“Intelligent” Crystal and Molecular Structure Search Algorithms

Isabelle.braems@cnrs-imn.fr

Before we start

not an astro-chemist Closest.. PostDoctorate in Bayesian Tools for Exoplanet detection (Terrestrial Planet Finder)

Sequential background in data science (ex « Signal Processing ») and atomistic modelling (Monte Carlo, DFT..)

THE ASTROPHYSICAL JOURNAL, 646:1260–1274, 2006 August 1
© 2006. The American Astronomical Society. All rights reserved. Printed in U.S.A.

LINEAR AND BAYESIAN PLANET DETECTION ALGORITHMS FOR THE *TERRESTRIAL PLANET FINDER*

N. JEREMY KASDIN
Mechanical and Aerospace Engineering, Princeton University,
Princeton, NJ 08544; jkasdin@princeton.edu

AND

ISABELLE BRAEMS
LEMHE / ICMMO, UMR 8182, Université Paris-Sud,
91405 Orsay Cedex, France; isabele.braems@lemhe.u-psud.fr
Received 2005 December 16; accepted 2006 April 5

Before we start

not an astro-chemist Closest.. PostDoctorate in Bayesian Tools for Exoplanet detection (Terrestrial Planet Finder)

Sequential background in data science (ex « Signal Processing ») and atomistic modelling (Monte Carlo, DFT..)

Computational Materials Design to accelerate discovery of new materials

Assisting experimentalists (academic and industrials) to identify their matter (crystal) sample structure *before* or *after* synthesis

Eco-responsible materials: soft-chemistry derived **mixed-anion compounds** (oxychalcogenides), **sulfides**, oxides, nitrites, and hydrated chlorides, for **new** greener batteries, pollution control devices, hydrogen generators, sensors and displays, (electronic) memories, understanding fission byproducts for recycling, space transportation, photovoltaics…

Before we start

not an astro-chemist Closest.. PostDoctorate in Bayesian Tools for Exoplanet detection (Terrestrial Planet Finder)

Sequential background in data science (ex « Signal Processing ») and atomistic modelling (Monte Carlo, DFT..)

Computational Materials Design to accelerate discovery of new materials

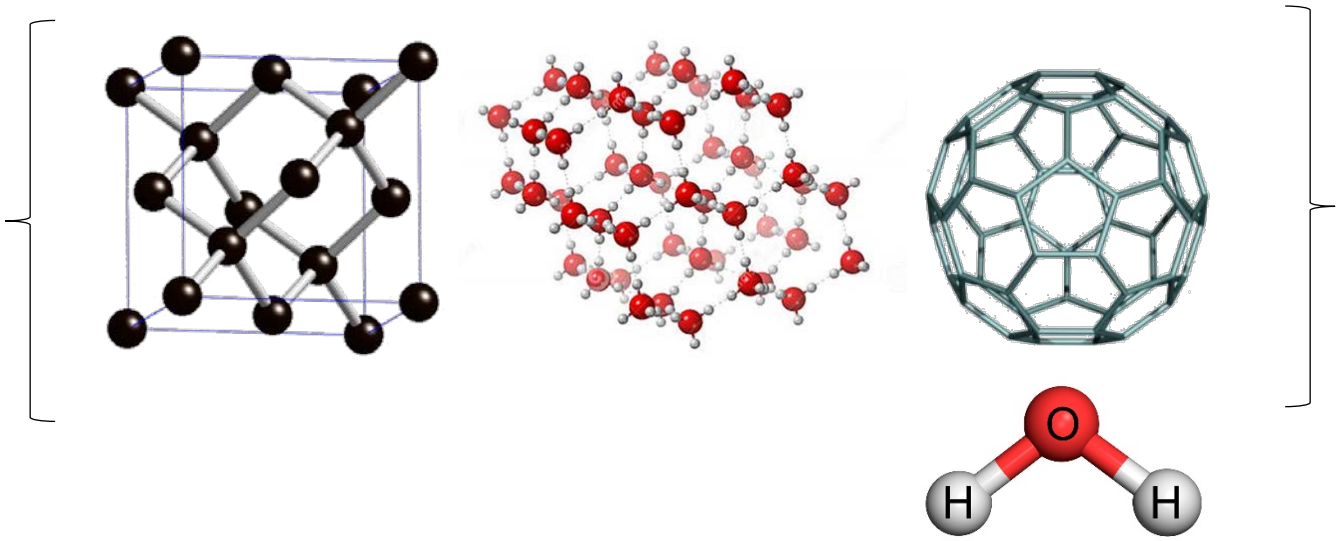
Assisting experimentalists (academic and industrials) to identify their matter (crystal) sample structure before or after synthesis

Eco-responsible materials: soft-chemistry derived **mixed-anion compounds** (oxychalcogenides), **sulfides**, oxides, nitrites, and hydrated chlorides, for **new** greener batteries, pollution control devices, hydrogen generators, sensors and displays, (electronic) memories, understanding fission byproducts for recycling, space transportation, photovoltaics…

Booming field : this talk mostly covering (2023-2025) is not exhaustive and may (will) be outdated in a year!

Structure?

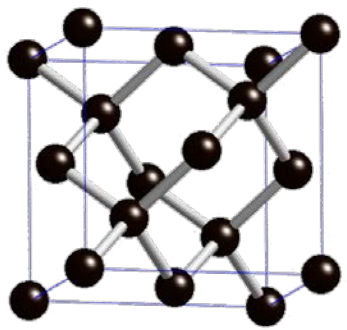
(Chemically labelled) Structure

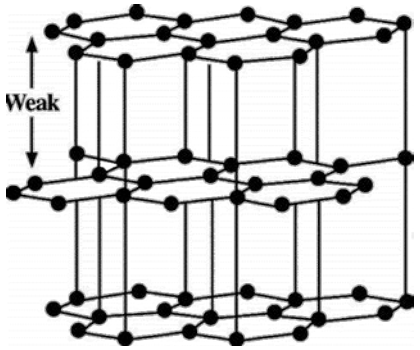


Only crystalline here (no amorphous, see M. Caro's talk)

Structure?

(Chemically labelled) Structure —explains→ Properties, Reactivity…

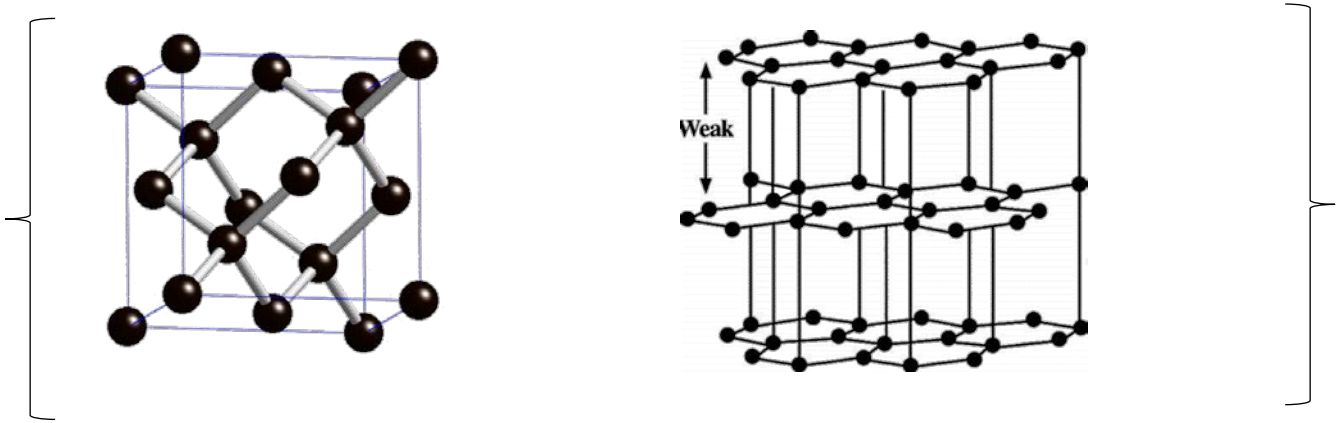




Only crystalline here (no amorphous, see M. Caro's talk)

Structure?

(Chemically labelled) Structure —explains→ Properties, Reactivity…



Only crystalline here (no amorphous, see M. Caro's talk)

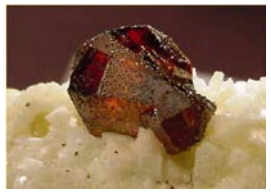
Structure search/**prediction**

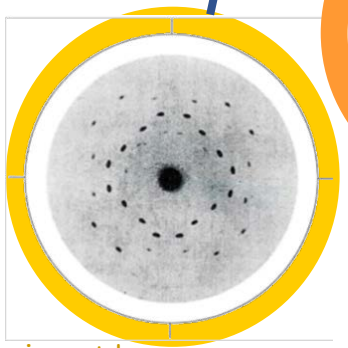


Property search/**prediction**



« Seeing is believing » Venkatraman Ramakrishnan (NP Chemistry, 2009)

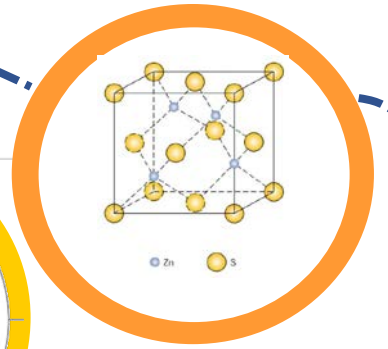


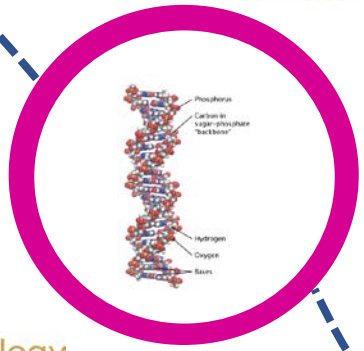


Friedrich, Knipping et Laue
Sitzungsberichte der Kgl.
Bayer. *Akad. der Wiss*. 1912

The Nobel Prize in Physics 1914

Max von Laue "for his discovery of the diffraction of X-rays by crystals"





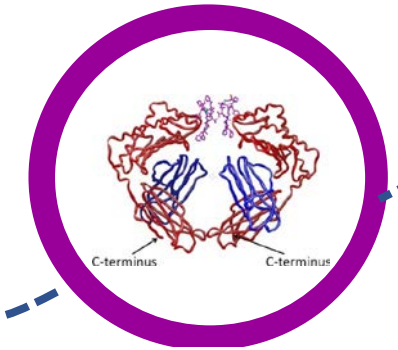
Nobel Prize in Physiology
or Medicine 1962

Francis H. C. Crick, J. D. Watson and M. H. F. Wilkins "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material"

(Rosalind Franklin and Maurice Wilkins)

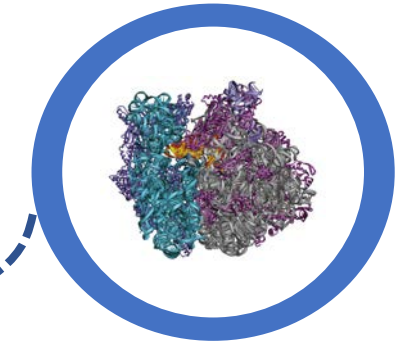
The Nobel Prize in Physics 1915

Sir William Henry Bragg and William Lawrence Bragg "for their services in the analysis of crystal structure by means of X-rays"



The Nobel Prize in Chemistry 1985

Herbert A. Hauptman and Jerome Karle "for their outstanding achievements in the development of direct methods for the determination of crystal structures"



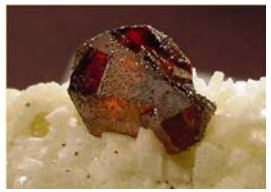
Nobel Prize in Chemistry
2009

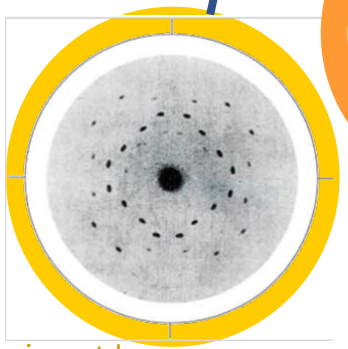


50

https://www.nobelprize.org/

« Seeing is believing » Venkatraman Ramakrishnan (NP Chemistry, 2009)

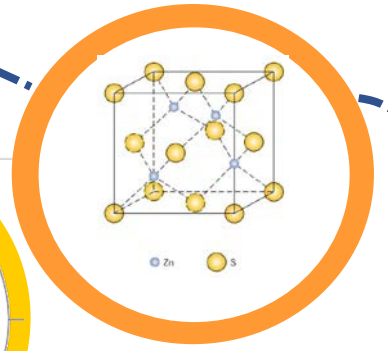




Friedrich, Knipping et Laue
Sitzungsberichte der Kgl.
Bayer. *Akad. der Wiss.* 1912

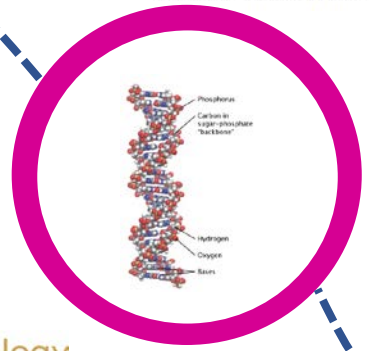
The Nobel Prize in Physics 1914

Max von Laue "for his discovery of the diffraction of X-rays by crystals"



The Nobel Prize in Physics 1915

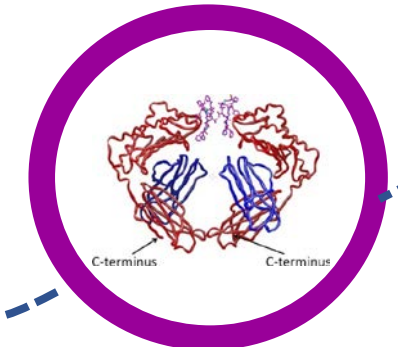
Sir William Henry Bragg and William Lawrence Bragg "for their services in the analysis of crystal structure by means of X-rays"



Nobel Prize in Physiology or Medicine 1962

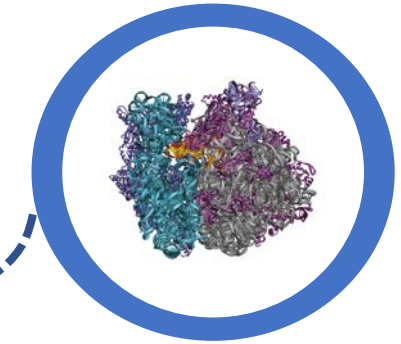
Francis H. C. Crick, J. D. Watson and M. H. F. Wilkins "for their discoveries concerning the molecular structure of nucleic acids and its significance for information transfer in living material"

(Rosalind Franklin and Maurice Wilkins)



The Nobel Prize in Chemistry 1985

Herbert A. Hauptman and Jerome Karle "for their outstanding achievements in the development of direct methods for the determination of crystal structures"



Nobel Prize in Chemistry 2009



50

https://www.nobelprize.org/







THE PRINCIPLES DETERMINING THE STRUCTURE OF COMPLEX IONIC CRYSTALS

The Nobel Prize in Chemistry 1954

L. C. Pauling "for his research into the nature of the chemical bond and its application to the elucidation of the structure of complex substances"



The Nobel Prize in Chemistry 2024

one half awarded to D. Baker "for computational protein design", the other half jointly to D. Hassabis and J. Jumper "for protein structure prediction"

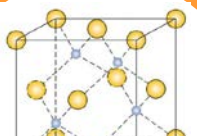


2

"RCSB Protein Data Bank" (CC-BY-4.0 license)

« Seeing is believing » Venkatraman Ramakrishnan (NP Chemistry, 2009)





Acc. Chem. Res. **1994,** *27,* 309−314

Are Crystal Structures Predictable?

ANGELO GAVEZZOTTI*

Dipartimento di Chimica Fisica ed Elettrochimica, Università di Milano, Milano, Italy

Received May 16, 1994

"No": by just writing down this concise statement, in what would be the first one-word paper in the chemical literature, one could safely summarize the

nd its

structure of complex substances"

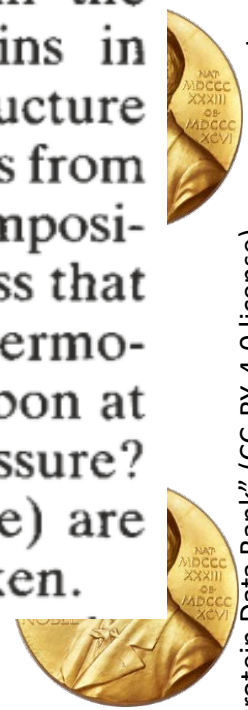
The Nobel Prize in Physics 1915

Sir William Henry Bragg and William Lawrence Bragg "for their services in the analysis of crystal structure by means of X-rays"

NATURE VOL. 335 15 SEPTEMBER 1988 NEWS AND VIEWS 201

Crystals from first principles

ONE of the continuing scandals in the physical sciences is that it remains in general impossible to predict the structure of even the simplest crystalline solids from a knowledge of their chemical composition. Who, for example, would guess that graphite, not diamond, is the thermodynamically stable allotrope of carbon at ordinary temperature and pressure? Solids such as crystalline water (ice) are still thought to lie beyond mortals' ken.



The Nobel Prize in Chemistry 2024

one half awarded to D. Baker "for computational protein design", the other half jointly to D. Hassabis and J. Jumper "for protein structure prediction"

https://www.nobelprize.org/

"RCSB Protein Data Bank" (CC-BY-4.0 license)

Nature does not obey Pauling rules

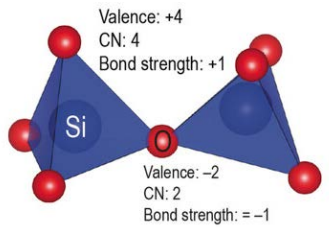
- Organic conformation rules (VSEPR Gillespie-Nyholm Theory, Torsion/conformation rules, environment effects…) , **see B. MacGuire's talk**
- 1920's crystal « geometrical restraint » rules : Goldschmidt (substitution, 1925), Hume-Rothery (alloys, 1926), Pauling (ionic crystals, 1929)

Solid-State Chemistry

Deutsche Ausgabe: DOI: 10.1002/ange.20200082
Internationale Ausgabe: DOI: 10.1002/anie.20200082

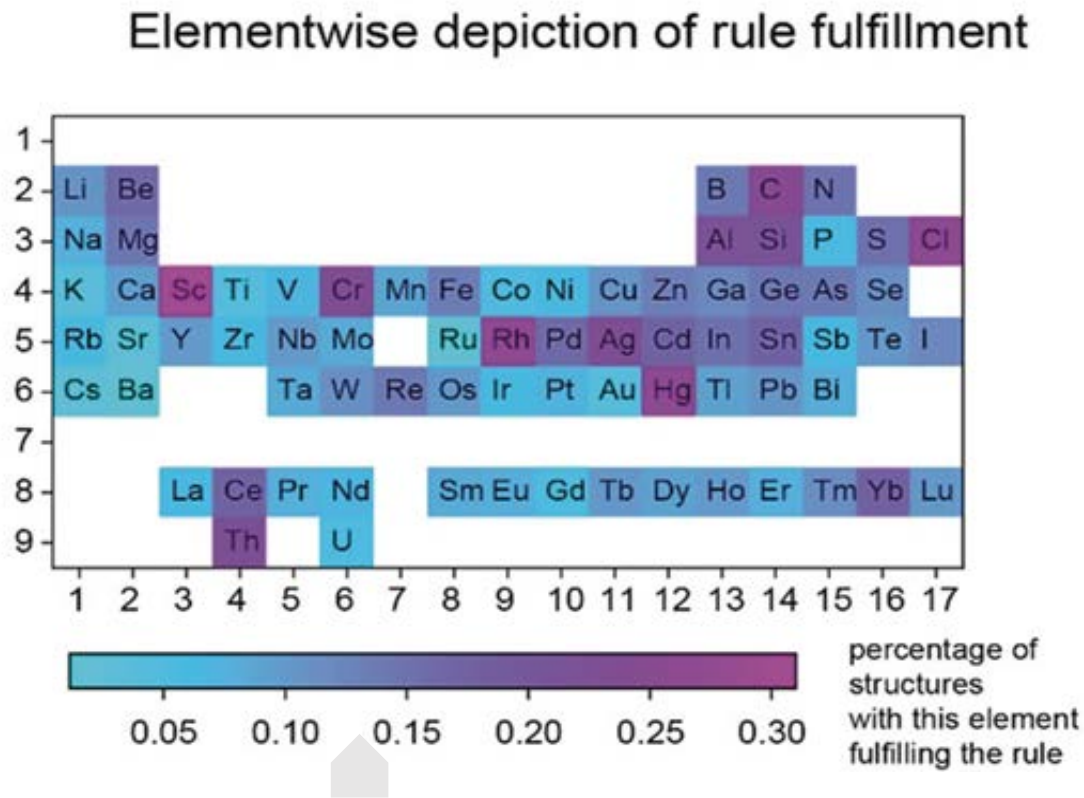
The Limited Predictive Power of the Pauling Rules**

*Janine George, David Waroquiers, Davide Di Stefano, Guido Petretto, Gian-Marco Rignanese und Geoffroy Hautier**

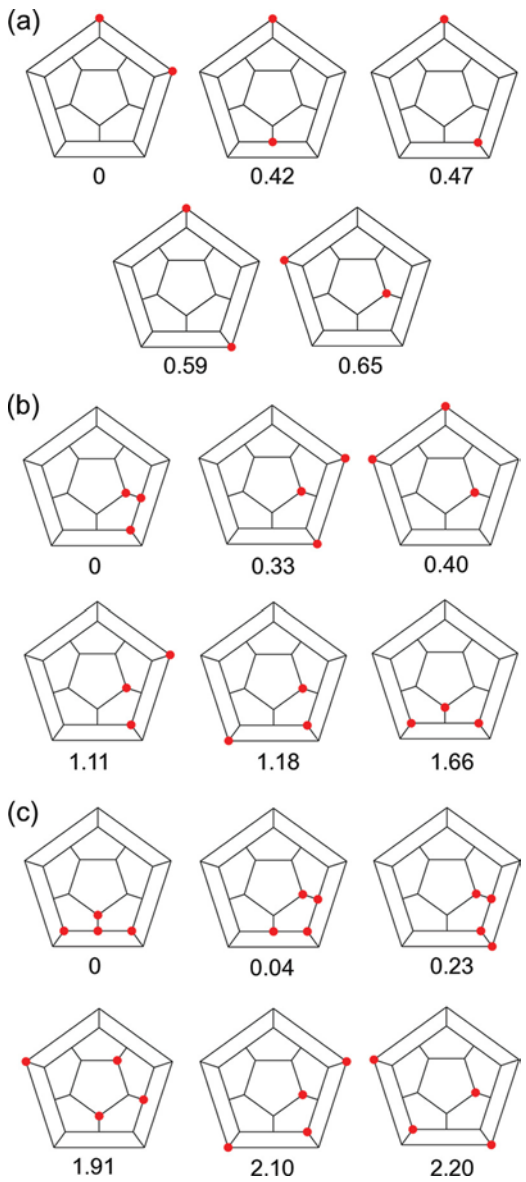


ICSD

13% of 5000 oxides obey rules **2-5**



Suppose we want… can we find them *all*?



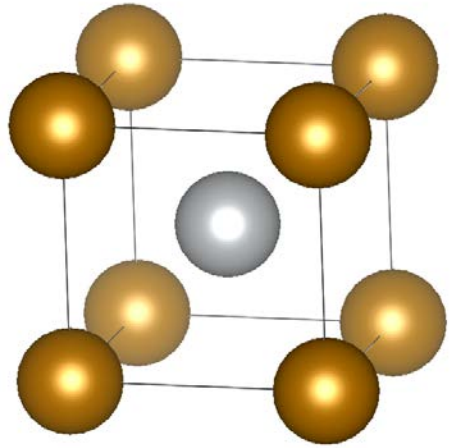
Crystal

N = total number of atoms in the unit cell of volume V,
d = relevant discretization parameter (for instance, 1 A)

$$C = \frac{N!}{\left(N - \frac{V}{\delta}\right)!\frac{V}{\delta}!}\prod_i \frac{N!}{n_i!\,(N-n_i)!} = 10^N$$

For small systems (N=10) with (d=1 A, V 10 A^3) $C=10^N$:
t = 317 years…. (if 1 energy/optimization cycle = 1s!!)

Oganov A.R. et al (Editors). Computational Materials Discovery. RSC



Molecule functionalization

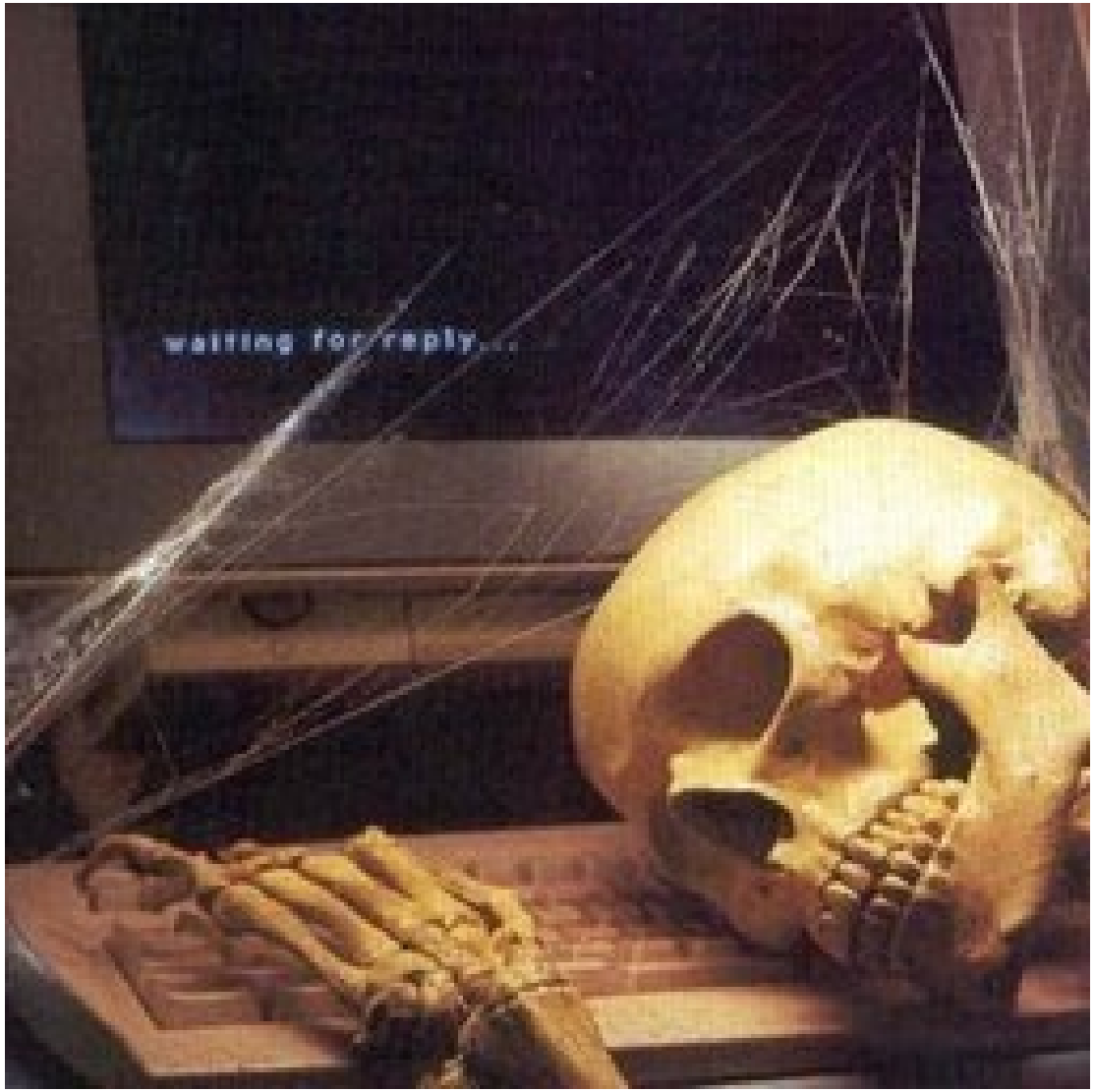
Number of distinct isomers is

$$C = \frac{M!}{n!\,(M-n)!}$$

M = size of the cage
n = number of exohedral functional groups,
For I_h–C60 (few distinct isomers as icosahedral symmetry):
C60X4 : 487,635 isomers (4,165 symmetrically distinct),
C60X5 : over **5,461,512** isomers (45,010 symmetrically distinct).

J. Chem. Theory Comput. 2012, 8, 1841–1851

Exhaustive search of conformers/isomers/polymorphs by combination is out of question



https://knowyourmeme.com/

Do we need to seek for them all?

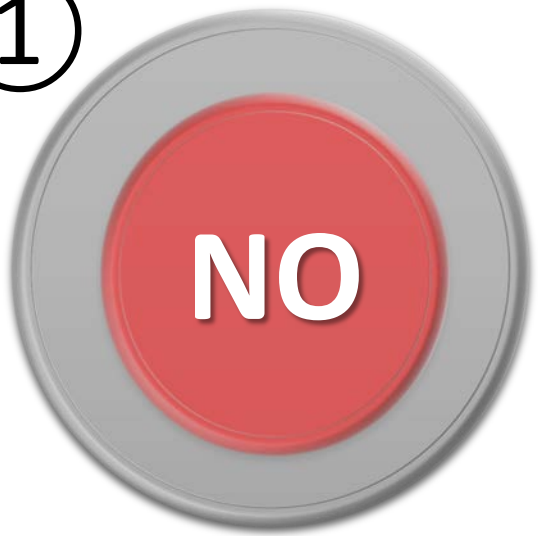
①



Discriminative AI

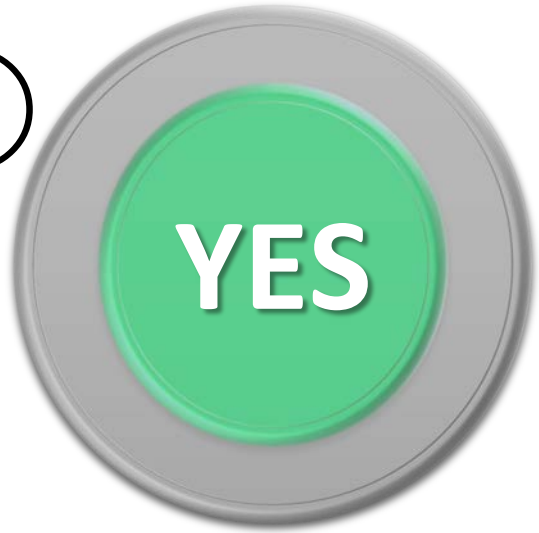
Do we *need* to seek for them all?

1



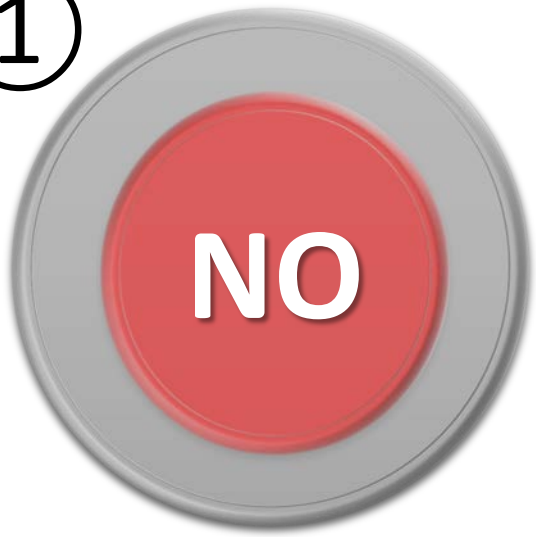
Discriminative AI

2



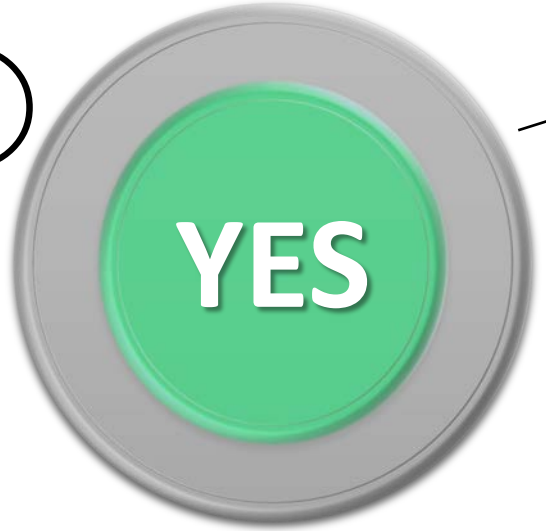
Do we *need* to seek for them all?

①



Discriminative AI

②



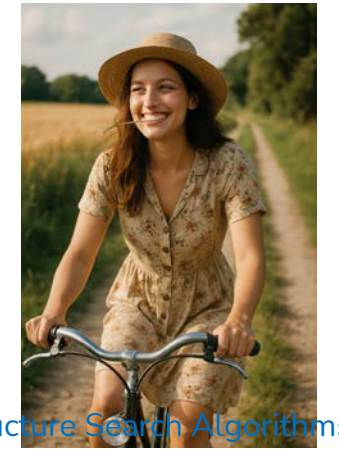
Generative *AI*

Atlanta « Spaghetti Junction »

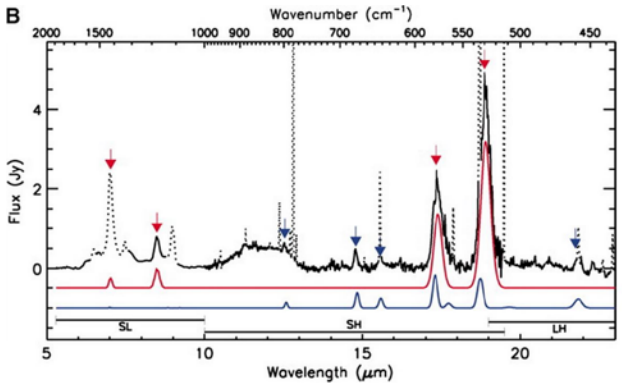


Google Maps

Global optimisation

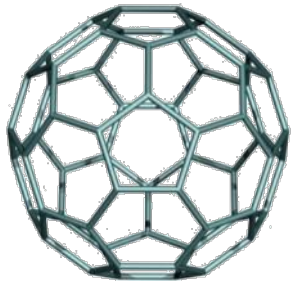


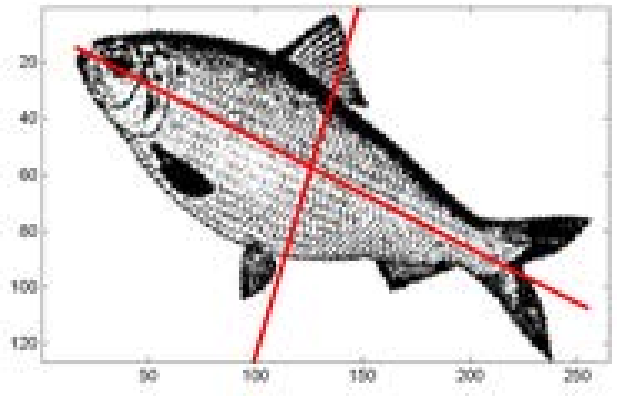
Discriminative AI?



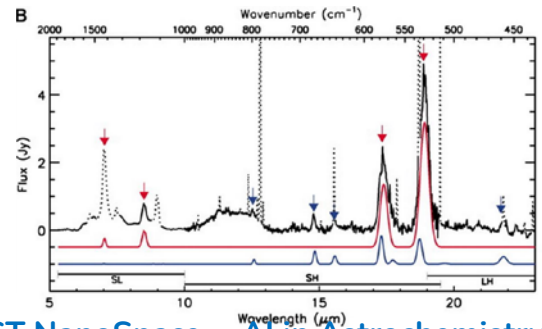
Inverse problem is ill-posed

(dimension augmentation)

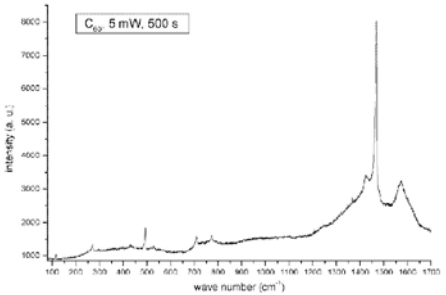




Combined multiple spectra

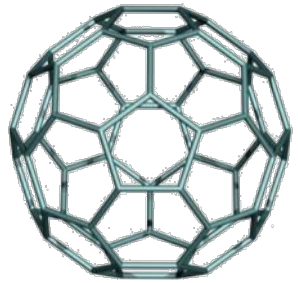


+....+



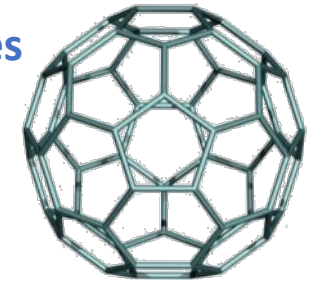
NI: routine

AI: still in R&D



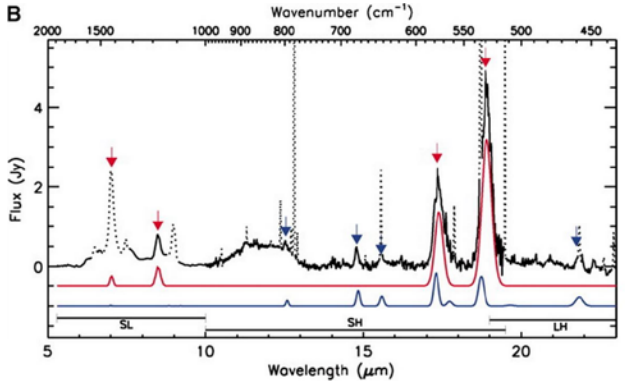
Discriminative AI?

Few candidate structures



Evaluation

(Forward method)



Similarity check
(match)
on the observable(s)

See J. Heyl's
Trinity of astrochemistry

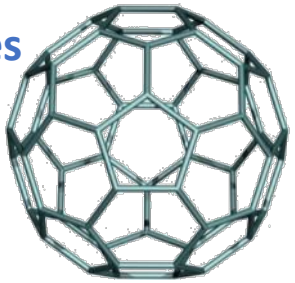
Structural « similarity »
+
Chemical « similarity »

Properties similarity

Goldschmidt substitution rule, chemist expertise
first superconducting pnictide oxide $LaFeAsO_{1-x}F_x$.
La->Sm substitution : new isostructural compound $SmFeAsO_{1-x}F_x$

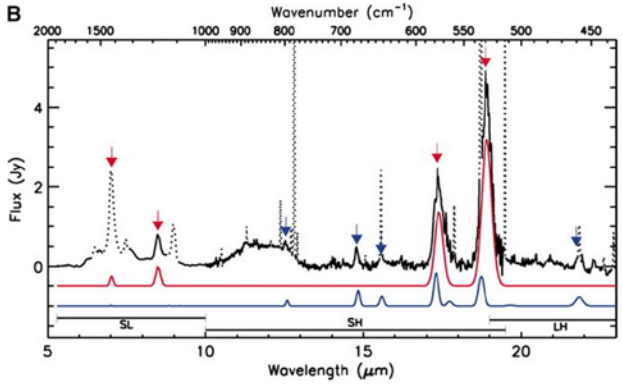
Discriminative AI?

Many candidate structures



High-Throughput

Virtual screening (HTVS)



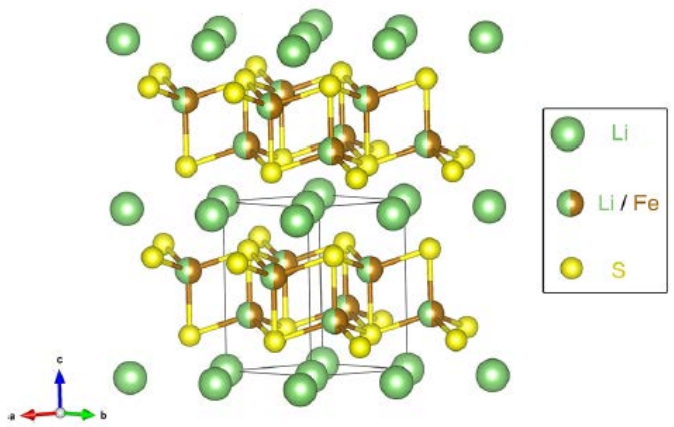
Similarity check
(match)
on the observable





Data-driven, automated, extensive, accelerated search
Yes/No answer
User-selected library **must contain promising structures**
➢ Wide (MLP over DFT)
see R. Ibragimova's talk
Otherwise Data Mining

Double delithiation of Li_2FeS_2 gives an long-term unknown FeS_2 structure



Li_2FeS_2

-Li

Unknown structure since 1989

$Fe^{3+}S^{2-}(S_2)^{2-}{}_{1/2}$

- ***IR spectroscopy :***
  - d(S–S)= 2.12 Å
- ***Mössbauer and EXAFS :***
  - Fe^{3+}HS ion in **tetrahedral sites**
  - d(Fe–S)=2.28 Å
  - S^{2-}/S_2^{2-} **= 2/1**
- **Lamellar structure** expected due to the stacking of $LiFeS_2$ and Li layers.

3D

ICSD 307,301

The Materials Project 154,718

OQMD The Open Quantum Materials Database 1,226,781

AFLOW Automatic-FLOW for Materials Discovery 3,530,33

MATERIALSCLOUD

NOMAD NOVEL MATERIALS DISCOVERY

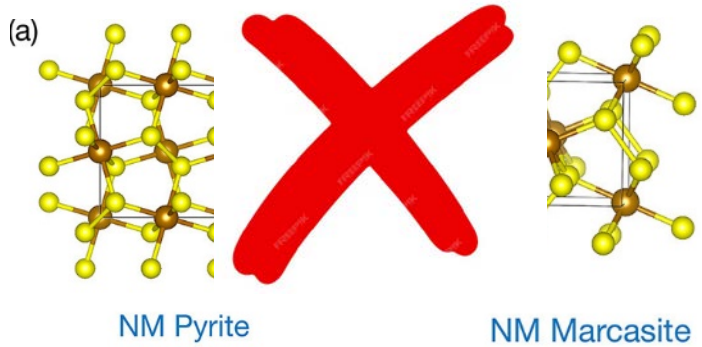
0 PAULING FILE

LEARN WORK DISCOVER EXPLORE ARCHIVE

39,037,528

(seek for interoperable platforms

and use API)



R. Brec et al. *J. Power. Sources.* **1989**, *26*, 325

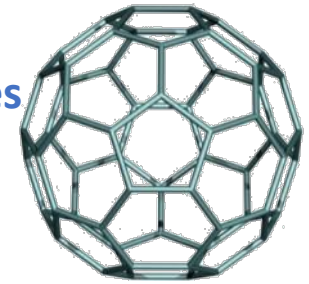
B. Wang, IB, S. Sasaki, G. Guegan, L. Cario, S. Jobic and G. Frapper, *J. Phys. Lett.* **2020**, *11*, 8861.

Discriminative AI?



Incomplete databases for exotic molecules/ novel crystals - YES/NO answer is not informative enough

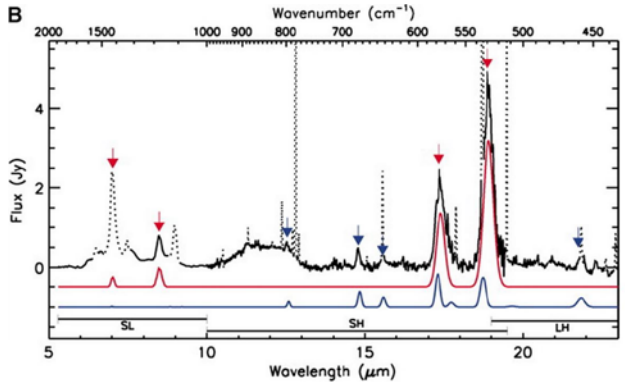
Many candidate structures
But no YES



Structure X

Substitution-based

Data Mining

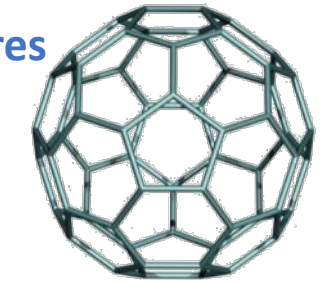


Property class Y

Conditional Probability
P(Y|X)

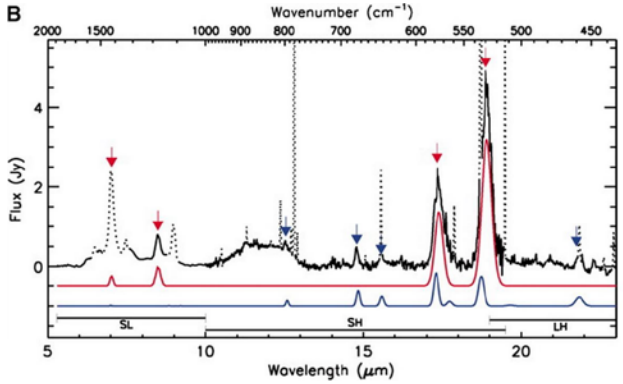
Discriminative AI?

Many candidate structures
But no YES



Substitution-based

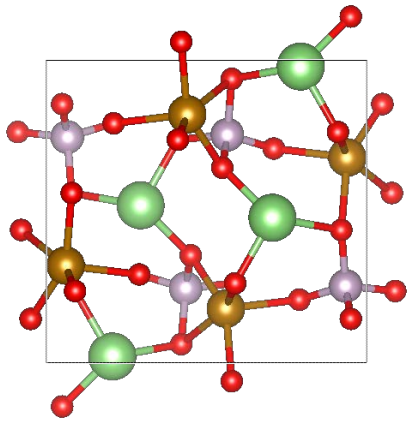
Data Mining



Conditional Probability
P(Y|X)

statistics-based criterion : probabilistic model that predicts ionic substitutions which keep the crystal structure of a compound inchanged.

G Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, *Inorg. Chem*. 2011, **50**, 656–663



$$p_5(Sm, Fe^{2+}, As, O^{2-}, F^{-} | La, Fe^{2+}, As, O^{2-}, F^{-})$$

How likely Fe^{2+} in a lithium transition metal phosphate is to be substituted by Ni^{2+}?

$p_4(Ni^{2+}, Li^{1+}, P^{5+}, O^{2-} | Fe^{2+}, Li^{1+}, P^{5+}, O^{2-})$:

Answer: high as Ni^{2+} and Fe^{2+} are both transition metals with similar charge and size.
Actually $LiNiPO_4$ and $LiFePO_4$ both form the olivine-like structure

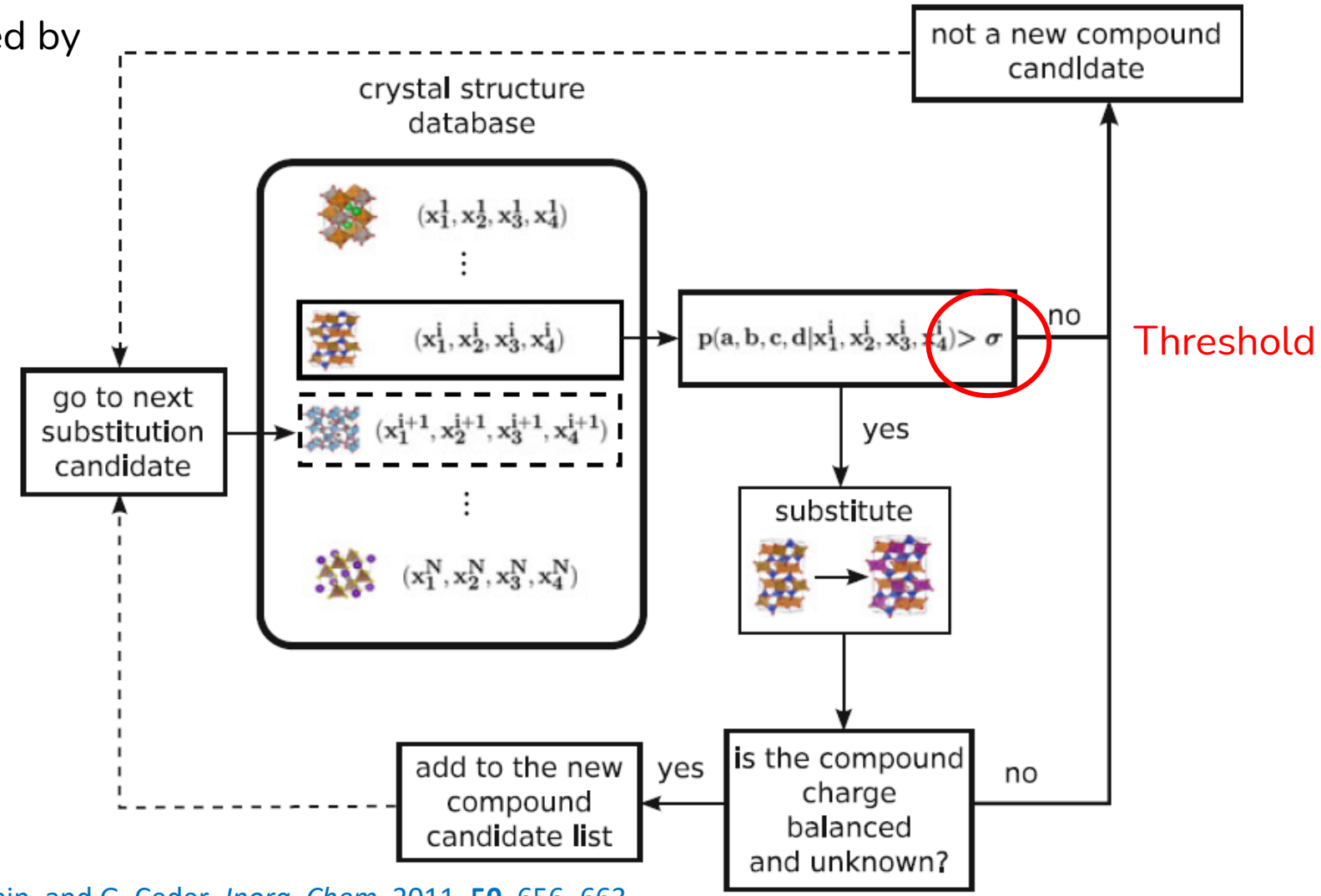
How likely Sr^{2+} in a lithium transition metal phosphate is to be substituted by Ni^{2+}.



$p_4(Sr^{2+}, Li^{1+}, P^{5+}, O^{2-} | Fe^{2+}, Li^{1+}, P^{5+}, O^{2-})$: Answer: low !

Data Mining based on fragment substitution on similar structures

Predict a compound formed by species a,b,c,d



G Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, *Inorg. Chem*. 2011, **50**, 656–663

Data Mining based on fragment substitution on similar structures

Setting the threshold

True positive rate
= fraction of existing compounds found back by the model

$$TP_{rate}(\sigma) = \frac{TP(\sigma)}{P}$$

Number of these existing compounds
Found with the given threshold σ

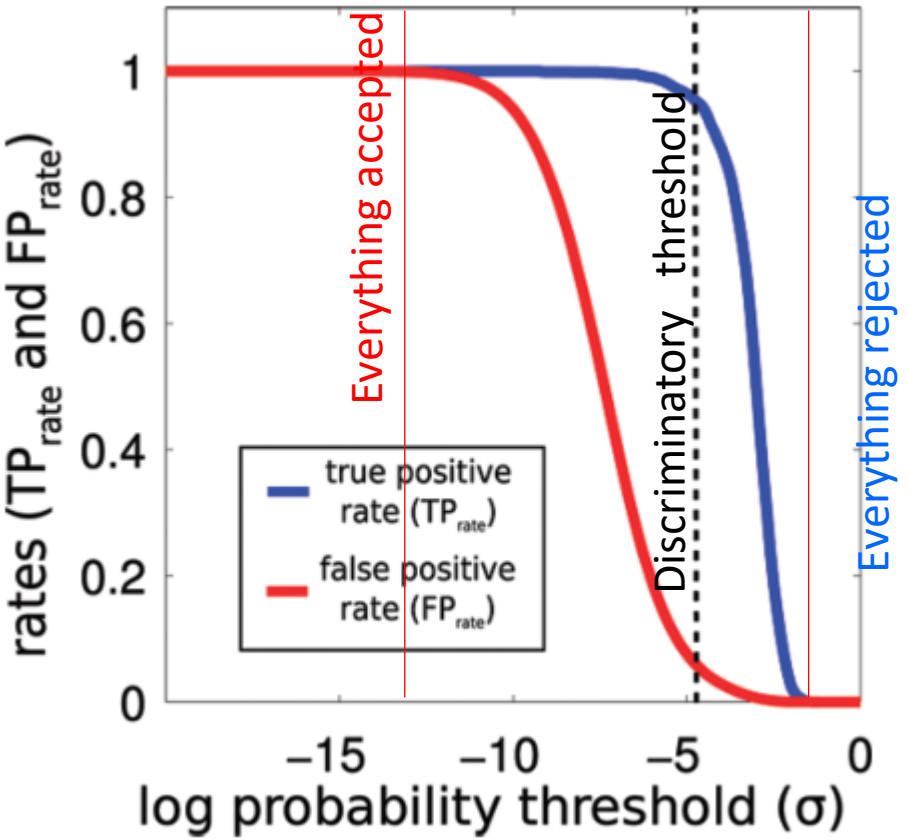
Number of existing compounds
during the cross-validation test

False positive rate
= fraction of compounds not existing but suggested by ML

$$FP_{rate}(\sigma) = \frac{FP(\sigma)}{N}$$

Number of these non-existing compounds
Found with the given threshold σ

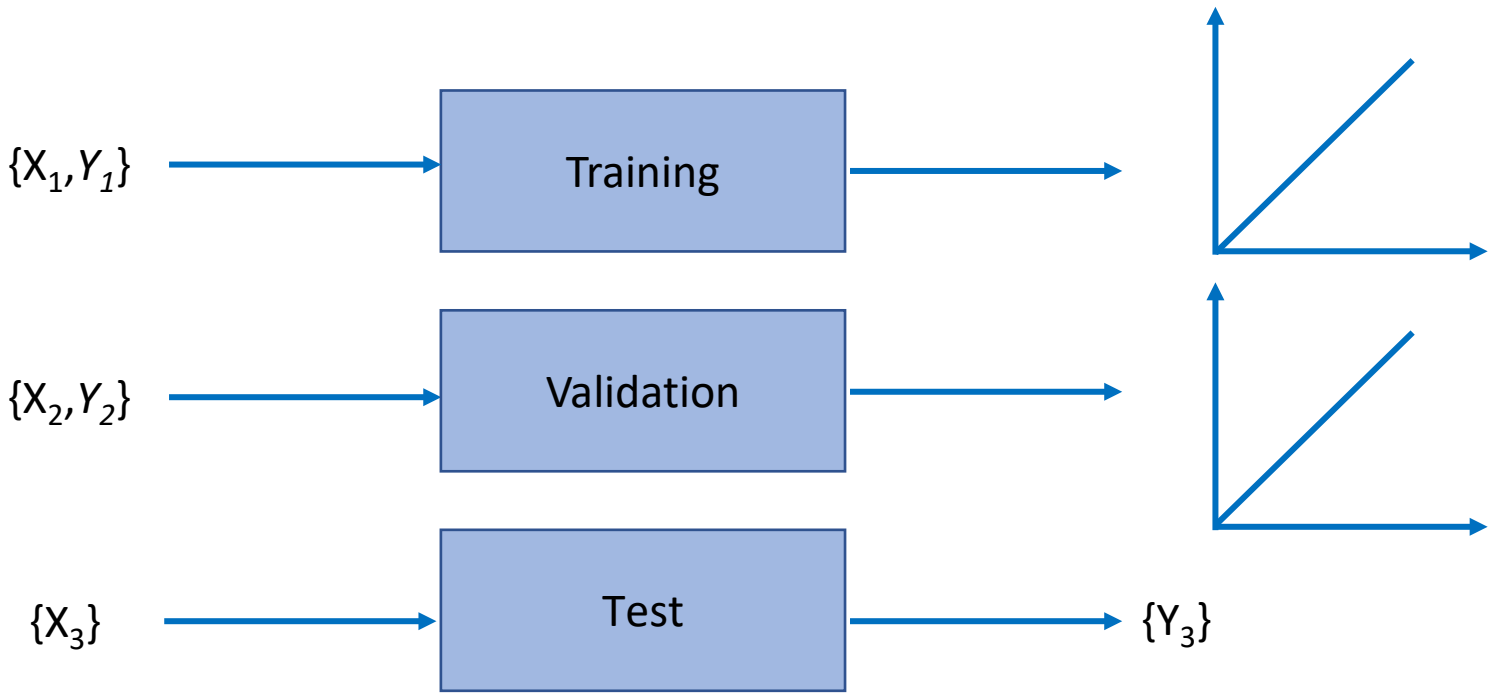
Number of proposed compounds non existing
But considered during the cross-validation



G Hautier, C. Fischer, V. Ehrlacher, A. Jain, and G. Ceder, *Inorg. Chem.* 2011, **50**, 656–663

Discriminative AI learns to distinguish (like an art critic)

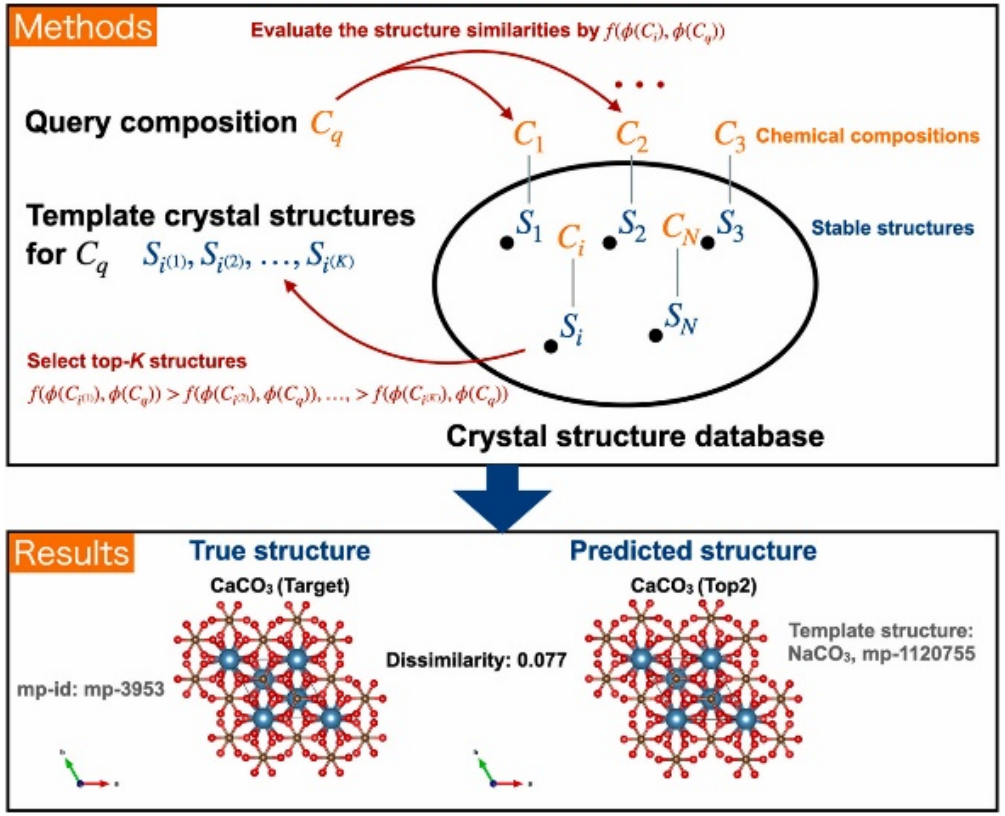
Same statistical problem with advanced capabilities:
Models the probability **P(Y|X)** of class Y (same structure) given structure X and **decides (classifies)**
Employs M/L learnings to account for more information (local neighbourhood, descriptors)



Discriminative AI learns to distinguish (like an art critic)

Crystal Structure Prediction with ML-based element substitution (CSPML) using more complex structure descriptors (hidden structure-element relationships)

A binary classifier, trained on a large number of known crystal structures, can determine the isomorphism of crystal structures formed for two given chemical compositions with an accuracy of around **96.4%**.



M. Kusaba, C. Liu, R. Yoshida, *Comp. Mater. Sci.* **211**, 111496 (**2022**)

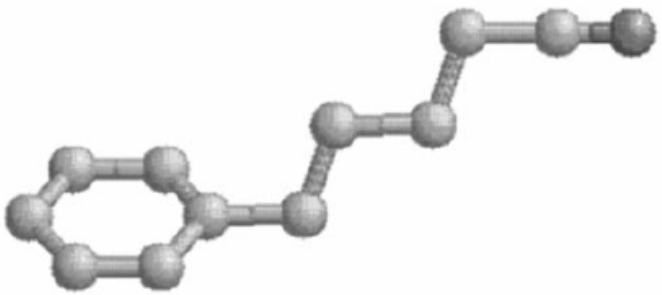
« Historical » NN (2000): to routinely analyze IR spectra (substitution/discrimination)

Molecule Structure

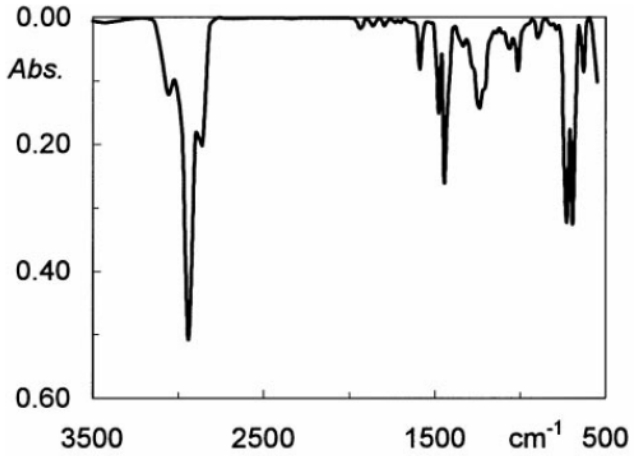
Spectral simulator

IR Spectrum

NN (CPN)

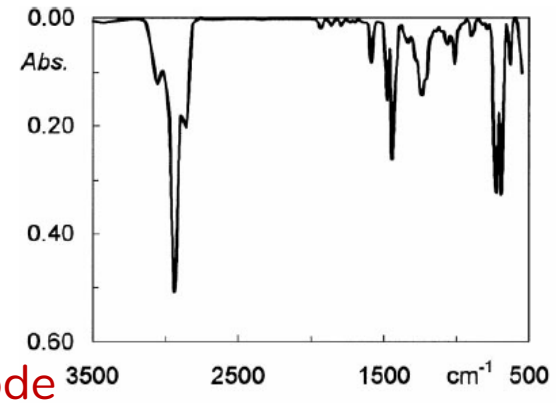


4-phenylbutannitril structure



Infrared Spectrum

Hadamard transformation



RDF code

λ_1	A_1
λ_2	A_2
...	...
λ_n	A_n

Absorbance values

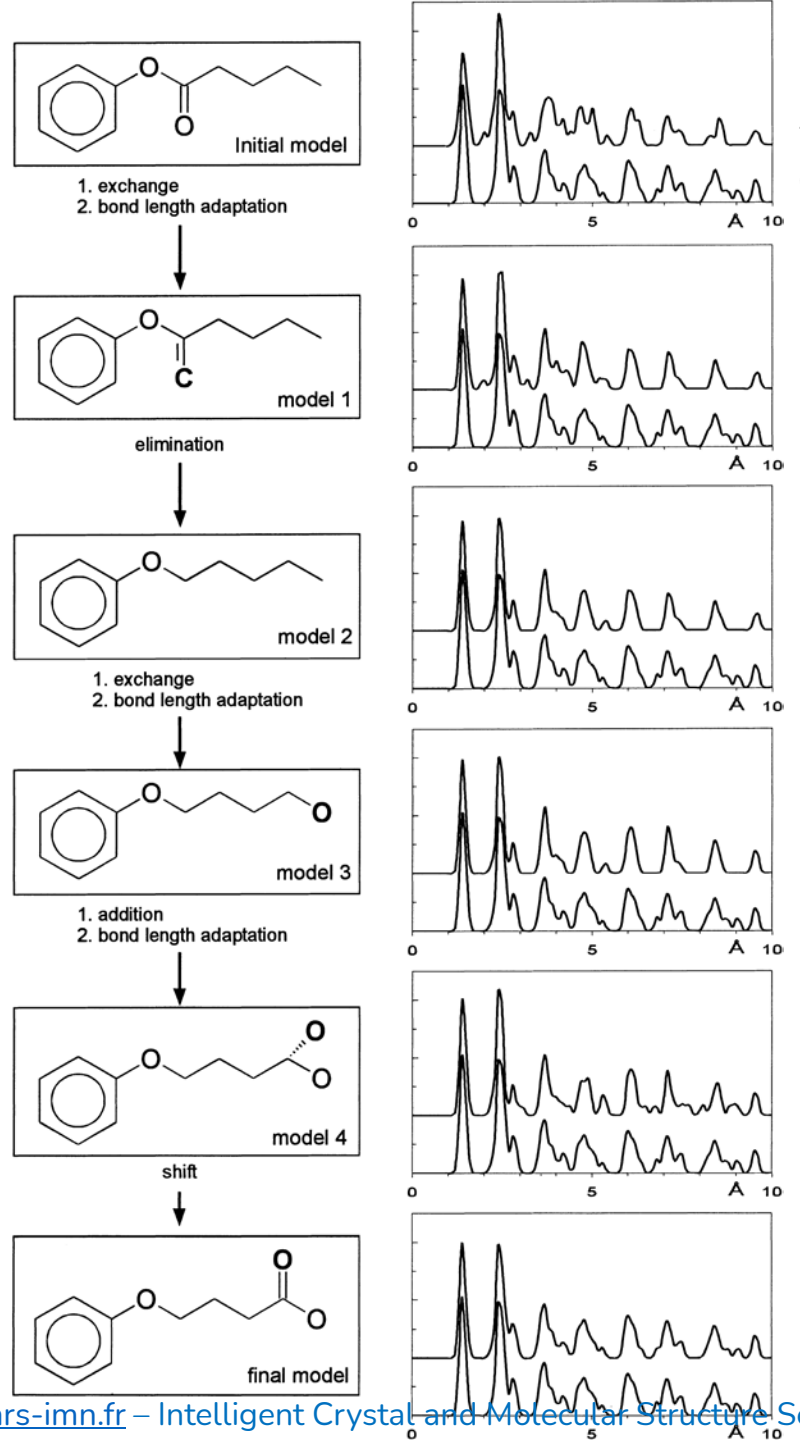
Structure Representation that contains the spatial arrangement of the atoms

Property Representation

M. C. Hemmer, J. Gasteiger, Prediction of three-dimensional molecular structures using information from infrared spectra, Analytica Chimica Acta 420 (2000) 145–154

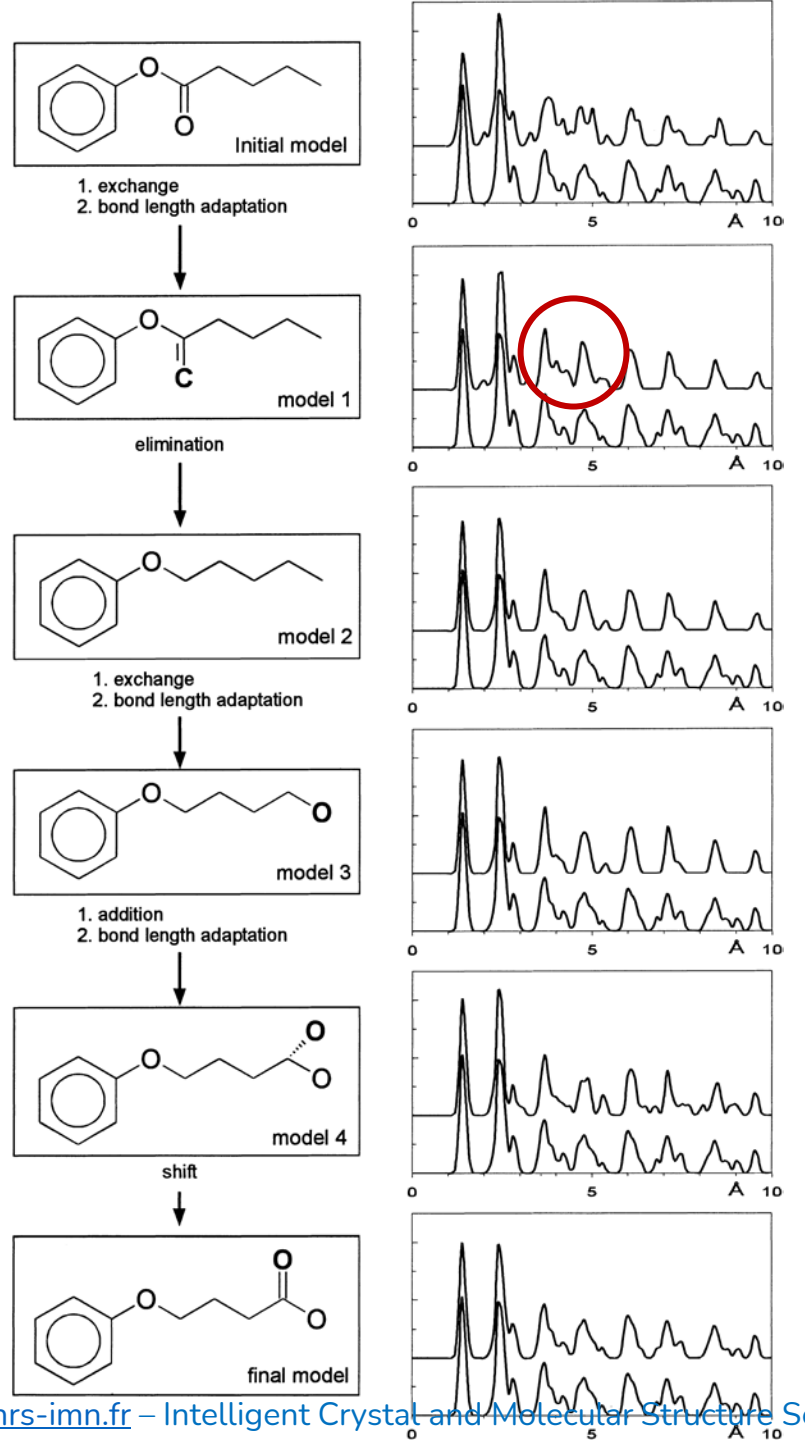
Step 0 Obtain an initial structure

the simulated code is compared with structure codes generated from a database of over 100,000 molecules.
The molecule with the most similar structure code defines the so-called initial model for the 3D structure (m- methyl derivative)
Computation of the IR spectrum of the query model



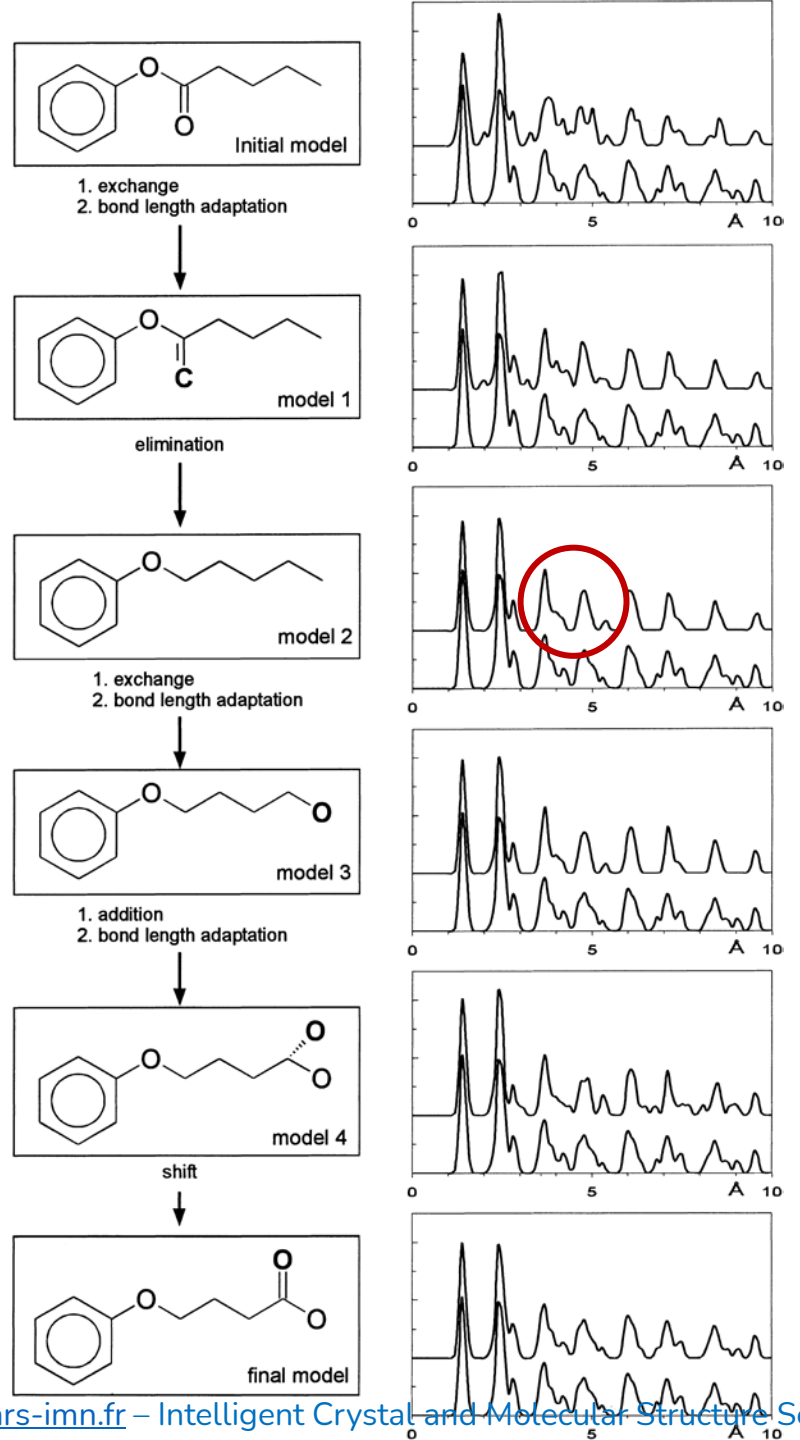
M. C. Hemmer, J. Gasteiger, Prediction of three-dimensional molecular structures using information from infrared spectra, Analytica Chimica Acta 420 (2000) 145–154

Step 1 Tries to adapt the peak heights in the RDF.
O in the carbonyl group will be replaced by C. After correction of the bond length, the number of peaks in the RDF of the model increases because of the shorter CDC bond distance.



M. C. Hemmer, J. Gasteiger, Prediction of three-dimensional molecular structures using information from infrared spectra, Analytica Chimica Acta 420 (2000) 145–154

Step 2 Tries to reduce the number of peaks for the actual model : C atom is removed and superfluous peaks disappear

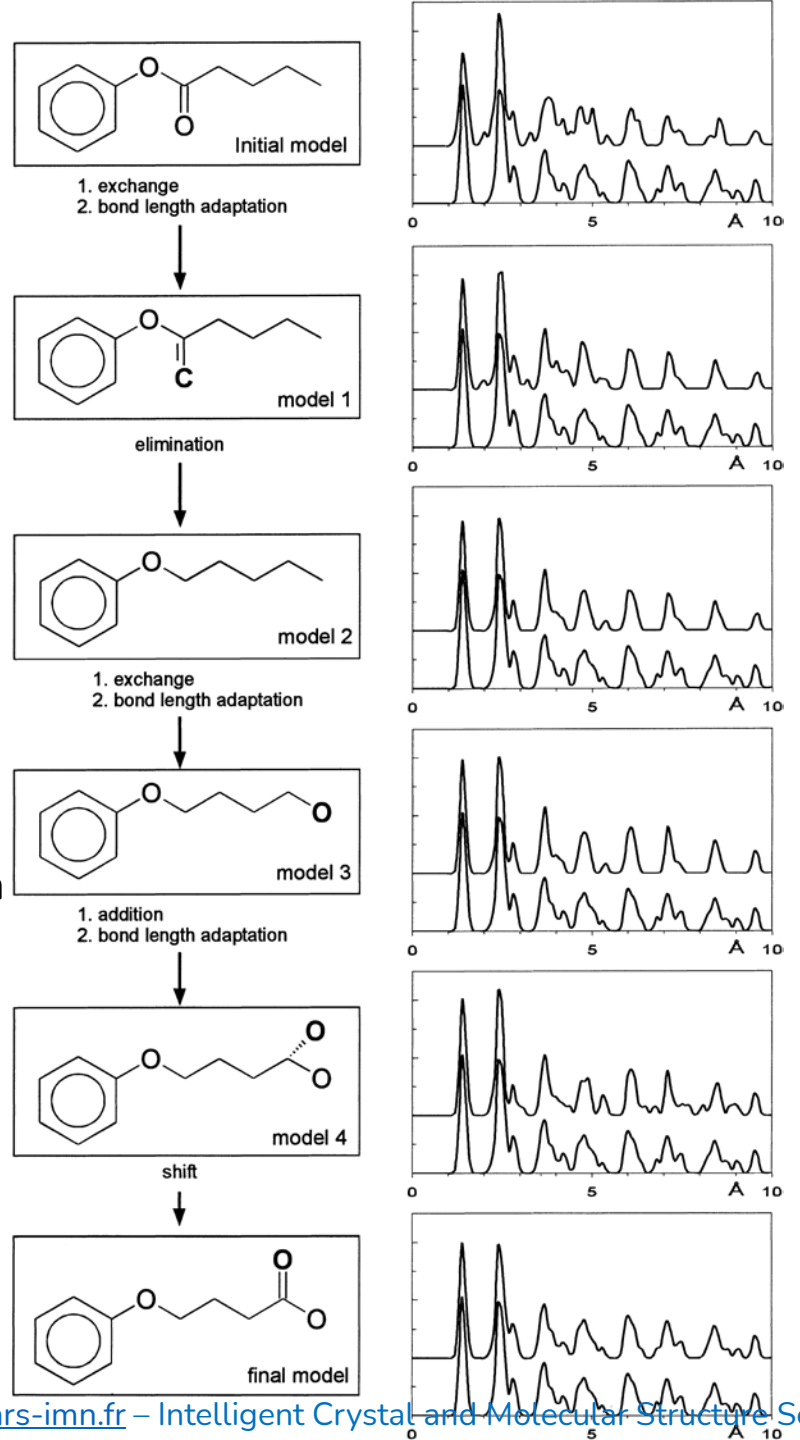


M. C. Hemmer, J. Gasteiger, Prediction of three-dimensional molecular structures using information from infrared spectra, Analytica Chimica Acta 420 (2000) 145–154

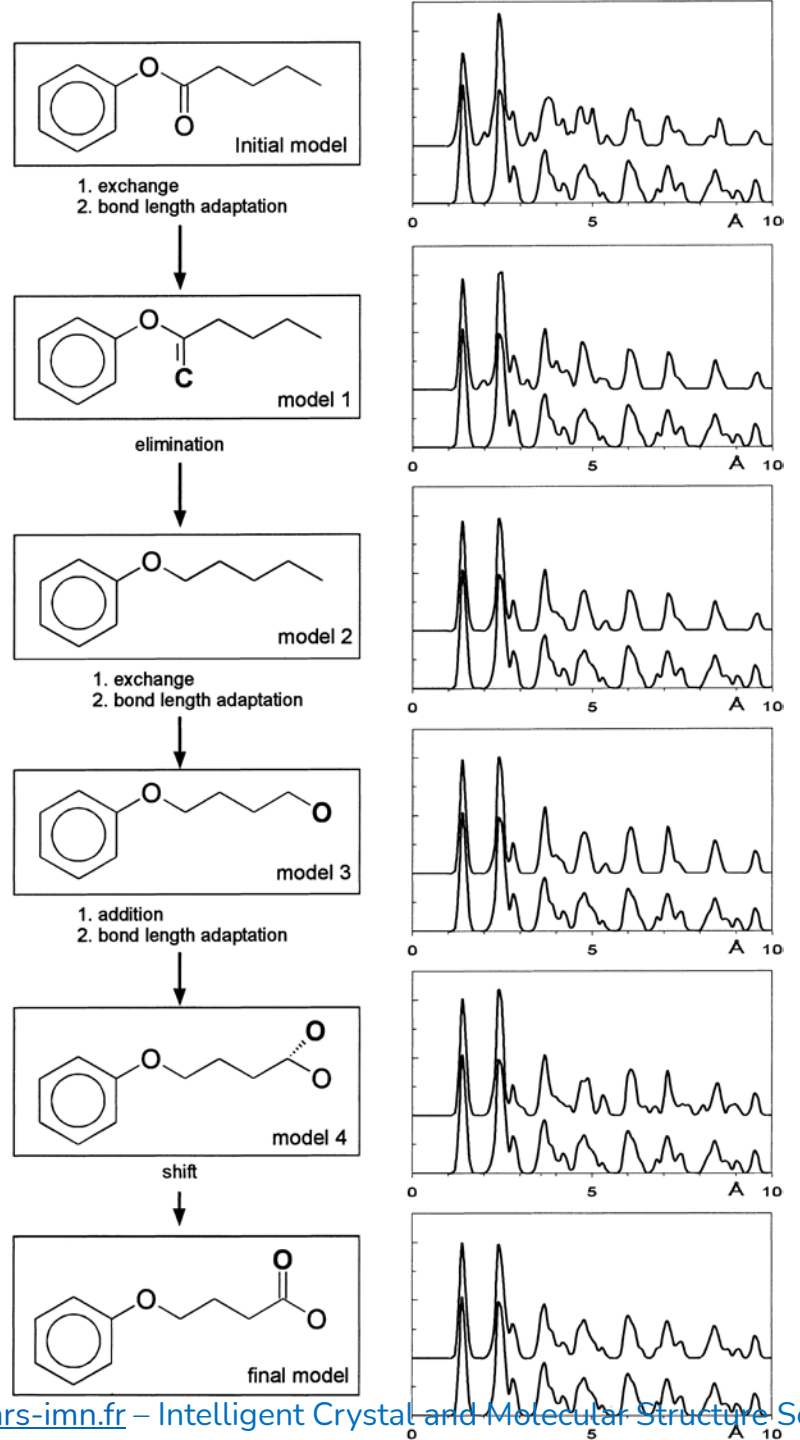
Step 3

Exchanges a terminal C→O

Number of peaks reduced compared with the RDF derived from the CPG network.



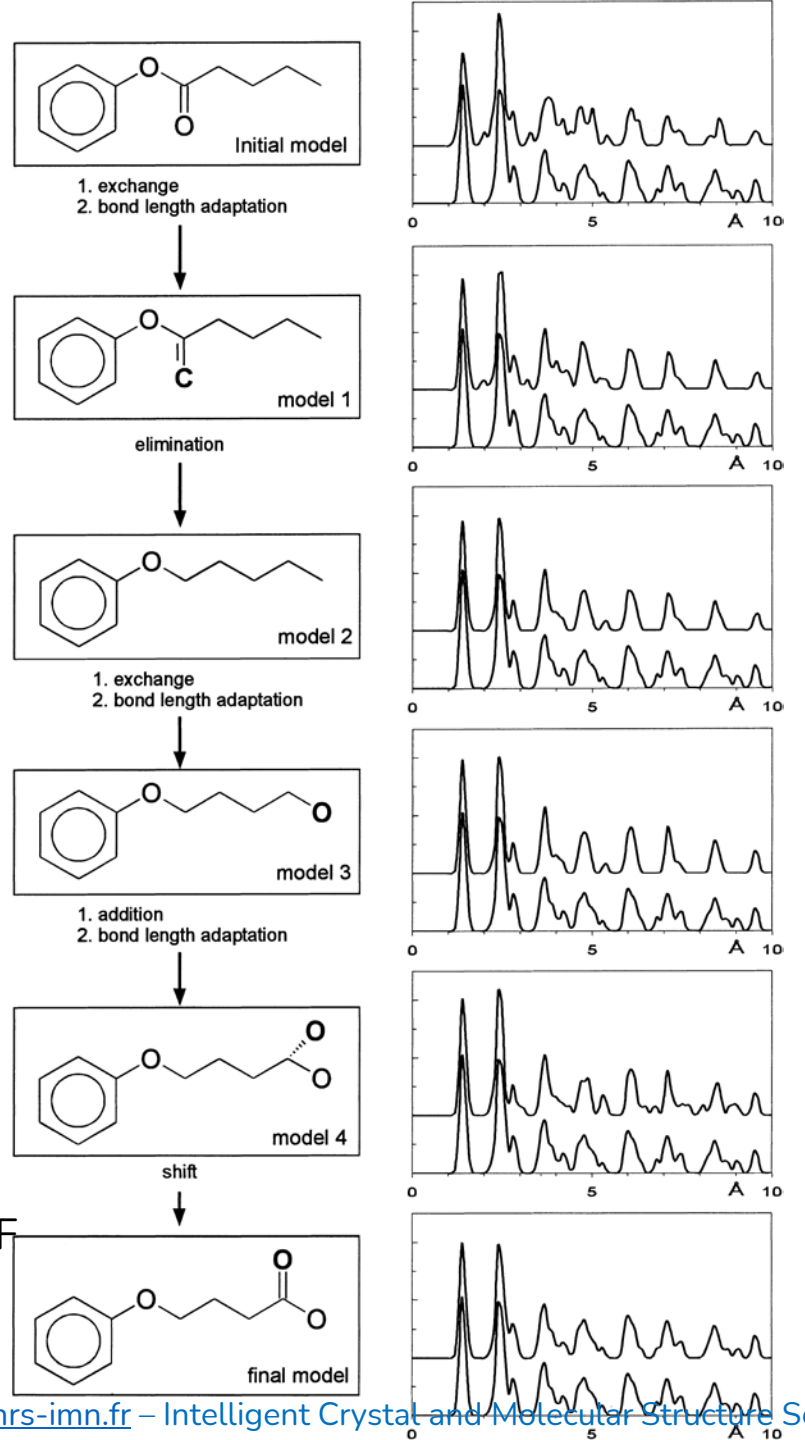
M. C. Hemmer, J. Gasteiger, Prediction of three-dimensional molecular structures using information from infrared spectra, Analytica Chimica Acta 420 (2000) 145–154



Step4

Addition of O at a position formerly held by H

M. C. Hemmer, J. Gasteiger, Prediction of three-dimensional molecular structures using information from infrared spectra, Analytica Chimica Acta 420 (2000) 145–154



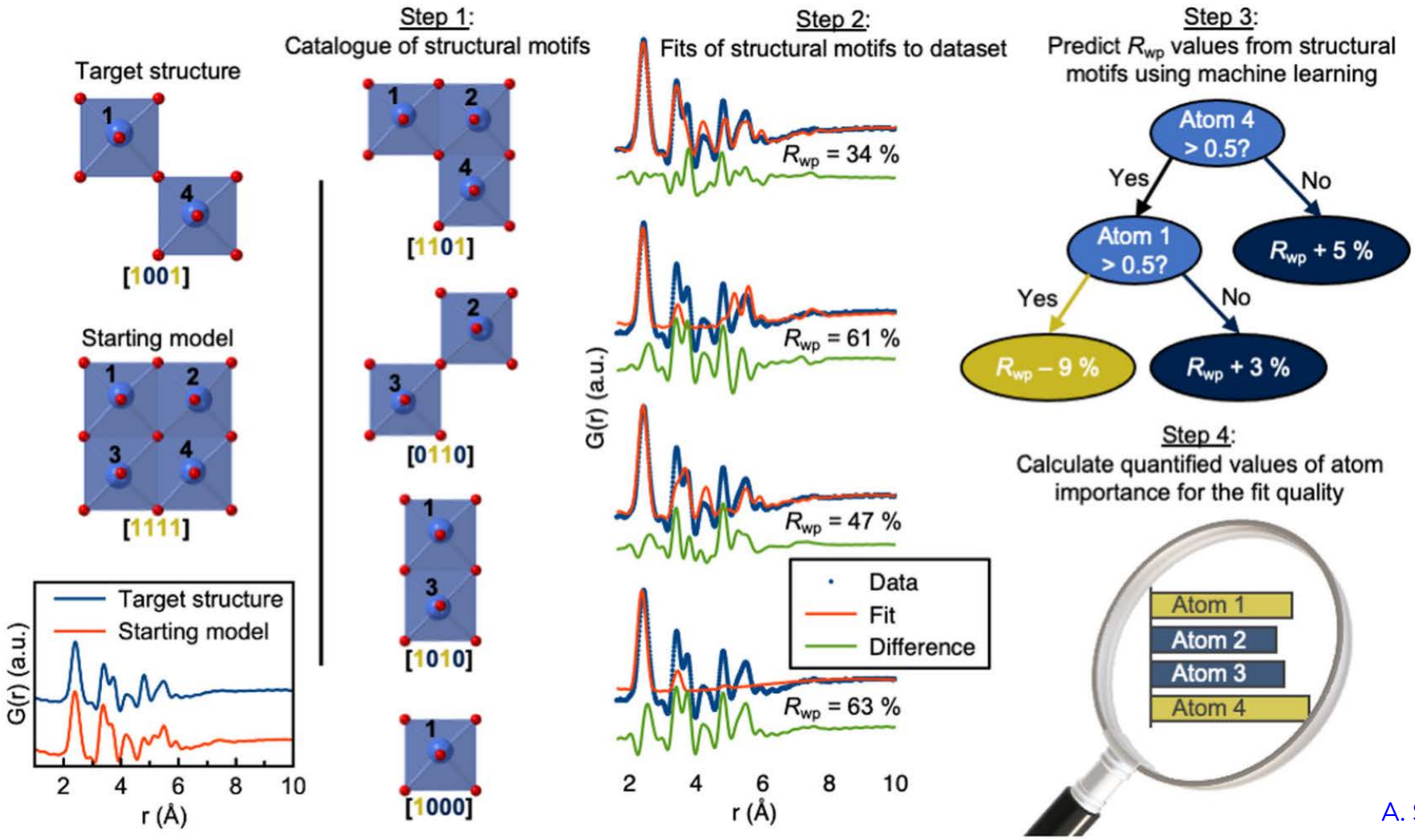
M. C. Hemmer, J. Gasteiger, Prediction of three-dimensional molecular structures using information from infrared spectra, Analytica Chimica Acta 420 (2000) 145–154

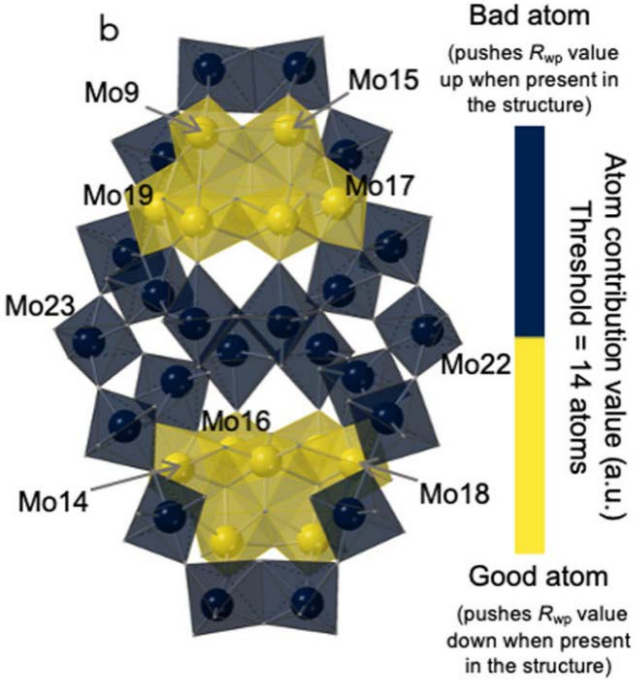
Step5 (final)

Next operation leads to a bond angle of 120 for the position of the added atom and a shortened bond length, resulting in a terminal carboxyl group.

Further modifications do not lead to an improvement of the RDF code.

ML-generated modelling for PDF technique

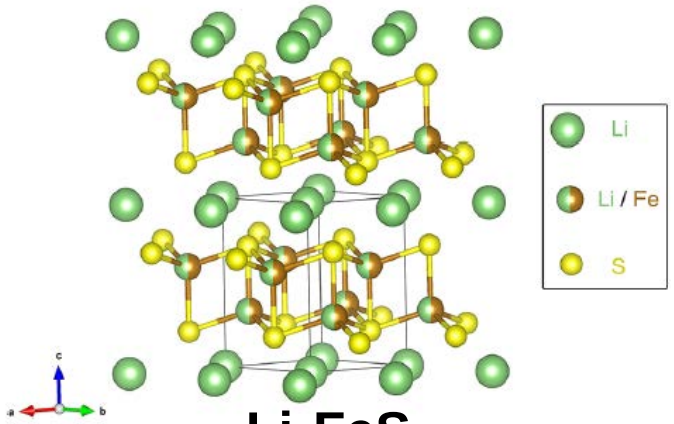




A. S. Anker et al. *NPJ. Comput. Mater.* **2022**, 213.

FeS_2 structure obtained by double delithiation of Li_2FeS_2 ?

R. Brec et al. *J. Power. Sources.* **1989**, *26*, 325



Li_2FeS_2

-Li

Unknown structure since 1989

$Fe^{3+}S^{2-}(S_2)^{2-}_{1/2}$

- ***IR spectroscopy :***
  - d(S–S)= 2.12 Å
- ***Mössbauer and EXAFS :***
  - Fe^{3+}HS ion in **tetrahedral sites**
  - d(Fe–S)=2.28 Å
  - S^{2-}/S_2^{2-} **= 2/1**
- **Lamellar structure** expected due to the stacking of $LiFeS_2$ and Li layers.

2D 2dmatpedia (6351 materials)

Materials Cloud two-dimensional crystals database (MC2D)

2D Materials Encyclopedia

Stanford ENGINEERING Materials Science & Engineering

MAGICS Materials Genome Innovation for Computational Software

UNREAL ENGINE 5

MATDAT



FM *P-3m1* 11-2D

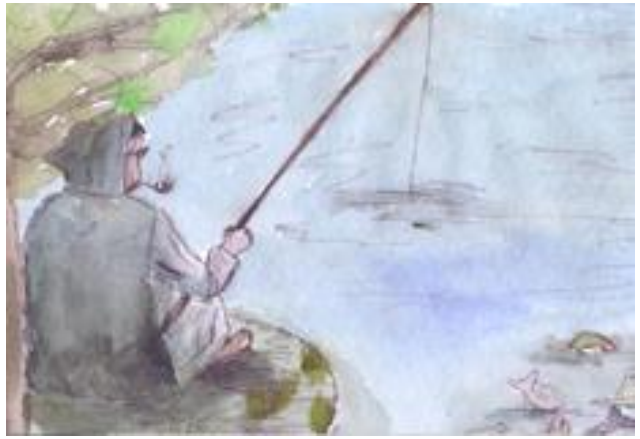
(MoS_2-like)

Invalid + dynamically unstable structure

Remarks

- Sequence of the manipulations depends optional parameters (similarity criterion, targeted physicochemical properties).

- Prediction quality mainly depends on the similarity of the selected initial model (chemical similarity + conformational similarity of skeleton)!!



"For a fisherman, effective data mining (or discriminative AI) means deducing where he's most likely to find fish, but it doesn't guarantee that he'll catch one."

The real limitation is that no discovery is possible with discriminative tools

- Discriminative AI by **substitution** does not always lead to **valid structures** (**structure representation** is key)

https://paulingfile.com/index.php

What if the structure is not known? *de novo* generation to improve database !

①



Discriminative AI



My Goal

Generative *AI*

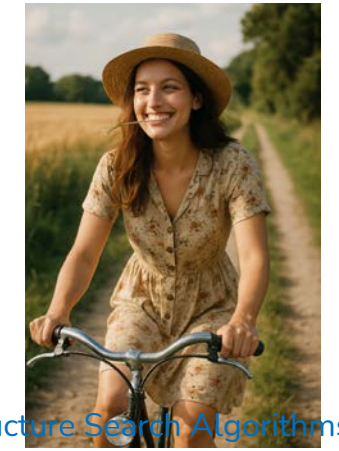


Google Maps

②



Global optimisation



Discriminative AI vs Generative AI model: « art critic » vs « artist »

X = input structure data to learn from, (retrieved from a data set for instance)



discriminate between different kinds of data instances
(tell a dog from a cat)

May learn the difference between "sailboat" or "not sailboat" by just looking for a few tell-tale patterns and **ignore most correlations**

ignores the question of whether a given instance (the currently investigated structure) is **likely**, and just tells you **how likely a label** is to apply to the instance

{X1,*Y1*} → Training (+Validation) → P(Y|X)

{X3} → Test → {Y3}



can generate **new** data instances (generate new photos of animals that look like real animals)

Must capture correlations and subsequent complicated data distributions like "things that look like boats are probably going to appear near things that look like water"

includes the distribution of the data itself, and tells you **how likely a given example** is

{X1,*Y1*} → Training → P(X|Y) or P(X)

{*Z*} → Generation → {X'}

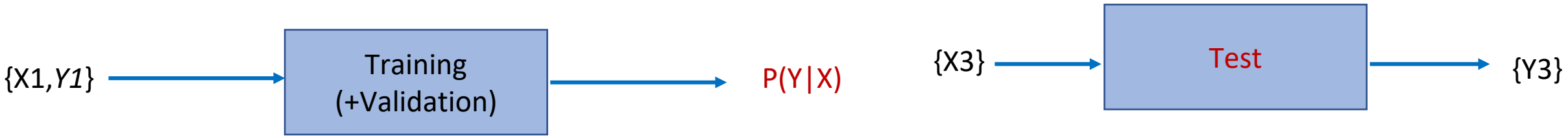
Dall-E 3 - https://www.innosight.com/insight/leading-into-the-age-of-ai/

Discriminative AI vs Generative AI model: « art critic » vs « artist »

X = input structure data to learn from, (retrieved from a data set for instance)

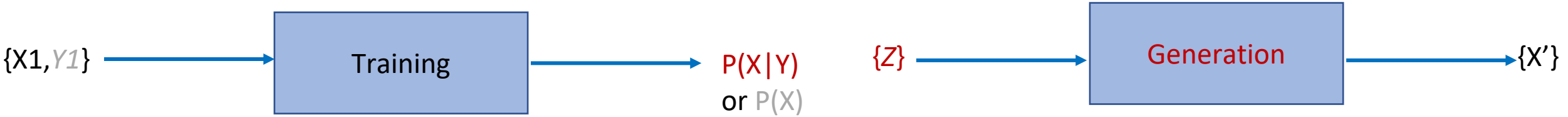


capture the **conditional probability p(Y|X)**
p(dog| this picture)
If the probability is high enough (σ), declare this picture to represent a dog
expand on existing molecules, use a set of hidden rules to link existing molecular fragments





Data + labels (structure with target properties): capture the joint probability p(X, Y) (input data +labels)
No labels : capture p(X)
Learn the underlying structure of the distribution of molecules
Generate molecules based on motifs learned from **large** existing sets of molecules.



Dall-E 3 - https://www.innosight.com/insight/leading-into-the-age-of-ai/

(Input) Data is still - and more than ever - the key…

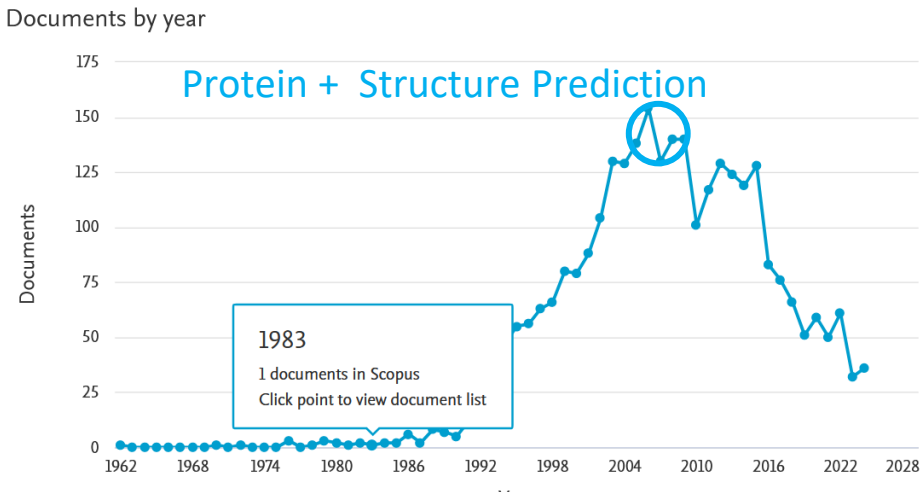
① Effectiveness relies on available **large-scale labeled training data**.

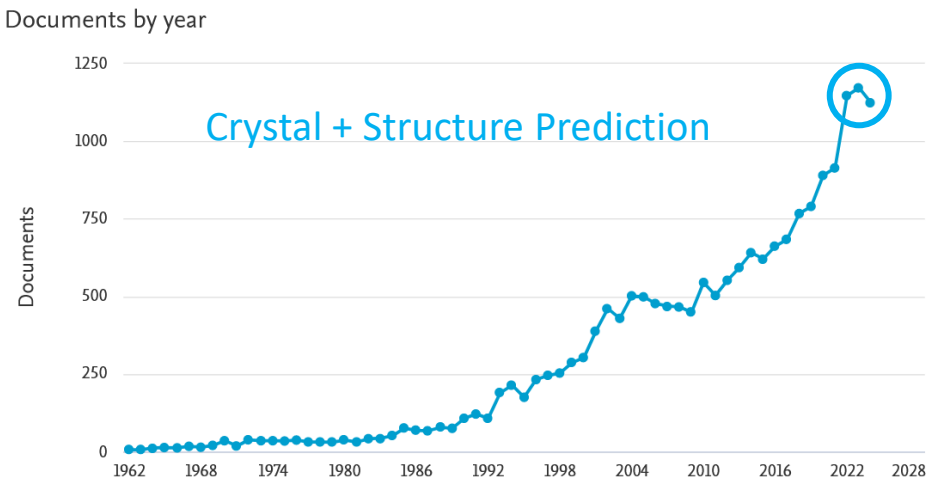
- In image processing: millions of labeled samples
- Gathering **labeled** molecules for structure and (molecular property) through screening experiments is exceptionally costly and labor-intensive data labeling is a resource-intensive and time-consuming task : abundance of unlabeled data and only a few is labeled.
- Solution : semi-supervised learning architecture:
  - First unsupervised **pre-training** of the model (few unsupervised fingerprint methods)
  - Second **fine-tuning** training with labeled data.

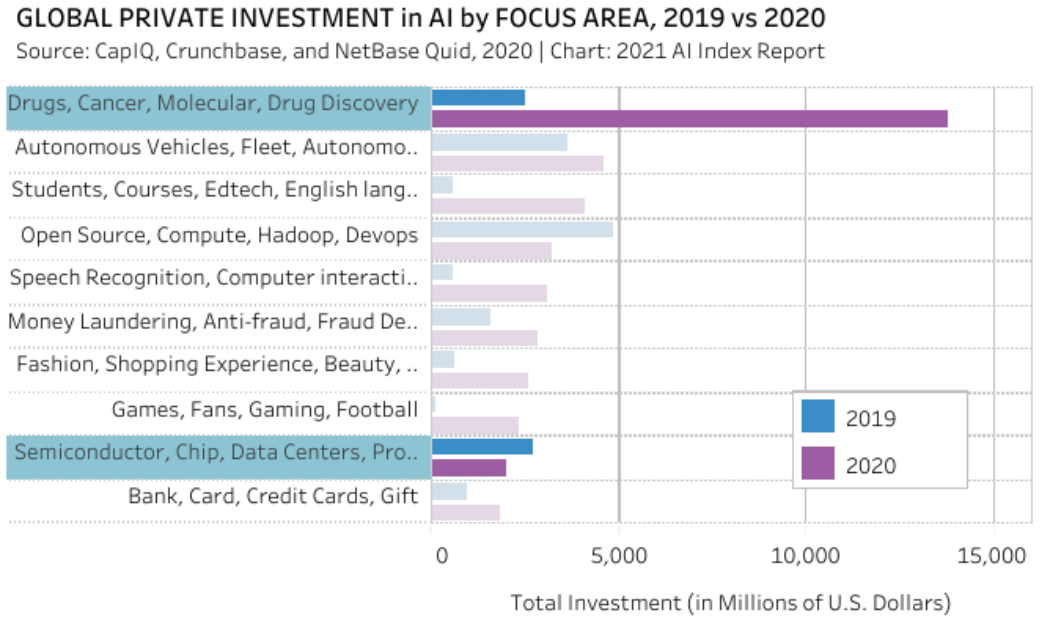
(Input) Data is still - and more than ever - the key…

① Effectiveness relies on available **large-scale labeled training data**.

② Crystal GM models are late vs molecular GM models







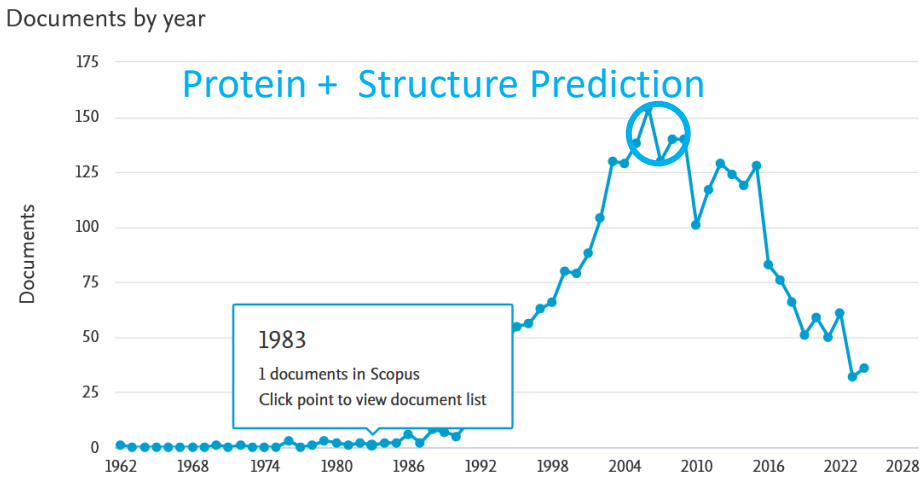
Artificial Intelligence Index

Stanford University Human-Centered Artificial Intelligence

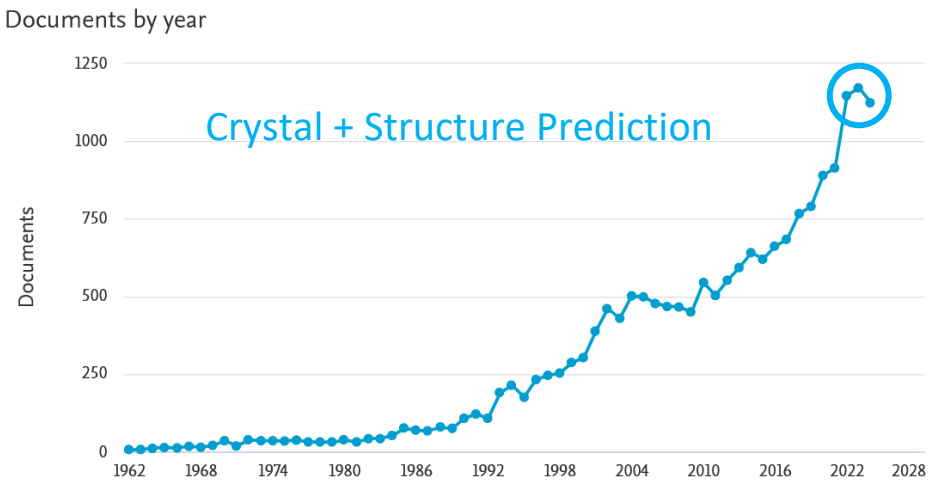
(Input) Data is still - and more than ever - the key…

① Effectiveness relies on available **large-scale labeled training data**.

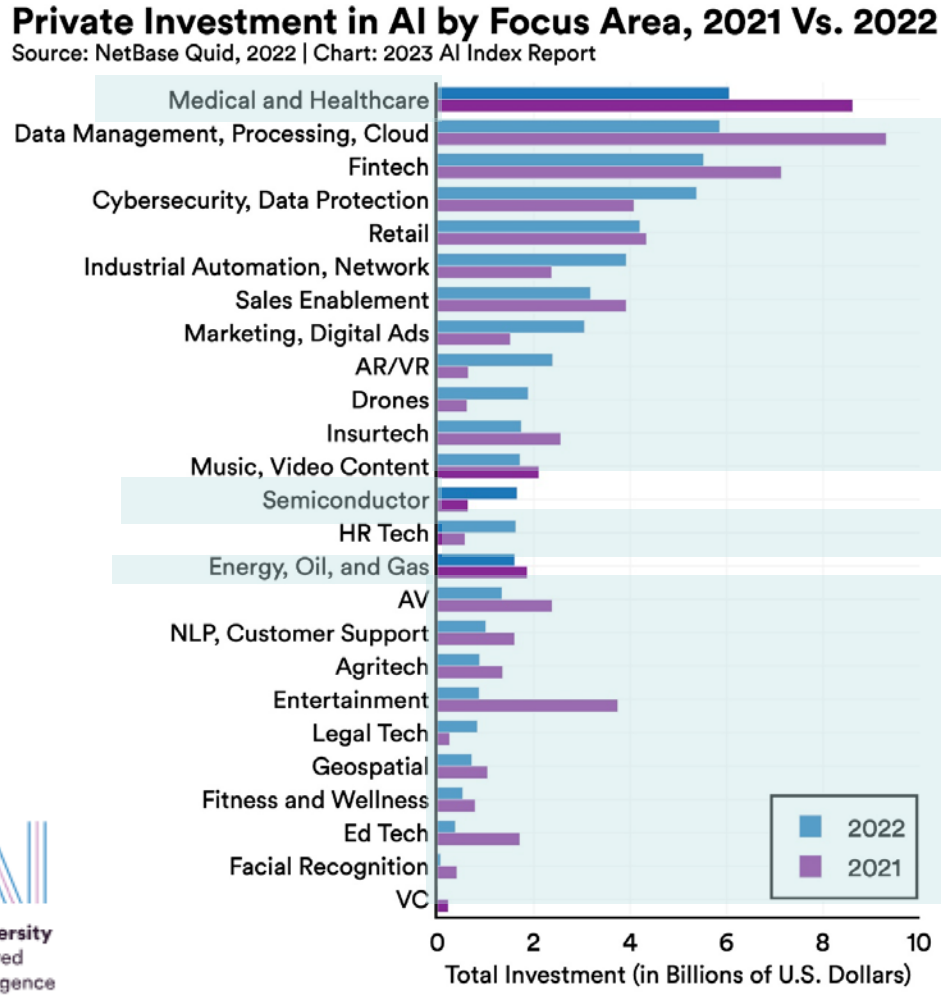
② Crystal GM models are late vs molecular GM models



AlphaFold



GNoME



Artificial Intelligence Index

Stanford University Human-Centered Artificial Intelligence

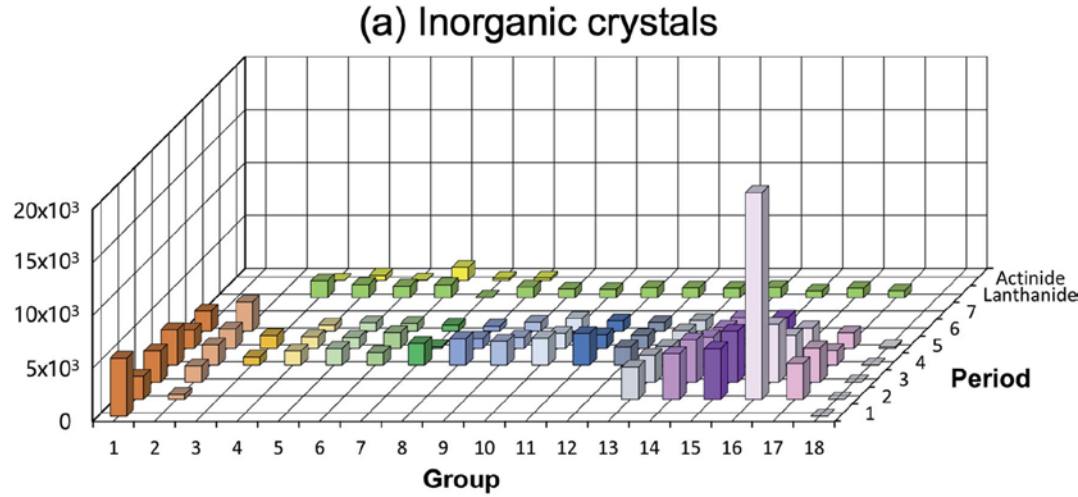
(Input) Data is still - and more than ever - the key…

① Effectiveness relies on available **large-scale labeled training data**.

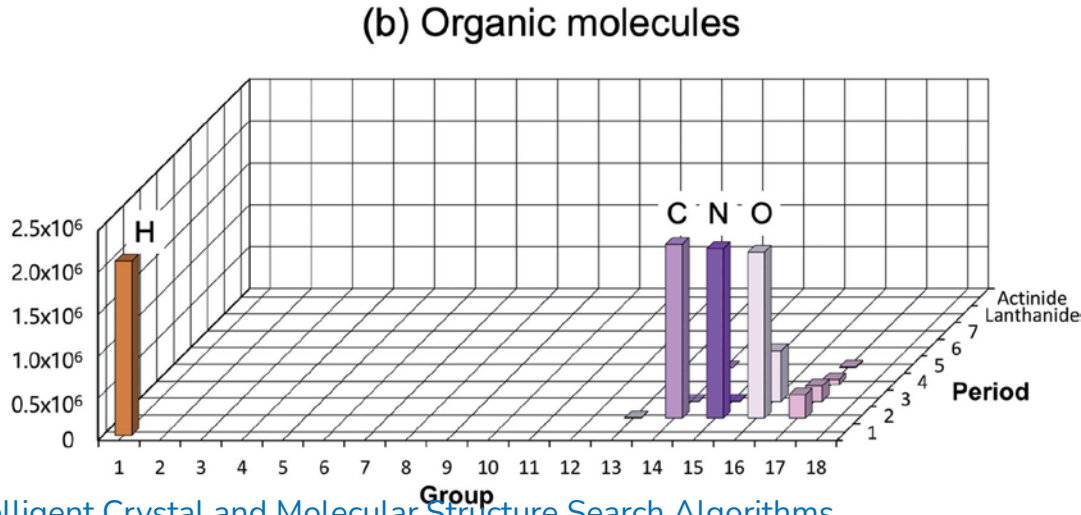
② Crystal GM models are late vs molecular GM models

- Partly because of Low structural diversity (data) per element on inorganic crystal structures compared to molecular chemical space

Crystal database (MP). Experimentally reported inorganic materials (# of data = 48 567).
High elemental diversity :
cover most elements
Low structural diversity :
Number of data per element is sparsely populated



Molecular database (subset of ZINC)
Organic molecules (# of data = 2 077 407).
Low elemental diversity:
cover very limited elements
High structural diversity :
They are densely populated



Chem. Sci., 2020, **11**, 4871–4881

(Input) Data is still - and more than ever - the key…

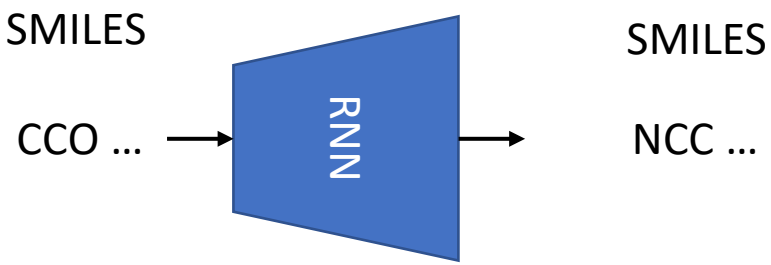
① Effectiveness relies on available **large-scale labeled training data**.

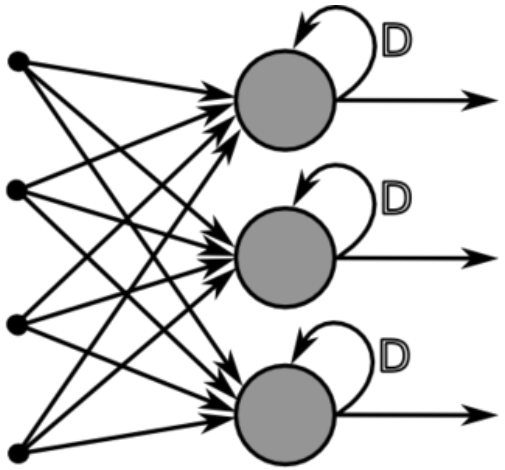
② Crystal GM models are late vs molecular GM models

- Partly because of Low structural diversity (data) per element on inorganic crystal structures compared to molecular chemical space
- **NN architecture** strongly depends on the **data representation**

Representation = digital encoding used for each datum that serves as input for training the deep learning model
Ex: How to represent a spectrum?

As ***vectors*** (intensities, wavenumbers) → ***Sequence models: RNN*** (recurrent neural networks)





(Input) Data is still - and more than ever - the key…

① Effectiveness relies on available **large-scale labeled training data**.

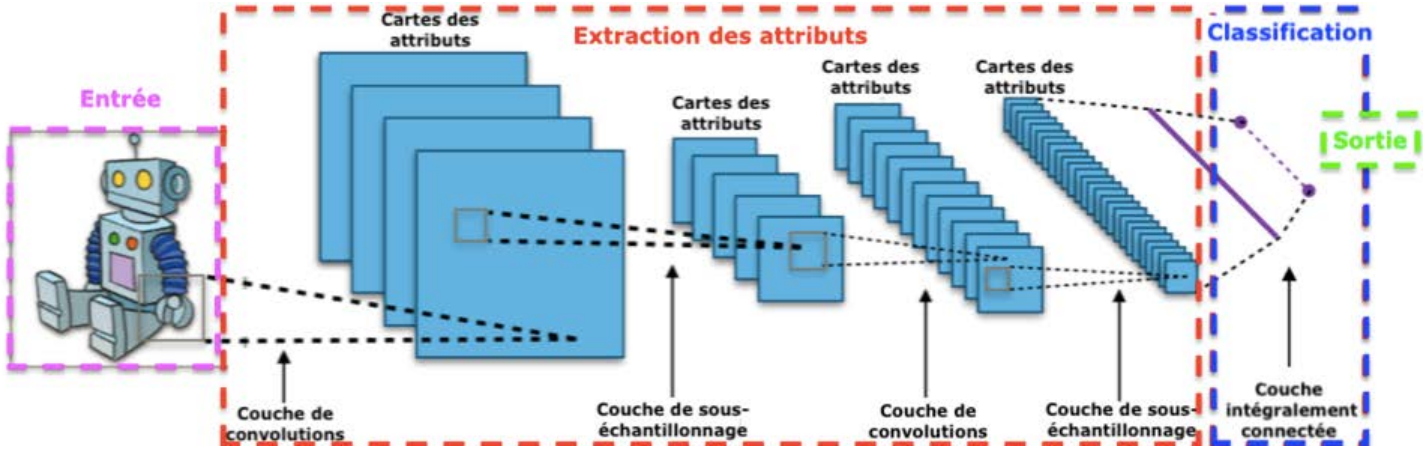
② Crystal GM models are late vs molecular GM models

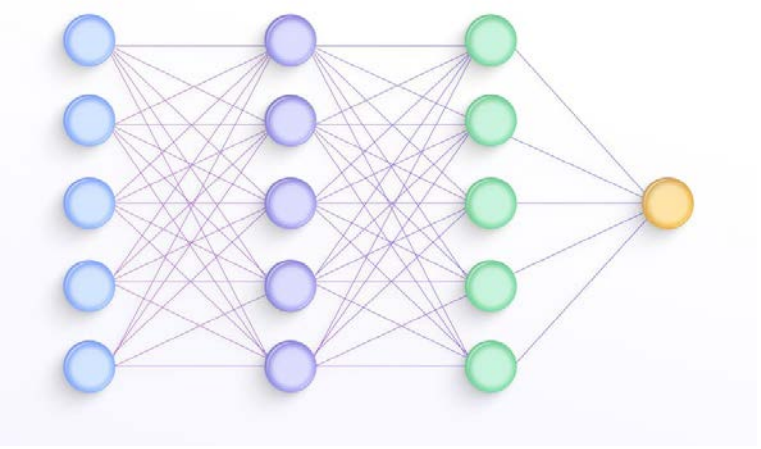
- Partly because of Low structural diversity (data) per element on inorganic crystal structures compared to molecular chemical space
- **NN architecture** strongly depends on the **data representation,**

finding a representation for a molecule for a computer was achieved much before than finding one for a crystal

Representation = digital encoding used for each datum that serves as input for training the deep learning model
Ex: How to represent a spectrum?

As ***2D image*** → ***CNN (convolution-based neural network)***





https://fr.wikipedia.org

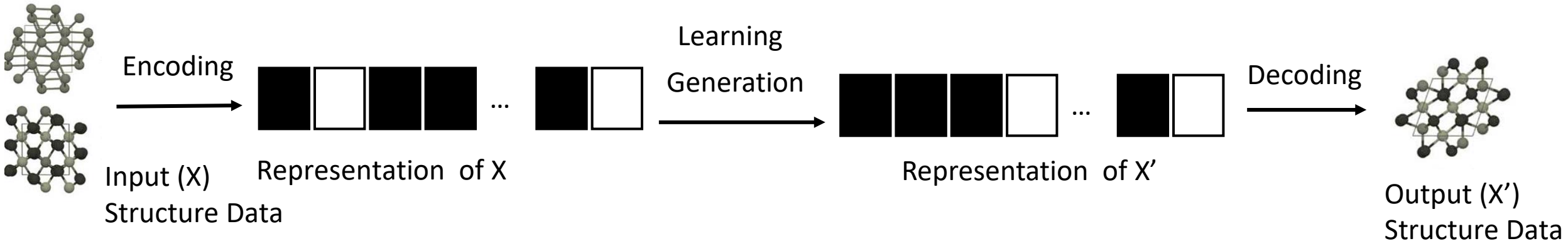
Structure representation



logos-world.net

How to write molecules and crystals (X) such that computers can easily process them and us recognize them later ?
Possible to implement directly with nuclear coordinates
but coordinates are **not invariant to molecular translation, rotation, and permutation of atomic indexing or crystal symmetry invariance** !
Can be useful in data augmentation

Featurization : creating an appropriate **representation scheme, i.e. that captures essential structural information** about each molecule/crystal



Structure representation



logos-world.net

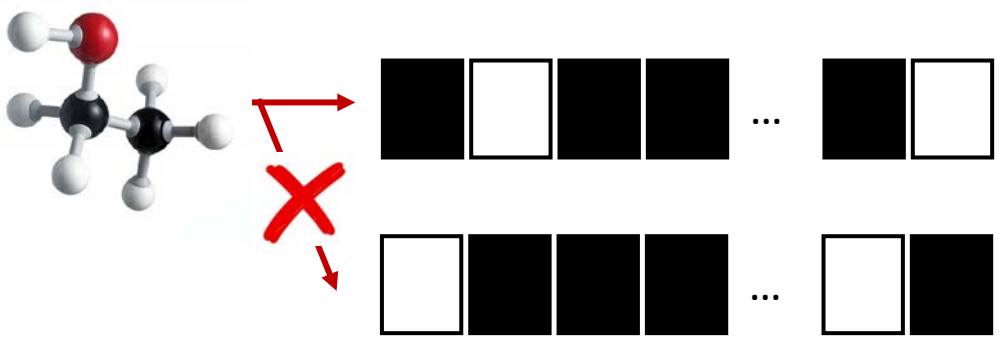
How to write molecules and crystals (X) such that computers can easily process them and us recognize them later ?
① But how to recognize the same molecule/crystal (or know they differ) ?

Need **metrics** adapted to a representation, that must be **unique**

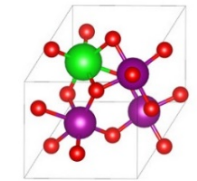
Each structure is associated with a single representation
(easy for crystals, many for molecules are not unique)

SMILES string representation

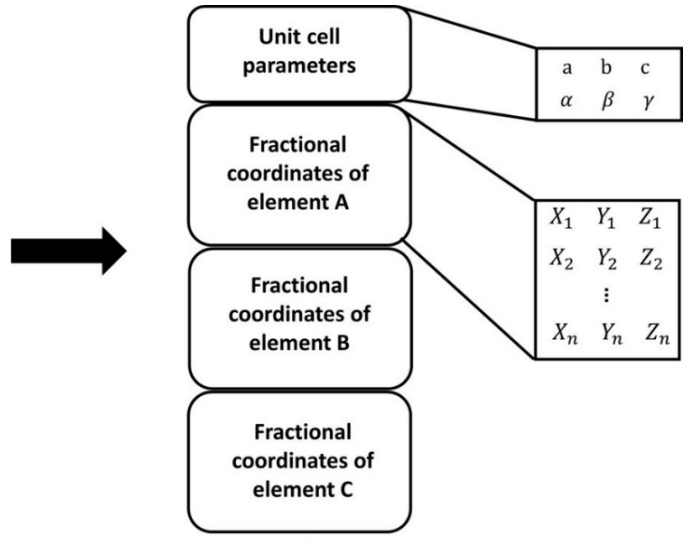
Ethanol → CCO,OCC, C(C)O, C(O)C



Matrix representation



$A_nB_mC_l$ ...



Structure representation



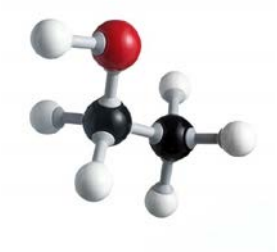
logos-world.net

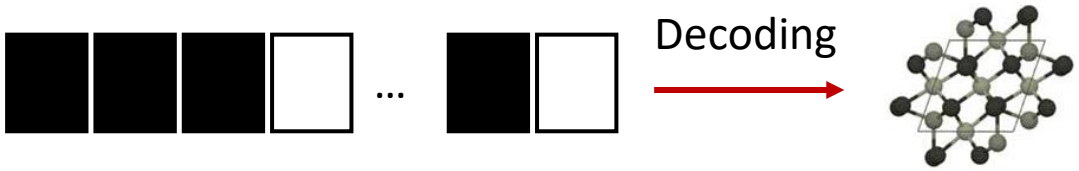
How to write molecules and crystals (X) such that computers can easily process them and us recognize them later ?
② How to ensure the generated output will lead to the molecule structure ?

Need **invertible** representation (minimum/optimized loss embedding)
Each representation is associated with a single structure (easy for molecules, complex for periodic crystal structures)

SMILES string representation

CCO → Ethanol





Output (X')
Structure Data

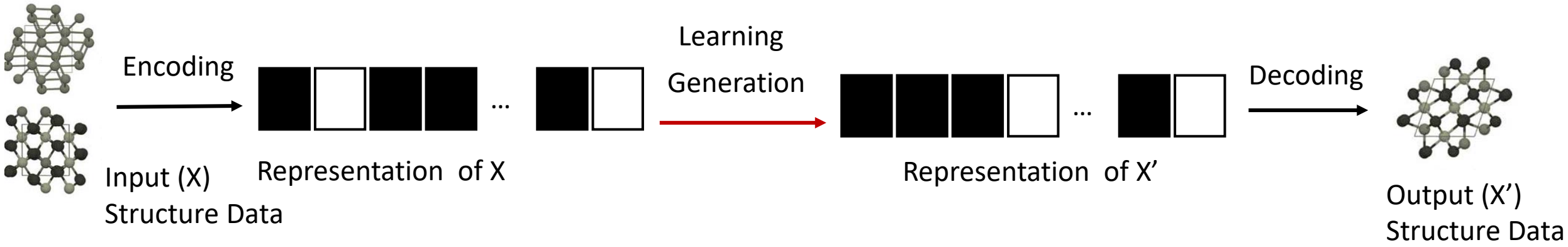
Structure representation



logos-world.net

How to write molecules and crystals (X) such that computers can easily process them and us recognize them later ?
③ How to ensure the generated output is a **valid** molecule structure ?

Need to encrypt relevant information (connectivity range? Hot topic for crystals)
and representation must be syntax error insensitive (if possible, slightly human interpretable)
But also **compact** as possible (remove invariances and degrees of freedom)



Structure representations

C. Elton et al, Mol. Syst. Des. Eng. 2019

1. **3D representations**
2. *Fingerprints*
3. **String-based (text)** representation : SMILES, InChI, SELFIES : Molecule, ~~Crystal~~
4. **Graph**-based representation : GNN
5. *Latent representations : Auto-encoders*
6. So many others (fingerprints, see Mads-Peter .V Christiansen's talk, image-based, tensor fields, **wavelet transform invariants** (see M. Hirn et al, Multiscale Model. Simul., 2017, 15, 827–863),**solid harmonic wavelet scattering transforms** (see M. Eickenber et al, Advances in Neural Information Processing Systems 30, 2017, pp. 6540–6549.), **tensor field networks** (see N. Thomas et al, Tensor field networks: Rotation- and translation-equivariant neural networks for 3D point clouds, 2018)



Not all of them are one-to-one (bijections) *i.e.* they may lose structural information (embedding) :
not all dedicated to structure prediction but property prediction !

Not all of them are insensitive to errors from generative NN !

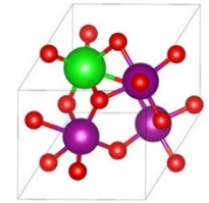
Structure representations: 3D representations

- **Matrix representation**

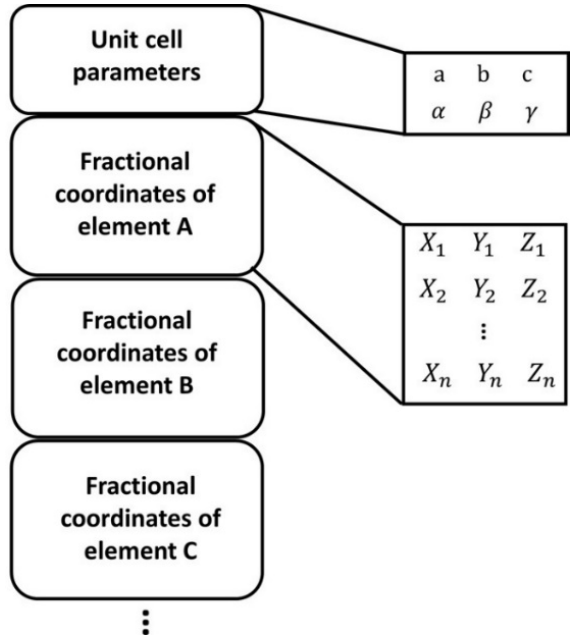
Cartesian coordinates or distance matrices

From files such as XYZ, PDB, CIF

- Represents the actual geometry of a molecule (no loss)
- Useful for simulating or predicting physical properties (*e.g.* energy, force)
- Requires geometric relaxation
- **Direct generation of valid 3D structures** is difficult



$A_nB_mC_l$ ...



Matrix representation

S. Kim, J. Noh, GH Gu, A. Aspuru-Guzik , Y. Jung: *ACS Central Sci*, 6:1412-1420 (2020)

Structure representations: 3D representations

➢ **Voxelization or 3D grids**

Divides space into voxels where each voxel encodes the presence of atoms (nuclear charge, atom type)

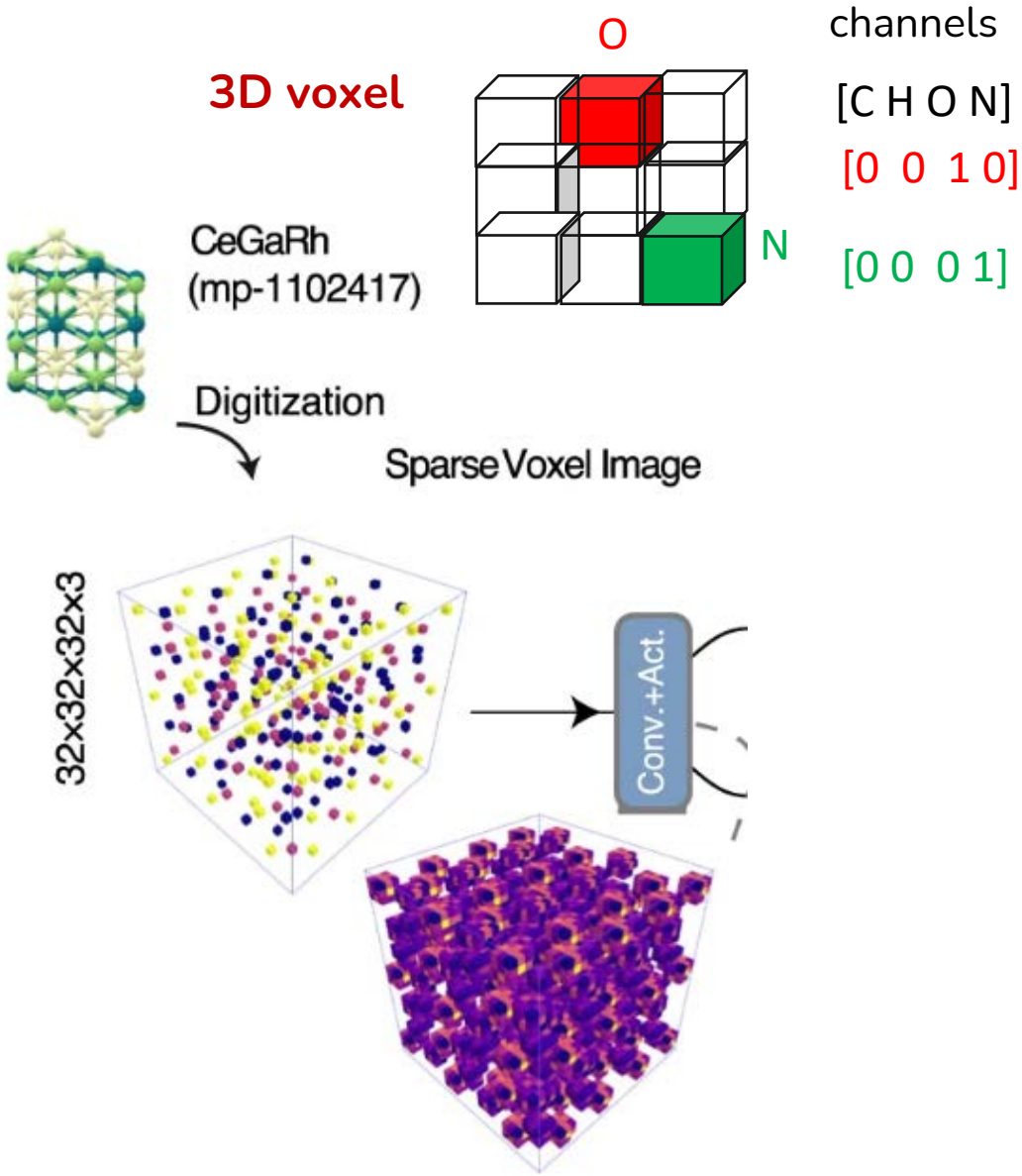
No loss

memory-hungry, limited resolution

Atom type specified by a **channel** (one-hot vector of dimension n)
→ lots of sparse data in high-D space → very large training sets

Solutions : spatial smoothing…

Needs alignment+PCA to ensure invariance



A Davariashtiyani et al, Nat Com Mat **4**, 105 (2023)
Amidi et al., PeerJ, 2018, 6, e4750
D. Kuzminykh et al, Mol. Pharmaceutics, 2018, 4378–4385.
M. Skalic et al, J. Chem. Inf. Model., 2019, 59(3), 1205–1214 (CNN-based AE).

Structure representations – string-based representations

SMILES, InChI, SELFIES : Molecule, ~~Crystal~~

Simplified Molecular Input Line Entry System (1980) : most popular

Inline notation base on a symbolic language using short ASCII **strings** (NIST Chemistry WebBook)
A string is a **sequence** of word ('**token**' = word elementary unit)
"tokens" are the different characters which are used to specify atom types, bond types, parentheses, and the start and stop points of rings. Allow to describe atoms, connectivity, bond order and chirality (but bond length, conformation… lost)

Example :
Ethanol is represented for instance by the string CCO, C and O are tokens.

M. Krenn et al, SELFIES and the future of molecular string representations, Patterns 3(10), 2022, 100588

Structure representations – string-based representations

Easy to manipulate, facilitates chemical data interoperability :
REAL database : #170 million compounds (2017) → over 5.5 billion compounds (2022): ZINC20 virtual screening database
No starting conformation
Compatible with Natural Language Models (RNN, Transformer, etc.)
Not one-to-one : Éthanol → CCO,OCC, C(C)O, C(O)C
Syntax error sensitive : most combinations of symbols lead to invalid results with no valid chemical interpretation (non closed cycles, C1=CC=CC):
SMILES based methods often struggle to achieve a high percentage of valid SMILES.

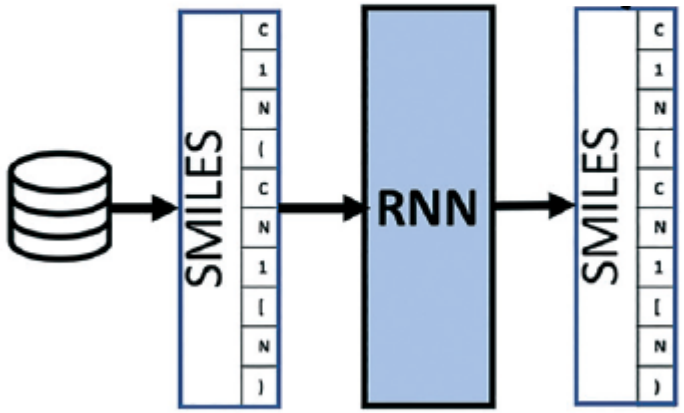
Solution :

- canonical SMILES (MolVS, RDKiT) to standardize SMILES
- **SELFIES Self-referencing Embedded Strings (2020) : avoid syntax error sensitivity:** all SELFIES sequence corresponds to a correct molecule (Chemical rules included within the grammar by using tokens embedding atom valency **Ethanol [CH3][CH2][OH]**
  - Several SELFIES can represent the same molecule : [CH3][CH2][OH], [OH][CH2][CH3] (possible to combine with unique SMILES)
  - Less compact, less human interpretable
- SMARTS (SMILES arbitrary target specification) to specify substructural patterns in molecules(wildcards)

M. Krenn et al, SELFIES and the future of molecular string representations, Patterns 3(10), 2022, 100588

String-based representations well-suited with recurrent neural networks (RNN)

Neural network sequence models **used to be the most popular models**
(most VAE and GAN implementations include an RNN for sequence generation)



Training : learn to predict the next character of a SMILES

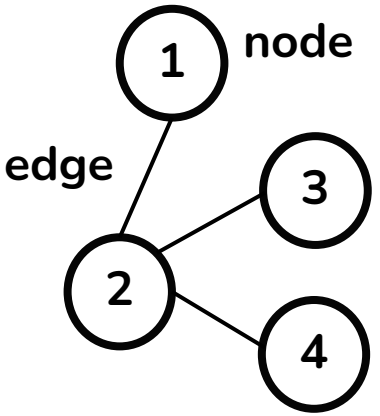
Generation : produces a complete SMILES, character by character

Graph-based representations

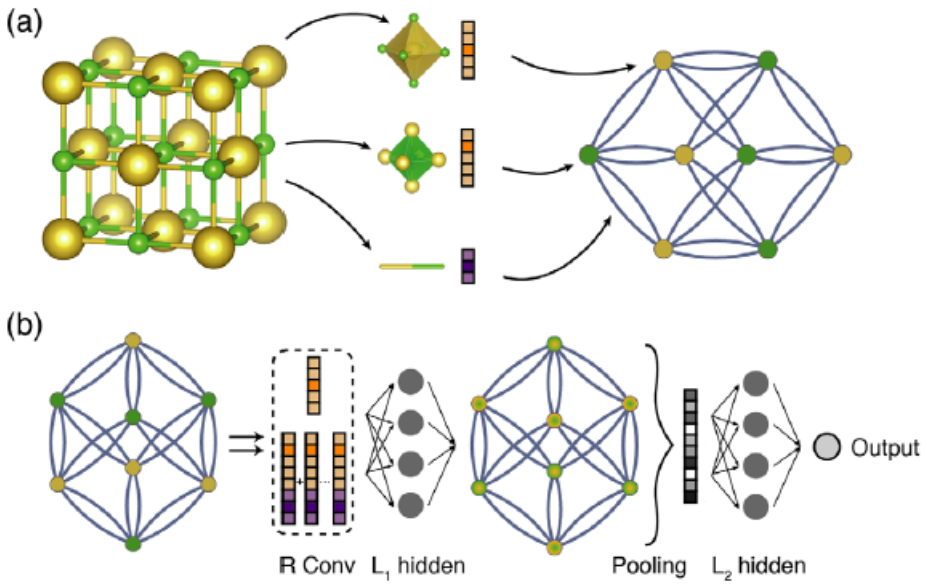
Very popular in recent deep learning models

Non oriented graph with a set of nodes (local environment of each atom) and a set of edges (encode *relationships* (bonds or distances))

keeps the chemical connectivity and structural validity



Embedding stores the information on the node



Gong et al arXiv:2208.05039 2022

Xie and Grossman, PRL 2018, 14, 145301

Crystal graph

Captures *average* local atomic environments
-Short periodicity ok
-Long periodicity not learnable : length and cluster size Matters

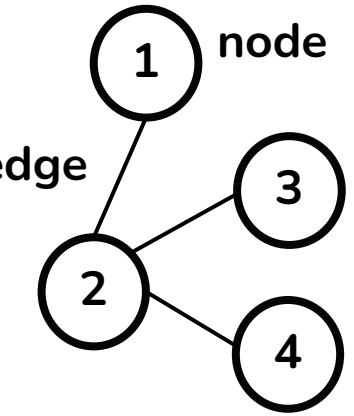
If not stored information about bond lengths is lost

Graph-based representations

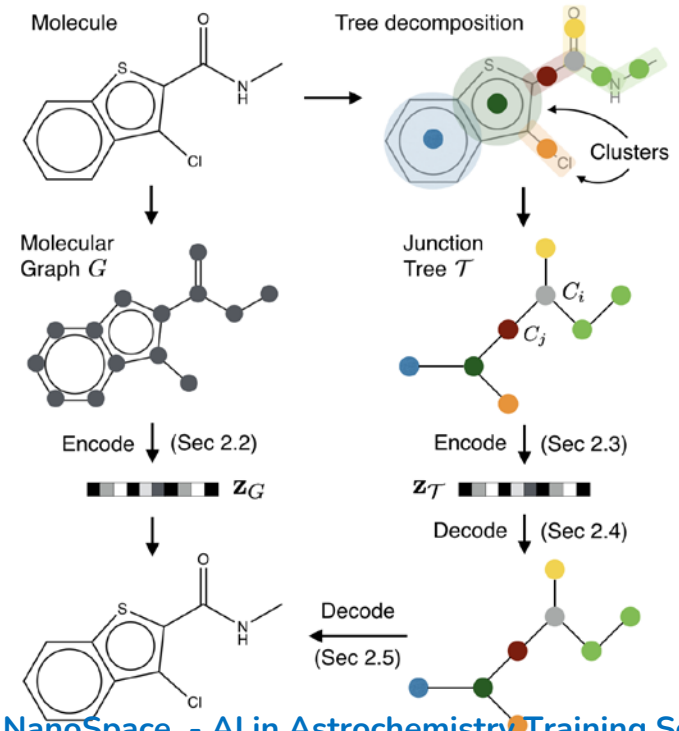
Very popular in recent deep learning models

Non oriented graph with a set of nodes (local environment of each atom) and a set of edges (encode *relationships* (bonds or distances))

keeps the chemical connectivity and structural validity



Embedding stores the information on the node



Molecular graph

Disadvantage : information about bond lengths and 3Dconformation is lost. (molecule 3D conformations may be important).

Decoding : molecular graphs can be embedded in 3D using distance geometry methods (see OpenBabel toolkit)+ energy minimization

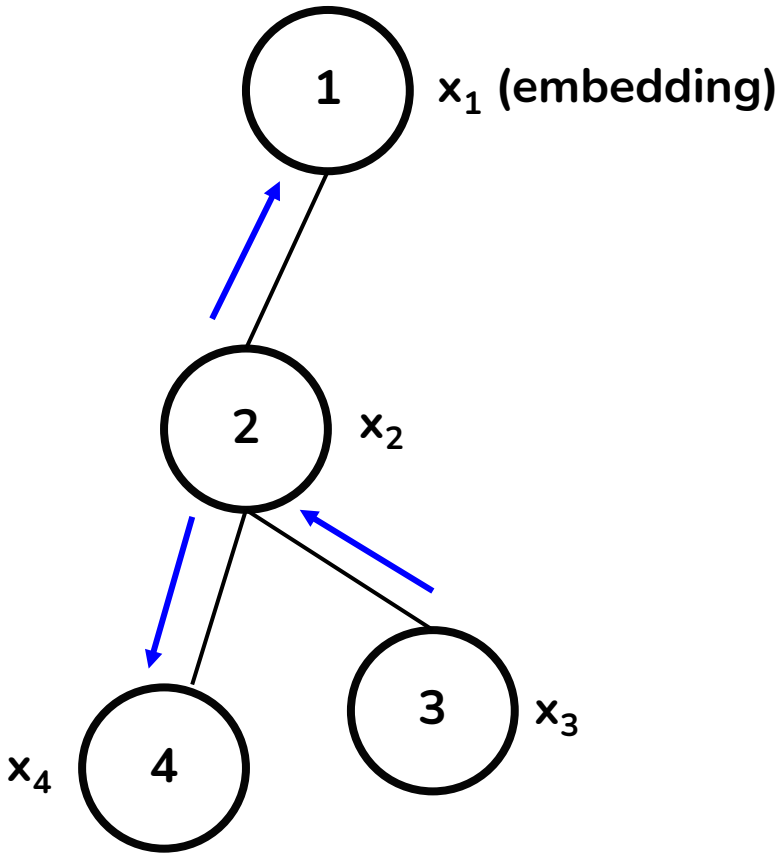
W. Jin, R Barzilay, T. Jaakkola, Junction Tree Variational Autoencoder for Molecular Graph Generation, Proceedings of the 35 th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018

Structure representations – (Message Passing) GNN (Graph Neural Networks)

Message Passing Graph Neural Network (MPGNN),
Information propagation between neighbors on the graph
Learn from the molecule/crystal topology

Graphs used in many architectures:
Graph Convolutional Neural Networks (GCNN) : crystal graph (when deep reinforcement learning to add new atoms and bonds in each action step from a list of predefined possible actions with rules ensure that basic atom valence is satisfied: **100% generated molecules chemically valid**)
Graph Attention Neural Networks (GAT)
Graph Variational AutoEncoder (GraphVAE)
Graph Generative Adversarial Network (GraphGAN)

MEGNet integrates internal/global properties (P,T) with atomic graphs
ALIGNN : advanced appproaches accounting for tensor aligment and edge angles. Powerful: Explicit, interpretable structure, Allows molecular topology to be respected
More complex to generate, Post-generation chemical validation often required, less understandable



Comparative summary

Representation	Ok for GM	Valid	Interpretable	Invertible to 3D Structure	AI architecture
SMILES	**yes**	**no**	quite	**no**	Sequence RNN
SELFIES	**yes**	**yes**	less	**no**	Sequence RNN
3D coordinates	Yes but hard	no	**yes**	**yes**	CNN
3D Voxels	Yes but rare	no	No	yes	CNN
Fingerprint	Yes but avoid	**yes**	no	**no**	Any, RNN
Latent space	yes	**yes**	**no**	sometimes	VAE
Graphs	yes	quite	yes	2D topology	GNN

Generative models for structure search

Unsupervised learning model to learn probability distributions over a large set of molecules to later generate new molecules by sampling from these distributions

GeoDiff
CGAN
condVAE
MolGPT
EDM
JTVAE
GDM
ORGAN
CrystalGAN…



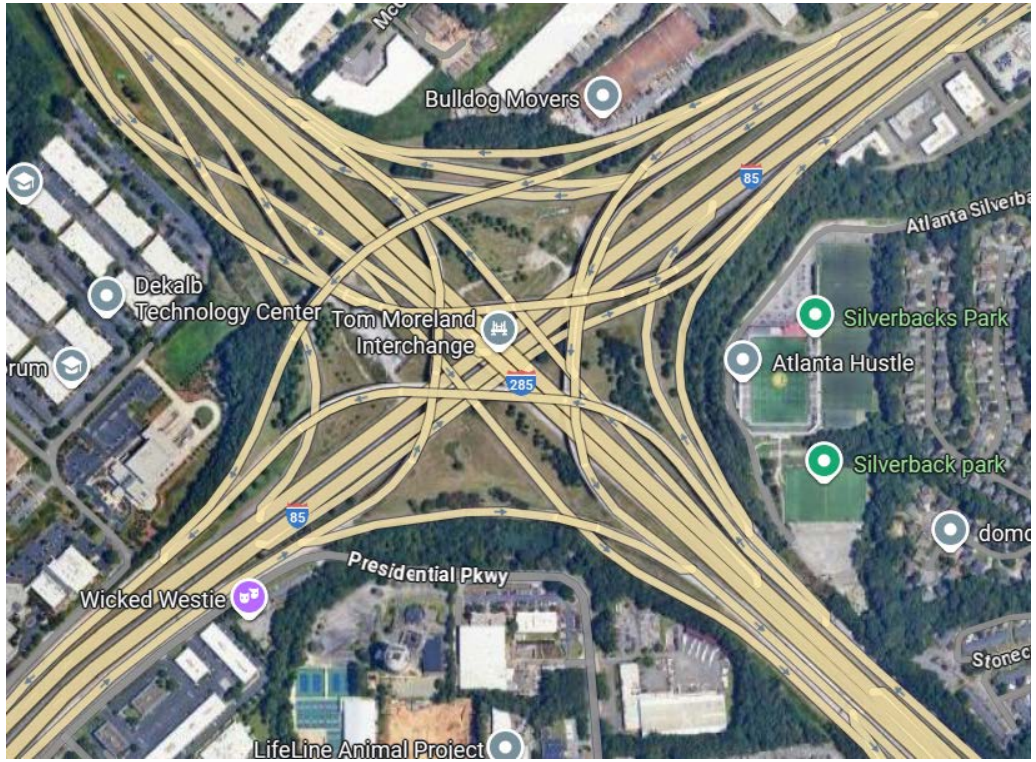
Generative models for structure search

Unsupervised learning model to learn probability distributions over a large set of molecules to later generate new molecules by sampling from these distributions

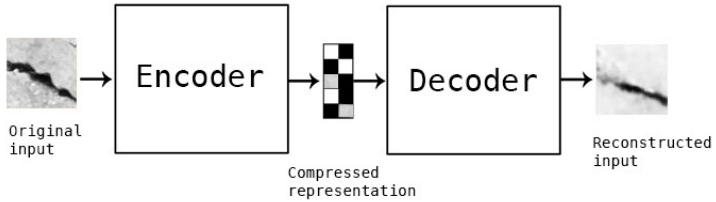
Latent models p(X) using latent or hidden space z

GAN





VAE



Direct models p(Y) only for molecules

GPT

Diffusion models

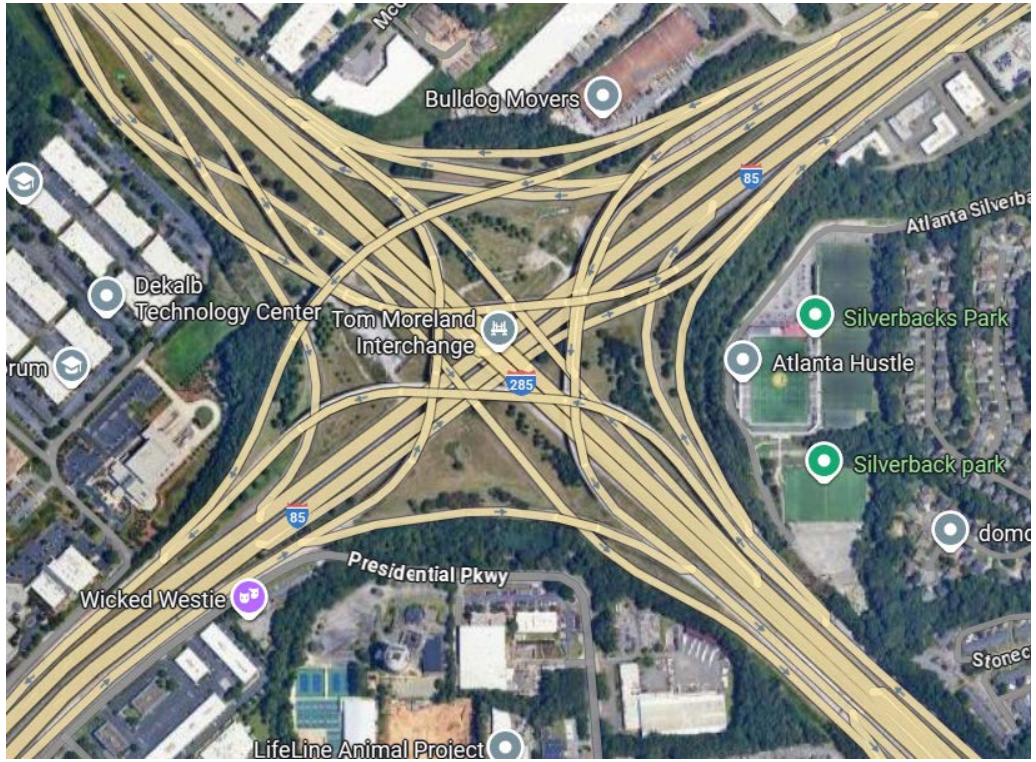


Generative models for structure search

GAN

1 road = 1 choice of data representation+..
crystal/molecule

VAE



GPT

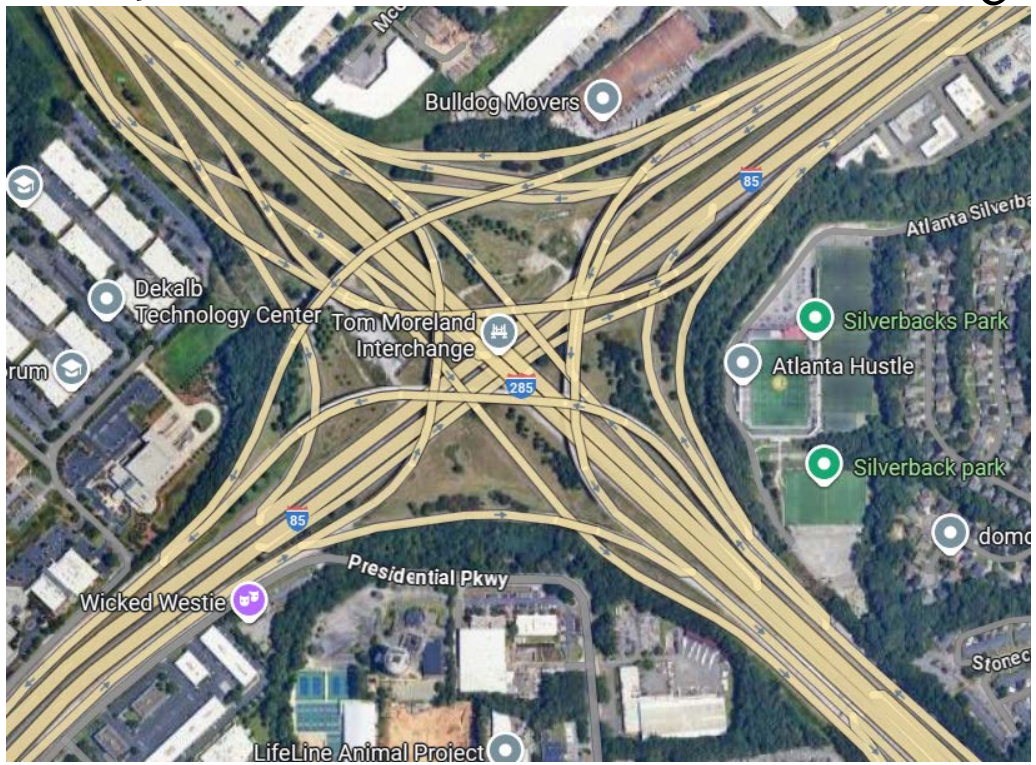
Diffusion models

Generative models for structure search

GAN

ORGAN (2020)
MolGAN (2018)
CrystalGAN

Junction Tree VAE (JT-VAE)
NP-VAE
Graph-VAE



VAE

Diffusion models

MolGPT

GPT

GeoDiff
EDM (E(3)-Equivariant Diffusion Models)
GDM (Generative Diffusion Models)

1 - Generative Adversarial Networks (GAN): a friendly competition between sparring partners NNs where both improve



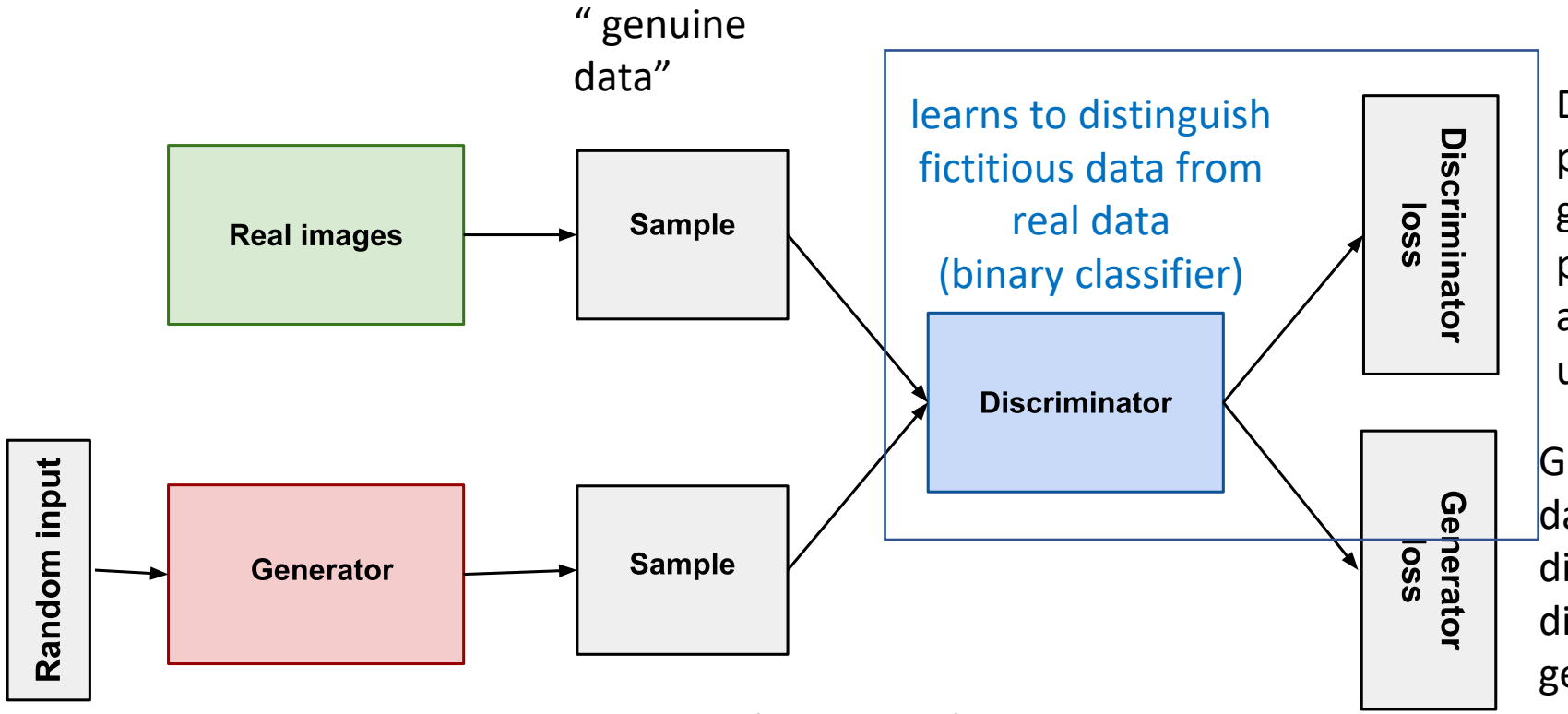
Alexis and Felix Lebrun, OG Gold medalists (Paris, 2024)





1 - Generative Adversarial Networks : competition between NNs (2014)

competition dynamics between two main neural networks : generator vs discriminator
Unsupervised learning algorithm (2014) of the distribution of the chemical space in an **implicit** manner



GAN : iterative training of the discriminator and discriminator

@

Generated Data | Discriminator | Real Data

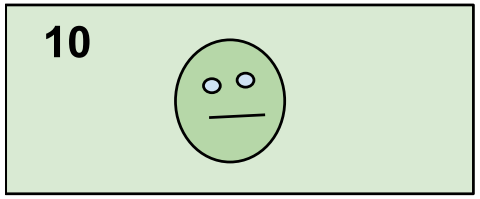
Obviously fake data



FAKE REAL



FAKE



FAKE REAL





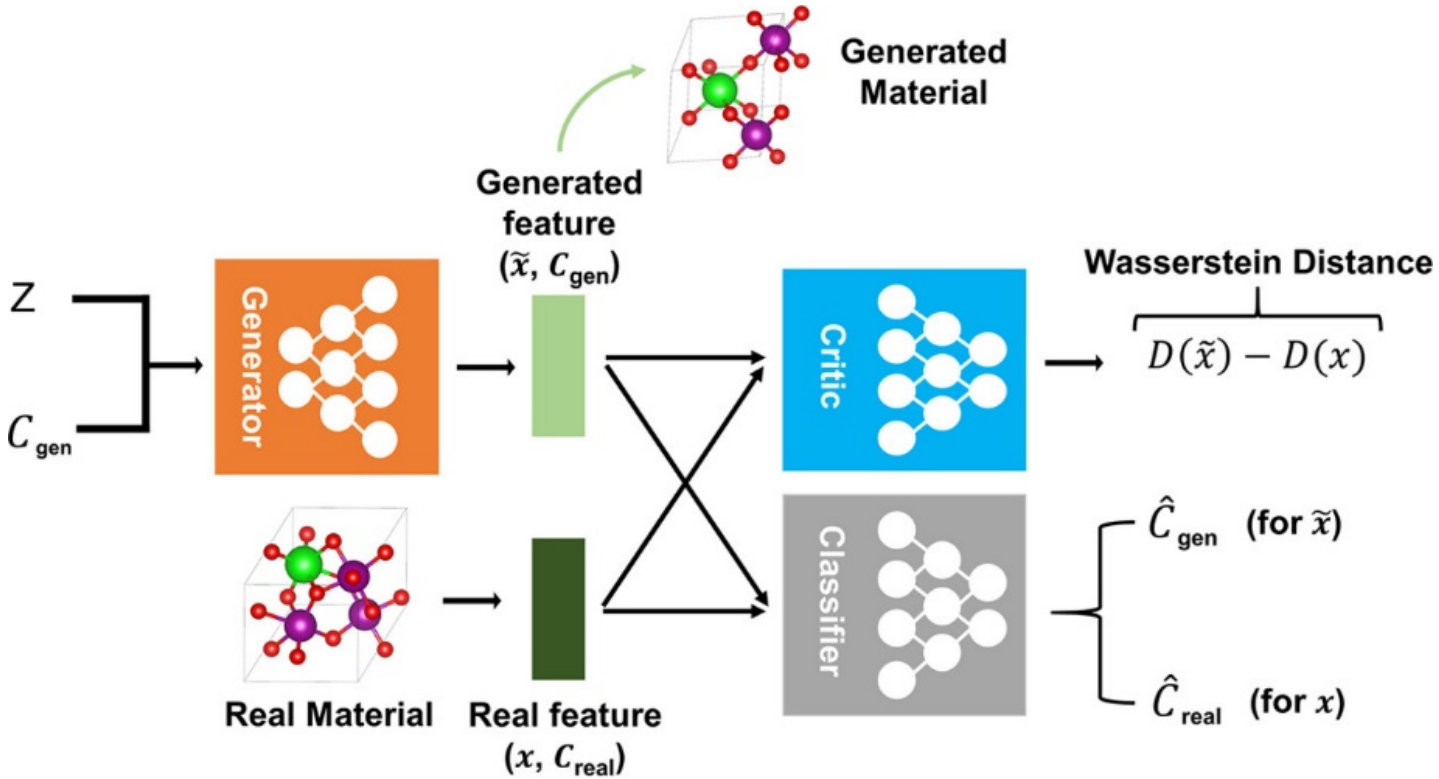
REAL REAL



Must penalize to improve discrimination

GAN for unsupervised structure prediction

- **MolGAN** (2018) : a GAN has learnt to generate directly molecular graphs
- **ORGAN** (Objective-Reinforced GAN) : GAN + reinforcement for targeted properties molecules
- **CrystalGAN (2020) to generate crystal polymorphs structures**



S. Kim, J. Noh, GH Gu, A. Aspuru-Guzik , Y. Jung: *ACS Central Sci*, 6:1412-1420 (2020)

Variational auto-encoders : SMILES, graph, 3D

X



Z

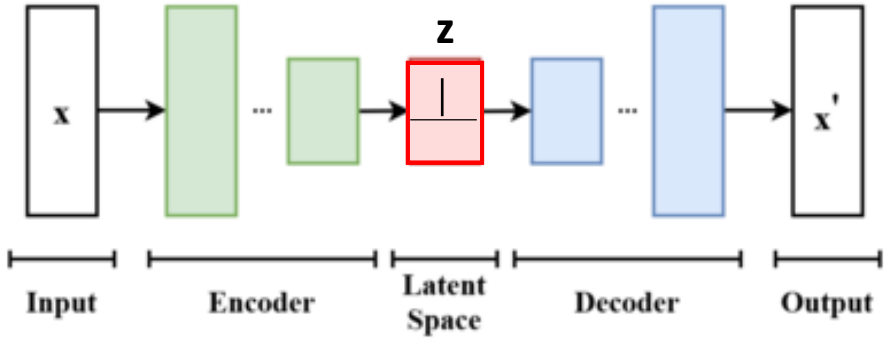


X’



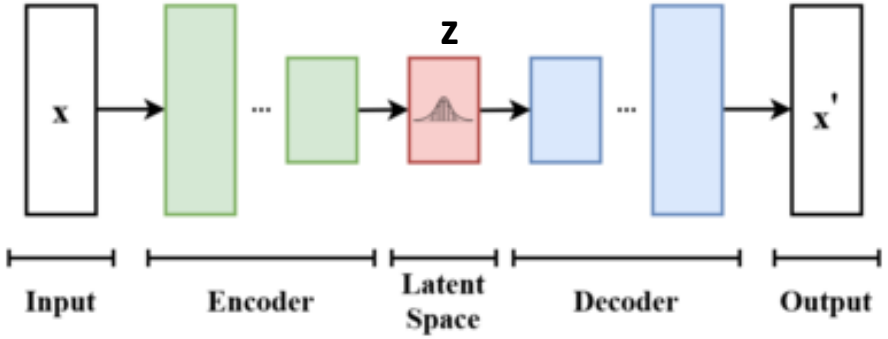
Variational auto-encoder (VAE)

Auto-encoder



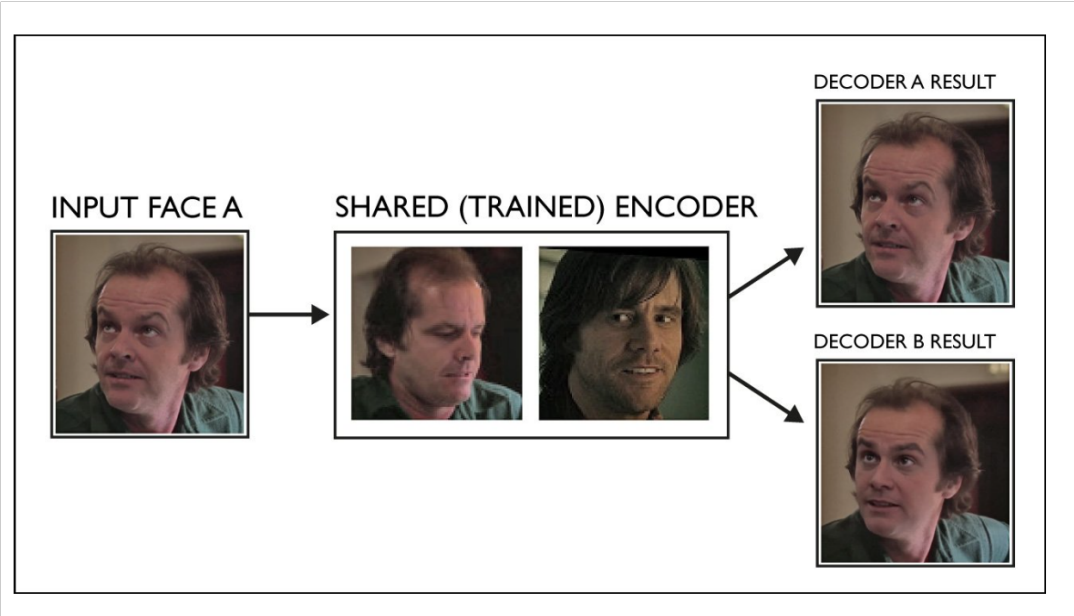
Encoder : Learns a as closest as possible representation X' of X **with loss of information**
Non recurrent neural network that propagates forward
z = (fixed, deterministic) discrete variable in latent space
See Lorenzo Branca's and Brett MacGuire's talk

Variational auto-encoder



z : continuous probabilistic representation
gives **a set of 'possible spaces' instead of one latent space.**
VAE randomly picks one of these possibilities to reconstruct X.
Gives slightly **different versions** of the same input data

Variational auto-encoder to mix faces/molecules



Variational autoencoder of DeepFaceLab

W. Jin, R Barzilay, T. Jaakkola, Junction Tree Variational Autoencode)r for Molecular Graph Generation, Proceedings of the 35 th International Conference on Machine Learning, Stockholm, Sweden, PMLR 80, 2018
T. Ochiai et al, Variational autoencoder-based chemical latent space for large molecular structures with 3D complexity, Nat. Comm. Chem, **6**, Article number: 249 (2023)

- introduced in 2016 by Bombarelli et al. : first machine learning based generative model for molecules
- **Learn a continuous latent representation of molecules (**the distribution of the chemical space in an **explicit** manner
- **Generate new ones from this latent representation**.
- **Advantage :** easy to control generation and easy to explore latent space to explore molecular families
- *Junction Tree VAE* (JT-VAE) improves the chemical validity by structuring the molecule as chemical subtrees, NP-VAE (Natural Product-oriented Variational Autoencoder), Graph-VAE

Variational auto-encoder for crystal structure prediction

npj | computational materials

Article

Published in partnership with the Shanghai Institute of Ceramics of the Chinese Academy of Sciences

https://doi.org/10.1038/s41524-024-01443-y

Deep learning generative model for crystal structure prediction

Check for updates

Xiaoshan Luo[1,2], Zhenyu Wang[1,2], Pengyue Gao[1], Jian Lv[1] ✉, Yanchao Wang[1] ✉, Changfeng Chen[4] ✉ & Yanming Ma[1,2] ✉

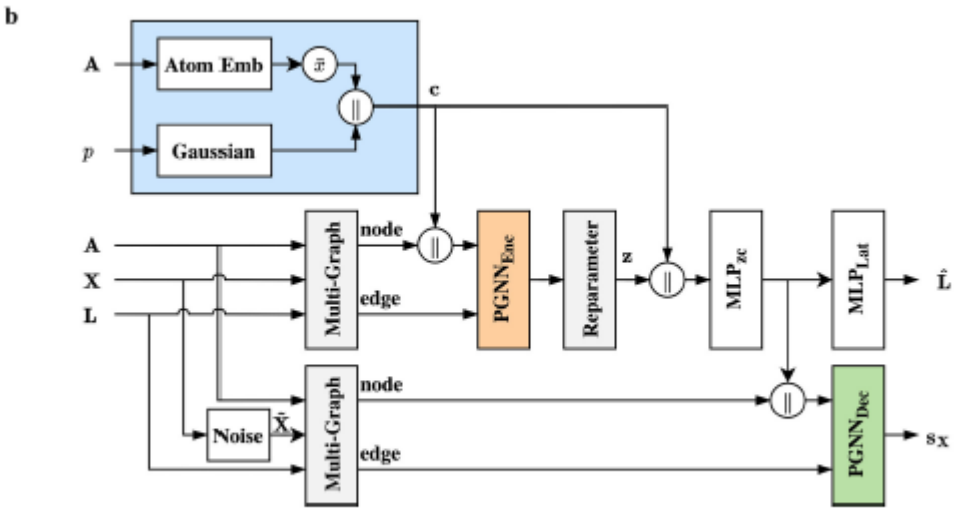
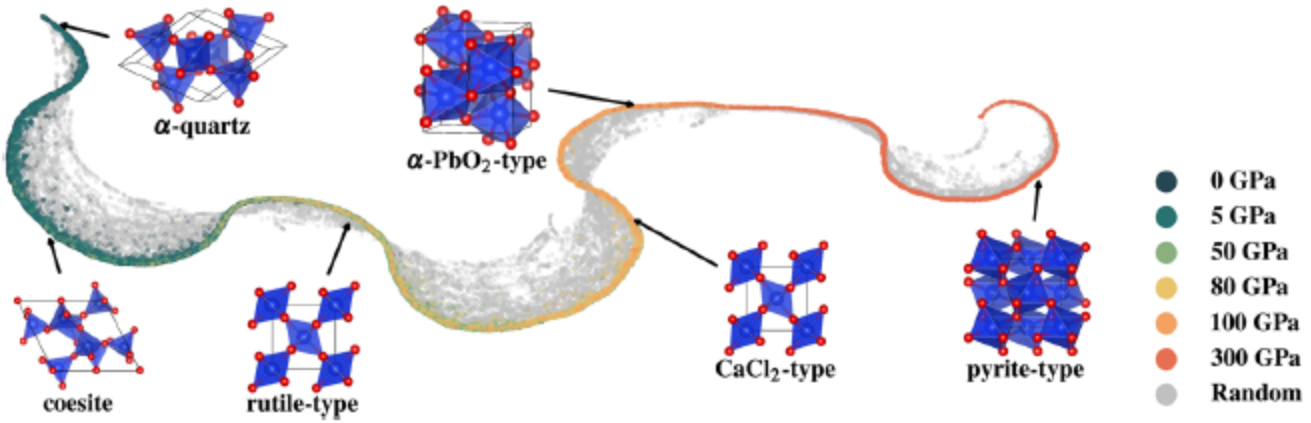


Fig. 1 | Dataset and model architecture. a Distribution of elements in the MP60-CALYPSO dataset. Color intensity represents the number of structures containing each element. b Model architecture of Cond-CDVAE. ‖ denotes weighted average of element embedding vectors based on the number of atoms of each type, ‖ denotes concatenation, and MLP$_{\square}$ denotes fully connected layers.

Fig. 4 | Two-dimensional projection of silica structures generated by the Cond-CAVAE-4M model alongside random symmetric structures. Data points representing the model-generated structures are color-coded to reflect the input pressure conditions, whereas the data points for the random structures are shown in gray. Ground-state structures of silica at various pressures are also depicted.



(Non equilibrium Thermodynamics-driven)Diffusion models



https://culturesciences.chimie.ens.fr/

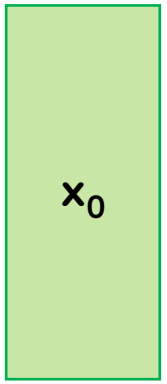
(Non equilibrium Thermodynamics-driven)Diffusion models

- **Fundamentally different** from all the previous generative methods. Recent for molecules.
- **Generate new molecular structures by inverting a progressive noise process.**
- Aim to decompose the image generation process (sampling) in many small “denoising” steps.
- The model **can correct itself over these small steps** and gradually produce a good sample.

- **AI Structures** : CNN, transformers
- **Examples :** GeoDiff, EDM (E(3)-Equivariant Diffusion Models), GDM (Generative Diffusion Models)
- **Promising for the generation of realistic 3D structures (continuous atomic coordinates).**
- Excellent/promising performance for generating realistic 3D conformal structures or molecular complexes.
- Unlike VAE models, diffusion models are learned with a fixed procedure and the **latent variable has high dimensionality (same as the original data**).

Diffusion Models learning strategy : gradually add (generally Gaussian) noise and then reverse

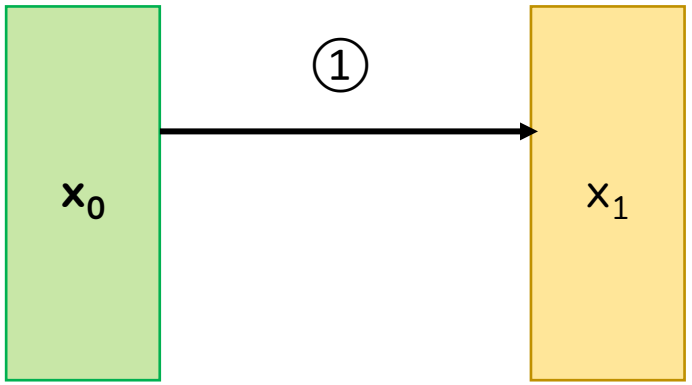
Select a sample from a generally gaussian distribution



Diffusion Models learning strategy: gradually add (generally Gaussian) noise and then reverse

Select a sample from a generally gaussian distribution

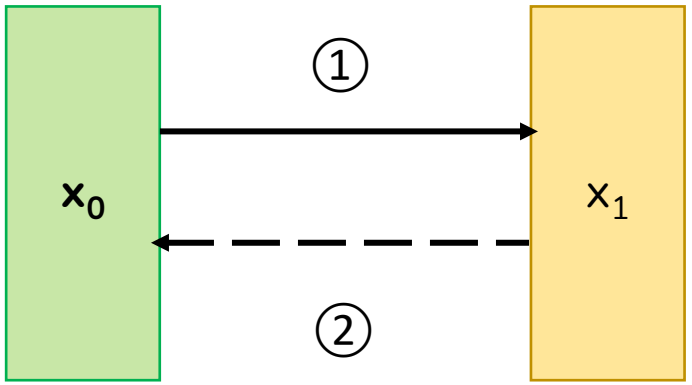
① **Forward Diffusion process : controlled addition of structured noise**



Diffusion Models learning strategy: gradually add (generally Gaussian) noise and then reverse

Select a sample from a generally gaussian distribution

① **Forward Diffusion process : controlled addition of structured noise**



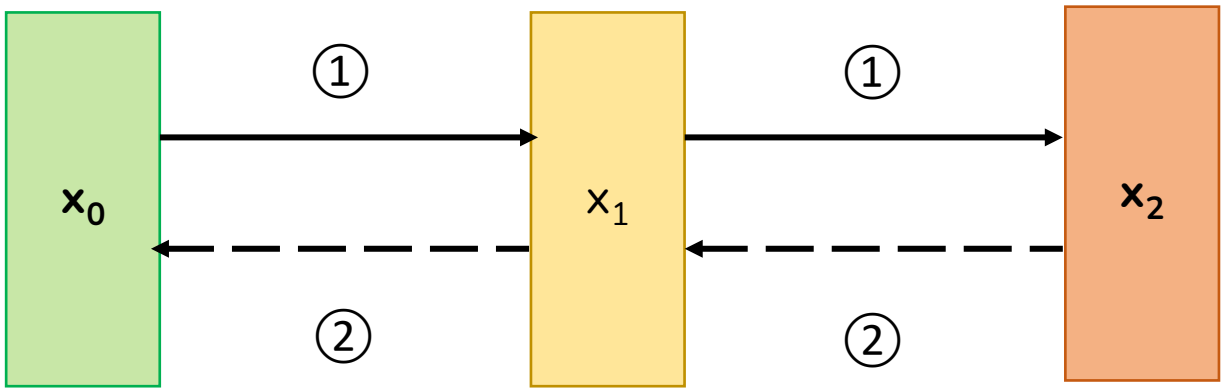
The model learns to capture and reproduce the hidden features of the initial data (variational inference) by recognizing the noise model

② **Inverse diffusion process: cancel the process to get the original data distribution from noisy ones**

Diffusion Models learning strategy: gradually add (generally Gaussian) noise and then reverse

Select a sample from a generally gaussian distribution

① **Forward Diffusion process : controlled addition of structured noise**



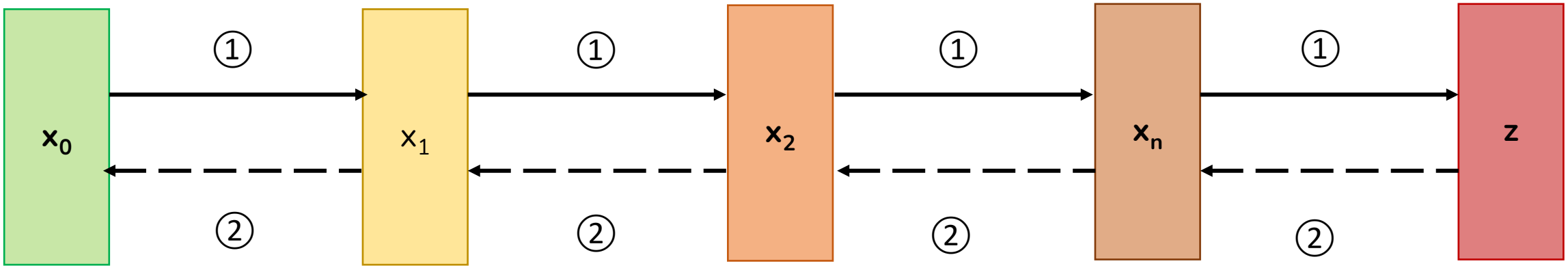
Data are progressively degradated.
The model uses its knowledge to predict the noise (complex Markov chain)

② **Inverse diffusion process: cancel the process to get the original data distribution from noisy ones**

Diffusion Models learning strategy: gradually add (generally Gaussian) noise and then reverse

Select a sample from a generally gaussian distribution

① **Forward Diffusion process : controlled addition of structured noise**



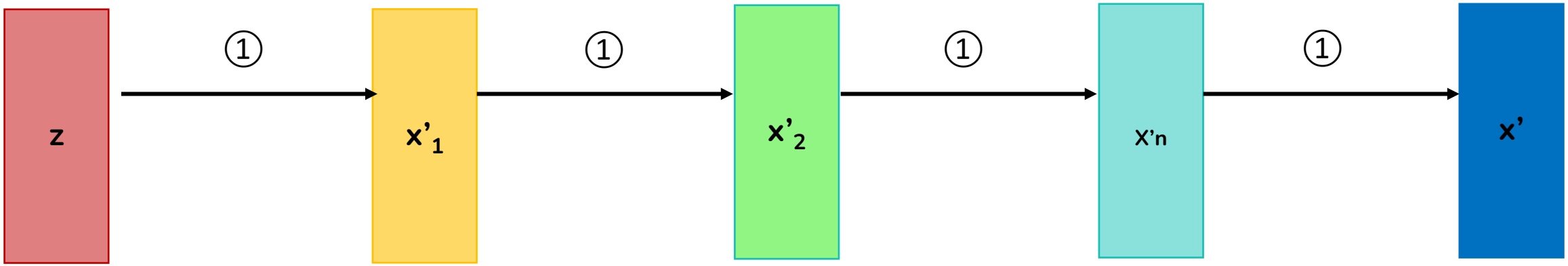
② **Inverse diffusion process: cancel the process to get the original data distribution from noisy ones.**

The last data set z is pure random noise (« white »)

Diffusion Models generation strategy: reconstruct data into something new from noise

Start from a pure noise data sample

Apply NN to reconstruct the initial data or something new



Diffusion Models

AI Structures : CNN, transformers

Examples : GeoDiff, EDM (E(3)-Equivariant Diffusion Models), GDM (Generative Diffusion Models)

Excellent/promising performance for generating realistic 3D conformal structures or molecular complexes.

Unlike VAE models, diffusion models are learned with a fixed procedure and the **latent variable has high dimensionality (same as the original data)**.

NeurIPS Proceedings

Crystal Structure Prediction by Joint Equivariant Diffusion

Part of Advances in Neural Information Processing Systems 36 (NeurIPS 2023) Main Conference Track

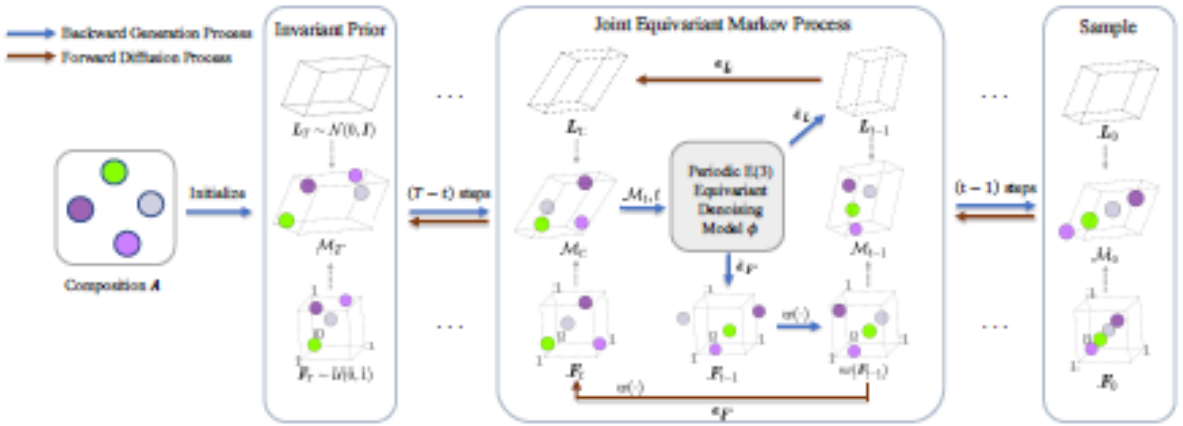


Figure 2: Overview of DiffCSP. Given the composition $\boldsymbol{A}$, we denote the crystal, its lattice and fractional coordinate matrix at time t as $\mathcal{M}_t$, $\boldsymbol{L}_t$ and $\boldsymbol{F}_t$, respectively. The terms $\boldsymbol{\epsilon}_{\boldsymbol{L}}$ and $\boldsymbol{\epsilon}_{\boldsymbol{F}}$ are Gaussian noises, $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{L}}$ and $\hat{\boldsymbol{\epsilon}}_{\boldsymbol{F}}$ are predicted by the denoising model ϕ.

5.2 Comparison with DFT-based Methods

We further select 10 binary and 5 ternary compounds in MP-20 testing set and compare our model with USPEX [59], a DFT-based software equipped with the evolutionary algorithm to search for stable structures.

Table 2: Overall results over the 15 selected compounds.

	Match rate (%)↑	Avg. RMSD↓	Avg. Time↓
USPEX [59]	53.33	**0.0159**	12.5h
DiffCSP	**73.33**	0.0172	**10s**

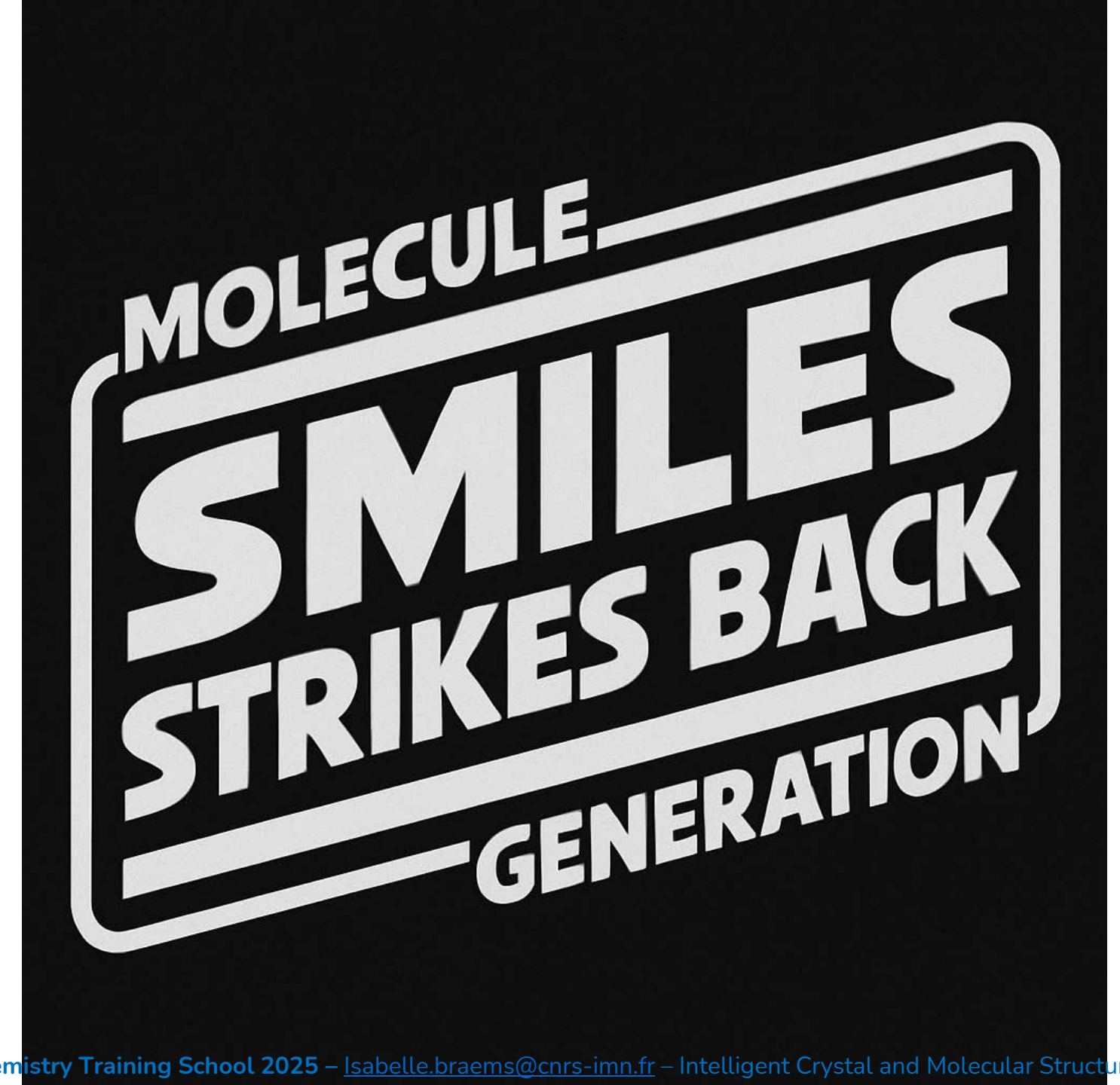
For our method, we sample 20 candidates for each compound following the setting in Table 1. We select the model trained on MP-20 for inference, with a training duration of 5.2 hours. For USPEX, we apply 20 generations, 20 populations for each compound, and select the best sample in each generation, leading to 20 candidates as well. We summarize the **Match rate** over the 15 compounds, the **Averaged RMSD** over the matched structures, and the **Averaged Inference Time** to generate 20 candidates for each compound in Table 10. The detailed results for each compound are listed in Appendix F. DiffCSP correctly predicts more structures with higher match rate, and more importantly, its time cost is much less than USPEX, allowing more potential for real applications.

This was 2019…

«Several important high level themes emerge as a result of this work, including the shift away from the SMILES string representation of molecules towards more sophisticated representations such as graph grammars and 3D representations»

D. C. Elton, Z. Boukouvalas, M. D. Fuge and P. W. Chung, Deep learning for molecular design—a review of the state of the art, Mol. Syst. Des. Eng., 2019

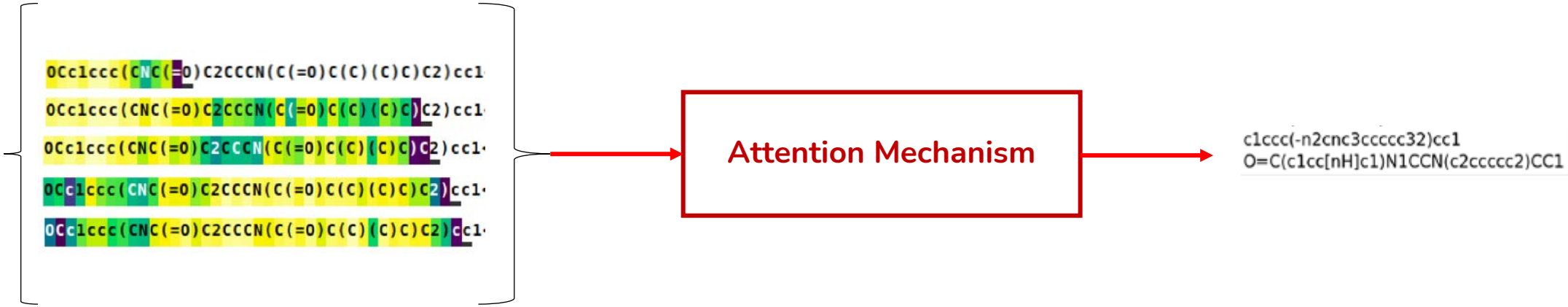
But **Transformers** generate sequences that align with what is required to interpret **SMILES**


MOLECULE
SMILES
STRIKES BACK
GENERATION

(Generative Pretrained) Transformers (for molecules only)

A **transformer** processes a **sequence input** (text, SMILES) **into a similar kind of output** (text, SMILES)

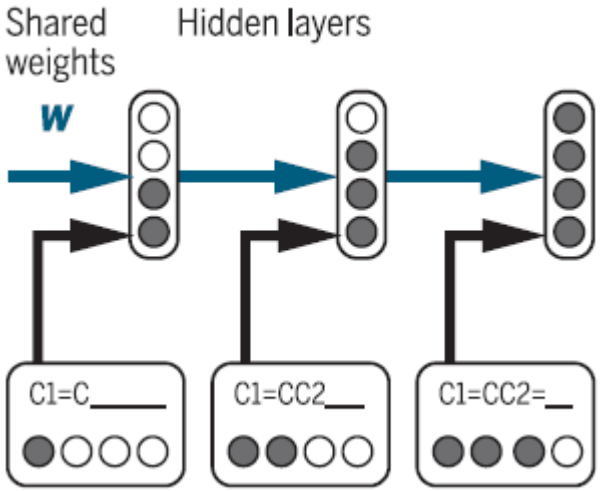
This without recursivity or convolution, but using an **(multi-head) attention mechanism.**



High learning capacity from large databases (unlabeled data, limited labeled data)
Able to capture **long-range dependencies and contextual information via dynamical learning**
Continual Improvement
Flexibility: an accommodate different types of molecular representations beyond SMILES, such as graph-based representations

SMILES can generate invalid strings if the model is not well trained or controlled.

From RNN to transformers



An **embedding** is a dense (256 or 768-length) vector of a token
It is **automatically learnt and associated** to each token during the training according to the context of use.

Sequence models are often trained to predict **just a single missing token** in a sequence (trying to predict more than one token leads to a **combinatorial explosion** of possibilities)

training input SMILES : “CC(=O)O”

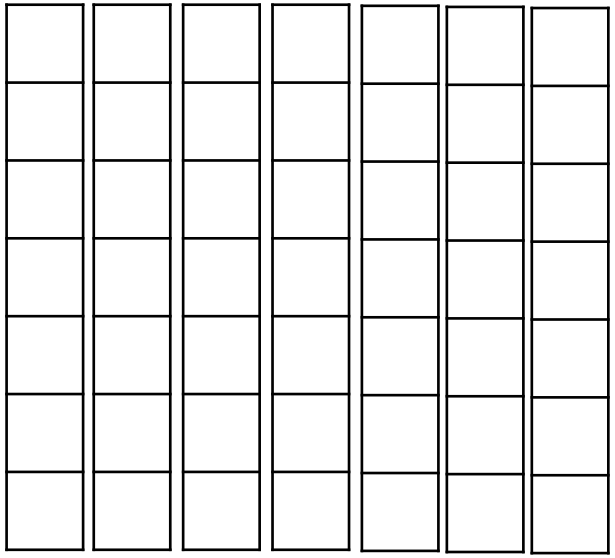
↓ **Tokenisation**

input = "C", "C", "(", "=", "O", ")", "O

output = "C", "(", "=", "O", ")", "O", "EOS“

↓ **Embedding**

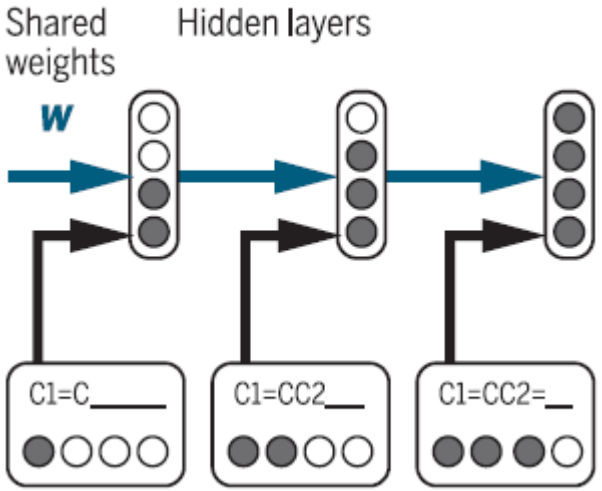
input = "C", "C", "(", "=", "O", ")", "O



These vectors allow the model to **compute relationships between tokens** (similarity, analogies, etc.) and identify the different context on the scale of a sentence.

A positionnal encoding is added for each token as the model is non sequential

From RNN …



Sequence models are often trained to predict **just a single missing token** in a sequence (trying to predict more than one token leads to a **combinatorial explosion** of possibilities)

training input SMILES : “CC(=O)O”

Tokenisation

input = "C", "C", "(", "=", "O", ")", "O

output = "C", "(", "=", "O", ")", "O", "EOS“

Embedding

"C" → RNN → “W"

« C » → RNN → “("

« (» → RNN → “O"

« O » → RNN → “O"

"C", "(", "=", "O", ")", "O", "EOS“

"C", "(", "=", "O", ")", "O", "EOS“

"C", "(", "=", "O", ")", "O", "EOS“

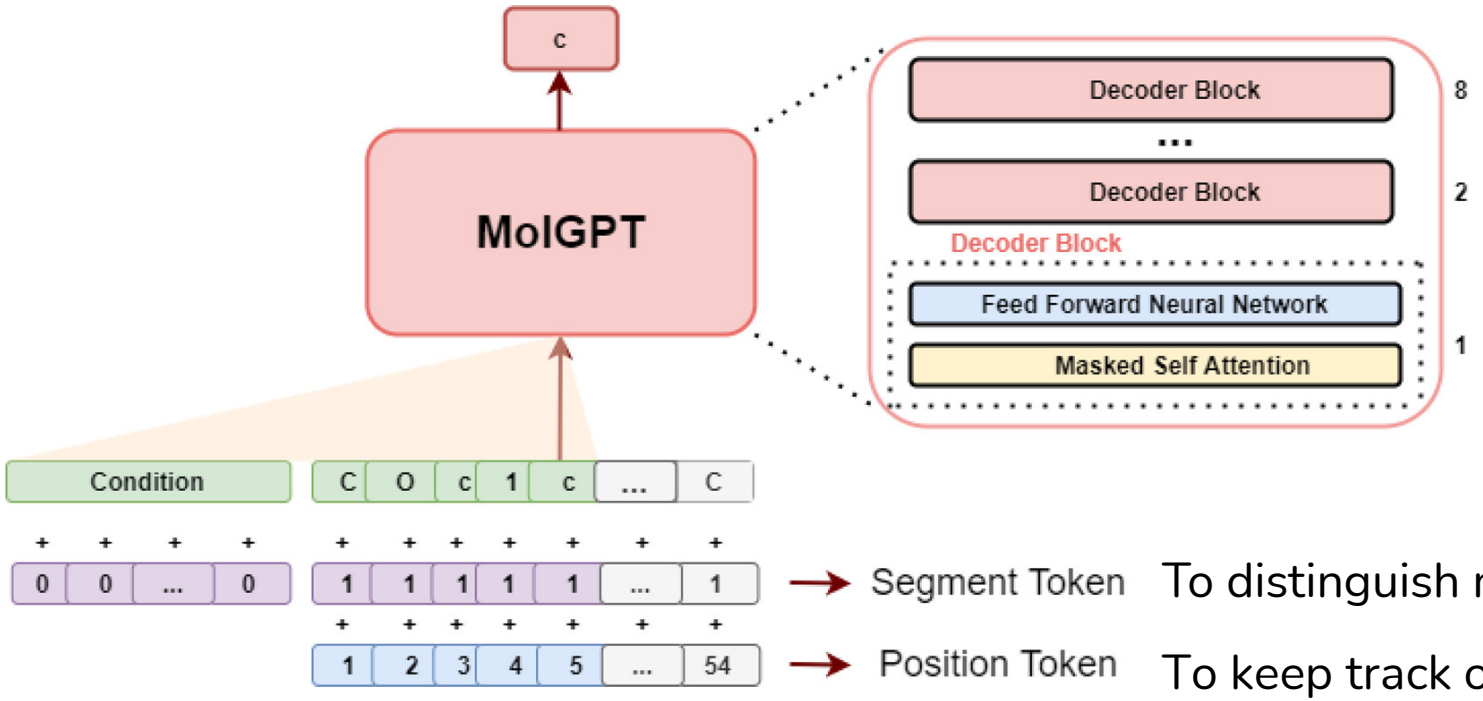
"C", "(", "=", "O", ")", "O", "EOS“

C → W → (→ O → O →) → O vs C → C → (→ = → O →) → O

RNN weights are adjusted to maximize the probability of true outputs $P(x_1, x_2, ..., x_n) = \prod_{i=1}^{n} P(x_i|x_1, ..., x_{i-1})$

… to (Generative Pretrained) Transformers : MolGPT

attention mechanism : no need data to be treated sequentially, highly paralellizable !



Only stacked decoder blocks
(Only) 6M parameters

In each step of the decoder,
we look to see if important information can be found in the past (rereading the sentence to be translated at each word/step)

To distinguish molecules from conditions

To keep track of the order of the input sequence

Trained to **predict (syntax and chemically) valid SMILES**

Can be conditionned on a targeted property

No graph.. **Easier to train that molecular graph (e.g. GrahMolGen)**

M.E. Mswahili, Y-S Jeong,Transformer-based models for chemical SMILES representation: A comprehensive literature review Heliyon, 10(20), e39038 (2024)

Attention Is All You Need

Ashish Vaswani* Google Brain avaswani@google.com

Noam Shazeer* Google Brain noam@google.com

Niki Parmar* Google Research nikip@google.com

Jakob Uszkoreit* Google Research usz@google.com

Llion Jones* Google Research llion@google.com

Aidan N. Gomez* † University of Toronto aidan@cs.toronto.edu

Łukasz Kaiser* Google Brain lukaszkaiser@google.com

Illia Polosukhin* ‡ illia.polosukhin@gmail.com

NeurIPS Proceedings

Periodic Graph Transformers for Crystal Material Property Prediction

Part of Advances in Neural Information Processing Systems 35 (NeurIPS 2022) Main Conference Track

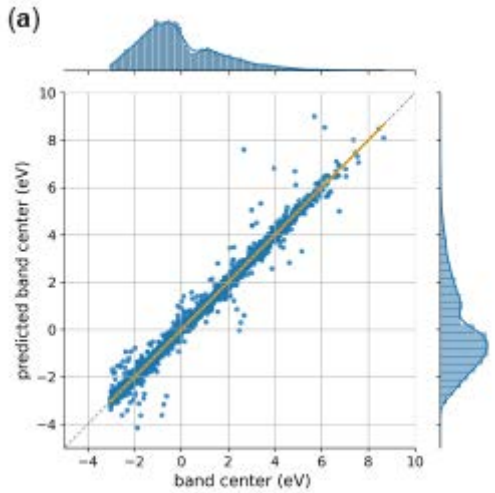
arXiv > cond-mat > arXiv:2411.16483

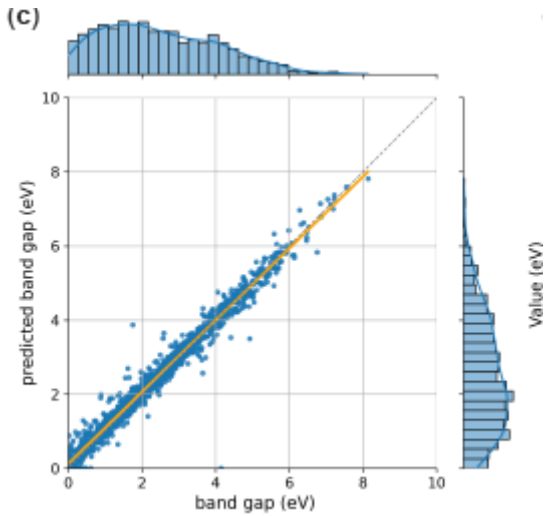
Condensed Matter > Materials Science

[Submitted on 25 Nov 2024]

Graph Transformer Networks for Accurate Band Structure Prediction: An End-to-End Approach

Weiyi Gong, Tao Sun, Hexin Bai, Jeng-Yuan Tsai, Haibin Ling, Qimin Yan





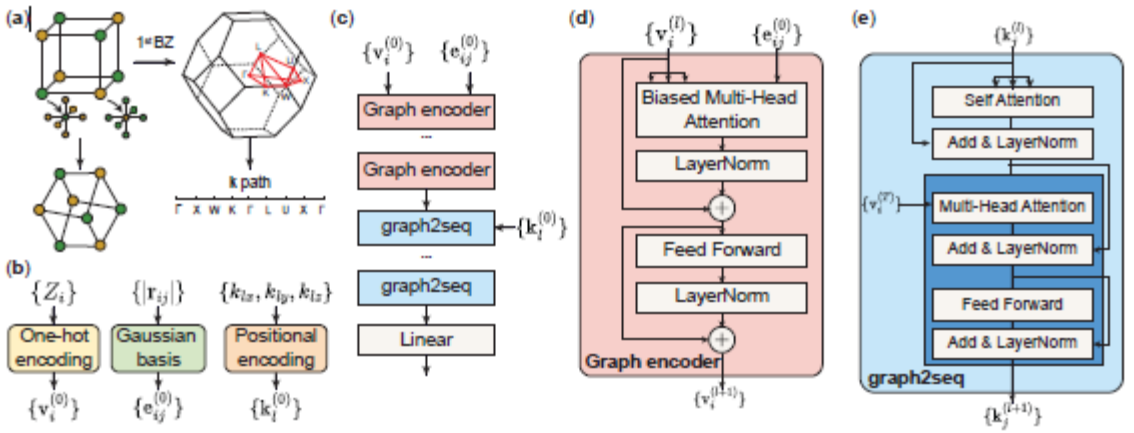


FIG. 1. **The workflow and architecture of Bandformer.** **(a)** For each input crystal, a crystal graph is constructed from local environment of atoms in the primitive cell, which is used as the input for the encoder. The high symmetry k-path in the first Brillouin zone is extracted and resampled and used as the input for the decoder graph2seq. **(b)** Atoms are given element features attached to graph nodes, and interatomic distance are given Gaussian expansions of the distance as features. K-points are given spatial positional encoding. **(c)** The architecture of Bandformer. **(d)** The architecture of graph encoder. The input atomic features and the edge features are used to generate a bias term when calculating the attention coefficients in the Biased Multi-head Attention module, which is connected to LayerNorm layer in a ResNet structure [23]. **(e)** For each k-point along high symmetry k-path, positional encoding of k-point coordinates is used as the initial feature, which is passed to a graph2seq decoder. Within the decoder, k-points learns positional information through Self Attention module. The output is passed to

What if the skeleton structure is not known? Needs to generature structures!

« The best way to have a good idea is to have lots of ideas »
Linus Pauling

Generative AI (2017)
Very large database required
Designed to become transferable

Generate new valid, and unique molecules and crystals that are only partially similar to data on which they were trained



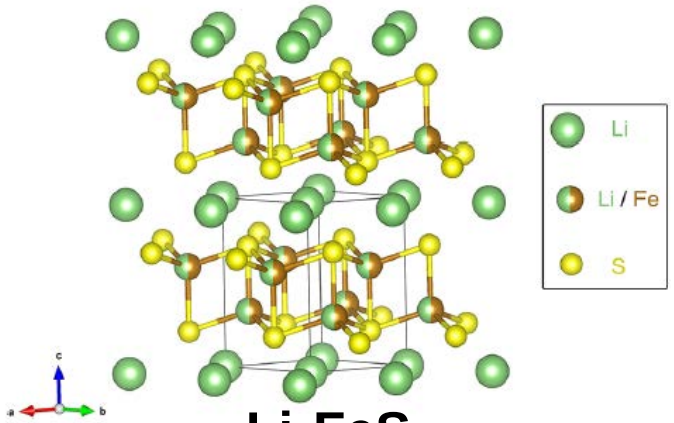
"Everything should be made as simple as possible, but not simpler"
Albert Einstein

« Less is more : let us stick to the question »
many (including me)

Global optimization (2006)
No database required
Not transferable



Double delithiation of Li_2FeS_2 gives an long-term unknown FeS_2 structure



Li_2FeS_2

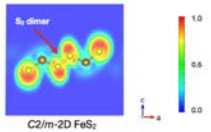
-Li

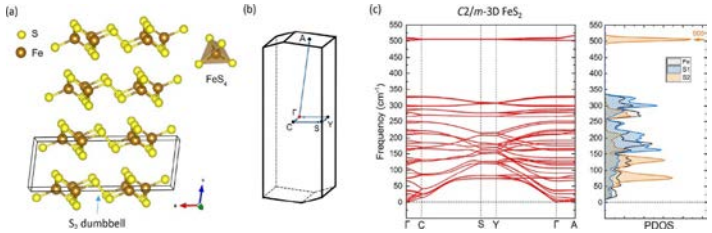
Unknown structure since 1989

$Fe^{3+}S^{2-}(S_2)^{2-}_{1/2}$

- ***IR spectroscopy :***
  - d(S–S)= 2.12 Å
- ***Mössbauer and EXAFS :***
  - Fe^{3+}HS ion in **tetrahedral sites**
  - d(Fe–S)=2.28 Å
  - **S^{2-}/S_2^{2-} = 2/1**
- **Lamellar structure** expected due to the stacking of $LiFeS_2$ and Li layers.





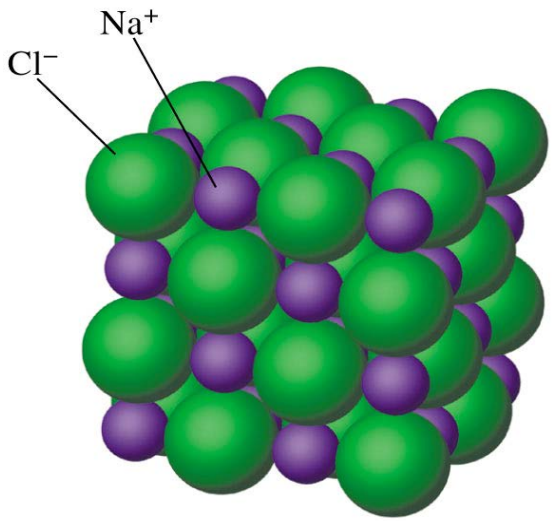


R. Brec et al. *J. Power. Sources.* **1989**, *26*, 325

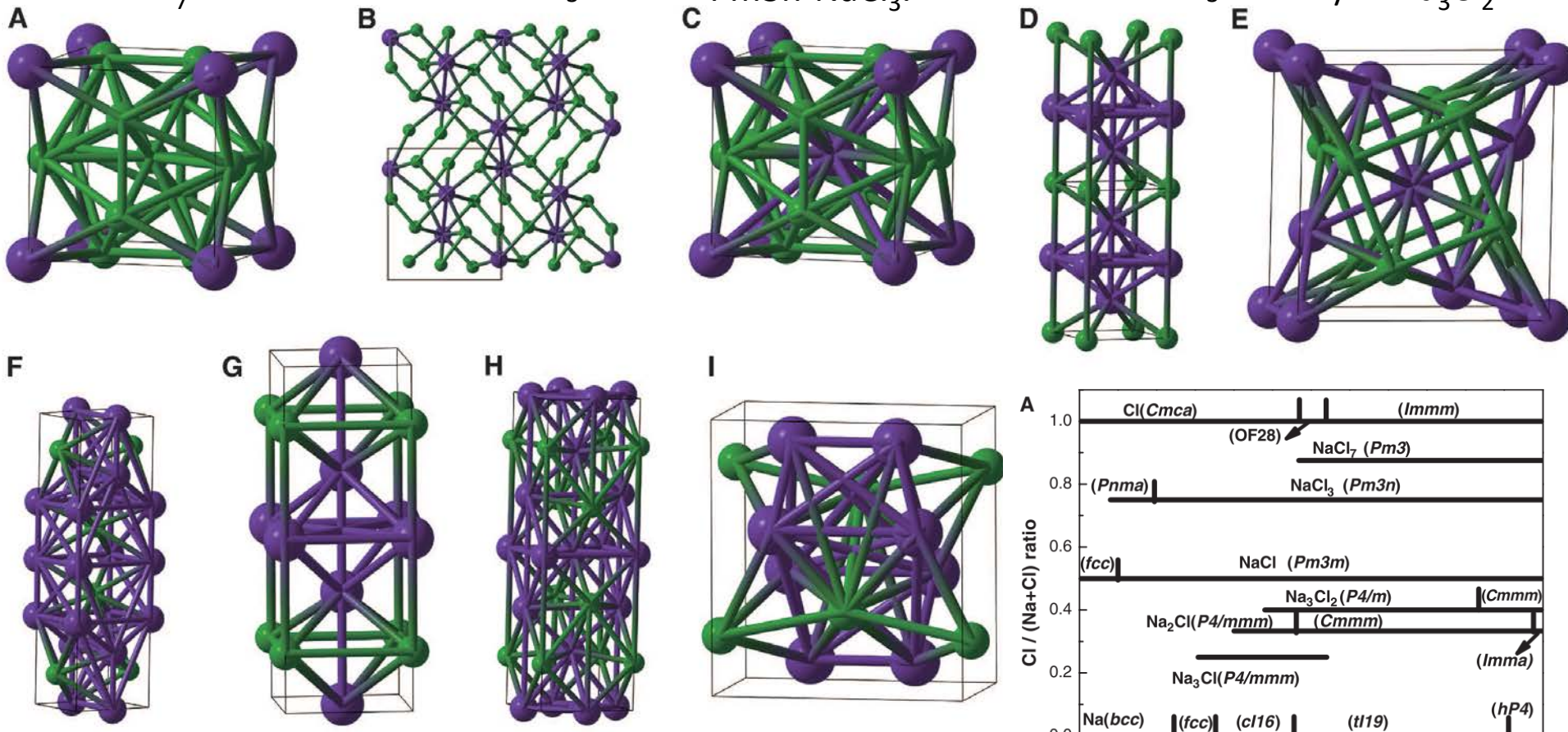
B. Wang, IB, S. Sasaki, G. Guegan, L. Cario, S. Jobic and G. Frapper, *J. Phys. Lett.* **2020**, *11*, 8861.

What if the underlying lattice is not known and no need/cannot use database ?

Pressure induced phases: computationally predicted and (some) later confirmed experimentally



W Zhang,A. R. Oganov et a*l, Science*, 342 (2013)



Pm3-$NaCl_7$.

Pnma-$NaCl_3$

Pm3n-$NaCl_3$.

P4/mmm-Na_3Cl.

P4/m-Na_3Cl_2.

Cmmm-Na_3Cl_2.

P4/mmm-Na_2Cl

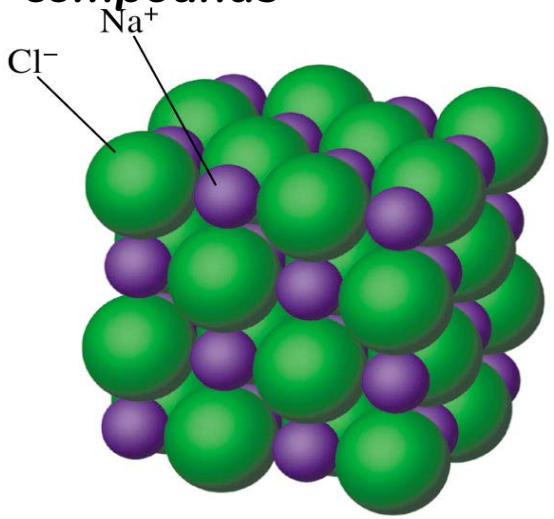
Imma-Na_2Cl

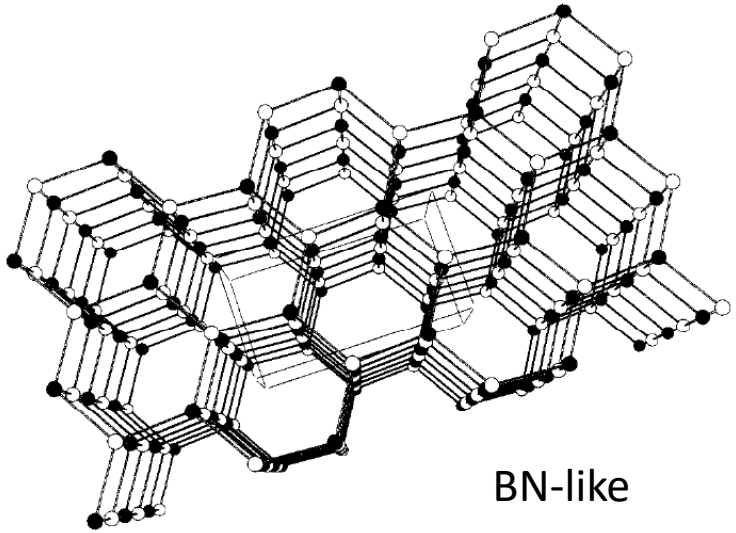
Cmmm-Na_2Cl.

What if the underlying lattice is not known and no need/cannot use database ?

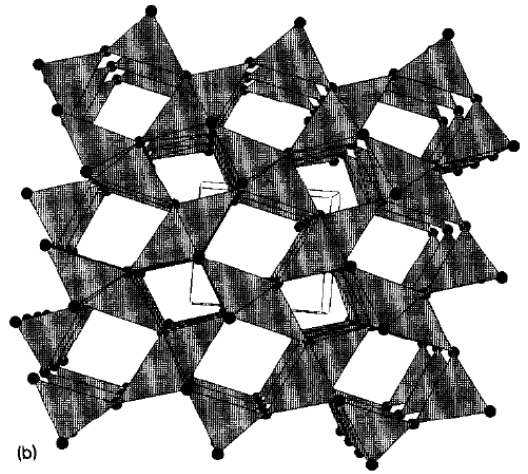
Metastable phases computationally predicted and (some) later confirmed experimentally

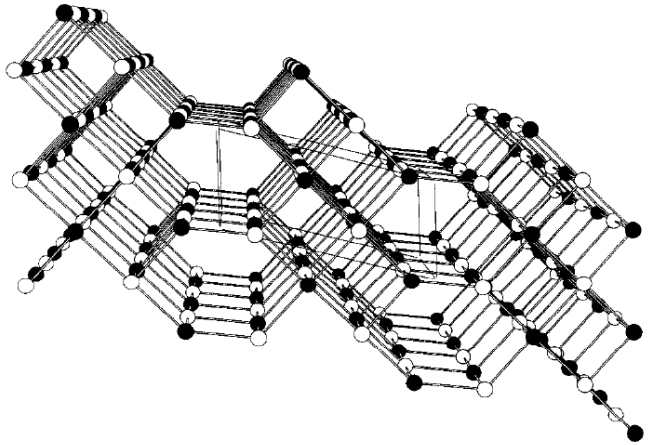
Metastable compounds





BN-like



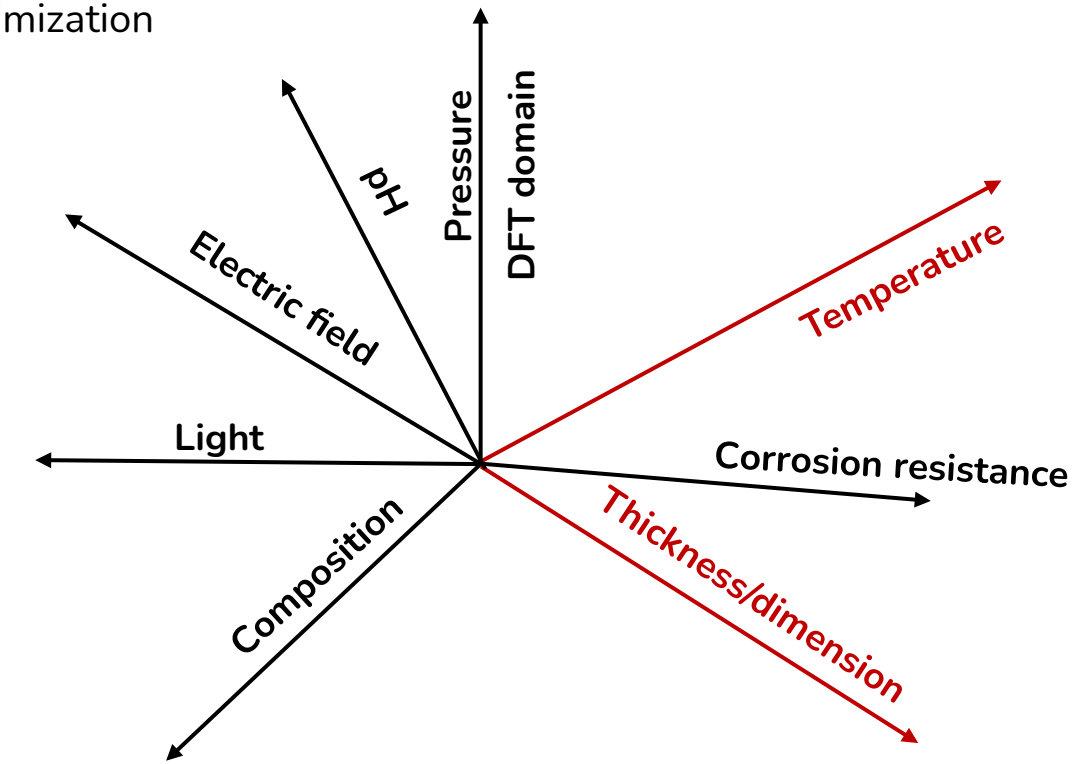




Requires a method functioning with **no prior knowledge** to allow unexpected discoveries **priori CSP**

Global optimisation (GO) is an alternate solution to learning(s)

- Seek novel structures that **do not belong** to databases
- account for ranking crystal structures as a function of synthesis conditions (labeled data)
- Analyses the Potential Energy Landscape associated with the composition (non-transferable)
- Combines efficient combinatorial search + energetic optimization
- **GO at the core of most advanced GM models!**



Q. Tong, L. Xue , J. Lv Y, Wang Y, Ma, *Faraday Discuss*, 211:31-43 (2018)



Chemistry 2024

Predicting the 3D structure that a protein will adopt based **solely on its amino acid sequence**



Demis Hassabis

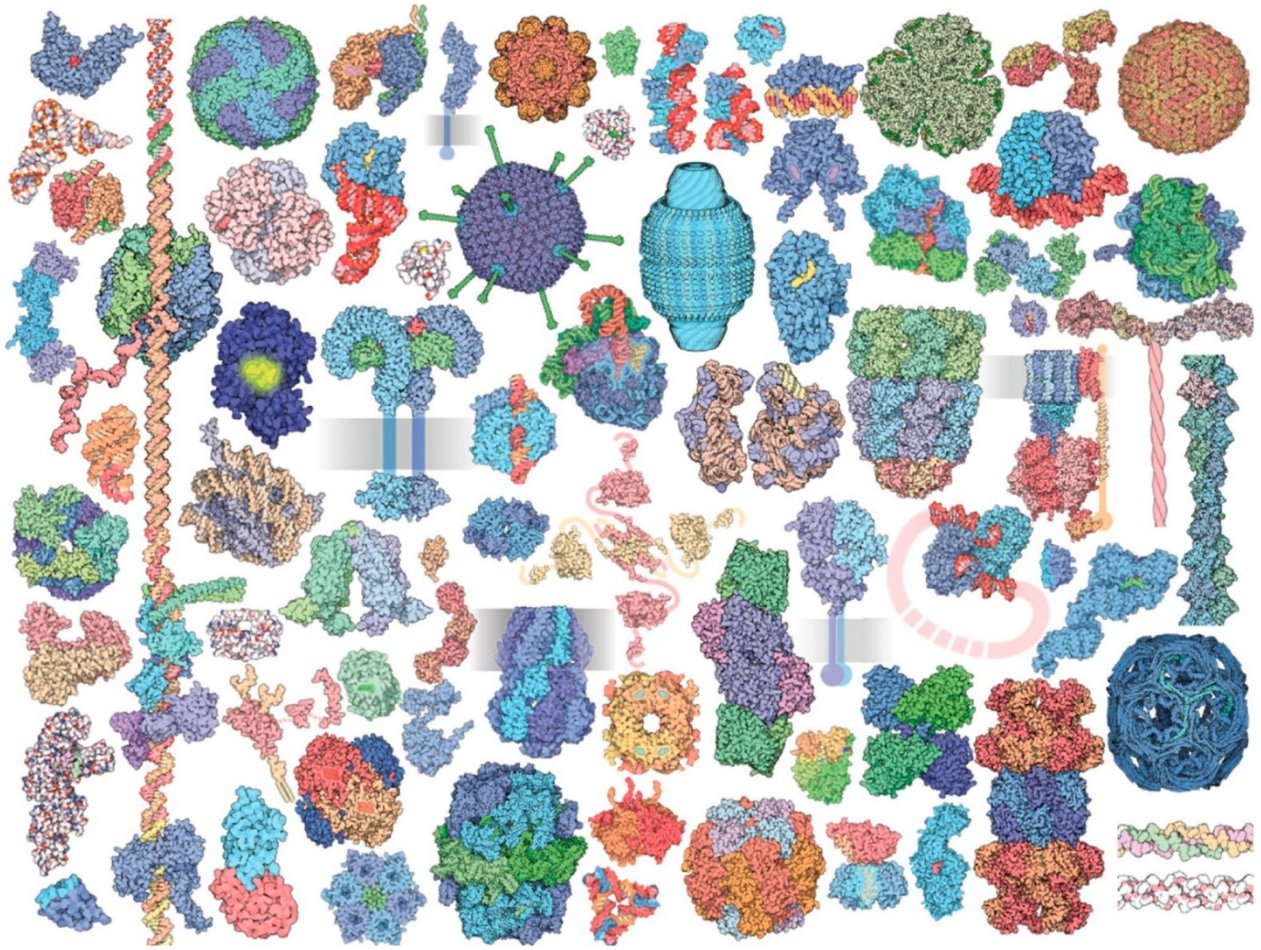


John Jumper

DeepMind

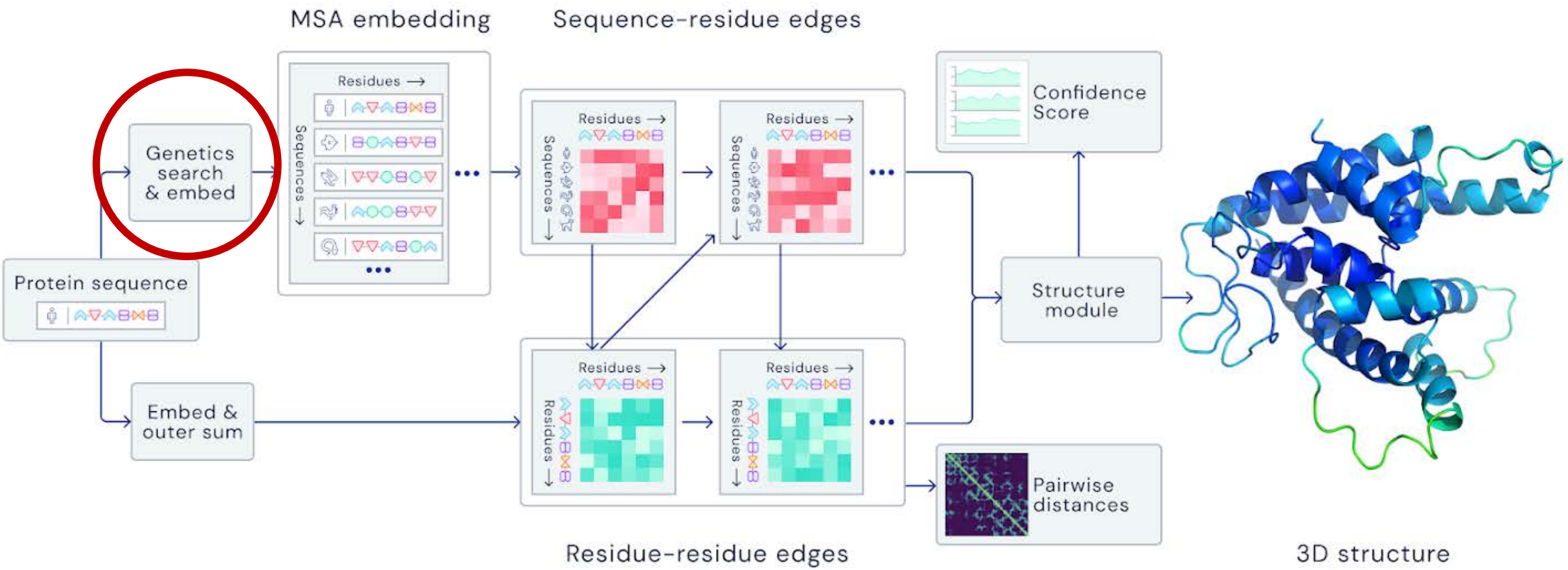


David Baker



The remarkable diversity and complexity of protein structures.
Credit to "RCSB Protein Data Bank" (CC-BY-4.0 license)

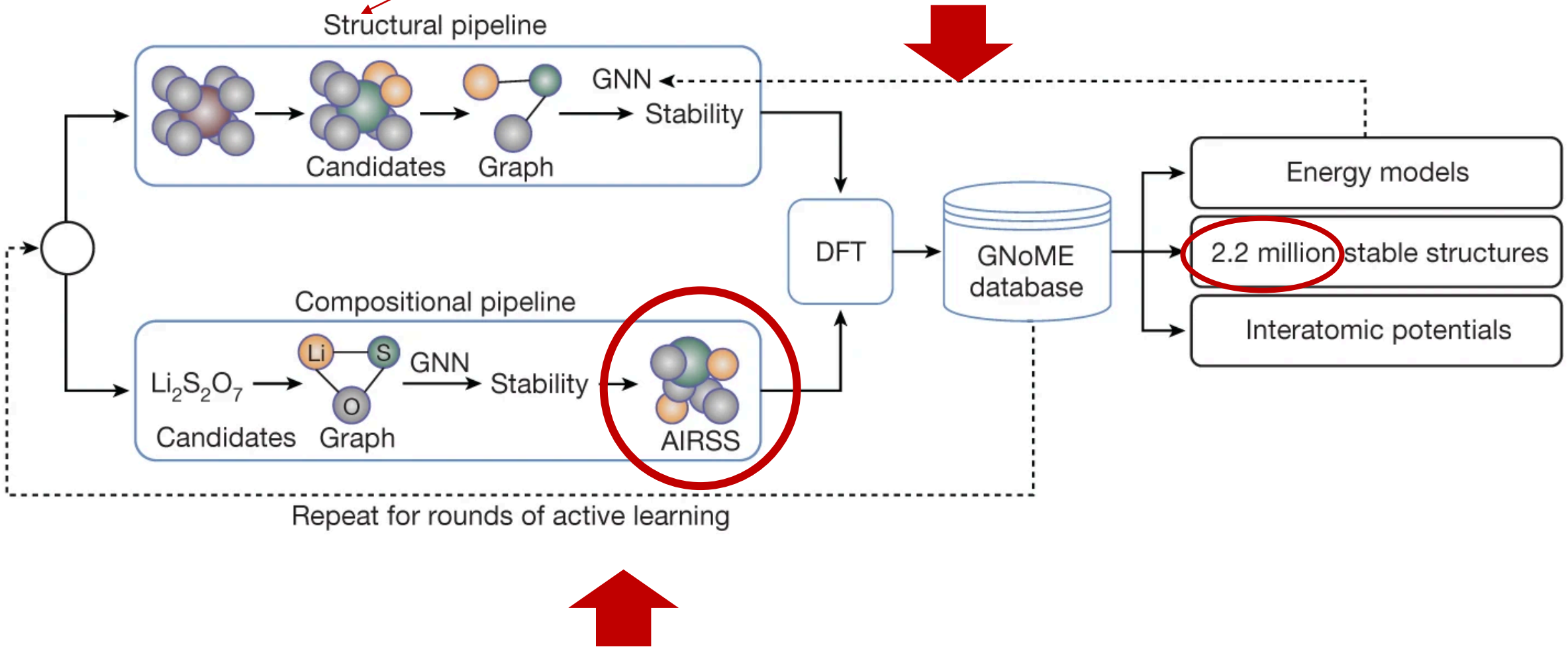
Deepmind AlphaFold2 architecture



J. Jumper et al, D. Hassabis, **Highly accurate protein structure prediction with AlphaFold, *Nature* 596**, pages 583–589 (2021)

DeepMind GNoME - graph neural networks for materials exploration

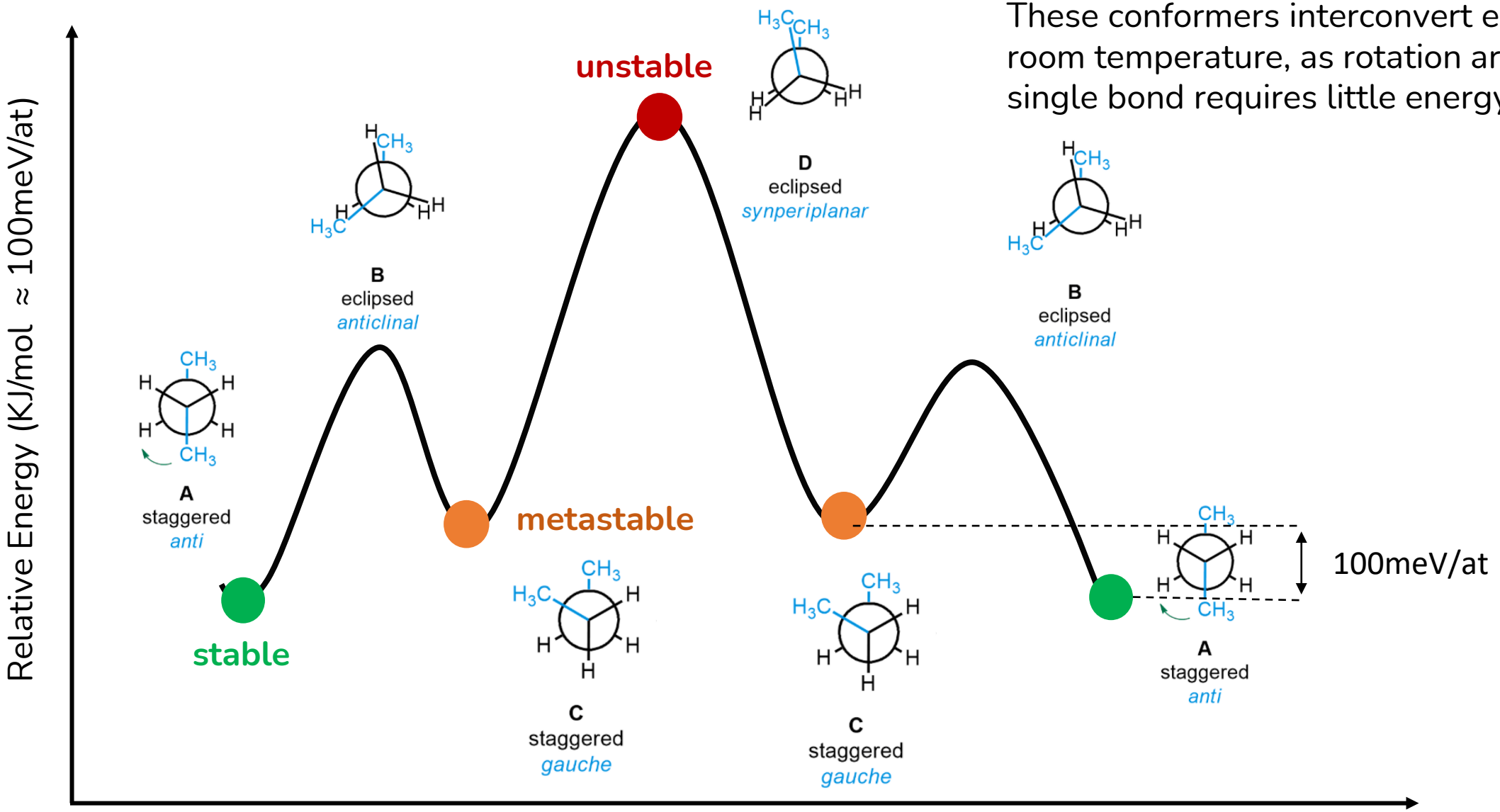
Generates candidate structures similar to known crystals



A. Merchant, S. Batzner, S. S. Schoenholz, M. Aykol, G. Cheon, E. D. Cubuk, Nature 624(7), 2023

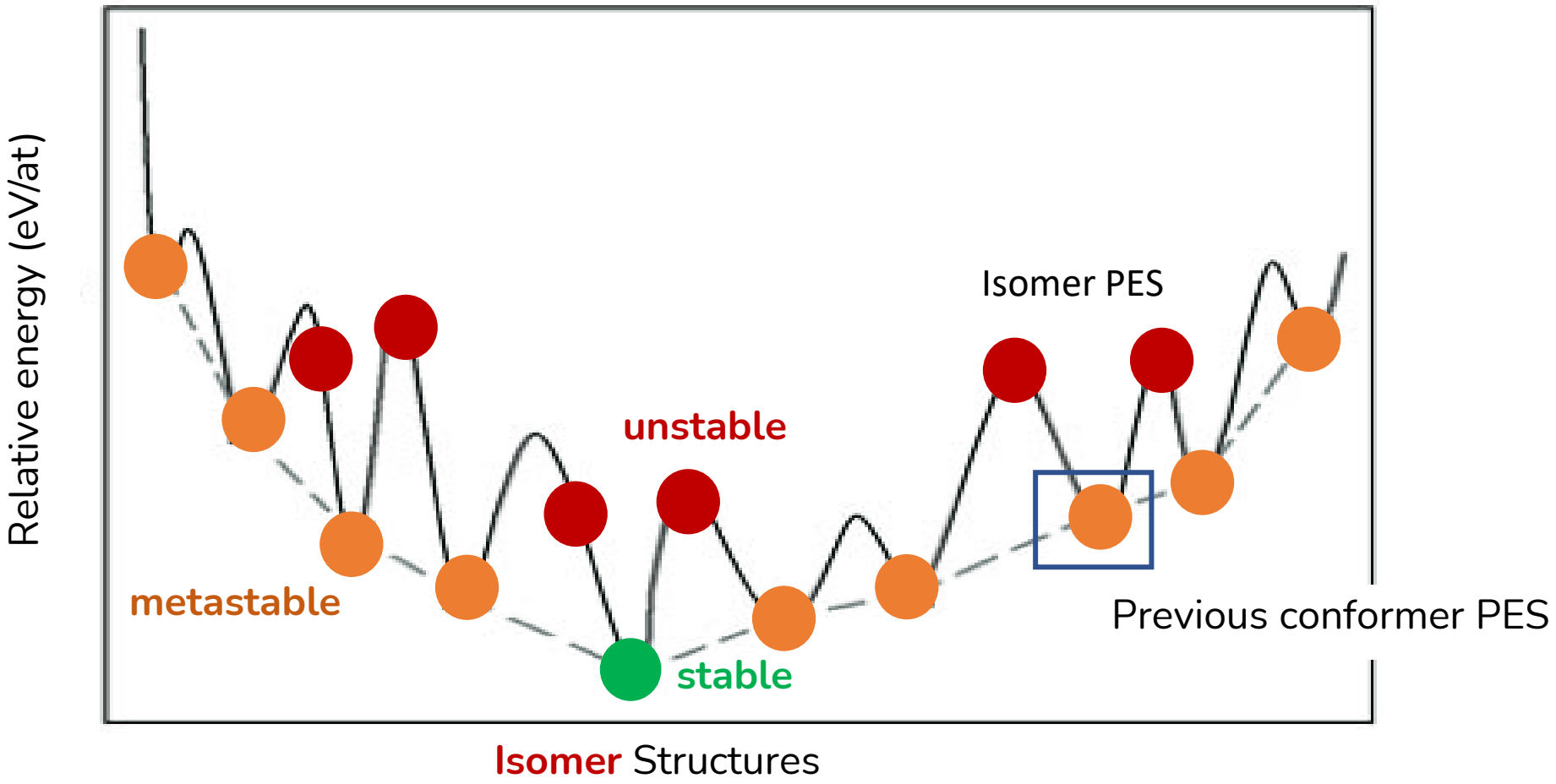
Potential energy landscape (conformers)

These conformers interconvert easily at room temperature, as rotation around a single bond requires little energy.



One step back : Potential energy landscape for isomers or polymorphs

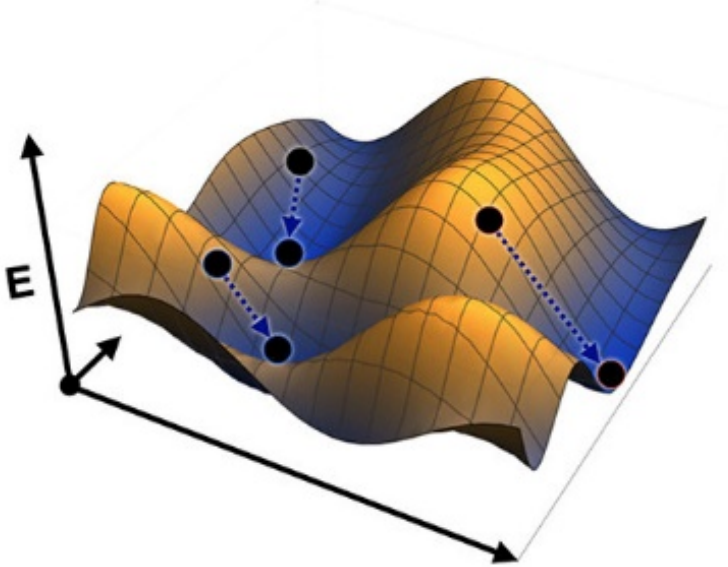
Funnel, superbasin…



Global optimization methods : when nothing or little about the structure is known

(a)

Random Searching

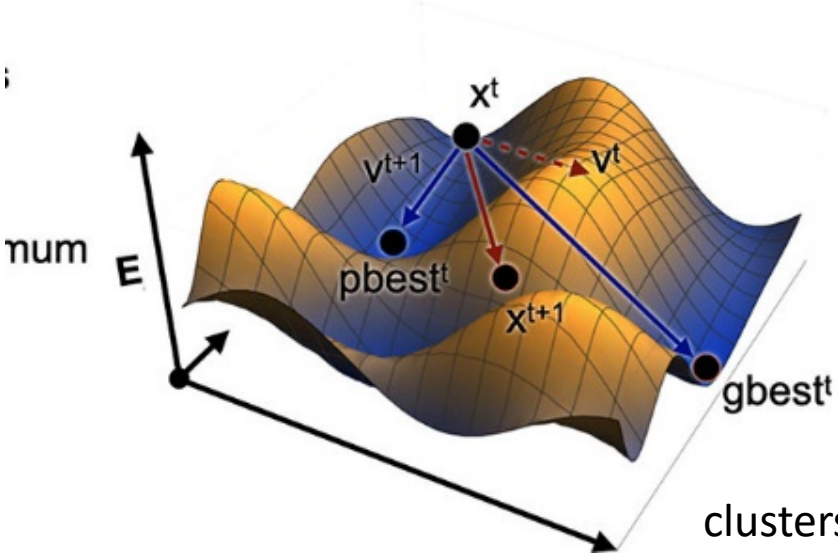


Battery materials



(e)

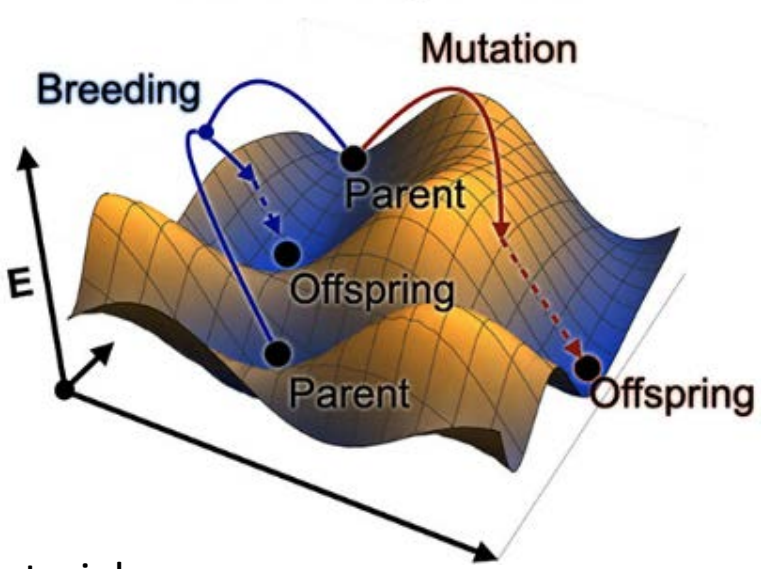
Particle Swarm Optimization



CALYPSO

(f)

Genetic Algorithm



clusters,
2D layered materials,
solids, electrides
and superhard
compounds.

USPEX Computational Materials Discovery

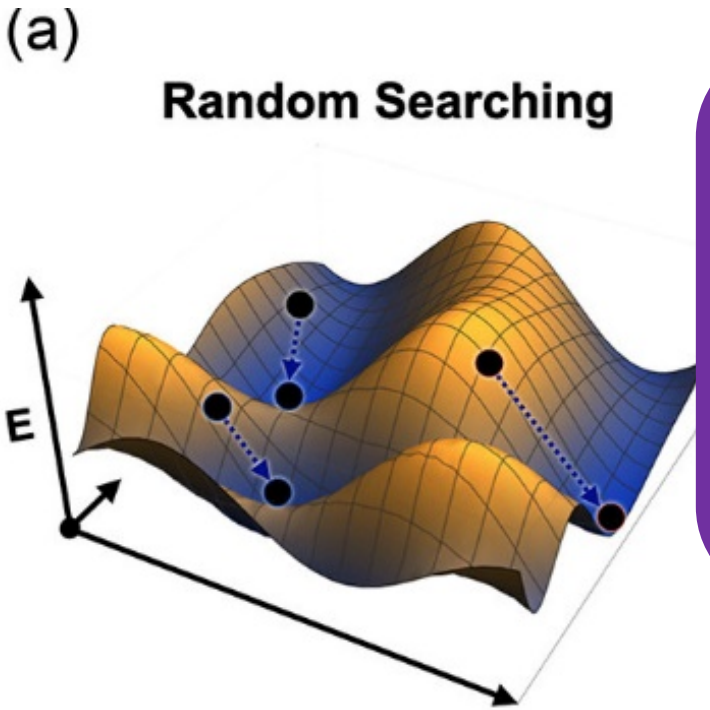
MUSE

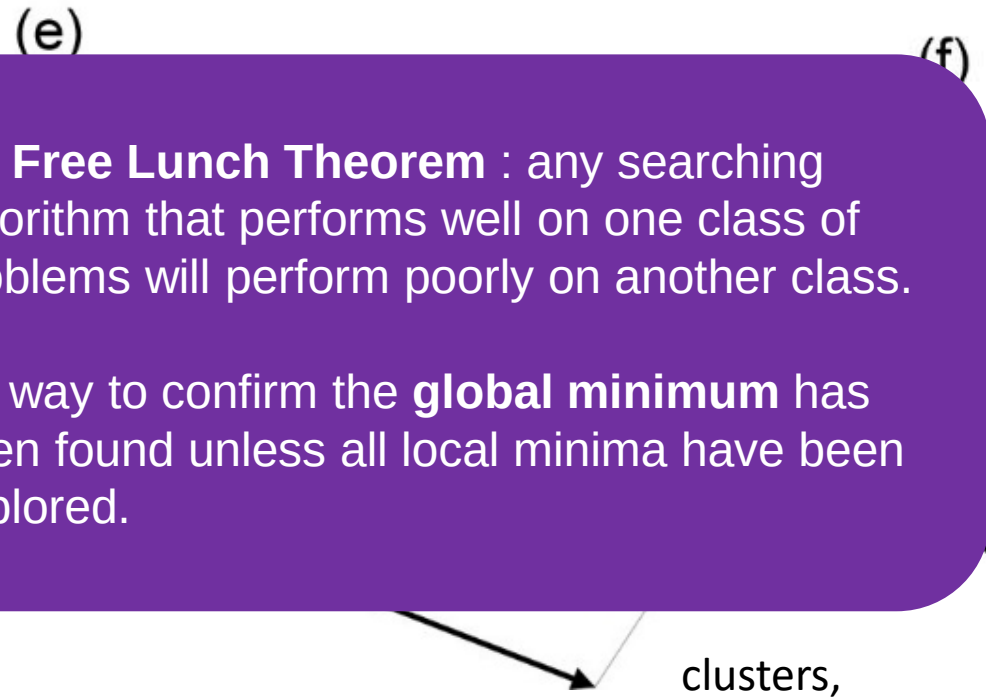
XtalOpt Evolutionary Crystal Structure Prediction

gasp EVO

Z. Falls, P. Avery, X. Wang, K. P. Hilleke, and E. Zurek, J. Phys. Chem. C 2021, 125, 1601−1620

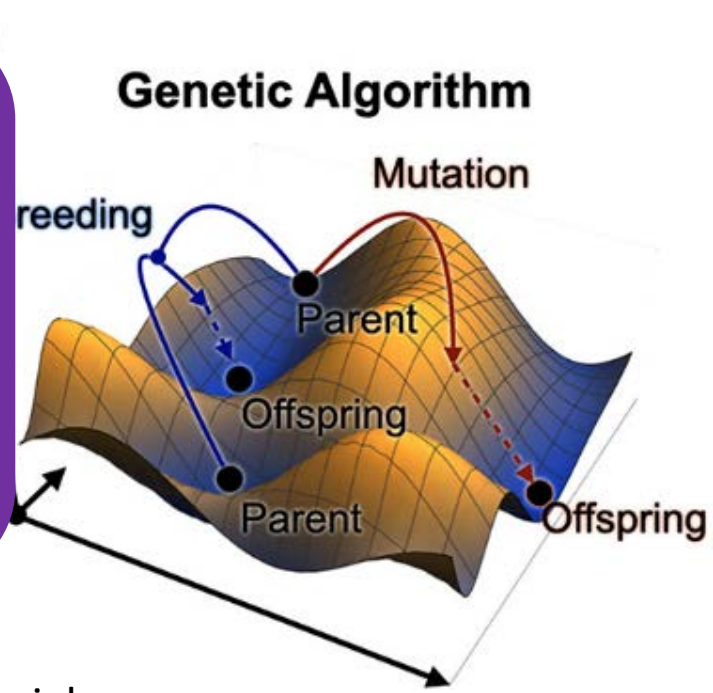
Global optimization methods : when nothing or little about the structure is known





No Free Lunch Theorem : any searching algorithm that performs well on one class of problems will perform poorly on another class.

No way to confirm the **global minimum** has been found unless all local minima have been explored.



Battery materials



clusters,
2D layered materials,
solids, electrides
and superhard
compounds.

CALYPSO

USPEX Computational Materials Discovery

MUSE

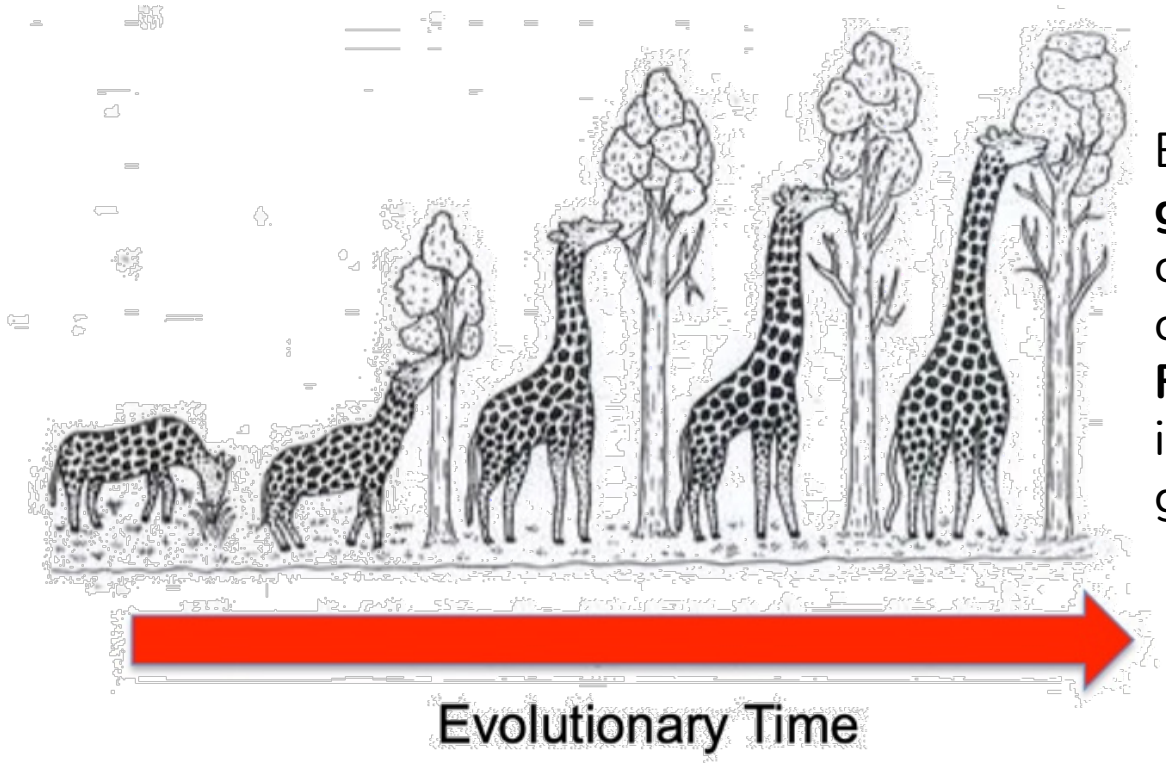
XtalOpt Evolutionary Crystal Structure Prediction

gasp EVO

Z. Falls, P. Avery, X. Wang, K. P. Hilleke, and E. Zurek, J. Phys. Chem. C 2021, 125, 1601−1620

Genetic Algorithms (GAs)

A GA is an 'intelligent' stochastic search method that operates using the principles of Darwinian evolution roughly resumed as "**survival of the fittest**" (1859)

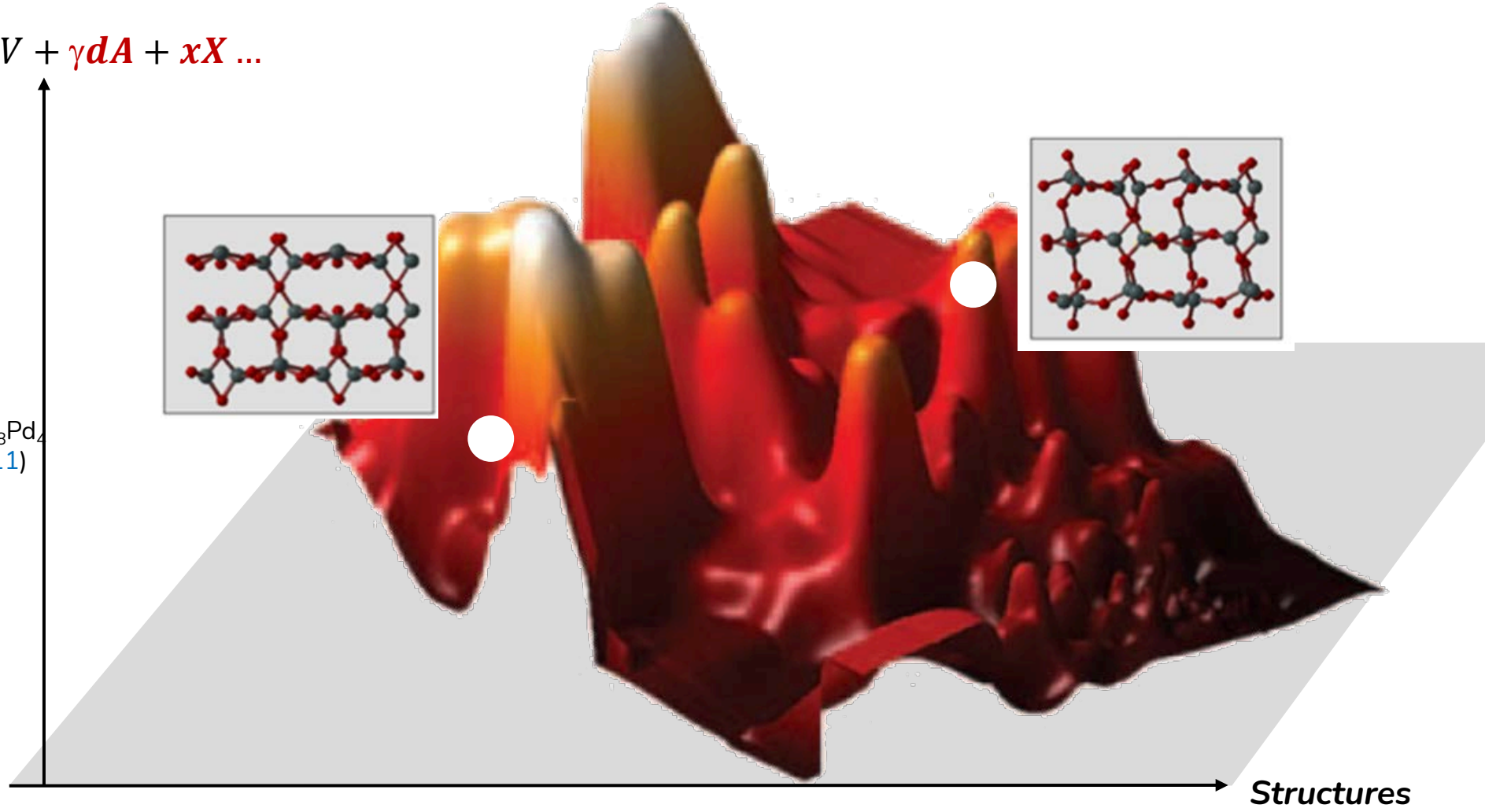


Evolving set (population) of genomes via **successive generations** created by biological **crossover** (i.e., a new, child genome is generated from the complementary parts of two or more parent genomes) and **mutation**.
Fitter individuals have a greater chance of reproducing - i.e. passing on some or all of their genes to the next generation.

Efficiently exploring the rough potential energy surface (PES)

$$G(P,T,\dots) = U - TS + PV + \gamma \boldsymbol{dA} + \boldsymbol{xX} \dots$$

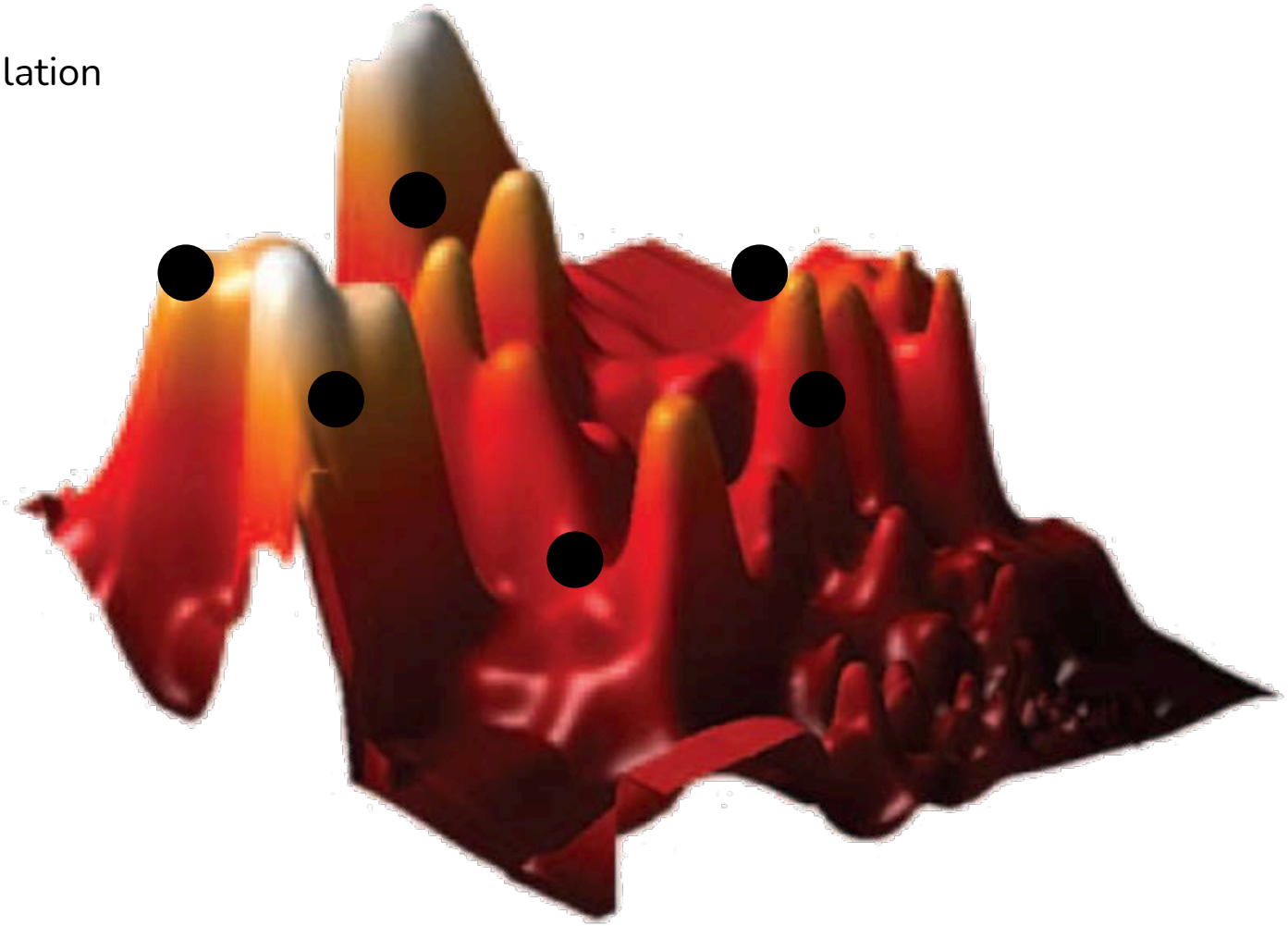
fitness function



2D-projection of the reduced landscape of Au_8Pd_4
(Oganov, Modern Methods of CSP, Wiley, 2011)

Global optimization via GA

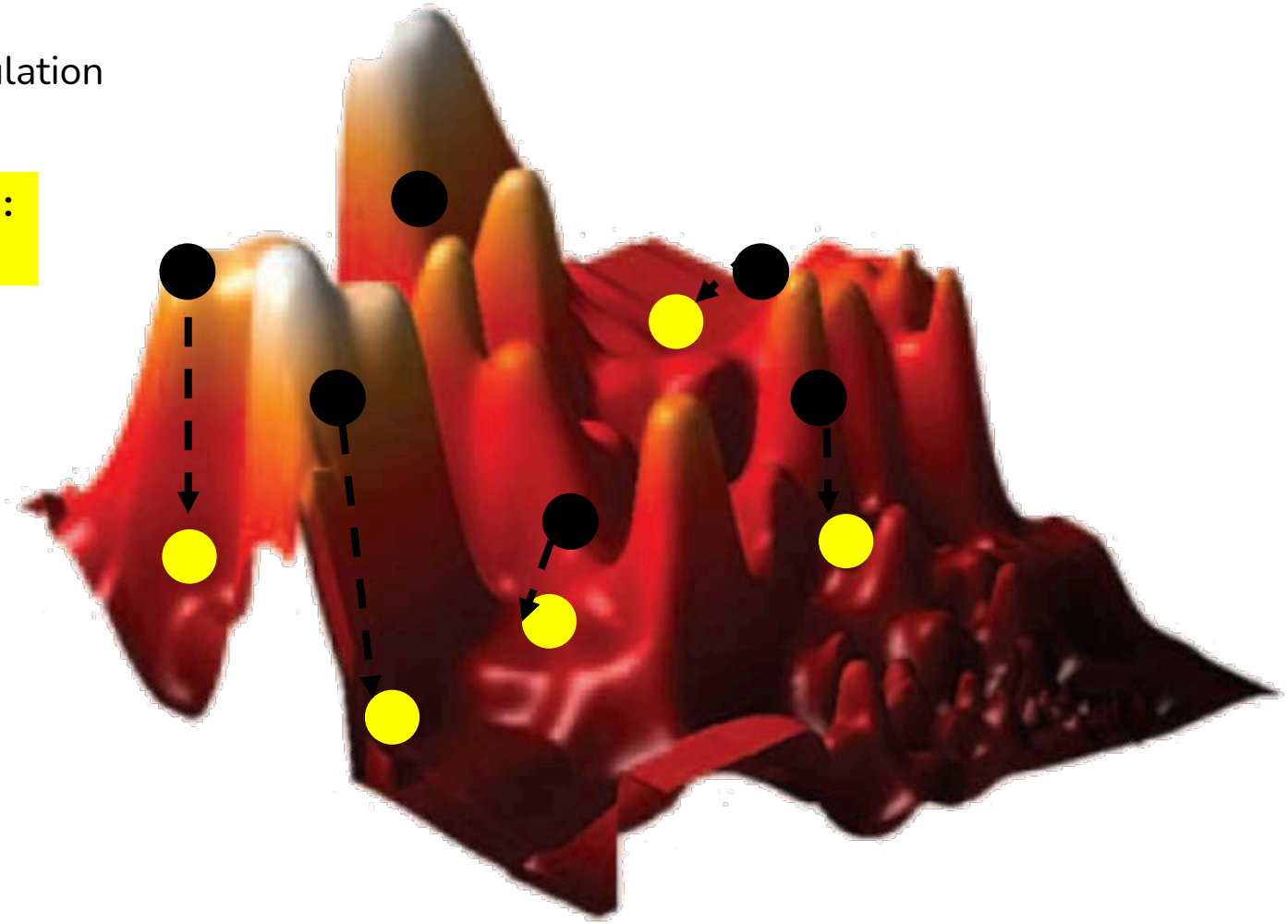
① Start from a generated **randomized** initial population



Global optimization via GA

① Start from a generated **randomized** initial population

② (Local) energy optimization of each structure :
DFT, FF, MLP

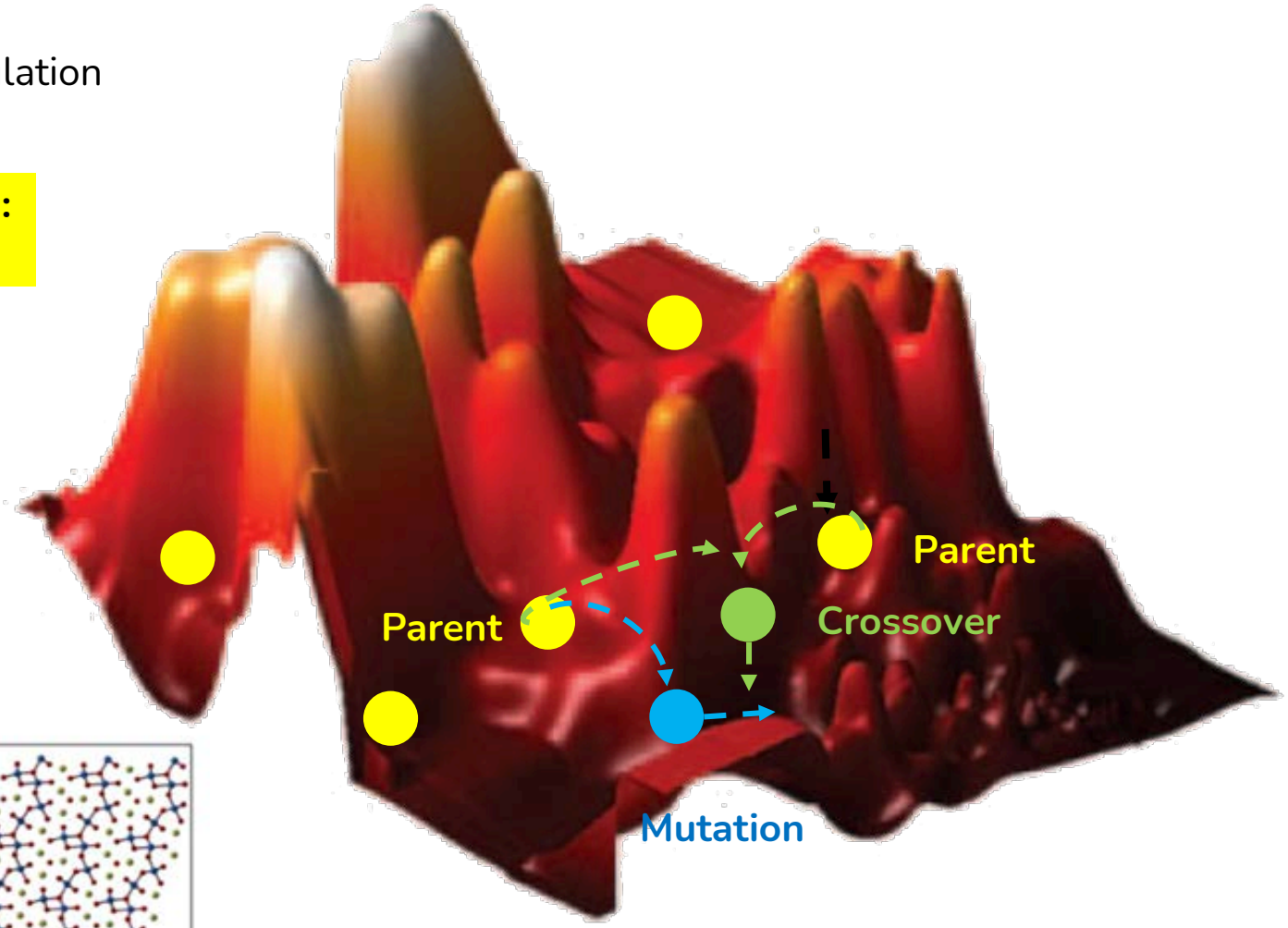


Global optimization via GA

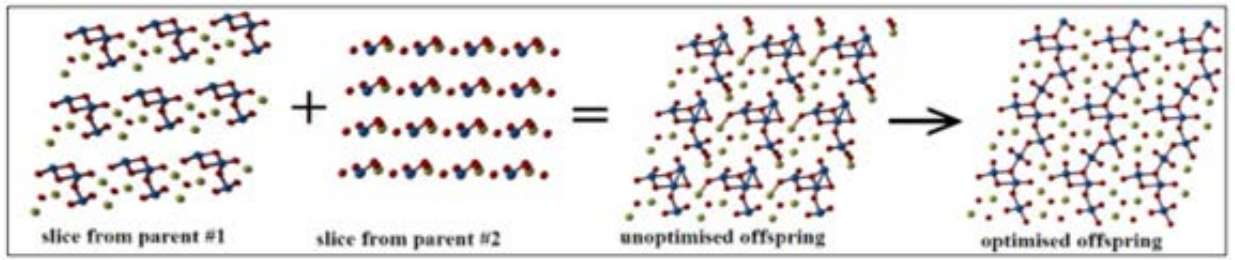
① Start from a generated **randomized** initial population

② (Local) energy optimization of each structure : DFT, FF, MLP

③ Global optimisation (seeking structures)
- Random exploration
- Particle swarm algorithms
- **Genetic/Evolutionary** algorithms



(1) Heredity (crossover)

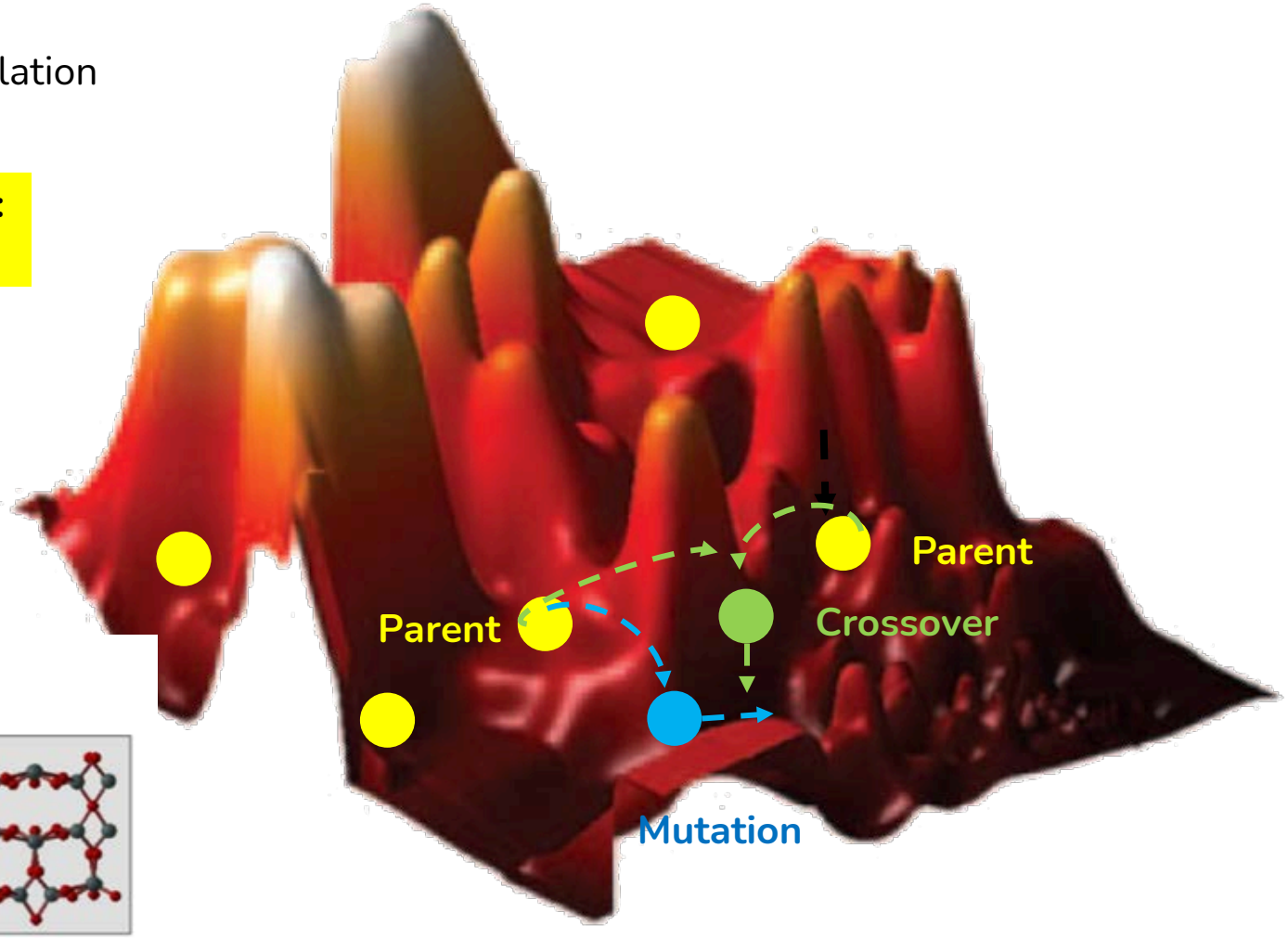


Global optimization via GA

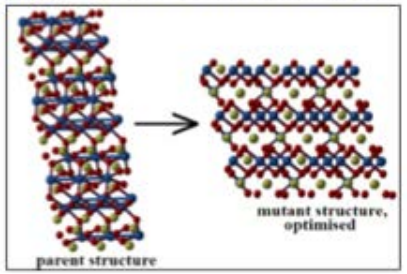
① Start from a generated **randomized** initial population

② (Local) energy optimization of each structure : DFT, FF, MLP

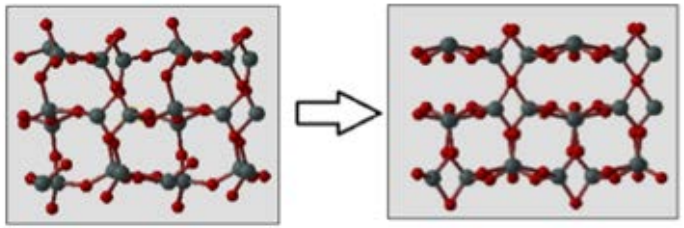
③ Global optimisation (seeking structures)
- Random exploration
- Particle swarm algorithms
- **Genetic/Evolutionary** algorithms



(2) Lattice mutation



(3) Softmutation

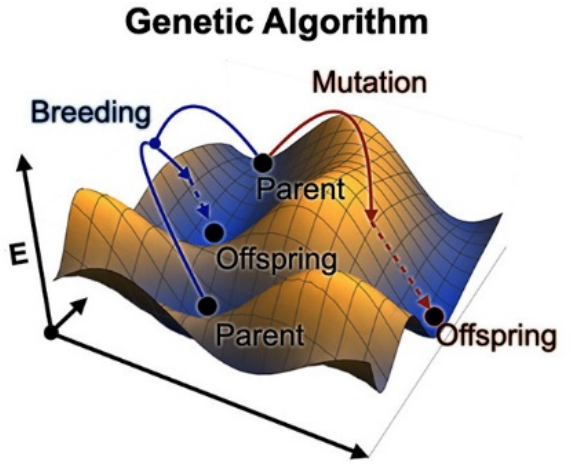


② An optimization method Evolutionary Algorithm Variation Operators

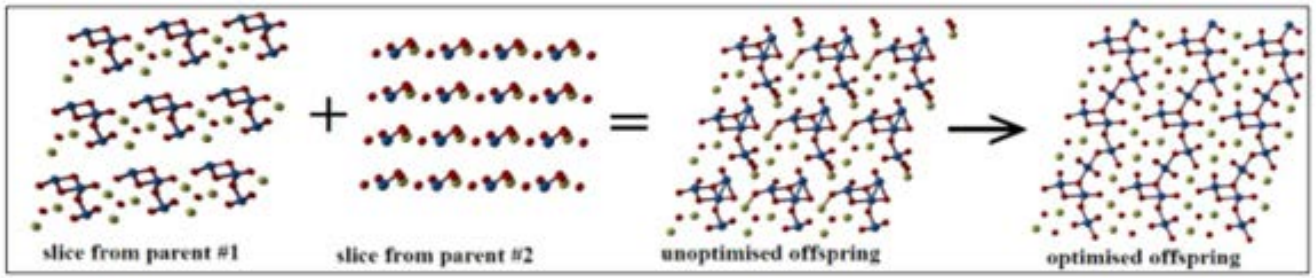
USPEX

Generation of a **Diverse** Population of Structures

« Self-Learning » tool

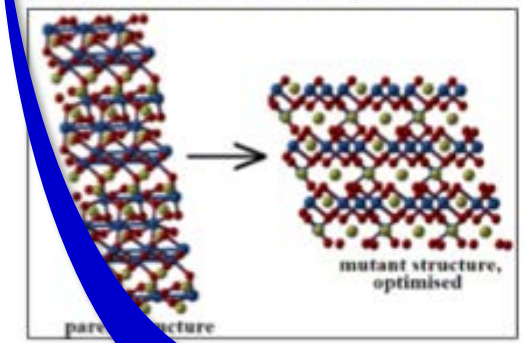


(1) Heredity (crossover)

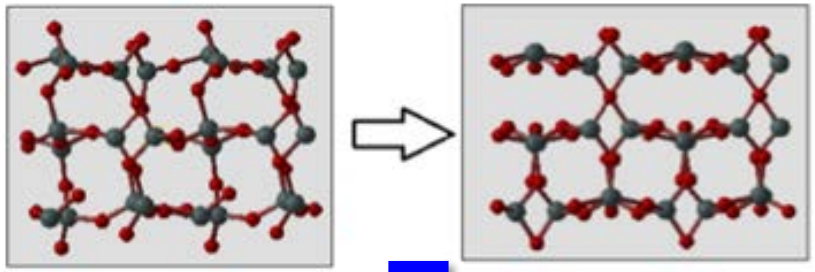


Mutation operators - Enhanced Search

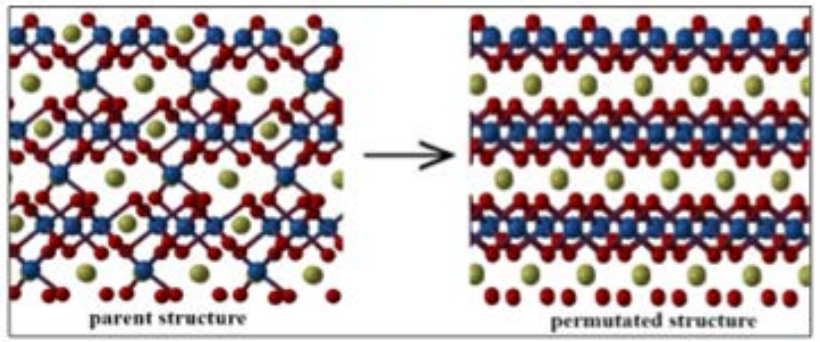
(2) Lattice mutation



(3) Softmutation



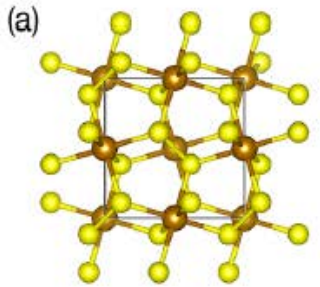
(4) Permutation



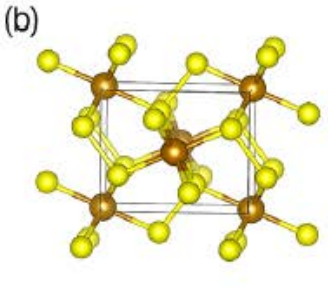
Eugenic selection of the fittest (survivors) and elimination of the others

uspex-te rg

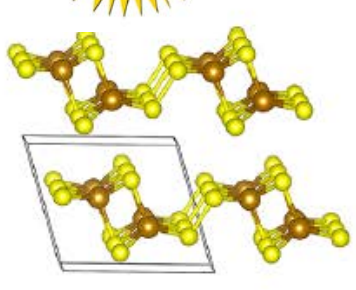
FeS_2 structure search by GA : input =“FeS_2 “



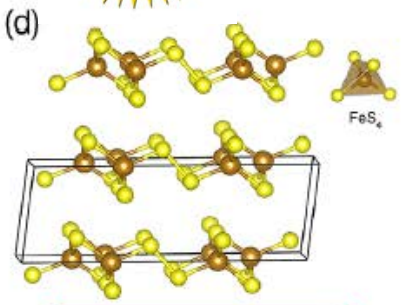
NM Pyrite
ΔE = 0.000 eV/f.u.
(Ground state)



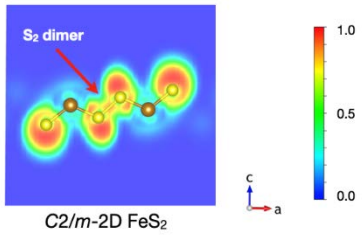
NM Marcasite
ΔE = 0.009 eV/f.u.



AFM *P*-1-3D
ΔE = 0.282 eV/f.u.



AFM *C*2/*m*-3D
ΔE = 0.425 eV/f.u.

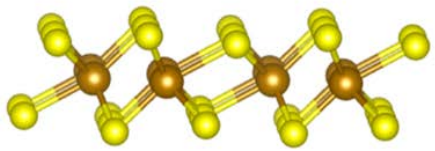


*C*2/*m*-2D FeS_2

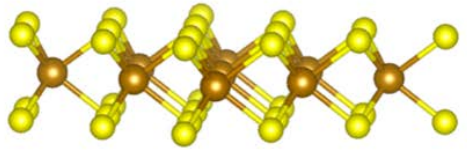
Strongly differ
from classical MS_2 structures

→

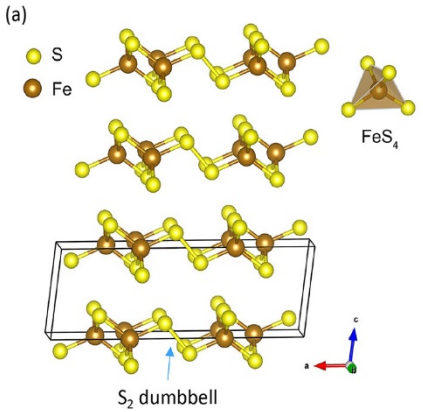


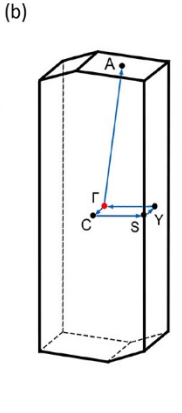


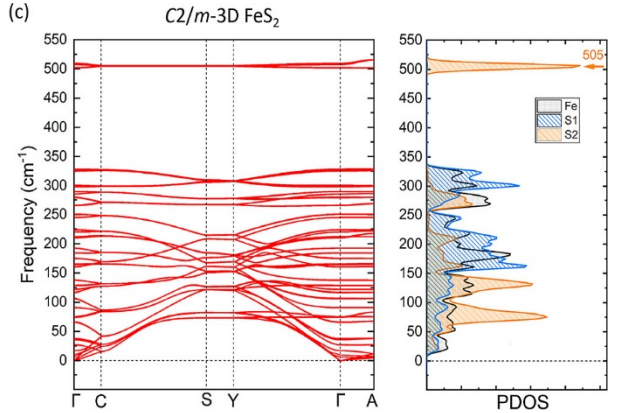
AFM $P2_1/m$ 1T’-2D
ΔE = 0.713 eV/f.u.



FM *P*-6*m*2 1H-2D
ΔE = 1.254 eV/f.u.





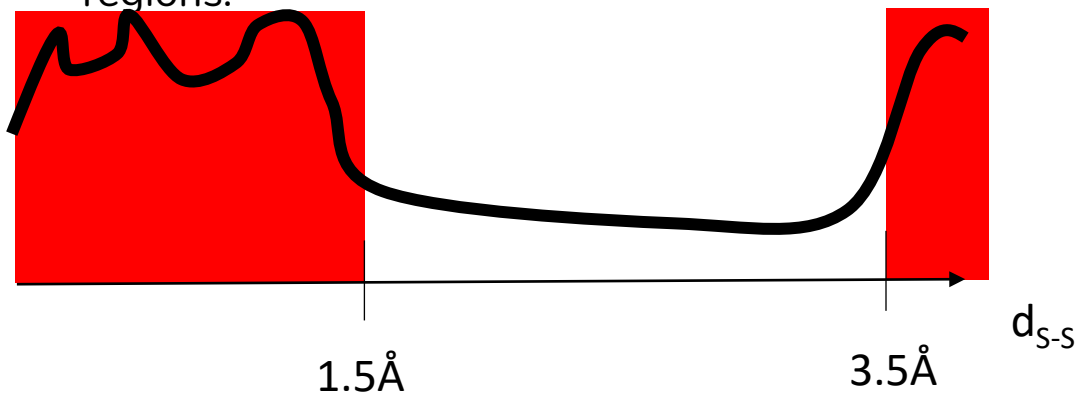


Here displayed Full search
Constrained search penalized by experimental information directly yields to the right structure

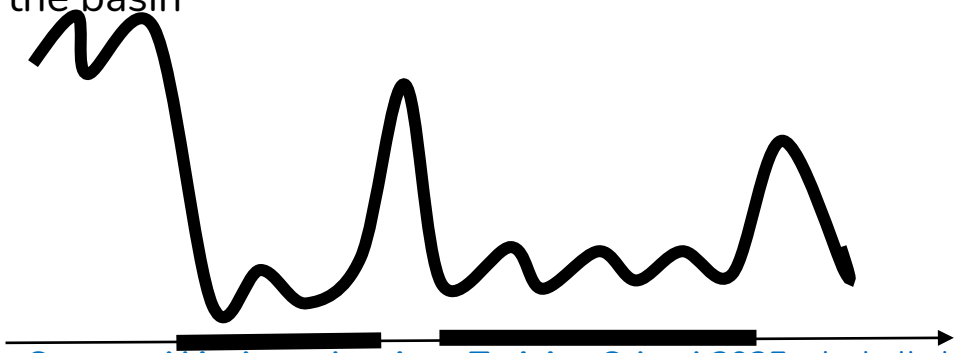
B. Wang, IB, S. Sasaki, G. Guegan, L. Cario, S. Jobic and G. Frapper, *J. Phys. Lett.* **2020**, *11*, 8861.

Why does it work ?

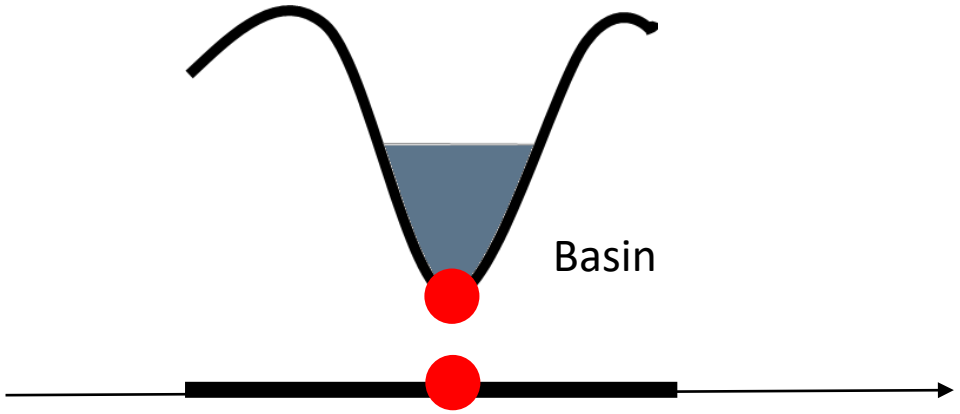
- Only certain regions of the PES are ***chemically relevant* (bond types, lengths..)**
  No need to explore chemically unreasonable regions.



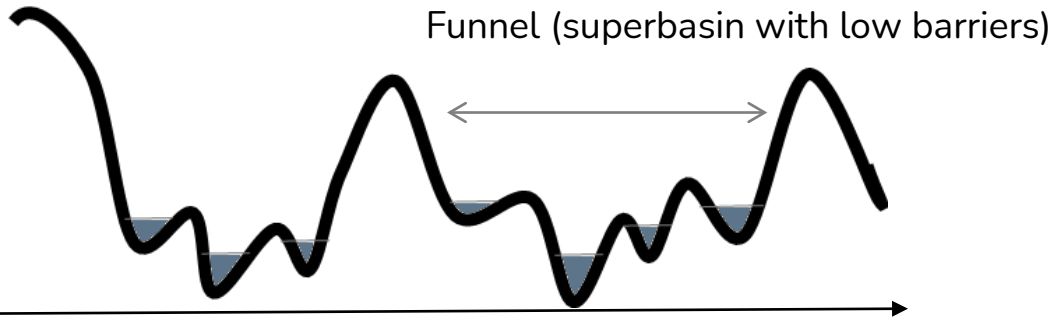
- Low-energy basins occupy ***the largest amount of*** the PES : a randomly generated structure has a high probability to fail in the basin



- ***Relaxation*** decreases the dimensionality of the PES

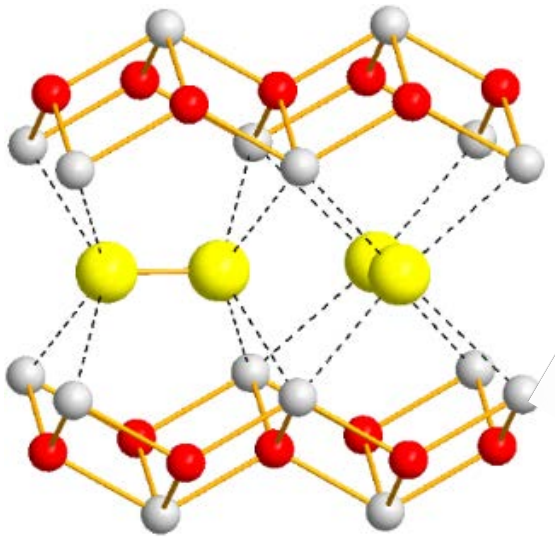


- ***Bell Evans Polanyi principle***: barriers between low-lying minima in a PES are expected to be small. Low-energy basins are likely to be close to one another.



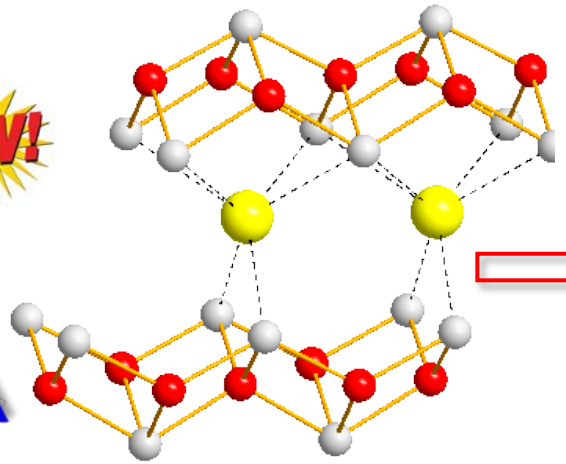
Topochemical S (de)intercalation of $La_2O_2S_2$

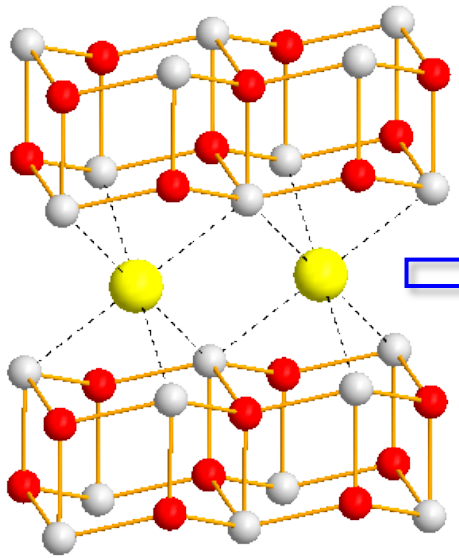
$La_2O_2S_2$

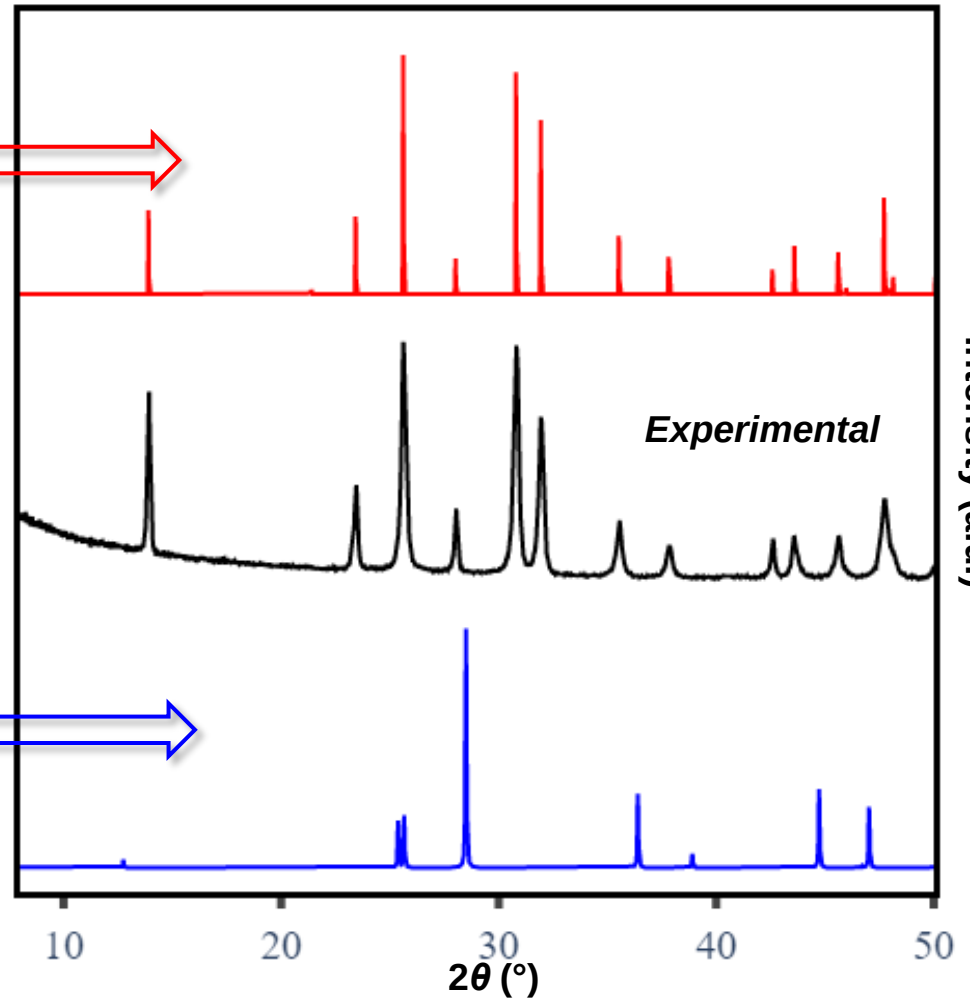


– 1.0 S

NEW!

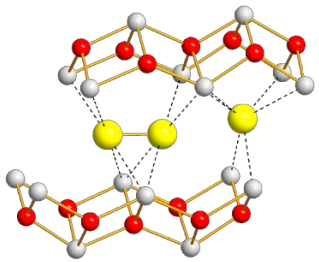






S. Sasaki, M. Caldes, C. Guillot-Deudon, IB, G. Steciuk, L. Palatinus, E. Gautron, G. Frapper, E. Janod, B. Corraze, S. Jobic and L. Cario, *Nat. Commun.* **12**, 3605, (2021)

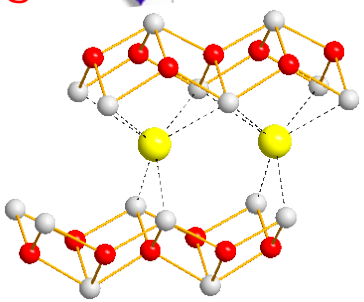
Possible competing reactions



oA - $La_2O_2S_{1.5}$

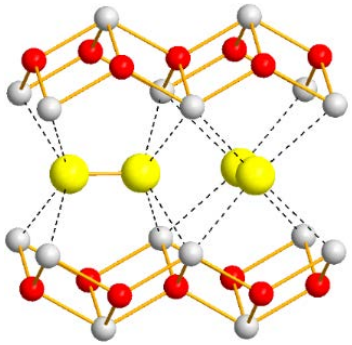
+ 0.5 S

Sulfide anion Topochemistry



oA - La_2O_2S

– 1.0 S

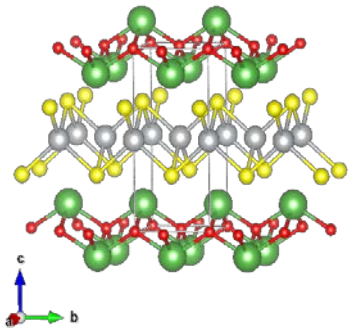


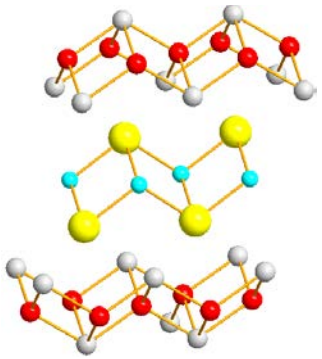
$La_2O_2S_2$

+ $2M^0$
(M = Cu, Fe, Co, Ni)

– M^0
(by Iodine)

Metal cation Topochemistry





$[La_2O_2]^+$

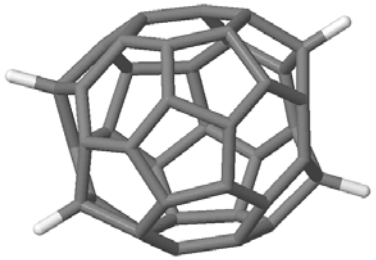
Cu^+

S^{2-}

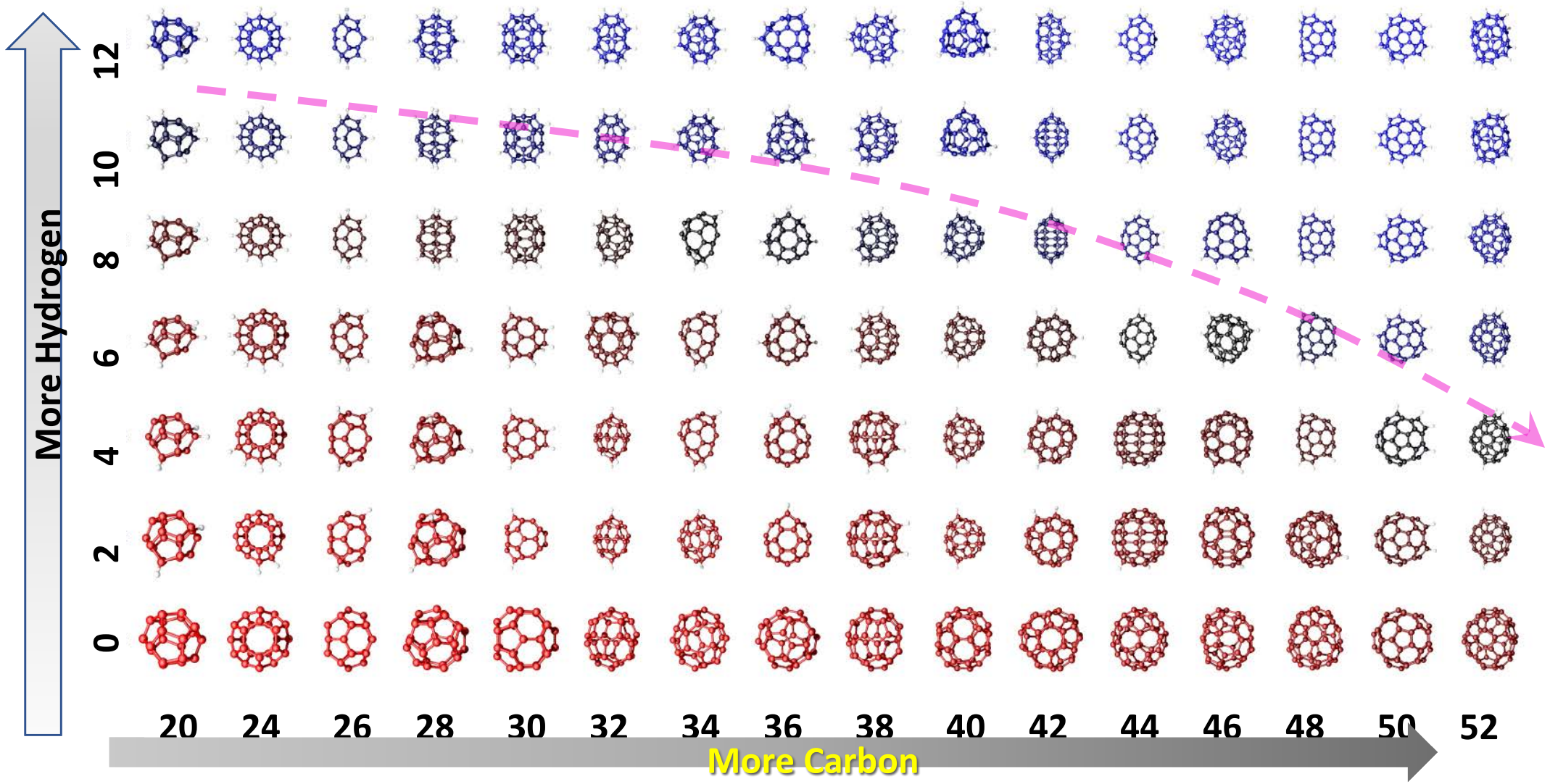
$La_2O_2Cu_2S_2$

Genetic Algorithms (GAs) for molecular search

- Broadly used for chemical search and minimization problems:
  - determination of cluster geometries,
  - protein folding
  - ligand docking
  - molecular design
  - conformational searching of floppy molecules.



  - timely in assisting the experimental characterization of an important class of molecular compounds and can rapidly be expanded toward the a priori prediction of low-lying functionalization isomers of carbon nanotubes, graphene flakes,



See C. Ewels et al : Locating the most thermodynamically stable isomer of an arbitrary functionalized fullerene (private communication). See also J. Chem. Theory Comput. 2012, 8, 1841−1851

Genetic Algorithms (GAs) for molecular search

- **Does not require any derivative information** (which may not be available for many real-world problems).
- Has very **good parallel capabilities**.
- Optimizes both continuous and discrete functions and also multi-objective problems (multimodal optimization)
- Provides a list of "good" solutions and not just a single solution: stable, metastable…
- Always gets an answer to the problem, which gets better over the time.
- Useful when the search space is very large and there are a large number of parameters involved.

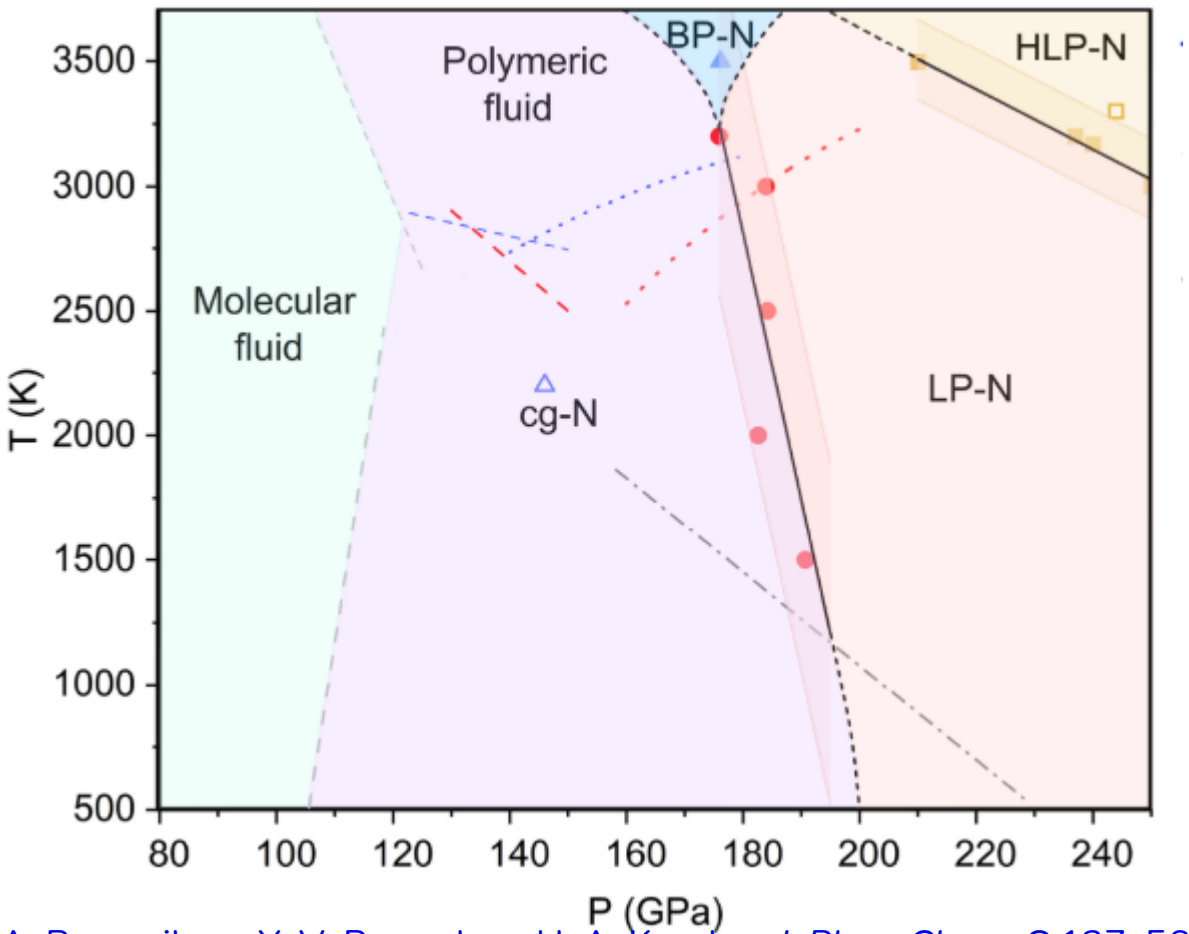
- Fitness value is calculated repeatedly which might be computationally expensive for some problems (DFT, MLP)
- Being stochastic, there are no guarantees on the optimality or the quality of the solution.

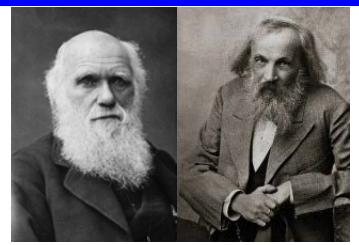
- **Meta-GGAs (**J. Chem. Theory Comput. 2012, 8, 1841−1851)

Time for merge ?

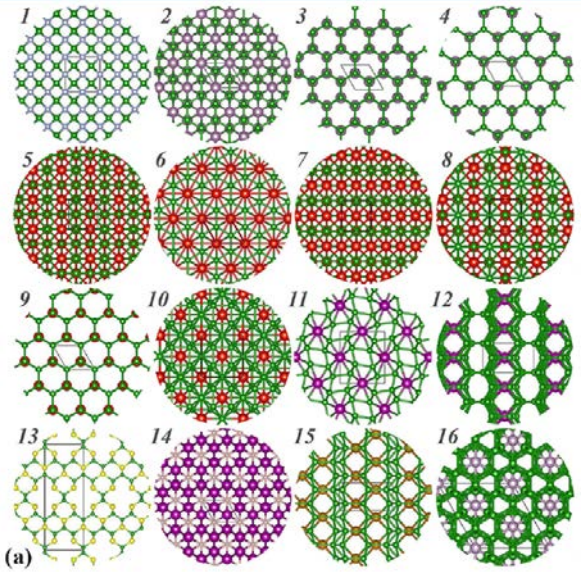
Time to merge

(P,T) prediction : MLIP potential – T-USPEX





Mendelevian Search : Pseudo Periodic Table Exploration by combining Data Mining and evolutionary algorithms



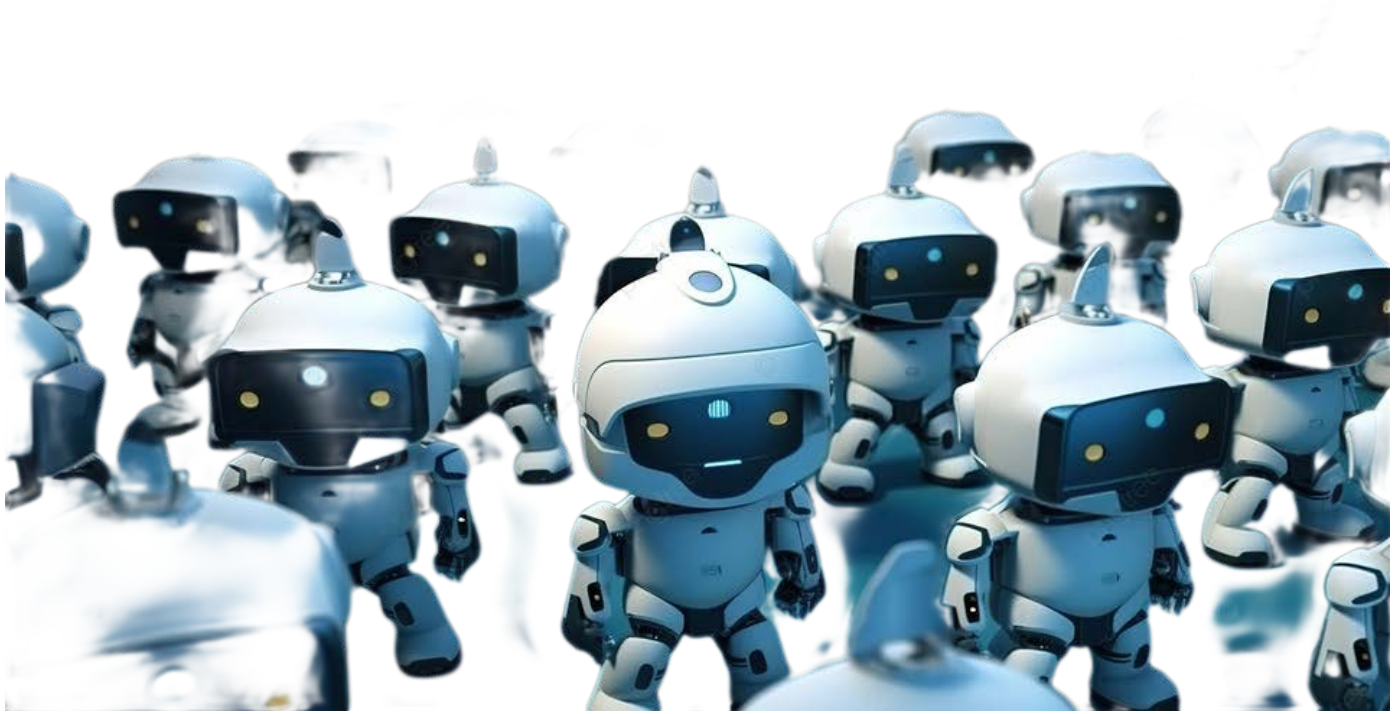
4950 binary systems,

161,700 ternary systems,

3,921,225 quaternary systems

Z. Allahyary and AR Oganov,
npj Computational Materials **6**, 55 (2020)

L. A. Bereznikova,Y. V. Propad, and I. A. Kruglov, *J. Phys. Chem.* C 127, 5683–5688 (2023)



Future is **(no) more** AI?

Mathematical optimisation : no PES but a proof…

Reformulate as a global optimization problem **with constraints**

$$\min E \; s.t. \{C_i = 1\}$$

Allows **non intuitive designs that are very easily transferable**

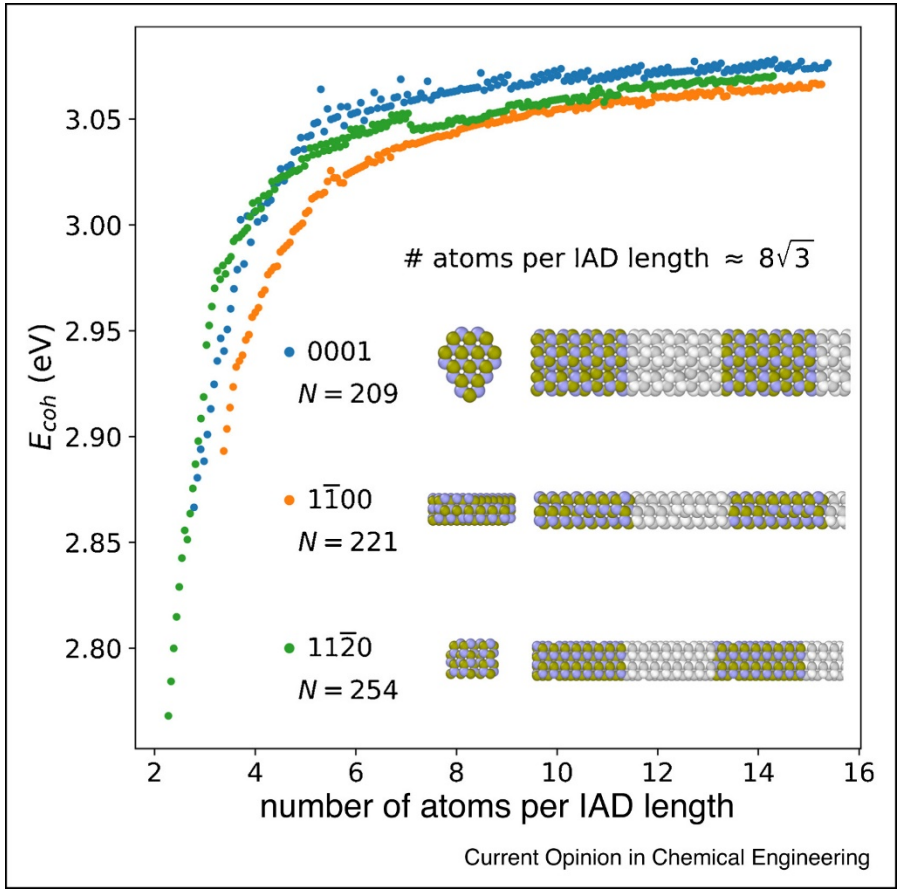
- Non linear or (MILP) programming
- Current algebraic form not (yet) consistent with first principle calculations or ML models
- Binary decisions + assembling blocs (*MatOpt*)

X. Yin and C. E Gounaris, *Current Opinion in Chemical Engineering*, **35**:100726 (2022)

Formal solution integrates combinatorial and **continuous** optimization, but its practical application hinges on **quantum computing** (not yet within immediate reach)

Gusev, V. V. et al. Optimality guarantees for crystal structure prediction. Nature 619, 68–72 (2023).

Nanofil InAs



cartoonstock



Take aways

Many stabilities (mechanical, dynamical, thermal…) : post-analysis is mandatory

Dynamical stability (phonon dispersion curve)

Mechanical, Thermal stability

Chemist's analysis : ELF…

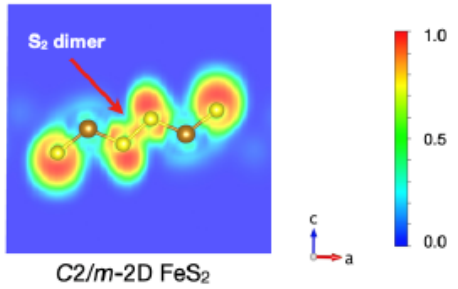


Figure S6. Electron Localization Function (ELF) for the 2D *C2/m* FeS_2 single layer.

Some properties or spectra cannot be directly targeted and are computed afterwards

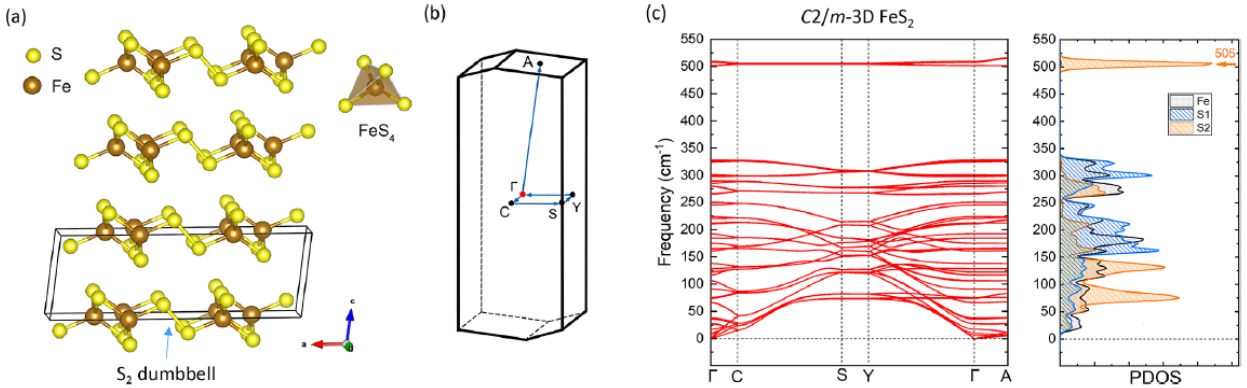


Figure 2. Layered metastable *C2/m*-3D FeS_2. (a) Crystal structure and enlightened FeS_4 tetrahedra and S_2^{2-} dumbbells. (b) Highlighted Brillouin zone for *C2/m*-3D and a selected k path. (c) Phonon dispersion curve (left) and phonon DOS (right). S_1 indicates the sulfur belonging to the FeS_4 tetrahedron, and S_2 indicates the sulfur belonging to the S_2^{2-} dimer. The arrow indicates a peak at 505 cm^{-1} which is attributed to S–S stretching modes of S_2^{2-} dimers (blue).

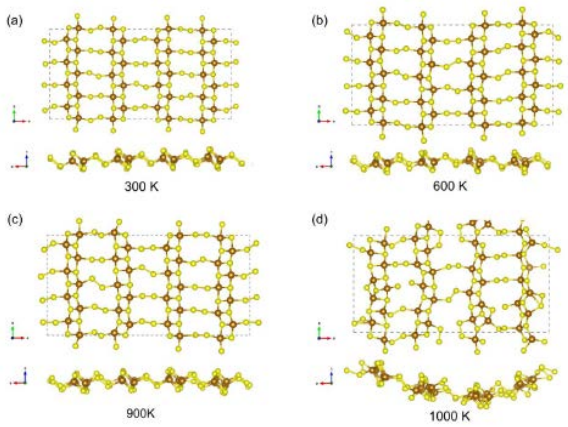
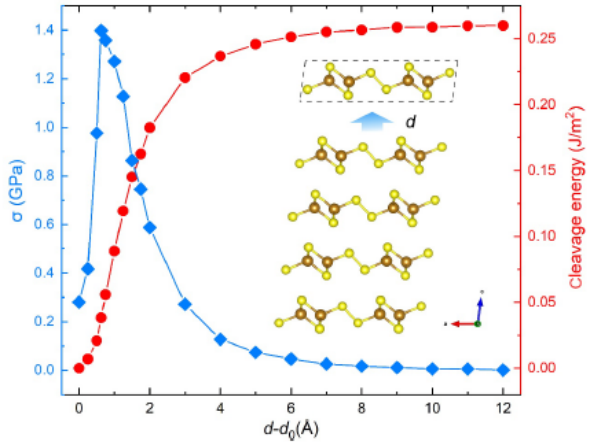
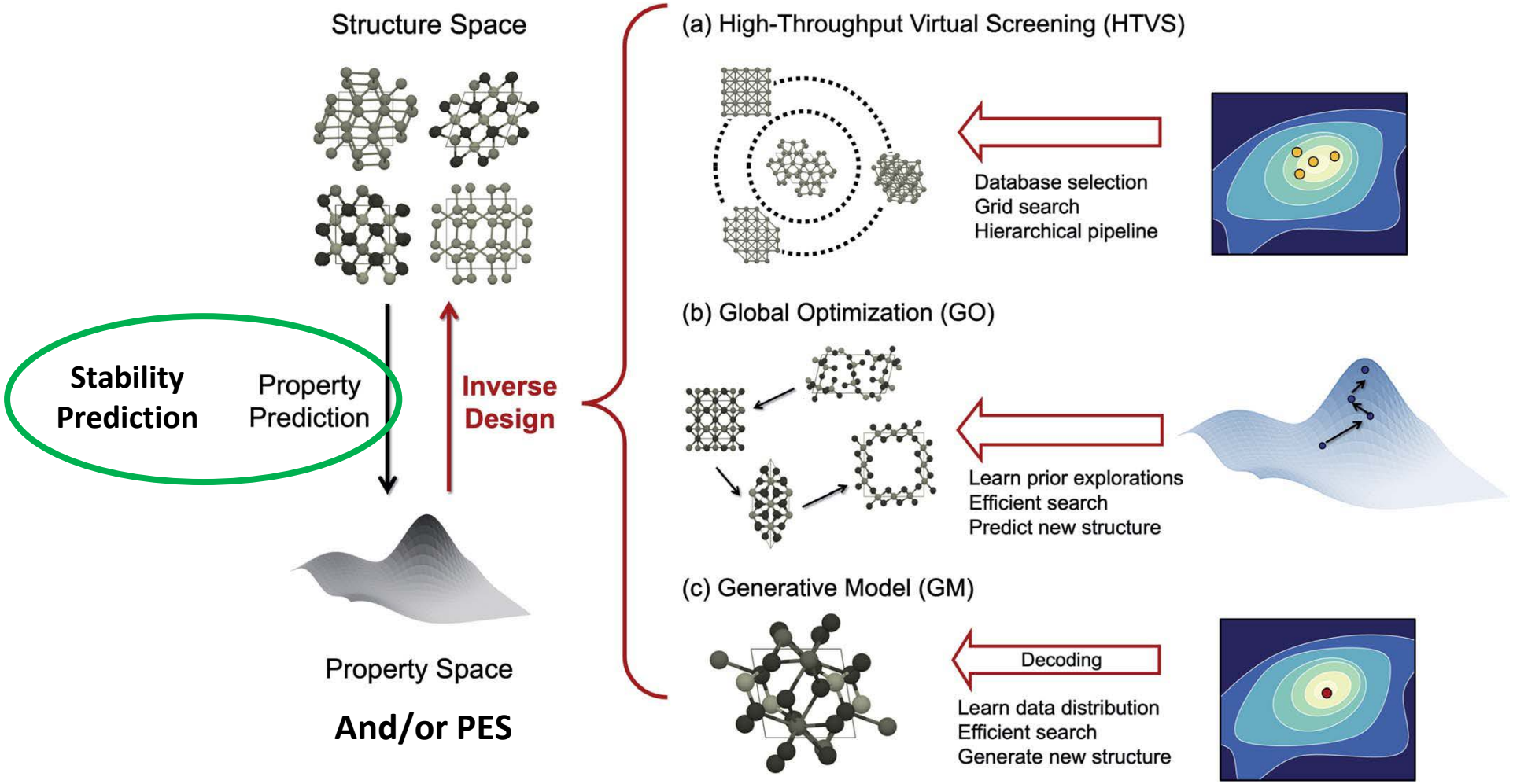


Figure S7. Snapshots of the 2D *C2/m* FeS_2 2 × 3 × 1 supercell at the end of 10 ps AIMD simulations at (a) 300 K, (b) 600 K, (c) 900 K and (d) 1000 K. Large brown spheres are iron atoms, and small yellow spheres are sulfur atoms.



(inverse) Structure prediction strongly differs from (direct) property prediction



Adapted from Chem Sci, 2020, **11**, 4871

Worth noting that the latest genetic algorithm based methods can still compete with “today’s” deep learning based methods (2019)

- All high level structure search algorithms (DeepMind) have them implemented at the core as they explore chemical space **in a very efficient manner**
- **Wide-range benchmarks still lack**
- GM mostly focuses on stable structures, metastable structures is a natural by-product of GO
- GM mostly treat molecules and bulks; GO can deal with **molecules, bulks, surfaces, interfaces, clusters**….

- GM : Lack of Guidelines to evaluate significance of these “discoveries” : synthetizability, reaction formation path and conditions, decomposition
- still great need of human expertise to validate and interpret AI propositions

Acknowledgements

Funded by the European Union

- This work is based upon work from COST Action CA21126 - Carbon molecular nanostructures in space (NanoSpace), supported by COST (European Cooperation in Science & Technology).

- COST (European Cooperation in Science & Technology) is a funding agency for research and innovation networks. Our Actions help connect research initiatives across Europe and enable scientists to grow their ideas by sharing them with their peers. This boosts their research, career and innovation.

- www.cost.eu
- https://www.cost.eu/actions/CA21126/
- https://research.iac.es/proyecto/nanospace/



This work was performed using HPC resources from GENCI-IDRIS