From Descriptors to Predicted Properties: Experimental Design by Using Applicability Domain Estimation

Stefan Brandmaier, Sergii Novotarskyi, Iurii Sushko and Igor V. Tetko

Helmholtz-Zentrum München — German Research Centre for Environmental Health (GmbH), Institute of Structural Biology, Munich, Germany; eADMET GmbH, Neuherberg, Munich, Germany

Summary — The importance of reliable methods for representative sub-sampling in terms of experimental design and risk assessment within the European Registration, Evaluation, Authorisation and Restriction of Chemicals (REACH) system is crucial. We developed experimental design approaches, by utilising predicted properties and the ‘distance to model’ parameter, to estimate the benefits of certain compounds to the quality of a resulting model. A statistical evaluation of four regression data sets and one classification data set showed that the adaptive concept of iteratively refining the representation of the chemical space contributes to a more efficient and more reliable selection in comparison to traditional approaches. The evaluation of compounds with regard to the uncertainty and the correlation of prediction is beneficial, and in particular, for regression data sets of sufficient size, whereas the use of predicted properties to define the chemical space is beneficial for classification models.

Key words: bagging, distance to model, representative subset selection.

Address for correspondence: Stefan Brandmaier, Helmholtz-Zentrum München — German Research Centre for Environmental Health (GmbH), Institute of Structural Biology, IngolstädterLandstrasse 1, Neuherberg D-85764, Munich, Germany.
E-mail: stefan.brandmaier@gmail.com

Introduction

The sampling of representative compounds delivers a sensitive subset of a defined chemical space. When used for experimental design, it contributes to financial and time efficiency, as it permits the calculation of highly predictive models at a low experimental cost (1). Furthermore, it is relevant for several other tasks in chemo-informatics and computational chemistry, such as drug design (2) and risk assessment (3).

Most applications in chemistry work in the descriptor space (4). After applying an orthogonal transformation by employing a principal component analysis (PCA) to increase the dimensionality of the chemical space, selection algorithms, such as the D-Optimal criterion (5), the Kennard–Stone algorithm (6) or the Most Descriptive Compound (MDC) selection (7), are applied. Bayesian approaches (8–10) that take preliminary information into account, or adaptive approaches that refine the representation of the chemical space in a step-wise manner, are rarely used. Our recent studies investigate iteratively optimised variables to span the search space for experimental design, by combining a partial least squares technique with a dissimilarity selection (11), and a simple descriptor selection with a similarity selection, to increase the efficiency of the experimental design. Having said that, both approaches still work on a representation of the chemical space that uses descriptors as the basis for the selection of compounds.

In this study, we propose a novel adaptive projection that moves from a descriptor-based construction of the chemical space toward a predicted property-based view. We investigate three strategies for experimental design — one of them uses the predicted properties to define the chemical search space, while the others use the concept of the ‘distance to model’ (DM) parameter, suggested by Tetko et al. (12–14), to estimate the uncertainty of prediction for each compound within a data collection. We are not aware of other studies in computational chemistry that use this parameter for compound selection.

The DM-based approaches decide whether to test a compound either exclusively on the basis of the prediction uncertainty, or they combine this parameter with a compound’s hypothetical contribution in decreasing the prediction uncertainty of other relevant compounds. The estimation of this contribution is based on the correlation in ensemble predictions, a concept that is the basis of Associative Neural Networks (ASNN; 15), and which is also used in the ASNN Library mode to make local corrections (16).

We statistically evaluated the new approaches on four regression data sets, each with a different endpoint, and on one classification data set. Furthermore, the performances of the newly developed approaches were compared to the performances