Chemical Compound Space Conference 2024 • Heidelberg • May, 21st to 24th • Germany

ROBERT: Bridging the Gap between Machine Learning and Chemistry

David Dalmau, Juan V. Alegre-Requena, Esteban Urriolabeitia

INTRODUCTION

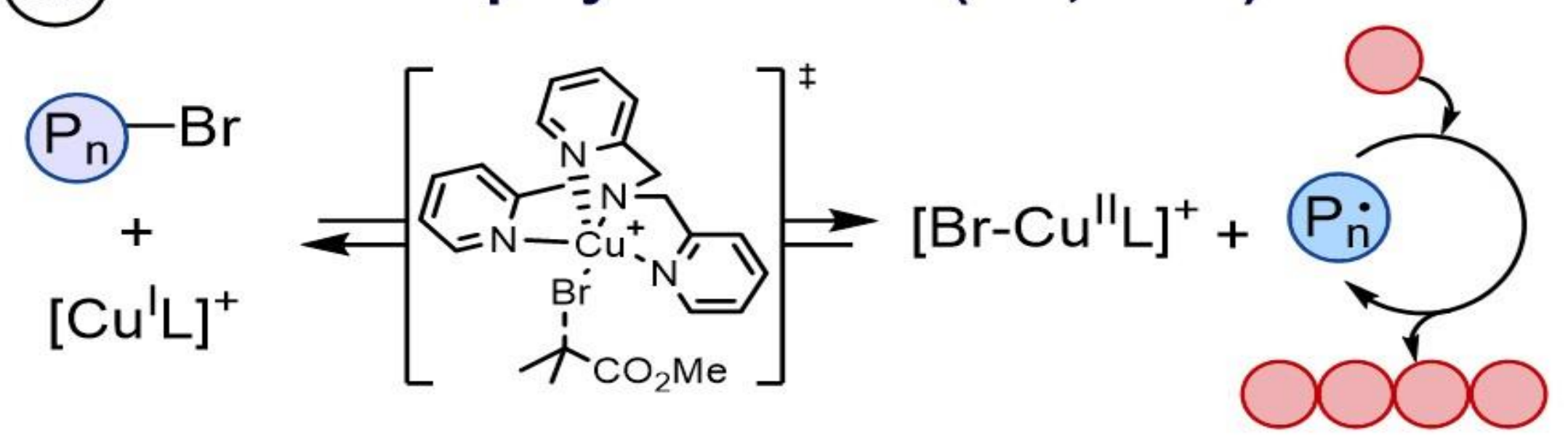
The rapid progress of machine learning (ML) has transformed chemical research. Its integration not only fulfills technological needs but also fosters sustainability through the adoption of digitalized procedures, yielding important benefits for a more environmentally conscious future. Despite this evolution, there are implementation gaps that hinder the widespread adoption of ML protocols among a significant portion of the chemistry community. Herein, we introduce ROBERT, [1] a program designed to make ML more accessible to chemists regardless of their level of programming. This software not only enables researchers to produce results comparable to experts in the field, but also adheres to strict reproducibility and transparency standards. [2]

ROBERT

AUTOMATED ML PROTOCOLS

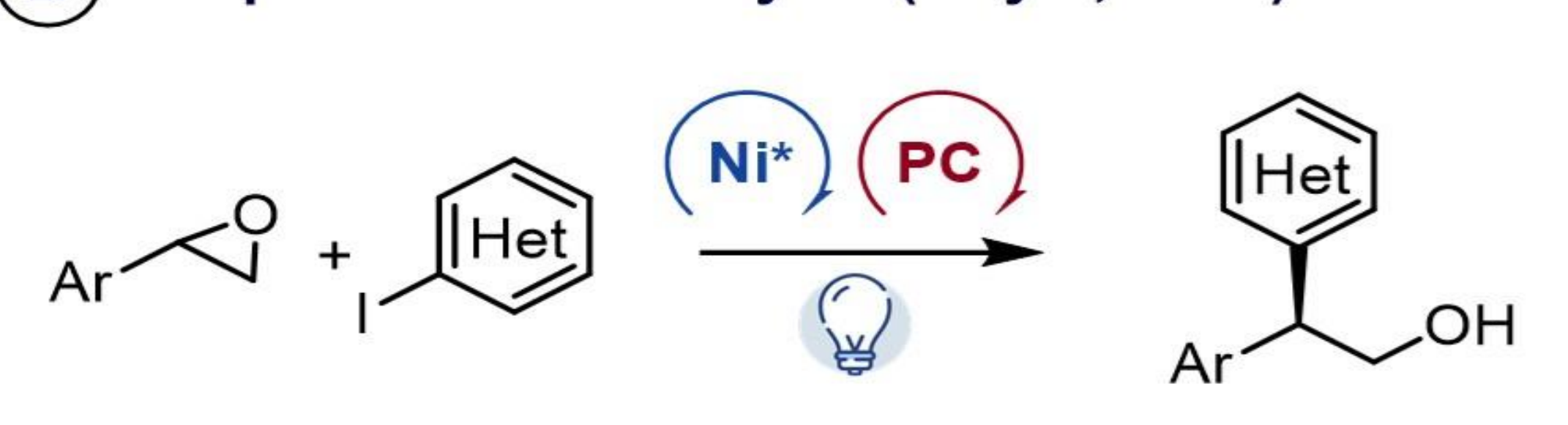
BENCHMARKING

A Cu radical polymerization (Liu, 2019)



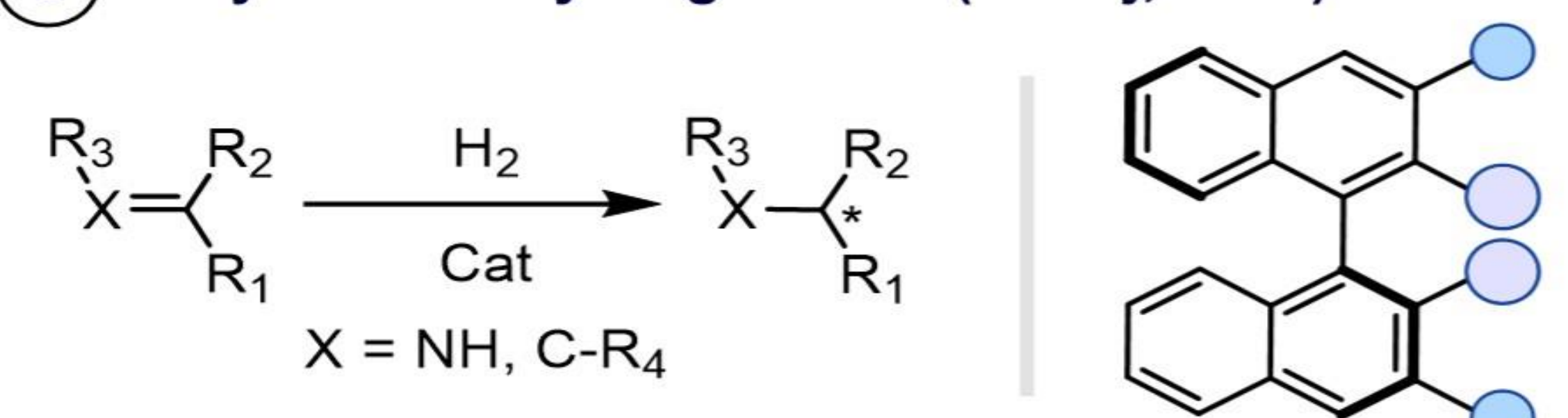
- 18 datapoints
- Regression ($y = \Delta G^‡$)
- Training size: 50% (original), 70% (ROBERT)
- Number of descriptors: 4 (original), 2 (ROBERT)

B Ni photoredox catalysis (Doyle, 2021)



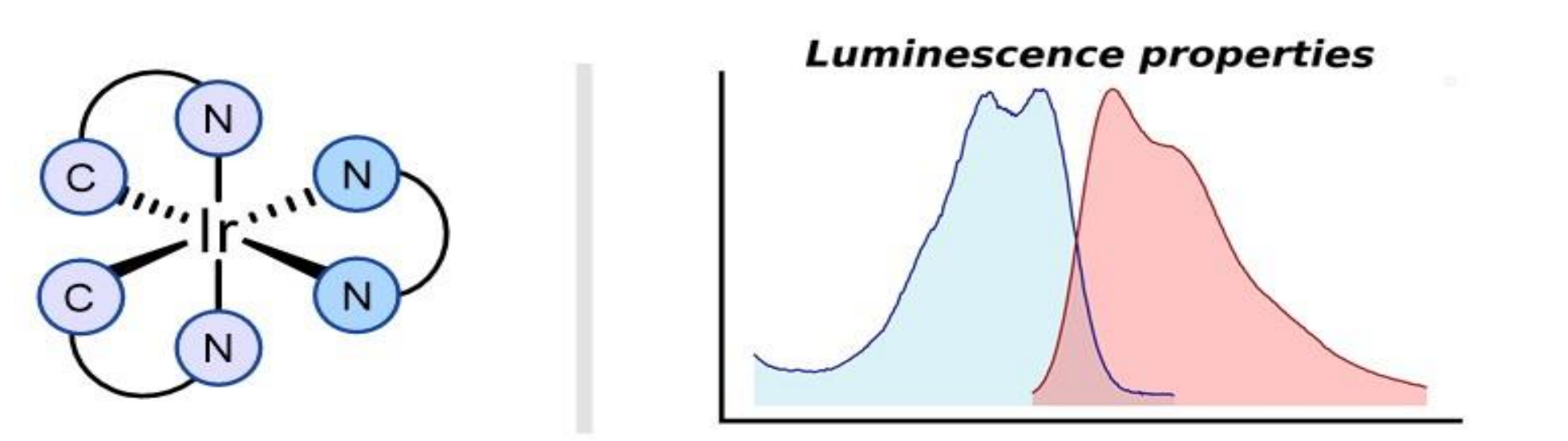
- 29 datapoints
- Regression ($y = \Delta\Delta G^‡$)
- Training size: 70% (original), 60% (ROBERT)
- Number of descriptors: 3 (original), 3 (ROBERT)

C Asymmetric hydrogenation (Sunoj, 2020)



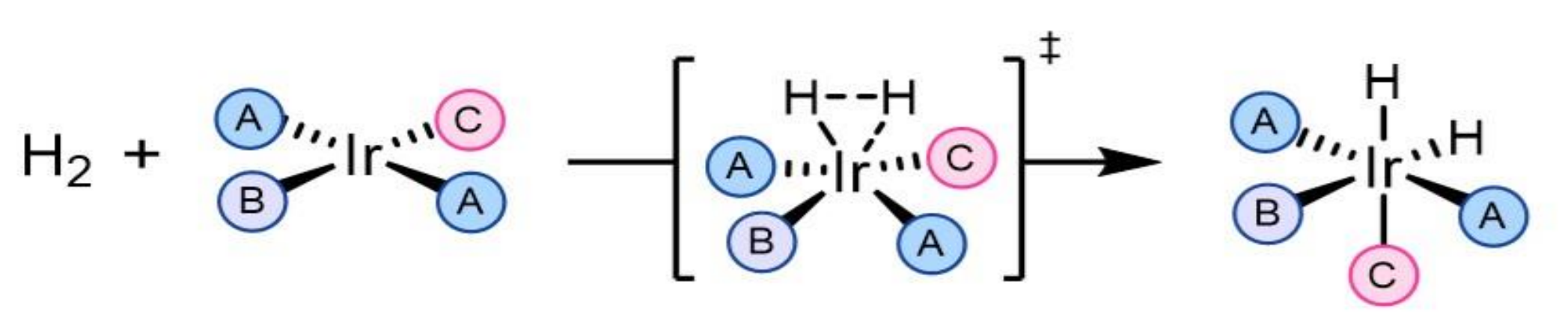
- 368 datapoints
- Regression (y = %ee)
- Training size: 80% (original), 70% (ROBERT)
- Number of descriptors: 101 (original), 3 (ROBERT)

D Ir photophysical properties (Kulik, 2023)



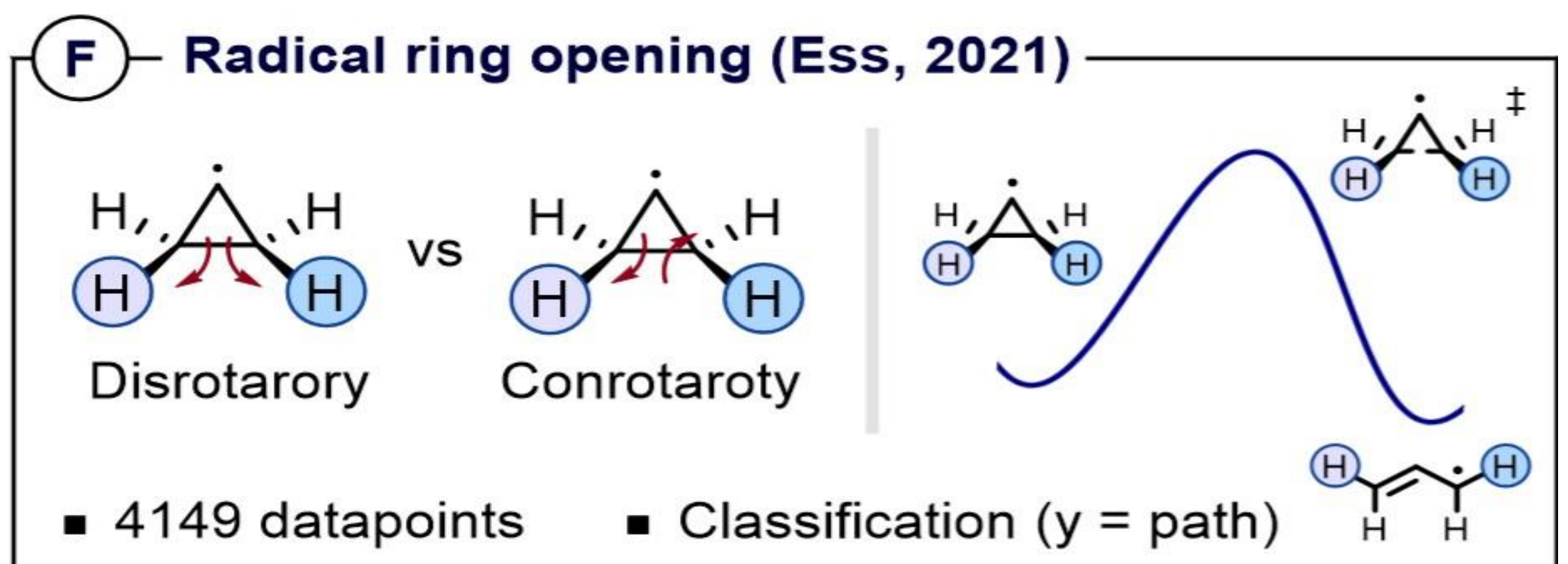
- 1173 datapoints
- Regression (y = spec. integ.)
- Training size: 85% (original), 90% (ROBERT)
- Number of descriptors: 12 (original), 11 (ROBERT)

E H_2 split (Aspuru-Guzik & Balcells, 2020)

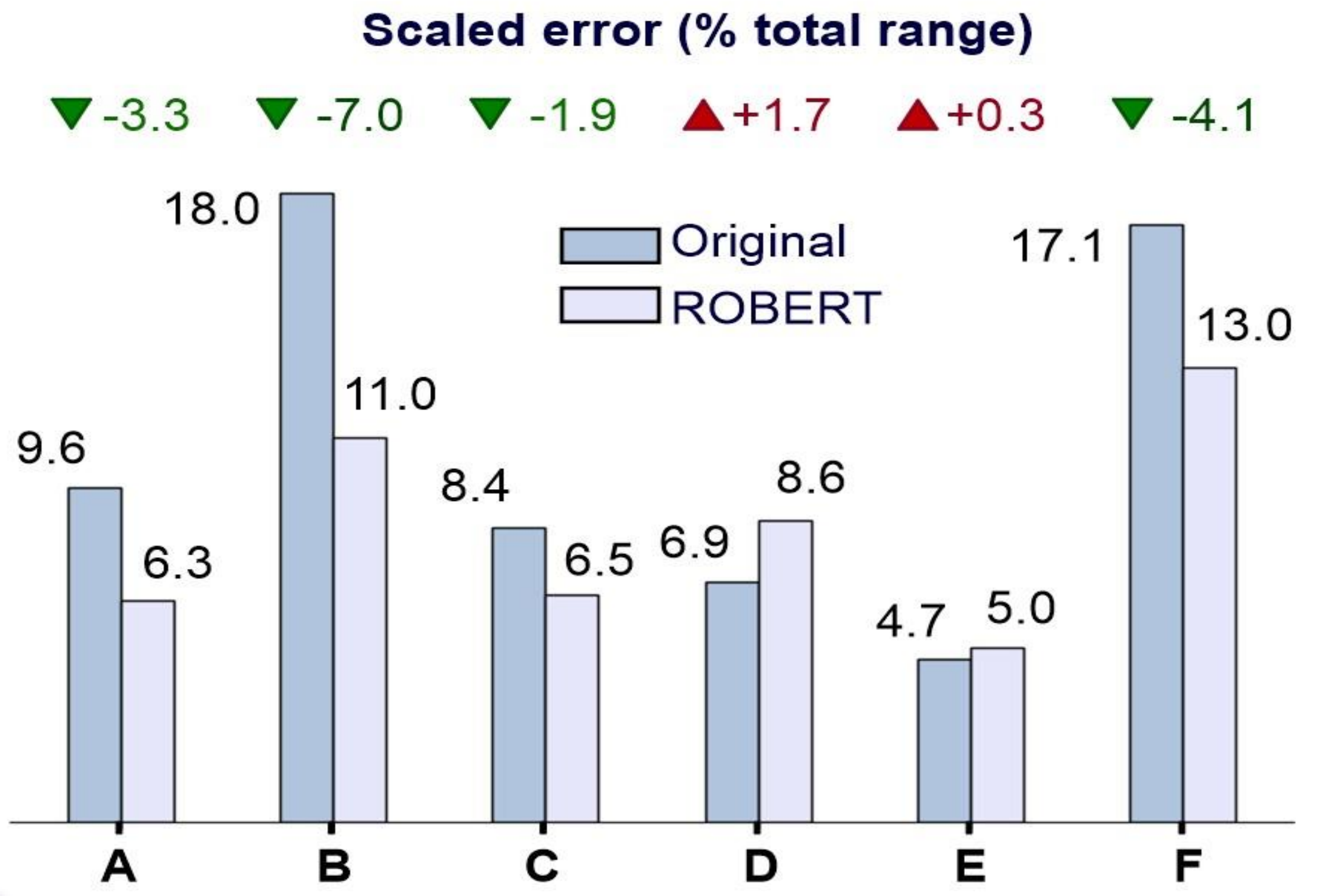


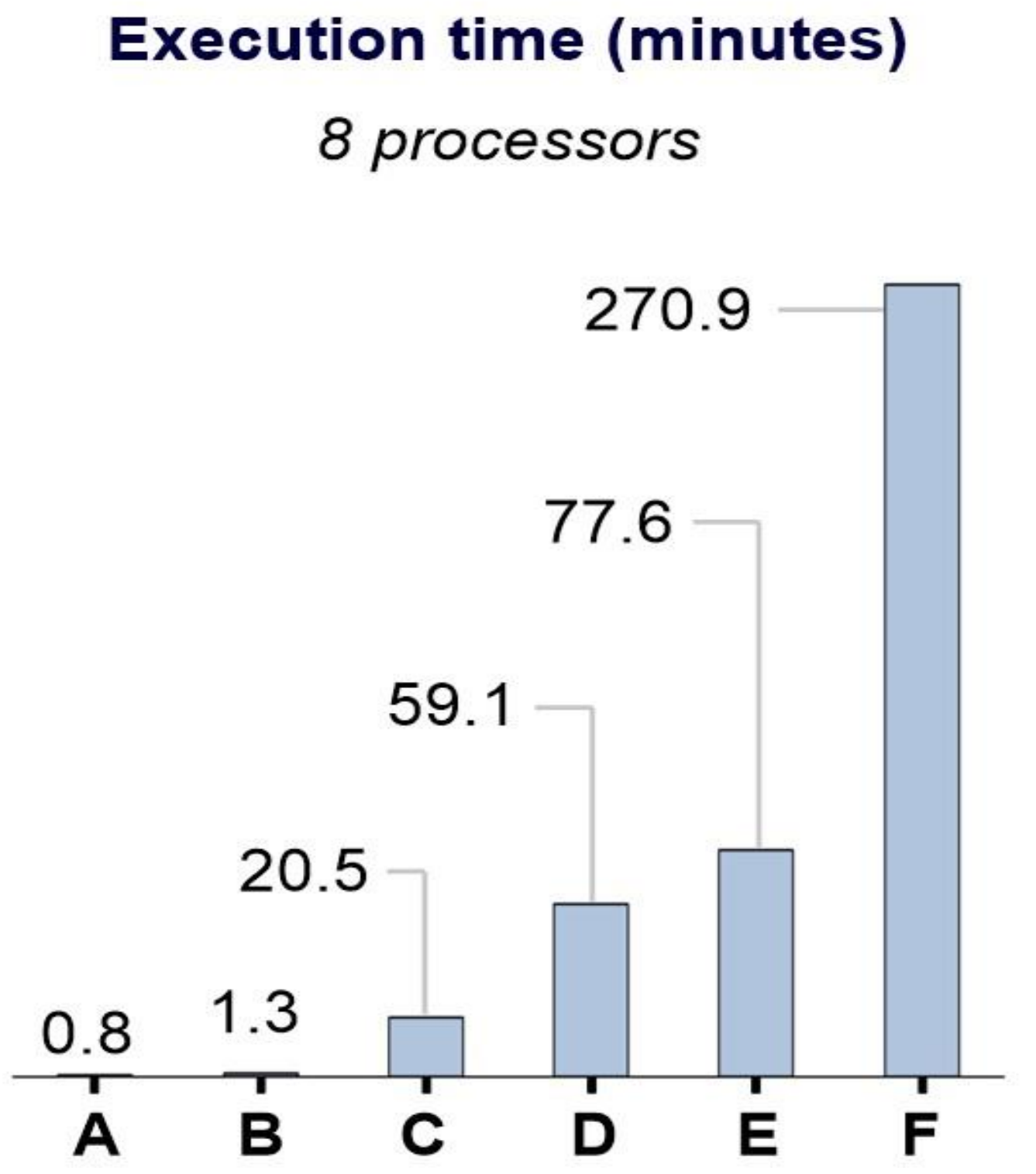
- 1947 datapoints
- Regression ($y = \Delta E^‡$)
- Training size: 80% (original), 90% (ROBERT)
- Number of descriptors: 135 (original), 32 (ROBERT)

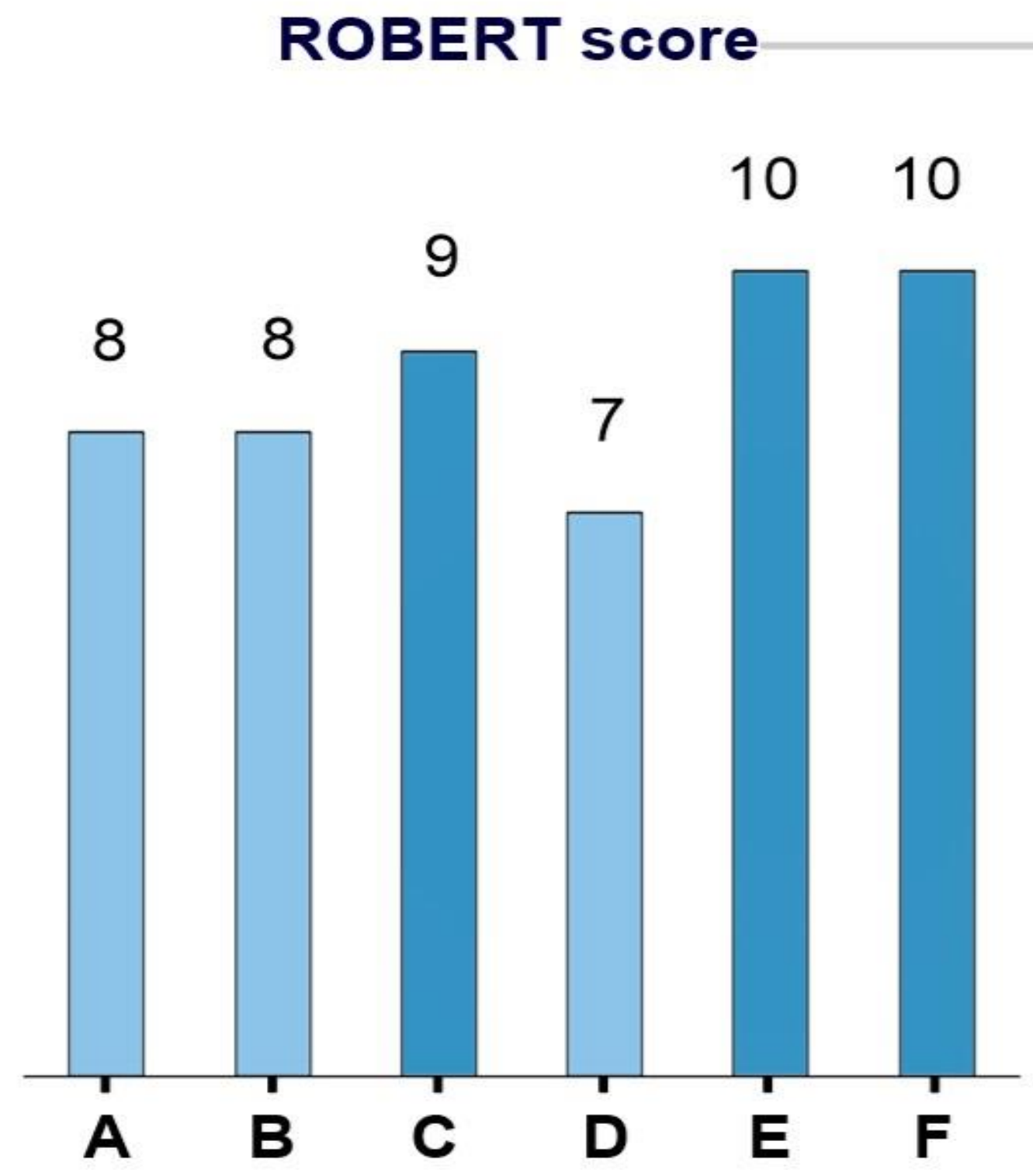
F Radical ring opening (Ess, 2021)



- 4149 datapoints
- Classification (y = path)
- Training size: 95% (original), 90% (ROBERT)
- Number of descriptors: 85 (original), 60 (ROBERT)



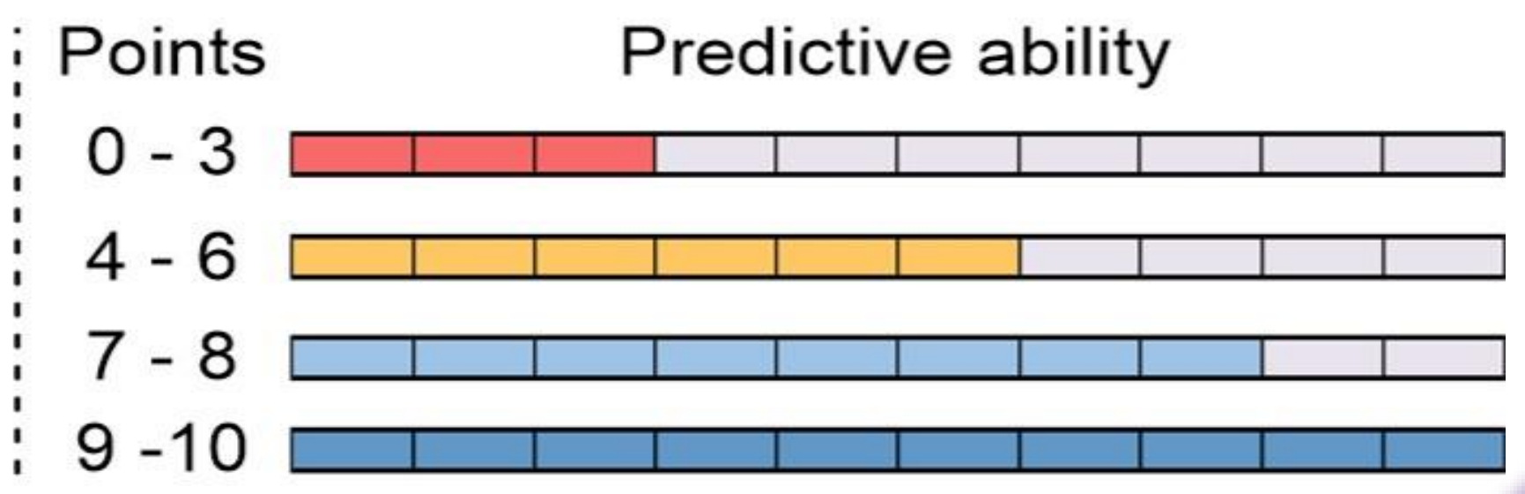




Score criteria (points)

Points	R^2	Outliers	Desc:pts
●●	> 0.85	< 7.5%	> 1:10
●	0.70 - 0.85	7.5 - 15%	1:3 - 1:10
–	< 0.70	> 15%	< 1:3

Verify tests (up to ●●●●)	● 5-fold CV	● y-shuffle
	● y-mean	● One-hot

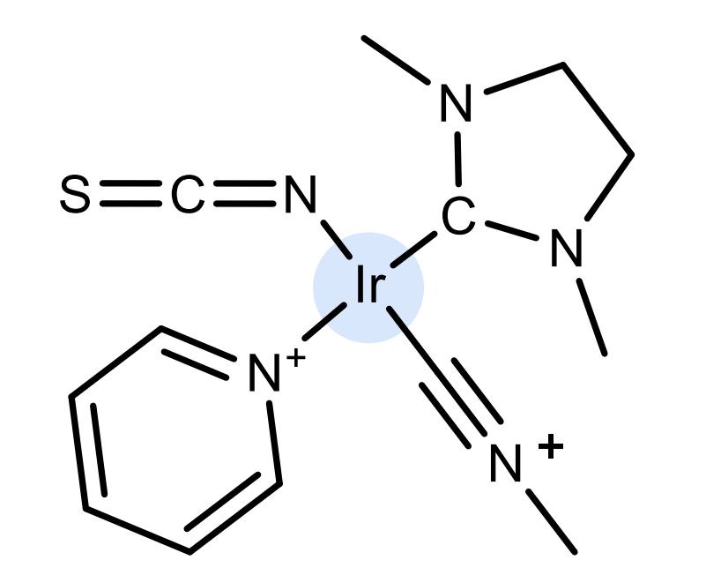


WORKFLOWS FROM SMILES

Vaska's complex database

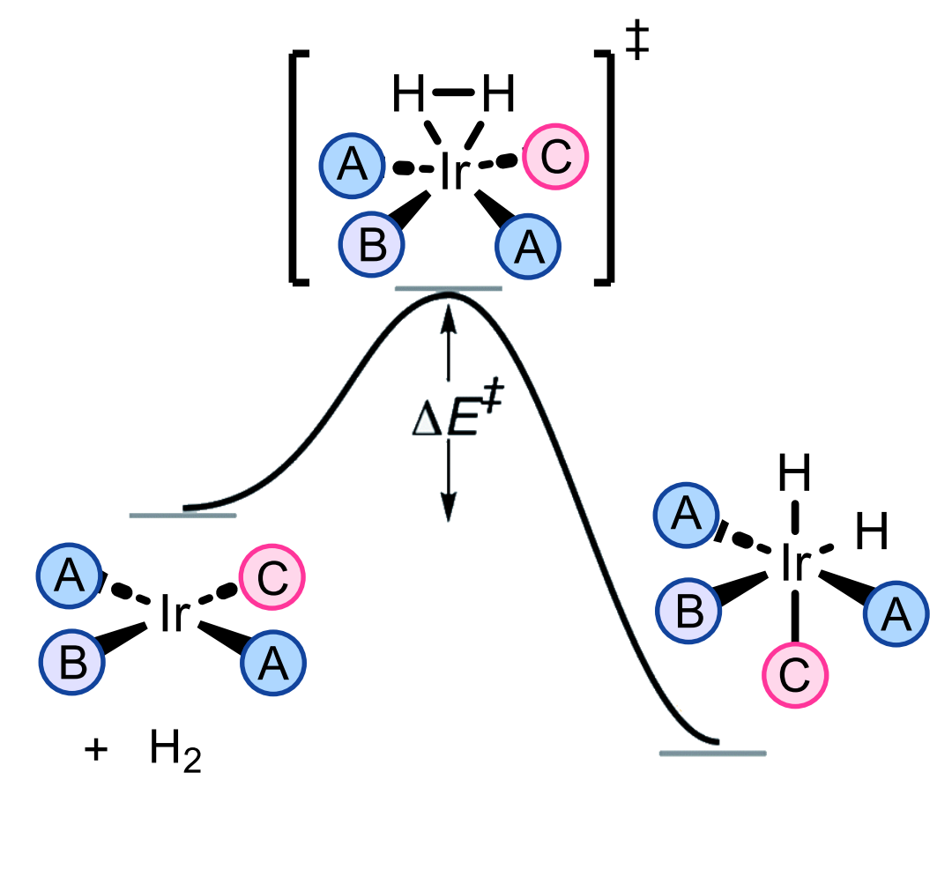
SMILES	code_name	$\Delta E^‡$
[Ir]([P+](CC)...	ir_tbp_1_dft-pet3...	8.9
[Ir]([P+](C)...	ir_tbp_1_dft-pme3...	6.5
[Ir]([As+](C)...	ir_tbp_1_dft-asme3...	17.9
[Ir]([n+]1ccn...	ir_tbp_1_dft-pyz...	22

Step 1: AQME

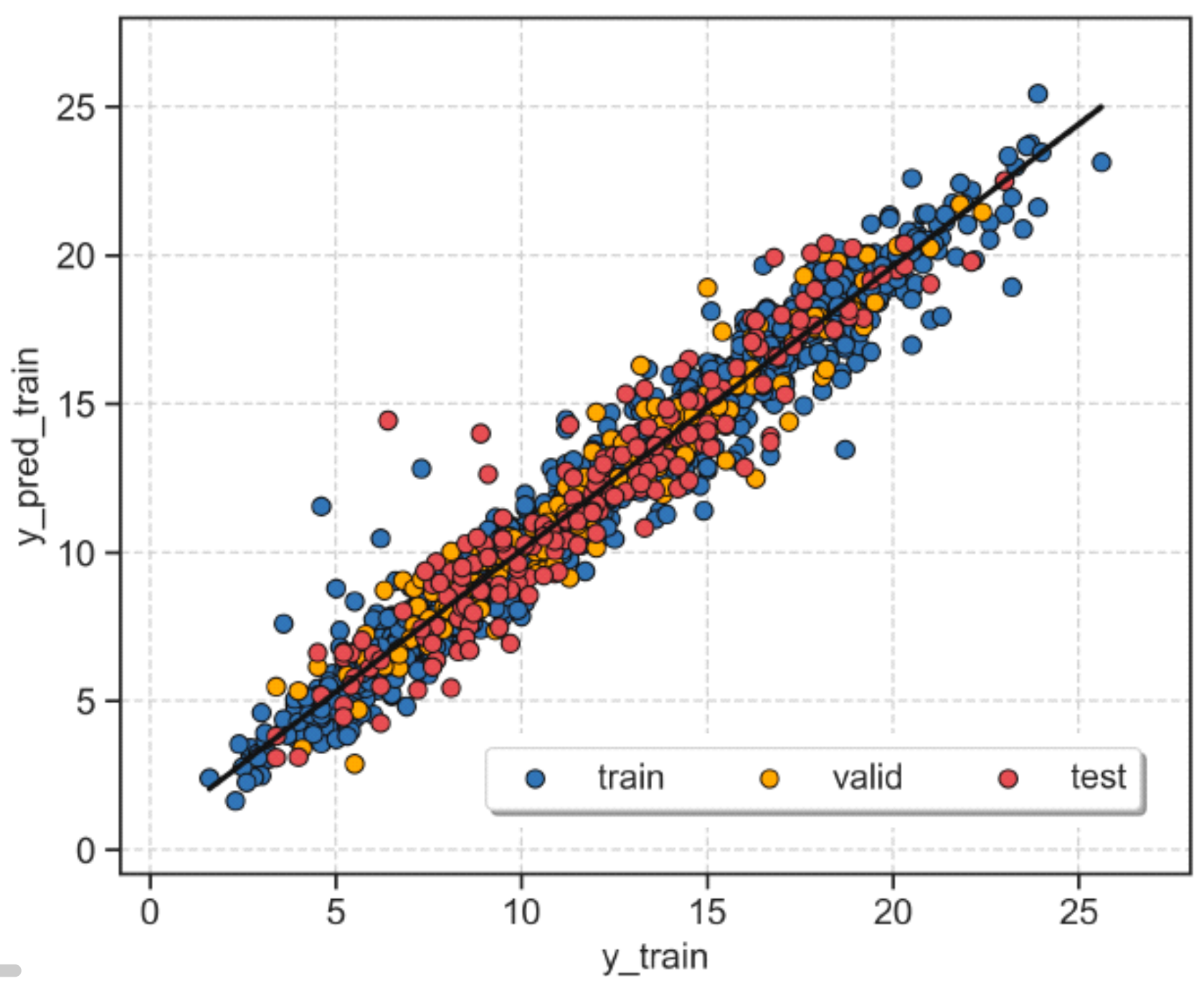


200+ electronic & steric atomic/molecular descps. (RDKit, xTB & DBSTEP)

Step 2: ROBERT



$\Delta E^‡$ prediction



ROBERT SCORE

ML model: NN

Proportion Train:Validation:Test = 81:9:10

The model has a score of 9/10

- ●● The test set shows an R^2 of 0.89
- ● The valid. set has 10.5% of outliers
- ●● Using 1711:21 points(train+valid.):descriptors
- ●●●● The valid. set passes 4 VERIFY tests

REFERENCES

[1] Dalmau, D.; Alegre Requena, J. V., *ChemRxiv* **2023**, DOI: 10.26434/chemrxiv-2023-k994h.

[2] Walsh, I.; Fishman, D., *Nat. Methods* **2021**, 18, 1122–1127.

ACKNOWLEDGEMENTS

This research was funded by the Aragón Government (Spain, DGAFSE research groups E17_23R and E07_23R) and the Spanish Government (MICIN, Projects No. PID2022-140159NA-I00, PID2019-104379RB-C21 and PID2019-106394GB-I00, funded by MCIN/AEI/10.13039/501100011033).

MINISTERIO DE CIENCIA E INNOVACIÓN

GOBIERNO DE ARAGON

Unión Europea

CONTACTING DETAILS: ddalmau@unizar.es ChemRxiv Documentation @DalmauDavid