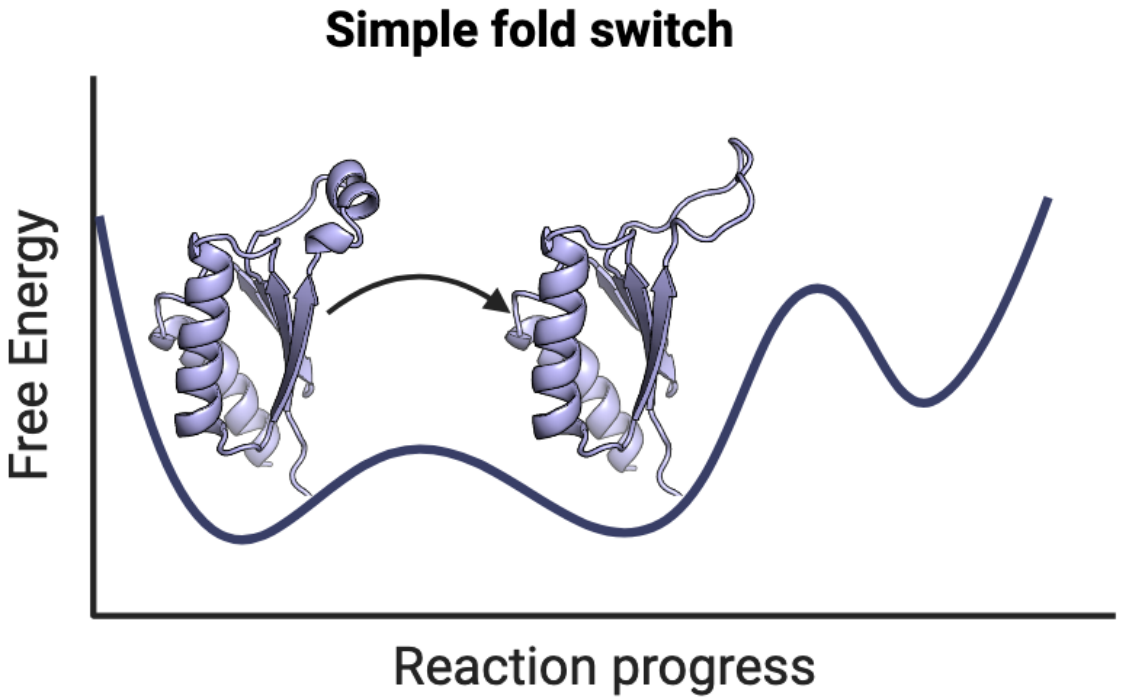
Rotation Group Meeting

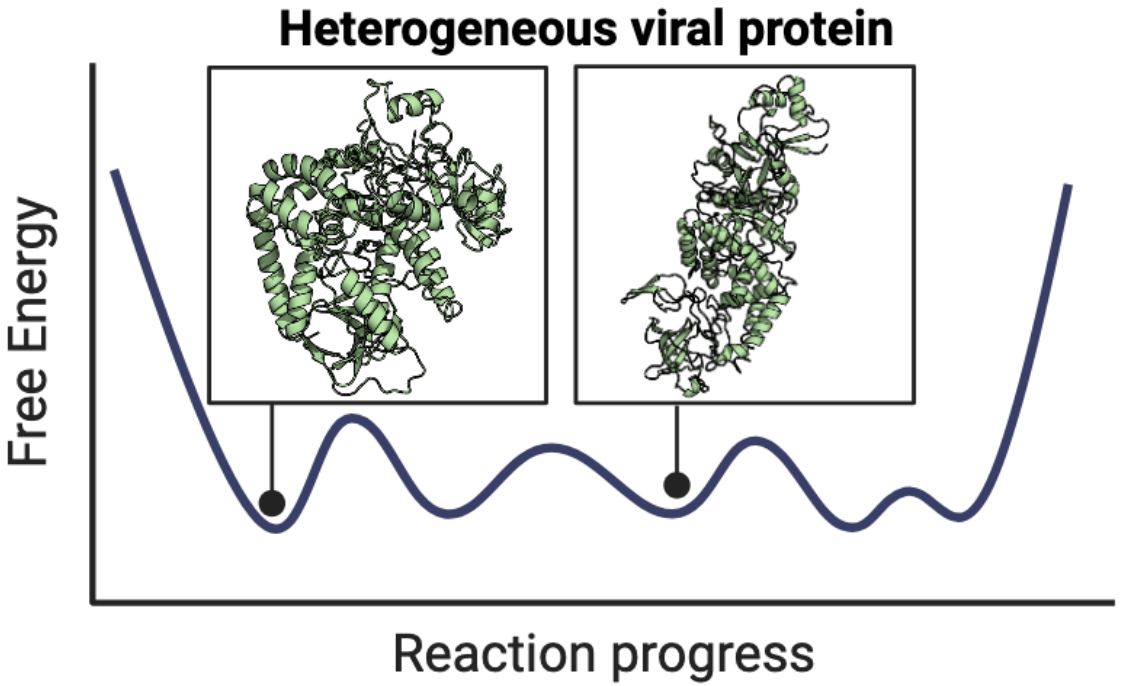
Yulia Gutierrez

Šali & Echeverria Labs

December 10, 2024

Proteins exist in a range of conformations



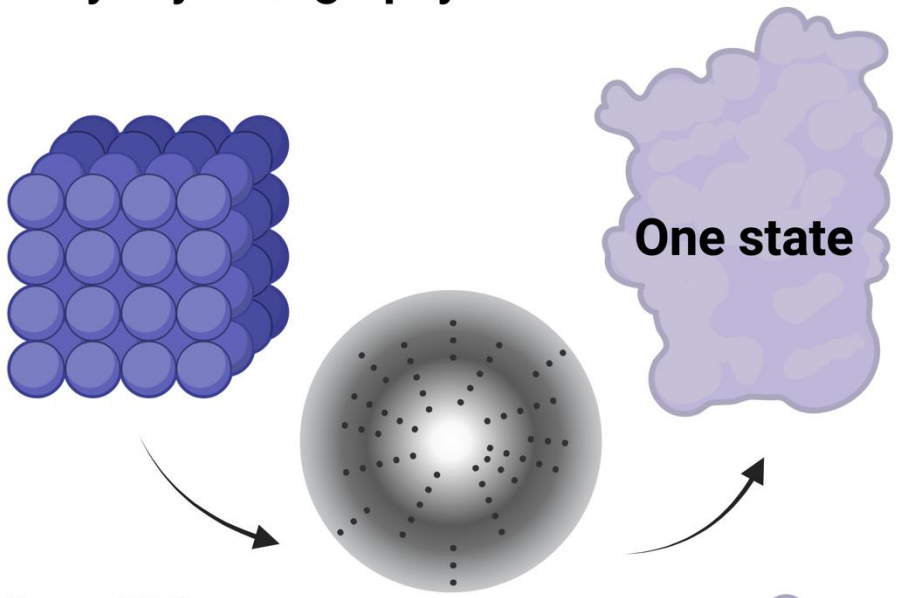


Viral proteins are particularly challenging to model with traditional methods, highlighting the need for ensemble modelling!

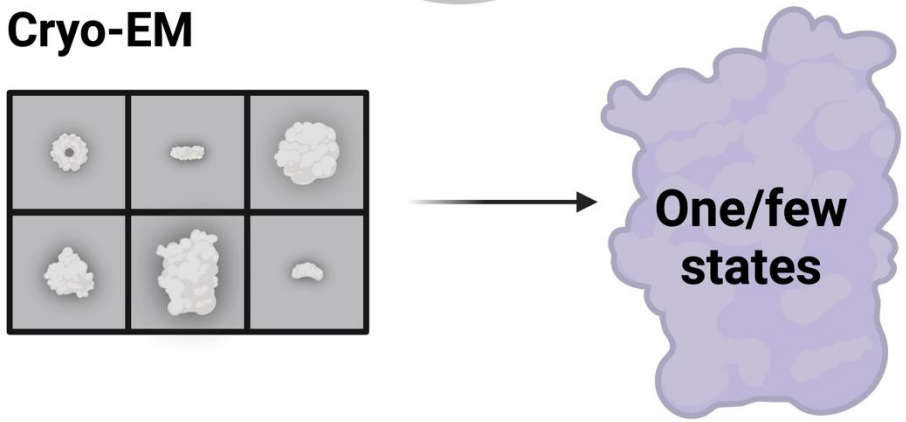
Kenneth Huang
BioRender

Modelling conformational ensembles is challenging with existing methods

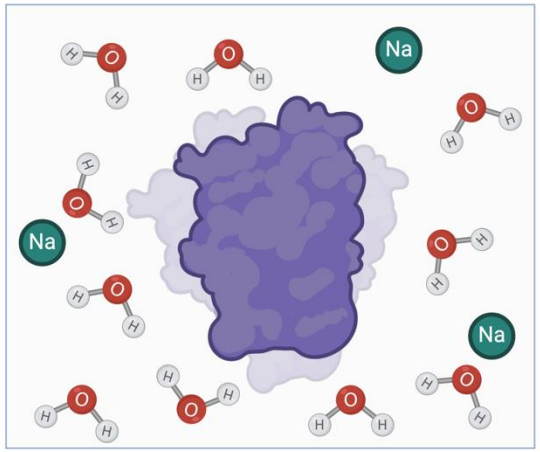
X-ray Crystallography



Cryo-EM



MD Simulations

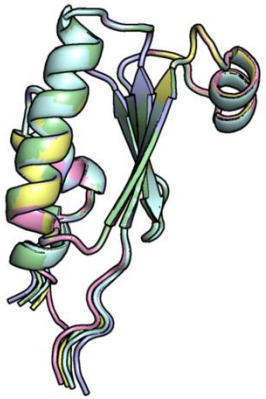


Can be limited by resources

AlphaFold

SEQWENCE

Structural homogeneity

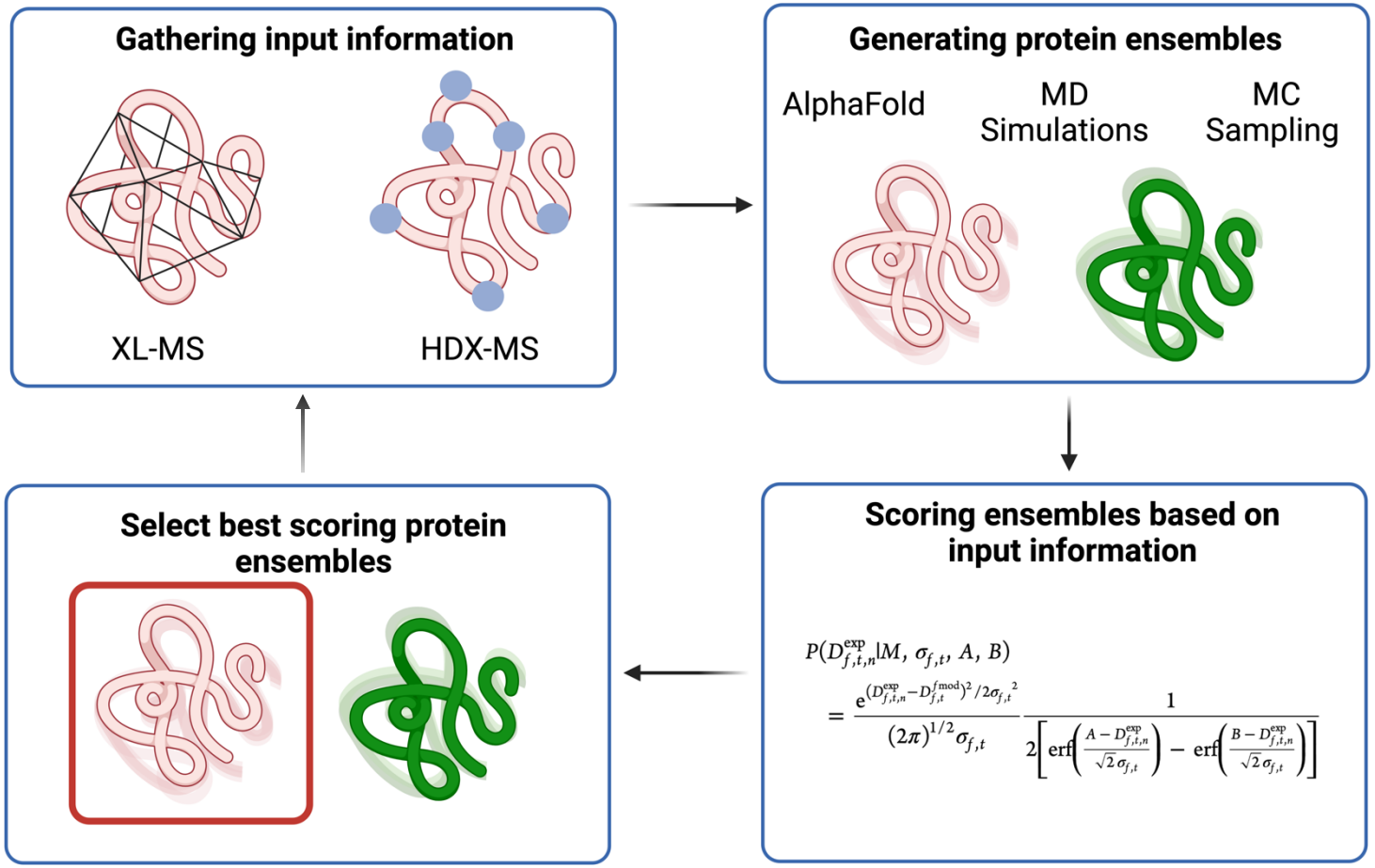


BioRender

Conformational ensemble refinement by integrative modelling of HDX-MS and XL-MS

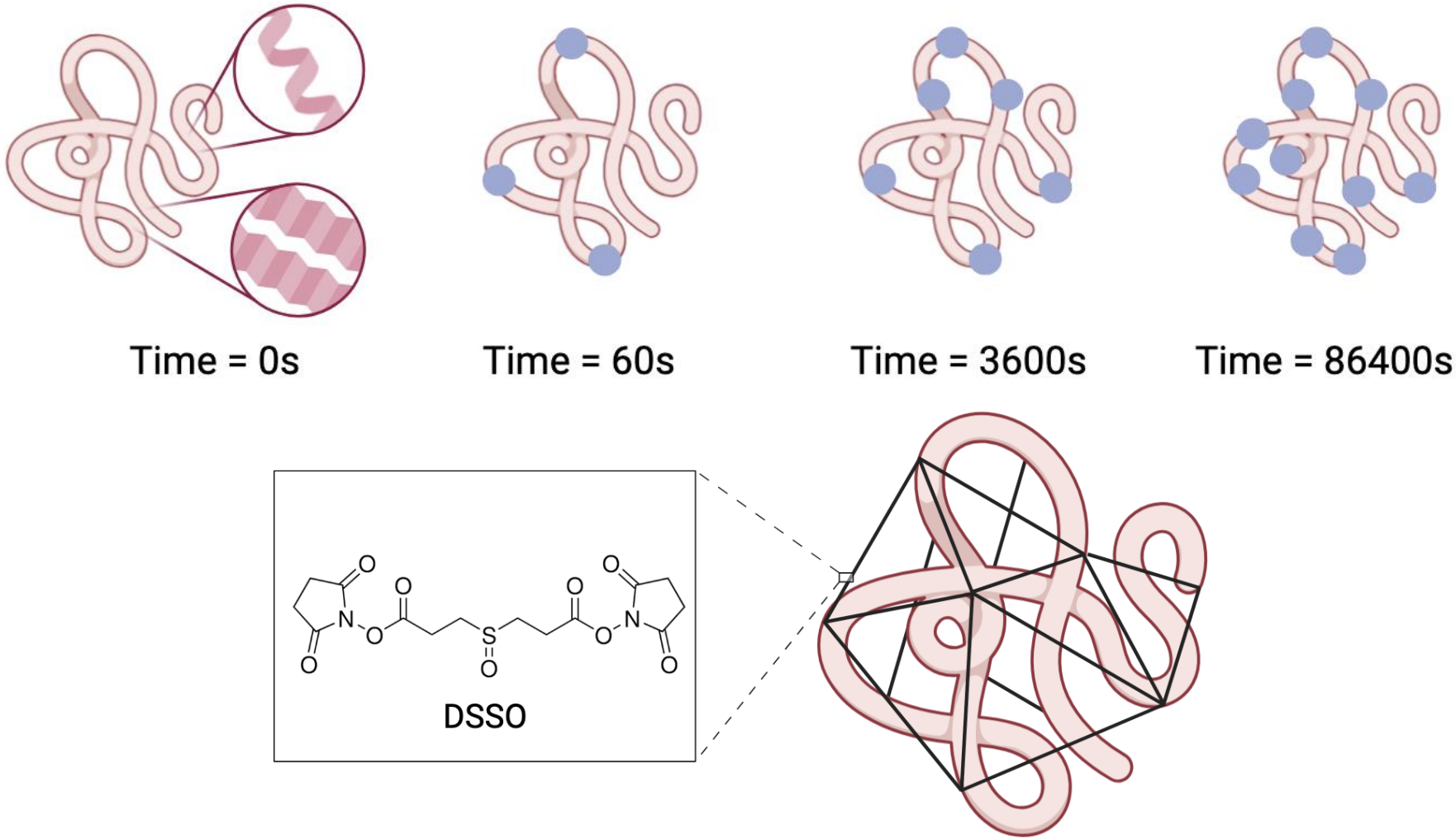
Four stages of modelling:

1. Gathering input information
2. Representation and Scoring
3. Sampling
4. Model Validation



BioRender

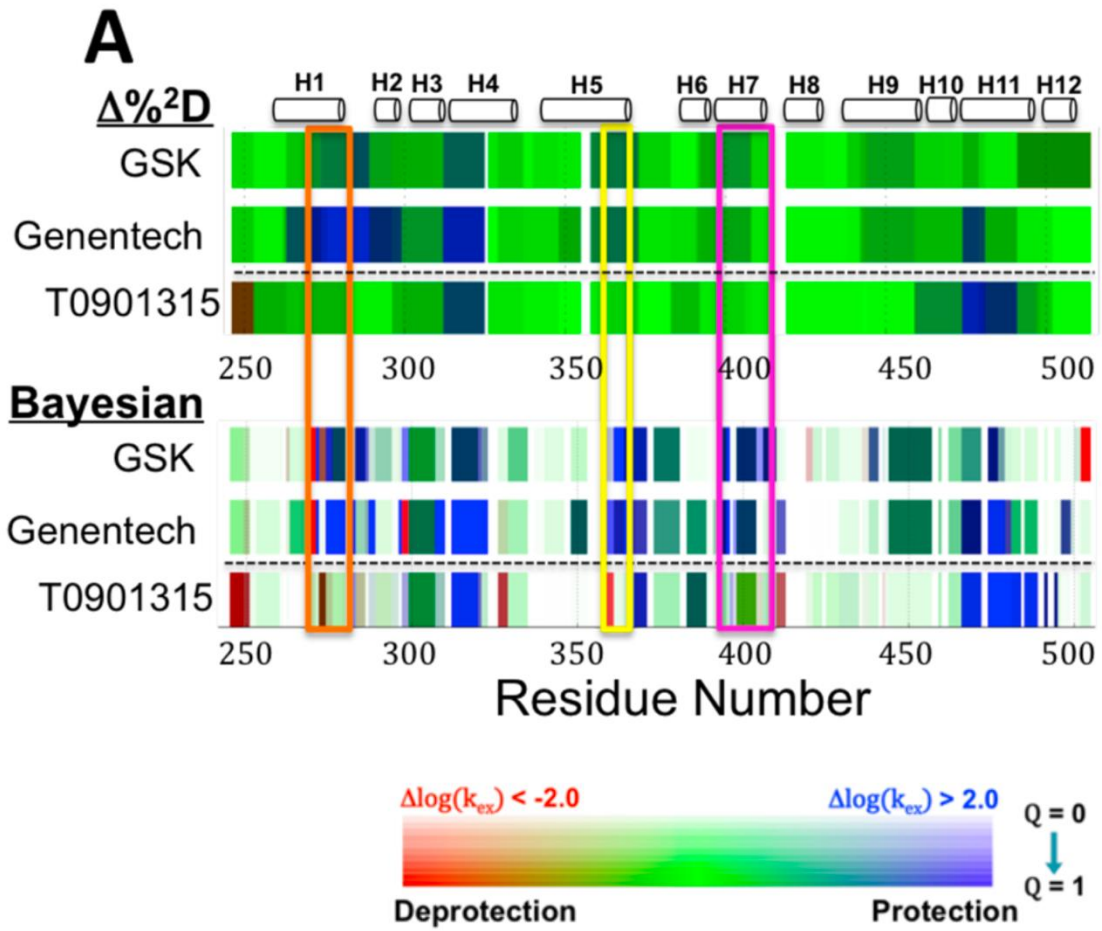
HDX-MS and XL-MS reveal structural and dynamic information



BioRender

Previous work by Saltzberg et al. 2016 introduces Bayesian method for HDX-MS data

- HDX-MS experiments record D incorporation over time on a peptide-level
- Understanding residue-level dynamics improves application of HDX-data for ligand binding and drug discovery



Saltzberg et al. 2016, Figure 4

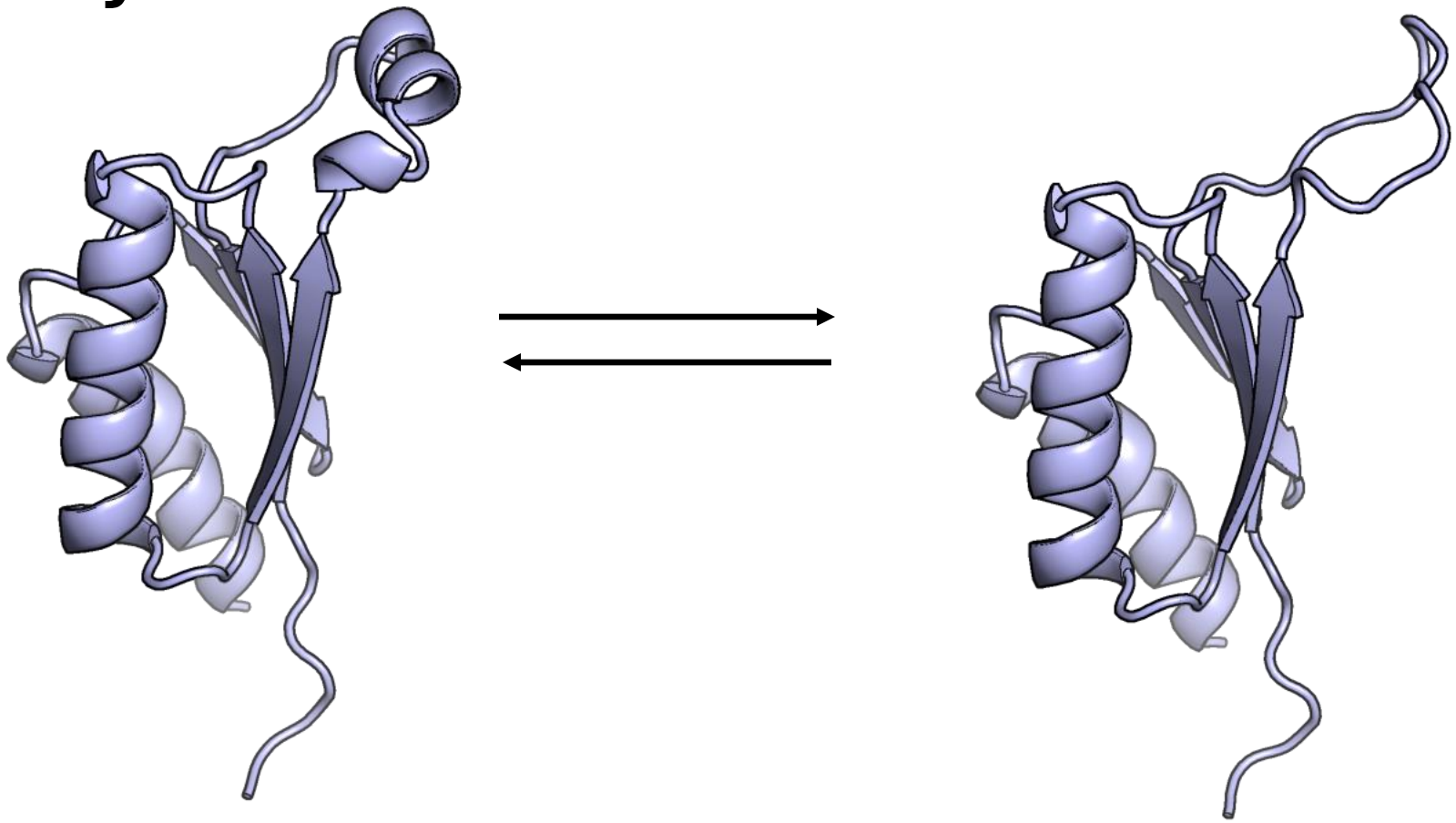
Scoring an ensemble of structures for its likelihood of representing HDX-MS data

- We will use a toy system to create a synthetic dataset to evaluate our scoring function

Overview:

1) Estimate protection factors of each residue
2) Calculate deuterium uptake per tryptic peptide over time
3) Create synthetic dataset and score using Daniel's likelihood function
4) Implement model selection to choose 1-state, 2-state, or 3-state model

We chose a 99-residue fold-switch as a model system



Well-studied fold-switch KaiB (1VGL_A)

The forward model depends on inferred chemical parameters

$$D_{f,t}^{f_{mod}} = F_{f,t}(\{k_i\}, \phi) = \phi\left(N_f - \sum_{i=n_{f,beg}}^{n_{f,end}} \partial_i e^{-\frac{k_i}{PF_i}t}\right)$$

ϕ – Deuterium fraction of exchange buffer

N_f – Number of exchangeable amide hydrogens

k_i – Intrinsic rate of deuterium exchange of residue i

PF_i – Protection factor of residue i

Adapted from Saltzberg et al. 2016

We estimate protection factors by the Best-Vendruscolo method

$$(N-H)_{cl} \underset{k_{cl}}{\overset{k_{op}}{\rightleftharpoons}} (N-H)_{op} \overset{k_{int}}{\rightleftharpoons} (N-D)_{op} \rightleftharpoons (N-D)_{cl}$$

$$PF = \frac{k_{cl}}{k_{op}} = \frac{k_{int}}{k_{obs}}$$

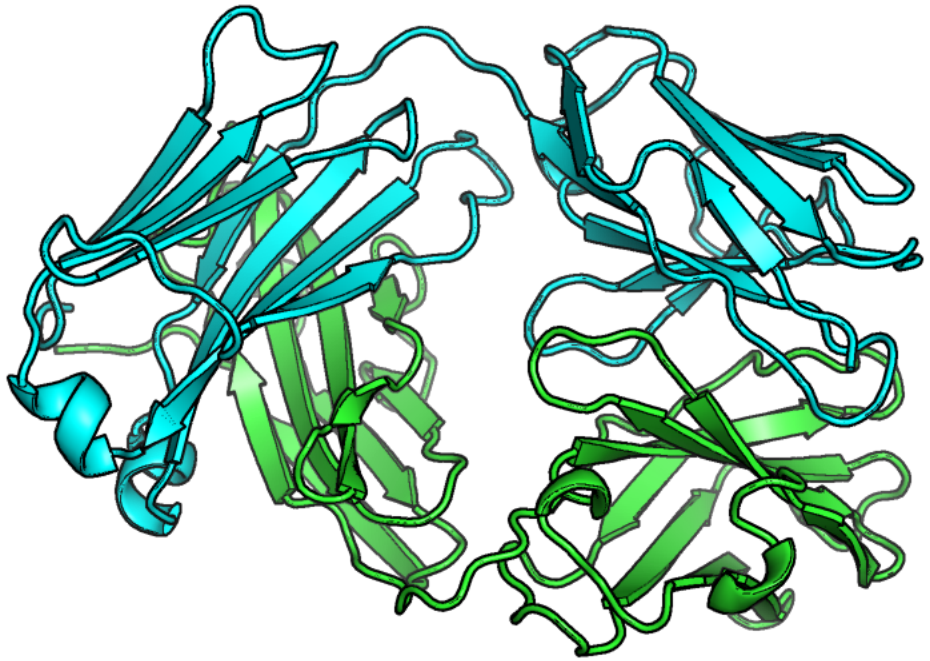
Best-Vendruscolo Estimation: $\ln P_i = \langle \beta_C N_{C,i} + \beta_H N_{H,i} \rangle$

$N_{C,i}$ - Heavy atom contacts

$N_{H,i}$ - Hydrogen bonds formed by backbone amide

Best and Vendruscolo, 2006

We evaluated PFs by comparing predicted deuterium uptake with experimental data

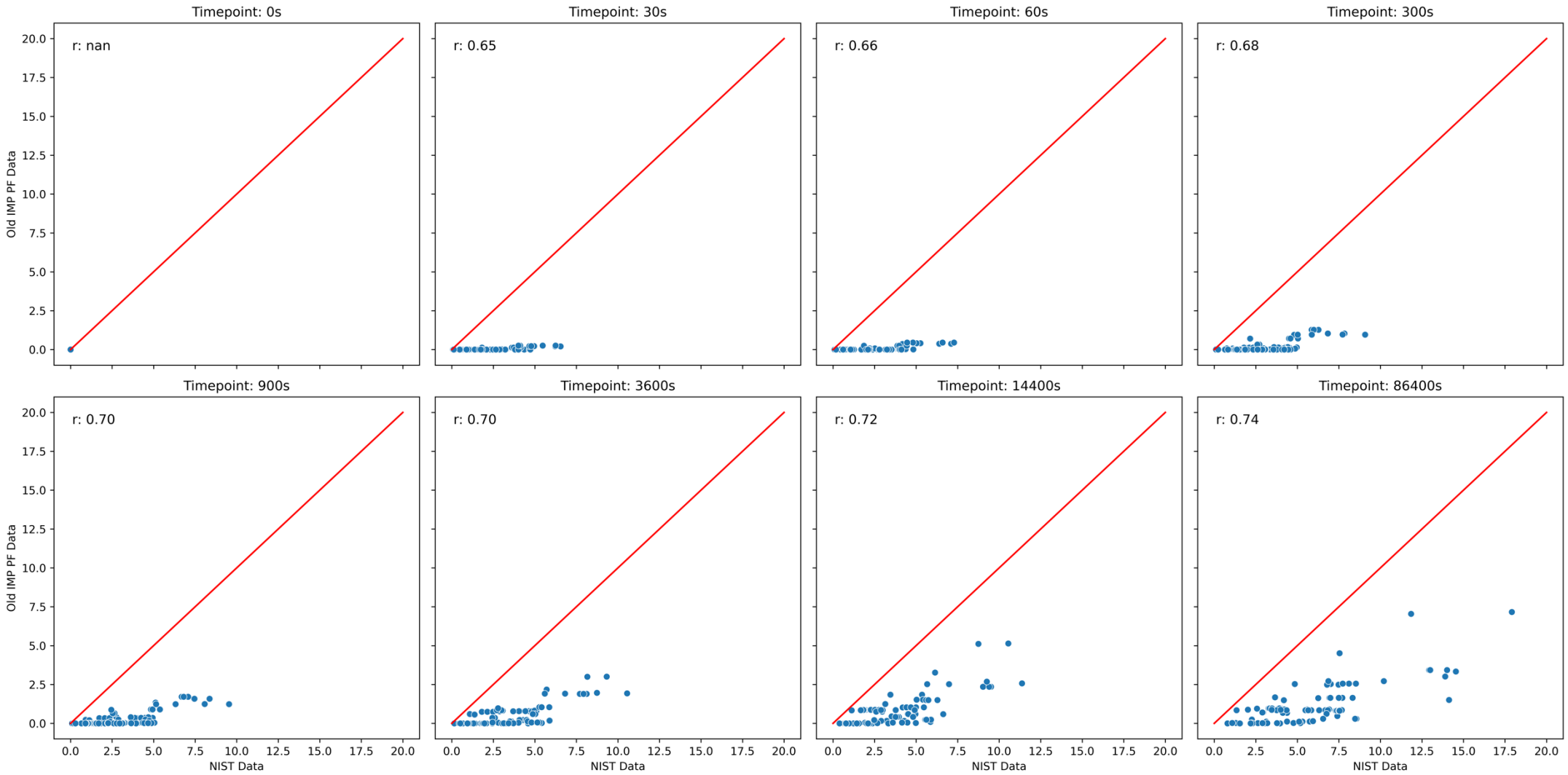


NIST-Fab

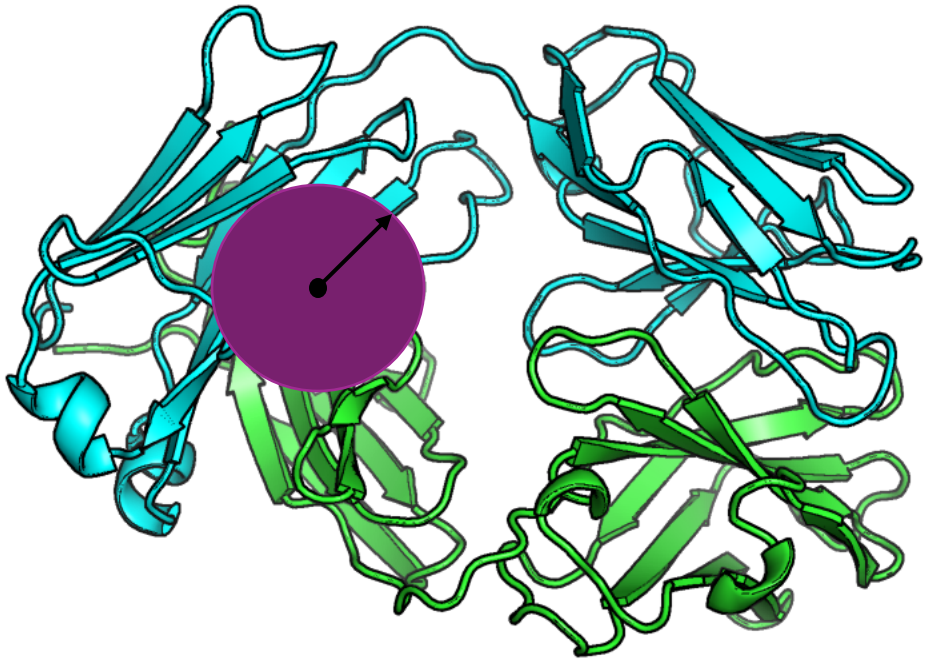
Original criteria in IMP code:

- Baker-Hubbard method for computing H-bonds
  - R(donor-acceptor) < 2.5 A
  - Θ > 120°
- Distance cutoff of 6.5A for computing heavy atom contacts

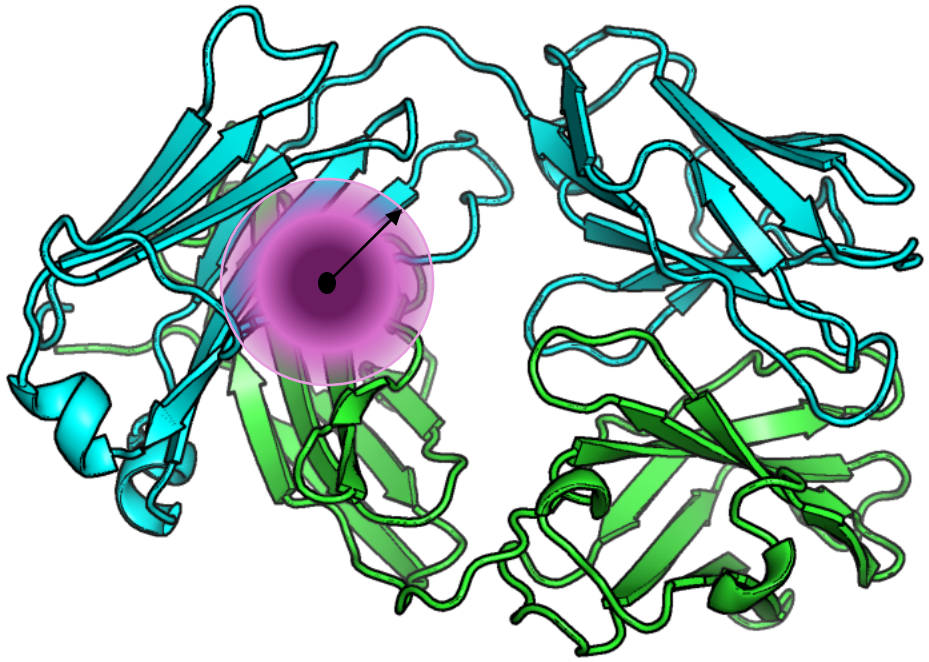
Previous IMP criteria for determining PFs leads to poor prediction of %D for NIST Fab



A sigmoidal decay for heavy atom counts was implemented

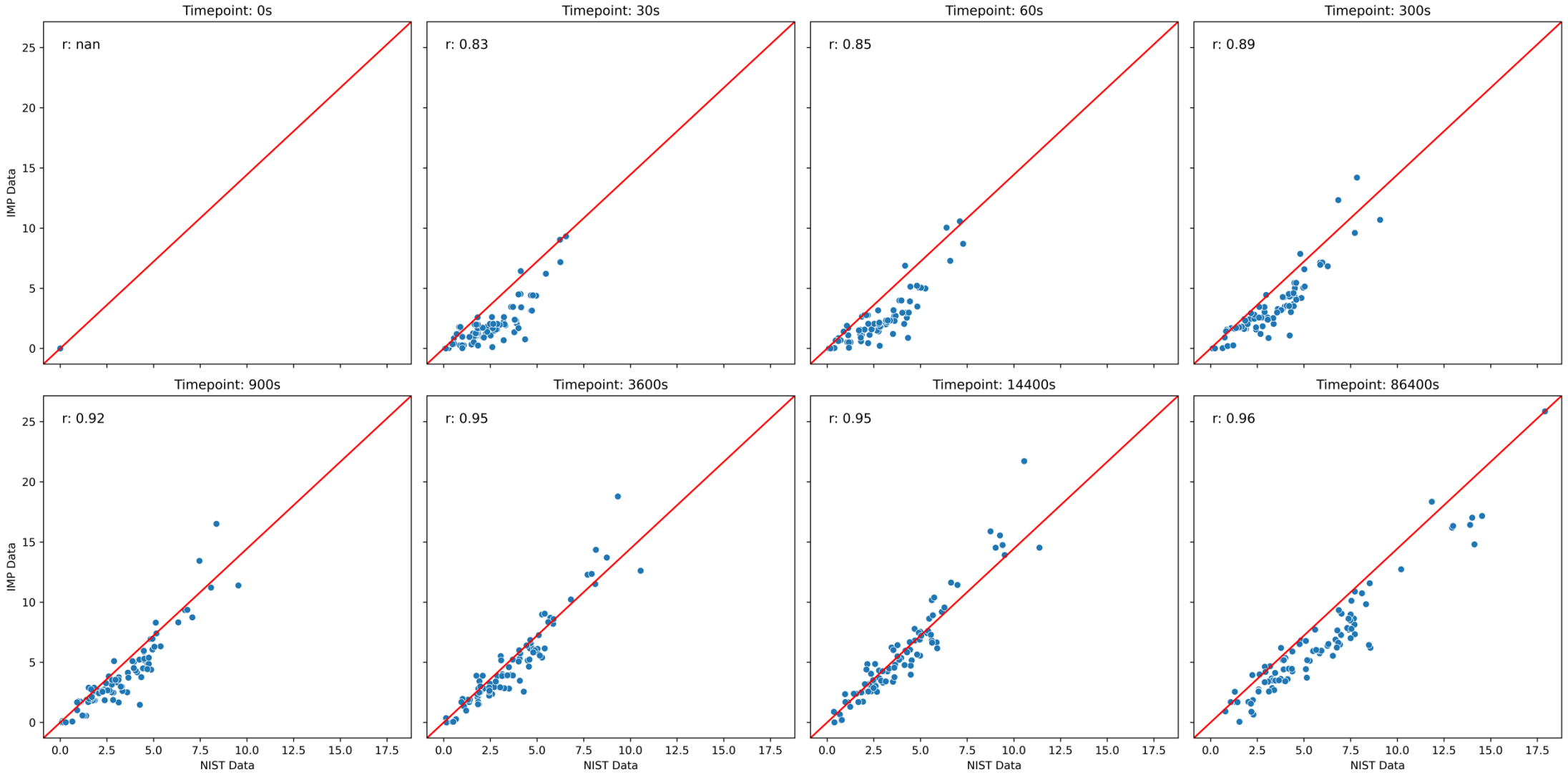


Distance cutoff of 6.5A for computing heavy atom contacts



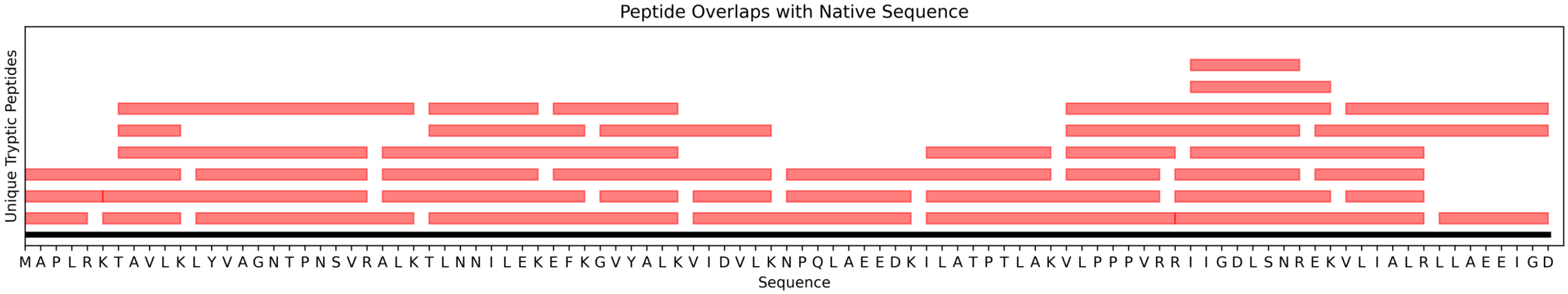
Sigmoidal decay for computing heavy atom contacts

IMP models experimental deuterium uptake fairly well for NIST Fab with sigmoidal decay



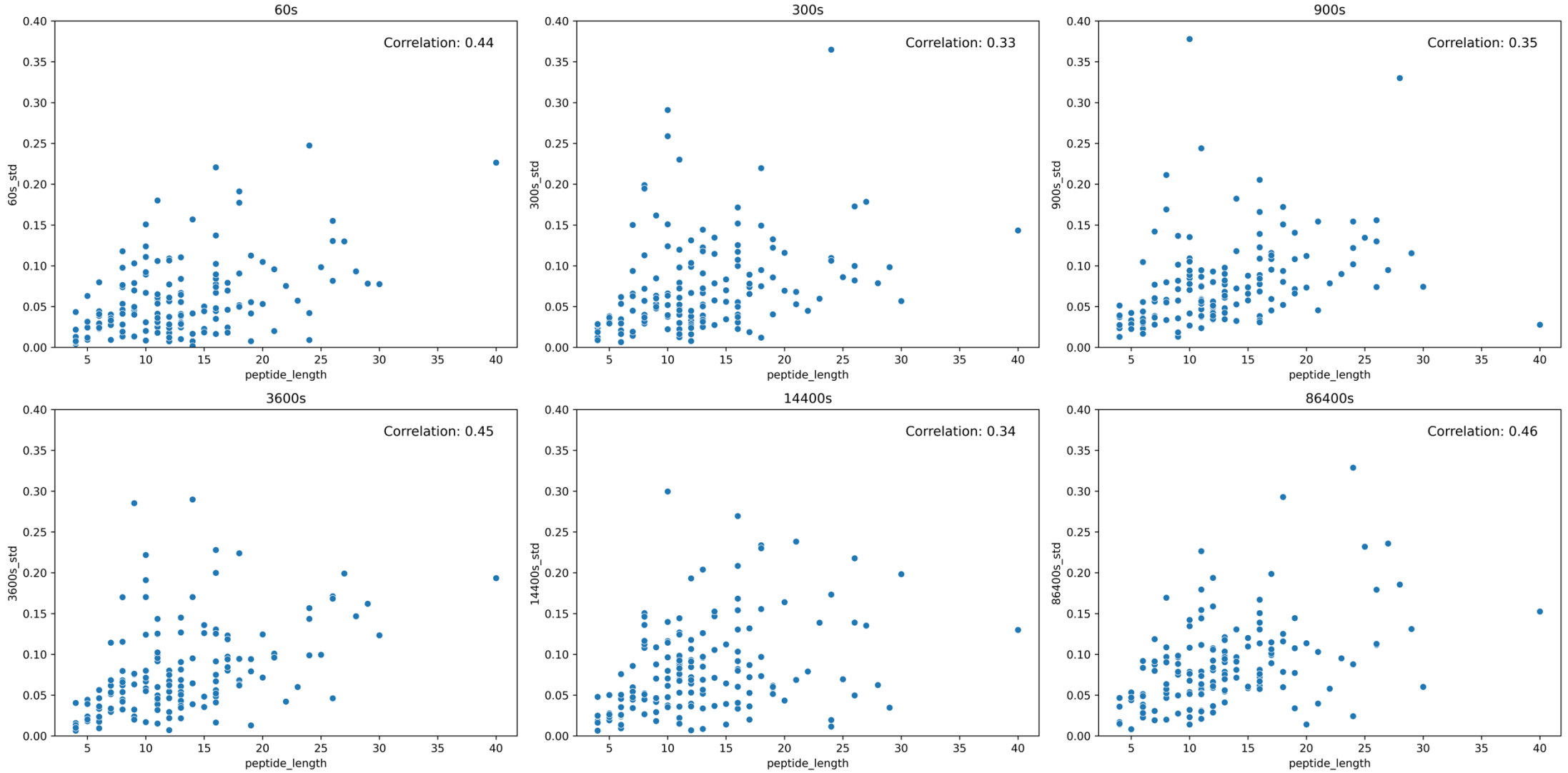
Considering experimental noise and coverage in synthetic data curation

- Noise between replicates
  - Depends on peptide length
  - Modelling as Gaussian noise
- Peptides detectable by MS
  - Between 5-20 residues
  - Varies widely (<20% - >80%)



All unique tryptic peptides of model fold-switch protein

Experimental data has length-dependent noise among replicates



I have created a synthetic dataset using our model fold-switch with random noise

- I am currently working on scoring the synthetic data (representing a mixture of the two states) using Daniel's likelihood function

$$P(D^{\exp}_{f,t,n}|M,\ \sigma_{f,t},\ A,\ B)$$

$$= \frac{e^{(D^{\exp}_{f,t,n} - D^{f\,\mathrm{mod}}_{f,t})^2 / 2{\sigma_{f,t}}^2}}{(2\pi)^{1/2}\sigma_{f,t}} \frac{1}{2\left[\mathrm{erf}\left(\frac{A - D^{\exp}_{f,t,n}}{\sqrt{2}\,\sigma_{f,t}}\right) - \mathrm{erf}\left(\frac{B - D^{\exp}_{f,t,n}}{\sqrt{2}\,\sigma_{f,t}}\right)\right]}$$

- Next: use model selection to choose between 1, 2, or 3 state models, incorporate XL-MS data

Saltzberg et al. 2016

Future directions include model selection and incorporating XL-MS data

- Compute odds ratio for a 1- or 2- state model (expect to select 2-state model for fold-switch) to update priors

$$O_{nn\prime} = \frac{\Pr(n|D,I)}{\Pr(n\prime|D,I)} = \frac{\int \mathrm{d}X\mathrm{d}\alpha \Pr(D|X,\alpha,n,I)\Pr(X,\alpha|n,I)}{\int \mathrm{d}X\mathrm{d}\alpha \Pr(D|X,\alpha,n\prime,I)\Pr(X,\alpha|n\prime,I)}$$

- Explore course-grain representations
- Creating XL-MS forward model and likelihood function will place distance restraints on protein ensemble
- **Conformational ensemble refinement for Sars-Cov2 NSP2 protein**

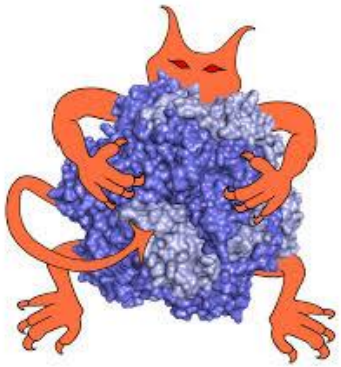
Carsetns et al. 2020

Thank you for a great rotation!

Acknowledgements

- **Andrej Šali**
- **Ignacia Echeverria**
- **Kenneth Huang**
- Arthur Zalevsky
- Neelesh Soni
- Andrew Latham
- Abantika Pal
- Eli Draizen
- Aji Palar
- Matthew Hancock
- Atreya Dey
- Vipul Kumar
- Sree Ganesh Balasubramani
- Surabhi Rathore
- Jared Sagendorf
- Vikas Tiwari
- Ben Webb

- My computer



UCSF