

# Stage M2: Développement d'un modèle probabiliste de diffusion pour la sélection phénotypique appliquée à la prédiction multi-environnements.

## Contexte

La sélection phénotypique est une approche consistant à prédire le phénotype (la taille, la résistance aux maladies, etc.) d'un individu (ici, une variété de plante) à partir d'un spectre dans le proche infra-rouge, appelé spectre NIRS, collecté sur l'un des organes de cet individu (la feuille, la racine) [1]. Pour se faire, on dispose d'un ensemble d'entraînement, *i.e.*, d'un échantillon d'individus pour lesquels le spectre NIRS et le phénotype d'intérêt ont été collectés. Le problème de prédiction est alors traité comme un problème classique d'apprentissage supervisé.

Dans le cas particulier de la génétique des plantes, un même individu (*i.e.*, un même génotype) peut être observé dans différents environnements afin d'évaluer ses capacités d'adaptation. Toutefois, il est généralement coûteux d'évaluer l'ensemble des individus dans tous les environnements. Les méthodes de prédiction existantes ne tiennent pas compte de la difficulté liée à la nécessité de prédire le phénotype d'un individu dans tous les environnements, alors qu'il n'a été observé (et son spectre collecté) que dans un ensemble d'environnements restreint.

## Sujet

Le(la) stagiaire contribuera à la mise en place et l'évaluation de deux méthodes permettant de reconstituer l'information sur les spectres NIRS manquants à partir de ceux disponibles :

- une première méthode basée sur une approche de type auto-encodeur associée à une perte garantissant la cohérence entre les spectres d'un même individu collectés dans différents environnements (*cycle consistency loss* [2]),
- une deuxième méthode où le réseau d'auto-encodage est remplacé par un modèle de diffusion probabiliste [3].

Les deux modèles seront appliqués à l'étude du déterminisme génétique de caractères d'intérêt chez la vigne. Les jeux de données portent sur un panel d'environ 300 variétés de vignes pour lesquelles le spectre NIRS et différents phénotypes d'intérêt ont été partiellement obtenus pour différents organes (feuille, raisin ou tige), dans des conditions contrastées (présence ou absence de stress hydrique, plantes en serre ou au champ), et sur différentes années d'expérimentation. Des premiers modèles de prédiction phénotypique ont été établis à partir de spectres NIRS et de caractères collectés au vignoble [4]. Ici, on recherchera spécifiquement à prédire ces caractères évalués au vignoble à partir de spectres NIRS collectés sur plantes en pots dans diverses conditions.

## Compétences requises

Ce stage s'adresse à un(e) étudiant(e) de Master 2 dans l'un de ces domaines : informatique, statistiques ou apprentissage automatique. Une maîtrise de la programmation en Python et/ou en R, une expérience de travail avec de grands ensembles de données, ainsi qu'un intérêt pour les applications en biologie, et plus particulièrement en génétique sont requis.

## Environnement de travail

Le(la) stagiaire sera financé(e) par le projet DeepPhenomics du métaprogramme DIGITBIO d'INRAE. Vous travaillerez dans l'équipe SOLsTIS de l'unité MIA Paris-Saclay, située à AgroParisTech (Palaiseau), sous la supervision de Julie Aubert, Hugo Gangloff et Tristan Mary-Huard, en collaboration avec

Vincent Segura (INRAE) pour les aspects de génétique de la vigne.

La durée du stage envisagée est de 5 à 6 mois, avec une date de début comprise entre février et avril 2024 suivant la disponibilité de l'étudiant(e).

### **Contact**

Les candidat(e)s intéressé(e)s doivent postuler en envoyant un CV et une lettre de motivation à :

julie.aubert@inrae.fr

hugo.gangloff@inrae.fr

tristan.mary-huard@agroparistech.fr

### **Références**

[1] Rincent, R., Charpentier, J. P., Faivre-Rampant, P., Paux, E., Le Gouis, J., Bastien, C., and Segura, V. (2018). Phenomic selection is a low-cost and high-throughput method based on indirect predictions: proof of concept on wheat and poplar. *G3: Genes, Genomes, Genetics*, 8(12), 3961-3972.

[2] Zhu, J.Y., Park, T., Isola, P., and Efros., A. A. (2020). Unpaired Image-to-Image Translation using Cycle-Consistent Adversarial Networks". *IEEE International Conference on Computer Vision (ICCV)*, 2017.

[3] Ho, J., Jain, A., and Abbeel, P. (2020). Denoising diffusion probabilistic models. *Advances in Neural Information Processing Systems*, 33, 6840-6851.

[4] Brault, C., Lazerges, J., Doligez, A., Thomas, M., Ecarnot, M., Roumet, P., ..., and Segura, V. (2022). Interest of phenomic prediction as an alternative to genomic prediction in grapevine. *Plant methods*, 18:108.