Automating model building in ligand-based predictive drug discovery using the Spark framework

Abstract
Automation of model building enables new predictive models to be generated in a faster, easier and more straightforward way once new data is available to predict on. Automation can also reduce the demand for tedious bookkeeping that is generally needed in manual workflows (e.g. intermediate files needed to be passed between steps in a workflow). The applicability of the Spark framework related to the creation of pipelines for predictive drug discovery was here evaluated and resulted in the implementation of two pipelines that serves as a proof of concept. Spark is considered to provide good means of creating pipelines for pharmaceutical purposes and its high level approach to distributed computing reduces the effort put on the developer compared to a regular HPC implementation.

Keywords
Machine learning, predictive drug discovery, Spark, Big Data, Hadoop, pharmaceutical bioinformatics, pipelining, cloud computing, OpenStack