UNIVERSITY OF CAMBRIDGE

Practical Machine Learning for Organic Small Molecule Modeling

---

Machine Learning Modalities for Materials Science Workshop

16 May 2024

Emma King-Smith

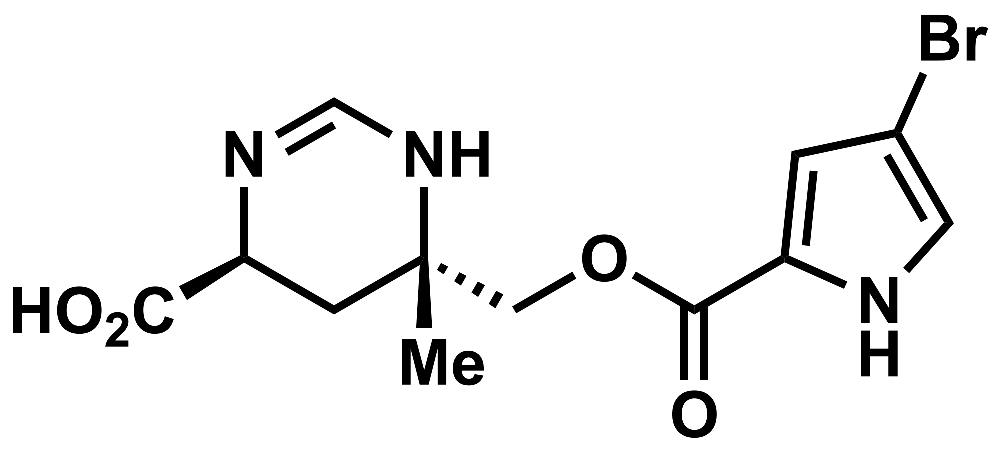
What Do Synthetic Chemists Want?

Make Molecules!

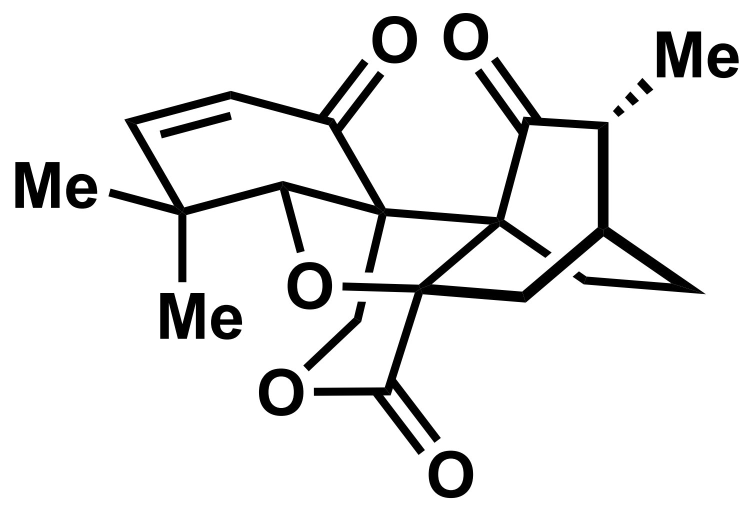
What Do Synthetic Chemists Want?

Make Molecules!

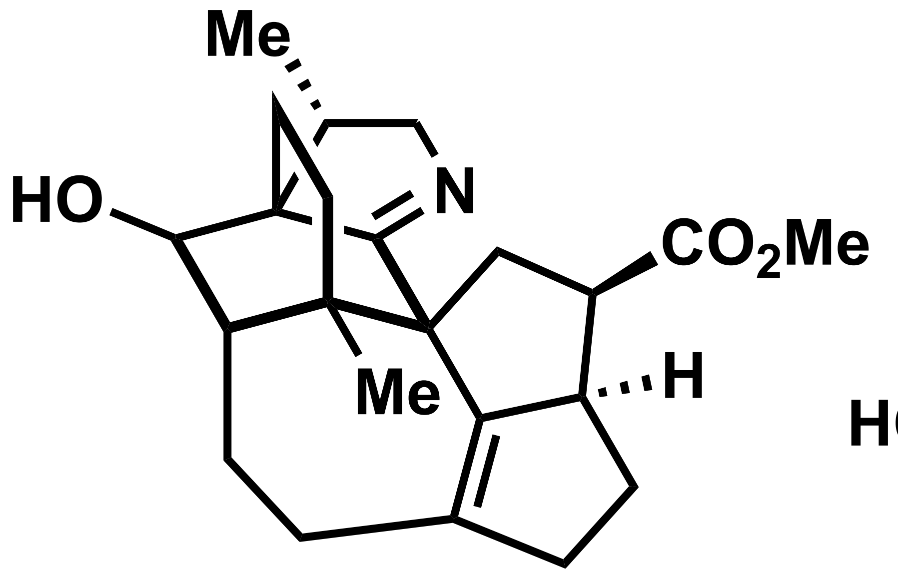
"Pretty" Molecules



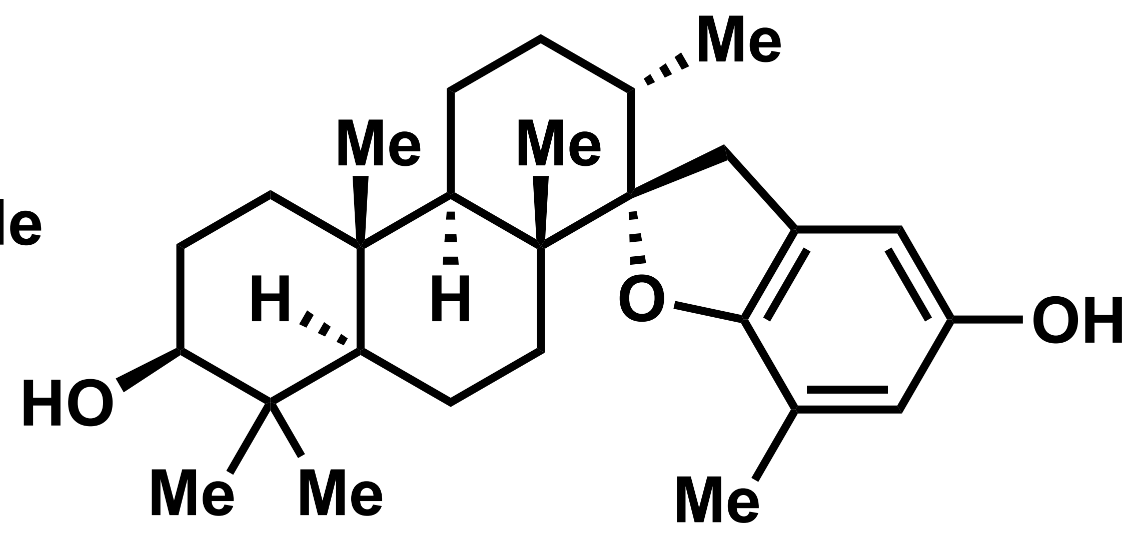
manzacidin C



maoecrystal V



(–)-calyciphylline N



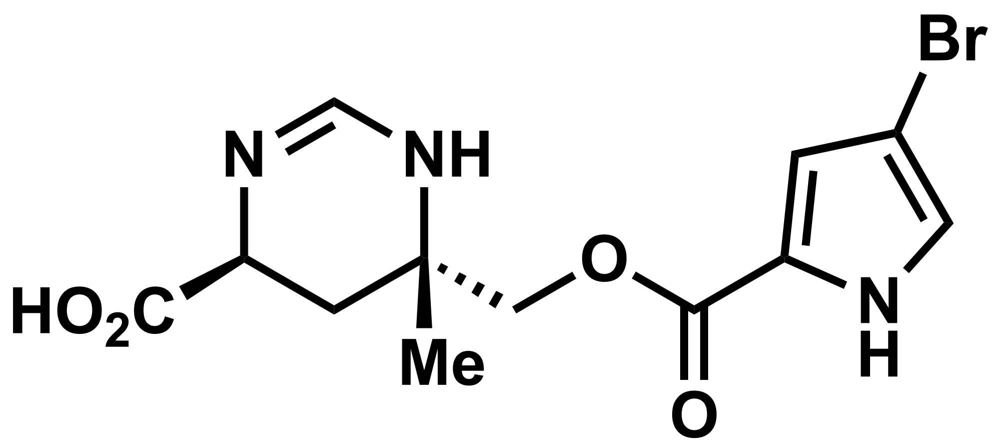
stypodiol

What Do Synthetic Chemists Want?

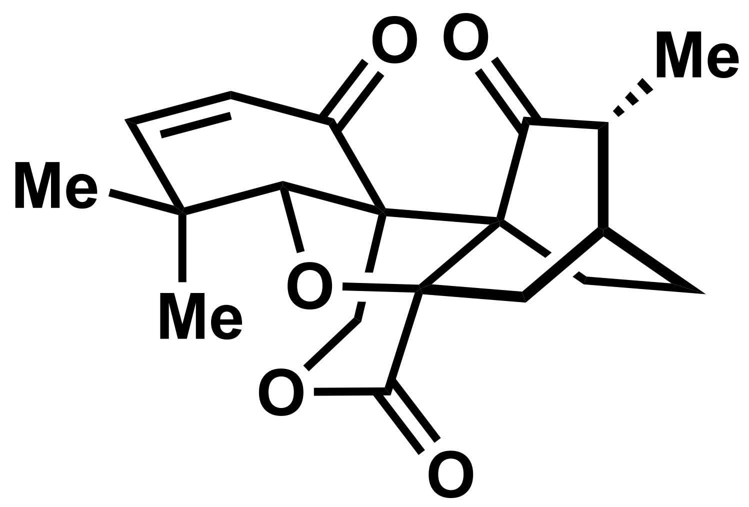
UNIVERSITY OF CAMBRIDGE

Make Molecules!

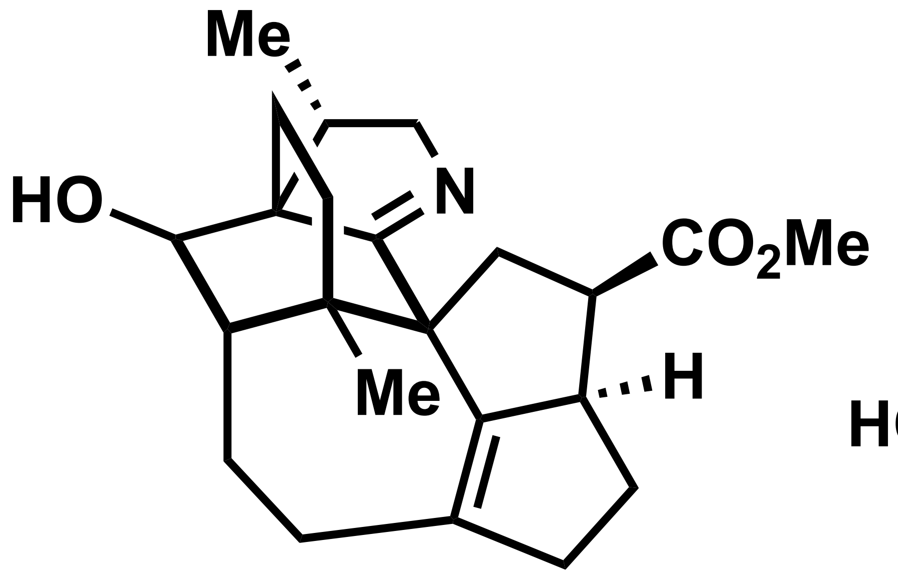
"Pretty" Molecules



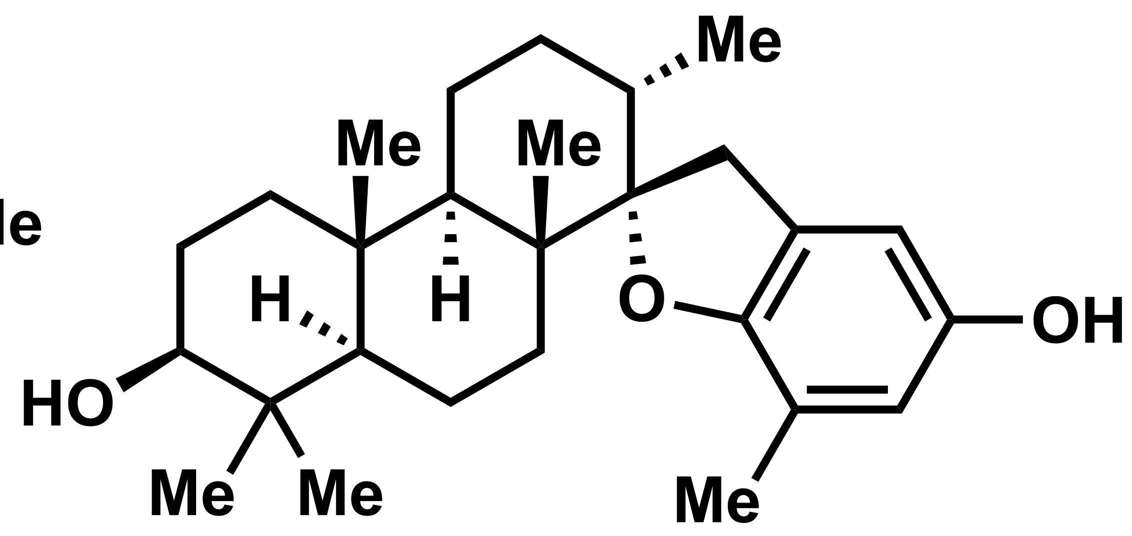
manzacidin C



maoecrystal V

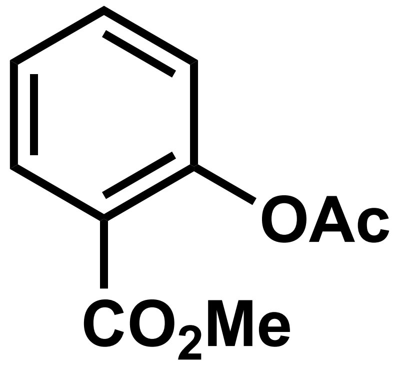


(–)-calyciphylline N

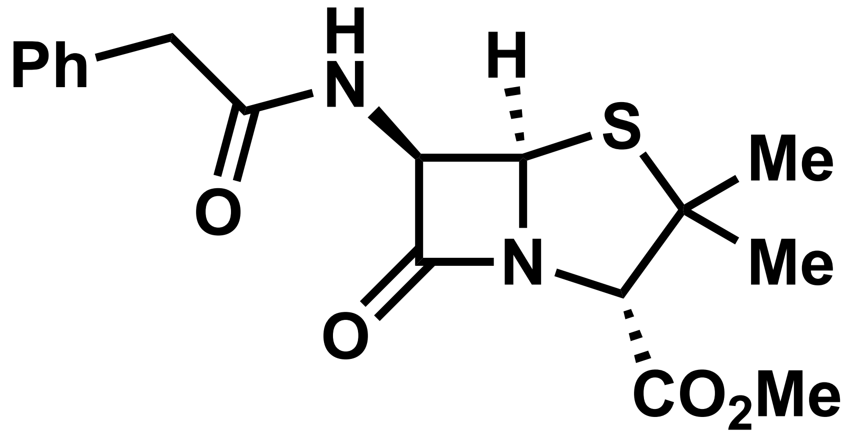


stypodiol

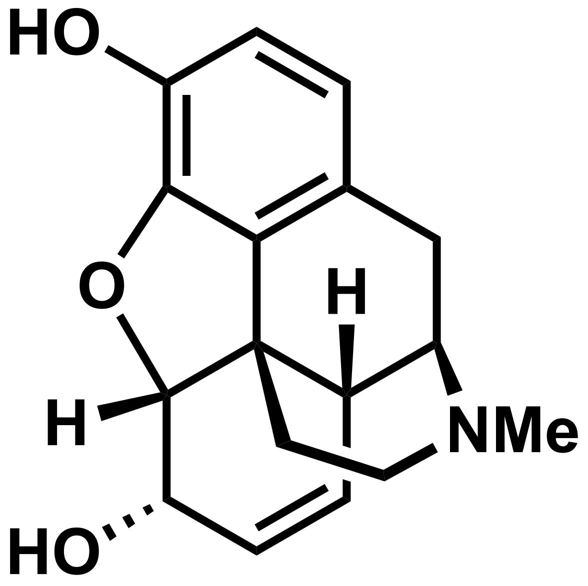
Pharmaceuticals



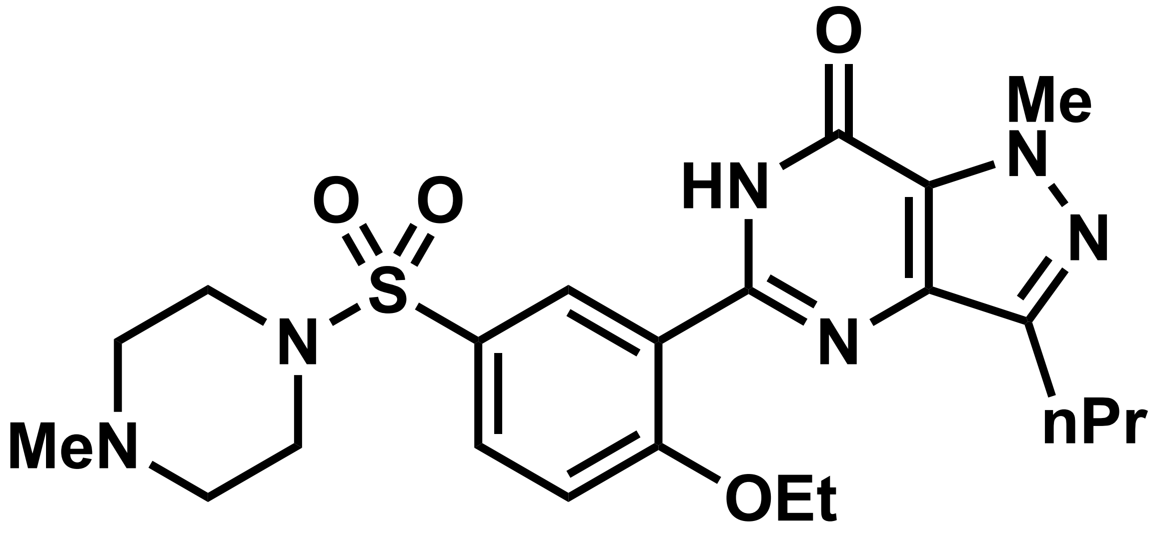
aspirin



penicillin G



morphine

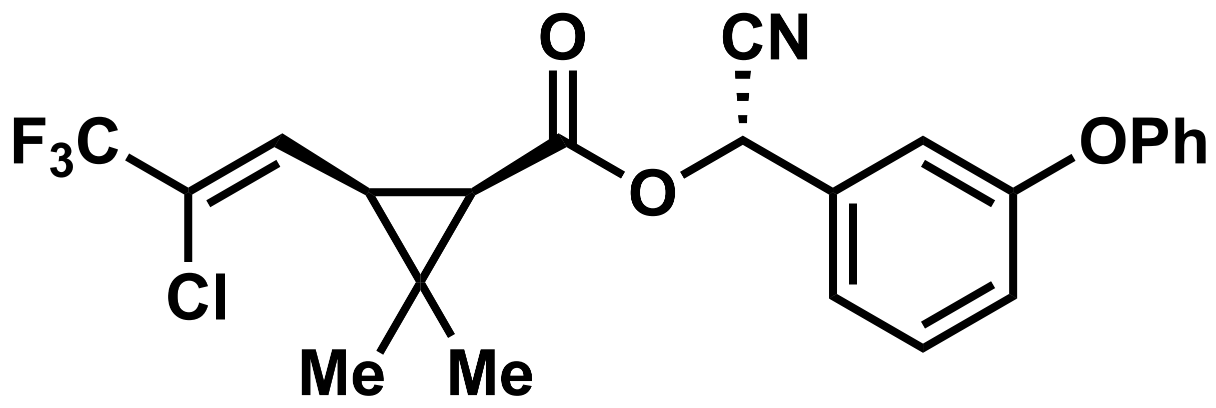


sildenafil

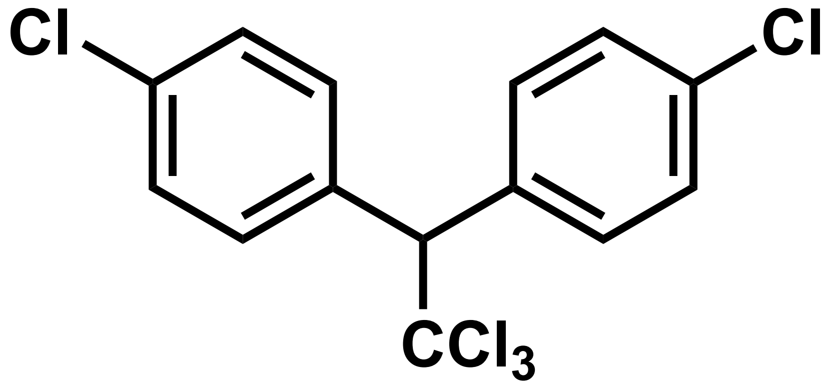
What Do Synthetic Chemists Want?

Make Molecules!

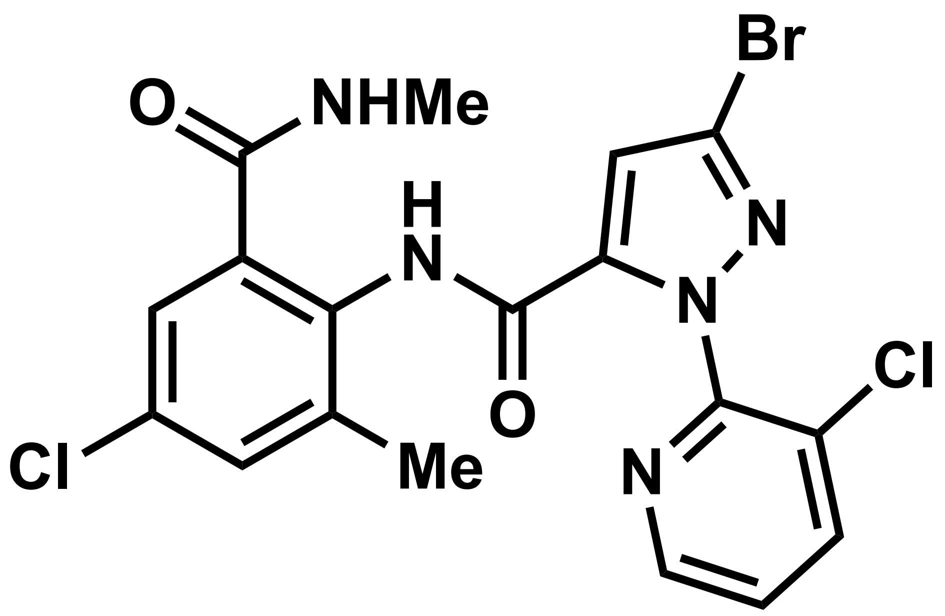
Agrochemicals



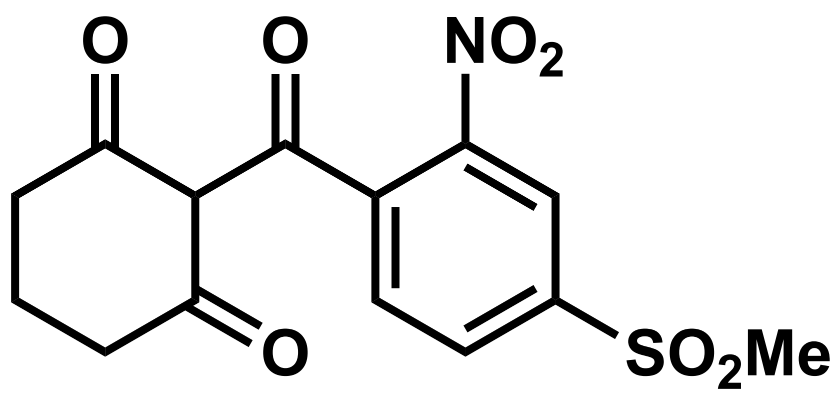
lambda-cyhalothrin



DTT



chlorantraniliprole

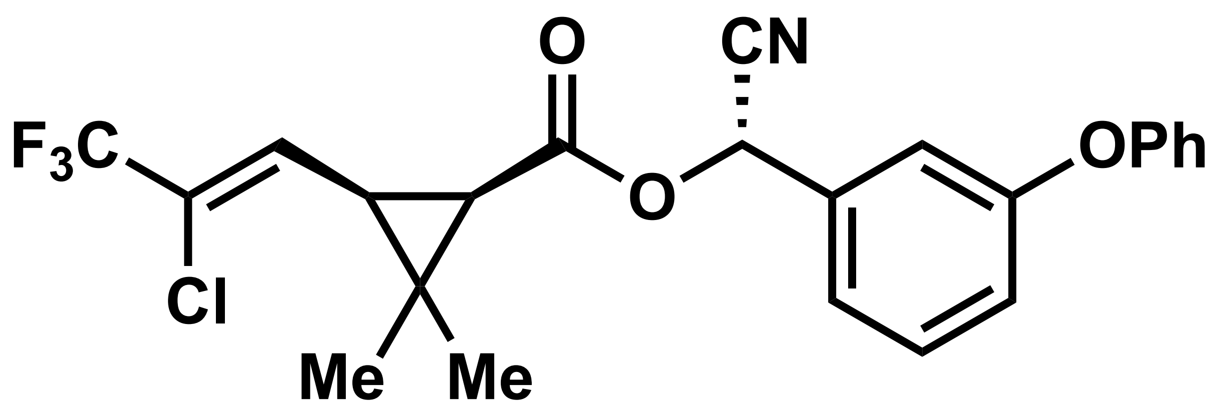


mesotrione

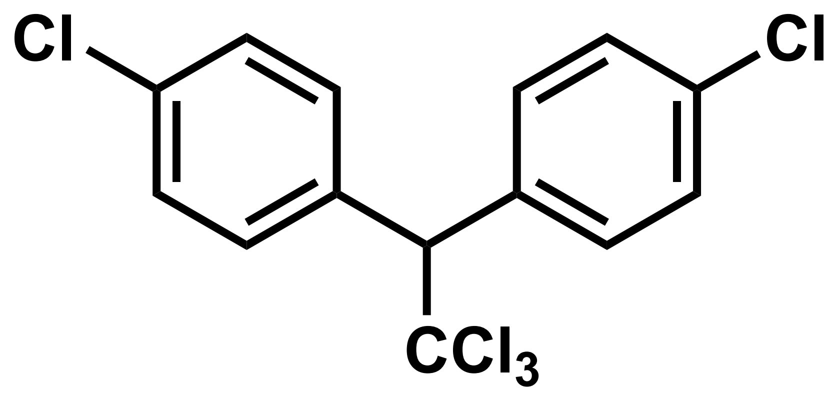
What Do Synthetic Chemists Want?

Make Molecules!

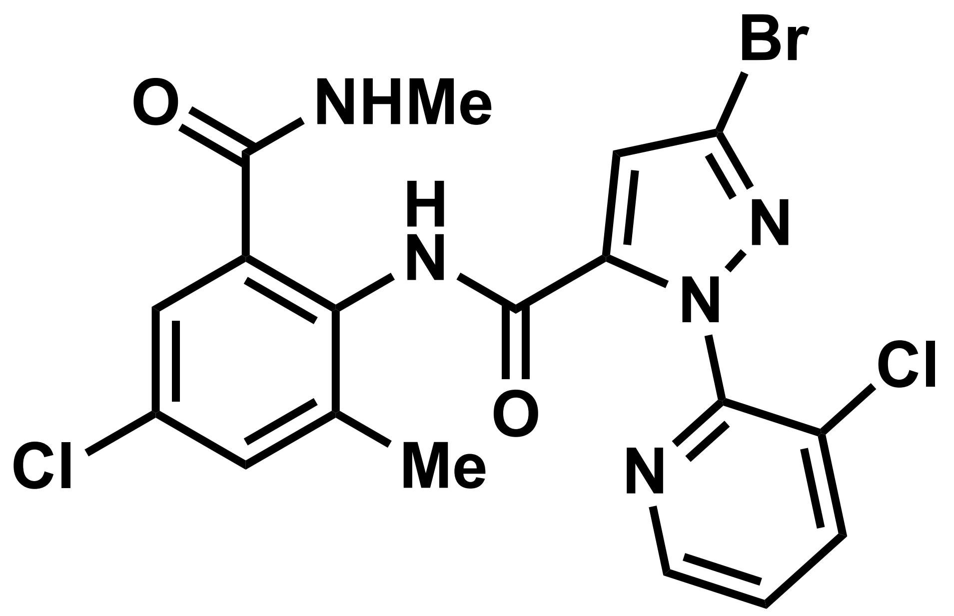
Agrochemicals



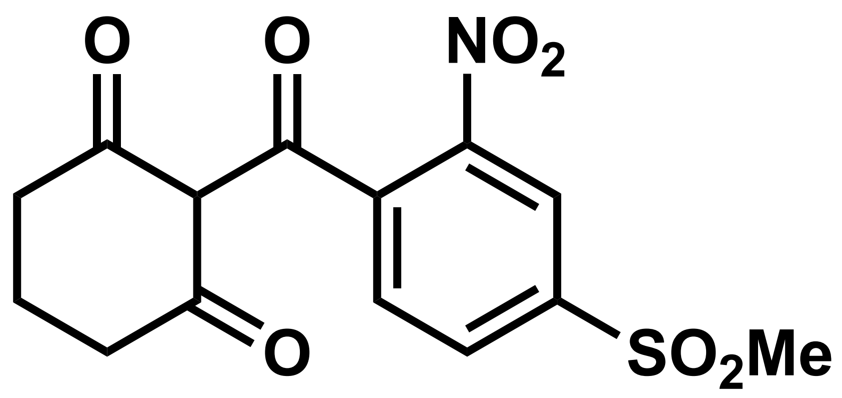
lambda-cyhalothrin



DTT

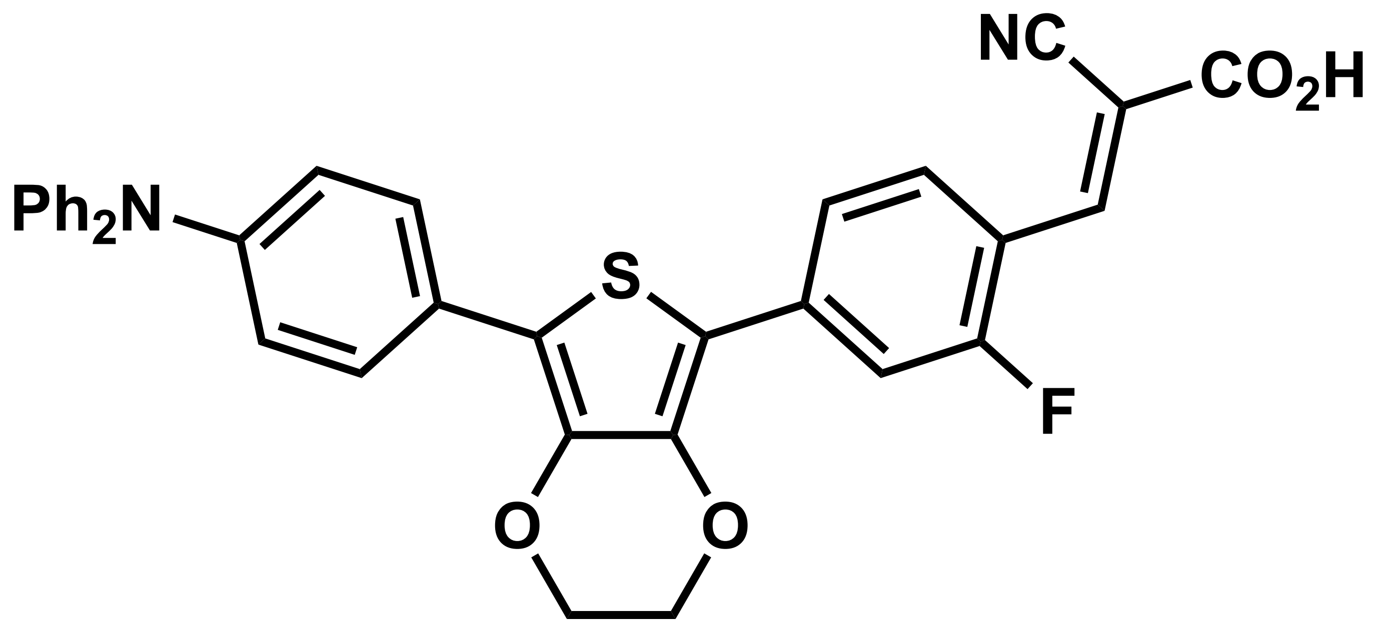


chlorantraniliprole

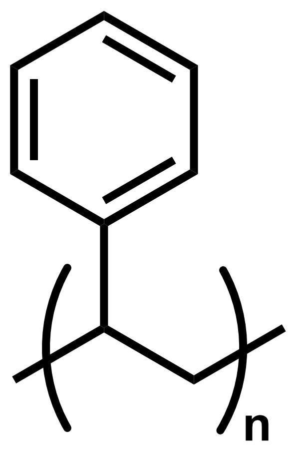


mesotrione

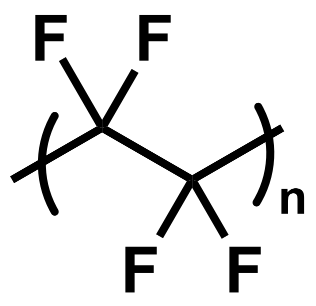
Organic Materials



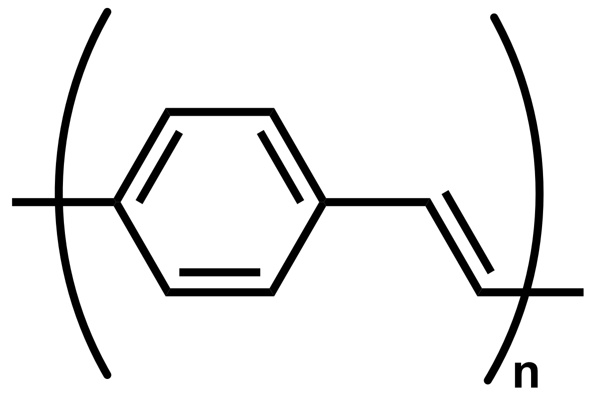
LBJ-F_o



polystyrene



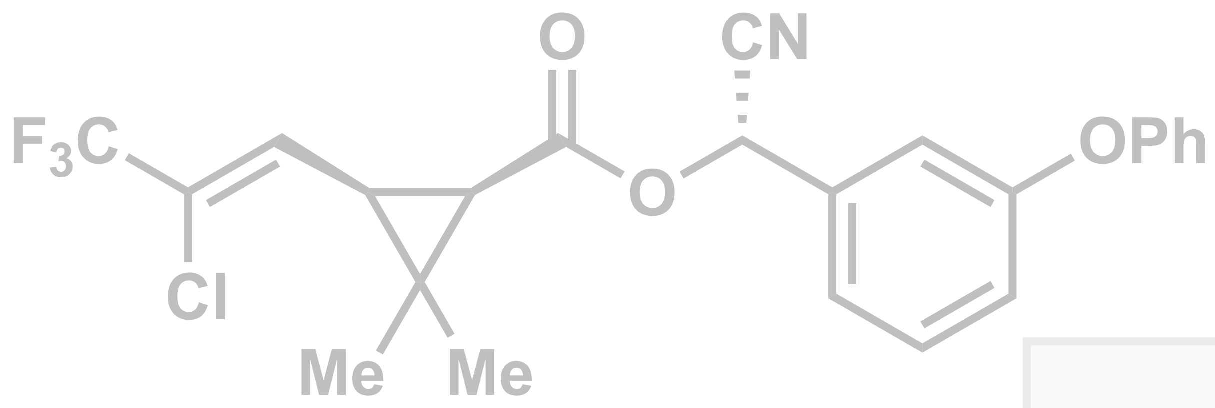
Teflon



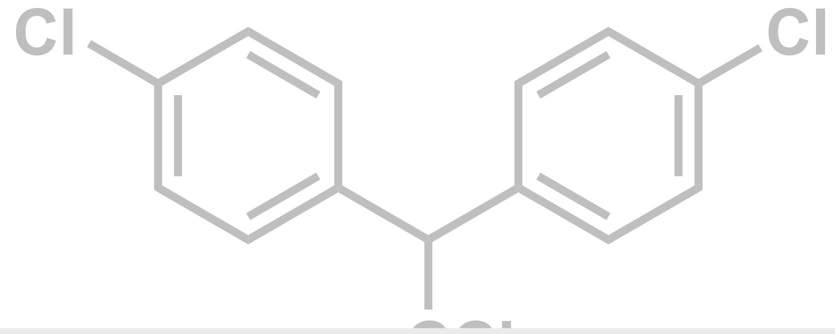
poly(*p*-phenylenevinylene)

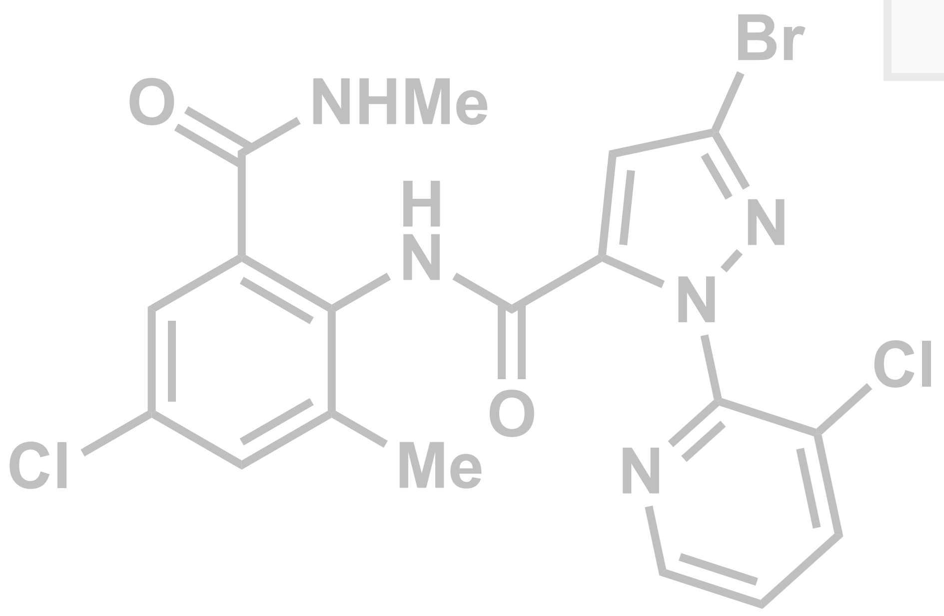
Make Molecules!

Agrochemicals

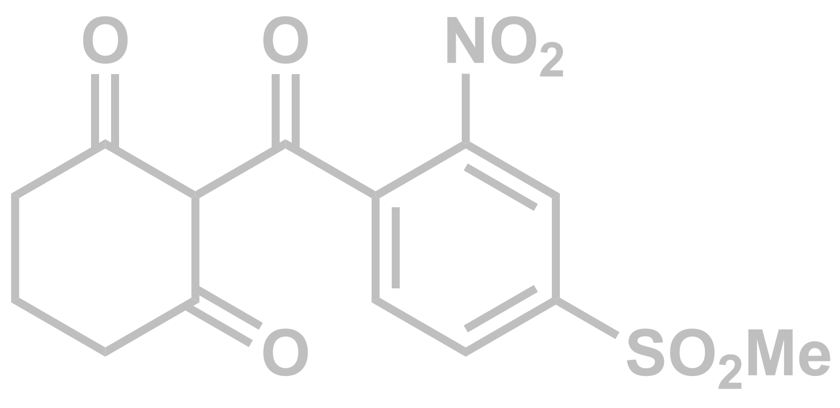


lambda-cyhalothrin



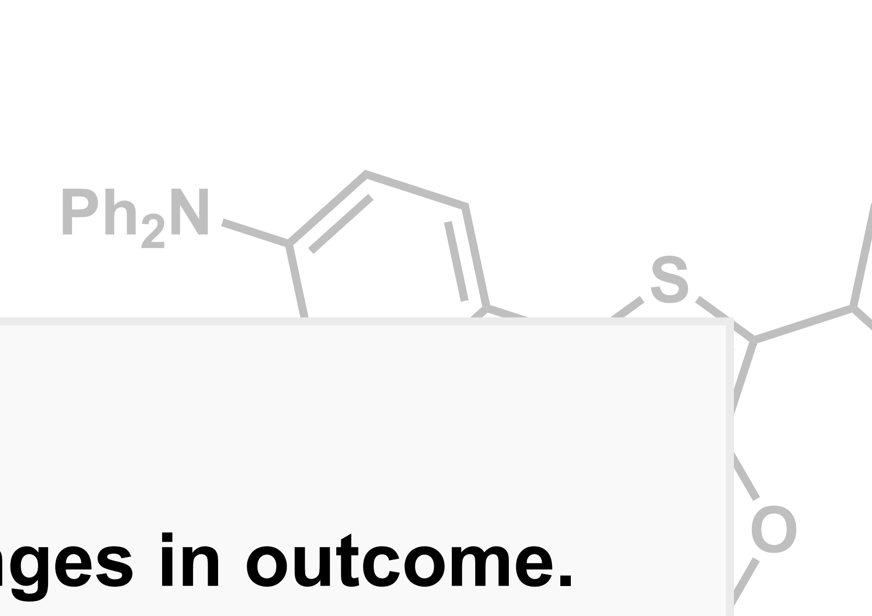


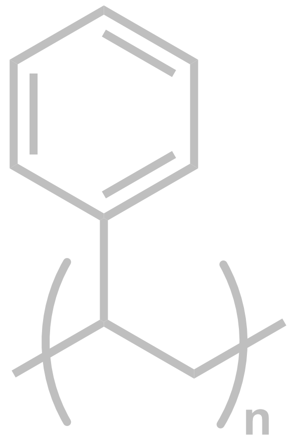
chlorantraniliprole



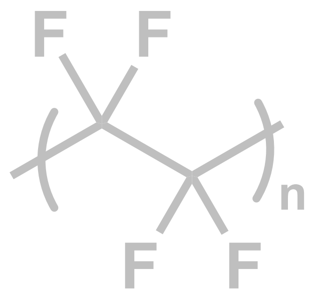
mesotrione

Organic Materials

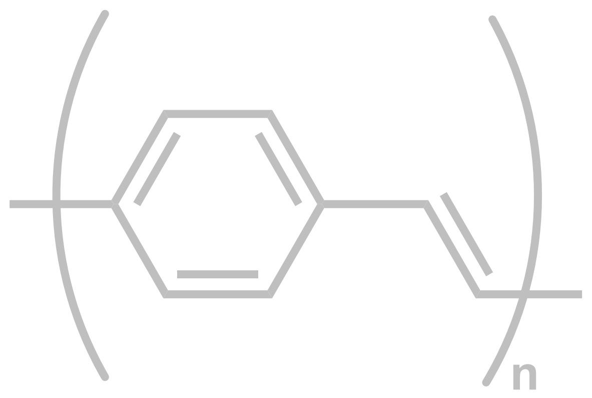




polystyrene



Teflon



poly(*p*-phenylenevinylene)

Small changes in structure ≠ Small changes in outcome.

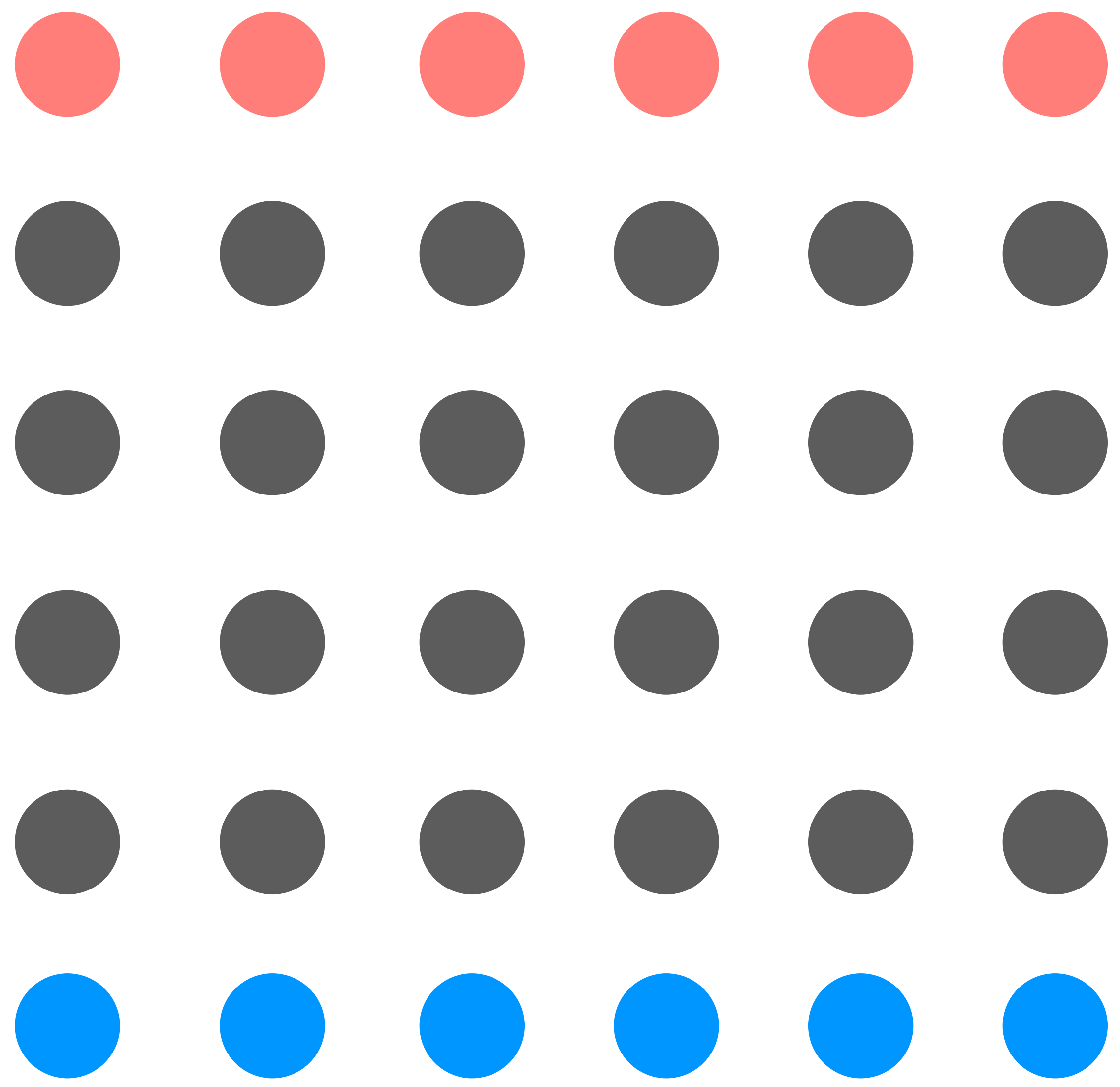
UNIVERSITY OF CAMBRIDGE

Part I:

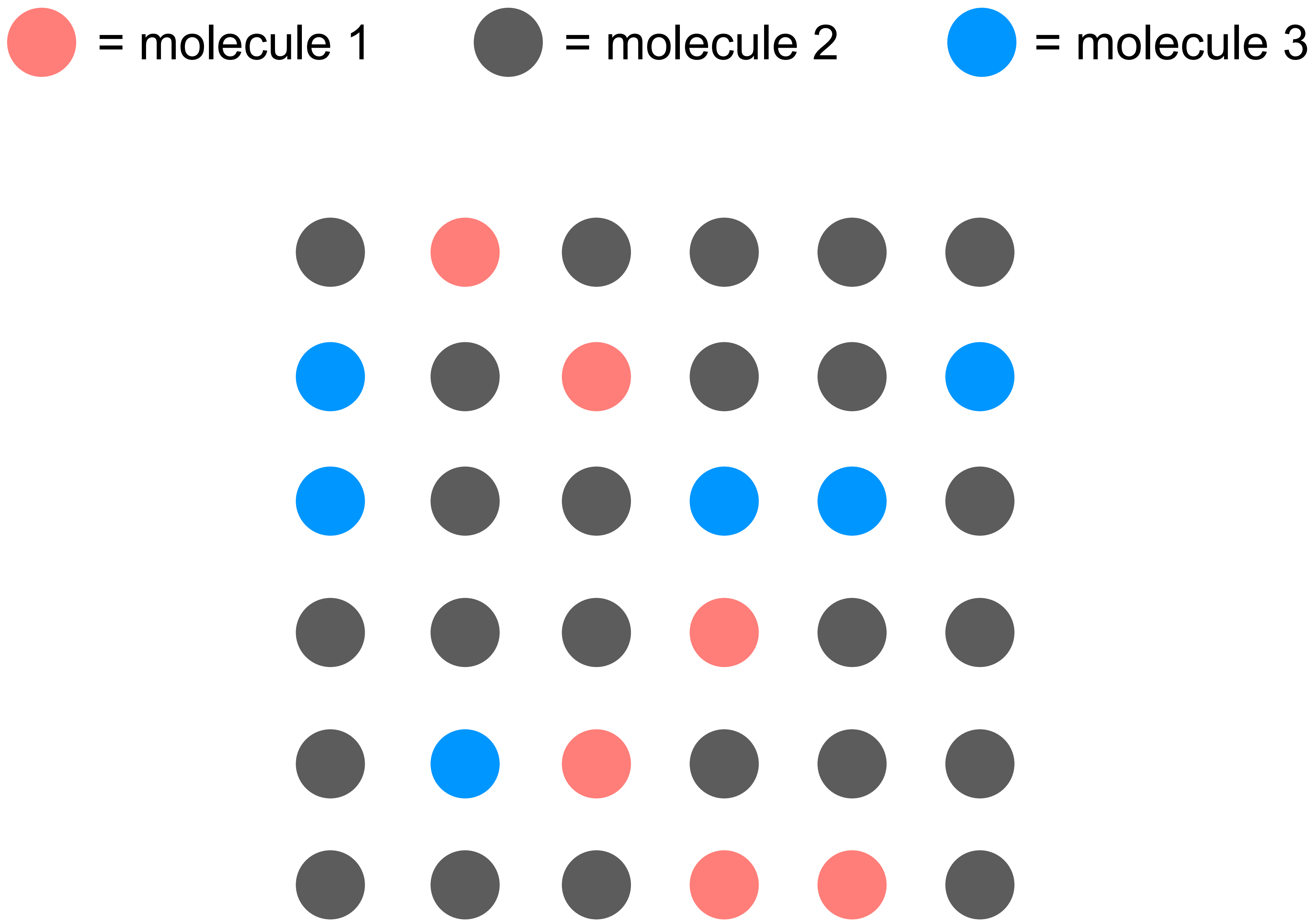
Transfer Learning to Unlock Chemical Predictions in Low Data Regimes

The Importance of Test Sets

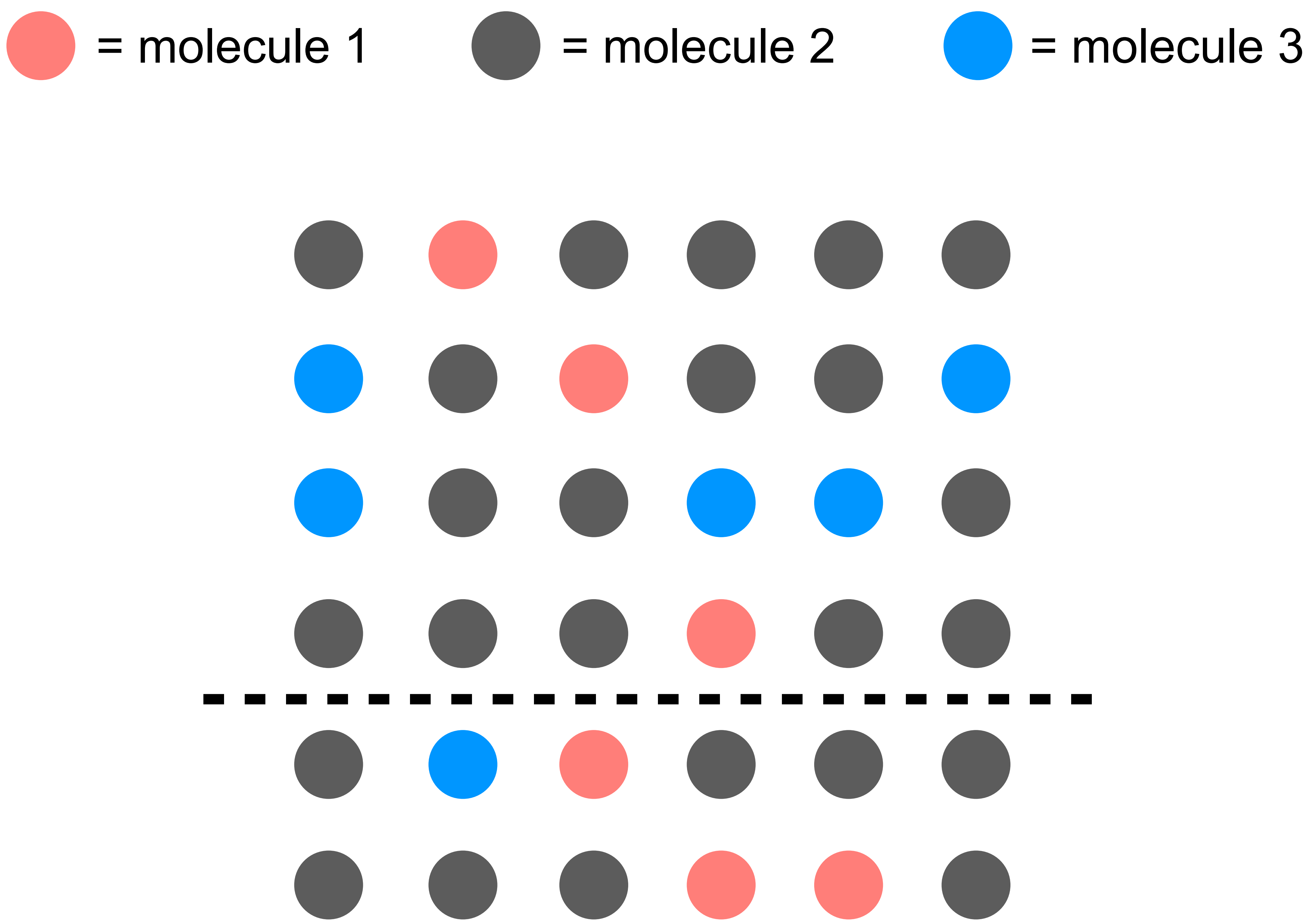
= molecule 1 = molecule 2 = molecule 3



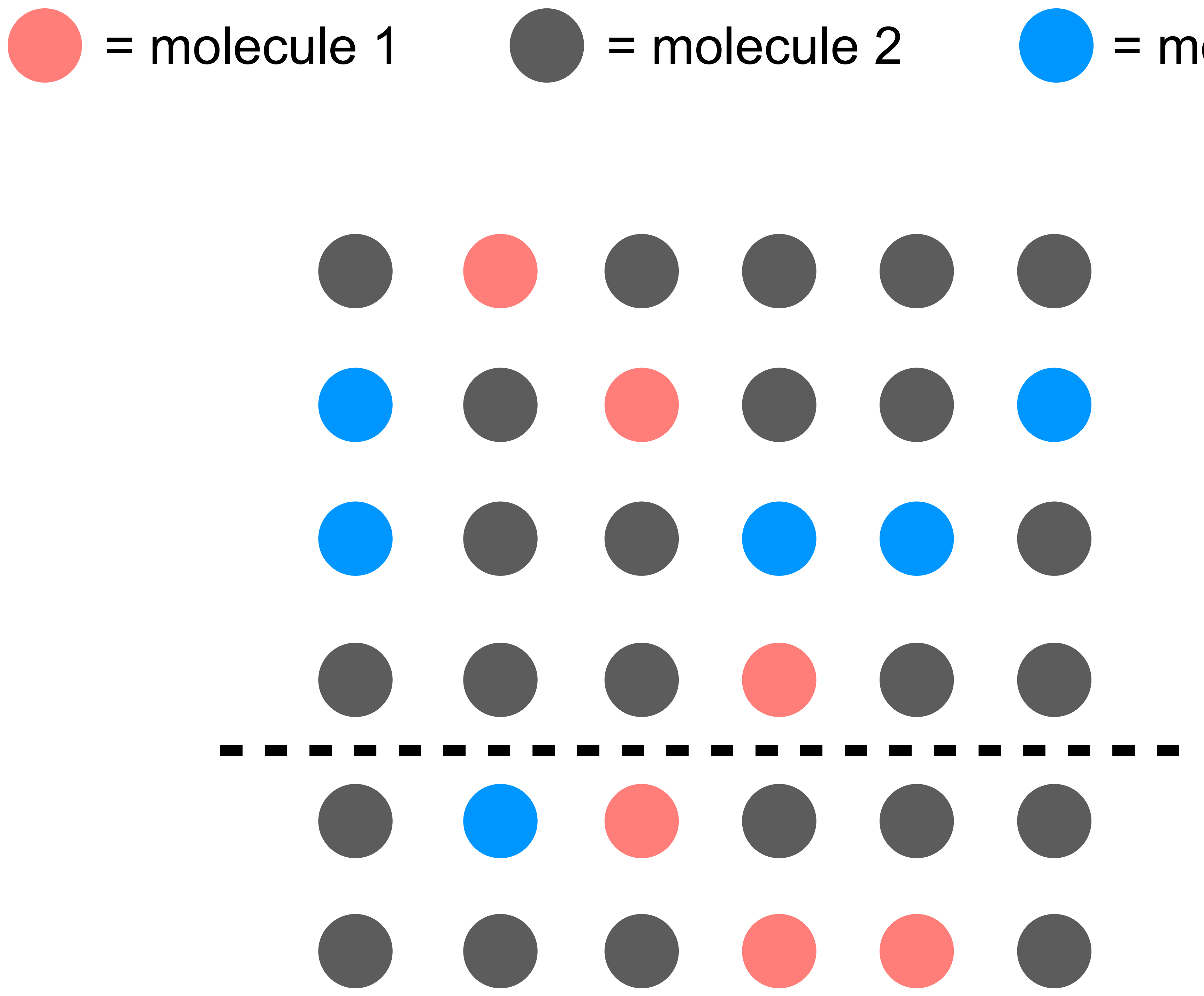
The Importance of Test Sets



The Importance of Test Sets



The Importance of Test Sets

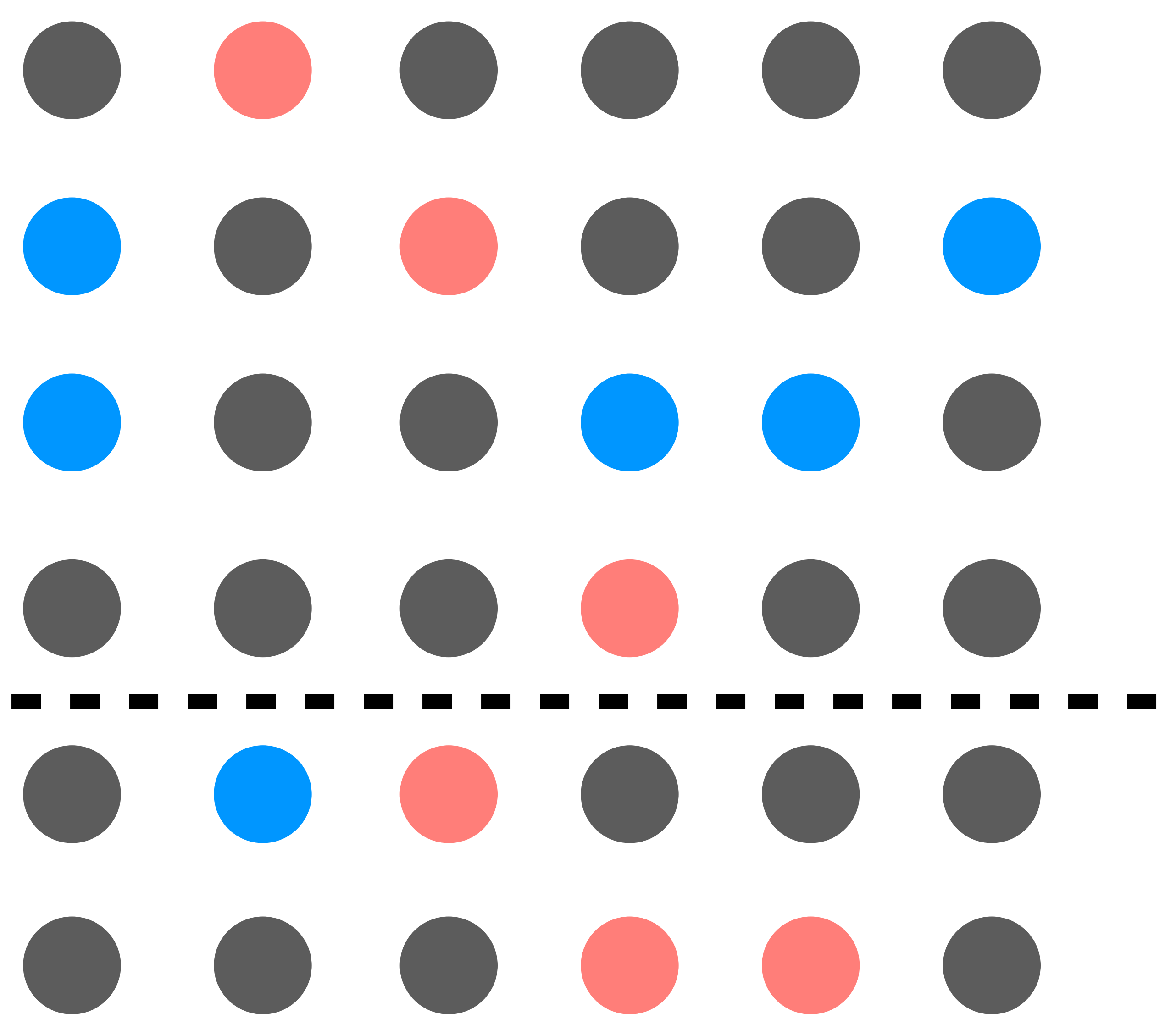


simple

The Importance of Test Sets

UNIVERSITY OF CAMBRIDGE

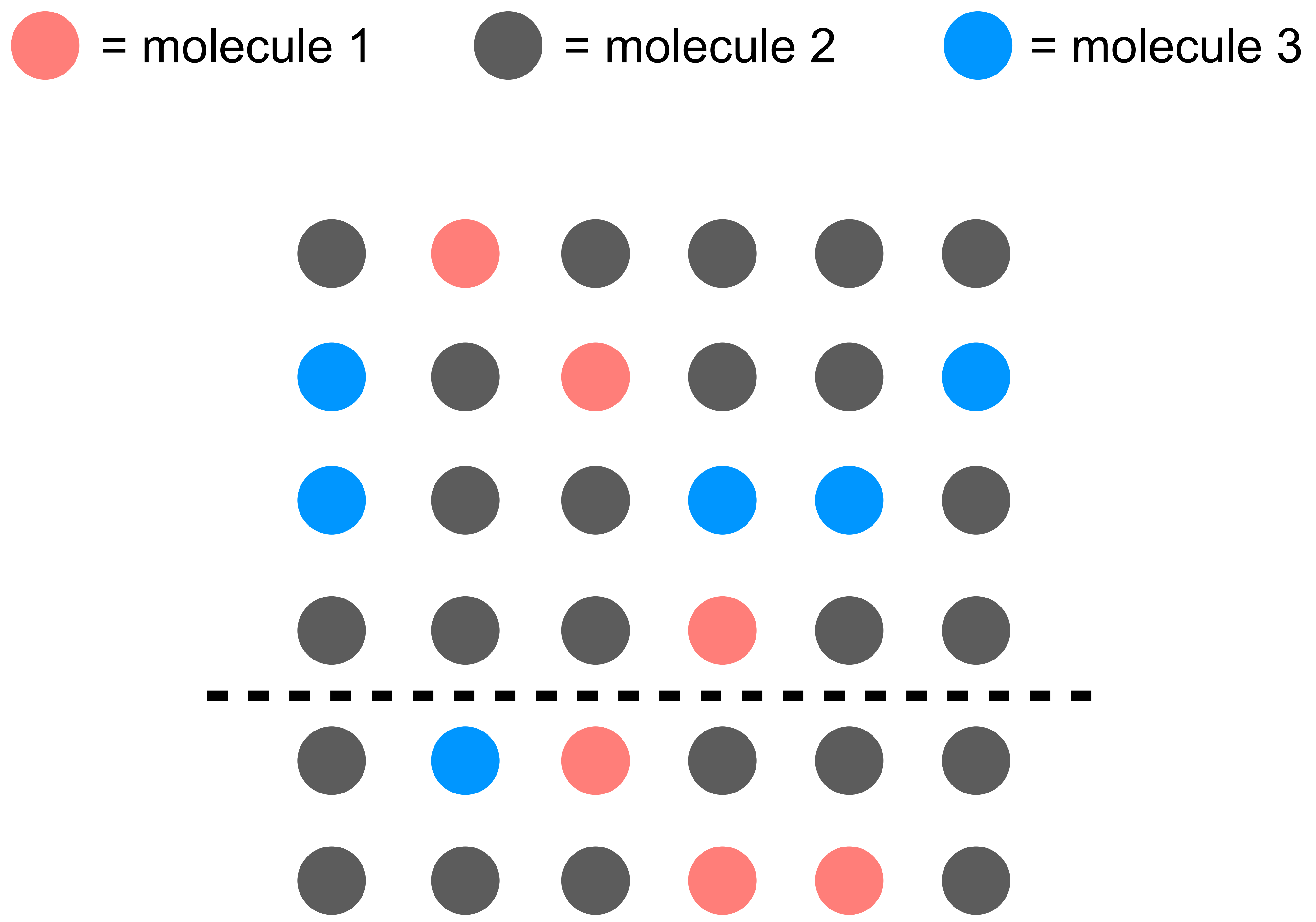
= molecule 1   = molecule 2   = molecule 3



simple

easy to implement

The Importance of Test Sets

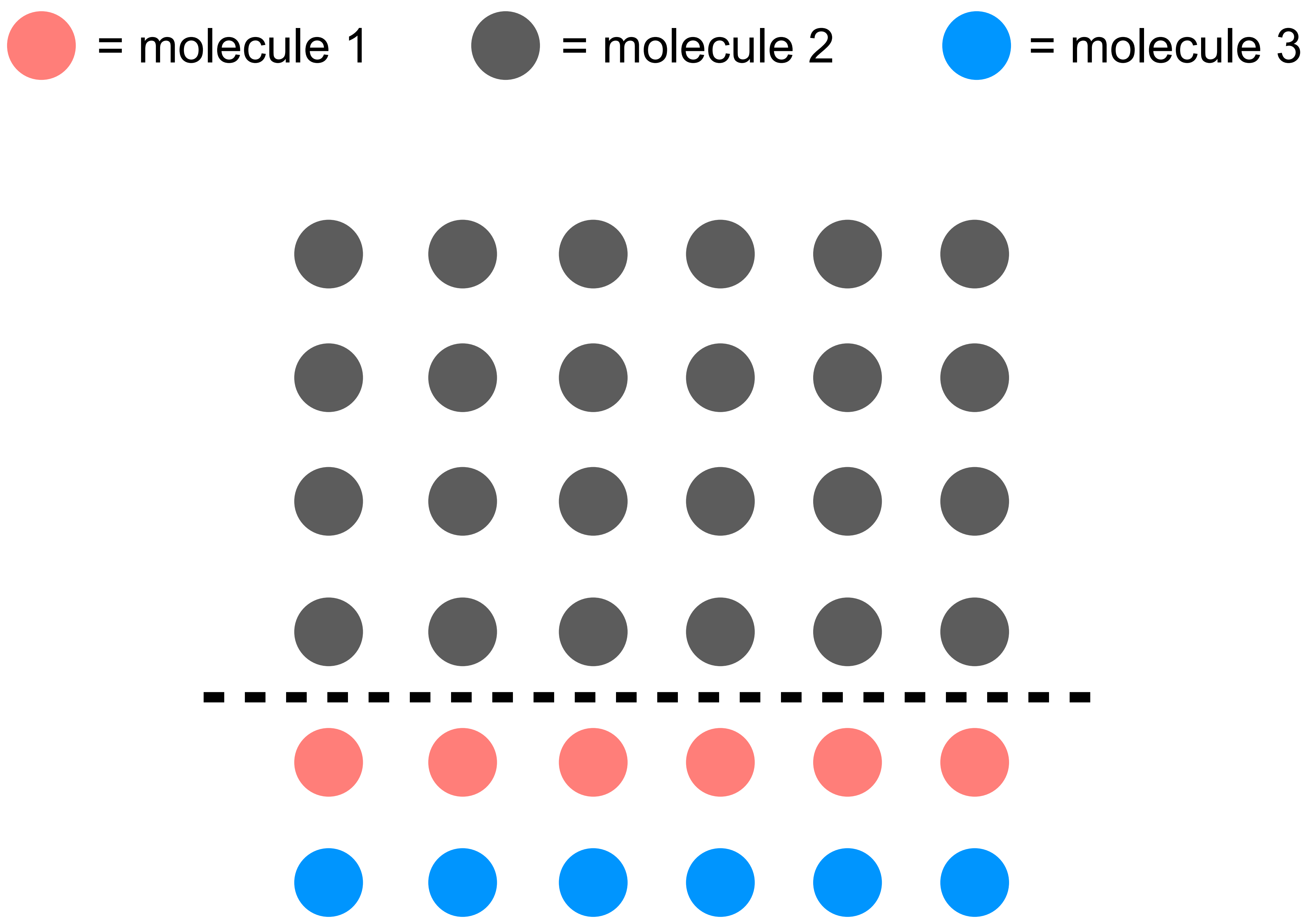


simple

easy to implement

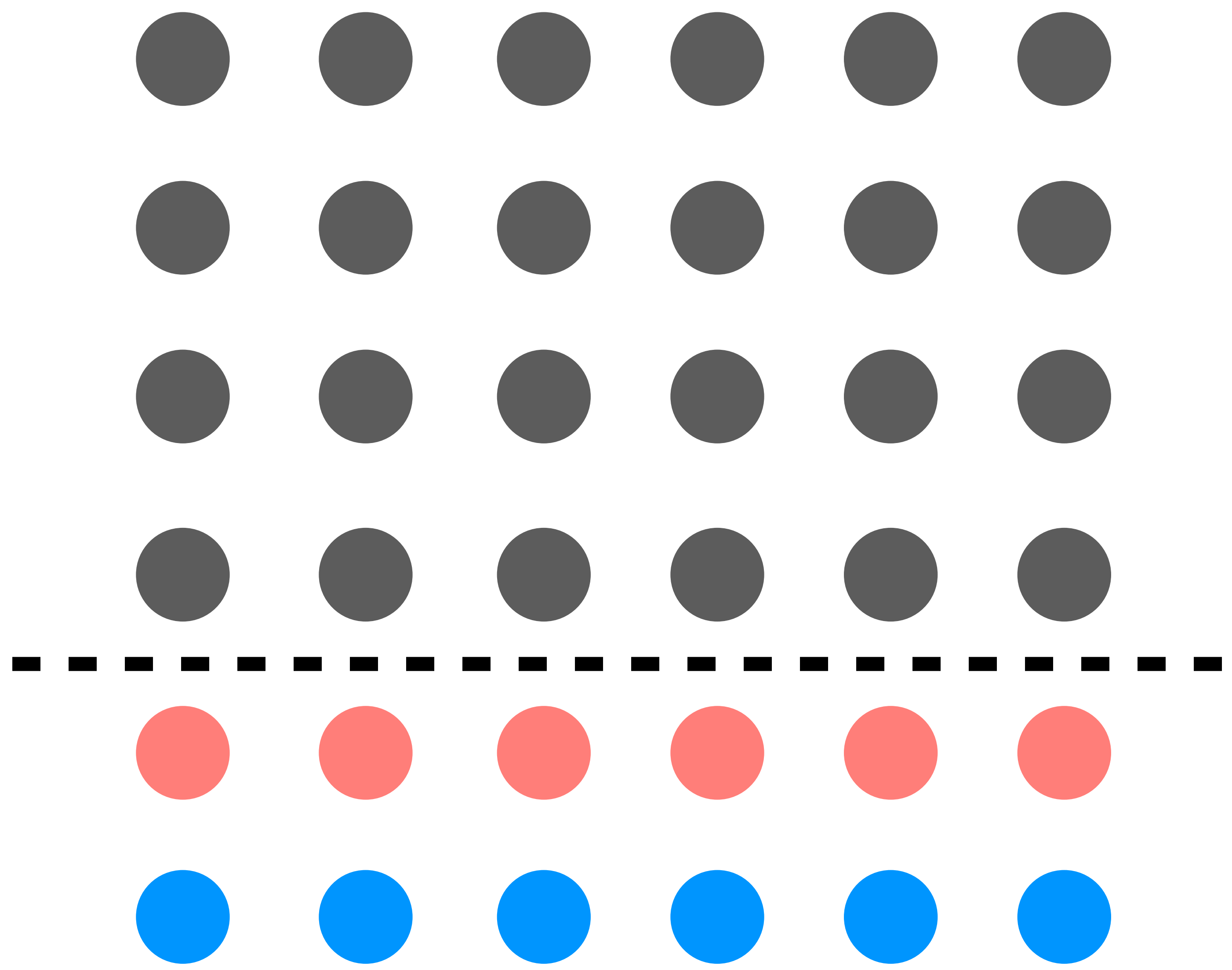
can be used with any dataset

The Importance of Test Sets



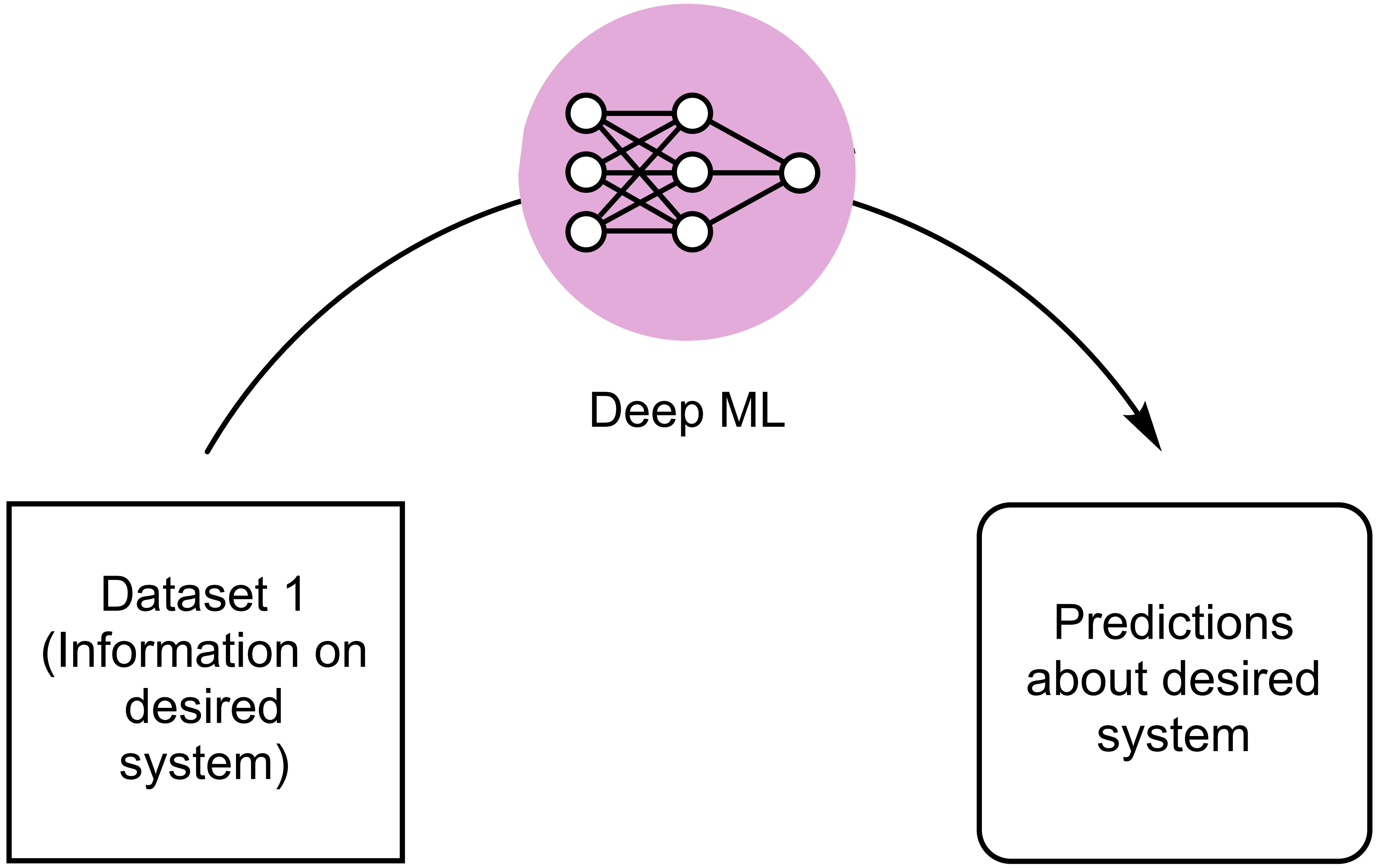
The Importance of Test Sets

= molecule 1 = molecule 2 = molecule 3

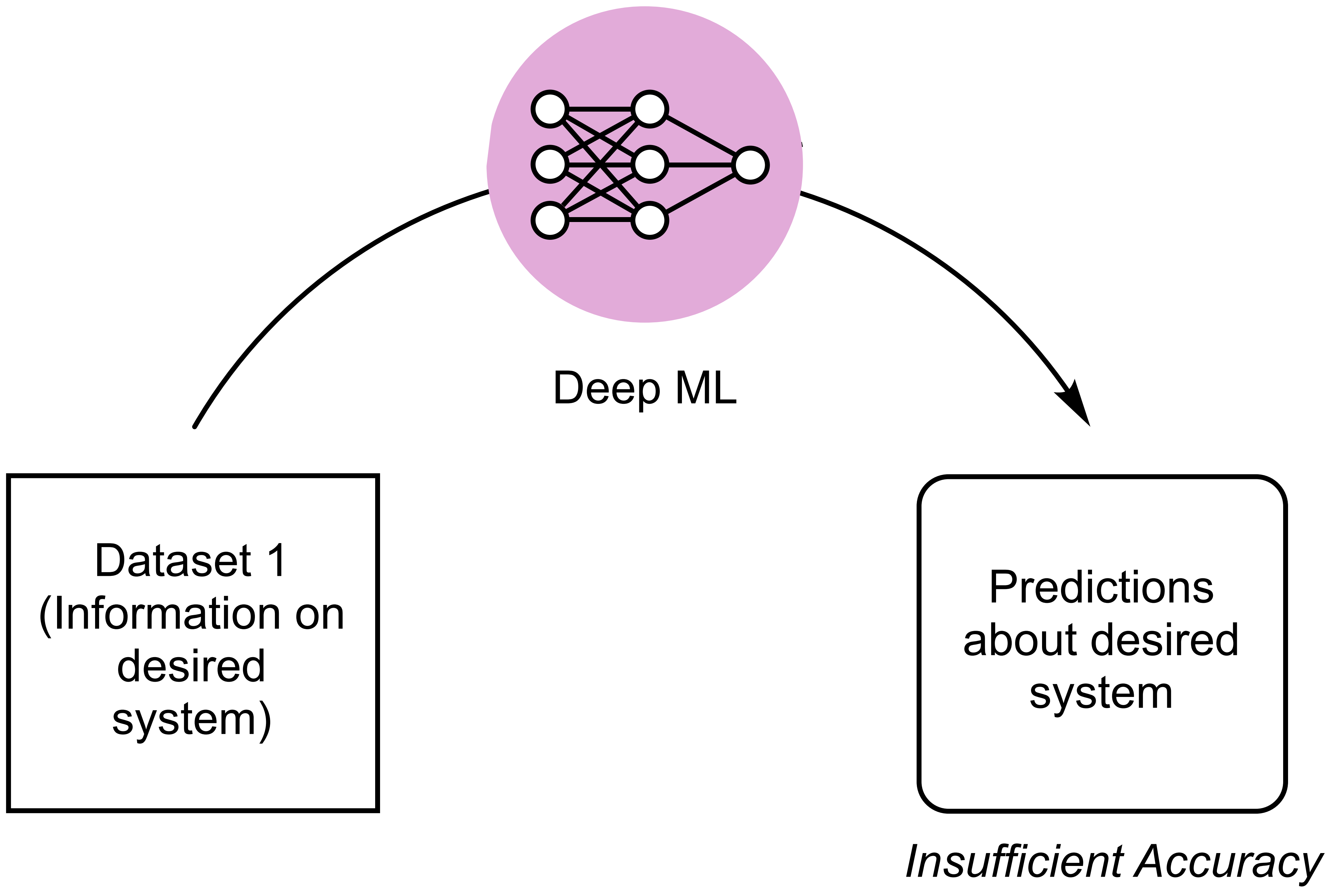


more challenging

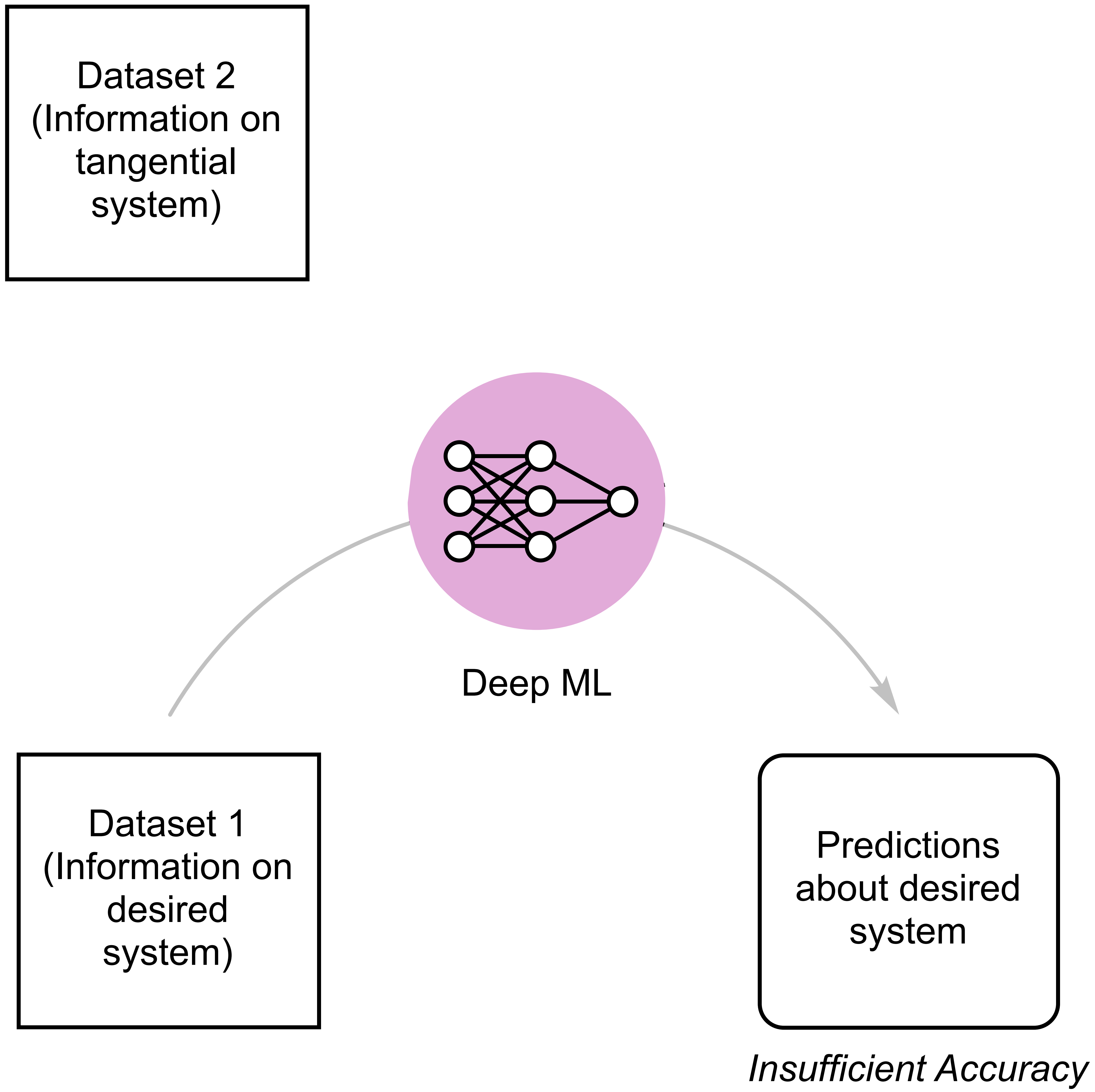
What Is Transfer Learning?



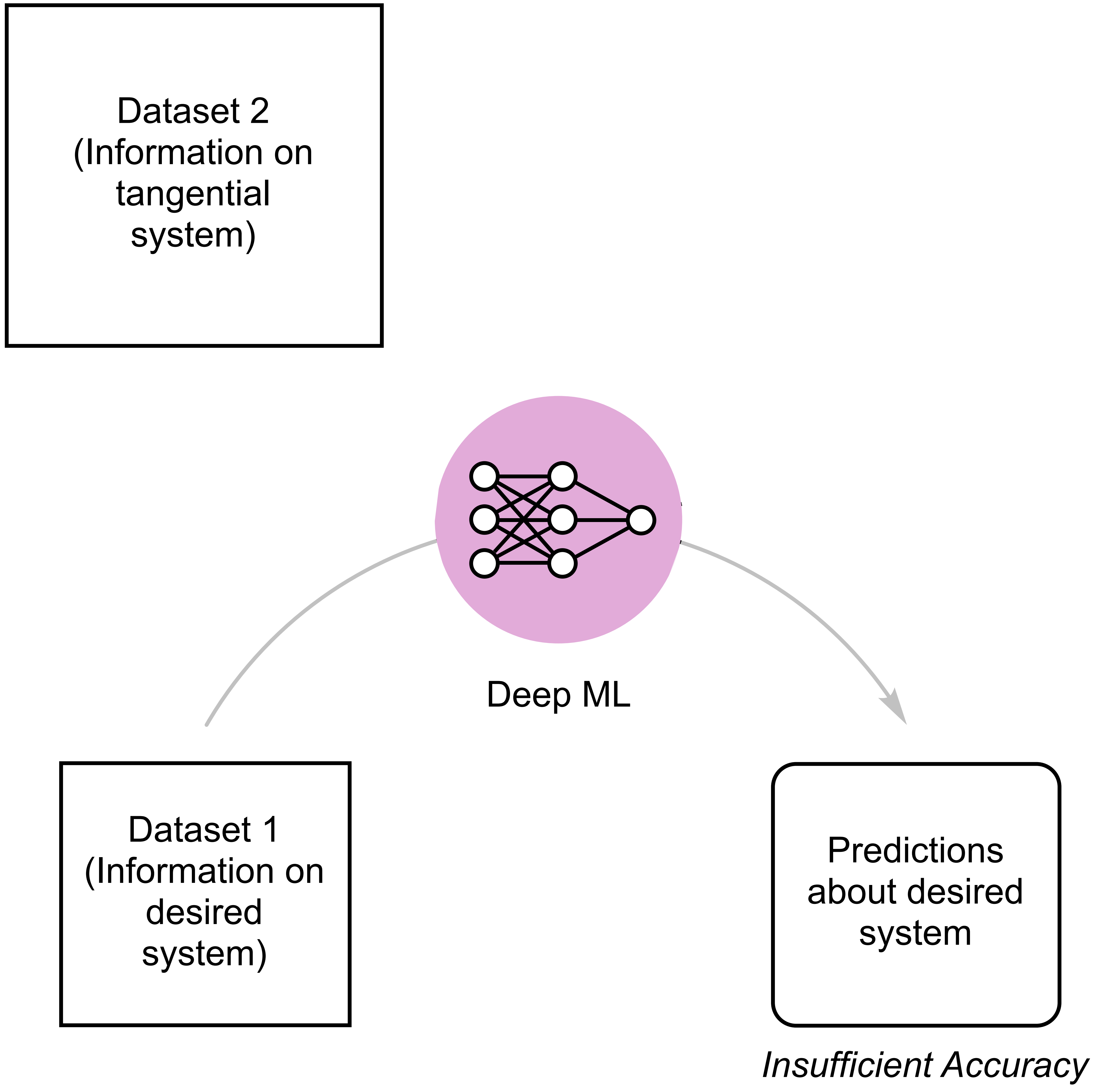
What Is Transfer Learning?



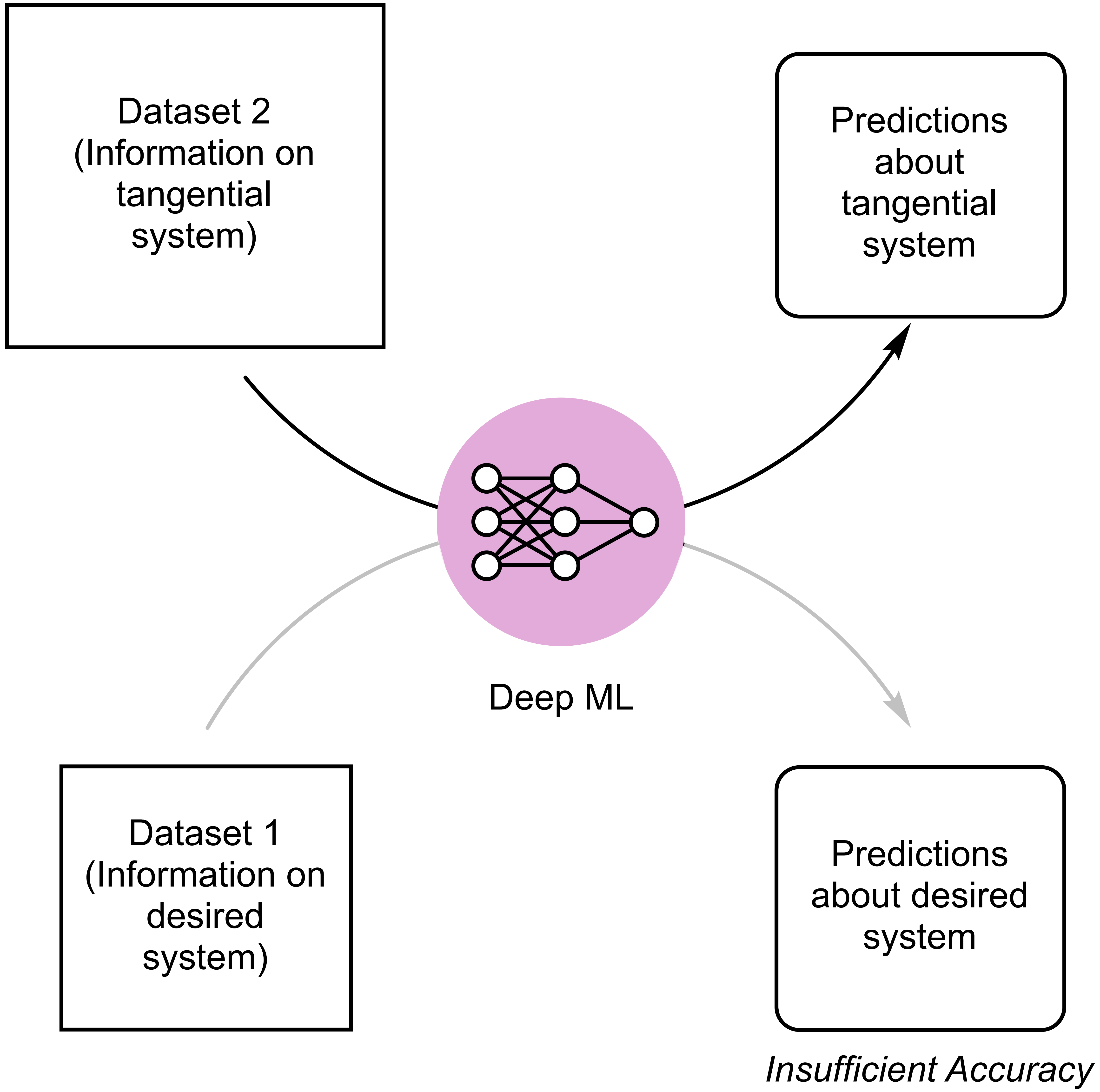
What Is Transfer Learning?



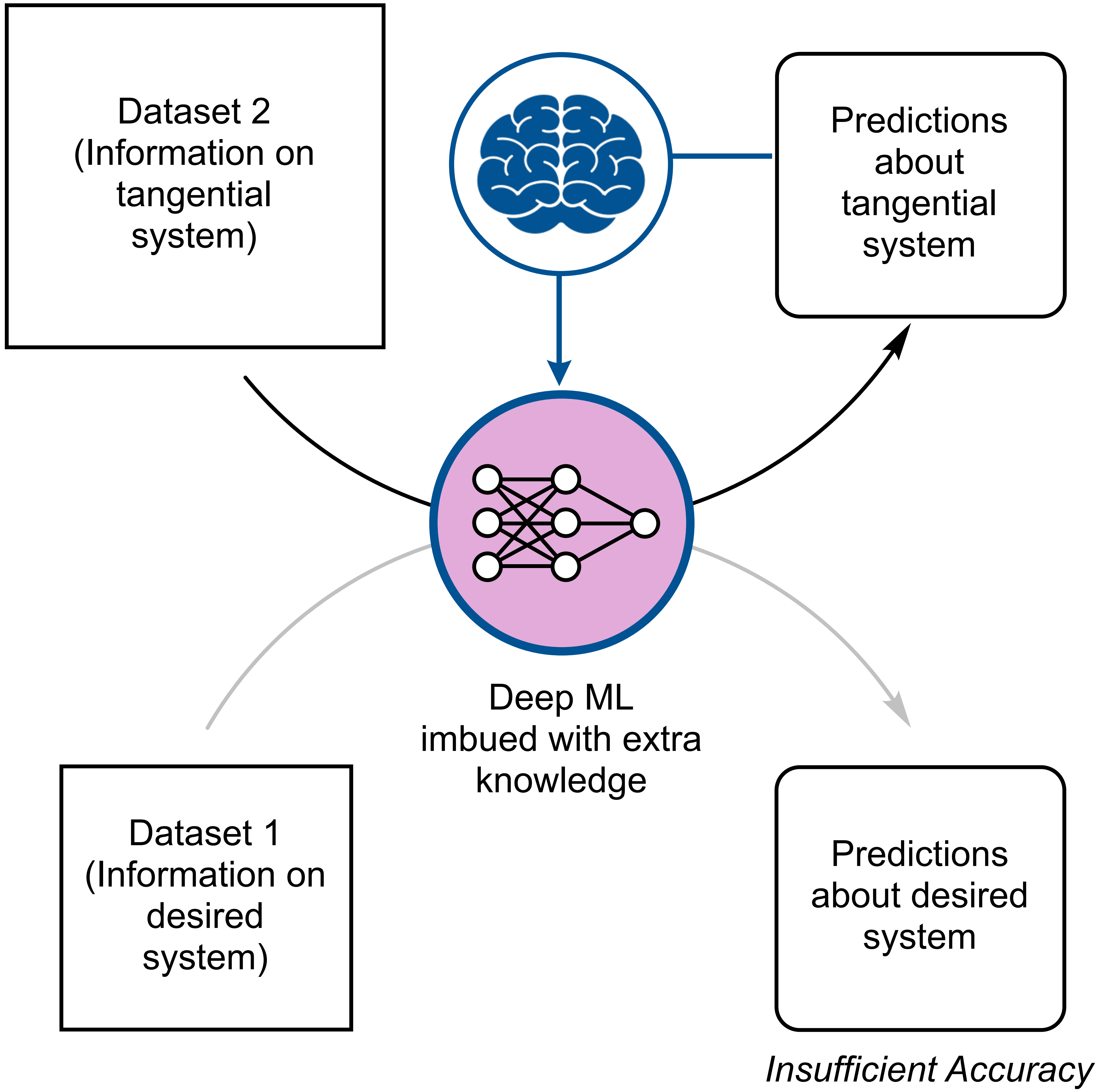
What Is Transfer Learning?



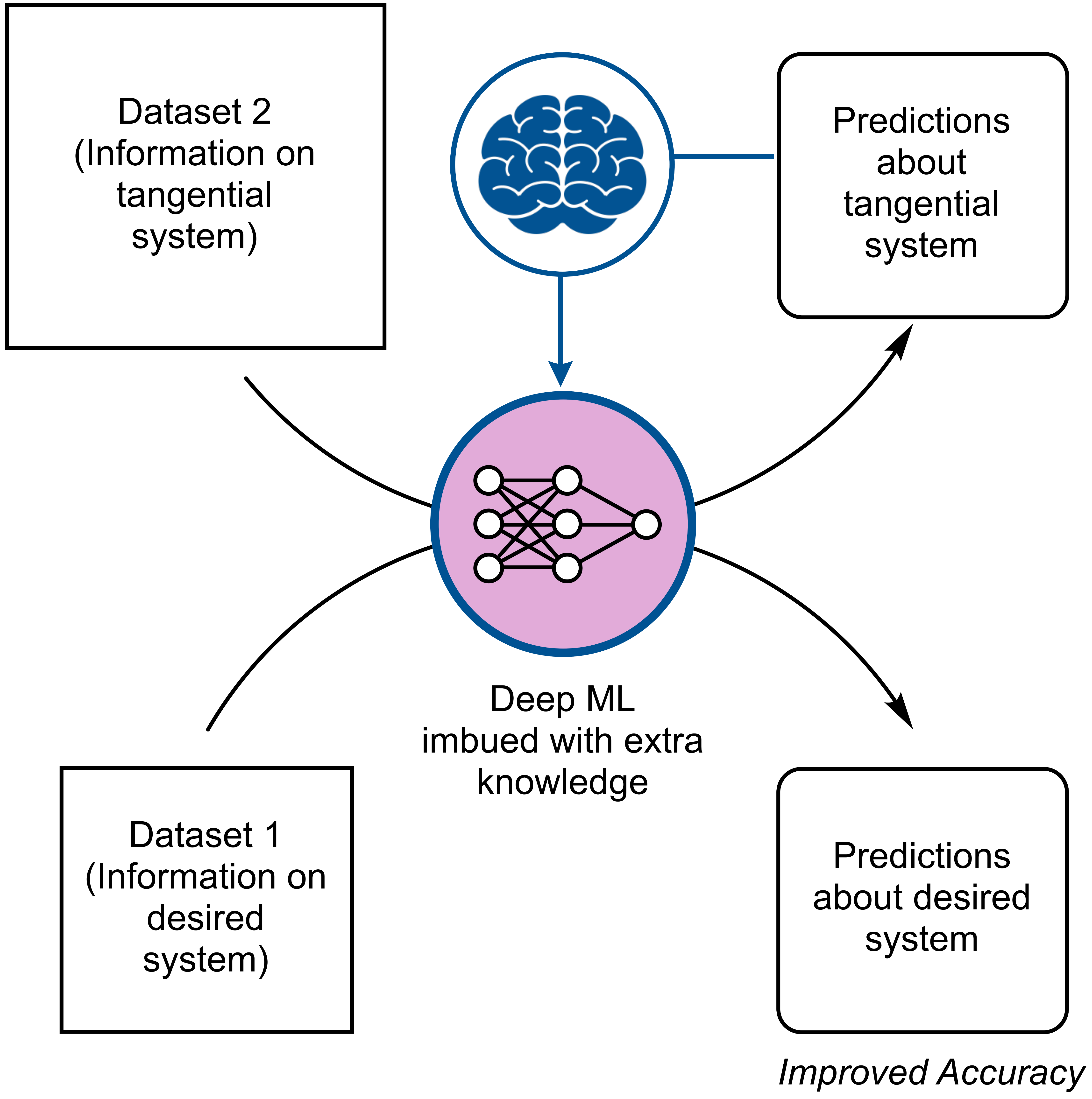
What Is Transfer Learning?

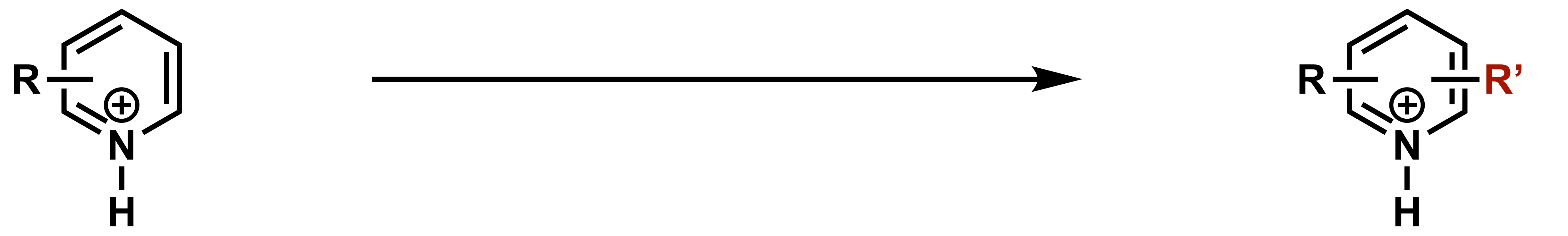


What Is Transfer Learning?



What Is Transfer Learning?

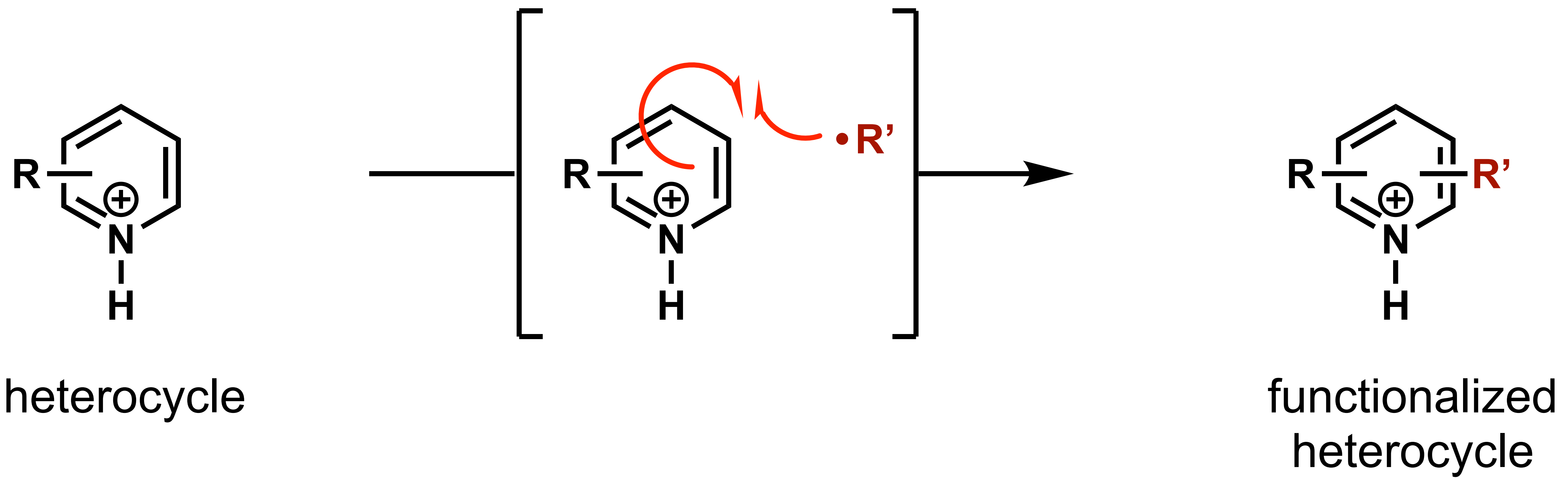




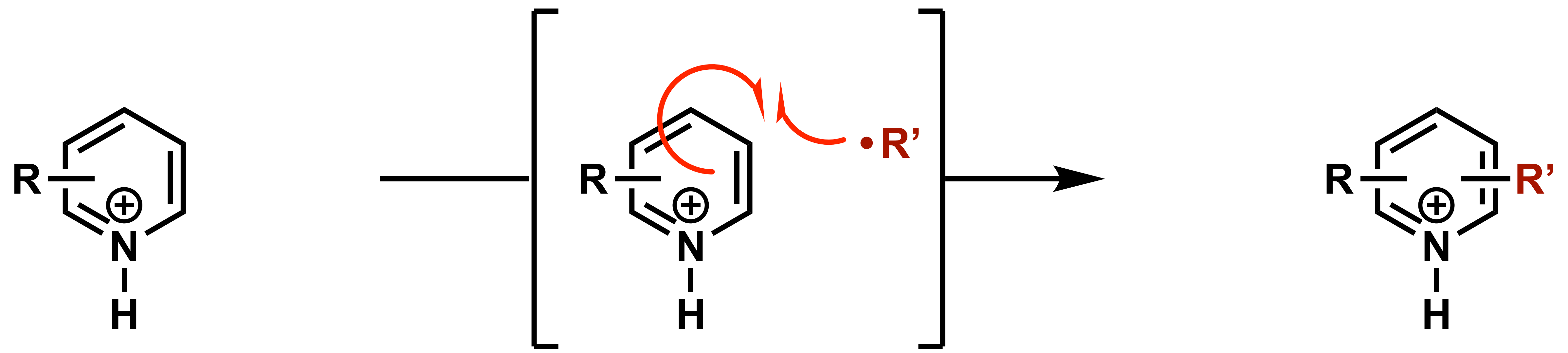
heterocycle

functionalized heterocycle

The Minisci Reaction

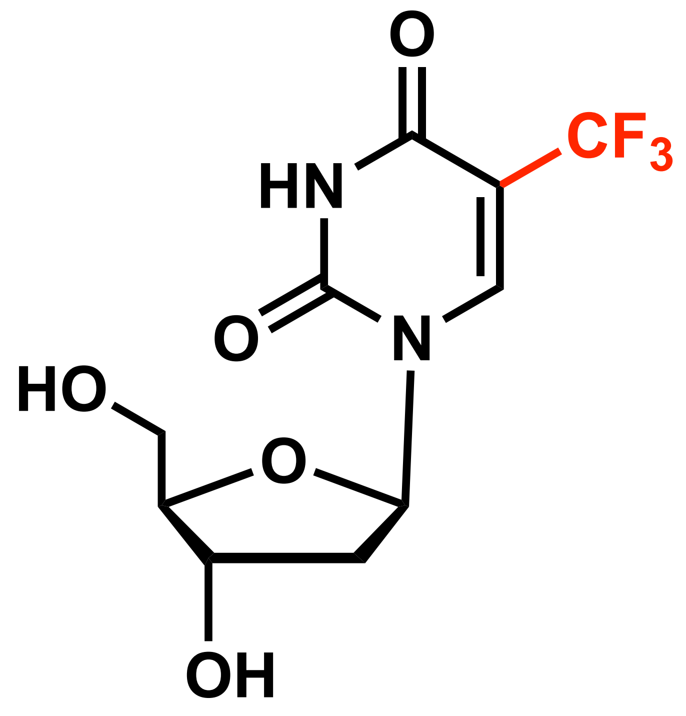


The Minisci Reaction



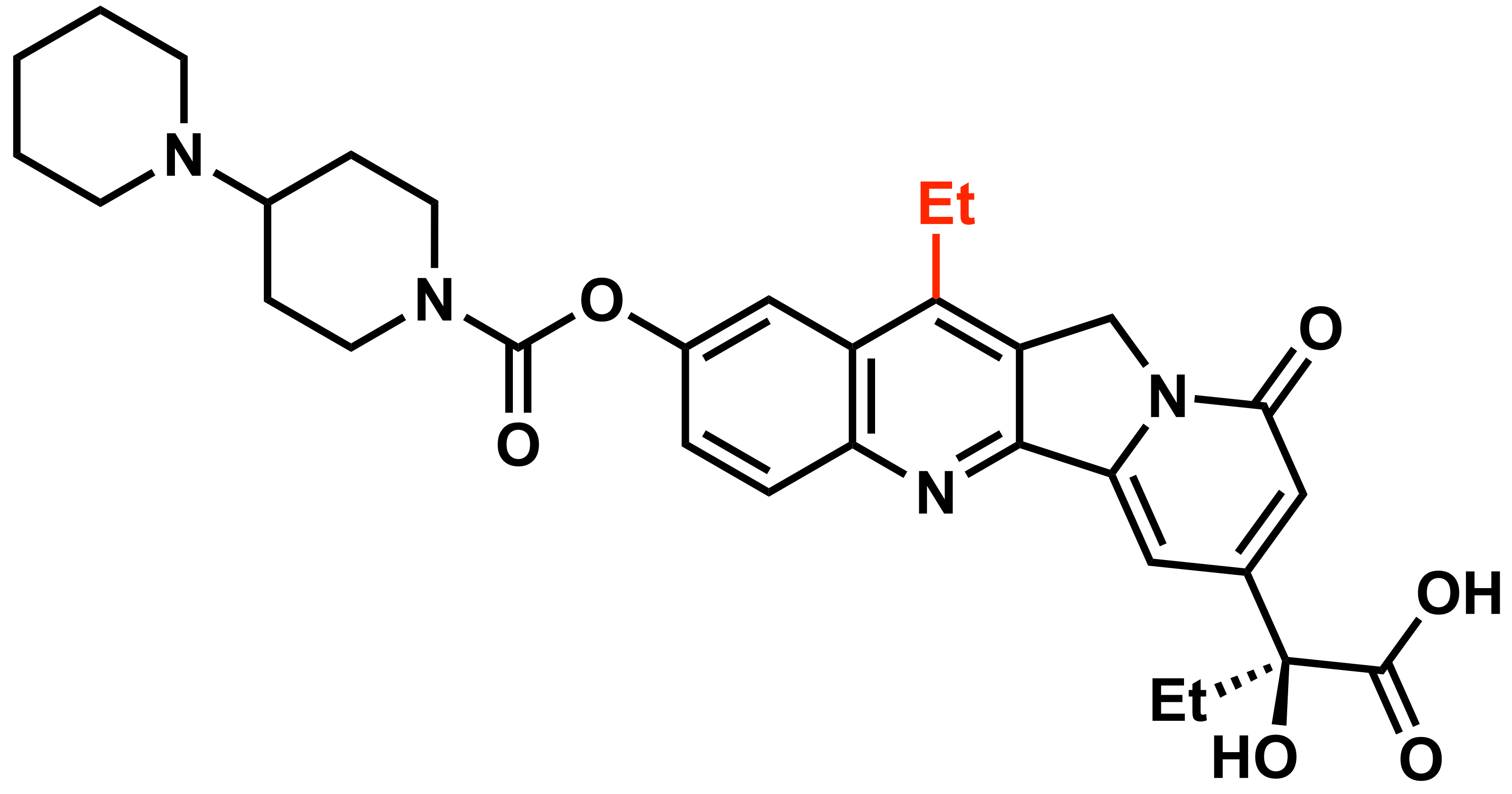
heterocycle

functionalized heterocycle



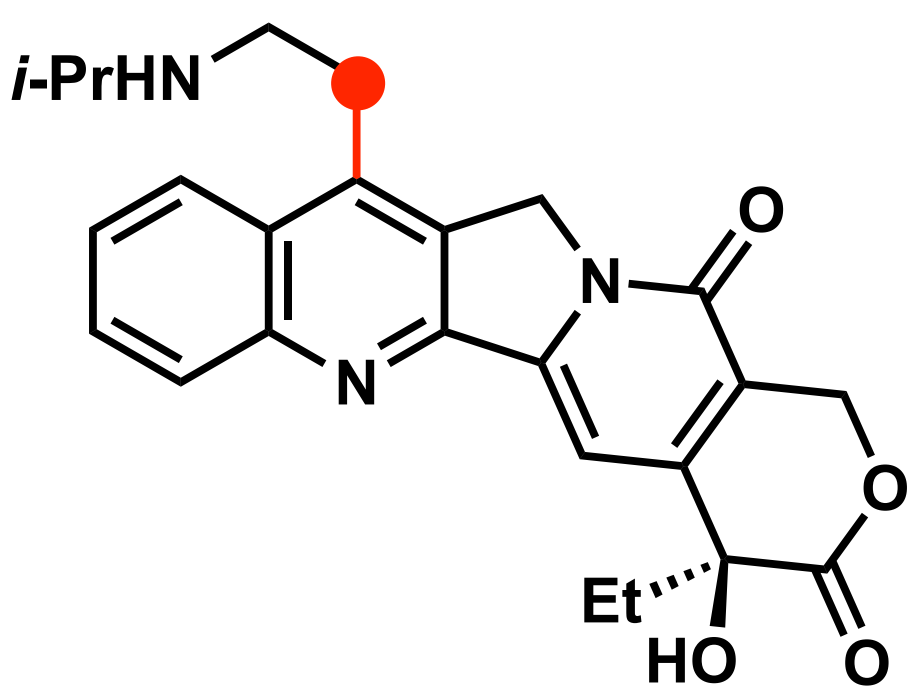
Viroptic

anti-viral eye drops



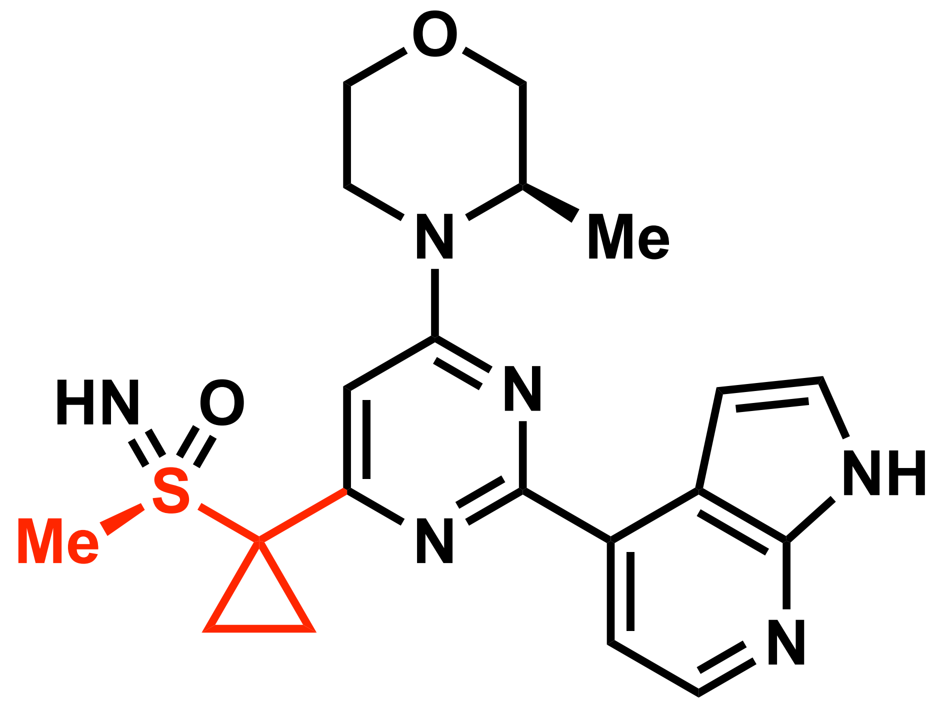
Irinotecan

colon, small-cell lung cancer



Belotecan

ovarian, small-cell lung cancer



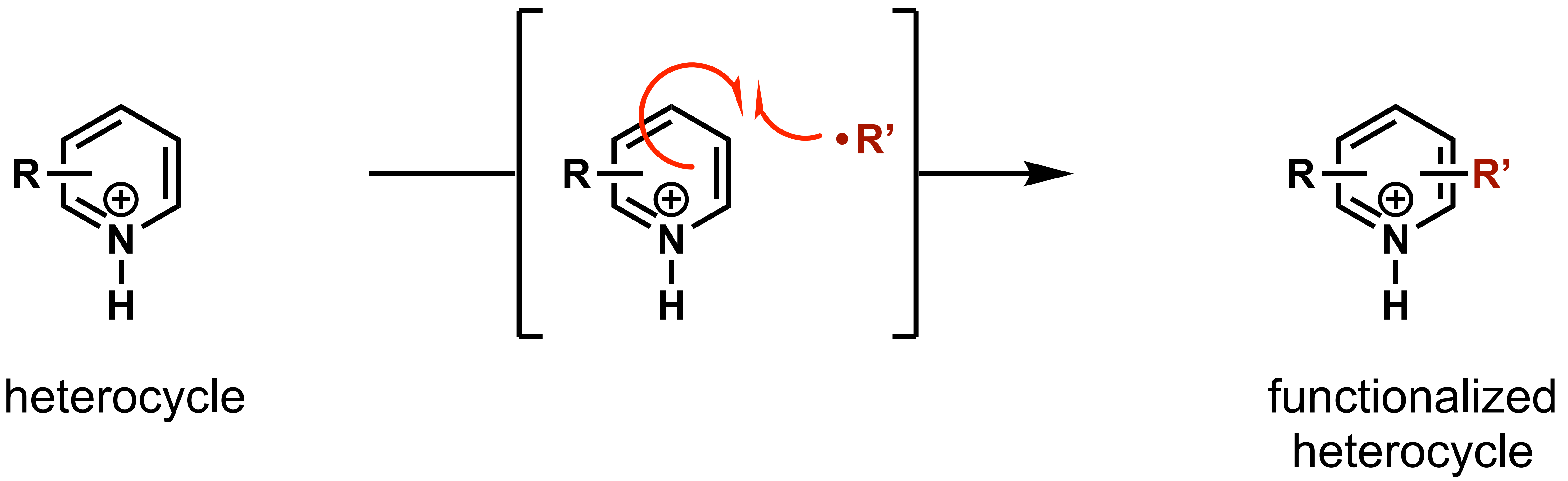
Ceralasertib

ovarian, small-cell lung, cervix cancer

J. Am. Chem. Soc. **2016**, *138*, 12692. *Med. Chem. Commun.* **2011**, *2*, 1135.

Med. Chem. Commun. **2011**, *2*, 1135. *Org. Process Res. Dev.* **2021**, *25*, 57.

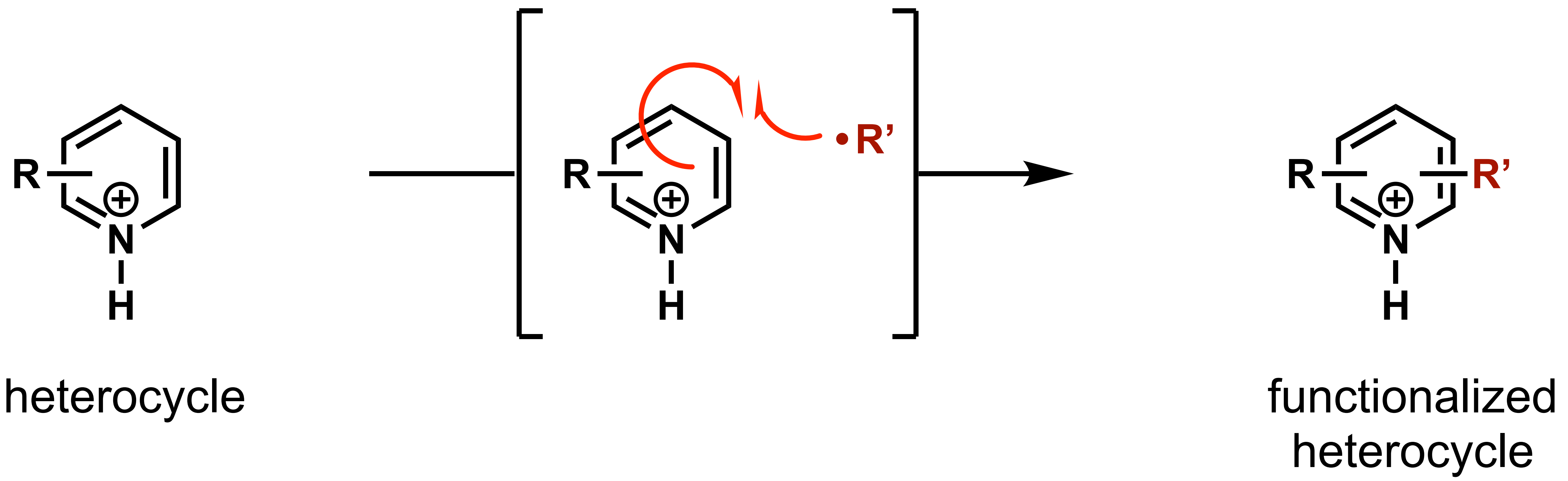
Regioselectivity Factors of the Minisci Reaction



Electronics, sterics, and longevity of •R'

J. Am. Chem. Soc. **2013**, *135*, 12122.

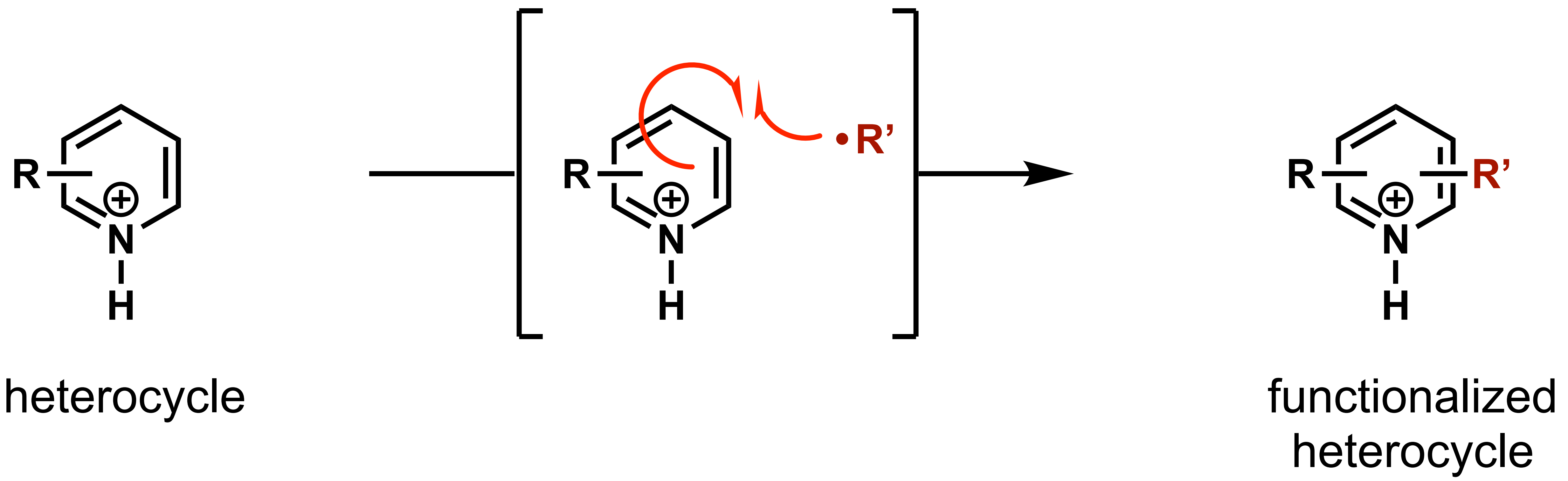
Regioselectivity Factors of the Minisci Reaction



Electronics, sterics, and longevity of **•R'**

Electronics of heterocycle

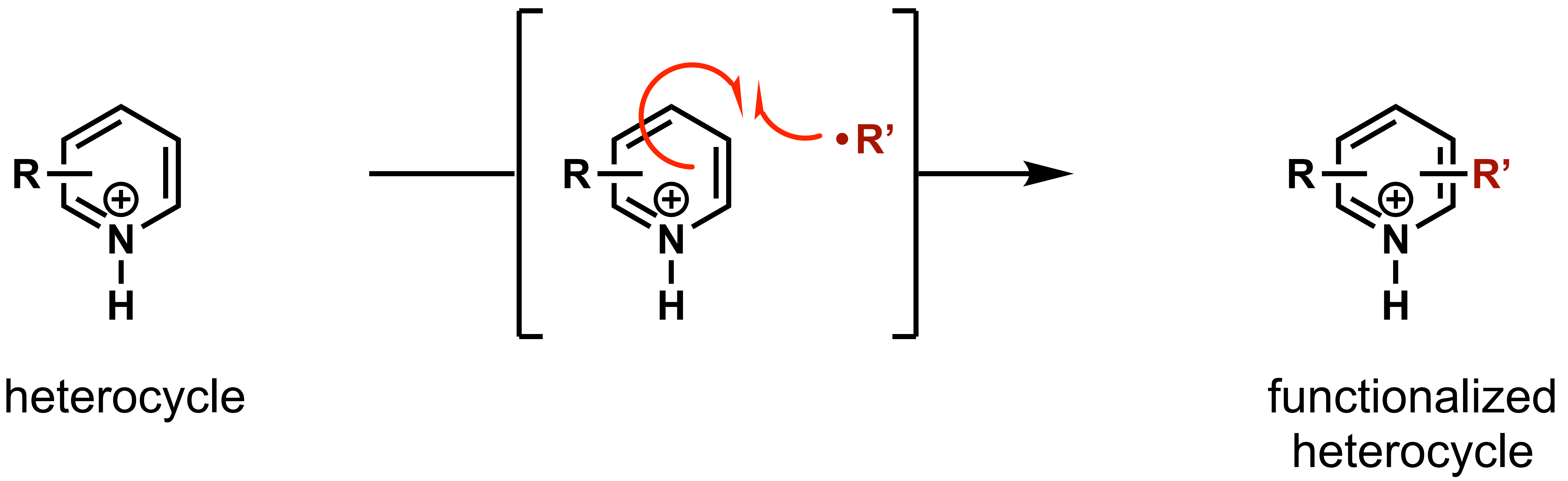
Prediction of Regiochemical Outcomes of the Minisci Reaction



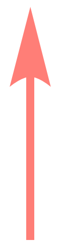
DFT-derived Fukui reactivity indices are ~90% accurate.

RSC Adv. **2014**, *4*, 17262. *ChemMedChem* **2018**, *13*, 983.

Prediction of Regiochemical Outcomes of the Minisci Reaction



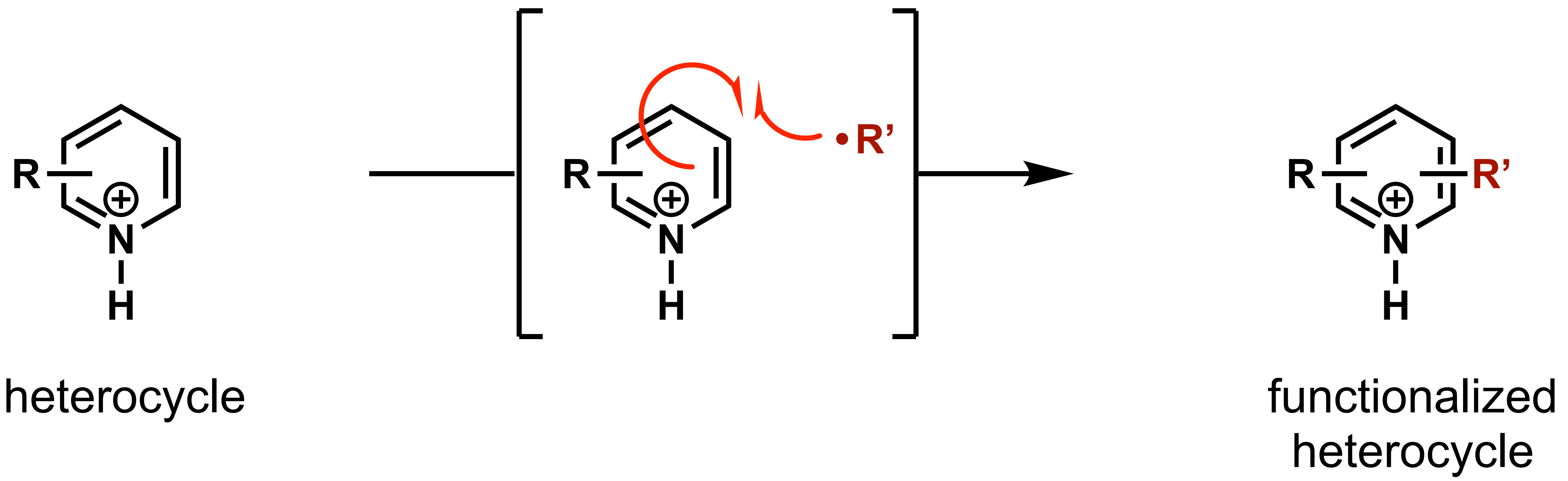
DFT-derived Fukui reactivity indices are ~90% accurate.



Molecule Complexity

RSC Adv. **2014**, *4*, 17262. *ChemMedChem* **2018**, *13*, 983.

Prediction of Regiochemical Outcomes of the Minisci Reaction

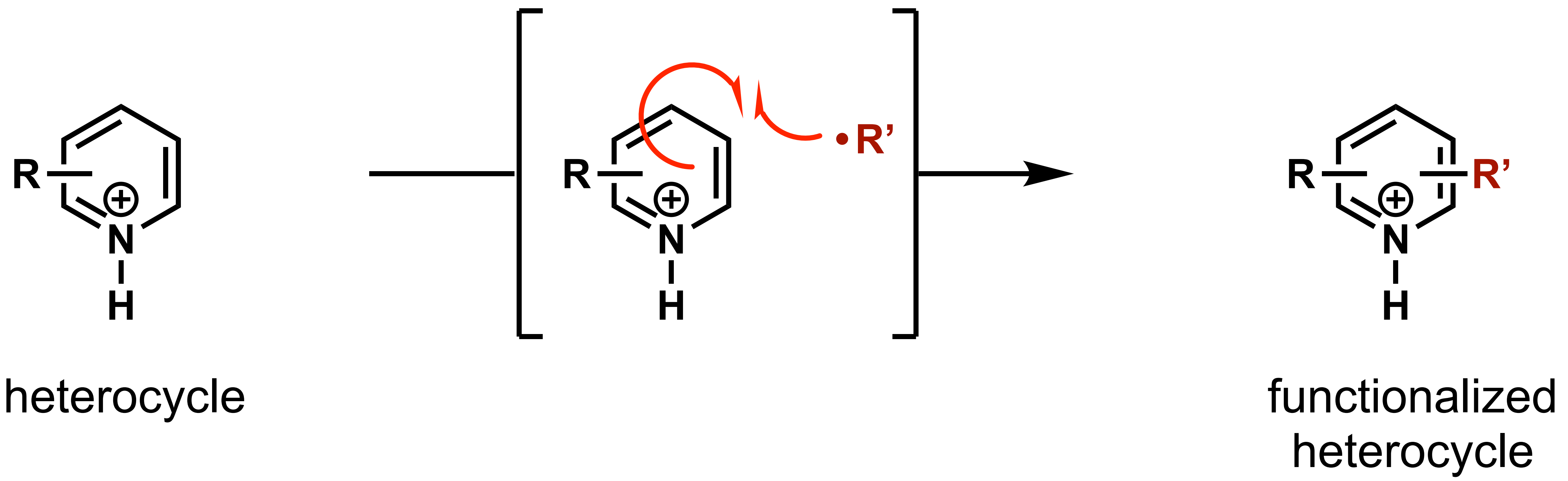


DFT-derived Fukui reactivity indices are ~90% accurate.

Molecule Complexity → **Possible Reaction Sites**

RSC Adv. **2014**, *4*, 17262. *ChemMedChem* **2018**, *13*, 983.

Prediction of Regiochemical Outcomes of the Minisci Reaction

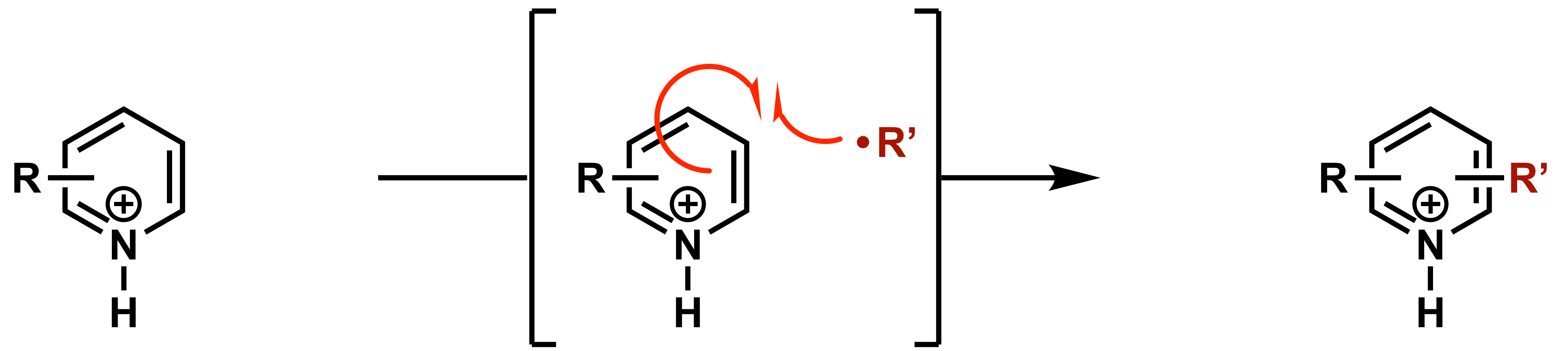


DFT-derived Fukui reactivity indices are ~90% accurate.

Molecule Complexity → **Possible Reaction Sites** → **Regiochemical Prediction Accuracy**

RSC Adv. **2014**, *4*, 17262. *ChemMedChem* **2018**, *13*, 983.

Prediction of Regiochemical Outcomes of the Minisci Reaction



heterocycle

functionalized heterocycle

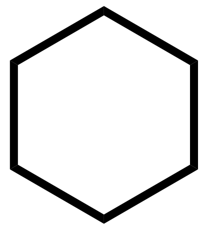
DFT-derived Fukui reactivity indices are ~90% accurate.

↑ **Molecule Complexity** → ↑ **Possible Reaction Sites** → ↓ **Regiochemical Prediction Accuracy**

Can machine learning provide some improvement?

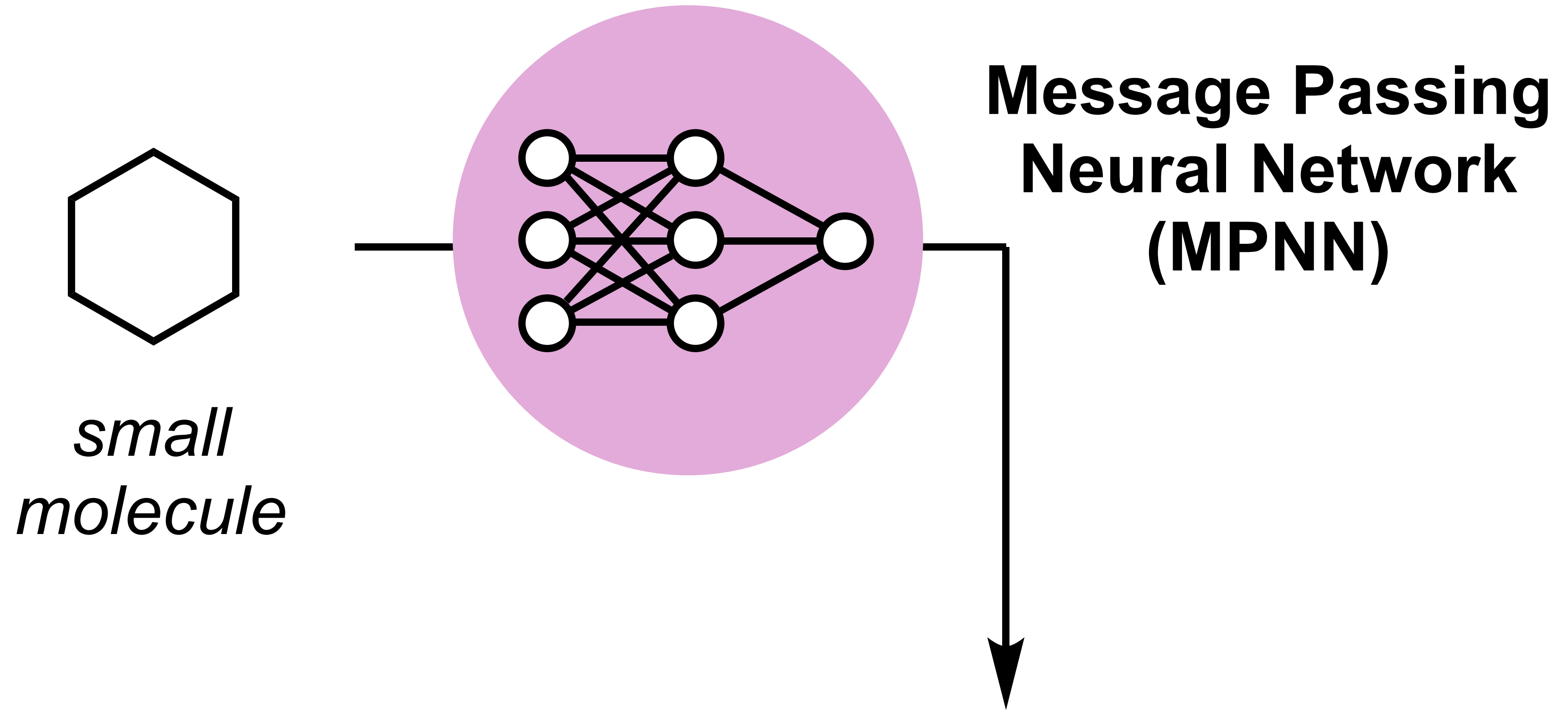
RSC Adv. **2014**, *4*, 17262. *ChemMedChem* **2018**, *13*, 983.

The Big Idea

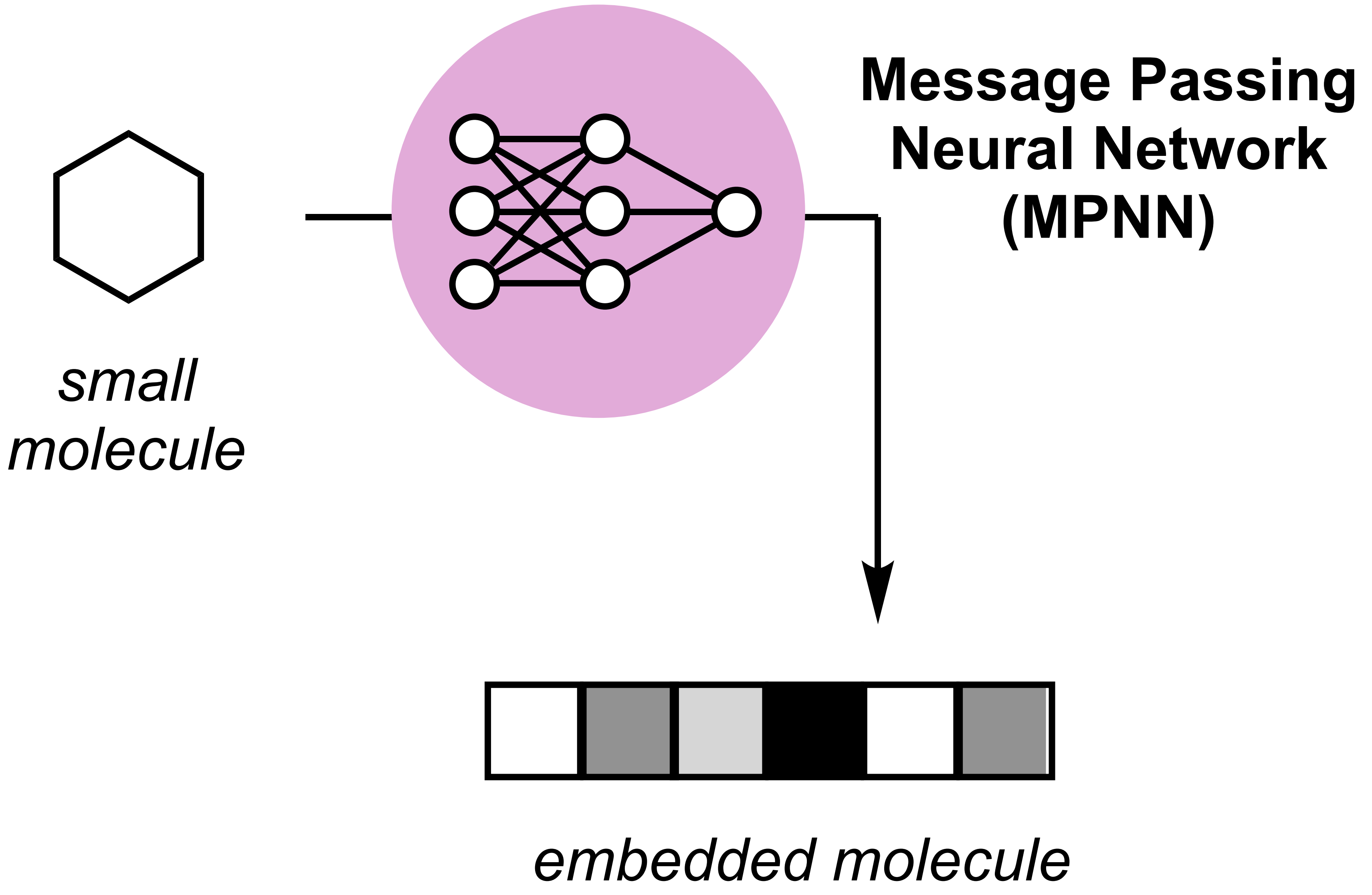


small molecule

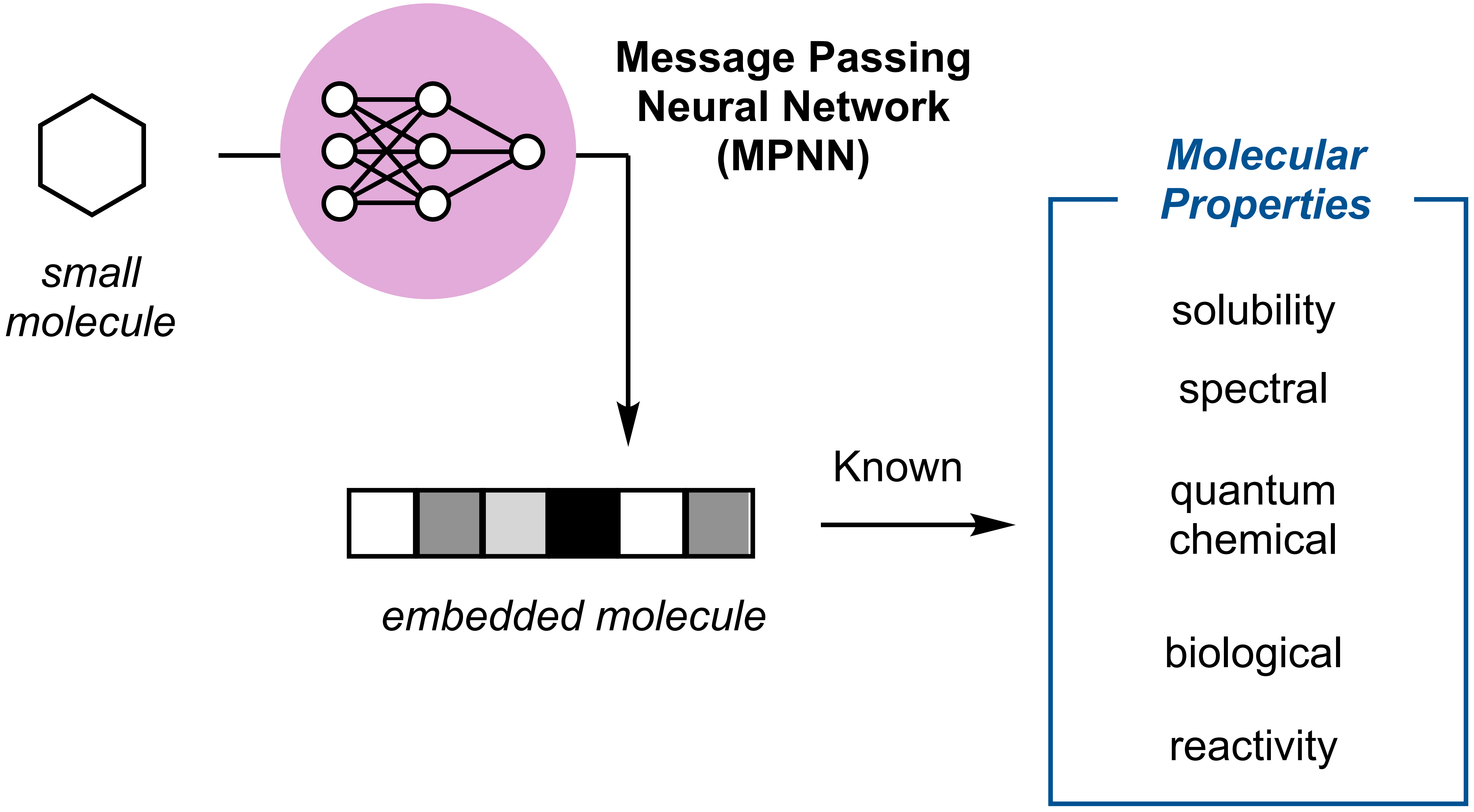
The Big Idea



The Big Idea

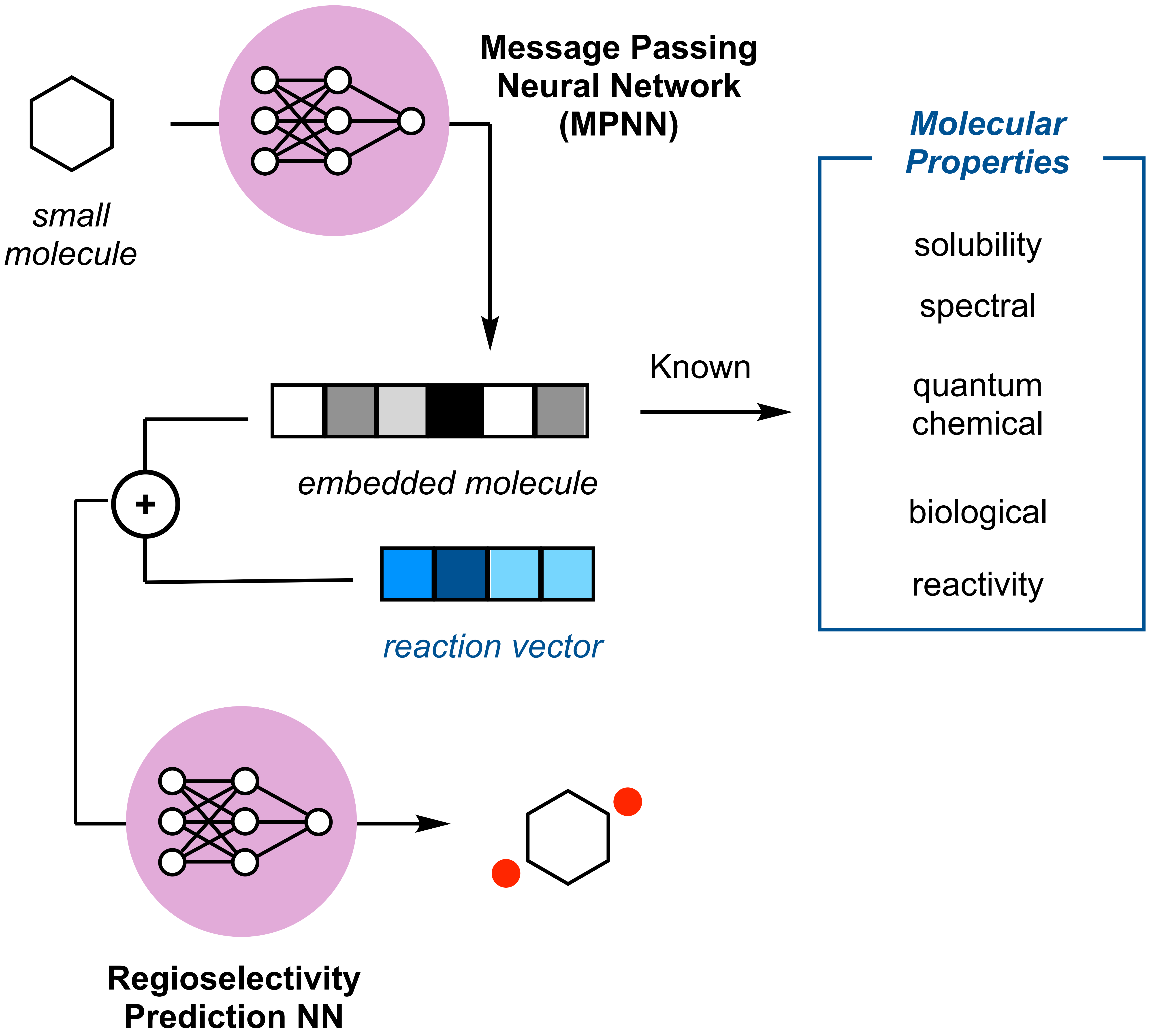

Message Passing
Neural Network
(MPNN)
small
molecule
embedded molecule

The Big Idea



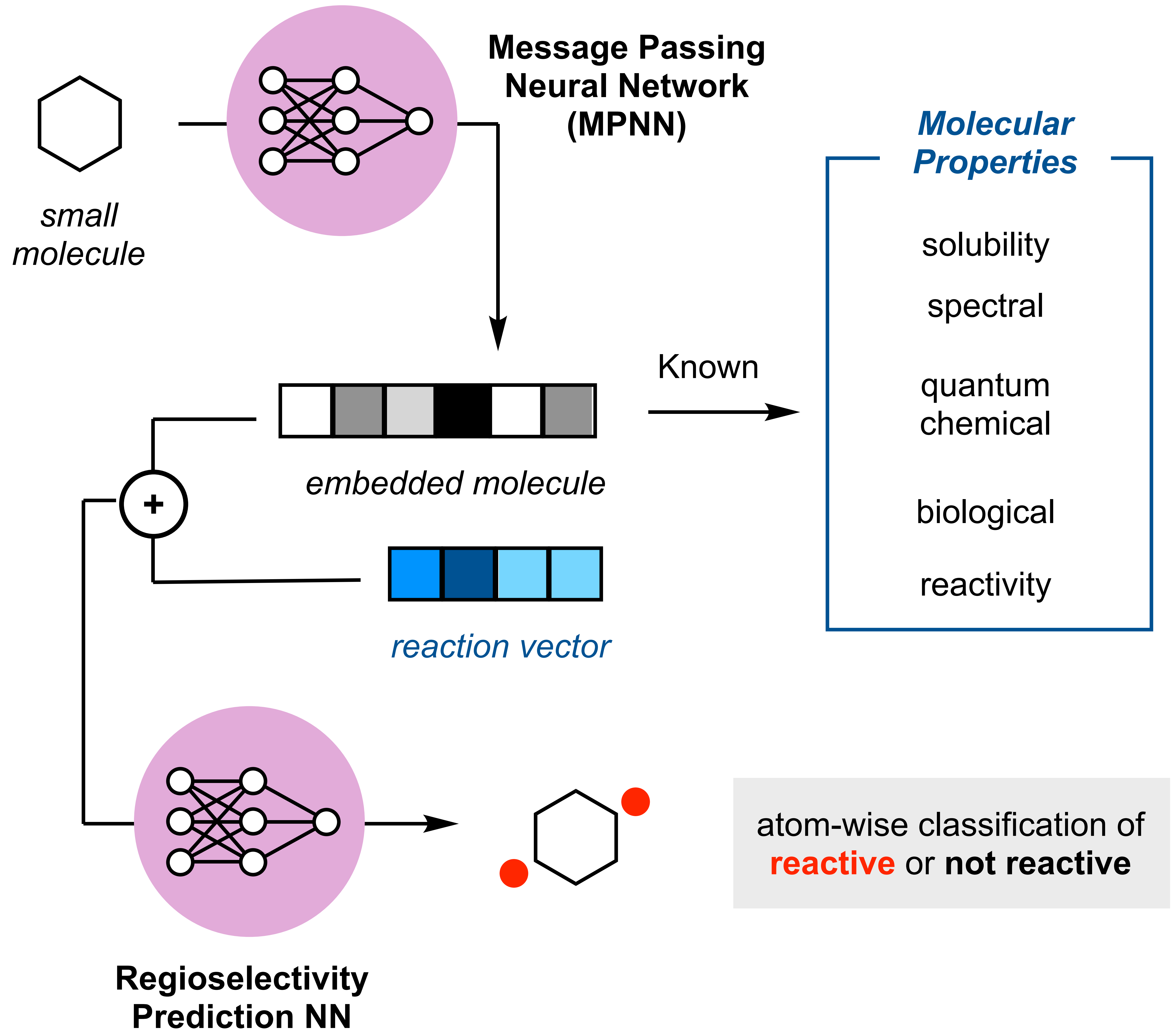
ChemRxiv **2022**, DOI: 10.26343/chemrxiv-2022-gkxm6-v2 *J Chemoinformatics* **2020**, *12*, 1. *J. Chem. Inf. Model.* **2021**, *61*, 2594.
J Chemoinformatics **2020**, *12*, 15. *Chem. Sci.* **2021**, *12*, 2198.

The Big Idea



ChemRxiv **2022**, DOI: 10.26343/chemrxiv-2022-gkxm6-v2 *J Chemoinformatics* **2020**, *12*, 1. *J. Chem. Inf. Model.* **2021**, *61*, 2594.
J Chemoinformatics **2020**, *12*, 15. *Chem. Sci.* **2021**, *12*, 2198.

The Big Idea

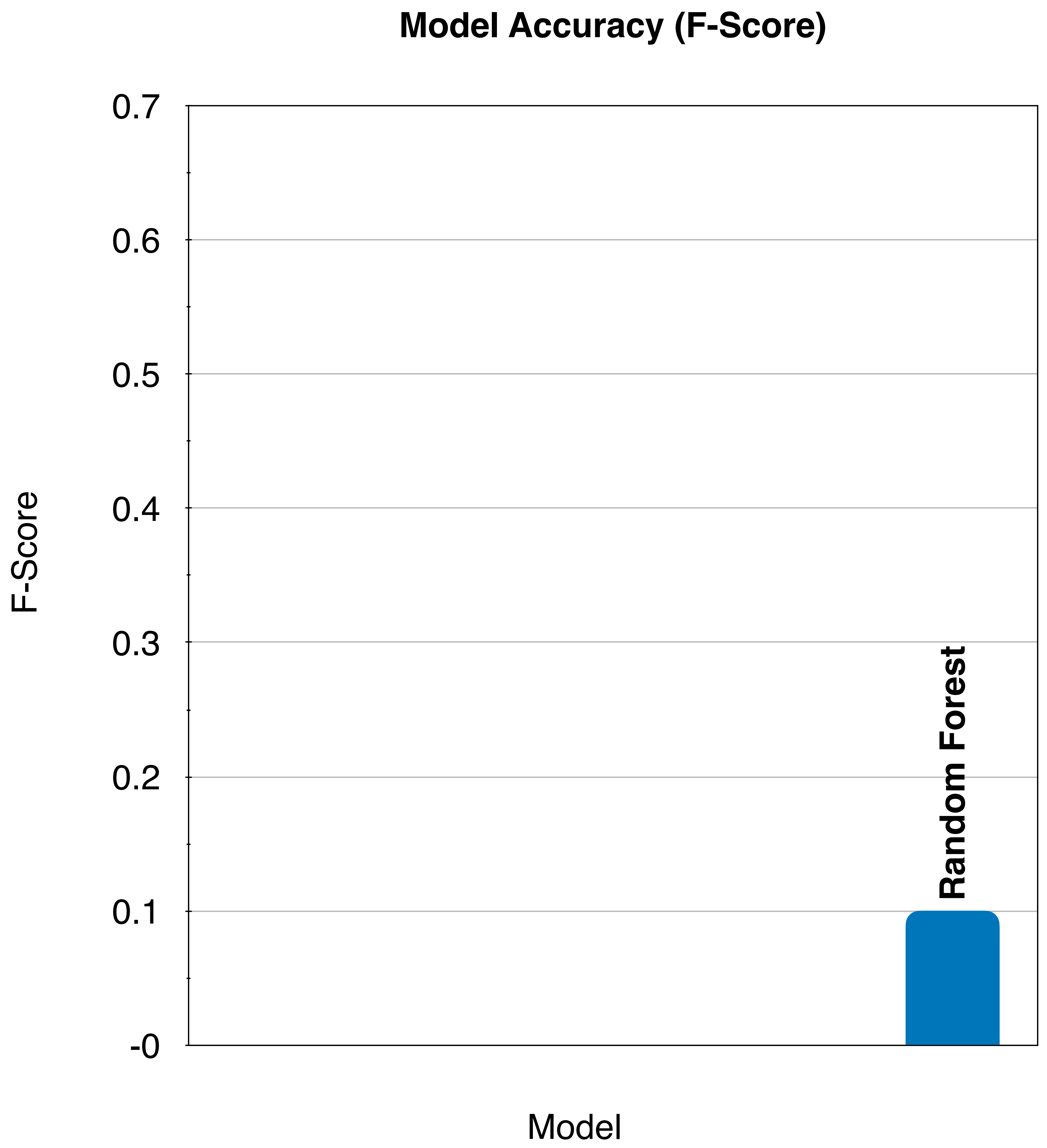


ChemRxiv **2022**, DOI: 10.26343/chemrxiv-2022-gkxm6-v2 *J Chemoinformatics* **2020**, *12*, 1. *J. Chem. Inf. Model.* **2021**, *61*, 2594.

J Chemoinformatics **2020**, *12*, 15. *Chem. Sci.* **2021**, *12*, 2198.

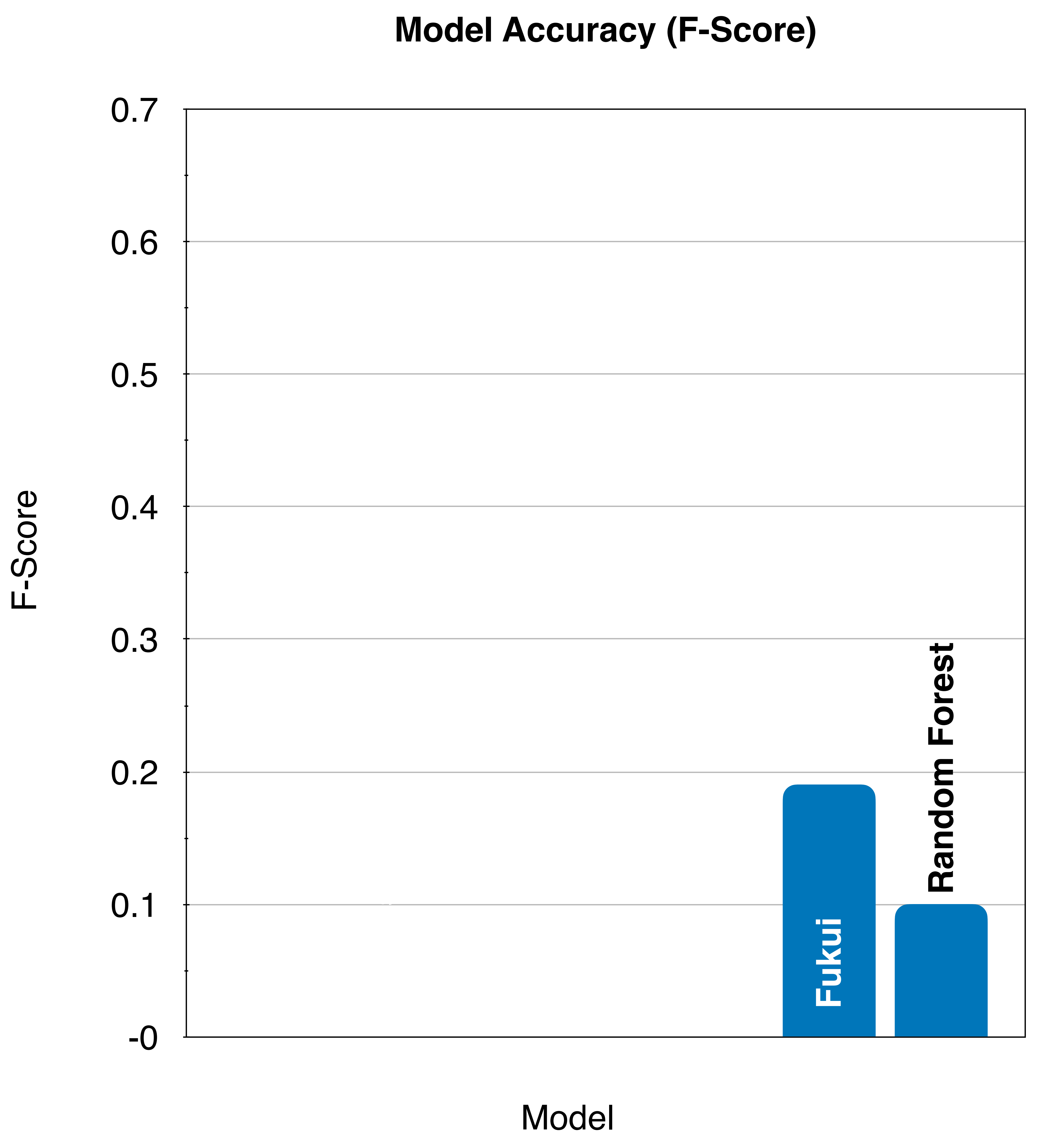
Baselines

UNIVERSITY OF CAMBRIDGE



Nat. Commun. **2024**, *15*, 426.

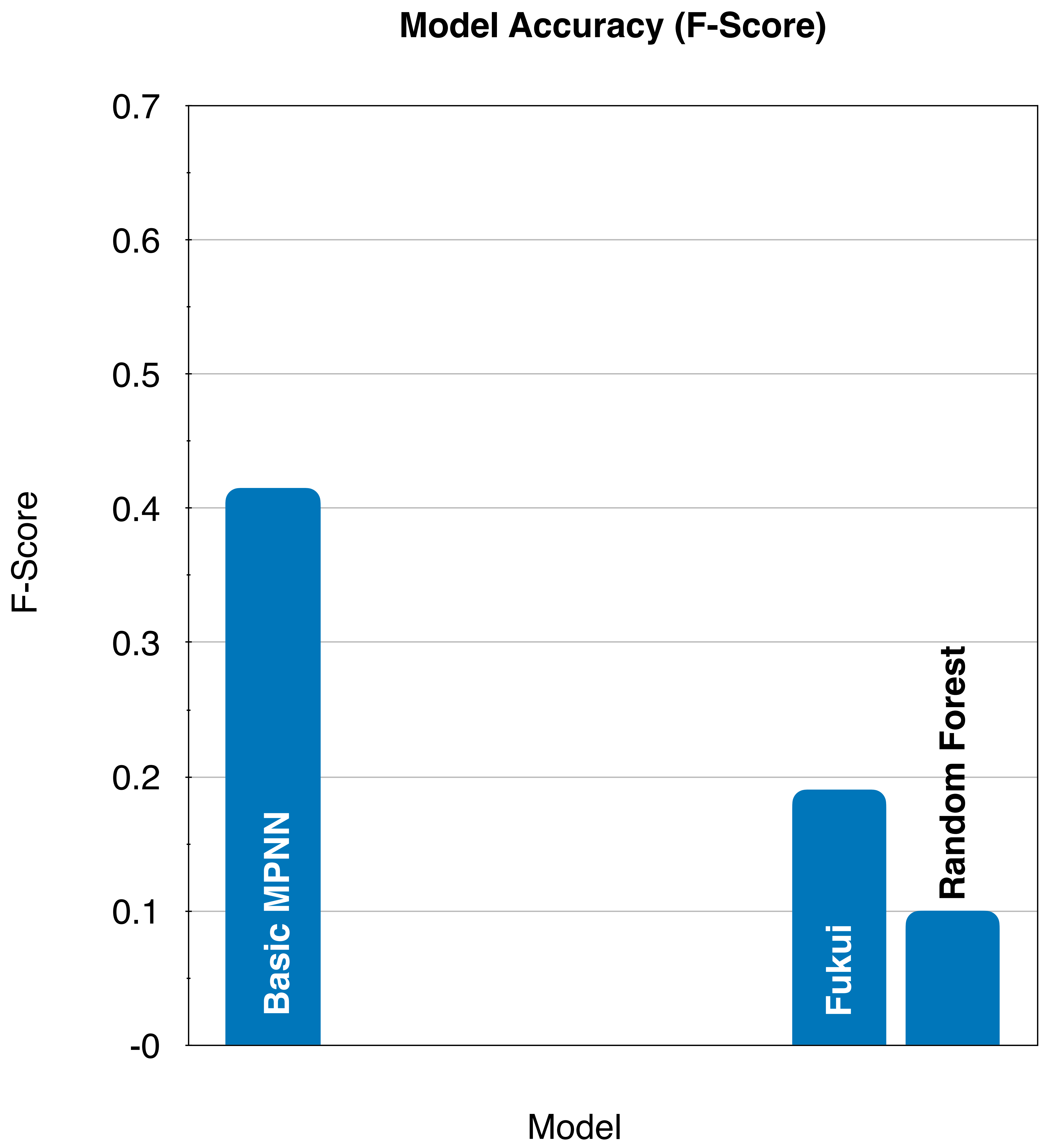
Baselines



Nat. Commun. **2024**, *15*, 426.

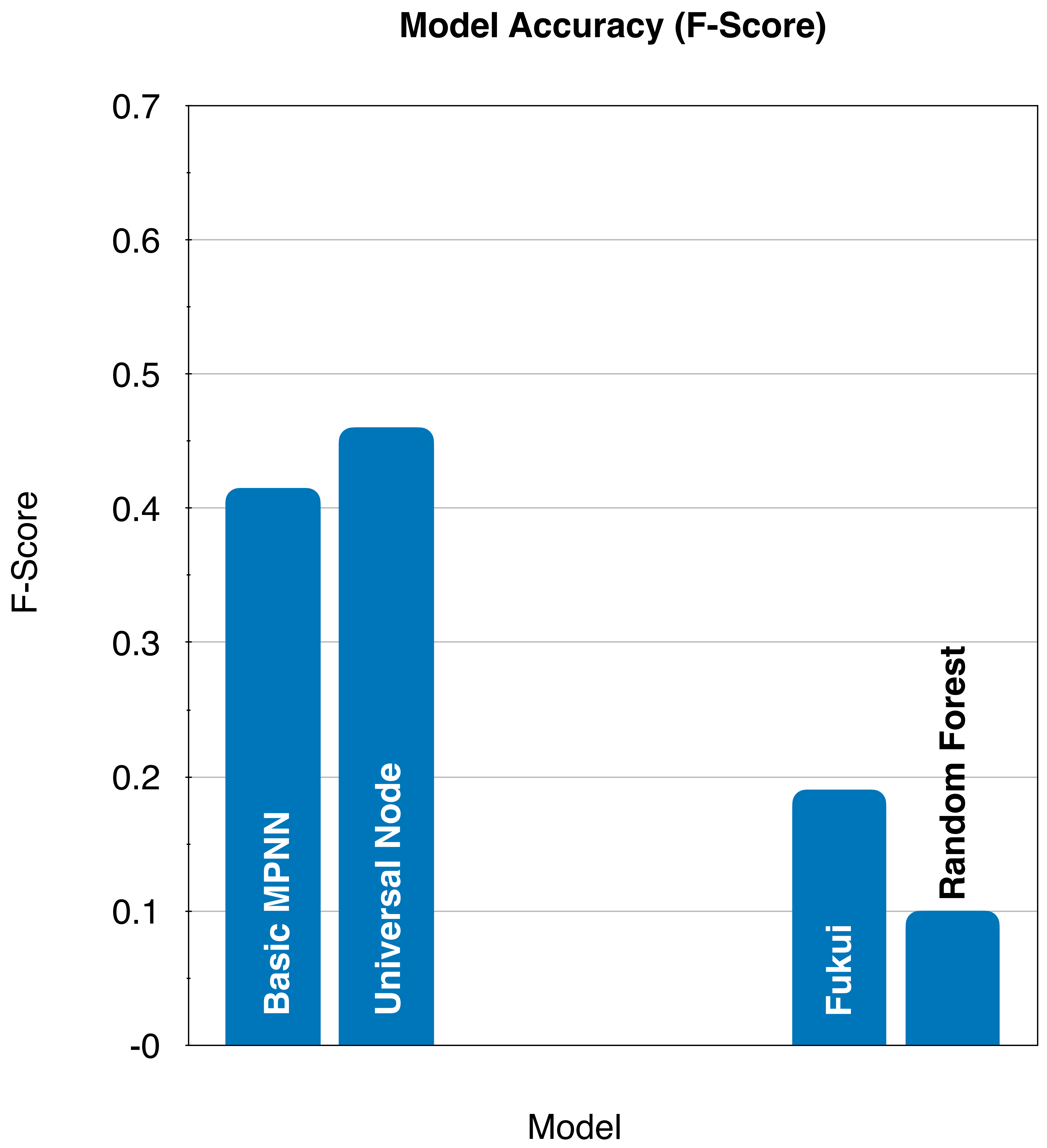
A Good Start

UNIVERSITY OF CAMBRIDGE



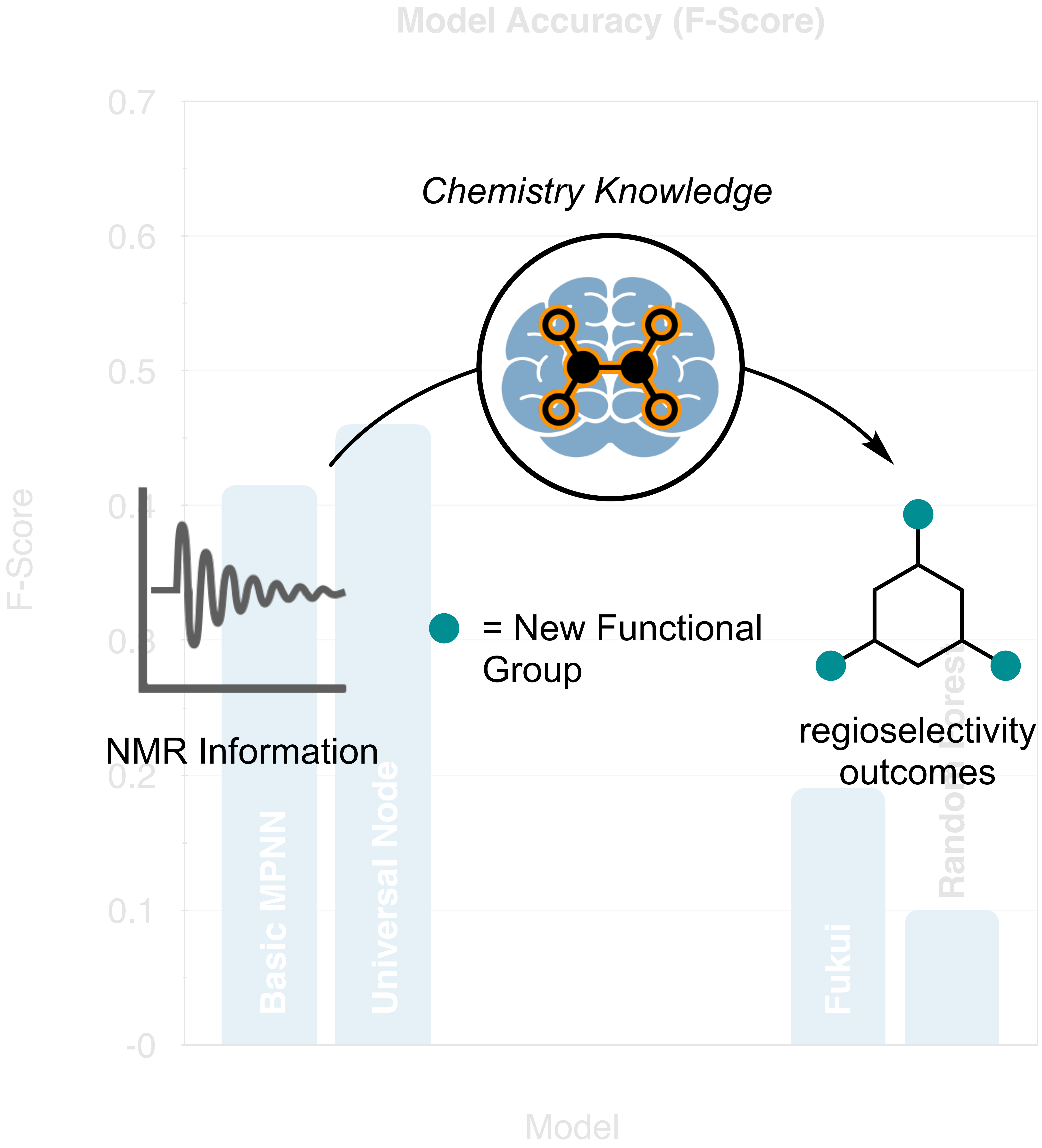
Nat. Commun. **2024**, *15*, 426.

A Modest Improvement



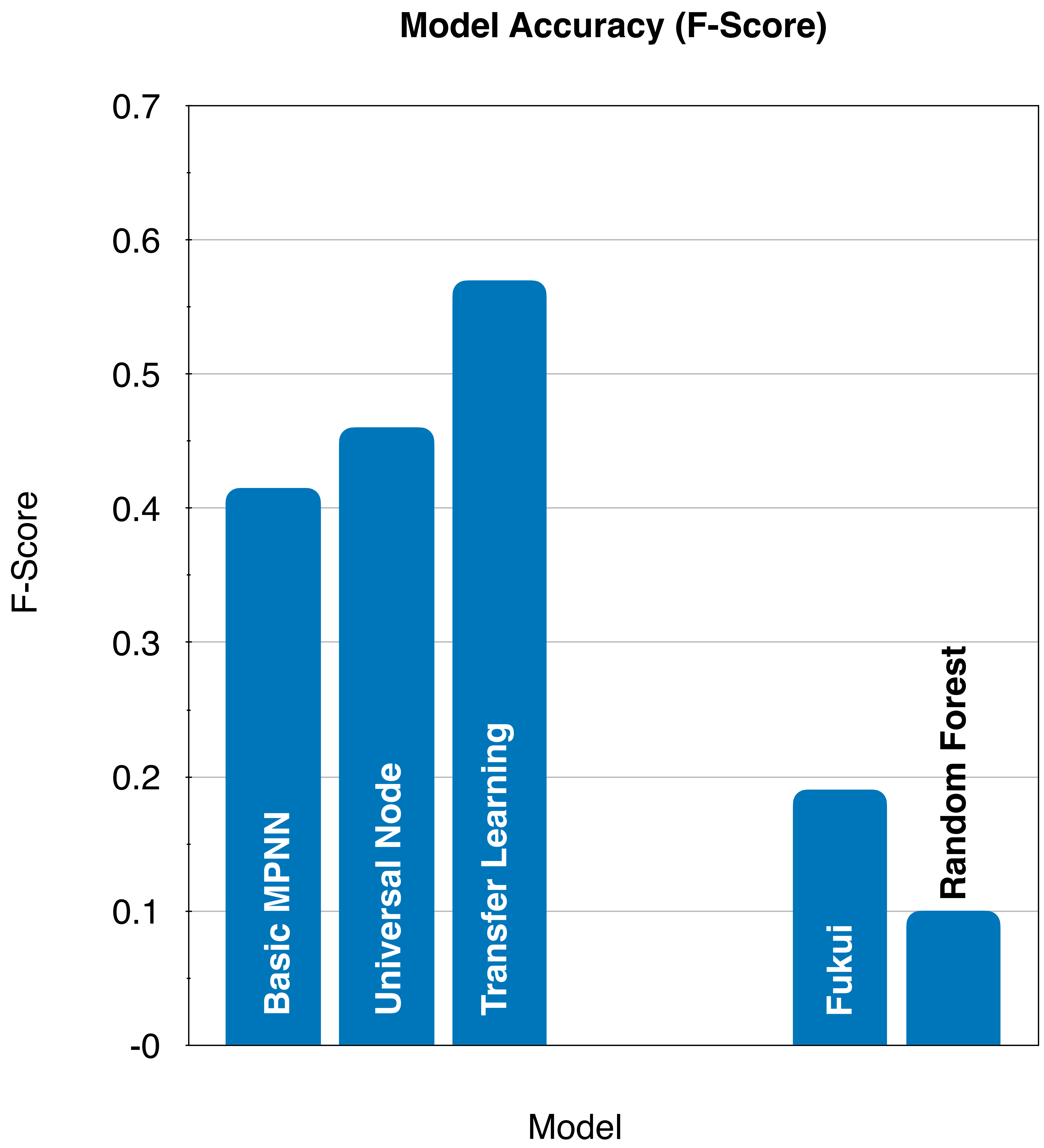
Nat. Commun. **2024**, *15*, 426.

A Modest Improvement



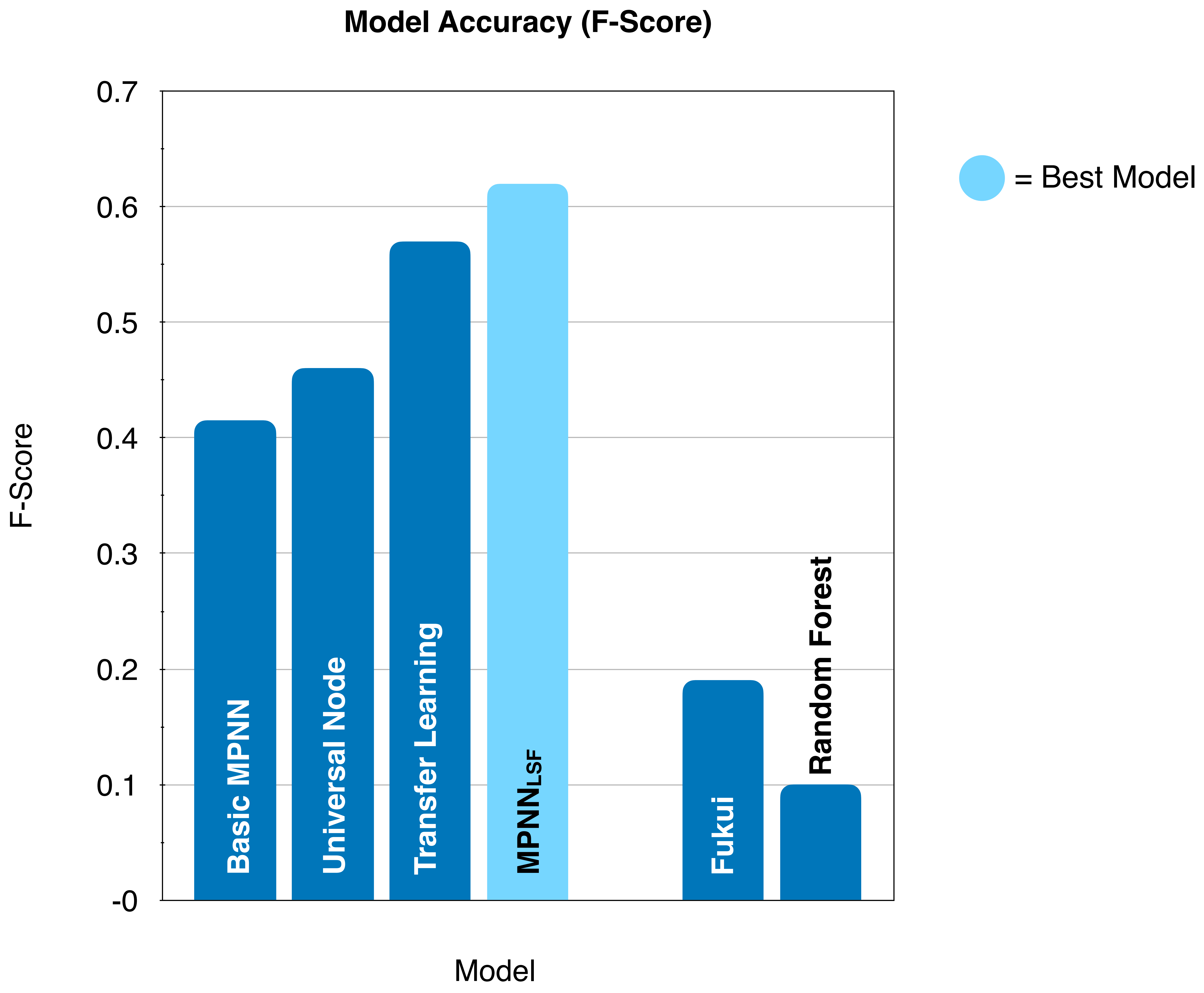
Nat. Commun. **2024**, *15*, 426.

Significant Improvement!



Nat. Commun. **2024**, *15*, 426.

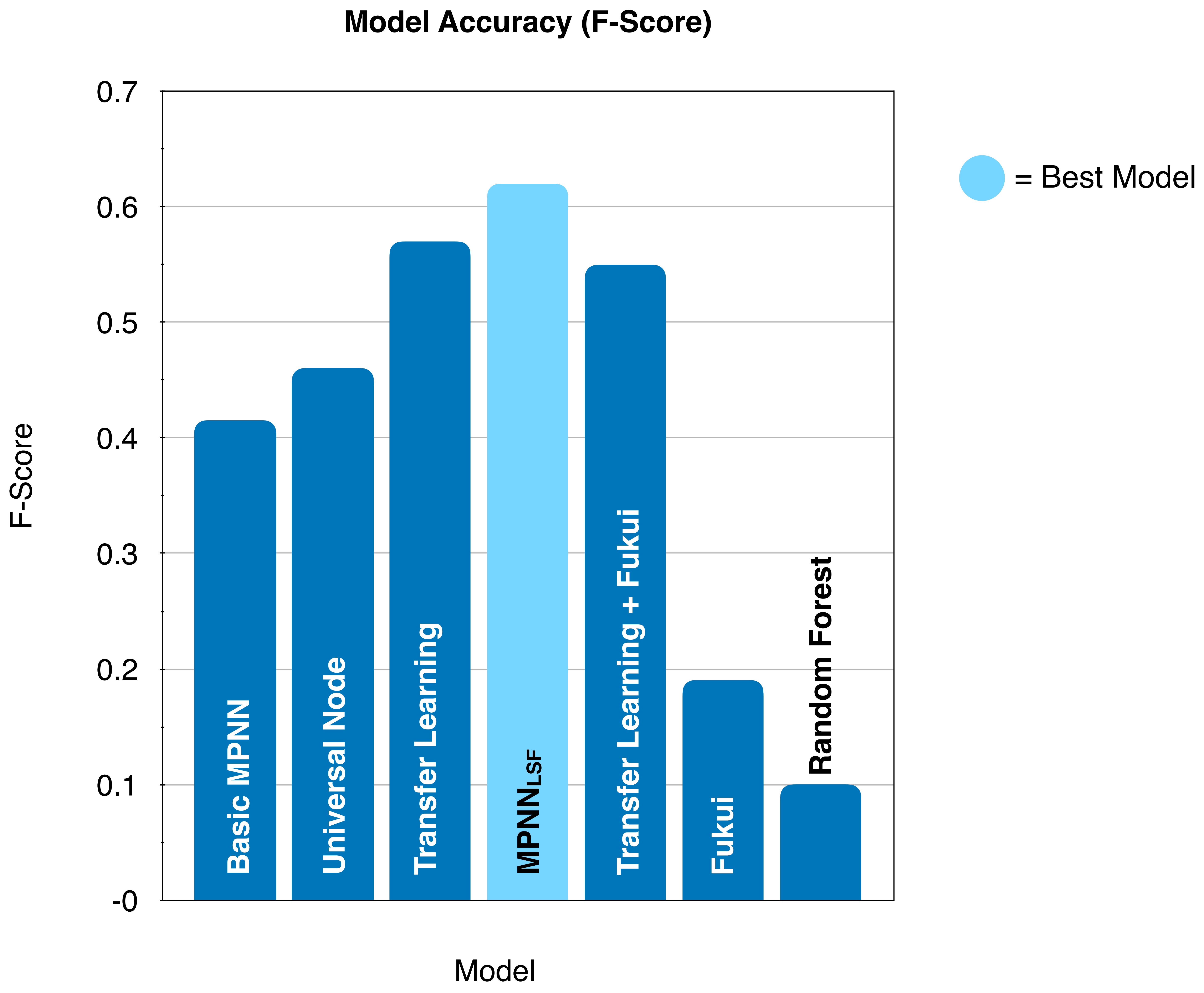
Significant Improvement!



Nat. Commun. **2024**, *15*, 426.

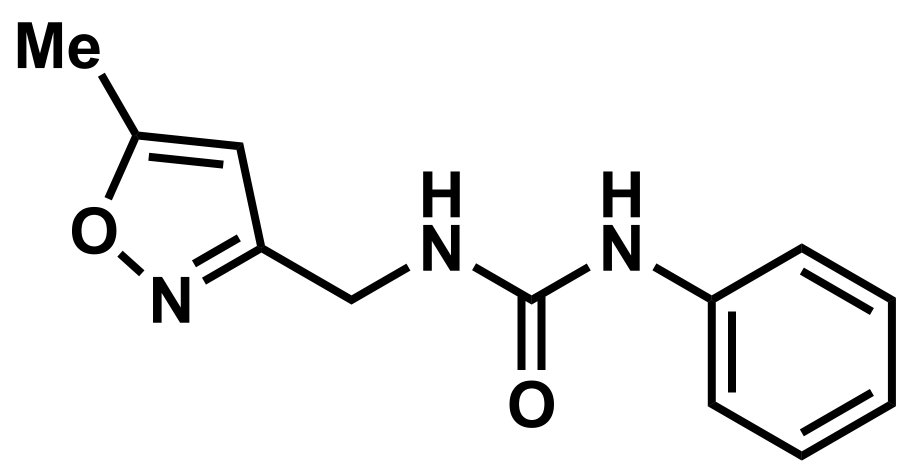
Inclusion of Fukui Indices as Atom Information

UNIVERSITY OF CAMBRIDGE

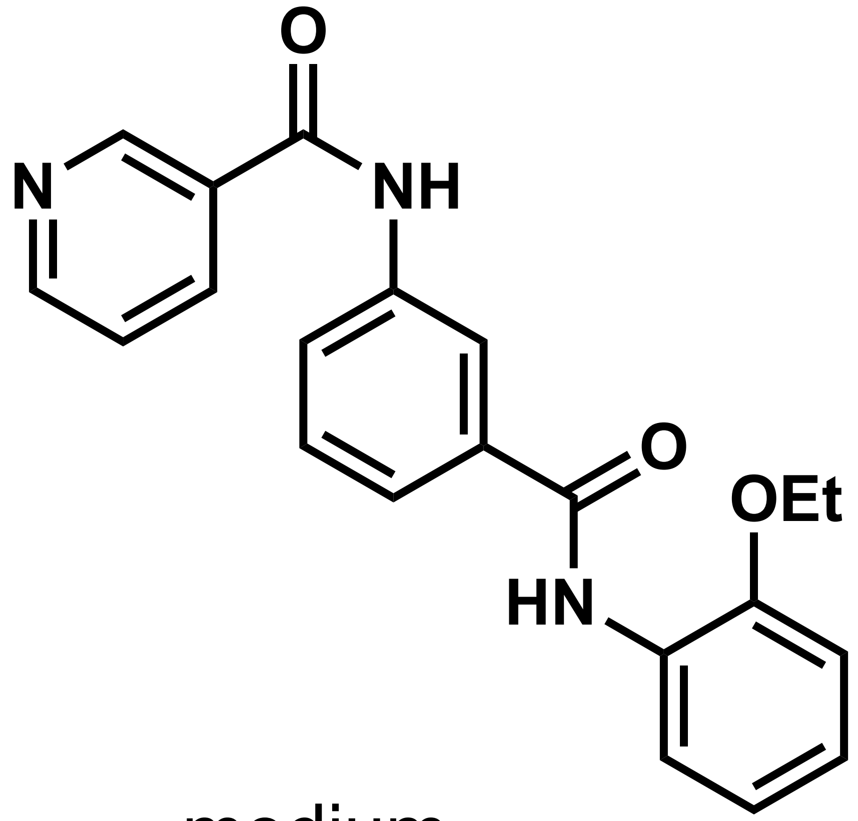


Nat. Commun. **2024**, *15*, 426.

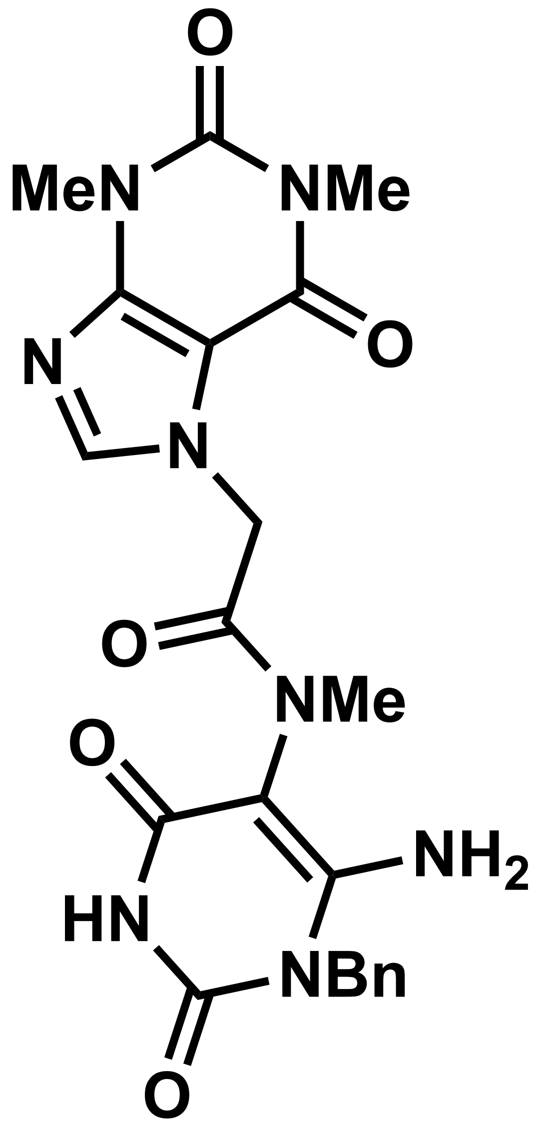
How Do We Perform on Completely Unseen Molecules?



small



medium

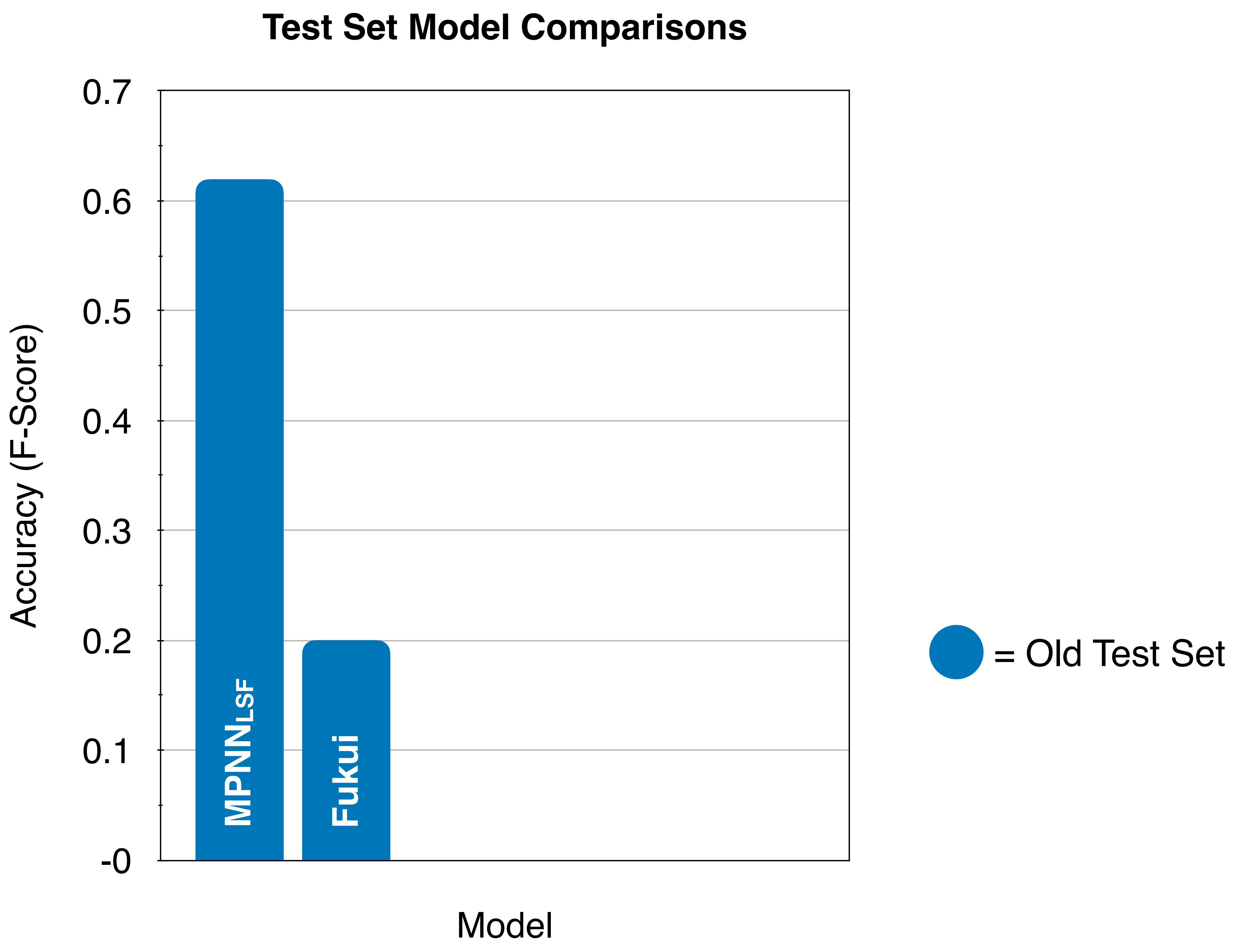


large

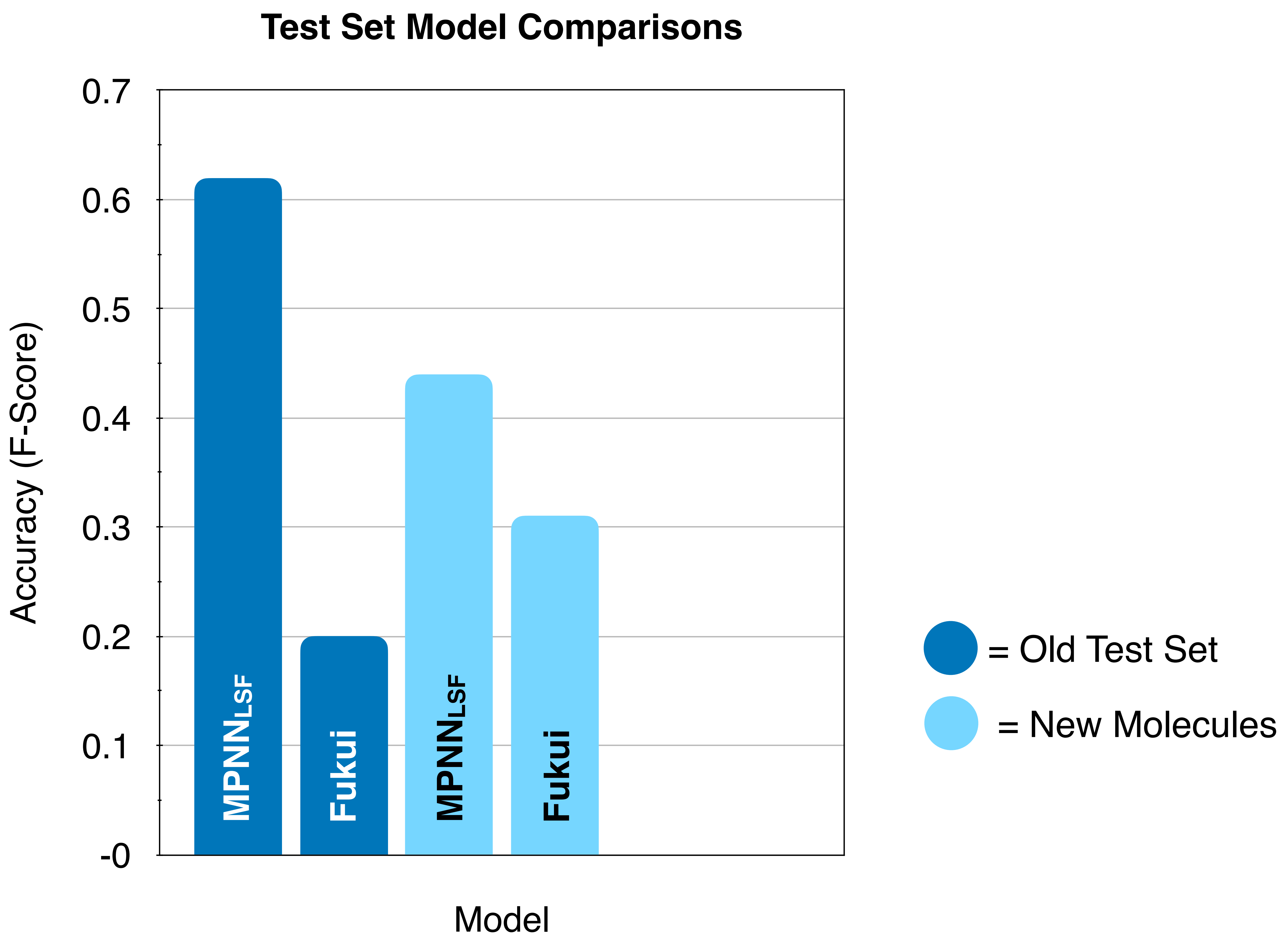
How do we do in a real-life scenario?

Nat. Commun. **2024**, *15*, 426.

How Do We Perform on Completely Unseen Molecules?

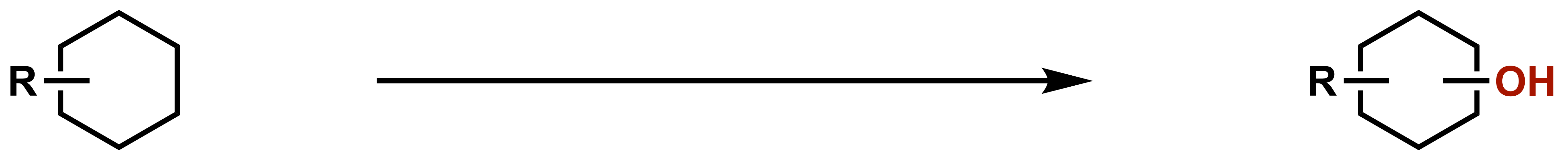


How Do We Perform on Completely Unseen Molecules?



Nat. Commun. **2024**, *15*, 426.

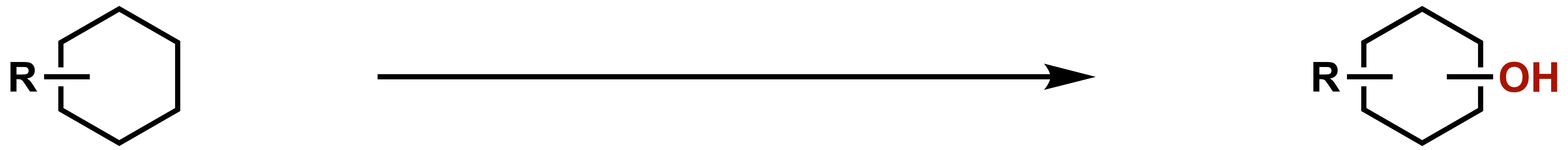
A Different Reaction Altogether



molecule

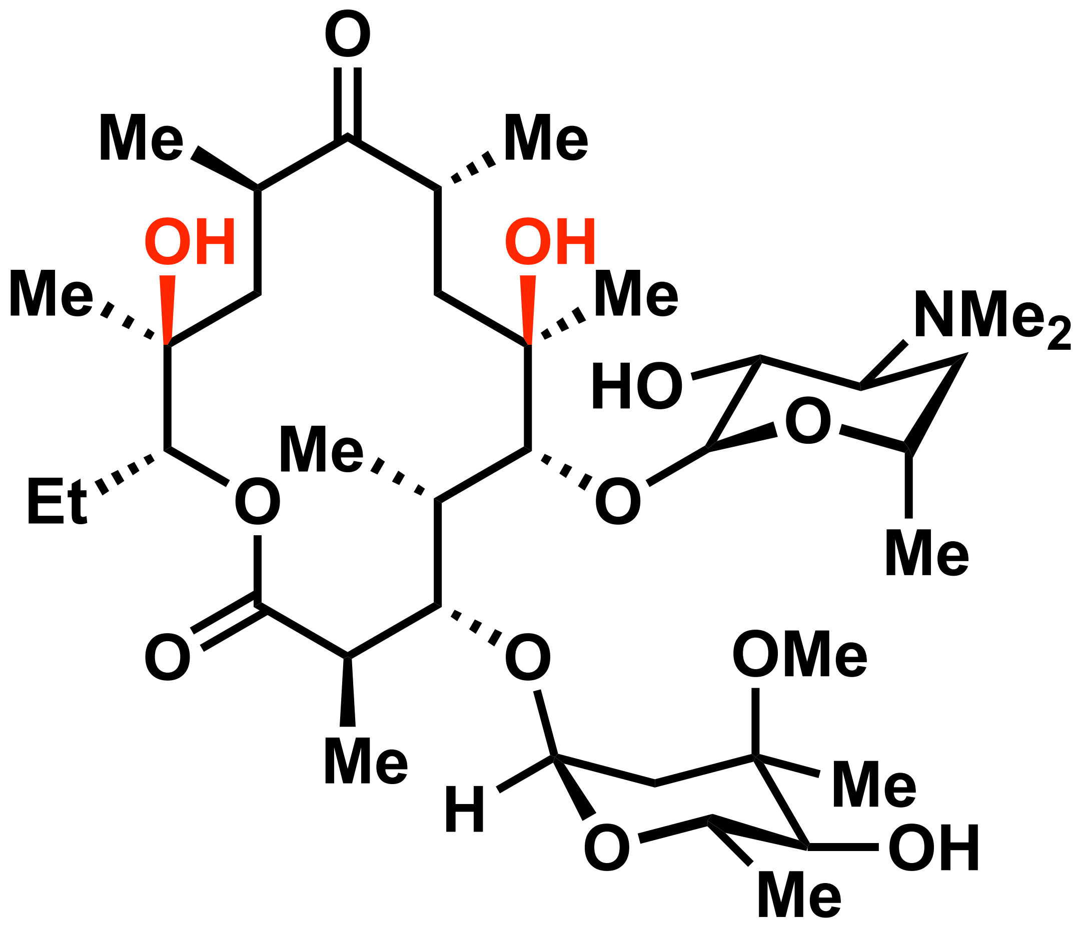
functionalized
molecule

A Different Reaction Altogether



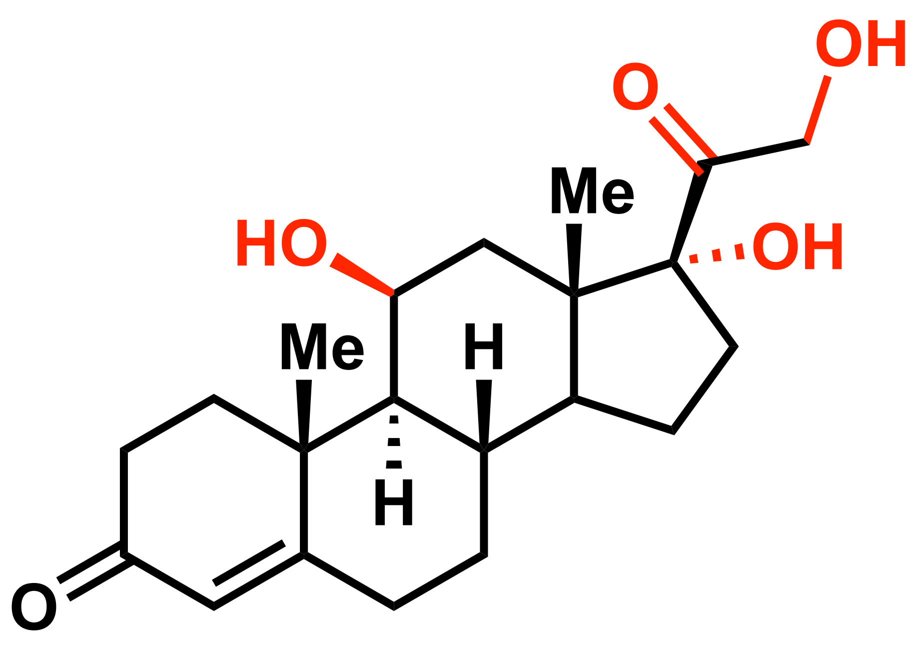
molecule

functionalized molecule



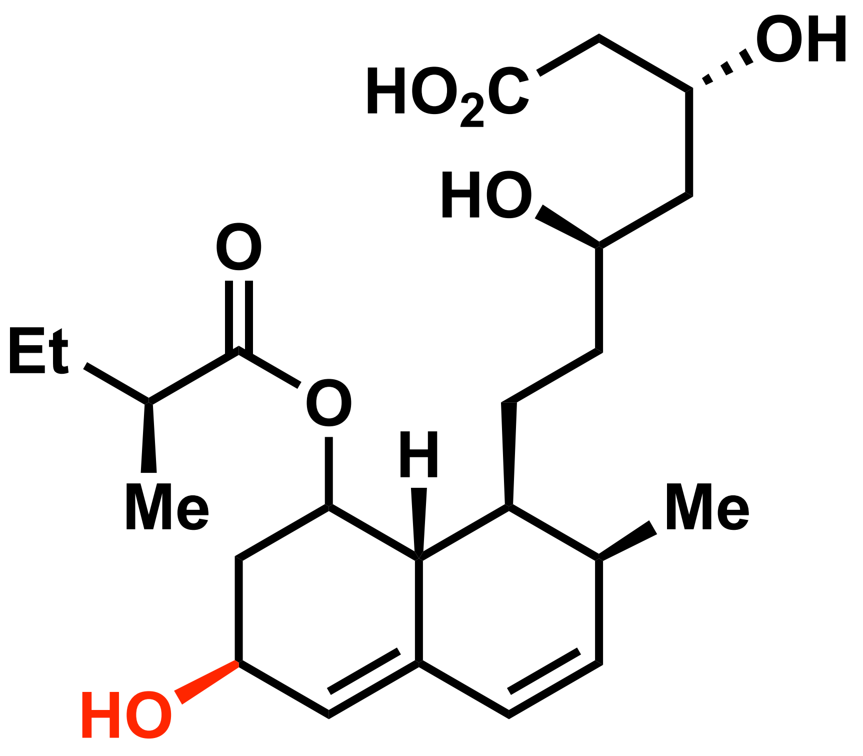
erythromycin

common anti-bacterial



hydrocortisone

common anti-inflammatory

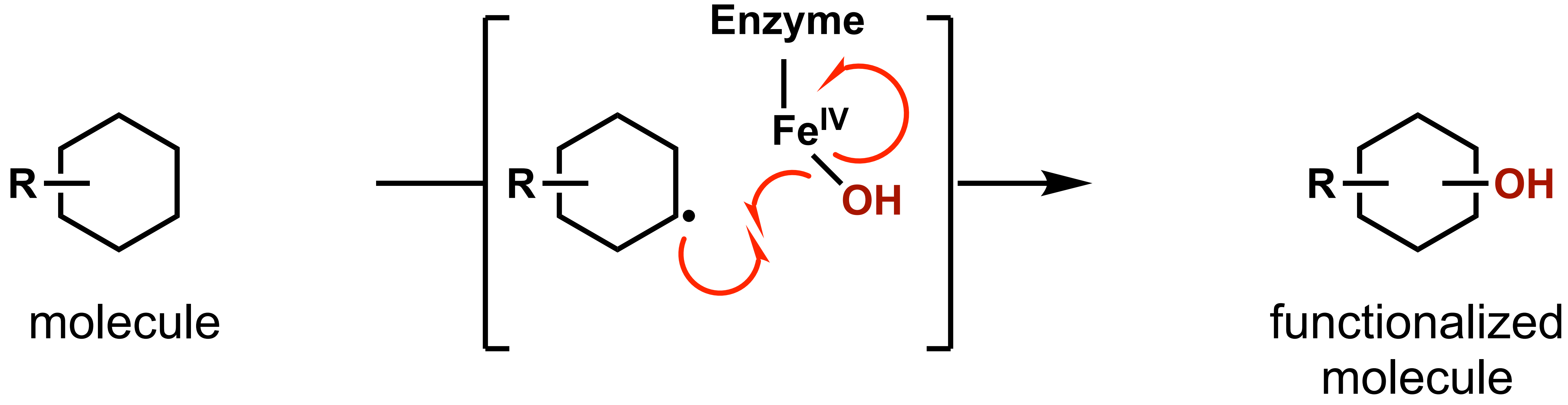


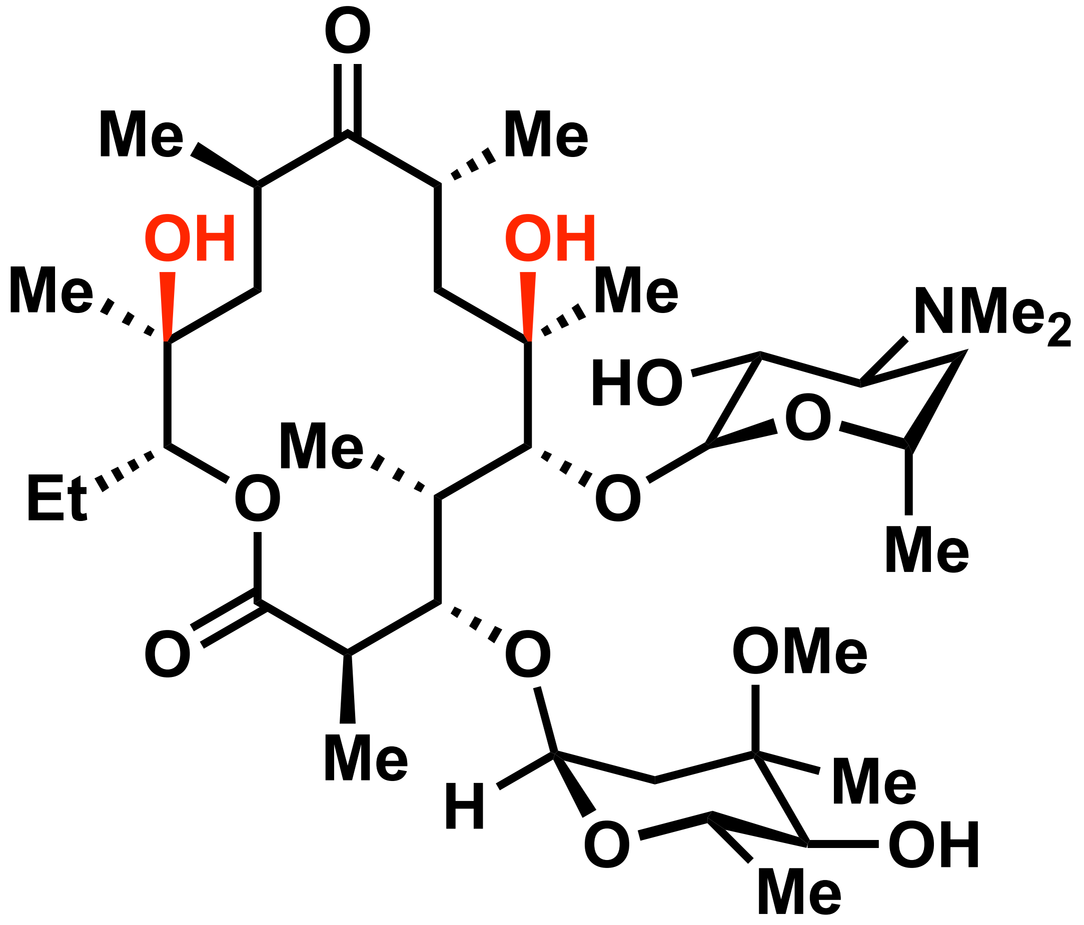
Pravachol

high cholesterol treatmeant

J. Biol. Chem. **2020**, *295*, 833.

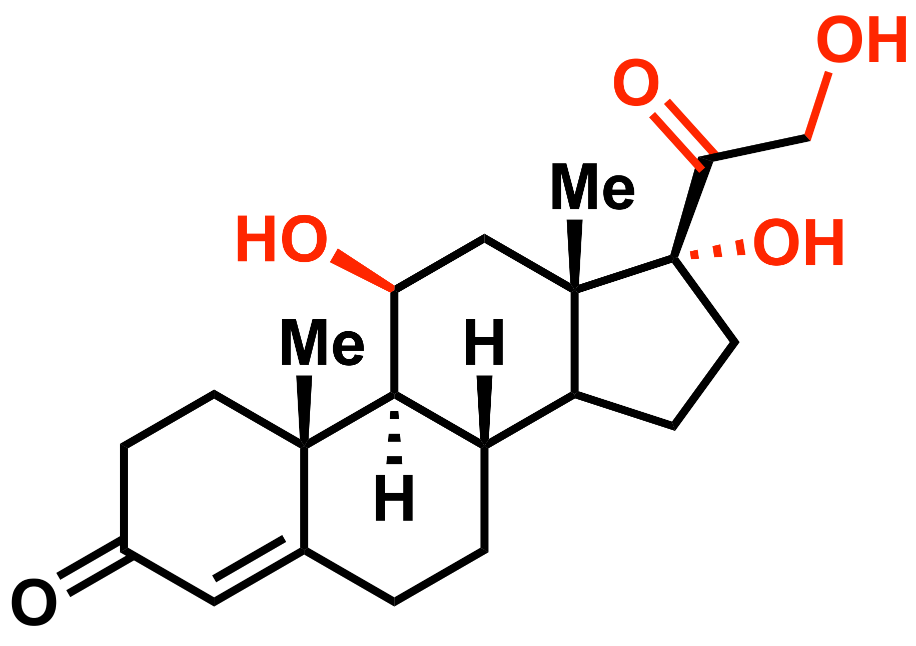
A Different Reaction Altogether





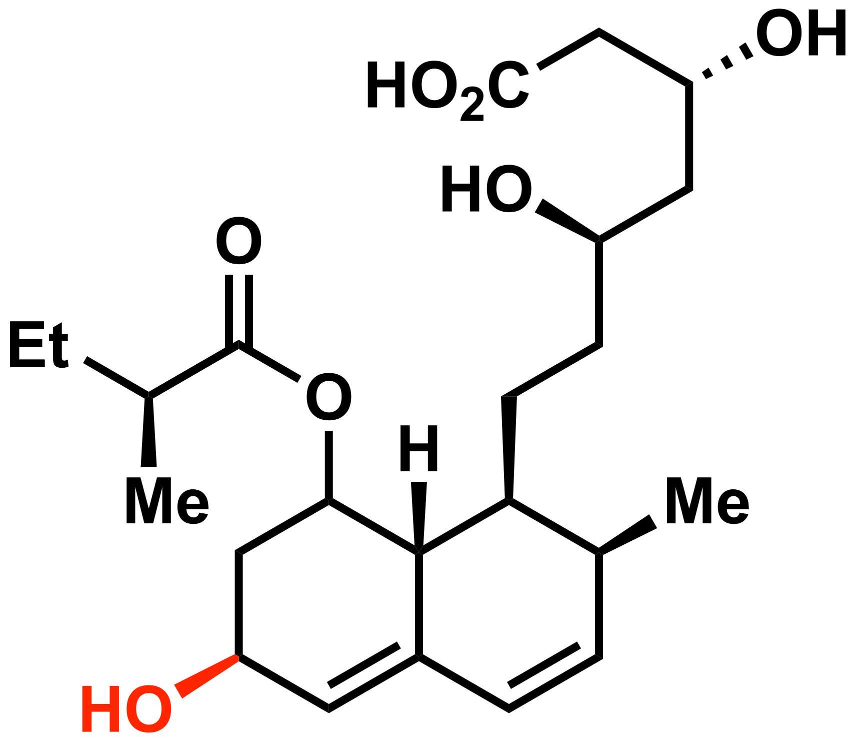
erythromycin

common anti-bacterial



hydrocortisone

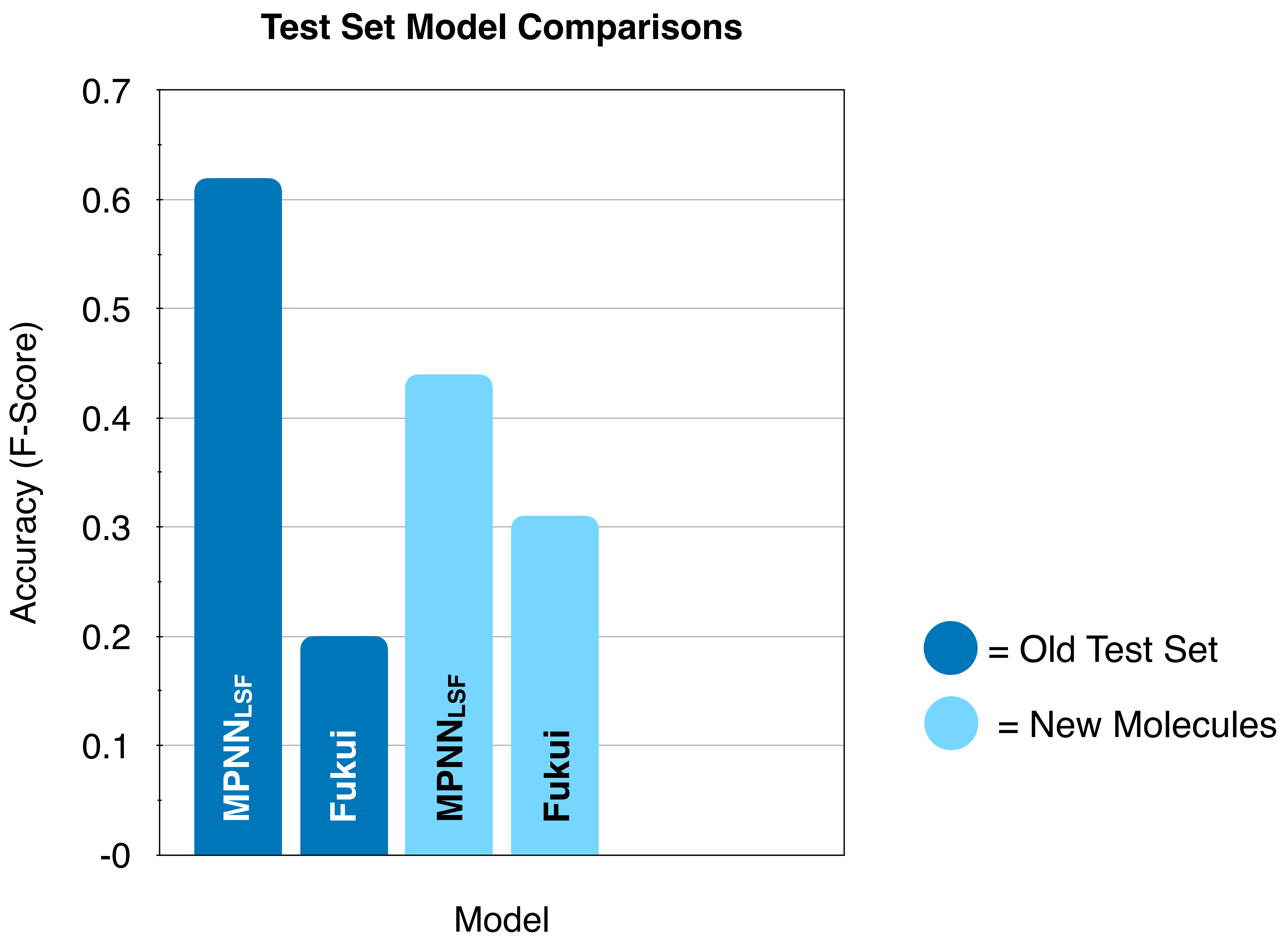
common anti-inflammatory



Pravachol

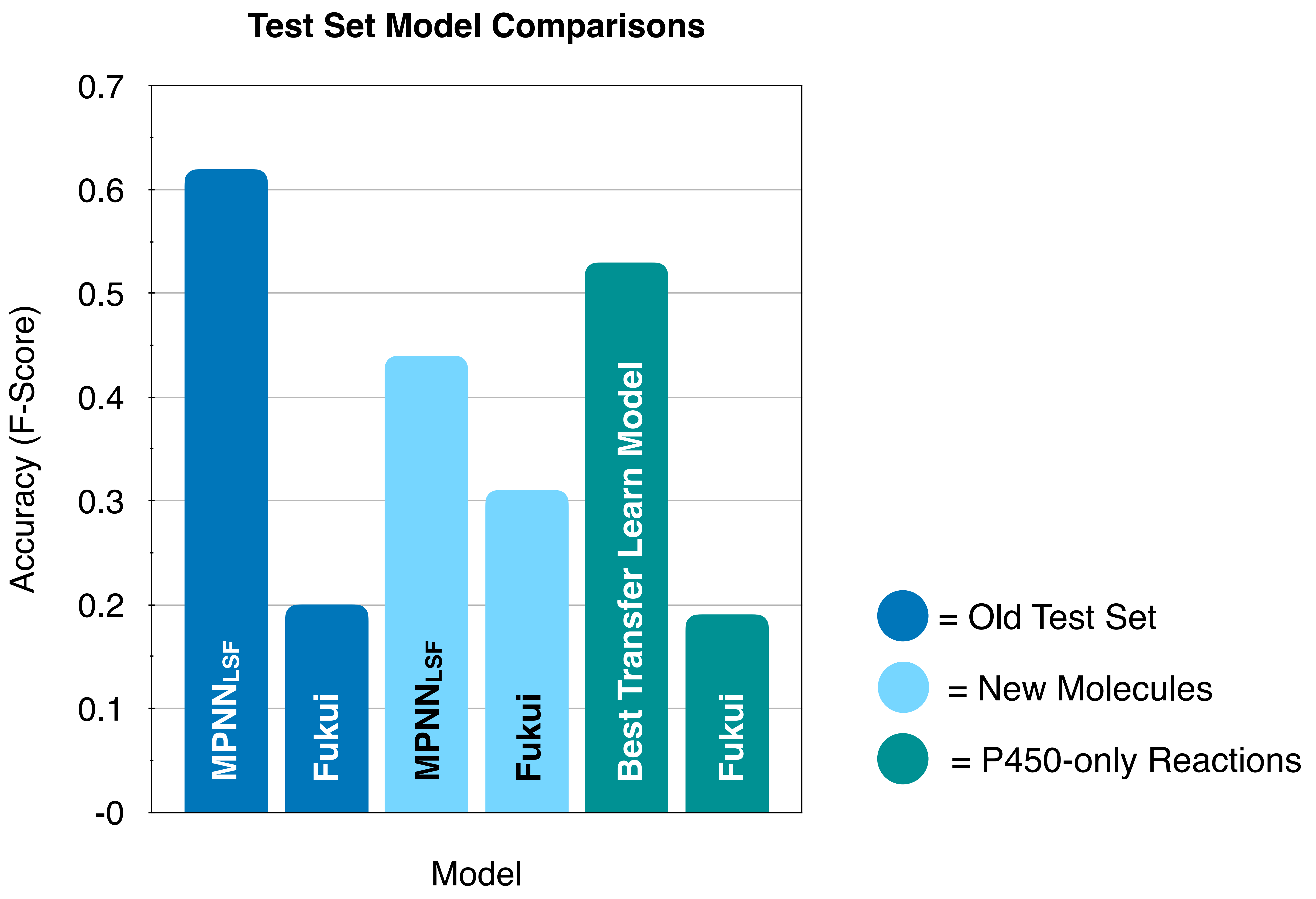
high cholesterol treatmeant

How Do We Perform on Completely Unseen Molecules?



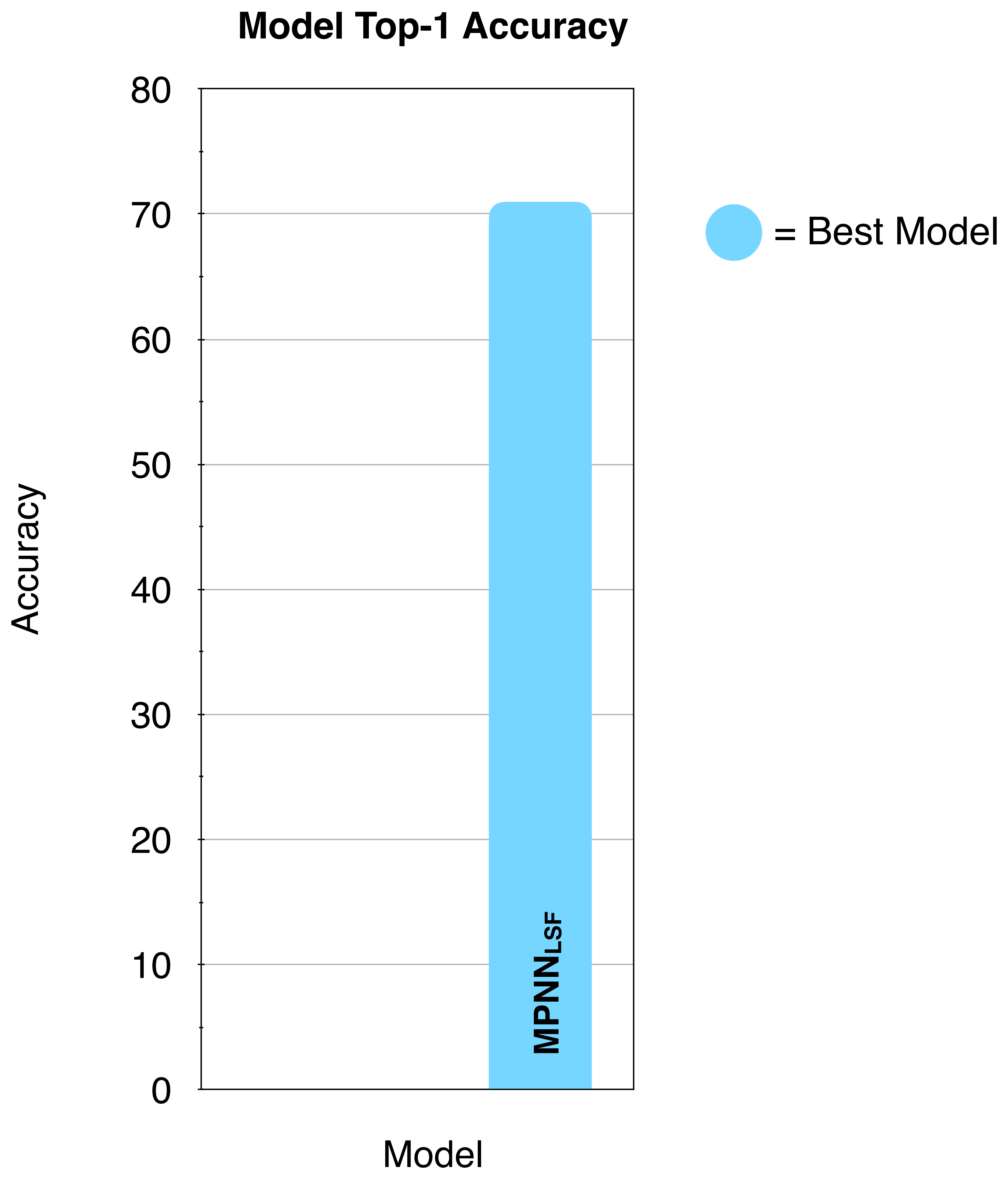
Nat. Commun. **2024**, *15*, 426.

How Do We Perform On a P450-Only Test Set?



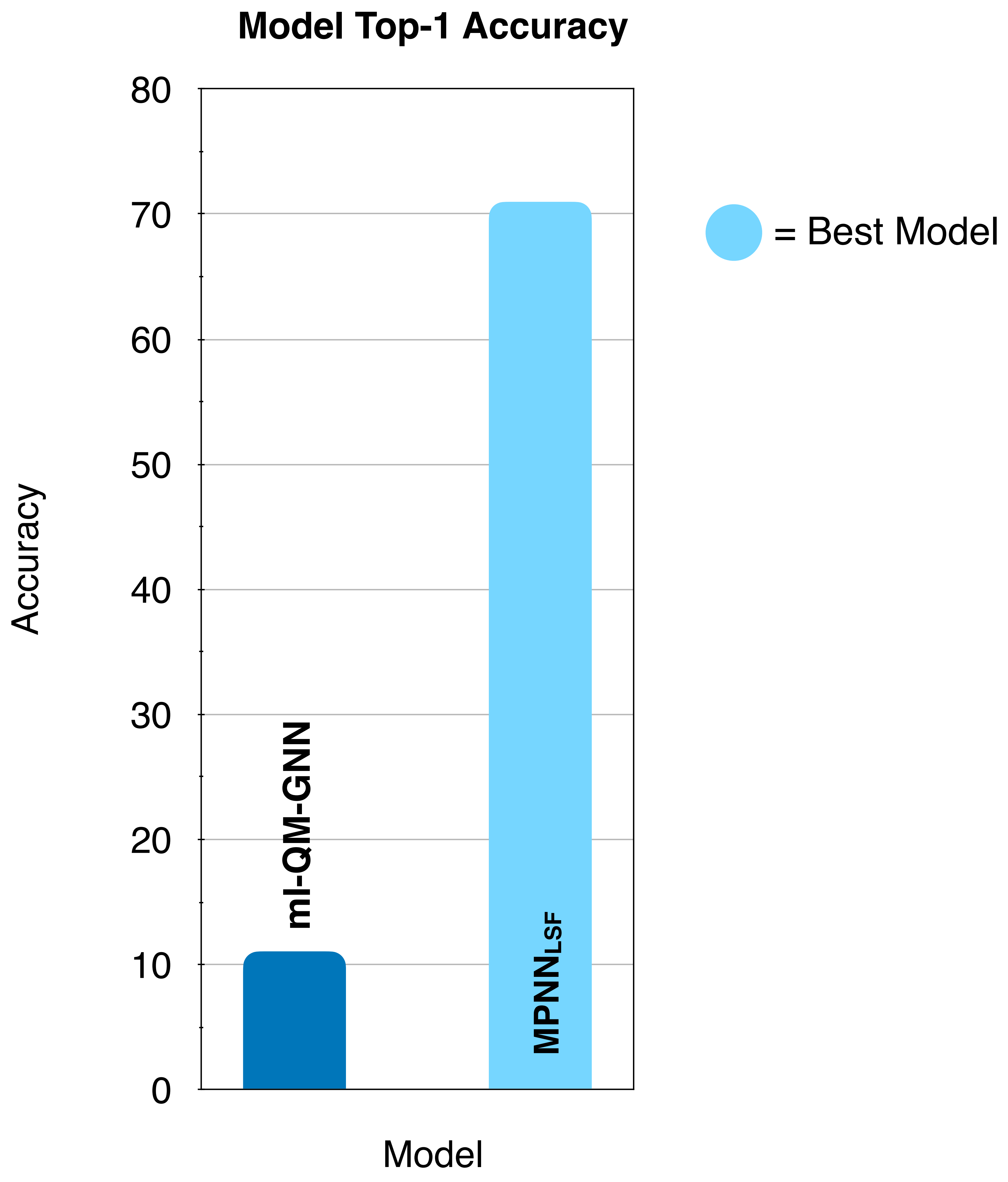
Nat. Commun. **2024**, *15*, 426.

Comparison To Other Reactivity-Based Models



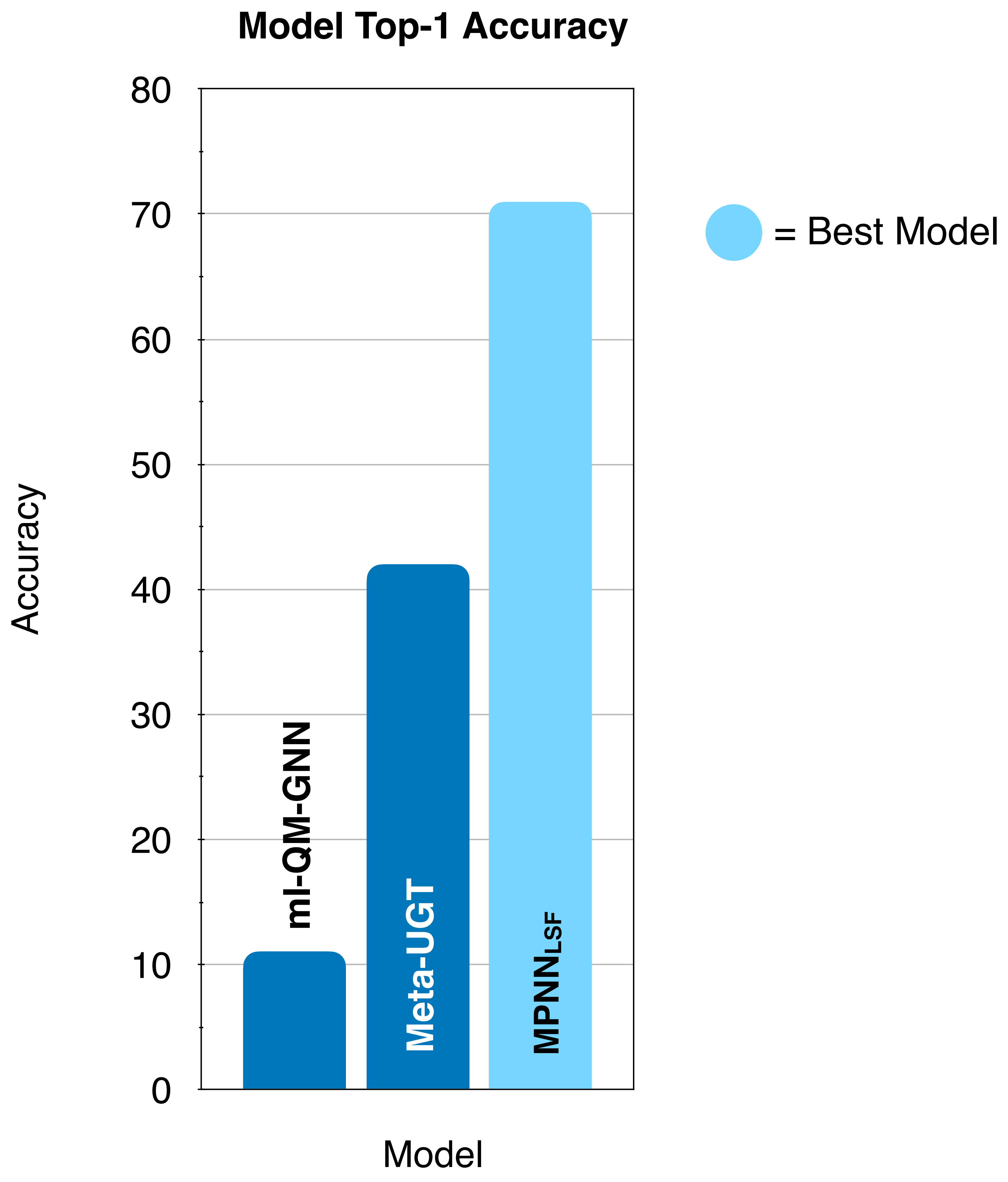
Chem. Sci. **2021**, *12*, 2198. *J Cheminform.* **2022**, *14*, 46.

Comparison To Other Reactivity-Based Models



Chem. Sci. **2021**, *12*, 2198. *J Cheminform.* **2022**, *14*, 46.

Comparison To Other Reactivity-Based Models



Chem. Sci. **2021**, *12*, 2198. *J Cheminform.* **2022**, *14*, 46.

Motivation

UNIVERSITY OF CAMBRIDGE

NMRShiftDB

CCDC

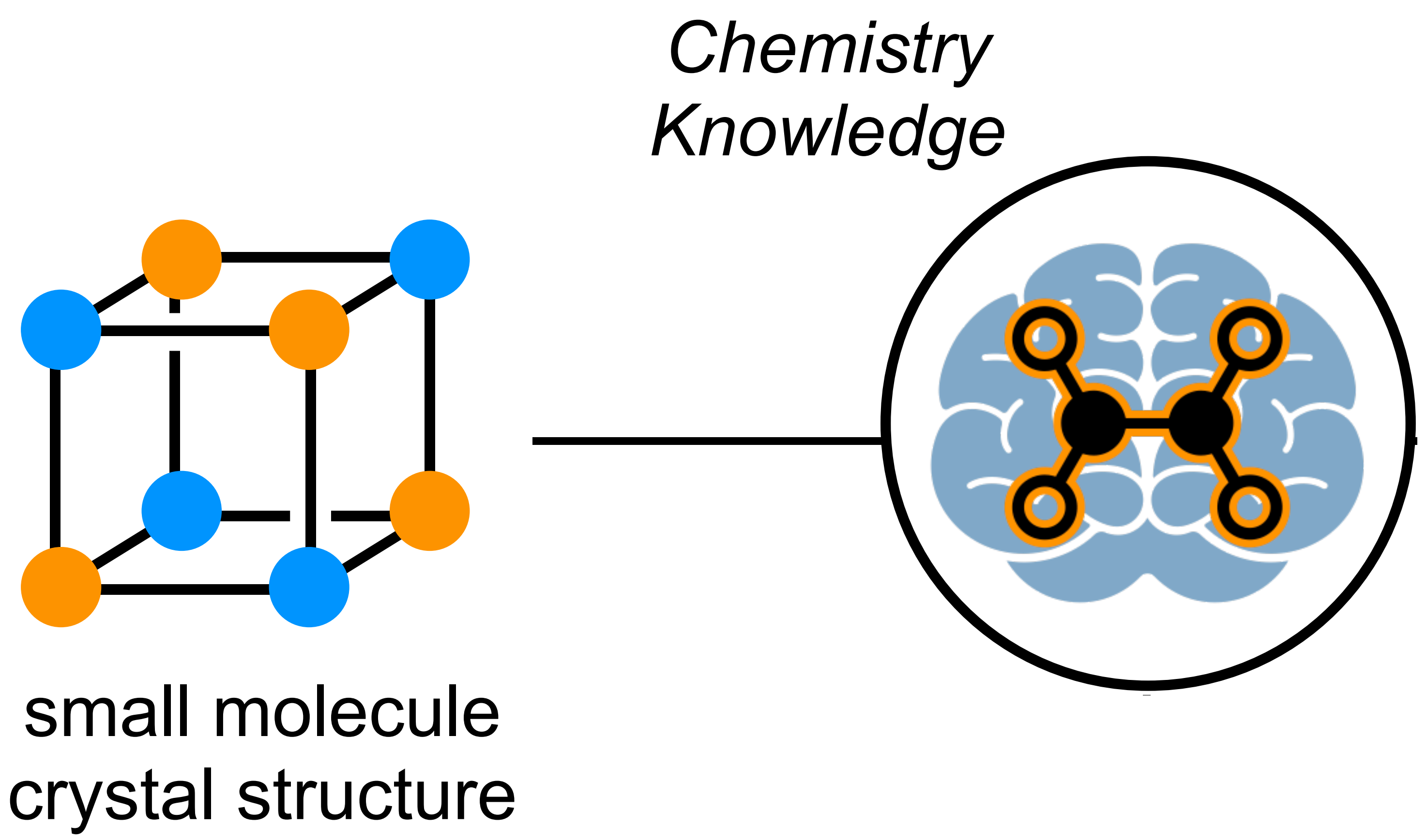
~27,000 spectra

1,000,000+ compounds

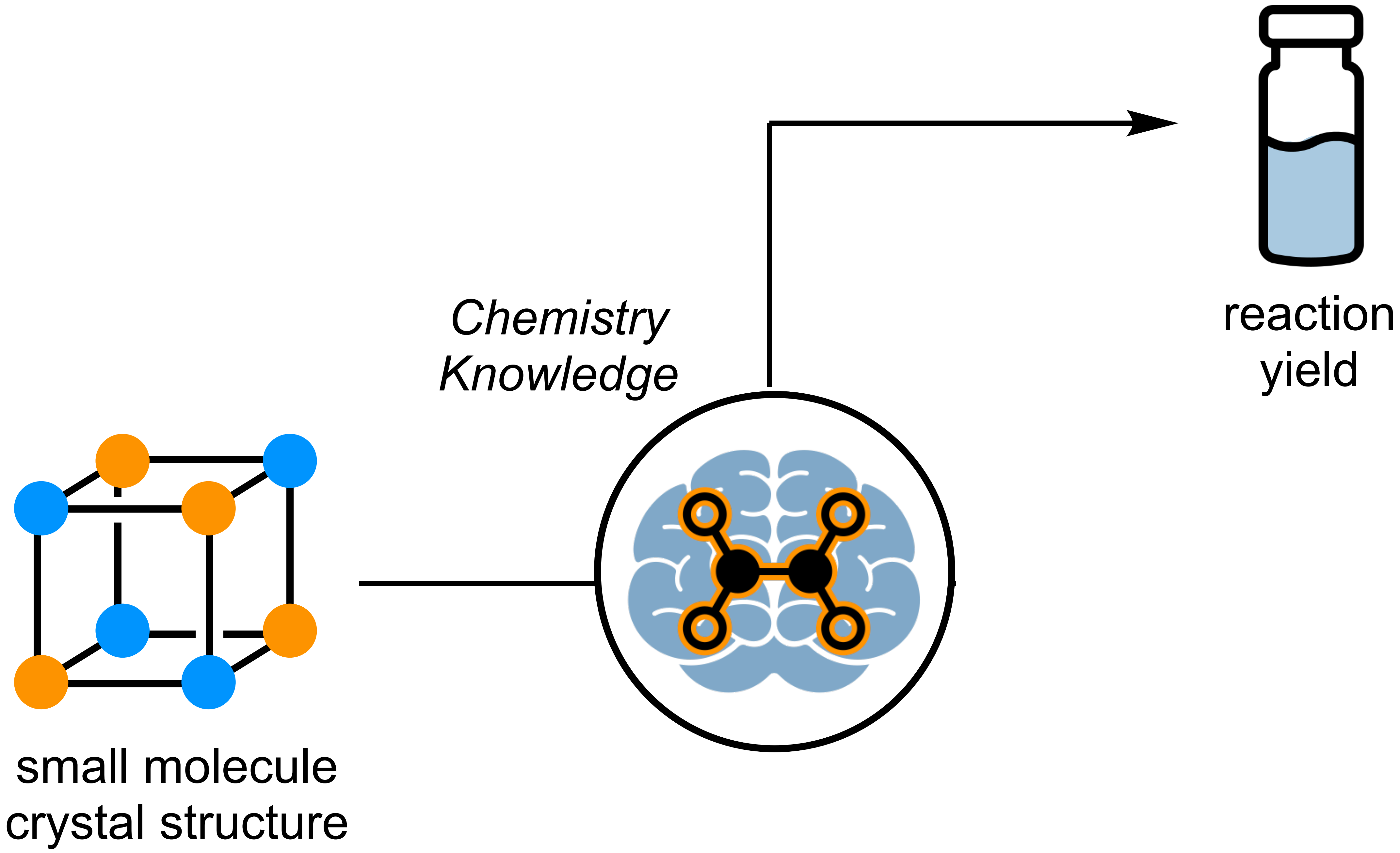
J. Chem. Inf. Comput. **2003**, *43*, 1733.

Acta Crystallogr. B: Struct. Sci. Cryst. Eng. Mater. **1979**, *35*, 2331.

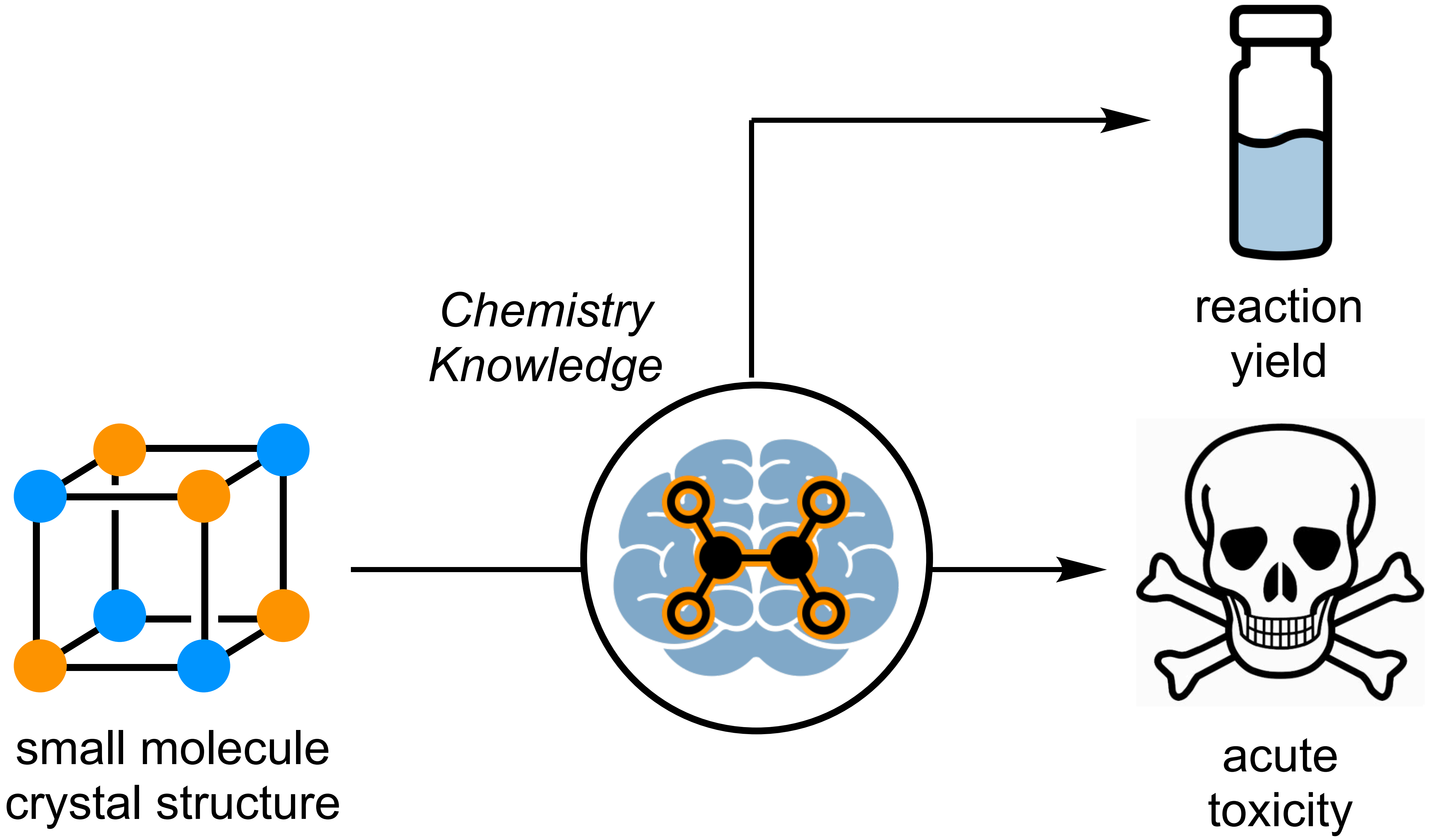
Motivation



Motivation

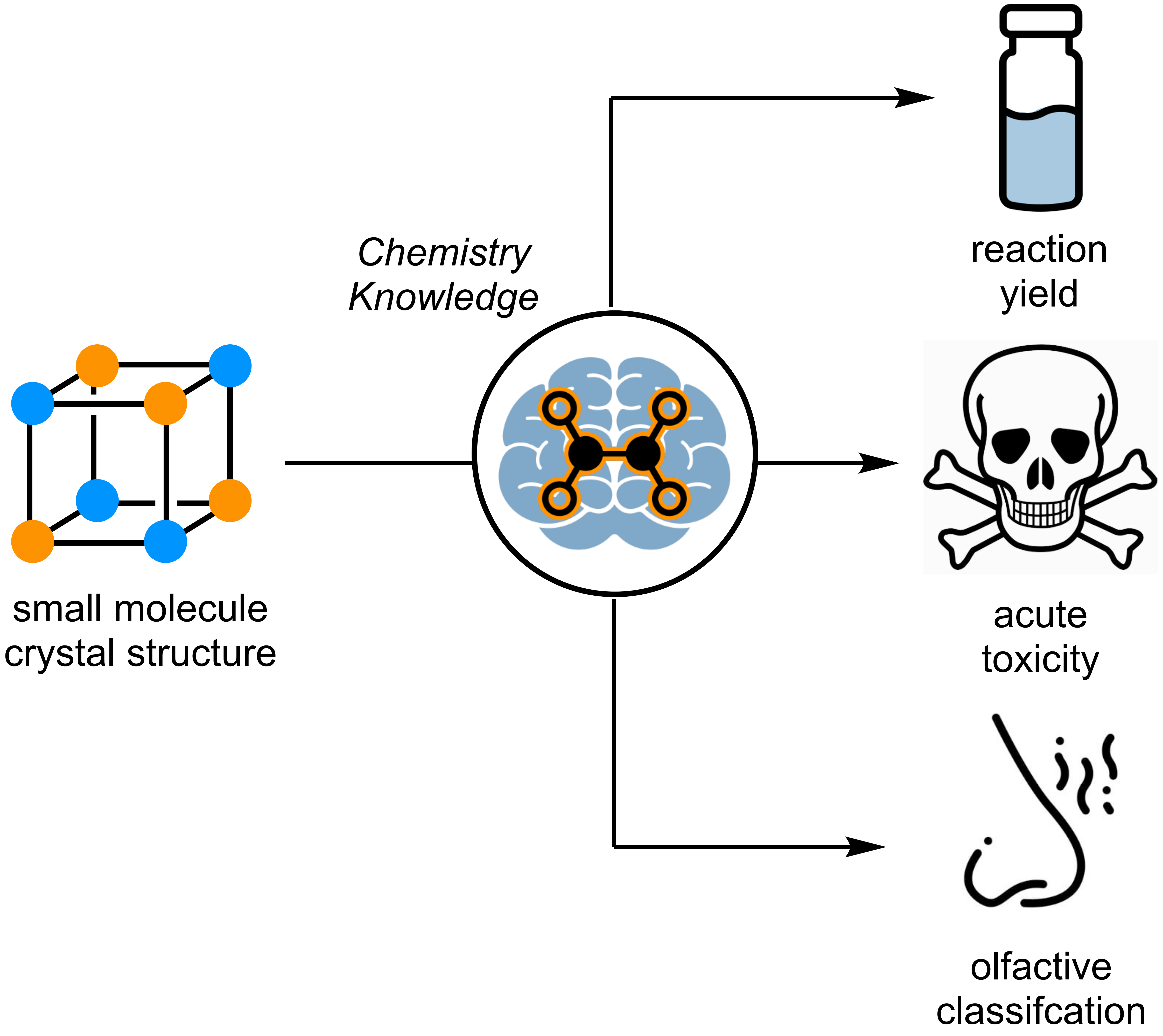

reaction
yield
Chemistry
Knowledge
small molecule
crystal structure

Motivation



Chemistry
Knowledge
reaction
yield
small molecule
crystal structure
acute
toxicity

Motivation



Chem. Sci. **2024**, *15*, 5143.

Suzuki and Buchwald-Hartwig Cross Coupling Yield Predictions

UNIVERSITY OF CAMBRIDGE

Model	**Suzuki Yield Error (MAE)** Unseen Boronic Acids	Unseen Aryl Halides	**Buchwald-Hartwig Yield Error (MAE)** Unseen Boronic Acids	Unseen Aryl Halides	Unseen Ligands	Unseen Additives
Random Forest						
Adaboost						
Yield-BERT						
GraphRXN						
Crystal-Yield						

*Increased Crystal-Yield's size to half of GraphRXN's parameters

Chem. Sci. **2024**, *15*, 5143.

UNIVERSITY OF CAMBRIDGE

Model	Suzuki Yield Error (MAE)		Buchwald-Hartwig Yield Error (MAE)			
	Unseen Boronic Acids	Unseen Aryl Halides	Unseen Boronic Acids	Unseen Aryl Halides	Unseen Ligands	Unseen Additives
Random Forest						
Adaboost						
Yield-BERT						
GraphRXN						
Crystal-Yield	**18.4 ± 0.3**	**18.5 ± 0.2**	**21.3 ± 3.3**	**13.4 ± 0.3**	**11.7 ± 2.2***	**16.2 ± 0.4**

*Increased Crystal-Yield's size to half of GraphRXN's parameters

Chem. Sci. **2024**, *15*, 5143.

Suzuki and Buchwald-Hartwig Cross Coupling Yield Predictions

	Suzuki Yield Error (MAE)		**Buchwald-Hartwig Yield Error (MAE)**			
Model	Unseen Boronic Acids	Unseen Aryl Halides	Unseen Boronic Acids	Unseen Aryl Halides	Unseen Ligands	Unseen Additives
Random Forest	19.5 ± 0.03	19.5 ± 0.03	25.2 ± 2.0	28.1 ± 4.1	28.5 ± 0.6	30.4 ± 1.5
Adaboost	21.6 ± 0.1	21.5 ± 0.1	24.7 ± 2.6	25.5 ± 2.9	27.9 ± 0.7	26.7 ± 0.5
Yield-BERT						
GraphRXN						
Crystal-Yield	**18.4 ± 0.3**	**18.5 ± 0.2**	**21.3 ± 3.3**	**13.4 ± 0.3**	**11.7 ± 2.2***	**16.2 ± 0.4**

*Increased Crystal-Yield's size to half of GraphRXN's parameters

Chem. Sci. **2024**, *15*, 5143.

Suzuki and Buchwald-Hartwig Cross Coupling Yield Predictions

UNIVERSITY OF CAMBRIDGE

Model	Suzuki Yield Error (MAE)		Buchwald-Hartwig Yield Error (MAE)			
	Unseen Boronic Acids	Unseen Aryl Halides	Unseen Boronic Acids	Unseen Aryl Halides	Unseen Ligands	Unseen Additives
Random Forest	19.5 ± 0.03	19.5 ± 0.03	25.2 ± 2.0	28.1 ± 4.1	28.5 ± 0.6	30.4 ± 1.5
Adaboost	21.6 ± 0.1	21.5 ± 0.1	24.7 ± 2.6	25.5 ± 2.9	27.9 ± 0.7	26.7 ± 0.5
Yield-BERT	21.9 ± 0.06	22.0 ± 0.03	24.7 ± 2.1	24.3 ± 1.6	24.3 ± 1.4	24.1 ± 0.7
GraphRXN	40.0 ± 3.0	37.8 ± 2.7	25.2 ± 7.0	17.9 ± 4.6	13.8 ± 1.7	17.5 ± 1.8
Crystal-Yield	**18.4 ± 0.3**	**18.5 ± 0.2**	**21.3 ± 3.3**	**13.4 ± 0.3**	**11.7 ± 2.2***	**16.2 ± 0.4**

*Increased Crystal-Yield's size to half of GraphRXN's parameters

Chem. Sci. **2024**, *15*, 5143.

Mach. Learn.: Sci. Technol. **2021**, *2*, 015016. *J. Chemoinformatics* **2023**, *15*, 72.

LD50 Toxicity Predictions

Model	Pharmaceuticals (MAE)
Random Forest	
Gaussian Process	
Adaboost	
Oloren Chem Engine	
Crystal-Tox	**0.52 ± 0.007**

Chem. Sci. **2024**, *15*, 5143.

LD50 Toxicity Predictions

Model	Pharmaceuticals (MAE)
Random Forest	0.62 ± 0.002
Gaussian Process	0.73 ± 0.002
Adaboost	0.71 ± 0.002
Oloren Chem Engine	
Crystal-Tox	**0.52 ± 0.007**

Chem. Sci. **2024**, *15*, 5143.

LD50 Toxicity Predictions

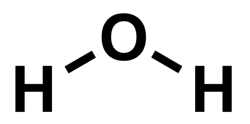
Model	Pharmaceuticals (MAE)
Random Forest	0.62 ± 0.002
Gaussian Process	0.73 ± 0.002
Adaboost	0.71 ± 0.002
Oloren Chem Engine	0.55 ± 0.009
Crystal-Tox	**0.52 ± 0.007**

Chem. Sci. **2024**, *15*, 5143.

ChemRxiv Preprint **2022**, DOI: 10.26434/chemrxiv-2022-zz776.

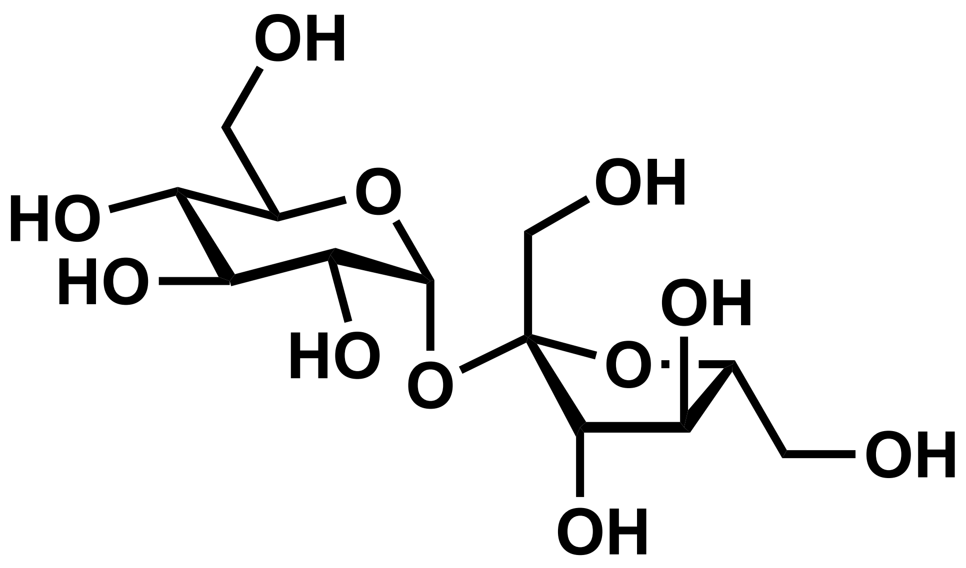
New Molecules for Testing

Benign Molecules



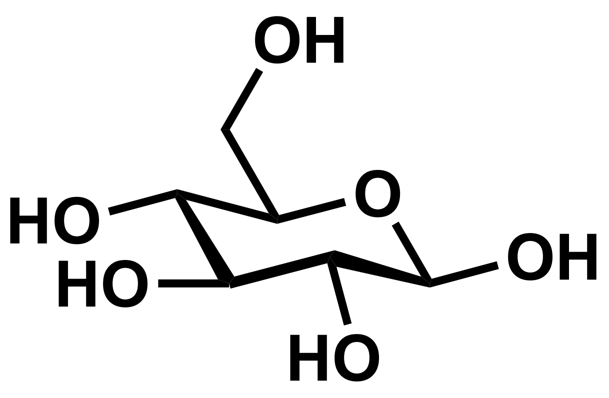
water (**1**)

LD50 90,000



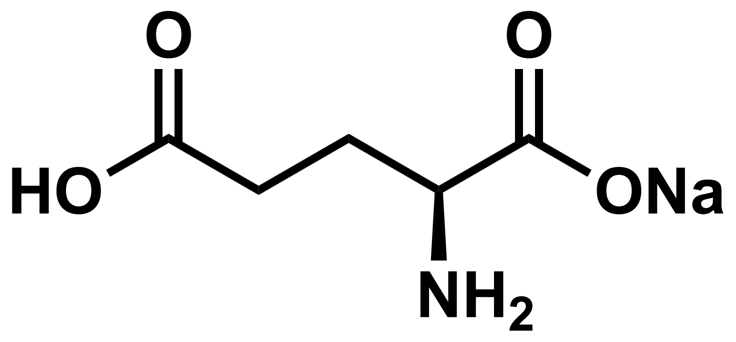
sucrose (**2**)

LD50 29,700



glucose (**3**)

LD50 = 25,800



monosodium glutamate (**4**)

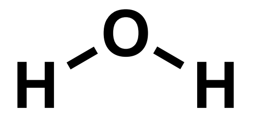
LD50 = 16,600

Chem. Sci. **2024**, *15*, 5143.

New Molecules for Testing

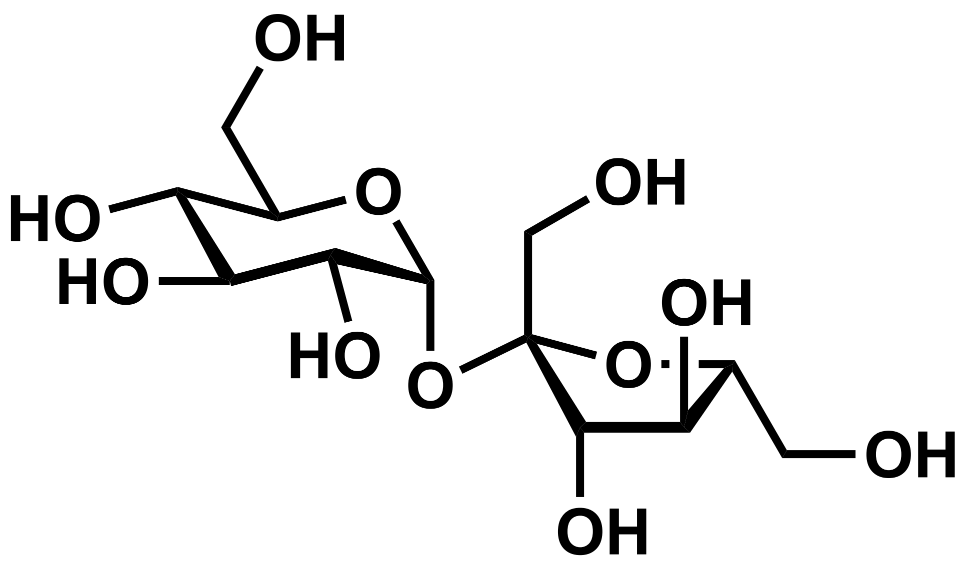
UNIVERSITY OF CAMBRIDGE

Benign Molecules | ***Natural Toxins***



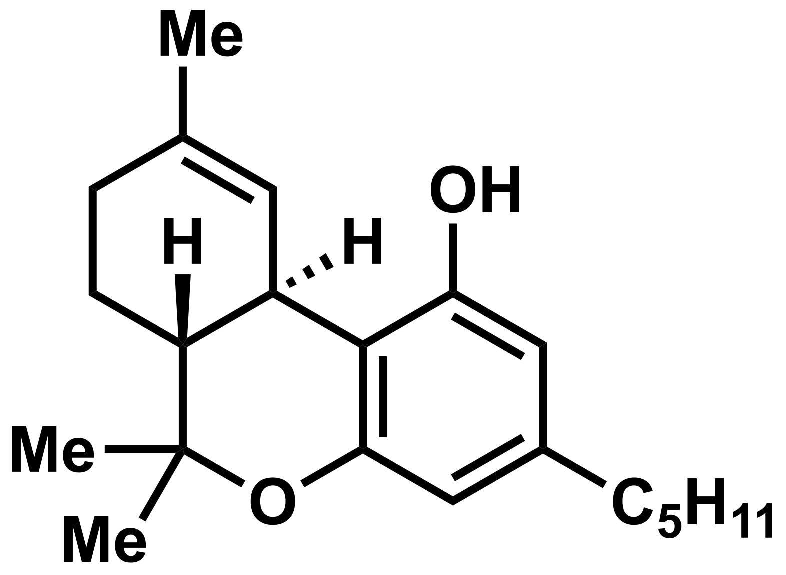
water (**1**)

LD50 90,000



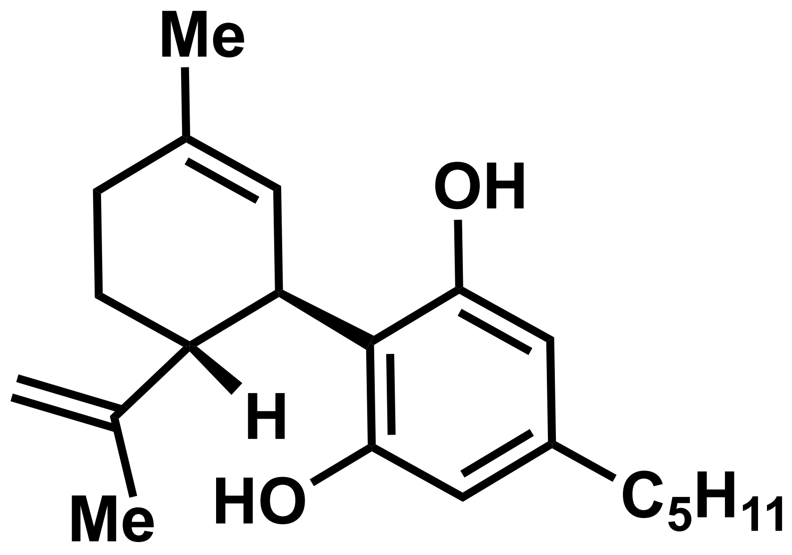
sucrose (**2**)

LD50 29,700



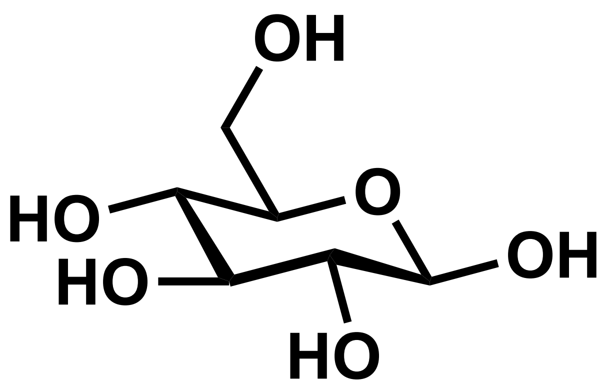
THC (**5**)

LD50 = 1,270



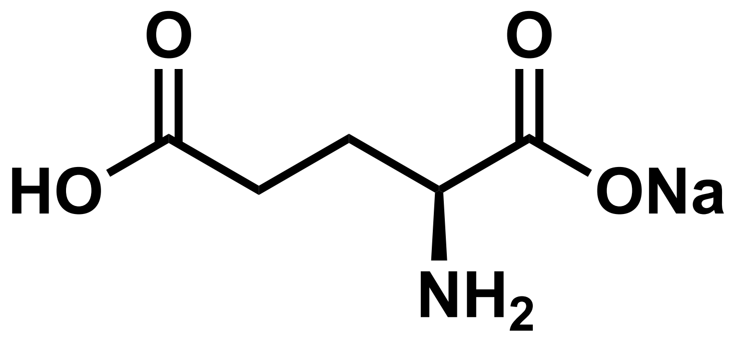
CBD (**6**)

LD50 = 980



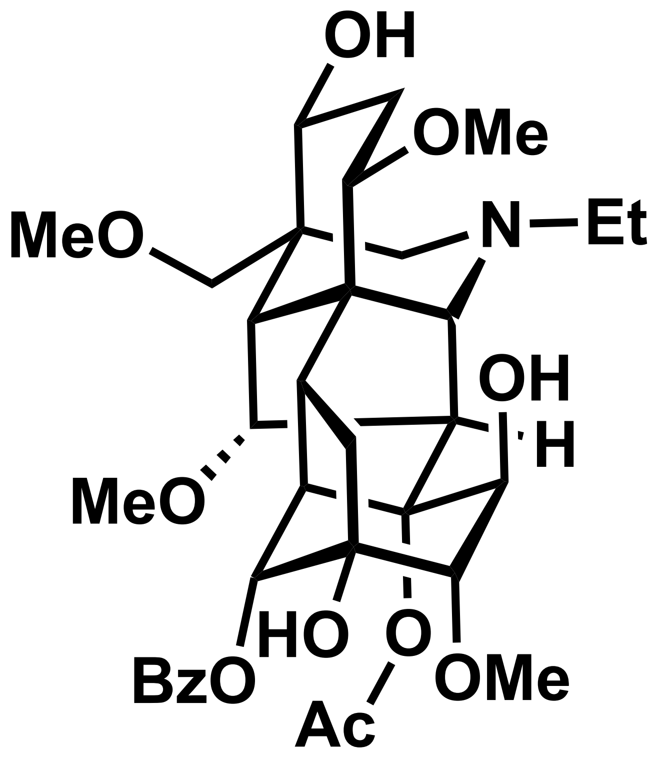
glucose (**3**)

LD50 = 25,800



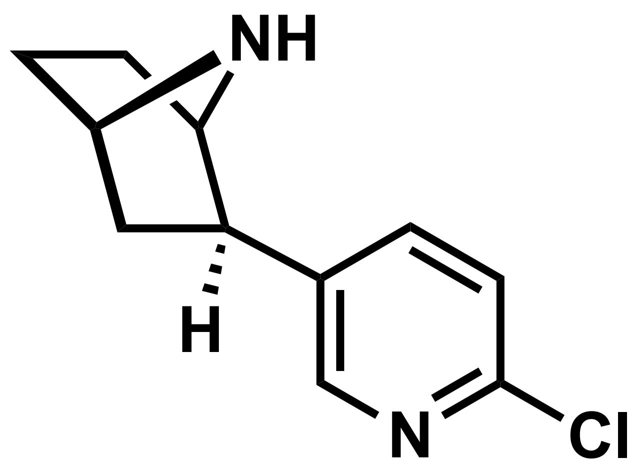
monosodium glutamate (**4**)

LD50 = 16,600



aconitine (**7**)

LD50 = 0.08



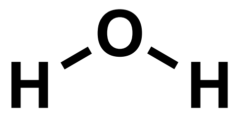
epibatidine (**8**)

LD50 = 0.0077

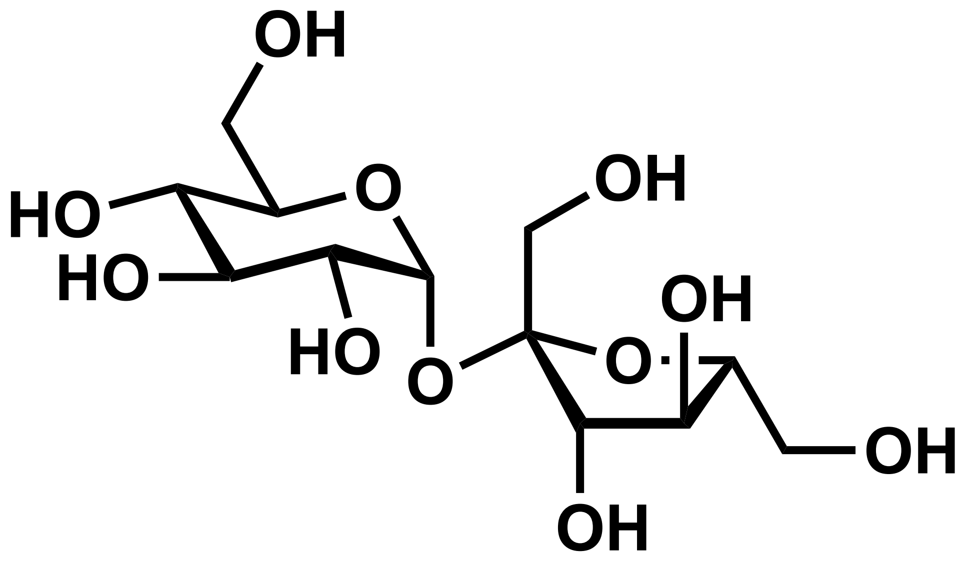
Chem. Sci. **2024**, *15*, 5143.

New Molecules for Testing

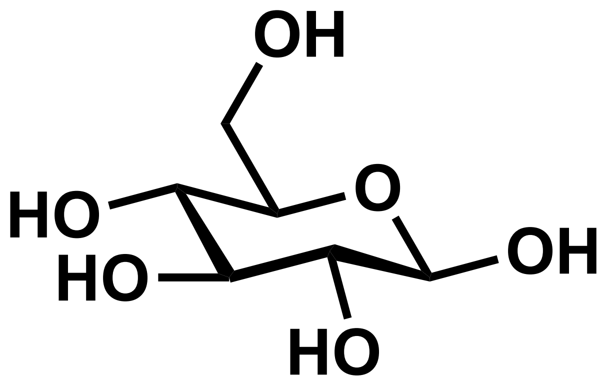
Benign Molecules



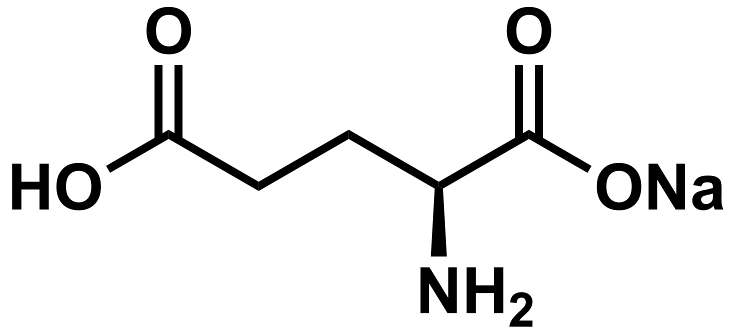
water (**1**)
LD50 90,000



sucrose (**2**)
LD50 29,700

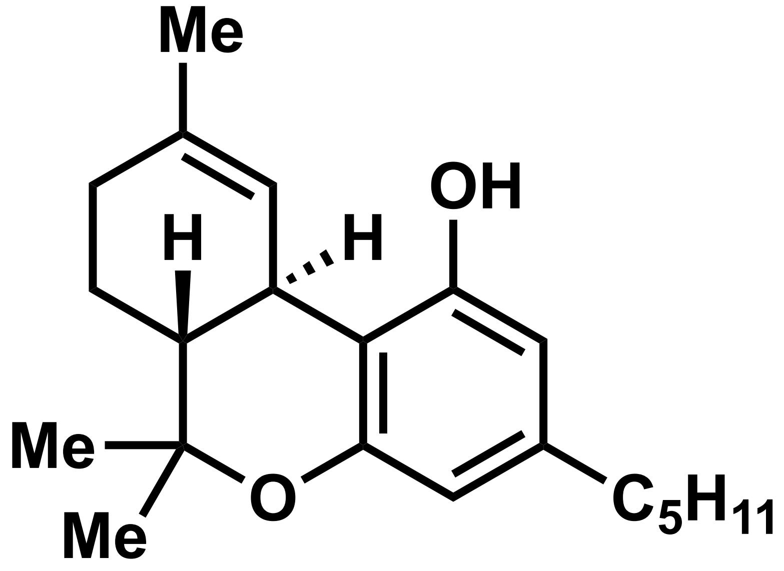


glucose (**3**)
LD50 = 25,800

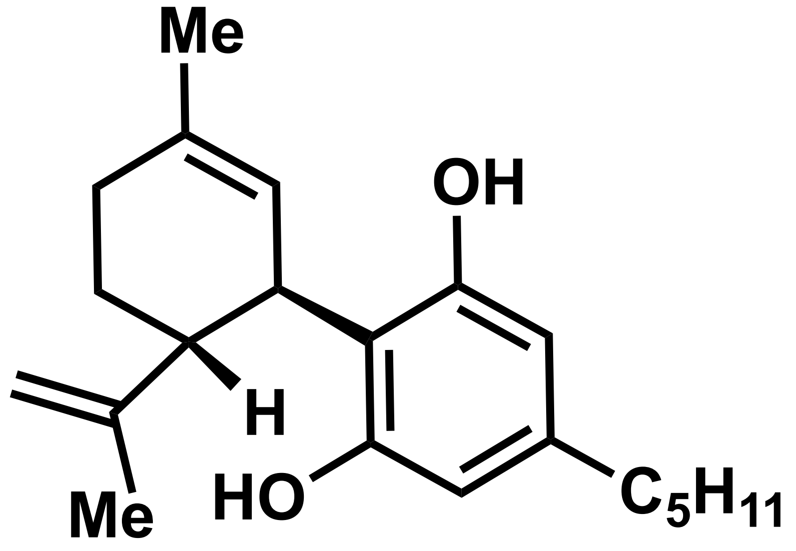


monosodium glutamate (**4**)
LD50 = 16,600

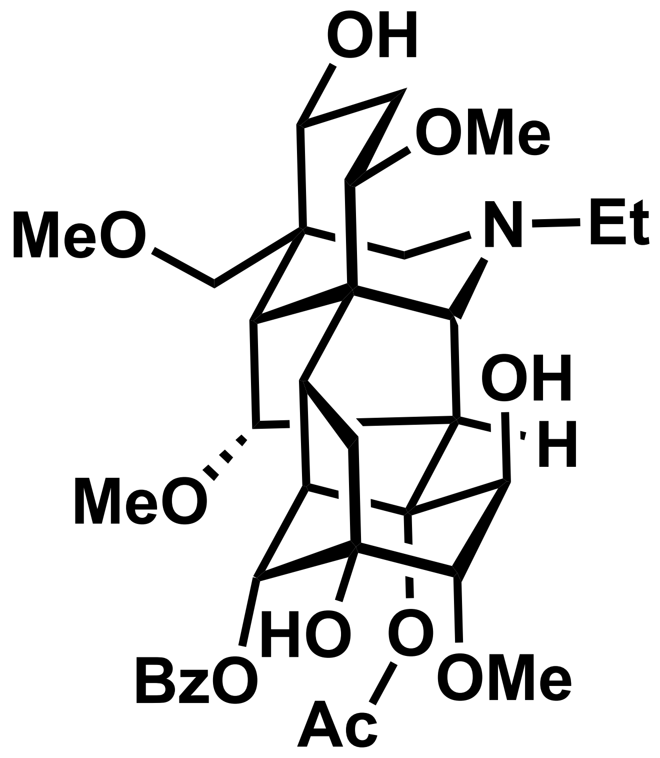
Natural Toxins



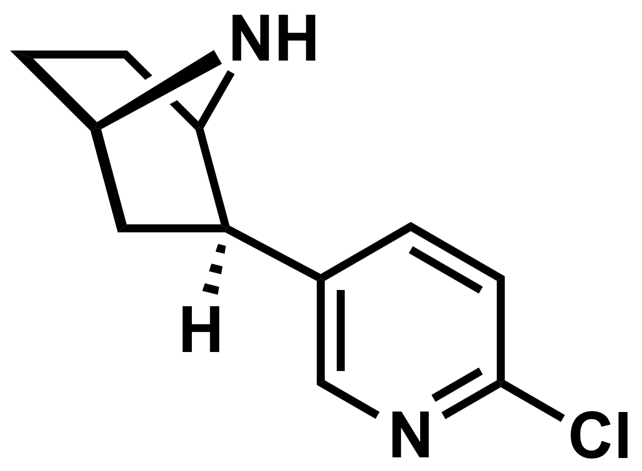
THC (**5**)
LD50 = 1,270



CBD (**6**)
LD50 = 980

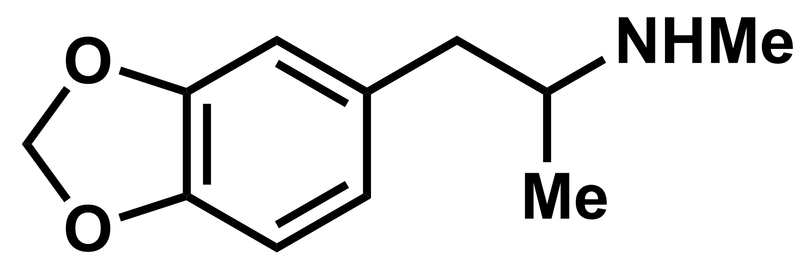


aconitine (**7**)
LD50 = 0.08

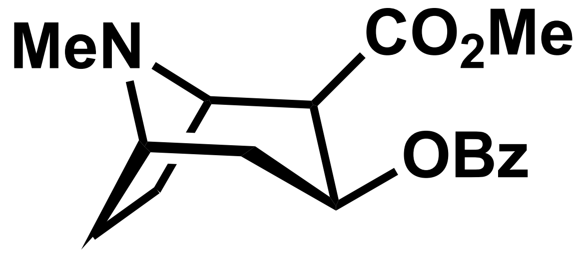


epibatidine (**8**)
LD50 = 0.0077

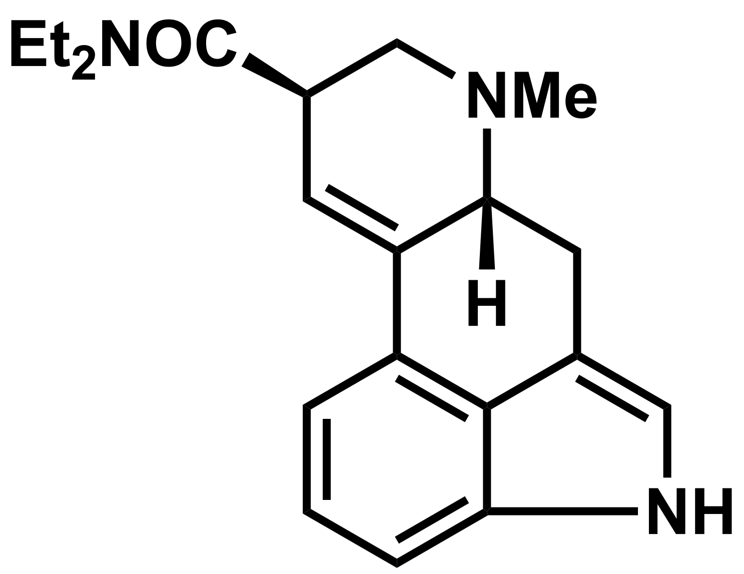
Illicit Substances



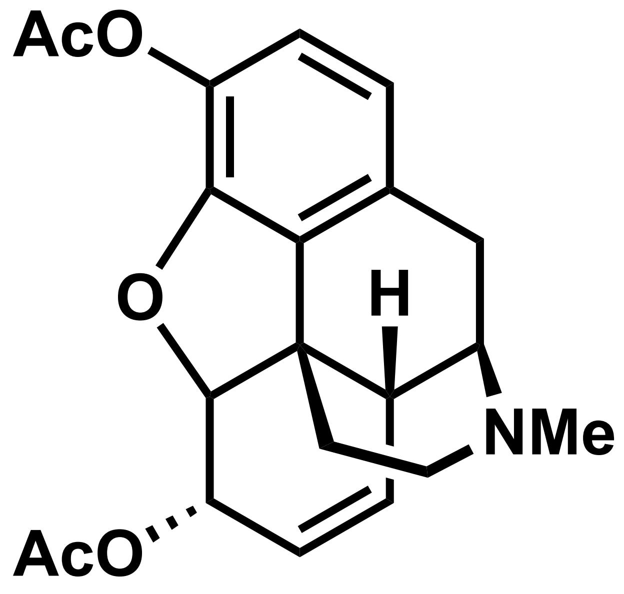
MDMA (**9**)
LD50 = 160.



cocaine (**10**)
LD50 = 96.0



LSD (**11**)
LD50 = 16.5



heroin (**12**)
LD50 = 21.8

Chem. Sci. **2024**, *15*, 5143.

LD50 Toxicity Predictions

Model	Pharmaceuticals (MAE)
Random Forest	0.62 ± 0.002
Gaussian Process	0.73 ± 0.002
Adaboost	0.71 ± 0.002
Oloren Chem Engine	0.55 ± 0.009
Crystal-Tox	**0.52 ± 0.007**

Chem. Sci. **2024**, *15*, 5143.

ChemRxiv Preprint **2022**, DOI: 10.26434/chemrxiv-2022-zz776.

LD50 Toxicity Predictions

UNIVERSITY OF CAMBRIDGE

Model	Pharmaceuticals (MAE)	Non-Pharmaceuticals (MAE)
Random Forest	0.62 ± 0.002	1.59 ± 0.02
Gaussian Process	0.73 ± 0.002	1.86 ± 0.002
Adaboost	0.71 ± 0.002	1.77 ± 0.002
Oloren Chem Engine	0.55 ± 0.009	1.48 ± 0.006
Crystal-Tox	**0.52 ± 0.007**	**1.38 ± 0.02**

Chem. Sci. **2024**, *15*, 5143.

ChemRxiv Preprint **2022**, DOI: 10.26434/chemrxiv-2022-zz776.

Olfactive Classification

Chiral & Non-Chiral

Model	Macro F-Score	Weighted F-Score
Random Forest	0.19 ± 0.1	0.32 ± 0.009
K-Nearest Neighbors	0.20 ± 0.002	0.33 ± 0.002
Crystal-Olfaction	**0.62 ± 0.004**	**0.92 ± 0.002**

Chem. Sci. **2024**, *15*, 5143.

Olfactive Classification

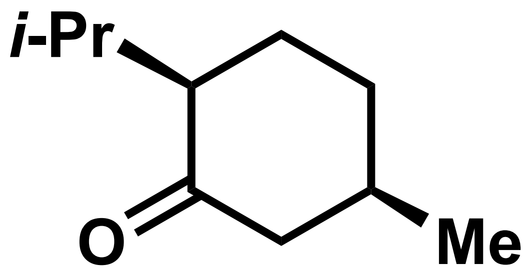
UNIVERSITY OF CAMBRIDGE

Model	Chiral & Non-Chiral: Macro F-Score	Chiral & Non-Chiral: Weighted F-Score	Enantiomer Differentiation: Macro F-Score	Enantiomer Differentiation: Weighted F-Score
Random Forest	0.19 ± 0.1	0.32 ± 0.009	0.069 ± 0.002	0.31 ± 0.003
K-Nearest Neighbors	0.20 ± 0.002	0.33 ± 0.002	0.31 ± 0.0002	0.20 ± 0.001
Crystal-Olfaction	**0.62 ± 0.004**	**0.92 ± 0.002**	**0.58 ± 0.003**	**0.93 ± 0.002**

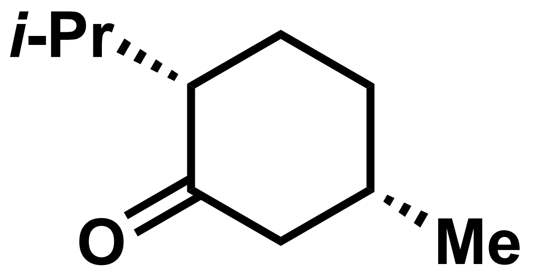
Chem. Sci. **2024**, *15*, 5143.

Olfactive Classification

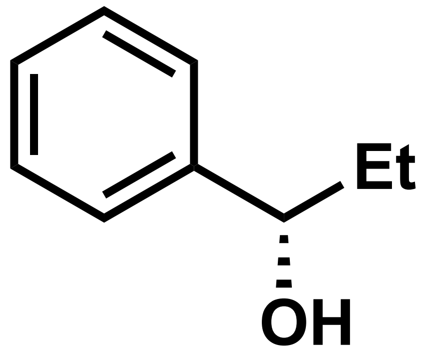
Identical Olfactive Profiles



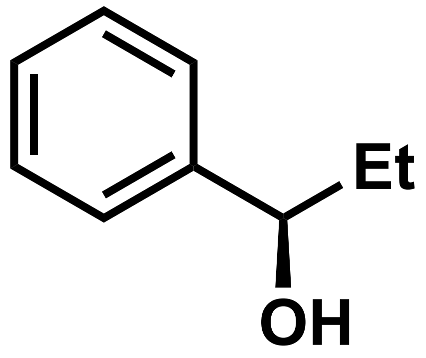
(*R*)-isomenthone (**13**)



(*S*)-isomenthone (**14**)



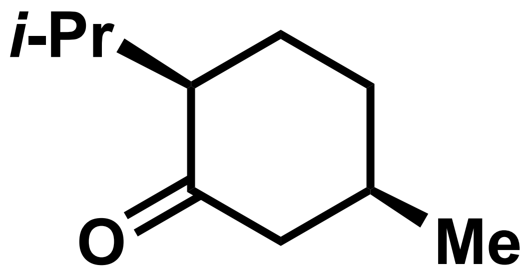
15



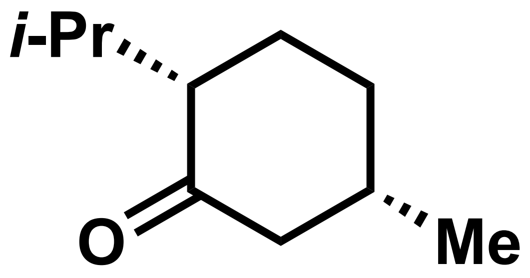
16

Chem. Sci. **2024**, *15*, 5143.

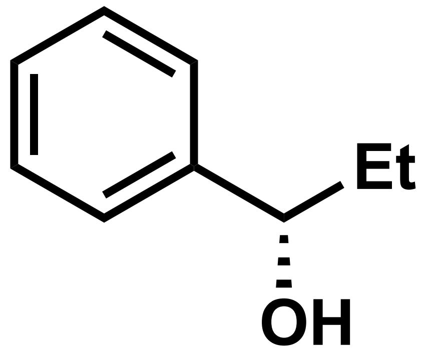
Identical Olfactive Profiles



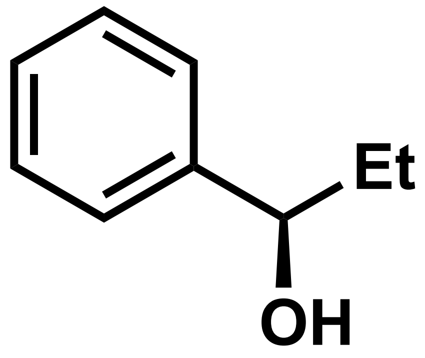
(*R*)-isomenthone (**13**)



(*S*)-isomenthone (**14**)

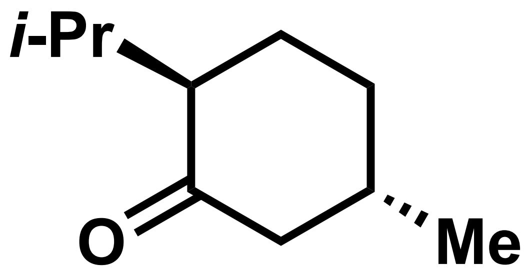


15

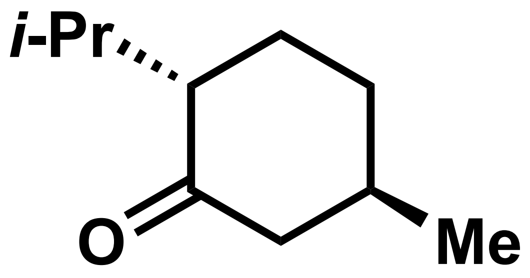


16

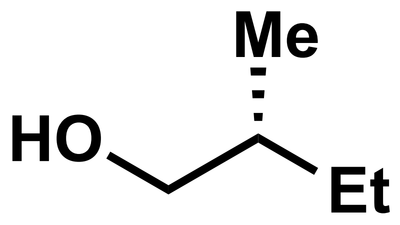
Dissimilar Olfactive Profiles



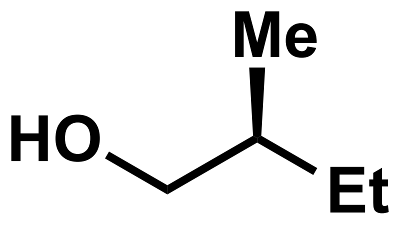
(*R*)-menthone (**17**)



(*S*)-menthone (**18**)



19



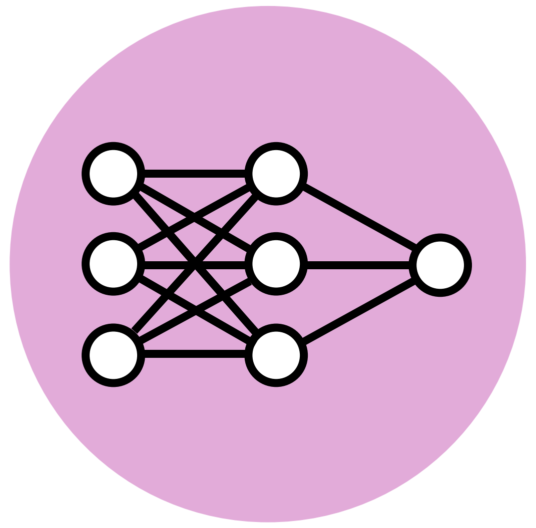
20

UNIVERSITY OF CAMBRIDGE

Part II:

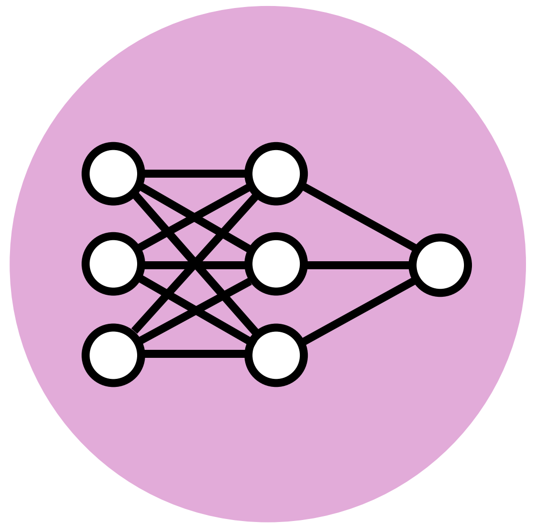
Hidden Chemical Insights from Lightweight Machine Learning

Why Lightweight ML?



Deep ML

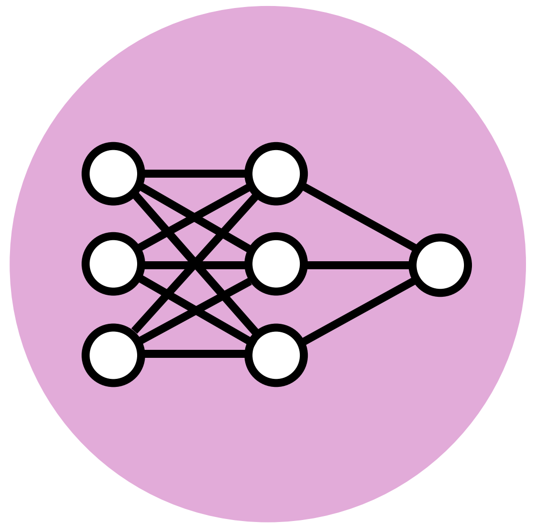
Why Lightweight ML?



Deep ML

Can get accurate predictions

Why Lightweight ML?

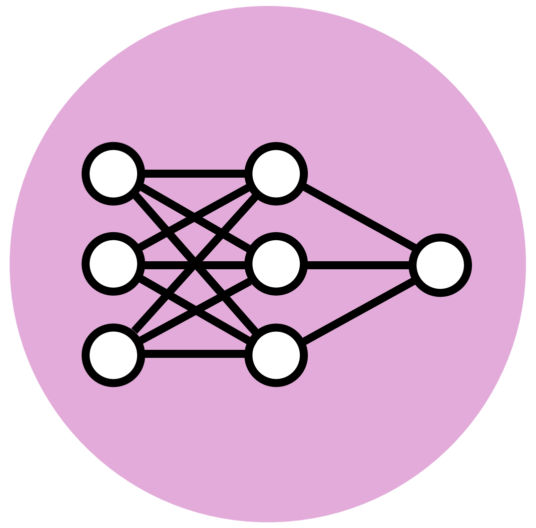


Deep ML

Can get accurate predictions

High computational resources

Why Lightweight ML?



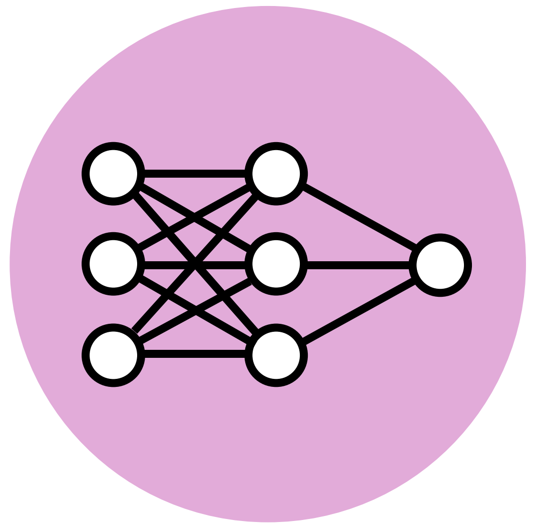
Deep ML

Can get accurate predictions

High computational resources

Careful optimization of learning architecture

Why Lightweight ML?



Deep ML

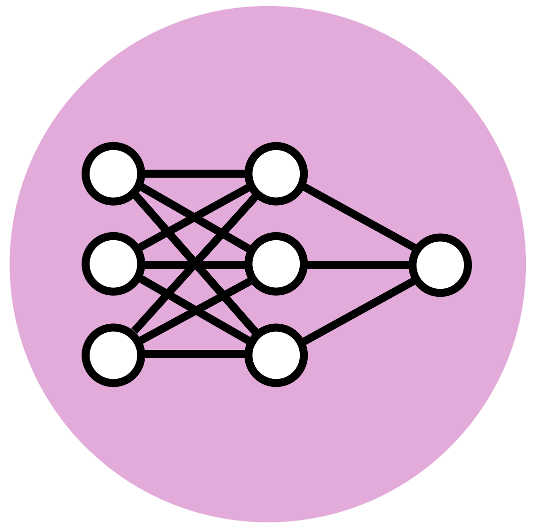
Can get accurate predictions

High computational resources

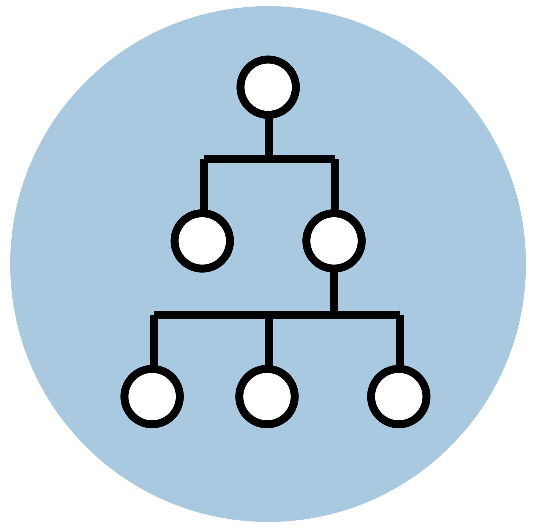
Careful optimization of learning architecture

High data requirement

Why Lightweight ML?



Deep ML



Lightweight ML

Can get accurate predictions

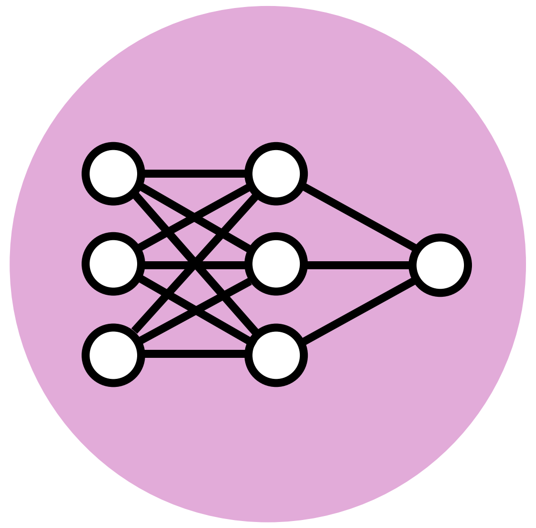
High computational resources

Careful optimization of learning architecture

High data requirement

Why Lightweight ML?

UNIVERSITY OF CAMBRIDGE



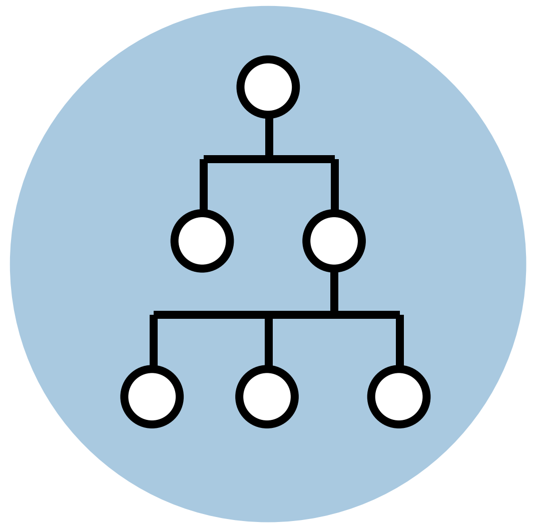
Deep ML

Can get accurate predictions

High computational resources

Careful optimization of learning architecture

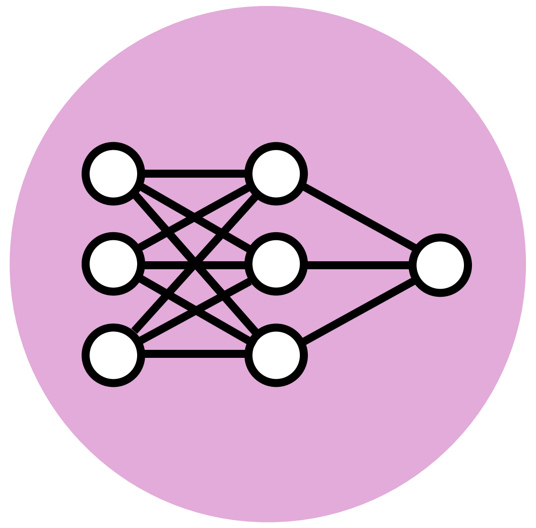
High data requirement



Lightweight ML

Run on your laptop

Why Lightweight ML?



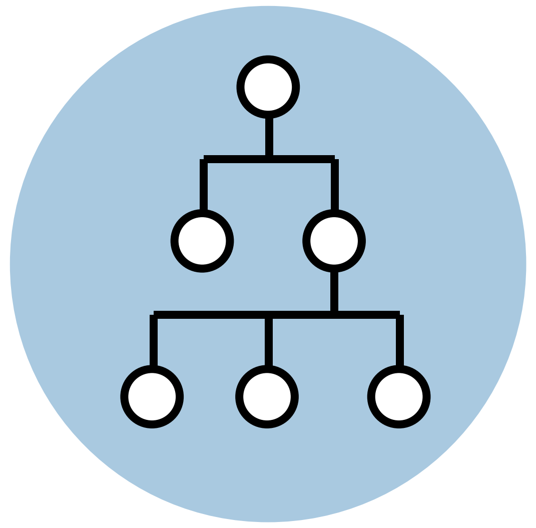
Deep ML

Can get accurate predictions

High computational resources

Careful optimization of learning architecture

High data requirement

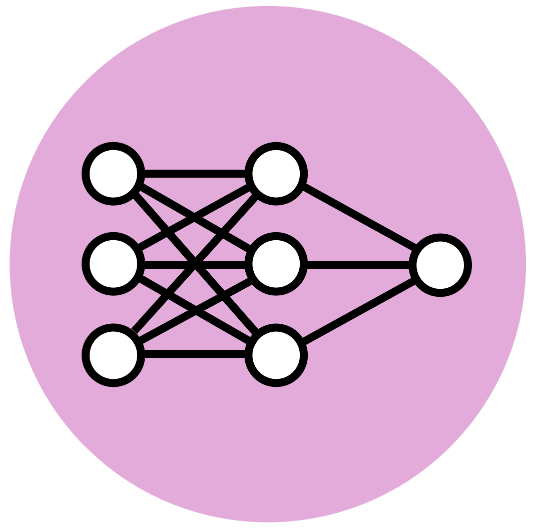


Lightweight ML

Run on your laptop

Ready to use systems

Why Lightweight ML?



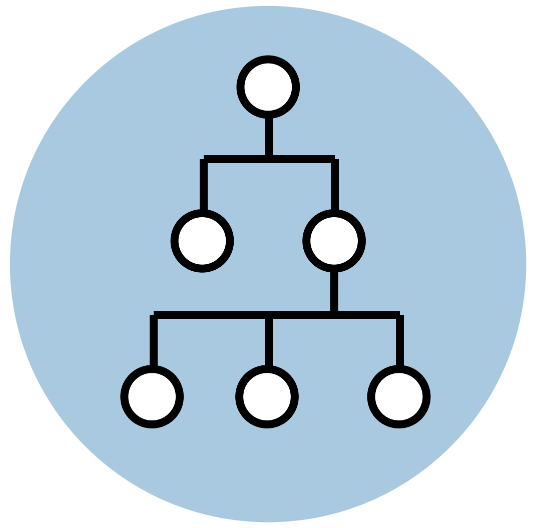
Deep ML

Can get accurate predictions

High computational resources

Careful optimization of learning architecture

High data requirement



Lightweight ML

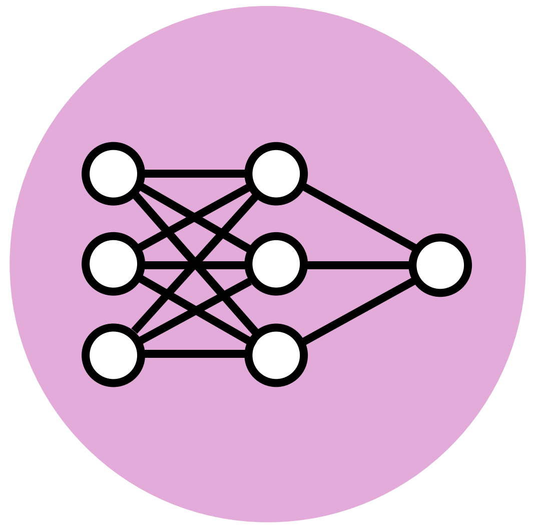
Run on your laptop

Ready to use systems

Lower data requirement

Why Lightweight ML?

UNIVERSITY OF CAMBRIDGE



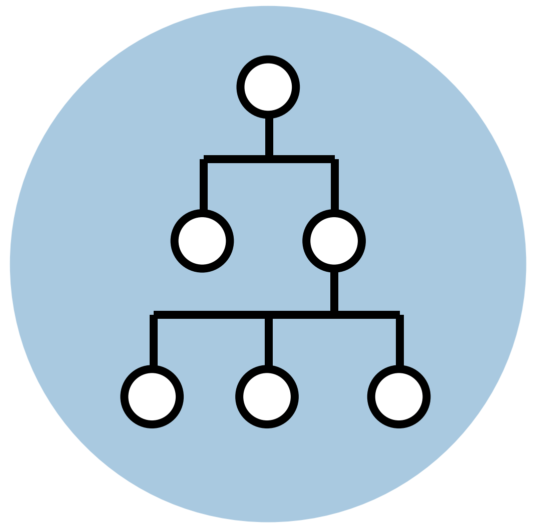
Deep ML

Can get accurate predictions

High computational resources

Careful optimization of learning architecture

High data requirement



Lightweight ML

Qualitative Predictions

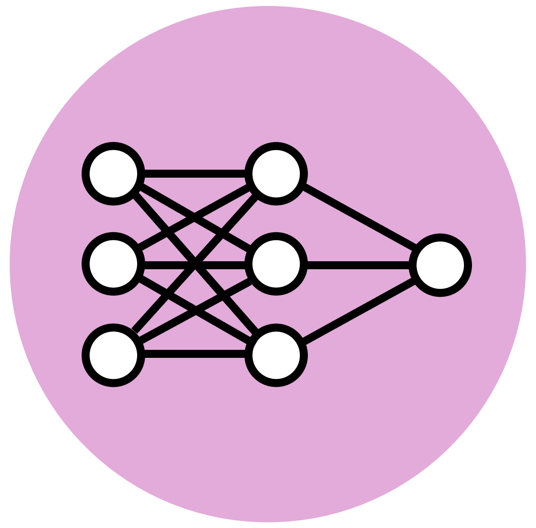
Run on your laptop

Ready to use systems

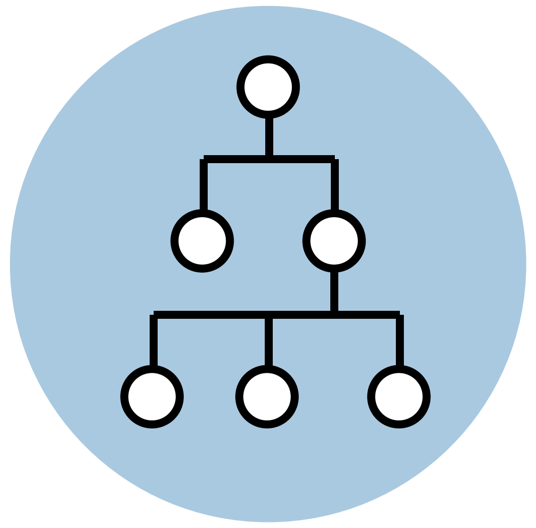
Lower data requirement

Why Lightweight ML?

UNIVERSITY OF CAMBRIDGE



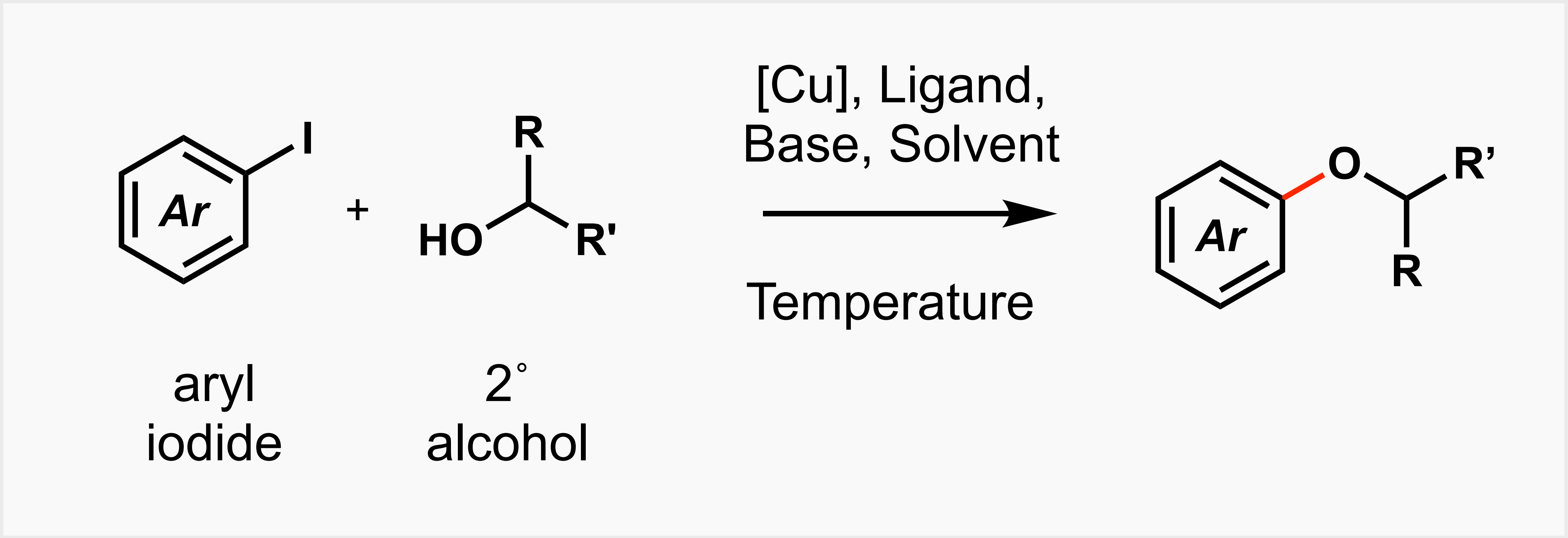
Deep ML



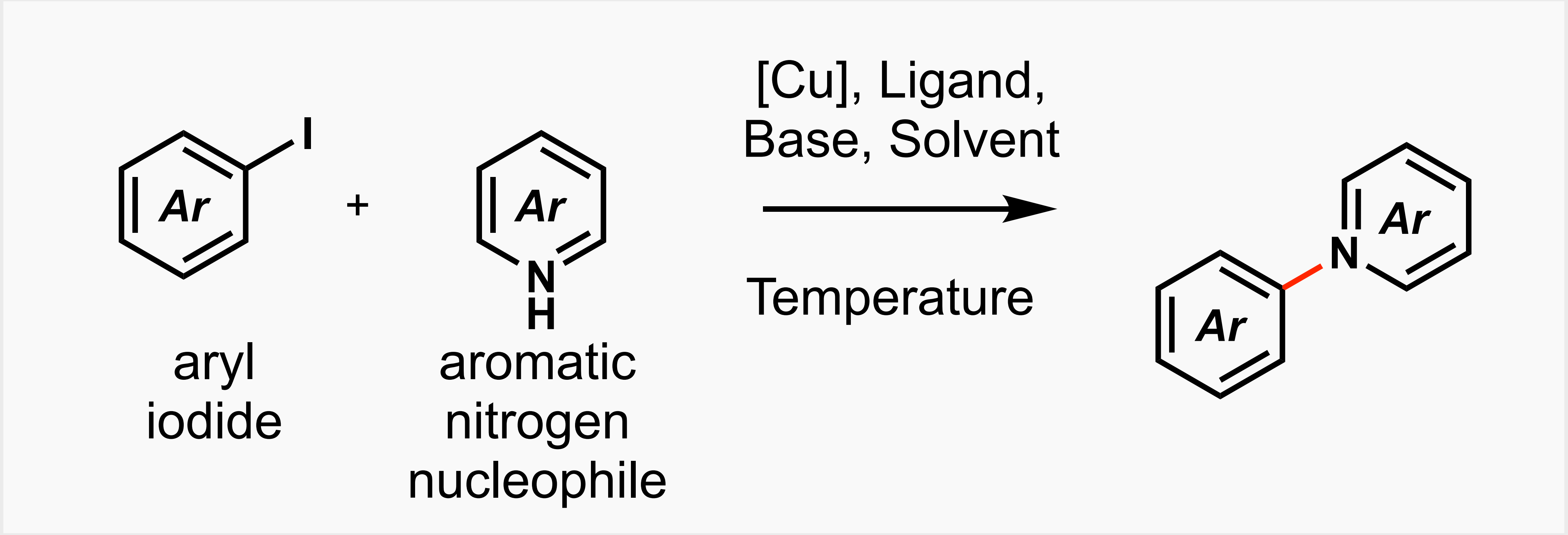
Lightweight ML

Deep ML	Lightweight ML
Can get accurate predictions	Qualitative Predictions
High computational resources	Run on your laptop
Careful optimization of learning architecture	Ready to use systems
High data requirement	Lower data requirement

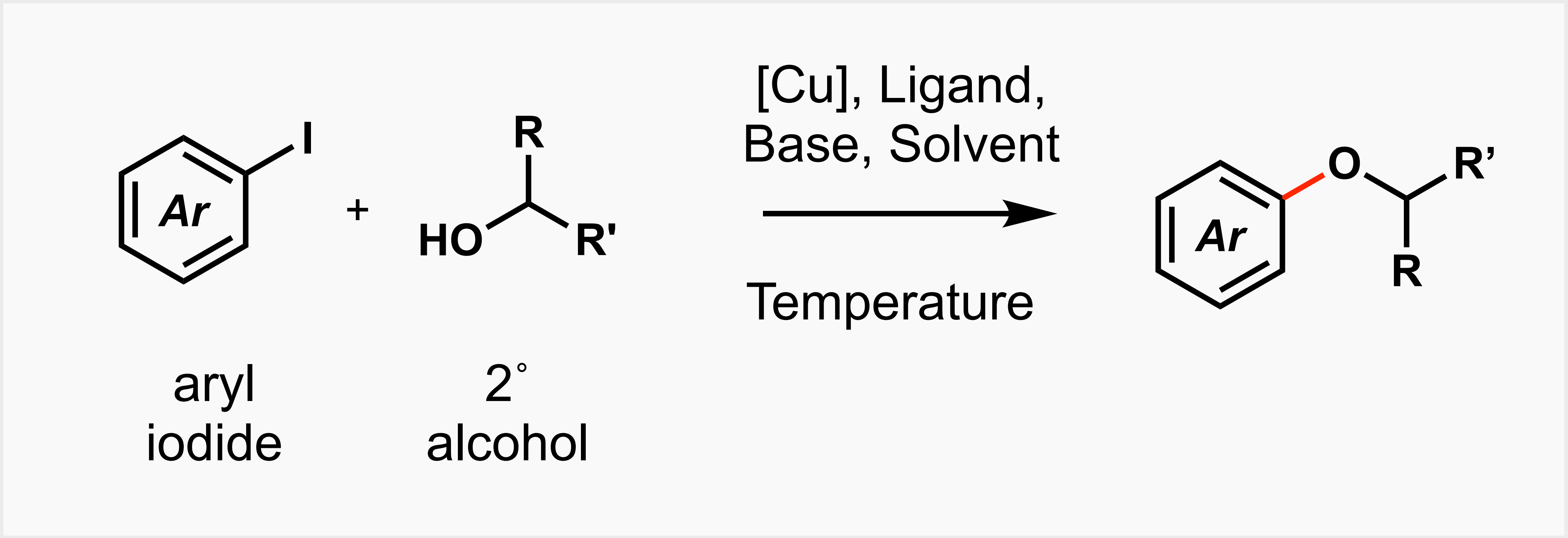
hidden chemical insights

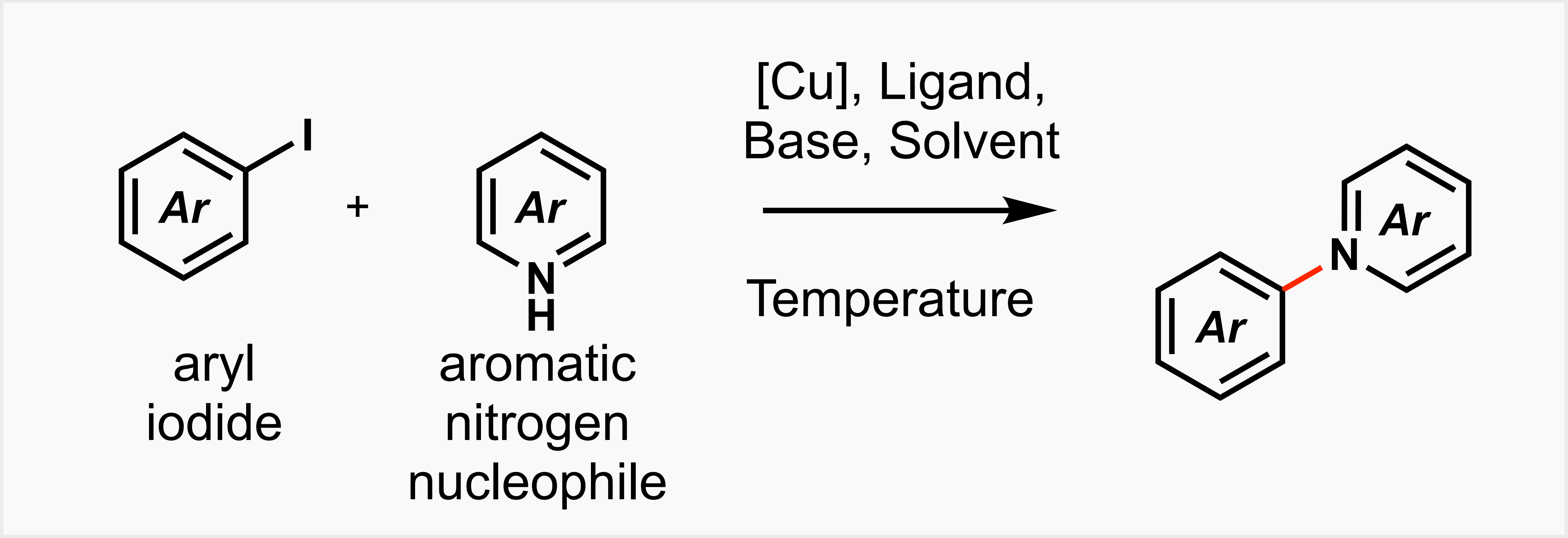


I
Ar
+
R
HO
R'
[Cu], Ligand,
Base, Solvent
Temperature
O
R'
R
Ar
aryl
iodide
2˚
alcohol



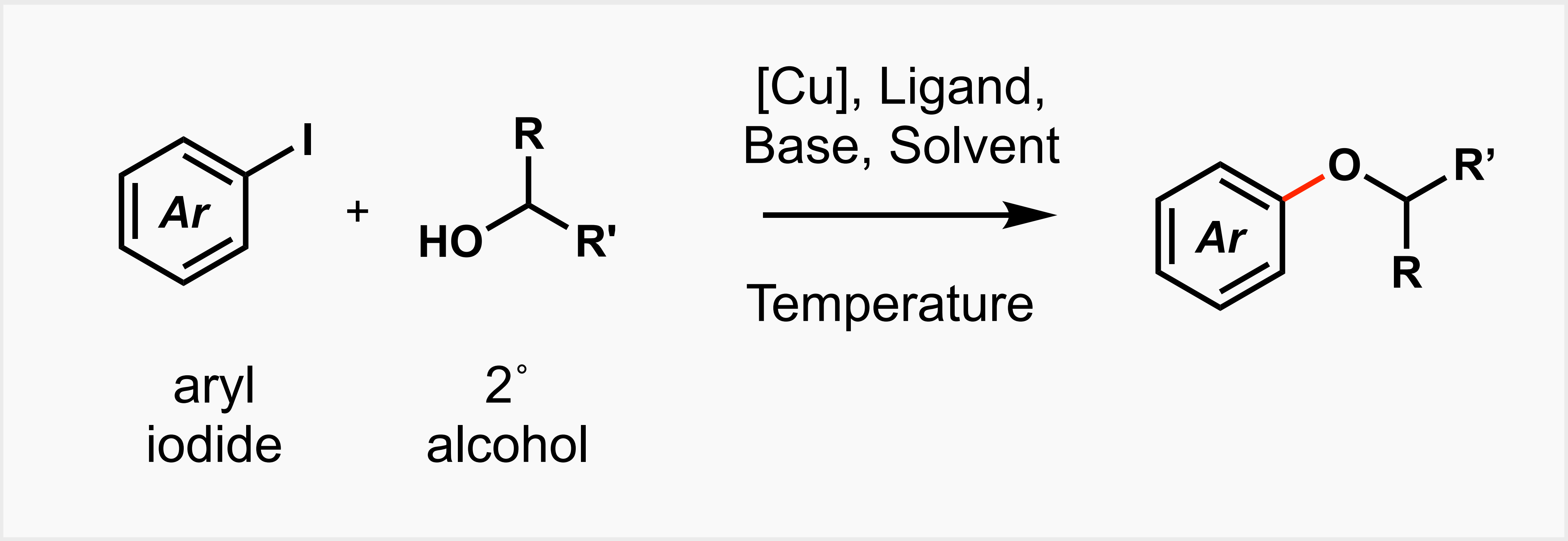
I
Ar
+
Ar
N
H
[Cu], Ligand,
Base, Solvent
Temperature
Ar
N
Ar
aryl
iodide
aromatic
nitrogen
nucleophile

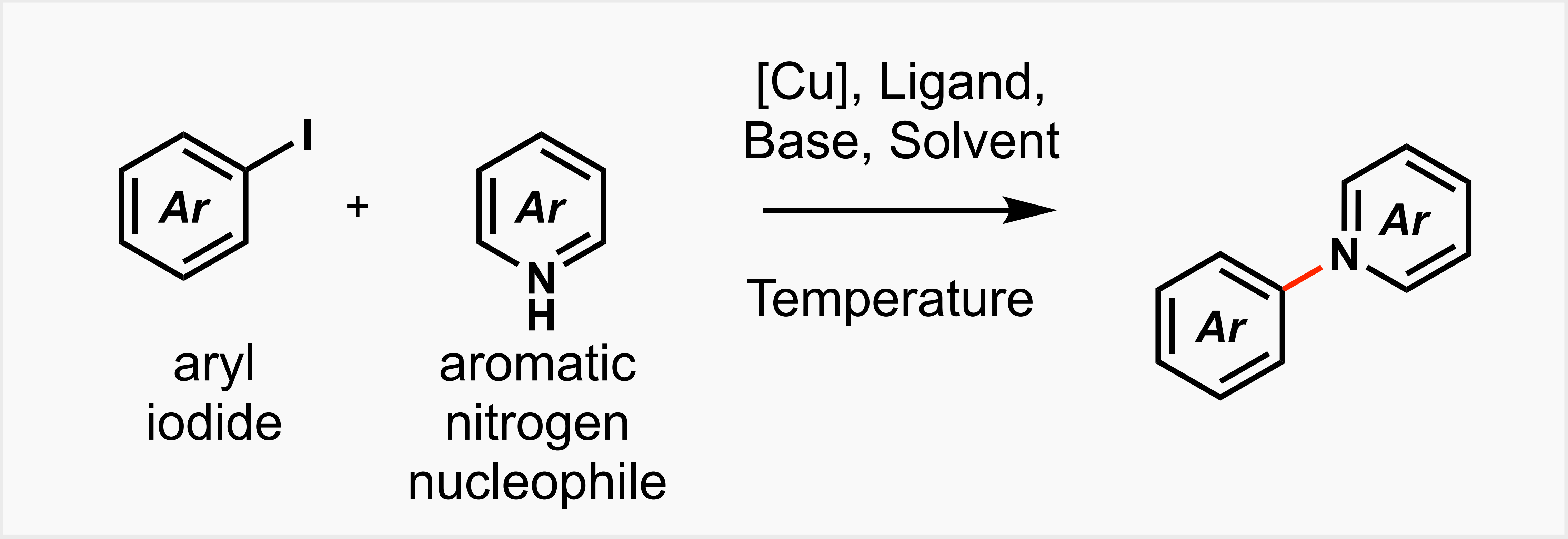




Historical data

Nature Chem. **2024**, 16, 633.

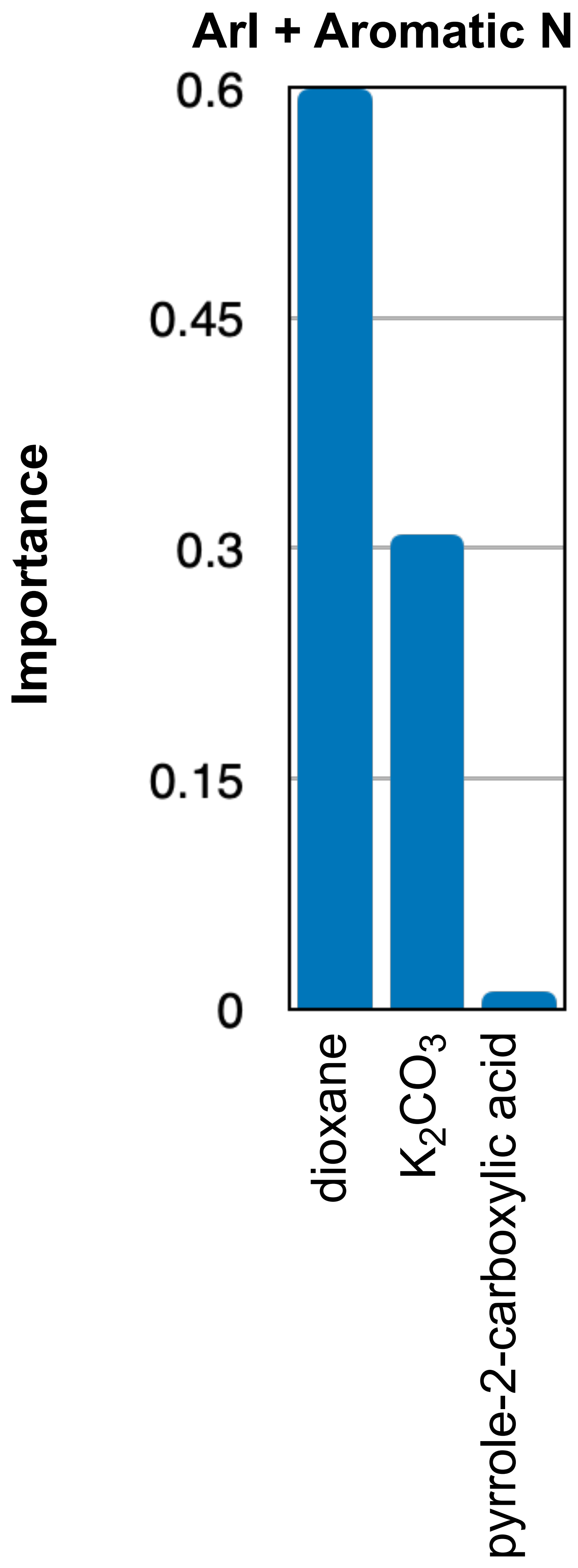


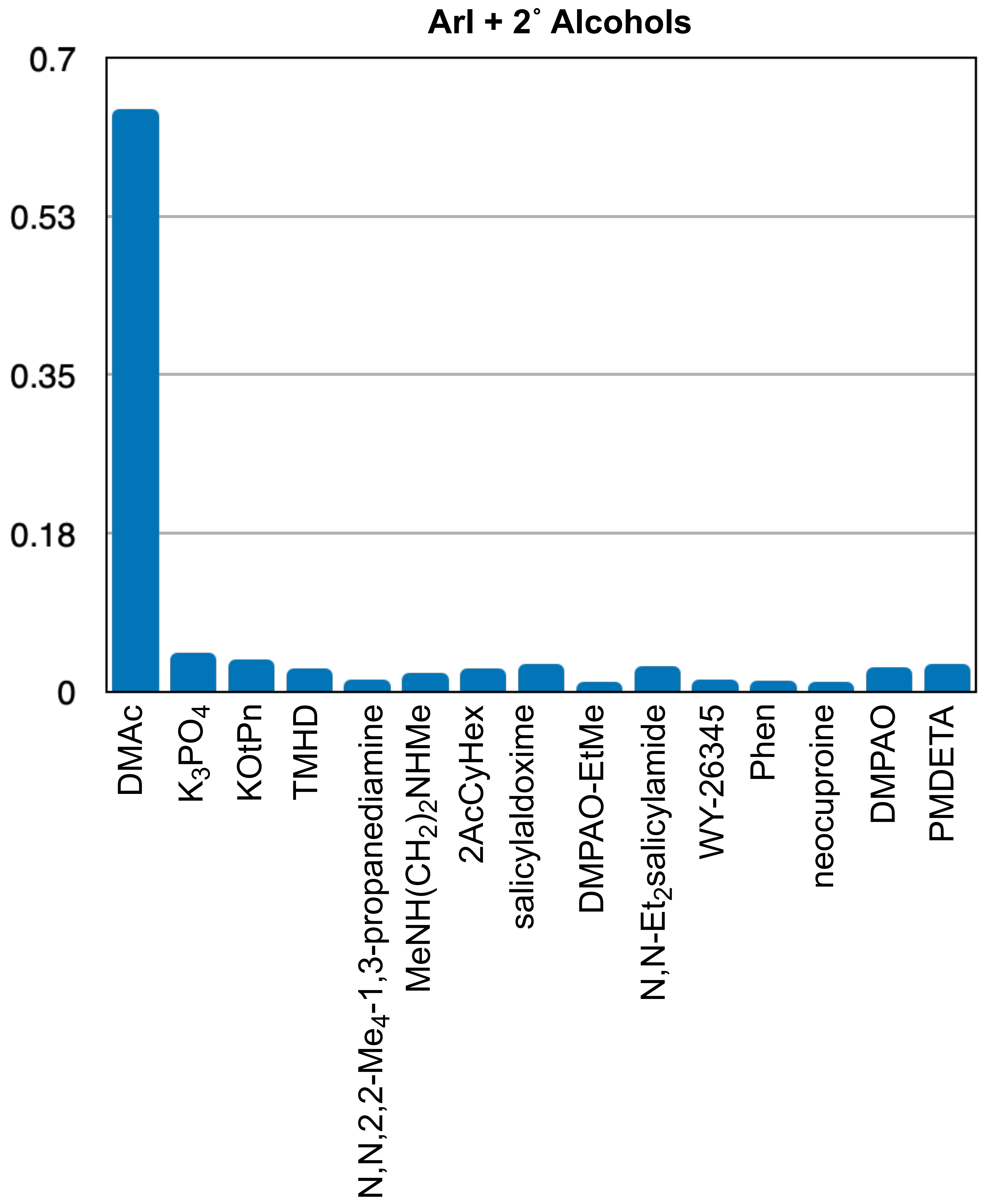


Historical data

What are the important factors for reaction yield in each specific reaction class?

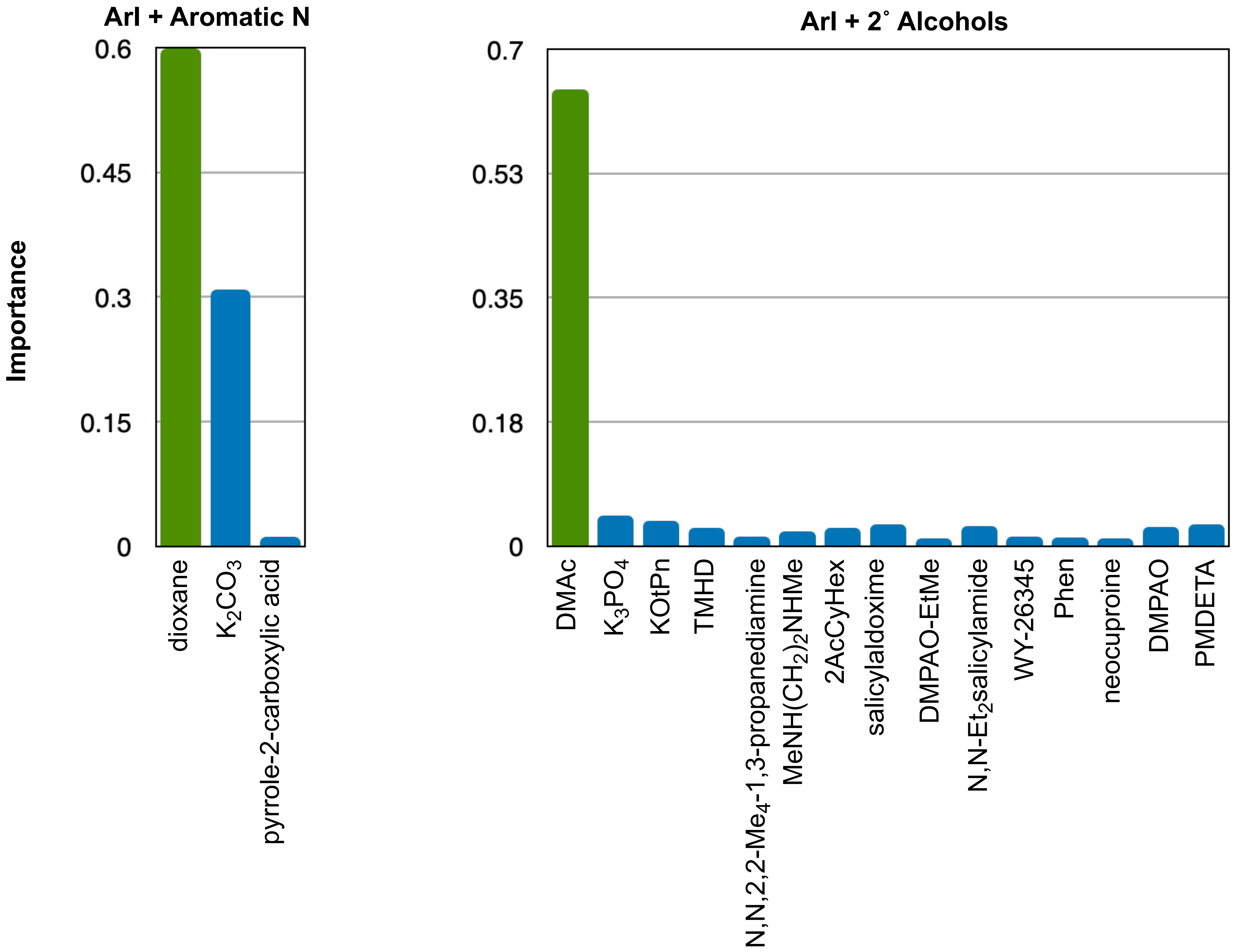
Important Features of Ullmann Condensation





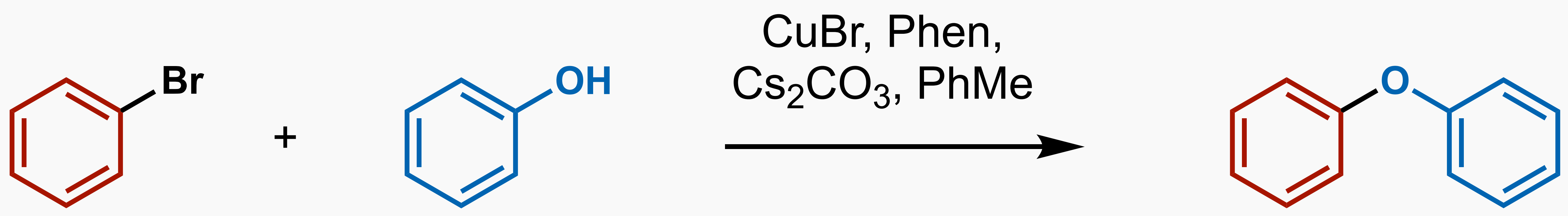
Nature Chem. **2024**, 16, 633.

Important Features of Ullmann Condensation



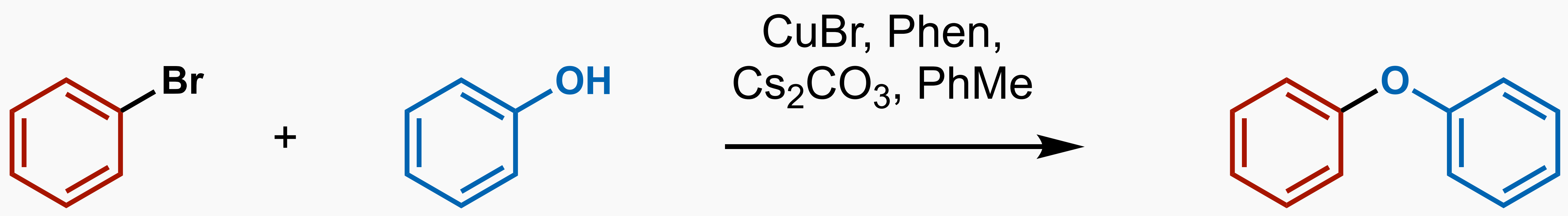
Nature Chem. **2024**, 16, 633.

UNIVERSITY OF CAMBRIDGE

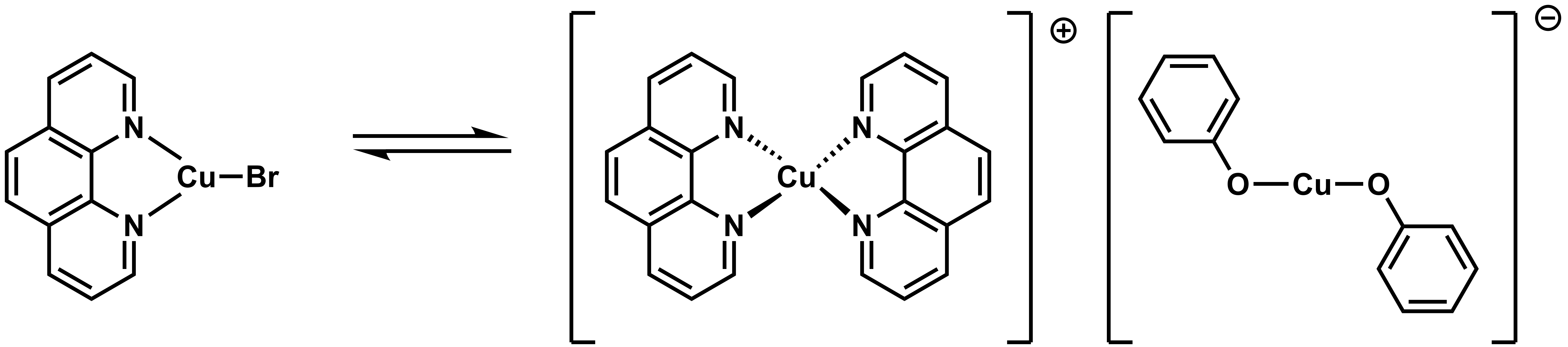


Dalton Trans. **2012**, *41*, 13832.

In the Literature…



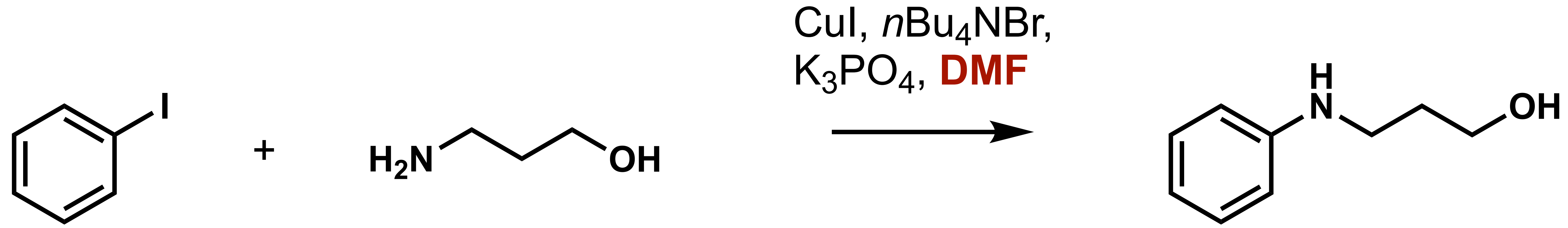
Solvent Polarity Controlled Equilibrium



Solvent can effect active catalytic species.

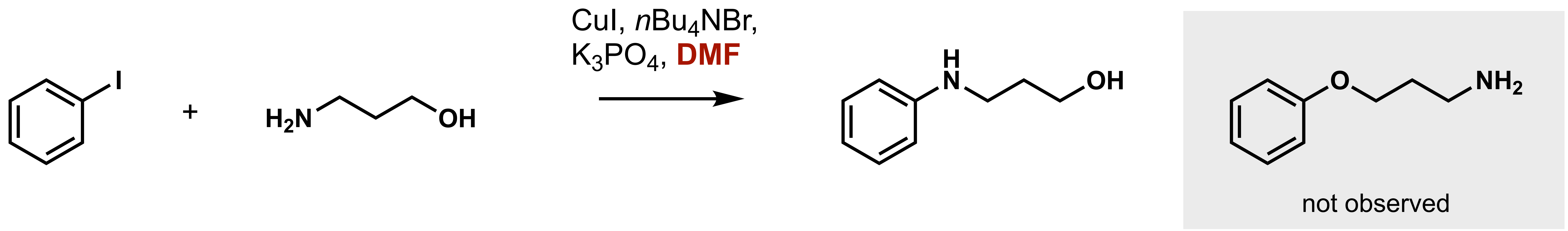
Dalton Trans. **2012**, *41*, 13832.

In the Literature…

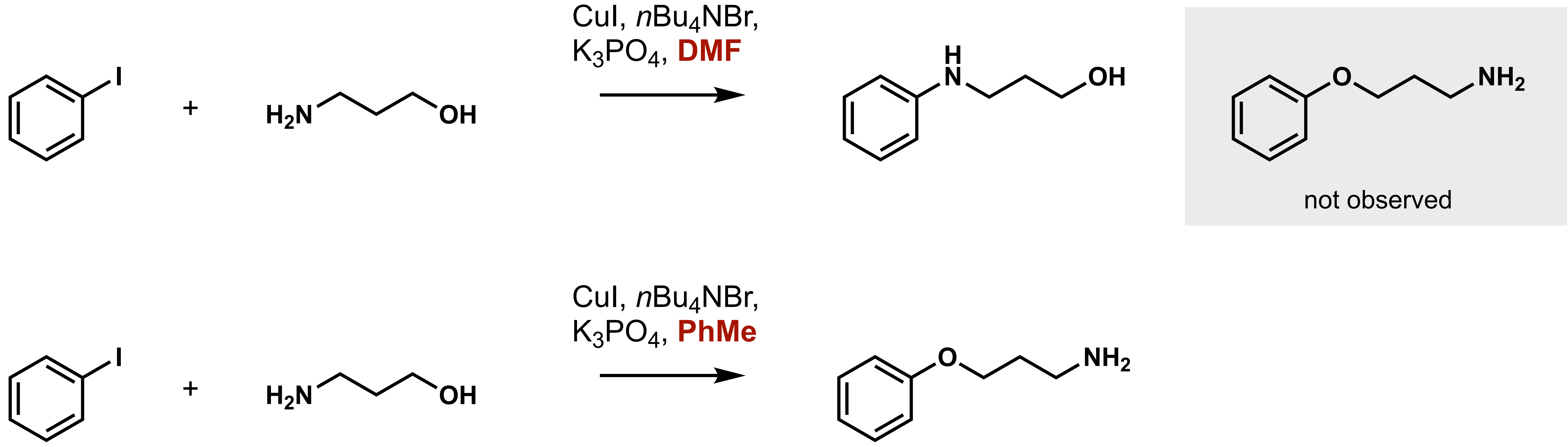


Tetrahedron Lett. **2008**, *49*, 2018.

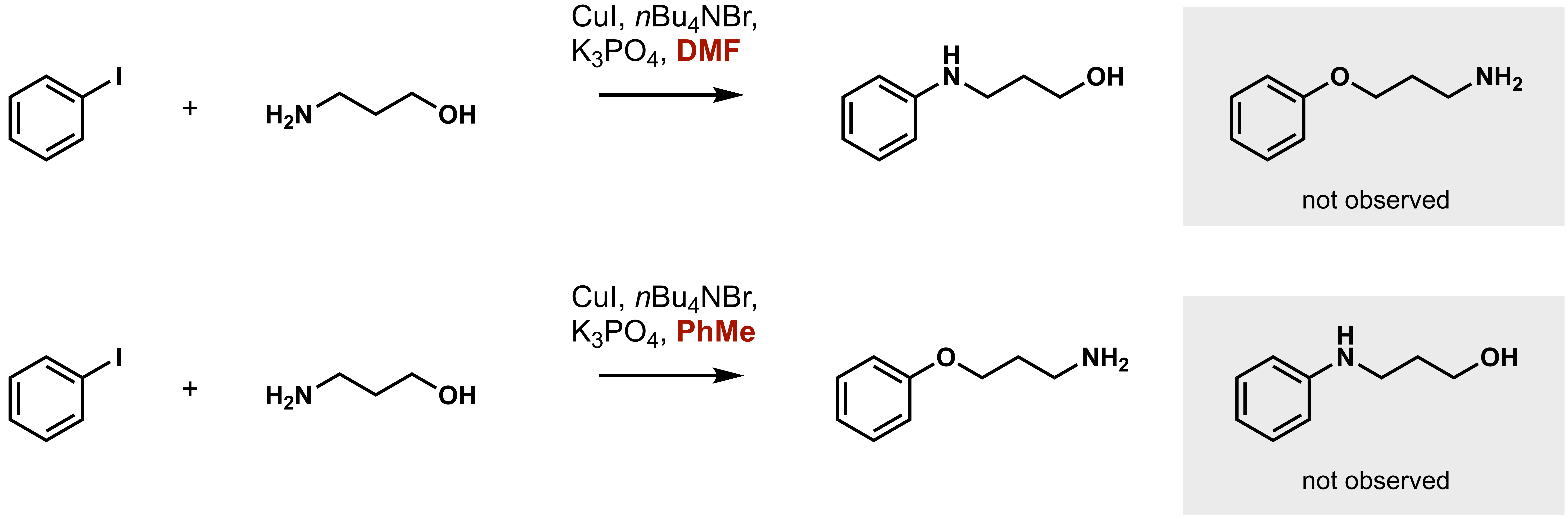
In the Literature…



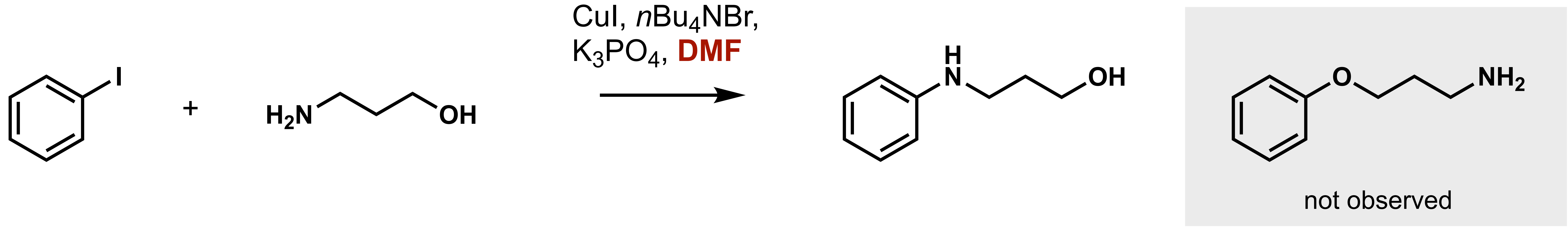
In the Literature…

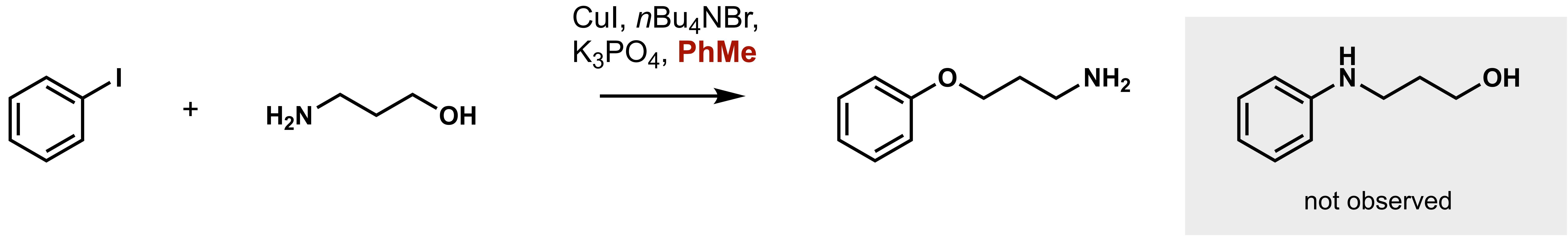


In the Literature…



UNIVERSITY OF CAMBRIDGE





No systematic review of solvent effects.

We can draw out the best solvents for a given reaction.

Tetrahedron Lett. **2008**, *49*, 2018.

Acknowledgements

UNIVERSITY OF CAMBRIDGE

Lee Group

Dr. Alpha Lee

Dr. Felix Faber

Rokas Elijošius

William McCorkindale

Pfizer

Dr. Joy Yang

Dr. Roger M. Howard

Dr. Simon Berritt

Dr. Anton V. Sinitskiy

Usa Reilly

Gaunt Group

Prof. Matt Gaunt

Dr Srimanta Manna

Markus Böcker

Wolfson College

Prof. Jane Clarke

Lyn Alcantara

Wolfson College Choir

Funding

Newton International Fellowship (Royal Society)

Pfizer

Coming Soon

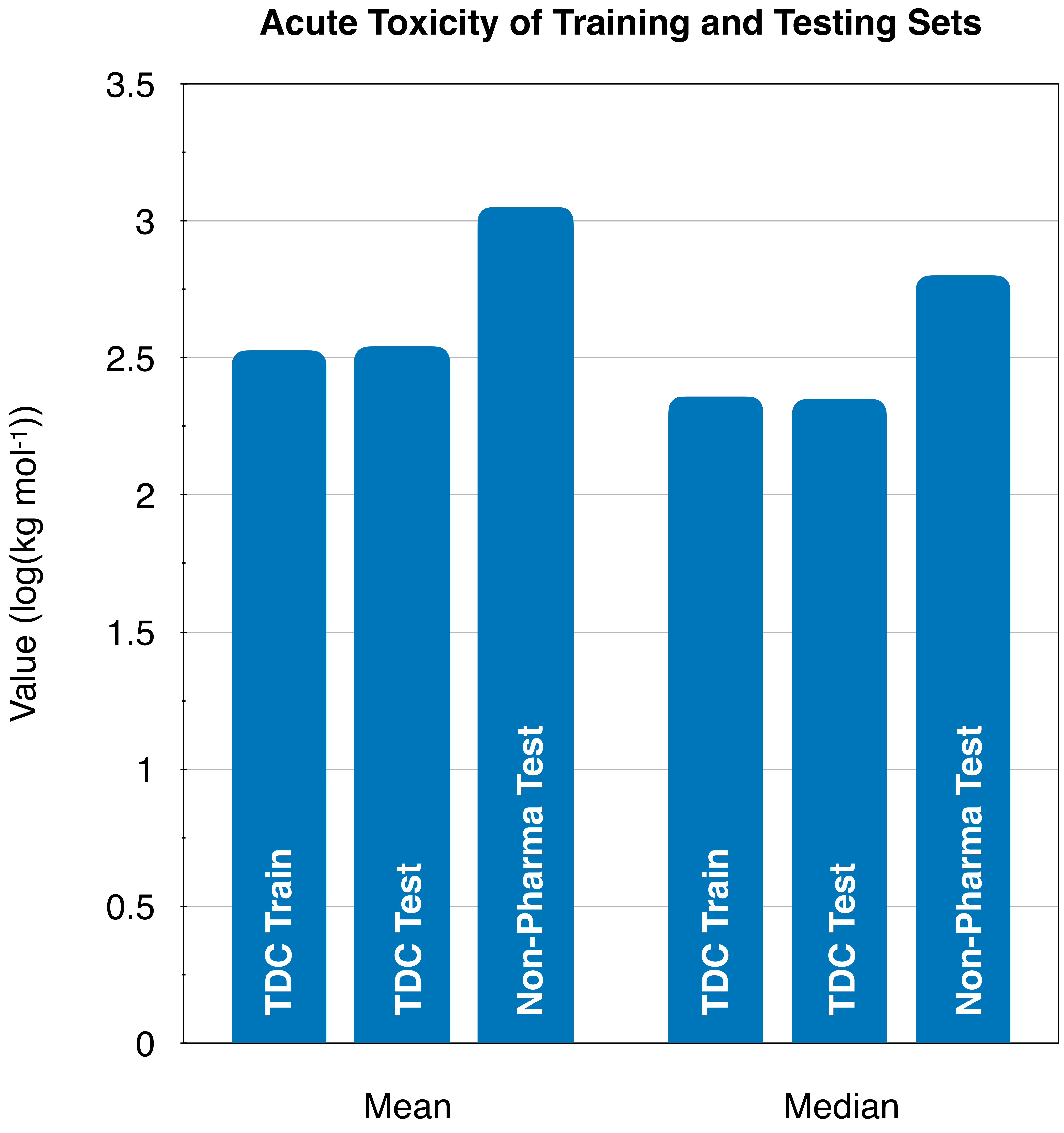
THE UNIVERSITY OF EDINBURGH

THE ROYAL SOCIETY

Pfizer

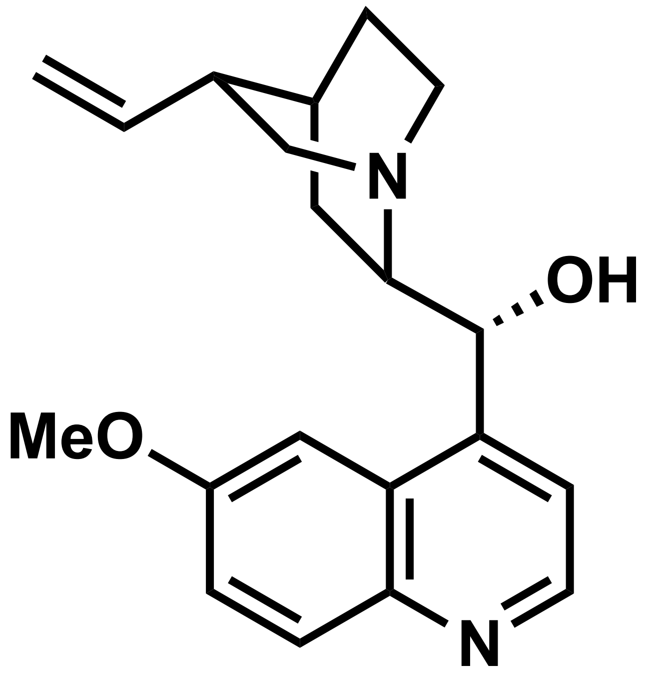
Have a Question? esk34@cam.ac.uk

Training and Test Set Comparisons



arXiv preprint **2021**, DOI: arXiv:2102.09548.

Backup: Retrospective Test Set

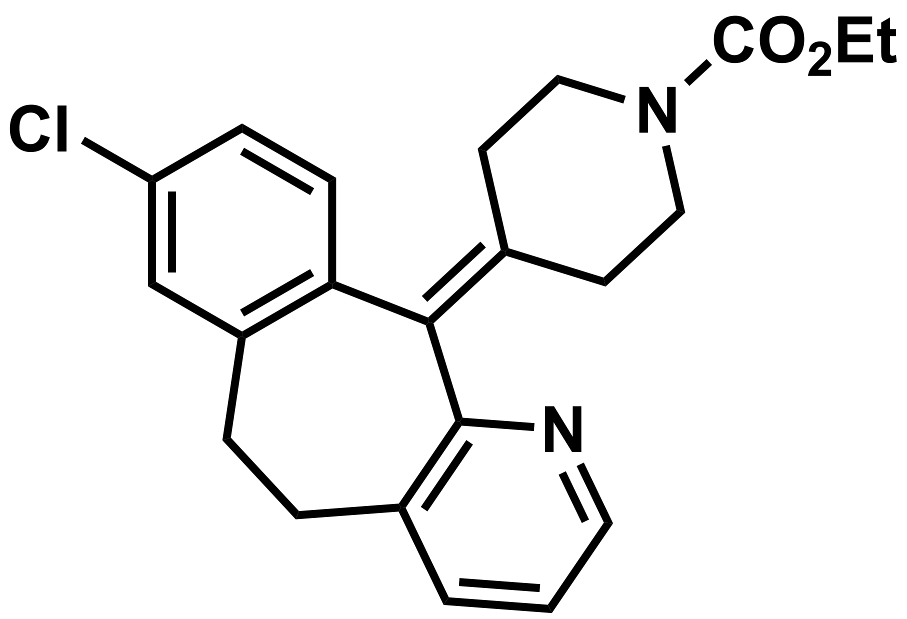


quinidine (**1**)

Unique LSF Reagents

$cBuBF_3K$

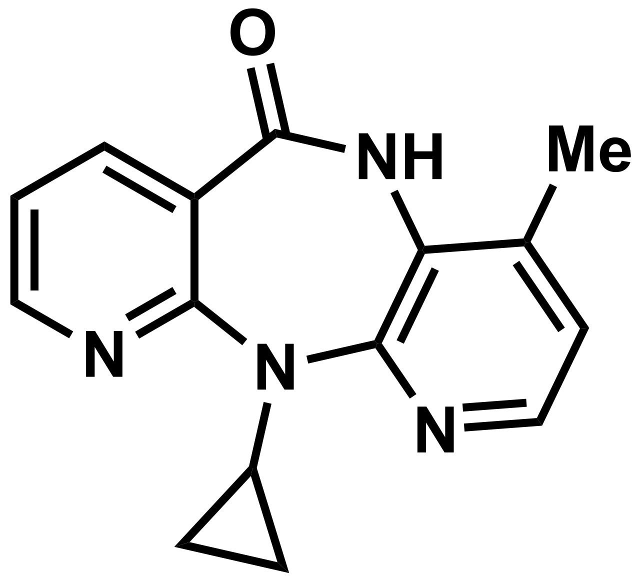
$(CF_3SO_2)_2Zn$



loratadine (**2**)

Unique LSF Reagents

1-CF_3-cPrSO_2Na	Metalloenzyme
CF_3SO_2Na	$(HCF_2SO_2)_2Zn$
P450	$HOCH_2SO_2Na$



nevirapine (**3**)

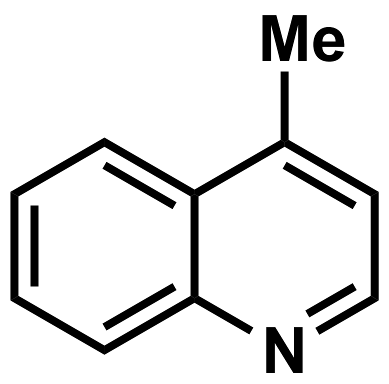
Unique LSF Reagents

$cBuBF_3K$

$(CF_3SO_2)_2Zn$

Nat. Commun. **2024**, *15*, 426.

Backup: Retrospective Test Set



lepidine (**4**)

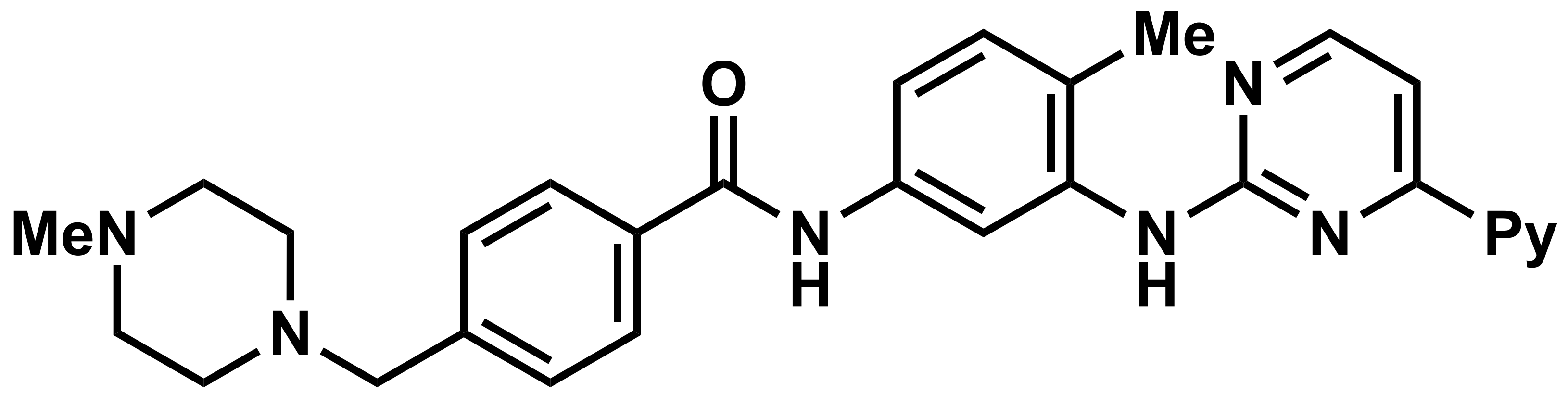
Unique LSF Reagents

$NHBocCH_2BF_3K$

$2\text{-}Me\text{-}cPrSO_2Na$

$cBuBF_3K$

$CF_3(CH_2)_2SO_2Na$



imatinib (**5**)

Unique LSF Reagents

$cBuBF_3K$

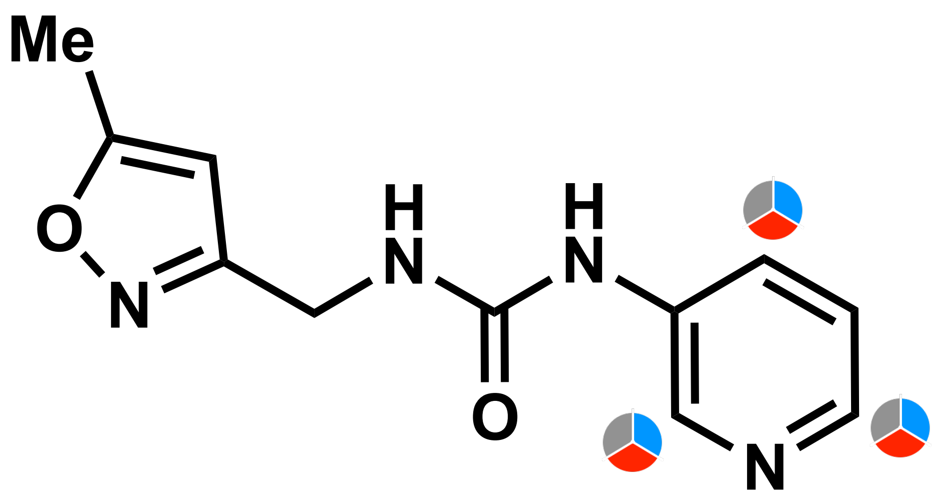
HCF_2SO_2Na

$(CF_3SO_2)_2Zn$

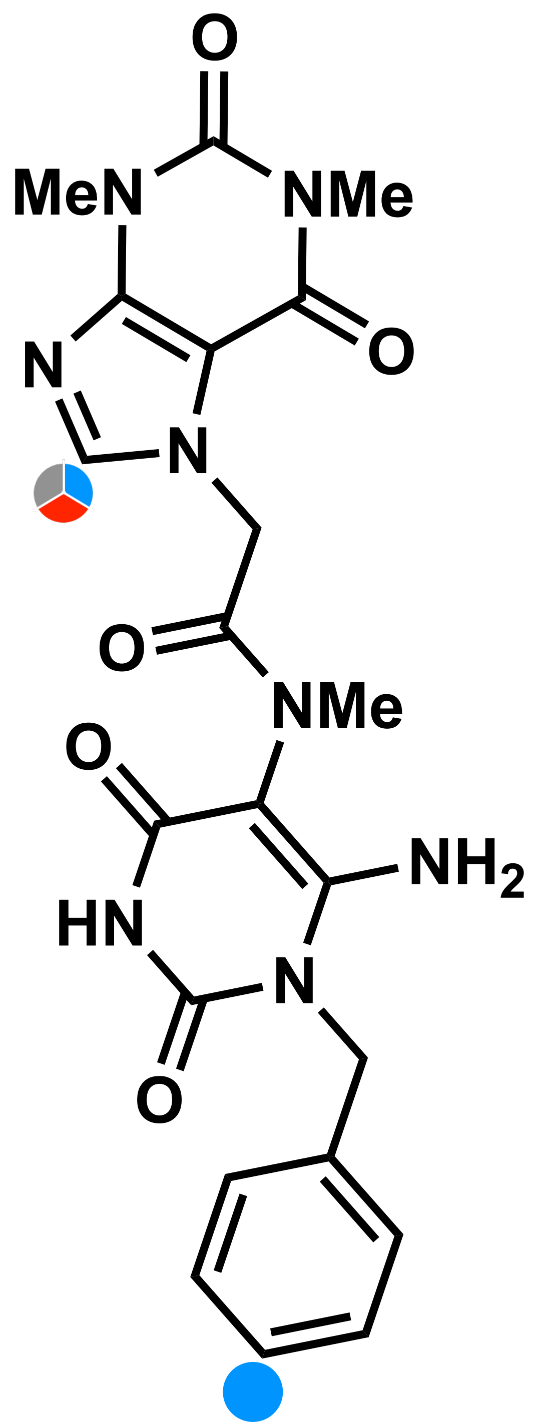
Nat. Commun. **2024**, *15*, 426.

Backup: Prospective Results

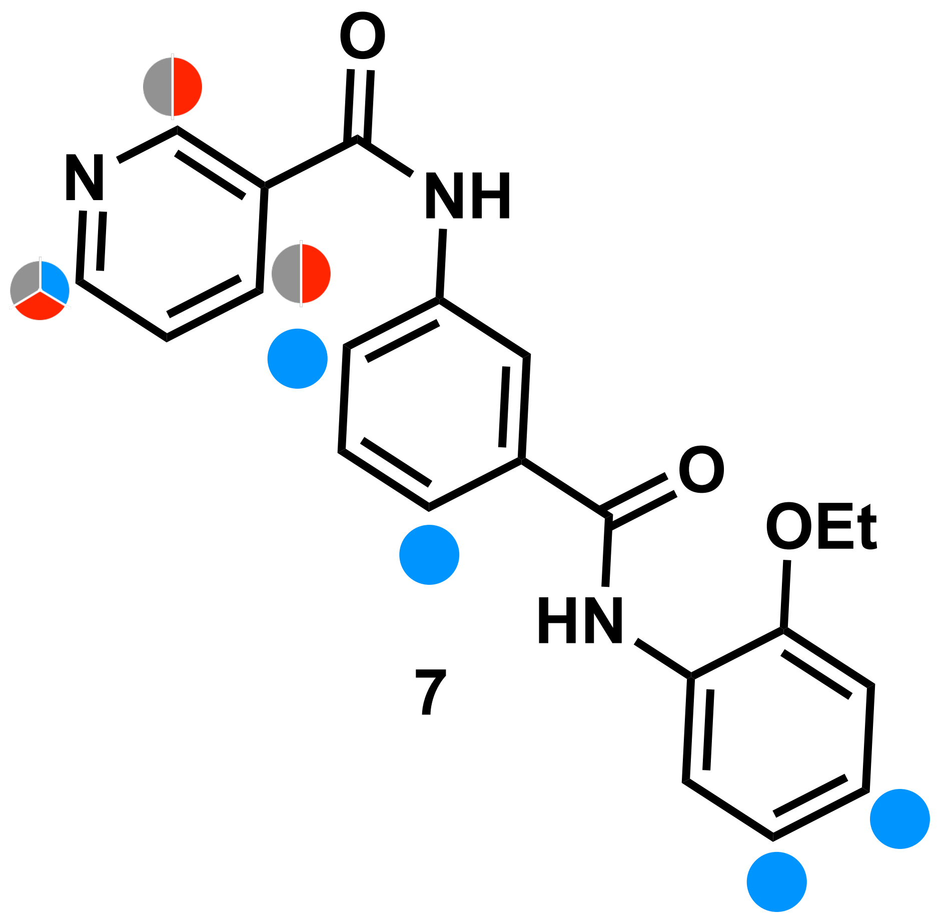
Prospective Test Set
Experimental Results



6

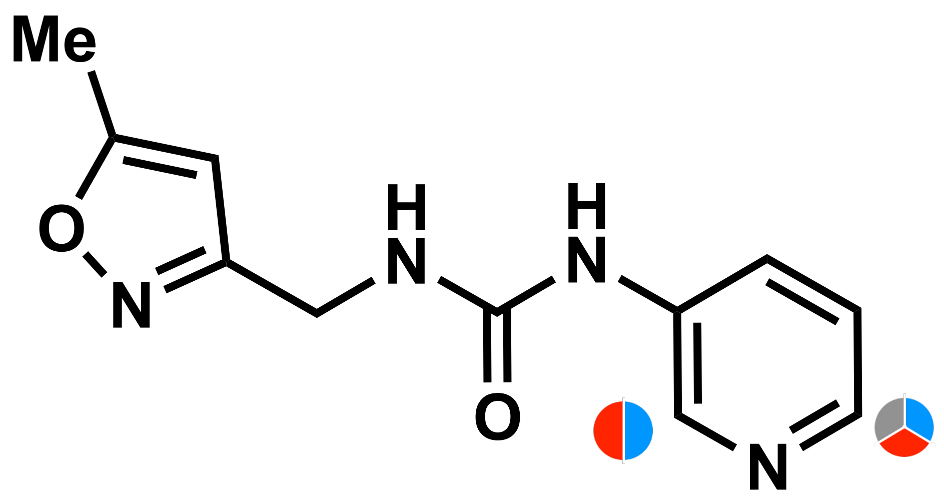


8

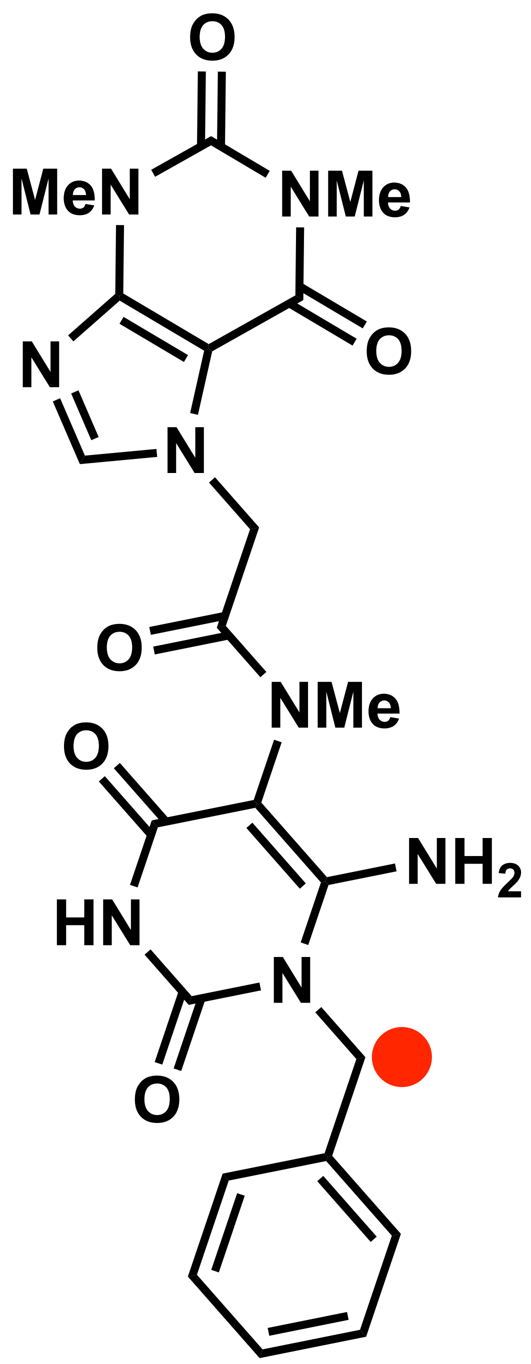


7

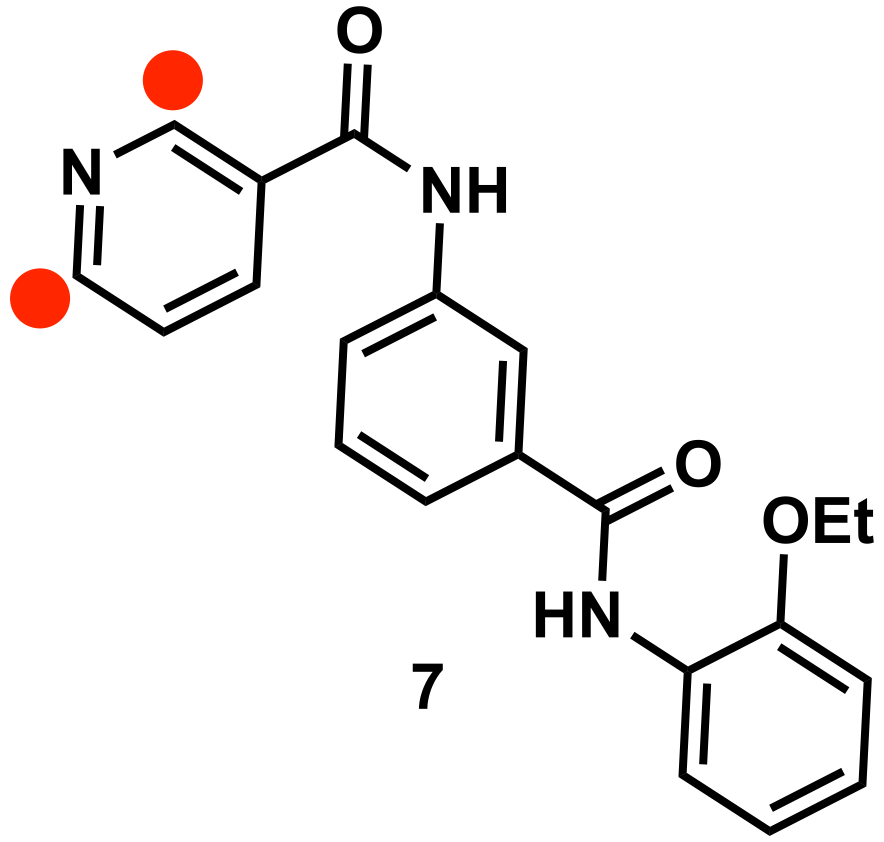
Prospective Test Set
Predictions MPNN$_{LSF}$



6



8



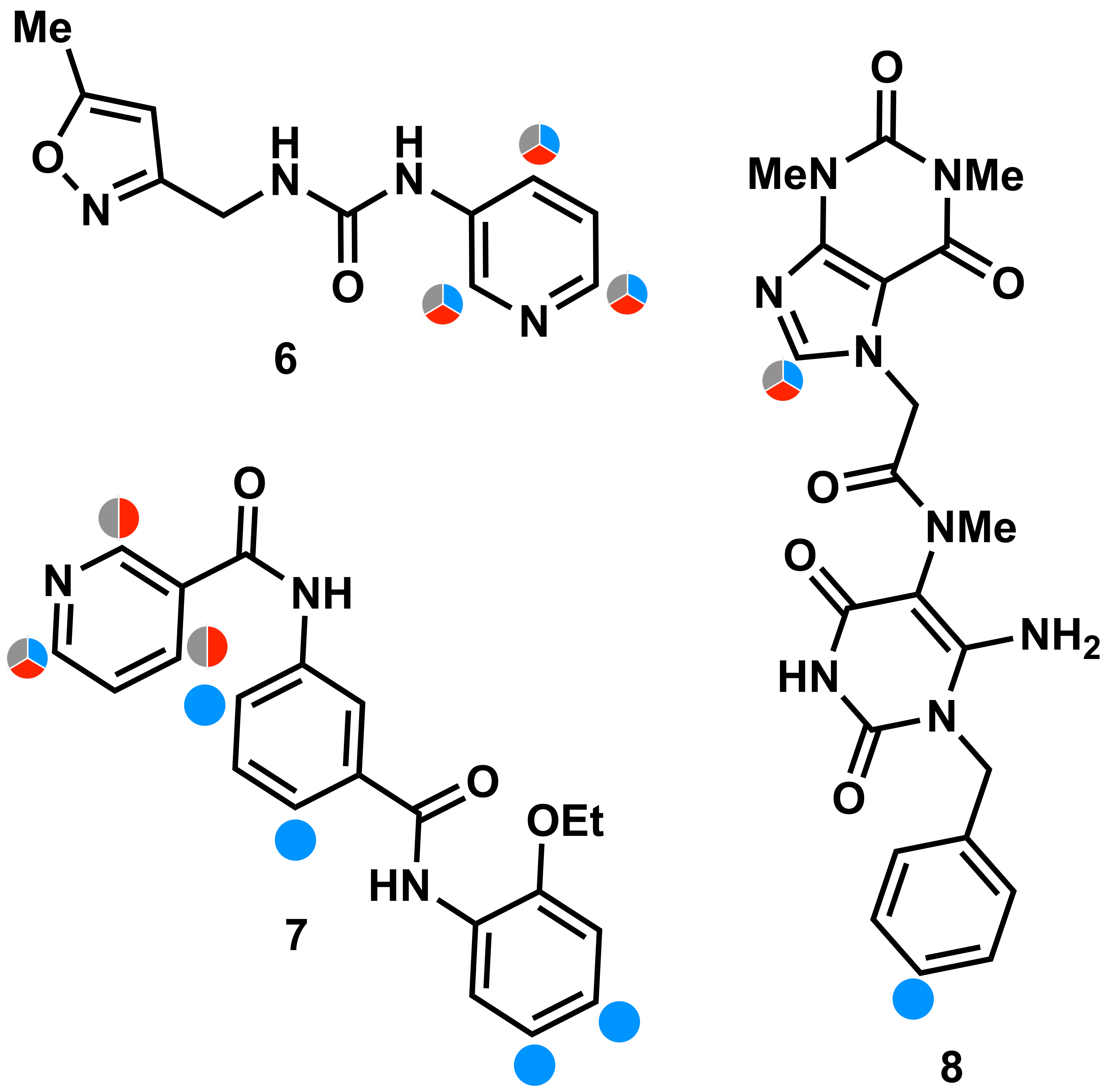
7

Nat. Commun. **2024**, *15*, 426.

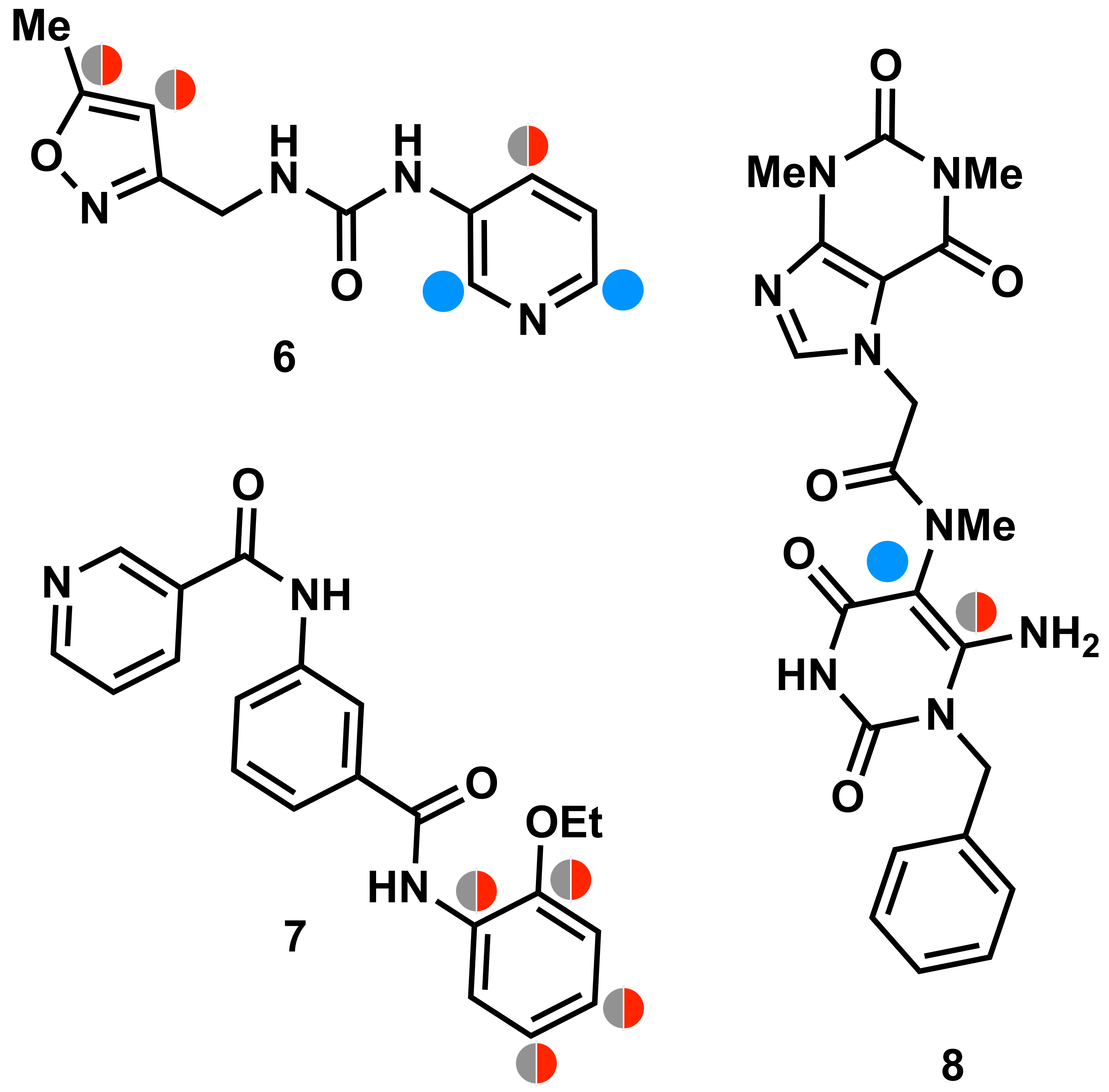
Backup: Prospective Results

UNIVERSITY OF CAMBRIDGE

Prospective Test Set
Experimental Results

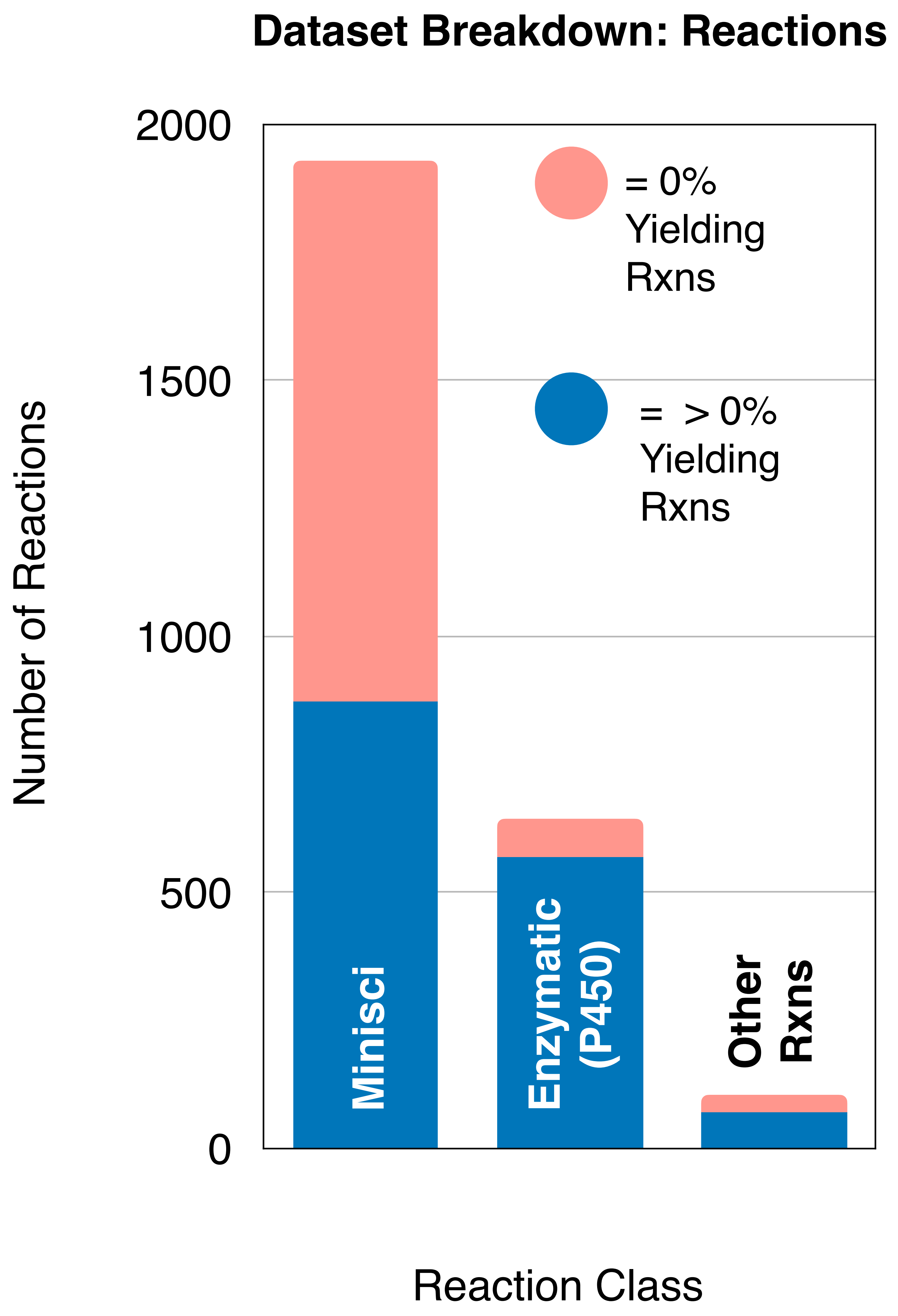


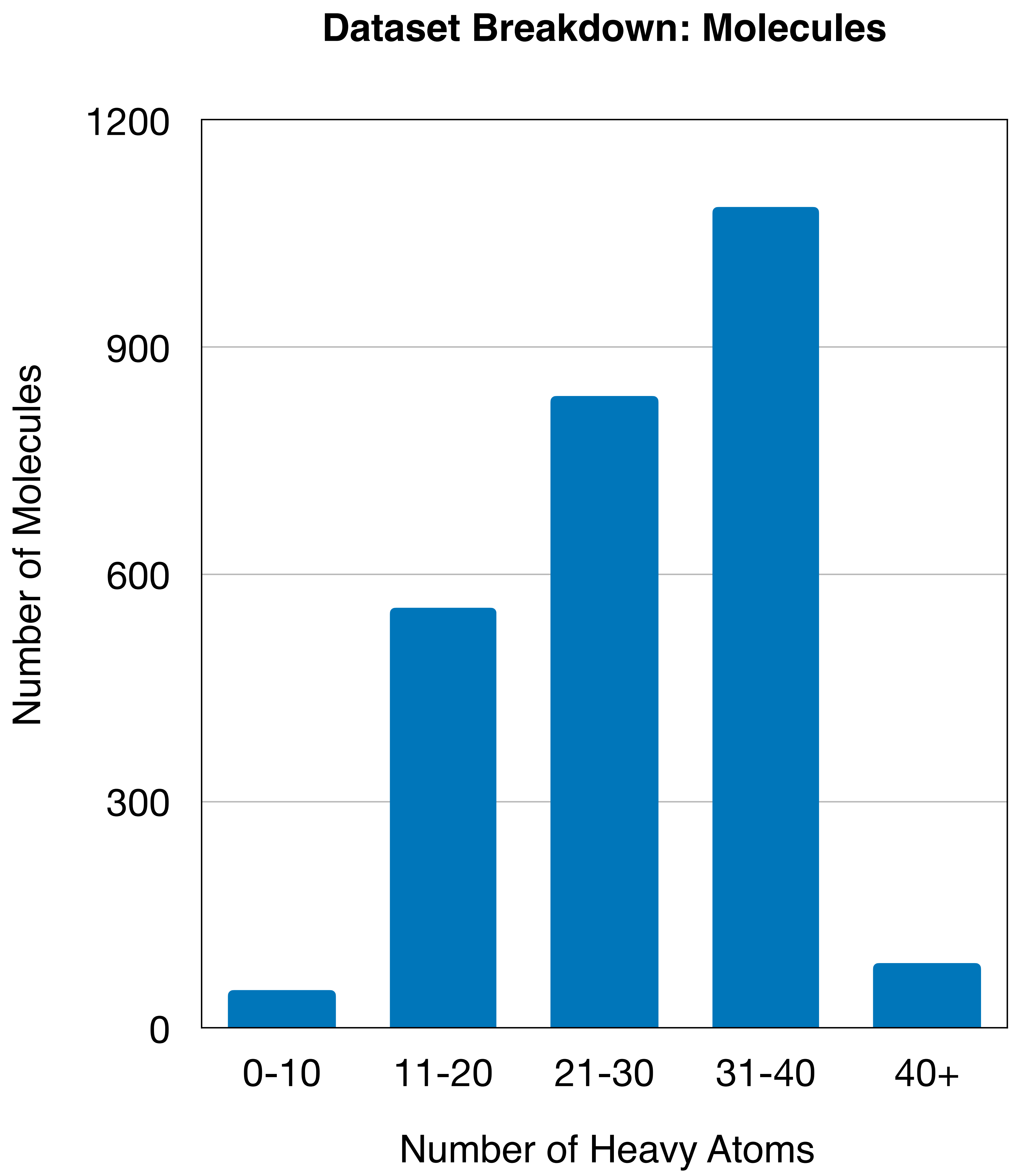
Prospective Test Set
Predictions Fukui



Nat. Commun. **2024**, *15*, 426.

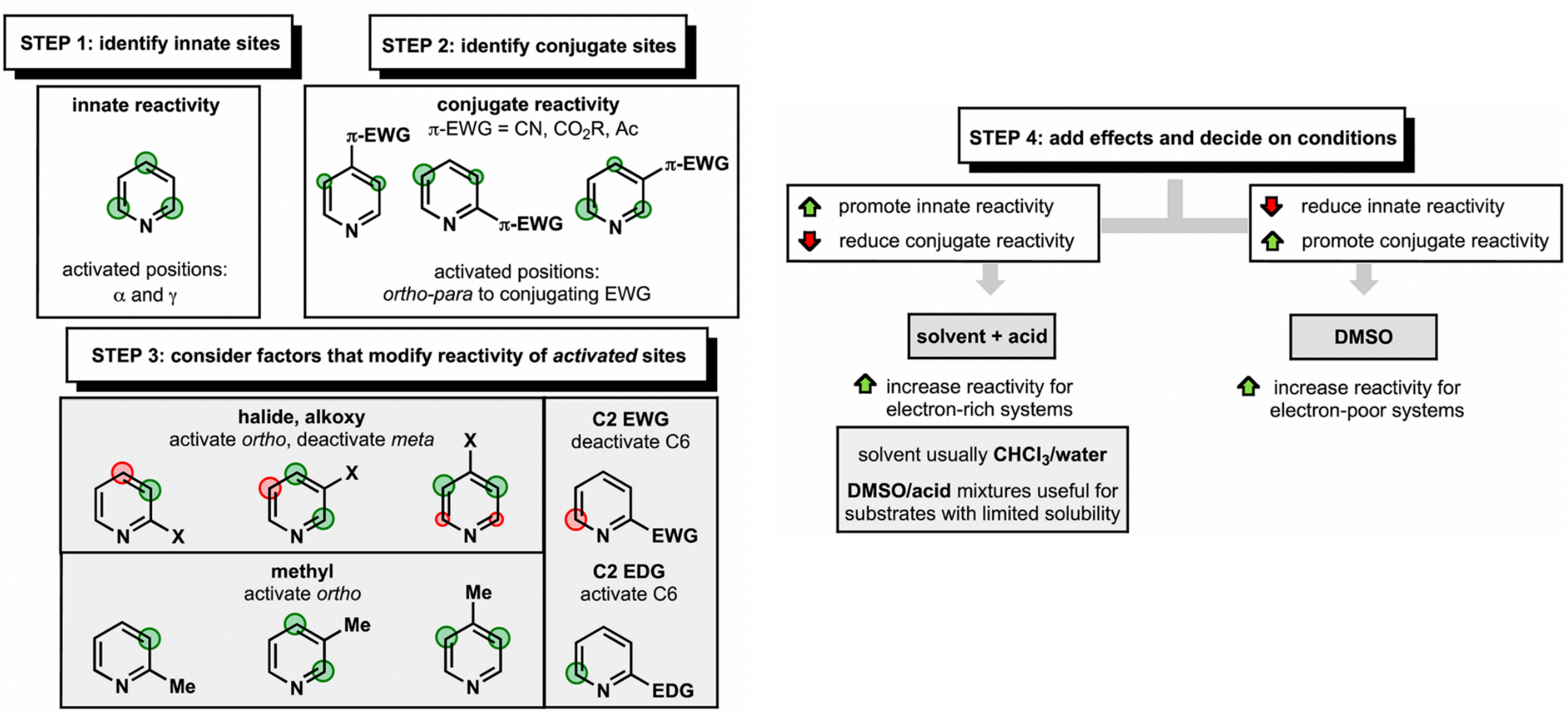
UNIVERSITY OF CAMBRIDGE





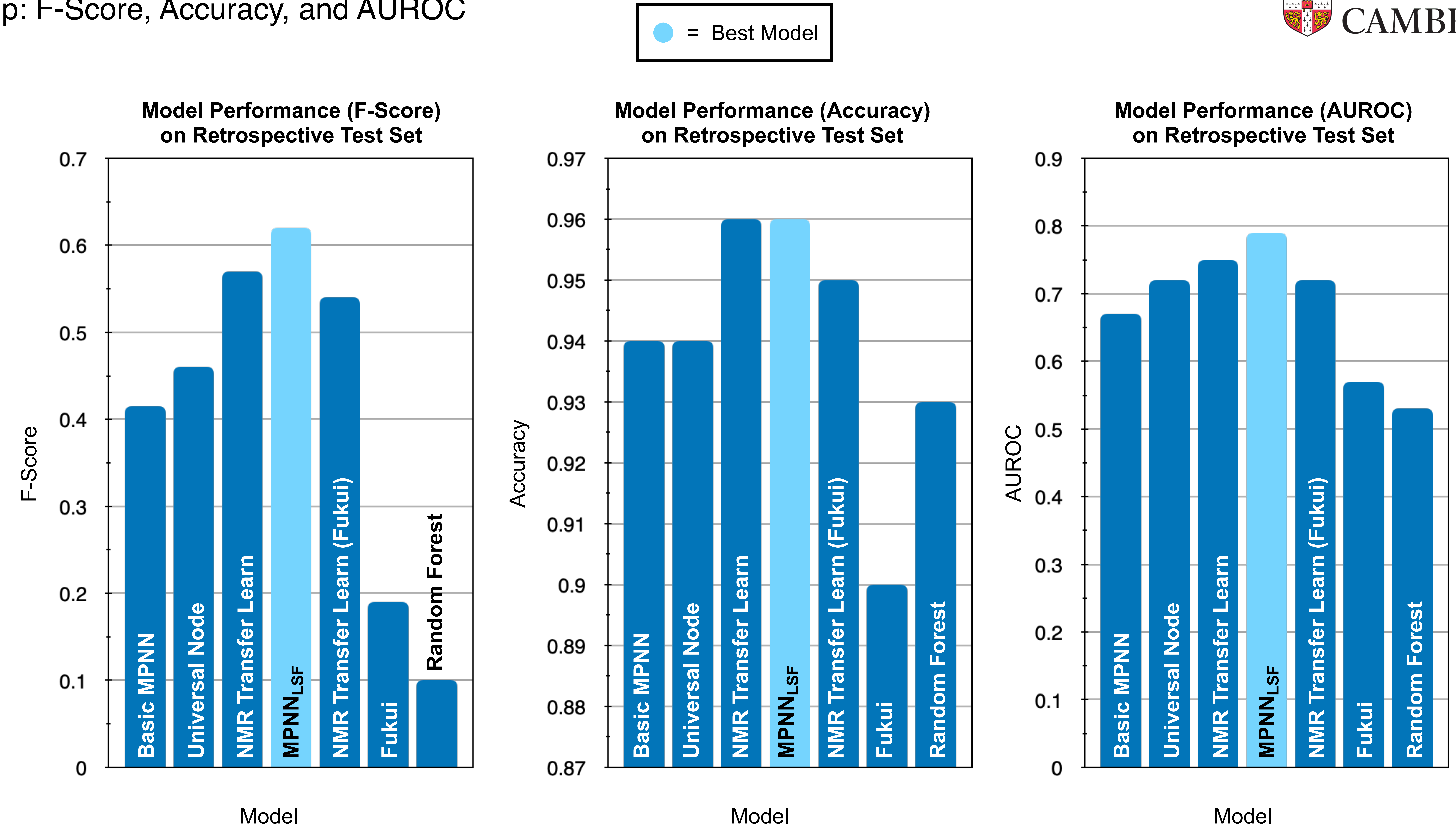
Nat. Commun. **2024**, *15*, 426.

Backup: Baran Heuristics



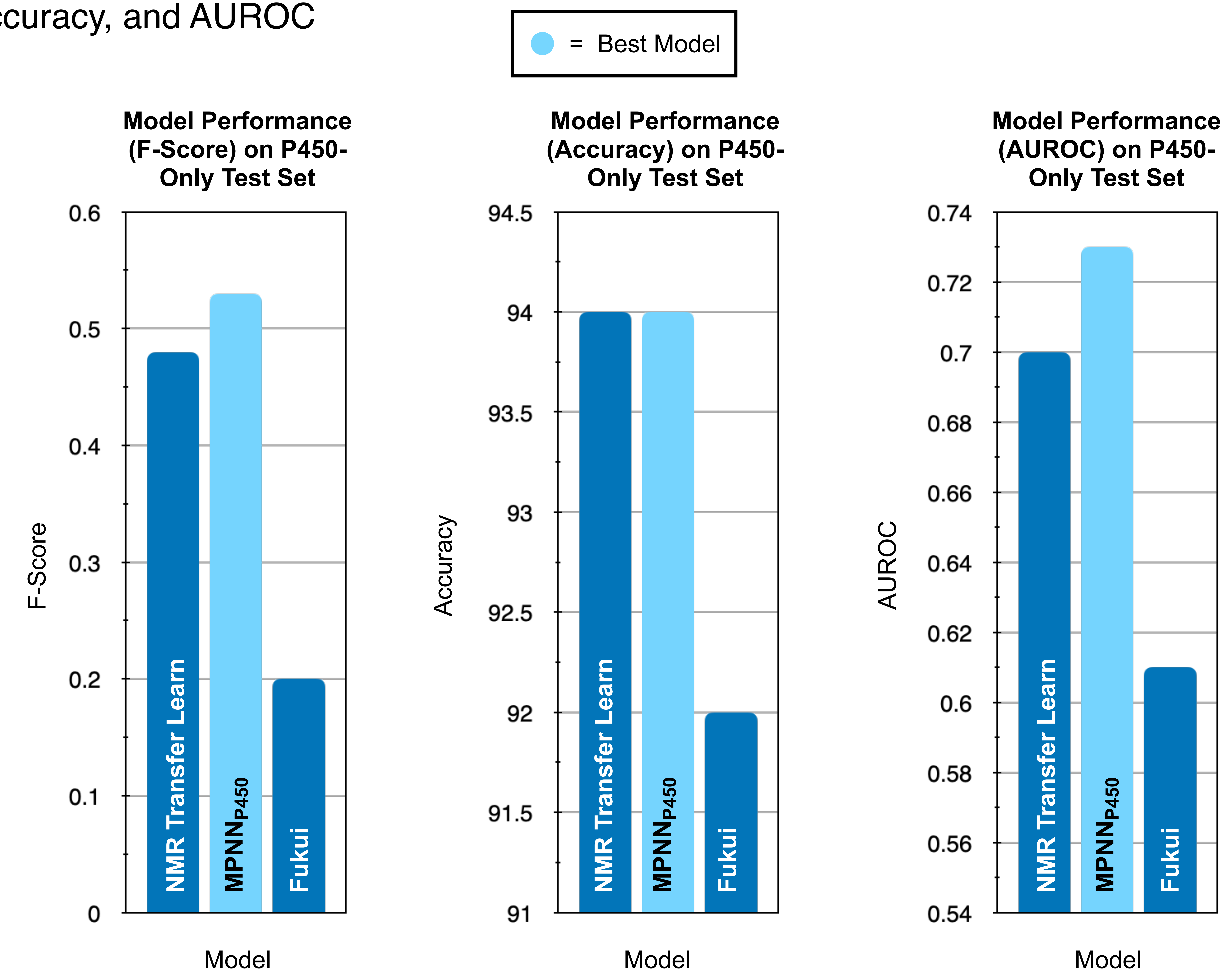
J. Am. Chem. Soc. **2013**, *135*, 12122.

Backup: F-Score, Accuracy, and AUROC



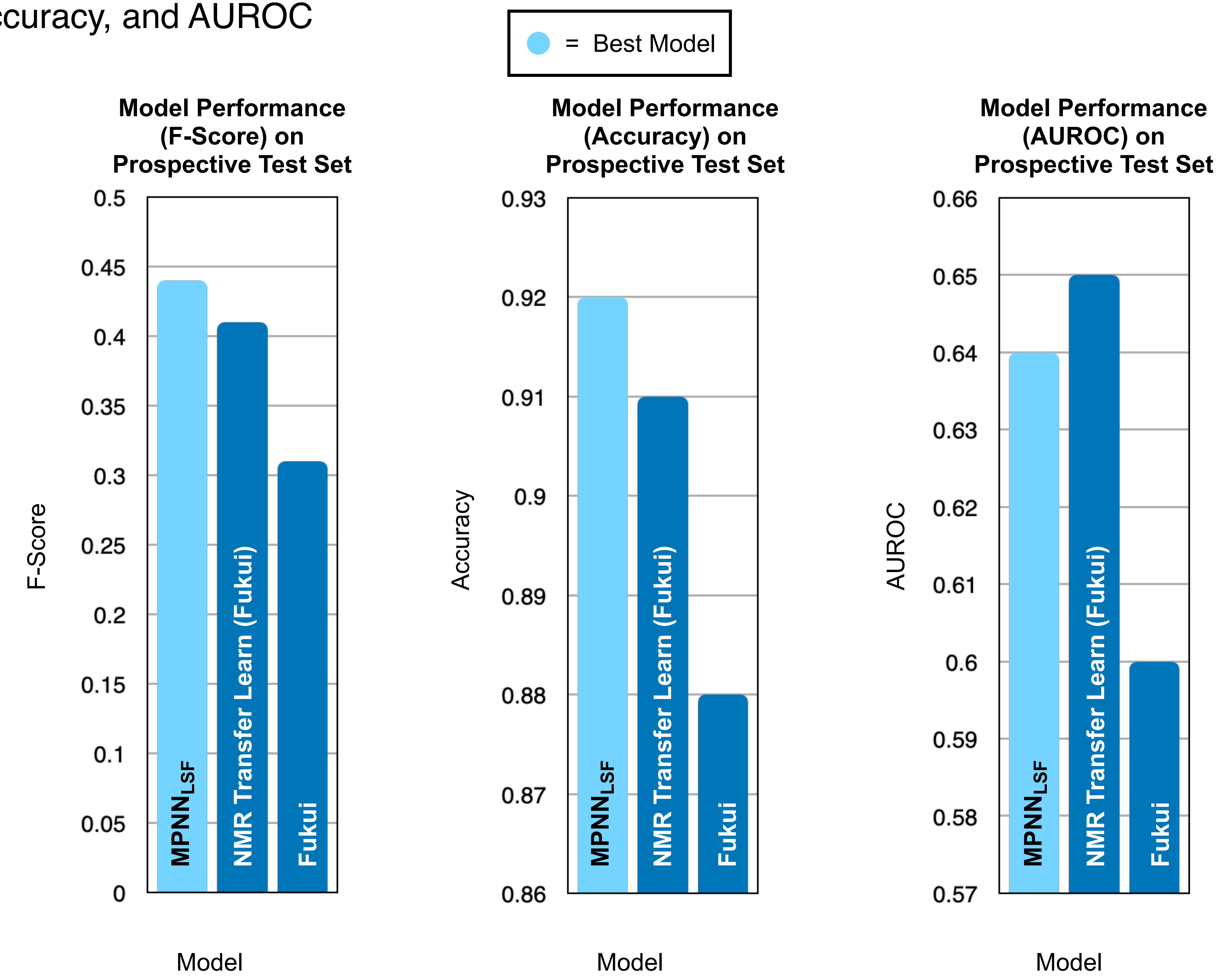
Nat. Commun. **2024**, *15*, 426.

Backup: F-Score, Accuracy, and AUROC



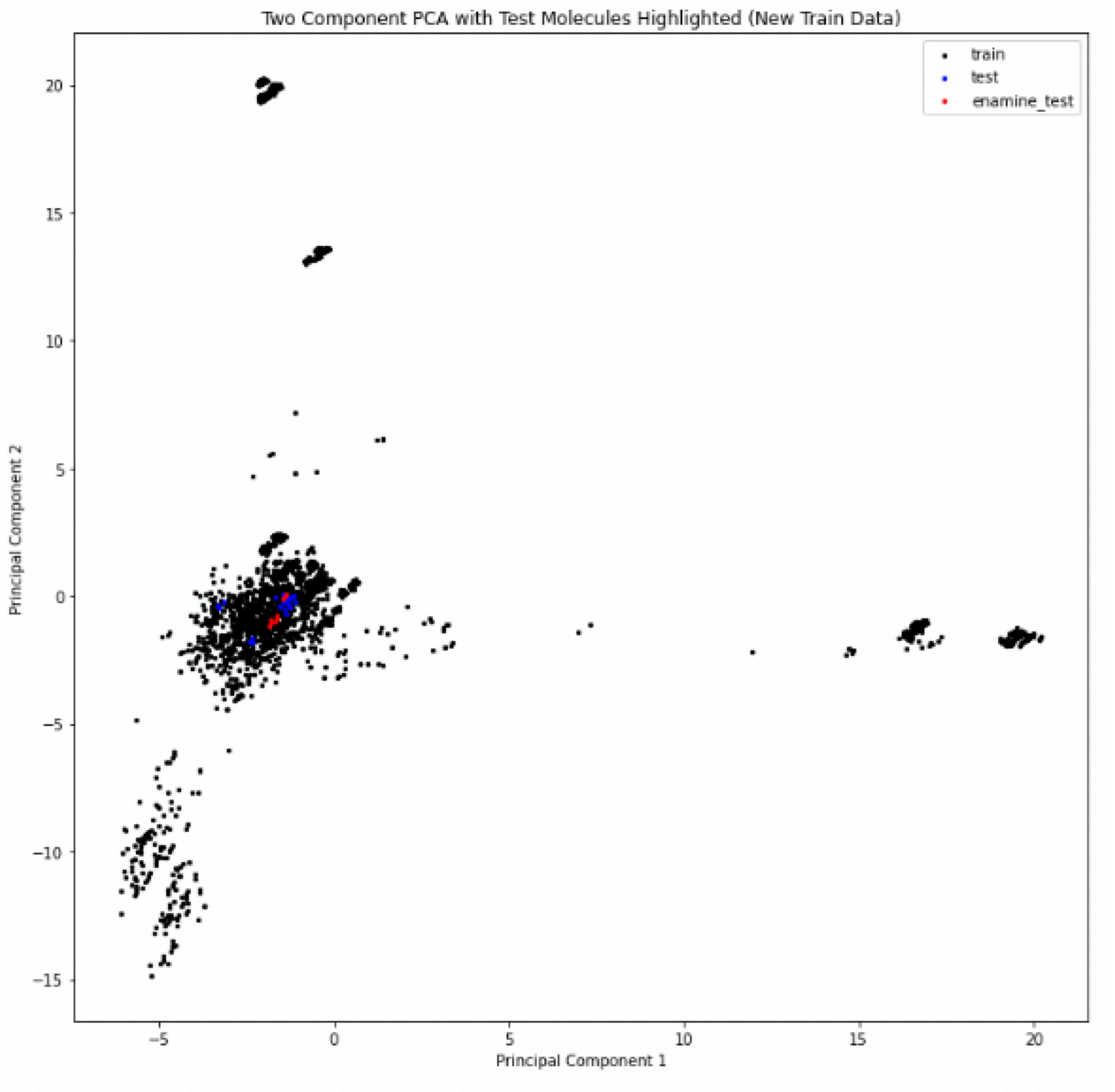
Nat. Commun. **2024**, *15*, 426.

Backup: F-Score, Accuracy, and AUROC



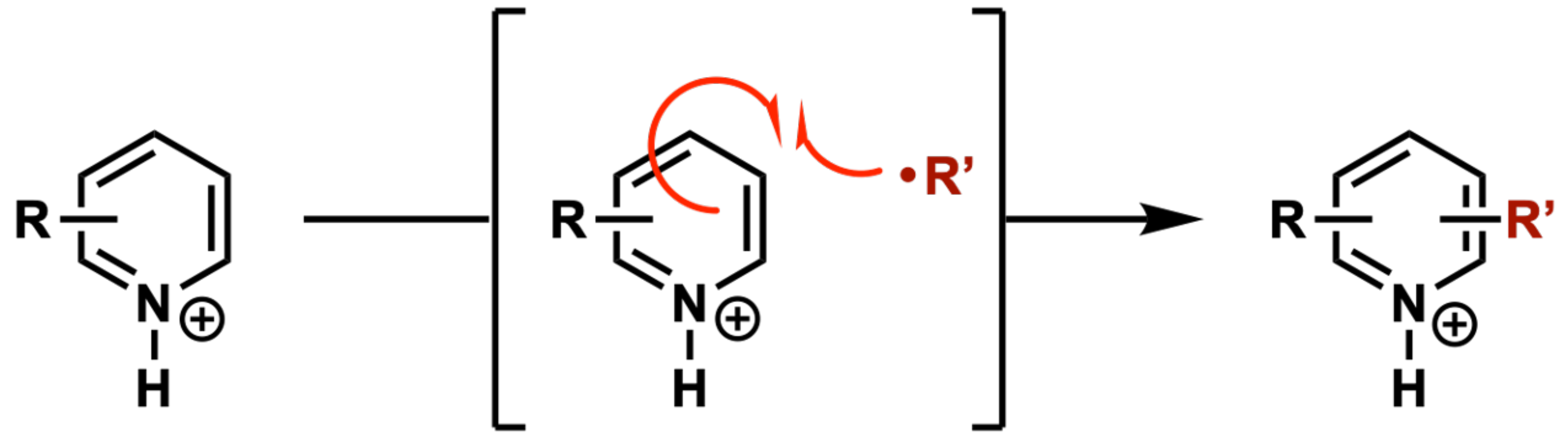
Nat. Commun. **2024**, *15*, 426.

Backup: PCA of the Dataset Chemical Space



Nat. Commun. **2024**, *15*, 426.

Backup: Fukui Indices



Fukui function-derived predictions

$$F_i(-) = q_i(N-1) - q_i(N)$$ (electrophilic radicals)

$$F_i(0) = \frac{q_i(N-1) - q_i(N+1)}{2}$$ (nucleophilic radicals)

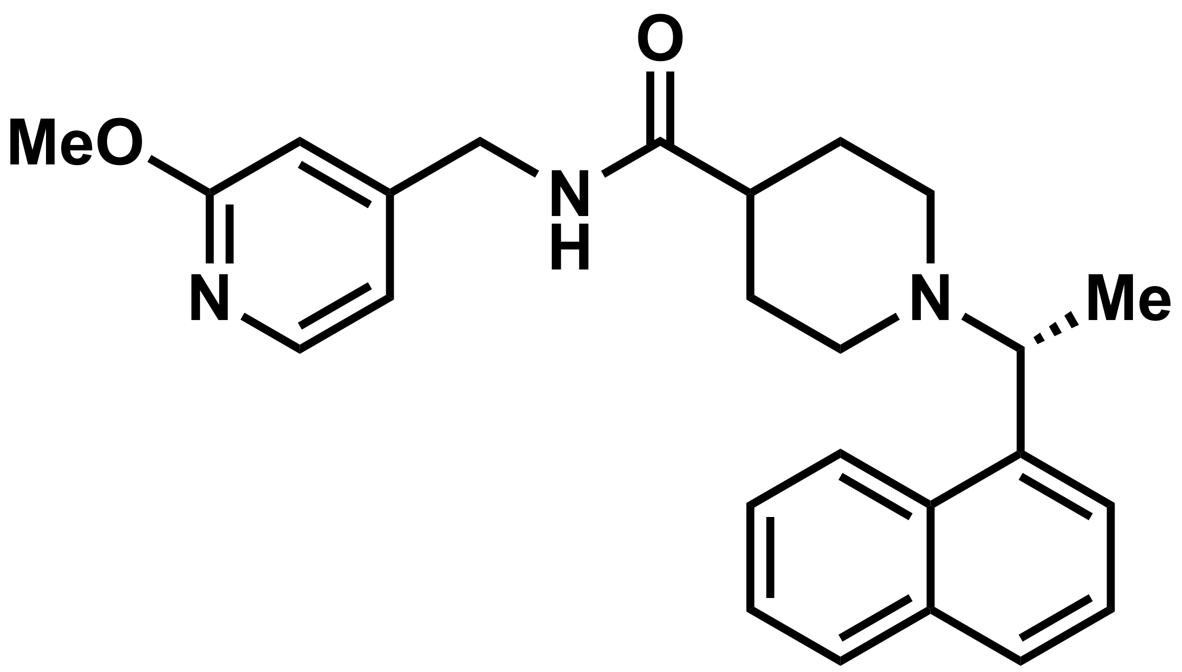
$q_i(N)$ = charge at atom i in a molecule with N electrons.

Nat. Commun. **2024**, *15*, 426.

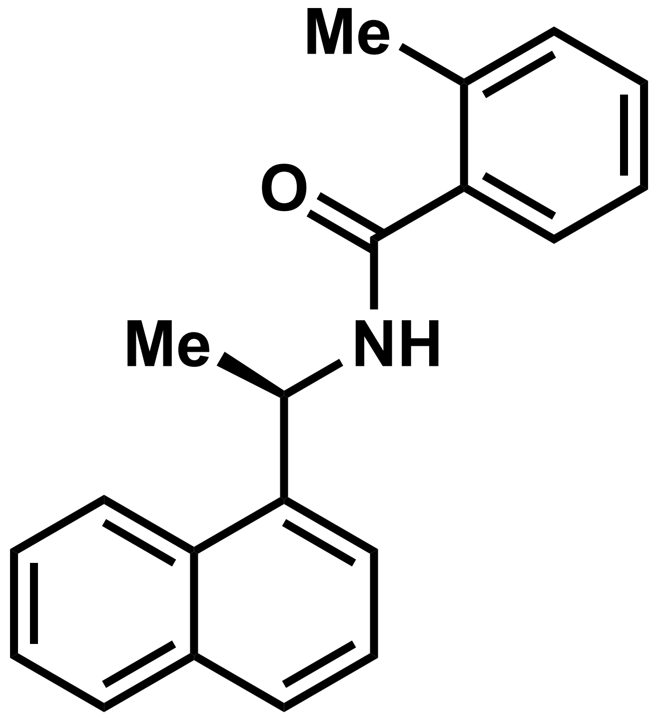
	F-Score / %	PVV / %	TPR / %	Accuracy / %
aGNN2D	38 (±5)	56 (±1)	30 (±6)	88 (±1)
aGNN2DQM	39 (±2)	54 (±2)	30 (±3)	87.6 (±0.3)
aGNN3D	59 (±3)	**62 (±2)**	56 (±4)	**90 (±1)**
aGNN3DQM	**60 (±4)**	**62 (±2)**	**59 (±6)**	**90 (±1)**

Nat. Chem. **2024**, *16*, 239.

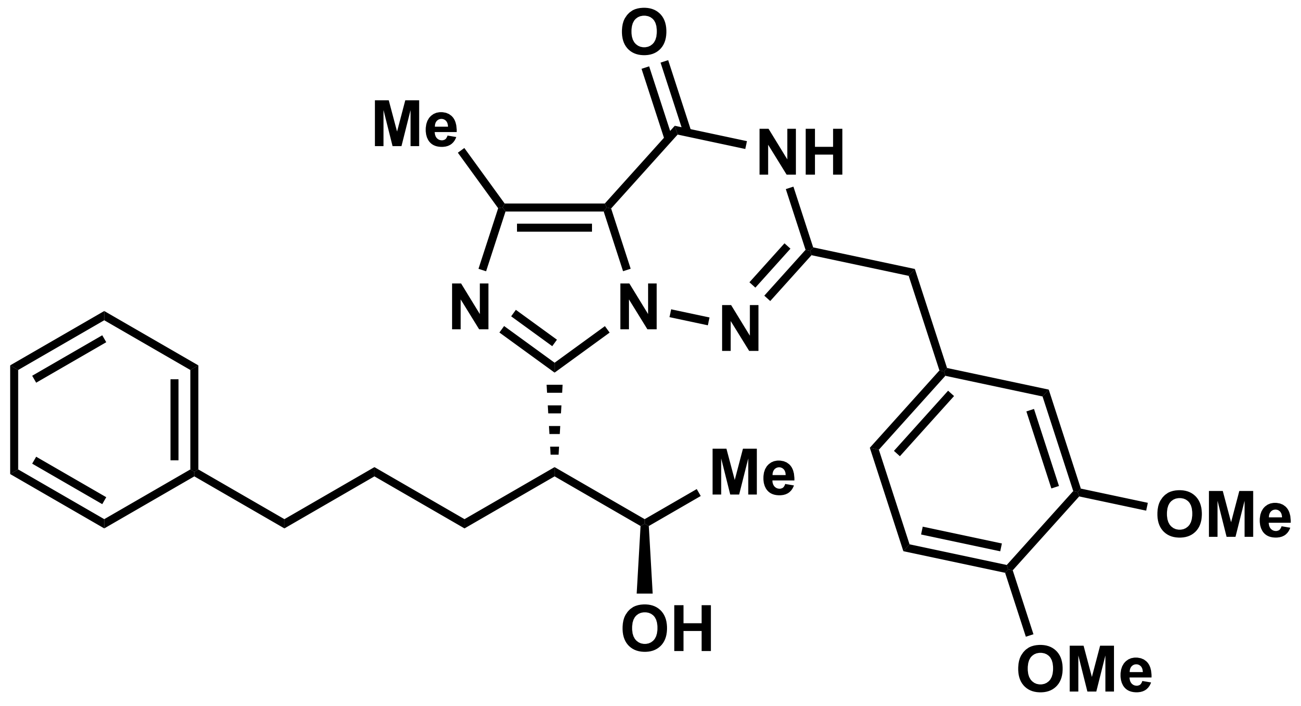
Backup: Representative P450 Test Set Molecules



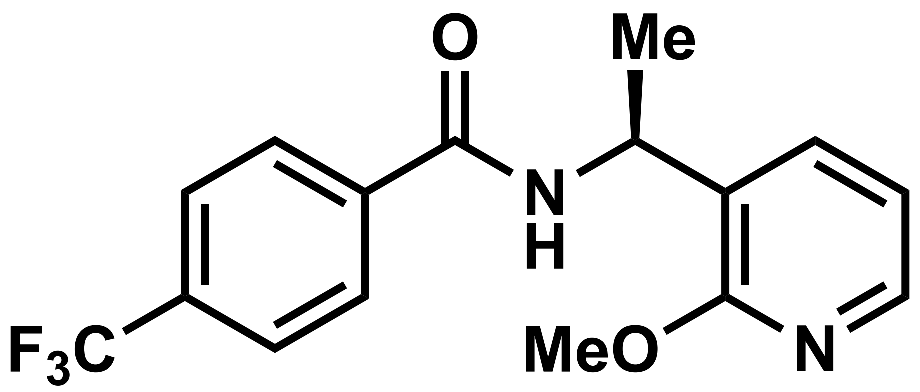
6



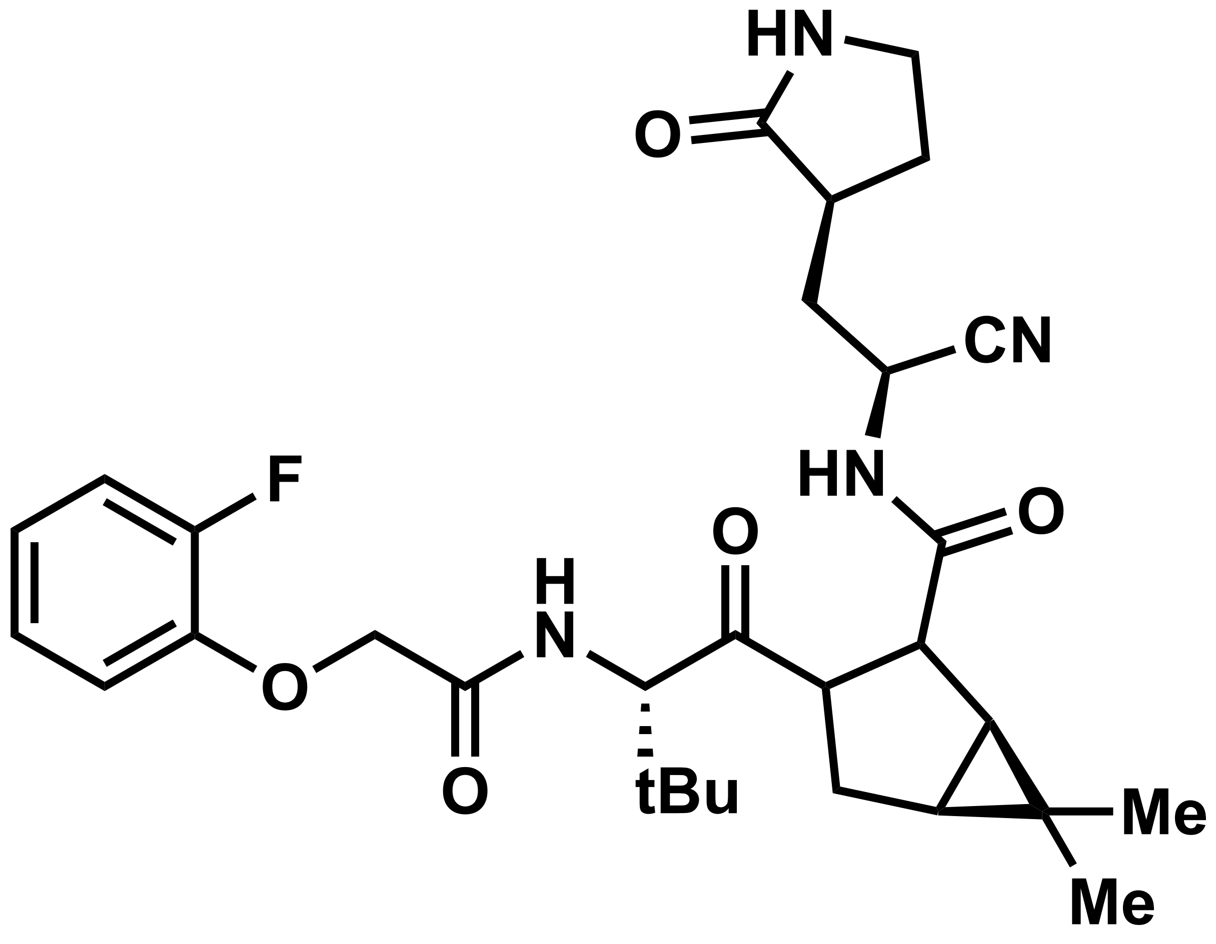
7



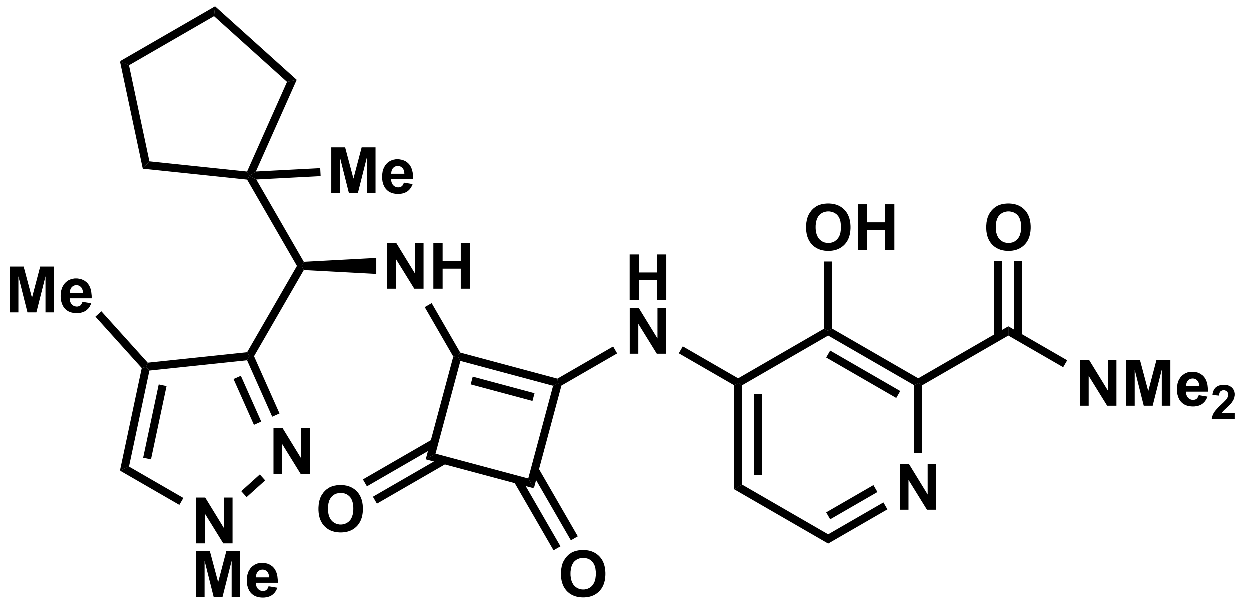
8



9

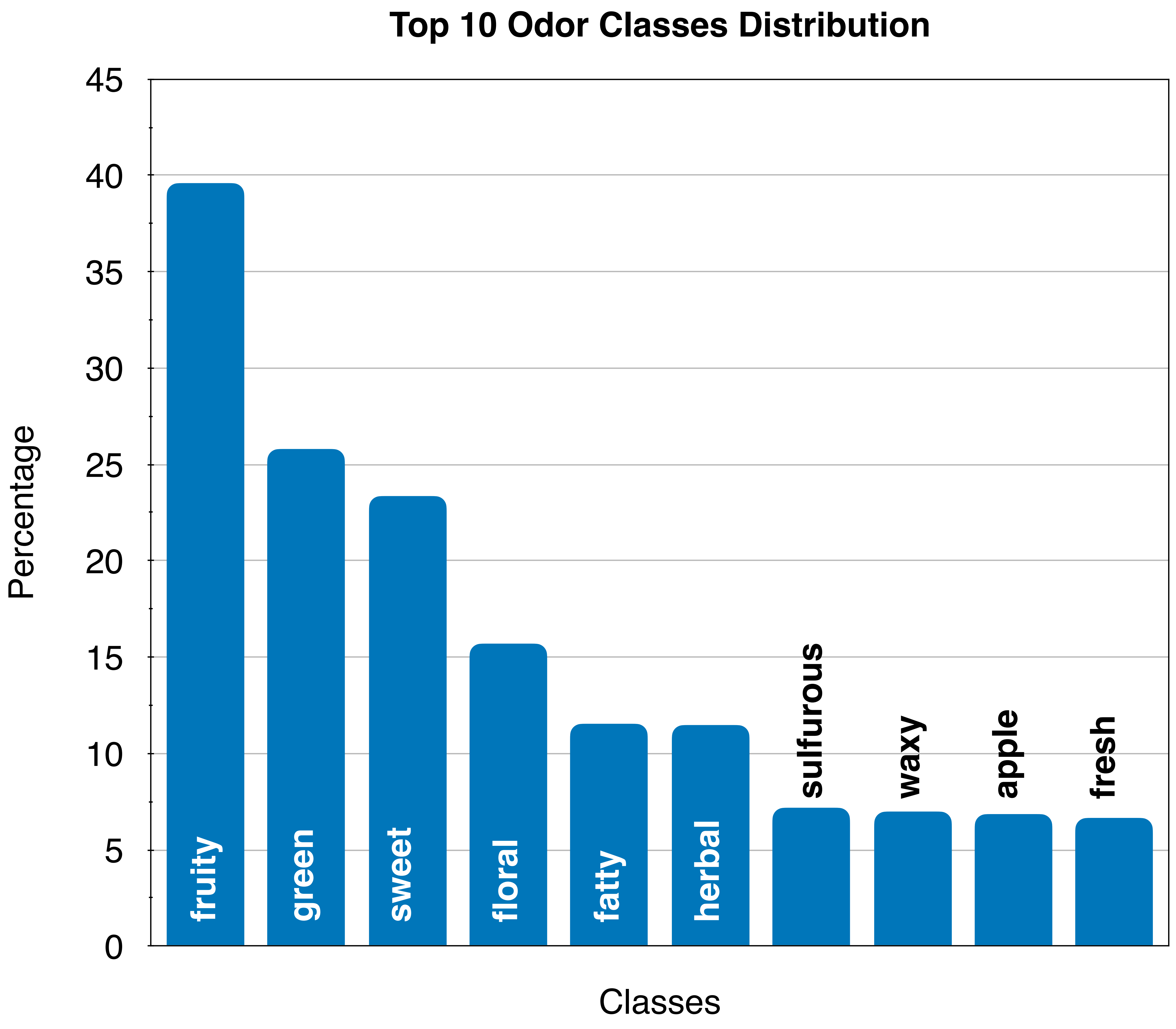


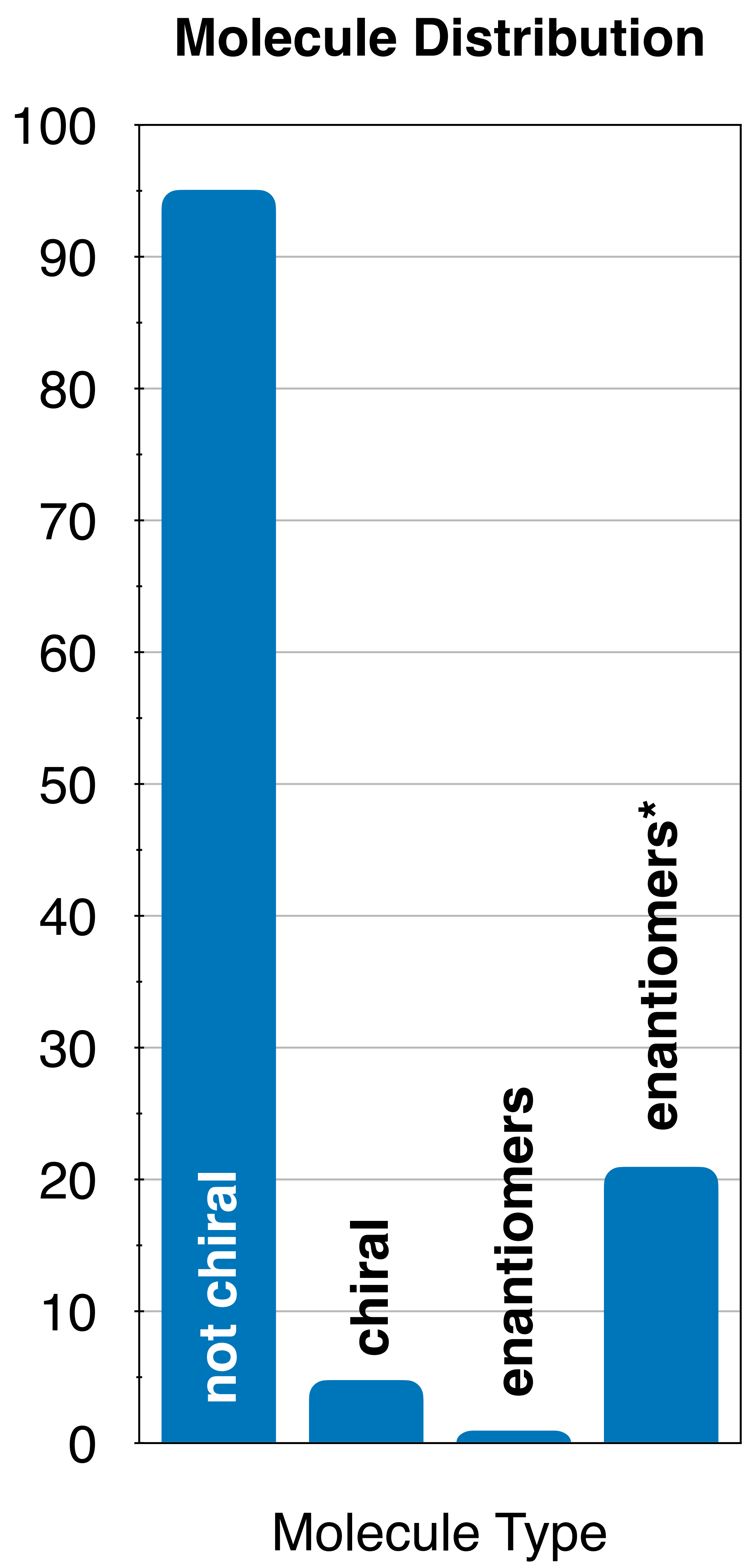
10



11

Fragrance Dataset Breakdown

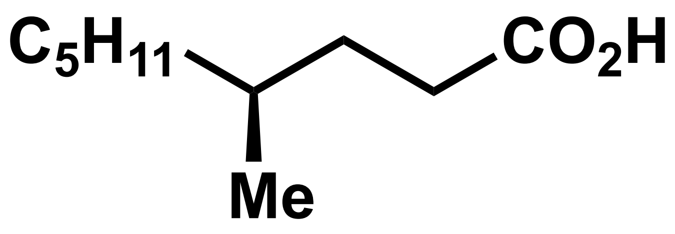




*enantiomers = % enantiomers in all chiral molecules

Chem. Sci. **2024**, *15*, 5143.

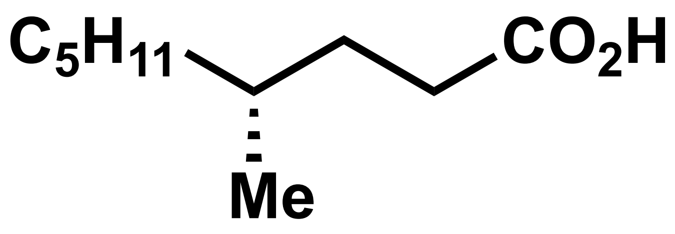
Enantiomeric Pairs Predictions: Similar Olfactive Notes



13

animal

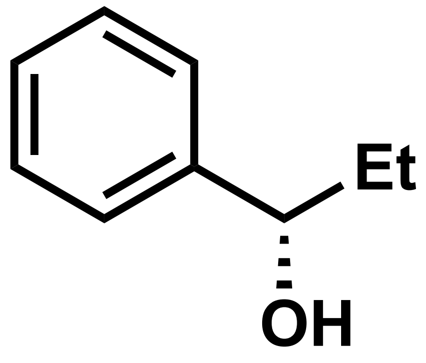
(—)



14

animal

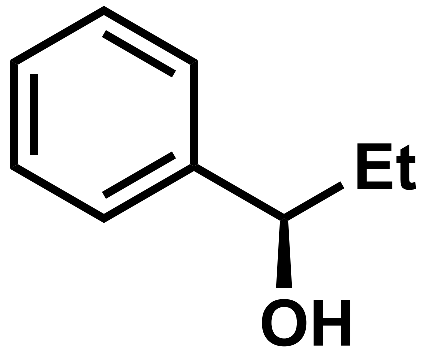
(—)



27

floral

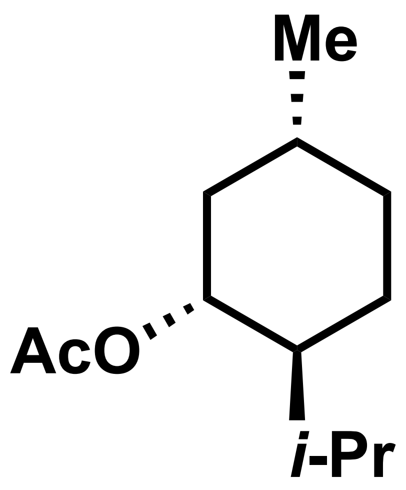
(—)



28

floral

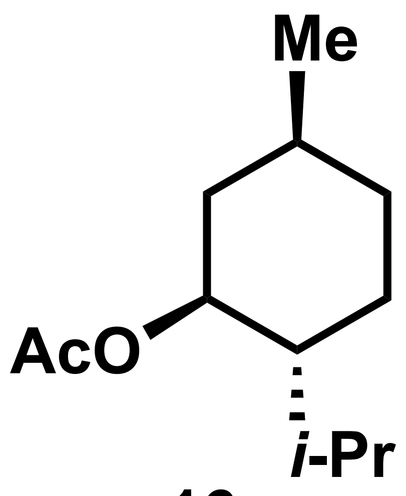
(—)



15

fresh, fruity, mint

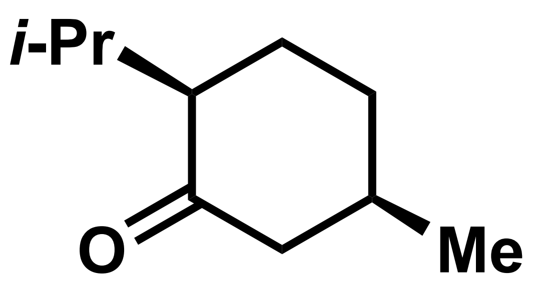
(fruity, mint)



16

fresh, fruity, mint

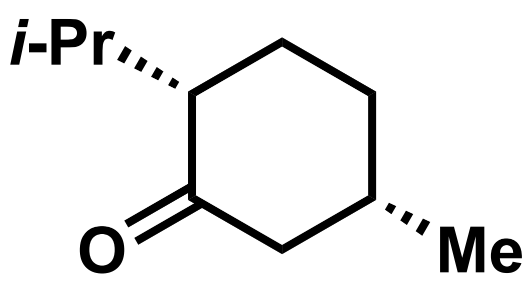
(fruity, mint)



(*R*)-isomethone (**17**)

mint

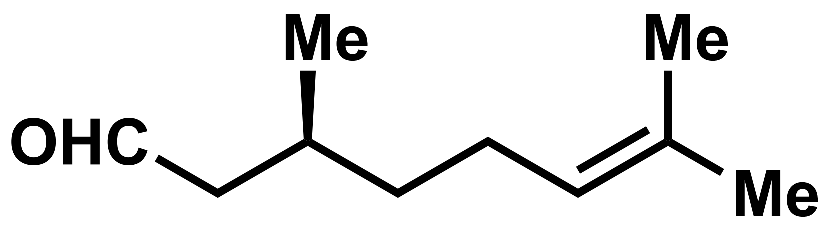
(mint)



(*S*)-isomethone (**18**)

mint

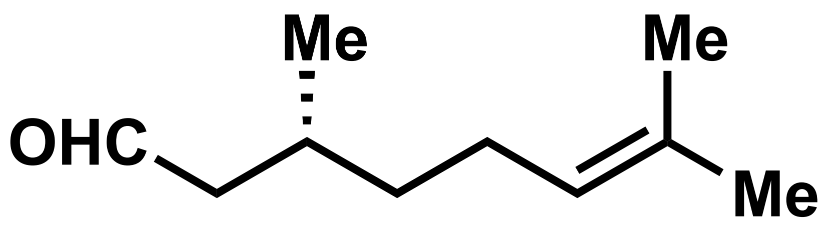
(mint)



(*S*)-citronellal (**25**)

citrus, fresh, herbal

(fresh, herbal)



(*R*)-citronellal (**26**)

citrus, fresh, herbal

(herbal)

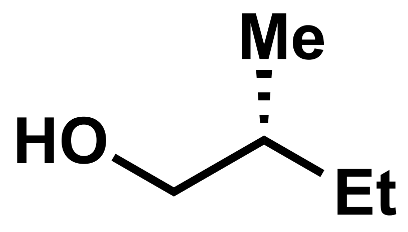
● = correctly identified similarity / dissimilarity in olfactive notes

● = incorrectly identified similarity / dissimilarity in olfactive notes

Chem. Sci. **2024**, *15*, 5143.

Enantiomeric Pairs Predictions: Dissimilar Olfactive Notes

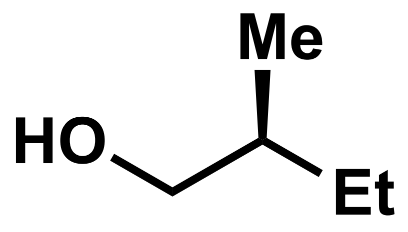
UNIVERSITY OF CAMBRIDGE



19

fatty, fermented

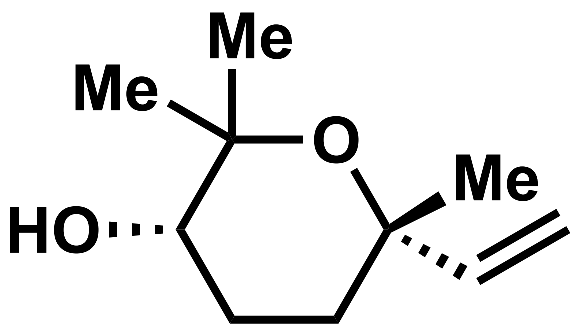
(cheesy*)



20

ethereal, fresh

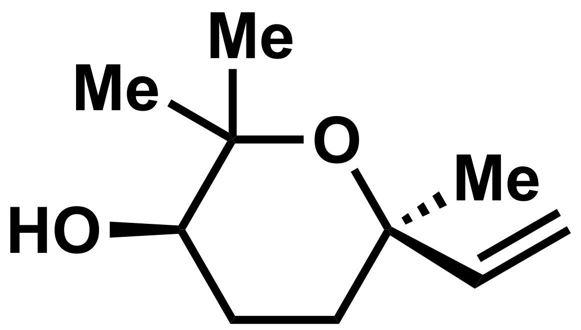
(ethereal)



33

dairy, floral, sweet

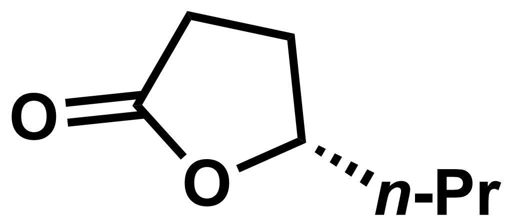
(—)



34

earthy

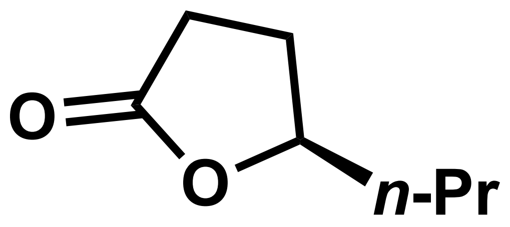
(—)



29

coconut, fruity, sweet

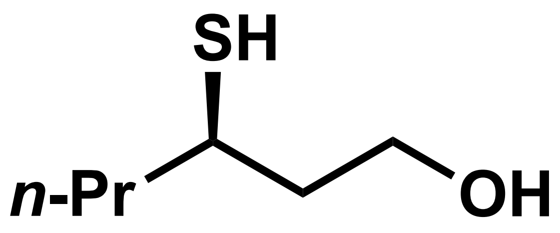
(—)



30

grassy, spicy, sweet, vanilla

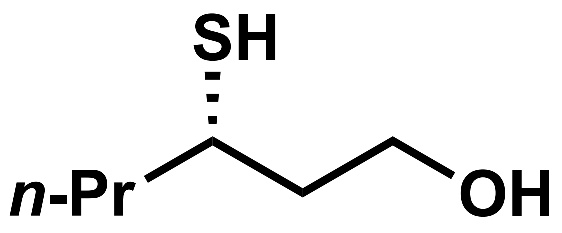
(—)



21

herbal, sulfurous

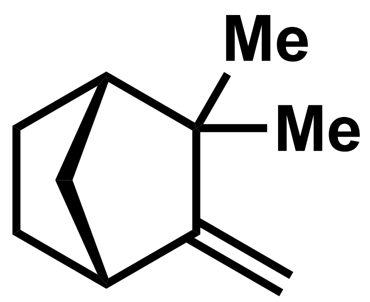
(sulfurous)



22

blackcurrant, fruity, sweet, topical, woody

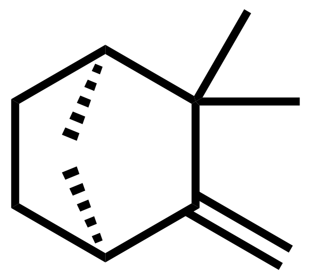
(sweet, fruity)



(*R*)-camphene (**31**)

balsamic, medicinal

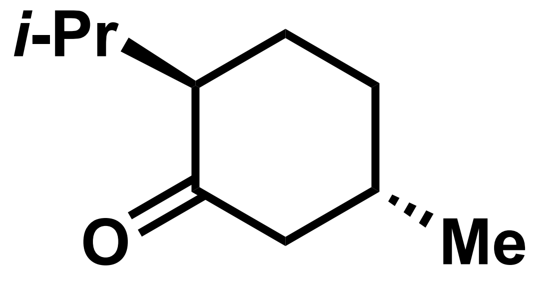
(—)



(*S*)-camphene (**32**)

camphoreous, pine

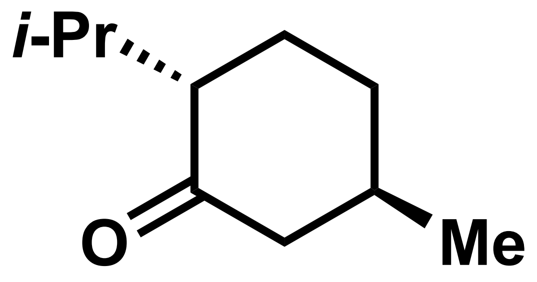
(camphoreous)



(*R*)-methone (**23**)

fresh, musty

(—)



(*S*)-methone (**24**)

camphoreous, fresh

(camphoreous)

● = correctly identified similarity / dissimilarity in olfactive notes

● = incorrectly identified similarity / dissimilarity in olfactive notes

Chem. Sci. **2024**, *15*, 5143.

Performance on CCDC Dataset

UNIVERSITY OF CAMBRIDGE

Model	Test Set MSE
Small MPNN	3.17
Large MPNN	2.93

Mean Squared Error (MSE) of total loss (bond distance loss + bond angles loss) on crystal structure data for a variety of message passing neural networks (MPNNs). Test set consisted of unseen molecules.

Chem. Sci. **2024**, *15*, 5143.

Per-Molecule Error Rate (Toxicity)

UNIVERSITY OF CAMBRIDGE

Compound	True Toxicity (log(mol kg^{-1}))	Crystal-Tox Predicted Toxicity (log(mol kg^{-1}))	Oloren ChemEngine Predicted Toxicity (log(mol kg^{-1}))
water	–0.70	1.53	1.98
sucrose	1.06	1.01	1.48
glucose	0.84	1.25	1.77
monosodium glutamate	1.00	1.66	2.10
THC	2.39	2.88	2.53
CBD	2.51	2.62	2.41
aconitine	6.90	3.84	3.38
epibatidine	7.43	2.88	2.93
MDMA	3.08	2.59	2.55
cocaine	3.50	2.09	2.67
LSD	4.29	2.65	2.89
heroin	4.23	2.80	3.19

Chem. Sci. **2024**, *15*, 5143.

Per-Molecule Set Error Rate (Yields)

Model	Split MAE			
	Halide Set 0	*Halide Set 1*	*Halide Set 2*	*Halide Set 3*
Random Forest	23.6	23.9	22.2	31.0
Gaussian Process	27.3	25.2	21.7	30.9
Adaboost	24.6	23.9	18.7	31.6
Yield-BERT	27.3	25.2	21.7	30.9
GraphRXN	**9.5**	41.6	30.9	**18.7**
Crystal-Yield	26.7	**14.8**	**16.3**	27.5
	Base 0	*Base 1*	*Base 2*	
Random Forest	32.0	32.4	19.9	
Gaussian Process	31.0	34.3	24.8	
Adaboost	27.2	29.5	19.9	
Yield-BERT	23.3	27.4	22.1	
GraphRXN	**12.8**	27.1	13.8	
Crystal-Yield	13.9	**13.0**	**13.4**	
	Ligand 0	*Ligand 1*	*Ligand 2*	*Ligand 3*
Random Forest	27.4	29.0	27.6	29.8
Gaussian Process	39.8	32.2	29.2	30.6
Adaboost	26.8	29.9	25.9	27.2
Yield-BERT	20.4	24.0	25.8	27.0
GraphRXN	**9.7**	17.6	12.7	15.2
Crystal-Yield	24.5	23.4	10.4	14.5
Crystal-Yield[a]	17.1	**12.2**	**6.5**	**10.8**
	Additive Set 0	*Additive Set 1*	*Additive Set 2*	*Additive Set 3*
Random Forest	34.0	31.3	26.7	29.4
Gaussian Process	32.7	29.0	24.5	27.9
Adaboost	29.0	27.3	26.7	27.5
Yield-BERT	25.2	22.9	22.8	25.3
GraphRXN	16.7	15.2	22.8	**15.4**
Crystal-Yield	**15.6**	**16.6**	**17.2**	15.5

Chem. Sci. **2024**, *15*, 5143.

Ranked Elements in Training Foundational Model

most common elements in dataset

C N O F S Cl P B S Cl P B Si Br Cu I Fe Zn Co

Os Re Au Rh K W Na Al Pt Ag Li Sn Pd Se Mo Mn Ru Cd Ni

V Cr Ti Pb Eu Ir As U Ga Ge Mg Sb Tb Gd Zr Te Hg Dy Nd

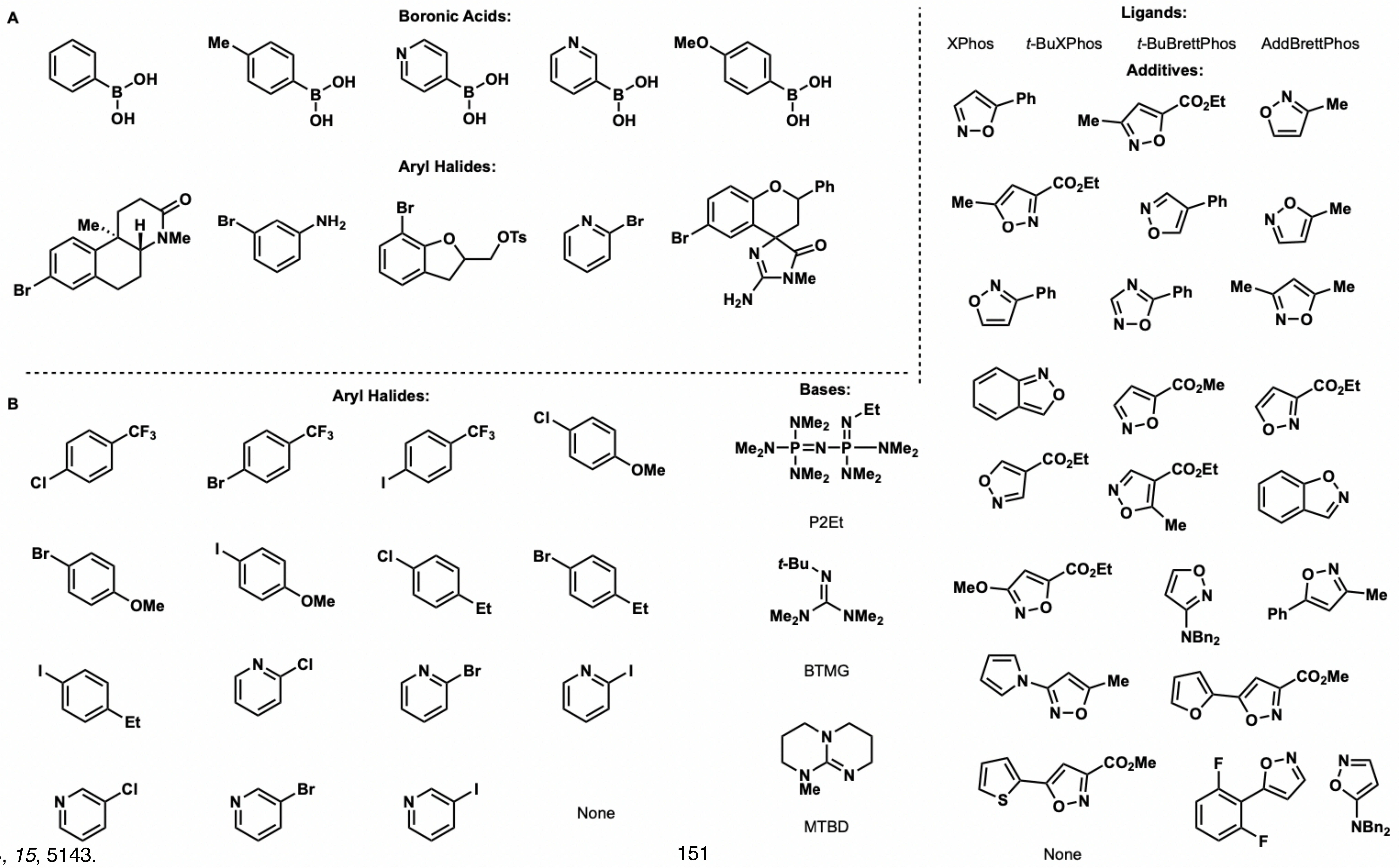
Tl Rb Sc Nb Ta Ho Cs Sr Yb Ce Ba Er Y Pr Bi Sm In La Ca

least common elements in dataset

Lu Hf Tm Th Be Tc Np Pu

Chem. Sci. **2024**, *15*, 5143.

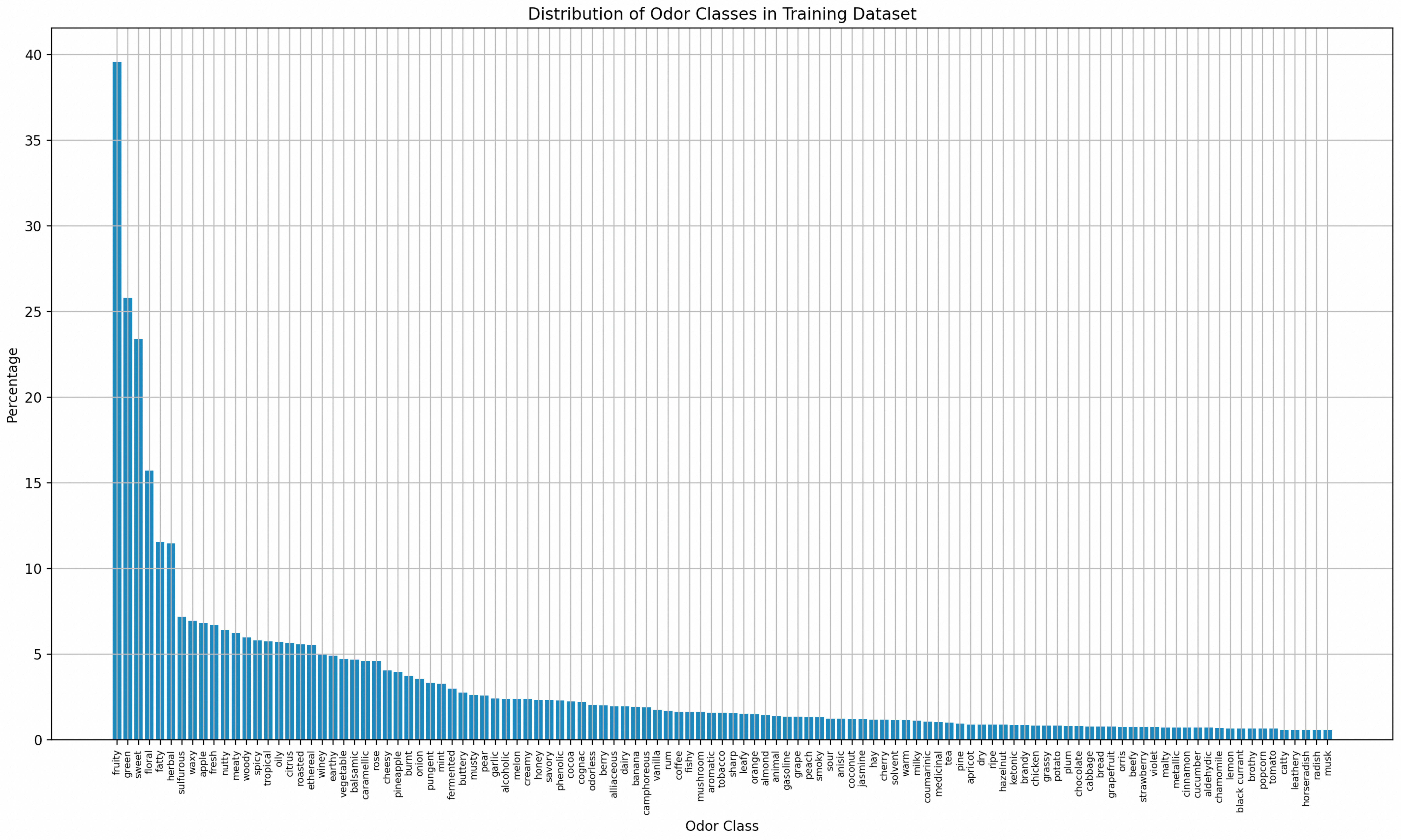
Selected Molecules in Yield Datasets



Chem. Sci. **2024**, *15*, 5143.

Selected Molecules in Yield Datasets

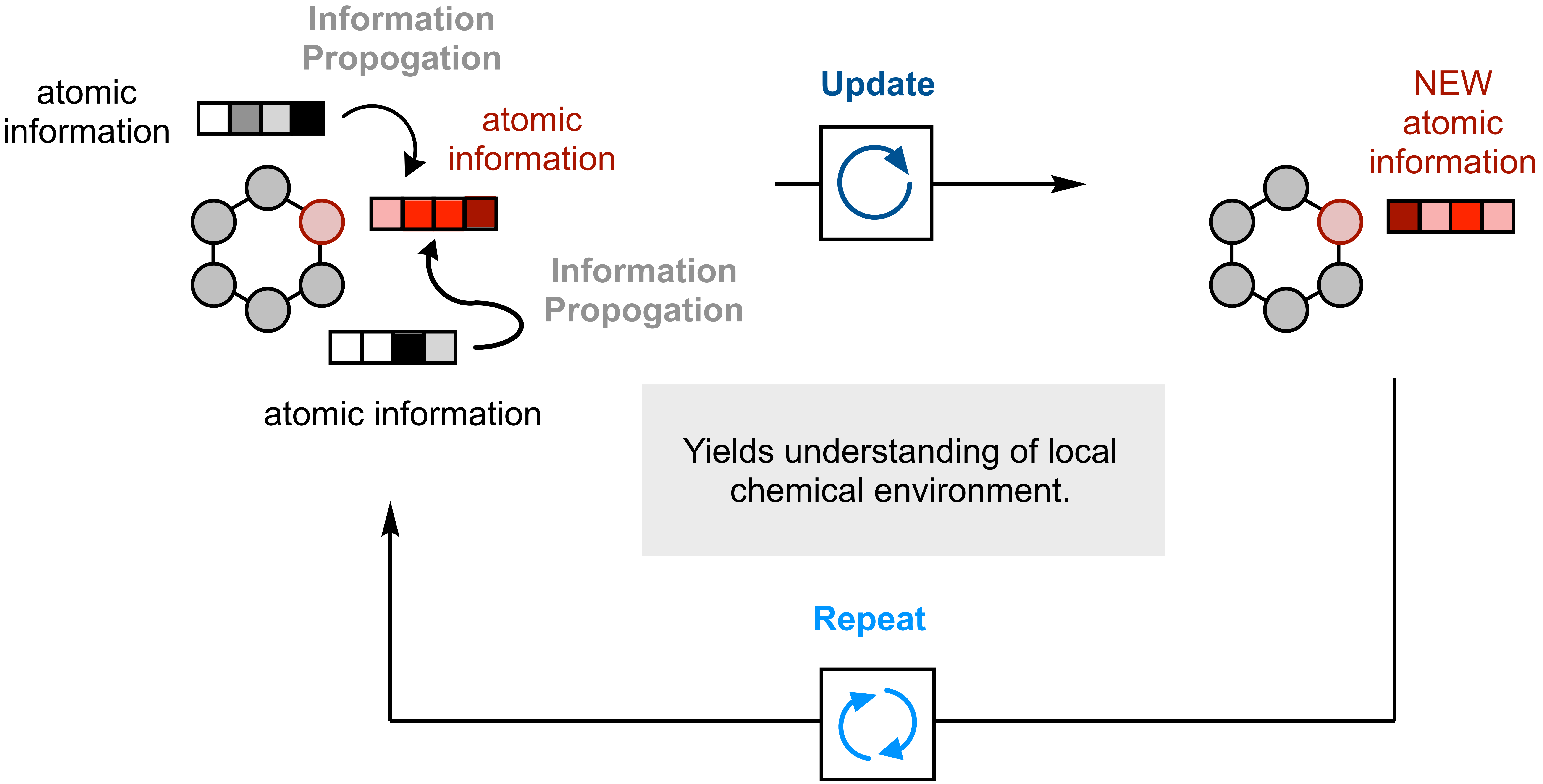
UNIVERSITY OF CAMBRIDGE



Chem. Sci. **2024**, *15*, 5143.

How Does the MPNN Work?

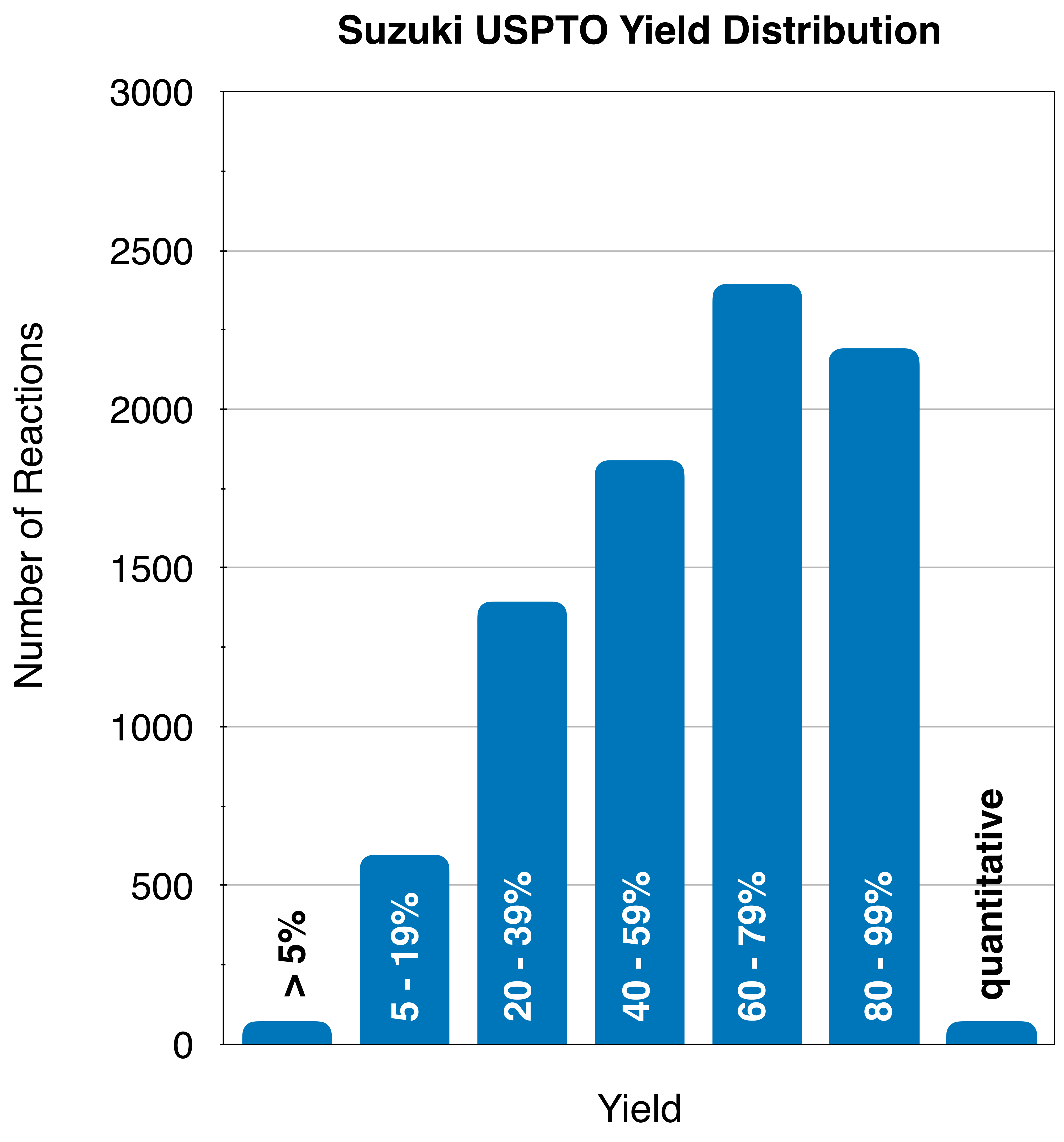
UNIVERSITY OF CAMBRIDGE

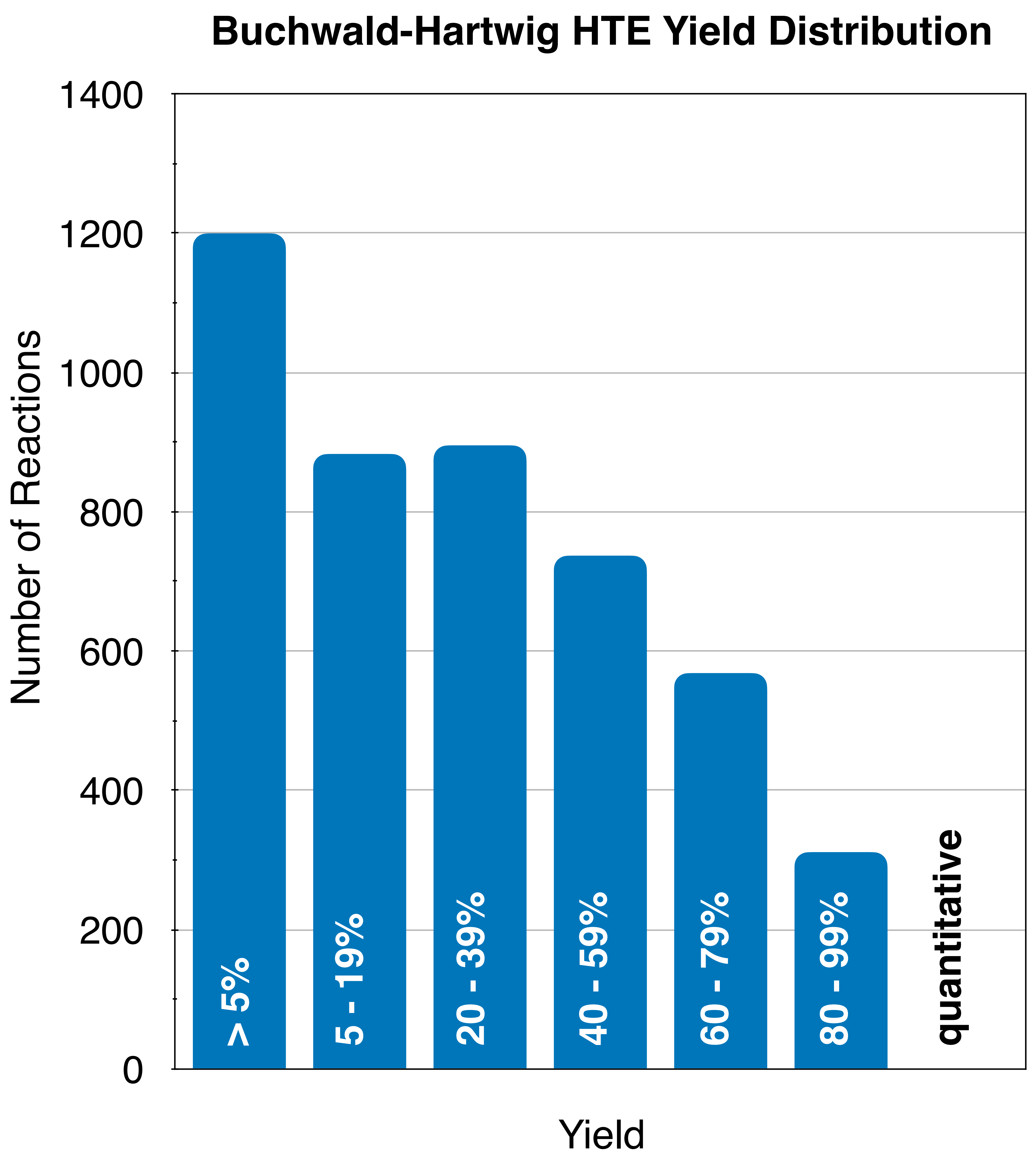


Yields understanding of local chemical environment.

Nat. Commun. **2024**, *15*, 426.

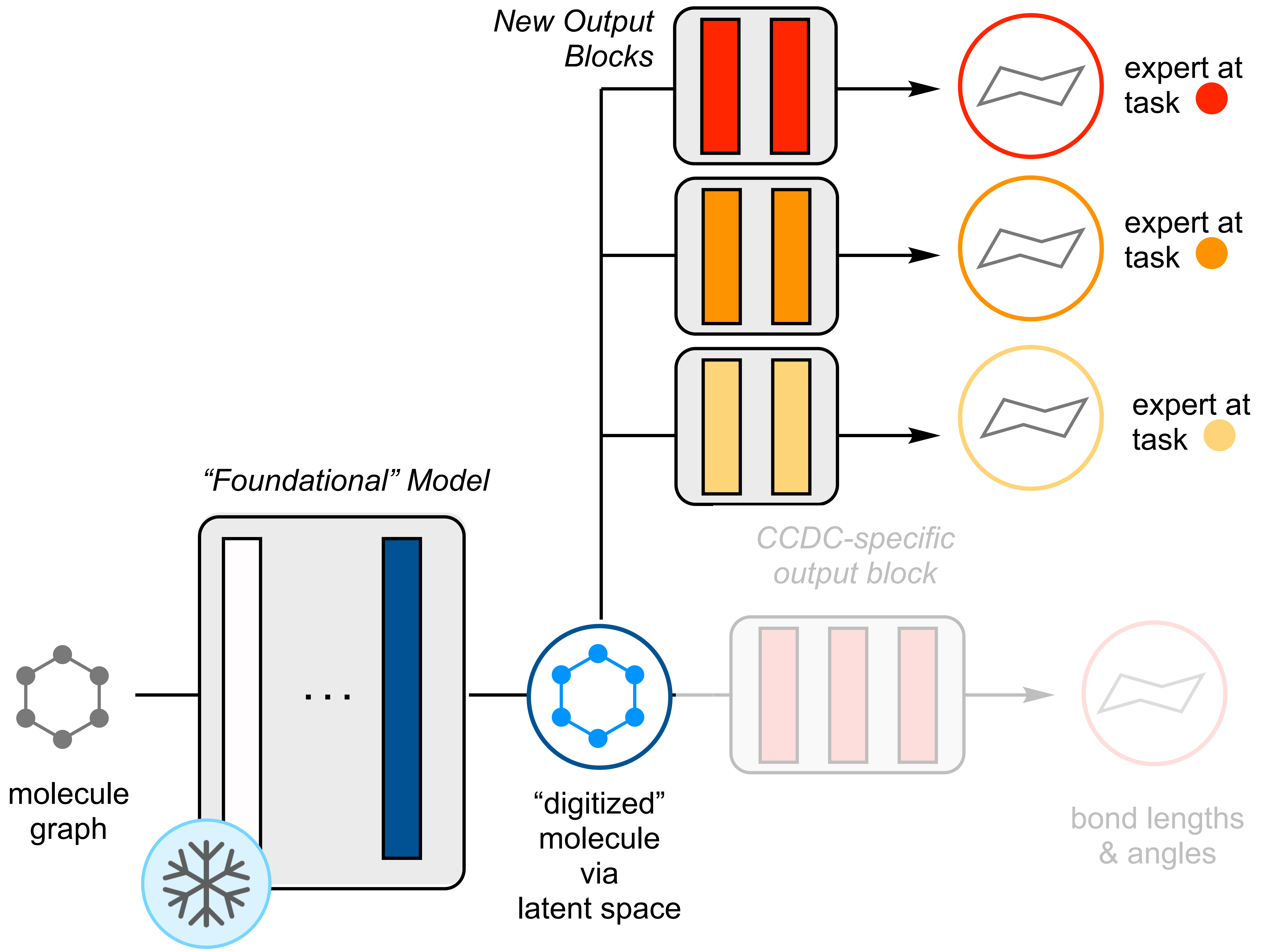
Yield Prediction





J. Am. Chem. Soc. **2022**, *144*, 4819. *Science* **2018**, *360*, 186.

The Big Idea



New Output Blocks
expert at task
expert at task
expert at task
"Foundational" Model
CCDC-specific output block
. . .
molecule graph
"digitized" molecule via latent space
bond lengths & angles