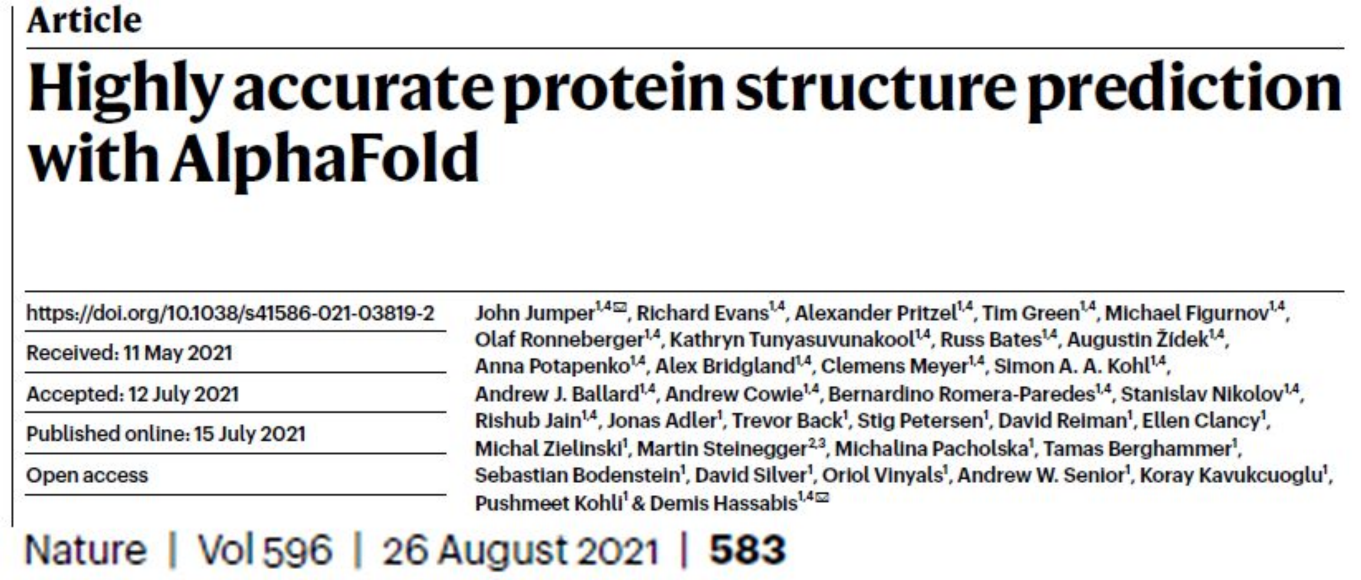
US-SOMO-AF: a database of hydrodynamic, circular dichroism, and SAXS-derived parameters for the AlphaFold-predicted protein structures

Emre Brookes
Department of Chemistry and Biochemistry, The University of Montana, Missoula, MT, USA

Mattia Rocco
Retired, Proteomics & Mass Spectrometry, IRCCS Ospedale Policlinico San Martino, National Institute for Research on Cancer, Genova, Italy

The AlphaFold “revolution”



Article

Highly accurate protein structure prediction with AlphaFold

https://doi.org/10.1038/s41586-021-03819-2

Received: 11 May 2021

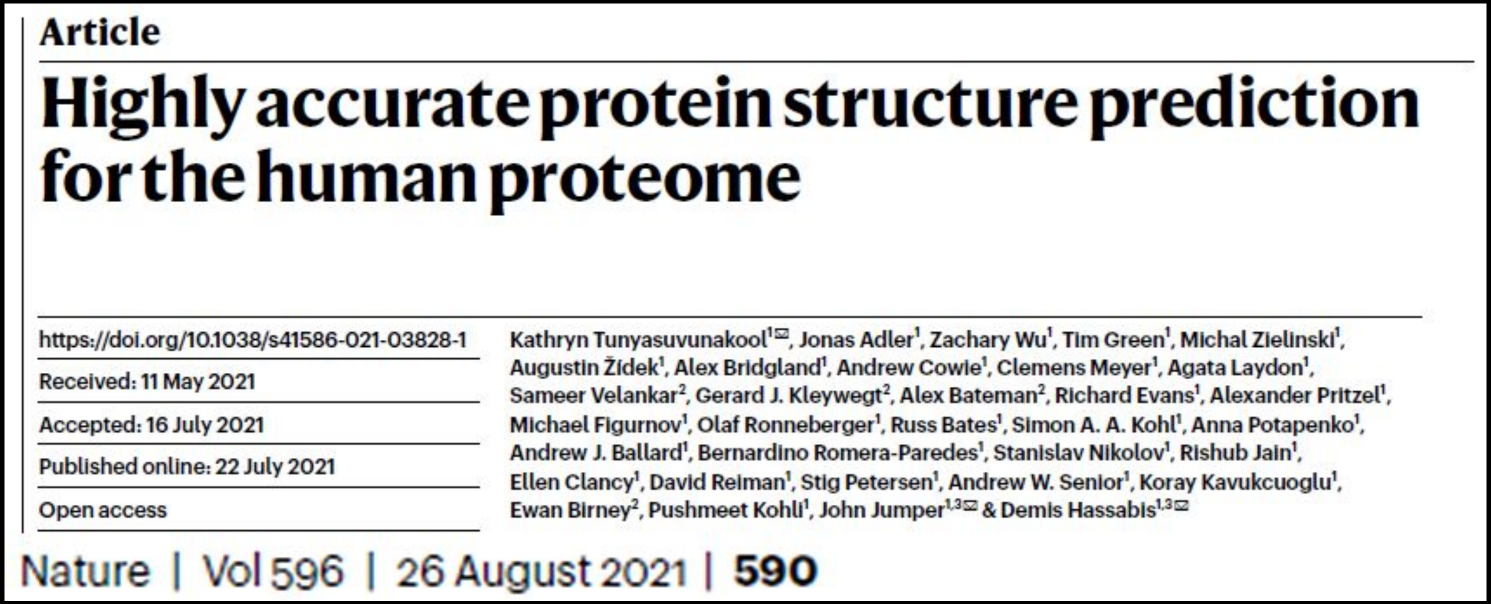
Accepted: 12 July 2021

Published online: 15 July 2021

Open access

John Jumper[1,4]✉, Richard Evans[1,4], Alexander Pritzel[1,4], Tim Green[1,4], Michael Figurnov[1,4], Olaf Ronneberger[1,4], Kathryn Tunyasuvunakool[1,4], Russ Bates[1,4], Augustin Žídek[1,4], Anna Potapenko[1,4], Alex Bridgland[1,4], Clemens Meyer[1,4], Simon A. A. Kohl[1,4], Andrew J. Ballard[1,4], Andrew Cowie[1,4], Bernardino Romera-Paredes[1,4], Stanislav Nikolov[1,4], Rishub Jain[1,4], Jonas Adler[1], Trevor Back[1], Stig Petersen[1], David Reiman[1], Ellen Clancy[1], Michal Zielinski[1], Martin Steinegger[2,3], Michalina Pacholska[1], Tamas Berghammer[1], Sebastian Bodenstein[1], David Silver[1], Oriol Vinyals[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Pushmeet Kohli[1] & Demis Hassabis[1,4]✉

Nature | Vol 596 | 26 August 2021 | 583



Article

Highly accurate protein structure prediction for the human proteome

https://doi.org/10.1038/s41586-021-03828-1

Received: 11 May 2021

Accepted: 16 July 2021

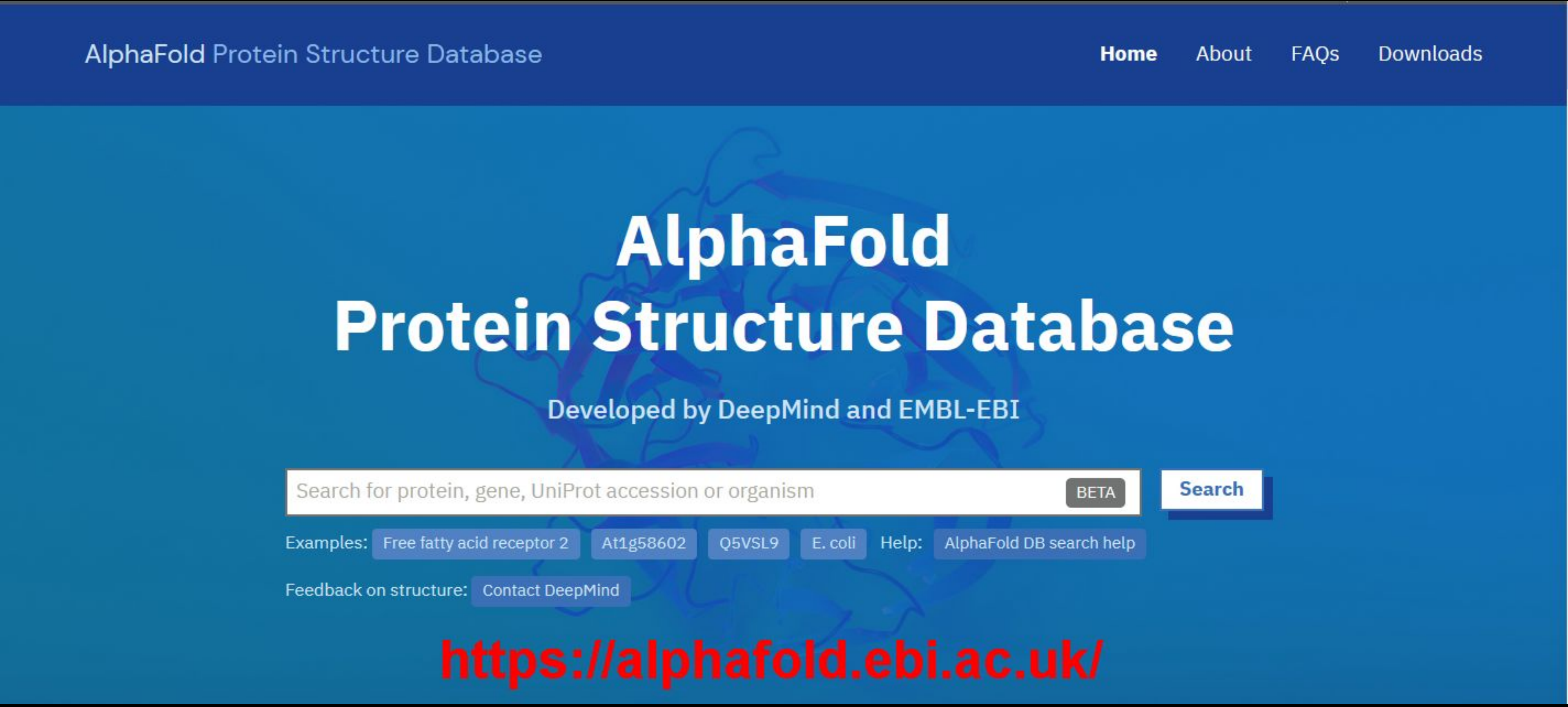
Published online: 22 July 2021

Open access

Kathryn Tunyasuvunakool[1]✉, Jonas Adler[1], Zachary Wu[1], Tim Green[1], Michal Zielinski[1], Augustin Žídek[1], Alex Bridgland[1], Andrew Cowie[1], Clemens Meyer[1], Agata Laydon[1], Sameer Velankar[2], Gerard J. Kleywegt[2], Alex Bateman[2], Richard Evans[1], Alexander Pritzel[1], Michael Figurnov[1], Olaf Ronneberger[1], Russ Bates[1], Simon A. A. Kohl[1], Anna Potapenko[1], Andrew J. Ballard[1], Bernardino Romera-Paredes[1], Stanislav Nikolov[1], Rishub Jain[1], Ellen Clancy[1], David Reiman[1], Stig Petersen[1], Andrew W. Senior[1], Koray Kavukcuoglu[1], Ewan Birney[2], Pushmeet Kohli[1], John Jumper[1,3]✉ & Demis Hassabis[1,3]✉

Nature | Vol 596 | 26 August 2021 | 590

The AF consortium database currently includes 992,316 predicted structures covering 48 organism proteomes and the majority of Swiss-Prot



AlphaFold Protein Structure Database

AlphaFold
Protein Structure Database

Developed by DeepMind and EMBL-EBI

Examples: Free fatty acid receptor 2 At1g58602 Q5VSL9 E. coli Help: AlphaFold DB search help

Feedback on structure: Contact DeepMind

https://alphafold.ebi.ac.uk/

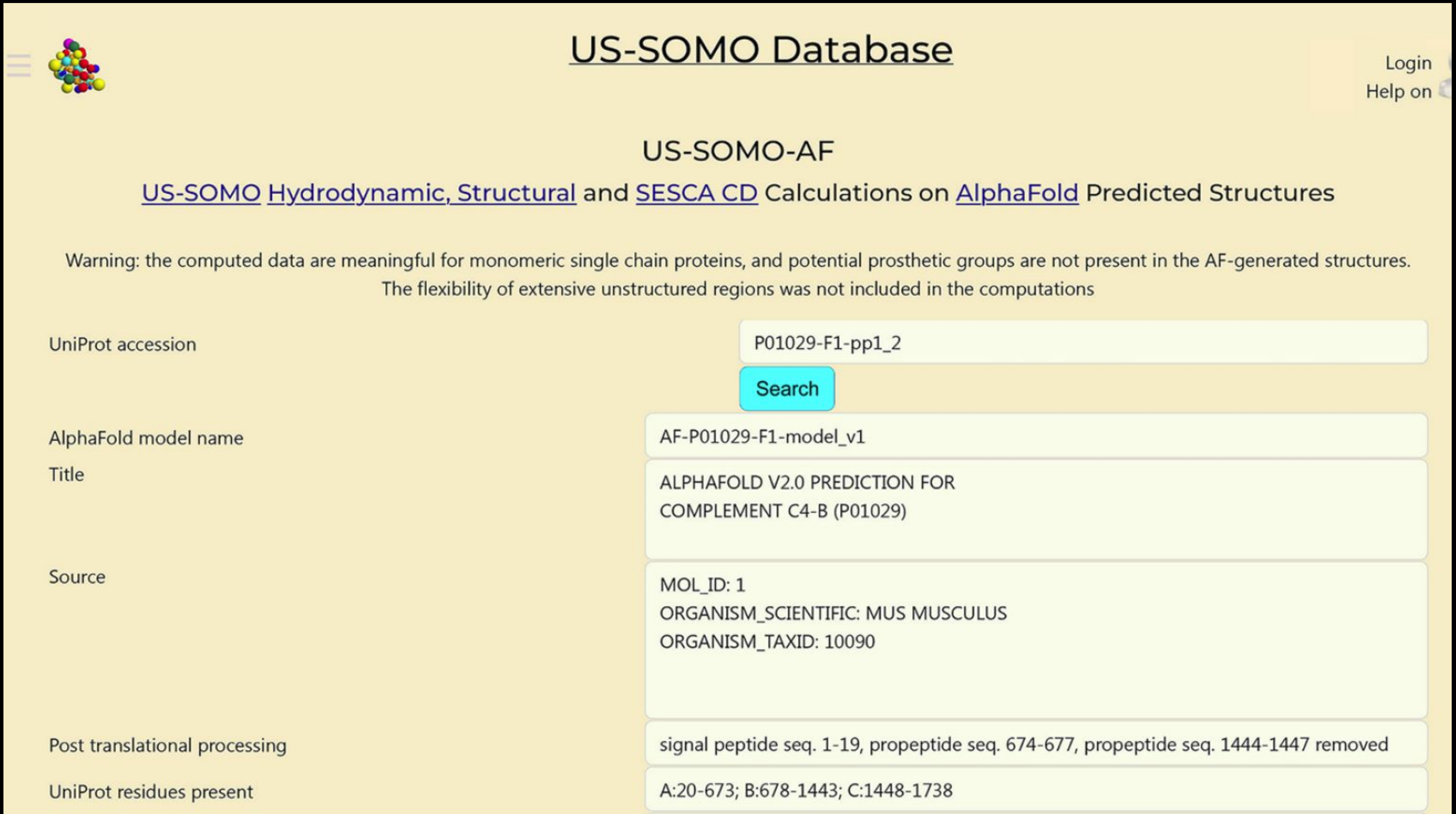
AF method and issues

- AF is based on an AI algorithm trained on the protein structures present in the PDB
- No thermodynamic/mechanistic approach, relies only on a deep learning process
- Potentially biased toward structures already present in the PDB
- Potentially unstructured regions are approximated with an “unique” conformation

Performing some rapid solution test on a predicted structure should be considered

- Verifying the secondary structure content by Circular Dichroism (CD) spectroscopy
- Assessing the overall shape compatibility by measuring hydrodynamic parameters, such as $D^0_{t(20,w)}$, $s^0_{(20,w)}$, $[\eta]$
- Using small-angle x-ray scattering (SAXS) methods to produce the pairwise distance distribution function $p(r)$ vs. r
- All these parameters/functions can be directly calculated from structures

To facilitate the comparison between measured and calculated parameters, we have computed them for the entire AF database, and placed them in the public-domain US-SOMO-AF database: https://somo.genapp.rocks/somoaf



US-SOMO Database

US-SOMO-AF

US-SOMO Hydrodynamic, Structural and SESCA CD Calculations on AlphaFold Predicted Structures

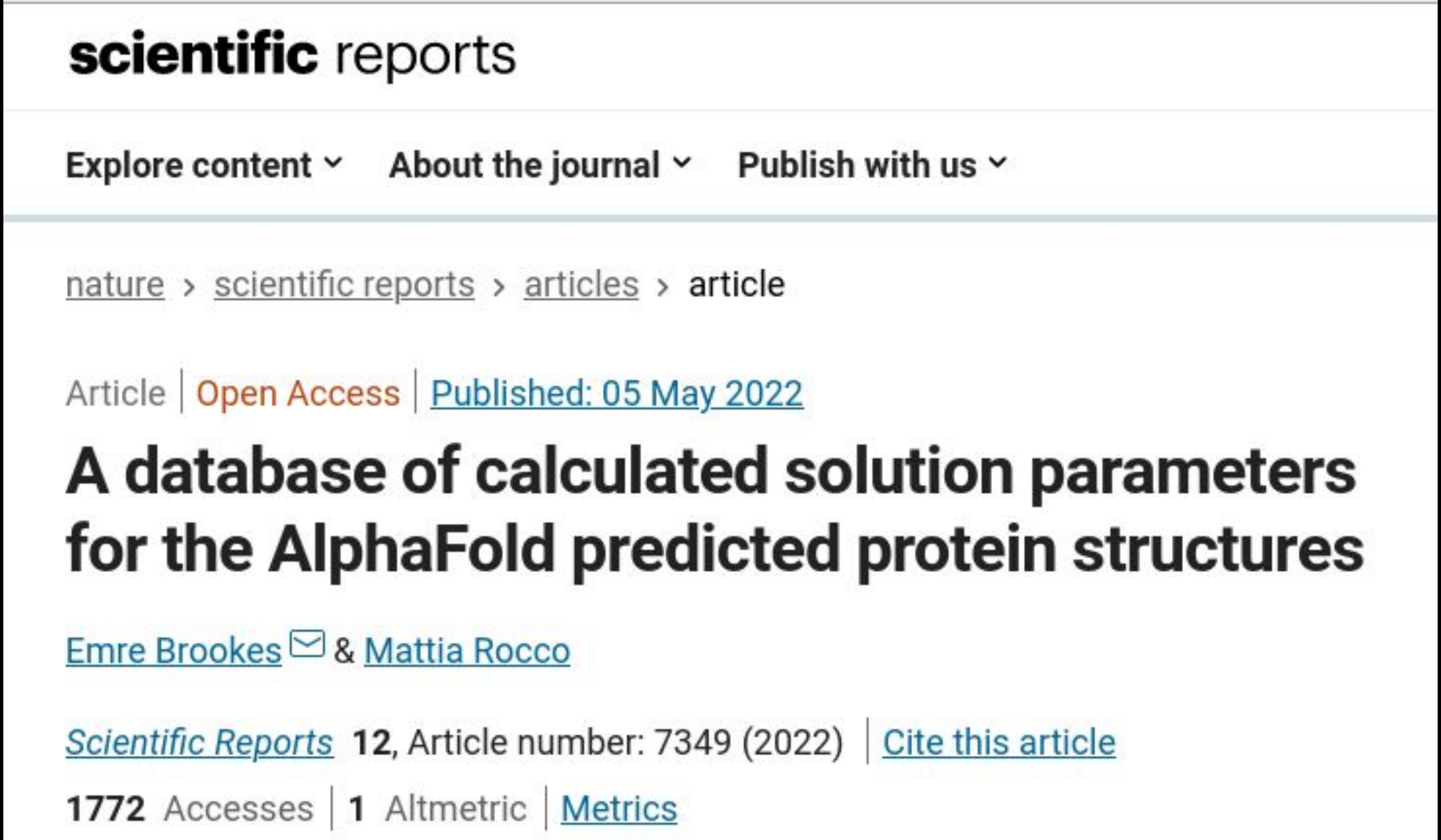
Warning: the computed data are meaningful for monomeric single chain proteins, and potential prosthetic groups are not present in the AF-generated structures. The flexibility of extensive unstructured regions was not included in the computations

UniProt accession	P01029-F1-pp1_2
AlphaFold model name	AF-P01029-F1-model_v1
Title	ALPHAFOLD V2.0 PREDICTION FOR COMPLEMENT C4-B (P01029)
Source	MOL_ID: 1 ORGANISM_SCIENTIFIC: MUS MUSCULUS ORGANISM_TAXID: 10090
Post translational processing	signal peptide seq. 1-19, propeptide seq. 674-677, propeptide seq. 1444-1447 removed
UniProt residues present	A:20-673; B:678-1443; C:1448-1738

To facilitate the comparison between measured and calculated parameters, we have computed them for the entire AF database, and placed them in the public-domain US-SOMO-AF database:

https://somo.genapp.rocks/somoaf

https://www.nature.com/articles/s41598-022-10607-z



scientific reports

Explore content ˅ About the journal ˅ Publish with us ˅

nature › scientific reports › articles › article

Article | Open Access | Published: 05 May 2022

A database of calculated solution parameters for the AlphaFold predicted protein structures

Emre Brookes & Mattia Rocco

Scientific Reports **12**, Article number: 7349 (2022) | Cite this article

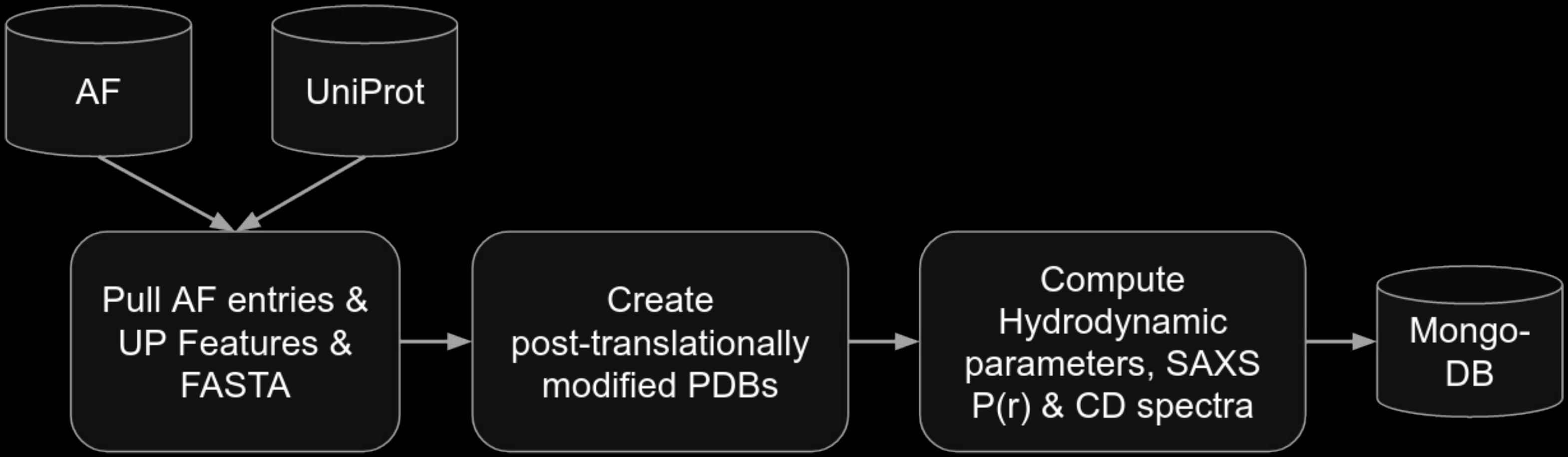
1772 Accesses | **1** Altmetric | Metrics

Methods

- The AlphaFold structures were predicted directly from the UniProt sequences, without any curing regarding post-translational modifications
- Based on the UniProt annotations, we have removed the Initiator Methionine, Signal Peptide, and Transit Peptide(s) from the AF structures. Permuted structures with/without Propeptide(s) were also generated (subtotal: ~110,000)
- CD spectra were computed using the SESCA program *https:// doi. org/10.1021/acs.jctc.9b00203*
- US-SOMO was used to compute the hydrodynamics (SoMo with overlaps + ZENO method) and the $p(r)$ vs. r using SAXS-related parameters

Methods

- Processing pipeline:

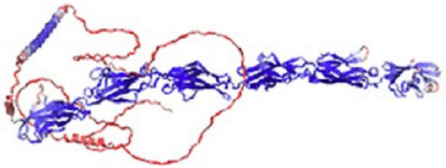

AF
UniProt
Pull AF entries & UP Features & FASTA
Create post-translationally modified PDBs
Compute Hydrodynamic parameters, SAXS P(r) & CD spectra
Mongo-DB

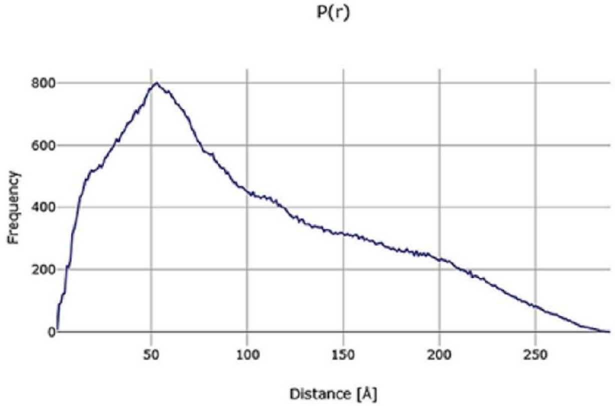
- Processing performed on resources:
  - University of Lethbridge & the Texas Advanced Computer Center
- Website:
  - Generated using the GenApp framework *https://genapp.rocks*
  - Hosted on NSF Jetstream2 *https://somo.genapp.rocks*
  - Allocated via NSF XSEDE

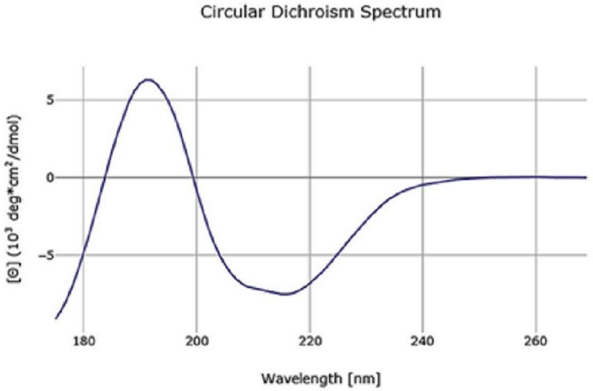
Some examples:

Organism	Mean AF % conf.	Signal peptide	Molecular mass [Da]	R_g [nm]	R_s [nm]	$[\eta]$ [cm^3/g]	Helix%	Sheet%
H. sapiens	75.64	1–28	98,141	8.42	6.56	23.3	9.2	25.5

Q9Y5H4:

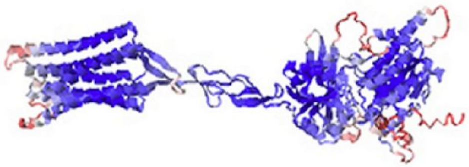


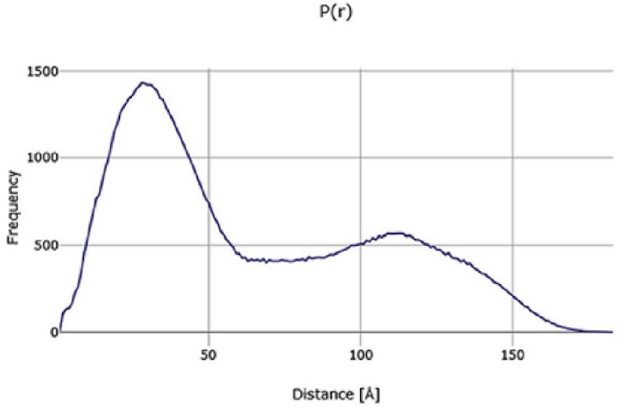


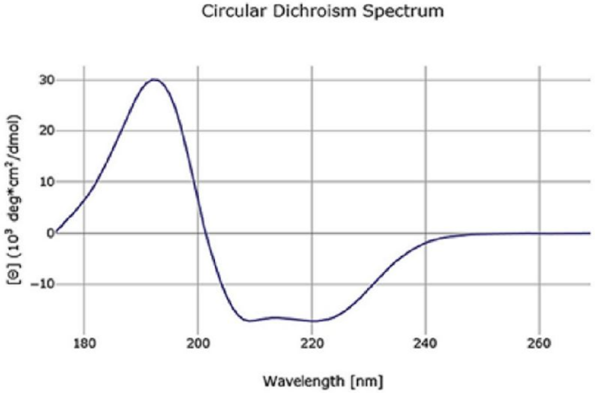


Organism	Mean AF % conf.	Signal peptide	Molecular mass [Da]	R_g [nm]	R_s [nm]	$[\eta]$ [cm^3/g]	Helix%	Sheet%
R. norvegicus	82.81	1–20	94,123	5.55	4.76	8.93	42.8	11.2

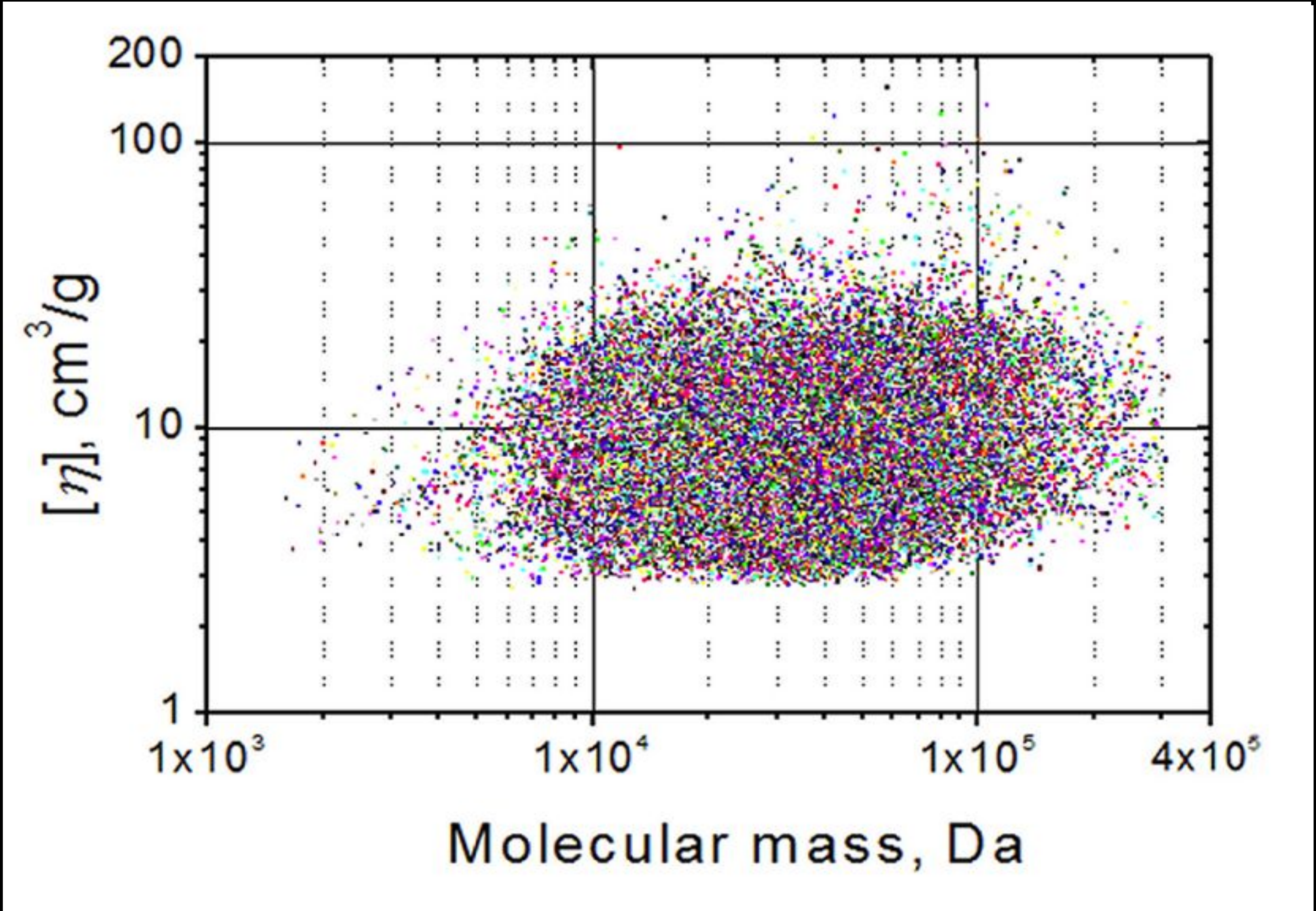
D3ZV97:



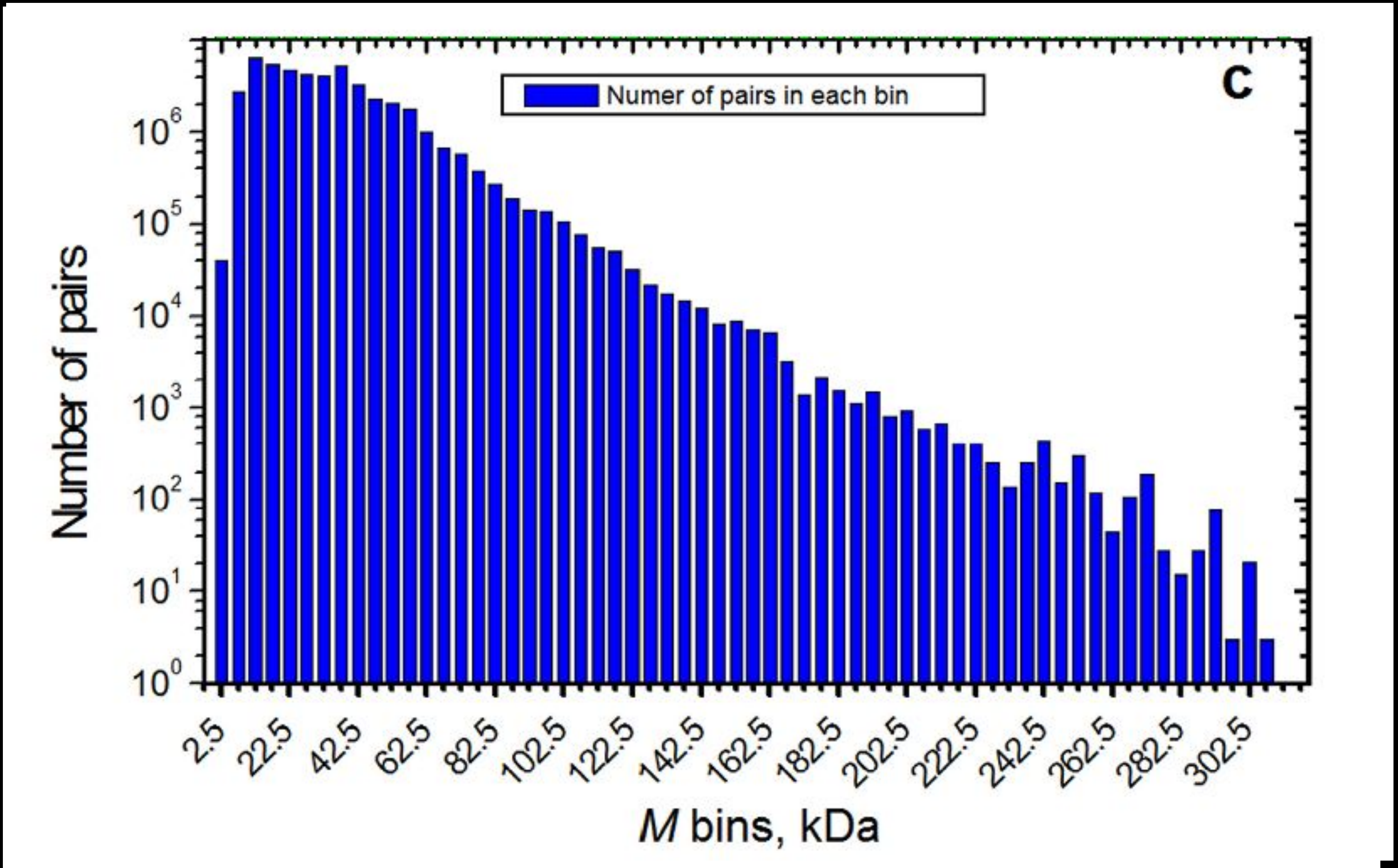




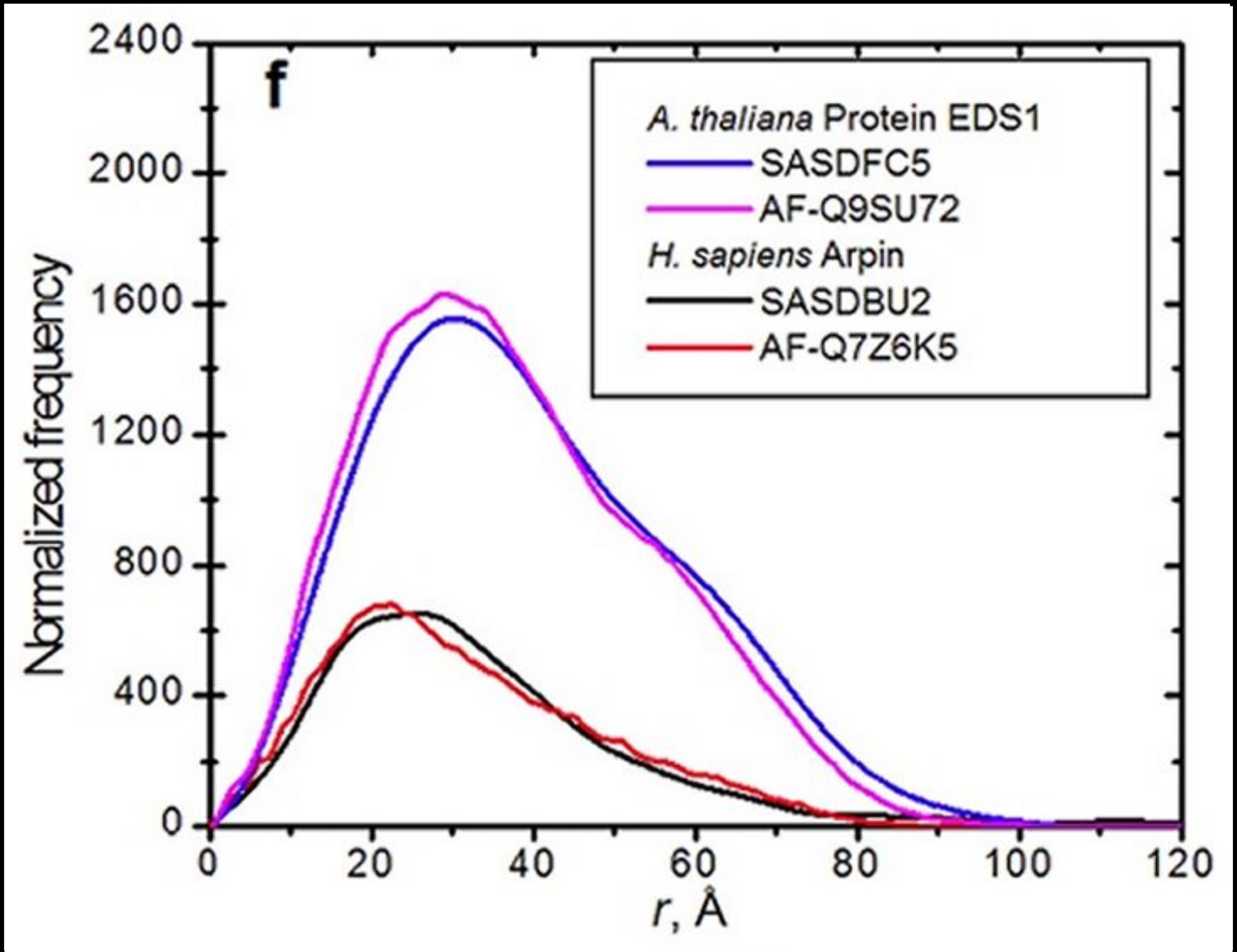
Analyzing the calculated hydrodynamic parameters for a subset of ~41,200 AF structures



Given an average experimental error of ±3%, what % of structures within 2x or 3x the average error can we distinguish within 5 kDa bins?

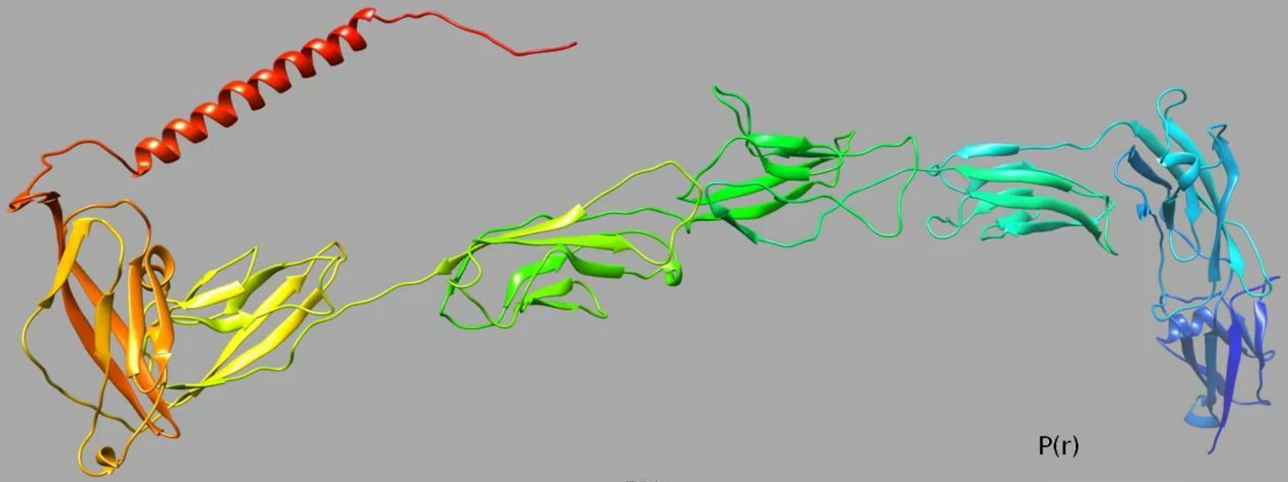


Comparison between *p*(*r*) vs. *r* derived from SAXS, and computed from AF (and PDB) structures



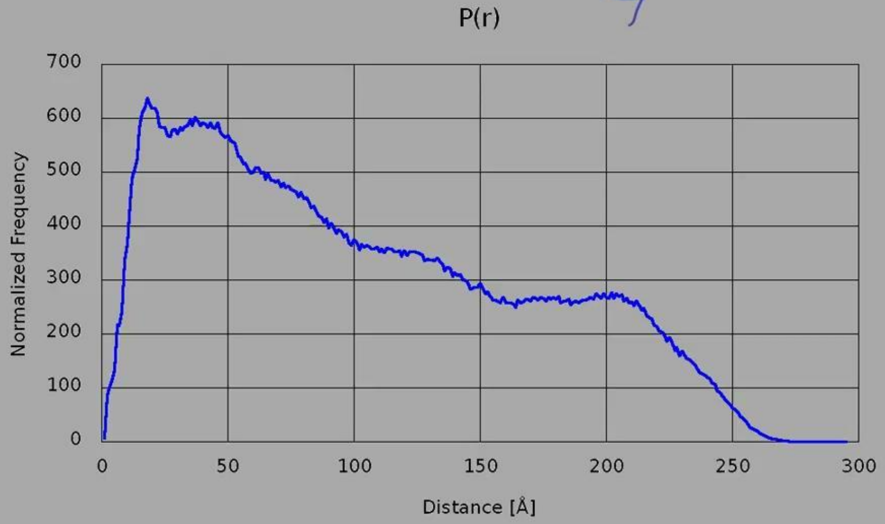
Effect of conformational variability on the hydrodynamic parameters and *p*(*r*) vs. *r*: a DMD simulation

O88338 Frame 001

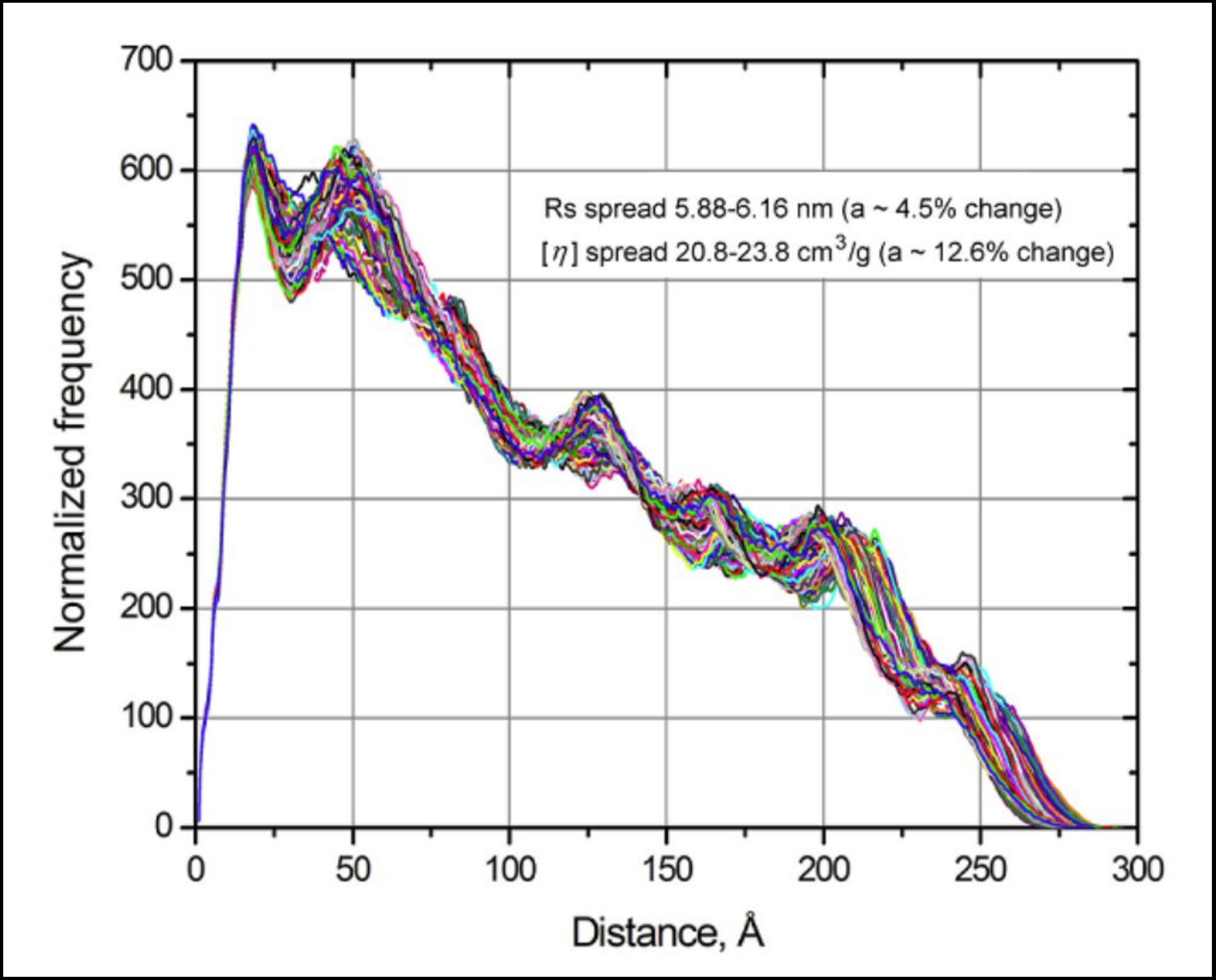


Rs = 5.88

[η] = 20.82 ± 0.33



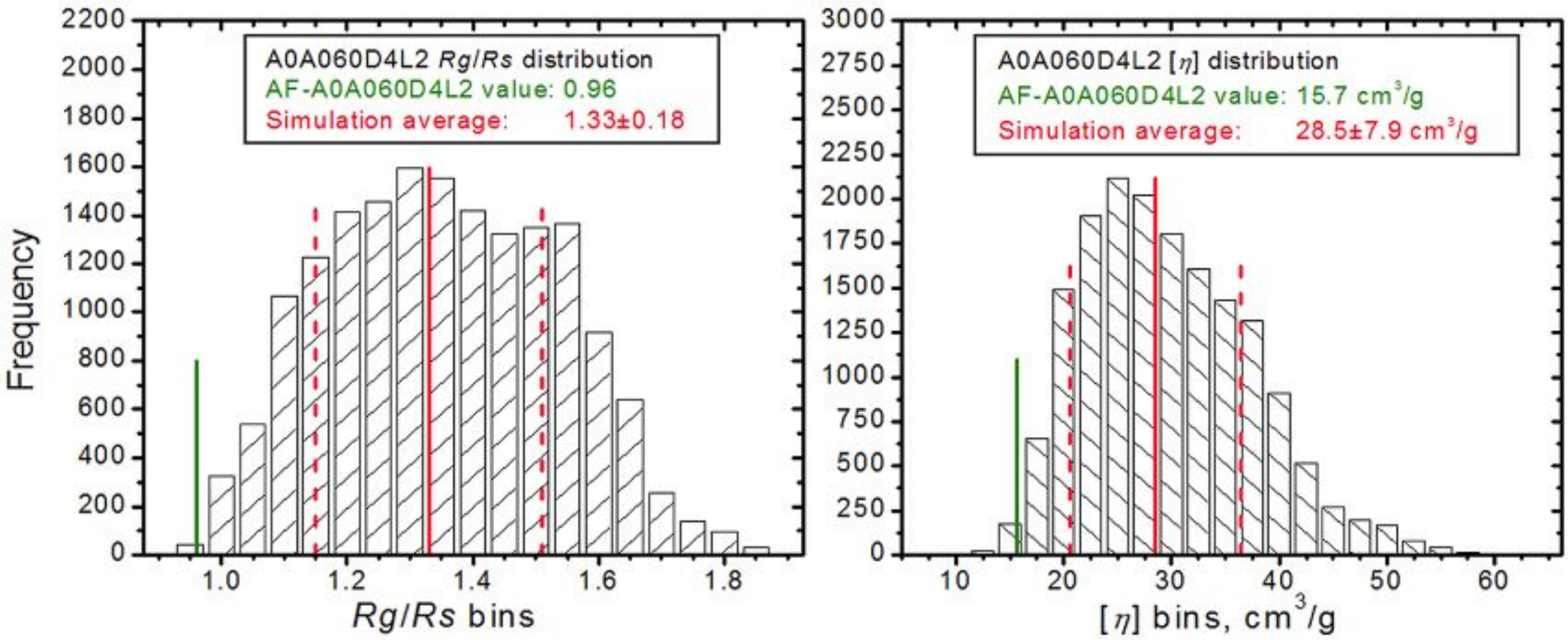
Effect of conformational variability on the hydrodynamic parameters and *p*(*r*) vs. *r*: a DMD simulation, summary



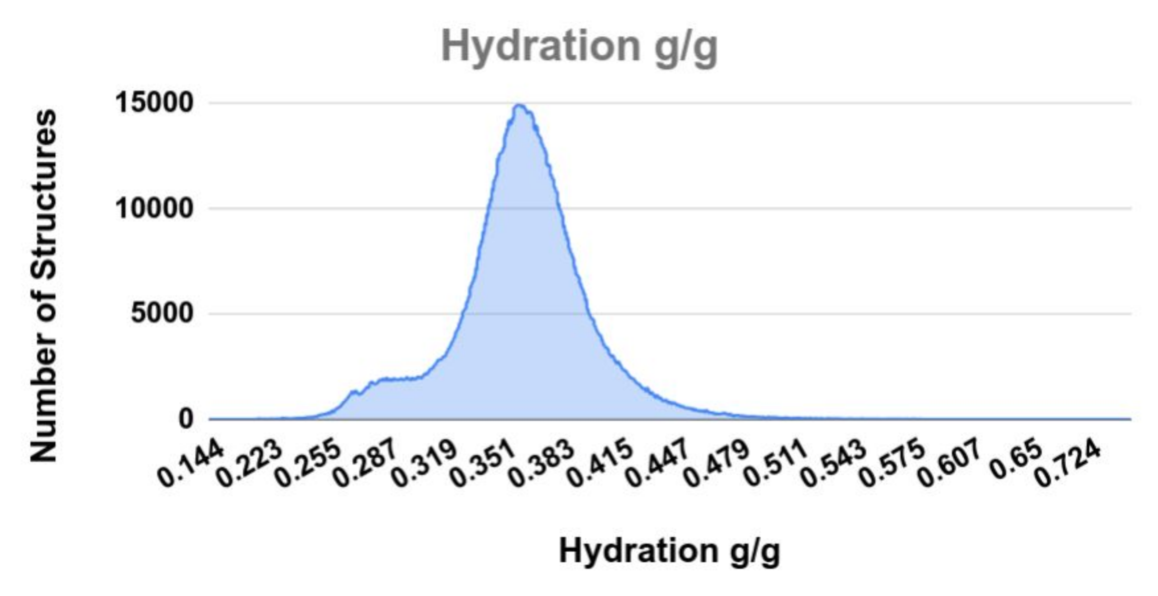
Effect of long unstructured regions on the hydrodynamic parameters. A Monomer Monte Carlo simulation on the 1-118 N-terminal residues of structure AF-A0A060D4L2

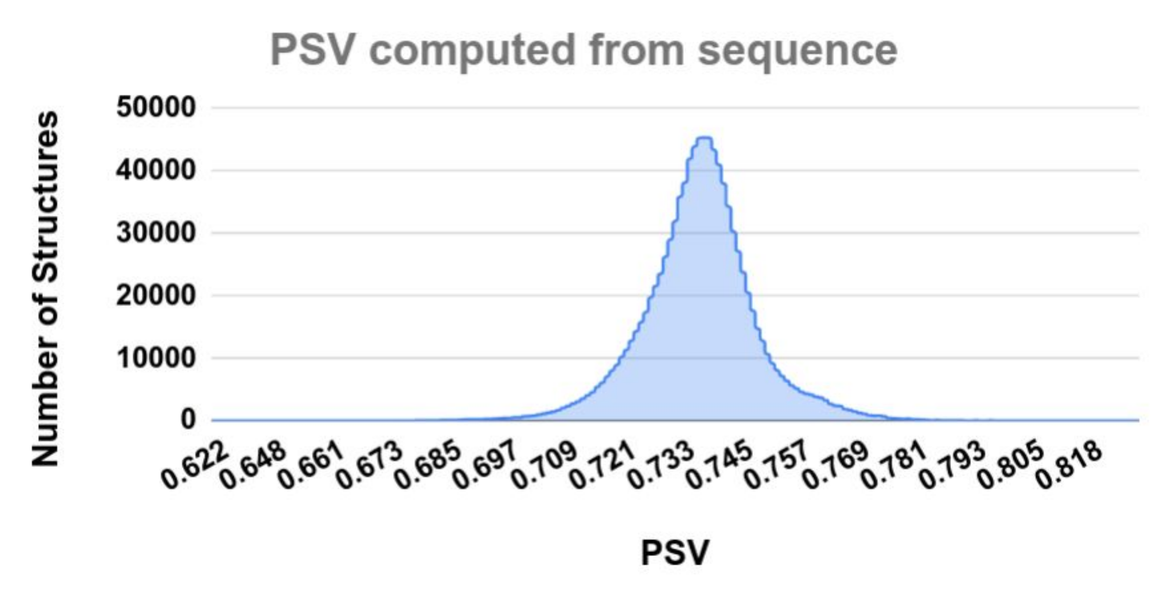
https://sassie-web.chem.utk.edu Curtis, et al. 2012 doi: 10.1016/j.cpc.2011.09.010

Results of the MMC simulation on structure AF-A0A060D4L2, >16,000 conformations

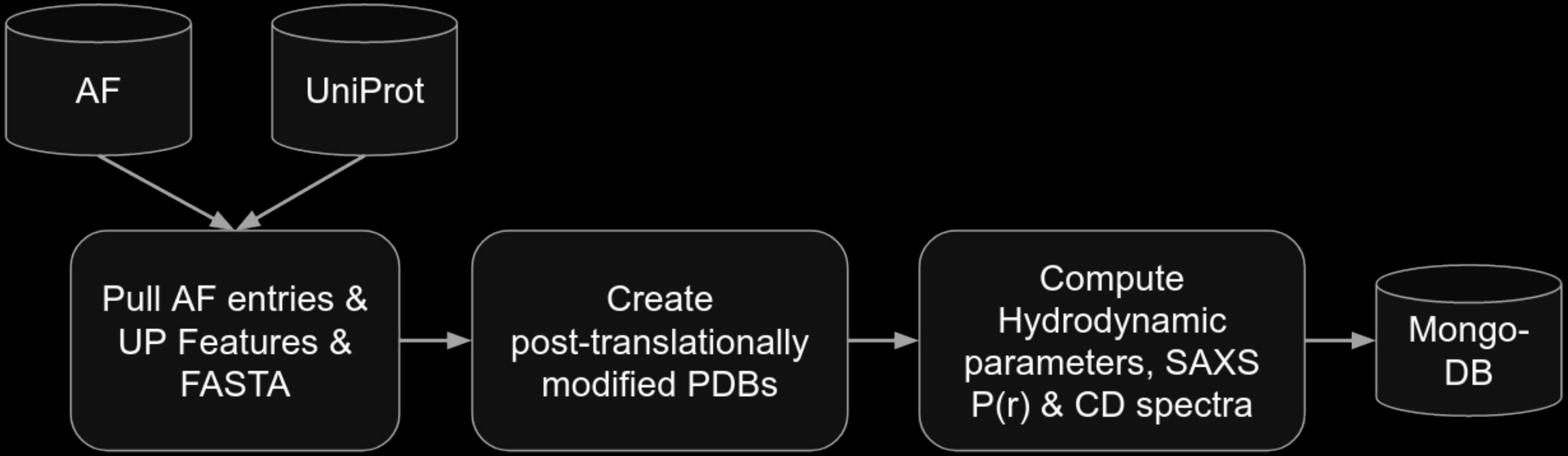


Database enables global studies





Processing pipeline enables additional calculations



AF

UniProt

Pull AF entries & UP Features & FASTA

Create post-translationally modified PDBs

Compute Hydrodynamic parameters, SAXS P(r) & CD spectra

Mongo-DB

Drawbacks & perspectives

- The current AF database release (v2) contains predictions for single chain structures only
- AF has released a program that can predict multiple-chains structures. If and when this will be generalized to produce an updated database, we can recalculate the parameters/functions for the new structures

Drawbacks & perspectives

- No prosthetic groups, such as carbohydrates, were taken into consideration by AF
- For carbohydrates, methods to predict their structure from composition are available and under continuous development in several laboratories. The biggest hurdle is to accurately predict the composition of carbohydrates and correctly store this information at the UniProt level. US-SOMO already handles carbohydrates, so updating the database will be possible
- The situation is obviously more complicated for the hundreds of other potential prosthetic groups

Drawbacks & perspectives

- Unstructured parts are represented as a single defined conformation in the AF predictions
- Correctly taking into account segmental or generalized flexibility is a much bigger issue. Molecular Dynamics - requiring huge computer power, Monte Carlo simulations or Brownian Dynamics, appear to be the best possibilities
- However, the data in the US-SOMO-AF database could raise “red flags”, and indicate that additional modeling work is required to further validate a predicted structure

Acknowledgments

- NIH/NIGMS R01GM120600 (Brookes, Demeler, Savelyev)
- NSF CHE-1265817, OAC-1740097, OAC-1912444 (Brookes).
- This work used the Extreme Science and Engineering Discovery Environment (XSEDE), which is supported by NSF Grant Number ACI-1548562 and utilized Jetstream at Indiana University through allocation TG-MCB170057 to Brookes.
- The authors acknowledge the Texas Advanced Computing Center (TACC) at The University of Texas at Austin for providing HPC resources that have contributed to the research results reported (http://www.tacc.utexas.edu).
- The computing infrastructure at the University of Lethbridge was funded by the Canada Foundation for Innovation (CFI-37589 to Demeler).
- We are grateful to P. Vachette (I2BC, Université Paris-Saclay, CEA, CNRS, Gif-sur-Yvette, France) for comments and suggestions.

Thank you for your attention

Emre Brookes
Department of Chemistry and Biochemistry, The University of Montana, Missoula, MT, USA
emre.brookes@umontana.edu

Mattia Rocco
Retired, Proteomics & Mass Spectrometry, IRCCS Ospedale Policlinico San Martino, National Institute for Research on Cancer, Genova, Italy
mattia.rocco@quipo.it