

# Représentation 3D des résultats

Je teste ici plusieurs paramètres pour voir quel est le meilleur combo pour avoir les meilleurs résultats.

Les réseaux présentés sont :

- Méthode Resize sur des maps non normalisées (valeur à 255). 1048 step par epoch (15 epoch) et training sur toutes les couches (sans poids pré-load) → **R255Full**
- Méthode Resize sur des maps normalisées (entre 0 et 1). 1048 step par epoch (15 epoch) et training sur toutes les couches → **R01Full**
- Méthode Resize sur des maps normalisées (entre 0 et 1). 1048 step par epoch (15 epoch) et training sur les dernières couches avec des poids COCO → **R01Heads**
- Méthode Padding sur des maps normalisées, 131 step par epoch, 15 epoch, training sur les dernières couches → **PAD**

Les réseaux avec le padding étant difficiles à obtenir (problème de mémoire GPU qui fait crash le training car je crois que cette méthode demande plus de ressources que le Resize normal), il n'y en a qu'un seul ici, et il est entraîné sur 15 epoch avec 131 step par epoch (contre 1048 pour les autres). Les protéines de références sont issues de l'article **An ambiguity principle for assigning protein structural domains**.

9PAP

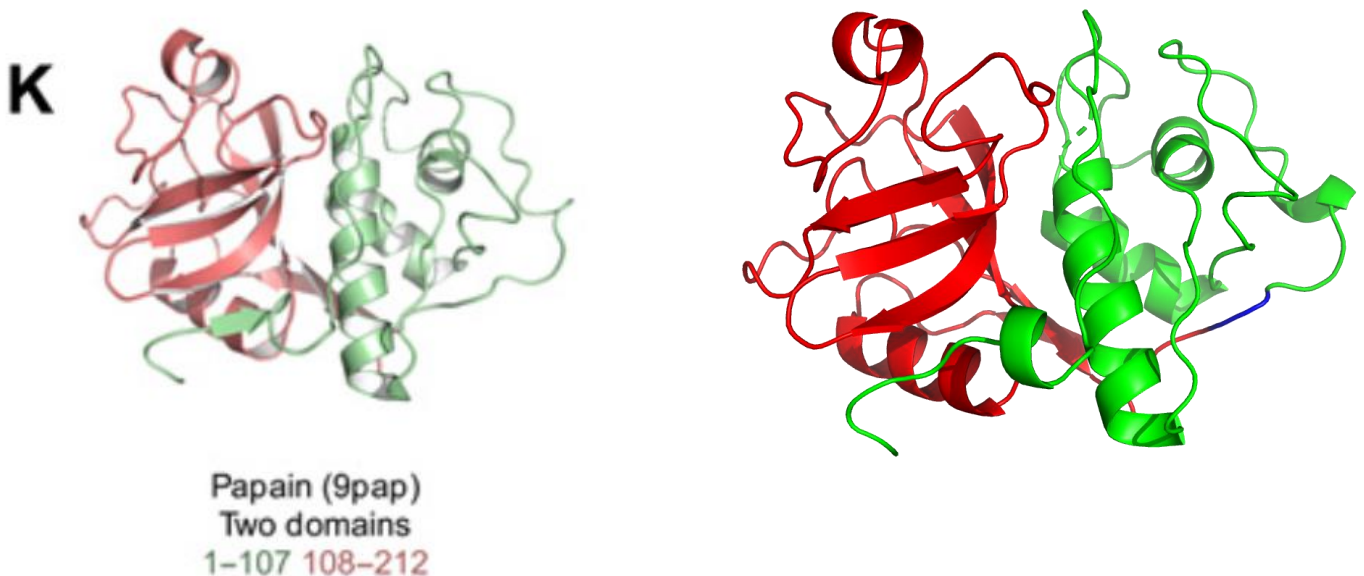


Figure 1 : Découpage avec le réseau R255Full de 9PAP

Domaines trouvés avec le réseau **R255Full** pour cette protéine :

**PU predicted : 106 208; score : 0.9798335433006287 → peut être associé au domaine 108-212**

**PU predicted : 1 103; score : 0.9713642001152039 → peut être associé au domaine 1-107**

PU predicted : 0 182; score : 0.9255672097206116

PU predicted : 0 55; score : 0.714890718460083

Parmi les 4 domaines trouvés par le réseau **R255Full**, seuls 2 sont « corrects » selon la référence, et cela a été déterminé manuellement. Il manque encore une étape pour pouvoir déterminer automatiquement les bons découpages.

Les délimitations vont donc de 1-103 et 106-208. Les résidus **208-212** et **103-106** ne sont pas pris en compte dans ces délimitations, ils sont donc colorés en **bleu** (Fig 1 à droite).

**(les scores représentent le score de confiance donné par le réseau et non à un score PU)**

Il y a donc deux questions ici :

- **Comment déterminer les bons des « mauvais » domaines ?** (Ici 2/4 domaines)  
Sachant que les domaines trouvés par le réseau sont souvent chevauchants, donc les colorations se mélangent et nécessite une intervention manuelle pour voir les domaines 1 à 1 et déterminer les meilleurs domaines.
- **Comment gérer les zones non sélectionnées par le réseau ?** En bleu sur la figure

Voici un résultat avec le même réseau mais en normalisant les données (**R01Full**) :

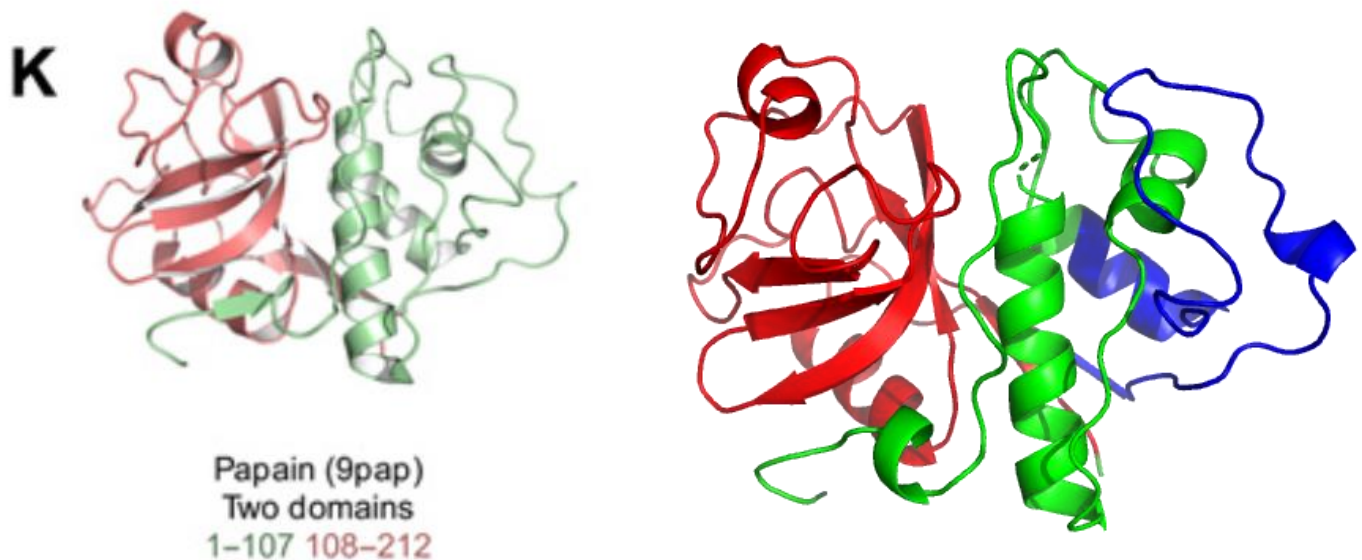


Figure 2 : Découpage avec le réseau R01Full de 9PAP

Domaines prédits :

PU predicted : 2 211; score : 0.8196399211883545

**PU predicted : 2 70; score : 0.7503053545951843**

**PU predicted : 111 211; score : 0.6194136738777161**

Les résultats de **R01Full** sont moins bons que pour le **R255Full**, car la zone non sélectionnée par le réseau est beaucoup plus grande → 70-111 en bleu.

Pour le Padding (**PAD**) :

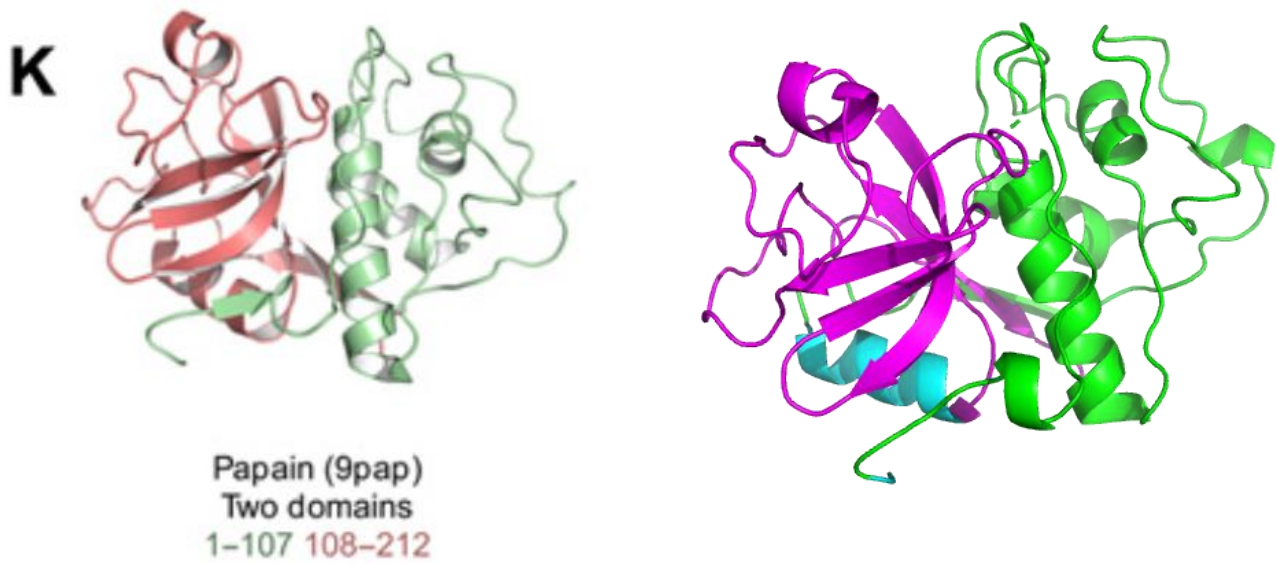


Figure 3 : Découpage avec le réseau PAD de 9PAP

Les résultats obtenus par le padding sont ok mais il y a toujours une zone morte (cyan) qui va de 117 à 127 qui n'est pas sélectionné par le réseau.

Domaines prédits :

PU predicted : 0 64; score : 0.9976034760475159

PU predicted : 143 211; score : 0.9879078269004822

PU predicted : 8 211; score : 0.9832943677902222

**PU predicted : 3 117; score : 0.9243836402893066**

**PU predicted : 127 211; score : 0.7387328743934631**

Les résultats avec **R01Heads** sont pas terribles et donc pas représentés.

1JHGA

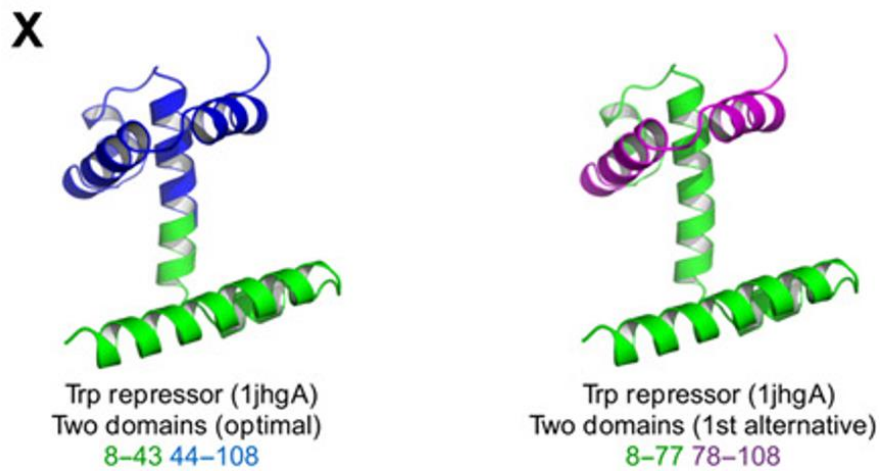


Figure 4 : Découpage de référence issu de l'article

Premier résultat avec le réseau **R255Full** :

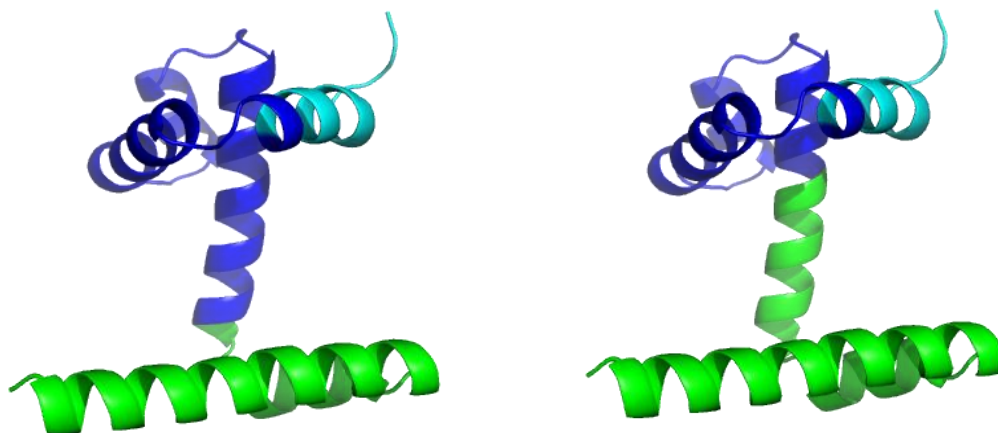


Figure 5 : Découpage prédit pour 1JHGA pour le réseau **R255Full**

Domaines prédits :

PU predicted : 11 108; score : 0.9609130024909973

**PU predicted : 8 55; score : 0.8796014785766602**

**PU predicted : 47 96; score : 0.7740593552589417**

Ici, le domaine en cyan représente une zone morte non sélectionnée par le réseau (96-108).

Le domaine en bleu (47-96) est chevauchant au domaine vert (8-55), si on colore le domaine bleu à partir du résidu 47, on obtient la figure de gauche, et si on colore le domaine bleu après avoir colorié le domaine en vert, soit au résidu 55, on a la figure de droite.

Une autre question se pose ici :

- Quel domaine privilégié quand il chevauche un autre domaine ?

Pour le réseau **R01Full**, **PAD** et **R01Heads**, les résultats sont assez similaires :

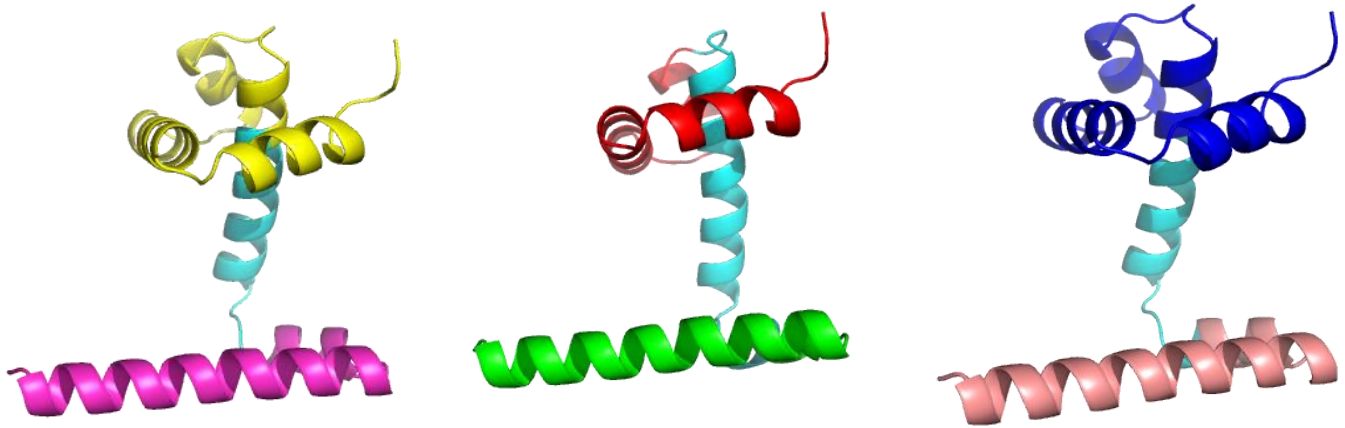


Figure 5 : Découpage prédit pour 1JHGA pour les réseaux **R01Full**, **PAD** et **R01Heads**

La zone morte est représentée en cyan pour les 3 réseaux. Cette dernière est plus importante pour le **Padding** (milieu, 38-67), et similaire entre **R01Full** (gauche, 41-59) et **R01Heads** (droite, 39-57).

Les domaines pour **R01Full** (Gauche) :

PU predicted : 8 108; score : 0.8196440935134888

PU predicted : 8 41; score : 0.7498291730880737

PU predicted : 59 108; score : 0.6188859343528748

**PAD** (milieu) :

PU predicted : 7 38; score : 0.9974536299705505

PU predicted : 75 108; score : 0.9878290891647339

PU predicted : 11 108; score : 0.98313969373703

PU predicted : 67 108; score : 0.7406870722770691

PU predicted : 70 108; score : 0.6087132692337036

**R01Heads** (droite) :

PU predicted : 7 108; score : 0.8898613452911377

PU predicted : 57 108; score : 0.729573667049408

PU predicted : 53 84; score : 0.5829322338104248

PU predicted : 8 39; score : 0.5360698699951172

## Conclusion :

Il y d'autres résultats qui sont moins bons, mais ceux là permettent de mettre en lumière les points à améliorer :

- **Comment déterminer les bons des « mauvais » domaines ?**
- **Comment gérer les zones non sélectionnées par le réseau ?**
- **Quel domaine privilégier quand il chevauche un autre domaine ?**

On a pu également voir que le **R255Full**, soit la méthode **Resize**, avec des données **non normalisé** (255) et en entraînant **toutes les couches** semble donner les meilleurs résultats. Mais pour confirmer cela, il faudrait plus de résultats de comparaison.

Pour rappel, le réseau **PAD** est entraîné sur 15 epoch avec 131 step par epoch contre 1048 step pour les autres réseaux, donc il y a déjà un problème d'équité.