



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



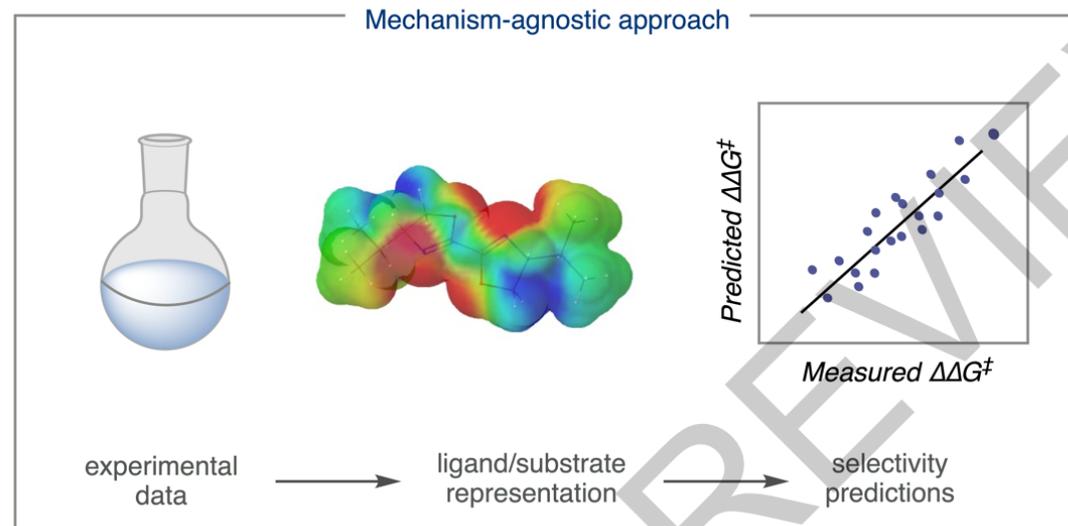
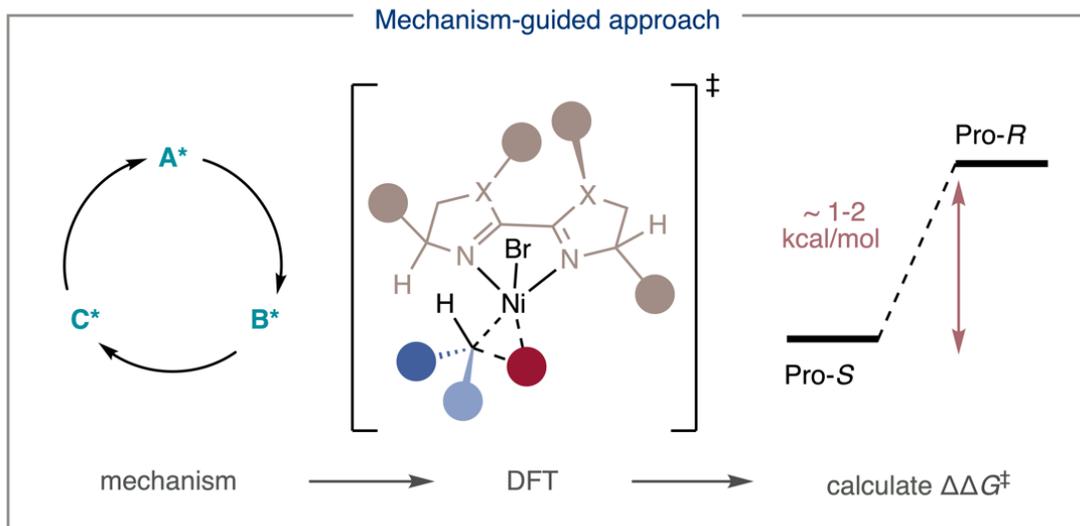
1. Introduction

2. Results

3. Summary

1. 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

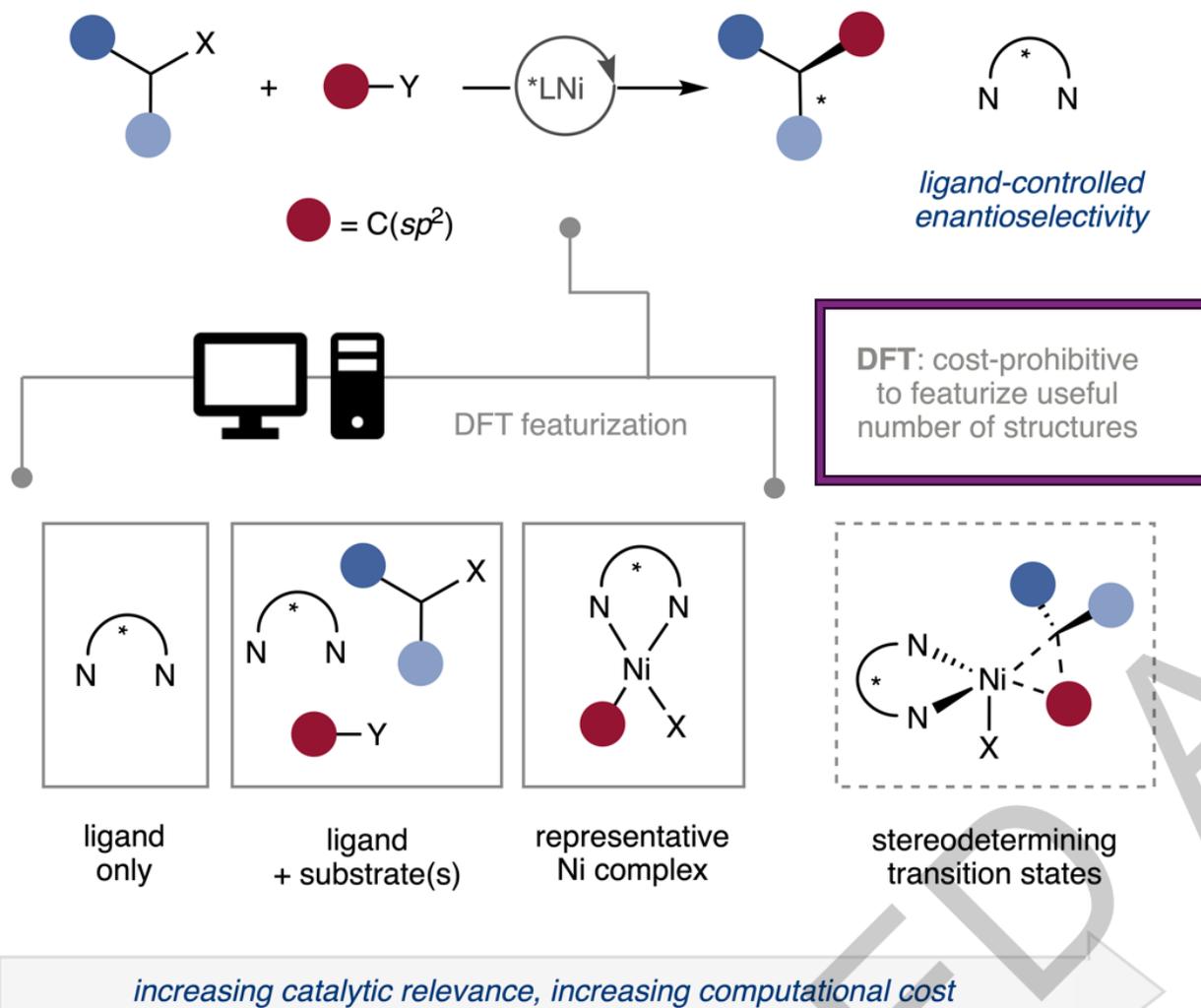
Sparse data (one ligand-multi subs / multi ligands-one sub)

Different mechanism (different enantio-ds)

Descriptor can't describe cat-sub interactions

Need DFT level results to train

b MLR for asymmetric Ni-catalyzed cross-electrophile coupling (XEC)

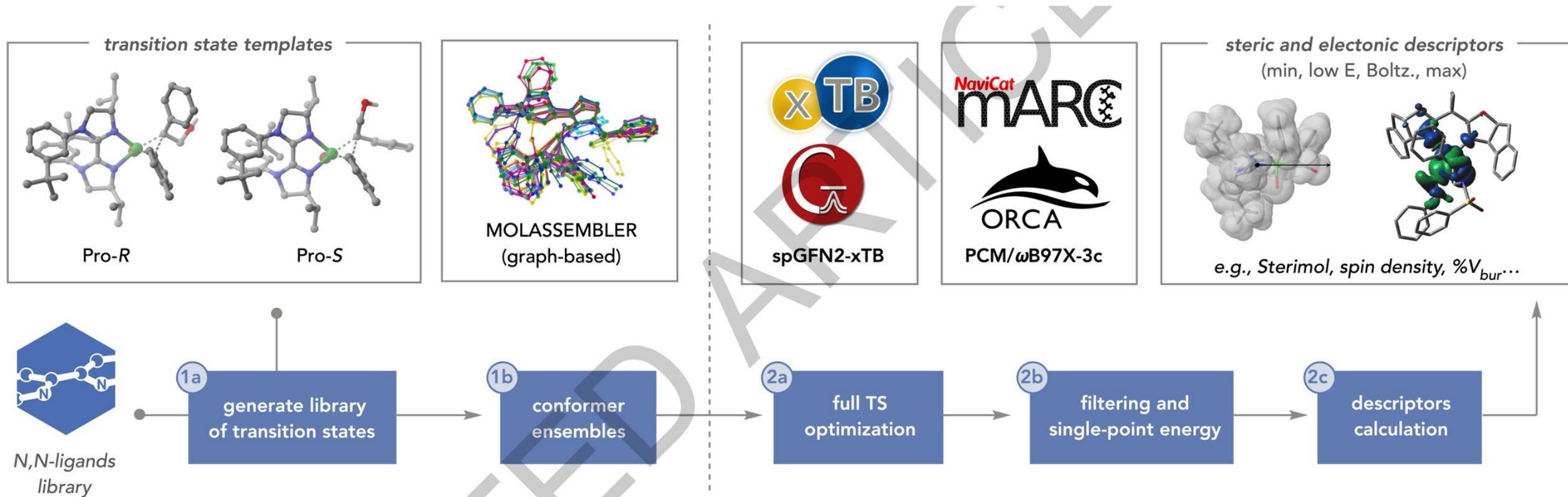


Different sub needs different ligand to control ee

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).

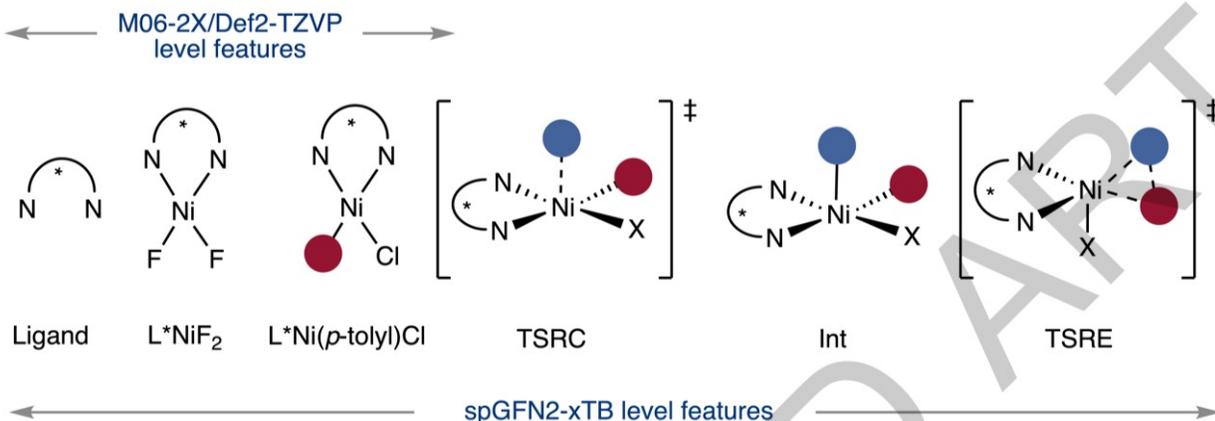
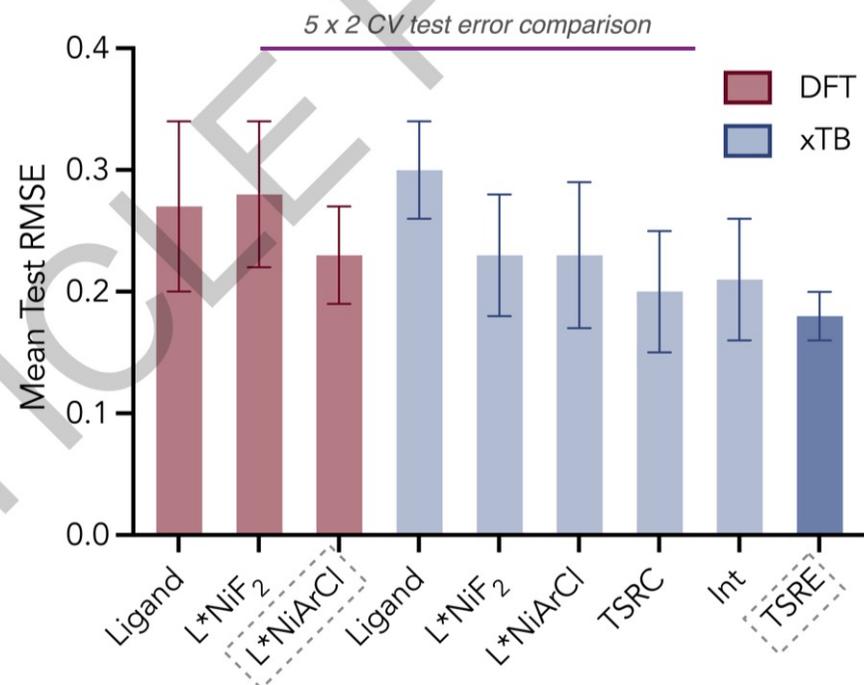
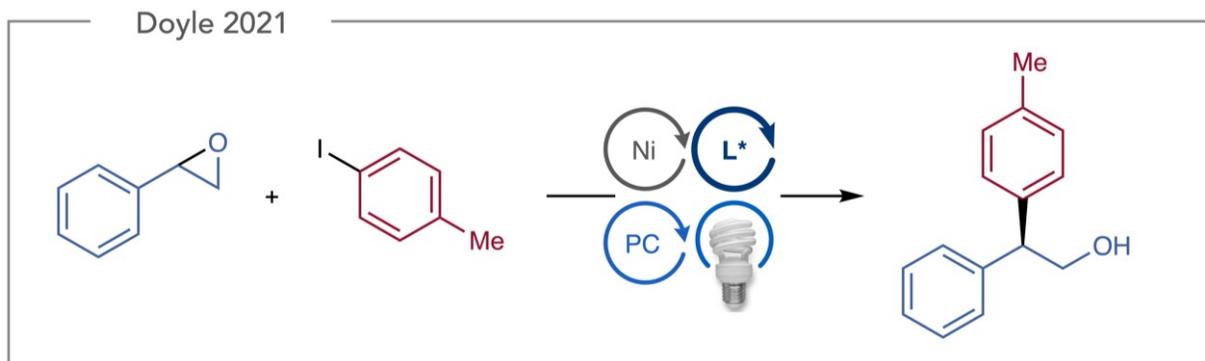
2a A script to use xtb-gaussian, <https://github.com/aspuru-guzik-group/xtb-gaussian>.

2b MARC: *J. Phys. Chem. Lett.* **15**, 7363–7370 (2024).

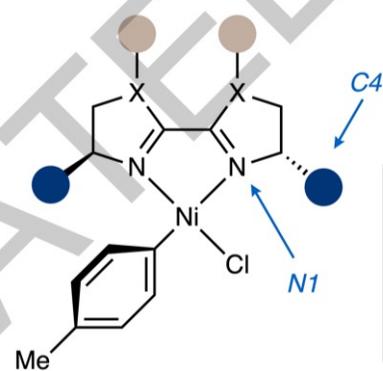
2c Get Properties and Jupyter notebook: *Digit. Discov.* **4**, 222–233 (2025).

Get 756 parameters

Results—Computational workflow and benchmarking

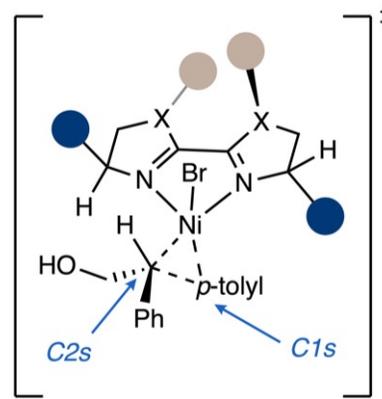


- highly predictive models obtained with low-level xTB features
- increased model performance with catalytically relevant structures
- reduced computational time and cost



L*Ni(p-tolyl)Cl
DFT-level features

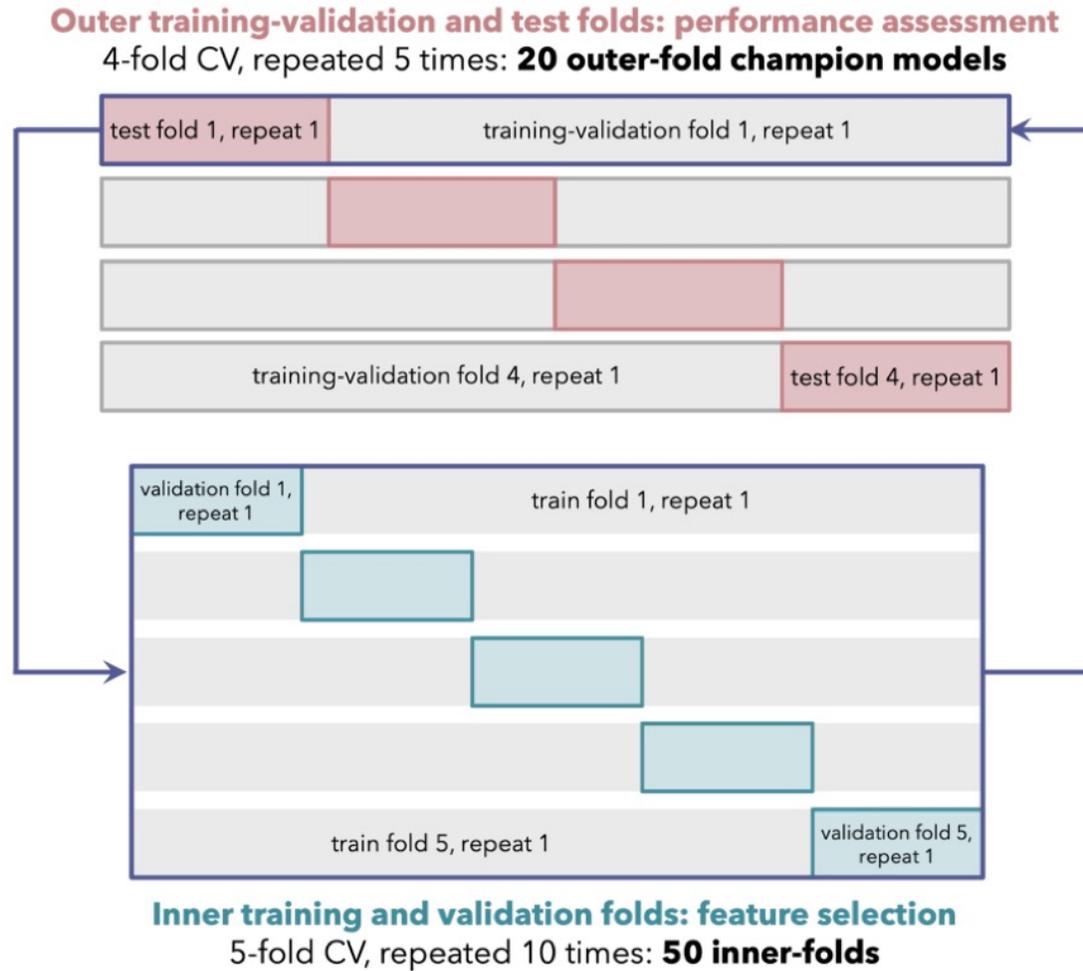
$$\Delta\Delta G^\ddagger = 1.16 - (0.09)Pol + (0.17)NBO_{N1} - (0.21)NBO_{C4}$$



TSRE
xTB-level features

$$\Delta\Delta G^\ddagger = 1.16 - (0.11)\eta^{Boltz} + (0.25)\%V_{bur\ 2.0\ \text{\AA}}(C1s)^{Boltz} - (0.26)\%V_{bur\ 4.0\ \text{\AA}}(C2s)^{Boltz}$$

Results—5 x 2 CV test



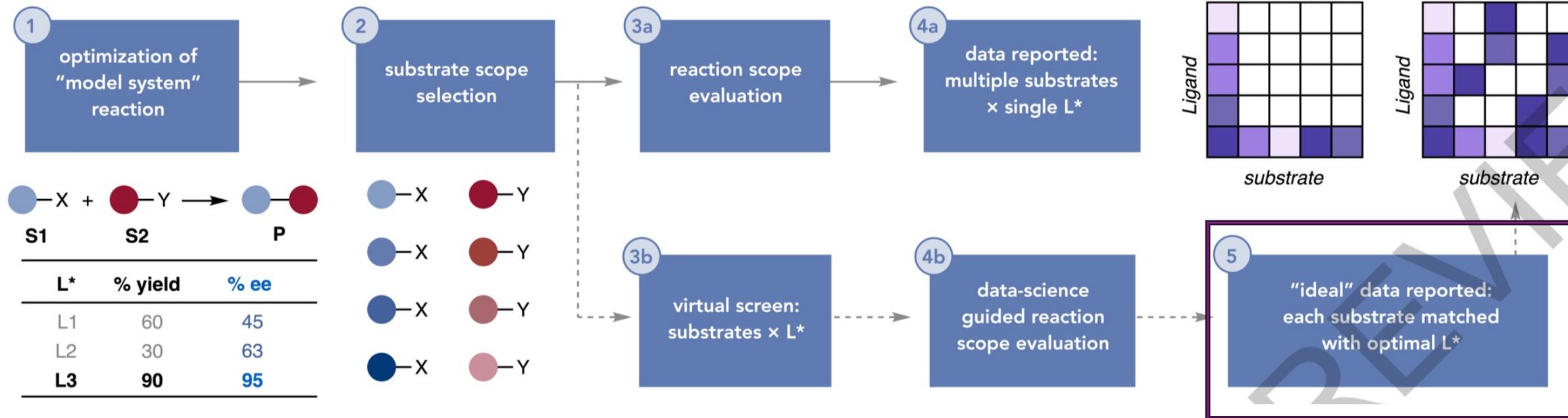
Good for sparse data

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

Results—Case study 1



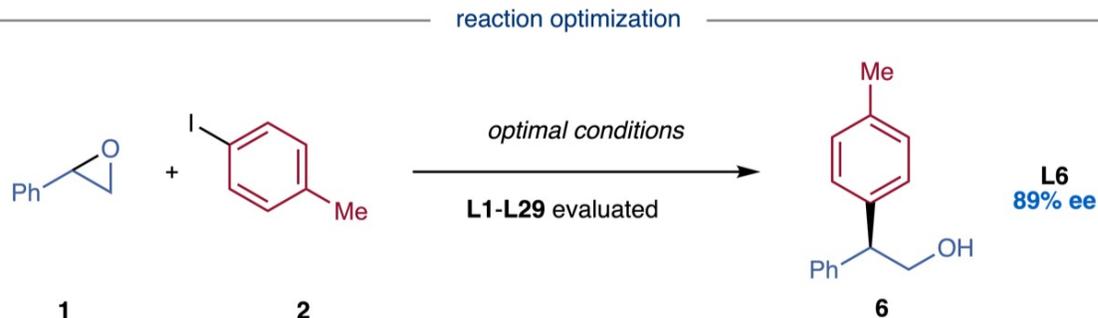
a Traditional workflow for reaction optimization



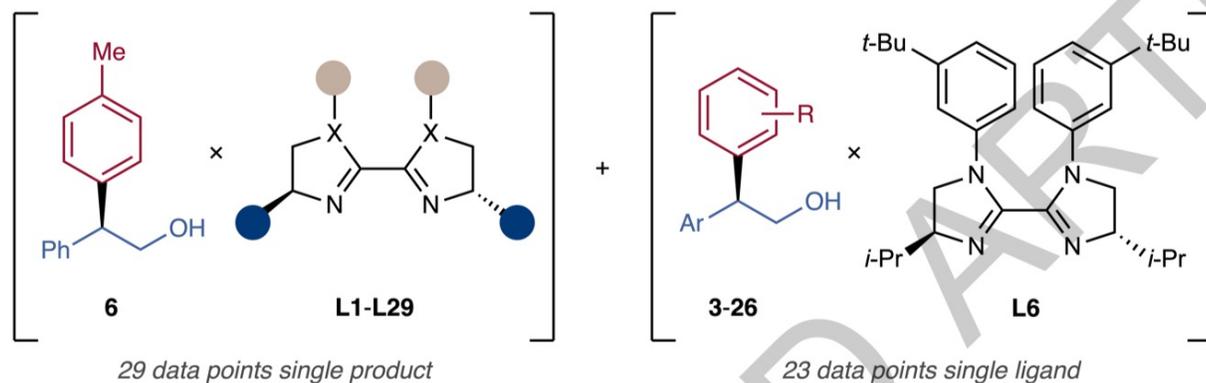
Results—Case study 1



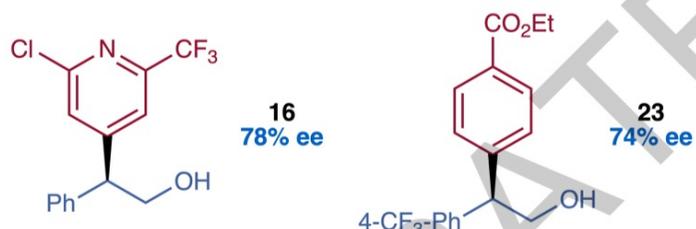
b Test case for proposed workflow



training set for ensemble modeling



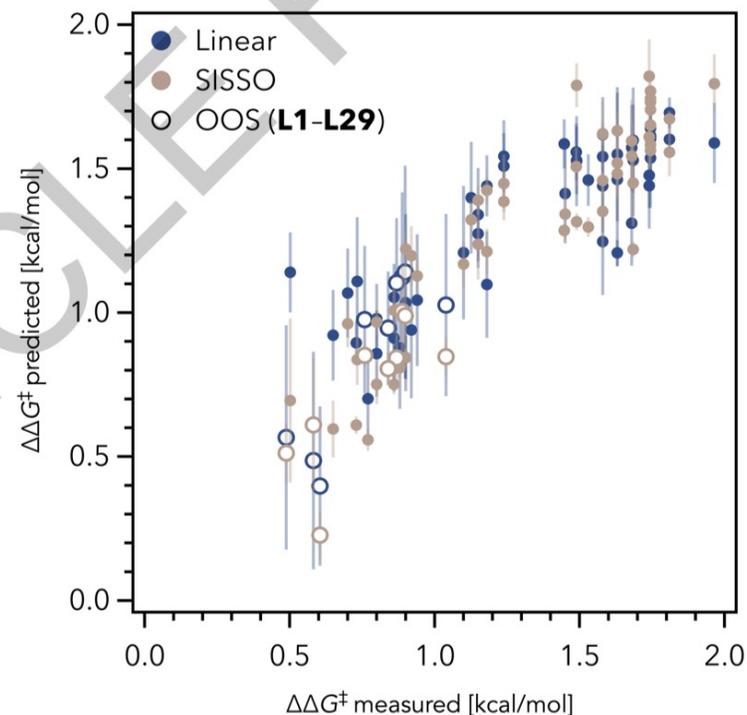
lower selectivity with L6



Predictions with **L1-L29**: none more selective than **L6**

OOS validation with **L1-L29**:
 16 × 3 **L***
 23 × 6 **L***

c Experimental validation of models



OOS validation (9 data points) confirms predictive accuracy of ensemble models

Linear Parameters OOS:
 MAE = 0.16, RMSE = 0.18 kcal/mol, $R^2 = 0.76$

SISSO Parameters OOS:
 MAE = 0.11, RMSE = 0.15 kcal/mol, $R^2 = 0.60$

SISSO

Sure
 Independence
 Screening and
 Sparsifying
 Operator

White-box,

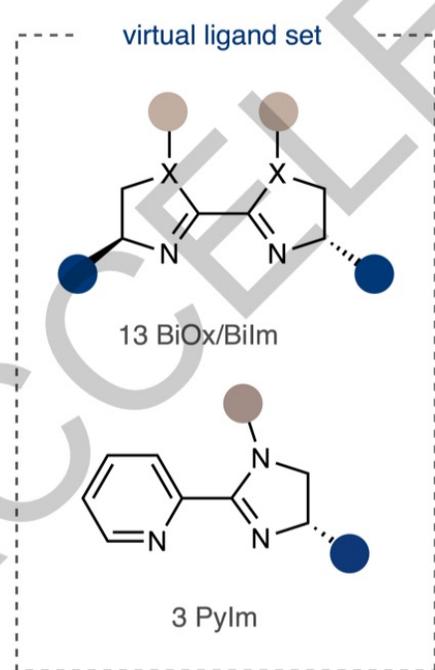
good for
 sparse data

give a formula

OOS

Out-of-Sample

d Using ensemble models for virtual screening of unseen chiral ligands



predict on
virtual L* set



$\Delta\Delta G^\ddagger$ predicted range
(0.32-1.04) kcal/mol

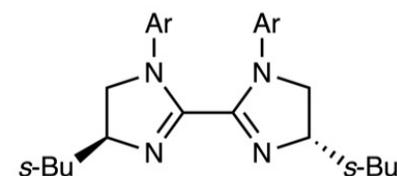
1 none of 16 L* predicted to be more selective than **L6**: no additional experiments performed



$\Delta\Delta G^\ddagger$ predicted range
(0.70-1.50) kcal/mol

2 more selective L* identified with 3 additional experiments

Experimentally evaluated 3 Bilm ligands predicted > 74% ee

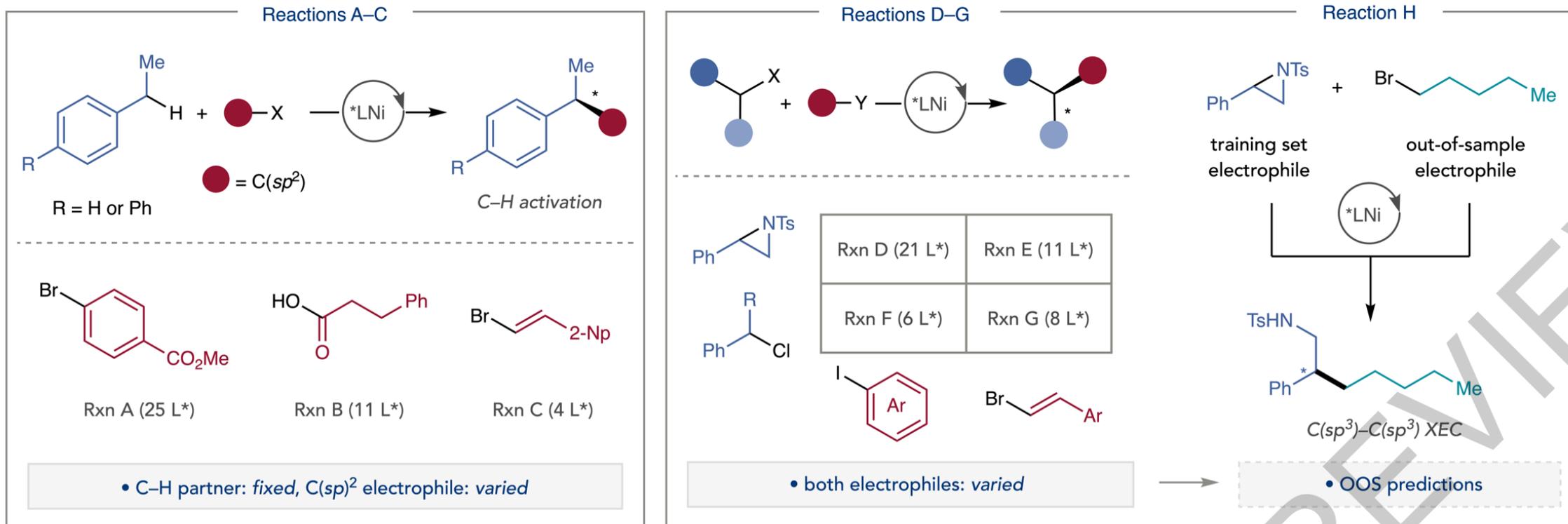


L65 (Ar = 4-Me-Ph)
L66 (Ar = 3,5-*t*-Bu-Ph)
L144 (Ar = 4-*t*-Bu-Ph)

Ligand	$\Delta\Delta G^\ddagger$ predicted (kcal/mol)	$\Delta\Delta G^\ddagger$ 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

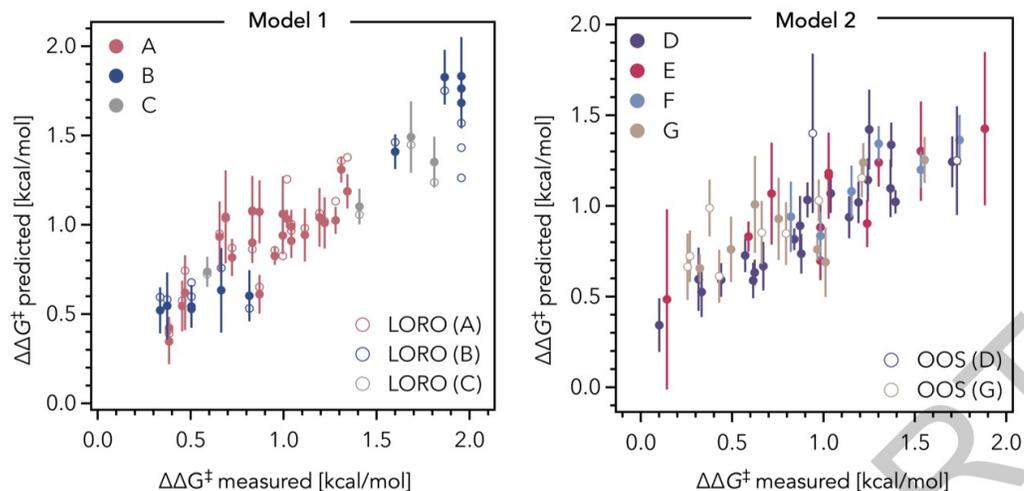
a Data set curation



Results—Case study 2



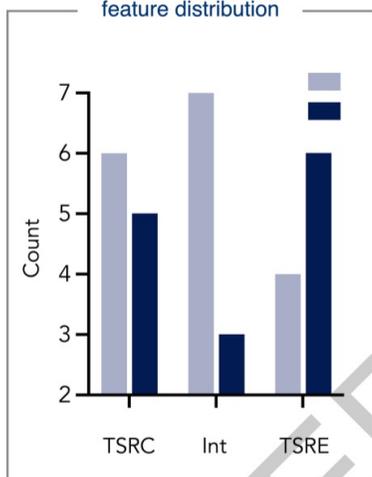
b Modeling results and feature analysis



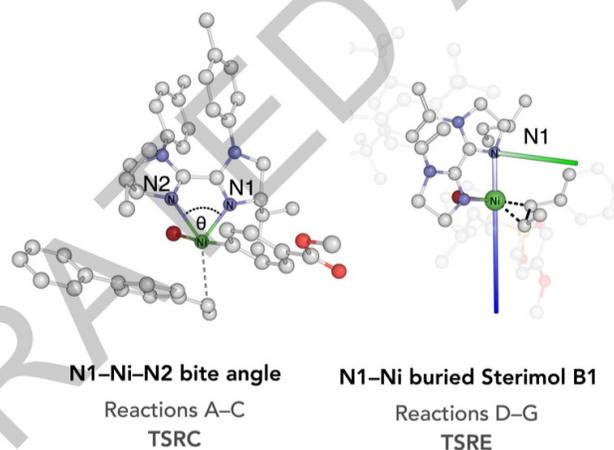
MAE: (train) = 0.15; (LORO) = 0.25

MAE: (train) = 0.20; (OOS) = 0.29

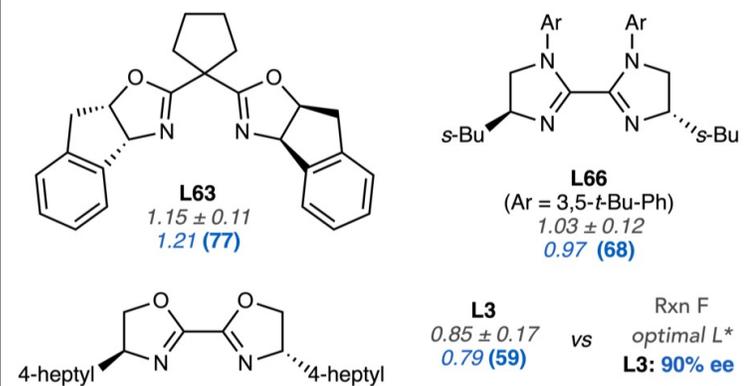
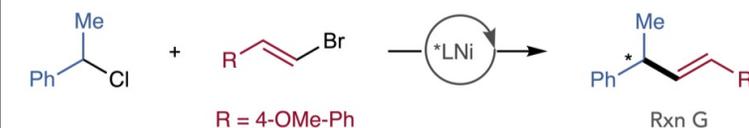
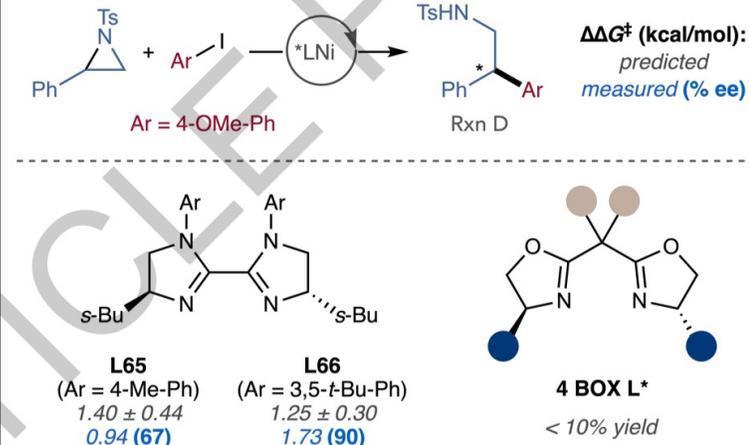
feature distribution



most important descriptors



c Experimental validation of Model 2



D
Different sub

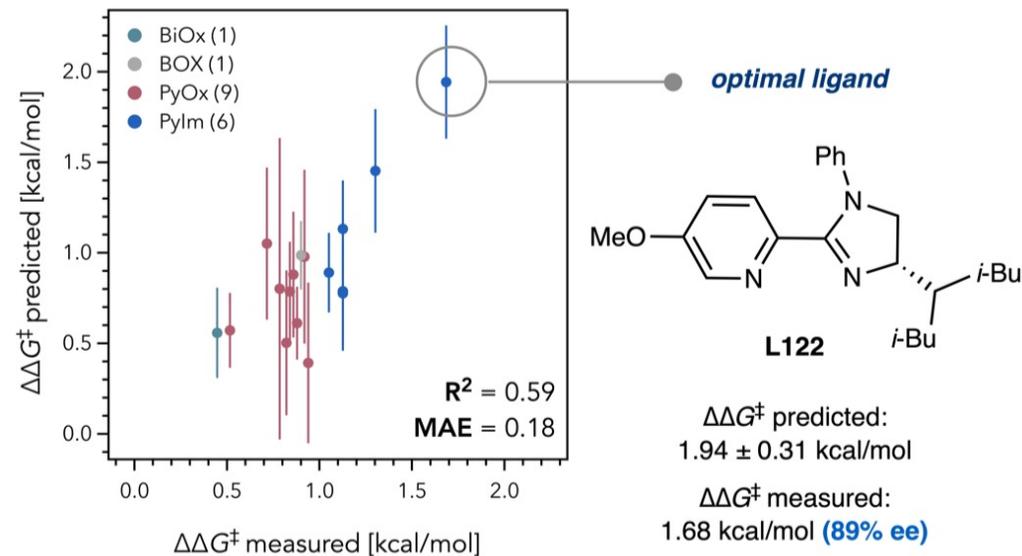
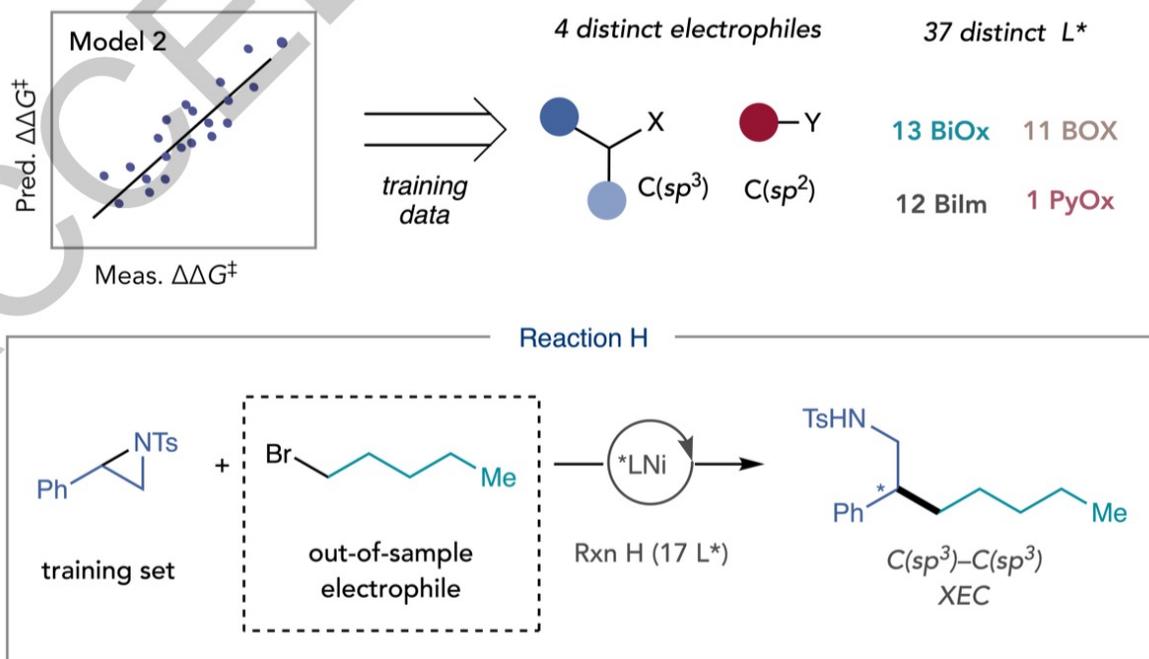
AB --- C
DEF --- G

LORO
leave-one-reaction-out

G

Unseen ligand Use DEFG --- H

d Out-of-sample predictions on unseen substrate and ligand type



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:

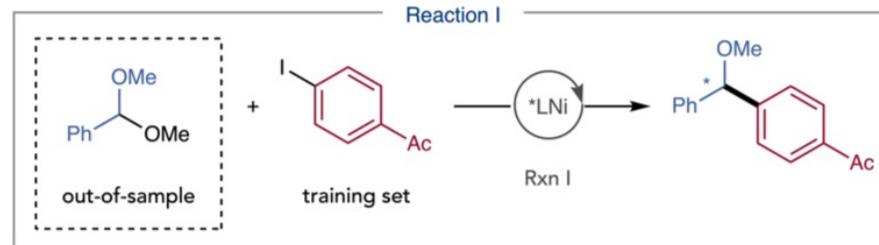
a probabilistic regression method that predicts both the expected value and the uncertainty of a target, making it well suited for Bayesian optimization with sparse data.

SHAP:

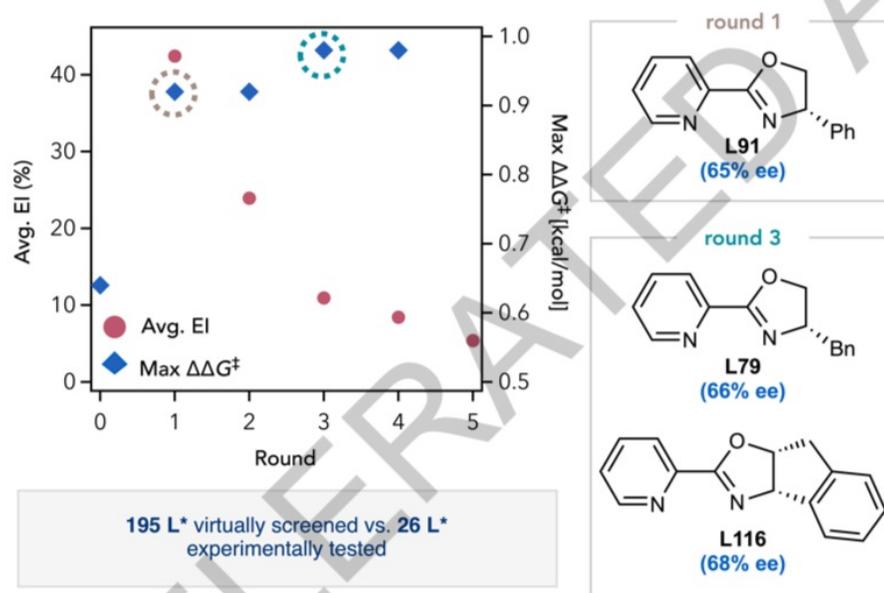
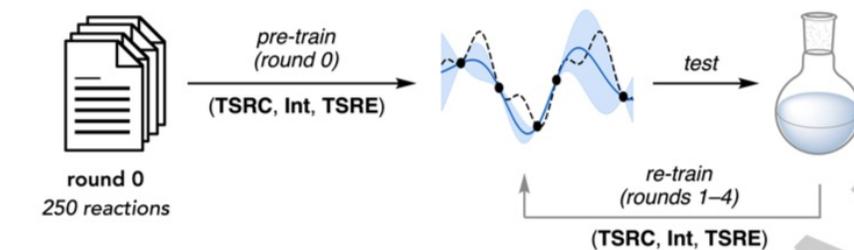
a model-interpretation method that quantifies how much each feature contributes to a prediction.

UMAP

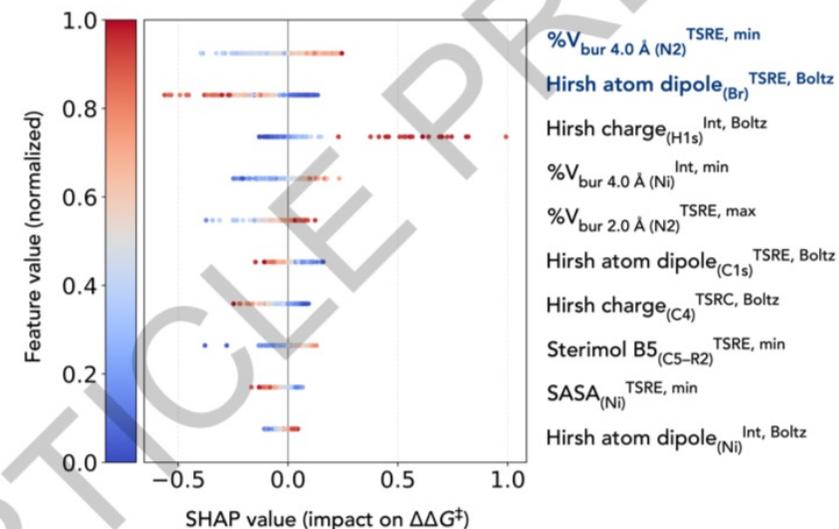
a Rapid identification of selective ligands using Bayesian optimization



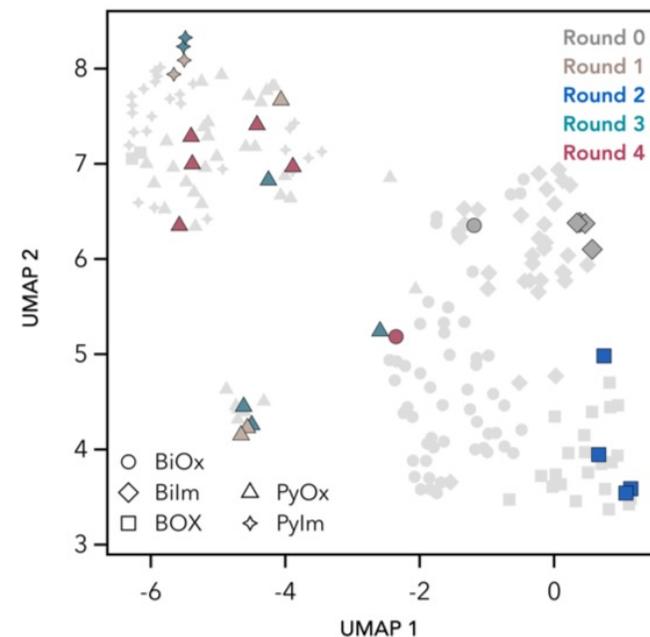
mechanistically relevant features inform reaction optimization algorithm



b Feature importance using SHAP analysis



c Dimensionality reduction of virtual ligand space



3. 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