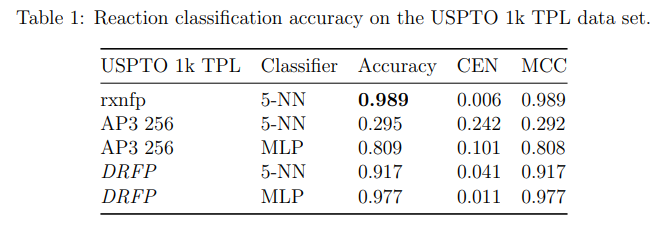
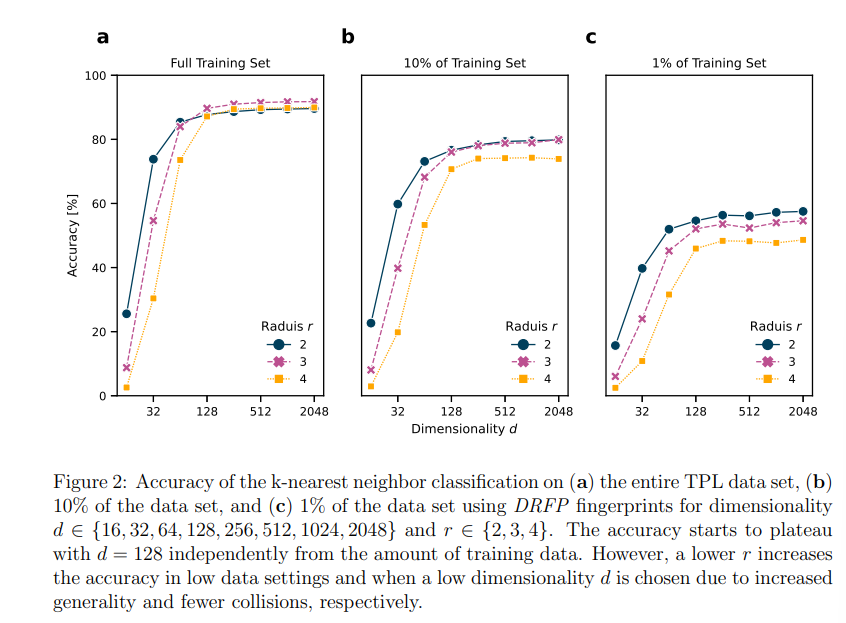
# Reaction Classification and Yield Prediction using the Differential Reaction Fingerprint DRFP

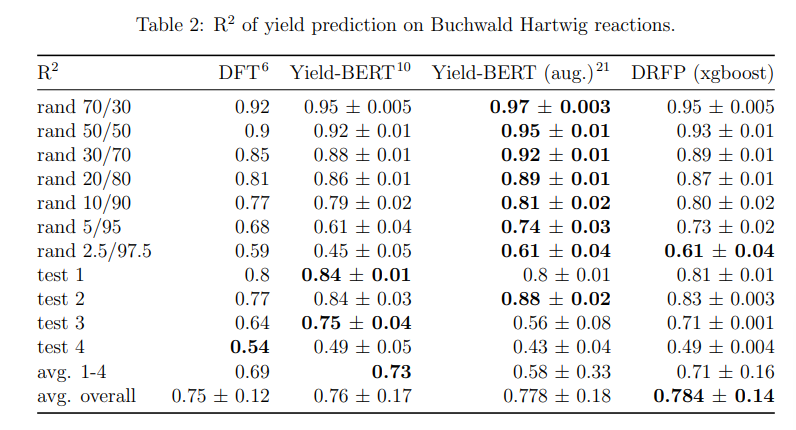
* Abstract: - The fingerprints required large training data sets, they are inherently biased, and are based on complex deep learning architecture.
* The researcher presents the differential reaction fingerprint *DRFP*. The *DRFP* algorithm takes a reaction SMILES as an input and creates a binary fingerprint based on the symmetric difference of two sets containing the circular molecular n-grams generated from the molecules listed left and right from the reaction arrow, respectively, without the need for distinguishing between reactants and reagents.
* Introduction: - an important outcome of a chemical reaction is its yield, the percentage of successfully converted reactants into the desired product.
* The availability of large data sets and the resurgence of artificial neural networks (ANN), deep learning-based learned fingerprints have been introduced as an alternative to earlier methods, outperforming them by considerable margins.
* It has challenges also like training a learned fingerprint requires large amounts of data of acceptable quality and must be retrained when new data becomes available, posing a challenge to accessibility and reproducibility.
* The researchers’ differential reaction fingerprint (*DRFP*) search and categorized the reaction as well as predict the yield.
* The reaction fingerprints *DRFP* borrows the categorization of circular substructures from a molecule and the subsequent hashing of their SMILES representations from the chemical fingerprints ECFP and MHFP, respectively.
* The researchers include three additional steps in REACTANTS>AGENTS>PRODUCTS, these are (I) the agents are added to the reactants, resulting in the representation REACTANTS+AGENTS>>PRODUCTS; (II) molecules on each side of the reaction representation are processed individually, resulting in two sets of SMILES *R* and *P*; (III0 the symmetric difference of the two sets *S = RP* is taken, hashed using an arbitrary hash function with a sufficiently low collision probability (BLAKE2), and then further hashed into a fix-length binary vector using *h(k)* = *k* mod *d*, where *k S*, and *d* is the desired dimensionality of the fingerprint.
* The fingerprint requires an unannotated, non-atom-mapped reaction SMILES as input and embeds this molecular representation from reaction SMILES space into an arbitrary low dimensional binary metric space through set operations and subsequent mod hashing.
* The researchers show that their model, based on a simple set operation and hashing scheme, can outperform both deep learning-based learned fingerprints and physics-based descriptors in yield prediction tasks.
* Results and Discussion: - Reaction Classification: - the reaction classification was done using KNN classifier. Initially, DRPF was evaluated on the USPTO 1k TPL set using different configurations like radius r {2, 3, 4} and dimensionality d {16, 32, 64, 128, 256, 512, 1024, 2048}.
* The accuracy increases strongly between *d* = 16 to *d* = 128, while only increasing slightly from *d* = 256 to *d* = 2048. The *r* = 2 variant performs significantly better than *r* {3, 4} for d {16, 32}.
* The *d* = 256, and *r* = 3 performs better than both the other variants.
* If the training set is reducing from 10% to 1% then it also leads to a better relative performance of the *r* = 2 variant across all dimensions *d*.
* The *r* = 2 performance better in low data settings, but the *r* = 3 variant performed best in the case of the complete training set of high *d*, therefore the *r* = 3 and *d* = 2048 variant is chosen for the benchmark, including reaction yield predictions.



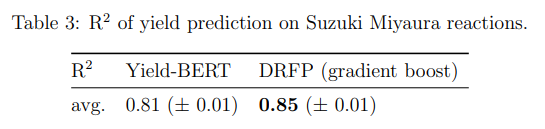
* Evaluating the kNN classification benchmark on the TPL data set, *DRFP* outperforms the structures-based fingerprint AP3 256 by a factor of 3.1 and reaches 93% of the performance of the learned fingerprint rxnfp. If the kNN classifier replace with a simple multilayer perceptron (MLP), then the *DRFP* reaches 9% of the performance of the rxnfp.
* Reaction Yield Prediction: - *DRFP* outperforms Yield-BERT, an augmented version of Yield-BERT, as well as a DFT-based method, in a yield prediction task on a data set of Buchwald Hartwig reactions.
* It also outperforms rxnfp in yield prediction of USPTO reaction data and a data set of Suzuki Miyaura reactions.
* Hyperparameter optimization was performed on five random splits (70/30). The resulting performance (*R2*) is then compared to the density functional theory (DFT) based fingerprint with a random forest regressor by Ahneman, Yield-BERT, an extension of the learned rxnfp fingerprint with a regression layer, and an augmented variant of the latter.

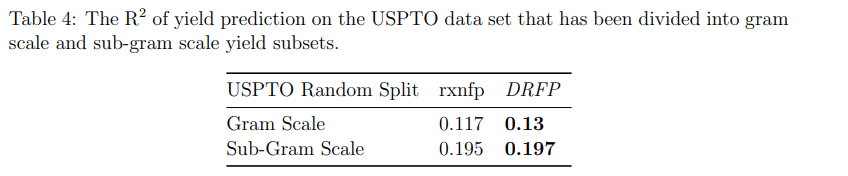


* *DRFP* performs better on the random splits than the DFT-based fingerprint with random forests and Yield-BERT but is outperformed by the augmented Yield-BERT by a narrow margin.

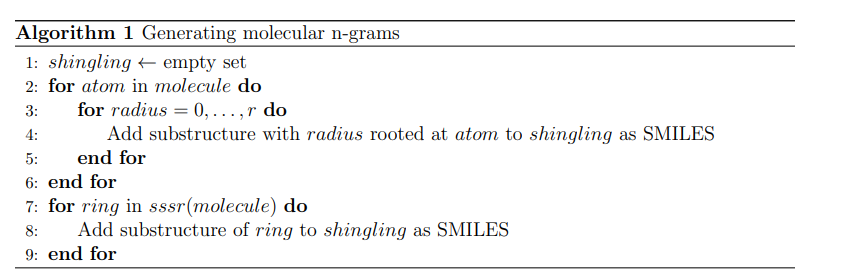


* The average of all 11 tests, *DRFP* performs best.





* The *DEFP* reaches a compelling performance in yield prediction using a gradient boosting regressor that does not require hyperparameter tuning between different sets.
* Conclusion: - *DRFP*, based on a simple 4-step process comprised of extracting circular n-grams, XORing, hashing, and folding.
* Methods: - Molecular n-grams: - Molecular n-grams are generated from SMILES using the RDKit library. The radius r, iterate over the heavy atoms in an input molecule and extract sub-SMILES centred on each atom with radii 0 to r, where a radius of 0 is the single central atom.
* Gradient Boosting: - for the gradient boosting, they use the python library xgboost, and for hyperparameter tuning was carried out on the rand 70/30 set of the Buchwald-Hartwing reaction data set.
* Algorithm: -



* For the test, 10% of the training data were randomly selected as the validation set an removed from the training set. The validation data sets were used as evaluation sets for early stopping.
* k-Nearest Neighbours Classifier: - the kNN has the value of k = 5.
* Multilayer Perceptron Classifier: - the *DRFP* + 5-NN classifier, *DRFP* + multilayer perceptron (MLP) classifier was applied to the USPTO 1k TPL data set.
* It has the input vector of 2048, a dense hidden layer of size 1664 and a tanh activation function, and a dense output layer with a softmax activation function.
* The loss function was sparse categorical cross-entropy. Adam was used as an optimizer. The model was trained over 10 epochs with a batch size of 64.
* For evaluation of AP3 256, the number of units in the hidden layer was changed to 1024, and the model was trained for 100 epochs.