Machine-learning assisted workflows for inorganic molecular discovery

Jon Paul Janet [1] Chenru Duan[2] Aditya Nandy[2]
Heather Kulik [1]

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

[2]Department of Chemistry, Massachusetts Institute of Technology

Machine Learning and Informatics for Chemistry and Materials

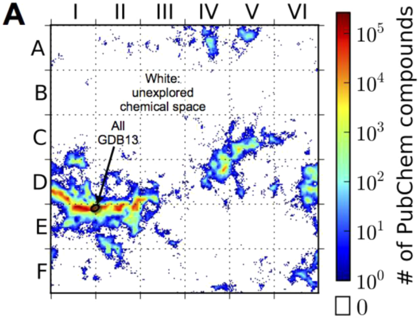
10.02.19

Motivation: chemical discovery

How can we design new materials using computers?

The space of possible chemistries is incredibly vast, with $\mathcal{O}(10^{60})$ small organic molecules.

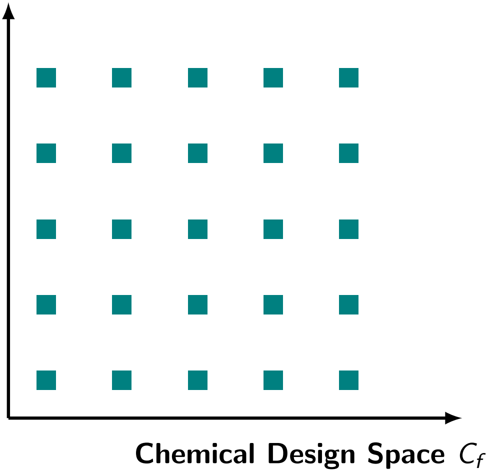
All potentially undiscovered medicines, catalysts and materials are somewhere, out in this huge space.



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

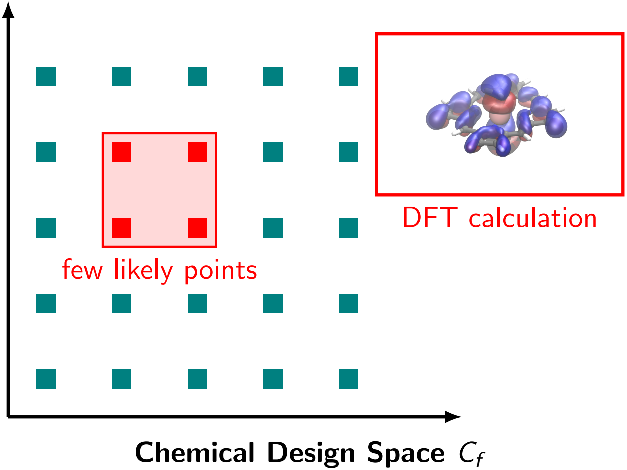
Motivation: chemical discovery

How can we design new materials using computers?



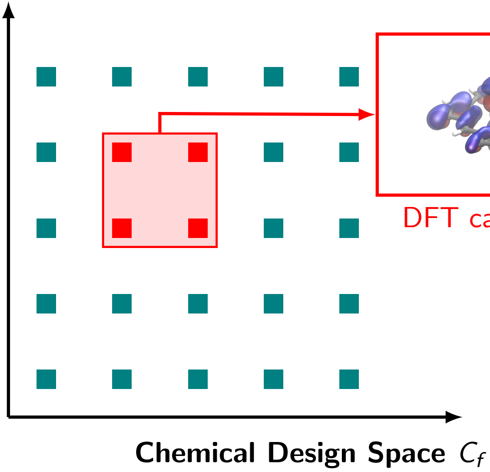
Motivation: chemical discovery

How can we design new materials using computers?



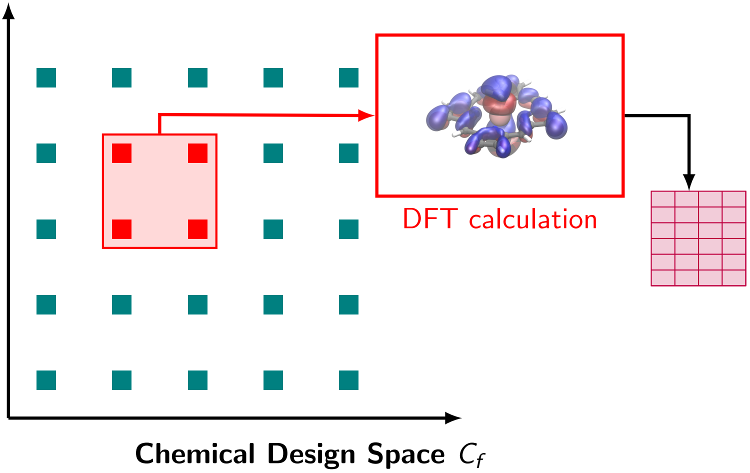
Motivation: chemical discovery

How can we design new materials using computers?



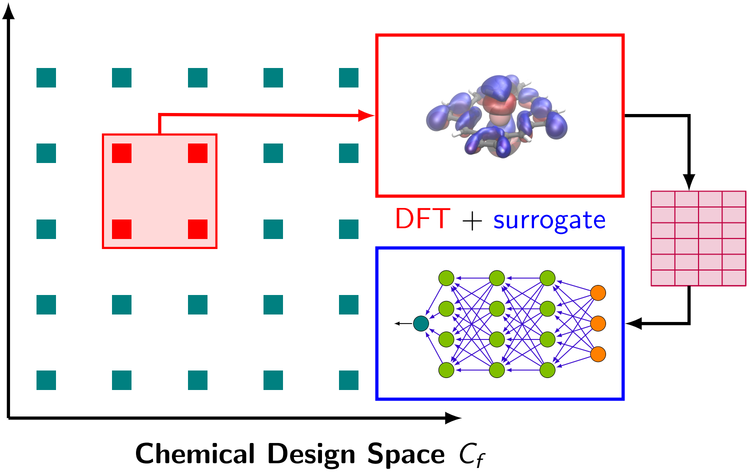
Motivation: chemical discovery

How can we design new materials using computers?



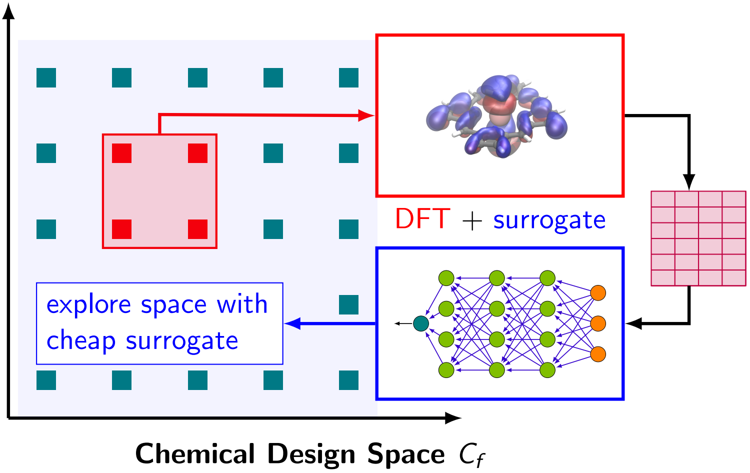
Motivation: chemical discovery

How can we design new materials using computers?



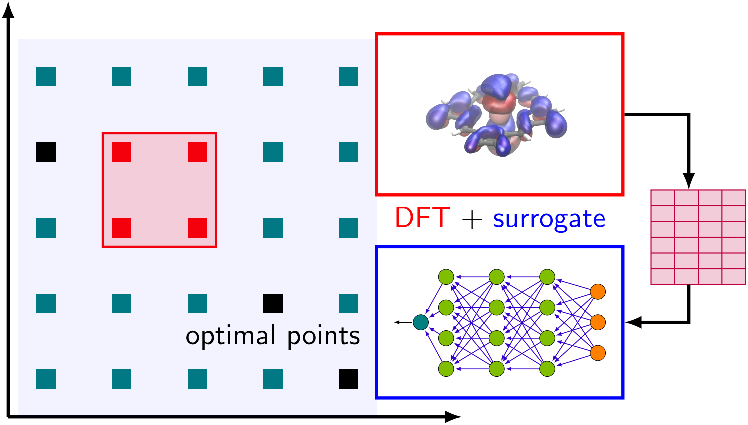
Motivation: chemical discovery

How can we design new materials using computers?



Motivation: chemical discovery

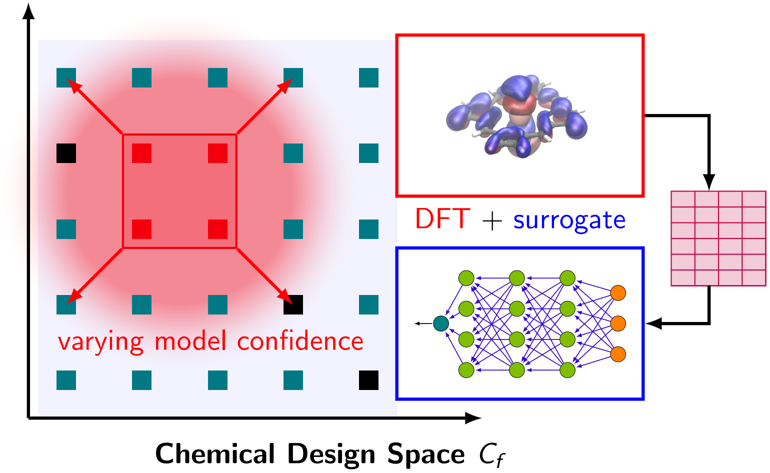
How can we design new materials using computers?



Chemical Design Space C_f

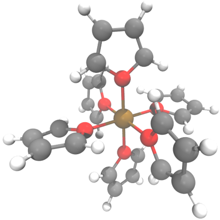
Motivation: chemical discovery

How can we design new materials using computers?

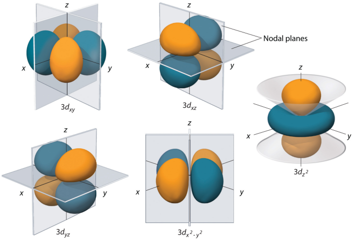


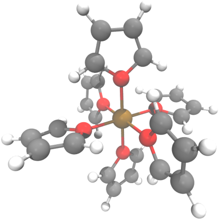
Transition metal complexes

Transition metal complexes

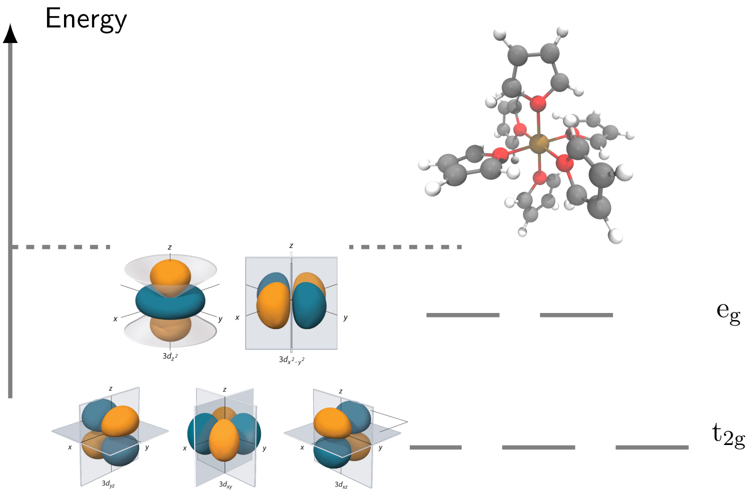


Transition metal complexes

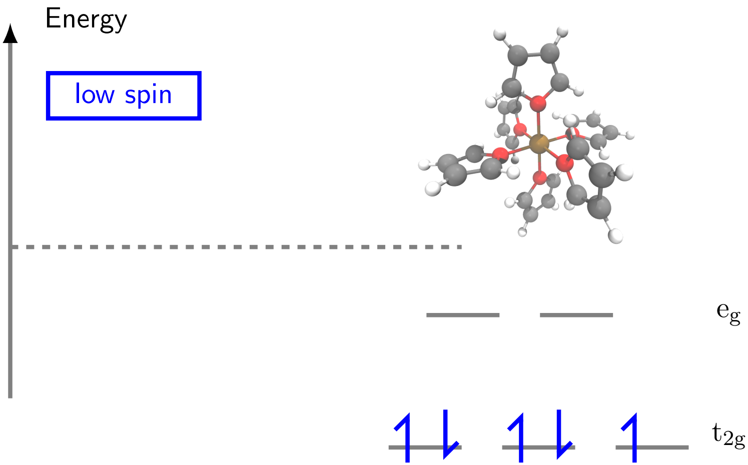




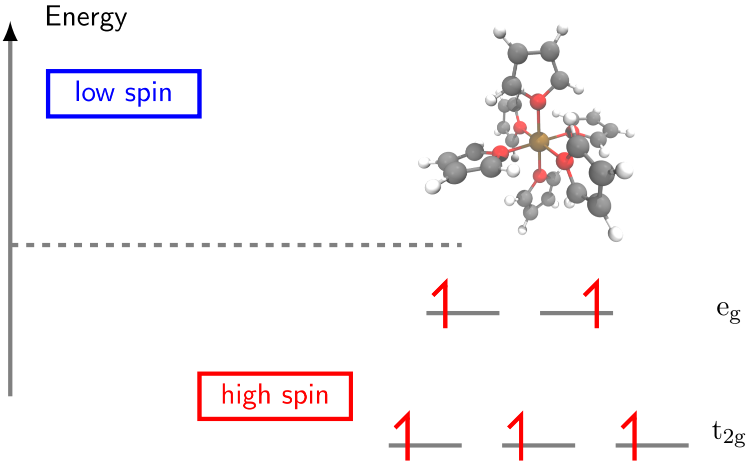
Transition metal complexes



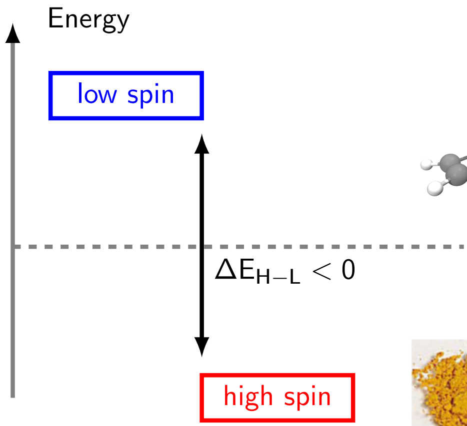
Transition metal complexes

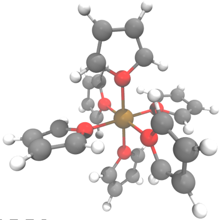


Transition metal complexes

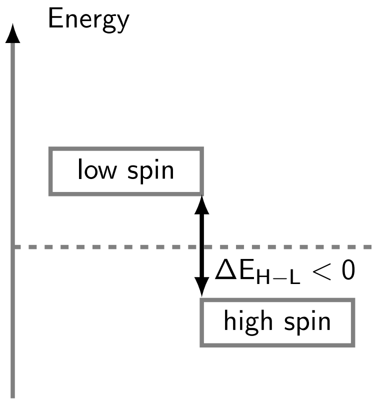


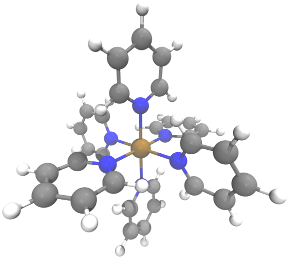
Transition metal complexes



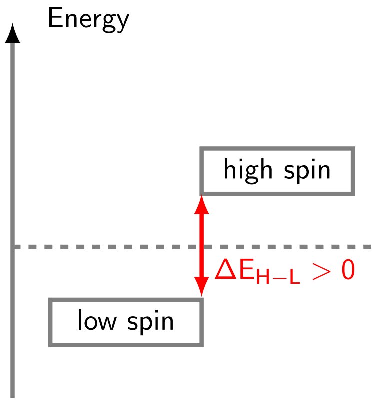


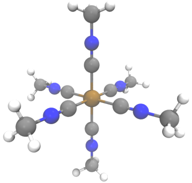
Transition metal complexes



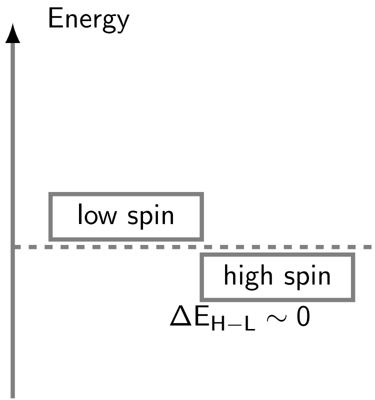


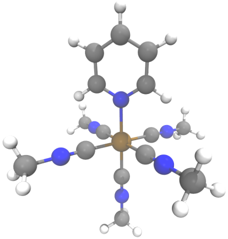
Transition metal complexes



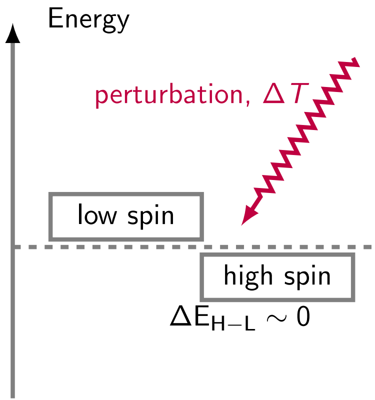


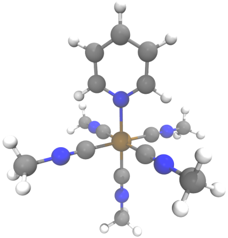
Transition metal complexes



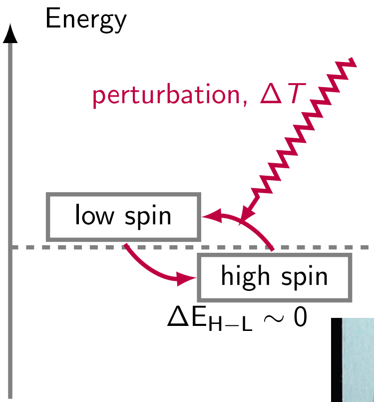


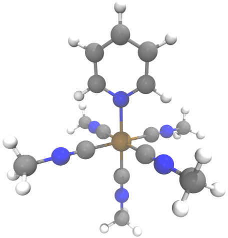
Transition metal complexes

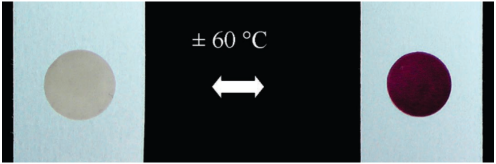




Transition metal complexes

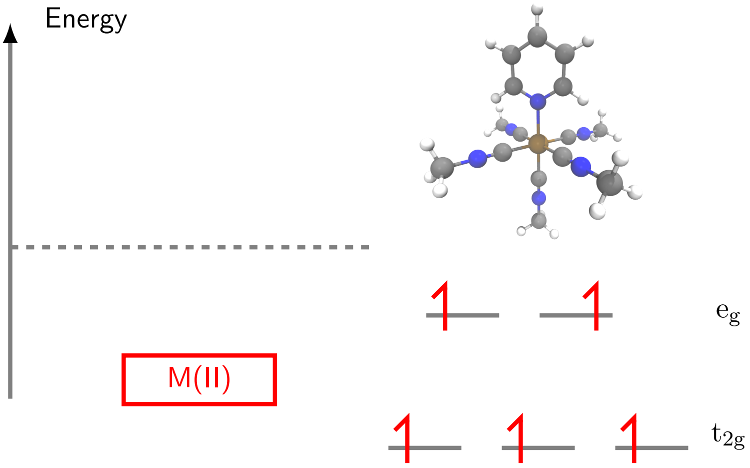




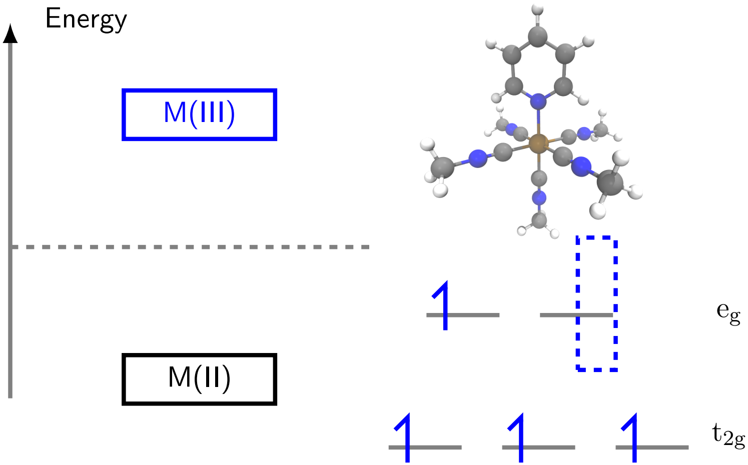


Seredyuk, M *et al.*, *Chem. Mater.*, 18(10):2513–2519, 2006.

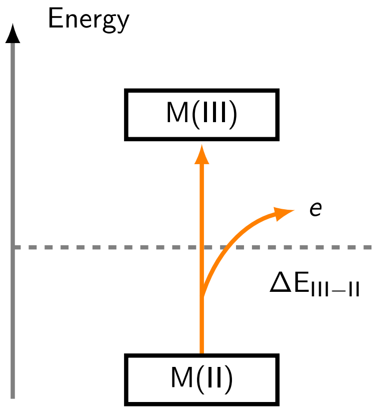
Transition metal complexes

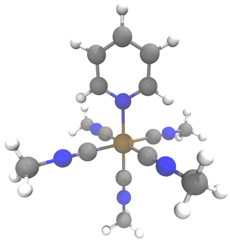


Transition metal complexes



Transition metal complexes





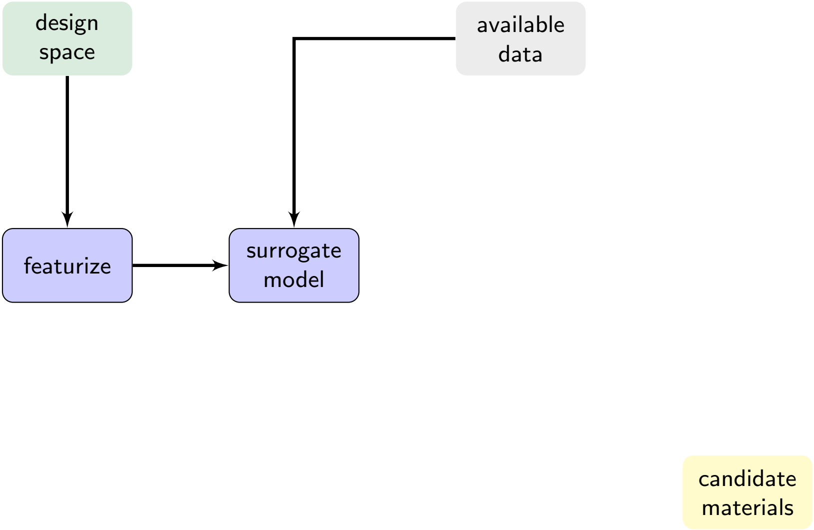
Algorithmic chemical discovery

design space

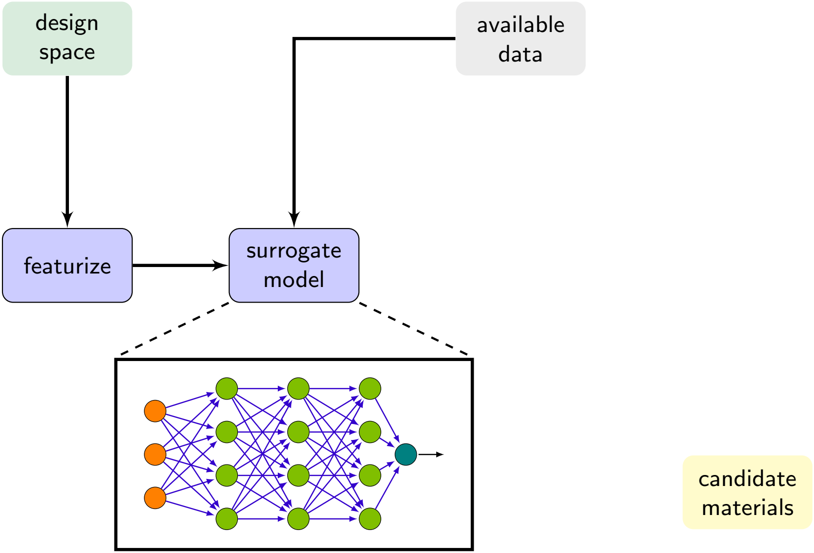
candidate materials

Algorithmic chemical discovery

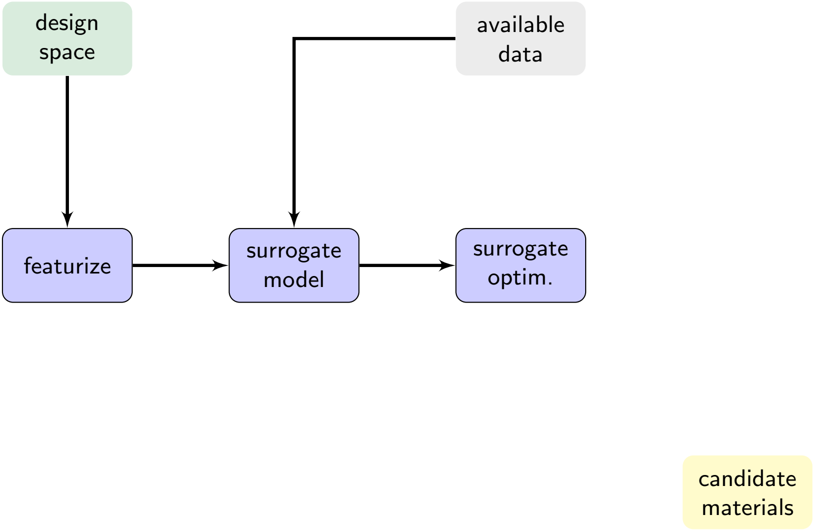
Algorithmic chemical discovery



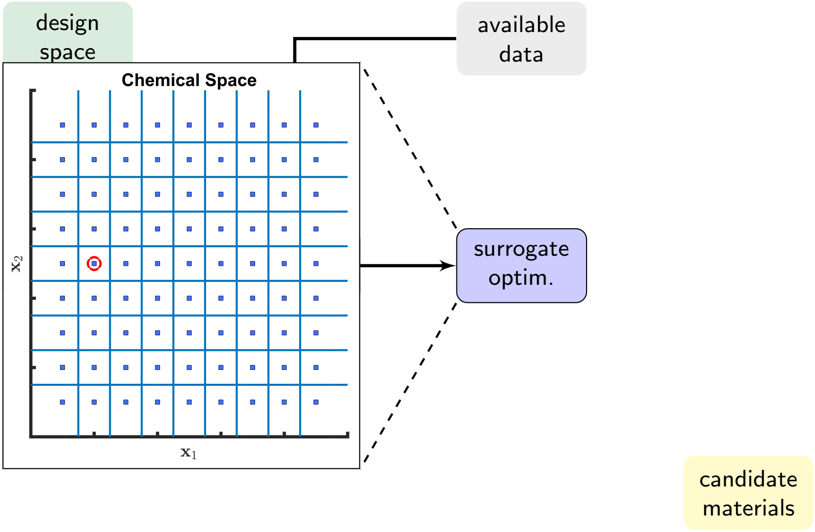
Algorithmic chemical discovery



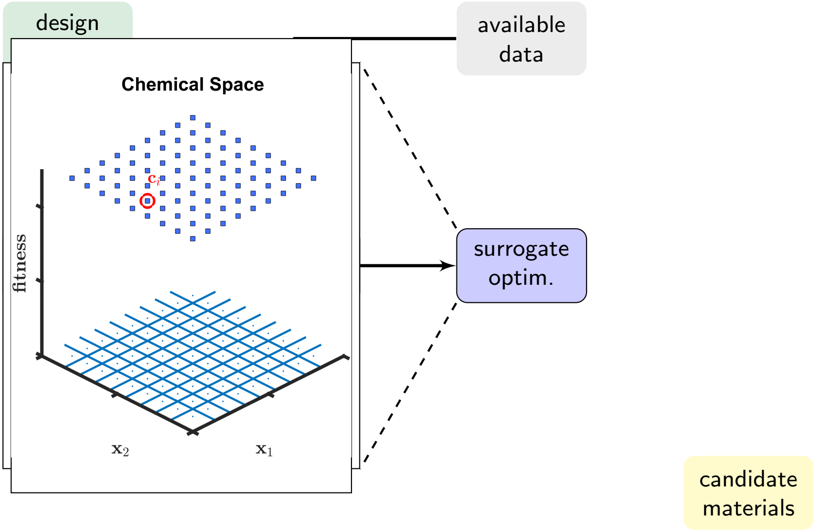
Algorithmic chemical discovery



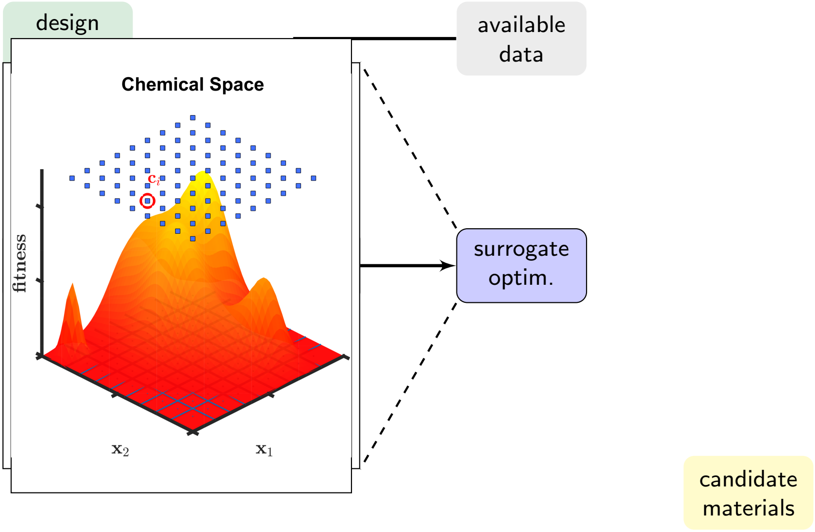
Algorithmic chemical discovery



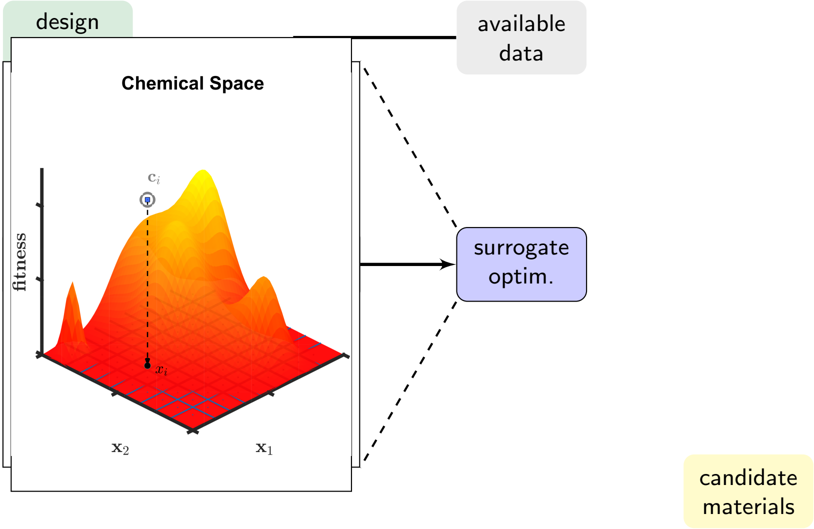
Algorithmic chemical discovery



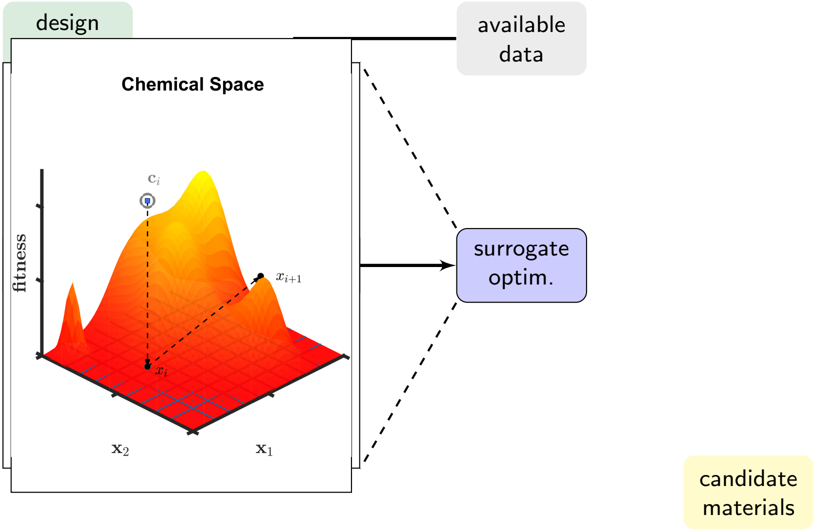
Algorithmic chemical discovery



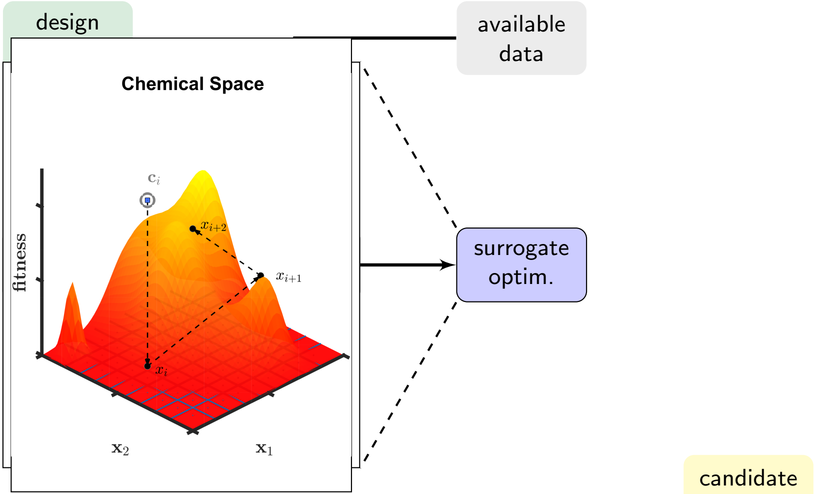
Algorithmic chemical discovery



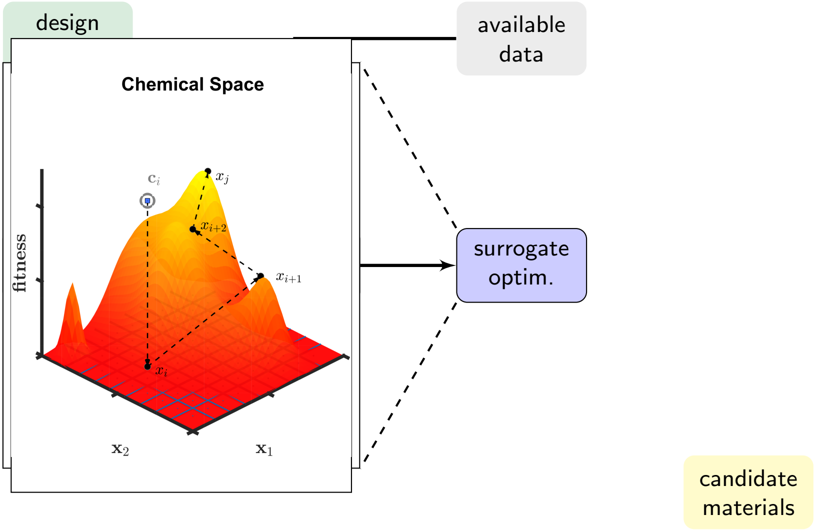
Algorithmic chemical discovery



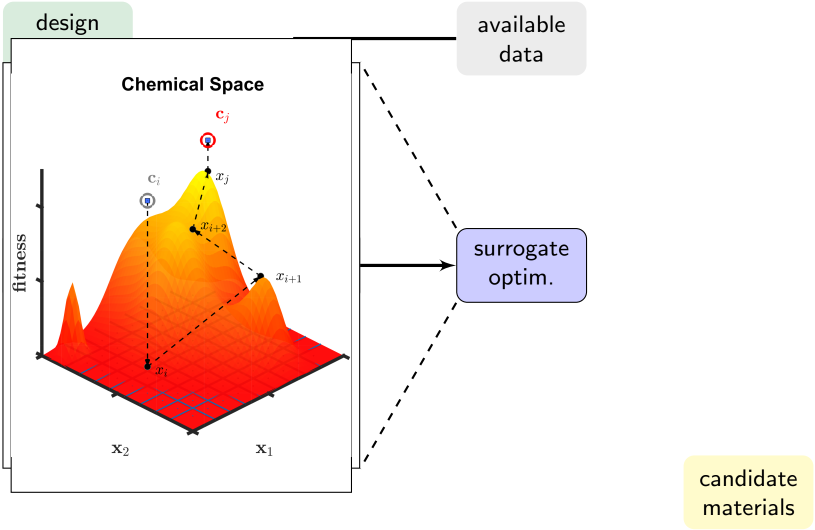
Algorithmic chemical discovery



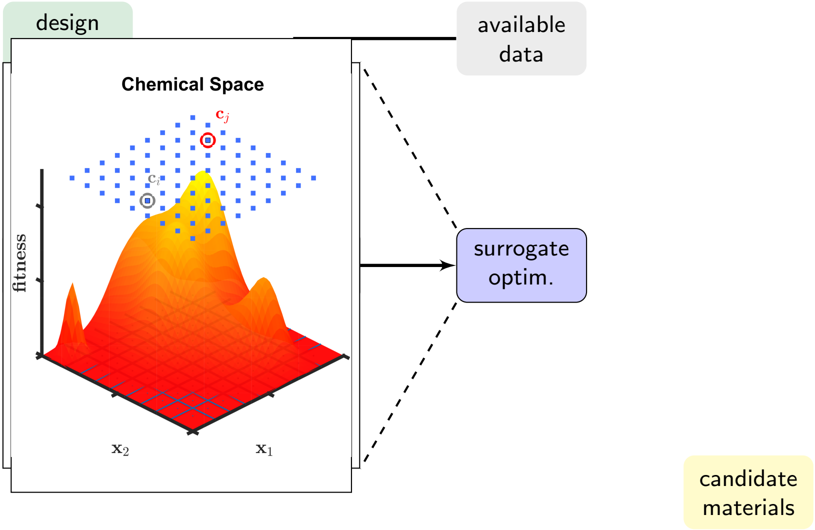
Algorithmic chemical discovery



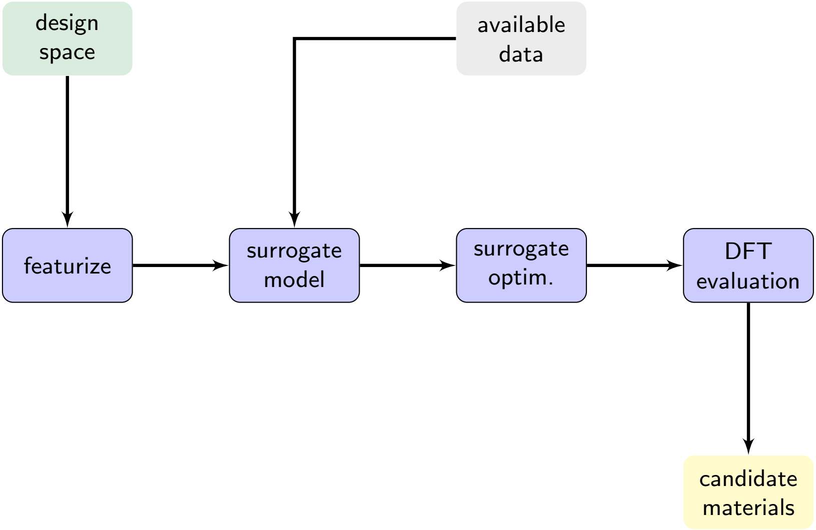
Algorithmic chemical discovery



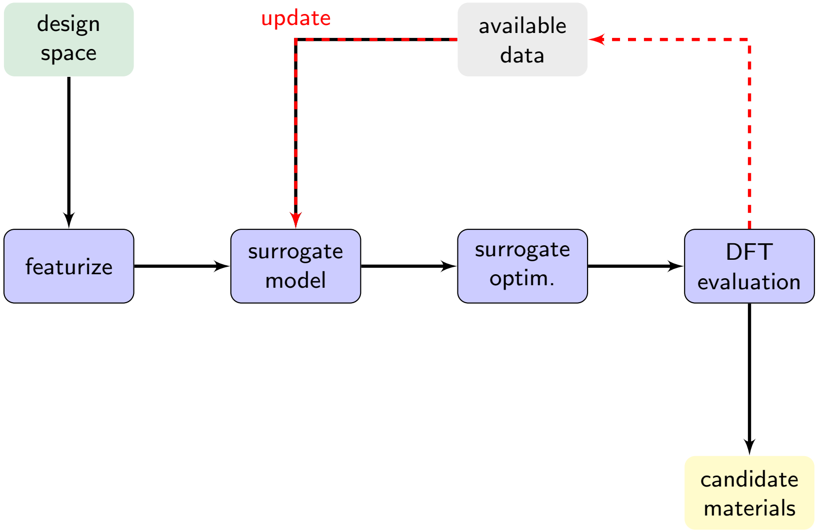
Algorithmic chemical discovery



Algorithmic chemical discovery



Algorithmic chemical discovery

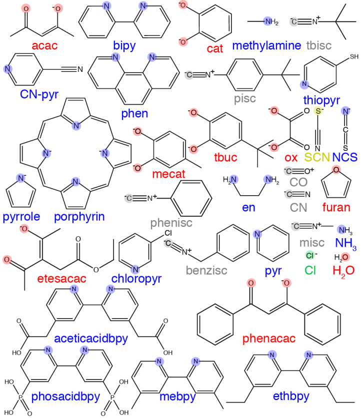


First-principles calculations

train on $\sim$ 100–2000 DFT calculations:

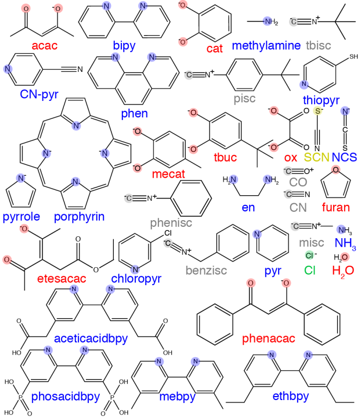
First-principles calculations

train on ∼ 100–2000 DFT calculations:



First-principles calculations

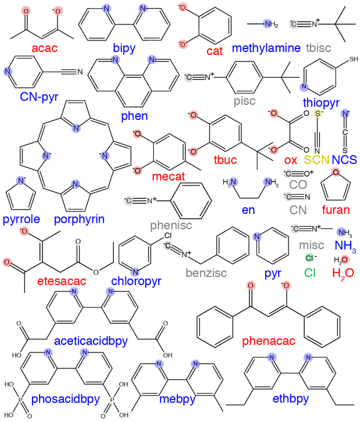
train on ∼ 100–2000 DFT calculations:

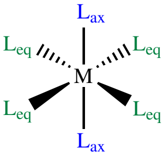


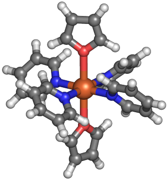
Cr	Mn	Fe	Co

First-principles calculations

train on ~ 100–2000 DFT calculations:



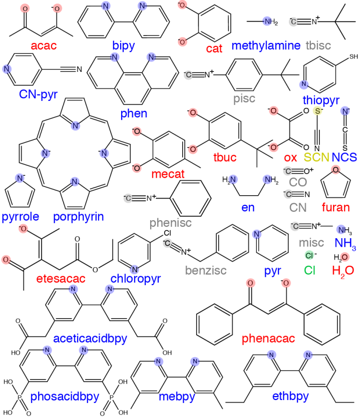


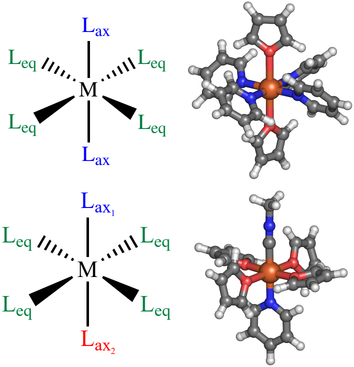


Cr	Mn	Fe	Co

First-principles calculations

train on ∼ 100–2000 DFT calculations:



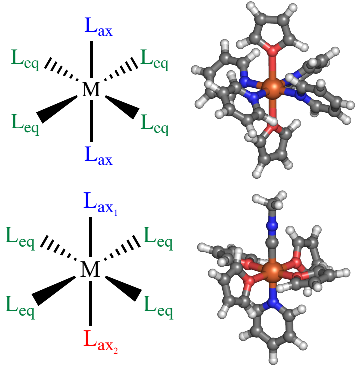


Cr	Mn	Fe	Co

First-principles calculations

train on ∼ 100–2000 DFT calculations:

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



Cr	Mn	Fe	Co

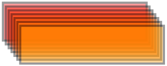
First-principles calculations

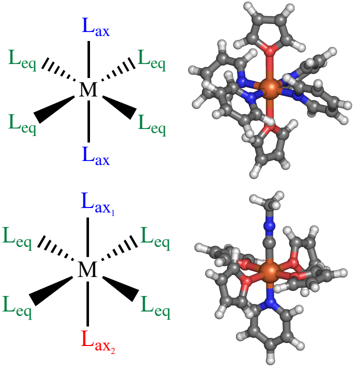
train on ~ 100–2000 DFT calculations:

Details:

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

HF exchange varied 0–30%

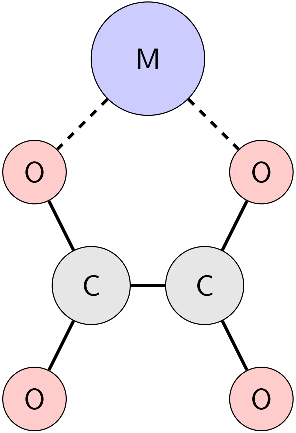




Cr	Mn	Fe	Co

Featurization with RACs

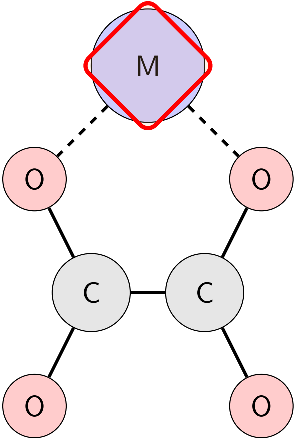
Graph-based features designed for TM complexes



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

Featurization with RACs

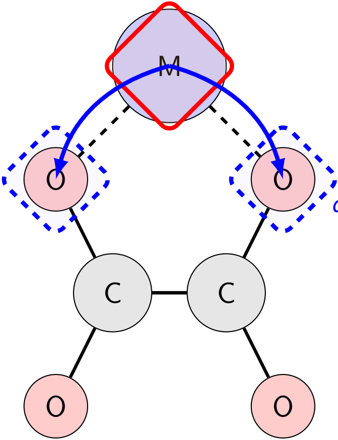
Graph-based features designed for TM complexes



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

Featurization with RACs

Graph-based features designed for TM complexes

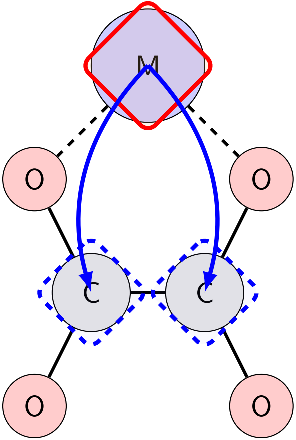


$d_1 : \sum_{M,O} Z_M Z_O$

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

Featurization with RACs

Graph-based features designed for TM complexes



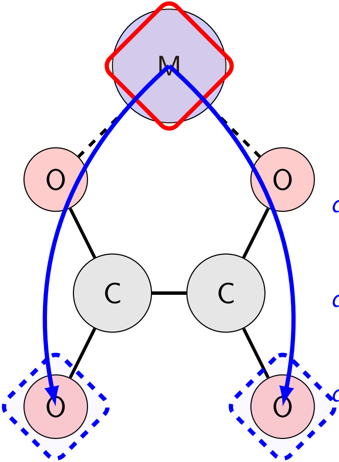
$$d_1 : \sum_{M,O} Z_M Z_O$$

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

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

Featurization with RACs

Graph-based features designed for TM complexes



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

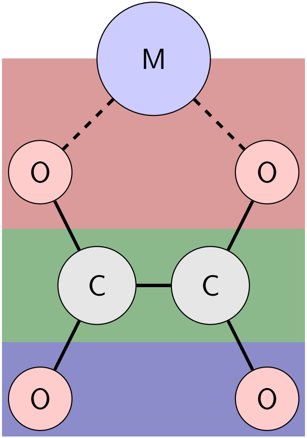
$d_2 : \sum_{M,C} Z_M Z_C$

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

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

Featurization with RACs

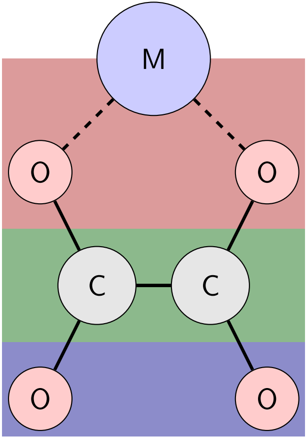
Graph-based features designed for TM complexes

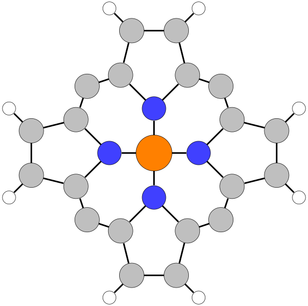


Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

Featurization with RACs

Graph-based features designed for TM complexes

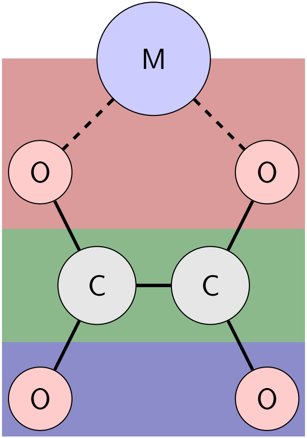




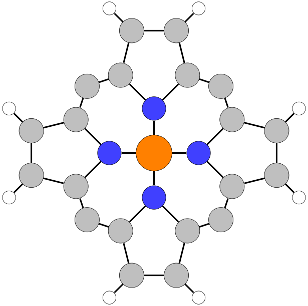
Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

Featurization with RACs

Graph-based features designed for TM complexes



spin splitting ΔE_{H-L}

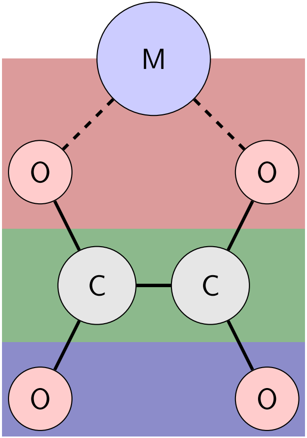


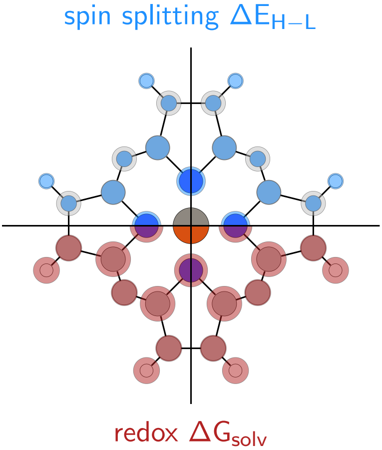
redox ΔG_{solv}

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

Featurization with RACs

Graph-based features designed for TM complexes

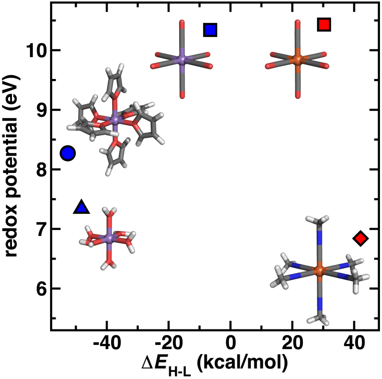


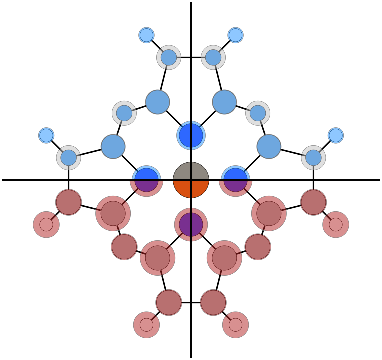


Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

Featurization with RACs

Graph-based features designed for TM complexes





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

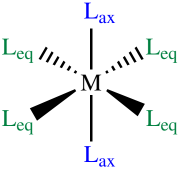
Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

Property inference with ANNs

Estimate properties using small artifical neural networks (ANNs)

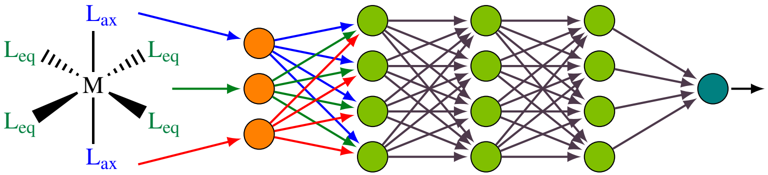
Property inference with ANNs

Estimate properties using small artifical neural networks (ANNs)



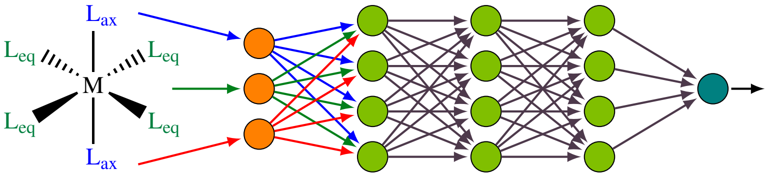
Property inference with ANNs

Estimate properties using small artifical neural networks (ANNs)

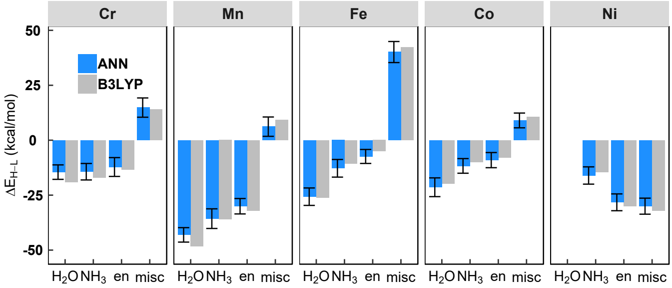


Property inference with ANNs

Estimate properties using small artifical neural networks (ANNs)



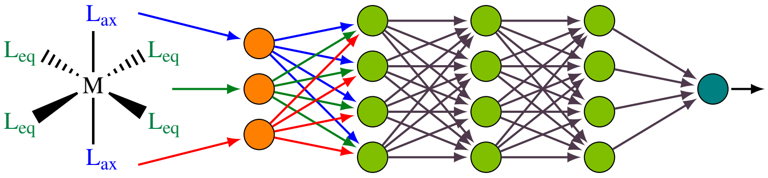
spin splitting energies



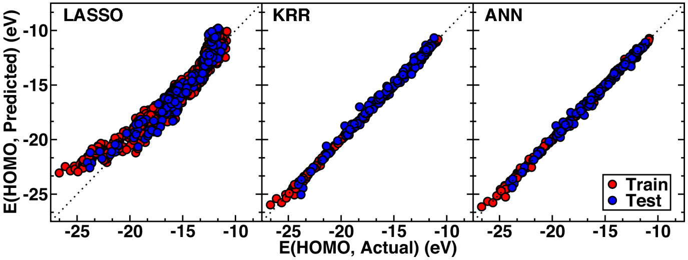
Janet, J.P. and Kulik, H.J., *Chem. Sci.*, 8:5137–5152, 2017. 50 × 50 tanh nodes fully connected ANN

Property inference with ANNs

Estimate properties using small artifical neural networks (ANNs)



frontier orbital properties

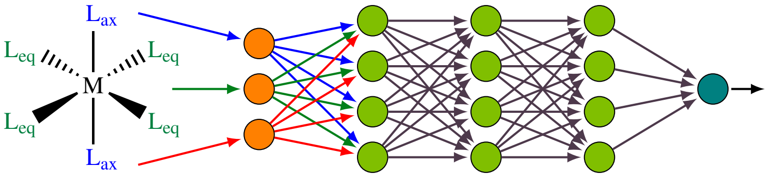


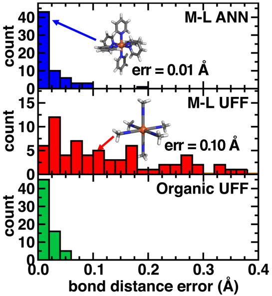
ANN: 500 × 500 ReLU nodes, fully connected

Nandy, A. *et al.*, *Ind. Eng. Chem. Res.*, 57(42):13973–13986, 2018. KRR: RFB + feature selection

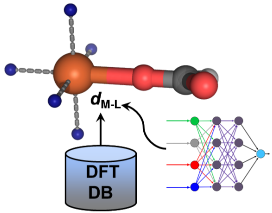
Property inference with ANNs

Estimate properties using small artifical neural networks (ANNs)





DFT equilibrium bond lengths



Janet, J.P. *et al.*, *Ind. Eng. Chem. Res.*, 56(17):4898–4910, 2017.

Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

300 × 200 × 200 tanh nodes, fully connected

Beyond prediction: live job management

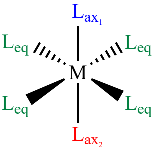
In high-throughput DFT screening, job failure is a frequent issue:

Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management

In high-throughput DFT screening, job failure is a frequent issue:

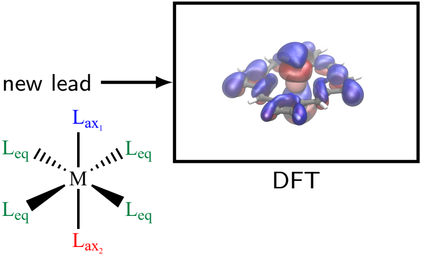
new lead



Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management

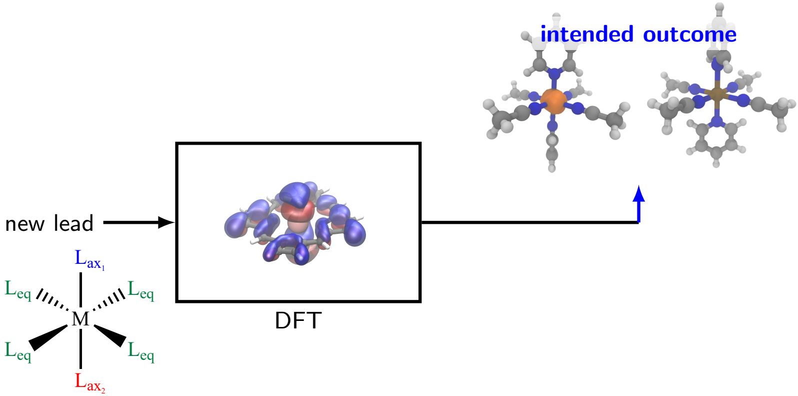
In high-throughput DFT screening, job failure is a frequent issue:



Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management

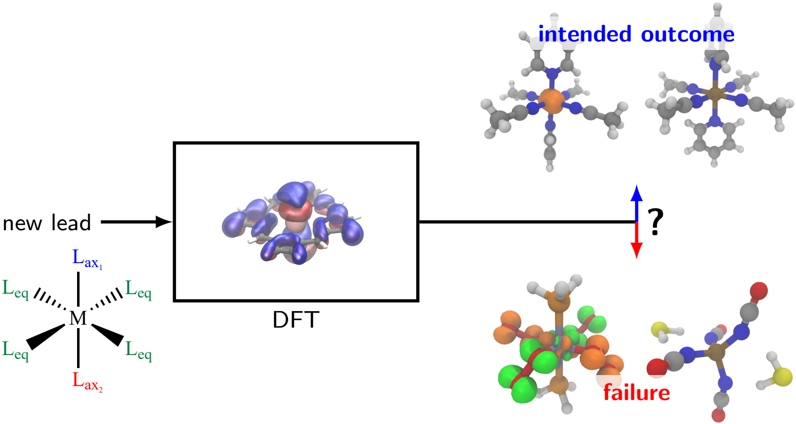
In high-throughput DFT screening, job failure is a frequent issue:



Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

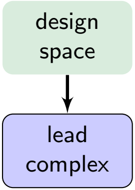
Beyond prediction: live job management

In high-throughput DFT screening, job failure is a frequent issue:



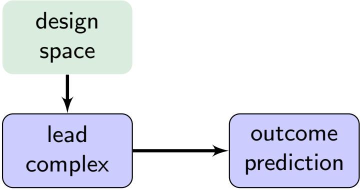
Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management



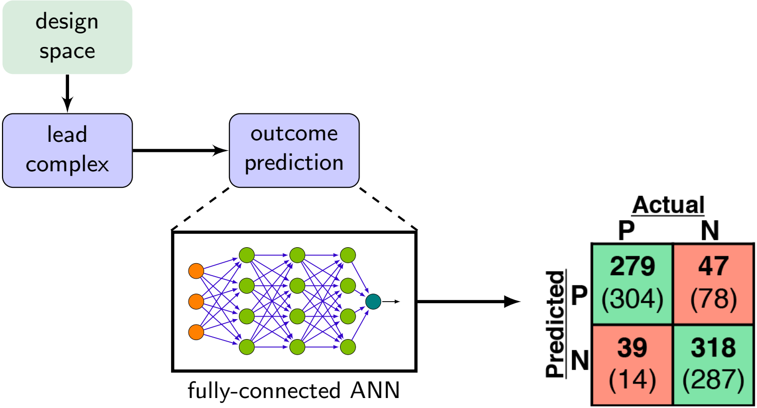
Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management



Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

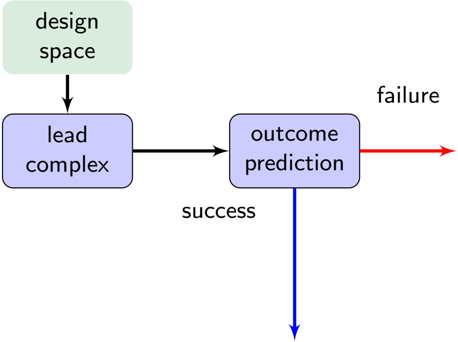
Beyond prediction: live job management



2 $\times$ 200 ReLU nodes, fully connected

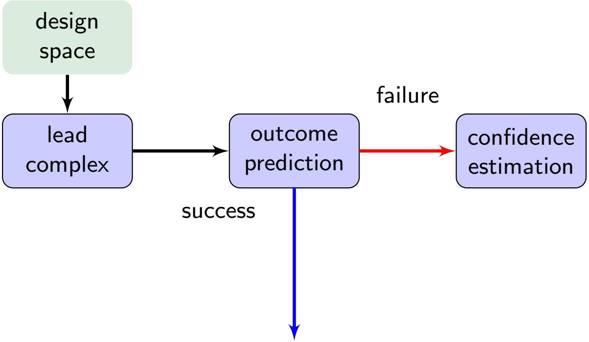
Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management



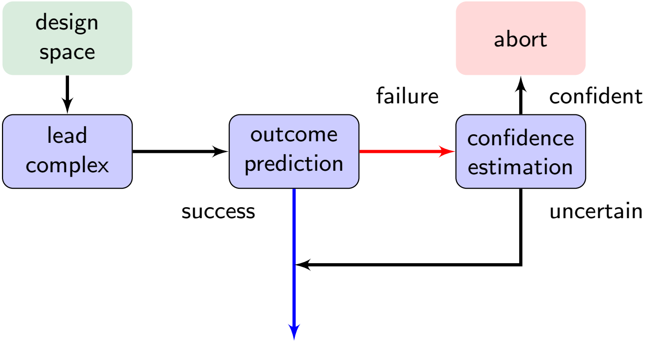
Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management



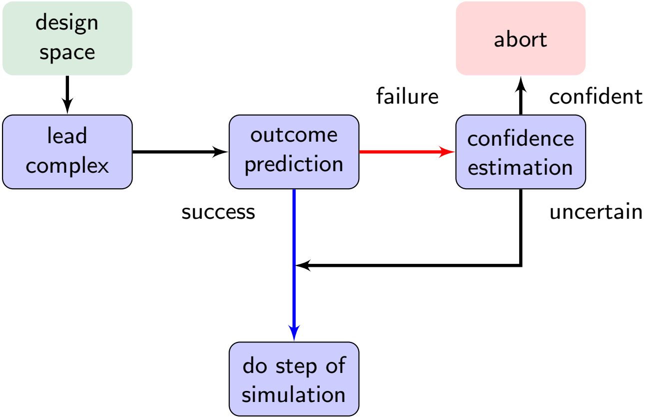
Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management



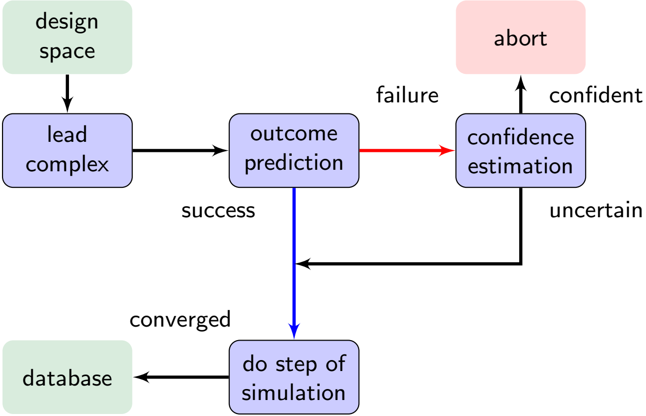
Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management



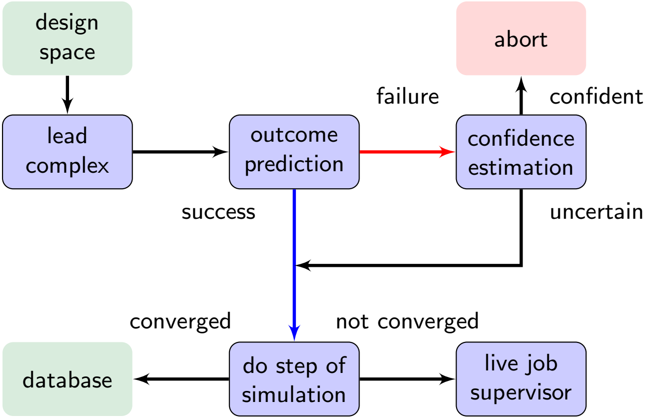
Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management



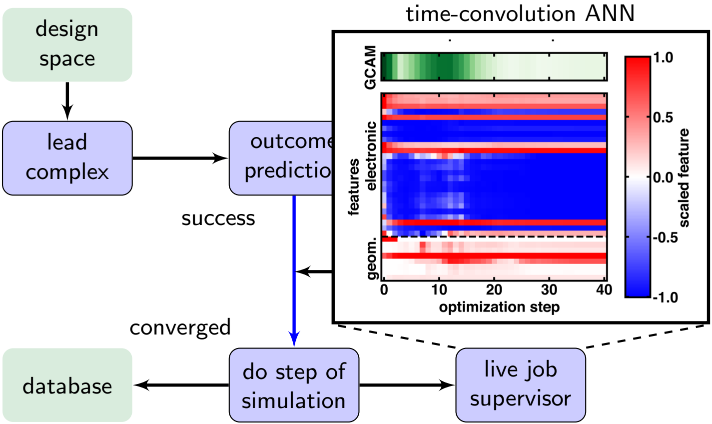
Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management



Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

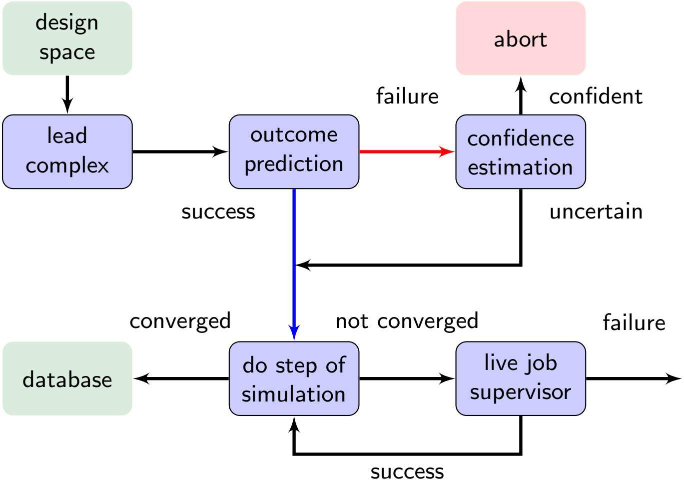
Beyond prediction: live job management



45 time convolution + 2 × 100 ReLU nodes

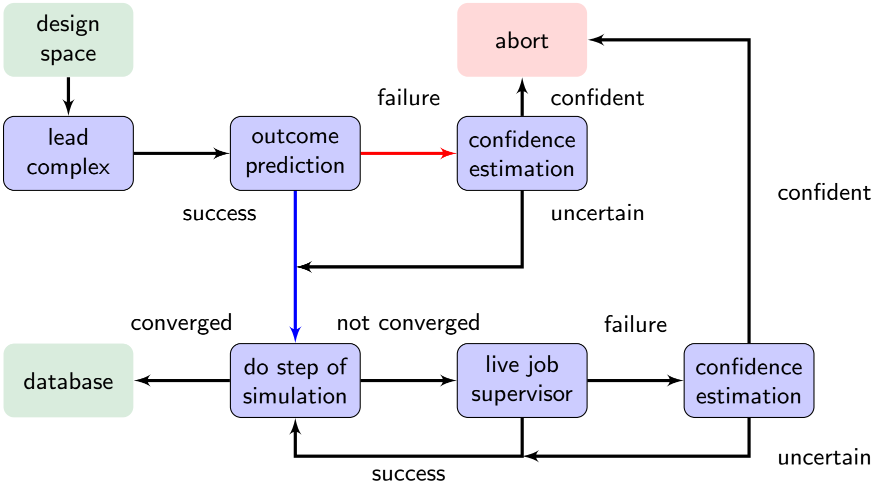
Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management



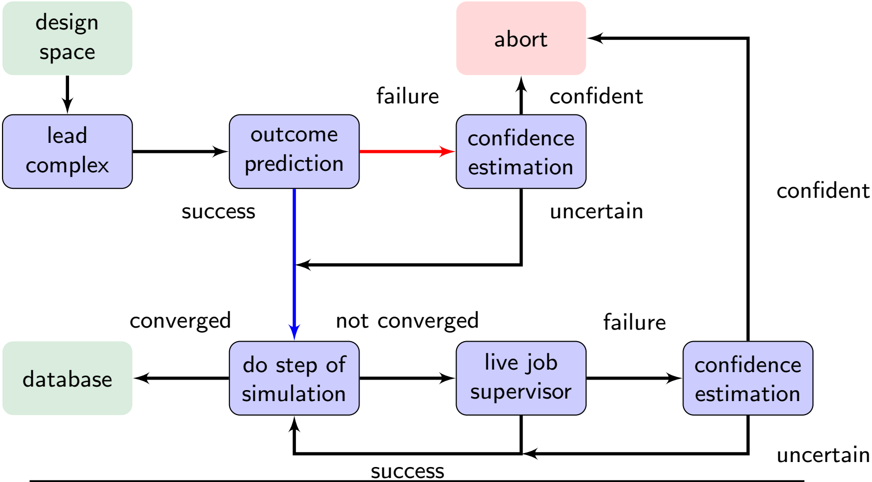
Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management



Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Beyond prediction: live job management



This leads to about **40% time savings** and can abort almost all failures.

Duan, C., Janet, J.P. *et al.*, *J. Chem. Theory. Comp.*, 15(4):2331–2345, 2019.

Model transferability

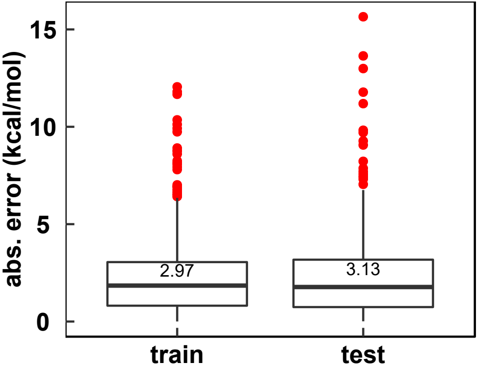
Test-set performance is not necessarily a good metric for general transferability:

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

50 $\times$ 50 tanh nodes fully connected ANN

Model transferability

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

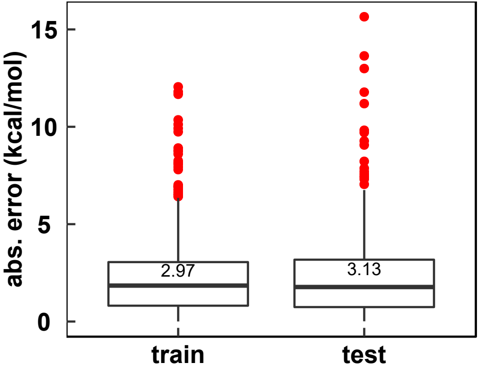


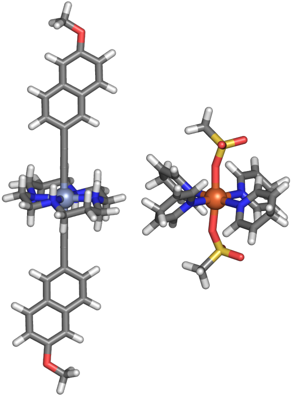
Janet, J.P., and Kulik, H.J., *Chem. Sci.*, 8:5137–5152, 2017.

50 × 50 tanh nodes fully connected ANN

Model transferability

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



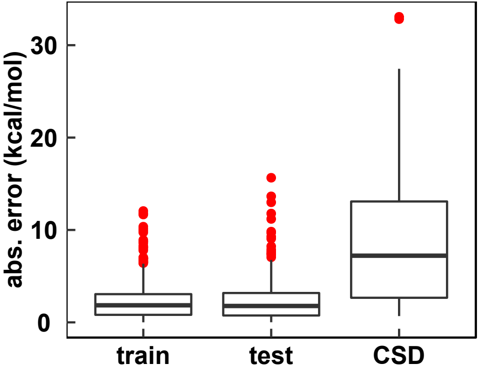


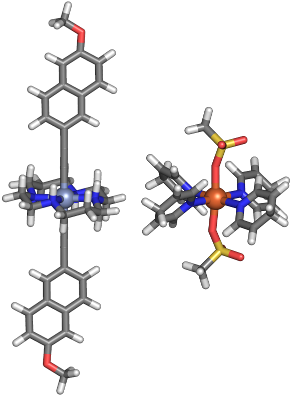
Janet, J.P., and Kulik, H.J., *Chem. Sci.*, 8:5137–5152, 2017.

50 × 50 tanh nodes fully connected ANN

Model transferability

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



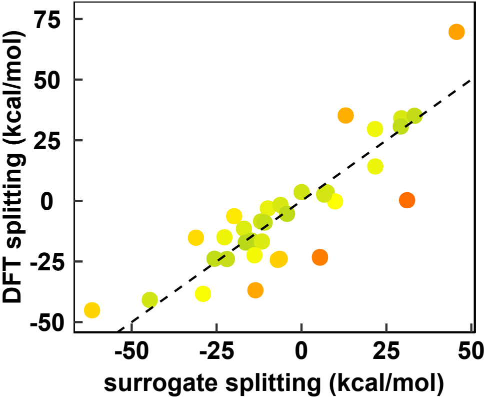


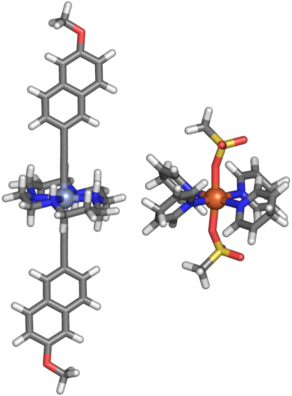
Janet, J.P., and Kulik, H.J., *Chem. Sci.*, 8:5137–5152, 2017.

50 × 50 tanh nodes fully connected ANN

Model transferability

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





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

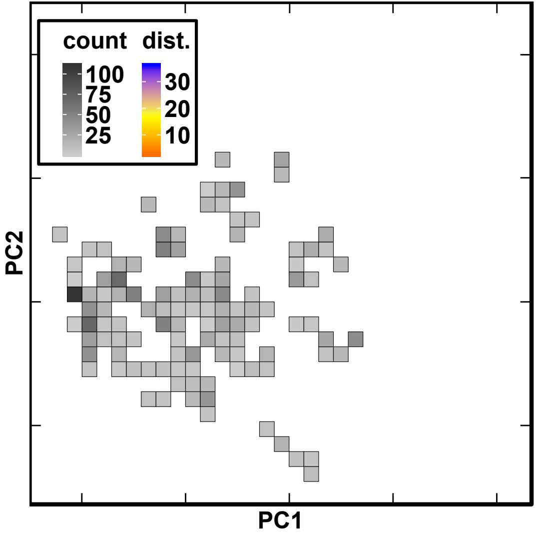
50 × 50 tanh nodes fully connected ANN

How far can we extrapolate?

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

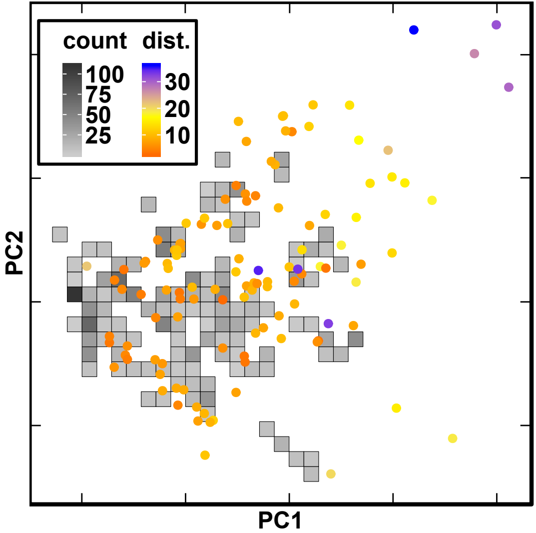
How far can we extrapolate?

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



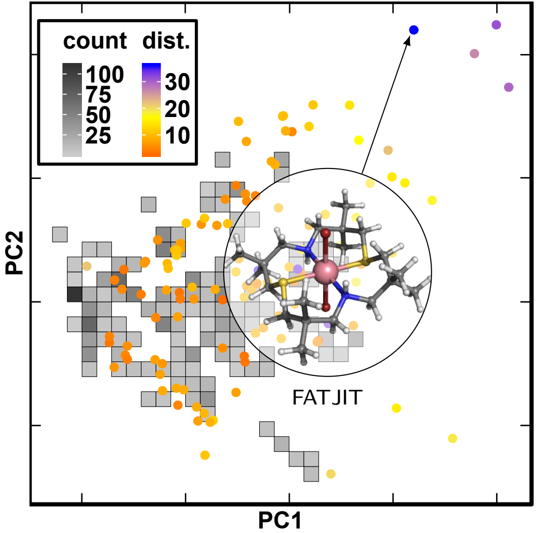
How far can we extrapolate?

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



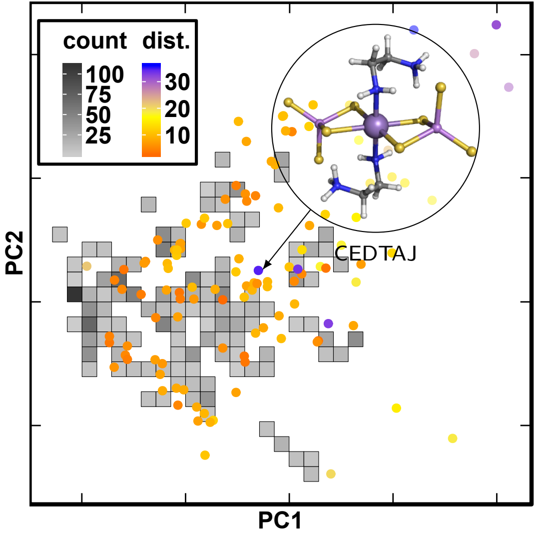
How far can we extrapolate?

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



How far can we extrapolate?

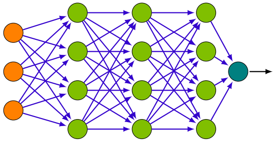
'Out-of-distribution' test: spin-splitting energies of 116 structures from the CSD, from training-like to very different.

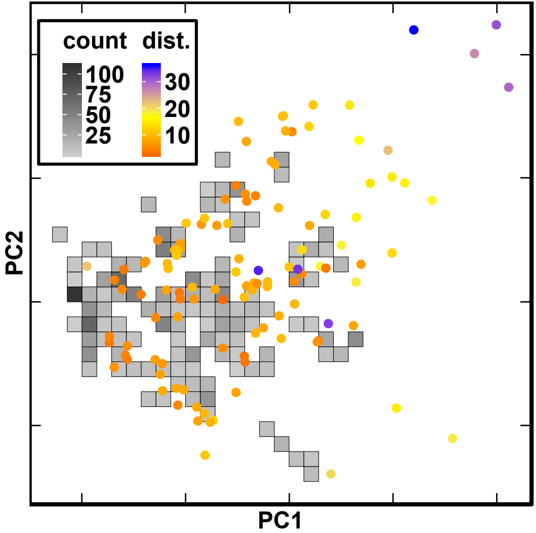


How far can we extrapolate?

'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:

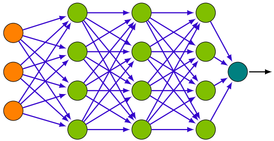


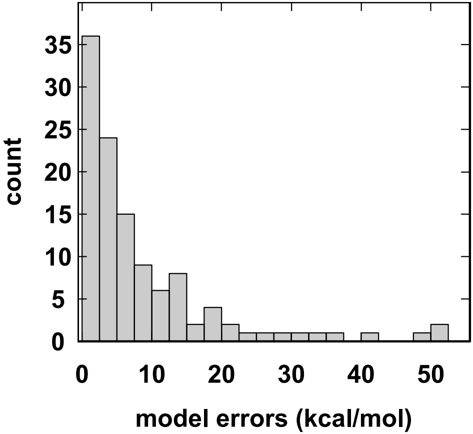


How far can we extrapolate?

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

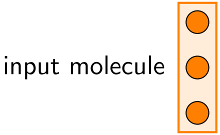
Train 3-layer fully connected ANN on 1900 DFT results on simple ligands:



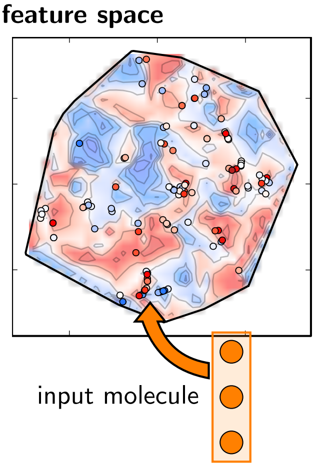


Latent distance similarity

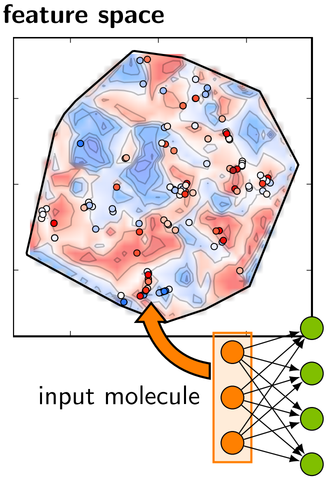
Latent distance similarity



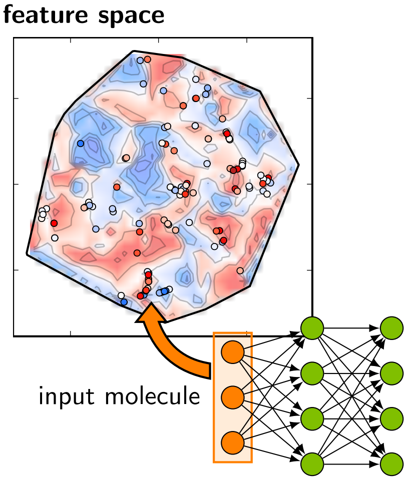
Latent distance similarity



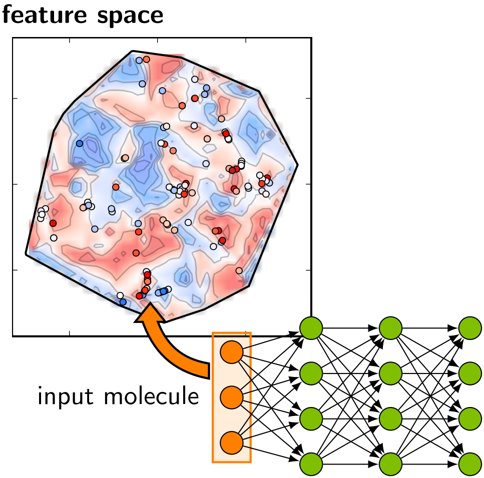
Latent distance similarity



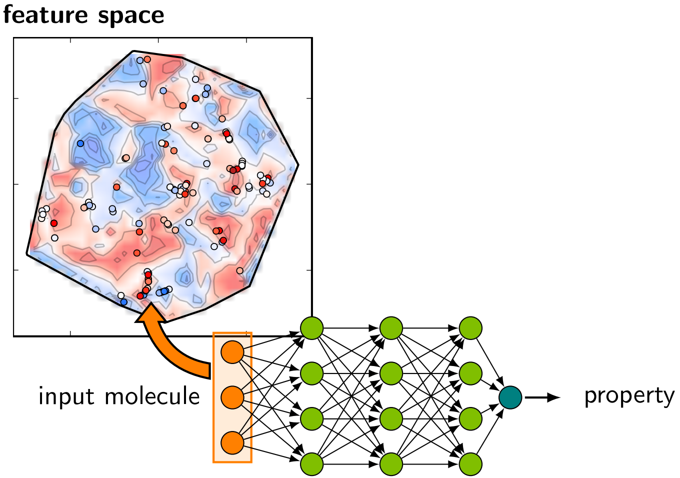
Latent distance similarity



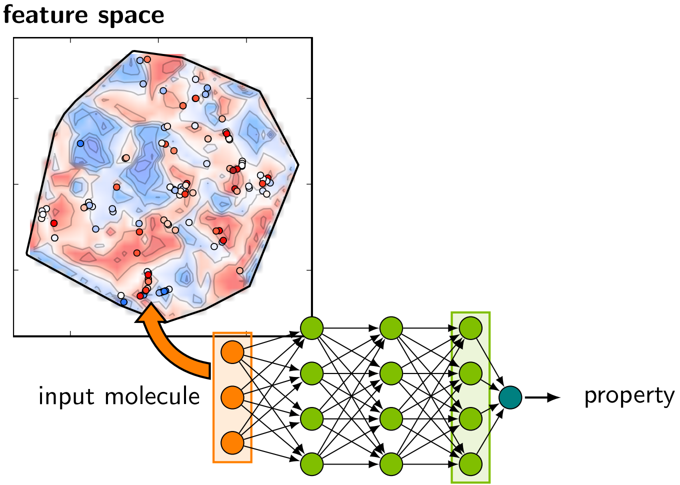
Latent distance similarity



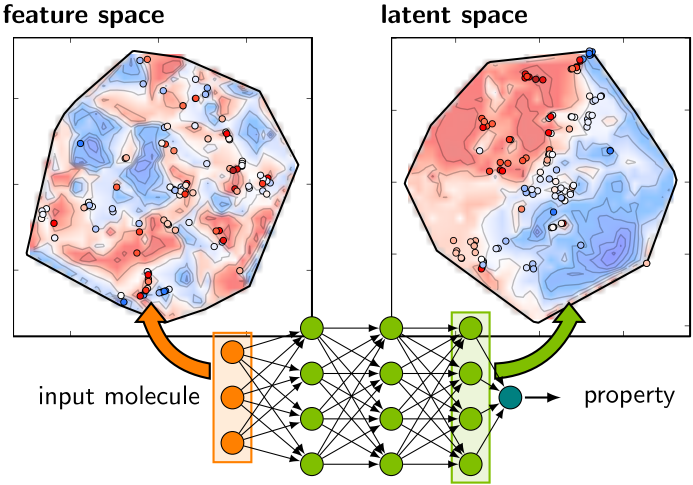
Latent distance similarity



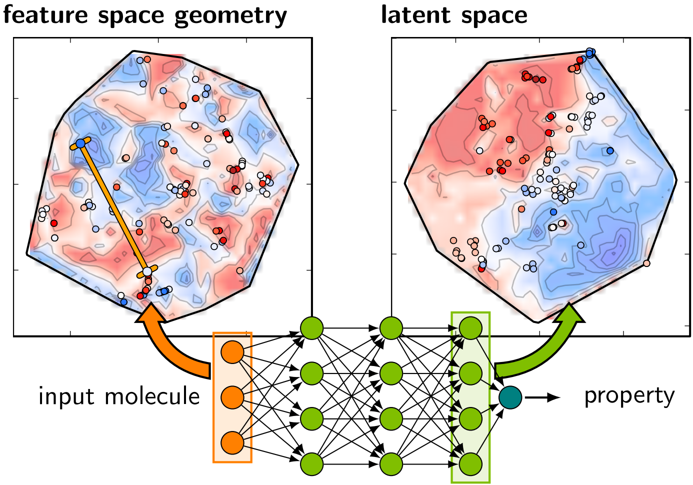
Latent distance similarity



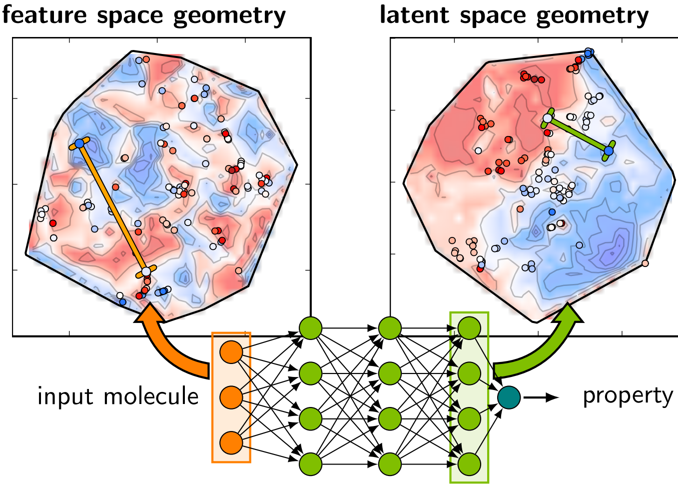
Latent distance similarity



Latent distance similarity



Latent distance similarity



From distance to energy

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

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

From distance to energy

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

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

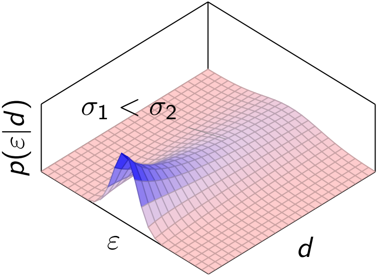
Estimate σ_1, σ_2 by max log likelihood, stably estimated even using few out-of-sample points

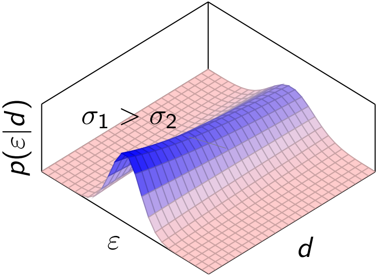
From distance to energy

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



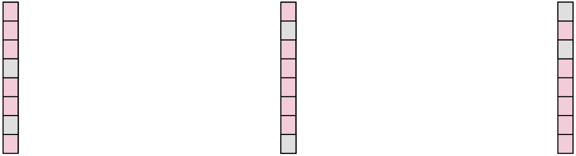


Other UQ metrics

1) Data-sampling ensembles:

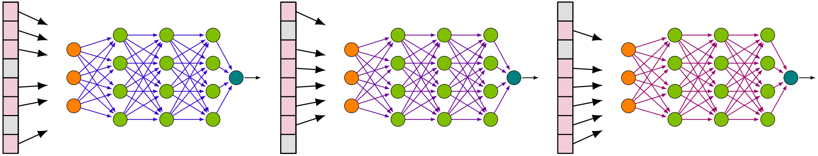
Other UQ metrics

1) Data-sampling ensembles:



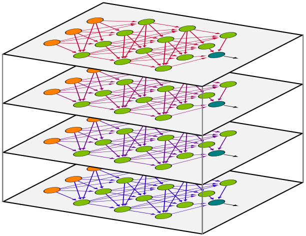
Other UQ metrics

1) Data-sampling ensembles:



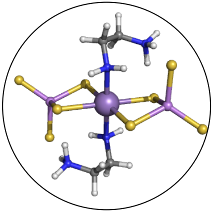
Other UQ metrics

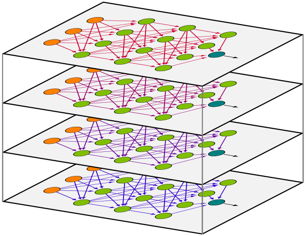
1) Data-sampling ensembles:



Other UQ metrics

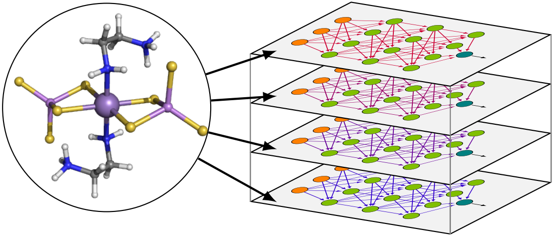
1) Data-sampling ensembles:





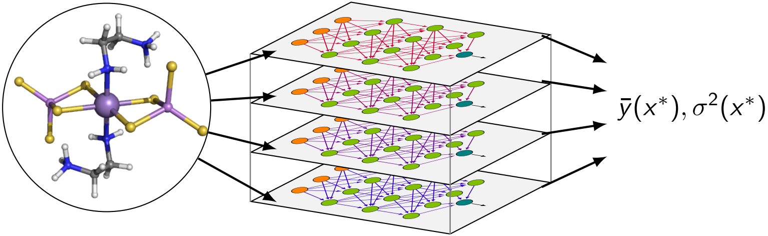
Other UQ metrics

1) Data-sampling ensembles:



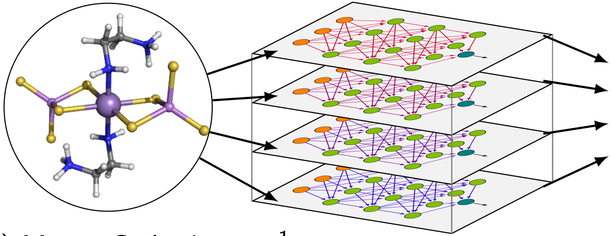
Other UQ metrics

1) Data-sampling ensembles:



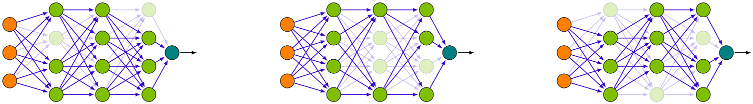
Other UQ metrics

1) Data-sampling ensembles:



$\bar{y}(x^*), \sigma^2(x^*)$

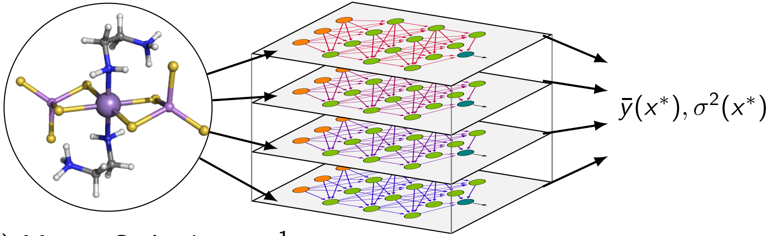
2) Monte Carlo dropout[1]:



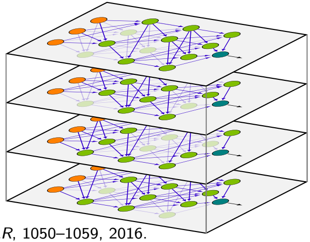
[1]:Gal, Y. and Ghahramani, Z., *ICMLR*, 1050–1059, 2016.

Other UQ metrics

1) Data-sampling ensembles:



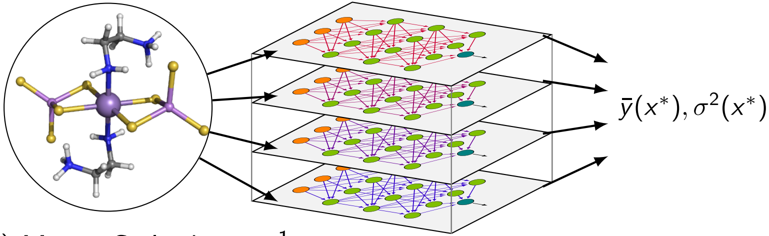
2) Monte Carlo dropout[1]:



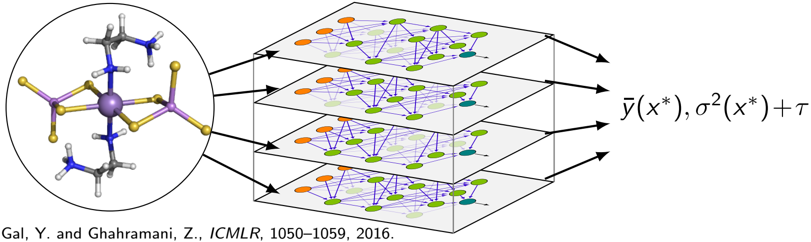
[1]:Gal, Y. and Ghahramani, Z., *ICMLR*, 1050–1059, 2016.

Other UQ metrics

1) Data-sampling ensembles:

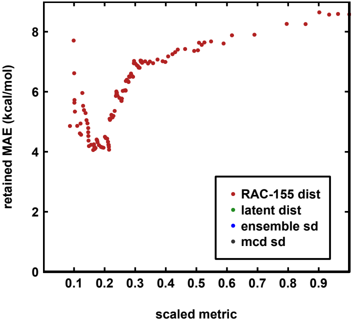


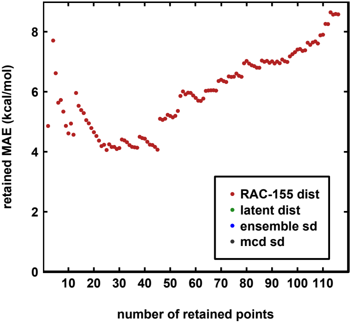
2) Monte Carlo dropout[1]:



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

How do these distributions compare?



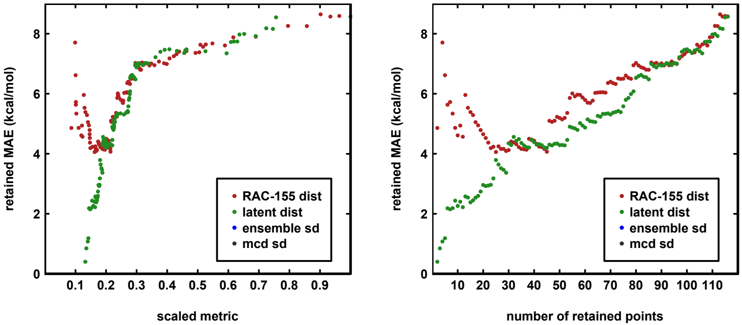


Janet, J.P., *et al.*, *Chem. Sci.*, 10:7913–7922 , 2019.

3×200 ReLU nodes, fully connected

How do these distributions compare?

latent distances are superior to feature space distances

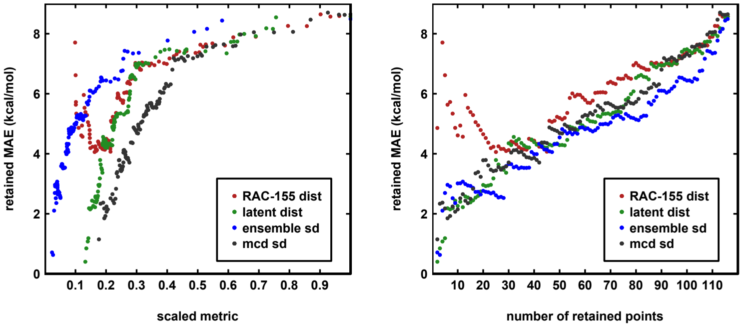


Janet, J.P., *et al.*, *Chem. Sci.*, 10:7913–7922 , 2019.

3×200 ReLU nodes, fully connected

How do these distributions compare?

latent distances are superior to feature space distances



comparable with ensembles and mc dropout

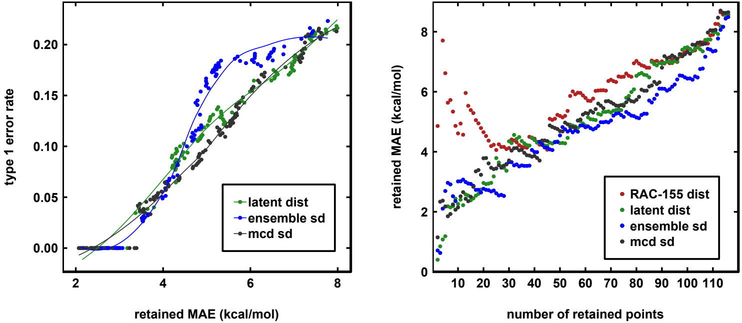
Janet, J.P., *et al.*, *Chem. Sci.*, 10:7913–7922 , 2019.

3 × 200 ReLU nodes, fully connected

How do these distributions compare?

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

latent distances are superior to feature space distances



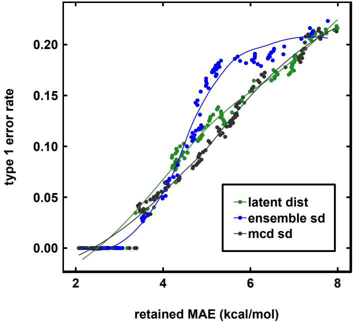
comparable with ensembles and mc dropout

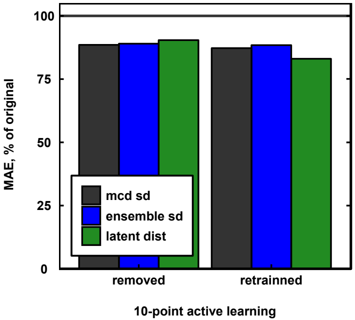
Janet, J.P., *et al.*, *Chem. Sci.*, 10:7913–7922 , 2019. 3 × 200 ReLU nodes, fully connected

How do these distributions compare?

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

latent distances are superior to feature space distances





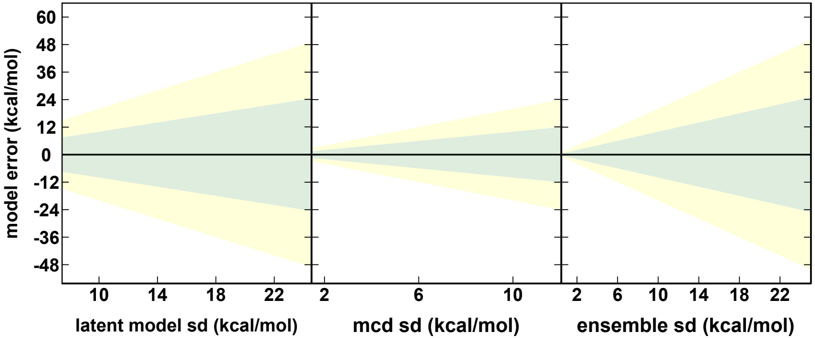
comparable with ensembles and mc dropout

Janet, J.P., *et al.*, *Chem. Sci.*, 10:7913–7922 , 2019.

3×200 ReLU nodes, fully connected

How do these distributions compare?

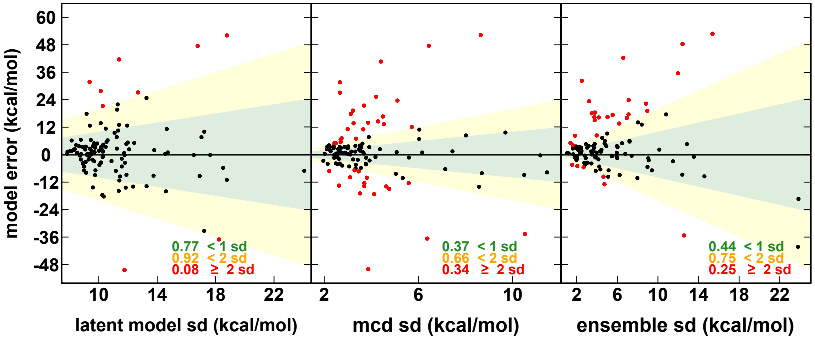
Comparison in energy units: $\varepsilon(d) \sim \mathcal{N}\left(0, \sigma_1^2 + d\sigma_2^2\right)$



Janet, J.P., *et al.*, *Chem. Sci.*, 10:7913–7922 , 2019. 3×200 ReLU nodes, fully connected

How do these distributions compare?

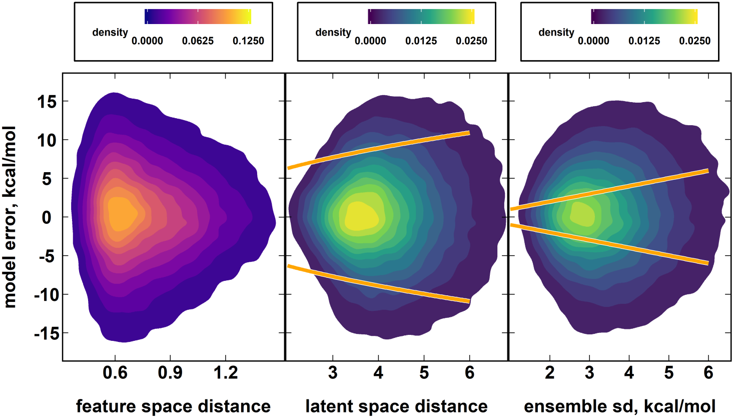
Comparison in energy units: $\varepsilon(d) \sim \mathcal{N}\left(0, \sigma_1^2 + d\sigma_2^2\right)$



Janet, J.P., *et al.*, *Chem. Sci.*, 10:7913–7922 , 2019. 3×200 ReLU nodes, fully connected

How do these distributions compare?

Comparison in energy units: QM9 atomization energy data



Janet, J.P., *et al.*, *Chem. Sci.*, 10:7913–7922 , 2019. 2 × 100 ReLU nodes, skip connections

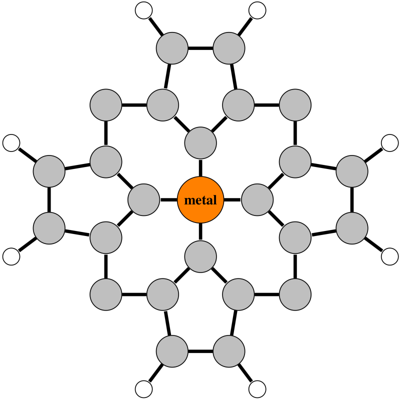
Targeted data acquisition

Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:

Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

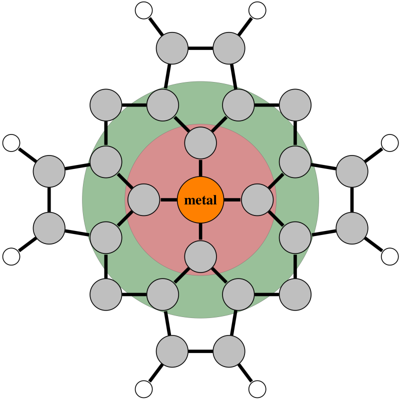
Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:



Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

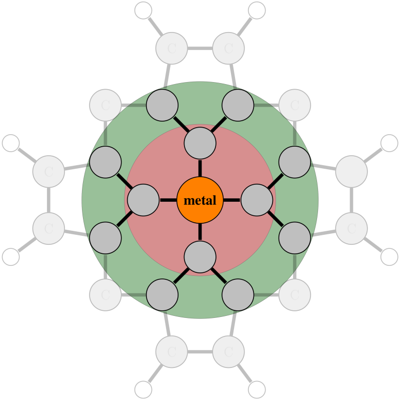
Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:



Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

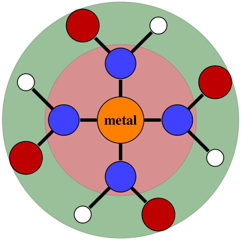
Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:



Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

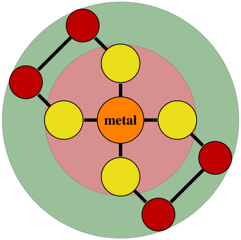
Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:



Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

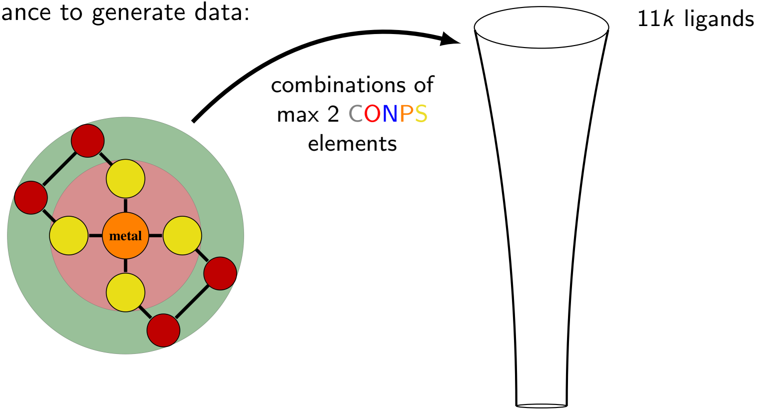
Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:



Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

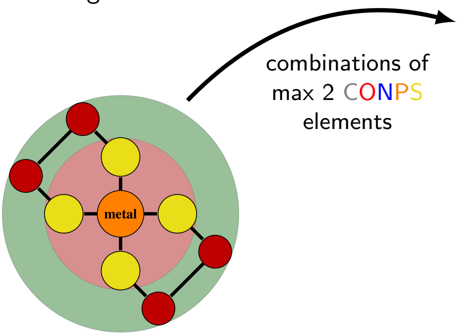
Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:



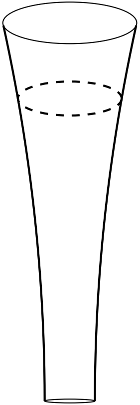
Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:



combinations of
max 2 CONPS
elements

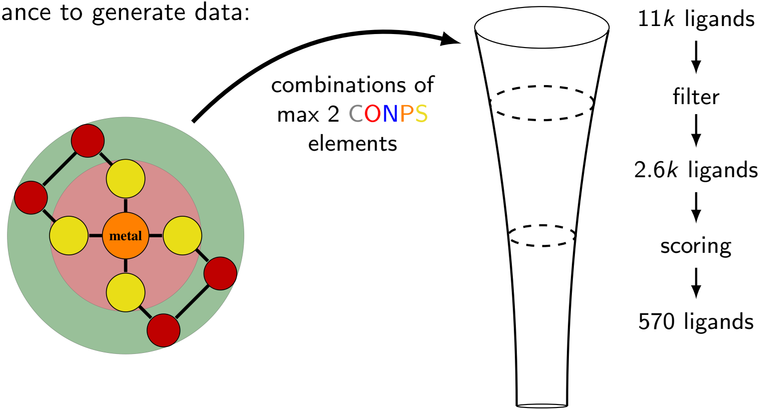


$11k$ ligands
filter
$2.6k$ ligands

Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

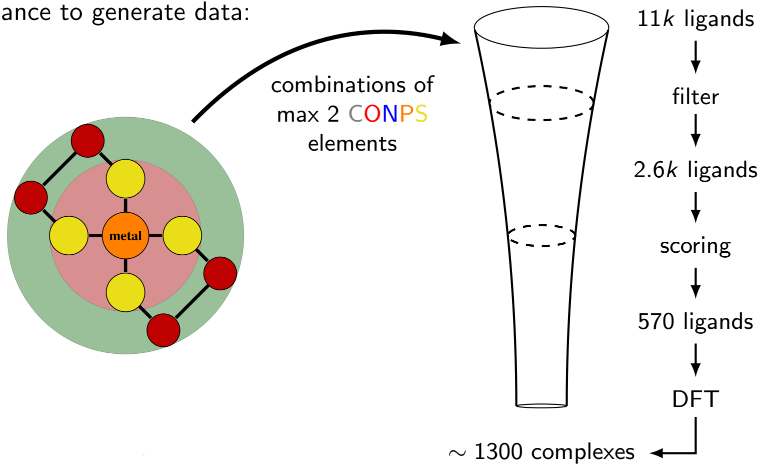
Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:



Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

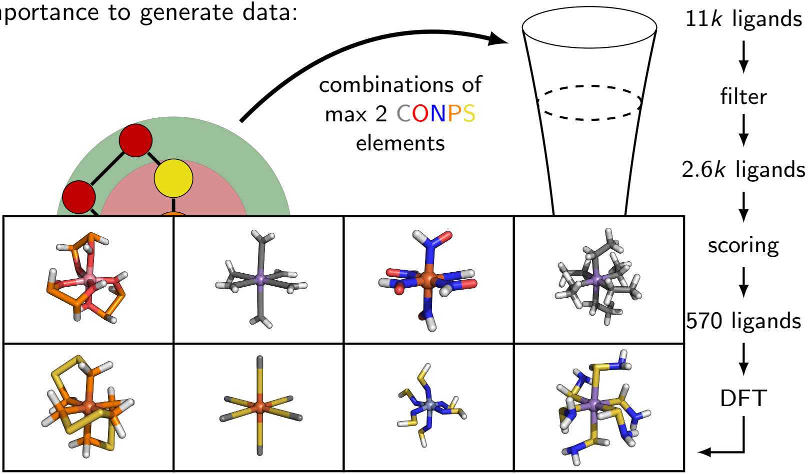
Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:



Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

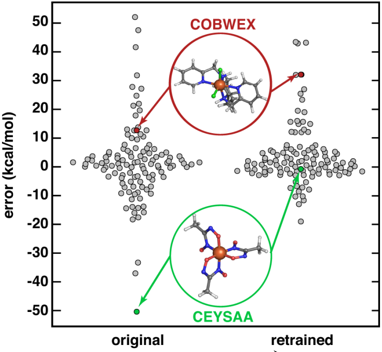
Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:

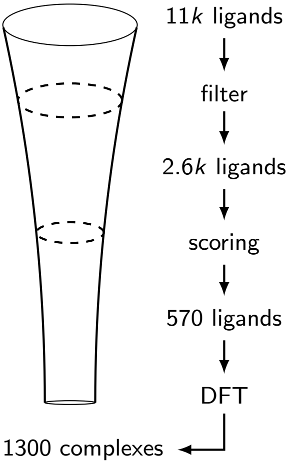


Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:





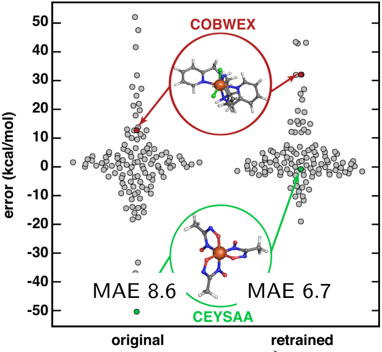
add to surrogate model

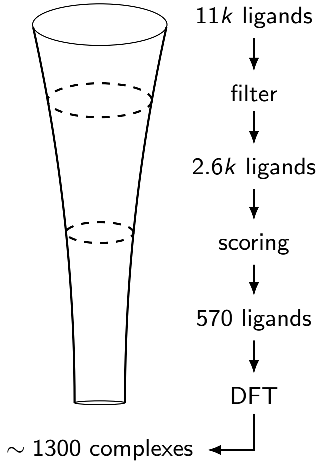
3×100 tanh nodes, fully connected ANN

Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Targeted data acquisition

Lack of data is a persistent issue. We can exploit knowledge of feature importance to generate data:





add to surrogate model

3 $\times$ 100 tanh nodes, fully connected ANN

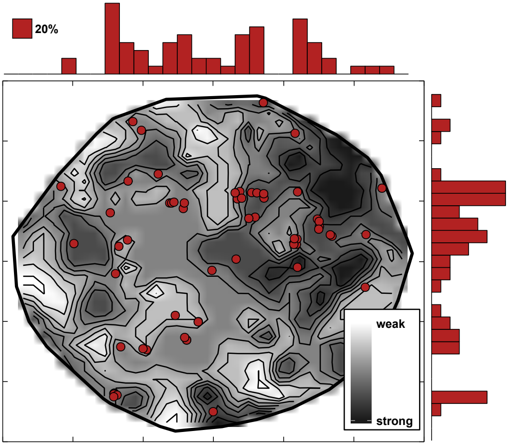
Gugler, S., Janet, J.P., and Kulik, H.J., *Mol. Sys. Des. Eng.*, Advance Article, 2019.

Hedging against DFT uncertainty

Because we have trained our models with varying exact exchange, we can tune functionals for design:

Hedging against DFT uncertainty

Because we have trained our models with varying exact exchange, we can tune functionals for design:

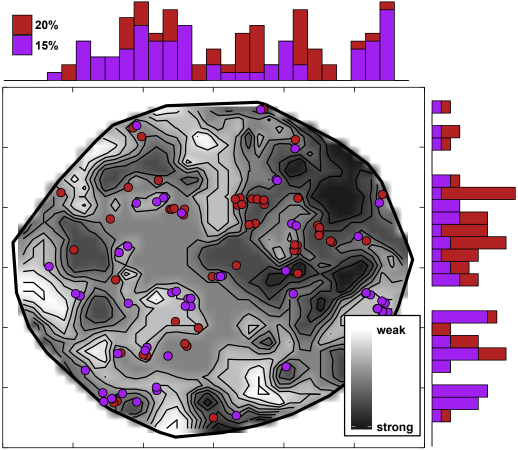


Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

50 × 50 tanh nodes fully connected ANN

Hedging against DFT uncertainty

Because we have trained our models with varying exact exchange, we can tune functionals for design:

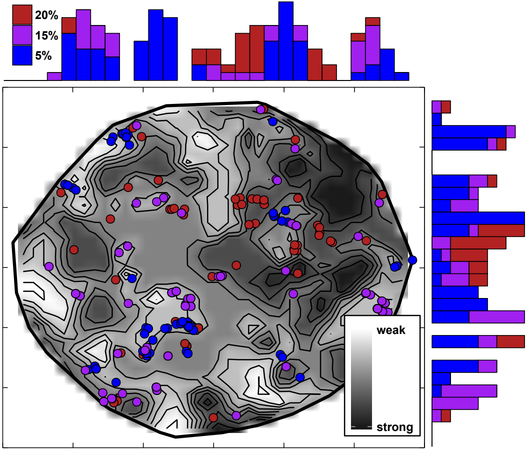


Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

50 × 50 tanh nodes fully connected ANN

Hedging against DFT uncertainty

Because we have trained our models with varying exact exchange, we can tune functionals for design:



Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

50 × 50 tanh nodes fully connected ANN

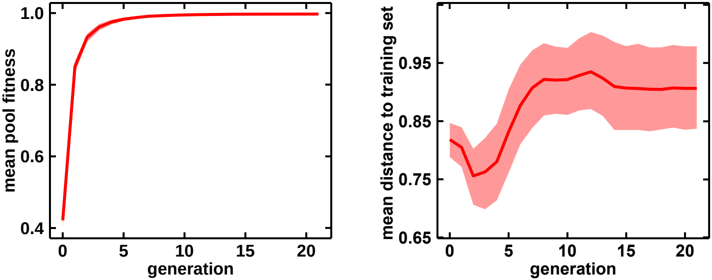
Using models for discovery

We utilize evolutionary algorithms to conduct design, guided by uncertainty metrics

Using models for discovery

We utilize evolutionary algorithms to conduct design, guided by uncertainty metrics

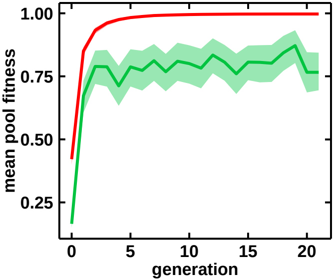
$$F_{s+d}(x) := \exp\left[-\left(\frac{\Delta E_{\text{H-L}}(x)}{p_{\Delta E_{\text{H-L}}}}\right)^2\right]$$

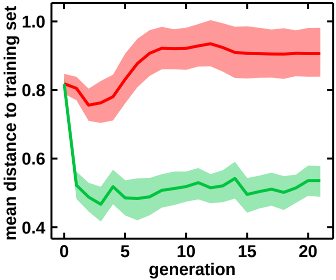


Using models for discovery

We utilize evolutionary algorithms to conduct design, guided by uncertainty metrics

$$F_{s+d}(x) := \exp\left[-\left(\frac{\Delta E_{\text{H-L}}(x)}{p_{\Delta E_{\text{H-L}}}}\right)^2\right] \exp\left[-\left(\frac{d(x)}{p_d}\right)^2\right]$$





Using models for discovery

We utilize evolutionary algorithms to conduct design, guided by uncertainty metrics

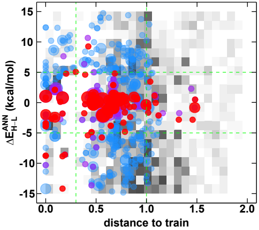
$$F_{s+d}(x) := \exp\left[-\left(\frac{\Delta E_{\text{H-L}}(x)}{p_{\Delta E_{\text{H-L}}}}\right)^2\right] \exp\left[-\left(\frac{d(x)}{p_d}\right)^2\right]$$

Using models for discovery

We utilize evolutionary algorithms to conduct design, guided by uncertainty metrics

$$F_{s+d}(x) := \exp\left[-\left(\frac{\Delta E_{\text{H-L}}(x)}{p_{\Delta E_{\text{H-L}}}}\right)^2\right] \exp\left[-\left(\frac{d(x)}{p_d}\right)^2\right]$$

spin crossover complexes



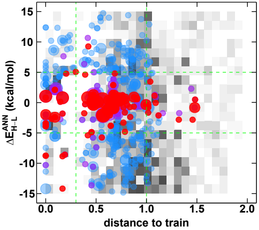
Janet, J.P., Chan, L. and Kulik, H.J., *J. Phys. Chem. Lett.*, 9(5):1064–1071, 2018.

Using models for discovery

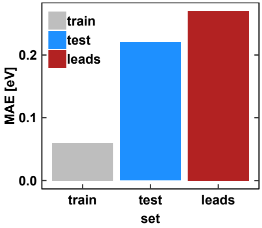
We utilize evolutionary algorithms to conduct design, guided by uncertainty metrics

$$F_{s+d}(x) := \exp\left[-\left(\frac{\Delta E_{\text{H-L}}(x)}{p_{\Delta E_{\text{H-L}}}}\right)^2\right] \exp\left[-\left(\frac{d(x)}{p_d}\right)^2\right]$$

spin crossover complexes



frontier orbital properties



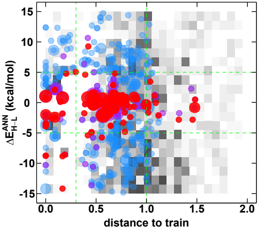
Janet, J.P., Chan, L. and Kulik, H.J., *J. Phys. Chem. Lett.*, 9(5):1064–1071, 2018.
Nandy, A. *et al.*, *Ind. Eng. Chem. Res.*, 57(42):13973–13986, 2018.

Using models for discovery

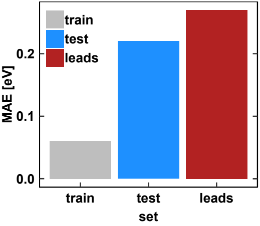
We utilize evolutionary algorithms to conduct design, guided by uncertainty metrics

$$F_{s+d}(x) := \exp\left[-\left(\frac{\Delta E_{\text{H-L}}(x)}{p_{\Delta E_{\text{H-L}}}}\right)^2\right] \exp\left[-\left(\frac{d(x)}{p_d}\right)^2\right]$$

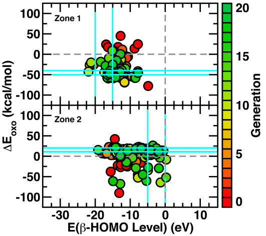
spin crossover complexes



frontier orbital properties



unusual catalytic reaction energies



Janet, J.P., Chan, L. and Kulik, H.J., *J. Phys. Chem. Lett.*, 9(5):1064–1071, 2018.
Nandy, A. *et al.*, *Ind. Eng. Chem. Res.*, 57(42):13973–13986, 2018.
Nandy, A. *et al.*, *ACS Catal.*, 9(9):8243–8255, 2019.

More than just ML

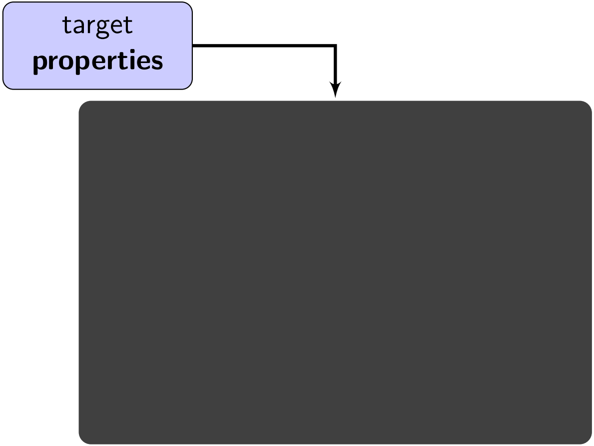
Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

More than just ML

target
properties

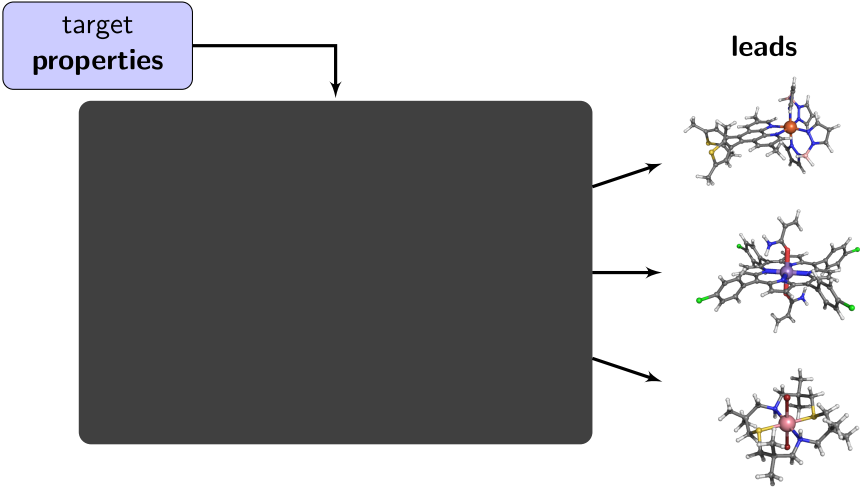
Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

More than just ML



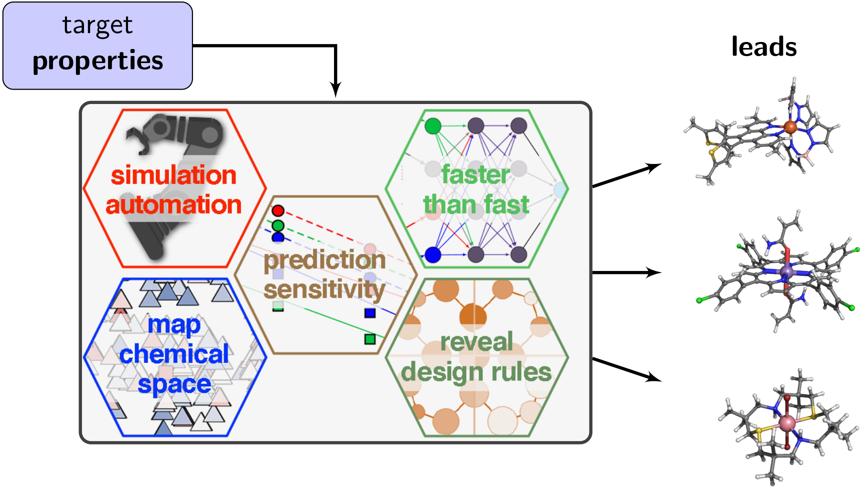
Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

More than just ML



Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

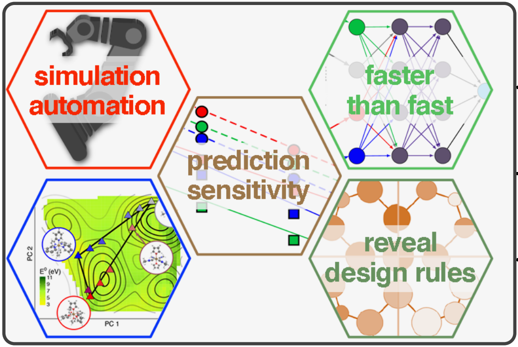
More than just ML



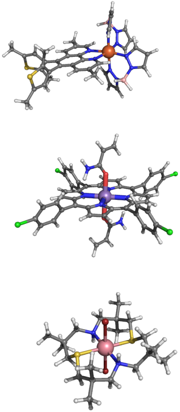
Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

More than just ML

target **properties**



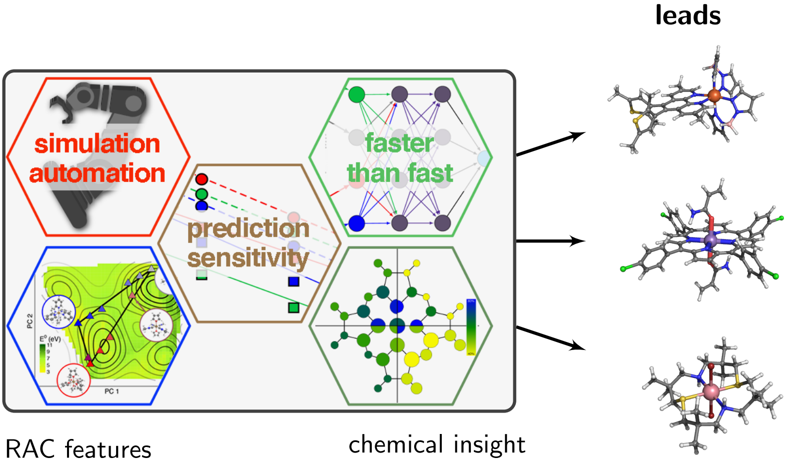
leads



RAC features

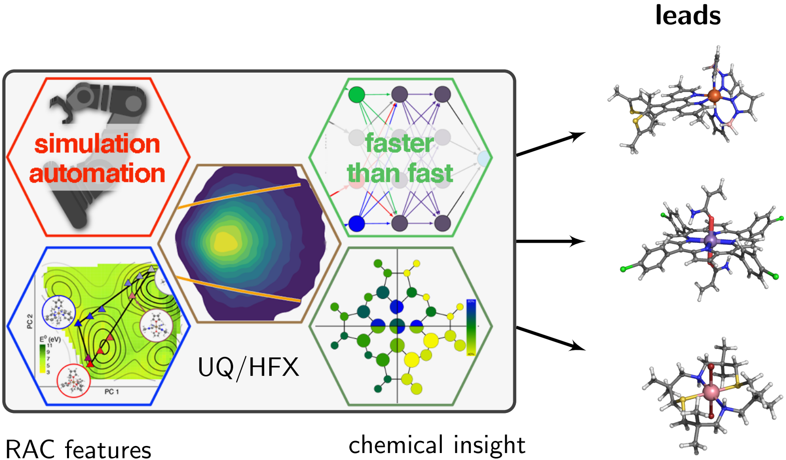
Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

More than just ML



Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

More than just ML



RAC features

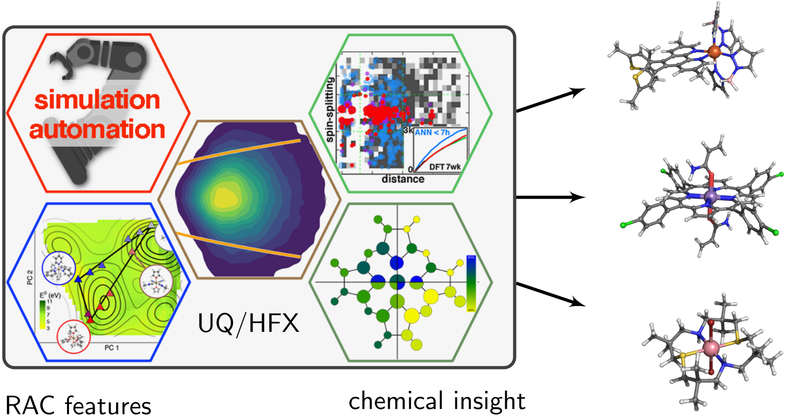
chemical insight

Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

More than just ML

design algorithms

leads



RAC features

chemical insight

Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

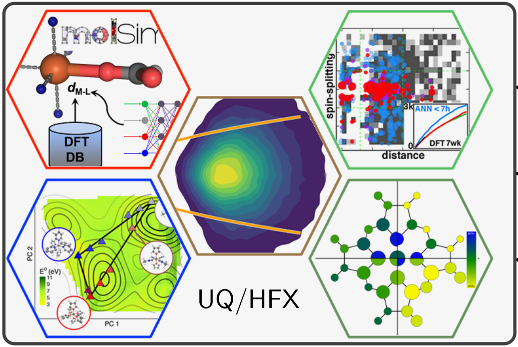
More than just ML

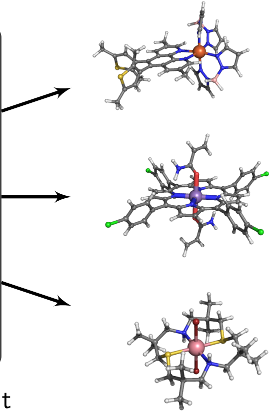
github.com/hjkgrp

mAD/molSimplify

design algorithms

leads





RAC features

chemical insight

Janet, J.P. *et al.*, *Inorg. Chem.*, 58(16):10592–10606, 2019.

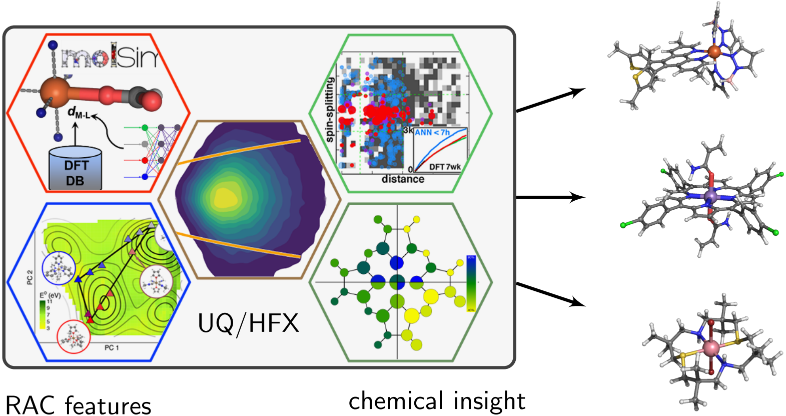
More than just ML

github.com/hjkgrp

mAD/molSimplify

design algorithms

leads



RAC features

chemical insight

An end-to-end approach for rational TM complex design

Case study: redox couples

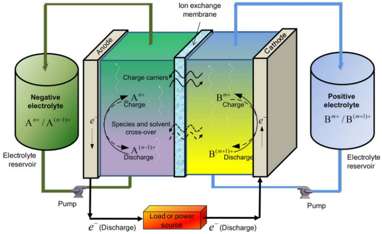
Consider a multiobjective optimization problem:

Redox flow batteries (RFBs)
are a promising option for
scalable energy storage:

Case study: redox couples

Consider a multiobjective optimization problem:

Redox flow batteries (RFBs) are a promising option for scalable energy storage:

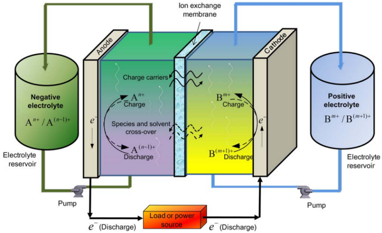


Perry, M.L. and Adam, Z., *J. Electrochem. Soc.*, 163(1):A5064–A5067, 2018.

Case study: redox couples

Consider a multiobjective optimization problem:

Redox flow batteries (RFBs) are a promising option for scalable energy storage:



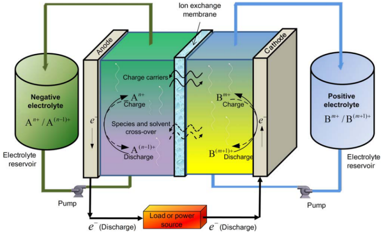
Perry, M.L. and Adam, Z., *J. Electrochem. Soc.*, 163(1):A5064–A5067, 2018.

Transition metal complexes make attractive redox couples for RFBs

Case study: redox couples

Consider a multiobjective optimization problem:

Redox flow batteries (RFBs) are a promising option for scalable energy storage:



Perry, M.L. and Adam, Z., *J. Electrochem. Soc.*, 163(1):A5064–A5067, 2018.

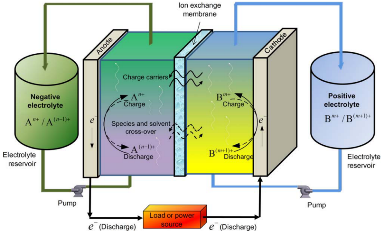
Transition metal complexes make attractive redox couples for RFBs

- good ion stability (compared to organics)

Case study: redox couples

Consider a multiobjective optimization problem:

Redox flow batteries (RFBs) are a promising option for scalable energy storage:



Perry, M.L. and Adam, Z., *J. Electrochem. Soc.*, 163(1):A5064–A5067, 2018.

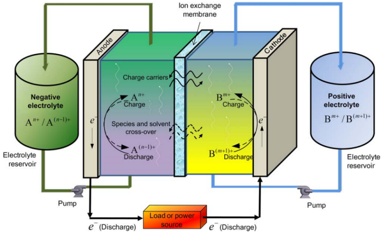
Transition metal complexes make attractive redox couples for RFBs

- good ion stability (compared to organics)
- good range of redox potentials available

Case study: redox couples

Consider a multiobjective optimization problem:

Redox flow batteries (RFBs) are a promising option for scalable energy storage:



Perry, M.L. and Adam, Z., *J. Electrochem. Soc.*, 163(1):A5064–A5067, 2018.

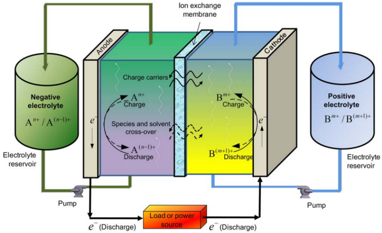
Transition metal complexes make attractive redox couples for RFBs

- good ion stability (compared to organics)
- good range of redox potentials available
- solubility is an issue!

Case study: redox couples

Consider a multiobjective optimization problem:

Redox flow batteries (RFBs) are a promising option for scalable energy storage:



Perry, M.L. and Adam, Z., *J. Electrochem. Soc.*, 163(1):A5064–A5067, 2018.

Transition metal complexes make attractive redox couples for RFBs

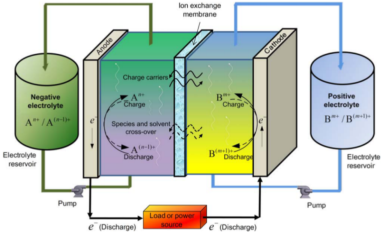
- good ion stability (compared to organics)
- good range of redox potentials available
- solubility is an issue!

$$E_{\text{cell}} = 0.5 \times V_{\text{cell}} \times C \times n \times F$$

Case study: redox couples

Consider a multiobjective optimization problem:

Redox flow batteries (RFBs) are a promising option for scalable energy storage:



Perry, M.L. and Adam, Z., *J. Electrochem. Soc.*, 163(1):A5064–A5067, 2018.

Transition metal complexes make attractive redox couples for RFBs

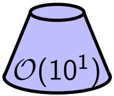
- good ion stability (compared to organics)
- good range of redox potentials available
- solubility is an issue!

$$E_{\text{cell}} = 0.5 \times V_{\text{cell}} \times C \times n \times F$$

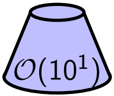
We need complexes that have high redox potential **and** good solubility

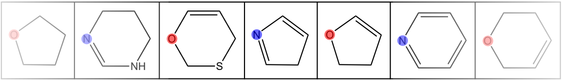
Design space construction

Design space construction



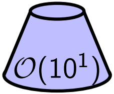
Design space construction

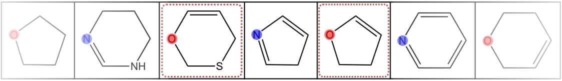




40 heterocycles

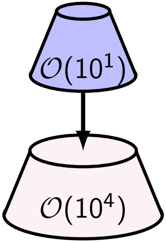
Design space construction

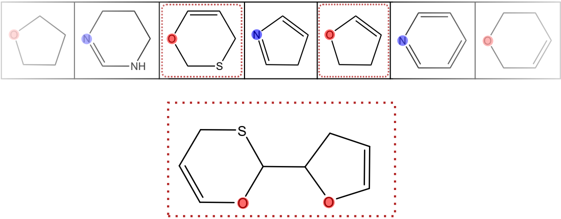




40 heterocycles

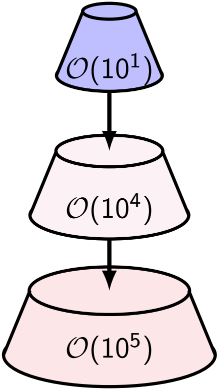
Design space construction

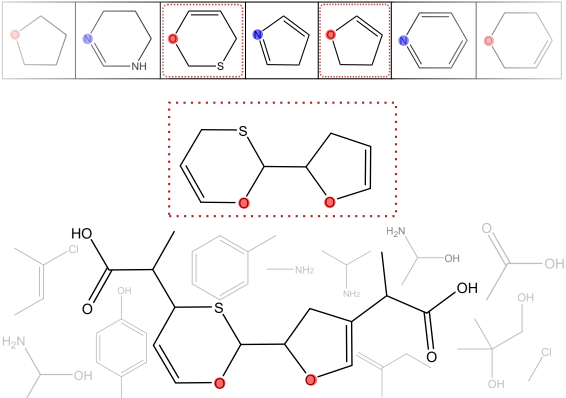




800 base ligands

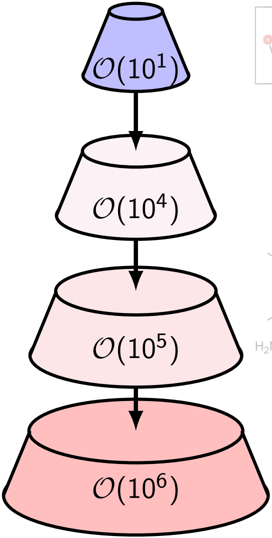
Design space construction

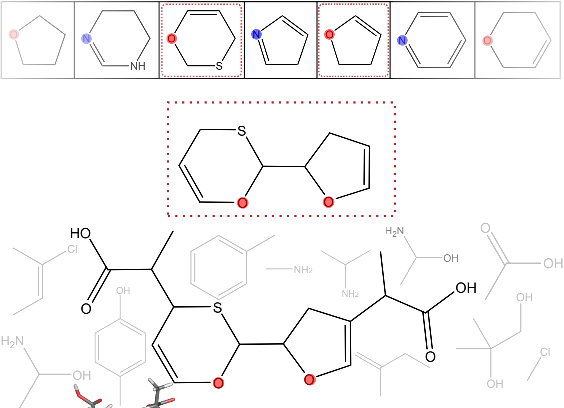


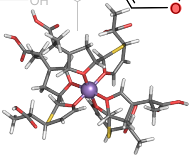


700k unique ligands

Design space construction







$\sim$ 3M complexes
with 4 metals

Multiobjective framework

We can predict quantites of interest for our RFBs:

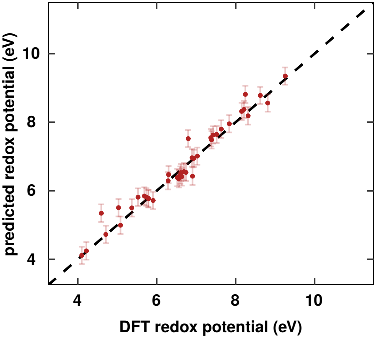
Janet, J.P., *et al.*, *in preparation*.

3×100 tanh nodes, multitask, fully connected

Multiobjective framework

We can predict quantites of interest for our RFBs:

$$\Delta G_{\text{solv}} \approx \Delta E_{\text{III}-\text{II}} + \Delta\Delta G_s$$



Janet, J.P., *et al.*, *in preparation*.

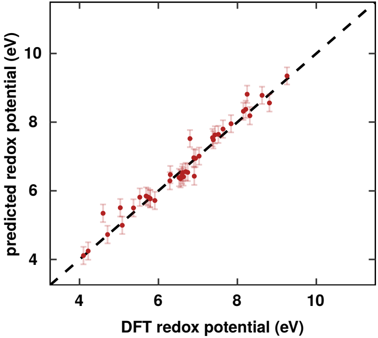
3 × 100 tanh nodes, multitask, fully connected

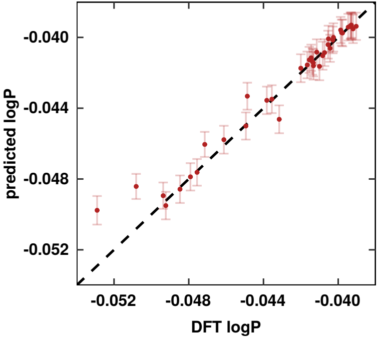
Multiobjective framework

We can predict quantites of interest for our RFBs:

$\Delta G_{\mathrm{solv}} \approx \Delta E_{\mathrm{III-II}} + \Delta\Delta G_s$

$\log P \approx \log \frac{\Delta G_{\mathrm{s,II,octanol}}}{\Delta G_{\mathrm{s,II,water}}}$





Janet, J.P., *et al.*, *in preparation*.

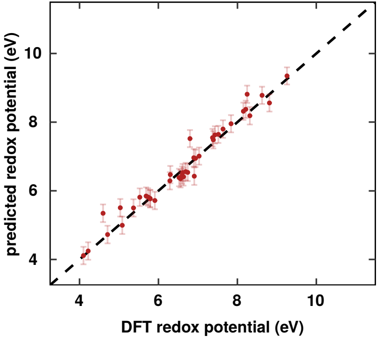
3 × 100 tanh nodes, multitask, fully connected

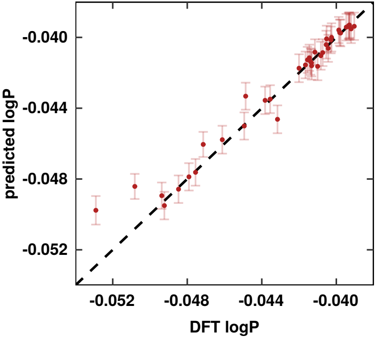
Multiobjective framework

We can predict quantites of interest for our RFBs:

$\Delta G_{\text{solv}} \approx \Delta E_{\text{III}-\text{II}} + \Delta\Delta G_s$

$\log \text{P} \approx \log \frac{\Delta G_{\text{s,II,octanol}}}{\Delta G_{\text{s,II,water}}}$





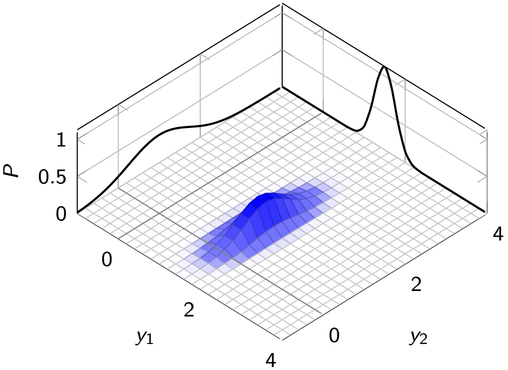
$$\begin{matrix}\Delta G_{\text{solv}} \\ \log \text{P}\end{matrix} = \begin{bmatrix}\hat{y}_1 \\ \hat{y}_2\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\hat{\mu}_1 \\ \hat{\mu}_2\end{bmatrix}, \begin{bmatrix}\hat{\sigma}_1^2 & 0 \\ 0 & \hat{\sigma}_2^2\end{bmatrix}\right)$$

Janet, J.P., *et al.*, *in preparation*.

3 × 100 tanh nodes, multitask, fully connected

Multiobjective framework

We can predict quantites of interest for our RFBs:

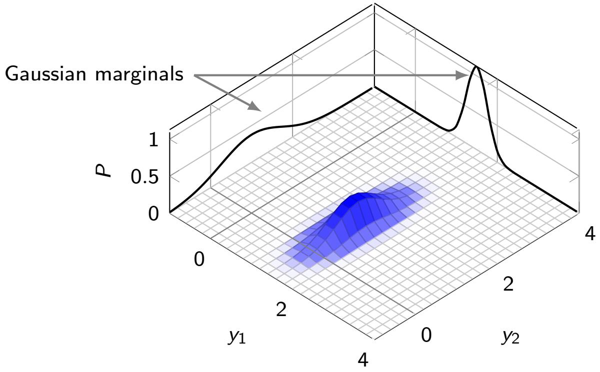


$$\begin{matrix}\Delta G_{\mathrm{solv}} \\ \log \mathrm{P}\end{matrix} = \begin{bmatrix}\hat{y}_1 \\ \hat{y}_2\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\hat{\mu}_1 \\ \hat{\mu}_2\end{bmatrix}, \begin{bmatrix}\hat{\sigma}_1^2 & 0 \\ 0 & \hat{\sigma}_2^2\end{bmatrix}\right)$$

Janet, J.P., *et al.*, *in preparation*.

Multiobjective framework

We can predict quantites of interest for our RFBs:

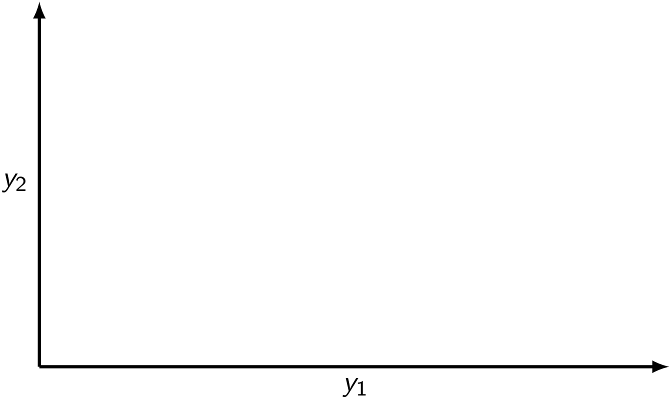


$$\begin{matrix}\Delta G_{\text{solv}} \\ \log \text{P}\end{matrix} = \begin{bmatrix}\hat{y}_1 \\ \hat{y}_2\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\hat{\mu}_1 \\ \hat{\mu}_2\end{bmatrix}, \begin{bmatrix}\hat{\sigma}_1^2 & 0 \\ 0 & \hat{\sigma}_2^2\end{bmatrix}\right)$$

Janet, J.P., *et al.*, *in preparation*.

2D EGO Illustration

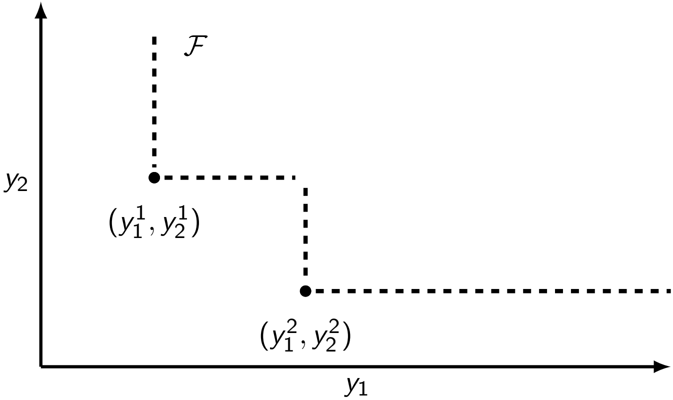
We will use a multiobjective expected improvement framework:



Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

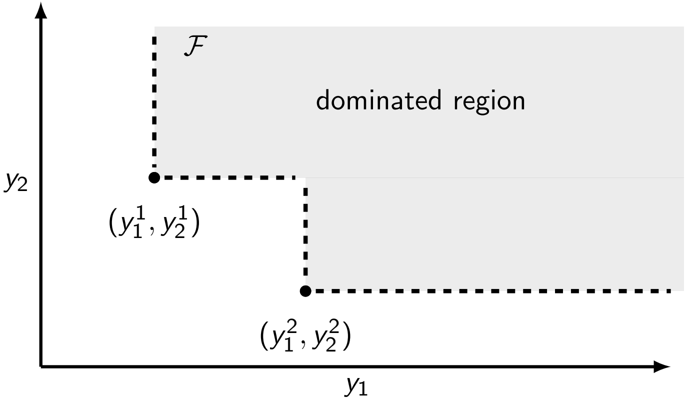
We will use a multiobjective expected improvement framework:



Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

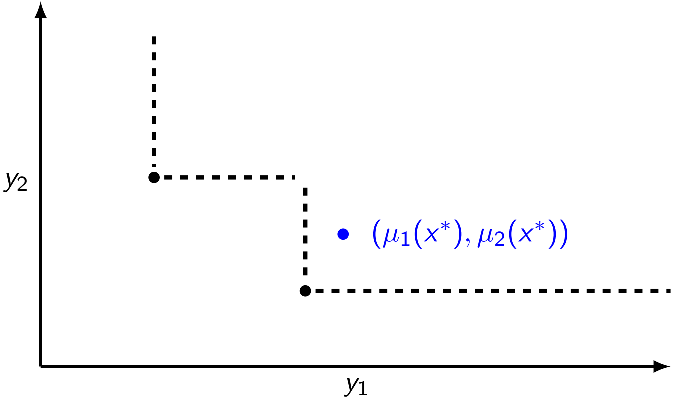
We will use a multiobjective expected improvement framework:



Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

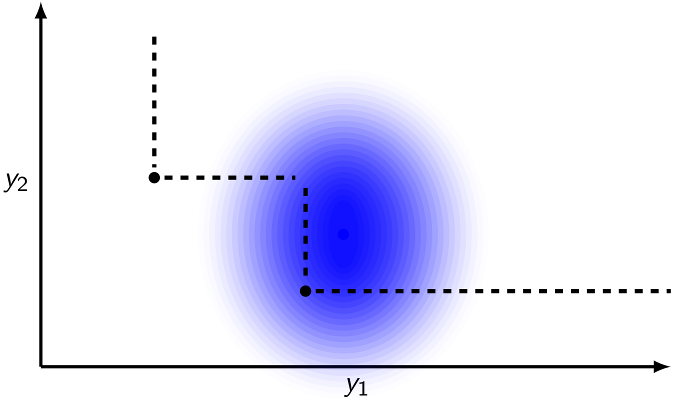
We will use a multiobjective expected improvement framework:



Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

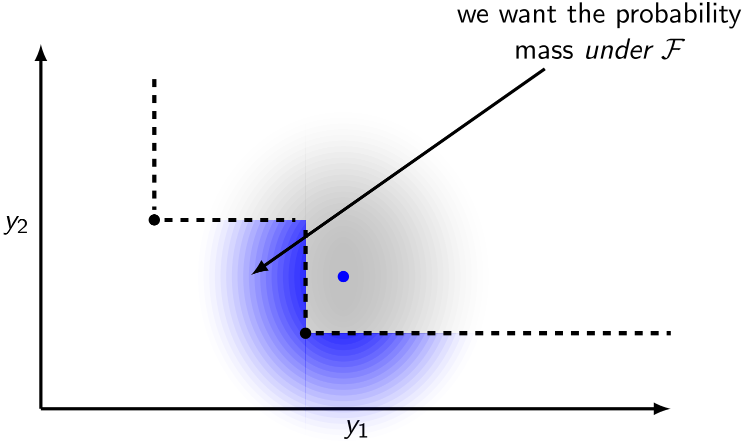
We will use a multiobjective expected improvement framework:



Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

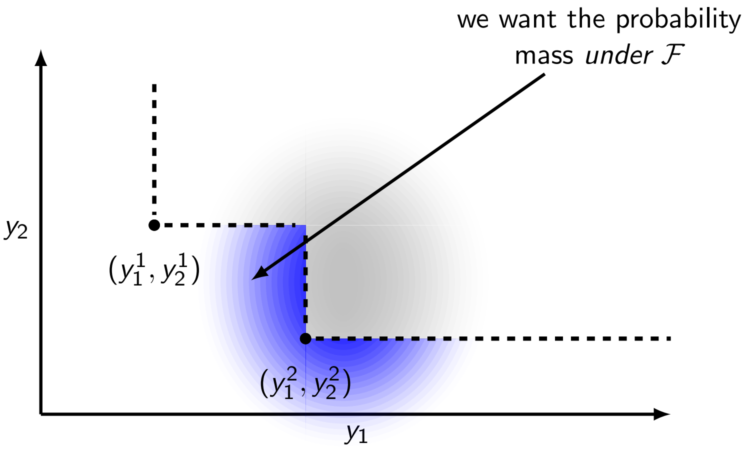
We will use a multiobjective expected improvement framework:



Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

We will use a multiobjective expected improvement framework:

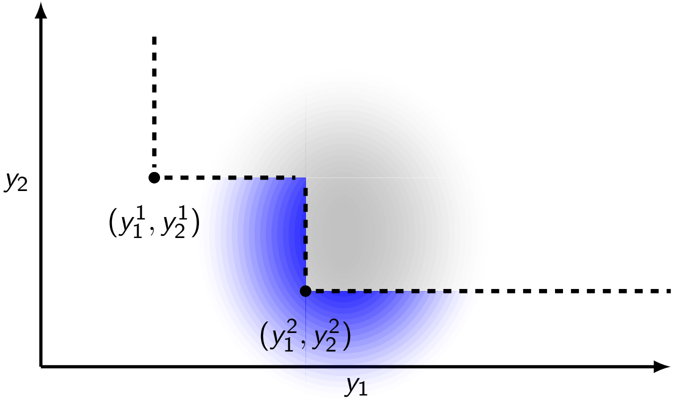


Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

We will use a multiobjective expected improvement framework:

$$P(I) =$$

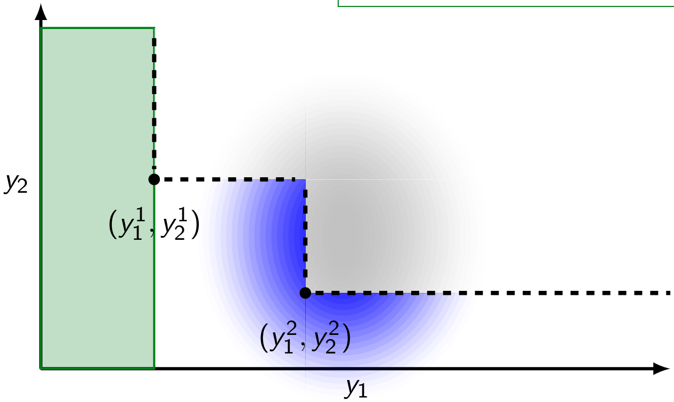


Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

We will use a multiobjective expected improvement framework:

$$P(I) = \boxed{\int_{-\infty}^{y_1^1} \int_{-\infty}^{\infty} p(\hat{y}_1(x^*), \hat{y}_2(x^*)) dy_1 dy_2}$$



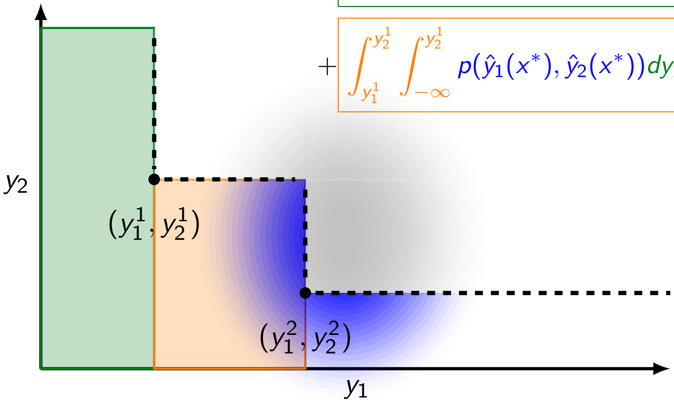
Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

We will use a multiobjective expected improvement framework:

$$P(I) = \boxed{\int_{-\infty}^{y_1^1}\int_{-\infty}^{\infty} p(\hat{y}_1(x^*), \hat{y}_2(x^*))dy_1 dy_2}$$

$$+ \boxed{\int_{y_1^1}^{y_2^1}\int_{-\infty}^{y_2^1} p(\hat{y}_1(x^*), \hat{y}_2(x^*))dy_1 dy_2}$$



Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

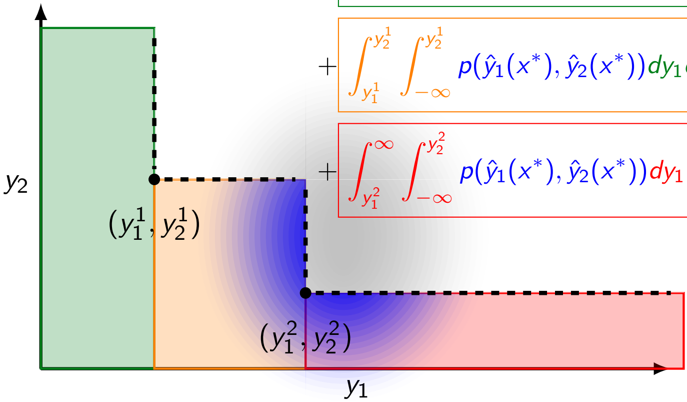
2D EGO Illustration

We will use a multiobjective expected improvement framework:

$$P(I) = \boxed{\int_{-\infty}^{y_1^1}\int_{-\infty}^{\infty} p(\hat{y}_1(x^*), \hat{y}_2(x^*))\,dy_1\,dy_2}$$

$$+ \boxed{\int_{y_1^1}^{y_2^1}\int_{-\infty}^{y_2^1} p(\hat{y}_1(x^*), \hat{y}_2(x^*))\,dy_1\,dy_2}$$

$$+ \boxed{\int_{y_1^2}^{\infty}\int_{-\infty}^{y_2^2} p(\hat{y}_1(x^*), \hat{y}_2(x^*))\,dy_1\,dy_2}$$



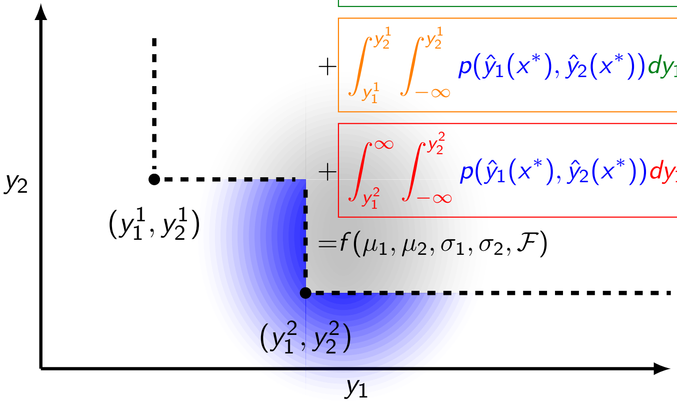
Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

We will use a multiobjective expected improvement framework:

$$P(I) = \boxed{\int_{-\infty}^{y_1^1}\int_{-\infty}^{\infty} p(\hat{y}_1(x^*), \hat{y}_2(x^*))\,dy_1\,dy_2} + \boxed{\int_{y_1^1}^{y_2^1}\int_{-\infty}^{y_2^1} p(\hat{y}_1(x^*), \hat{y}_2(x^*))\,dy_1\,dy_2} + \boxed{\int_{y_1^2}^{\infty}\int_{-\infty}^{y_2^2} p(\hat{y}_1(x^*), \hat{y}_2(x^*))\,dy_1\,dy_2}$$

$$= f(\mu_1, \mu_2, \sigma_1, \sigma_2, \mathcal{F})$$

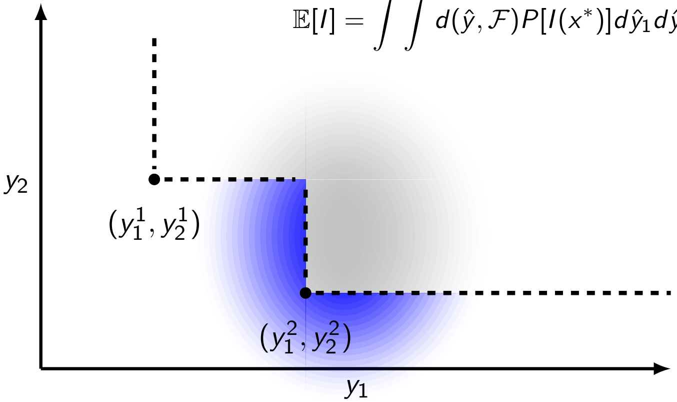


Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

We will use a multiobjective expected improvement framework:

$$\mathbb{E}[I] = \int\int d(\hat{y}, \mathcal{F}) P[I(x^*)] d\hat{y}_1 d\hat{y}_2$$



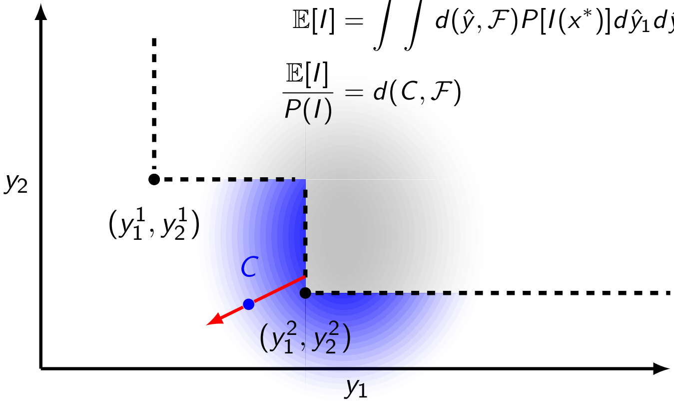
Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

We will use a multiobjective expected improvement framework:

$$\mathbb{E}[I] = \int\int d(\hat{y}, \mathcal{F}) P[I(x^*)] d\hat{y}_1 d\hat{y}_2$$

$$\frac{\mathbb{E}[I]}{P(I)} = d(C, \mathcal{F})$$



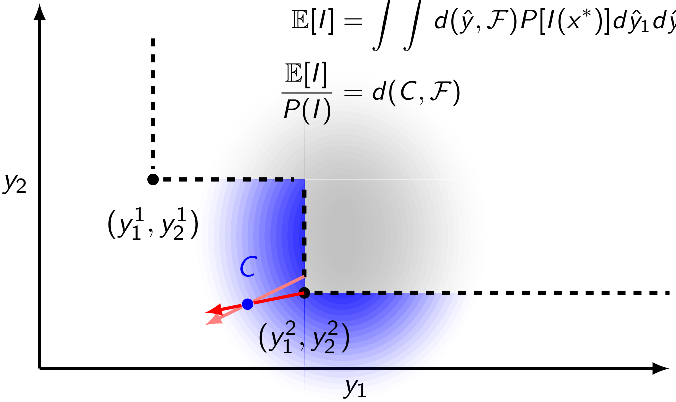
Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

2D EGO Illustration

We will use a multiobjective expected improvement framework:

$$\mathbb{E}[I] = \int\int d(\hat{y}, \mathcal{F}) P[I(x^*)] d\hat{y}_1 d\hat{y}_2$$

$$\frac{\mathbb{E}[I]}{P(I)} = d(C, \mathcal{F})$$



Keane, A. J., AIAA Journal, 44(4):879–891, 2006.

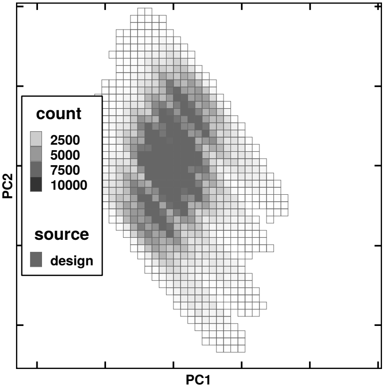
Design space and existing data

We observe poor overlap between existing data and design space:

Janet, J.P., *et al.*, *in preparation*.

Design space and existing data

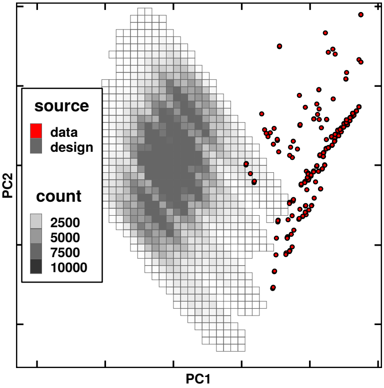
We observe poor overlap between existing data and design space:



Janet, J.P., *et al.*, *in preparation*.

Design space and existing data

We observe poor overlap between existing data and design space:

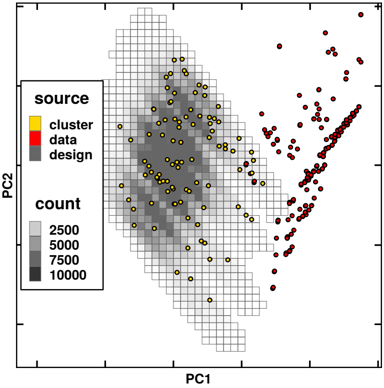


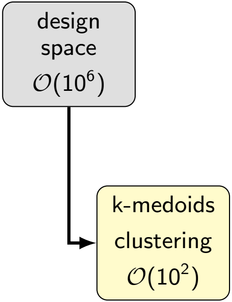
design space $\mathcal{O}(10^6)$

Janet, J.P., *et al.*, *in preparation*.

Design space and existing data

We observe poor overlap between existing data and design space:

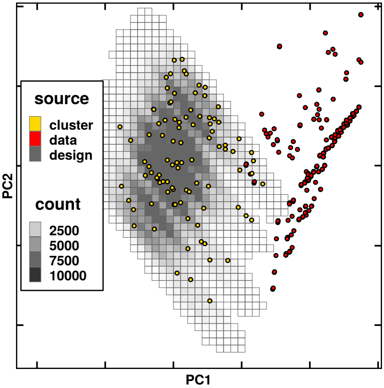


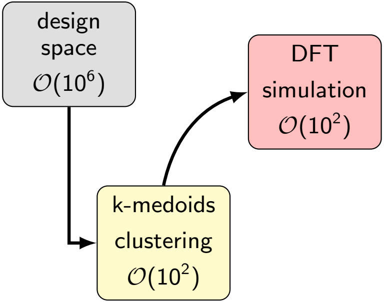


Janet, J.P., *et al.*, *in preparation*.

Design space and existing data

We observe poor overlap between existing data and design space:

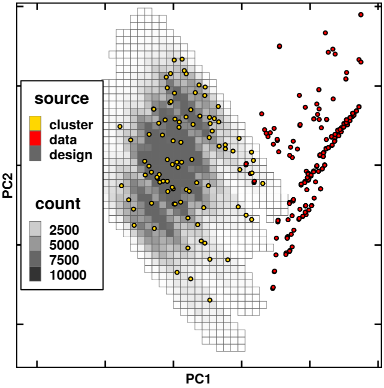


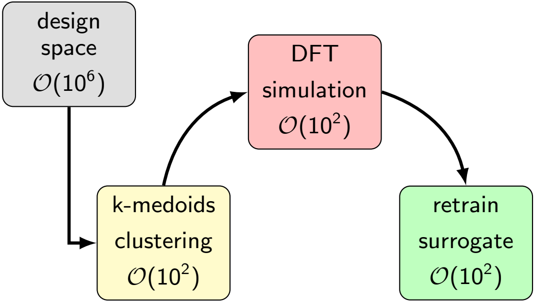


Janet, J.P., *et al.*, *in preparation*.

Design space and existing data

We observe poor overlap between existing data and design space:

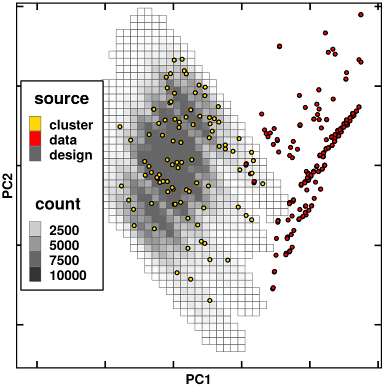


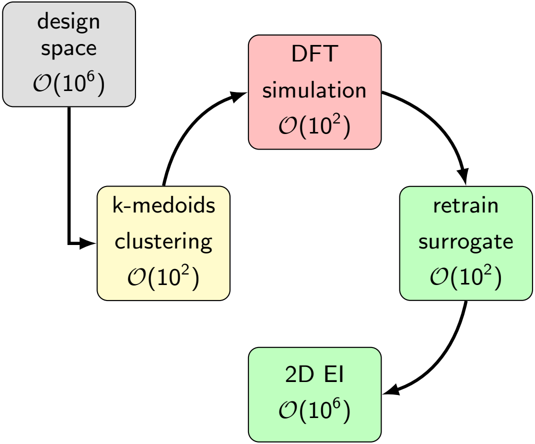


Janet, J.P., *et al.*, *in preparation*.

Design space and existing data

We observe poor overlap between existing data and design space:

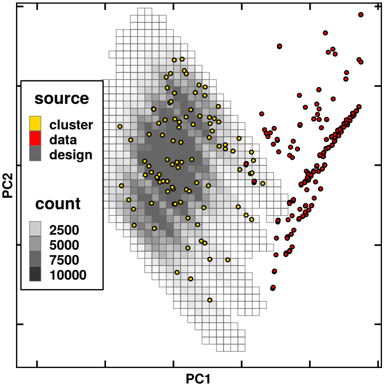


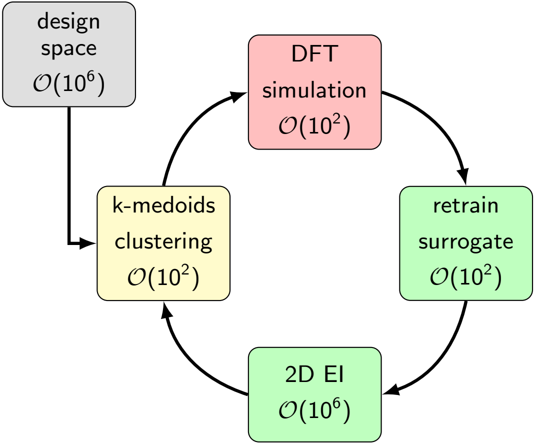


Janet, J.P., *et al.*, *in preparation*.

Design space and existing data

We observe poor overlap between existing data and design space:

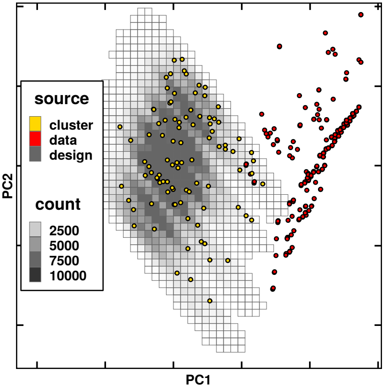


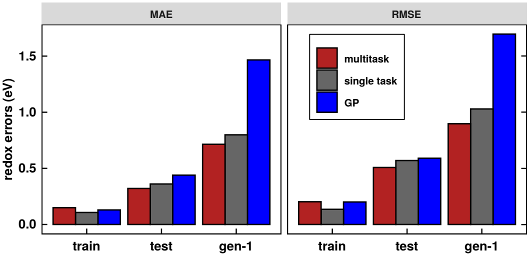


Janet, J.P., *et al.*, *in preparation*.

Design space and existing data

We observe poor overlap between existing data and design space:



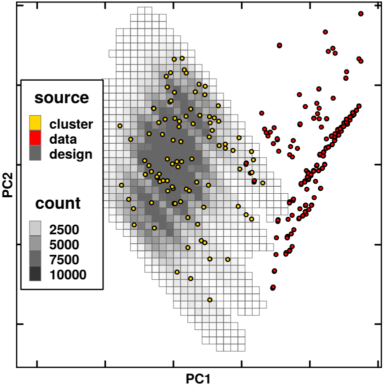


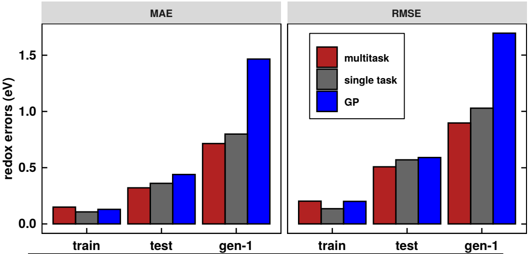
single task redox: 3 × 100 tanh nodes, fully connected
single task log P: 2 × 50 ReLU nodes, skip + residual connections
multitask: 2 × 100 tanh nodes, fully connected

Janet, J.P., *et al.*, *in preparation*.

Design space and existing data

We observe poor overlap between existing data and design space:





Therefore, we will use multitask ANN to drive the design process

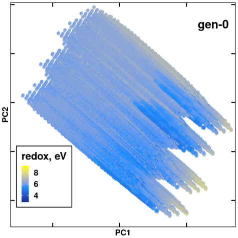
single task redox: 3 × 100 tanh nodes, fully connected
single task log P: 2 × 50 ReLU nodes, skip + residual connections
multitask: 2 × 100 tanh nodes, fully connected

Janet, J.P., *et al.*, *in preparation*.

Active learning

We can monitor the evolution of the model as data is acquired

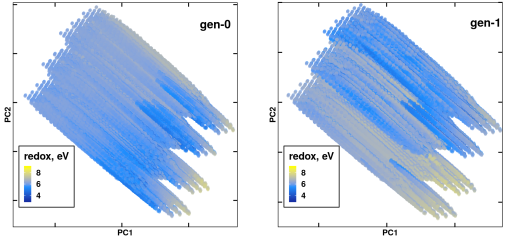
feature space



Active learning

We can monitor the evolution of the model as data is acquired

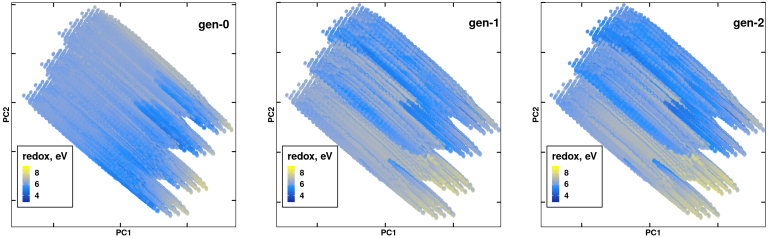
feature space



Active learning

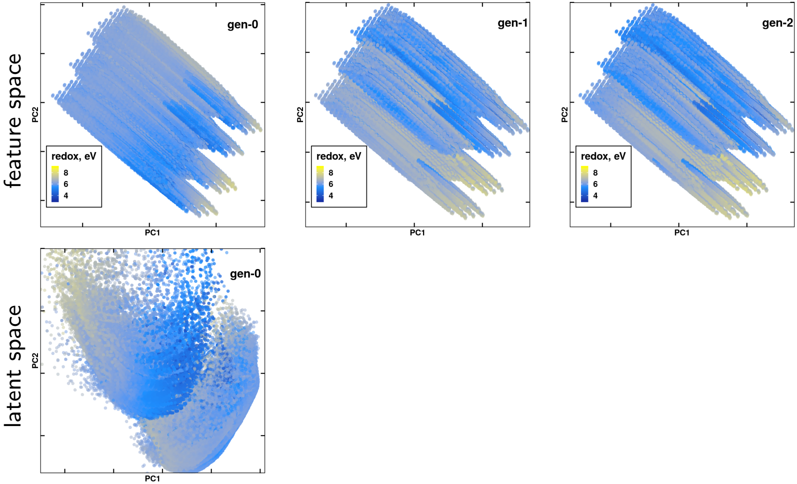
We can monitor the evolution of the model as data is acquired

feature space



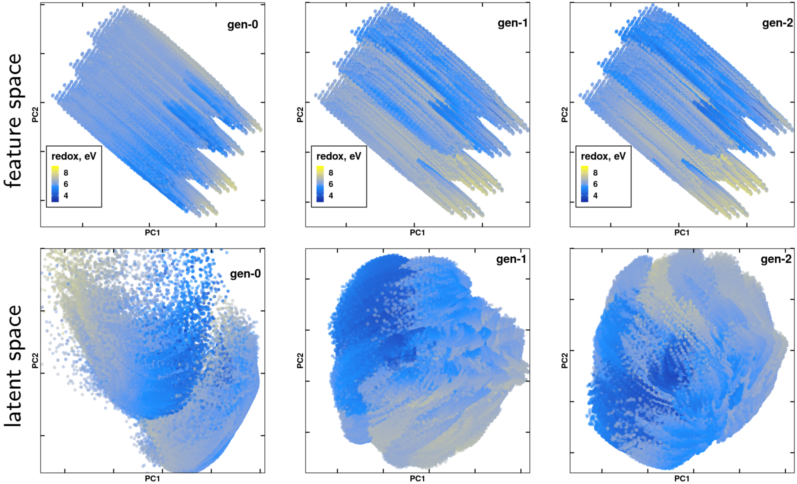
Active learning

We can monitor the evolution of the model as data is acquired



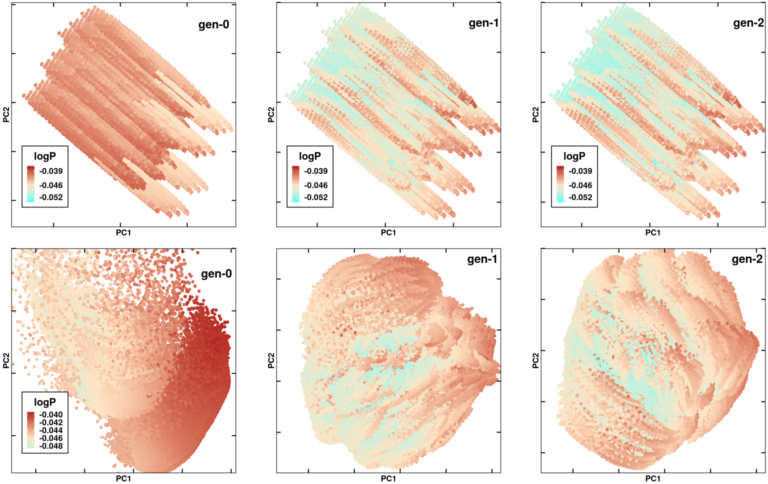
Active learning

We can monitor the evolution of the model as data is acquired



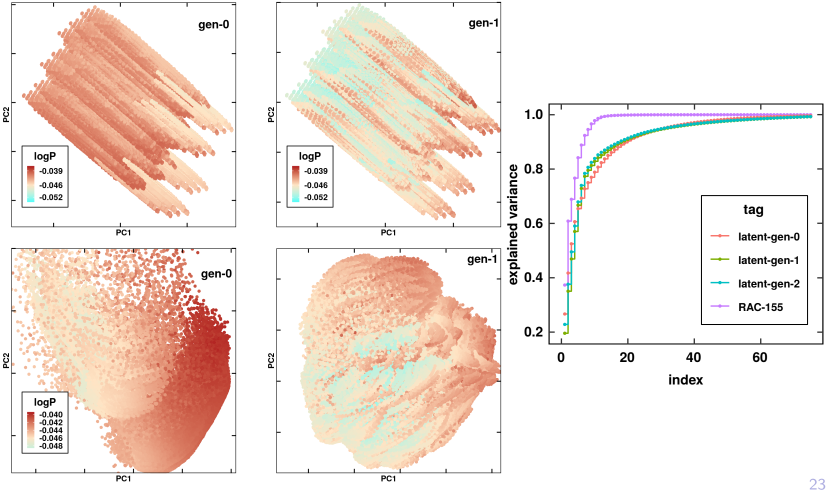
Active learning

We can monitor the evolution of the model as data is acquired

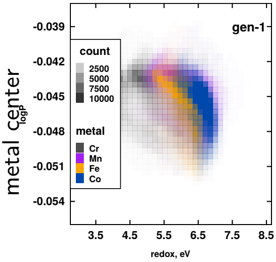


Active learning

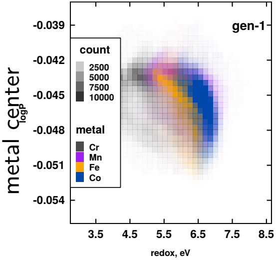
We can monitor the evolution of the model as data is acquired

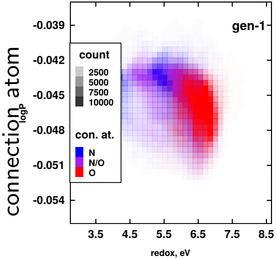


Mapping output space

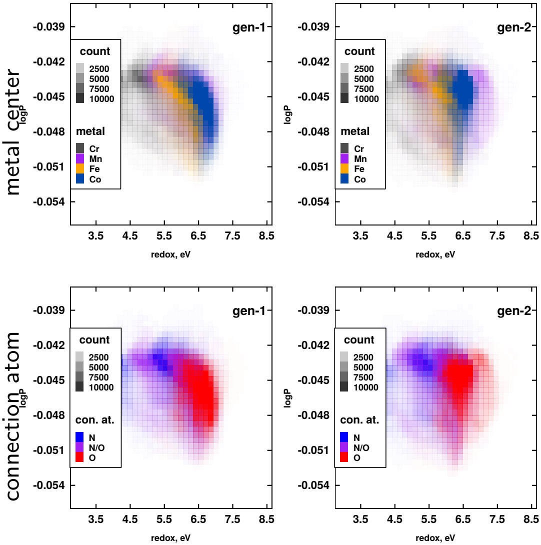


Mapping output space





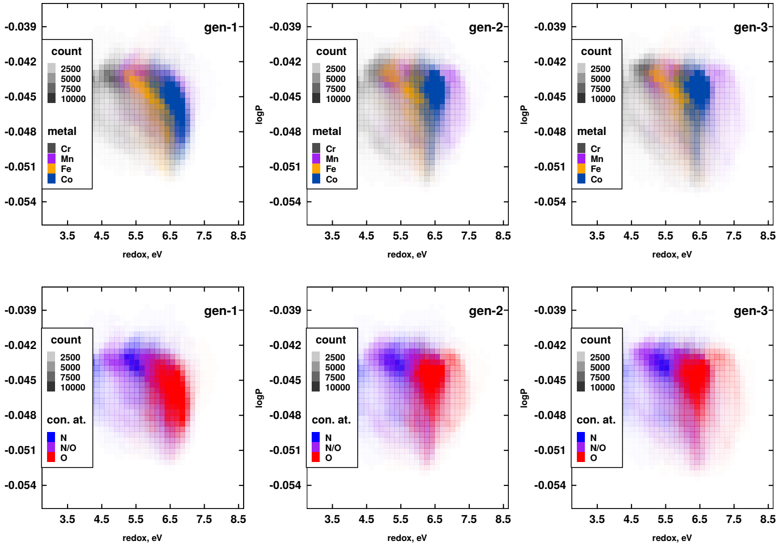
Mapping output space



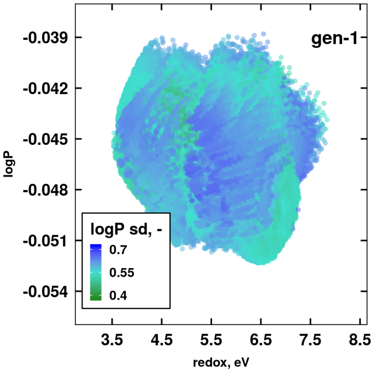
Mapping output space

metal center

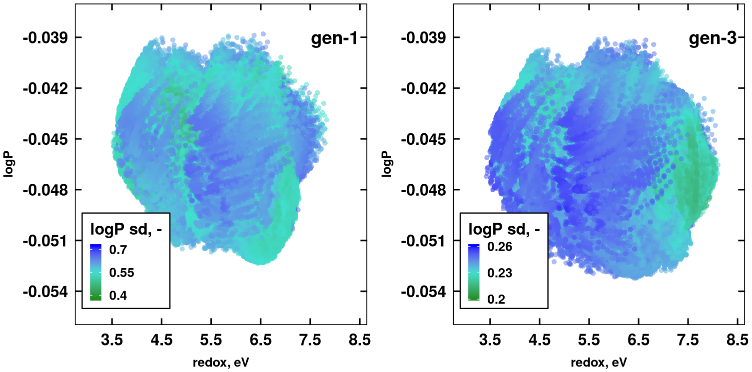
connection atom



Mapping output space



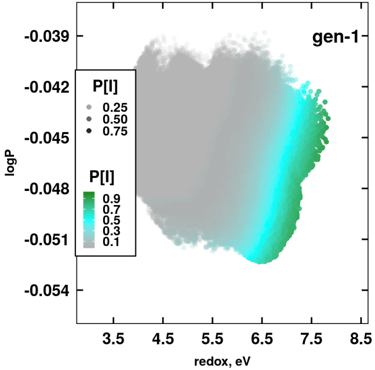
Mapping output space



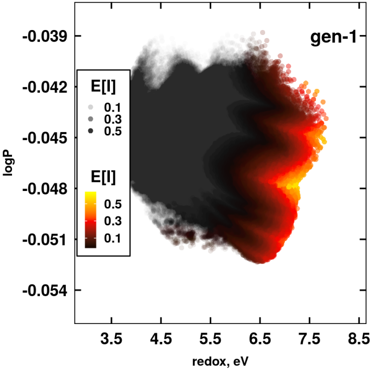
model confidence is localized in target region!

EGO results

probability of improvement



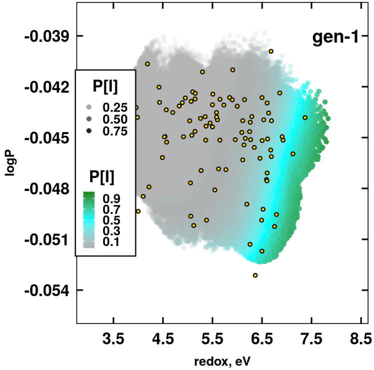
expected improvement



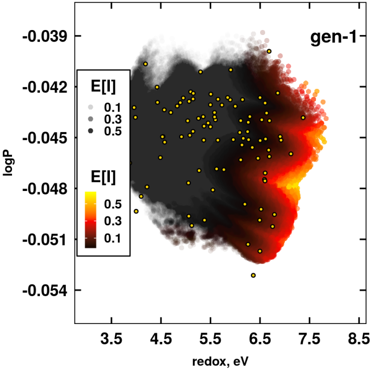
Janet, J.P., *et al.*, *in preparation*.

EGO results

probability of improvement

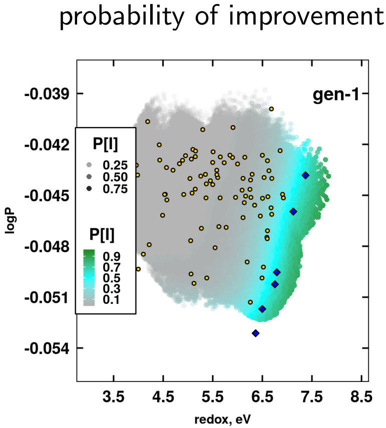


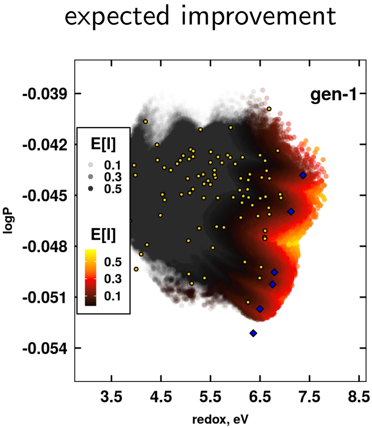
expected improvement



Janet, J.P., *et al.*, *in preparation*.

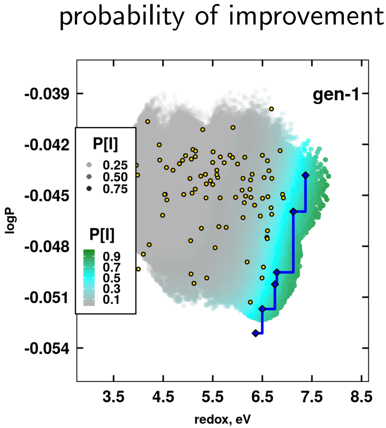
EGO results

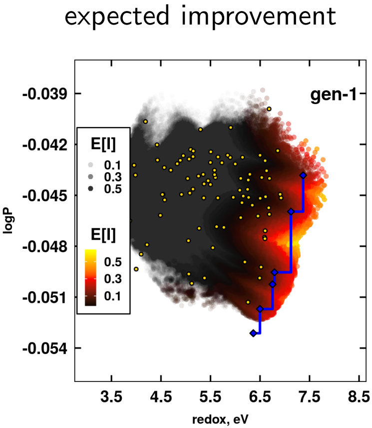




Janet, J.P., *et al.*, *in preparation*.

EGO results

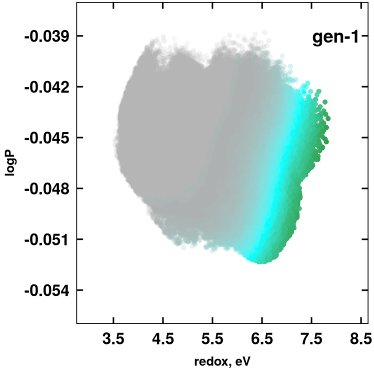




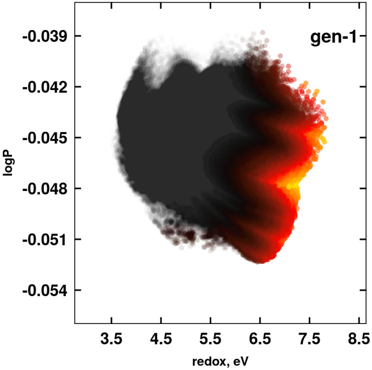
Janet, J.P., *et al.*, *in preparation*.

EGO results

probability of improvement

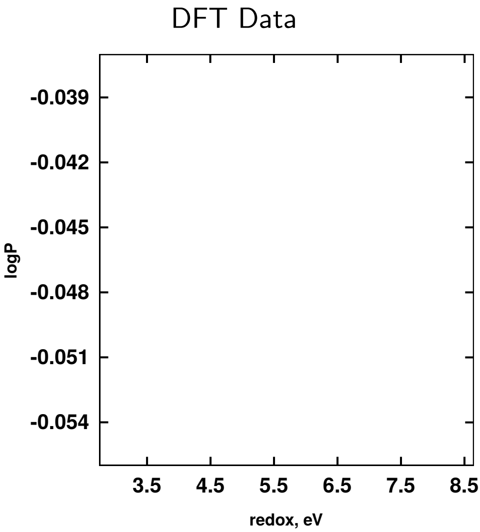


expected improvement



Janet, J.P., *et al.*, *in preparation*.

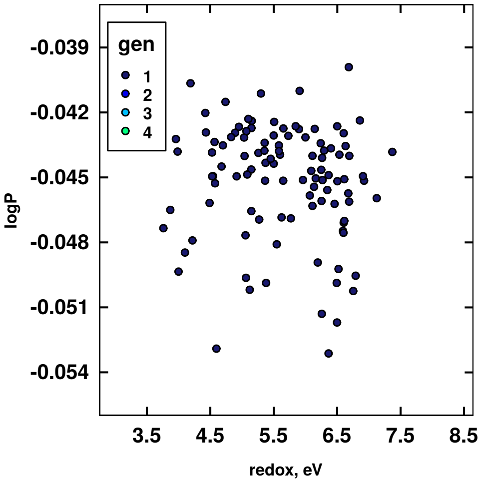
DFT results



Janet, J.P., *et al.*, *in preparation*.

DFT results

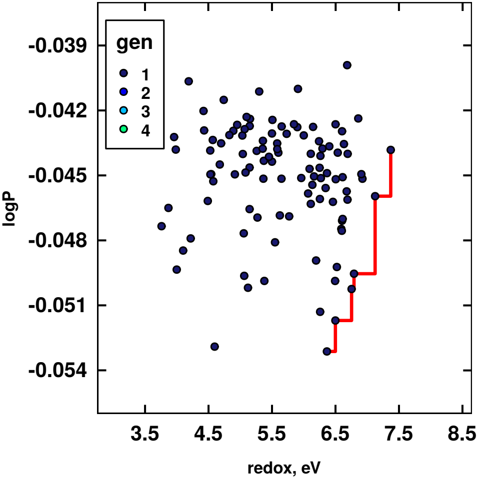
k-medoids points (gen 1: 107 complexes)



Janet, J.P., *et al.*, *in preparation*.

DFT results

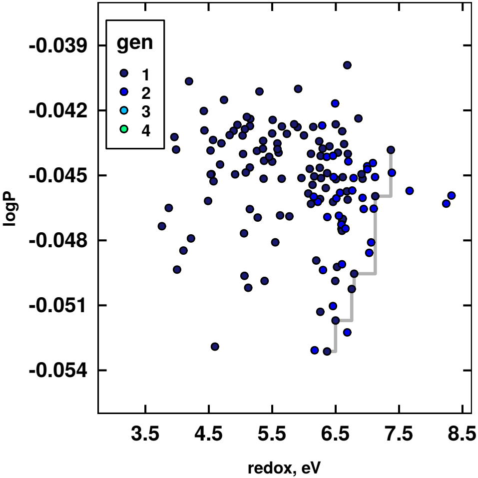
pareto front (gen 1)



Janet, J.P., *et al.*, *in preparation*.

DFT results

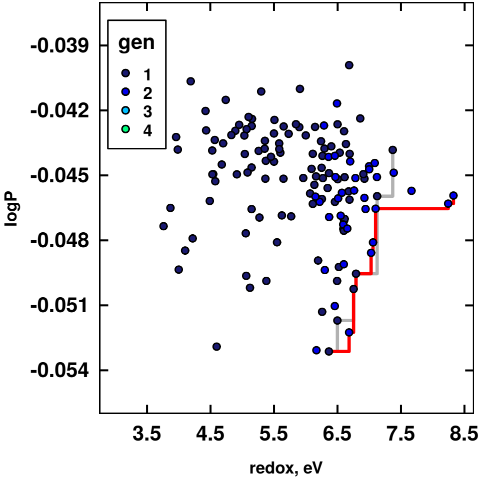
EI points (gen 2: 34 complexes)



Janet, J.P., *et al.*, *in preparation*.

DFT results

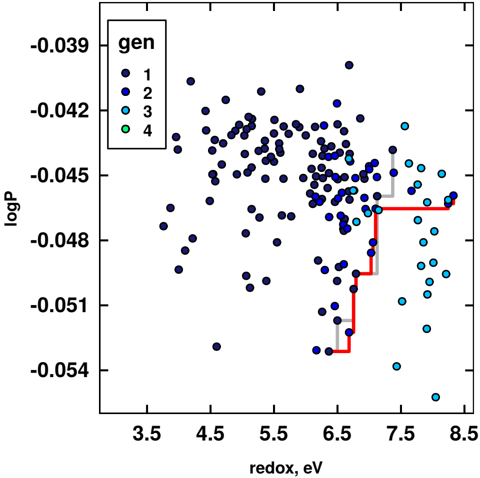
pareto front (gen 2)



Janet, J.P., *et al.*, *in preparation*.

DFT results

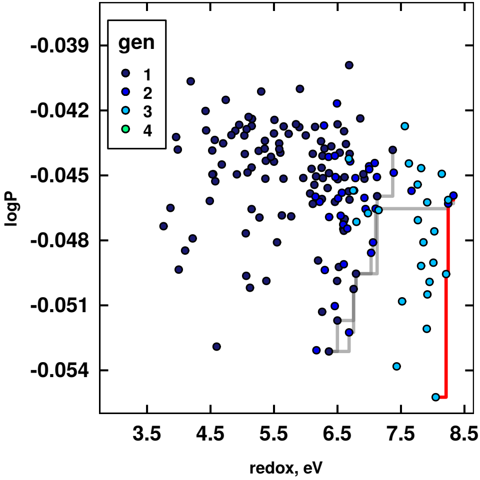
EI points (gen 3: 24 complexes)



Janet, J.P., *et al.*, *in preparation*.

DFT results

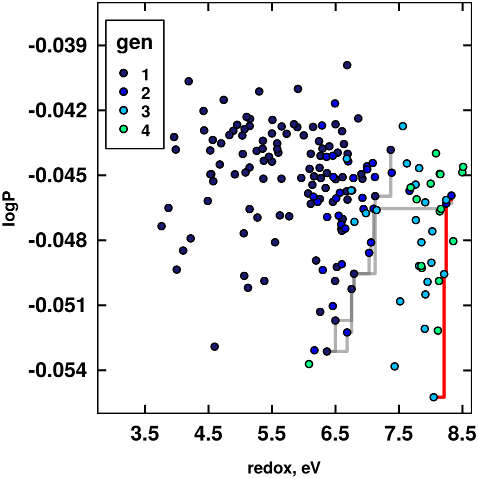
pareto front (gen 3)



Janet, J.P., *et al.*, *in preparation*.

DFT results

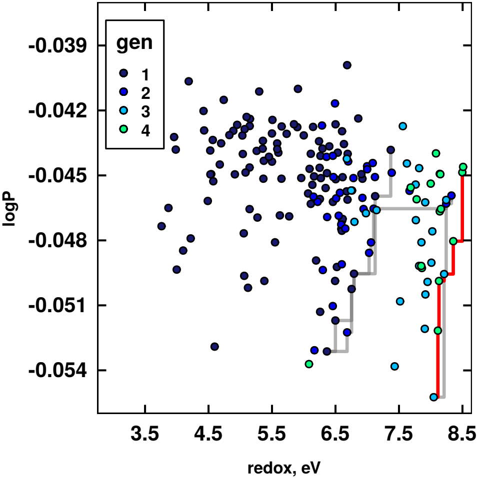
EI points (gen 4: 15 complexes)



Janet, J.P., *et al.*, *in preparation*.

DFT results

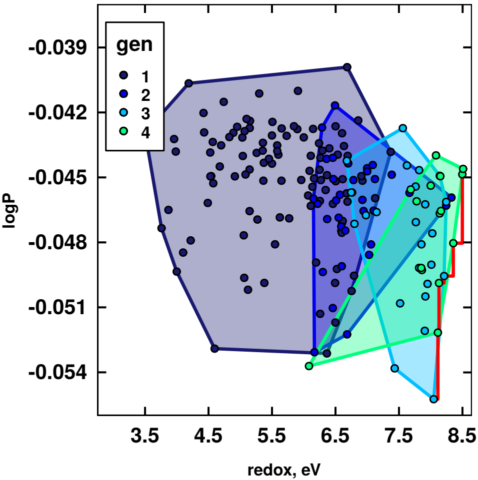
pareto front (gen 4)



Janet, J.P., *et al.*, *in preparation*.

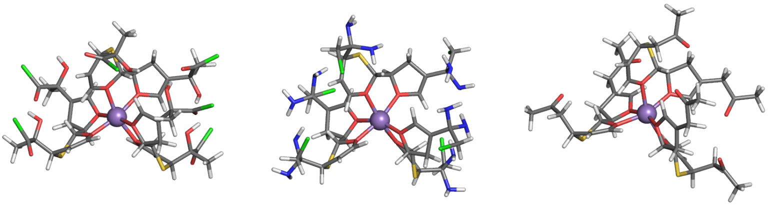
DFT results

convex hulls



Janet, J.P., *et al.*, *in preparation*.

DFT results



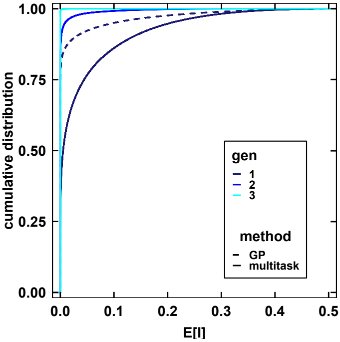
Janet, J.P., *et al.*, *in preparation*.

Case study conclusions

Janet, J.P., *et al.*, *in preparation*.

Case study conclusions

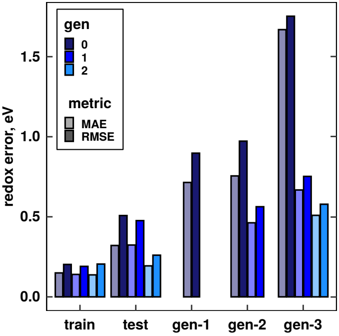
- EI framework provides high resolution in the region of interest (c.f. maximum uncertainty), converges quickly



Janet, J.P., *et al.*, *in preparation*.

Case study conclusions

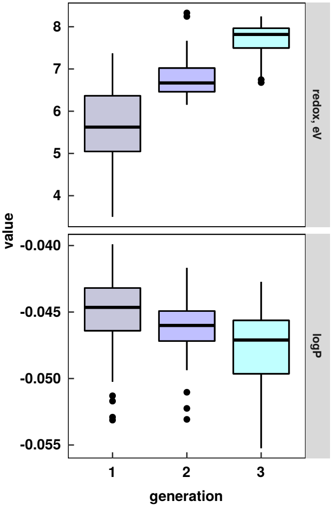
- EI framework provides high resolution in the region of interest (c.f. maximum uncertainty), converges quickly
- We are able to identify fruitful regions from large chemical spaces based on few DFT evaluations



Janet, J.P., *et al.*, *in preparation*.

Case study conclusions

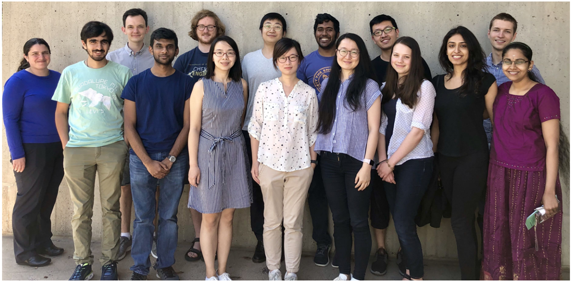
- EI framework provides high resolution in the region of interest (c.f. maximum uncertainty), converges quickly
- We are able to identify fruitful regions from large chemical spaces based on few DFT evaluations
- Multiobjective DFT optimization guided by data-driven method efficiency generates lead complexes



Janet, J.P., *et al.*, *in preparation*.

Acknowledgments

Thanks to the Kulik group and funding partners:



DARPA
MITei
DEPARTMENT OF ENERGY
UNITED STATES OF AMERICA
eni
Office of Naval Research
ONR
Science & Technology
NSF
BURROUGHS
WELLCOME
FUND

Thermodynamic cycle

$$\mathrm{M(III)}_s + e^- \xrightarrow[\Delta G_{\mathrm{solv}}]{} \mathrm{M(II)}_s$$

Thermodynamic cycle

$$M(III)_g + e^- \xrightarrow{\Delta G_{gas}} M(II)_g$$

$$M(III)_s + e^- \xrightarrow[\Delta G_{solv}]{} M(II)_s$$

Thermodynamic cycle

$$
\begin{tikzcd}
\mathrm{M(III)}_g + e^- \arrow[r, "\Delta \mathrm{G}_{\mathrm{gas}}"] \arrow[d, "\Delta \mathrm{G}_{\mathrm{s,III}}"'] & \mathrm{M(II)}_g \arrow[d, "\Delta \mathrm{G}_{\mathrm{s,II}}"] \\
\mathrm{M(III)}_s + e^- \arrow[r, "\Delta \mathrm{G}_{\mathrm{solv}}"'] & \mathrm{M(II)}_s
\end{tikzcd}
$$

Thermodynamic cycle

$$\begin{tikzcd}
M(III)_g + e^- \arrow[r, "\Delta G_{gas}"] \arrow[d, "\Delta G_{s,III}"'] & M(II)_g \arrow[d, "\Delta G_{s,II}"] \\
M(III)_s + e^- \arrow[r, "\Delta G_{solv}"'] & M(II)_s
\end{tikzcd}$$

$$\Delta G_{solv} = \Delta G_{gas} + \Delta\Delta G_s$$

Thermodynamic cycle

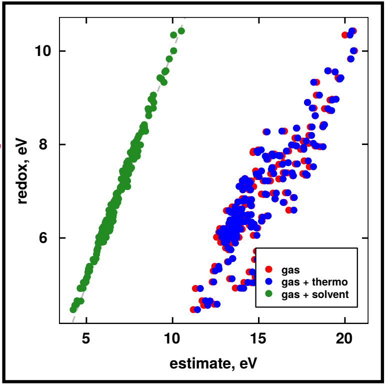
$$
\begin{tikzcd}
\mathrm{M(III)}_g + e^- \arrow[r, "\Delta \mathrm{G_{gas}}"] \arrow[d, "\Delta \mathrm{G_{s,III}}"'] & \mathrm{M(II)}_g \arrow[d, "\Delta \mathrm{G_{s,II}}"] \\
\mathrm{M(III)}_s + e^- \arrow[r, "\Delta \mathrm{G_{solv}}"'] & \mathrm{M(II)}_s
\end{tikzcd}
$$

$$
\begin{aligned}
\Delta \mathrm{G_{solv}} &= \Delta \mathrm{G_{gas}} + \Delta\Delta \mathrm{G}_s \\
&\approx \Delta \mathrm{E}_{\mathrm{gas},III-II} + \Delta\Delta \mathrm{G}_s
\end{aligned}
$$

Thermodynamic cycle

$$\begin{tikzcd}
M(III)_g + e^- \arrow[r, "\Delta G_{gas}"] \arrow[d, "\Delta G_{s,III}"'] & M(II)_g \arrow[d, "\Delta G_{s,"] \\
M(III)_s + e^- \arrow[r, "\Delta G_{solv}"'] & M(II)_s
\end{tikzcd}$$

$$\Delta G_{solv} = \Delta G_{gas} + \Delta\Delta G_s$$
$$\approx \Delta E_{gas,III-II} + \Delta\Delta G_s$$



Thermodynamic cycle

$$
\begin{tikzcd}
\mathrm{M(III)}_g + e^- \arrow[r, "\Delta \mathrm{G}_{\mathrm{gas}}"] \arrow[d, "\Delta \mathrm{G}_{\mathrm{s,III}}"'] & \mathrm{M(II)}_g \arrow[d, "\Delta \mathrm{G}_{\mathrm{s,II}}"] \\
\mathrm{M(III)}_s + e^- \arrow[r, "\Delta \mathrm{G}_{\mathrm{solv}}"'] & \mathrm{M(II)}_s
\end{tikzcd}
$$

$$
\begin{aligned}
\Delta \mathrm{G}_{\mathrm{solv}} &= \Delta \mathrm{G}_{\mathrm{gas}} + \Delta\Delta \mathrm{G}_s \\
&\approx \Delta \mathrm{E}_{\mathrm{gas},III-II} + \Delta\Delta \mathrm{G}_s \\
\log \mathrm{P} &\approx \log \frac{\Delta \mathrm{G}_{\mathrm{s,II,octanol}}}{\Delta \mathrm{G}_{\mathrm{s,II,water}}}
\end{aligned}
$$

- Spin state taken as the M(II) ground state
- Need 3 geometry optimizations at minimum