Uncertainty in molecular deep learning

Alpha Lee

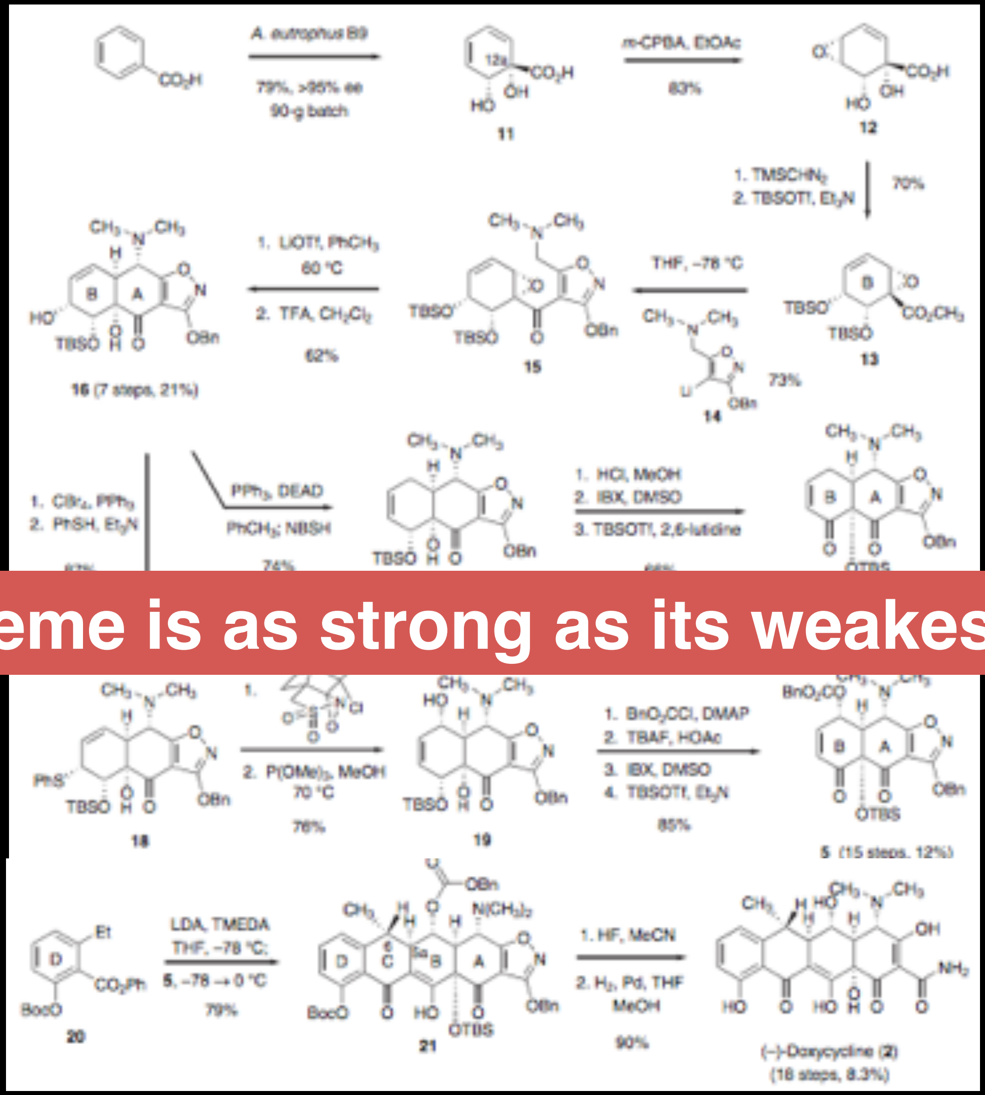
Department of Physics, University of Cambridge
aal44@cam.ac.uk

Two stories on uncertainty

- **Reaction prediction and uncertainty calibration**
- Bayesian graph neural networks for uncertainty quantification

The challenge of synthesis

- It is all well and good to suggest promising hits *in silico*, but making molecules is an unsolved challenge



A scheme is as strong as its weakest link!

Charest et al., *Science*, 308, 395 (2005)

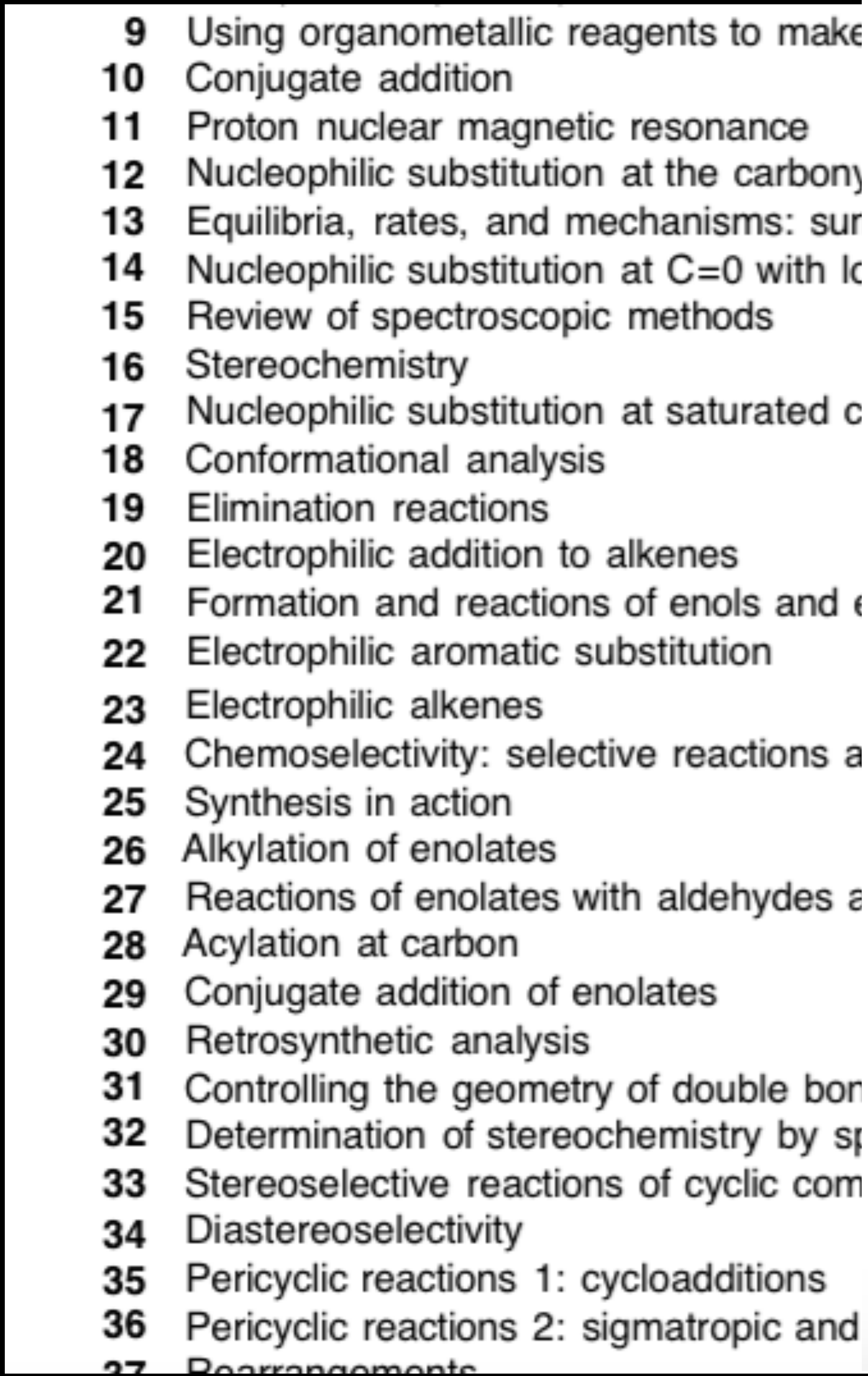
The law of compound interest

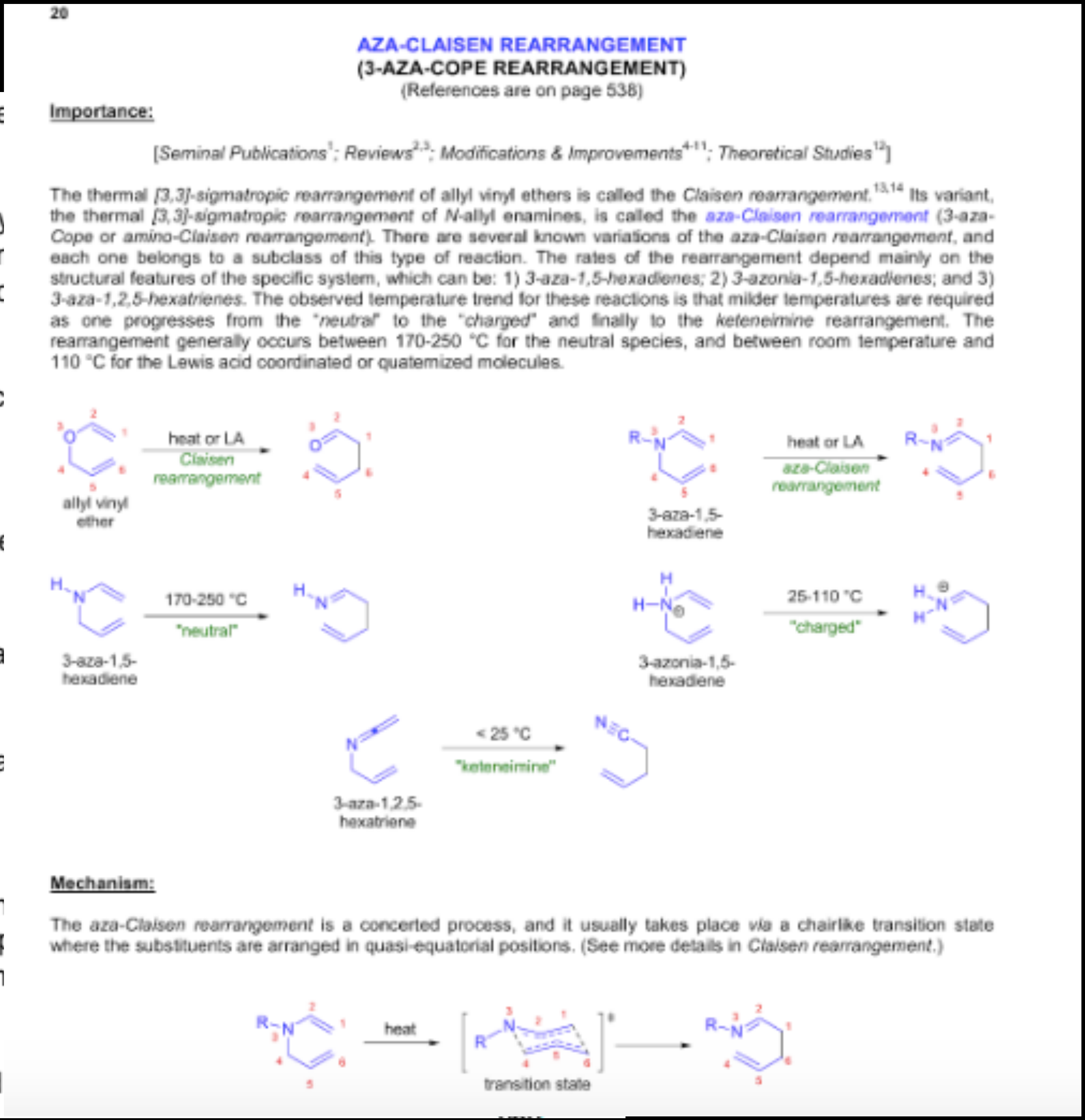
$$A \to B \to C \to D \to E$$

Probability of each prediction being right: p

Probability of a N-step scheme being right (assuming independence): p^N

A chemist's view of reactions





AZA-CLAISEN REARRANGEMENT
(3-AZA-COPE REARRANGEMENT)

(References are on page 538)

Importance:

[Seminal Publications[1]; Reviews[2,3]; Modifications & Improvements[4-11]; Theoretical Studies[12]]

The thermal *[3,3]-sigmatropic rearrangement* of allyl vinyl ethers is called the *Claisen rearrangement.*[13,14] Its variant, the thermal *[3,3]-sigmatropic rearrangement* of *N*-allyl enamines, is called the *aza-Claisen rearrangement* (*3-aza-Cope* or *amino-Claisen rearrangement*). There are several known variations of the *aza-Claisen rearrangement*, and each one belongs to a subclass of this type of reaction. The rates of the rearrangement depend mainly on the structural features of the specific system, which can be: 1) *3-aza-1,5-hexadienes;* 2) *3-azonia-1,5-hexadienes*; and 3) *3-aza-1,2,5-hexatrienes*. The observed temperature trend for these reactions is that milder temperatures are required as one progresses from the "*neutral*" to the "*charged*" and finally to the *keteneimine* rearrangement. The rearrangement generally occurs between 170-250 °C for the neutral species, and between room temperature and 110 °C for the Lewis acid coordinated or quaternized molecules.

Mechanism:

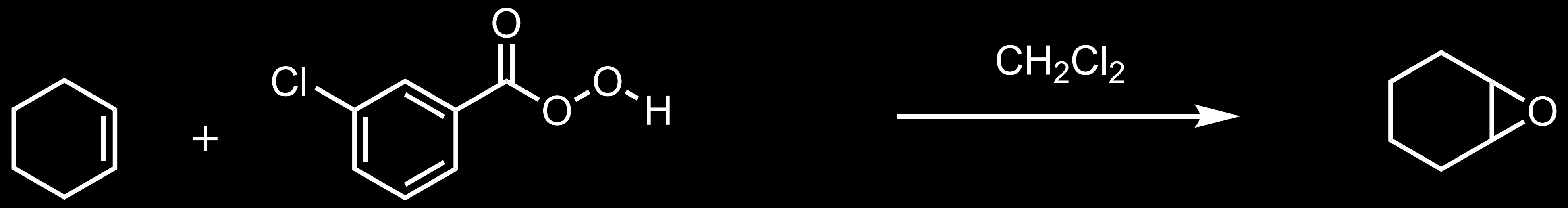
The *aza-Claisen rearrangement* is a concerted process, and it usually takes place *via* a chairlike transition state where the substituents are arranged in quasi-equatorial positions. (See more details in *Claisen rearrangement*.)

Can we infer chemical reactivity by correlation analysis?

reactants

reagent & conditions

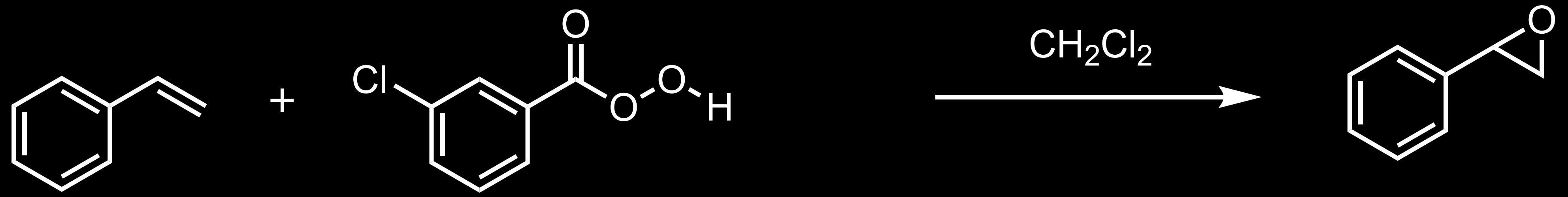
product



C1=CCCCC1.O=C(OO)c1cccc(Cl)c1

ClCCl

C1CCC2OC2C1

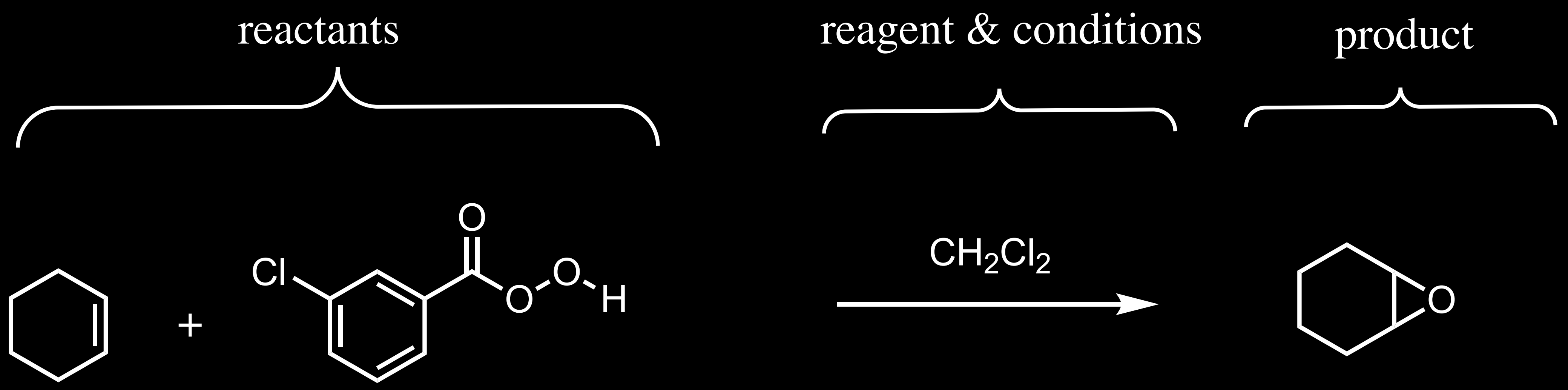


C=Cc1ccccc1.O=C(OO)c1cccc(Cl)c1

ClCCl

c1ccc(C2CO2)cc1

Can we infer chemical reactivity by correlation analysis?



reactants

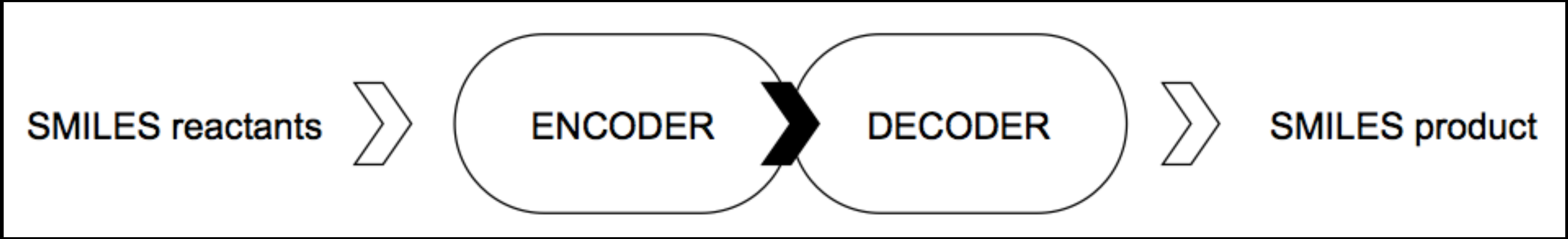
reagent & conditions

product

CH_2Cl_2

C1=CCCCC1.O=C(OO)c1cccc(Cl)c1.ClCCl	C1CCC2OC2C1
C=Cc1ccccc1.O=C(OO)c1cccc(Cl)c1.ClCCl	c1ccc(C2CO2)cc1
CC=C(C)C.O=C(OO)c1cccc(Cl)c1	CC1OC1(C)C
C/C=C(/C)c1ccccc1.O=C(OO)c1cc(Cl)cc(Cl)c1	CC1OC1(C)c1ccccc1

A machine translation approach



- Taking a leaf out of Google Translate's book
- Chemistry-specific knowledge: new architecture for long-ranged token-token correlations
- **rdkit SMILES canonisation**
- Augment dataset by simple reaction template to strengthen the model's performance on "simple reactions"

Predicting reaction by correlation of SMILES tokens

Model benchmarked on a freely available set of reactions reported in US patents (~500,000 reactions)

	Jin et al. (2017)	Schwaller et al. (2018)	Bradshaw et al. (2018)	Our work
Test set accuracy (%)	79.6	80.3	87.0 (after eliminating all complex reactions)	89.1

Explicitly considering reaction centre

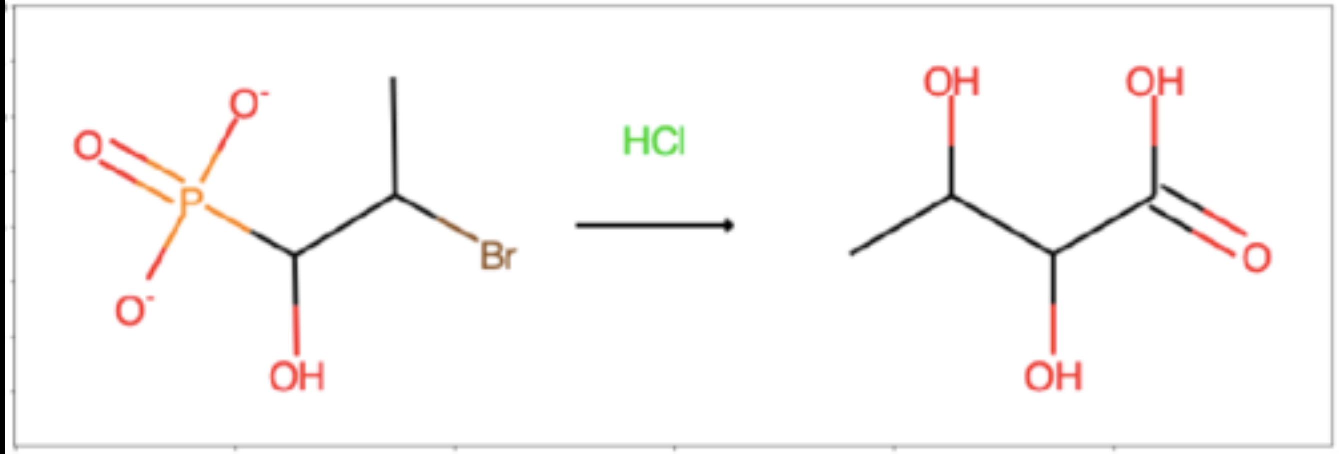
Separating reagent and reactants

Predicting electron paths

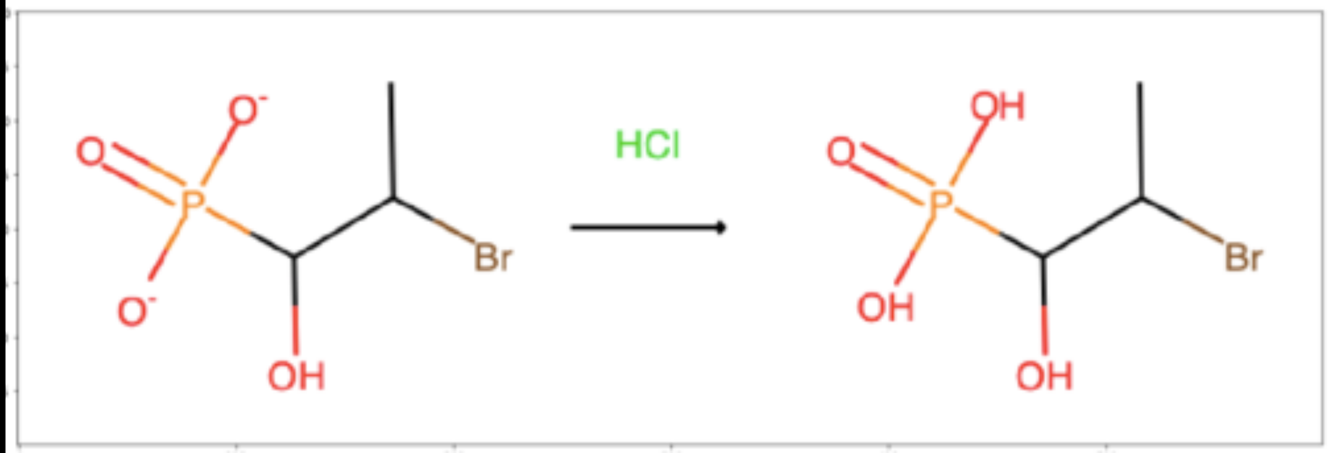
Learning the grammar of chemistry is important!

Digital alchemy….

Schwaller et al. (2018)



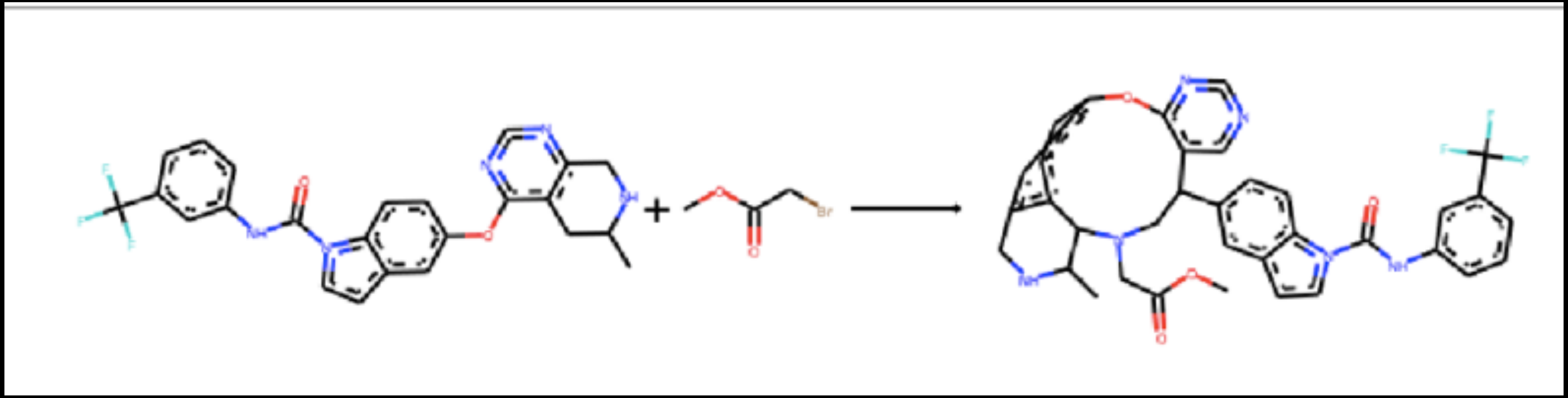
Our model



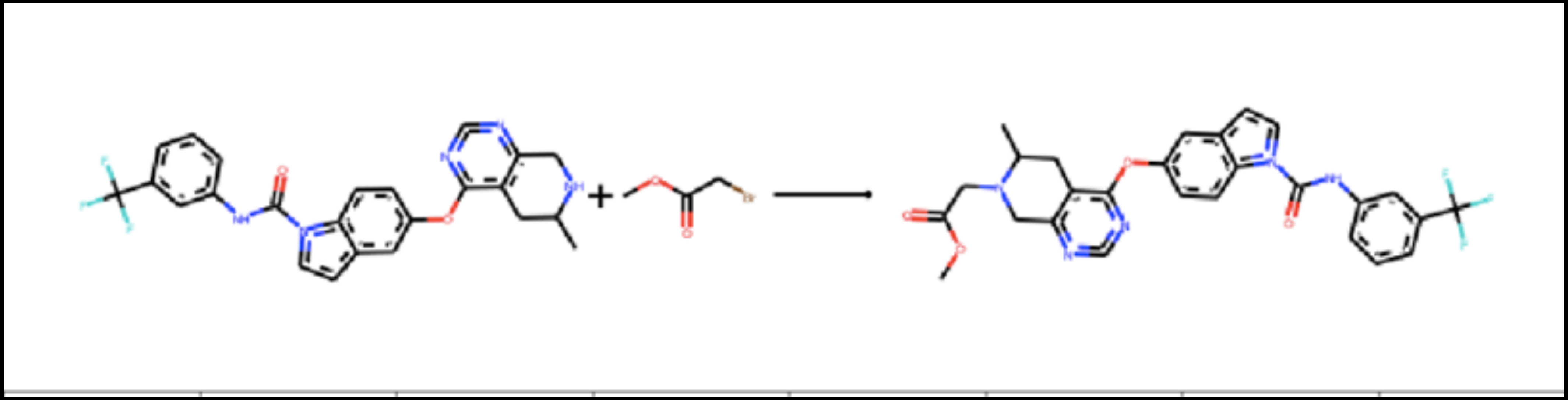
Learning the grammar of chemistry is important!

Handling big molecules

Schwaller et al. (2018)

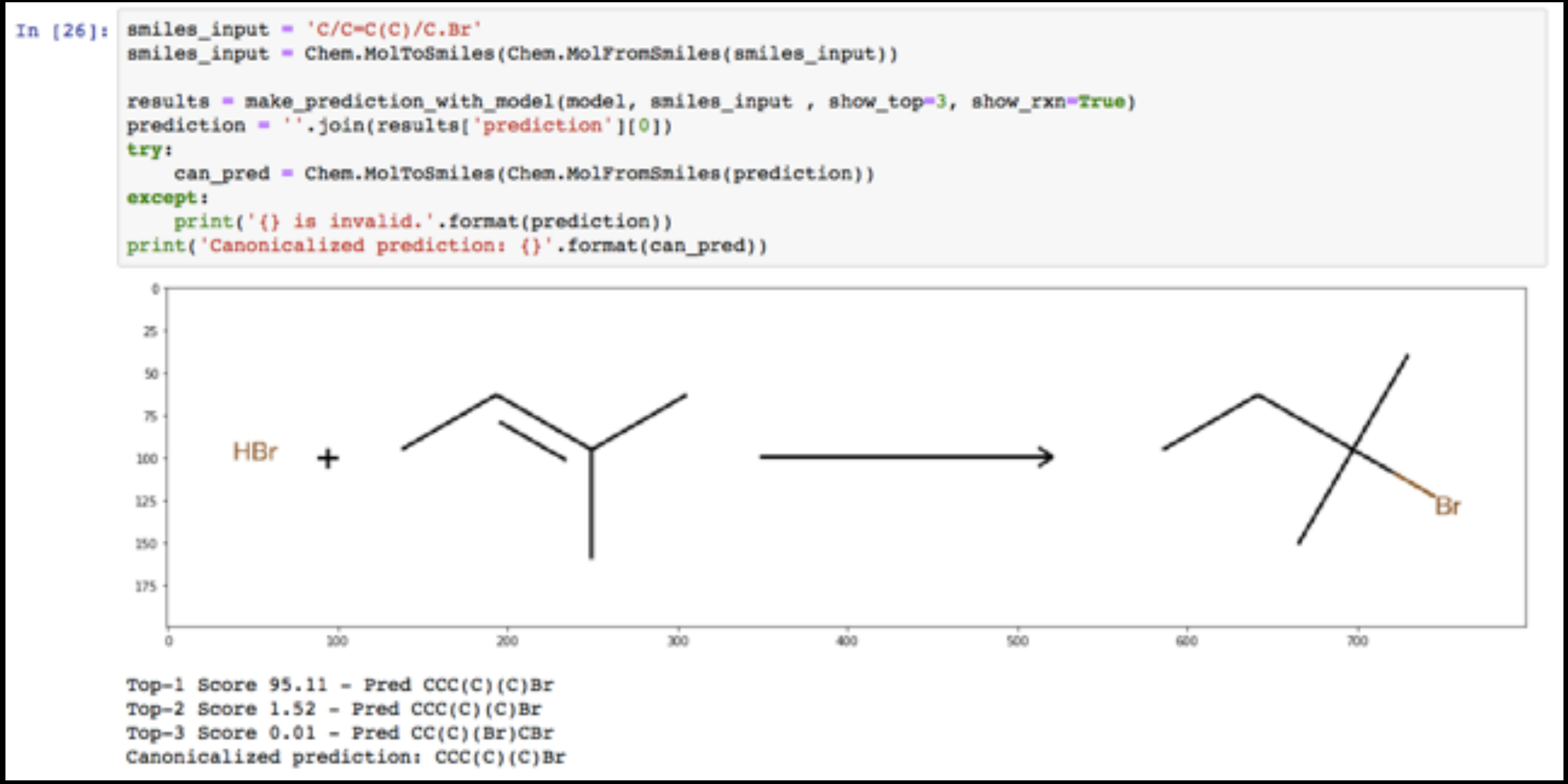


Our model



The model predicts subtle chemical selectivities

Regioselectivity: selecting amongst multiple reactive positions in a molecule



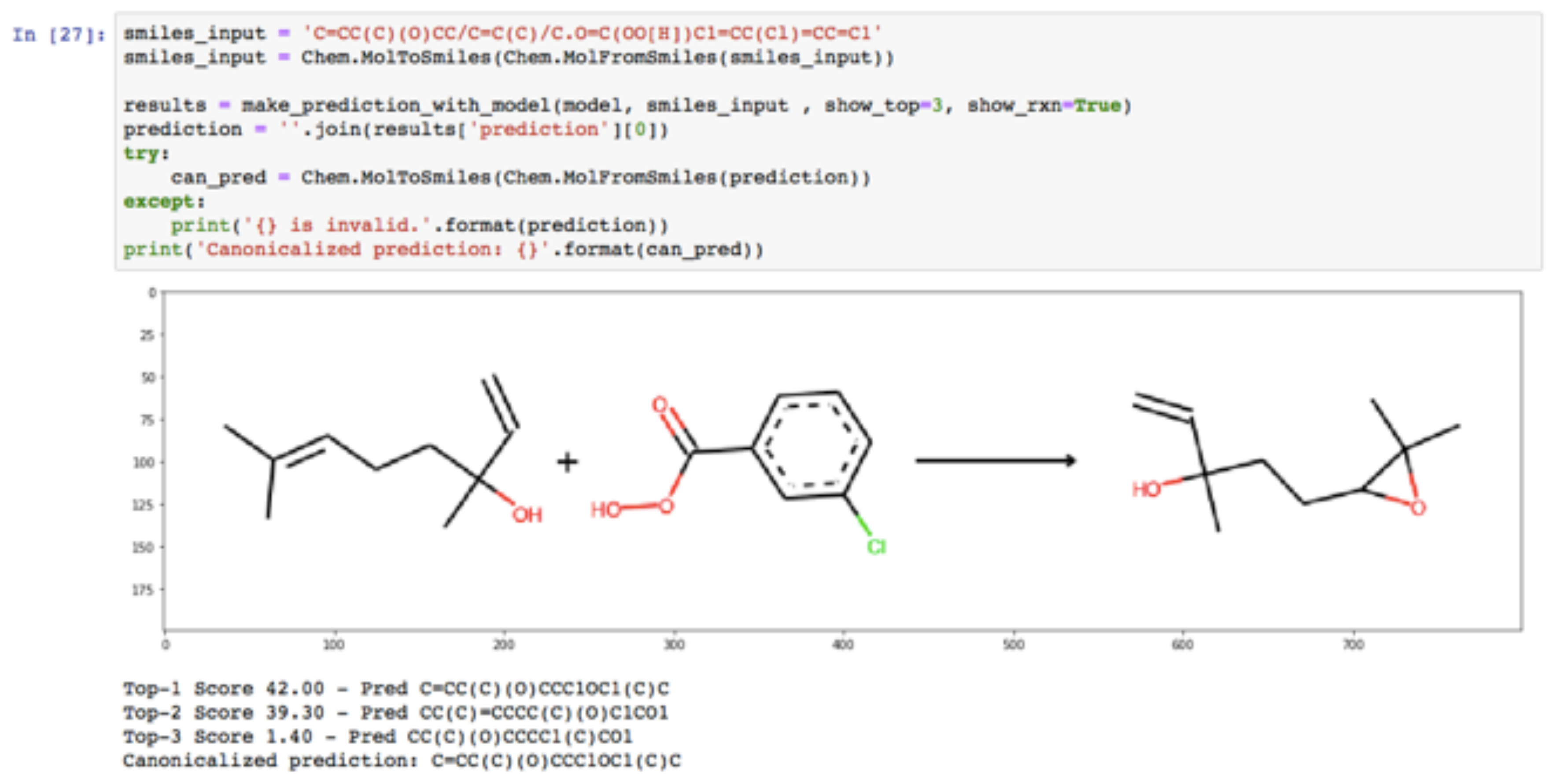
```
In [26]: smiles_input = 'C/C=C(C)/C.Br'
         smiles_input = Chem.MolToSmiles(Chem.MolFromSmiles(smiles_input))

         results = make_prediction_with_model(model, smiles_input , show_top=3, show_rxn=True)
         prediction = ''.join(results['prediction'][0])
         try:
             can_pred = Chem.MolToSmiles(Chem.MolFromSmiles(prediction))
         except:
             print('{} is invalid.'.format(prediction))
         print('Canonicalized prediction: {}'.format(can_pred))
```

```
Top-1 Score 95.11 - Pred CCC(C)(C)Br
Top-2 Score 1.52 - Pred CCC(C)(C)Br
Top-3 Score 0.01 - Pred CC(C)(Br)CBr
Canonicalized prediction: CCC(C)(C)Br
```

The Markovnikov rule is inferred from data without us coding the rule into the algorithm

The model predicts subtle chemical selectivities

Regioselectivity: selecting amongst multiple reactive positions in a molecule



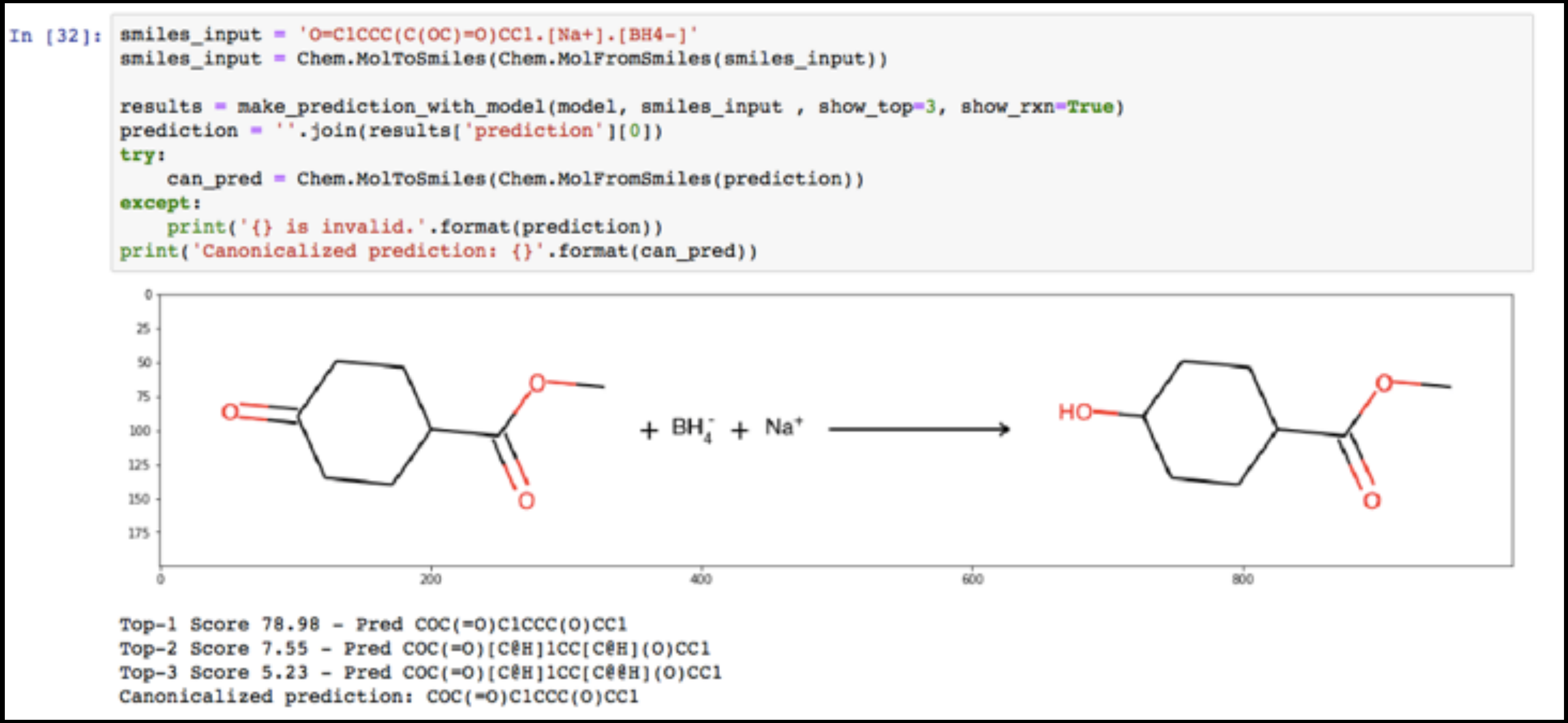
```
In [27]: smiles_input = 'C=CC(C)(O)CC/C=C(C)/C.O=C(OO[H])C1=CC(Cl)=CC=C1'
         smiles_input = Chem.MolToSmiles(Chem.MolFromSmiles(smiles_input))

         results = make_prediction_with_model(model, smiles_input , show_top=3, show_rxn=True)
         prediction = ''.join(results['prediction'][0])
         try:
             can_pred = Chem.MolToSmiles(Chem.MolFromSmiles(prediction))
         except:
             print('{} is invalid.'.format(prediction))
         print('Canonicalized prediction: {}'.format(can_pred))
```

```
Top-1 Score 42.00 - Pred C=CC(C)(O)CCC1OC1(C)C
Top-2 Score 39.30 - Pred CC(C)=CCCC(C)(O)C1CO1
Top-3 Score 1.40 - Pred CC(C)(O)CCCC1(C)CO1
Canonicalized prediction: C=CC(C)(O)CCC1OC1(C)C
```

Epoxidation of the more electron-rich alkene; reaction taken from J. Org. Chem., 57, 1198 (1992).

The model predicts subtle chemical selectivities

Chemoselectivity: selecting amongst multiple functional groups in a molecule



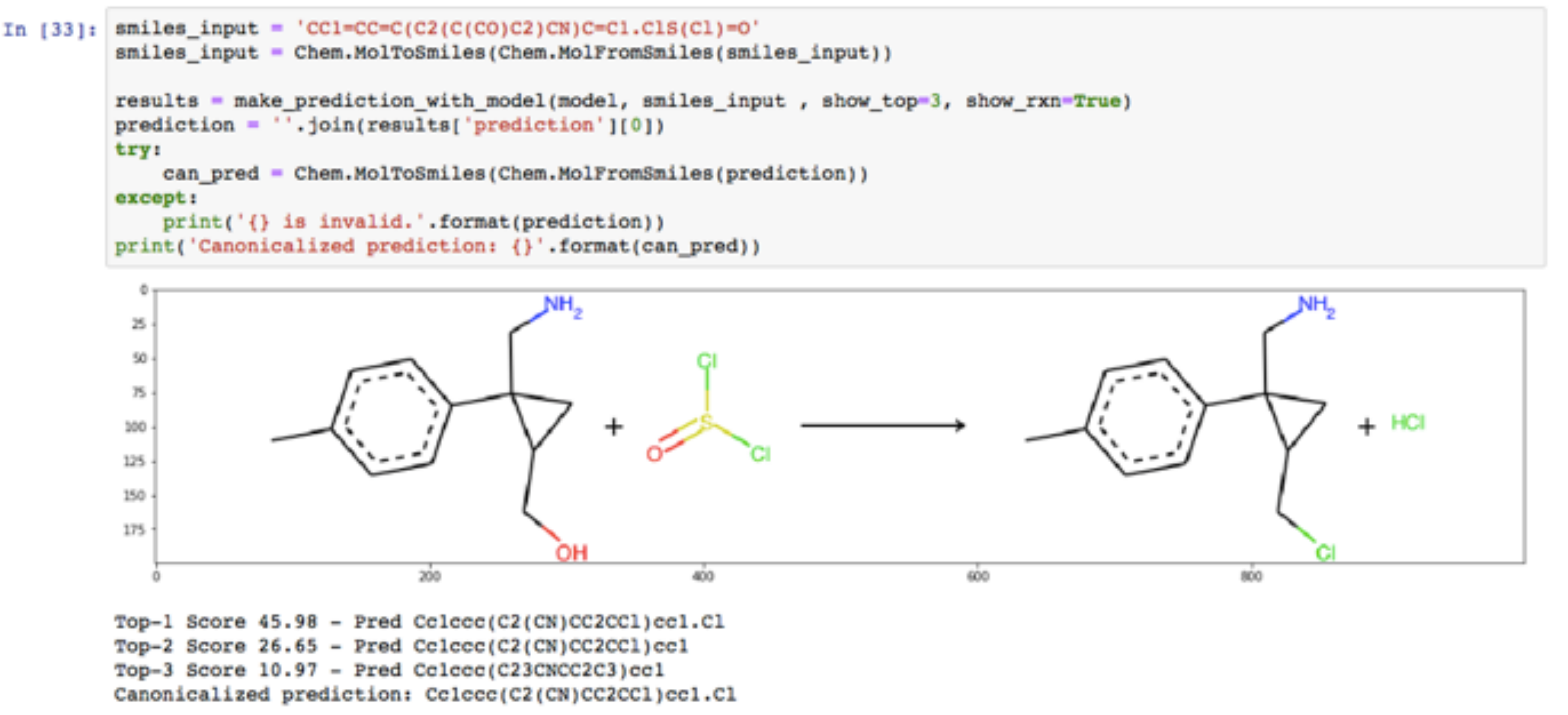
```
In [32]: smiles_input = 'O=C1CCC(C(OC)=O)CC1.[Na+].[BH4-]'
         smiles_input = Chem.MolToSmiles(Chem.MolFromSmiles(smiles_input))

         results = make_prediction_with_model(model, smiles_input , show_top=3, show_rxn=True)
         prediction = ''.join(results['prediction'][0])
         try:
             can_pred = Chem.MolToSmiles(Chem.MolFromSmiles(prediction))
         except:
             print('{} is invalid.'.format(prediction))
         print('Canonicalized prediction: {}'.format(can_pred))
```

```
Top-1 Score 78.98 - Pred COC(=O)C1CCC(O)CC1
Top-2 Score 7.55 - Pred COC(=O)[C@H]1CC[C@H](O)CC1
Top-3 Score 5.23 - Pred COC(=O)[C@H]1CC[C@@H](O)CC1
Canonicalized prediction: COC(=O)C1CCC(O)CC1
```

Chemoselective reduction of ketones in the presence of esters

The model predicts subtle chemical selectivities

Chemoselectivity: selecting amongst multiple functional groups in a molecule

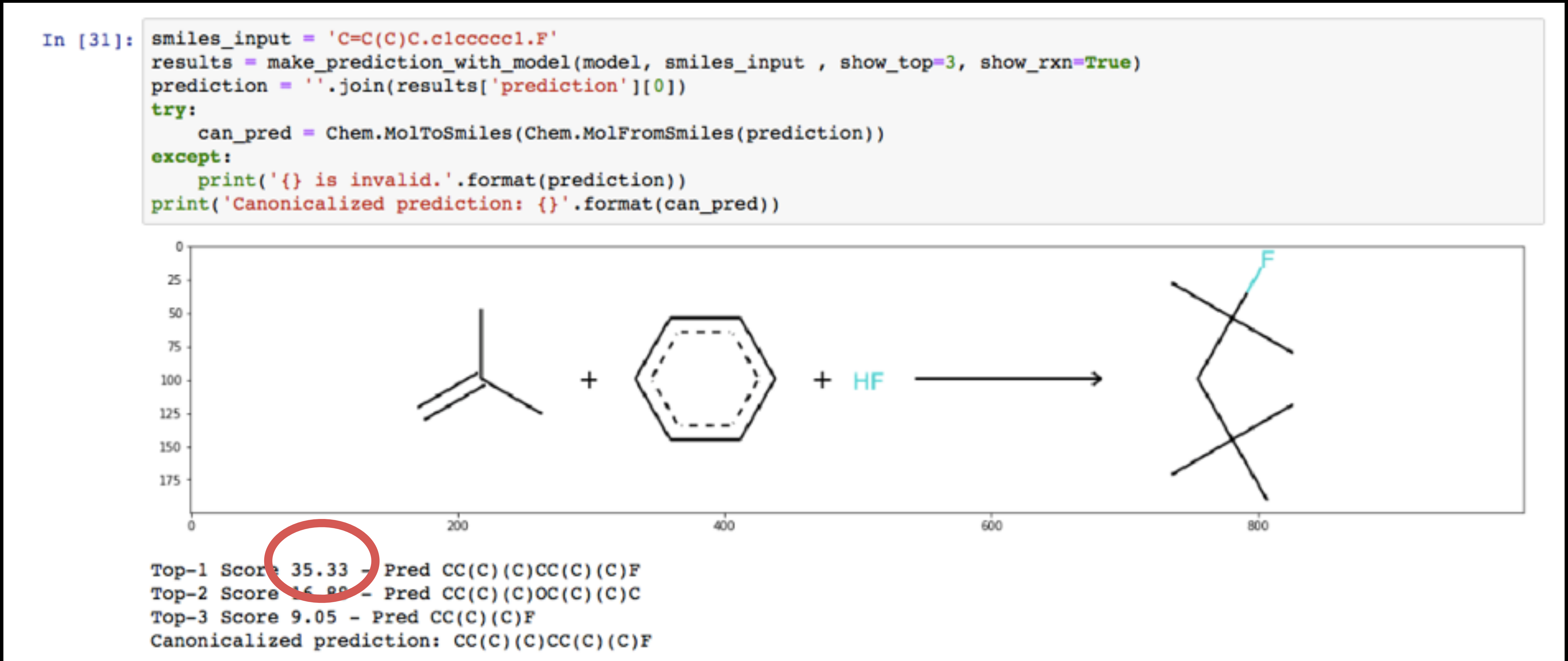


```
In [33]: smiles_input = 'CC1=CC=C(C2(C(CO)C2)CN)C=C1.ClS(Cl)=O'
         smiles_input = Chem.MolToSmiles(Chem.MolFromSmiles(smiles_input))

         results = make_prediction_with_model(model, smiles_input , show_top=3, show_rxn=True)
         prediction = ''.join(results['prediction'][0])
         try:
             can_pred = Chem.MolToSmiles(Chem.MolFromSmiles(prediction))
         except:
             print('{} is invalid.'.format(prediction))
         print('Canonicalized prediction: {}'.format(can_pred))
```

```
Top-1 Score 45.98 - Pred Cc1ccc(C2(CN)CC2CCl)cc1.Cl
Top-2 Score 26.65 - Pred Cc1ccc(C2(CN)CC2CCl)cc1
Top-3 Score 10.97 - Pred Cc1ccc(C23CNCC2C3)cc1
Canonicalized prediction: Cc1ccc(C2(CN)CC2CCl)cc1.Cl
```

The hydroxy group is chlorinated rather than the amine group; reaction taken from Angew. Chemie, 49, 262 (2010)

Even the wrong predictions are “chemically plausible”

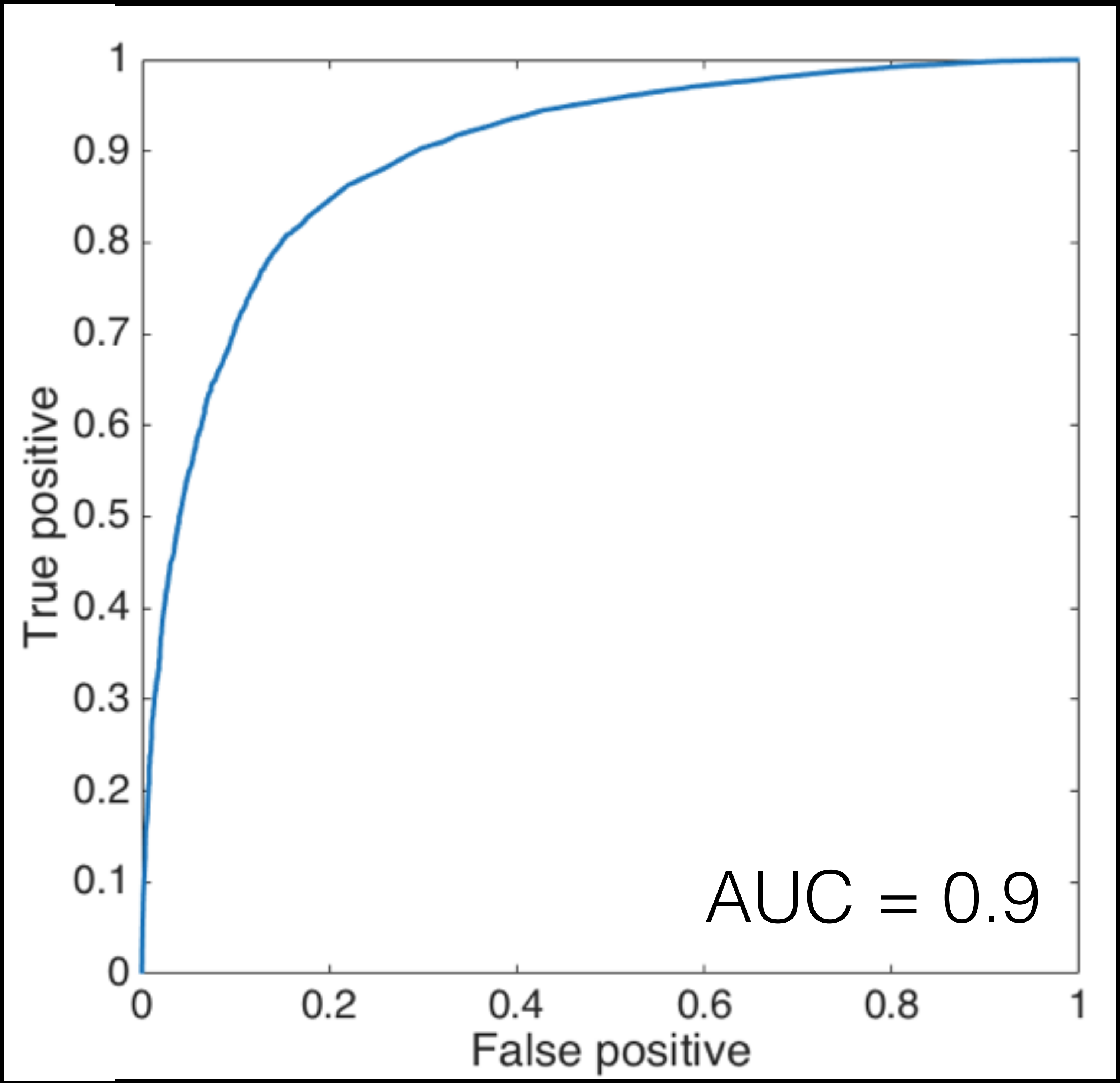


```
In [31]: smiles_input = 'C=C(C)C.c1ccccc1.F'
         results = make_prediction_with_model(model, smiles_input , show_top=3, show_rxn=True)
         prediction = ''.join(results['prediction'][0])
         try:
             can_pred = Chem.MolToSmiles(Chem.MolFromSmiles(prediction))
         except:
             print('{} is invalid.'.format(prediction))
         print('Canonicalized prediction: {}'.format(can_pred))
```

```
Top-1 Score 35.33 - Pred CC(C)(C)CC(C)(C)F
Top-2 Score 16.89 - Pred CC(C)(C)OC(C)(C)C
Top-3 Score 9.05 - Pred CC(C)(C)F
Canonicalized prediction: CC(C)(C)CC(C)(C)F
```

The model gives an estimate of likelihood

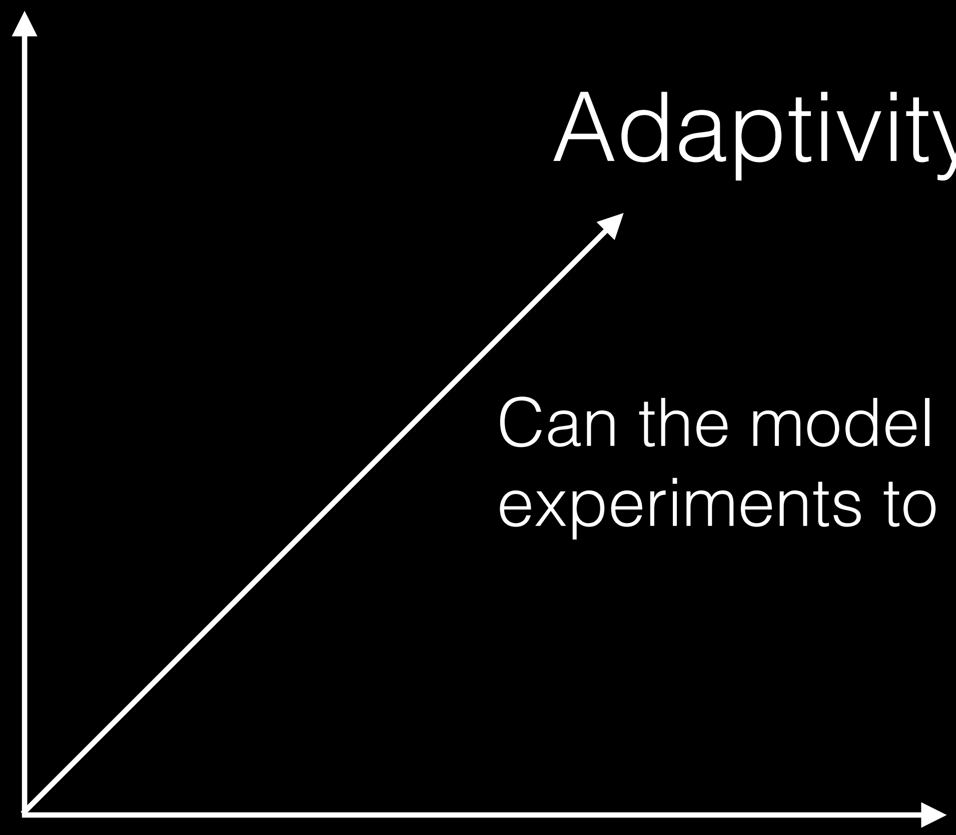
The model knows when it fails



Road map

- Reaction prediction and uncertainty calibration
- **Bayesian graph neural networks for uncertainty quantification**

Blackbox bioactivity prediction - what do we need?



Accuracy

RMSE, AUC etc.

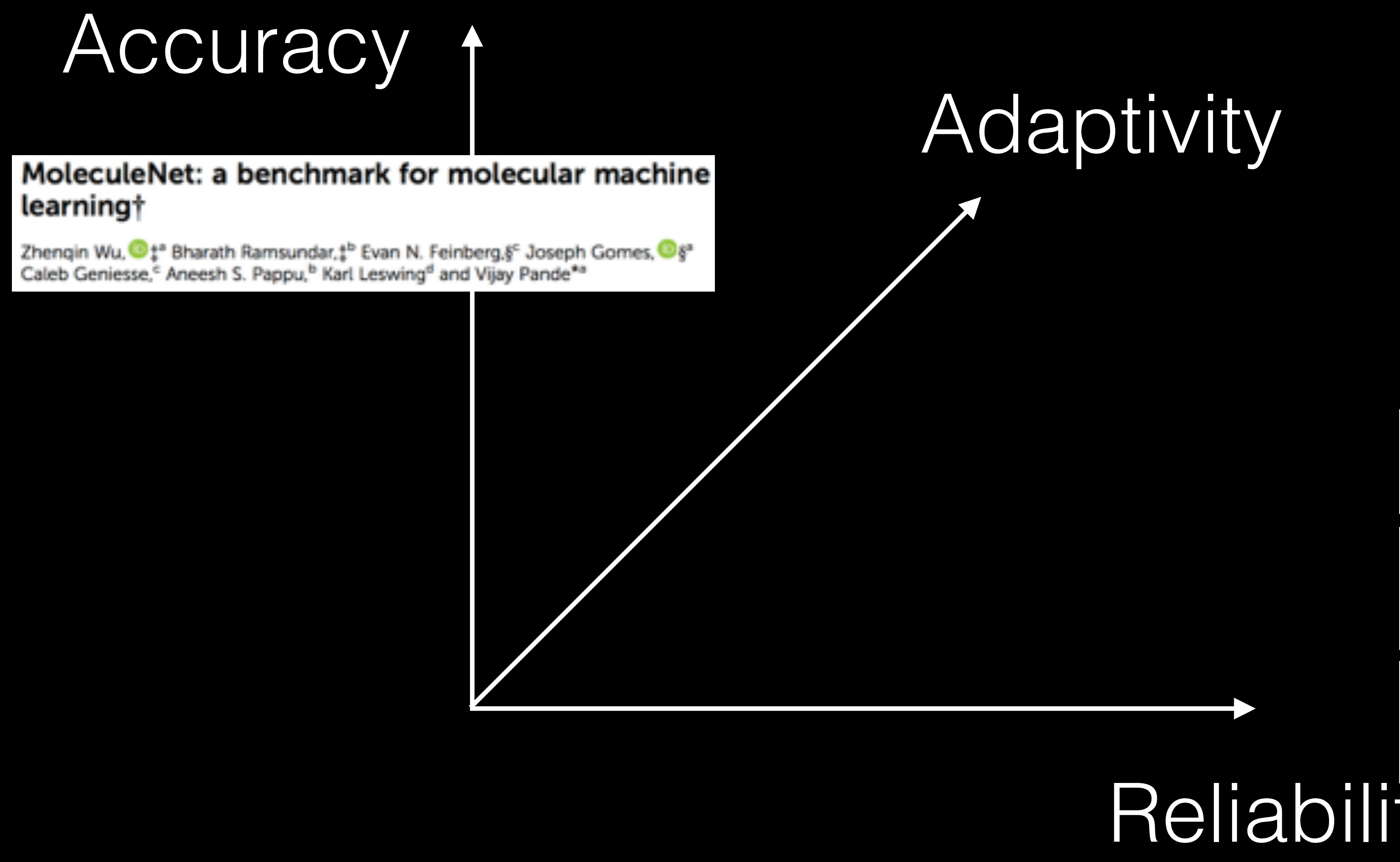
Adaptivity

Can the model design experiments to improve itself?

Reliability

Can the model predict when it will fail?

Blackbox bioactivity prediction - what do we need?



MoleculeNet: a benchmark for molecular machine learning†

Zhenqin Wu,[‡a] Bharath Ramsundar,[‡b] Evan N. Feinberg,[§c] Joseph Gomes,[§a] Caleb Geniesse,[c] Aneesh S. Pappu,[b] Karl Leswing[d] and Vijay Pande*[a]

ACS central science

Research Article

http://pubs.acs.org/journal/acscii

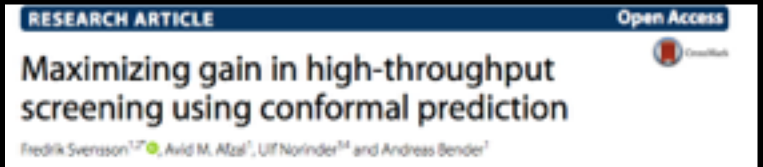
Low Data Drug Discovery with One-Shot Learning

Han Altae-Tran,[†,#] Bharath Ramsundar,[‡,#] Aneesh S. Pappu,[†] and Vijay Pande*[,§]

Active-learning strategies in computer-assisted drug discovery

Daniel Reker and Gisbert Schneider

Swiss Federal Institute of Technology (ETH), Department of Chemistry and Applied Biosciences, Vladimir-Prelog-Weg 4, 8093 Zürich, Switzerland



RESEARCH ARTICLE Open Access

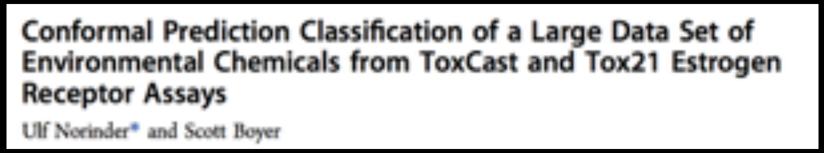
CrossMark

Maximizing gain in high-throughput screening using conformal prediction

Fredrik Svensson[1,2*], Avid M. Afzal[1], Ulf Norinder[3,4] and Andreas Bender[1]

Three Useful Dimensions for Domain Applicability in QSAR Models Using Random Forest

Robert P. Sheridan*



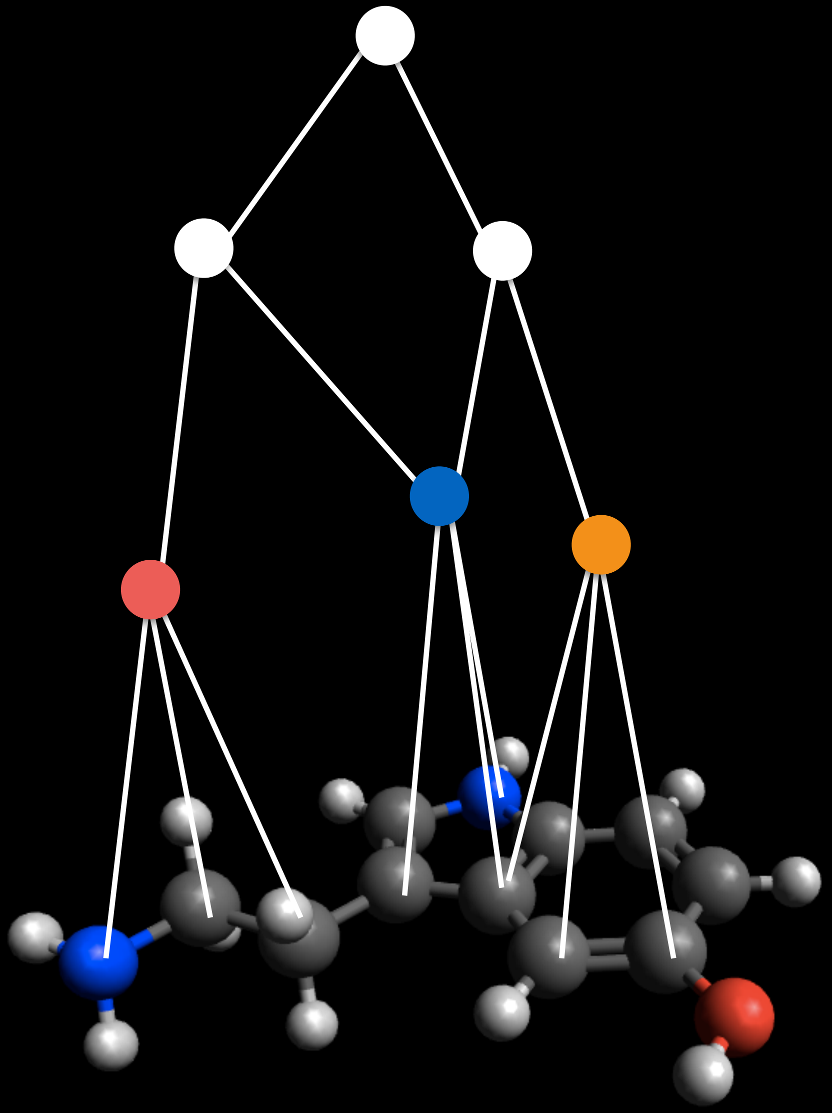
Conformal Prediction Classification of a Large Data Set of Environmental Chemicals from ToxCast and Tox21 Estrogen Receptor Assays

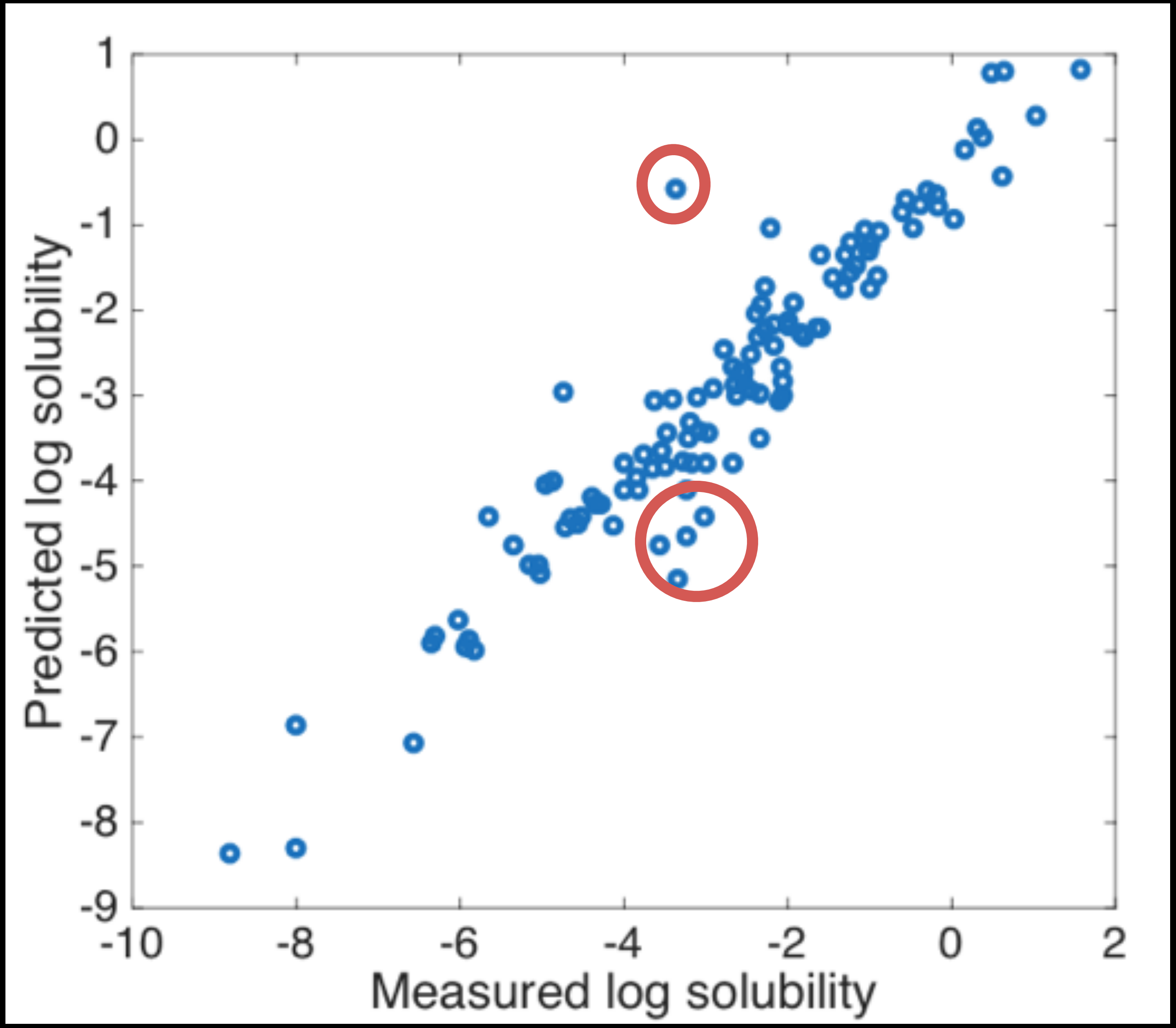
Ulf Norinder* and Scott Boyer

Conformal Regression for Quantitative Structure–Activity Relationship Modeling—Quantifying Prediction Uncertainty

Fredrik Svensson,[*,†,‡] Natalia Aniceto,[†] Ulf Norinder,[§,∥] Isidro Cortes-Ciriano,[†] Ola Spjuth,[⊥] Lars Carlsson,[#,∇] and Andreas Bender[†]

Uncertainty in graph neural networks





e.g. D. Duvenaud et al., NIPS 2015

The Bayesian idea

- Determine the distribution of parameters that conform to the data rather than best-fit parameters

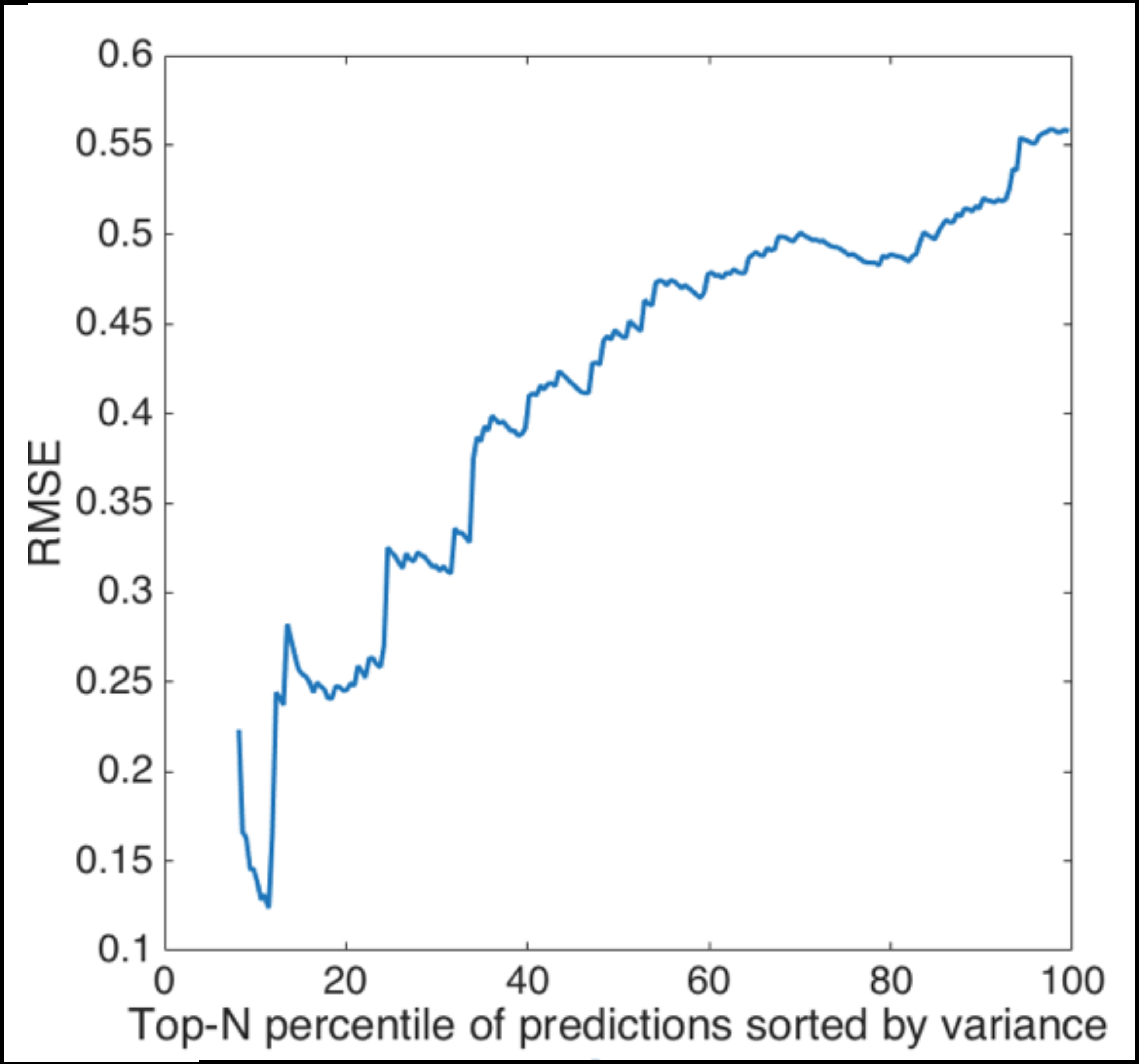
$$p(\mathbf{\Theta}|\mathrm{Data}) = \frac{1}{Z} p(\mathrm{Data}|\mathbf{\Theta})p(\mathbf{\Theta})$$

- The model prediction and error are simply the mean and variance with respect to the posterior

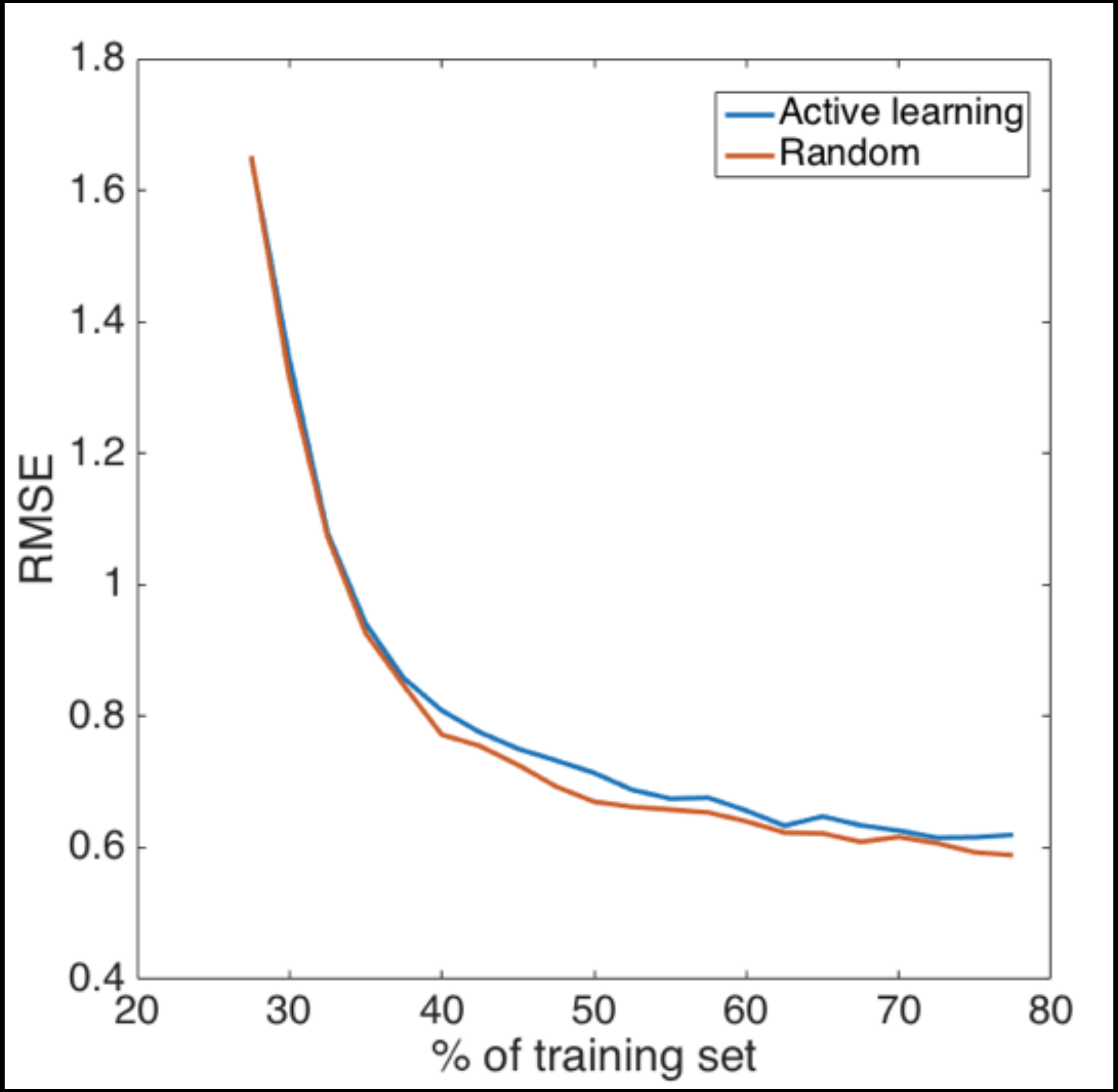
$$\langle y_{\mathrm{pred}} \rangle = \int f_{\Theta}(\mathbf{x}) p(\mathbf{\Theta}|\mathrm{Data}) \boxed{\mathrm{d}\mathbf{\Theta}}$$

$$\sigma_y^2 = \langle y_{\mathrm{pred}}^2 \rangle - \langle y_{\mathrm{pred}} \rangle^2$$

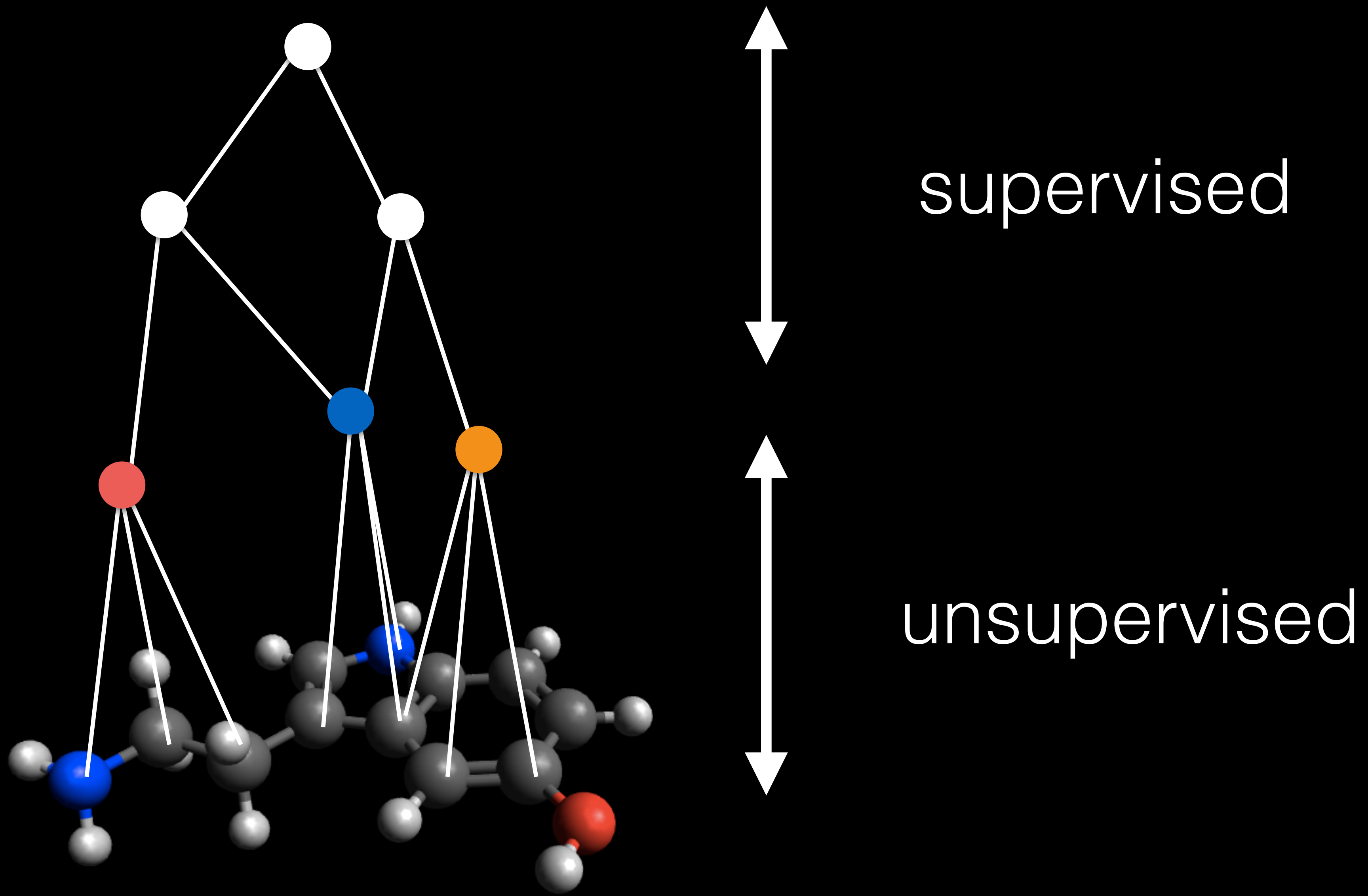
Simple Bayesian approach with Dropout



The good news becomes the bad news



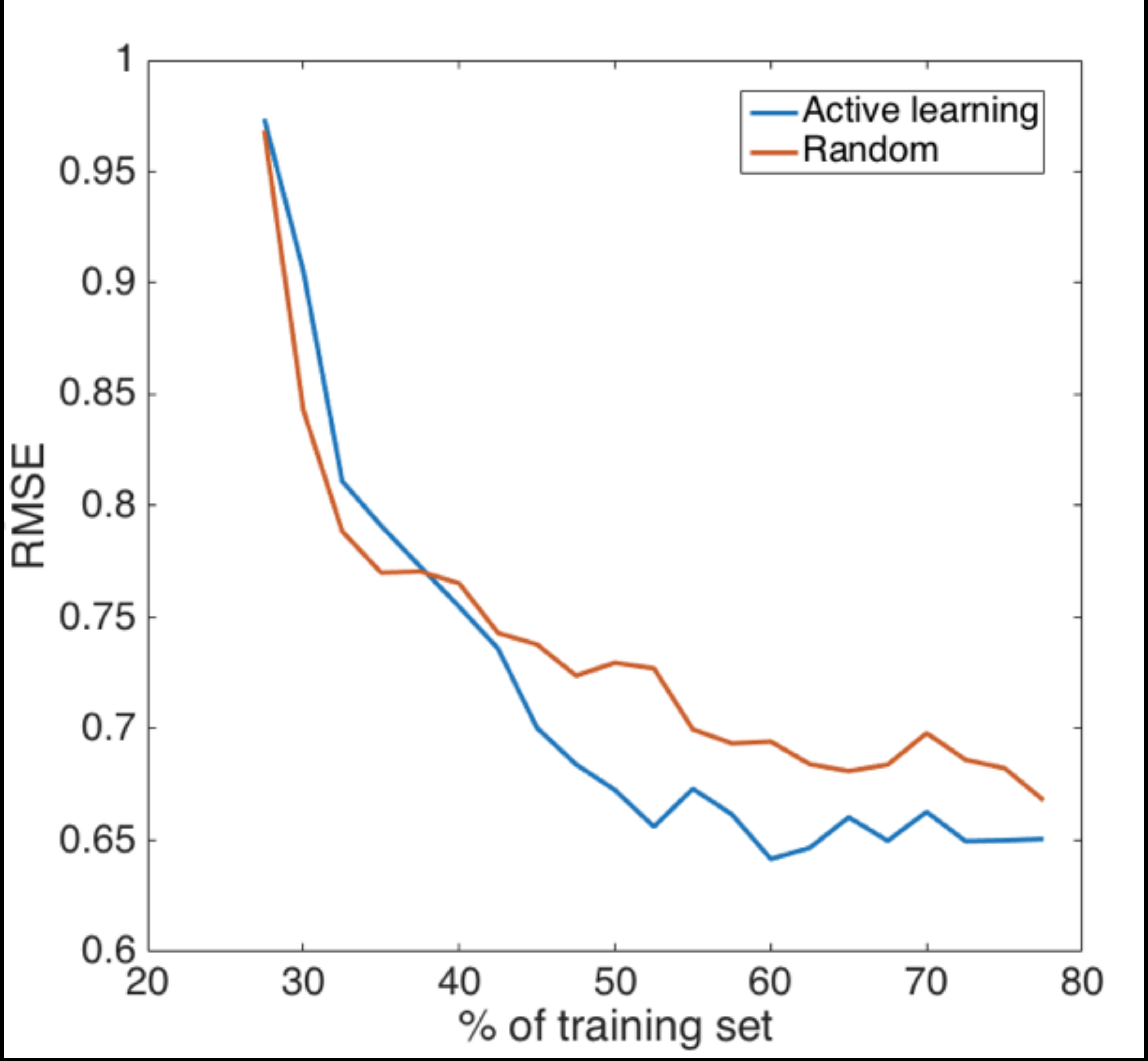
Finding robust representation in the low-data limit



supervised

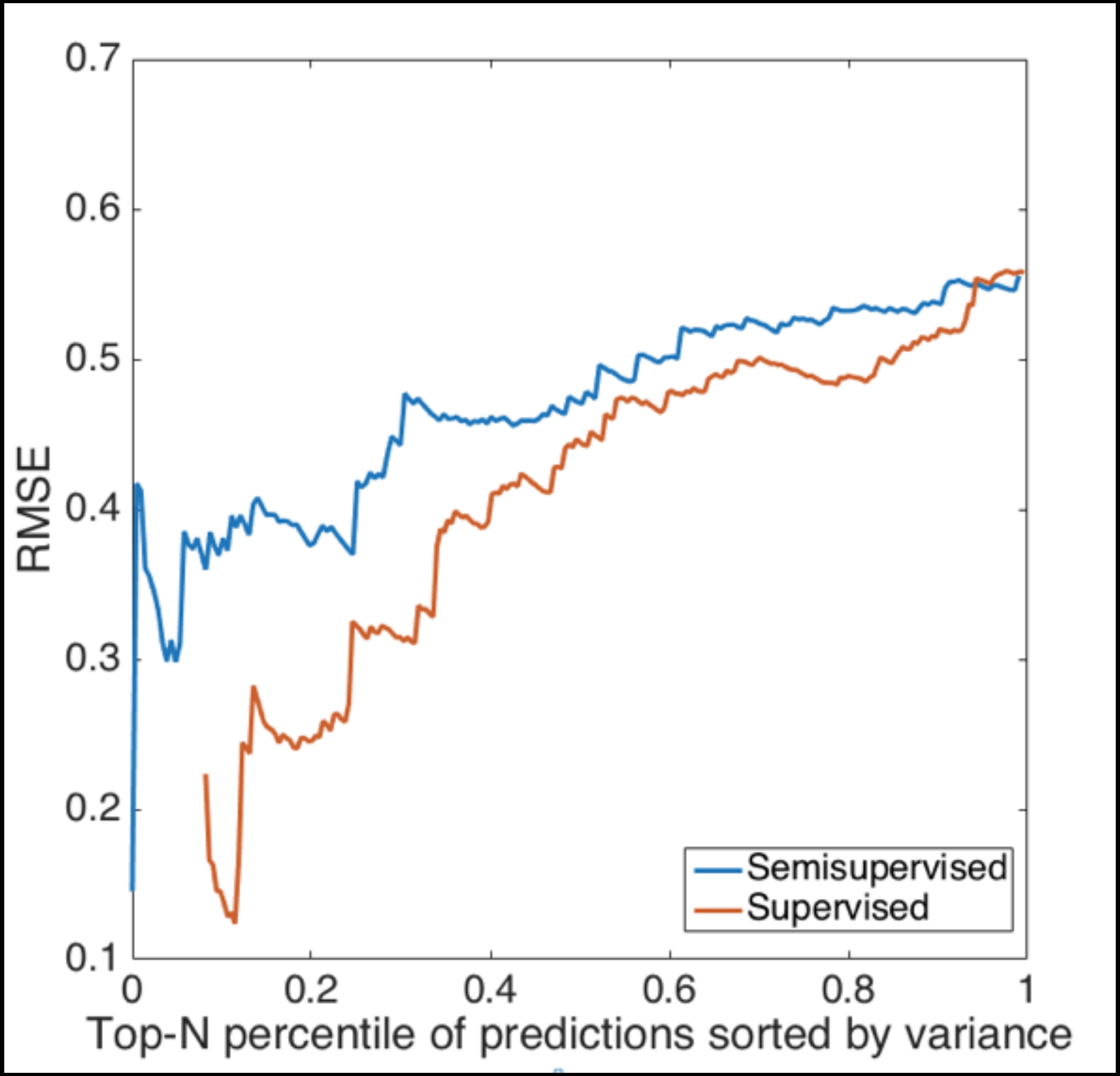
unsupervised

Active learning



~10%

Uncertainty estimation



Conclusions

- Spotting correlations in SMILES tokens between reactant, reagent and product is an accurate and reliable way to predict the outcome of organic reactions

- For molecular properties prediction, Bayesian semi-supervised deep learning appears to give a balanced performance in terms of accuracy, reliability, and adaptivity

Acknowledgements

- Philippe Schwaller
- Yao Zhang

THE WINTON PROGRAMME FOR THE
Physics of Sustainability

We are hiring!

contact me: aal44@cam.ac.uk