Co-funded by the Horizon 2020 Framework Programme of the European Union

Grant agreement No. 671555

ExCAPE
Exascale Compound Activity Prediction Engine

Future and Emerging Technologies (FET)

Call: H2020-FETHPC-2014
Topic: FETHPC-1-2014
Type of action: RIA

Deliverable D1.8

Development report

Due date of deliverable: 31.08.2018
Actual submission date: 28.08.2018

Start date of Project: 1.9.2015 Duration: 36 months

Responsible Consortium Partner: JP
Name of author(s) and contributors: Vladimir Chupakhin (JPNV)
Revision: V2
Internal reviewer(s): Andreas Mayr (UL)

Project co-funded by the European Union within the Horizon 2020 Framework Programme (2014-2020)		
Dissemination Level		
PU	Public	**PU**
PP	Restricted to other programme participants (including the Commission Services	
RE	Restricted to a group specified by the consortium (including the Commission Services)	
CO	Confidential, only for members of the consortium (including the Commission Services)	

Document revision tracking

This page is used to follow the deliverable production from its first version until it has been reviewed by the assessment team. Please give details in the table below about successive releases.

Release number	Date	Reason of this release and/or validation	Dissemination of this release (task level, WP/ST level, Project Office Manager, Industrial Steering Committee, etc)
V1	28/08/2018	First draft	None
V2	19/09/2018	Update after final meeting	Public

Glossary

NCV	Nested cross-validation
DNN	Deep neural network
SNN	self-normalizing neural network
GFA	group factor analysis
BMFPP	Bayesian matrix factorization with posterior propagation,
MICP	Mondrian inductive Conformal Prediction
VAP	Venn-Abers predictions
BPMF	Bayesian Matrix Factorization with Side Information
SMURFF	Package for BPMF with side information
L1000	L1000 gene expression
HCI	High content imaging

Link to Tasks

Task number	Work from task carried out	Deviations from task technical content, with motivation and summary of impact
T1.1	Coordination of the interactions between industry and academy partners	None

Table of contents

1 Executive summary

This deliverable summarizes algorithmic impact that methods developed in ExCAPE project had on industry challenges. Most of them were successfully addressed and produced pipelines that were tested on pharma industry data. Some of the new challenges related to previous were also underlined.

2 Introduction – Aim

Six main industry challenges relevant to pharma industry data were described in the ExCAPE proposal and more details in deliverable D1.1. Namely, (1) large, sparse and highly imbalanced data sets; (2) complexity and intrinsic noise of the different types of biological data quantitively describing protein-ligand interaction; (3) lack of widely accepted data standards; (4) improvement of chemogenomics models using multi-task information compared to single task models; (5) heterogenous input space, and (6) model and/or results interpretability. This deliverable will address them in more details in section 3.

3 Algorithmic challenges of pharma industry addressed by algorithm developed in ExCAPE

3.1 Large, noisy and highly imbalanced data

Two multi-task algorithms were successfully tested on the industry like dataset (ExCAPEdb): Bayesian matrix factorization with side information (SMURFF) and deep learning models. Both showed that usable and successful models could be developed using datasets with around 1 million compounds, with a significant sparsity of the data with < 10% of all possible interactions known and highly imbalanced toward one class of the biological activities (active vs. inactive). Challenge of imbalanced and sparse data was successfully resolved with those two algorithms. In one case batch processing nature of deep learning helped to solve the imbalanced issue, while in matrix factorization case global block-wise optimization using matrix decomposition algorithms helped with data imbalance and sparsity. A successful application example showed that testing DNN models developed on ExCAPEdb on Astra Zeneca internal dataset gave an average ROC AUC of 0.7 (details see report D3.18) .

Two additional methods for matrix factorization did not show improvement compared to baseline method for subset of ExCAPEdb dataset and artificial data: sparse group factor analysis (D1.3) and distributed Bayesian matrix factorization (D1.4) with Markov chain Monte Carlo methods and random projections.

Issue with intrinsic noise was addressed implicitly in all studies algorithms.

3.2 Data standards

Despite the lack of widely accepted data standards for compound preparation for modeling, it was not an issue. All partners agree on the common set of rules for compound standardization that was followed later for all project related activities.

3.3 Heterogenous input space

Chemical compounds can be represented using different chemical descriptors: 2-dimensional representation like fragment-based or graph, 3-dimensional – shape, pharmacophore, etc. All those descriptors encode only specific property of the chemical compound, while it is assumed that the best descriptor is a combination of several ones. Another type of descriptors is so-called experimental descriptors derived from the real biological experiment: like high-content-imaging or gene expression analysis, which found already a valuable resource.

The successful combination of them requires a different type of handling to extract only relevant features for supervised modeling. While this drawn significant attention in a domain of image and video analysis [1],[2], this is still underdeveloped for computational chemogenomics. In ExCAPE we tested GFA that gave us mixed results, showing that the combination of the different data sources did not bring improvement on cancer cell line datasets (details see report D3.10). For large internal datasets of 200 000 compounds, a combination of the gene expression and chemical compounds descriptors used in SMURFF (matrix factorization with side information) gave only very little or no improvement in comparison with the models developed using only chemical descriptors, this part of work was not reported under ExCAPE project (internal research). Application of the Self-Normalizing Neural Networks on the same dataset showed that combination of the biological signatures (L1000, HCI) to chemical descriptors (ECFP6, chem2vec) increased the quality for 61% of the targets, while for 39% of compounds model quality decreased. This trigger further questions: How to correctly combine different data sources to get the best out of each of them? How to correctly estimate a benefit of additional data using non-standard model quality metrics, especially in the case when the model quality increased/decreased non-significantly or remained the same?

3.4 Model interpretation

Interpretation of the developed models was attempted to solve using conformal and probabilistic Venn-Abers predictions (D1.4 and D1.6). The results allow relatively good interpretation thou required additional post-processing to be used for non-experienced end-user, a chemist or a biologist. While conformal, Venn-Abers and Platt scaling (D1.9) serve as prediction calibration algorithms all of them were designed for single tasks, and comparison of calibrated values of one target vs. another will provide incorrect ranking (Chupakhin et al., unpublished). Calibration for a combination of the results of different models was addressed in D1.7.

3.5 New challenges?

One of the significant issues was an initial setup of the supervised modeling that requires nested cross-validation (NCV) to select the best performing hyperparameter and to evaluate the quality of the models for individual targets. This require an initial split of the compounds into train, test, and validation, in all cases, we have to extract some part of the data for evaluation, that in turn influence final model as all estimates is done only on the part of the data, this can be biased. Another issue is model quality estimation that for NCV setup is not always possible as individual targets do not always represent in all three sets (train, test,

validation). This, in turn, can be solved by excessive splits to find the best ones that will include all individual targets. For large scale models, this is hardly feasible as it would require significant compute time and non-practical per se. Ideal algorithm needs to include local and global model quality estimation as an intrinsic property.

Completely unexpected challenge is usage of the developed models in prediction mode: for example in case of SMURFF package predictions were unexpectedly slow, and still require significant optimization on C++ side, while Python interface developed by Janssen postdoctoral researcher is slow and original one is even slower. For DNN models, prediction on large virtual chemical spaces can still be a challenge and probably distributed prediction model of the developed models will be of a great use.

4 Conclusion

Deep learning and matrix factorization machine learning algorithm can be successfully used for computational chemogenomics supervised model development, means applicable for large and heavily imbalanced datasets with a high level of data scarcity, a low level of intersections between different modeled activities (targets). In Janssen we successfully developed and applied SMURFF for modeling on internal data for ~2 million compounds on chemical descriptors, predictions were further used for drug design and discovery both as individual predictions per target and as a whole vector, so-called biosignature. In AZ, the Conformal prediction, Venn-Abers prediction have been deployed to give confidence estimation for in-house in-silico model predictions. The machine learning models built on ExCAPE dataset have been used to give prediction on AZ in-house data for evalution.

Overall most of the challenges underlined in the D1.1 were addressed, means a possible algorithmic solution was proposed, or addressed and successfully applied to ExCAPEdb or industry data, that is summarized in table 1. Confidence estimation was not addressed in neural networks (DNN, SNN) and heterogeneity of the data was not addressed for MICP, for all other challenges some form of a solution was proposed. Due to various reasons, only several algorithms were also deployed on pharma industry data or ExCAPEdb: neural networks and SMURFF, while SNN has some limitations on the processing of the large datasets and algorithm need probably further optimization.

	imbalance	sparsity	volume	heterogeneity	confidence
DNN	2	2	2	1	0
SNN	2	2	1	1	0
SMURFF	2	2	2	1	1
GFA	1	1	1	1	1
BMFPP	1	1	2	1	1
MICP	2	1	2	0	2
VAP	1	1	2	1	2

Table 1. Summary of the challenges (see details in section 4). Notations used in the table: 2 (green) – challenge was addressed and successfully applied on industry or industry-like dataset (ExCAPEdb), 1 (yellow) - challenge was addressed but solution was either not optimized to be run on industry scale or other reasons (lack of computer time, prioritization of other algorithms, etc), 0 (red) – challenge was not addressed. Abbreviations used: *DNN - Deep neural network; SNN - self-normalizing neural network; BPMF - Bayesian Matrix Factorization with Side Information; SMURFF - package for BPMF with side information; GFA - group factor analysis; BMFPP - Bayesian matrix factorization with posterior propagation; MICP - Mondrian inductive Conformal Prediction; VAP - Venn-Abers predictions;*

5 Submission Delay

There was no delay in deliverable submission.

6 Bibliography

[1] N. Srivastava and R. Salakhutdinov, “Multimodal Learning with Deep Boltzmann Machines,” *Adv. neural Inf. Process. Syst.*, pp. 2222–2230, 2012.

[2] J. Ngiam, A. Khosla, M. Kim, J. Nam, H. Lee, and A. Y. Ng, “Multimodal Deep Learning,” *Proc. 28th Int. Conf. Mach. Learn.*, pp. 689–696, 2011.