2021

DOSSIER DE CANDIDATURE

APPLICATION

Cochez le concours sur lequel vous candidatez
Check the competition exam for which you are applying

☒ **ISFP - (Inria Starting Faculty Position / Inria Starting Faculty Position)**

☒ **CRCN - (Chargés de recherche de classe normale / Young graduate scientist position)**

□ **DR2 - (Directeurs de recherche de deuxième classe / Senior researcher position)**

Nom[1] **:** FEYDY
Last name

Prénom : Jean
First name

Sexe : □ **F** ☒ **M**
Sex

Nom utilisé pour vos publications (facultatif) :
Name used for your publications (optional):

[1] Il s'agit du nom usuel figurant sur vos pièces d'identité
It is the name appearing on your identity cards

DEPOT DE VOTRE CANDIDATURE
SUBMITTING YOUR APPLICATION

Le dossier de candidature doit comprendre :

- Formulaire 1 : Parcours professionnel
- Formulaire 2 : Description synthétique de l'activité antérieure
- Formulaire 3 : Contributions majeures
- Formulaire 4 : Programme de recherche
- Formulaire 5 : Liste complète des contributions

CRCN & ISFP :

- Les rapports de thèse ou de doctorat (si disponibles)
- Une copie des derniers titres et diplômes
- Une photographie récente de la candidate / du candidat (facultative)

DR2 :

- Les rapports d'habilitation à diriger des recherches (si applicable)
- Une copie des derniers titres et diplômes
- Une photographie récente de la candidate / du candidat (facultative)

The application file must include:

- *Form 1: Professional history*
- *Form 2: Summary of your past activity*
- *Form 3: Major contributions*
- *Form 4: Research program*
- *Form 5: Complete list of contributions*

CRCN & ISFP:

- *PhD dissertation reports (when available)*
- *A copy of most recent titles and diplomas*
- *A recent photography of the applicant (optional)*

DR2:

- *Habilitation dissertation reports (if applicable)*
- *A copy of most recent titles and diplomas*
- *A recent photography of the applicant (optional)*

SOMMAIRE / *SUMMARY*

Formulaire 1 — PARCOURS PROFESSIONNEL
Form 1 — PROFESSIONAL HISTORY

1) Parcours Professionnel / *Professional history*

Situation professionnelle actuelle / ***Current professional status***

Statut et fonction[2] / *Position and Status*[2]: research assistant (postdoc).
Etablissement (ville - pays) / *Institution (city - country)*: Imperial College London (London - United Kingdom).
Date d'entrée en fonction / *Start*: 1 September 2019.
[ ] Sans emploi / *Without employment*

Expériences professionnelles antérieures /***Previous professional experiences***

Date début *Start*	Date fin *End*	Etablissement *Institution*	Fonction et statut[2] *Position and status*[2]
1 September 2019	31 August 2021/22	Imperial College London	Postdoctoral researcher, research assistant
1 September 2016	31 August 2019	École Normale Supérieure	Enseignant/chercheur contractuel
1 September 2016	31 August 2019	ÉNS Paris-Saclay	PhD student

Nombre d'années d'exercice des métiers de la recherche après la thèse / *Number of years of professional research experience after the PhD:* 1.

Between 2016 and 2019, I worked as a tutor ("Caïman", "agrégé préparateur") in the Department of Mathematics and Applications (DMA) of the École Normale Supérieure. This position involved significant teaching duties (detailed below) and superseded my doctoral contract.
Please also note that I defended my PhD thesis on July the 2nd, 2020, several months after the start of my postdoc at Imperial College. This is due to the extra care that I took to write an accessible dissertation and to the Covid pandemic. My current contract runs until August 2022, with the understanding that I may leave London this summer if I get offered a permanent position.

2) Interruptions de carrière/*Career breaks*

None.

3) Encadrement d'étudiants et de jeunes chercheurs / *Supervision of students and early-stage researchers*

- February – May 2018: **full supervision** of Hugo Malamut and Noé Sotto for their **bachelor thesis (L3 level)** in the department of mathematics and applications of the École Normale Supérieure (Université PSL). This thesis, titled *Shapes spaces*, provides a short introduction to a family of Riemannian metrics that are commonly used for the diffeomorphic registration of medical images. The aim of this thesis was to introduce these two students to a genuine research topic with a balanced mix of algorithmic, geometric and experimental work.

- September 2020 – February 2021: **partial supervision** (30%) of the **pre-doctoral internship** of Nemo Fournier at Télécom Paristech, with Pietro Gori (the main supervisor) and Pierre Roussillon. This internship on using optimal transport for the segmentation and anatomical study of brain tractograms extends a MICCAI 2019 paper that I wrote with P. Roussillon, A. Trouvé and P. Gori. Nemo is currently improving this methodology and applying it to large-scale data from the Human Connectome Project, handling millions of 3D curves at a time. We expect this work to lead to a publication in a medical imaging journal, while Nemo pursues a PhD on a related subject in neuro-anatomy under the supervision of Stanley Durrleman in the Aramis INRIA team.

4) Encadrement de développements technologiques (logiciel, matériel, robotique) / *Supervision of technological development (software, hardware, robotics)*

- January 2021 – May 2021: **full supervision** of a **software development project (M2 level)** for the Master of Science in Artificial Intelligence at Imperial College London. I am currently working with a group of six students (Yaniel Cabrera, Yihang Chew, Anna Hlédiková, Monika Jotautaite, Stefan Sturluson and Hudson Yeo) on fast approximation methods for geometric learning. The aim of this project is to provide efficient and versatile GPU support for two families of acceleration strategies in numerical analysis:

1. Clustering schemes and iterative methods for approximate nearest neighbor search (IVF-like methods as in the FAISS library, NN-Descent algorithm as in the UMAP package).
2. Nyström and Fourier-based decompositions for point convolutions and kernel matrix-vector products.

For its GPU backend, this work relies extensively on the KeOps library that is presented below as my first major contribution. On top of an implementation and benchmarking work, this project involves in-depth documentation, testing and will eventually be merged in the main branch of the KeOps repository. Our main target is to provide access to a collection of advanced numerical schemes for the wider scientific community. We intend to provide a reference implementation of these methods that is well-documented, efficient and **easy-to-use with any distance or kernel function**.

This project is part of a gradual shift in the development of the KeOps library: after four years of internal development, we are currently opening the door to external contributors on major features. Going forward, I intend to keep supervising students on similar projects. As the KeOps library matures and gains traction in the scientific community, management and maintenance work is bound to become one of my major development activities.

5) Responsabilités collectives / *Responsibilities*

- I am a member of the organizing commitee for the colloquium on "Medical Imaging in an age of Articial Intelligence", organized by the Paris Brain institute (ICM, University Hospitals Pitié–Salpêtrière) in October 2020 (3rd edition) and June 2021 (4th edition). This day-long event was first set up in the Collège de France in 2018; it attempts to bring together academics, clinicians, jurists and industry partners to discuss the challenges that are most relevant to the future of medical imaging in France.

- I regularly review papers for the MICCAI (2019, 2020) and AISTATS (2020) conferences, the Journal of Mathematical Imaging and Vision (JMIV), the SIAM Journal on Imaging Sciences (SIIMS), the IEEE Transactions on Medical Imaging (T-MI), the IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI) and the Journal of Machine Learning Research (JMLR).

6) Management (si pertinent) / *Management (if relevant)*

None.

7) Mobilité (si pertinent) / *Mobility (if relevant)*

- From April to September 2015, I did an **internship at Siemens Healthcare in Princeton, NJ**. Under the supervision of Boris Mailhe and Mariappan Nadar, I wrote my Master's thesis on the real-time denoising of medical images using steerable wavelets.

  This experience in the industry had a **deep influence** on my vision of academic research in healthcare. Most importantly, it opened my eyes on the amount of work that is needed to turn a neat mathematical idea into a genuine product. I discovered some of the inner workings of one of the major constructors of medical imaging devices and remember two major lessons from this stay in the US:

  1. Clinical-grade devices are developed by engineers who are open to novel ideas but have a limited amount of time to get familiar with new theories and tools. High-quality **software development** is therefore an **integral part of academic research** in our field: to have an impact on clinical practice, an original method must first be put "on the shelf" through a well-supported implementation.

  2. In medical sciences, the processing of raw sensor data into a useful piece of information is a complex process that involves many different fields. Medical imaging pipelines result from the work of dozens of engineers and technicians: they can hardly be understood as well-defined "end-to-end" machine learning models. In this context, we must strive to develop **robust**, **interpretable** and **modular tools**. Methods that are well-documented and produce stable results across patients and devices are much more likely to be adopted for widespread clinical practice.

- Since September 2019, I have been working as a postdoc fellow in the **Department of Computing at Imperial College London**, in the team of Michael Bronstein. I am primarily working on geometric deep learning problems: applications to the processing of protein structures are detailed below as my third major contribution.

  Apart from the scientific benefits of this experience (which are detailed in my research project), I remember two main lessons from my stay in London:

1. Even when they work on the same topics, researchers can have vastly **different backgrounds and motivations**. Notably, I have learned to work with students who have a solid background in computer science but little knowledge in mathematics and physics. Interactions with the industry are also much more common and supported across the Channel. After years of training in the French school of mathematics at the ÉNS, this experience in London is bringing me some balance and perspective on the global research community.
2. **Cross-field interactions** are key to scientific research. First of all, in shape analysis: the computer vision, graphics and medical imaging communities have developed complementary ideas over the last two decades and have a lot to learn from each other. Going further, exchanges with Michael Bronstein have allowed me to appreciate the value of unexpected collaborations with e.g. biologists or food scientists. Building bridges from the ground up with communities that do not share our mathematical training requires a significant amount of work, but can have a deep and lasting impact.

8) Enseignement (si pertinent) / *Teaching (if relevant)*

Between 2016 and 2019, I worked as a tutor ("caïman", "agrégé préparateur") in the Department of Mathematics and Applications (DMA) of the École Normale Supérieure. This position superseded my doctoral contract and involved significant teaching duties in the L3 and M1 programs of mathematics at the ÉNS (Université PSL):

Class name, format.	Years taught, hours, level.	Subjects.	Original teaching material.
Mathematical culture for non-mathematicians, **full lectures.**	2017, 2018, 24 hours per year, **L3-M1.**	Elementary logic and number theory, Fourier and shape analysis, Riemannian geometry, finite element method.	Lecture notes (245 pages).
Introduction to Riemannian geometry through the study of shapes spaces, **reading group.**	2017, 2019, 24 hours per year, **L3**.	Riemannian geometry, computational anatomy.	Lecture notes (236 pages), written with the students.
Mathematical foundations of data sciences, **workshop sessions** for Gabriel Peyré's lectures.	2018, 2019, 24 hours per year, **M1**.	Wavelets, convex optimization, optimal transport, neural networks.	Workshop notes (122 pages), 5 notebooks to supplement the Numerical Tours website.

I felt a great responsability towards my students, who got selected through a competitive exam and are likely to become part of a new generation of French researchers in mathematics and computer science. In this context, the ÉNS puts an emphasis on original classes that let students engage with modern research topics. I was given *carte blanche* for these lectures and was invited to showcase **applied topics with strong theoretical foundations** that could appeal to young mathematicians.

I must note that the ÉNS has a long history of promoting pure mathematical research, which can have an adverse psychological effect on students who feel inclined to work on applied problems. This is a concern for the teaching staff, who intend to let students walk their own paths without prejudice or latent peer pressure. For these reasons, tutors are asked to act as relatable "role models" and honest advisors for students in the maths department.
On top of the lectures above, I thus devoted one to two hours per week to one-on-one meetings with young tutees: I worked as a **tutor for six L3 students** per year who had chosen a dual "mathematics + computer science" curriculum and generally acted as a **mentor for M1 students** who were thinking about pursuing a career in data sciences. For a majority of students, this involved giving advice on applied M2 programs, internships and discussing career plans. Some students with deeper confidence issues or administrative problems also required greater care: I was especially considerate of female and foreign students, who are in the minority in the department of mathematics. In all cases, I tried my best to foster a positive atmosphere around these students and remind them of the many open paths that are available to them after graduation.

Overall, during my PhD, I dedicated **a third of my time** to teaching and tutoring activities at the ÉNS. This experience allowed me to appreciate the **diversity** of students' profiles, hone my **communication** skills and broaden my **mathematical background**. Even if my primary research interests remain related to geometry (in a broad sense, from shape analysis to high-dimensional data sciences), these years of teaching now allow me to hold productive discussions on e.g. convex optimization and imaging sciences.

9) Diffusion de l'information scientifique (si pertinent) / *Dissemination of scientific knowledge (if relevant)*

Making sure that medical doctors are not mystified by "artificial intelligence" and "deep learning" methods is a necessary first step to engage productive collaborations. To this end, I have given accessible tutorials on image processing and computational anatomy in several events:

- I am in charge of a tutorial on computational anatomy at the "Artificial intelligence and medical imaging" masterclass. This 3 days long event is held every year by the Université de Paris since February 2019. It primarily targets **radiology interns** in the wider Paris area, with about 30 students per year. In collaboration with Tom Boeken from the Georges Pompidou Hospital (Paris), I created original notebooks to present the main ingredients of image processing programs to our end-users: **medical doctors and technicians**.

- In May 2020, I presented a short introduction to machine learning research in the "DIU Neuroradiologie vieillissement", an inter-university program for the continuous training of French **radiologists**.

- In September 2018, I was in charge of a tutorial on computational anatomy at the GeomData summer school. This week-long event was attended by **graduate students and engineers** as part of a vocational training program.

In the same vein, I have also been invited to speak in several radiology conferences:

- In June 2018, in a duo with Dr Francis Besse from the Centre Cardiologique du Nord (Saint-Denis), I gave the **opening talk** of the annual congress of the **French Society for CardioVascular Imaging** (JFICV 2018, Beaune).

- In October 2020, I presented an introduction to scientific computing in medical imaging at the 3rd Colloquium on Medical Imaging in an age of Artificial Intelligence, organized by the **Paris Brain Institute.**

- I gave shorter presentations in the annual congress of the **French Society for NeuroRadiology** (SFNR 2019, Paris), in the Harvey Cushing symposium (June 2019, **American Hospital of Paris**) and in a training day for medical technicians that was organized by **Canon Medical Systems**. Talks at the annual congress of the **French Society of Medical Imaging for Women** (SIFEM 2021, Ajaccio) and at the **French Academy of Medicine** have been postponed due to the Covid pandemic and will be held in 2021.

10) Visibilité (si pertinent) / *Visibility (if relevant)*

I have been invited to several international colloquiums, organized both as week-long workshops (which is common in mathematics):

- The 5 days workshop for people in optimal transportation and applications, organized in June 2019 by the Scuola Normale Superiore in Cortona, Italy.

- The 5 days workshop on shape analysis, stochastic mechanics and optimal transport, organized in December 2018 by the BIRS center in Banff, Alberta.

- The 5 days seminar on optimal transport and hydrodynamics, organized in October 2018 by the Oberwolfach Research Institute for Mathematics in Germany.

- The 5 days workshop on shape analysis and computational anatomy, organized in November 2017 by the Isaac Newton Institute in Cambridge, UK.

And as collections of parallel sessions (which is common in computer science):

- The SIAM-IS20 conference on imaging science (Toronto/online), in a session on optimal transport.

- The Curves and Surfaces 2018 conference (Arcachon), in a contributed session.

- The SIAM-IS18 conference on imaging science (Bologna), in a session on geometric methods for shape analysis.

11) Eléments divers / *Other relevant information*

None.

Formulaire 2 — DESCRIPTION SYNTHÉTIQUE DE L'ACTIVITÉ ANTÉRIEURE
Form 2 — SUMMARY OF YOUR PAST ACTIVITY

I am primarily interested in **shape analysis** and **data sciences** for **medical applications**. I work at the intersection of geometry and machine learning, optimal transport and computer vision, measure theory and high performance computing. In the wider context of data sciences, medical problems are charaterized by the need for **robust, interpretable methods** and the **heterogeneity of data sources** – from electronic health records to the many different types of medical images. Most often, **medical samples are small and high-dimensional**. As each patient is unique in terms of genetic and physiological characteristics, illness trajectories and treatments, even large-scale datasets often have to be stratified into smaller "homogeneous" cohorts. Medical records are then encoded as time series of clinical examinations, diagnoses and prescriptions that contain a significant amount of **missing data**. Dealing reliably with the **heavy tail** of our populations is also a priority. First, because we tend to be more interested in the detection of varied pathologies than by the condition of average "healthy" patients. Second, because some rare outcomes of medical decisions may well be deadly.
These constraints put our field at the intersection of statistics and mathematical modelling. We strive to develop **domain-specific methods** that leverage **expert medical knowledge** to generalize reliably outside of our challenging datasets.

The deep learning revolution. Over the last decade, the emergence of deep learning frameworks (Theano, TensorFlow, PyTorch, etc.) has had a tremendous impact on the field. These Python libraries provide a convenient interface for automatic differentiation engines, with full support for Graphics Processing Units (GPU). They enable the fast prototyping and optimization of complex models, unlocking significant progresses in e.g. image processing, computer vision and natural language processing. In medical imaging, the most impactful work of the past few years is probably the U-Net architecture for tissue segmentation (MICCAI 2015), which has paved the way for fully automatic methods in computational anatomy.
At a deeper level, the emergence of these numerical tools is completely changing the way we code and share ideas. Research teams around the world now favour **modular, inter-operable Python libraries** over grand C++ or MATLAB toolkits. In computational anatomy, this new development paradigm allows students to get started in days instead of months and **unlocks fruitful interactions** with neighboring research communities. This stimulating environment has acted as a catalyst for research in the field, inducing a growing interest for "artificial intelligence" among medical doctors.

Limitations. Surprisingly though, these progresses are built upon numerical foundations that remain relatively narrow. Even if they are backed by major industry players, deep learning frameworks **cannot provide cutting edge support for the full range of operations** that are of interest to applied mathematicians. In practice, they tend to focus on image convolutions and linear algebra routines for dense matrices, to the detriment of other algorithmic structures. There are good reasons for this design choice, but the situation is progressively putting our field under severe constraints: a widening performance gap is growing between models that benefit from first-rate industry support, and other methods. This forces us to compromise between numerical efficiency and the need to **encode medical hypotheses within our models**.

Providing efficient and versatile tools for the community. Getting the best out of the deep learning toolbox while **retaining complete modelling freedom** is therefore a key challenge for the community. In particular, I am convinced that geometric, non-smooth and non-Euclidean methods have an important role to play in data sciences. In light of the points above, I thus decided to focus on **breaking through some of the major computational bottlenecks** in the field:

1. My KeOps library provides **efficient support for distance- and kernel-like matrices.** It is having a game-changing impact in many applied fields and provides the numerical foundations of my work.
2. My GeomLoss library puts **scalable and robust optimal transport** on the shelf for applications to machine learning and computational anatomy. After four years of work on theoretical and numerical aspects, I am currently applying these tools to genuine clinical problems and packaging them for a wide user base.
3. Going further, I am now relying on the tools above to design **data-driven shape models**. This work is still very much in progress: it is presented in Chapters 4 and 5 of my PhD thesis and in a recent pre-print on protein structures.

Bridging the gap between mathematical research and clinical practice. These contributions build on top of each other and are detailed in Form 3. My primary target is to **bring effective geometric ideas to a wider audience** in medical sciences. As detailed in Form 1, I thus believe that **reaching out to engineers** and students who do not share my mathematical background is a top priority. I am investing a significant amount of time in making my work easily accessible, from the documentation of my libraries to my PhD thesis that is written as a textbook for newcomers in the field.
All of this work is part of an **ambitious research project** that I started to plan more than five years ago, and which I intend to pursue at INRIA. It motivated my long-term investment in **both theoretical and numerical skills**, with the results that are presented below. As these novel geometric tools progressively mature and gain traction in the community, I am now feeling ready to start direct collaborations with medical doctors on genuine clinical problems.

Formulaire 3 — CONTRIBUTIONS MAJEURES
Form 3 — MAJOR CONTRIBUTIONS

Taille maximum de cette partie / Maximum size for this part :

- **CRCN & ISFP – 3 fiches**
  Taille maximum de cette partie : 3 pages
  Maximum size for this part: 3 pages
- **DR2 – 5 fiches**
  Taille maximum de cette partie : 5 pages
  Maximum size for this part: 5 pages

Remplir une fiche par contribution majeure (5 au plus pour les candidatures DR2 — 3 au plus pour les candidatures CRCN & ISFP).

Il peut s'agir d'une contribution scientifique donnant lieu à un ensemble de publications (voire une publication majeure), ou à un développement technologique (logiciel, matériel, robotique ou autre), d'une action de transfert industriel ou sociétal, d'une responsabilité collective, d'une activité d'animation d'une communauté de recherche, ou tout autre élément relevant des missions d'un chercheur ou d'une chercheuse. Les critères importants sont la créativité, l'originalité et l'impact. Chaque fiche suivra le plan indiqué ci-dessous. Dans l'ensemble du texte, pensez à donner, le cas échéant, les références permettant de consulter sur le Web les documents mentionnés (articles, thèses, logiciels, etc.).

Pour les logiciels, fournissez une autoappréciation selon le canevas disponible dans le document ń Criteria for Software Self-Assessment ż disponible à l'URL
https://www.inria.fr/sites/default/files/2021-01/Criteria%20software%20self%20assessment.pdf.

Pour les actions de transfert (transfert technologique ou sociétal), fournissez une description selon le canevas disponible dans le document ń Evaluation des contributions scientifiques en matière de transfert / Guide méthodologique ż disponible à l'URL
https://www.inria.fr/sites/default/files/2020-01/2018-06-GuideMethodologique_EvaluationTransfert%281%29.pdf.

Fill in one form for each major contribution (at most 5 for the DR2 candidates — 3 for the CRCN & ISFP candidates).

It may be a scientific contribution expressed through a set of publications (or a single major publication) or through a technological development (software, hardware, robotic, or other); it may also be an industrial or a societal transfer, a participation to the management of research or to the animation of a scientific community, or any other element. The main criteria are creativity, originality and impact. Each form should follow the guidelines given below. In the body of the text, give the Web references for quoted documents (articles, dissertations, software,...), if available.

For software, please use the ń Self-assessment software criteria ż guideline, available at the URL
https://www.inria.fr/sites/default/files/2021-01/Criteria%20software%20self%20assessment.pdf.

For transfer actions (technology or society transfer) please describe your achievements following the guidelines ń Evaluating scientific contributions in relation to transfer / Methodological guide ż available at the URL
https://www.inria.fr/sites/default/files/2020-01/2018-06-GuideMethodologique-EvaluationTransfert_EN%281%29.pdf.

Fiche 1 : KeOps: fast geometric learning with symbolic matrices

1. Description de la contribution / *Description of the contribution*

Purpose. The KeOps library is an extension for PyTorch, NumPy, Matlab and R that provides efficient **GPU support** for **"symbolic" matrices** $M_{i,j} = F(x_i, y_j)$ whose coefficients are given by a mathematical formula F that is evaluated on (relatively) small data arrays $(x_1, \ldots, x_N)$ and $(y_1, \ldots, y_M)$. Among other examples, we think of distance and kernel matrices: tensors that are not sparse in the traditional sense, but can nevertheless be represented efficiently thanks to their low Kolmogorov complexity. As far as users are concerned, *symbolic* arrays are as easy to use as *sparse* matrices through a transparent "`LazyTensor`" wrapper. KeOps supports generic mathematical formulas F, an arbitrary number of variables "x_i, y_j" and batch dimensions. Under the hood, it combines a just-in-time compilation engine with a collection of efficient map-reduce schemes on CPU and GPU, written in C++/CUDA.
Main results. KeOps speeds up a wide range of computations with a transparent Python interface and full support for automatic differentiation. It is especially useful for small- and medium-sized geometric problems, when dealing with clouds of 1K to 1M samples in dimension 1 to 100. In this setting, KeOps generally provides a **x10-x100 speed-up** when compared with PyTorch, TensorFlow or JAX GPU baselines while keeping a **linear** $O(N+M)$ **instead of a quadratic** $O(NM)$ **memory footprint**. It allows researchers to use **arbitrary metrics and formulas** with run times that are competitive with handcrafted C++/CUDA kernels. Going further, KeOps supports advanced features such as blockwise sparsity masks: it lets users implement complex numerical schemes from the comfort of a differentiable and tensor-centric interface.

2. Contribution personnelle de la candidate ou du candidate / *Personal contribution of the applicant*

Team. I have been developing KeOps since 2017, in close collaboration with Benjamin Charlier (Université de Montpellier) and Joan Glaunès (Université de Paris). We received help from François-David Collin (continuous integration) and Ghislain Durif (R interface), two research engineers from the Université de Montpellier. Since 2020, we have started to welcome external contributors on our GitHub repository, mostly for minor bug fixes and extensions to our mathematical engine.
Personal contributions. B. Charlier, J. Glaunès and myself have contributed equally to the library. We meet every two weeks to take design decisions and have always worked on an equal footing. Of course, some features are closer to some of the team members: for instance, I implemented most of the support for block-wise sparsity masks, generic reduction operators and batch processing. I also wrote the vast majority of our documentation and of our NeurIPS and JMLR papers.

3. Originalité et difficulté / *Originality and difficulty*

State of the art. The acceleration and compilation of scientific code is an active research topic that attracts significant investment from industry players. In practice, all tools available face a **trade-off** between **speed**, **ease of use** and compatibility with varied **hardware** accelerators. Some notable projects in this field are TVM and Halide, that support a wide range of hardware chips but require significant expertise on parallel programming to be used efficiently; Numba, which is a reference library on CPU but does not abstract the intricacies of low-level GPU programming to end-users; the PyTorch JIT engine, TensorFlow and JAX/XLA that are both versatile and easy to use but do not implement optimal C++/CUDA schemes for e.g. kernel matrix-vector products and nearest neighbor search.
Our approach. In sharp contrast with the generalist frameworks above, KeOps only targets a "single" problem: reductions on symbolic matrices. This choice is motivated by our **background in applied mathematics**, as we acknowledge that this family of computations is the bottleneck for a very wide range of mathematical methods. This focus allows us to reach **optimal run times** in many settings while being **easy to use** for both beginners and advanced programmers.
From a technical perspective, the development of this library posed three major challenges:

1. Unlike most handcrafted C++/CUDA kernels, we support operations on symbolic matrices in **full generality** – from K-Nearest Neighbors queries in hyperbolic spaces to generic point convolutions in geometric deep learning.
2. To fit in the deep learning era, **automatic differentation** is a must-have feature. Since we cannot use e.g. the PyTorch autograd module to differentiate our C++ routines, we must implement an autodiff engine of our own.
3. To be accepted by the wider scientific community, our work has to **blend seamlessly** with the standard software stack and be **fully documented** with an extensive collection of benchmarks, demos and tutorials.

Overcoming these issues was only made possible by a years-long investment in a **tailor-made compilation and mathematical engine**, which is fully described on our website. This effort was motivated by a strong belief in the relevance of advanced geometric methods, and the desire to **see elegant theoretical ideas scale up to real clinical data**. Most importantly, working on this project allowed me to gain a significant expertise in GPU programming and software development while broadening my knowledge in computational mathematics. KeOps now **fills an important niche** in the open source ecosystem and is bound to stimulate research on e.g. non-Euclidean machine learning methods. Going forward, our main priorities will be to keep up with technological advances (such as the NVRTC compilation framework or the new Tensor core instructions) and progressively add support for approximate reduction schemes and advanced numerical methods.

4. Validation et impact / *Validation and impact*

KeOps has been very well received by other **library developers**. For instance, it has been adopted as an efficient GPU backend by GPyTorch (from the Universities of Pennsylvania, Cornell, Columbia) and Falkon (from the University of Genoa and the Sierra INRIA team), two libraries for large-scale Gaussian Processes; the Deformetrica software for computational anatomy (from the Aramis INRIA team); the Gudhi library for topological data analysis (from the DataShape INRIA team). This library has also had a **game-changing impact** on my research and the work of my collaborators. Going further, KeOps is progressively gaining traction in a wide range of applied and theoretical fields, from the study of dynamical multi-agent systems to graph matching algorithms. Our papers have recently been accepted in the two most prestigious venues for machine learning research: we expect that the awareness around this package will grow steadily over the next few years.

5. Diffusion / *Dissemination*

KeOps is distributed under the permissive MIT license. It is freely available on our website (www.kernel-operations.io), on GitHub and on the PyPI (`pip install pykeops`) and CRAN (`install.packages("rkeops")`) repositories. As detailed in Form 5, our papers have recently been presented at NeurIPS 2020 (spotlight talk) and accepted for publication in the Journal of Machine Learning Research. `Family=research; Audience=community; evolution=lts; Duration=4; contribution=leader, devel, softcont; Url=https://www.kernel-operations.io/`

Fiche 2 : GeomLoss: fast, scalable and robust optimal transport solvers

1. Description de la contribution / *Description of the contribution*

Context. Optimal Transport (OT) generalizes sorting to spaces of dimension $D \geqslant 1$. It is a fundamental tool in pure and applied mathematics, inducing the *Wasserstein* or *Earth Mover's* distance between probability distributions. The effective resolution of the OT problem has attracted significant interest since the 1940's: depending on the context, efficient methods may **leverage the structure of the distributions** or compromise on accuracy to reach optimal run times.
Results. On the **theoretical** side, I showed that entropy-regularized OT and the associated Sinkhorn algorithm induce pseudo-distances on spaces of probability measures that interact well with optimization routines. On top of being differentiable, **debiased** Sinkhorn divergences are **positive, definite, convex and metrize the convergence in law**. They behave as smooth regularizations of the OT cost, with properties that interpolate between the Wasserstein distance and dual Hilbert–Sobolev norms (known as kernel MMDs in the machine learning literature). On the **practical** side, I designed a multiscale Sinkhorn method that behaves as a **generalized "Quicksort" algorithm** for weighted distributions of samples in metric spaces. For typical use cases in machine learning and shape analysis, my reference implementation GeomLoss for PyTorch outperforms standard baselines by **one to three orders of magnitude**. It is easy to use and scales up to millions of samples in seconds, opening the door to the widespread use of OT as a foundational tool for higher-level methods.

2. Contribution personnelle de la candidate / du candidat / *Personal contribution of the candidate*

Team. My work builds upon a decade of active research on computational optimal transport in the Paris region. I keep close ties with the NORIA ERC project managed by Gabriel Peyré (École Normale Supérieure) and the MoKaPlan INRIA team, in Paris. My closest collaborators on this project are Gabriel Peyré, François-Xavier Vialard (Université Gustave Eiffel) and their PhD student, Thibault Séjourné. I developed applications to computational anatomy with Alain Trouvé (ÉNS Paris-Saclay), Pierre Roussillon (École Normale Supérieure) and Pietro Gori (Télécom ParisTech).
Personal contributions. I benefitted immensely from this stimulating environment, but always worked with a great deal of autonomy. We split our informal group in non-overlapping projects, with e.g. the focus of Thibault Séjourné on extending these results to the unbalanced setting: **the work that is presented here is mostly my own.** I proved our core Theorems on debiased Sinkhorn divergences (positivity, definiteness) in a duo with François-Xavier Vialard; I pioneered applications of OT theory to diffeomorphic registration, and analysed brain tractograms that were provided by Pietro Gori with Pierre Roussillon. I worked alone on numerical aspects, with results that exceeded by far the expectations of our community.

3. Originalité et difficulté / *Originality and difficulty*

State of the art. Over the last decade, **multiscale solvers** have been studied extensively by e.g. Quentin Mérigot, Bruno Lévy and Bernhard Schmitzer; these methods get close to **log-linear run times** in spaces of dimension 2 or 3 but are hard to implement on the GPU. In the machine learning literature, **entropy-regularized OT** and a family of related Sinkhorn algorithms have become standard; these methods **stream well on the GPU**, but induce several geometric artifacts and become prohibitively slow when trying to approximate a genuine OT problem with a relative accuracy of 5% to 1% or better.
My approach. I built upon my **cross-field knowledge** of the literature to get the best of both lines of work. Thanks to a deep understanding of the **geometric structure** of OT problems and my **expertise in GPU computing**, I was able to propose a multiscale Sinkhorn solver that outperforms previous approches by orders of magnitude on typical problems

in shape and data analysis. My pragmatic GPU implementation leverages the KeOps library to its full extent (with block-wise sparsity masks, stabilized log-sum-exp reductions, etc.) and can now be used with any metric space whose distance function is known in closed form. On the theoretical side, my **original geometric proofs** have brought important guarantees of **robustness** to practitioners. After more than 20 years of use of entropy-regularized OT for measure- and shape-fitting applications, I showed for the very first time that a divergence which is derived from the Sinkhorn algorithm could be positive, definite and be used in optimization pipelines **without introducing any shrinkage** of the measures' supports.

4. Validation et impact / *Validation and impact*

My theoretical results on entropy-regularized OT have been well-received by the machine learning community. They have cemented the use of **debiased** Sinkhorn divergences as reliable approximations of the Wasserstein distance, especially in low-dimensional spaces. On the practical side, my GeomLoss package is quickly **redefining the state of the art** for discrete OT. I intend to integrate it as a backend for higher-level packages such as the popular POT library.

5. Diffusion / *Dissemination*

GeomLoss is distributed under the permissive MIT license. My documentation, source code and Python package (`pip install geomloss`) are all freely available online. As detailed in Form 5, we presented our theoretical results on debiased Sinkhorn divergences at AISTATS 2019. Applications to computational anatomy were presented at MICCAI 2017 (oral presentation), at the ShapeMI 2018 MICCAI workshop (oral presentation) and at MICCAI 2019. Chapters 3 and 4 of my PhD thesis contain a detailed presentation and evaluation of the multiscale Sinkhorn algorithm that will be published in a future journal paper. `Family=research; Audience=community; evolution=lts; Duration=2; contribution=leader, devel, softcont; Url=https://www.kernel-operations.io/geomloss`

Fiche 3 : differentiable-MaSIF: fast end-to-end learning on protein surfaces

1. Description de la contribution / *Description of the contribution*

Proteins' biological functions are defined by the geometric and chemical structure of their 3D molecular surfaces. Our deep learning method processes raw atomic point clouds to predict the locations of binding sites on these surfaces.

2. Contribution personnelle de la candidate / du candidat / *Personal contribution of the candidate*

Team. Prior to my arrival in the team, my co-authors Bruno E. Correia (EPFL), Michael Bronstein (Imperial College) and their students had already published a first version of this work, MaSIF, which relied on mesh convolutional networks. It has been well-received, being presented on the cover of Nature Methods in February 2020.
Personal contributions. I contributed to a complete revamp of MaSIF, replacing all the mesh-based modules with fast differentiable layers that require no pre-processing. Using the KeOps library, **I designed and implemented the three main building blocks** of our new method: data-driven chemical and geometric descriptors, that extract relevant features from the raw atomic point clouds; a fast sampling layer for protein surfaces; a quasi-geodesic convolution layer on oriented point clouds. The setup of the global network architecture and experimental evaluation were then done by Freyr Sverrisson.

3. Originalité et difficulté / *Originality and difficulty*

Recent approaches to the binding problem must choose between representations of proteins that have complementary strengths and weaknesses: surface meshes let models focus on the relevant parts of a protein, while full 3D point clouds or volumetric representations stream very well on GPUs. Thanks to the KeOps library, which allowed me to implement an original and complex network architecture from scratch, our method can now get the best of both worlds. We encode **relevant modelling hypotheses** on the protein binding problem without compromising on precision and run times.

4. Validation et impact / *Validation and impact*

I include this work in progress as an example of a **productive collaboration with domain experts, on real data**. In practice, our new architecture compares favourably with the state of the art method MaSIF while being orders of magnitude faster, lighter and easier to use. We are now working on extending this method for *de novo* protein design. Experimental evaluation for some of our protein docking predictions is currently under way in Bruno Correia's wet lab at EPFL.

5. Diffusion / *Dissemination*

Our pre-print and public code release are freely available online. `Family=vehicle; Audience=partners; evolution=basic; Duration=1; contribution=leader; Url=https://github.com/FreyrS/dMaSIF`

Formulaire 4 — PROGRAMME DE RECHERCHE
Form 4 — RESEARCH PROGRAM

☒ Je souhaite candidater dans l'équipe-projet, ou les équipes-projets suivante(s) : HeKA.

Intitulé du programme de recherche : **Geometric data analysis for healthcare.**

Introduction. As discussed in Form 2, my work has always been motivated by medical applications. Going forward, I intend too keep working on **fundamental** theoretical and algorithmic challenges in the field while starting **direct collaborations with medical doctors**. I organize my project in the HeKA team in three major axes: (1) A sustained effort on the numerical foundations of our field, to provide new modelling options for exciting research works. (2) A pragmatic project on shape analysis, with a focus on designing data-driven yet robust and reproducible baselines in computational anatomy. (3) A long-term transition towards data analysis in non-Euclidean spaces. As always, I put a strong emphasis on making my results easy to use by non-specialists, both within and outside of medical sciences.

1. The KeOps and GeomLoss libraries: fast and robust numerical foundations

KeOps: project. The development, impact and short-term future of the KeOps library are detailed in Form 3 (first major contribution) and Form 5 (Section 3). Needless to say, I am **responsible towards our sizeable user base** and intend to provide long-term support for this project. Some of the related tasks are straightforward: extensions of our mathematical engine, interactions with users on GitHub, etc. They could be handled in part by a research engineer.
On the other hand, other improvements will rest directly upon my shoulders. For instance, the progressive integration of **approximation strategies** for nearest neighbor queries and symbolic matrix-vector products will be most challenging. These methods require a significant expertise in both numerical analysis and GPU computing to be implemented efficiently, and have often never been tested and studied in **full generality**. In the short term, as discussed in Form 1 (Section 4), I intend to focus on relatively simple schemes such as the Nyström method and IVF-type strategies. In the long term, I target the integration of more complex strategies such as the Fast and Free Memory method and Hierarchical matrices.
KeOps: collaborations. Thanks to years of prior development, the acceleration (by orders of magnitude) of common methods that are relevant to the HeKA team is now a relatively low-hanging fruit. For instance, KeOps will allow us to scale up computations on Weighted Cumulative Exposure models to large datasets from the Assurance Maladie, while retaining complete modelling freedom on possible extensions to e.g. personalized illness trajectories. This opens the door to the development of complex exposure models to varied diseases, which is of utmost interest to **Sarah Zohar**, **Anne-Sophie Jannot** and **Pierre Sabatier** for their research in etiology. Beyond these first projects, KeOps is integral to all the work that is discussed below: keeping this package at the cutting edge of computational mathematics remains a major priority.

GeomLoss: project. The short-term future of GeomLoss is detailed in Form 3 and Form 5 (Section 3). As with KeOps, I am **responsible towards a growing user base** and intend to provide long-term support for this project. Progresses in computational optimal transport are key to my research program and leverage efficient solvers on grid images, generalized point clouds and histograms with a fixed support. In the long run, two major settings that I intend to study in depth are optimal transport on graphs and surface meshes, endowed with the geodesic distance.
GeomLoss: collaborations. I will remain close to the optimal transport scene in the Paris region – including Rémi Flamary (École Polytechnique), the main developer of the popular POT library. On the theoretical side, a major point that remains to understand rigorously is the stunning effectiveness of annealing strategies for the Sinkhorn method: this is a question that I am currently investigating with Jean-David Benamou (MoKaPlan INRIA team, Paris).

2. Computational anatomy

Project. A primary motivation for my work has always been to unlock the full potential of data-driven ("deep learning") methods in computational anatomy. Even if KeOps and GeomLoss are of interest to the wider scientific community, I have primarily designed them to provide numerical foundations for a **new generation of methods in computational anatomy**. As discussed in Form 2 as well as Chapters 1 and 5 of my PhD thesis, statistical shape analysis has undergone a major transition over the past five years. On the one hand, the advent of reliable segmentation networks for anatomical tissues has made **shape analysis increasingly relevant** in medical imaging. On the other hand, GPU-based Python codes have progressively made obsolete C++ and Matlab toolboxes – but no user-friendly library has yet truly emerged to make recent methodological advances available to a wide user base. In this context, one of my major future projects will be to develop a pragmatic and **modular "scikit-shapes" library** that can be used both as a reference tool by practitioners and as a platform for interactions with neighboring fields. Through my PhD and postdoc, I now have first-hand experience of complementary approaches to shape analysis in **computational anatomy** (splines, diffeomorphic registration, etc.) and **computer vision** (morphable models, autoencoders, etc.), while I keep an acute interest in related works from the **computer graphics** literature (functional maps, elastic models, etc.). I am thus in a perfect position to take on this project and attempt a partial **convergence between these three lines of work**.

In the short term, I intend to focus on **registration methods** that we decompose in three successive steps: **feature extraction** from the raw shape or image data using e.g. a convolutional neural network; **feature matching** using a (soft) nearest neighbor projection or an optimal transport solver in a high-dimensional feature space; **regularization** using a simple smoothness prior. I have already started preliminary work on this question, that covers most state of the art methods for **large deformations**. It will provide a solid foundation for ulterior works and collaborations.

In the long term, a key target will be to investigate complex and data-driven regularization priors – which are often understood as **shape metrics** in computational anatomy. The work of Marc Niethammer is one of several sources of inspiration; I intend to build upon these ideas as well as my work on **geometric deep learning** for feature extraction to design models that are more expressive than translation-invariant baselines while remaining trainable on relatively small medical samples.

Collaborations. The applications of this project in the HeKA team are two-fold. First of all, my expertise in image registration and shape analysis will allow me to solve minor but recurrent geometric issues in the use of e.g. multi-modal images. A major target of this project is to establish a **clear and modern baseline** for standard tasks in computational anatomy, thus allowing us to focus our efforts on key challenges and data types.

Second, and most importantly, efficient shape models will allow me to describe the anatomy of patients using reliable low-dimensional vectors (known as "codes" in the deep learning literature). Following a methodology that is now standard in the field (but whose results depend entirely on the **quality and robustness of the underlying atlas model**), this compact description of shape data will then let us **handle anatomical images with robust statistical methods and optimization routines**. This is a line of work that I intend to develop extensively with **Stéphanie Allassonnière** and her students on e.g. the early detection of the agenesis of the corpus callosum from fetal MRI data, in collaboration with the Trousseau Hospital.

This project will also be of utmost interest to the Aramis INRIA team (in Paris) and the wider community for shape analysis in medical imaging. I will keep close ties with my collaborators at EPFL and UNC Chapel Hill. Long-term, I also intend to interact with the Geometric and Visual Computing lab at École Polytechnique and the Epione INRIA team, in Sophia.

3. Data sciences in non-Euclidean spaces

Project. As discussed in Form 2, medical data is **highly heterogeneous**: we strive to handle time series of images, tensors, graphs, texts or drug prescriptions within a common framework to deliver a reliable piece of medical information. In this context, an appealing strategy is to model data samples as points in a known metric space, on which we can then apply a standard machine learning algorithm. This allows us to **encode expert knowledge** in the structure of a data space while **retaining strong guarantees** of statistical consistency and convergence of our optimization methods. We can understand the framework of geometric statistics as an extension of the standard "kernel method" to non-linear embedding spaces: it is flexible enough to let us describe e.g. hierarchical data (in hyperbolic spaces or discrete graphs) and anatomical shapes (in a high-dimensional Riemannian manifold) with a **common, principled language**.

Going forward, I intend to build upon the **game-changing flexibility** that is provided by KeOps and GeomLoss to study the suitability of optimal transport distances and domain-specific Riemannian metrics for large-scale medical problems. An inspiring example is the UMAP library for non-linear dimension reduction, which has succeeded in bringing advanced geometric ideas to life for e.g. the analysis of single-cell sequencing data. Long-term, I intend to leverage fast solvers for the Wasserstein barycenter problem to propose **simple and robust baselines for machine learning with shapes and histograms**. In the spirit of recent works on structured learning by Giulia Luise (Imperial College) and Alessandro Rudi (Sierra INRIA team, Paris), I intend to push forward the study of deterministic, convex and reliable methods in the field.

Collaborations. This project fits well with the work of the HeKA team on patient representation. From a deep learning perspective, geometric data spaces can be understood as "**mathematical auto-encoders**" that are usually known in closed form. They are best suited to cases where the relationship between any two data points can be described using a semi-explicit model, e.g. for drugs that are naturally represented in a hierarchical ATC classification or anatomical shapes that we can compare with each other using either an optimal transport distance or a more accurate Riemannian metric.

Going further, this geometric perspective on data sciences is being actively studied abroad, with growing interest in the machine learning community. In France, medical applications of geometric statistics have been primarily studied in the Epione INRIA team (ex-Asclepios, in Sophia), with inspiring and highly impactful work on diffusion tensor imaging and computational anatomy. The integration of my work within its GeomStats library would be a natural outcome.

Closer to Paris, this project could also lead to exchanges with the Parietal and DataShape teams at INRIA Saclay, where KeOps is already being used. Even if the research interests of these teams differ from my own focus on clinical practice with HeKA, creating **long-term synergies** on time series, geometric optimization or high-dimensional data analysis will certainly be relevant for all parties involved.

Relationship with network analysis. From a broader perspective, I understand "Riemannian" geometric data analysis as an **intermediate setting** between standard (Euclidean) data sciences and graph processing. Curved metric spaces whose distance function is known in closed form strike a good balance between performance and flexibility. They allow us to model fairly complex situations while retaining the excellent numerical properties of standard Euclidean methods.
I am currently getting first-hand experience on the fruitful interactions that can take place at the junction of **Riemannian geometry and network analysis**. My supervisor at Imperial College, Michael Bronstein, has built a large part of his career on cross-field exchanges between 3D shape analysis and graph processing within the common framework of geometric deep learning – the now standard dynamic graph CNN being one of the most illustrative examples. Even if our research interests currently have little overlap, I am thus looking forward to discussions with **Adrien Coulet**, whose expertise on knowledge graphs and embeddings will certainly have a stimulating effect on my work in the HeKA team.

4. Conclusion.

I have been planning and refining this project since 2017. Pursuing a career at the intersection of geometric data analysis and medical sciences allows me to reconcile a sincere desire to **contribute to our healthcare system** with my ever-present taste for Riemannian geometry. Crucially, my long-term investment in numerical tools is finally bearing fruit: KeOps and GeomLoss now provide me with a **reliable platform** to turn original geometric ideas into scalable methods.

A right fit for INRIA. Building bridges between modern mathematical research and real-life clinical practice is a generational project, that must involve a wide range of research profiles. To address the countless challenges that are encountered in healthcare, our community will need to open both surprising "mountain trails" and reliable "highways".
In this context, I decided early on in my career to focus on building **innovative software tools for geometric data analysis**. Faced with the limitations of the current development ecosytem in medical data sciences, I decided to take the matter into my own hands – in the best interest of all my colleagues. The ongoing deep learning revolution, that was kickstarted by **a first generation of academic libraries** such as Caffe or Theano before the advent of industry-wide support with TensorFlow or PyTorch has been a source of motivation. Nevertheless, I am well aware of the fact that this time-consuming work is at odds with standard evaluation guidelines in our field. Transparent and well-designed software libraries are **taken for granted** by most users and seldom acknowledged as important scientific contributions. This is especially true for toolboxes such as SciPy or **KeOps**, which are designed to be integrated in the foundations of higher-level packages.
Fortunately, I understand that INRIA intends to provide a space for such research projects within French academia. Some of its most impactful results stem from fruitful exchanges at the junction of applied mathematics and computer science, to be distributed worldwide as highly innovative software. As discussed throughout this application, my libraries have already attracted **significant interest from several INRIA teams**, especially in the Paris and Saclay centers. I believe that this is a consequence of the strategic positioning of my work, which strikes a **delicate balance** between ease of use, performance, mathematical depth and adequacy with modern research directions.

A right fit for the HeKA team. More specifically, my skill set and research interests are a good fit for HeKA. First of all, I bring to the team a strong expertise in computational anatomy, geometric data analysis and software development. These will be key to handle heterogeneous medical data and disseminate our results in the wider scientific community. Going further, my diverse background should allow me to act as a **bridge between different research profiles**, both within the team and within the Paris INRIA center. After years of training in mathematics at the ÉNS (both as a student and as a teacher), a 2-year postdoc in the department of computing at Imperial College and past experience in the industry, I have become used to juggling between different languages and evaluation standards. On the one hand, I am well integrated in the applied maths community and fully abide to its standards of rigor. On the other hand, I am familiar with common empirical practices in the deep learning literature. This is important both for the dissemination of our work in the healthcare industry and for the effective training of **students who come from very diverse backgrounds**. In order to **lower the barrier of entry** to my main research ideas, I have written both my PhD thesis and the documentation of my libraries as key pedagogical resources on e.g. applied measure theory for future students and colleagues.
As a final note, I would like to stress that the **strong ties of HeKA with clinical practice** have been a **decisive factor** in my application. Remarkably, half of the members of the team (both permanent researchers and students) are medical doctors, with valuable access to data and priceless insight on public health through their daily practice. I am all too aware of the gap that exists between idealized machine learning problems and real-life healthcare, of the amount of work that is required to meet the actual needs of practitioners. I am thus convinced that the exceptional situation of the HeKA team will be key in allowing me to turn promising research ideas into **genuine clinical tools**. After five years of planning and exchanges in radiology events, I look forward to finally taking up this challenge.

Formulaire 5 — LISTE COMPLÈTE DES CONTRIBUTIONS[1]
Form 5 — COMPLETE LIST OF CONTRIBUTIONS[1]

1. Publications caractéristiques/*Representative publications*

The publications below provide a balanced overview of my scientific profile: I always try to combine a deep understanding of theoretical concepts (1) with good programming skills (2) in order to bring tangible value to domain experts (3). Note that my PhD thesis contains a significant amoung of unpublished material and is the piece of work that I am the most proud of. However, it is is probably too long and verbose to be included in this short list of publications:

1. Interpolating between optimal transport and MMD using Sinkhorn divergences, Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun'ichi Amari, Alain Trouvé and Gabriel Peyré. Presented at AISTATS 2019 and published in the corresponding proceedings (22nd International Conference on Artificial Intelligence and Statistics).
2. Fast geometric learning with symbolic matrices, Jean Feydy*, Joan Alexis Glaunès*, Benjamin Charlier* and Michael Bronstein, presented as a spotlight talk at NeurIPS 2020 and published in the corresponding proceedings (Advances in Neural Information Processing Systems 33).
3. Fast end-to-end learning on protein surfaces, Freyr Sverrisson*, Jean Feydy*, Bruno E. Correia and Michael M. Bronstein, which is currently available on bioRxiv and under review at CVPR 2021.

2. Publications

Publication strategy. I work at the intersection of medical imaging, machine learning research and applied mathematics: three communities that have very different practices and traditions with respect to the diffusion of scientific ideas.
To publish every step of my work in the most suitable venue, I rely on the following roadmap:
(1) Publish **proof of concept papers** in medical imaging conferences (MICCAI) to get a first feedback from practitioners.
(2) Publish **stable releases** in mainstream machine learning conferences (AISTATS, NeurIPS), to gain visibility and stimulate cross-field interactions. (3) Once all ideas have become mature and tools have stood the test of time with an active user base, publish **final results** in a relevant journal for posterity.

This publication cycle goes hand-in-hand with the time-consuming documentation of my libraries (presented in Section 3). It is still in progress for all my projects but the KeOps package, which means that all but one of my papers so far have been published in machine learning and medical imaging conferences. Authors are thus listed by decreasing order of contribution, with the principal investigator of the project as last author.
A star "*" denotes co-first authorship, with equal contribution.

2.1 Revues internationales/*International journals*

- Kernel operations on the GPU, with autodiff, without memory overflows, Benjamin Charlier*, Jean Feydy*, Joan Alexis Glaunès*, François-David Collin, Ghislain Durif. On January 26, 2021, this short presentation of the KeOps library has been accepted for publication in the **Journal of Machine Learning Research**, by the Open Source Software track (action editor: Alexandre Gramfort).

2.2 Conférence internationales avec comité de lecture/*Reviewed international conferences*

- Fast geometric learning with symbolic matrices, Jean Feydy*, Joan Alexis Glaunès*, Benjamin Charlier* and Michael Bronstein. Presented as a **spotlight talk at NeurIPS 2020** and published in the corresponding proceedings (Advances in Neural Information Processing Systems 33).
- Fast and scalable optimal transport for brain tractograms, Jean Feydy*, Pierre Roussillon*, Alain Trouvé and Pietro Gori. Presented at **MICCAI 2019** and published in the corresponding proceedings (International Conference on Medical Image Computing and Computer-Assisted Intervention, 2019).
- Interpolating between optimal transport and MMD using Sinkhorn divergences, Jean Feydy, Thibault Séjourné, François-Xavier Vialard, Shun'ichi Amari, Alain Trouvé and Gabriel Peyré. Presented at **AISTATS 2019** and published in the corresponding proceedings (22nd International Conference on Artificial Intelligence and Statistics).
- Global divergences between measures: from Hausdorff distance to optimal transport, Jean Feydy and Alain Trouvé. Presented as an oral talk at the ShapeMI workshop of MICCAI 2018 and published in the corresponding proceedings (International Workshop on Shape in Medical Imaging, 2018).

- Optimal transport for diffeomorphic registration, Jean Feydy, Benjamin Charlier, François-Xavier Vialard and Gabriel Peyré. Presented as an **oral talk at MICCAI 2017** and published in the corresponding proceedings (International Conference on Medical Image Computing and Computer-Assisted Intervention, 2017).

2.3 Livres et chapitres de livre/*Books and book chapters*

2.4 Autres publications internationales (posters, articles courts)/*Other international publications (posters, short papers)*

- Distortion minimizing geodesic subspaces in shape spaces and computational anatomy, Benjamin Charlier, Jean Feydy, David W. Jacobs and Alain Trouvé. Presented at VipIMAGE 2017 and published in the corresponding proceedings (Lecture Notes in Computational Vision and Biomechanics, volume 27).

2.5 Revues nationales/*National journals*

2.6 Conférence nationales avec comité de lecture/*Reviewed national conferences*

2.7 Rapports de recherche et articles soumis/*Research reports and publications under review*

- Fast end-to-end learning on protein surfaces, Freyr Sverrisson*, Jean Feydy*, Bruno E. Correia and Michael M. Bronstein. This pre-print is currently available on bioRxiv and under review at CVPR 2021.

3. Développements technologiques : logiciel ou autre réalisation / *Technology development : software or other realization*

- **KeOps:** `Family=research; Audience=community; evolution=lts; Duration=4; contribution=leader, devel, softcont; Url=https://www.kernel-operations.io/`

  This library is presented in Form 3 as my first major contribution: **fast geometric learning with symbolic matrices**. It is primarily coded in C++/CUDA (20K lines) and Python (15K lines), with additional bindings for R (3K lines) and MATLAB (1K lines). Its main purpose is to enable the use of efficient numerical schemes in the wider scientific community. Notably, we focus on the simple but powerful abstraction of **symbolic matrices** as a third major class of numerical tensors that complements the usual **dense** arrays and **sparse** matrices. This allows KeOps to support cutting edge algorithms on the GPU while being remarkably easy to use by mathematicians and data scientists.

  Benjamin Charlier, Joan Glaunès and myself have been working on this project since 2017, with a first stable release in April 2019. Following our publications in 2020, it is quickly gaining traction in the machine learning literature. Our priority is now to foster the development of a genuine community around this toolbox: we focus on building bridges with other fields in numerical analysis and open the door for new contributors on major features.

- **GeomLoss:** `Family=research; Audience=community; evolution=lts; Duration=2; contribution = leader, devel, softcont; Url=https://www.kernel-operations.io/geomloss`

  This library is presented in Form 3 as part of my second major contribution: **fast, scalable and robust optimal transport solvers**. It is coded in Python (5K lines), with all the necessary C++/CUDA code being factored out in the KeOps library. I have been working on this project since 2019, with a break in 2020 that was devoted to getting familiar with geometric deep learning applications in Michael Bronstein's team. I am now actively working on updating this library with generic solvers for both grid images and unstructured sampled data, Wasserstein barycenters, an automated test suite and a comprehensive documentation. I target a first stable release by mid-2021.

  As detailed in my research project, the main purpose of this library is to **bring efficient optimal transport methods to a general audience**. I intend to work on the integration of GeomLoss in the foundations of standard libraries for data sciences (e.g. POT, UMAP or PyTorch_Geometric) as well as advanced research codes for computational physics, the simulation of partial differential equations and computational anatomy.

- **dMaSIF:** `Family=vehicle; Audience=partners; evolution=basic; Duration=1; contribution=leader; Url=https://github.com/FreyrS/dMaSIF` (This public release does not keep track of our commit history.)

  This GitHub repository supports my work on protein surfaces, presented in Form 3 as my third major contribution. It is coded in Python (4K lines), with all the necessary C++/CUDA code being factored out in the KeOps library. As detailed in Form 3, **I designed and implemented the core geometric layers of this deep learning model**. These were then put together by Freyr Sverrisson, who designed a full data processing pipeline and was in charge of our numerical experiments. We are currently working on improving this method with geometric steps that model both rigid and flexible docking, as we target applications to protein design.

Going forward, this code will replace the previous MaSIF architecture as a fast and versatile model for deep learning on protein structures. The long-term future of this project will be handled by the Laboratory of Protein Design and Immunoengineering at EPFL: if this line of work delivers on its current promises, our pre-trained model will eventually be packaged as a web service (following common practice in bioinformatics).

4. Impact socio-économique et transfert / *Socio-economic impact and transfer*

I wrote my Master's thesis during a 5-month internship at Siemens Healthcare in Princeton, NJ. As described in Form 1, Section 7 (Mobility), this industry experience had a **deep and lasting impact on my research**. It confirmed my desire to work in the medical field, motivated my extensive numerical work and opened me the doors of the booming Parisian tech scene in AI for healthcare (with visits at Canon Medical Systems, Therapixel, Gleamer, Owkin, Incepto, etc.).

Apart from KeOps (which is already attracting significant interest through our support for GPyTorch), my work is still several years away from being mature enough for widespread adoption in the industry. Looking back on my progresses since 2015, however, I believe that I am on a track that leads to **fruitful interactions with healthcare companies**. This is of interest to me both as a scientist (as a source of stimulating questions) and as an end-user of medical devices (hopefully, not too soon). My long-term goal is to enable the use of geometric methods as standard tools in medical sciences, with a significant impact on clinical practice. In this context, the active support that INRIA provides for transfer works is a major motivation for my application to this job offer.