# Mapping transition metal chemical space with continuous descriptors – feature selection and implications for machine learning models

**Jon Paul Janet** [1] Terry Z.H. Gani [1] Heather Kulik [1]

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


254th American Chemical Society National Meeting
Washington, DC

08.23.17

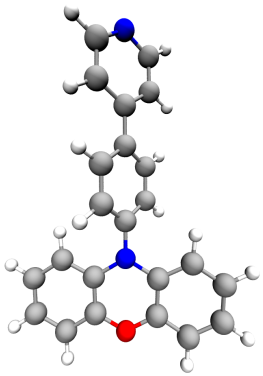
# Data-driven molecular design

**Machine learning is transforming how we design new materials**

# Data-driven molecular design

**Machine learning is transforming how we design new materials**

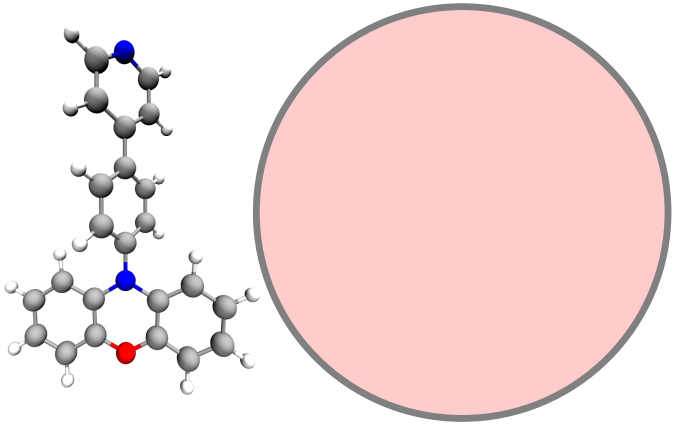
Gomez-Bombarelli, R. *et al.*. *Nat. Mater.*, 15(10):1120-1127, 2016.



# Data-driven molecular design

**Machine learning is transforming how we design new materials**

Gomez-Bombarelli, R. *et al.*. *Nat. Mater.*, 15(10):1120-1127, 2016.

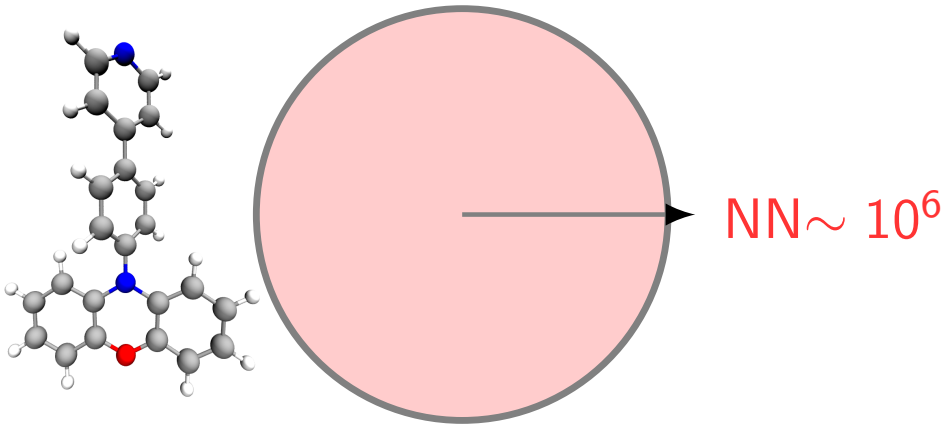


OLED chemical space

# Data-driven molecular design

**Machine learning is transforming how we design new materials**

Gomez-Bombarelli, R. *et al.*. *Nat. Mater.*, 15(10):1120-1127, 2016.

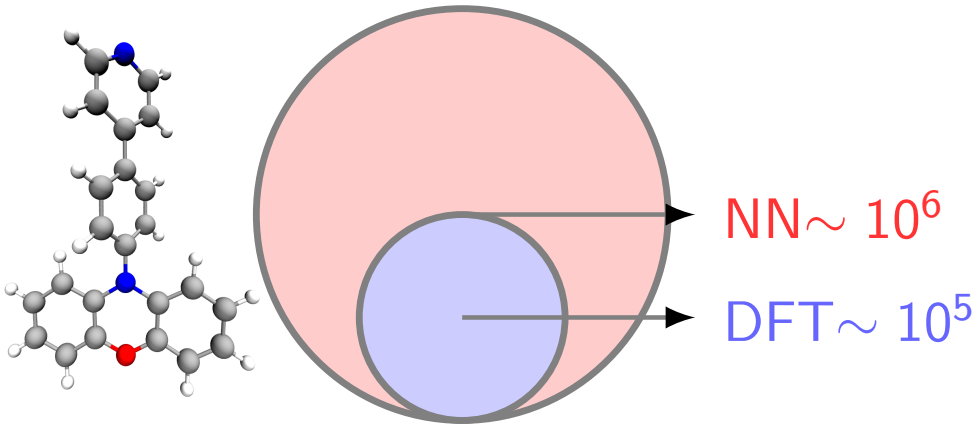



OLED chemical space

# Data-driven molecular design

**Machine learning is transforming how we design new materials**

Gomez-Bombarelli, R. *et al.*. *Nat. Mater.*, 15(10):1120-1127, 2016.

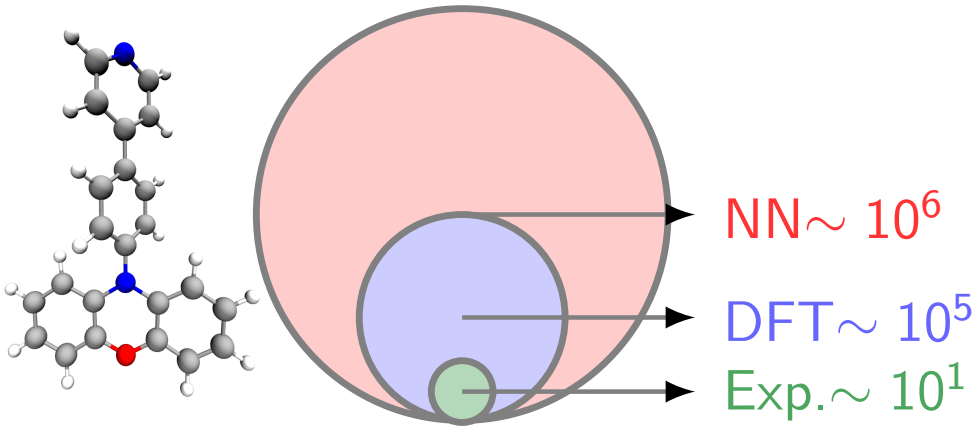



OLED chemical space

# Data-driven molecular design

**Machine learning is transforming how we design new materials**

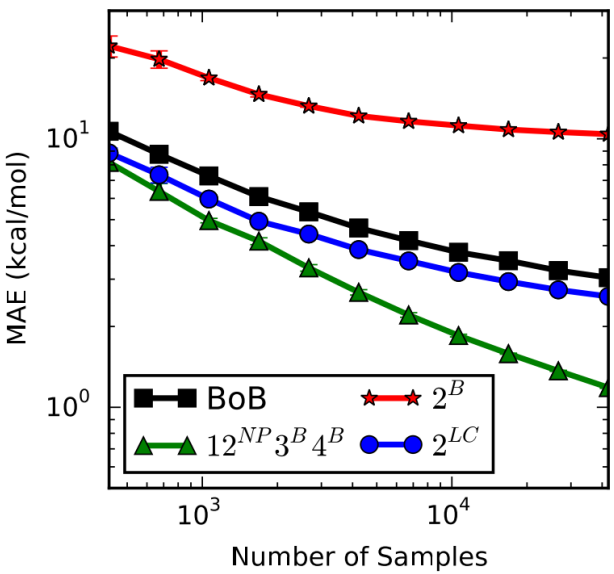
Gomez-Bombarelli, R. *et al.*. *Nat. Mater.*, 15(10):1120-1127, 2016.




OLED chemical space

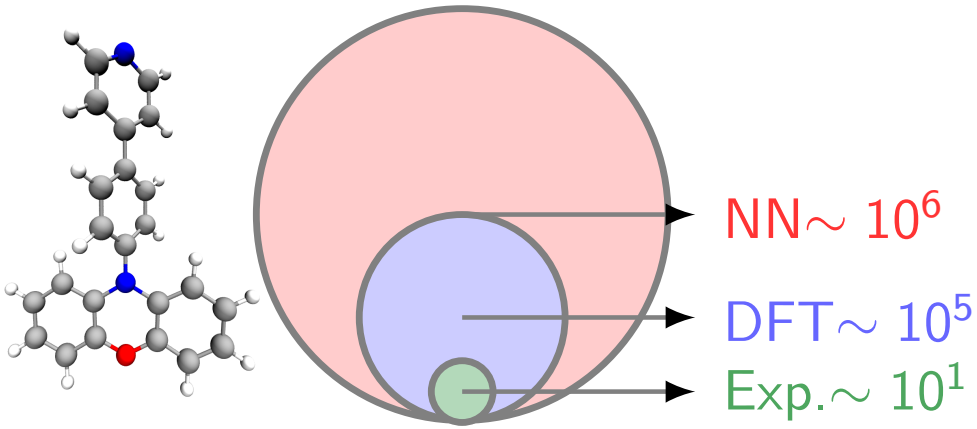
# Data-driven molecular design

**Machine learning is transforming how we design new materials**




Collins *et al.* *arXiv*, 1701.06649, 2017.

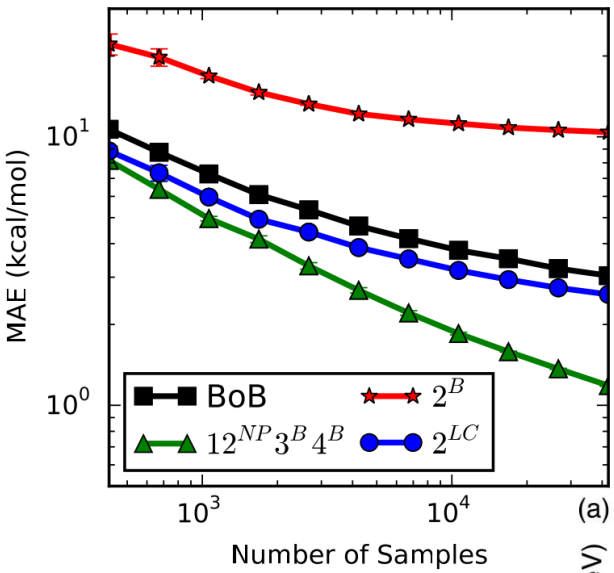
Gomez-Bombarelli, R. *et al.*. *Nat. Mater.*, 15(10):1120-1127, 2016.




OLED chemical space

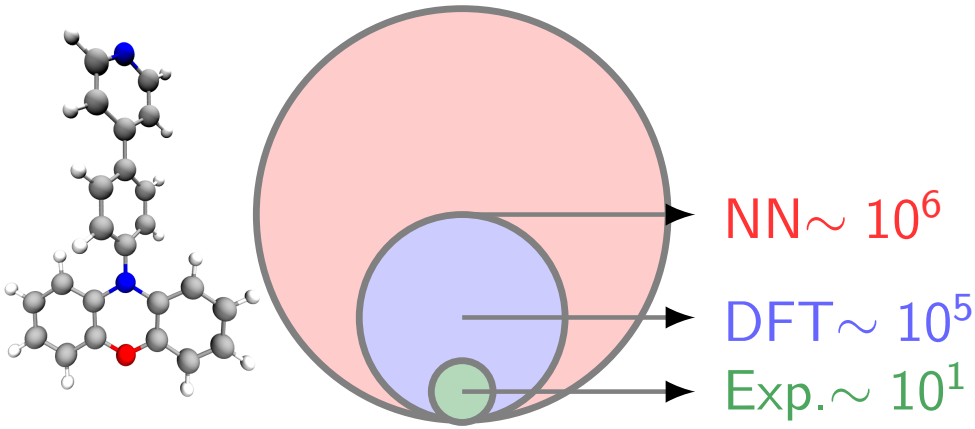
# Data-driven molecular design

**Machine learning is transforming how we design new materials**



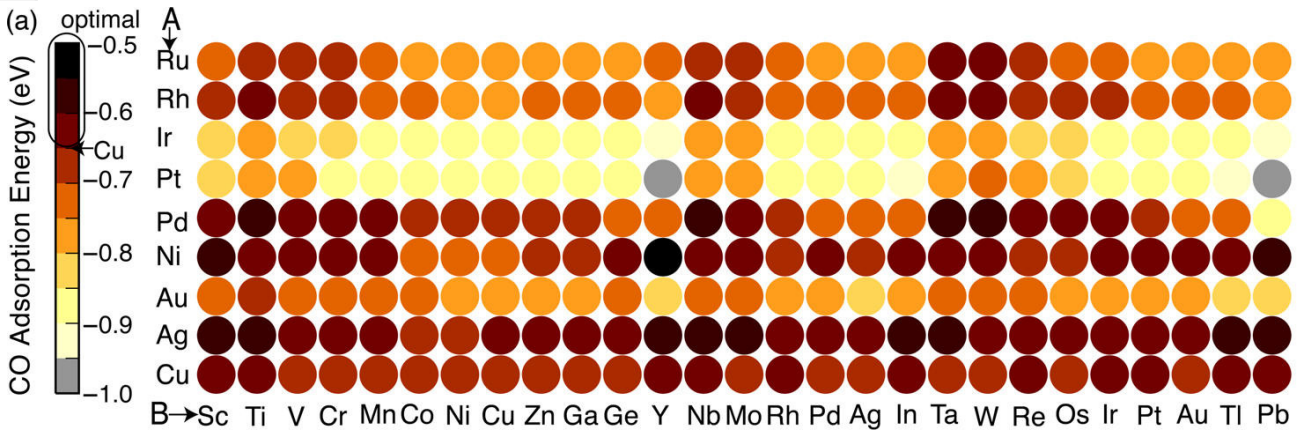

Collins *et al.* *arXiv*, 1701.06649, 2017.

Gomez-Bombarelli, R. *et al.*. *Nat. Mater.*, 15(10):1120-1127, 2016.



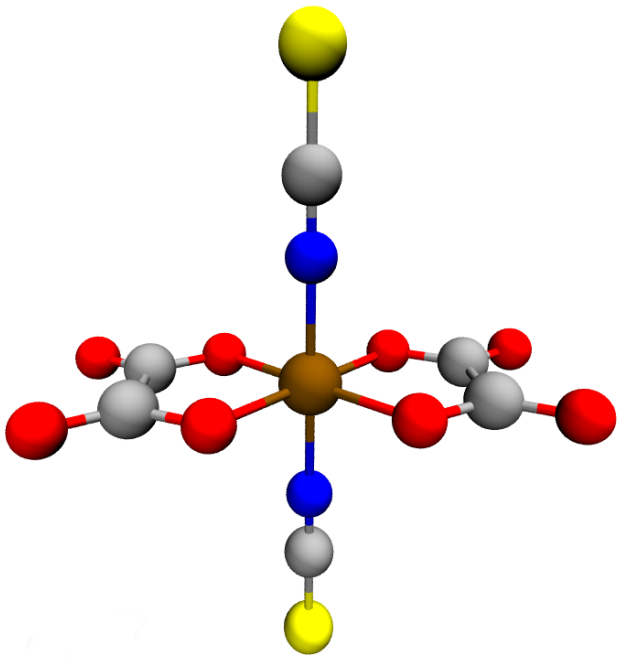

OLED chemical space

Ma, X. *et al.* J. Phys. Chem. Lett., (18):3528-3533, 2015.



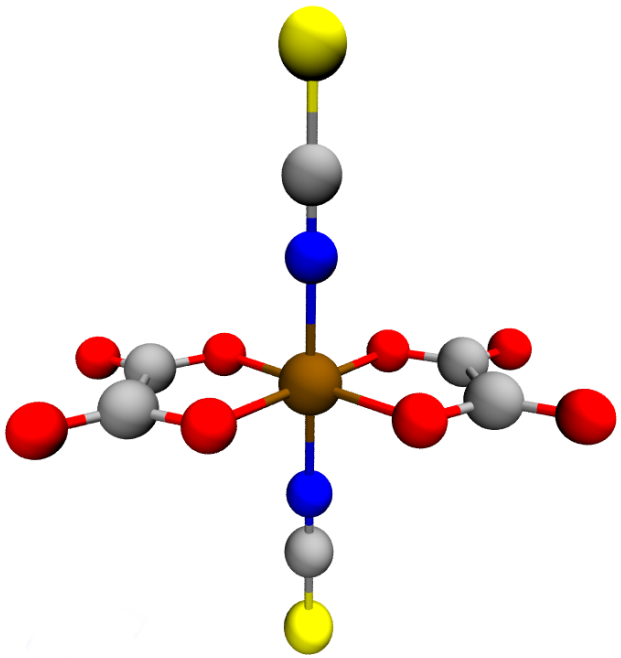
# Transition metal complexes

**We study transition metal complexes**:

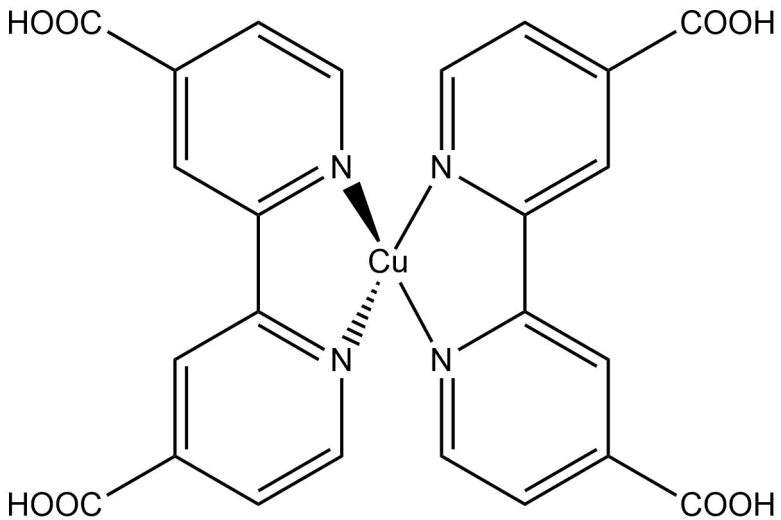


# Transition metal complexes

**We study transition metal complexes**:



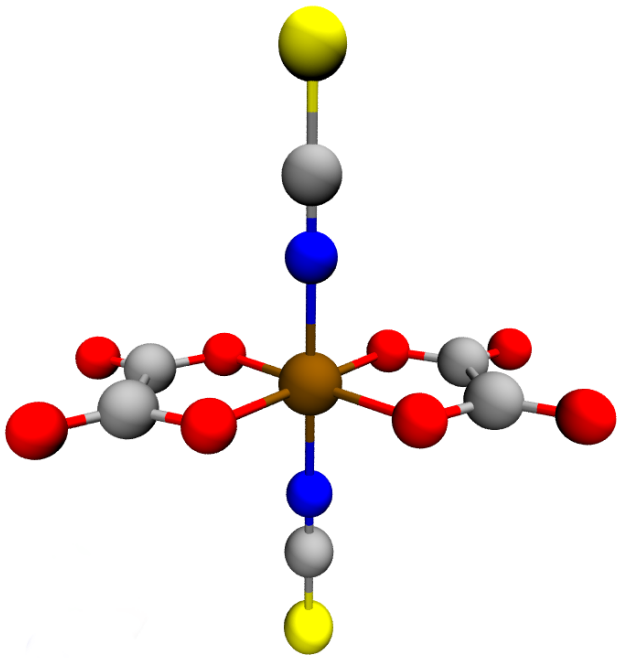
Dye sensitizers:



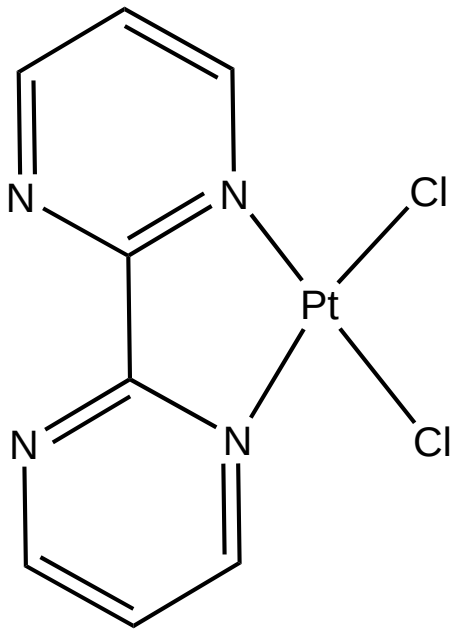

Bignozzi, C. *et al. Coord. Chem. Rev.*, 257(9): 1472–1492, 2013.

# Transition metal complexes

**We study transition metal complexes**:



Catalysts:

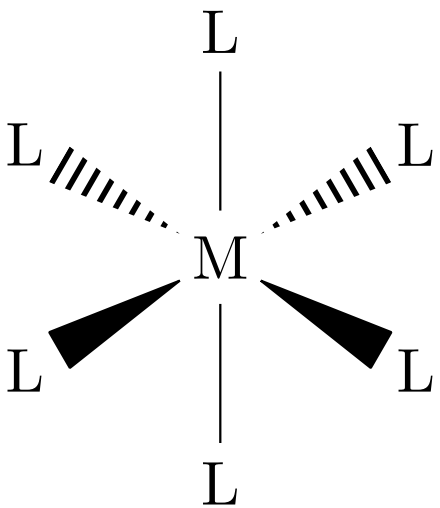



Periana, R. A. *et al.* *Science*, 280(5363):560–564, 1998.

# Transition metal complexes

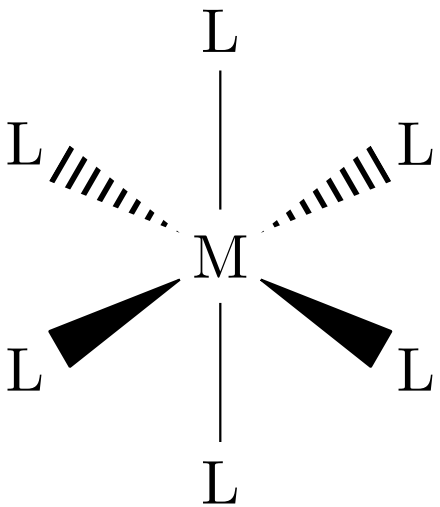
**We study transition metal complexes**:

✓ tunable properties



# Transition metal complexes

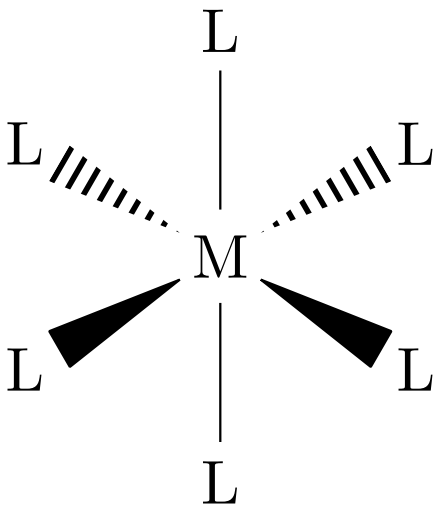
**We study transition metal complexes**:




- ✓ tunable properties
- ✓ great potential for molecular devices/catalysis

# Transition metal complexes

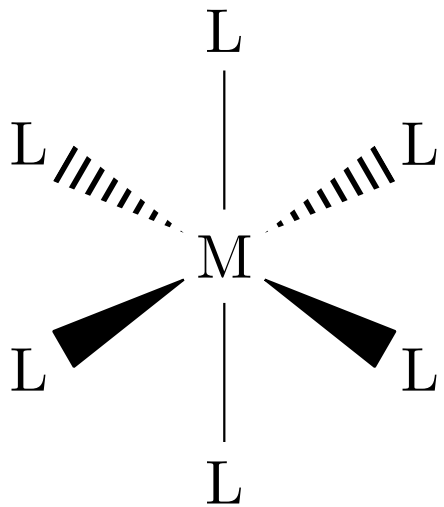
**We study transition metal complexes**:




✗ multiple spin states –
**ground state unknown**

# Transition metal complexes

**We study transition metal complexes**:

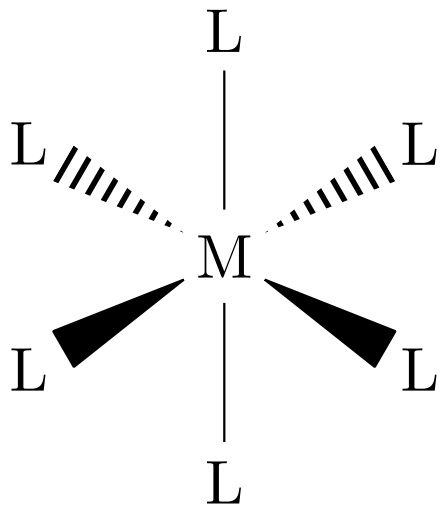



✗ multiple spin states – **ground state unknown**

✗ DFT calculations expensive – **need ML methods**

# Transition metal complexes

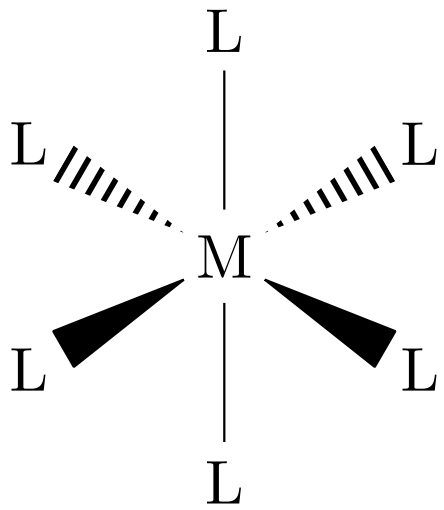
**We study transition metal complexes**:




- ✗ multiple spin states – **ground state unknown**
- ✗ DFT calculations expensive – **need ML methods**
- ✗ and not always reliable – **uncertainty in functional choice**

# Transition metal complexes

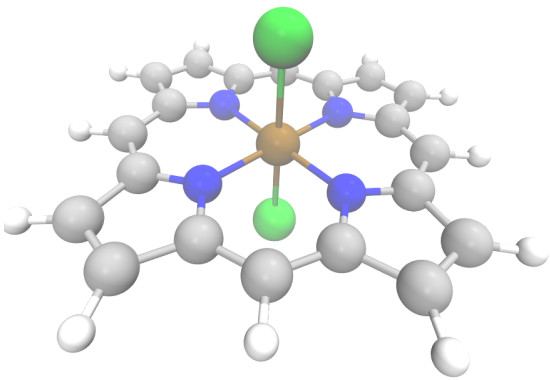
**We study transition metal complexes**:




- ✗ multiple spin states – **ground state unknown**
- ✗ DFT calculations expensive – **need ML methods**
- ✗ and not always reliable – **uncertainty in functional choice**
- ✗ difficult to get geometries a priori – **starting structures difficult**

# ML for TM complexes

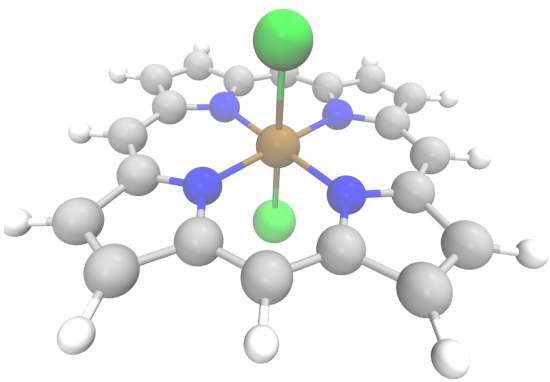
How can ML help?

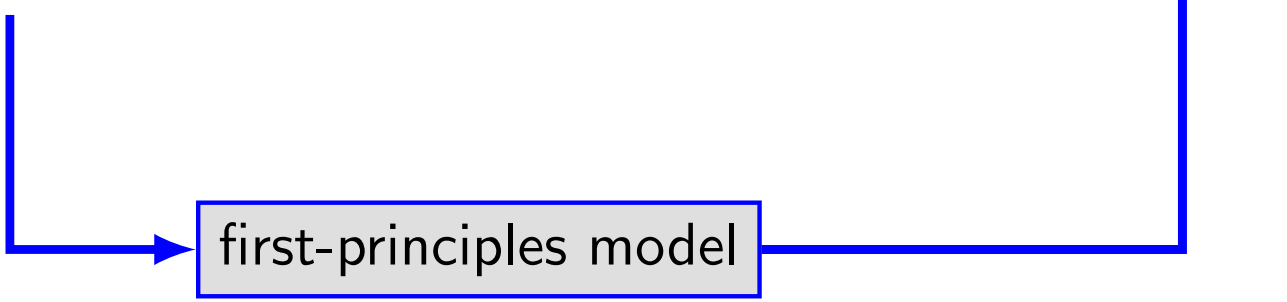


property

# ML for TM complexes

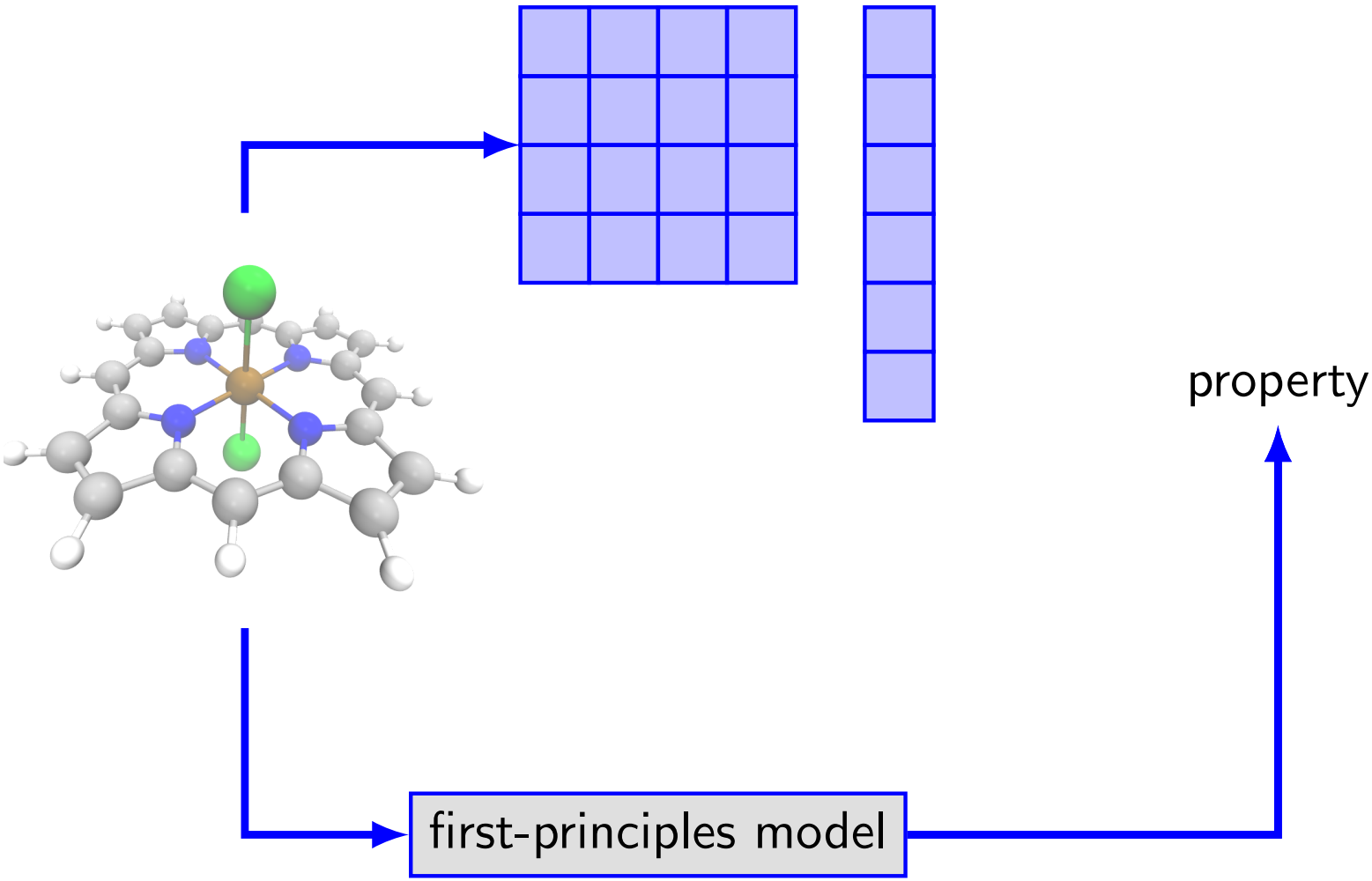
How can ML help?





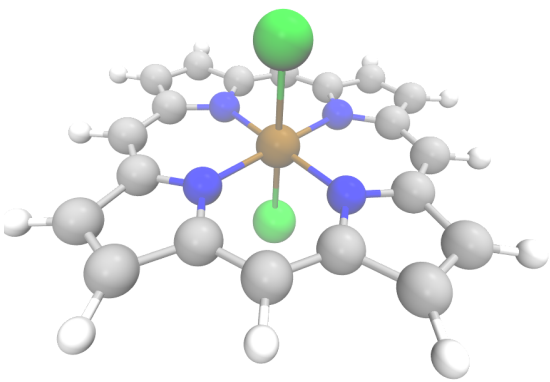
# ML for TM complexes

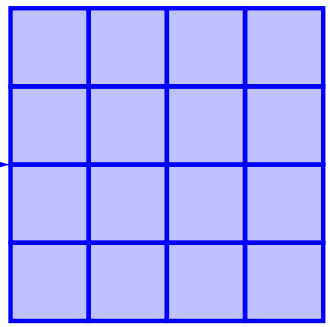
How can ML help?

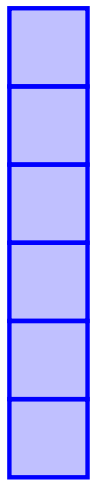


# ML for TM complexes

How can ML help?







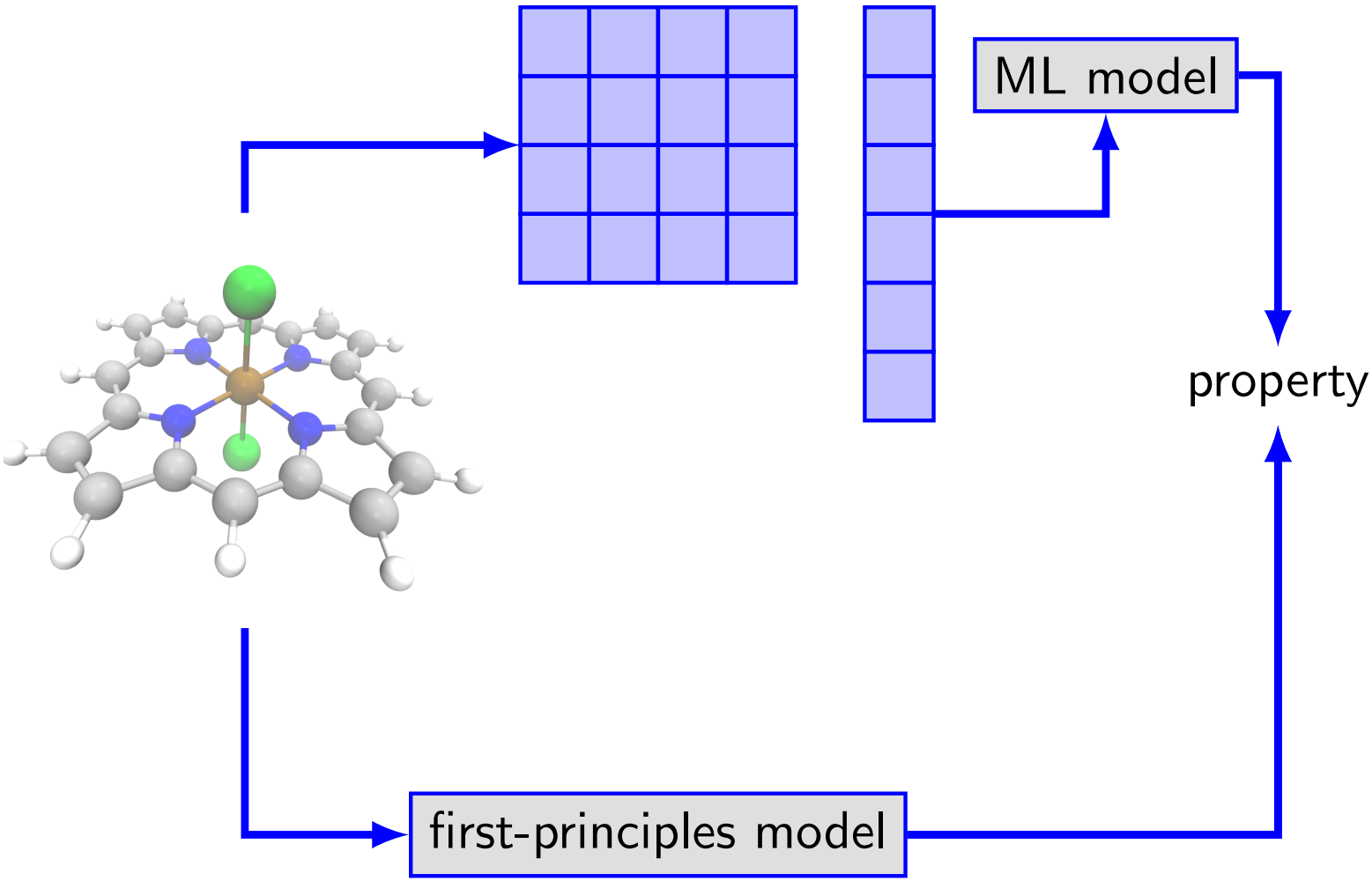
dimension? type?

property

first-principles model

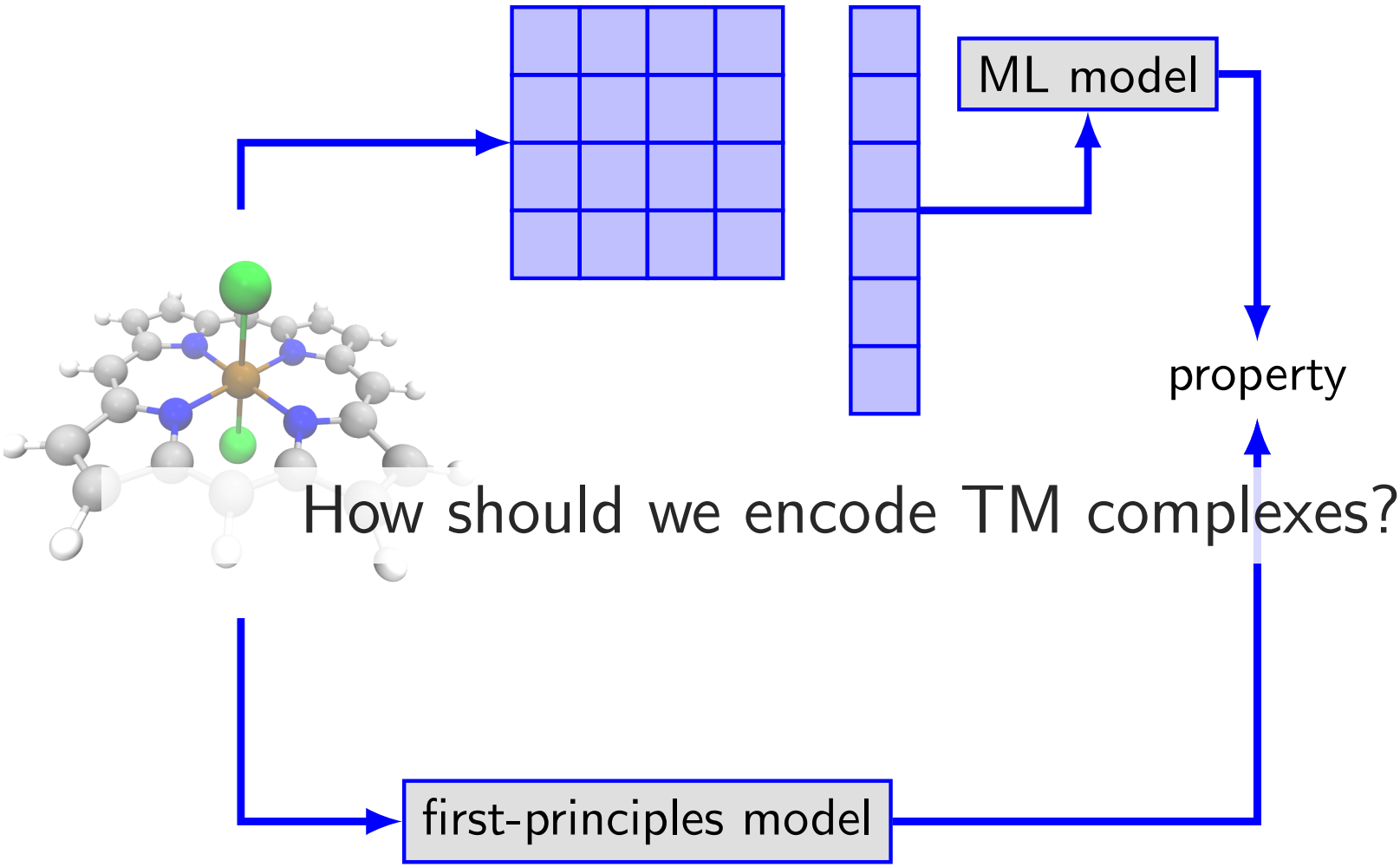
# ML for TM complexes

How can ML help?

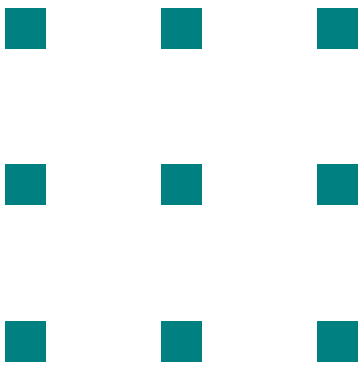


# ML for TM complexes

How can ML help?

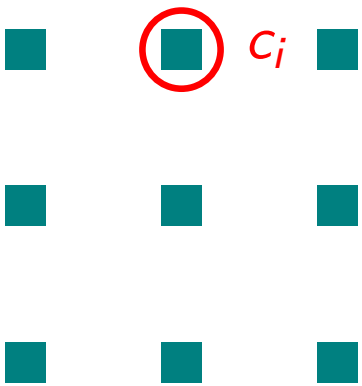


# Descriptors define chemical space



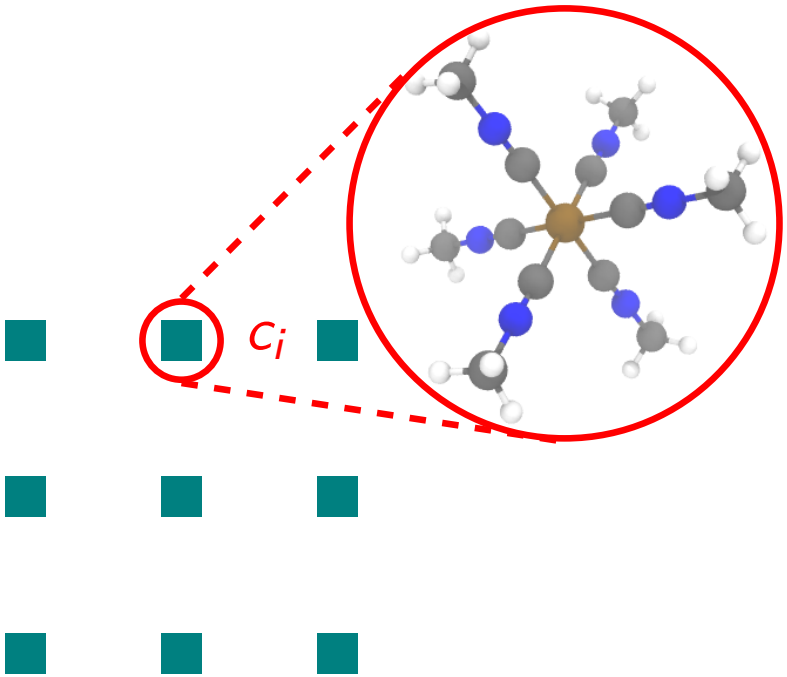
**Chemical Space** $C_f$

# Descriptors define chemical space



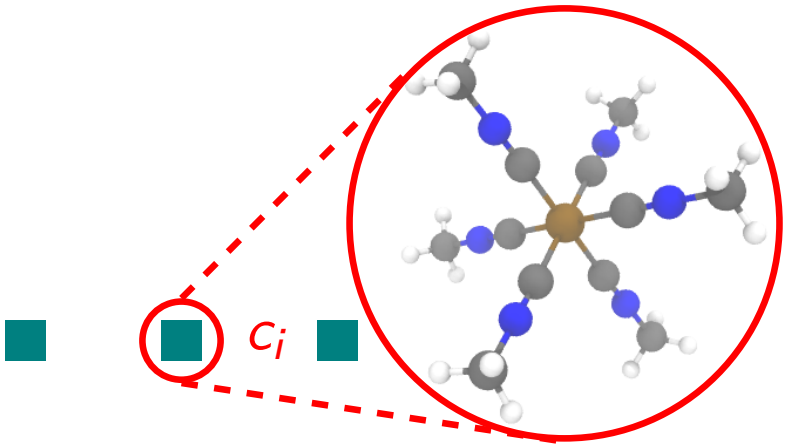

**Chemical Space** $C_f$

# Descriptors define chemical space




**Chemical Space** $C_f$

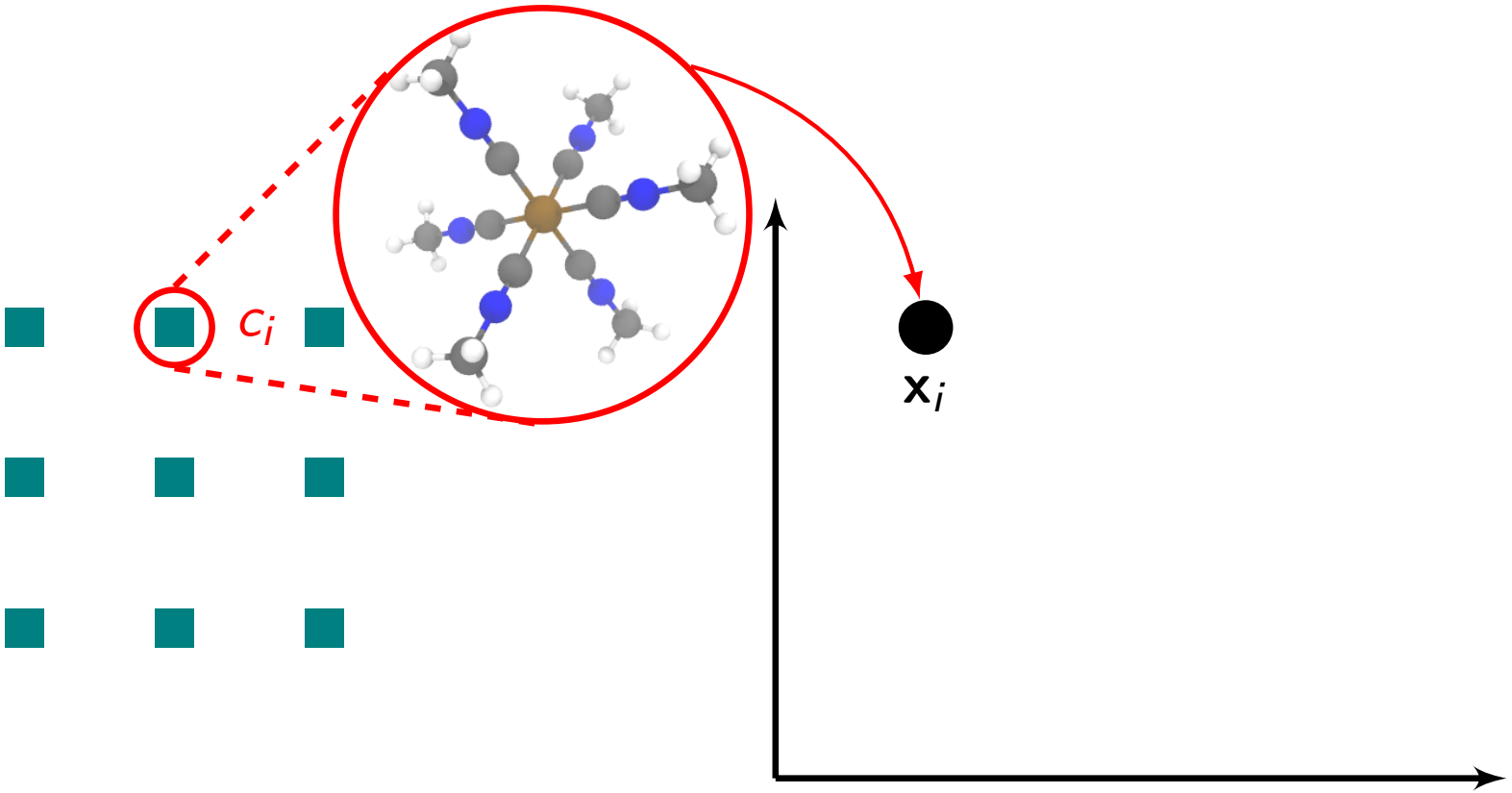
# Descriptors define chemical space




**Chemical Space** $C_f$

**Descriptor Space** $\mathcal{X} \subset \mathbb{R}^d$

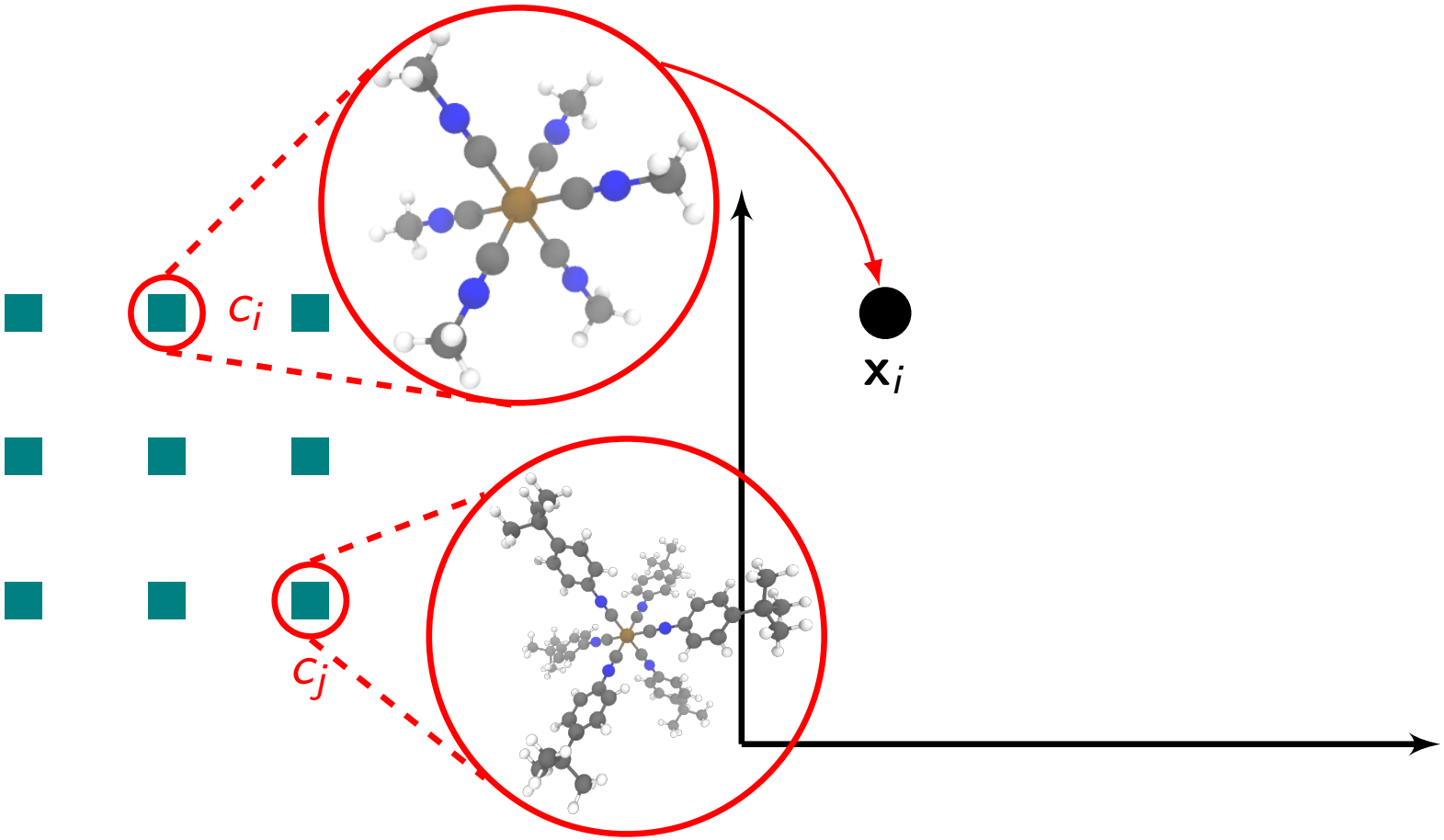
# Descriptors define chemical space




**Chemical Space** $C_f$

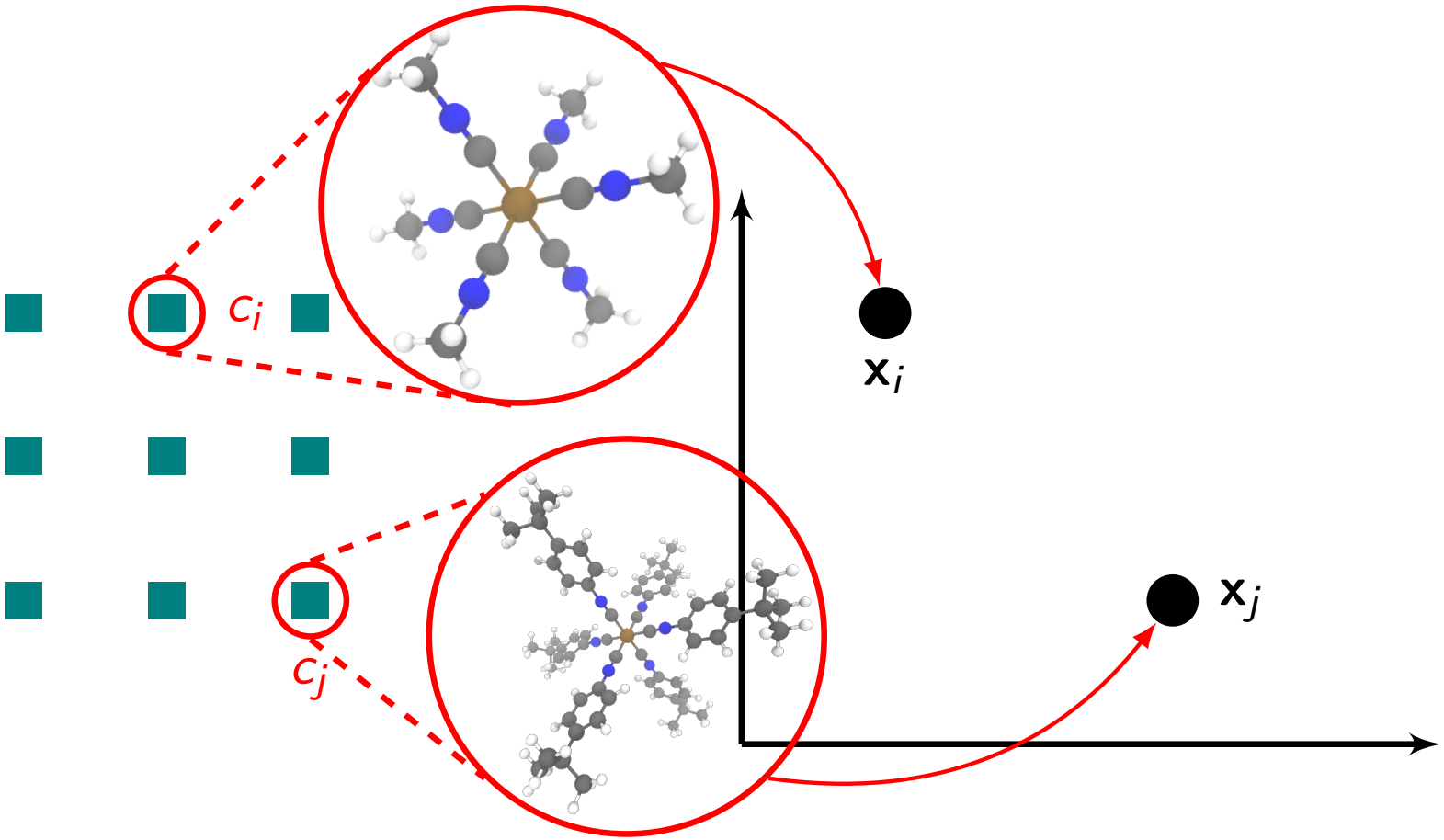
**Descriptor Space** $\mathcal{X} \subset \mathbb{R}^d$

# Descriptors define chemical space



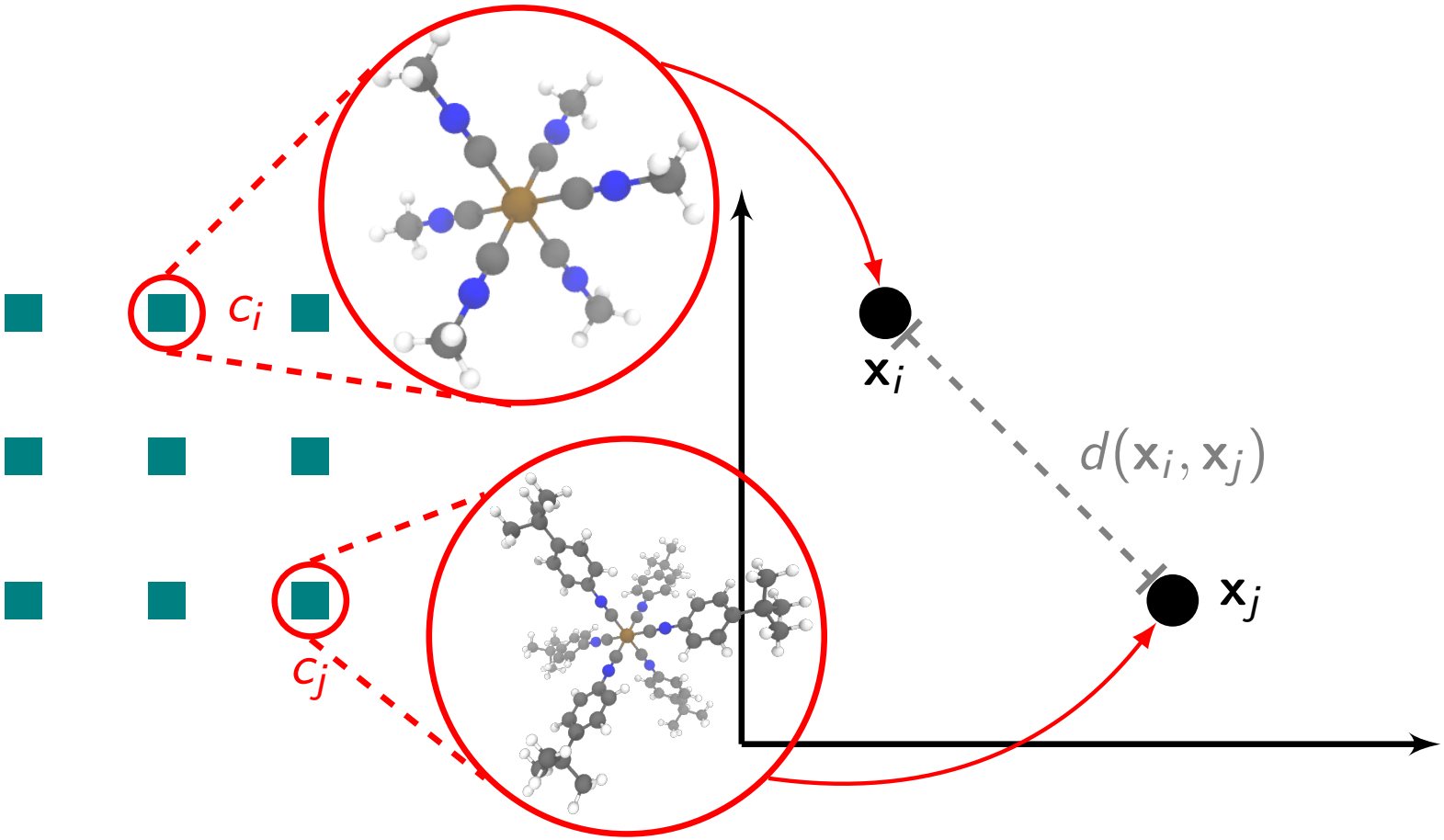

**Chemical Space** $C_f$ **Descriptor Space** $\mathcal{X} \subset \mathbb{R}^d$

# Descriptors define chemical space



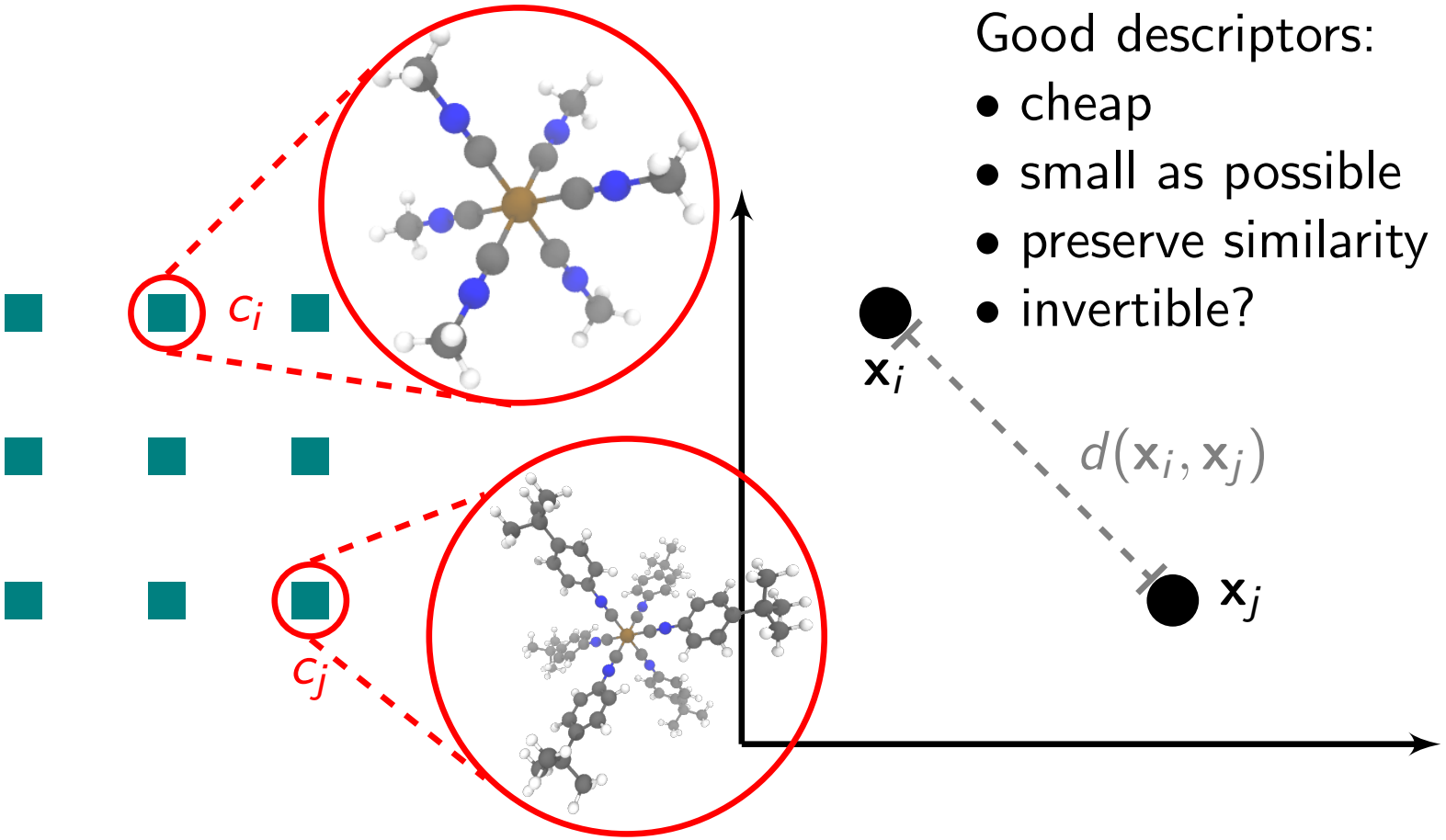

**Chemical Space** $C_f$ **Descriptor Space** $\mathcal{X} \subset \mathbb{R}^d$

# Descriptors define chemical space




**Chemical Space** $C_f$ **Descriptor Space** $\mathcal{X} \subset \mathbb{R}^d$

# Descriptors define chemical space




Good descriptors:

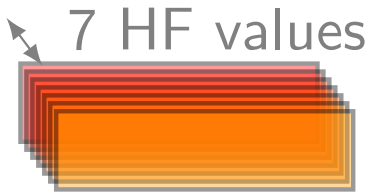
- cheap
- small as possible
- preserve similarity
- invertible?

**Chemical Space** $C_f$

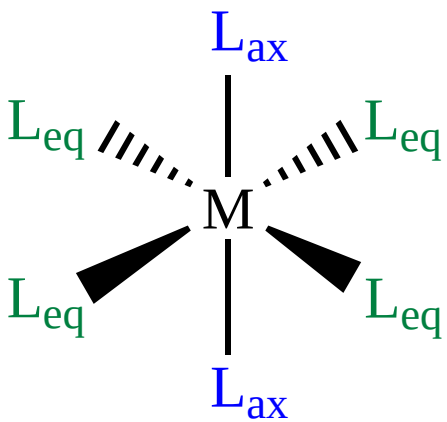
**Descriptor Space** $\mathcal{X} \subset \mathbb{R}^d$

# Predictive modeling

Data for octahedral complexes[1]:



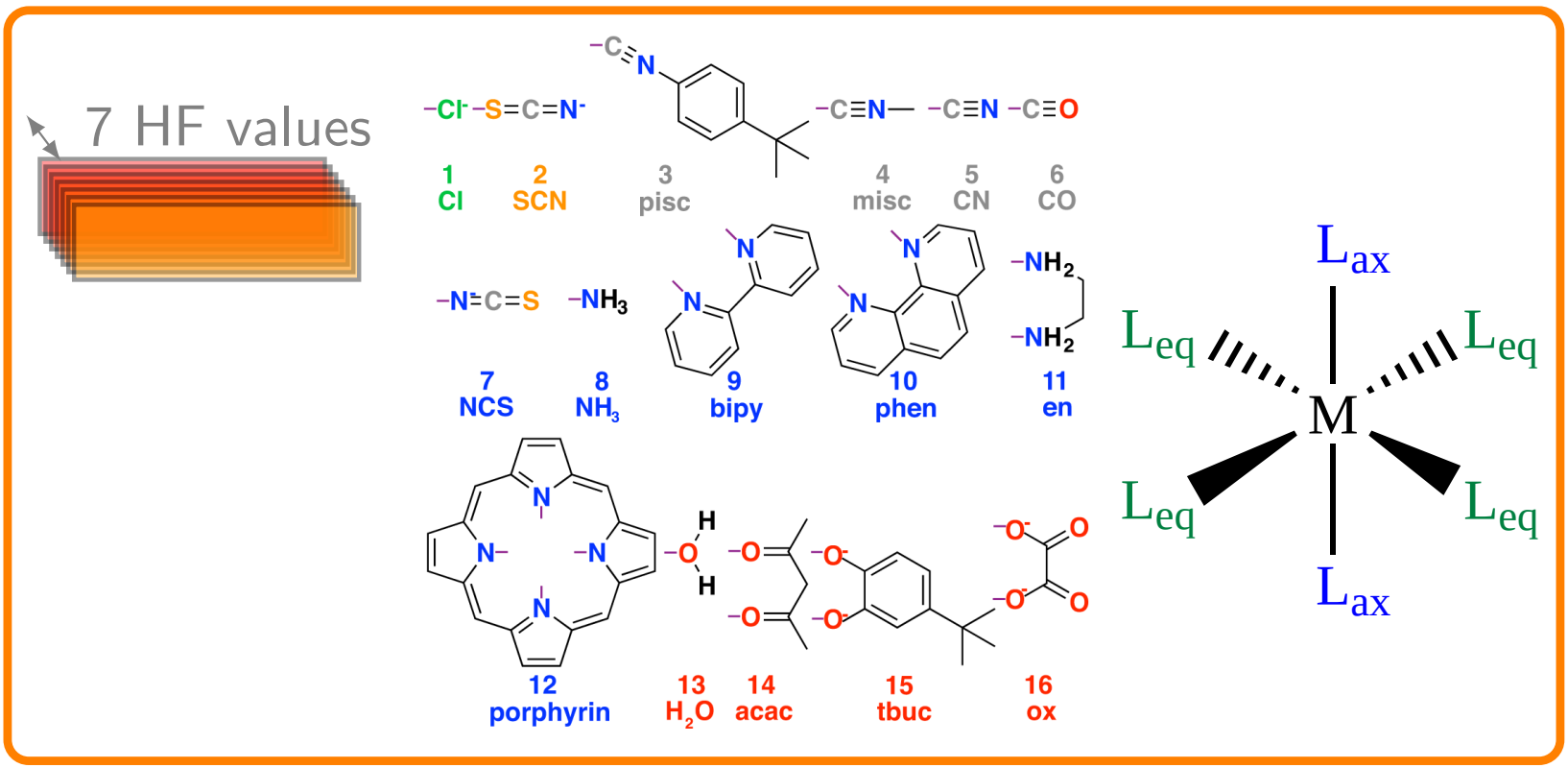

1345 (194) complexes




[1] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

# Predictive modeling

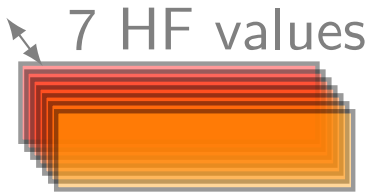
Data for octahedral complexes[1]:



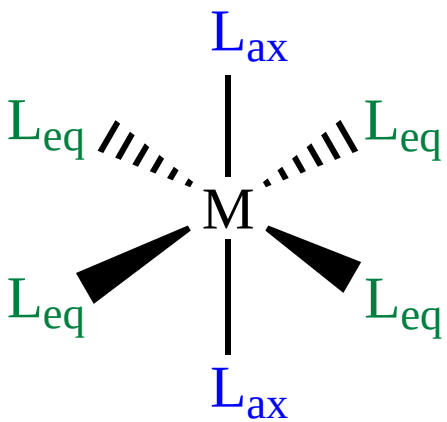

[1] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

# Predictive modeling

Data for octahedral complexes[1]:



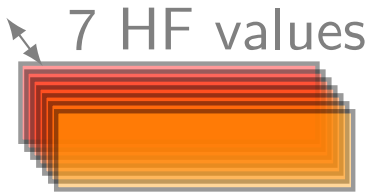

1345 (194) complexes




[1] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

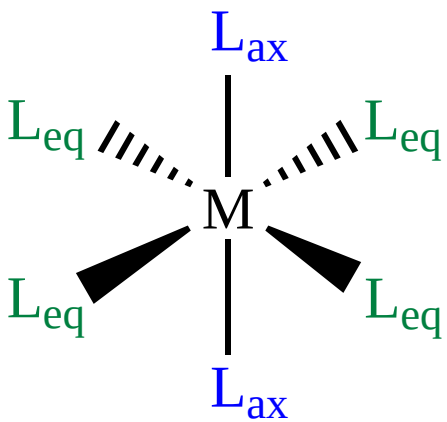
# Predictive modeling

Data for octahedral complexes[1]:




1345 (194) complexes

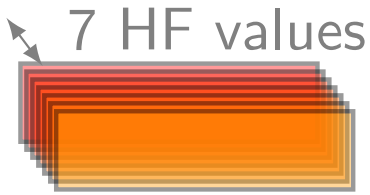
B3LYP-like DFT
HF exchange in 0-30%
gas phase optimizaton
LANL2DZ/6-31G*
high- and low-spin
M(II)/(III)




[1] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

# Predictive modeling

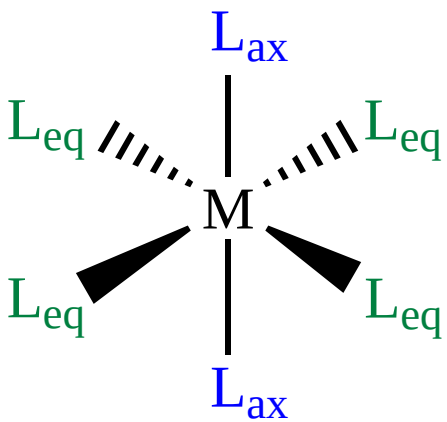
Data for octahedral complexes[1]:




1345 (194) complexes

built by molSimplify

`molsimplify.mit.edu`




[1] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

# ML for TM complexes

Can we (computationally) describe TM complexes with off-the-shelf tools? Coulomb matrix descriptor[2],KRR:

[2]Rupp, M., *et al.*. *Phys. Rev. Lett.*, 108(5), 58301-58311, 2012.

# ML for TM complexes

Can we (computationally) describe TM complexes with off-the-shelf tools? Coulomb matrix descriptor[2],KRR:

| | test data RMSE, kcal/mol |
|---|---|
| CM-ES | 19.2 |

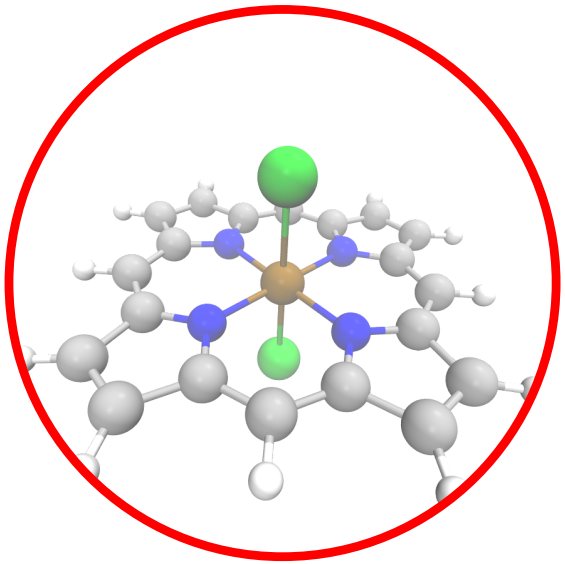
[2] Rupp, M., *et al.*. *Phys. Rev. Lett.*, 108(5), 58301-58311, 2012.

# ML for TM complexes

Can we (computationally) describe TM complexes with off-the-shelf tools? Coulomb matrix descriptor[2],KRR:

| | test data RMSE, kcal/mol |
|---|---|
| CM-ES | 19.2 |

Let's try to understand why...



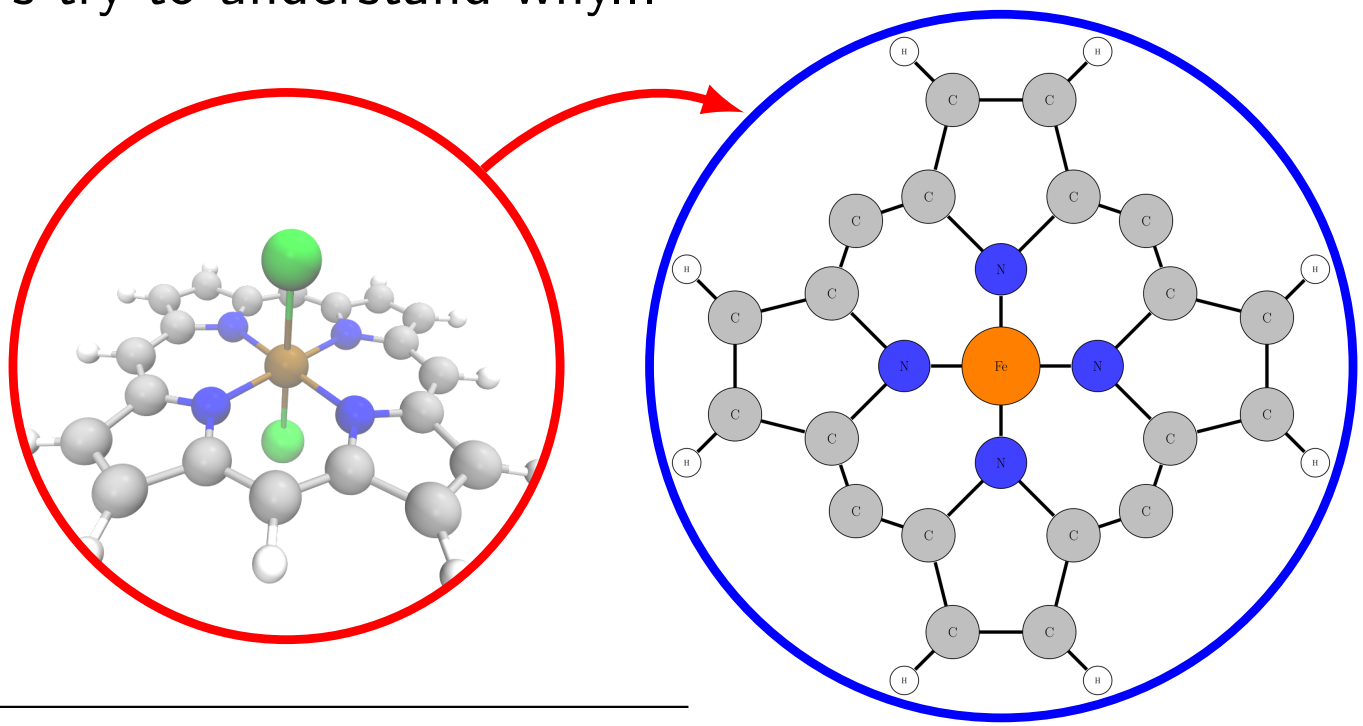
[2]Rupp, M., *et al.*. *Phys. Rev. Lett.*, 108(5), 58301-58311, 2012.

# ML for TM complexes

Can we (computationally) describe TM complexes with off-the-shelf tools? Coulomb matrix descriptor[2],KRR:

| | test data RMSE, kcal/mol |
|---|---|
| CM-ES | 19.2 |

Let's try to understand why...



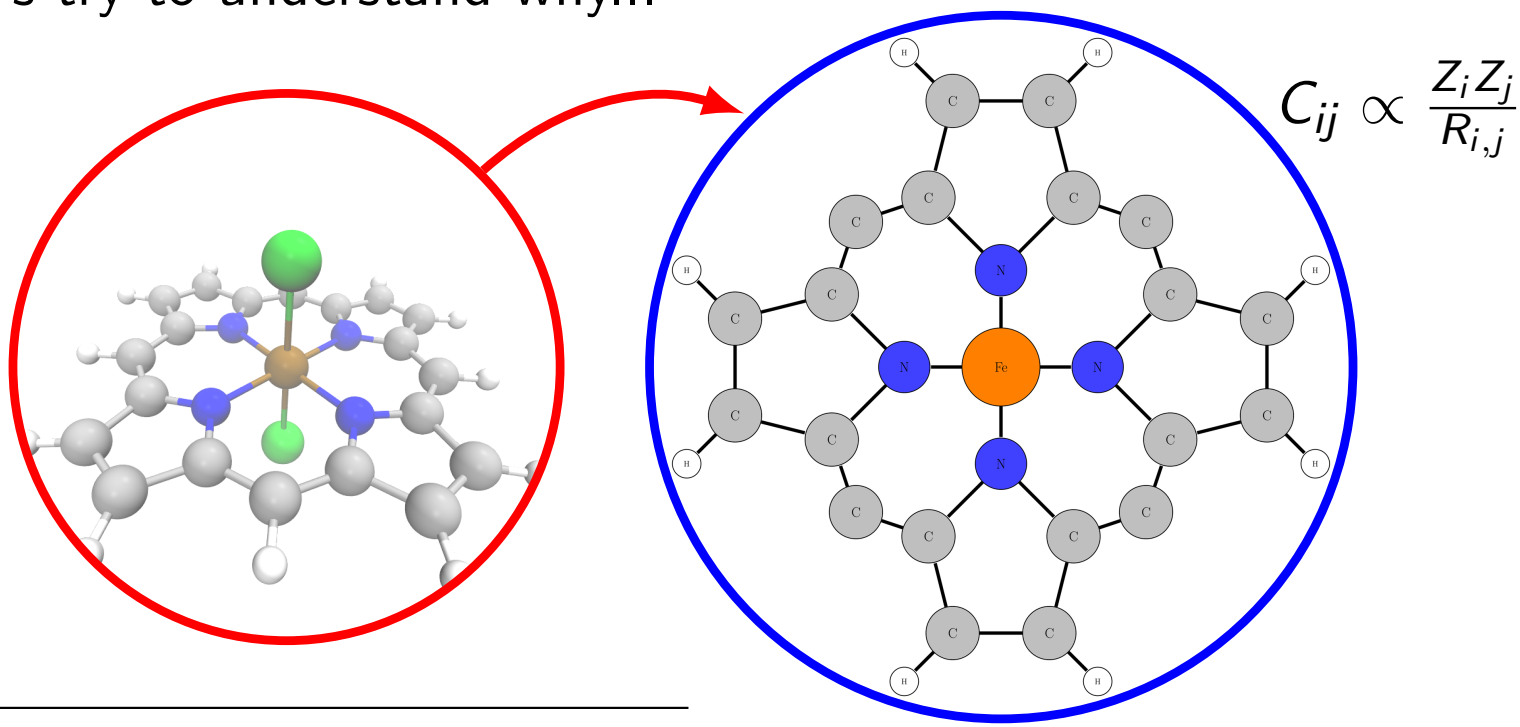

[2]Rupp, M., *et al.*. *Phys. Rev. Lett.*, 108(5), 58301-58311, 2012.

# ML for TM complexes

Can we (computationally) describe TM complexes with off-the-shelf tools? Coulomb matrix descriptor[2],KRR:

| | test data RMSE, kcal/mol |
|---|---|
| CM-ES | 19.2 |

Let's try to understand why...



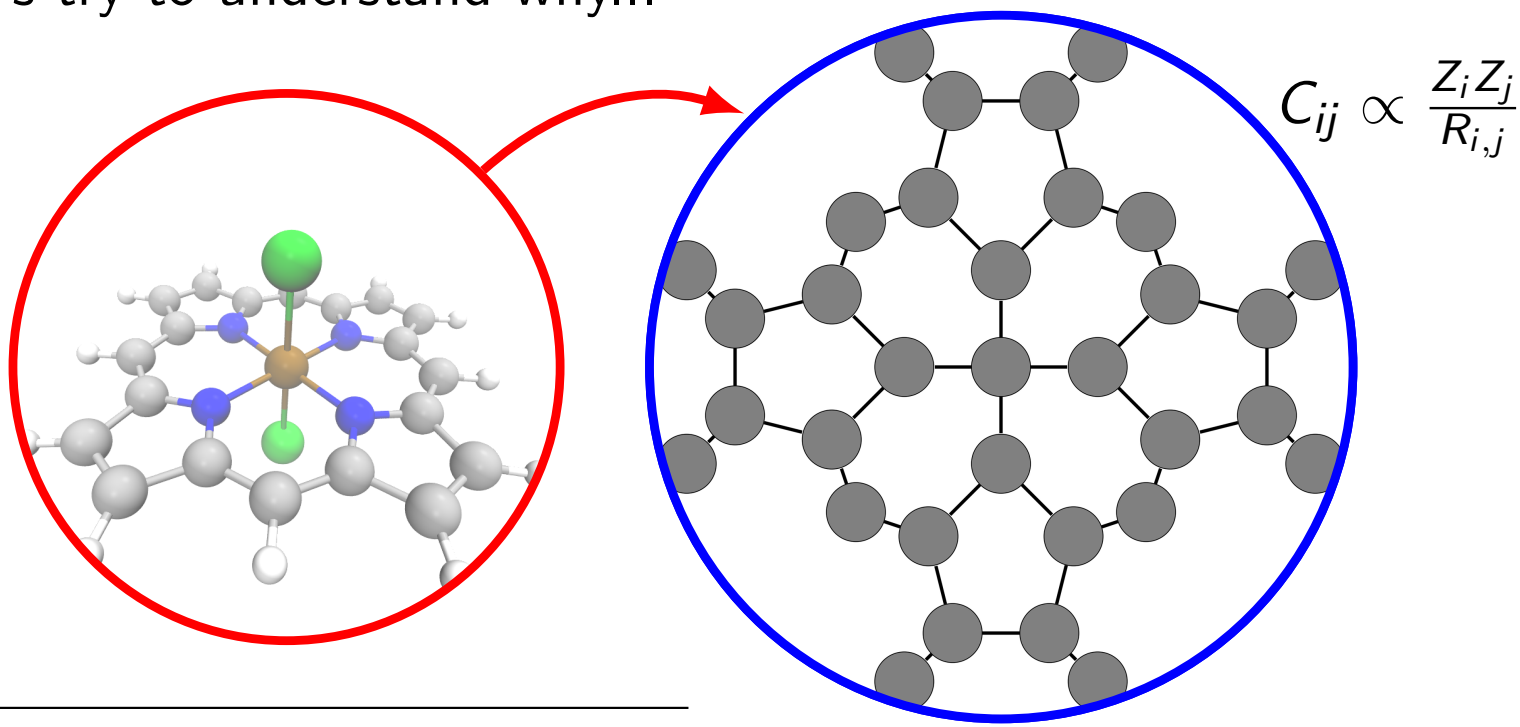

[2]Rupp, M., *et al.*. *Phys. Rev. Lett.*, 108(5), 58301-58311, 2012.

# ML for TM complexes

Can we (computationally) describe TM complexes with off-the-shelf tools? Coulomb matrix descriptor[2],KRR:

| | test data RMSE, kcal/mol |
|---|---|
| CM-ES | 19.2 |

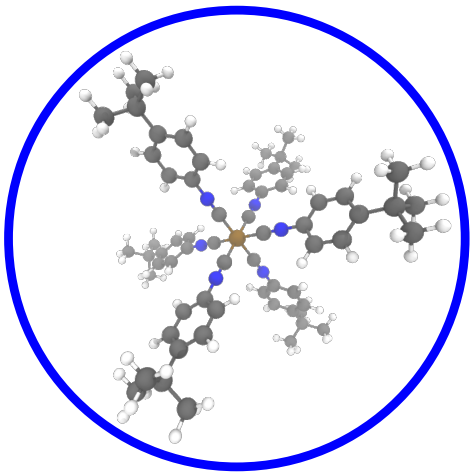
Let's try to understand why...

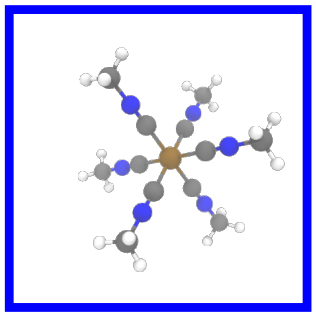



[2]Rupp, M., *et al.*. *Phys. Rev. Lett.*, 108(5), 58301-58311, 2012.

# A tale of two complexes

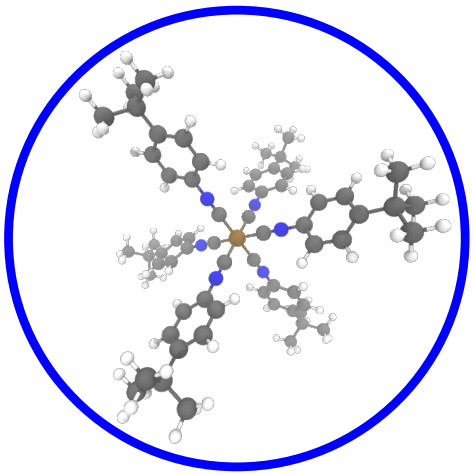
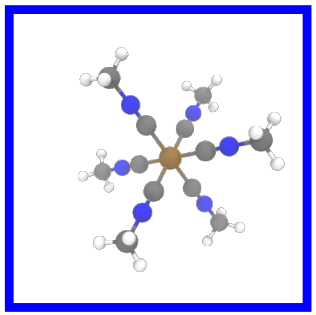
$Fe[pisc]_6^{3+}$





$Fe[misc]_6^{3+}$

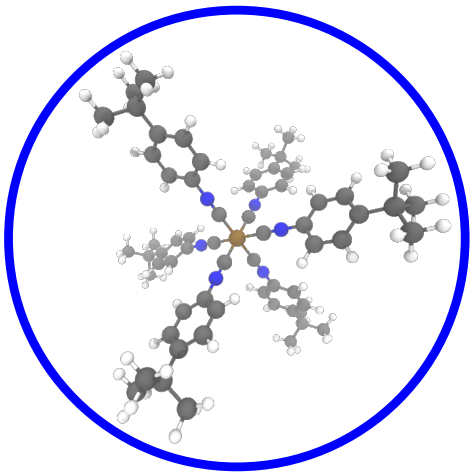
# A tale of two complexes

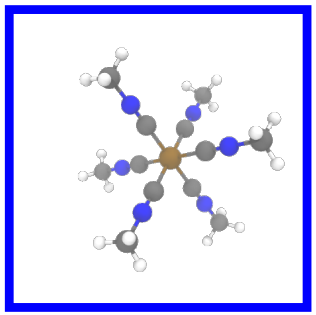
$Fe[pisc]_6^{3+}$   $Fe[misc]_6^{3+}$

$\Delta E_{H\text{-}L} = 37.7$ kcal/mol $\Delta E_{H\text{-}L} = 40.7$ kcal/mol

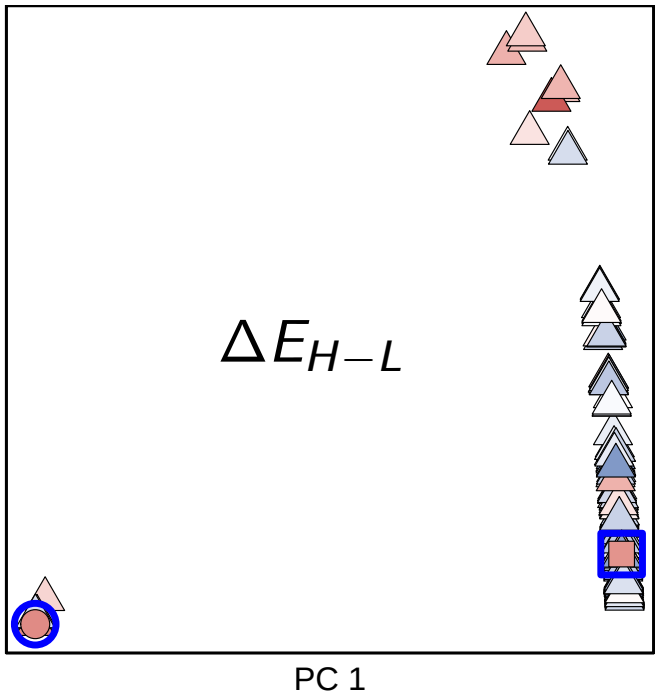
# A tale of two complexes

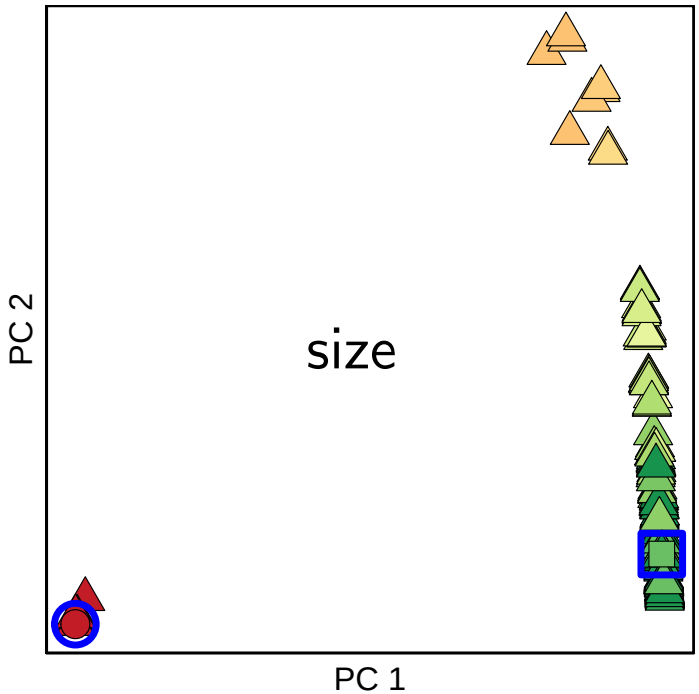
$Fe[pisc]_6^{3+}$



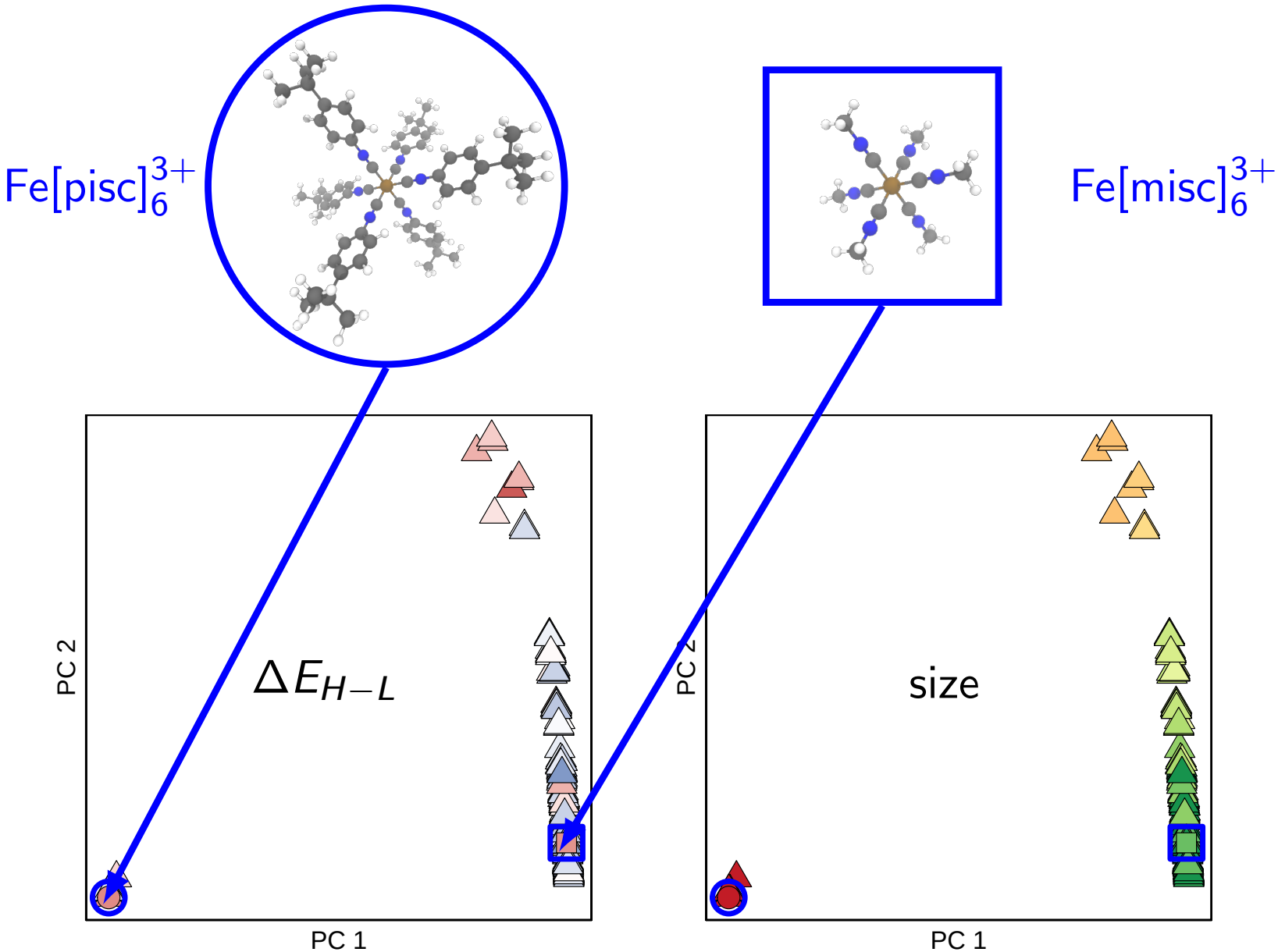


$Fe[misc]_6^{3+}$

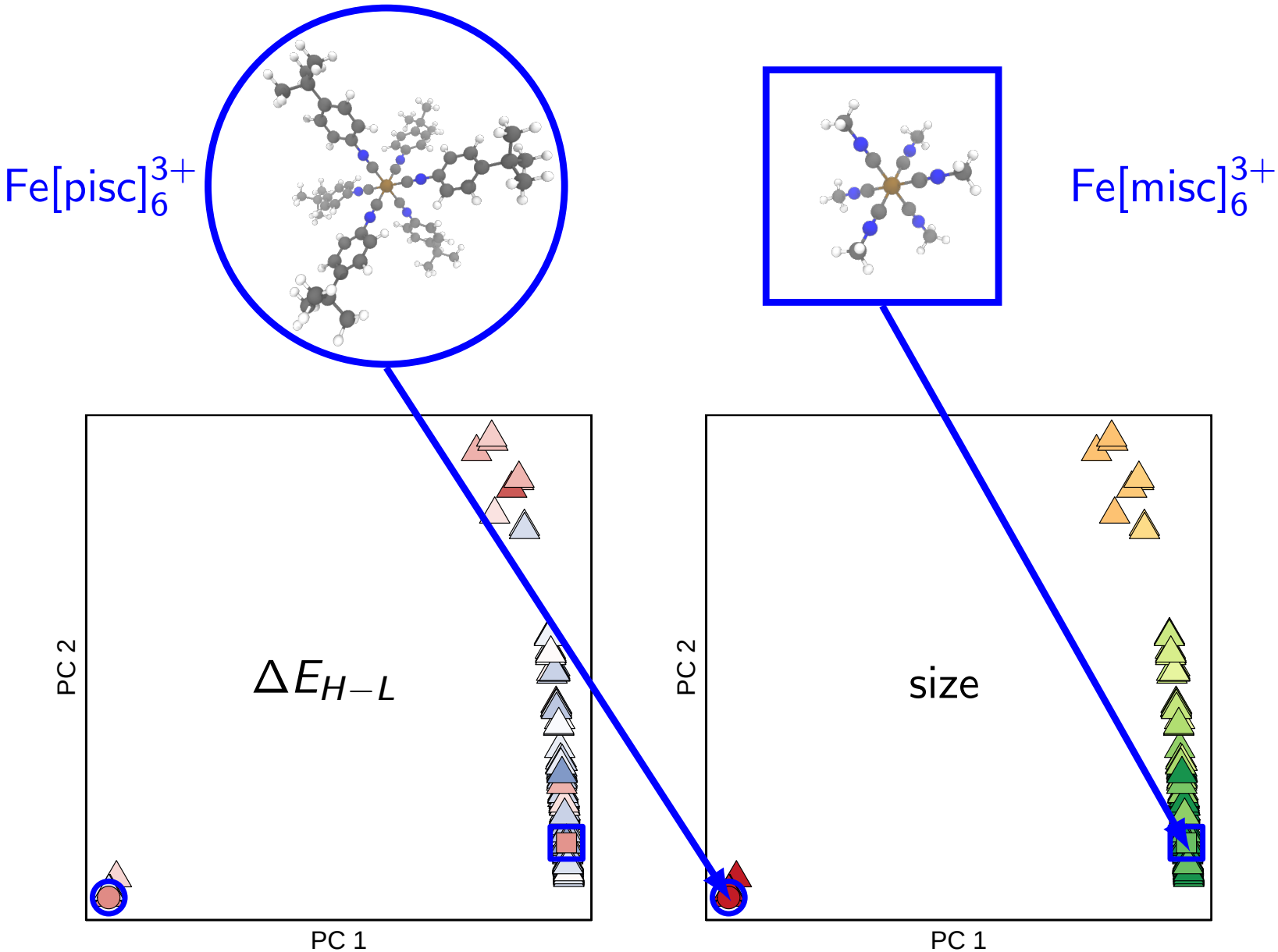


# A tale of two complexes



# A tale of two complexes



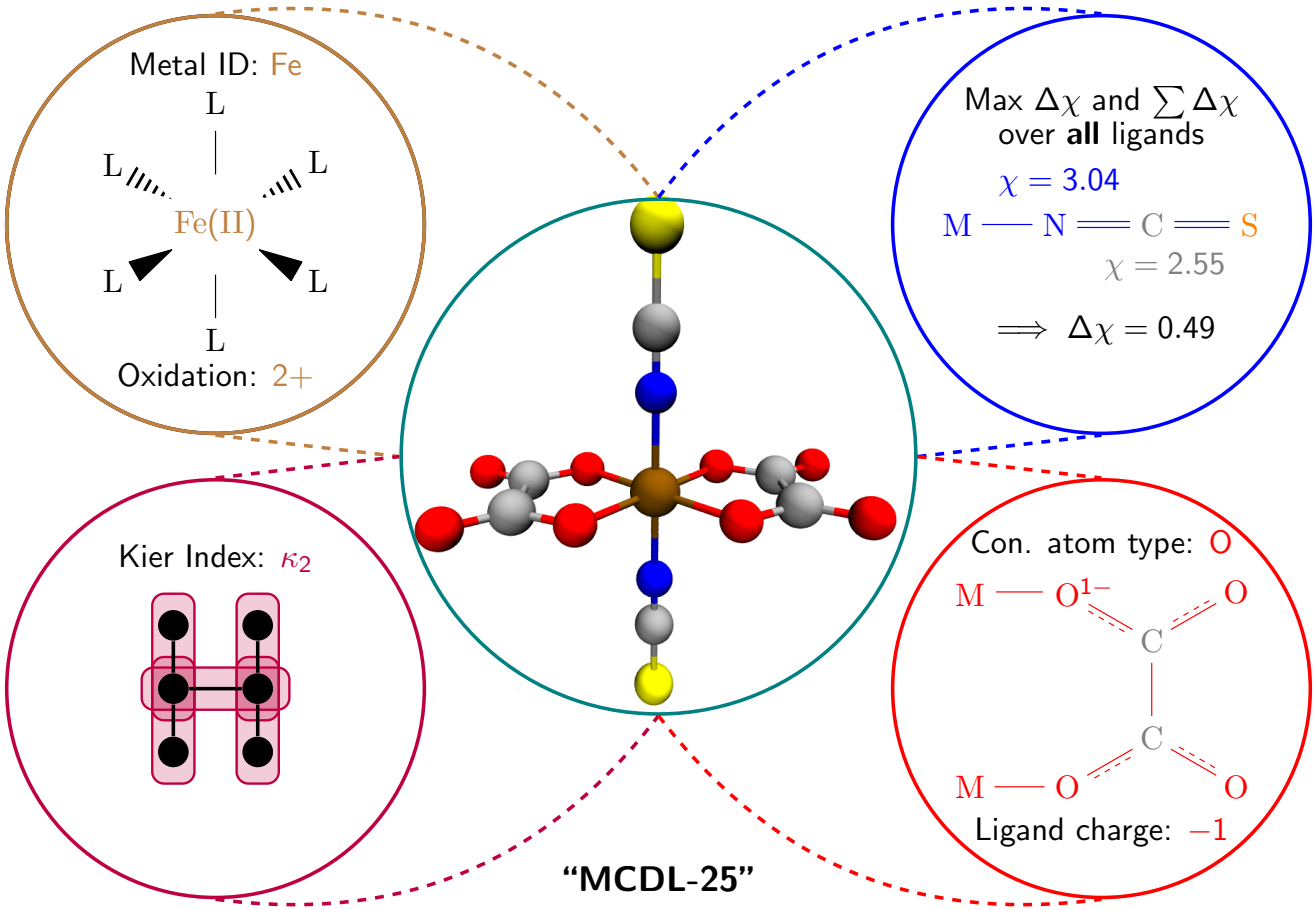
# Describing TM complexes

We need specialized descriptors for TM complexes[3]...

[3] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

# Describing TM complexes

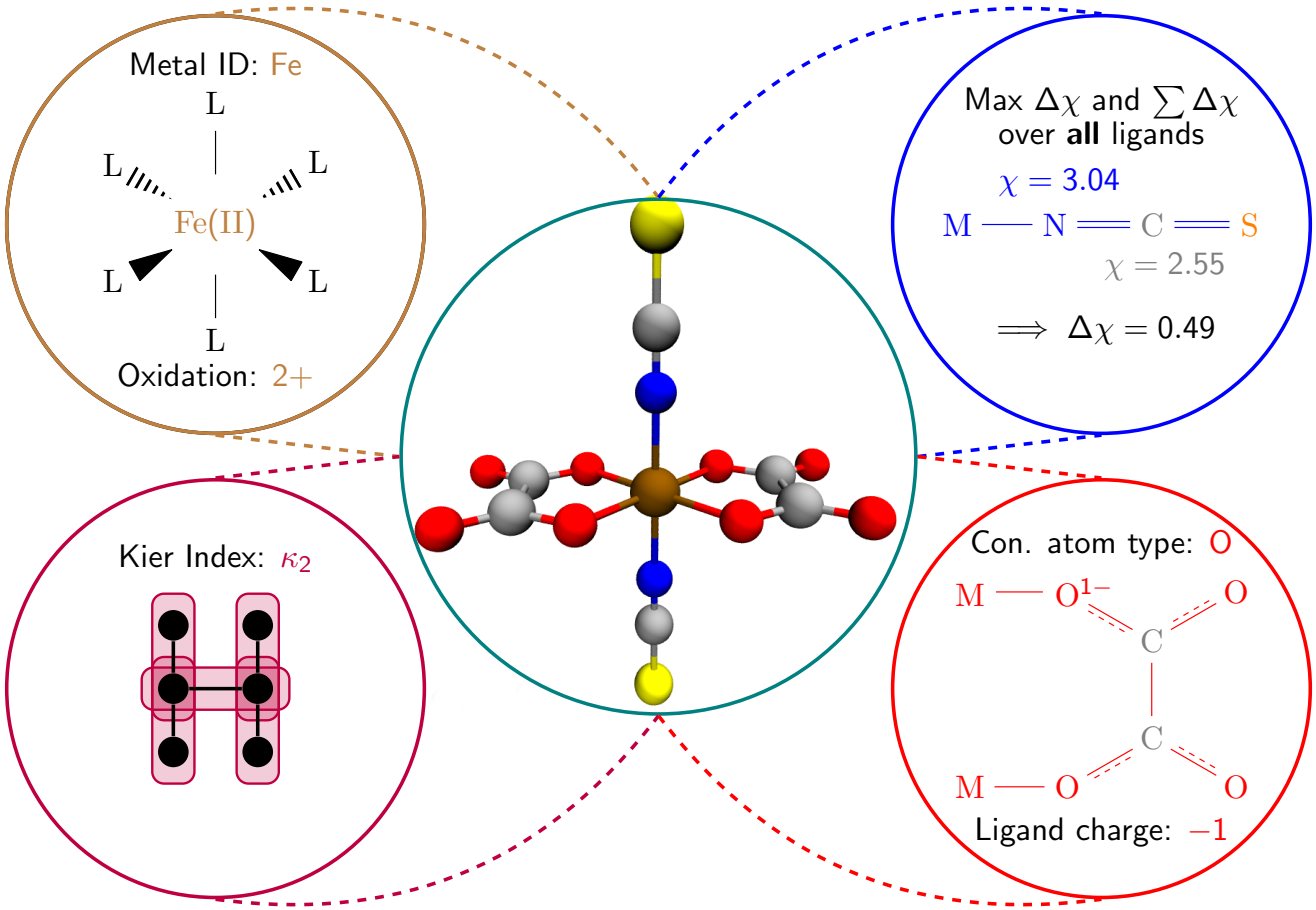
We need specialized descriptors for TM complexes[3]...




[3] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

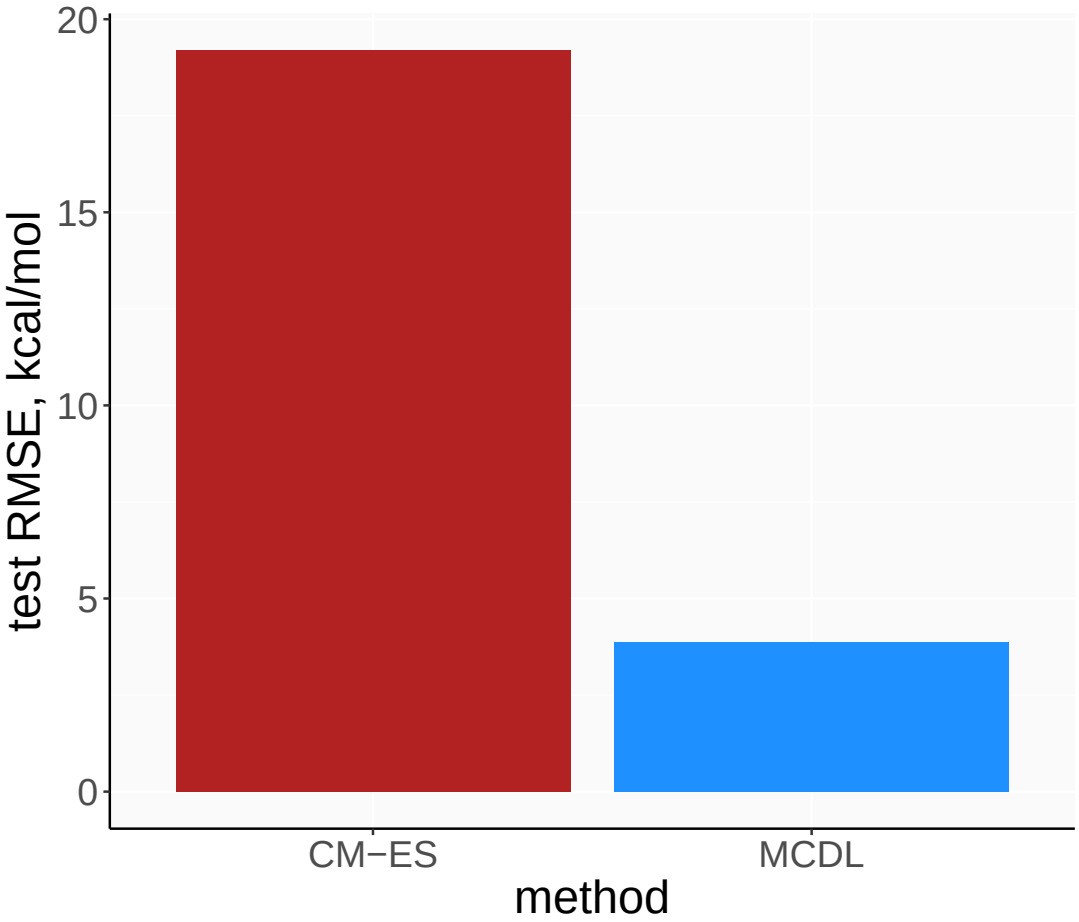
# Describing TM complexes

## MCDL: First shell effects are emphasized!



# Describing TM complexes

Greatly improves KRR performance:



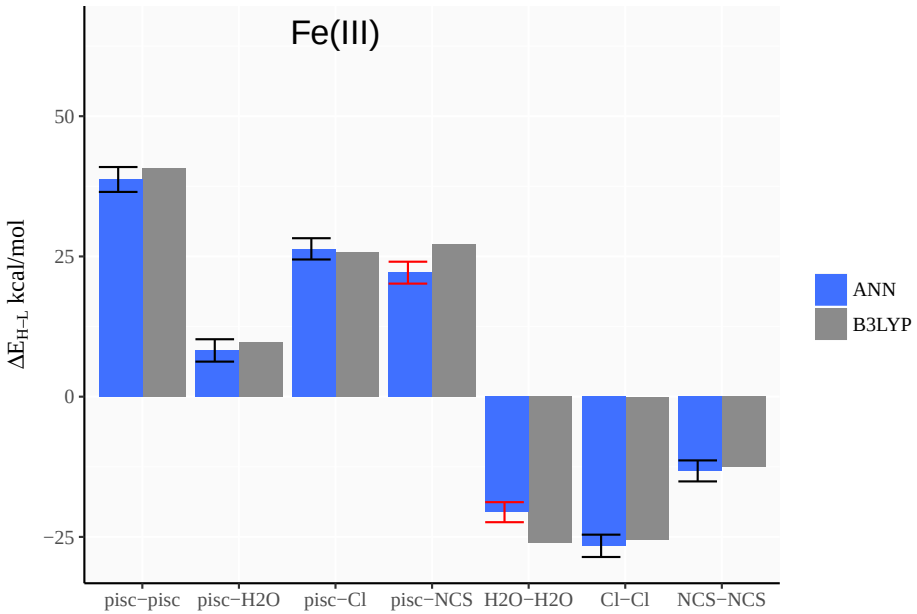
# Predictive modeling

Can use to train ANNs[3]:

[3] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

# Predictive modeling

Can use to train ANNs[3]:

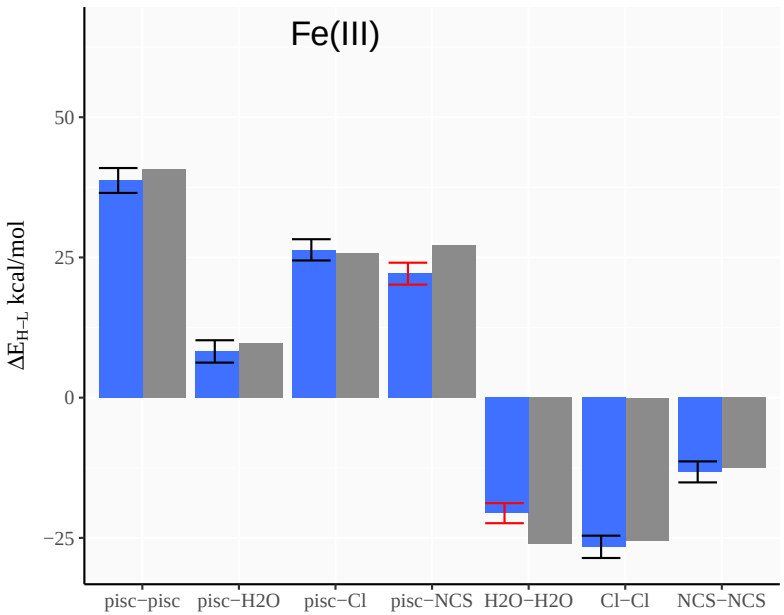



heteroleptic fields

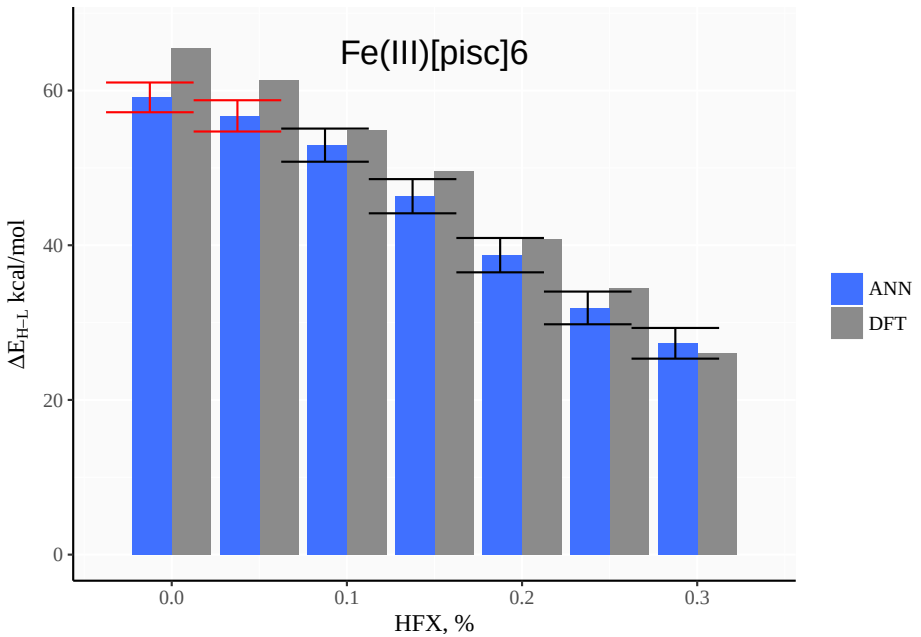
[3] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

# Predictive modeling

Can use to train ANNs[3]:



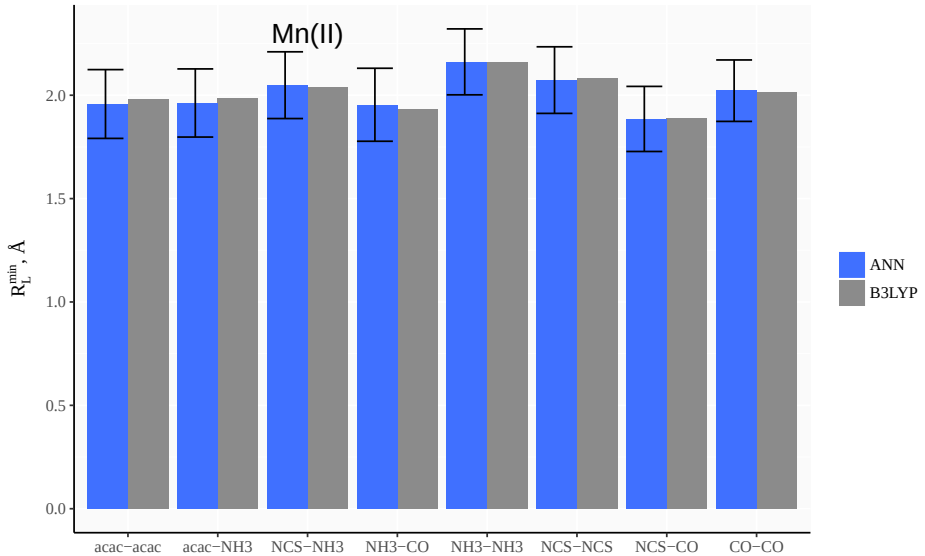

heteroleptic fields




[3] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

# Predictive modeling

Can use to train ANNs[3]:

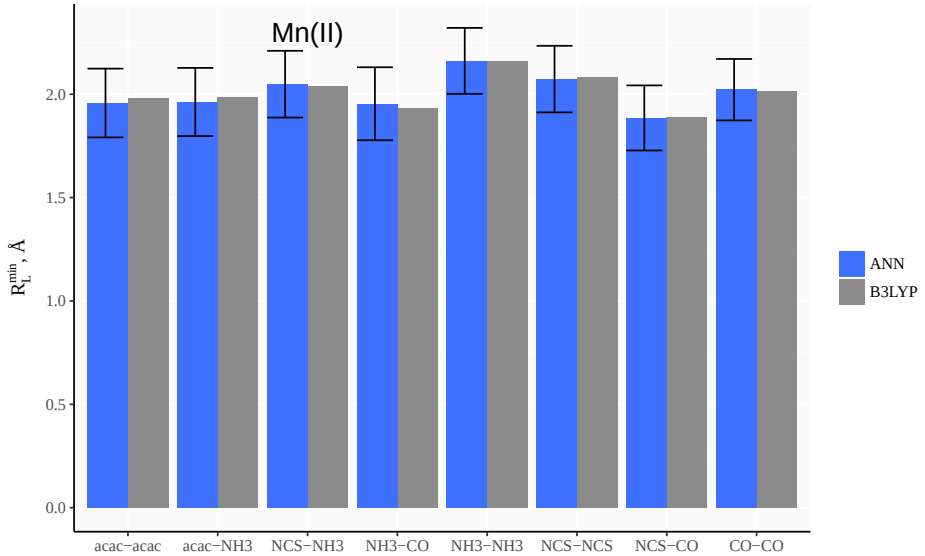



bond lenghts in a **spin** and **oxidation state** dependent manner

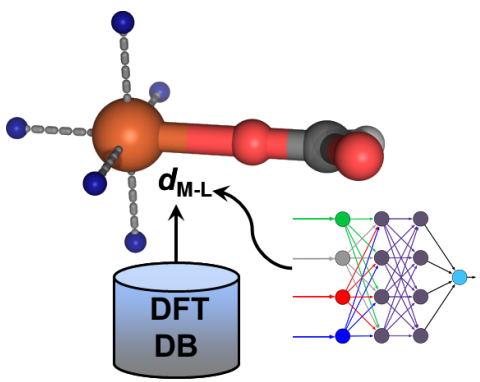
[3] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

# Predictive modeling

Can use to train ANNs[3]:



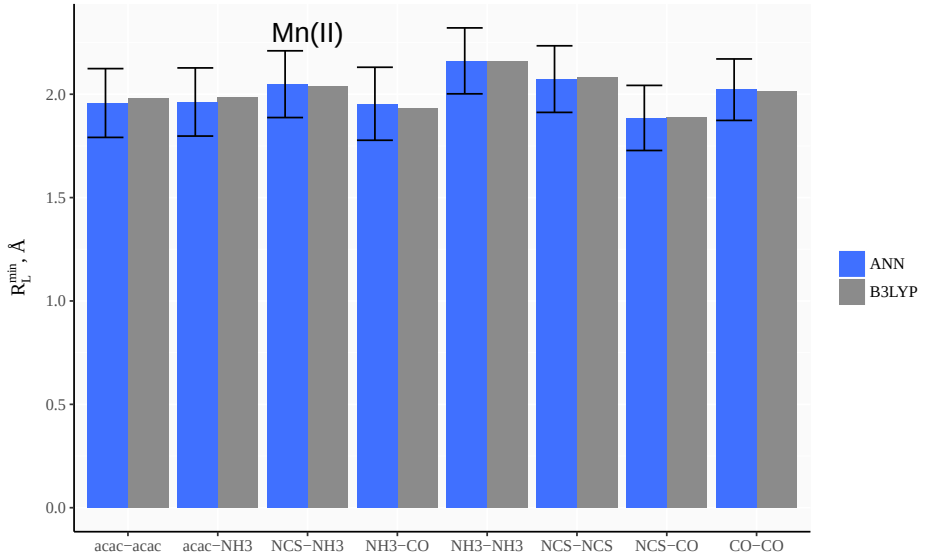

bond lenghts in a **spin** and **oxidation state** dependent manner



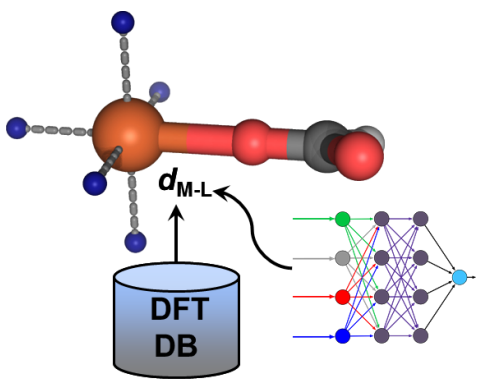

[3] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

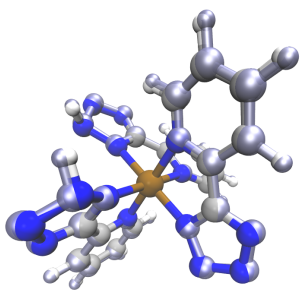
# Predictive modeling

Can use to train ANNs[3]:




bond lenghts in a **spin** and **oxidation state** dependent manner

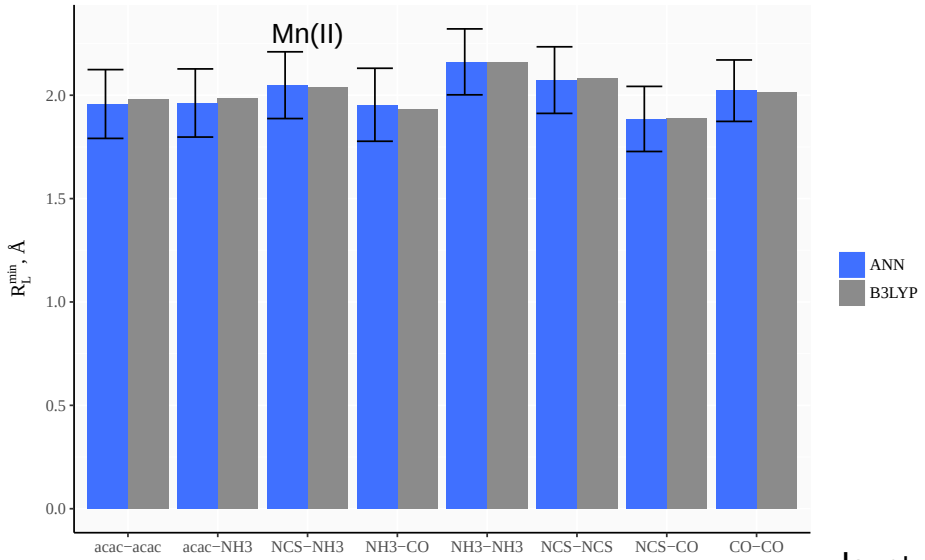


[3] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

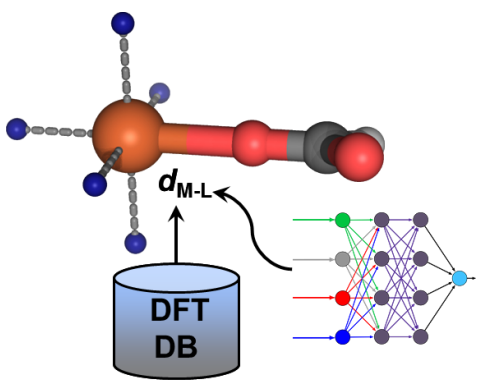
# Predictive modeling

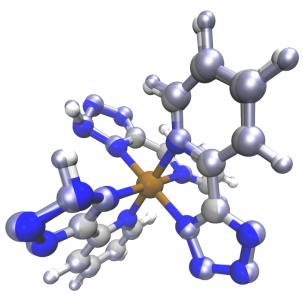
Can use to train ANNs[3]:

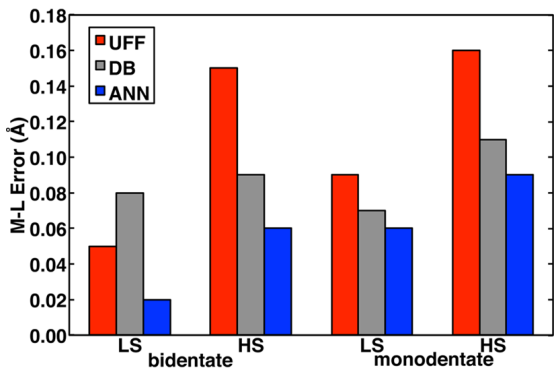



bond lenghts in a **spin** and **oxidation state** dependent manner

Janet, J.P. *et al.* *Ind. Eng. Chem. Res.* 56.17, 2017.



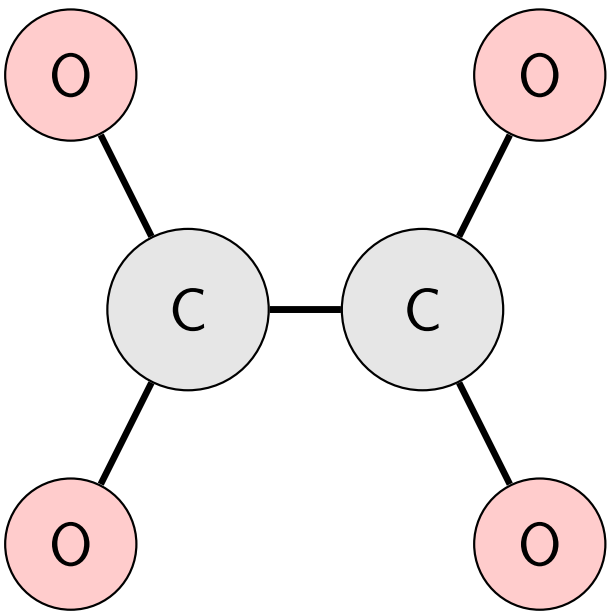






[3] Janet, J.P., and Kulik, H.J. *Chemical Science*, 2017, 8, 5137-5152.

# New descriptors

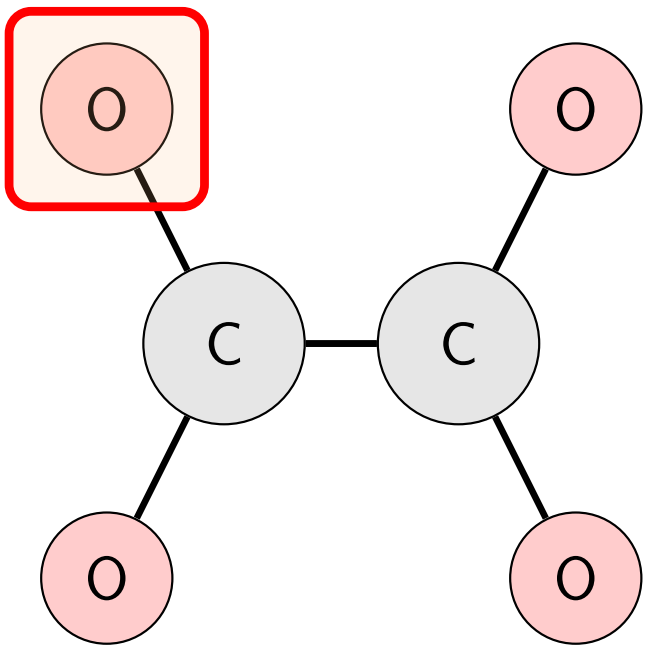
New descriptors based on autocorrelations[4]




[4] Broto, P., Moreau, G. and Vandycke, C. *Eur. J. Med. Chem.*, 19(1):71-78, 1984.

# New descriptors

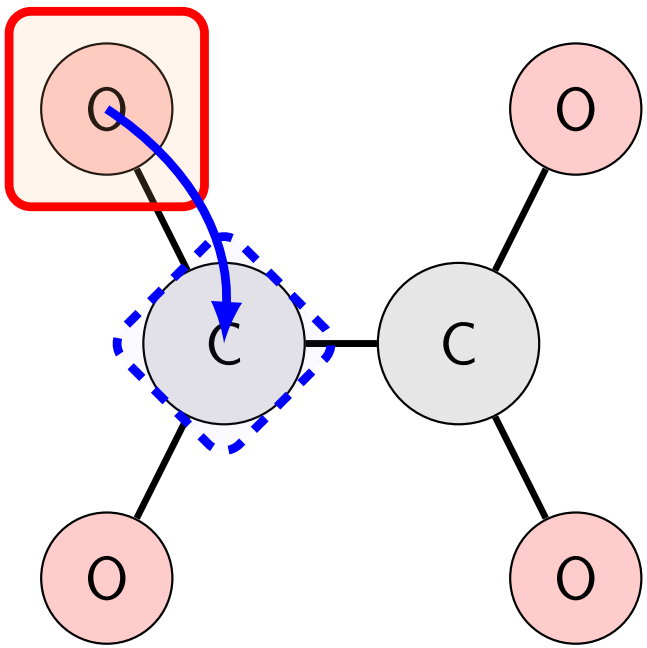
New descriptors based on autocorrelations[4]




[4] Broto, P., Moreau, G. and Vandycke, C. *Eur. J. Med. Chem.*, 19(1):71-78, 1984.

# New descriptors

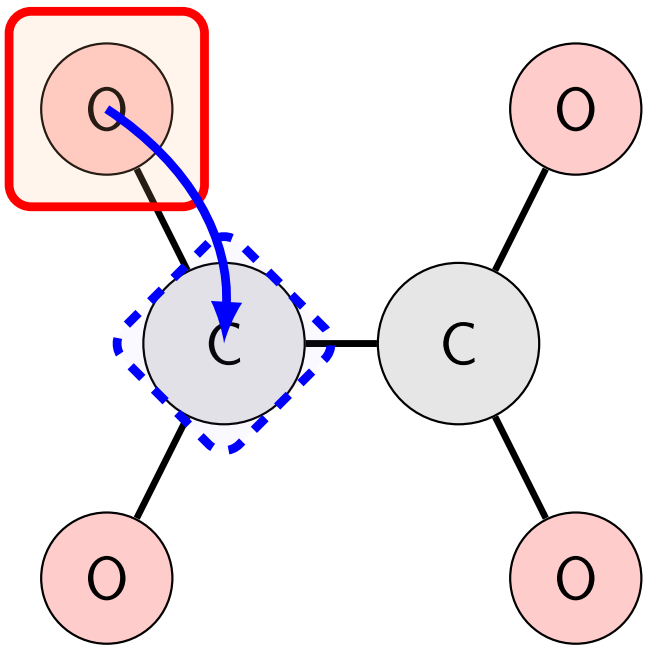
New descriptors based on autocorrelations[4]




[4] Broto, P., Moreau, G. and Vandycke, C. *Eur. J. Med. Chem.*, 19(1):71-78, 1984.

# New descriptors

New descriptors based on autocorrelations[4]

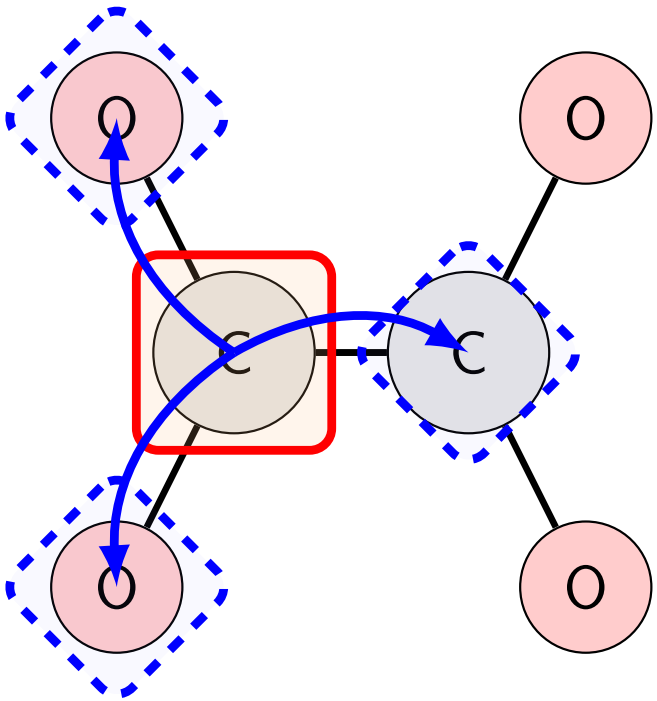



$$d_1 : \sum_{O,C} Z_O Z_C = 48$$

[4] Broto, P., Moreau, G. and Vandycke, C. *Eur. J. Med. Chem.*, 19(1):71-78, 1984.

# New descriptors

New descriptors based on autocorrelations[4]

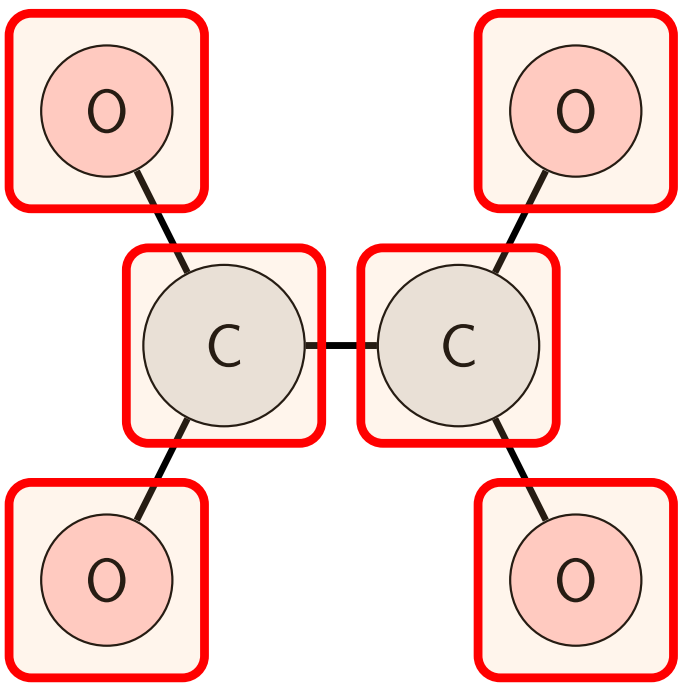



$$d_1 : 48 + \sum_{C,O} Z_O Z_C = 144 + 48$$

[4] Broto, P., Moreau, G. and Vandycke, C. *Eur. J. Med. Chem.*, 19(1):71-78, 1984.

# New descriptors

New descriptors based on autocorrelations[4]

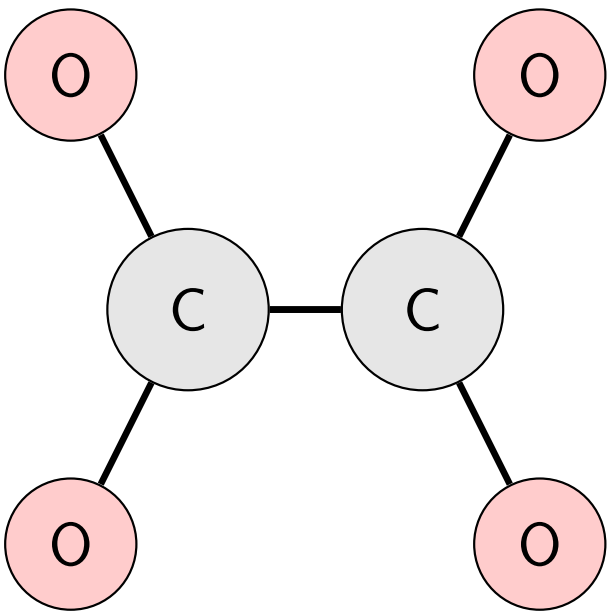



$$d_1 : \sum_i \sum_j Z_i Z_j \delta(d_{i,j}, 1)$$

[4] Broto, P., Moreau, G. and Vandycke, C. *Eur. J. Med. Chem.*, 19(1):71-78, 1984.

# New descriptors

New descriptors based on autocorrelations[4]

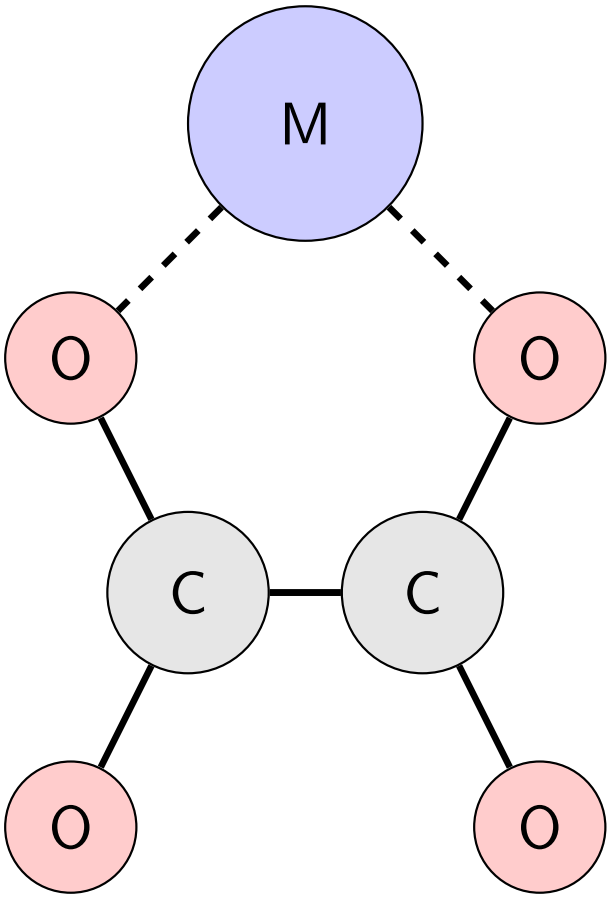



$$d_x : \sum_i \sum_j Z_i Z_j \delta(d_{ij}, x)$$

[4] Broto, P., Moreau, G. and Vandycke, C. *Eur. J. Med. Chem.*, 19(1):71-78, 1984.

# New descriptors

New descriptors based on autocorrelations[4]

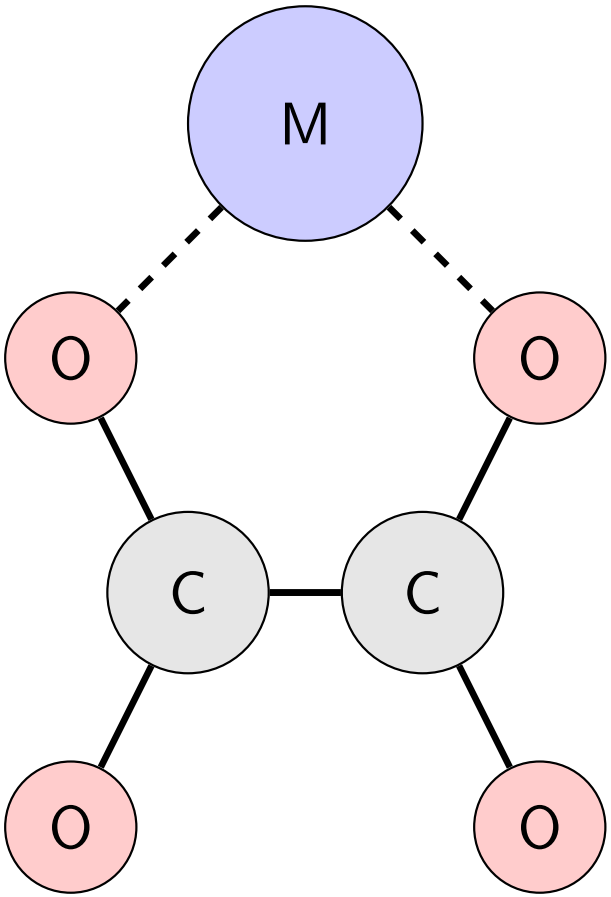



How to adapt to TM complexes?

[4] Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# New descriptors

New descriptors based on autocorrelations[4]

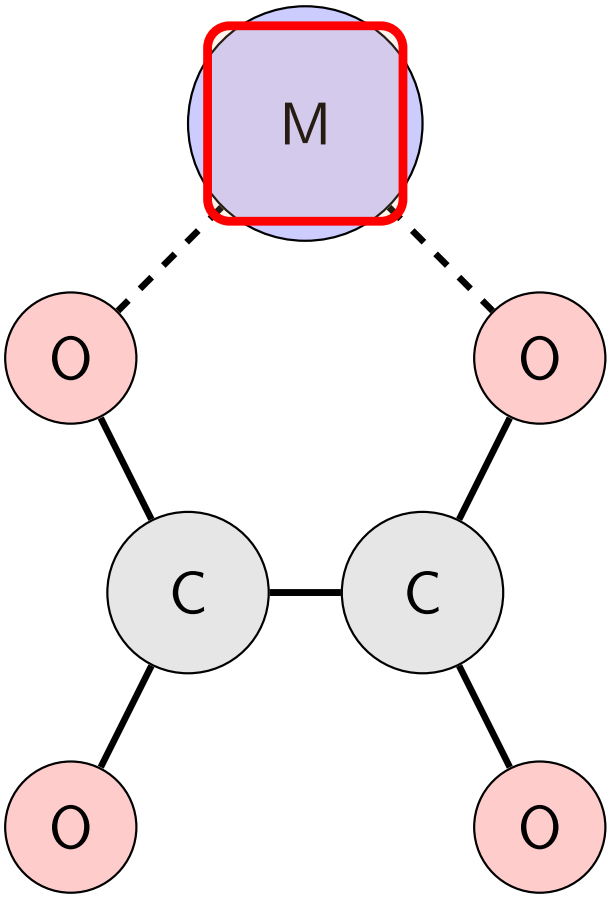



How to adapt to TM complexes?
restrict the scope to focus on
*near-metal atoms*

[4] Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# New descriptors

New descriptors based on autocorrelations[4]

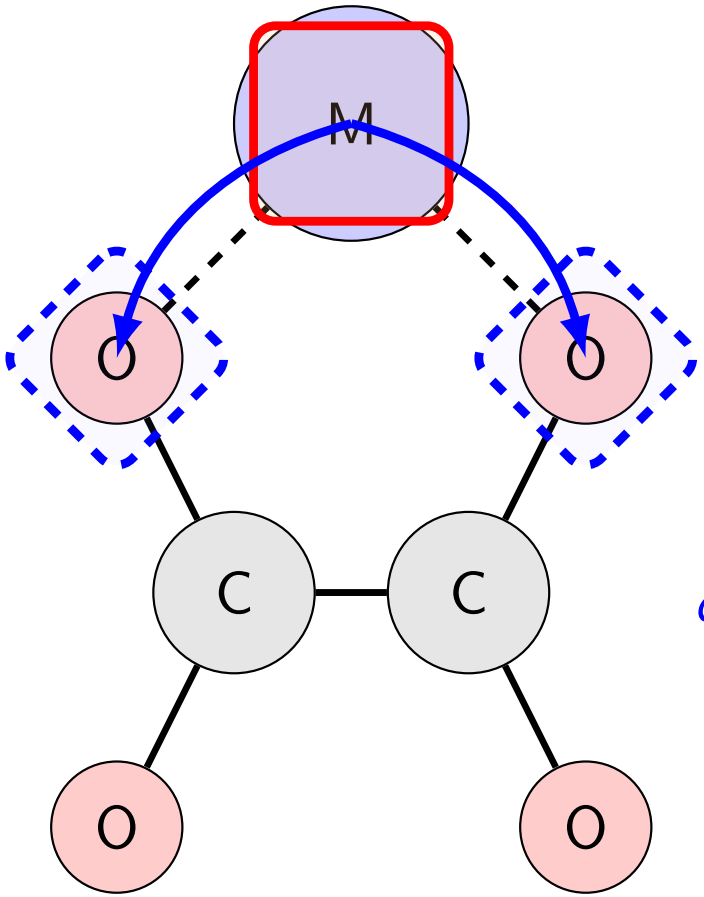



How to adapt to TM complexes?
restrict the scope to focus on
*near-metal atoms*

[4] Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# New descriptors

New descriptors based on autocorrelations[4]




How to adapt to TM complexes?

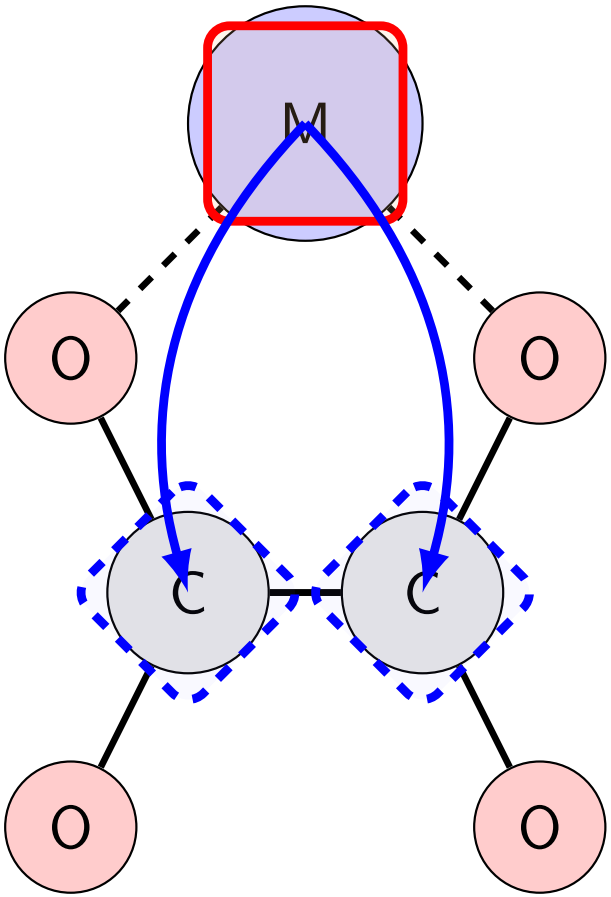
restrict the scope to focus on *near-metal atoms*

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

[4] Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# New descriptors

New descriptors based on autocorrelations[4]




How to adapt to TM complexes?

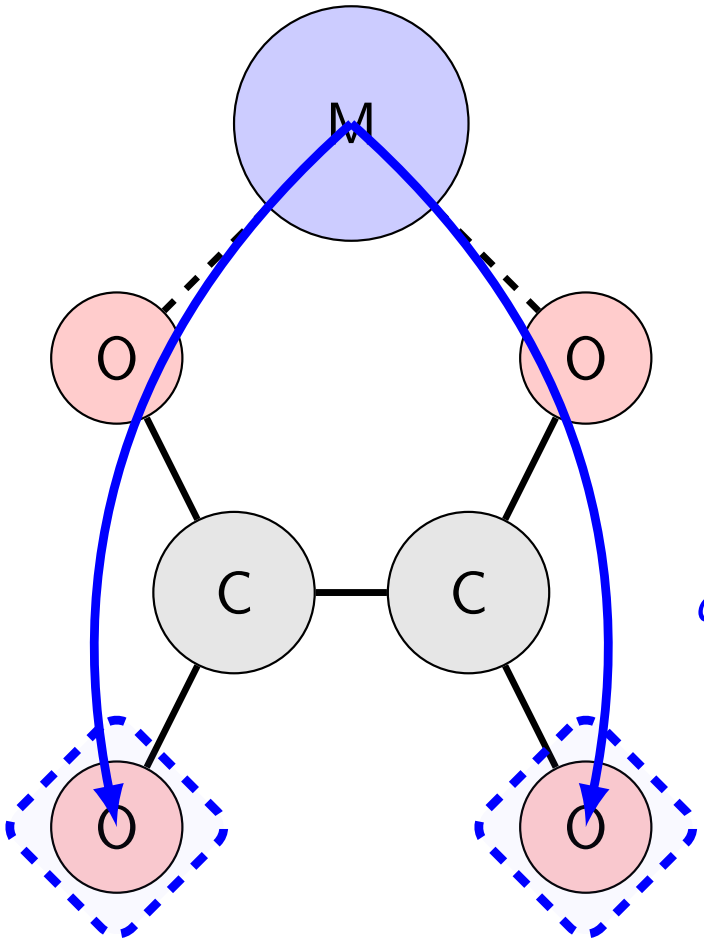
restrict the scope to focus on *near-metal atoms*

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

[4] Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# New descriptors

New descriptors based on autocorrelations[4]




How to adapt to TM complexes?

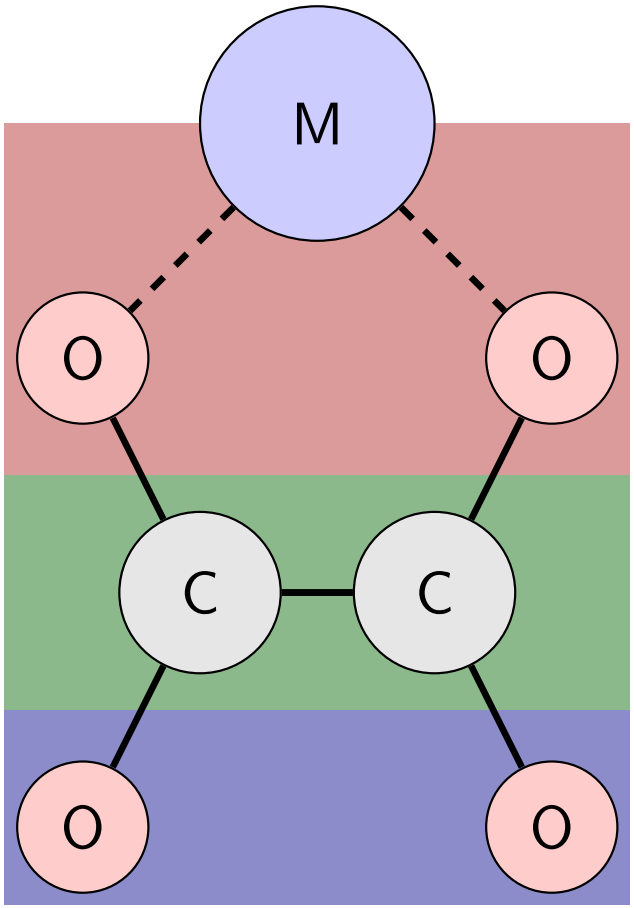
restrict the scope to focus on *near-metal atoms*

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

[4] Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# New descriptors

New descriptors based on autocorrelations[4]




How to adapt to TM complexes?

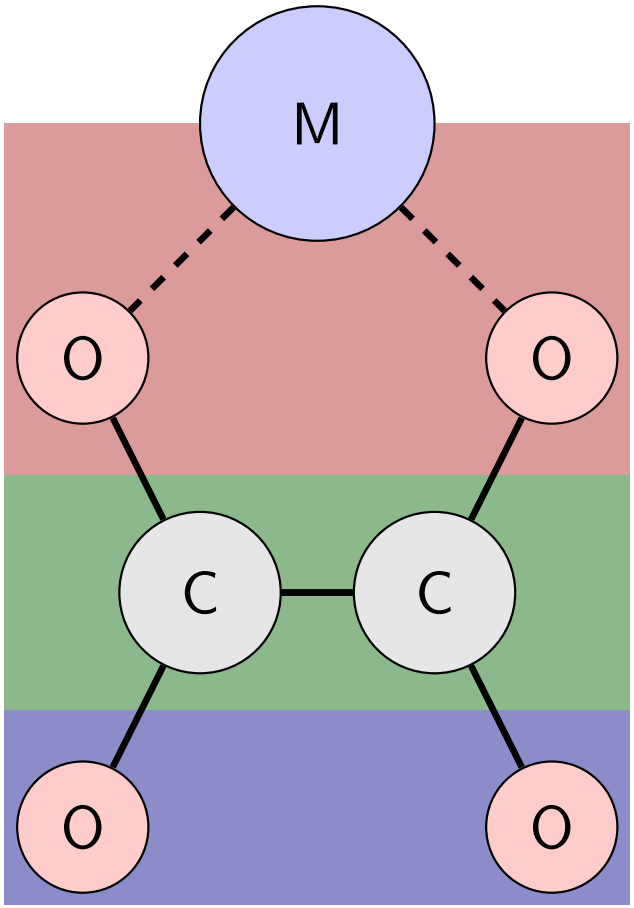
restrict the scope to focus on *near-metal atoms*

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

[4] Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# New descriptors

New descriptors based on autocorrelations[4]




How to adapt to TM complexes?

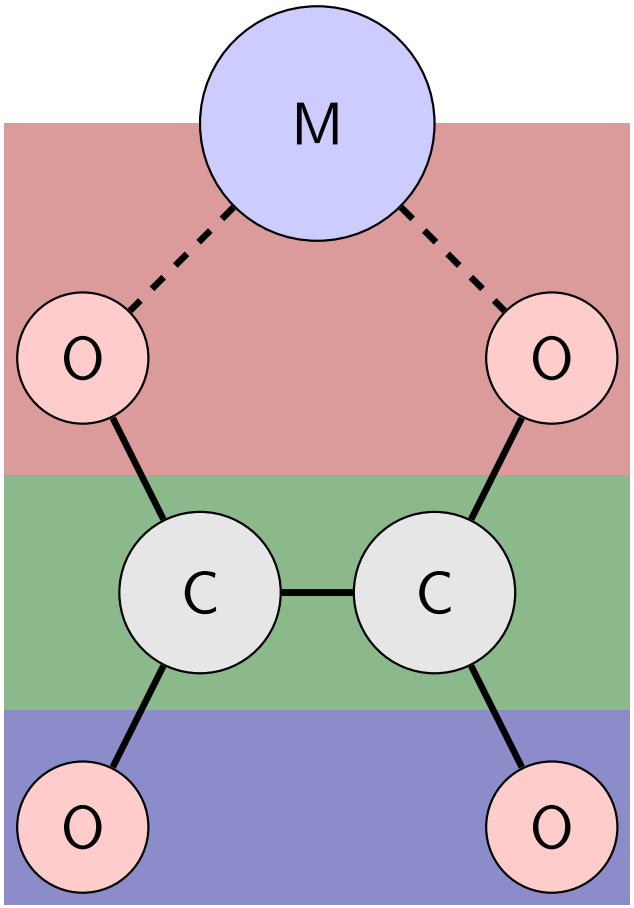
restrict the scope to focus on *near-metal atoms*

$$d_3 : \sum_{M,O} Z_M Z_O \, (Z_i - Z_j)$$

[4] Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# New descriptors

New descriptors based on autocorrelations[4]




How to adapt to TM complexes?

restrict the scope to focus on *near-metal atoms*

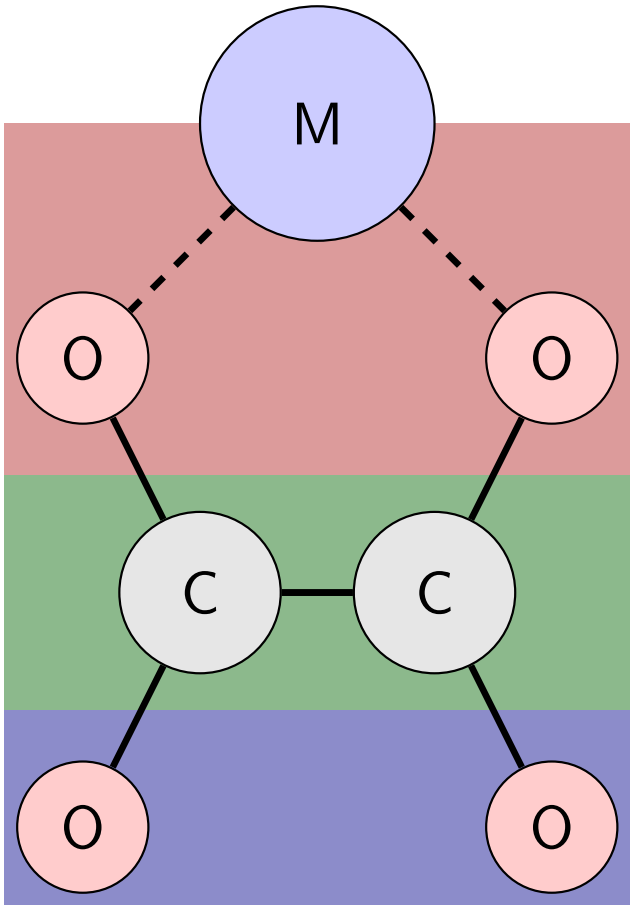
$$d_3 : \sum_{M,O} Z_M Z_O \, (Z_i - Z_j)$$

properties: $T, \chi, Z, I, S$

[4] Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# New descriptors

New descriptors based on autocorrelations[4]




How to adapt to TM complexes?

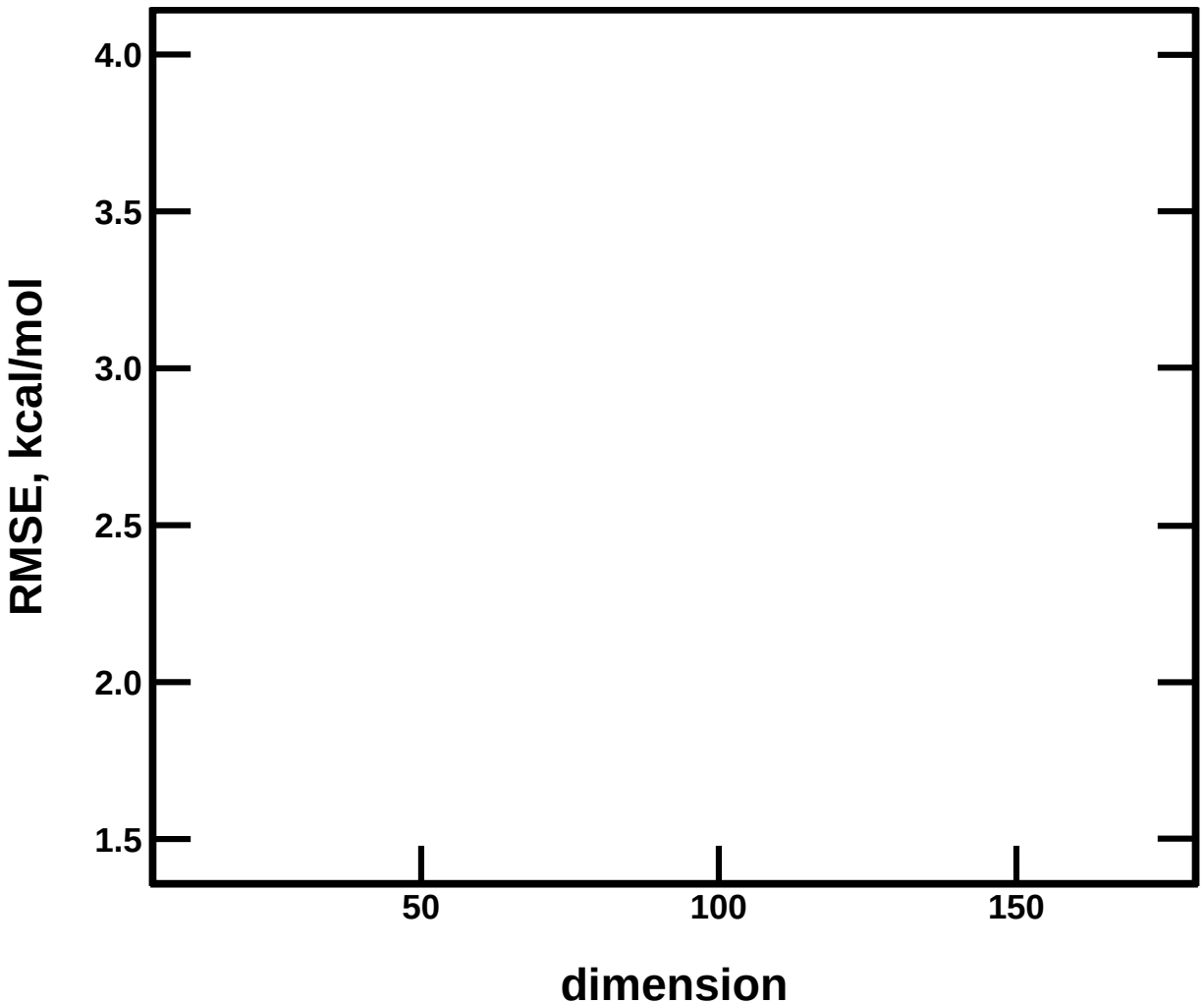
restrict the scope to focus on *near-metal atoms*

$$d_3 : \sum_{M,O} Z_M Z_O \, (Z_i - Z_j)$$

$\sim$ 160 features in total

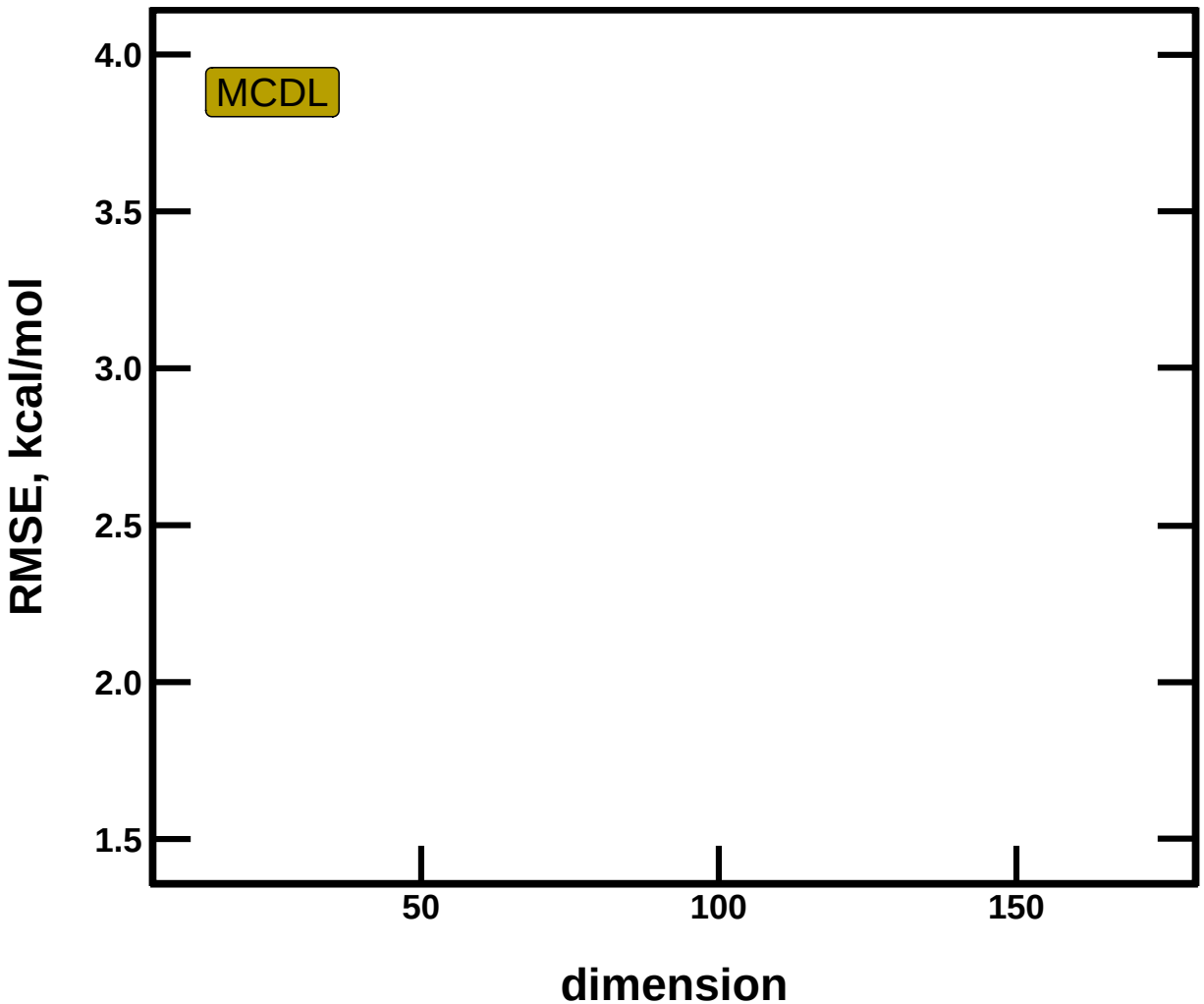
[4] Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



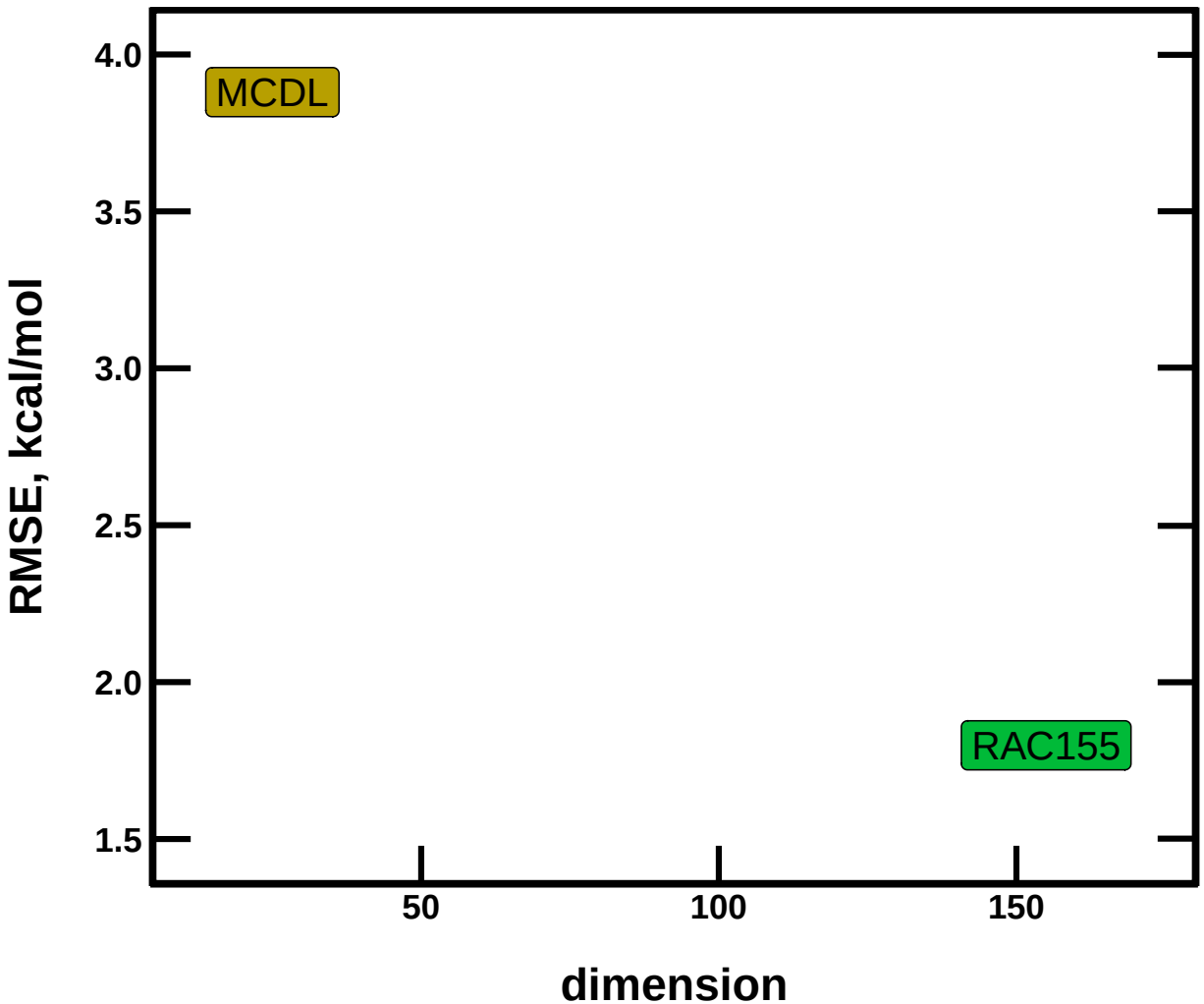

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



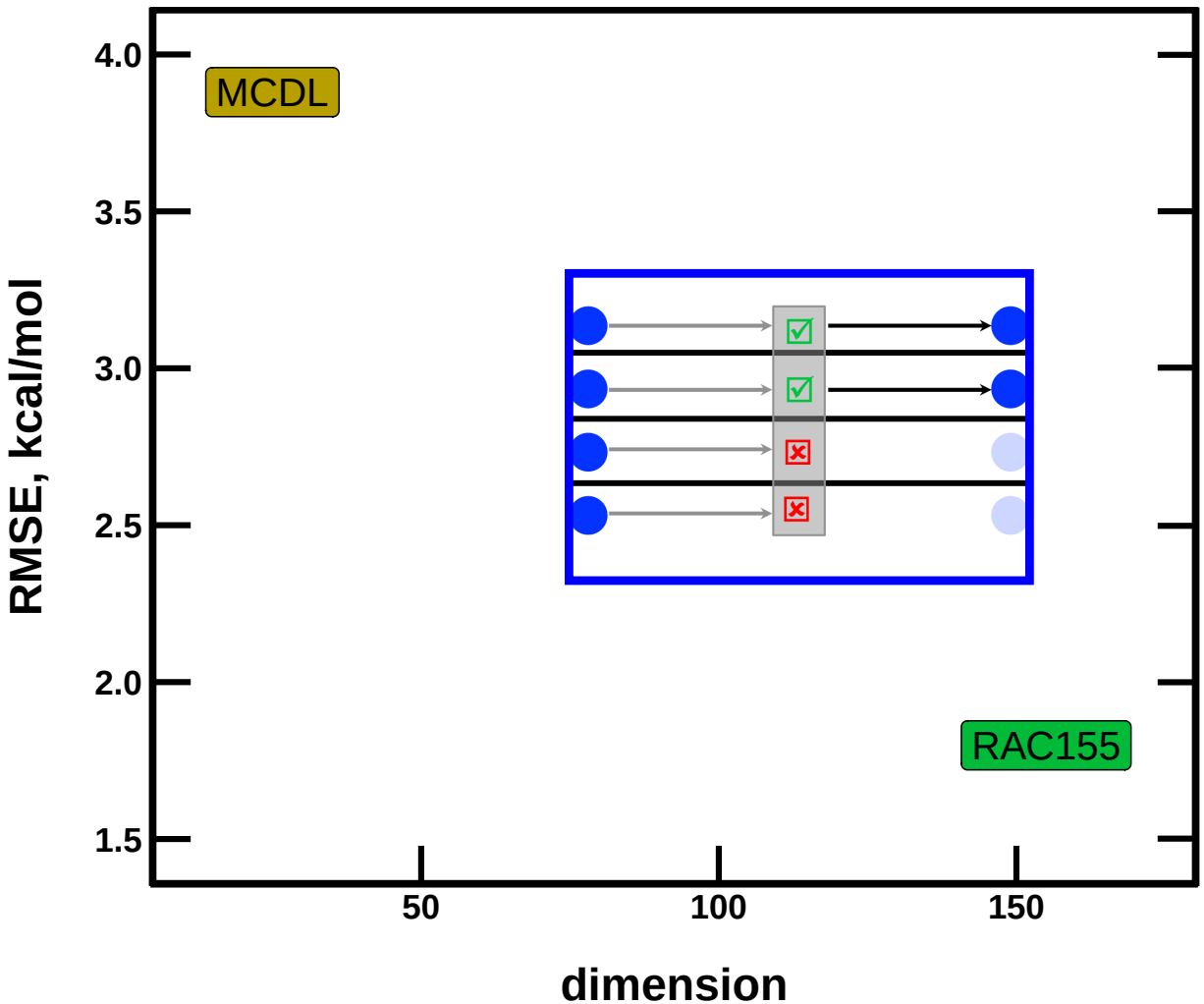

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



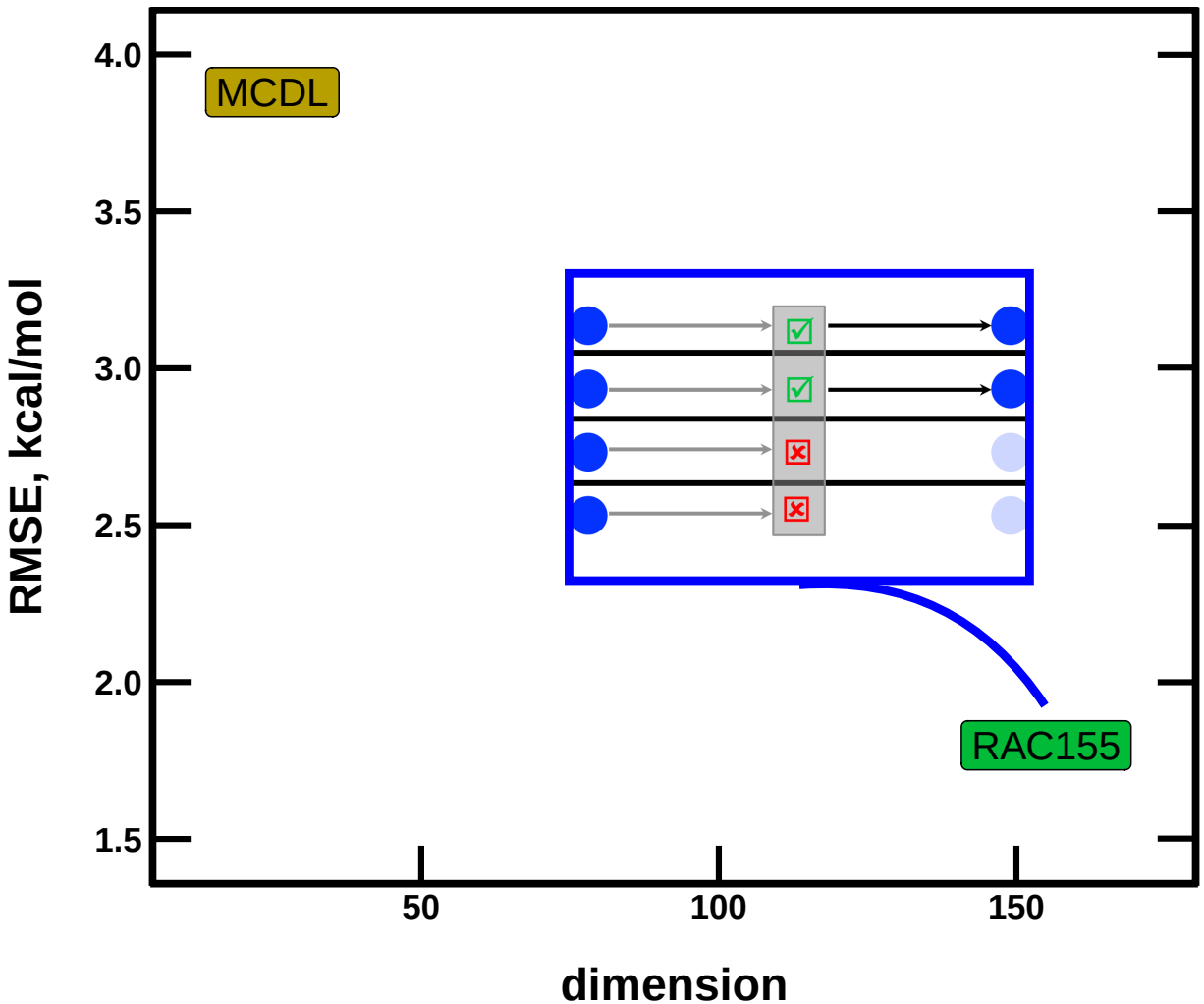

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



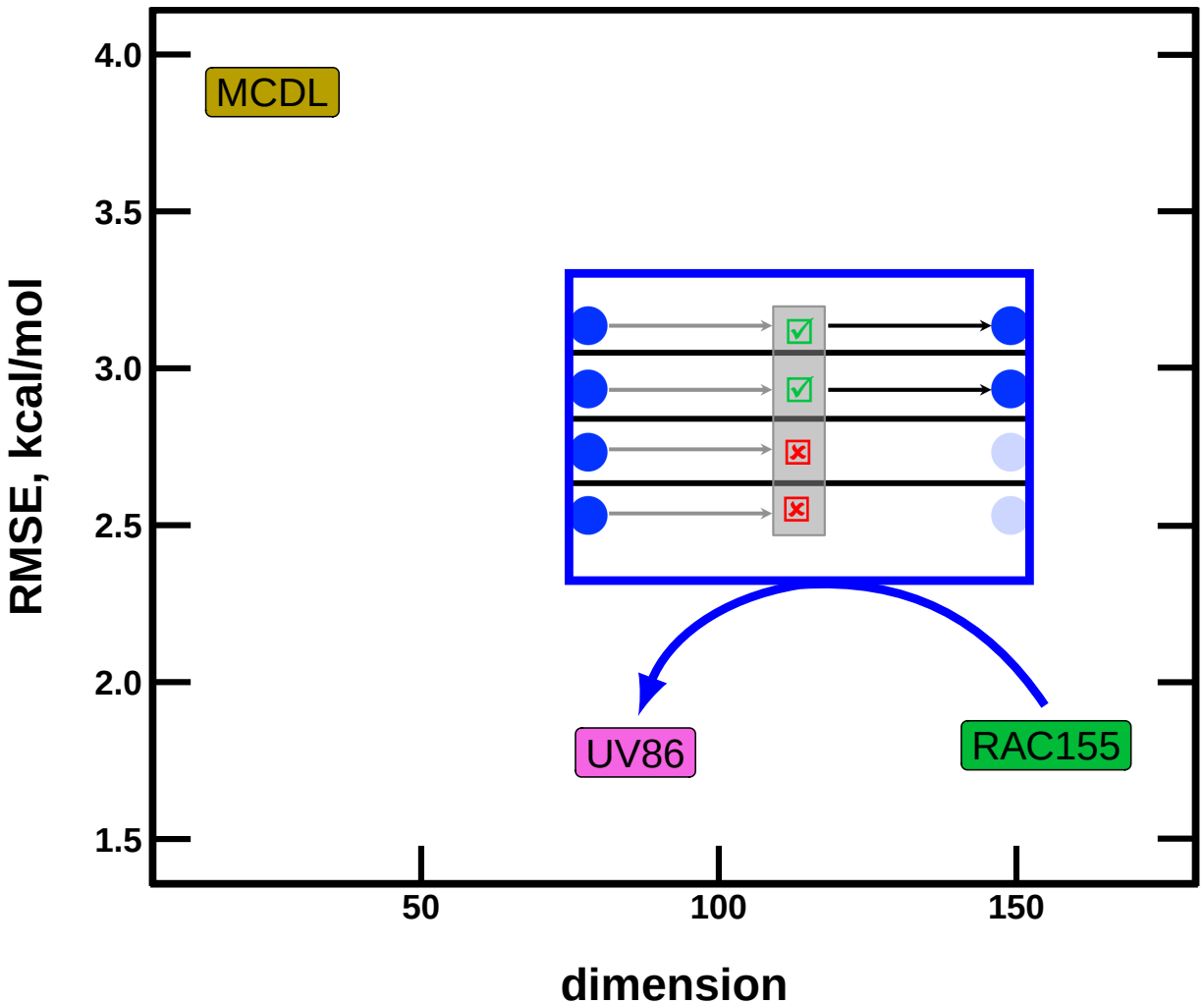

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



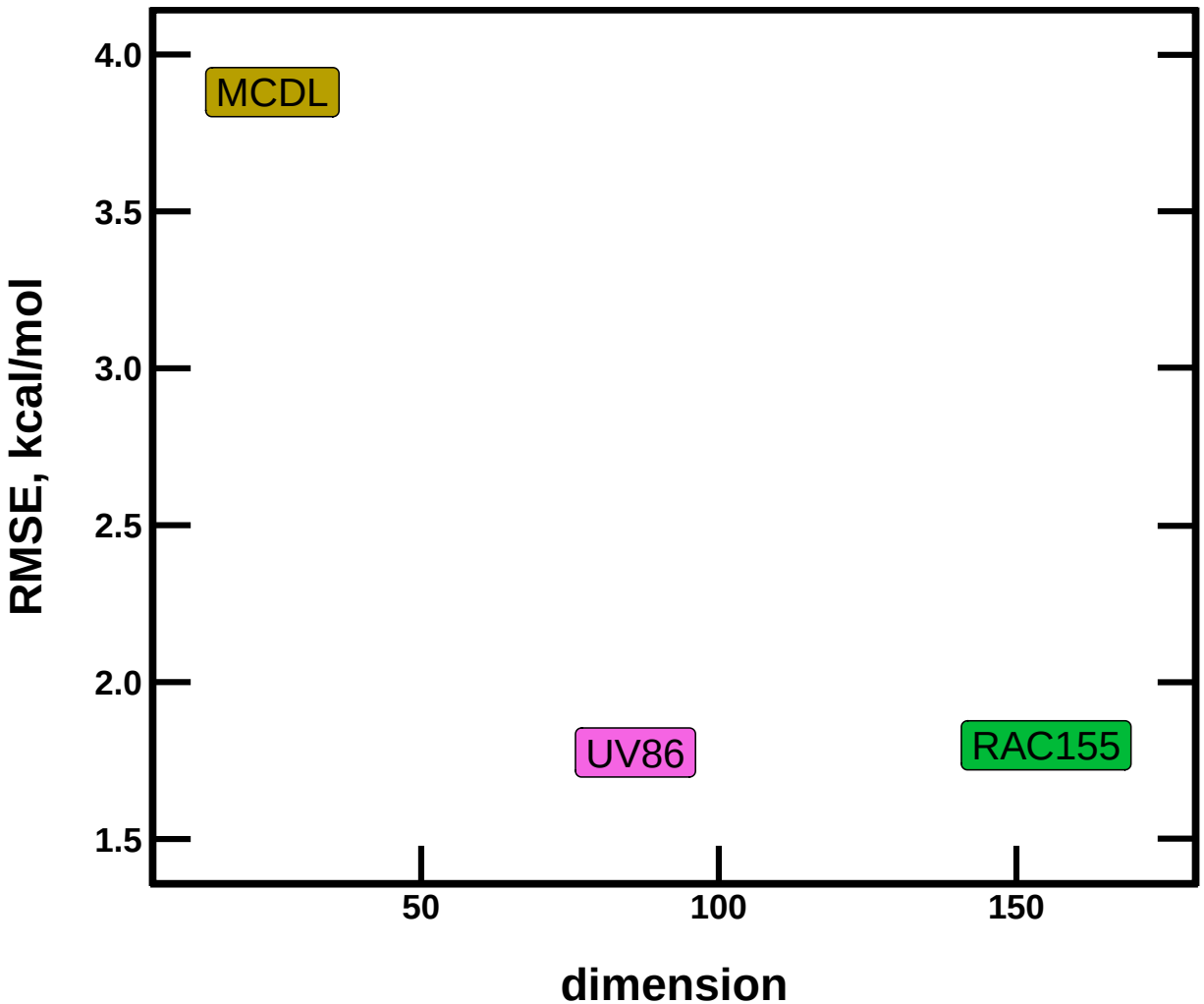

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



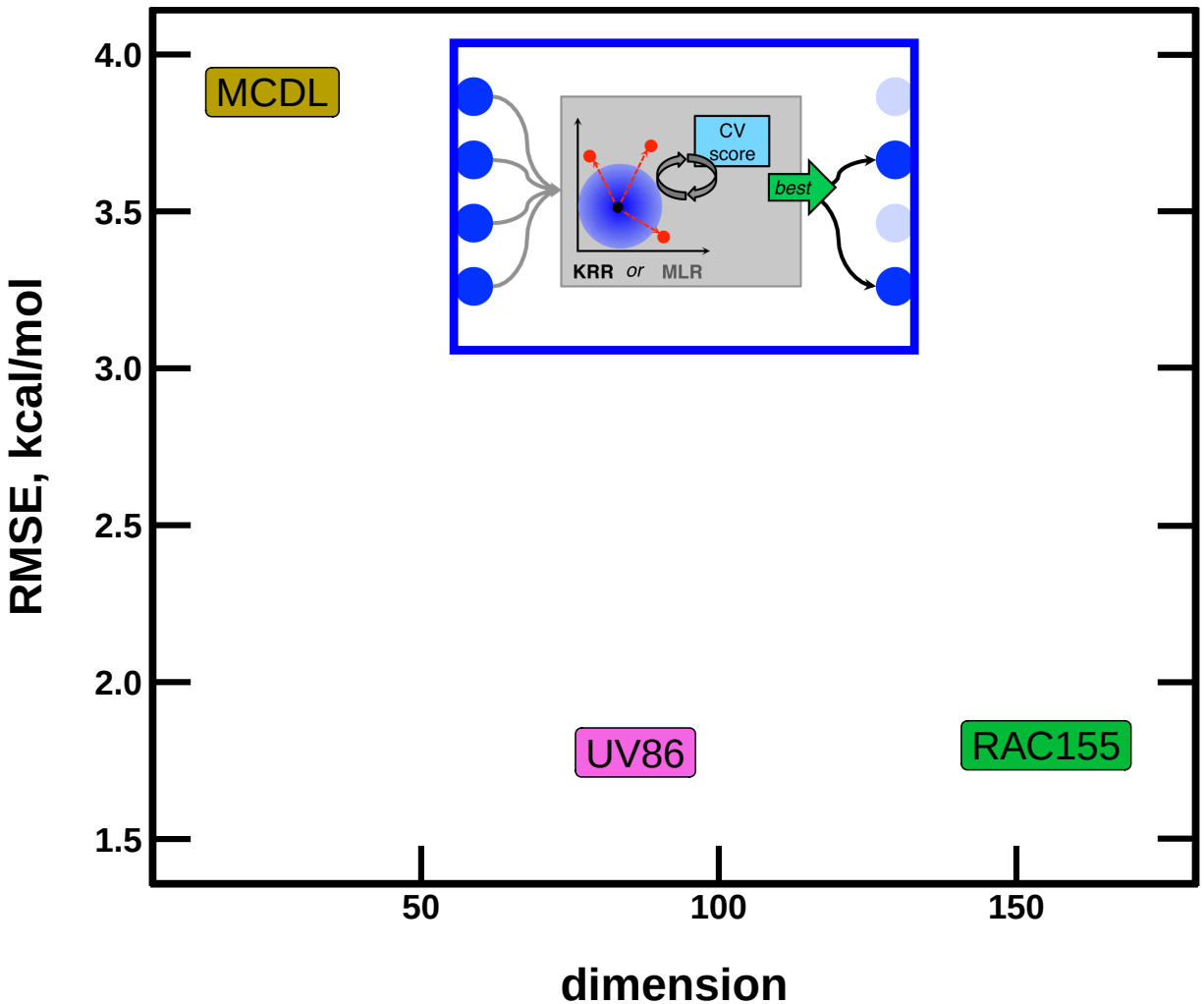

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



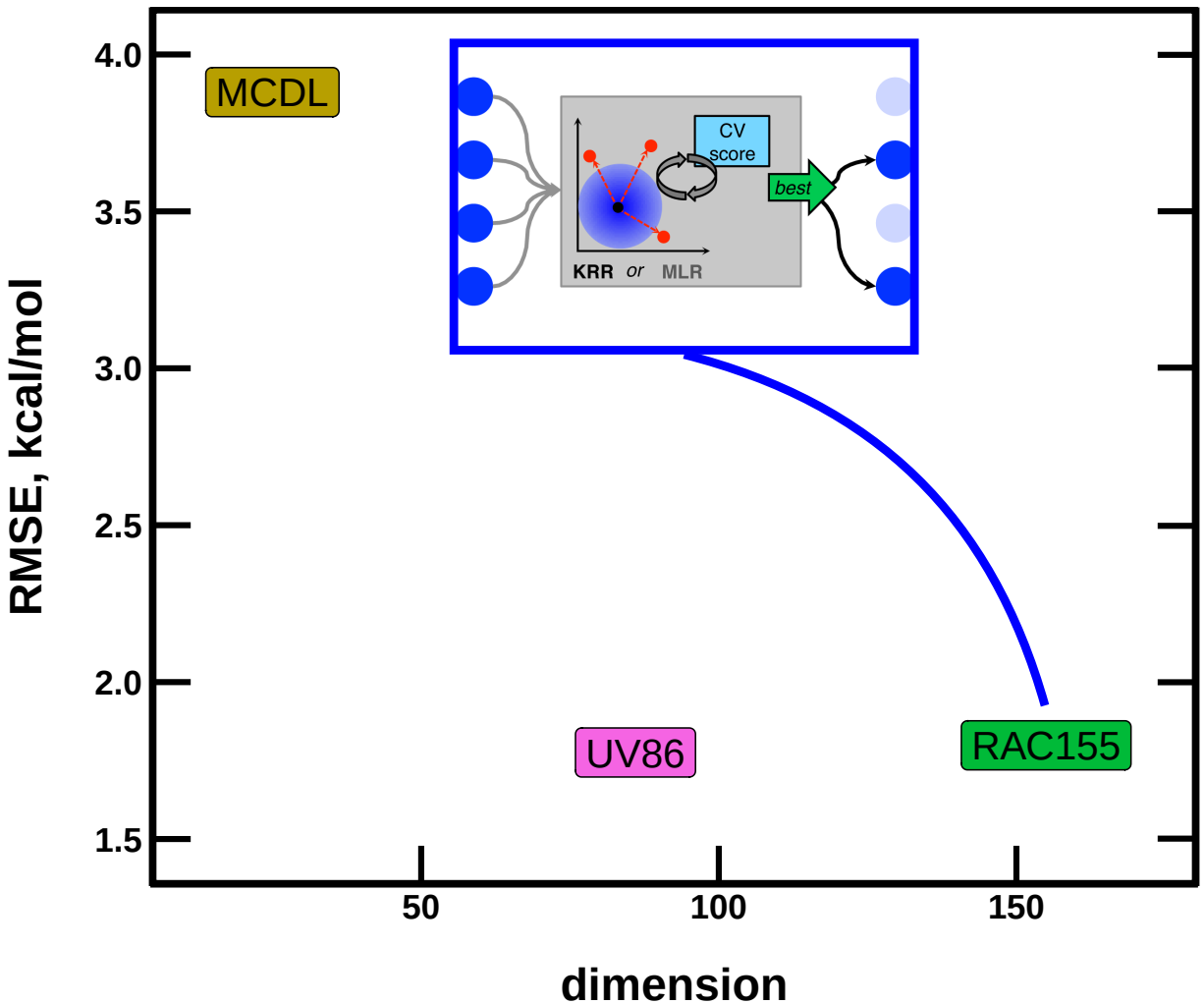

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



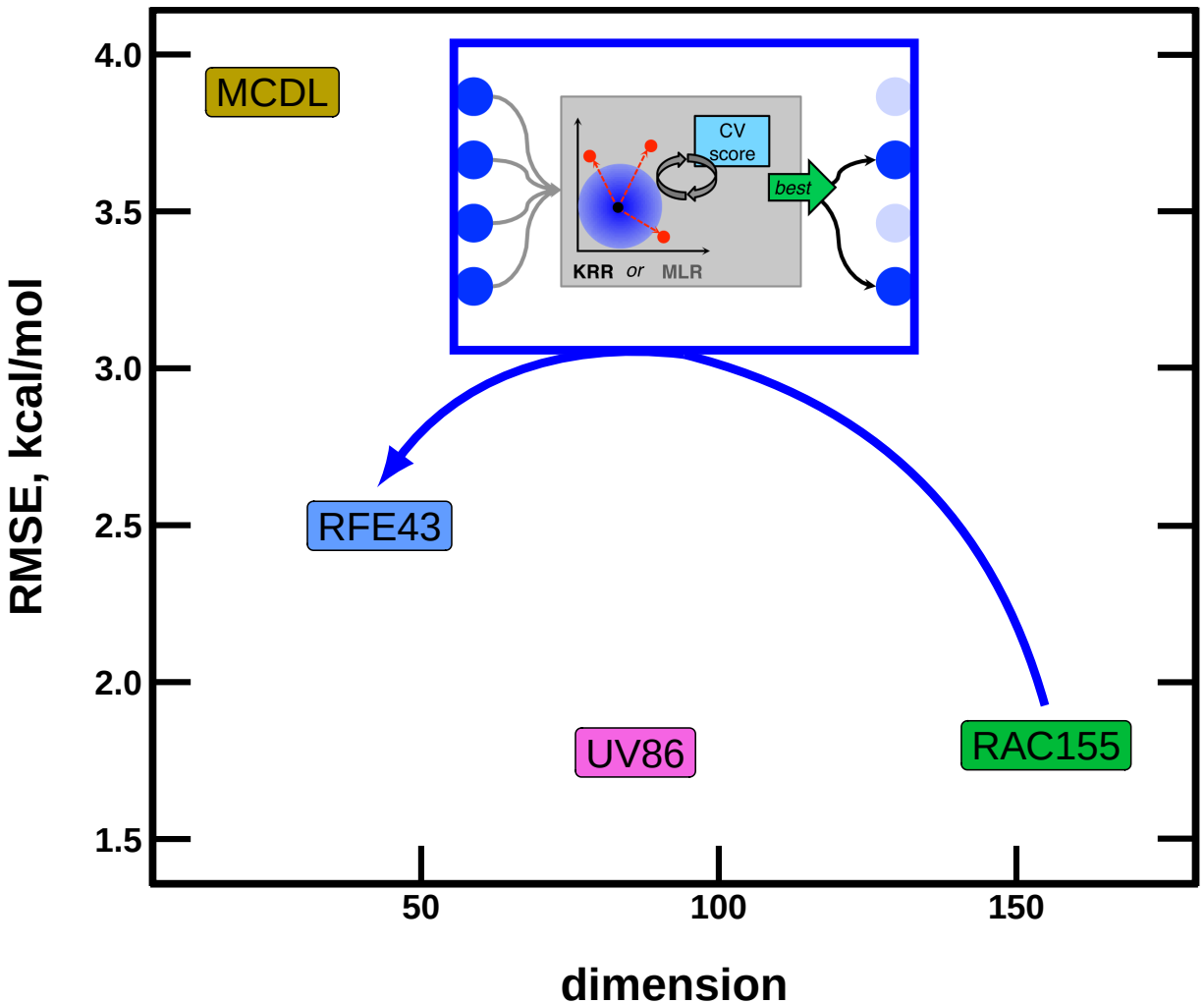

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



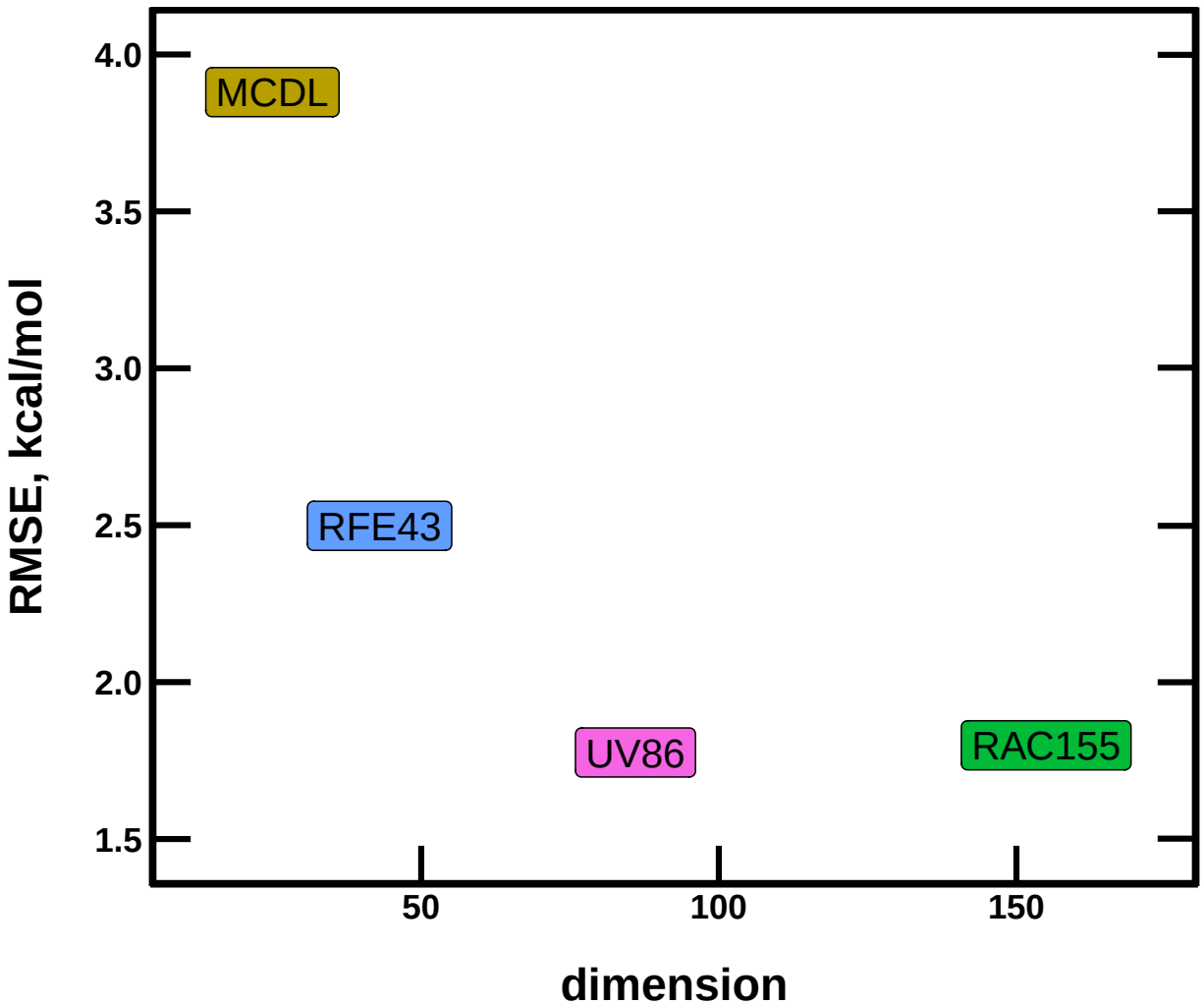

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



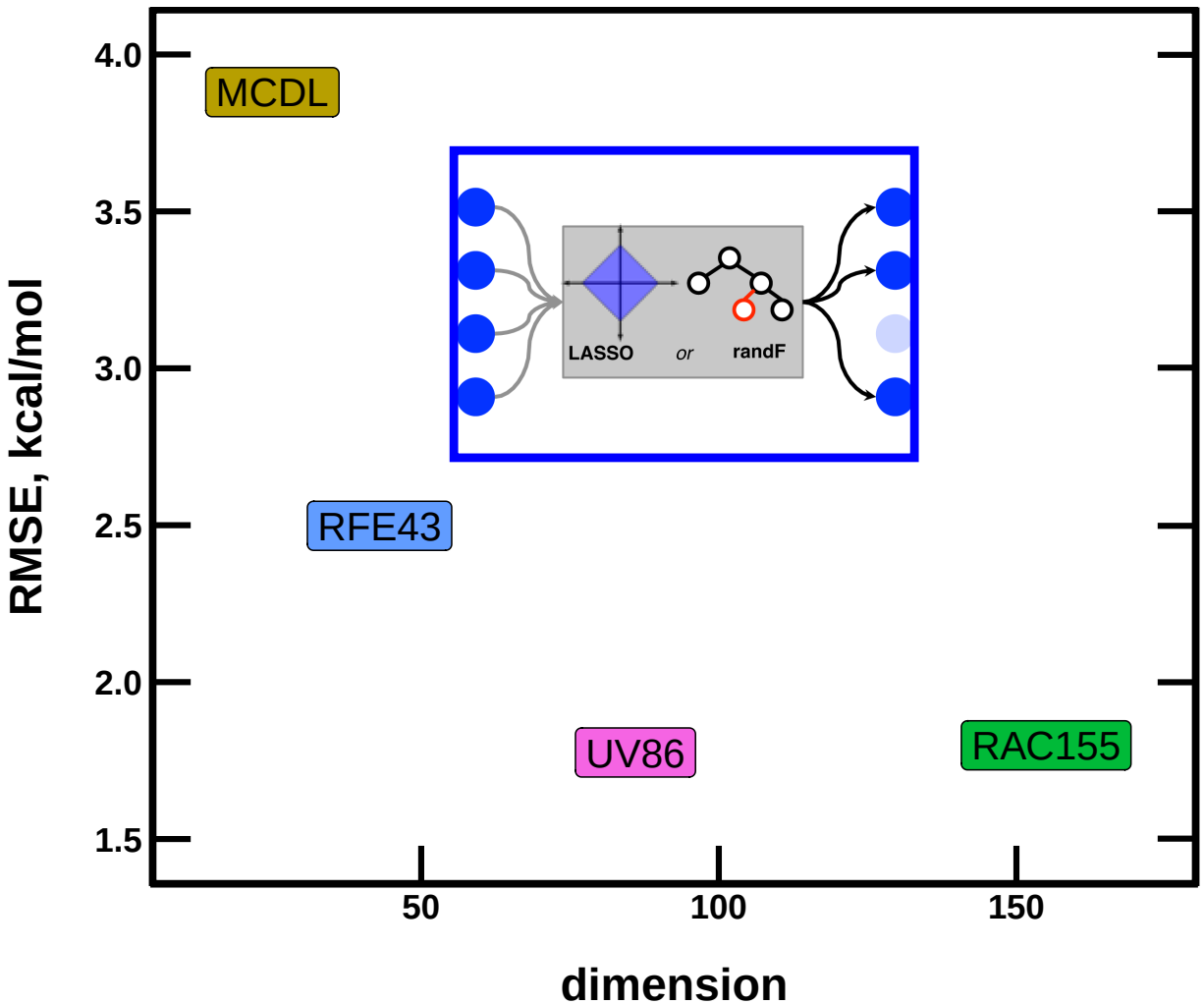

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



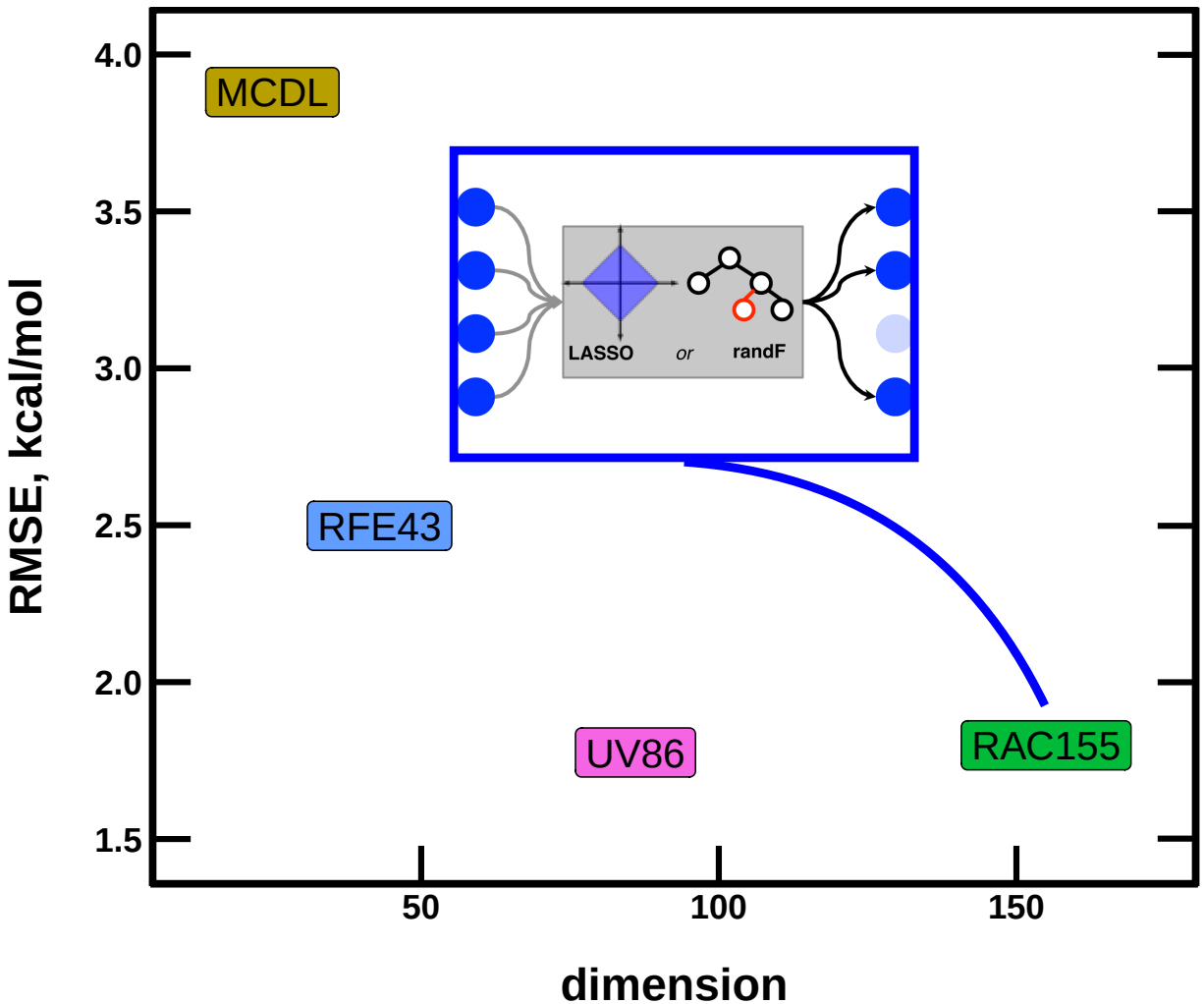

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



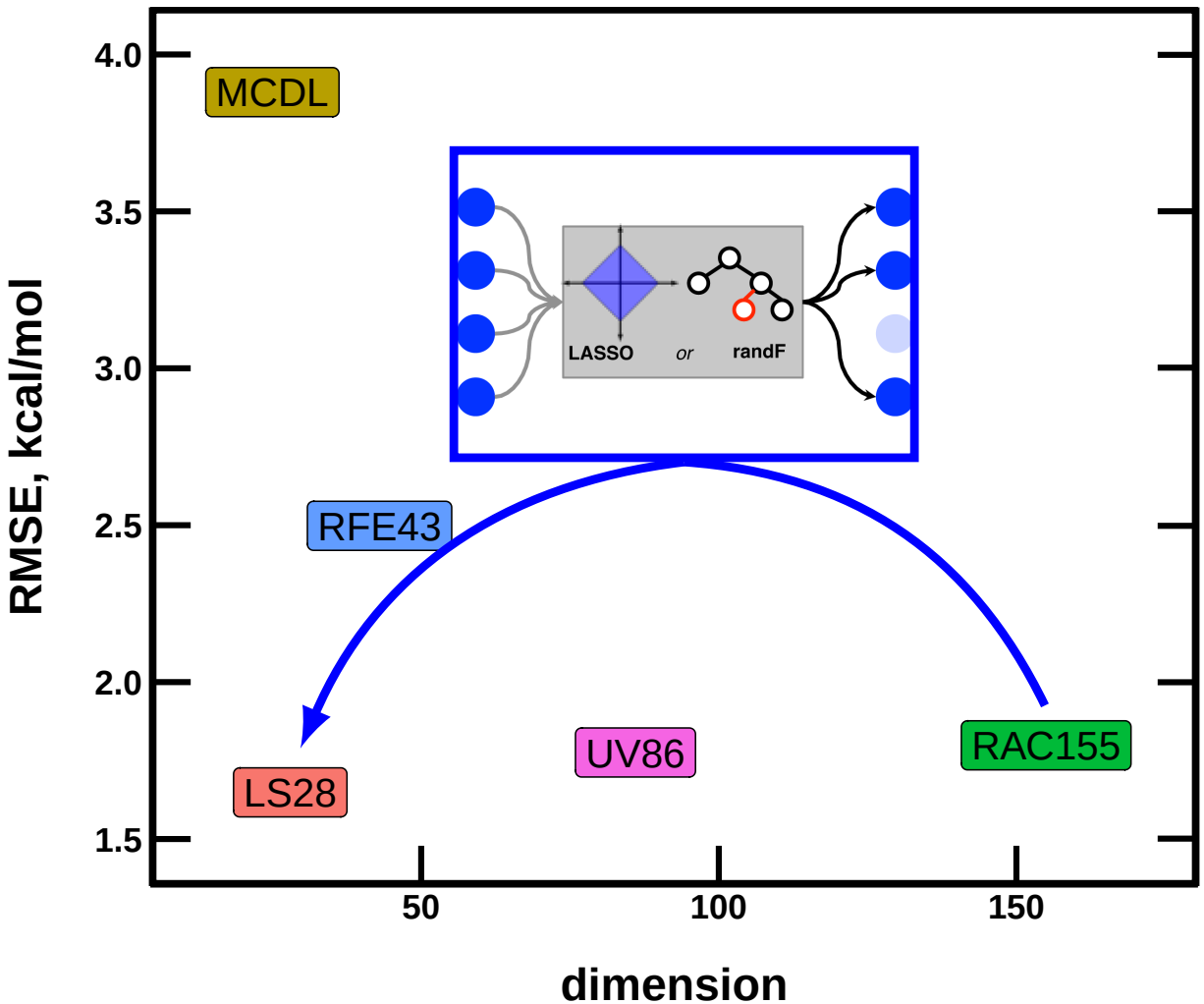

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



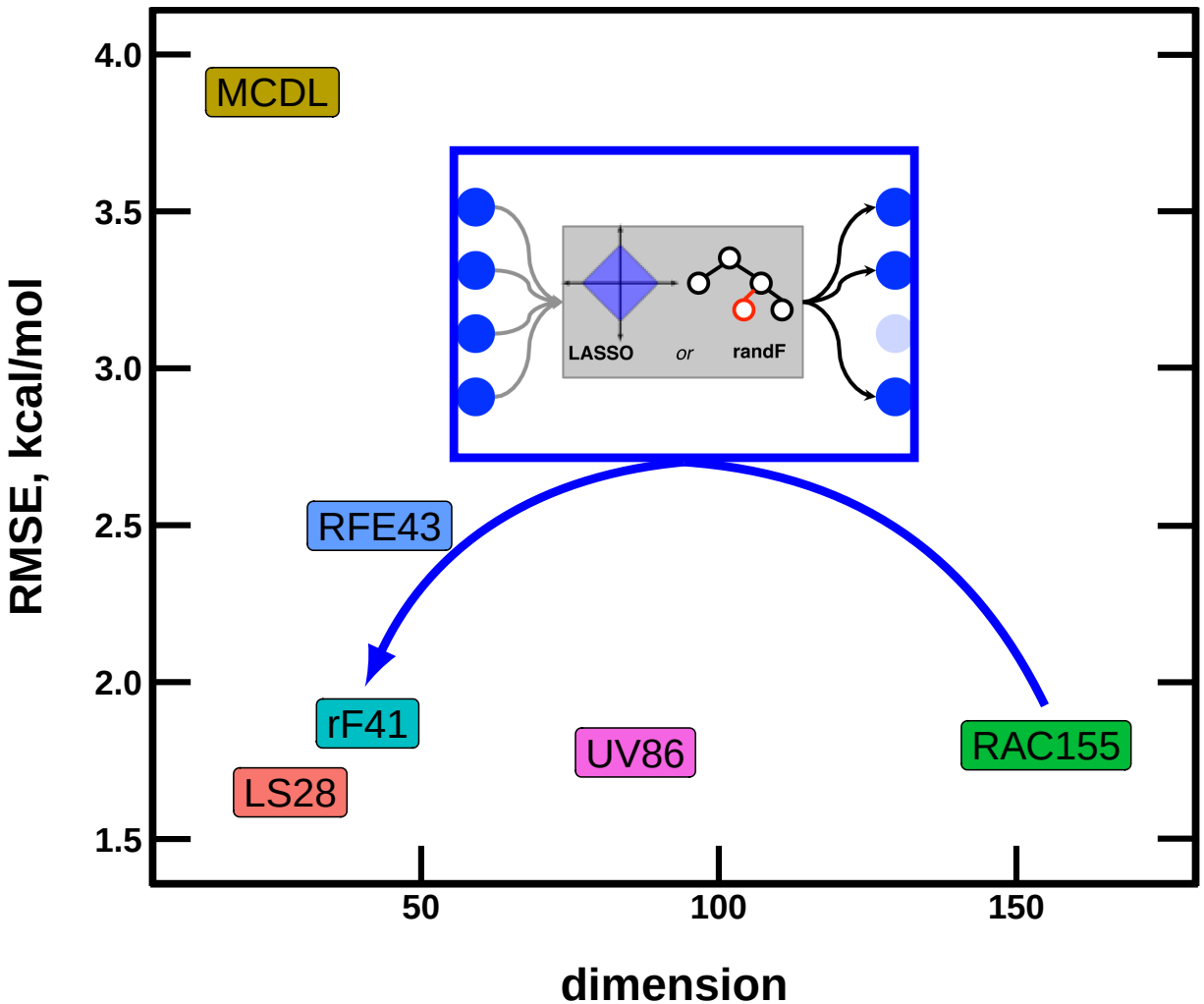

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# Feature selection



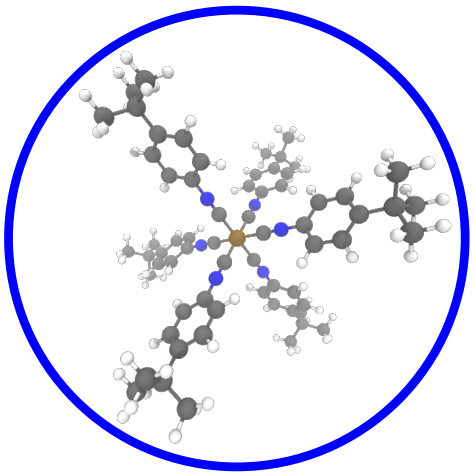

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

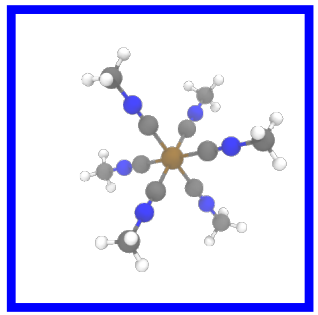
# Feature selection

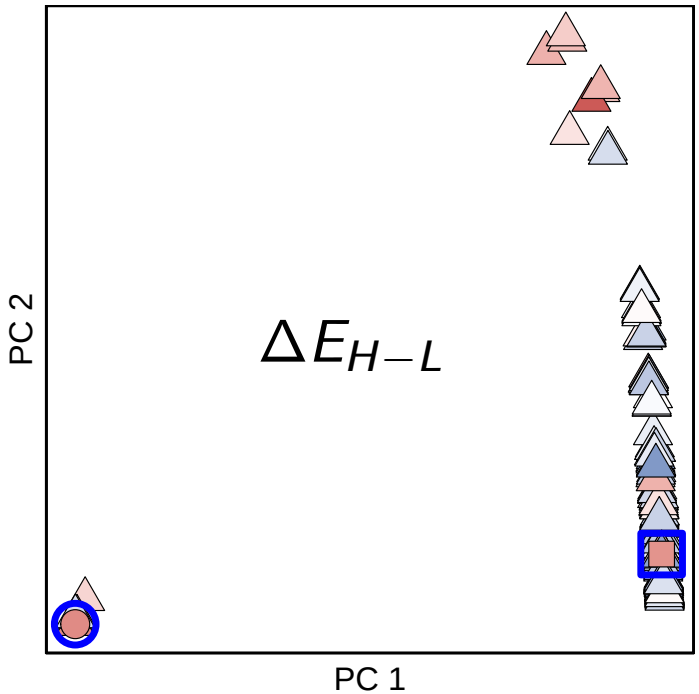


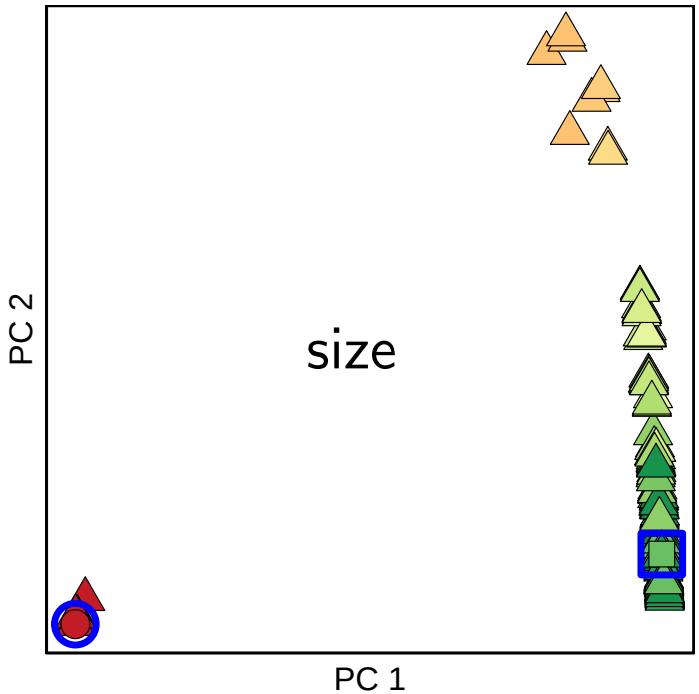

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# A tale of two complexes, II

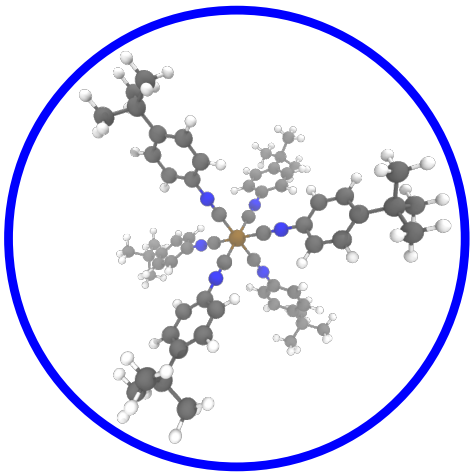


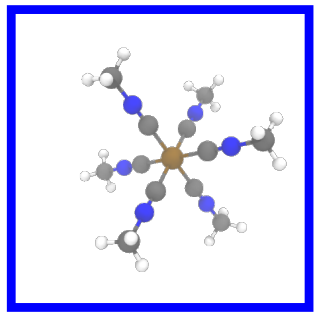


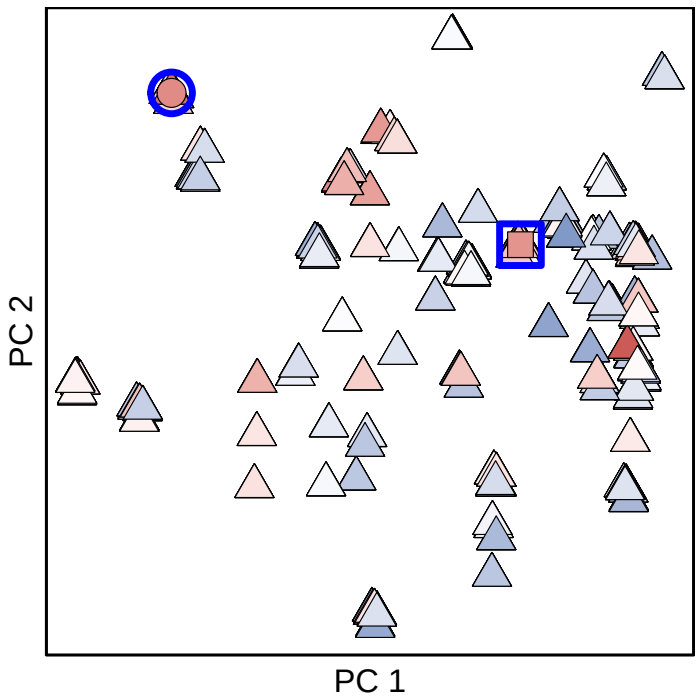


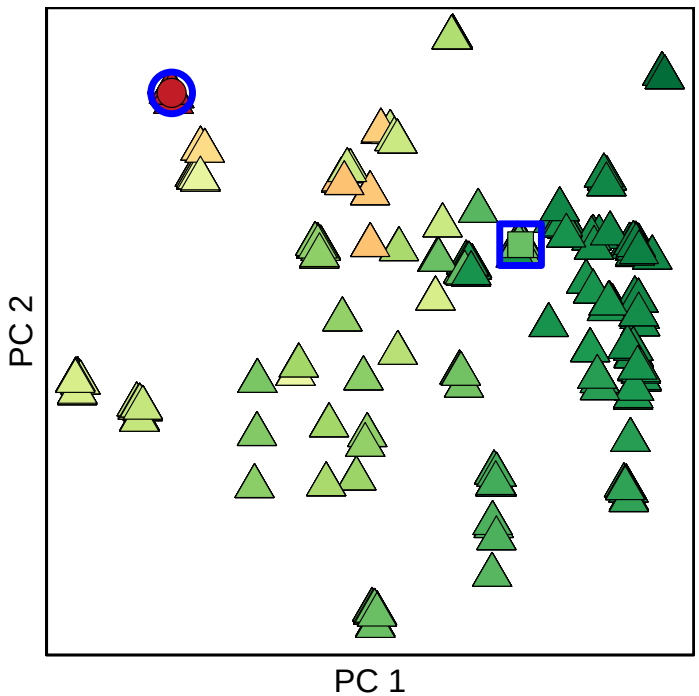



# A tale of two complexes, II

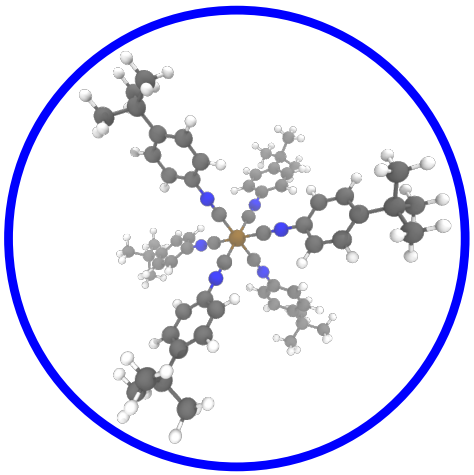


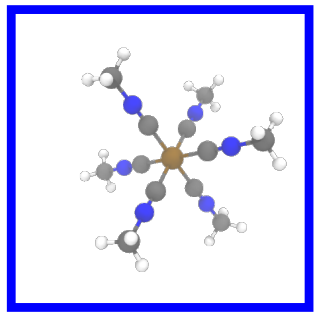


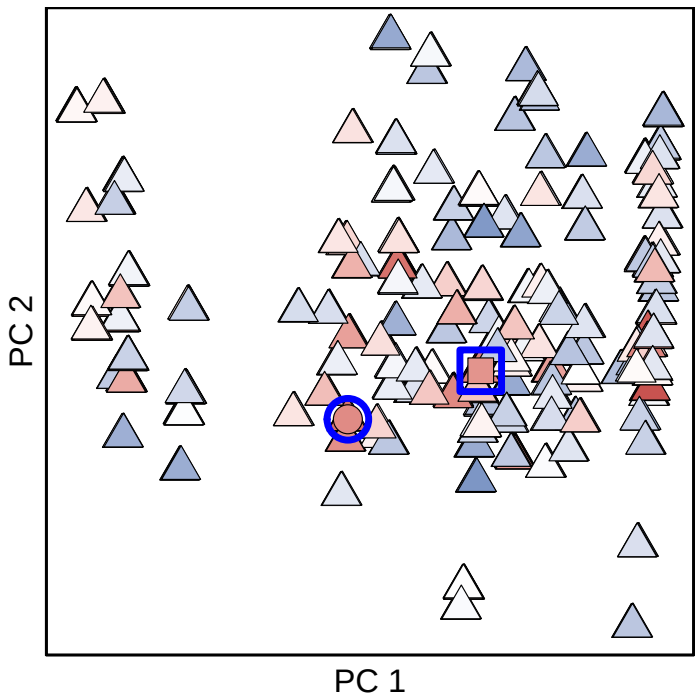


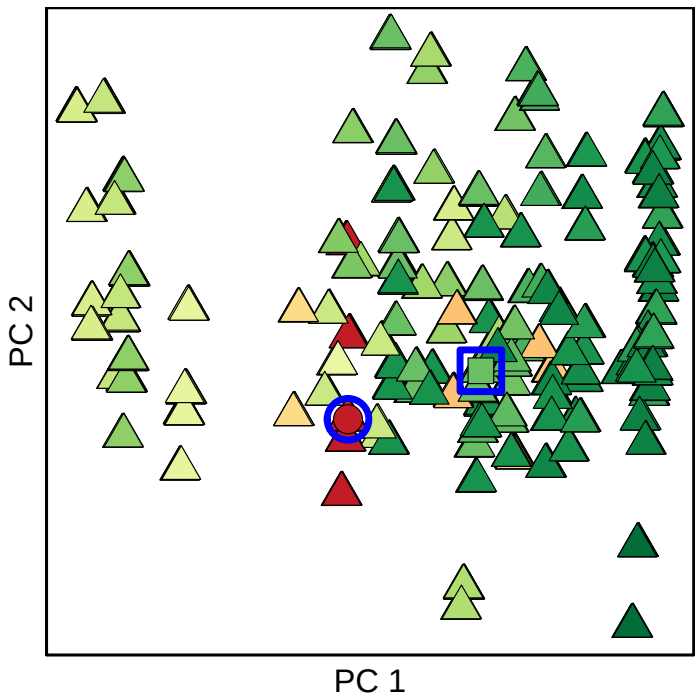



# A tale of two complexes, II

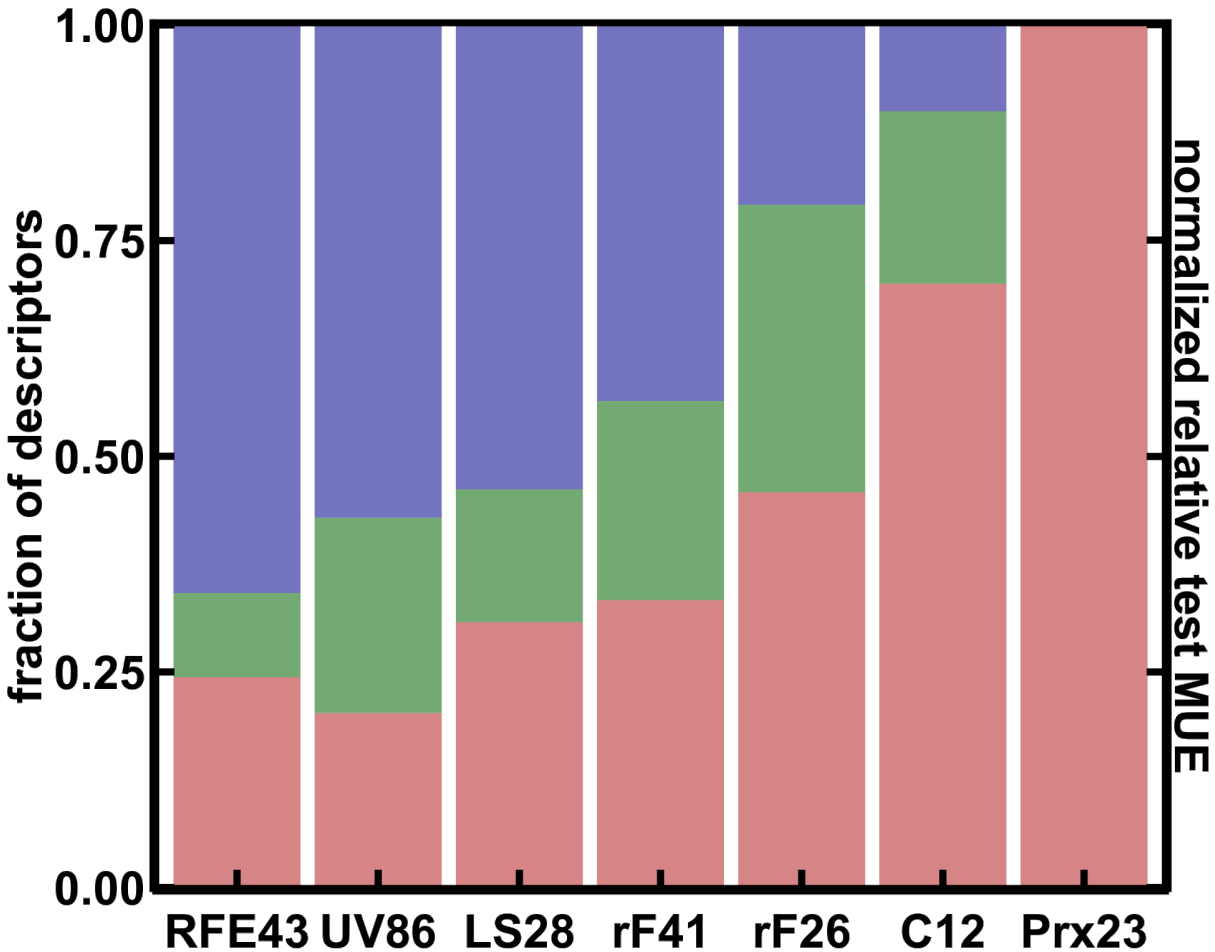






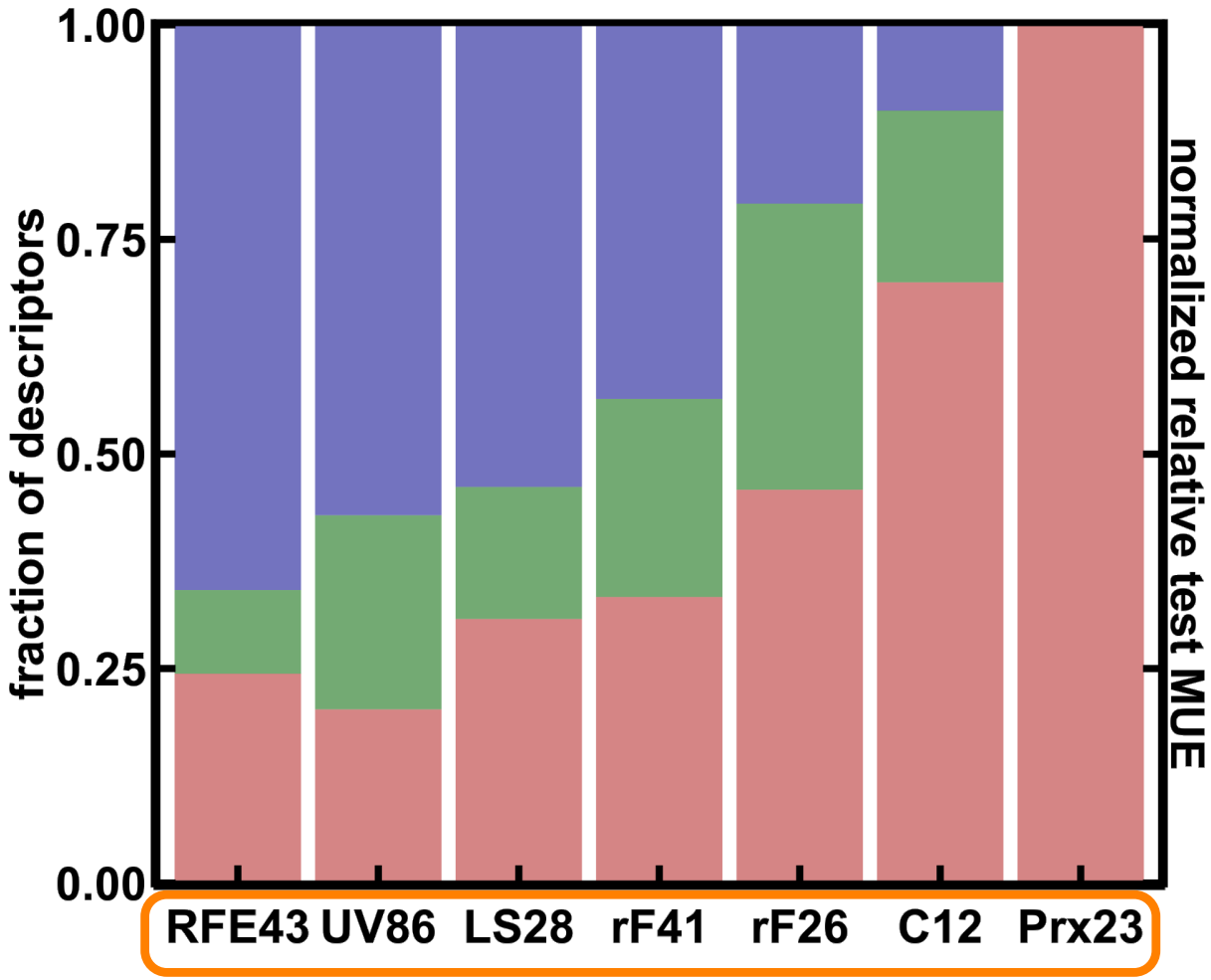



# How local is too local?



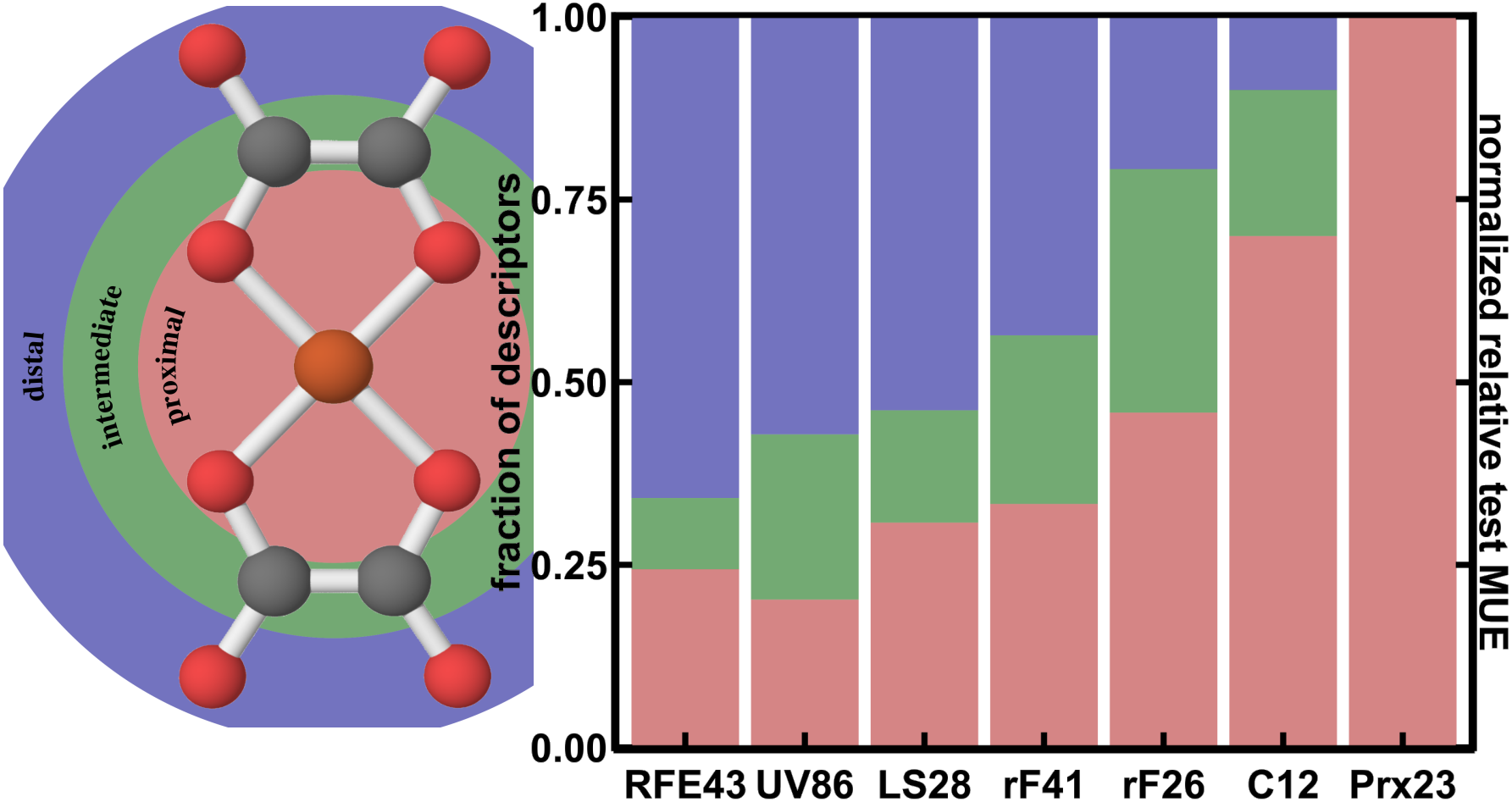

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# How local is too local?



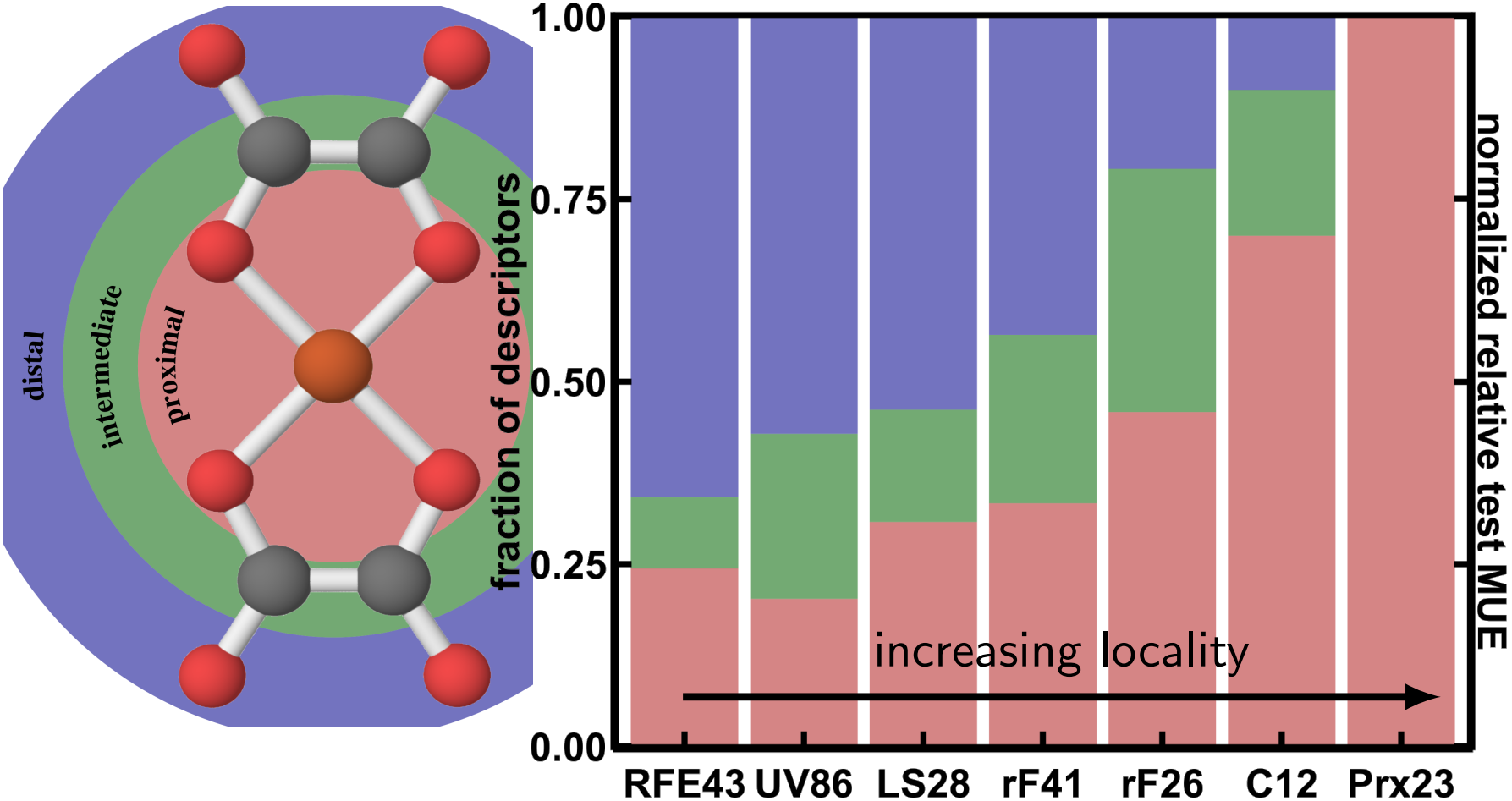

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# How local is too local?



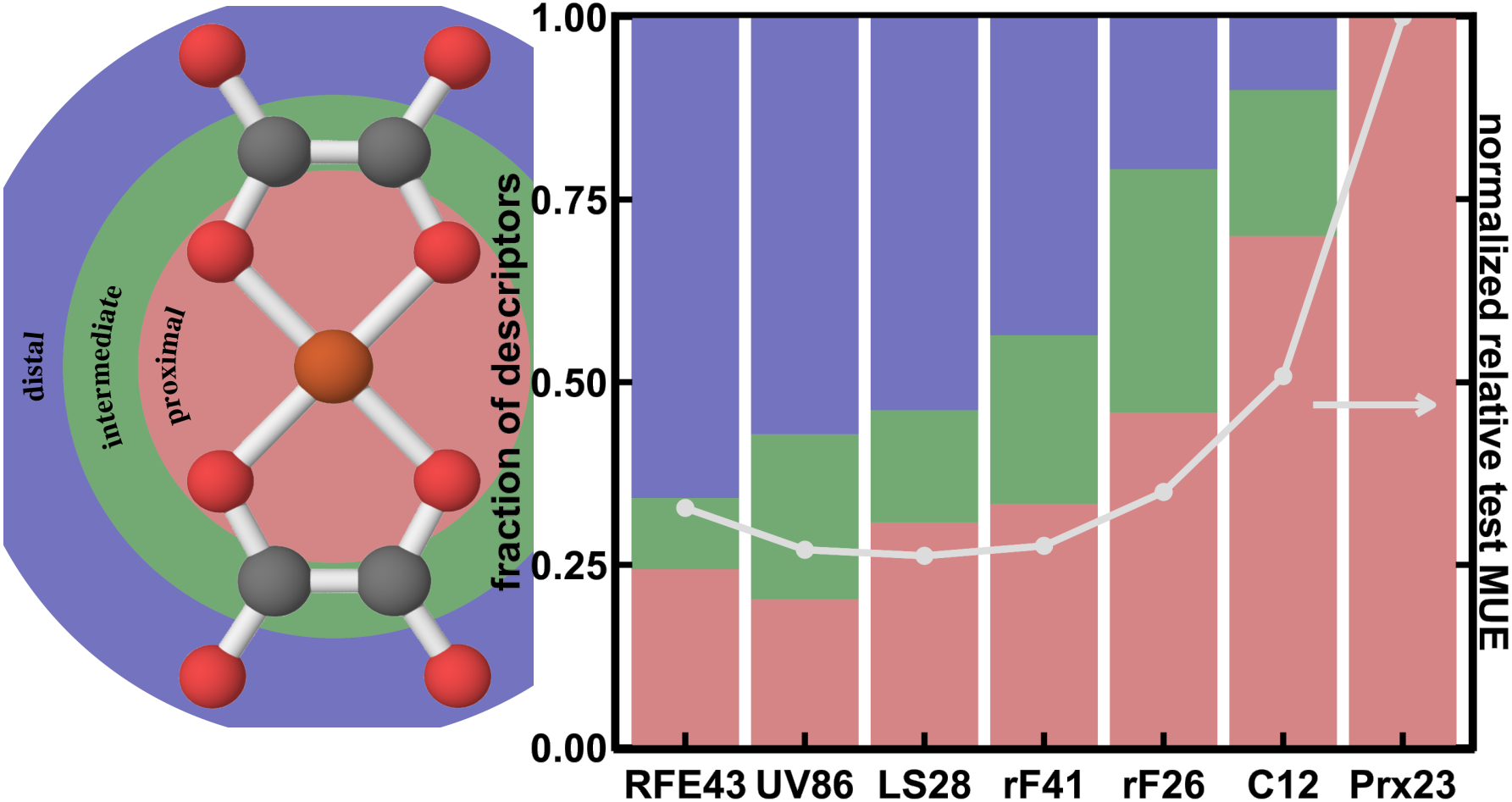

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# How local is too local?



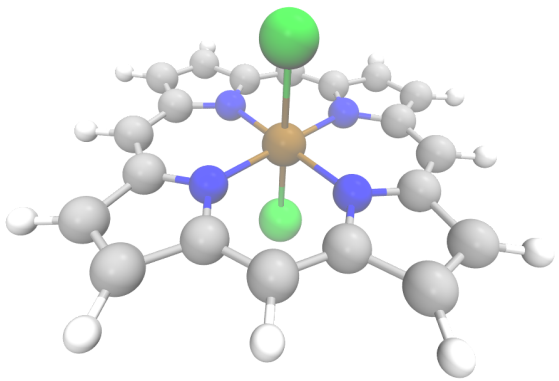

Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

# How local is too local?

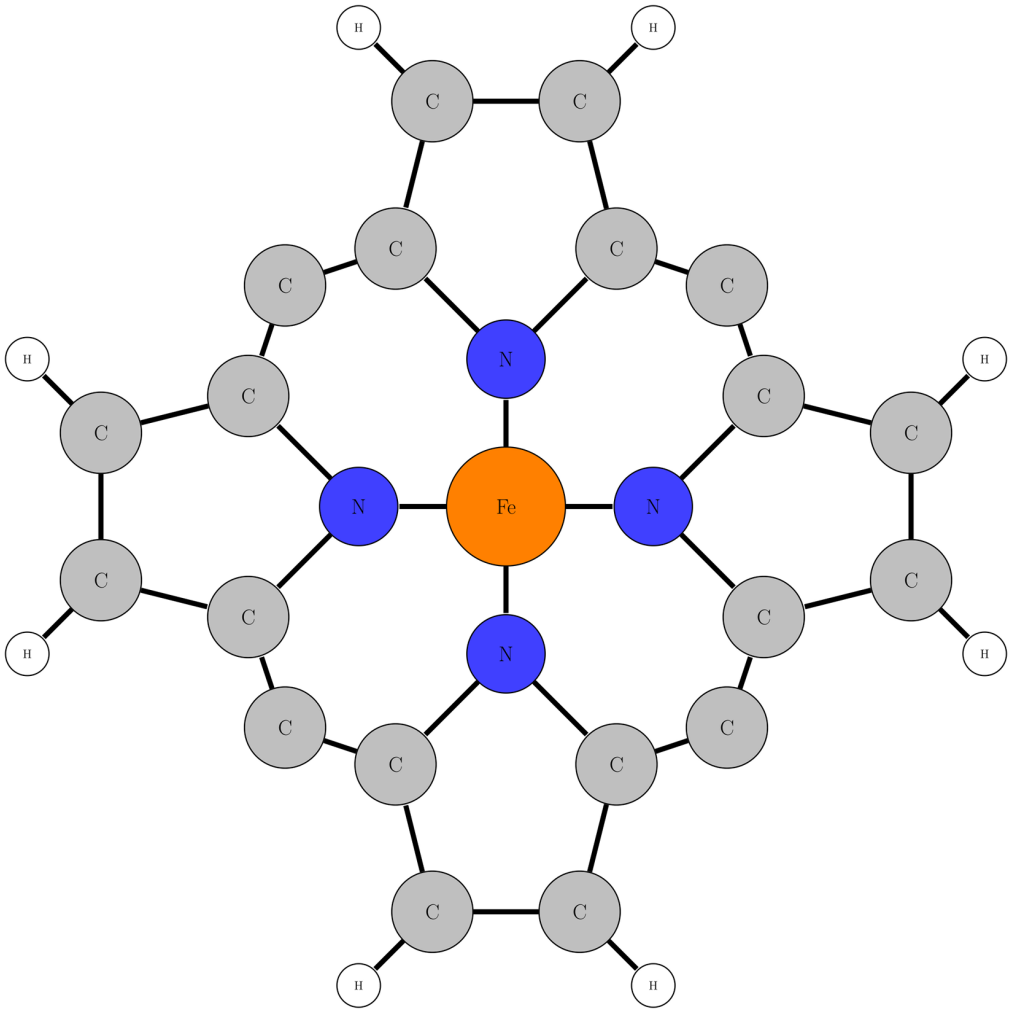



Janet, J.P., and Kulik, H.J., *arXiv*, 1708.06017, 2017.

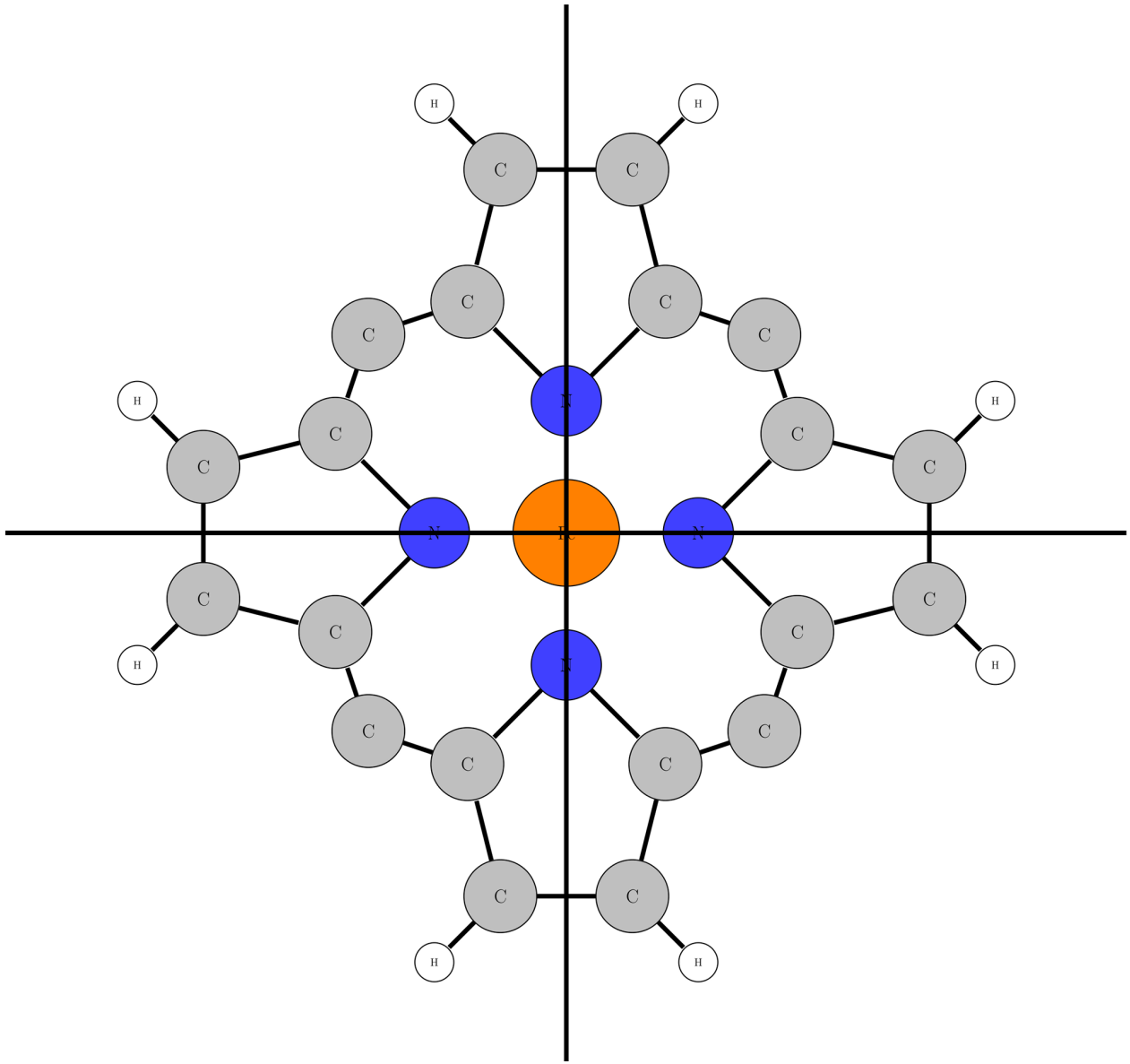
# How does the information localize?



# How does the information localize?

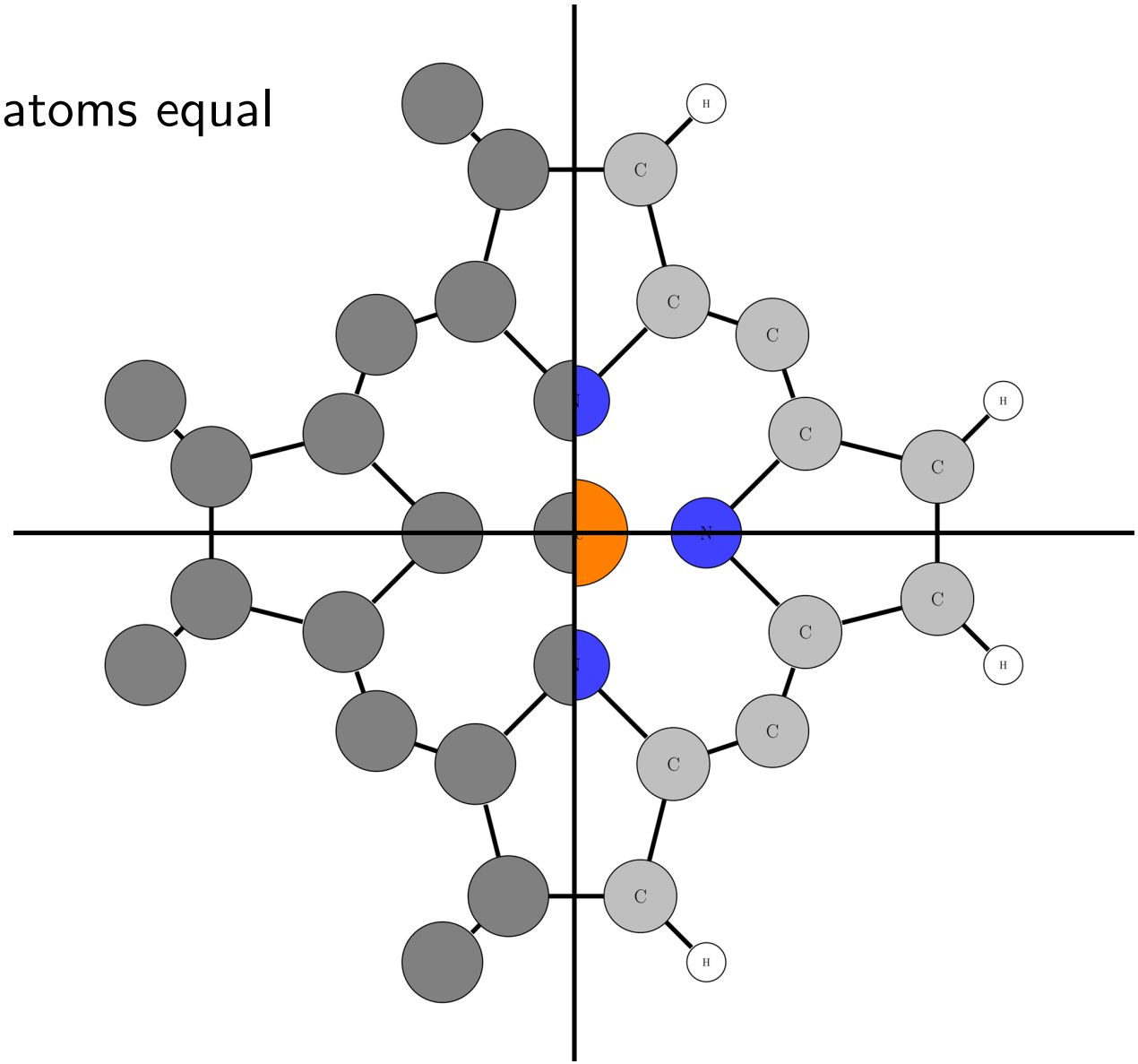


# How does the information localize?



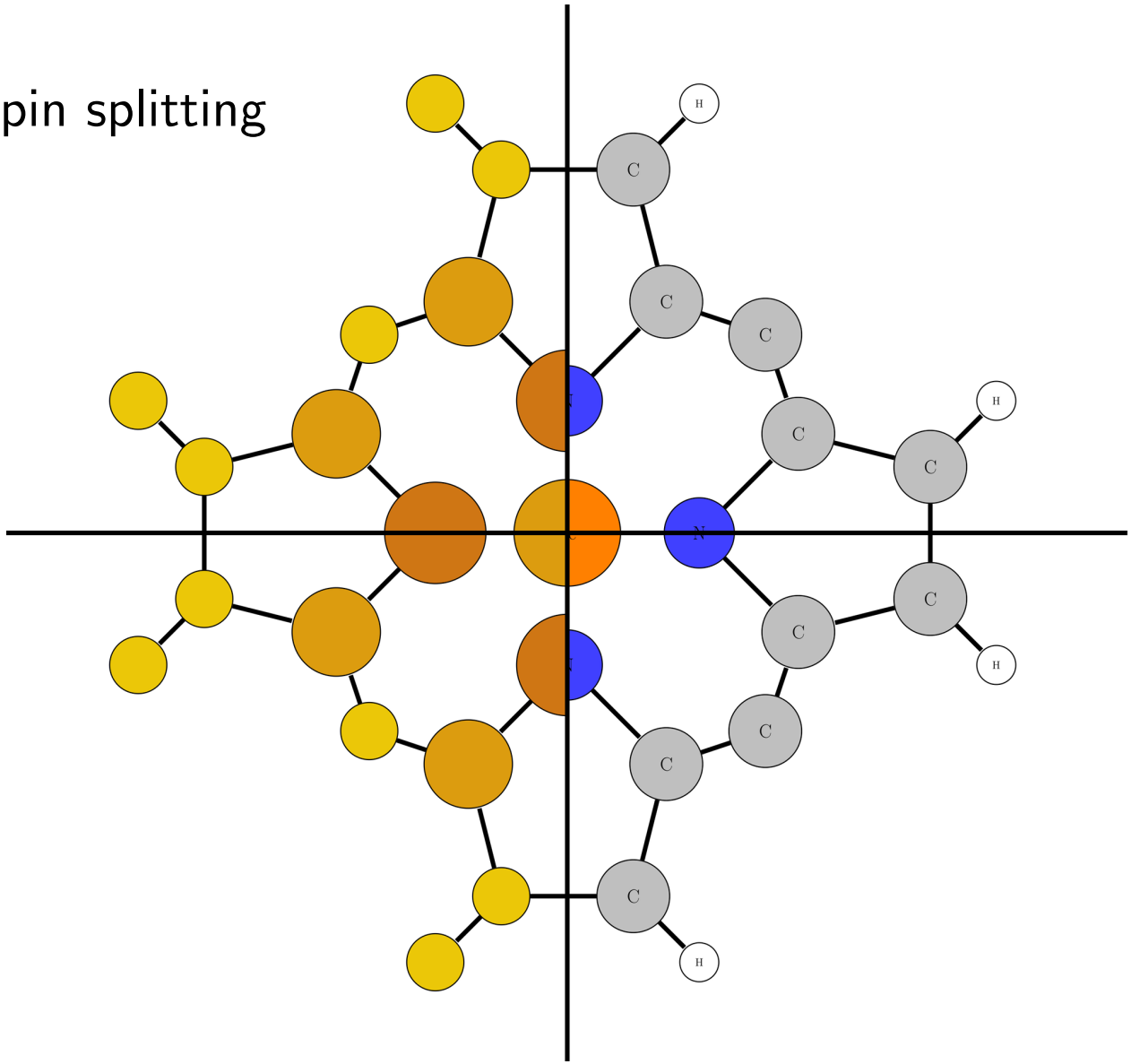
# How does the information localize?

All atoms equal



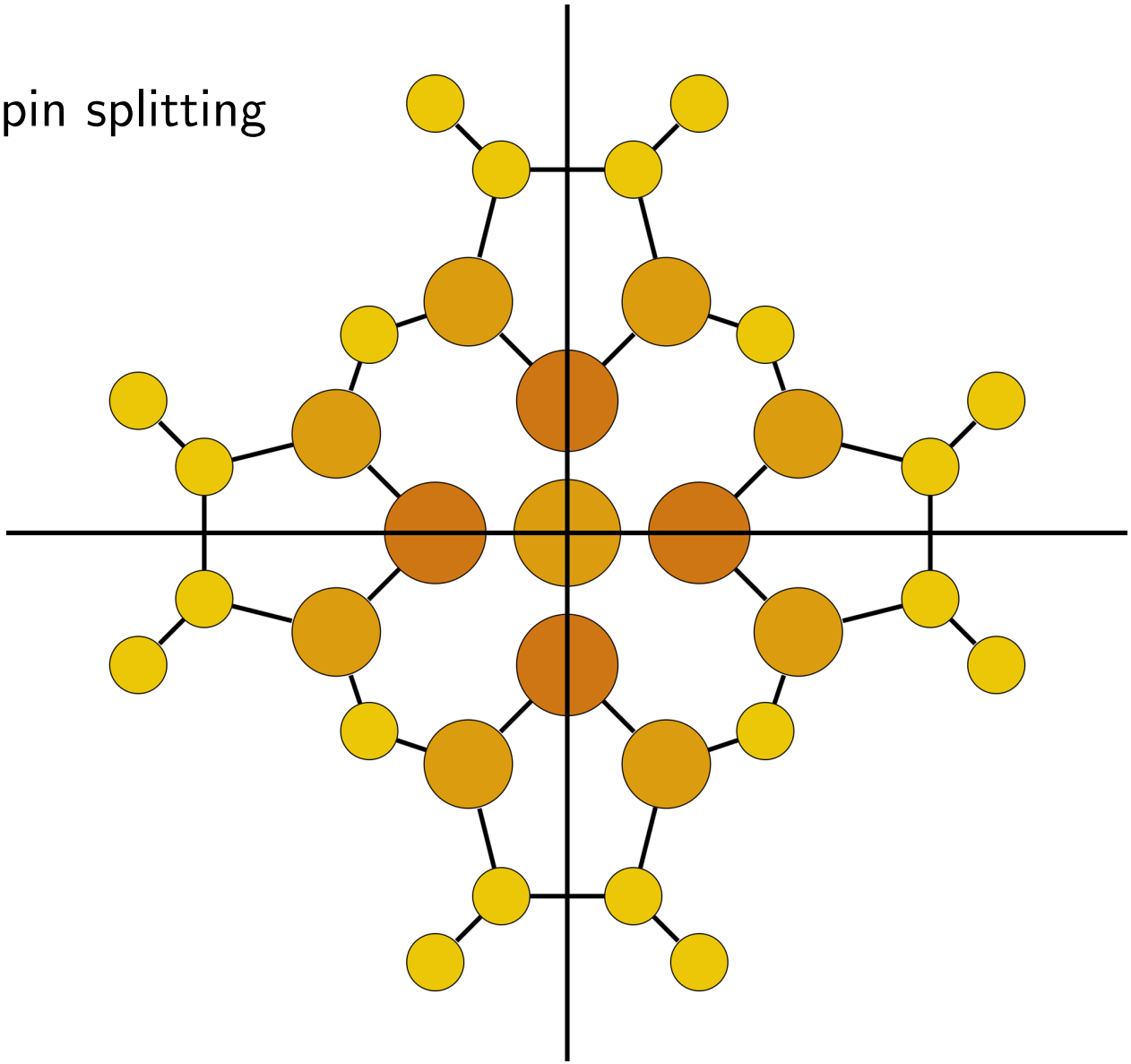
# How does the information localize?

rF41 - spin splitting

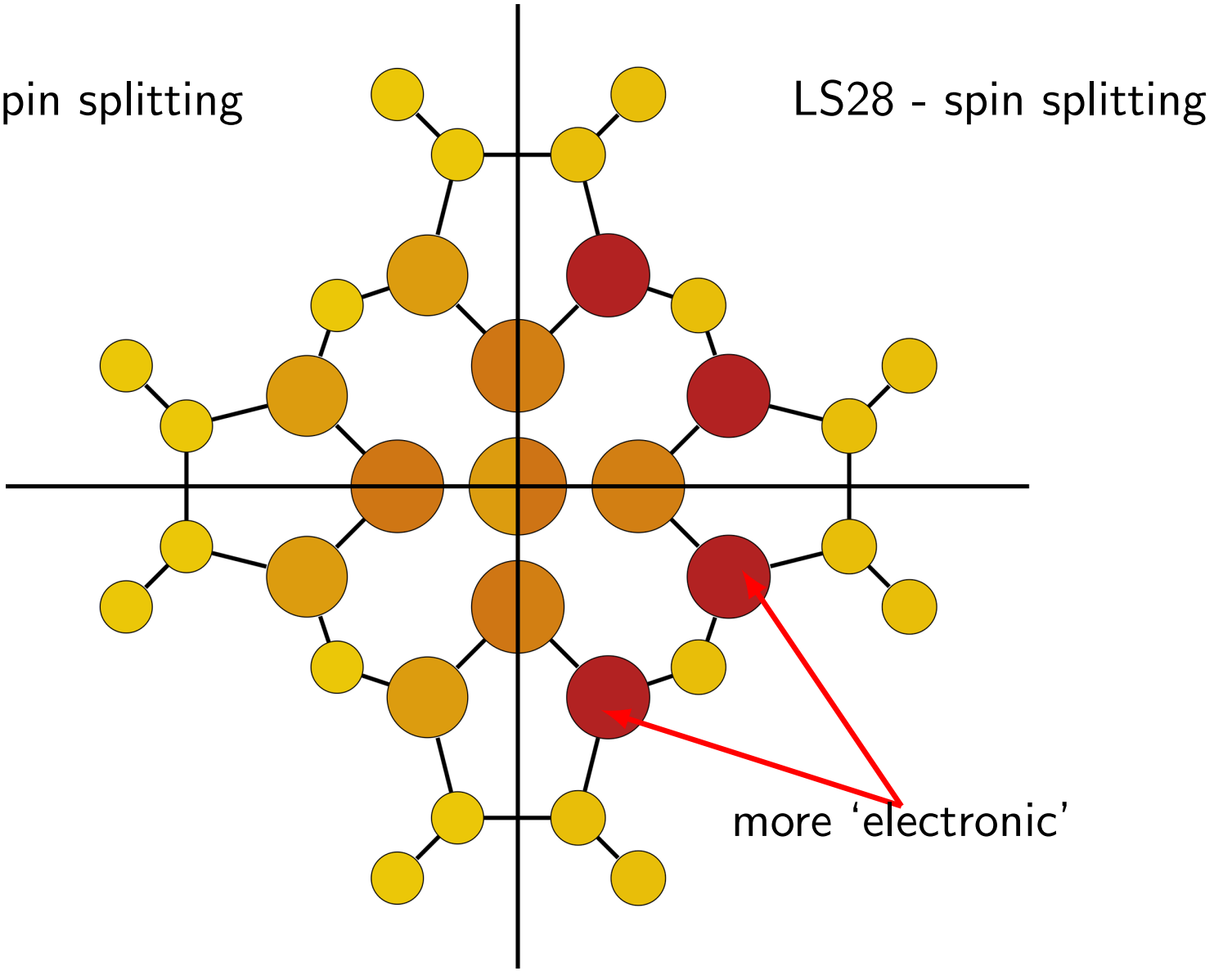


# How does the information localize?

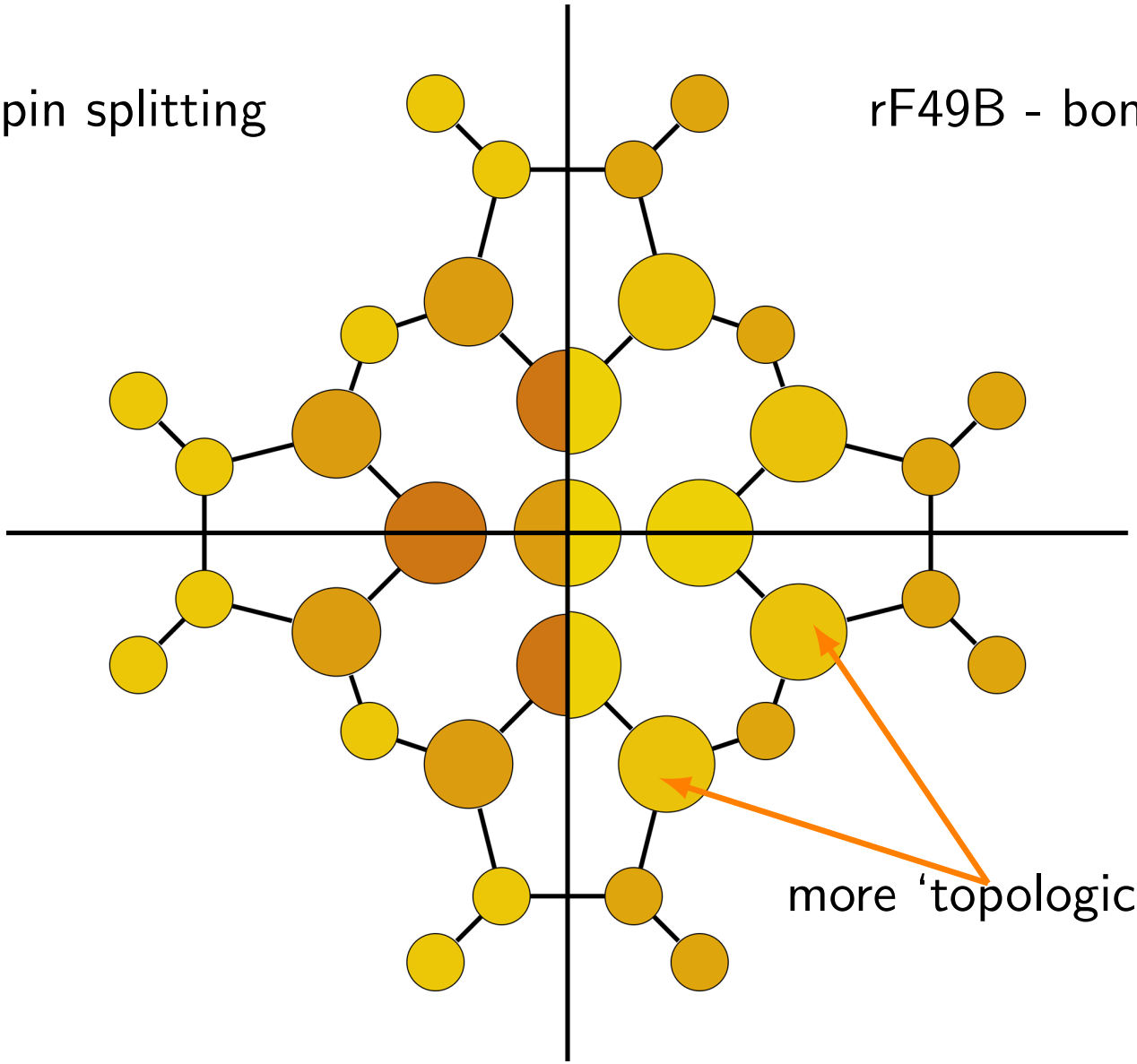
rF41 - spin splitting



# How does the information localize?



# How does the information localize?



# Outlook

Conclusions:

- machine learning TM complexes faces unique challenges but has enormous potential
- descriptors from organic ML are poorly suited to inorganic systems with moderate data
- ACs are a promising starting point for low-cost descriptors
- imbuing 'chemical intuition' to descriptor construction can drastically improve learning

# Outlook

Conclusions:

- machine learning TM complexes faces unique challenges but has enormous potential
- descriptors from organic ML are poorly suited to inorganic systems with moderate data
- ACs are a promising starting point for low-cost descriptors
- imbuing 'chemical intuition' to descriptor construction can drastically improve learning

Future work:

- use ACs to drive exploration/rational design in TM complex space

# Acknowledgments

Thanks to the Kulik group

- Prof. Heather Kulik
- Akash Bajaj
- Terry Gani
- Dr. Jeong-yun Kim
- Rimsha Mehmood
- Helena Qi
- Dr. Adam Steeves
- Qing Zhao