POSTER 1

QSARStudio: An enterprise platform for building and monitoring machine learning models

David Marcus, GSK

Continuously supporting drug discovery projects requires a modelling platform that can combine state of the art molecular representations, predictive models, and uncertainty quantifications. For this purpose, we have built an in-house machine learning platform called QSARstudio that supports our projects teams to apply local and global QSAR models for bioactivity, DMPK and toxicity data that drive our drug discovery efforts from early stage to lead optimisation and embedded in projects continuous DMTA cycles. the platform also supports our model drift and usage monitoring to flag and address ML degradation and domain drift in real time and suggest fast action to overcome model drift by model recalculation or data generation to overcome these issues.

POSTER 2

Synergistic application of QuanSA & physics-based simulation in affinity prediction

Peter Hunt, Optibrium

Binding affinity prediction continues to be a challenge for CADD, especially where no high-resolution crystal structure of the target protein or large sets of affinity data exist. Here we describe the application of QuanSATM in a typical lead optimisation scenario and compare the results to both an FEP+ strategy alone and the combination of QuanSA and FEP+ predictions[1]. QuanSA constructs physically sensible binding site models using multiple instance machine learning, ligand structure and affinity data. The QuanSA pocket field allows scoring of new molecules either from within the training set chemotypes or novel molecules enabling the technique to be used in a screening scenario. The average error for affinity predictions using QuanSA was 0.4 pKi units whilst for the FEP+ set the average error was marginally worse. However, due to partial error cancellation, the simple combination of the two methods produced predictions with an average error less than 0.3 pKi units and a noticeable reduction in large errors.

[1] Ann Cleves, Stephen Johnson, Ajay Jain, J. Chem. Inf. Model. (2021), 61, 5948−5966.

POSTER 3

Models of unbound volume of distribution, Vdss and fraction unbound

Mark Gardner, AMG Consultants Ltd.

Predictions of dose are important in understanding the potential of drug candidates. A single dose cure is a desirable characteristic of parasitic infectious diseases such as malaria and schistosomiasis. Two important parameters that feed into dose prediction are volume of distribution and fraction unbound which can be combined to unbound volume of distribution. This poster will consider experimental variation and modelling of these properties.

POSTER 4

Filling the gaps: Data Imputation Methods for Drug Discovery

Jiahao Yu, GSK & University of Bristol

Drug discovery datasets are often shown as sparse, noisy, and heterogeneous. To facilitate drug discovery projects and to ensure the effectiveness of Machine Learning (ML) algorithms and predictive models, it is necessary sometimes to find methods that can fill in the gaps in this data. In this poster, we discuss several classic and state-of-the-art data imputation methods and compare their performance with classic QSAR modelling. We found that data imputation models can usually outperform classic QSAR models, however some are unsuitable for data imputation in drug discovery, and some will require extensive calculation time.

POSTER 5

How do autoencoders help explore the conformational space of MD simulations of cyclic peptides?

***L. M. Windeln**, C. Holdship, J. Frey, J.W. Essex, University of Southampton*

α-Conotoxins are a class of disulfide-rich cyclic peptides produced by marine cone snails that target human nicotinic acetylcholine receptors (nAChRs). The specificity of these toxins against different isoforms of nAChRs make them attractive pharmacophore candidates. To elucidate the mechanism of action, the solution conformations of the conotoxins must be determined before modelling their interactions with the receptor. We have carried out molecular dynamics (MD) simulations of five conotoxins using enhanced sampling methods. In MD simulations, large datasets with high dimensionality (many variables) are generated. These variables are the cartesian coordinates of each atom for each time-step of the simulation. From this, other variables such as the backbone torsion angles can be derived. To extract meaningful movement of the system over time and thus obtain stable solution structures, the dimensionality of these data has to be reduced significantly. A popular method for dimensionality reduction is principal component analysis (PCA) which uses a linear combination of all input variables to calculate the orthogonal collective variables which maximally capture the data covariance. In the case of our conotoxin simulations, the first two principal components of the PCA captured less than 50 % of the variance. Here we compare the use of PCA to analyse the conformational space of conotoxins to using autoencoders for the same task. Standard autoencoders have a symmetric neural network architecture with a bottleneck in the centre (encoder layer) whilst the network is trained to reduce the error between input and output (decoder) layer. We were able to perform reduction of the simulation data to two dimensions (2 encoder nodes) and extract diverse conformations with better separation than for the PCA workflow. Furthermore, we achieved high model accuracy (defined by low reproduction error) throughout the trajectory for autoencoders, which was not the case for PCA. Going forward autoencoders will enable us to more accurately model interactions of conotoxins and their targets. Our findings in using autoencoders for analysing simulations of cyclic peptides will allow us to expand their use in structural biology and help identify more biologically relevant conformations of other cyclic peptides.

POSTER 6

Database AutoPH4: Pharmacophore Analysis of Multiple Protein Structures

Andrew Henry, Chemical Computing Group

Drug discovery projects often involve determination, assessment and analysis of multiple protein (“apo”) or protein-ligand (“holo”) structures – these may have been produced by X-ray crystallography or cryo-EM measurements, or may be conformational ensembles from MD simulations. Such studies can reveal structural factors which can be used to find new potential binding sites, binders, or advance a medicinal chemistry campaign in some other way. Here, we present extensions and further developments of the AutoPH4 method(1) in the Molecular Operating Environment (MOE) software system which, in conjunction with the application of MOE’s Site Finder and other tools, enable the analysis of such multiple structures to produce consensus pharmacophores, binding site information and classification and new ways to assess docking results. The utility of these methods is illustrated by an analysis of a database of Abl kinase structures.

(1). S. Jiang, M. Feher, C. Williams, B. Cole, D.E. Shaw; *J. Chem. Inf. Model*. 60 (2020) 4326–4338.

POSTER 7

Handling Ultra-large Chemical Spaces in Structure-Based Drug Design

__Noel O'Boyle__, Jon Tyzack, Daniel Santos-Stone, Chris de Graaf: Sosei Heptares

Chemical space is "vastly, hugely, mindbogglingly big" [1] with some estimates suggesting that there are 1060 drug-like molecules. This is both a cause for optimism ("the right molecule must surely exist!") and a potential challenge ("how will I ever find the right molecule?"). Within the context of Structure-Based Drug Design the challenge is to find a match between the chemical space of small molecules and the conformational flexibility of orthosteric and allosteric protein binding sites that can be targeted by small molecules [2-4].

Here we explore the challenges and opportunities of navigating ultra-large chemical spaces to target diverse protein binding sites in the context of GPCR Structure-Based Drug Design (SBDD) at Sosei Heptares. Combinatorial ultra-large virtual libraries such as Enamine REAL are growing exponentially and present a particular challenge to established techniques for virtual screening such as protein-ligand docking [5,6]. Recently, approaches based around the components of the combinatorial library ('synthons') have been developed [7], as well as machine-learning approaches [for example, 8].

We describe the derivation of synthons for Enamine REAL via Mol2Synthon and their use to find entries in Enamine REAL with high docking scores via three distinct approaches: SynthonConnect, Gabby and generative design [9-10].

References:

1. Adams DN. The Hitchhiker's Guide to the Galaxy. Pan Books, 1979.

2. Vass M. Chemical Diversity in the G Protein-Coupled Receptor Superfamily. Trends Pharmacol Sci. 2018, 39, 494.

3. Congreve M. Applying Structure-Based Drug Design Approaches to Allosteric Modulators of GPCRs. Trends Pharmacol Sci. 2017, 38, 837.

4. Congreve M, de Graaf C, Swain NA, Tate CG. Impact of GPCR Structures on Drug Discovery. Cell. 2020, 181, 81.

5. Ballante F. Structure-Based Virtual Screening for Ligands of G Protein-Coupled Receptors: What Can Molecular Docking Do for You? Pharmacol Rev. 2021, 73, 527.

6. Bender BJ. A practical guide to large-scale docking. Nat Protoc. 2021, 16, 4799.

7. Sadybekov AA. Synthon-based ligand discovery in virtual libraries of over 11 billion compounds. Nature. 2022, 601, 452.

8. Gentile Fl. ACS Cent Sci. 2020, 6, 939.

9. Thomas M. Comparison of structure- and ligand-based scoring functions for deep generative models: a GPCR case study. J Cheminform. 2021, 13, 39.

10. Thomas M. Augmented Hill-Climb increases reinforcement learning efficiency for language-based de novo molecule generation. J Cheminform. 2022, 14, 68

POSTER 8

Towards early prediction of human pharmacokinetics using AI approaches

***Olga Obrezanova*[1], *Anton Martinsson*[2], *Filip Miljković*[2], *Beth Williamson*[3], *Susanne Winiwarter*[4], *Martin Johnson*[5], *Andy Sykes*[5], *Graham Smith*[1], *Andreas Bender*[1], *Nigel Greene*[6]**

1 Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Cambridge, UK

2 Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Gothenburg, Sweden

3 Drug Metabolism and Pharmacokinetics, Research and Early Development, Oncology R&D, AstraZeneca, Cambridge, UK

4 Drug Metabolism and Pharmacokinetics, Research and Early Development, Cardiovascular, Renal and Metabolism (CVRM), Biopharmaceutical R&D, AstraZeneca, Gothenburg, Sweden

5 Clinical Pharmacology & Quantitative Pharmacology, Clinical Pharmacology & Safety Sciences, R&D AstraZeneca, Cambridge, UK

6 Imaging and Data Analytics, Clinical Pharmacology & Safety Sciences, R&D, AstraZeneca, Waltham, USA

Drug metabolism and pharmacokinetic data are routinely used in drug discovery to understand absorption, disposition, metabolism, and elimination (ADME) of candidate drugs in animal species and human. A multitude of in vitro assays, as well as in silico models, have been developed for this purpose and are usually followed up with in vivo PK experiments in preclinical species. AI-driven approaches to predict rat in vivo and human PK parameters from chemical structure can guide the design of molecules with optimal PK profiles, enable the prediction of virtual compounds, and help to prioritise compounds for in vivo assays.

We present a machine learning model for prediction of rat in vivo PK parameters, including rat oral bioavailability and clearance, which utilises the molecular chemical structure and either measured or predicted in vitro ADME parameters [1]. The model was trained on in vivo rat PK data for 4,000 diverse compounds from multiple therapeutic areas and employed a state-of-the-art AI approach based on multi-task graph convolutional neural networks.

Furthermore, we present machine learning models for prediction of human in vivo PK parameters, including peroral Cmax and intravenous volume of distribution, using chemical structural information and dose as inputs [2]. The models are based on PharmaPendium human PK data for a set of 1,000 compounds and use chemical descriptors as well as in silico ADME and rat PK predictions.

The developed AI approaches are leading towards an accurate prediction of human PK early in drug discovery and are a valuable addition to AI-assisted drug design and prioritisation tools, with the aim to increase efficiency of Design-Make-Test-Analyse cycles, to reduce the number of animal PK experiments and to lower compound attrition.

References

1. O. Obrezanova et al. Prediction of in vivo pharmacokinetic parameters and time-exposure curves in rats using machine learning from chemical structure. Mol. Pharmaceutics 2022, 19, 1488-1504. https://pubs.acs.org/doi/pdf/10.1021/acs.molpharmaceut.2c00027

2. F. Miljković et al. Machine Learning Models for Human In Vivo Pharmacokinetic Parameters with In-House Validation. Mol. Pharmaceutics 2021, 18, 4520–4530. https://doi.org/10.1021/acs.molpharmaceut.1c00718

POSTER 9

3D-QSAR meets ML for Binding Affinity Prediction

***Mireille Krier*; Katarína Stančiaková; Lukas Eberlein; Gunther Stahl; Jingyi Chen; Ali Mozaffari; Shyamal Nath: OpenEye**

We report the development of a 3D-QSAR methodology for predicting binding affinity using ROCS and EON based similarities as descriptors in combination with two distinct machine learning approaches. Our results demonstrate that our models perform comparably to or even surpass the performance of other methods, including CoMFA and the latest techniques in the literature. Moreover, our models provide prediction confidence, ensuring the reliability of the predictions. In addition, our developed model offers valuable design insights for lead optimization.

POSTER 10

The new SureChEMBL in a nutshell

***Nicolas Bosc*, Tevfik Kiziloren, Ricardo Arcila, Eloy Felix, Barbara Zdrazil, Andrew R. Leach**

EMBL-EBI, Wellcome Genome Campus, Hinxton, Cambridge, CB10 1SD

The SureChEMBL database (www.surechembl.org) gathers and makes easily accessible more than 160 million patents from worldwide authorities, for free (1). Having recently celebrated its 10th anniversary since it was acquired by EMBL-EBI, SureChEMBL continues with its main objective of extracting and delivering compound structures extracted from the document text and images, on a daily basis. As a result, 26.5 million compounds have been extracted from 28 million patents. In parallel, using a language model retrained on a corpus of patents, the documents are annotated for genes/proteins, diseases and mode of action. SureChEMBL offers different ways of accessing the leveraged data with a modern UI and several download formats.

Recently the new SureChEMBL was introduced. This represents the deepest change in a decade with a complete refoundation of the system offering a revamped interface and a container-based architecture. This results in system stability and response time improvements. More importantly for us, it is less difficult to maintain and makes the development and delivery of new functionalities much easier. Guided by recent user survey results, we are actively working on the development of further improvements focused on improving the data quality, coverage and accessibility. Efforts will also be given to push further the implementation of the SureChEMBL public API.

The poster will focus on the recent SureChEMBL changes, provided extra details on the architecture and the patent language model. We will also discuss the current development.

(1) George Papadatos, Mark Davies, Nathan Dedman, Jon Chambers, Anna Gaulton, James Siddle, Richard Koks, Sean A. Irvine, Joe Pettersson, Nicka Goncharoff, Anne Hersey, John P. Overington, SureChEMBL: a large-scale, chemically annotated patent document database, Nucleic Acids Research, Volume 44, Issue D1, 4 January 2016, Pages D1220-D1228, https://doi.org/10.1093/nar/gkv1253

POSTER 11

Robust Automated Equilibration Detection for Molecular Simulation

Finlay Clark[a]***, Graeme Robb***[b]***, Daniel J. Cole***[c]***, Julien Michel***[a]

a EaStCHEM School of Chemistry, University of Edinburgh, David Brewster Road, Edinburgh EH9 3FJ, United Kingdom.

b Oncology R&D, AstraZeneca, Cambridge CB4 0WG, United Kingdom.

c School of Natural and Environmental Sciences, Newcastle University, Newcastle upon Tyne NE1 7RU, United Kingdom.

Molecular simulations are widely used to inform structure-based drug design. They form the foundation of techniques such as alchemical binding free energy calculations, which are routinely used to prioritise candidate compounds for synthesis and testing.1 However, quantities calculated from molecular simulations, such as binding affinities, are often subject to an initial bias due to unrepresentative configurations of the input system. This systematic error can be reduced by truncating data from the start of the simulation and calculating the quantity of interest on the remaining "equilibrated" data. However, discarding too much data unnecessarily increases random error. It is common practice to select a fixed truncation point, which can introduce additional error into the calculated quantities, or to select by visually inspecting the data, which is only feasible for a few simulations.

In this work, we investigate automated approaches to select the optimal truncation point. Firstly, we create a synthetic dataset modelled on the output from a long absolute binding free energy calculation. This allows us to obtain robust statistics while avoiding the substantial cost of molecular dynamics simulations. Subsequently, we use this dataset to evaluate the performance of a range of new and reported methods for equilibration detection from within and outside the field of molecular simulation. We illustrate that methods found to be optimal on non-molecular simulations perform poorly here due to the long autocorrelation times often associated with molecular simulation.2 Additionally, we show that Chodera's method,3 which was designed for molecular simulation, is susceptible to discarding an excessive fraction of the data. In conclusion, we recommend the adoption of the best-performing method identified through our analysis. This method reduces bias more effectively than traditional non-molecular simulation approaches and introduces less variance than Chodera's method. It also offers a substantial speed advantage over the latter.

(1) Schindler, C. E. M. et al. Large-Scale Assessment of Binding Free Energy Calculations in Active Drug Discovery Projects. J. Chem. Inf. Model. 2020, 60 (11), 5457–5474. https://doi.org/10.1021/acs.jcim.0c00900.

(2) Hoad, K.; Robinson, S.; Davies, R. Automating Warm-up Length Estimation. J. Oper. Res. Soc. 2010, 61 (9), 1389–1403. https://doi.org/10.1057/jors.2009.87.

(3) Chodera, J. D. A Simple Method for Automated Equilibration Detection in Molecular Simulations. J. Chem. Theory Comput. 2016, 12 (4), 1799–1805. https://doi.org/10.1021/acs.jctc.5b00784.

POSTER 12

The Application of Machine Learning to Predict and Solve NMR Spectra in Structure Elucidation

Ben Honoré*, *Calvin Yiu, Craig Butts: University of Bristol

At Bristol we have developed IMPRESSION [1], a machine learning tool for predicting solution state chemical shifts and coupling constants of 3D molecular structures. The gold standard of IMPRESSION is a multitask graph transformer network that predicts 1H and 13C chemical shifts to within 0.1 and 1ppm, respectively, of quantum mechanical level (DFT) calculations. Predicted NMR parameters provide an important comparison when analysing NMR spectra and IMPRESSION takes a matter of seconds, rather than the hours/days that DFT can take to achieve the same result.

The inverse problem of generating the correct molecular structure from a set of experimental NMR chemical shifts is much more challenging but would be ground-breaking in synthetic chemistry. We have taken an approach that makes best use of IMPRESSION and our ability to rapidly predict chemical shifts. We obtain large amounts of data by evolving valency-valid structures of various molecules and measuring the closeness of the IMPRESSION-predicted chemical shifts with the known true chemical shifts at each iteration.

A reinforcement learning agent is trained on many instances of changing a molecular structure by addition or removal of chemical bonds and how this change affects the agreement between the predicted chemical shifts and the ground truth experimental data. Over training time, this forms a model capable of generating chemical structures to fit with a given set of experimental NMR data. There are challenges associated with quality of experimental data and chemical diversity of the training set, but this reinforcement learning workflow is highly promising in the move towards automated structure elucidation.

[1] W. Gerrard, L. A. Bratholm, M. Packer, A. J. Mulholland, D. R. Glowacki and C. P. Butts, Chem Sci, (2020), 11, 508-515.

POSTER 13

Phosphatidylcholine-specific phospholipase C as a promising drug target

Chatchakorn Eurtivong[1], Euphemia Leung[2], Ivanhoe K. H. Leung[3] and Jóhannes Reynisson[4]

1 Department of Pharmaceutical Chemistry, Faculty of Pharmacy, Mahidol University, Bangkok 10400, Thailand

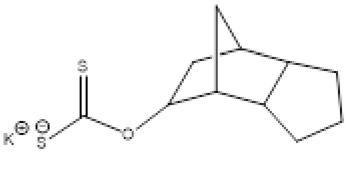
2 Auckland Cancer Society Research Centre, The University of Auckland, Auckland 1142, New Zealand

3 School of Chemistry, The University of Melbourne, VIC 3052, Australia

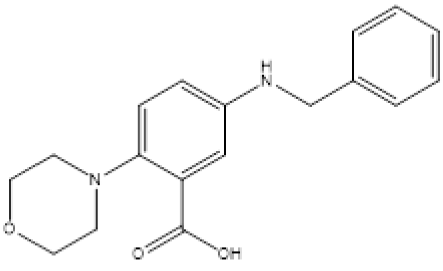
4 School of Pharmacy and Bioengineering, Keele University, Newcastle-under-Lyme, ST5 5BG, United Kingdom

Phosphatidylcholine-specific phospholipase C (PC-PLC) is an enzyme that catalyses the formation of the important secondary messengers phosphocholine and diacylglycerol (DAG) from phosphatidylcholine. Although PC-PLC has been linked to the progression of many pathological conditions including cancer, atherosclerosis, inflammation and neuronal cell death, studies on PC-PLC are relatively scarce. To date, the human gene expressing PC-PLC has not yet been found, and the only protein structure is from Bacillus cereus (PC-PLCBc). Nonetheless, there is evidence for PC-PLC activity has a human functional equivalent.1

A handful of inhibitors have been developed against PC-PLC; however, none of them are drug-like, e.g., tricyclodecan-9- yl-xanthogenate (D609). To find drug-like ligands a virtual screen was conducted using the PC-PLCBc crystal structure, in conjunction with the Amplex-Red biochemical activity assay and an intrinsic fluorescence binding assay. The most potent series identified was morpholinobenzoic acid (1).2



D609



Morpholinobenzoic acid (1)

The SAR profile was established by exchanging the morpholine with piperidine and piperazine derivatives; the carboxyl acid was removed as well as substituted for its methyl ester counterpart. Also, the substitution pattern on the right-hand-site phenyl group was explored. The results were that the morpholine ring is optimal, the removal of the carboxylic acid is beneficial, and finally, the ortho position, with a small electron withdrawing group is optimal.3 Based on molecular modelling then carboxylic acid chelates with the zinc ions in the binding pocket, therefore a hydroxylamine group was introduced, an excellent chelator, but with no increase activity.4 An MALDI-TOF protocol was developed to strengthen the SAR, which results revealed that the carboxylic acid, its methyl ester and the hydroxylamine give similar results.5 In conclusion, drug-like lead series has been developed for PC-PLC.

1. Eurtivong et al. Molecules, 2023, 28, 5637.

2. Eurtivong et al. Eur. J. Med. Chem., 2020, 187, 111919.

3. Pilkington et al. Eur. J. Med. Chem., 2020, 191, 112162.

4. Rees, et al. Bioorg. Chem., 2021, 114, 105152.

5. Sharma, et al. Anal. Meth., 2021, 13, 491.

POSTER 14

QITB: An interactive open-source web app for cheminformatics

Syed Zayyan Masud

The field of cheminformatics has been active under various guises for almost half a century, with brilliant innovation over the decade. These methods aid the process of drug discovery simultaneously accelerating and reducing the cost. However, the barrier to entry to use these diverse techniques remains in the form of programming know-how. Often it is a collaborative process across multiple disciplines including theoretical and experimental scientists to integrate cheminformatics tasks into a drug discovery pipeline. Various efforts have been made to mitigate this issue through comprehensive User Interfaces, but there is a lot left to be desired. There are issues like cross-platform compatibility; and the availability of resources to run such software or web servers. There are also questions of data privacy especially when sending user data on third-party servers. Therefore, we introduce a static web app, QSAR In The Browser (QITB), that runs all cheminformatics functions solely on the consumer's device, with no external server attached and, on any device, capable of running any modern web browser. It includes tools to fetch data from external services like ChEMBL or load one's data; data pre-processing; interactive chemical space visualisation and QSAR modelling. The code for this data is Open Source and the web app itself is hosted through GitHub Pages. In this work, the architecture, design decisions and implementation details of QITB are discussed and how it can facilitate the application of cheminformatics without any prior coding knowledge.

POSTER 15

The Danish (Q)SAR Database and modelling software

***Nikolai G. Nikolov**, Eva B. Wedebye*

Technical University of Denmark, National Food Institute

Henrik Dams Allé, Building 202, 2800 Kgs Lyngby, Denmark

The Danish (Q)SAR database (https://qsar.food.dtu.dk) is a free online searchable repository of structural information and (Q)SAR predictions, comprising a total of over 150 million data points. The predicted properties from over 200 (Q)SAR models, developed in-house or obtained from external sources, cover many physical-chemical, environmental fate, bioaccumulation, eco-toxicity, absorption, metabolism, molecular and toxicity endpoints, and are instantly available for each of the database's more than 650,000 substances, including >80,000 REACH substances, as a downloadable chemical profile. For single substances, the database integrates many predictions for weight-of-evidence assessment, contributing to reducing overall uncertainty. In addition, the database can be screened across all contained predictions/substances, for prioritization purposes or sorted by structural similarity to a target to find structural analogs etc. The database website provides an extensive search system including search by predictions, structure, similarity and experimental data from training sets, and logical combinations of searches. All in-house QSAR models are documented in QMRFs (QSAR Model Reporting Format).

The database is developed by the QSAR team at the DTU Food Institute, with financial support from the Danish EPA and Nordic Council of Ministers, and is furthermore supported by the European Chemicals Agency. Since the publication of the database in 2015, it has been used by more than 10,000 unique IP addresses worldwide running more than 200,000 searches and requesting download of over 100,000 (Q)SAR profiles.

For a number of categorical endpoints we developed separate models in three software systems, using different descriptors, variable selection methods and modelling algorithms. With the battery approach, it is in many cases possible to reduce "noise" from the individual model estimates and thereby improve accuracy and/or broaden the applicability domain.

In addition to using commercially available and open-source software for QSAR modelling, we are now finishing the development of a first version of an in-house QSAR modeling software, integrating descriptor calculation, descriptor selection, model development, validation and application. A short description of the software and the algorithms, as well as model performance results for 20 diverse endpoints from the Danish (Q)SAR Database will be presented.

POSTER 16

Computational study of water network influencing potency of BCL6 inhibitors

Daniella Hares, ICR

Water networks can have a critical role in small molecules binding to a target protein and are an important consideration in structure-based drug design. When modifying the molecular structure of the lead compound, the rearrangement of water networks in the binding site can impact potency but this contribution is impossible to measure experimentally. Therefore, computational methods can be used to study the interplay between ligand optimisation and water displacement, by predicting the effect of structural changes on both the activity of the compound and the stability of neighbouring water molecules.

We used Grand Canonical Monte Carlo simulations and free energy calculations to rationalise the trends in potency observed in a previous project at the Institute of Cancer Research on B-cell Lymphoma 6 (BCL6) inhibitors. As part of the drug design process, the inhibitor structure was modified to displace water molecules that were part of a network within the protein binding site.

Using the BCL6 project as an example, this poster will demonstrate how computational approaches can shine a light into an aspect that is often overlooked but essential to guide the design of better compounds when working with solvent-filled protein pockets. We hope to show the power of these methods and encourage their use more widely, particularly in prospective applications on drug discovery projects.

POSTER 17

Prioritization of new molecule design using QSAR models: 2D- and 3D-QSAR studies on SARS-CoV-2 M^{pro} inhibitors

Oliver Hills, Natercia Braz

Cresset, New Cambridge House, Bassingbourn Road, Litlington, Cambridgeshire, SG8 0SS, UK.

The viral main protease M^{pro} is a crucial enzyme for the replication of the severe acute respiratory syndrome coronavirus 2 (SARS-CoV-2), the cause of the COVID-19 global pandemic. In addition to the established vaccination programs against COVID-19, antiviral drugs are seen as essential to control the inevitable future epidemics of coronaviruses. Because of its key role, M^{pro} has received much attention as a potential target for novel therapeutic agents.

Robust and predictive Quantitative Structure Activity Relationships (QSAR) models of activity against the M^{pro} of SARS-CoV-2 were developed to elucidate observed activity and inform new antiviral molecule design. Using a dataset of 76 compounds with known experimental activity and a common binding mode, machine learning (ML) and Field-QSAR models were constructed, within 2D and 3D descriptor space, establishing an ensemble of accurate and predictive QSAR models for new antiviral bioactivity prediction. In addition, interrogation of the Field-QSAR electrostatic and steric coefficients assisted in rationalizing inhibitor potency, highlighting molecular functionality, located in critical regions about the molecular frame, crucial for favourable activity against M^{pro}

POSTER 18

An MD-based workflow for predicting relative affinities of series of congeneric ligands

Adriana Coricello,[a] ***Maria Musgaard,***[b] ***Benjamin G. Tehan,***[b] ***Giovanni Bottegoni***[a,c]

a Dipartimento di Scienze Biomolecolari, Università degli Studi di Urbino "Carlo Bo", 61029 Urbino, Italy

b OMass Therapeutics Ltd, Building 4000, John Smith Dr, Oxford Business Park, ARC, Oxford OX4 2GX, UK

c Institute of Clinical Sciences, University of Birmingham, Edgbaston, B15 2TT, Birmingham, United Kingdom

Herein, we present an Adiabatic Bias Molecular Dynamics (ABMD)-based1,2,3 workflow for predicting relative affinities of a series of thirteen congeneric compounds against Thrombin.4 The key assumption here was that, for congeneric compounds, the key contribution to binding affinity comes from koff. Starting from the protein-ligand complex, we applied the harmonic bias to the distance between the center of mass of the ligand and that of the binding site residues. For each ligand-protein system, after equilibration, ABMD simulations in twelve separate replicas were performed. Each ABMD simulation lasted until the unbinding event was observed or until a ceiling simulation time of 500 ns was reached. The calculated mean times of unbinding of the ligands against their experimental pKi displayed a promising correlation.

While encouraging, these results came at a remarkable computational cost. We hence devised a modified protocol to speed up the calculations without compromising the reliability of the method. The new protocol consisted of running an initial set of twenty separate replicas of ABMD in parallel for 30 ns. The resulting trajectories were subsequently evaluated and, if unbinding was not observed, the replica which advanced the most along the CV was selected to start five new independent 10 ns-long replicas with randomized initial velocities. This last step was iterated until unbinding was observed. The new method still provided a good correlation between the time of unbinding and the pKi of the simulated ligands, with a negligible loss of accuracy and a significant gain in terms of efficiency.

While preliminary and limited to a single system, these results are encouraging. Simulations on other systems of pharmaceutical interest are currently on-going.

1. Marchi, M. and Ballone, P., J. Chem. Phys. 1999, 110, 3697-3702.

2. Bortolato, A., et al. JCIM 2015, 55, 1857-1866.

3. Gobbo, D., et al. JCTC 2019, 15, 4646-4659.

4. Baum, B., et al. J. Mol. Biol. 2009, 390, 56-69

POSTER 19

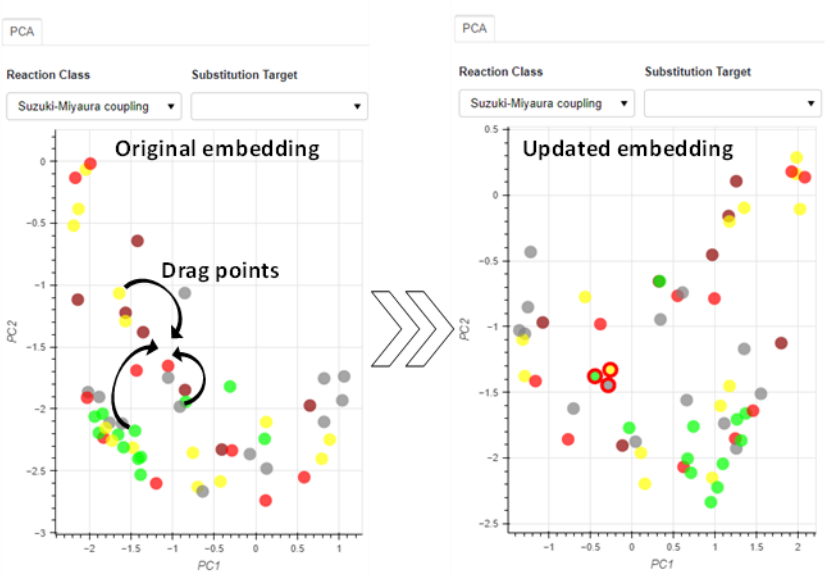
Solvent Surfer: interactive PCA for solvent selection

Joe Heeley, Samuel Boobier, Thomas Gärtner, Jonathan D. Hirst

School of Chemistry, University Park, Nottingham NG7 2RD

Unsustainable chemical processes can have devastating impacts on safety, health, and the environment. To encourage making sustainable choices at the laboratory level we have developed AI4Green,1 an open-source electronic laboratory notebook (ELN) that utilises cutting-edge machine learning methods to augment greener decision-making during reaction design.

The Solvent Surfer is a feature of AI4Green that can be used to identify greener alternatives to common laboratory solvents, an important challenge for sustainable chemistry.2 Interactive knowledge-based kernel PCA3 allows users to dynamically explore many 2D projections of solvents by defining control points that shape the embedding. This is done by directly interacting with the visualisation; dragging similar data points towards one another automatically updates the embedding and relocates the remaining points in a mathematically rigorous fashion (Figure 1). Our semi-supervised embedding technique exploits the influence of the control points in order to enable the user to shape and steer a live-updating embedding.



Reaction Class: Suzuki-Miyaura coupling

Substitution Target:

Figure 1: The Solvent Surfer's interactive knowledge-based kernel PCA as part of AI4Green.

The Solvent Surfer integrates this knowledge-based PCA into AI4Green, allowing chemists to explore various projections of solvent space during reaction design. Experimental data can be used to cluster control points and tailor embeddings for specific use cases: guiding solvent screening, identifying high-performance solvents, and highlighting sustainable substitutions. Imparting such expertise into the embeddings can provide non-intuitive insights and potentially accelerate the development of greener processes at the laboratory level.

References

[1] Boobier, S., Davies, J. C., Derbenev, I. N., Handley, C. M., Hirst, J. D. J Chem Inf Model, 2023, 63, 2895–2901.

[2] Byrne, F. P., Jin, S., Paggiola, G., Petchey, T. H. M., Clark, J. H., Farmer, T. J., Hunt, A. J., McElroy, R. C., Sherwood, J. Sustainable Chemical Processes, 2016, 4, 7.

[3] Oglic, D., Paurat, D., Gärtner, T. In Machine Learning and Knowledge Discovery in Databases, Calders, T., Esposito, F., Hüllermeier, E., Meo, R., Eds. Springer Berlin Heidelberg: Berlin, Heidelberg, 2014, pp 501–516.

POSTER 20

Machine Learning Driven Development of a Coarse-Grained Water Model Using ForceBalance

Jack Sawdon, University of Southampton

Without a doubt, water is one of the most important molecules in biology and chemistry, with computational studies being routinely used to model water and aqueous solutions. Water has a unique set of properties, such as a density maximum at 277 K and compression increasing diffusivity, which highlight its more complex dynamics compared to other solvents. Owing to the unique properties and importance of water, the development of accurate water models is an important challenge in molecular dynamics. Water models commonly used in molecular dynamics simulation (such as TIP3P and SPC) leave room for improvement in accurately reproducing water thermodynamic properties. ForceBalance, an iterative automated machine learning approach to forcefield parameterisation, has been shown to improve atomistic water models and likely achieves the accuracy limit for the 3-site functional form. Coarse-grained molecular dynamics offers opportunities for researchers to investigate larger systems over longer periods of time, however, this is often at the cost of accuracy. Indeed, coarse-grained water models are commonly less accurate compared to their atomistic counterparts. Here we use ForceBalance to generate a novel coarse-grained water model, BMW-FB. BMW-FB achieves atomistic model-like accuracy of water properties while retaining a high degree of coarse-graining, and a simple functional form, allowing for efficient implementation into molecular dynamics code.

POSTER 21

Predicting Bioactivity by Traversing Knowledge Graphs

Terence Egbelo[1], Vlad Sykora[2], Michael Bodkin[2], Zeyneb Kurt[1], Val Gillet[1]

[1]Information School, University of Sheffield, The Wave, 2 Whitham Rd, Sheffield S10 2AH, United Kingdom

[2]Insilico R&D, Evotec (UK) Ltd, 114 Park Drive, Abingdon OX14 4RZ, United Kingdom

A biomedical knowledge graph is a heterogeneous information network integrating the relationships between entities such as genes, proteins, compounds and diseases. A variety of properties relevant to drug discovery are encoded as direct links between entities, for example chemical similarities between pairs of druglike compounds. These properties may correlate with more complex patterns within the graph such as the tendency of similar compounds to have similar biological effects.

This poster summarises a study that tackles the prediction of compound bioactivity in protein assays as a knowledge graph completion problem. Here, knowledge graph completion is a classification task where an Active or Inactive label is to be assigned to each prospective compound-assay pairing. The objective is thus to learn classifier models that can separate Compound-*Active*-Assay knowledge graph "triples" from Compound-*Inactive*-Assay triples and generalise to correctly infer activity and inactivity in unseen compound-assay pairings.

Inspired by previous research (Lao et al 2011, Fu et al 2016, Himmelstein et al 2017), the approach used in this study leverages observable knowledge graph topological properties to tackle the knowledge graph completion problem. Each Compound-*Active*(*Inactive*)-Assay triple in the data set is characterised by a feature vector whose elements are the counts of a set of distinct path types ("metapaths") in the sample KG that connect the compound and the assay. The main connecting edge types for these metapaths are compound similarity edges, based on Chemoinformatics descriptors of molecular structure and shape. By traversing the resulting graph representation of chemical space, paths predictive of bioactivity may be found between compounds and assays.

The data set used is an extract of Evotec's comprehensive proprietary biomedical knowledge graph created by integrating public data sources from the areas of proteomics, chemistry and pharmacology. The extract graph draws together experimental data from a set of predominantly binding and functional assays from ChEMBL (Gaulton et al 2017) describing diverse protein targets.

In this poster, we describe the process of generating feature vectors from the knowledge graph that form the inputs for the training, validation and testing of Random Forest models. We then:

- Ensure activity class balance in individual assays to prevent model-distorting biases previously traced to imbalanced individual assay data
- Assess the effect of different similarity edge creation thresholds on the predictive power of compound similarity edges in the knowledge graph
- Benchmark the performance of metapath-based activity prediction classifiers against molecular fingerprint-based QSAR classifiers using the same training and test instances, showing that the metapath-based approach is well-suited to dealing with sparse data

In multi- and single-assay activity classification alike (including single assays with fewer than 20 training compounds), metapath-based models that exploit chemical similarity-based paths and paths expressing the query compound's prior experimental record outperform fingerprint-based QSAR classifiers. Classifiers trained on a combination of metapath features and fingerprints see additional performance gains.

References

1. Lao, N., Mitchell, T., & Cohen, W. (2011, July). Random walk inference and learning in a large scale knowledge base. In Proceedings of the 2011 conference on empirical methods in natural language processing (pp. 529-539).
2. Fu, G., Ding, Y., Seal, A., Chen, B., Sun, Y., & Bolton, E. (2016). Predicting drug target interactions using meta-path-based semantic network analysis. BMC bioinformatics, 17(1), 1-10.
3. Himmelstein, D. S., Lizee, A., Hessler, C., Brueggeman, L., Chen, S. L., Hadley, D., ... & Baranzini, S. E. (2017). Systematic integration of biomedical knowledge prioritizes drugs for repurposing. Elife, 6, e26726.
4. Gaulton, A., Hersey, A., Nowotka, M., Bento, A. P., Chambers, J., Mendez, D., ... & Leach, A. R. (2017). The ChEMBL database in 2017. *Nucleic acids research*, *45*(D1), D945-D954.

5. Galárraga, L. A., Teflioudi, C., Hose, K., & Suchanek, F. (2013, May). AMIE: association rule mining under incomplete evidence in ontological knowledge bases. In Proceedings of the 22nd international conference on World Wide Web (pp. 413-422).
6. Horn, A. (1951). On sentences which are true of direct unions of algebras. The Journal of Symbolic Logic, 16(1), 14-21.

POSTER 22

Integrating sustainability into reaction planning for the AI4Green electronic laboratory notebook

Joe Davies, University of Nottingham

Traditionally, sustainability has not been a priority during the discovery phase of research and development programs. Reasons for this include the small scale of the chemistry being performed and the significant additional overheads that would have to be introduced to measure and improve sustainability. There is now a wider recognition that designing experiments to be sustainable in the first instance can aid later process development and reduce the environmental impact and waste disposal costs in the discovery phase. We have integrated a sustainability framework into a retrosynthetic planner and condition prediction tool into the AI4Green electronic laboratory notebook (ELN). We use the AiZynthFinder retrosynthetic planner, ASKCOS condition prediction tool, and the CHEM21 sustainability framework. The ELN is a web-based application hosted at ai4green.app which requires no download and is accessible by browser. A user can enter their target molecule by drawing the structure or entering the SMILES string. This is used to generate a series of routes with conditions including solvent, reagent, catalyst, and temperature. Sustainability is compared by a combination of ranking routes, and supporting visualisation, such as color-coding. Data are sourced from a single Postgres database connected to the ELN, containing CHEM21 sustainability data and chemical data extracted from PubChem. The sustainability is automatically assessed by the backend Python code and the processed results are posted to the frontend. This provides a quick method for the chemist to generate a wide range of possible routes that they could then assess the feasibility of using their intuition and experience and combine this with the computer-generated sustainability to select the most appealing option. The work here presents an example of integrating existing machine learning tools and combining these with a sustainability framework to provide automatic sustainability assessments at the route planning stage with minimal overhead for the chemist.

POSTER 23

ChemXpander: Exploration of Ultra Large Small Molecule Libraries for Hit Discovery

Peter Curran, Pharmenable Tx

Generating novel molecules in silico is a fundamental component of modern hit discovery. Enumerating reaction-based networks is a powerful method to access vast chemical spaces that, crucially, are synthesisable. Through leveraging the principles of diversity-oriented synthesis (DOS), reaction-based networks can be designed to access more complex, more 3D molecules that can assist drug hunters in solving common drug discovery challenges such as the discovery of novel bioactive chemotypes, improving ADME properties or manoeuvring through IP space. The major limitation of this approach is the time-intensive, expert-driven design process of the reaction network. Herein, we present ChemXpander, a prototype web application that assists synthetic experts in assembling and encoding reaction networks. We demonstrate the value of this approach by the rediscovery of Celecoxib via a genetic-based multiparametric optimisation of the input reactants to the network.

POSTER 24

Fusing spectral geometry and content-based image retrieval techniques to generate a 3D alignment-invariant shape and electrostatic molecular descriptor

James Middleton[1]*, Gian Marco Ghiandoni[2], Martin J. Packer[3], Mengdie Zhuang[1], Valerie J. Gillet[1]

1 Information School, University of Sheffield, The Wave, 2 Whitham Rd, Sheffield S10 2AH, United Kingdom

2 AstraZeneca R&D IT, Academy House, 136 Hills Road, Cambridge CB2 8PA

3 AstraZeneca Early Oncology R&D, Alderley Park, Macclesfield, SK10 4TG

It has been well established that shape complementarity plays an important role in molecular recognition. However, shape information alone does not suffice for describing accurately the binding process between a drug molecule and a therapeutic target, which also requires accounting for their electrostatic complementarity (Rathi et al., 2020).

In our work, the MolSG spectral geometry-based molecular descriptor workflow developed by Seddon et al. (2019) has been modified to include electrostatic information whilst retaining the beneficial alignment-invariance property of the original shape descriptor. Electrostatic potential (ESP) surfaces are computed using the Graph Convolutional Network developed by Rathi et al. (2020) which has been shown to produce potentials of comparable quality to computationally expensive density-functional theory (DFT) derived ESP surfaces whilst taking a fraction of the time.

The electrostatic information is represented using alignment-invariant 1D histograms which take inspiration from the colour histograms popularized in content-based image retrieval. The ESP energy information is initially projected onto a spherical surface which is subsequently separated into distance-based bins with respect to an intrinsic surface centroid. This centroid has been approximated using gradient descent with momentum alongside an implementation of the Riemannian center of mass as a cost function. The application benefit of using an intrinsic centroid as opposed to an extrinsic centroid based on the surface volume, means that the alignment-invariance property of the original MolSG descriptor is retained. A new algorithm, called Electro-MolSG, was therefore implemented by augmenting the MolSG 3D shape information with spatially enriched electrostatic information using 1D ESP energy histograms.

Electro-MolSG was benchmarked against a series of established descriptors in a ligand-based virtual screening setting using the DUD-E dataset. Initial findings suggest that the Electro-MolSG descriptor can outperform the shape-only implementation of MolSG, leading to enhanced enrichment in ligand-based virtual screening applications.

References

Prakash Chandra Rathi, R. Frederick Ludlow, and Marcel L. Verdonk, Journal of Medicinal Chemistry 2020 63 (16), 8778-8790, DOI: 10.1021/acs.jmedchem.9b01129

Matthew P. Seddon, David A. Cosgrove, Martin J. Packer, and Valerie J. Gillet, Journal of Chemical Information and Modeling 2019 59 (1), 98-116, DOI: 10.1021/acs.jcim.8b00676