Machine Learning Approaches for Targeted Fragment Library Design and Experimental Optimisation in Cardiometabolic Drug Development

Oxford supervisors: Dr. Fergus Imrie[1] and Professor Charlotte Deane[1]

Novo Nordisk supervisors: Dr Christos A Nicolaou

Departments: 1. Department of Statistics
2. Novo Nordisk

Background. Cardiovascular and cardiometabolic diseases remain the leading cause of morbidity and mortality worldwide, accounting for a substantial proportion of healthcare costs and premature death [1,2]. While existing treatments have led to improved patient outcomes, there remains considerable unmet need for novel pharmaceutical interventions capable of preventing, managing, or reversing disease progression. Recent advances in artificial intelligence and machine learning offer promising avenues for addressing the challenges inherent in drug discovery for cardiometabolic diseases.

In the Oxford Protein Informatics Group (OPIG), we have pioneered machine learning-based approaches for structure-based drug design, including methods for predicting small molecule binding (e.g. [3-5]) and generating novel compounds (e.g. [6-8]). Key features of our work are the ability to incorporate 3D structural information explicitly into compound design and enabling more control of the design process, thus allowing users to incorporate prior knowledge and design hypotheses.

Hypothesis. In contrast with method development, limited effort has been placed in establishing how to practically use such techniques most effectively to accelerate and improve drug discovery. Adopting these methods as drop-in replacements promises some improvements; however, such techniques have the ability to unlock new drug discovery paradigms and can be much more effectively harnessed.

This project seeks to develop techniques that improve the way we use machine learning tools in drug discovery. In particular, it will focus on (1) ways to use machine learning to inform experimental design and maximise information gain, and (2) techniques that can learn effectively from this information. Consequently, this project will be centred around the following 3 work packages (WP):

WP1: Machine learning approaches for fragment library design

Fragment-based approaches have become increasingly important tools for finding hit compounds for difficult protein targets, and have led to successes for targets that proved otherwise intractable. First, a library of fragments is screened to identify low-potency, high-quality leads. Currently, the most common strategy for library design is to maximise structural diversity. Recently, we showed that structurally diverse fragment libraries do not necessarily exhibit more functional diversity than randomly selected libraries and, further, functionally diverse fragment libraries recover substantially more information about novel targets compared with using randomly selected or structurally diverse fragments [9].

One limitation of this analysis was the requirement for data from numerous fragment screens, which also restricts analysis to previously screened fragments and protein targets. Additionally, this does

not allow new fragments to be introduced. Furthermore, a large number of fragments have never bound to a target, but are still included, thus providing unknown experimental information.

WP1 will develop computational structure-based approaches to fragment library design, representing a novel way of designing fragment libraries. To achieve this, we will develop fragment-specific scoring functions as the basis for computationally performing library optimisation. Furthermore, this will enable fragment libraries to be designed for a specific protein target, which has not yet been possible.

WP2: Protein-specific structure-based scoring functions

Computational assessment of protein-ligand complexes is a critical element of structure-based approaches to developing potent, selective small-molecule binders, with machine learning models showing substantial promise (e.g. [3-5]). A major challenge is the heterogeneity of binding between different targets. As a result, targeted scoring functions will often outperform a universal model [10] and current approaches perform substantially less well when assessed on unseen protein targets, hindering their prospective use. Furthermore, a small quantity of target-specific data is often available, either at inception or during a project, but is currently under-utilised by structure-based methods.

It is currently not clear how best to construct targeted scoring functions. Previously, we demonstrated how transfer learning could be used to incorporate domain-specific knowledge to construct protein family-specific models [3]. However, this approach is rudimentary and requires representative data from other members of the same protein family, which is often unavailable.

WP2 will explore methods for developing target-specific machine learning scoring functions, including both domain-driven approaches and machine learning techniques, such as unsupervised domain adaptation, few-shot learning, and test time training.

WP3: Active learning and experimental design

Substantial focus has been placed on techniques to issue predictions and propose molecular designs. However, how to select which compounds to experimentally evaluate next has received relatively limited attention. The Department of Statistics is at the forefront of research in uncertainty quantification, such as conformal prediction, active learning, and experimental design. For example, previously active learning sought to target samples with the greatest uncertainty; the RainML lab recently introduced a novel acquisition objective, the expected predictive information gain [11,12], that measures information gain in the space of predictions rather than parameters.

WP3 will develop new estimators and acquisition functions for active learning to select the most informative compounds. Critically, unlike many existing active learning settings, drug discovery requires the optimisation of multiple, often interdependent, factors, which will require methodological advances to solve effectively.

References

[1] M. Vaduganathan, et al. (2022). The global burden of cardiovascular diseases and risk. J. Am. Coll. Cardiol. 80 (25), 2361–2371.

[2] GBD 2021 (2023). Diabetes Collaborators. Global, regional, and national burden of diabetes from 1990 to 2021, with projections of prevalence to 2050: a systematic analysis for the Global Burden of Disease Study 2021. Lancet, 402(10397), 203–234.

[3] F. Imrie,..., C.M. Deane (2018). Protein Family-Specific Models Using Deep Neural Networks and Transfer Learning Improve Virtual Screening and Highlight the Need for More Data. J. Chem. Inf. Model. 58 (11), 2319-2330

[4] J. Scantlebury,..., C.M. Deane (2023). A Small Step Toward Generalizability: Training a Machine Learning Scoring Function for Structure-Based Virtual Screening. J. Chem. Inf. Model. 63 (10), 2960-2974.

[5] I. Valsson, M. Warren, C.M. Deane et al. (2025). Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data. Commun. Chem. 8, 41.

[6] F. Imrie,…, C.M. Deane (2020). Deep Generative Models for 3D Linker Design. J. Chem. Inf. Model. 60, 4, 1983–1995.

[7] F. Imrie,..., C.M. Deane (2021). Deep Generative Design with 3D Pharmacophoric Constraints. Chem. Sci. 12, 14577-14589.

[8] Y. Ziv, B. Marsden, C.M. Deane (2024). MolSnapper: Conditioning diffusion for structure based drug design. bioRxiv.

[9] A. Carbery,..., C.M. Deane (2022). Fragment Libraries Designed to Be Functionally Diverse Recover Protein Binding Information More Efficiently Than Standard Structurally Diverse Libraries. J. Med. Chem. 65 (16), 11404-11413.

[10] G.A. Ross, et al. (2013). One Size Does Not Fit All: The Limits of Structure-Based Models in Drug Discovery. J. Chem. Theory Comput. 9 (9), 4266-4274

[11] F. Bickford Smith, et al. (2023). Prediction-Oriented Bayesian Active Learning. 26th International Conference on Artificial Intelligence and Statistics (AISTATS), 206, 7331-7348.

[12] F. Bickford Smith, et al. (2024). Making Better Use of Unlabelled Data in Bayesian Active Learning. 27th International Conference on Artificial Intelligence and Statistics (AISTATS), 238, 847-855.

Supervisor's recent relevant publications

J. Scantlebury,..., C.M. Deane (2023). A Small Step Toward Generalizability: Training a Machine Learning Scoring Function for Structure-Based Virtual Screening. J. Chem. Inf. Model. 63 (10), 2960-2974

I. Valsson, M. Warren, C.M. Deane et al. (2025). Narrowing the gap between machine learning scoring functions and free energy perturbation using augmented data. Commun. Chem. 8, 41.

F. Imrie,…, C.M. Deane (2020). Deep Generative Models for 3D Linker Design. J. Chem. Inf. Model. 60, 4, 1983–1995.

F. Imrie,…, C.M. Deane (2021). Deep Generative Design with 3D Pharmacophoric Constraints. Chem. Sci., 12, 14577-14589.

A. Carbery,..., C.M. Deane (2022). Fragment Libraries Designed to Be Functionally Diverse Recover Protein Binding Information More Efficiently Than Standard Structurally Diverse Libraries. J. Med. Chem. 65 (16), 11404-11413