



Data and biophysics

Oswaldo Burastero

ARISE Fellow, Garcia-Alai Team

MOSBRI Course - Quality control for Integral Membrane Proteins

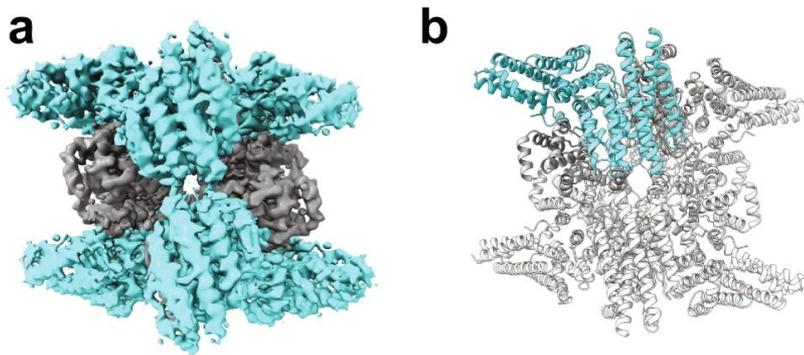
2022

14 September 2022

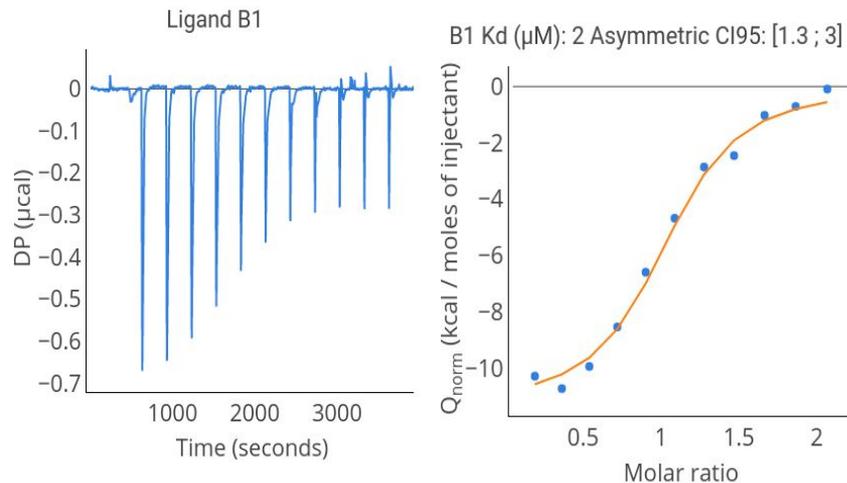


The final objective

Cryo-EM structure of a 16-mer AENTH complex



Isothermal titration calorimetry example

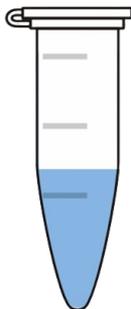


How to get there



1. Experiment planning

- Data management plan
- Predict possible outcomes



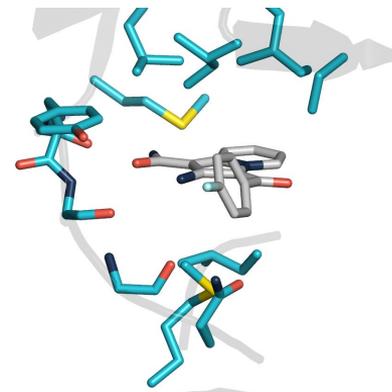
2. Sample preparation

- Protein expression & purification



3. Quality check

- Homogeneity
- Integrity
- Identity



4. Measure & analysis

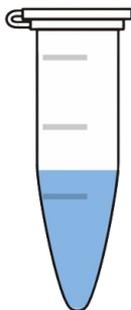
- Activity / inhibition assays
- Structural studies

How to get there



1. Experiment planning

- Data management plan
- Predict possible outcomes



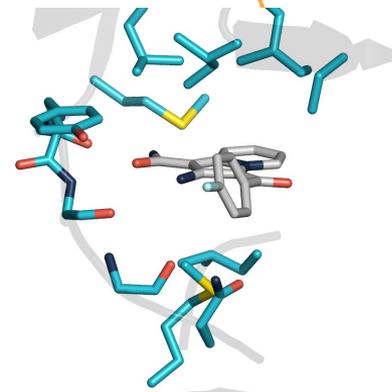
2. Sample preparation

- Protein expression & purification



3. Quality check

- Homogeneity
- Integrity
- Identity



4. Measure & analysis

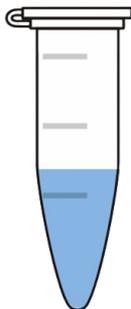
- Activity / inhibition assays
- Structural studies

How to get there



1. Experiment planning

- Data management plan
- Predict possible outcomes



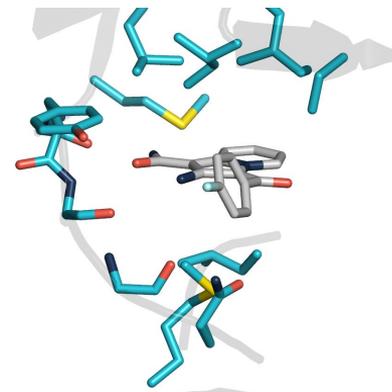
2. Sample preparation

- Protein expression & purification



3. Quality check

- Homogeneity
- Integrity
- Identity



4. Measure & analysis

- Activity / inhibition assays
- Structural studies

What is research data?

Data	Metadata
<ul style="list-style-type: none">● Digital data generated during and after the research project<ul style="list-style-type: none">● Observations● Acquired data (raw & processed): text files, videos, ...● Software (code, algorithms)	<ul style="list-style-type: none">● Data about the data (provides context)<ul style="list-style-type: none">● Origin of the data● Who, when, why, how● Used resources● Licenses



“Good metadata will save us
precious time in the future”

Data management plan (DMP)

What is a DMP?

- Formal document that describes how the data will be handled during and after the project

Why we need it?

- Good scientific practice
- Required by funders and/or institution



Horizon 2020
European Union Funding
for Research & Innovation



Data management plan (DMP)

What is a DMP?	Why we need it?	Benefits	FAIR Principles
<ul style="list-style-type: none">Formal document that describes how the data will be handled during and after the project	<ul style="list-style-type: none">Good scientific practiceRequired by funders and/or institution	<ul style="list-style-type: none">Save time and resourcesImproved reproducibility and reusability	<ul style="list-style-type: none">FindableAccesibleInteroperableReusable



Horizon 2020
European Union Funding
for Research & Innovation

scientific **data**

[Explore content](#) ▾ [About the journal](#) ▾ [Publish with us](#) ▾

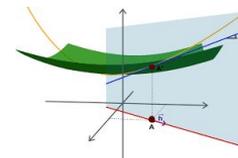
[Open Access](#) | [Published: 15 March 2016](#)

The FAIR Guiding Principles for scientific data management and stewardship



DMP checklist

Project	What is the project?
Data	What type, format and size does the used and produced data have?
Interpretation	Which information is required to understand the data?
Procedures	Which procedures will be used to create, process and quality control the (meta)data?



DMP checklist

Documentation	How the data processing steps will be recorded?
Access	Are there any security or access control requirements?
Project end	What happens to the data after the project finishes?



DMP checklist

Documentation	How the data processing steps will be recorded?
Access	Are there any security or access control requirements?
Project end	What happens to the data after the project finishes?
Intellectual property (IP)	How will be the IP managed?
Responsibilities	Who is responsible for which part of the data management?



DMP checklist - zoom in

Interpretation	
<ul style="list-style-type: none">• Can the data be read only with specific software?• Where is the data documentation to be found? Lab information management system?• How is data going to be documented?<ul style="list-style-type: none">• Metadata, identifiers (of biological entities) & ontologies	



DMP checklist - zoom in

Interpretation	Procedures
<ul style="list-style-type: none">• Can the data be read only with specific software?• Where is the data documentation to be found? Lab information management system?• How is data going to be documented?<ul style="list-style-type: none">• Metadata, identifiers (of biological entities) & ontologies	<ul style="list-style-type: none">• How will data and files be named and organised?• How will changes be tracked and propagated?<ul style="list-style-type: none">• How will metadata and provenance be preserved?• How will derived data be updated?

DMP checklist - zoom in

Processing

- Manual data processing steps
- Configuration parameters
- Analysis versions
- Scientific workflow management system
- Open source software
- etc.



DMP in real life - a living document

Data generated for analysis of protein X stability and homogeneity

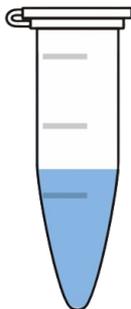
Dataset	Origin	Size	Format	Availability
DSF data	measurements performed in a nDSF Prometheus® (Nanotemper)	<100 MB	.xlsx .csv	Open
DLS data	measurements performed in a DynaPro® Plate Reader (Wyatt Technology)	<10 MB	.csv	Open

How to get there



1. Experiment planning

- Data management plan
- Predict possible outcomes



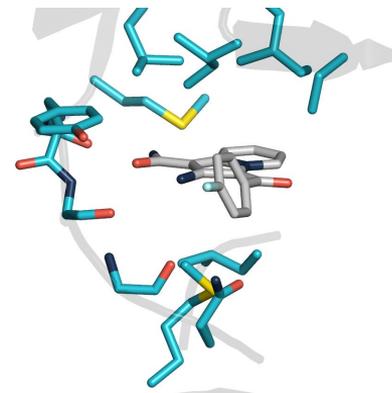
2. Sample preparation

- Protein expression & purification



3. Quality check

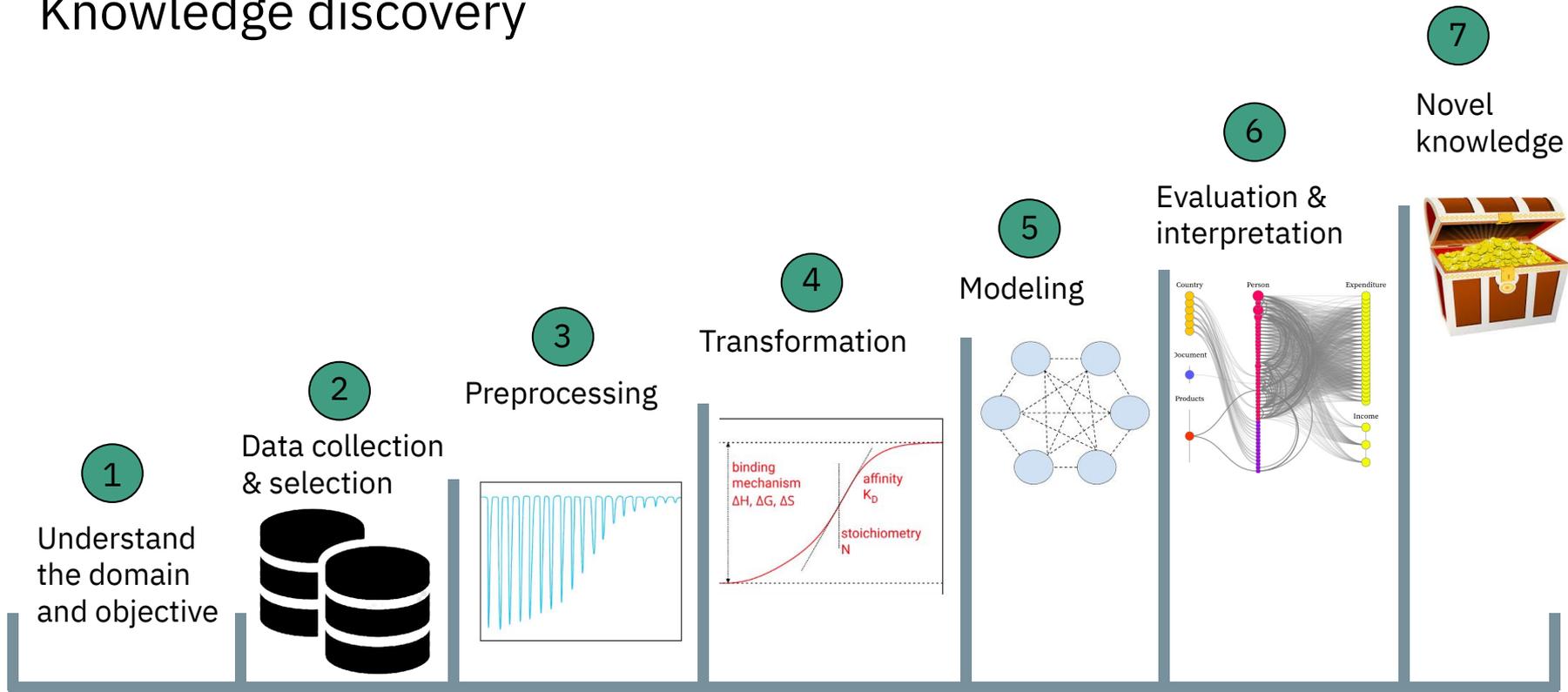
- Homogeneity
- Integrity
- Identity



4. Measure & analysis

- Activity / inhibition assays
- Structural studies

Knowledge discovery



Sample Preparation & Characterization (SPC) Facility

- Optimisation
- Quality control
- Characterisation (thermodynamics & kinetics)



MALDI TOF



Differential Scanning Fluorimetry



Circular Dichroism



Mass Photometry



Isothermal Titration Calorimetry



MicroScale Thermophoresis

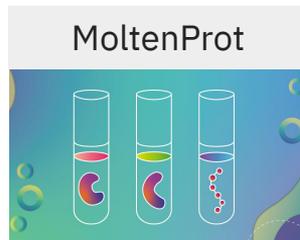


Dynamic Light Scattering

and much more!

eSPC, enriching service provision

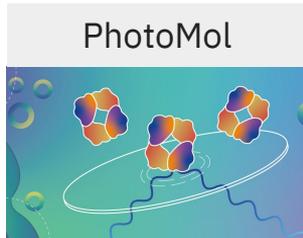
Differential Scanning
Fluorimetry (DSF)



MicroScale
Thermophoresis
(MST)



Mass Photometry
(MP)



Sample Preparation & Characterization (SPC) Facility

- Optimisation
- Quality control
- Characterisation (thermodynamics & kinetics)



MALDI TOF



Differential Scanning Fluorimetry



Circular Dichroism



Mass Photometry



Isothermal Titration Calorimetry



MicroScale Thermophoresis



Dynamic Light Scattering

and much more!

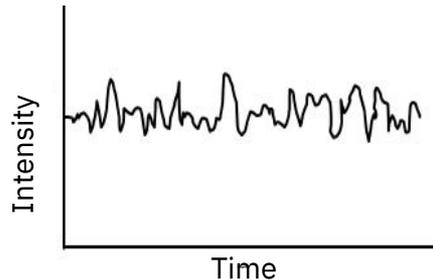
Dynamic light scattering (DLS) in a nutshell

Why?

- Homogeneity of a sample
- Estimation of the hydrodynamic radius (Hr)

How does it work?

- It measures the autocorrelation of the scattered light



Dynamic light scattering (DLS) in a nutshell

Why?

- Homogeneity of a sample
- Estimation of the hydrodynamic radius (Hr)

How does it work?

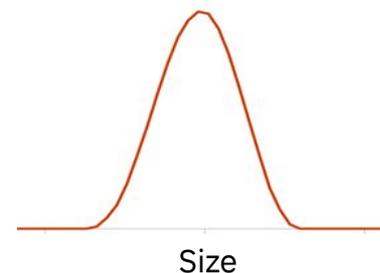
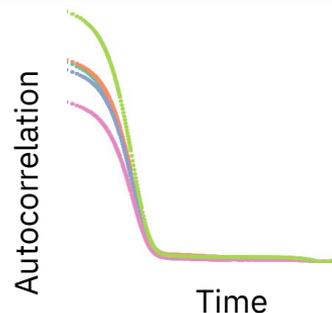
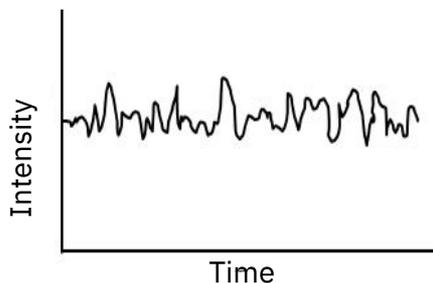
- It measures the autocorrelation of the scattered light

How do we extract valuable information?

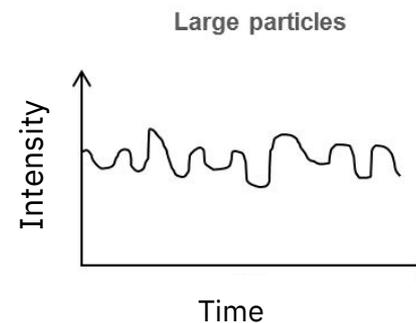
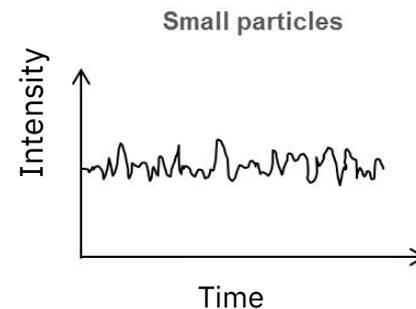
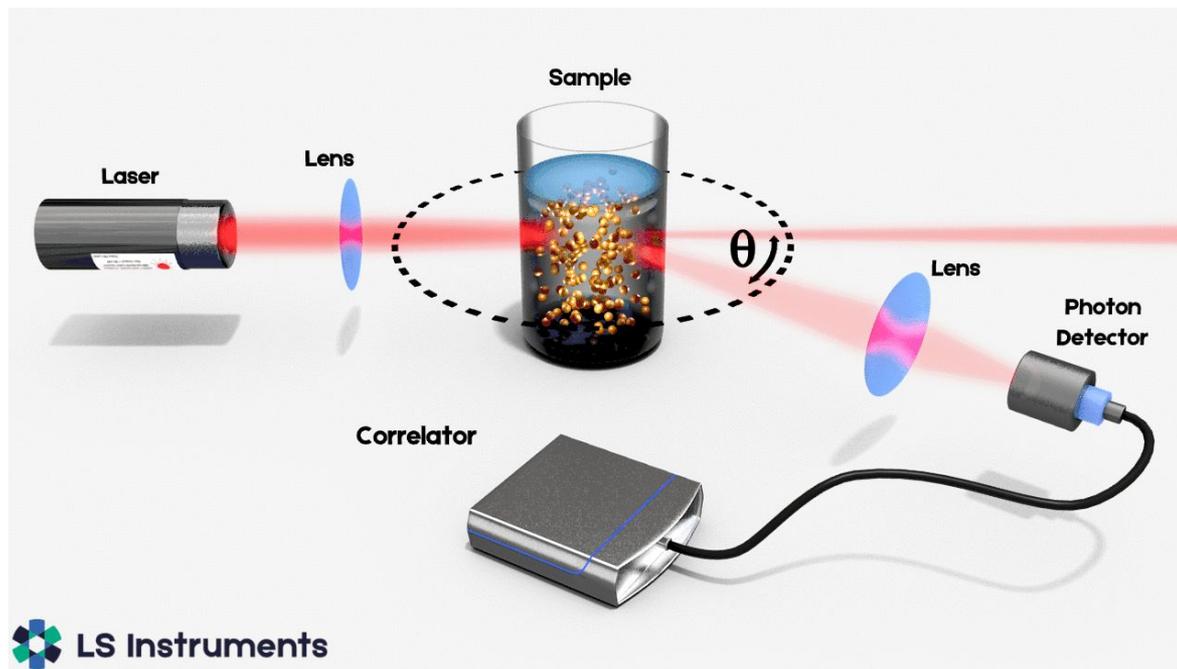
- We fit one/two, or a distribution of decay rates

Limitations

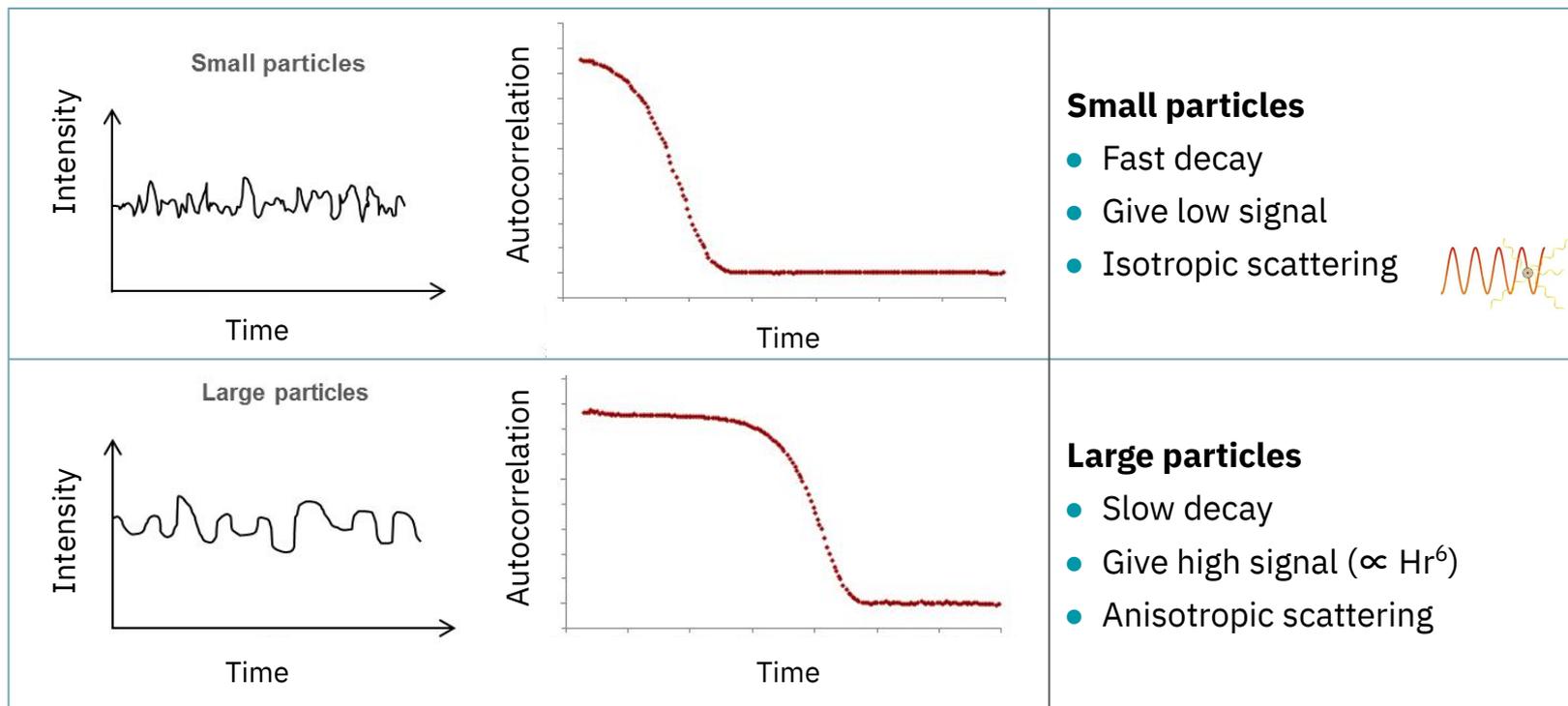
- Signal $\propto Hr^6$
- Semi-quantitative
- Scattering isn't isotropic for large particles



DLS - Theory



DLS - Theory



“Any model is at best, an
useful fiction”

DLS - Model & fitting

What we measure

Second order correlation
function G_2

$$G_2(\tau) = \langle I(t)I(t + \tau) \rangle$$

Integral over the product of
intensities at time t and delayed
time $t+\tau$

$$g_2(\tau) = \frac{\langle I(t)I(t + \tau) \rangle}{\langle I(t) \rangle^2}$$

DLS - Model & fitting

What we measure

Second order correlation function G_2

$$G_2(\tau) = \langle I(t)I(t + \tau) \rangle$$

Integral over the product of intensities at time t and delayed time $t+\tau$

$$g_2(\tau) = \frac{\langle I(t)I(t + \tau) \rangle}{\langle I(t) \rangle^2}$$

Relationship with particle motion

Normalised first-order correlation function g_1

$$g_2(\tau) = 1 + \beta |g_1(\tau)|^2$$

Coherence factor $\beta \propto$
instrument & molecules

DLS - Model & fitting

What we measure

Second order correlation function G_2

$$G_2(\tau) = \langle I(t)I(t + \tau) \rangle$$

Integral over the product of intensities at time t and delayed time $t+\tau$

$$g_2(\tau) = \frac{\langle I(t)I(t + \tau) \rangle}{\langle I(t) \rangle^2}$$

Relationship with particle motion

Normalised first-order correlation function g_1

$$g_2(\tau) = 1 + \beta |g_1(\tau)|^2$$

Coherence factor $\beta \propto$ instrument & molecules

Distribution of decay rates

Second order correlation function

$$g_1(\tau) = \int_0^\infty G(\Gamma) \exp(-\Gamma\tau) d\Gamma$$

Intensity-weighted integral over a distribution of decay rates

DLS - Model & fitting

Decay rate & diffusion rates

Each decay rate can be associated to a certain diffusion factor D

$$D(s, q) = 1/(s * (q^2))$$

where s is the inverse of the decay rate and q is the Bragg wave vector (\propto angle of detector & refractive index)

Conversion to hydrodynamic radius

Sphere-like model allows estimating the Hr

$$Hr = \frac{k_b * temperature}{D * viscosity * 6\pi}$$

The ISO recommended cumulants approach

- Moment analysis of the linear form of the measured correlogram
- Assumes a single particle family (Gaussian)
- Gives the Z-average (mean value) and the PdI (polydispersity index, relative variance of the Gaussian)

DLS - Model & fitting

Limitations of the cumulants approach

- Extremely sensitive to small amounts of aggregates
- Unsuitable for a polydisperse sample (polydispersity > 20 %)

Fitting a distribution of decay rates

- We need to define a decay rate space

$$g_1(t) = \sum_{i=1}^{200} c_i \exp\left(-\frac{t}{s_i}\right)$$

- Ill-posed problem that requires regularization

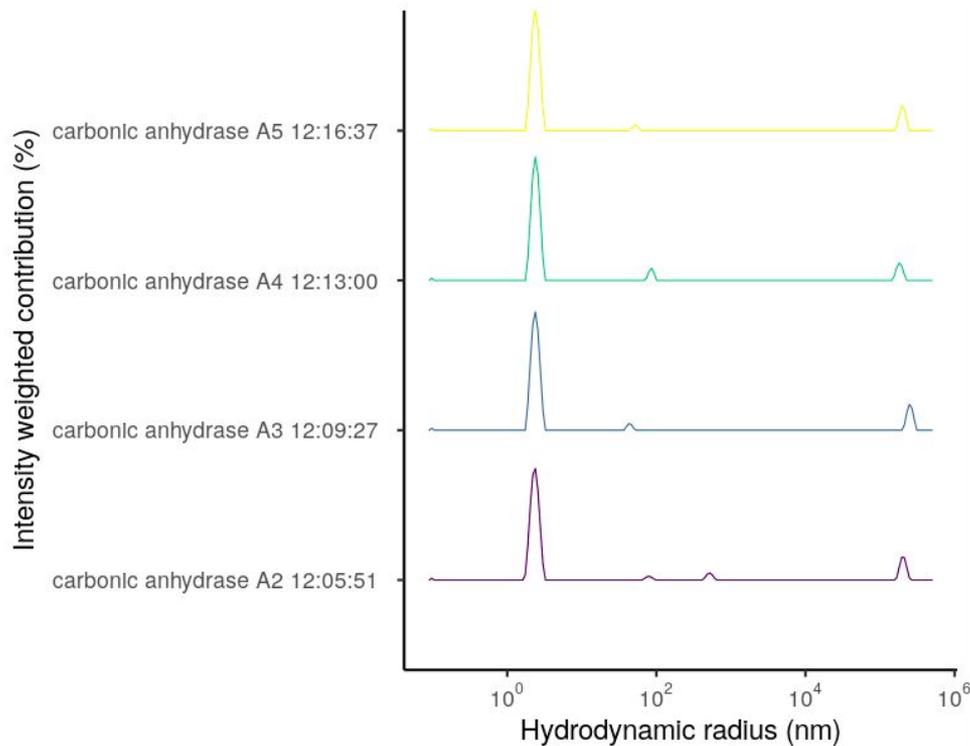
$$\|Ax - b\| + \alpha \|Mx\| + \beta \|Ix\|$$

$$\alpha \left\| \sum_{i=2}^{199} 2c_i - c_{i-1} - c_{i+1} \right\|$$



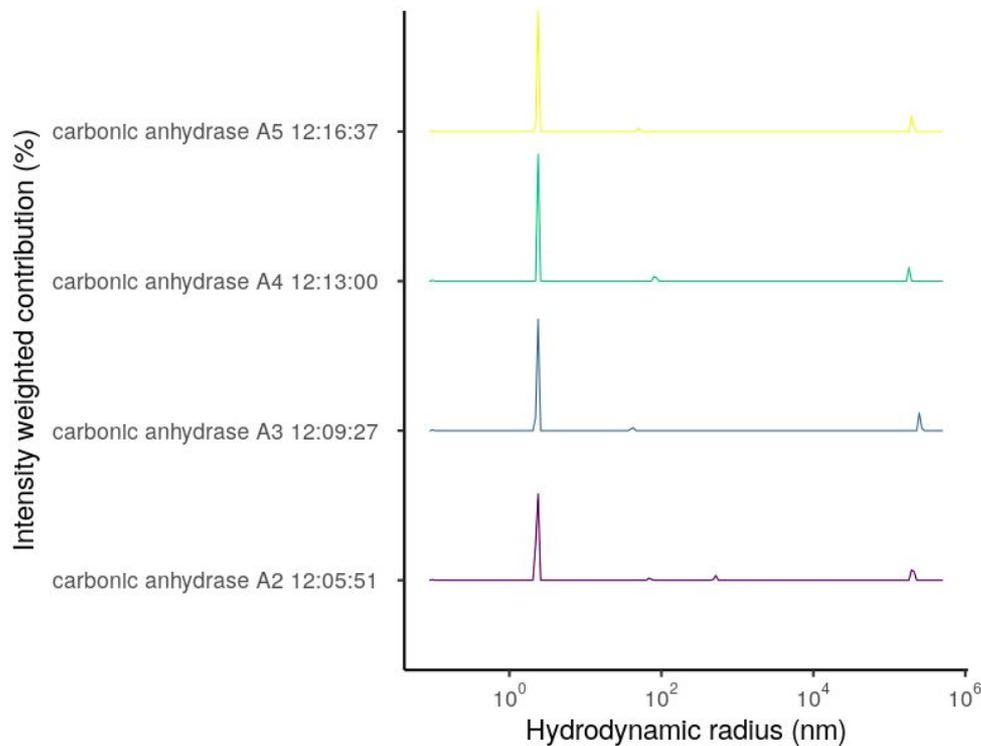
“DLS is (almost always)
semi-quantitative”

DLS - Fitting in practice



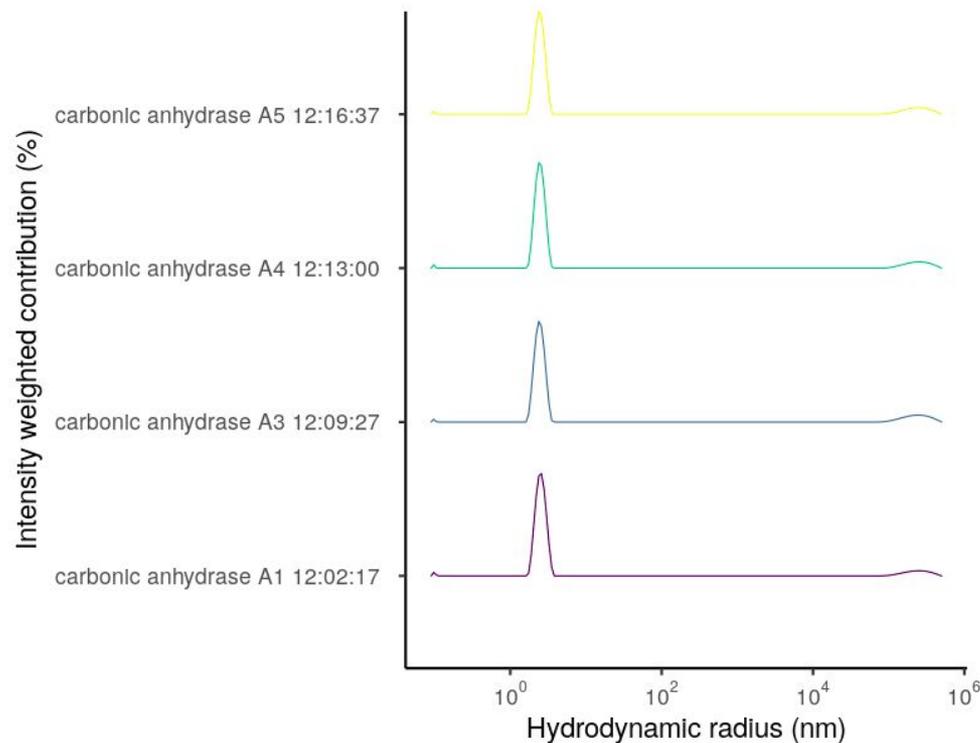
- Hr min value 0.09 nm
- Hr max value $5e5$ nm
- Number of Hr points 200
- Max time 1 sec
- Alpha regularization 0.1
- Beta regularization 0

DLS - Fitting in practice



● Hr min value	0.09 nm
● Hr max value	5e5 nm
● Number of Hr points	200
● Max time	1 sec
● Alpha regularization	0
● Beta regularization	0

DLS - Fitting in practice



- Hr min value 0.09 nm
- Hr max value $5e5$ nm
- Number of Hr points 200
- Max time **100 μ s**
- Alpha regularization 0.1
- Beta regularization 0

DLS - Interpretation

Good samples

- Cumulants PdI < 20 % (Malvern)
- One peak in region 1-20 nm with mass > 99.9 %

Oligomerization

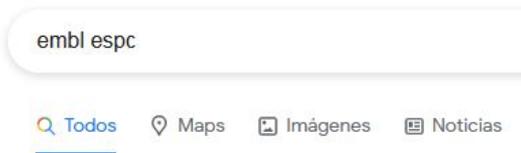
- Only big differences in size (factor 3-5) are detected, i.e., monomer to hexamer

Hr values

- The Hr values are semi-quantitative
- Hr estimation assumes sphere-like model

Raynals, an app for DLS analysis

Step 1. Google “embl espc”



Cerca de 6,760 resultados (0,37 segundos)

<https://spc.embl-hamburg.de> Traducir esta página

SPC Webserver - EMBL Hamburg

eSPC, an Online Data Analysis Platform for Molecular Biophysics
easy to use software for the understanding of biophysical experiments.
Visitaste esta página varias veces. Última visita: 12/09/2024

<https://www.embl.org/groups/se...> Traducir esta página

Services and Resources – Sample Preparation

eSPC data analysis platform for molecular biophysics ... |
resources that SPC at EMBL-Hamburg could offer you at

Step 2. Access spc.embl-hamburg.de

Directory Search

Characterisation > SPC Services > SPC Data Analytics Related: EMBL Hamburg

eSPC, an Online Data Analysis Platform for Molecular Biophysics

The eSPC platform provides easy to use software for the understanding of biophysical experiments.

FoldAffinity MoltenProt

Step 3. Access Raynals spc.embl-hamburg.de/app/raynals



Raynals, analysis code available under request

```
W      = np.arange(1, len(data)+1) * 0 + 0.1 # all weights are equal, except the initial and last value
W      = W / np.max(W)
W      = np.append(W, np.array([1e2, 1e2, 1e2])) # weight to force the initial and last values equal to 0, and the sum of contributions equal to 1

rowToForceInitialValue = np.zeros(kernel.shape[1])
rowToForceInitialValue[0] = 1
rowToForceLastValue    = np.flip(rowToForceInitialValue)

data   = np.sqrt(W) * np.append(data, np.array([1, 0, 0]))
data   = data.reshape(-1, 1)

kernel = np.vstack([kernel, np.ones(kernel.shape[1]), rowToForceInitialValue, rowToForceLastValue])
kernel = np.sqrt(W[:, None]) * kernel

cols   = kernel.shape[1]

M = np.zeros((cols, cols))
for i in range(1, M.shape[1]-1):
    M[i, i-1] = -1
    M[i, i]   = 2
    M[i, i+1] = -1

L      = alpha * M
C      = np.concatenate([kernel, L], axis=0)
d      = np.concatenate([data, np.zeros(cols).reshape(-1, 1)])

I      = beta * np.eye(*kernel.shape)
C      = np.concatenate([C, I], axis=0)
d      = np.concatenate([d, np.zeros_like(data)])
x, _   = nnls(C, d.flatten())

return x
```

DLS - beyond the monodisperse / polydisperse sample

Global Analysis of Dynamic Light Scattering Autocorrelation Functions

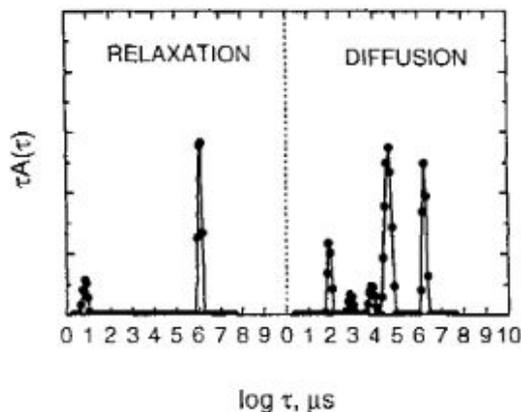
Stephen W. Provencher*, Petr Štěpánek**

$$\hat{g}_1(t) = \int A(\tau) e^{-t/\tau} d\tau.$$

$A(\tau)$ is related to decay rates that can be converted into diffusion coefficients

$$\beta(q)\hat{g}_1(t; q) = \int_0^\infty A_r(\tau) e^{-t/\tau} d\tau + \int_0^\infty A_d(D) e^{-q^2 D t} dD, \quad (2)$$

$$A(\tau; q) = A_r(\tau) + A_d[1/(q^2\tau)] \quad (3)$$



Decomposition into “diffusive” and “relaxational” (independent of angle) components

Sample Preparation & Characterization (SPC) Facility

- Optimisation
- Quality control
- Characterisation (thermodynamics & kinetics)



MALDI TOF



Differential Scanning Fluorimetry



Circular Dichroism



Mass Photometry



Isothermal Titration Calorimetry



MicroScale Thermophoresis



Dynamic Light Scattering

and much more!

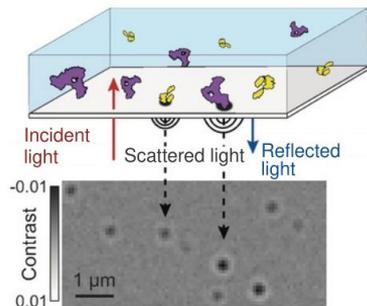
Mass photometry (MP) in a nutshell

Why?

- Homogeneity of a sample
- Estimation of the molecular masses of different species

How does it work?

- Interference between scattered and reflected light combined with ratiometric imaging



Mass photometry (MP) in a nutshell

Why?

- Homogeneity of a sample
- Estimation of the molecular masses of different species

How does it work?

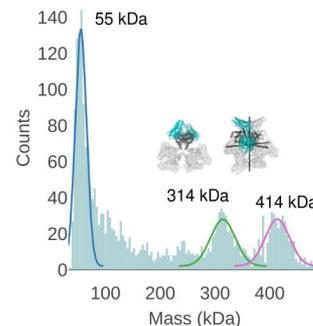
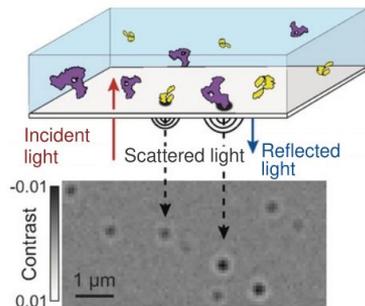
- Interference between scattered and reflected light combined with ratiometric imaging

How do we extract valuable information?

- We compare contrasts with known samples
- We fit n distributions of masses

Limitations

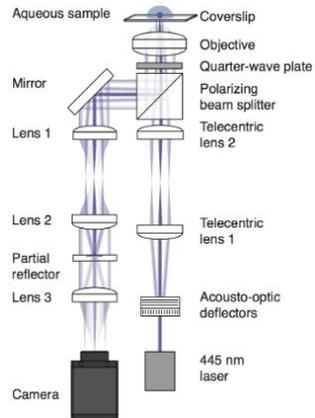
- nM concentration is required
- Detergent produces high background
- Accurate only with soluble proteins



MP - Theory

Experimental setup

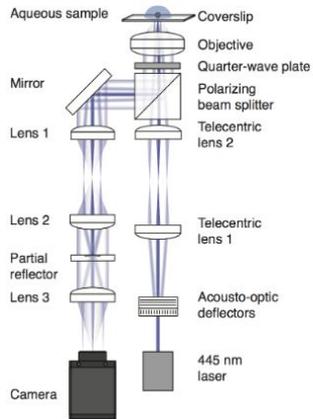
- Separation of incident from scattered and reflected light



MP - Theory

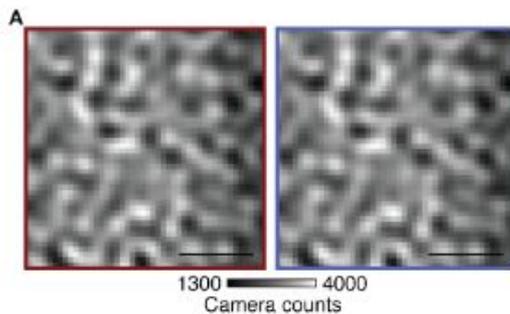
Experimental setup

- Separation of incident from scattered and reflected light



Raw images

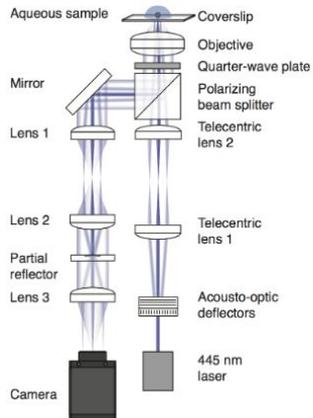
- Interference: scattered - reflected
- High background



MP - Theory

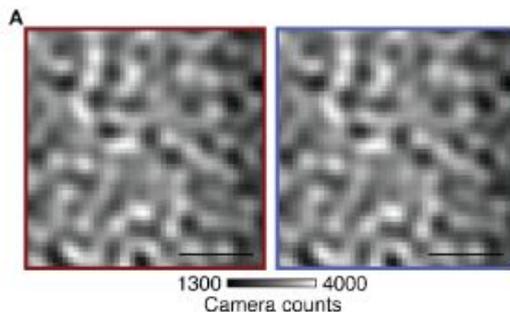
Experimental setup

- Separation of incident from scattered and reflected light



Raw images

- Interference: scattered - reflected
- High background



Preprocessing

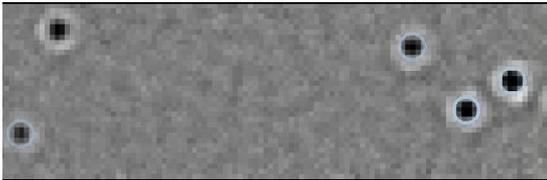
- Background is removed by conversion to ratiometric images



MP - Theory

Binding events

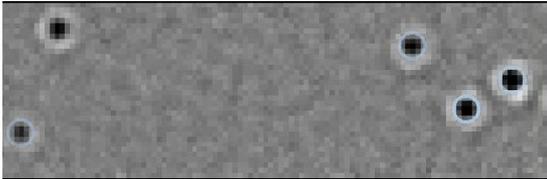
- Appear as a (dark) point spread function (PSF)
- The number should be \gg unbinding events (bright spots)



MP - Theory

Binding events

- Appear as a (dark) point spread function (PSF)
- The number should be \gg unbinding events (bright spots)



Particle detection & quantification

- Automated spot detection routine
- Fitting of candidate pixels ($772 \times 772 \text{ nm}^2$) to a 2D concentric gaussian model

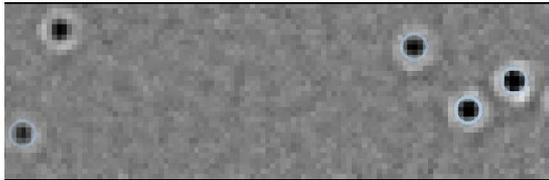
$$f(x, y) = A \left(e^{-\left[\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2} \right]} - \frac{(1-T)}{s} e^{-\left[\frac{(x-x_0)^2}{2(s\sigma_x)^2} + \frac{(y-y_0)^2}{2(s\sigma_y)^2} \right]} \right) + b$$

$$A \left(1 - \frac{(1-T)}{s} \right)$$

MP - Theory

Binding events

- Appear as a (dark) point spread function (PSF)
- The number should be \gg unbinding events (bright spots)



Particle detection & quantification

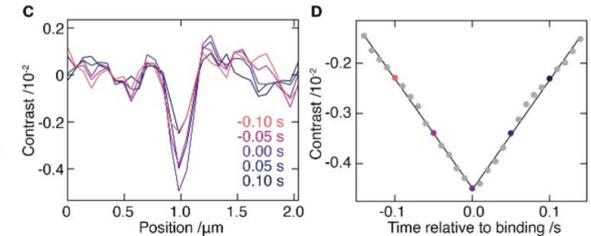
- Automated spot detection routine
- Fitting of candidate pixels ($772 \times 772 \text{ nm}^2$) to a 2D concentric gaussian model

$$f(x, y) = A \left(e^{-\left[\frac{(x-x_0)^2}{2\sigma_x^2} + \frac{(y-y_0)^2}{2\sigma_y^2} \right]} - \frac{(1-T)}{s} e^{-\left[\frac{(x-x_0)^2}{2(s\sigma_x)^2} + \frac{(y-y_0)^2}{2(s\sigma_y)^2} \right]} \right) + b$$

$$A \left(1 - \frac{(1-T)}{s} \right)$$

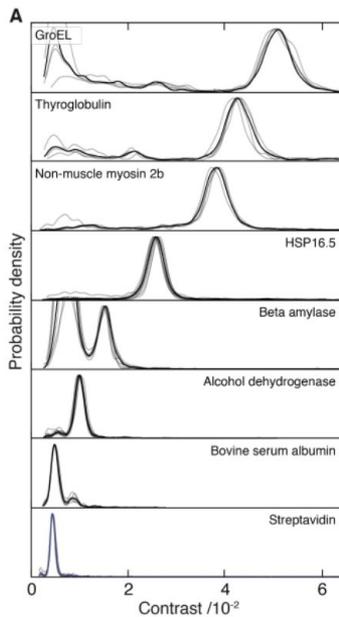
Contrast of a given particle

- Fitted from the contrast as a function of time



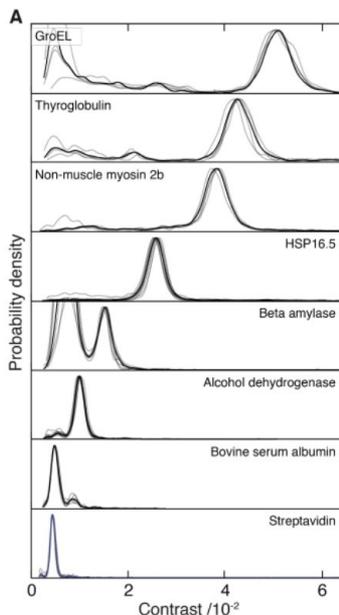
MP - Model & fitting

Contrast to molecular weight



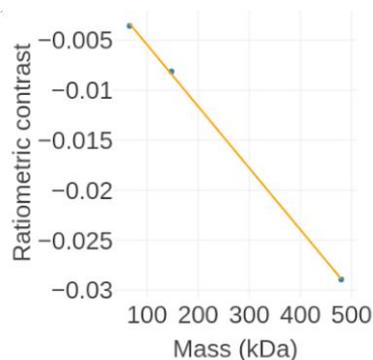
MP - Model & fitting

Contrast to molecular weight



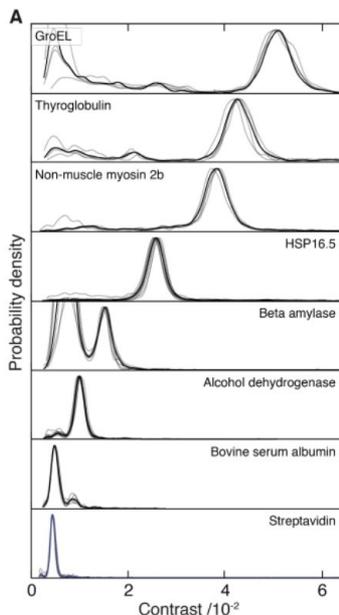
Calibration

- Using known protein standards
- Quantitative for soluble proteins without post translational modifications!



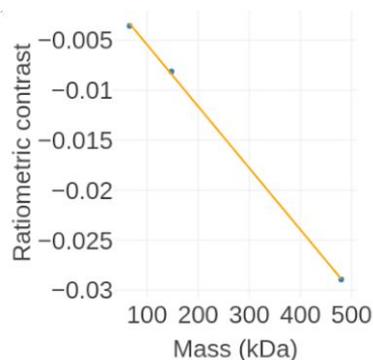
MP - Model & fitting

Contrast to molecular weight



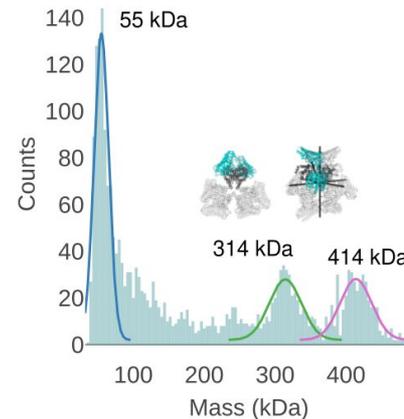
Calibration

- Using known protein standards
- Quantitative for soluble proteins without post translational modifications!



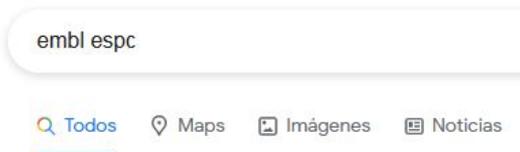
Unknown sample

- Fitting of gaussian distributions



PhotoMol, an app for MP analysis

Step 1. Google “embl espc”



Cerca de 6,760 resultados (0,37 segundos)

<https://spc.embl-hamburg.de> Traducir esta página

SPC Webserver - EMBL Hamburg

eSPC, an Online Data Analysis Platform for Molecular Biophysics. An easy to use software for the understanding of biophysical experiments. Visitaste esta página varias veces. Última visita: 12/09/21

<https://www.embl.org/groups/se...> Traducir esta página

Services and Resources – Sample Prep

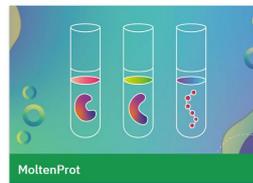
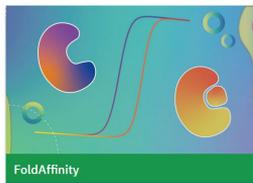
eSPC data analysis platform for molecular biophysics ... | resources that SPC at EMBL-Hamburg could offer you a

Step 2. Access spc.embl-hamburg.de

Characterisation > SPC Services > SPC Data Analysis Related: EMBL Hamburg

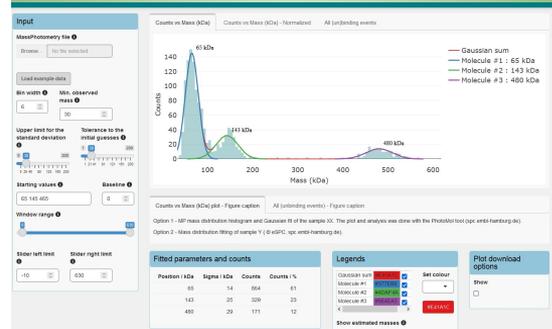
eSPC, an Online Data Analysis Platform for Molecular Biophysics

The eSPC platform provides easy to use software for the understanding of biophysical experiments.



Directory Search

Step 3. Access PhotoMol spc.embl-hamburg.de/app/photoMol

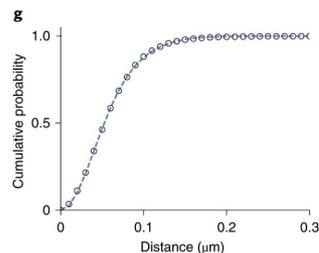
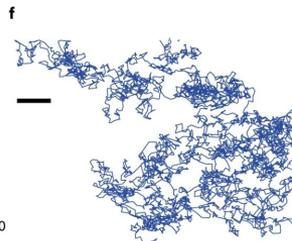
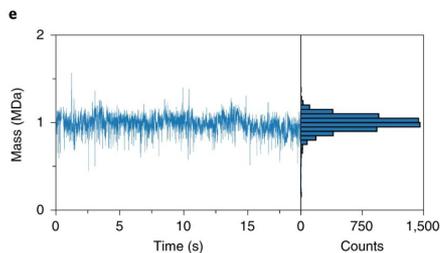
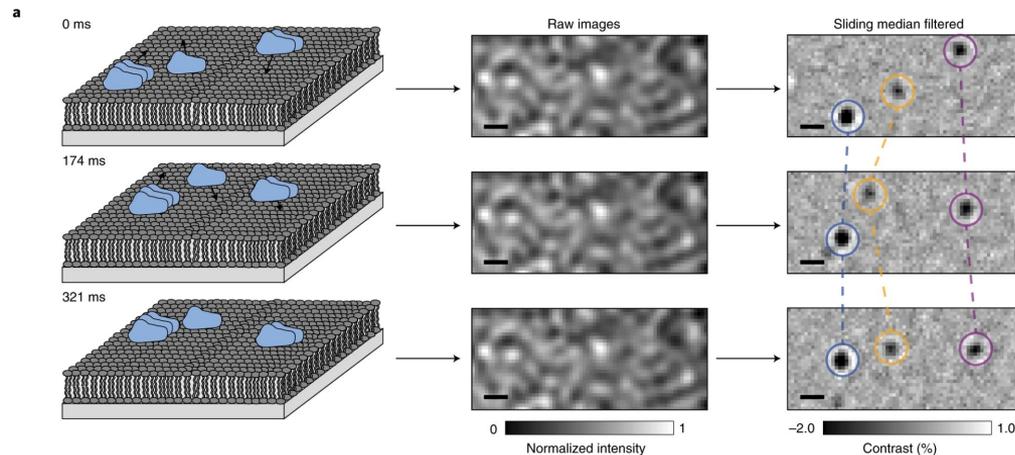


MP - beyond the yes/no complex formation question

Mass photometry enables label-free tracking and mass measurement of single proteins on lipid bilayers

[Eric D. B. Foley](#), [Manish S. Kushwah](#), [Gavin Young](#) & [Philipp Kukura](#) 

Simultaneously imaging, tracking and measuring the mass of diffusing biomolecular complexes on supported lipid bilayers (SLBs)

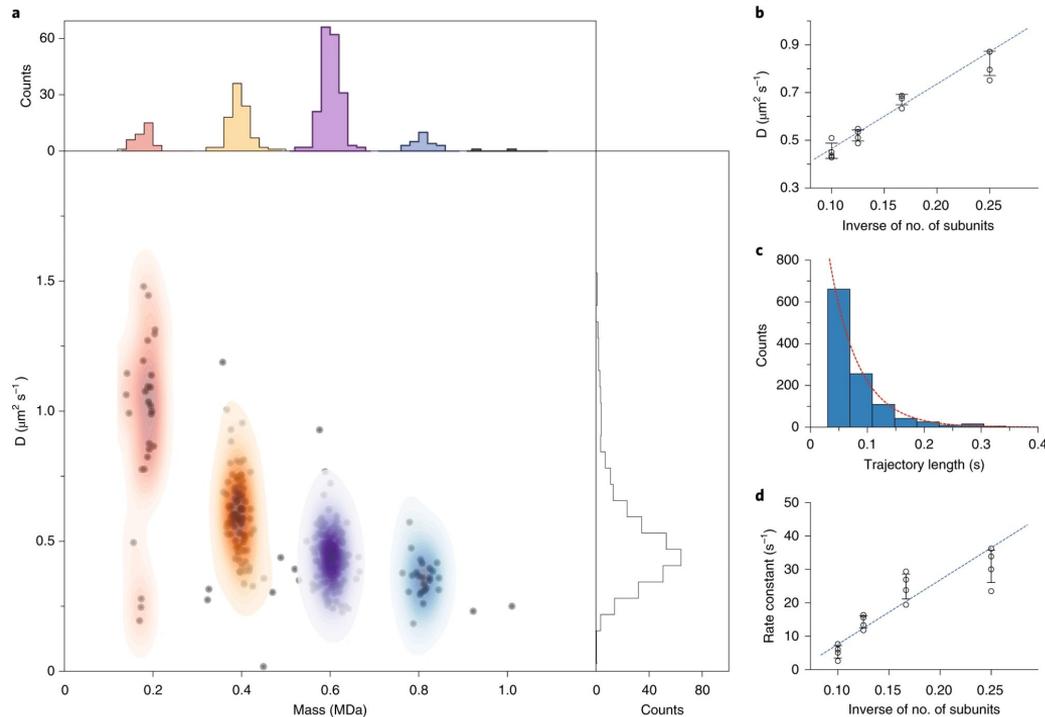


MP - beyond the yes/no complex formation question

Mass photometry enables label-free tracking and mass measurement of single proteins on lipid bilayers

[Eric D. B. Foley](#), [Manish S. Kushwah](#), [Gavin Young](#) & [Philipp Kukura](#) 

Simultaneously imaging, tracking and measuring the mass of diffusing biomolecular complexes on supported lipid bilayers (SLBs)



Sample Preparation & Characterization (SPC) Facility

- Optimisation
- Quality control
- Characterisation (thermodynamics & kinetics)



MALDI TOF



Differential Scanning Fluorimetry



Circular Dichroism



Mass Photometry



Isothermal Titration Calorimetry



MicroScale Thermophoresis



Dynamic Light Scattering

and much more!

Nano Differential scanning fluorimetry (nDSF) in a nutshell

Why?

- Protein stability or ligand binding

How does it work?

- By heating the sample and measuring the fluorescence

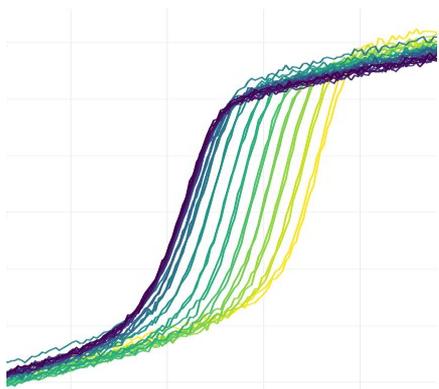


Photo credit: Vadim Kotov

Nano Differential scanning fluorimetry (nDSF) in a nutshell

Why?

- Protein stability or ligand binding

How does it work?

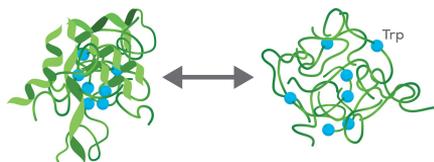
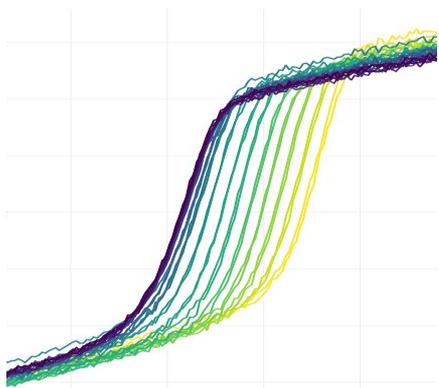
- Heating the sample and measuring the fluorescence

How do we extract valuable information?

- Fitting unfolding models

Limitations

- Autofluorescence
- Multiple transitions
- No transitions



Native \rightleftharpoons Unfolded

Photo credit: Vadim Kotov

Protein image: unchainedlabs.com

nDSF - Unfolding models

n-mer	Protomers	States	Intermediates
Monomer	A	2	none
Monomer	A	3	1 monomer
Monomer	A	2+p	p monomers
Homodimer	A ₂	.	.
.	.	.	.
Heterodimer	AB	.	.
.	.	.	.
Heterotrimer	ABC	.	.
.	.	.	.

Native → Unfolded

Native ⇌ Unfolded

Native ⇌ Intermediate ⇌ Unfolded

nDSF - How to build a model

Step 1. Equation of the signal with n states

$$Y_i(x) = Y_n[\text{N}] + \sum_j Y_j[\text{I}_j] + Y_u[\text{U}] + Y_d(x), \quad j = 1, \dots, p$$

Step 2. Pre/post transition dependence

$$Y_n = y_n + m_n x;$$

Step 3. States interconversion

Native \rightleftharpoons Unfolded

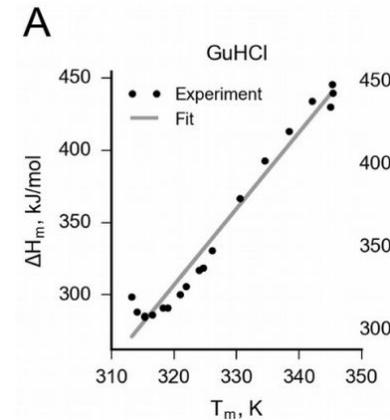
Send us code/equations & data and we will add them into the server!

nDSF - The two state equilibrium model

- Reversible equilibrium between native & unfolded states

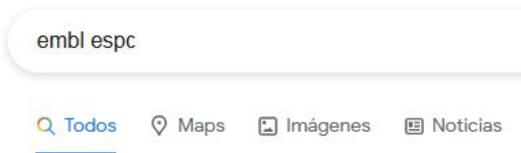
$$\Delta G(T) = -RT \ln [K(T)] = \Delta H_m (1 - T/T_m) - \Delta C_p [(T_m - T) + T \ln (T/T_m)]$$

- ΔH and T_m can be precisely determined (but not ΔC_p)
- ΔC_p can be determined using different chemical denaturant concentrations



MoltenProt & FoldAffinity, two apps for DSF analysis

Step 1. Google “embl espc”



Cerca de 6,760 resultados (0.37 segundos)

<https://spc.embl-hamburg.de> Traducir esta página

SPC Webserver - EMBL Hamburg

eSPC, an Online Data Analysis Platform for Molecular Biophysics
easy to use software for the understanding of biophysical experiments.
Visitaste esta página varias veces. Última visita: 12/09/2012

<https://www.embl.org/groups/se...> Traducir esta página

Services and Resources – Sample Preparation

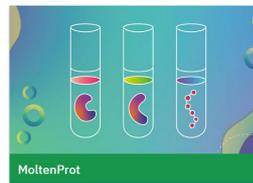
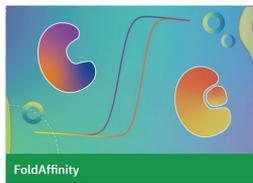
eSPC data analysis platform for molecular biophysics ... |
resources that SPC at EMBL-Hamburg could offer you at

Step 2. Access spc.embl-hamburg.de

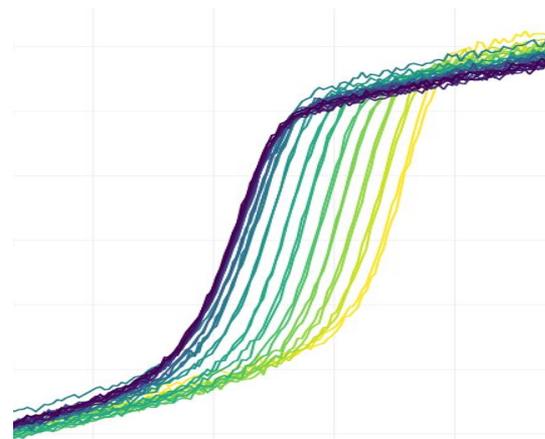
Characterisation > SPC Services > SPC Data Analytics Related: EMBL Hamburg

eSPC, an Online Data Analysis Platform for Molecular Biophysics

The eSPC platform provides easy to use software for the understanding of biophysical experiments.



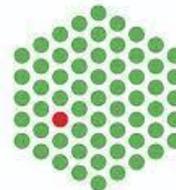
Step 3. Access MoltenProt or FoldAffinity



SPC Team



EMBL



European Molecular Biology Laboratory



EMBL Hamburg IT Team

This project has received funding from the European Union's Horizon 2020 research and innovation programme under the Marie Skłodowska-Curie grant agreement No 945405.