PhD Project Proposal

Funder details

Studentship funded by:	MRC DTP

Project details

Project title:	Combining AI and physics-based simulations for fragment-based cancer drug discovery

Supervisory team

Primary Supervisor:	Dr. Gary Newton
Associate Supervisor(s):	Dr. Andrea Scarpino
Secondary Supervisor:	Dr. Rob van Montfort

Divisional affiliation

Primary Division:	Cancer Therapeutics
Primary Team:	In Silico Medicinal Chemistry – Medicinal Chemistry Team 3
Site:	Centre for Cancer Drug Discovery

Project background

Fragment-based drug discovery has delivered multiple FDA-approved drugs including capivasertib, the first-in-class AKT inhibitor co-developed by the ICR, Astex and AstraZeneca. Despite the track record and recent progress in the underlying design principles, fragment-to-lead optimisation remains challenging primarily due to two aspects: deciding which chemical modifications to pursue among billions of possibilities, and predicting whether these modifications will maintain the desired binding to the target protein. This project aims to develop a computational framework ideally suited to address both challenges.

Recent advances in machine learning offer promising solutions to the combinatorial problem, with the development of active learning strategies enabling efficient exploration of vast chemical spaces accessible via the rapid expansion of purchasable compound libraries. Furthermore, state-of-the-art generative AI models can propose entirely novel molecules with desirable properties. Computational methods that can reliably predict which fragment modifications will succeed are therefore essential.

The inherently weak binding affinity (mM-μM) and transient interactions in protein cavities make fragments particularly challenging to evaluate using traditional computational methods like docking. However, molecular dynamics simulations are often used to investigate the structural stability of protein-ligand complexes. Furthermore, enhanced sampling methods such as metadynamics have been shown as reliable tools for assessing binding stability in a limited timescale, providing an efficient physics-based approach to evaluate multiple hypotheses. Pose stability metrics (e.g., residence time, RMSD fluctuations, and hydrogen bond persistence) will be used to assess whether experimentally validated binding modes can be maintained through subsequent fragment elaborations.

Overall, this project seeks to bridge the gap between fragment hit identification and lead development by introducing pose stability as a selection criterion, enabling rational fragment expansion in vast chemical spaces. By integrating

machine learning, generative AI, and physics-based simulations, the approach promises to accelerate the development of new cancer therapeutics while reducing costs and improving success rates.

Project aims

- Evaluate simulation protocols to quantify fragment pose stability, comparing standard and enhanced sampling techniques and establishing metrics that predict successful elaborations.
- Create automated pipelines for pose stability assessments on virtual elaboration pools from commercial libraries or AI-generated molecules, with hierarchical filtering to focus physics-based simulations only on viable candidates, and validate this framework retrospectively on internal data or published fragment optimisation campaigns.
- Apply this methodology to increase the potency of weak binding fragments against a validated target of interest to the Centre of Cancer Drug Discovery; synthesising and testing the molecules.
- Implement active learning strategy to efficiently navigate the chemical space through iterative subset selection, considering reinforcement learning-active learning combinations to further boost sampling efficiency.
- Develop open-source tools integrating pose stability assessment with active learning, enabling community adoption for fragment-based drug discovery.

Research proposal

Fragment-based drug discovery (FBDD) offers a systematic approach to developing cancer therapeutics and has been successfully applied in campaigns leading to FDA-approved drugs (**Figure 1**). However, optimising weak-binding fragments into potent drugs remains challenging. This project aims to accelerate the application of FBDD strategies to cancer targets through a novel computational framework that integrates AI and physics-based simulations. The framework will be applied to predict fragment hits and their successful elaborations by assessing pose stability and leveraging active learning to prioritise promising compounds. This innovative approach addresses one of the fundamental challenges in FBDD: selecting elaborations from billions of possibilities that preserve validated interactions while improving affinity. We will subsequently apply this to increase the potency of fragment hits against a target of interest to the Centre for Cancer Drug Discovery.

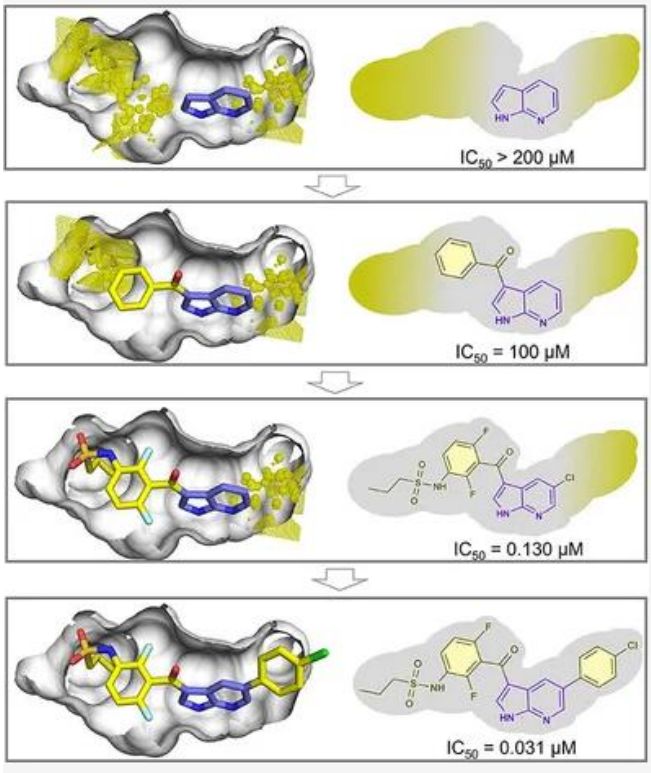

Figure 1: The iterative process that lead from a fragment hit bound to the oncogenic B-RAF kinase to vemurafenib, the first marketed drug from fragment-based drug design. The binding site surface is shown in grey, and the space predicted to be available for expansion as yellow spheres and surface. Figure from de Souza Neto LR et al., Front. Chem. 8:93, 2020.

Stage 1: Method development and validation

The first stage focuses on developing and benchmarking computational methods that assess fragment pose stability as a means of identifying fragment hits.

We will take validated fragment hits with crystallographic data from internal projects or public datasets (e.g., XChem screens) and evaluate the ability of different methods to recall hits within sets containing inactives or decoys. In particular, we will analyse the scope and performance of existing computational protocols for pose stability assessments, probing their suitability for fragment-sized molecules. Different approaches will be benchmarked, including standard molecular dynamics and metadynamics-based simulations (e.g., BPMD), in comparison to docking. The goal

is to implement and validate a methodology providing an optimal balance between accuracy, hit enrichment and computational efficiency.

We will then build automated pipelines that allow us to assess pose stability following iterative virtual elaborations of these hits using commercial libraries. Hierarchical filters will be employed (e.g., pharmacophore matching) to ensure selected molecules maintain key fragment features upon elaboration, preventing computational resources being wasted on unproductive candidates. Once again, we will test the ability of this approach to select for modifications that are known to increase potency. Furthermore, as we increase the molecular size during elaborations, we will assess whether pose stability still provides a suitable metric for assessing larger compounds, or if these could benefit from more expensive but accurate free energy calculations. An adaptive decision framework will be evaluated to select the most appropriate methods based on the optimisation stage, available computational resources, and required accuracy.

This retrospective validation uses well-documented cases from literature where complete data exists from fragment to lead candidate, as well as targets from internal programmes with available crystallographic fragment data. This provides advantages over published studies in terms of complete structure-activity relationships, consistent experimental protocols, and broader access to negative data (i.e., failed elaborations). The validation will allow to establish baseline performance metrics and identifies method strengths and limitations.

Stage 2: Application to real-world cancer drug discovery

In this stage, we will apply our chosen method to increase the potency of a fragment hit on at least one oncology target where we have established expertise in terms of assays and crystallography.

The computational framework will be applied prospectively in cancer drug discovery projects. Using the computational approaches established in Stage 1, we will select compounds for synthesis and biological testing, and this will serve as the basis to guide further iterations. Experimental validation will be carried out through orthogonal techniques to confirm binding and predicted binding modes.

This prospective validation will demonstrate the framework's real-world utility and provide critical feedback for method refinement. Results from synthesised compounds will inform model improvements, with special attention to understanding cases where computational predictions diverge from experimental results.

Stage 3: Protocol refinement

To enable virtual screening of large compound libraries from either commercial suppliers (e.g., Enamine REAL) or from multiple chemical elaborations, we will seek to implement active learning approaches to optimise computational efficiency. These approaches will balance exploration of novel chemical space with exploitation of promising regions, learning to prioritise compounds with favourable features. The implementation will require comparing different acquisition functions and their impact on discovery efficiency.

We will also explore integration with generative models to expand chemical space exploration. Implementations combining the active learning framework with reinforcement learning-based generative models (REINVENT4) will enable the exploration of novel regions of the chemical space beyond commercial libraries, potentially providing a further boost in sampling efficiency. This dual approach, i.e. sampling from commercial libraries and generating novel molecules, maximises the chemical diversity available for compound optimisation.

Stage 4: Dissemination

The final stage focuses on dissemination to the scientific community. We will publish retrospective validation results demonstrating the framework's ability to recover known successful elaborations and identify alternative optimisation paths. A second publication will present the prospective medicinal chemistry optimisation, including synthesised compounds, experimental characterisation, and lessons learned from computational predictions. We will evaluate packaging the framework as open-source tools to enable broader adoption of these methods, democratising access to advanced computational approaches for FBDD.

Overall, the developed framework aims to make fragment optimisation more efficient, ultimately accelerating the development of new cancer therapeutics. Successful applications will result in multiple high-impact publications and pave the way for applying this methodology to novel systems of interest to the ICR and the wider community.

Supervision of the project

The student will be supervised by Dr. Andrea Scarpino and Dr. Gary Newton (ICR). The student will carry out the proposed research within the In Silico Medicinal Chemistry team and the Medicinal Chemistry Team 3 at the Centre for Cancer Drug Discovery (CCDD) and benefit from the significant experience in drug design and medicinal chemistry in both teams. In addition, the interdisciplinary environment of CCDD will facilitate collaboration with other computational scientists, medicinal chemists, and experts in assay and structural biology. For example, the student

will have the opportunity to obtain high-impact intellectual and scientific input from Dr. Rob van Montfort (HDSD Team) and Prof. Swen Hoelder (Medicinal Chemistry 4).

Literature references

1. Murray, C.W. and Rees, D.C. (2009) 'The rise of fragment-based drug discovery', Nature Chemistry, 1(3), pp. 187-192.
2. Erlanson, D.A., et al. (2016) 'Twenty years on: the impact of fragments on drug discovery', Nature Reviews Drug Discovery, 15(9), pp. 605-619.
3. Lukauskis, D., et al. (2022) 'Open Binding Pose Metadynamics: An Effective Approach for the Ranking of Protein−Ligand Binding Poses', Journal of Chemical Information and Modeling, 62, pp. 6209-6216.
4. Graff, D.E., Shakhnovich, E.I. and Coley, C.W. (2021) 'Accelerating high-throughput virtual screening through molecular pool-based active learning', Chemical Science, 12(22), pp. 7866-7881.
5. Correy, G.J., et al. (2025) 'Exploration of structure-activity relationships for the SARS-CoV-2 macrodomain from shape-based fragment linking and active learning', Science Advances, 11, p. eads7187.
6. Dodds, P., et al. (2024) 'Sample efficient reinforcement learning with active learning for molecular design', Chemical Science, 15, pp. 4146-4161.

Candidate profile

Note: the ICR's standard minimum entry requirement is a relevant undergraduate Honours degree (First or 2:1).

Pre-requisite qualifications of applicants:

- MSc in Computational Chemistry, Computer Science with chemistry background, Medicinal chemistry, or related field
- Experience with molecular modelling and basic understanding of protein-ligand interactions
- Programming skills in Python and familiarity with scientific computing
- Interest in drug discovery and cancer research
- Organic chemistry wet-lab experience is desirable

Intended learning outcomes:

- Develop expertise in modern AI/ML approaches for molecular design
- Master molecular dynamics simulation techniques and understand their application to drug discovery
- Gain practical experience in fragment-based drug discovery workflows from computational prediction to experimental validation
- Build skills in scientific tool development and workflow automation
- Develop abilities in project management and scientific communication
- Understand the drug discovery pipeline and requirements for translational research

Advertising details

Project suitable for a student with a background in:

- [ ] Biological Sciences
- [ ] Physics or Engineering
- [x] Chemistry
- [ ] Maths, Statistics or Epidemiology
- [x] Computer Science
- [x] Life Sciences