POSTER 1

Mondrian Conformal Prediction
– Confidence Predictions with Excellent Handling of Imbalanced Data

Fredrik Svensson[†⊥], Ulf Norinder[‡‖]

[†] Centre for Molecular Informatics, Department of Chemistry, University of Cambridge, Lensfield Road, Cambridge CB2 1EW, UK

[⊥] IOTA Pharmaceuticals, St Johns Innovation Centre, Cowley Road, Cambridge CB4 0WS, UK

[‡] Swetox, Karolinska Institutet, Unit of Toxicology Sciences, Forskargatan 20, SE-151 36 Södertälje, Sweden

[‖] Department of Computer and Systems Sciences, Stockholm University, Box 7003, SE-164 07 Kista, Sweden

Quantification of prediction confidence and the ability to handle very imbalanced datasets are two key features when making bioactivity predictions. Aggregated Mondrian Conformal Prediction (AMCP) is a type of confidence predictor and represents an effective alternative for modelling very imbalanced datasets. In contrast to other methods, AMPC requires no additional balancing measures and deliver predictions with an associated confidence.[1,2]

This poster describes the application of AMCP to model a variety of extremely imbalanced cytotoxicity and high-throughput screening (HTS) datasets where both the prediction confidence and the accurate modelling of the minority class are very important.[3,4]

Cytotoxicity predictions were based on PubChem data spanning 16 cell lines and comprising more than 440,000 unique compounds. The data was heavily imbalanced with only 0.8 % of the tested compounds being cytotoxic. The HTS predictions were comprised of 4 datasets where the imbalance ranged between 1:1 – 911:1. Despite the high level of imbalance, the models delivered accurate predictions with a balanced performance with respect to the two classes. On one external dataset used for the cytotoxicity measures the model had a sensitivity of 74 % and a specificity of 65 % at the 80 % confidence level among the compounds assigned to a single class. This represents superior performance compared to previous studies on datasets that are extremely imbalanced. The results from the HTS datasets showed, in a similar manner, that valid and well-balanced models could be developed, achieving a sensitivity and specificity of 79-89%.

Regardless of degree of imbalance, AMCP gracefully retrieves the minority class to a very high degree

with respect to the set significance level. Together with the controllable error rate and ease of implementation this makes AMCP a useful tool to model bioactivity in imbalanced datasets.

References

(1) Vovk, V.; Gammerman, A.; Shafer, G. Algorithmic Learning in a Random World; Springer: New York, 2005; pp 1–324.

(2) Löfström, T.; Boström, H.; Linusson, H.; Johansson, U. Bias Reduction through Conditional Conformal Prediction. *Intell. Data Anal.* **2015**, *19*, 1355–1375.

(3) Svensson, F.; Norinder, U.; Bender, A. Modelling Compound Cytotoxicity Using Conformal Prediction and PubChem HTS Data. *Toxicol. Res. (Camb).* **2017**, *6*, 73–80.

(4) Norinder, U.; Boyer, S. Binary classification of imbalanced datasets using conformal prediction. *Journal of Molecular Graphics and Modelling.* **2017**, 72, 256–265.

POSTER 2

Predicting Ligand-Protein binding affinities in Leucyl-tRNA Synthetases.

Edmund, GHC, Charlton, MH.

Oxford Drug Design, Oxford Centre for Innovation, New Road, Oxford, OX1 1BY, UK.
E-mail: grace.edmund@oxforddrugdesign.com.

Binding affinities have been predicted for a set of inhibitors of *E. coli* leucyl-tRNA synthetase using the Molecular Mechanics/Generalized Born - Volume Integral (MM/GB-VI) approach. This method uses molecular dynamic simulations to sample the conformational space of the complex combined with molecular mechanics energies and the Generalized Born implicit solvent model to predict a binding affinity.[1] A variety of simulation conditions have been investigated to generate a model for the LeuRS system. Replicate simulations have been performed to achieve better statistical accuracy, with identical simulations typically showing a standard deviation of ~3-4 $kcalmol^{-1}$. The results have been compared with experimental binding energies from isothermal titration calorimetry studies (ITC) and half maximal inhibitory concentrations (IC_{50}) from biochemical assays.

1. Labute, P. The Generalized Born/Volume Integral Implicit Solvent Model: Estimation of the Free Energy of Hydration Using London Dispersion Instead of Atomic Surface Area. *J. Comput. Chem.*, **2008,** 29 (10), 1693-1698.

POSTER 3

Active search for computer-aided drug design

Steven Oatley, University of Nottingham

Chemical space is large, to the point of precluding its explicit enumeration. Thus, it represents a so-called intensionally defined design space. Search strategies for intensionally designed spaces are a current area of interest in machine learning. In the context of a drug design problem, we have investigated the application of a data-driven adaptive Markov chain approach [1], where the acceptance probability is given by a probabilistic surrogate of the target property, modelled with a maximum entropy conditional model. We have applied the approach [2] to a lead development search for an antagonist of an α_v integrin, using a molecular docking score as the optimisation function. The RGD integrin receptors are thought to play a key role in fibrosis. Antagonism of $\alpha_v\beta_6$ is one promising avenue for the development of a novel therapeutic treatment and some success has been reported [3] in discovering compounds with significant activity against $\alpha_v\beta_6$ and physicochemical properties commensurate with oral bioavailability. Molecular docking was performed using OpenEye FRED, which uses a rigid ligand approach, where a large number of conformations are generated and each of those are docked successively. The adaptive Markov chain algorithm is (i) soundly based in machine learning; (ii) proposes structures from an implicitly defined space of potential designs; (iii) is guaranteed to converge; and (iv) achieves a large structural variety of new compounds predicted to be active, some of which provoke significant interest from a medicinal chemistry perspective.

[1] D. Oglic, R. Garnett & T. Gärtner. Active search in intensionally specified structured spaces. *Proc. 31st AAAI Conf. Artif. Intell*., **2017**, 2449.
[2] Oglic D. et al. Active search for computer-aided drug design. *Molecular Informatics*, **2018**, in press.
[3] Adams J., et al. Structure activity relationships of αv integrin antagonists for pulmonary fibrosis by variation in aryl substituents. *ACS Med. Chem. Lett*., **2014**, *5*, 1207.

POSTER 4

Augmenting De novo Drug Design using Reaction Classification

G. Ghiandoni[1], B. Chen [2], M. J. Bodkin [3], V. J. Gillet. [1]

[1] *Information School, University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, United Kingdom*
[2]*Chemistry Department, University of Sheffield, Dainton Building, Brook Hill, Sheffield, S3 7HF, United Kingdom*
[3]*Evotec (U.K.) Ltd, 114 Innovation Drive, Milton Park, Abingdon, OX14 4RZ,United Kingdom*

The rational design of tailored chemical structures, with desired pharmacodynamic and pharmacokinetic properties, can be achieved through the application of de novo molecular design methodologies.[1]
The potential number of chemical structures that meet Lipinski's Rule-of-Five (RO5) requirements for drug-likeness, has been estimated at 10^{60}molecules.[2]Nevertheless, the introduction of fragment-based construction design techniques, that permit the evaluation of the synthetically accessible chemical space, has drastically reduced the drug-like space into a smaller number of structures.[3]More specifically, a novel structure generation tool has been implemented using reaction vectors, which are descriptors that are able to encode accurately the topological changes occurring in real reaction examples, in order to design molecules that are likely meet the criteria of synthetic feasibility.[4,5]
Herein, we present a machine learning model for reaction classificationthat is based on reaction vectors and is trained towards the prediction of 336 medicinal chemistry reaction classes. The modelenablesthe direct selection and application of desired sets of reactionsfor de novo design applications.For example, the model can be applied in order to differentiate or prioritize particular reaction classes such as functionalizations or protections/deprotections; or simply to enhancethe exploration of ELNs (electronic laboratory notebooks).Furthermore, we present an example of augmented de novo design, where reaction classification is applied in order to exploit the knowledge within a database of ninety-three thousand pharmaceutical reactions.

1. Hartenfeller, M. & Schneider, G., Enabling future drug discovery by de novo design. *Wiley Interdisciplinary Reviews: Computational Molecular Science*. **2011**, 1, 742-759.

2. Bohacek, R.S., McMartin, C. & Guida, W.C., The art and practice of structure-based drug design: A molecular modeling perspective. *Medicinal Research Reviews*. **1996**, 16, 3-50.

3. Hartenfeller, M., Reaction-Driven De Novo Design: aKeystone for Automated Design of Target Family-Oriented Libraries. In*De novo Molecular Design*; Schneider, G., Ed.; Wiley-VCH Verlag GmbH & Co. KGaA, Weinheim, Germany, **2013**; 245-266.

4. Patel, H., Bodkin, M. J., Chen, B. & Gillet, V. J., Knowledge-Based Approach to De Novo Design Using Reaction Vectors. *Journal of Chemical Information and Modeling*. **2009**, 49, 1163-1184.

5. Hristozov, D., Bodkin, M., Chen, B., Patel, H., & Gillet, V. J., Validation of Reaction Vectors for de Novo Design.*Library Design, Search Methods, and Applications of Fragment-Based Drug Design*, **2011**, 29-43.

POSTER 5

Machine Learning: Predicting Chemical Ames Mutagenicity

Charmaine S.M. Chu, *Neil G. Berry*
Department of Chemistry, University of Liverpool, Crown Street, Liverpool L69 7ZD

For modern drug discovery, mutagenicity detection is one of the key element to try and avoid toxicity. However, the traditional method of mutagenicity detection using the Ames test[1] is costly and time consuming as the compound which need to be tested for toxicity need to be synthesised and the test result is not 100% accurate and reproducible. Therefore, it is necessary to develop accurate and robust *in silico* models which can predict accurately the mutagenicity of a compound before synthesis and to overcome the hurdles associated with the Ames test. Using a previously defined compound mutagenicity library (~7000 compounds), chemical fingerprints and molecular properties were calculated and 11 classification modelling algorithms, including random forest (RF), support vector machine (SVM), mixture discriminant analysis (MDA) and k-nearest neighbour (KNN) have been developed and tested and has identified models with very good performance. These models have comparable or better performance in comparison to the model defined previously in by the Congying Xu group.[2] It was found that the RF models built during this study have excellent performance, followed by some of the SVM, MDA and KNN models. Their validity has been rigorously tested using external test sets (e.g. MDA AUC ROC = 0.95) and y-randomisation approaches. In the future, we will investigate and interpret the models to discover molecular features which are important in Ames positive and negative compounds.

References:

[1] B. N. Ames, E. G. Gurney, J. A. Miller and H. Bartsch, *Proceedings of the National Academy of Sciences*, (19720, **69**, 3128-3132.
[2] C. Xu, F. Cheng, L. Chen, Z. Du, W. Li, G. Liu, P. W. Lee and Y. Tang, *Journal of Chemical Information and Modeling*, (2012), **52**, 2840-2847.

POSTER 6

On the mechanisms of DNA G-quadruplexes targeting by gold(I) N-heterocyclic compounds: new insights by combined meta-dynamics and biophysical methods

Darren Wragg,[1] Andreia de Almeida,[1] Riccardo Bonsignore,[1] Stefano Leoni[1,] and Angela Casini[1,]

School of Chemistry, Cardiff University, Park Place, CF10 3AT Cardiff, United Kingdom

WraggDD@cardiff.ac.uk

G-quadruplex (G4s) are secondary DNA structures, formed by guanine-rich sequences coupled by Hoogsteen hydrogen bonds and stabilized by cations (e.g. K^+), which are present in telomeres and promotor regions of oncogenes.[1] G4s are known to be involved in a number of biological processes, such as telomere maintenance and replication as well as oncogenes regulation[2,3] and have thus become an important targets for cancer therapy.

A number of recent experimental[4] and *in silico*[5] studies report on the stabilisation of G4s by small molecules, leading to the blockage of G4s capability of unwinding, consequently inhibiting DNA polymerase function. This effect is of particular importance in cases where the G4 structure is in a promotor area of an oncogene.

Within this context, an organometallic Au(I) N-heterocyclic carbene (NHC) compound featuring caffeine moieties - Au(9-methylcaffein-8-ylidene)$_2$ (AuTMX$_2$)(Fig 1) - has been recently observed not only to be able to stabilize G4s structures, but to be also highly selective for binding to G4s DNA, when compared to duplex or single-stranded DNA[4,6].

By combining state-of-the-art free energy calculation methods, such as metadynamics, with FRET (fluorescence resonance energy transfer) DNA melting studies, we aim here to elucidate the mechanism of G4s stabilisation by this gold NHC complex at a molecular level, and to identify its favourite binding poses and related binding energies. Overall, our integrated approach will enable the development of highly selective gold-based chemotherapeutic agents targeted to G4s structures.

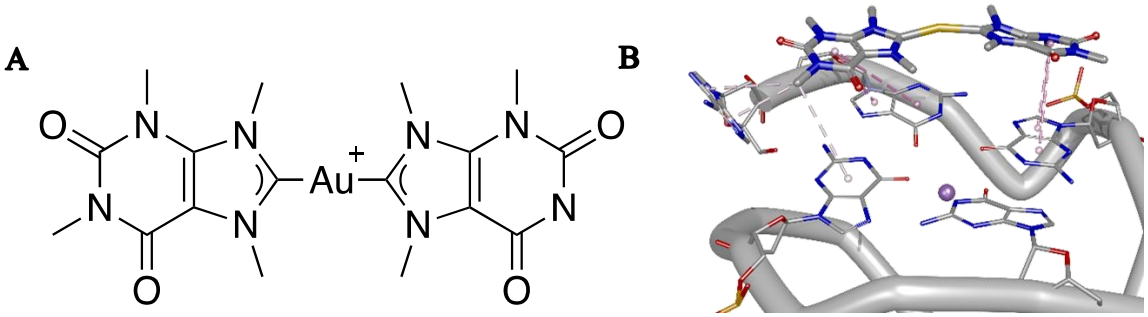

Fig 1. (**A**) Au(9-methylcaffein-8-ylidene)$_2$ (AuTMX$_2$). (**B**) Interaction of AuTMX$_2$ with ckit-01.

[1] V.S. Chambers et al., *Nat. Biotechnol.*, 2015, *33*, 877
[2] S. Neidle, *Nat. Rev. Chem*. 2017, 1, 10.
[3] D. Rhodes, H. J. Lipps, *Nucleic Acids Res*. 2015, 43, 8627 – 8637.
[4] Ö. Karaca et al, *Chem. Commun*., 2017, 53, 8249–8260.
[5] F. Moracae al., *Proc. Natl. Acad. Sci*., 2017, 114, E2136–E2145.
[6] B. Bertrand, A. Casini, *Dalt. Trans*., 2014, 43, 4209–4219.

POSTER 7

Using advanced computational methods to model the binding of flexible proteins: a case study from the coagulation cascade

Martin Rosellen (UCL), Prof. Flemming Hansen (UCL), Prof. Adrian Shepherd (Birkbeck)

Understanding the impact of a single (or a few) mutations on antibody binding may be useful - both
mutations within an antibody (which may aid the identification of higher-affinity therapeutic antibodies) as well as within an antigen (which may help us identify potential escape mutations of a pathogen, or engineer therapeutic proteins that are less immunogenic with respect to host antibodies
considered immunodominant). There are a range of experimental techniques, such as alanine scanning mutagenesis, that can characterise the contribution of particular residues to the binding between antibody (Ab) and antigen (Ag), but they are generally time consuming and expensive. In
this context, (reasonably) accurate computer simulations are a very appealing option.

Recently it has been shown that accurate simulations of Abs bound to the stalk of influenza A hemagglutinin are possible using molecular dynamics simulation. The aim of this project is to adapt this approach in the context of the binding of human monoclonal Ab BO2C11 to the C2 domain of Factor VIII, a monoclonal Ab known to inhibit the effectiveness of therapeutic FVIII. Accurate simulation of substitutions that have not been characterised experimentally may engender useful insights into the likelihood of BO2C11-like Abs cross-reacting with the endogenous FVIII of
patients with disease-causing substitutions in this region or ways to modify the therapeutic to make it less immunogenic.

Results for a set of substitution of the epitope of BO2C11 show very good agreement with published experimental values in 7 cases, with additional 3 clustered in one region of the epitope. Further, with comparisons between the C2-domain in complex with BO2C11 versus in the unbound form we are investigating the structural and energetic foundations behind the non-binding mutations R2220A/Q as reported by Pratt et. al..

POSTER 8

WhichP450 – A Multi-class Categorical Model to Predict the Major Metabolising CYP450 Isoform for a Compound

Peter A. Hunt, Matthew D. Segall, Jonathan D. Tyzack; Optibrium Ltd.

In the development of novel pharmaceuticals, the knowledge of how many, and which, Cytochrome P450 isoforms are involved in the phase I metabolism of a compound is important. Potential problems can arise if a compound is metabolised predominantly by a single isoform in terms of drug-drug interactions or genetic polymorphisms that would lead to variations in exposure in the general population. Combined with models of regioselectivities of metabolism by each isoform, such a model would also aid in the prediction of the metabolites likely to be formed by P450-mediated metabolism. We describe the generation of a multi-class random forest model to predict which, out of a list of the 7 leading Cytochrome P450 isoforms, would be the major metabolising isoforms for a novel compound. The model has a 76% success rate with a top-1 criterion and an 88% success rate for a top-2 criterion and shows significant enrichment over randomised models.

POSTER 9

Water Networks and Topological Data Analysis: A Data Science Approach to Understanding Intermolecular Structures

Lee Steinberg, *Jeremy Frey*

School of Chemistry, University of Southampton, Highfield, Southampton, SO17 1BJ, UK

The nature of water networks is of fundamental importance to the field of drug discovery. For example, understanding how the presence of a solute affects the native water structure enables us to build better models for the prediction of free energy changes of solvation, and would provide more consistent models for solubility. Previous investigations of such systems have tended to include radial or spatial distribution functions or a graph theoretical approach. However, such methods suffer from difficulties in understanding behavior beyond nearest neighbor interactions, and the requirement for a chemical heuristic determining correlation respectively.

Topological data analysis (TDA), a series of recent developments in the field of data science, seeks to understand the "shape" of data. In particular, persistent homology, one of the main branches of TDA, is an attempt to determine the underlying structure of a data set by searching it for n-dimensional holes. Such techniques have found uses in chemical fields – in particular materials science. Examples include the determination of similarity in nanoporous materials through pore recognition[1], and the understanding of various motifs in amorphous structures[2].

Here, we develop a persistent homology approach to understand the structure of water networks, calculated from simulations. We improve pre-existing techniques to construct an 'average' persistence, and use these techniques to build a size-independent persistent homology model for a fluid system. We construct a descriptor which is well-suited to machine learning models, and demonstrate its use in a support vector machine formalism. We then apply this method to four commonly used water models (TIP3P, TIP4P/Ew, SPC/E, OPC) and demonstrate how we can use persistent homology to understand the tetrahedrality of water networks. We lastly show that a persistent homology model enables us to recover well-known statements of similarity between these water models.

References:
[1] Y. Lee, S. Barthel, P. Dłotko, S. M. Moosavi, K. Hess, B. Smit, *Nature Communications*, 2017, DOI: 10.1038/ncomms15396
[2] Y. Haraoka., T. Nakamura, A. Hirata, E. G. Escolar, K. Matsue, Y. Nishiura, *PNAS*, 2016, **113**, 7035-7040

POSTER 10

Comparison of multitask prediction methods for chemical data

Antonio de la Vega de León, Valerie J Gillet

University of Sheffield, Regent Court, 211 Portobello, S1 4DP Sheffield, United Kingdom

Multitask prediction, where several outputs are predicted using one model, has become more common in the chemoinformatics field during the last decade. The popularity of deep neural networks have been central in achieving this feat. They have been used on large scale predictive modelling of bioactivity and screening data.[1–3] Deep neural networks have also become a popular basis for generative models in drug discovery.[4]

Deep neural networks are not the only multitask prediction method available. Random Forests, a commonly used technique in chemoinformatics, are able to perform multitask prediction. There has been work done to better adapt these tree ensemble models for multitask prediction.[5] Additionally, methods based on probabilistic matrix factorization have been applied to chemoinformatics problems.[6] These methods provide alternatives to deep neural networks and they do not require large GPU resources that might not be available in all research environments.

In this work, we perform a comparison of multitask methods on different chemical data sets. We compare deep neural networks to Macau, a new and novel method based on Bayesian probabilistic matrix factorization. We used different regression data sets in our comparisons. Our results show that Macau can be competitive to deep neural networks and provides a valid alternative to this frequently used method.

The research leading to these results has received funding from the European Union's Seventh Framework Programme (FP7/2007-2013) under grant agreement n°612347.

1. Dahl, G. E., Jaitly, N. & Salakhutdinov, R. Multi-task Neural Networks for QSAR Predictions. *arXiv* 1406.1231 (2014).
2. Unterthiner, T. *et al.* Deep Learning as an Opportunity in Virtual Screening. *Deep Learn. Represent. Learn. Work. NIPS 2014* 1–9 (2014).
3. Ramsundar, B. *et al.* Massively Multitask Networks for Drug Discovery. *arXiv* 1502.02072 (2015).
4. Schneider, G. Generative Models for Artificially-intelligent Molecular Design. *Mol. Inform.* **37,** 1880131 (2018).
5. Simm, J. & Magrans De Abril, I. Tree-Based Ensemble Multi-Task Learning Method for Classification and Regression. *IEICE Trans. Inf. Syst.* 1677–1681 (2014).
6. Simm, J. *et al.* Macau: Scalable Bayesian Multi-relational Factorization with Side Information using MCMC. *arXiv* 1509.04610 (2015).

POSTER 11

Investigating the importance of solid state descriptors for statistical models of temperature dependent aqueous solubility

R.L. Marchese Robinson,[a] Chris Morris,[b] Rebecca Mackenzie,[b] K.J. Roberts,[a] E.B. Martin[a]

a. School of Chemical and Process Engineering, University of Leeds, Leeds LS2 9JT, United Kingdom

b. Science and Technologies Facilities Council, Daresbury Laboratory, Sci-Tech Daresbury, Warrington WA4 4AD, United Kingdom

Over the years, a plethora of computational models, including quantitative structure-property relationships (QSPRs), have been developed to predict aqueous solubility, at a single temperature, as an indicator of active pharmaceutical ingredient (API) bioavailability [1]. Few studies [2, 3] have developed QSPR models for temperature dependent solubility profiles. However, these predictions are needed to support the digital design of unit operations in pharmaceutical manufacturing, e.g. cooling crystallization and wet granulation. The work reported in this presentation builds upon the earlier QSPR studies of temperature dependent aqueous solubility by critically examining whether the incorporation of descriptors explicitly capturing solid state contributions to solubility and its temperature dependence actually add value to the model. This work builds upon the debate in the recent literature regarding the importance of (explicitly) capturing the solid state contribution in models of aqueous solubility and the extent to which molecular descriptors can capture this [4, 5, 6]. This presentation will situate this work within the context of the UK ADDoPT (Advanced Digital Design of Pharmaceutical Therapeutics) project (www.addopt.org) and discuss the results obtained to date.

1. Skyner et al., *Phys.Chem.Chem.Phys.,* **2015** (17), p.6174
2. Avdeef, *ADMET & DMPK*, **2015** (3), p.298
3. Klimenko et al., *J. Comput. Chem.*, **2016** (37), p.2045
4. Salahinejad et al., *Mol Pharmaceutics*, **2013** (10), p. 2757
5. Emami et al., *J. Solution Chem.*, **2015** (44), p. 2297
6. Abramov, *Mol Pharmaceutics*, **2015** (12), p.2126

POSTER 12

Meta-QSAR: a large-scale application of meta-learning to drug design and discovery

Ivan Olier, Noureddin Sadawi, G. Richard Bickerton, Joaquin Vanschoren, Crina Grosan, Larisa Soldatova, Ross D. King
Liverpool John Moores University

We investigate the learning of quantitative structure activity relationships(QSARs) as a casestudy of meta-learning. This application area is of the highest societal importance, as it is a key step in the development of new medicines. The standard QSAR learning problem is: given a target (usually a protein) and a set of chemical compounds (small molecules) with associated bioactivities (e.g. inhibition of the target), learn a predictive mapping from molecular representation to activity. Although almost every type of machine learning method has been applied to QSAR learning there is no agreed single best way of learning QSARs, and therefore the problem area is well-suited to meta-learning. We first carried out the most comprehensive ever comparison of machine learning methods for QSAR learning: 18 regression methods, 3 molecular representations, applied to more than 2700 QSAR problems. We then investigated the utility of algorithm selection for QSAR problems. We found that this meta-learning approach outperformed the best individual QSAR learning method (random forests using a molecular fingerprint representation) by up to 13%, on average. We conclude that meta-learning outperforms base-learning methods for QSAR learning, and as this investigation is one of the most extensive ever comparisons of base and meta-learning methods ever made, it provides evidence for the general effectiveness of meta-learning over base-learning.

This research work was recently published in Machine Learning, 2018:

Olier, I., Sadawi, N., Bickerton, G. R., Vanschoren, J., Grosan, C., Soldatova, L., & King, R. D. (2018). Meta-QSAR: a large-scale application of meta-learning to drug design and discovery. Machine Learning, 107(1), 285–311. https://doi.org/10.1007/s10994-017-5685-x

POSTER 13

DL_ANALYSER Notation for Atomic Interactions (DANAI): A Natural Annotation System for Molecular Interactions, Using Carboxylic Acids as Test Cases

Chin W. Yong

Scientific Computing Department, Science and Technology Facilities Council, Daresbury Laboratory, Sci-Tech Daresbury, Warrington WA4 4AD, UK

Manchester Pharmacy School, Faculty of Medical and Human Sciences, University of Manchester, Manchester M13 9NT, UK

Abstract

The DL_ANALYSER Notation for Atomic Interactions**,** DANAI, is the notation syntax to describe interactions between molecules [1]. This notation can annotate precisely the detailed atomistic interactions without having to resolve to diagrammatic illustrations, and yet can be interpreted easily by both human users and computational means. By making use of the DL_F Notation [2], a universal atom typing scheme for molecular simulations, DANAI contains the expression of atomic species in a natural chemical sense. It is implemented within DL_ANALYSER, a general analysis software program for DL_POLY molecular dynamics simulation software. By making references to the molecular dynamics simulations of pure carboxylic acid liquids, it is shown that DL_ANALYSER can identify and distinguish a variety of hydrogen bond and hydrophobic contact networks, through the use of the DANAI expression. From such, statistical analysis such as the correlation coefficients can be carried out. It was found that the carboxylic groups preferentially orientated in a "head-to-tail" conformation to form hydrogen bonds between the carbonyl oxygen and hydroxyl hydrogen, resulting in a series of linear structures that intertwined with pockets of methyl clusters.

[1] C. W. Yong & I. T. Todorov, *Molecules* (2018), **23**, 36
[2] C. W. Yong, *J. Chem. Inf. Model.* (2016), **56**, 1405-1409

POSTER 14

G-protein coupled receptor-KNIME Automated Modelling Platform (GPCR_KAMP)

A. Pal, Dublin/IE, G.K. Kinsella, Dublin/IE, A.J. Chubb, Dublin/IE
Ajay Pal, Dublin Institute of Technology, Dublin 1, Dublin
Dr. Gemma K Kinsella, Dublin Institute of Technology, Dublin 1, Dublin
Dr. Anthony J Chubb, Royal College of Surgeons in Ireland, Dublin 2, Dublin

G-protein coupled receptors (GPCRs) are widely expressed cell surface receptors and the most successfully exploited drug targets with approximately 30% of currently marketed drugs targeting human GPCRs. Prediction of three dimensional structures of GPCRs can help in the identification of new small molecule therapeutics, biomarkers, and better understanding of the protein ligand interactions. [1]
Homology modelling coupled with virtual screening (molecular docking) is one of the leading structure based drug discovery approaches utilised to screen large databases *in silico* for hit compounds which can be optimised to lead molecules. [1] However, the large number of computational steps deriving from high-throughput studies required to pre-process and analyse the results obtained from homology modeling and molecular docking simulations represents a bottleneck. Indeed, several programs need to be used to accomplish a number of tasks (such as homology model building, predicted models' score analysis, protein file formatting with charge and torsion angles, ligand file formatting, defining the ligand binding pocket in protein and molecular docking), requiring computational competences and resources not always present in an experimental group.
In this context, the KNIME Analytics Platform [2] was used to develop a pipeline joining the MODELLER [3], GHECOM [4], SMINA [5] with in-house Python scripts, and KNIME functionalities to perform the aforementioned steps even by users unfamiliar with programming. Here, the pipeline developed for linux cluster was benchmarked with 37 x-ray crystallized GPCRs available in RCSB database (www.rcsb.org) with bound small molecules. GPCR_KAMP will be an opensource KNIME pipeline available via Google code and GitHub.

Literature:
[1] Carlsson J, Coleman RG, Setola V, Irwin JJ, Fan H, Schlessinger A, Sali A, Roth B, Shoichet BK. Nat Chem Biol. 2011, 7(11):769-78.
[2] Berthold M, Cebron N, Dill F, Gabriel T, Kötter T, Meinl T, Ohl P, Springer Berlin Heidelberg, 2007, 319−326.
[3] Webb B, Sali A, John Wiley & Sons, Inc., 2014, 5.6.1-5.6.32.
[4] Kawabata T, Proteins, 2010, 78, 1195-1121
[5] Koes D, Baumgartner M, Camacho J, Journal of Chemical Information and Modeling 2013, 53 (8), 1893-1904

POSTER 15

From exploration to exploitation: how do models depict trends during lead optimisation?

David Marcus, Chris Luscombe, Stephen Pickett, Stefan Sanger and Darren Green

Lead optimisation is a multiple cycle effort to improve one or more properties of a hit series and produce the best clinical candidates. To accelerate this effort and reduce costs, molecular models can suggest structural modifications by predicting multiple parameters of newly generated molecules. However, initial models could suffer from low applicability or high overfitting that could result in low performance and over-exploitation of the initial hit series. Active-learning has the potential to overcome these hurdles by controlling the number of similar/novel molecules and suggest molecules that could also improve the models iteratively. Shifting between exploration and exploitation has long been implemented during lead optimisation when medicinal chemists analyse the effects of molecular modification on the SAR. In this poster, we analyse the shifts in chemical space and the trends of exploration and exploitation during the life cycle of several lead optimisation efforts. We find that molecular models serve as a good measure to evaluate uncertainty in chemical space and provide good estimate when exploration and exploitation should be applied. They can also suggest the best timing to apply active-learning approach to a lead optimisation effort.

POSTER 16

Adaptation of a Matched Molecular Pair Identification Algorithm for Solid State Informatics Analysis of the Cambridge Structural Database

Jakub Janowiak [a], .E.B. Martin [a], K.J Roberts [a], R.L. Marchese Robinson,[a], A. Maloney,[b] I. Giangreco,[b] K. Pencheva[c]

a. School of Chemical and Process Engineering, University of Leeds, Leeds LS2 9JT, United Kingdom

b. Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge CB2 1EZ, United Kingdom

c. Pfizer Worldwide R&D, Sandwich, Kent, United Kingdom

Solid state properties such as polymorphism are an important consideration for product development in the pharmaceutical industry. The Cambridge Structural Database (CSD) contains an ever-growing amount of crystal related data. The Matched Molecular Pair (MMP) methodology [1] is one of the techniques that has been developed to mine large datasets of molecular data. The Hussain and Rea Fragmentation (HRF) method [2] is a computationally efficient algorithm for the identification of MMPs without the need for pre-defined transformations. However, the original implementation in the Rdkit has several limitations including (i) the inability to add new molecules without re-running the entire script, and (ii) the large number of MMPs identified that do not occur in sufficient number to obtain statistically significant results. In this poster, an adaptation of the HRF algorithm – an MMP database approach – is presented that leverages database capabilities to address these issues. New molecules can be added without the need to re-run the identification algorithm on the entire dataset. This is especially useful when working with CSD due to it receiving periodic updates. For a given pair of molecules, the most “chemically meaningful” transformation is selected based on the size of the change. This step approximately halves the number of MMPs. An interactive Jupyter notebook has also been developed to aid in the analysis process. The poster compares the two methods and presents a case study using the CSD to study polymorph propensity.

[1] C. Tyrchan, E. Evertsson, Matched Molecular Pair Analysis in Short: Algorithms, Applications and Limitations, Comput. Struct. Biotechnol. J. 15 (2017) 86–90. doi:10.1016/j.csbj.2016.12.003.

[2] J. Hussain, C. Rea, Computationally Efficient Algorithm to Identify Matched Molecular Pairs (MMPs) in Large Data Sets, J. Chem. Inf. Model. 50 (2010) 339–348. http://pubs.acs.org/doi/abs/10.1021/ci900450m.

POSTER 17

The mechanism of aquaporin inhibition by gold compounds elucidated by biophysical and computational methods

Andreia de Almeida [1]*, Andreia Mósca [2,3], Darren Wragg [1], Margot Wenzel [1], Stefano Leoni [1], Graça Soveral [2,3] and Angela Casini [1]

[1] *School of Chemistry, Cardiff University, Main Building, Park Place, CF103AT Cardiff, UK*
[2] *Research Institute for Medicines (iMed.ULisboa), Faculty of Pharmacy, Universidade de Lisboa, 1649-003 Lisboa, Portugal.*
[3] *Dept. Bioquímica e Biologia Humana, Faculty of Pharmacy, Universidade de Lisboa, 1649-003 Lisboa, Portugal*

deAlmeidaA@cardiff.ac.uk

Aquaporins (AQPs) are membrane channels that facilitate the transport of water and/or glycerol across cellular membranes and are crucial to cell physiology. Mechanisms of water flux gating through classical AQPs have been described, however, less is known about the regulation of water and glycerol transport through members of the aquaglyceroporin subfamily.[1,2] Aquaporin-3 (AQP3), an aquaglyceroporin isoform, has been shown to be over-expressed in several cancer types and to have a crucial role in tumour progression, making it an important target for cancer therapeutics.[1]

Recently, Au(III) compounds have been found to selectively inhibit glycerol permeation of AQP3.[3] In this work, the inhibition of water and glycerol permeation via human AQP3 by Au(III) complexes has been studied by stopped-flow spectroscopy and, for the first time, its mechanism has been described using molecular dynamics (MD), combined with density functional theory (DFT) and electrochemical studies. The detailed molecular mechanism of inhibition of the most potent compound [Au(PbImMe)Cl2]PF6 (Fig. 1) was studied and, for the first time, important structural changes leading to pore closure upon gold binding were identified. [4]

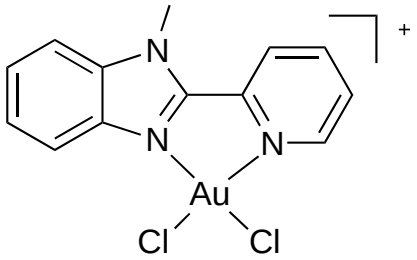

Fig.1. Structure of [Au(PbImMe)Cl2]PF6, a potent and selective AQP3 inhibitor.

References

1. Aquaporins in health and disease: new molecular targets for drug discovery, G. Soveral, S. Nielsen, A. Casini, CRC Press, Taylor & Francis Group **2016**
2. S. Verkman, M. O. Anderson and M. C. Papadopoulos, Nat. Rev. Drug Discov., **2014**, 13, 259–77
3. A. P. Martins, A. Ciancetta, A. de Almeida, A. Marrone, N. Re, G. Soveral and A. Casini, ChemMedChem, **2013**, 8, 1086–1092
4. A. de Almeida, A. F. Mósca, D. Wragg, M. Wenzel, P. Kavanagh, G. Barone, S. Leoni, G. Soveral and A. Casini, Chemical Communications, **2017**, 53, 3830-3033.

POSTER 18

Regression conformal prediction of ADME data using error models for normalisation

Christina Founti[1], Val Gillet[1] and Jonathan Vessey[2]

1: Information School, The University of Sheffield, Regent Court, 211 Portobello, Sheffield, S1 4DP, UK
2: Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds LS11 5PS, UK

Predictive models are integral tools for decision-making and routinely used in drug development for the identification of new leads and ADMET property optimisation. However, despite the availability of large amounts of data models often fail to deliver accurate predictions. Although the accuracy of future predictions is unknown, the reliability of predictions can assessed using error estimation methods, such as domain applicability models, error models or resampling. The error estimates obtained may then be used for the calculation of confidence intervals or prediction intervals. While the former assess the range of values where the prediction lies, prediction intervals assess a range of values that contains the true value. A number of methods for prediction interval estimation are available but rely on the assumption that residual errors are normally distributed.

Conformal prediction is a machine learning framework that integrates prediction interval estimation in QSAR modelling without making assumptions about the error distribution [1]. This method is used to construct valid prediction intervals that represent the reliability of the individual predictions at a user-defined level of confidence. The nonconformity score, i.e. the error margin, for the calculation of prediction intervals is inferred from the empirical error distribution of calibration data. The calculation of compound-specific prediction intervals requires prior normalisation of the errors with the expected model accuracy.

Although conformal prediction guarantees the validity of prediction intervals, high efficiency is difficult to obtain and particularly at higher confidence levels [1-3]. Efficiency can be improved by the normalisation of nonconformity scores with suitable error estimates. In this study, the performance of regression conformal predictors using different error models [4, 5] as normalisation factors is assessed for ADME datasets. Further considerations include the effect of dataset size, QSAR and error model performance.

Abbreviations
ADMET: Absorption, Distribution, Metabolism, Excretion, Toxicity
QSAR: Quantitative Structure Activity Relationship

[1] Eklund M., Norinder U., Boyer S., Carlsson L. (2012) Application of Conformal Prediction in QSAR. In: Iliadis L., Maglogiannis I., Papadopoulos H., Karatzas K., Sioutas S. (eds) Artificial Intelligence Applications and Innovations. AIAI 2012. IFIP Advances in Information and Communication Technology, Vol 382. Springer, Berlin, Heidelberg

[2] Carlsson, L., Eklund, M., Norinder, U. Aggregated Conformal Prediction. Iliadis, L., Maglogiannis, I., Papadopoulos H., Sioutas, S., Makris, C. 10th IFIP International Conference on Artificial Intelligence Applications and Innovations (AIAI), Sep 2014, Rhodes, Greece. Springer, IFIP Advances in

Information and Communication Technology, AICT-437, pp.231-240, 2014, Artificial Intelligence Applications and Innovations.

[3] Lindh, M., Karlén, A., Norinder, U. (2017) Predicting the Rate of Skin Penetration Using an Aggregated Conformal Prediction Framework. Molecular Pharmaceutics, 14: 1571-1576.

[4] Sheridan, RP. (2012) Three useful dimensions for domain applicability in QSAR models using random forest. Journal of Chemical Information and Modelling, 52 (3): 814-23.

[5] Toplak M., Močnik R., Polajnar M., et al. (2014) Assessment of Machine Learning Reliability Methods for Quantifying the Applicability Domain of QSAR Regression Models. Journal of Chemical Information and Modelling, 54 (2): 431-441.

POSTER 19

Application of quantum mechanics and/or structural fingerprinting for the prediction of glutathione conjugation site-specificity?

Martin Payne, Jeff Plante and David Ponting
Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, United Kingdom

Glutathione conjugation of drugs or xenobiotics is an important metabolic process with repercussions (both positive and negative) for chemical toxicity. However the potential for such reactions may go unrecognised and the specific site of reaction may influence the toxicity *in vivo*. Quantum mechanical calculations have long been seen as a means of identifying electrophilicity and more recently as a means of distinguishing reactions with hard and soft nucleophiles. However, enzyme mediated reactions complicate predictive modelling of metabolite products. In this work we look at the contribution that quantum mechanics can make over and above the use of sub-structural fingerprinting.

An in-house dataset of glutathione conjugation reactions, generated from an earlier careful extraction from published literature, contained 1038 reactions for 608 starting molecules. Using NWChem 6.6, a range of molecular and atom-based descriptors were generated including HOMO and LUMO energies and their atomic populations, Wondrousch and conventional local electrophilicities, and Fukui reactivity indices. Two fingerprints were also calculated; firstly MACCS keys, which relate to functional groups and may thus be useful in a reactivity analysis, and an atom typer fingerprint developed in-house, modified from the Ghose and Crippen atom types.

Using the experimental data and calculated descriptors, statistical models for the prediction of the sites of conjugation were developed using a Support Vector Machine (SVM) methodology. The performance of models globally and for individual reactivity classes have been considered. Comparison was also made with predictions using Meteor Nexus (Lhasa Limited). The local fingerprint was originally included as a counter-hypothesis to the use of quantum-mechanical descriptors, but showed better performance, alone, than the latter. However, the best models for many classes included the quantum-mechanical descriptors indicating that these are useful when searching for models regardless of computational cost. The strengths and weaknesses of significant descriptors are considered with a view to improving performance and giving mechanistic interpretations of models.

POSTER 20

Conformational variation from a spectral geometry perspective

M Seddon [1], D Cosgrove[2], M Packer [3], V Gillet [1]

[1] University of Sheffield, Sheffield, UK, [2] CozChemIx Limited, Macclesfield, UK, [3] AstraZeneca, Cambridge, UK

Three-dimensional molecular shape is a key determinant of molecular interactions[1]. To date, widespread use of 3D similarity methods in drug development has been hampered by computational complexity surrounding structure alignment and molecular flexibility. Typically, 3D shape comparison treats the molecules as rigid bodies and flexibility is taken into account using conformation ensembles to sample conformational space, which significantly increases the computational cost of 3D similarity searching on large molecular databases.

Spectral geometry provides a framework for exploring concepts of flexible 3D shape[2]. In brief, spectral geometry treats the surface of a 3D shape as a curved 2D surface and encodes the geometric properties in the spectrum of the Laplace-Beltrami Operator over that surface. These methods are of particular interest for high throughput virtual screening because they produce rich descriptors of 3D shape that are alignment-invariant and also invariant to a specific class of flexibility, called *isometric deformation*. Furthermore, the conceptual framework has a large amount of promise for investigating the relationship between 3D molecular shape and conformational variation in a mathematically robust manner.

In this poster we have used the spectral geometry framework to explore the relationship between conformational variation and molecular shape. In particular, the extent to which the variation in conformational shape can be captured by the isometric deformation assumption. Our results suggest the spectrum of the Laplace-Beltrami Operator is sufficient to identify conformations of a single molecule and we propose a method for determining when two conformations are different shapes from a spectral geometry perspective.

(1) Nicholls, A.; McGaughey, G. B.; Sheridan, R. P.; Good, A. C.; Warren, G.; Mathieu, M.; Muchmore, S. W.; Brown, S. P.; Grant, J. A.; Haigh, J. A.; et al. Molecular Shape and Medicinal Chemistry: A Perspective. *J. Med. Chem.* **2010**, *53* (10), 3862–3886.

(2) Biasotti, S.; Cerri, A.; Bronstein, A.; Bronstein, M. Recent Trends, Applications, and Perspectives in 3D Shape Similarity Assessment. *Comput. Graph. Forum* **2015**, n/a-n/a.

POSTER 21

Large-scale comparison between robust methods for QSAR

Philippa G. McCabe, Sandra Ortega-Martorell, Andrew Leach, Ivan Olier
Liverpool John Moores University

Predicting the ability of chemical compounds to activate or inhibit proteins would facilitate the process of drug discovery and development. This process is known as learning quantitative-structure activity relationships (QSARs) for the interaction between compounds and protein targets. Aimed at understanding and improving the technology of QSAR building, we recently used public chemoinformatics databases to generate ~10,000 datasets using freely available molecular descriptors and fingerprints. We identified that around half of the datasets were too small or not diverse enough hence becoming unsuitable for QSAR modelling. This is a common problem in QSAR research but more data acquisition means more wet lab assays, which are usually costly, time-consuming and hazardous. Deep learning, which has been recently suggested as an excellent method for QSAR modelling, may not be of help in this context. In this research, we are performing a large quantitative comparison between deep learning and other traditionally robust machine learning methods such as random forest and gradient boost machines. Preliminary results suggest that there is no unique way of learning QSARs and that the benefit of using deep learning in small datasets is rather limited or non-existent.

POSTER 22

Modelling ChEMBL protein targets using conformal prediction

Nicolas Bosc, EBI

Over the years, QSAR modelling has become a standard tool to estimate the activity of molecular compounds on protein targets. Despite many efforts to develop new molecular descriptors and to introduce more accurate and efficient machine learning techniques, the reliability of QSAR predictions remains a major pitfall. Recently, conformal prediction has emerged as an extension of QSAR that aims not only to return a prediction but also a confidence associated with the output.[1]
With a clean and reliable data source, it is now possible to build a large collection of models and to assess their reliability. Using the ChEMBL database in its version 23[2], we developed conformal predictive models for a set of nearly 800 targets. After having selected activity data from the database for each target, the compounds were classified as active or inactive according to the protein family-specific thresholds introduced by the Illuminating Druggable Genome consortium.[3] Class conformal models were then generated for each protein using the Random Forest method. Using different confidence values allowed us to estimate how many models could be applicable to achieve reliable predictions.
When it comes to evaluate the performance of a conformal model, validity and efficiency are often considered. In our study, at 80% of confidence we observed that 686 models are valid and 565 efficient. In parallel, we developed standard QSAR models for the same panel of proteins with more than half of the models being sufficiently robust to be used for activity prediction. Nevertheless, some proteins are excluded despite the fact that their equivalent conformal prediction model shows it is still possible to achieve predictions with acceptable confidence. We therefore conclude that conformal prediction is a valuable methodology when developing large-scale predictive models for target binding.

(1) Norinder, U.; Carlsson, L.; Boyer, S.; Eklund, M. Introducing Conformal Prediction in Predictive Modeling. A Transparent and Flexible Alternative to Applicability Domain Determination. *J. Chem. Inf. Model.* **2014**, 140521120138005.
(2) Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrián-Uhalte, E.; et al. The ChEMBL Database in 2017. *Nucleic Acids Res.* **2017**, *45* (D1), D945–D954.
(3) Illuminating the Druggable Genome - Protein Families Web Page. https://druggablegenome.net/ProteinFam