

Audition INRIA Paris 2021 – équipe HeKA

Jean Feydy
Imperial College London

26 mai 2021

CV : un parcours “maths-info” dirigé vers la médecine

Formation en **mathématiques** et **sciences des données** :

2012–2016 ENS Paris, mathématiques.

2014–2015 M2 mathématiques, vision, apprentissage à l’ENS Cachan.

2016–2019 Thèse en **imagerie médicale** avec Alain Trouvé de l’ENS Cachan.

Enseignement en **maths appliquées** à l’ENS Paris.

2019+ **Deep learning géométrique** avec Michael Bronstein de l’Imperial College.

Liens étroits avec la **médecine** :

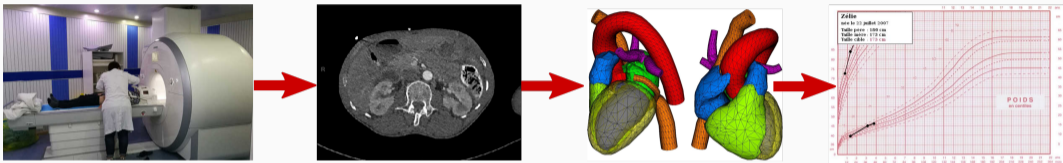
2015 Débruitage d’images chez **Siemens Healthcare** à Princeton.

2018+ 5 présentations invitées à des **congrès de radiologie**.

2019+ MasterClass IA–Imagerie, pour les **internes de radio** à l’Université de Paris.

2020+ Colloque sur l’**Imagerie médicale à l’heure de l’IA**, à l’Institut du Cerveau.

Ma motivation : l'analyse de données médicales



Trois caractéristiques :

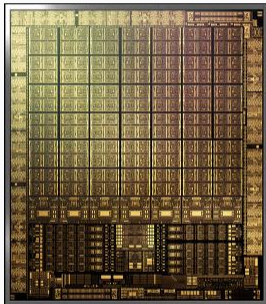
- Données **hétérogènes** : parcours de soin, images, etc.
- Petits échantillons stratifiés : 10 - 1 000 patients par groupe.
- La gestions des **outliers** et de la **queue lourde** des distributions est une priorité.

Enjeu. Développer des modèles qui combinent **expertise médicale** et jeux de **données**.

Défi. Arrivée des **cartes graphiques (GPU)** :

- Excellent rapport **qualité-prix** :
 $1\ 000\text{€} \simeq 1\ 000\ \text{cœurs} \simeq 10^{12}\ \text{opérations/s}$.
- **Difficulté** : contraintes sur la gestion des **registres**.

Écosystème Python “grand public” consolidé autour d’un **petit nombre d’opérations élémentaires**.



7,000 cœurs en batterie sur un GPU.

Solution. Étendre la boîte à outils standard en sciences des données pour répondre aux défis posés par le monde médical.

Faciliter le développement de **modèles avancés**, sans compromis sur les performances numériques.

Activités antérieures, **fondations** numériques → applications de **haut niveau** :

1. Manipulation efficace de **matrices “symboliques”** (distances, noyaux, etc.).
2. **Transport optimal** : méthodes de tri généralisées (focus technique).
3. **Deep learning** géométrique et applications **biomédicales**.

Projet de recherche et **intégration** dans l'équipe HeKA.

1. Matrices symboliques

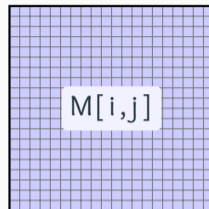
Les outils de calcul scientifique manipulent (surtout) des matrices

Contexte. Accès mémoires contraignants sur GPU :

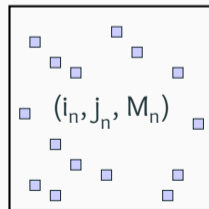
- Temps de transfert vers les registres pénalisent l'utilisation de matrices **denses**.
- Prime aux accès mémoires **contigus** pénalise l'utilisation de matrices **creuses**.

Défi. Pour atteindre des performances optimales :

- Se **restreindre** aux opérations supportées directement par le constructeur : convolutions, FFT, etc.
- Développer soi-même de nouvelles routines numériques en C++/CUDA (**plusieurs mois de travail**).



Matrice dense



Matrice creuse

Ma contribution : support efficace des matrices symboliques

Solution. KeOps – www.kernel-operations.io :

- Pour PyTorch, NumPy, Matlab et R sur **CPU et GPU**.
- **Différentiation automatique**.
- **Compilation** à la volée de **schémas optimisés** en C++, exécutés pour chaque **réduction** (somme, min, etc.).

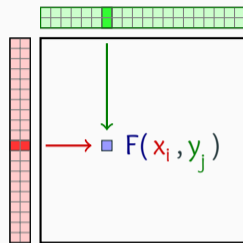
Si la formule “F” est simple (≤ 100 opérations arithmétiques) :

calcul “100k \times 100k” \rightarrow 10ms – 100ms,

calcul “1M \times 1M” \rightarrow 1s – 10s.

Limite matérielle de 10^{12} opérations/s.

Exécution **10 à 100 fois plus rapide** / code GPU standard pour une large gamme de problèmes.



Matrice symbolique

Formules + données

- Distances $d(x_i, y_j)$.
- Noyaux $k(x_i, y_j)$.
- Nombreuses transformées.

2. Transport optimal

Le transport optimal (OT) généralise le tri aux espaces de dimension $D > 1$

Contexte. Si $A = (x_1, \dots, x_N)$ et $B = (y_1, \dots, y_N)$ sont deux nuages de N points dans \mathbb{R}^D , on définit :

$$\text{OT}(A, B) = \min_{\sigma \in \mathcal{S}_N} \frac{1}{2N} \sum_{i=1}^N \|x_i - y_{\sigma(i)}\|^2$$

Généralise le **tri** aux espaces métriques.

Problème linéaire sur la matrice de permutation P :

$$\text{OT}(A, B) = \min_{P \in \mathbb{R}^{N \times N}} \frac{1}{2N} \sum_{i,j=1}^N P_{i,j} \cdot \|x_i - y_j\|^2,$$

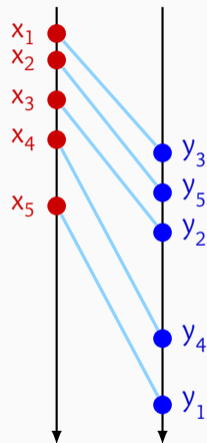
$$\text{s.t. } P_{i,j} \geq 0$$

$$\underbrace{\sum_j P_{i,j}} = 1$$

Chaque point au départ...

$$\underbrace{\sum_i P_{i,j}} = 1.$$

est transporté jusqu'à l'arrivée.



assignement

$$\sigma : [1, 5] \rightarrow [1, 5]$$

Régularisation et biais entropique

Défi. Assignment linéaire : difficile à résoudre en toute généralité.

Structure de la matrice des distances $\|x_i - y_j\|$

\implies **Accélérer** les calculs.

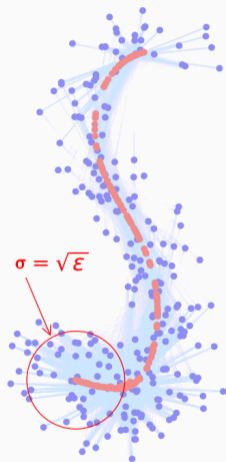
Outil fondamental : le **transport régularisé**

$OT_\varepsilon(A, B) \simeq OT(A, B) +$ pénalité entropique d'intensité $\varepsilon > 0$.

Approximation **lisse** et strictement **convexe** : plus facile à étudier,
très populaire algorithme de **Sinkhorn** (ou "SoftAssign").

En revanche, elle ne **vérifie pas les axiomes d'une distance** :

$$OT_\varepsilon(B, B) > 0.$$



$$\arg \min_A OT_\varepsilon(A, B).$$

Solution théorique : garantie de robustesse au biais entropique

Solution. Les divergences de Sinkhorn sont définies par :

$$S_\varepsilon(A, B) = \text{OT}_\varepsilon(A, B) - \frac{1}{2}\text{OT}_\varepsilon(A, A) - \frac{1}{2}\text{OT}_\varepsilon(B, B)$$

afin de retrouver une valeur nulle lorsque $A = B$.

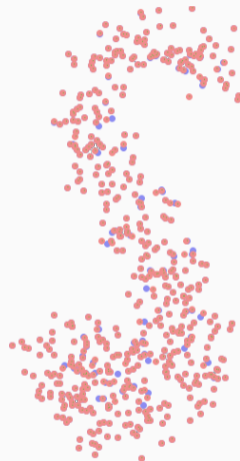
Théorème (S_ε se prête bien à l'optimisation)

Pour tous échantillons A et B :

$$S_\varepsilon(A, B) \geq 0 \text{ avec égalité ssi. } A = B,$$

$A \mapsto S_\varepsilon(A, B)$ est convexe au sens des mesures,
différentiable et métrise la convergence en loi.

On généralise ce résultat aux **mesures** de Radon positives,
à des **métriques** $\|x_i - y_j\|$ générales et au cas “**unbalanced**”.



$$\arg \min_A S_\varepsilon(A, B).$$

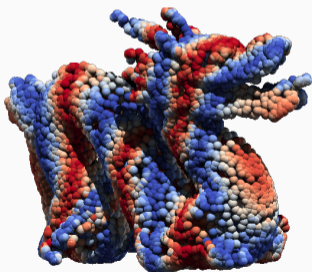
Solution pratique : l'algorithme de Sinkhorn multi-échelle

Solution. J'ai amélioré l'état de l'art pour le transport optimal discret par **un à trois ordres de grandeur** – selon le régime d'utilisation :

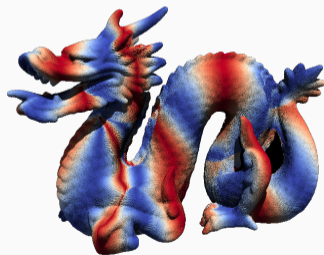
Sinkhorn GPU $\xrightarrow{\times 10}$ + KeOps $\xrightarrow{\times 10}$ + Recuit simulé $\xrightarrow{\times 10}$ + Multi-échelle

Résoudre un problème de transport avec une précision de 1% :

`pip install
geomloss
+
GPU récent
(1 000 €)`



10k points en 30-50ms



100k points en 100-200ms

3. Deep learning géométrique

Concevoir des modèles entraînaibles et adaptés aux experts

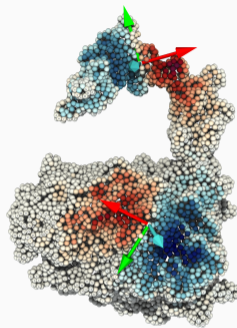
Enjeu. Modèles profonds sur des **domaines non-euclidiens** (nuages de points, surfaces, graphes, etc.), au-delà des images 2D/3D.

Défi. En dépit d'un intérêt industriel croissant, ces modèles restent **mal supportés** numériquement. Le C++/CUDA reste (souvent) incontournable.

Solution. En quelques lignes de Python, avec KeOps :

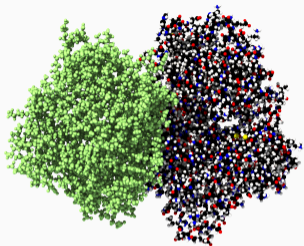
- Interactions **locales** : K-plus-proches voisins.
- Interactions **globales** : convolutions généralisées.

Liberté de modélisation
⇒ **Spécificités** des domaines étudiés.



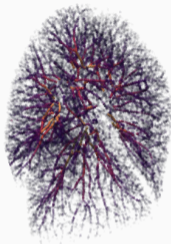
Convolution
quasi-géodésique
sur un nuage de
points orientés.

Quelques applications biomédicales depuis mon arrivée à Londres



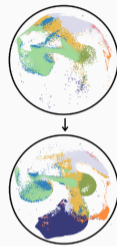
Docking de protéines :
deep learning géométrique
avec KeOps.

Michael Bronstein
et Bruno Correia (EPFL).



Recalage de poumons :
deep learning géométrique,
transport optimal et
recalage difféomorphique.

Marc Niethammer
(UNC Chapel Hill).



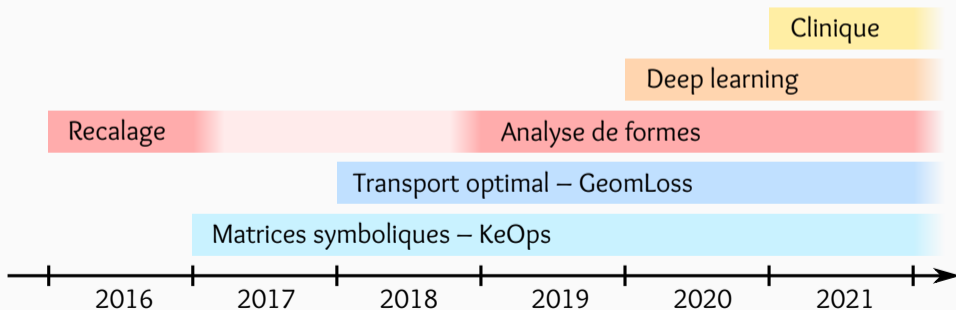
**Différenciation de cellules
souches du sang :**
transport optimal,
modélisation.

Dominique Bonnet
(Francis Crick Institute).

Bilan : un investissement de long terme qui porte ses fruits

Deux évolutions majeures :

- “Gros” problème géométrique : $N > 10k \rightarrow N > 1M$.
- Transport optimal : **problème** linéaire + **tri rapide** généralisé.



Projet de recherche

Contexte. L'équipe **HeKA** (commune Inria – Inserm – Université de Paris) :

- Affiliée à l'UFR de Médecine et au Centre Inria de Paris.
- Spectre de compétences étendu, **7 permanents hospitalo-universitaires**.

Défi. HeKA est centrée sur la **représentation des patients** et de leurs trajectoires : santé publique, médecine de précision, essais cliniques.

Mon projet. Analyse géométrique de données de **santé** :

1. **Passage à l'échelle** de modèles statistiques complexes.
2. Méthodes **géométriques** pour l'analyse de données structurées.
3. Nouvelle génération de modèles robustes pour l'**anatomie** computationnelle.

1. Passage à l'échelle de modèles statistiques complexes

Contexte. En pharmacovigilance l'estimation d'effets indésirables est un problème de santé publique.

Défi. Modèles standards : bibliothèques R conçues pour des **essais cliniques** à petite ou moyenne échelle, \neq “**dépistage massif**” de la population française.

Limitations numériques

⇒ Dépendance aux signaux d'alerte en médecine de ville.

Clinical Therapeutics/Volume 43, Number 2, 2021

Commentary

Artificial Intelligence in Pharmacovigilance: Scoping Points to Consider

Manfred Hauben, MD, MPH^{1,2};
and Craig G. Hartford, MBCh, PhD³

¹Safety Sciences Research, Pfizer Inc;

²Department of Medicine, NYU Langone Health;

³Pharmaceutical Physician, Kent, United Kingdom

ABSTRACT

Artificial intelligence (AI), a highly interdisciplinary science, is an increasing presence in pharmacovigilance (PV). A better understanding of the scope of artificial intelligence in pharmacovigilance (AIPV) may be advantageous to more sharply defining, for example, which terms, methods, tasks, and data sets are suitably subsumed under the application of AIPV. Accordingly, this article explores relevant points to consider regarding defining the scope of AIPV and offers a potential working definition of the scope of AIPV. (*Clin Ther.* 2021;43:372–379) © 2020 Elsevier Inc.

1. Passage à l'échelle de modèles statistiques complexes

Solution. KeOps démocratise l'analyse de jeux de données d'envergure nationale.

Avec **HeKA** – modèle de survie *WCE* (*Weighted Cumulative Exposure* = Cox PH + splines)
données de l'**assurance maladie** (660k parcours de soin) :

R-WCE avec 32 cœurs CPU $\xrightarrow{\times 10 - \times 100}$ R-WCE + R-KeOps avec 1 GPU.

\implies **Veille continue** à l'échelle nationale.

Verrou à débloquer : **modèles** plus fins, **structure** des espaces de données.

Pour *WCE* : classification hiérarchique ATC des **médicaments**.

2. Méthodes géométriques pour l'analyse de données structurées

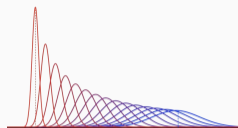
Solution. KeOps + GeomLoss

- ⇒ Passage à l'échelle de géométries arbitraires
- ⇒ Méthodes de **haut niveau**, au-delà des ACP tangentes : modèles Gaussiens, analyse topologique, *deep learning*...

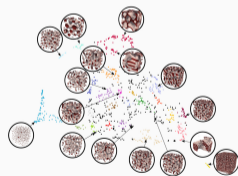
Focus sur la métrique de **Wasserstein** (transport optimal) :
mesures et **histogrammes**.

Verrous à débloquer :

- Algorithmique : stratégies multi-échelles **non-euclidiennes**.
- Mathématique : métriques moins **homogènes**, mais proches des **données** et des **experts**.
- Pratique : **diffusion** auprès du monde médical.



Géodésique de
Wasserstein.



Analyse UMAP
de $N = 1\,000$
diagrammes.

3. Intégration de données anatomiques

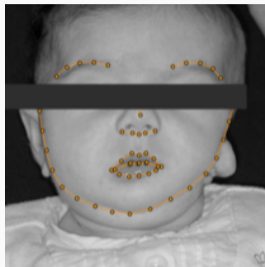
Contexte. Expertise en anatomie computationnelle
⇒ Nouvelle compétence essentielle pour HeKA.

Défi. Problèmes **cliniques** issus des Hôpitaux de Paris.

Solution. Nouvelle génération de modèles en morphométrie,
combinant pour la première fois :

- Extraction **automatique** de descripteurs.
- Mise en correspondance à l'aide du **transport optimal**.
- Analyse **riemannienne** des déformations.

Verrou. Imagerie médicale \cap Vision par ordinateur \cap Graphics.



L'un des 40 000 jeunes patients documentés depuis 1995 par le service de chirurgie maxillo-faciale de l'Hôpital Necker.

Un projet à l'interface pour contribuer à moderniser notre système de santé

J'apporte à HeKA : expertise en **géométrie** \longrightarrow données **hétérogènes**.

En retour : réflexion **mathématique** \longleftarrow problèmes **cliniques**.

Logiciel \implies dynamique **au-delà** du monde médical.

Court terme :

- **WCE-GPU**
(Projet Epi-Phare).
- **Dismorphies** infantiles
(Hôpital Necker).
- **Spondylarthrites**
ankylosantes
(Opis, Hôpital Cochin).

Moyen terme :

- **Transport optimal**
(MoKaPlan, Parietal).
- **scikit-shapes** prend la
relève de Deformetrica
(Paris + Saclay, JHU,
Imperial, UNC).

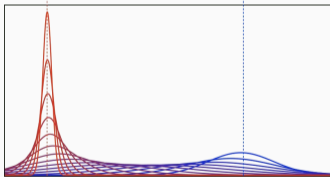
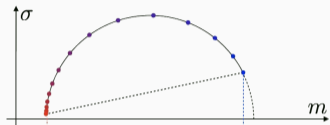
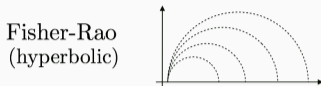
Long terme :

- **Chirurgie pariétale**
(Hôpital Louis Mourier).
- **Statistiques**
géométriques
(Epione, Sierra).
- Transfert **industriel**,
au-delà de KeOps.

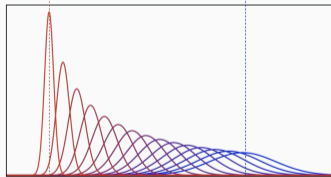
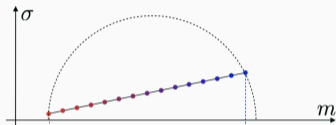
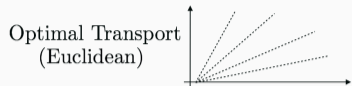
Quelques éléments de mon dossier

- **Prix de thèse** de l'association française pour la reconnaissance et l'interprétation des formes.
- KeOps : **Journal** of Machine Learning Research.
- 2 × MICCAI (dont un oral) + ShapeMI (oral), NeurIPS (spotlight), CVPR, AISTATS.
- **50k téléchargements** pour KeOps (matrices symboliques), 25k pour GeomLoss.
- Intégration dans des **logiciels de référence**, développés à INRIA et à l'étranger :
GPyTorch (Cornell, Facebook...), Deformetrica ([Aramis](#)), Falkon (Université de Gênes + [Sierra](#)),
Gudhi ([DataShape](#)), Python Optimal Transport (CMAP + [Parietal](#)).
- **Enjeu** : **débloquer** l'utilisation de méthodes **non-euclidiennes** dans le monde médical.

Entropie relative ou transport optimal? Le cas des lois normales.

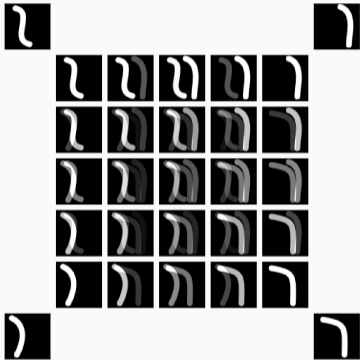


Géodésique pour la métrique induite
par l'entropie relative.

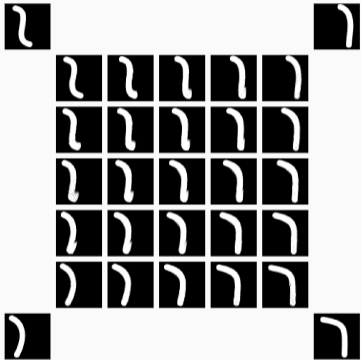


Géodésique pour la métrique induite
par le **transport optimal**.

Le transport optimal induit une géométrie “particulière”

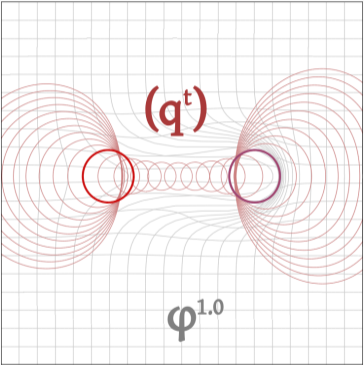


Barycentres euclidiens.
("eulériens")

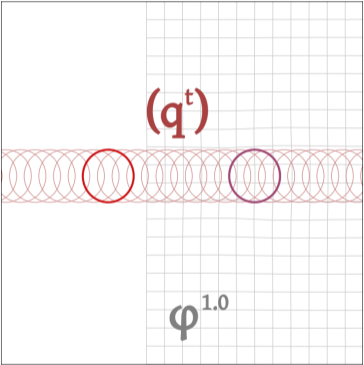


Barycentres de Wasserstein.
("lagrangiens")

Limite des modèles de déformation implémentés par Deformetrica

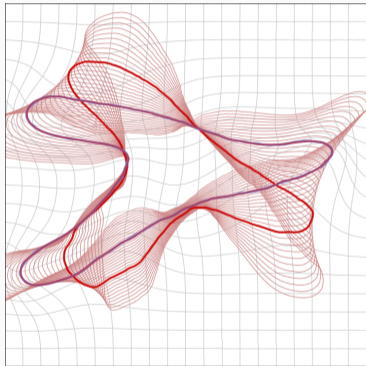


LDDMM.

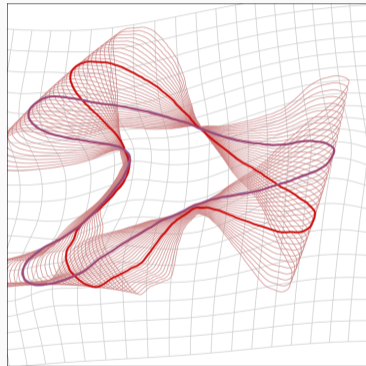


LDDMM + normalisation.

Limite des modèles de déformation implémentés par Deformetrica

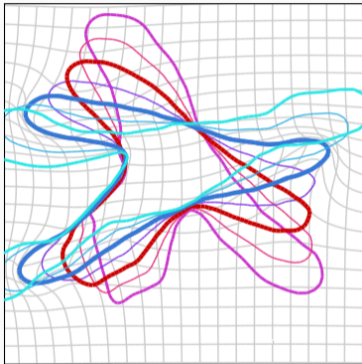


LDDMM.

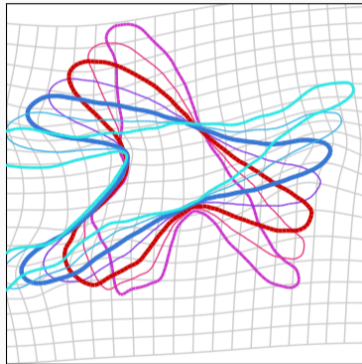


LDDMM + normalisation.

Limite des modèles de déformation implémentés par Deformetrica

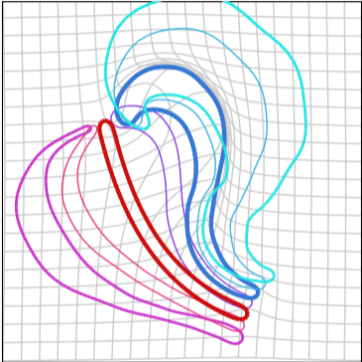


LDDMM.

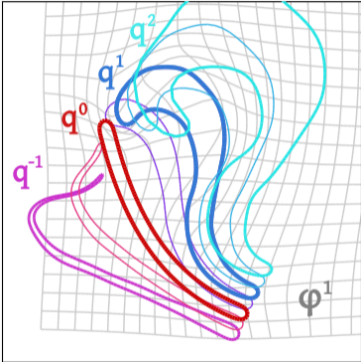


LDDMM + normalisation.

Limite des modèles de déformation implémentés par Deformetrica



LDDMM.



LDDMM + normalisation.