When yield prediction does not yield prediction and approaches to fix it

Varvara Voinarovska[1,2], Mikhail Kabeshov[2], Samuel Genheden[2], Dmytro Dudenko[3] and Igor Tetko[4]

[1]Technical University of Munich, Germany; TUM Graduate School, Faculty of Chemistry
[2]Molecular AI, Discovery Sciences, AstraZeneca R&D., Pepparedsleden 1, 43150 Mölndal, Sweden.
[3]Enamine Ltd., 78 Chervonotkatska str., 02094 Kyiv, Ukraine.
[4]Institute of Structural Biology, Molecular Targets and Therapeutics Center, Helmholtz Munich – Deutsches Forschungszentrum für Gesundheit und Umwelt (GmbH), 85764 Neuherberg, Germany.

Previous work:

Why do we want to predict yields:

- Crucial for multistep synthesis where yield impacts overall success, yield decrease in a step can significantly affect synthesis success
- Cost reduction in synthesis, making drugs more affordable
- Minimization of unwanted byproducts, enhancing sustainability

What have we investigated?

- Overview of current challenges in data recording regarding yield
- A typical pipeline of yield prediction in SoTA in a regression manner
- The dependence in the representation of reactions based on a record source

Findings:

- Models investigated (Yield-BERT, classical models trained on RXNFP, DRFP) do not perform well in a regression manner on real-world Buchwald-Hartwig reaction data and have poor generalization.
- Different data recording representation can have drastic effects on data encoding

Conclusions:

- We need more standardization and homogeneity in data recording
- Current widely used descriptors are not well-suited for reactivity predictions
- The more chemically relevant descriptors should be used

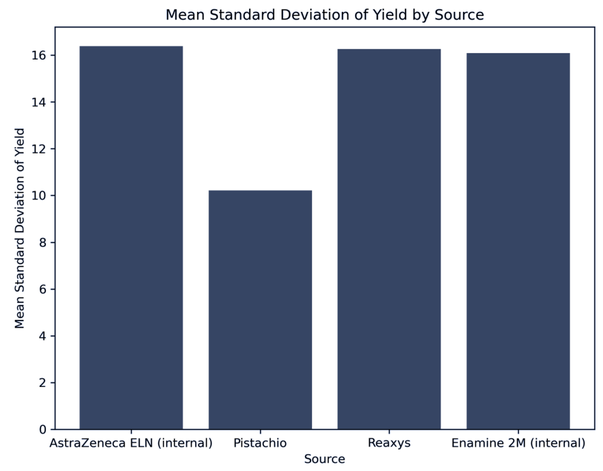

Fig. 1. Yield is a source of uncertainty in different datasets

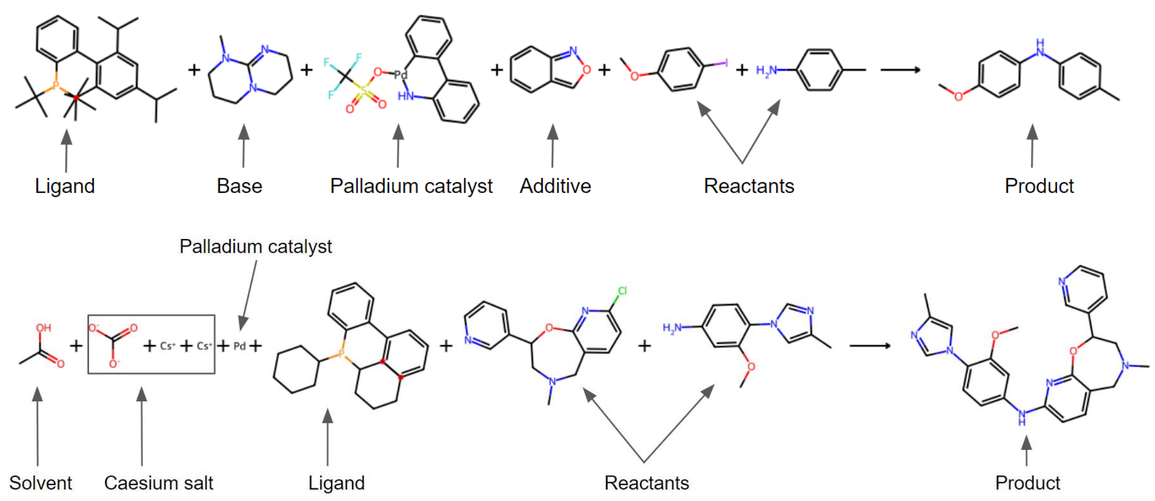

Fig .2. A different recording of one class of a reaction (top: BH dataset, bottom AZ ELN 750 dataset)

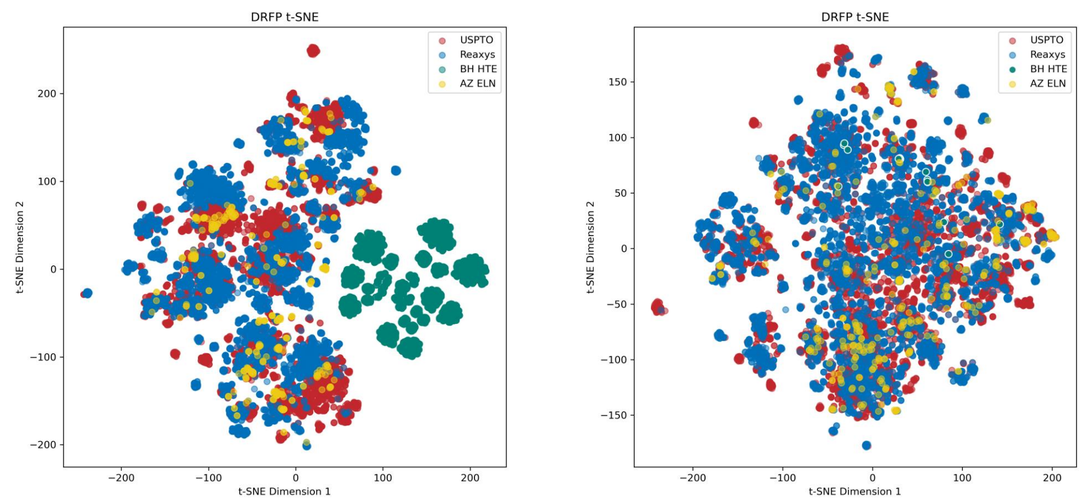

Fig. 3. Deviation in the encoding due to different recording

Current work:

Multi-bin prediction

Rationale:

- Divide available yield data into bins to translate the problem from regression to a classification problem
- Classification can be used to predict whether a reaction is good or poor-yielding

How to:

- Find the optimal bin thresholds using the Optuna package
- Using thresholds, train models and compare performance between various methods, fingerprints and reactions

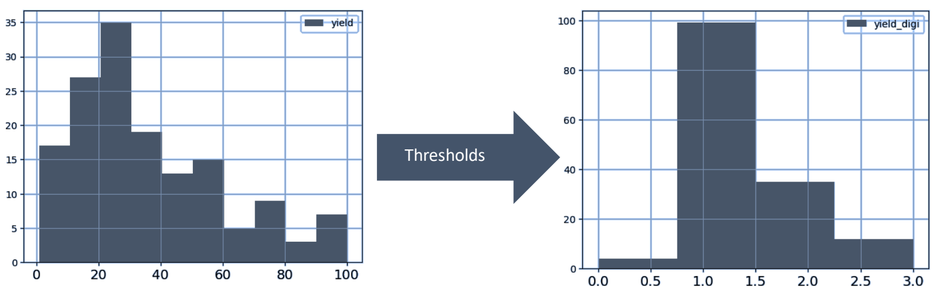

Fig. 4. Ease the problem by binning the data

Data:

- 5 Reaction types: Amide coupling, Reductive amination, Buchwald-Hartwig coupling, Suzuki coupling, SnAr
- Inner AZ Electronic Lab Notebooks (ELN)
- Reaxys
- Small libraries from iLAB team in AZ
- Enamine

Methods:

- Yield-BERT
- Random Forest Classification (RFC)
- Fingerprints: ECFP and Proximity shells + Charges (generated with Kallisto)

Findings:

- Data gets divided into common sense bins during the bin optimization
- Thorough data curation improves the results of classification
- RFC models developed on pure ELN training set perform better than models developed on ELN+Reaxys
- Kallisto-developed models perform better than ECFP and Yield-BERT

Limitations:

- Poor performance on imbalanced datasets and limited generalizability
- Data quality is still problematic

Overcoming limitations:

- Include encodings of the essential reagent to the reaction
- Use structural split to further estimate generalization capabilities
- Add the purification type as a descriptor
- Investigate other factors that could influence yield from the theory to the practice

Food for thought:

- *What are we missing in data recording that could be made machine-readable?*
- *Can we extract this data from the current recorded data?*
- *Which factors are the most important between the thermodynamical accessibility of a reaction and the synthesized product?*

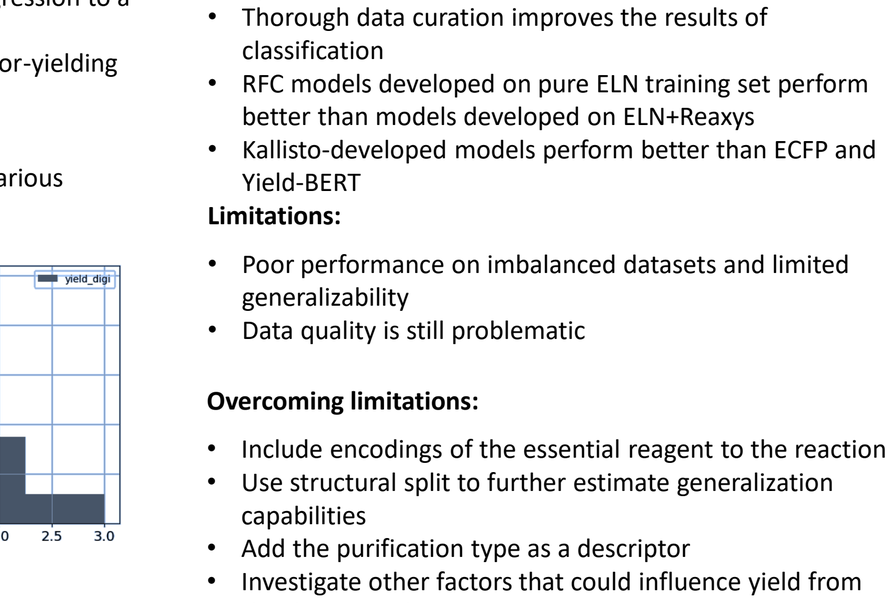

Fig. 7. Balanced accuracy and Precision micro for 3 and 4 class amide coupling models Kallisto FP vs ECFP FP

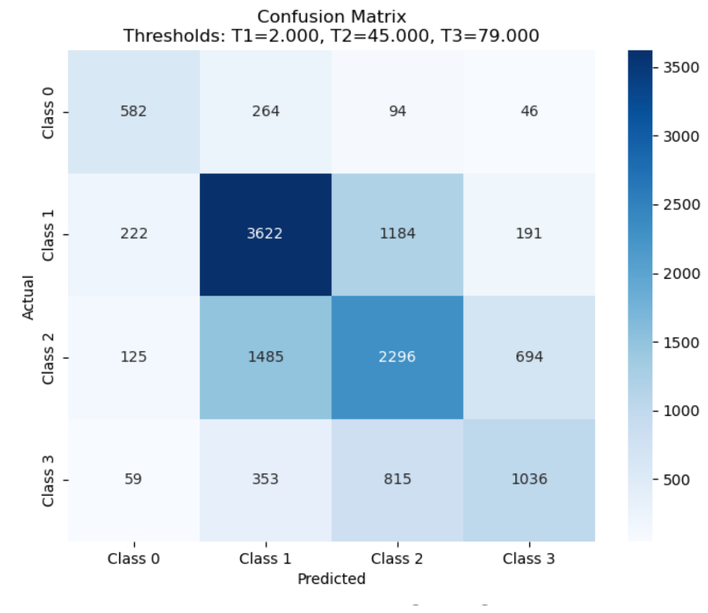

Fig. 5. Amide coupling model trained on ELN data prediction on ELN hold-out test set

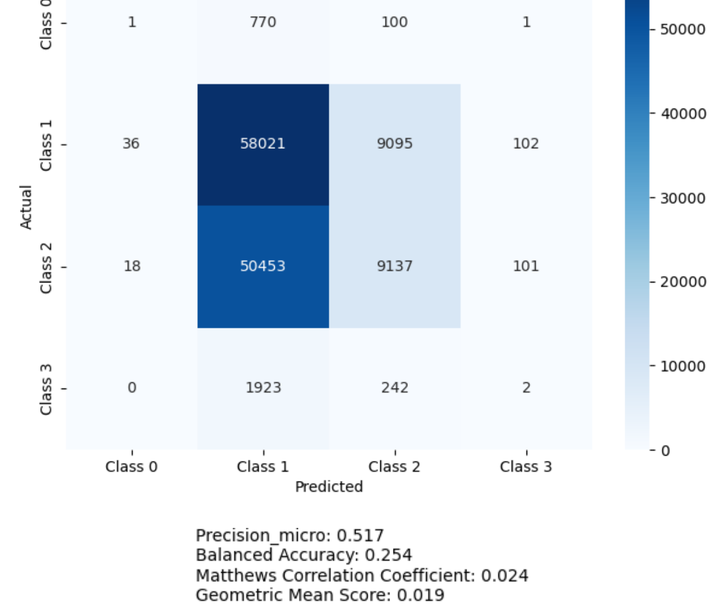

Fig. 6. Amide coupling model trained on ELN data prediction on Enamine dataset

Author's information
Varvara Voinarovska
3rd year PhD student
AstraZeneca, Technische Universität München
Fellow of Marie Skłodowska-Curie Excellence program
varvara.voinarovska@az.com

References:
[1] Voinarovska, V., Kabeshov, M., Dudenko, D., Genheden, S., and Tetko, I.V. When Yield Prediction Does Not Yield Prediction: An Overview of the Current Challenges. Journal of Chemical Information and Modeling. , 64(1), 42–56. (2023).
[2] Schwaller, P., Vaucher, A.C., Laino, T., and Reymond, J.-L. Prediction of chemical reaction yields using deep learning. Machine Learning: Science and Technology. , 2(1), 015016. (2021).
[3] Akiba, T., Sano, S., Yanase, T., Ohta, T., and Koyama, M. Optuna: A Next-generation Hyperparameter Optimization Framework. (2019).
[4] Caldeweyher, E. kallisto: A command-line interface to simplify computational modelling and the generation of atomic features. Journal of Open Source Software. , 6(60) , 3050. (2021).
Acknowledgements:
This study was partially funded by the European Union's Horizon 2020 research and innovation program under the Marie Skłodowska-Curie Innovative Training Network European Industrial Doctorate grant agreement No. 956832.

"The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data." - *John Tukey*.