Rostislav Fedorov | 21.03.2024

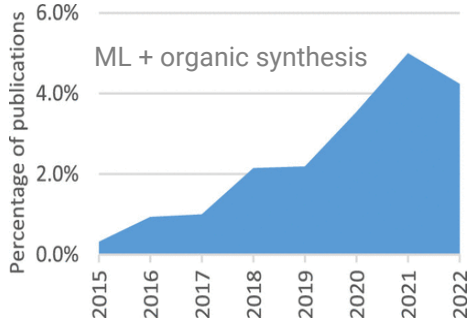
The Importance of Data in Chemistry

Challenges

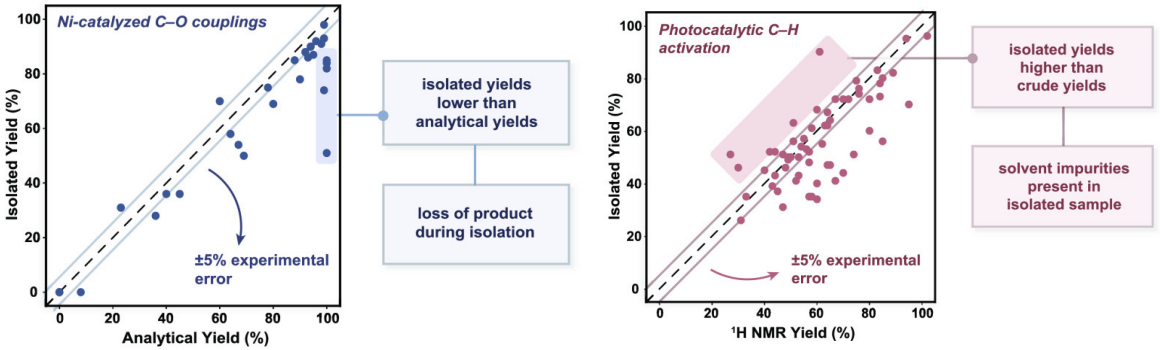
There is much more information in low-yielding reactions than is commonly accepted, and simply stating that a reaction gave 0% yield is insufficient to learn from. Possible reasons/limitations:

- no remaining starting material and no product;
- most or all of the starting material remains;
- the reaction was not performed as intended;
- Isolated/crude yields reported (see below[2]);
- Conversion as a proxy for yield.

Solution: mandatory statement before closing the experiment (drop-down menu).

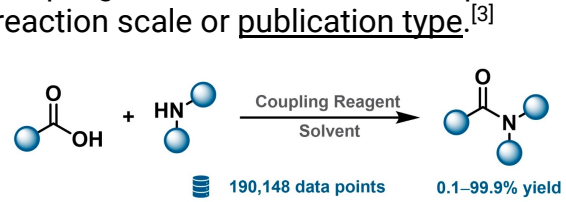


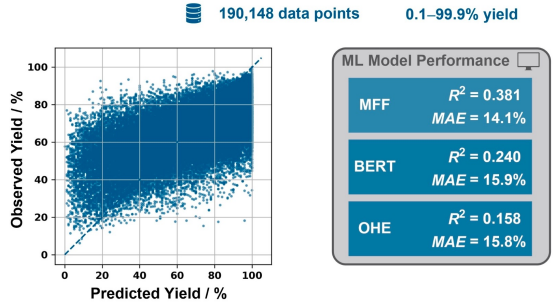
[1] Reprinted with permission from *Org. Lett.* **2023**, *25*, 2945. © 2023 American Chemical Society.



[2] Reprinted from *ACS Cent. Sci.* **2023**, *9*, 2196 under CC-BY 4.0 OA license.

Studies on the literature-extracted dataset of ~2,000 nickel-catalysed C–O couplings found that the most important model features implicitly encoded the reaction scale or publication type.[3]



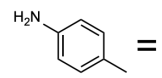


[3] *J. AM. Che. Soc.* **2022**, *144*, 14722.
[4] Reprinted with permission from *Angew. Chem. Int. Ed.* **2022**, *61*, e202204647. © 2022 Wiley-VCH GmbH.

Data augmentations:

{aryl_halide}.{methylaniline}.{pd_catalyst}.{ligand}.{base}.{additive}>>{product}
{ligand}.{base}.{methylaniline}.{additive}.{pd_catalyst}.{aryl_halide}>>{product}
{base}.{methylaniline}.{pd_catalyst}.{aryl_halide}.{additive}.{ligand}>>{product}
{additive}.{base}.{aryl_halide}.{ligand}.{methylaniline}.{pd_catalyst}>>{product}
{aryl_halide}.{pd_catalyst}.{base}.{ligand}.{methylaniline}.{additive}>>{product}

Molecule permutations

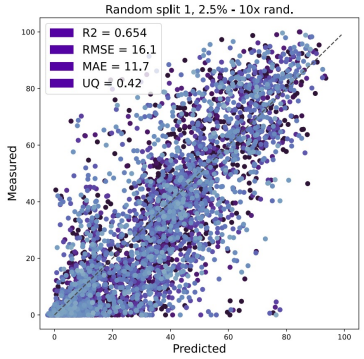


=

c1c(N)ccc(C)c1
c1cc(C)ccc1N
c1c(C)ccc(N)c1
c1(C)ccc(N)cc1
Cc1ccc(N)cc1
c1(N)ccc(C)cc1
Nc1ccc(C)cc1
c1cc(N)ccc1C

Molecule SMILES randomizations

Small data regimes (98 points!):



R^2	2.5/97.5	5/95	10/90	20/80	30/70	50/50
can	0.45 ± 0.05	0.61 ± 0.04	0.79 ± 0.02	0.86 ± 0.01	0.88 ± 0.01	0.92 ± 0.01
permuted	0.47 ± 0.13	0.70 ± 0.06	0.81 ± 0.02	0.87 ± 0.02	0.90 ± 0.01	0.94 ± 0.01
randomized	**0.61 ± 0.04**	**0.74 ± 0.03**	**0.81 ± 0.02**	**0.89 ± 0.01**	**0.92 ± 0.01**	**0.95 ± 0.01**
perm&rand	0.57 ± 0.08	0.71 ± 0.04	0.81 ± 0.02	0.89 ± 0.01	0.91 ± 0.01	0.95 ± 0.01
DFT+RF (19)	0.59	0.68	0.77	0.81	0.85	0.9

[5] Reprinted from 10.26434/chemrxiv.13286741.v1 under CC-BY NC ND 4.0 OA license. Machine Learning for Molecules Workshop at NeurIPS 2020.

Simulating the errors [4]