nature portfolio

https://doi.org/10.1038/s41557-021-00716-z

Supplementary information

Best practices in machine learning for chemistry

In the format provided by the authors and unedited

Checklist for reporting and evaluating machine learning models	
1. Data sources	
1a. Are all data sources listed and publicly available?	
1b. If using an external database, is an access date or version number provided?	
1c. Are any potential biases in the source dataset reported and/or mitigated?	
2. Data cleaning	
2a. Are the data cleaning steps clearly and fully described, either in text or as a code pipeline?	
2b. Is an evaluation of the amount of removed source data presented?	
2c. Are instances of combining data from multiple sources clearly identified, and potential issues mitigated?	
3. Data representations	
3a. Are methods for representing data as features or descriptors clearly articulated, ideally with software implementations?	
3b. Are comparisons against standard feature sets provided?	
4. Model choice	
4a. Is a software implementation of the model provided such that it can be trained and tested with new data?	
4b. Are baseline comparisons to simple/trivial models (for example, 1-nearest neighbour, random forest, most frequent class) provided?	
4c. Are baseline comparisons to current state-of-the-art provided?	
5. Model training and validation	
5a. Does the model clearly split data into different sets for training (model selection), validation (hyperparameter optimization), and testing (final evaluation)?	
5b. Is the method of data split (data splitting (for example, random, cluster- or time-based splitting, forward cross-validation) clearly stated? Does it mimic anticipated real-world application?	
5c. Does the data splitting procedure avoid data leakage (for example, is the same composition present in the training and test sets)?	
6. Code and reproducibility	
6a. Is the code or workflow available in a public repository?	
6b. Are scripts to reproduce the findings in the paper provided?	