POSTER 1

Investigating the impact and mitigation of noise in QSAR models

Adelaide Punt[1], Thierry Hanser[2], Stephane Werner[2], and Garrett M. Morris[1]

[1]Oxford Protein Informatics Group, Department of Statistics, University of Oxford, Oxford OX1 3LB, UK; [2]Lhasa Limited, Granary Wharf House, Leeds, LS11 6PS

Quantitative Structure-Activity Relationship (QSAR) models are used across the pharmaceutical, cosmetic, and agricultural industries to predict molecular properties and biological activities of various compounds. However, they are sensitive to experimental noise—stemming from measurement errors, environmental variability, and inherent biological fluctuations. This work addresses QSAR models' ability to perform under noisy conditions, focusing on improving the accuracy of predictions in real-world drug discovery scenarios.
We explored the impact of noise on QSAR models using a diverse set of molecular representations and machine learning approaches. We introduced controlled artificial noise, mimicking real-world conditions, into datasets and assessed its effects on model performance. We benchmarked various QSAR models, including Random Forest (RF), Support Vector Machine (SVM), XGBoost, and Gaussian Process (GP), paired with molecular representations such as ECFP4, SMILES strings, and molecular graphs. Quantum mechanical (QM) properties, specifically HOMO and LUMO, served as the basis for evaluating model accuracy, as they are deterministic with respect to a given molecular representation and therefore provide a source of clean data. RF/ECFP4 and XGBoost/SMILES combinations demonstrate considerable resilience to noise, maintaining relatively stable performance even as noise levels increase. Conversely, SVMs exhibited the most pronounced performance degradation under similar conditions. Graph-based models, particularly GINs, displayed potential with high R^2 values but suffered from high epistemic uncertainty, limiting their applicability in noisy environments. This research provides a guideline for both constructing noise-robust QSAR models as well as testing an existing model's performance with noise. Future work will extend these analyses to activity prediction and explore the implications of noise in federated learning, aiming to refine the QSAR modeling process further.

==

POSTER 2

Generative AI as a tool for metabolite identification

Alexander Porter & Nick Mulholland - Syngenta

Understanding metabolism plays a crucial role in the development and registration of agrochemicals and pharmaceuticals.1,2 Identifying the structures of metabolites generated in complex real-world studies remains a significant challenge, particularly when those metabolites are the product of multiple successive transformations from the starting material. In this study, we propose a novel approach for metabolite identification by combining generative chemistry, reinforcement learning (RL) and cheminformatics pipelines to tackle this problem.

Our method employs a generative model trained on known metabolite structures and enumerated chemical space around a specific unknown metabolite. The enumerated structures are generated using the opensource molecular graph generator Surge3 and then filtered using NextMove's Arthor4. The generative agent then iteratively constructs metabolite candidates, and the generated structures are evaluated using a scoring function. This scoring function consists of: the presence of known substructures; the total mass of the species, and other relevant structural information from spectroscopic studies of the unknown metabolite. Through reinforcement learning the agent learns to generate high-scoring metabolite structures through trial-and-error exploration guided by the scoring function.

By leveraging the power of generative chemistry and reinforcement learning, our approach demonstrates the potential of this technology to provide predictions of metabolite structures.

===

POSTER 3

Implementation of R-BIND Nearest Neighbour Search and ROBIN Machine Learning methods to generate a RNA focused Virtual Library

C.Chu, I. Proietti Silvestri - Liverpool ChiroChem

RNA targeting small molecules had gained a growing interest as the understanding of RNA in diseases and RNA related technologies improve. With the increased attention, QSAR-based tools such as R-BIND Nearest Neighbour Search (NNS) [1] and ROBIN Machine Learning (ML) [2] have been developed based on RNA/nucleic acid binder databases. By adapting a combination of R-BIND NNS and ROBIN ML method, a 5.2 million-member virtual lead-like space based upon difunctional scaffolds had been analysed to extract 1.5 million potential RNA targeting compounds. A comparison of the two methods, along with the analysis of the combined result will also be discussed.

1. Morgan BS, Sanaba BG, Donlic A, Karloff DB, Forte JE, Zhang Y, Hargrove AE. R-BIND: An Interactive Database for Exploring and Developing RNA-Targeted Chemical Probes. ACS Chem Biol. 2019 Dec 20;14(12):2691-2700. doi: 10.1021/acschembio.9b00631. Epub 2019 Oct 29. PMID: 31589399; PMCID: PMC6925312.
2. Yazdani K, Jordan D, Yang M, Fullenkamp CR, Calabrese DR, Boer R, Hilimire T, Allen TEH, Khan RT, Schneekloth JS Jr. Machine Learning Informs RNA-Binding Chemical Space. Angew Chem Int Ed Engl. 2023 Mar 6;62(11):e202211358. doi: 10.1002/anie.202211358. Epub 2023 Feb 6. PMID: 36584293; PMCID: PMC9992102.

===

POSTER 4

KLig Enumerator: a Tool for Library Creation Based on Kinase X-Ray Structures.

F. Ricci,[a] D. Sotillo,[a] G. Bottegoni[a]
[a]University of Urbino, Dept. of Biomolecular Sciences, Piazza Rinascimento 6, Urbino

Kinases encompass one of the widest groups of proteins targeted in drug discovery campaigns. Not only due to their involvement in several diseases, kinases are often object of study in the context of multi – target drug discovery programs.[1]
Despite the different pockets and binding modes reported in literature for specific kinases or chemotypes, the so-called hinge region of the ATP binding site results is a conserved hot – spot for inhibitor design.[2] Furthermore, a plethora of co – crystal structures are available in the RCSB PDB and several information about ligand binding are now encoded in the KLIFS database.[3] Therefore, embedding moieties, whose binding mode in the proximity of the hinge is already known by means of crystallography, into another scaffold might be a strategy for increasing the chance of obtaining tailored kinase inhibitors and also for developing “fused” multi-target compounds.
Herein, KLig compound enumerator, a tool able to extract the fragments binding the hinge region from a list of kinase co-crystals and to connect them to a user defined scaffold, is presented.
The computational pipeline has been indeed employed to generate two libraries of potential dual inhibitors for the treatment of several multifactorial diseases, starting from a list of 4381 kinase complexes and two privileged cores. Nevertheless, further code implementations aimed at performing fully combinatorial library generation or at targeting different kinase regions are under evaluation.

References:
[1] RMC. Di Martino et al., ChemMedChem 2020, 15(11), 949 - 954.
[2] L. Xing et al., J Comput Aided Mol Des 2014, 28(1), 13 - 23.
[3] OP. van Linden et al., J Med Chem 2014, 57(2), 249 - 77.

==

POSTER 5

Aminoacyl-tRNA synthetases (aaRSs) as drug targets

Eunice S. H. Gwee, Jagmohan S. Saini, Peter E. G. F. Ibrahim, Mike J. Bodkin, Drug Discovery Unit, University of Dundee

Aminoacyl-tRNA synthetases (aaRSs), a family of enzymes that catalyses the reaction between tRNA and its cognate amino acid with high levels of fidelity, play a vital role in protein synthesis. The ability to target various sites of the aaRS, such as the amino acid or the adenosine triphosphate (ATP) active sites, and disrupt protein translation makes them suitable drug targets. Mupirocin and borrelidin are some examples of aaRS inhibitors, specifically for threonyl- and isoleucyl-tRNA synthetases, respectively. The differences in enzyme sequences and structures between human, plasmodium falciparum (Pf), mycobacterium tuberculosis (Mtb) and fungi make it possible to selectively target the latter three without cross reacting with the former. This has motivated the need to design an aaRS focused library for hit identification and fragment screening. To achieve this, a thorough investigation of the aaRSs' active sites of all four species must be conducted. Previously published crystal structures and AlphaFold structures of the aaRSs will be used for this study. Firstly, receptor-based docking studies of enzymes against a library of ligands will be performed to identify ligands who would bind with high affinity. Subsequently, molecular dynamics (MD) simulations will be performed on the selected protein-ligand complexes to better understand their interaction and the rigidity of the active site. Thereafter, interaction energies between the ligand and residues at the active site, calculated using quantum mechanics (QM), specifically via Fragment Molecular Orbital Theory (FMO). Cavity comparison analysis will also be done based on FMOphore, an inhouse program, to recognize the presence of conserved interaction types. These would pinpoint the essential residues required at the active site for protein-ligand binding to occur. Unique ligand-residue interactions will be exploited using FraGrow, another inhouse program, for fragment-based drug discovery (FBDD) and design new lead ligands that are selective for the different aaRSs of different species.

===

POSTER 6

Synthesis directed elaborations by automated chemistry to maximally exploit the fragment-design space

Kate Fieseler[1,2], Max Winokan[2,3], Daniel Muñoz Reyes[4], Matteo Ferla[1], XChem[3], Maria Jose Sanchez-Barrena[4], Frank von Delft[1,2,3,5], Charlotte M. Deane[1], and Warren Thompson[2,3]

1. University of Oxford
2. Research Complex at Harwell
3. Diamond Light Source
4. Instituto de Química Física
5. University of Johannesburg

A fragment screen offers an information rich starting point for derivative compounds that can recapitulate fragment interactions[1]. The synthetic accessibility of fragment inspired compound designs originating from in silico methods varies significantly and the number of designs proposed are usually limited to <20[2–4]. While successful in these use-cases, currently no tools are available that propose 1,000s of compounds designed to be synthesized by high-throughput robotic chemistry with shared routes and starting materials. Our approach Syndirella (synthesis directed elaborations) elaborates fragment merges, through digitised multi-step synthetic routes, to maximize the exploration of the merge-design space. From the explicit restriction to synthetic accessibility, and consideration of lead-time and reactant budget, rapid (<10 weeks) fragment progression to experimental elaborated structures was achieved for a calcium binding protein with our novel pipeline.

Syndirella proposes elaborations of fragment merged designs, from in silico synthesis of superstructures of reactants using tractable synthesis routes, that are energy minimised in the protein with restraints to experimental data, and scored on recapitulation of experimental interactions. Synthesis requirements are tunable to intuitive chemical rules that are expandable to the user's preferences.

Applied to a fragment screen of the calcium binding protein (neuronal calcium sensor-1), our approach produced in silico compound sets that used inexpensive building blocks (<$7/mg) and produced 60% greater coverage of interactions when compared to the average random compound set within the same $8000 budget. The final compound set was synthesized in four days, using tractable synthesis routes, resulting in 14 experimentally determined crystal structures being found. With tunable synthetic and logistical constraints at every stage of the design process combined with high-throughput synthesis and structural biology, we were able to rapidly explore the merge-design space to fast-forward the initial phases of fragment development campaigns.

References

(1) Keserű, G. M.; Erlanson, D. A.; Ferenczy, G. G.; Hann, M. M.; Murray, C. W.; Pickett, S. D. Design Principles for Fragment Libraries: Maximizing the Value of Learnings from Pharma Fragment-Based Drug Discovery (FBDD) Programs for Use in Academia. J. Med. Chem. 2016, 59 (18), 8189–8206. https://doi.org/10.1021/acs.jmedchem.6b00197.
(2) Tang, Y.; Moretti, R.; Meiler, J. Recent Advances in Automated Structure-Based De Novo Drug Design. J. Chem. Inf. Model. 2024. https://doi.org/10.1021/acs.jcim.4c00247.
(3) Sommer, K.; Flachsenberg, F.; Rarey, M. NAOMInext – Synthetically Feasible Fragment Growing in a Structure-Based Design Context. European Journal of Medicinal Chemistry 2019, 163, 747–762. https://doi.org/10.1016/j.ejmech.2018.11.075.
(4) Chevillard, F.; Rimmer, H.; Betti, C.; Pardon, E.; Ballet, S.; van Hilten, N.; Steyaert, J.; Diederich, W. E.; Kolb, P. Binding-Site Compatible Fragment Growing Applied to the Design of Β2-Adrenergic Receptor Ligands. J. Med. Chem. 2018, 61 (3), 1118–1129. https://doi.org/10.1021/acs.jmedchem.7b01558.

==

POSTER 7

Comprehensive machine learning boosts structure-based virtual screening for PARP1 inhibitors

Klaudia Caba[1], Viet-Khoa Tran-Nguyen[2], Taufiq Rahman[3], Pedro J. Ballester[1]*
[1]Department of Bioengineering, Imperial College London, London, SW7 2AZ, UK
[2]Unité de Biologie Fonctionnelle et Adaptative, UFR Sciences du Vivant, Université Paris Cité, 75013 Paris, France
[3]Department of Pharmacology, University of Cambridge, Cambridge, CB2 1PD, UK
*Corresponding author: p.ballester@imperial.ac.uk

Poly (ADP-ribose) polymerase 1 (PARP1) plays a key role in the microhomology-mediated end joining (MMEJ) pathway, critical for DNA damage repair. In addition to its function in DNA repair, PARP1 is involved in processes like transcription regulation, chromosome stability, cell division, differentiation, and apoptosis, making it an attractive target for cancer

therapy, particularly in breast and ovarian cancers with BRCA mutations. Although several PARP1 inhibitors have received FDA approval, issues like non-selective activity, poor solubility, and toxicity persist, necessitating the development of new inhibitors.
This study investigates the use of machine learning (ML) to enhance structure-based virtual screening (SBVS) for identifying potent and novel PARP1 inhibitors. Classical scoring functions (SFs) used in molecular docking often fall short in capturing the complex interactions within protein-ligand binding, leading to suboptimal performance. In contrast, ML-driven docking can account for nonlinear interactions and has successfully identified molecules with affinity for various targets.
We developed ML SFs tailored to PARP1, leveraging extensive experimental and synthetic data, using various ML models and featurisation strategies. Experimental bioactivity data was retrieved from ChEMBL and PubChem and property-matched decoys were generated as synthetic data. Ligands were docked into the PARP1 structure (PDB ID 7KK4) using Smina. The dataset was split into training and test sets, with an additional dissimilar test set to evaluate performance on diverse compounds. Five supervised learning algorithms were employed in both binary classification and regression formats, including random forest, extreme gradient boosting, support vector machines, artificial neural networks, and deep neural networks. Featurisation techniques included protein-ligand complex-based features like protein-ligand extended connectivity fingerprints (PLEC) and 3D grid-based features, along with ligand-only Morgan fingerprints.
Our results identified highly predictive ML SFs, demonstrating the potential of integrating ML techniques with SBVS for discovering novel and potent PARP1 inhibitors.

===

POSTER 8

Automated Identification of Cryptic Pockets for Drug Discovery

Lukas Eberlein, Neha Vithani, Gunther Stahl, David Lebard; OpenEye, Cadence Molecular Sciences

Identification of cryptic pockets has the potential to open new therapeutic opportunities by discovering ligand binding sites that remain hidden in static apo structures of a target protein. Moreover, allosteric cryptic pockets can become valuable for designing target-selective ligands when the known binding sites are conserved in variants of a protein. Here, we present a newly developed approach for the exploration of cryptic pockets using weighted ensemble molecular dynamics simulations with inherent normal modes as progress coordinates applied to the wild type KRAS and the G12D isoform. We performed all-atomic simulations with and without several co-solvents (xenon, ethanol, benzene), and analyzed trajectories using three distinct methods to search for potential binding pockets.

===

POSTER 9

iScore: Ultra-fast Virtual screening

Sayyed Jalil Mahdizadeh at Gothenburg University

In the quest for accelerating de novo drug discovery, the development of efficient and accurate scoring functions represents a fundamental challenge. This study introduces iScore, a novel machine learning (ML)-based scoring function designed to predict the binding affinity of protein-ligand complexes with remarkable speed and precision. Uniquely, iScore circumvents the conventional reliance on explicit knowledge of protein-ligand interactions and full picture of atomic contacts, instead leveraging a set of ligand and binding pocket descriptors to evaluate binding affinity. This approach avoids the inefficient and slow conformational sampling stage, thereby enabling the rapid screening of ultra-huge molecular libraries, a crucial advancement given the practically infinite dimensions of chemical space. iScore was rigorously trained and validated using the PDBbind 2020 refined set, CASF 2016, and CSAR NRC-HiQ Set1/2, employing three distinct ML methodologies: Deep Neural Network (iScore-DNN), Random Forest (iScore-RF), and eXtreme Gradient Boosting (iScore-XGB). A hybrid model, iScore-Hybrid, was subsequently developed to incorporate the strengths of these individual base learners. The hybrid model demonstrated a Pearson correlation coefficient (R) of 0.78 and a root mean square error (RMSE) of 1.23 in cross-validation, outperforming the individual base learners and establishing new benchmarks for scoring power (R = 0.814, RMSE=1.34), ranking power (ρ = 0.705), and screening power (success rate at top 10% = 73.7%).

===

POSTER 10

Advancing Fragment-Based Drug Discovery in the ASAP Consortium with Fragmenstein

Matteo P. Ferla [Department of Statistics, University of Oxford; Centre for Medicine Discovery, University of Oxford], Charlotte M. Deane [Department of Statistics, University of Oxford] and Frank von Delft [Centre for Medicine Discovery, University of Oxford]

The ASAP Consortium (AI-driven Structure-enabled Antiviral Platform, asapdiscovery.org) is a collaborative, global effort uniting diverse groups to accelerate antiviral drug development against overlooked viral pathogens. The ASAP Consortium was born from the Covid Moonshot project and leverages the diverse combined expertise in drug discovery alongside high-throughput techniques, including crystallographic fragment screening and biochemical assays, and advanced computational methods (molecular mechanics, deep learning).
For the targets at the hit discovery stage in the ASAP consortium, fragment-based drug discovery (FBDD) is performed via an initial library screen at the XChem facility in Diamond Light Source, followed by testing of in silico derivations. This step uses a variety of approaches with the common aim of strictly following the conformation of template hits used. A common element in these processes is Fragmenstein, a Python-based tool for structure-based drug discovery. Fragmenstein operates under the principle that the binding mode remains consistent between the original fragment hit and the larger designed

molecule. Fragmenstein creates a stitched together conformer, a monster, which is energy minimised in contrasts to traditional methods that select pre-generated conformers. This approach has demonstrated superior accuracy in predicting protein-ligand complex conformations compared to existing pharmacophore-constrained docking techniques. It creates novel compounds by amalgamating one or more template hits or predicting bound conformers for virtual compounds based on template hits.
In the ASAP consortium, Fragmenstein synergises with other tools in five distinct pipelines:

- Fragmenstein Proper: Merges/links fragment hits, performs catalogue searches against SmallWorld server, and places analogues.
- Fragment Knitting: Enhances novelty by enumerating catalogue compounds related to template hits and placing them via Fragmenstein.
- SILVR and STRIFE: Utilise deep learning for de novo compound generation, followed by chemistry correction, SmallWorld search, and placement.
- Arthorian Quest: Addresses specific hypotheses by creating flexible SMARTS patterns, searching in Arthor server, and utilising Fragmenstein for placement.

The resulting virtual compounds undergo a multi-parametric scoring and interaction-based clustering, and upload to Fragalysis (fragalysis.diamond.ac.uk) for review by medicinal chemists prior to purchasing, thus facilitating efficient hit progression and discovery. This comprehensive approach within the ASAP Consortium underscores the potential of Fragmenstein in advancing FBDD, paving the way for swift, reliable antiviral drug development.

==

POSTER 11

Active learning FEP using 3D-QSAR for prioritizing bioisosteres in medicinal chemistry

Matthew Habgood, Venkata K. Ramaswamy, and Mark D. Mackey
Cresset, New Cambridge House, Bassingbourn Road, Litlington, Cambridgeshire, SG8 0SS, UK

Bioisostere replacement is a powerful and popular tool used to optimize the potency and selectivity of candidate molecules in drug discovery. Using human aldose reductase inhibitors we demonstrate how two rigorous computational methods, 3D-quantitative structure activity relationships (3D-QSAR) and relative binding free energy free energy perturbation (FEP), can be combined into an active learning workflow to prioritize molecules from a pool of several hundred bioisostere replacement candidates (generated by Cresset's Spark). This workflow can rapidly locate the strongest-binding bioisosteric replacements with a relatively modest computational cost. The ROC-AUC for selection of known actives in 80 top-ranked candidates improved to 0.88 from 0.64, and the top picks were enriched with highly potent ALR2 inhibitors, including the well-known Zopolrestat.

==

POSTER 12

FP-score for hotspot identification and efficient fragment-to-lead growth strategies

Peter E.G.F. Ibrahim,[1] Ulrich Zachariae,[1] Ian Gilbert[1] and Mike Bodkin[1*]
[1]Drug Discovery Unit, Division of Biological Chemistry and Drug Discovery, University of Dundee, Dow St, Dundee DD1 5EH, UK

Fragment based drug design (FBDD) is like a chess game in that a good or a bad move can dramatically influence the outcome. At the start of the design process, it is important to determine the key binding site residues (hotspots) that can have a substantial impact on ligand efficiency (LE) and binding. Here, we introduce a novel, fully automated algorithm named FMOPhore, performing Quantum Mechanics Fragment Molecular Orbital (QM-FMO) calculations on 3D-protein-ligand pharmacophore models. This is implemented in a novel scoring function namedFP-score to classify binding site residues in two classes: 1) Hotspot residues (divided into three categories; Anchor, Transient, and Accessible) and 2) Non-hotspot residues. Protein binding site flexibility using Dy-FMOPhore improves hotspots detection. We apply our algorithm in two different scenarios: apo-structure and holo-complex scenarios (290 protein-ligand complexes), testing its robustness on sixteen different protein targets including the application for fragment growing and target selectivity. The FMOPhore algorithm can be a powerful tool in identifying and quantifying binding site hotspots to enable an efficient design strategy for fragment-to-lead optimization.

==

POSTER 13

Benchmarking Generic Classical SFs For Structure Based Virtual Screening Against TRPM8 Ion Channel

Nivya James[1], Hannah Okesade[2], Taufiq Rahman[3] and Pedro J Ballster[1*]
1) Department of Bioengineering, Imperial College London
2) Department of Chemistry, Imperial College London
3) Department of Pharmacology, University of Cambridge

Growing evidence of the involvement of Transient Receptor Potential Melastatin 8 (TRPM8) ion channels in pain and cancer has opened new directions for treatment strategies. With advancements in computational power, virtual screening (VS) is increasingly used in the early stages of drug discovery as a fast and cost-effective alternative to wet-lab high-throughput screening (HTS) experiments. In this study, we evaluate the performance of existing generic classical scoring functions (SF), for use in a retrospective structure-based virtual screening (SBVS) study for TRPM8. Specifically, we evaluated the empirical scoring function (SF) of Smina, a fork of AutoDock Vina, with ChemEM, a recently developed flexible docking software for cryo-EM structures. Two high resolution cryo-EM structures of TRPM8 from Protein Data Bank were used as the protein structure template for the study. Docking studies were conducted on a 96,930 molecule dataset, which included activity-labelled data from public repositories and property-matched decoys generated using the DeepCoy algorithm. A rigorous analysis of the Smina SF, with five parameter sets, revealed poor performance for TRPM8 ion channel. However, preliminary results from a pilot study with

ChemEM SF suggest that it is promising for further full-fledged docking experiments. Given that TRPM8 ion channel drugs have consistently failed in clinical trials, SBVS campaigns like this could revitalize the stagnant field of pain therapy and accelerate drug discovery for multiple cancers.

==

POSTER 14

FraGrow: Fully automated software for fragment growth and optimization

Peter E. G. F. Ibrahim, Ulrich Zachariae, Ian Gilbert & Mike Bodkin.
[1]Drug Discovery Unit, Division of Biological Chemistry and Drug Discovery, University of Dundee, Dow St, Dundee DD1 5EH, UK

The growth of fragment hits to sub-micromolar is a bottleneck in drug discovery pipelines. For the initial optimization of the fragment, it is important to optimize the core scaffold as well as the growth vectors. However, the challenge is to obtain growth vectors with a low binding interaction energy, and appropriate 3D-spatial geometry of optimal functional groups to be added on the core scaffold of hit molecule. For this purpose, we introduce a novel fully automated software platform FraGrow, that combines DA-QM-FMO (Dynamical Average Quantum Mechanics Fragment Molecular Orbital) calculations, with protein-ligand pharmacophore model generation predicting the binding pocket hotspots and 3D spatial growth vectors - the FMOphore. This information is fed to deep generative models, for hit fragments linkage and core scaffold hopping. Fragments generated are subjected to binding pose prediction and scoring using the P-score protocol and ranking, to validated new hits both enthalpic and entropic terms. We apply FraGrow to main-protease protein (Mpro) for SARS-CoV-2 coronavirus as a case in a process of hit-to-lead optimization.

==

POSTER 15

Artificial Intelligence for Phenotypic Virtual Screening

Qianrong Guo, Department of Bioengineering, Imperial College London

Virtual screening (VS) is a process that screens potential hit compounds from large chemical libraries using computational tools. Current VS integrates Artificial Intelligence (AI) for more accurate predictions in drug identification. To make the AI black box prediction more reliable and robust, data is typically split into training, validation, and testing sets for the models to learn and evaluate. However, the data splitting methods might not always be trustworthy due to the data distribution shifts in real-world situations. The data collected for model building and evaluation can differ greatly from the molecule libraries used in subsequent wet lab experiments. Such discrepancy has led to a scarcity of promising results in the field of AI drug discovery due to models' limited generalizability.

Our study aims to identify robust data splitting and machine learning models in Phenotypic VS. We benchmarked the different machine learning models on the NCI-60 cancer cell line growth inhibition datasets under various splitting methods. We have concluded that the widely used scaffold split can overestimate the models' performance in more challenging real-world settings.
Additionally, to facilitate real-world drug discovery with our findings, we are further conducting a case study of anti-tuberculosis drug discovery using phenotypic virtual screening. With data collected from previous studies for model development and external testing, we obtained preliminary results benchmarking 11 machine learning models. As the next steps, we plan to benchmark more methods under realistic splitting methods, identify potential hit compounds, and verify via resazurin reduction assay.

==

POSTER 16

Prioritisation and Ranking of fragments from a high-throughput screen

Ronald Cvek (Oxford, Diamond), Max Winkoan (Diamond), Frank von Delft (Oxford, Diamond, CMD)

Here we develop a robust computational pipeline to process and prioritise fragment hits from high-throughput crystallography screens, integrating techniques like dynamic undocking (DUck), molecular dynamics (MD), interaction fingerprinting (PLIFs), and quantum mechanics (QM) calculations to allow a more quantitative approach to the selection of fragments for progression.

==

POSTER 17

Towards Automated Reassignment of NMR Spectra

Kotlyarov, Ruslan[1], Maltby, Thomas[1], Howarth, Alexander[1], Goodman, Jonathan M. [1]
[1] University of Cambridge, Department of Chemistry

Nuclear magnetic resonance (NMR) spectroscopy is central to small molecule characterisation. However, the deduction of structural information from NMR spectra remains a challenging task to automate, especially for natural products. The structural assignment benefits from extra information, such as prior knowledge of the species involved in the reaction or data from the technically sophisticated 2D NMR experiments.

Inspired by open-source advances in generative models, inverse design, and neural network-based NMR spectrum simulation, we treat structure elucidation from NMR spectra as a molecular optimisation problem. The fitness criterion incorporates how well the generated molecules match the experimental spectrum.

In our studies, we use the REINVENT framework to generate molecules and the CASCADE neural net to simulate their 13C NMR spectra. We investigated the use of DP5 and corrected mean absolute error as fitness criteria. We achieved structure reassignment automatically, based solely on 13C NMR spectra.

==

POSTER 18

DEEPGRID: 3D image recognition using deep learning and grid molecular interaction fields

Simon Cross (1), Loriano Storchi (2)
1) Molecular Discovery, Kinetic Business Centre, Theobald Street, Borehamwood, Herts, WD6 4PJ, U.K.
2) Dipartimento di Farmacia, Università G. D’Annunzio, Via dei Vestini 31, 66100 Chieti, Italy

We report a novel method for building regression and classification models using GRID Molecular Interaction Fields in combination with a Deep Learning using an image recognition approach. Typical image recognition uses two-dimensional images described by red, green, and blue channels; with our method molecules are described by three-dimensional images using a number of channels that describe their interactions with diverse molecular probes at each position in the image. Molecular Interaction Fields can be generated to describe both small molecules and macromolecules and hence used as inputs for the method; here we report good models for both a small molecule blood-brain barrier dataset and an allosteric protein cavity dataset, demonstrating that the method is broadly applicable to diverse datasets. The classical problematic requirement of aligning molecules in the dataset is avoided through the use of data augmentation and the generation of diverse ‘viewpoints’ for each object molecule; the Deep Learning approach learns to extract and recognise important features regardless of the image orientation, making dataset preparation trivial and the resulting models robust. Whilst dataset preparation is trivial, model building is relatively slow as we explore thousands of potential models via an automated hyperparameter search, however this is a one-off cost and prediction of new objects is relatively fast. One limitation we encountered is that the data size of the molecular images are relatively large, meaning that for the larger datasets we explored we were unable to load all of the data concurrently onto the graphics processing unit,
however this was solved using a batch loading approach. The main remaining limitation is the ‘black-box’ nature of the models and their interpretability, which is something we are currently exploring.

==

POSTER 19

Comparative Study of Allosteric GPCR Binding Sites and Their Ligandability Potential

Sonja Peter[1,2], Lydia Siragusa[3,8], Morgan Thomas[1,6#], Tommaso Palomba[4], Simon Cross[4], Noel M. O’Boyle[1], Dávid Bajusz[5], György G. Ferenczy[5], György M. Keserű[5], Giovanni Bottegoni[2,7], Brian Bender[1], Ijen Chen[1*], Chris De Graaf[1*]

[1] Nxera Pharma UK, Computational Chemistry, Steinmetz Building, Granta Park, Cambridge CB21 6DG, UK. *ijen.chen@nxera.life; *chrisdgrf@gmail.com;
[2] Department of Biomolecular Sciences, University of Urbino Carlo Bo, Piazza Rinascimento 6, 61029 Urbino (Italy)
[3] Molecular Discovery Ltd., Kinetic Business Centre, Theobald Street, Elstree, Borehamwood, WD6 4PJ Hertfordshire, UK.
[4] Molecular Discovery Ltd., via Montelino 30, 06084 Bettona (PG), Italy
[5] Medicinal Chemistry Research Group, HUN-REN Research Centre for Natural Sciences, Magyar tudósok krt. 2, 1117 Budapest, Hungary
[6] University of Cambridge Yusuf Hamied Department of Chemistry Cambridge, UK CB2 1EW
[7] Institute of Clinical Sciences, University of Birmingham, Edgbaston, B15 2TT, Birmingham, United Kingdom
[8] Molecular Horizon, via Montelino 30, 06084 Bettona (PG), Italy
#current address: Computational Science Laboratory, Universitat Pompeu Fabra, Barcelona, Spain

The steadily growing number of experimental G-protein-coupled receptor (GPCR) structures has revealed diverse locations of allosteric modulation and yet few drugs target them. This gap highlights the need for a deeper understanding of allosteric modulation in GPCR drug discovery. The current work introduces a systematic annotation scheme to structurally classify GPCR binding sites, based on receptor class, transmembrane helix contacts, and for membrane-facing sites, membrane sublocation. This GPCR specific annotation scheme was applied to 107 GPCR structures bound by small molecules contributing to 24 distinct allosteric binding sites for comparative evaluation of three binding site detection methods (BioGPS, FTMap and SiteMap). BioGPS identified the most in 22 out of 24 sites. In addition, our property analysis showed that extrahelical allosteric ligands and binding sites represent a distinct chemical space characterized by shallow pockets with low volume, and the corresponding allosteric ligands showed an enrichment of halogens. Furthermore, we demonstrated that combining receptor and ligand similarity can be a viable way for ligandability assessment. One challenge regarding site prediction is the ligand shaping effect on the observed binding site, especially for extrahelical sites where ligand induced effect was most pronounced. To our knowledge, this is the first study presenting a binding site annotation scheme standardized for GPCRs and it allows comparing allosteric binding sites across different receptors in an objective way. The insight from this study provides a framework for future GPCR binding site studies and highlights the potential of targeting allosteric sites for drug development.

==

POSTER 20

Cheminformatics/QSAR approach for personal care formulation optimization

Swathi S*, Saswati Pujari, Joe Jankolovits, Connor Walsh, Joe Carnali, Georgia Shafer, Ian Stott

Personal care formulations are complex mixtures of ingredients that have to meet various criteria such as product stability, viscosity, antimicrobial efficacy, and preservation. The choice of ingredients depends on the format and brand of the product, as well as the surfactant type, pH window, and the incorporation of functional actives. In this poster, we present a cheminformatics approach based on similarity searching with molecular fingerprints to identify alternative molecules within the formulation space that can satisfy the desired properties and constraints. Also, this approach, integrated with ML modelling, allows us to predict the antimicrobial efficacy of a formulation. The examples demonstrated here show that cheminformatics/QSAR can be a useful tool for personal care formulation design and optimization: 1) predicting the antimicrobial efficacy 2) finding molecular analogues through solution properties like viscosity & solubility.

==

POSTER 21

Molrus – A cheminformatics library written in Rust

Syed Zayyan Masud

Molrus is a freely available open-source Rust crate for cheminformatics. The aim is to provide methods for many common tasks in molecular informatics, including 2D and 3D rendering of chemical structures, I/O routines, SMILES parsing and generation, ring searches, isomorphism checking, structure diagram generation, etc. Built with performance and safety in mind, Molrus leverages Rust's ecosystem to provide an efficient, reliable solution for researchers and developers. While there may be many similar tools available, Molrus aims to offer an alternative that is robust, modern, and adaptable to various workflows in cheminformatics. The question of its necessity is left to the community—it exists to provide choice and innovation.

==

POSTER 22

In Silico Identification of Readily Available Target Pairs for Dual-Target-directed Ligand Development

Vittorio Lembo[1,2] and Giovanni Bottegoni[1,3]

[1] Department of Biomolecular Sciences, Università degli Studi di Urbino Carlo Bo, Piazza Rinascimento 6, 61029, Urbino, Italy

[2] Computational and Chemical Biology, Istituto Italiano di Tecnologia, Via Morego 30, 16163 - Genova (Italy)
[3] Institute of Clinical Sciences, University of Birmingham, Edgbaston, B15 2TT, Birmingham, United Kingdom

In a previous work (1), we systematically analyzed the Dual-Target-Directed Ligands (DTDLs) present in the literature, drawing the following conclusions:

- the molecular structures of the analyzed DTDLs are extensively based on pre-existing chemistry of single target compounds;
- Target association was systematically made based on one or more of the following factors:
- a strong pre-existing pathology-targets association
- structural similarity of the two binding pockets
- pre-existing overlap between the chemistry of single target modulators

Our investigation revealed that a significant portion of the therapeutic potential of DTDLs remains untapped.

Building on these insights, this work seeks to determine whether all the readily available target combinations for designing dual-active compounds have been fully explored. We present an in silico pipeline for identifying target pairs that exhibit the key association features, namely common target-disease associations, pocket similarity, and overlapping areas of the chemical space.

The pipeline consists of two main steps:

1. Identification of target pairs with a common strongly associated disease
   Human targets from ChEMBL(2) with at least 10 compounds meeting specific criteria (pChEMBL_value > 5, assay_confidence_score > 8, 198>MW>800, only Binding or Functional assay_type) were collected.
   Target-disease associations were retrieved from Open Targets Platform (3), a web-based platform which integrates multiple data sources, and returns the Overall Association Score (OAS), of the association of a protein with a disease. For each target only the strongly associated diseases (OAS in the 99th percentile) were retrieved. Targets were compared to identify pairs sharing at least one common strongly associated disease.
2. Chemogenomic analysis of identified target pairs.
   Each identified target pair will be evaluated to determine the overlap in their ligand chemical space, and DeeplyTough to quantify the structural distance between their binding pockets. (4) Identified target pairs presenting these features will be reported and discussed as case-studies.

1. Lembo V, Bottegoni G. Journal of Medicinal Chemistry. 2024, 67, 12, 10374–10385
2. Zdrazil B. et al., Nucleic Acids Research, 2024,52 1180-1192
3. Ochoa D, et al. Nucleic Acids Research, 2023, 51, 1353–1359
4. Simonovsky M, Meyers J. Journal of Chemical Information and Modeling. 2020,4,60.

==

POSTER 23

Cheminformatics and Machine Learning Approaches for GPCR Computer-Aided Drug Design

Wei Dai (qmul), Arianna Fornili (qmul), Noel O'Boyle (Nxera), Chris de Graaf (Nxera)

In recent years, Artificial Intelligence (AI) has exhibited significant potential in computer-aided drug design, particularly through the utilization of generative models. This project explores the development and use of generative AI in the context of structure-based drug design. An extended module is currently being developed based on an existing de novo generative tool that has the capability to generate molecules which could be potential ligands for one given target from scratch. This module allows for the simultaneous utilization of multiple receptor structures during training. The effectiveness of this enhancement when utilizing multiple structures of the same receptor is currently under investigation.