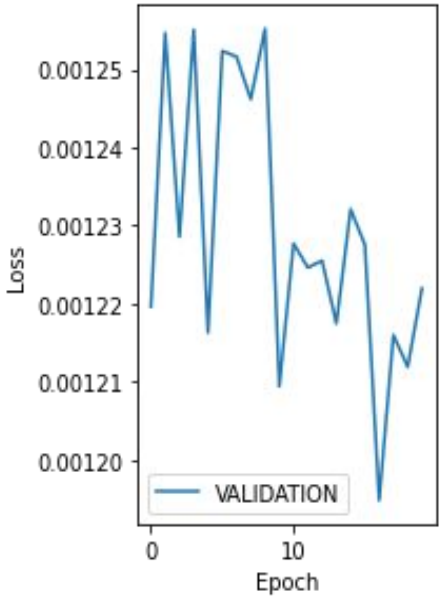
Résultat CNN simplifié Learn

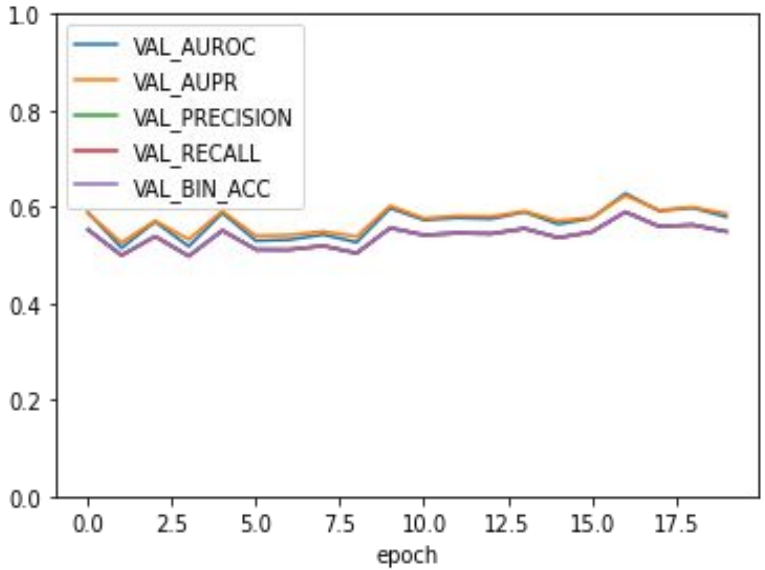




Matrices de confusion brut et normalisé du modèle CNN simplifiée lors de la phase d'apprentissage

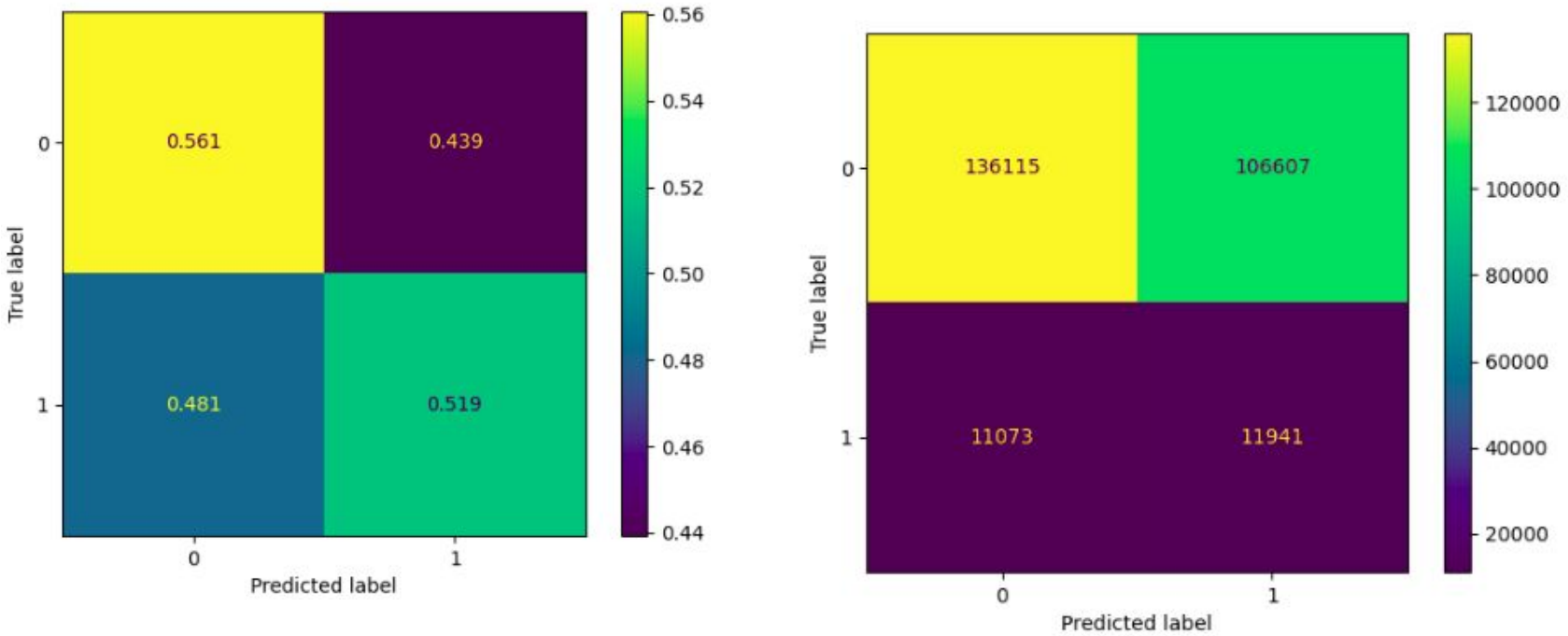
Résultat CNN simplifié Validation





Graphiques des performances du modèle CNN simplifié en fonction des époques lors de la phase de validation

Résultat CNN simplifié Validation



Matrices de confusion brut et normalisé du modèle CNN simplifiée lors de la phase de validation

Perspectives

Simple peut être meilleur

Optimiser le réseau simple

Réalisé plus d’époques

Tester les autres réseaux compilés Inception / GRU

Tester notre meilleur modèle sur le test set

MERCI DE VOTRE ATTENTION

Références

(1) Onuma Y., Tateno H., Tsuji S., Hirabayashi J., Ito Y., Asashima M.. A lectin-based glycomic approach to identify characteristic features of xenopus embryogenesis. PLoS One. 2013; 8:e56581

(2) Maverakis E., Kim K., Shimoda M., Gershwin M.E., Patel F., Wilken R., Raychaudhuri S., Ruhaak L.R., Lebrilla C.B.. Glycans in the immune system and the altered glycan theory of autoimmunity: a critical review. J. Autoimmun. 2015; 57:1–13.

(3) Hauri H.-P., Nufer O., Breuza L., Tekaya H.B., Liang L.. Lectins and protein traffic early in the secretory pathway. Biochem. Soc. Symp. 2002; 69:73–82.

(4) Zuverink M., Barbieri J.T.. Protein toxins that utilize gangliosides as host receptors. Prog. Mol. Biol. Transl. Sci. 2018; 156:325–354.

(5) Chen L., Li F.. Structural analysis of the evolutionary origins of influenza virus hemagglutinin and other viral lectins. J. Virol. 2013; 87:4118–4120.

(6) Watanabe Y.; Allen J. D.; Wrapp D.; McLellan J. S.; Crispin M. Site-Specific Glycan Analysis of the SARS-CoV-2 Spike. Science 2020, eabb9983. 10.1126/science.abb9983.

(7) Casalino, L.; Gaieb, Z.; Goldsmith, J. A.; Hjorth, C. K.; Dommer, A. C.; Harbison, A. M.; Fogarty, C. A.; Barros, E. P.; Taylor, B. C.; McLellan, J. S.; Fadda, E.; Amaro, R. E. Beyond Shielding: The Roles of Glycans in the SARS-CoV-2 Spike Protein. *ACS Cent. Sci.* **2020**, *6* (10), 1722–1734. https://doi.org/10.1021/acscentsci.0c01056.

(8) Sztain, T.; Ahn, S.-H.; Bogetti, A. T.; Casalino, L.; Goldsmith, J. A.; Seitz, E.; McCool, R. S.; Kearns, F. L.; Acosta-Reyes, F.; Maji, S.; Mashayekhi, G.; McCammon, J. A.; Ourmazd, A.; Frank, J.; McLellan, J. S.; Chong, L. T.; Amaro, R. E. A Glycan Gate Controls Opening of the SARS-CoV-2 Spike Protein. *Nat. Chem.* **2021**, *13* (10), 963–968. https://doi.org/10.1038/s41557-021-00758-3.

(9) Yan R.; Zhang Y.; Li Y.; Xia L.; Guo Y.; Zhou Q. Structural Basis for the Recognition of SARS-CoV-2 by Full-Length Human ACE2. Science (Washington, DC, U. S.) 2020, 367 (6485), 1444–1448. 10.1126/science.abb2762.

(10) Lütteke T., Frank M., von der Lieth C.-W.. Data mining the protein data bank: automatic detection and assignment of carbohydrate structures. Carbohydr. Res. 2004; 339:1015–1020.

(11) Burley S.K., Berman H.M., Bhikadiya C., Bi C., Chen L., Di Costanzo L., Christie C., Dalenberg K., Duarte J.M., Dutta S. et al... RCSB Protein Data Bank: biological macromolecular structures enabling research and education in fundamental biology, biomedicine, biotechnology and energy. Nucleic Acids Res. 2019; 47:D464–D474.

(12) Copoiu, L.; Torres, P. H. M.; Ascher, D. B.; Blundell, T. L.; Malhotra, S. ProCarbDB: A Database of Carbohydrate-Binding Proteins. *Nucleic Acids Res.* **2020**, *48* (D1), D368–D375. https://doi.org/10.1093/nar/gkz860.

(13) Cheng H, Schaeffer RD, Liao Y, Kinch LN, Pei J, Shi S, Kim BH, Grishin NV. ECOD: an evolutionary classification of protein domains. PLoS Comput Biol. 2014 Dec 4;10(12):e1003926. doi: 10.1371/journal.pcbi.1003926

(14) https://github.com/facebookresearch/esm

(15) https://scikit-learn.org/dev/index.html