

Projet de recherche – Programme interdisciplinaire

Demande de financement Ra&D

Interdisciplinarité (Indiquez ici les deux ou trois domaines scientifiques touchés démontrant l'interdisciplinarité du projet)

Agronomie, Biologie moléculaire, Bioinformatique, Machine Learning

Titre du projet en français (1-2 lignes, destinées à la publication)

Séquençage à ultra-haut débit, bioinformatique et Machine Learning comme outil de Diagnostic de la santé des Sols

Acronyme

MaLDiveS (Machine Learning Diagnostic Sol)

Mots clés en anglais (séparés par un point-virgule)

Bioinformatic; Bioindicators; Computational biology; Environmental diagnostic; High-throughput sequencing; Illumina MiSeq; Machine learning; Microorganisms; Pesticides; Protist; Soil; Sustainable agriculture; Viticulture

Résumé

(3'000 caractères maximum espaces compris, destinés à la publication, sans photographie ni annexe).

Le but général de notre projet interdisciplinaire est de concevoir **une nouvelle méthode et un prototype d'outil de diagnostic permettant d'évaluer l'impact des pesticides et du système d'entretien du sol sur la santé et la qualité des sols de vigne, basé (i) sur des données de séquençage à ultra-haut débit et (ii) sur leur traitement par des méthodes prédictives de machine learning.**

En effet l'impact des modes d'exploitation agricoles et des pesticides sur la qualité et la santé des sols est une préoccupation croissante chez les consommateurs, exploitants et gestionnaires des sols. Pour évaluer cet impact, des bioindicateurs tels que les protistes ont un grand potentiel, mais leur utilisation reste limitée car les méthodes actuelles ne permettent pas d'analyser les échantillons de sol de manière détaillée et rapide. Pour surmonter ces inconvénients, l'identification des espèces basée sur des séquences d'ADN («code-barres») couplée aux nouvelles techniques de séquençage à ultra-haut débit représente une approche prometteuse, mais l'énorme quantité de séquences et leur grande complexité rend difficile leur traitement par des moyens conventionnels. Il devient donc essentiel de développer des méthodes mariant bioinformatique et machine learning pour (i) quantifier, analyser et traiter les séquences de protistes; (ii) identifier et sélectionner les bioindicateurs (un sous ensemble de protistes (OTUs)) associés à différents facteurs de stress; mais également de (iii) modéliser leur abondance relative en fonction des différentes conditions, menant ainsi à la construction de modèles diagnostiques. Pour ce projet, nous proposons de développer une approche de biomonitoring dans les sols des vignes basée sur la quantification du metabarcoding de protistes et sur le pouvoir prédictif du machine learning.

Les données servant de base à ce développement proviendront, d'une part, d'une expérience menée en laboratoire où l'impact d'un mélange de pesticides, la température et la teneur en eau du sol sur les communautés de protistes seront testés dans des microcosmes et d'autre part, d'une étude de terrain où l'impact des facteurs environnementaux sur les communautés de protistes sera évalué dans un réseau de 33 vignobles valaisans.

Le caractère interdisciplinaire de ce projet devrait fournir des résultats d'une grande valeur scientifique et permettre le développement d'un prototype de diagnostic environnemental. L'étape suivante consistera à généraliser l'utilisation de cet outil à d'autres sols de vignes puis de l'étendre à d'autres cultures. Des collègues et entreprises telle que la startup genevoise ID-Gene LTD nous ont déjà indiqué leur volonté de collaborer avec nous pour le développement d'un tel projet.

Requérant-e-s (La première personne est le requérant principal – un requérant par institut)

Nom	Prénom	Email	Ecole – Institut (acronymes)
Heger	Thierry	thierry.heger@changins.ch	HES-SO Changins
Peña	Carlos Andrés	carlos.pena@heig-vd.ch	HEIG-VD

Début du projet	Novembre 2017	Durée	24 mois
------------------------	---------------	--------------	---------

Compétences clés des requérants	
Nom Prénom	Trois à cinq compétences clés (en anglais)
Heger Thierry	<ul style="list-style-type: none"> • Microbial ecology • Biomonitoring • High-throughput environmental sequencing
Peña Carlos Andrés	<ul style="list-style-type: none"> • Machine learning • Bioinformatics • Computational biology

1. Contexte du projet

Quelle est la problématique ?

Une préoccupation croissante chez les consommateurs, ainsi que chez les exploitants et les gestionnaires des sols est l'impact des modes d'exploitation agricoles et des pesticides sur la santé et la qualité des sols. Pour évaluer la qualité des sols, les organismes des sols peuvent être utilisés comme bioindicateurs. Parmi les différents groupes d'organismes des sols, les protistes (protozoaires et microalgues) représentent de bons candidats. Abondants et diversifiés, ils sont aussi très sensibles aux perturbations naturelles et anthropogènes (Foissner 1999; Pawlowski et al. 2016). Cependant, l'identification des protistes (Fig. 1), traditionnellement basée sur des analyses morphologiques, est problématique car chronophage, coûteuse, et requérant la compétence de spécialistes. De plus, dans les sols, l'identification au microscope est souvent perturbée car les protistes sont souvent masqués par des particules minérales et organiques.

Pour surmonter ces inconvénients, l'identification des espèces basée sur des séquences d'ADN («code-barres») couplée aux nouvelles techniques de séquençage à ultra-haut débit (=metabarcoding) représente une approche très prometteuse. Cette approche a déjà été validée pour le diagnostic environnemental en milieu aquatique (Apothéloz-Perret-Gentil et al. 2017) mais n'a, par contre, pas encore été implémentée dans les sols agricoles.

Néanmoins, l'énorme quantité de séquences et leur grande diversité rend difficile leur traitement par des moyens conventionnels. De plus, les outils statistiques conventionnels peuvent s'avérer limités pour l'identification de groupes de bioindicateurs comportant des interactions complexes. Il devient donc essentiel de développer des méthodes, mariant bioinformatique, statistique et machine learning, pour quantifier, analyser et traiter ces séquences; mais surtout pour modéliser le comportement de leur abondance relative en fonction des différentes conditions. Pour ce projet, nous proposons de développer une approche de biomonitoring dans les sols des vignes basée sur la quantification du metabarcoding de protistes et sur le pouvoir prédictif du machine learning.

Les sols de vignes (Fig. 1) représentent un système agricole idéal pour tester une telle approche car l'impact environnemental de la viticulture est relativement élevé. Presque 30% de la quantité des pesticides utilisés en Suisse sont dédiés à la viticulture alors que le vignoble représente moins de 2% de la surface agricole totale. En outre, il y a une réelle volonté provenant de nombreux viticulteurs, d'agences environnementales et de la Confédération de réduire l'impact environnemental des modes d'exploitations sur les sols de vigne et les autres écosystèmes touchés. Par exemple, le plan d'action sur les produits phytosanitaires en cours d'élaboration de la Confédération vise à réduire de 50% l'utilisation de produits phytosanitaires en agriculture.

Dans ce contexte, où des efforts considérables sont réalisés dans le but de développer des modes d'exploitation durables en viticulture et dans d'autres types de cultures, il est urgent de développer des outils de diagnostic environnementaux, complémentaires aux analyses physico-chimiques de routine, afin d'évaluer la qualité des sols de vigne de manière détaillée, rapide et relativement peu onéreuse.

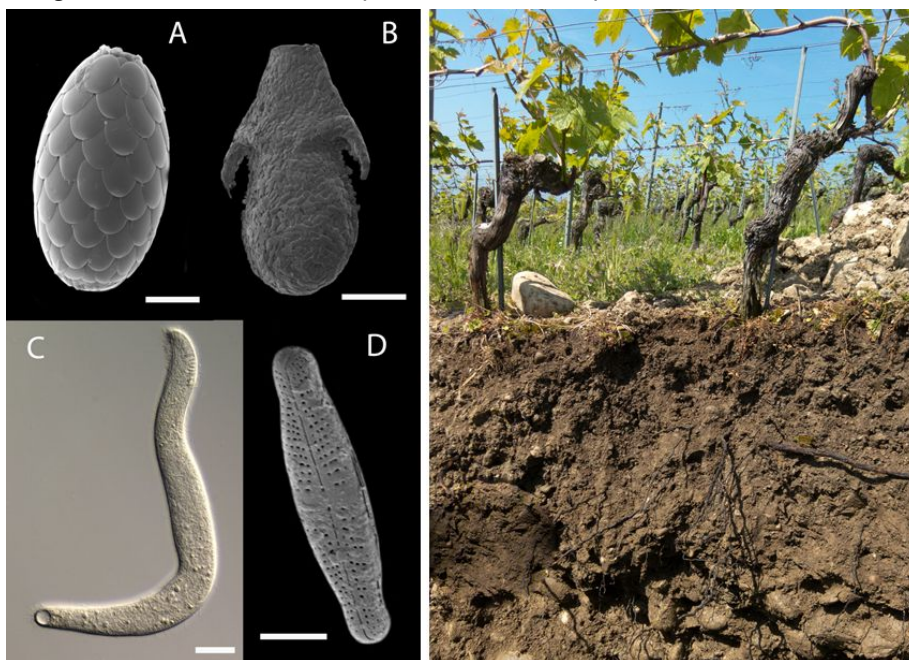


Figure 1: A gauche, images de microscopie optique et électronique illustrant quatre différents groupes de protistes des sols : A) amibe à thèque du groupe des euglyphidés, B) amibe à thèque du groupe des arcellinidés, C) cilié et D) diatomée, images: T. Heger and W. Bourland. A droite, vigne et profil du sol, photo: M. Motta (Changins).

Changins et la HEIG-VD ont décidé d'allier leurs expertises pour mener à bien ce projet qui possède un grand potentiel d'application et nécessite des compétences interdisciplinaires. Des collaborations avec un laboratoire de l'Université de Fribourg et un laboratoire de l'Université de Strasbourg sont également prévues. Ces deux laboratoires mettront à disposition un très grand nombre d'échantillons et de données nécessaires à la réalisation du projet MaLDiveS.

2. Objectifs

Quels sont les objectifs du projet (scientifiques, techniques) ?

MaLDiveS est un projet interdisciplinaire qui vise à développer une nouvelle méthode **de biomonitoring, basée sur des données de séquençage à ultra-haut débit et sur leur traitement par des méthodes de machine learning, permettant d'évaluer l'impact du système d'entretien du sol sur la santé et la qualité des sols de vignes.**

Pour cette étude, deux jeux de données complémentaires seront analysés :

- Au laboratoire, l'effet d'un mélange de 20 pesticides (herbicides et fongicides) sera analysé, dans des microcosmes, avec des conditions contrôlées de température et de teneur en eau du sol.
- Sur le terrain, l'effet du mode d'entretien du sol et d'autres facteurs environnementaux seront évalués dans 33 vignobles valaisans.

Les objectifs spécifiques du projet sont les suivants:

1. Faire un état des lieux de la santé de différents sols viticoles valaisans, en se basant sur la vie microbiologique, et en particulier celle des protistes.
2. Concevoir et appliquer des méthodes de machine learning permettant d'identifier et de caractériser, sur la base de l'abondance relative des séquences des différents taxa, des bioindicateurs servant à évaluer l'impact des pesticides et d'autres facteurs de stress sur les sols de vigne.
3. Implémenter un prototype d'outil informatique capable de traiter les données de séquençage (détecter et quantifier les bioindicateurs identifiés) dans l'objectif de proposer un diagnostic environnemental.

L'acquisition des données se fera avec la collaboration de deux instituts de recherche

Les échantillons de sol de l'expérience de laboratoire nous seront mis à disposition par l'[équipe du Dr. Gwenaél Imfeld](#) du laboratoire d'HYdrologie et de GeoChimie de Strasbourg (LHyGes), une Unité Mixte de Recherche entre le CNRS et l'Université de Strasbourg. Ces échantillons découlent d'une étude dont le but était d'analyser la dynamique de dégradation d'un mélange de pesticides de synthèse en fonction de différentes conditions de température et teneur en eau du sol. En effectuant des analyses moléculaires des protistes, il nous sera possible d'évaluer les liens entre les pesticides, les composés de dégradations des pesticides et les communautés de protistes dans différentes conditions de température et d'humidité.

Les **données provenant des vignobles valaisans** nous seront mises à disposition par [l'équipe du Prof. Sven Bacher](#) de l'Université de Fribourg. Dans le cadre du projet européen PromESSinG ([Promoting EcoSystem Services in Grapes](#)), cette équipe mène une ambitieuse étude ayant pour but d'identifier des modes de gestions qui favorisent la biodiversité et réduisent les intrants externes. Leur projet inclut l'analyse des communautés de microorganismes et des protistes avec l'aide de séquençage à ultra-haut débit et la récolte de nombreuses métadonnées susceptibles d'aider à comprendre l'influence du mode d'exploitation (y compris l'utilisation de pesticides) sur les communautés de protistes. Comme les objectifs de notre projet sont complémentaires à ceux du Prof. Bacher, il a répondu très favorablement à notre demande de collaboration. En conséquence, les données moléculaires des communautés de protistes ainsi que les métadonnées seront disponibles dès le début du projet.

La collaboration entre le consortium de la HES-SO et les laboratoires du Dr. Imfeld et Bacher sera extrêmement précieuse. En effet, dans le cadre de ce projet interdisciplinaire de 24 mois, nous n'aurions ni le temps, ni les ressources financières pour générer une telle quantité de données par nous-mêmes. Il sera aussi extrêmement bénéfique de pouvoir interagir avec ces deux équipes possédant des compétences complémentaires aux nôtres. Le Prof. Bacher est un biologiste de renom, expert dans les domaines de l'écologie du paysage, de la lutte biologique et des interactions entre espèces. Le Dr Imfeld est un spécialiste de la

dégradation et la toxicité des pesticides dans les sols de vigne. Il est important de signaler que les Dr Imfeld et Heger ont déjà collaboré ensemble avec succès. Ils ont notamment co-édité, avec le Prof. Mitchell, un [numéro spécial sur la bioindication](#) dans les sols pour le “*European Journal of Soil Biology*”.

3. État de l'art

3.1 État de la recherche dans le domaine des travaux projetés avec mention des principales réalisations / publications.

3.1.1 Les protistes et leur utilisation comme bioindicateurs

Les protistes (par ex. amibes, ciliés et diatomées) sont des organismes unicellulaires présents parmi tous les principaux groupes d'eucaryotes. Environ 200'000 espèces de protistes ont été décrites jusqu'à présent mais leur diversité totale est probablement bien plus élevée. Les protistes sont utilisés avec succès comme bioindicateurs dans les écosystèmes terrestres et aquatiques. Outre le fait qu'ils répondent rapidement aux changements environnementaux, les protistes ont l'avantage d'être abondants, diversifiés et ont de larges aires de distributions. Dans les milieux aquatiques, les protistes sont couramment utilisés pour évaluer la qualité des eaux de rivière et des lacs. Par contre, dans les sols, même si de nombreuses études ont montré qu'ils répondent à de nombreux stress tels que les pesticides ou le mode d'exploitations agricole (Foissner 1999), (Heger et al. 2012), ils sont rarement intégrés dans des projets de biomonitoring de grande ampleur, probablement car les méthodes d'identification classiques des protistes des sols sont particulièrement longues et fastidieuses (Geisen et al. 2017).

3.1.2 Le metabarcoding représente un outil puissant pour l'analyse des communautés de protistes

L'identification des espèces basée sur des séquences d'ADN («code-barres») couplée aux nouvelles techniques de séquençage à ultra-haut débit (metabarcoding) fournit un nouvel outil pour identifier les protistes. En produisant plusieurs millions de séquences par run, les techniques de séquençage à ultra-haut débit tel qu'Illumina MiSeq, représentent une approche puissante pour identifier de manière rapide, peu onéreuse et systématique la diversité des protistes. Ces nouvelles approches sont en train de révolutionner notre perception de la diversité des protistes et ouvrent de nouvelles perspectives dans le domaine du diagnostic environnemental. Par exemple, le prof. Pawlowski et son équipe du laboratoire de systématique moléculaire et la start up genevoise [ID-Gene](#) ont récemment réussi à établir [un index de qualité des eaux de rivières genevoises basé uniquement sur les séquences d'ADN de diatomées](#). Cette étude a montré qu'il est possible d'analyser un grand nombre d'échantillons en peu de temps et pour un faible coût (Apothéloz-Perret-Gentil et al. 2017). Une telle approche n'a par contre pas encore été testée dans des sols agricoles.

3.1.3 Application du machine learning dans la découverte et modélisation de bioindicateurs

Le Machine Learning (apprentissage automatique) est une discipline scientifique centrée sur le développement, l'analyse et l'implémentation de méthodes prédictives automatisables offrant la possibilité à une machine d'acquérir des comportements grâce à un algorithme d'apprentissage. Par rapport aux méthodes statistiques usuelles, le machine learning peut potentiellement atteindre des niveaux de prédiction et d'adaptation supérieures dans des contextes complexes. Ces méthodes sont de plus en plus utilisées dans différents domaines pour interpréter le nombre important de données qu'il est possible d'obtenir aujourd'hui.

Des techniques de machine learning ont déjà été appliquées pour l'identification et la caractérisation d'Unités Taxonomiques Opérationnelles (OTUs, d'après l'anglais Operational Taxonomic Unit). Une OTU désigne un regroupement d'organismes, phylogénétiquement proches, regroupés sur la base de la similarité de la séquence d'ADN d'un gène donné, qui sert de marqueur taxonomique utilisées pour regrouper des individus. Dans le domaine de la santé, par exemple, des OTUs ont été identifiés comme biomarqueurs de maladies du système digestif (Tedjo et al. 2016), (Saulnier et al. 2011) et des modèles prédictifs ont été construits pour estimer le risque de développer une maladie (Ai et al. 2017) ou pour la caractériser plus précisément (Ganz et al. 2017). Parmi différentes approches, testées dans plusieurs études, l'algorithme Random Forest (basé sur des arbres de décision) a permis d'obtenir des très bonnes performances prédictives basées sur l'abondance relative des OTUs.

Par exemple, cet algorithme a permis, en se basant sur les communautés bactériennes (16S), de détecter les sites contaminés par des déchets nucléaires (Smith et al. 2015) (Fig. 2). Cette étude, ainsi que d'autres récents travaux nous persuadent qu'en analysant les communautés de protistes du sol avec des méthodes de machine

learning, il sera également possible d'identifier des bioindicateurs, et de réaliser des prédictions sur la qualité des sols de vigne.

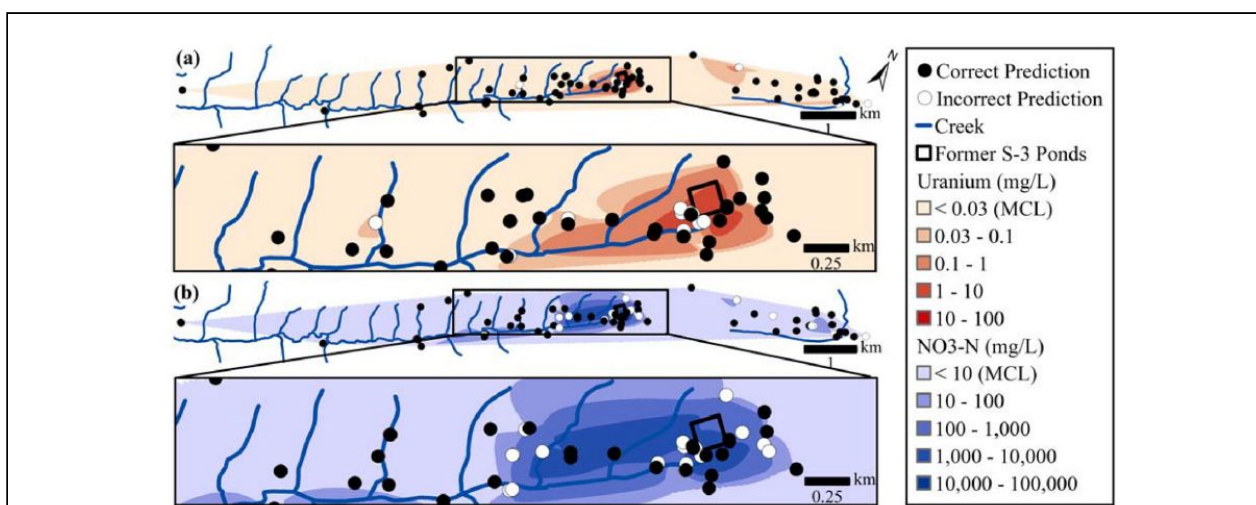


Figure 2 : Résultats de la prédiction de la contamination en uranium et en nitrate des puits à partir d'un RandomForest basé sur l'abondance relative des OTUs bactériens (Smith et al. 2015).

3.2 État des recherches effectuées par le(s) requérant(s) dans le domaine des travaux projetés avec mention des principales réalisations / publications.

3.2.1 Prof. Thierry Heger, PhD, Haute Ecole Spécialisée Suisse occidentale (Changins)

Depuis janvier 2017, le Prof. Thierry Heger est responsable du laboratoire des sciences du sol de Changins qui comprend un total de quatre collaborateurs. Il a plus de 10 ans d'expérience dans le domaine de l'écologie microbienne, la diversité des protistes et de la bioindication. Il est familier avec différentes techniques moléculaires incluant les méthodes de séquençage massif pour l'analyse de la composition et les dynamiques des communautés microbiennes. Les résultats de ses recherches ont été publiés dans plus de 25 articles présentés lors de nombreuses conférences et symposiums.

Le Prof. Heger a obtenu une thèse de doctorat en 2010 dans le laboratoire des systèmes écologiques (ECOS) de l'EPFL où il a notamment évalué l'impact des différents modes d'exploitations agricoles sur les communautés d'amibes et d'algues des sols de grandes cultures (Heger et al. 2012). Il a ensuite mené une expérience professionnelle de plus de cinq ans au Canada en tant que post-doctorant puis associé de recherche à l'Université de Colombie Britannique. En collaboration avec l'équipe multidisciplinaire « [Kwakshua Watershed Program](#) », il a notamment contribué à mieux comprendre les relations entre organismes du sol, végétation et fonctionnement des écosystèmes. Depuis son retour en Suisse en juin 2016, le Prof. Heger a travaillé quelques mois dans le laboratoire de la biodiversité du sol à l'Université de Neuchâtel avant d'obtenir le poste de professeur des sciences du sol à Changins.

Durant sa carrière, Thierry Heger a obtenu plusieurs bourses et subventions pour financer sa recherche (bourse FNS de recherche pour chercheurs débutants (PBELP2-122999), bourse FNS de recherche pour chercheurs avancés (PA00P3 145374), fonds de la fondation canadienne TULA) et a participé à la co-supervision de plusieurs étudiants.

Les connaissances de Thierry Heger dans le domaine de la microbiologie des sols ont été également reconnues au niveau éditorial. En 2010-11, il a servi comme éditeur invité pour [un numéro spécial sur la bioindication dans les sols](#) dans la revue scientifique « [European Journal of Soil Biology](#) » puis il a été nommé éditeur associé de ce journal pour une période de plus de trois ans (2013- début 2016). Son intérêt pour la recherche appliquée dans le domaine des sciences du sol peut aussi s'illustrer par l'organisation d'un workshop en 2010 à l'Université de Neuchâtel sur la bioindication des sols qui a réuni de nombreux étudiants, chercheurs et collègues impliqués dans le suivi de l'état de la qualité des sols en Suisse.

A Changins, le Prof. Heger peut compter sur une solide équipe ayant plusieurs années d'expérience dans le domaine des sols de vigne et possédant d'excellents contacts avec les praticiens et les responsables de la gestion des sols. Dorothea Noll est spécialisée dans le domaine de la cartographie des sols, de l'érosion et de la problématique des pesticides dans les sols. Frédéric Lamy un spécialiste de la physique des sols et Matteo Mota est un expert de la flore viticole et étudie l'influence du modes d'entretien du sol sur la flore et le fonctionnement du sol. Tous les membres du groupe sol de Changins seront impliqués dans ce projet à des

degrés divers.

3.2.2 Prof. Ordinaire Carlos Peña, PhD, Haute Ecole Spécialisée Suisse occidentale (HEIG-VD).

Le Professeur Carlos Peña dirige l'Axe des applications biomédicales du département TIC qui appartient au groupe Health Engineering and Economics (HE&E). Il est également responsable de l'équipe Computational Intelligence for Computational Biology (CI4CB, <http://iict-space.heig-vd.ch/cpn/>) affiliée, depuis 2016, à l'Institut Suisse de Bioinformatique (ISB, <http://www.sib.swiss/pena-reyes-carlos-andres/>). La recherche menée par le groupe CI4CB, se concentre sur le développement de techniques de machine learning et d'intelligence artificielle pour la résolution computationnelle de problèmes complexes en sciences de la vie, dans des domaines aussi divers que l'agriculture, l'électrophysiologie, la bioinformatique et l'ingénierie biomédicale.

L'équipe du Prof. Peña possède une forte expertise dans l'analyse de données et la modélisation prédictive (supervisée, semi-supervisée et non supervisée), la découverte de biomarqueurs, le diagnostic clinique et l'aide à la prise de décision. L'équipe a développé des compétences particulièrement fortes dans les modèles dits de 'boîte grise' qui combinent un pouvoir prédictif élevé à un pouvoir explicatif élevé, et qui se révèlent particulièrement importants pour la résolution de problèmes biologiques et cliniques. Certaines de ses recherches récentes incluent:

- Entre 2009 et 2014, plusieurs projets impliquant: (i) une analyse basée sur des modèles d'une très grande quantité de données électrophysiologiques (PharMEA, projet européen); (ii) le développement de modèles prédictifs de la concentration de médicaments dans le sang permettant la personnalisation optimale du dosage du médicament (ISyPeM, nano-tera); (iii) des méthodes de modélisation floue pour sélectionner et caractériser un pool de biomarqueurs comme outil de diagnostic de 18 sous-types de leucémies (nanoFUGE, financés par le SNSF) et (iv) le développement d'une méthodologie brevetée et exclusive dédiée à la découverte, au développement, au déploiement et à l'exploitation de biomarqueurs diagnostiques (DiagnoSuite, financé par CTI).
- Depuis 2015, l'équipe participe au projet Européen Eurostars-2 "Fishguard: Improving fish viral diseases monitoring through the development of fast, cost-effective and infield screening tests", en partenariat avec deux PME européennes. L'équipe contribue par son expertise en bioinformatique et machine learning à la découverte et à la caractérisation de biomarqueurs. (<http://www.fishguard-eurostars.eu>)
- Depuis 2016, l'équipe participe au projet FNS «INPHINITY: In silico prediction of phage-bacteria infection networks as a tool to implement personalized phage therapy» (2016-19), menée en collaboration avec l'Université de Lausanne et l'Inselspital de Bern. L'équipe développe des modèles computationnels (méthode supervisée et semi-supervisée) pour prédire et mieux comprendre les réseaux d'infections phages-bactéries en fonction de l'information contenue dans leurs génomes (Carvalho Leite et al. 2017).
- Le projet CTI "BOSS Explorer: Development and integration of new methods for exploration, selection, and visualization of diagnostic signatures" (2016-17) en collaboration avec SimplicityBio spin-off du groupe, vise à développer de nouvelles méthodes de recherche et de visualisation pour améliorer les capacités de découverte et d'interprétation de biomarqueurs au sein de la plateforme de SimplicityBio.
- Finalement, «D-Rex: Rule extraction from deep neural networks» (2016-18), est un projet exploratoire financé par la Fondation Hasler. Il a pour objectif d'explorer et de développer des méthodes permettant de comprendre, de hiérarchiser et d'expliquer la connaissance acquise par un réseau de neurones profonds, en les représentant sous forme de règles logiques.

4. Descriptif de la recherche proposée

Quelles sont les méthodes de recherche envisagées et/ou l'innovation recherchée dans le cadre du projet ?

Le but du projet MaLDiveS est de développer un prototype d'outil de diagnostic environnemental basé sur des séquences moléculaires de protistes traitées et analysées avec des méthodes de machine learning.

Ce projet nous permettra de mieux comprendre l'impact des pesticides et des autres facteurs environnementaux sur les communautés de protistes puis d'identifier des bioindicateurs associés à différents stress environnementaux et d'y caractériser le comportement de leur abondance relative, menant ainsi à la construction de modèles diagnostiques.

Les données servant de base à ce développement proviendront, d'une part, d'une expérience menée en laboratoire où l'impact d'un mélange de pesticides, la température et la teneur en eau du sol sur les

communautés de protistes seront testés dans des microcosmes et d'autre part, une étude de terrain sur un réseau de 33 vignobles valaisans, permettra d'évaluer l'impact des facteurs environnementaux sur les communautés de protistes. L'étude en laboratoire et celle sur le terrain sont complémentaires mais pas interdépendantes. Ainsi, leur analyse pourra être menée de façon séparée et non bloquante fournissant, les deux, des résultats d'une grande valeur scientifique.

L'un des principaux défis de cette étude est de pouvoir prendre en compte l'énorme abondance de protistes ainsi que leur grande complexité, facteurs qui rendent difficile leur traitement par des moyens conventionnels. Il devient, donc, essentiel de développer des méthodes pour (i) quantifier, analyser et traiter les séquences de protistes; (ii) identifier et sélectionner les bioindicateurs (un sous ensemble de protistes (OTUs)) qui reflètent la qualité des sols de vigne; mais également de (iii) modéliser leur abondance relative en fonction des différentes conditions. Pour ce faire, les deux jeux de données seront traités et analysés de manière similaire avec des méthodes de machine learning.

La figure 3 décrit schématiquement (Partie **A**) la chaîne d'acquisition et de traitement des données menant à l'identification des bioindicateurs et à la création de modèles prédictifs, ainsi que (Partie **B**) le fonctionnement ultérieur de l'outil diagnostique dont le prototype sera réalisé et qui intègre les deux produits principaux du projet: les bioindicateurs identifiés et le(s) modèle(s) prédictif(s).

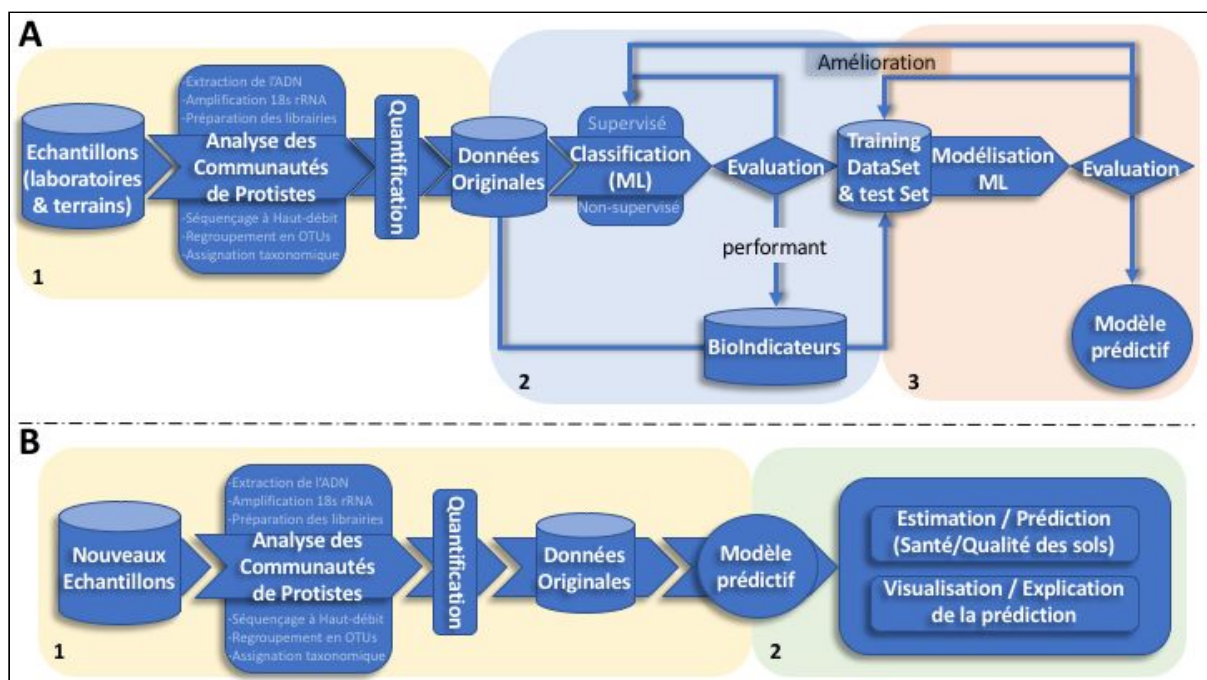


Figure 3: A. Schéma du processus de sélection des bioindicateurs et de création du modèle prédictif (1. Extraction des données, 2. Apprentissage et sélection des features, 3. Création et sélection du modèle prédictif) **B.** Schéma du fonctionnement du prototype envisagé (1. Extraction des données, 2. Utilisation du modèle établi pour estimer la qualité des sols)

Nos hypothèses sont les suivantes :

H1 : Dans les microcosmes, les communautés de protistes sont fortement influencées par le mélange de pesticides. Des bioindicateurs associés à un taux élevé de pesticides et peu sensibles aux conditions climatiques peuvent être identifiés.

H2 : Dans les 33 vignobles valaisans étudiés, une diversité de protistes très élevée est observée dans les vignes où peu de pesticides sont utilisés et où le sol est enherbé.

H3 : L'expérience de laboratoire valide les résultats du terrain. Les bioindicateurs associés aux pesticides de synthèse sont les mêmes sur le terrain et au laboratoire. De plus, l'étude de terrain permet d'identifier des bioindicateurs associés à d'autres stress naturels et anthropogènes tel que la concentration en cuivre du sol.

H4 : Pour le traitement des grands jeux de données de protistes, les méthodes de sélection de variables, issues du machine learning mais aussi de la statistique, permettent d'identifier et de sélectionner des bioindicateurs de manière rapide et sensible.

H5 : L'abondance relative des communautés de protistes permettent de prédire quantitativement

certaines facteurs du sol (Cu, pH, rapport C/N, fongicides et herbicides de synthèse).

Les deux jeux de données de ce projet MaLDIveS proviendront de :

I) Expérience de laboratoire dans des microcosmes (collaboration avec le Dr. [Imfeld](#), Université de Strasbourg)

Cette expérience permettra d'évaluer l'impact d'un mélange de 20 pesticides de synthèse (herbicides et fongicides) sur les communautés de protistes du sol et de comprendre comment les communautés évoluent en fonction du temps et de la dégradation des pesticides dans trois conditions différentes de température et de la teneur en eau du sol. Les sols proviennent de l'observatoire viticole de Rouffach en Alsace (France) et sont bien caractérisés et contextualisés (Babcsányi et al. 2016). Ce travail représente une extension de l'expérience menée par le G. Imfeld et sa doctorante Fatima Meite dans lesquelles les cinétiques et voies de dégradation des pesticides en fonction du temps ont été caractérisée par des méthodes de chimie analytique (GCMS, LCMS) et isotopique (GC-C-IRMS).

Brièvement, le protocole expérimental de l'équipe du Dr. Imfeld était le suivant. Du sol a été prélevé dans un vignoble alsacien puis incubé jusqu'à 200 jours dans des microcosmes (récipients confinés permettant les échanges d'oxygène) dans des conditions de températures et teneur en eau différentes (Fig. 4). Le mélange de pesticides utilisés pour cette expérience incluait des herbicides et fongicides communément utilisés en viticulture aussi bien en France qu'en Suisse. Lors de cette expérience, des sous-échantillons ont été prélevés puis conservés dans des conditions appropriées afin de pouvoir caractériser les communautés de microorganismes à posteriori.

Les résultats préliminaires de cette étude ont montré des cinétiques de dégradation très variable en fonction des pesticides tandis que les conditions de température et d'humidité du sol jouent un rôle apparemment secondaire.

Dans le cadre du projet MaLDIveS, les communautés de protistes de 108 échantillons provenant de cette expérience seront analysés (2 traitements (mélange de 20 pesticides vs sans pesticides) * 3 réplicats * 6 temps d'incubation (0, 1, 10, 50, 100 et 200 jours) * 3 conditions de température et teneur en eau (Fig. 4). A Changins, l'ADN sera extrait des échantillons puis le fragment V4 du 18S sera amplifié avec des amorces générales pour les eucaryotes. Les produits amplifiés seront ensuite envoyés au Canada (Genome Québec) pour le séquençage (Illumina MiSeq). Les séquences seront regroupées en OTUs (Operational Taxonomic Units) puis l'assignation taxonomique se fera avec la base de référence PR2. Ces traitements bioinformatiques seront réalisées par l'équipe du Dr. Heger avec le soutien de l'équipe du Dr. Peña. Thierry Heger est familier avec ces analyses et il a déjà analysé d'autres jeux de données de manière similaire (Heger et al. 2017). Les échantillons ainsi que les métadonnées provenant de cette expérience seront disponibles dès le début de l'expérience (voir lettre d'intention du Dr. Imfeld).

Afin de s'assurer que les données du laboratoire et du terrain soient parfaitement comparables, les différentes étapes bioinformatiques mentionnées ci-dessus se feront selon le protocole développé dans le projet PromESSinG (voir ci-dessous).

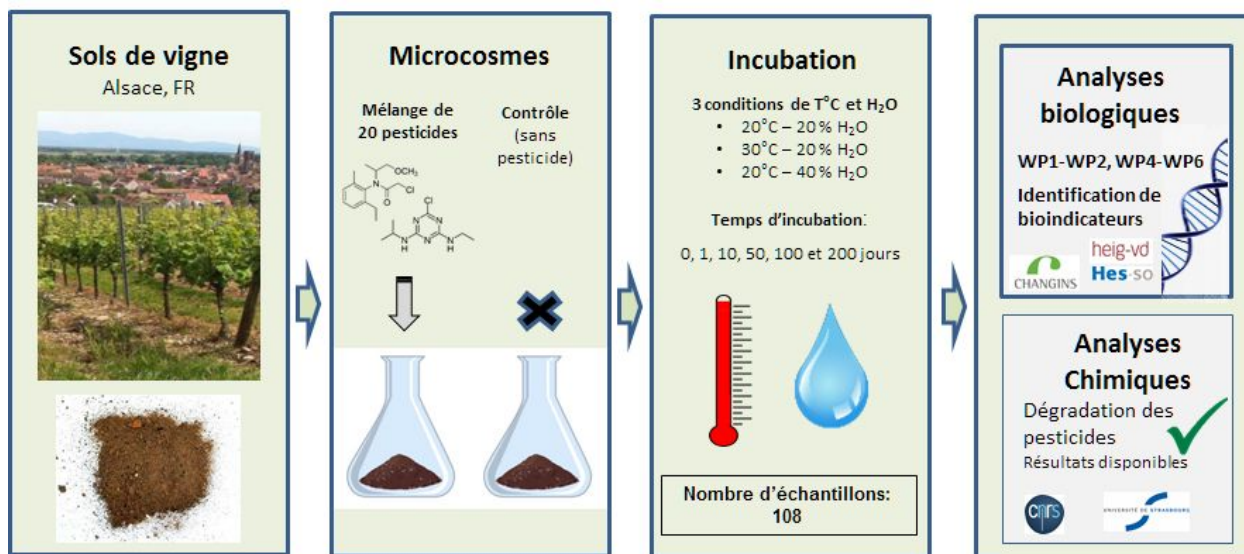


Figure 4: Schéma décrivant les principales étapes de l'expérience de laboratoire menées par l'équipe du Dr. Imfeld. Les analyses chimiques permettant de quantifier la dégradation des pesticides en fonction de la teneur en eau et de la température ont déjà été réalisées.

II) Etude de terrain réalisée dans un réseau de 33 vignobles en Valais (collaboration avec l'équipe du Prof Sven Bacher UniFR, du projet PromESSinG)

Un important nombre d'échantillons et de métadonnées ont été récoltés dans 33 vignobles durant trois années successives. Dans chaque vignoble, deux échantillons (provenant chacun d'un mélange de 8 sous-échantillons) ont été collectés entre 2015 et 2017, juste après la floraison des vignes. Le nombre total d'échantillons récolté s'élève à 330 (33 vignobles * 2 réplicats * 5 (sous les vignes et dans l'inter-rang en 2015 et en 2016 et sous les vignes uniquement en 2017, voir Figure 5). L'ADN est extrait avec un kit MoBio puis le fragment V4 du 18S est amplifié avec des amorces générales pour les eucaryotes puis séquencé avec la plateforme Illumina MiSeq. Les amplifications, séquençage, puis les analyses bioinformatiques pour produire les données annotées sont réalisées par Genome Québec (Canada).

De nombreuses métadonnées ont été mesurées dans le cadre de cette étude. Elles seront d'une grande valeur pour interpréter l'influence des facteurs anthropogènes et naturels sur les communautés de protistes. Ces métadonnées incluent des informations concernant les paramètres du sol (matière organique, rapport C/N, azote, pH, argile, limon, sable, carbone total, cuivre, teneur en eau), la végétation (biomasse, diversité et couverture du sol), le fonctionnement du sol (taux de décomposition de la matière organique, respiration du sol) ainsi que le mode d'exploitation (produits phytosanitaires utilisés, bêcheage).

Les données moléculaires des protistes (OTUs) et les métadonnées de 2015 et 2016 sont déjà disponibles. Les échantillons de 2017 ont déjà été récoltés et sont en cours de traitement. **Toutes les données moléculaires de protistes et métadonnées seront disponibles dès le début du projet MaLDiveS à la fin de l'année 2017.**

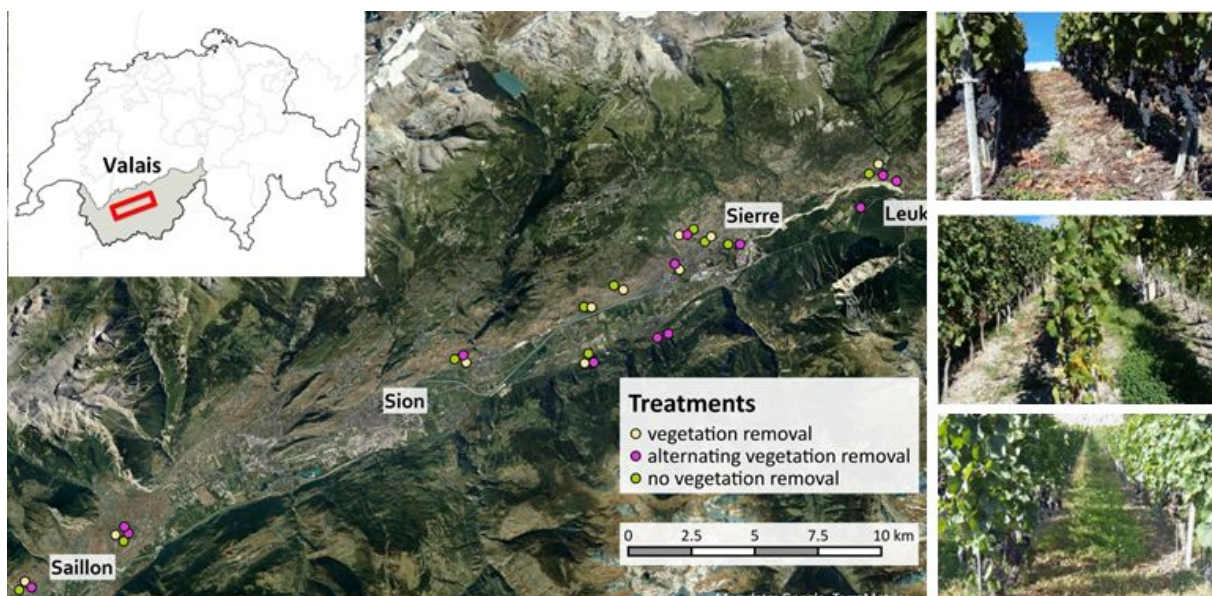


Figure 5: Carte illustrant les 33 vignobles valaisans étudiés par l'équipe du Prof. Bacher dans le cadre du projet européen PromESSinG ([Promoting EcoSystem Services in Grapes](#)). Au total, 330 échantillons ont été collectés durant trois saisons successives (2015-2017) dans des vignobles exploités de trois manières différentes : 1) sols nus, 2) inter-rangs avec alternance de sol nu et enherbé et 3) inter-rangs enherbés (photos sur la droite). Images : M. Steiner, A-L. Fragnière et S. Bacher.

Predictive modelling

En termes généraux, nous concevons une méthode, ou un ensemble de méthodes, d'apprentissage supervisé permettant de construire des modèles prédictifs (classificateurs et/ou estimateurs), pour évaluer l'impact des pesticides et du système d'entretien du sol sur la santé et la qualité des sols de vigne. Ces méthodologies en outre permettront d'identifier et sélectionner les bioindicateurs (un sous ensemble de protistes (OTUs)) associés à différents facteurs de stress (Fig. 3); mais également de modéliser leur abondance relative en fonction des différentes conditions, menant ainsi à la construction de modèles diagnostiques

D'une manière générale, l'apprentissage supervisé consiste à utiliser les caractéristiques d'un ensemble de données connues, validées, annotées et correctement classifiées (également appelé ensemble d'apprentissage) pour inférer des règles permettant de classer des nouvelles données jusque-là inconnues. La classification des objets en plusieurs classes est un problème typique d'apprentissage automatique, où une combinaison "intelligente" de caractéristiques (features), existantes ou extraites des données d'origine, est identifiée par le modèle et utilisée pour discriminer les objets (les cas) entre différentes classes (Fig. 6). Ces caractéristiques (features) transmettent les informations essentielles aux processus de classification, et sont au-moins aussi importantes pour le pouvoir prédictif des modèles d'apprentissage automatique que la qualité et l'annotation des données d'origine.

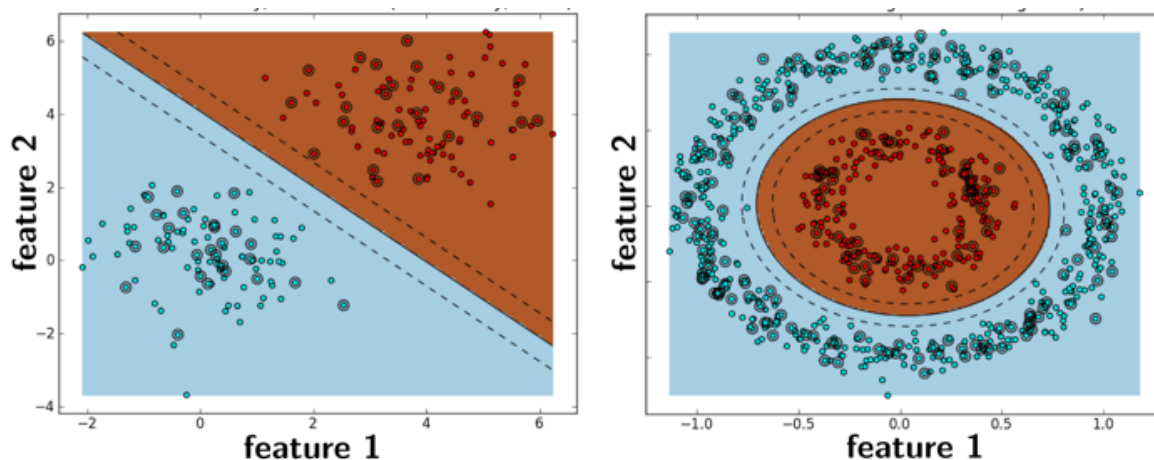


Figure 6: Exemple d'un problème à deux classes avec une séparation linéaire (gauche) et non-linéaire (droite) des données.

Feature selection - Détermination des bioindicateurs

Nous cherchons à évaluer l'impact des pesticides et d'autres facteurs anthropogènes et naturels sur les communautés de protistes des sols en fonction de la biodégradation et de la persistance des pesticides pour développer un système de diagnostic de la qualité des sols de vigne. Dans le cadre de ce projet, le terme "feature" désigne l'ensemble des protistes (OTUs). Les features d'intérêt (= bioindicateurs) désignent un sous-ensemble de protistes qui se voient le plus clairement affectés par les facteurs mentionnés, reflétant ainsi la qualité des sols de vigne. Compte tenu de l'abondance et de la grande diversité des protistes (features), des modèles d'apprentissage automatique prenant compte de toutes les features pourraient être construits. Néanmoins, l'énorme quantité de données rend la modélisation délicate et pourrait, surtout, affecter la qualité de la prédiction. Toutes les features n'ont pas la même importance, certaines s'avèrent essentielles pour une classification précise, alors que d'autres n'ont aucun impact ou très peu. Ainsi, réduire le nombre de features en sélectionnant les plus pertinentes améliorera la précision et la rapidité des systèmes de prédiction. Lors de la 'sélection de feature', étape préliminaire, mais étroitement liée, à la modélisation prédictive proprement dite, nous cherchons à réduire la haute dimensionnalité de nos données en identifiant le sous-ensemble de features (bioindicateurs) le plus pertinent pour la classification.

Les techniques de sélection de features qui ont pour objectif d'obtenir un sous-ensemble des features d'origine sans les modifier ou transformer, préservent la sémantique originale des variables et peuvent être interprétées plus facilement par un expert du domaine. Ces méthodes peuvent aussi bien être non-supervisées (attributs de classe inconnue) que supervisées (connaissance des attributs de classe). Les méthodes non-supervisées peuvent comporter des critères simples, comme par exemple, l'élimination de features avec une variance extrêmement faible et des critères plus complexes allant de l'entropie aux scores laplaciens. Les méthodes supervisées reposent sur une liste de features, un modèle général pour les combiner (mathématiques et/ou computationnelles), et une méthode d'apprentissage automatique pour déterminer l'importance de chaque feature dans la classification des instances de l'ensemble d'apprentissage. Les algorithmes de sélection de features peuvent être classés en trois catégories principales : "filter", "wrapper" et "embedded".

Bien que les étapes de sélection et de modélisation soient étroitement liées, la sélection de features pourrait être considérée comme une étape préliminaire permettant d'obtenir un ensemble de données final (bioindicateurs) pour la phase de modélisation prédictive (voir section Machine-Learning Model ci-dessous). L'ensemble de données sélectionné sera ensuite divisé en un ensemble d'apprentissage et un ensemble de test ou l'on s'assurera que la diversité des données dans les deux sous-ensembles soit respectée. Nous pourrions même réserver un troisième sous-ensemble de données de validation pour être utilisé comme contrôle aveugle, uniquement à des moments cruciaux pour évaluer et améliorer la pertinence du processus de modélisation. Il sera important d'évaluer la performance de notre méthodologie d'une manière statistiquement robuste.

Machine-Learning Models

Sur la base de notre expérience en apprentissage automatique, nous mettrons en œuvre des méthodes de modélisation avancées (p. ex. réseaux de neurones artificiels (ANN), machines à vecteur de support (SVM), forêts aléatoires (RF), logique floue (FL), ou des modèles bayésiens) - pour construire des modèles prédictifs à partir de l'ensemble de données raffinées. La nature des données présente une difficulté particulièrement exigeante: la haute dimensionnalité.

Haute dimensionnalité. Ce défi a déjà été discuté dans la section 'sélection de features' pour la sélection/recherche des bioindicateurs de la santé et qualité des sols de vignes. Cependant, certaines stratégies supplémentaires pourraient également être utilisées spécifiquement lors de la construction de modèles. En utilisant des algorithmes de type ensembliste ou 'Mixture of Experts' (ME) la sélection de features pourrait être appliquée localement, au niveau de la modélisation pour chaque modèle individuel.

Nos modèles prédictifs seront développés en suivant essentiellement deux stratégies : (i) en construisant des prédicteurs indépendants, (modèle multi-classes et/ou modèle de régression), un mélange de produits phytosanitaires étudié (pesticides) ; (ii) en construisant des prédicteurs intégrés (modèle multi-classes et/ou de régression) qui associent tous les produits phytosanitaires dès le début. Afin de maximiser les performances du modèle, nous allons combiner les différents prédicteurs en appliquant des approches appelées apprentissage par ensemble [p. Ex. association algébriques, méthodes basées sur le vote].

L'apprentissage automatique est généralement utilisé pour son pouvoir prédictif, mais il peut également

être utilisé pour expliquer les données. En effet, comme tout autre modèle, il peut être utilisé à la fois pour prédire et pour mieux comprendre les processus sous-jacents. Ainsi, après avoir développé un modèle prédictif robuste et statistiquement significatif, nous veillerons à ce que les caractéristiques (features) explicatives puissent être extraites aisément pour identifier précisément les facteurs déterminants d'une prédiction, soulignant les relations entre les pesticides de synthèse sur les communautés des protistes et l'état de la qualité/santé des sols de vignes.

De ce point de vue, les modèles prédictifs sont généralement décrits comme allant de la boîte blanche (par exemple, les modèles mécanistes), à la boîte noire (p. Ex. ANN), en passant par des niveaux intermédiaires d'opacité ou des modèles dits de boîtes grises. Ces deux derniers types de modèles seront au cœur de ce projet. Les modèles de boîtes noires sont des modèles conçus pour être hautement prédictifs, leur capacité à apprendre et à classer précisément de grandes quantités de données complexes, dépassent les capacités humaines, mais n'offrent aucune explication (modèles ou mécanismes) sous-jacente derrière leurs prédictions et restent donc très opaques. Par conséquent, afin de comprendre les processus impliqués dans leurs décisions, les modèles de boîtes noires devraient être utilisés en association avec d'autres méthodes capables d'en extraire de l'information. Les modèles de boîtes grises sont des modèles dotés d'une représentation schématique de la connaissance. En effet, ces méthodes peuvent définir des règles compréhensibles par l'homme (règles floues ou Booléennes) pour classer les données, qui peuvent ensuite être utilisées pour mieux comprendre les caractéristiques et les interactions qui sont à la base des prédictions. Néanmoins, les modèles de boîtes grises souffrent du compromis nécessaire entre la performance prédictive et le pouvoir explicatif. Pour faire face à ce compromis, nous appliquerons certaines de nos propres méthodes de modélisation floue (Peña-Reyes and Sipper 2003).

Évaluer les performances prédictives

Une validation croisée (p. Ex., k-folding, bootstrapping) sera effectuée afin d'éviter le sur-apprentissage et d'optimiser la sélection du modèle. La performance finale de notre système sera évaluée de manière approfondie à partir d'un test indépendant et des ensembles de validation, extraits des ensembles de données initiaux et déterminés a priori. Il pourrait également être validé sur un tout nouvel ensemble de données composé d'échantillons obtenus et annotés au cours de notre projet et après extraction du jeu de données initial. Les performances prédictives seront évaluées à l'aide de plusieurs mesures (métriques) en fonction du type de modèle et des méthodes d'apprentissage automatique utilisées.

- Les mesures de distance telles que l'erreur absolue moyenne (the Mean Absolute Error), la racine de l'erreur absolue moyenne (Root Mean Absolute Error), l'erreur quadratique relative (Relative Squared Error) ou la similarité cosinus (Cosine similarity) (pour les espaces à haute dimension) pourraient être utilisées pour évaluer les modèles de régression et, dans les modèles de classification, utilisés avec des versions adaptées qui mesurent la marge de discrimination afin de renforcer la robustesse des prédictions.
- Des outils graphiques et différentes métriques seront utilisés pour respectivement représenter et mesurer la capacité d'un test à discriminer entre un sol pollué (en bonne santé) ou non. La courbe ROC (ROC curve), la sensibilité (sensitivity), la spécificité (specificity), la précision (accuracy), le taux de fausses découvertes (false discovery rate), le score F1 (F1 score) permettent d'estimer précisément et facilement la qualité des modèles de prédiction diagnostic. Pour la classification multi-classes, nous utiliserons des figures en fonction de la matrice de confusion et/ou des mesures statistiques comme le coefficient kappa de Cohen (Cohen's kappa coefficient).

Outil informatique diagnostic environnemental: prototype

Nous développerons le prototype d'une plate-forme web conviviale en utilisant le framework python-Django afin de présenter prospectivement la valeur ajoutée de l'outil de diagnostic aux biologistes avant la validation sur le terrain en collaboration avec des viticulteurs et des agences environnementales. En effet, le langage de programmation Python assurera l'interopérabilité et facilitera la maintenance tout en profitant des nombreuses fonctionnalités déjà implémentées dans Django. Nous nous concentrerons particulièrement sur les aspects de la visualisation des données, car l'objectif principal de notre outil diagnostic est de communiquer efficacement avec les biologistes, viticulteurs ou/et autres agences environnementales. Notre objectif est de rendre les données complexes accessibles, compréhensibles et facilement utilisables.

Les biologistes de Changins seront fortement impliqués tout au long du développement du prototype afin de s'assurer que la solution proposée réponde à leurs besoins et à leurs attentes, en particulier pour être intuitive, pertinente et conviviale. Cela devrait, nous l'espérons, aboutir à une acceptabilité élevée et à

l'adoption de l'outil diagnostic par tous les acteurs concernés (biologistes, viticulteurs agences environnementales). A cette fin, nous utiliserons une approche par étapes selon laquelle les besoins, la portée et les critères d'acceptation sont discutés conjointement pour définir les exigences pour le développement du prototype de l'outil de diagnostic. L'utilité et les fonctionnalités de l'outil informatique seront régulièrement évaluées par les biologistes impliqués dans le projet afin de renforcer son adéquation.

5. Planification du projet

5.1 Quelles sont les activités envisagées ?

Pour chaque activité, merci d'identifier les risques, de les évaluer et de définir le cas échéant des mesures préventives ou correctives.

Le tableau 2 résume l'ordonnancement des lots de travail (workpackages) dans le temps et leur répartition entre les différents partenaires du projet (section 5.2). Les livrables et les jalons sont **résumés** dans le **tableau 2** (diagramme de Gantt) et le tableau 4 (voir la section 6). Le projet sera réalisé en collaboration entre deux groupes partenaires, chacun, apportant l'expertise nécessaire pour accomplir les différentes tâches. Le groupe du prof. Heger sera responsable de l'acquisition, l'assignation, l'annotation et de la validation des données provenant de l'expérience de laboratoire et de l'acquisition des données provenant du réseau de 33 vignobles. Le groupe du prof. Peña sera chargé de concevoir et d'implémenter les tâches de calcul (informatique), comprenant la détermination des modèles prédictifs (apprentissage automatique) et l'implémentation de l'outil de diagnostic des sols. Les deux groupes seront profondément impliqués dans la conception du programme de prédiction de la qualité des sols. Il est important de noter que le responsable principal du projet sera chargé d'organiser des réunions régulières, de coordonner l'effort de transfert de technologie et la mise en œuvre du projet. Le consortium vise à se réunir tous les deux mois entre les groupes participants (Professeurs, personnel et étudiants impliqués) pour discuter des avancées et ou des difficultés rencontrées pour une collaboration optimale. S'appuyant sur l'expérience acquise lors de nos précédents projets de recherche financés par le FNS ou EU, les deux Co-requérants s'efforceront de stimuler les interactions et les échanges scientifiques entre les ingénieurs et les biologistes afin de faire avancer le programme dans les meilleures conditions possibles.

WP1. Périmètre du projet (Mois 1 – Mois 2)

- ❖ Coordonner l'acquisition des échantillons de sol et des métadonnées (collaborations avec l'équipe du Dr. Imfeld et l'équipe du Prof. Sven Bacher)
- M1 : Définition du périmètre du projet
 - D1 - Définition du périmètre du projet

WP2. Acquisition des données de l'expérience en laboratoire (Mois 2 – Mois 6)

- ❖ Acquisition des échantillons et des métadonnées
- ❖ Extraction de l'ADN
- ❖ Amplification du fragment V4 du 18S rRNA avec amorces eucaryotes universelles (protocole identique à celui utilisé pour les échantillons du Valais, voir ci-dessous)
- ❖ Préparation des bibliothèques
- ❖ Séquençage Illumina Miseq (par Génome Québec)
- ❖ Regroupement des séquences en OTUs (Operational Taxonomic Units)
- ❖ Assignation taxonomique des séquences d'ARNr 18S (base de référence PR 2)
- M2 - Données de laboratoires annotées
 - D2 - Données de laboratoires annotées

WP3. Acquisition des données du terrain (réseau de 33 vignobles) (Mois 2 – Mois 3)

- ❖ Sélection des données nécessaires pour le projet
- ❖ Établissement des tableaux de données
- M3 - Données de terrain annotées
 - D3 - Données de terrain annotées

WP4. ML Modèles hautement prédictifs (Mois 7 – Mois 19)

- ❖ Sélection des Features
 - D4 - Liste des bioindicateurs
- ❖ Sélection du Modèle
- ❖ Modélisation
 - D5 -Modèle(s) de prédiction
- ❖ Amélioration du modèle

- D6 - Modèle explicatif
- M4 - Évaluation du modèle
- ❖ Évaluation des performances prédictives du modèle

WP5. Outil informatique diagnostic environnemental (prototype) (Mois 20 – Mois 24)

- ❖ Définitions des besoins, de la portée, des critères d'acceptation
 - D7 - Cahier des charges (document interne fixant les modalités d'exécution de l'outil prototype)
- ❖ Visualisation des solutions
- ❖ Implémentation du prototype
- ❖ Définition des tests à effectuer
- ❖ Tests utilisateurs
- ❖ Contrôle de conformité avec le cahier des charges
- M5 - Evaluation de l'outil
 - D8 - Prototype de l'outil de prédiction

WP6. Communication & dissémination (Mois 1 – Mois 24)

- ❖ Coordination et suivi du projet
 - D9 - Les résultats de ce projet seront valorisés par la rédaction d'une publication dans une revue à comité de lecture.
- ❖ Transfert de technologies, préparation (réunions)
 - D10 - Évaluation des avantages et des limitations du système
- ❖ Valorisation des résultats (par exemple, attirer l'attention des médias)

Tableau 1 : Prévion et gestion des risques

Risque	Cause(s) possibles	Conséquence	Mesure
Risque organisationnel	Mauvaise gestion Problème de communication	Retard, erreurs	-Planifier d'avance les rôles et les devoirs. -Partager les ressources. Partage équitable des tâches pour tous les partenaires du projet
Accès limité aux données	Problèmes pour produire les données moléculaires provenant des microcosmes (par ex. problème avec extraction de l'ADN, PCR)	Retard, risque de sur-apprentissage	-Multiplication des sources de données (Les données du terrain proviendront de l'équipe du prof. Sven Bacher. Les échantillons de l'expérience en microcosmes seront fournis par le Dr. Gwenaël Imfeld) -Commencer à collecter de manière prospective de nouvelles données pendant le projet.
Manque de personnel qualifié pour certaines tâches du projet	Absence d'un collaborateur (pour cause de maladie ou pour d'autres raisons)	Retard	-Profiter de l'expertise des différents partenaires impliqués -Recruter un nouveau collaborateur - Mutualiser les compétences
Prédiction pas assez précise	Données pas assez prédictives	Difficulté de sélectionner des bioindicateurs associés aux	-Élargir l'étude à d'autres groupes d'organismes (bactéries, champignons)