Biomolecules Random Walks, Heterogeneities and Model Selection:

What Information is accessible from experimental Biomolecules Random Walk?

Jean-Baptiste Masson
Physics of Biological Systems, Institut Pasteur, CNRS UMR 3525, Paris, Framce
Janelia Research Campus, Ashburn,
Laboratoire Physico-Chimie, Institut Curie, CNRS UMR168, Paris France

Introduction

- Order of Magnitudes
- Analysing Random Walks
- Pipeline
- Not Tracking
- Mapping
- Application

Biological Media

- Localization precision ~ 10nm
- Temporal precision ~ 10ms
- Static Density of tagging ~ 1-10000 #.μm^{-2}
- Static Density of tagging ~ 1-100 #.μm^{-3}
- Dynamic Density of Tagging ~ 0.01-10 #.μm^{-2}
- Dynamic Density of Tagging ~ 0.01-2 #.μm^{-3}
- Duration of recording ~ 1-1000s
- Field of View ~ 10-2500 μm^2

Time Scales:

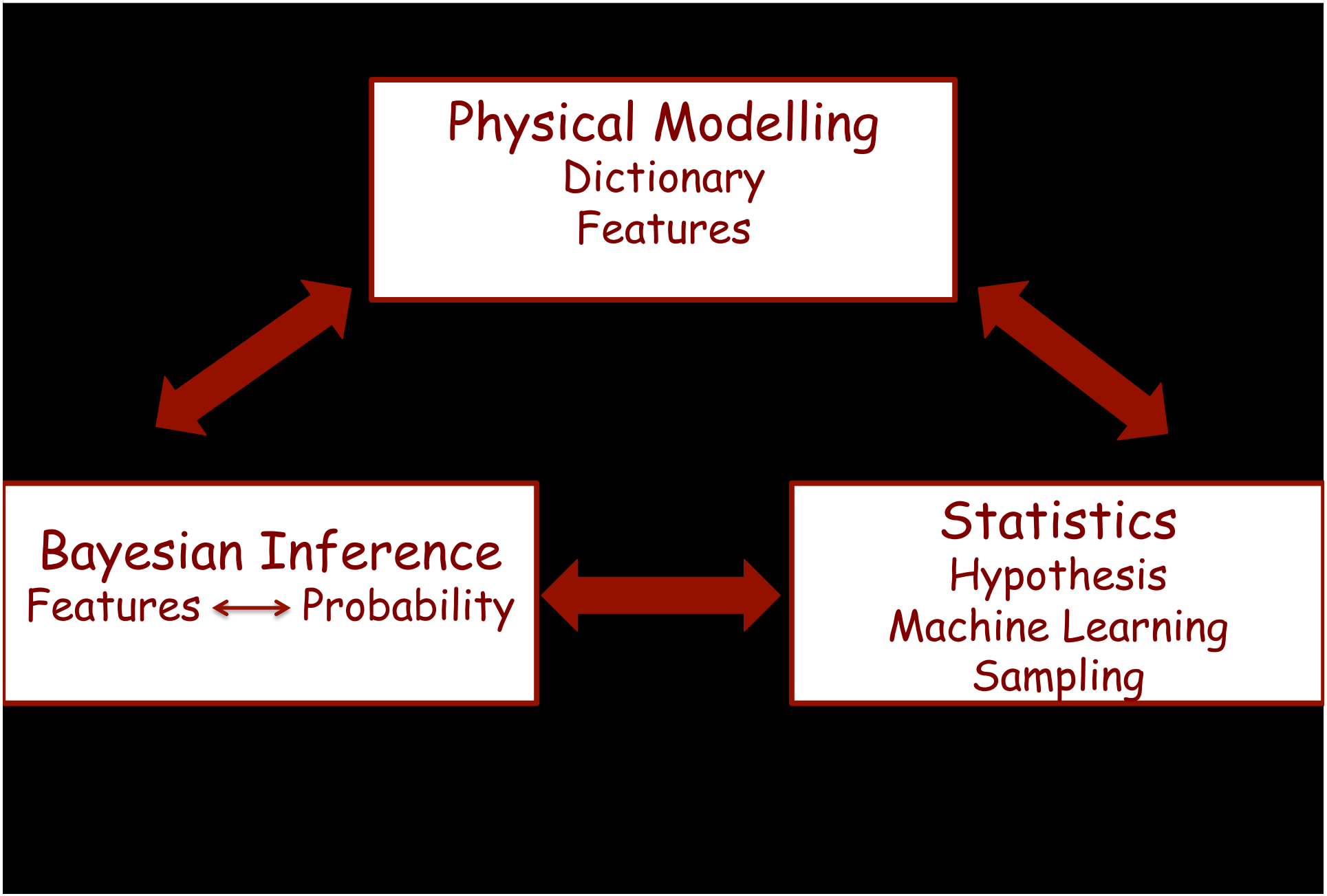
- Transcription ~ 1 - 7.10^5 s
- Traduction ~ 1 - 7.10^5 s
- Methylation ~ 0.1 - 100 s
- Phosphorilation ~ 0.01 - 10 s
- Scaffold ~ 10 - 10^5 s
- Lipid Exchange ~ 0.1 – 500s
- Gephyrin ~ 300 s
- Un/binding ~ 0.1-100 s

- Diffusion ~ 0.005-5 $\mu m^2.s^{-1}$
- Drifts ~ 0.001-10 $\mu m.s^{-1}$
- Forces ~ 0.001-10pN
- Interaction Energy ~ 0.1-10 k_BT
- Local Gradients
  - ∇D ~ 10^{-3}-1 $\mu m.s^{-1}$
  - ∇F ~ 1-500 $k_BT.\mu m^{-2}$
  - ∇V ~ 1-100 $k_BT.\mu m^{-1}$
  - ∇I ~ 0.001-10

Spatial Scales:

- DNA ~ 0.3nm
- Actin-Myosin ~ 10nm
- Lipid Rafts ~ 20nm
- Caveolin ~ 50-70nm
- Lipid Platforms ~ 100-1000nm
- Scaffolds ~ 50-500nm
- PIP2/PIP3 ~ 200-3000nm
- Cytoskeleton Mesh ~ 50-1000nm
- Cytosol subdomain ~ 20-2000nm

Analysing Random Walks

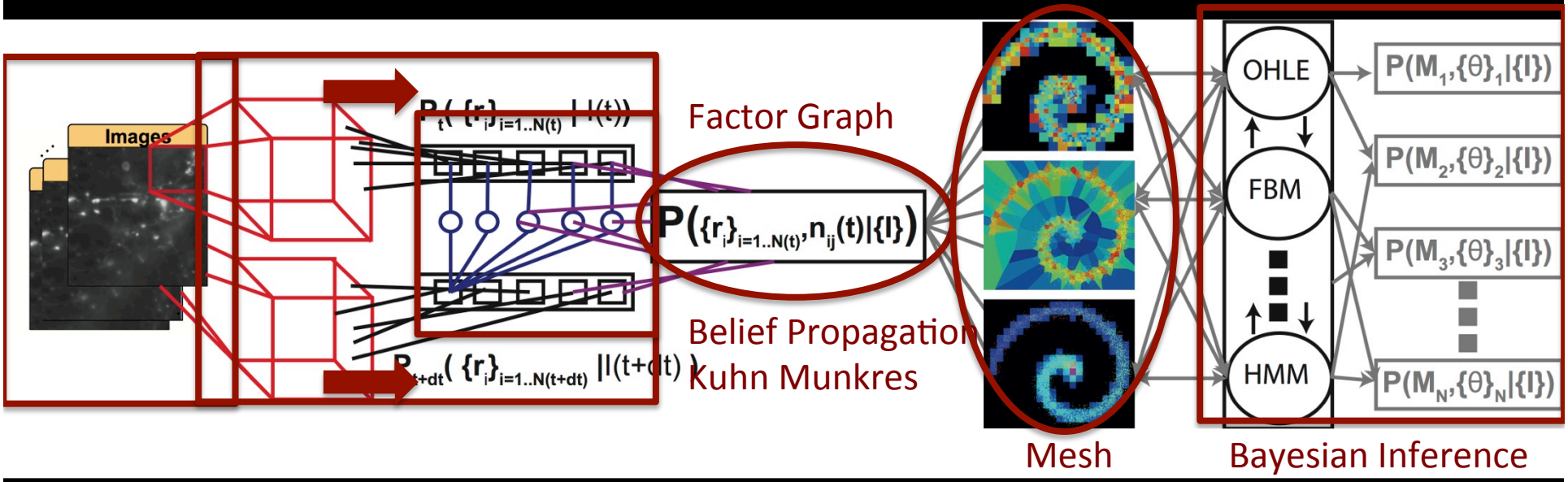


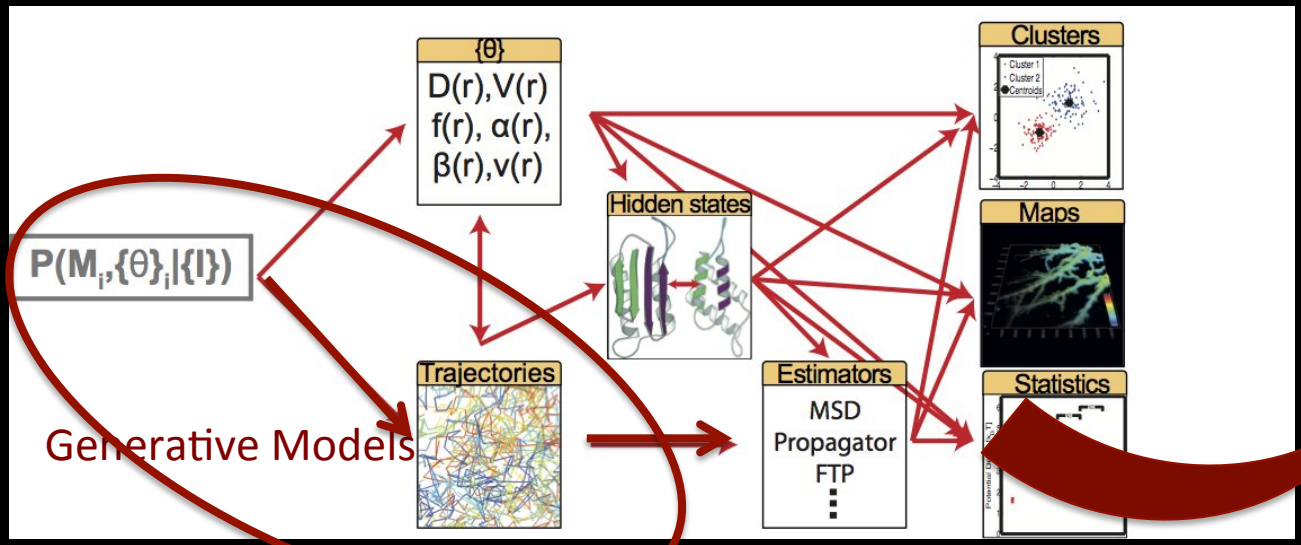
Statistical Hypothesis

- Global distribution analysis (forced i.i.d.)
  - Heterogeneous
  - Statistical testing of pooling procedure, sub families?
- Single trajectory associated properties
  - Heterogeneous
  - Identical Environments?
- Clustered analysis
  - Spatial, internal states etc
  - Pooling neighbouring areas? Collapsing states?
- Time Evolution
  - Complex Windowing? Information?
  - Out of Equilibrium?

D. Koller and N. Friedman, Probabilistic Graphical Models: Principles and technics
Y. Abu-Mostafa & M. Magdon-ismail, Learning from Data
C. Bishop, Pattern Recognition and Machine Learning
G. Saporta, Probabilites, Analyses de données et statistiques
T. Hastie & R. Tibshirani , The Elements of Statistical Learning: Data Mining, Inference, and Prediction

Pipeline

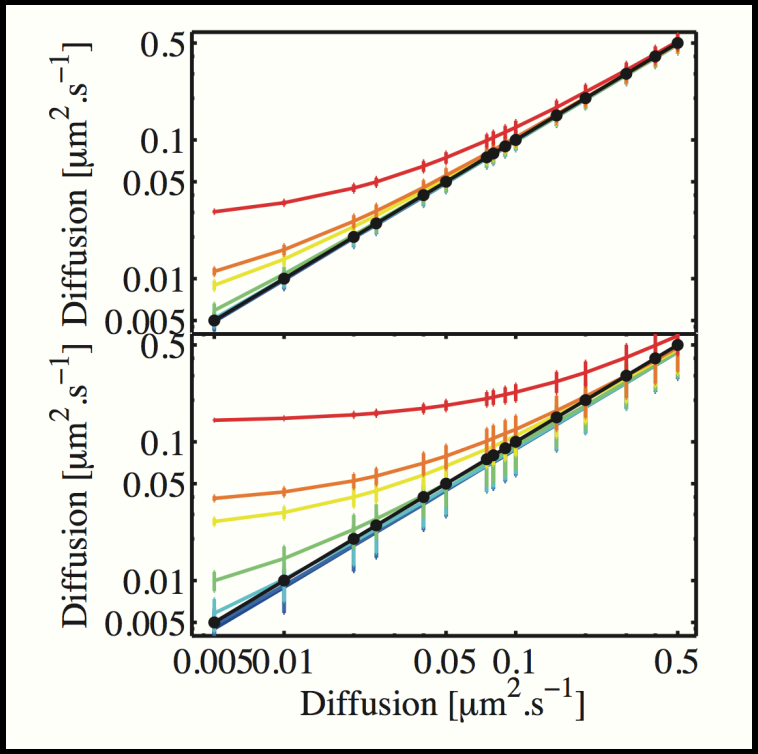

Images
Factor Graph
Belief Propagation
Kuhn Munkres
Mesh
OHLE
FBM
HMM
Bayesian Inference


{θ}
D(r),V(r)
f(r), α(r),
β(r),v(r)
Hidden states
Clusters
Maps
Trajectories
Estimators
MSD
Propagator
FTP
Statistics
Generative Models

Estimators and Acquisition Method

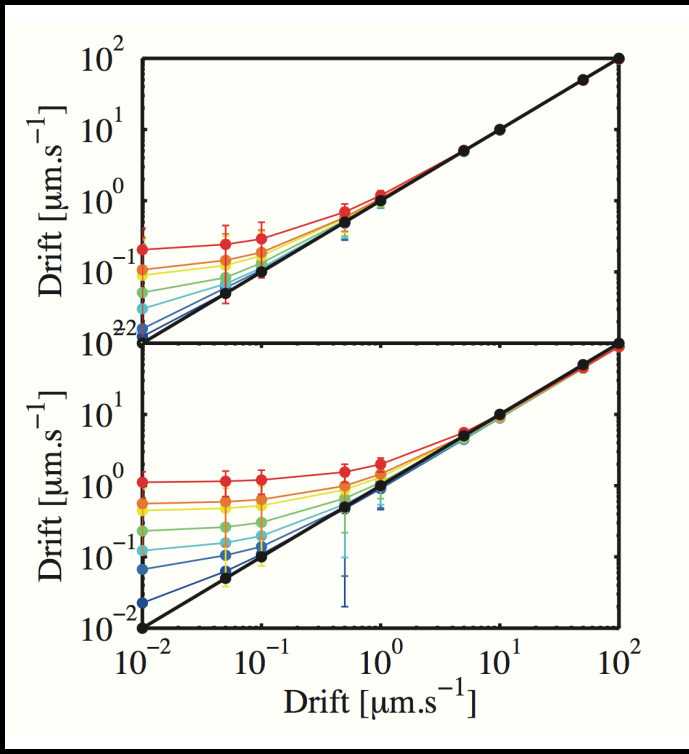
- Estimators have different level of robustness
- Acquisition methods matters
- Tracking error

-500nm
-10nm



50pts

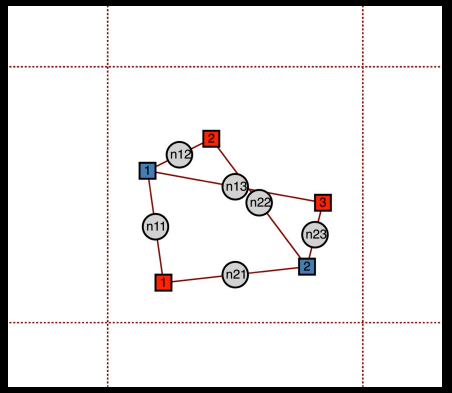
10pts



- If Memoryless trajectories, High density trajectories

No Tracking-Graph Matching: 2 Choices

- High Density ($\rho<5\#.\mu m^{-2}$, $\rho<2\#.\mu m^{-3}$) or highly variable number of Biomolecules
  - Best Assignment captures most of the dynamics
  - 2 steps
    - Min-Sum algorithm on distance based model
    - Bayesian Inference of the Parameters



- Very High Density ($\rho>5\#.\mu m^{-2}$, $\rho>2\#.\mu m^{-3}$)
  - Tracking should not be performed
  - 1 step
    - Solve simultaneously the assignment and the inference Problem: Belief Propagation

S. Abrahamsson *et al*, *Nature Methods* 10, 60–63 (2013)

High Density Step 1: Min-Sum

- Distance based model

$$P(\{n\}|\theta) = \prod_{i=1}^{N_A}\left[\delta\left(\sum_{j=1}^{N_B} n_{ij},1\right)+e^{-\mu}\delta\left(\sum_{j=1}^{N_B} n_{ij},0\right)\right]\times\prod_{j=1}^{N_B}\left[\delta\left(\sum_{i=1}^{N_A} n_{ij},1\right)+e^{-\mu}\delta\left(\sum_{i=1}^{N_A} n_{ij},0\right)\right]\times\prod_{(i,j)} e^{-\beta n_{i,j} d_{ij}^2}$$

- β->∞, only the best Assignment remains
- Iteratively find the best Assignment using Min-Sum Scheme

$$\begin{cases} x_{k\to i}^{j} = \min\limits_{l\in I(j)\setminus i}\left(-\log\left(d_{l\to k}^{j\to j+1}\right)-x_{l\to k}^{j+1}\right) \\ x_{i\to k}^{j+1} = \min\limits_{l\in I(j+1)\setminus k}\left(-\log\left(d_{i\to l}^{j\to j+1}\right)-x_{l\to i}^{j}\right) \end{cases} \qquad \begin{cases} \pi^{j}(i) = \operatorname{argmin}\limits_{l}\left(-x_{l\to k}^{j+1}-\log\left(d_{i\to l}^{j\to j+1}\right)\right) \\ \pi^{j+1}(k) = \operatorname{argmin}\limits_{i}\left(-x_{i\to k}^{j}-\log\left(d_{i\to k}^{j\to j+1}\right)\right) \end{cases}$$

- With $\pi^{j}\left(\pi^{j+1}(i)\right) = i$

M. Mezard & A. Montanari, Information, Physics , and computation
Yedidia *et al,* IEEE TRANS. on inf. theo., vol. 51, no. 7, (2005)

Very High density #10.µm^{-2}

$$P\left(\{n\}|\theta\right)=\prod_{i=1}^{N_A}\left[\delta\left(\sum_{j=1}^{N_B}n_{ij},1\right)+e^{-\mu}\delta\left(\sum_{j=1}^{N_B}n_{ij},0\right)\right]\times\prod_{j=1}^{N_B}\left[\delta\left(\sum_{i=1}^{N_A}n_{ij},1\right)+e^{-\mu}\delta\left(\sum_{i=1}^{N_A}n_{ij},0\right)\right]\times\prod_{(i,j)}e^{-\beta n_{i,j}\left(-4\pi(D+\frac{\sigma^2}{\delta t})\delta t-\frac{d_{ij}^2}{4(D+\frac{\sigma^2}{\delta t})\delta t}\right)}$$

$$\left\{\left(D_l\right)\right\}=Arg\max_{(D_l)}\sum_{\{n\}}P\left(\{n\}\left|\left(D_l\right)\right.\right)=Arg\max_{(D_l)}Z\left(D_l\right)=Arg\max_{(D_l)}\log\left(Z\left(D_l\right)\right)=Arg\min_{(D_l)}\left(F\right)$$

- Approximating the Free Energy: Bethe Free Energy

$$b\left(\{n\}\right)\propto\frac{\prod_i b_i\left(n_i\right)\prod_j b_j\left(n_j\right)}{\prod_{(i,j)}b_{ij}\left(n_{ij}\right)}$$

$$x_{j\to i}^{L}=-\frac{1}{\beta}\log\left(\sum_{k\in A\setminus i}e^{\log\left(P_{kj}\right)+\beta x_{k\to j}^{R}}+e^{-\mu}\right)$$

$$x_{i\to j}^{R}=-\frac{1}{\beta}\log\left(\sum_{k\in B\setminus j}e^{\log\left(P_{ik}\right)+\beta x_{k\to i}^{L}}+e^{-\mu}\right)$$

$$F\left(\beta\right)=-\sum_{(i,j)}\log\left(1+e^{\log\left(P_{kj}\right)+\beta x_{i\to j}^{R}+\beta x_{j\to i}^{L}}\right)+\sum_{i\in A}\log\left(e^{-\mu}+\sum_{j\in B}e^{\log\left(P_{kj}\right)+\beta x_{j\to i}^{L}}\right)+\sum_{j\in B}\log\left(e^{-\mu}+\sum_{i\in A}e^{\log\left(P_{kj}\right)+\beta x_{i\to j}^{R}}\right)$$

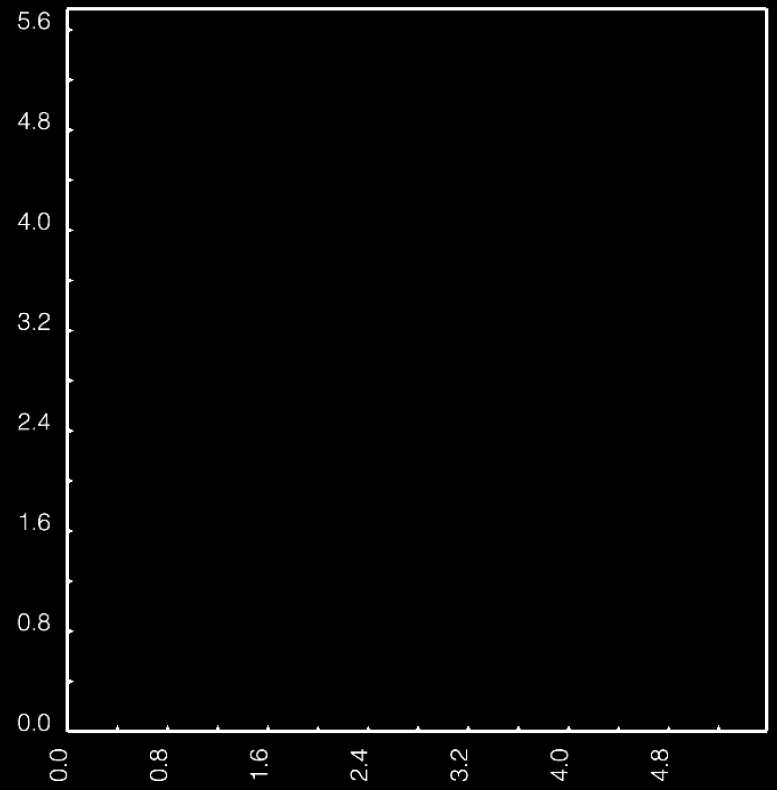
M. Mezard & A. Montanari, Information, Physics , and computation
Yedidia *et al,* IEEE TRANS. on inf. theo., vol. 51, no. 7, (2005)
Cherktkov *et al, PNAS* vol. 107, no. 17, p7663–7668 (2010

Selective Graph Cuts

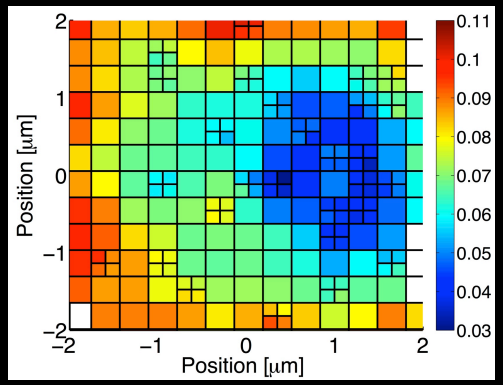
- Highly spatially varying parameters
- Global Solution overestimate parameters and induce high spatial smoothing
- Recursive Solving
  - Until Convergence (or predefined)
  - Parameters initialisation (Over estimation)
  - Mesh generation (Hierarchical)
  - Graph Cuts Based on the Mesh (O(N))
  - Belief Propagation to extract parameters

Temporal evolution of diffusive trap

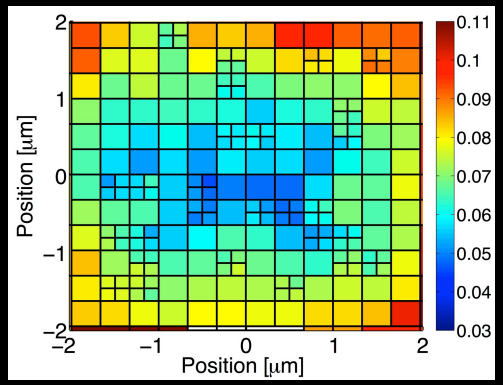
Double Trap



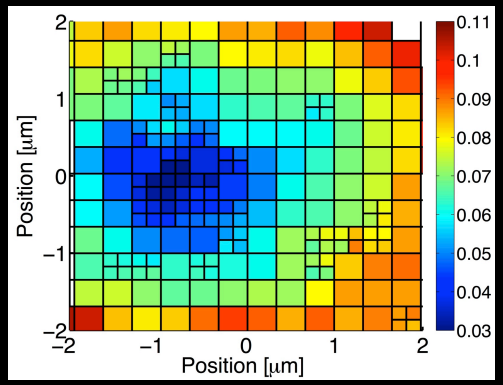
Numerical Simulations
Scaffolding proteins evolution



t=0s



t=30s



t=60s

Bayesian Inference

$$P(\{U_k\} \mid \{T_l\}, M_i) = \frac{P(\{T_l\} \mid \{U_k\}, M_i)\, P(\{U_k\} \mid M_i)}{P(\{T_l\} \mid M_i)}$$

- $P(\{U_k\}|\{T_l\},M_i)$ posterior distribution of the parameters $\{U_k\}$ having observed $\{T_l\}$ (set of trajectories) with Model M_i
- $P(\{T_l\}|\{U_k\},M_i)$ likelihood of model
- $P(\{U_k\}|M_i)$ prior distribution of the parameters
- $P(\{T_l\}|M_i)$ Evidence of the Model
- Bayesian Choice

$$\gamma_{ij} = \frac{P(\{T_l\}, M_i)}{P(\{T_l\}, M_j)} = \frac{P(\{T_l\} \mid M_i)\, P(M_i)}{P(\{T_l\} \mid M_j)\, P(M_j)}$$

- Bayesian Choice Systematically fails for most Random Walks analysis

H.L. Harney, Bayesian Inference, Parameter Estimation and Decision,
Voisinne *et al*, Biophysical Journal Volume 98 February 2010 596–605
S. Turkcan *et al*, PLoS ONE 8(12): e82799. doi:10.1371/journal.pone.0082799

G. E. P. Box & G. C, Tiao, Bayesian Inference in Statistical Analysis.

Biomolecules Random Walk

Overdamped Langevin Equation

Interaction Terms

Diffusive Term

$$\frac{d\vec{r}}{dt} = \frac{F_t(\vec{r})}{\gamma_t(\vec{r})} + \sqrt{2D_t(\vec{r})}\vec{\xi}(t)$$

$$F_t(\vec{r}) = -\nabla V(\vec{r})$$

$$\gamma_t(\vec{r}) = \frac{k_B T}{D_t(\vec{r})}$$

$$\langle \xi_i(t)\xi_j(t) \rangle = \delta(t-s)\delta_{ij}$$

- The Fokker-Planck Equation, Methods of solutions and Applications, H. Risken
- J.-B M *et al*, Phys. Rev. Lett 102 , 048103 (2009)
- J.-B M *et al*, Biophys J. vol 106, issue 1, p74-83 (2014)
- Hanggi *et al, J. Stat. Phys. Vol 18, No. 2 (1978)*
- O. Farago & N. Grønbech-Jensen, J. Stat. Phys. Vol 156, p1093 (2014)
- O. Farago & N. Grønbech-Jensen, Phys. Rev. E 89, 013301 (2014)
- R. Fox, J. Stat. Phys. Vol 46 issue 5/6 (1987)

Priors

- Experimental Priors: statistics from previous experiments

- Simulation Priors: Statistics and Cost Functions

C. Salvatico et al, Nat. Neuro. (submitted)

- Jefferey's Prior

$$P(\{U\}) \propto \sqrt{|J|} \quad J = \left(\partial_U \partial_{U'}^T \int d\mathbf{r} \sqrt{P(\mathbf{r}|\{U\}) P(\mathbf{r}|\{U'\})} \right)_{U=U'}$$

$$P(D(\vec{r})) \propto \frac{D^2(\vec{r})}{\left(D(\vec{r})\Delta t + \sigma^2\right)^2}$$

JBM *et al*, Biophys. J., vol. 106 issue 1 (2014)

- Field Priors

$$P(D(\vec{r}), V(\vec{r})) \propto \exp\left(-\mu \iint \|\nabla D(\vec{r})\|^2 d^2r - \lambda \iint \|\nabla V(\vec{r})\|^2 d^2r \right)$$

- Non-Local Priors

$$P(D(\vec{r})) \propto \exp\left(-\sigma\left(D(\vec{r}), \vec{F}(\vec{r}) \right) \right)$$

(complex optimization)

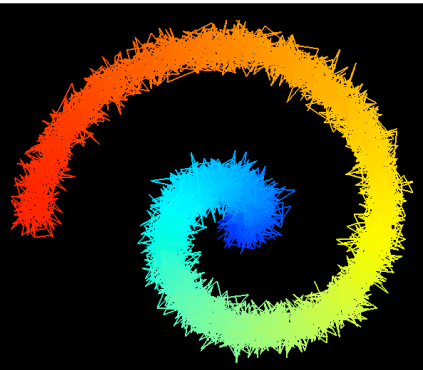
Field Theories for learning probability distribution, W. Bialek *et al*, PRL 77, 4693, (1996)
Bayesian Inference, Parameter Estimation and Decision, H.L. Harney

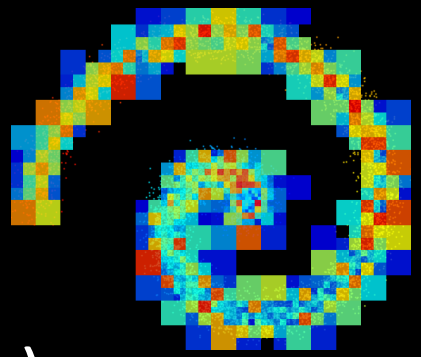
Multi-Scale Heterogeneities

- Unknown level or scale of heterogeneities
- Level : $\Delta D \sim 0.01 - 0.2\ \mu m^2.s^{-1}$ (100nm) synapses
- Scales : l~1nm-1µm. Full Cell Scale.
- Measures at Small Scales: dt ~ 10ms , dr ~ 10nm
- Partially Unsupervised Learning Problem
- Local single Molecule Density might be a clue (Biology)

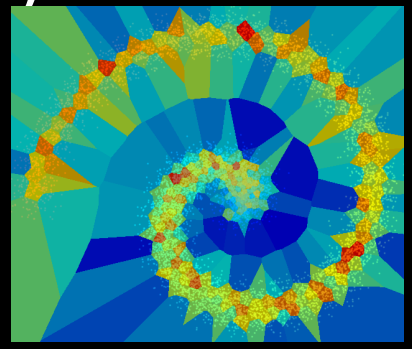
2 Main Methods:

- Quadtree (Octree for 3D)
  - Hierarchical Graph
  - Subdivide space
  - 2 criteria
    - Capacity -> Information
    - Length scale -> local diffusivity scale





- K-Mean Clustering (Gaussian Mixture,…)
  - Unsupervised Learning
  - Initialization (Random, Density Based, Bubbling)
  - K-Mean clustering
  - Hierarchical updates (divide, fusion etc)



C. Bishop, Pattern Recognition and Machine Learning; R. Finkel & J.L. Bentley, Acta Informatica, vol 4, issue 1, p1-9 (1974).

Fokker-Planck and the Ito/Stratonovitch/Hanggi Dilemma

- Different Definition of the Stochastic Integrals

$$\frac{\partial P\left(\mathbf{r},t\left|\mathbf{r}_0,t_0\right.\right)}{\partial t}=-\nabla\bullet\left[\left(\frac{F_t\left(\mathbf{r}\right)}{\gamma_t\left(\mathbf{r}\right)}+\lambda\nabla D_t\left(\mathbf{r}\right)\right)P\left(\mathbf{r},t\left|\mathbf{r}_0,t_0\right.\right)\right]+\nabla\bullet\left[D_t\left(\mathbf{r}\right)\nabla P\left(\mathbf{r},t\left|\mathbf{r}_0,t_0\right.\right)\right]$$

- Ito λ=0, Stratonovitch λ=1/2 and Hanggi λ=1

- $D(\mathbf{r}\to\mathbf{r}_0)\sim D(\mathbf{r}_0)$,
- $\mathbf{F}(\mathbf{r}\to\mathbf{r}_0)\sim\mathbf{F}(\mathbf{r}_0)$,
- $\nabla V(\mathbf{r}\to\mathbf{r}_0)\sim\nabla V(\mathbf{r}_0)$

- $D(\mathbf{r}\to\mathbf{r}_0)\sim D(\mathbf{r}_0)+\nabla D.(\mathbf{r}-\mathbf{r}_0)$
- $\mathbf{F}(\mathbf{r}\to\mathbf{r}_0)\sim\mathbf{F}(\mathbf{r}_0)$,
- $\nabla V(\mathbf{r}\to\mathbf{r}_0)\sim\nabla V(\mathbf{r}_0)$

$$P\left(\mathbf{r},t;\mathbf{r}_0,t_0\left|D,F\right.\right)=\frac{\exp\left(-\left(\mathbf{r}-\mathbf{r}_0-DF\Delta t/k_BT\right)^2\Big/4\left(D+\frac{\sigma^2}{\Delta t}\right)\Delta t\right)}{4\pi\left(D+\frac{\sigma^2}{\Delta t}\right)\Delta t}$$

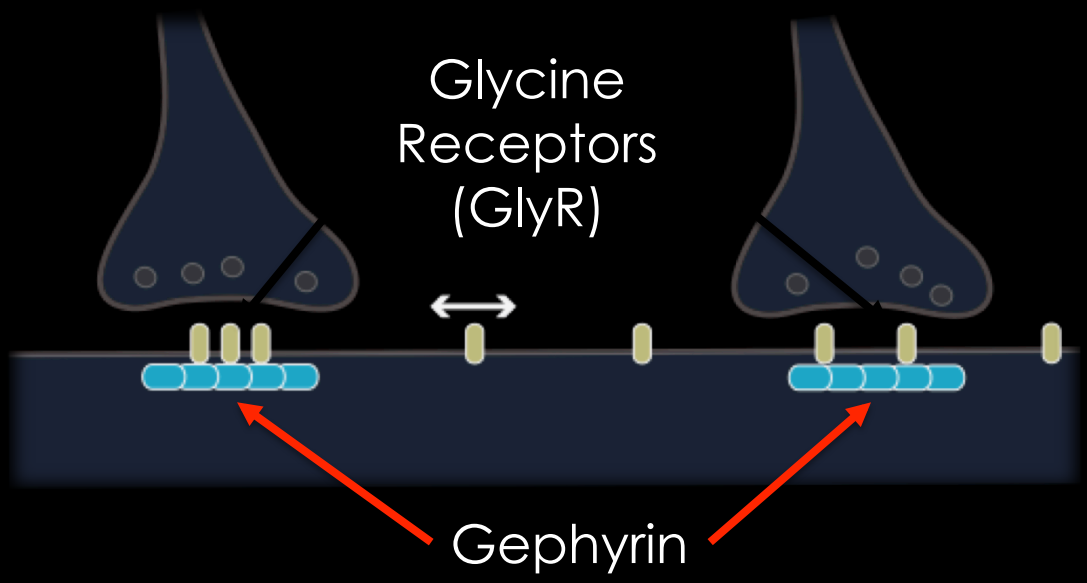
$$P\left(\mathbf{r},t;\mathbf{r}_0,t_0\left|D,\nabla D,F\right.\right)=$$

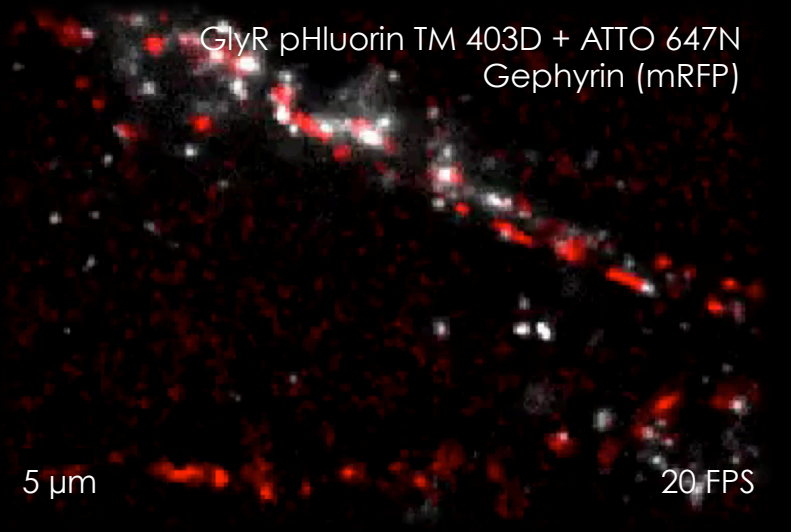
$$\frac{1}{2\pi}\int_{-\infty}^{\infty}\frac{1}{\left(1+ik\nabla D\Delta t\right)^{\left(\frac{F/\gamma+(1+\lambda)\nabla D}{\nabla D}\right)}}e^{-\frac{k^2\Delta t\left(D+\sigma^2/\Delta t-\nabla Dx_0\right)+ikx_0}{1+ik\nabla D\Delta t}}e^{ikx}$$

The Fokker-Planck Equation, Methods of solutions and Applications, H. Risken
O. Farago, N. Grønbech-Jensen, J. Stat. Phys. 156, 1093, (2014)
P. Hanggi, Helvetica Physica Acta, Vol. 51 (1978)
R. L. Fox,, j. Stat. Phys., Voi. 46, Nos. 5/6, 1987

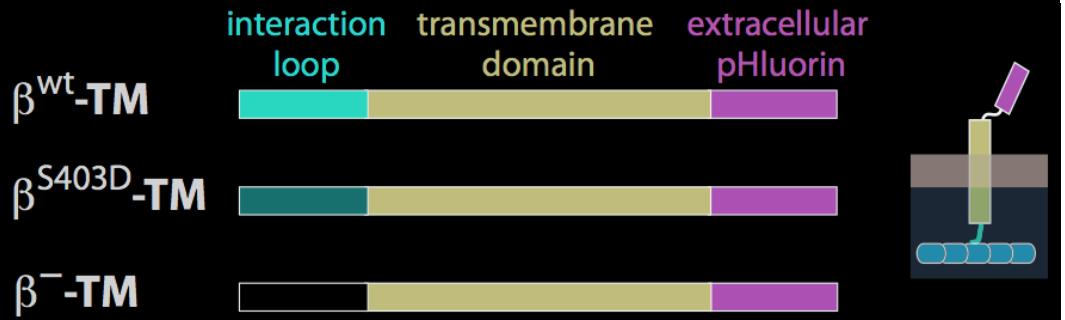
Inhibitory synapses: GlyR and Gephyrin

- Gephyrin is a critical scaffold protein in the inhibitory neuronal synapse
- Synapses in neurons and artificial



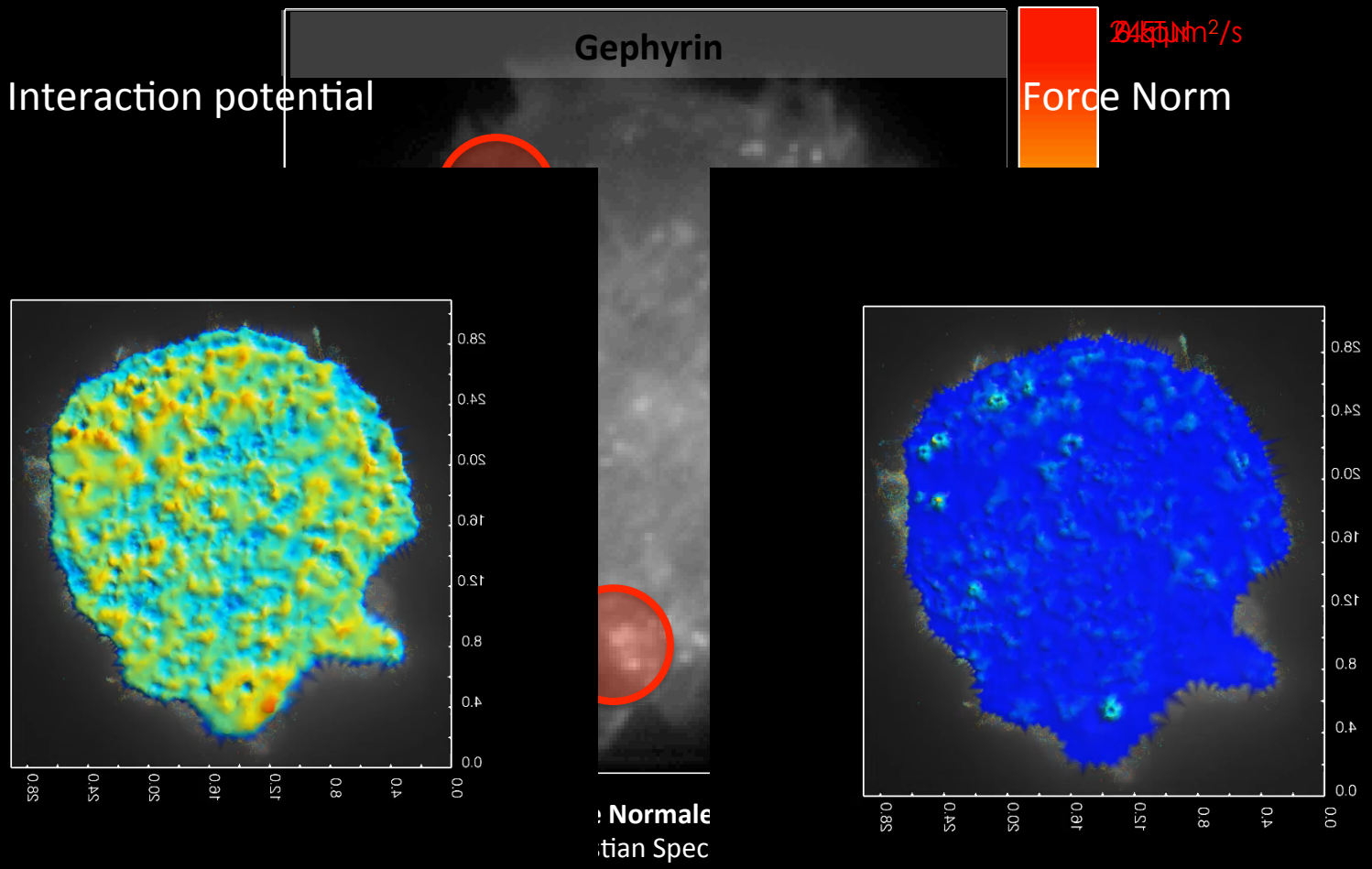


GlyR Constructions



J.-B Masson *et al, Biophys. J*, vol. 106, no. 1, 74–83 (2014).
C. Salvatico *et al,* Nat. Neuro. (*In Preparation*)

Artificial Synapses in Cos Cells



Generative Model: Simulation in The Maps

- Fokker Planck ⟶ Master Equation
- Masters Equations

$$\frac{dP_{(i,j)}(t)}{dt} = \sum_{(i',j')\in N(i,j)} W_{(i,j),(i',j')} P_{(i',j')}(t) - \sum_{(i',j')\in N(i,j)} W_{(i',j'),(i,j)} P_{(i,j)}(t) \qquad a_v = W_{(i,j),(i',j')}$$

$$W_{(i,j),(i',j')} = \frac{D(i',j')}{\Delta x^2} \exp\left(-\frac{\Delta x F^x_{(i,j),(i',j')}}{2\gamma(i',j') D_{(i',j')}}\right) \qquad a_0 = \sum_v a_v$$

- Multi-scale Space Structure
- Time to wait

$$\tau = \frac{1}{a_0}\log\left(\frac{1}{r_1}\right) \qquad r_1 = [0..1]$$

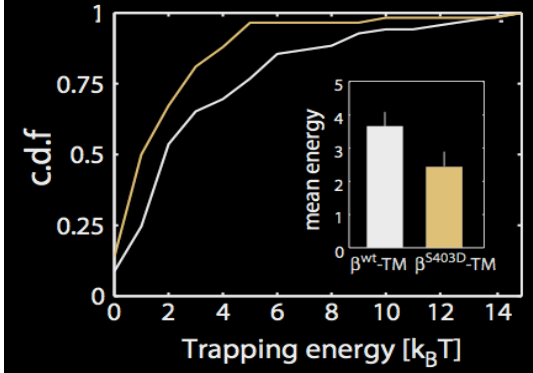
- Site to chose, k,

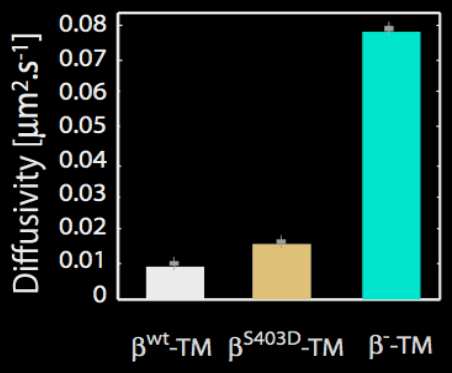
$$\sum_{v=0}^{k-1} a_v \le r_2 a_0 < \sum_{v=0}^{k} a_v \qquad r_2 = [0..1]$$

Gillespie, D. ,Exact Stochastic Simulation of Coupled Chemical Reactions. J. of Phys. Chem. 81, 2340-2361.(1977)

Few Results

In the synapses

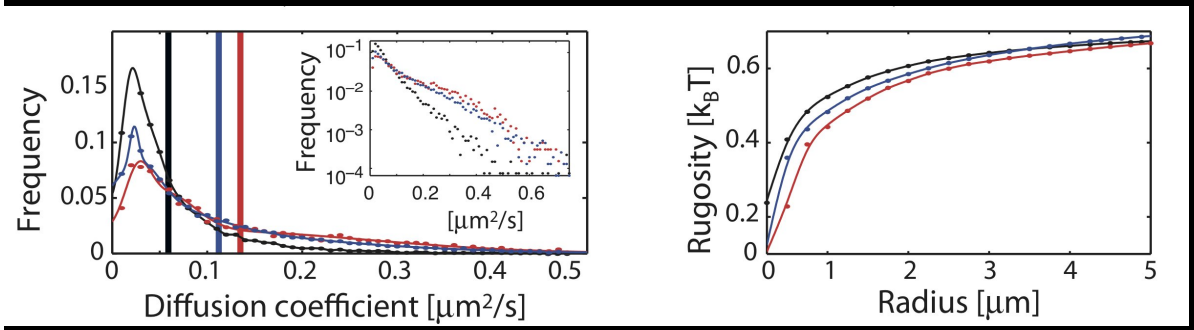


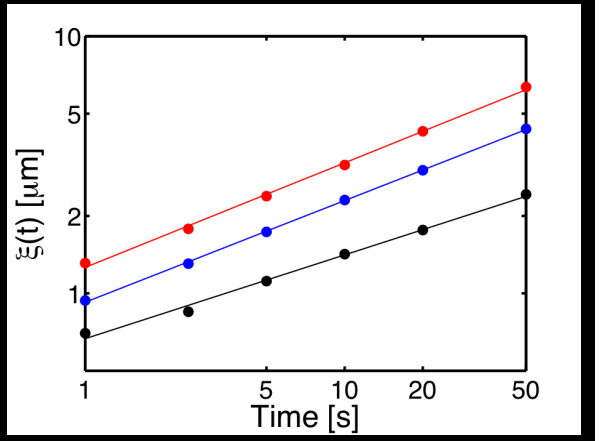


$$P(r) = \frac{1}{\pi \xi^2(t)} e^{-\frac{r^2}{\xi^2(t)}}$$

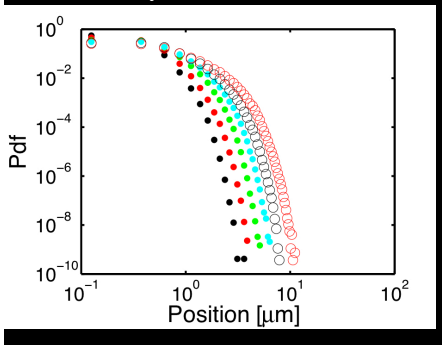
$$\rho = \frac{r}{\xi(t)}$$

β^{WT}-TM, β^{S403D}-TM, β^{-}-TM

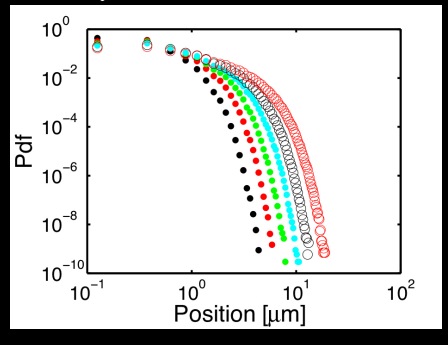




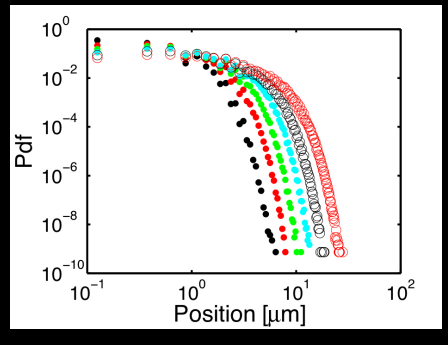
β^{wt}-TM



β^{S403D}-TM



β^{-}-TM



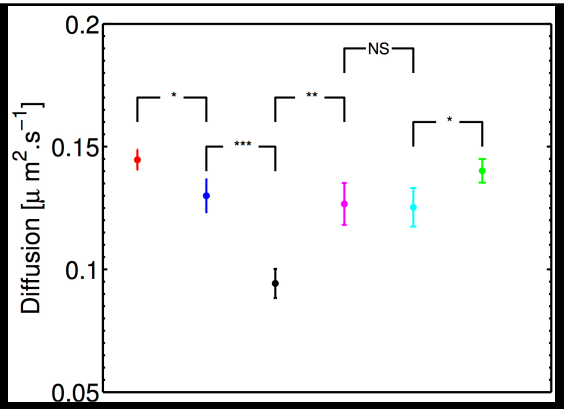
$$\xi(t) \propto t^{\gamma}$$

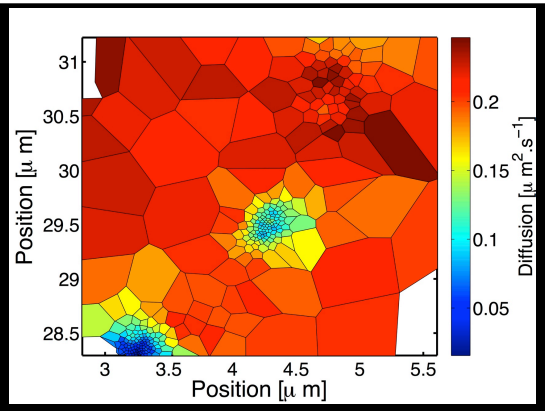
β^{-}-TM: γ=0.41

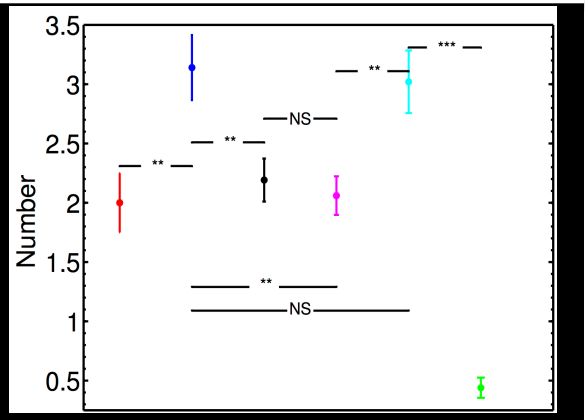
β^{S403D}-TM: γ=0.39

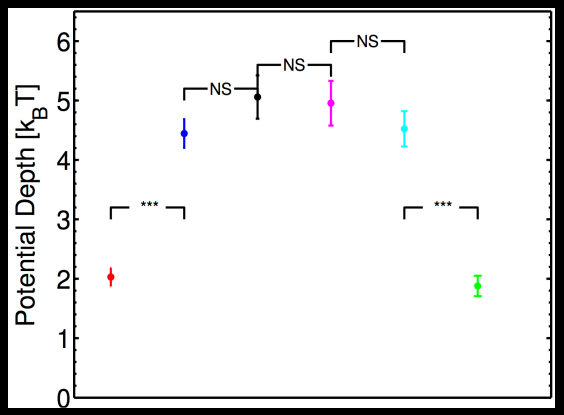
β^{wt}-TM: γ=0.33

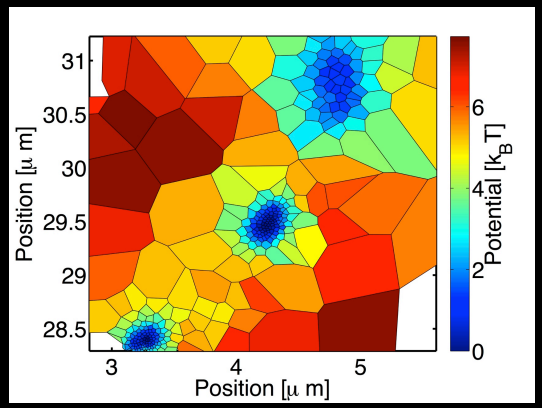
Pipeline On Neurons: 450 Go -> 670 synapses (1 Hour)

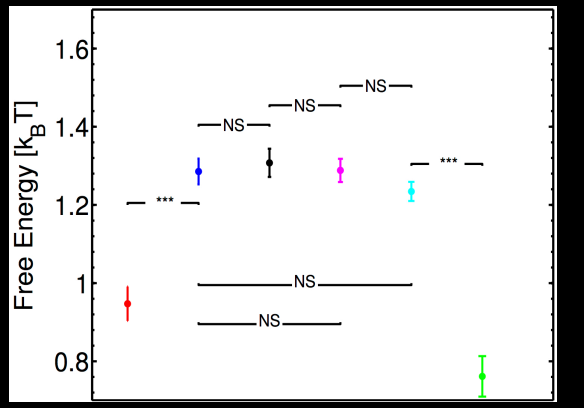


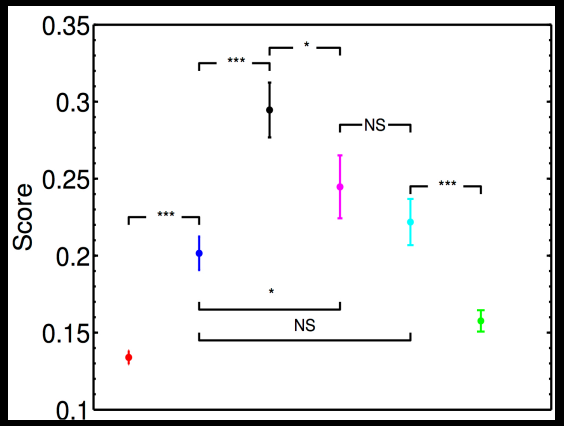


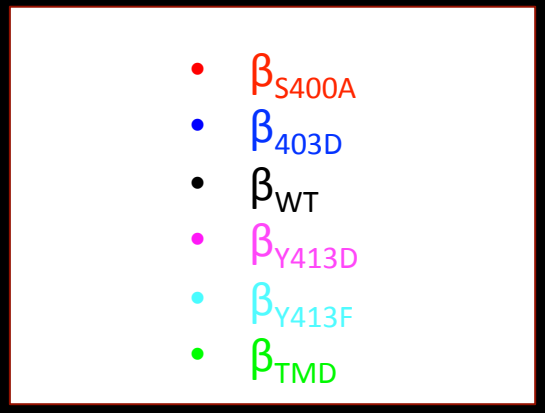


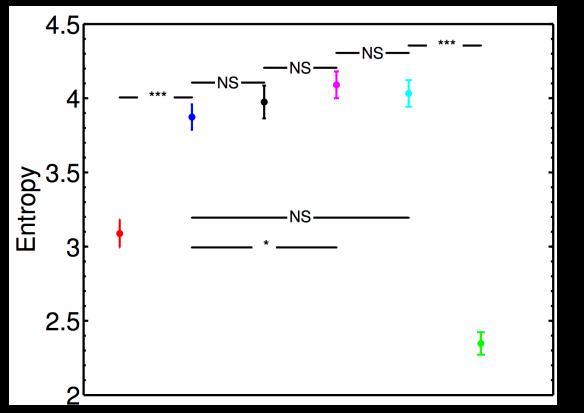












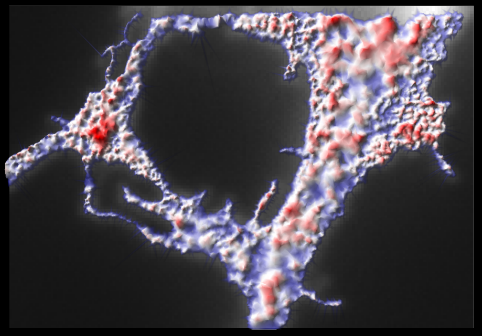
Equilibrium?

- Is the process at equilibrium?
- Methods to detect Characteristics of Non-Equilibrium dynamics
  - Work of the force fields $\Delta W = \oint \vec{F} \cdot d\vec{r}$ on random close contour in the Maps
    $\vec{F} = -\vec{\nabla} V + \vec{\nabla} \times \vec{A}$ Ill posed problem
  - Asymmetries in the paths (A->B, B->A)
    - Metric on the paths
    - Relative Entropy
  - Detailed Equilibrium measure at various scales of the mesh $\chi = P_i P_{i \to j} / P_j P_{j \to i}$
- All compared to simulations (information accumulated)

R. Kawai *et al, PRL* **98**, 080602 (2007)
A. Comez-Marin *et al, EPL 82 50002 (2008)*
M. Esposito *et al, N. J. Phys., 12 013013 (2010)*
J. Horowitz & C. Jarzynski, PRE 79 021106 (2009)
A. V. Popov & R. Hernandez, PRL

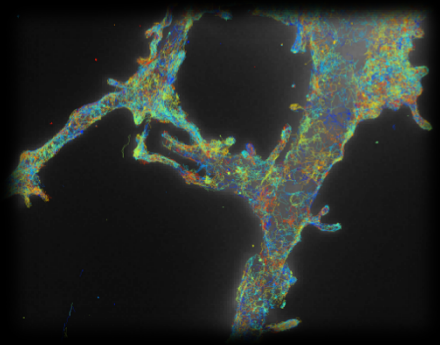
InferenceMAP

Transmembrane Protein Diffusion in Neurons

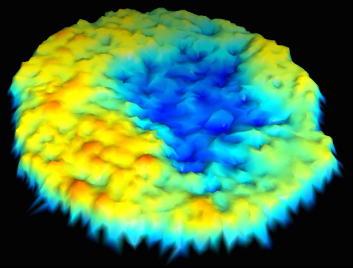


P. Dionne, Ecole Normale Supérieure

Tailored for Big Data



Nuclear Protein Diffusion



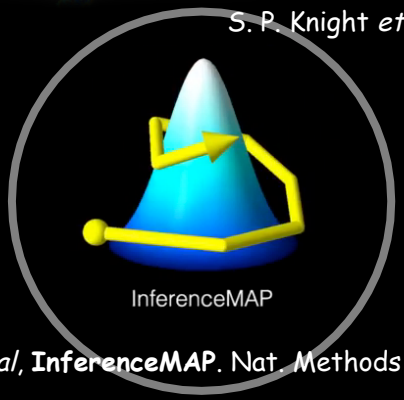
J. Liu, Janelia Research Campus

Calculation Flexibility

$$P\left(D(\vec{r}), \vec{F}(\vec{r})|T\right)$$

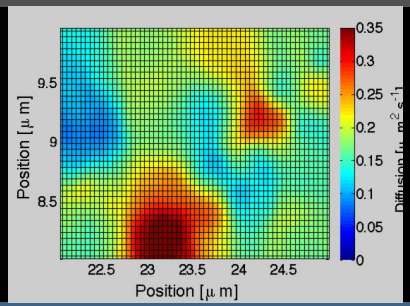
$$P\left(D(\vec{r}), V(\vec{r})|T\right)$$

S. P. Knight *et al*, *Science 350, 6262 p823 (2015)*



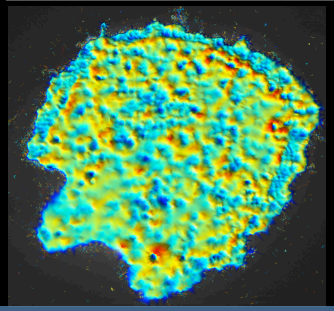
M. El Beheiry *et al*, **InferenceMAP**. Nat. Methods 12, 594, (2015)

Rac$_1$ dynamics



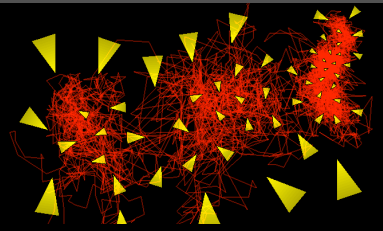
A. Remorino, Insitut Curie

Membrane Protein Interaction Energy



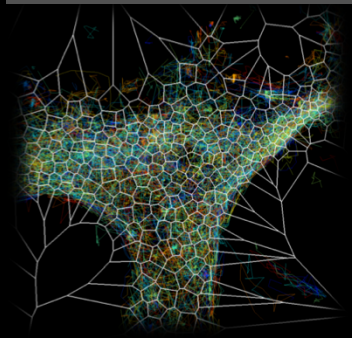
M. El Beheiry, Institut Curie

Membrane Confinement Zones

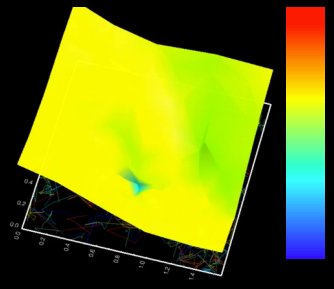


S. Türkcan, Ecole Polytechnique

Adaptive Multi-Scale Meshing

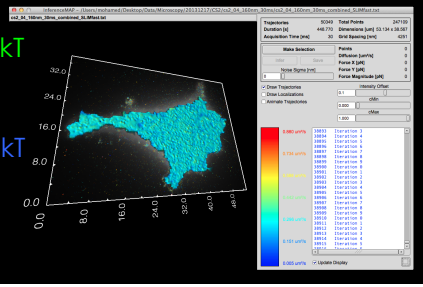


Energy of Virion Formation



C. Floderer, CNRS CPBS

User-Friendly Interface



Conclusion

- Global interdisciplinary Approach
  - Physics
  - Bayesian inference
  - Statistics
- Challenges
  - Sampling complex distributions
  - Heterogeneity
  - Transition between models
  - Multi-Method Model selection
- Bayesian Inference as a global framework
  - Bacterial Chemotaxis (JBM *et al*, PNAS vol 109, No 5, p1802-1807 (2012))
  - Leucocyte search strategies 22, p2375-2382 (Sarris *et al*, Current Biology (2012))
  - Searching in turbulent environemnt (JBM, PNAS vol 110, No 28, p11261-11266 (2013))

Acknowledgments

Pasteur Institute
- Massimo Vergassola (now at UCSD)
- Jerome Wong-Ng (now at UCSD)
- Guillaume Voisinne (now at Sloan Kettering)
- David Digregorio Lab
- Christophe Zimmer Lab
- Sven von Teeffelen Lab

Ecole Polytechnique (LOB)
- Antigoni Alexandrou Team
- Max Richly

Stanford
- Silvan Tuerkcan

Colorado State University
- Diego Krapf Lab

- Institut Curie
- Maxime Dahan Lab
- Mohamed El Beheiry
- Mathieu Coppey
- Amanda Remorino
- Bassam Hajj
- Patricia Bassereau Lab

Janelia Research Campus
- Marta Zlatic Lab
- Brian English
- Timothée Lionnet
- TIC Consortium

- Ecole Normale Superieure
- Antoine Triller Lab
- Charlotte Salvatico
- Marianne Renner
- Christian Specht

Grants:
ANR Grand Emprunt: Pherotaxis (2011-2015), ANR 07 NANO 062 03 (2010-2013), ANR 09 PIRI 0025 01 (2011-2014), C'NANO Ile de France (2010-2012), Masson Visitng Project Janelia Reseach Campus (2014-2016)