Agenda

Meeting title:	CCP4 Working Group 2 meeting		
Date:	Wednesday 8th February 2023	**Time:**	11:30 – 18.00
Location:	Birkbeck College, BCB308	**Zoom:**	link see email
Circulation:	ccp4wg2@stfc.ac.uk		
Attending:	Jonas Adler, Jon Agirre, Charles Ballard, Lucrezia Catapano, Grzegorz Chojnowski, Paul Emsley, Maria Fando, Deborah Harrus, John Jumper, Ronan Keegan, Oleg Kovalevskiy, Andrew Lebedev, Martin Malý, Stuart McNichoals, Garib Murshudov, Rob Nicholls, Andy Purkiss (AP), Randy Read, Dan Rigden, Pietro Roversi, Ivo Tews, Mihali Varadi, Sameer Velankar, Keitaro Yamashita		
Online:	Arnaud Basle, David Briggs, Ralf Flaig, Luiso Fuentes Montero, Vilmos Fulop, Robbie Joosten, Petr Kolenko, Gustavo Lima, Juan Mateos, Martin Noble, Nicholas Pearce, Mark Roe, Alan Roseman, Roberto Steiner, Johan Turkenburg, Ville Uski, Mihali Varadi, Melanie Vollmar, Pamela Williams, Keith Wilson, Marcin Wojdyr, Briony Yorke, Diangquan Yu		

1. Approval of minutes from the WG2 meetings 12/01/23
2. Chairs report (Ivo)
3. Machine learning, deep learning, AI and models

   Dan Rigden – Introduction: Lessons from CASP
   Jonas Adler – AlphaFold
   Randy Read – Phenix: synergistic AF model improvement

Lunch

4. AI and new models

   Dingquan Yu – AlphaPulldown
   Grzegorz Chojnowski – Check/FindMySequence for nucleic acids
   Ronan Keegan – What we have in CCP4: AI methods and MR

Discussion + tea

5. Extending refinement

   Nick Pearce – Multi-model-multi-data scenarios
   Helen Ginn – Torsion angle refinement with RoPE

6. Validation and refinement

   Paul Emsley – The new era with WebCoot and DesktopCoot
   Sameer Velankar – New deposition models and challenges

7. Take note of the date for the next meeting (proposal 9.3.22)
8. AOB

Minutes

1. **Approval of minutes from the online WG2 meeting 12/1/23 (Ivo Tews)**

The minutes from the online WG2 meeting 12/1/23 were approved.

2. **Chairs report (Ivo Tews)**

IT has chaired WG2 since 2014 and is now stepping down as he becomes WG1 chair.

The WG2 meetings have, during Covid, been converted to shorter and more frequent online formats. In 2023, we have had six meetings: four meetings in January, February, June and October dealt with Study Weekend organisation and news from the CCP4 core group (amongst other things); two meetings in March and April were focussed on Alphafold/MR and Data Processing; the focussed meeting planned for November was cancelled. In 2023, we already had one meeting on study weekend feedback and one on science, the latter in person. The suggestion is to keep such a rota, with 6-8 meetings over the year, and some of these in person.

The role of WG2 meetings is to oversee the organisation of the CCP4 study weekend, the dissemination in teaching crystallography activities such as meetings, workshops and schools, as well as documentation and web page presentation. As the work of WG2 is important for demonstrating novel software in order to provide a path of software into the suite, WG2 features various scientific content. A need is identified for WG2 to better engage with models we produce, their validation and deposition, incorporating our efforts on ligands and interaction with the PDB.

The tasks that WG2 now performs have grown over the years. Because of the increased workload the suggestion discussed in CCP4exec in January was to split the work and appoint two WG2 chairs. We are happy that Jon Agirre and Arnaud Basle have agreed to head WG2 into the future.

3. **Machine learning, deep learning, AI and models (chair: Dan Rigden)**

a) **Introduction: Lessons from CASP (Dan Rigden)**

Dan gave an overview of the results of CASP15. DeepMind did not take part in the competition, however they did send their predictions in. Most of the top scoring solutions were based on AlphaFold or some other deep learning software (RoseTTAFold). Performance seems to have reached a plateau. There was talk that perhaps the targets of CASP15 were a bit too hard.

b) **AlphaFold (Jonas Adler)**

In this talk, Jonas focuses on the elements that make AlphaFold work. He discussed details of the AlphaFold architecture (those fine details that are often overlooked in the appendix of the paper), e.g. how AF is using statistical patterns to predict protein structure, and explains why AF can often generalise in surprising ways. He focussed on parts of the neural network that contribute most to the successful prediction, such as trunk, and mentioned other parts that receive heightened attention but are actually less important. Jonas gave an outlook on AlphaFold development.

Folding protein structures by simulating physics is hard. Folding@home did not produce great advances. There are many factors that affect the final product, including interaction with the many things present in the cell (e.g. chaperones). We can do better by using evolutionary information.

About 40k unique structures/proteins in the PDB (170k total at the time of training). Very little data, so requires some complex processing, not as straightforward as other projects which require just a few lines of code. Evoformer uses Transformers (Jonas thinks it's the main innovation behind AlphaFold). AlphaFold and other models fail quite badly using only sequence data; genomics data is very important. The only way to break AlphaFold is to have no genomics data. Structure module catches people's attention at ML conferences, but it's not very important; directly projecting the structure from the trunk produces almost as good results.

Next steps include understanding dynamics/multistate and coping with disruptive mutations.

c) Phenix: synergistic AF model improvement (Randy Read)

CASP14: almost all models that had diffraction data could be solved with the predicted AF2 models. Construct design is now informed by AF2. MR is now easier and will almost always succeed. Still some problems: too many copies in AU, low resolution, crystal pathologies. Cryo-EM: accelerate initial map interpretation.
Randy is presenting new work on making AF work better by feeding experimental information. Dataset: old SAD structures. First prediction, then rebuild the structure using Cryo-EM map, then feed that rebuilt model as template to AF2. First model was 90% complete, the second prediction with 99% of Calphas within 2 Å.

Garib: one structure was not solved, it would be interesting to see which one.
Rob: there is a paper from Stephen Burley saying that 3.5 Å resolution structures are more reliable than AF2 models. Many people in the room would love to see the evidence.

Tom says you have to treat AF2 models as a really good hypothesis.
There is talk about having AF2 process experimental data from other techniques, such as NoEs from NMR, images, classes or maps from Cryo-EM.
Pietro: "I don't find that colab gives me access to the MSA, I don't know what sequences went into the prediction".
John Jumper: "we designed our colab for delivering the kind of precision we had achieved at CASP14". Also: AlphaFold-multimer does not rely on evolution.

4. AI and new models (chair: Ronan Keegan)

a) AlphaPulldown (Dingquan Yu)

AlphaFold has a few shortcomings: pipeline is hard on CPU, sometimes fails to predict interfaces that are known. There is a need for a tool that analyses the properties of the models and summarises them. And there is a need to streamline the pipeline, allowing for modelling fragments of a protein conveniently.

AlphaPullDown runs a regular AF2 process on CPU, then either goes into pulldown mode, homo-oligomer mode, custom mode or all_vs_all mode, and then creates a Jupyter notebook with the results, plus some tables.

Advantages: no need to recalculate MSA, the original residue index in the final model, Domains can contain multiple discontinuous regions.
Interfaces are evaluated by third-party tools (iPTM, pDockQ, PI_score).

Need to work on better scoring functions.

b) Check/findMySequence for nucleic acids (Grzegorz Chojnowski)

findMySequence uses sequence reassignment as model validation. Slides the input sequence over the built model and tries to validate sequence assignment, e.g. finding gaps/breaks in the chain. Grzegorz shows an example of a proline in the middle of a helix, which introduces a register error. Also another MX example where even with the sequence being very wrong, the sigmaA calculation still yields a map that is usable by the neural network. There are quite a few very big structures, where it is obvious that the depositors could not pay attention to all residues. Many of these errors can get corrected by using AF2.
There is a modified workflow for DNA/RNA where base pairing is introduced. Same procedure.

Garib: have you tested nucleic acids on anything other than ribosomes? Mainly ribosomes.
Keith: Any way fixed structures could be made available?
Sameer: 3D-beacons provides this service.

c) What we have in CCP4 - AI methods and MR (Ronan Keegan)

Generating search models for MR is something very useful for CCP4. ESMFold, Colabfold, RoseTTAFold can provide this.
Processing models for MR involves conversion of confidence scores into B-factors, elimination of low confidence regions and splitting of domains depending on PAE.
RoseTTAfold used to offer RMSDs, but now are producing pLDDT in the newer versions (due to compatibility with CASP?).
MrParse searches the PDB, AFDB and ESMAtlas databases for search models.
There is an interface to Randy's tools for processing AF2 models in CCP4i2, there is SliceNDice (clustering using Scikit Learn for machine learning) in CCP4 Cloud.
Ronan demos the CCP4 Cloud (remote version, with computations done at Harwell) tools for AF2 models.
File import -> Define AU -> MRParse.

5. Extending refinement (chair Garib Murshudov)

a) Multi–model–multi–data scenarios (Nick Pearce)

Multiple sources of uncertainty in crystallography. Nick talks about his new group in Sweden, and his research aims, which include working on expanding PanDDa, and working on multi-conformer modelling.

Mulch is a multi-dataset analysis platform for structural biology. Done in collaboration with Helen Ginn. Pandemonium does multi dataset modelling and refinement.

b) Torsion angle refinement with RoPE (Helen Ginn)

Local effects propagate through the protein backbone upon ligand binding. In an atom-centric model, atoms are assigned positions but motions are described clumsily. Helen takes a bond-centric approach where torsion angles are refined. Torsions make it easier to propagate long range movements. Helen says this approach reduces model bias, although R-factors are still higher than those of atomic models in most cases. However some low resolution structures show a benefit from using this approach.
Current shortcomings include naivety in the heuristics used for tweaking the torsion angles. RoPE stands for Representation of Protein Entities. Written in order to support refinement using multiple datasets.
URL: rope.hginn.co.uk
Garib: how do you deal with disulfides and sugars? (needs more discussion). Also, how many significant SVD values do you get? Answer: case dependent.
Requirements: SDL2 is not yet in CCP4 – it's small and relatively simple. FFTW3 is already in the suite. Gemmi and nlohmann-json (header-only).

6. Validation and refinement (Sameer Velankar)

a) The new era with WebCoot and DesktopCoot (Paul Emsley)

Push: moving to Python 3. Huge infrastructure change, involving moving to GTK3, newer openGL, etc. Development of Coot 1: much better graphics, faster. Problems with Mac OS X, Charles discovered an environment variable that when tweaked sped up the whole thing. Acceptable performance now. Many new representations that make use of the new graphics engine include clashes, rama scores and rota scores. All clickable. Rama plot updates in real time. Ability to colour maps by correlation coefficients.
Moorhen (web coot): the standard Coot module is very heavy – libcootsumo is 370Mb and links 195 other libraries. Now there is a headless API that can be compiled with webassembly. Includes a reduced set of functions, but new ones are being added daily.
Coming: integration with CCP4 Cloud and CCP-EM Doppio.

URL: https://moorhen.org (beware, it's addictive)

b) New deposition models and challenges (Sameer Velankar)

ORCID based login for deposition. Users need to log in to ORCID to first generate a key for the PDB/EMDB Deposition API. The key is a JSON token containing the ORCID ID of the user authenticated. The key will expire too.
REST API calls to Create, Get deposition info, List depositions by user, Manage files, Process deposition, Manage access.
Linking multiple structures – "investigation". Good for fragment screening, probing dynamics, time-resolved, XFEL. Data standards based on the mmCIF framework. Facilitate deposition and biocuration of multiple structures and associated datasets.
Martin: great progress, Sameer.

7. **Take note of the date of the next meeting.**

Proposed date: proposal 9th March (online). << This has been brought forward to the 2nd of March.

8. **AOB**

Ivo Tews thanks WG2 for the trust and the opportunity to chair WG2.