

This is a preprint of the article published in J Comput Aided Mol Des. 2008 Jun-Jul;22(6-7):431-40. Epub 2008 Feb 14.

Abstract

In this article, we present an automatic model generation process for building QSAR models combined with Gaussian Processes, a powerful machine learning modeling method. We describe the stages of the process that ensure models are built and validated within a rigorous framework: descriptor calculation, splitting data into training, validation and test sets, descriptor filtering, application of modeling techniques and selection of the best model. We apply this automatic process to data sets of blood-brain barrier penetration and aqueous solubility data sets and compare the resulting automatically generated models with 'manually' built models using external test sets. The results demonstrate the effectiveness of the automatic model generation process for two types of data sets commonly encountered in building ADME QSAR models, a small set of *in vivo* data and a large set of physico-chemical data.

You can read a copy of this article as a [PDF](#) file