

**Artificial Intelligence and Augmented
Intelligence for Automated Investigations
for Scientific Discovery**

OPEN SOURCE MALARIA

Looking for New Medicines



**RSC INTEREST GROUP
CHEMICAL INFORMATION
AND COMPUTER APPLICATIONS**



Predicting the activity of Drug Candidates when there is no target
31/01/2020
Burlington House, Royal Society of Chemistry

Dr Chris Swain
Cambridge MedChem Consulting

10/02/2020

Predicting the activity of Drug Candidates when there is no target

AI3SD-Event-Series:Report-16

10/02/2020

DOI: 10.5258/SOTON/P0020

Published by University of Southampton

Network: Artificial Intelligence and Augmented Intelligence for Automated Investigations for Scientific Discovery

This Network+ is EPSRC Funded under Grant No: EP/S000356/1

Principal Investigator: *Professor Jeremy Frey*

Co-Investigator: *Professor Mahesan Niranjan*

Network+ Coordinator: *Dr Samantha Kanza*

Table of Contents

| | |
|---|----|
| 1. Workshop Details | 1 |
| 2. Event Summary and Format..... | 1 |
| 3. Talks | 1 |
| 3.1. Competition Entry – Benedict Irwin..... | 6 |
| 3.2. Competition Entry – Willem Van Hoorn | 7 |
| 3.3. Competition Entry – David Guan..... | 9 |
| 3.4. Competition Entry – Giovanni Cincilla..... | 10 |
| 3.5. Competition Entry – Ho Leung Ng..... | 11 |
| 3.6. DeeplyTough: Learning to structurally compare protein binding sites – Joshua Myers..... | 12 |
| 3.7. The AssayNet Project: A Directed Graph of Bioassays – Professor John Overington | 13 |
| 3.8. Explainable AI for the Medicinal Chemist - Al Dossetter (@MedChemica)..... | 15 |
| 3.9. Multitask bioactivity prediction by comparing chemical and cell morphology information - Maria-Anna Trapotsi (University of Cambridge) | 15 |
| 4. Participants..... | 15 |
| 5. Summary..... | 15 |

1. Workshop Details

| | |
|-----------------|--|
| Title | Predicting the activity of Drug Candidates when there is no target |
| Organisers | AI3SD, OSM & RSC-CICAG |
| Dates | 31/01/2020 |
| Programme | Programme |
| No Participants | 56 |
| Location | Burlington House, Royal Society of Chemistry |
| Committee | Professor Mat Todd, Dr Edwin Tse, Profesor Jeremy Frey, Dr Samantha Kanza & Dr Chris Swain |
| Sponsors | RSC-CICAG & CCDC |

2. Event Summary and Format

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 on Jan 31st, 2020, at the RSC headquarters at Burlington House in London and was supported by CICAG (<http://www.rscicag.org>) and CDD (<https://www.collaboratedrug.com>).

3. Talks

This meeting opened with a talk by Matthew Todd giving the background to the Open Source Malaria (OSM) project, this meeting centred on a real example - a competition run by the Open Source Malaria project, (<http://opensourcemalaria.github.io/NewSite/#>), funded by a grant from the EPSRC/AI3SD+ Network (<http://www.ai3sd.org>). Data on active and inactive compounds in one OSM antimalarial series (Series 4) were available online (<https://docs.google.com/spreadsheets/d/1WWP8fE3X2BLzZ7jOm6bRWpnHZqVJBf8XgIYeb7hshXU/edit#gid=1950012249>), and anyone was able to submit a model able to predict the actives.

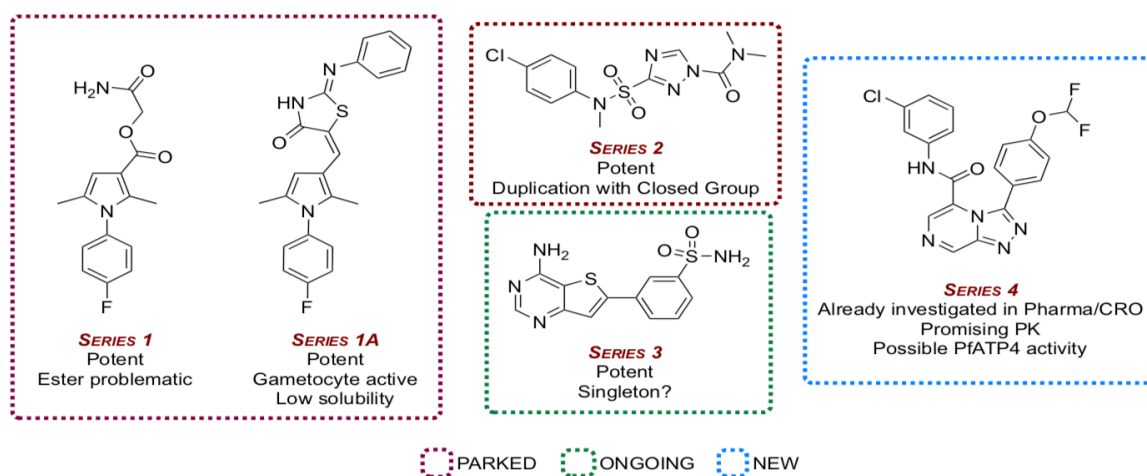
The Open Source Malaria project is trying a different approach to curing malaria. Guided by open source principles, everything is open and anyone can contribute. Because everything is in the public domain there are no issues with intellectual property. All experiments are conducted using publicly accessible 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 (<https://github.com>) 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.

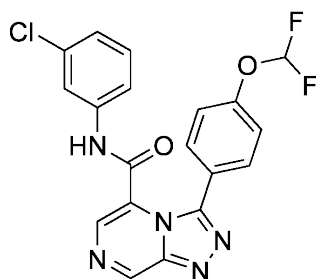
Originally all the data was reported via static html pages but this became cumbersome to maintain and impossible to search. All the data was subsequently transferred to a Google sheet (https://docs.google.com/spreadsheets/d/1Rvy6OiM291d1GN_cyT6eSw_C3ISuJ1jaR7AJa8hgGsc/edit#gid=510297618). This low tech solution provides easy access either via a web browser or programmatic access via the Google sheet api (e.g. Jupyter notebook access <https://www.macinchem.org/reviews/osm/osmipython.php>). 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 project information is also communicated via social media (https://twitter.com/O_S_M).

The Open Source Malaria project has worked on several series



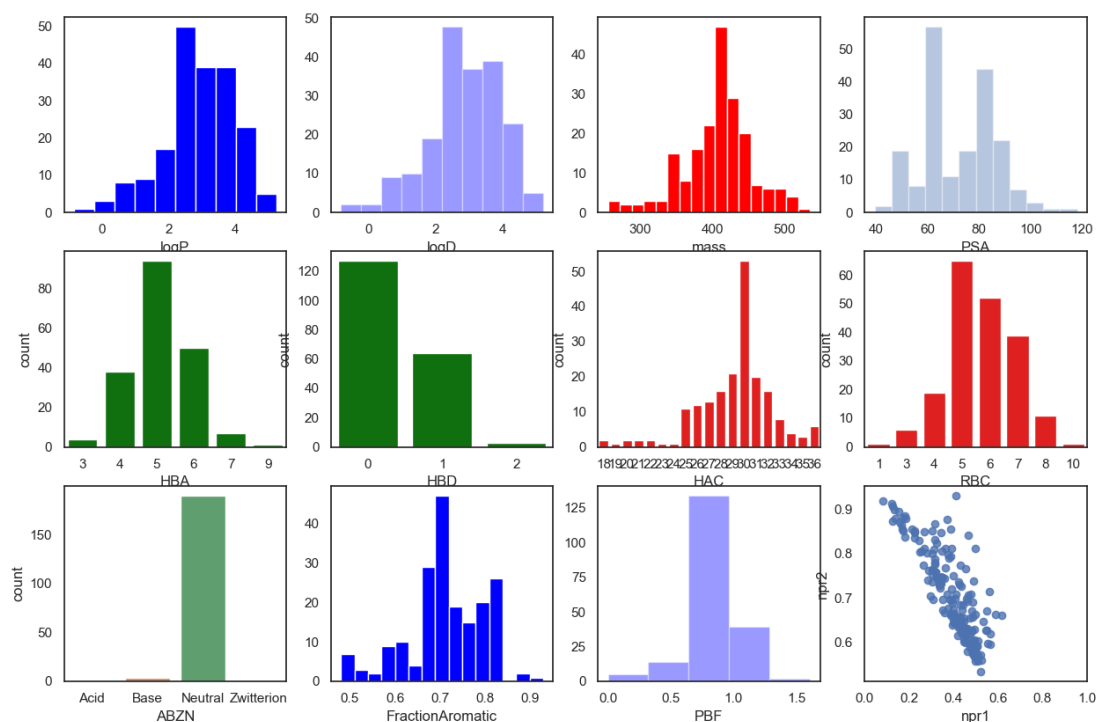
The Triazolopyrazine (TP) Series, or *Series 4*, is the latest of the OSM series an example of which is shown below. This series originated from Pfizer (Sandwich) and was donated to Medicines for Malaria Venture (<https://www.mmv.org>) who put the data in the public domain via OSM.



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 (<https://www.macinchem.org/reviews/ipython/calcpromerties2.php>). 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, this coupled with calculated LogP between 2 and 5, and high aromatic atom content results in poor solubility for many compounds.

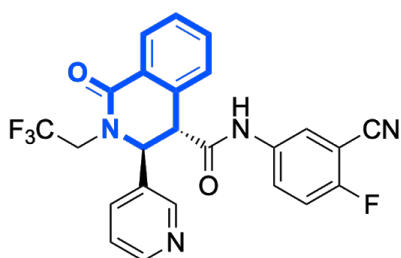


Whilst the chemistry to make the compounds is well established the routes are lengthy and resource intensive. There is a pressing need to improve the efficiency of molecular target selection, both for increasing biological activity but also improving solubility.

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.

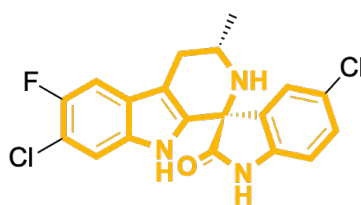
An initial attempt was made to build a pharmacophore model assuming all the known PfATP4 inhibitors bind to a common site, however this proved to be unsuccessful. Indeed, the diversity of known ligands is quite notable (shown below) and it seems likely that there are multiple binding sites on the protein.

DIHYDROISOQUINOLONES



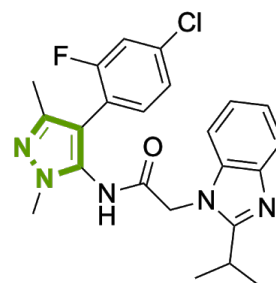
(+)-SJ733

SPIROINDOLONES



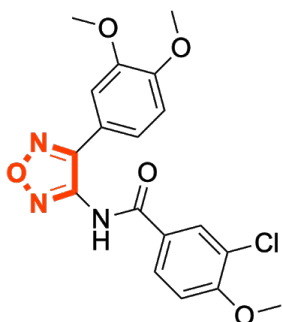
KAE609

PYRAZOLES



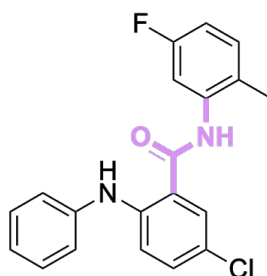
21A092

OXADIAZOLES



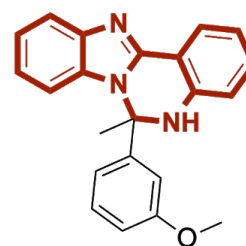
MMV011567

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 method using 21 quantum | 82% | Winner |

| | | | |
|---|--|-----|----------------------|
| | mechanical descriptors calculated at the Hartree Fock level of theory and LogP optimised for Mean Absolute Error. | | (non-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: $\text{precision} = \frac{x}{x+y}$, 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.

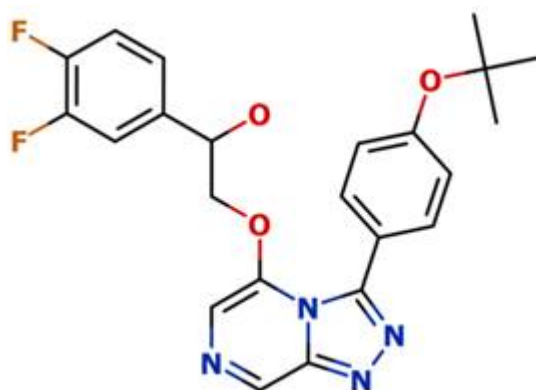
3.1. Competition Entry – Benedict Irwin

The first talk from a competition entrant was Benedict Irwin (Optibrium, <https://www.optibrium.com>) who described their collaboration with Intellegens (<https://intellegens.ai>). 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 (<https://www.optibrium.com/stardrop/>).

They found some assays were highly predictive, others less so Pfa1(GSK) R^2 0.95, Pfa1(Dundee) R^2 0.59, they can improve the models if they remove compounds with lower confidence.

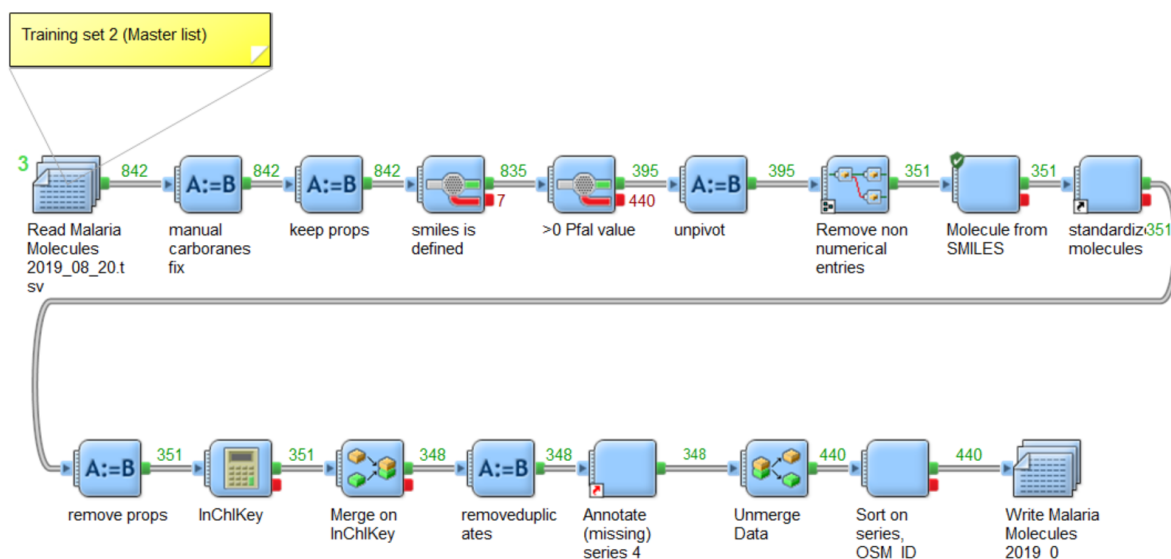
Stardrop provides several ways to generate ideas for novel molecules (MedChem transforms, Nova module, Matched molecular series, fragments), these generated lots of suggestions which were scored to be active across all assays and solubility (using Stardrop module). These were felt to be fairly conservative 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.

3.2. Competition Entry – Willem Van Hoorn

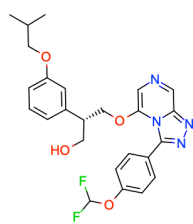
Laksh Aithani/Bill Tatsis/Willem van Hoorn (Exscientia, <https://www.exscientia.ai>) next presented their work. This group were new to the OSM project and the challenges they faced in trying to understand the data underline the need for domain knowledge. They gave an excellent description of the data processing/cleaning needed before model building. After downloading the spreadsheet they used Pipeline Pilot (<https://www.3dsbiovia.com/products/collaborative-science/biovia-pipeline-pilot/>) to standardise the structures from the SMILES strings, then generated the InChIKey to compare with the InChKey in the spreadsheet.



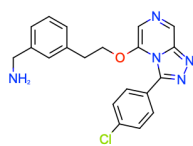
They identified several structures where the (unusual) carboranes employed in some structures were represented as one large ring system while others had the correct fully connected structure. Edwin Tse (chemist working in OSM) fixed the instances where the structure was wrongly assigned. This resulted 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, they noted that the molecules are very similar. They removed features common to all molecules.

They 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 regularisation term is used to prevent coefficients becoming too large.

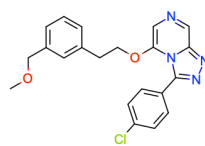
The four most potent compounds from series 4 were submitted to Medchemica MCexpert (<https://www.medchemica.com>), a tool based on a matched molecular pairs analysis of over 5000 assays, to generate ideas that increase solubility, these were then scored using the predictive model. The top eight ranked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



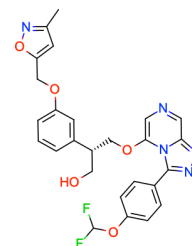
design 1



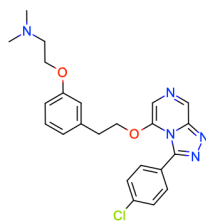
design 2



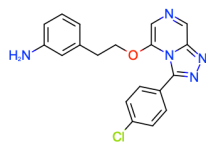
design 3



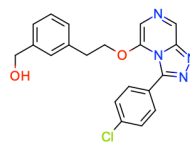
design 4



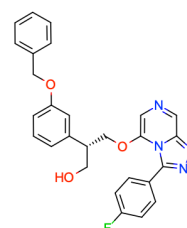
design 5



design 6



design 7



design 8

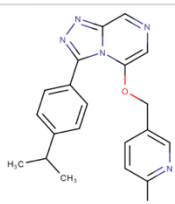
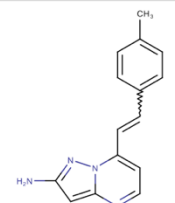
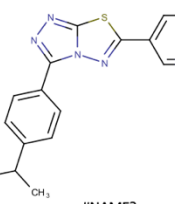
3.3. Competition Entry – David Guan

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 (<https://chemaxon.com>). 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 this work are on the issue tracker on the OSM site (https://github.com/OpenSourceMalaria/Series4_PredictiveModel/issues/19).

3.4. Competition Entry – Giovanni Cincilla

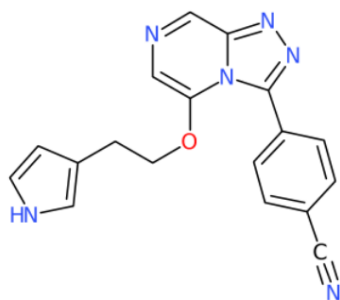
Giovanni Cincilla (Molomics, <https://molomics.com>) described their approach combining Artificial Intelligence (AI) with Human Collective Intelligence (HCI).

They used the Exscientia curated dataset, and tried two strategies, regression modelling and classification modelling, with ECFP4 descriptors.

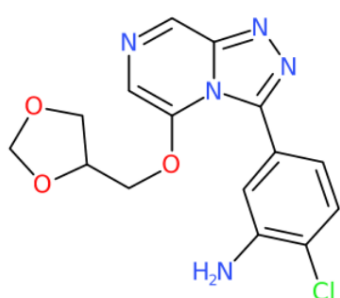
They found no success with linear regression models, but did using random forest, logical regression classification.

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.

M295



M4250



Anyone can access the models via the website (<https://osm.molomics.com/osm>).

3.5. Competition Entry – Ho Leung Ng

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 (*Pf*ATP4) transporter. The binding site for the ligands is not known and there is no published crystal structure. However a homology model (<https://doi.org/10.1016/j.ijpddr.2015.07.001>) is known this 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.

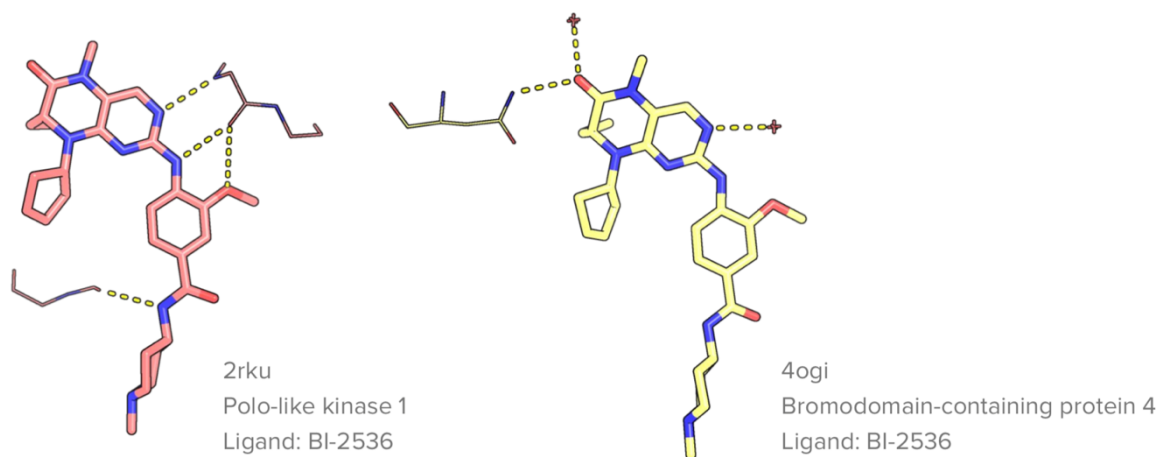
The Ng lab built their own homology model using I-TASSER (<https://zhanglab.ccmb.med.umich.edu/I-TASSER/>) and the structure refined using Yasara (<http://www.yasara.org>). The 5 most active ligands were docked using SMINA (<https://sourceforge.net/projects/smina/>) looking for a consensus binding site. The ligands were then minimised into the consensus site. The remainder of the ligands were then docked using POSIT (<https://docs.eyesopen.com/applications/oedocking/posit/posit.html>). 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 (<https://github.com/mordred-descriptor/mordred>) and used gradient boosted trees to (XGBOOST, <https://xgboost.readthedocs.io/>) try and develop a model. Unfortunately, the models had modest predictive power ($R^2 = 0.33$)

3.6. DeeplyTough: Learning to structurally compare protein binding sites – Joshua Myers

The first speaker after lunch was Joshua Meyers (BenevolentAI, <https://benevolent.ai>) 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, elucidating orphan protein function.

The dataset used was Tough-M1 (<https://www.brylinski.org/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 (<https://github.com/Discngine/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 (<https://github.com/htmiquel/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)

The SiteHopper tool represents pockets as 3D *patches* encoded with spatial information concerning the local molecular surface (shape) and chemical properties (color) 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.

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.

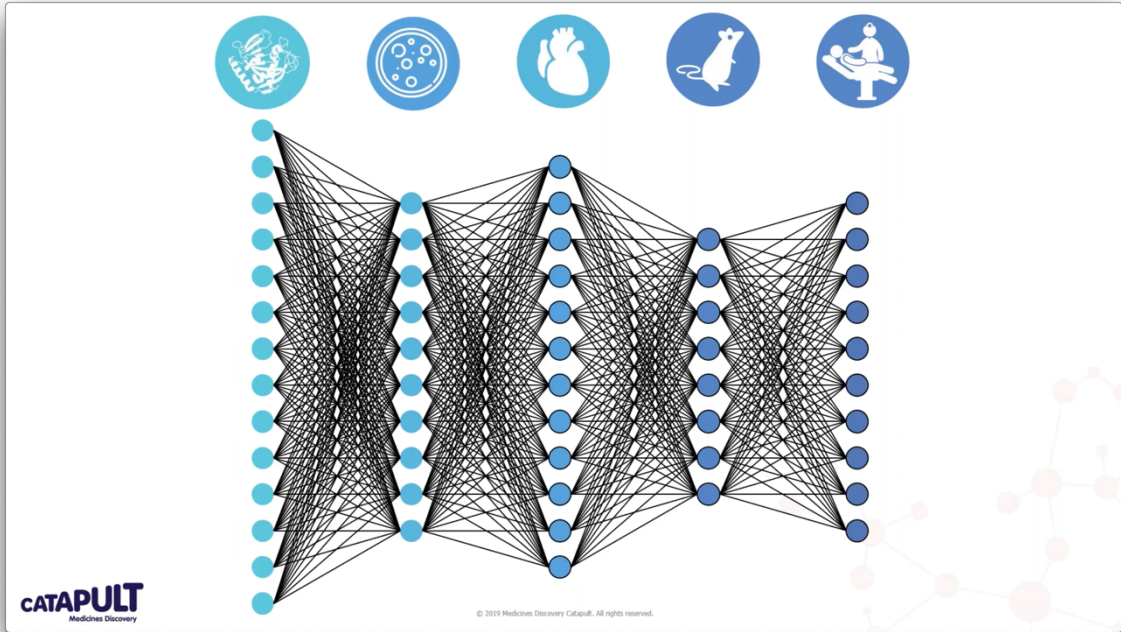
A preprint is available describing this work is on bioRxiv (<https://www.biorxiv.org/content/10.1101/600304v1>), in addition all code is available on GitHub (<https://github.com/BenevolentAI/DeeplyTough>).

Joshua also described a use case, Privileged Structures and Polypharmacology within and between Protein Families (<https://pubs.acs.org/doi/10.1021/acsmchemlett.8b00364>). Illustrated by the of active site matching to identify HSF1 as a potential protein target for a known CDK9 inhibitor.

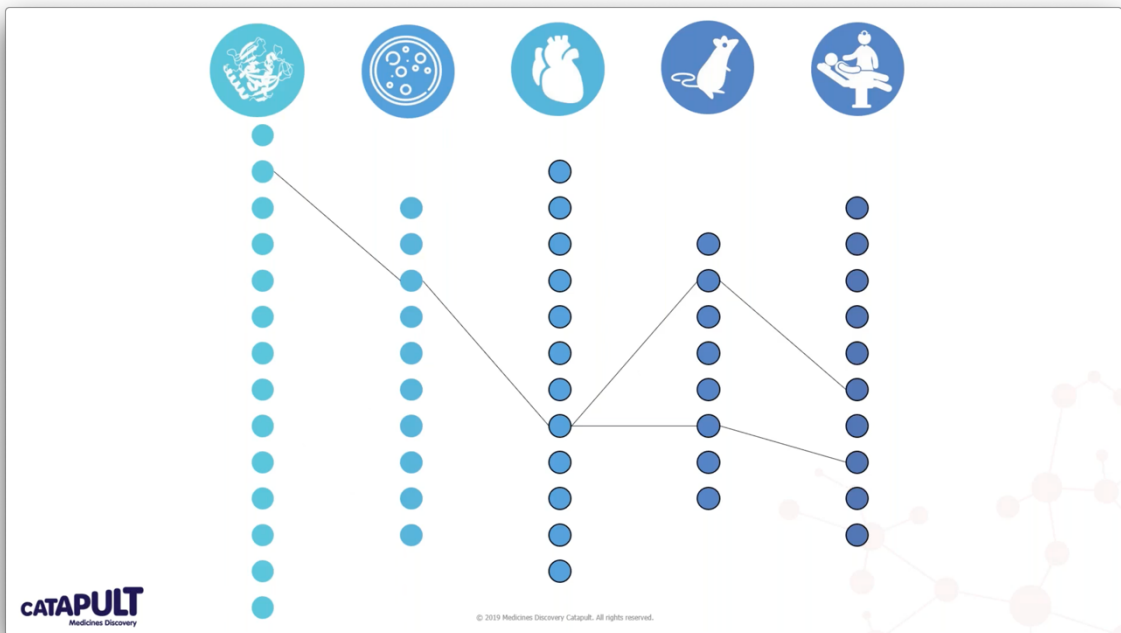
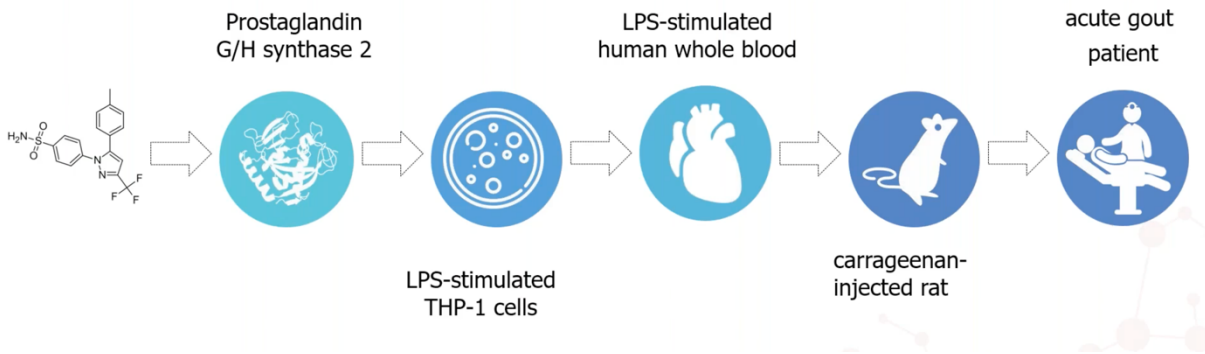
3.7. The AssayNet Project: A Directed Graph of Bioassays – Professor John Overington

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.

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

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 (<https://doi.org/10.1021/acs.jcim.7b00335>). 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 (<https://www.slideshare.net/AIDossetter>), 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.

3.9. Multitask bioactivity prediction by comparing chemical and cell morphology information - Maria-Anna Trapotsi (University of Cambridge)

Maria-Anna Trapotsi (University of Cambridge) gave a talk on in silico target prediction, this is 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 (<https://doi.org/10.1093/gigascience/giw014>).

The data was modelled using Macau, a Bayesian Matrix Factorization method that can also incorporate side information.

4. Participants

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 M:F.

5. 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