Basics of Chemical Machine Learning

September 17, 2025

Dr. Juan Vicente Alegre Requena

The AI “hype”

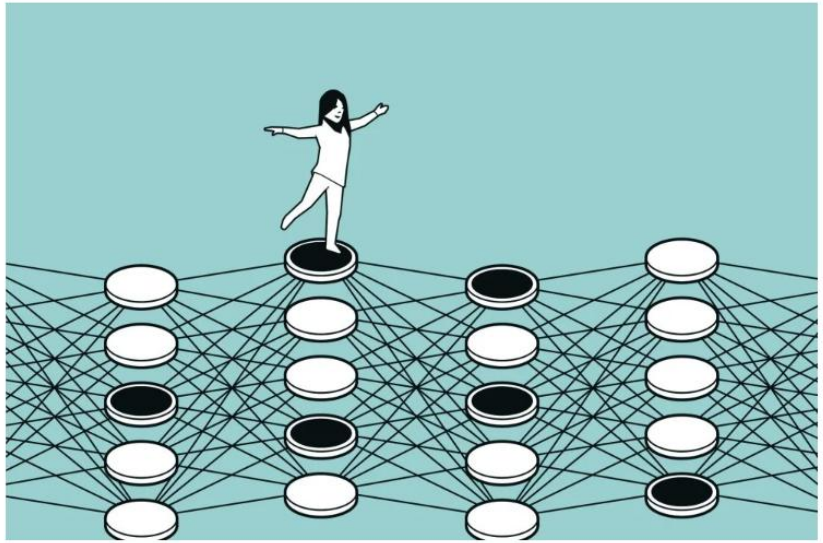
The Nobel Prize in Physics 2024

They used physics to find patterns in information

This year’s laureates used tools from physics to construct methods that helped lay the foundation for today’s powerful machine learning. John Hopfield created a structure that can store and reconstruct information. Geoffrey Hinton invented a method that can independently discover properties in data and which has become important for the large artificial neural networks now in use.

Related articles

Press release
Popular information: They used physics to find patterns in



© Johan Jarnestad/The Royal Swedish Academy of Sciences

The AI “hype”

The Nobel Prize in Chemistry 2024

They cracked the code for proteins’ amazing structures

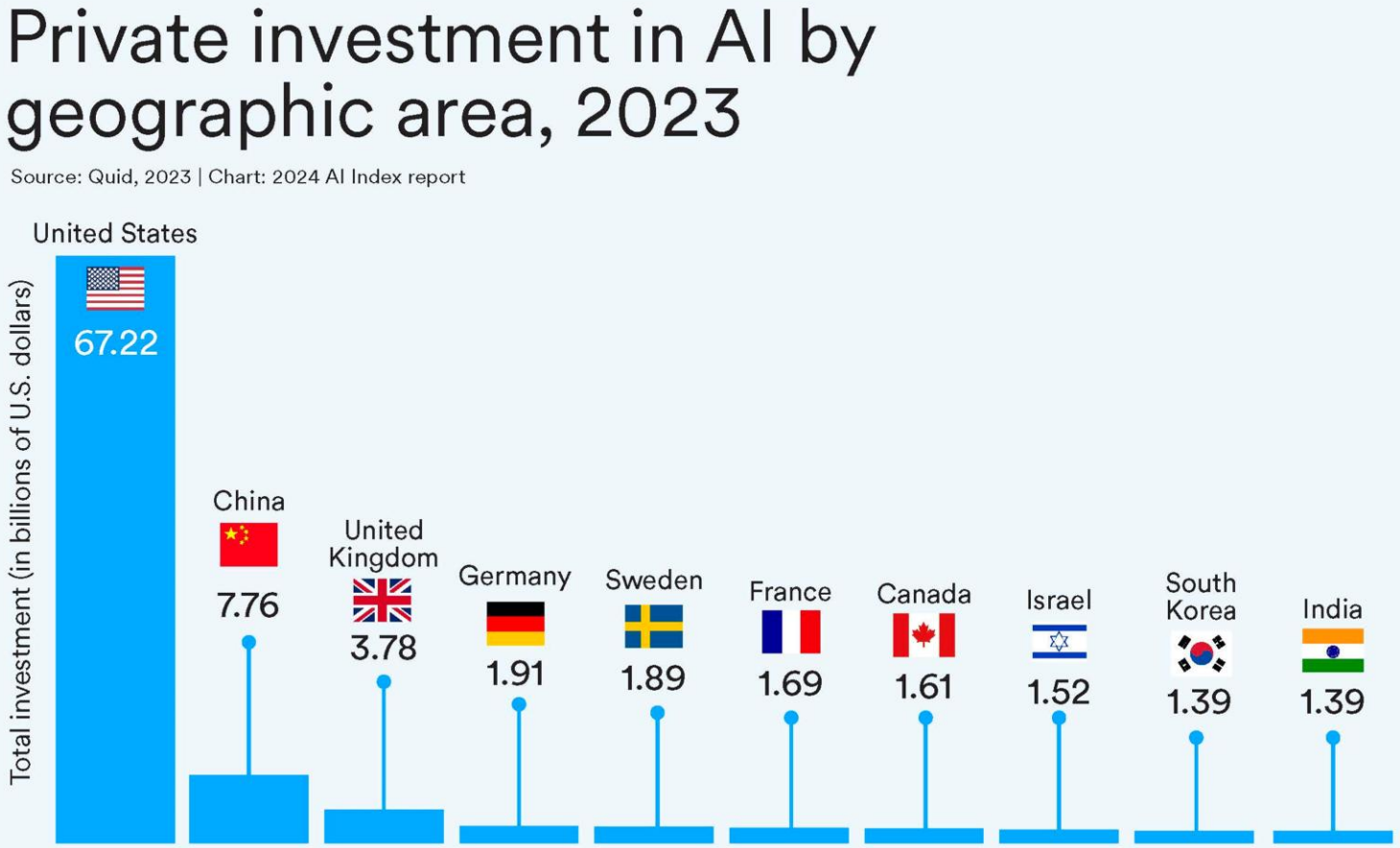
The Nobel Prize in Chemistry 2024 is about proteins, life’s ingenious chemical tools. David Baker has succeeded with the almost impossible feat of building entirely new kinds of proteins. Demis Hassabis and John Jumper have developed an AI model to solve a 50-year-old problem: predicting proteins’ complex structures. These discoveries hold enormous potential.

Related articles

Press release
Popular information: They have revealed proteins’ secrets through

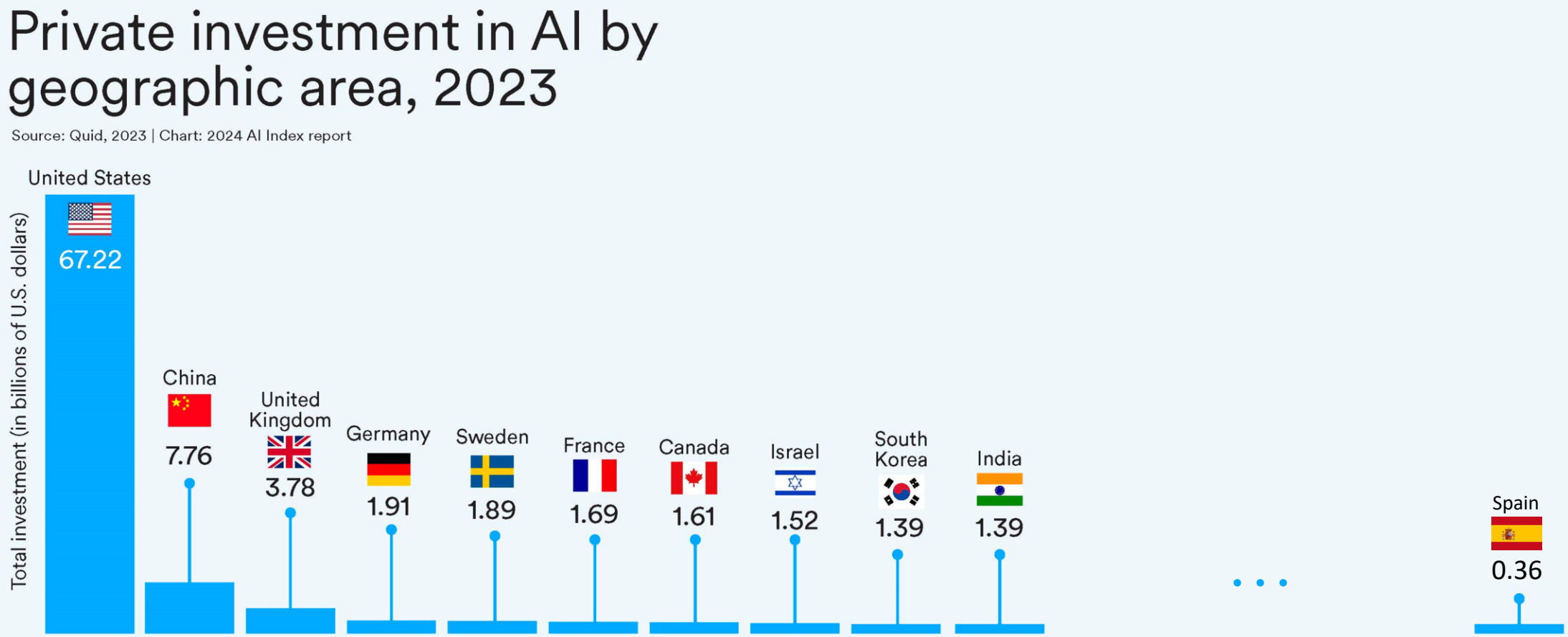
© Johan Jarnestad/The Royal Swedish Academy of Sciences

The AI “hype”



Stanford AI Index Report 2024

The AI “hype”

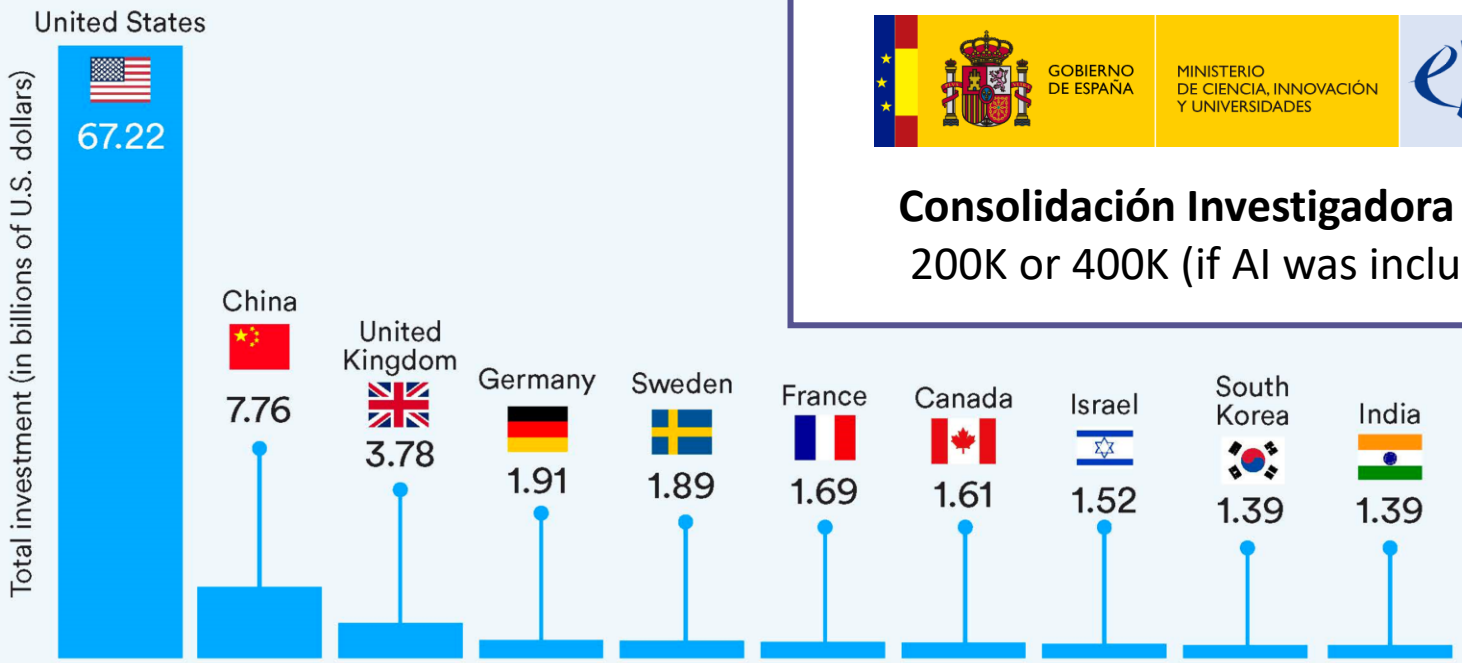


Stanford AI Index Report 2024

The AI “hype”

Private investment in AI by geographic area, 2023

Source: Quid, 2023 | Chart: 2024 AI Index report



Consolidación Investigadora 2023
200K or 400K (if AI was included)

Momentum 2024
2 contracts (AI or digital.)

Stanford AI Index Report 2024

Basics of Chemical ML

How does machine learning work?

Data in

> *"In a village in la Mancha, whose name I do not care to remember, an hidalgo lived not long ago, one of those who keeps a lance on the rack, an old leather shield, skinny nag..."*

> *"To be, or not to be: that is the question: Whether 'tis nobler in the mind to suffer the slings and arrows of outrageous fortune, or to take arms against a sea of troubles, and by opposing end them? To die: to sleep…*

ChatGPT, personal communication, February 11, 2023

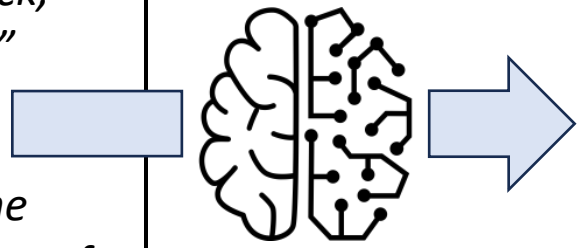
Basics of Chemical ML

How does machine learning work?

Data in

"In a village in la Mancha, whose name I do not care to remember, an hidalgo lived not long ago, one of those who keeps a lance on the rack, an old leather shield, skinny nag..."

"To be, or not to be: that is the question: Whether 'tis nobler in the mind to suffer the slings and arrows of outrageous fortune, or to take arms against a sea of troubles, and by opposing end them? To die: to sleep…



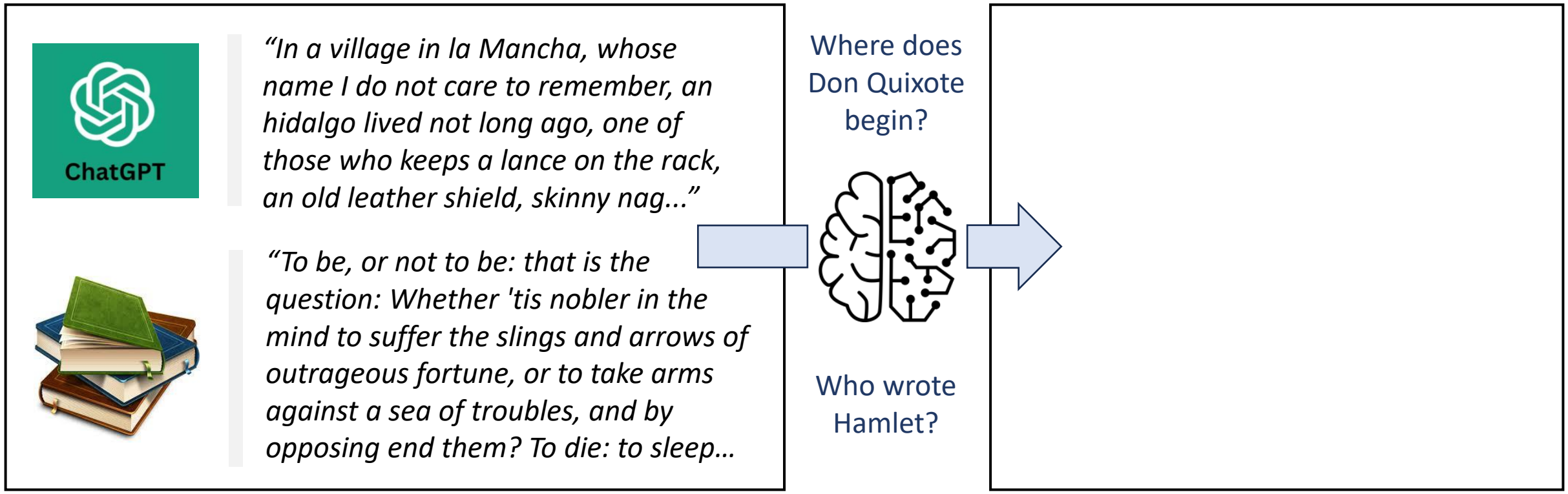
ChatGPT, personal communication, February 11, 2023

Basics of Chemical ML

How does machine learning work?

Data in

Data out



ChatGPT, personal communication, February 11, 2023

Basics of Chemical ML

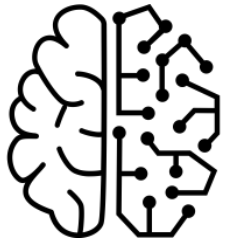
How does machine learning work?

Data in

"In a village in la Mancha, whose name I do not care to remember, an hidalgo lived not long ago, one of those who keeps a lance on the rack, an old leather shield, skinny nag..."

"To be, or not to be: that is the question: Whether 'tis nobler in the mind to suffer the slings and arrows of outrageous fortune, or to take arms against a sea of troubles, and by opposing end them? To die: to sleep…

Where does Don Quixote begin?



Who wrote Hamlet?

Data out

"Don Quixote begins in a place in La Mancha. Don Quixote of La Mancha is a novel written by the Spanish author Miguel de Cervantes Saavedra. Its first part was published with the title 'The Ingenious Gentleman Don Quixote'…"

ChatGPT, personal communication, February 11, 2023

Basics of Chemical ML

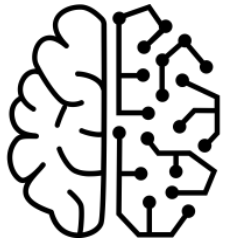
How does machine learning work?

Data in

"In a village in la Mancha, whose name I do not care to remember, an hidalgo lived not long ago, one of those who keeps a lance on the rack, an old leather shield, skinny nag..."

"To be, or not to be: that is the question: Whether 'tis nobler in the mind to suffer the slings and arrows of outrageous fortune, or to take arms against a sea of troubles, and by opposing end them? To die: to sleep…

Where does Don Quixote begin?



Who wrote Hamlet?

Data out

"Don Quixote begins in a place in La Mancha. Don Quixote of La Mancha is a novel written by the Spanish author Miguel de Cervantes Saavedra. Its first part was published with the title 'The Ingenious Gentleman Don Quixote'…"

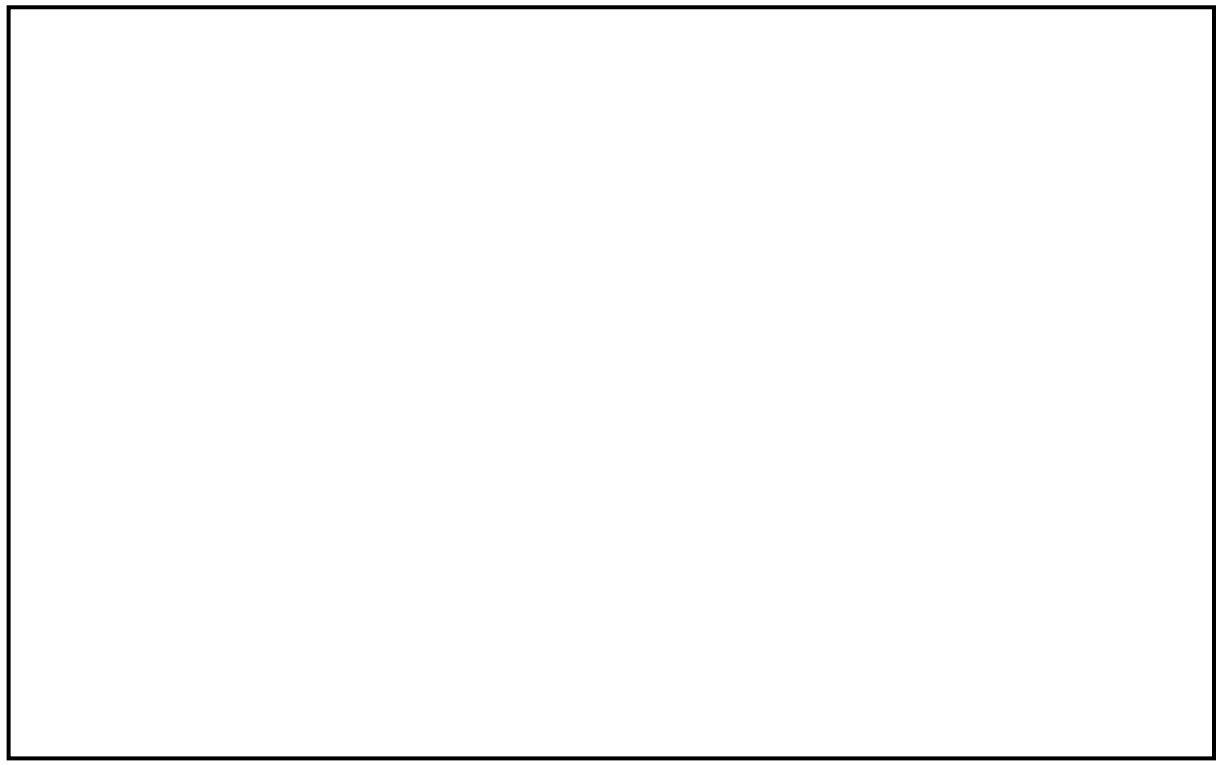
"William Shakespeare wrote Hamlet. The tragedy of Hamlet, Prince of Denmark (original title in English: The Tragedy of Hamlet, Prince of Denmark), is a play by the…"

ChatGPT, personal communication, February 11, 2023

Basics of Chemical ML

How does machine learning work in chemistry?

Data in

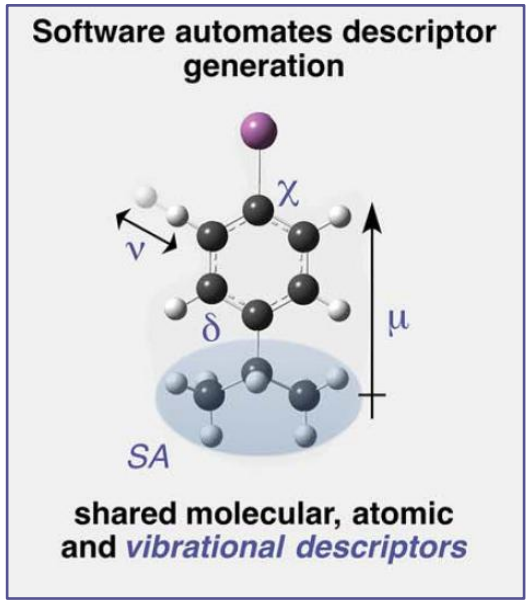


Science **2018**, *360*, 186

Basics of Chemical ML

How does machine learning work in chemistry?

Data in



Molecular properties

- *Solubility*
- *m.p.*

Atomic properties

- *Charge*
- *Sterics*

Material properties

- *Component ratio*

Process properties

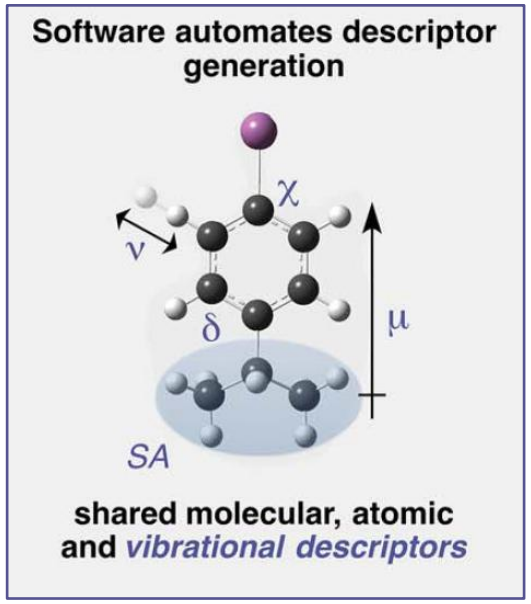
- *Temperature*
- *Pressure*

Science **2018**, *360*, 186

Basics of Chemical ML

How does machine learning work in chemistry?

Data in



Molecular properties

- *Solubility*
- *m.p.*

Atomic properties

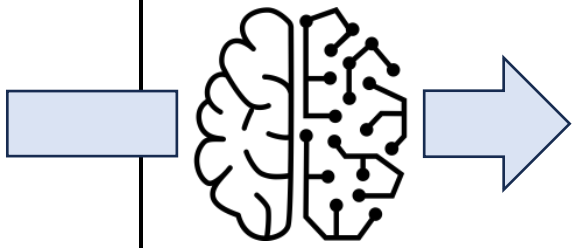
- *Charge*
- *Sterics*

Material properties

- *Component ratio*

Process properties

- *Temperature*
- *Pressure*

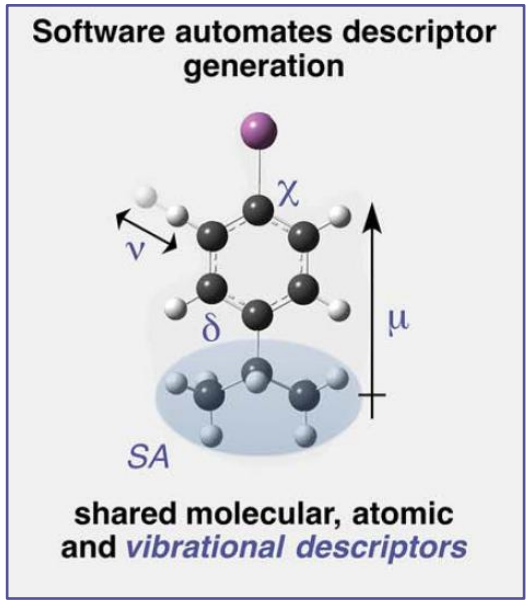


Science **2018**, *360*, 186

Basics of Chemical ML

How does machine learning work in chemistry?

Data in



Molecular properties
- *Solubility*
- *m.p.*

Atomic properties
- *Charge*
- *Sterics*

Material properties
- *Component ratio*

Process properties
- *Temperature*
- *Pressure*

Design an effective catalyst

Design an optimal solar cell

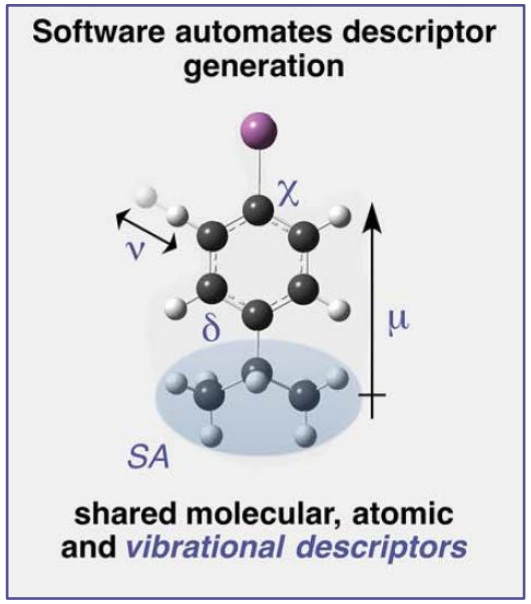
Data out

Science **2018**, *360*, 186

Basics of Chemical ML

How does machine learning work in chemistry?

Data in



Molecular properties
- *Solubility*
- *m.p.*

Atomic properties
- *Charge*
- *Sterics*

Material properties
- *Component ratio*

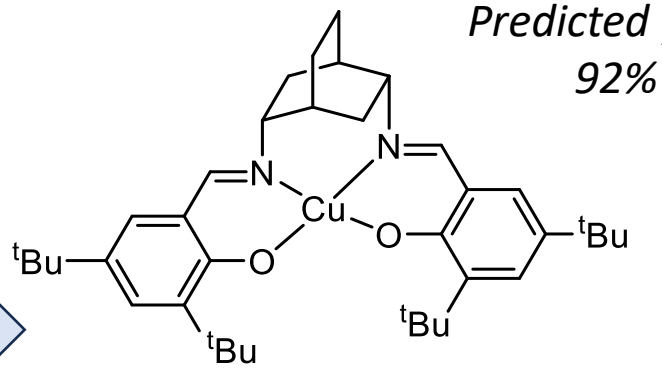
Process properties
- *Temperature*
- *Pressure*

Design an effective catalyst

Design an optimal solar cell

Data out

Predicted yield: 92%

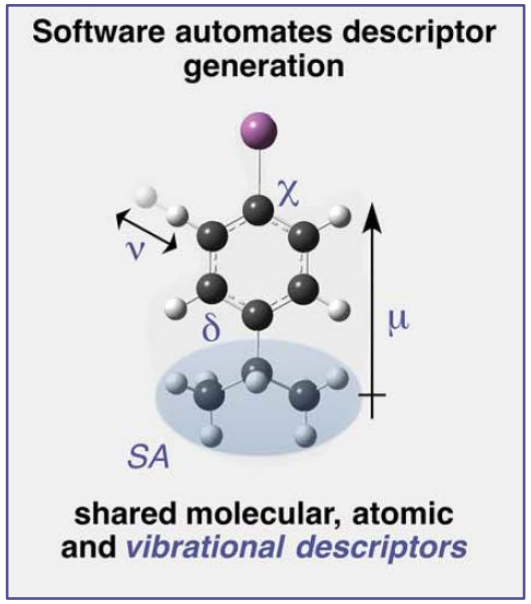


Science **2018**, *360*, 186

Basics of Chemical ML

How does machine learning work in chemistry?

Data in



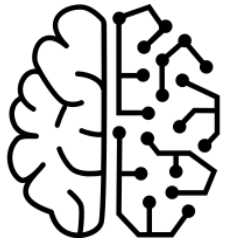
Molecular properties
- *Solubility*
- *m.p.*

Atomic properties
- *Charge*
- *Sterics*

Material properties
- *Component ratio*

Process properties
- *Temperature*
- *Pressure*

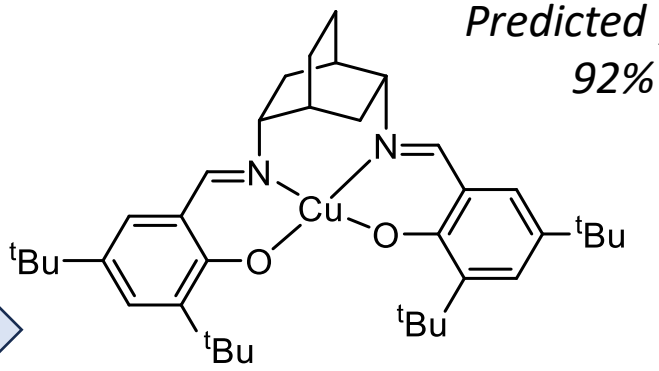
Design an effective catalyst



Design an optimal solar cell

Data out

Predicted yield: 92%



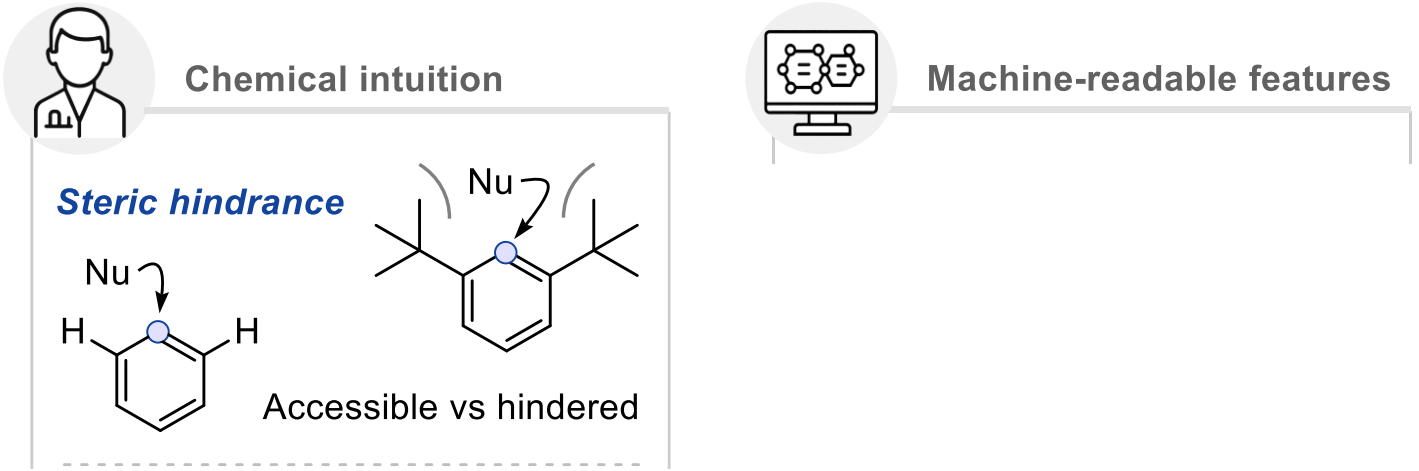
20% graphene + 40% Si + 40% GaAs

Predicted solar-cell efficiency: 21%

Science **2018**, *360*, 186

Basics of Chemical ML

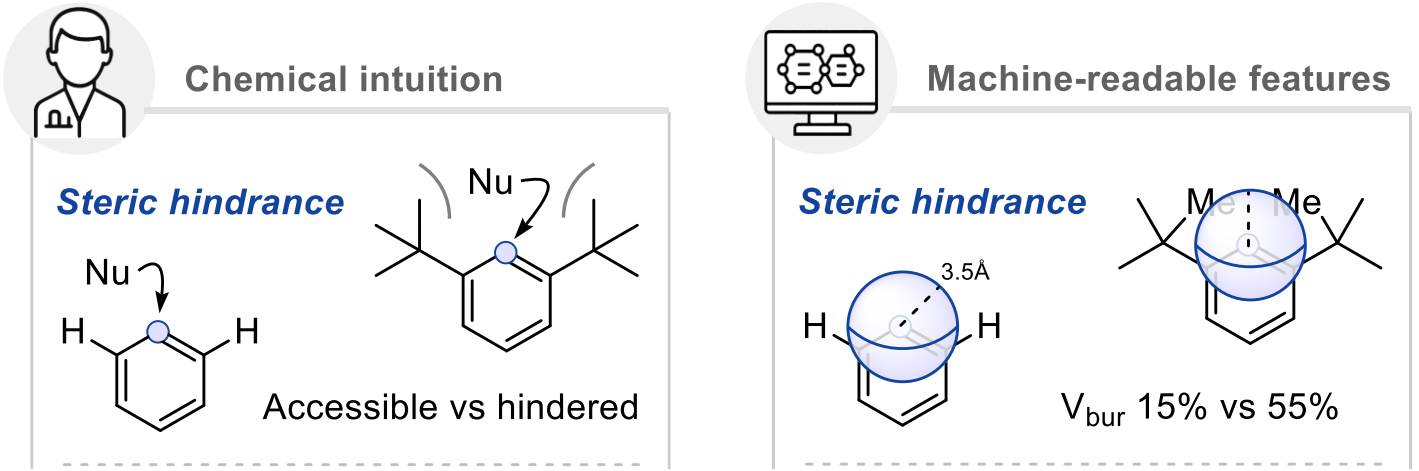
First step: digitalization of molecules



Submitted

Basics of Chemical ML

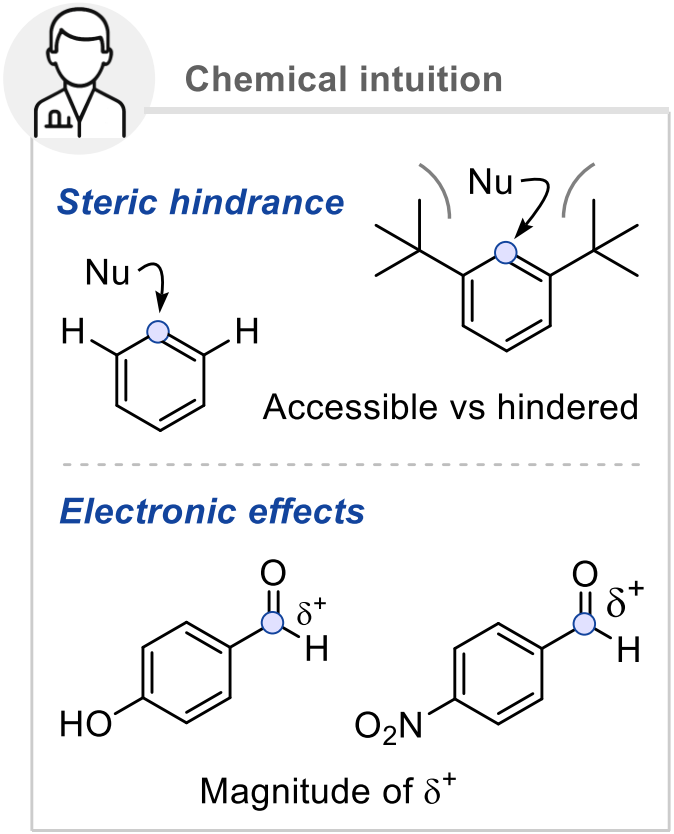
First step: digitalization of molecules

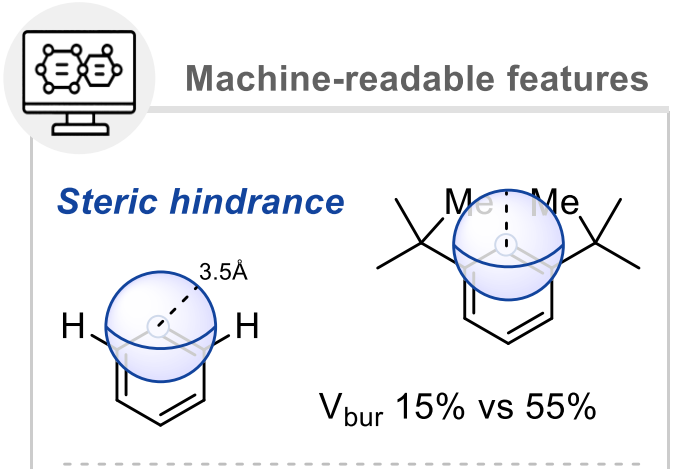


Submitted

Basics of Chemical ML

First step: digitalization of molecules

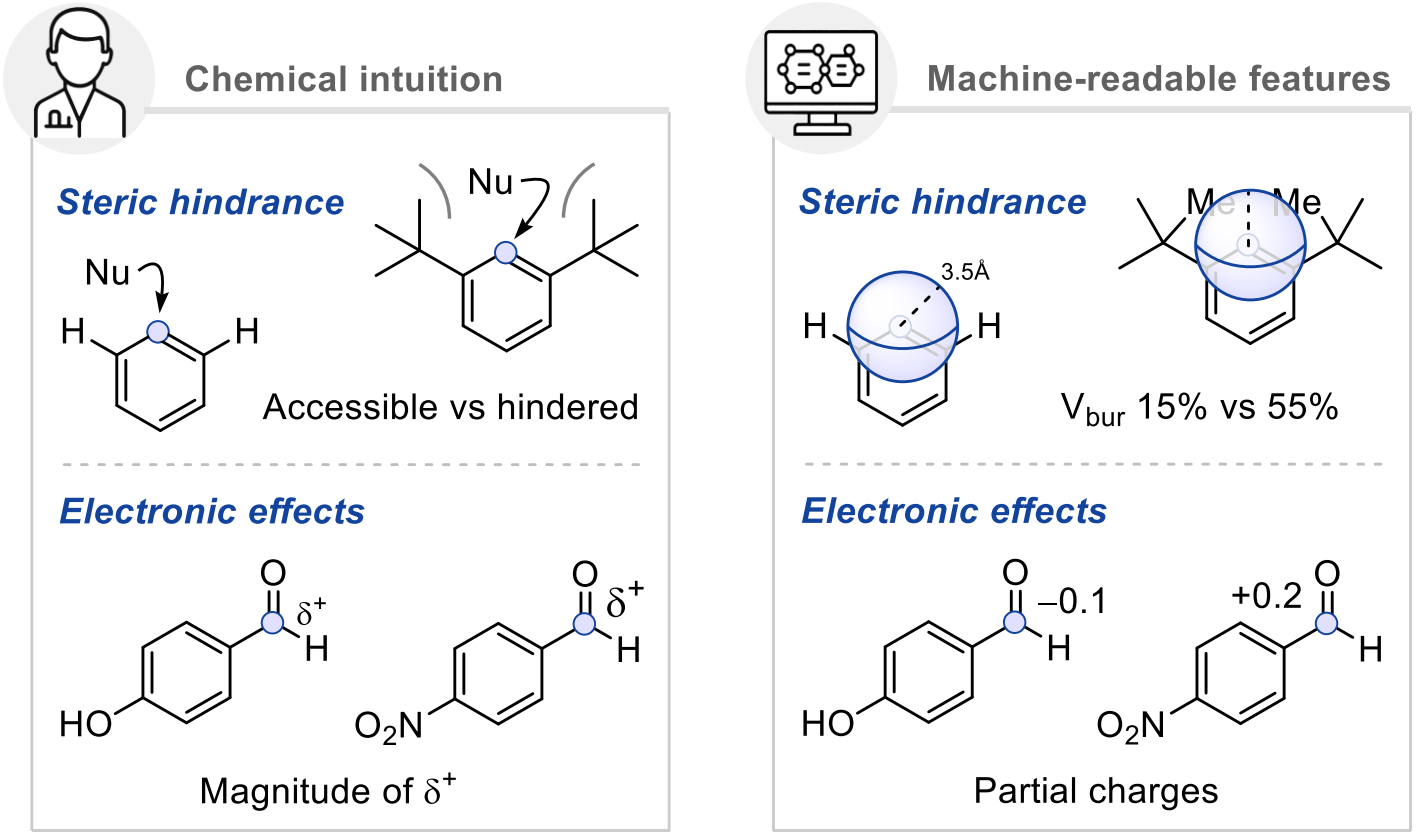




Submitted

Basics of Chemical ML

First step: digitalization of molecules



Submitted

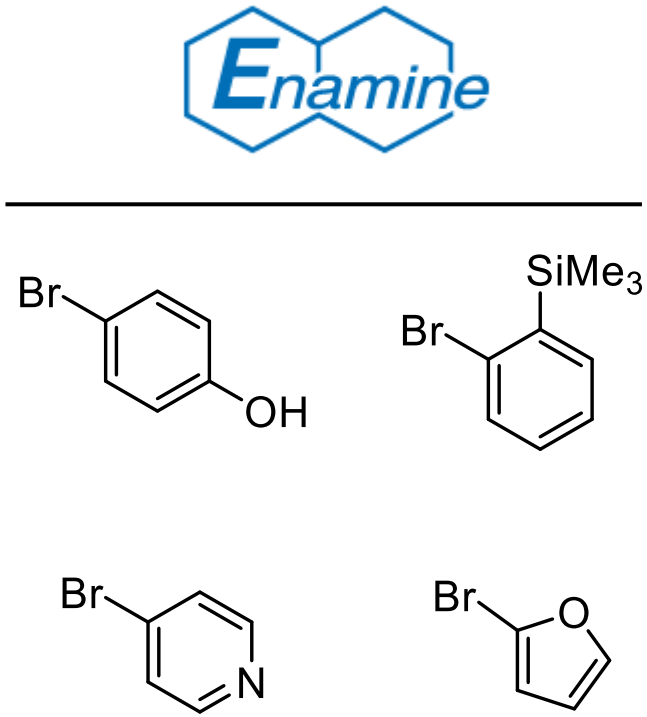
Basics of Chemical ML

First step: digitalization of molecules

Guidelines
Generate or collect relevant atomic and molecular descriptors to represent your system. A few options: • Databases: KRAKEN,[65] ioChem-BD,[66] OSCAR,[67] and OMol25.[68] • Topological/MM descriptors and fingerprints: RDKit,[17] CDK,[69] and PaDELPy.[70] • QM descriptors: AQME,[71] ObeLiX,[72] Autoqchem,[73] and molli.[74] • Steric descriptors: MORFEUS,[75] DBSTEP,[76] and SambVca.[77] • ML-predicted properties: Chemprop[78] and MolPROP.[79]

Basics of Chemical ML

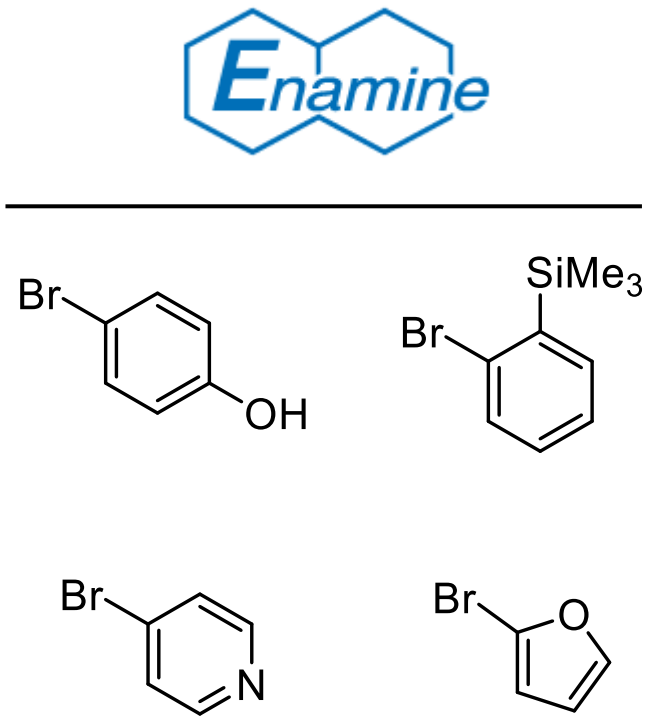
Unsupervised learning: Chemical space and clustering of molecules for sampling



25000 Ar–Br (dataset Pd cross-couplings)

Basics of Chemical ML

Unsupervised learning: Chemical space and clustering of molecules for sampling



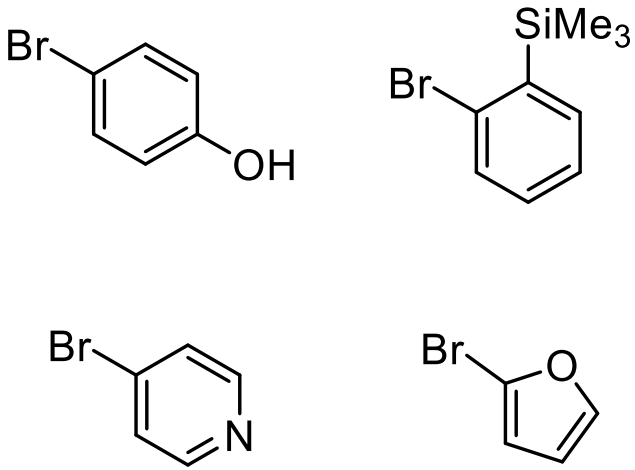
25000 Ar–Br (dataset Pd cross-couplings)

Descriptor generation

Basics of Chemical ML

Unsupervised learning: Chemical space and clustering of molecules for sampling

Enamine



25000 Ar–Br (dataset Pd cross-couplings)

Descriptor generation

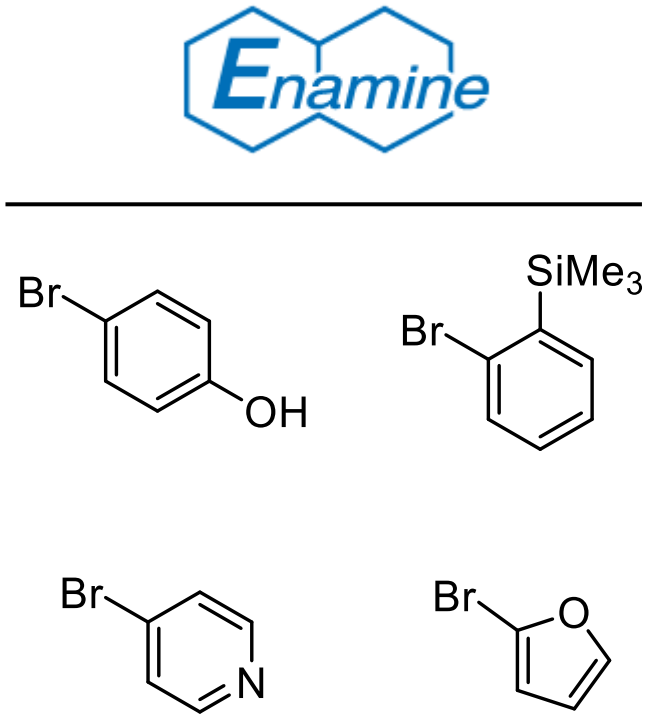
AQME
Automated QM Environments

MORFEUS
molecular features for machine learning

Version 1.8 coming October!
Better descriptor collection, x5-10 faster

Basics of Chemical ML

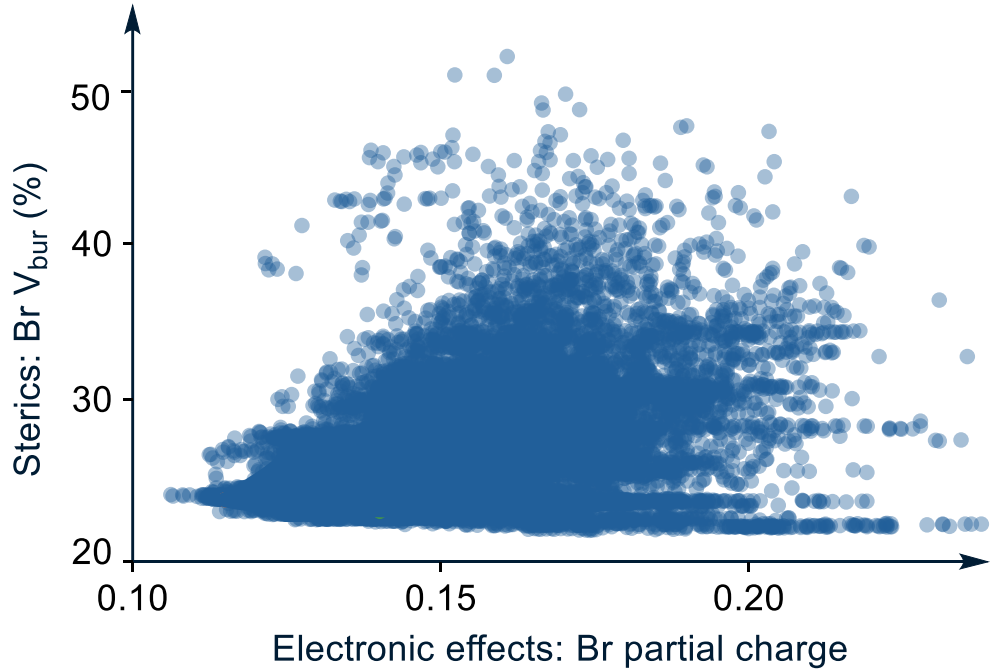
Unsupervised learning: Chemical space and clustering of molecules for sampling



25000 Ar–Br (dataset Pd cross-couplings)

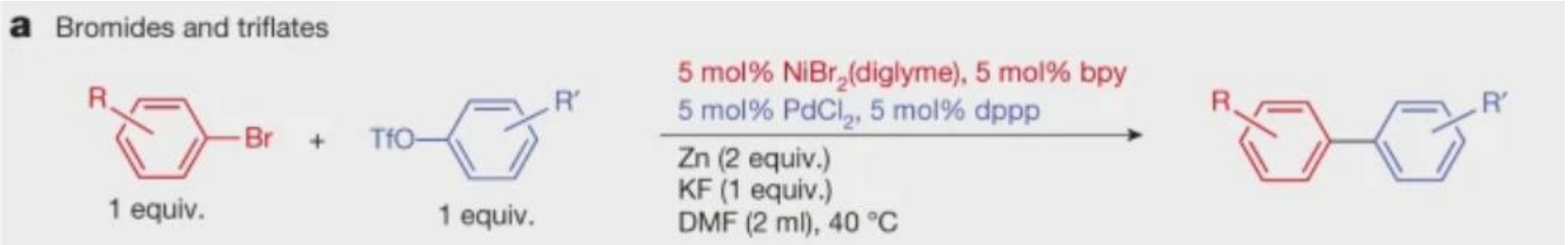
Descriptor generation

AQME
Automated QM Environments

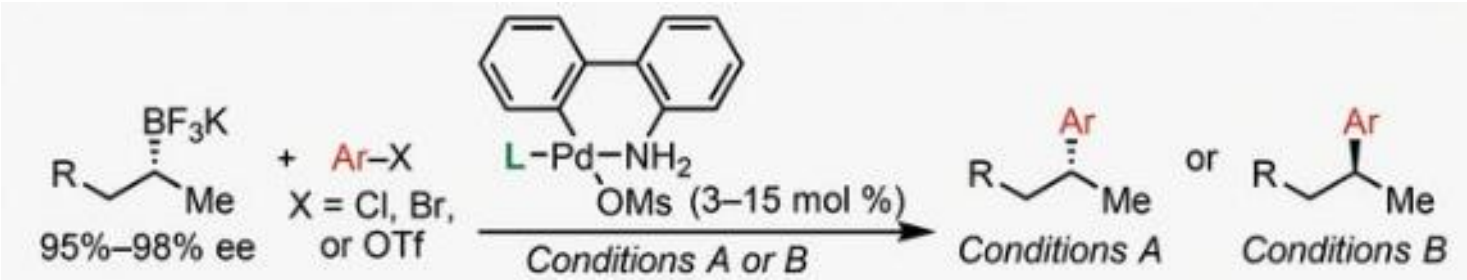


Basics of Chemical ML

Unsupervised learning: Chemical space and clustering of molecules for sampling



19 Ar-Br substrates *Nature* **2015**, *524*, 454–457



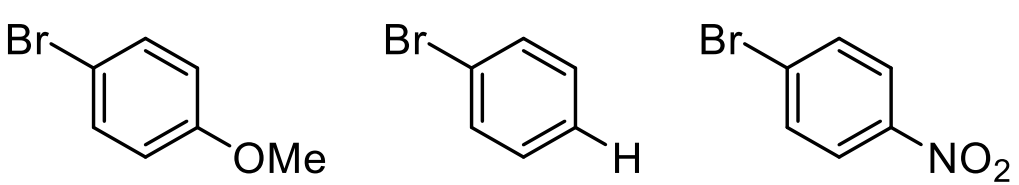
12 Ar-Br substrates *Science* **2018**, *362*, 670–674

Basics of Chemical ML

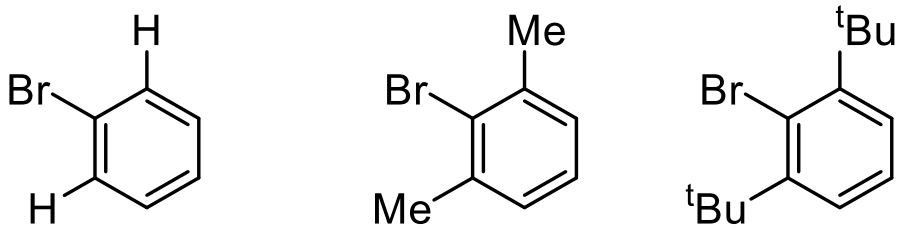
Unsupervised learning: Chemical space and clustering of molecules for sampling

- One-variable-at-a-time strategy
- Usually same substrates used

Electronic effects

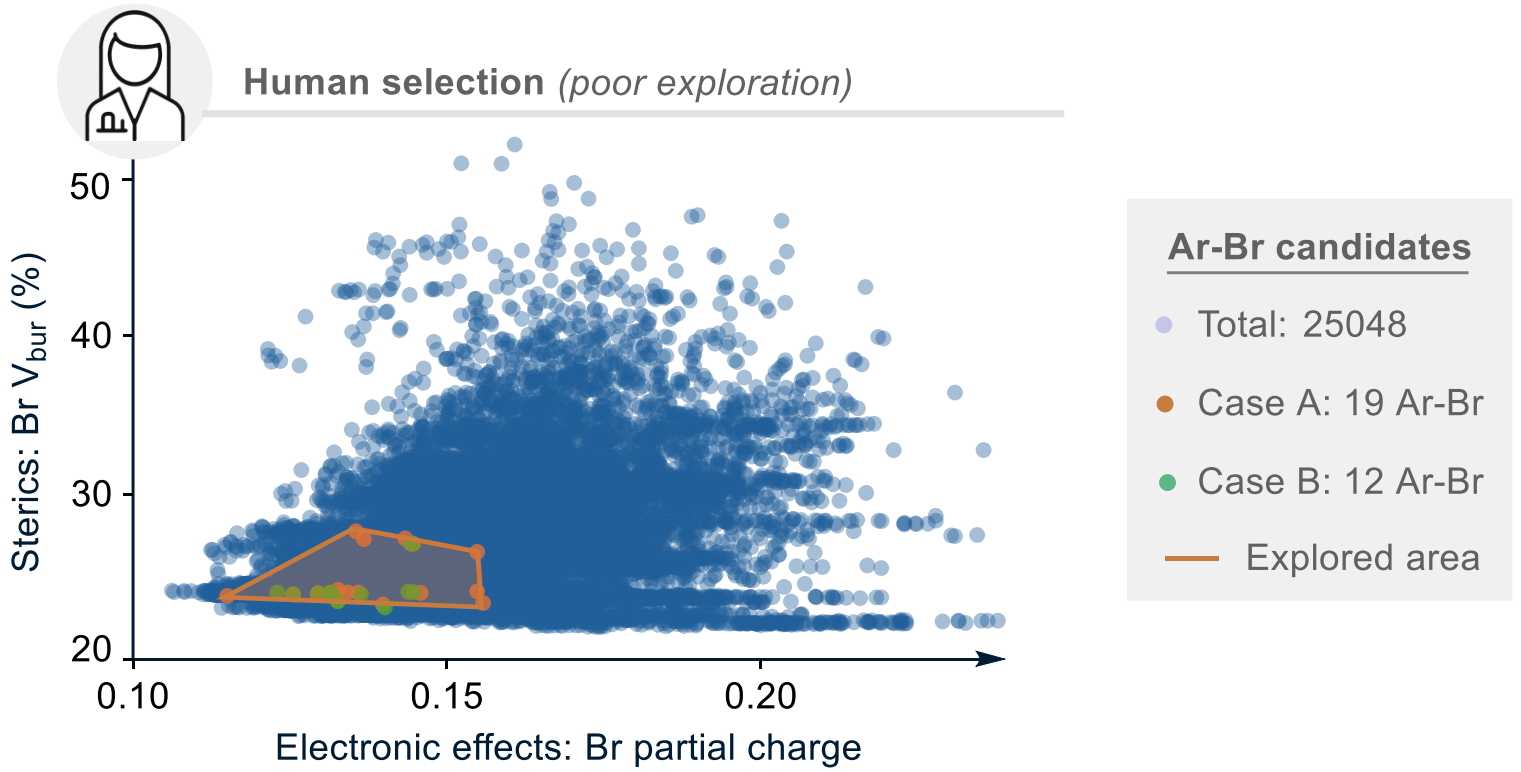


Steric effects



Basics of Chemical ML

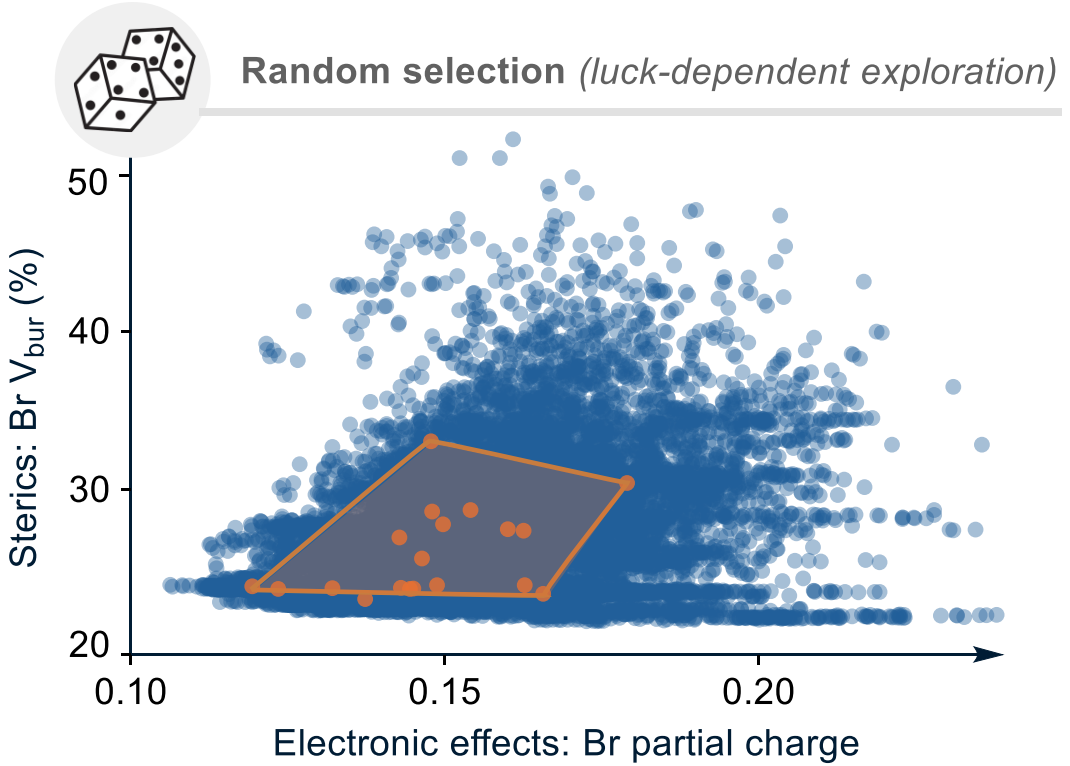
Unsupervised learning: Chemical space and clustering of molecules for sampling

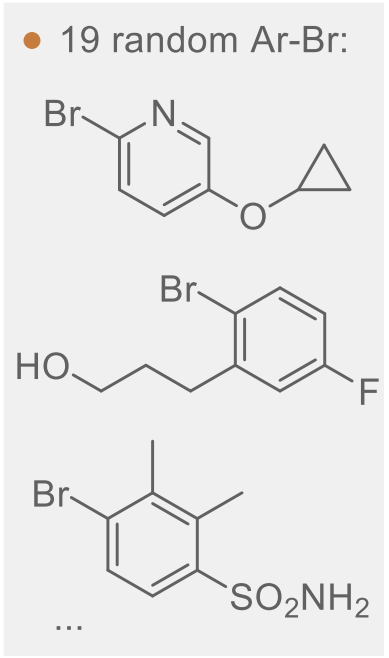


- Very poor exploration, low generality assessment

Basics of Chemical ML

Unsupervised learning: Chemical space and clustering of molecules for sampling

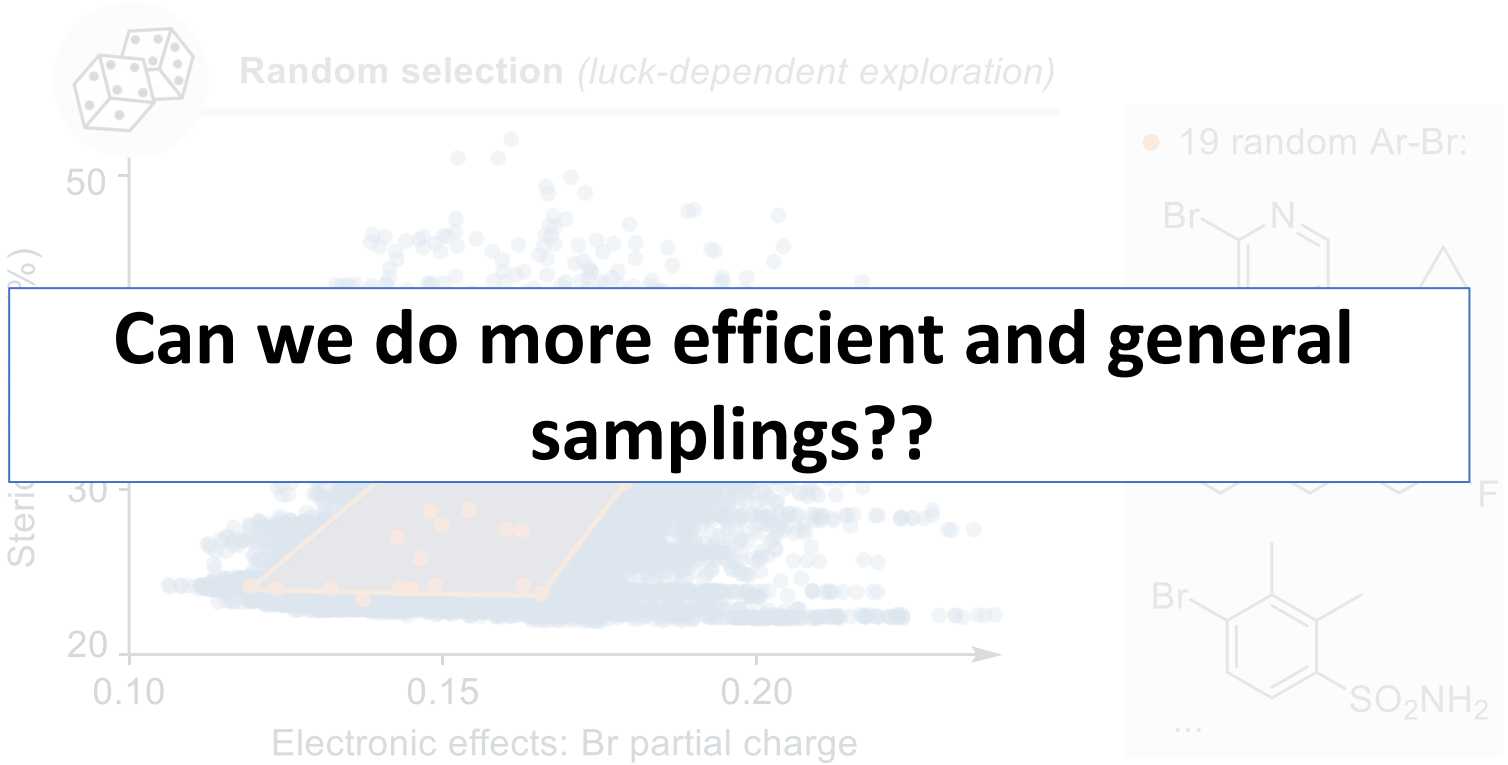




- Even a random selection does better!

Basics of Chemical ML

Unsupervised learning: Chemical space and clustering of molecules for sampling



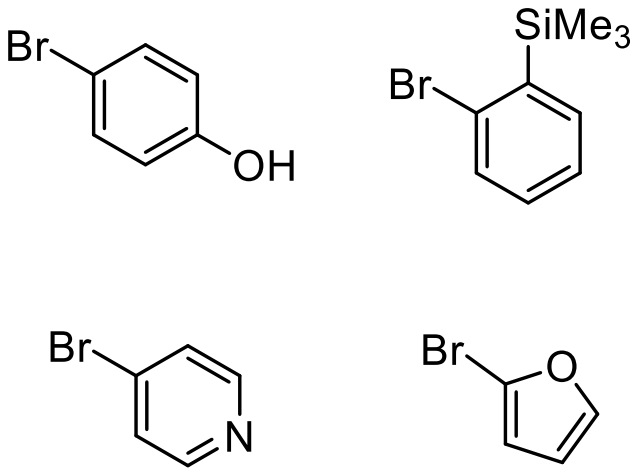
Can we do more efficient and general samplings??

- Even a random selection does better!

Basics of Chemical ML

Unsupervised learning: Chemical space and clustering of molecules for sampling

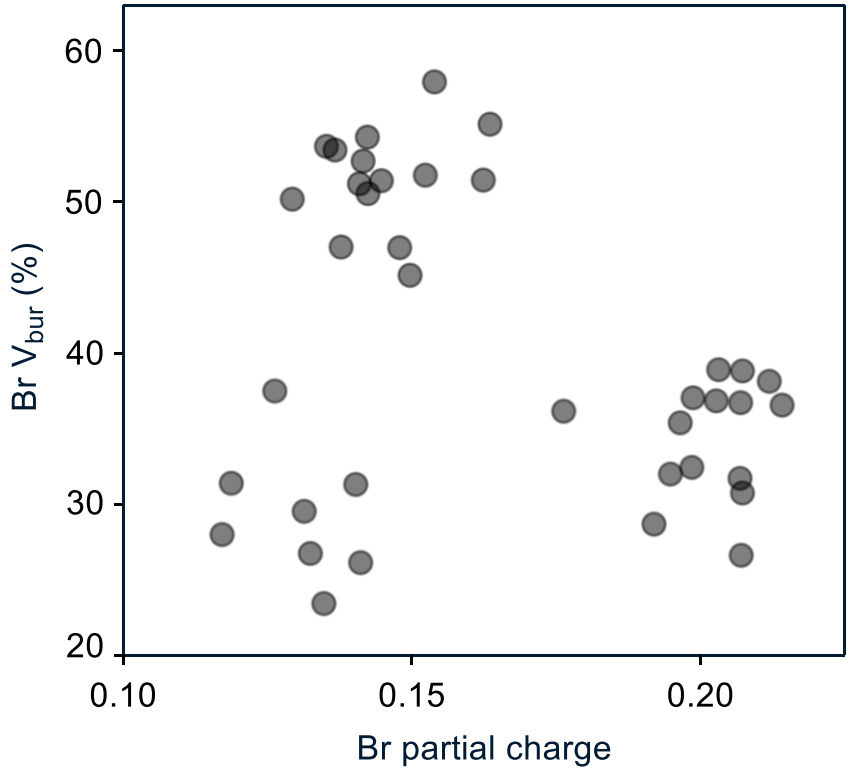
Enamine



38 Ar–Br

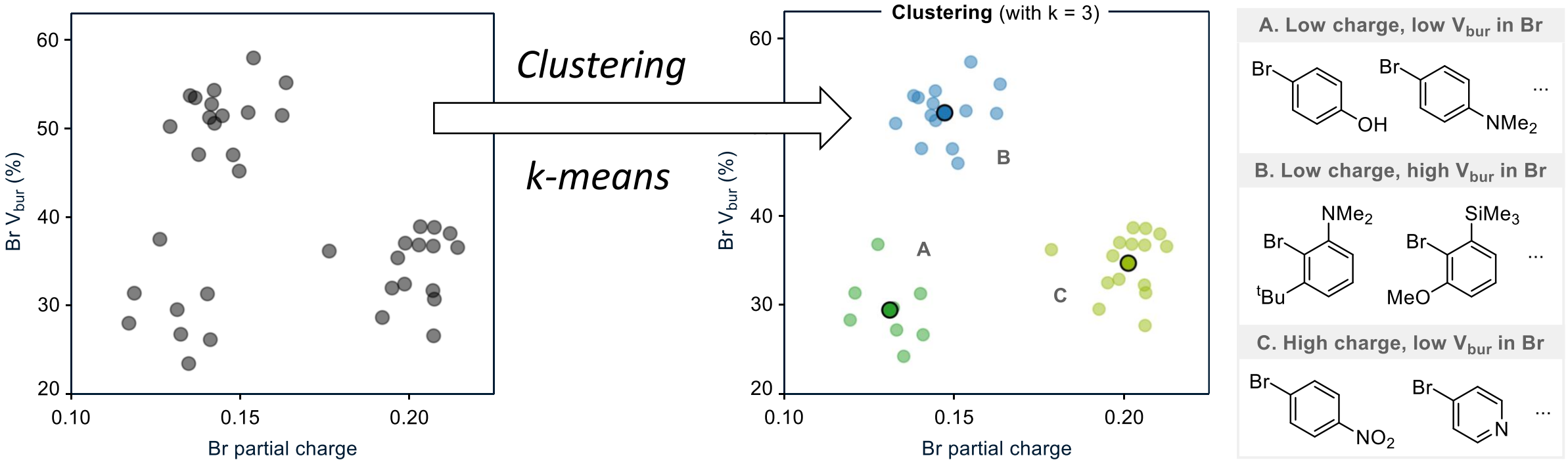
Descriptor generation

AQME
Automated QM Environments



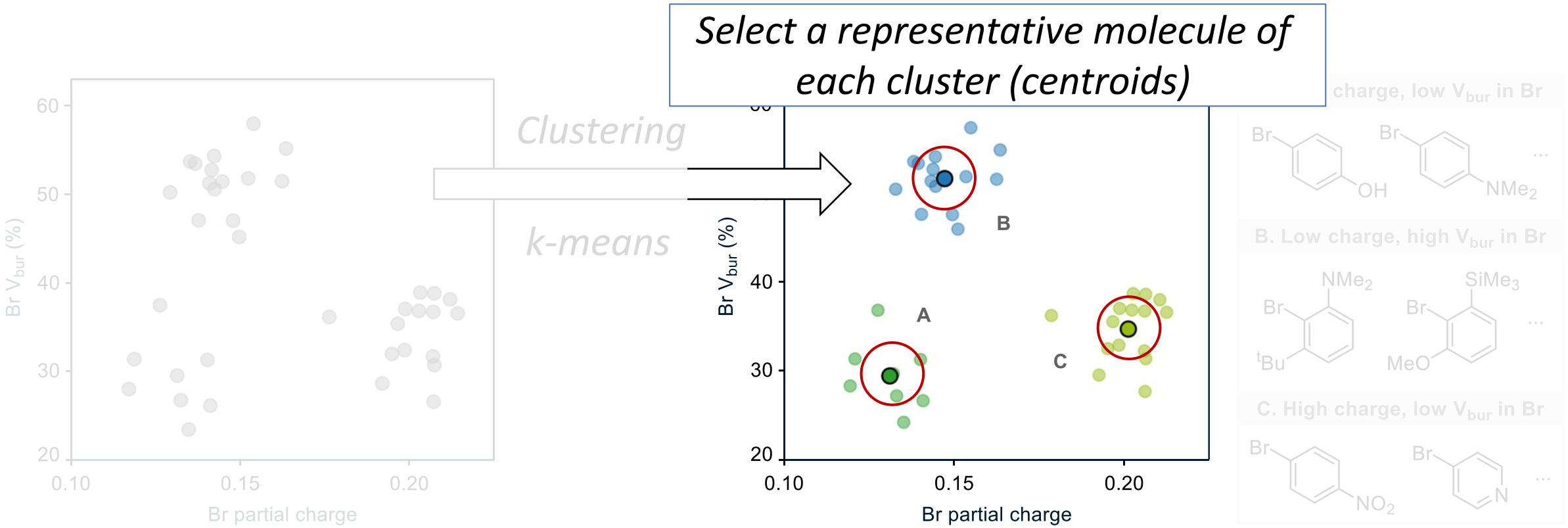
Basics of Chemical ML

Unsupervised learning: Chemical space and clustering of molecules for sampling



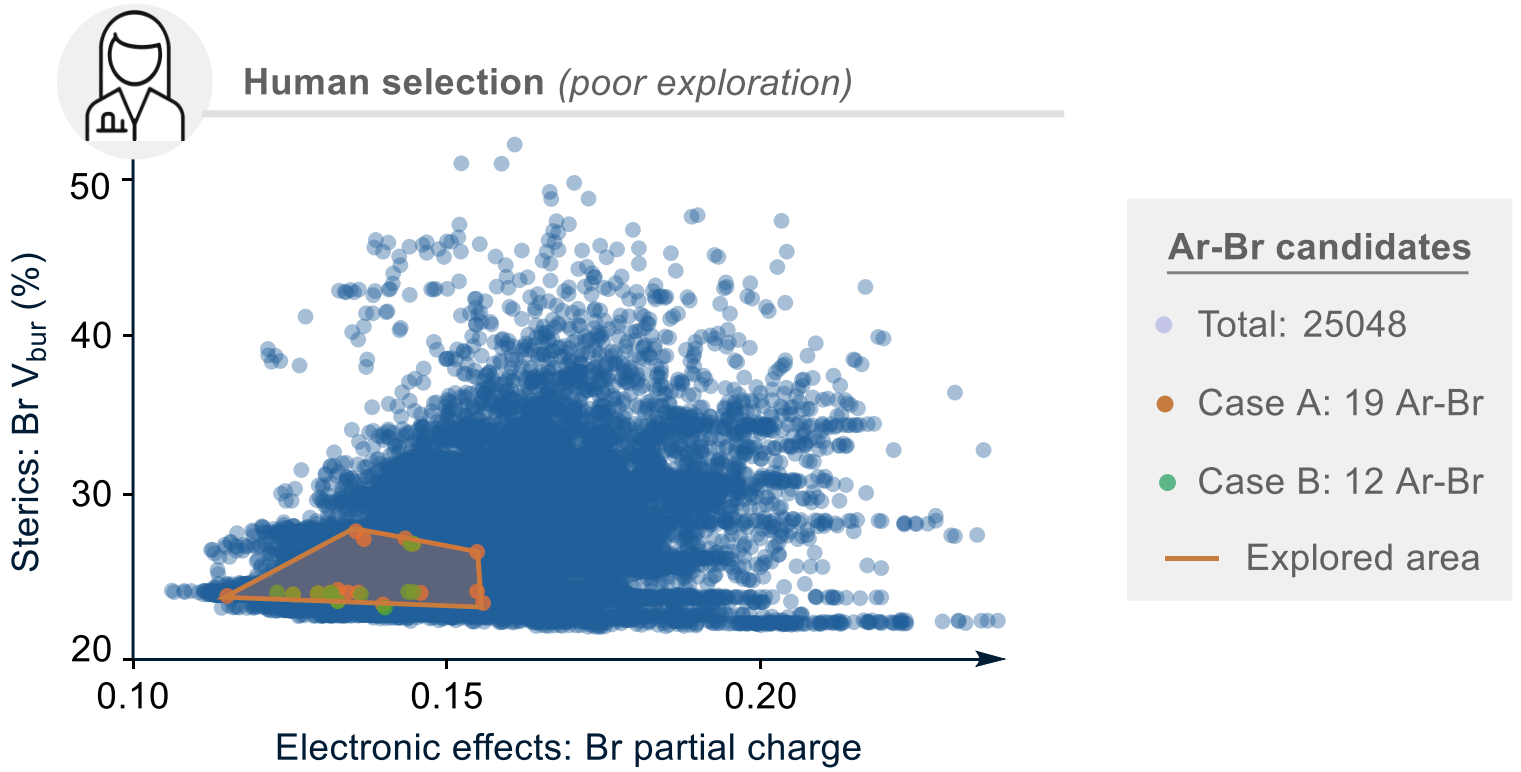
Basics of Chemical ML

Unsupervised learning: Chemical space and clustering of molecules for sampling



Basics of Chemical ML

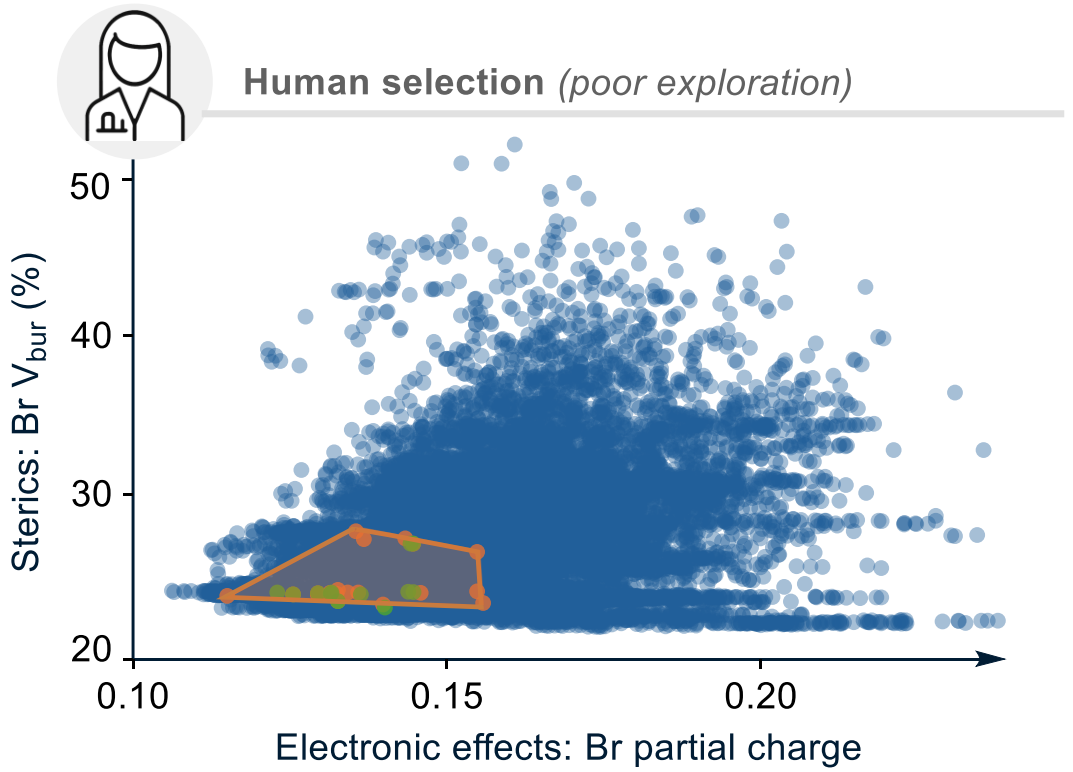
Unsupervised learning: Chemical space and clustering of molecules for sampling

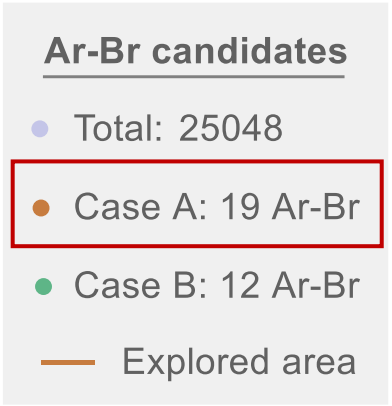


- Very poor exploration, low generality assessment

Basics of Chemical ML

Unsupervised learning: Chemical space and clustering of molecules for sampling



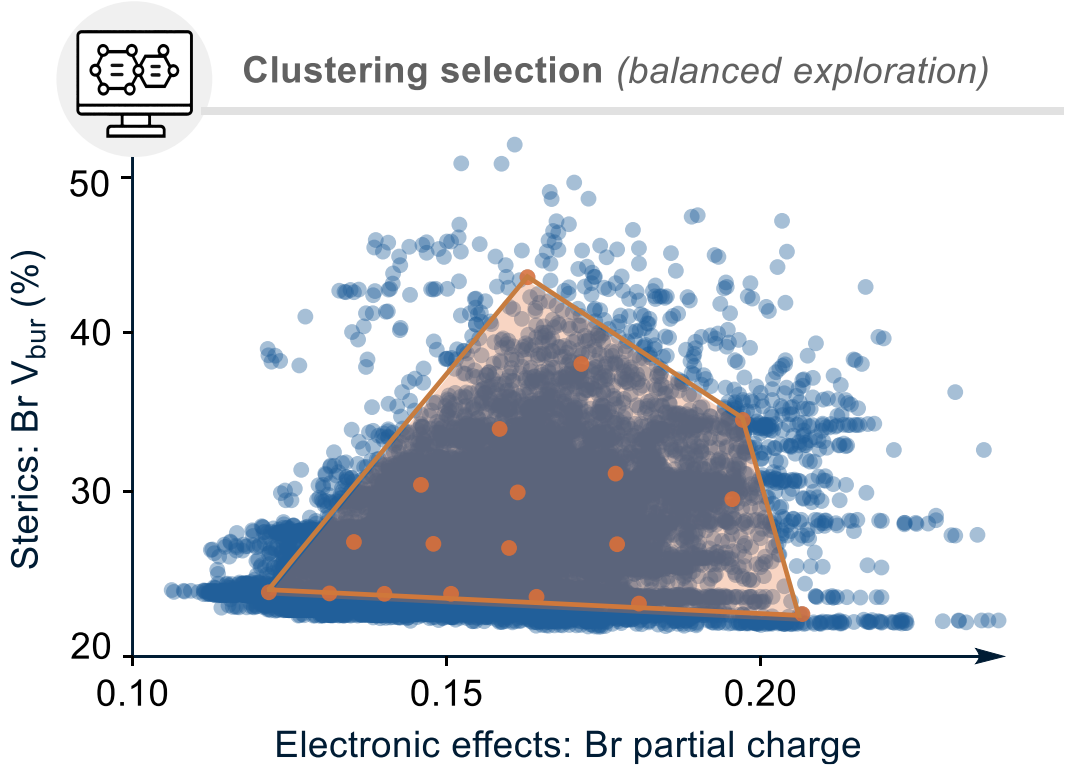


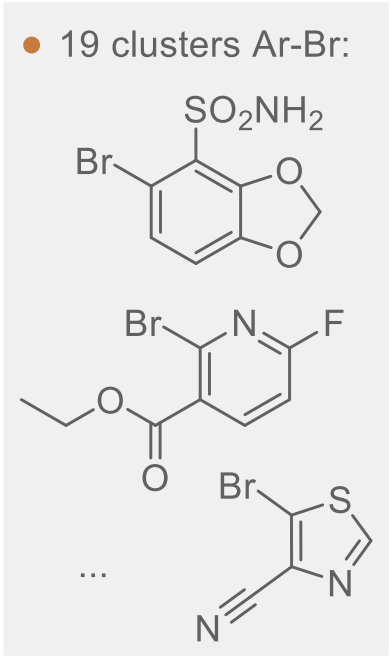
19 clusters

- Very poor exploration, low generality assessment

Basics of Chemical ML

Unsupervised learning: Chemical space and clustering of molecules for sampling





19 clusters

- Good exploration, high generality assessment

Unsupervised learning: Chemical space and clustering of molecules for sampling

Key points:

- Clustering is useful for chemical space exploration
- Quite easy to run (generate descriptors, run clustering algorithm – ChatGPT?)
- Examples: substrate scopes, catalyst screening, starting batch of active learning, etc.
- **CAREFUL!** You need to define your chemical space with meaningful descriptors

Basics of Chemical ML

Unsupervised learning: Chemical space and clustering of molecules for sampling

11:00 - 12:45 *Tutorial*

Automation of QM and clustering

Prof: Garcia

Assist: Stuyver

Dr. Susana Garcia

Supervised learning: Data-driven discovery

Unsupervised learning (clustering, dimensionality reduction)

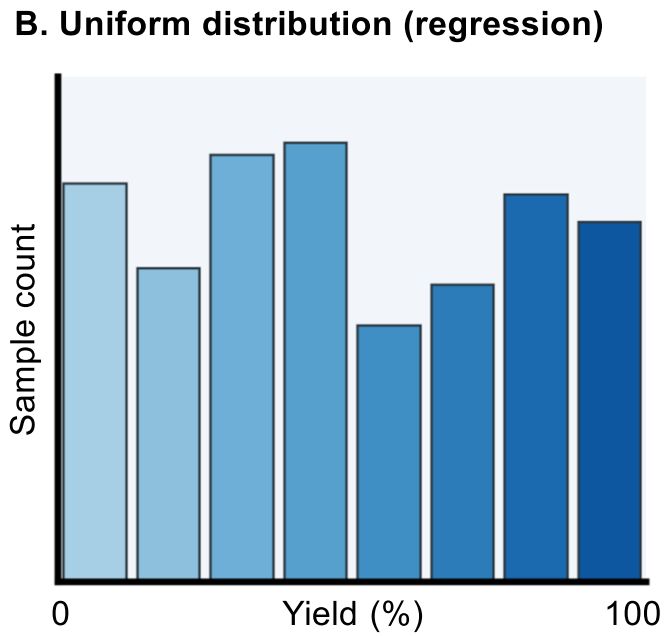
Name	SMILES	Descriptors (X_1, ..., X_n)
Mol_1	CO	X_{11}, ..., X_{1n}
...	...	...
Mol_n	CCO	X_{n1}, ..., X_{nn}

Supervised learning: Data-driven discovery

Supervised learning (regression and classification)

Name	SMILES	Descriptors (X_1, ..., X_n)	Target value (y)
Mol_1	CO	X_{11}, ..., X_{1n}	y_1
...	...	...	...
Mol_n	CCO	X_{n1}, ..., X_{nn}	y_n

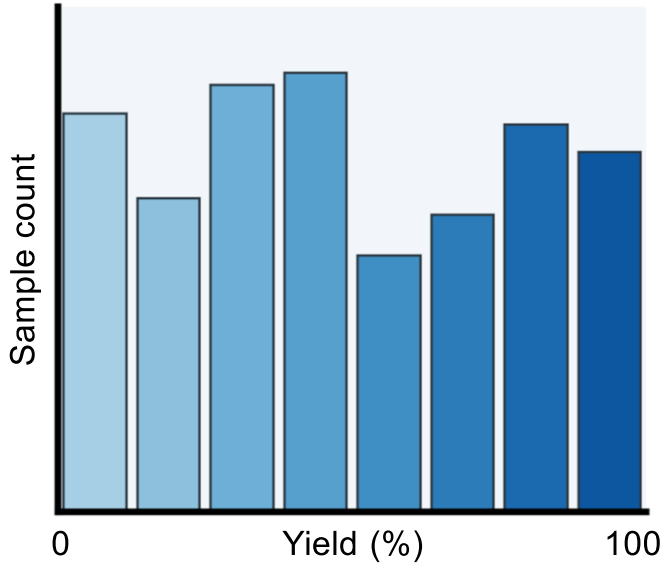
Supervised learning: Data-driven discovery



Yield, enantioselectivity, wavelength, ^{1}H NMR shift, etc

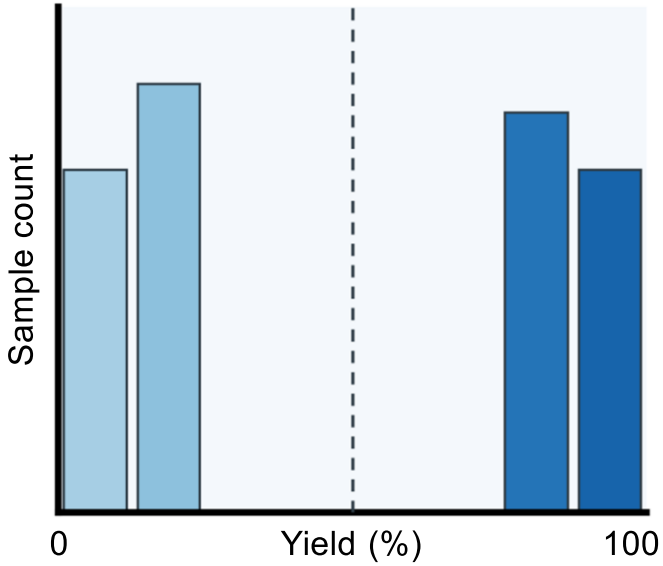
Supervised learning: Data-driven discovery

B. Uniform distribution (regression)



Yield, enantioselectivity, wavelength, ^{1}H NMR shift, etc

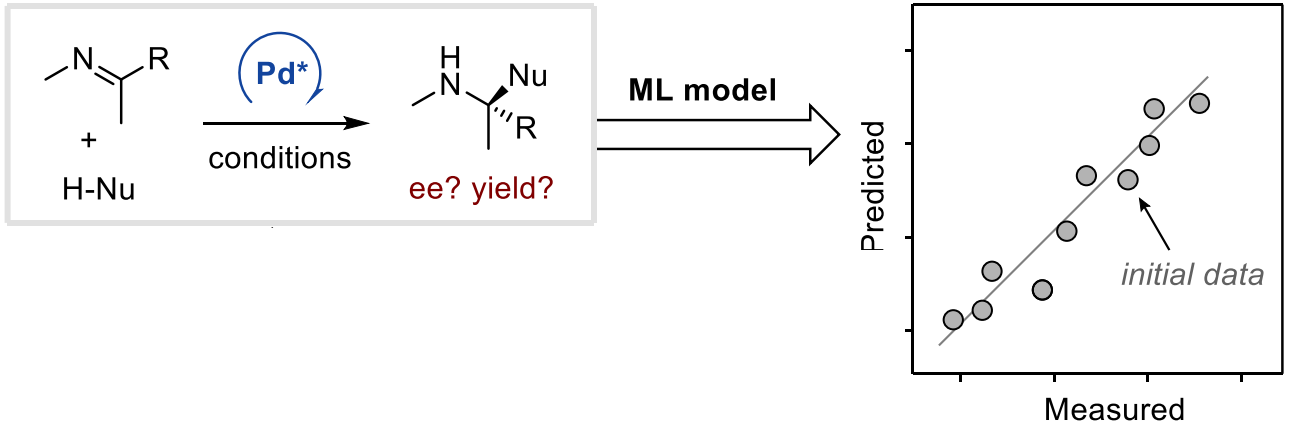
C. Balanced distribution (classification)



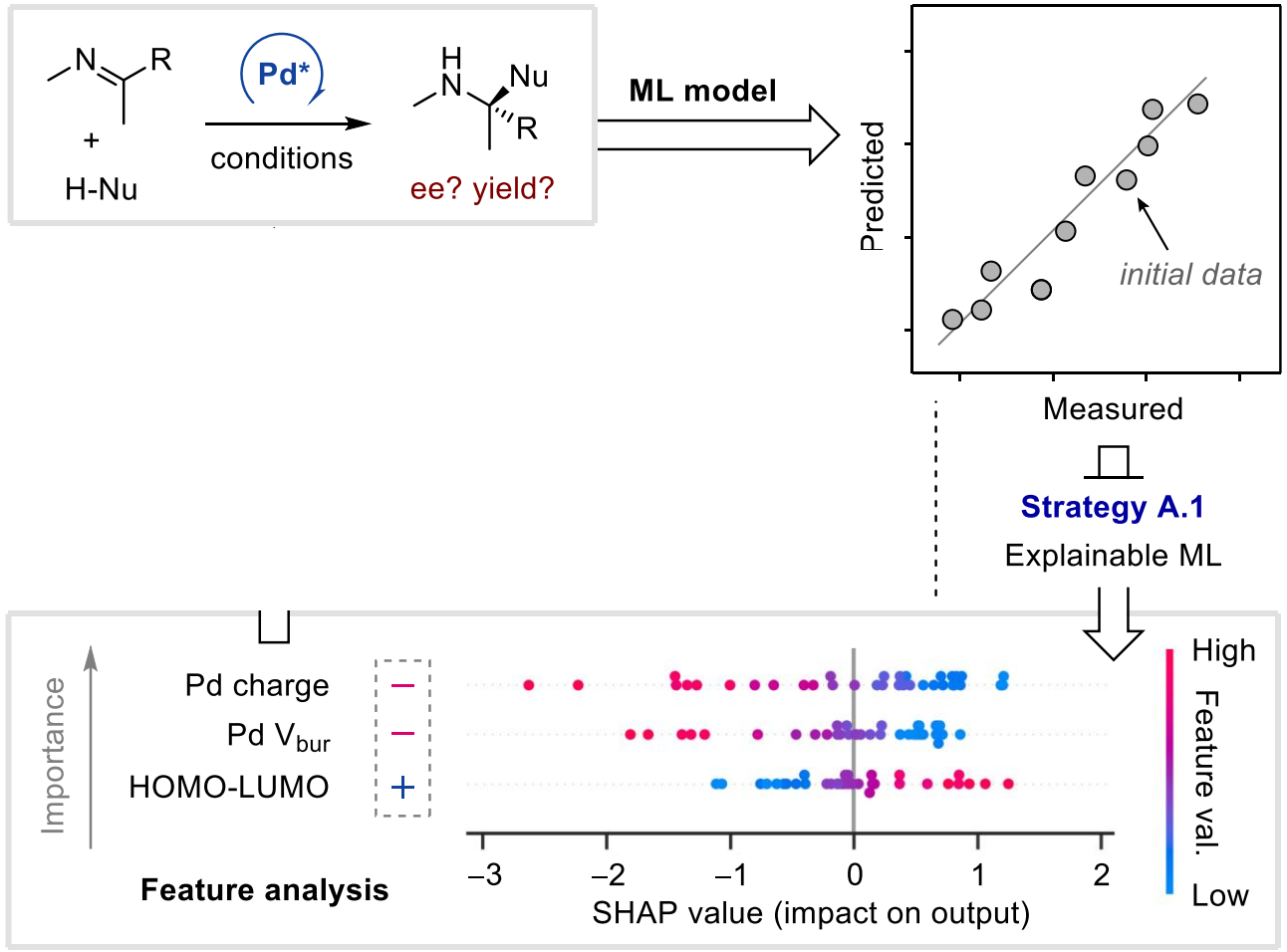
Active vs Inactive, Stable vs Unstable, etc.

Basics of Chemical ML

Supervised learning: Two main strategies

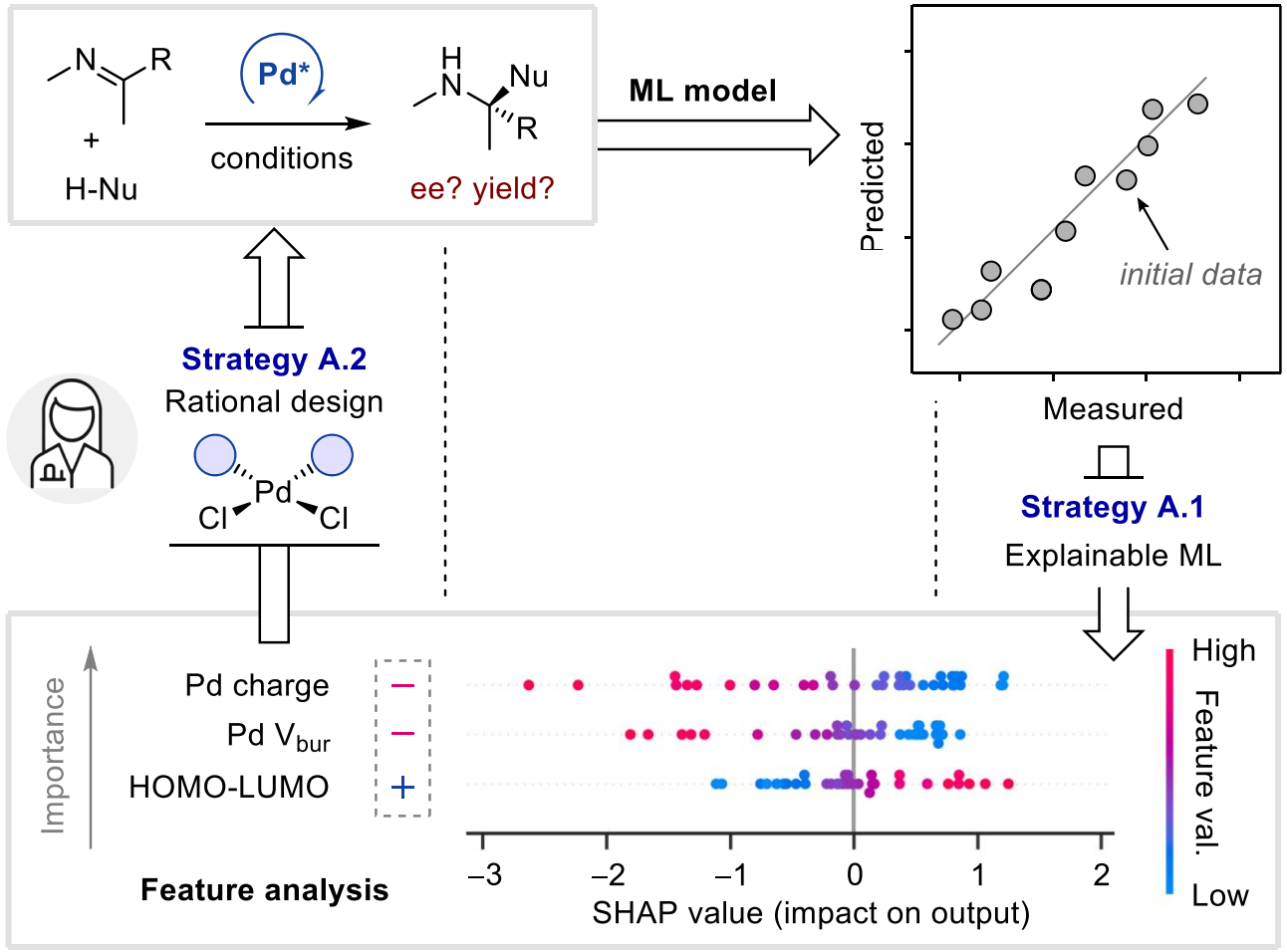


Supervised learning: Two main strategies



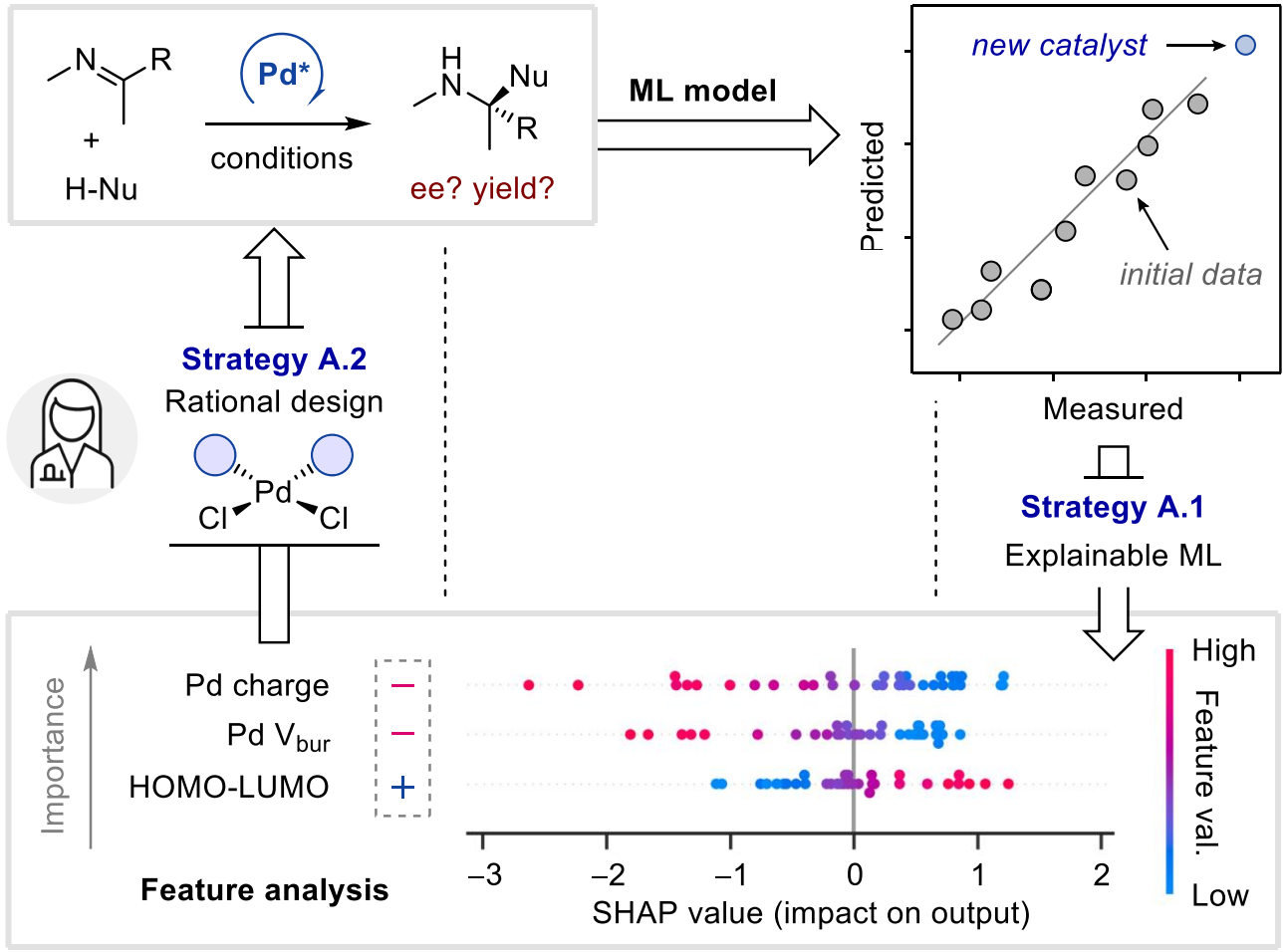
Basics of Chemical ML

Supervised learning: Two main strategies

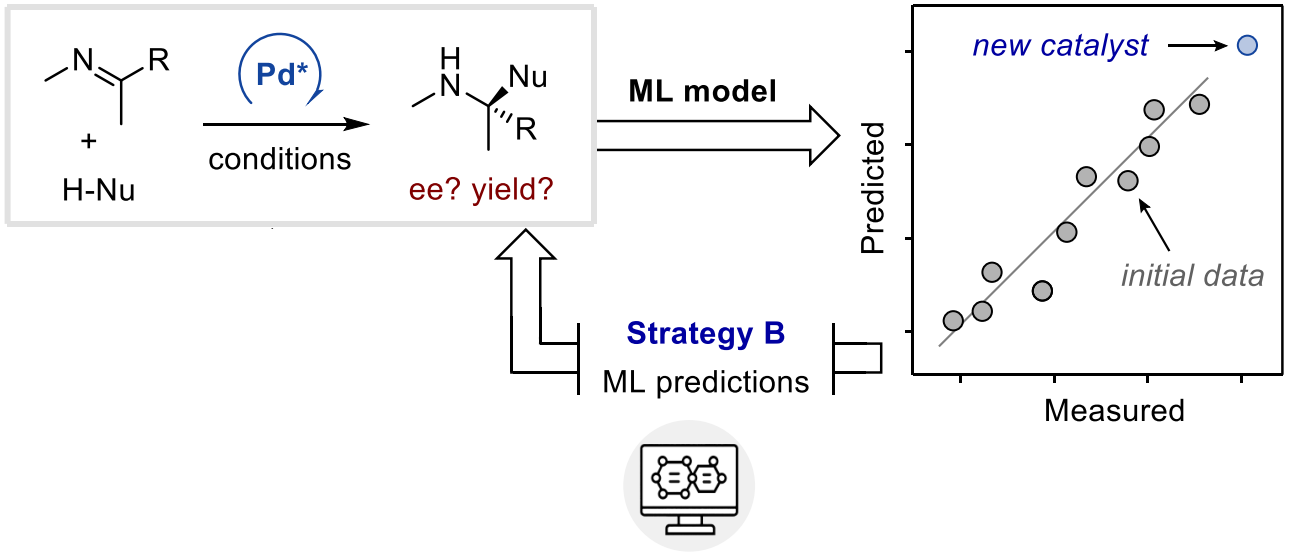


Basics of Chemical ML

Supervised learning: Two main strategies



Supervised learning: Two main strategies



Supervised learning: Data-driven discovery

Key points:

- Supervised learning is useful for understanding processes and designing new molecules
- Examples: catalyst discovery, chromophore design, new materials, etc.
- **CAREFUL!** Supervised learning gets tricky real fast…..

Supervised learning: Data-driven discovery

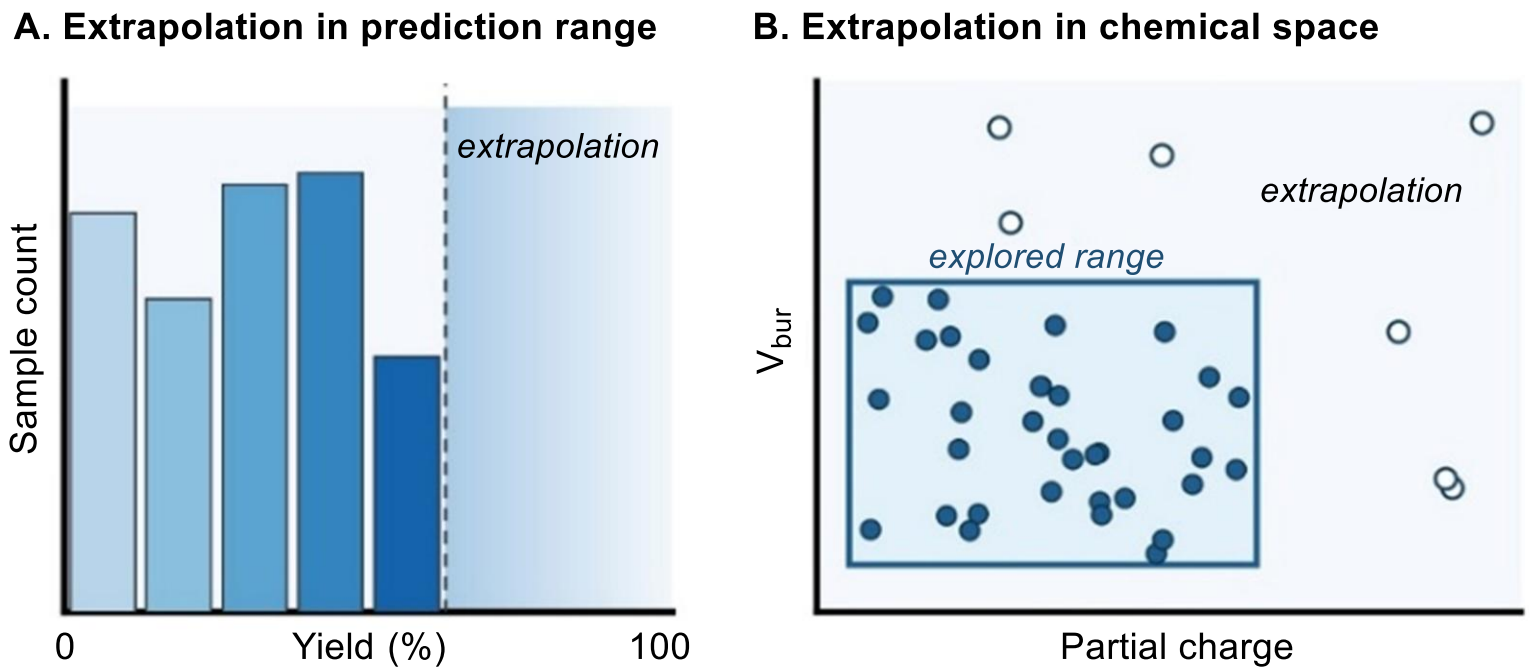
Key points:

- Supervised learning is useful for understanding processes and designing new molecules
- Examples: catalyst discovery, chromophore design, new materials, etc.
- **CAREFUL!** Supervised learning gets tricky real fast…..

Basics of Chemical ML

Supervised learning: Data-driven discovery

ML is not magic, beware of its limitations (easy to mess up)



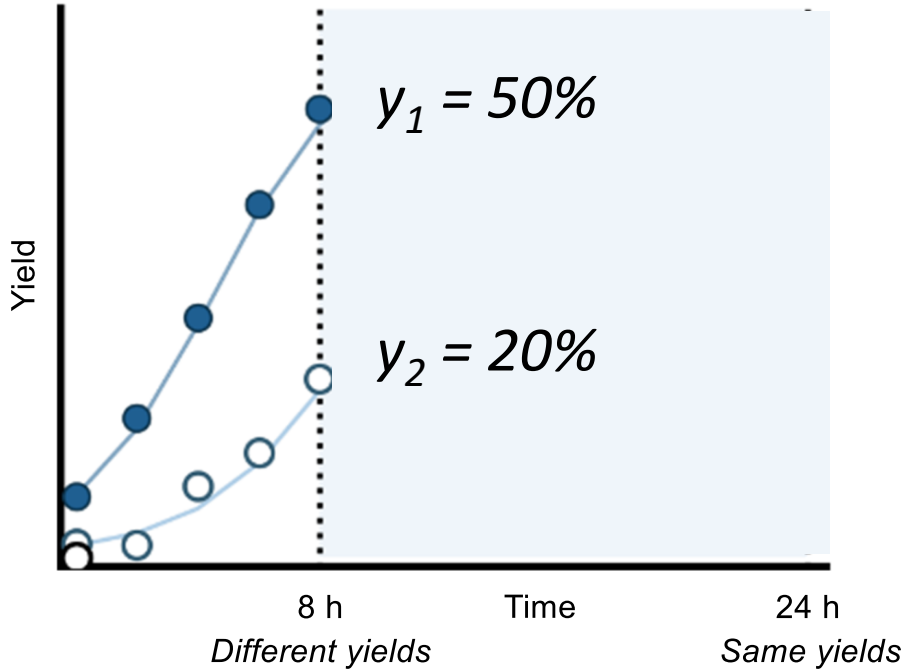
Basics of Chemical ML

Supervised learning: Data-driven discovery

ML is not magic, beware of its limitations (easy to mess up)

Name	SMILES	Descriptors (X_1, ..., X_n)	Target value (y)
Mol_1	CO	X_{11}, ..., X_{1n}	y_1
...	...	...	...
Mol_n	CCO	X_{n1}, ..., X_{nn}	y_n

Catalyst 1, NaF – X_1 (F^- charge) = -0.56
Catalyst 2, NaCl – X_1 (Cl^- charge) = -0.26



Basics of Chemical ML

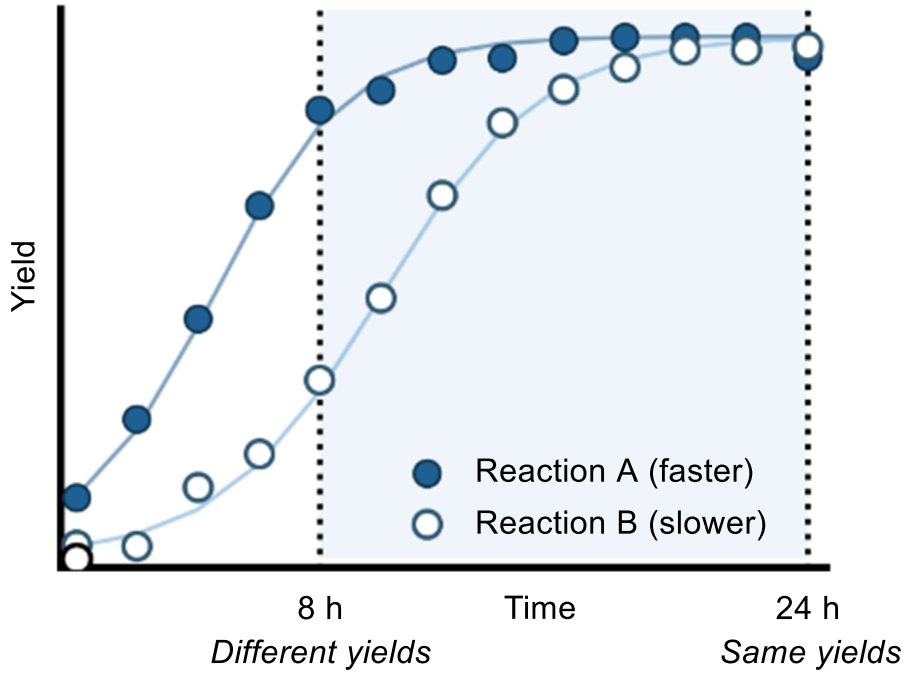
Supervised learning: Data-driven discovery

ML is not magic, beware of its limitations (easy to mess up)

Name	SMILES	Descriptors ($X_1, ..., X_n$)	Target value (y)
Mol_1	CO	$X_{11}, ..., X_{1n}$	y_1
...	...	...	...
Mol_n	CCO	$X_{n1}, ..., X_{nn}$	y_n

Catalyst 1, NaF – X_1 (F^- charge) = -0.56
Catalyst 2, NaCl – X_1 (Cl^- charge) = -0.26

$y_1 = y_2 = 60\%$



Basics of Chemical ML

Supervised learning: Data-driven discovery

14:45 - 16:15 Tutorial

ML automation with ROBERT

Prof: Dalmau

Assist: Luchini

Dr. David Dalmau

Basics of Chemical Machine Learning

September 17, 2025

Dr. Juan Vicente Alegre Requena