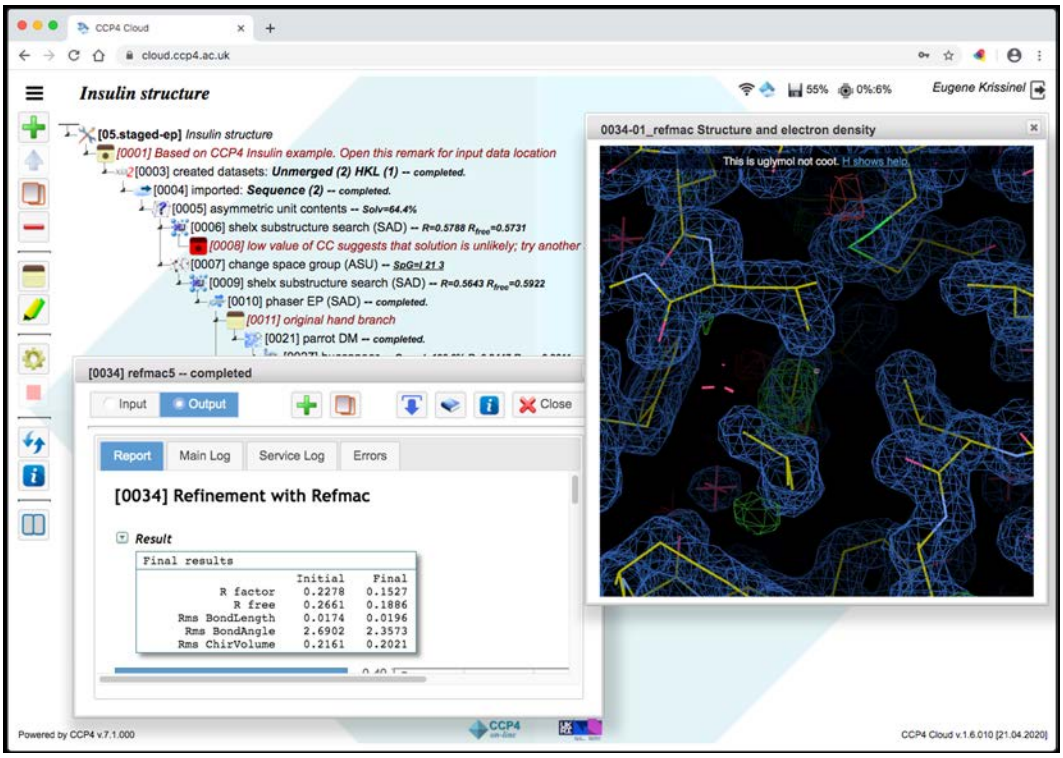
RSC INTEREST GROUP
CHEMICAL INFORMATION AND
COMPUTER APPLICATIONS GROUP

NEWSLETTER summer 2020



Insulin structure

```
[05.staged-ep] Insulin structure
  [0001] Based on CCP4 Insulin example. Open this remark for input data location
    [0003] created datasets: Unmerged (2) HKL (1) -- completed.
      [0004] imported: Sequence (2) -- completed.
        [0005] asymmetric unit contents -- Solv=64.4%
          [0006] shelx substructure search (SAD) -- R=0.5788 Rfree=0.5731
            [0008] low value of CC suggests that solution is unlikely; try another
          [0007] change space group (ASU) -- SpG=I 21 3
            [0009] shelx substructure search (SAD) -- R=0.5643 Rfree=0.5922
              [0010] phaser EP (SAD) -- completed.
                [0011] original hand branch
                  [0021] parrot DM -- completed.
```

[0034] refmac5 -- completed

[0034] Refinement with Refmac

Result

```
Final results
                    Initial    Final
         R factor    0.2278   0.1527
           R free    0.2661   0.1886
 Rms BondLength      0.0174   0.0196
  Rms BondAngle      2.6902   2.3573
 Rms ChirVolume      0.2161   0.2021
```

0034-01_refmac Structure and electron density

This is uglymol not coot. H shows help.

Above: The CCP4 Cloud GUI (see *The Collaborative Computational Project Number 4 (CCP4) Release 7.1*, page 10)

CICAG aims to keep its members abreast of the latest activities, services, and developments in all aspects of chemical information, from generation through to archiving, and in the computer applications used in this rapidly changing area through meetings, newsletters and professional networking.

Chemical Information & Computer Applications Group Websites:

http://www.rsccicag.org
http://www.rsc.org/CICAG

QR Code

Linked in® http://www.linkedin.com/groups?gid=1989945

https://twitter.com/RSC_CICAG

Table of Contents

Contributions to the CICAG Newsletter are welcome from all sources - please send to the Newsletter Editor:
Stuart Newbold, FRSC, email: stuart@psandim.com

Chemical Information & Computer Applications Group Chair's Report

Contributed by RSC CICAG Chair Dr Chris Swain, email: swain@mac.com

The COVID pandemic has cast its shadow over the start of 2020, and our thoughts are with all those who have been affected. On a more practical note the RSC has a Community Fund that can be used for support in these difficult times.

CICAG social media is becoming an increasingly active way for communicating with members during lockdown. Our Twitter feed, (now being run by new recruit Jack Simpson from Liverpool University) with nearly 1000 followers, and LinkedIn with 400 followers gaining popularity. Importantly, the social media feeds provide an opportunity for communication with both RSC and non-RSC members around the world. The CICAG website is often updated and we would be very interested to hear suggestions for additional content.

CICAG have cancelled three events that were due to take place this year. The SCI-RSC Workshop on Computational Tools for Drug Discovery due to take place in Leeds has been converted into a regular "Wednesday Workshop" which was delivered as a free online event every Wednesday afternoon between 3 and 5 pm until very recently. The first two workshops proved really popular, with a couple of hundred people logged in. We are looking at adding additional workshops. Online registration is free and a number of the vendors offer limited time licences at the time of the workshops.

The CICAG and University of Cambridge Postgraduate Cheminformatics/CompChem Day in July has now been postponed to 2021. I'd just like to thank the students and post-docs in Cambridge who put a lot of effort behind the scenes organising the event, and I'm sure their efforts will bear fruit in 2021.

We have recently taken the decision to change the AI in Chemistry Meeting in September from a physical meeting to a virtual event. We are going to explore how we might take advantage of some of the online technologies to offer different opportunities for interaction. There will be more details in the future #AIChem20.

We have continued planning a November event on Open Chemical Science, originally to be at Burlington House, London, UK, but now reconstituted as a 5-day online event, further details of which are outlined later in this newsletter.

In the last Newsletter, I mentioned that we were thinking of reverting to a spring and autumn timetable for publication. With the advent of COVID-19, we have decided to go ahead with a summer edition to try and provide members with something of interest during these difficult times. This edition once again includes external contributions and we are grateful to both Tudor I Oprea, as well as Eugene Krissinel and other members of the CCP4 team. Once again, I'd like to invite contributions that would be of interest to the CICAG community.

Whilst RSC members can join up to 3 interest groups for free, in practise many members do not take up this opportunity. You can make a request to join a group via email (membership@rsc.org) or telephone (01223 432141).

CICAG Planned and Proposed Future Meetings

The table below provides a summary of CICAG's planned and proposed future scientific and educational meetings. For more information, please contact CICAG's Chair, Dr Chris Swain.

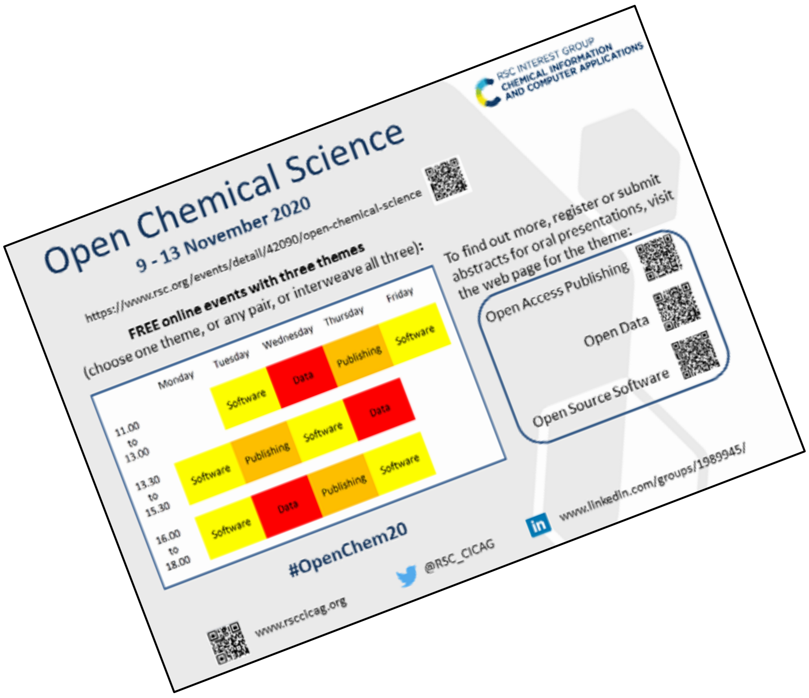
Meeting	Date	Location	Further Information
3rd Artificial Intelligence in Chemistry Meeting	28-29 Sep 2020	Virtual event	Joint event from RSC-CICAG and RSC-BMCS division
'In Silico Toxicology' Network Meeting 2020	30 Sep 2020	Virtual event	Meeting supported by RSC-CICAG
Open Chemical Science	9-13 Nov 2020	Virtual event	Five days of online events exploring sharing and collaborative developments
Big Data	TBD	TBD	Proposed joint meeting with the SCI)

Open Chemical Science – Free Online Webinars and Workshops in November 2020

Contribution from CICAG Committee member Dr Helen Cooke, email: helen.cooke100@gmail.com

There is currently much discussion about how access to data, information, knowledge and the software tools essential for 21st century chemical science research can be increased and simplified while ensuring that quality is maintained. Significant progress is being made, but the myriad of policies, repositories, standards, sources, tools and subscription models is at the same time complicating the landscape.

To address the many questions concerning the rapidly evolving open research landscape, towards the end of last year (before Covid-19), the CICAG Committee started to plan a series of three one-day events in November 2020, covering open access publishing, open data and open source software. The aim was to examine the benefits, risks and likely future developments and their impact on research. The aim hasn't changed but, due to Covid-19, the events will now be run as a series of online webinars to take place over five days, from 9-13 November 2020. We have decided to make them free of charge - this has been possible because the cost of running online events is much less than for face-to-face meetings.



RSC INTEREST GROUP
CHEMICAL INFORMATION AND COMPUTER APPLICATIONS

Open Chemical Science

9 - 13 November 2020

https://www.rsc.org/events/detail/42090/open-chemical-science

FREE online events with three themes
(choose one theme, or any pair, or interweave all three):

	Monday	Tuesday	Wednesday	Thursday	Friday
11.00 to 13.00		Software	Data	Publishing	Software
13.30 to 15.30	Software	Publishing	Software	Data	
16.00 to 18.00	Software	Data	Publishing	Software	

To find out more, register or submit abstracts for oral presentations, visit the web page for the theme:

Open Access Publishing

Open Data

Open Source Software

#OpenChem20

@RSC_CICAG

www.linkedin.com/groups/1989945/

www.rsccicag.org

CICAG realise that online events will reduce the networking opportunities provided by face-to-face meetings, but we hope they will appeal to a wider range of participants due to time savings and the absence of travel expenses. The limit on the number of participants will also be higher. A specific advantage for the open-source software sessions, which will be practical workshops, is that attendees can join as many sessions as they wish - the face-to-face events would have necessitated restrictions due to the limitations on physical space and time available.

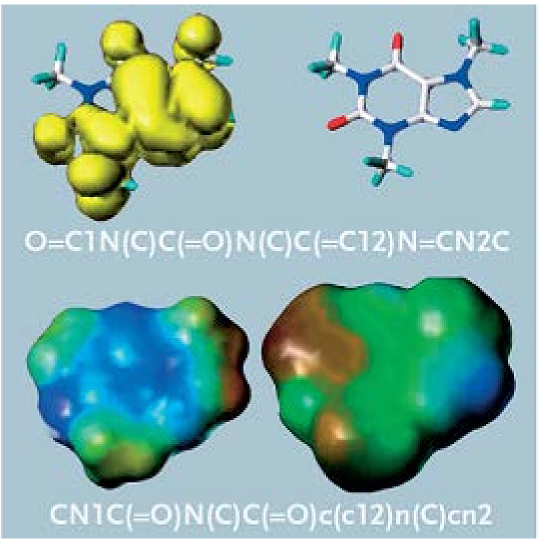
The names of speakers confirmed so far are available on the web pages linked to below. We are also inviting abstracts from potential speakers for the Open Access Publishing and Open Data sessions – a form is available via the web pages. As the programme evolves over the coming months, sub-themes for individual sessions will be confirmed and more speakers' names added to the web pages. Delegates will be able to register to join as many or as few sessions as they wish.

For further information and to register see:
Open Access Publishing: https://www.rsc.org/events/detail/43178/open-access-publishing-for-chemistry
Open Data: https://www.rsc.org/events/detail/43179/open-data-for-chemistry
Open Source Software: https://www.rsc.org/events/detail/43180/workshop-on-open-source-tools-for-chemistry

If you have specific questions please contact Gillian Bell cicageventsmanager@gmail.com

Time to add "Cheminformatics" (*) to Keywords Indexing Science

Contribution from ACS CINF/COMP/MEDI member Tudor I Oprea, MD PhD, email: toprea@salud.unm.edu


O=C1N(C)C(=O)N(C)C(=C12)N=CN2C
CN1C(=O)N(C)C(=O)c(c12)n(C)cn2

Above: Book cover for *Chemoinformatics in Drug Discovery*, edited by TI Oprea, Wiley-VCH 2005, illustrating the many faces of caffeine.

Our ability to rapidly and reliably process chemical information using computers has progressed significantly in the past five decades. Rapid processing of chemical information is needed when handling chemical inventories and when performing virtual screening; indeed regulatory agencies and legislators also need to process chemical information. Particularly in the last 25 years, cheminformatics and computer-aided processing of chemical information have emerged as a critical activity, rooted in science and requiring specialised skills.

We have (in no particular order):

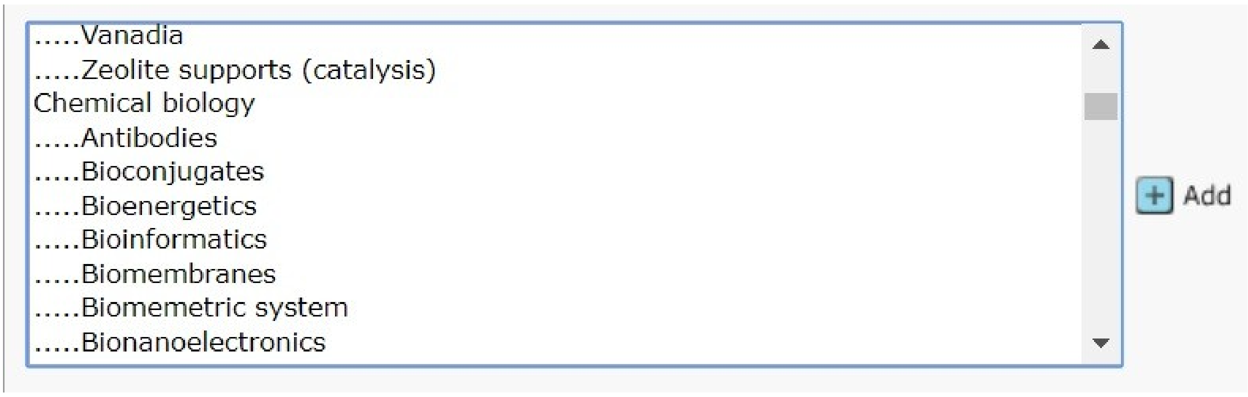
1. Specialised scientific journals, such as Journal of Chemical Information and Modeling, Journal of Computer-aided Molecular Design, Journal of Cheminformatics (to name a few) and (many) specialised books

2. Dedicated scientific societies, e.g., the QSAR, Chemoinformatics and Modeling Society, or divisions such as the ACS Division of Chemical Information (CINF) and the RSC Chemical Information and Computer Applications Group
3. Specialised on-line chemistry resources such as PubChem, SciFinder, ChemSpider and ZINC
4. Bioactivity resources like ChEMBL and GuideToPharmacology
5. Drug databases such as DrugBank and DrugCentral
6. Chemically aware online patent search systems such as SureChEMBL and GooglePatents
7. Specialised software, both under the license/subscription and open-access models
8. Vocabulary terms specific to cheminformatics (SMILES, SDF, InChI, QSAR, LogP, Lipinski rules – to name a few)

   ... and more.

Yet *cheminformatics (chemoinformatics, chemiinformatics and sometimes chemical informatics, etc.)* remains absent as a keyword in scientific journals. Below is a screenshot from one of the main chemistry journals (via the ManuscriptCentral™ interface). It matters not which journal, because the problem is – to the best of my knowledge – ubiquitous. In this particular example, one can find "bioinformatics" under "chemical biology" (one has to wonder why) but cheminformatics (or its variants) is not listed under the same heading.

As Frank K Brown wrote in 1998, "*Chemoinformatics is the mixing of information resources to transform data into information and information into knowledge for the intended purpose of making better decisions faster in the arena of drug lead identification and optimization*" [1] so, it would stand to reason that "chemoinformatics" is listed under "Chemical biology" as well.


.....Vanadia
.....Zeolite supports (catalysis)
Chemical biology
.....Antibodies
.....Bioconjugates
.....Bioenergetics
.....Bioinformatics
.....Biomembranes
.....Biomemetric system
.....Bionanoelectronics

It gets worse: Cheminformatics is **not** a keyword in the Clarivate Science Citation Index, Expanded ("bioinformatics"). Both cheminformatics and chemoinformatics are mappable in GoogleScholar. However, GoogleScholar is not directly linked to the scientific publishing industry.

Scientific publishers reading this, please take note; and those of you working in the areas of chemical documentation, library services and scientific communication – kindly request that scientific publishers remedy this omission.

Otherwise, chem(o)informatics will remain the Cinderella in the story of Chemistry.

Reference:

[1]. F.K. Brown. "Ch. 35. Chemoinformatics: What is it and How does it Impact Drug Discovery". *Annual Reports in Medicinal Chemistry*. 33, **1998** pp. 375–384

COVID-19 and the Identification of "Drug Candidates"

Contributed by RSC CICAG Chair Dr Chris Swain, email: swain@mac.com

One of the really heartening things to come out of the current pandemic is the willingness of many scientists to put aside their own research and throw themselves into the efforts to find a treatment. However, lack of domain expertise is always a problem when scientists enter a new field, so here are a few things to consider.

With *in-silico* screening, for docking experiments you need to put considerable effort into ensuring the protein structure used is appropriate; you can't simply download a PDB file from the Protein Data Bank and use it. It will undoubtedly contain errors, so you will need check protonation, hydrogen bonds, etc. Then there is the issue of deciding which solvent molecules are important. Binding energies, docking scores, etc. are not as accurate as many seem to assume, and are no substitute for an experienced medicinal chemist looking at the bound poses. I've tried to summarise the different types of molecular interactions here. Remember to also think about the impact of solvation. For other virtual screening approaches, you need to be very careful about the quality of the input data. In many cases it will be heavily biased towards actives.

In-silico predictions are no substitute for biological data, and if you are using repurposed drugs or available chemicals there is really no excuse for not generating the appropriate in vitro biological data, and there are many labs who would be happy to collaborate. If the molecules are novel, many custom synthesis companies may offer help. Remember that the IC50 is probably not that useful – it is likely that you will want to block the target 100% so you need to be above the IC95. *In vitro* biochemical assays using isolated enzymes will often give a false sense of potency, therefore you should also determine activity in a cell-based assay in the presence of plasma.

If you are proposing a repurposed drug there will be a lot of information about the drug in the public domain, and so remember to also search for compound codes and various drug name synonyms. UniChem is a very useful web service for cross-referencing between chemical structure identifiers.

There are now many free, web-accessible databases; some useful starting points are shown in the table below.

Name	Description
ChEMBL	A database of bioactive drug-like small molecules, it contains 2-D structures, calculated properties (e.g. logP, Molecular Weight, Lipinski Parameters, etc.) and abstracted bioactivities (e.g. binding constants, pharmacology and ADMET data).
PubChem	Three linked databases within the NCBI's Entrez information retrieval system. These are PubChem Substance, PubChem Compound, and PubChem BioAssay. Many compounds have links to primary literature and patents.
Guide to Pharmacology	An expert-driven guide to pharmacological targets and the substances that act on them.
DrugBank	The DrugBank database is a unique bioinformatics and cheminformatics resource that combines detailed drug data with comprehensive drug target information.
NCI Thesaurus	NCI Thesaurus (NCIt) provides reference terminology for many NCI and other systems. It covers vocabulary for clinical care, translational and basic research, and public information and administrative activities.
Clinical Trials	A database of privately and publicly funded clinical studies conducted around the world.
FDA	Food and Drug Administration responsible for safety and efficacy of

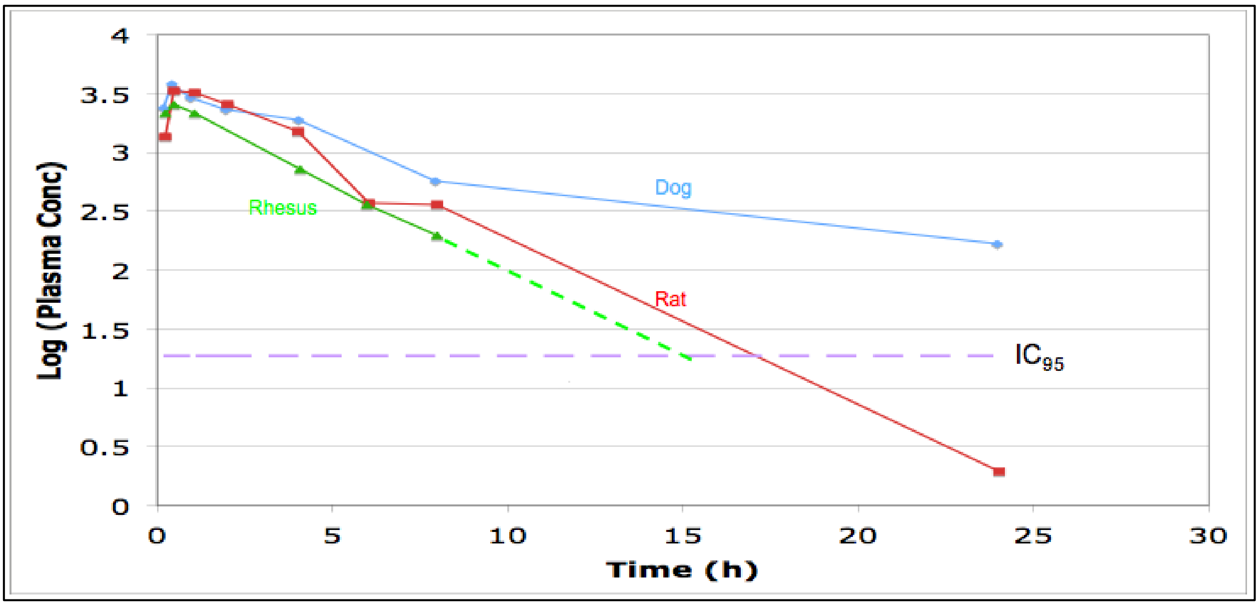
	drugs.
WIPO	The World Intellectual Property Organisation's one-stop shop for IP resources and databases.

Do find out the original target and mode of action. I've seen a couple of proposed compounds that are known prodrugs – the parent compound is designed to either break down or be modified *in vivo* to yield the active compound. The prodrug may have negligible systemic exposure. Covalent modifiers may look attractive but selectivity is always a concern and they may have narrow therapeutic windows.

Look at the original indication, many anticancer drugs are extremely toxic and could not be given to other patients. Similarly drugs that reduce blood pressure or other physiological changes may be problematic. You may well be able to find counter-screening data: this could highlight problematic off-target activities.

Look at the approved dosing regime: if a drug is only approved for doses of e.g. 2 ug/kg there might well be good reasons, and if your proposed drug only has mu activity in the *in vitro* assays you won't be able to generate sufficient plasma concentrations. Check what safety studies have been undertaken; are they sufficient to support multi-day dosing?

Look at the pharmacokinetics – you should be able to model the dosing regimen needed to maintain plasma concentrations above IC95, and this may well need to be maintained 24 hours a day. Check protein binding and distribution and use the information in the predictive modelling.



Look for the routes of administration. For in intensive care I suspect many will need the drug to be administered intravenously. If there is no i.v. formulation, is the drug soluble enough for one to be developed; and bear in mind the limitations of intravenous formulations.

Many of the patients will be on multiple drugs, both to treat the viral infection but also adventitious bacterial infections, and since many are elderly and have pre-existing medical conditions, they may have a cocktail of drugs prescribed. Drug-Drug interactions thus become a major concern, and any proposed drug to treat the virus that has major interactions with CYP450 enzymes (induction, inhibition or metabolism) is likely to hugely complicate the overall dosing-regime.

Check for any toxicity information, particularly black box warnings. HERG inhibition and QT prolongation is an issue that most drug discovery projects have to address at some point. This is particularly worrying if coupled with potential drug-drug interaction described above. You should also be able to find the data from safety studies; these may describe the dose limiting toxicities.

All of this information should be in the public domain, and if you are proposing a compound as a "Drug Candidate" you should not be expecting someone else to pull it all together to decide whether it is worth pursuing clinically.

DP4-AI: a Major Step Forward in High-Throughput Automatic Assignment of NMR Spectra

Contribution from CICAG Committee member Professor Jonathan Goodman, email: jmg11@cam.ac.uk

How can we get the most out of our NMR data? NMR is expensive to collect and can be difficult to analyse. NMR spectra can be calculated fairly readily using DFT (density functional theory) methods. The calculations are quite accurate, but not perfectly so. We are often interested in molecules with similar structures, such as diastereoisomers or regio-isomers. Such groups of compounds usually have rather similar NMR spectra, reflecting the similarity in structure. Calculating these small differences, with sufficient accuracy to be useful, and interpreting the results with confidence has been a major challenge.

DP4 is an algorithm which combines experimental and computational data to generate an assignment with a quantified level of confidence. DP4 usually assigns close isomers with a high level of confidence, and is right to do so. From time to time, it will conclude that the differences in spectral data are too small to provide a confident assignment, and this is also a useful conclusion. When a reaction might give a mixture of products, a chemist usually has a clear sense of the outcome that they want. If the spectrum more or less fits the product you expect or the product that you want, how confident can you be? Are you confident enough to go on to the next step of the process, or do you need to devote more resources to increase the certainty of the assignment? DP4 provides quantitative information to help make this decision, and so makes it possible to use resources more effectively.

It is now ten years since the original DP4 paper[1]. Since then, many studies have been published basing their conclusions on DP4 analyses. This has been useful for the structural assignment of large natural products as well as small molecules. For example, the stereochemistry of the C16-C28 subunit of a complex polypropionate, hemicalide, was assigned with the help of DP4[2]. Improvements in DFT methods and conformation analysis have made the results even more reliable. Kris Ermanis has made major contributions to making the process more reliable and also easier to use. PyDP4 provides an integrated workflow for NMR analysis, making the whole process more accessible. We are very grateful for sponsorship from *Medivir*, the *Leverhulme Trust* and the *Newton Trust*, which has made this possible. Most recently, Kris, working with Alex Howarth, a PhD student in the group, have achieved a major step forward. A bottleneck in the DP4 process has been the transfer of NMR data from the spectrometer to the DP4 algorithm. A new algorithm, DP4-AI, can take free induction decays directly from an NMR machine, automatically assign all of the peaks in the spectrum and produce an analysis without user intervention[4]. This is a major step forward in convenience, and it also makes high-throughput experimentation possible. A multi-parallel experiment may generate hundreds of NMR spectra. Analysing all of these by hand is a very big job, which may well be the rate-limiting step of the process. DP4-AI provides a means to automate this.

The code for DP4-AI is freely available on GitHub: https://github.com/KristapsE/DP4-AI. As well as the code and documentation, this site also has information on the issues people have encountered using DP4-AI and how they have been solved. What are the next steps? Can we expand DP4 to more NMR experiments and to more analytical methods? This could play an important role in the increasingly automated future of organic synthesis.

References:

[1] Assigning Stereochemistry to Single Diastereoisomers by GIAO NMR Calculation: The DP4 Probability
S. G. Smith and J. M. Goodman J. Am. Chem. Soc. 2010, 132, 12946-12959.
DOI: 10.1021/ja105035r

[2] Toward the stereochemical assignment and synthesis of hemicalide: DP4f GIAO-NMR analysis and synthesis of a reassigned C16–C28 subunit
C. I. MacGregor, B. Y. Han, J. M. Goodman and I. Paterson
Chem. Commun. 2016, 52, 4632-4635.
DOI: 10.1039/c6cc01074a

[3] Expanding DP4: Application to drug compounds and automation
K. Ermanis, K. E. B. Parkes, T. Agback and J. M. Goodman
Org. Biomol. Chem. 2016, 14, 3943-3949.
DOI: 10.1039/C6OB00015K

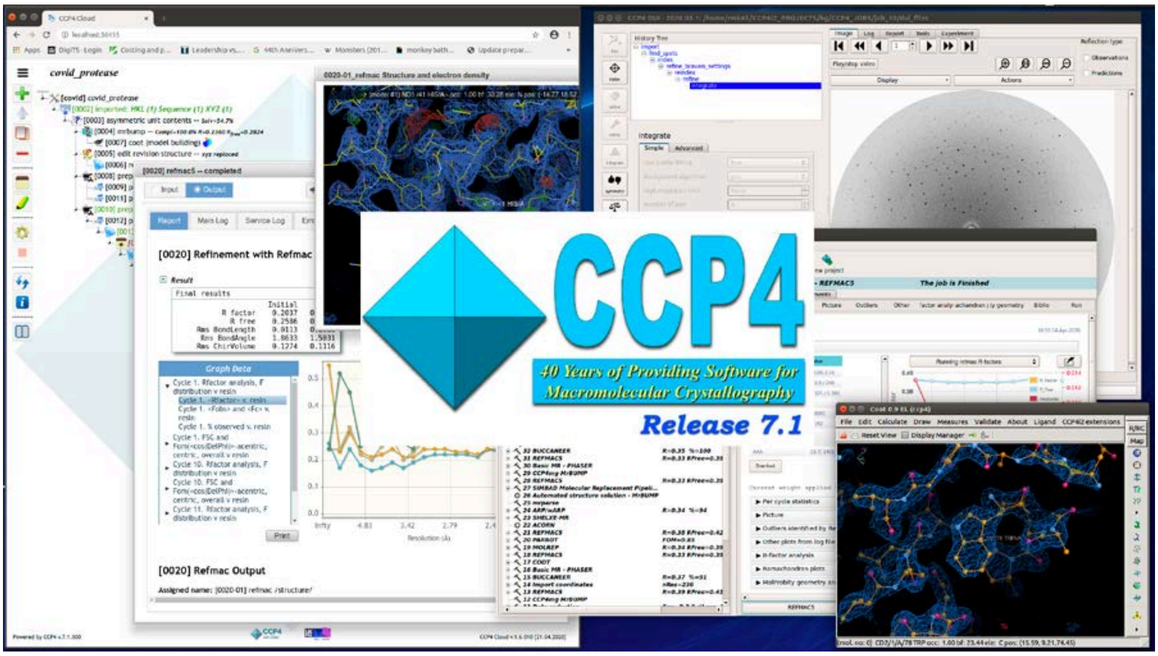
[4] DP4-AI Automated NMR Data Analysis: Straight from Spectrometer to Structure
A. Howarth, K. Ermanis and J. M. Goodman
Chemical Science 2020, 11, 4351-4359.
DOI: 10.1039/D0SC00442A

- Jonathan M Goodman PhD FRSC CChem, is Professor of Chemistry at Cambridge University Department of Chemistry, Lensfield Road, Cambridge

---

The Collaborative Computational Project Number 4 (CCP4) Release 7.1

Contribution from Eugene Krissinel, Charles Ballard, Andrey Lebedev, Ronan Keegan, Ville Uski, Oleg Kovalevskyi, David Waterman, Kyle Stevenson and Tarik Drevon, from the Research Complex at Harwell, Scientific Computing Department, Science and Technology Facilities Council, Harwell, OX11 0FA United Kingdom, email eugene.krissinel@stfc.ac.uk (CCP4 web-site: https://www.ccp4.ac.uk)


CCP4
40 Years of Providing Software for Macromolecular Crystallography
Release 7.1
covid_protease
[0020] Refinement with Refmac
[0020] Refmac Output
Graph Data
The job is Finished
REFMAC5
Integrate

Background

Collaborative Computational Project Number 4 (CCP4) was established in 1979 with the aim to facilitate the development, use and distribution of software related to the determination of 3-dimensional structures of biological macromolecules by means of X-ray diffraction on macromolecular crystals, as well as by other biophysical techniques. Those were relatively young years of Protein Crystallography, when the Protein Data Bank (PDB) contained only 53 structures, as against today's 163,141 and counting, and yet, the importance of an addressed approach to the development and maintenance of associated software tools was fully recognised. This was due to many factors. In general, reconstruction of 3D structures from a set of diffraction images is a complex theoretical and mathematical problem, the solution of which requires a considerable effort and level of expertise. The case of protein crystallography is particularly challenging due to the lower quality of macromolecular crystals, the method often being used at the edge of what is possible. Therefore, the combined and coordinated effort of many scientists was required to make advances in the field.

Over the course of its more than 40-year history, the CCP4 Software Suite underwent many transformations. Having started as a mere collection of useful utilities, it has grown into a comprehensive, world-leading package, delivering a complete software solution for macromolecular crystallography (MX). Today, the Suite contains world-leading tools for all stages of structure solution: image processing, data scaling and reduction, phasing, density modification, model building and refinement, validation and deposition, totalling some 500 executable modules and scripts exposed to end users. In order to facilitate ease of access to such a variety of sophisticated tools, the Suite comes equipped with graphical interfaces, which streamline many operations and organise the user's work in logically developed projects, important for analysis and communication.

It would not be an exaggeration to state that, today, CCP4 underpins the world's research in structural biology in both academia and industry. The total number of CCP4 users across the world is conservatively estimated at 20-25 thousand, based on some 24,000 unique software downloads and setups counted in 2019.

Traditionally, MX software was developed and used on UNIX platforms, the typical workstations used for computations in the 1990s. Since then, the impressive progress in computing hardware has made MX computations feasible with commodity PCs as well, and CCP4 software was ported to MS Windows systems in the early 2000s. Today, the CCP4 Software Suite is distributed for all MS Windows, Linux and Mac OSX platforms, with an approximately equal number of end users for each of them.

CCP4 Releases and Updates

Since the 1990s, the CCP4 Software Suite has been distributed to users via regular releases, usually once every 1-2 years. This frequency was barely sufficient given the highly dynamic nature of the project. CCP4 setups were routinely updated between releases by manually replacing files, or by periodic recompilation from source codes by more experienced users.

By 2010, the CCP4 distribution had reached a few hundred megabytes in size with some 30,000 files in the package, close to the scale of operating system distributions (the latest CCP4 series 7 requires 6.5 GB disk space and contains well over 100,000 files). Simultaneously, the Suite became considerably more integrated with strong interdependencies between program components and various libraries. This made manual building and updating unsuitable, and CCP4 invested a considerable effort in the modernisation of the Suite's infrastructure. Newly created building and packaging automated pipelines made it possible to provide dynamic updates between the releases, conceptually similar to updates in operating systems such as MS Windows, Mac OSX and Linuces. As a result, CCP4 is now able to provide the MX community with bi-weekly updates, containing the most recent program features, improvements and bug fixes, within 2-4 weeks of their release by CCP4 contributors and developers.

CCP4 Release 7.1

On the 23 April 2020, CCP4 announced Release 7.1, the second in CCP4 Series 7. The previous Release 7.0 was probably the longest running in CCP4's history, being first released in January 2016. The efficient update mechanism was sufficient for delivering all required improvements during more than 4 years (78 updates issued in total), until essential infrastructural changes became necessary in order to accommodate software components based on modern software techniques.

In the most part, CCP4 Release 7.1 is a smooth continuation of the 7.0 line starting from update 78. All of the main advances from update 78 are summarised in the release announcement. Some components in the Release are new:

GEMMI

GEMMI is a new library for handling coordinate and reflection data (by Marcin Wojdyr, https://github.com/project-gemmi). The library is developed as a joint project by CCP4 and Global Phasing Ltd., aimed at the essential modernisation of related functionality in CCP4. It replaces a few similar libraries developed 20 and more years ago and provides a modern API in the C++ and Python languages.

MRparse

MRparse is an assistant application for making and analysing search models for molecular replacement (by Jens Thomas from Prof. Daniel Rigden group in the University of Liverpool, https://github.com/rigdenlab/MrParse). This useful program will search for suitable models in the Protein Data Bank and provides the Expected Log Likelihood Gains scores of their suitability using the Phaser software. The results are represented in graphical way, convenient for making decisions on model suitability and use for molecular replacement.

CCP4build

CCP4build is a new protein model building pipeline, which combines several CCP4 applications: CParrot, CBuccaneer, Refmac, Coot and EDStats, for the enhancement of automatic model building. CCP4build explores several model building scenarios in an iterative approach with rollbacks, and adjusts automatically to building in both experimental phases and phases obtained in the course of molecular replacement.

CCP4 Cloud

CCP4 Cloud is a framework for distributed MX computations. CCP4 Cloud represents a conceptually new approach to organising and maintaining crystallographic projects in CCP4 and running CCP4 tasks. CCP4 Cloud allows a user to keep their data and structure solution projects online and manage them with any modern browser running on a desktop PC, laptop, tablet/ipad or even a smartphone, from any geographic location.

Interface to BUSTER software from Global Phasing Ltd.

The interface allows a user to refine macromolecular structures with the BUSTER software from Global Phasing Ltd. The interface is available through the CCP4 Cloud - both in local setups as well as through the CCP4 Cloud service at the RCaH at Harwell (UK) (use of the interface within your local CCP4 Cloud installation requires a separate, local installation from http://www.globalphasing.com).

CCP4 Cloud: concepts and main features

CCP4 Cloud represents an endeavour to meet modern trends in computing practices and explore alternative ways to deliver and maintain MX Software for end users. There are many benefits that online computing offers in general and in the MX case specifically.

Computational resources: Although most of CCP4 programs are not particularly demanding in terms of CPU time and memory, automated structure solvers are based on the exploration of a significant field of options and structure solution scenarios and may take hours, days and sometimes weeks of CPU time. In CCP4, the first automatic pipelines for Molecular Replacement appeared in 2008. Since then, their quality and efficiency has improved dramatically. The trend toward using automated methods, both for Molecular Replacement and experimental phasing, as well as model building, albeit at the expense of computational time, is obvious nowadays and in the majority of cases, the most practical strategy. At the same time, the local management of computational resources can be costly, and sustaining such infrastructure is not efficient given the irregular usage pattern typical of MX projects. By its very nature, cloud computing is particularly suited to accessing significant computational resources, which may be located anywhere and acquired temporarily on demand.

Software management and databases: In general, the CCP4 Software Suite is not too demanding in terms of overall management. The Suite takes about 6.5 GB of disk space in its minimal configuration and is not particularly difficult to maintain thanks to the dynamic update mechanism introduced in 2012. However, the full CCP4 setup, recommended for MX centres, includes several databases, requiring some 50-100GB of disk space, and 3rd party software, making it more demanding of IT support. In addition, MX projects usually have a highly collaborative nature, and researchers often exchange intermediate results, which may have been obtained on different platforms and with different versions of the Suite. To facilitate this, the cloud setup offers zero maintenance burden and a uniform computing experience for end users.

Data management: With the continuous development and improvement of X-ray detectors, X-ray diffraction experiments are producing an ever-increasing amount of data. For example, Diamond Ltd. produces 15-20TB of data daily. Traditionally, researchers are moving collected diffraction images to in-house storage after experiments, which are not optimal in terms of the associated cost, time for moving data, data security and sustaining availability on long timescales. Major data-producing facilities, such as Diamond Ltd., have invested considerable resources in data archives to tackle these problems. The cloud setup offers the possibility to conveniently access such stored data without the need to move it to user locations.

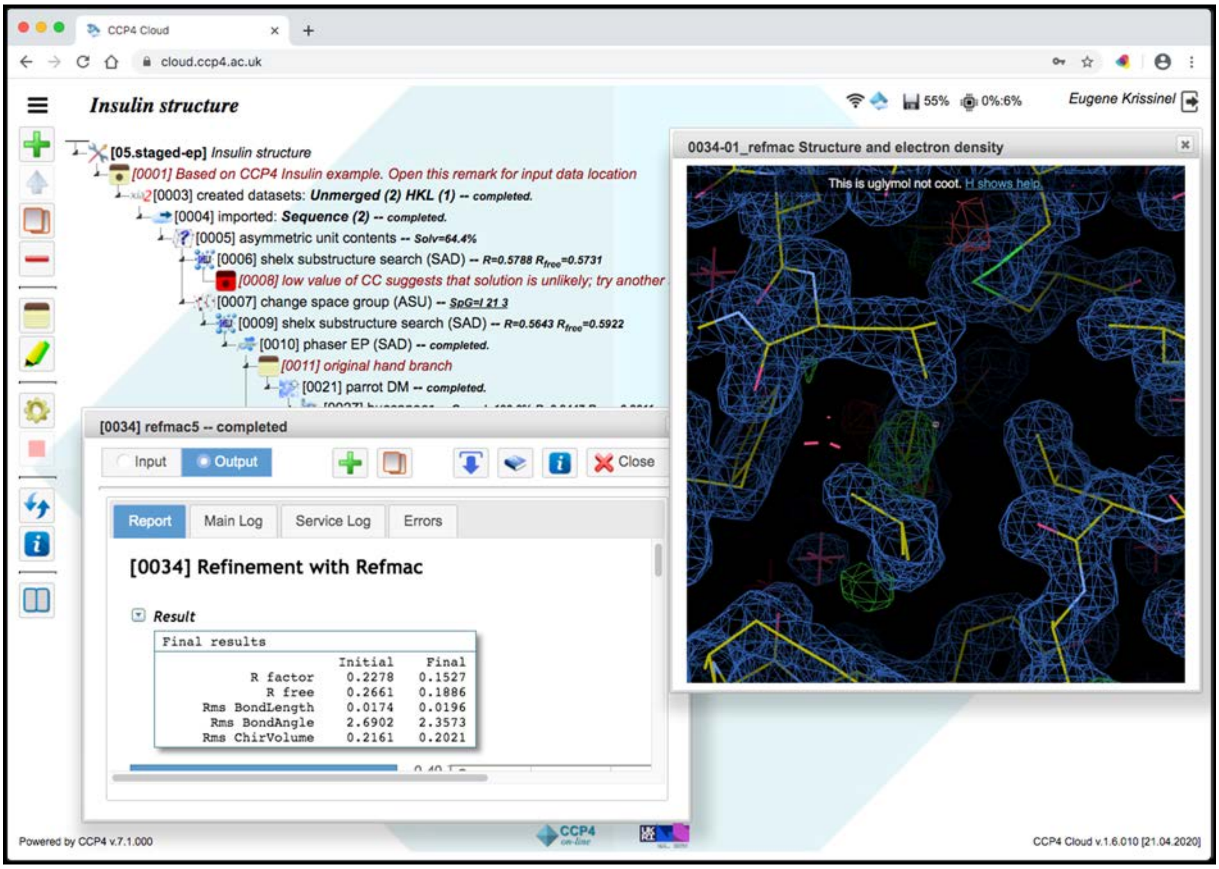
Structure solution projects: The reconstruction of 3-dimensional macromolecular structures from X-ray diffraction images is a complex branching process, often involving hundreds of computational jobs. Therefore, bookkeeping all steps in a logical, concise and reproducible manner is very important. The structure solution project by itself represents a scientific evidence. Very often, several researchers work on different aspects of the same Project, which needs to be shared and periodically moved, along with all or some data, between participating teams. The cloud setup is particularly suitable for sharing an MX Project within a group of collaborators, while keeping it in a single instance.

Diversity of computing platforms: Computing hardware and practices change constantly with time. 40 years ago, MX computing was concentrated in dedicated national-scale centres. The appearance of productive workstations in the 1990s made it possible to perform MX computations within labs. From 2000, MX computations could be reasonably conducted using budget hardware. Nowadays, ordinary PCs are being supplanted by mobile devices, especially among younger researchers. We are approaching a situation where any device – PC, laptop, tablet/iPad, even a smartphone becomes merely a terminal for accessing various resources, located elsewhere. Which, in essence, is the cloud model.

CCP4 Cloud was developed to deliver the above benefits to CCP4 users, but also to remain a community-oriented framework, which can be easily setup in any particular location. Integration of CCP4 Cloud in CCP4 Release 7.1 allows for several deployment scenarios. In the simplest case, working out-of-the-box, CCP4 Cloud runs on a user's machine without any external connection, just as an ordinary GUI. The second scenario, which also works out-of-the-box, makes use of the remote CCP4 Cloud setup based at RAL (Harwell Campus) for all computations apart from some, like interactive model building with Coot, which is performed locally on the user's device. In the third type of scenario, a local CCP4 7.1 setup is converted into a CCP4 Cloud instance on its own, using in-house resources and may be limited, if necessary, to use only by local staff and only on an in-house network. All deployment scenarios deliver an identical user experience and produce compatible MX Projects, which can be shared between CCP4 Cloud users and instances. To date, 5 separate CCP4 Cloud instances were successfully deployed at CCP4-Harwell, the Francis Crick Institute in London, the University of Newcastle, the University of Exeter and the EMBL Outstation in Hamburg (DESY facility).

CCP4 Cloud GUI

CCP4 Cloud delivers a rich graphical experience to users. All operations are done exclusively via a graphical user interface, working in all modern web browsers, including their versions for mobile devices, such as iPads, tablets and smartphones. The new GUI is conceptually different from other CCP4 Graphical Interfaces in several ways. Firstly, the cloud paradigm does not assume access to the remote file system, therefore, the new GUI operates with data objects, which may correspond to a data file, or to a part of a file, or be a combination of several files logically related to each other. Secondly, most data operations occur in the cloud and the ideal development of a structure solution project assumes data exchange with the user's device only at the outset (upload of experimental data) and at the end of the project (download of final results). Thirdly, projects are presented as branched trees of tasks, similar to the one shown in the Figure below. In most cases, a task involves several CCP4 programs, providing a high level of abstraction by eliminating many routine operations such as format conversions and preliminary data analysis.



[05.staged-ep] *Insulin structure*
[0001] Based on CCP4 Insulin example. Open this remark for input data location
[0003] created datasets: **Unmerged (2) HKL (1)** -- *completed.*
[0004] imported: **Sequence (2)** -- *completed.*
[0005] asymmetric unit contents -- *Solv=64.4%*
[0006] shelx substructure search (SAD) -- *R=0.5788 R_{free}=0.5731*
[0008] low value of CC suggests that solution is unlikely; try another
[0007] change space group (ASU) -- *SpG=I 21 3*
[0009] shelx substructure search (SAD) -- *R=0.5643 R_{free}=0.5922*
[0010] phaser EP (SAD) -- *completed.*
[0011] original hand branch
[0021] parrot DM -- *completed.*

[0034] refmac5 -- completed

[0034] Refinement with Refmac

Result

```
Final results
                   Initial    Final
         R factor   0.2278   0.1527
           R free   0.2661   0.1886
  Rms BondLength   0.0174   0.0196
   Rms BondAngle   2.6902   2.3573
   Rms ChirVolume   0.2161   0.2021
```

This is uglymol not coot. H shows help.

Representation of Projects as branching trees makes the structure solution pathway perfectly transparent and easily reconcilable. All main stages of structure solution in MX: image processing, phasing, model building, refinement, validation and PDB deposition are presented in CCP4 Cloud as a collection of more than 75 tasks.

CCP4 Cloud as a collaboration platform for the MX community

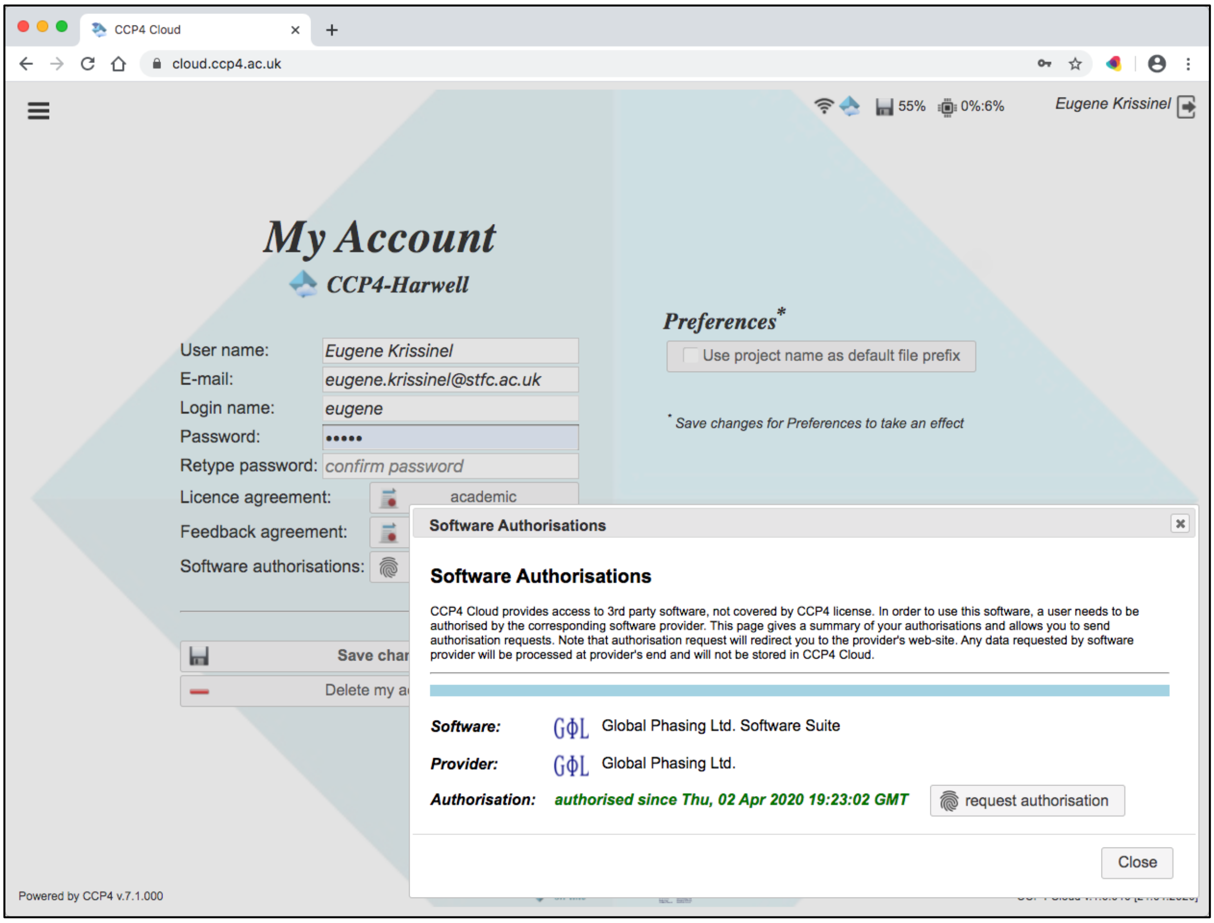
MX researchers are a good example of a sizeable, extremely geographically-diverse community with a high level of collaborative links and communication. CCP4 was always one of main communication centres in MX, owing to its role as a major software provider, organiser of CCP4 Study Weekends, International MX Schools and Workshops, and hosting the CCP4 Bulletin Board with over 7000 subscribed members. In acknowledgement of its role and achievements, CCP4 was given the *Rita and John Cornforth Award 2011* from the RSC "*for providing a resource that underpins macromolecular structural chemistry worldwide and for exemplar team-ethos over many years*". We strongly believe that the development of CCP4 Cloud is the next step in this direction, aimed at elevating the collaboration spirit in MX to the next level. There are several obvious reasons for this, and probably more will be discovered in the future.

Firstly, sharing structure solution projects in CCP4 Cloud is easy as never before, and requires nothing more than a mouse click. This may seem to be a mere technical convenience, however, combined with the central repository for projects and data, this brings about a powerful educational potential. There are about 160,000 structures in the PDB today, and probably a million structure solution projects, successful and not, were performed to solve them. Some of these projects are known as exemplars of sophisticated techniques, but it is a question how many of them are lost or forgotten. CCP4 Cloud makes it easier to retain knowledge and expertise in the MX community by providing a facility for keeping not just the final results, but also the methodologies, tricks and practical patterns. From day one, CCP4 Cloud comes with a library of several introductory projects for beginners, which can be expanded into a precious resource by community effort in the form of voluntary project and knowledge contributions.

Secondly, structure solution software is a resource of global value, which always develops faster than its documentation. It is generally accepted that community-driven documentation projects, such as Wiki, are most suitable for highly dynamical projects with limited resources. CCP4 Cloud comes with a parallel documentation project, based in Gitlab, which is open to all users, and already enjoys contributions from enthusiastic individuals.

Thirdly, CCP4 Cloud represents an open server architecture, extremely suitable for the contribution of 3rd party resources. Its computational nodes are agnostic of the particular task dispatching centres and can serve any number of CCP4 Cloud instances simultaneously. This means that CCP4 Cloud can use whatever resources a community may donate, from small workstations to major clusters, in a highly dynamic manner, irrespectively of their geographic locations.

Fourthly, CCP4 Cloud is designed for the inclusion of 3rd party software, which is not included in the CCP4 Software Suite. Such inclusion requires only the setup of a CCP4 Cloud authorisation module on the software provider's web-site. An example of such a module for the BUSTER Refinement Software from Global Phasing Ltd. is shown in the Figure below. The module is included in CCP4 Release 7.1.



User name: Eugene Krissinel
E-mail: eugene.krissinel@stfc.ac.uk
Login name: eugene
Password: •••••
Retype password: confirm password
Licence agreement: academic
Feedback agreement:
Software authorisations:

* Save changes for Preferences to take an effect

Software Authorisations

CCP4 Cloud provides access to 3rd party software, not covered by CCP4 license. In order to use this software, a user needs to be authorised by the corresponding software provider. This page gives a summary of your authorisations and allows you to send authorisation requests. Note that authorisation request will redirect you to the provider's web-site. Any data requested by software provider will be processed at provider's end and will not be stored in CCP4 Cloud.

Software: GΦL Global Phasing Ltd. Software Suite
Provider: GΦL Global Phasing Ltd.
Authorisation: authorised since Thu, 02 Apr 2020 19:23:02 GMT

This facility makes it possible to provide a uniform access, from a user experience point of view, to many relevant resources on the Web, currently scattered over individual web-sites, while preserving their distinct identity, important for proper acknowledgement, and increasing their user base.

Acknowledgement

The effort of all CCP4 developers and contributors toward making and testing Release 7.1 is highly appreciated. Many users donated their effort and time for the trial use of earlier versions of CCP4 Cloud, with extremely helpful feedback, which is greatly appreciated. Maria Fando from the Institute of Protein Research, Pushchino, Russian Federation, took an effective lead on CCP4 Cloud documentation. Global Phasing Ltd. have kindly provided their BUSTER Refinement Software for the free use by academic users of CCP4 Cloud and made it available by designing and setting up the corresponding online authorisation module. Karen McIntyre provides excellent administrative support in all CCP4-related matters, which releases considerable effort for many CCP4 endeavours.

CCP4 is supported by BBSRC UK Grant BB/S006974/1 "CCP4 Advanced integrated approaches to macromolecular structure determination" and commercial users via software licensing. CCP4 Cloud development was supported by Work Package 0 of the BBSRC UK Grant BB/L007037/1 "CCP4 Grant Renewal 2014-2019: Question-driven crystallographic data collection and advanced structure solution".

Meeting Report: AI & ML in Drug Discovery Meeting

'Predicting the Activity of Drug Candidates when there is no Target'

31 January 2020, Burlington House, London

Contributed by RSC CICAG Chair Dr Chris Swain, email: swain@mac.com

This one-day meeting brought together two highly topical areas of drug discovery. Firstly, the application of machine learning/artificial intelligence (ML/AI) approaches to the discovery of new drug leads, and secondly, programs where the biological target is not clearly established - so-called phenotypic drug discovery. Whilst there have been a number of publications describing AI or machine learning approaches in drug discovery, many use historical data sets that have been carefully cleaned and validated. Unfortunately, real world experimental data from a phenotypic screen is rarely clean and tidy and presents significant challenges to model builders.

The meeting was held at the RSC headquarters at Burlington House in London and was supported by CICAG and CDD (Collaborative Drug Design). Over 70 people registered mainly from the UK, 57 people attended the meeting of which 18 were students. The ratio of academic to industry was around 50:50, and the gender balance was 70:30 male-female.

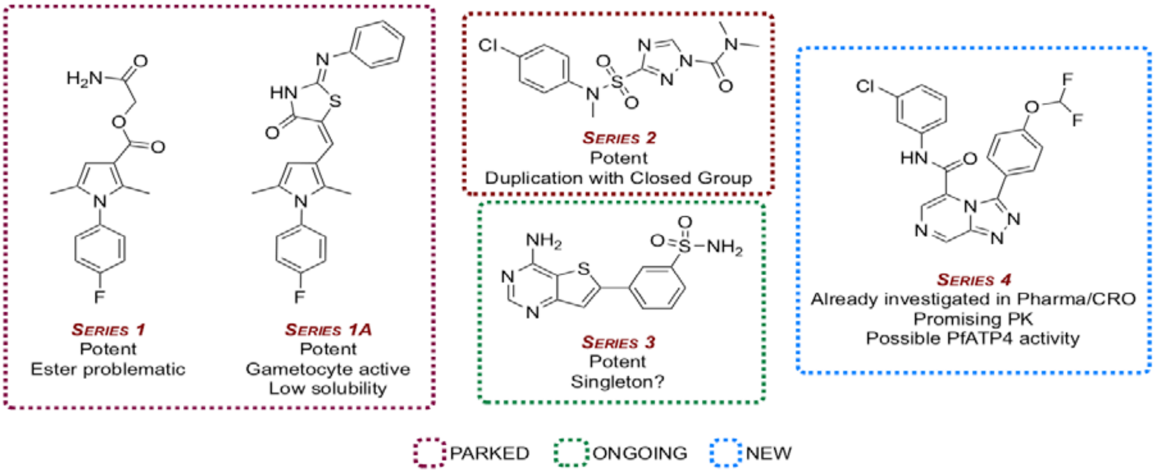
This meeting opened with a talk by Matthew Todd giving the background to the Open Source Malaria Project (OSM), and this meeting centred on a real example - a competition run by the project and funded by a grant from the EPSRC/AI3SD+ Network. Data on active and inactive compounds in one OSM antimalarial series (Series 4) were available online, and anyone was able to submit a model able to predict the actives.

The Open Source Malaria project is trying a different approach to curing malaria. Guided by open source principles, everything is open and anyone can contribute. Because everything is in the public domain there are no issues with intellectual property. All experiments are conducted using publicly accessible electronic notebooks whenever possible using open source software. To facilitate discussions in an open forum the project has made use of the issue tracker in GitHub for collaborative discussion. Whilst software developers will be familiar with GitHub the OSM team have used the GitHub wiki to provide information about the project and the tracker as a public to-do list.

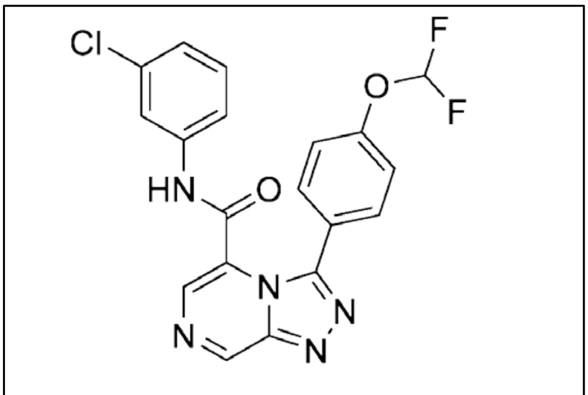
Originally all the data was reported via static html pages but this became cumbersome to maintain and impossible to search. All the data was subsequently transferred to a Google sheet. This low tech solution provides easy access either via a web browser or programmatic access via the Google sheet api (e.g. Jupyter notebook access). Whilst this arrangement has provided reliable access for many years, as more assays have been added to the worksheet it has become rather cumbersome, underlining the need for an open-source, chemically intelligent database accessible via a web interface.

General OSM project information is also communicated via social media.

The Open Source Malaria project has worked on several series:



The Triazolopyrazine (TP) Series, or Series 4, is the latest of the OSM series an example of which is shown below. This series originated from Pfizer (Sandwich) and was donated to Medicines for Malaria Venture, who put the data in the public domain via OSM.

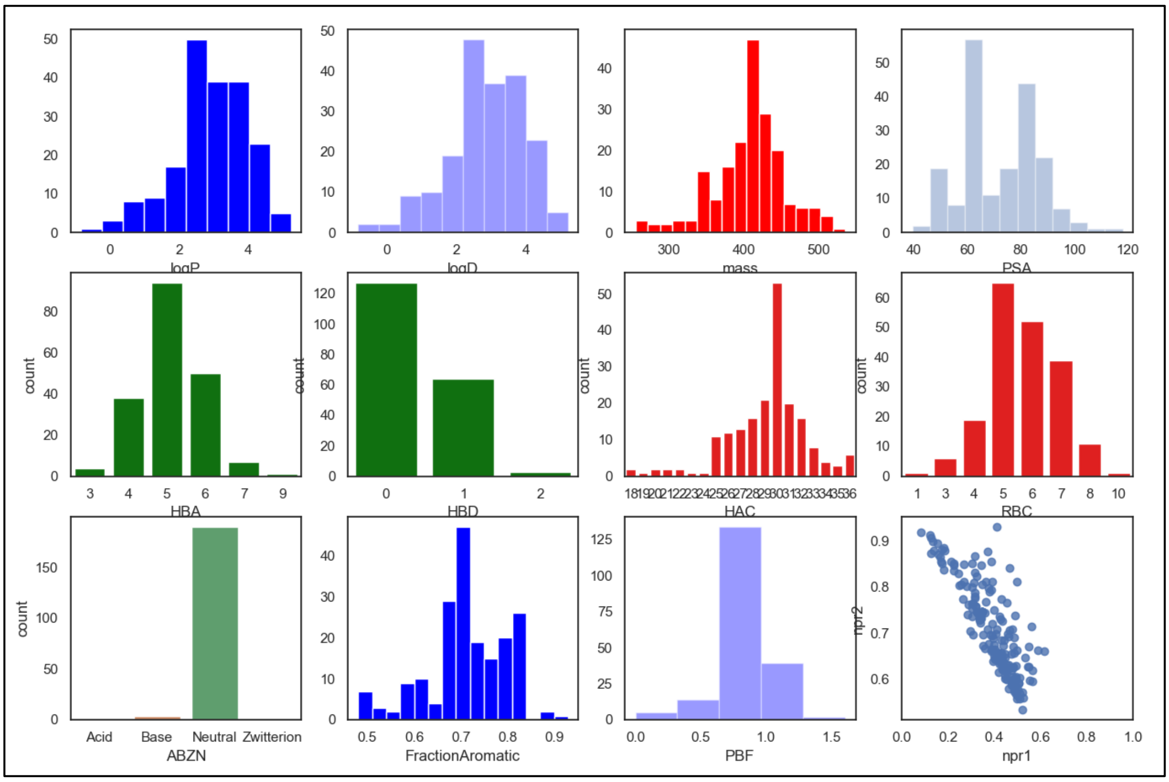


OSM Series 4
The Triazolopyrazine (TP) Series
(Representative Compound)
Already investigated in Pharma/CRO
Promising PK
Possible PfATP4 activity

The series includes many potent compounds, some with promising physicochemical properties. Most importantly, members of the series have proven to be potent in vivo, with several members having been found to be able to cure malaria in the *in vivo* mouse model of the disease.

The OSM team have made a number of related analogues and now several hundred analogues have been tested.

The physicochemical properties of the series 4 compounds are shown in the plots below calculated using a Jupyter notebook. The existing compounds span a wide range of molecular weights but most are greater than 400. However, the lack of ionisable groups at physiological pH is notable, and this coupled with calculated LogP between 2 and 5, and high aromatic atom content results in poor solubility for many compounds.

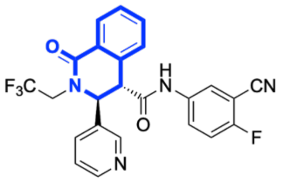


Whilst the chemistry to make the compounds is well established, the routes are lengthy and resource intensive. There is a pressing need to improve the efficiency of molecular target selection, both for increasing biological activity but also improving solubility.

The primary assay for the project is a phenotypic assay, whilst the biological molecular target has not been unambiguously identified there is mounting evidence that the series 4 compounds inhibit the Plasmodium falciparum P-type Na+-ATPase (PfATP4) transporter.

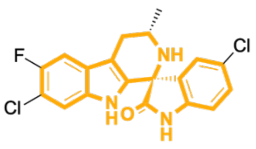
An initial attempt was made to build a pharmacophore model, assuming all the known PfATP4 inhibitors bind to a common site, however this proved to be unsuccessful. Indeed, the diversity of known ligands is quite notable (shown below) and it seems likely that there are multiple binding sites on the protein.

DIHYDROISOQUINOLONES



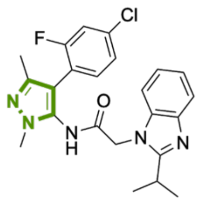
(+)-SJ733

SPIROINDOLONES



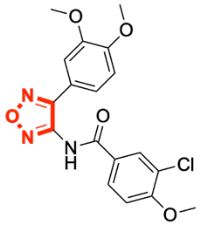
KAE609

PYRAZOLES



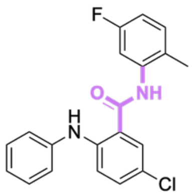
21A092

OXADIAZOLES



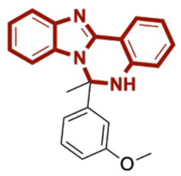
MMV011567

CARBOXAMIDES



MMV007275

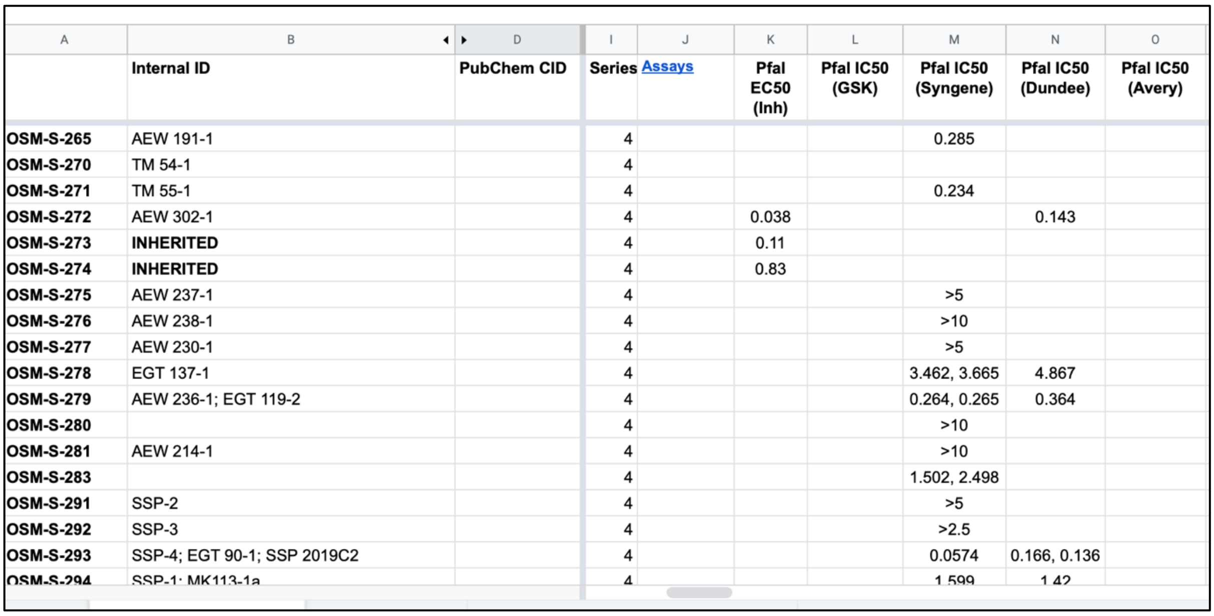
DIHYDROBENZOIMIDAZOQUINAZOLINES



MMV396719

A predictive modelling competition was run between 2016 and 2017 and elicited 6 entries, again using data from all the diverse structural classes. The predictive models were evaluated using the top twenty compounds from these rankings against undisclosed experimental data. Whilst successful in encouraging participation in an open science competition the models proved to have limited predictive utility.

With the current explosion of interest in AI/ML modelling in drug discovery, this seemed an opportunity to evaluate these new technologies using real experimental data. The current competition aims to produce a predictive model for series 4 molecules only, with the aim of improving biological activity and solubility within this series. All data for existing molecules is stored in a public spreadsheet as shown below.



A	B	D	I	J	K	L	M	N	O
	Internal ID	PubChem CID	Series	Assays	Pfal EC50 (Inh)	Pfal IC50 (GSK)	Pfal IC50 (Syngene)	Pfal IC50 (Dundee)	Pfal IC50 (Avery)
OSM-S-265	AEW 191-1		4				0.285		
OSM-S-270	TM 54-1		4						
OSM-S-271	TM 55-1		4				0.234		
OSM-S-272	AEW 302-1		4		0.038			0.143	
OSM-S-273	INHERITED		4		0.11				
OSM-S-274	INHERITED		4		0.83				
OSM-S-275	AEW 237-1		4				>5		
OSM-S-276	AEW 238-1		4				>10		
OSM-S-277	AEW 230-1		4				>5		
OSM-S-278	EGT 137-1		4				3.462, 3.665	4.867	
OSM-S-279	AEW 236-1; EGT 119-2		4				0.264, 0.265	0.364	
OSM-S-280			4				>10		
OSM-S-281	AEW 214-1		4				>10		
OSM-S-283			4				1.502, 2.498		
OSM-S-291	SSP-2		4				>5		
OSM-S-292	SSP-3		4				>2.5		
OSM-S-293	SSP-4; EGT 90-1; SSP 2019C2		4				0.0574	0.166, 0.136	
OSM-S-294	SSP-1; MK113-1a		4				1.599	1.42	

The models were judged against an OSM dataset that was temporarily kept private, and the winners were asked to use their models to predict novel molecules. The aim is that these novel molecules will be made and evaluated and the results be made public, thereby validating the models.

The AI3SD-supported competition which ran in 2019 received 10 entries using a variety of computational technologies. The competition results are shown Table 1 below.

- Table 1: Summary of the results of the predictive modelling competition

Entrant (Affiliation)	Description of Model	Precision of Accurate Predictions (Active and Inactive)	Result
Jonathan Cardoso-Silva (KCL)		36%	Runner-up
Giovanni Cincilla (Molomics)	Logistic regression classifier model using a stochastic average gradient as solver, a uniform regularisation and a learning step size = 0.01.	91%	**Winner (company)**
Mykola Galushka (Auromind)		58%	Runner-up
Davy Guan (USyd)	Automated machine learning	82%	**Winner (non-**

	method using 21 quantum mechanical descriptors calculated at the Hartree Fock level of theory and LogP optimised for Mean Absolute Error.		**company)**
Ben Irwin/Mario Öeren/Tom Whitehead (Optibrium/Intellegens)	Deep imputation [Whitehead2019] with quantum mechanical StarDrop6.6 Automodeller and pKa descriptors [Hunt2020].	81%	**Second place**
Raymond Lui (USyd)	Automated machine learning method using 59 permutation feature importance selected Mordred and quantum mechanical descriptors optimised for Mean Absolute Error	58%	Runner-up
Slade Matthews (USyd)	Random forest model using 200 Mordred descriptors based on optimised 3D structures. Training RMSE = 0.805.		Runner-up
Ho-Leung Ng (KSU)	QSAR model based on detailed homology modeling of PfATP4 and docking. 3D features are combined with 1D/2D QSAR features using XGBoost (gradient boosted trees) to make a regression model.	71%	Runner-up
Vito Spadavecchio (Interlinked TX)		36%	Runner-up
Laksh Aithani/Bill Tatsis /Willem van Hoorn (Exscientia)	Ridge regression model with alpha = 1. ECFP4 fingerprints with (Morgan radius 2) were the input to the model.	81%	**Second place**

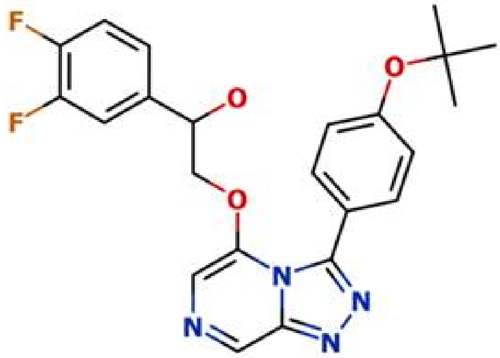
The precision of each model was calculated according to: precision =x/x+y, where x is the number of correct predictions (active and inactive combined) and y is the number of false positive predictions.

Four entrants (first and second place winners) were tasked with generating two new structures that were predicted to be active using their models, giving a total of eight molecules to be synthesised and validated experimentally

The first talk from a competition entrant was from Benedict Irwin (Optibrium) who described their collaboration with Intellegens. Unlike many machine-learning technologies, their method is designed to work with sparse bioactivity data enabling it to learn directly from correlations between activities measured in different assays. In this case rather than trying to combine the results from the same assay run in different labs into a single "average" result, they use assay to assay correlation to impute the missing data. This allows them to keep the data from each lab's assay separate, and thus calculate confidence measures for each assay. They used 330 descriptors from StarDrop.

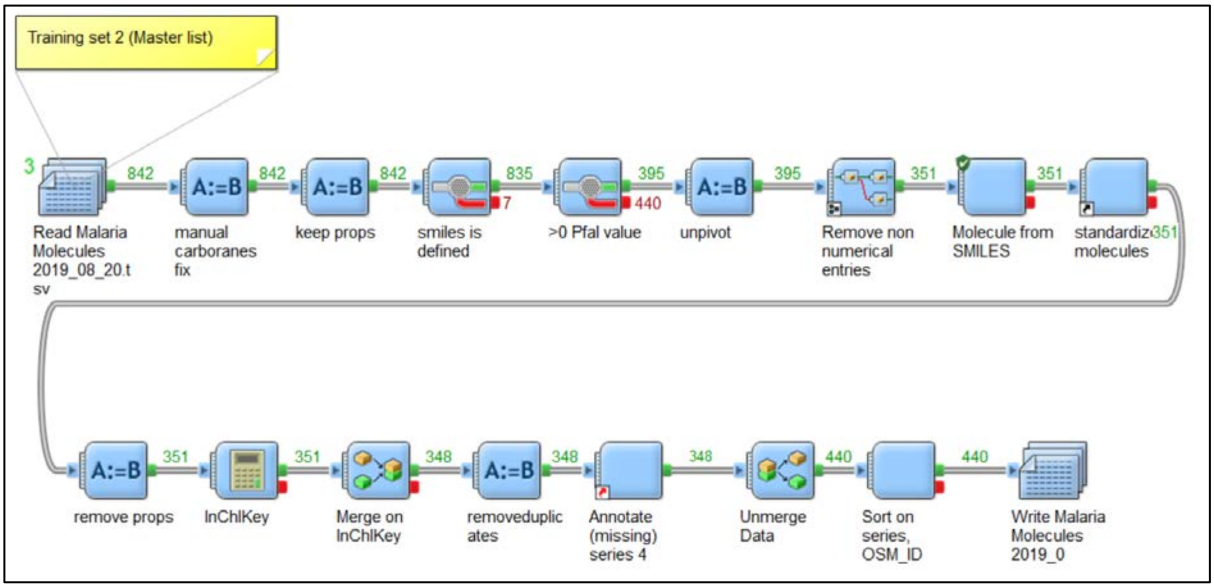
They found some assays were highly predictive, others less so Pfal(GSK) R^2 0.95, Pfal(Dundee) R^2 0.59, and they can improve the models if they remove compounds with lower confidence.

Stardrop provides several ways to generate ideas for novel molecules (MedChem transforms, Nova module, Matched molecular series, fragments), these generated lots of suggestions which were scored to be active across all assays and solubility (using Stardrop module). These were felt to be fairly conservative ideas (perhaps due to limited chemical space in the original data set). They also tried a Recurrent Neural Network (RNN) approach to generate novel structures using descriptors generated using structures from ChEMBL. This was felt to be generating "wacky" structures probably because the OSM descriptor space was not within the ChEMBL descriptor space. The primary structure that was actually synthesised is shown below. The introduction of the tert-butyl was of particular interest as this was predicted by the human OSM chemists as unlikely to be active!



They predicted pIC50: 6.4; the experimental measurement in that assay was pIC50: 6.2.

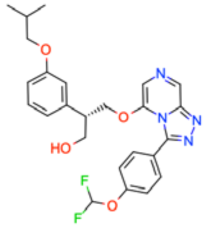
Laksh Aithani/Bill Tatsis/Willem van Hoorn (Exscientia) next presented their work. This group were new to the OSM project and the challenges they faced in trying to understand the data underline the need for domain knowledge. They gave an excellent description of the data processing/cleaning needed before model building. After downloading the spreadsheet they used Pipeline Pilot to standardise the structures from the SMILES strings, then generated the InChiKey to compare with the InChKey in the spreadsheet.



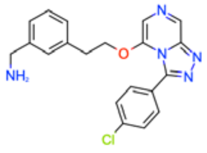
They identified several structures where the (unusual) carboranes employed in some structures were represented as one large ring system while others had the correct fully connected structure. Edwin Tse (chemist working in OSM) fixed the instances where the structure was wrongly assigned. This resulted in 440 structures, which were then annotated by hand to convert >10 to separate operator and number, and to remove some structures that did not fall within the series 4 scope. This refined data table they very generously shared with the community. They used ECFP4 fingerprints, LogP and pKa as descriptors, and they noted that the molecules are very similar. They removed features common to all molecules.

The group tried a variety of machine learning techniques but found Ridge Regression gave the best results. Ridge Regression is a technique for analysing multiple regression data that suffer from multicollinearity; a regulisation term is used to prevent coefficients becoming too large.

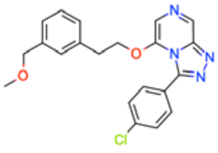
The four most potent compounds from series 4 were submitted to Medchemica MCexpert, a tool based on a matched molecular pairs analysis of over 5000 assays, to generate ideas that increase solubility. These were then scored using the predictive model. The top eight ranked compounds aim to maintain/improve potency and improve solubility. Ultimately, design 2 was chosen as a compromise between unusualness and synthetic accessibility.



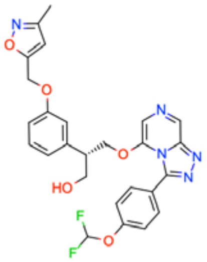
design 1



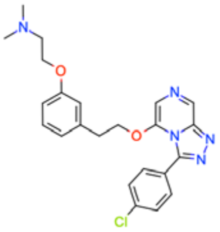
design 2



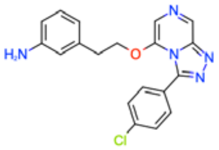
design 3



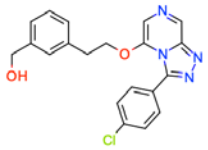
design 4



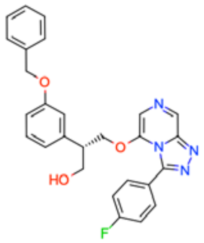
design 5



design 6



design 7

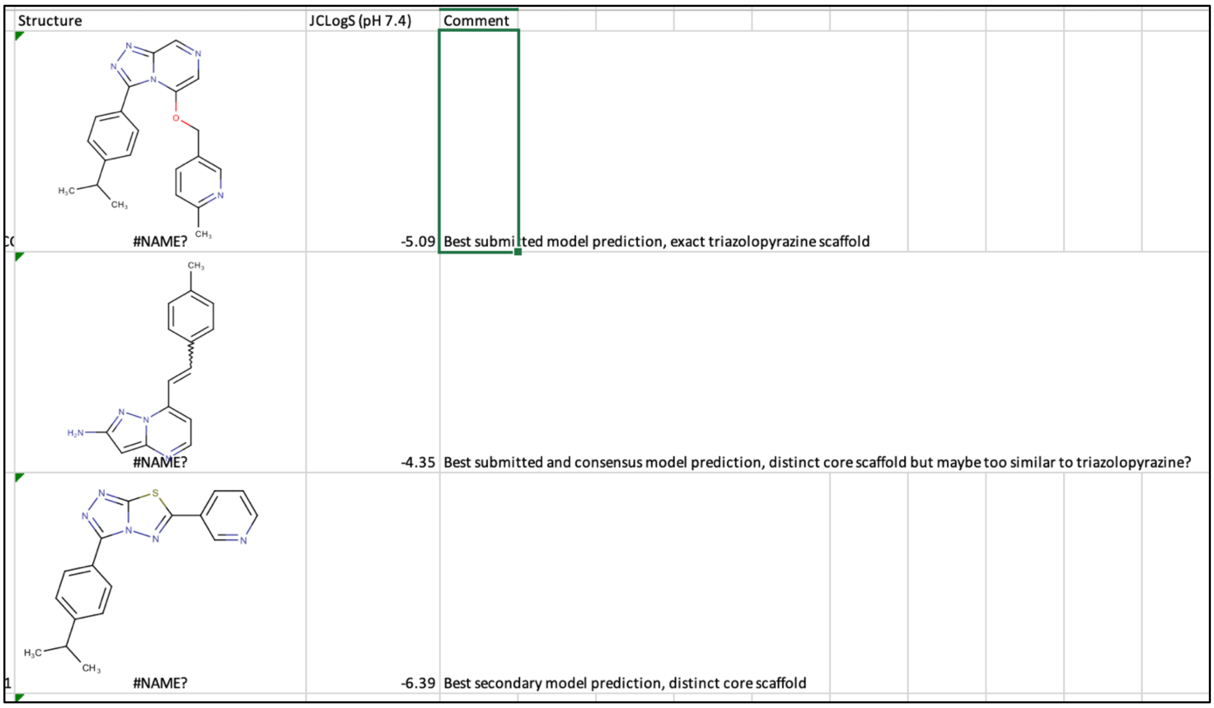


design 8

David Guan (University of Sydney) gave the next presentation remotely. David participated in the previous competition, using HTS data to try and expand the structural space explored, however this proved to be unsuccessful.

This time they used 21 QM electronic descriptors, the workflow generated minimised structures from SMILES strings using the UFF forcefield and PM7, and the QM descriptors were generated at the Hartree-Fock level of theory (Hf-3c). They also added LogP as a descriptor generated using JChem LogP. The models were built using Automated Machine Learning, optimising the models using Darwinian evolutionary theory. One audience member commented it can be difficult for medicinal chemists to interpret QM descriptors.

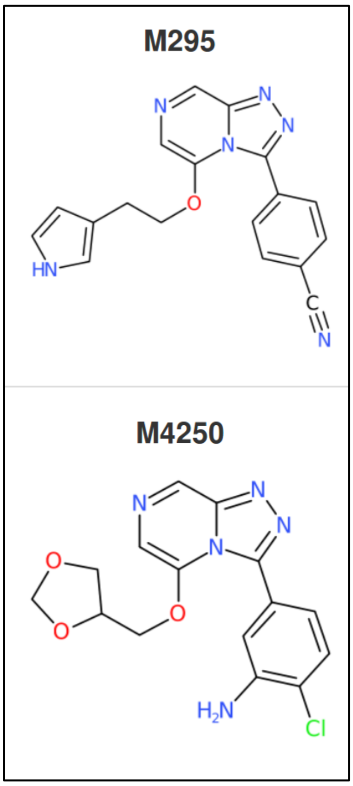
The next phase was to use generative modelling to design novel ligands predicted to be active based on the model. The series 4 SMILES were broken up into the smallest parts (tokens) and these were used to generate new SMILES strings, all valid SMILES were subjected to the predictive model. However, the molecules generated were felt to too similar to existing series 4 molecules. In an effort to generate more varied structures the SMILES from ZINC (250,000 structures), ChEMBL (40,000) were also tokenised and used to generate novel molecules.



Structure	JCLogS (pH 7.4)	Comment
#NAME?	-5.09	Best submitted model prediction, exact triazolopyrazine scaffold
#NAME?	-4.35	Best submitted and consensus model prediction, distinct core scaffold but maybe too similar to triazolopyrazine?
#NAME?	-6.39	Best secondary model prediction, distinct core scaffold

More details of all this work are on the issue tracker on the OSM site.

Giovanni Cincilla (Molomics) described their approach combining Artificial Intelligence (AI) with Human Collective Intelligence (HCI). They used the Exscientia curated dataset, and tried two strategies, regression modelling and classification modelling, with ECFP4 descriptors. They found no success with linear regression models, but did using random forest, logical regression classification. For novel molecules they used suggestions from human input, then scored them for PfaI, LogS and Caco-2 and also provided a consensus score. The final suggestions are shown below.



Anyone can access the models via the website.

The final competition entrant was Ho Leung Ng (Kansas State Univ) who described their docking model. The molecular target is thought to be Plasmodium falciparum P-type Na^+-ATPase (PfATP4) transporter. The binding site for the ligands is not known and there is no published crystal structure. However a homology model is known which was created using the full length sequence as input and was generated using the I-TASSER server.

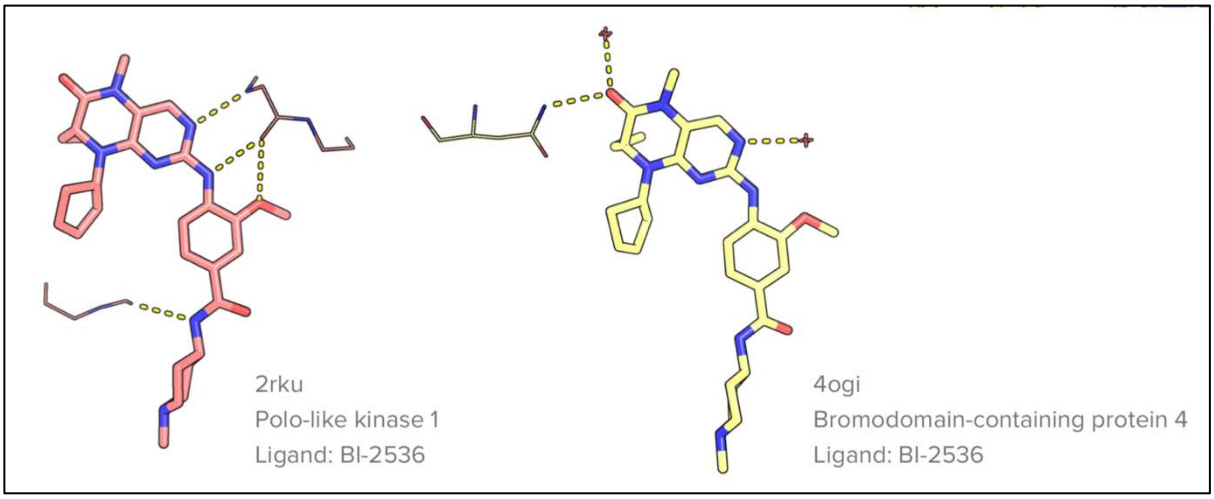
The aim of this work was to use the known SAR to try and improve the homology model by predicting binding energies. The longer-term goal is to use SAR data to gain information about binding mechanisms.

The Ng lab built their own homology model using I-TASSER and the structure refined using Yasara. The 5 most active ligands were docked using SMINA, looking for a consensus binding site. The ligands were then minimised into the consensus site. The remainder of the ligands were then docked using POSIT. Very weak ligands (>10 uM) were annotated by hand in a range 30 – 1000 uM.

They then used the docking scores together with a variety of 1-3D descriptors generated using Mordred and used gradient boosted trees XGBOOST to try and develop a model. Unfortunately, the models had modest predictive power (R^2 =0.33)

The first speaker after lunch was Joshua Meyers (BenevolentAI) who gave a talk entitled *DeeplyTough: Learning to structurally compare protein binding sites*. Comparison of binding sites has potential utility in the repurposing of known ligands, identifying potential off target interactions, and elucidating orphan protein function.

The dataset used was Tough-M1, classified into Positive binding sites known to bind similar ligands, and Negative binding sites presumed not to bind similar ligands. Pockets on the protein were identified using fpocket.



Whilst most image machine learning methods use pixels as input, the three dimensional structure of the pocket was converted to voxels for input into the CNN adapted from Deepsite and then encoded into vectors. These were compared efficiently in an alignment-free manner by computing pairwise Euclidean distances. The aim was to minimise the distance between positive binding sites and maximise the distance between negative binding sites. The resulting model was then tested on two independent data sets, Vertex (Chen et al., 2016) and ProSPECCTs (Ehrt et al., 2018).

DeeplyTough was competitive with other methods (SiteHopper and TM-Align) using the Vertex set but gave variable with the ProSPECCTs. Inspection of the false negatives identified similar ligands that bound in different conformations and since the binding sites are different these should probably not be regarded as true false negatives.

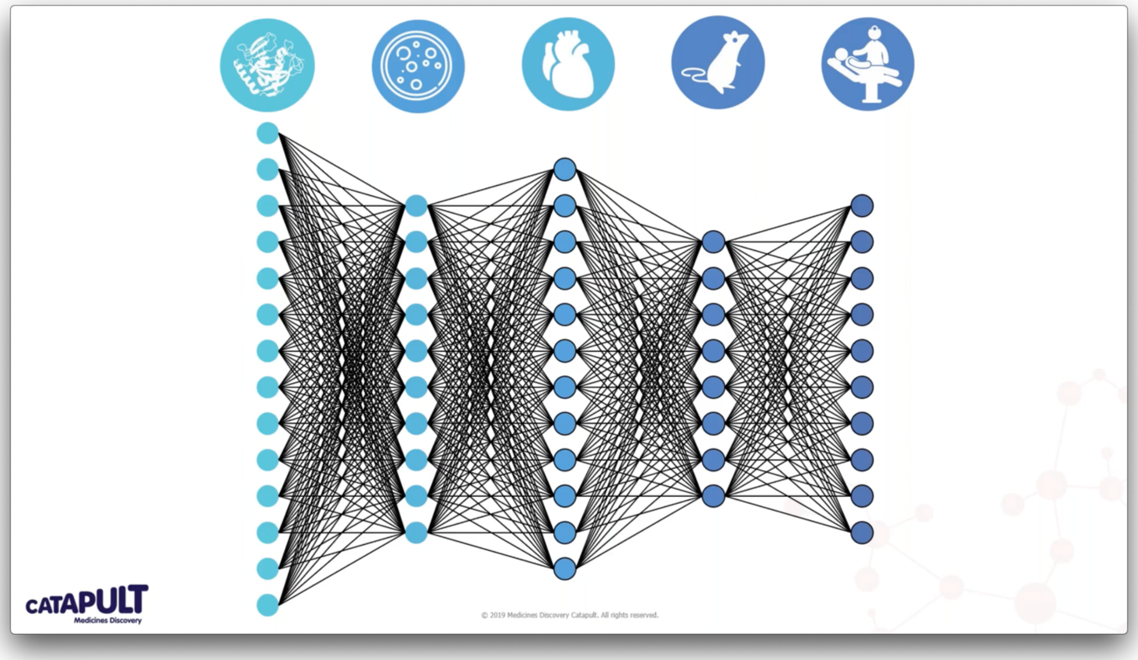
- The SiteHopper tool represents pockets as 3D patches encoded with spatial information concerning the local molecular surface (shape) and chemical properties (colour) of residues lining protein binding sites
- TM-align is an algorithm for sequence independent protein structure comparisons. For two protein structures of unknown equivalence, TM-align first generates optimized residue-to-residue alignment based on structural similarity using heuristic dynamic programming iterations

A preprint is available describing this work and is on bioRxiv, in addition all code is available on GitHub.

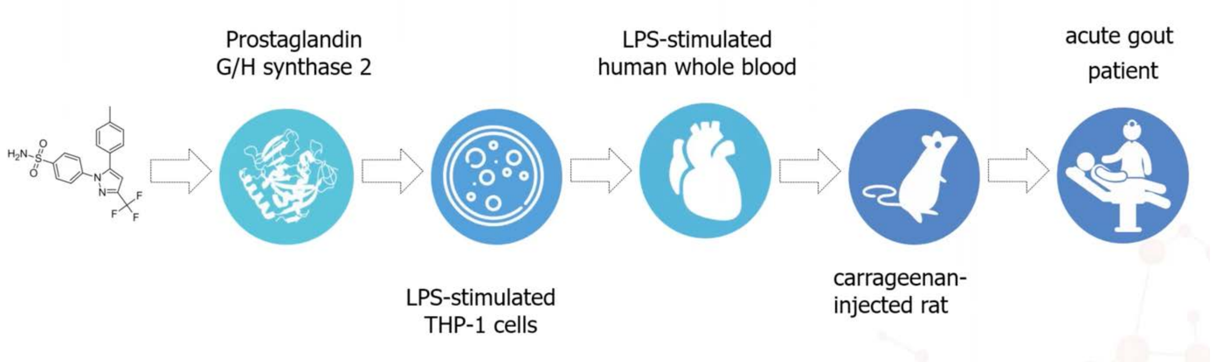
Joshua also described a use case, Privileged Structures and Polypharmacology within and between Protein Families, illustrated by the use of active site matching to identify HSF1 as a potential protein target for a known CDK9 inhibitor.

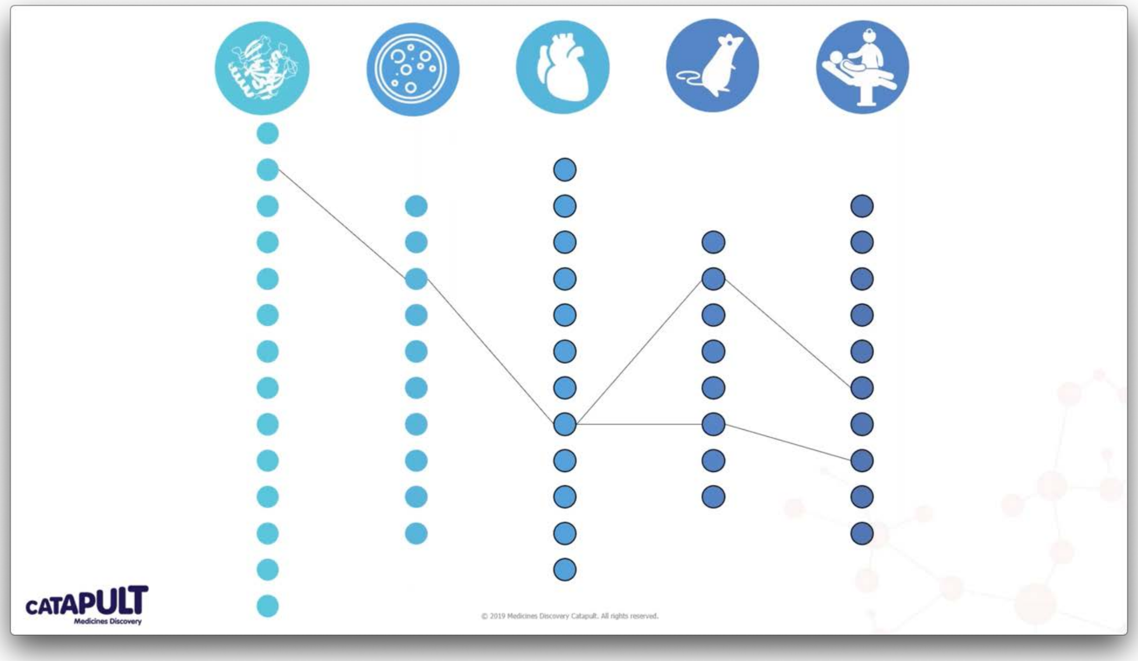
Professor John Overington @johnpoverington (Medicines Discovery Catapult) gave an interesting talk describing the *AssayNet Project: A Directed Graph of Bioassays*.

AssayNet is a directed network linking compound, protein, cell-based, tissue, animal and human data, created by text mining across papers, patents etc., building custom dictionaries and classifying experiments.



An example of a path across the network is shown below:





This network can now be explored, assays involving closely related biological targets are clustered together and can be mined to suggest novel therapeutic mechanisms that can be tested in animal models.

Explainable AI for the Medicinal Chemist - Al Dossetter (@MedChemica)

From this talk we heard that the MedChemica strategy is to provide tools to augment the chemist, enhancing data, highlighting conflicting data, placing issues in context, in a platform that is continuously updating as new data is generated. The system is based on fully automated matched molecular pairs. A statistical analysis of the transformations was then used to generate a set of medchem rules with a high probability of being beneficial. These rules have been used on multiple projects; importantly the rules can be tracked back to the literature examples used to generate the rule so the chemist can evaluate how applicable it is.

Maria-Anna Trapotsi (University of Cambridge) gave a talk on in silico target prediction – an area of increased interest. It is a potential route for the identification of the molecular targets for hits from phenotypic screens, and the prediction of potential off-target activities.

For input data they used 70 million SAR datapoints from ChEMBL and PubChem databases, including structure, target information and activity annotations. In addition, they used image data from cell morphology studies generated from treatment of 30,000 compounds. Data was then collected using the cell painting assay. The data was modelled using Macau, a Bayesian Matrix Factorization method that can also incorporate side information.

Summary

Whilst there have been a number of publications describing AI or machine learning approaches in drug discovery, many use historical data sets that have been carefully cleaned and validated. One of the major facets of this competition was that the entrants were using real world data from a public phenotypic assay. The significant efforts needed to clean up the data were nicely described by Willem van Hoorn and these efforts were acknowledged by other entrants. It was particularly interesting to see such a variety of approaches and techniques used on the same dataset. The prospective use of the models to design new molecules together with the commitment to actually make and test the molecules was also particularly important. I suspect these efforts will continue, and since everything is in the public domain, others can build on this work.

At the close of the meeting Matthew Todd summarised some themes of the day:

1. The prediction of what molecules to make next is a common and important problem in phenotypic drug discovery. Given that one of the two compounds predicted and made so far was active, it is going to be essential in the future to use AI/ML methods in research projects such as this.
2. It is hoped that this project, and this competition, will help provide a real case study of the capabilities of AI/ML, as an antidote to some of the hype in the area.
3. The language used by the two communities present – the maths/software people and the chemistry people – can often seem different. It is important that these communities communicate at meetings such as this so that we can understand each other better.
4. While OSM has benefitted greatly over the years from spontaneous inputs, this competition provided other examples of the benefits to be gained from active participation, for example the construction of a sanitised dataset by the Exscientia team, which benefitted all entrants.
5. The role of AI/ML models demonstrates the potential of such approaches towards not the replacement of scientists but the creation of allies.
6. Several people won cash prizes from this competition, but they have all generously agreed to donate their winnings to either OSM or a malaria charity.

The Details of the OSM competition can be found on GitHub.

GitHub Repository: https://github.com/OpenSourceMalaria/Series4_PredictiveModel

---

News from CAS

Contribution from Dr Anne Jones, email: ajones2@acs-i.org

No doubt you will all be hearing the phrase 'unprecedented times' and indeed, no other phrase can be used to describe the world we live in right now. We are all feeling the effects in our personal and professional lives and wondering if we will get back to the old 'normal' again. It is a worrying time for everyone but here's hoping that when we get through the other side of the tunnel we will approach our lives with a changed view of what really is important and what perhaps is not.

Like many other organisations, CAS is still working, albeit in a completely different way to what we were a matter of months ago. There is little to nothing in the way of face to face meetings with our customers, and way more phone, WebEx and video interactions. Our mission right now is to try and support our customers in the best ways possible and ensuring we can do whatever we can to help.

COVID-19

When it became apparent that the COVID situation was a global issue in need of full scientific attention and resource, the President of CAS, Manny Guzman, sent a request that the full anti-viral candidate compound dataset and well as the COVID-19 protein target thesaurus, should be made freely available to any organization that would benefit from them. These can be found at the COVID-19 section of our website or also on the landing page of SciFinder^n.

Additionally, our information science team have produced open access publications reviewing the COVID-19 landscape with the objective that it helps further the research and understanding in this area. We are of the view that we all need to do whatever we can to help with this situation and of course to help the many companies trying to find a means to control and protect ourselves against this virus.

Please visit https://www.cas.org/COVID-19 for more information.

Public Webinars

SciFinder^n continues to evolve and grow with new features and functionality being added on a regular basis. To ensure we help customers get the best from this tool, we have launched a series of public webinars addressing different aspects of searching on SciFinder^n. These webinars are recorded and made available after the scheduled date. To see what webinars are coming up soon, please visit our events page.

A Selection of new Enhancements in SciFinder^n

The following are newly incorporated since the last CICAG newsletter –

- Retrosynthesis 'Custom Scoring' - This feature allows users to modify the scoring of their retrosynthesis plans in accordance with their personal preferences
- Addition of deuterium and tritium to the selectable atoms in the CAS Draw interface
- Boolean and truncation searching
- Combine answer sets
- Generation of substance SMILES - Substances with stereochemistry or isotopes will provide isomeric SMILES and substances without stereochemistry or isotopes will provide canonical SMILES
- Improved reaction result set displays
- Group reactions by document

For more details, feel free to get in touch and we'll be happy to discuss any of these with you.

Formulus

We introduced the newest search tool in the previous edition of the newsletter, and as with SciFinder^n, we are continuing to improve the content and functionality in order to meet the needs of those searching information around formulations. We have introduced some advanced search options to allow a search to combine aspects of formulations such as ingredients, physical forms, delivery routes, targets and determine whether they are 'required' 'optional' or to be 'excluded'.

In the last two weeks, Formulus was further enhanced to allow a search to filter to formulations with a given process, experimental results, effective dose and those containing component amounts. Additional content has appeared such as the CosIng inventory (for Cosmetics), and well as the Orange Book and Green Book content with direct links to the corresponding US patents. Once more, please feel free to contact us about this and we will be happy to discuss it with you.

Collaboration with Patsnap

Back in March this year, CAS announced its collaboration with Patsnap, enabling both organisations to quickly augment existing solutions with new content and technology-driven functionality as well as leveraging our collective expertise to jointly develop new capabilities. "*The CAS and PatSnap teams share a passion for accelerating innovation across industries including healthcare, agriculture, diversified chemicals and energy to improve people's lives*," said CAS President Manuel Guzman. "*Partnerships are a critical aspect of our strategy to rapidly advance that vision. By working together, we can deliver greater value across the complete discovery and commercialization lifecycle.*"

CAS Services

CAS is offering custom services based on our content collection, subject matter expertise and specialised technologies. We have grouped these into a few different categories: *Technology, Content, Knowledge* and *Professional Services*. If you or your organisation needs any assistance in these areas, then CAS can help. Please refer to https://www.cas.org/services for further information.

News from AI3SD

Contribution from AI3SD Network+ Coordinator Dr Samantha Kanza, email: s.kanza@ai3sd.org

Our website has a wealth of information including a number of reports and articles are available detailing activities and information on what has been going on since the last newsletter. Here are some highlights:

1. AI for Allergens Blog Post and report

2. AI and ML in Drug Discovery meeting (do also see the report from Chris Swain on page 15 of the newsletter)

3. AI3SD, Dial-a-Molecule & Directed Assembly: AI for Reaction Outcome and Synthetic Route Prediction
The network did manage to hold this AI for Reaction Prediction Meeting just before lockdown, and the report if not already published, will be appearing soon.

4. AI3SD has put together a COVID-19 resources page. Please contact us if readers know of other resources that should be listed.

5. The network is also currently working, building up a list of data, learning and software resources, and now offer to host relevant datasets for our members.

6. Do you work with data but experience issues with obtaining, using or sharing data? Maybe you want to publish your data but feel overwhelmed by the complexities of getting it right? The network has launched a survey to help shape our series of events that you can get involved in to help discuss how to get data sharing right.

Other Chemical Information Related News

Contributed by RSC CICAG Member Dr Keith White and RSC CICAG Newsletter Editor Stuart Newbold

[*All hyperlinks correct & working as of 6 July 2020*]

Universities will never be the same after the Coronavirus Crisis
How virtual classrooms and dire finances could alter academia: Part 1 in a series on science after the pandemic.
https://www.nature.com/articles/d41586-020-01518-y
Source: Nature

Taylor & Francis Acquires F1000 Research
Taylor & Francis (part of Informa plc) has acquired open research publisher F1000 Research Ltd. In 2013 F1000 Research launched the world's first open research publishing platform, combining the opportunities offered by technology with a desire to identify new ways to validate and share research. F1000 Research also provides fully managed, open research publishing services directly to research funders and institutions,

including Wellcome, the Bill & Melinda Gates Foundation and the Health Research Board Ireland and to other scholarly publishers such as Emerald Publishing. Taylor & Francis will provide investment, expertise and ongoing support to enable F1000 Research to continue to develop its range of publishing services, supporting continued innovation in scholarly communication and accelerating impact across the whole research ecosystem.
https://www.infotoday.eu/Articles/News/Featured-News/Taylor-and-Francis-Acquires-F1000-Research-136024.aspx
Source: Information Today / Taylor & Francis

Springer Nature Commits to Transition Majority of Journals
The group has committed to transition the vast majority of its Springer Nature-owned English language journals that are not already open access, including Nature and the Nature Research journals, to become Transformative Journals. The approach means that Plan S-funded authors can continue to submit research to these journals, subject to acceptability of transparency requirements to be published by cOAlition S.
https://www.researchinformation.info/news/springer-nature-commits-transition-majority-journals
Source: Research Information

UK Eliminates VAT on Digital Publications
The VAT on digital publications in the UK stood at 20% and was originally scheduled to be scrapped on 1 December 2020. However, the UK government announced that would instead institute the elimination of the tax on 1 May 2020.
https://www.infotoday.eu/Articles/News/Featured-News/UK-eliminates-VAT-on-digital-publications-141049.aspx
Source: Information Today

Frontiers partners with Clarivate on new Reviewer Recognition Service
New service provides greater recognition for reviewers and their contributions to scholarly publishing.
https://blog.frontiersin.org/2020/04/28/frontiers-partners-with-clarivate-on-new-reviewer-recognition-service/
Source: Frontiers Science News

RSC and Jisc Consortium Agree Transformative new Journals Deal
The two organisations have collaborated on a transformative national deal that will provide UK institutions with access to read, and publish open access in the Royal Society of Chemistry's peer-reviewed portfolio of hybrid journals in the chemical sciences and related fields.
https://www.rsc.org/news-events/articles/2020/05-may/jisc-2020-agreement/
Source: RSC News and Events

Startup of the Week: Scimagine — a Platform for Material Scientists
Scimagine is a startup that functions as a materials-related, experimental data cloud-storage and management platform.
https://www.arabnews.com/node/1683371/saudi-arabia
Source: Arab News

Dame Ottoline Leyser appointed new CEO of UK Research and Innovation
Professor Ottoline Leyser is a leading British plant biologist and Director of the Sainsbury Laboratory at the University of Cambridge. As UKRI CEO, Professor Leyser will guide the delivery on the government's ambitions to increase investment in research and development (R&D) to 2.4% of GDP by 2027, establishing the UK as a global hub for science and technology, now and far into the future.
https://www.gov.uk/government/news/dame-ottoline-leyser-appointed-new-ceo-of-uk-research-and-innovation-ukri
Source: UKRI

Cochrane Announces New 10-Year Publishing Agreement with Wiley
Cochrane has entered into a new contract with John Wiley & Sons to publish the Cochrane Library for the next 10 years from January 2021. The agreement guarantees major investment into future development of

the Library to sustain Cochrane as the world's pre-eminent collection of high-quality evidence to inform global healthcare decision making.
https://newsroom.wiley.com/press-releases/press-release-details/2020/Cochrane-Announces-New-10-Year-Publishing-Agreement-with-Wiley/default.aspx
Source: Wiley

WIPO Launches Tool to Track IP Policy Information in Member States during COVID-19 Pandemic
https://www.wipo.int/pressroom/en/articles/2020/article_0010.html
Source: WIPO

Materials Informatics: The Time for Adoption is now Reveals IDTechEx
Materials Informatics is the key to enabling a paradigm shift in our approach to materials science R&D. Significant investment, notable adoption from key end-users, and technology leaps make it evident that its time has come. Through primary-interview based analysis, IDTechEx has introduced the most comprehensive market report on the topic: Materials Informatics 2020-2030.
https://www.prnewswire.com/news-releases/materials-informatics-the-time-for-adoption-is-now-reveals-idtechex-301070087.html
Source: PR Newswire

Publishers 'must drop fees in wake of Covid-19' – Jisc/UUK
Major academic publishers are being urged to reduce their prices by 25 per cent on all agreements in light of the severe financial impact institutions are facing because of the Covid-19 pandemic.
https://www.researchinformation.info/news/publishers-must-drop-fees-wake-covid-19-jiscuuk
Source: Research Information

Frontiers and Clarivate Partner on Peer Review Service
Frontiers is partnering Clarivate on a new Reviewer Recognition Service, in an effort to to give peer reviewers greater recognition for their work. The service is made possible by a feature from Publons, acquired by Clarivate in 2017, and now an integral part of the Web of Science. The Web of Science Reviewer Recognition Service allows reviewers to easily track, verify and record all their review and editorial contributions. By integrating the service into Frontiers' Collaborative Review Platform, reviewers can also upload contributions automatically to their Publons profile.
https://www.researchinformation.info/news/frontiers-and-clarivate-partner-peer-review-service
Source: Research Information

The Institution of Engineering and Technology and Wiley Announce Open Access Publishing Partnership
Under the terms of the publishing agreement, the IET will transition its entire hybrid subscription journals portfolio to a gold OA model, joining its existing gold open access journals, to create a leading collection of engineering and technology open access journals. The IET is working with its existing stakeholders to make this transition. The move to the open access publishing model allows researchers and practitioners around the world immediate and free access to the IET Engineering and Technology hub through the Wiley Online Library.
https://newsroom.wiley.com/press-releases/press-release-details/2020/The-Institution-of-Engineering-and-Technology-and-Wiley-Announce-Open-Access-Publishing-Partnership/default.aspx
Source: Wiley

Pistoia Alliance Partners with the America Chemical Society
CAS has partnered with the Pistoia Alliance to develop and host the Pistoia Alliance Chemical Safety Library.
https://www.scientific-computing.com/news
Source: Scientific Computing World

AAAS and Wiley Collaborate to Drive the STEM Workforce Forward
John Wiley and the American Association for the Advancement of Science (AAAS) have announced their collaboration to strengthen and support the Science Careers job board. The AAAS is the non-profit publisher of the Science family of journals and Science Careers website and job board. As part of the agreement, Wiley

will now manage the AAAS Science Careers job board that supports organizations and professionals in science, technology, engineering and math (STEM) fields.
https://newsroom.wiley.com/press-releases/press-release-details/2020/AAAS-and-Wiley-Collaborate-to-Drive-the-STEM-Workforce-Forward/default.aspx
Source: Wiley

ShanghaiTech University and ACS Partner on new Journal, Accounts of Materials Research
ShanghaiTech University and the Publications Division of the American Chemical Society (ACS) have announced their first publishing collaboration, creating a new journal for the worldwide research community, Accounts of Materials Research. Both organizations held a virtual ceremony to mark their collaboration on June 1, during which the agreement was signed. The journal will begin to accept submissions in the summer of 2020.
https://www.acs.org/content/acs/en/pressroom/newsreleases/2020/june/shanghaitech-university-and-acs-partner-on-new-journal-accounts-of-materials-research.html
Source: ACS

Wiley and ResearchGate Announce Cooperation Agreement Enhancing Research Collaboration
As part of the cooperation agreement, Wiley and ResearchGate will:

- Experiment with new models of journal article discovery to better serve the companies' shared objective of advancing research communication and collaboration
- Develop and share insights about the usage of Wiley content on the platform
- Collaborate on educating users about their rights in relation to copyright-protected content by providing clear and relevant information about how and when they may share their journal articles on the network
- Work to promptly identify and address any copyright-infringing public sharing of Wiley content on the ResearchGate platform

https://newsroom.wiley.com/press-releases/press-release-details/2020/Wiley-and-ResearchGate-Announce-Cooperation-Agreement-Enhancing-Research-Collaboration/default.aspx
Source: Wiley

Coronavirus Turmoil Fuels the rise of AI-Powered Companies
https://www.reuters.com/article/us-health-coronavirus-automation/coronavirus-turmoil-fuels-the-rise-of-ai-powered-companies-idUSKBN23O1NQ
Source: Reuters

World-leading publishers join RSC in Action-Focussed Commitment to make Research Publishing more Inclusive and Diverse
Publishers responsible for tens of thousands of peer-reviewed journals and books have signed an agreement to take a proactive stance against bias, as the RSC commits to working together to better reflect the diversity of our communities and to remove barriers for under-represented groups.
https://www.rsc.org/news-events/articles/2020/jun/publishers-join-us-in-id-commitment/
Source: RSC

De Gruyter teams up with 67 Bricks to build new Digital Publishing Platform
67 Bricks, a trusted partner to some of the most respected names in academic publishing, is currently working with De Gruyter's technology team to develop a highly customized digital platform that will allow the publisher to become more user-centered and data-driven and react much more flexibly to the needs of its customers.
https://www.stm-publishing.com/de-gruyter-teams-up-with-67-bricks-to-build-new-digital-publishing-platform/
Source: STM Publishing News

Wiley Acquires Bio-Rad's Informatics Spectroscopy Software and Spectral Databases
As one of the largest providers of spectral libraries in mass spectrometry and long-standing partner of Wiley, Bio-Rad's Sadtler spectra databases in IR, Raman, UV-Vis and NMR allows Wiley to expand its offering in the spectroscopy market. With the acquisition of Bio-Rad's KnowItAll desktop spectroscopy data system (SDS), server SDS, web-server SDS and ChemWindow chemical structure drawing software, Wiley is poised to enhance its support in vital areas of research data interpretation.

https://newsroom.wiley.com/press-releases/press-release-details/2020/Wiley-Acquires-Bio-Rads-Informatics-Spectroscopy-Software-and-Spectral-Databases/default.aspx
Source: Wiley

Copyright Clearance Center Partners with Editage to Offer Research Promotion Solutions to Authors through RightsLink®
This strategic collaboration between CCC and Editage gives RightsLink publishers the immediate ability to offer high-impact research promotion solutions to authors at the time of manuscript acceptance without incurring additional overhead.
https://www.stm-publishing.com/copyright-clearance-center-partners-with-editage-to-offer-research-promotion-solutions-to-authors-through-rightslink/
Source: STM Publishing News

The Strengths and Weaknesses of Third-Party Patent Databases
There are many databases compiling data from a variety of sources and collated in a way that users can access, interpret, search and use the information.
https://www.iam-media.com/the-strengths-and-weaknesses-of-third-party-patent-databases
Source: IAM

How Combining Blockchain Technology and AI could benefit Patent Analysis
With ongoing technical advancements, IP processes are becoming swifter and easier. There have been numerous developments with regard to the integration of data from intellectual property into blockchain technology. A combination of AI and blockchain technologies would be very exciting from an IP standpoint. The main purpose of the AI would be to assist in the analysis of various applications submitted at the patent office.
https://www.effectualservices.com/how-combining-blockchain-technology-and-ai-could-benefit-patent-analysis/
Source: Effectual Services

Springer Nature Reaches new Milestone with Publication of 1000th Open Access Book
As the largest OA publisher, Springer Nature launched a dedicated OA book programme in 2012 to give authors the opportunity to publish scholarly books OA. Six years later, in 2018, the OA books programme comprised 500 OA books with over 30 million chapter downloads. Only two years later, the total OA book output has now doubled, and the number of chapters downloaded nearly tripled.
https://group.springernature.com/gp/group/media/press-releases/new-open-access-milestone-with-the-publication-of-1000th-oa-book/17990474
Source: Springer Nature

Cambridge Launches Open Research Platform
Cambridge University Press (CUP) has officially launched Cambridge Open Engage, an early and open content and collaboration platform, which is now open to direct submissions from researchers. Developed in-house and in consultation with researchers, the platform builds on the technology behind Cambridge Core, the online home for CUP's academic books and journals, to publish early and open research outputs.
https://www.researchinformation.info/news/cambridge-launches-open-research-platform
Source: Research Information

Anaqua and Clarivate Announce Strategic Partnership
Trademark search integration increases workflow efficiency for IP professionals.
https://clarivate.com/news/anaqua-and-clarivate-announce-strategic-partnership/
Source: Clarivate

Dear USPTO: Patents for Inventions by AI Must be Allowed
https://www.ipwatchdog.com/2020/05/21/dear-uspto-patents-inventions-ai-must-allowed/id=121784/
Source: IP Watchdog

2020 Journal Citation Reports now Available
Clarivate has released the 2020 update to its annual Web of Science Journal Citation Reports (JCR). The annual JCR release enables the research community to evaluate the world's high-quality academic journals

using a range of indicators, descriptive data and visualisations. The reports are used by academic publishers across the globe to evaluate the impact of their journals relative to their field and promote them to the research community.
https://www.researchinformation.info/news/2020-journal-citation-reports-now-available
Source: Research Information

Optibrium Adopts Cheminformatics Toolkits from OpenEye Scientific
Optibrium and OpenEye Scientific (OpenEye), have announced integration between their two software platforms to provide scientists with new added funcionality for small molecule drug discovery. StarDrop, Optibrium's software for small molecule design, optimisation and data analysis, is now powered by OpenEye Scientific's cheminformatics toolkits. These cheminformatics libraries will assits StarDrtop users to more effectively harness computational software suite to meet the demands posed by future drug discovery projects.
https://www.scientific-computing.com/news/optibrium-adopts-cheminformatics-toolkits-openeye-scientific
Source: Scientific Computing World

Data Conversion Laboratory Announce Major Update to Harmonizer Software Solution
Harmonizer analyzes document collections and incorporates artificial intelligence (AI) into its text analysis, using natural language processing (NLP) to identify redundant and near redundant content in the collection.
https://www.dataconversionlaboratory.com/press-release-2020-harmonizer
Source: DCL