Machine Learning – now and in the future

Jon Paul Janet [1]

[1]Medicinal Chemistry, Early CVRM,
R&D BioPharmaceuticals, AstraZeneca, 431 83 Mölndal, Sweden

25.05.2021

Table of Contents

Rise of the (chemical) machines

Rise of the (chemical) machines

Something interesting happened at the **CASP 13** protein folding prediction competition in Mexico in December 2018...

Rise of the (chemical) machines

Something interesting happened at the **CASP 13** protein folding prediction competition in Mexico in December 2018...

A new entry, competing in their first CASP, dominated in the no-information category, **winning 25 out of 43 tests**.

Rise of the (chemical) machines

Something interesting happened at the **CASP 13** protein folding prediction competition in Mexico in December 2018...

A new entry, competing in their first CASP, dominated in the no-information category, **winning 25 out of 43 tests**.The next best team won 3 of the remaining tests.

Rise of the (chemical) machines

Something interesting happened at the **CASP 13** protein folding prediction competition in Mexico in December 2018...

A new entry, competing in their first CASP, dominated in the no-information category, **winning 25 out of 43 tests**.The next best team won 3 of the remaining tests.

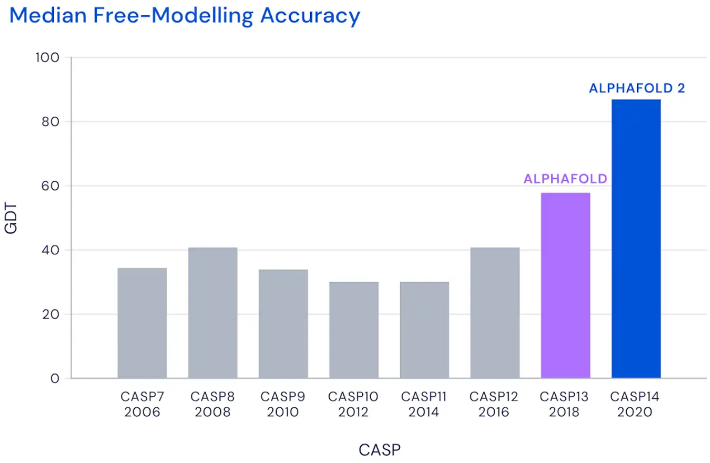
The same team ran away with the competition in **CASP 14** in 2020, leading CASP co-founder John Moult to conclude "In some sense the problem is solved"

Rise of the (chemical) machines

The team was Alphafold, by DeepMind.

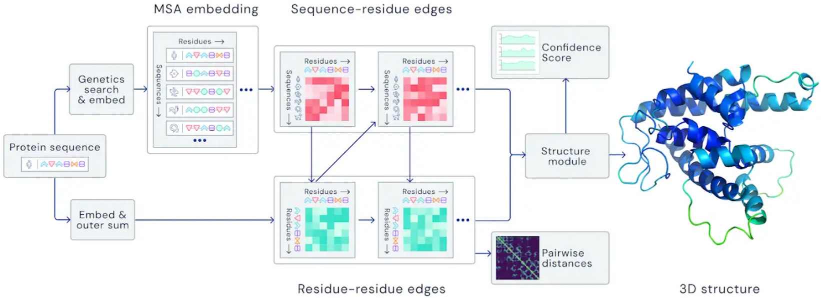
Rise of the (chemical) machines

The team was Alphafold, by DeepMind.



Rise of the (chemical) machines

The team was Alphafold, by DeepMind.



Senior, A.W., *et al.*, *Nature*, 577: 706–710, 2020.

Rise of the (chemical) machines

The team was Alphafold, by DeepMind.

"*It is not that machines are going to replace chemists. It's that the chemists who use machines will replace those that do not*"
-Derek Lowe, In the Pipeline

Rise of the (chemical) machines

The team was Alphafold, by DeepMind.

"*It is not that machines are going to replace chemists. It's that the chemists who use machines will replace those that do not*"
-Derek Lowe, In the Pipeline

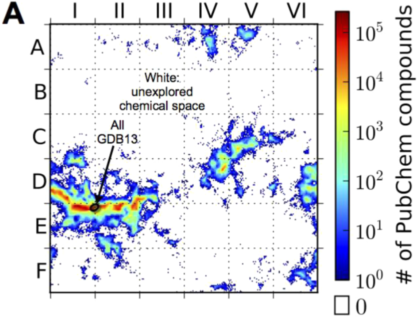
This is probably a bit strong, but all scientists generate data as a product. ML provides new, powerful ways to exploit this information.

Motivation: chemical discovery

Why is ML transforming chemisty?

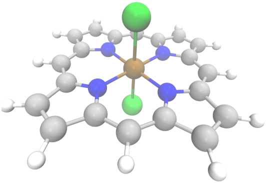
The space of possible chemistries is incredibly vast, with $\mathcal{O}(10^{60})$ small organic molecules.

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

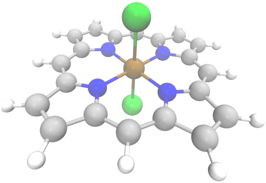


Virshup *et al.*, *J. Am. Chem. Soc.*, 135(19): 7296–7303, 2013.

Why ML in chemical sciences?

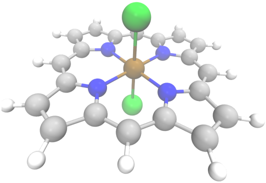


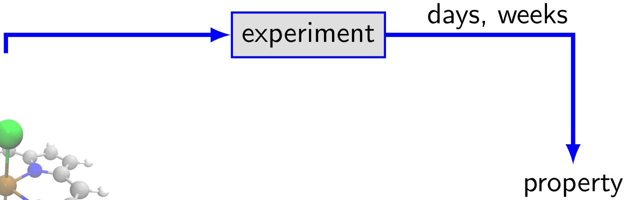
Why ML in chemical sciences?



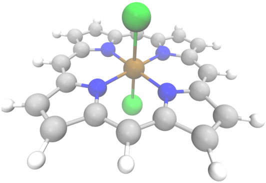
property

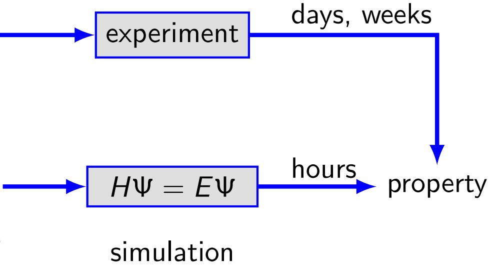
Why ML in chemical sciences?



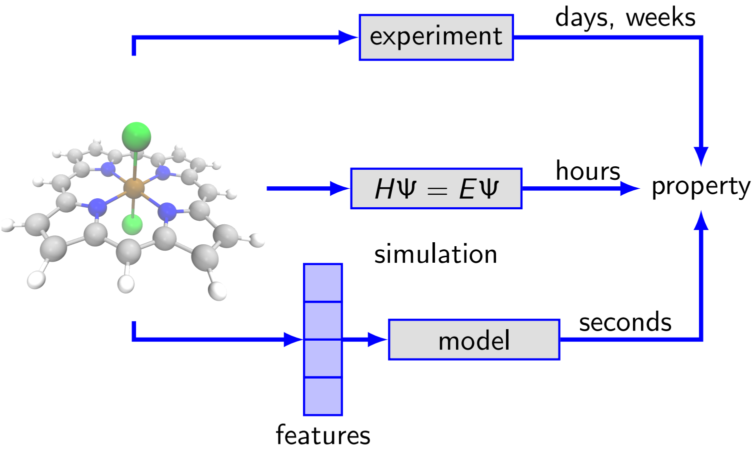


Why ML in chemical sciences?





Why ML in chemical sciences?



Why does ML seem to be taking over?

machine learning methods

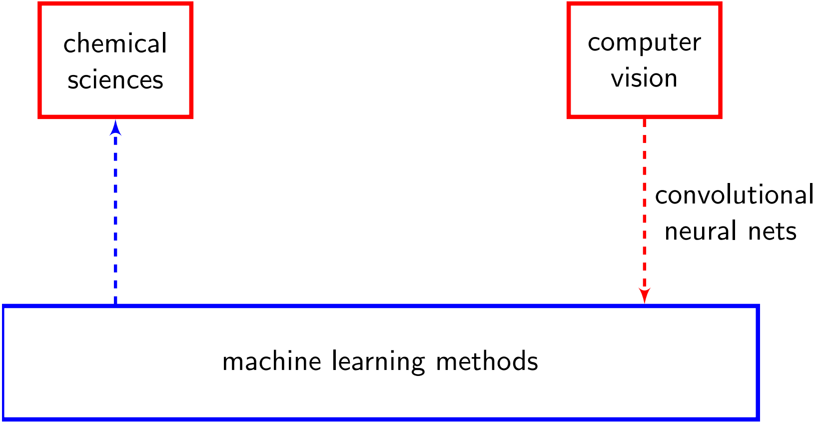
Why does ML seem to be taking over?

chemical sciences

computer vision

machine learning methods

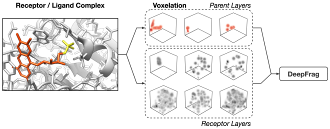
Why does ML seem to be taking over?



Why does ML seem to be taking over?

chemical sciences

computer vision

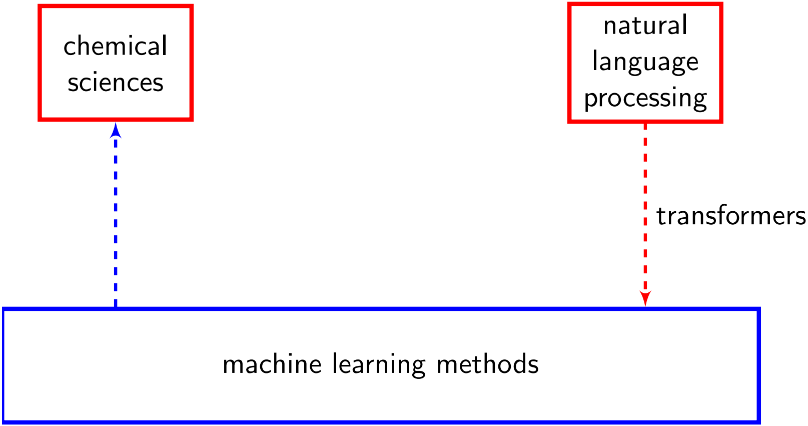


convolutional neural nets

Green, H., *et al.*, bioRxiv 2021.01.07.425790, 2021.

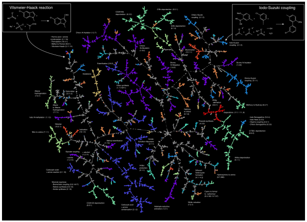
machine learning methods

Why does ML seem to be taking over?



Why does ML seem to be taking over?

chemical sciences



Schwaller, P., *et al.*, *Nat. Mach. Intell.*, 3: 144–152, 2021.

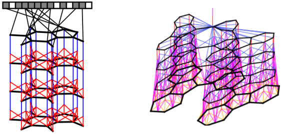
natural language processing

transformers

machine learning methods

Why does ML seem to be taking over?

chemical sciences



Duvenaud, D., *et al.*, *NeurIPS 28*, 2015.

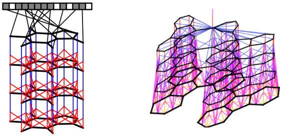
graph convolutions

machine learning methods

Why does ML seem to be taking over?

chemical sciences

diverse applications from traffic networks to medicine



Duvenaud, D., *et al.*, *NeurIPS 28*, 2015.

graph convolutions

machine learning methods

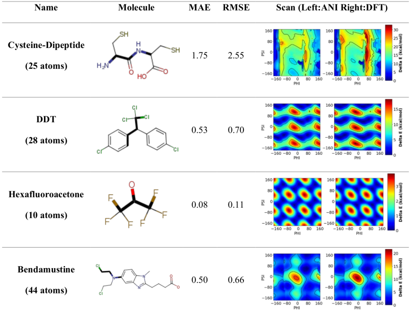
Future directions for ML in chemistry

Some areas of high current interest:

Future directions for ML in chemistry

Some areas of high current interest:

- Neural network potentials - quantum accuracy, force field cost. Reactive dynamics on your laptop!



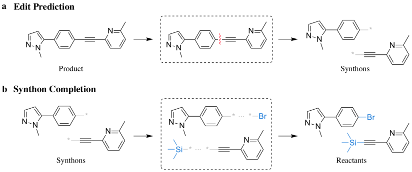
Name	Molecule	MAE	RMSE	Scan (Left:ANI Right:DFT)
Cysteine-Dipeptide (25 atoms)		1.75	2.55	
DDT (28 atoms)		0.53	0.70	
Hexafluoroacetone (10 atoms)		0.08	0.11	
Bendamustine (44 atoms)		0.50	0.66	

Devereux, C., *et al.*, *J. Chem. Theory Comput.*, 16(7):4192–4202, 2020

Future directions for ML in chemistry

Some areas of high current interest:

- Neural network potentials - quantum accuracy, force field cost. Reactive dynamics on your laptop!
- Synthesis planning and optimization. Fully automated chemistry!

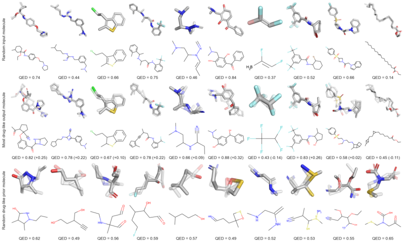


Somnath, V.R., *et al.*, arXiv:2006.07038v1, 2020

Future directions for ML in chemistry

Some areas of high current interest:

- Neural network potentials - quantum accuracy, force field cost. Reactive dynamics on your laptop!
- Synthesis planning and optimization. Fully automated chemistry!
- Generative models. Designing new drugs directly into the pocket, *de novo*!



Ragoza, M., *et al.*, arXiv:2010.08687v3, 2020

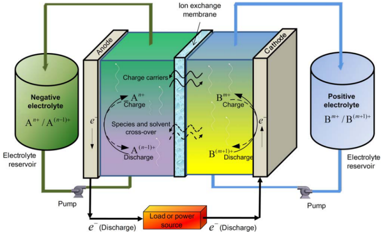
Table of Contents

Redox flow batteries

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

Redox flow batteries

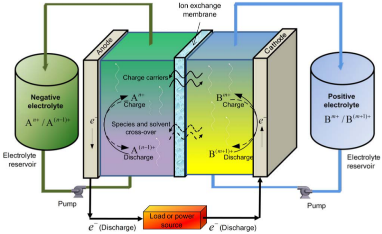
Redox flow batteries (RFBs) are a promising option for scalable energy storage:



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

Redox flow batteries

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

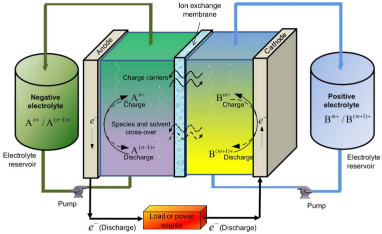


Perry, M.L. and Adam, Z., *J. Electrochem. Soc.*, 163(1):A5064–A5067, 2018.

Transition metal complexes make attractive redox couples for RFBs

Redox flow batteries

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



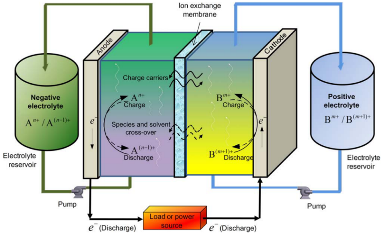
Perry, M.L. and Adam, Z., *J. Electrochem. Soc.*, 163(1):A5064–A5067, 2018.

Transition metal complexes make attractive redox couples for RFBs

- good ion stability (compared to organics)

Redox flow batteries

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



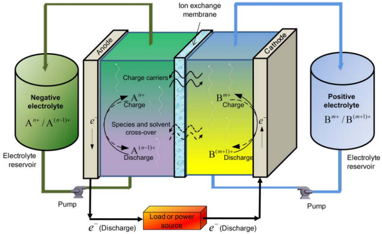
Perry, M.L. and Adam, Z., *J. Electrochem. Soc.*, 163(1):A5064–A5067, 2018.

Transition metal complexes make attractive redox couples for RFBs

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

Redox flow batteries

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



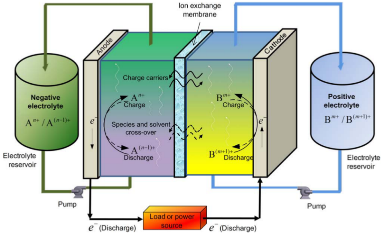
Perry, M.L. and Adam, Z., *J. Electrochem. Soc.*, 163(1):A5064–A5067, 2018.

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!

Redox flow batteries

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.

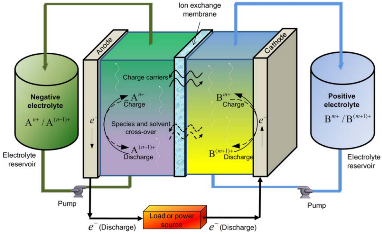
Transition metal complexes make attractive redox couples for RFBs

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

$$E_{\text{cell}} = 0.5 \times \Delta G_{\text{solv}} \times C \times n \times F$$

Redox flow batteries

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

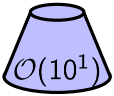
- good ion stability (compared to organics)
- good range of redox potentials available
- solubility is an issue!

$$E_{\text{cell}} = 0.5 \times \Delta G_{\text{solv}} \times C \times n \times F$$

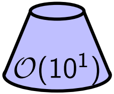
We need complexes that have high redox potential **and** good solubility

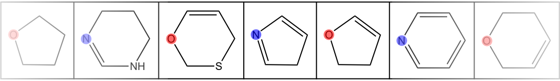
A design space for RFBs

A design space for RFBs



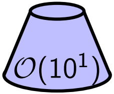
A design space for RFBs

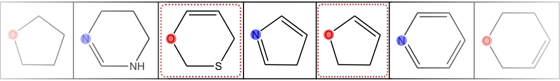




38 heterocycles

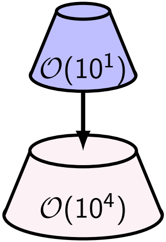
A design space for RFBs

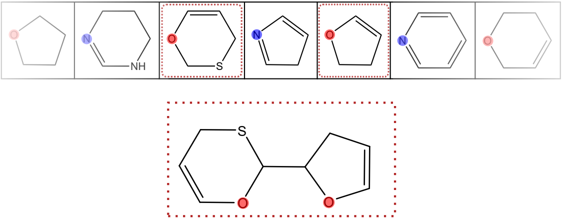




38 heterocycles

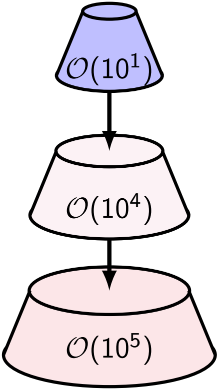
A design space for RFBs

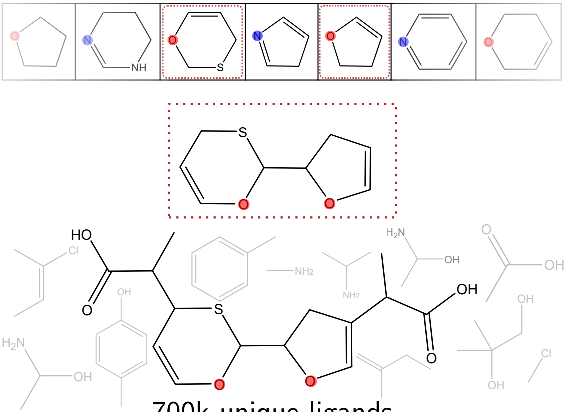




779 base ligands

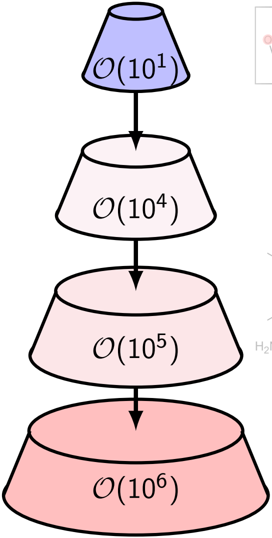
A design space for RFBs

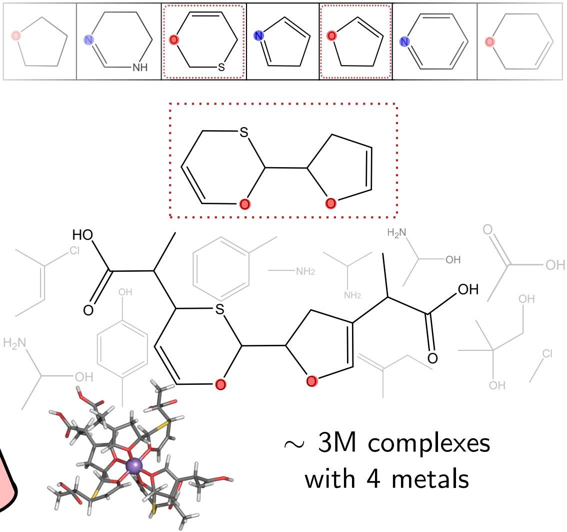




700k unique ligands

A design space for RFBs

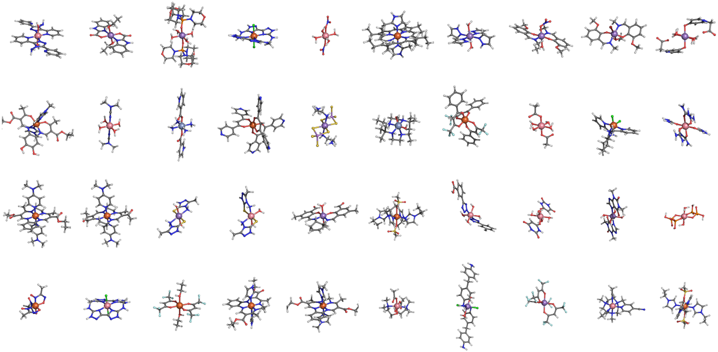




$\sim$ 3M complexes
with 4 metals

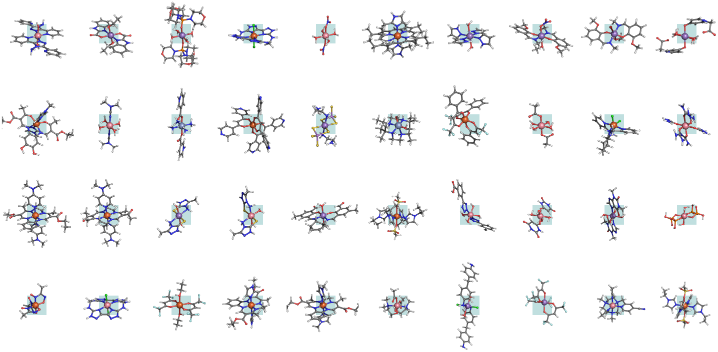
Computational approaches to chemical discovery

Computational methods can search for suitable complexes



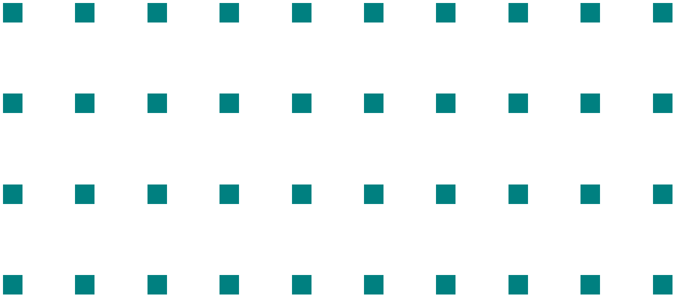
Computational approaches to chemical discovery

Computational methods can search for suitable complexes



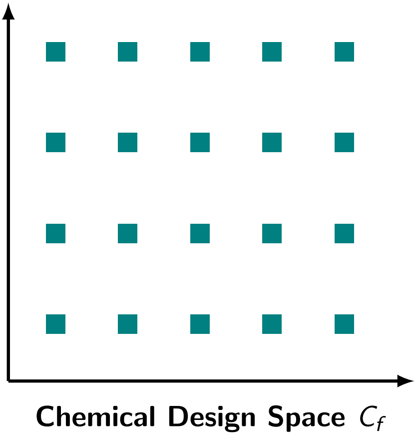
Computational approaches to chemical discovery

Computational methods can search for suitable complexes

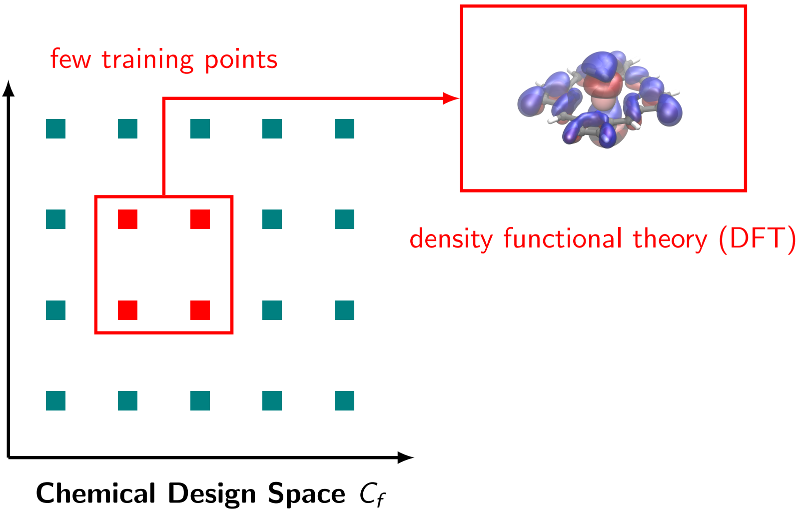


Computational approaches to chemical discovery

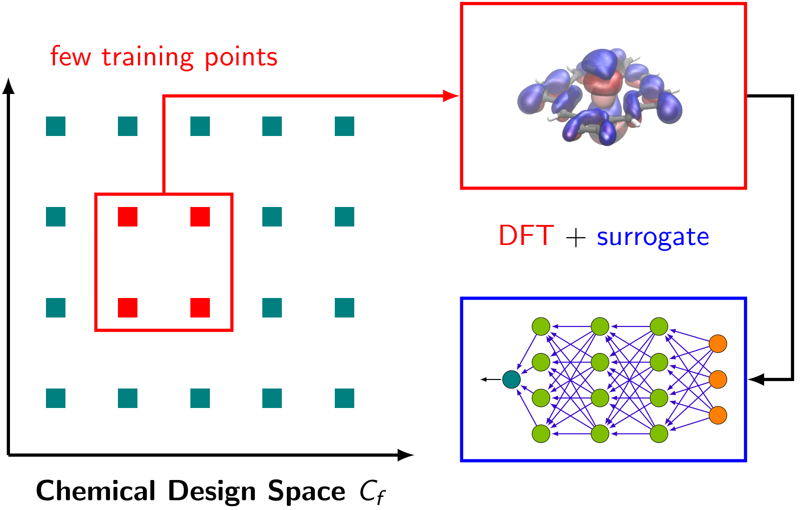
Computational methods can search for suitable complexes



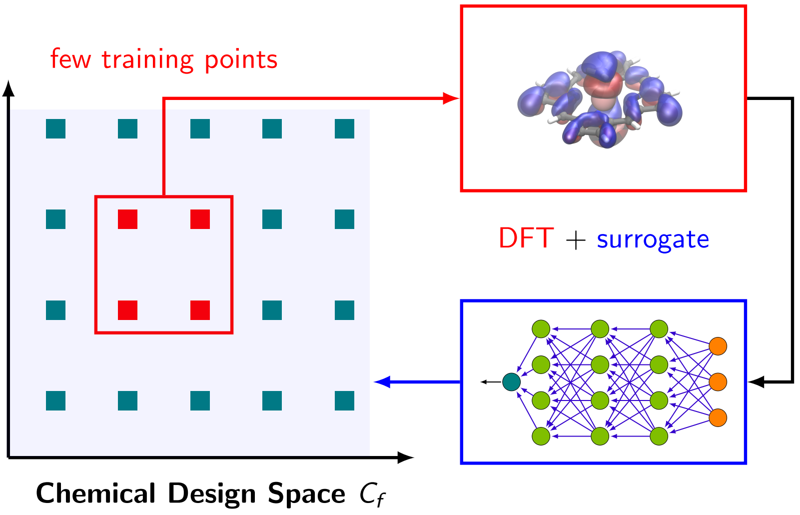
Computational approaches to chemical discovery



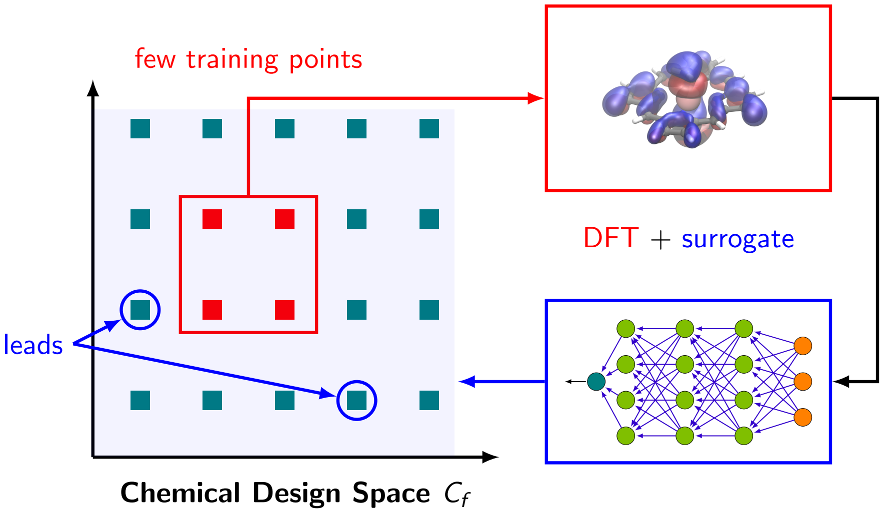
Computational approaches to chemical discovery



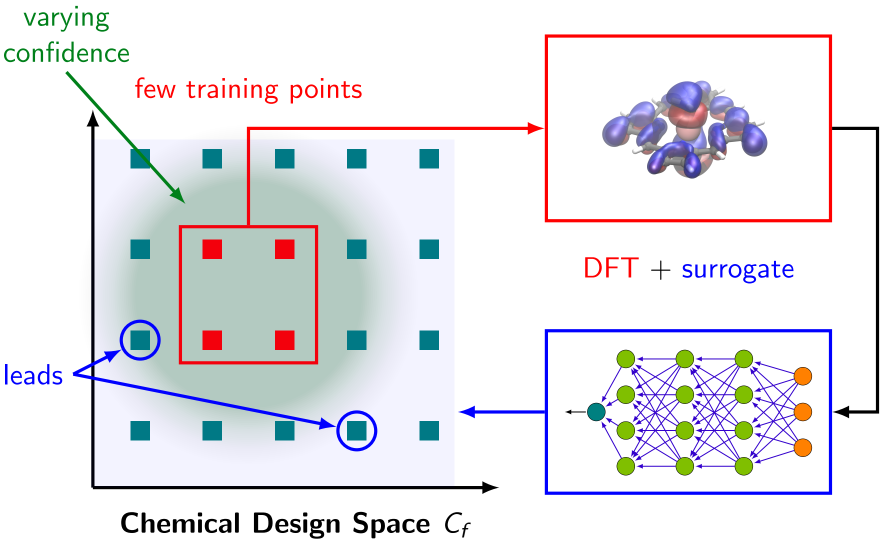
Computational approaches to chemical discovery



Computational approaches to chemical discovery



Computational approaches to chemical discovery



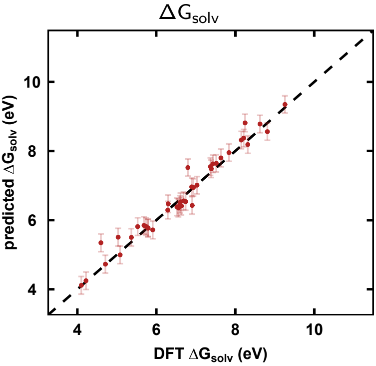
Multiobjective optimization

We can predict quantites of interest for our RFBs with ANNs

Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Multiobjective optimization

We can predict quantites of interest for our RFBs with ANNs

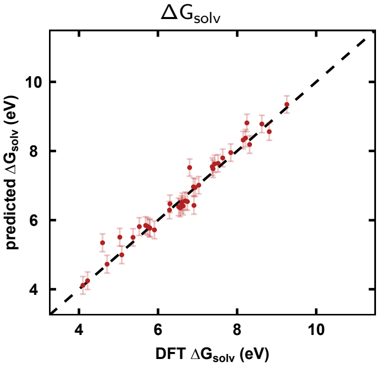


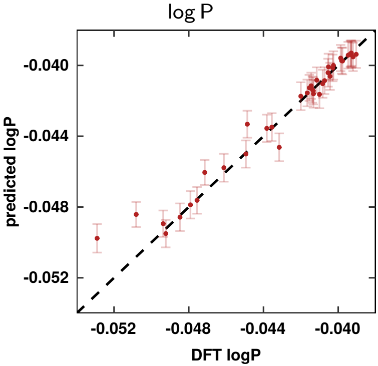
Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

3 × 100 tanh nodes, multitask, fully connected

Multiobjective optimization

We can predict quantites of interest for our RFBs with ANNs



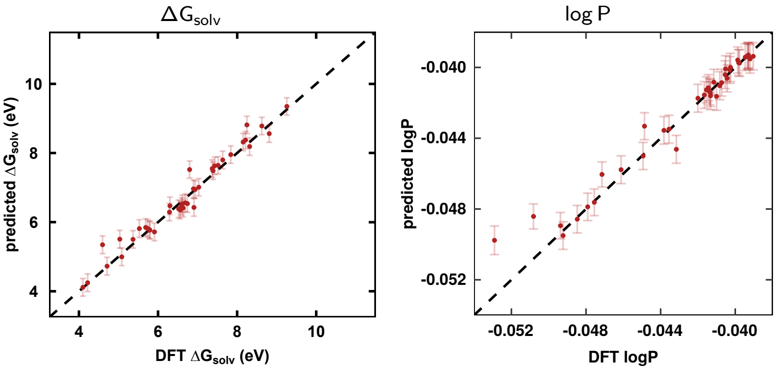


Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

3×100 tanh nodes, multitask, fully connected

Multiobjective optimization

We can predict quantites of interest for our RFBs with ANNs



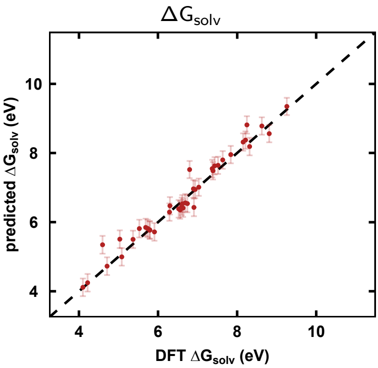
Screen $3M$ complexes in < 4 **minutes** on a regular workstation, c.f. 50 **GPU-years** with DFT

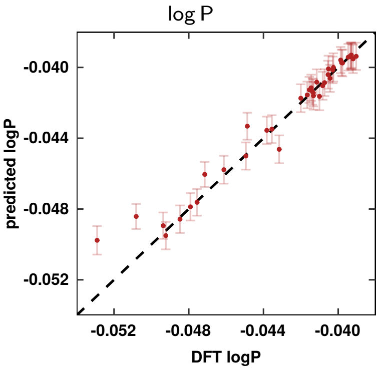
Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

3×100 tanh nodes, multitask, fully connected

Multiobjective optimization

We can predict quantites of interest for our RFBs with ANNs





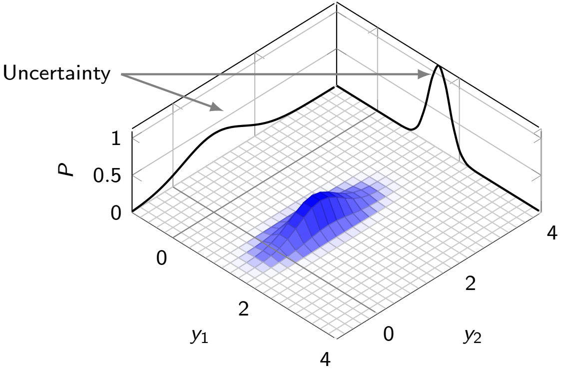
$$\begin{matrix}\Delta G_{solv} \\ \log P\end{matrix} = \begin{bmatrix}\hat{y}_1 \\ \hat{y}_2\end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix}\hat{\mu}_1 \\ \hat{\mu}_2\end{bmatrix}, \begin{bmatrix}\hat{\sigma}_1^2 & 0 \\ 0 & \hat{\sigma}_2^2\end{bmatrix}\right)$$

Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

3×100 tanh nodes, multitask, fully connected

Multiobjective optimization

We can predict quantites of interest for our RFBs with ANNs



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

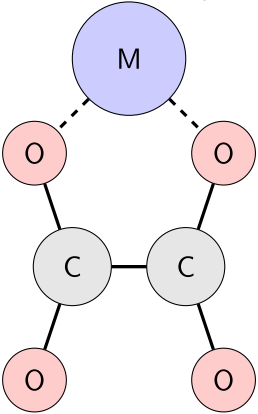
Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

3×100 tanh nodes, multitask, fully connected

Machine learning methods

Featurization:

Graph-based features (RACs) designed for TM complexes:

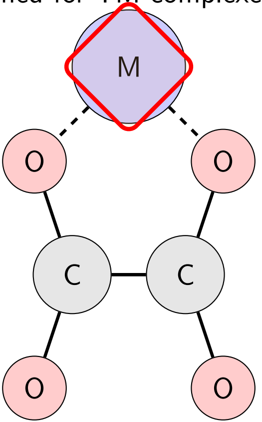


Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

Machine learning methods

Featurization:

Graph-based features (RACs) designed for TM complexes:

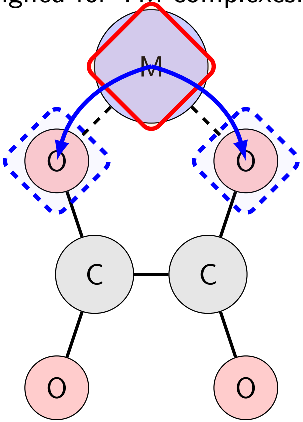


Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

Machine learning methods

Featurization:

Graph-based features (RACs) designed for TM complexes:



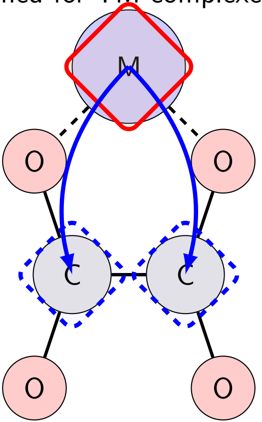
$$d_1 : \sum_{M,O} Z_M Z_O$$

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

Machine learning methods

Featurization:

Graph-based features (RACs) designed for TM complexes:



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

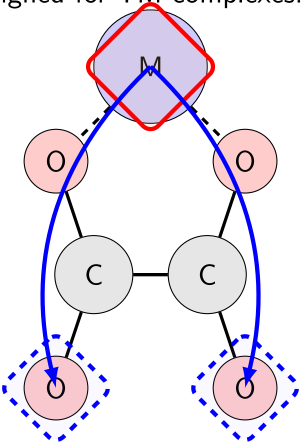
$$d_2 : \sum_{M,C} Z_M Z_C$$

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

Machine learning methods

Featurization:

Graph-based features (RACs) designed for TM complexes:



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

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

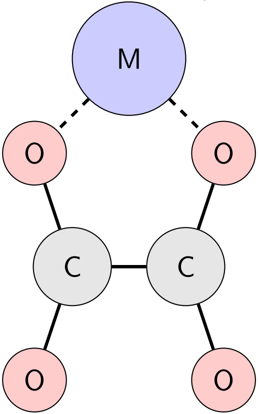
$$d_3 : \sum_{M,O} Z_M Z_O$$

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

Machine learning methods

Featurization:

Graph-based features (RACs) designed for TM complexes:



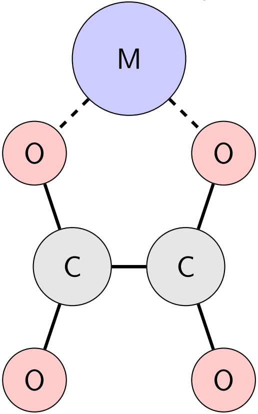
Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

Regression:

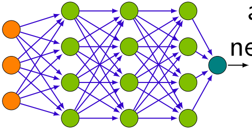
Machine learning methods

Featurization:

Graph-based features (RACs) designed for TM complexes:



Regression:



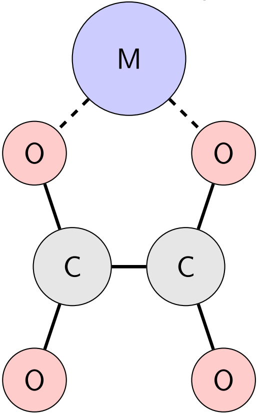
artifical neural networks (ANNs)

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

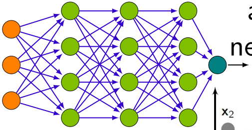
Machine learning methods

Featurization:

Graph-based features (RACs) designed for TM complexes:

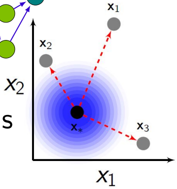


Regression:



artifical neural networks (ANNs)

Kernel methods (KRR)

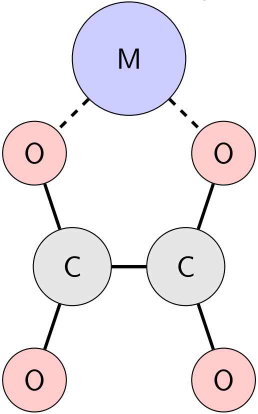


Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

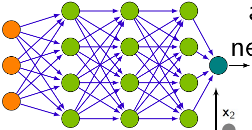
Machine learning methods

Featurization:

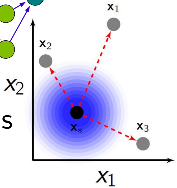
Graph-based features (RACs) designed for TM complexes:



Regression:

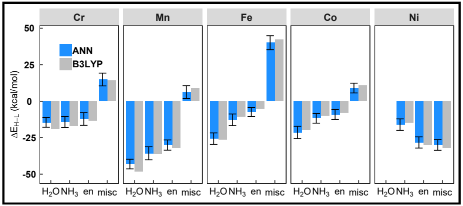


artifical neural networks (ANNs)



Kernel methods (KRR)

spin splitting energies



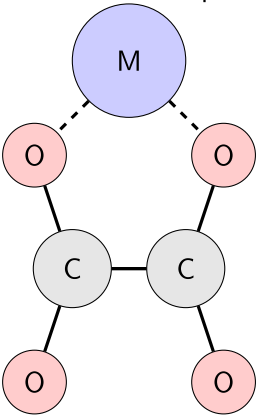
Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

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

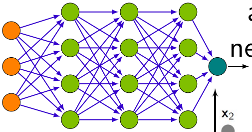
Machine learning methods

Featurization:

Graph-based features (RACs) designed for TM complexes:

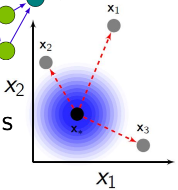


Regression:

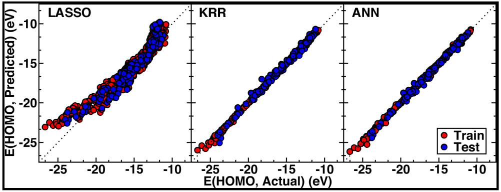


artifical neural networks (ANNs)

Kernel methods (KRR)



frontier orbital properties

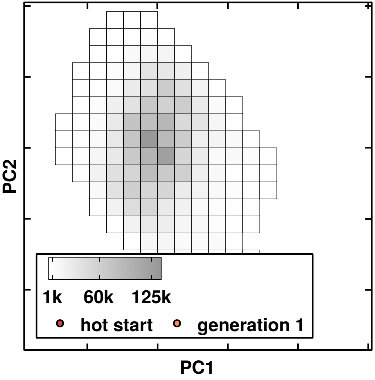


Janet, J.P., and Kulik, H.J., *J. Phys. Chem. A*, 121(46):8939–8954, 2017.

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

Design space and clustering

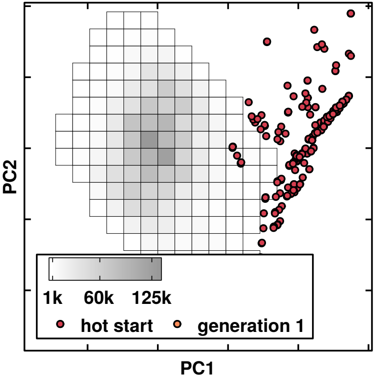
Jump start the design with diversity-oriented cluster:



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Design space and clustering

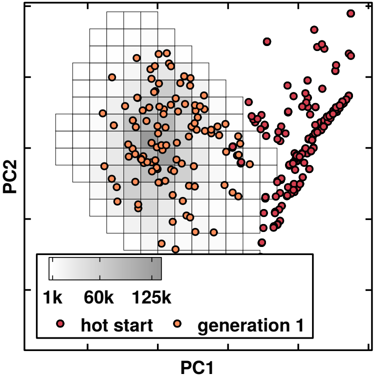
Jump start the design with diversity-oriented cluster:



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Design space and clustering

Jump start the design with diversity-oriented cluster:

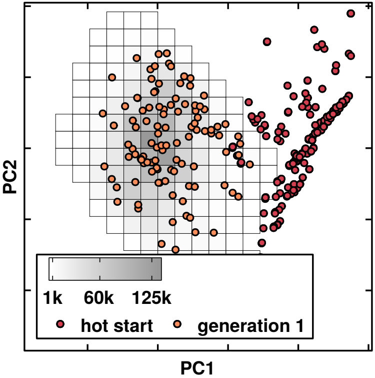


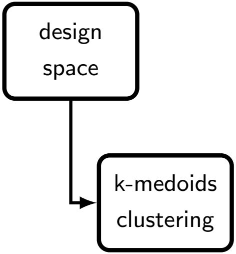
design space

Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Design space and clustering

Jump start the design with diversity-oriented cluster:

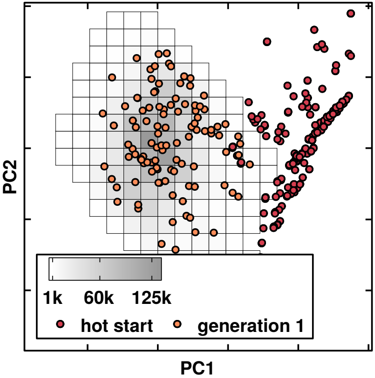


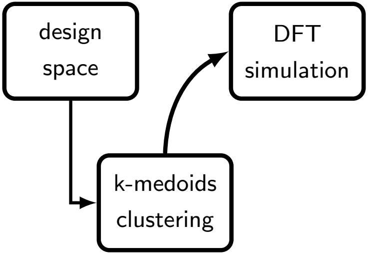


Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Design space and clustering

Jump start the design with diversity-oriented cluster:

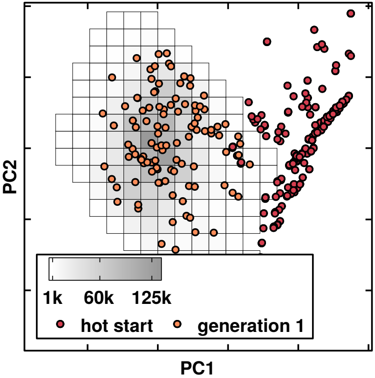


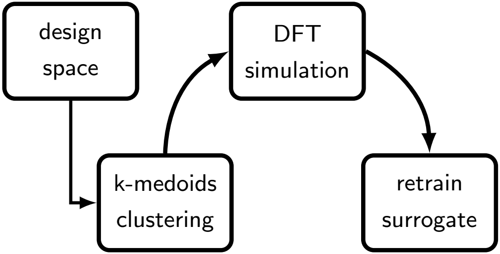


Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Design space and clustering

Jump start the design with diversity-oriented cluster:

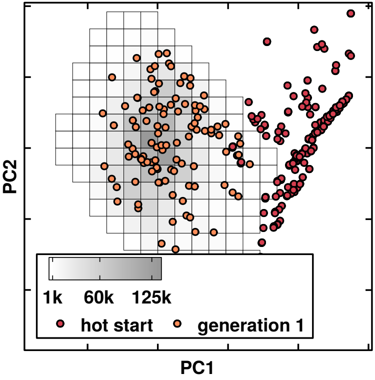


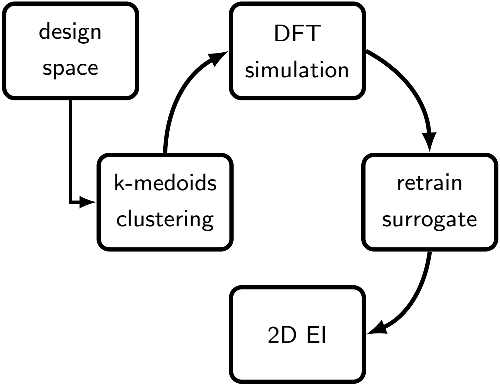


Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Design space and clustering

Jump start the design with diversity-oriented cluster:



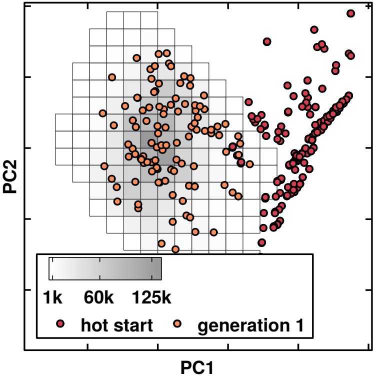


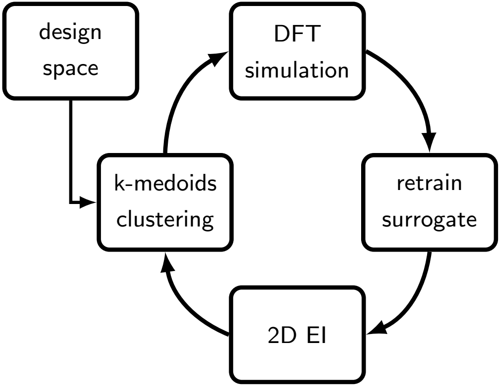
Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

multitask: 2 × 100 tanh nodes, fully connected

Design space and clustering

Jump start the design with diversity-oriented cluster:



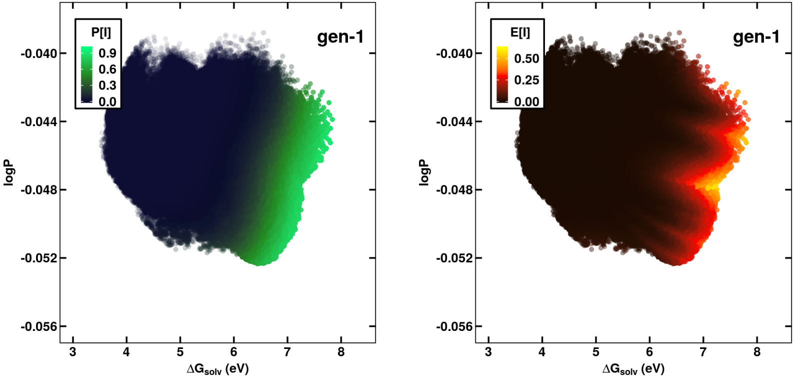


Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

multitask: 2 × 100 tanh nodes, fully connected

Evolution of PI and EI

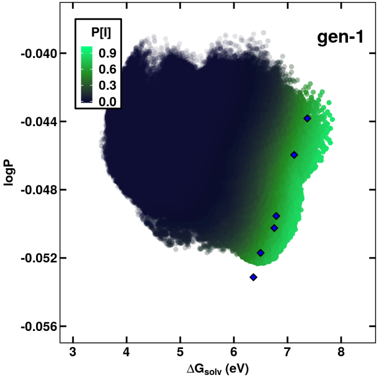
probability of improvement expected improvement



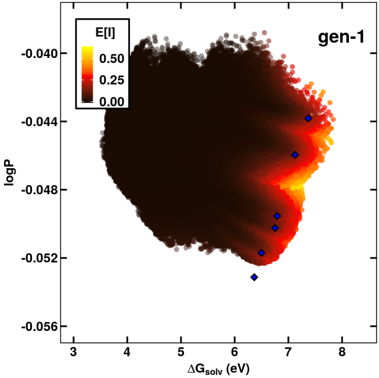
Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Evolution of PI and EI

probability of improvement

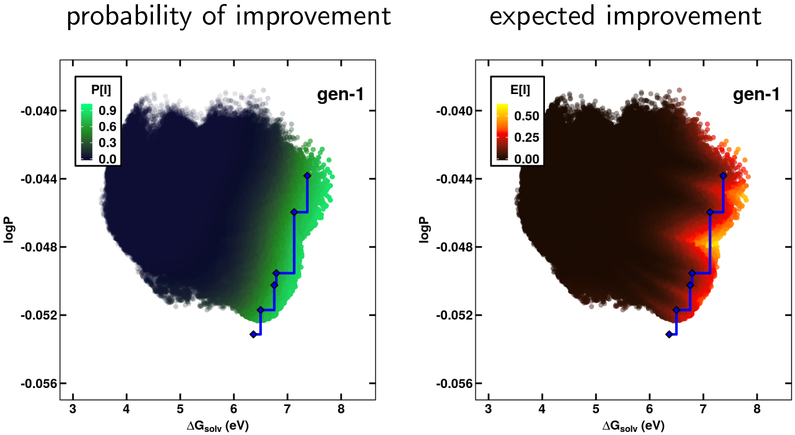


expected improvement



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

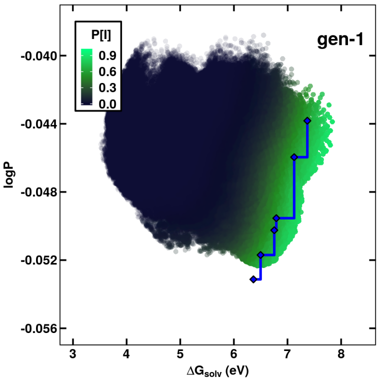
Evolution of PI and EI



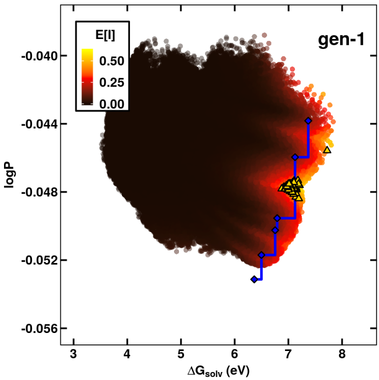
Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Evolution of PI and EI

probability of improvement



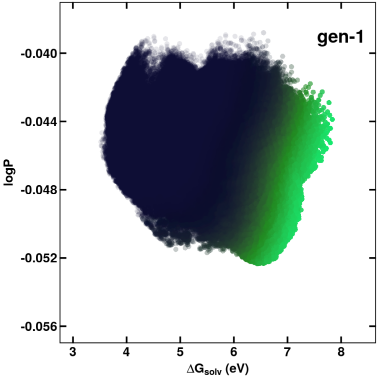
expected improvement



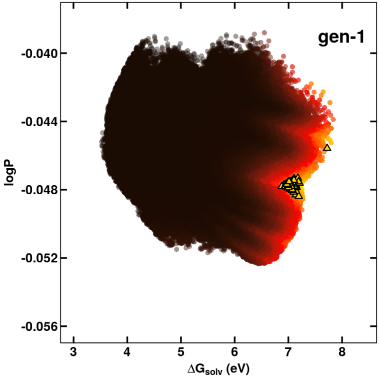
Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Evolution of PI and EI

probability of improvement



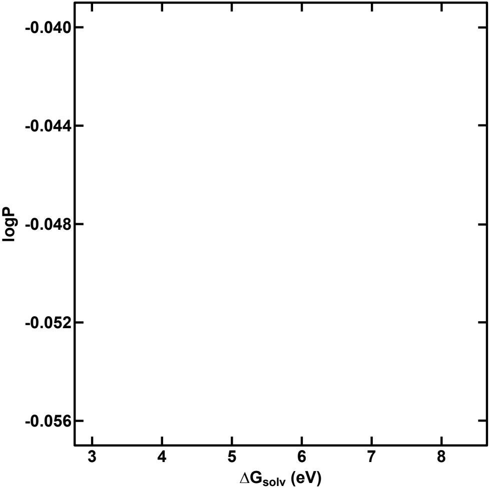
expected improvement



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Simulation results

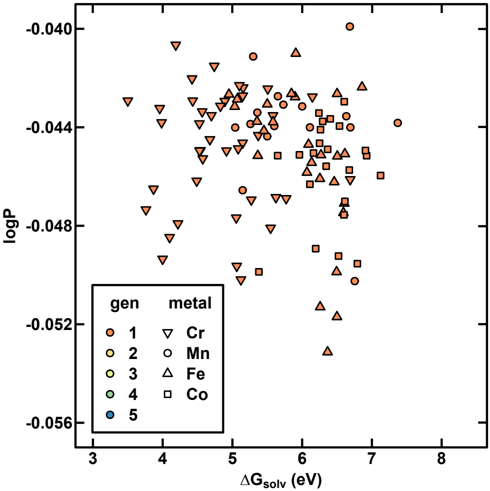
DFT Data



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Simulation results

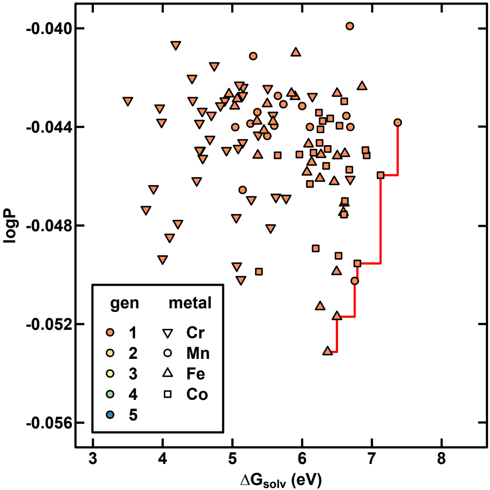
k-medoids points (generation 1)



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Simulation results

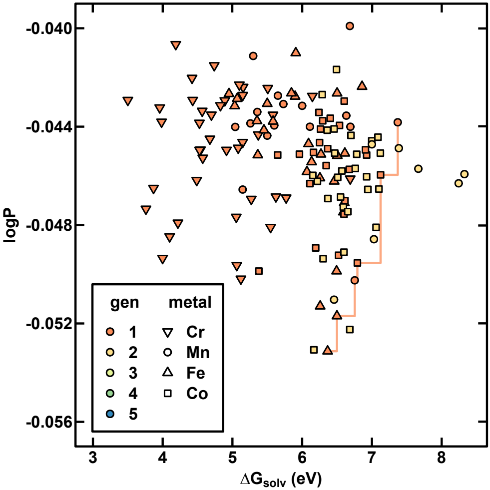
pareto front (generation 1)



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Simulation results

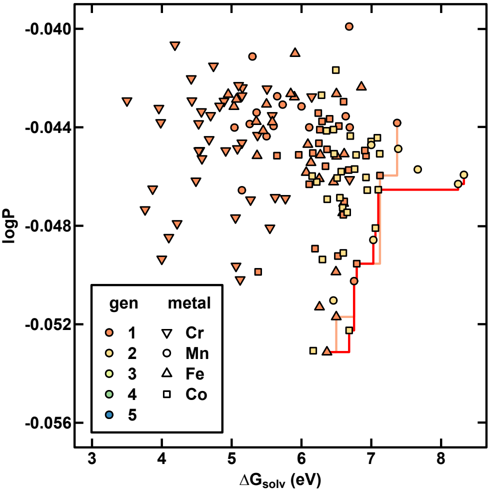
EI points (geneneration 2)



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Simulation results

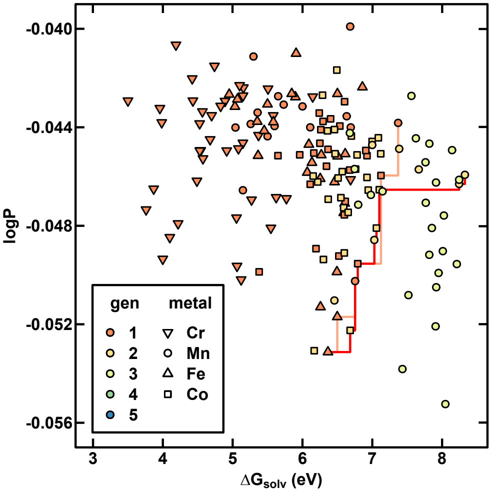
pareto front (generation 2)



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Simulation results

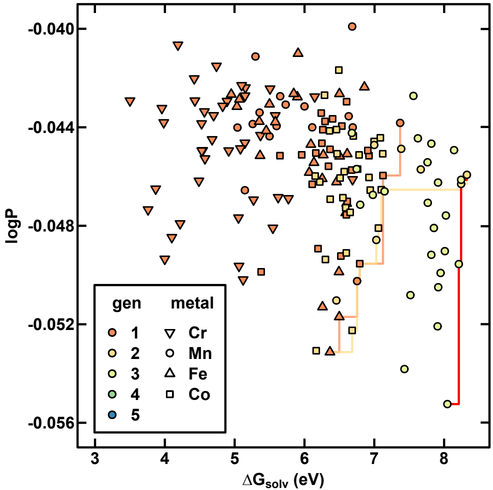
EI points (generation 3)



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Simulation results

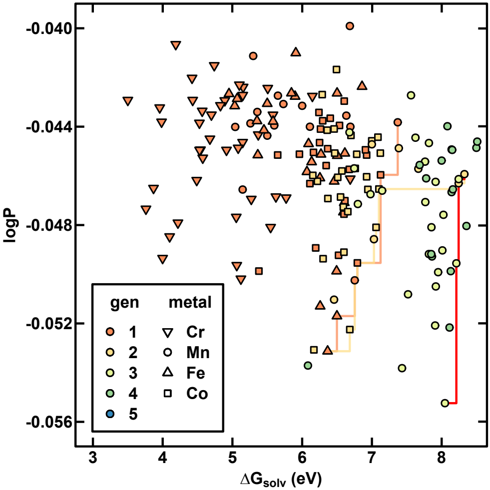
pareto front (generation 3)



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Simulation results

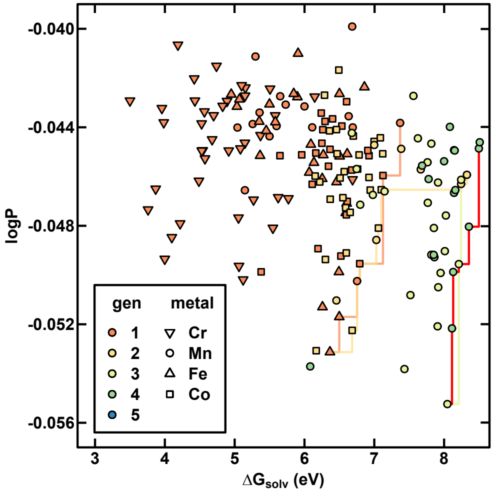
EI points (generation 4)



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Simulation results

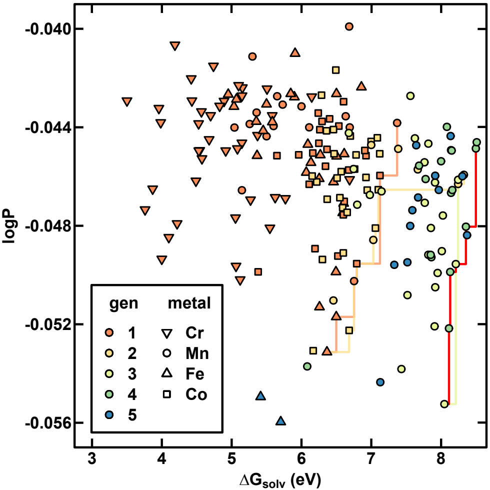
pareto front (generation 4)



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Simulation results

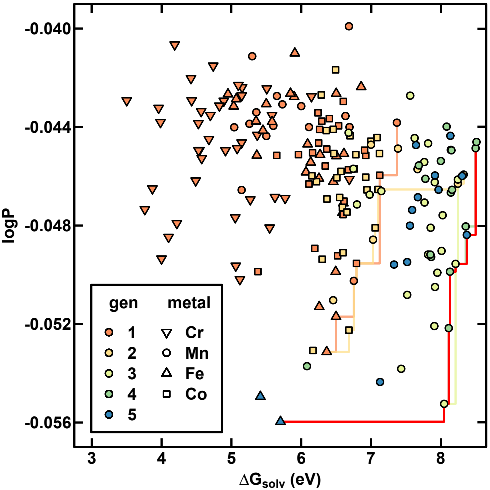
EI points (generation 5)



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

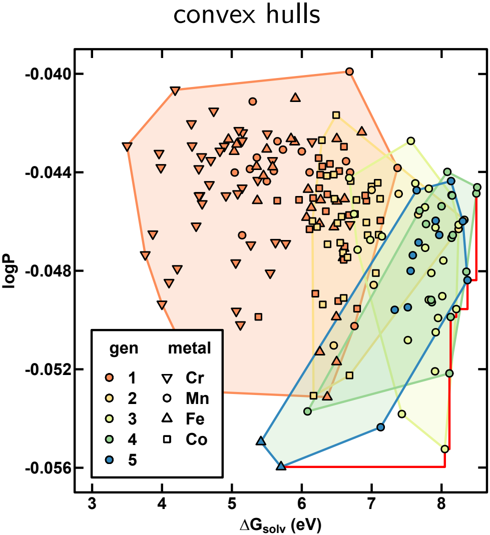
Simulation results

pareto front (generation 5)



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Simulation results

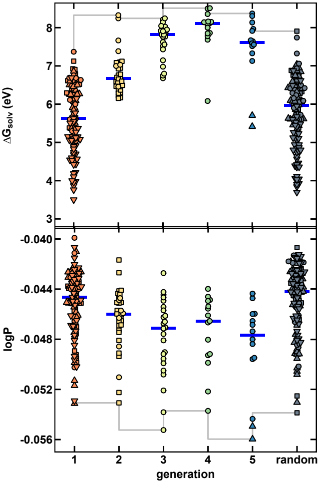


Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Conclusions

Conclusions

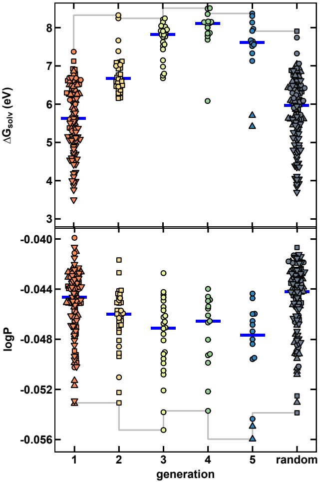
- EI framework provides high resolution in the region of interest (c.f. maximum uncertainty), converges quickly



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Conclusions

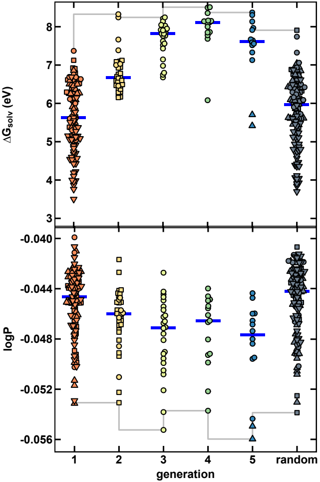
- EI framework provides high resolution in the region of interest (c.f. maximum uncertainty), converges quickly
- We are able to identify fruitful regions from large chemical spaces based on few DFT evaluations



Janet, J.P., *et al.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

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.*, *ACS Cent. Sci.*, 6(4):513–524, 2020

Acknowledgments

This work is thanks to the Kulik group and funding partners:

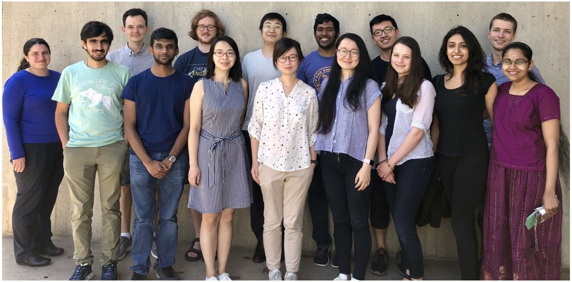
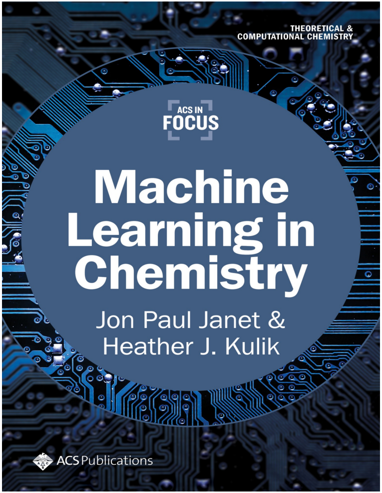


Table of Contents

Machine learning in chemistry book

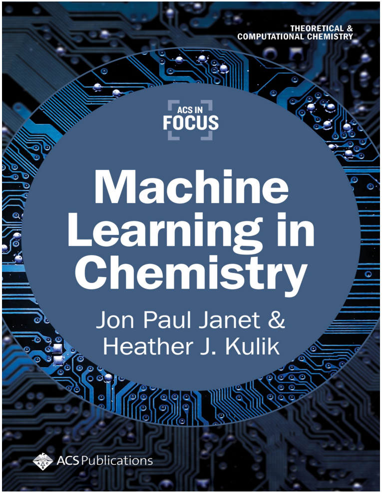
Introduces everything needed to work with common machine learning tools in the context of chemical sciences:



Machine learning in chemistry book

Introduces everything needed to work with common machine learning tools in the context of chemical sciences:

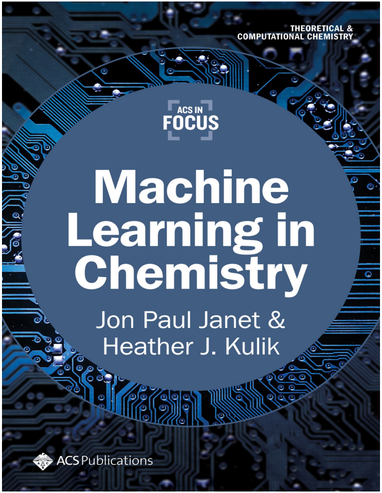
1. History and context



Machine learning in chemistry book

Introduces everything needed to work with common machine learning tools in the context of chemical sciences:

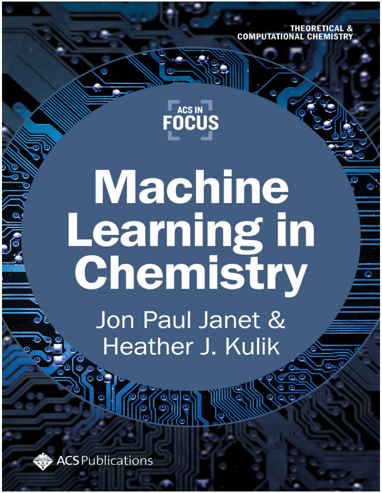
1. History and context
2. Statistical learning



Machine learning in chemistry book

Introduces everything needed to work with common machine learning tools in the context of chemical sciences:

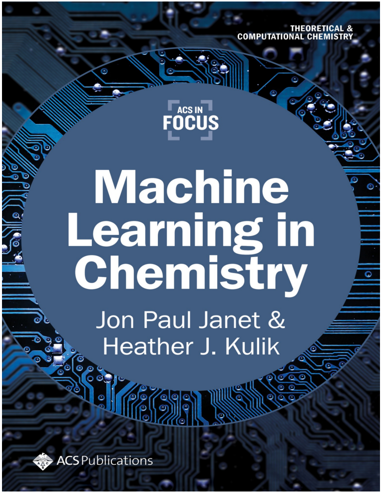
1. History and context
2. Statistical learning
3. Linear and kernel models



Machine learning in chemistry book

Introduces everything needed to work with common machine learning tools in the context of chemical sciences:

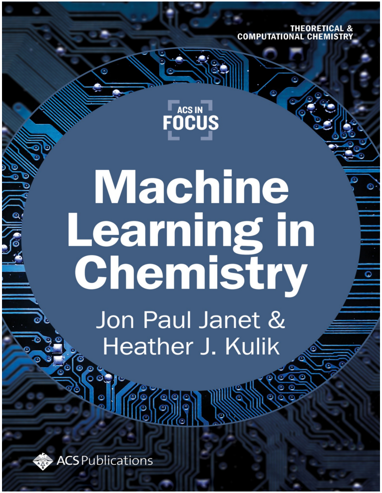
1. History and context
2. Statistical learning
3. Linear and kernel models
4. Representations and feature Selection



Machine learning in chemistry book

Introduces everything needed to work with common machine learning tools in the context of chemical sciences:

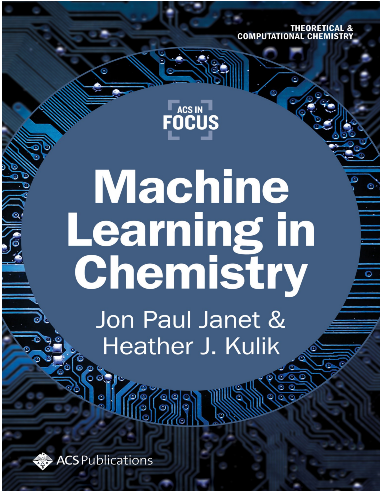
1. History and context
2. Statistical learning
3. Linear and kernel models
4. Representations and feature Selection
5. Neural networks and representation learning



Machine learning in chemistry book

Introduces everything needed to work with common machine learning tools in the context of chemical sciences:

1. History and context
2. Statistical learning
3. Linear and kernel models
4. Representations and feature Selection
5. Neural networks and representation learning
6. Practical advice



C2: Supervised learning

Supervised learning methods attempt to connect patterns in data to known endpoints by learning model parameters that reproduce the observed relationship.

C2: Supervised learning

Supervised learning methods attempt to connect patterns in data to known endpoints by learning model parameters that reproduce the observed relationship.

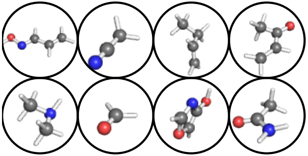
observation property

C2: Supervised learning

Supervised learning methods attempt to connect patterns in data to known endpoints by learning model parameters that reproduce the observed relationship.

observation

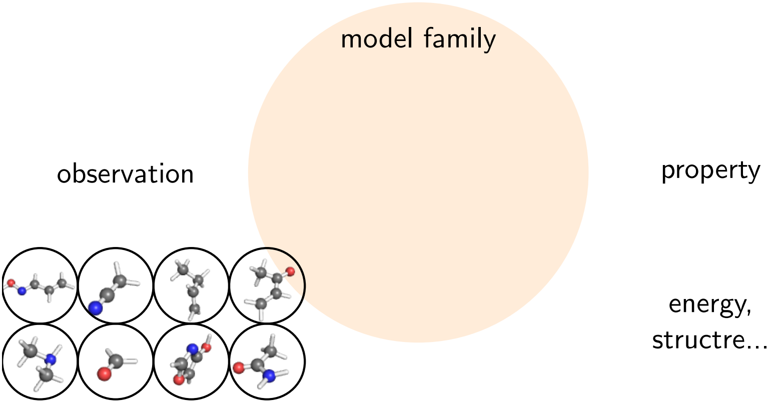
property



energy, structre...

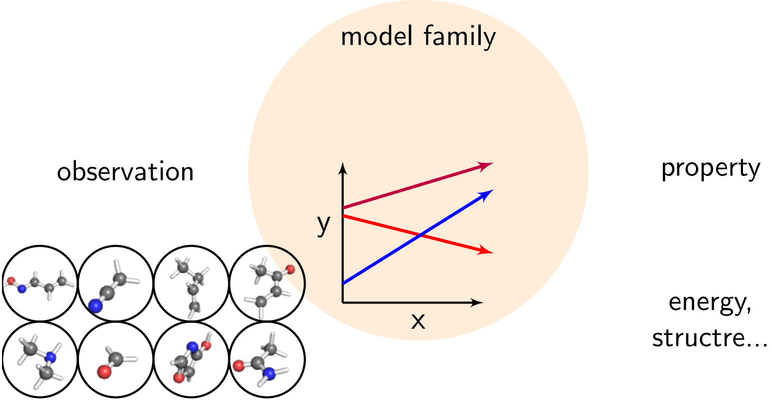
C2: Supervised learning

Supervised learning methods attempt to connect patterns in data to known endpoints by learning model parameters that reproduce the observed relationship.



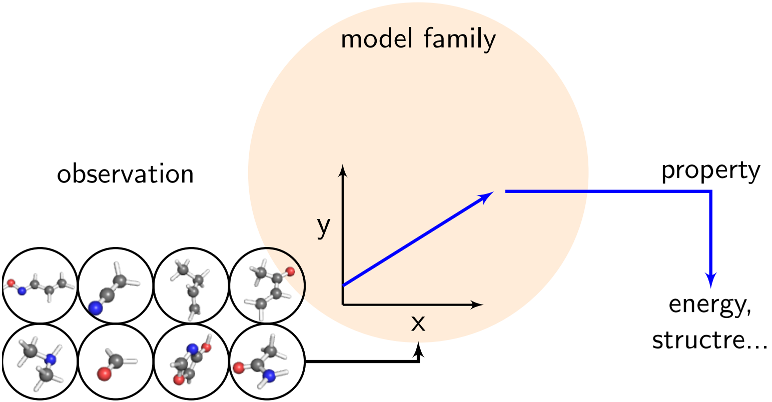
C2: Supervised learning

Supervised learning methods attempt to connect patterns in data to known endpoints by learning model parameters that reproduce the observed relationship.



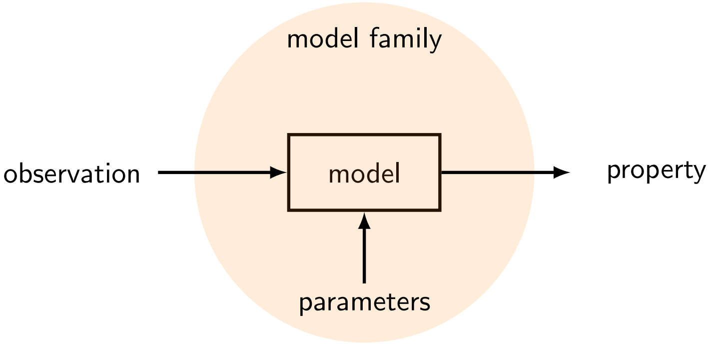
C2: Supervised learning

Supervised learning methods attempt to connect patterns in data to known endpoints by learning model parameters that reproduce the observed relationship.



C2: Supervised learning

Supervised learning methods attempt to connect patterns in data to known endpoints by learning model parameters that reproduce the observed relationship.



C2: Statistical learning and generalization

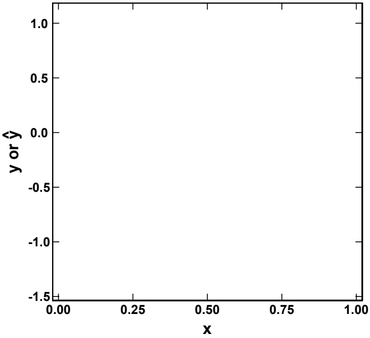
We need to understand how models can generalize, i.e. predict previously unseen data (or not). *Statistical learning theory* allows us to study this behaviour.

C2: Statistical learning and generalization

We need to understand how models can generalize, i.e. predict previously unseen data (or not). *Statistical learning theory* allows us to study this behaviour.

Let us use **polynomials** to estimate:

$$y(x) = \sin(2\pi x)$$



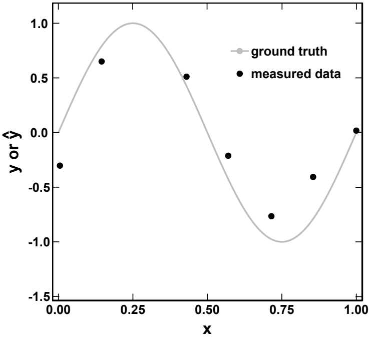
C2: Statistical learning and generalization

We need to understand how models can generalize, i.e. predict previously unseen data (or not). *Statistical learning theory* allows us to study this behaviour.

Let us use **polynomials** to estimate:

$$y(x) = \sin(2\pi x)$$

Assume 8 measurements with noise $\mathcal{N}(0, 0.2)$



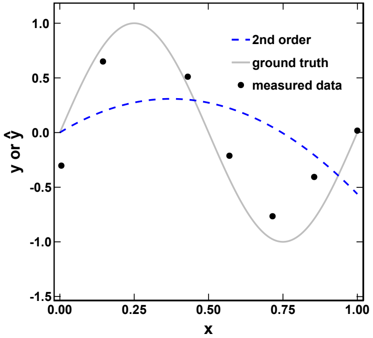
C2: Statistical learning and generalization

We need to understand how models can generalize, i.e. predict previously unseen data (or not). *Statistical learning theory* allows us to study this behaviour.

Let us use **polynomials** to estimate:

$$y(x) = \sin(2\pi x)$$

Start with degree 2...



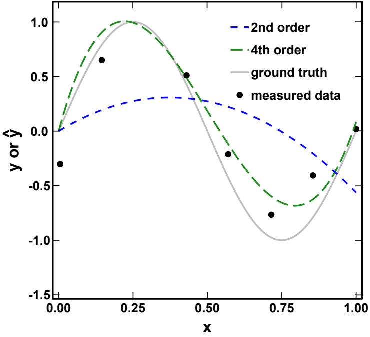
C2: Statistical learning and generalization

We need to understand how models can generalize, i.e. predict previously unseen data (or not). *Statistical learning theory* allows us to study this behaviour.

Let us use **polynomials** to estimate:

$$y(x) = \sin(2\pi x)$$

Start with degree 2...
What happens when we increase the degree ?



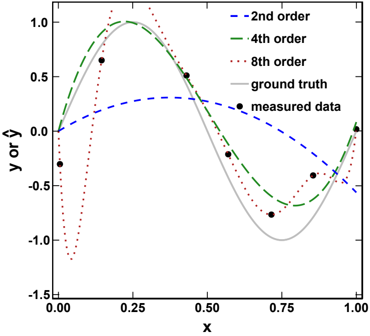
C2: Statistical learning and generalization

We need to understand how models can generalize, i.e. predict previously unseen data (or not). *Statistical learning theory* allows us to study this behaviour.

Let us use **polynomials** to estimate:

$$y(x) = \sin(2\pi x)$$

Start with degree 2...
What happens when we increase the degree ?



C2: Statistical learning and generalization

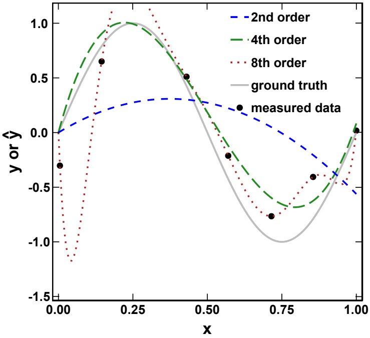
We need to understand how models can generalize, i.e. predict previously unseen data (or not). *Statistical learning theory* allows us to study this behaviour.

Let us use **polynomials** to estimate:

$$y(x) = \sin(2\pi x)$$

Empirical risk: error on training data

True risk: error over the whole domain



C2: Statistical learning and generalization

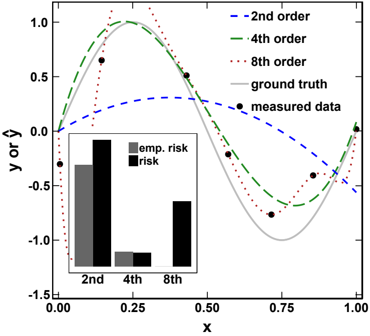
We need to understand how models can generalize, i.e. predict previously unseen data (or not). *Statistical learning theory* allows us to study this behaviour.

Let us use **polynomials** to estimate:

$$y(x) = \sin(2\pi x)$$

Empirical risk: error on training data

True risk: error over the whole domain



C2: Statistical learning and generalization

We need to understand how models can generalize, i.e. predict previously unseen data (or not). *Statistical learning theory* allows us to study this behaviour.

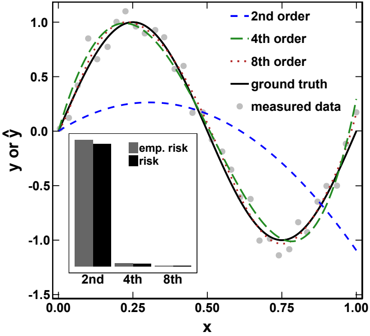
Let us use **polynomials** to estimate:

$$y(x) = \sin(2\pi x)$$

Empirical risk: error on training data

True risk: error over the whole domain

What happens if we add more data?

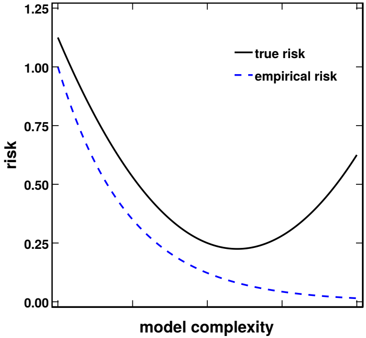


C2: Statistical learning and generalization

We need to understand how models can generalize, i.e. predict previously unseen data (or not). *Statistical learning theory* allows us to study this behaviour.

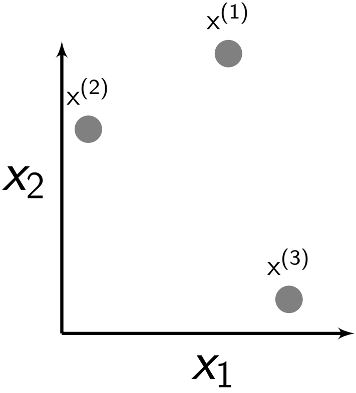
We cannot choose model complexity (hyperparameters, regularization) based on training data.

Cross-validation (and related techniques) must be used to compare models.



C3: Linear and nonlinear kernels

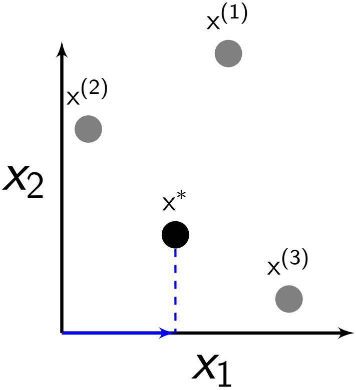
Linear models serve a tool to understand nonlinear models, regularization



linear model

C3: Linear and nonlinear kernels

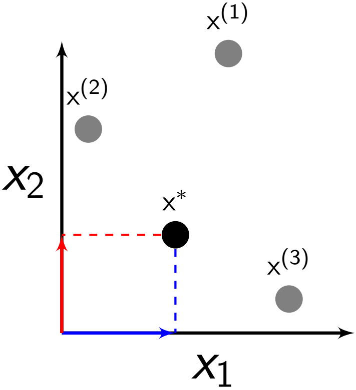
Linear models serve a tool to understand nonlinear models, regularization



linear model

C3: Linear and nonlinear kernels

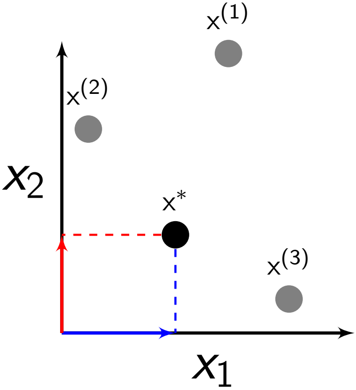
Linear models serve a tool to understand nonlinear models, regularization



linear model

C3: Linear and nonlinear kernels

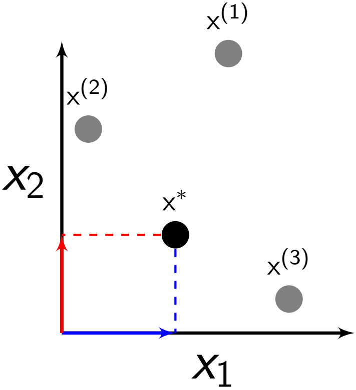
$$y(x^*) = w_1 x_1^* + w_2 x_2^*$$



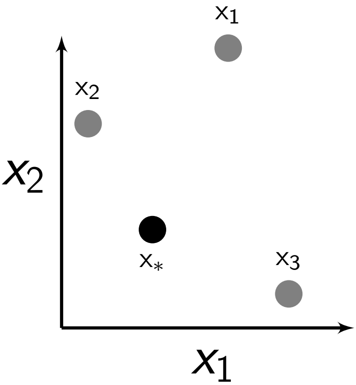
linear model

C3: Linear and nonlinear kernels

$$y(x^*) = w_1 x_1^* + w_2 x_2^*$$



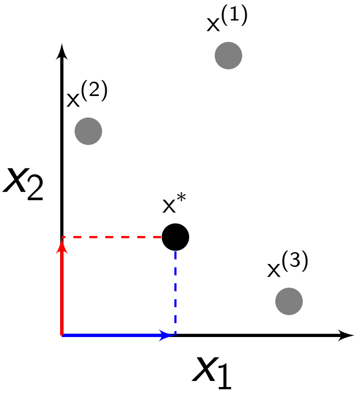
linear model



Gaussian kernel

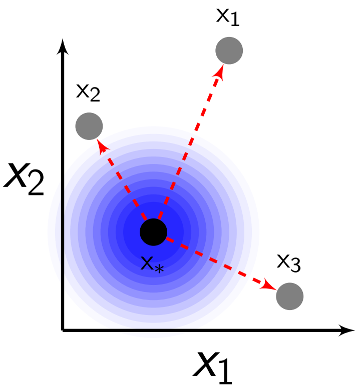
C3: Linear and nonlinear kernels

$$y(x^*) = w_1 x_1^* + w_2 x_2^*$$



linear model

$$y(x^*) = \sum_{i=1}^{n} a_i k(x^*, x^{(i)})$$

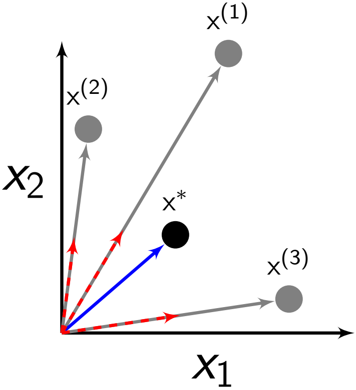


Gaussian kernel

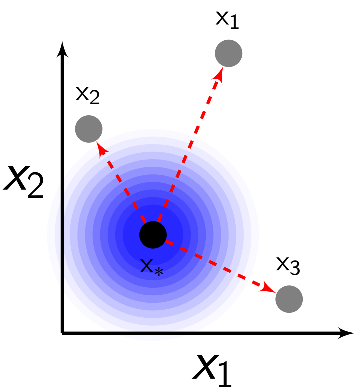
C3: Linear and nonlinear kernels

$y(x^*) = \sum_{i=1}^{n} a_i k(x^*, x^{(i)})$

$y(x^*) = \sum_{i=1}^{n} a_i k(x^*, x^{(i)})$

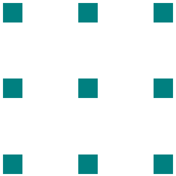


linear kernel



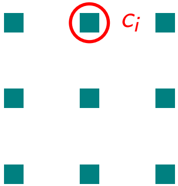
Gaussian kernel

C4: Representing chemical systems



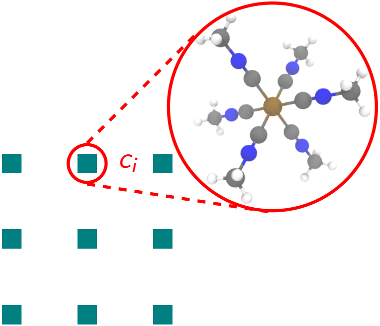
Chemical Space C_f

C4: Representing chemical systems



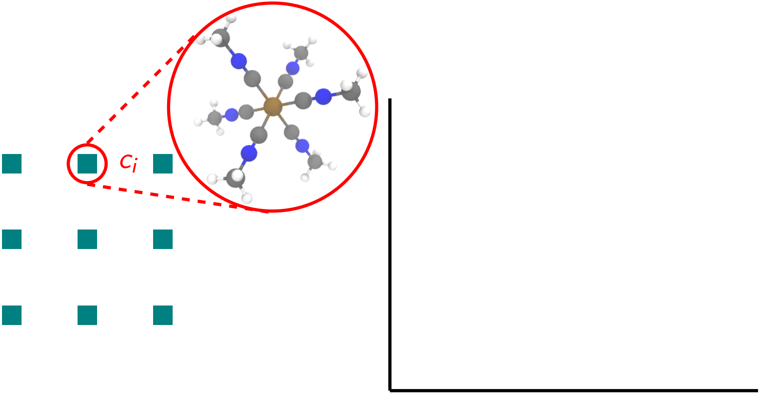
Chemical Space C_f

C4: Representing chemical systems



Chemical Space C_f

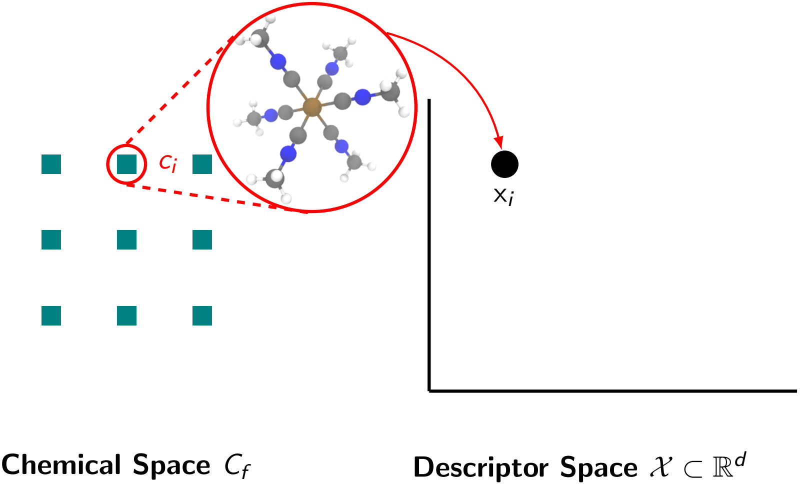
C4: Representing chemical systems



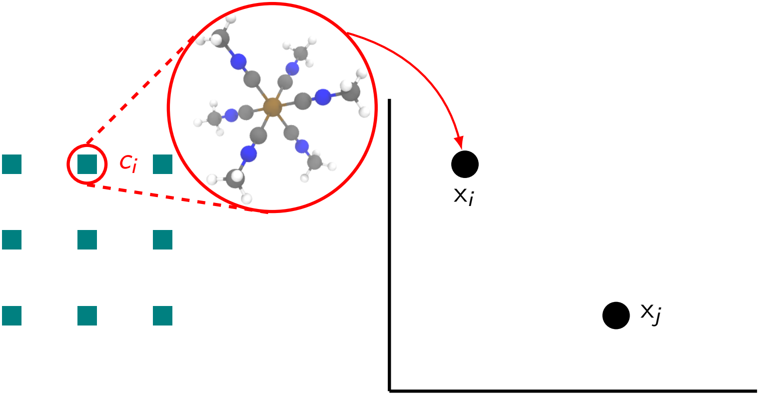
Chemical Space C_f

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

C4: Representing chemical systems



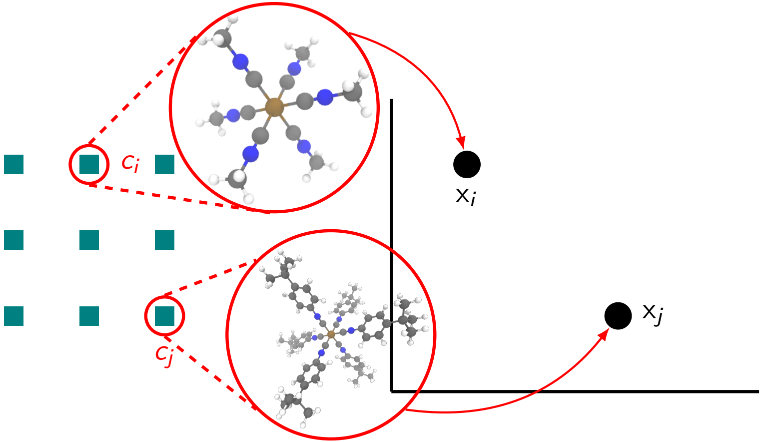
C4: Representing chemical systems



Chemical Space C_f

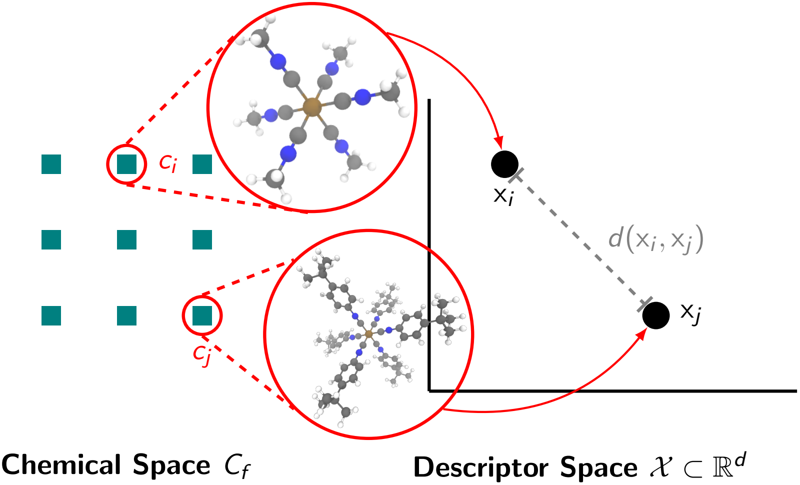
Descriptor Space $\mathcal{X} \subset \mathbb{R}^d$

C4: Representing chemical systems

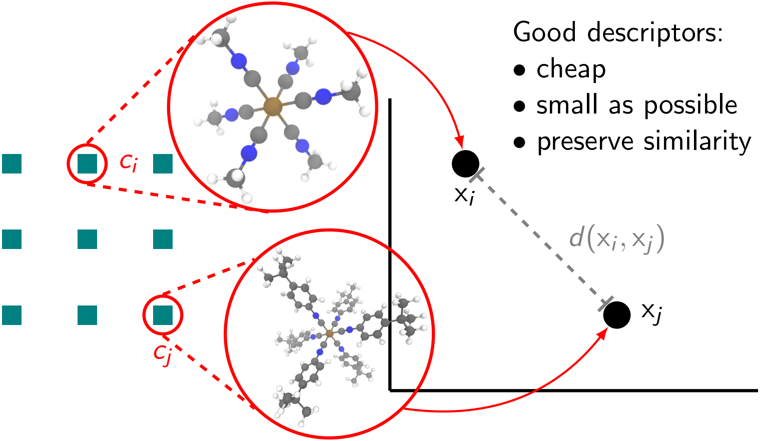


Chemical Space C_f **Descriptor Space** $\mathcal{X} \subset \mathbb{R}^d$

C4: Representing chemical systems



C4: Representing chemical systems



Good descriptors:

- cheap
- small as possible
- preserve similarity

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

C4: Types of representation

complexity

C4: Types of representation

complexity

Fingerprints

– considerable use in drug design

– no information related to molecular topology

– cheap to compute

C4: Types of representation

complexity

Fingerprints

– considerable use in drug design

– no information related to molecular topology

– cheap to compute

Graph-theoretic

– *topological* and *topochemical* representations

– includes all connectivity but no 3D information

– relatively easy to compute

C4: Types of representation

complexity →

Fingerprints

– considerable use in drug design

– no information related to molecular topology

– cheap to compute

Graph-theoretic

– *topological* and *topochemical* representations

– includes all connectivity but no 3D information

– relatively easy to compute

3D structure

– fine-grained structural information in 3D

– mimic input to a quantum chemistry code

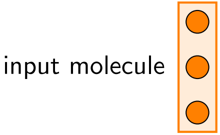
–expensive to compute, rich information

C5: How neural networks work

Simple neural networks can be understood as learned, continuous maps from the input space to a latent space, followed by linear regression

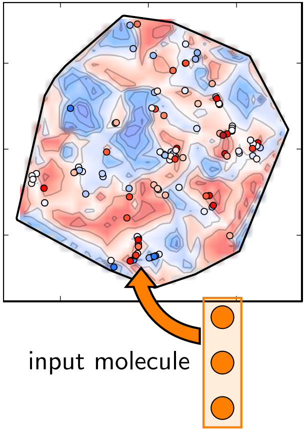
C5: How neural networks work

Simple neural networks can be understood as learned, continuous maps from the input space to a latent space, followed by linear regression

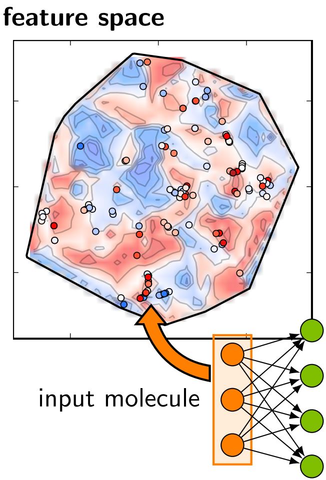


C5: How neural networks work

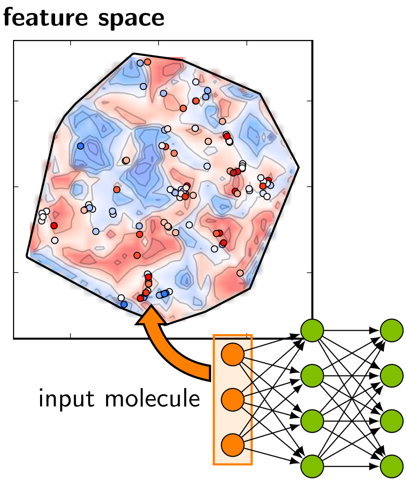
feature space



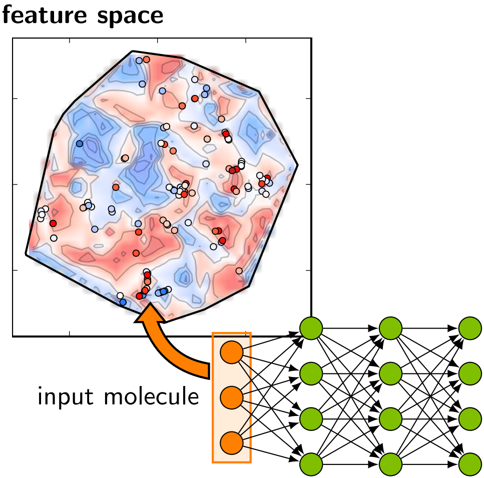
C5: How neural networks work



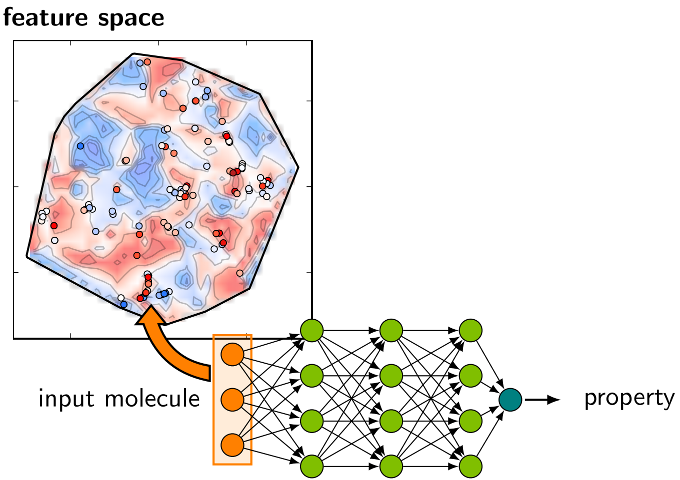
C5: How neural networks work



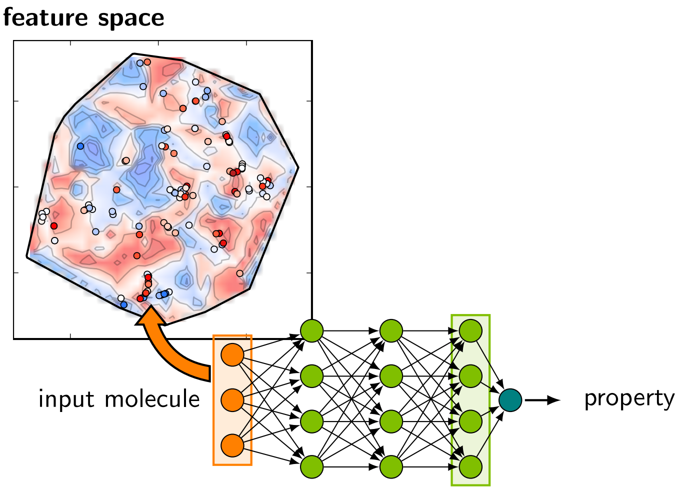
C5: How neural networks work



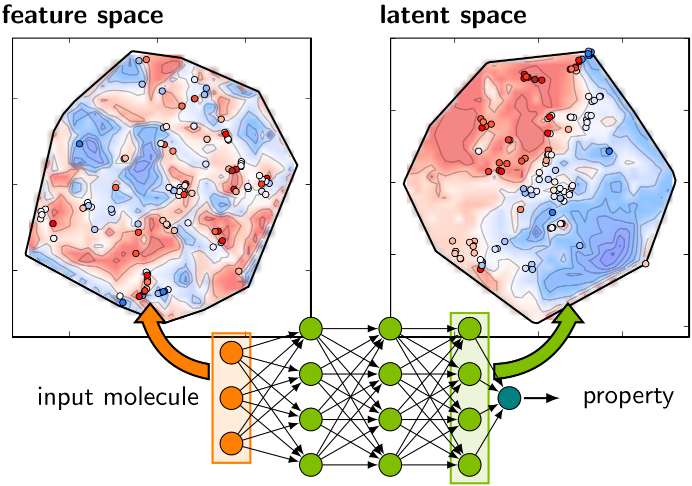
C5: How neural networks work



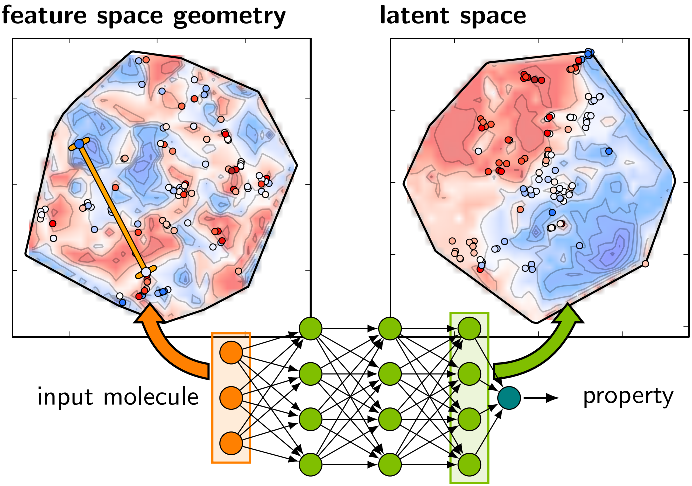
C5: How neural networks work



C5: How neural networks work



C5: How neural networks work



C5: How neural networks work

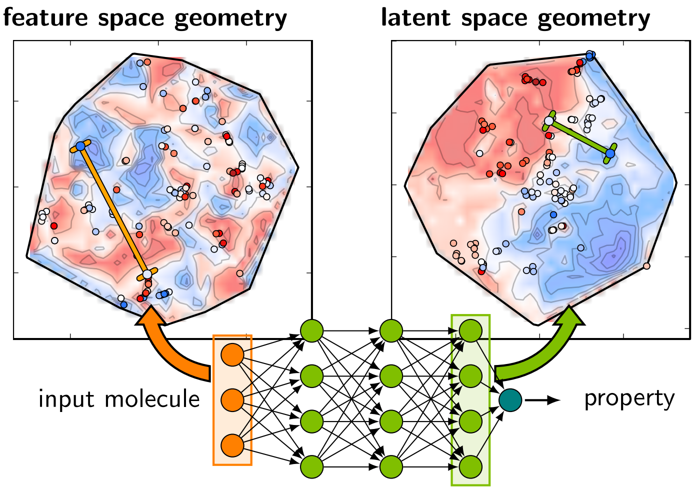


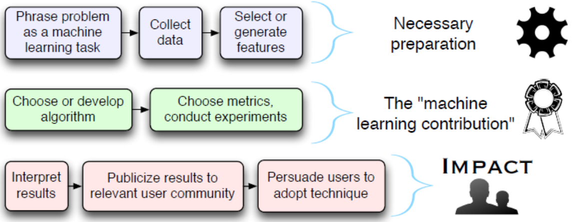
Table of Contents

Final thoughts

It is increasingly important to be literate about ML concepts. Even if/when the hype lessens, ML tools will continue to have a large impact on our science.

Final thoughts

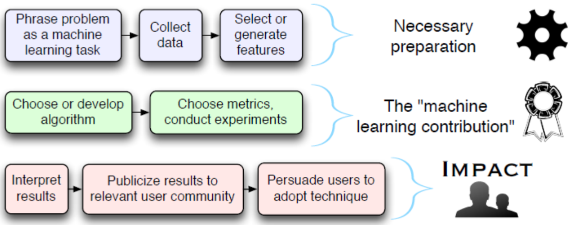
It is increasingly important to be literate about ML concepts. Even if/when the hype lessens, ML tools will continue to have a large impact on our science.



Wagstaff, K., "Machine Learning that Matters", ICML 29, 16(7):529–536, 2012

Final thoughts

It is increasingly important to be literate about ML concepts. Even if/when the hype lessens, ML tools will continue to have a large impact on our science.



Wagstaff, K., "Machine Learning that Matters", ICML 29, 16(7):529–536, 2012

Conversely, there is a growing need for domain experts to engage and derive impact from advances in ML, and you have a lot of value to contribute to interpreting and exploiting the results.