Controlling generalization errors for ML-informed molecular discovery

Jon Paul Janet [1] Heather Kulik [1]

[1]Department of Chemical Engineering, Massachusetts Institute of Technology

257th ACS National Meeting, Orlando

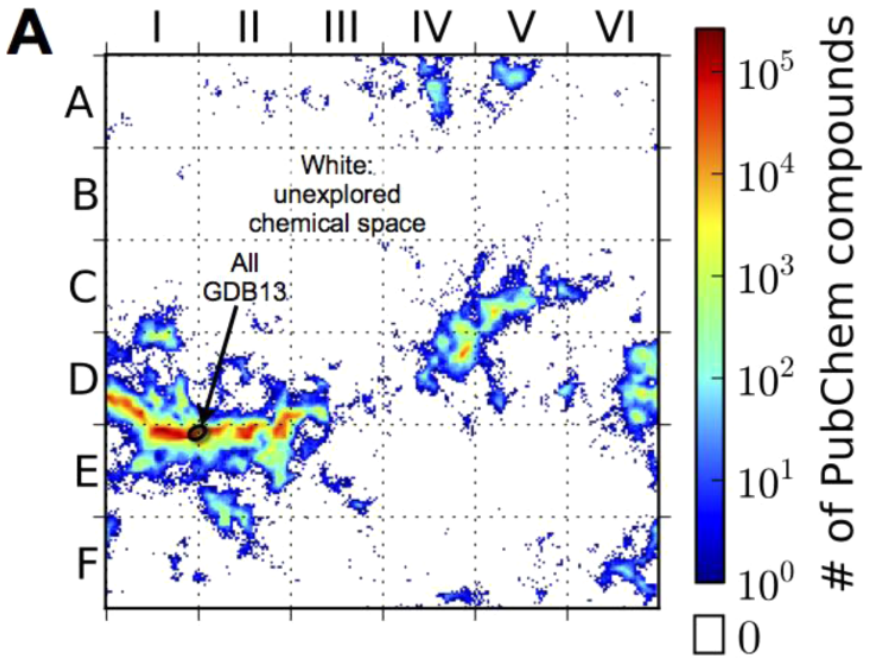
03.31.19

Motivation: chemical discovery

How can we design new materials using computers?

The space of possible chemistries is incredibly vast, with $\mathcal{O}(10^{50})$ small organic molecules.

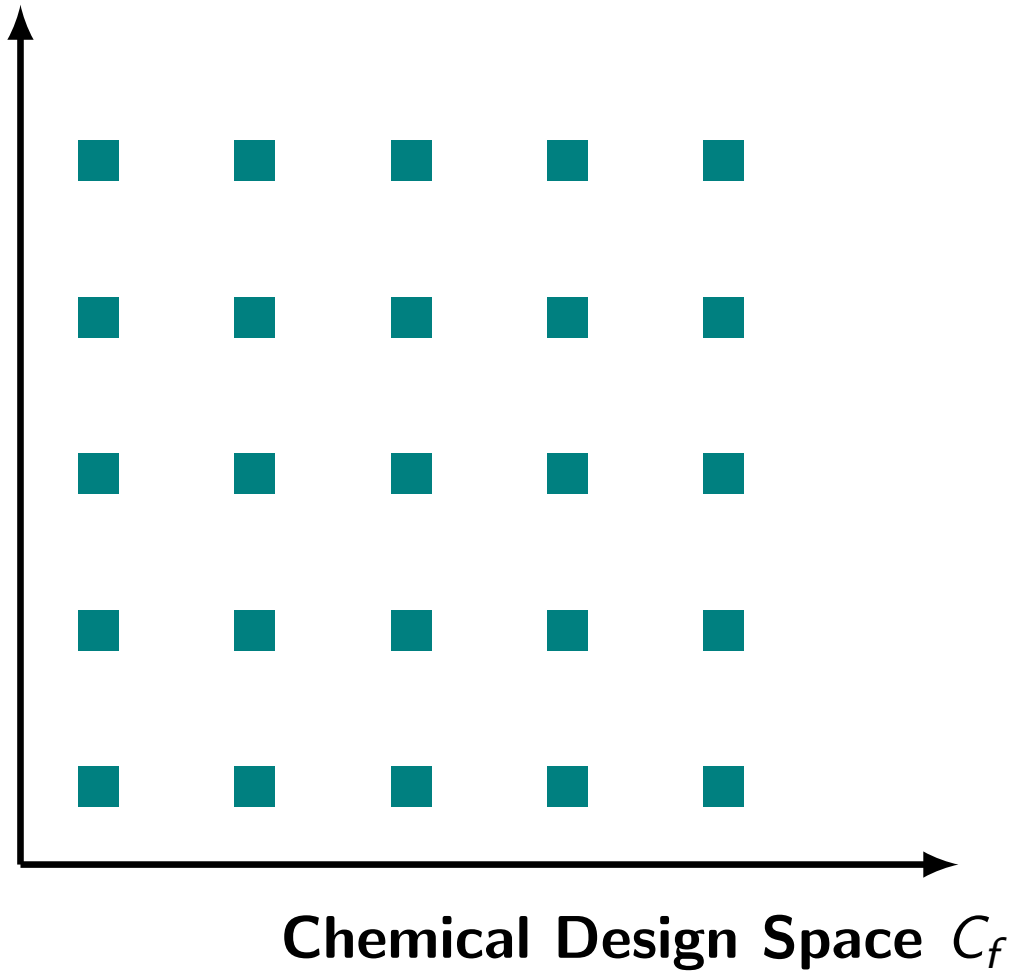
All potentially undiscovered medicines, catalysts and materials are somewhere, out in this huge space.



Virshup *et al.*, *J. Am. Chem. Soc.*, 135(19): 7296–7303, 2013.

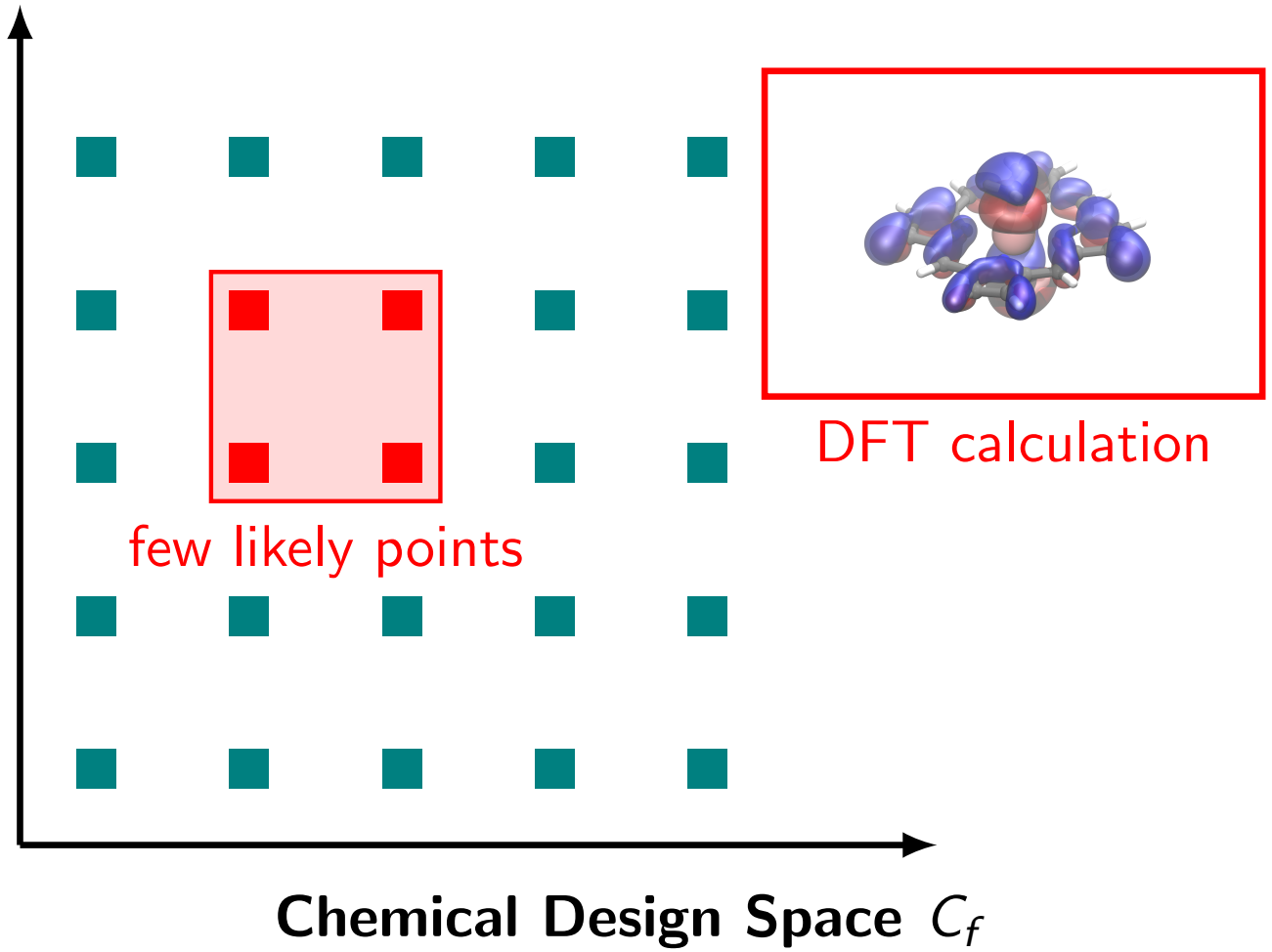
Motivation: chemical discovery

How can we design new materials using computers?



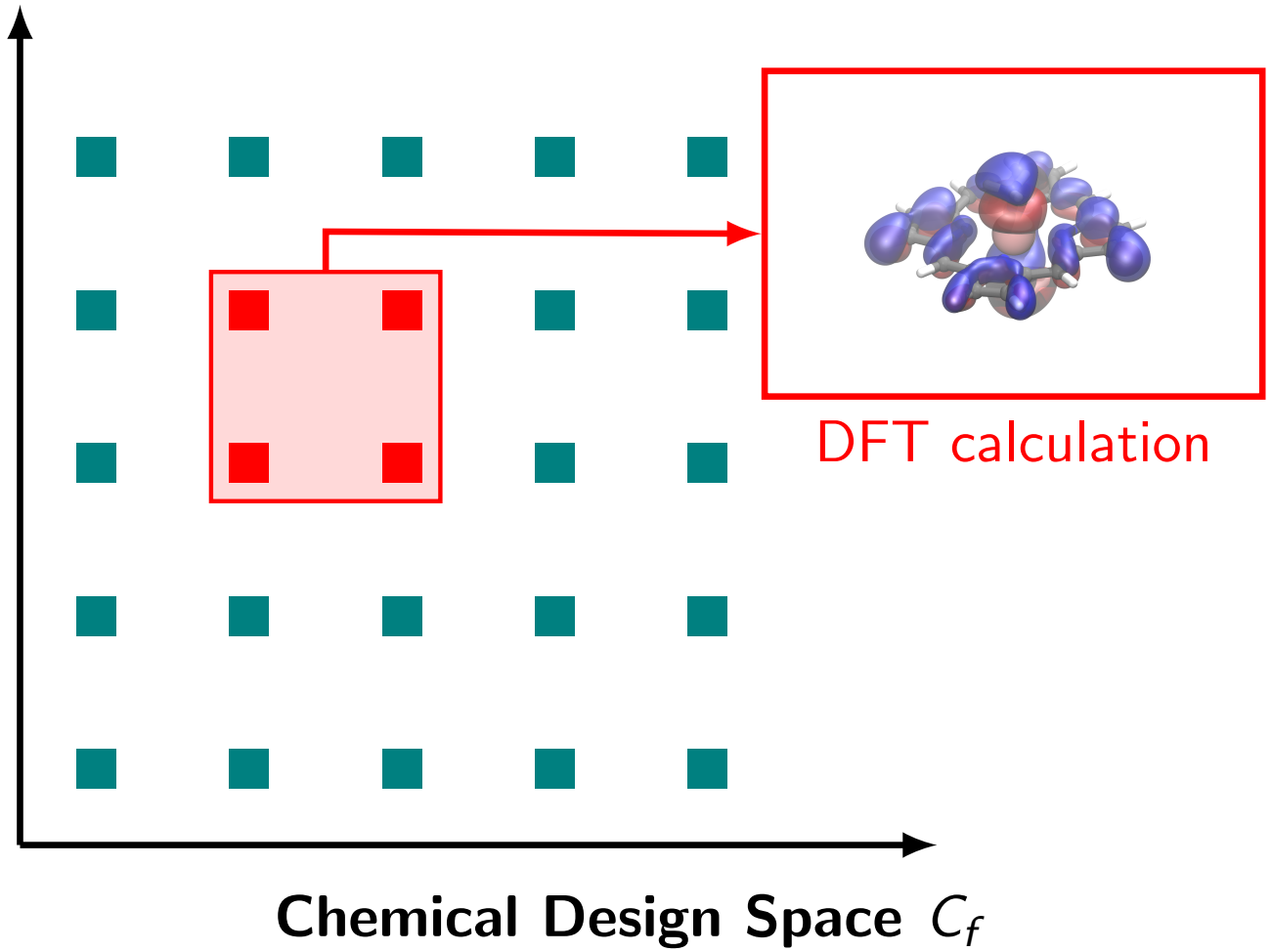
Motivation: chemical discovery

How can we design new materials using computers?



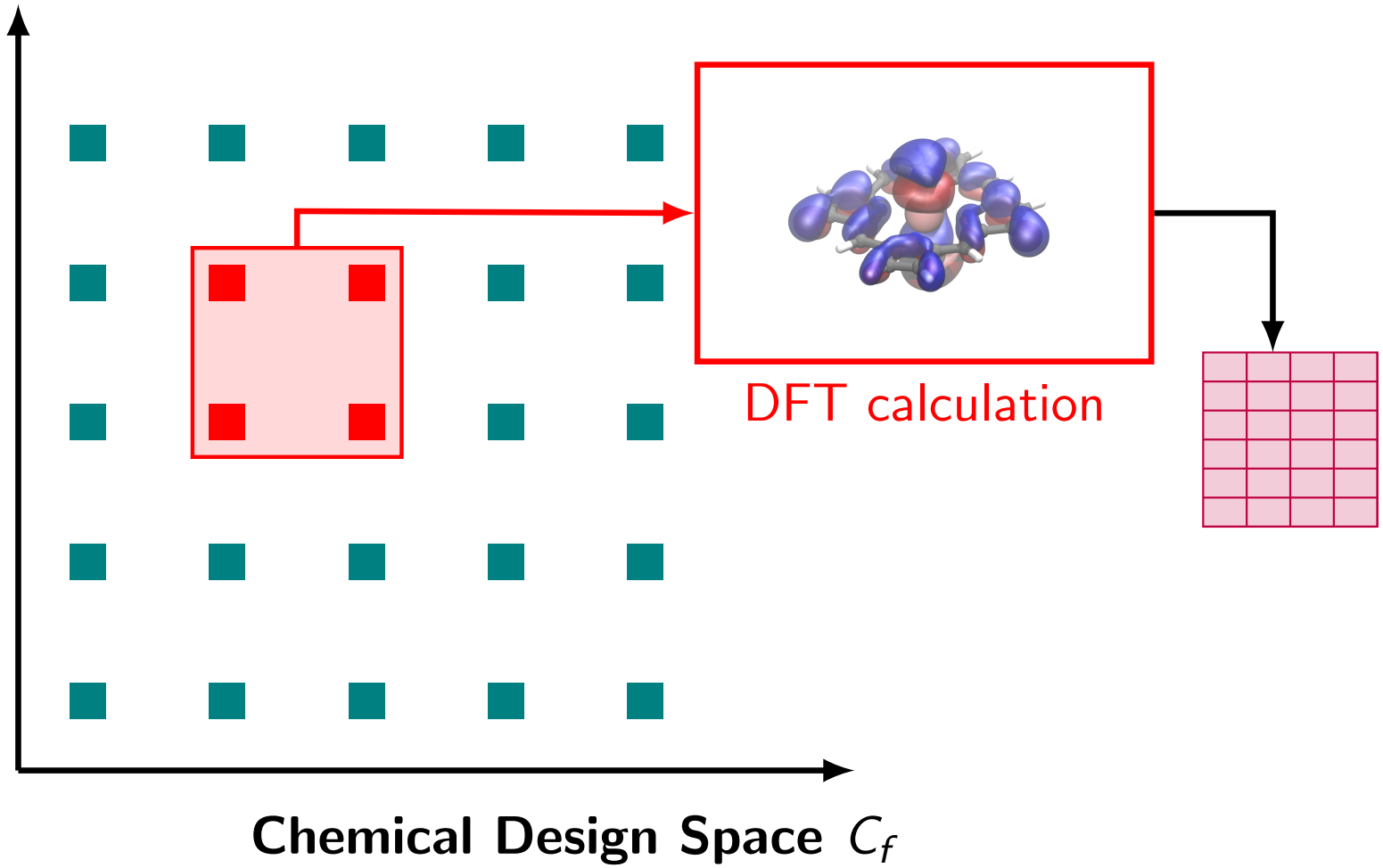
Motivation: chemical discovery

How can we design new materials using computers?

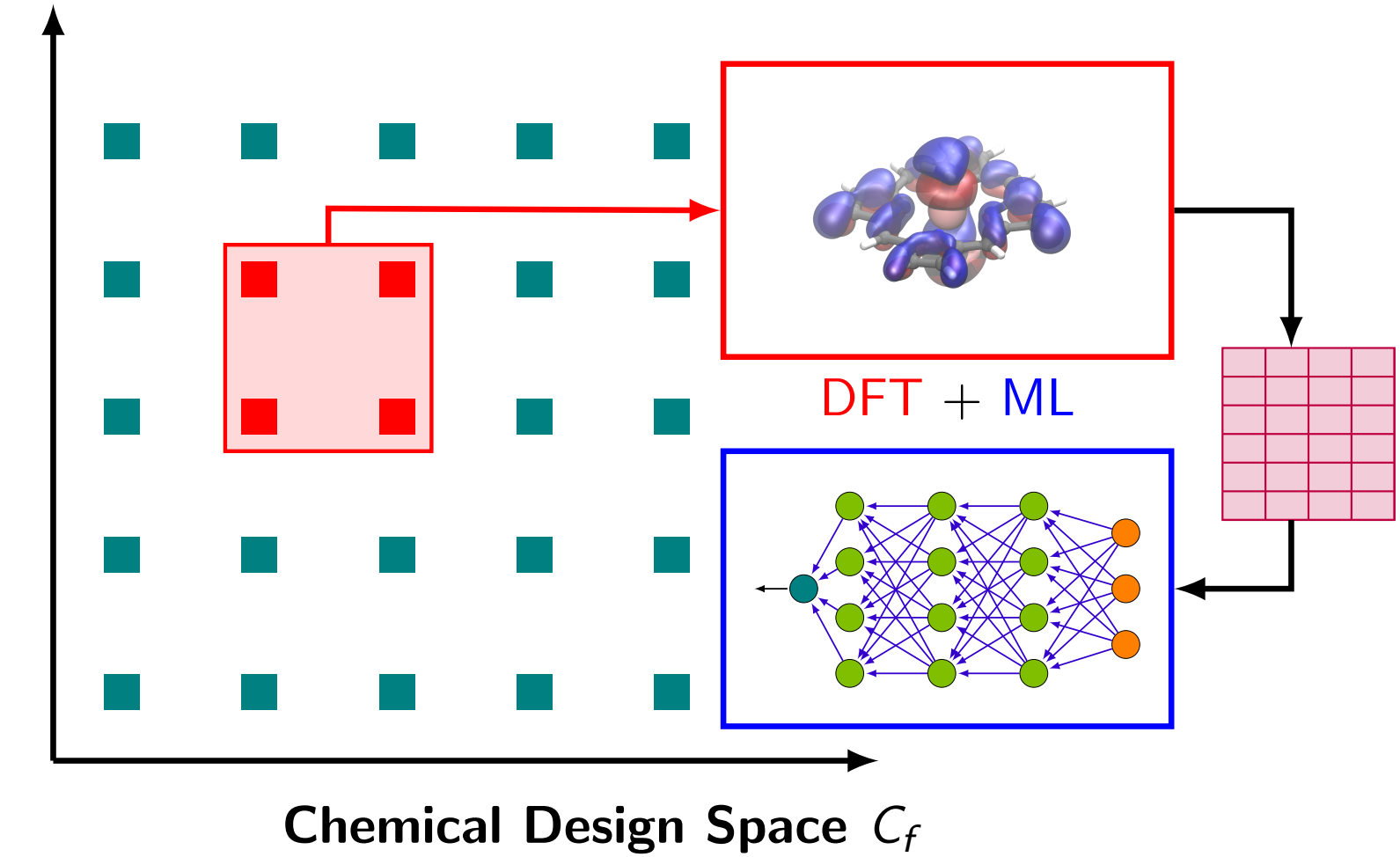


Motivation: chemical discovery

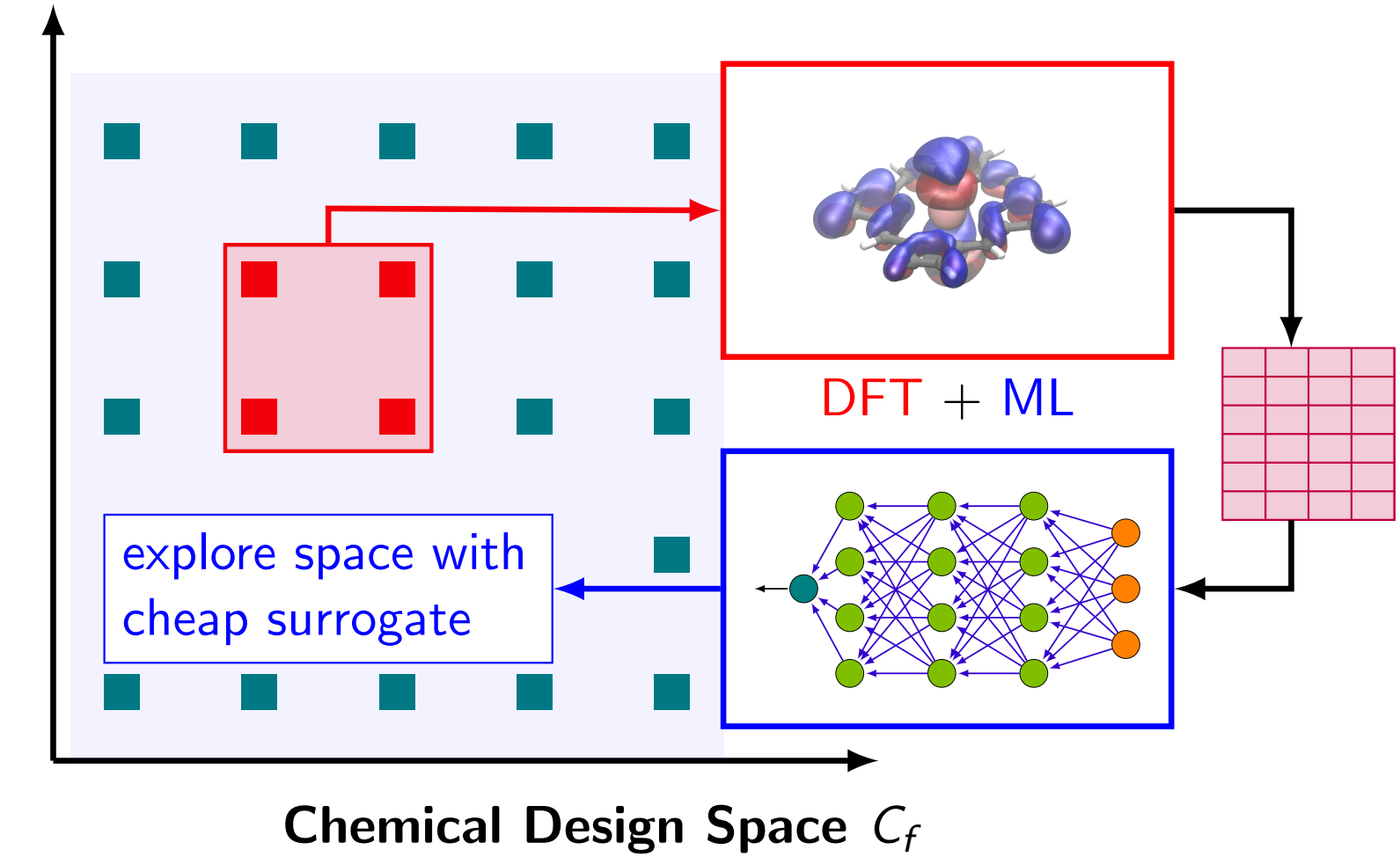
How can we design new materials using computers?



Motivation: chemical discovery

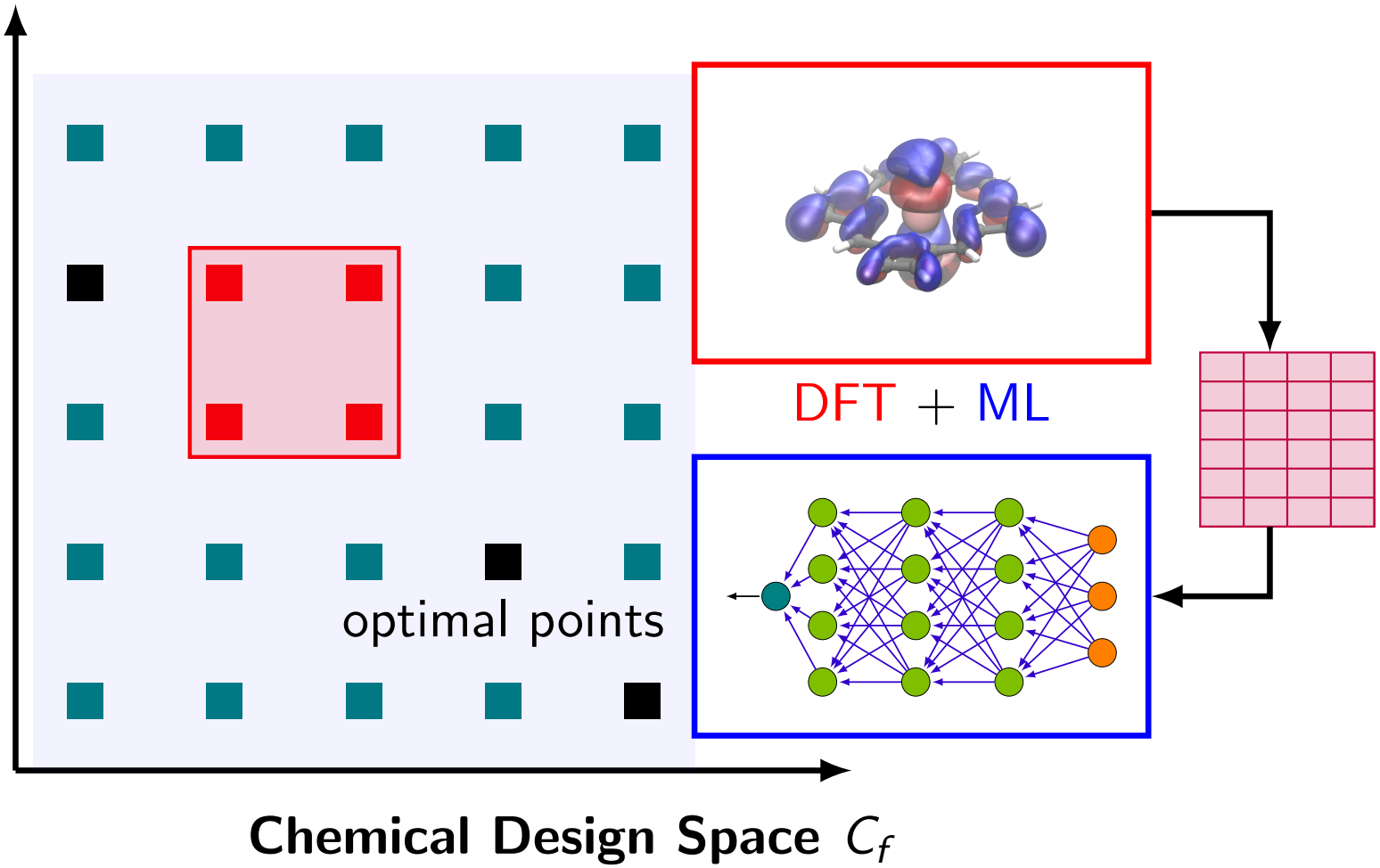


Motivation: chemical discovery



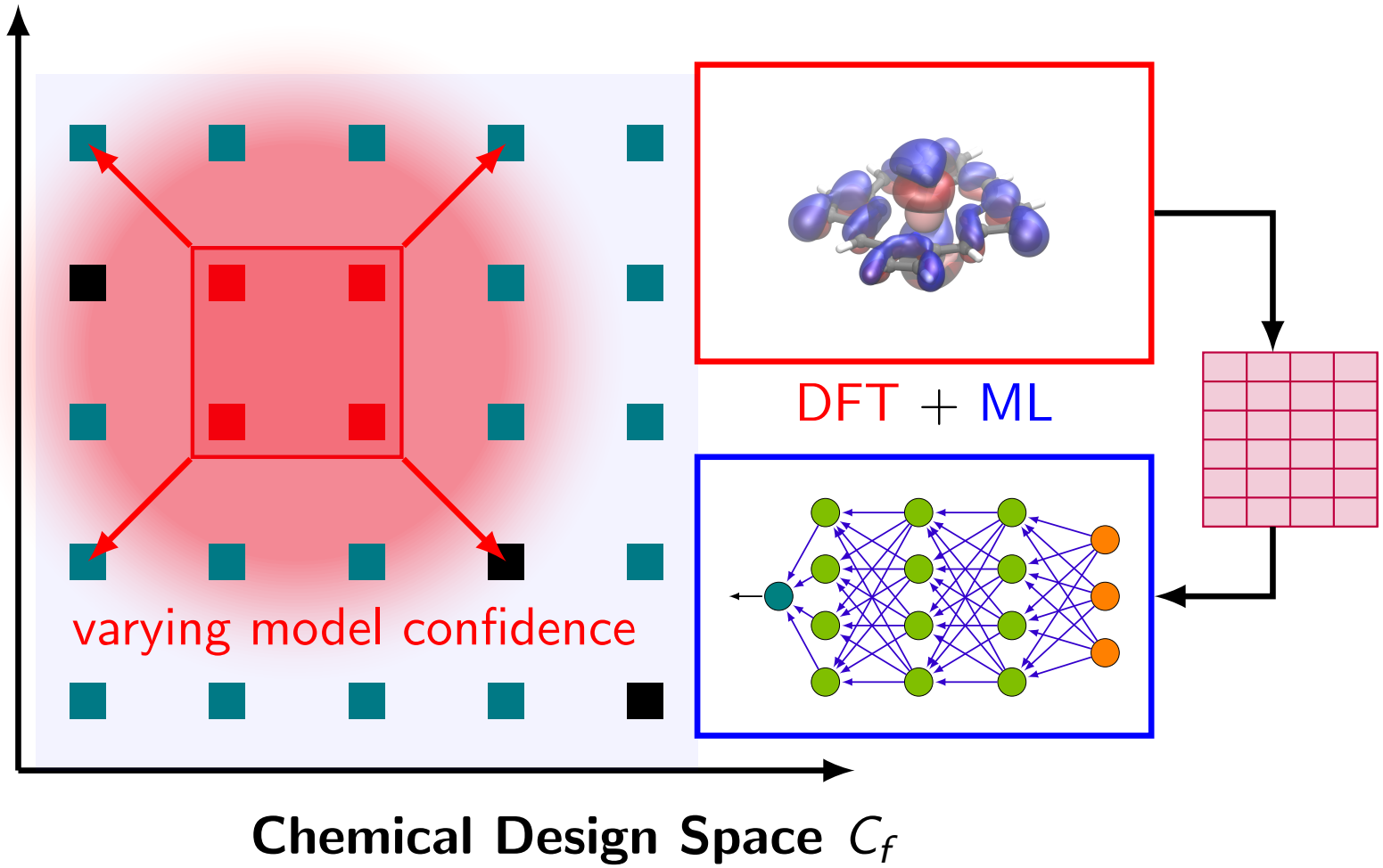
Motivation: chemical discovery

How can we design new materials using computers?



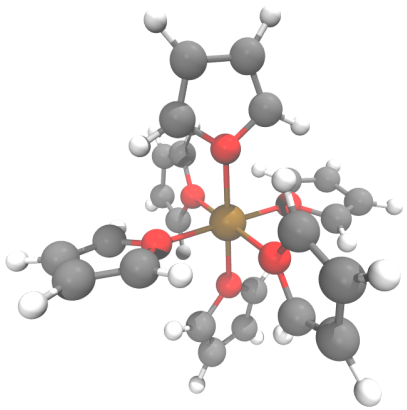
Motivation: chemical discovery

How can we design new materials using computers?

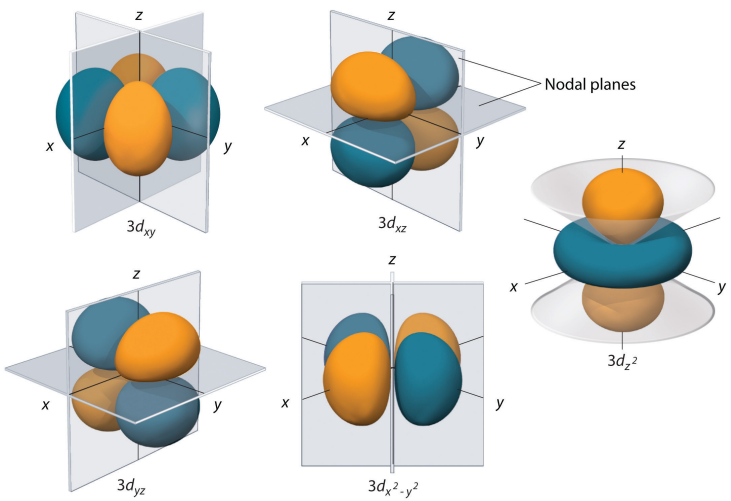


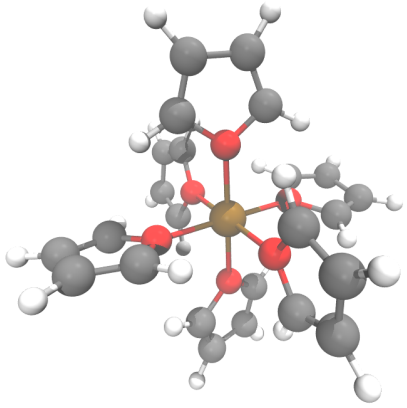
Transition metal complexes

Transition metal complexes



Transition metal complexes

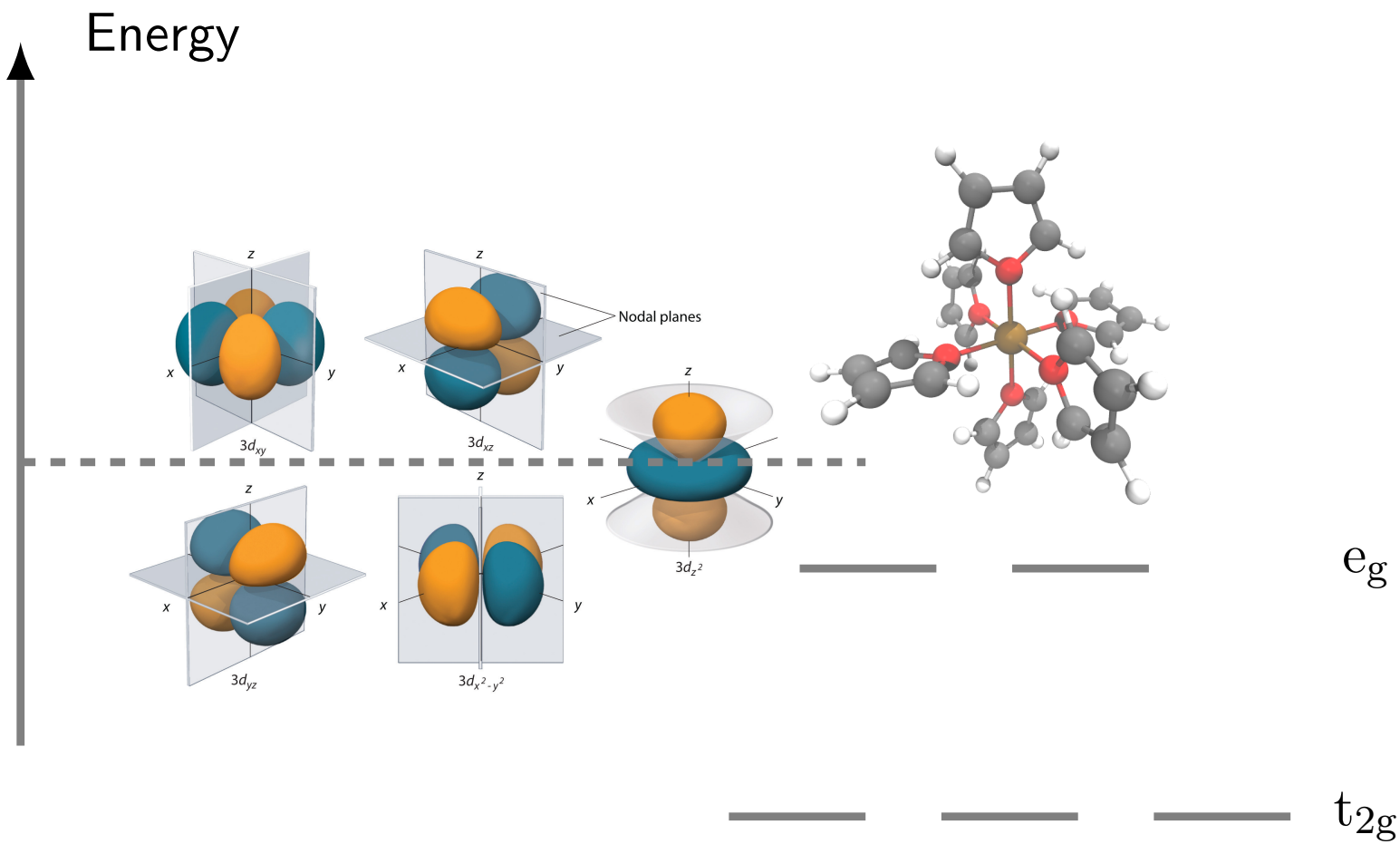




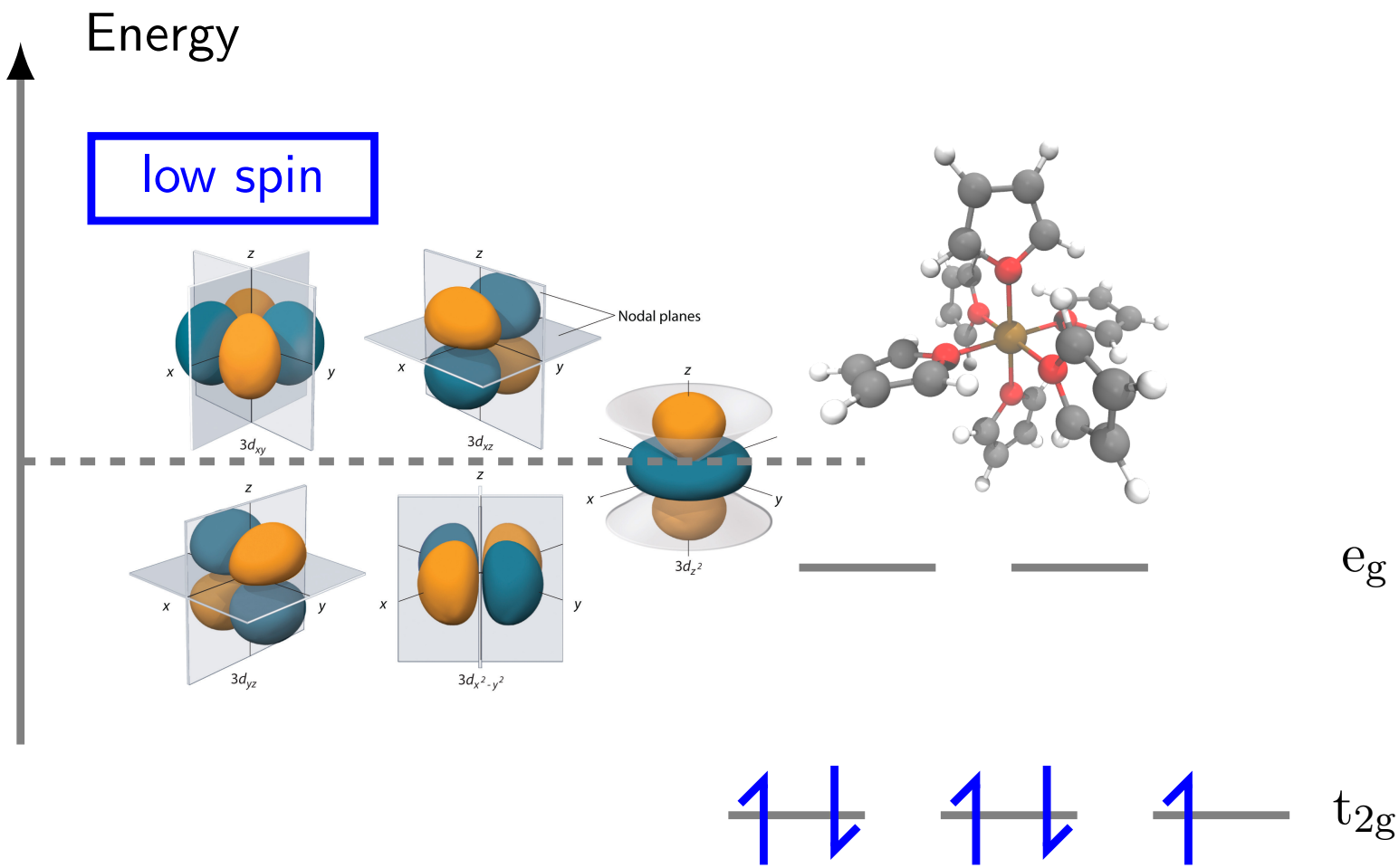
e_g

t_{2g}

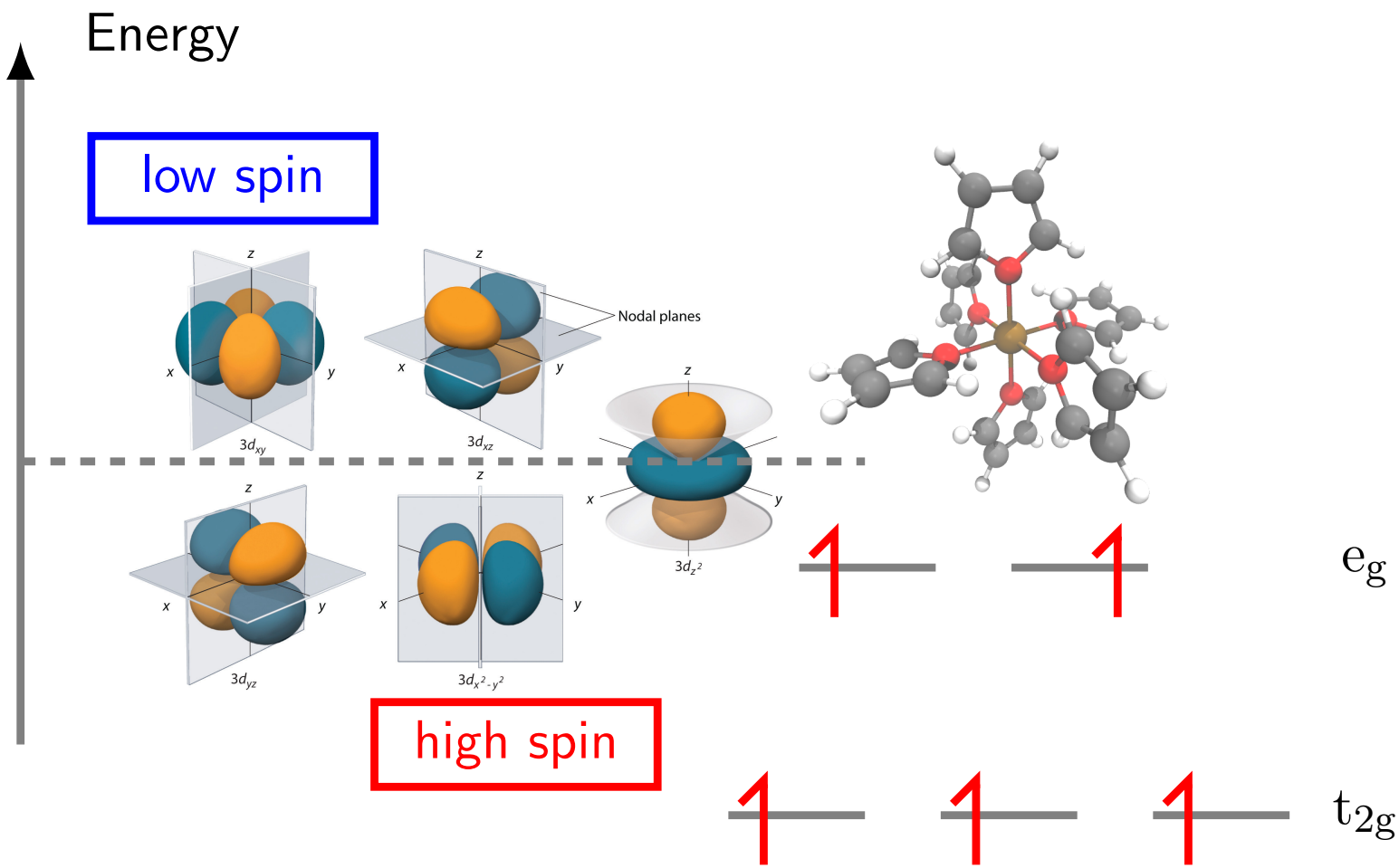
Transition metal complexes



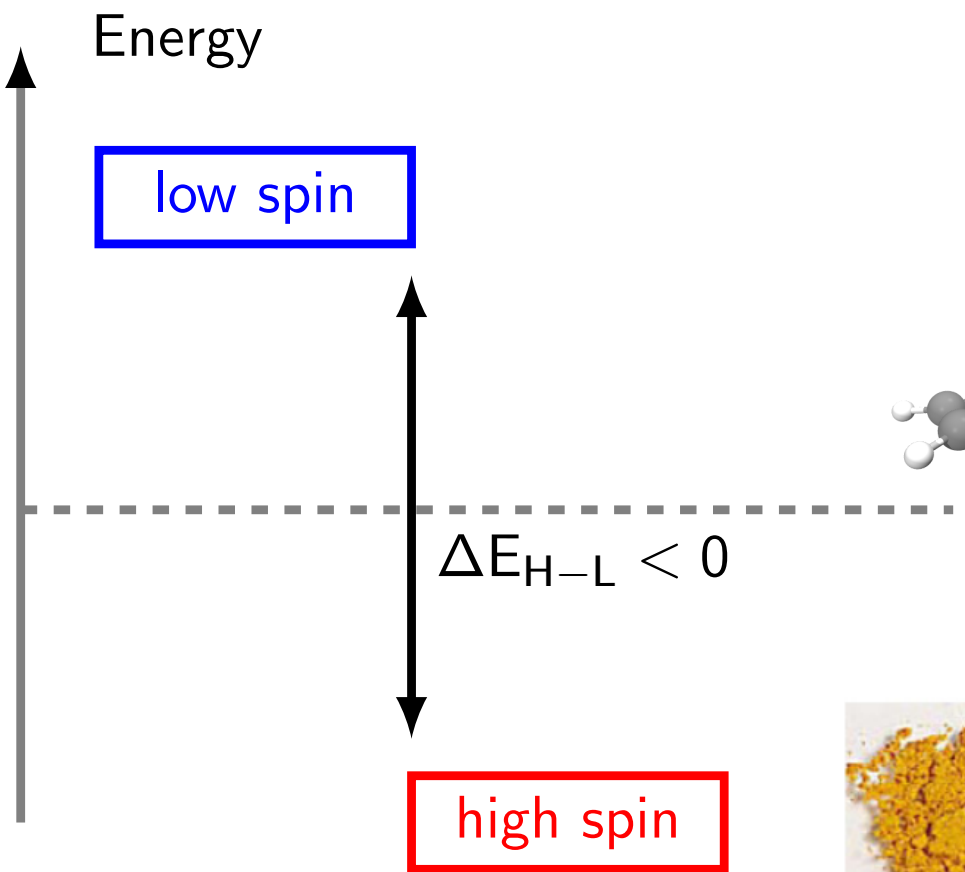
Transition metal complexes

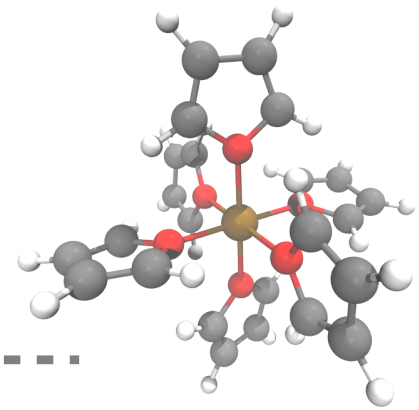


Transition metal complexes

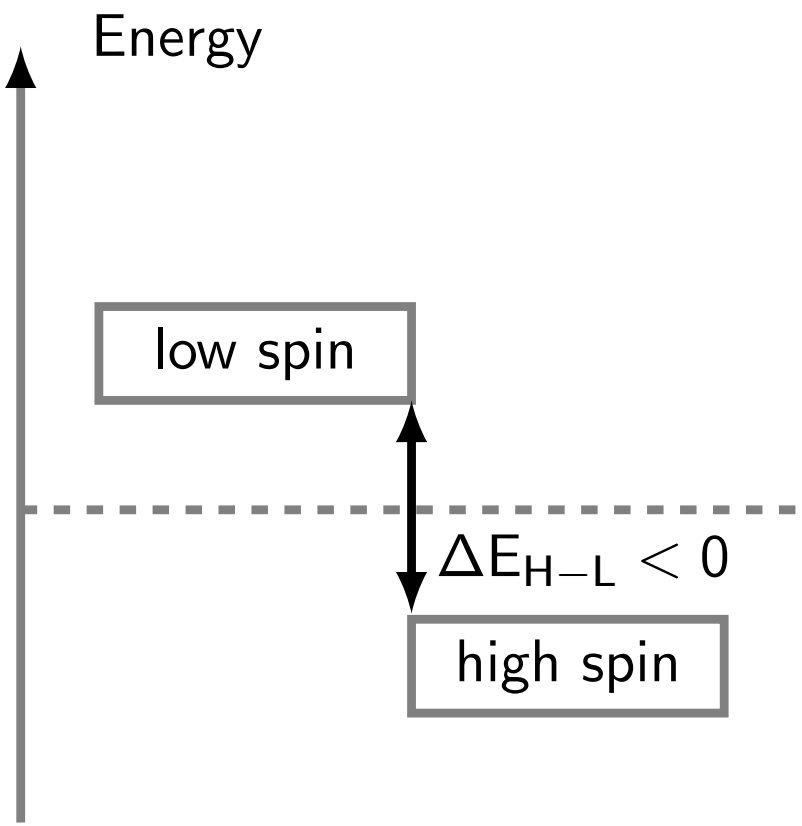


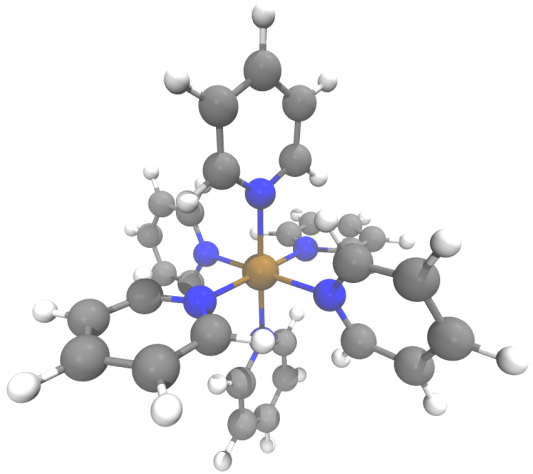
Transition metal complexes



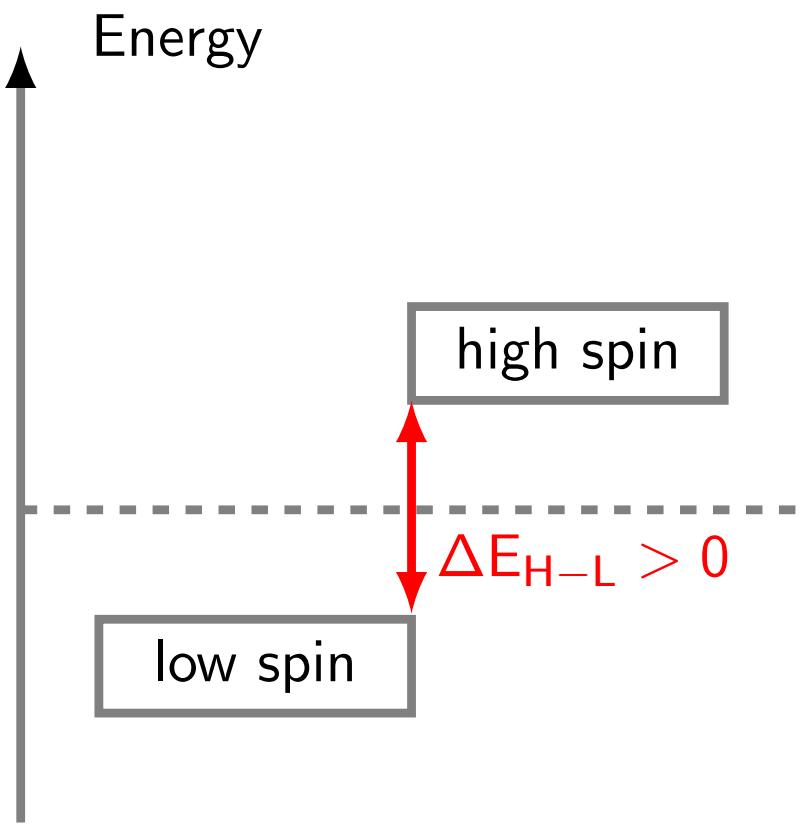


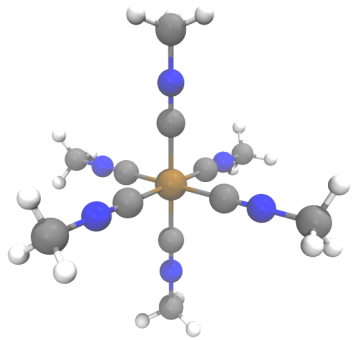
Transition metal complexes



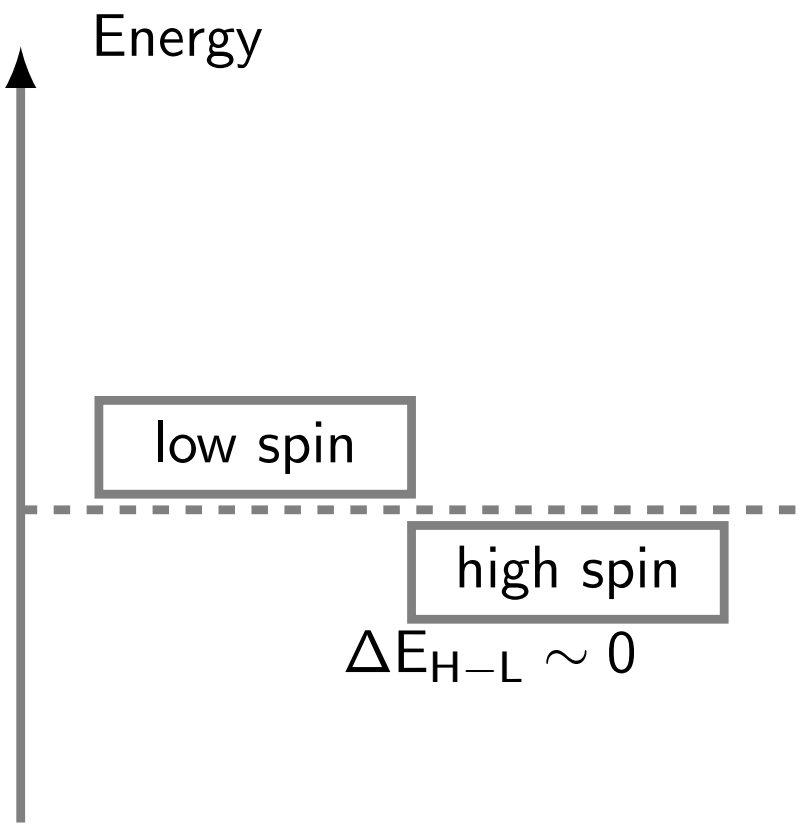


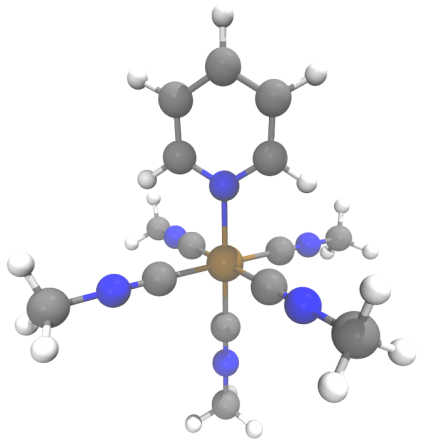
Transition metal complexes



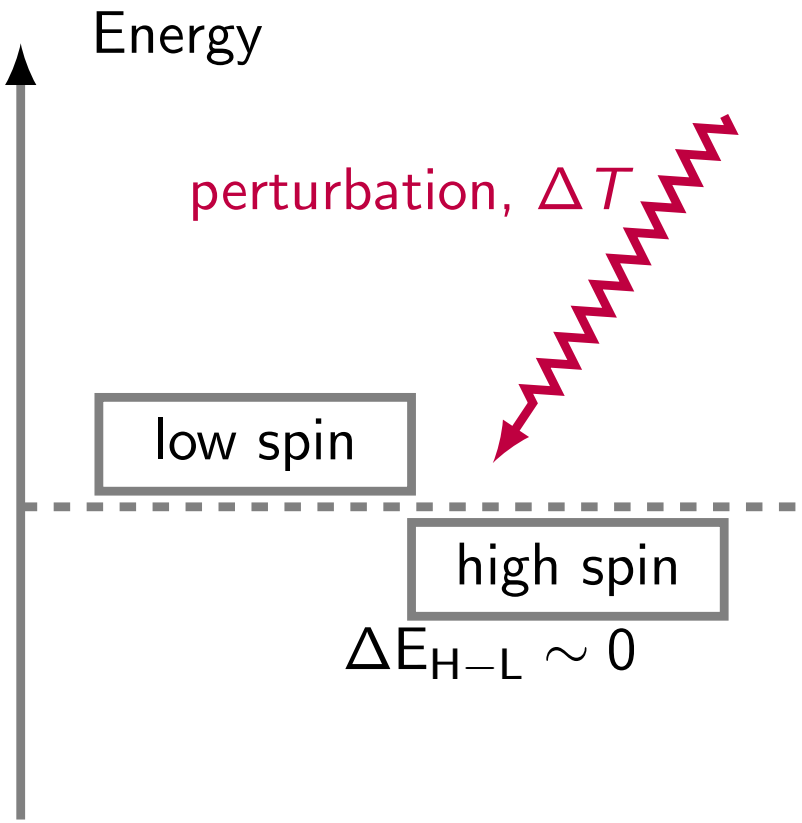


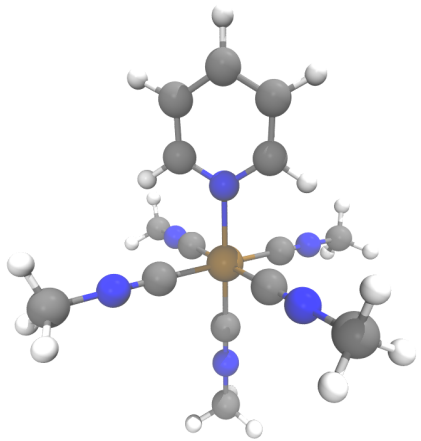
Transition metal complexes



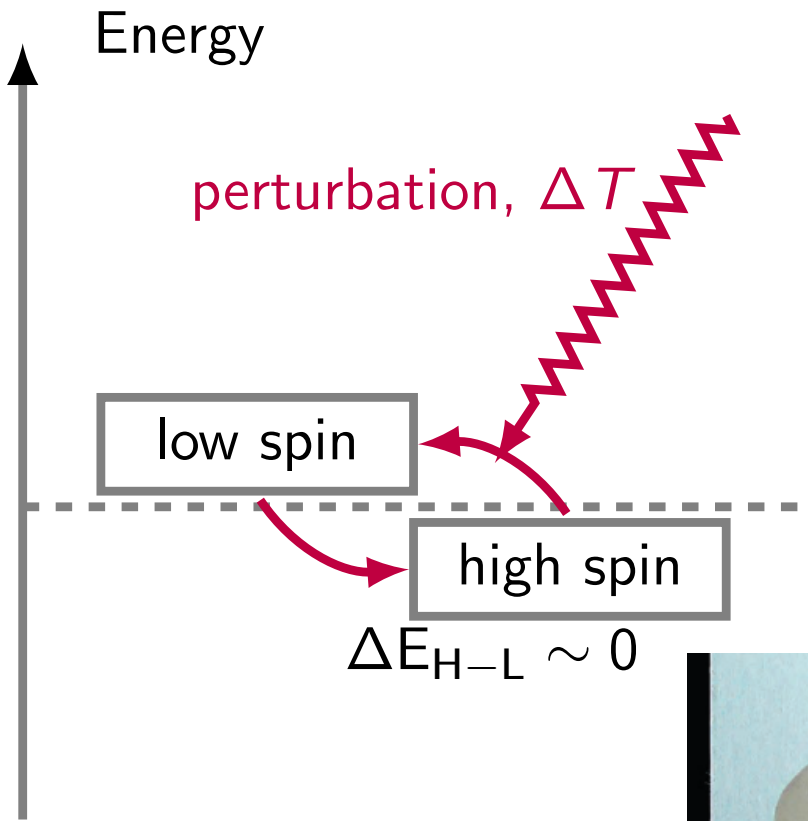


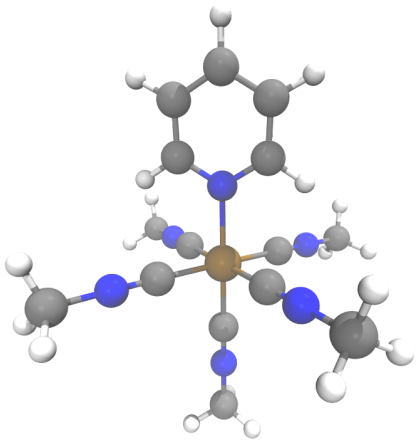
Transition metal complexes

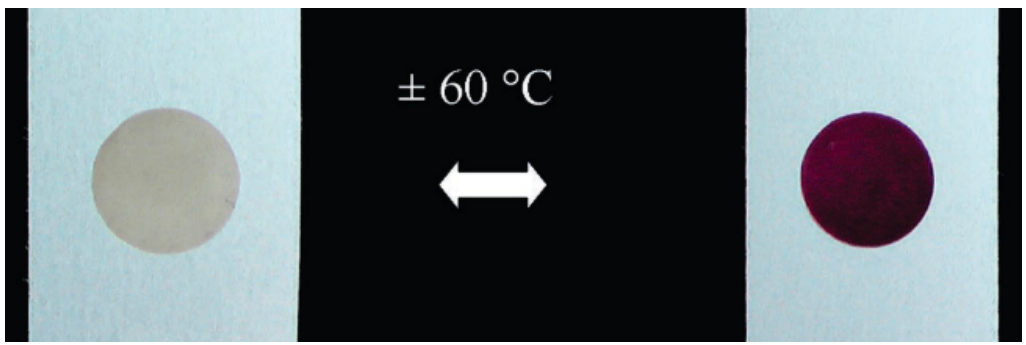




Transition metal complexes







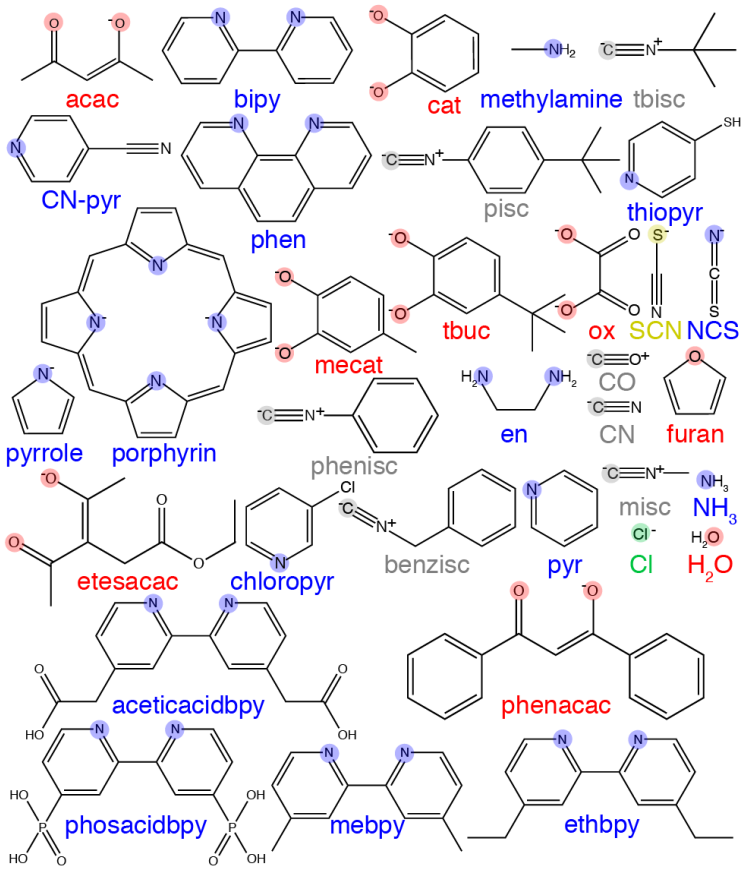
Seredyuk, M *et al.*, *Chem. Mater.*, 18(10):2513–2519, 2006.

DFT data

train on $\sim 100 - 2000$ of DFT calculations:

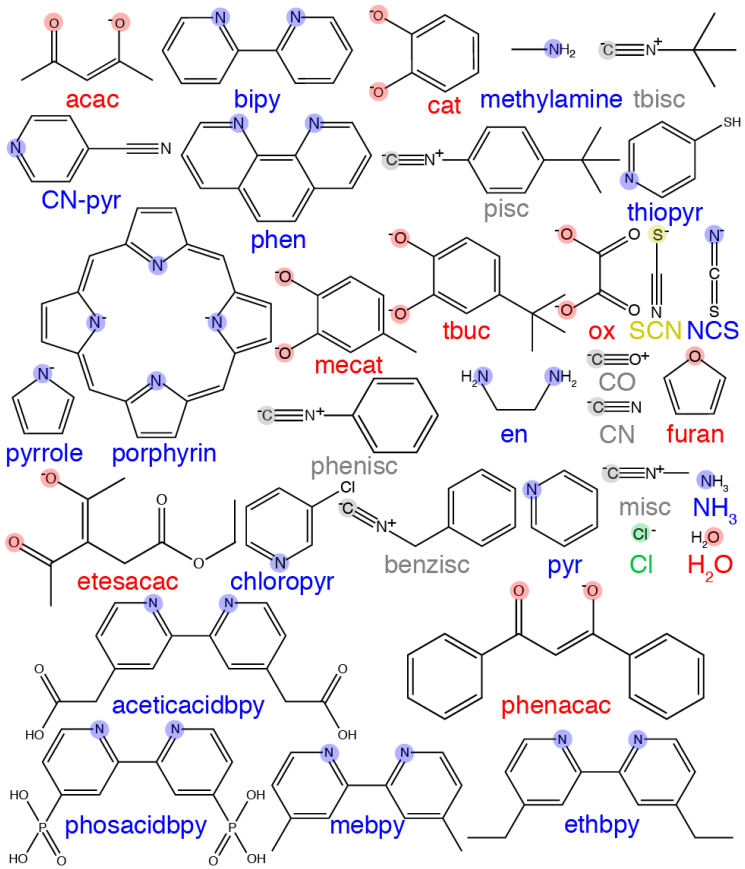
DFT data

train on $\sim 100 - 2000$ of DFT calculations:



DFT data

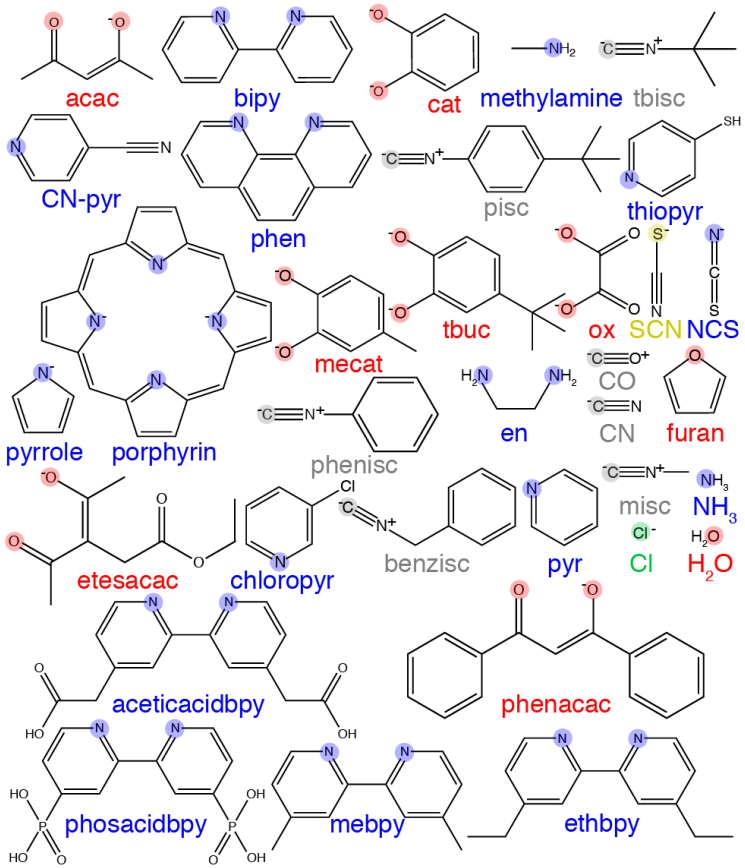
train on $\sim 100 - 2000$ of DFT calculations:

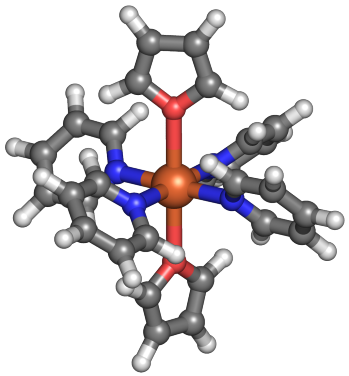


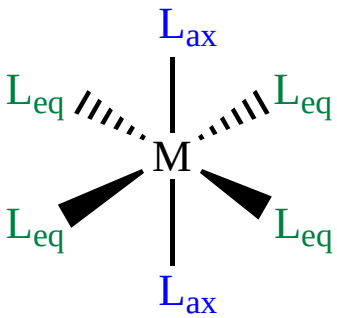
Cr	Mn	Fe	Co

DFT data

train on $\sim 100 - 2000$ of DFT calculations:



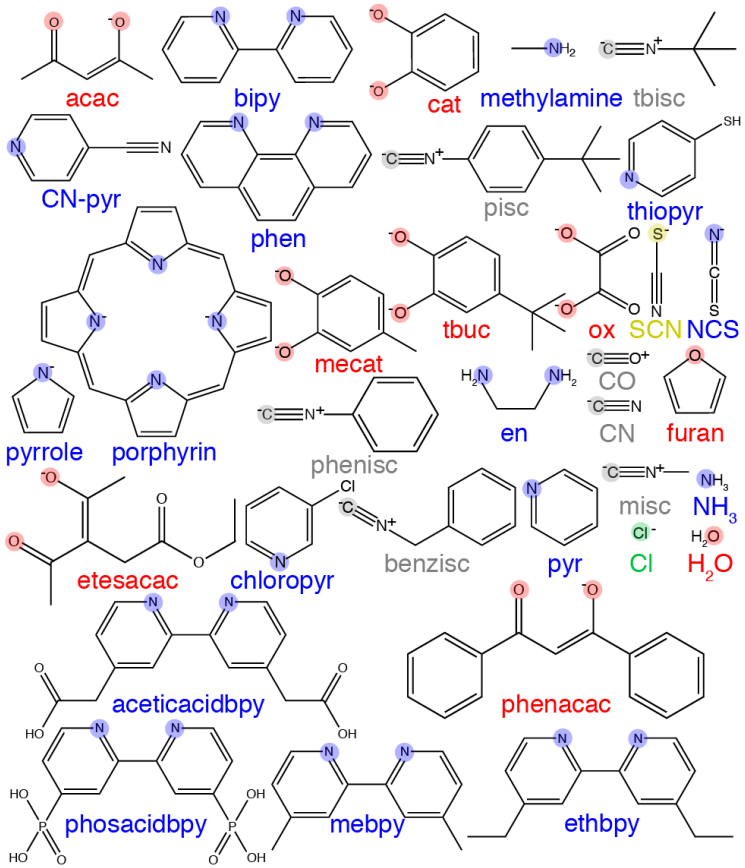


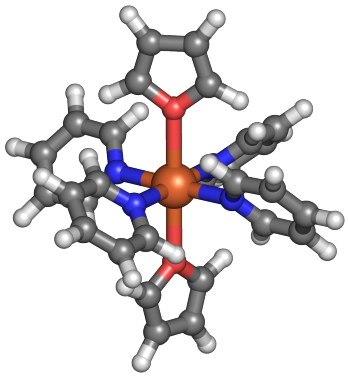


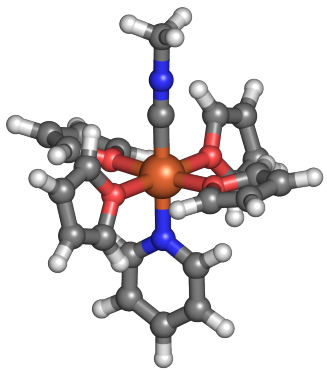
Cr	Mn	Fe	Co

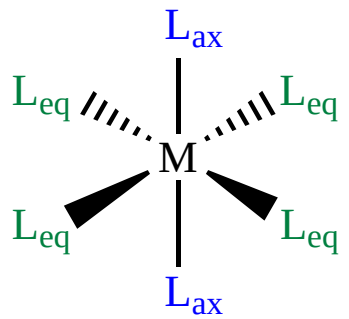
DFT data

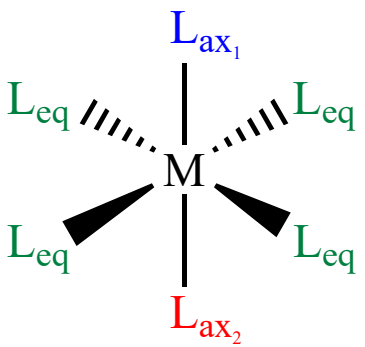
train on $\sim 100 - 2000$ of DFT calculations:









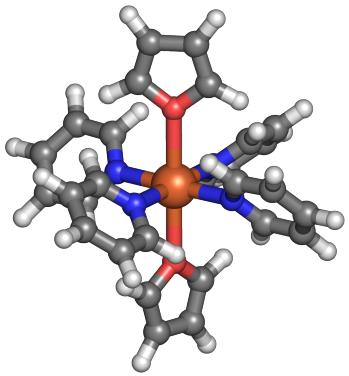


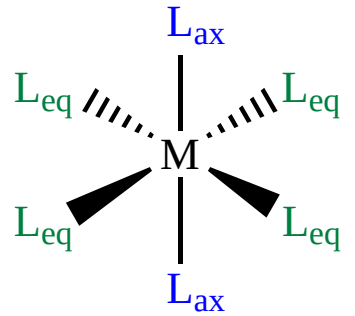
Cr	Mn	Fe	Co

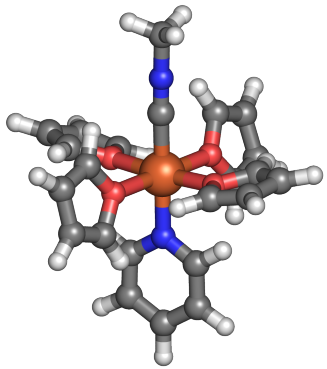
DFT data

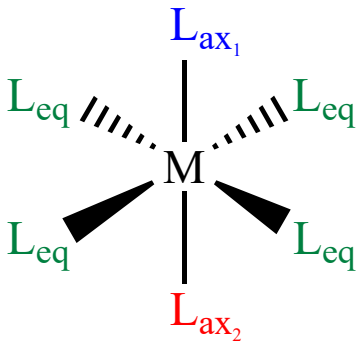
train on $\sim 100 - 2000$ of DFT calculations:

Details:
B3LYP-like DFT
gas phase optimizaton
LANL2DZ/6-31G*
high- and low-spin
M(II)/(III)









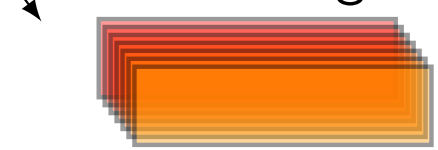
Cr	Mn	Fe	Co

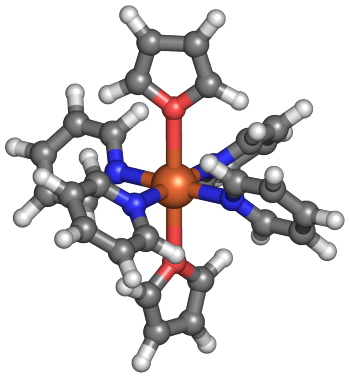
DFT data

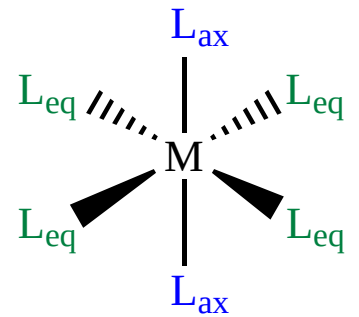
train on $\sim 100 - 2000$ of DFT calculations:

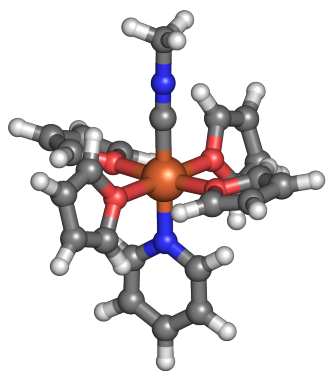
Details:
B3LYP-like DFT
gas phase optimizaton
LANL2DZ/6-31G*
high- and low-spin
M(II)/(III)

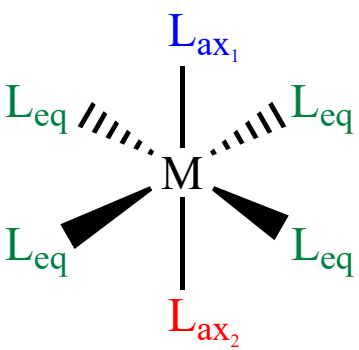
HF exchange varied $0 - 30\%$











Cr	Mn	Fe	Co

Predictive modeling of TM complexes

First attempt using simple features inspired by inorganic chem:

Predictive modeling of TM complexes

First attempt using simple features inspired by inorganic chem:

metal properties

identity

Fe(II)

oxidation state

Predictive modeling of TM complexes

First attempt using simple features inspired by inorganic chem:

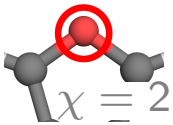
metal properties local ligand properties

identity

Fe(II)

oxidation state

$\chi = 3.44$



$\chi = 2.55$

$\max \Delta\chi$

Predictive modeling of TM complexes

First attempt using simple features inspired by inorganic chem:

metal properties

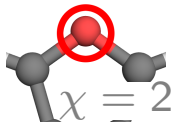
identity

Fe(II)

oxidation state

local ligand properties

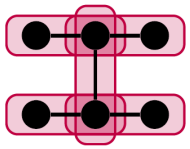
$\chi = 3.44$



$\chi = 2.55$

$\max \Delta\chi$

global properties

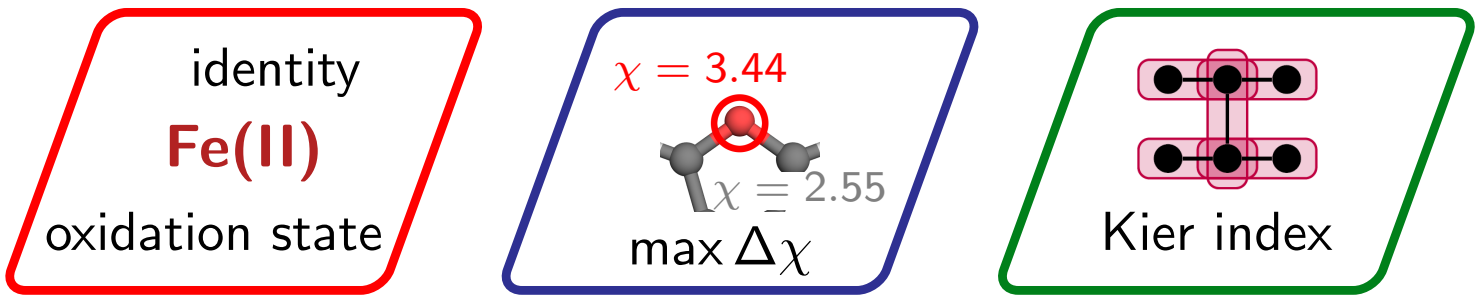


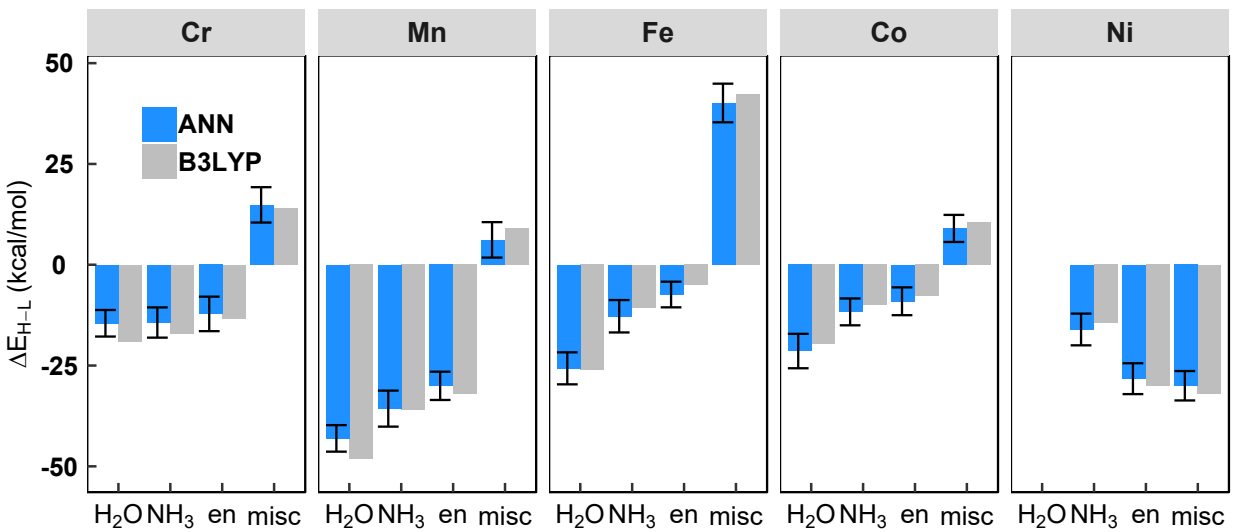
Kier index

Predictive modeling of TM complexes

First attempt using simple features inspired by inorganic chem:

metal properties | local ligand properties | global properties



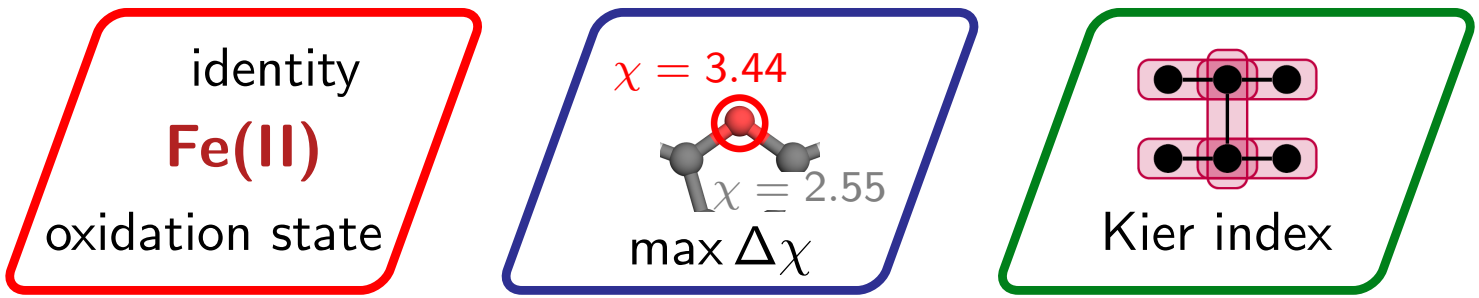


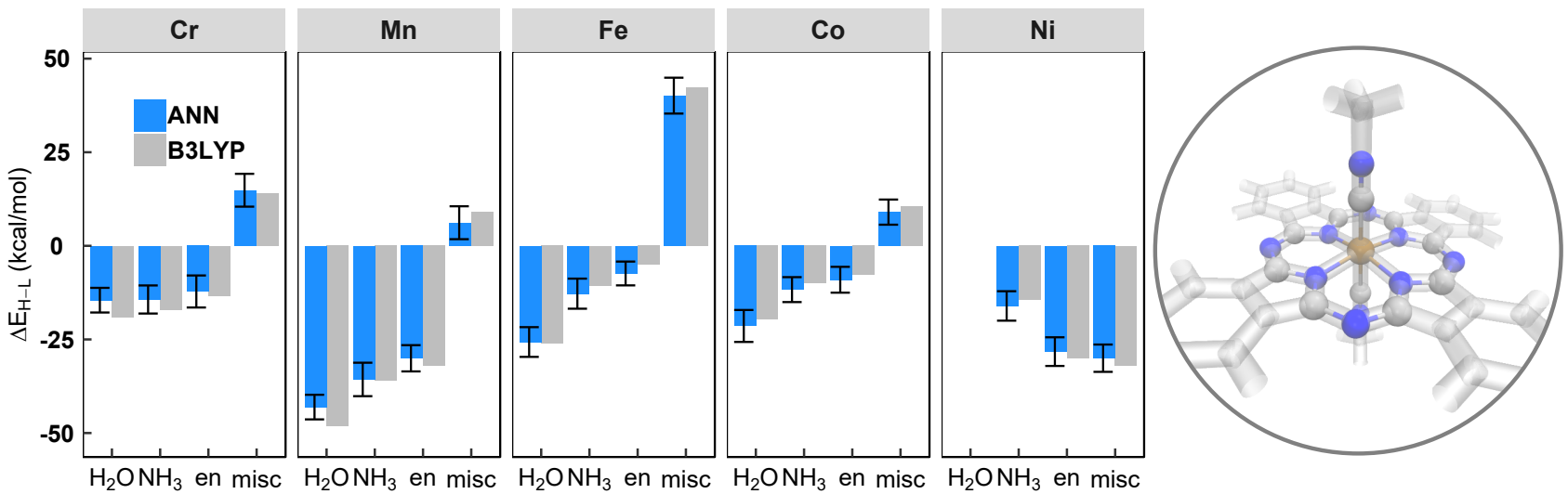
Janet, J.P. and Kulik, H.J., *Chem. Sci.*, 8:5137–5152, 2017.

Predictive modeling of TM complexes

First attempt using simple features inspired by inorganic chem:

metal properties | local ligand properties | global properties





Janet, J.P. and Kulik, H.J., *Chem. Sci.*, 8:5137–5152, 2017.

Predictive modeling of TM complexes

First attempt using simple features inspired by inorganic chem:

metal properties | local ligand properties | global properties

identity

Fe(II)

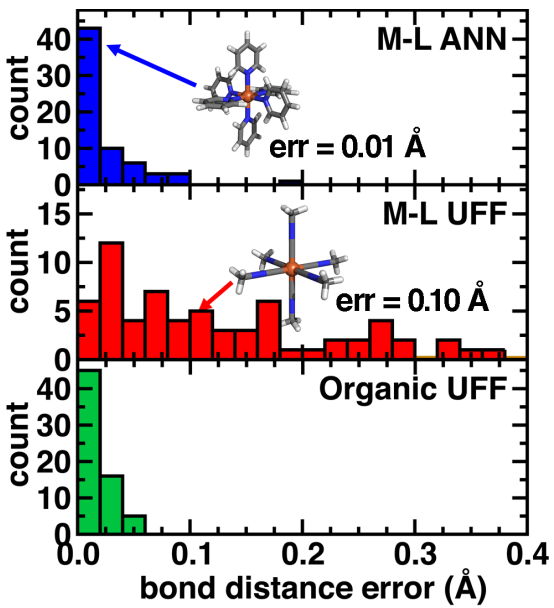
oxidation state

$\chi = 3.44$

$\chi = 2.55$

$\max \Delta\chi$

Kier index



we can predict bond lengths

Janet, J.P. *et al.*, *Inorg. Chem.*, Article ASAP, 2019

Predictive modeling of TM complexes

First attempt using simple features inspired by inorganic chem:

metal properties | local ligand properties | global properties

identity

Fe(II)

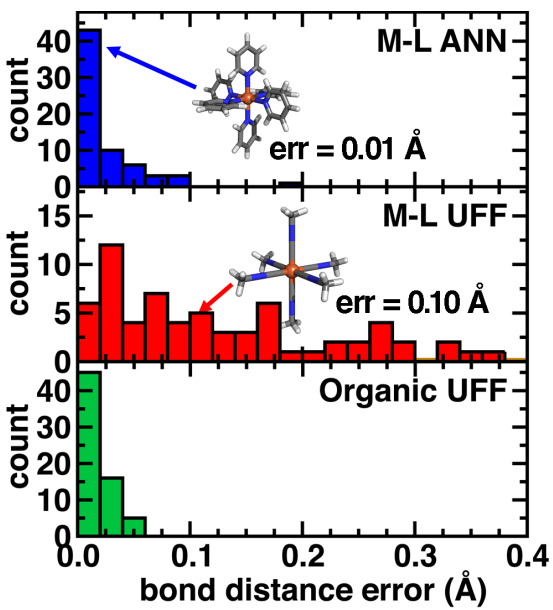
oxidation state

$\chi = 3.44$

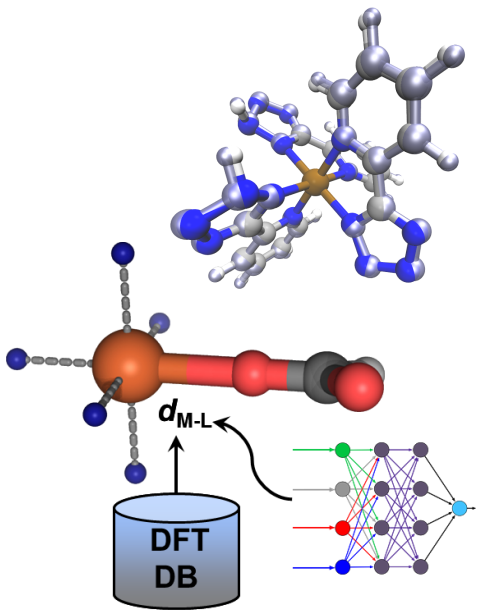
$\chi = 2.55$

$\max \Delta\chi$

Kier index



we can predict bond lengths and use this to intialize new calculations



Janet, J.P. *et al.*, *Inorg. Chem.*, Article ASAP, 2019

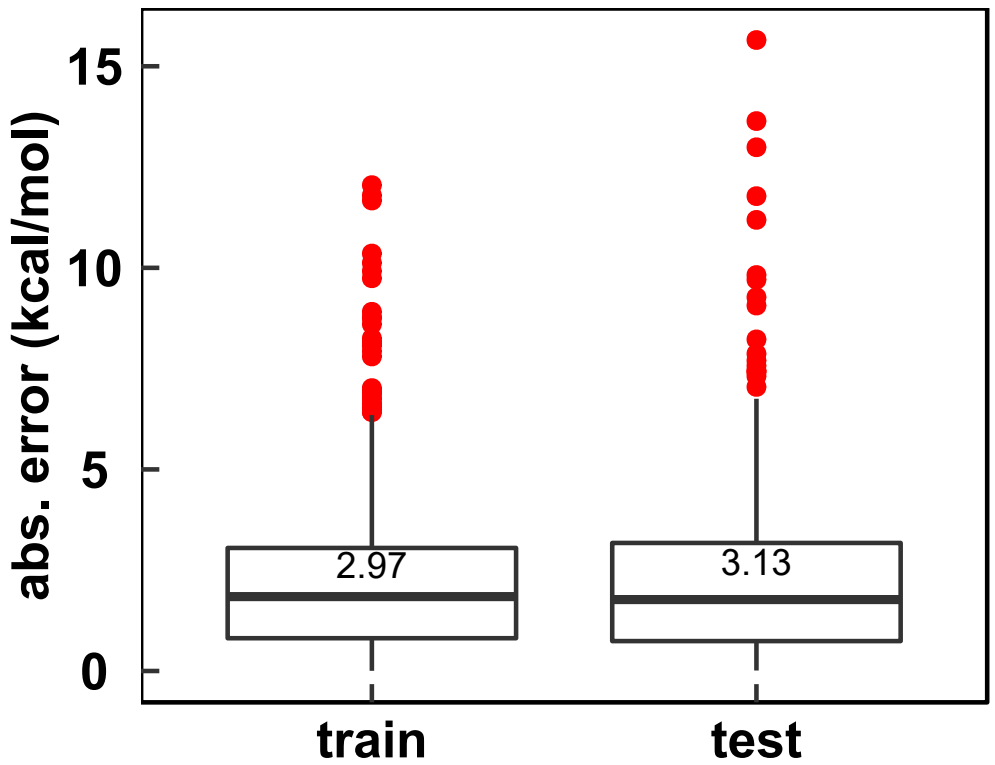
Model transferability

Test-set performance is not necessarily a good metric for general transferability[1]:

[1]:Janet, J.P., and Kulik, H.J., *Chem. Sci.*, 8:5137–5152, 2017.

Model transferability

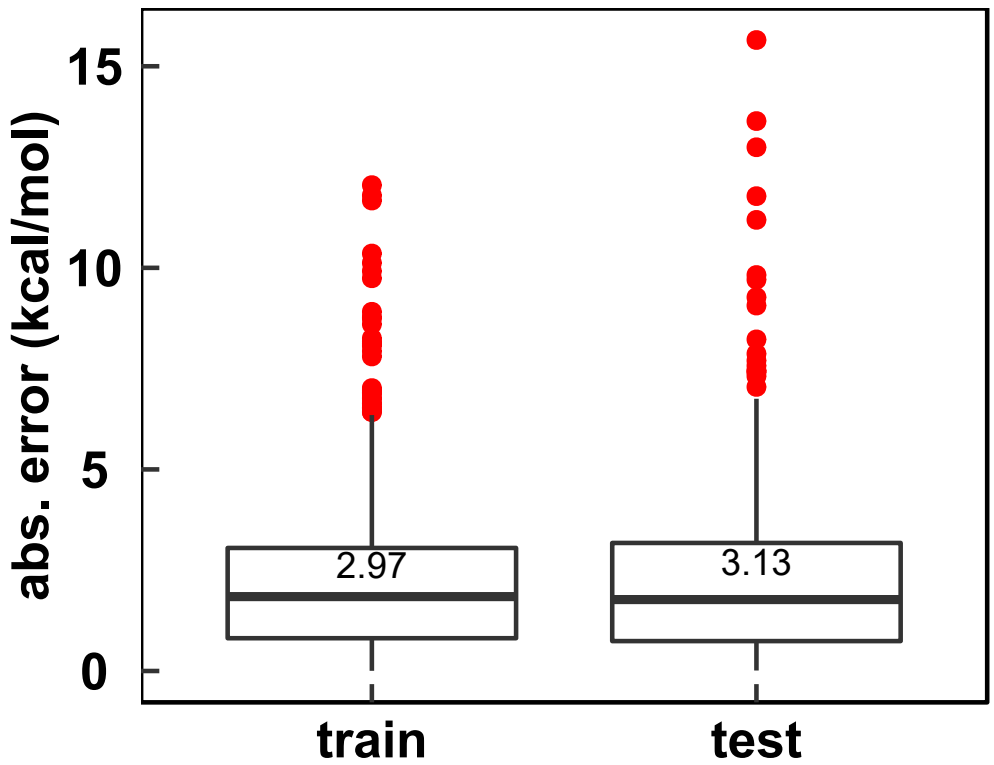
Test-set performance is not necessarily a good metric for general transferability[1]:

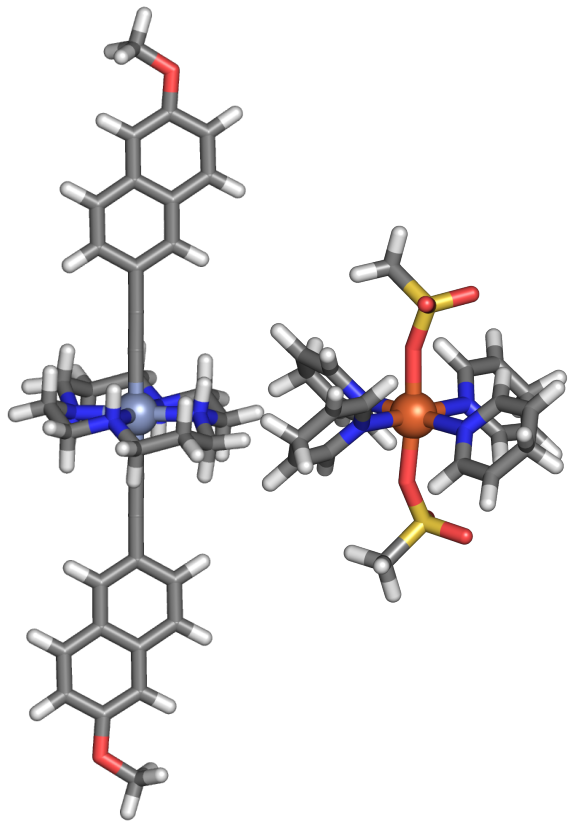


[1]:Janet, J.P., and Kulik, H.J., *Chem. Sci.*, 8:5137–5152, 2017.

Model transferability

Test-set performance is not necessarily a good metric for general transferability[1]:

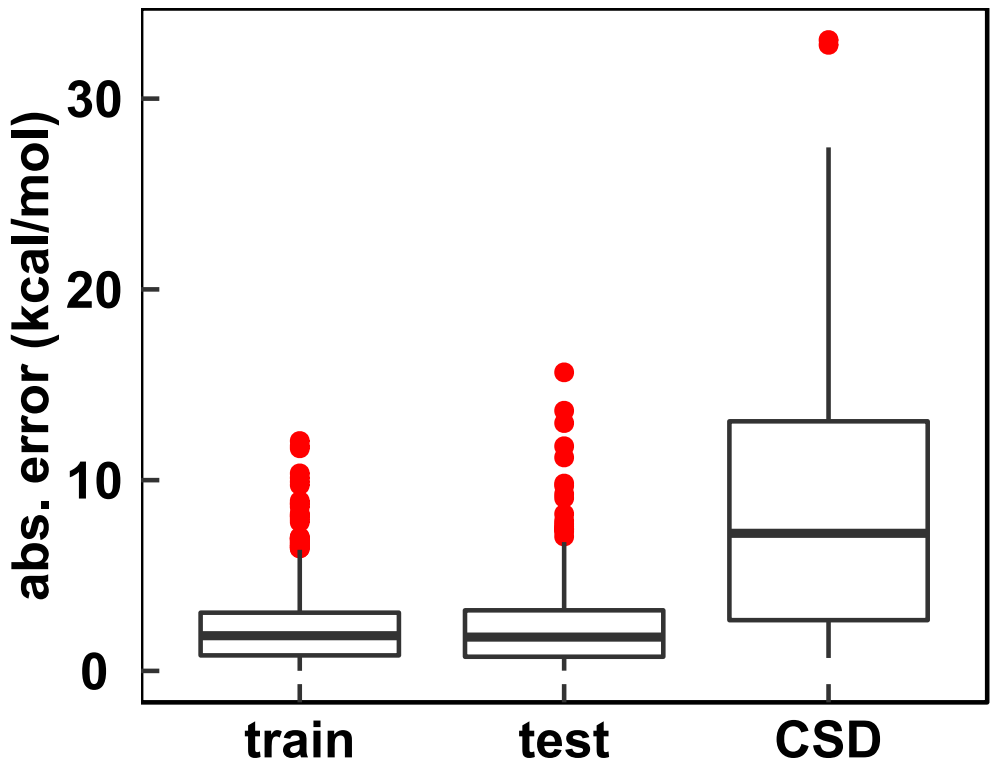


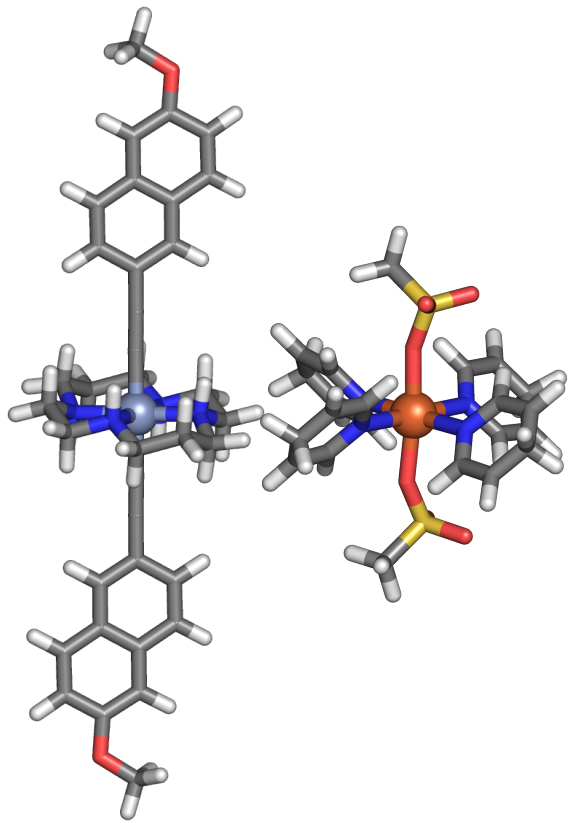


[1]:Janet, J.P., and Kulik, H.J., *Chem. Sci.*, 8:5137–5152, 2017.

Model transferability

Test-set performance is not necessarily a good metric for general transferability[1]:





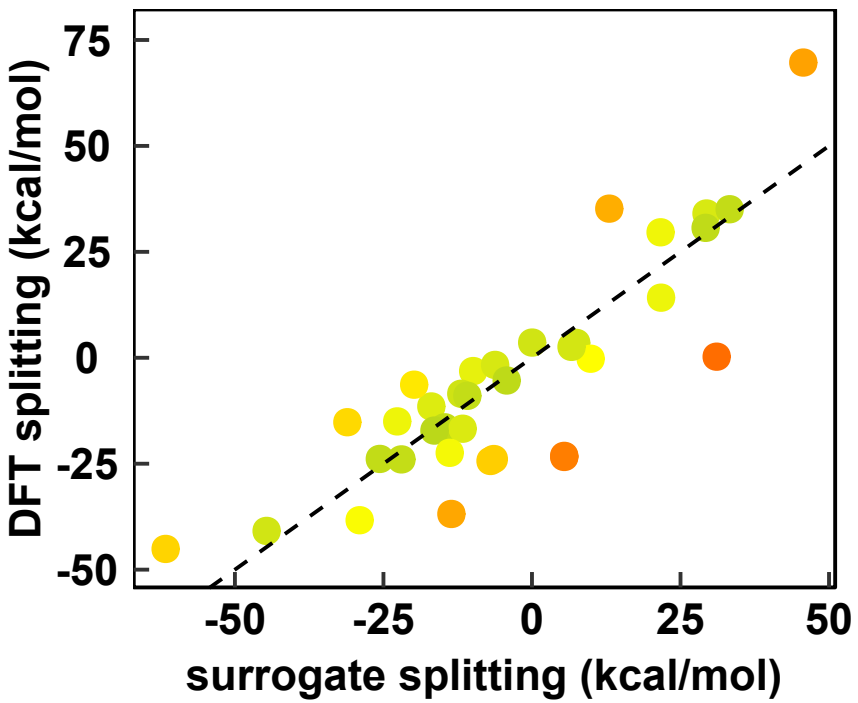
[1]:Janet, J.P., and Kulik, H.J., *Chem. Sci.*, 8:5137–5152, 2017.

System-specific generalization

In practice, model performance is highly variable:

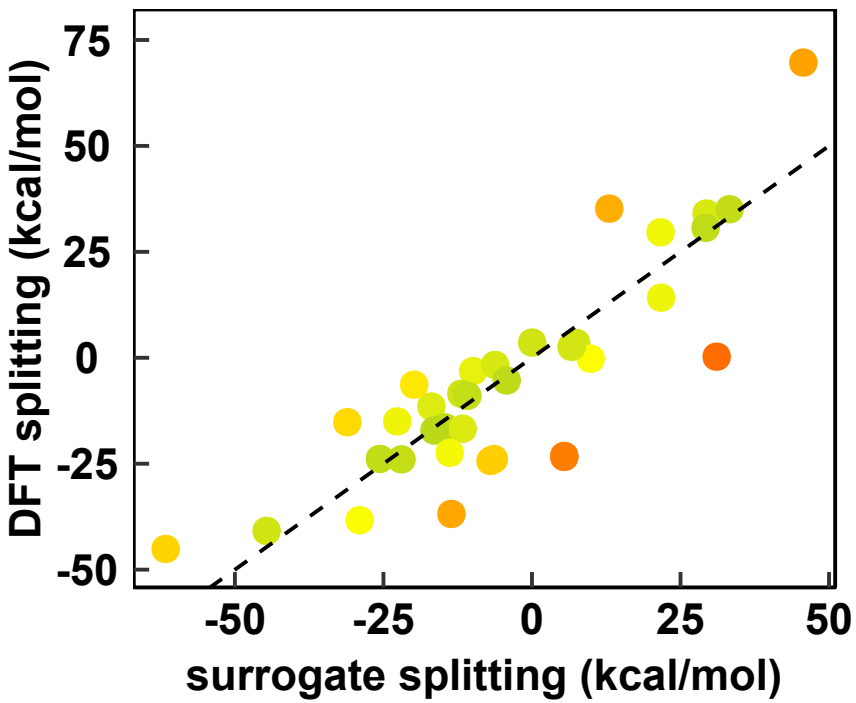
System-specific generalization

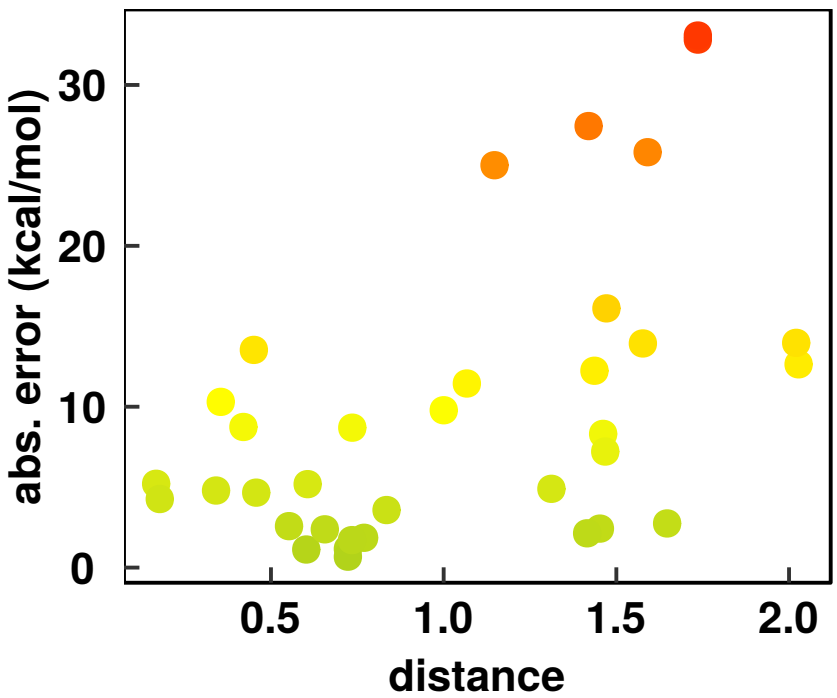
In practice, model performance is highly variable:



System-specific generalization

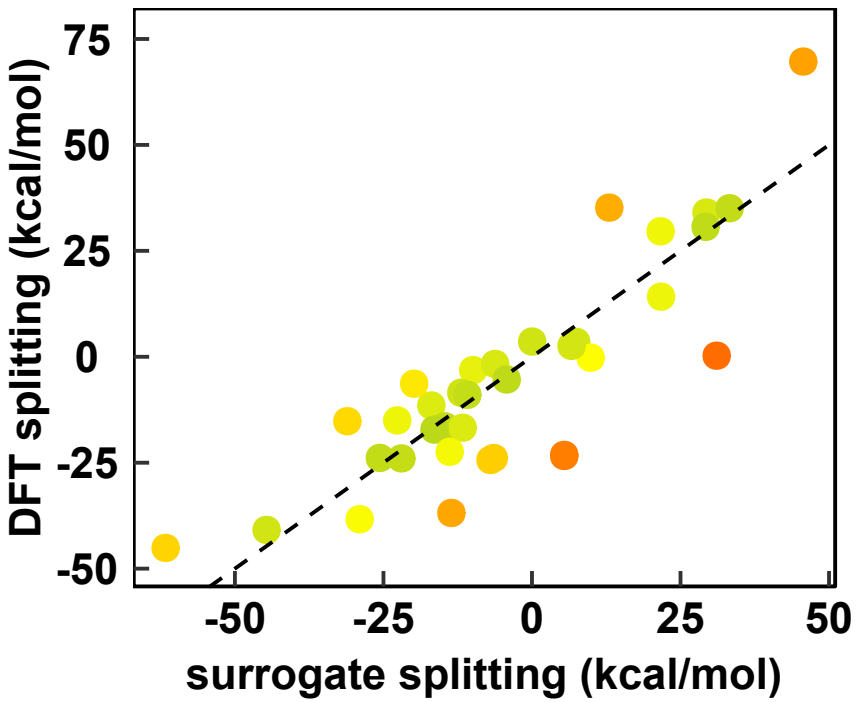
In practice, model performance is highly variable:

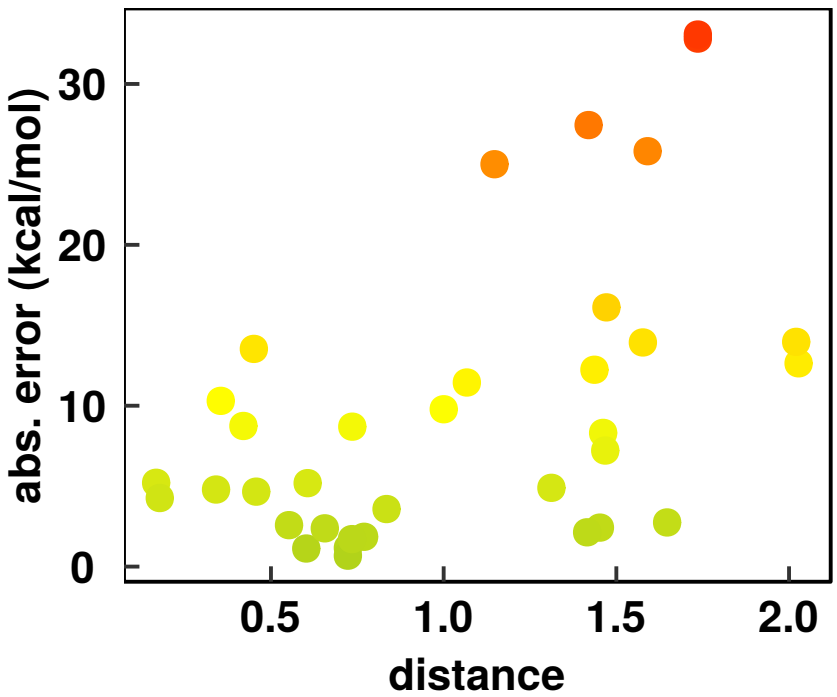




System-specific generalization

In practice, model performance is highly variable:



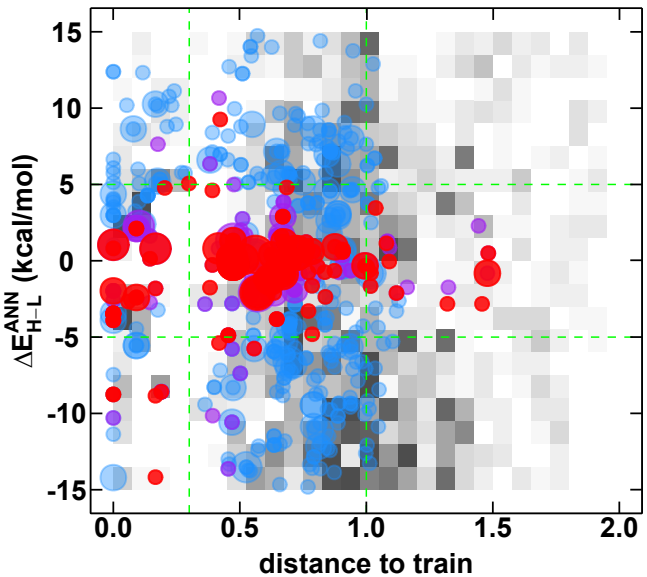


using simple distance worked pretty well!

TM complex discovery with ANNs

Spin splitting design:

We combine ANN predictions and uncertainties using an evolutionary algorithm.
Error control allows 60% of leads to be validated with DFT.[1]

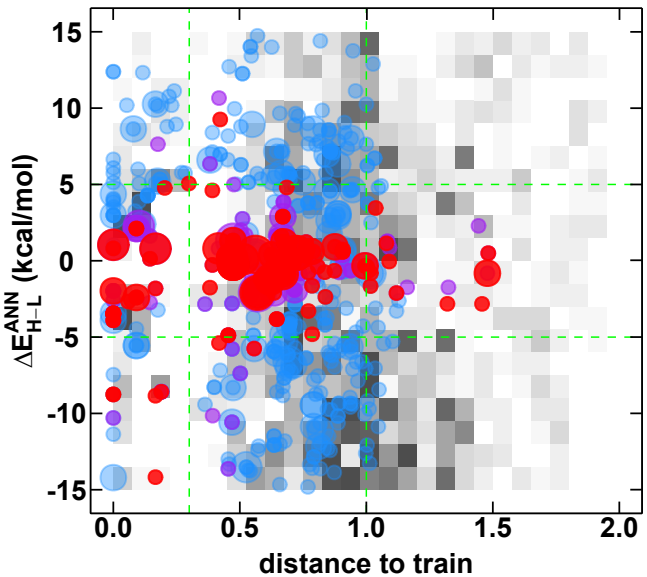


[1]: Janet, J.P., Chan, L. and Kulik, H.J., *J. Phys. Chem. Lett.*, 9(5):1064–1071, 2018.

TM complex discovery with ANNs

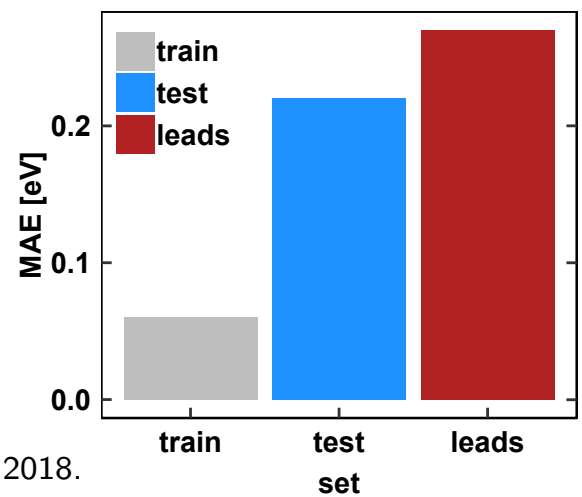
Spin splitting design:

We combine ANN predictions and uncertainties using an evolutionary algorithm.
Error control allows 60% of leads to be validated with DFT.[1]



Frontier orbital properties:

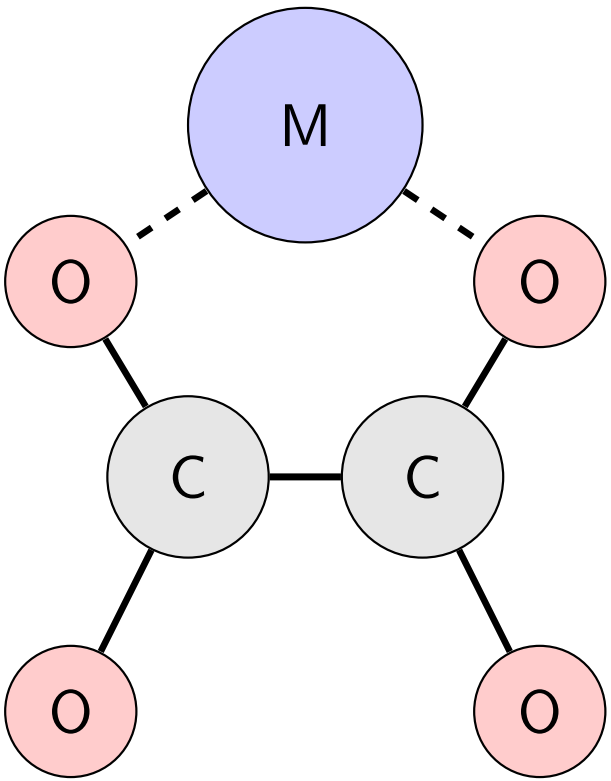
This approach also works for frontier orbtial design[2], obtaining average HOMO of 3.98 eV compared to target 4.00 eV.



[1]: Janet, J.P., Chan, L. and Kulik, H.J., *J. Phys. Chem. Lett.*, 9(5):1064–1071, 2018.
[2]: Nandy, A. *et al.*, *Ind. Eng. Chem. Res.*, 57(42):13973–13986, 2018.

More complex representations

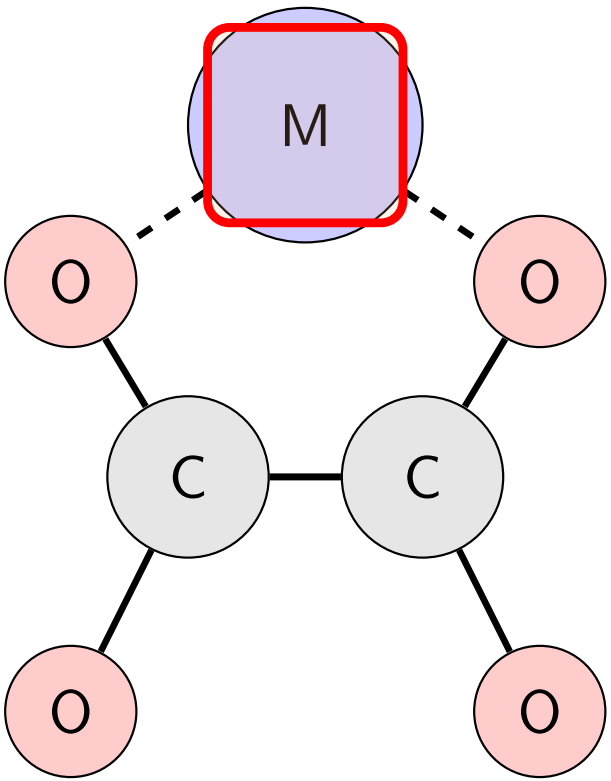
We introduced a new representation based on graph information designed for TM complexes[1]:



[1]:Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A* 121(46):8939–8954, 2017.

More complex representations

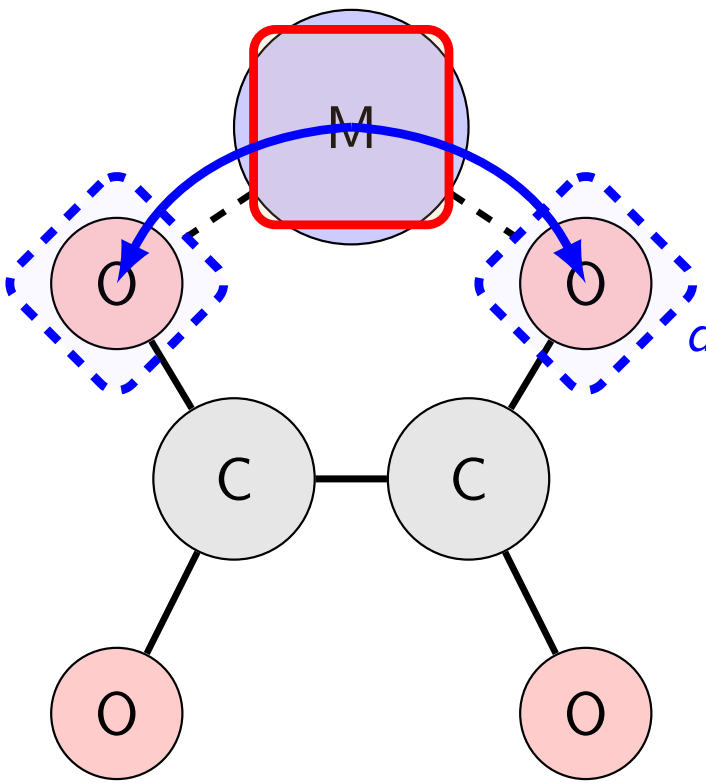
We introduced a new representation based on graph information designed for TM complexes[1]:



[1]:Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A* 121(46):8939–8954, 2017.

More complex representations

We introduced a new representation based on graph information designed for TM complexes[1]:

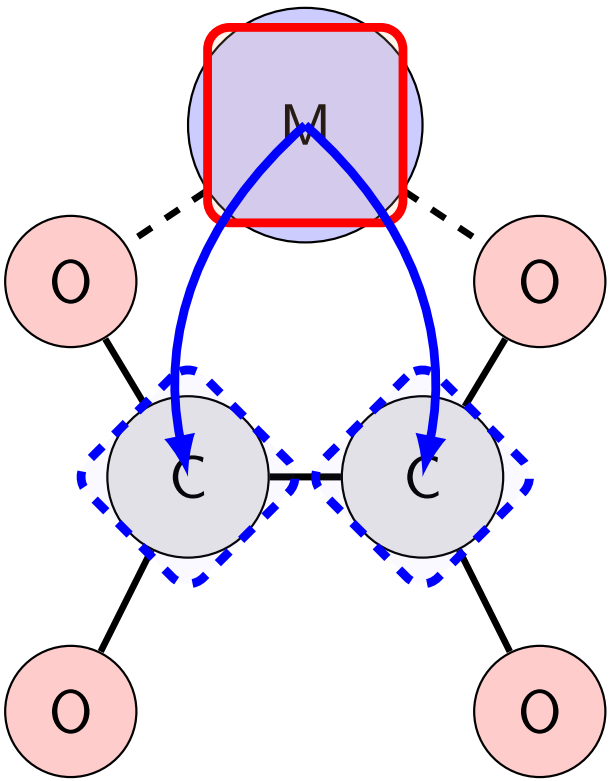


$$d_1 : \sum_{M,O} Z_M Z_O$$

[1]:Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A* 121(46):8939–8954, 2017.

More complex representations

We introduced a new representation based on graph information designed for TM complexes[1]:



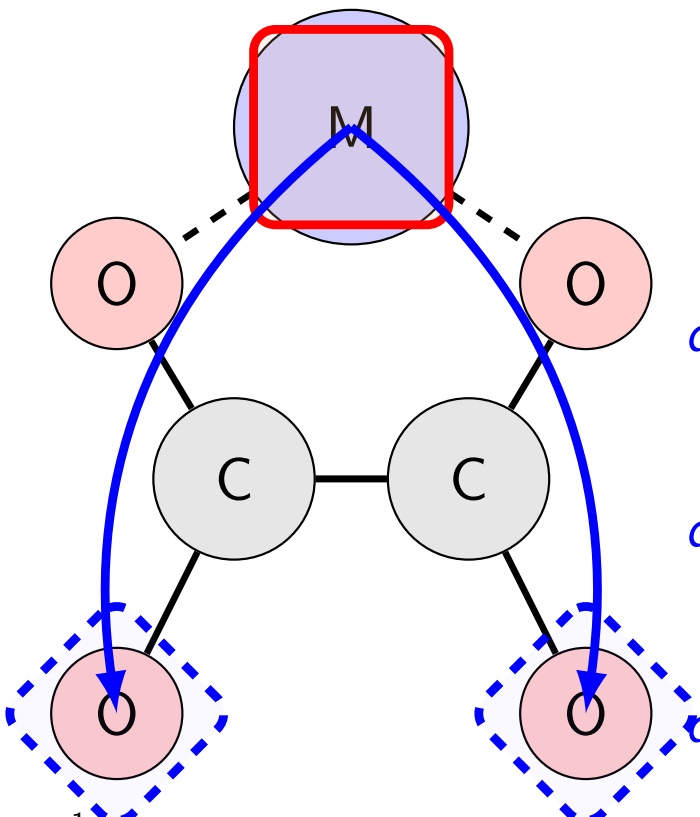
$$d_1 : \sum_{M,O} Z_M Z_O$$

$$d_2 : \sum_{M,C} Z_M Z_C$$

[1]:Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A* 121(46):8939–8954, 2017.

More complex representations

We introduced a new representation based on graph information designed for TM complexes[1]:



$$d_1 : \sum_{M,O} Z_M Z_O$$

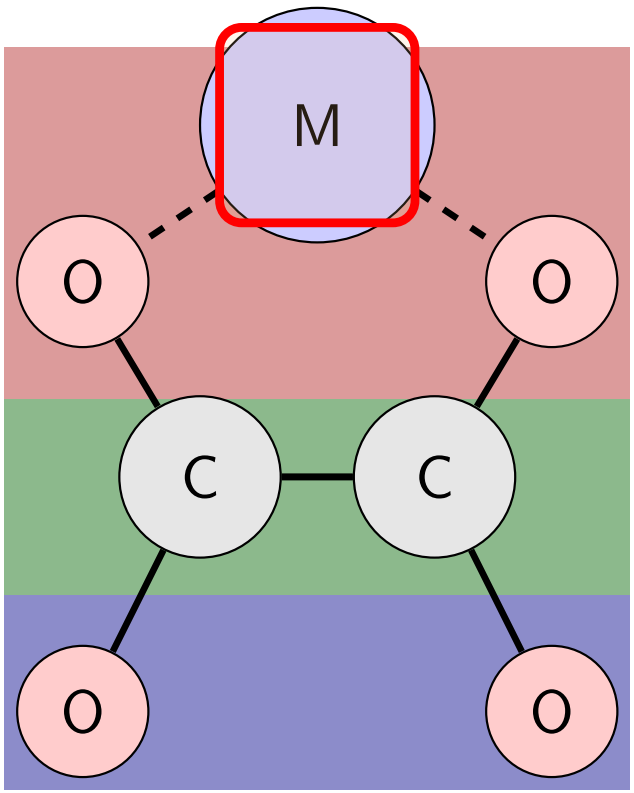
$$d_2 : \sum_{M,C} Z_M Z_C$$

$$d_3 : \sum_{M,O} Z_M Z_O$$

[1]:Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A* 121(46):8939–8954, 2017.

More complex representations

We introduced a new representation based on graph information designed for TM complexes[1]:



$$d_1 : \sum_{M,O} Z_M Z_O$$

$$d_2 : \sum_{M,C} Z_M Z_C$$

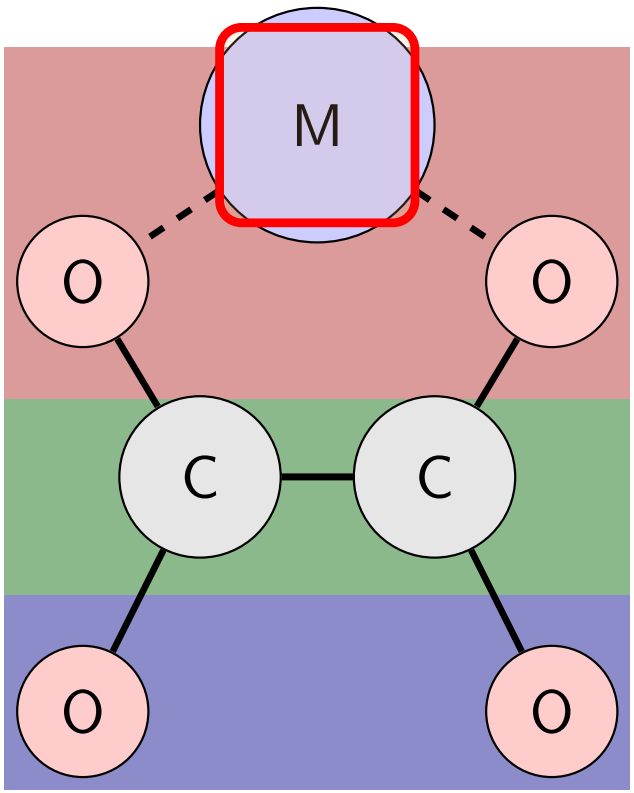
$$d_3 : \sum_{M,O} Z_M Z_O$$

[1]:Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A* 121(46):8939–8954, 2017.

More complex representations

We introduced a new representation based on graph information designed for TM complexes[1]:

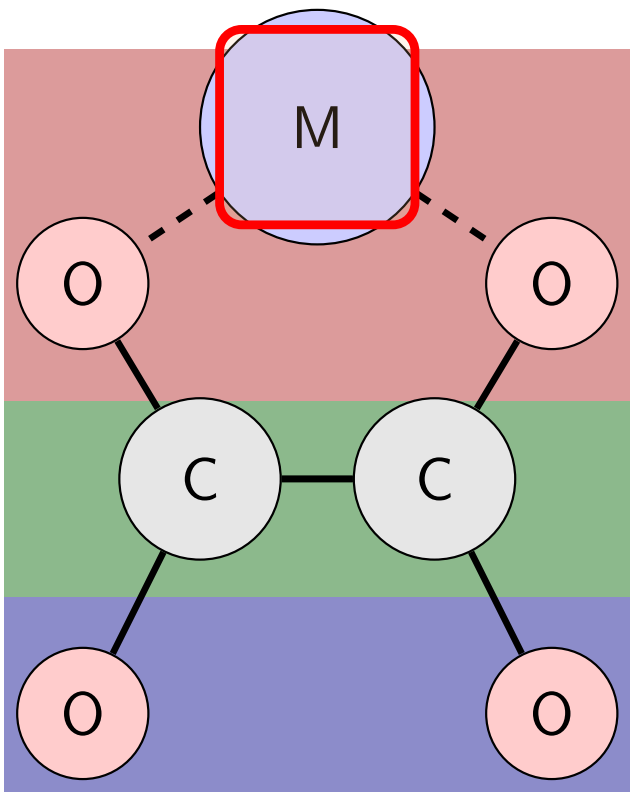
$\sim$ 160 features in total



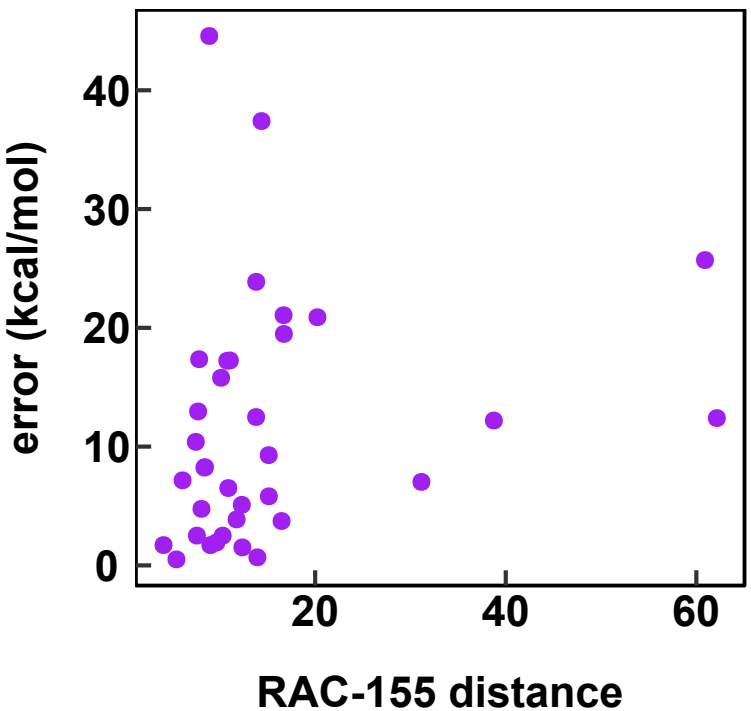
[1]:Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A* 121(46):8939–8954, 2017.

More complex representations

We introduced a new representation based on graph information designed for TM complexes[1]:



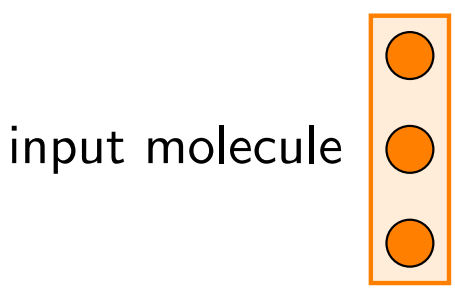
$\sim$ 160 features in total



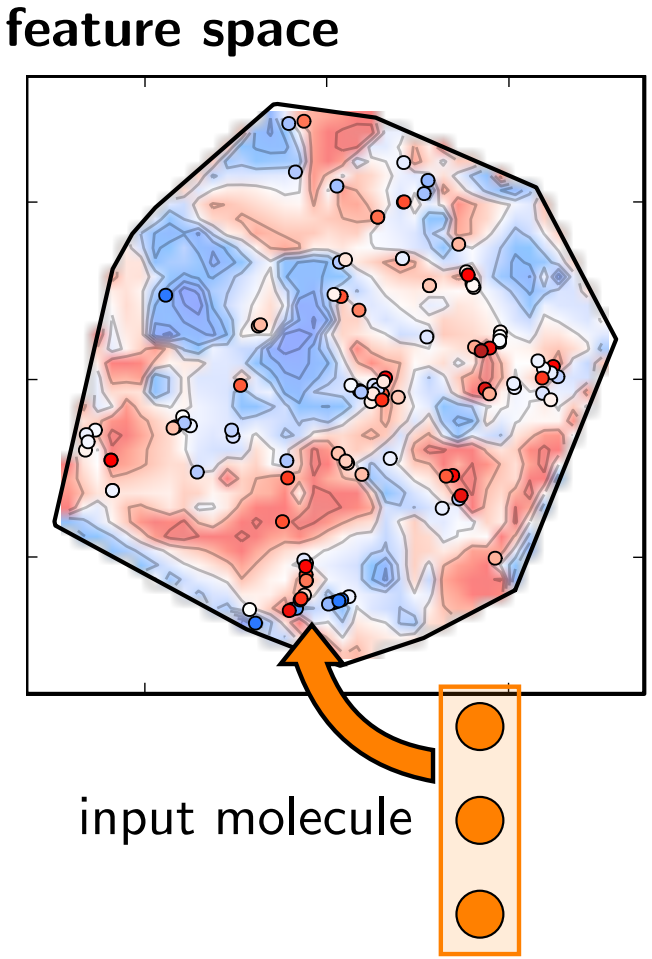
[1]:Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A* 121(46):8939–8954, 2017.

How ANNs work

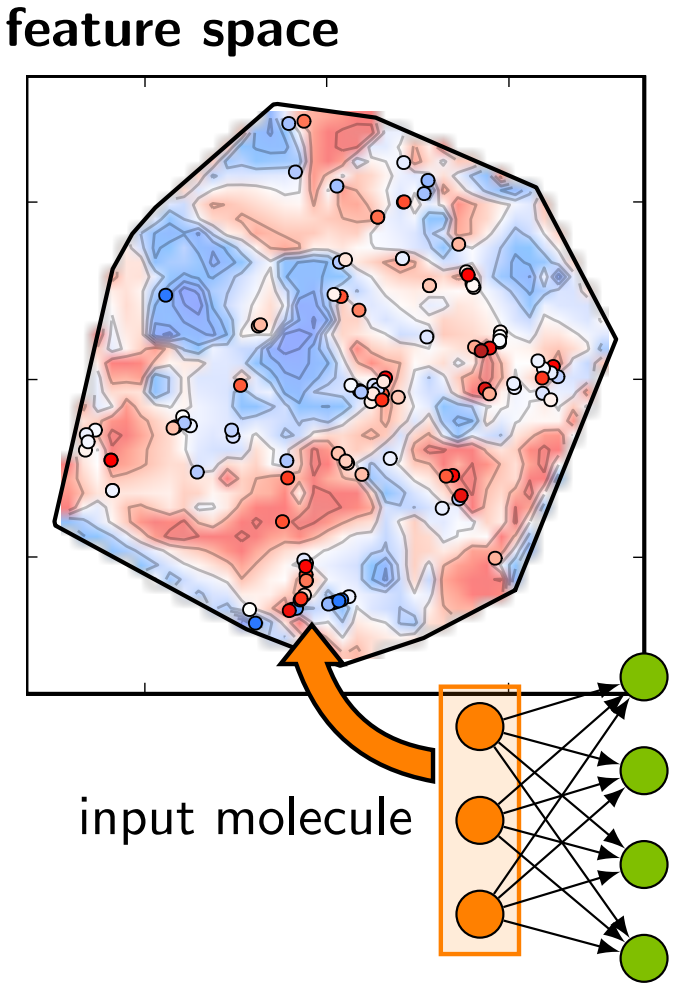
How ANNs work



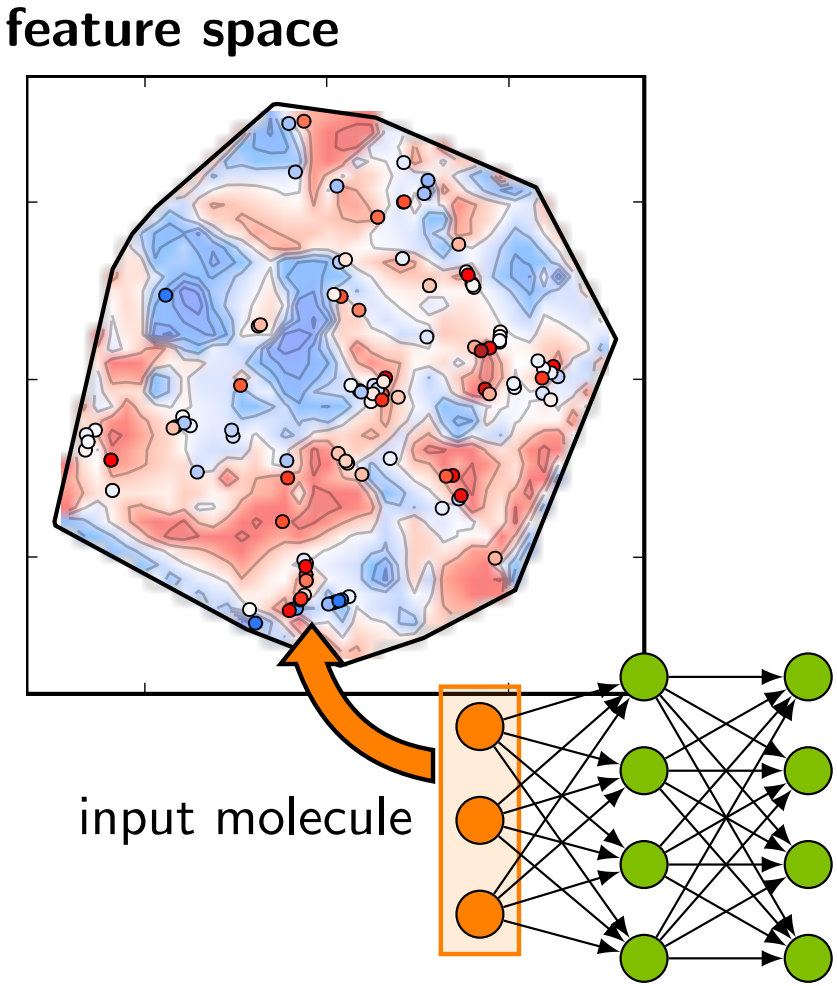
How ANNs work



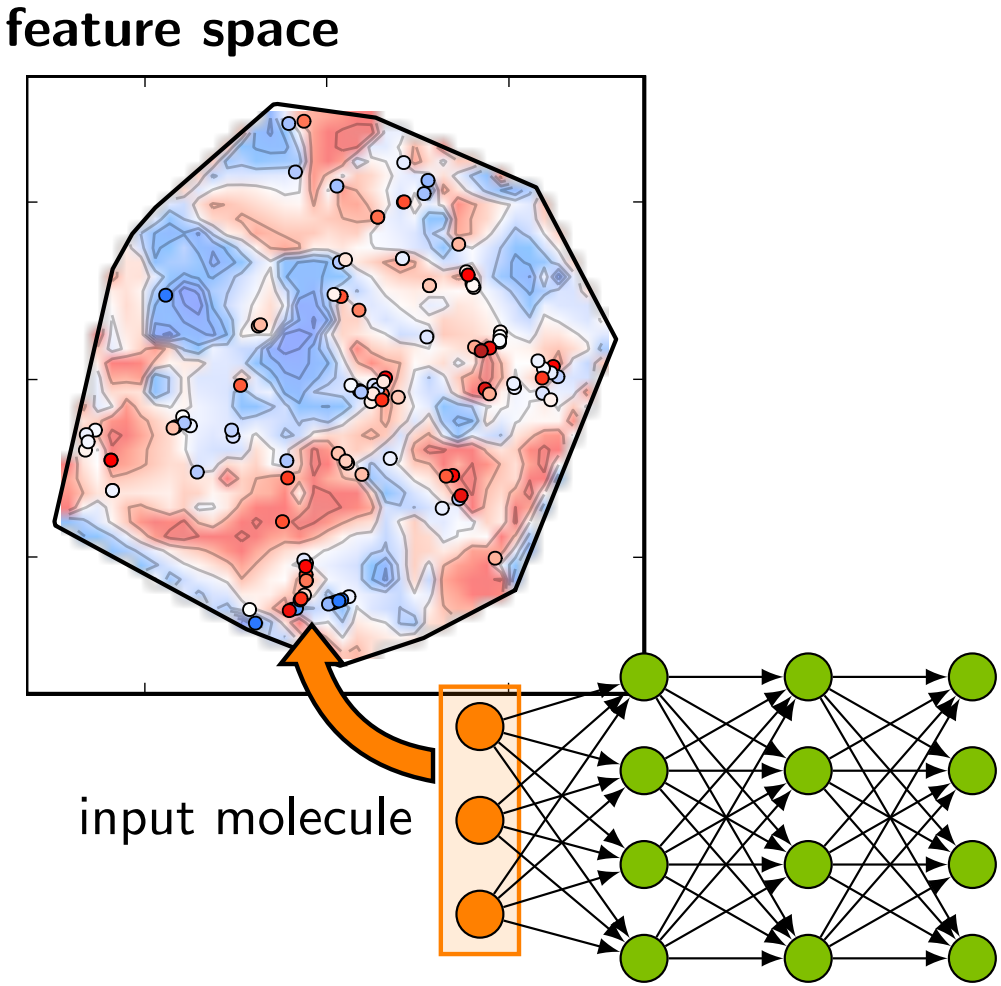
How ANNs work



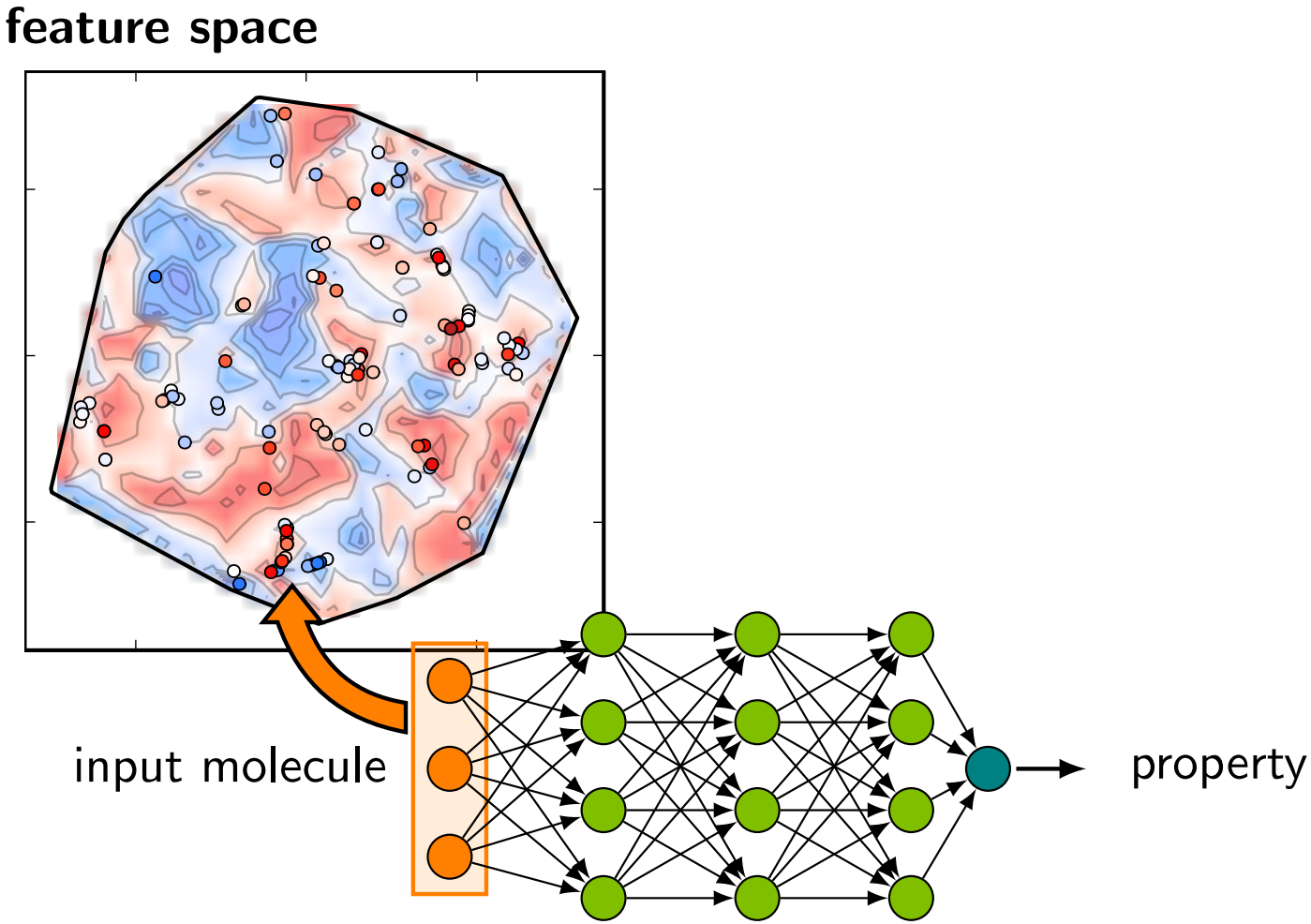
How ANNs work



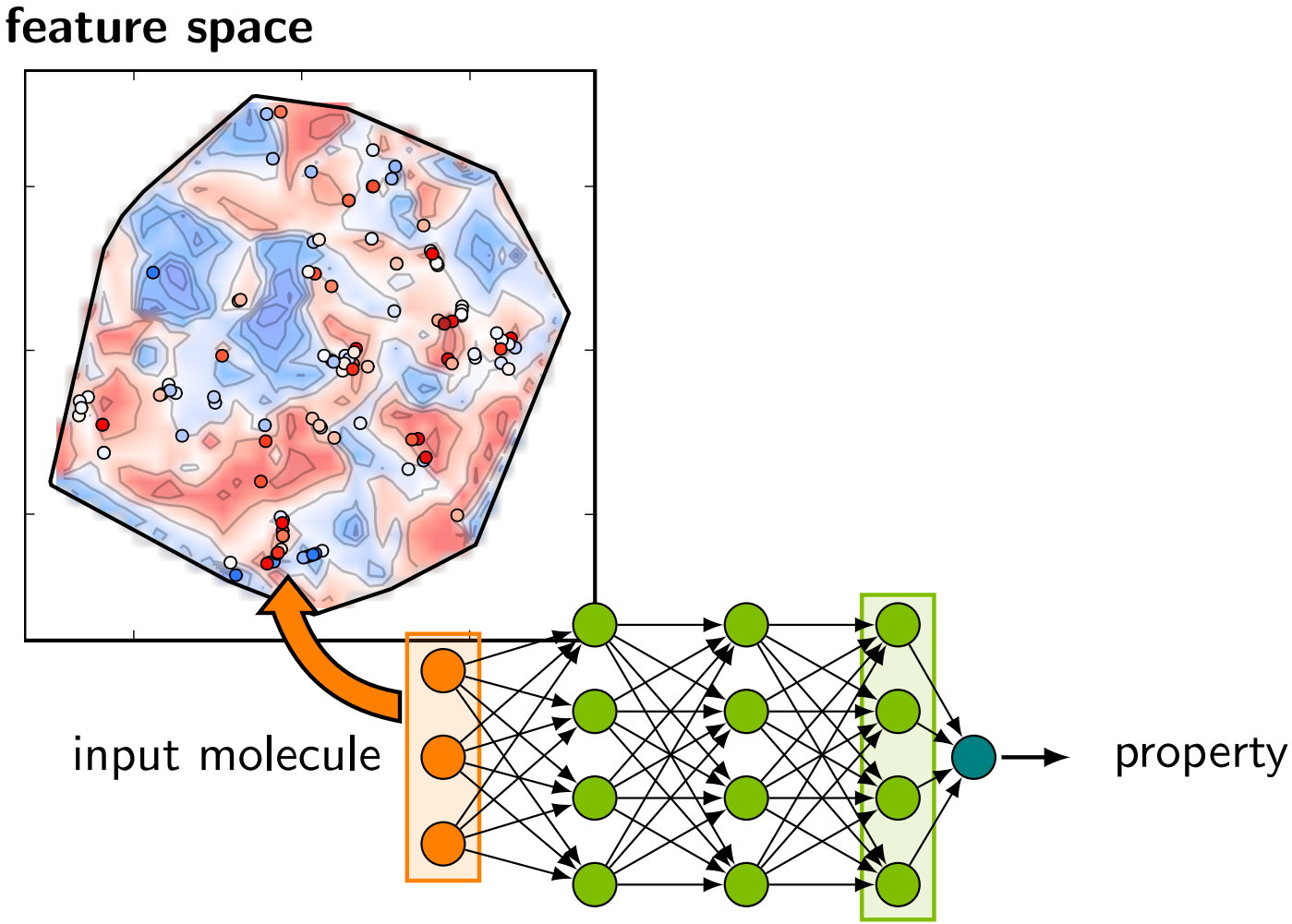
How ANNs work



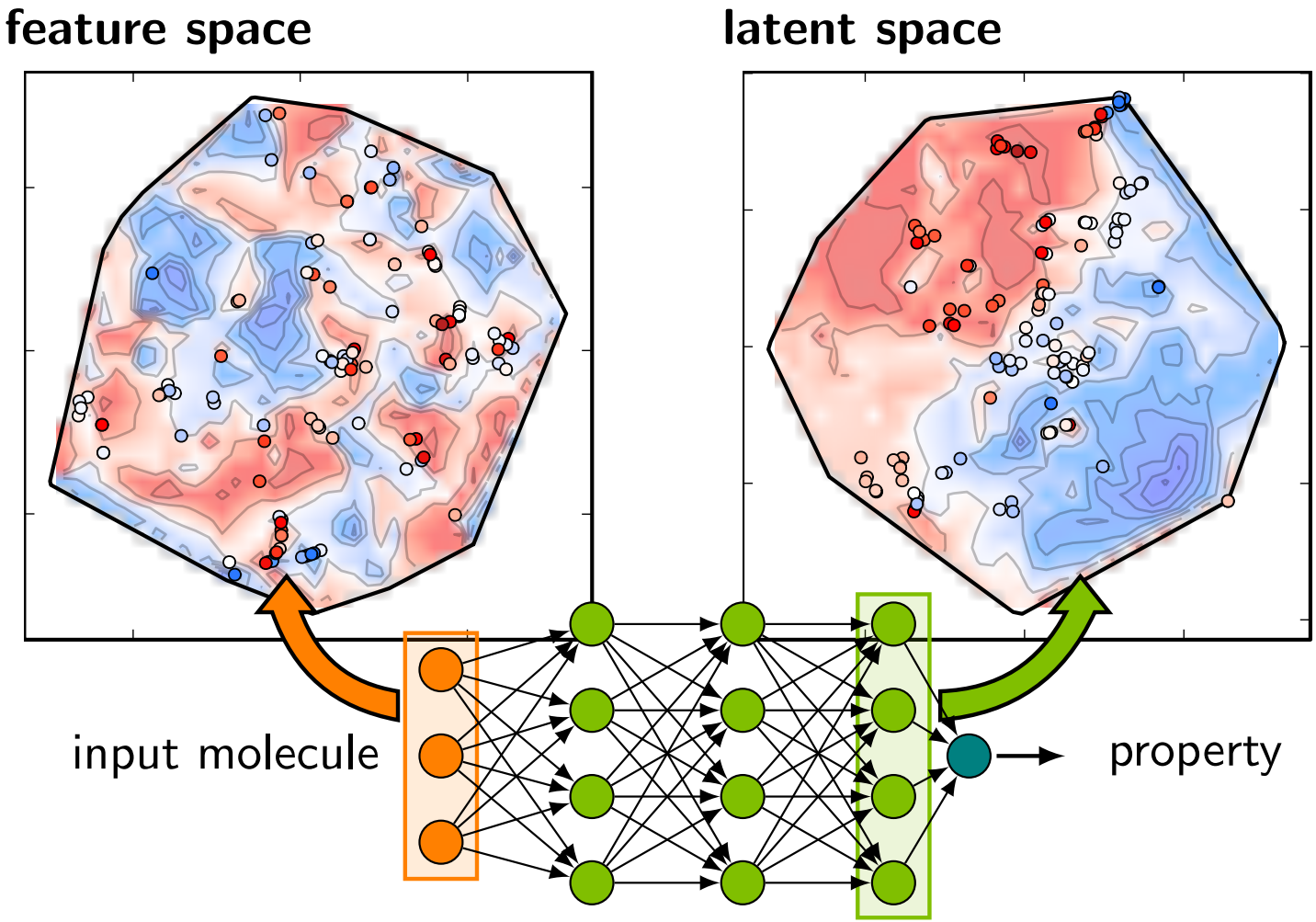
How ANNs work



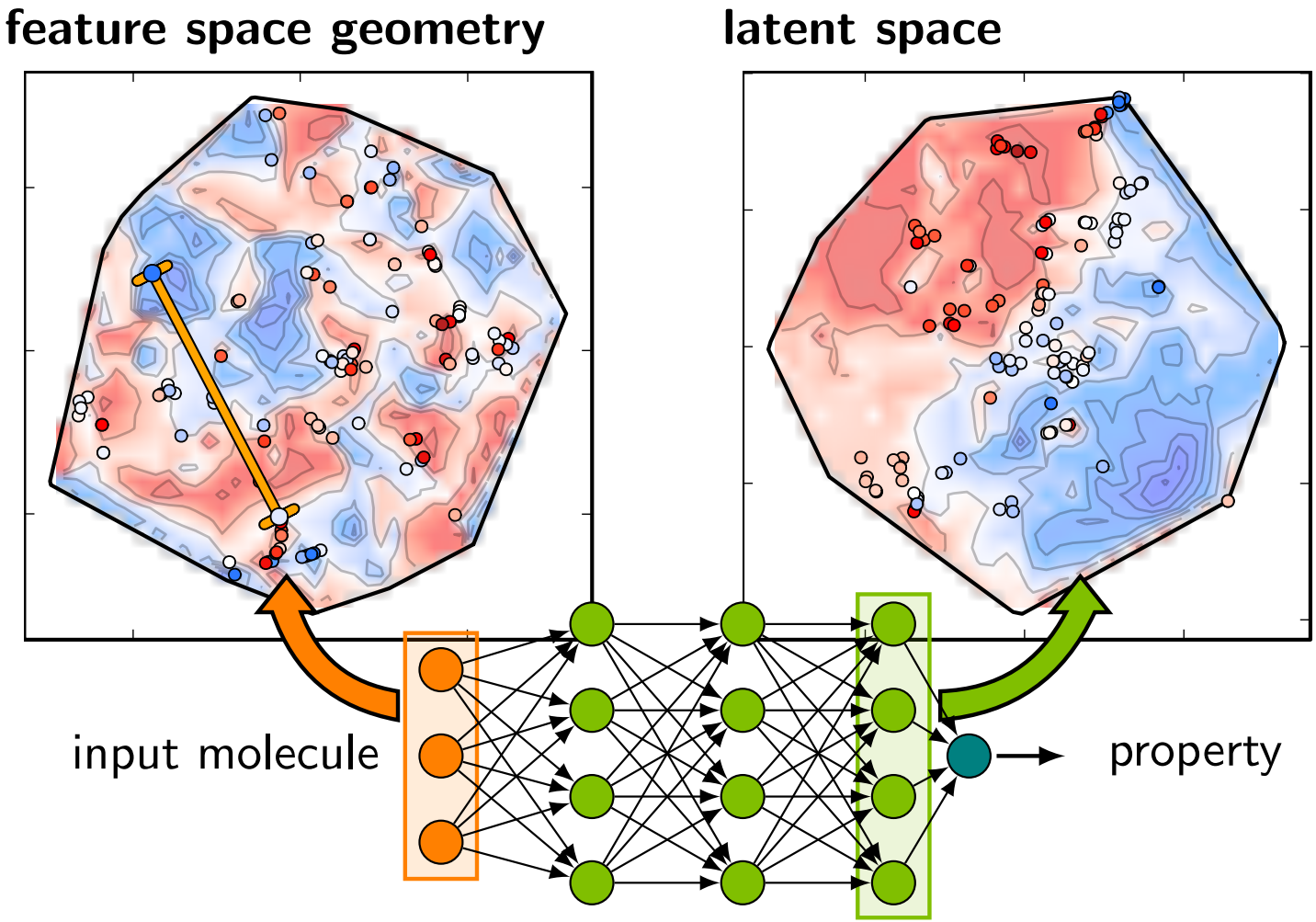
How ANNs work



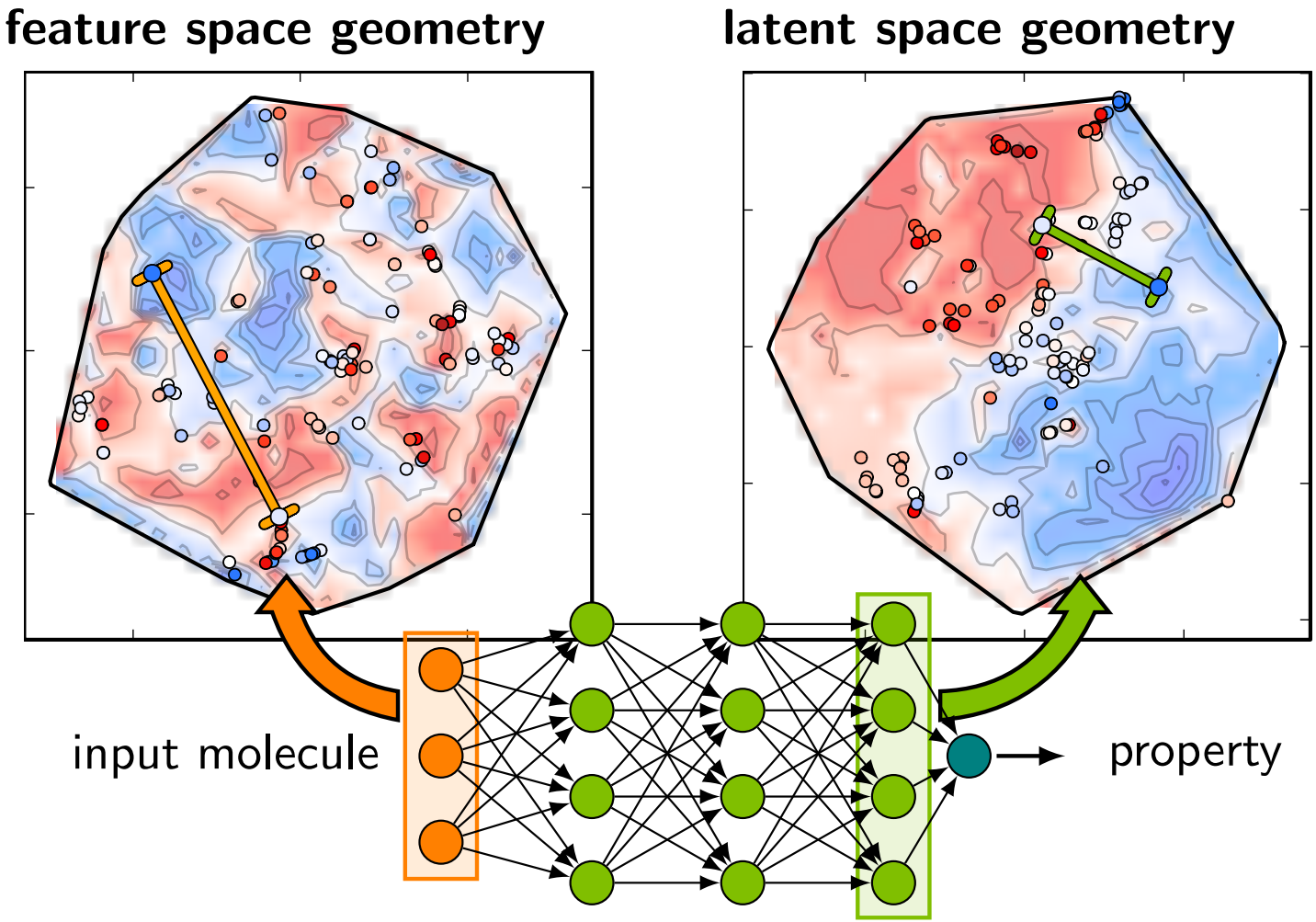
How ANNs work



How ANNs work



How ANNs work

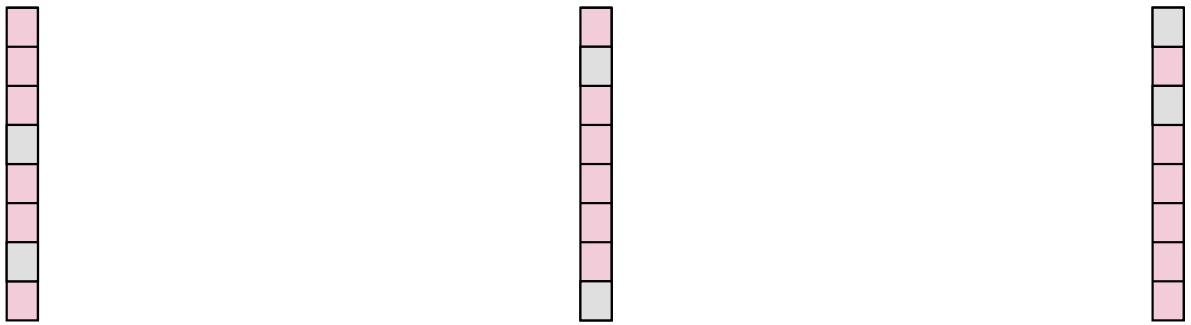


Other UQ metrics

1) Data-sampling ensembles:

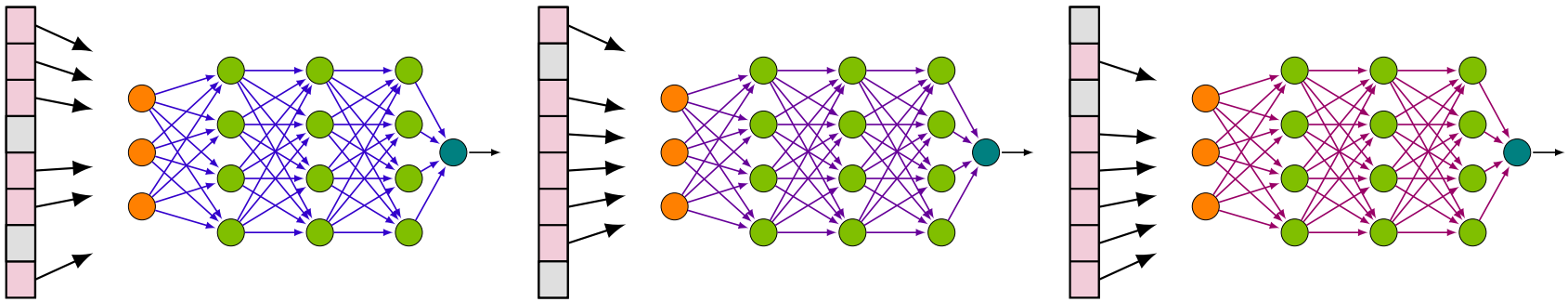
Other UQ metrics

1) Data-sampling ensembles:



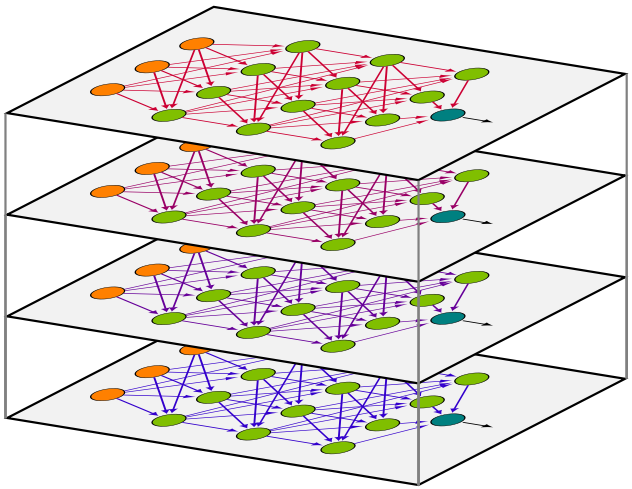
Other UQ metrics

1) Data-sampling ensembles:



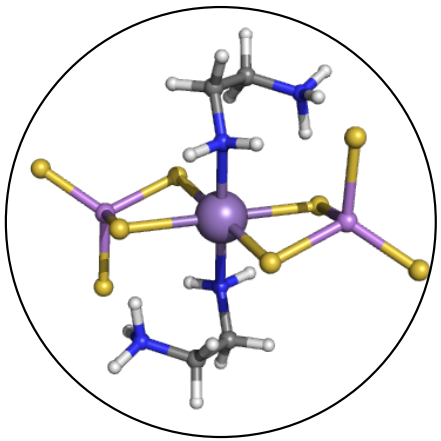
Other UQ metrics

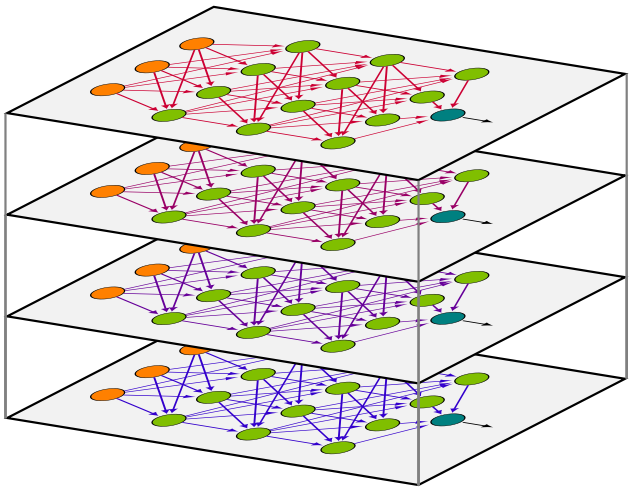
1) Data-sampling ensembles:



Other UQ metrics

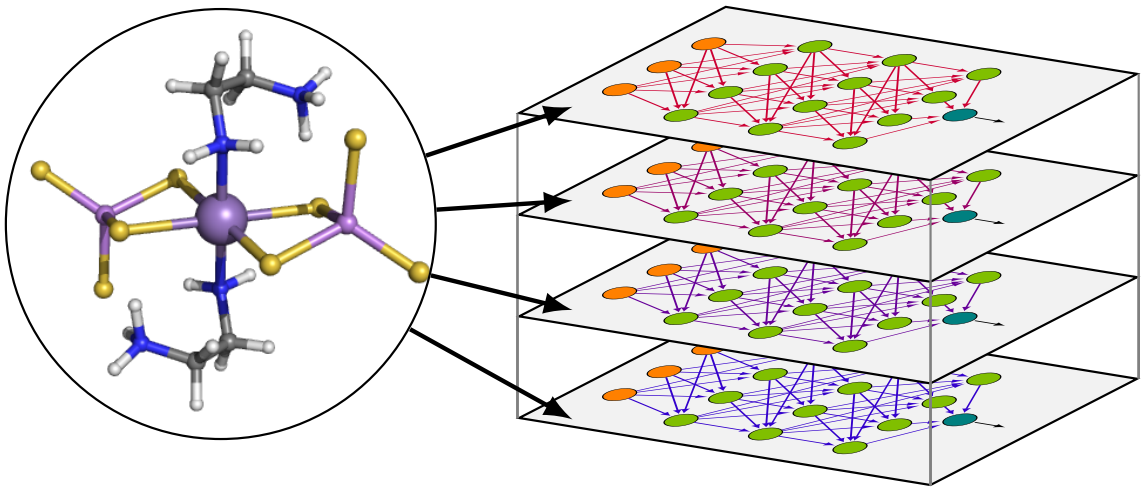
1) Data-sampling ensembles:





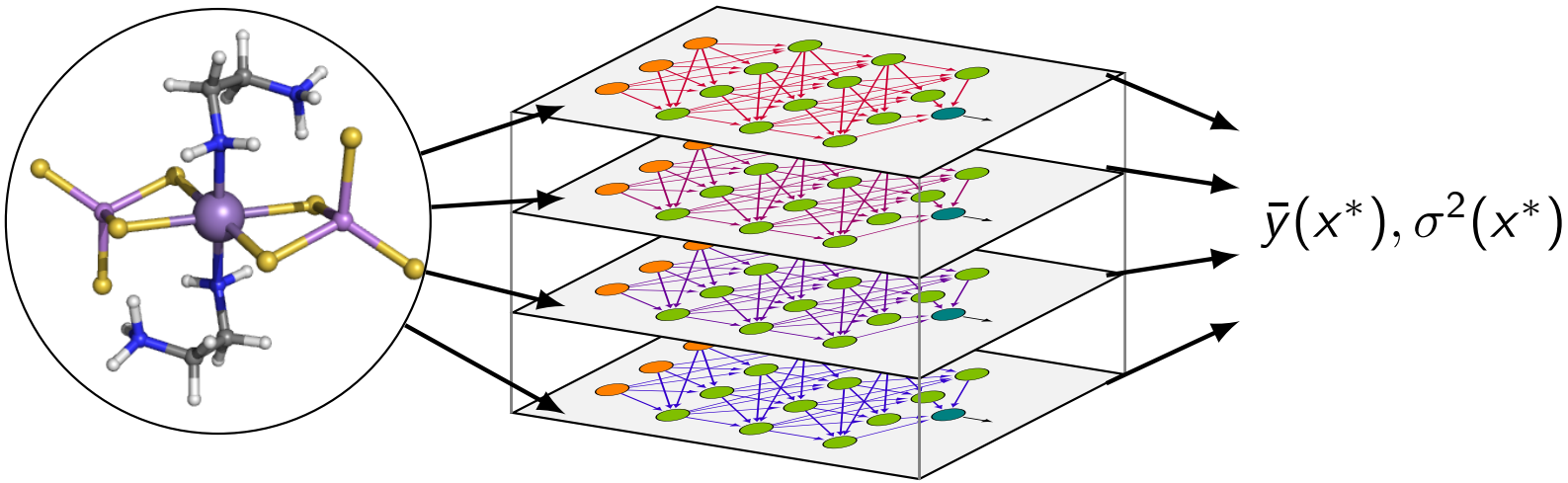
Other UQ metrics

1) Data-sampling ensembles:



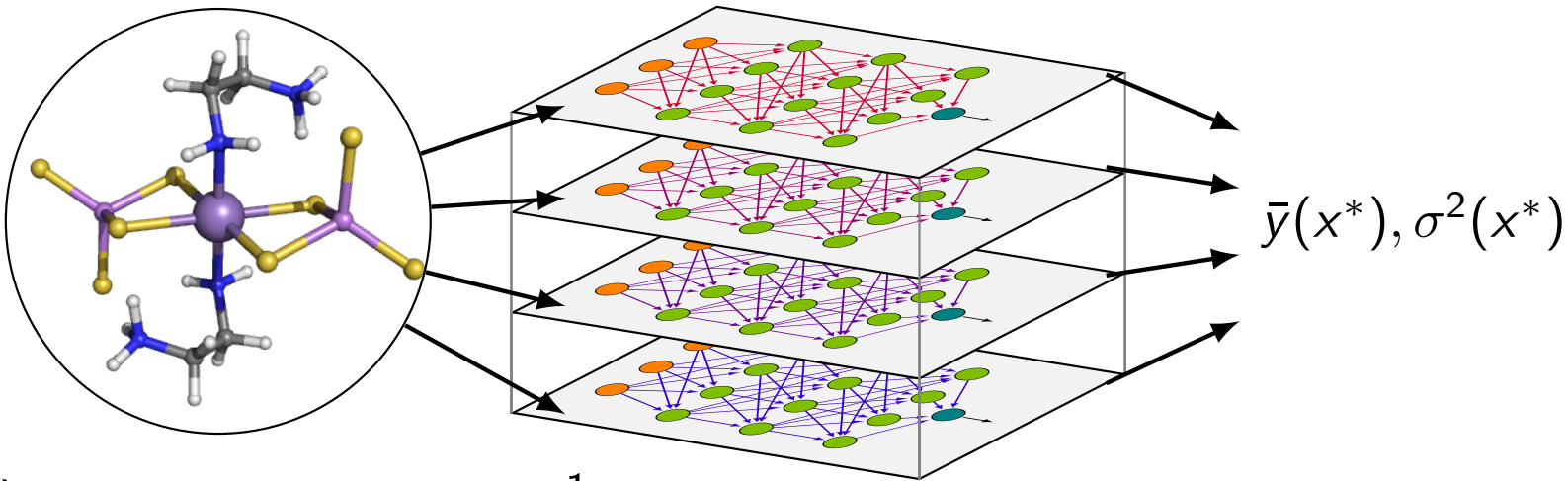
Other UQ metrics

1) Data-sampling ensembles:

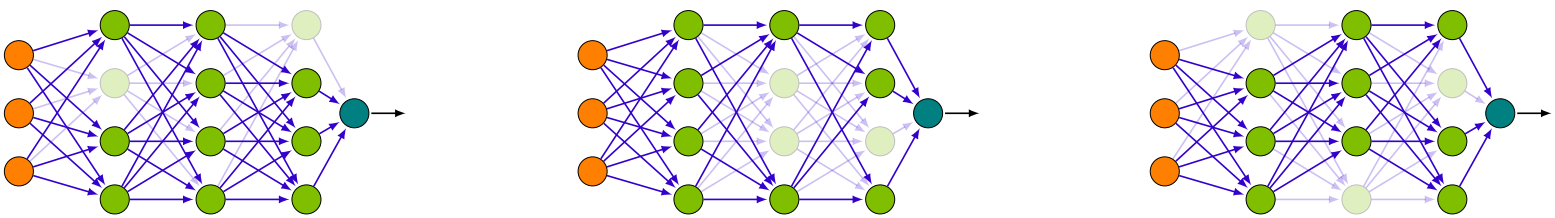


Other UQ metrics

1) Data-sampling ensembles:



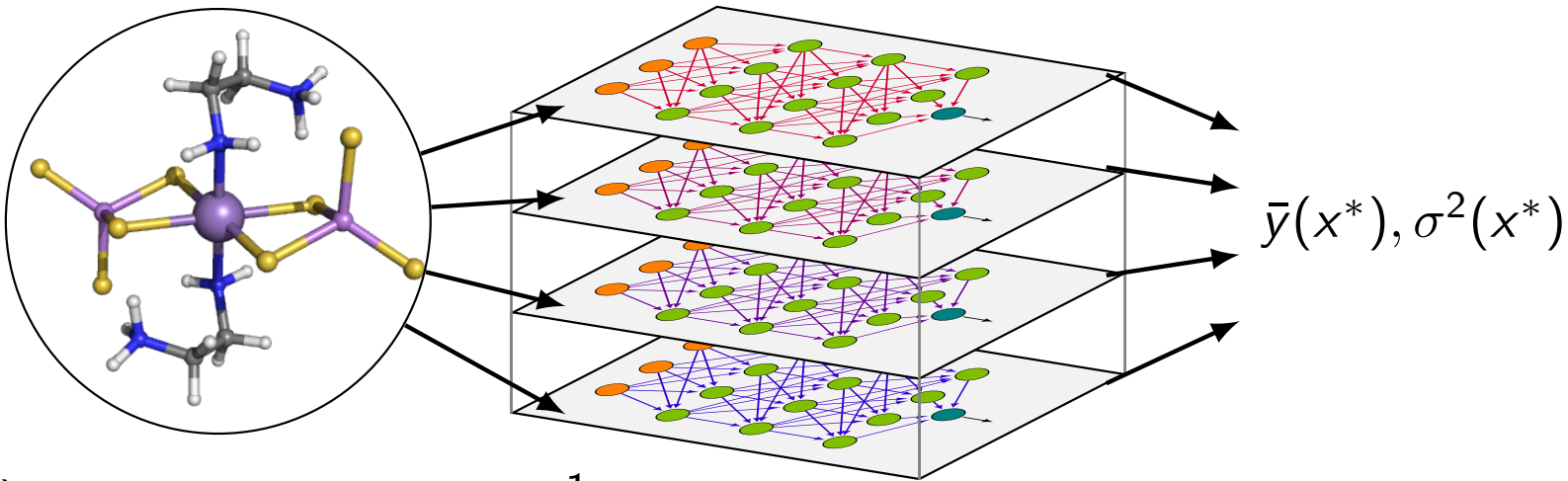
2) Monte Carlo dropout[1]:



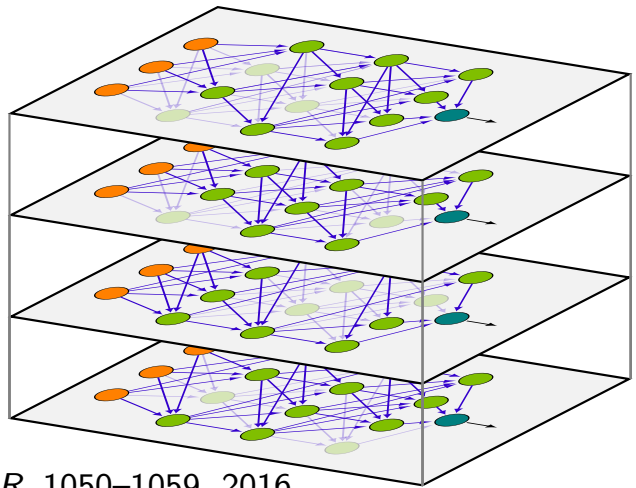
[1]:Gal, Y. and Ghahramani, Z., *ICMLR*, 1050–1059, 2016.

Other UQ metrics

1) Data-sampling ensembles:



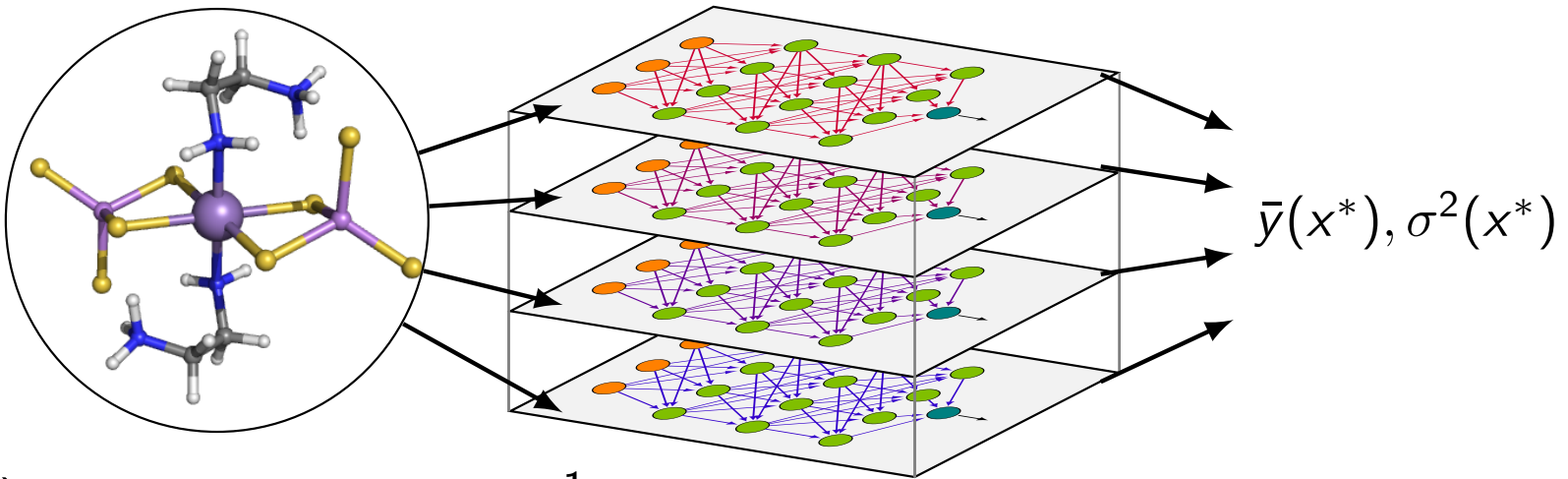
2) Monte Carlo dropout[1]:



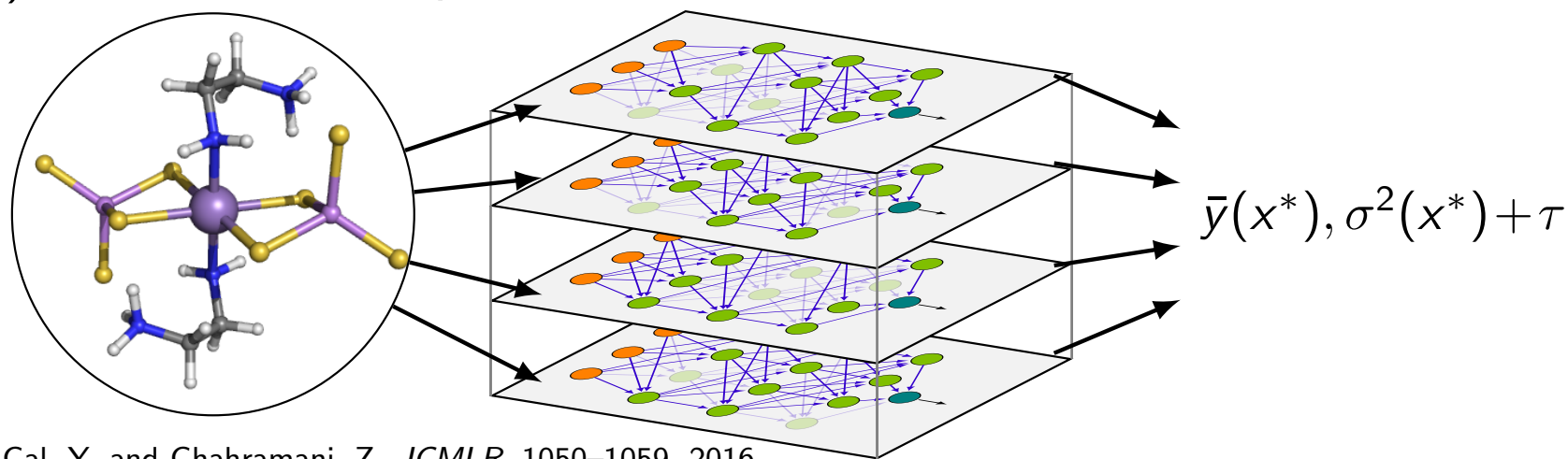
[1]:Gal, Y. and Ghahramani, Z., *ICML*, 1050–1059, 2016.

Other UQ metrics

1) Data-sampling ensembles:



2) Monte Carlo dropout[1]:



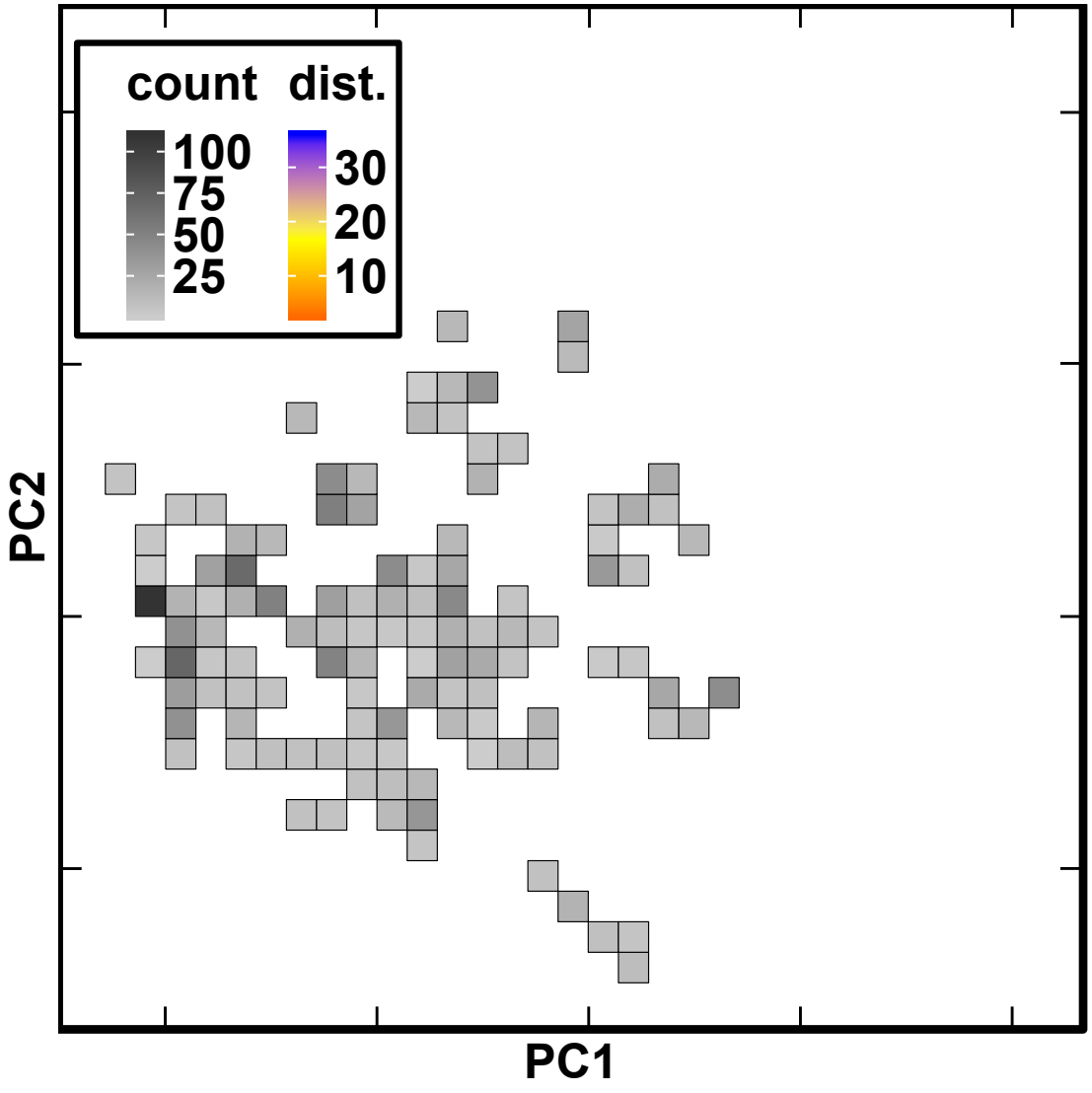
[1]:Gal, Y. and Ghahramani, Z., *ICMLR*, 1050–1059, 2016.

A challenging test case: CSD II

'Out-of-distribution' test: spin-splitting energies of 116 structures from the CSD, from training-like to very different.

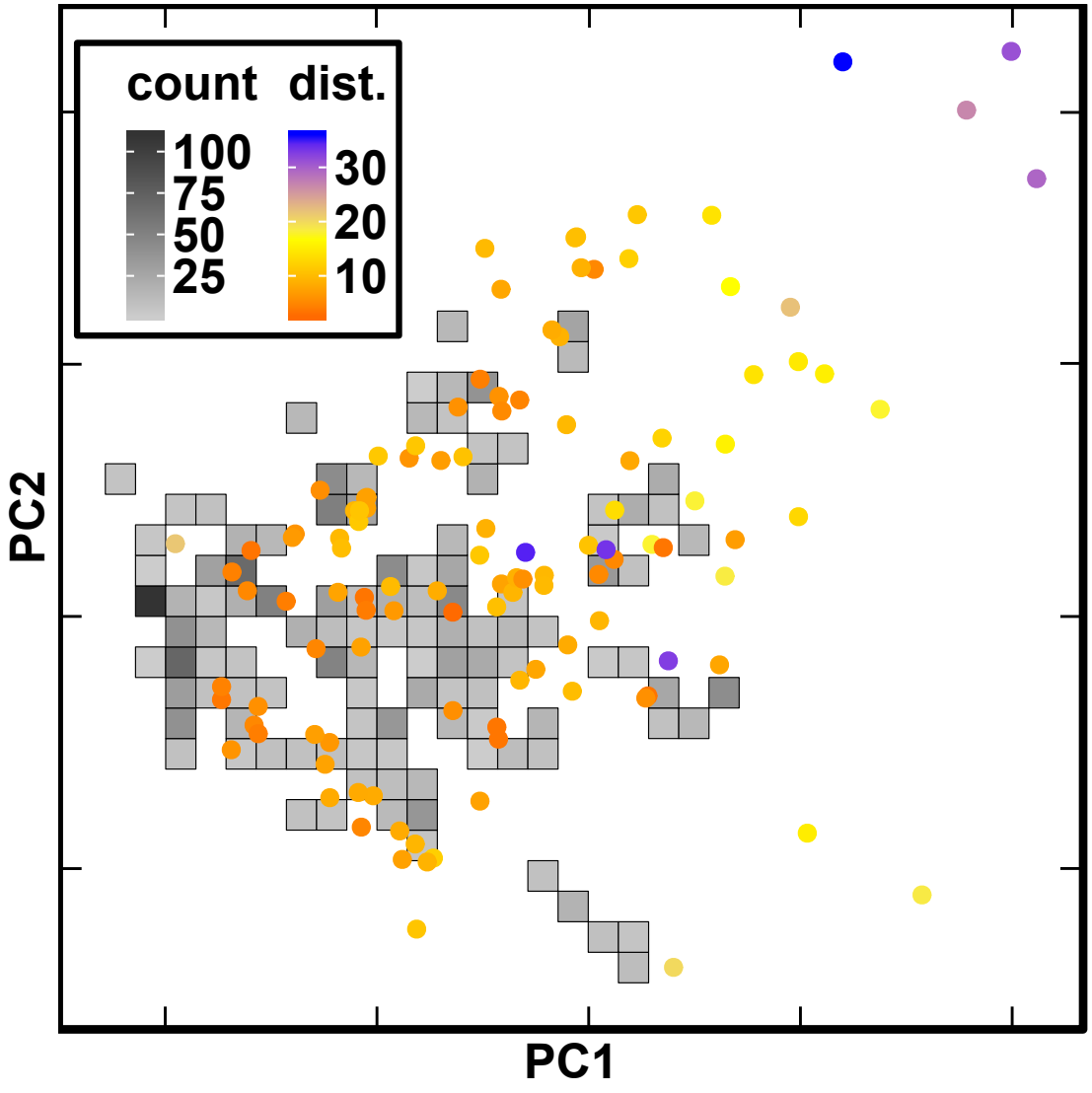
A challenging test case: CSD II

'Out-of-distribution' test: spin-splitting energies of 116 structures from the CSD, from training-like to very different.



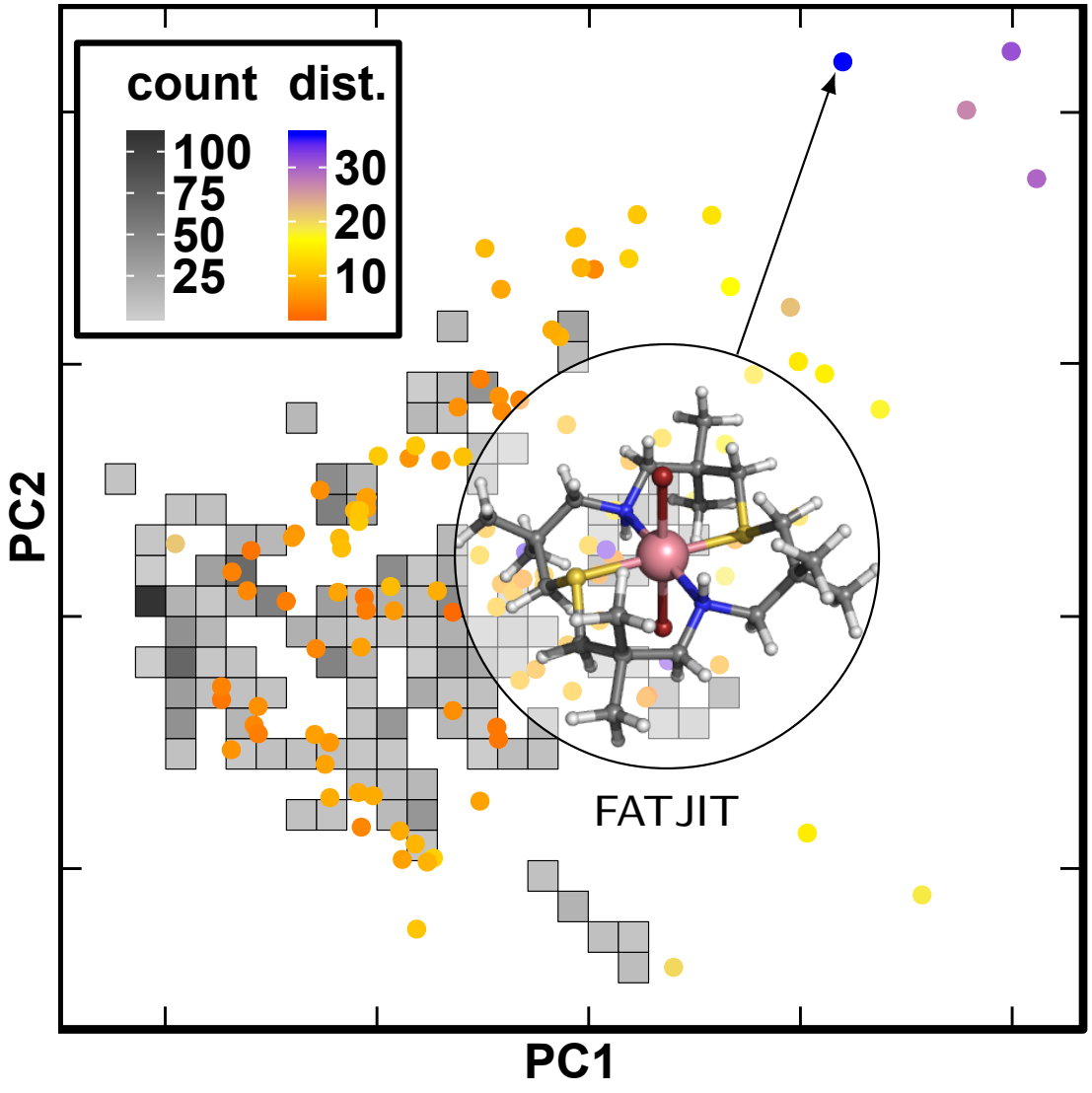
A challenging test case: CSD II

'Out-of-distribution' test: spin-splitting energies of 116 structures from the CSD, from training-like to very different.



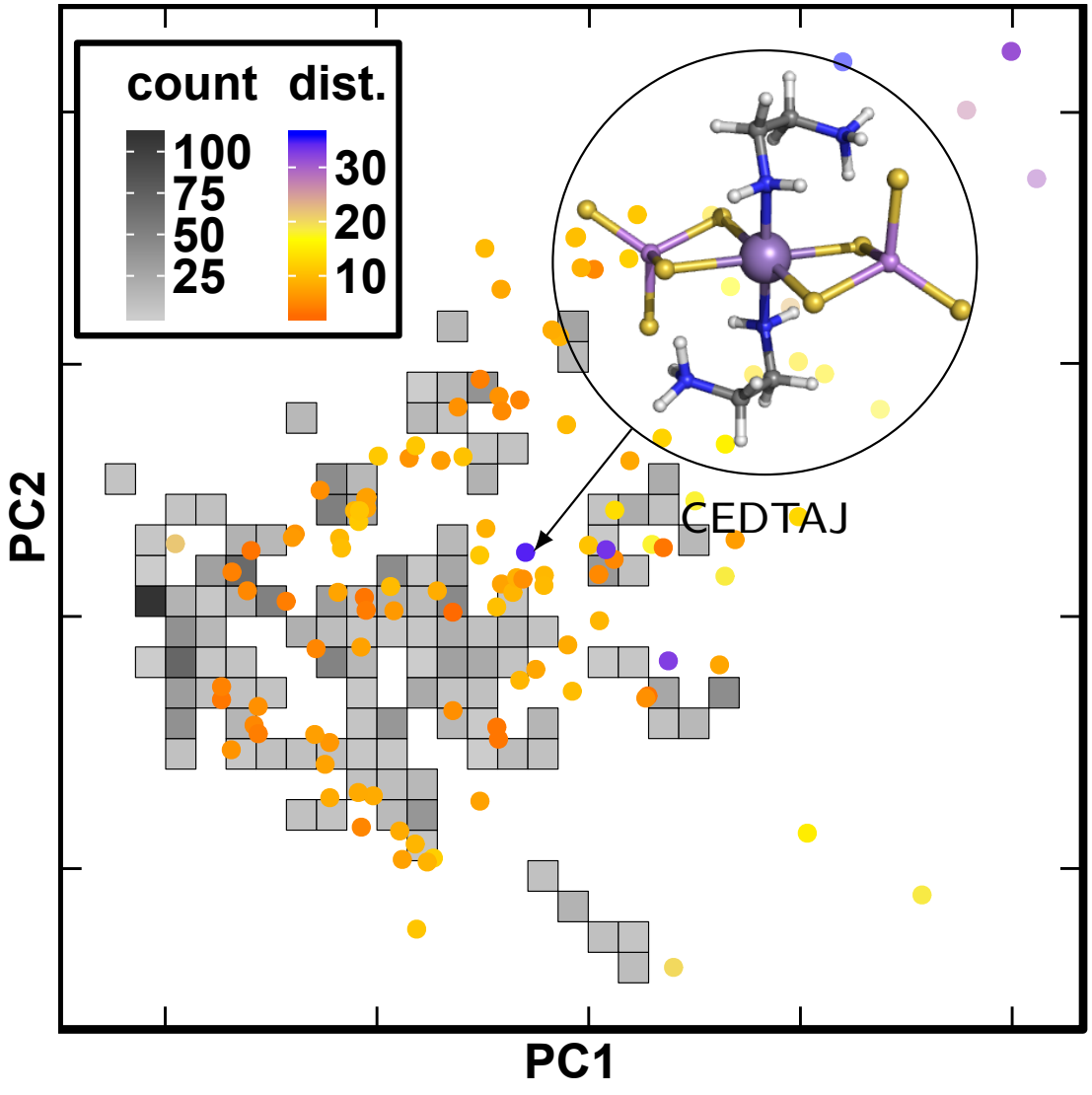
A challenging test case: CSD II

'Out-of-distribution' test: spin-splitting energies of 116 structures from the CSD, from training-like to very different.



A challenging test case: CSD II

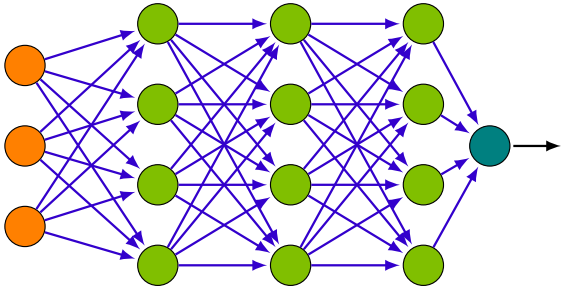
'Out-of-distribution' test: spin-splitting energies of 116 structures from the CSD, from training-like to very different.

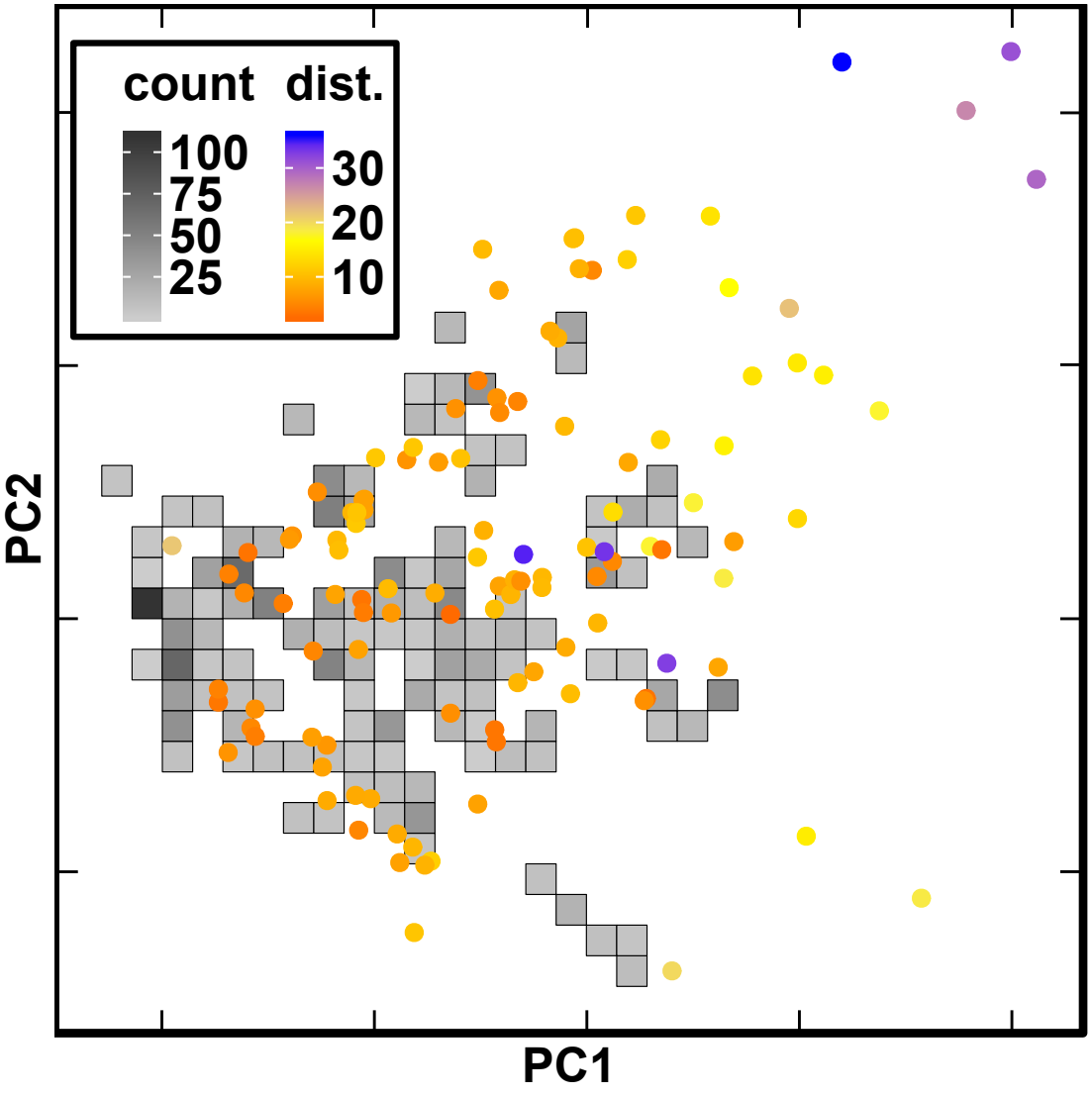


A challenging test case: CSD II

'Out-of-distribution' test: spin-splitting energies of 116 structures from the CSD, from training-like to very different.

Train 3-layer fully connected ANN on 1900 DFT results on simple ligands:

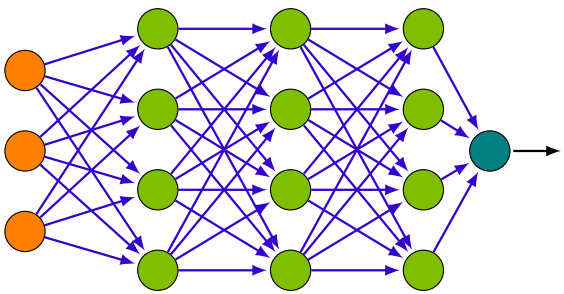


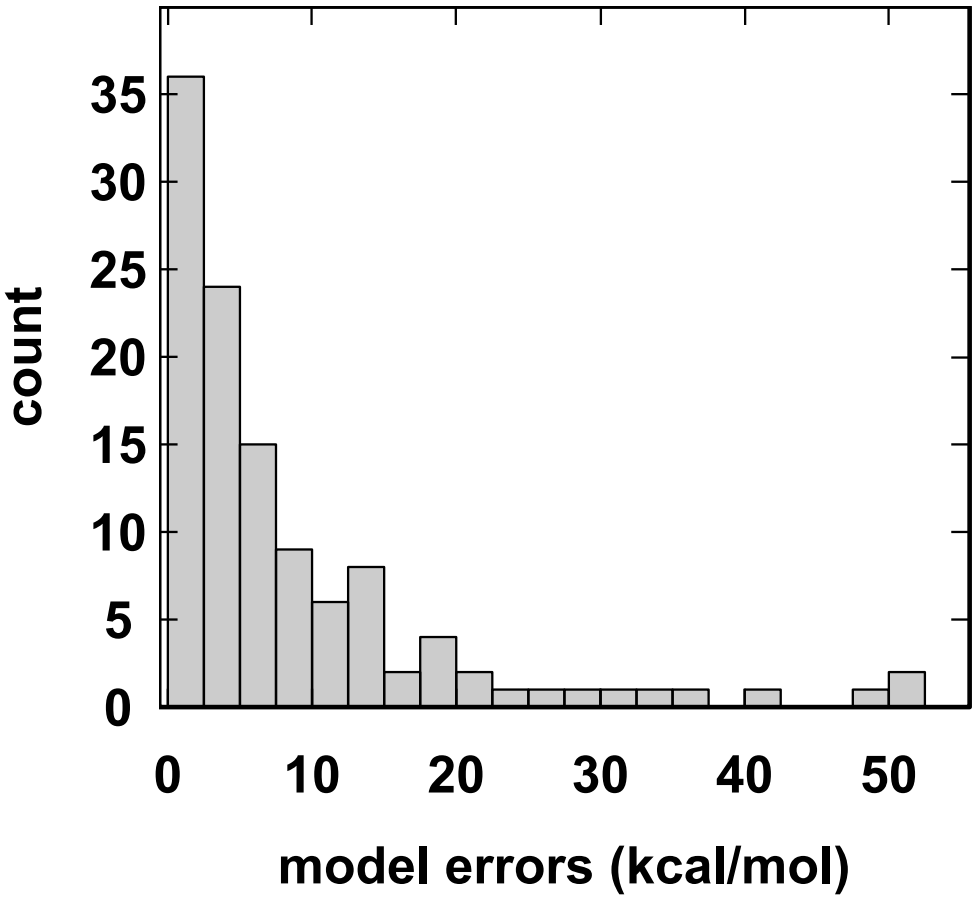


A challenging test case: CSD II

'Out-of-distribution' test: spin-splitting energies of 116 structures from the CSD, from training-like to very different.

Train 3-layer fully connected ANN on 1900 DFT results on simple ligands:





Which distance to measure

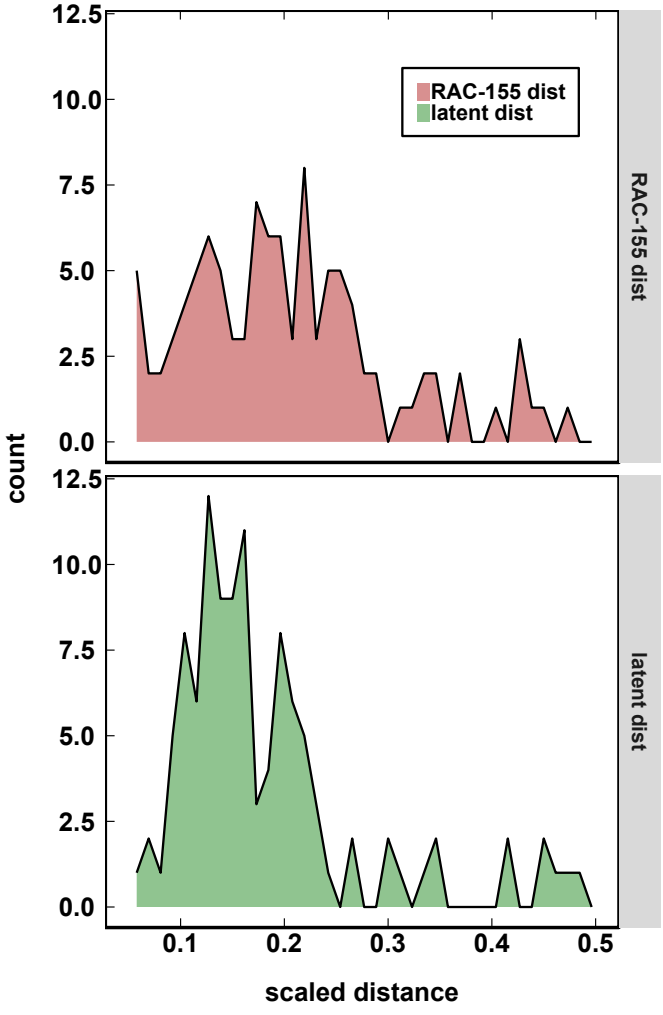
Key questions:

1. Are latent distances meaningful?
2. How can we compare distances?
3. How should we measure distance?

Which distance to measure

Key questions:

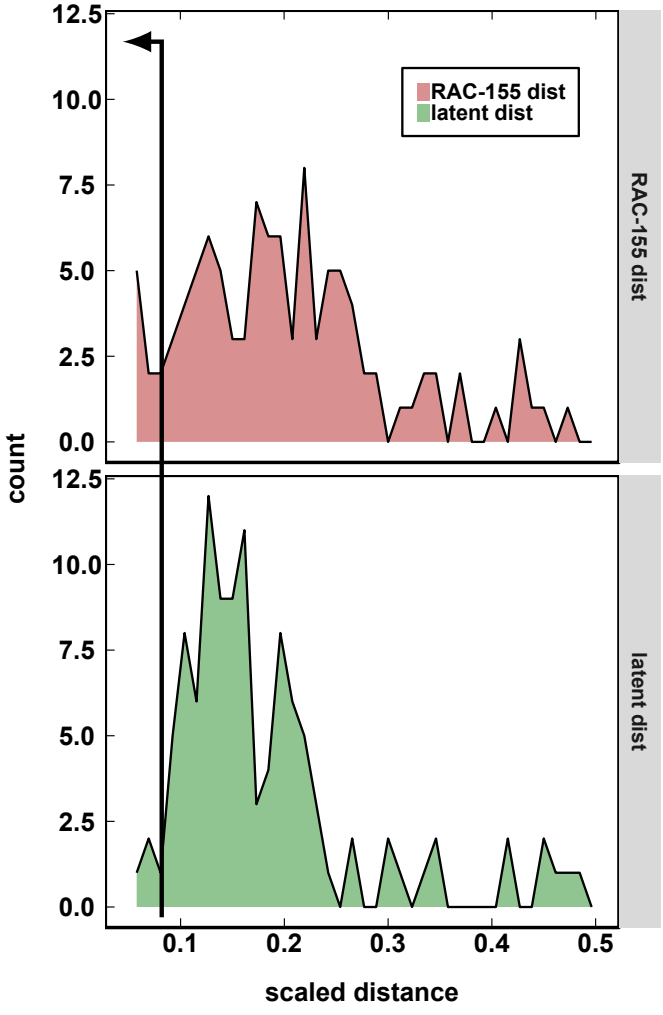
1. Are latent distances meaningful?
2. How can we compare distances?
3. How should we measure distance?



Which distance to measure

Key questions:

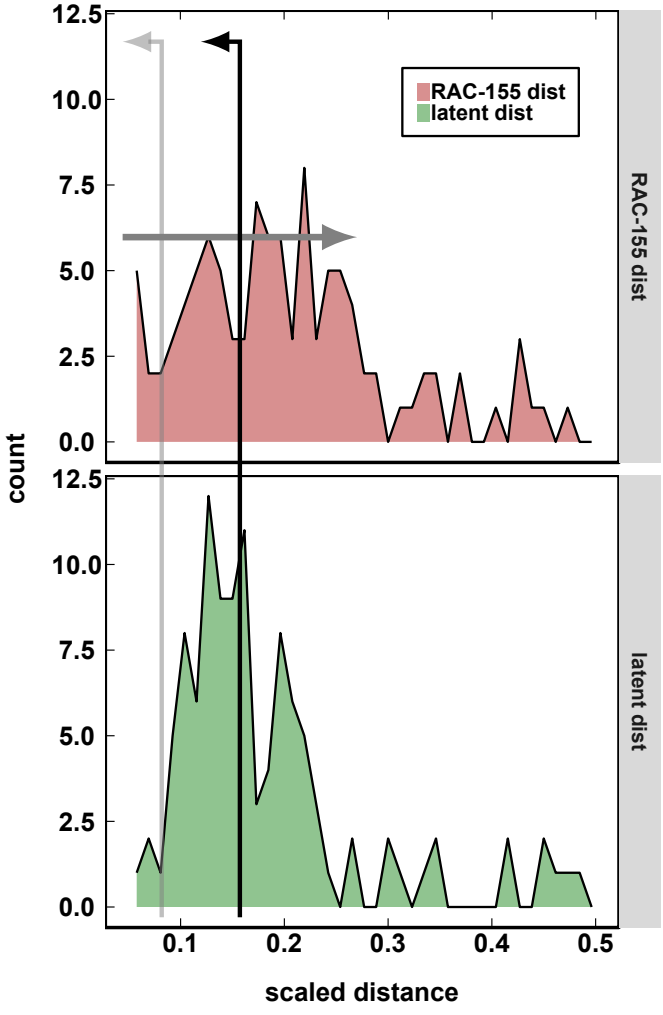
1. Are latent distances meaningful?
2. How can we compare distances?
3. How should we measure distance?



Which distance to measure

Key questions:

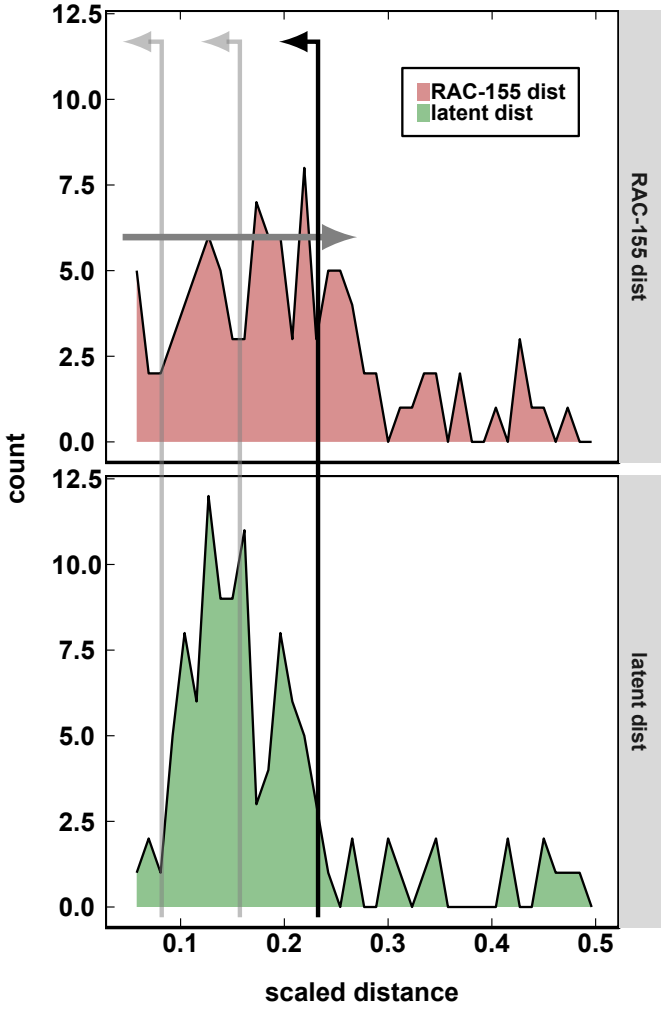
1. Are latent distances meaningful?
2. How can we compare distances?
3. How should we measure distance?



Which distance to measure

Key questions:

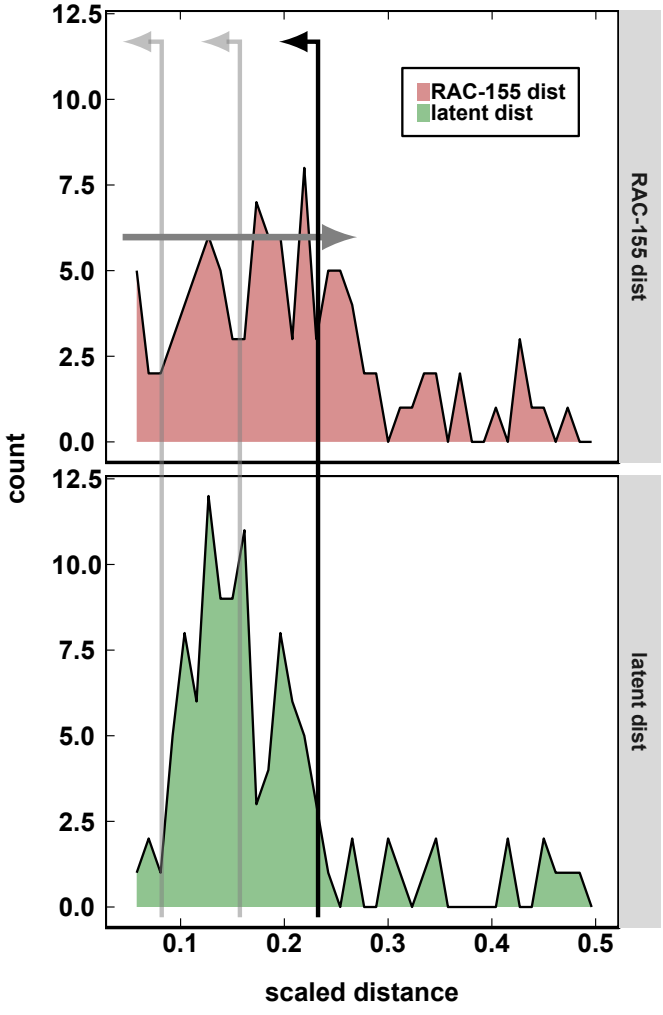
1. Are latent distances meaningful?
2. How can we compare distances?
3. How should we measure distance?



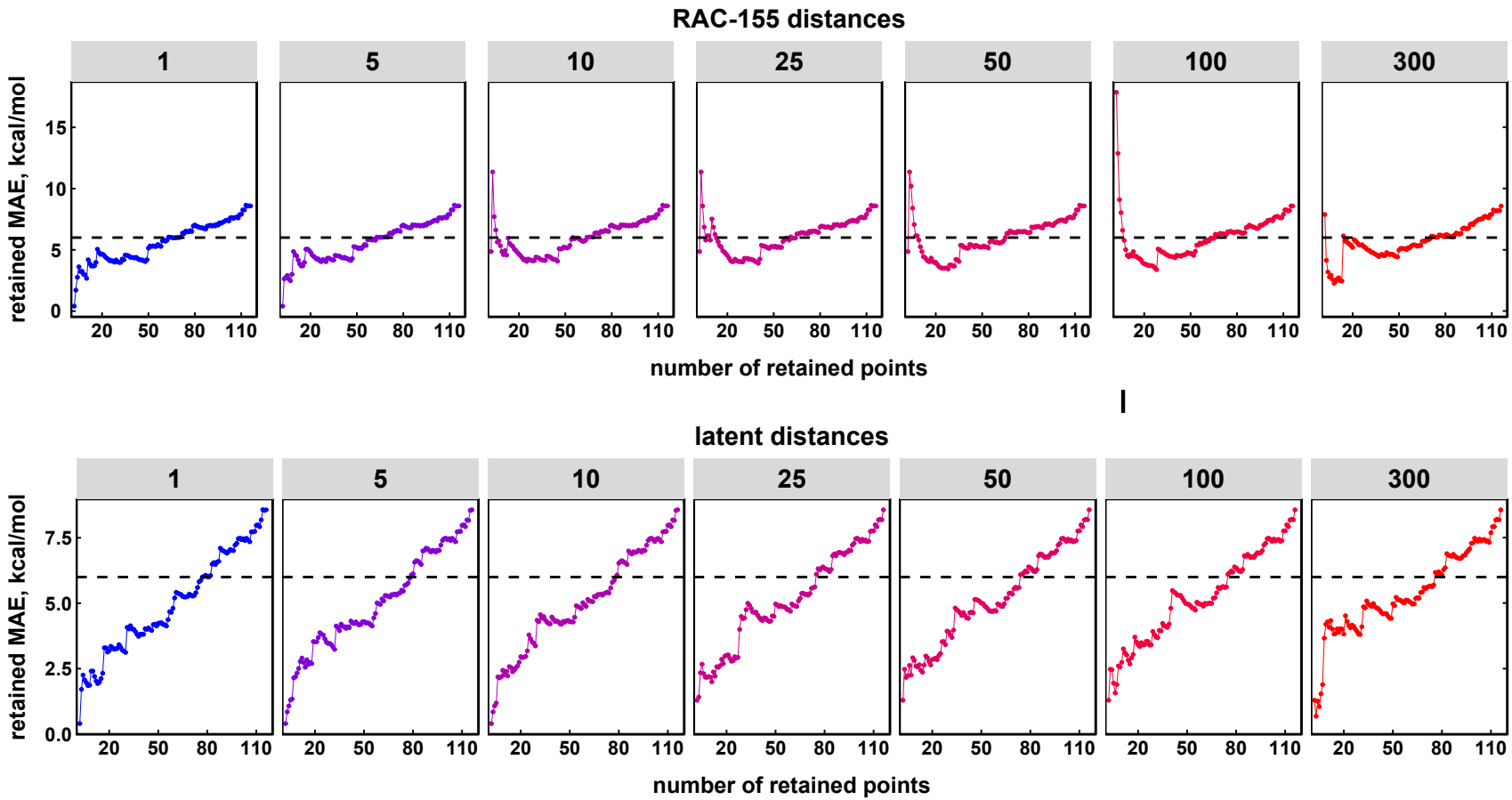
Which distance to measure

Key questions:

1. Are latent distances meaningful?
2. How can we compare distances?
3. How should we measure distance?

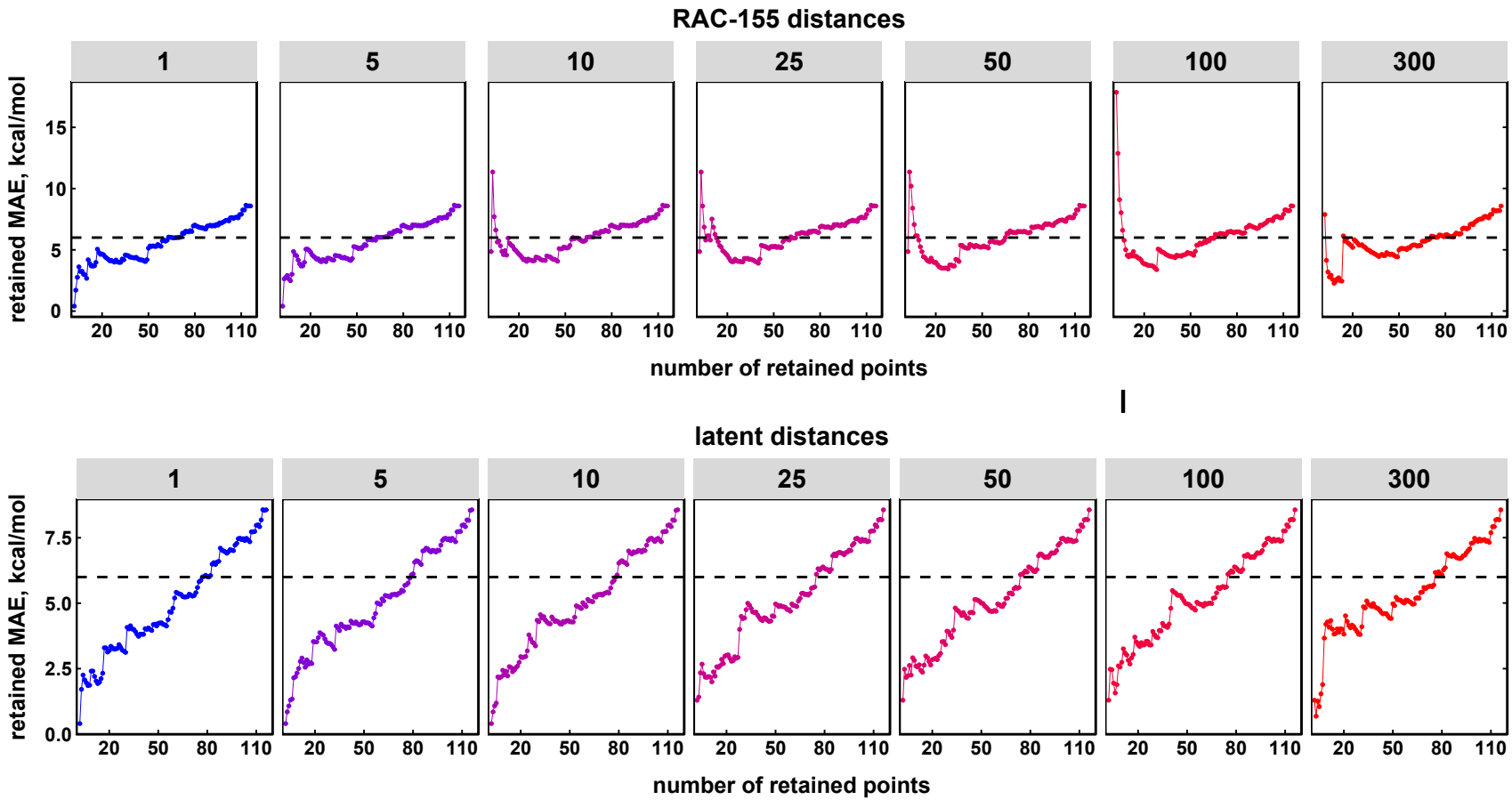


Which distance to measure



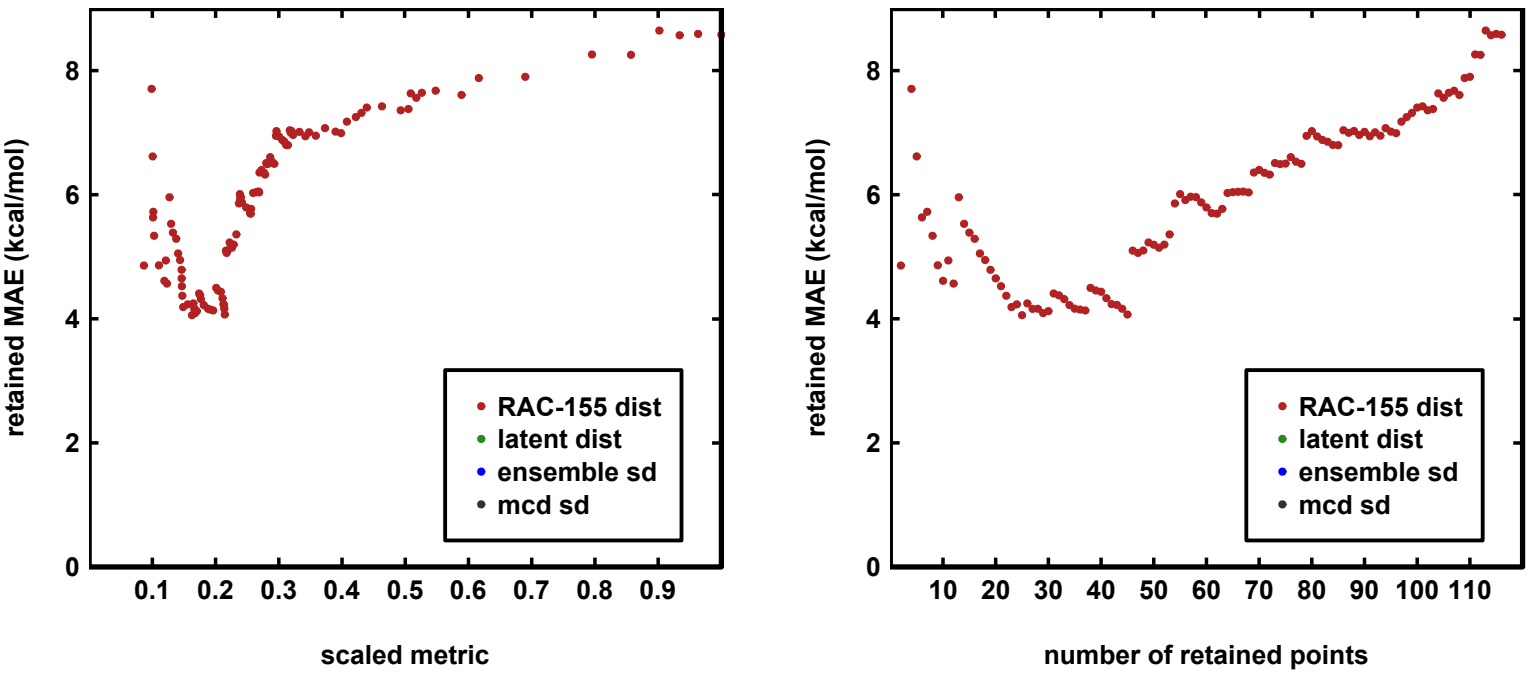
Which distance to measure

We choose to average over 10 neighbors, insensitive to this choice



Latent distances give stable error control

Make a comparison of discriminative power[1]:

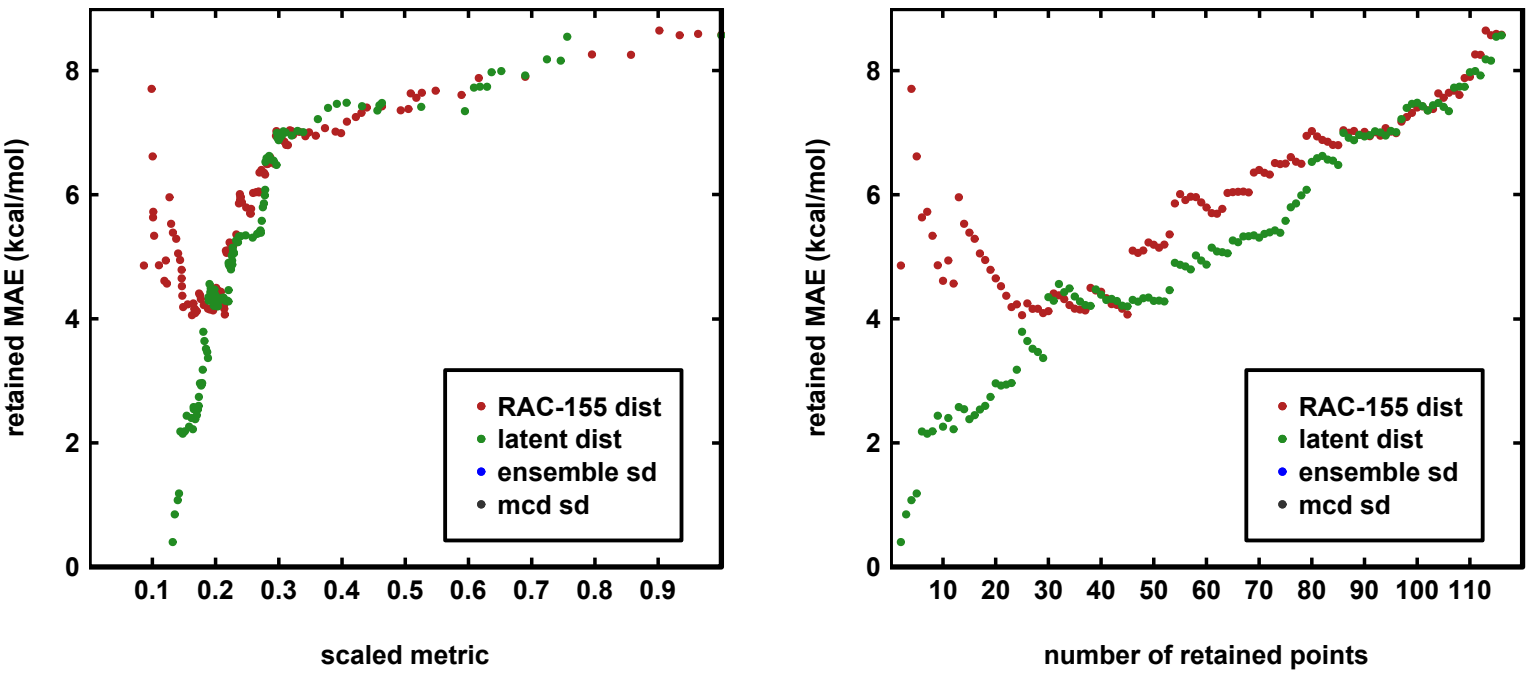


[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

Latent distances give stable error control

Make a comparison of discriminative power[1]:

latent distances are superior to feature space distances

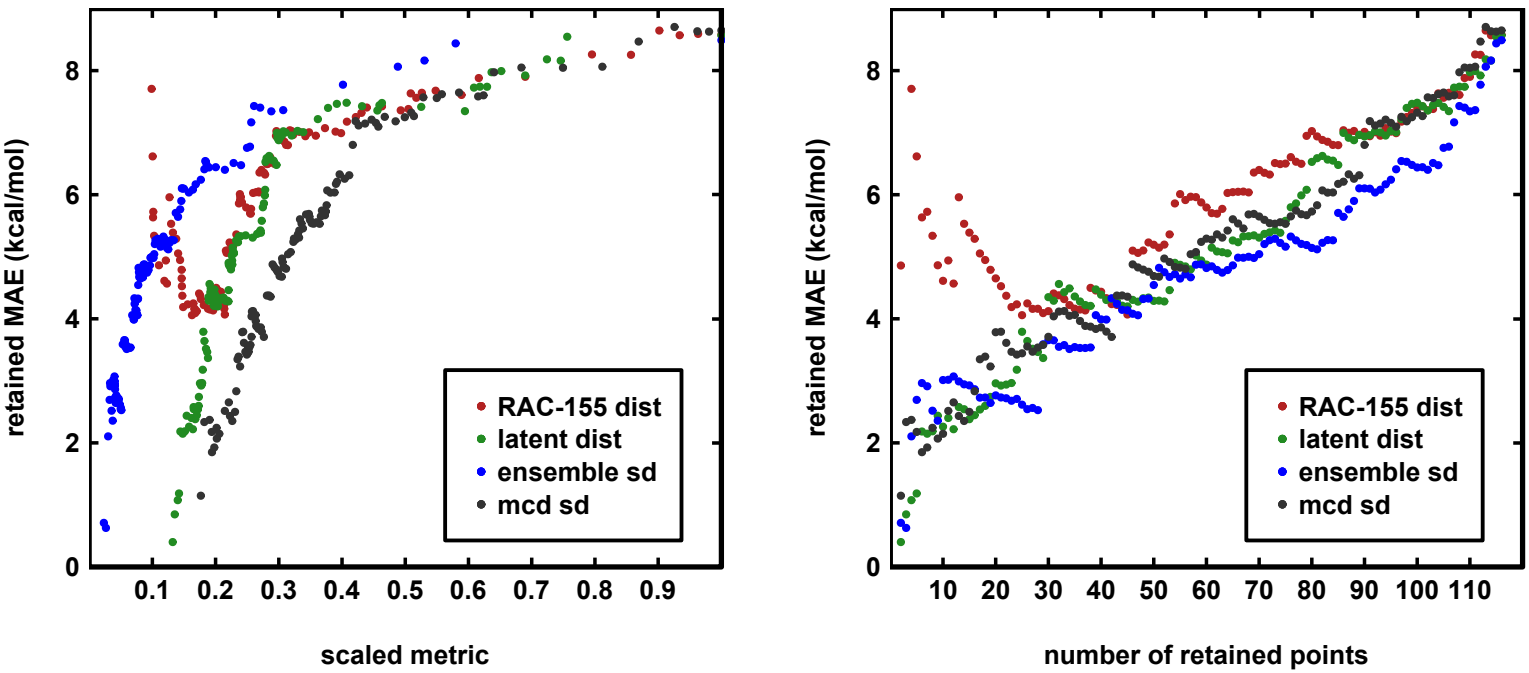


[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

Latent distances give stable error control

Make a comparison of discriminative power[1]:

latent distances are superior to feature space distances



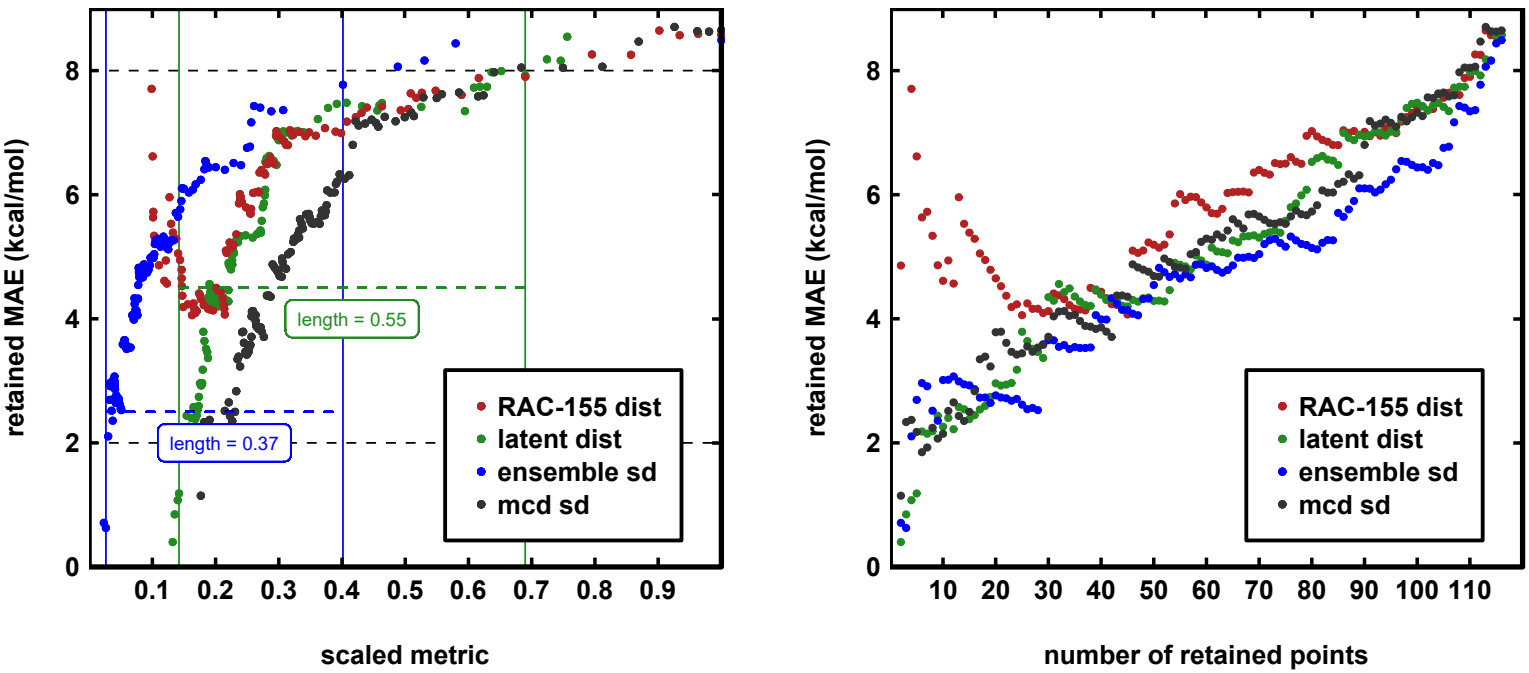
comparable with ensembles and mc dropout

[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

Latent distances give stable error control

Make a comparison of discriminative power[1]:

latent distances are superior to feature space distances



comparable with ensembles and mc dropout

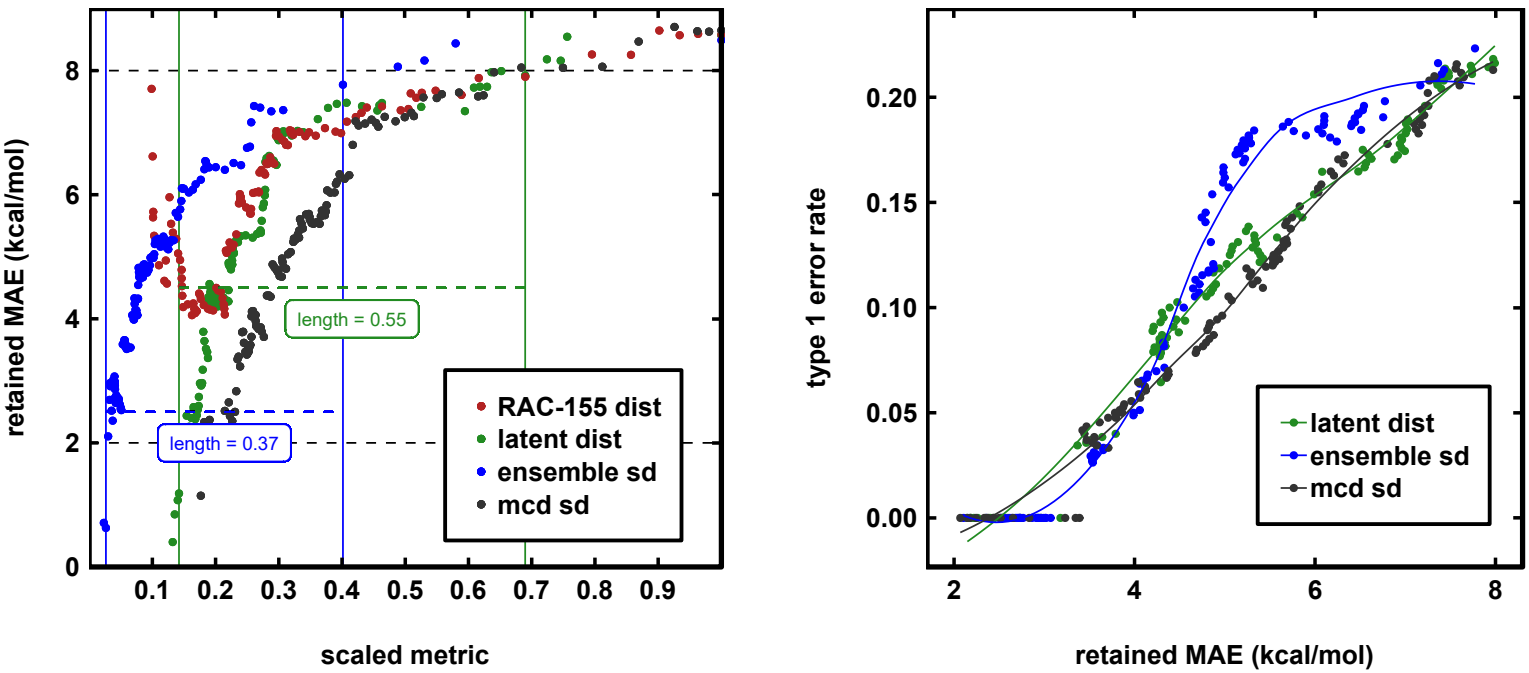
stability is important

[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

Latent distances give stable error control

Make a comparison of discriminative power[1]:

latent distances are superior to feature space distances



comparable with ensembles and mc dropout

[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

Converting distances to errors

Propose a simple conditionally-Gaussian model for predicting error distribution with latent distance, d:

$$\varepsilon(d) \sim \mathcal{N}\left(0, \sigma_1^2 + d\sigma_2^2\right)$$

Converting distances to errors

Propose a simple conditionally-Gaussian model for predicting error distribution with latent distance, d:

$$\varepsilon(d) \sim \mathcal{N}\left(0, \sigma_1^2 + d\sigma_2^2\right)$$

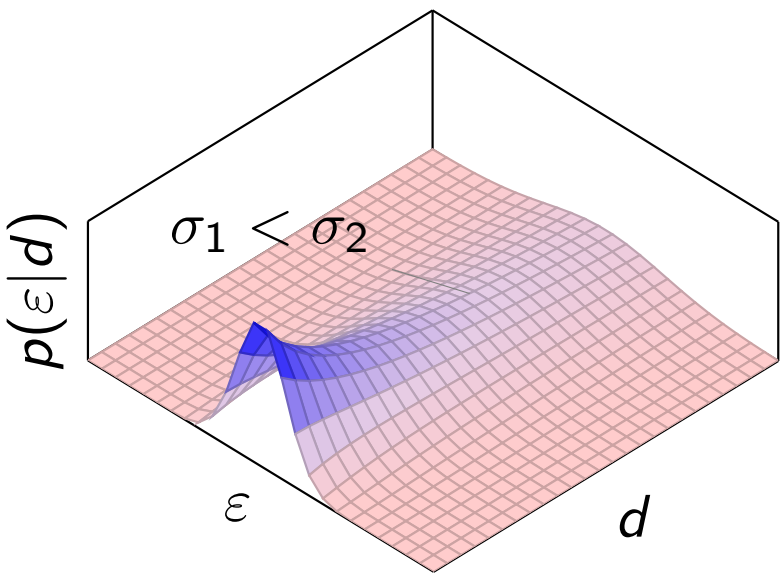
Estimate σ_1, σ_2 by max log likelihood, stably estimated even using few out-of-sample points

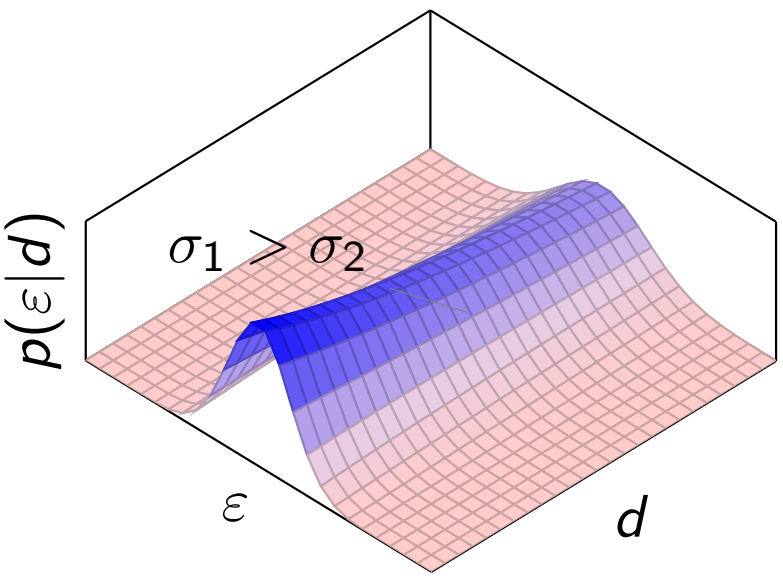
Converting distances to errors

Propose a simple conditionally-Gaussian model for predicting error distribution with latent distance, d:

$$\varepsilon(d) \sim \mathcal{N}\left(0, \sigma_1^2 + d\sigma_2^2\right)$$

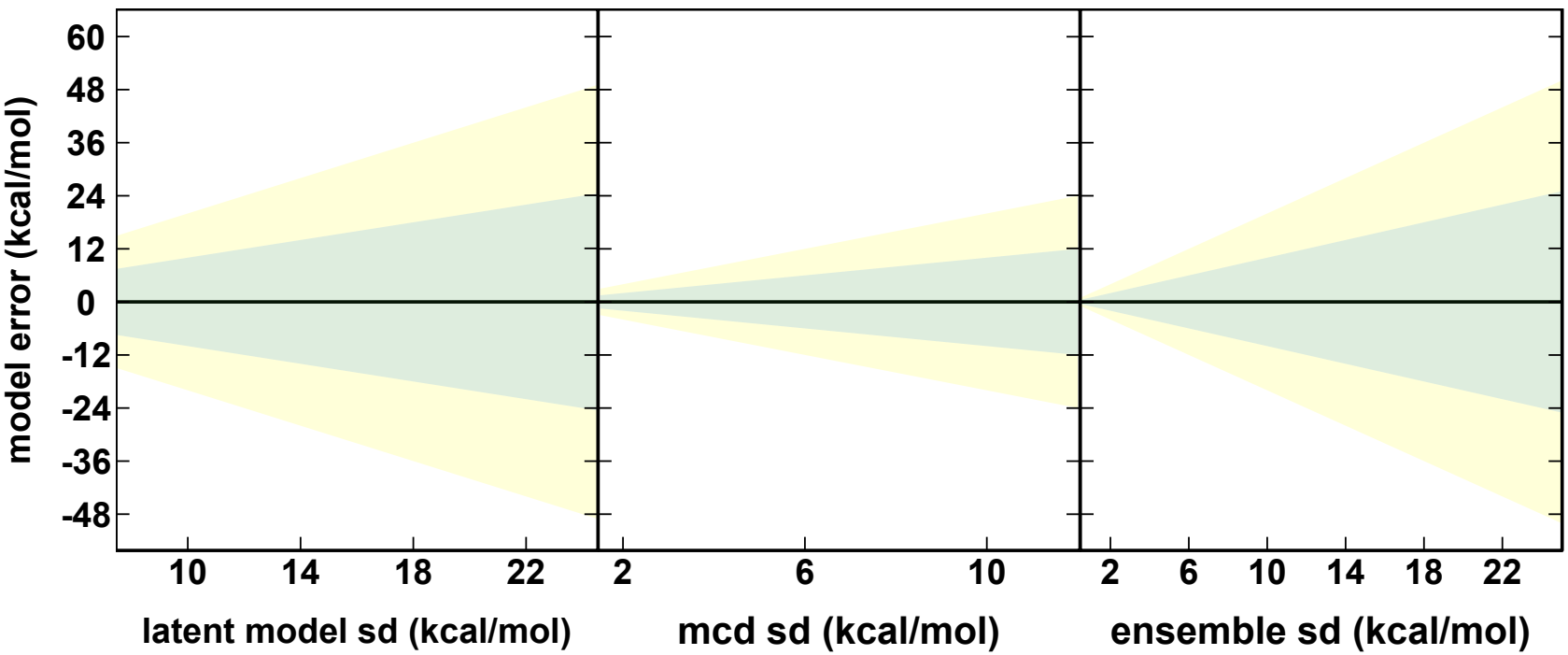
Estimate σ_1, σ_2 by max log likelihood, stably estimated even using few out-of-sample points





How do these distributions compare?

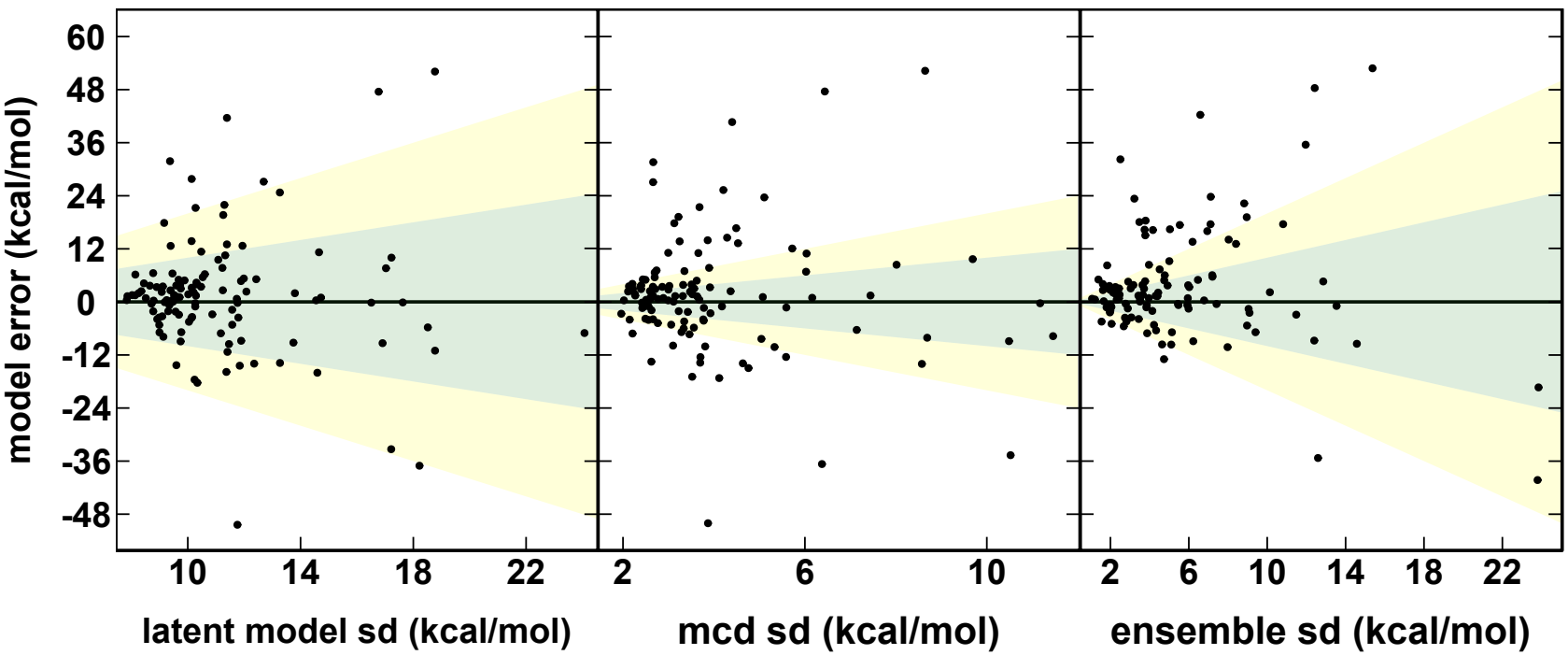
Comparison in energy units[1]:



[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

How do these distributions compare?

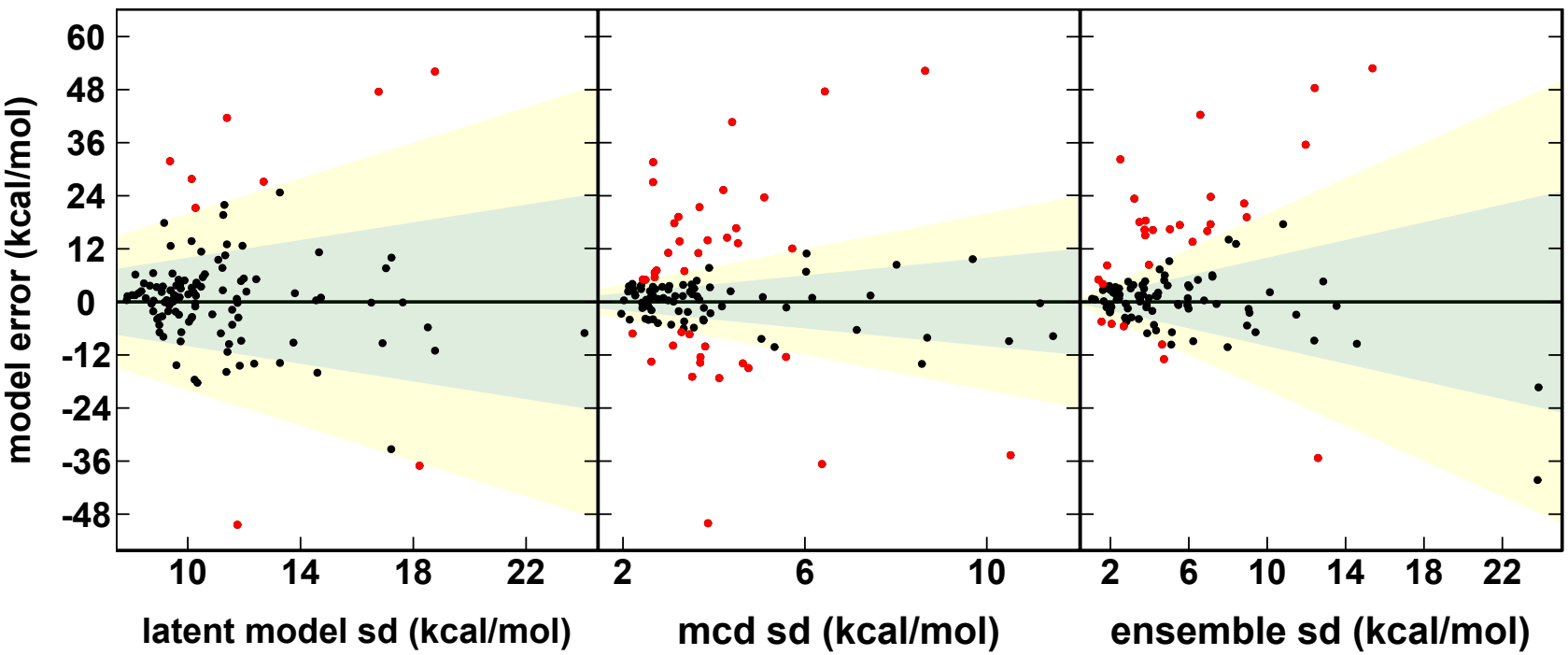
Comparison in energy units[1]:



[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

How do these distributions compare?

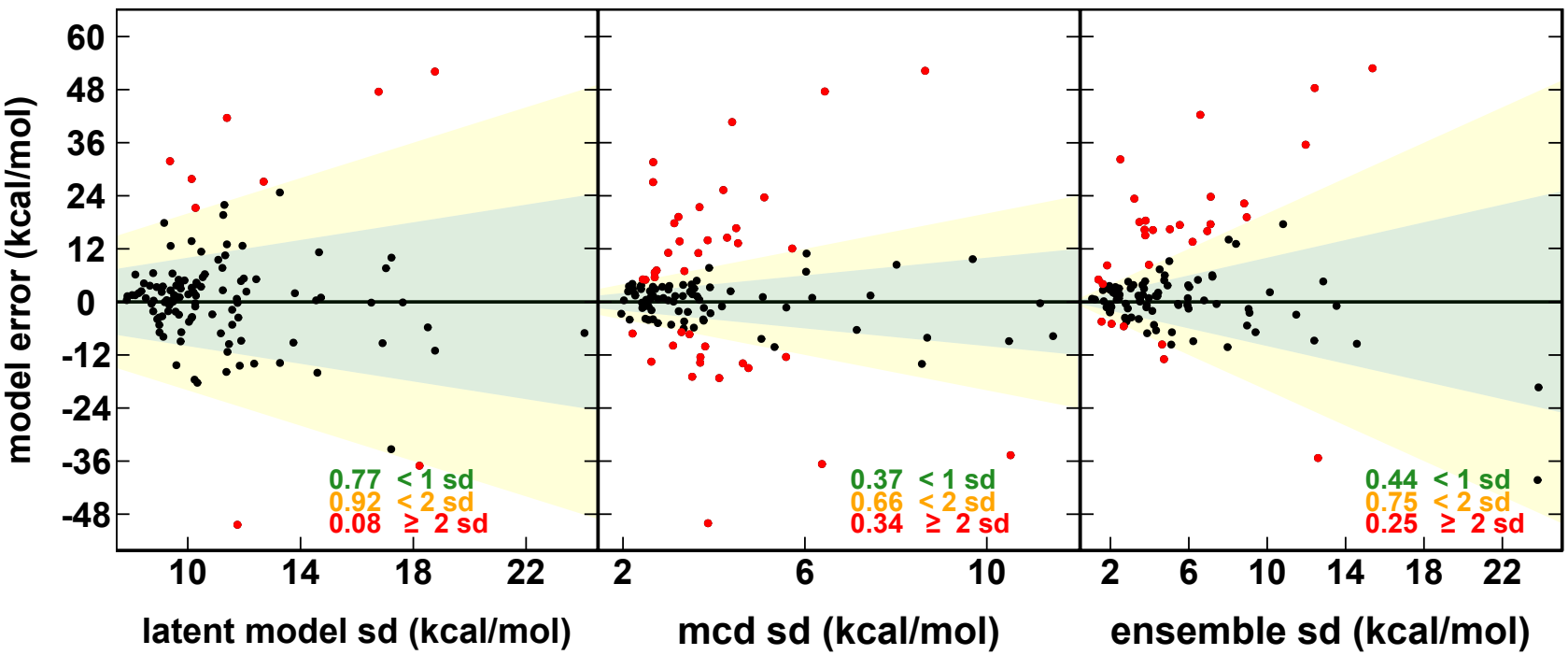
Comparison in energy units[1]:



[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

How do these distributions compare?

Comparison in energy units[1]:



[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

QM9 atomization benchmark

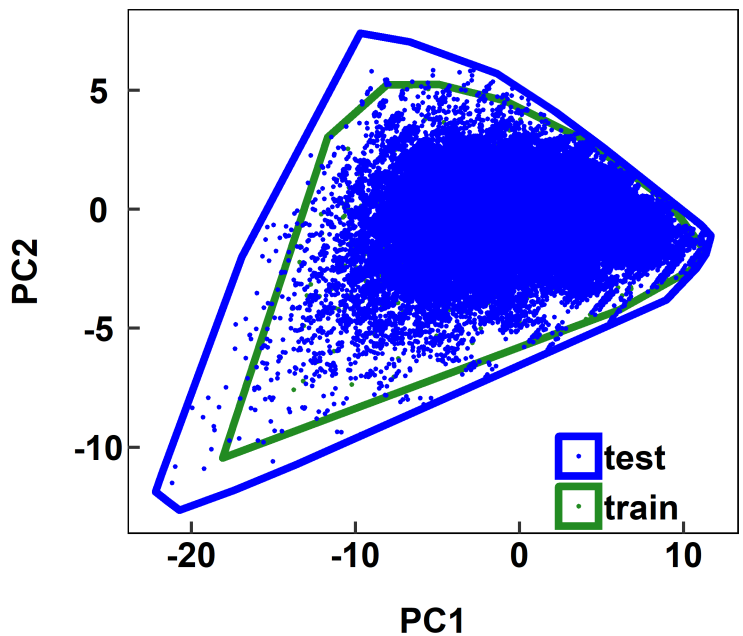
QM Atomization energy benchmark[1]: 130k small organic molecules commonly used to test ML.

[1]: Ramakrishnan, R., *et al.*, *Sci. Data*, 1, 2014.

QM9 atomization benchmark

QM Atomization energy benchmark[1]: 130k small organic molecules commonly used to test ML.

- 5% train and 95% test
- simple connectivity-only AC[2] descriptors
- Obtain MAE $\sim$ 6.5kcal/mol on test, performance uneven

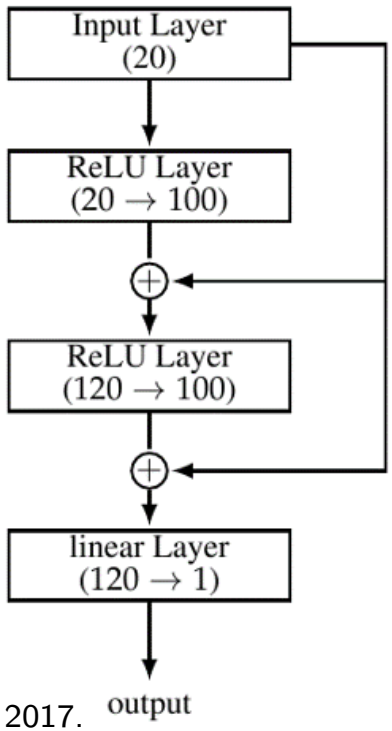


[1]: Ramakrishnan, R., *et al.*, *Sci. Data*, 1, 2014.
[2]: Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

QM9 atomization benchmark

QM Atomization energy benchmark[1]: 130k small organic molecules commonly used to test ML.

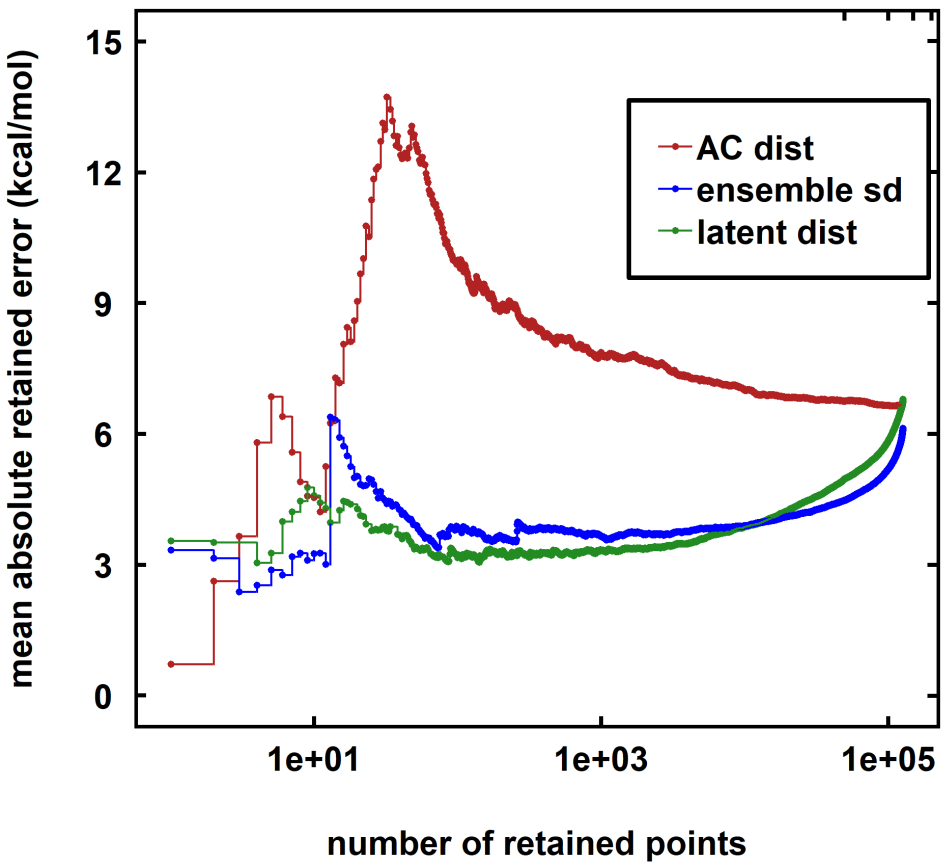
- 5% train and 95% test
- simple connectivity-only AC[2] descriptors
- Obtain MAE $\sim$ 6.5kcal/mol on test, performance uneven



[1]: Ramakrishnan, R., *et al.*, *Sci. Data*, 1, 2014.
[2]: Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

QM9 results

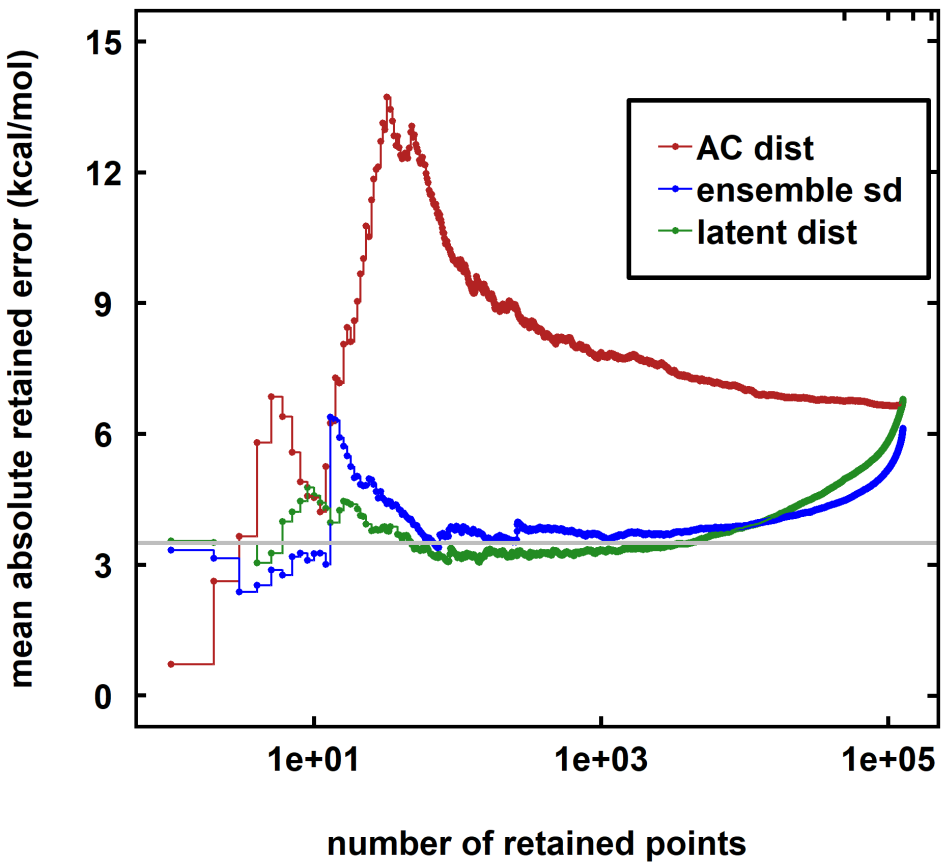
Similar error control can be obtained for this organic data[1]:



[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

QM9 results

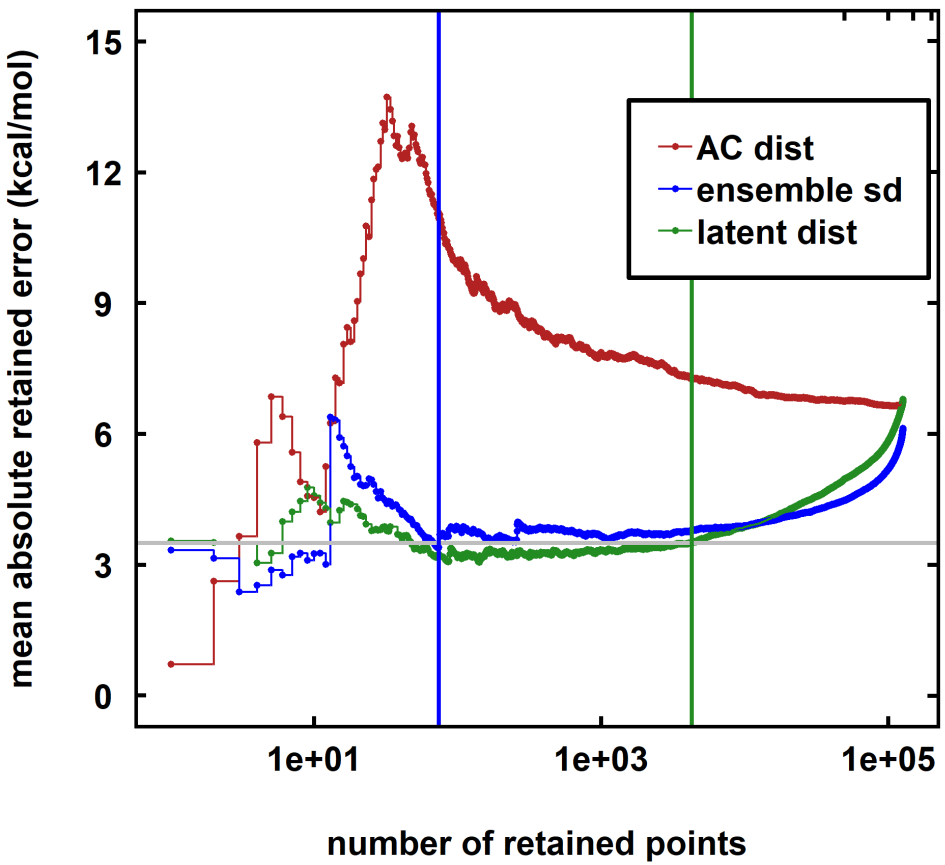
Similar error control can be obtained for this organic data[1]:



[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

QM9 results

Similar error control can be obtained for this organic data[1]:

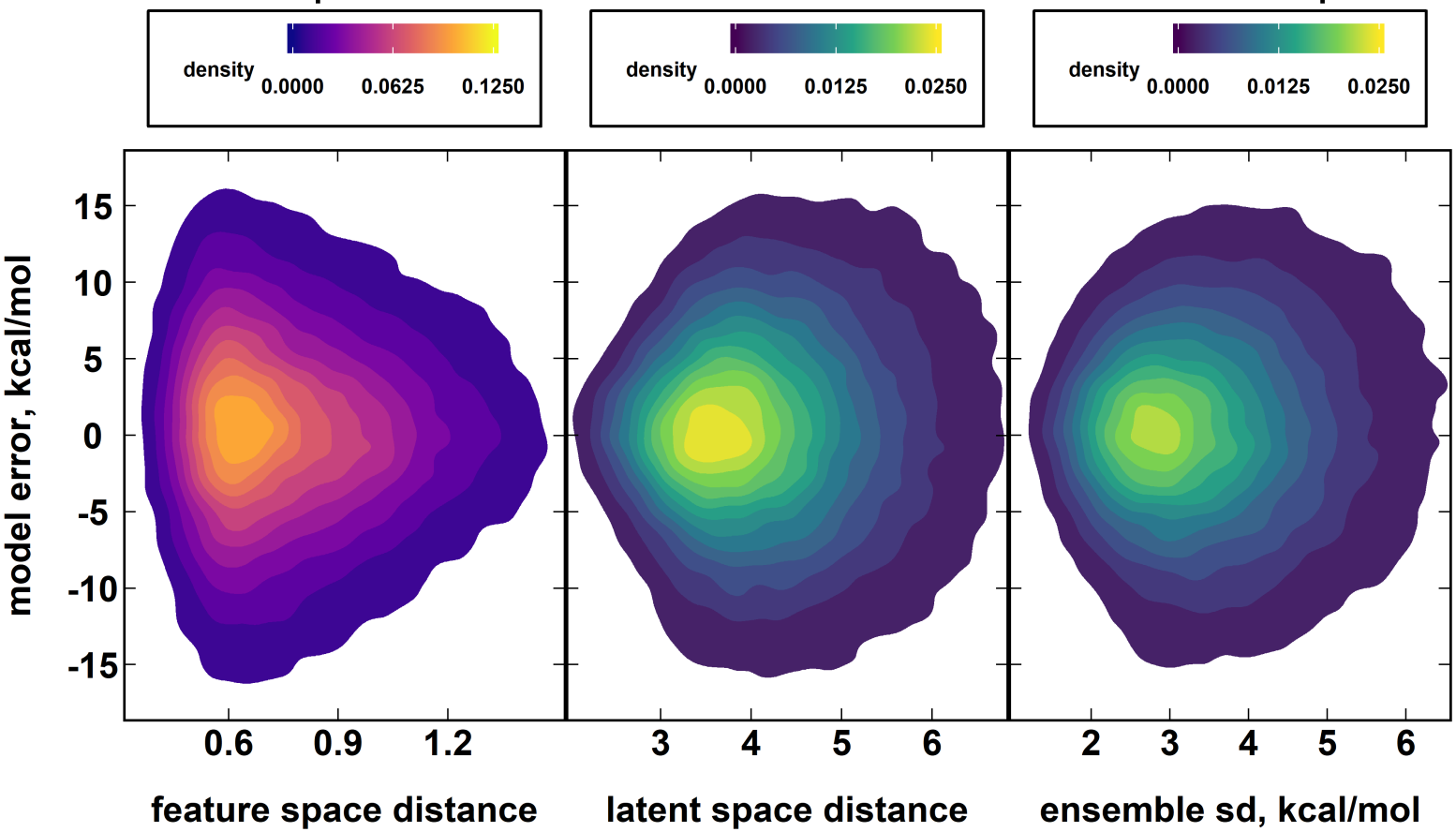


[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

QM9 results

Similar error control can be obtained for this organic data[1]:

We can compute distributions as before, based on 500 points:

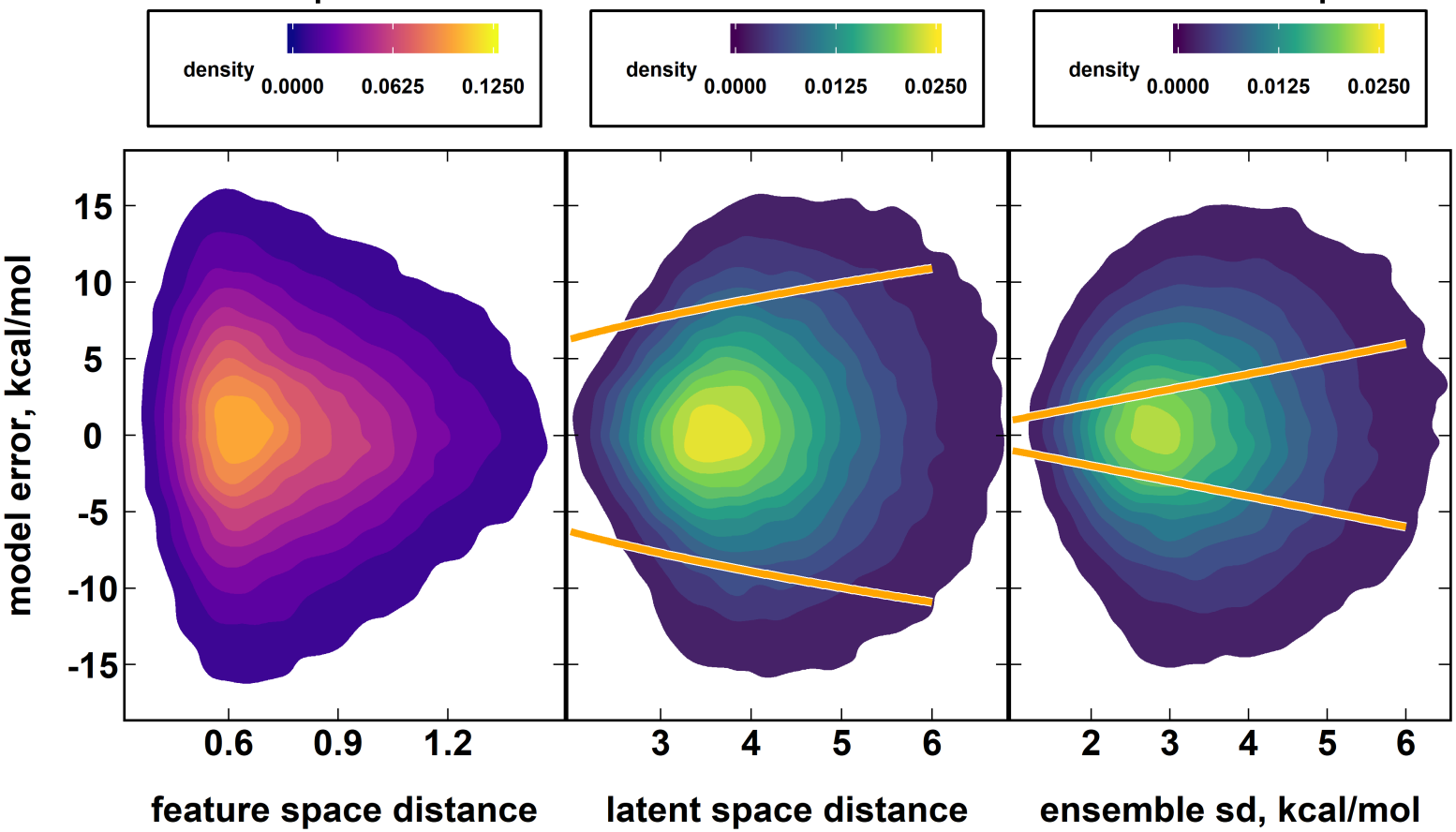


[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

QM9 results

Similar error control can be obtained for this organic data[1]:

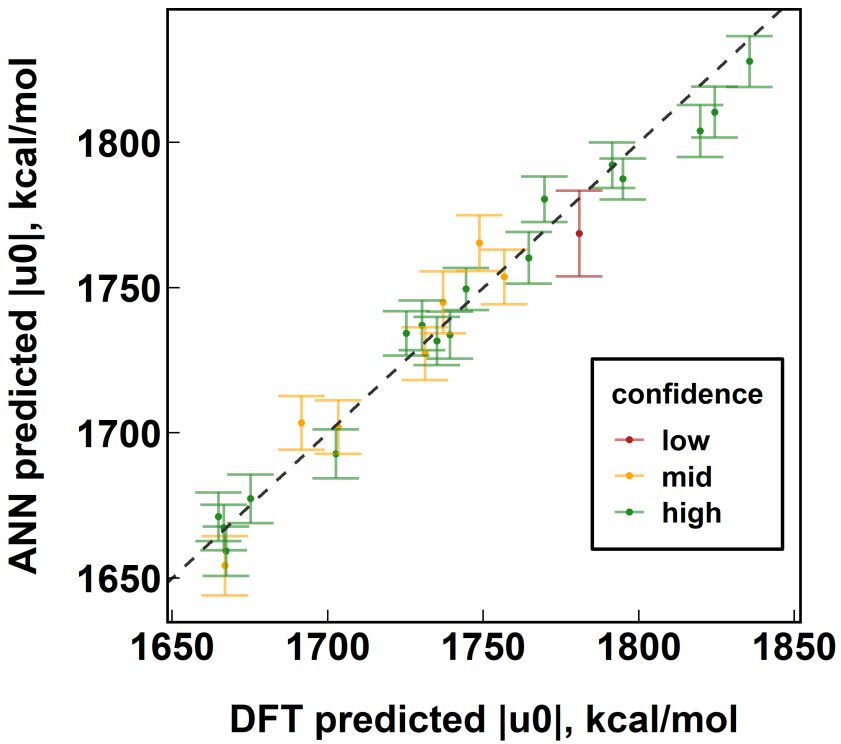
We can compute distributions as before, based on 500 points:



[1]: Janet, J.P., *et al.*, ChemRxiv, 10.26434/chemrxiv.7900277.v1

Summary

- The latent space provides substantially better results compared to raw feature space
- Out-of-sample errors can be controlled
- This method can provide useful errorbars on predictions and recover near-normal error quartiles



Acknowledgments

Thanks to the Kulik group and funding partners: