HPC-as-a-Service for Driving Artificial Intelligence for Drug Discovery

Václav Svatoň, Vojtěch Cima, Jan Martinovič
{vaclav.svaton, vojtech.cima, jan.martinovic}@vsb.cz
IT4Innovations, VŠB – Technical University of Ostrava
Czech Republic

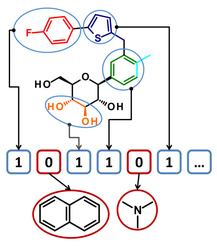
Nina Jeliazkova, Vedrin Jeliazkov, Luchesar Iliev
{jeliazkova.nina, vedrin.jeliazkov, luchesar.iliev}@gmail.com
IDEAconsult Ltd.
Bulgaria

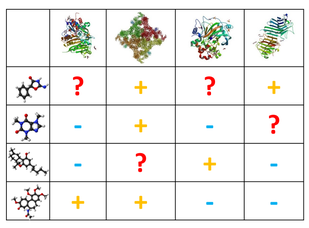
Vladimir Chupakhin
vchupakh@its.jnj.com
Janssen Pharmaceutica NV
Belgium

Motivation

Real-world pharma industry applications often encompass end-to-end data processing pipelines composed of a large number of interconnected tasks of various granularity. Most of the common tasks in the prediction of activity and toxicity of chemical compounds consist of several typical steps, such as compiling, cleaning and combining datasets, feature calculation, feature selection, model training and validation and applying models to predict properties of new compounds. Building and executing such pipelines on HPC systems can be challenging tasks for domain specialists who do not have sufficient level of experience in distributed computing. Therefore, we introduce a drug discovery web platform that enables large-scale machine learning applications being executed on supercomputing facilities via specialized middleware.

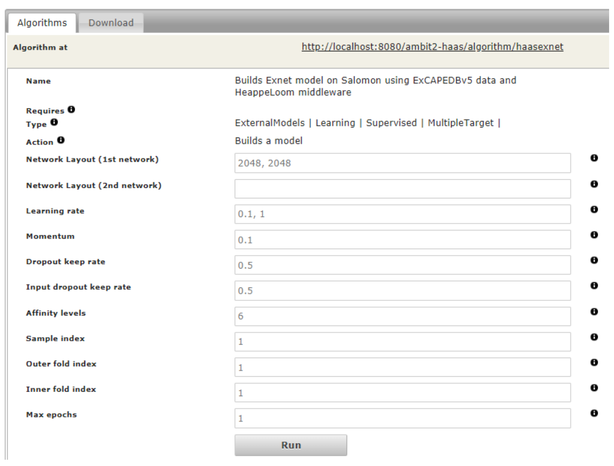
Machine Learning for Pharma Industry



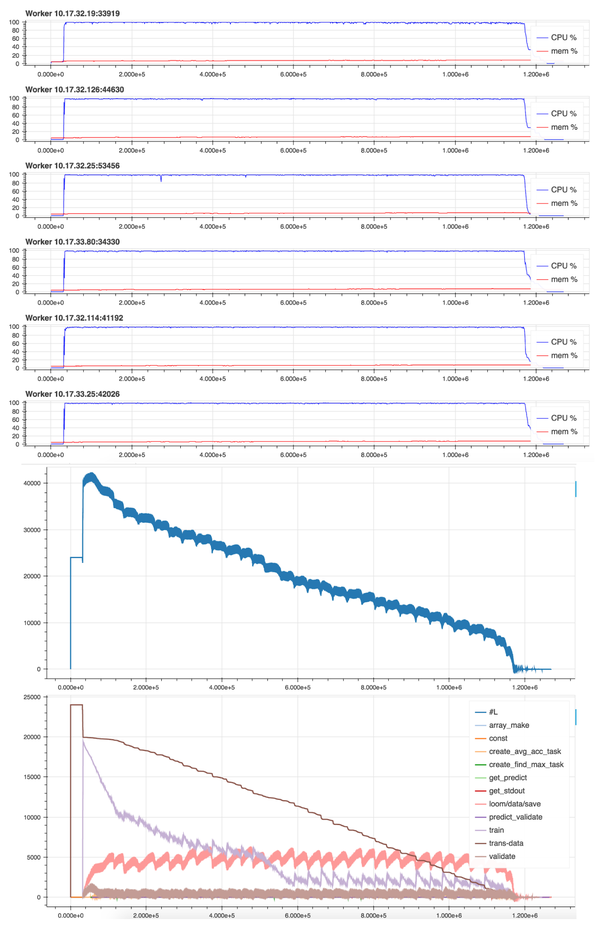


Pharma companies collected significant amount of protein-ligand interactions forming so-called chemogenomics matrix: interactions between compounds and proteins, but this matrix is very sparse, less 1% of this matrix is filled. Predictive modeling can help to fill this matrix using classification or regression model, predictions in turn are used to speed-up drug design and development process, can help to cut cost and also reduce animal use. While machine learning is widely used on every step of the drug design and discovery process it is still a hurdle to use it on big data, taking into account all modeling steps needed: hyperparameter search, model and predictions storage, etc.

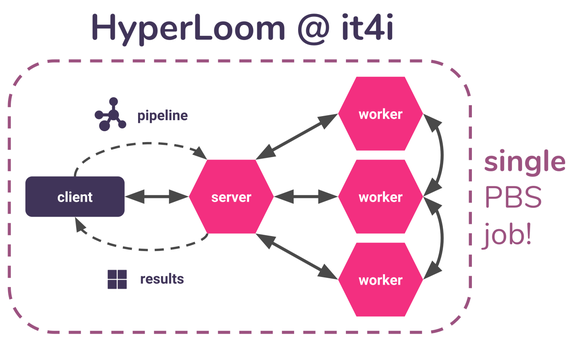
Web GUI

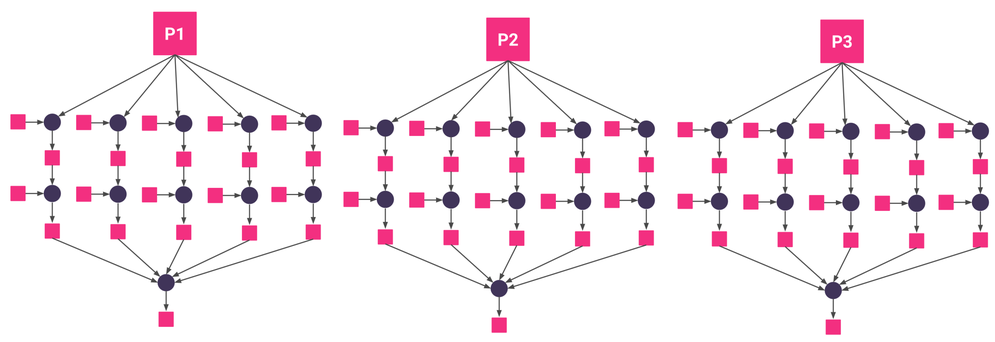


Monitoring



Building and Executing Scientific Pipelines

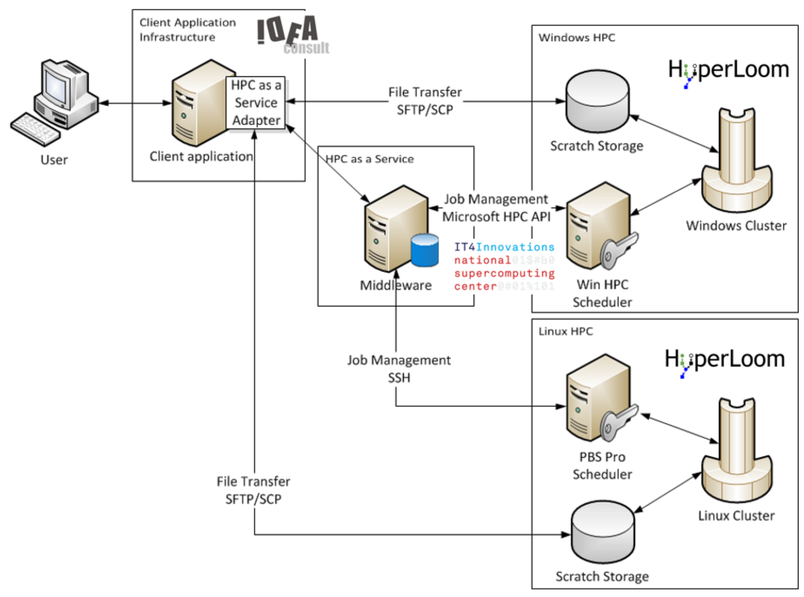




HyperLoom is an open-source platform for an efficient definition and execution of scientific pipelines in distributed environments. HyperLoom enables to chain large number of computational tasks into a complex end-to-end data processing pipelines using a simple Python interface as a gateway to the high-performance backend of HyperLoom.

- In-memory data storage
- Reactive scheduling
- Direct worker-to-worker communication
- Powerful task abstraction
- Performance visualization

Platform Architecture



- Web graphical user interface with REST API
- Providing HPC capabilities as a service to client applications and their users
- Unified middleware interface for different operating systems and schedulers
- Authentication and authorization to provided functions
- Monitoring and reporting of executed jobs and their progress
- Current information about the state of the clusters
- Job accounting and job reporting for user or user group
- Secure data migration between different jobs
- Prepared job templates for drug discovery pipelines (modelling, prediction, statistics)

Acknowledgements

This work was supported by The Ministry of Education, Youth and Sports from the National Programme of Sustainability (NPS II) project "IT4Innovations excellence in science - LQ1602", by the IT4Innovations infrastructure which is supported from the Large Infrastructures for Research, Experimental Development and Innovations project "IT4Innovations National Supercomputing Center - LM2015070" and by the ExCAPE project - the European Union's Horizon 2020 research and innovation programme under grant agreement No. 671555.