



*Dr. María Rodríguez Martínez
Technical Leader of Systems Biology
IBM Research – Zurich
 Säumerstrasse 4
 8803 Rüschlikon
 Switzerland*

Zürich, December 30, 2022

Report of the thesis presented by Krzysztof Koras

To whom it may concern,

My name is Dr María Rodríguez Martínez, and I am the Technical Leader of Computational Systems Biology at IBM Research Europe (Switzerland). This is a report on the thesis presented by Krzysztof Koras about his PhD work.

Krzysztof's work focused on the development of machine-learning (ML) approaches to tackle the problem of predicting drug sensitivity in cancer cell lines, an important therapeutic problem. From the ML point of view, the challenge is to predict the activity of a compound in a cell line by integrating disparate types of data regarding the cell lines, e.g. gene expression information and mutation information, and the compounds, e.g. putative drug targets or inhibitory profiles (in the case of kinase inhibitors). The challenge here is that data is high-dimensional while the data cohorts are typically smaller. The lack of transparency of most ML and deep learning models is another major concern that limits their usability in clinical settings, where model understanding is necessary to generate trust in the model's predictions.

The thesis is structured in 7 chapters. Two chapters present an introduction to cancer biology and ML models with an application to drug sensitivity prediction. Both chapters are well-written and present a concise summary of the state-of-the-art in these areas. They both provide the necessary background to position and understand Krzysztof's work.

Chapters 4, 5 and 6 present Krzysztof's work. Each one describes work that led to a publication where Krzysztof is the first author. The first piece of work focuses on the problem of feature selection in drug sensitivity models. While integrating all available data would maximise the amount of information a model has available to make a prediction, such a model would be untrainable with existent datasets. Krzysztof performs a systematic investigation to identify the most informative set of features using two approaches, one based on biological prior knowledge, e.g. compound drug targets and target pathways, and a data-driven approach that automatically selects the most informative features. The analyses presented show that the optimal feature selection strategy must consider the context, for instance, cell line mutation information can be highly predictive in smaller models, but it becomes redundant in larger models trained on genome-wide gene expression features.

Chapter 6 describes an interpretable recommender system for the prediction of the efficacy of kinase inhibitors. Recommender systems are a class of information filtering system that provides recommendations to a user. Amazon, providing recommendations of additional items a user might want to purchase, or Netflix, suggesting additional movies a user might like to watch, are two well-known examples. Here, Krzysztof explores the idea of building a recommender system based on deep learning models to predict kinase inhibitor efficacy. Namely, the model exploits two autoencoders to encode the information about the cell line and drug features, and a feed-forward neural network to combine them and make a prediction. Krzysztof demonstrates that the model achieves good performance and, importantly, model interpretability.

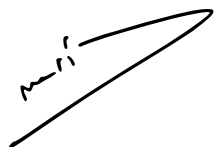
Chapter 7 explores the possibility of enhancing the prior architecture with generative capabilities. Namely, the autoencoder used to encode information about the compounds is now replaced by a variational autoencoder, which can be used to generate new data given some prior distribution about the latent space. Krzysztof explores the possibility of using a non-standard prior, concretely a Gaussian Mixture Model (GMM), which assumes that data comes from a mixture of Gaussians that might represent structurally different groups. This is an intelligent way to enforce clustering in the latent space, and might prove to be a very useful approach to generate data in situations where we expect the presence of heterogeneous groups. Once again, Krzysztof shows that the model achieves good performance and, importantly, enables the generation of new compounds with similar inhibitory profiles to a target cluster of compounds.

Interpretability is a major concern in this thesis, and Krzysztof thoroughly describes approaches to extract biological insight from the built ML models by examining the features selected for each predictive task. This is a nice and much-needed addition to this field, where too many black-box models are being developed that demand blind trust from the user.

Of course, all work can be extended with further discussions. I will be looking forward to discussing with Krzysztof during his dissertation additional extensions to the work presented here, such as probabilistic methods to include uncertainty about the therapeutic compounds. It has been shown that many putative targets of drug compounds are incorrect, and nevertheless, ML models built on this information still achieve good accuracy. I will be keen to hear Krzysztof's opinion on why this might be. Similarly, biological data, e.g. omics profiles of cell lines, are typically noisy and plagued with missing values. How can we adapt ML models to work with these data despite their limitations?

Besides these collegial discussions, the thesis is well written, the work is solid, and of high quality and originality. As the first author of 3 publications, Krzysztof has demonstrated scientific maturity and independence and deserves in my opinion to be awarded the title of doctor.

Sincerely,

A handwritten signature in black ink, appearing to read 'María', written over a long, sweeping horizontal line that extends across the width of the signature.

Dr. María Rodríguez Martínez