A perspective on the contribution of spectroscopy to characterising proteins for quality control

Alison Rodger, Macquarie University

23/5/22

FAAB
Facilitated Advancement of Australia's Bioactives
ARC TRAINING CENTRE

Simplistic view of Protein analysis

Mass spectrometry tells the **what** of proteins – sometimes with some interactions

Crystallography tells what it could look like (usually without any natively unfolded & flexible regions, perhaps with some extra bits)

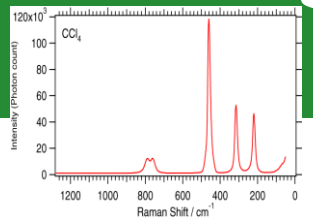
NMR tells what a smallish (~30 kDa but nearly up to 1 MDa) significantly ^{15}N and ^{13}C labelled sample looks like in an idealised solution

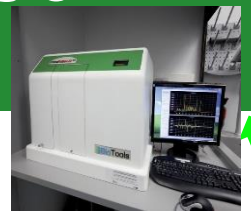
But what about in the **formulation vehicle** at the formulated concentration?

….. **Spectroscopy** (not NMR)

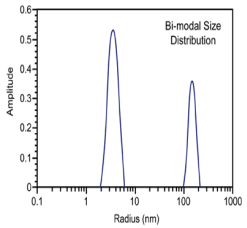
PDB 7M7W
Snell et al.

Spectroscopy is interaction between light & matter





Raman Spectroscopy

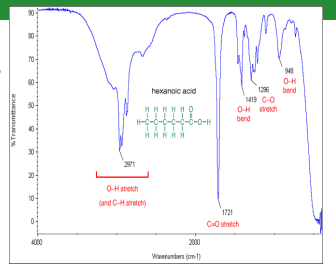


DLS

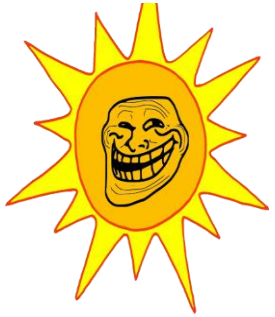
Scattering Spectroscopy

Absorption Spectroscopy

Vibrational realm



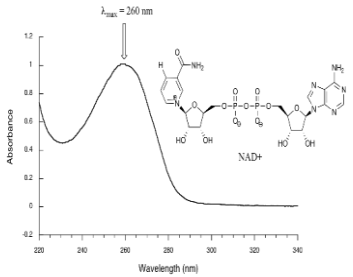
Infrared spectroscopy



Light source!

Sample!!!

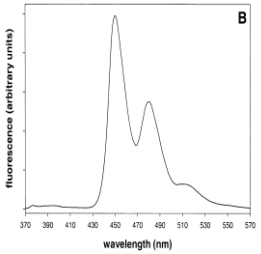
Electronic realm

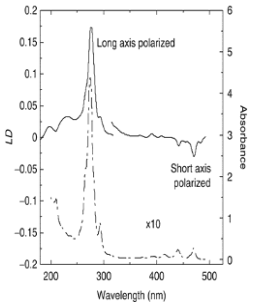


UV-vis spectroscopy

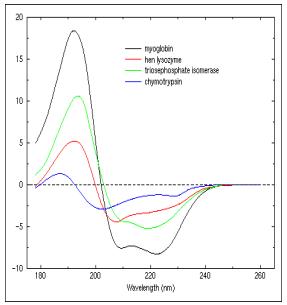
Emission Spectroscopy

Fluorescence Spectroscopy





LD



CD

Polarised UV-vis spectroscopy

Goal to read back from spectrum to something about the molecule

Absorption:

- Circular Dichroism
- Infrared Spectroscopy

Scattering:

- Raman Spectroscopy

(Emission:

- Fluorescence)

In passing

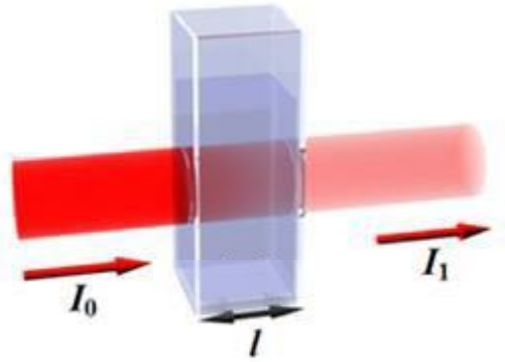
Absorbance Spectroscopies
UV-visible absorbance

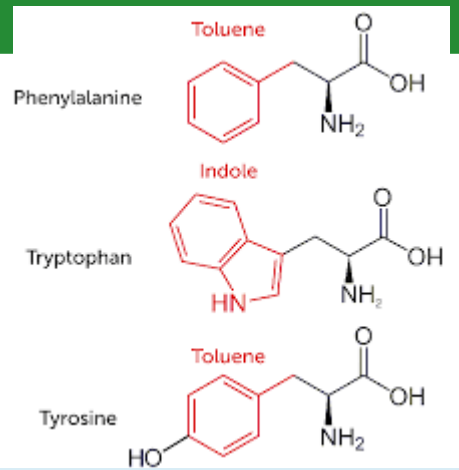
The Beer-Lambert Law:

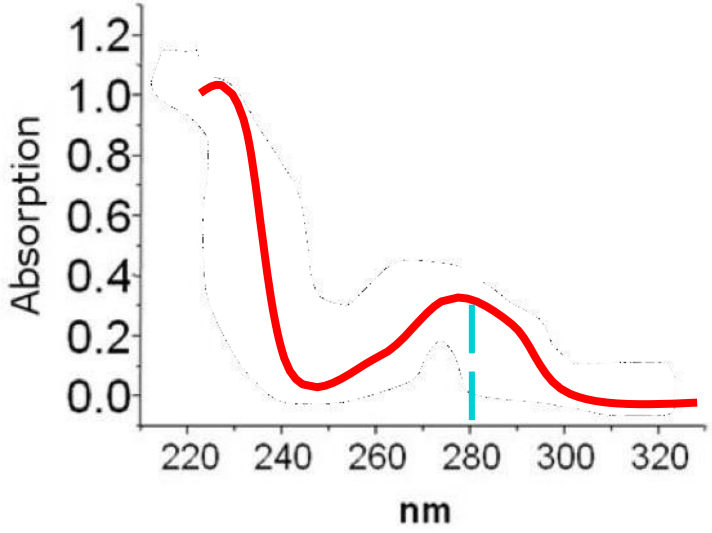
$$A_{280\,nm} = -\log(\frac{I_{out}}{I_{in}}) = \varepsilon c l$$

A = absorbance c = concentration

l = pathlength ε = extinction coefficient



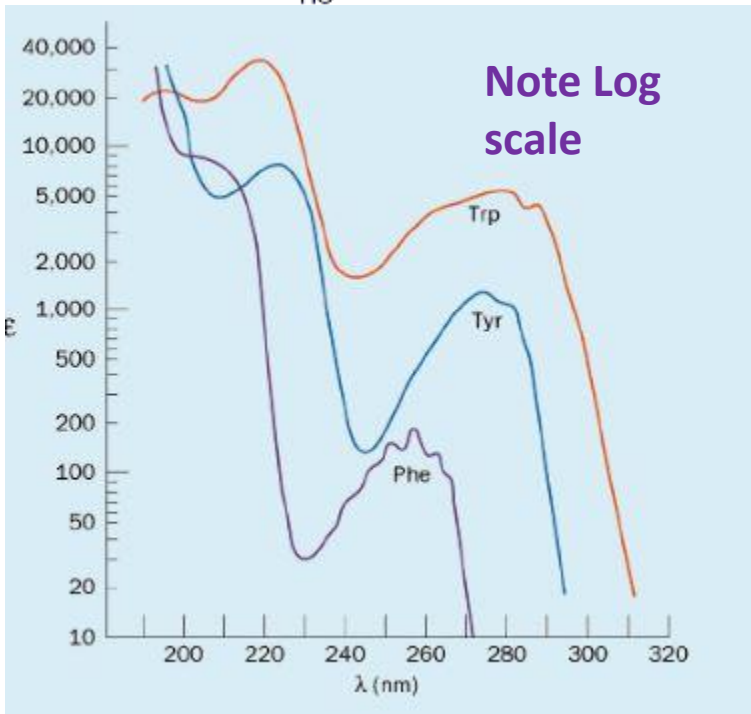




Estimates of ε_{280nm} based on primary sequence are generally quite accurate, especially for unfolded proteins.

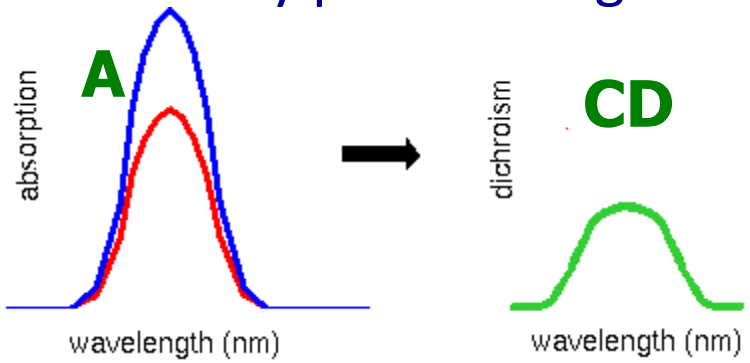
Need to guess S=S.

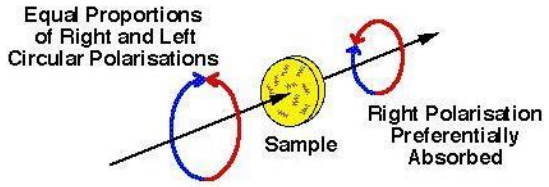
Absorbance of 1 mg/mL=1 is not very accurate.



Circular Dichroism: absorption spectroscopy

- CD is the difference between the absorption of left and right handed circularly polarized light as a function of wavelength.





- The difference very small (~<<1/1000 of total)

$\Delta A(\lambda) = A_L(\lambda) - A_R(\lambda) = [\varepsilon_L(\lambda) - \varepsilon_R(\lambda)]lc$ *or*

$\Delta A(\lambda) = \Delta\varepsilon(\lambda)lc$

- $\Delta\varepsilon$ ~ typically < 10 $mol^{-1}dm^3cm^{-1}$ *vs.* ε ~20,000 $M^{-1}cm^{-1}$

CD probes helicity — chirality — asymmetry

and hence molecular structure

Empirical analysis with CD

Identify something is chiral

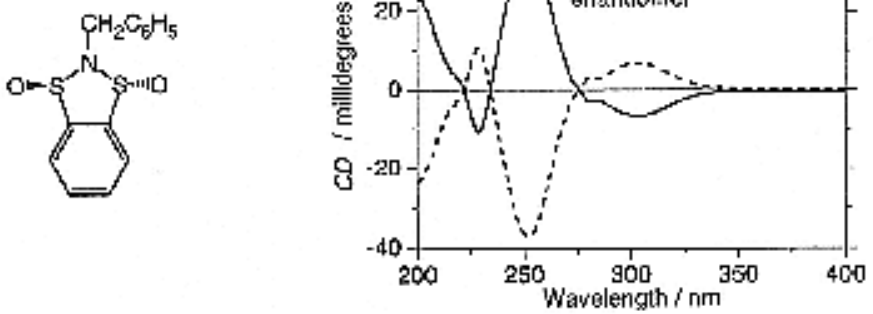
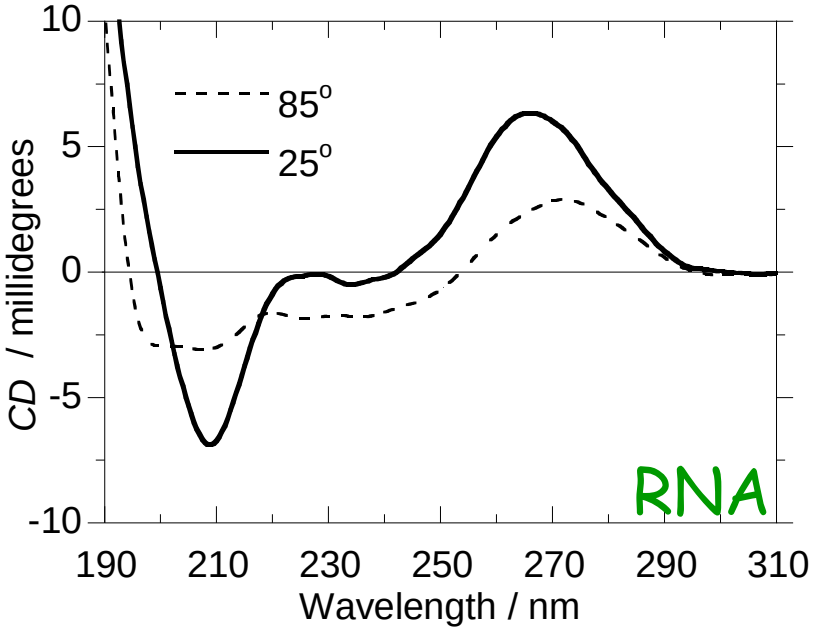


Fig. 5.3 *CD* spectra in acetonitrile for two 1,3,2-benzodithiazole-*S*-oxide enantiomers eluting successively off a chiral HPLC column, thus confirming the successful resolution of the enantiomers from a racemic mixture by HPLC.[2] The 1*R*,3*R* enantiomer, which is illustrated, is eluted from the column first.

HPLC chiral detector

Spot a structural change

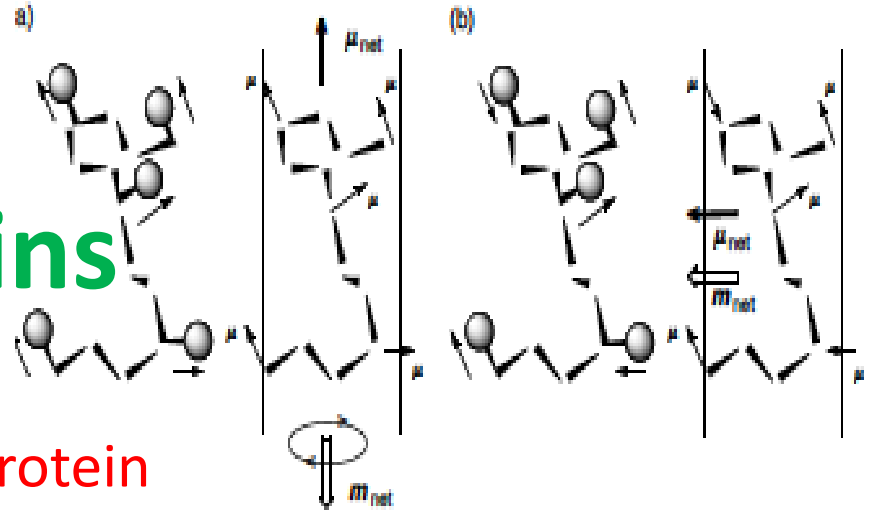


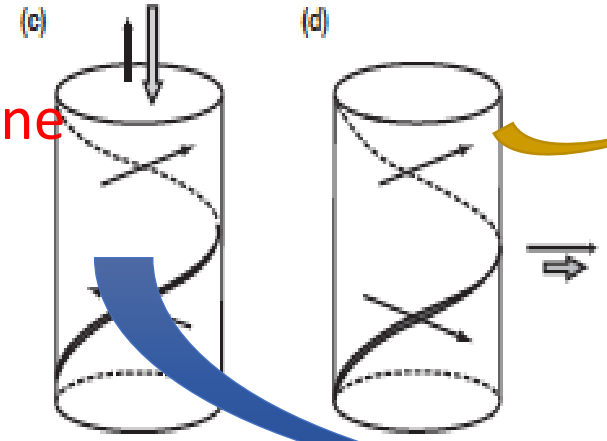
CD of 372 base mRNA as a function of temperature

Proteins

Different protein secondary structures have different backbone spectra.

Side chain CD probes environment.

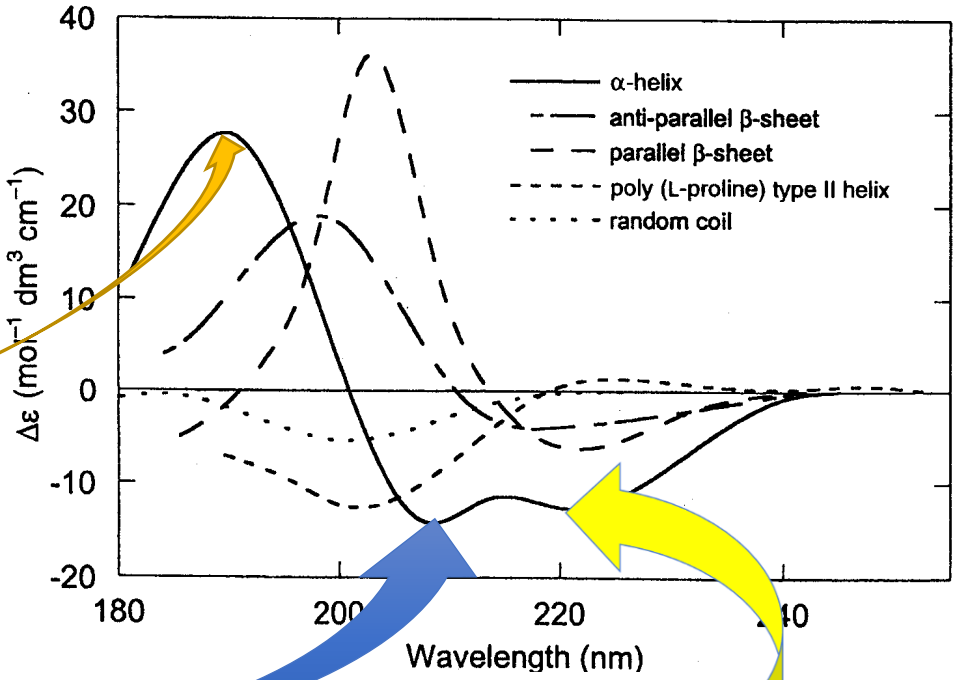




gure 8.8 Schematic illustration of the backbone and carbonyl groups (oxygens indicated by balls) of an α-helical pep proximately indicating (a) head-to-tail and (b) head-to-head couplings of the peptide electric dipole transition moments e resulting net electric and magnetic dipole transition moments. An alternative somewhat simplified model is to consider π* transitions of only two amide chromophores on a right-handed α-helix as shown in (c) and (d). By applying the six si ctlined in §4.7, the qualitative features of the exciton *CD* and *LD* spectra immediately emerge.

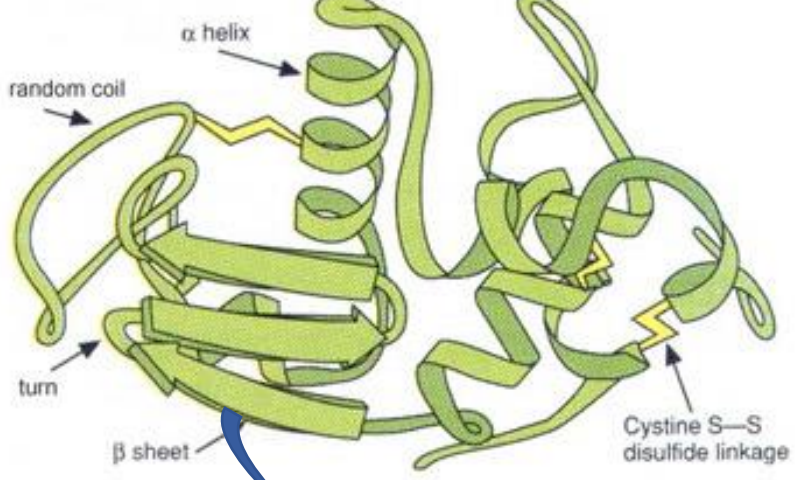
π-π amide transition

In α-helix



n-π amide transition: has magnetic component, 'borrows' electric

Protein CD



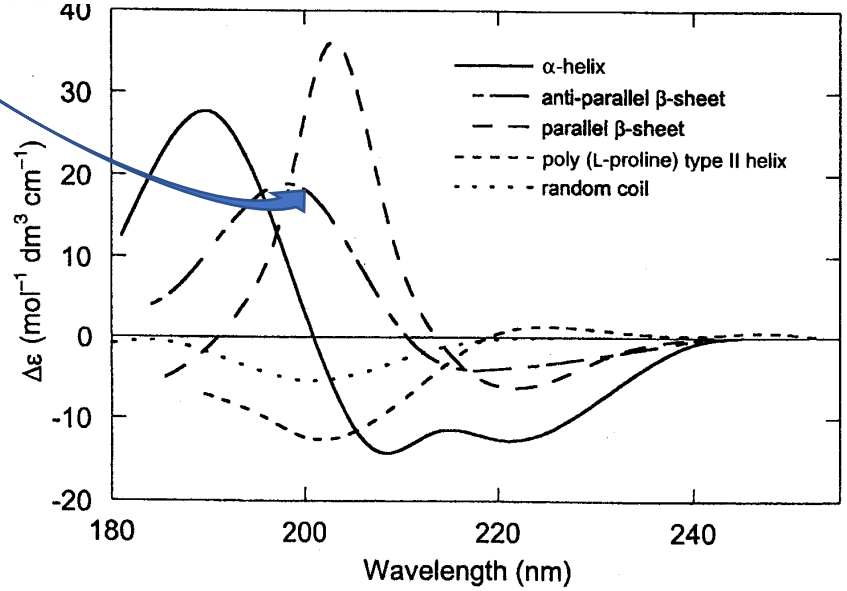


Figure 15 Typical protein CD spectra for particular secondary structural motifs protein structure fitting programs.

α-helical protein spectra are distinctive :222(–),208(–),190(+) nm

β-strand: ~216 nm (–), ~199 (+)

Other motifs also have well-defined spectra

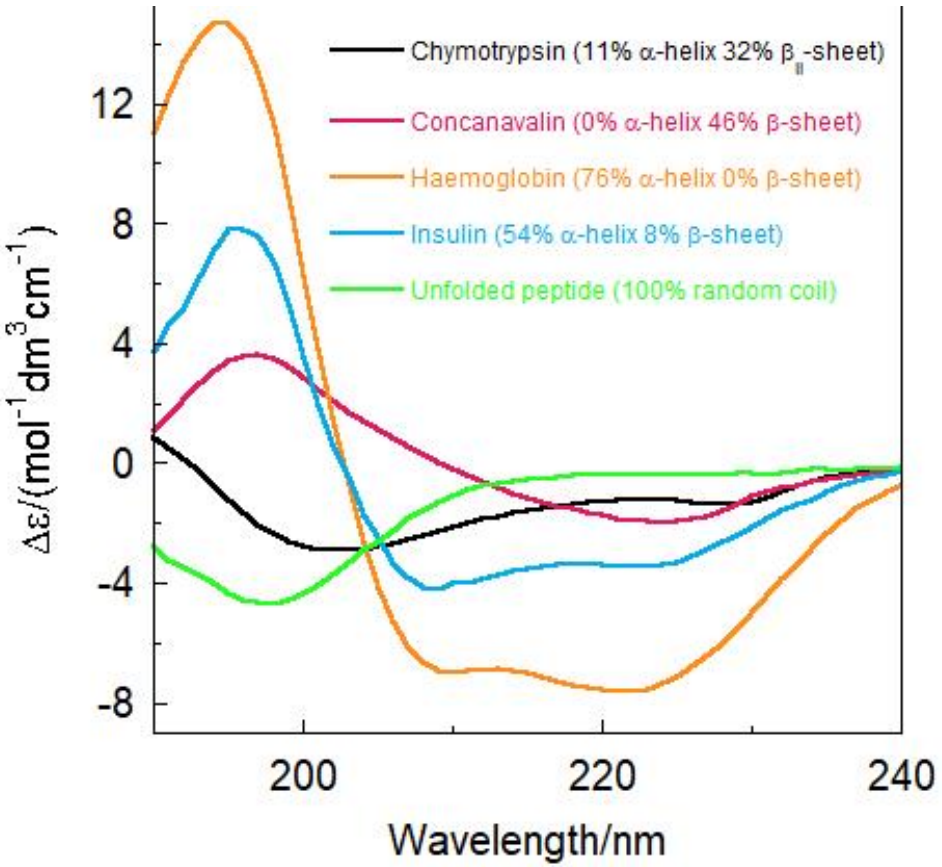
CD spectra depend on **AVERAGE** solution phase protein structure

Use to determine how environment — temperature, pH, solvent, ionic strength, denaturing agents — alters protein structure. Also binding constants.

Quick easy experiment that does not consume sample

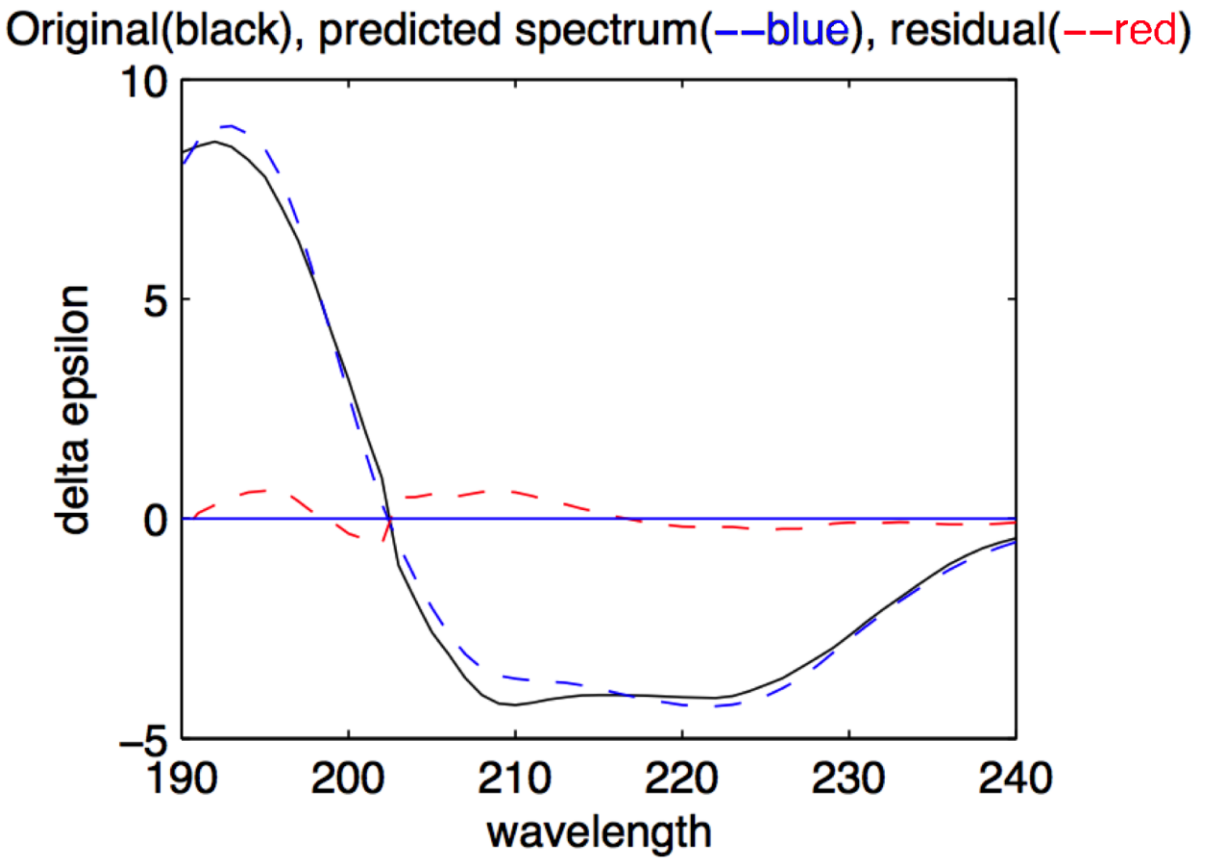
How to extract a structure summary?

- CDsstr (Johnson et al.) – spectra fits too good
- SELCON3 (Woody et al.) – self consistent variable selection approach, works well, needs a good reference set, available on DichroWeb
- SOMSpec (Rodger et al.) – self organising map approach, works well, needs a good reference set
  - Advantages of SOMSpec – we have the code, more detailed output is provided so it can be interrogated.



A SOM — SomSpec
(Secondary Structure Neural Network)

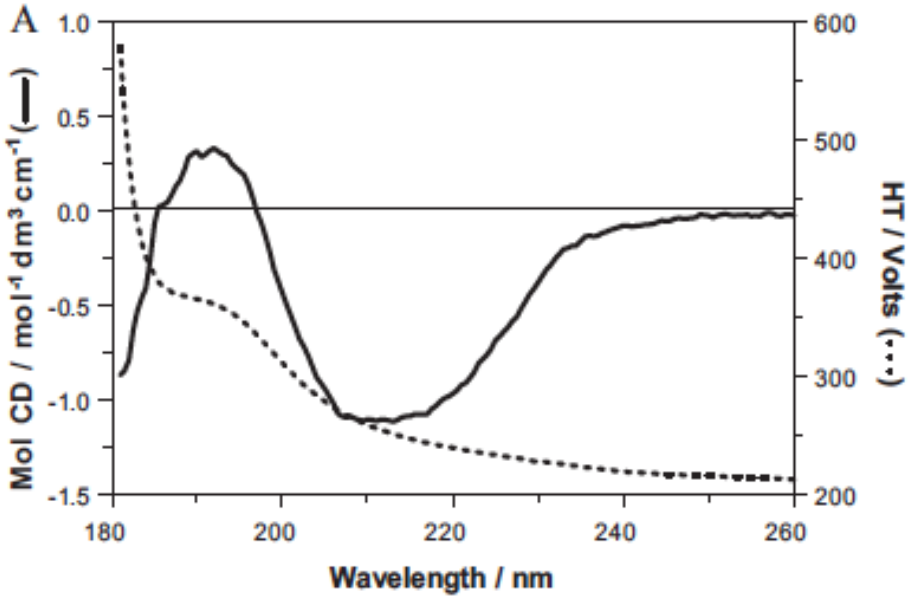
- A program to sort CD data into regions of like spectra/structure
  1. Make a spectra map of random values, unordered
  2. Take some reference CD spectra
  3. For each spectrum find vector with most similar numbers
  4. Modify numbers in random map to resemble spectrum more.
  5. Do same for neighbourhood of selected spectrum
  6. Fill in the missing regions with virtual spectra of intermediate values
  7. Create a matching structure map



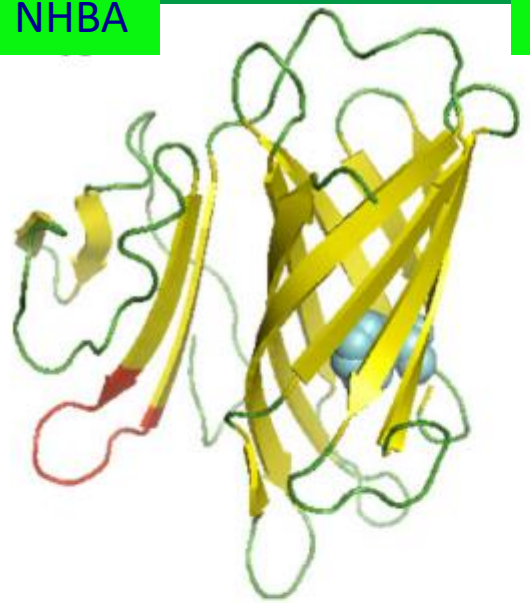
A new generation multicomponent meningococcal serogroup B vaccine

Fusion protein of two proteins NHBA and GNA1030

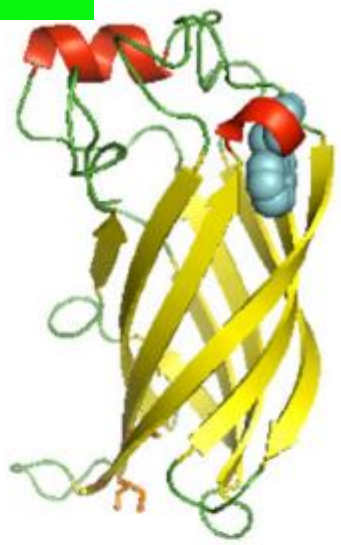
- **Homology modelling** analysis suggests each forms an 8-stranded β-barrel
- **CD is consistent** with this: 4% helix, 33% sheets, 61% other (inc. unfolded



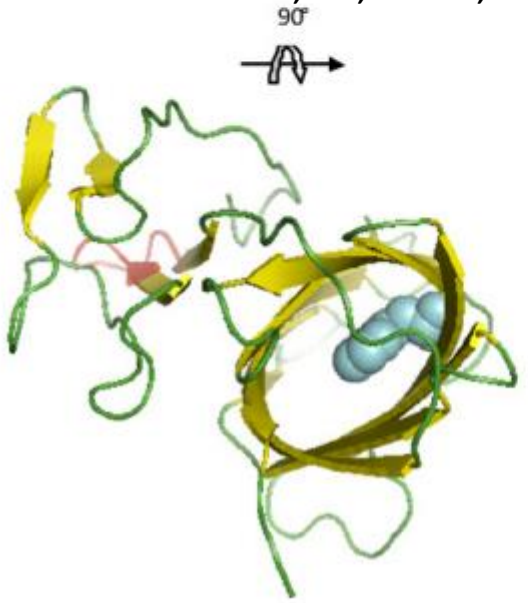
NHBA



GNA1030



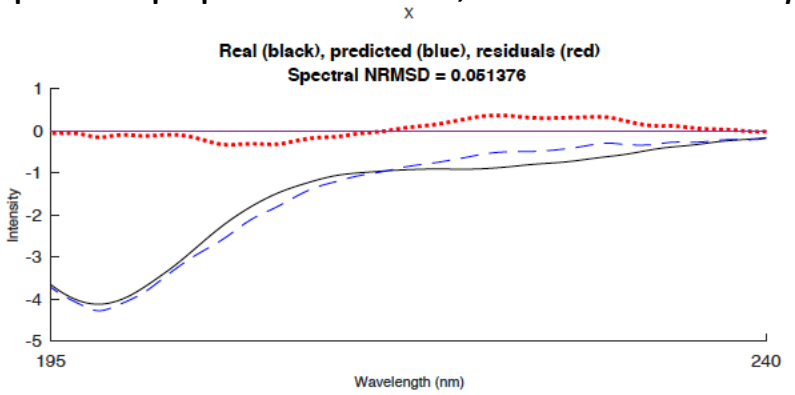
Martino, A.; et al., *Vaccine* **2012,** *30* (7), 1330-42.



MACQUARIE University

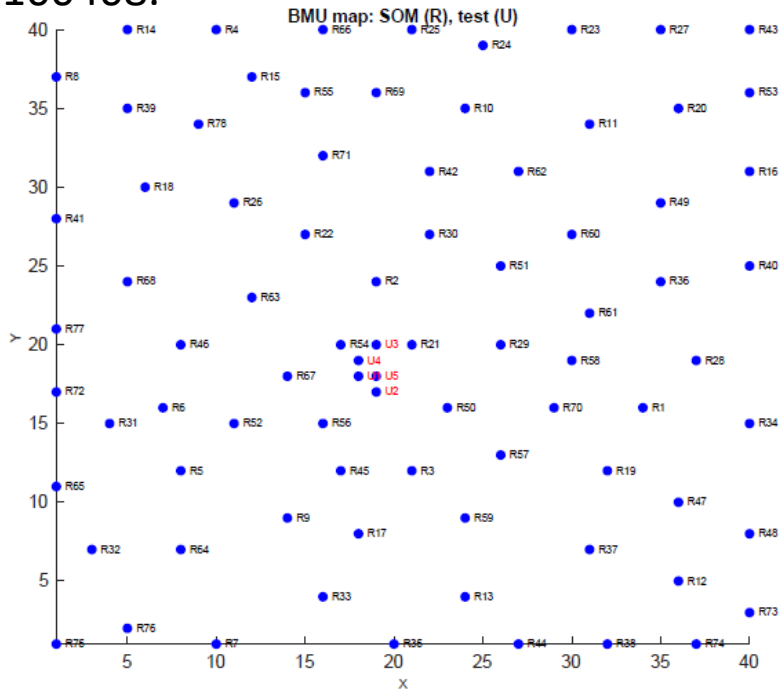
Structure prediction, concentration estimate or random coil removal then regeneration

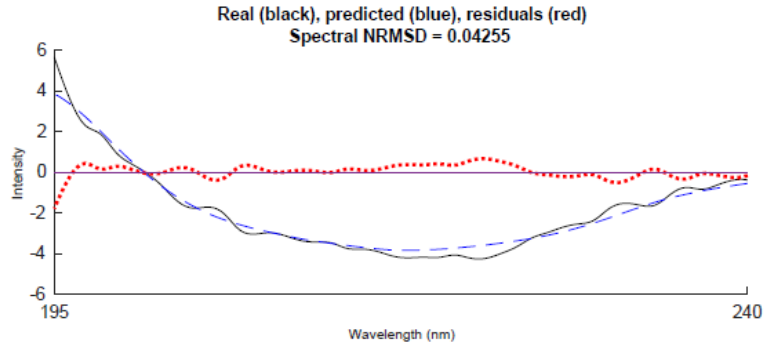
Uperin3 peptide: Martin, L. L. et al. *Chempluschem* **2022,** *87* (1), e202100408.



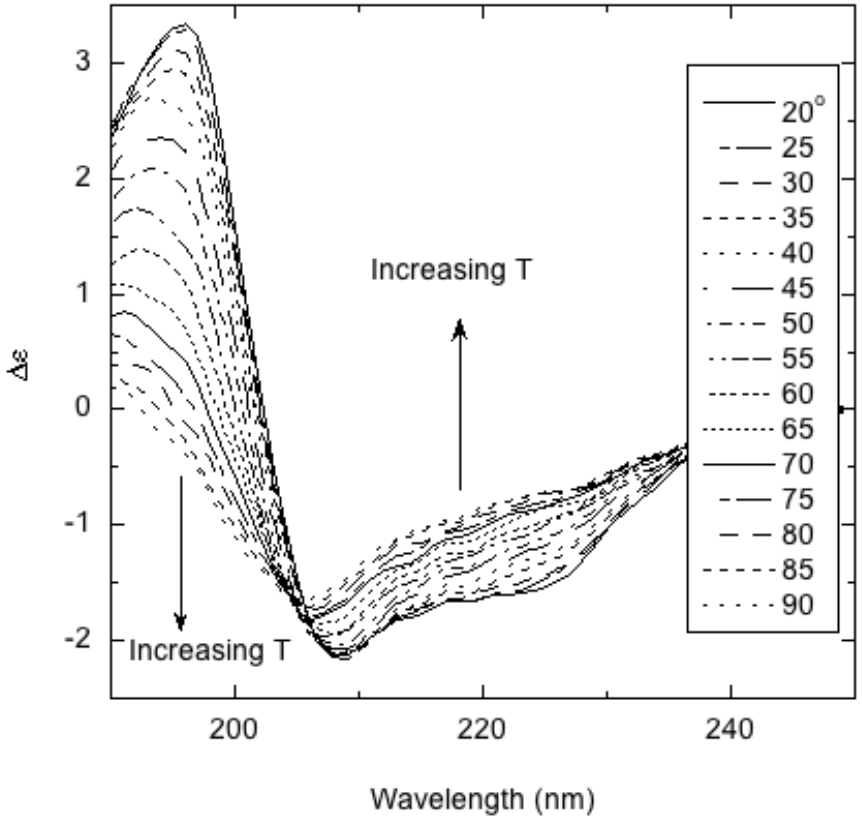
- Okish fit: 7% helix, 22% sheet, 59% RC BUT a β_{II} protein is a BMU, so RC may be underestimated.

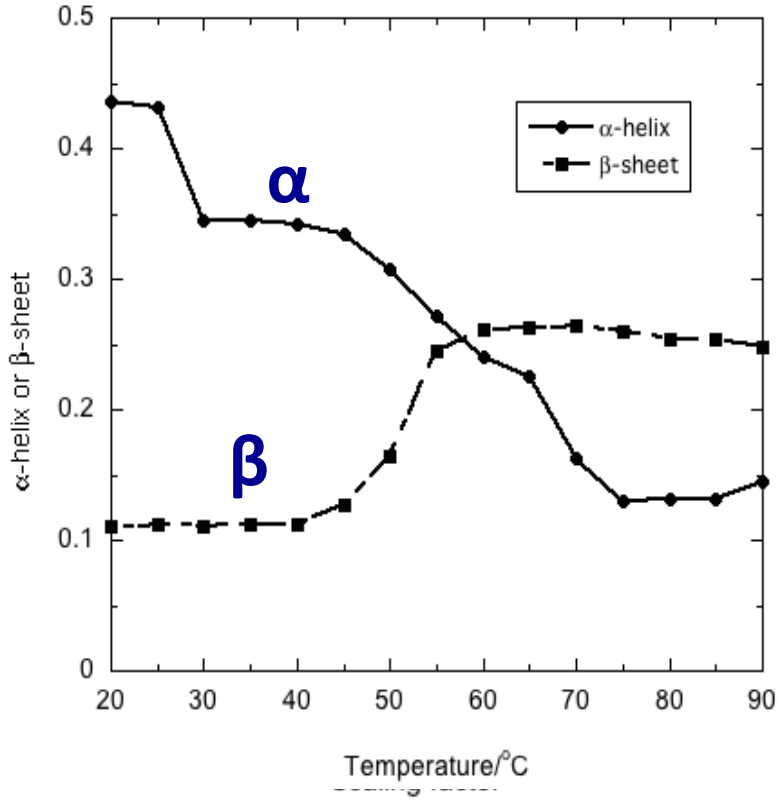
- So we removed increasing amounts of random coil
- Fitted the derandomized spectrum (best is 85% RC)
  - NRMSD only a bit better, but mainly due to noise.
- Added back random coil that was removed:
- 4% helix, 4% sheet, 89% RC





Insulin: unknown concentration no method fits REALLY well



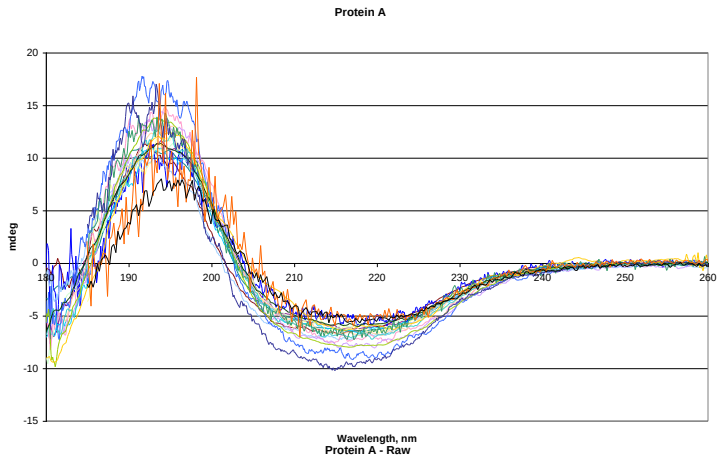


Temperature dependence of CD provides another dimension for stability and batch-to-batch comparison

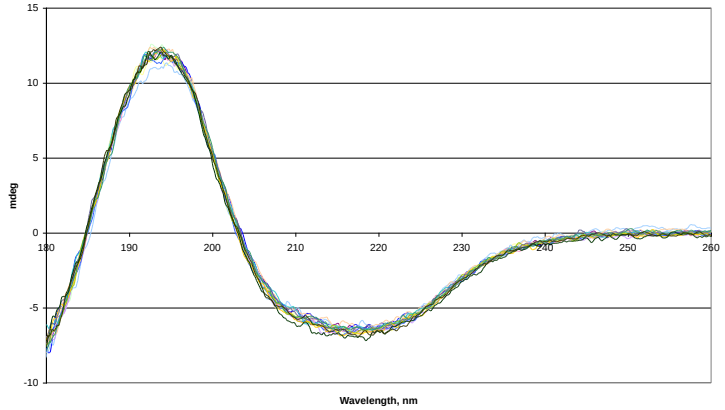
Round robin protein CD results

Protein ‘A’

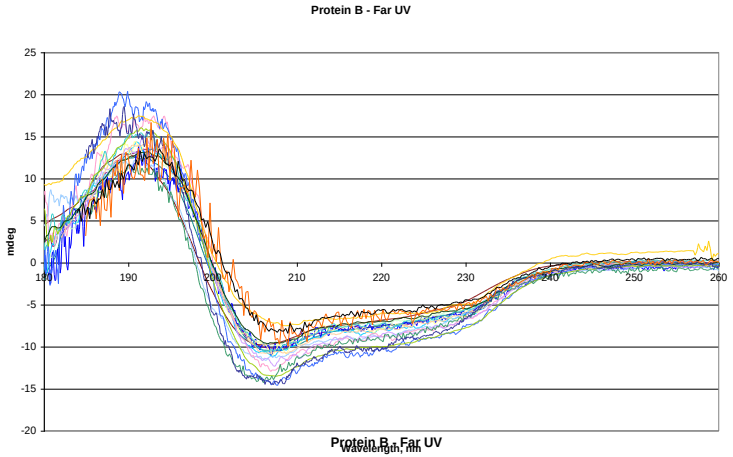
US

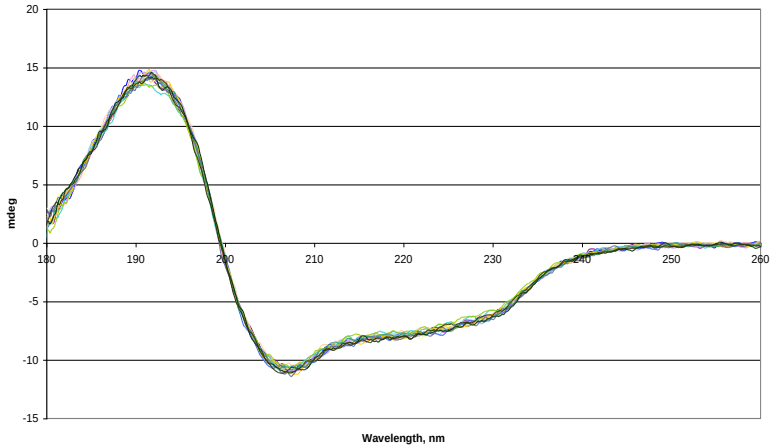


NPL



Protein ‘B’)

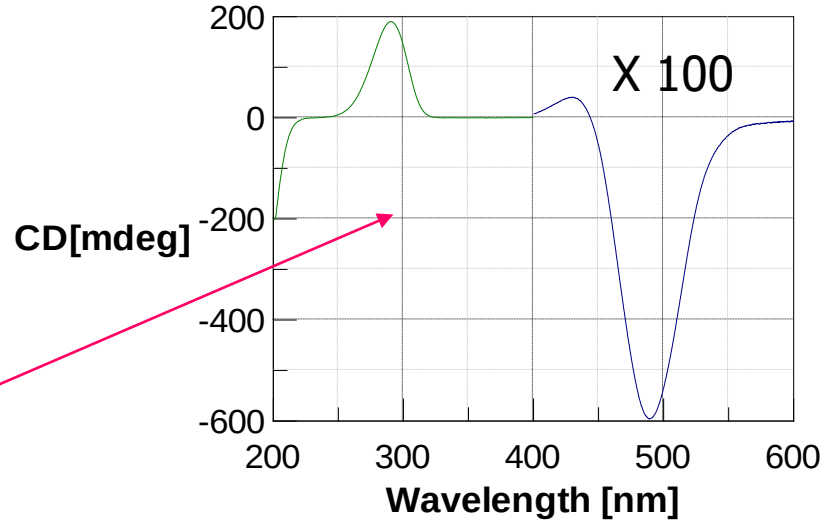




Proteins OK, instruments or operators not.
Perhaps not fair as demountable cells used

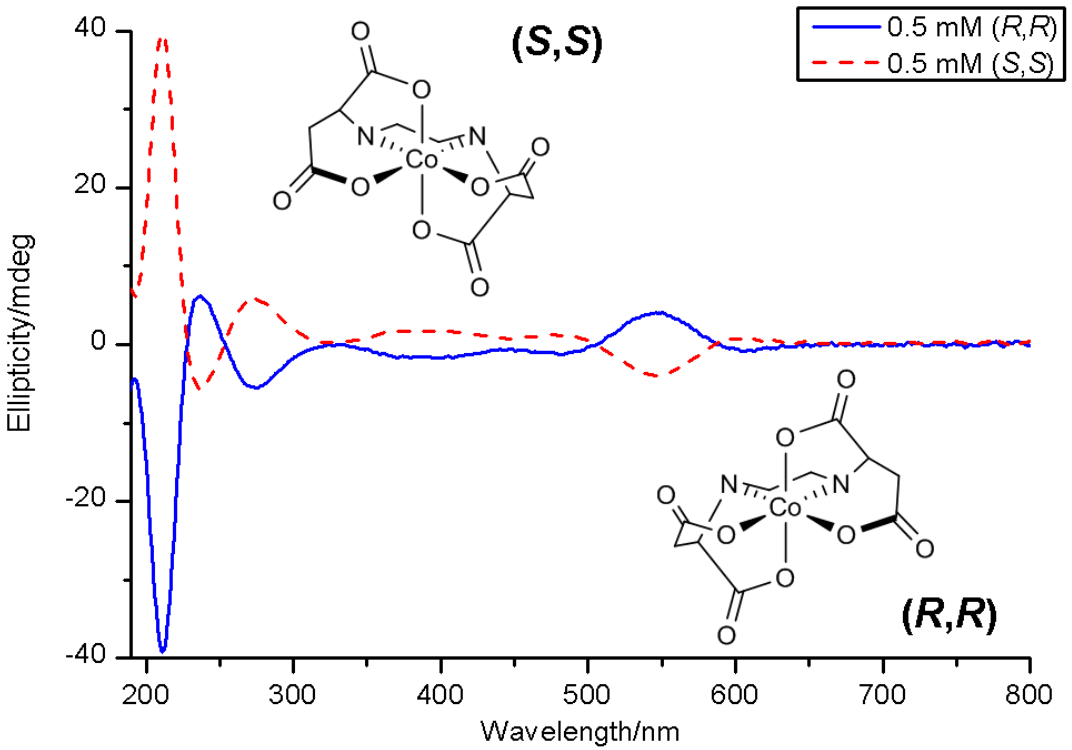
Any fitting is only as good as the data: instrument calibration and path length very important

- CD instruments are usually calibrated using chemical chiral compounds, e.g.:
  - ACS (ammonium-d-10-camphorsulfonic acid) – **single point**
  - CSA (camphorsulfonic acid)
  - Pantolactone
  - cobalt (III) tris-ethylenediamine



- There are several limitations with this approach:
  - Chemical (and enantiomeric) purity
  - Availability of reference data & appropriate wavelength features
  - Stability
  - The need to make reference solutions

CD data no use if calibration poor



With Starna, Jasco

Based on 3 or 5 matched sealed cuvettes – so expensive

Enantiomeric purity based on that of starting D-aspartic acid

Protein infra red absorbance

Use cm^{-1} as energy unit
Amide I C=O stretch: solution ~ 1600–1700 cm^{-1}
Amide II N-H bend: solution ~ 1550 cm^{-1}
α-helix + unordered: 1650 cm^{-1}
β-sheet:1618, 1632,
1661 cm^{-1}
β-turns: 1660–1679 cm^{-1}
non H-bonded C=O:
1700 cm^{-1}

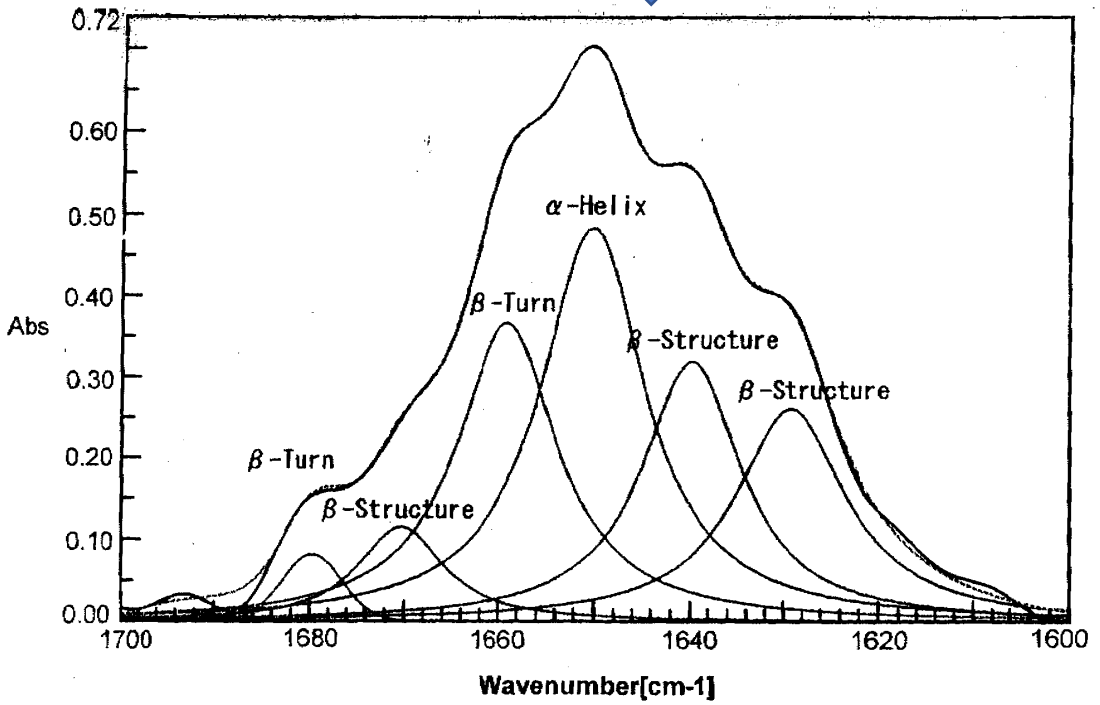
Water absorbance

CaF_2 cells
BSA 20 mg/mL
0.1 mm pathlength
D_2O

Originally used to use D_2O
But not for biopharmaceuticals

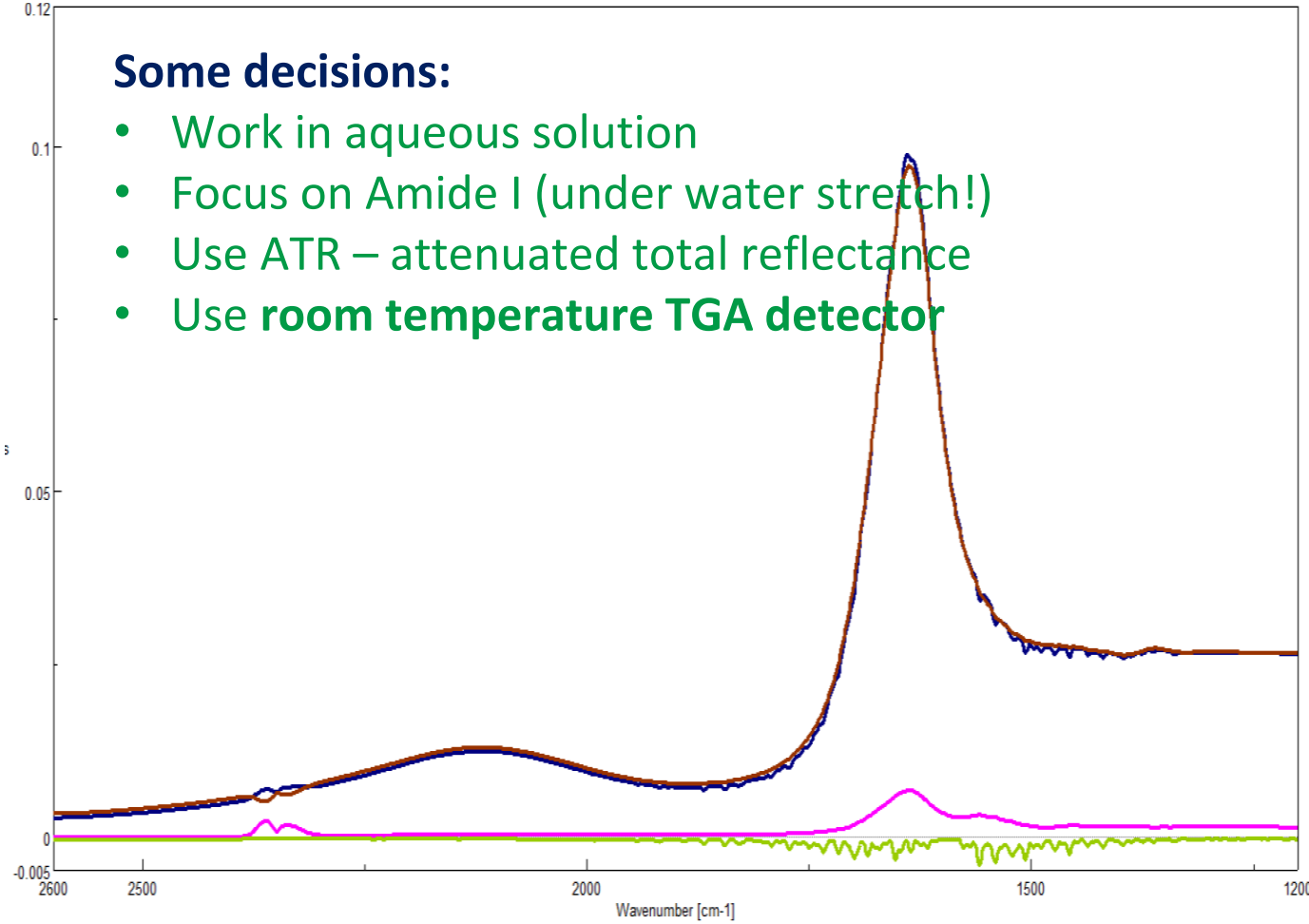
Protein IR spectroscopy

Water absorbance
1643 cm^{-1}



Some decisions:

- Work in aqueous solution
- Focus on Amide I (under water stretch!)
- Use ATR – attenuated total reflectance
- Use **room temperature TGA detector**



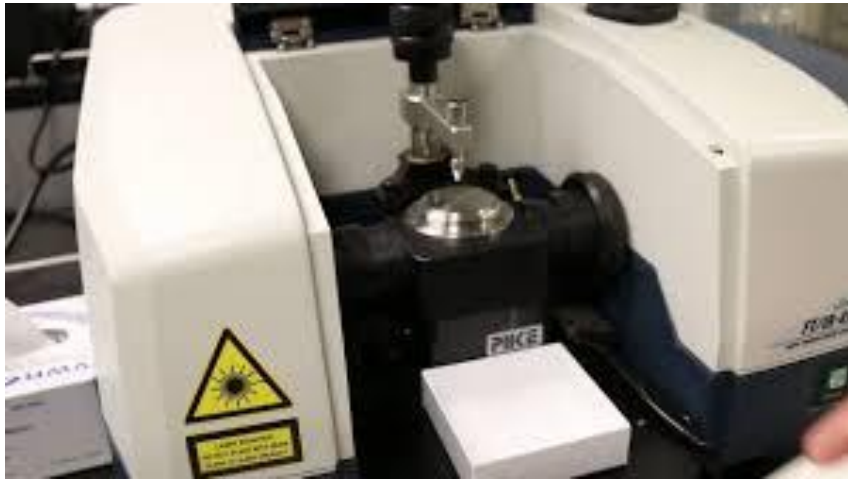
Quantitative IR spectroscopy

$$A = \epsilon C \ell$$

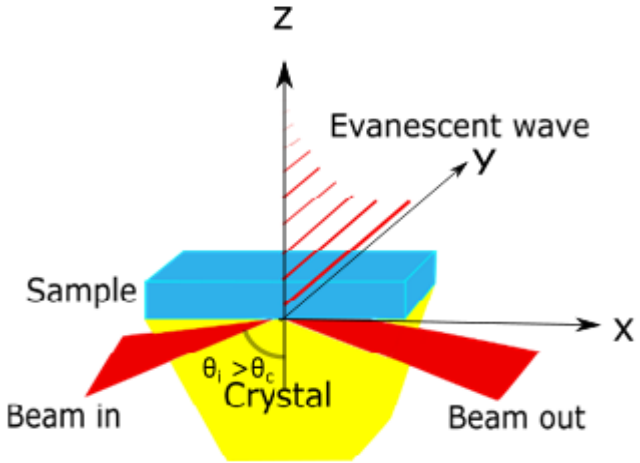
Challenges for transmission IR:

- What is pathlength? Typically 1–10 μm is needed. (Hair is 30–100 μm)
- How to present sample:
  - NaCl windows – water??
  - KBr pellets – grind analyte & KBr up and squash
  - Assemble CaF_2 for biomolecules

Attenuated total reflectance is a possible solution: the pathlength is defined by the instrument and the sample – but it varies with wavelength and sample.



Consider water absorbance at 1645 cm^{-1}
12 μm path length
A=εC?=21.7*55*12/10000=1.4

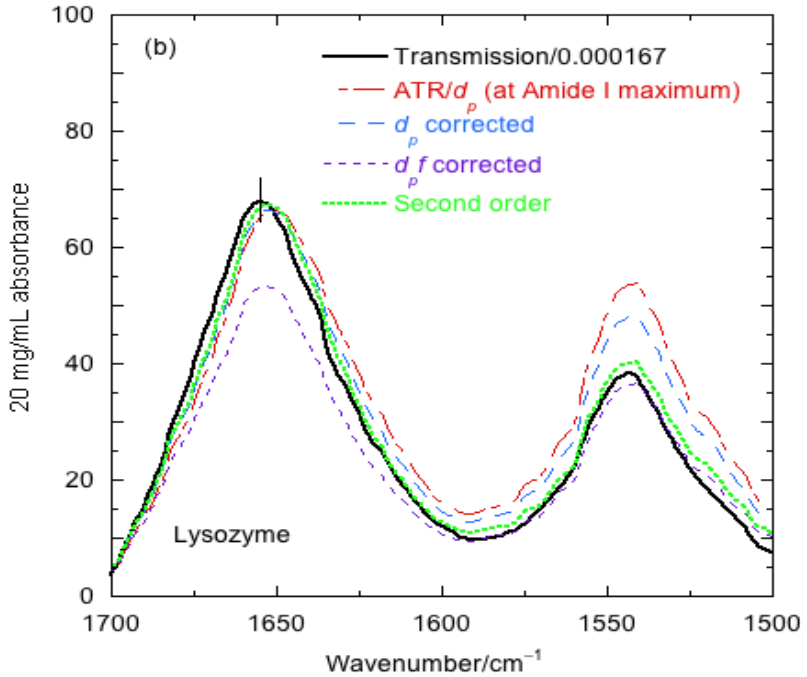


Protein ATR IR: what to do with the data

- Pretend it is transmission (2-3% extra helix error)
- Transform to transmission using

$$A^{ATR}_{protein} = (\varepsilon C)_{protein}\left(a d_p f\right)\left(1 - \left(ln\,10 d_p f\right)(\varepsilon C)_{water}\right)$$

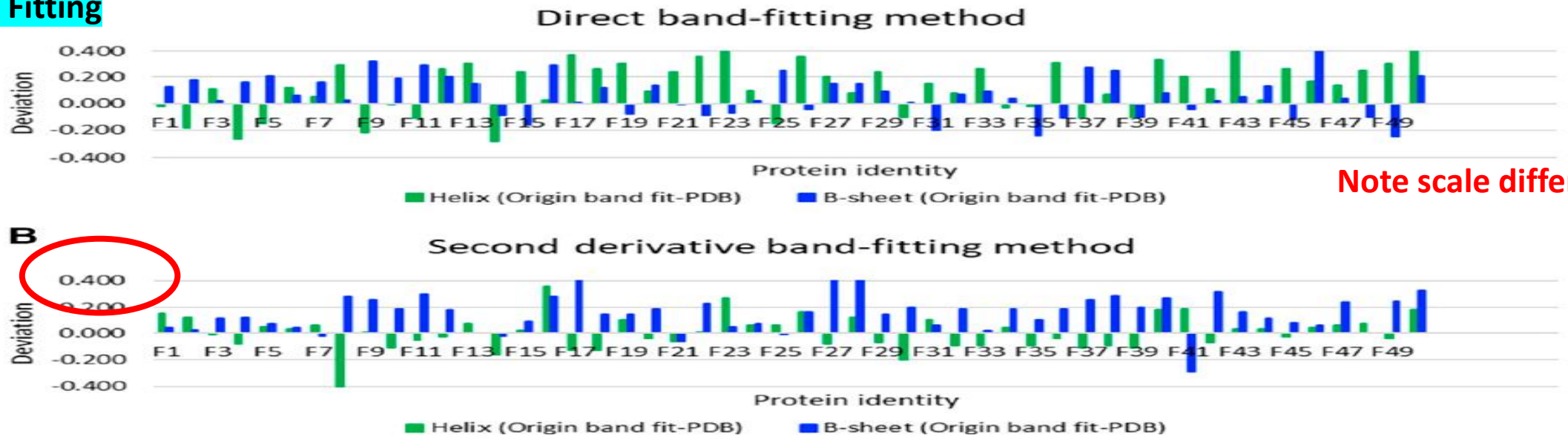
$$A^{Transmission}_{protein} = \frac{A^{ATR}_{protein}\ell}{\left(\left(d_p f\right)\left(1 - \left(ln\,10 d_p f\right)(\varepsilon C)_{water}\right)\right)}$$



- Sort of Beer Lambert with extra factor for penetr sample and decay of intensity with absorbance and distance (f)
- Then fit – usually with data normalised to 1, though with ATR we could use intensity accurately and get more information out
- Band fit or SOMSpec

QRB Discovery **2020,** *1*, e8
Frontiers in Chemistry **2022,** *9*

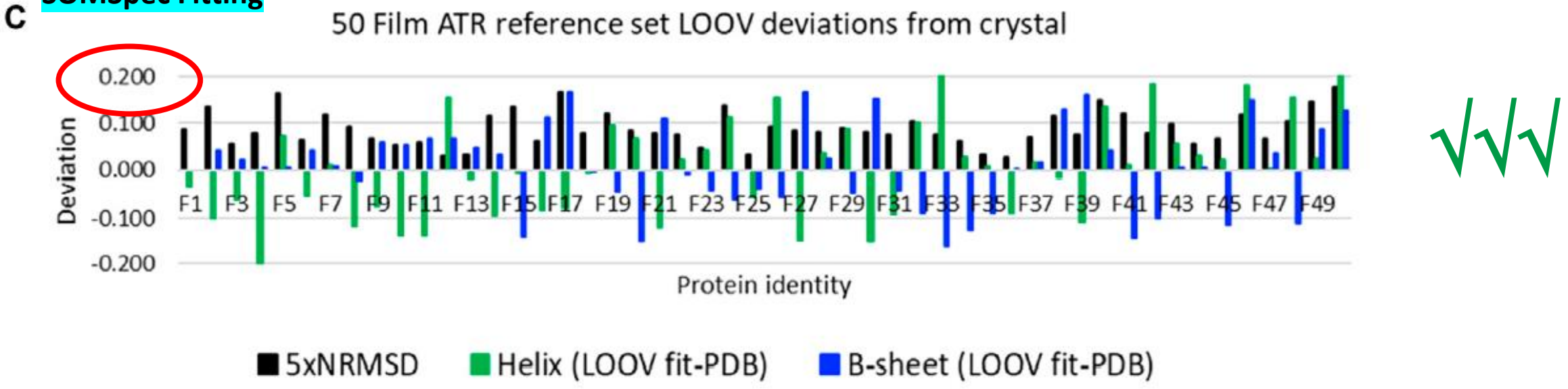
Band Fitting



Note scale difference

4 | Deviations of secondary structure prediction from PDB structures for helix and β-sheet for the Amide I band of the 50-protein film reference s
d in order of decreasing helix content from left to right for **(A)** direct Gaussian band-fitting and **(B)** the second derivative fitting approach reported in re
al., 2015). See Supplementary Material for protein identities. Deviations of the Other category deviations are minus the sum of helix and β-sheet dev

SOMSpec Fitting

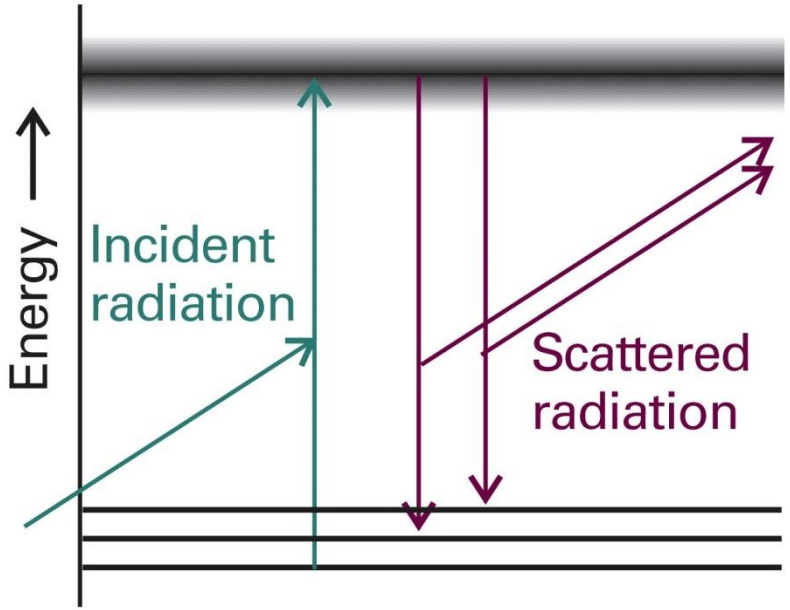


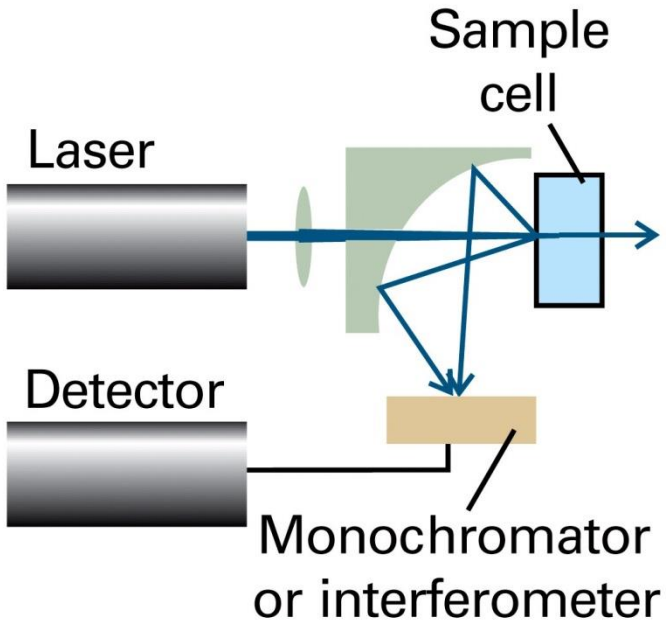
√√√

Raman Spectroscopy

- Alternative technique for monitoring vibrational properties of molecules.
- Measure frequencies present in the radiation **scattered** by molecules.

Formally: Excitation of molecule to wide range of states



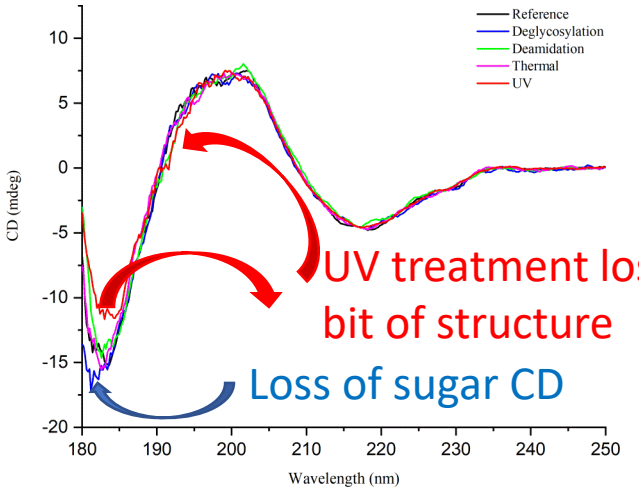


Stokes radiation – photons emerge with lower energy due to scattering.

anti-Stokes radiation – photons emerge with higher energy due to scattering.

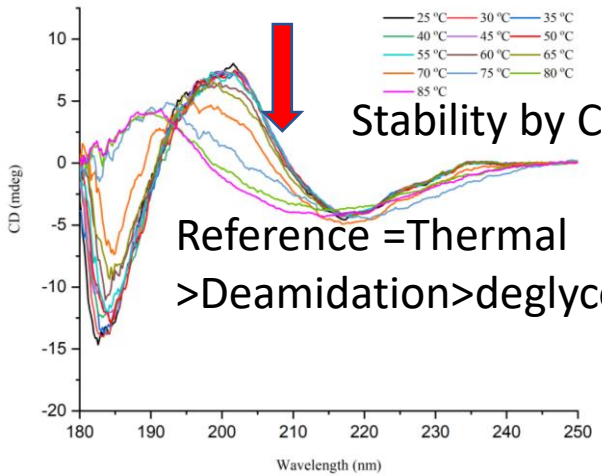
Rayleigh radiation – radiation scattered without change of frequency.

PTMs in a stressed monoclonal antibody



UV treatment loses bit of structure

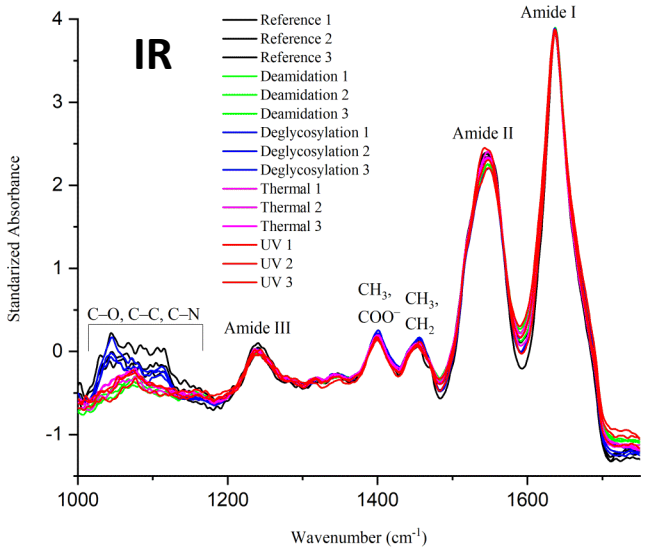
Loss of sugar CD



Stability by CD

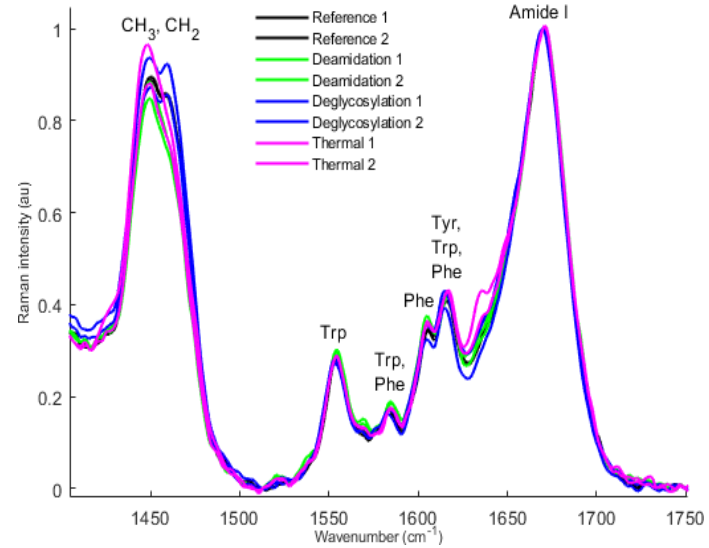
Reference =Thermal >Deamidation>deglycosylation>UV

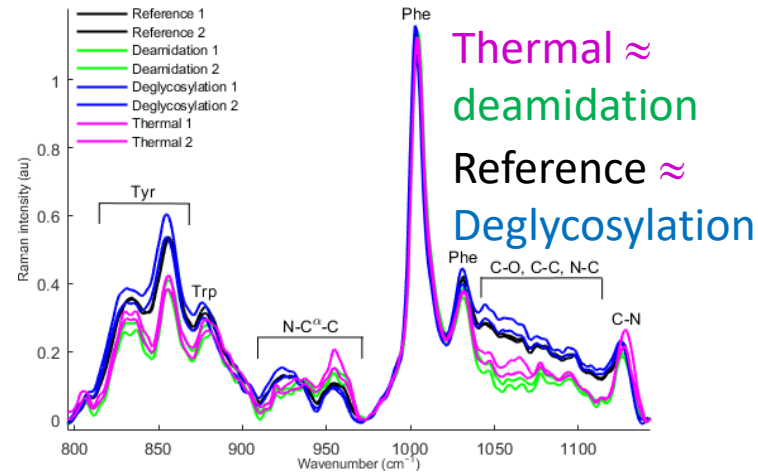
Thermal = 4 weeks 55°C (deamidation+)
UV = 2 weeks 160–450 nm (various)
Deamidation = pH 8 4 weeks 40°
Deglycosylation = PNGase



Potential here but needs more work

Raman

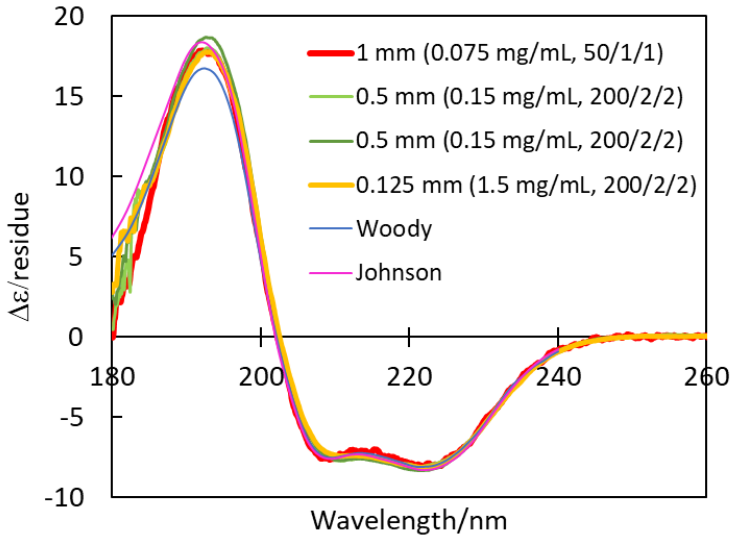


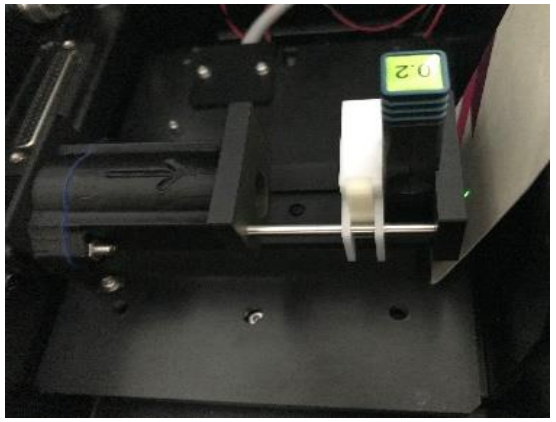


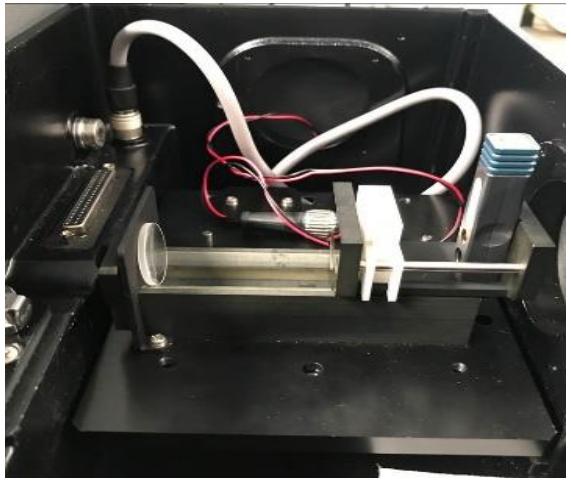
Thermal ≈ deamidation

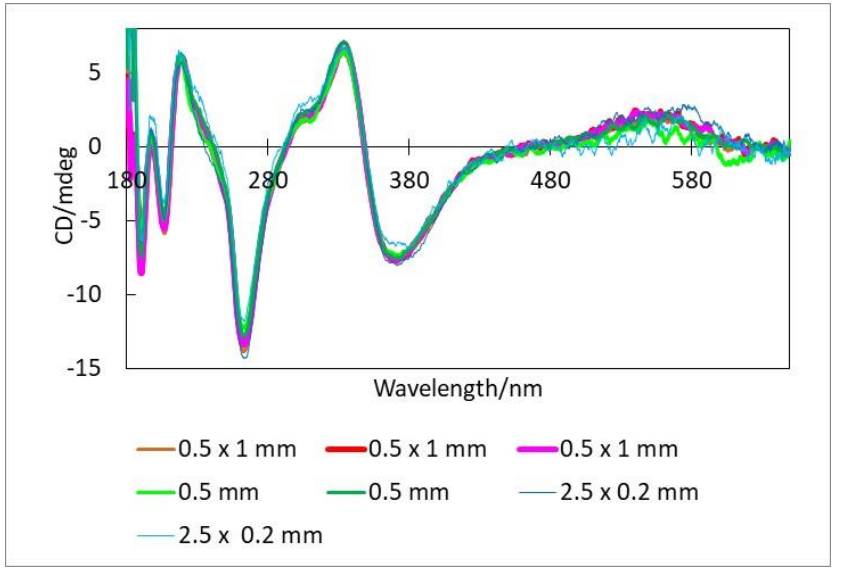
Reference ≈ Deglycosylation

Small volume CD: DMV Bio-cell & Jasco MSD-462









DMV Bio-cell: 500 μm 5 μL
200 μm 2 μL
125 μm 1.3 μL
Very easy to assemble (magnet) but limited path lengths.

Jasco MSD-462 (no spacers): 7 μm 1 μL

Why spectroscopy?

Relative to MS

- Quicker
- Data easier to interpret
- Cheaper
- Gives secondary structures
- Shows structure changes

But

- Not atom specific
- Care must be taken for comparability
- Some buffers stop signal
- Requires ~10 µg protein
- CD (now) requires ~200 ng protein.

What next?

- **Bioactive products – naturally present in food, exert a beneficial or toxic effect?**
- **Their complex matrices are often essential for activity**
- **How do we analyse them???**
- **Industrial Transformation Training Centre to try to answer that question!**

Thank you to

- Andrew Reason (BiopharmaSpec) – who has motivated me for years by refusing to understand instrumentation limits
- Viv Lindo (MedImmune) – who sees the big picture
- Marco Pinto (PhD with MedImmune, now with Agilent)
- Mike Steel (Macquarie) – understands Maxwell's equations properly
- Jason Peterson (BiopharmaSpec) – understands data quality
- Dale Ang and Vince Hall (PhD) – for coding versions of SOMSpec
- Nikola Chmel (Warwick) – excels at pulling things together, most lately SOMSpec
- Angela Martino, Lisa Martin, Meropi Sklepari Adewale Olamoyesan – for data collection and analysis

Does interpretation matter???
Or is pattern matching good enough??

It depends…

1. In R&D it is helpful to be able to say whether the protein in its formulation vehicle is the same as studied by MS/χ/NMR
2. People (including regulators) like having structure summarised in a simple number (such as %α-helix, %β-sheet, %other)

But for Batch-to-batch comparisons assuming we have the original data and

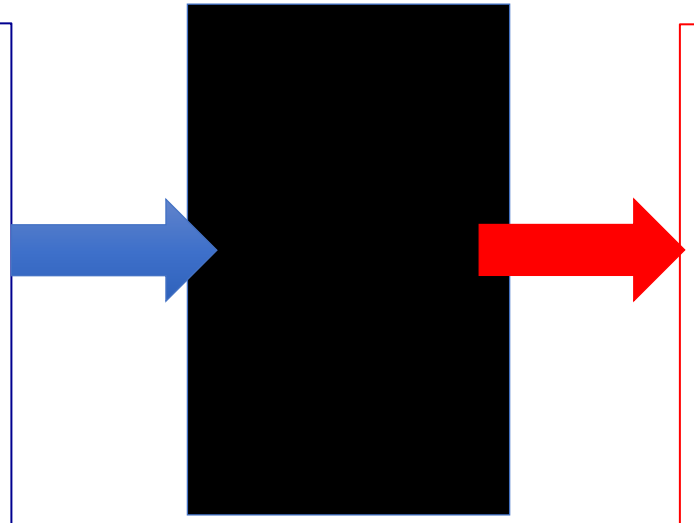
- We ensure data comparability between users? (Calibration and Traceability)
- We can objectively compare spectra/data for batch to batch differences? (Regulation)

May not need it – but I still like it!

Vision/Dream

0.01 mg/mL – 100 mg/mL

- Data from *any/many* technique
  - Concentration
  - Chirality
  - Buffers
  - Chromophores



- Structure/Activity
  - Handedness
  - 2° structure
  - 3° structure
  - Purity
  - Post-trans mod
  - Your desire!

Neural Networks
Independent component analysis
Clever statistics
?????