A systematic DNN-based QSPR modeling methodology for rapid and reliable prediction on flashpoints of chemicals

Huaqiang Wen[1], Yang Su[2], Zihao Wang[3], saimeng Jin[1], Jingzheng Ren[4], Weifeng Shen[1], and Mario Eden[5]

[1]Chongqing University
[2]Chongqing University of Science and Technology
[3]Max Planck Institute for Dynamics of Complex Technical Systems
[4]The Hong Kong Polytechnic University
[5]Auburn University

May 26, 2021

Abstract

Quantitative structure-property relationship (QSPR) studies based on deep neural networks (DNN) are receiving increasing attention due to their excellent performances. A systematic methodology coupling multiple machine learning technologies is proposed to solve vital problems including applicability domain and prediction uncertainty in DNN-based QSPRs. Key features are rapidly extracted from plentiful but chaotic descriptors by principal component analysis (PCA) and kernel PCA. Then, a detailed applicability domain (AD) is defined by K-means algorithm to avoid unreliable predictions and discover its potential impact on uncertainty. Moreover, prediction uncertainty is analyzed with dropout-embedded DNN by thousands of independent tests to assess the reliability of predictions. The prediction of flashpoint temperature is employed as a case study demonstrating that the model accuracy is remarkably improved comparing with the referenced model. More importantly, the proposed methodology breaks through difficulties in analyzing the uncertainty of DNN-based QSPRs and presents an AD correlated with the uncertainty.

Hosted file

`2021.5.19_wen.docx` available at https://authorea.com/users/415953/articles/523705-a-systematic-dnn-based-qspr-modeling-methodology-for-rapid-and-reliable-prediction-on-flashpoints-of-chemicals