Candidature au poste de
Maître-assistant de classe normale
Science des données
MINES ParisTech

Chloé-Agathe Azencott

CBIO MINES ParisTech
60 boulevard Saint-Michel
75272 Paris Cedex 06 – France
Téléphone: 06 25 53 80 71
Email : chloe-agathe.azencott@mines-paristech.fr
Homepage : http://www.cazencott.info

Née le : 7 janvier 1985 Nationalité : française

Table des matières

1. Lettre de motivation

Chloé-Agathe Azencott
CBIO MINES ParisTech
60 bd Saint-Michel
75272 Paris Cedex 06 – France

MINES ParisTech
60 bd Saint-Michel
75272 Paris Cedex 06 – France

23 mai 2018

Madame, Monsieur,

C'est avec enthousiasme que je présente ma candidature au poste de maître assistant de classe normale en sciences des données, avec affectation au Centre de Bio-Informatique (CBIO). Je suis actuellement déjà en poste dans ce même Centre, en tant que chargée de recherche contractuelle. Ce changement de statut me permettrait de renforcer mon investissement dans l'enseignement à l'École, et de consolider ma position au sein d'un Centre en pleine mutation.

Ma recherche porte sur le développement d'outils d'apprentissage statistique pour les données biologiques structurées, motivé par la recherche thérapeutique ; les applications sur lesquelles je travaille vont de la génétique des cancers à la prédiction d'effets secondaires. Ces thématiques s'intègrent parfaitement à la stratégie scientifique du CBIO, que j'ai en effet participé à définir avec les autres membres permanents du Centre au cours de ces dernières années.

De plus, elles me conduisent naturellement à proposer des enseignements sur des sujets d'actualité pour les ingénieurs formés à l'École. Je suis ainsi co-responsable avec T. Walter de l'enseignement spécialisé d'introduction à la génétique et la bioinformatique (proposé en deuxième année du cursus ingénieur civil), et avec J.-P. Vert et F. Moutarde de l'enseignement spécialisé sur le machine learning à grande échelle (proposé en semaine bloquée aux élèves de deuxième et troisième année du cursus ingénieur civil et ouvert plus largement aux élèves et personnels de PSL). Je suis aussi chargée d'une séance dans l'enseignement spécialisé d'apprentissage automatique de F. Moutarde (lui aussi proposé en semaine bloquée aux mêmes publics).

À l'extérieur de l'École, je suis actuellement responsable du cours électif d'introduction au machine learning de CentraleSupélec (proposé en deuxième année du cursus ingénieur civil et en anglais), dont le nombre d'inscrits est passé de 40 en 2015 à 149 en 2017. En partenariat avec CentraleSupélec, j'ai également piloté avec N. Paragios le parcours Data Scientist d'OpenClassrooms, la première plateforme francophone d'éducation en ligne. J'ai assuré quatre des cours proposés dans ce parcours, qui sont actuellement disponibles en ligne. Ces deux expériences m'ont amenée à écrire un manuel d'introduction au machine learning à destination d'élèves de L3/M1 ou de deuxième année d'un cycle ingénieur qui paraîtra chez Dunod à la rentrée 2018. Elles m'ont aussi permis de réfléchir en profondeur à l'enseignement de la science des données. J'ai à cœur de mettre cette expertise au service de la réflexion menée au sein de l'École sur la question, et ai pour cela rejoint le groupe de travail mené par D. Ryckelynck.

À ces expériences d'enseignement s'ajoutent enfin des interventions ponctuelles en bioinformatique dans le Master 2 en bioinformatique de Paris-Diderot ou dans le Master PSL Chimie et Sciences du Vivant, et des formations dans des écoles d'été internationales destinées à des chercheurs. Au delà de l'apprentissage statistique, je suis donc aussi prête à assurer des enseignements autour des applications en santé. Le cours mentionné ci-dessus dont je suis co-responsable avec T. Walter s'inscrit dans cette optique.

En ce qui concerne mes activités de recherche, le CBIO a été pour moi un excellent environnement où conduire mes travaux et continuera je pense de l'être pour les années à venir. Depuis mon arrivée, j'ai ainsi publié six articles dans des revues ou conférences internationales à comité de lecture et en ai deux autres en préparation. J'ai aussi pu co-encadrer quatre doctorants, dont un a soutenu sa thèse avec succès en mai 2017. Enfin, j'ai pu y créer le Idea DREAM Challenge, qui a permis en collaboration avec A. Goldenberg (U. Toronto) and Sage Bionetworks de financer deux projets proposés par des modélisateurs ayant un fort potentiel de retombées pour des problématiques santé.

Si les interactions du CBIO avec le secteur privé peuvent être moins fortes que dans d'autres Centres, je co-supervise néanmoins avec J.-P. Vert un doctorant financé par une thèse CIFRE avec Sanofi, et m'apprête à encadrer un post-doctorant financé par Sancare.

Vous trouverez ci-joint mon CV, une synthèse de mes activités de recherche, une liste de mes publications, ainsi que les documents administratifs requis pour candidater. Je me tiens à votre disposition pour toute information complémentaire, et vous remercie de l'attention que vous porterez à ma candidature.

Veuillez accepter, Madame, Monsieur, l'expression de mes sentiments les plus respectueux.

Chloé-Agathe Azencott

2. Curriculum Vitae

Situation actuelle

Depuis 2013 **Chargée de recherche**, *CBIO MINES ParisTech*, Paris (75).

Formation universitaire

2005 – 2010 **Doctorat en Informatique**, *University of California Irvine (UCI)*, États-Unis.
« Statistical data mining and machine learning for chemoinformatics and drug discovery »
Directeur : P. Baldi (UCI – Informatique et sciences de l'information)

2004 – 2005 **Master Mathématique et Informatique**, *École Nationale Supérieure des Télécommunications de Bretagne*, Brest (28), Mention Bien.
Spécialité Logiciel et méthodes formelles

2002 – 2005 **Diplôme d'Ingénieur**, *École Nationale Supérieure des Télécommunications de Bretagne*, Brest (28).
Option Informatique des télécommunications.

Expérience professionnelle: Enseignement

Depuis 2014 **Responsable de cours**.
MINES ParisTech, Paris (75)
- Depuis 2014 : Introduction à la génomique et à la bioinformatique (2A)
- Depuis 2017 : Large scale machine learning (2A – 3A)

CentraleSupélec, Chatenay-Malabry (92) puis Gif-sur-Yvette (91)
Depuis 2015 : Introduction to machine learning (2A, en anglais)
OpenClassrooms, Paris (75)
Depuis 2017 : Parcours data scientist (cours en ligne).

Depuis 2012 **Chargée de cours**.
MINES ParisTech, Paris (75)
Depuis 2015 : Apprentissage automatique (2A)
Université Paris-Diderot, Paris (75)
- Depuis 2017 : Découverte multi-locus de biomarqueurs (M2)
- 2015 – 2016 : Fouille de données graphe et chémoinformatique (M2)

PSL University, Paris (75)
Depuis 2017 : Machine learning for therapeutic research (M1 – M2)
Universität Tübingen, Allemagne
- 2012 : Fouille de données en bioinformatique (M2)
- 2013 : Fouille de données en bioinformatique (M2).

Depuis 2017 **Écoles d'été**.
- 2018 : Data Science Summer School, « Hands on machine learning for genetics data », Palaiseau (91)
- 2017 : École d'été du GDR BioComp, « Around machine learning in 90 minutes », Roscoff (28)
- 2017 : Microbiome Summer School, « Machine learning for efficient biomarker discovery », Université de Laval, Québec (Canada).

2006 – 2012 **Monitorat**.

Universität Tübingen (Allemagne)
- 2012 : Séminaire de bioinformatique (M2)

University of California Irvine (États-Unis)
- 2008 : Introduction aux statistiques pour l'informatique (licence)
- 2007 : Statistiques (licence)
- 2006 : Introduction à l'intelligence artificielle (licence).

Depuis 2017 **Comités pédagogiques**.
- Depuis 2017 : Master PSL Chimie et Sciences du Vivant, Paris (75)
- Depuis 2018 : Master PSL Sciences du Vivant, Paris (75).

Expérience professionnelle: Recherche

Depuis 2013 **Chargée de recherche**, *MINES ParisTech*, Paris (75).

Centre de Bio-Informatique
- Apprentissage automatique pour les données génomiques et la médecine de précision avec applications au cancer
- Sélection de variables et sparsité structurée
- Apprentissage multi-tâche
- Prédiction d'interactions protéine-ligand
- Associations génotype-phénotype.

2011 – 2013 **Chercheuse postdoctorante**, *Max Planck Institute for Developmental Biology & Max Planck Institute for Intelligent Systems*, Tübingen (Allemagne).

Groupe « Machine Learning and Computational Biology » dirigé par K. Borgwardt
- Méthodes statistiques pour études d'associations génome entier multilocus
- Prédictions d'associations gène-maladie à l'aide de graphes
- Intégration de données biologiques partiellement disponibles.

2005 – 2010 **Graduate Student Researcher**, *University of California Irvine*, États-Unis.

« Institute for Genomics and Bioinformatics » dirigé par P. Baldi
- Prédiction de propriétés biologiques, chimiques et physiques de molécules (noyaux pour graphes et SVM)
- Criblage haut-débit virtuel (méthodes à noyaux, réseaux de neurones)
- Prédiction de réactions chimiques (SVM, réseaux de neurones)
- Docking moléculaire.

Été 2009 **Stage de Recherche**, *IBM R&D Tel-Aviv*, Israël.

Groupe « Machine Learning and Data Mining » dirigé par M. Rosen-Zvi
- Analyse statistique de données SNP pour le projet européen HyperGene.

2005 – 2010 **Stage de Master Recherche**, *École des Mines de Paris*, Fontainebleau (78).

Centre for Computational Biology dirigé par J.-P. Vert
- Implémentation et validation de noyaux pour séquences biologiques
- Interface web pour le test de noyaux.

Logiciels

Le code écrit pour mes projets est disponible librement sur `http://github.com/chagaz`.

3. Synthèse des travaux

Mes travaux de recherche s'inscrivent dans le cadre du développement de techniques d'apprentissage automatique et statistique (« *machine learning* ») pour la recherche thérapeutique. Il s'agit d'extraire de quantités toujours plus larges de données biologiques, chimiques, médicales et pharmaceutiques des informations qui facilitent le développement de nouvelles thérapies. Les applications sont nombreuses, de la compréhension de mécanismes biologiques à la découverte et la synthèse de médicaments-candidats.

3.1 Médecine de précision

Mes travaux sont motivés par leurs applications à la *médecine de précision*. Celle-ci prend son origine dans le constat que des patients présentant les mêmes symptômes peuvent ne pas avoir le même pronostic ou répondre différemment au même traitement. La médecine de précision, ou médecine personnalisée, vise à identifier les causes de ces différences afin de faciliter l'adaptation des traitements aux caractéristiques personnelles des patients.

Parmi ces caractéristiques, les différences génomiques entre patients expliquent en grande partie pourquoi ceux-ci vivent la même maladie différemment. La mise en œuvre de la médecine de précision requiert donc d'identifier les régions du génome associées avec une prédisposition, un pronostic ou une réponse thérapeutique. Cet enjeu est reconnu par les pouvoirs publics, en particulier en France via le plan « Médecine France génomique 2025 », piloté par l'alliance nationale pour les sciences de la vie et la santé (Aviesan) et soutenu par l'État.

Cependant, si les avancées technologiques dans le domaine du séquençage génétique nous permettent de recueillir des données moléculaires de plus en plus riches, les méthodes permettant de les interpréter sont beaucoup plus limitées. En effet, alors que l'analyse de données massives (« *Big Data* ») est un domaine en plein essor, les données génétiques présentent des défis spécifiques auxquels les méthodes mises au point pour d'autres applications ne répondent souvent pas. Il s'agit en effet de données en très grande dimension, contenant largement plus de variables descriptives que d'échantillons ; pour pouvoir suggérer de nouvelles hypothèses biologiques, il est indispensable que les modèles construits soient interprétables ; enfin, les données sont généralement hétérogènes, de par la nature tant des variables considérées (mutations génétiques, profils de méthylation, quantités d'ARNs) que des échantillons observés (patients issus de populations génétiquement inhomogènes, diversité des cellules cancéreuses au sein d'une même tumeur).

Dans ce contexte, mes travaux consistent à proposer et mettre en œuvre de nouveaux outils de machine learning qui permettent de répondre à ces défis. Ces travaux bénéficient de l'environnement de recherche exceptionnel du CBIO MINES ParisTech, qui permet des interactions fréquentes et soutenues avec des experts en machine learning, bio-informatique, et génétique, au travers de son association avec l'Institut Curie et de sa situation géographique dans le quartier Latin.

3.2 Analyse de données structurées

Les données dont nous disposons comportent généralement des centaines de milliers de variables (par exemple, des mutations d'un seul nucléotide) pour seulement quelques milliers d'individus. Dans un contexte où le nombre d'échantillons est largement inférieur à celui de variables, les méthodes classiquement utilisées pour la sélection de variables manquent de puissance statistique [8].

Pour y remédier, il est possible de combiner ces données avec des connaissances accumulées par ailleurs concernant le système biologique qui nous intéresse. Ces connaissances peuvent souvent être représentées sous la forme de *structures* sur les variables, et en particulier de *réseaux biologiques*. Le problème de sélection de variables peut alors être formulé en ajoutant aux approches classiques une contrainte qui encourage les variables sélectionnées à être connectées sur un réseau donné. C'est ce que permet le cadre de la *pertinence régularisée*, que j'ai proposé avec la méthode SConES [CA5] puis développé dans d'autres travaux [CA12, CA17, CA16]. Ce cadre permet notamment de formuler un problème de sélection de variables sous contraintes structurelles comme un problème de coupe sur un graphe, ce qui permet de le résoudre efficacement par des méthodes de flux.

Les travaux de doctorat de H. Climente, que je co-encadre avec V. Stoven (CBIO), ont notamment permis de montrer l'applicabilité de ce type de méthodes à une cohorte concernant le cancer du sein, en collaboration avec N. Andrieu, C. Lonjou et F. Lesueur (Institut Curie).

Les travaux de doctorat de C. Le Priol, que je co-encadre avec X. Gidrol (CEA Grenoble), visent à étudier un type particulier de réseau biologique, celui des interactions entre ARN messagers et les micro-ARN qui les ciblent. Nous nous intéressons à comprendre, par ces réseaux, le rôle de la variabilité des micro-ARN dans le développement des cancers.

Perspectives : Le cadre de la pertinence régularisée offre de nombreuses perspectives de développement. En particulier, j'ai l'intention de l'exploiter pour mettre au point des méthodes de sélection de variables plus robustes (c'est-à-dire qui identifient les mêmes régions du génomes sur des jeux de données seulement légèrement différents) et non-linéaires (voir aussi la partie 3.4). C'est le sujet de deux demandes de financement que j'ai déposées en 2018, l'une auprès de l'Agence Nationale de la Recherche (projet SCAPHE, programme « Jeune Chercheur Jeune Chercheuse ») et l'autre auprès du European Research Council (programme « International Training Network », en collaboration avec un consortium d'experts européens mené par W. Huber de l'EMBL Heidelberg en Allemagne). J'ai aussi l'intention de continuer à appliquer cette méthodologie à de nouveaux types de problèmes biologiques, comme par exemple pour étudier le rôle des plaquettes dans la réponse immunitaire, dans le cadre d'une collaboration avec B. Ramkhelawon du NYU Medical Center aux États-Unis, pour laquelle nous avons déposé une demande de financement auprès du Human Frontier Science Program.

3.3 Apprentissage multi-tâche

Les méthodes dites *multi-tâche* permettent de réduire l'impact du manque relatif d'échantillons dans nos données. Ces méthodes consistent à résoudre simultanément des problèmes proches. Dans le cadre de la médecine de précision, il peut s'agir par exemple de rechercher les facteurs de risque pour des maladies similaires, ou d'étudier plusieurs traitements pour la même maladie [CA3, CA11]. Avec M. Sugiyama, j'ai montré comment étendre la pertinence régularisée à ce contexte [CA12]. Cependant, nos travaux ne prenaient pas en compte de notion de similarité entre tâches, qui permettrait d'imposer que les variables sélectionnées pour deux tâches soient d'autant plus similaires que les deux tâches sont semblables.

Cela semble particulièrement pertinent dans l'étude de la réponse à plusieurs traitements, car on

peut alors se baser sur la structure moléculaire de ces traitements pour les comparer. Ce sujet de recherche est historique au CBIO [9, 7], et aussi celui de ma thèse [CA10, CA19]. Les travaux de doctorat de Víctor Bellón, que j'ai co-encadrés avec V. Stoven (CBIO), ont démontré l'intérêt de ces approches dans un cadre plus classique [CA11].

Les approches multi-tâche sont par ailleurs particulièrement intéressantes dans le cadre de la prédiction d'interactions entre molécules thérapeutiques et protéines humaines, où chaque molécule (ou chaque protéine) peut former une tâche. Prédire ces interactions permet de proposer de nouvelles indications thérapeutiques ou de détecter des effets secondaires potentiels. C'est un sujet que j'étudie en collaboration avec V. Stoven et B. Playe (CBIO) [CA15].

Perspectives J'ai l'intention de poursuivre le développement d'approches de pertinence régularisée multi-tâches utilisant une similarité entre tâches, qui fait aussi partie du projet SCAPHE que j'ai déposé auprès de l'ANR. Dans le cadre de la prédiction d'interactions entre molécules et protéines, je m'intéresse aussi avec B. Playe (CBIO) à l'utilisation d'approches d'apprentissage profond pour apprendre des représentations informatives de molécules et protéines.

3.4 Non-linéarités

La grande majorité des approches de sélection de variables utilisées dans le contexte de la génomique sont *linéaires*, c'est-à-dire qu'elles ne peuvent expliquer le phénomène d'intérêt que comme une combinaison linéaire des variables sélectionnées. Cela vaut aussi pour SConES. Cependant, il est raisonnable de supposer que des régions du génome puissent interagir de façon non-linéaire. Modéliser de telles interactions, que l'on qualifie d'épistatiques, aggrave cependant les problèmes statistiques déjà rencontrés précédemment, et crée aussi des problèmes computationnels : il devient difficile d'évaluer toutes les combinaisons possibles de variables.

Quand on se limite au cas quadratique (les variables interagissent deux par deux), il est possible de résoudre le problème calculatoire grâce à des calculs sur carte graphique [CA6]. Les travaux de doctorat de L. Slim, que je co-encadre avec J.-P. Vert (CBIO) et C. Chatelain (Sanofi), s'inspirent de techniques développées pour déterminer l'effet d'un traitement dans le cadre d'études observationnelles pour proposer des méthodes ayant une meilleure puissance statistique. Ceux de H. Climente, que je co-encadre avec V. Stoven (CBIO), visent à intégrer les effets épistatiques à l'approche structurée discutée dans la partie 3.2. Nous explorons aussi, en collaboration avec M. Yamada (Université de Kyoto, RIKEN), l'idée d'utiliser la structure des données pour définir l'architecture d'un réseau de neurones [1].

Perspectives J'ai l'intention dans les années à venir d'étudier l'applicabilité d'autres méthodes de sélection de variables non-linéaires aux données génétiques. Les *set covering machines*, qui ont été appliquées avec succès à des données de dimensionalité comparable à celle qui nous intéresse ici [6], sont une piste particulièrement prometteuse. J'aimerais aussi, comme je l'ai proposé dans SCAPHE (déposé auprès de l'ANR), étudier l'applicabilité de techniques statistiques récentes comme les filtres knock-off [4], les p-filtres multi-couche [5] ou des algorithmes d'inférence sélective pour des interactions entre variables [2] peuvent être utilisées pour la détection d'effets épistatiques.

3.5 Intégration de données hétérogènes

Les méthodes discutées ci-dessus permettent de traiter un type de données biologiques à la fois. Cependant, il est fréquent de disposer de mesures de natures très différentes pour les mêmes échantillons. Ces mesures caractérisent souvent différentes échelles biologiques. Par exemple, des données de mutation décrivent la séquence d'ADN, tandis que des données d'expression de gène décrivent les transcrits produits et, en première approximation, les protéines présentes ; enfin, des données de microscopie peuvent caractériser l'intégralité d'une cellule, tandis que les dossiers patient électroniques permettent de décrire l'individu lui-même. Ces données sont ainsi complémentaires et les étudier simultanément devrait permettre d'améliorer nos analyses.

Perspectives L'intégration de ce type de données très hétérogènes est encore en ce qui me concerne une perspective de recherche. Les expertises présentes au sein du CBIO (V. Stoven pour les protéines, T. Walter pour les images, et J.-P. Vert et moi-même pour les données génétiques) devraient pouvoir nous permettre de proposer des solutions nouvelles et pertinentes sur le sujet. J'ai aussi commencé avec J.-P. Vert à m'intéresser aux données disponibles dans les dossiers patients électroniques, à travers un projet avec J. Guérin et A. Livartowsky de l'Institut Curie. Une collaboration avec la startup Sancare me permettra de poursuivre dans cette direction, en étudiant plus particulièrement les aspects liés à la confidentialité des données (une question que j'ai commencé à étudier dans le cadre des données génomiques [CA14]). Cet axe de recherche fait aussi l'objet d'une demande de financement que j'ai déposée auprès du European Research Council (programme « International Training Network », en collaboration avec un consortium d'experts européens mené par K. Borgwardt de l'ETH Zurich en Suisse.)

J'ai aussi commencé à travailler avec A. Rausell de l'Institut Imagine sur l'intégration de données via des réseaux biologiques. Il s'agit ici de construire plusieurs réseaux sur des gènes, issus de types de données différentes, et d'utiliser ces réseaux simultanément, à travers une approche d'apprentissage profond [3], pour prédire lesquels de ces gènes sont associés à une maladie donnée.

Réferences

[1] J. Ma, et al. **Using deep learning to model the hierarchical structure and function of a cell**. *Nature methods*, 15(4):290, 2018.

[2] S. Suzumura, K. Nakagawa, Y. Umezu, K. Tsuda, I. Takeuchi. **Selective inference for sparse high-order interaction models**. *PMLR*, 3338–3347, 2017.

[3] M. Schlichtkrull, et al. **Modeling Relational Data with Graph Convolutional Networks**. *arXiv:1703.06103*, 2017.

[4] D. Brzyski, et al. **Controlling the Rate of GWAS False Discoveries**. *Genetics*, 205(1):61–75, 2017.

[5] R. F. Barber, A. Ramdas. **The p-filter: multilayer false discovery rate control for grouped hypotheses**. *J. R. Stat. Soc. B*, 79(4):1247–1268, 2017.

[6] A. Drouin, et al. **Predictive computational phenotyping and biomarker discovery using reference-free genome comparisons**. *BMC Genomics*, 17:754, 2016.

[7] L. Jacob, B. Hoffmann, V. Stoven, J.-P. Vert. **Virtual screening of GPCRs: an in silico chemogenomics approach**. *BMC bioinformatics*, 9(1):363, 2008.

[8] R. Clarke, et al. **The properties of high-dimensional data spaces: implications for exploring gene and protein expression data**. *Nature Reviews Cancer*, 8(1):37–49, 2008.

[9] P. Mahé, L. Ralaivola, V. Stoven, J.-P. Vert. **The pharmacophore kernel for virtual screening with support vector machines**. *Journal of Chemical Information and Modeling*, 46(5):2003–2014, 2006.

4. Publications

Articles dans des revues avec comité de lecture

[CA1] Chloé-Agathe Azencott, Tero Aittokallio, Sushmita Roy, et al. **The inconvenience of data of convenience: computational research beyond post-mortem analyses**. *Nature methods*, 14(10):937, 2017.

[CA2] Solveig K. Sieberts et al. **Crowdsourced assessment of common genetic contribution to predicting anti-TNF treatment response in rheumatoid arthritis**. *Nature Communications*, 7:12460, 2016.

[CA3] Federica Eduati et al. **Opportunities and limitations in the prediction of population responses to toxic compounds assessed through a collaborative competition**. *Nature Biotechnology*, 33(9):933–940, 2015.

[CA4] Dominik Grimm, Chloé-Agathe Azencott, Fabian Aicheler, Udo Gieraths, Daniel MacArthur, Kaitlin Samocha, David Cooper, Peter Stentson, Mark Daly, Jordan Smoller, Laramie Duncan, and Karsten Borgwardt. **The evaluation of tools used to predict the impact of missense variants is hindered by two types of circularity**. *Human Mutation*, 36(5):513–523, 2015.

[CA5] Chloé-Agathe Azencott, Dominik Grimm, Mahito Sugiyama, Yoshinobu Kawahara, and Karsten M. Borgwardt. **Efficient network-guided multi-locus association mapping with graph cuts**. *Bioinformatics*, 29(13):i171–i179, 2013. Proceedings of the 21st Annual International Conference on Intelligent Systems for Molecular Biology (ISMB 2013).

[CA6] Tony Kam-Thong, Chloé-Agathe Azencott, Lawrence Cayton, Benno Pütz, André Altmann, Nazanin Karbalai, Philipp G. Sämann, Bernhard Schölkopf, Betram Müller-Myhsok, and Karsten M. Borgwardt. **GLIDE: GPU-based linear regression for the detection of epistasis**. *Human Heredity*, 73:220–236, 2012.

[CA7] Matthew A. Kayala, Chloé-Agathe Azencott, Jonathan H. Chen, and Pierre Baldi. **Learning to predict chemical reactions**. *Journal of Chemical Information and Modeling*, 51(9):2209–2222, 2011.

[CA8] S. Joshua Swamidass, Chloé-Agathe Azencott, Kenny Daily, and Pierre Baldi. **A CROC stronger than ROC: measuring, visualizing and optimizing early retrieval**. *Bioinformatics*, 26(10):1348–1356, 2010.

[CA9] S. Joshua Swamidass, Chloé-Agathe Azencott, Ting-Wan Lin, Hugo Gramajo, Sheryl Tsai, and Pierre Baldi. **The Influence Relevance Voter: an accurate and interpretable virtual high throughput screening method**. *Journal of Chemical Information and Modeling*, 49(4):756–766, 2009.

[CA10] Chloé-Agathe Azencott, Alexandre Ksikes, S. Joshua Swamidass, Jonathan H. Chen, Liva Ralaivola, and Pierre Baldi. **One- to four-dimensional kernels for virtual screening and the prediction of physical, chemical and biological properties**. *Journal of Chemical Information and Modeling*, 47(3):965–974, 2007.

Actes de conférences avec comité de lecture

[CA11] Victor Bellon, Véronique Stoven, and Chloé-Agathe Azencott. **Multitask feature selection with task descriptors**. In *Pacific Symposium on Biocomputing*, volume 21, pages 261–272, 2016.

[CA12] Mahito Sugiyama, Chloé-Agathe Azencott, Dominik Grimm, Yoshinobu Kawahara, and Karsten M. Borgwardt. **Multi-task feature selection on multiple networks via maximum flows**. In *Proceedings of the 2014 SIAM International Conference on Data Mining*, pages 199–207, 2014.

[CA13] Chloé-Agathe Azencott Pierre Baldi and S. Joshua Swamidass. **Bridging the Gap Between Neural Network and Kernel Methods: Applications to Drug Discovery**. In *20th Italian Workshop on Neural Nets*, pages 3–13, 2011.

Pré-publications

[CA14] Chloé-Agathe Azencott. **Machine learning and genomics: precision medicine vs. patient privacy**. *arXiv:1802.10568*, 2018. To appear in Philosophical Transactions of the Royal Society A.

[CA15] Chloé-Agathe Azencott Benoit Playe and Véronique Stoven. **Efficient multi-task chemogenomics for drug specificity prediction**. *bioRxiv:10.1101/193391*, 2018. Accepted with revisions at PLoS One.

Paquet logiciel

[CA16] Héctor Climente and Chloé-Agathe Azencott. **martini: GWAS incorporating networks in R**, 2017.

Chapitres de livre

[CA17] Chloé-Agathe Azencott. **Network-guided biomarker discovery**. In *Machine Learning for Health Informatics*, number 9605 in Lecture Notes in Computer Science. Springer, 2016.

[CA18] Chloé-Agathe Azencott and Pierre Baldi. **Virtual high-throughput screening with two-dimensional kernels**, In *Hands-On Pattern Recognition: Challenges in Machine Learning*, volume 1, pages 131–146. Microtome, 2011.

Thèse de doctorat

[CA19] Chloé-Agathe Azencott. **Statistical machine learning and data mining for chemoinformatics and drug discovery**. PhD thesis, University of California, Irvine, 2010. ProQuest/UMI, AAT 3422105.

Présentations dans des conférences avec comité de lecture

[CA20] Chloé-Agathe Azencott Christophe Le Priol, Laurent Guyon and Xavier Gidrol. **Analysis of microRNA sequences identifies conserved families of microRNAs**. 17th Open Days in Biology, Computer Science and Mathematics (poster), 2016.

[CA21] Chloé-Agathe Azencott, Dominik Grimm, Jordan Smoller, Laramie Duncan, and Karsten M. Borgwardt. **Beware of circularity: A critical assessment of the state of the art in deleteriousness prediction of missense variants**. 64th Annual Meeting of The American Society of Human Genetics, 2014.

[CA22] Víctor Bellón, Chloé-Agathe Azencott, Véronique Stoven, Valentina Boeva, and Jean-Philippe Vert. **DREAM Rheumatoid Arthritis Responder Challenge: Team Lucia**. RECOMB/ISCB Conference on Regulatory & Systems Genomics; DREAM Challenges & Cytoscape Workshops (poster), 2014.

[CA23] Chloé-Agathe Azencott, Dominik Grimm, Yoshinobu Kawahara, and Karsten M. Borgwardt. **Efficiently mapping phenotypes to networks of genetic loci**. Machine Learning and Computational Biology Workshop (poster), 2012.

[CA24] Chloé-Agathe Azencott, Matthew A. Kayala, and Pierre Baldi. **PropOrb: a frontier molecular orbital interaction proposer**. 239th American Chemistry Society National Meeting, 2010.

[CA25] Chloé-Agathe Azencott, Matthew A. Kayala, and Pierre Baldi. **Combining quantitative data and qualitative knowledge to score reaction energies**. 237th American Chemistry Society National Meeting, 2009.

[CA26] Chloé-Agathe Azencott, S. Joshua Swamidass, and Pierre Baldi. **Performance prediction of the Influence Relevance Voter**. The Learning Workshop (poster), 2009.

[CA27] Chloé-Agathe Azencott, S. Joshua Swamidass, and Pierre Baldi. **Virtual high-throughput screening and early recognition**. Women in Machine Learning Workshop (poster), 2009.

[CA28] Chloé-Agathe Azencott and Pierre Baldi. **Virtual high-throughput screening with two-dimensional kernels**. Agnostic Learning vs. Prior Knowledge Workshop, International Joint Conference on Neural Networks, 2007.

[CA29] Chloé-Agathe Azencott and Pierre Baldi. **Kernels for predictive regression–physical, biological and chemical properties of small molecules**. Workshop for Women in Machine Learning, 2006.

5. Complément du Curriculum Vitae

Encadrement

Depuis 2014 **Encadrement de doctorats**, *MINES ParisTech*, Paris (75).

- Depuis 2016 : Lotfi Slim, « Detection of epistasis in genome-wide association studies with machine learning methods for biomarkers and therapeutic target identification ». Co-supervision avec J.-P. Vert (CBIO) et C. Chatelain (Sanofi).
- Depuis 2016 : Héctor Climente González, « Integrating structural constraints in multi-locus genome-wide association studies ». Co-supervision avec V. Stoven.
- Depuis 2016 : Christophe Le Priol, « Systemic analysis of the micro-RNAs involved in epithelial cancers ». Co-supervision avec X. Gidrol (CEA Grenoble).
- 2014 – 2017 : Víctor Bellón, « Adverse drug reaction discovery ». Co-supervision avec V. Stoven (CBIO).

Depuis 2014 **Encadrement de stages et projets**, *MINES ParisTech*, Paris (75).

- 2018 : Weiyi Zhang, « Convolutional neural networks for multiplex biological nework ». M1 Shanghai Jiao Tong et MINES ParisTech. Co-supervsion avec A. Rausell (Imagine).
- 2018 : Victor Sorreau, « Machine learning and prediction of variant-induced RNA splicing defects ». L3 Université de Rouen. Co-supervision avec A. Martins (Université de Rouen).
- 2017 : Liyang Sun, « Convolutional neural networks for protein representation ». 2A CentraleSupélec.
- 2017 : Adrien Galamez, Paul Magon de la Villehuchet, Olivier Pham et Manon Revel, « Identifying recurrence in electronic health records ». 2A CentraleSupélec. Co-supervision avec J.-P. Vert (CBIO) et J. Guérin (Institut Curie).
- 2016 : Athénaïs Vaginay, « Multi-phenotype identification of biomarkers in a biological network ». M1 Paris-Diderot.
- 2015 : Killian Poulaud, « Multitask feature selection in a graph ». 2A Supinfo.
- 2014 : Jean-Daniel Granet, « Development and parallelization of the SConES tool for graph-guided GWAS ». 1A École 42.

2011 – 2013 **Encadrement de stages de M2**, *Universität Tübingen*, Allemagne.

- 2013 – 2014 : Udo Gieraths, « Machine Learning for identification of autosomal recessive genomic variants ». Co-supervision avec K. Borgwardt (MPI Tübingen).
- 2012 – 2013 : Fabian Aicheler, « Disease status prediction based on SNP annotation ». Co-supervision avec K. Borgwardt (MPI Tübingen).
- 2011 – 2012 : Valeri Velkov, « Mining correlated loci at a genome-wide scale ». Co-supervision avec K. Borgwardt (MPI Tübingen).

Financements obtenus

2018 – 2019 Financement d'un postdoctorat par Sancare

2016 – 2019 Financement d'une thèse CIFRE par Sanofi

2016 – 2019 Financement d'une thèse ERC COFUND (Institut Curie)

2011 – 2013 Bourse de recherche post-doctorale Alexander von Humboldt

2009 – 2010 Bourse de doctorat IBM

2009 Bourse d'excellence scientifique CINF-Symyx.

Prix

2014 Deuxième position dans la phase I du DREAM 8.5 Rheumatoid Arthritis Responder Challenge

2013 Deuxième position dans le défi 2 du DREAM 8 NIEHS-NCATS-UNC Toxicogenetics Challenge

2007 Premier prix du défi Agnostic Learning vs. Prior Knowledge pour le jeu de données HIVA et présentation invitée au workshop correspondant à IJCNN (International Joint Conference on Neural Networks).

Services professionnels

Depuis 2010 **Relectures pour des revues**.
Parmi lesquelles Annals of Applied Statistics, Bioinformatics, Journal of Chemical Information and Modeling, Journal of Machine Learning Research, IEEE Transactions on Pattern Analysis and Machine Intelligence.
Membre de F1000 Prime.

Depuis 2011 **Relectures pour des conférences internationales**.
Parmi lesquelles International Conference on Machine Learning (ICML), European Conference on Machine Learning (ECML), Knowledge Discovery and Data mining (KDD), Neural Information and Processing Systems (NIPS), Pacific Symposium on Biocomputing (PSB).

Depuis 2016 **Comités de conférences internationales**.
- Responsable du programme (« aera chair ») pour Neural Information and Processing Systems (NIPS) 2016, Women in Machine Learning (WiML) 2017, la Conférence d'Apprentissage (CAp) 2018, les Journées Ouvertes en Biologie, Informatique et Mathématiques (JOBIM) 2018.
- Responsable de la publication pour Artificial Intelligence and Statistics (AISTATS) 2018.

Depuis 2016 **Expertises pour des agences de financement**.
Agence Nationale de la Recherche (ANR, France), US-Israel Binational Science Foundation (BSF)BSF (US-Israel), Mathematics of Information Technology and Complex Systems (Mitacs, Canada), Consortium Québécois sur la Découverte du Médicament (CQDM, Canada).

Depuis 2013 **Organisation d'événements scientifiques**.
- 2017 : Machine Learning in Systems Biology Workshop, ISMB, Prague (République Tchèque)
- 2016 : Compétition internationale DREAM Idea Challenge
- 2015 : Workshop on Features and Structures, ICML, Lille (59)
- 2014 : École d'Été Machine Learning in Personalized Medicine, Paris (75)
- 2013 : École d'Été Machine Learning in Personalized Medicine, Tübingen (Allemagne).

Depuis 2016 **Comités de recrutement**.
- 2018: Concours maître de conférence ENSIMAG Grenoble (38)
- 2016: Concours chargés de recherche (CR1/CR2) INRIA Bordeaux (33).

Depuis 2016 **Comités de thèse**.
- 2017 : Yunlong Jiao, « Rank-based Molecular Prognosis and Network-guided Biomarker Discovery for Breast Cance », PSL Mines ParisTech (75).
- 2016 : Flore Harlé, « Multiple change-point detection in multivariate time series: application to the inference of dependency networks », Université Grenoble Alpes (38).

Exposés invités

Conférences et symposiums.

11/2018 : Soph.IA, Antibes Sophia-Antipolis (06)

10/2018 : AI and Biology, Montpellier (34)

06/2018 : Session « Statistics in genetics research and diagnostics », European Society of Human Genetics (ESHG), Milan (Italie)

05/2018 : dotAI, Paris (75)

05/2018 : Machine learning strategies for disease prediction, Copenhague (Danemark)

05/2018 : Big data approaches in health, disease and treatment trajectories, Novo Nordisk Foundation, Copenhague (Danemark)

02/2018 : Panel « AI & Health », Nuit de l'IA, Paris (75)

12/2017 : Imaging in Paris, Paris (75)

11/2017 : Séance commune Académie des Sciences – Académie de Médecine sur « Mathématiques, mégadonnées et santé : l'exemple du cancer », Paris (75)

11/2017 : Machine learning and molecules, Copenhague (Danemark)

10/2017 : The growing ubiquity of algorithms in society, Royal Society, Londres (UK)

10/2017 : Panel « Women in AI », France is AI, Paris (75)

09/2017 : France-Japan Machine Learning Workshop, Paris (75)

07/2017 : Symposium étudiant ISCB BeNeLux–France, Lille (59)

05/2017 : Workshop « Big data in human genetics », European Society of Human Genetics (ESHG), Copenhague (Danemark)

04/2017 : International Conference on Learning Representations (ICLR), Toulon (84)

03/2017 : Optimization, machine learning, and pluridisciplinarity, Grenoble (38)

10/2016 : From machine learning to personalized medicine, Munich (Allemagne)

06/2016 : Scikit-Learn Day, Paris (75)

05/2016 : Workshop « Big data in human genetics », European Society of Human Genetics (ESHG), Barcelone (Espagne)

04/2016 : Workshop « Machine learning and society », Data Learning and Inference (DALI), Sestri Levante (Italie)

01/2016 : Meetup Paris Machine Learning, Paris (75)

10/2015 : Journées de l'école doctorale Pierre Louis de santé publique, Saint-Malo (35).

Visites.

12/2015 : DIBRIS, Université de Gênes (Italie)

11/2015 : DIKU, Université de Copenhague (Danemark)

06/2015 : DBL, TU Delft (Pays-Bas)

04/2014 : UNATI, CEA Paris-Saclay (91)

02/2014 : MIA-T, INRA Toulouse (31)

07/2013 : SPL-2, EPFL, Lausanne (Suisse)

06/2013 : LIF, Universités Aix-Marseille (13)

03/2013 : EMBL-EBI, Hinxton (Royaume-Uni)

05/2012 : Memorial Sloan-Kettering Cancer Center, New-York (États-Unis)

05/2012 : Université de Notre-Dame (États-Unis)

05/2012 : Broad Institute, Boston (États-Unis).

Diffusion scientifique

11/2018 Conférence grand public sur le machine learning sur le site de MINES ParisTech à Fontainebleau (78).

05/2018 Introduction au machine learning à destination des lycéens membres du Cercle de Mathématiques de Strasbourg (69).

Depuis 2017 Co-fondatrice de la branche parisienne du Meetup « Women in Machine Learning and Data Science » qui organise des rencontres mensuelles à Paris (75).

Depuis 2017 Membre du conseil de France is AI, dont le but est de favoriser le développement d'un écosystème IA français dynamique, par l'organisation d'événements et le soutien à des initatives locales.

Depuis 2017 Membre du jury du Tournoi Français des Jeunes Mathématiciens et Mathématiciennes (TFJM^2) à Strasbourg (75) et en région parisienne (91).

2014 Journée « Filles et maths, une équation lumineuse » à l'EPITA (94).

Langues

Français **Langue maternelle**

Anglais **Bilingue** *5 ans en Californie*

Allemand **C1** *3 ans en Allemagne*

6. Documents administratifs

Sont joints à ce dossier les pièces suivantes :

1. Copie du diplôme de doctorat, délivré par University of California Irvine ;
2. Traduction en français de ce diplôme ;
3. Copie du diplôme d'ingénieur, délivré par l'École Nationale des Télécommunications de Bretagne ;
4. Copie du diplôme de master recherche, délivré par l'École Nationale des Télécommunications de Bretagne ;
5. Copie du passeport ;
6. Dossier administratif de candidature.