**Wednesday April 11th; 11:00-11:30**

# Predicting with Confidence:

# *In silico* model building and prediction using Conformal Prediction

**Ulf Norinder**

**Senior Research Specialist, Swetox, Sweden**

Predictions from a QSAR model are frequently confounded by a poor understanding of the reliability of the estimation for a compound of interest. What most users of *in silico* models would like to know is that a particular prediction is derived from an area of model property space where reliable estimations can be expected. Conformal Prediction (CP) represents a framework for current applications of machine-learning for *in silico* model building where the focus is on predictions with a pre-defined confidence level (fixed error rate).

The CP framework will be presented and applied to areas such as HTS screening and ADMET predictions. Examples related to classification as well as regression problems will be presented.

**11:30-12:00**

# Read-across as Structure Activity Relationship in Predictive Toxicology

**Nora Aptula**

Unilever Safety and Environmental Assurance Centre, Colworth Science Park, Sharnbrook, Bedford MK44 1LQ, United Kingdom

Read-across is gaining popularity as an alternative method to animal testing. The work over the past several years has revealed that, while it is conceptually simple, in practice it is difficult, especially for complex health endpoints. Read-across is the use of toxicity data from a tested chemical, to fill a data-gap for a "similar" untested chemical. The key step in this process is how to define "similar", as acceptance of the read-across outcome often hinges on the argument put forth to justify the similarity. Seminal features of a good read-across will be described and examples given.

**14:00-14:30**

# Application of conformal prediction in a more formal definition of applicability domain.

Sebastien Guesne, Lilia Fisk and Thierry Hanser
Lhasa Limited, Granary Wharf House, 2 Canal Wharf, Leeds, LS11 5PS, United Kingdom

Defining the applicability domain (AD) is the art of quantifying the uncertainty or confidence of predictions produced by classification and regression techniques in the field of machine learning. This quantification is of great importance when assessing the potential liability of a chemical to cause toxic/adverse effect(s) and promoting the use of *in silico* predictions. Hence it requires solutions so that the *in silico* model user is confident and has the possibility of setting up acceptable thresholds in accordance with one's use case.

Domain of applicability has many aspects that need to be considered. They include interpolation of the training set, the density and the quality of information of the training set around a query compound as well as the distance of the query compound to the decision boundary of an *in silico* model. Attempting to merge these aspects into a single metric is highly complex and may results in a "black box" framework where some aspects of AD may not be considered. We present a conceptual framework in which the AD is broken down into 3 steps: applicability, reliability and decidability domains, making the quantification of AD stepwise, intuitive and transparent.

A conformal predictor is an algorithmic framework which complements an underlying machine learning algorithm that allows the resulting system to produce predictions with information on their confidence. In the context of a classification problem such information includes *p*-values which delimit prediction regions where one of them contains a unique label. Intuitively the larger this region is the more a user can trust the conclusion of the prediction made by a conformal predictor and its underlying *in silico* model. This approach was used for the decidability domain of our framework of AD.

Firstly this framework will be introduced. Secondly this presentation will describe how conformal prediction can be projected onto the Lhasa Limited AD framework at the decidability level with a dataset of inhibitors of the bile salt export pump (BSEP).

**14:30-15:00**

# Exploring pharmacokinetic SARs early in drug discovery.

Robert D. Clark

Simulations Plus, Inc., Lancaster CA 93534 USA

Attrition due to lack of efficacy in clinical trials continues to be a problem in drug discovery and development. Piecemeal rules of thumb such as those that make up Lipinski's Rule of Five [1] have helped identify candidates that are liable to fail due to solubility and absorption problems, there are many combinations of properties that can also lead to development failure. The best way to identify these are the full-scale physiologically-based pharmacokinetic (PBPK) simulations available in programs like GastroPlus [2]. Such programs are designed primarily to analyze pre-clinical or clinical trial data on individual subjects, and so have to be configurable to take variations in subject sex, weight, and disease state into account. For optimal performance, they also need to know details about transporters and metabolism that are only rarely available early in a project. Such details are much less relevant in discovery and lead optimization, where the important thing is to anticipate generic problems with a class of chemistry and to differentiate such problems from ones that are specific to a particular molecular structure.

We have created a streamlined, high-throughput version of the GastroPlus gastrointestinal absorption model for the HTPK Simulation Module in ADMET Predictor 8.5 [2] that enables users to easily estimate percentage absorbed (%Fa) and percentage available after first-pass metabolism (%Fb) for thousands of compounds for a typical rat or human subject. Experimental values for solubility and *in vitro* metabolism can be used if available, but reasonable results are obtained from purely *in silico* property predictions in many cases. Moreover, experimental deviations from those predictions can provide important insight into complicating factors – e.g., unexpected precipitation, transporters or non-CYP metabolism – that are best addressed early in development.

[1] C.A. Lipinski. Lead- and drug-like compounds: the rule-of-five revolution. *Drug Discovery Today: Technologies* 1, 337-341.

[2] GastroPlus™ is distributed by Simulations Plus, Inc. (http://www.simulations-plus.com), as is ADMET Predictor™ and its HTPK Simulation Module.

**15:00-15:30**

# Exploring and Exploiting SAR in Lead Optimisation

**Stephen Pickett**, Chris Luscombe, David Marcus, Darren Green

Computational and Modelling Sciences, GlaxoSmithKline, Stevenage, Herts, UK

Small molecule drug discovery involves multi-disciplinary teams coming together to discover a molecule with the appropriate profile. It is often cast as a complex multi parameter optimisation problem with cycles of design, make, test. Within this context machine learning methods, experimental design and de novo structure generation have all found a place. However, they have tended to be used in a reactive manner to solve problems posed by the program team. In this presentation we will describe how data-driven chemoinformatics methods can automate much of what has historically been done by a medicinal chemist, leading to a radical rethinking of the design process. The implications of automation for the human-machine interface will be explored and illustrated with examples from Bradshaw, GSK's experimental automated design environment.

**16:00-16:30**

# Gaussian Processes: We demand rigorously defined areas of doubt and uncertainty

**Peter Hunt, Optibrium Ltd.**

A quantitative structure-activity relationship (QSAR) model is a mathematical function of molecular descriptors. The parameters of this function are found by maximizing the fit of this function to the observed activities of a training set of compounds, using a statistical or machine learning method. Following validation of the resulting model, most methods for estimation of the uncertainty in a prediction focus measures of the 'domain of applicability' or 'distance to model' to identify new compounds that differ significantly from the training set and hence for which the confidence in a prediction will be low.

In contrast to this, the Gaussian Processes method estimates a probability distribution over possible models that fit the observed data. The predicted value for a new compound is the mean of this distribution, while the standard deviation provides a well-defined estimate of the uncertainty for each individual prediction. This naturally takes into account the distance to the training set compounds and also identifies cases where variability in the training set data limits the ability to make a confident prediction, even if the new compound lies within the domain of applicability.

In this talk we will describe the Gaussian Processes method, discuss its strengths and weaknesses and compare its results with other QSAR modelling methods. This will be illustrated by several examples applications to different QSAR modelling problems.

**16:30-17:00**

# 3D-e-Chem: Structural Cheminformatics Workflows for Computer-Aided Drug Discovery

**Chris de Graaf, VU Amsterdam**

In the presentation I will discuss how the integration of structural interaction fingerprints in chemogenomics workflows can be used for the identification of function and protein-specific interaction hotspots to guide structure-based virtual screening and the construction of structure-based medicinal chemistry toolkits for ligand design. Computer-aided drug discovery workflows will be presented that combine structural cheminformatics tools and databases (https://3d-e-chem.github.io). I will show how developments in structural biology can be complemented by computational and experimental studies for a more accurate description and prediction of structural determinants of ligand binding kinetics and the identification and design of chemical modulators of protein function.

**Thursday April 12th; 09:00-09:30**

# Applying structural informatics approaches to pharmaceutical supply chain processes

Andrew G. P. Maloney[1], Mathew J. Bryant[1] and Ian Bruno[1]

[1]*The Cambridge Crystallographic Data Centre, 12 Union Road, Cambridge, CB2 1EZ, UK.*

E-mail: maloney@ccdc.cam.ac.uk

Pharmaceutical development currently stands poised to benefit from the big data revolution, but not equipped with a platform to reap the potential rewards. Central to this opportunity is the Cambridge Structural Database (CSD), the world's most comprehensive database of small molecule crystal structures.

The Advanced Digital Design of Pharmaceutical Therapeutics (ADDoPT) consortium is made up of a variety of pharmaceutical companies, small and medium enterprises and research organisations, each with vast repositories of drug product property and processing data. By linking this wealth of manufacturing data gained from across the ADDoPT consortium back to knowledge of molecular and crystal structures, an unprecedented perspective of drug product design can be obtained.

To this end, the CCDC is making use of the thousands of drug crystal structures in the CSD to establish the key links between structural features and potential formulation issues. By applying a solid-form informatics approach, the CCDC is directing the development and application of predictive tools that exploit our understanding of crystalline structures. These tools will then be applied to support the design of more robust manufacturing processes and the identification of the most appropriate formulation decisions.

Using the range of software available in the CSD Materials suite, crystal structures of drug molecules have been evaluated at the molecular, intermolecular and supramolecular level. Alongside the application of new bespoke analytical tools developed using the CSD Python API, a robust structural analysis protocol for the drug formulation scientist is emerging, which offers valuable insights into the potential downstream behaviour of drug candidates during formulation.

**09:30-10:00**

# Energy-Structure-Function maps for functional molecular crystals

Graeme Day, University of Southampton

**10:00-10:30**

# Towards A Digital Definition of Drug Product Design

Klimentina Pencheva, Robert Docherty, Tiffany Lai and Ernest Chow*

*Materials Science Drug Product Design*

*Pfizer Global Research & Development*
*Pharmaceutical R & D (i.p.c 612)*
*Sandwich, Kent*
*CT13 9NJ*

The selection of the solid form and particle attributes for development is a key milestone in the conversion of any new chemical entity into a drug product. An understanding of the materials science of a new active pharmaceutical is crucial at the interface of medicinal chemistry and pharmaceutical development.

The physical and chemical properties (e.g. solubility) of a new chemical entity that impact product performance and product robustness are strongly influenced by the solid state structure. Product performance can only be assured when the new chemical entity is delivered to the patient in a well-defined and understood chemically and physically stable solid form.

In recent times the transformation of molecules to medicines has been the subject of much interest especially the impact of digital technologies as epitomised in the Advanced Digital Design of Pharmaceutical Therapeutics (ADDoPT) initiative.

In this presentation we will attempt to integrate progress with cutting edge computational and QSPR tools in academia to the best current industrial practices so that medicinal chemists and pharmaceutical scientists can, through an unprecedented structural perspective, transform the journey from molecule to medicine.

**11:00-11:30**

# Machine-learning scoring functions to improve structure-based binding affinity prediction and virtual screening

Pedro Ballester, INSERM, Marseille, France

Docking tools to predict whether and how a small molecule binds to a target can be applied if a structural model of such target is available. The reliability of docking depends, however, on the accuracy of the adopted scoring function (SF). Despite intense research over the years, improving the accuracy of SFs for structure-based binding affinity prediction or virtual screening has proven to be a challenging task for any class of method. New SFs based on modern machine-learning regression models, which do not impose a predetermined functional form and thus are able to exploit effectively much larger amounts of experimental data, have recently been introduced. These machine-learning SFs have been shown to outperform a wide range of classical SFs at both binding affinity prediction and virtual screening. The emerging picture from these studies is that the classical approach of using linear regression with a small number of expert-selected structural features can be strongly improved by a machine-learning approach based on nonlinear regression allied with comprehensive data-driven feature selection. Furthermore, the performance of classical SFs does not grow with larger training datasets and hence this performance gap is expected to widen as more training data becomes available in the future.

**11:30-12:00**

# Chemical and Biological Data - from Compound Selection to Mode of Action Analysis (and Back Again)

Andreas Bender, Centre for Molecular Informatics, Department of Chemistry, University of Cambridge

More and more chemical and biological information is becoming available, both in public databases as well as in company repositories. However, how to make use of this information in chemical biology and drug discovery settings is much less clear. In this work, we will discuss how chemical and biological information from different domains – such as compound bioactivity data, pathway annotations from the bioinformatics domain, and gene expression data – can be used for a variety of purposes.

Examples related to *understanding* compound action include the mode-of-action analysis from phenotypic readouts,[1,2] and anticipating compound toxicities in early discovery and during lead optimization based on gene expression data[3]. Applications of *selecting* compounds with the desired activities include proteochemometrics modelling against a range of related protein targets such as enzymes in HIV[4] and against serine proteases[5] , using gene expression data to select compounds which modulate biological pathways used in cellular differentiation to generate cardiomyocytes,[6] and models for differential activity against particular cell lines[7].

More recent research in the groups includes the modelling compound combinations in the antibacterial context[8] as well as of cancer cell line screens[9], and learning from data to perform iterative screening[10], such as by utilizing conformal prediction methods.[11]

This presentation will go through some case studies selected from the above areas of our research.

1. Koutsoukas A, *et al. J. Proteomics* **2011**, 74, 2554 – 2574.
2. Drakakis G, *et al.* ACS Chem. Biol. **2017**, 12, 1593 – 1602.
3. Verbist B, *et al. Drug Discov. Today* **2015**, 20, 505 - 513.
4. Van Westen GJP, *et al. PLoS Comp. Biol.* **2013**, 9, e1002899.
5. Subramanian V., *et al. MedChemComm* **2017**, 8, 1037 – 1045.
6. KalantarMotamedi, Y, *et al. Cell Death Discovery* **2016**, 2, 16007.
7. Cortes-Ciriano I, *et al. Bioinformatics* **2016**, 32, 85 – 95.
8. Mason DJ, *et al. J. Med. Chem.* **2017**, 60, 3902 – 3912.
9. *under revision*
10. Paricharak S., *et al.* **2016**, 11, 1255 – 1264.
11. Svensson F, *et al. J. Chem. Inf Model.* **2017**,57, 439 – 444.

**12:00-12:30**

# Graph structured neural networks for machine learning with molecules

Alex Gaunt, Microsoft Research, Cambridge

**13:30-14:00**

# Probing conformational landscapes using swarm-enhanced sampling molecular dynamics

Irfan Alibay, Universities of Manchester and Oxford

Molecular dynamics (MD) is a useful tool in describing conformational changes at a level of atomistic detail not readily available via experimental methods. While the use of hardware accelerators has led to large improvements in MD simulation costs, exhaustively probing conformational changes which occur on multi-microsecond timescales remains a prohibitively time consuming task. Even for small molecule systems, simulations frequently require weeks to months of computation. A potential solution to this issue is the use of enhanced sampling MD methods. In this talk, we present a biased MD method which improves sampling efficiency by coupling the time evolution of a swarm of replica trajectories via a pairwise biasing potential acting on their conformational proximity in dihedral space. We evaluate the performance of this method, which we term multidimensional swarm enhanced sampling molecular dynamics (msesMD), in its ability to explore rare conformational events in explicitly solvated systems, comparing against other enhanced sampling methods and long unbiased MD simulations.

**14:00-14:30**

# Water networks in protein-ligand complexes using Grand Canonical ensemble methods

Hannah E. Bruce Macdonald[a,b], Chris Cave-Ayland[a], Marley Samways[a], *Richard D. Taylor[b], Jonathan W. Essex[a]*

[a)] School of Chemistry, University of Southampton, Highfield, Southampton S017 1BJ, UK

[b)] UCB, 216 Bath Road, Slough SL1 3WE, UK

Understanding the location and energetics of water molecules is helpful for rational drug design. We present our grand canonical Monte Carlo[1,2] (GCMC) method for predicting water locations and binding free energies in proteins.

The energetics of an active site water molecule can be vital to drug binding, through either stabilising the bound complex, or releasing entropy upon water displacement. Rationalising how water molecules should be treated is difficult – whether a water molecule should be retained in a bound structure, or if it should be displaced to recover entropy and allow for direct protein-ligand interactions, is unclear. GCMC is able to calculate the free energy of multiple water networks in a single simulation, and as the water location is predicted automatically as part of the simulation, no experimental knowledge of hydration site location is required. The binding free energies determined are rigorous and have been shown to be consistent with other gold-standard methods.

In this presentation, the GCMC method will be described and demonstrated on targets of significant pharmaceutical interest. The water locations calculated will be compared to good quality crystallographic data. The method has been optimised, both through reducing the number of chemical potentials needed to be simulated, as well as introducing replica exchange moves between neighbouring chemical potentials. The implications of the water networks observed in terms of ligand optimisation and binding will be discussed.

References:

[1] G. A. Ross, M. S. Bodnarchuk, J. W. Essex, *JACS,* (2015), **137**, 14930–14943

[2] G. A. Ross, H. E. Bruce Macdonald, C. Cave-Ayland, A. I. Cabedo-Martinez, J. W. Essex, *JCTC*, (2017), **13**, 6373–6381

**14:30-15:00**

# Beyond the Happy Water – recent advances in the quantification of contributions

Ben Tehan, Heptares Therapeutics

Water molecules and their networks are now recognized to have crucial functions for both proteins and ligands alike. Here we will be highlighting how water and its perturbation and displacement affects the mediation, strength and selectivity of ligand binding. Showing retrospectively where different techniques and levels of analyses have explained SAR which had previously confounded us, and prospectively how we use these techniques for drug discovery.