香港中文大學
The Chinese University of Hong Kong

Journal Club

Transferable enantioselectivity models from sparse data

Abigail G. Doyle & Matthew S. Sigman

Zihan Li

12th Mar 2026

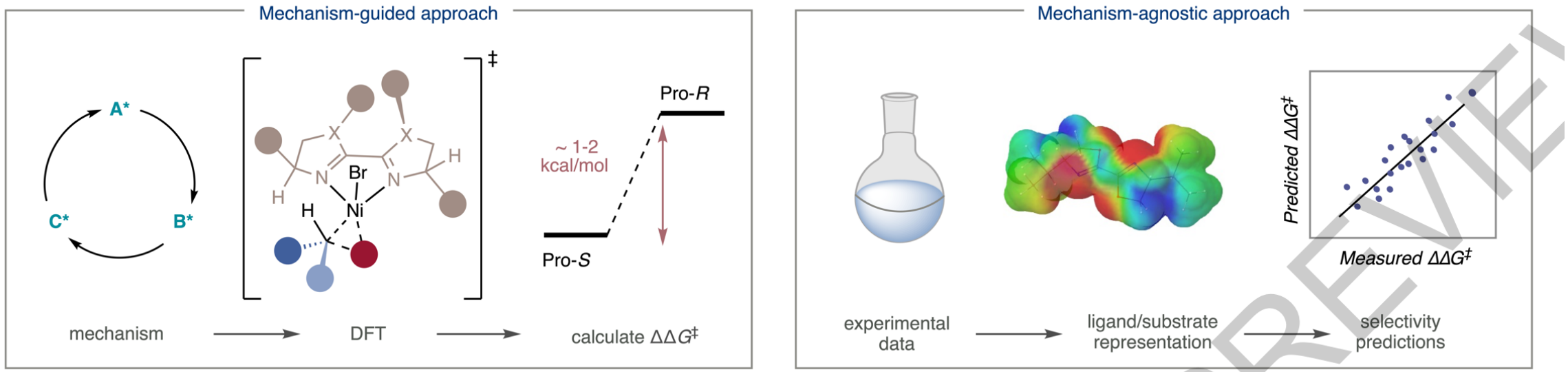
Nature https://doi.org/10.1038/ s41586-026-10239-7 (2026)

Content

1. Introduction

Introduction

a Computational approaches for understanding and predicting enantioselectivity in asymmetric catalysis

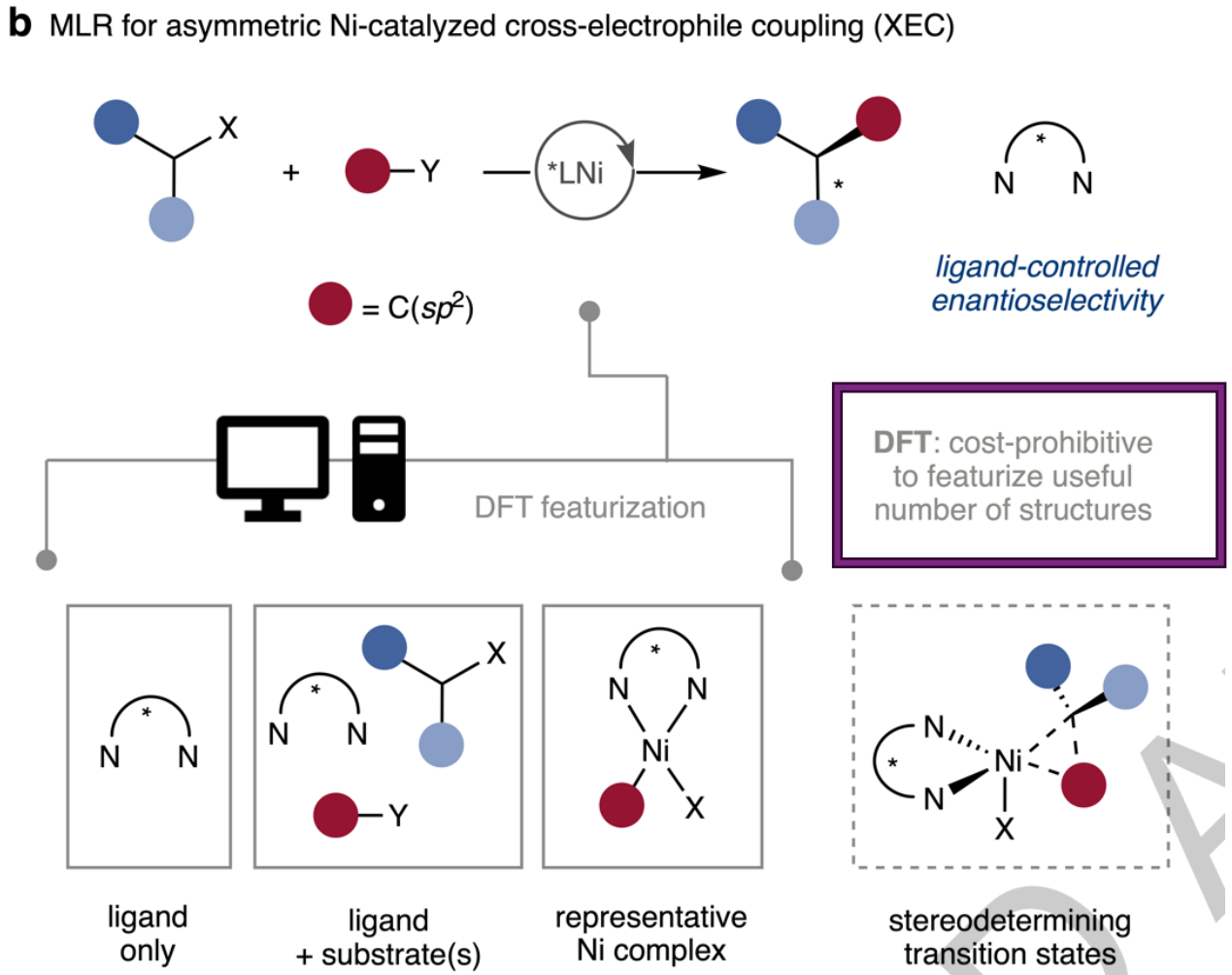


Difficulty: use DFT to predict enantioselectivity ➡ **MLR** (multivariate linear regression)

1. Small energy difference
2. Conformers
3. Competing and off-cycle pathway

- Sparse data (one ligand-multi subs / multi ligands-one sub)
- Different mechanism (different enantio-ds)
- Descriptor can't describe cat-sub interactions
- Need DFT level results to train

Introduction

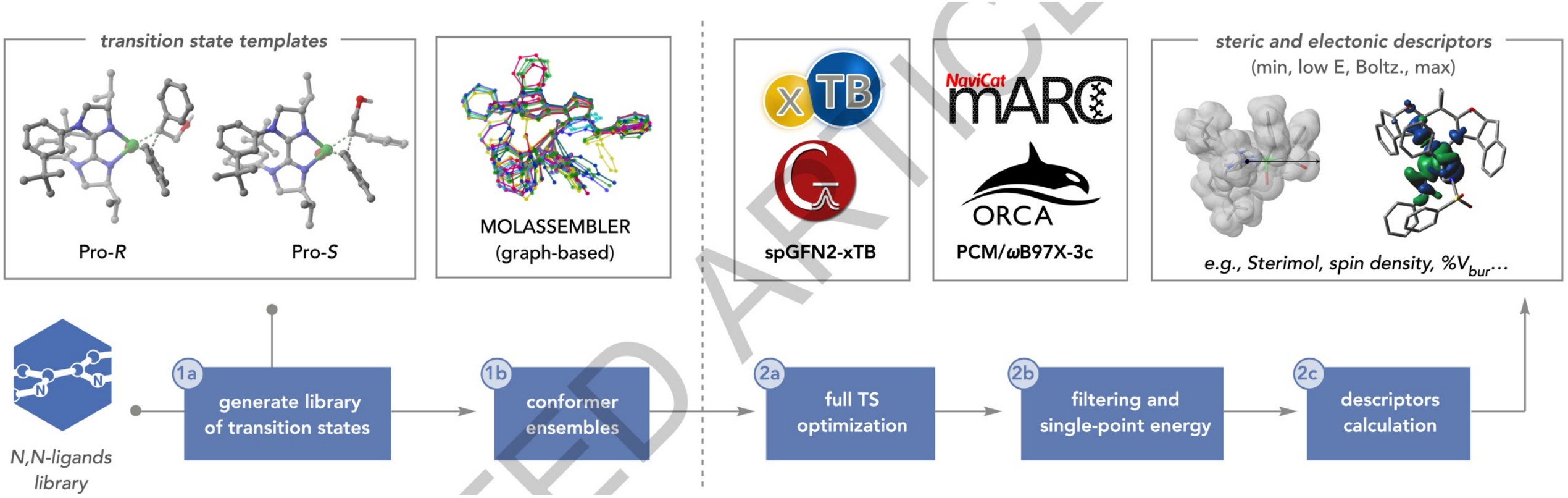


Different sub needs different ligand to control *ee*

The goal of this work is to maximize useful information extracted from sparse data, especially for emerging reactions where only limited data are available. And can also used for other reactions with same mechanism

2. Results

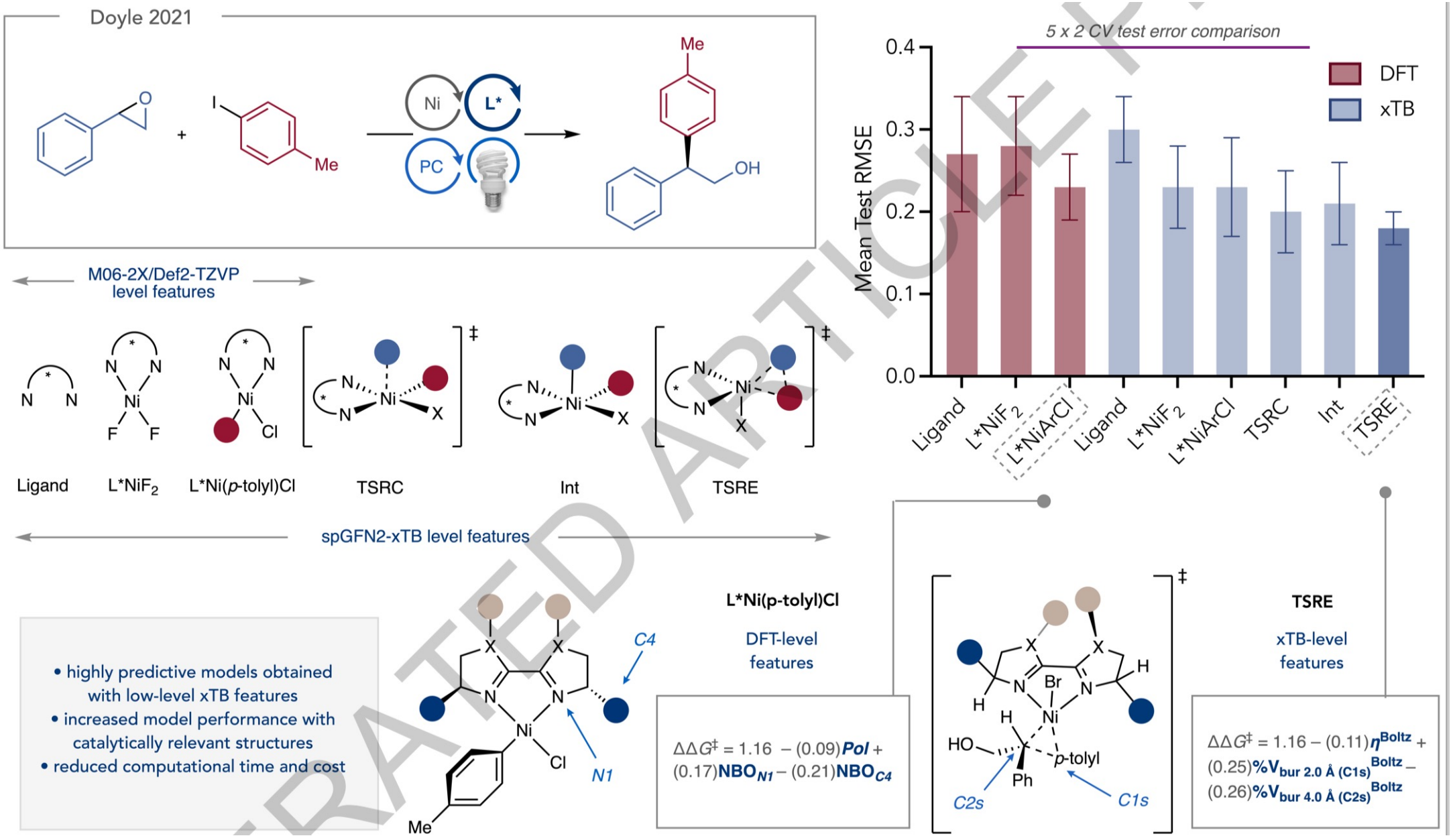
Results—Computational workflow and benchmarking



[1a] AaronTools: *WIREs Comput. Mol. Sci.* **11**, e1510 (2021).
[1b] Molassembler: *J. Chem. Inf. Model.* **60**, 3884–3900 (2020).
[2a] A script to use xtb-gaussian, https://github.com/aspuru-guzik-group/xtb-gaussian.
[2b] MARC: *J. Phys. Chem. Lett.* **15**, 7363–7370 (2024).
[2c] Get Properties and Jupyter notebook: *Digit. Discov.* **4**, 222–233 (2025).

Get 756 parameters

Results—Computational workflow and benchmarking



Results—5 x 2 CV test

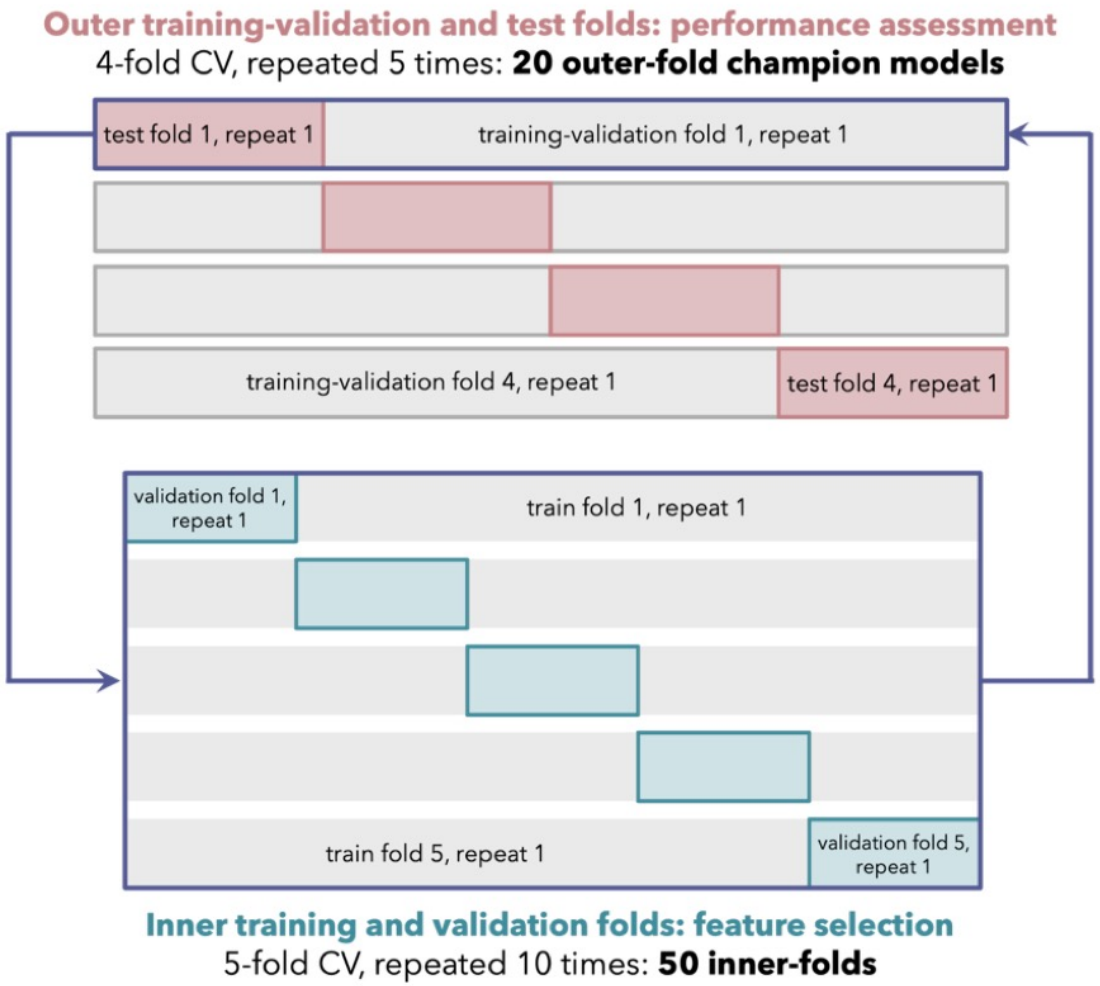
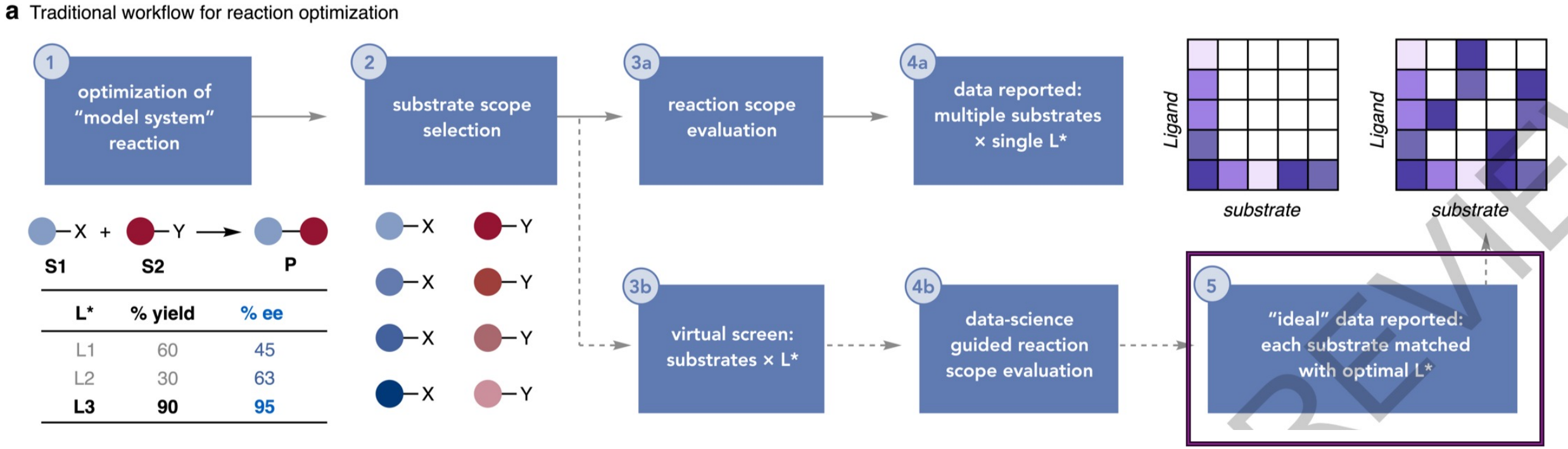


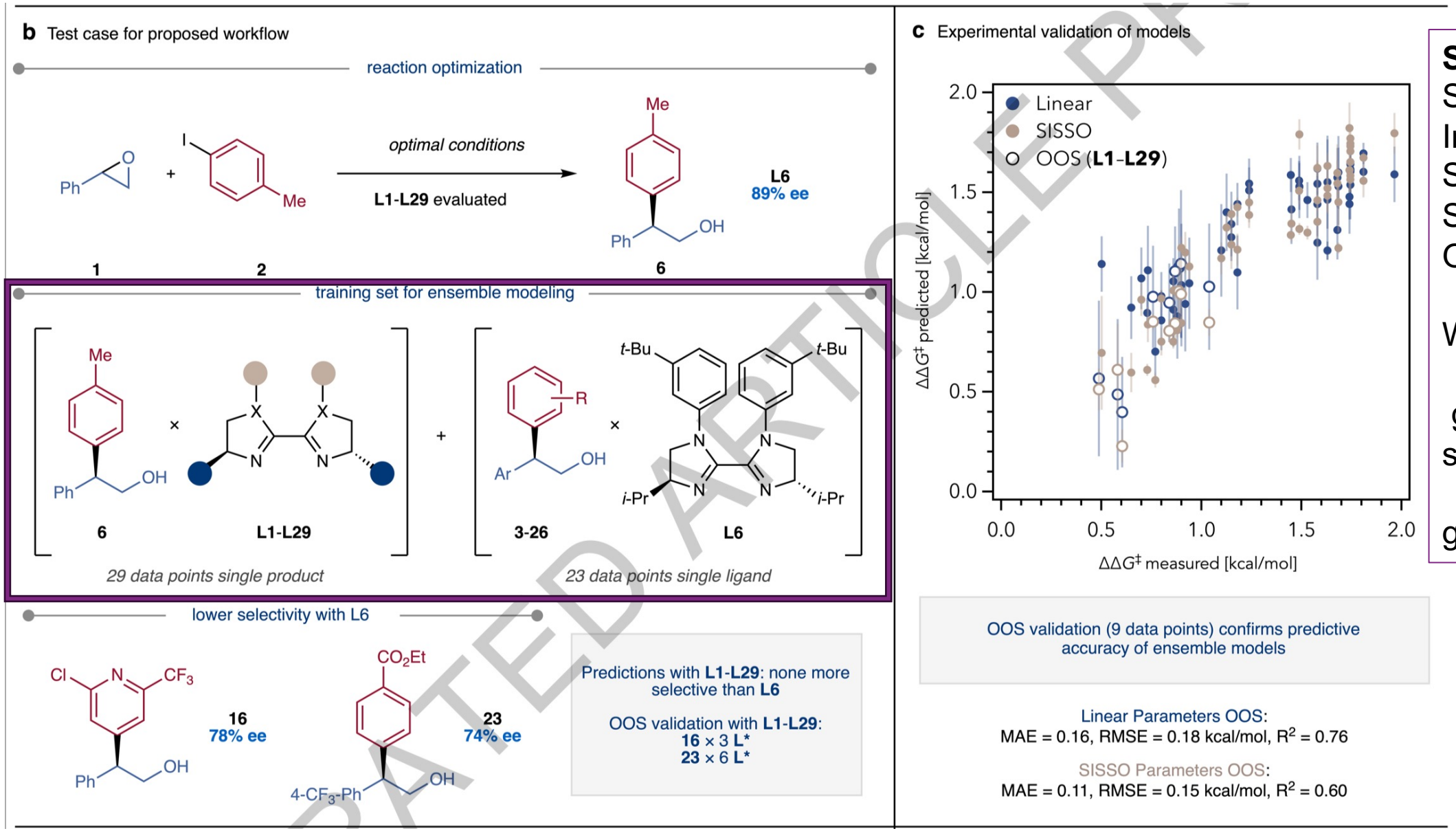
Figure S3. Repeated, stratified, nested *k*-fold CV scheme.

Good for sparse data

Results—Case study 1



Results—Case study 1



SISSO
Sure Independence Screening and Sparsifying Operator

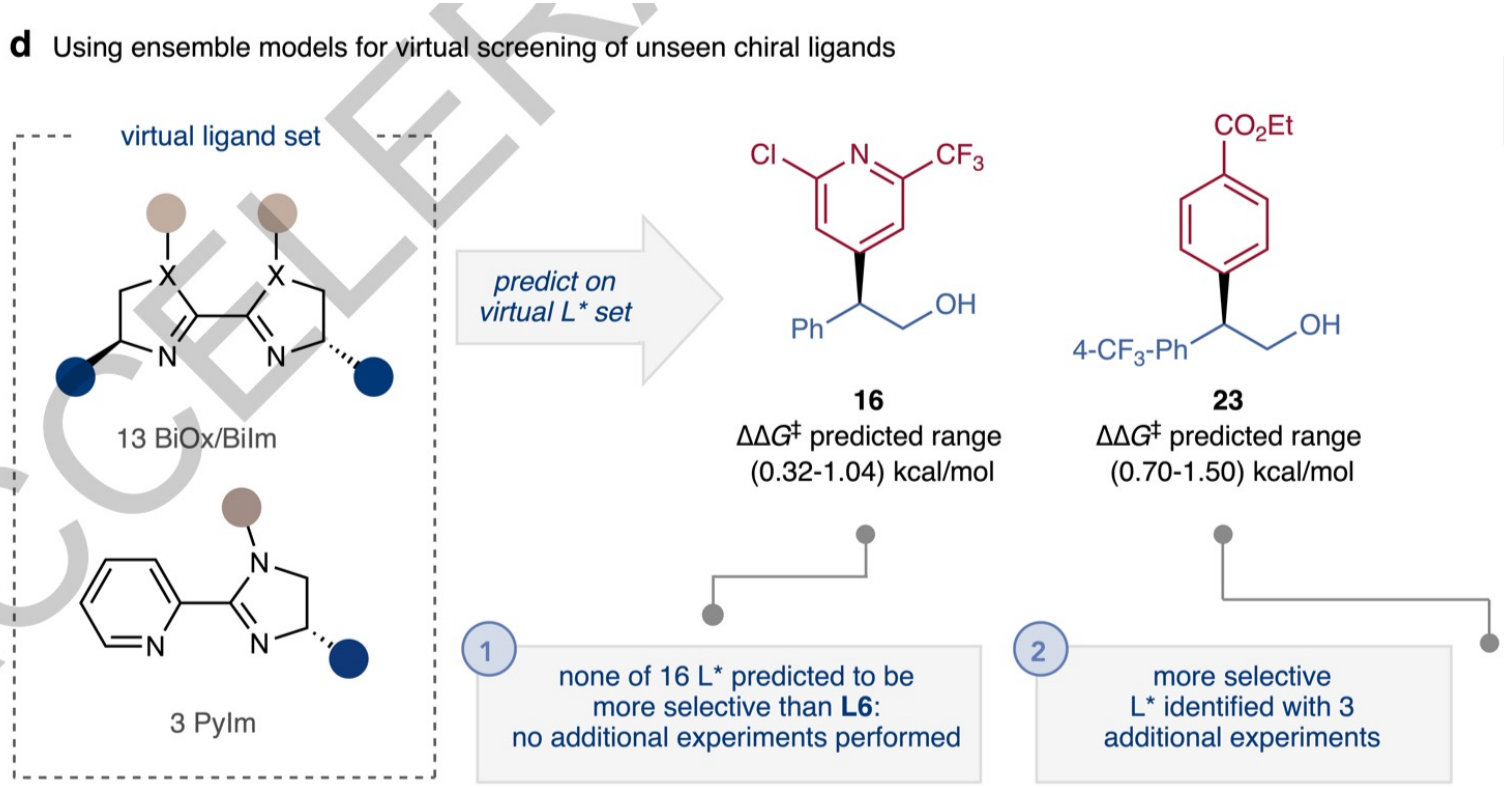
White-box,

good for sparse data

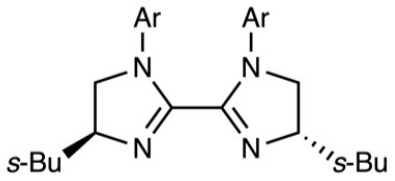
give a formula

OOS
Out-of-Sample

Results—Case study 1



Experimentally evaluated 3 BiIm ligands predicted > 74% ee

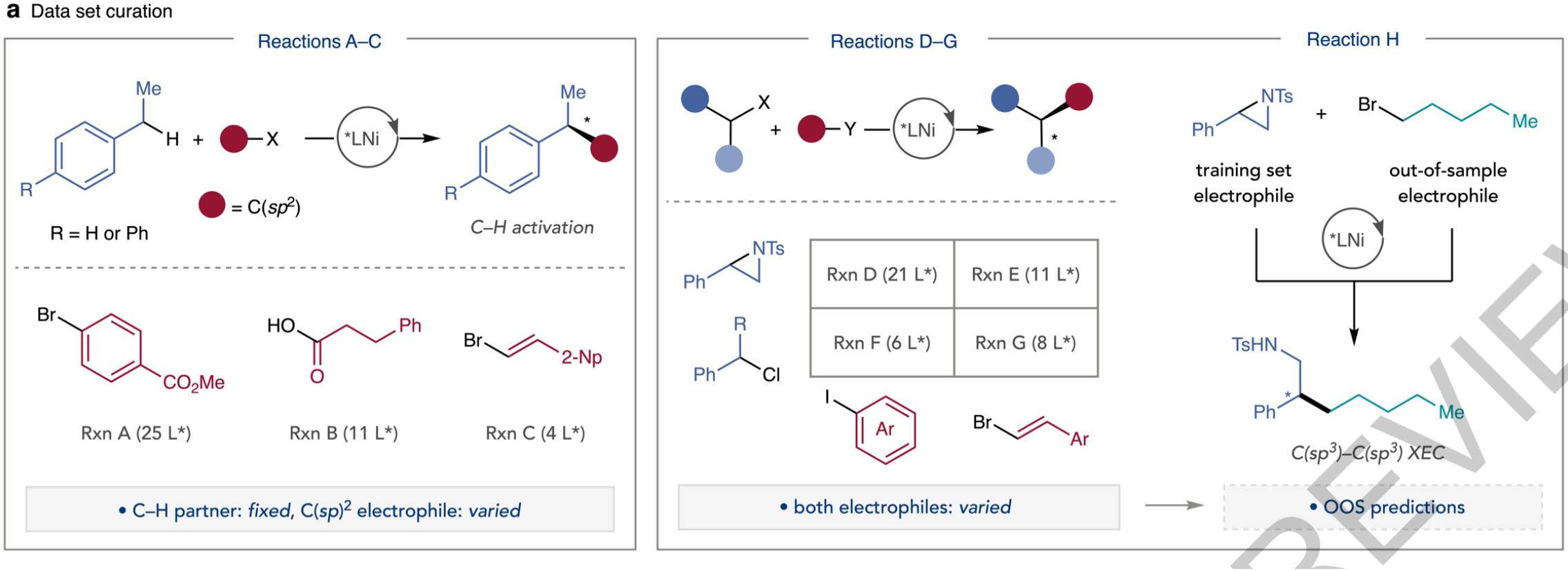


L65 (Ar = 4-Me-Ph)
L66 (Ar = 3,5-*t*-Bu-Ph)
L144 (Ar = 4-*t*-Bu-Ph)

Ligand	ΔΔG‡ predicted (kcal/mol)	ΔΔG‡ measured (kcal/mol)	% ee (measured)
L65	1.45 ± 0.33	1.06	72
L66	1.50 ± 0.18	1.47	85
L144	1.43 ± 0.21	1.02	70

From 74% to 85% *ee*

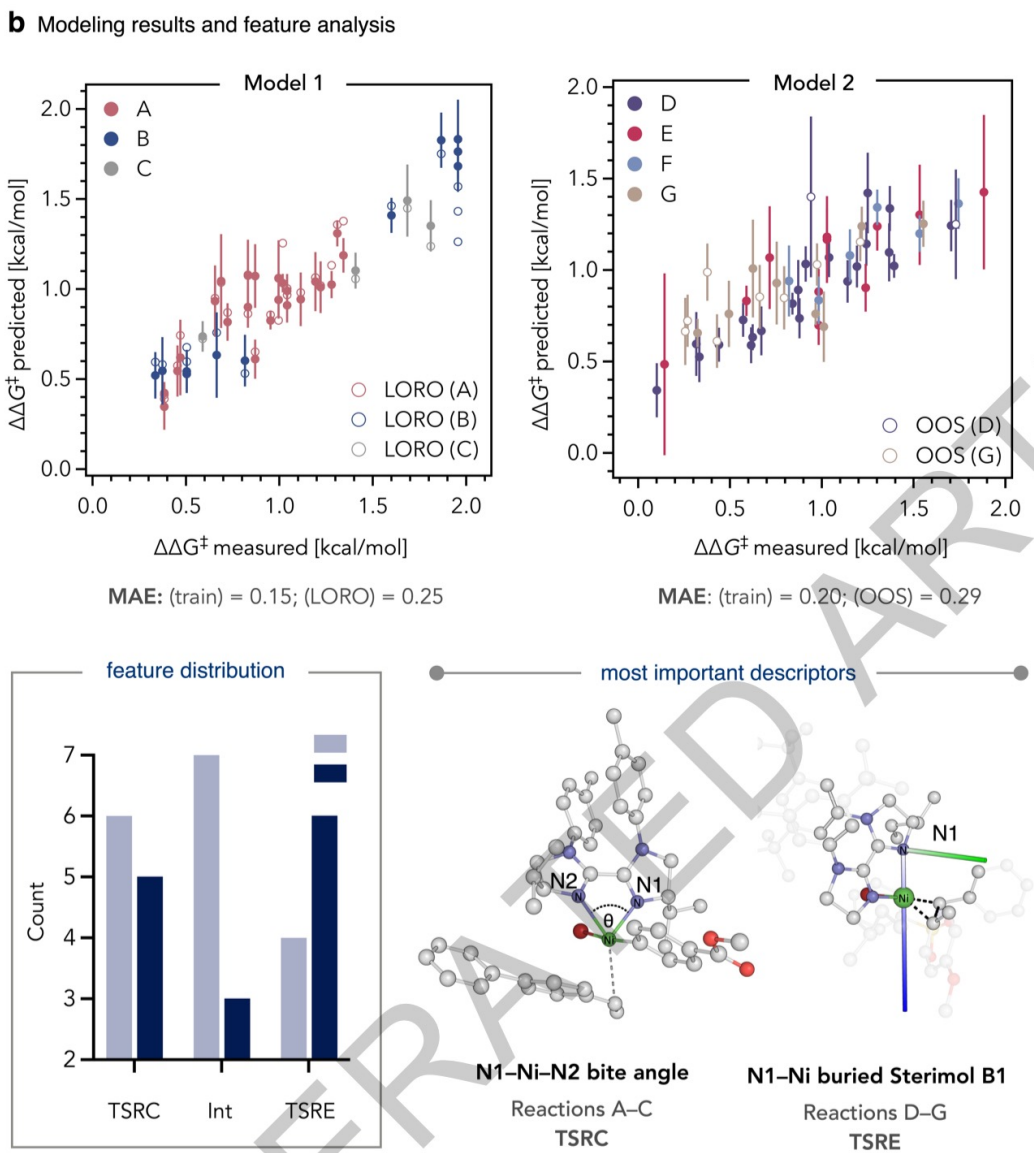
Results—Case study 2

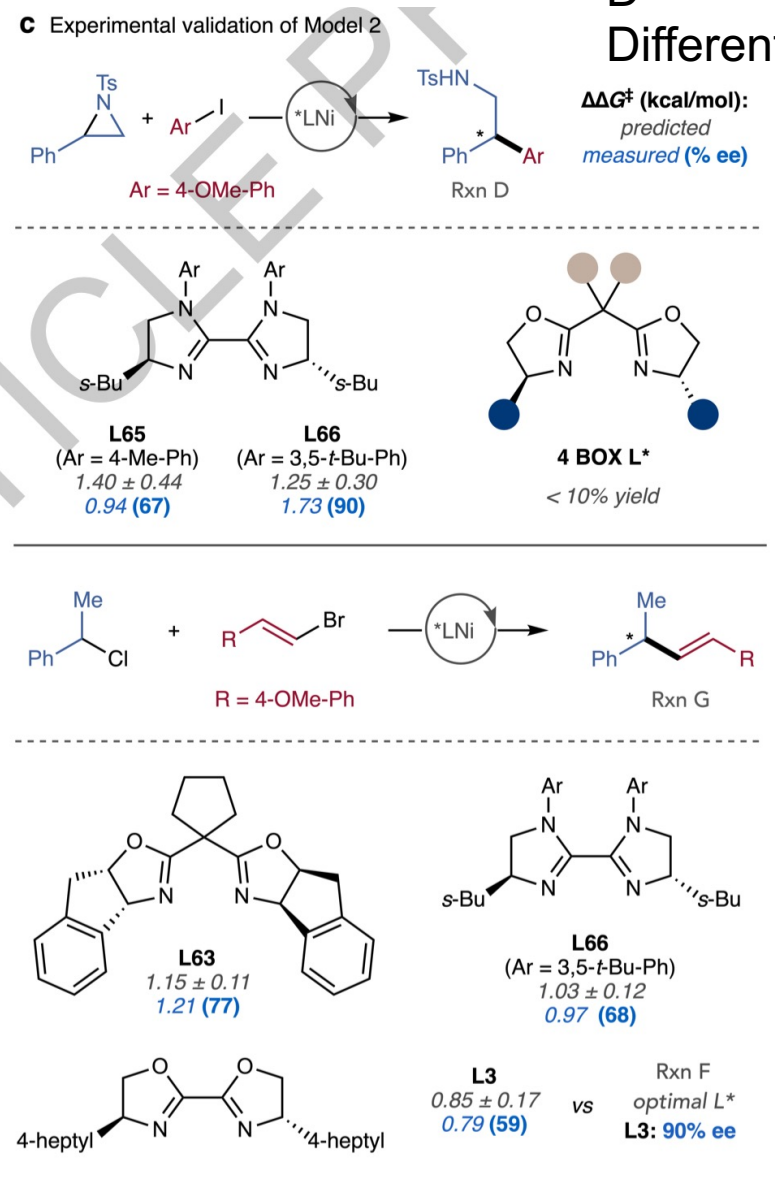


Results—Case study 2

A B --- C
D E F --- G

LORO
leave-one-reaction-out



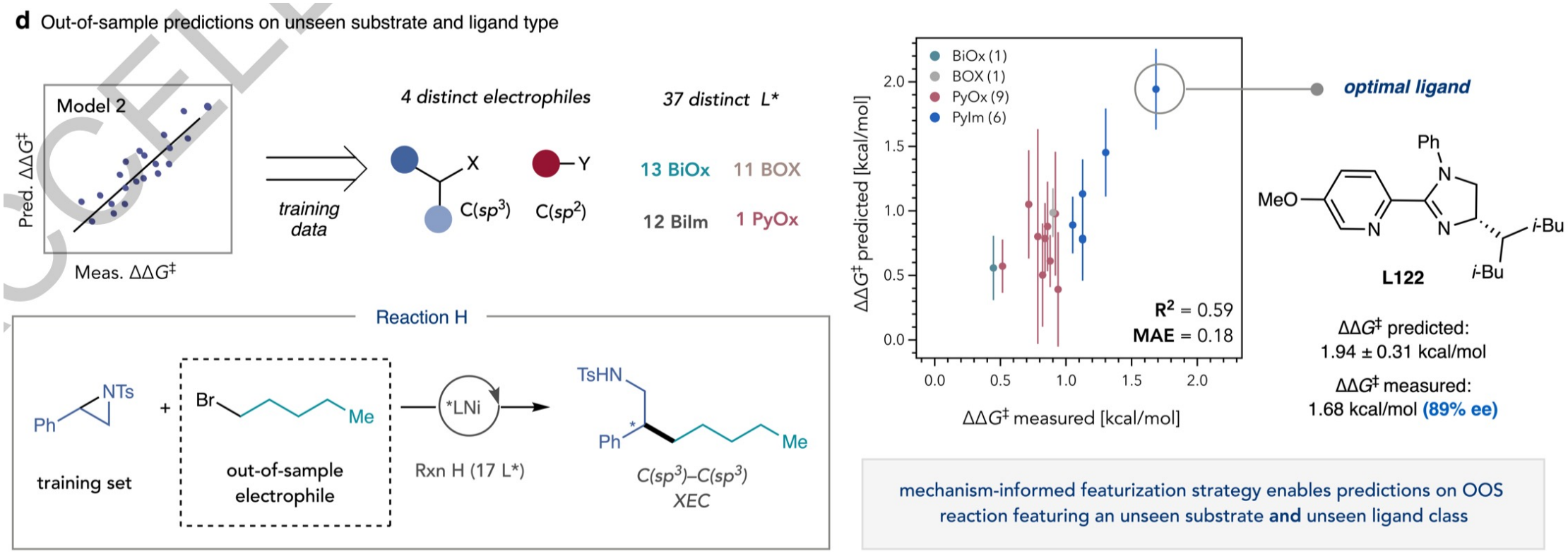


D
Different sub

G

Results—Case study 2

Unseen ligand Use DEFG --- H



mechanism-informed featurization strategy enables predictions on OOS reaction featuring an unseen substrate **and** unseen ligand class

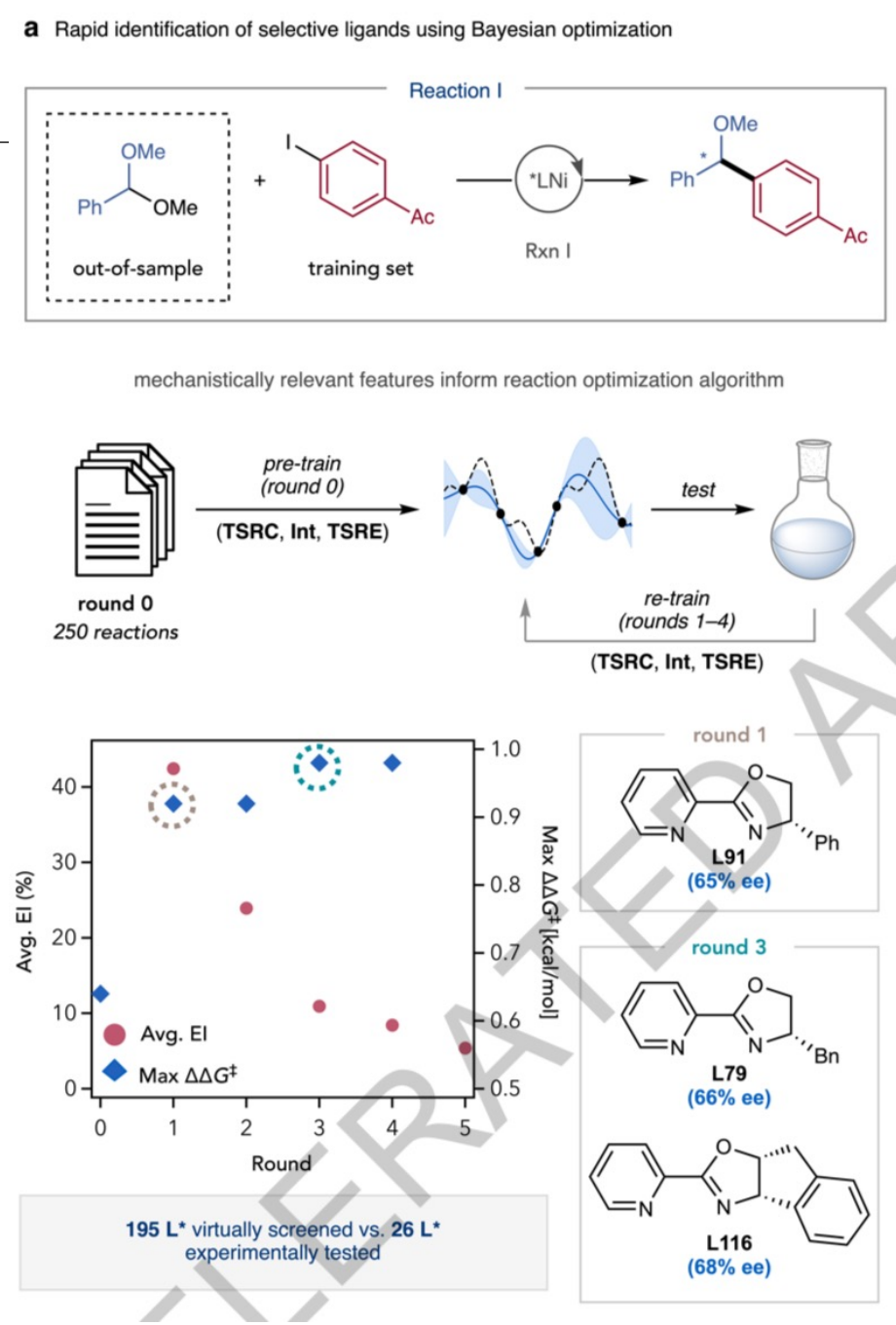
Results—Case study 3

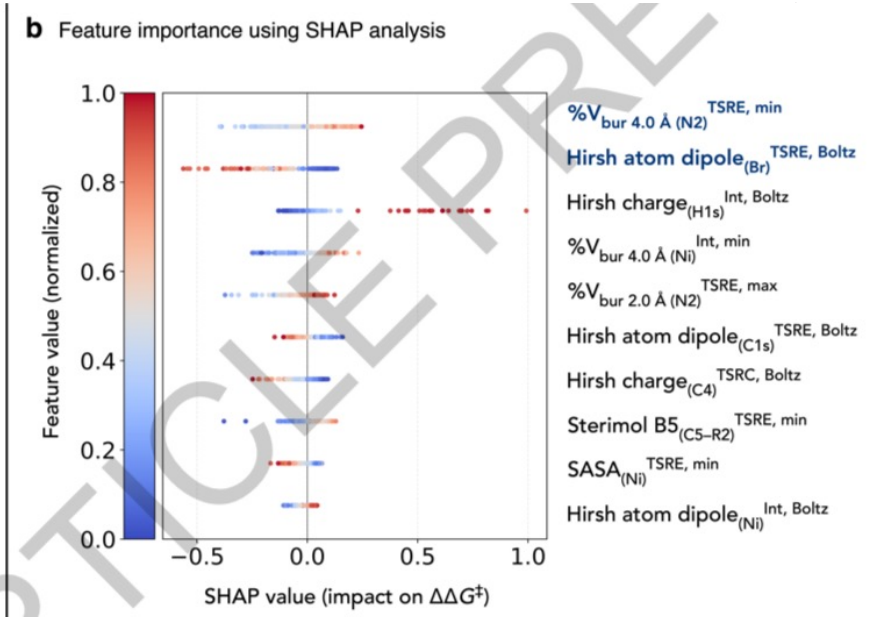
Bayesian optimization:
a sequential strategy uses a probabilistic model to efficiently identify optimal conditions with as few experiments as possible.

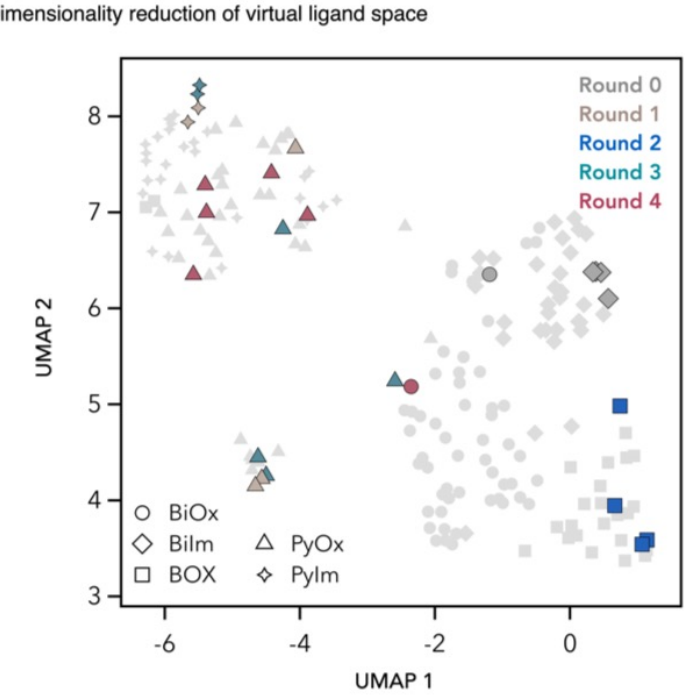
GPR:
a probabilistic regression method that predicts both the expected value and the uncertainty of a target, making it well suited for Bayesian optimization with sparse data.

SHAP:
a model-interpretation method that quantifies how much each feature contributes to a prediction.

UMAP







3. Summary

Summary

We have developed a workflow to predict the enantioselectivity of Ni-catalyzed C(sp^3)-couplings where multiple reaction components are varied simultaneously using only sparse data for training. Parametrizing regression models with descriptors extracted from catalytically relevant structures increases the models' accuracy and domain of applicability. Owing to the mechanistic complexity of Ni-catalyzed C(sp^3)-couplings, the input representation is large, which makes identifying the most important descriptors challenging. We have addressed this problem *via* repeated, nested *k*-fold cross-validation and ensemble modeling, but less expensive approaches are still desired. The low-cost, mechanism-informed features obtained *via* our workflow allowed us to identify optimal catalyst–substrate combinations and make accurate predictions on out-of-sample ligand and substrate classes. The added value of this approach is that limited data of known reactions may be used to train models applicable to unknown transformations. While we envision it will facilitate the development of transfer learning approaches for the discovery of novel synthetic methods, alternative solutions to extract detailed mechanistic information and unambiguously identify the nature of the enantiodetermining step for a given combination of ligand and substrates will further expand the scope of this workflow.

Sparse data prediction and useful methods

Workflow

Thank You