Combining molecular simulation and machine learning to INSPIRE improved cancer therapy.

Wright, D.W.[1], Devitt-Lee, A.[1], Clyde, A.[2], Palani, K.[3], Xia, F.[2], Turilli, M.[4], Karanicolas J.[3], Jha, S.[4,5], Stevens, R.[2], Chodera, J.D.[5], Coveney, P.V.[1,7]

[1]Centre for Computational Science, UCL, London, UK; [2]Argonne National Laboratory, USA; [3]Program in Molecular Therapeutics, Fox Chase Cancer Center, Philadelphia, USA; [4]Electrical & Computer Engineering, Rutgers, Piscataway, New Jersey, USA; [5] Brookhaven National Laboratory, USA; [6]Memorial Sloan Kettering Cancer Center, New York, USA; [7]Informatics Institute, University of Amsterdam, NL

1. The INSPIRE project

Cancer is the second leading cause of death in the United States (accounting for nearly 25% of all deaths). Targeted kinase inhibitors play an increasingly prominent role in the treatment of cancer and account for a significant fraction of the $37 billion U.S. market for oncology drugs in the last decade. Unfortunately, the development of resistance limits the amount of time patients derive benefits from their treatment. The INSPIRE project is laying the foundations for the use of molecular simulation and machine learning (ML) to guide precision cancer therapy, in which therapy is tailored to provide maximum benefit to individual patients based on genetic information about their particular cancer. It is vital that such an approach is based on predictive methods as the vast majority of clinically observed mutations are rare, rendersing catalog-building alone insufficient.

2. Predictive modelling

In employing predictive, physical modelling techniques we have focussed on two primary challenges: (1) generating realistic starting models of proteins (and variants) for which no experimental structure exists; and (2) estimating binding strength from a model of a given protein-ligand complex. To build models of inhibitor/kinase complexes, we draw from the many available structures in the PDB database as input for our comparative modelling pipeline. Our pipeline models the kinase of interest by analogy to related kinase structures, and positions the inhibitor in the active site through reference to other inhibitors. This initial model is then refined using the Rosetta modelling suite. This approach has yielded accurate models in previous test cases [1], making us confident to apply it at larger scale in these studies. Molecular dynamics (MD) based free energy calculations represent a practical, quantitative, generalizable approach to predicting the impact of clinically observed mutations on kinase inhibitor affinity. In this project we have developed and refined protocols based on both cheap end-point methods (ESMACS) and more expensive and potentially accurate alchemical methods (TIES and YANK) [2,3]. Key to refining our existing protocols is careful uncertainty quantification, avoiding the common issue that the calculated statistical error underestimates the true variation among independent experiments. As part of the project we have also investigated the utility of enhanced sampling methods for situations where mutations may alter the binding mode of ligands.

3. Machine Learning

INSPIRE leverages the models and techniques developed in the synergistic, "Exascale Deep Learning and Simulation Enabled Precision Medicine for Cancer" (CANDLE) project supported by the US Department of Energy and National Cancer Institute. Key to our approach is the use of a wide range of data sources and guided data collection from simulation and experiment in order to accurately span the known space of drug-like molecules. This space is currently not fully understood, and for that reason is difficult to span by common machine learning models. Standard virtual screening techniques have shown the difficulty of attempting to apply individual modelling to each and every molecule within a set of filters. Active learning is an approach to data sampling which weighs the model uncertainty with the computational cost to continue training the model. Active learning is used to train a surrogate molecular dynamics model, which can be run in seconds — a fraction of the time it takes to run a standard MD simulation. Using this surrogate model, it is possible to test previously intractable sets of molecules. Generative neural networks are capable of producing novel molecules with certain constraints and properties [4]. This approach allows molecules to be generated based on desirable properties and embedded in continuous representations to draw new molecules from a distribution. Our approach combines the benefits of active learning with generated molecular training data.

4. Combining MD and ML Workflows

The INSPIRE approach relies upon the creation of workflows which combine expensive but accurate free energy calculations with fast ML models. While ultimately we intend to predict the affinity of compounds to wide ranges of disease-relevant kinases and clinically identified kinase mutants, we have initially focussed on a subgoal; prioritizing the use of MD simulations to assign binding affinities to small molecule on a large set of small molecule drug candidates. Given a vast set of candidate drugs, what is the optimal ordering of simulating candidates to improve overall predictive screening performance using limited computer resource? Addressing this question is the basis of our prototype workflow (which initially targets a single kinase - Abl1), described below.

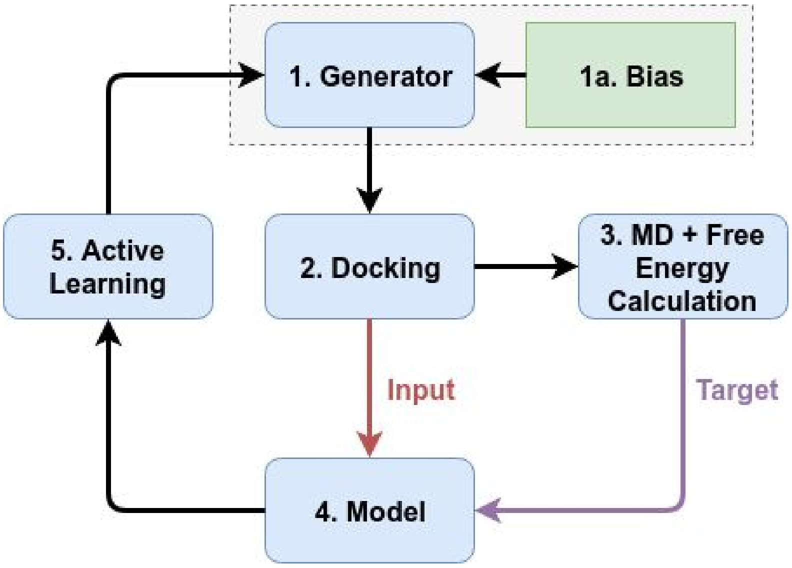

Figure 1 *Schematic overview of the INSPIRE lead identification workflow.*

1 & 2: The generation module samples from a known dataset (producing candidates as SMILES strings), but we will scale this to sampling from a variational autoencoder guided by a biasing filter. Bias is an optional module which restricts the generation module based on a particular subspace, dataset, or biochemical feature. Allowing explicit filtering using functions available in RDKit or OpenEye. 3D compound coordinates are generated from the SMILES, and docked into the pre-prepared protein conformation. The docking score is the first (and cheapest) binding strength estimate passed to the ML model.

3. The structure of the protein-ligand complex is prepared for simulation using one of our chosen free energy computations, ranging from ESMACS to the more expensive and rigorous TIES and YANK, which provides trajectories and binding free energy estimates (with associated uncertainties).

4. Model is a deep neural network which predicts the binding free energy

of a ligand. Initially the only input is the featurized SMILES string, though we will extend this to include topologies and trajectories.
5. The Active Learning module ingests the SMILES, free energy estimate and Model output and returns information to the generator either in the form of the next sample or a space to continue sampling.
Execution of a prototype workflow requires the coordination not only of the overall workflow but multi-stage pipelines of molecular simulations. To support the scalable, adaptive and automated calculation of the binding free energies concurrently with ML method on HPC resources, we are developing workflow automation tools based on the RADICAL-Cybertools middleware building block approach [5]. This allows us to to attain both workflow flexibility and performance.
The target supercomputer for the INSPIRE project is Summit (Oak Ridge National Laboratory), currently the world's fastest supercomputer. The NVIDIA Volta GPUs employed allow single OpenMM runs to generate 700+ nanoseconds of trajectory per day. However, the novel architecture of the system means tools that we have previously relied upon are currently unavailable. Consequently, our workflow has been adapted to make use of communication with a cluster running containers for docking and ligand preparation.

5. Conclusions

The INSPIRE project is preparing the way for the use of physical modelling methods in making clinical decisions in cancer therapy. Our first step is the production of a computational architecture, incorporating predictive modelling and ML, to guide the development of next-generation inhibitors able to circumvent drug resistance.

References

1. H. Zhang, *et al.*, "Targeting CDK9 Reactivates Epigenetically Silenced Genes in Cancer", Cell., 2018, 175(5) DOI: 10.1016/j.cell.2018.09.051
2. S. Wan, *et al.*, "Rapid and Reliable Binding Affinity Prediction of Bromodomain Inhibitors: A Computational Study", J. Chem. Theory Comput., 2017,13(2), DOI: 10.1021/acs.jctc.6b00794
3. https://github.com/choderalab/yank
4. T. Dimitrov, *et al.*, "Autonomous Molecular Design: Then and Now", ACS Appl. Mater. Interfaces, Article ASAP, DOI: 10.1021/acsami.9b01226
5. V. Balasubramanian, *et al.,* "RADICAL-Cybertools: Middleware Building Blocks for Scalable Science", arXiv:1904.03085