When yield prediction does not yield prediction: an overview of the current challenges

Varvara Voinarovska, PhD student,
AstraZeneca MAI group

19.09.2023

Plan:

Introduction

Data side

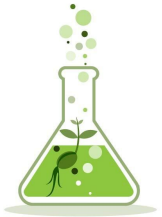
SotA side

Benchmarking

Outlook

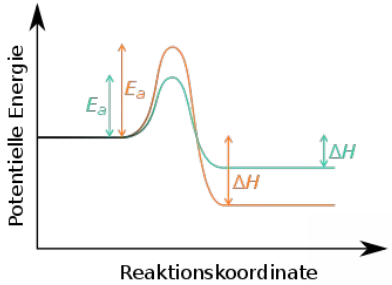
Why do we care: Importance of Reaction Yield Prediction

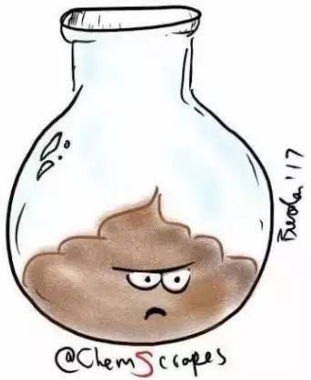
- **In CASP**: We could use yield prediction to filter out unsuccessful reactions in Computer-Aided Synthesis Planning (CASP). This could significantly cut costs and boost sustainability.
- **In HTE**: In High-Throughput Experimentation (HTE) accurate yield prediction can help us optimize synthesis processes, saving time and resources.



Factors Influencing Yield of a Chemical Reaction

- Low Reactivity
- Side Reactions
- Reactant/Reagent/Catalyst Deactivation
- Thermodynamic and Kinetic Factors
- Contaminants
- Sensitivity to Environment
- Product Degradation/Reactivity
- Product Isolation



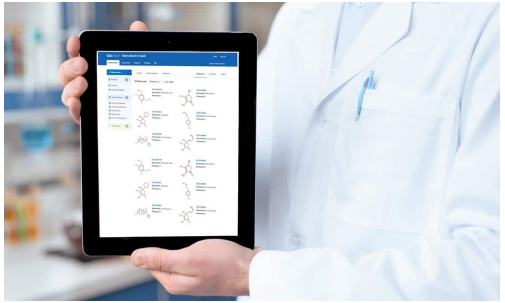


Data generation and sources

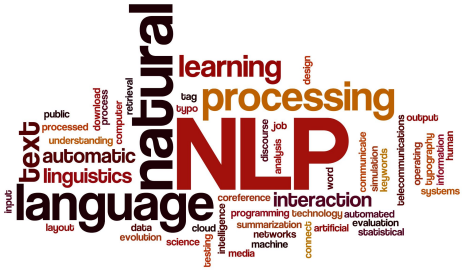
High-Throughput Experimentation

ORD

Flow Chemistry



Electronic Lab Notebooks

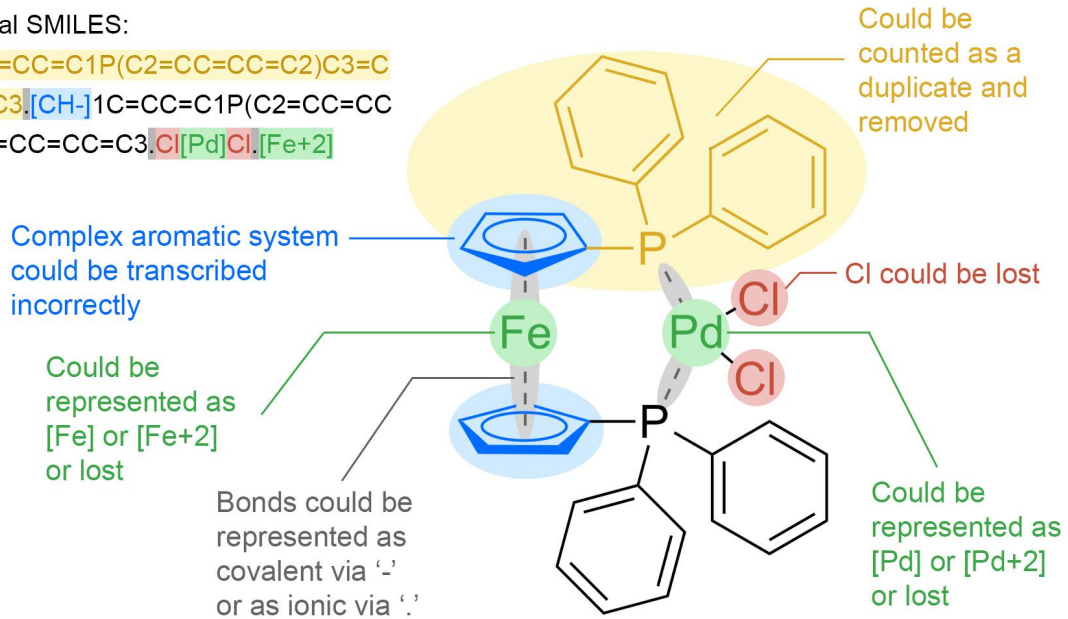


Reaction mining

What’s the problem with the data storage?

Canonical SMILES:

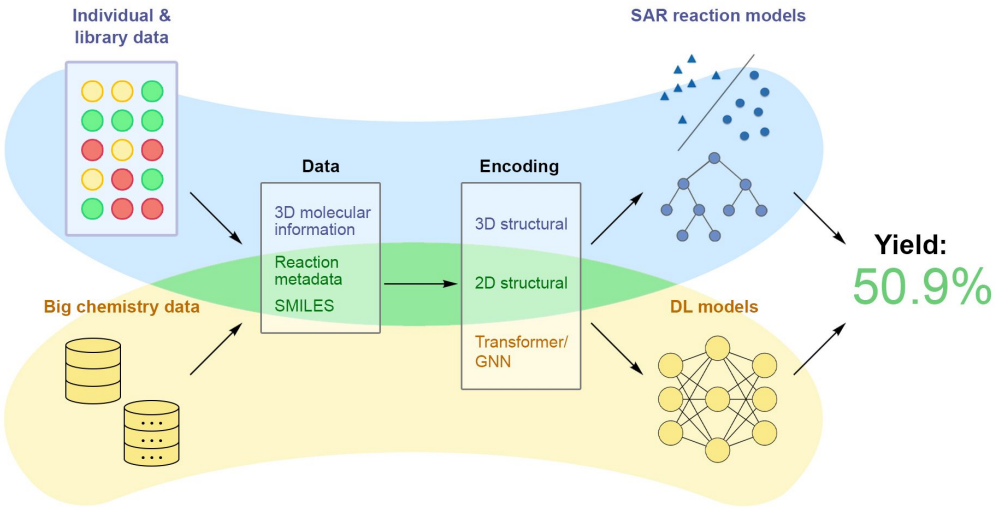
[CH-]1C=CC=C1P(C2=CC=CC=C2)C3=CC=CC=C3.[CH-]1C=CC=C1P(C2=CC=CC=C2)C3=CC=CC=C3.Cl[Pd]Cl.[Fe+2]



What is the SotA?

For *low-scale* data the Active Learning approach is popular. The design of fingerprints, including advanced ones is popular.

For *large-scale* data the Deep Learning (Transformers, Graph Neural Networks) are current SoTA.



SotA overview

Data encoding:

- DFT
- Structural fingerprints (ECFP, DRFP)
- Learned representations
- Graph-based Encodings
- One-hot

Methods:

- XGBoost, Random Forest and other classics
- Bayesian modeling
- Active learning approach

Natural Language Processing:

- Yield-BERT
- Multimodal Transformer
- Augmented Transformer

Graph-based DL:

- MPNN with self-attention
- Uncertainty-aware MPNN
- YieldGNN

Schwaller, P.; Vaucher, A. C.; Laino, T.; Reymond, J.-L. Prediction of chemical reaction yields using deep learning. Machine Learning: Science and Technology 2021, 2, 015016
Saebi, M.; Nan, B.; Herr, J. E.; Wahlers, J.; Guo, Z.; Zura´nski, A. M.; Kogej, T.; Norrby, P.-O.; Doyle, A. G.; Chawla, N. V.; Wiest, O. On the use of real-world datasets for reaction yield prediction. Chemical Science 2023, 14, 4997–5005
Baraka, S.; Kerdawy, A. M. E. Multimodal Transformer-based Model for Buchwald-Hartwig and Suzuki-Miyaura Reaction Yield Prediction. 2022; https://arxiv.org/abs/2204.14062
Kwon, Y.; Lee, D.; Choi, Y.-S.; Kang, S. Uncertainty-aware prediction of chemical reaction yields with graph neural networks. Journal of Cheminformatics 2022, 14
Yarish, D.; Garkot, S.; Grygorenko, O. O.; Radchenko, D. S.; Moroz, Y. S.; Gurbych, O. Advancing molecular graphs with descriptors for the prediction of chemical reaction yields. Journal of Computational Chemistry 2022, 44, 76–92
Neves, P.; McClure, K.; Verhoeven, J.; Dyubankova, N.; Nugmanov, R.; Gedich, A.; Menon, S.;zhicai Shi,; Wegner, J. Global Reactivity Models are Impactful in Industrial Synthesis Applications. 2022

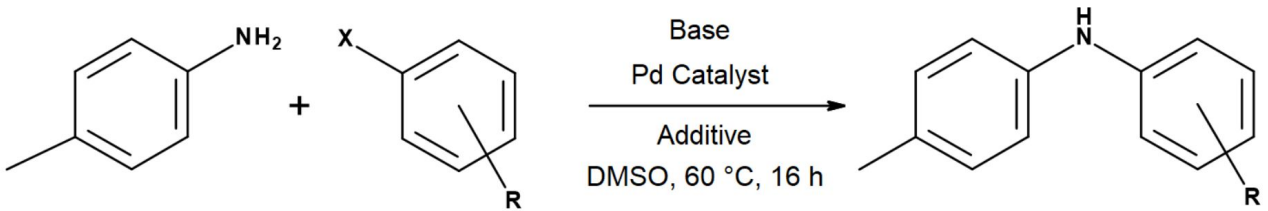
“Solved” Case

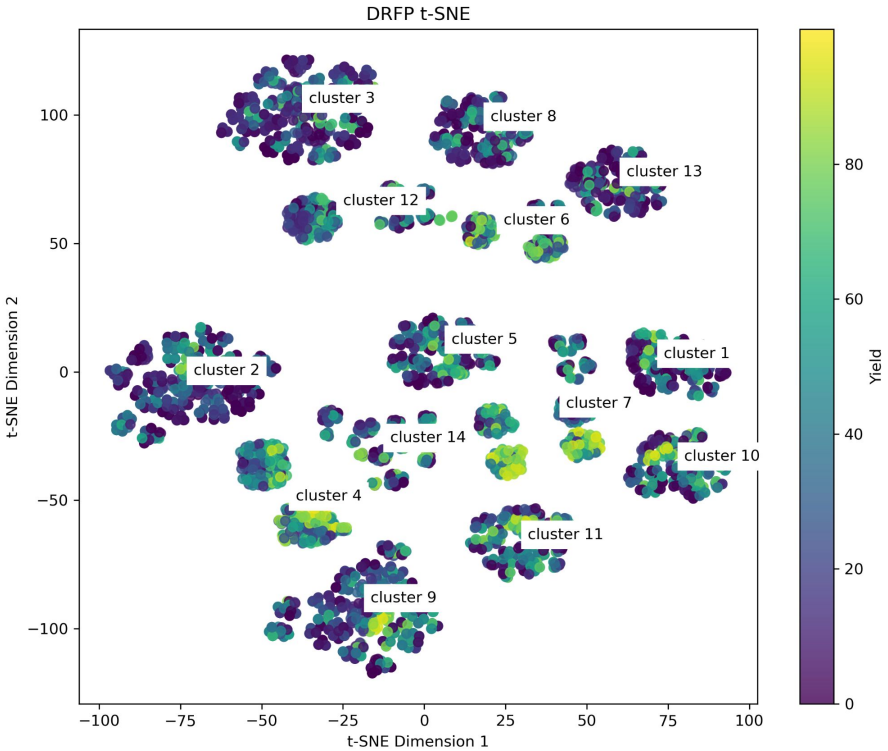
Buchwald-Hartwig HTE dataset:

15 aryl halides, 23 additives, 4 palladium catalysts, and 3 bases

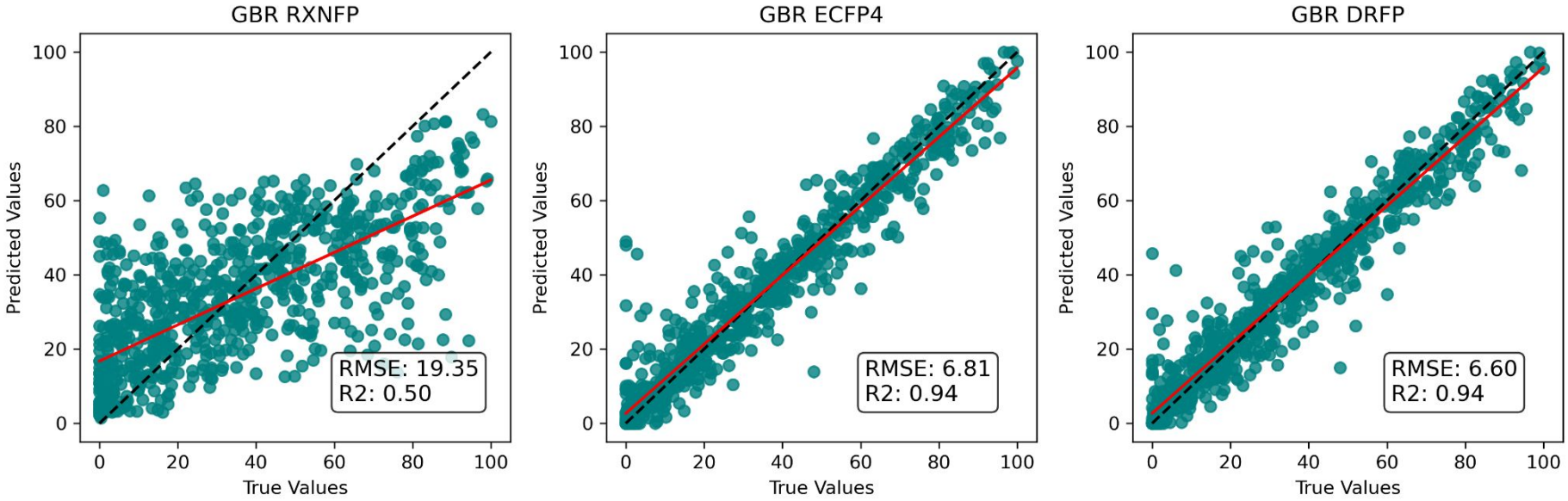
Dense and consistent

Toy dataset for yield prediction

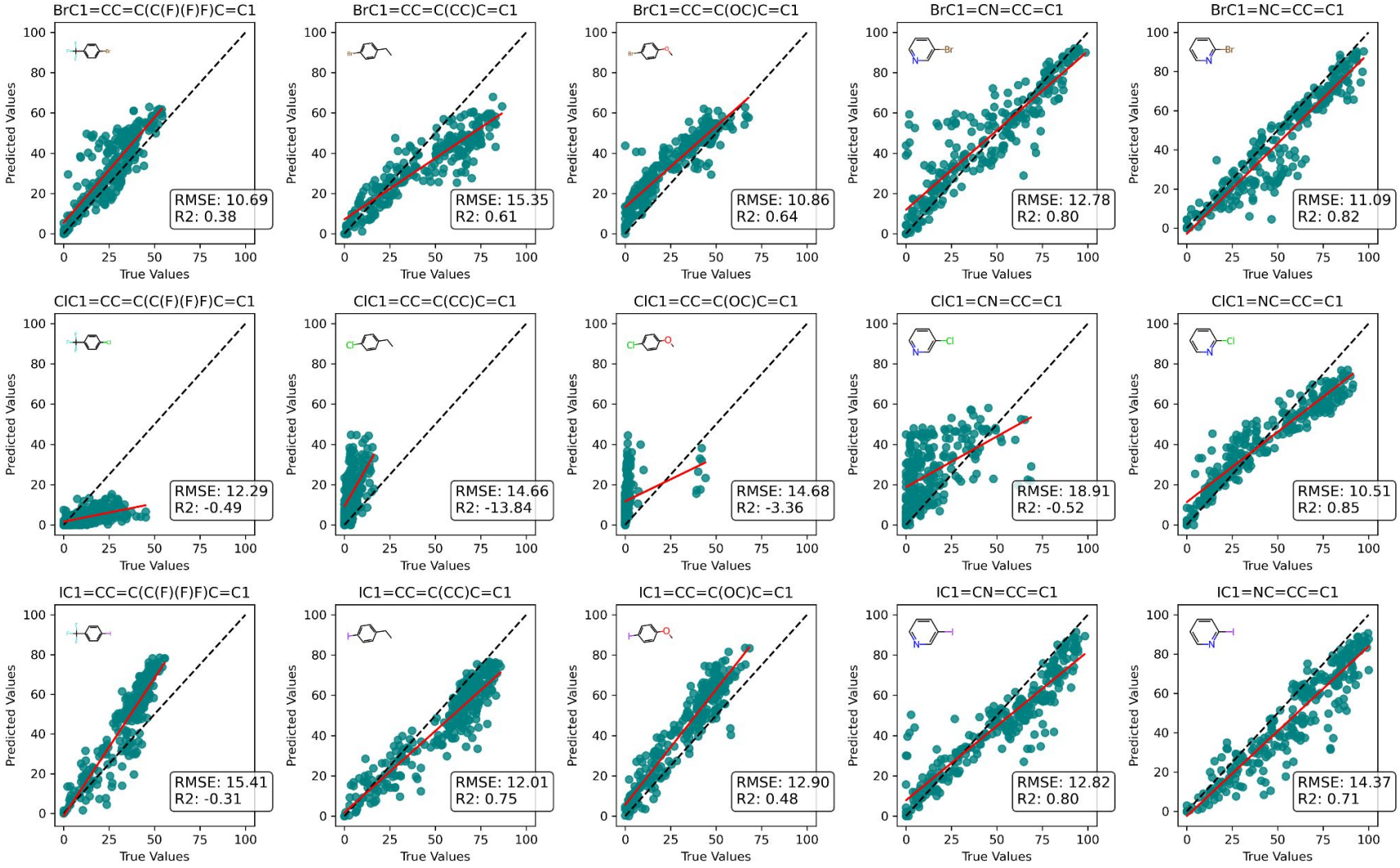




t-SNE of BH HTE dataset demonstrates that there are distinctive clusters from which one could see the areas with lower or higher yield.



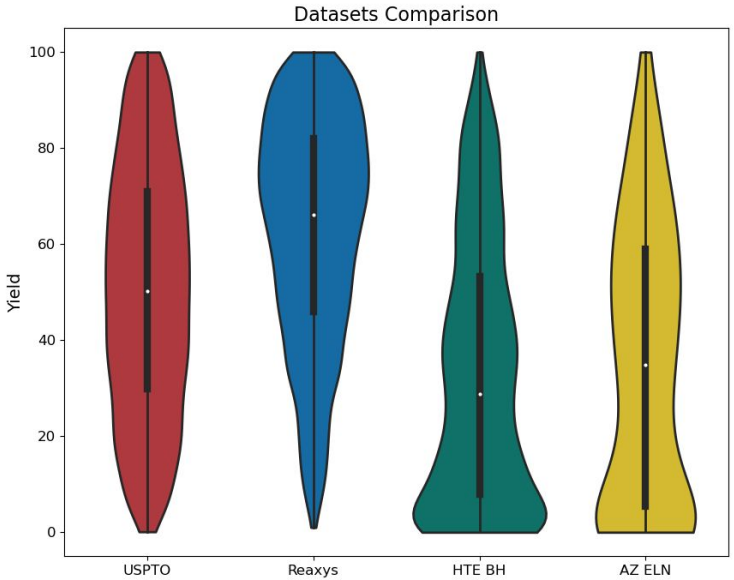
Gradient Boost Regression for BH HTE dataset in different featurizations

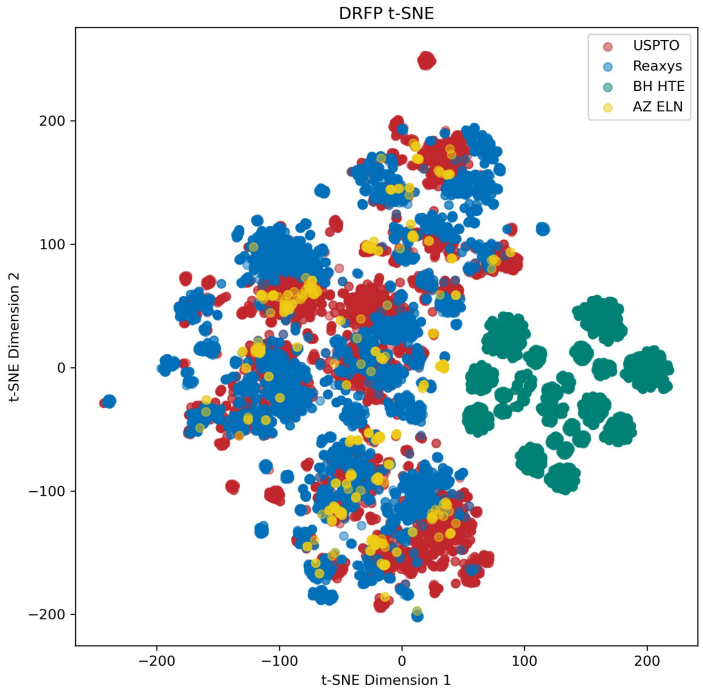

BrC1=CC=C(C(F)(F)F)C=C1
RMSE: 10.69
R2: 0.38
BrC1=CC=C(CC)C=C1
RMSE: 15.35
R2: 0.61
BrC1=CC=C(OC)C=C1
RMSE: 10.86
R2: 0.64
BrC1=CN=CC=C1
RMSE: 12.78
R2: 0.80
BrC1=NC=CC=C1
RMSE: 11.09
R2: 0.82
ClC1=CC=C(C(F)(F)F)C=C1
RMSE: 12.29
R2: -0.49
ClC1=CC=C(CC)C=C1
RMSE: 14.66
R2: -13.84
ClC1=CC=C(OC)C=C1
RMSE: 14.68
R2: -3.36
ClC1=CN=CC=C1
RMSE: 18.91
R2: -0.52
ClC1=NC=CC=C1
RMSE: 10.51
R2: 0.85
IC1=CC=C(C(F)(F)F)C=C1
RMSE: 15.41
R2: -0.31
IC1=CC=C(CC)C=C1
RMSE: 12.01
R2: 0.75
IC1=CC=C(OC)C=C1
RMSE: 12.90
R2: 0.48
IC1=CN=CC=C1
RMSE: 12.82
R2: 0.80
IC1=NC=CC=C1
RMSE: 14.37
R2: 0.71
Predicted Values
True Values

“Unsolved” Case

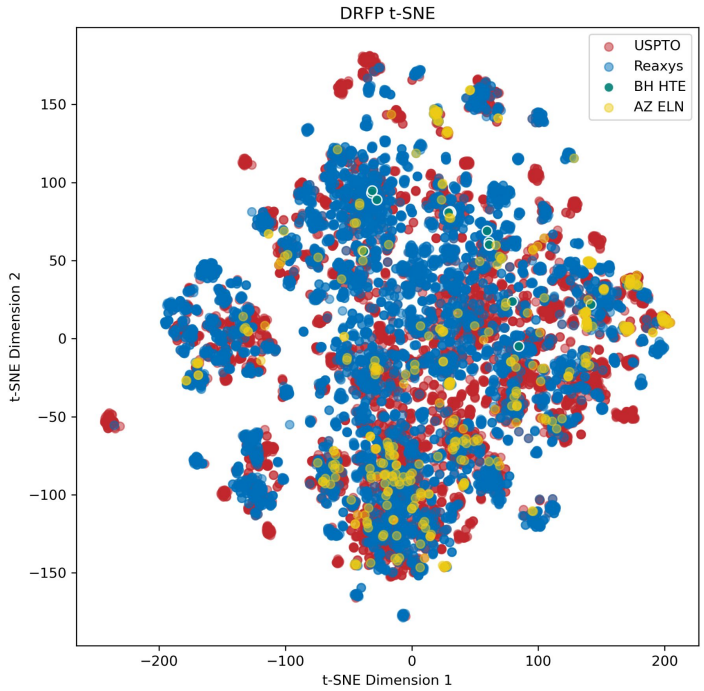
Buchwald-Hartwig Amination reaction

- Reaxys - 7K entries
- AZ ELN 750 - 500 entries
- Ahneman’s HTE Buchwald-Hartwig (BH HTE) - 4K entries
- USPTO - 6K entries

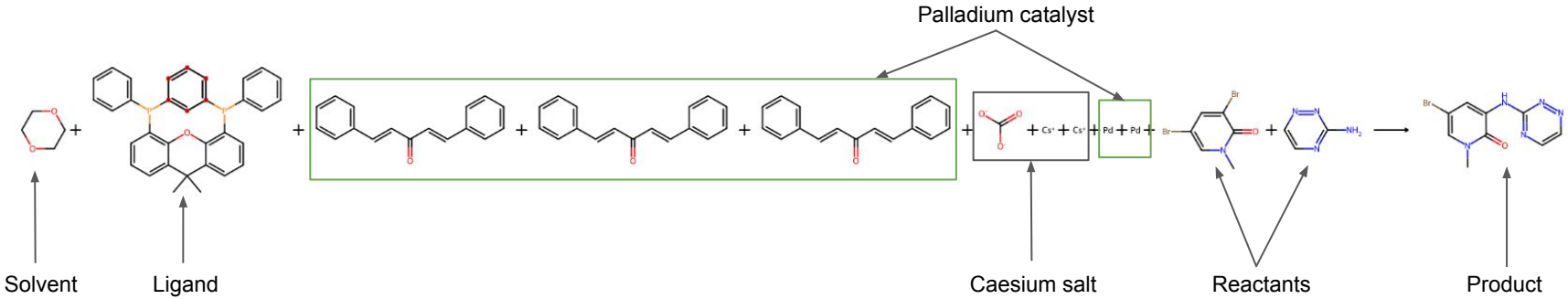




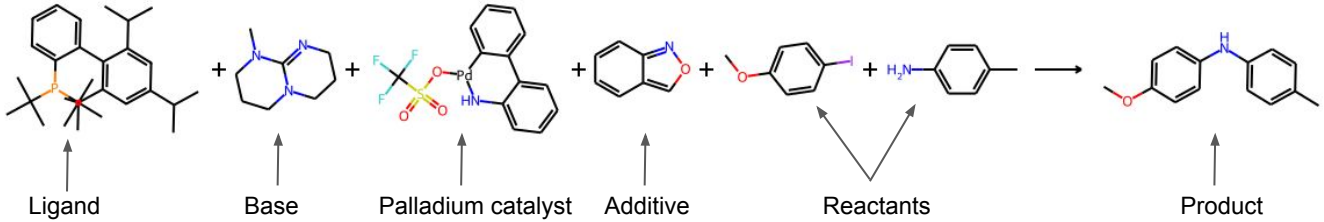
t-SNE with conditions



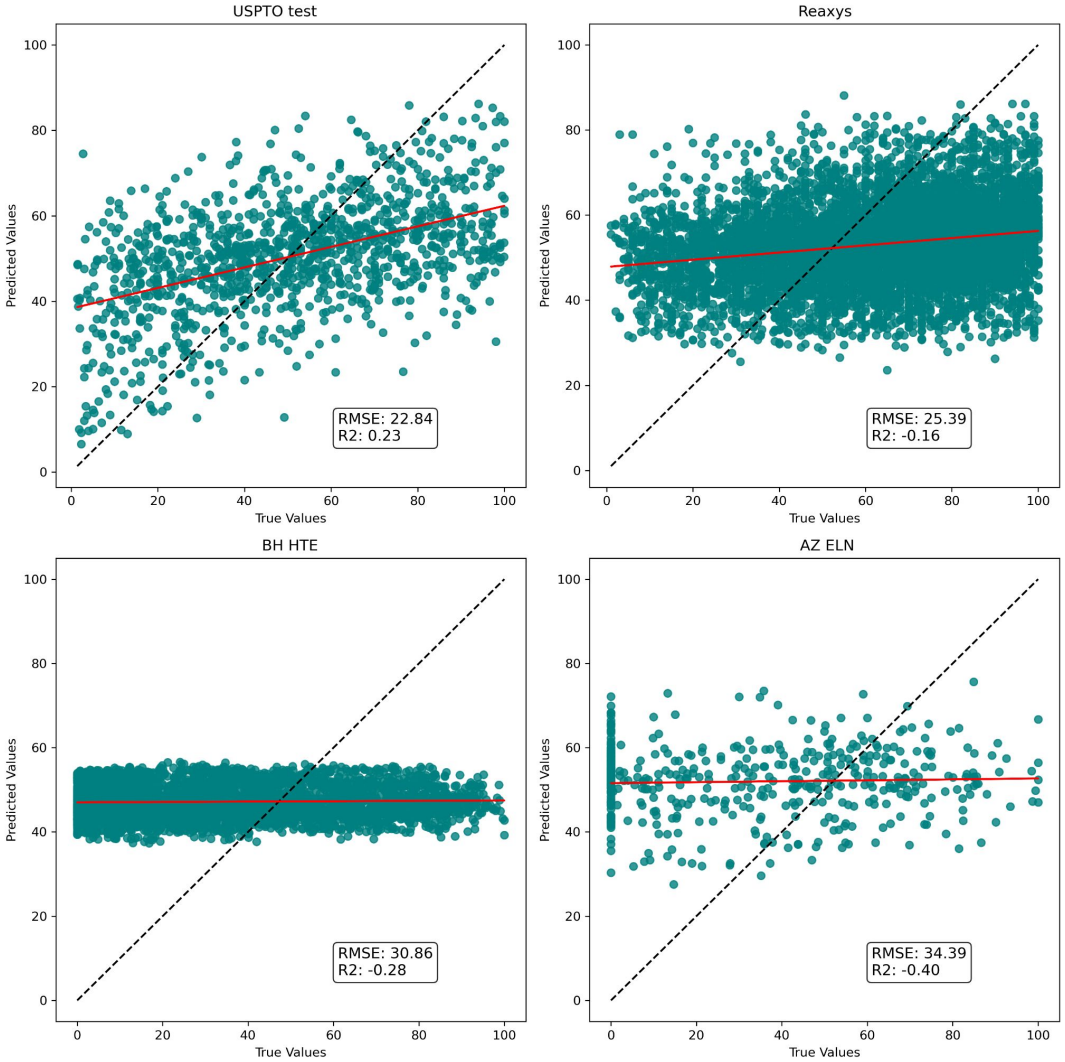
t-SNE without conditions

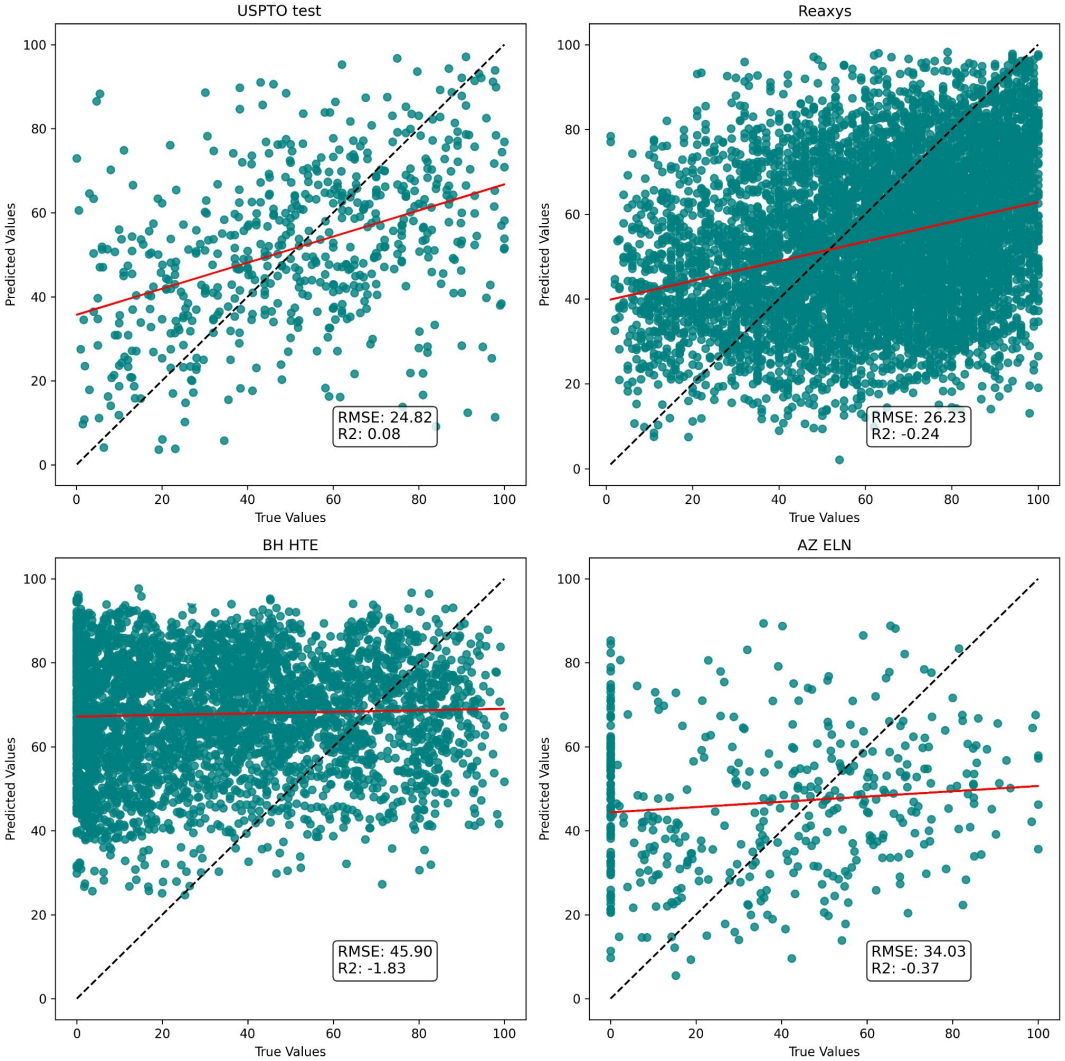


USPTO ID01456115



BH HTE example


USPTO test
Predicted Values
True Values
RMSE: 22.84
R2: 0.23
Reaxys
Predicted Values
True Values
RMSE: 25.39
R2: -0.16
BH HTE
Predicted Values
True Values
RMSE: 30.86
R2: -0.28
AZ ELN
Predicted Values
True Values
RMSE: 34.39
R2: -0.40


USPTO test
Predicted Values
True Values
RMSE: 24.82
R2: 0.08
Reaxys
Predicted Values
True Values
RMSE: 26.23
R2: -0.24
BH HTE
Predicted Values
True Values
RMSE: 45.90
R2: -1.83
AZ ELN
Predicted Values
True Values
RMSE: 34.03
R2: -0.37

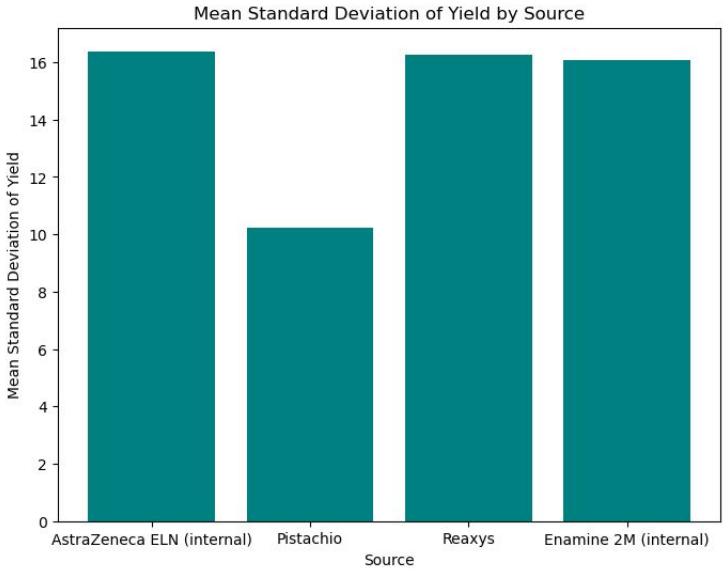
What could we do about that?

Data Standardization

Deep Integration of Chemistry

Yield Variability: a need for classification models with multiple bins to address data complexity.

Future Trajectory: The future of yield prediction involves *enhanced datasets*, uncertainty-based predictions, and the development of *reaction-specific descriptors.*



Yield prediction is influenced by inherent data noise. An analysis revealed standard deviations of around 16% in general datasets.

Acknowledgments

Supervisors:


Dr. Igor Tetko


Dr. Samuel Genheden


Dr. Dmytro Dudenko

Collaborator:


Dr. Mikhail Kabeshov

Partners:

AstraZeneca

HELMHOLTZ MUNICH

TUM

Advanced machine learning for Innovative Drug Discovery (AIDD)
Horizon 2020 Marie Skłodowska-Curie Innovative Training Network - European Industrial Doctorate

The AIDD Team

Early-Stage Researchers (ESRs)

ESR1

Peter Hartog

ESR2

Emma Svensson

ESR3

Paula Torren Peraire

ESR4

Varvara Voinarovska

ESR5

Julian Cremer

ESR6

Son Hà

ESR7

Alan Kai Hassen

ESR8

Ana Sanchez

ESR9

Yasmine Nahal

ESR10

Rosa Friesacher

ESR11

Vincenzo Pamacci

ESR12

Mikhail Andronov

ESR13

Allesio Fallani

ESR14

Muhammad Arslan Masood

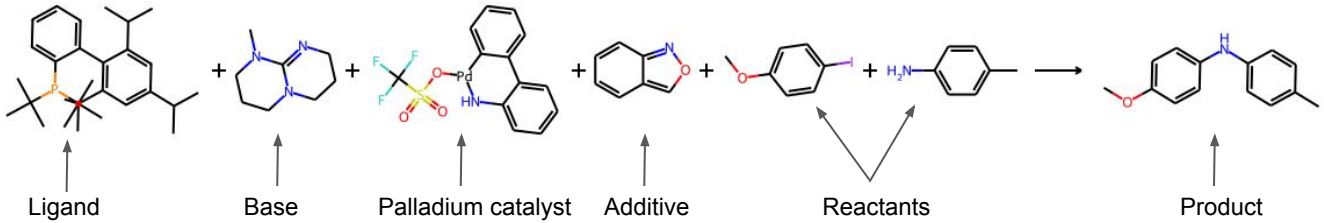
ESR15

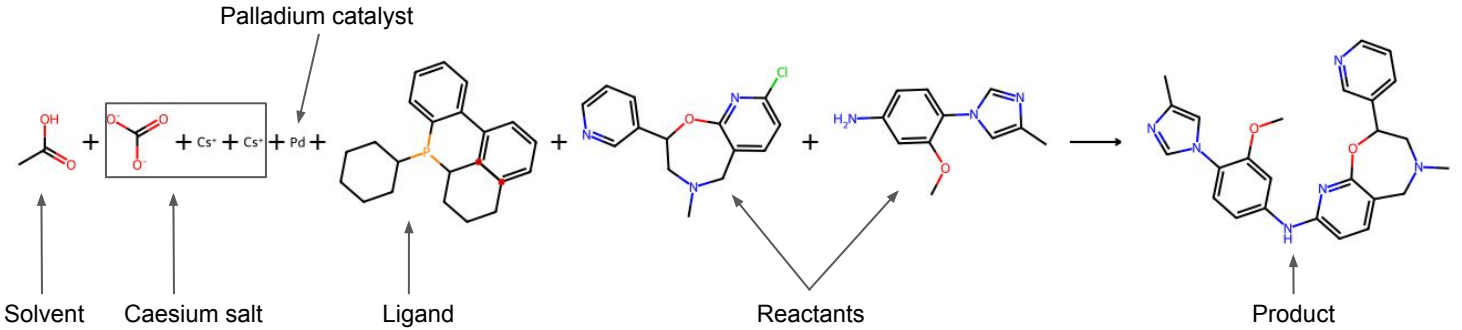
Mathias Hilfiker

ESR16

Mariia Radaeva

Thank you for your attention!


Ligand
Base
Palladium catalyst
Additive
Reactants
Product


Palladium catalyst
Solvent
Caesium salt
Ligand
Reactants
Product