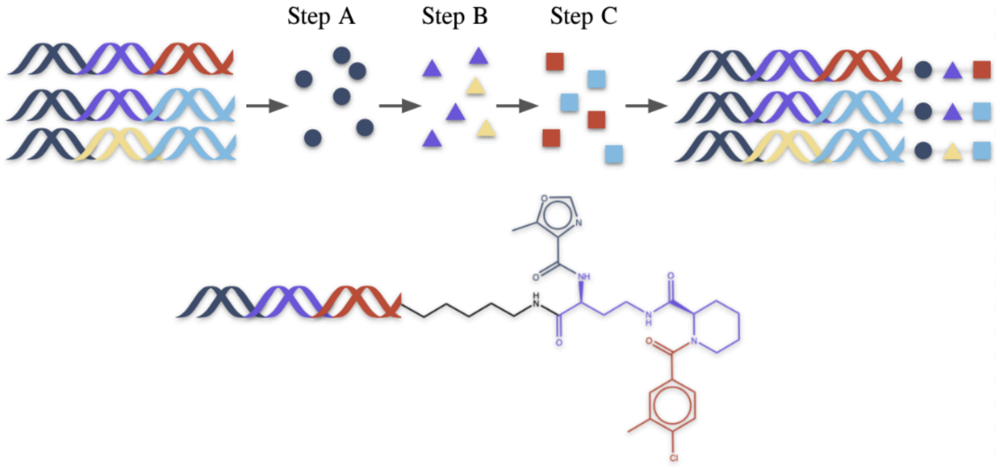
Predicting Kinase Inhibitor Activity with the KinDEL Dataset

Introduction

This project is proposed by Loka, a tech consultancy specializing in life sciences and healthcare projects. Many of Loka's clients come from the biotech industry, particularly in drug discovery, where the goal is to find and develop new therapeutic drugs. Drug development is a costly and time-consuming process, often taking up to 15 years to bring a new drug to market. At Loka, we help clients speed up this process by optimizing their AI pipelines and applying state-of-the-art models, such as large language models. This project specifically focuses on using DNA-encoded libraries, a technology we have explored in collaboration with a biotech client.

DNA-encoded libraries (DELs) are a powerful tool used in drug discovery, allowing researchers to screen vast numbers of chemical compounds simultaneously to identify those that interact with specific biological targets, such as disease-related proteins. In a DEL, each compound is linked to a unique DNA sequence that acts as a 'barcode', encoding the compound's structure and enabling high-throughput screening to quickly identify potential drug candidates.

The compounds in DELs are constructed from smaller building blocks known as synthons, which can be combined in different ways to generate diverse molecular structures (see image below). During a DEL experiment, the DNA barcodes are used to track which compounds bind to the target by measuring the enrichment of specific DNA sequences. Enrichment scores indicate how often a molecule interacts with the target compared to a control, giving an initial indication of the compound's binding strength. However, these scores can be noisy due to various factors, such as nonspecific binding or experimental variability.



To obtain more accurate measurements, candidates with high enrichment values undergo follow-up experiments to directly measure their binding affinity. A common method for this is determining the dissociation constant (Kd), which quantifies how tightly a compound binds to the target; lower Kd values indicate stronger binding interactions. These additional experiments validate the initial findings from DEL screening and help prioritize the most promising candidates for further development. Note the distinction between enrichment vs affinity: enrichment pertains to the relative increase of DEL molecules in the screening assay after several selection cycles due to their higher tendency to bind to the target, while affinity is a direct biochemical measurement of the interaction between the molecule and its target. The two concepts are adjacent, but not the same.

Dataset

The KinDEL dataset contains DNA-encoded library (DEL) data for two kinase targets: DDR1 and MAPK14. The data is available on GitHub, along with a set of benchmark models and evaluation scripts. We recommend using the subsampled version of the dataset, which includes only the top 1 million compounds with the highest enrichment scores from the library. The test set features data from real binding experiments, where binding affinities (Kd values) were measured. The dataset also includes predefined splits using two methods: a random split and a disynthon (building block) split that groups compounds based on shared building blocks. For a more detailed description of the dataset, please refer to the accompanying paper.

Project

The objective of this project is to build a machine learning model to predict enrichment scores of DEL compounds and evaluate how well it generalizes to real binding affinities (Kd). Since benchmark models are already provided in the GitHub repository, the focus will be on innovation, encouraging students to explore novel architectures and methods for representing the molecules. We recommend that students leverage the existing codebase developed by the authors of the paper.

Some research questions to guide your solution:

- Can you encode the structure/geometry of the molecules in your model?
- How could you use the information from different building blocks that make a molecule to inform your model?
- Could you use the SMILES strings directly in your model?
- Can you take advantage of biological large language models in your solution?
- Is your model equally effective for both data splits?
- Can you generalize across targets?

We will follow the evaluation strategy outlined in the paper, comparing the predicted enrichment scores with experimental Kd values for the molecules in the held-out test set. Using Spearman correlation as the primary metric, we will assess how well each model ranks the compounds by their binding affinity.

Papers

Chen, B., Danel, T., McEnaney, P. J., Jain, N., Novikov, K., Akki, S. U., ... & Watts, R. E. (2024). KinDEL: DNA-Encoded Library Dataset for Kinase Inhibitors. arXiv preprint arXiv:2410.08938.

Ross, J., Belgodere, B., Chenthamarakshan, V., Padhi, I., Mroueh, Y., & Das, P. (2022). Large-scale chemical language representations capture molecular structure and properties. Nature Machine Intelligence, 4(12), 1256-1264.

Wang, K., Zhou, R., Tang, J., & Li, M. (2023). GraphscoreDTA: optimized graph neural network for protein–ligand binding affinity prediction. Bioinformatics, 39(6), btad340.