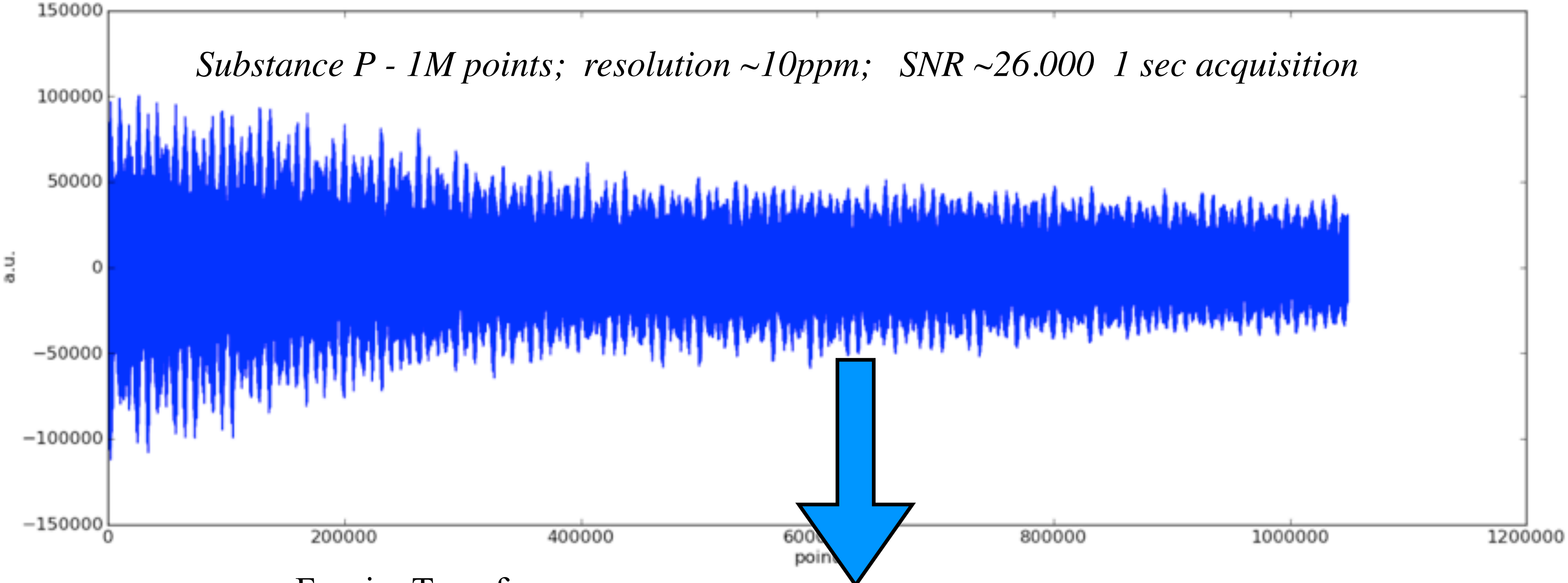
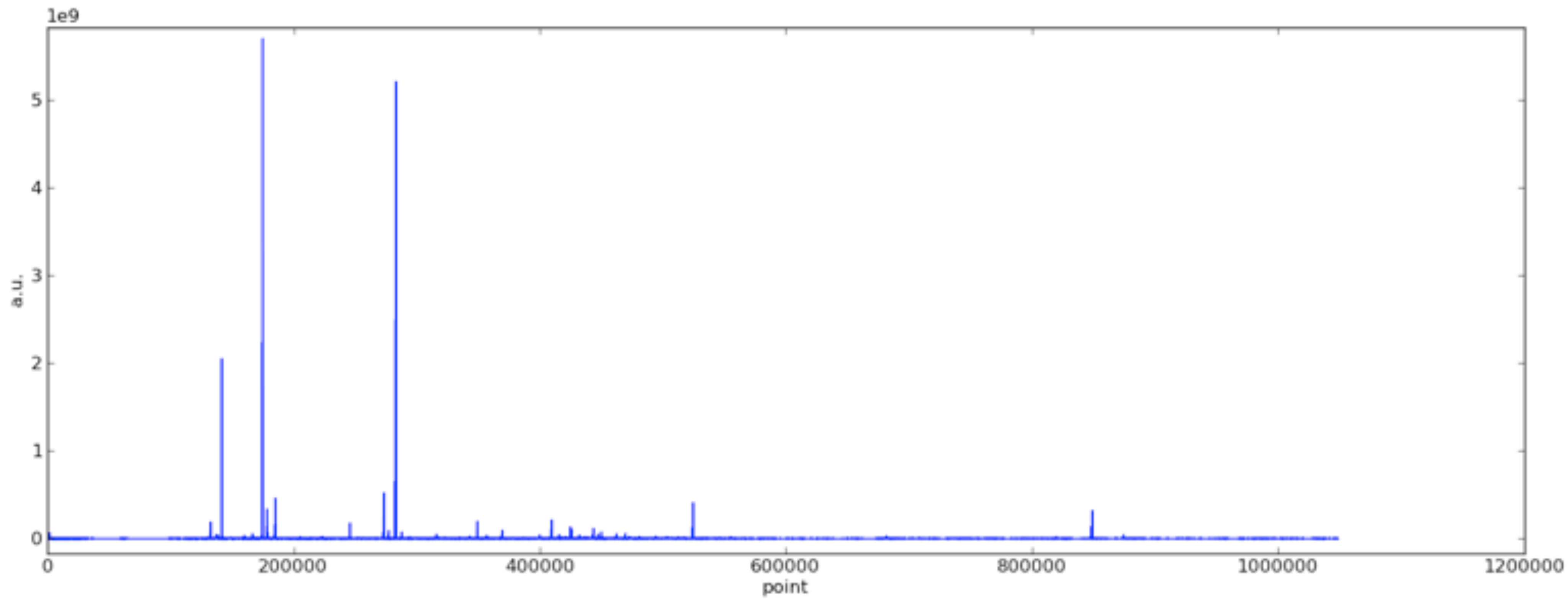
Extensions aux approches par transformée de Fourier

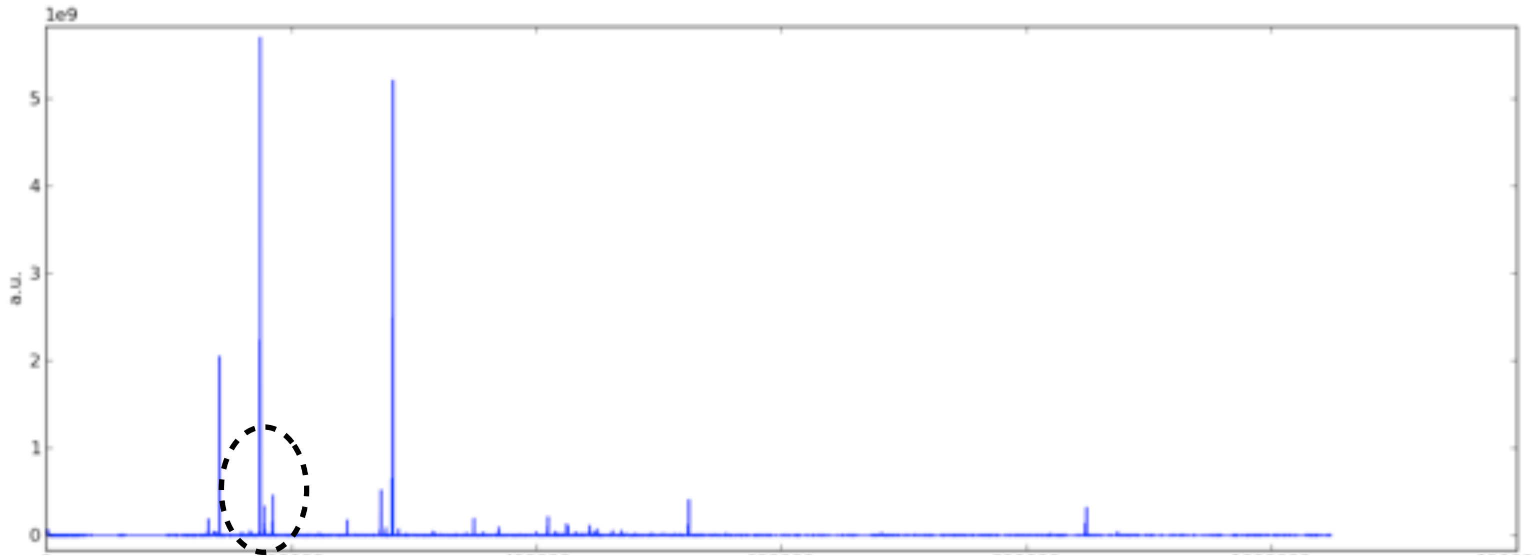
Marc-André Delsuc
IGBMC

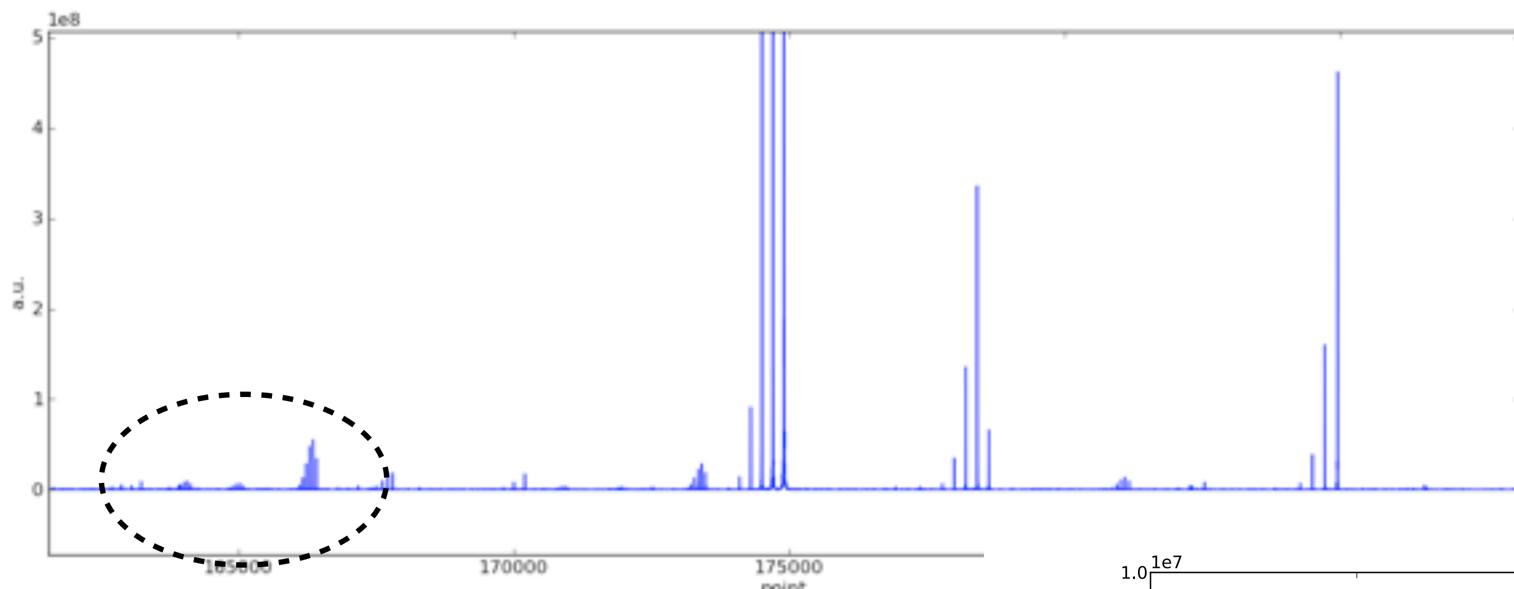


Fourier Transform
Modulus

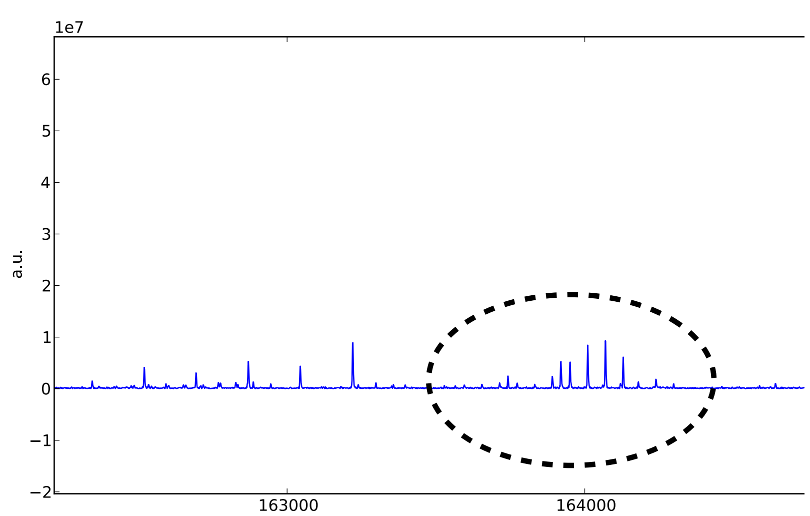


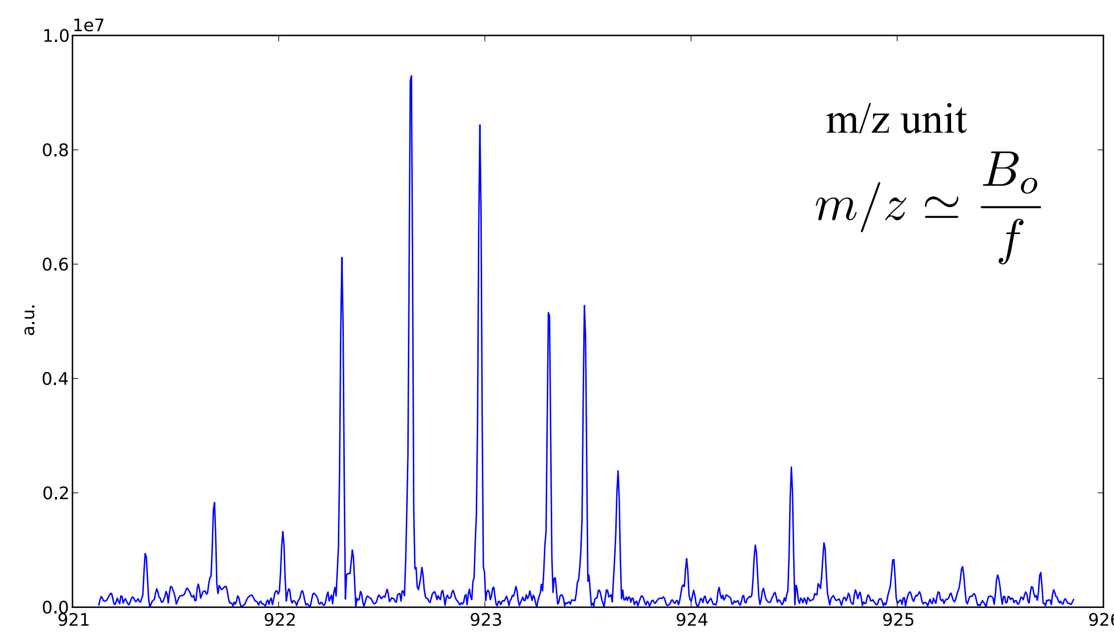
Zooming in





Substance P
1M points;
resolution ~10ppm
SNR ~26.000
1 sec acquisition





Quel traitement de données ?

- Approche générale :
  - N points de mesures
  - P paramètres à extraire

$$y^{mes} = T_f(s) + \epsilon$$

$N \qquad\qquad P$

- $N > P$
  - **l'ajustement** de paramètres (le *fit*)
  - <u>modélisation</u> du phénomène

FDM

- $N = P$
  - les **transformations** de données $T_f(\,)$ inversible ou non
  - <u>modélisation</u> de la mesure

Zero-Filling X

•Sur-résolution

•débruitage

•2D

- $N < P$
  - la **reconstruction** de données
  - <u>modélisation</u> de la mesure et de la connaissance

MaxEnt

Compressed-Sensing

sur-resolution

- La résolution de la mesure est dominée par la relation de Gabor-Heisenberg

$$t_{max} \Delta F = 1$$

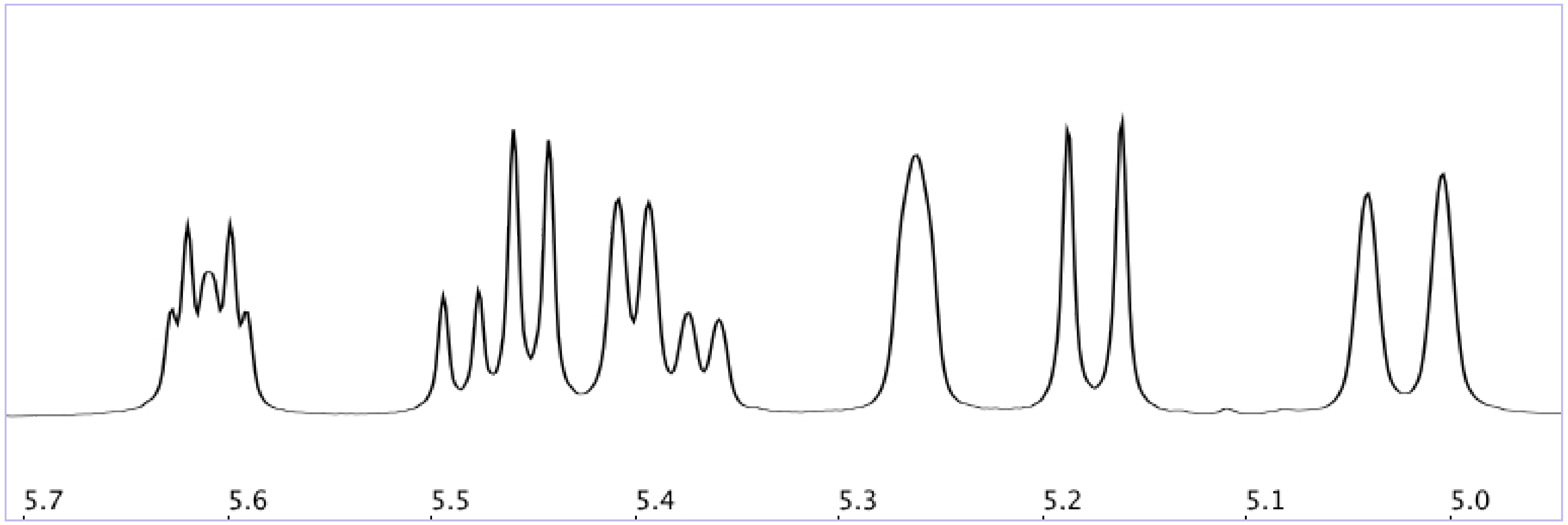
- maximiser la résolution pour une acquisition donnée

  Trois approches

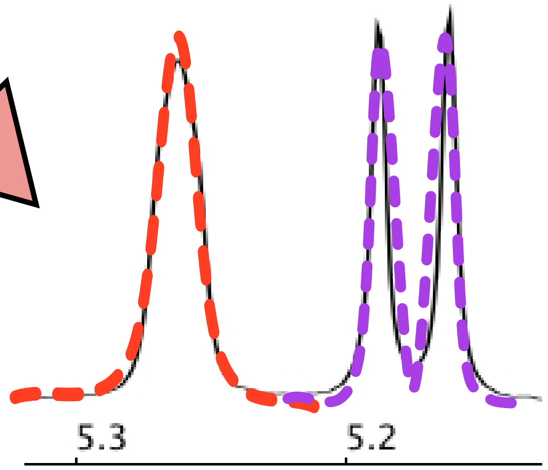
  - - ajustement
  - - transformation
  - - reconstruction / régularisation

- dépend du rapport signal à bruit

N>P L'ajustement



fit



#	pos.ppm	width Hz
1	5.24 ± 0.02	23 ± 5
2	5.18 ± 0.03	3.5 ± 0.3
...		

N>P L'ajustement

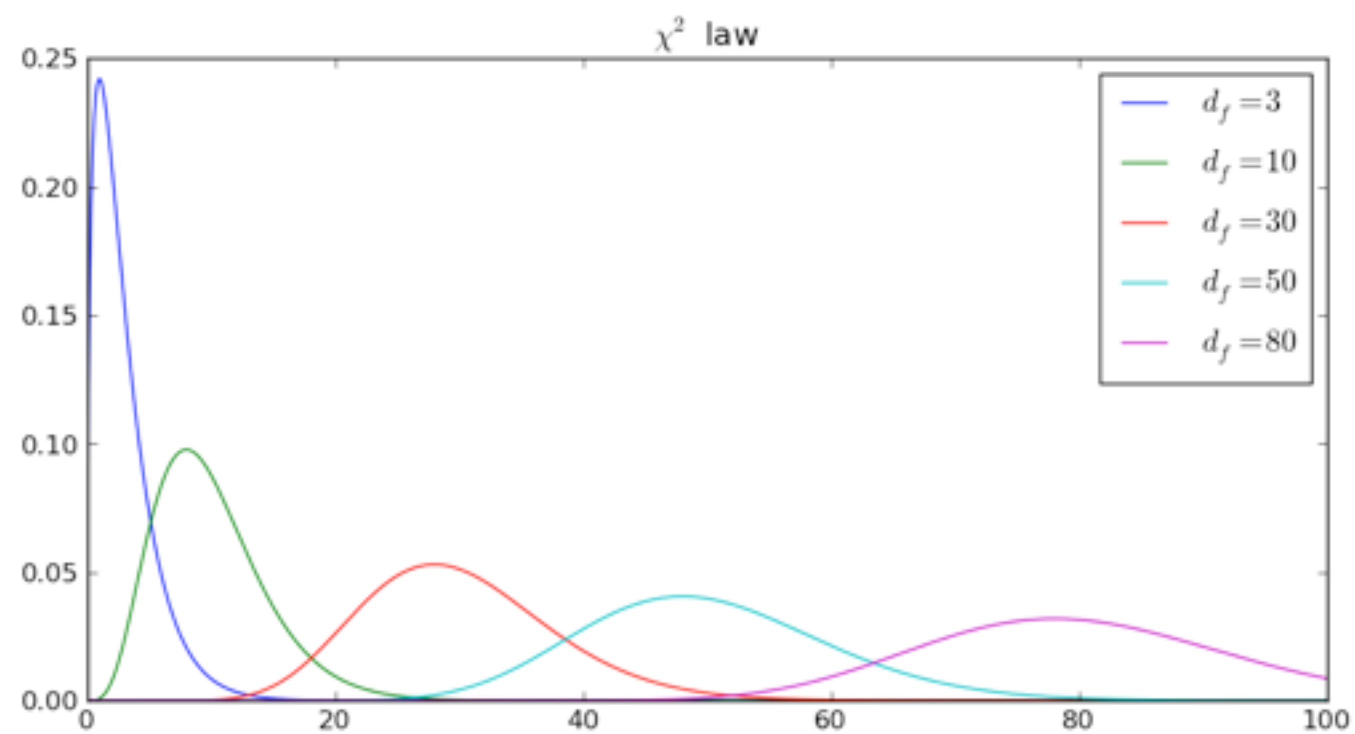
- modélisation du phénomène
  - modèle physique du phénomène étudié
  - on «ajuste» les P paramètres du modèle aux données mesurées
  - écart mesuré par maximum de vraisemblance

    solution des «moindres carrés»

$$y_i^{calc} = f(P_j)$$

$$\chi^2 = \sum_i \left(\frac{y_i^{calc} - y_i^{mes}}{\sigma_i} \right)^2$$

- degrés de liberté
  - df = N-P
- Loi du Chi2
  - moyenne = df
  - maximum = df-2
  - variance = 2df



Principes

- minimiser le Chi2
  - c'est une fonction convexe/quadratique ⇒ convergence rapide
  - $f(P_i)$ peut-être complexe ⇒ minimum non-unique

    ⇒ importance du choix des valeurs initiales
- Dans la pratique
  - utiliser une bibliothèque toute faite !

`curve_fit`(f, xdata, ydata, **kw[, p0, sigma])	Use non-linear least squares to fit a function, f, to data.

  - pour chaque paramètre ajusté on peut estimer l'erreur à partir de la matrice de covariance

Returns

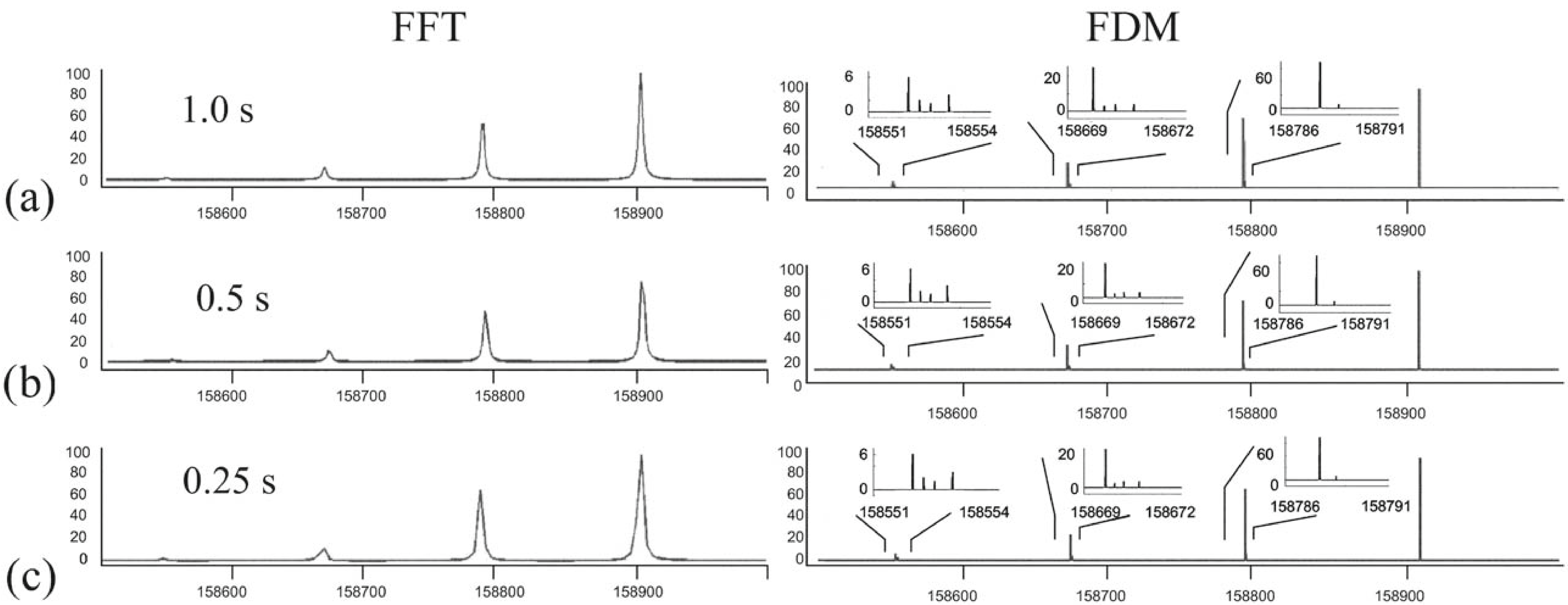
popt : array

Optimal values for the parameters so that the sum of the squared error of `f(xdata, *popt) - ydata` is minimized

pcov : 2d array

The estimated covariance of popt. The diagonals provide the variance of the parameter estimate.

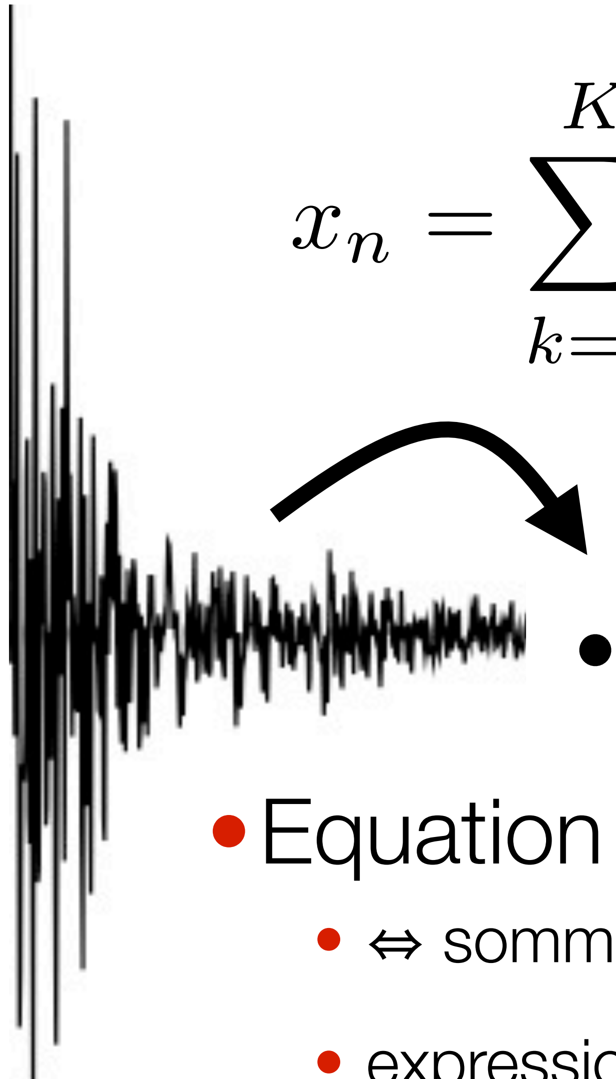
un exemple moins trivial



Aizikov et O'Connor. J.Am. Soc.Mass Spec. (2006) ***17*** *(6) pp. 836-843*

- Modèle AR : autoregressif
  - signal : somme de **n sinusoides** amorties **exponentiel**lement
- Filter Diagonalization Method
  - equivalent à un fit dans les données temporelles
  - méthode intelligente pour choisir les valeurs initiales; convergence optimisée

Modèle AutoRegressif



$$x_n = \sum_{k=1}^{K} b_r x_{n-k}$$

$$\begin{bmatrix} x_1 & x_2 & \ldots & x_P \\ x_2 & x_3 & \ldots & x_{P+1} \\ x_3 & x_4 & \ldots & x_{P+2} \\ x_4 & x_5 & \ldots & . \\ . & . & \ldots & . \\ x_{L-P} & x_{L-P+1} & \ldots & x_{L-1} \end{bmatrix} \begin{bmatrix} b_P \\ b_{P-1} \\ b_{P-2} \\ . \\ b_1 \end{bmatrix} = \begin{bmatrix} x_{P+1} \\ x_P \\ x_{P-1} \\ . \\ x_L \end{bmatrix}$$

Matrice de Hankel

- Equation de Prédiction Linéaire (LP)
  - ⇔ somme de K sinusoïdes amorties
  - expression matricielle du problème LP

    $$Hb = x$$
  - H est décomposée pour extraire les fréquences

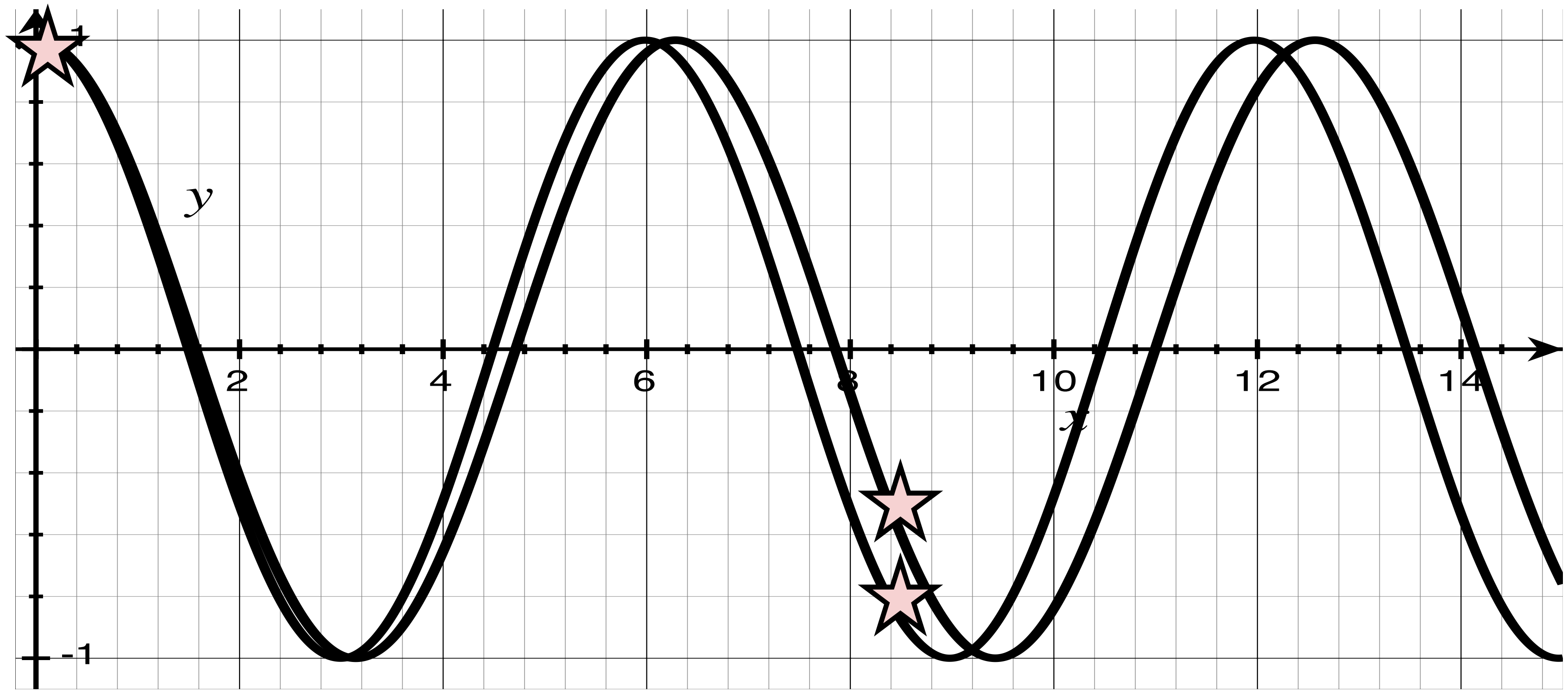
    valeurs propres, valeurs singulières,
  - Les intensités sont mesurées dans un deuxième temps

    par un moindre carré standard (Méthode de Prony)

Prony, R. Essai Expérimental et Analytique... *J. de l'Ecole Polytechnique (Paris)* **1,** 24–76 (1795).

Et le principe d'incertitude ?

- $$t_{max} = \frac{1}{\Delta F} \quad ?$$



- modélisation du signal
  - nombre de signaux
  - forme des signaux
- Information *a-priori*

$N{=}P$ la notion de transformée

- changement de point de vue
  - Typique des spectroscopies par FT (mais plus..)

    spectroscopies : NMR, FT-ICR, FT-IR, ...

    images : IRM, jpeg

$$y^{mes} = T_f(s) + \epsilon$$

- On **modèle** le processus de mesure

  notions de spectre / de fonction de mesure

  notion de fonction d'instrument (réponse impulsionelle - points manquants) -...

  hypothèse linéaire

- le problème au moindre carré
  - trouver $\tilde{s}$ le plus proche possible de *s*

    tel que

$$\chi^2 = \sum \left(\frac{T_f(s) - y^{mes}}{\sigma} \right)^2$$

$$\tilde{s} = T_f^{-1}(y^{mes})$$

  - la solution «triviale» est - *ou n'est pas* - la solution

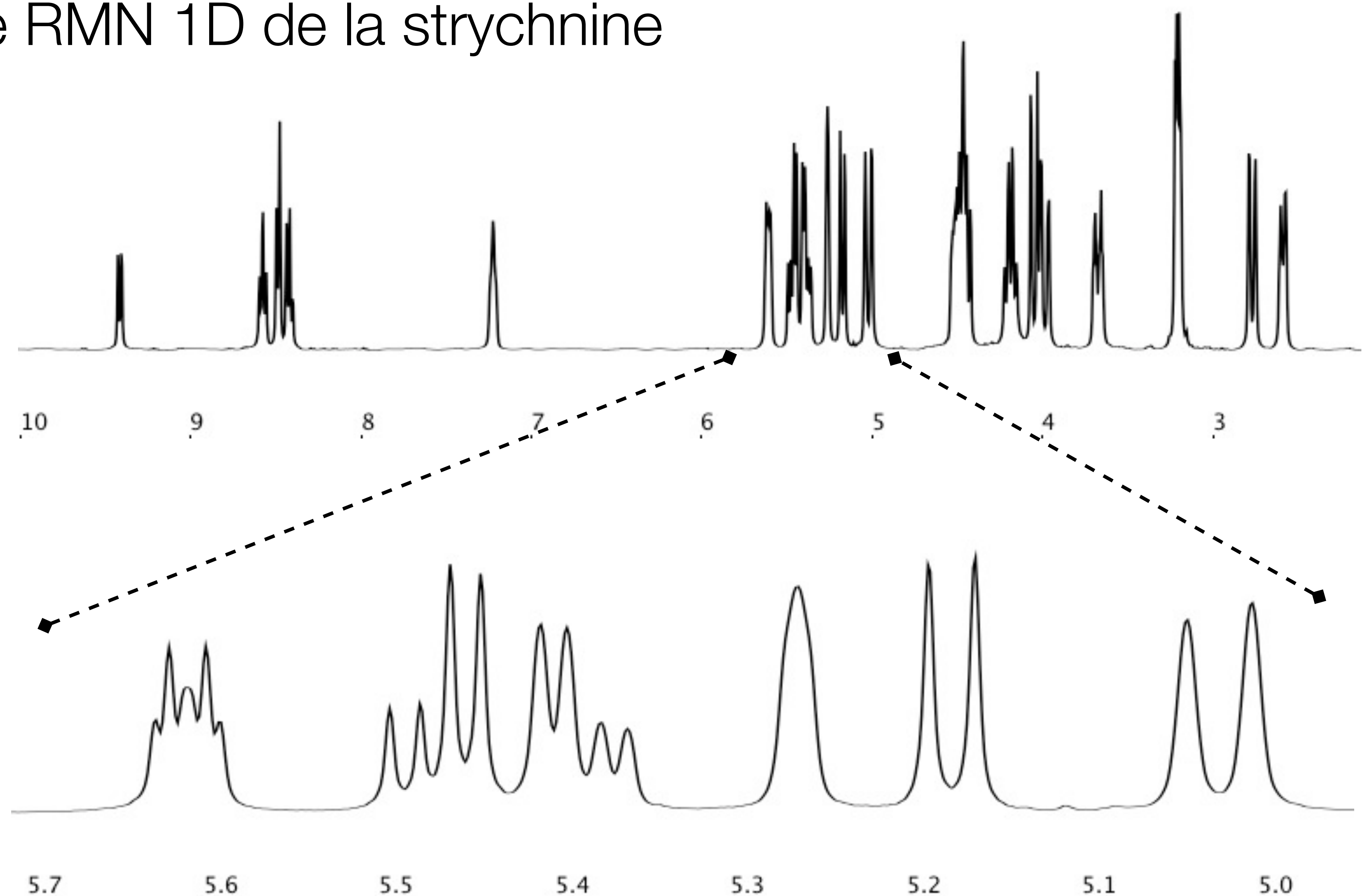
il existe différentes transformées

- transformée de Fourier
  - inversible
- transformée d'Hadamard
  - sorte de transformée de Fourier sur {-1, 1}
  - inversible
- transformée de Hilbert
  - la transformée qui transforme un signal réel en la partie imaginaire du signal analytique
  - calculée via la transformée de Fourier
  - inversible
- transformée de Laplace
  - transformée sur les fonctions exponentielles
  - non-inversible
- transformée de Radon
  - transformée sur les fonctions projection
  - non-inversible
- ...

$$\begin{pmatrix} + & + & + & + & + & + & + & + \\ + & + & + & + & - & - & - & - \\ + & + & - & - & + & + & - & - \\ + & + & - & - & - & - & + & + \\ + & - & + & - & + & - & + & - \\ + & - & + & - & - & + & - & + \\ + & - & - & + & + & - & - & + \\ + & - & - & + & - & + & + & - \end{pmatrix}$$

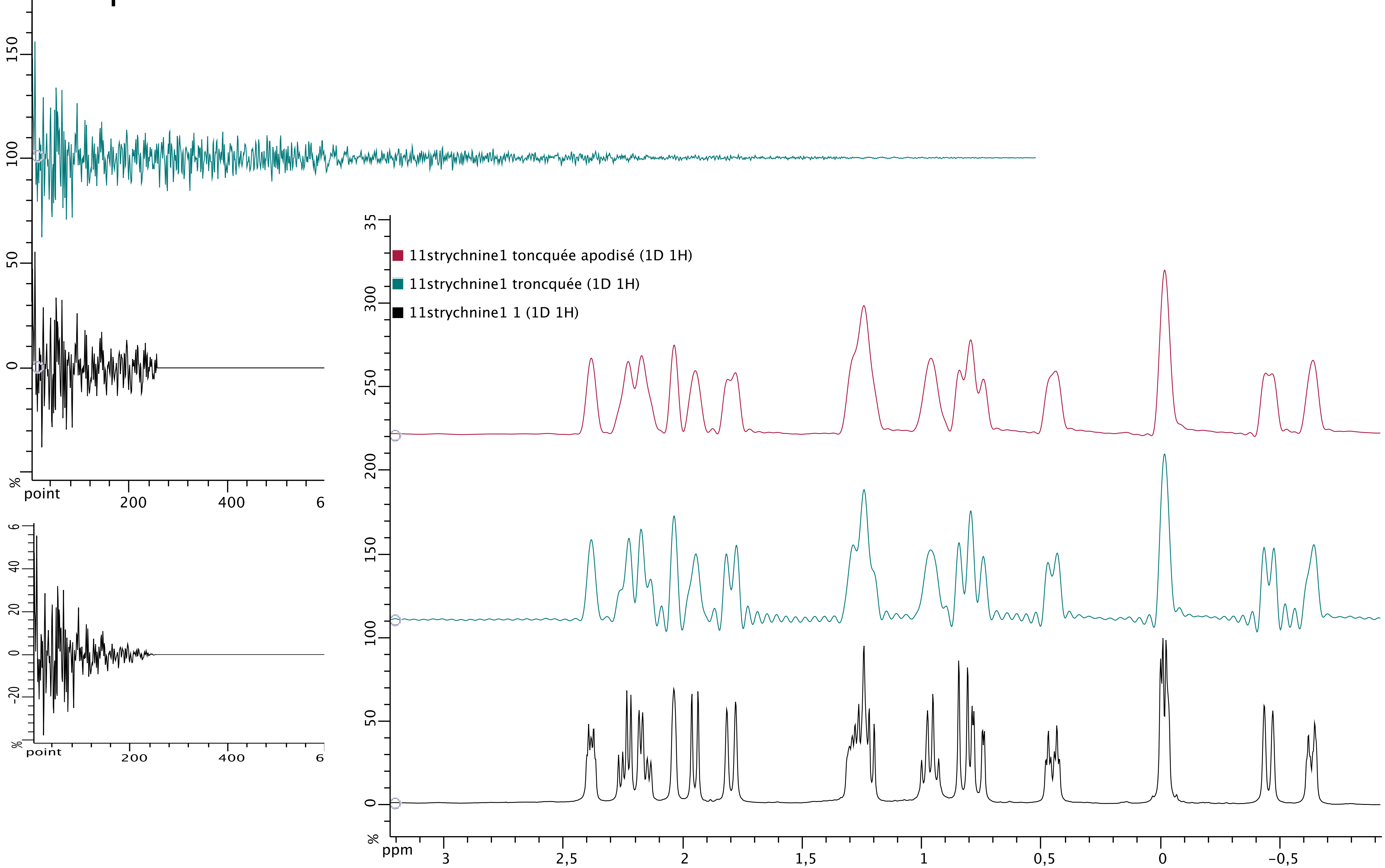
quelques exemples…

spectre RMN 1D de la strychnine



Zero-filling et apodisation

- apodisation arche de sinus



Petite considération sur le zero-filling

- comment ne pas perdre d'information

  1 valeur == 1 élément d'information

  - ne pas faire

FID : N points réels	N valeurs indépendantes
FT(FID) : N/2 points complexes	N valeurs indépendantes
ou N points complexes	2N valeurs non-indépendantes
module : N/2 points réels	N/2 valeurs indépendantes

  - faire

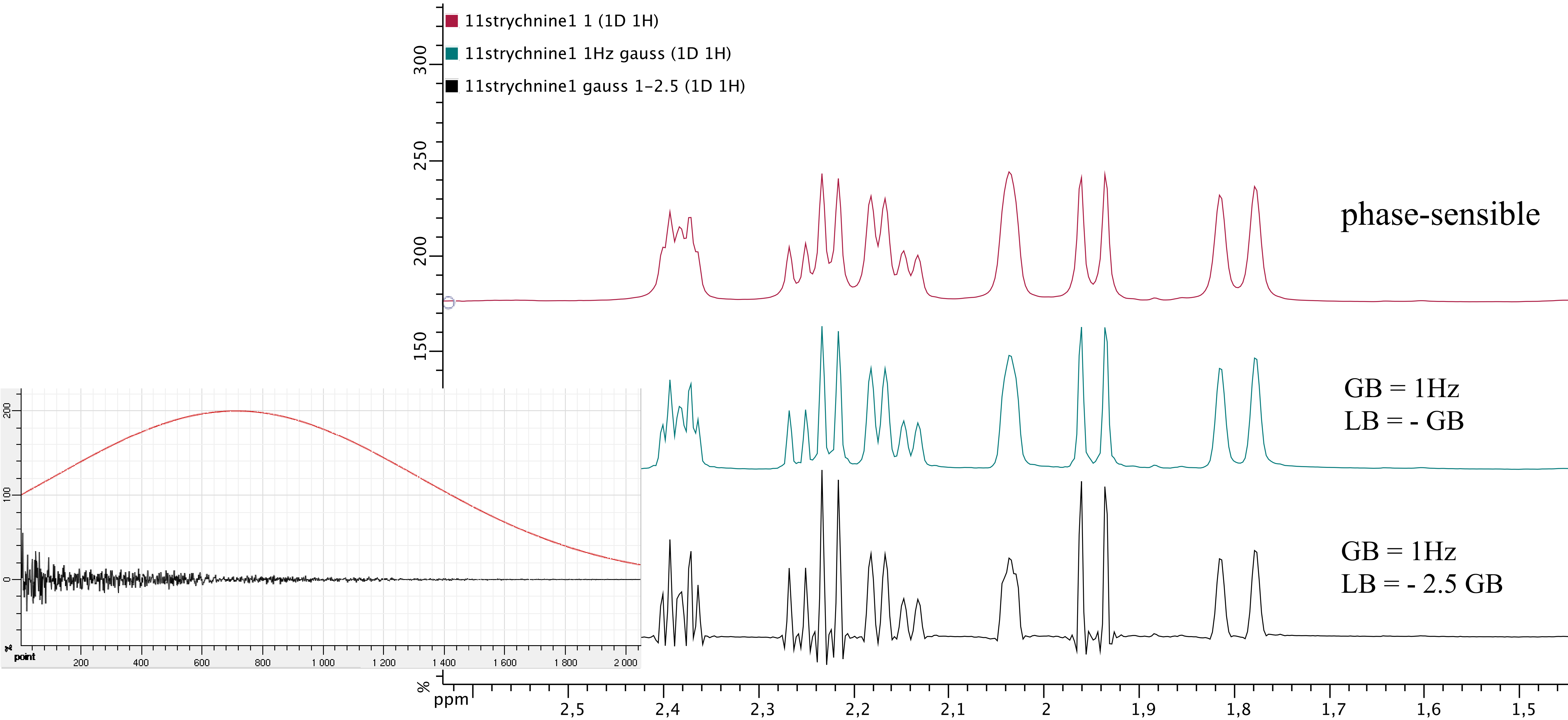
FID : N points réels	N valeurs indépendantes
Zero-Filling + N zéro	N non-nulles
FT(FID) : 2N points complexes	2N valeurs non-indépendantes
module : N points réels	N valeurs indépendantes

rule of thumb :
always zero-fill at least once

autres apodisations

- en mode phase sensible
  - apodisation “gaussienne”
  - déconvolution par une lorentzienne - reconvolution par une gaussienne

$$e^{\frac{t}{\mathrm{LB}}} e^{\frac{-t^2}{\mathrm{GB}}}$$



$N<P$ la régularisation

Moins de points dans les données (N) que dans l'analyse (P) ???

- En fait très courant
  - en FT : «zerofilling» : faire la transformée en rajoutant des zeros
  - en image : interpolation

    nouvelles télés !

- On reconstruit l'image $\tilde{s}$ la plus «jolie» possible

- Il faut un critère de beauté
  - information *a-priori* sur la mesure / sur les données

Position du problème

- modèle de mesure

$$y^{mes} = T_f(s) + \epsilon$$

- fonction cible à minimiser
  - insuffisant pour trouver une solution unique

$$\chi^2 = \sum \left(\frac{T_f(s) - y^{mes}}{\sigma} \right)^2$$

- critère supplémentaire
  - fonction de régulation à optimiser : *R(s)*

- Optimisation
  - miniser χ^2 sous la contrainte de *R(s)* minimum

$$\chi^2(s) + \lambda R(s)$$

Plusieurs possibilités:

- minimiser l'énergie **E** $$E = \sum(s_i^2)$$
- régularisation de Tikhonov **T** $$T = \sum(\Gamma(s)_i^2)$$
  - ($\mathbf{\Gamma}$ opération linéaire) : courbure - smoothness - ...
- maximiser l'entropie **S** $$S = -\sum(s_i \log(s_i))$$
  - le spectre «le plus probable»
  - le spectre qui a le moins d'«information» (sens de Shanon)
- maximiser la simplicité
  - la parcimonie : le nombre de signaux au dessus du bruit.

principe de l'approche inverse

données experimentales

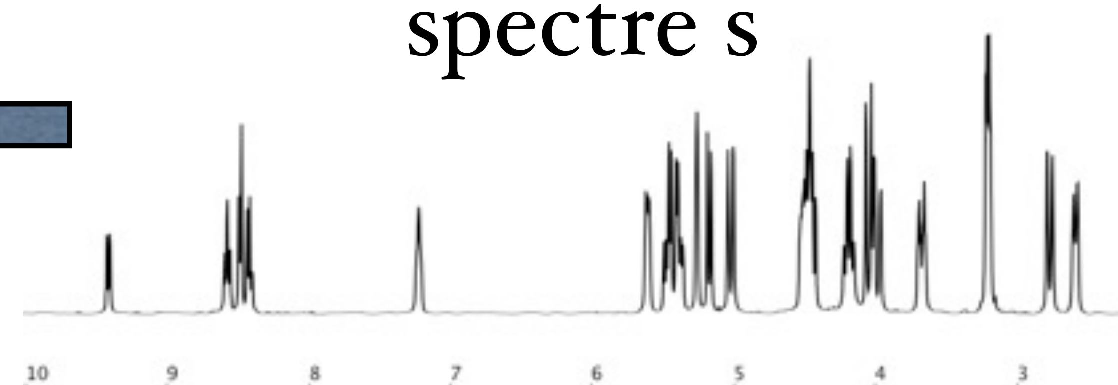
distance

$$\chi^2 = \sum \left(\frac{d_i^r - d_i^e}{\sigma_i}\right)^2$$

Tf

données recalculées

spectre s



Signal Entropy

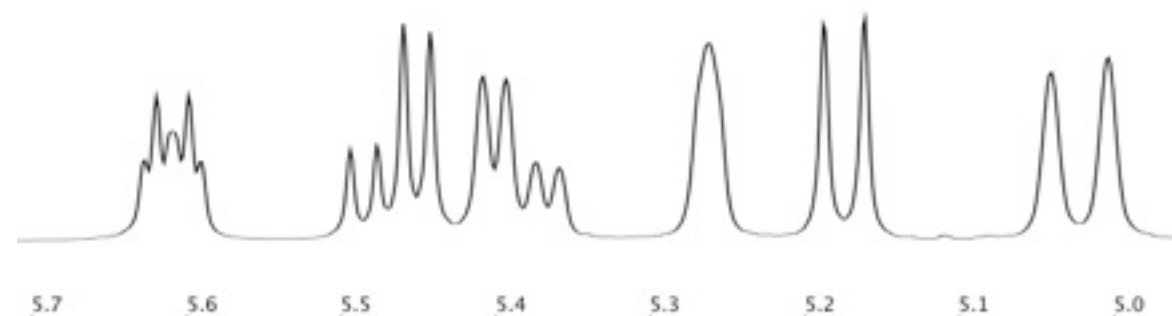
- Among all the spectra which adapt the data down to the noise, I choose :
  - the one with the minimum information (Shannon sense)
  - the most probable
  - => the one which maximize the signal entropy

$$\boldsymbol{S} = -\sum \boldsymbol{p_i} \log(\boldsymbol{p_i}) \quad \text{with} \qquad \boldsymbol{p_i} = \frac{\boldsymbol{f_i}}{\sum \boldsymbol{f_i}}$$

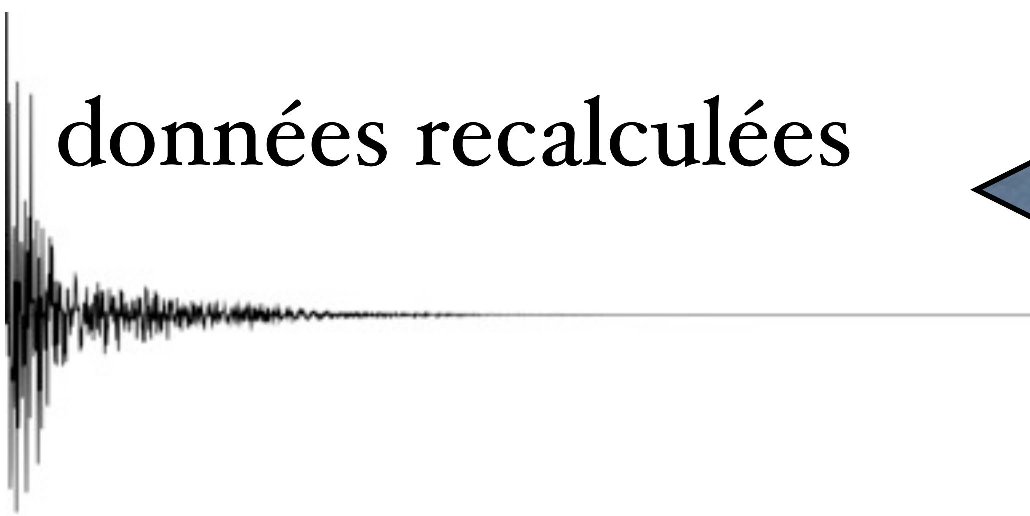
extension par approche inverse

données expérimentales tronquées

Spectre MaxEnt



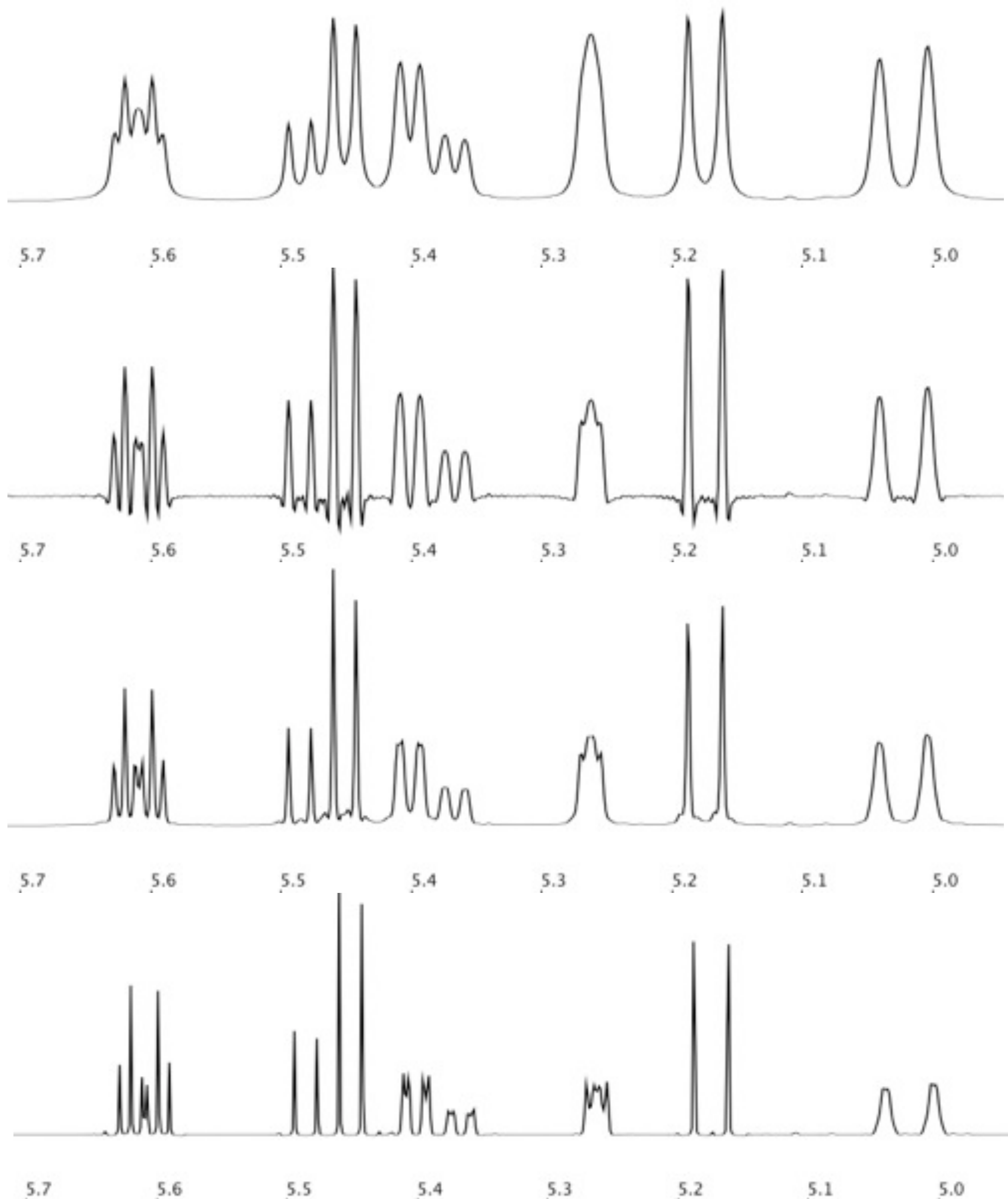
Tf



données recalculées

χ^2 calculé sur les données mesurées

quelques exemples...

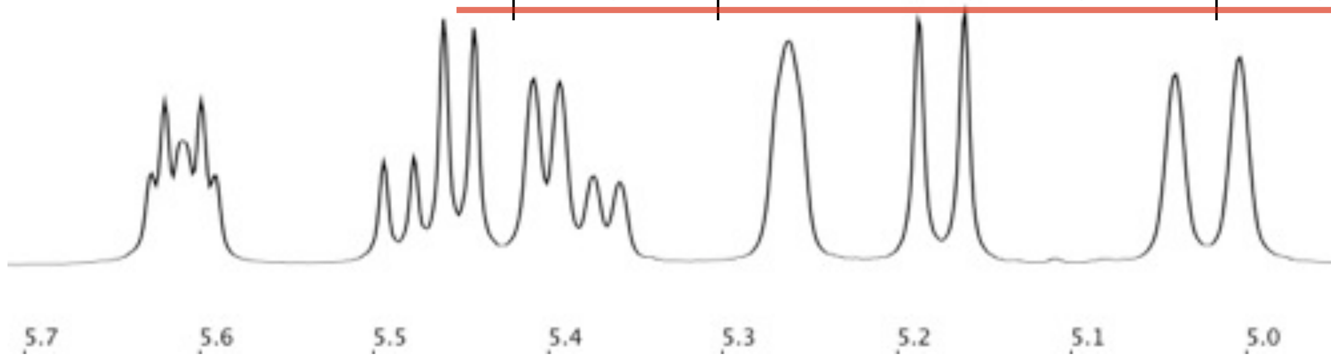


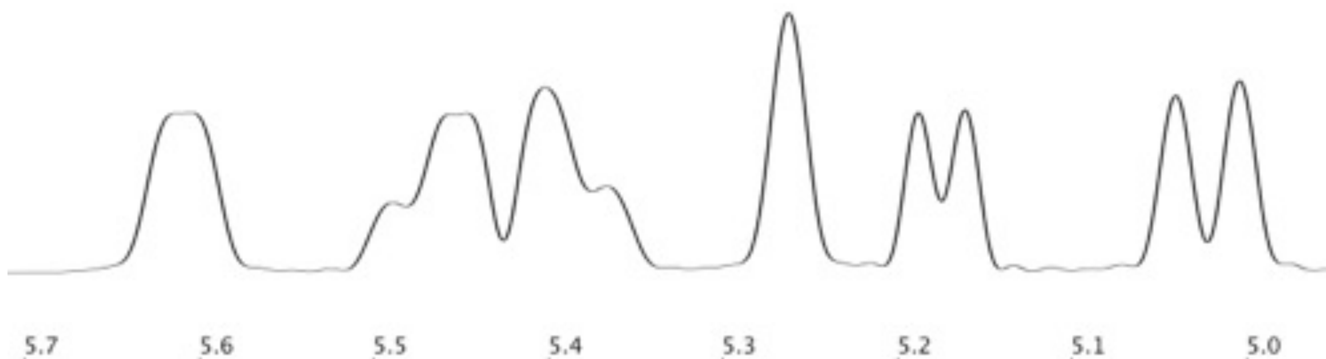
Gauss 2 Hz

MaxEnt 1 Hz

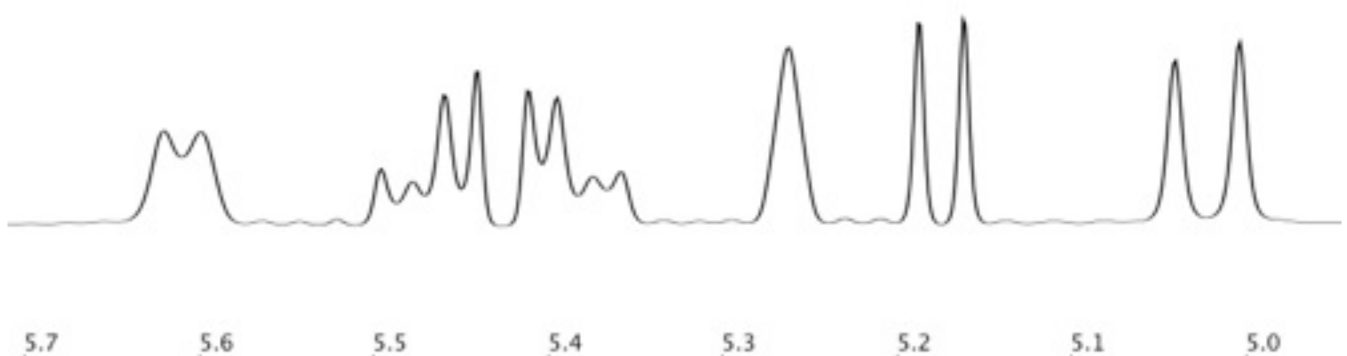
MaxEnt 2 Hz

quelques exemples…

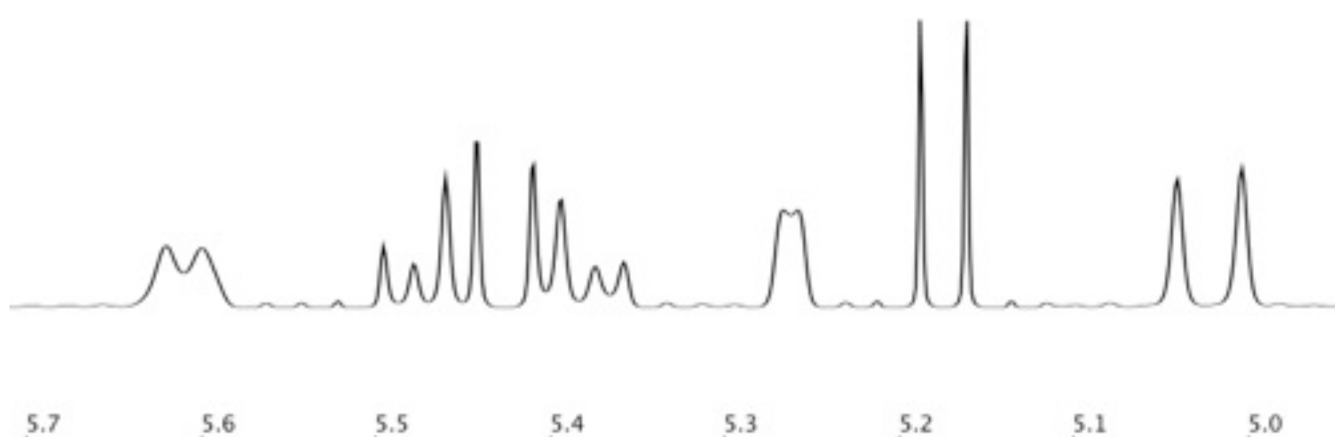




FID trunc. to 1k

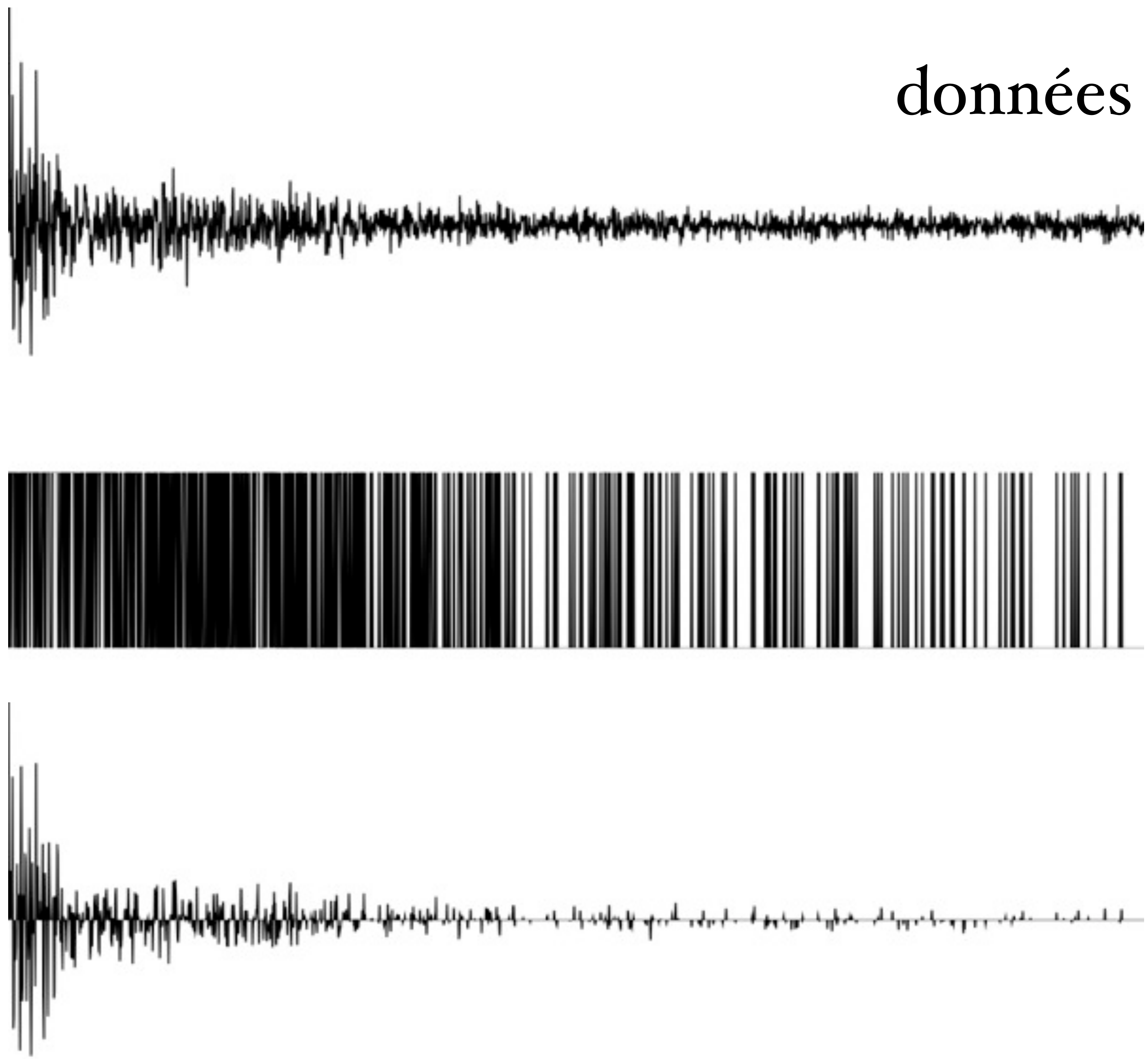


MaxEnt 1 Hz



MaxEnt 2 Hz

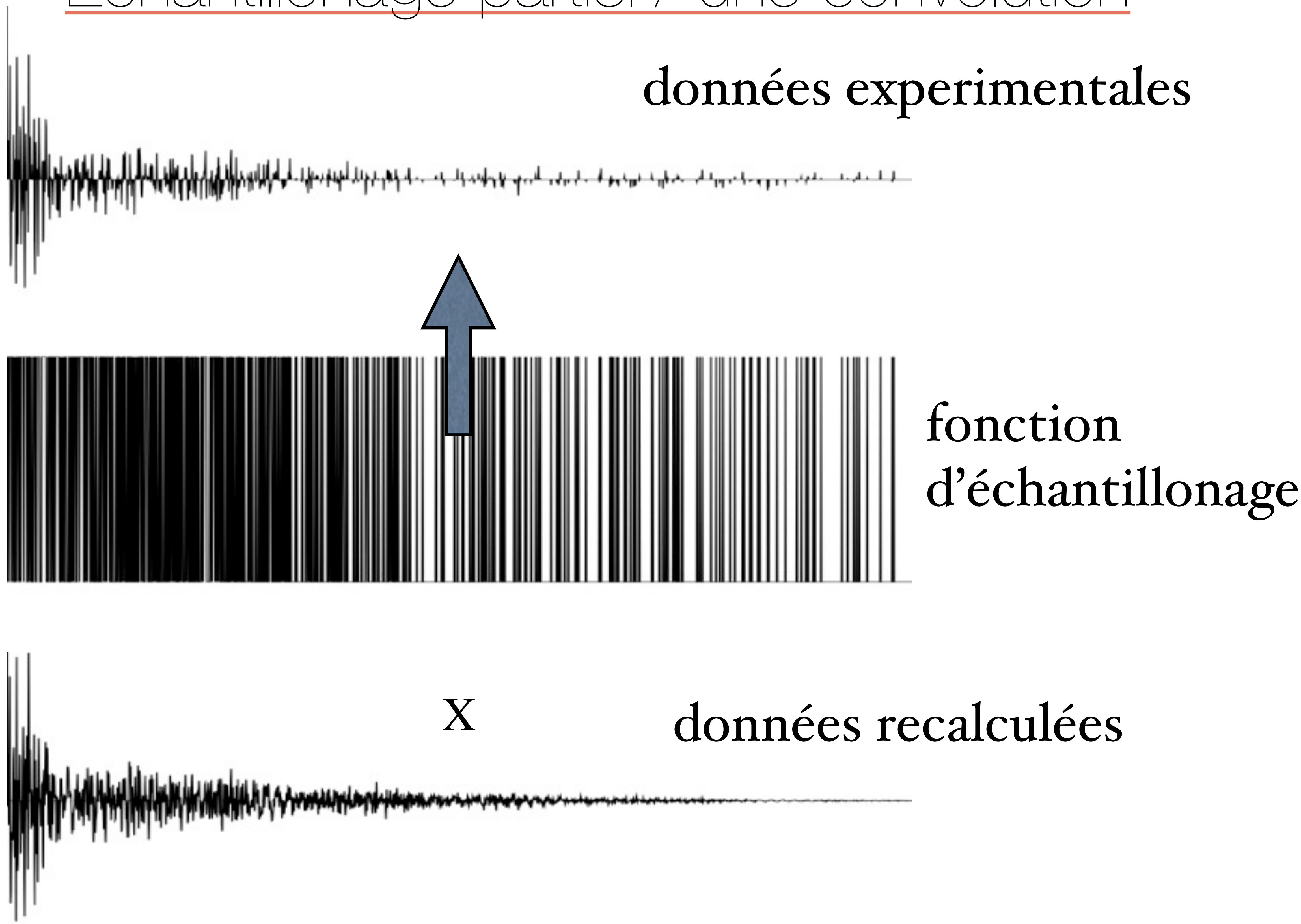
Échantillonage partiel vu comme une convolution



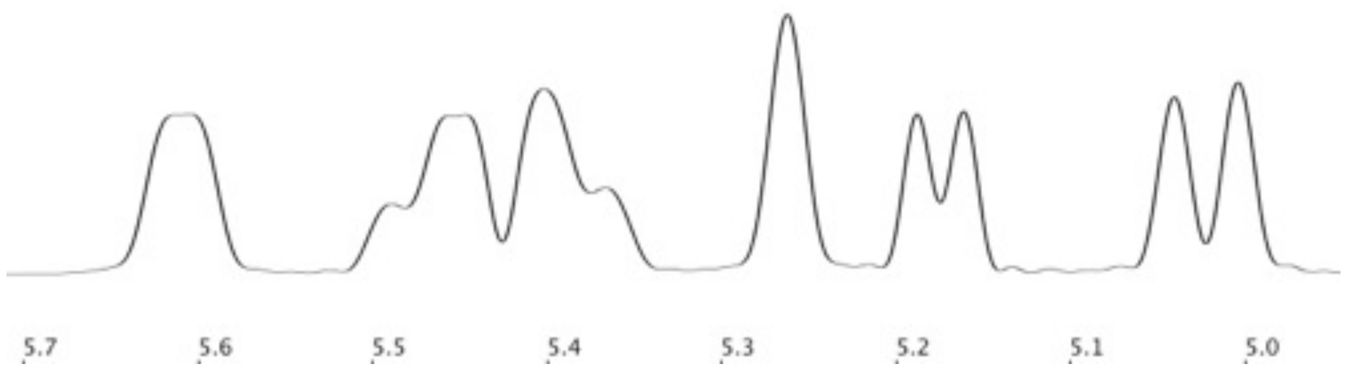
données classiques

fonction d'échantillonage

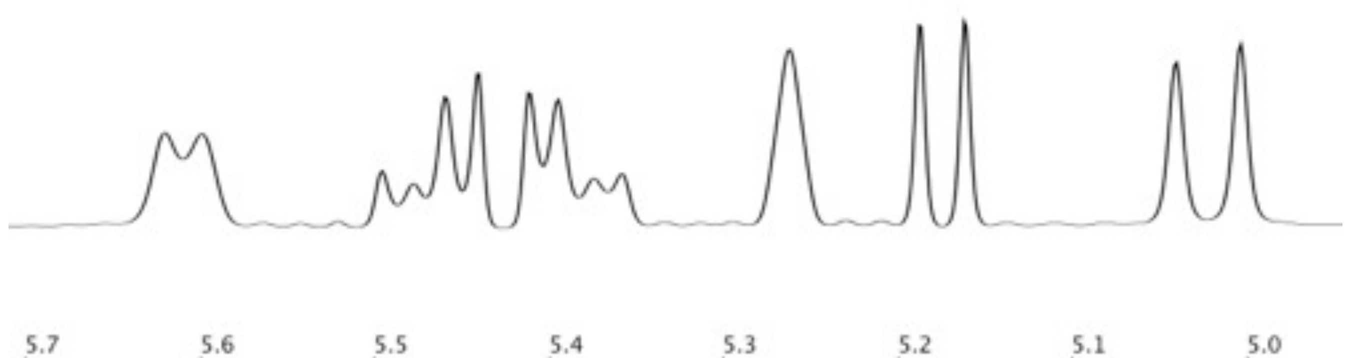
Échantillonage partiel / une convolution



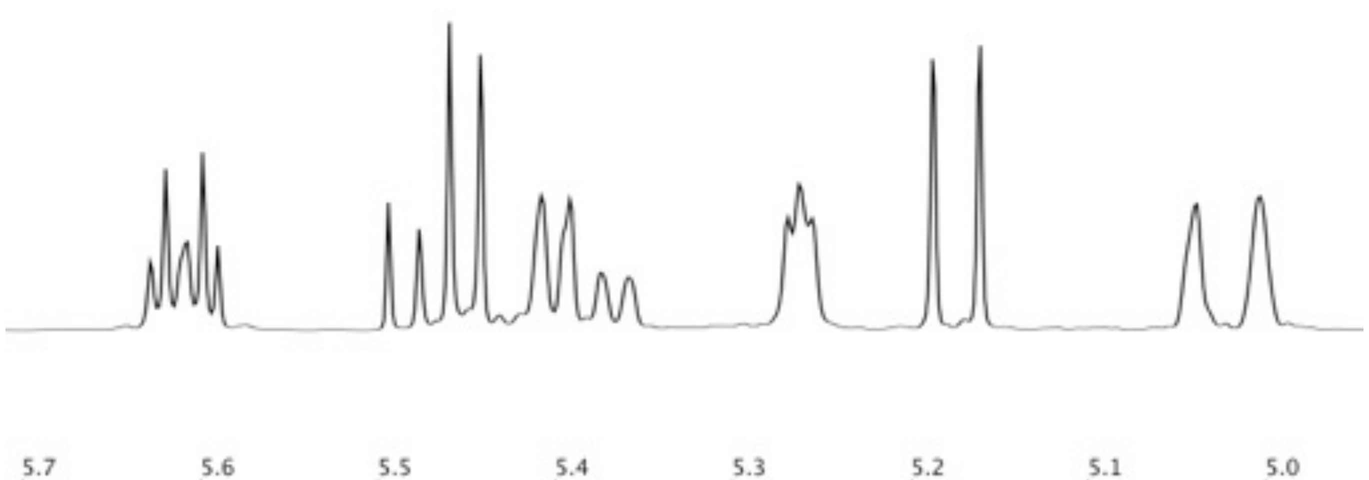
strychnine *encore* ...



FT
FID trunc. to 1k



MaxEnt - 1Hz
FID trunc. to 1k



MaxEnt - 1Hz
FID partially sampled
1k out of 4k

An automated tool for maximum entropy reconstruction of biomolecular NMR spectra

NATURE METHODS | VOL.4 NO.6 | JUNE 2007 | 467

Mehdi Mobli[1–3], Mark W Maciejewski[1,3], Michael R Gryk[1] & Jeffrey C Hoch[1]

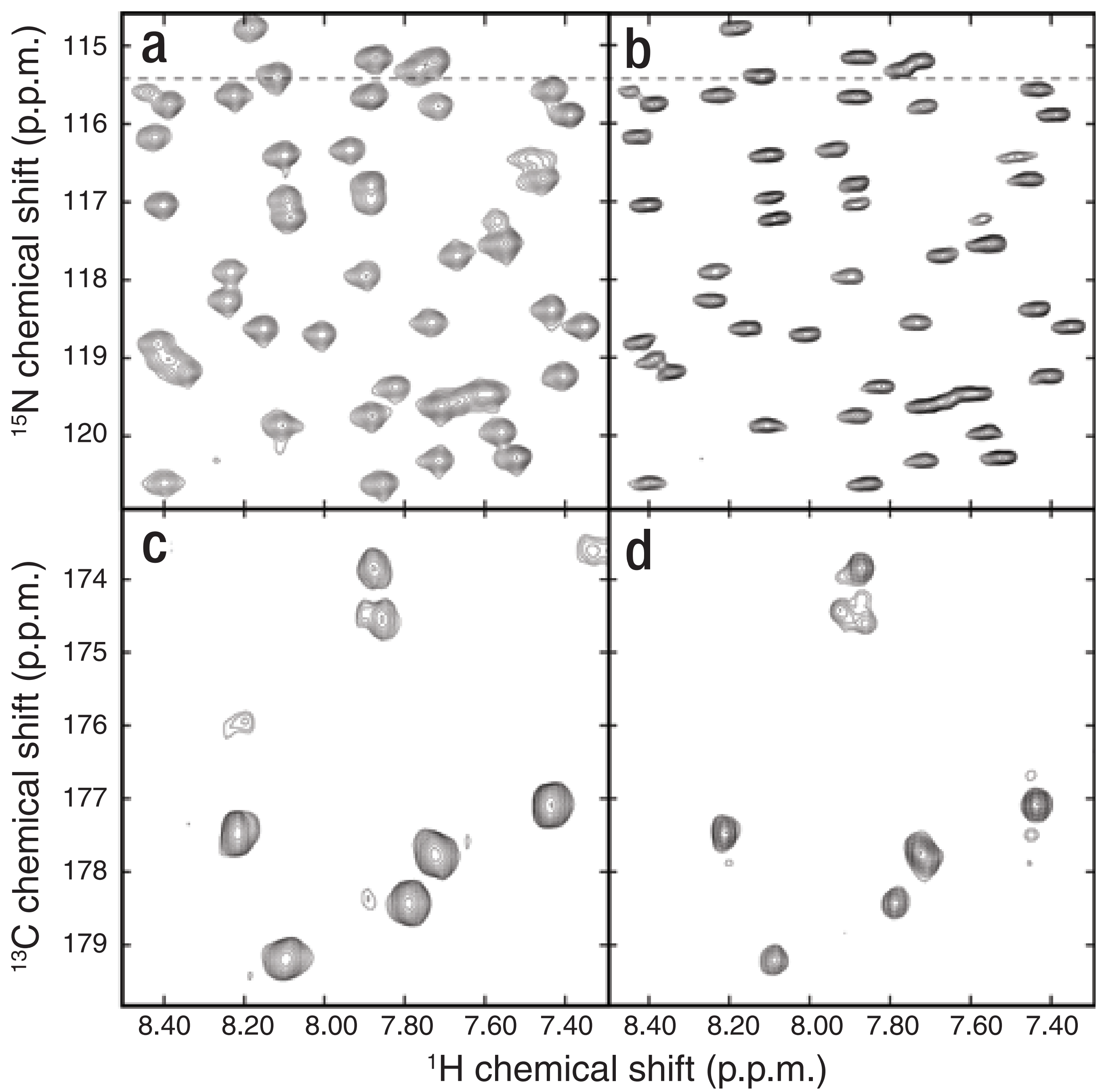


Figure 1 | Automated MaxEnt reconstruction can dramatically reduce data collection time or improve resolution. (**a**–**d**) ^{15}N heteronuclear single quantum correlation (HSQC; **a**,**b**) and HNCO (**c**,**d**) spectra are shown for DNA polymerase X obtained using conventional processing (linear-prediction extrapolation and sinebell apodization; **a**,**c**), and automated MaxEnt reconstruction using linewidth deconvolution to improve resolution without the sensitivity losses characteristic of apodization (**b**), and nonuniform sampling to achieve a sevenfold reduction in data acquisition time (**d**). Two-dimensional cross-sections of the three-dimensional spectrum in **c** and **d** correspond to the ^{15}N frequency indicated by the dashed lines in **a** and **b**.

Compressed Sensing

- Maximiser la “parcimonie” du signal
  - la simplicité
  - minimiser la somme de *abs*(x_i)
  - *K* : nombre de signaux non nul

$$K << N < P$$

- Nécessite certaine conditions
  - sur la fonction de transfer

    inversible

    RIP

- Alors la reconstruction peut-être exacte (si pas de bruit)
  - papier E.Candès et T.Tao 2006
  - example
  - SL0

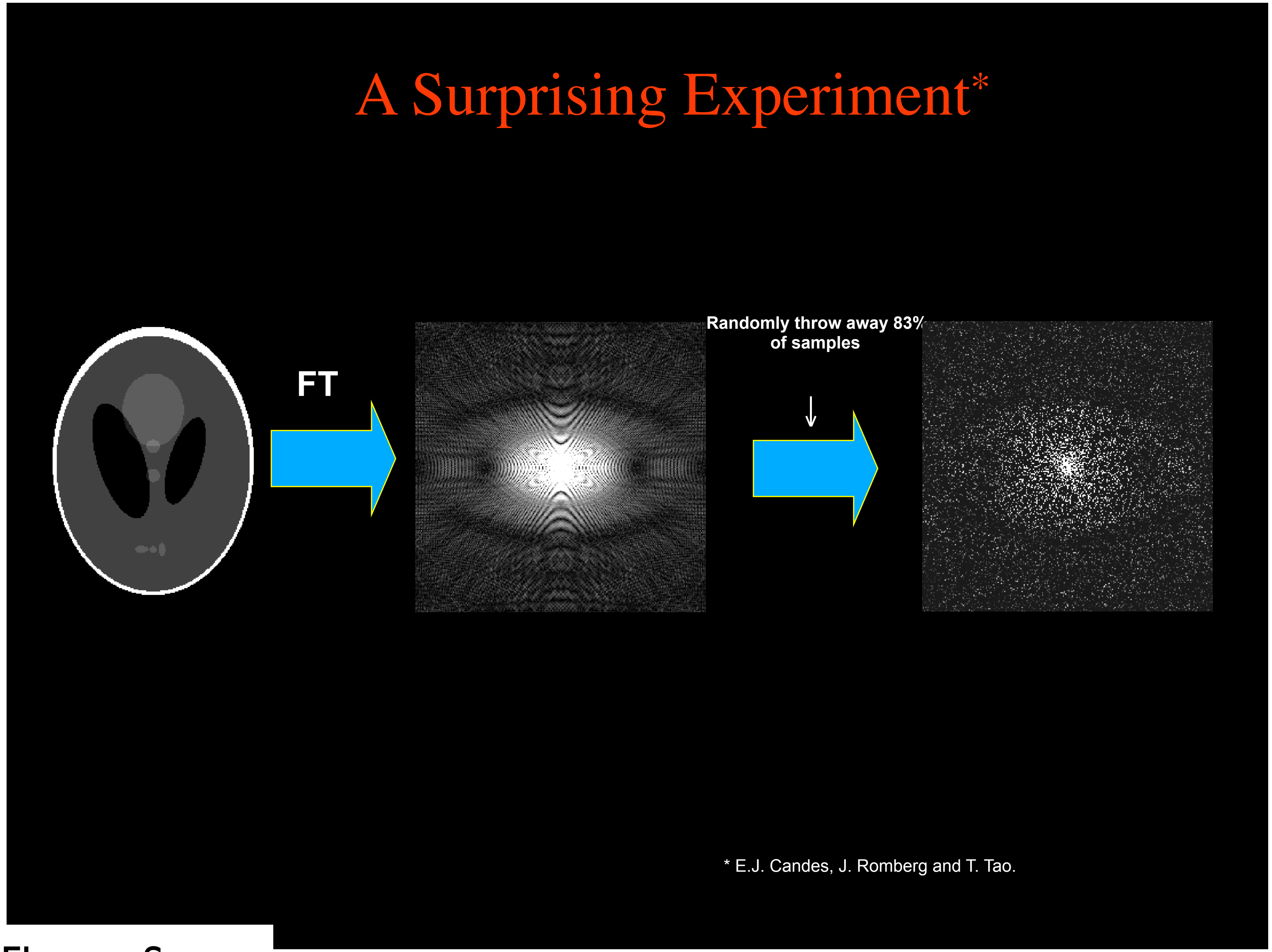
E.Candès T.Tao 2006

Robust Uncertainty Principles:
Exact Signal Reconstruction from Highly Incomplete
Frequency Information

Emmanuel Candes†, Justin Romberg†, and Terence Tao♯

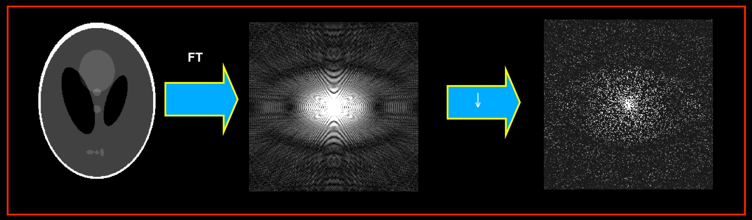
† Applied and Computational Mathematics, Caltech, Pasadena, CA 91125
♯ Department of Mathematics, University of California, Los Angeles, CA 90095

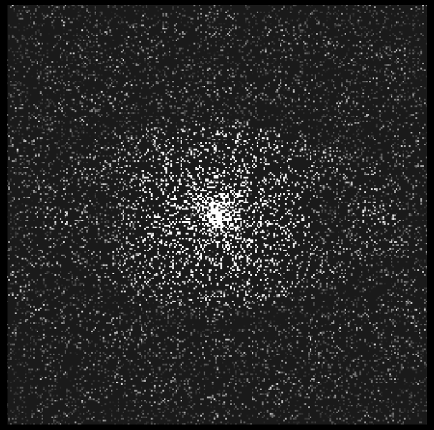
E. J. Candès, J. Romberg, and T. Tao, *Robust uncertainty principles: Exact signal reconstruction from highly incomplete frequency information*, IEEE Trans. on Information Theory, 52 (2006), pp. 489–509.



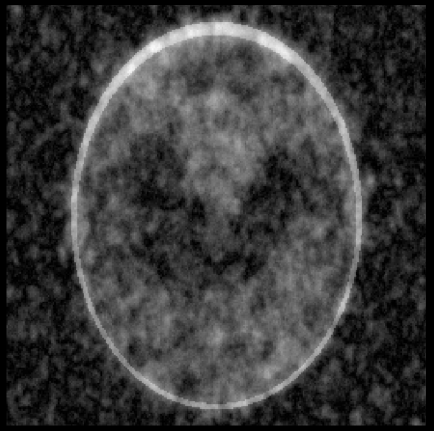
Jean-Luc Starck, Florent Sureau
J. Bobin, N. Barbey, A. Woiselle

A Surprising Result*



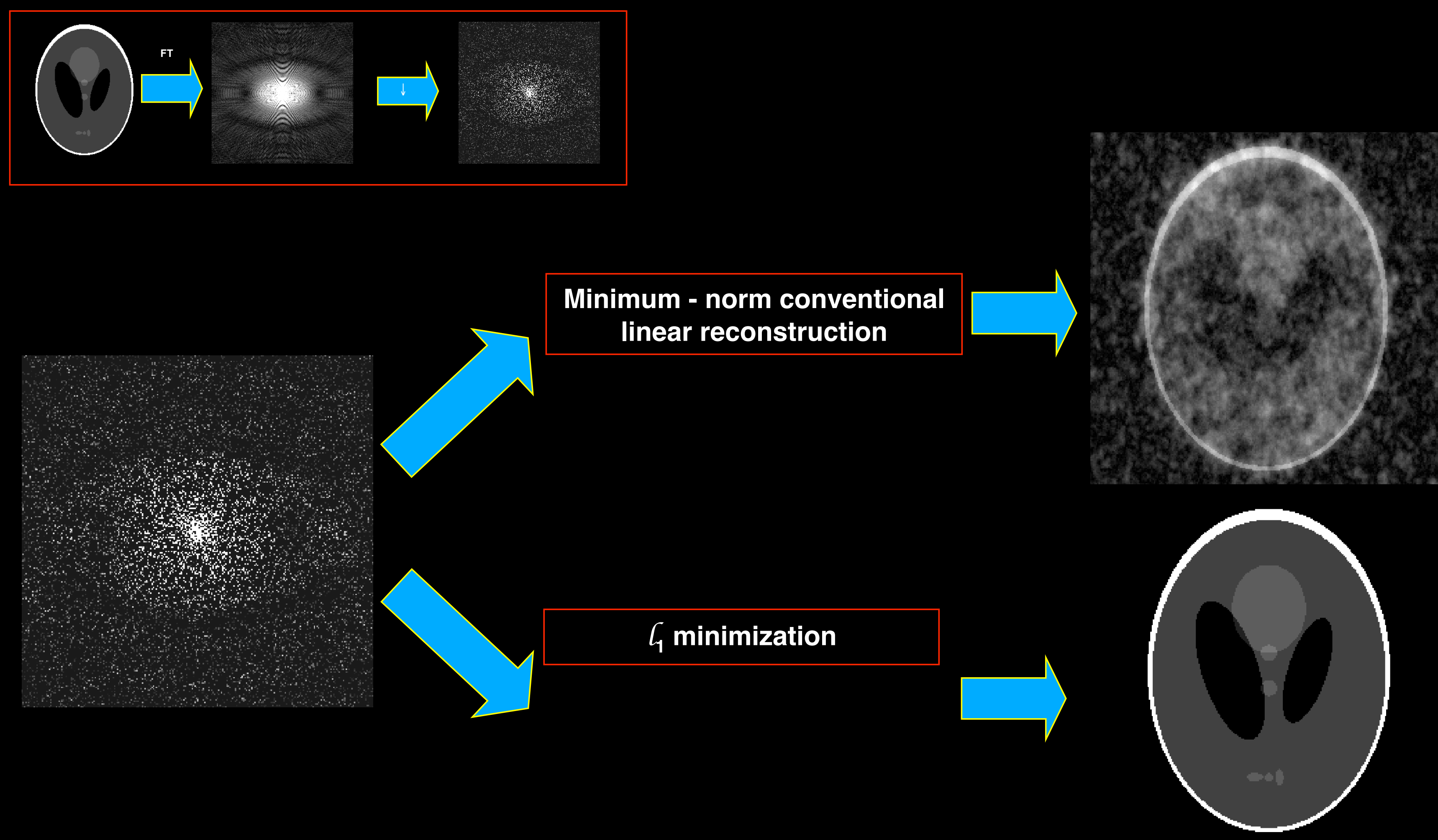


Minimum - norm conventional linear reconstruction



* E.J. Candes, J. Romberg and T. Tao.

A Surprising Result*



E.J. Candes

Compressed Sensing

- régulation : données simples/creuses - *sparsity* -
  - *K signaux - N mesures - P points dans le spectres*
  - hypothèse de peu de signaux :   K<<N<<P

- Le problème devient une simple optimisation convexe

$$\min(\|s\|_1) \quad \text{avec} \quad \|y - Ts\|_2 < \epsilon$$

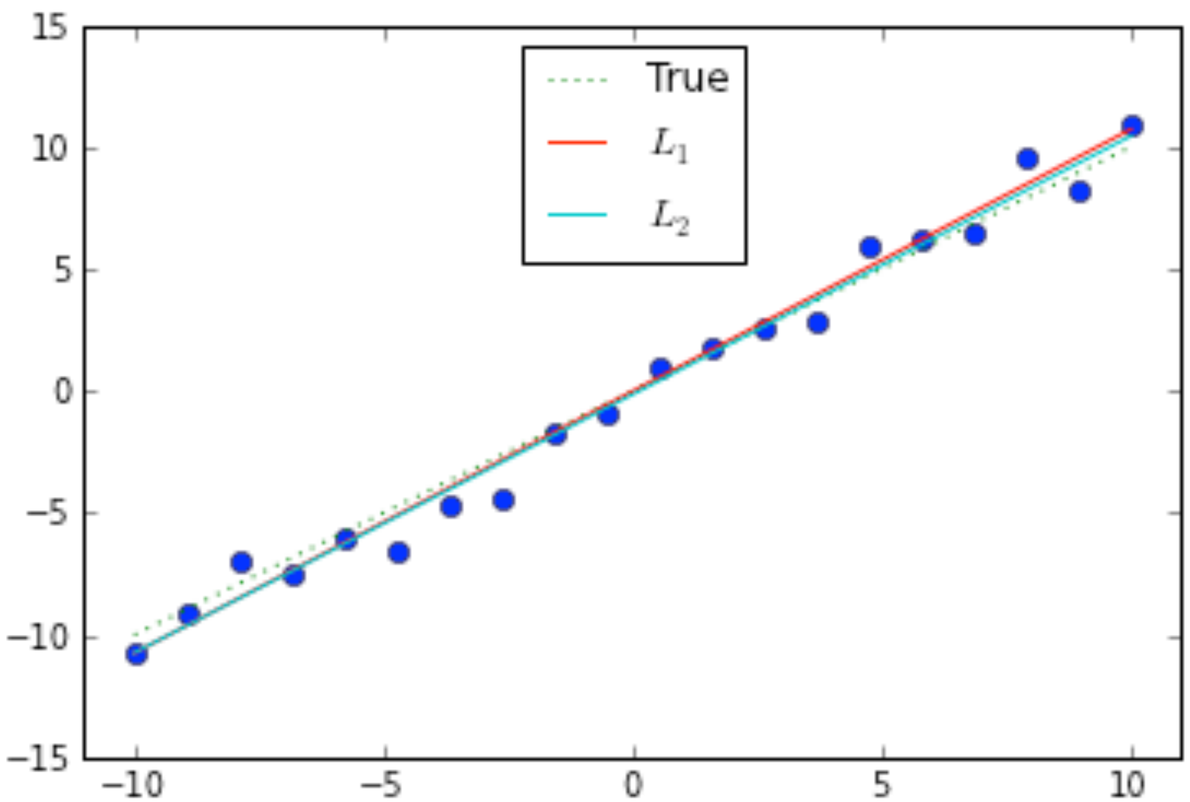
  - norme l_1 ou l_O mais pas l_2

$$\|s\|_1 = \sum |s_i| \qquad \|s\|_p = \left(\sum s^p\right)^{\frac{1}{p}}$$

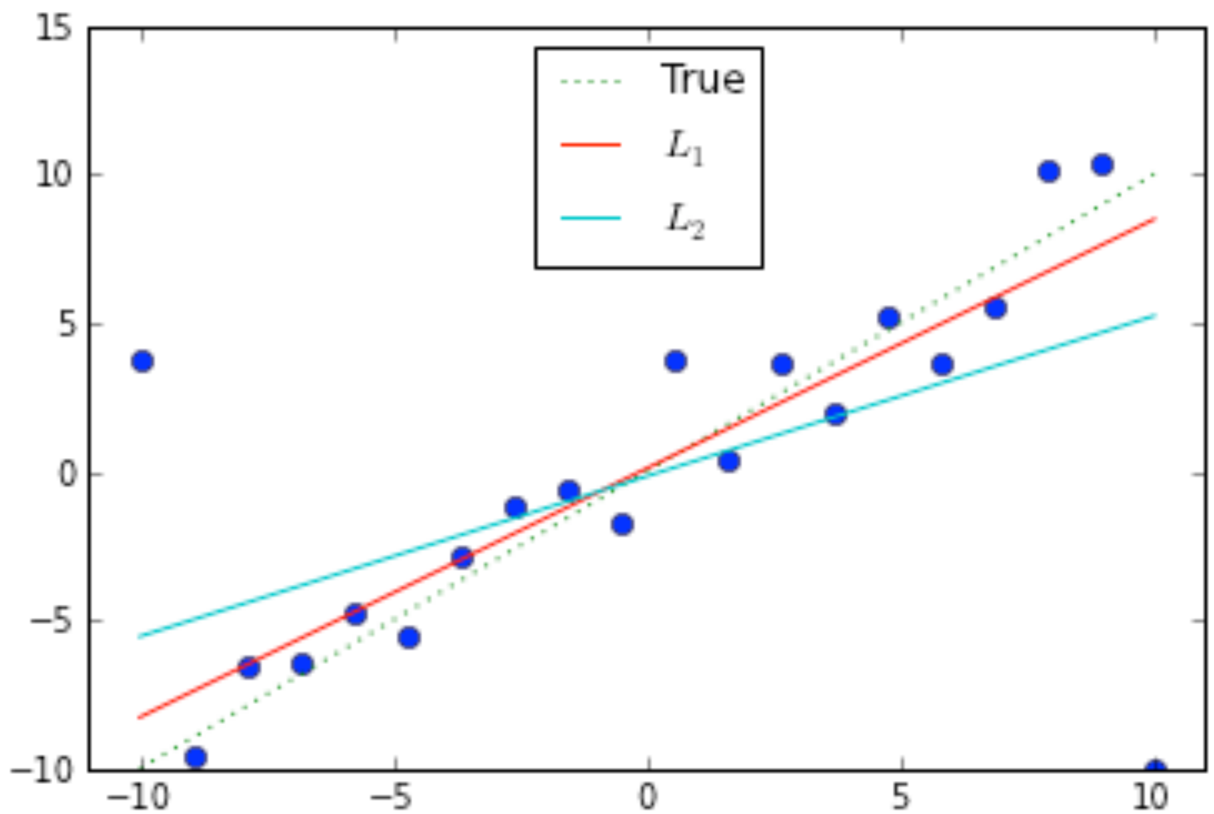
- Dans certaines conditions
  - R.I.P.   *Restricted Isometry Property*
  - dispersion   (FT par exemple)
  - pseudo-inversible
  - linéaire

En exemple LIVE

Comparaison minimisation L_1 vs L_2

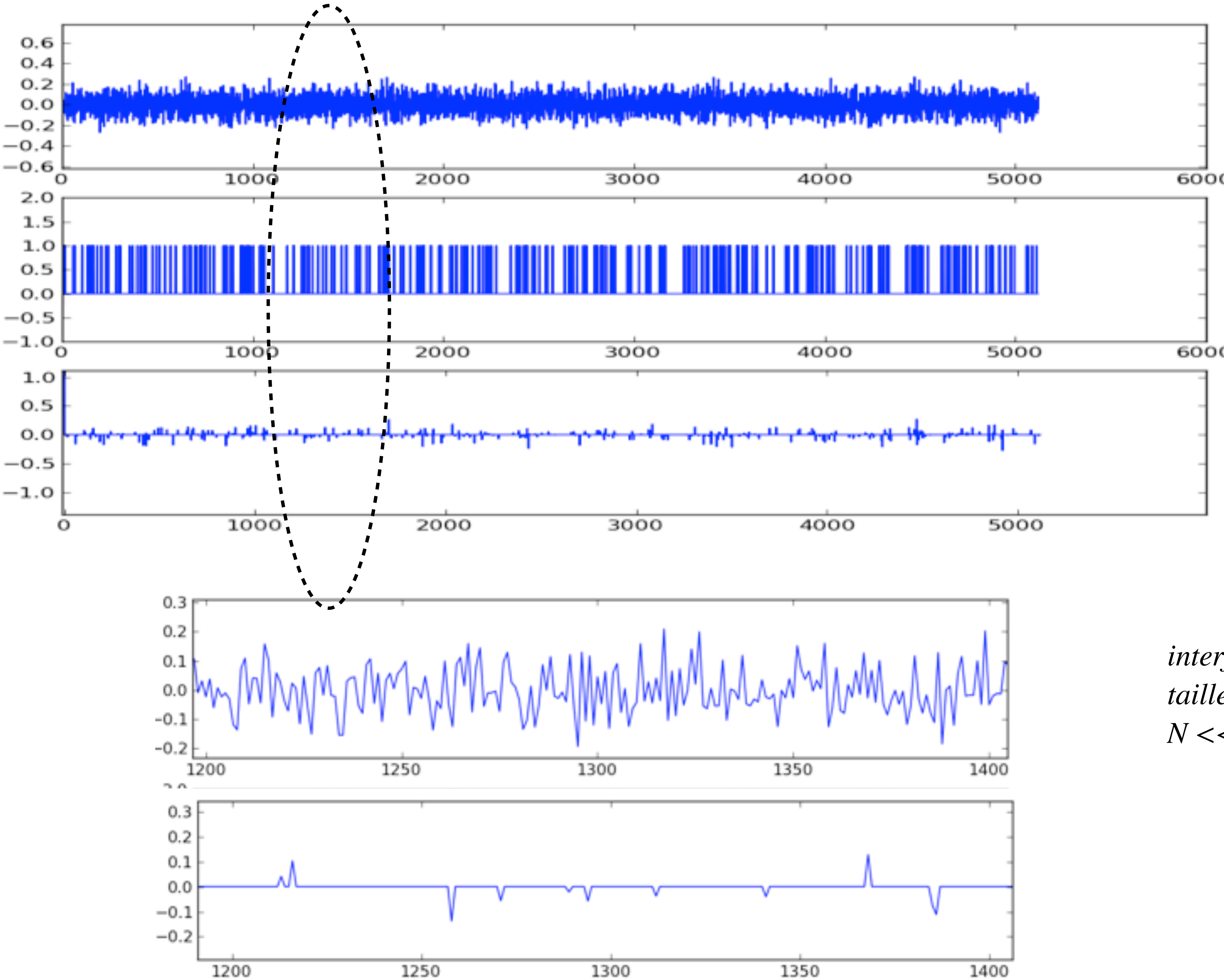


avec *bruit* gaussien



avec *bruit* gaussien
+ points abérants

échantillonnage extrême

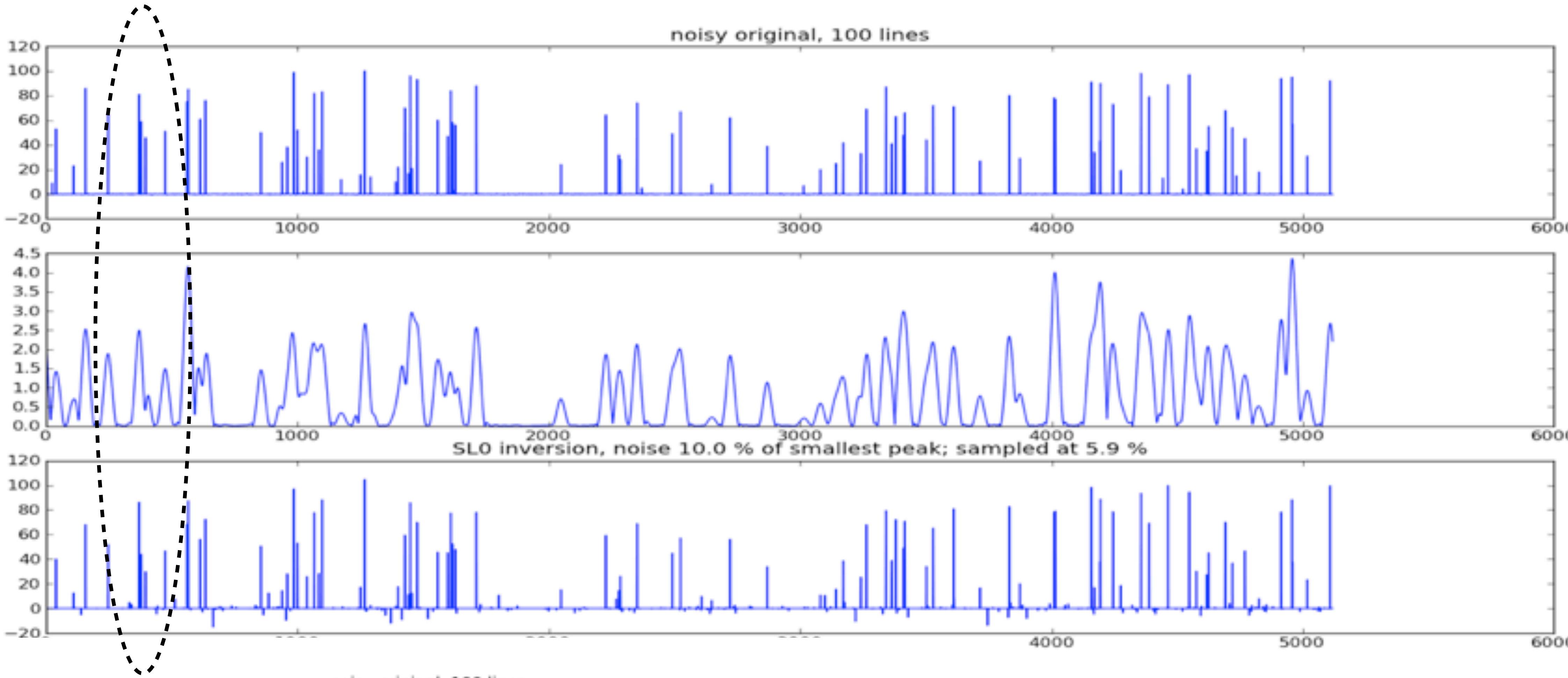


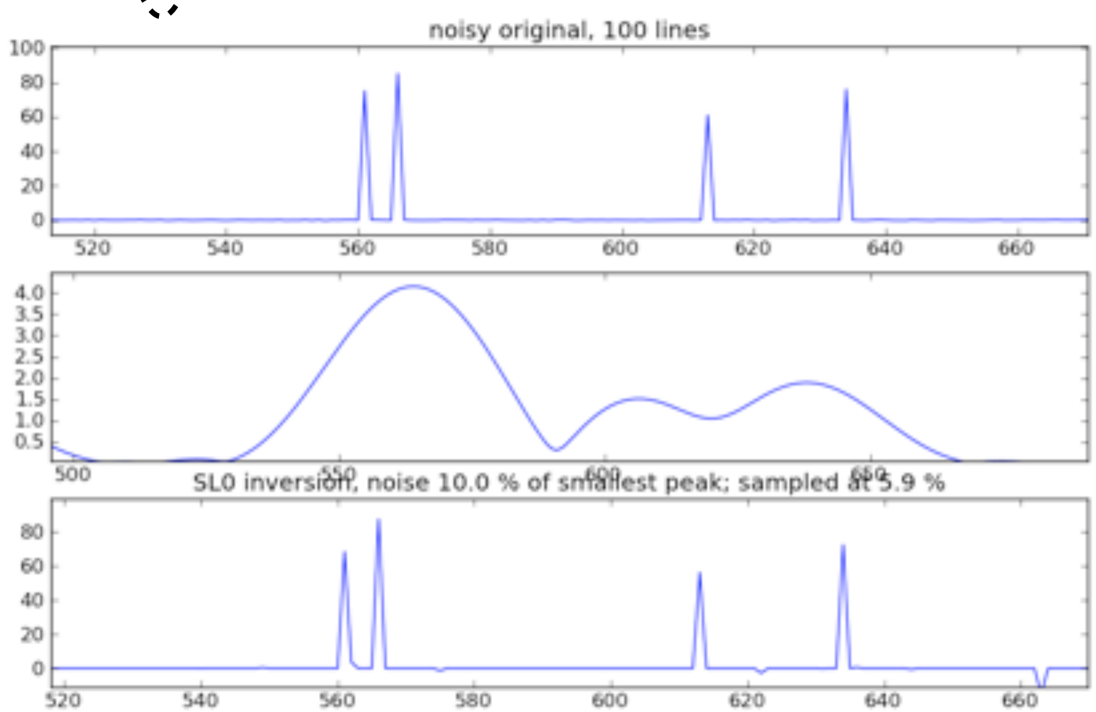
interferogramme $P = 5000$

taille de la mesure $N=300$

$N << P$

Un exemple



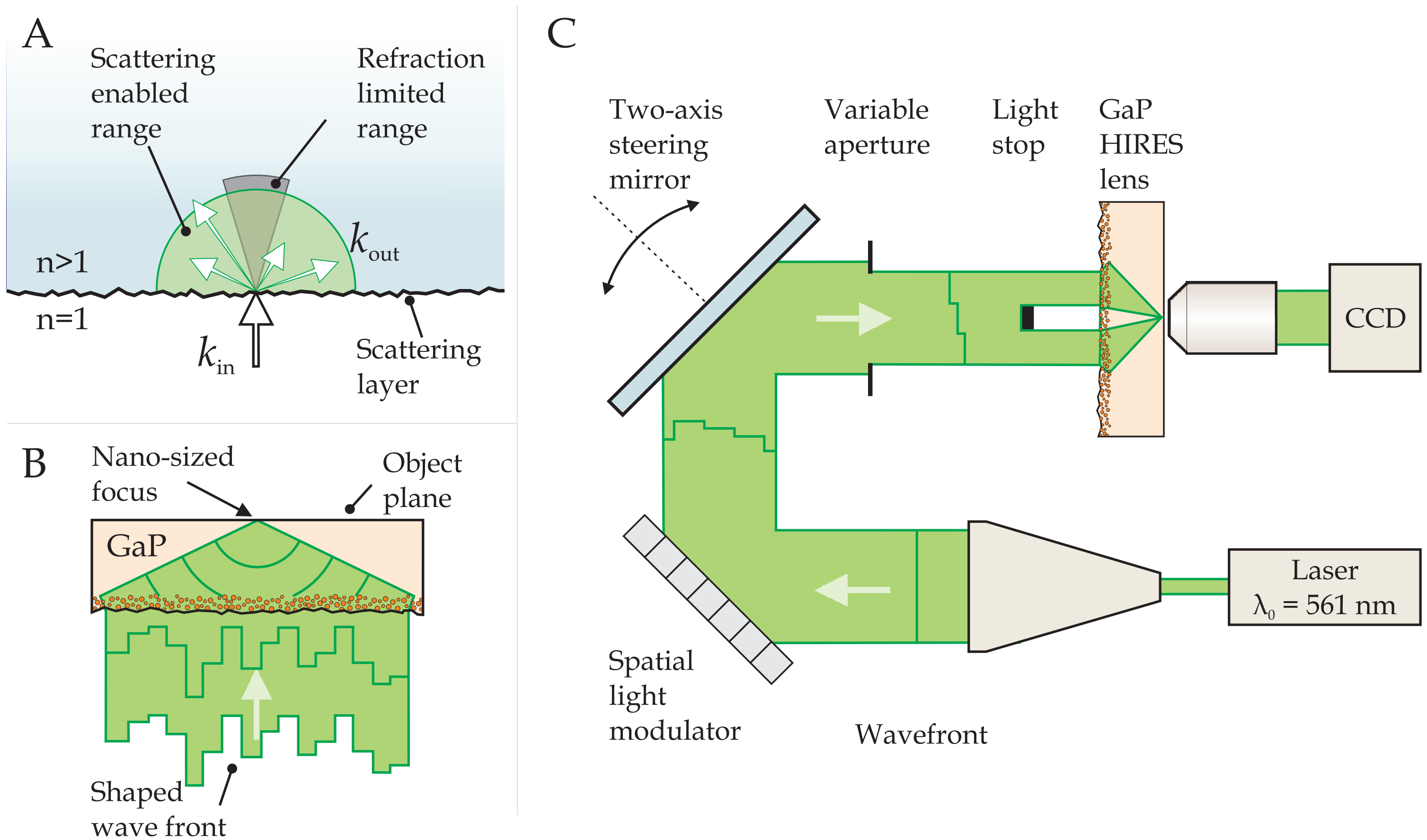


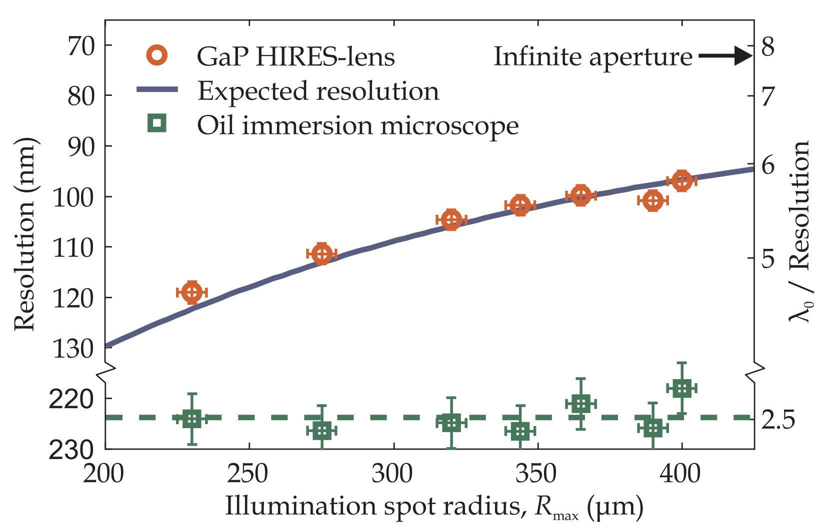
échantillonage d'un interférogramme très lacunaire
reconstruction de données par algorithme SL0
G Mohimani, M Babaie-Zadeh, C Jutten
ICASSP 2008. IEEE International Conference on (2008) pp. 3881-3884

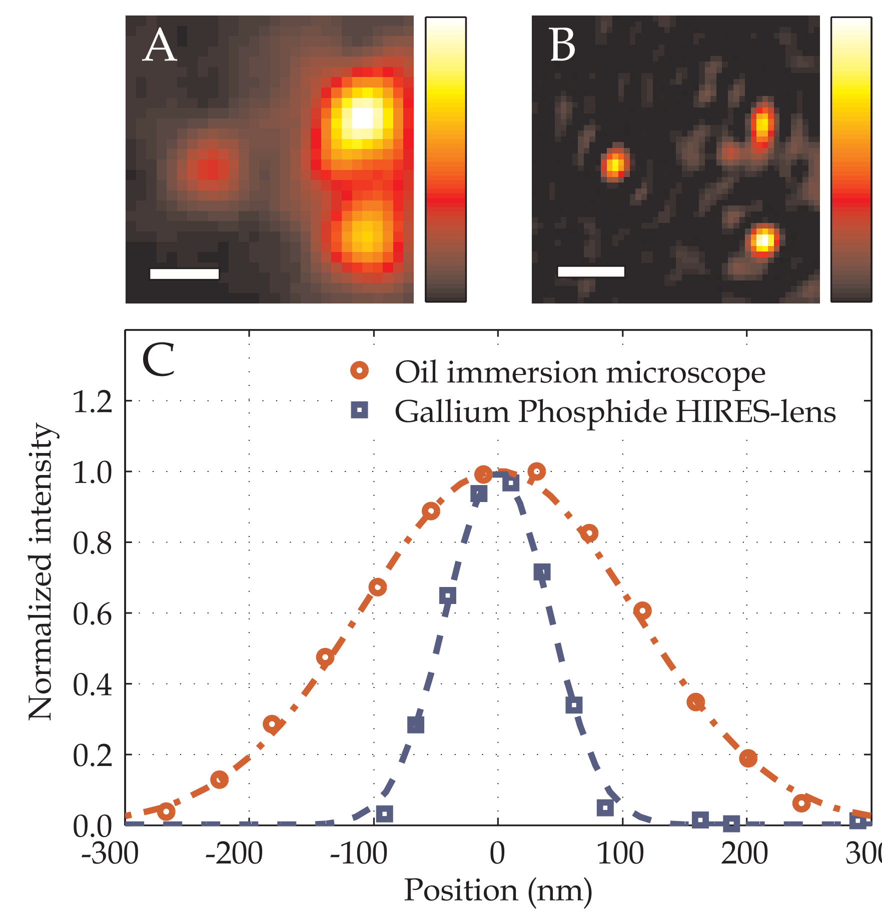
taille du spectre $P = 5000$
taille de la mesure $N=300$
nombre de signaux $K=100$

$K < N << P$

C'est assez fort !







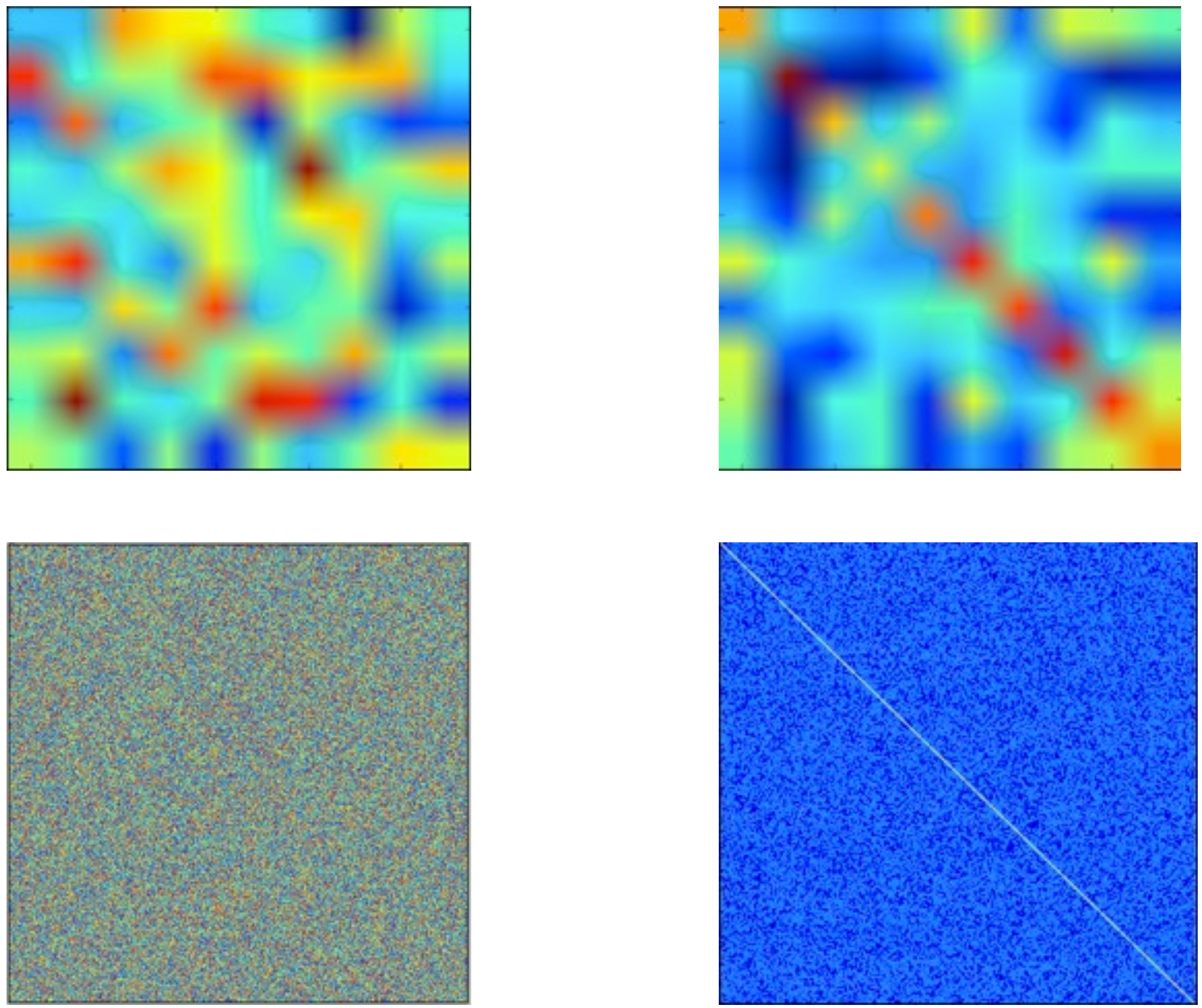
Putten et al. Scattering Lens Resolves sub-100 nm Structures with Visible Light. arXiv (2011) vol. physics.optics

matrices aléatoires

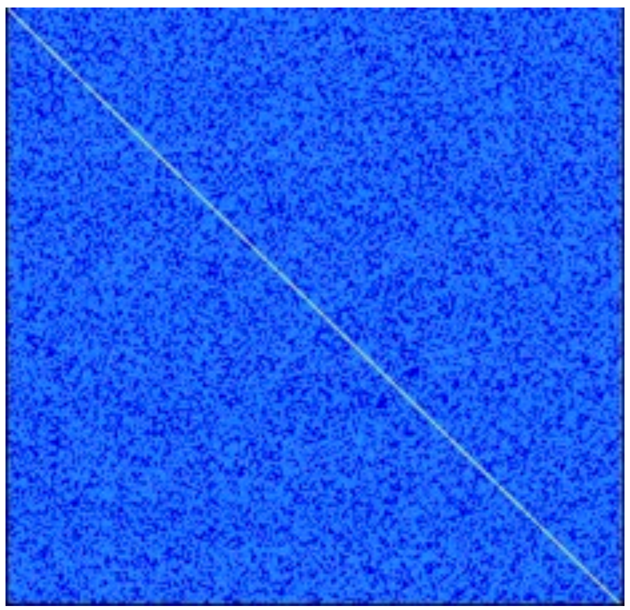
A $\quad$ $A\,{}^tA$

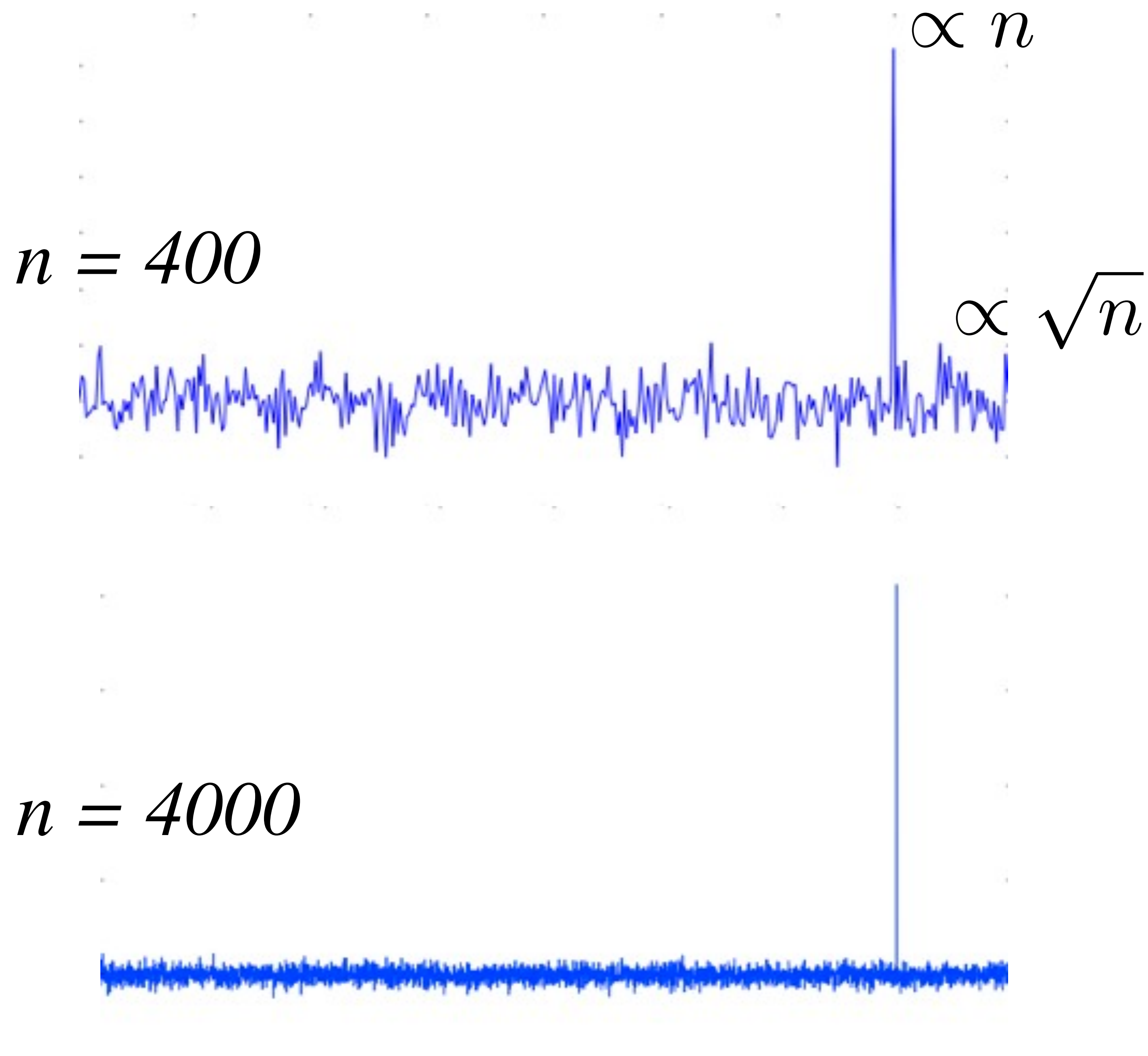
$n = 10$

$n = 400$

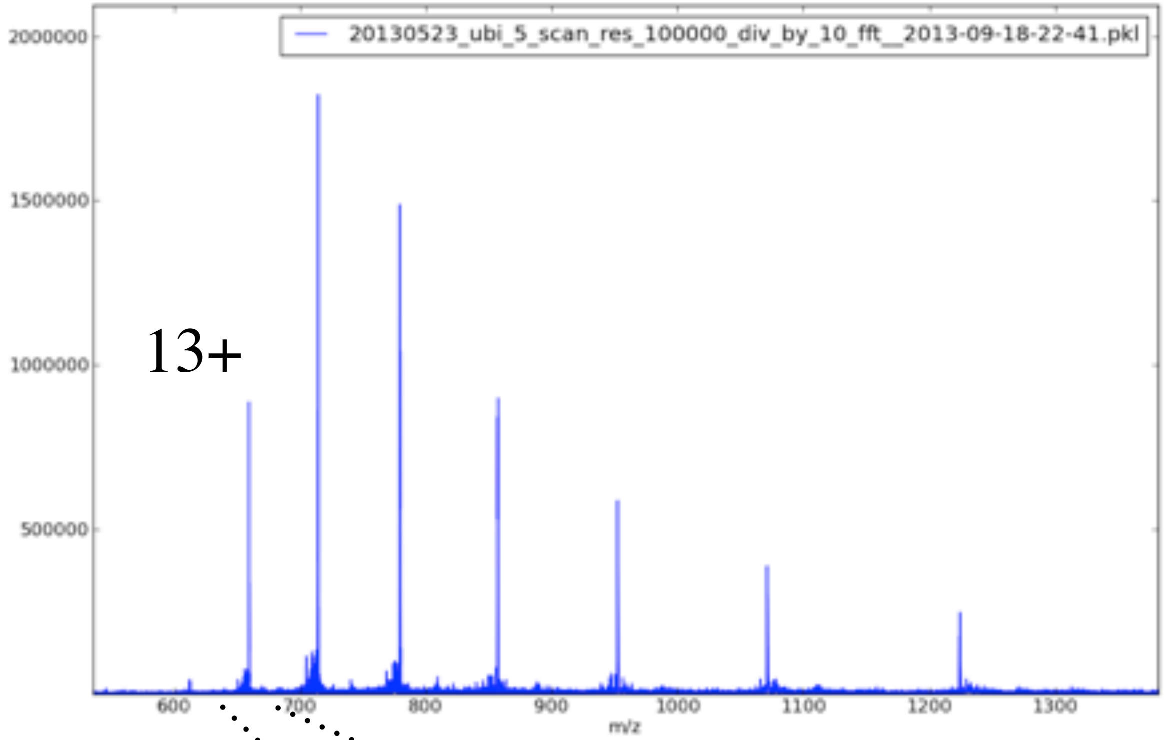


grandes matrices aléatoires



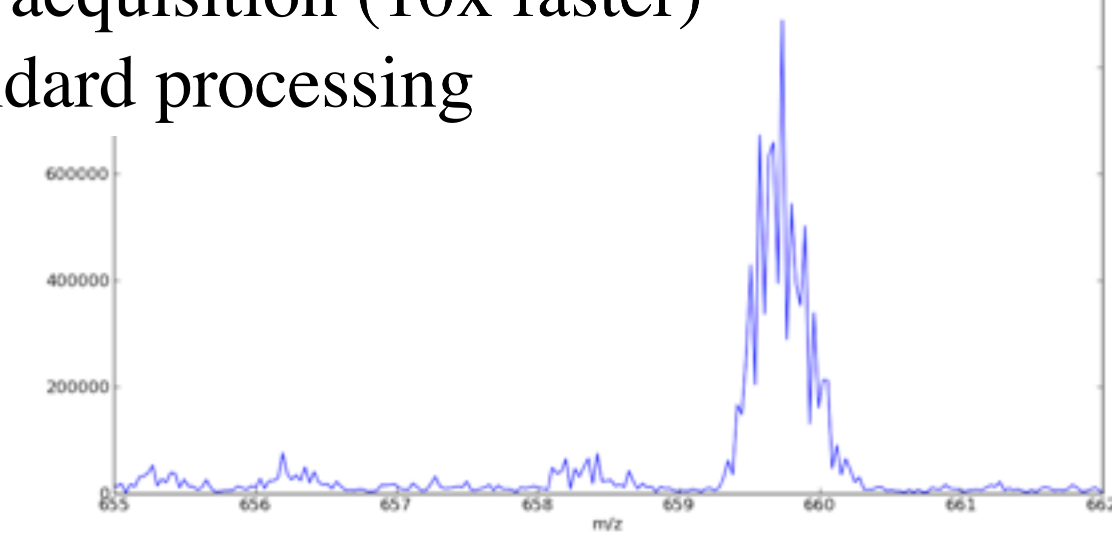


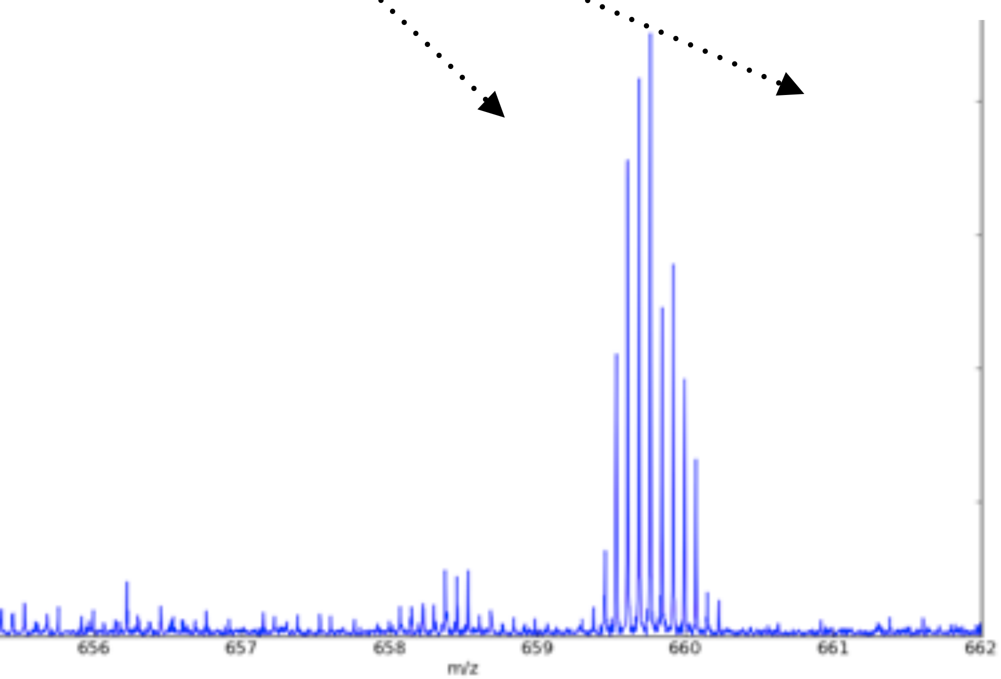
Super-resolution / fast acquisition



Orbitrap Ubiquitin spectrum
(*col. J. Chamot-Rooke Institut Pasteur*)

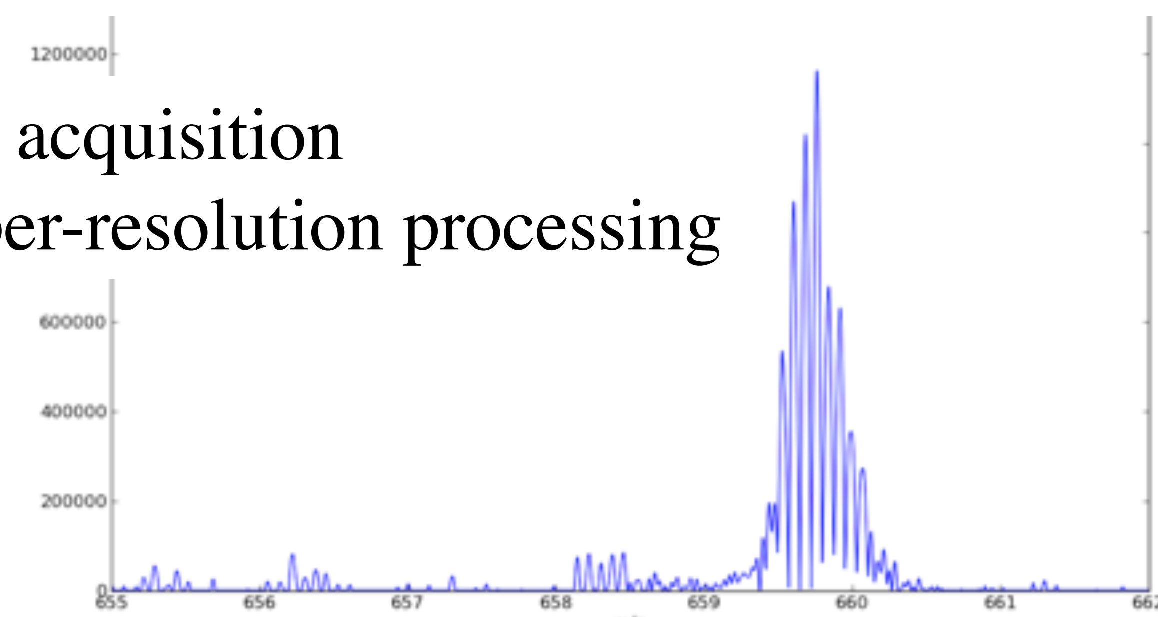
fast acquisition (10x faster)
standard processing





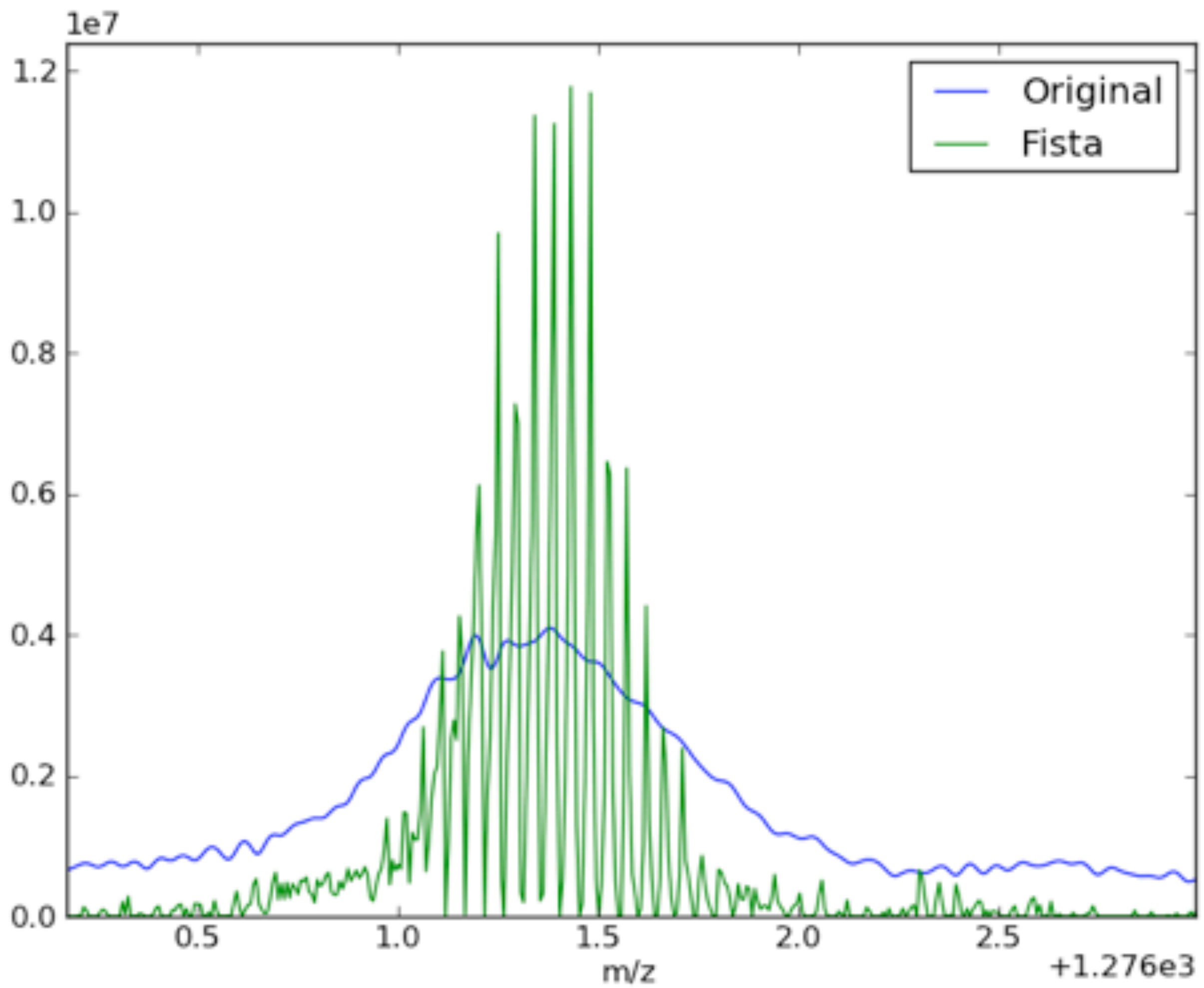
HiRes acquisition
standard processing

fast acquisition
super-resolution processing

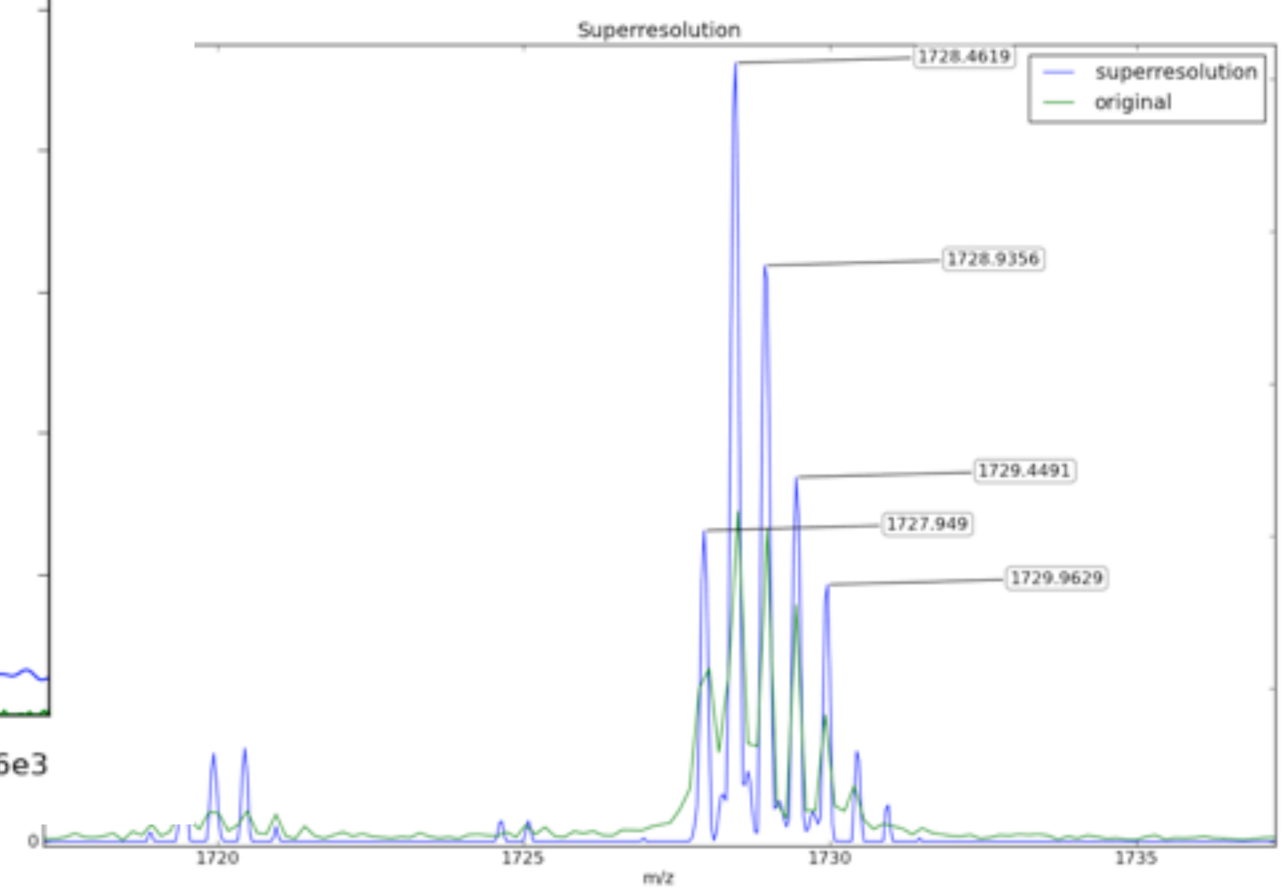


other example

- Standard FT vs Sur-resolution processing



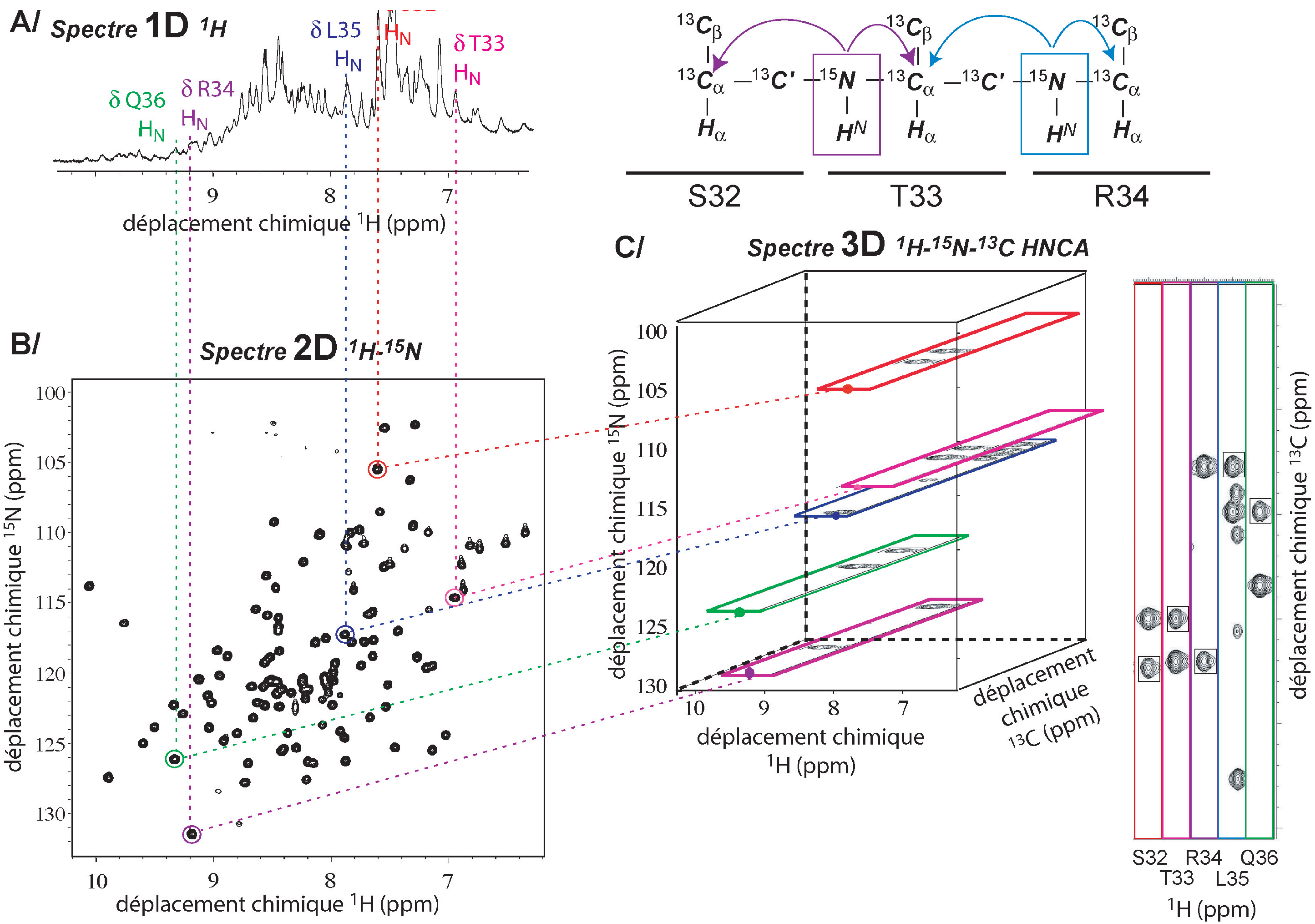
Fast Spectrum of Apolipoproteine A1
FT-ICR data
(30 kD z=24+ m/z 1271)



Fast Spectrum of Angiotensin
FT-ICR data
(m/z 1727)

(*col. C. Rolando Univ Lille*)

FTICR-2D


A/ Spectre 1D 1H
δ Q36 HN
δ R34 HN
δ L35 HN
HN
δ T33 HN
déplacement chimique 1H (ppm)
S32
T33
R34
B/ Spectre 2D 1H-15N
déplacement chimique 15N (ppm)
déplacement chimique 1H (ppm)
C/ Spectre 3D 1H-15N-13C HNCA
déplacement chimique 15N (ppm)
déplacement chimique 1H (ppm)
déplacement chimique 13C (ppm)
S32 T33 R34 L35 Q36
1H (ppm)
déplacement chimique 13C (ppm)

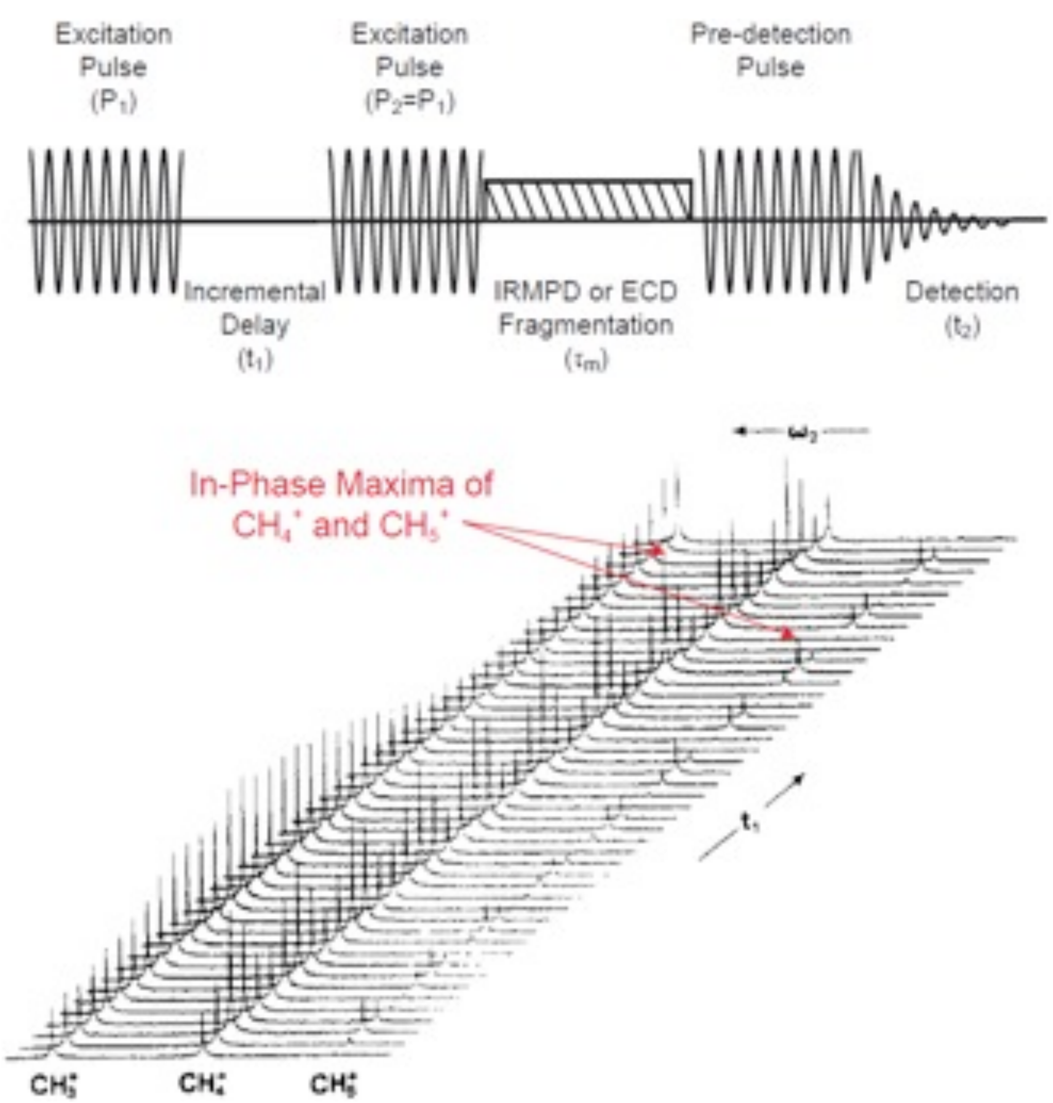
FTICR-2D

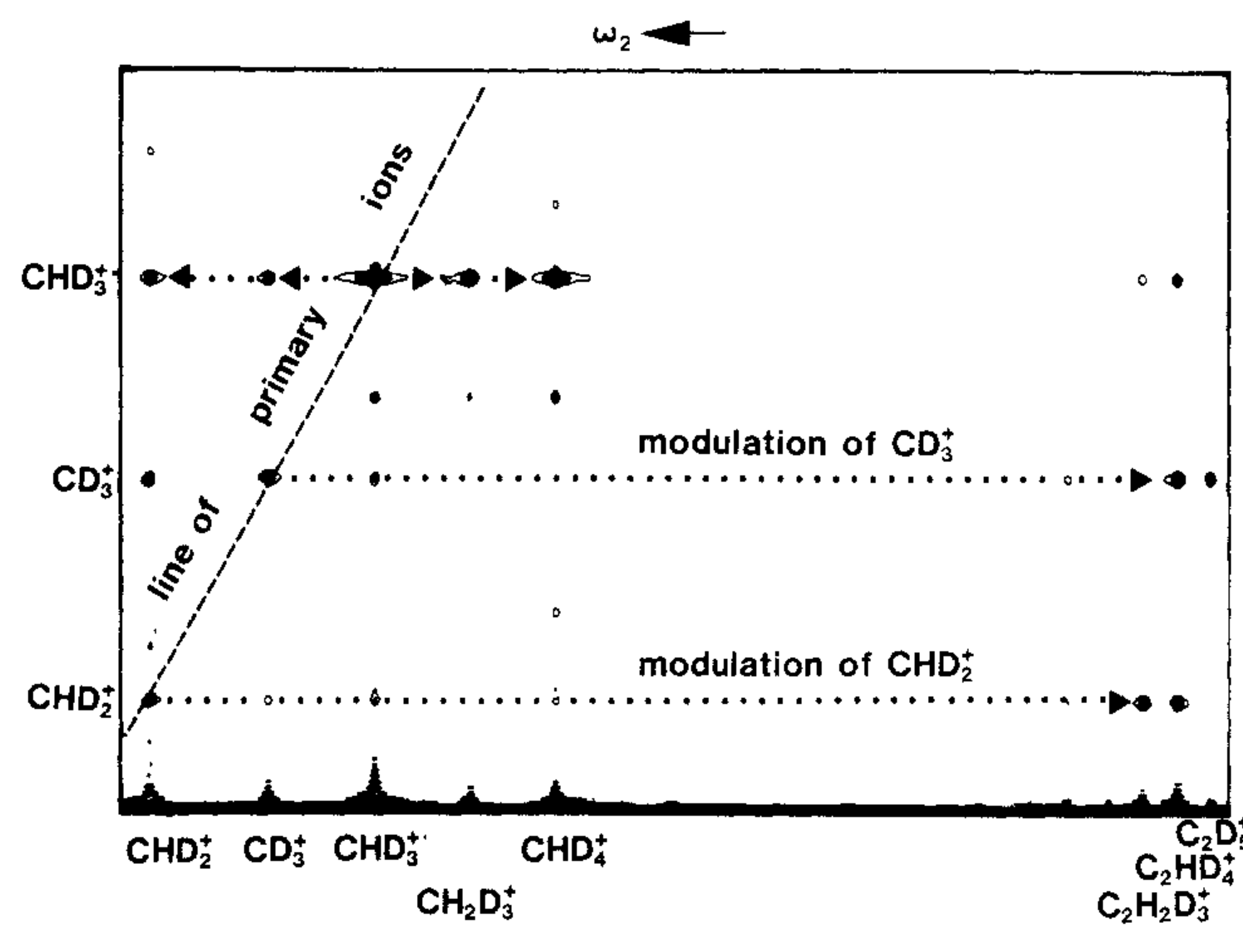
- principle of 2D FTICR proposed in 1987-88

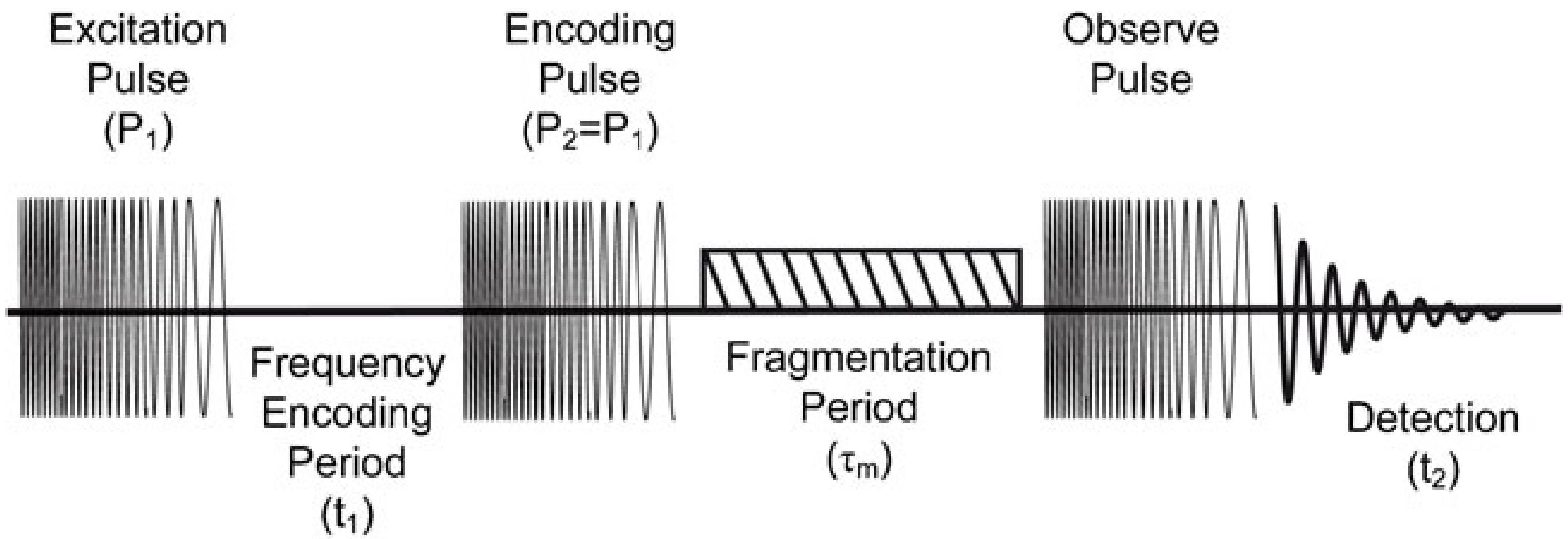
  nearly as old as 2D NMR

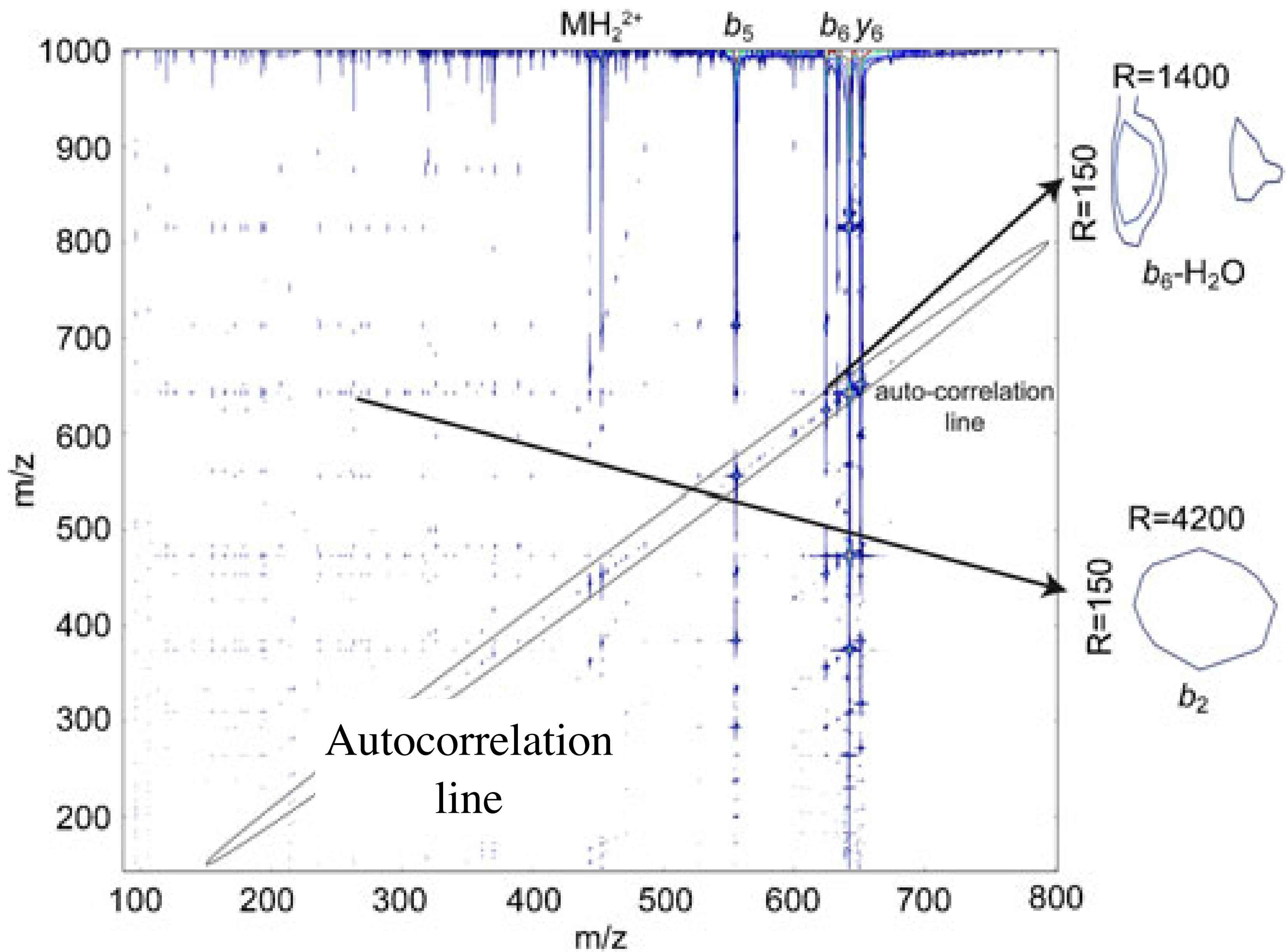
P Pfändler and G Bodenhausen and J Rapin and R Houriet and T Gäumann
Chem Phys Let (1987) vol. 138 (2) 195-200

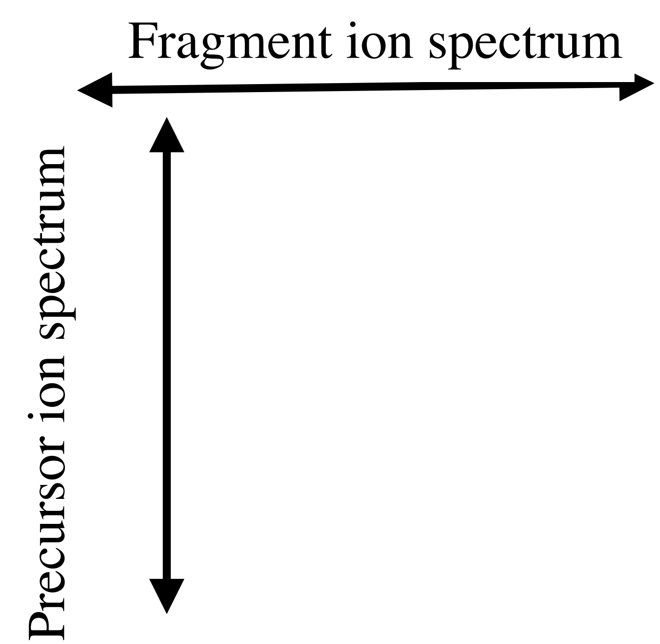
P Pfaendler, G Bodenhausen, J Rapin, M Walser, T Gaümann
J Am Chem Soc (1988) vol. 110 (17) 5625-5628









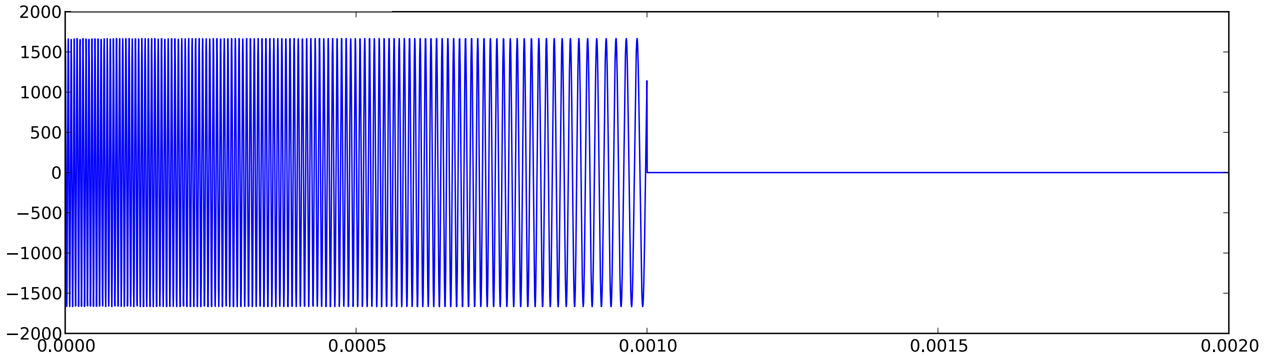


Two-dimensional FT-ICR MS spectrum of bradykinin using IRMPD fragmentation and Cadzow denoising (30 lines). Inserts enlargements of the b6 →b6-H2O and the b6 →b2 peaks.

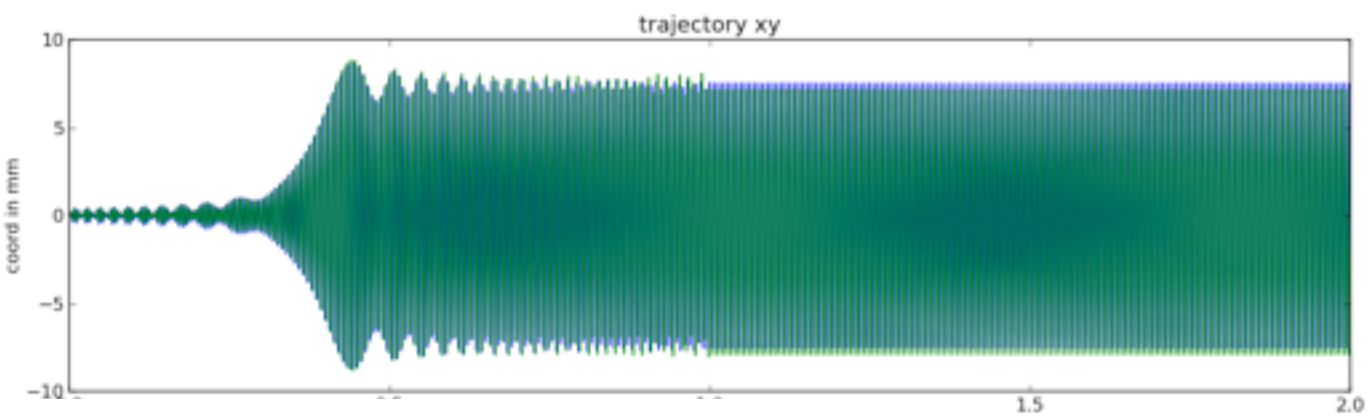
van Agthoven, M. A., Delsuc, M.-A., Bodenhausen, G. & Rolando, C. *Anal Bioanal Chem* **405,** 51–61 (2013).

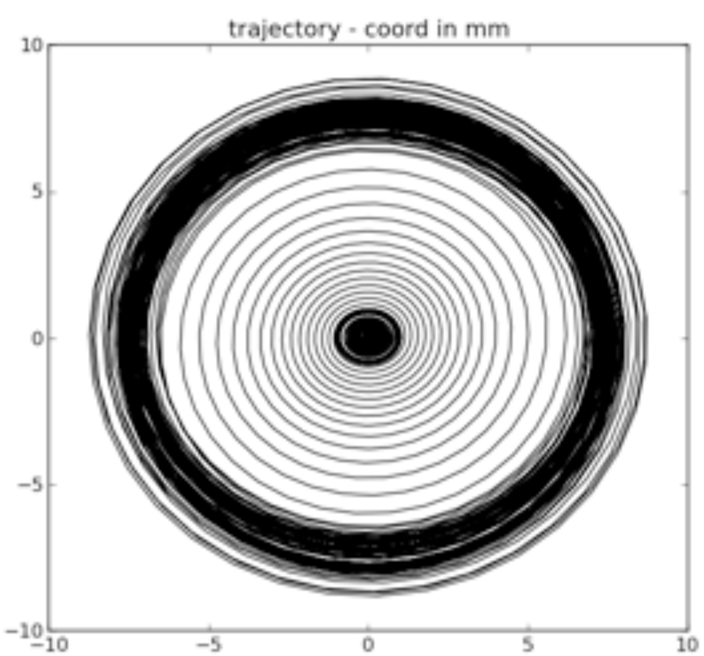
FT-ICR simulator

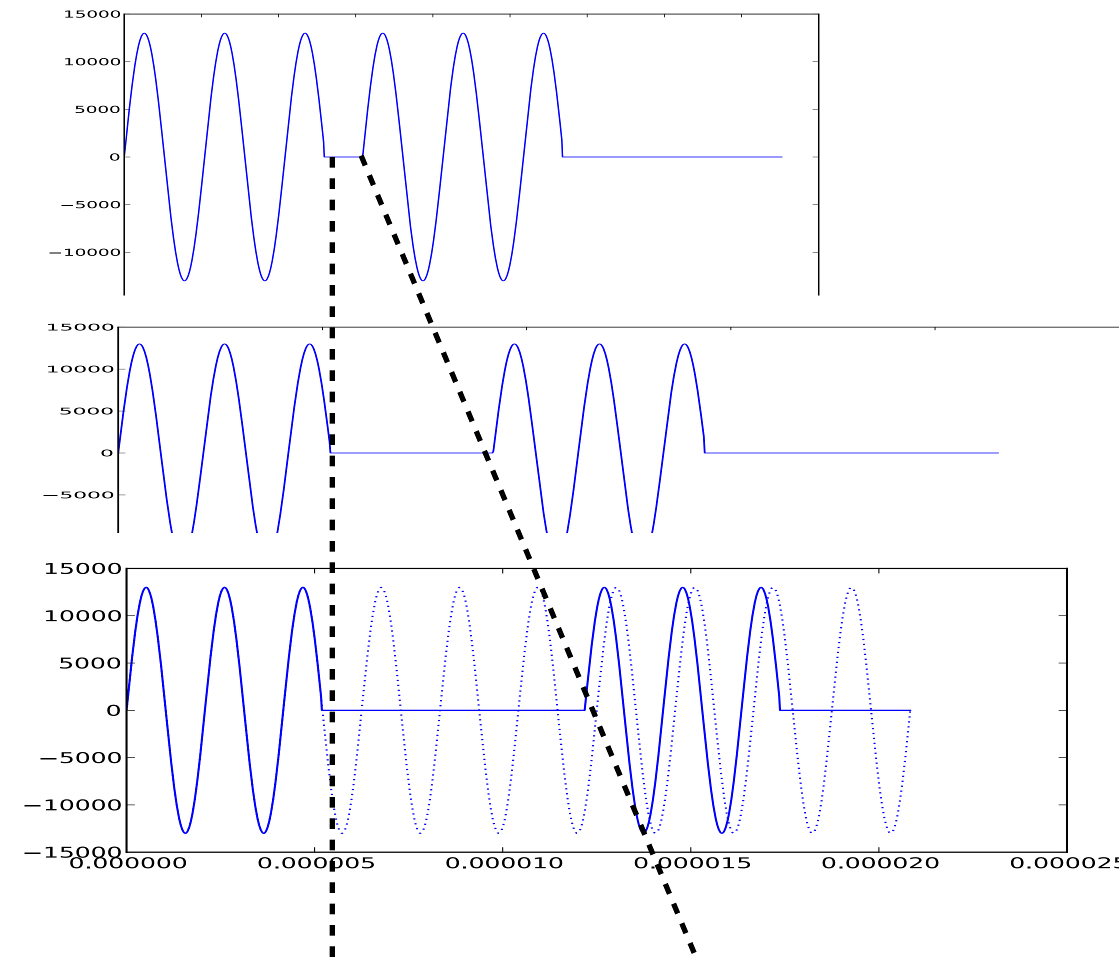
Chirp pulse



ion trajectory

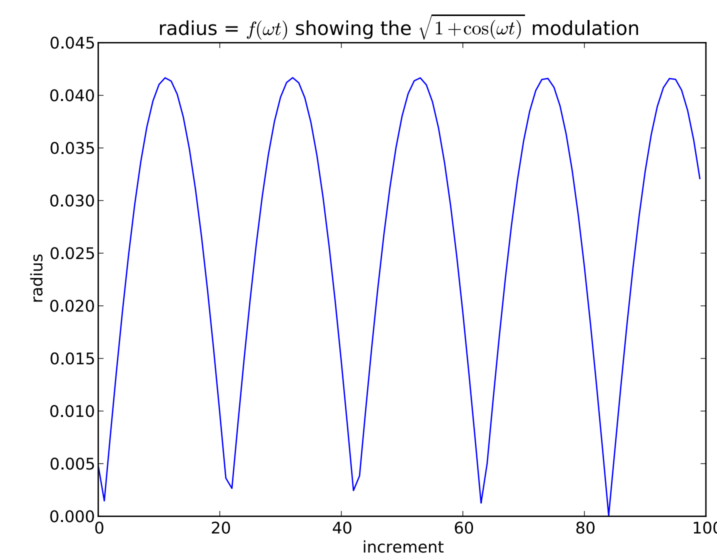




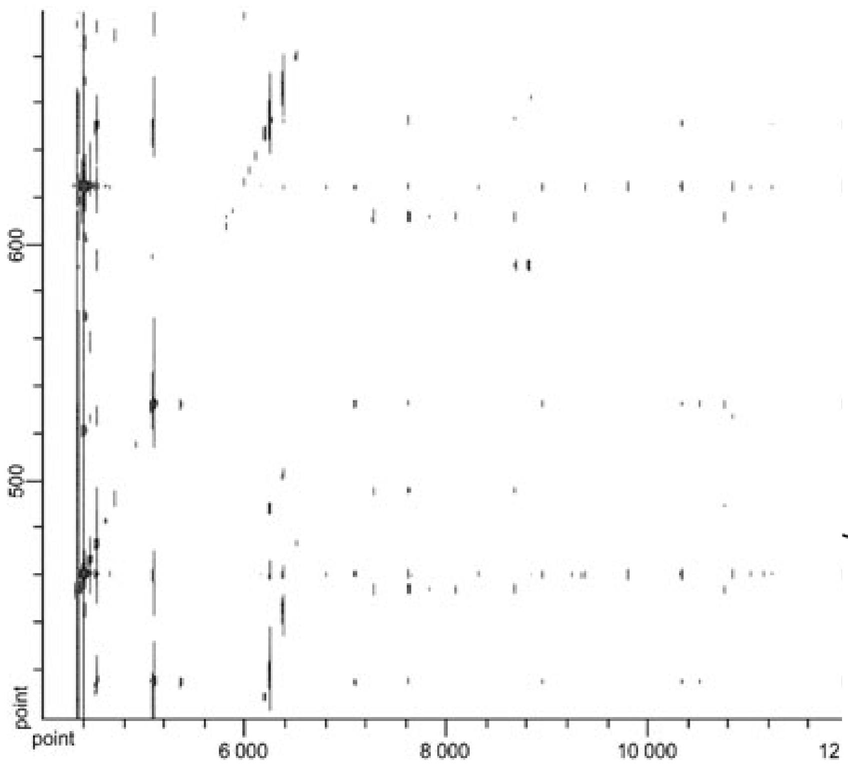


```
Swept pulse
Frequency : 200.000-50.000 kHz
sweep width : 150.000 kHz    sweep steps : 1000
duration : 1.000 msec
Epp : 1666.67 V/m
approx excitation radius : 11.82 mm

resonant frequency 144151.41 Hz
final radius : 7.56 mm
```

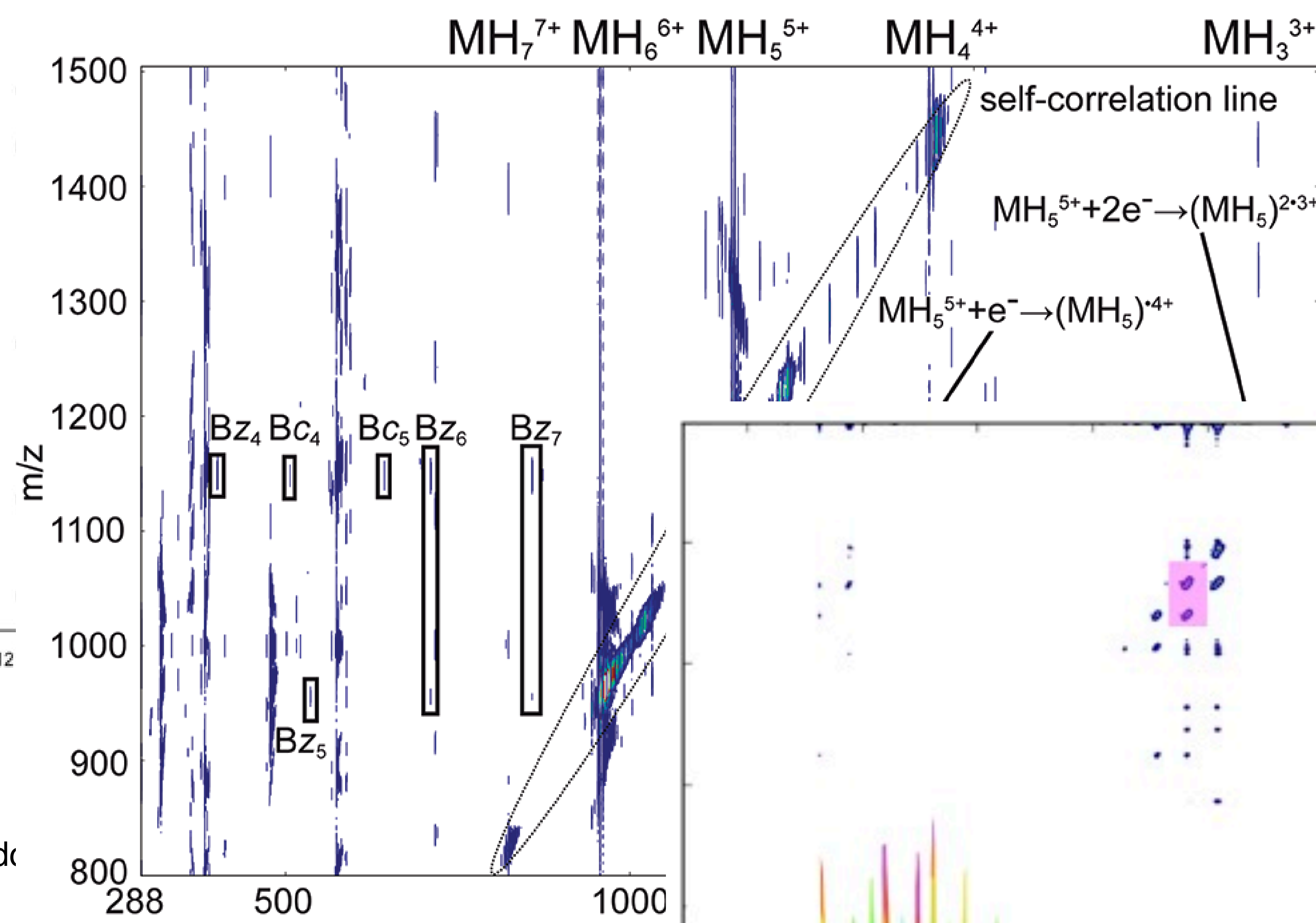


Improvements over time



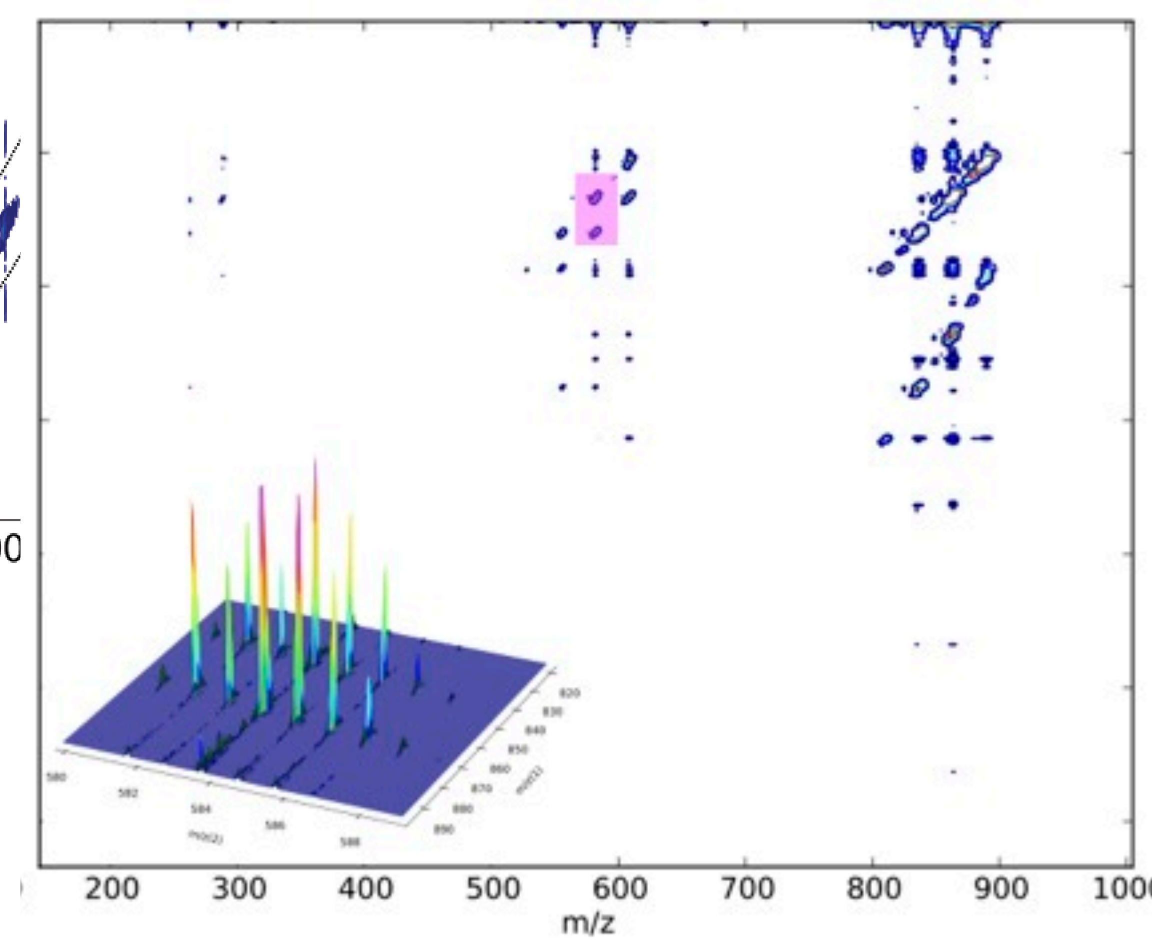
2011 : bradykinin - IRMPD

van Agthoven, M. A. V., Delsuc, M.-A. & Rolando, C.. *Int J Mass Spect.* **306,** 196–203 (2011).



2012 : BPTI - ECD

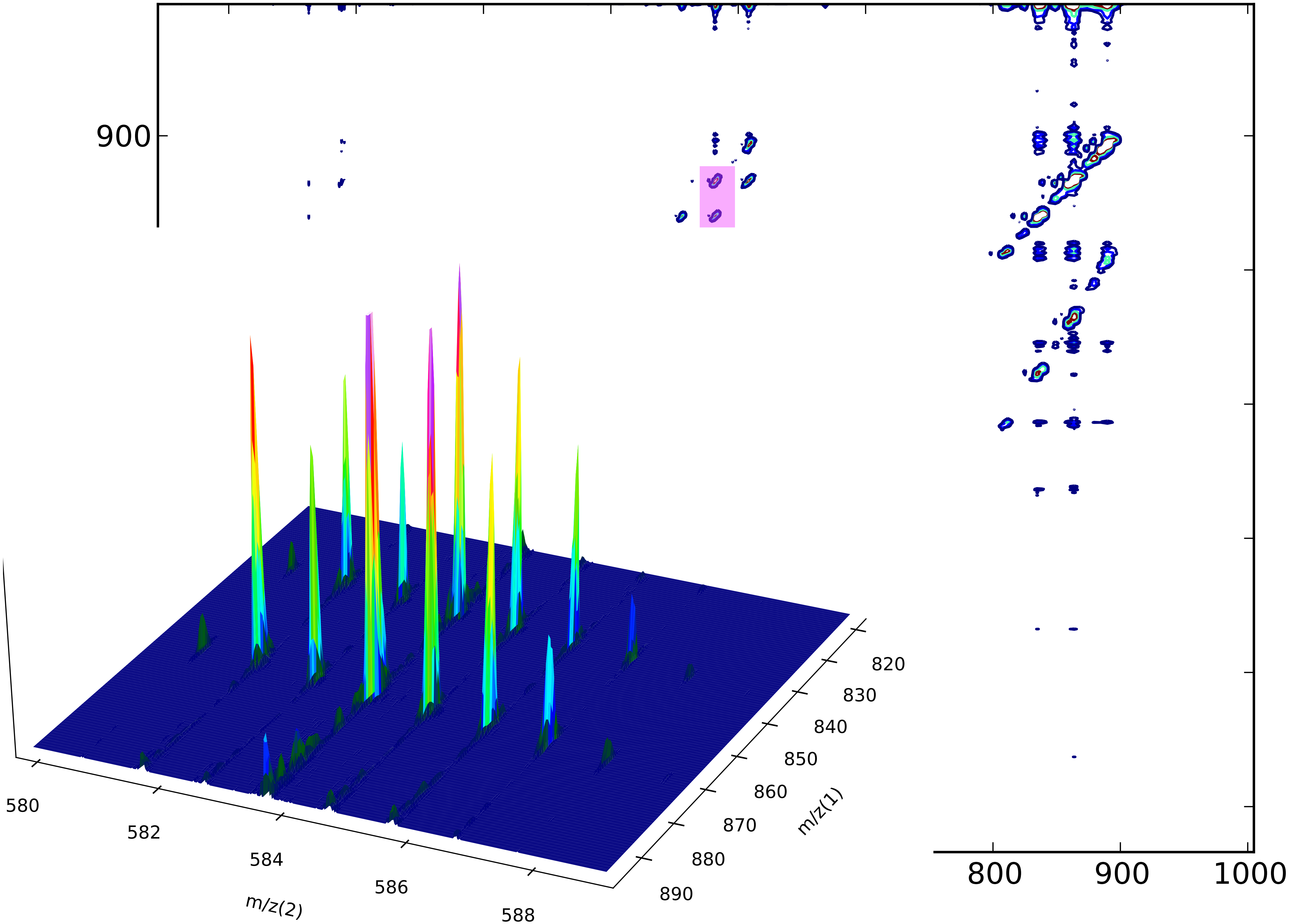
van Agthoven, M. A., Chiron, L., Coutouly, M.-A., Delsuc, M.-A. & Rolando, C.
Anal Chem **84,** 5589–5595 (2012).



2014 : triglycerid - high resolution

Chiron, L., van Agthoven, M. A., Kieffer, B., Rolando, C. & Delsuc, M.-A.
Proc Natl Acad Sci USA **111,** 1385–1390 (2014).

Tri Acyl Glycerol (TAG)



Le bruit

- Classique
  - additif - blanc - gaussien - centré

  - hétéroscédasticité

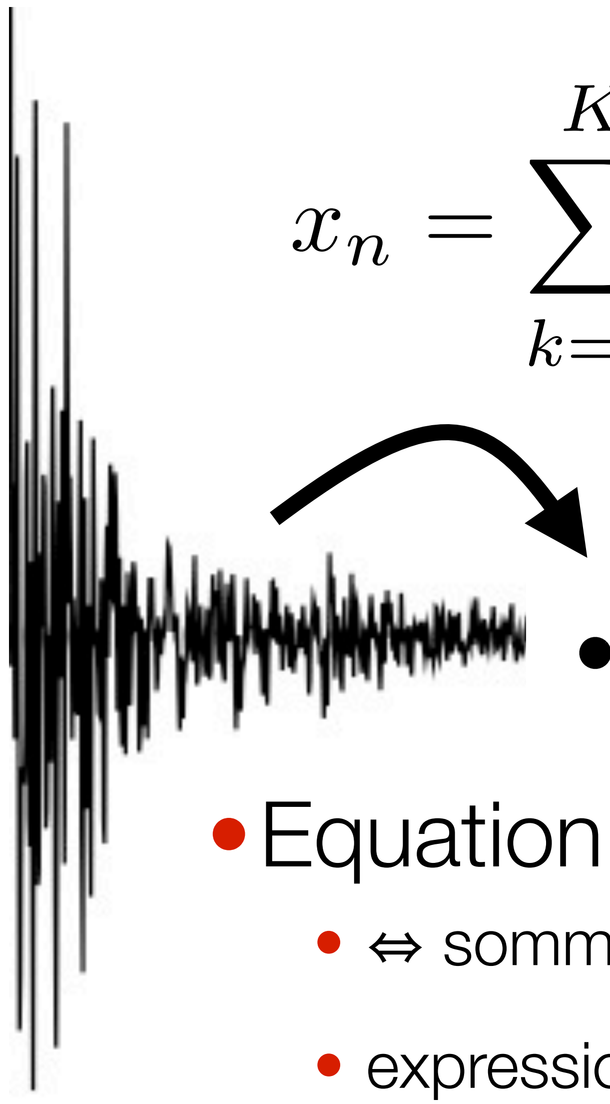
    bruit multiplicatif

    bruit de scintillation

    bruit d'instrument
  - traitement explicite dans les modèles mathématiques

- ce que je ne connais pas
  ce qui est en dehors de mon modèle
  - erreur
  - biais

Débruitage par la Procédure de Cadzow



$$x_n = \sum_{k=1}^{K} b_r x_{n-k}$$

$$\begin{bmatrix} x_1 & x_2 & \dots & x_P \\ x_2 & x_3 & \dots & x_{P+1} \\ x_3 & x_4 & \dots & x_{P+2} \\ x_4 & x_5 & \dots & . \\ . & . & \dots & . \\ x_{L-P} & x_{L-P+1} & \dots & x_{L-1} \end{bmatrix} \begin{bmatrix} b_P \\ b_{P-1} \\ b_{P-2} \\ . \\ b_1 \end{bmatrix} = \begin{bmatrix} x_{P+1} \\ x_P \\ x_{P-1} \\ . \\ x_L \end{bmatrix}$$

- Equation de Prédiction Linéaire (LP)
  - ⇔ somme de K sinusoïdes amorties
  - expression matricielle du problème LP

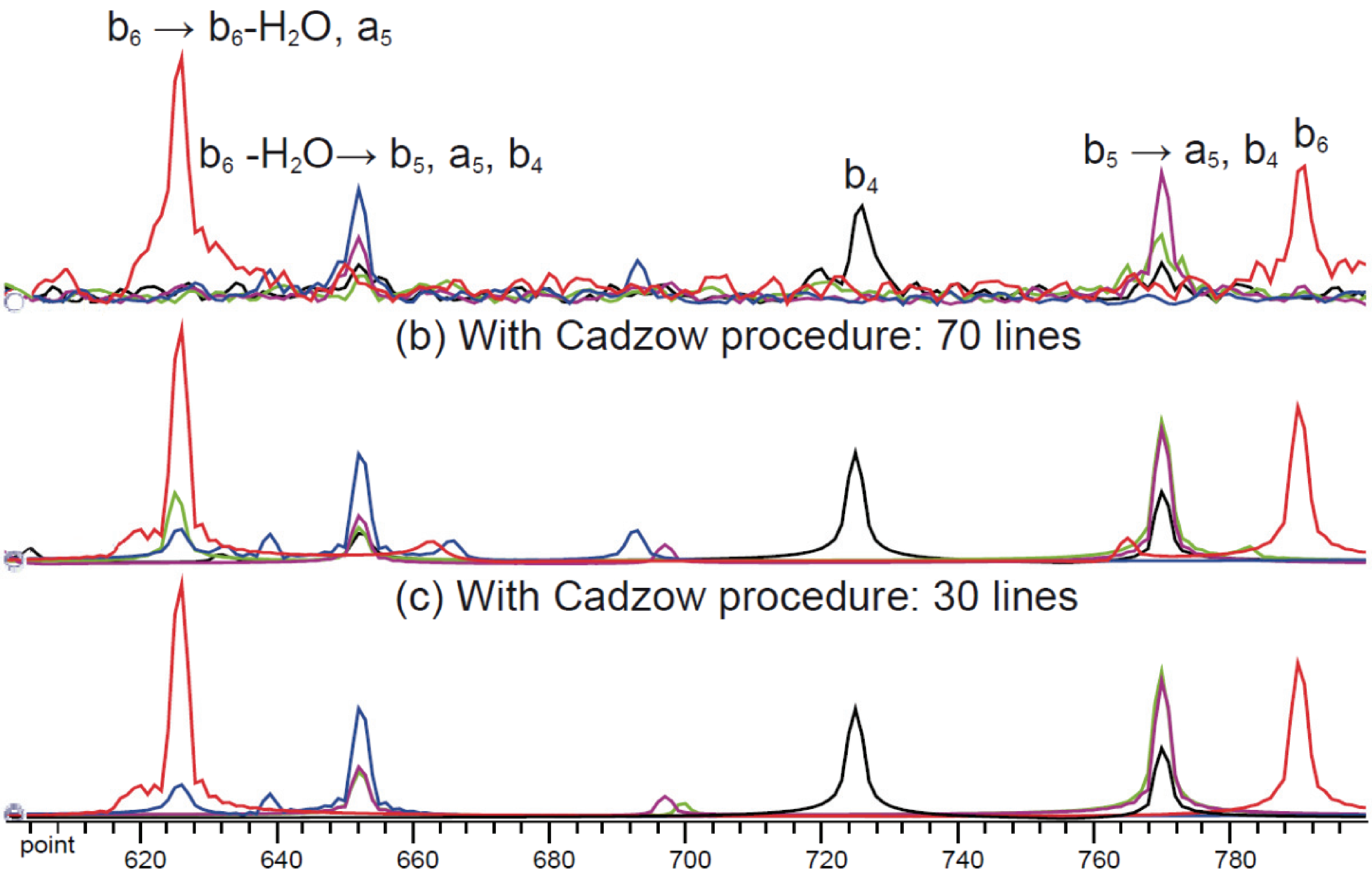
    $$Hb = x$$
  - H est décomposée en valeurs singulières

    $$H = U \times \Sigma \times V^*$$
  - puis tronquée à K valeurs propres

    $$\tilde{H} = U \times \tilde{\Sigma} \times V^*$$

$$\text{équivalent à } \|\tilde{H}\|_o = K$$

Cadzow



Vertical precursor ions spectra from the 2D IRMPD FT-ICR MS spectrum of bradykinin: *b6* (red), *b6*-H2O (blue), *b5* (pink), *a5* (green), *b4* (black) (a) without Cadzow procedure, (b) with Cadzow algorithm for 70 lines and (c) for 30 lines [32]

Improving the detection

- Sensitivity of the measure is governed by Signal/Noise ratio

  ⇒ increase signal

  ⇒ reduce noise

- Noise sources
  - “standard”
    - coming from the electronic on the apparatus

      ⇒ acquire more scan = takes time
  - scintillation noise
    - comes from the sample

      ⇒ no counter action during acquisition
    - preponderant in 2D (*t1-noise* in NMR)

$$S/N = \sqrt{N_{scan}}$$

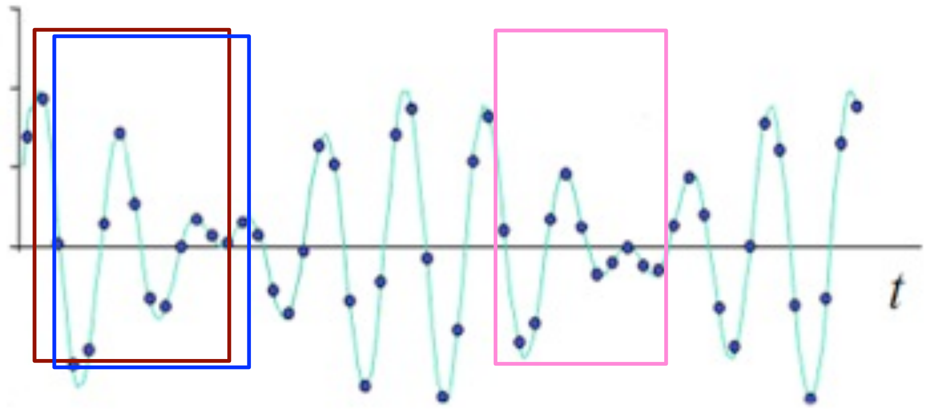
$$S/N \sim \text{invariant}$$

- Impact
  - better detection of weaker compounds
  - better coverage in bottom-up proteomics
  - better detection of PTM
  - faster acquisition

Statistical treatment

Signal time-series : $= (x_1, x_2, x_3, ...x_L)$

P frequencies



Uniform sampling

Hankel matrix

$$\begin{array}{cc} H = \begin{bmatrix} x_1 & x_2 & \cdots & x_N \\ x_2 & x_3 & \cdots & x_{N+1} \\ x_3 & x_4 & \cdots & x_{N+2} \\ x_4 & x_5 & \cdots & . \\ . & . & \cdots & . \\ x_M & x_{M+1} & \cdots & x_{M+N-1} \end{bmatrix} & M \\ N & \end{array}$$

M x *N*

$L = M + N + 1$

$M < N$

Hankel matrix: Same terms on antidiagonals

Cadzow procedure

- The idea is to decompose *H*
  - using **S**ingular **V**alue **D**ecompostion **SVD**

$$H = U \Sigma V^*$$

$$\Sigma = \begin{bmatrix} \sigma_1 & 0 & \dots & 0 \\ 0 & \sigma_2 & \dots & 0 \\ 0 & 0 & \dots & 0 \\ 0 & 0 & \dots & \sigma_N \\ \dots & & & \\ 0 & 0 & \dots & 0 \end{bmatrix}$$

  - singular values

$$\sigma_1 > \sigma_2 > \sigma_3 > \dots$$

- we keep only the *k* largest singular values
  - and reconstruct a **denoised** signal from the **rank-reduced** *H* matrix
  - projection of H on a subspace

$$\tilde{H} = U \Sigma_k V^*$$

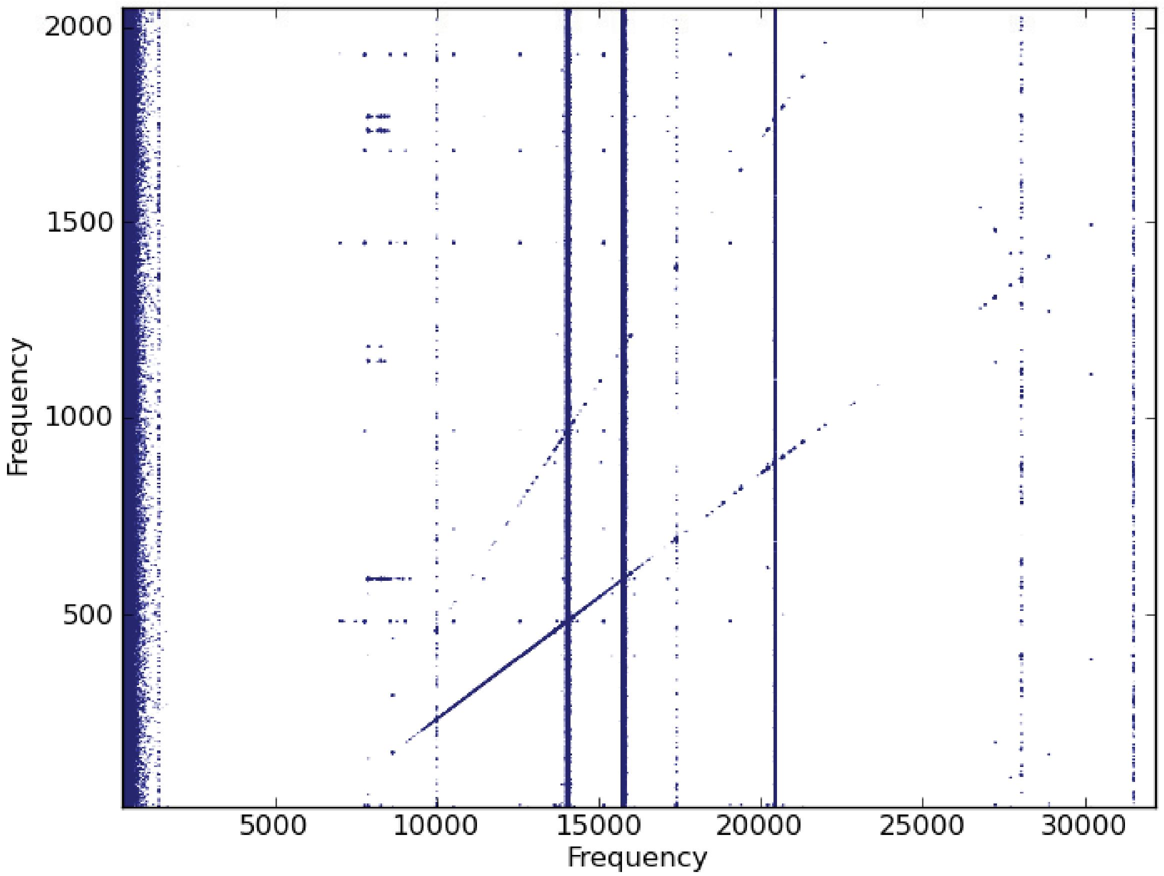
$$\tilde{H} = U_k U_k^* H$$

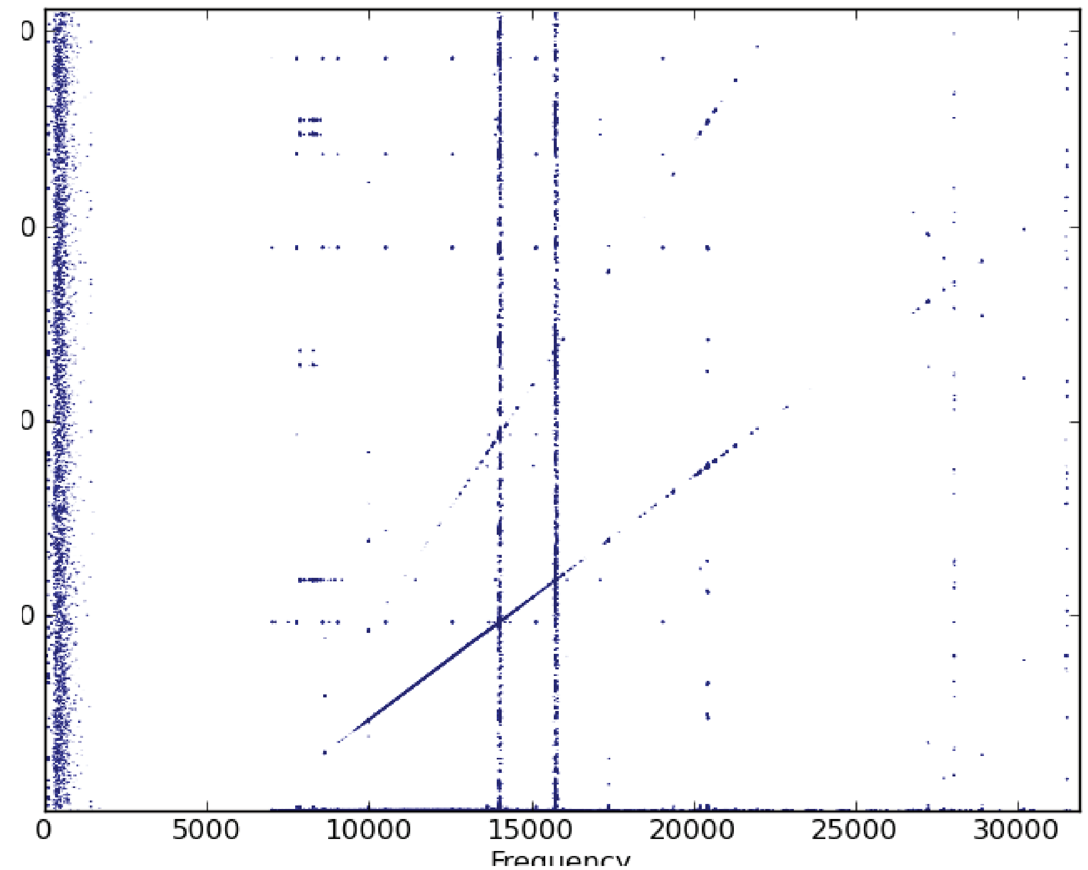
  - then averaging on H antidiagonals

$$\tilde{x}_p = \langle \tilde{H}_{ij} \rangle_{i+j-1=p}$$

Cadzow, J.A. (1988) *IEEE Trans. Acous. Speech Signal Proc.*, **36,** 49-62.

Effect of Cadzow cleaning





ECD - Substance P
1 hour acquisition

dataset : 512Mb
1 week processing
on departemental cluster !

Classical method : Cadzow
based on SVD analysis
does not scale well (N^3)

Sensitivity == noise reduction
Resolution == Size == processing time

Agthoven, M. A. V., Coutouly, M.-A., Rolando, C. & Delsuc, M.-A.
Rapid Commun Mass Spectrom **25**, 1609–1616 (2011).

Approximate by random sampling

combine several new mathematical ideas

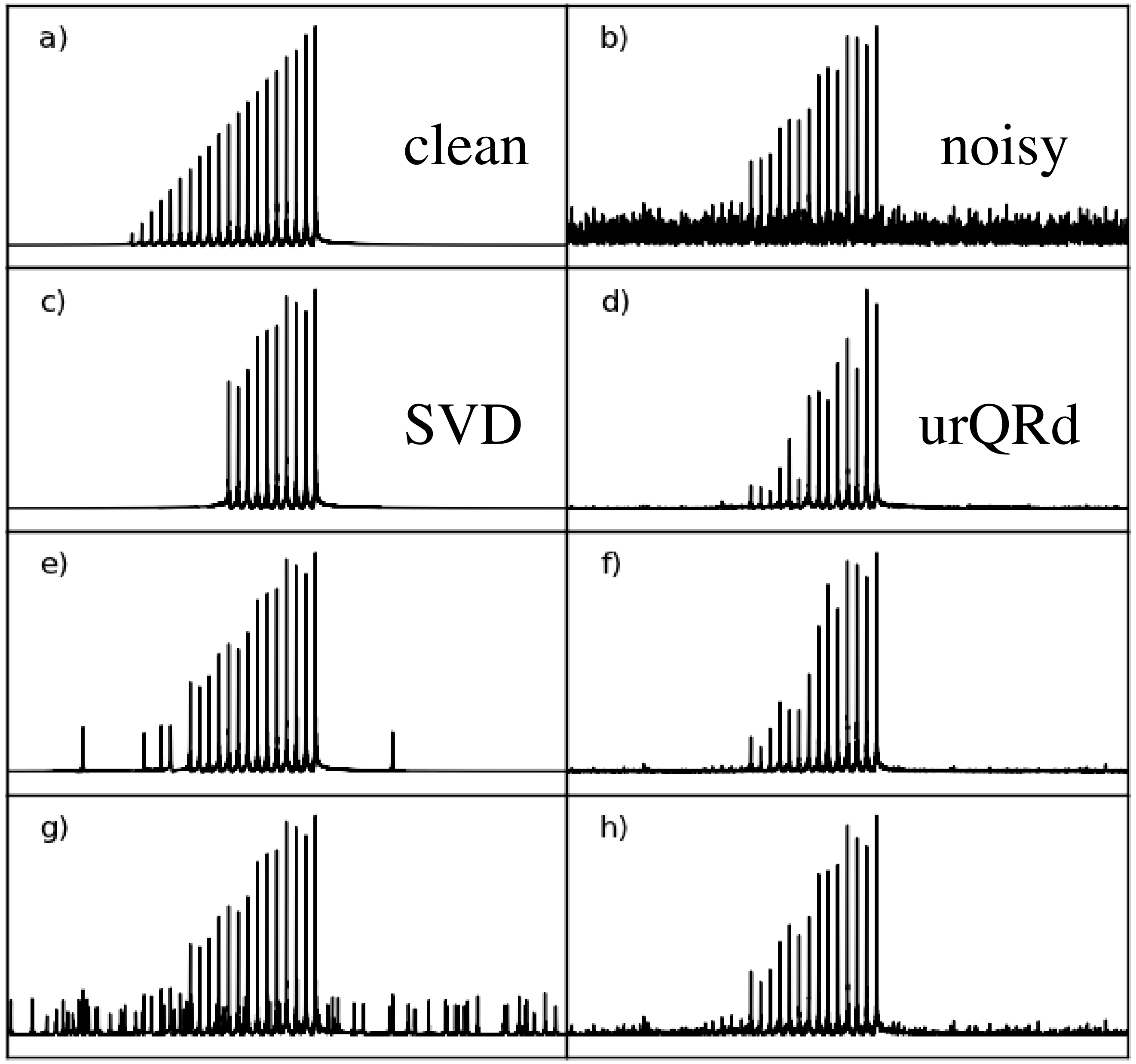
- use new developments linking between algebra and statistics
  - Johnson Linderstrauss Lemma (1984)
  - Compress Sensing approaches (Candès 2006, Donoho-Tanner 2007)

- Apply matrix approximation rather than complete matrices
  - Tygert, Martinsson (2007)

- ⇒ Estimate values rather than determining them
  - **SVD** can then be replaced by **QR** decomposition (faster)
  - precision and efficiency grows as the square root of the size
    hence efficient for Big Data

- **u**ncoiled **r**andom **QR d**enoising : **urQRd**
  - *noise reduction from random sampling !*

Example of urQRd on synthetic data



K : *estimate of number of signals*

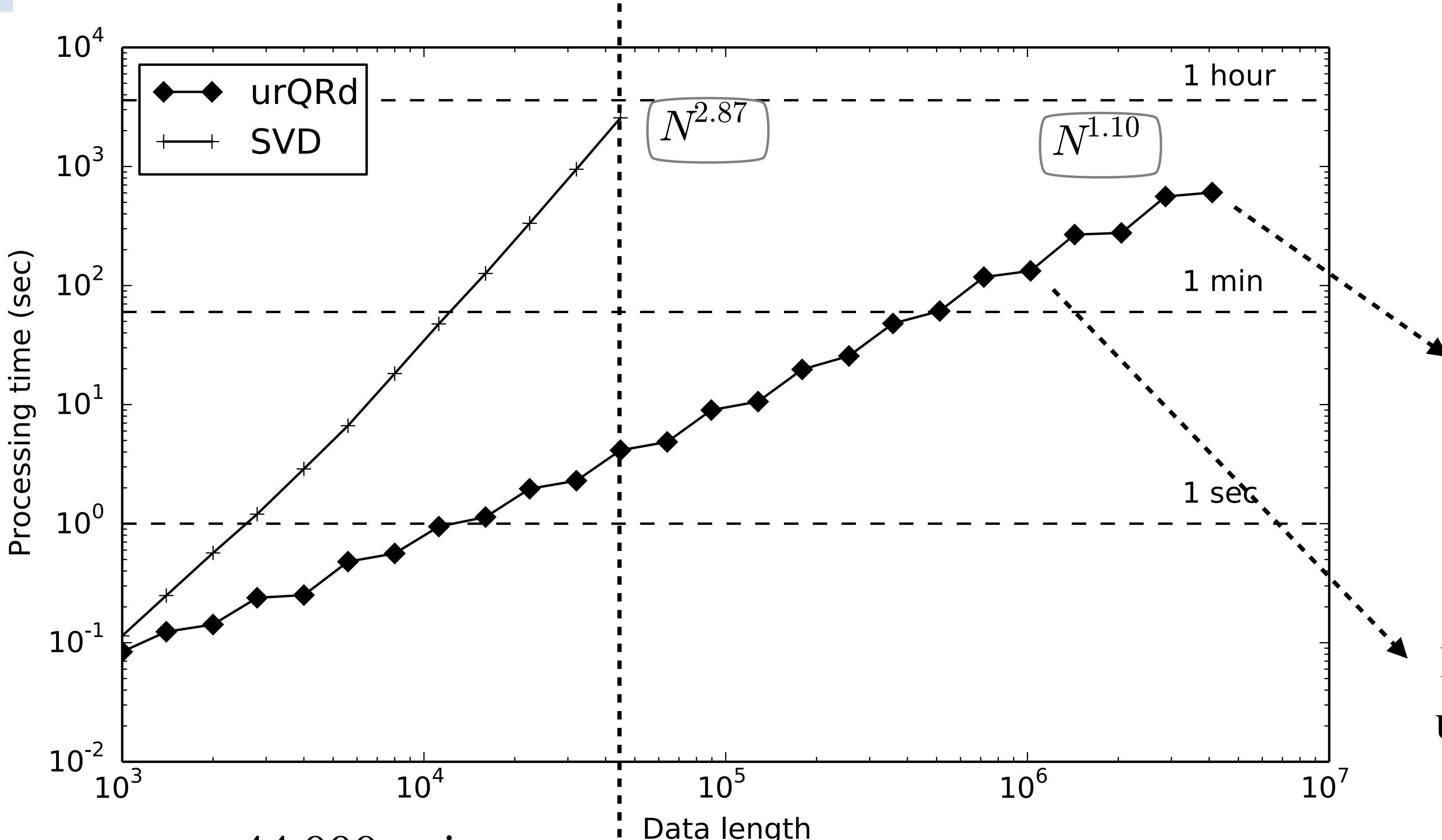
K=10

K=20

K=80

Fig. 1. Comparison of the the SNR gain afforded by the de-noising methods as a function of the rank. The computations are performed here on a synthetic complex 2000 points data-set containing 20 frequencies. a) Fourier Transform (FT) of the initial synthetic data-set composed of 20 lines of varying intensity. b) FT of the test data-set, with an added Gaussian white noise. SNR of the time-domain data-set is -0.14 dB. c-e-g) FT of the SVD processed of the synthetic data-set with with varying K. d-f-h) FT of the rQRd processed of the synthetic data-set with with varying K. c-d) rQRd and SVD processed of the synthetic data-set with $K = 10$ SNR gains : SVD 8.23 dB rQRd 2.91 dB, e-f) idem with $K = 20$ SVD 12.00 dB rQRd 5.13 dB. g-h) idem with $K = 80$ SVD 6.91 dB rQRd 9.95 dB.

much Faster - *much* Lighter



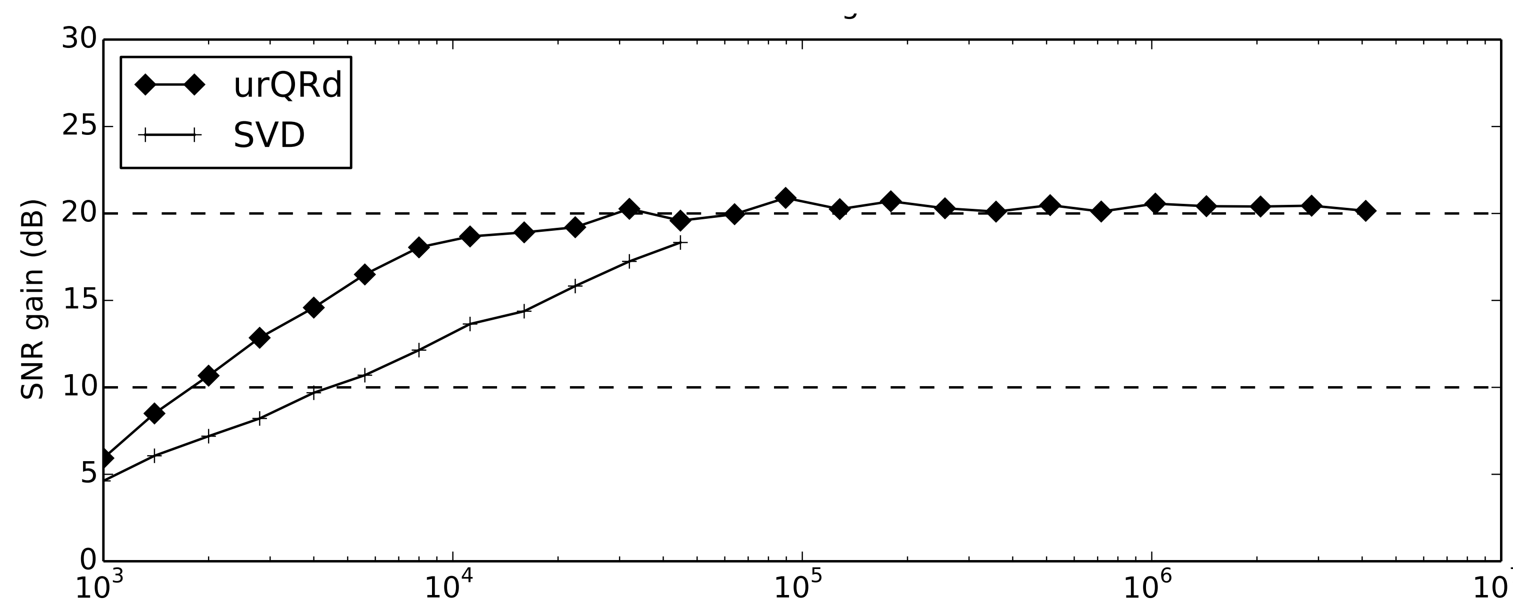
processing times
SVD in MN^2
urQRd in $KL\log(L)$
$K \ll M < N < L$

4.096.000 points
urQRd : 10 min

1.024.000 points
urQRd : 132 sec

44.000 points
SVD : 42 min
urQRd : 4.1 sec

better noise rejection



Applied to 1D FT-ICR data-set

http://urqrd.igbmc.fr

Tryptic digest of CytC
512k transient
1 second acquisition

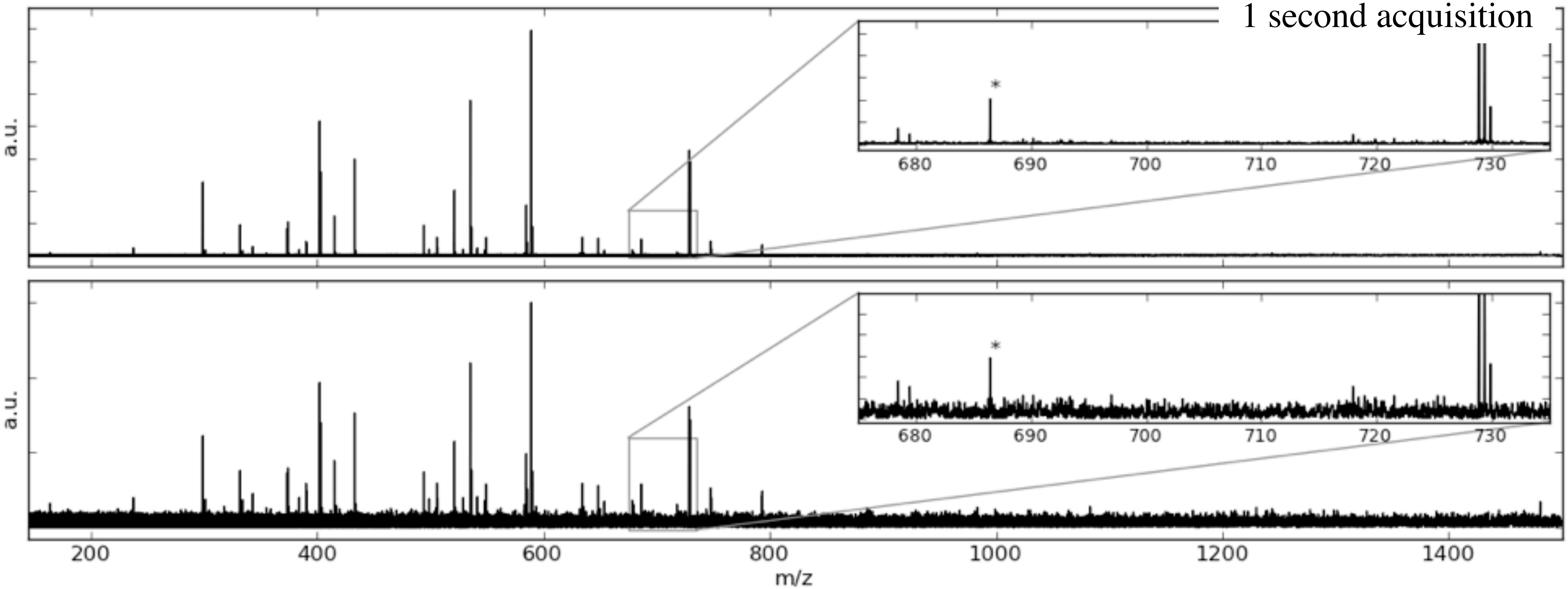
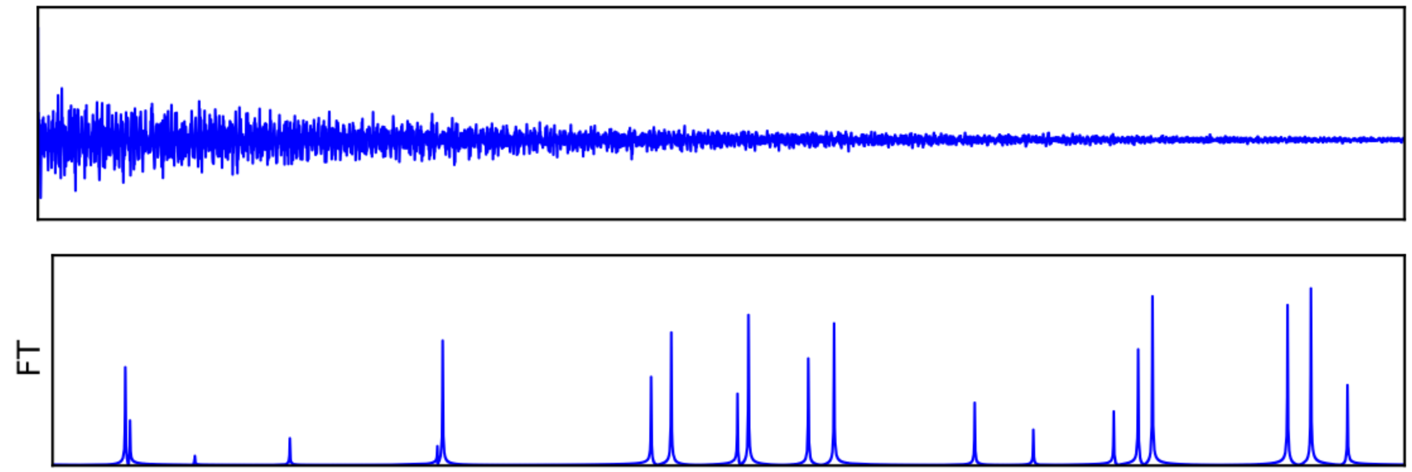


Fig. 4. Processing of a single-scan FT-ICR mass spectrum of a trypsin digest of Cytochrome C. *Bottom* original spectrum, SNR measured on the m/z 728.8388 peak is 24.0 dB. *Top* same spectrum after urQRd processing ($K = 1000$), SNR measured on the m/z 728.8388 peak is 40.7 dB. *inset*) the m/z 728.8388 peak corresponds to the TGQAPGFSTDANK^{2+} ion, m/z 678.3821 to YIPGTK+ and m/z 717.9012 to GEREDLIAYLKK^{2+}. The peak labeled with a star at m/z=686.390, lacking isotopic structure, is likely to be an experimental artifact. The processed interferogram is 512k points, processed here with $K = 1000$.

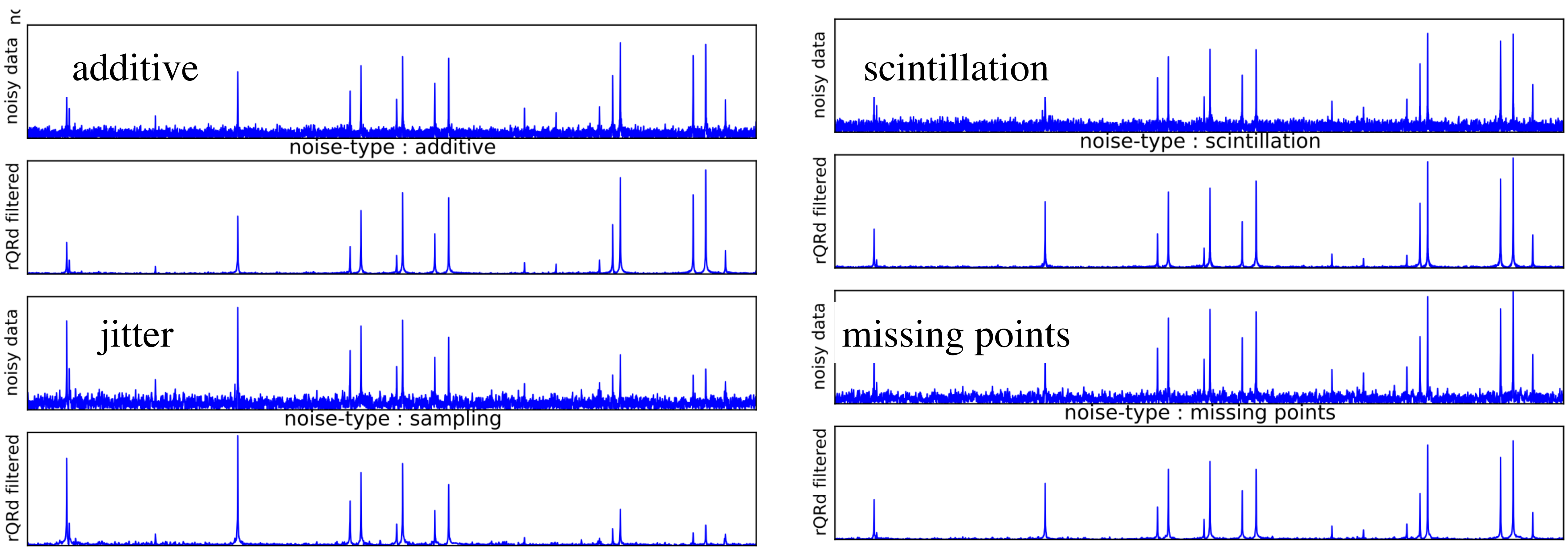
SVD	(*~45 days*	*512Go memory)*
urQRd	25 min	4Go memory

Chiron, L., van Agthoven, M. A., Kieffer, B., Rolando, C. & Delsuc, M.-A. *Proc Natl Acad Sci USA* **111,** 1385–1390 (2014).

Efficient on many kinds of noise



noise free data



effect of urQRd denoising

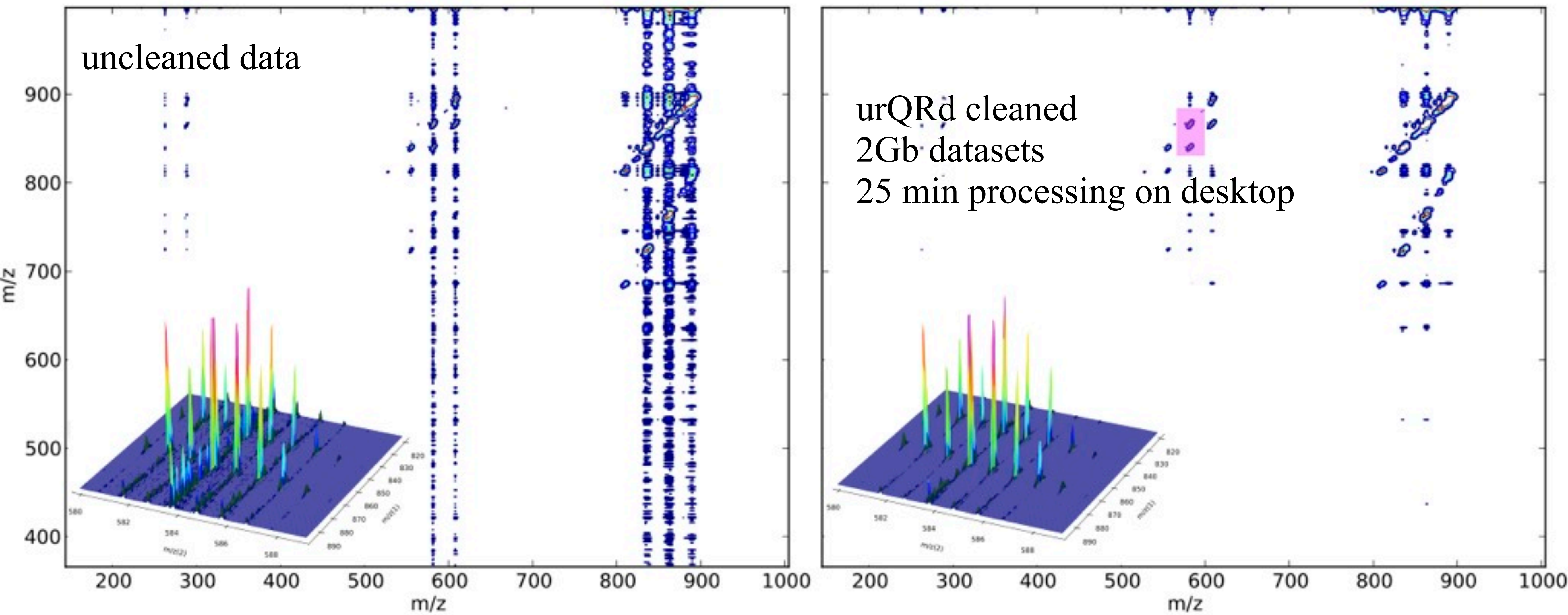
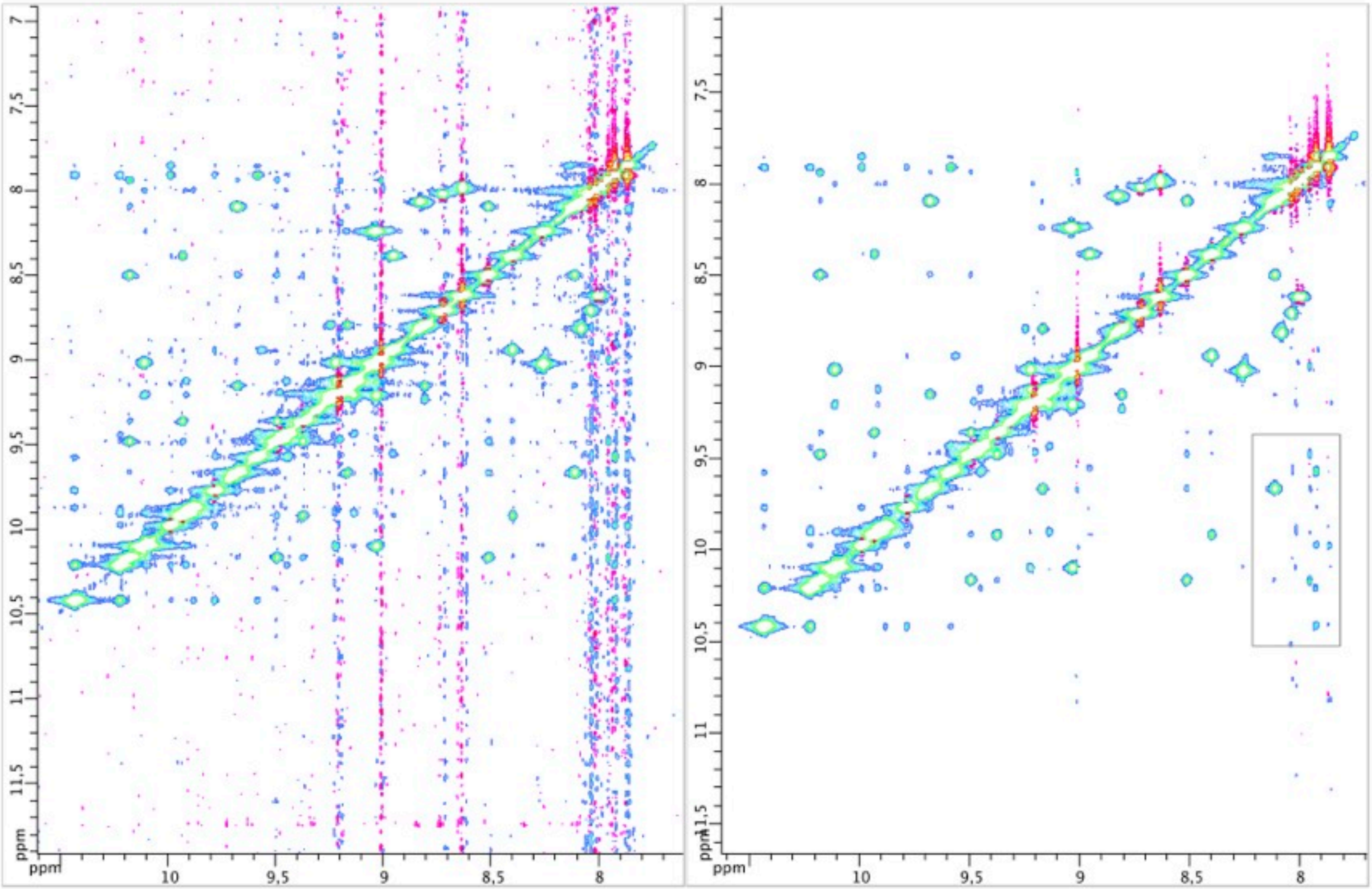
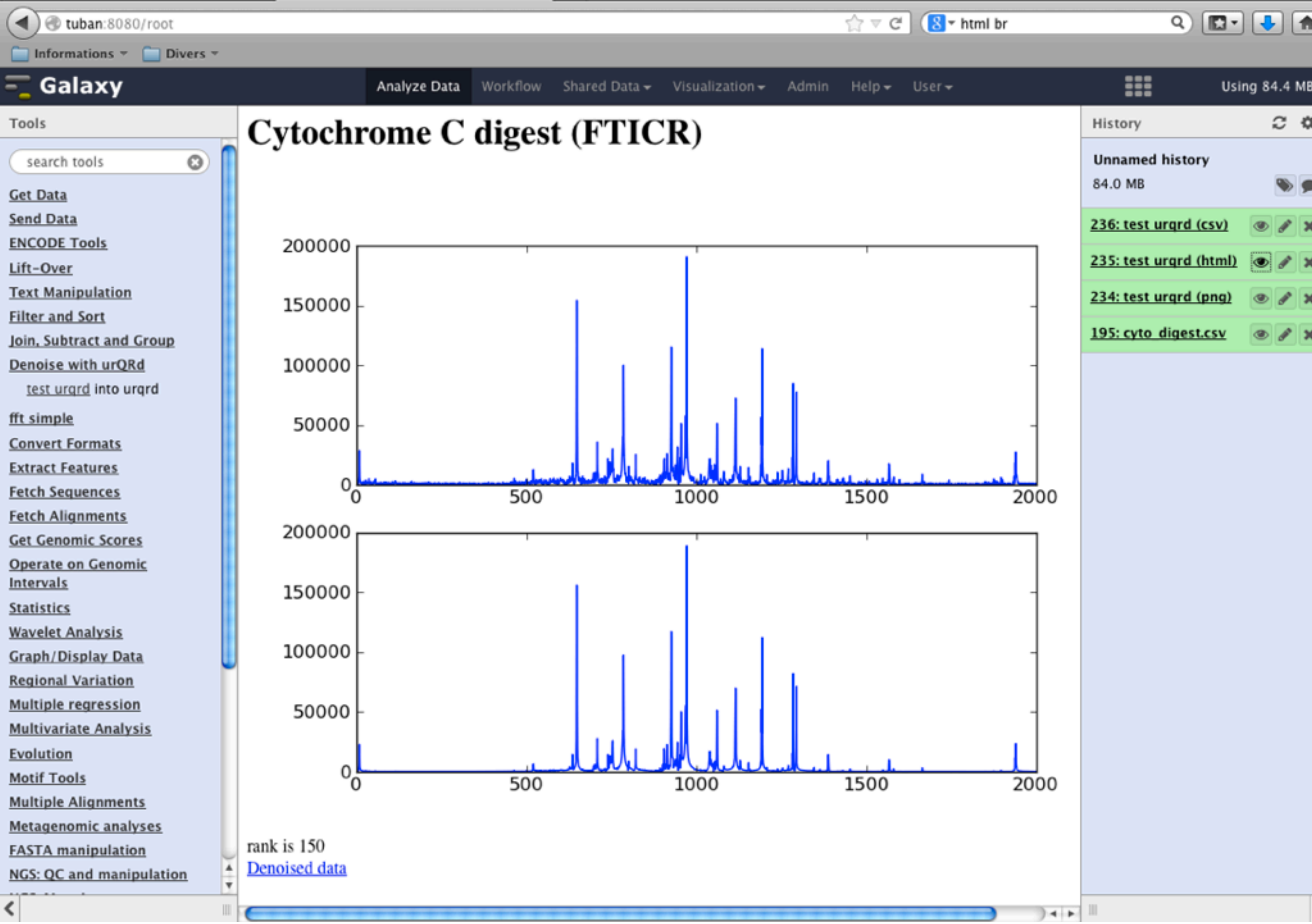


Fig. 5. 2D IRMPD FT-ICR MS spectrum of triacylglycerols extracted from human plasma showing strong scintillation noise. The data-set is 2k $\times$ 128k points. *inset*) zoom on the pattern centered at m/z(F1) 845 and m/z(F2) 584 (highlighted in pink). The two groups of peaks give the isotopic patterns of lithiated TAG(16 :0/16 :0/18 :1) at m/z 839.7674 and lithiated TAG(16 :0/18:1/18 :1) at m/z 865.7831 respectively losing a palmitic acid (MW 256.2396) and an oleic acid (MW 282.2553) in order to yield a lithiated diacylglycerol DAG(16:0/18:1) at m/z 583.5278(33). SNR was measured on the zoomed zone to 22.2 dB and 42.8 dB for the standard and de-noised datasets respectively.

Very efficient



Reduction of t1-noise on a 2D NOESY spectrum ~2minutes


tuban:8080/root
html br
Informations
Divers
Galaxy
Analyze Data
Workflow
Shared Data
Visualization
Admin
Help
User
Using 84.4 MB
Tools
search tools
Get Data
Send Data
ENCODE Tools
Lift-Over
Text Manipulation
Filter and Sort
Join, Subtract and Group
Denoise with urQRd
test urqrd into urqrd
fft simple
Convert Formats
Extract Features
Fetch Sequences
Fetch Alignments
Get Genomic Scores
Operate on Genomic Intervals
Statistics
Wavelet Analysis
Graph/Display Data
Regional Variation
Multiple regression
Multivariate Analysis
Evolution
Motif Tools
Multiple Alignments
Metagenomic analyses
FASTA manipulation
NGS: QC and manipulation
Cytochrome C digest (FTICR)
200000
150000
100000
50000
0
0
500
1000
1500
2000
rank is 150
Denoised data
History
Unnamed history
84.0 MB
236: test urqrd (csv)
235: test urqrd (html)
234: test urqrd (png)
195: cyto_digest.csv

Acknowledgments

- IGBMC NMR group
  - B.Kieffer
- IGBMC Bioinformatic
  - J. Seiler
  - S. Legras
  - M. Distelzwey
- NMRTEC
  - L.Chiron
  - J-P Starck
  - M-A Coutouly
- Lille - MSAP
  - C. Rolando
  - M. van Agthoven
- Paris - Pasteur
  - J Chamot-Rooke
- Paris - ENS
  - G. Bodenhausen
  - P. Pelupessy
  - A. Sehgal
- MesureHD group - Big-data
  - P.L. Combette (UPMC)
  - C. Pesquet (Paris-Est)
  - J Chamot-Rooke (Pasteur)
  - C. Rolando (Lille)
  - P. Roy (Soleil)
  - P. Collet (iCube)

fundings

- ANR
  - grant 2010 FT-ICR-2D
  - grant 2010 FRISBI
- CNRS
  - grant 2013 Mesure-HD
- NMRTEC S.A.S