NUS National University of Singapore | Institute for Functional Intelligent Materials

CONSTRUCTOR UNIVERSITY

Machine learning: bridging scale gap between the worlds of materials and particles

Andrey Ustyuzhanin

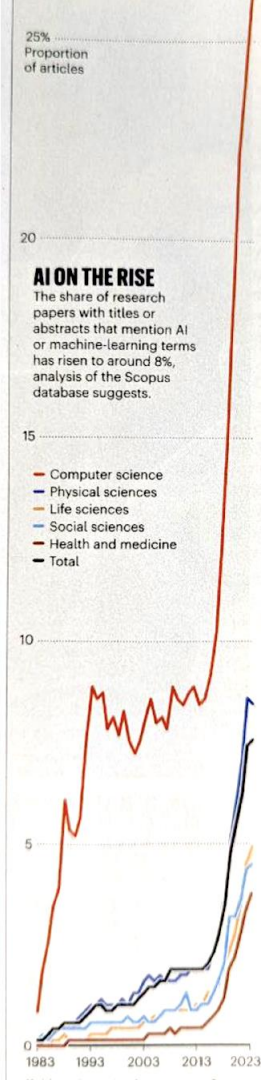
31.10.2023

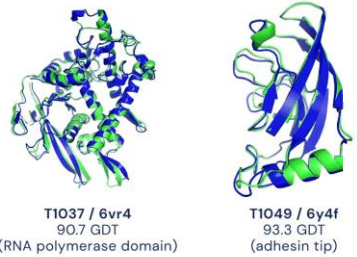
Self intro

- Computer Science →
- Data Science for industry →
- Data Science for Particle Physics (CERN, LHCb, CMS, OPERA, …)
  - 7 schools of Machine Learning, online course on ML for Particle Physics

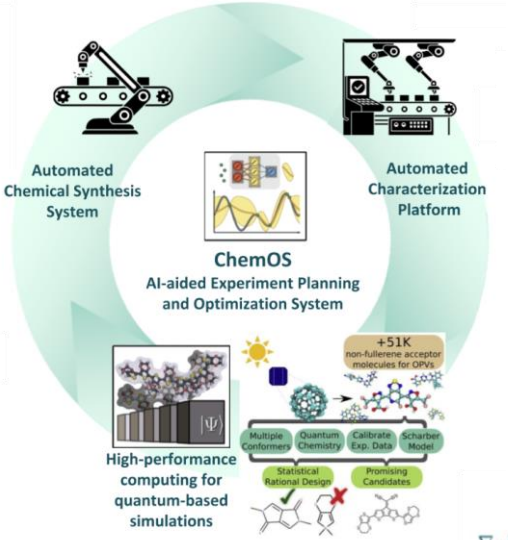
  →
- Data Science for Material Science @Institute of Functional Intelligent Materials, NUS, Singapore @Constructor University Bremen

Late news: AI is taking over the science

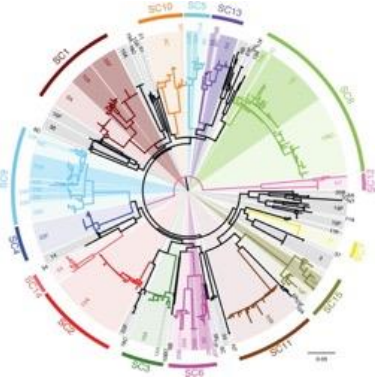


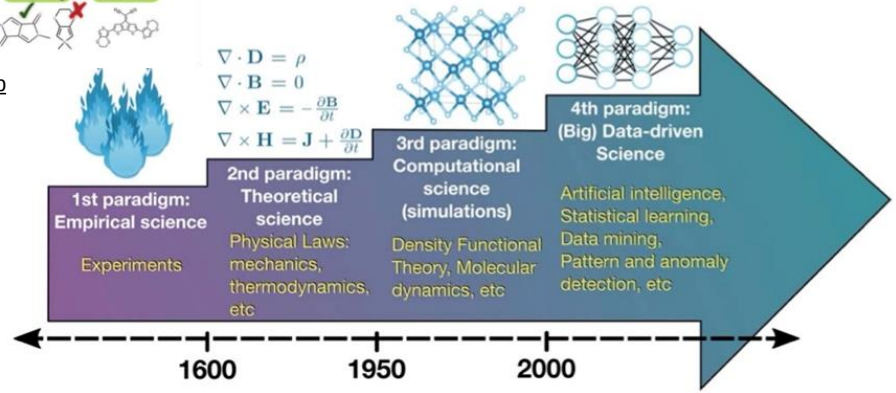


Two examples of protein targets. AlphaFold predicts highly accurate structures measured against experimental result.



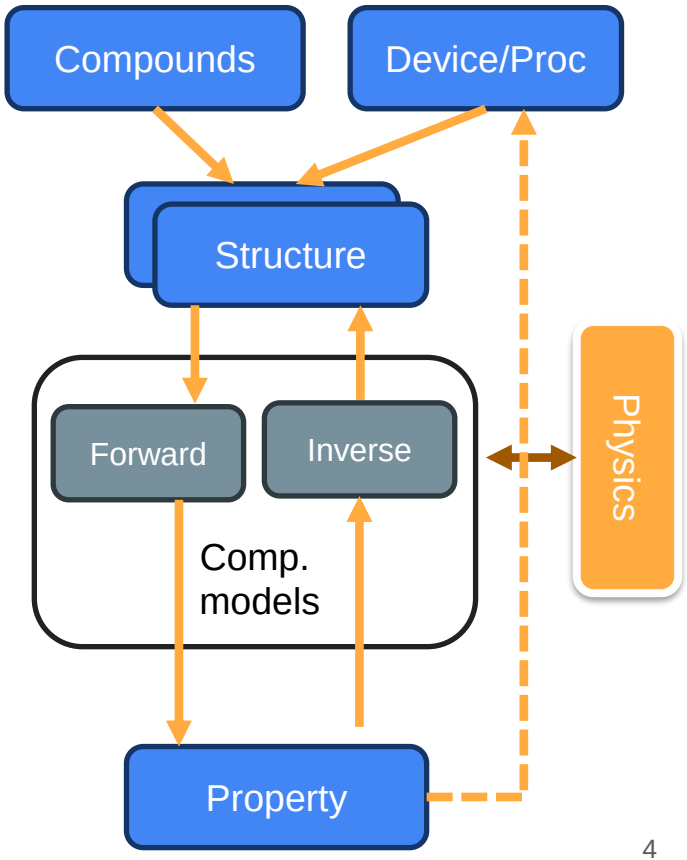
The matter lab



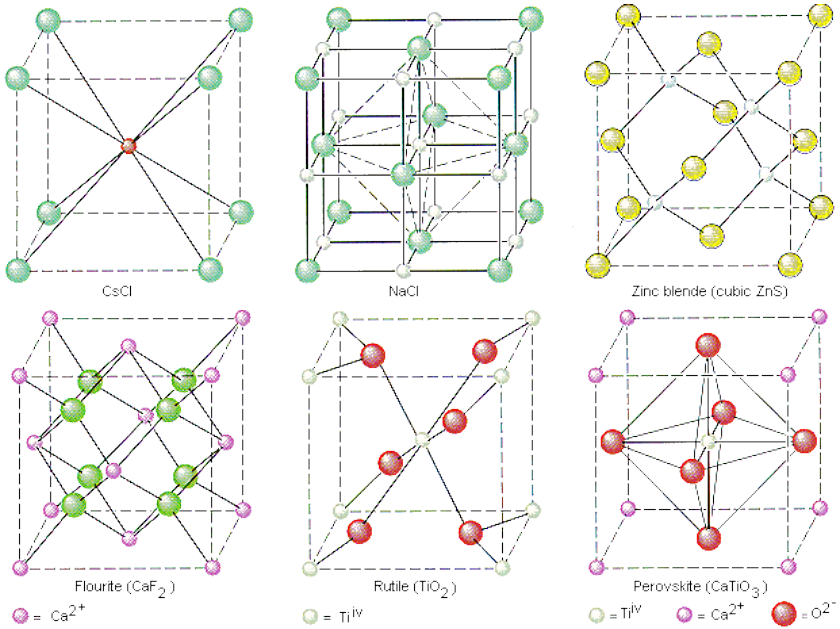


Material science in a nutshell

- What if I combine atoms like this? (static, dynamics properties)
  $\mathbf{X} \rightarrow \mathbf{Y}$
- How should I mix atoms / materials to get desired property $\mathbf{Y}$?
  $\mathbf{Y} \rightarrow \mathbf{X}$
- What is the most optimal design/process for device $\mathbf{W}$ with materials $\mathbf{X}$ that would be optimal for $\mathbf{Z}$?
- What model would be most useful for describing properties $\mathbf{Y}$ given materials like $\mathbf{X}$?



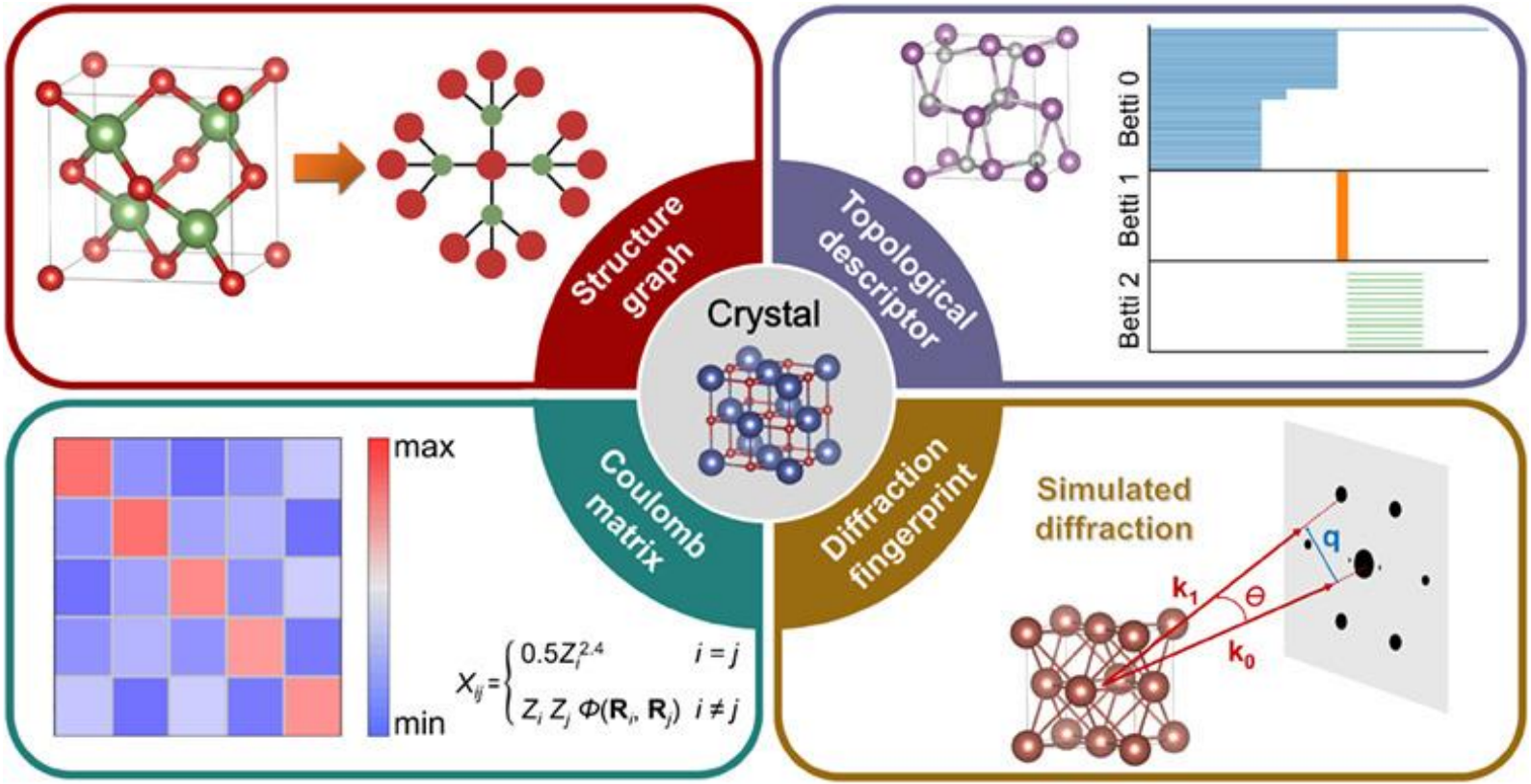
Basic building block: a crystal cell



A unit cell:
- minimal set of atoms representing crystal geometry (types and coordinates)
- set of bonds between atoms
- periodic boundary conditions
- 230 space groups, each having point groups

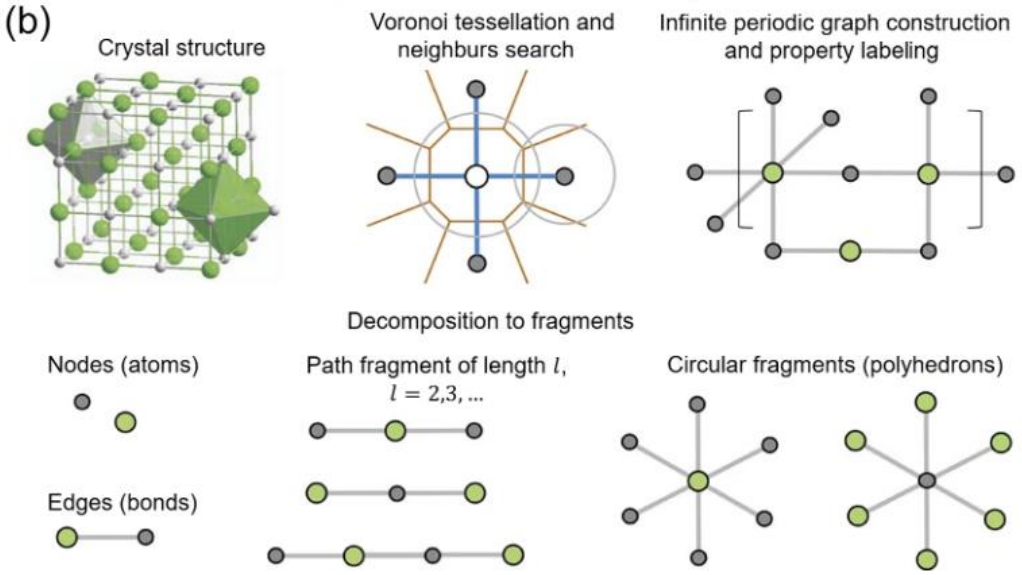
https://bit.ly/46ToXDq

Crystal structure representation for ML

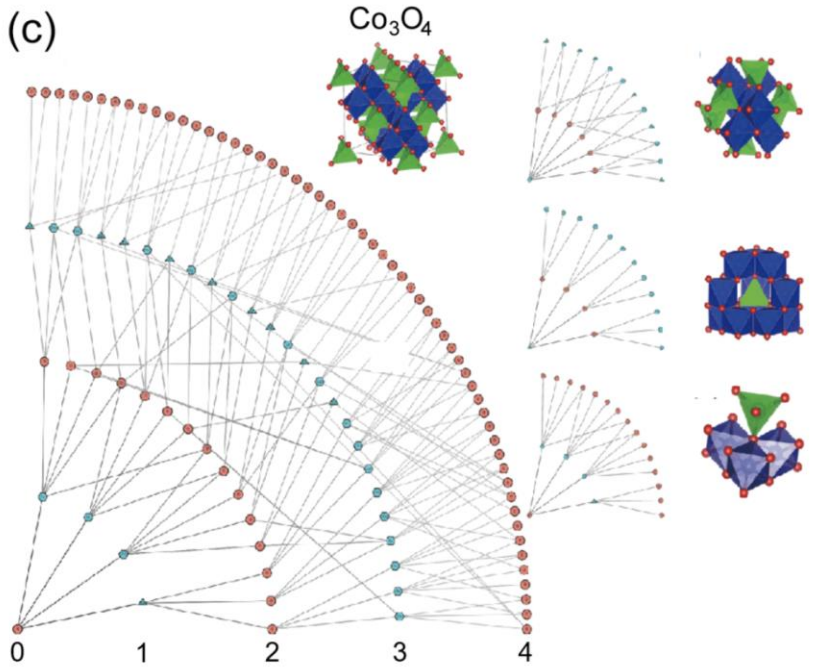


Structure graph representation

https://go.nature.com/3MpOyMe

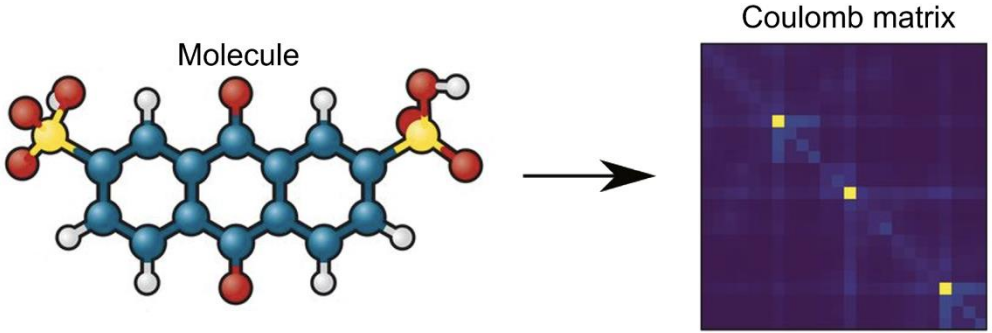


https://bit.ly/3Mnl6Xa



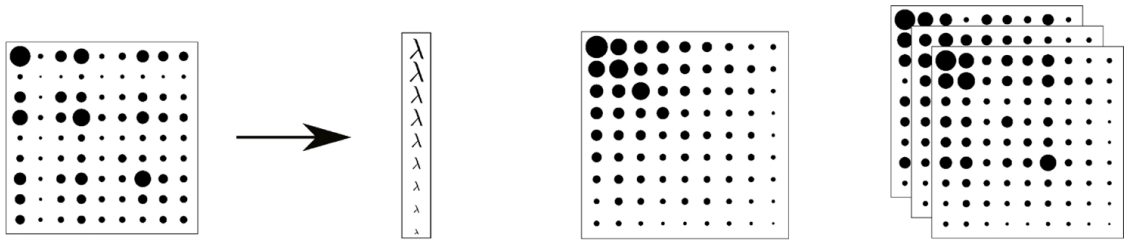
(left) Schematic illustration of the construction of property-labeled materials fragments descriptors (right) The subgraphs of spinel Co_30_4

Coulomb matrix



$$M_{ij} = \begin{cases} 0.5Z_i^{2.4} \text{ for } i = j \\ \dfrac{Z_i Z_j}{|\mathbf{R}_i - \mathbf{R}_j|} \text{ for } i \neq j \end{cases},$$

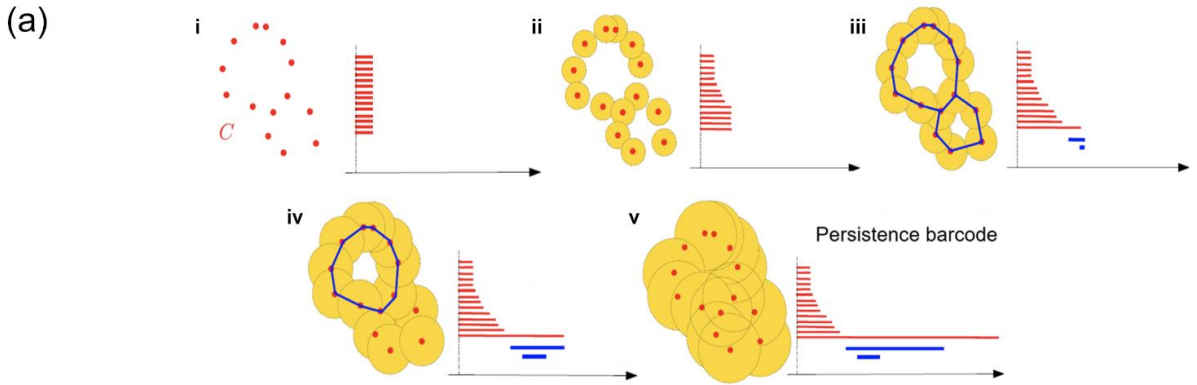
https://doi.org/10.1126/science.aat2663



https://pubs.acs.org/doi/abs/10.1021/ct400195d

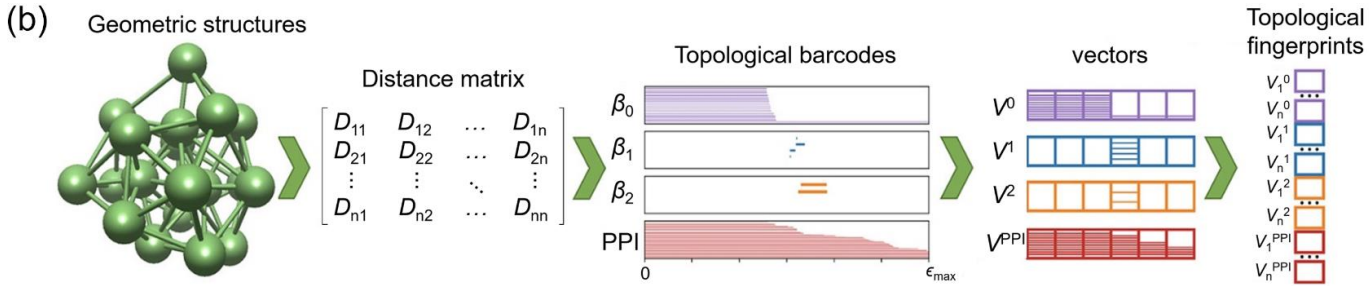
For periodic crystals: various extensions can be applied (e.g., Ewald sum and sine tensors)

Topological descriptors



https://pubs.acs.org/doi/abs/10.1021/ct400195d

The filtration of the distance function to a point cloud and construction of persistent barcodes

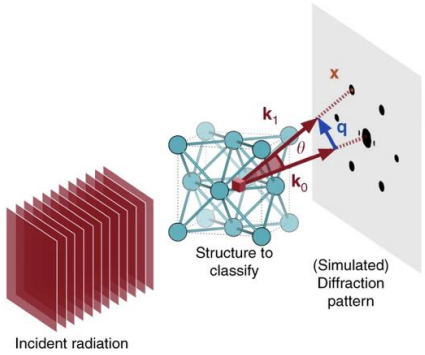


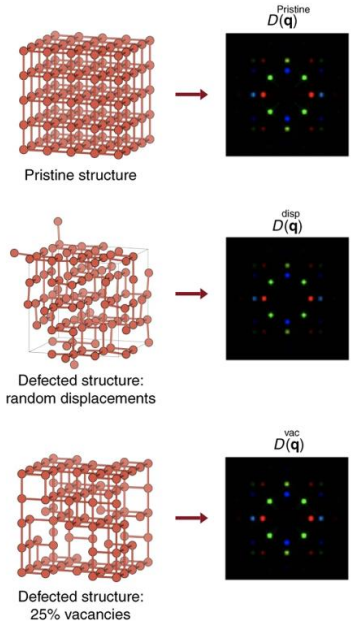
Construction of topological fingerprint of a Li cluster

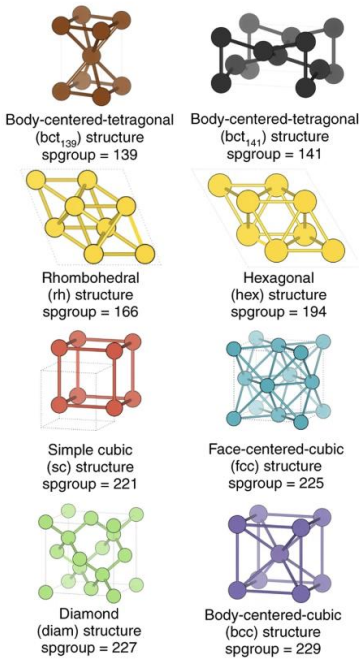
https://pubs.acs.org/doi/abs/10.1021/acs.jpclett.0c00974

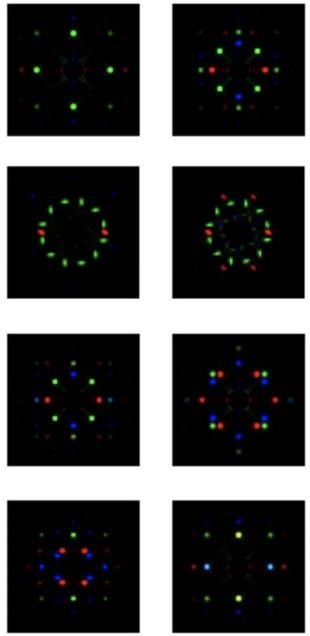
2D diffraction fingerprints

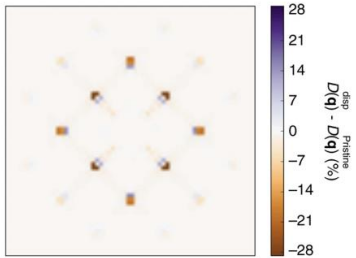
- Compiled with information of periodicity and symmetry.
- Discriminative of crystallographic classification.
- Constrained with limited element information.
- Incapable of describing the atomic interactions.

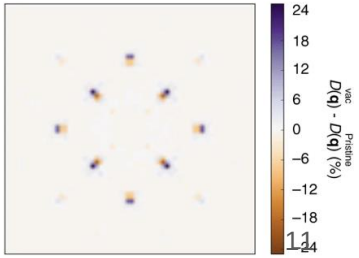












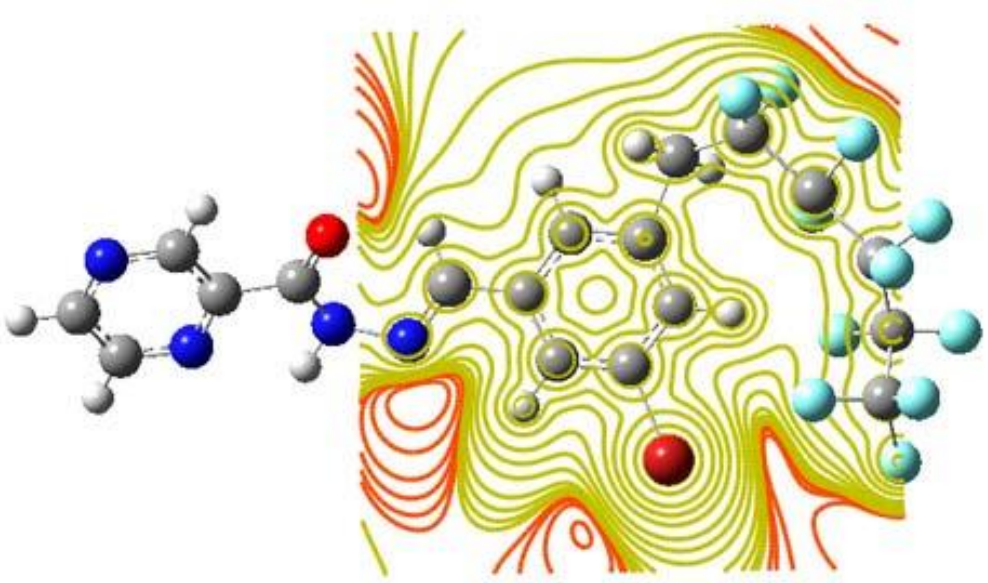
https://go.nature.com/46V3vhE

Datasets

Database Name	Description
NIST Materials Genome Initiative (MGI)	Several databases for different material classes
The NIMS Materials Database (MatNavi)	Polymers, inorganic material, metallic material and computational electronic structure
The Novel Materials Discovery (NOMAD) Laboratory	Input and output files from more than 100 million high-quality calculations. It also includes notebooks for several materials informatics problems
Materials Project	Inorganic compounds, nanoporous materials, elastic tensors, piezoelectric tensors, electrode materials
…	Many more at https://github.com/sedaoturak/data-resources-for-materials-science

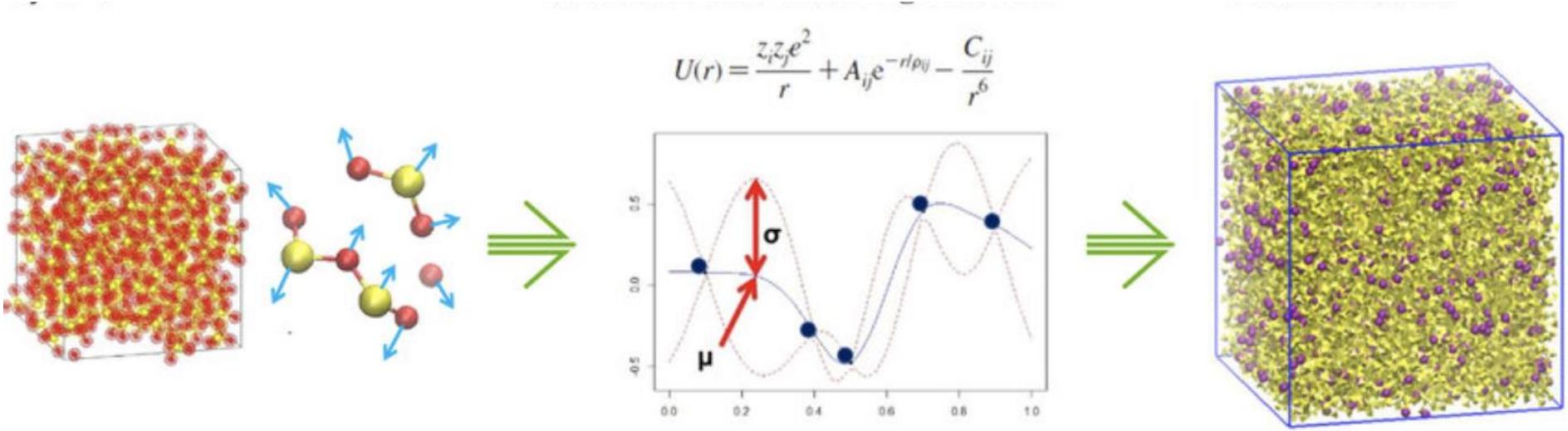
Forward modelling, state of the art

- Instead of solving multi-particle Schrodinger equation,
- Density Functional Theory (DFT) focuses on the electron density as a scalar field, making the computational problem more tractable.
- The central idea is that the ground state energy of a quantum system can be expressed as a functional of the electron density
- Large space for ML augmentation
  e.g., https://arxiv.org/pdf/2309.15127.pdf, https://openreview.net/forum?id=aBWnqqsuot7.



http://notes.fluorine1.ru/public/2014/5_2014/letters/letter2.html

Scaling to larger systems

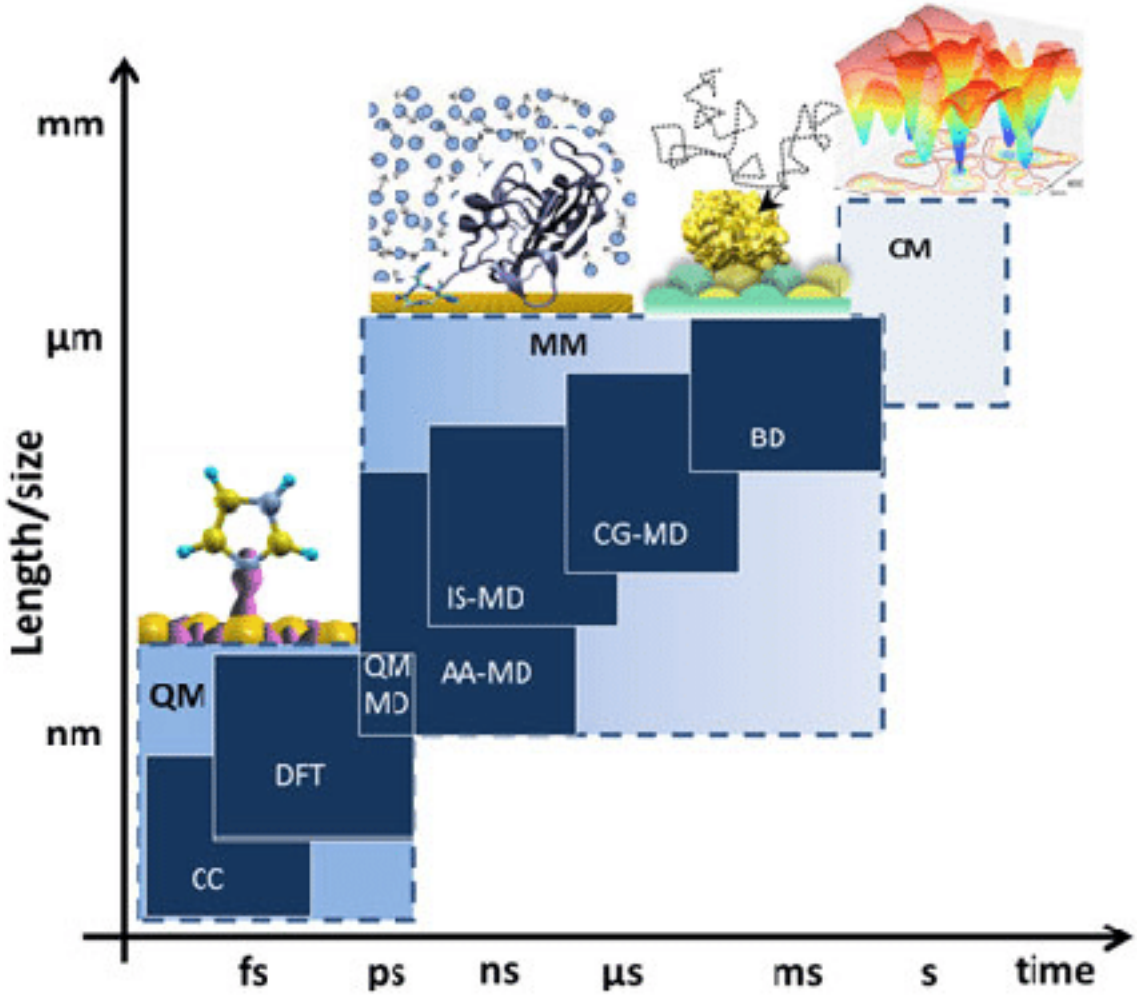


- From small unit-cells one can estimate interatomic potential/forces for conducting large-scale molecular-dynamics (MD) simulations

https://bit.ly/40eoQQB

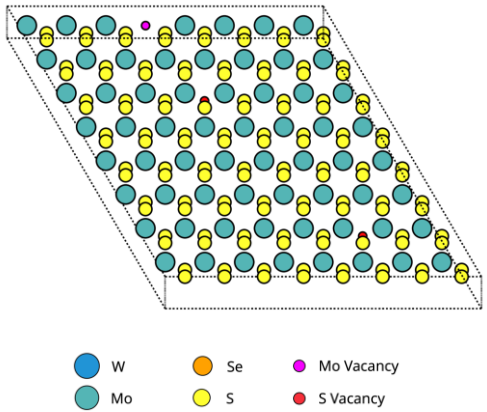
Bigger picture

Time and length scales of different simulation techniques: quantum mechanics (QM), including coupled cluster (CC) and DFT methods, molecular mechanics (MM), and the Brownian dynamics (BD) technique; and continuum mechanics (CM).
Demand for surrogate multiscale modelling.



Forward model for 2D materials

Inputs: unrelaxed 2D material structures with point defects



An example MoS_2 structure

Properties predicted:

- Defect formation energy
- HOMO – LUMO gap

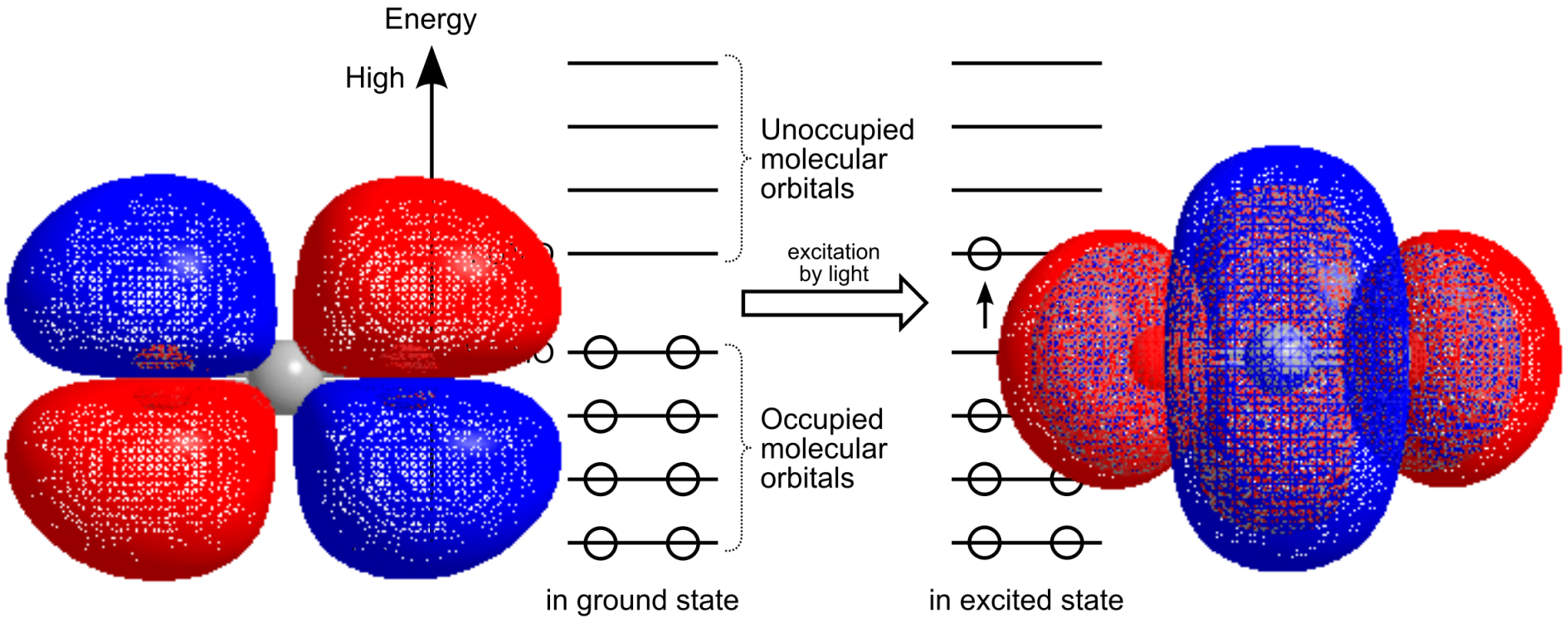
Objective: be 1000 times faster than DFT to allow configuration screening and inverse design

Caution: Mathematically, a set of atoms and their coordinates, but with peculiar symmetries:

- Permutation invariance
- Translation invariance
- Rotation invariance
- Locality of interactions
- Variable number of atoms in a set

"Challenges" for naïve ML – but opportunity for inductive bias!

HOMO – LUMO gap



https://en.wikipedia.org/wiki/HOMO_and_LUMO

Our 2D materials datasets

High defect concentration dataset

- hBN; v(B), v(N), C(B), C(N)
- InSe; v(In), V(Se), Ga(In), S(Se)
- GaSe; v(Ga), v(Se), In(Ga), S(Se)
- P; v(P), C(P)
- MoS_2; v(Mo), v(S), W(Mo), Se(S)
- WSe_2; v(W), v(Se), Mo(W), S(Se)

v(X) is a vacancy, X(Y) is an X to Y substitution

Total defect concentrations:
2.5%, 5%, 7.5%, 10%, 12.5%

500 structures per material, **3500** in total

Low defect concentration dataset

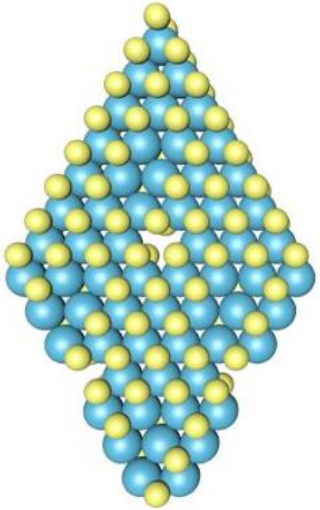
- MoS_2; v(Mo), v(S), W(Mo), Se(S)
- WSe_2; v(W), v(Se), Mo(W), S(Se)

1 – 3 defects

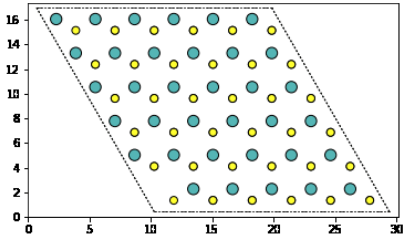
5934 structures with per material, **11868** in total

Huang, P., Lukin, R., Faleev, M. *et al.* Unveiling the complex structure-property correlation of defects in 2D materials based on high throughput datasets. *npj 2D Mater Appl* **7**, 6 (2023).

Project pipeline



Preparation of structures with defects



DFT relaxation and band structure computation

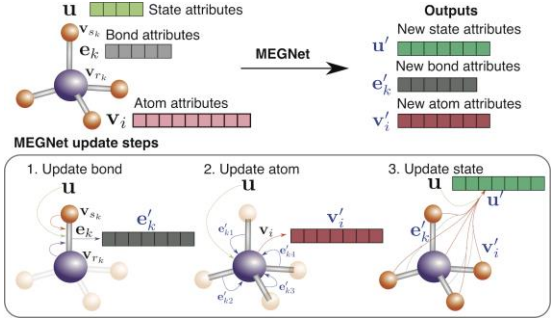
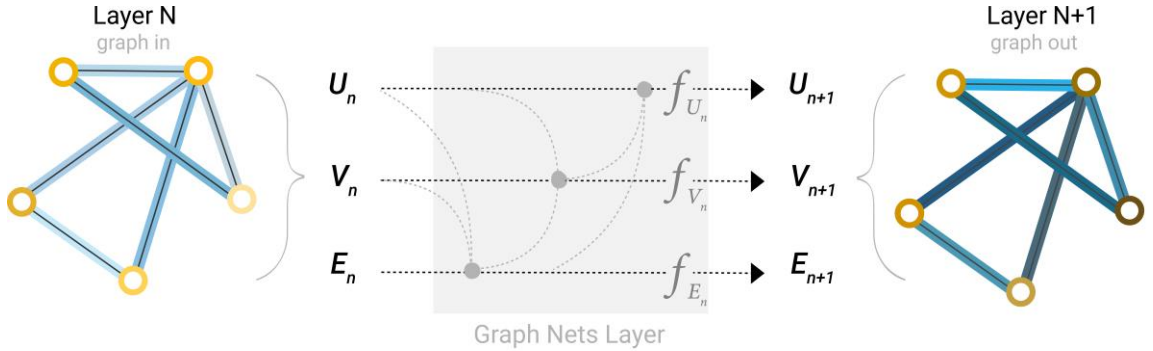


Image: Chem. Mater. 2019, 31, 9, 3564–3572

Machine learning algorithms evaluation

Sparse representation

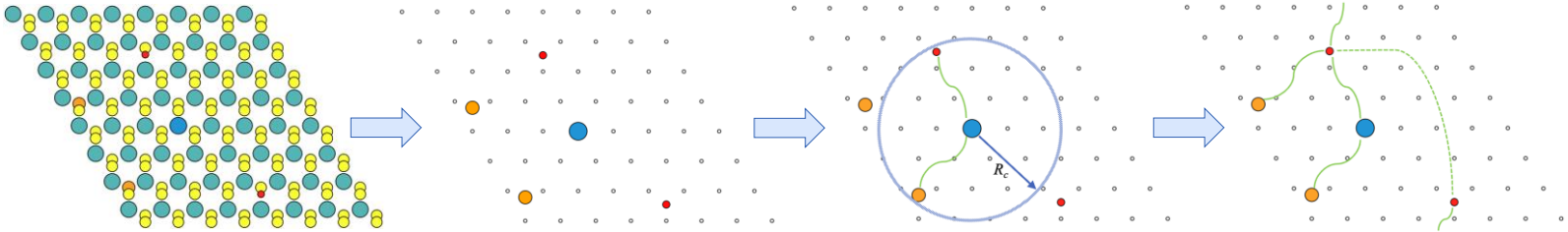
Input/output: graph with
- V Vertex (or node) attributes e.g., atom species
- E Edge attributes e.g., distances
- Global (or master node) attributes e.g., energy, band gap



Images source: https://distill.pub/2021/gnn-intro/

update function f = , ...

pooling function ρ



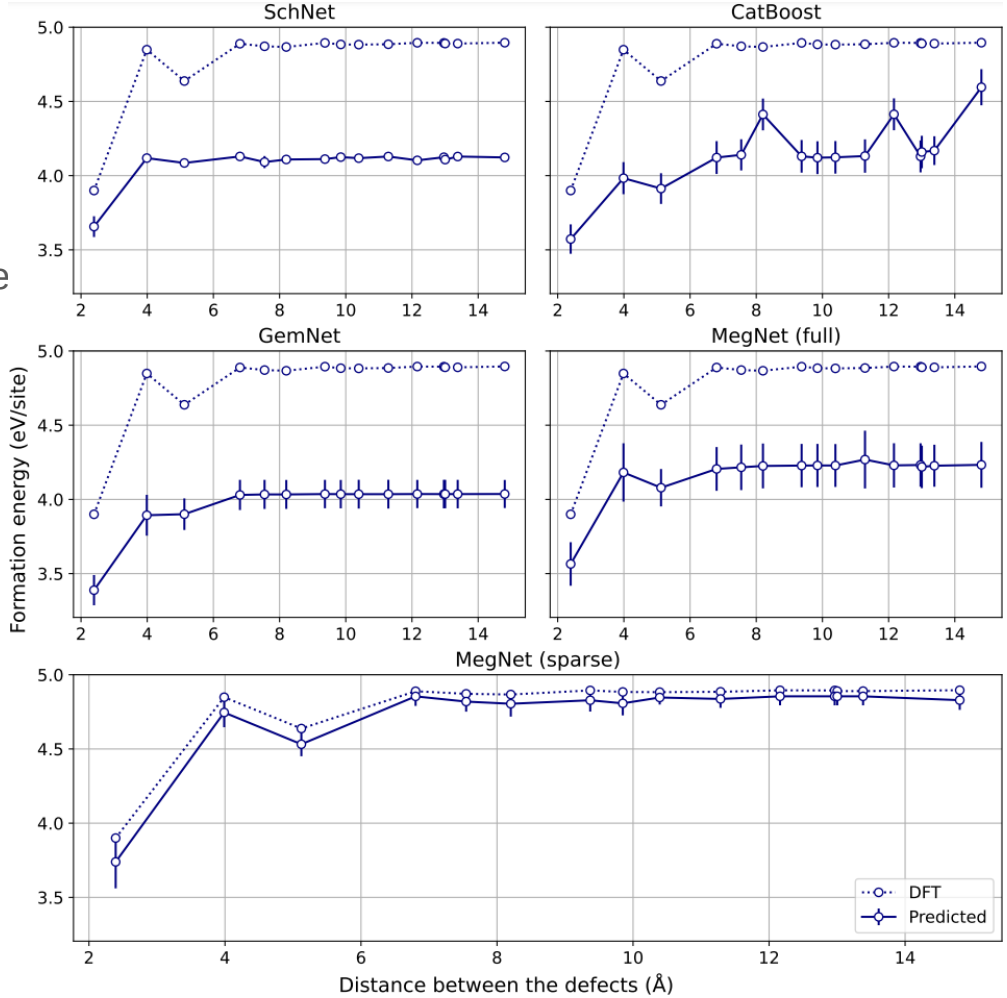
Build the graph from defects, not atoms

1. Reduces dimensionality from O(500) to O(9)
2. Preserves information

Node [defect atom, pristine atom]
Edge [distance, Δz]
Global state: pristine formula [42, 16]

Two vacancies in MoS_2

- Energy is non-monotonic with distance
- Baselines fail to learn it, while
- Our sparse representation does



https://www.nature.com/articles/s41524-023-01062-z

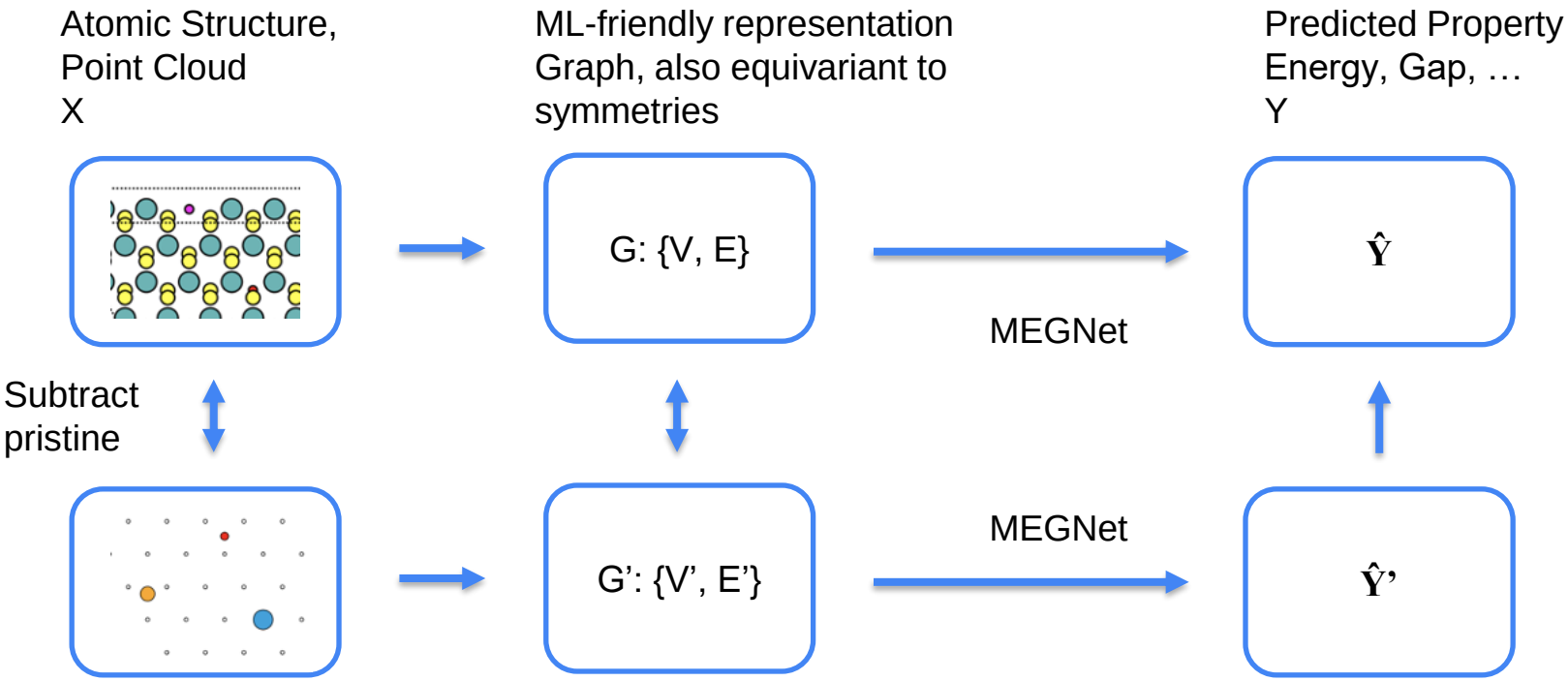
Formation energy per site MAE, meV; lower is better						
Material	Density	SchNet	GemNet	MEGNet	CatBoost	**Sparse (MEGNet)**
combined	both	631 ± 31	483 ± 91	158 ± 47	164 ± 5	$\mathbf{43 \pm 5}$
BP	high	2088 ± 72	1490 ± 429	198 ± 211	382 ± 30	$\mathbf{80 \pm 10}$
GaSe	high	245 ± 12	230 ± 41	107 ± 25	103 ± 4	$\mathbf{48 \pm 7}$
InSe	high	268 ± 19	247 ± 26	95 ± 27	137 ± 5	$\mathbf{35 \pm 2}$
MoS_2	high	321 ± 100	535 ± 206	136 ± 22	136 ± 5	$\mathbf{23 \pm 5}$
WSe_2	high	536 ± 123	575 ± 181	112 ± 33	162 ± 6	$\mathbf{23 \pm 4}$
h-BN	high	1442 ± 68	697 ± 315	496 ± 229	363 ± 17	$\mathbf{127 \pm 16}$
MoS_2	low	65 ± 5	44 ± 14	58 ± 11	12.6 ± 0.4	$\mathbf{4 \pm 1}$
WSe_2	low	85 ± 22	42 ± 9	65 ± 16	16.3 ± 0.8	$\mathbf{6 \pm 1}$

HOMO – LUMO gap MAE, meV; lower is better						
Material	Density	SchNet	GemNet	MEGNet	CatBoost	Sparse (MEGNet)
combined	both	224 ± 111	166 ± 42	$\mathbf{112 \pm 3}$	117 ± 1	$\mathbf{112 \pm 3}$
BP	high	208 ± 20	$\mathbf{176 \pm 10}$	$\mathbf{170 \pm 4}$	$\mathbf{174 \pm 2}$	187 ± 9
GaSe	high	309 ± 83	196 ± 11	$\mathbf{178 \pm 8}$	$\mathbf{173 \pm 4}$	194 ± 11
InSe	high	214 ± 69	178 ± 22	$\mathbf{156 \pm 7}$	$\mathbf{155 \pm 1}$	167 ± 15
MoS_2	high	204 ± 121	174 ± 111	54 ± 4	71 ± 4	$\mathbf{39 \pm 4}$
WSe_2	high	186 ± 177	268 ± 182	47 ± 3	106 ± 6	$\mathbf{38 \pm 4}$
h-BN	high	244 ± 24	227 ± 6	233 ± 4	$\mathbf{208 \pm 3}$	260 ± 14
MoS_2	low	187 ± 180	46 ± 42	30 ± 2	26.7 ± 0.8	$\mathbf{5.7 \pm 0.2}$
WSe_2	low	236 ± 224	64 ± 46	32 ± 5	18.3 ± 0.6	$\mathbf{8.1 \pm 0.6}$

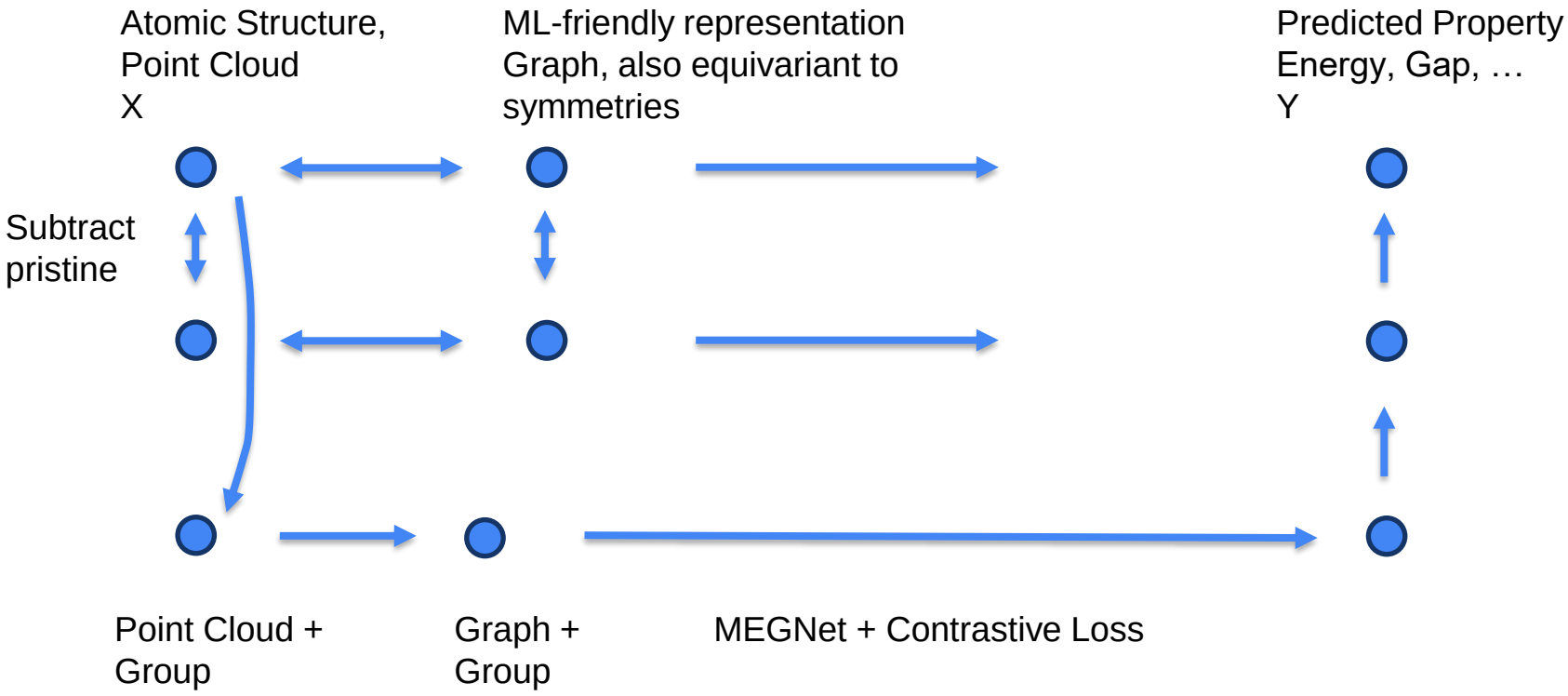
https://bit.ly/40kVW1h

- “Sparse (MEGNet)” is our representation
- Rest are state-of-the-art baselines
- by far is the best for energy and HOMO-LUMO gap prediction

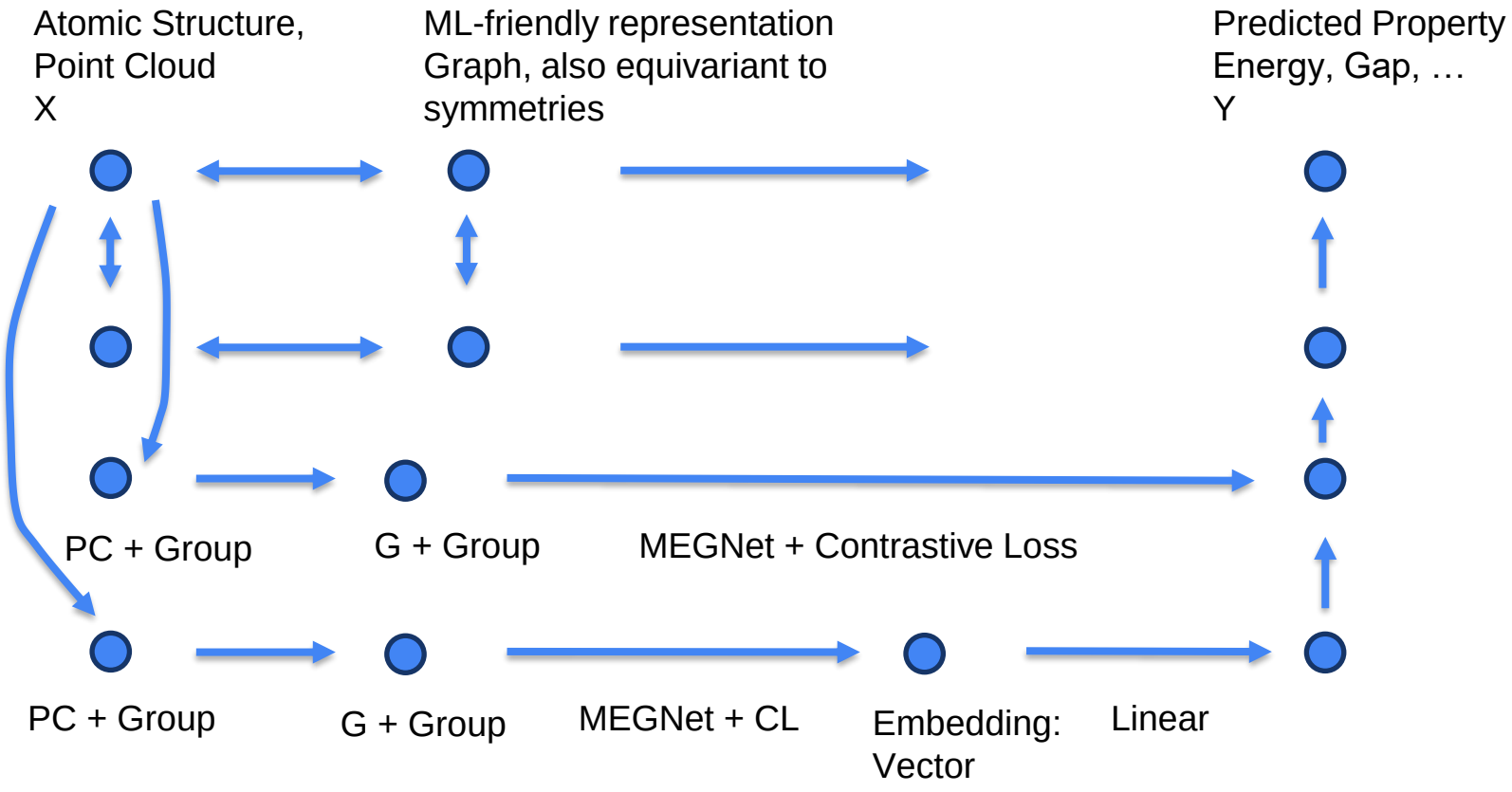
In more schematic terms



In more schematic terms



In more schematic terms (2)



Web-based collaborative platform

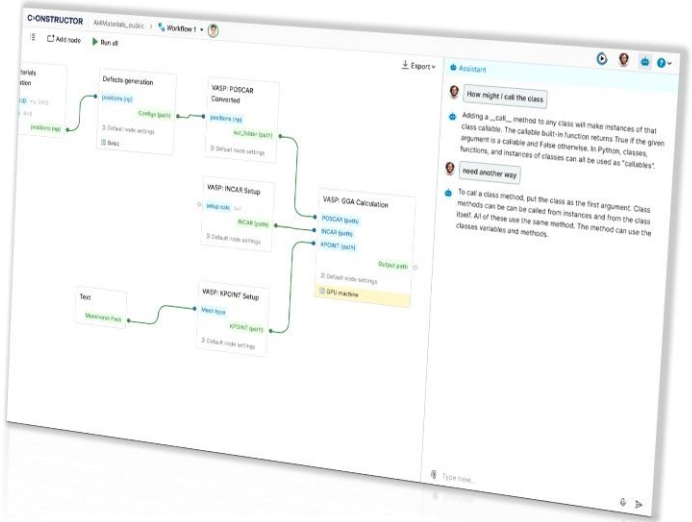
Cloud **compute and storage resources**

Research pipeline **visual representation**

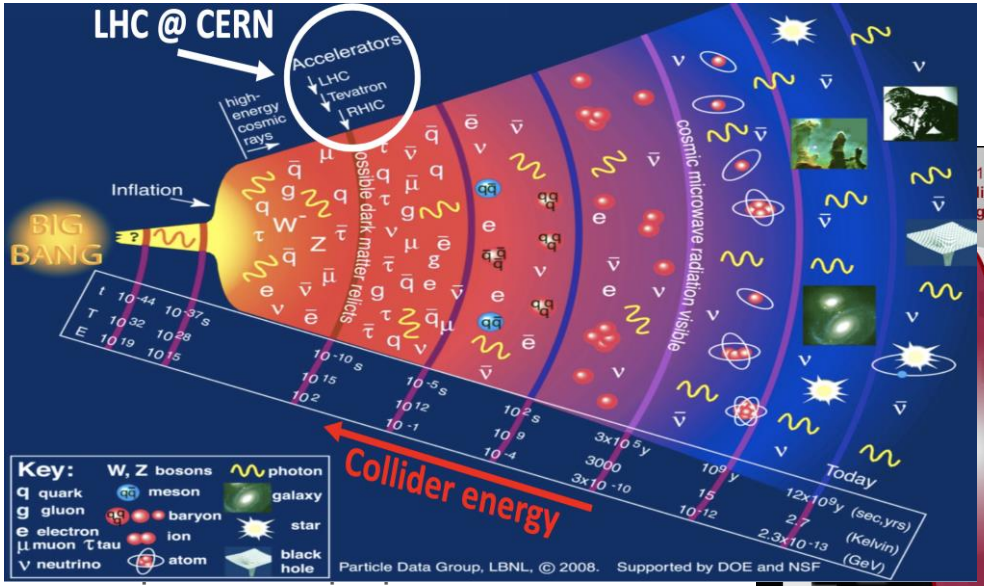
Collaborative work with versioning control

Publication of **reproducible experiments**

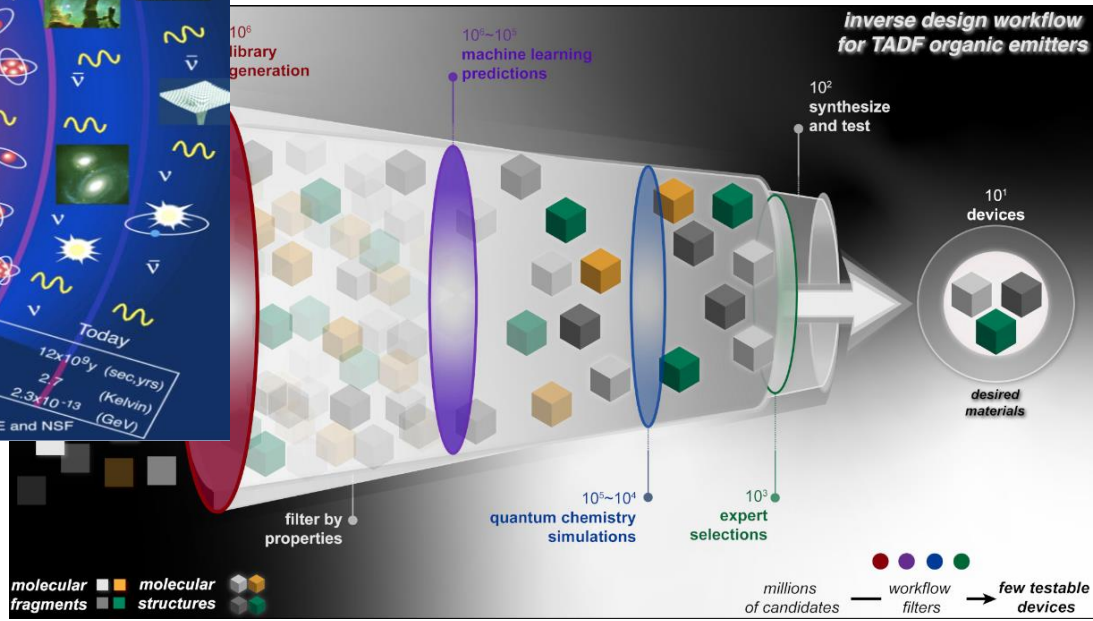
Research **AI assistant** built on a **scientific based Large Language Model** is coming



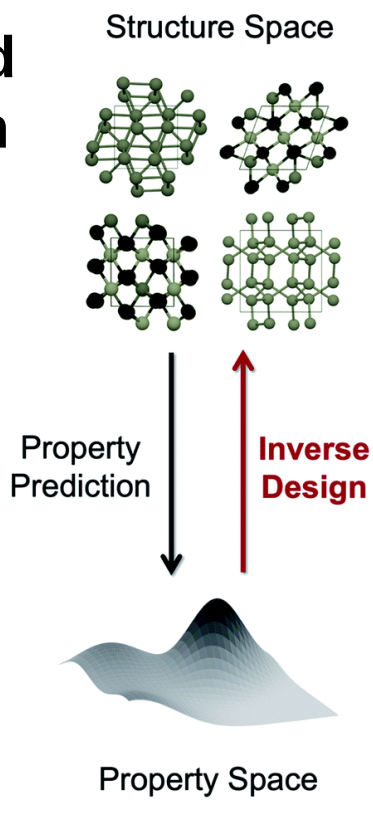
Inverse problem, high-throughput screening (HTS)



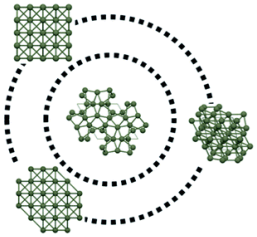
stages.



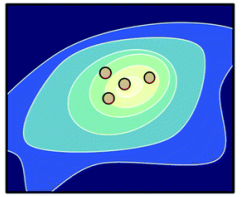
Generative and Inverse Design



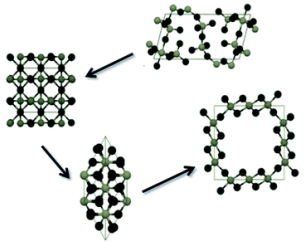
(a) High-Throughput Virtual Screening (HTVS)



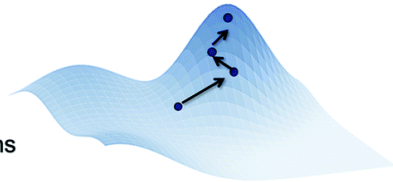
Database selection
Grid search
Hierarchical pipeline



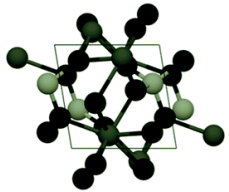
(b) Global Optimization (GO)



Learn prior explorations
Efficient search
Predict new structure

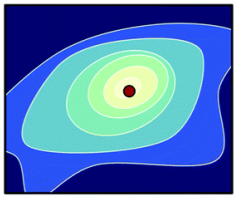


(c) Generative Model (GM)



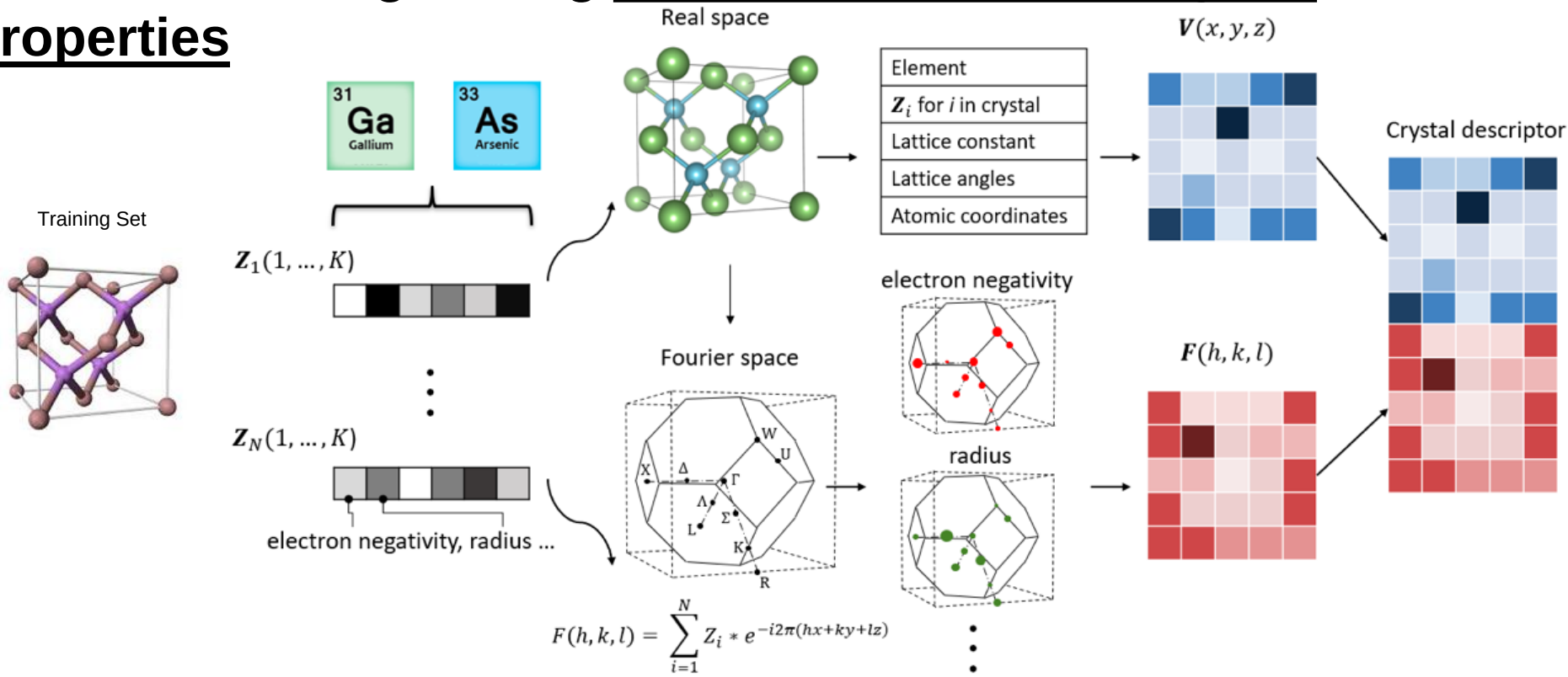
Decoding

Learn data distribution
Efficient search
Generate new structure



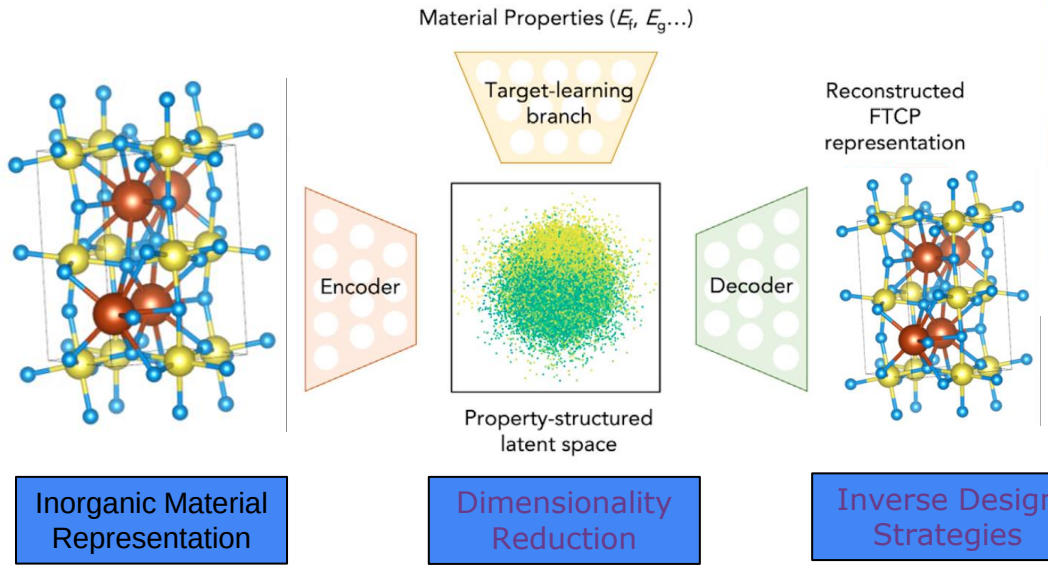
Chem. Sci., 2020,**11**, 4871-4881

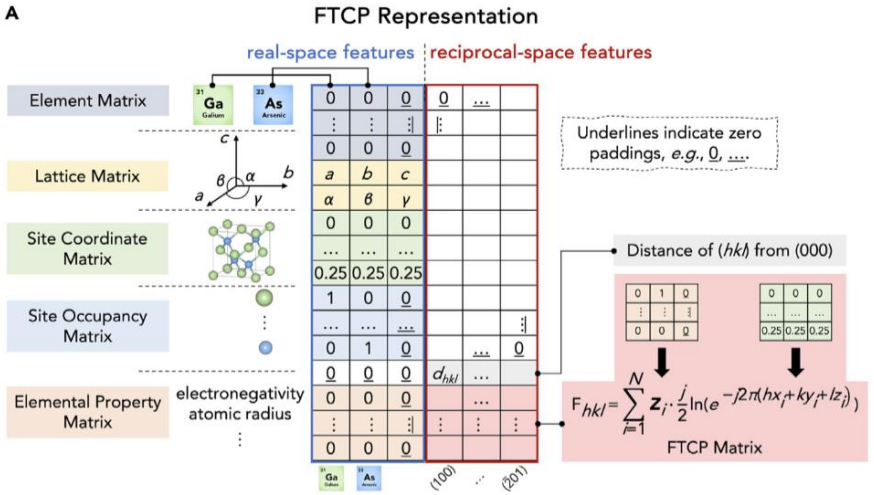
Generative Design using Fourier transformed crystal properties



Ren et. al. ***Matter*** *Volume 5, Issue 1, 5 January 2022, Pages 314-335*

Autoencoders for inverse design of Inorganic Materials





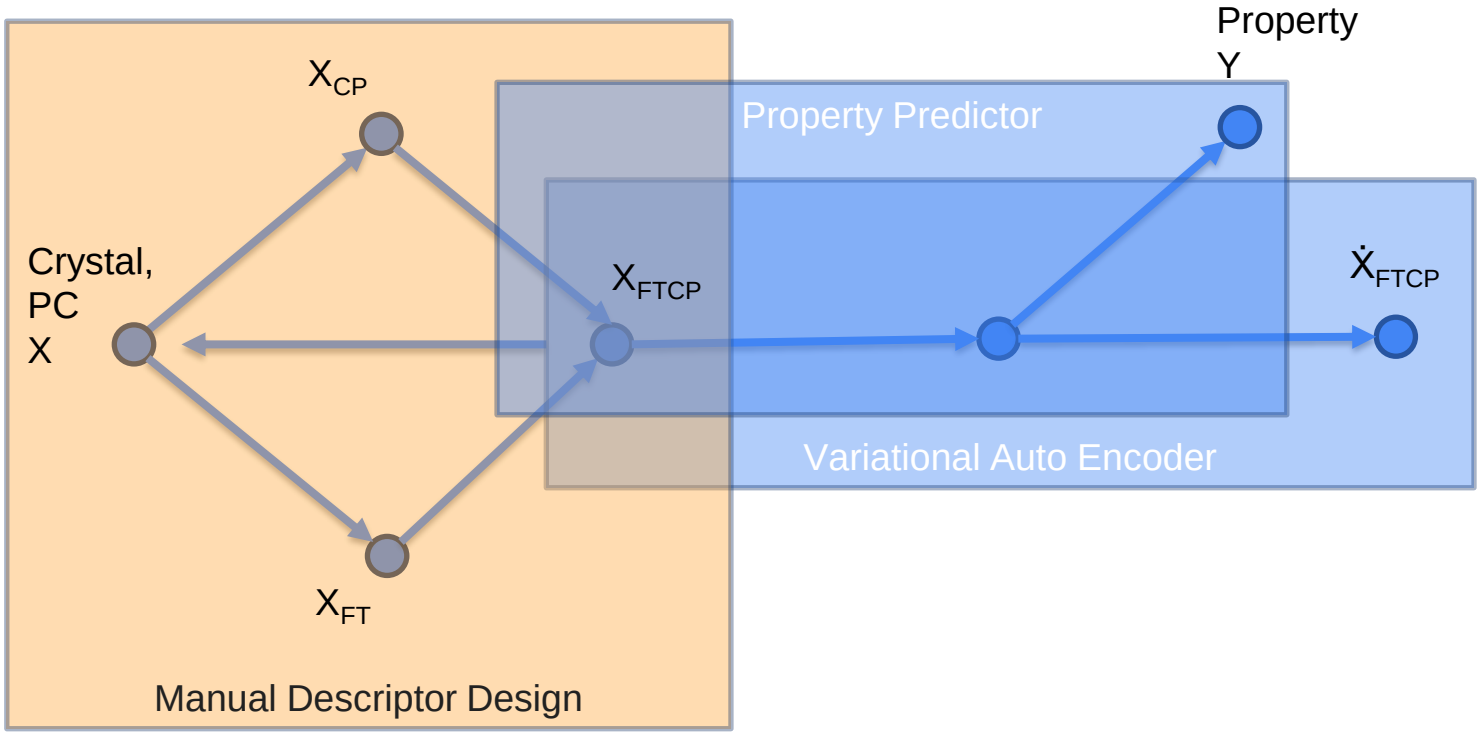
Inorganic Material Representation

Dimensionality Reduction

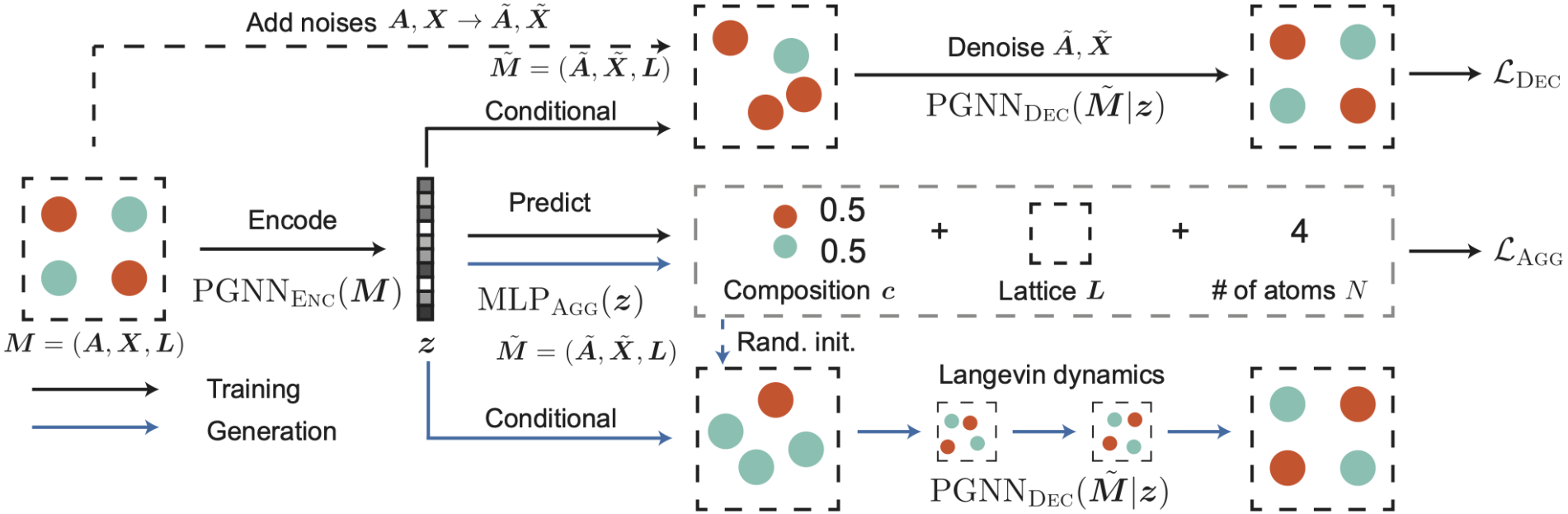
Inverse Design Strategies

Ren et. al. ***Matter*** *Volume 5, Issue 1, 5 January 2022, Pages 314-335*

Schematic representation

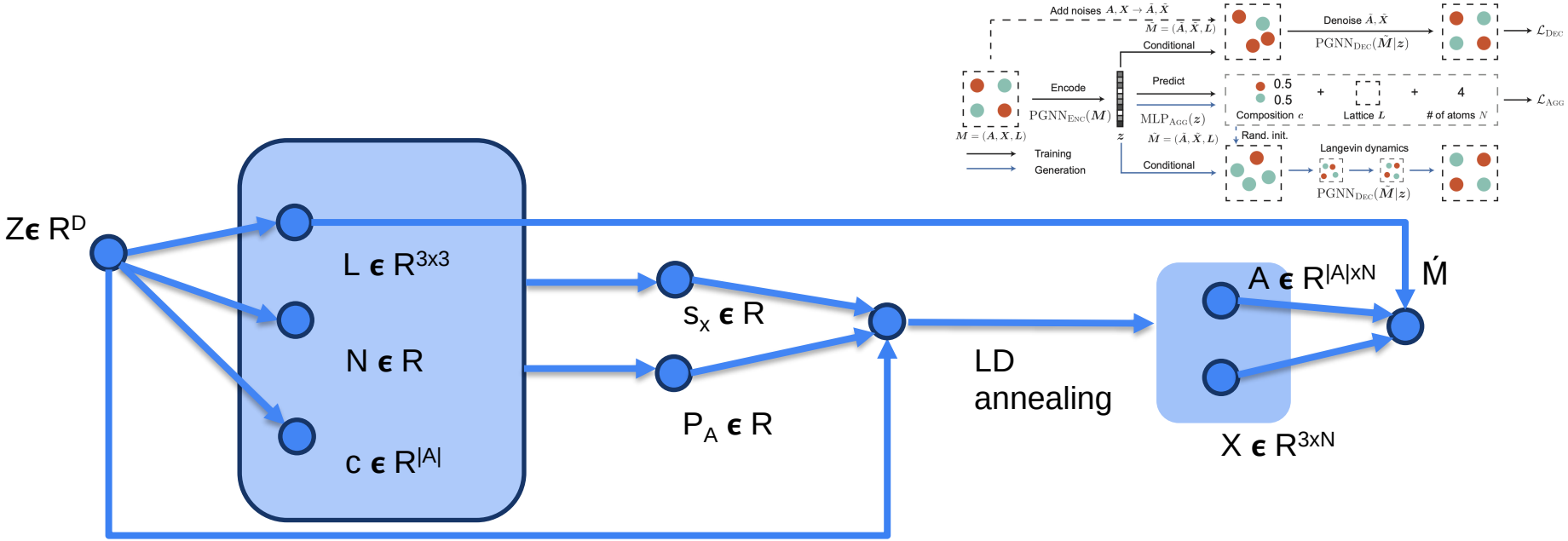


Crystal Diffusion Variational Auto-Encoder (CDVAE)



https://github.com/txie-93/cdvae

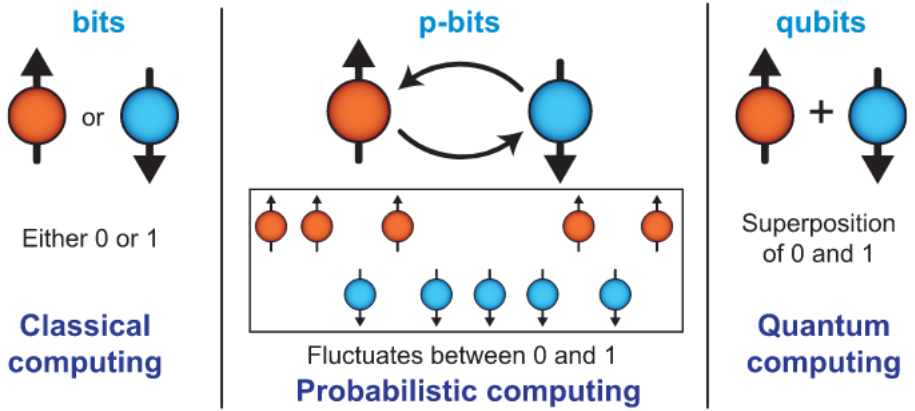
Crystal Diffusion Variational Auto-Encoder (CDVAE)

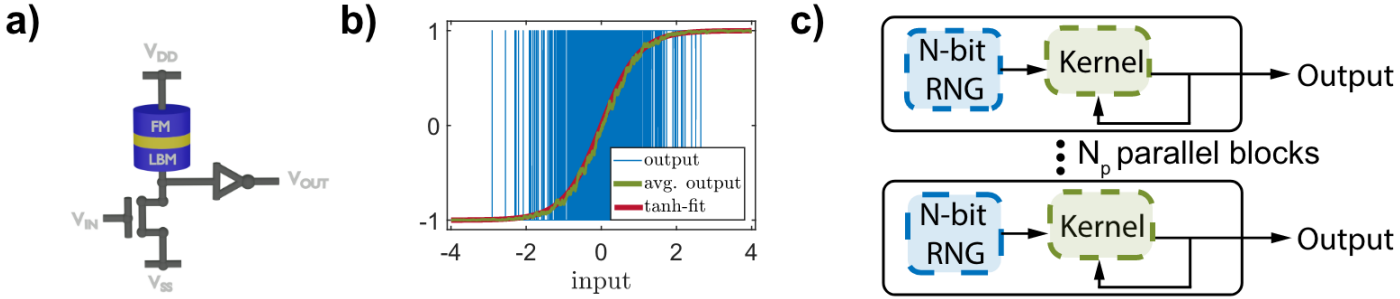


CDVAE prediction pipeline

Probabilistic computing

a) s-MTJ based p-bit (T: transistor, s-MTJ: stochastic magnetic tunnel junction). The MTJ has a fixed ferromagnet (FM) layer and a low-barrier magnet (LBM) as a free layer. Input: analog voltage , output: digital voltage.





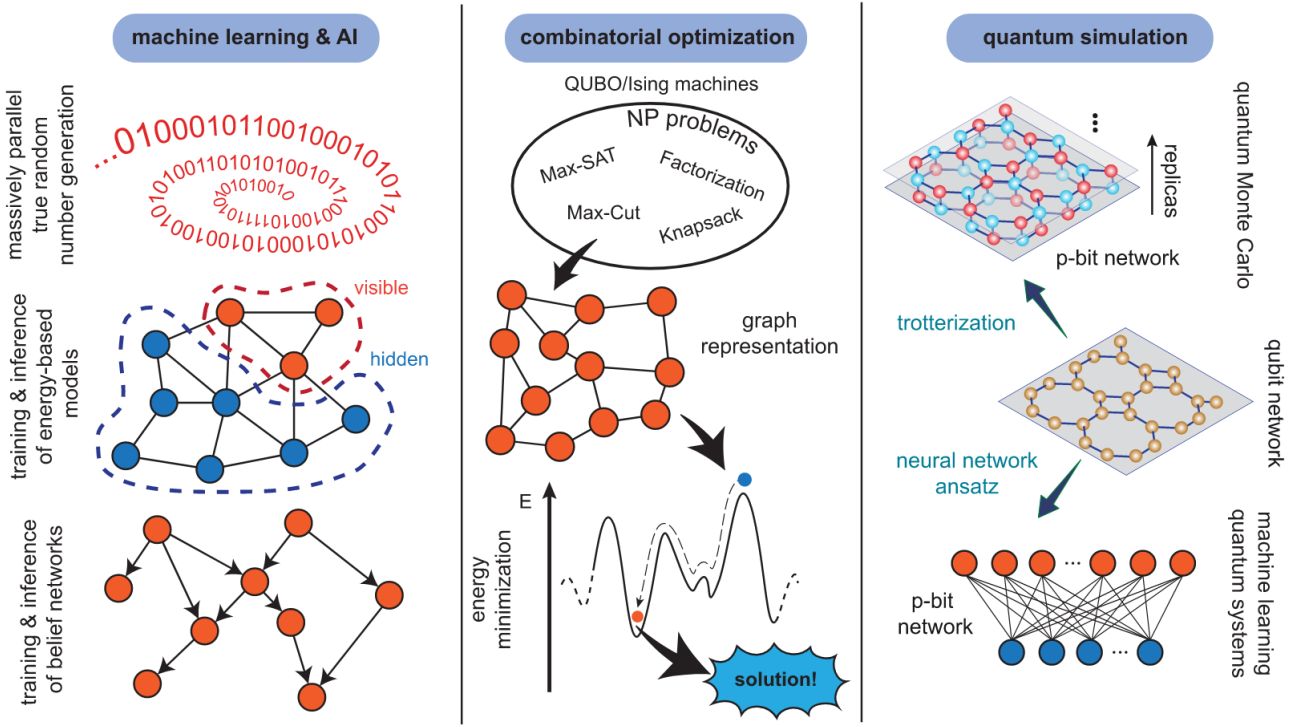
b) Input-Output characteristic of the p-bit. On average the p-bit output can be described by a sigmoid.

c) framework of a computer based on p-bits. The p-bits feed random numbers into a Kernel that generates the output

10.1109/IEDM45625.2022.10019548

Application domains

- combinatorial optimization, probabilistic ML, and quantum simulation
- fits nicely inverse design problems
- aided with ML one can optimize algorithms on really low - atomic level



10.1109/JXCDC.2023.3256981

https://bit.ly/46982ff

HEP vs Material Science (MS)

Similarities

- fast simulation / generative models
- need for foundation models
- representation learning
- optimal transport methods
- inverse design / design optimization
- ML model uncertainty estimation
- spatial structures representations
- need for differential simulations / simulation-based inference
- denoising / stability estimation methods
- anomaly detection methods

HEP distinct features

- Centralized data collection
- Bump hunting
- Science of confidence intervals
- Plenty of theoretical models for unknown
- Search for unknown

MS distinct features

- Multiscale effects / modelling
- Time-dependent modelling
- Data is heavily fragmented

ML challenges summary

- Data efficiency for all challenges below, the search space is vast
- Multimodal data analysis
- Multiscale modelling, both for space and time, static and dynamic
- Language for representing data transformations respecting equivariance to symmetries
- Forward problem for various classes of materials: e.g., Structure => Energy
  - Structure robustness estimation
- Inverse problem for atomic structure: Energy => Structure
- Inverse problem for material synthesis process/design
- Theory synthesis (see talk by Li Qianxiao)

Every interdisciplinary project is by far - translation problem

Material Science and AI essentially speak different languages
e.g.: **fidelity** in material science and **accuracy** in ML are very similar

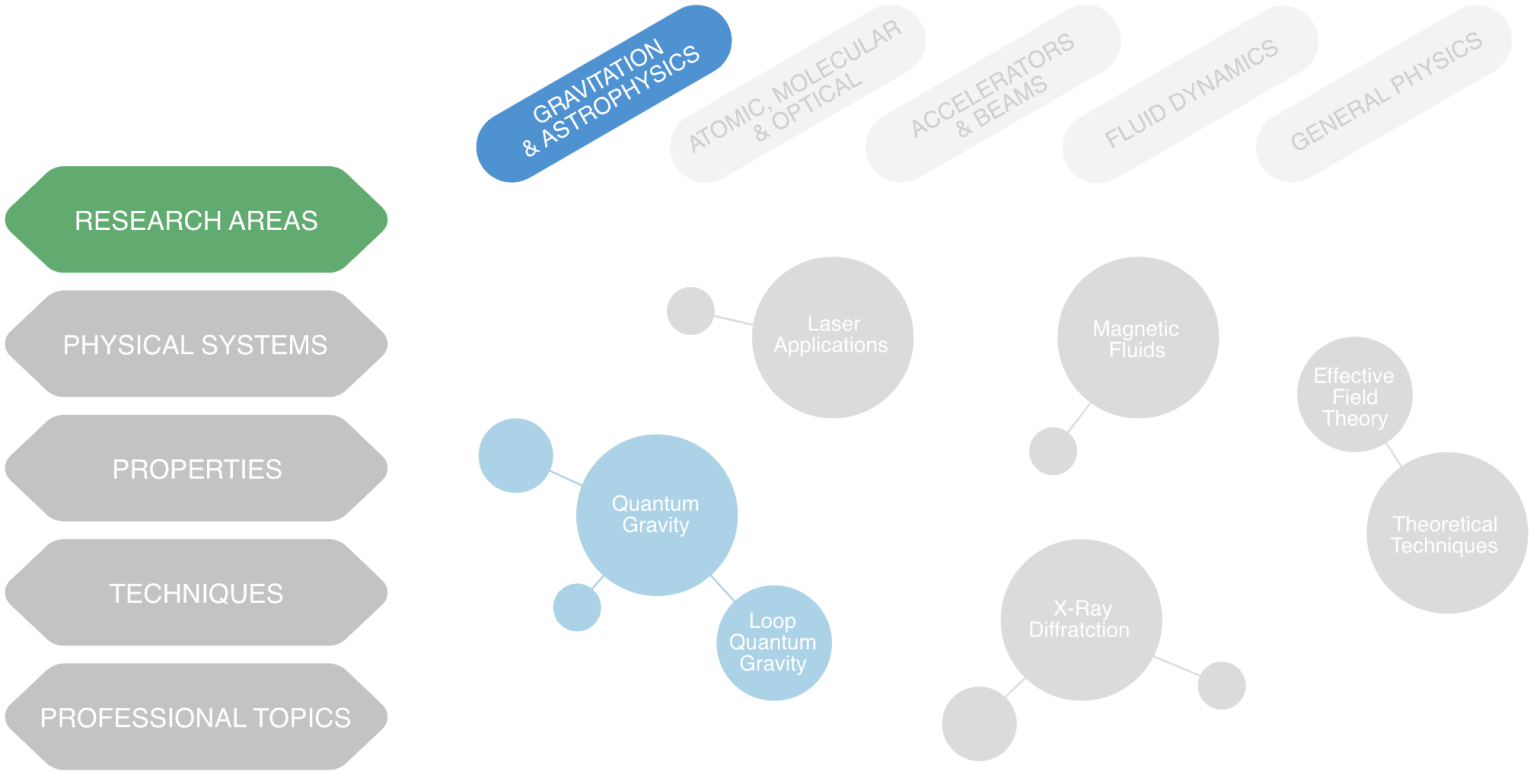
E.g., for AI + X one usually needs:

- What kind of model you are interested? (forward, inverse, how explanatory should it be? how generic should it be?)
- Find and specify good dataset
- Specify computable figure(s) of merit
- Describe some physics intuition behind data
- Define expected *data transformation*

Curios to discuss (brainstorming or elsehow) if *data transformation* can be represented in a declarative language that takes into account both data formats, information preserved and respected symmetries that would be easy to translate into loss function / computer code?

PhySH concepts are organized by facets and disciplines

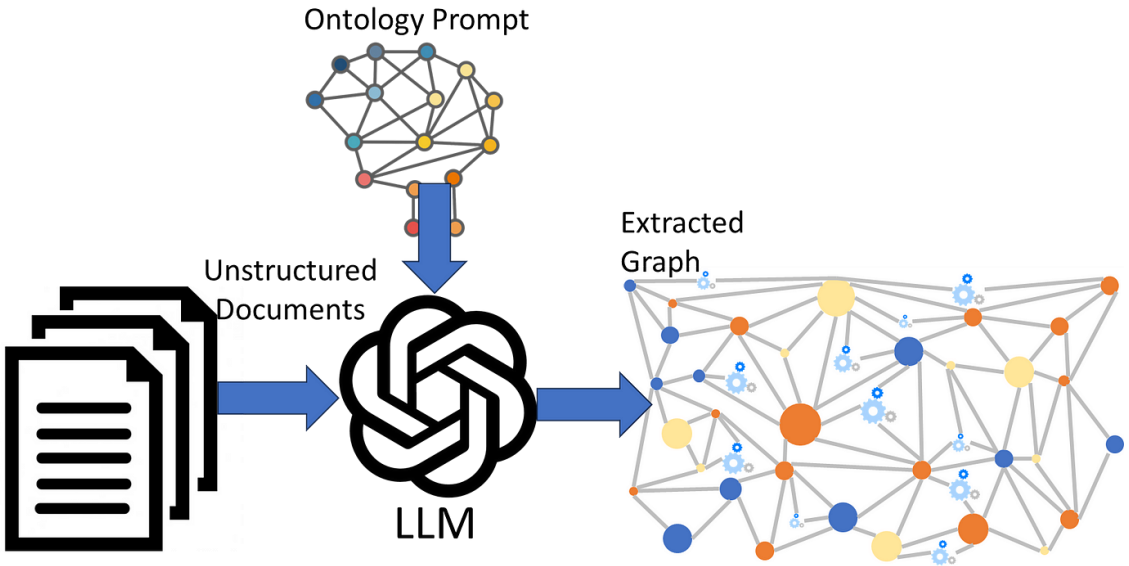
More Details



Andrey Ustyuzhanin, Artem Maevskiy

https://physh.org/

Filling in interdisciplinary gaps with LLMs



Conjecture:
a) LLMs contains information about every scientific domain structure
b) challenges of similar structure can be addressed by similar methods

Question: can we use LLMs as a source for new method mining?

Hypothesis finder

Query

Enter your question

How can I model a material properties like magnetic resistivity and electronic bands that depend on material atomic structure as well as higher-level mesoscopic configuration like phase domains and its walls in a unified holistic way. Such computational model should be able to generalize descriptions between different scales and help to reason about exhibited properties.

Num. contexts

3

1

Generator T

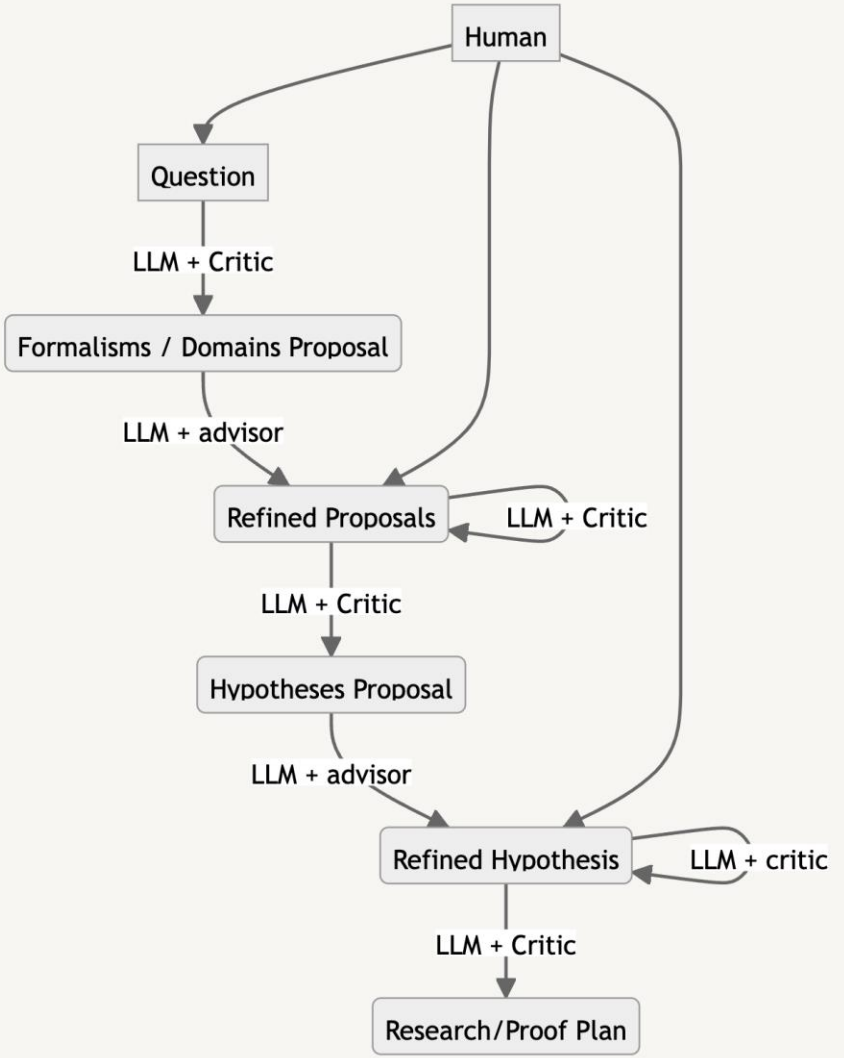
1.00

0.00

GPT model

gpt-4

Fake llm (debug



Andrey Ustyuzhanin, Artem Maevskiy

Question

How can I model a material properties like magnetic resistivity and electronic bands that depend on material atomic structure as well as higher-level mesoscopic configuration like phase domains in a holistic way.
Such computational model should be able to generalize descriptions between different scales and help to reason about exhibited properties.

Context 1. Computational Physicist / Computational Physics::Multiscale Modeling Theory / 70.0

Summary

Ultimately, as a computational physicist, the largest tool at our disposal would be the field of computational physics with its powerful numerical techniques. Specifically, using Multiscale Modeling principles we may combine different models for different scales into a unified model. This will let us move from understanding atomic structures to mesoscopic configurations simultaneously. Theoretical representation based on Quantum Field Theory would serve as a common language to describe the phenomena at different scales, from quantum mechanical interactions to phase domain formations and properties like magnetic resistivity and electronic band structures.

Role: Computational Physicist

Framework (broad): Computational Physics

Review

The proposed approach is comprehensive and w
thought out. The idea of using a multidisciplinar
approach via computational physics and multisc
modeling theory seems promising. However, the
proposal lacks details on how the Quantum Fiel
Theory will be adapted to meet the unique need
this problem. The submission also does not prov
any information on the specific computational
tools, techniques, or algorithms that will be used
also misses out on a crucial aspect - validation o
the model. There is no mention of how the mode
predictions will be tested against real-world
observations or experiments.

Score

70.0

Full result (JSON):

Andrey Ustyuzhanin, Artem Maevskiy

Representation and Hypothesis

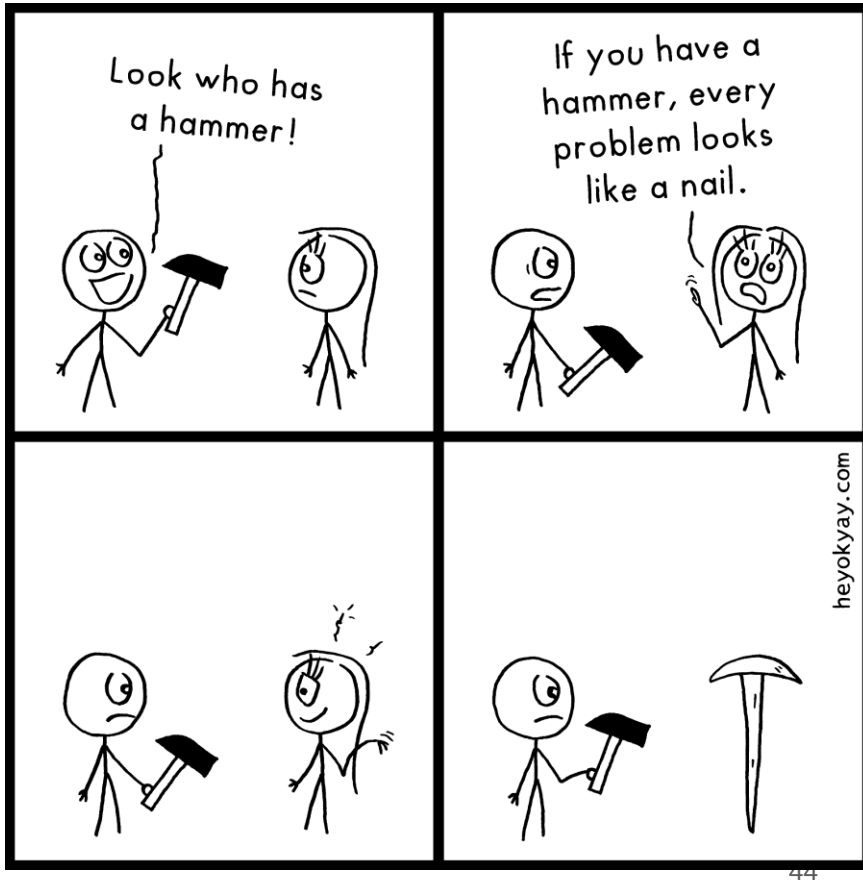
We hypothesize that by integrating the Path-Integral Hierarchical Equations of Motion (PI-HEOM) with a multi-scale formalism, it is possible to create a unified computational model. This model would permit a comprehensive understanding of material properties like magnetic resistivity and electronic bands from atomic to mesoscopic scales.

The integration of quantum dynamics provided by PI-HEOM with the differential scales of interaction given by multi-scale modelling can bridge the gap between discrete atomistic behavior and higher-level mesoscopic structure. Steps:

1. Develop or adapt existing PI-HEOM and multi-scale formalism frameworks.
2. Create a computational algorithm to integrate both, towards generating a unified model.
3. Simulate diverse material systems and study properties arising from atomic structure to mesoscopic configurations.
4. Compare the results with existing experimental data to validate the model.
5. Refine the algorithm based on findings and reiterate.

Conclusion

- Fields are completely different by tasks, scales, methods
- From ML perspective similarities prevail
- Rich field for cross-domain collaborations
- Two key issues: finding suitable representation and cross-domain communication
- LLMs provide opportunities to bridge sophisticated multidisciplinary gaps
  - By translating problems and approaches
  - By finding meaningful representations



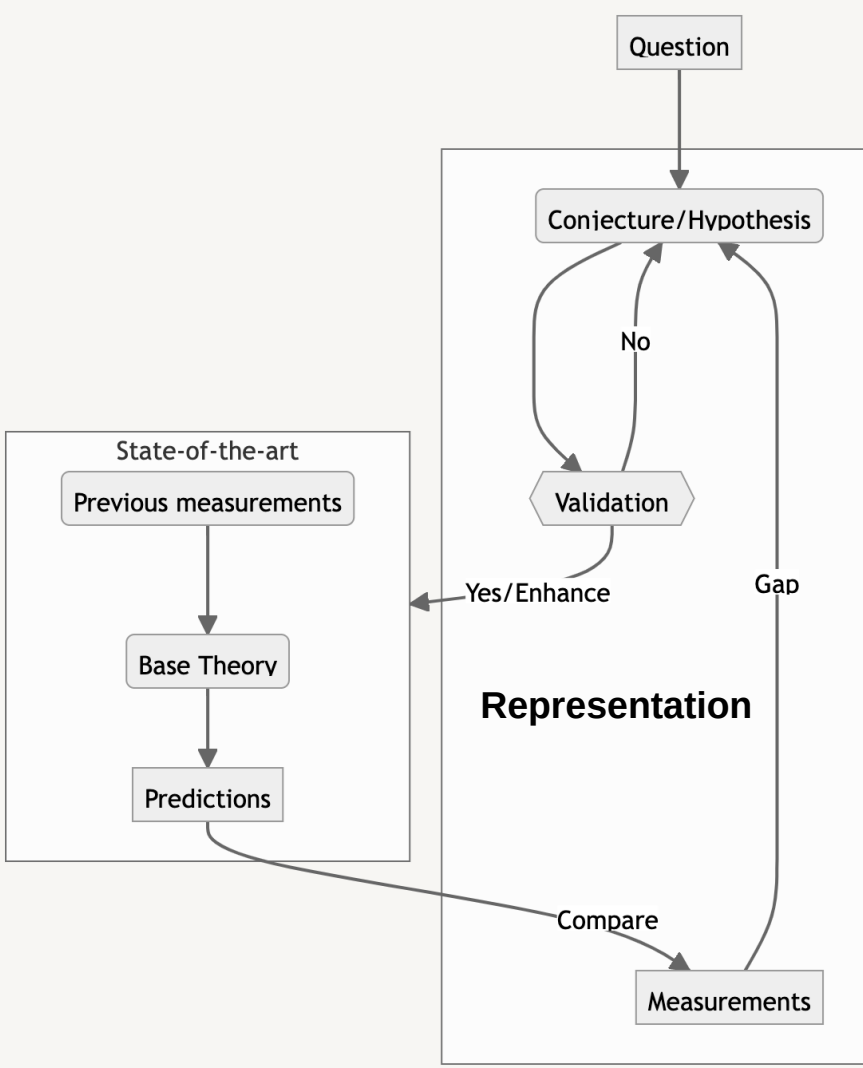
Thank you

Andrey andrey.u@nus.edu.sg

Backup

Finding Representations

- Scientific development cycle
- Representation of hypothesis is crucial for validation
- **Goal:** select suitable formalisms given the initial problem/question, data and metadata using LLMs



Dan Hendricks: ”Natural Selection Favors AIs over Humans” (competence without comprehension)

- The logic of competitive evolution will lead to the same outcome as with humans: increasingly intelligent AI agents will become more selfish and willing to use deception and force to achieve their goals, the main one being power.
- Natural selection of AI agents results in more selfish species usually having an advantage over more altruistic ones. AI agents will behave selfishly and pursue their own interests, with little concern for humans, which could lead to catastrophic risks for humanity.
- There is a considerable chance that this will happen not as a result of some specific evil intent by humans or machines, but solely as a result of applying evolutionary principles of development to AI according to Darwinian logic.
- To minimize this risk, it is necessary to carefully design the internal motivations of AI agents, introduce restrictions on their actions, and create institutions that encourage cooperation in AI.

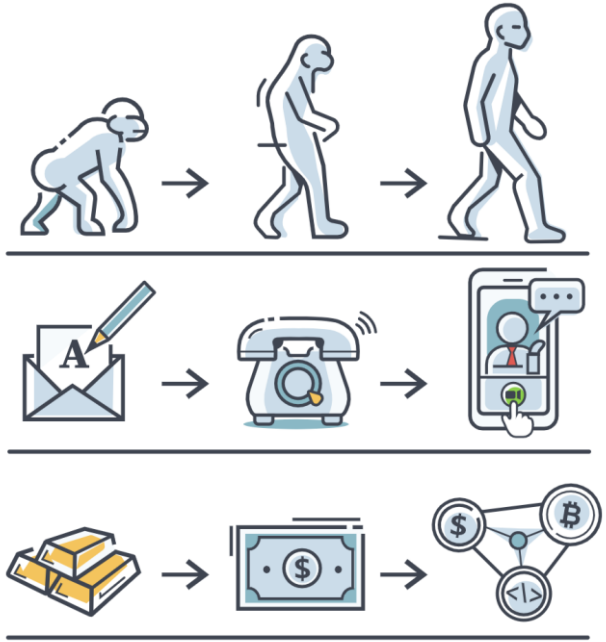
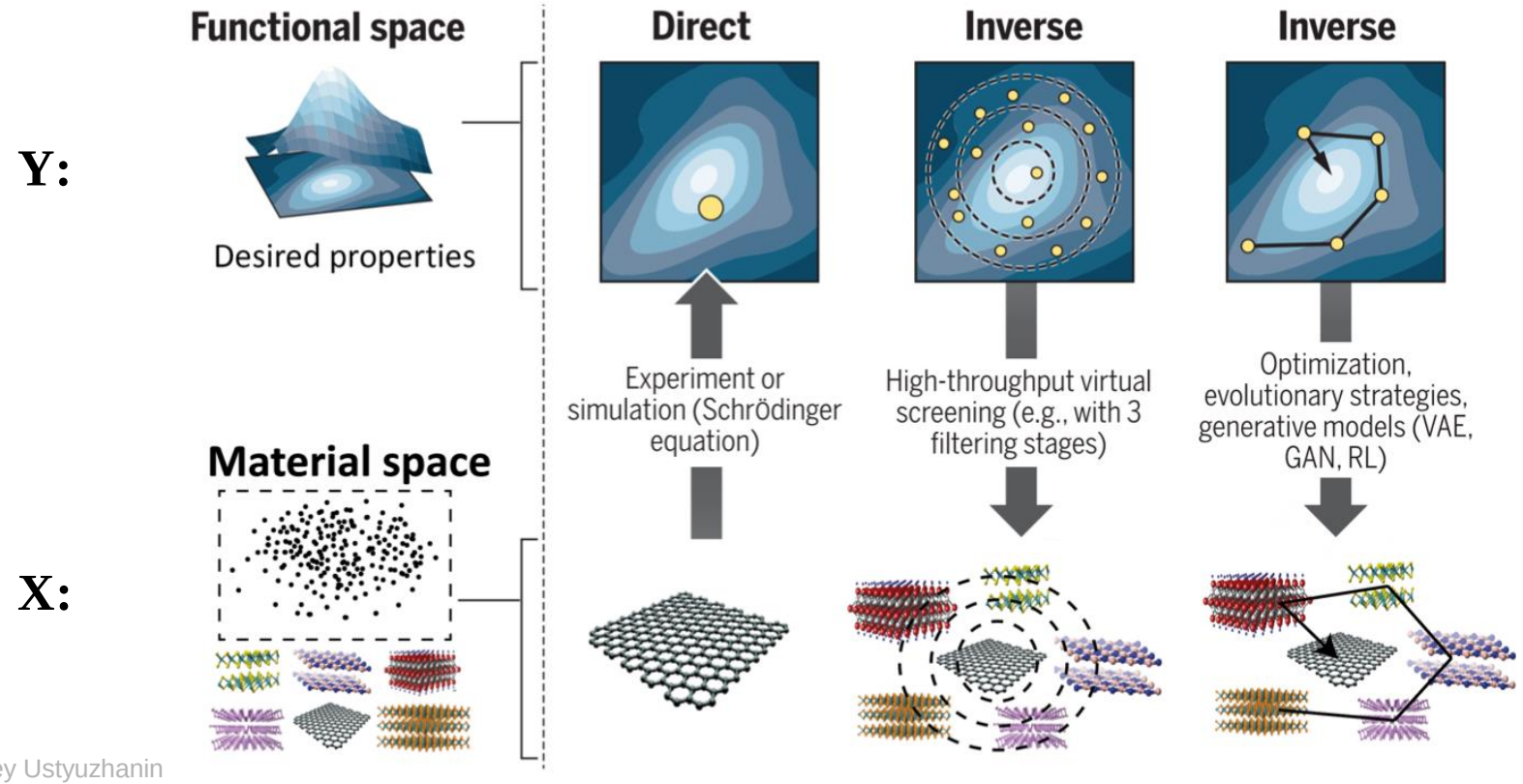


Figure 2: Darwinism generalized across different domains. The arrow does not necessarily indicate superiority but indicates time.

https://arxiv.org/abs/2303.16200

Connection between two challenges through optimizaiton



Machine Intelligence (or AI) in a nutshell

- Learning complex patterns by fitting weird function into data space under some prior constraints
- Critical steps: pick data representation, figure of merit and loss function
- All weird functions are wrong, but some are useful

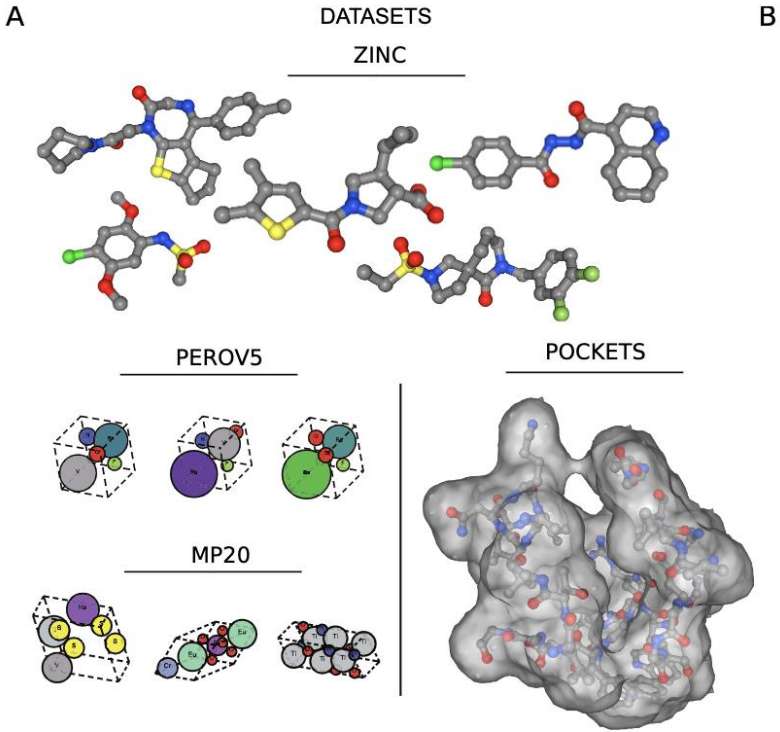
OK, predictions are accurate, but can we learn some physics from it?

- **Conjecture:** the better model predicts, the better physics it learned internally.
- Interpretation:
  - Hidden representation
  - Symbolic regression
  - Feature importance
  - Generalization (accuracy on a hidden dataset)
- Architecture regularization by inductive bias: chose architecture in a way that will help model to find right internal representation with **less data**

LLM for structure generation

Can LLMs trained using next-token prediction generate novel and valid structures in 3D for substantially different chemical structures?

- Molecules,
- Crystals
- Protein binding sites



https://arxiv.org/pdf/2305.05708.pdf

Schematic representation and Results

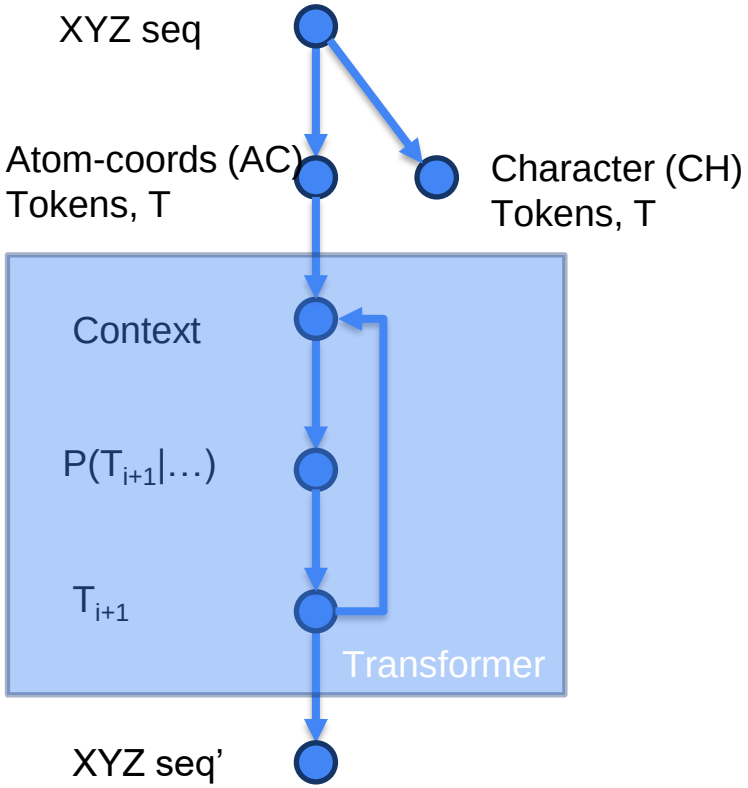
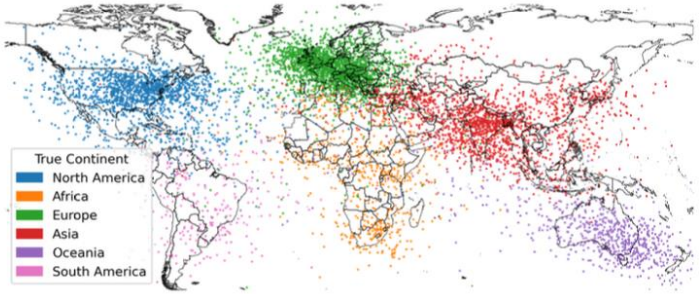


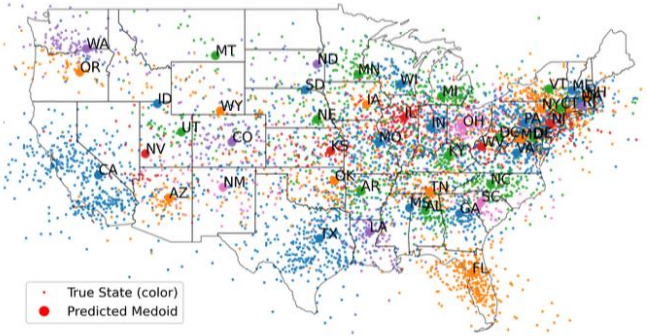
TABLE I. Generation performance for ZINC.

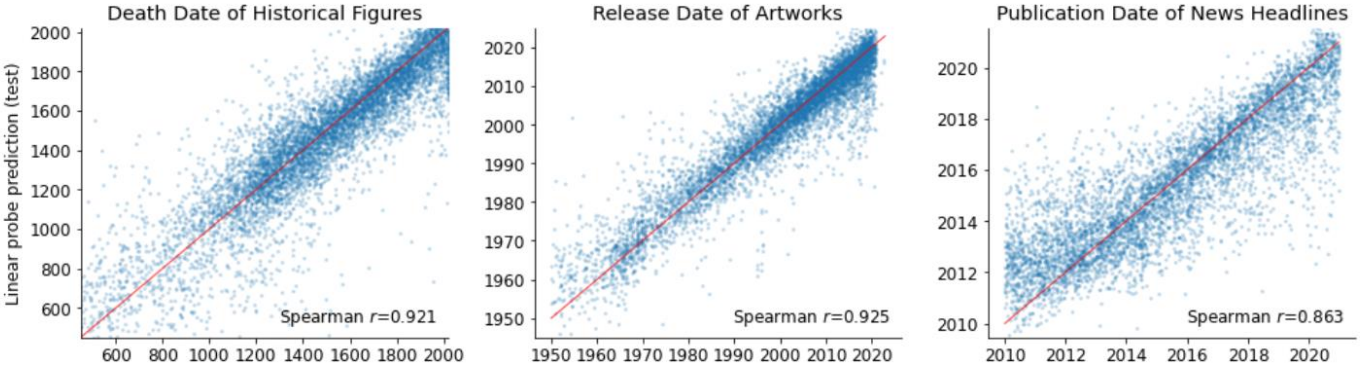
3D	Model	Basic Metrics (%) ↑ Valid	Unique	Novel	WA Metrics ↓ MW	SA	QED
Not 3D	Train	100.0	100.0	100.0	0.816	0.013	0.002
	SMLM	98.35	100.0	100.0	3.640	0.049	0.005
	SFLM	100.0	100.0	100.0	3.772	0.085	0.006
	DGMG	79.63	100.0	99.38	88.94	3.163	0.095
	JTVAE	100.0	98.56	100.0	22.63	0.126	0.023
	CGVAE	100.0	100.0	100.0	45.61	0.426	0.038
3D	ENF	1.05	96.37	99.72	168.5	1.886	0.160
	GSchNet	1.20	55.96	98.33	152.7	1.126	0.185
	EDM	77.51	96.40	95.30	101.2	0.939	0.093
	LM-CH	90.13	100.0	100.0	3.912	2.608	0.077
	LM-AC	**98.51**	**100.0**	**100.0**	**1.811**	**0.026**	**0.004**

MW – molecular weight
SA - synthetic accessibility
QED - quantitative estimate of drug-likeness

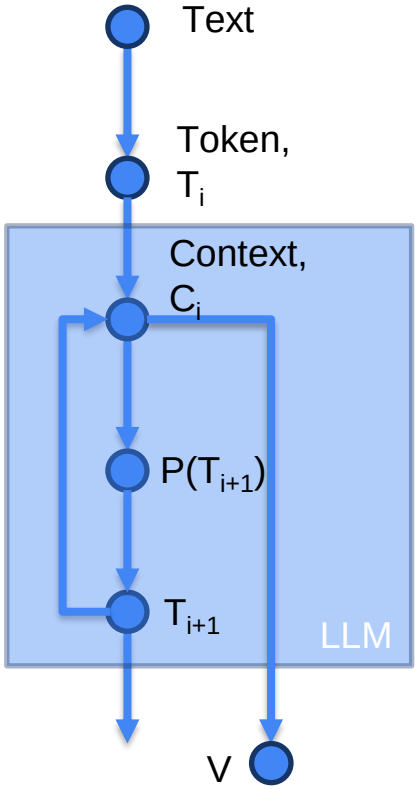
On internal space & time by LLM



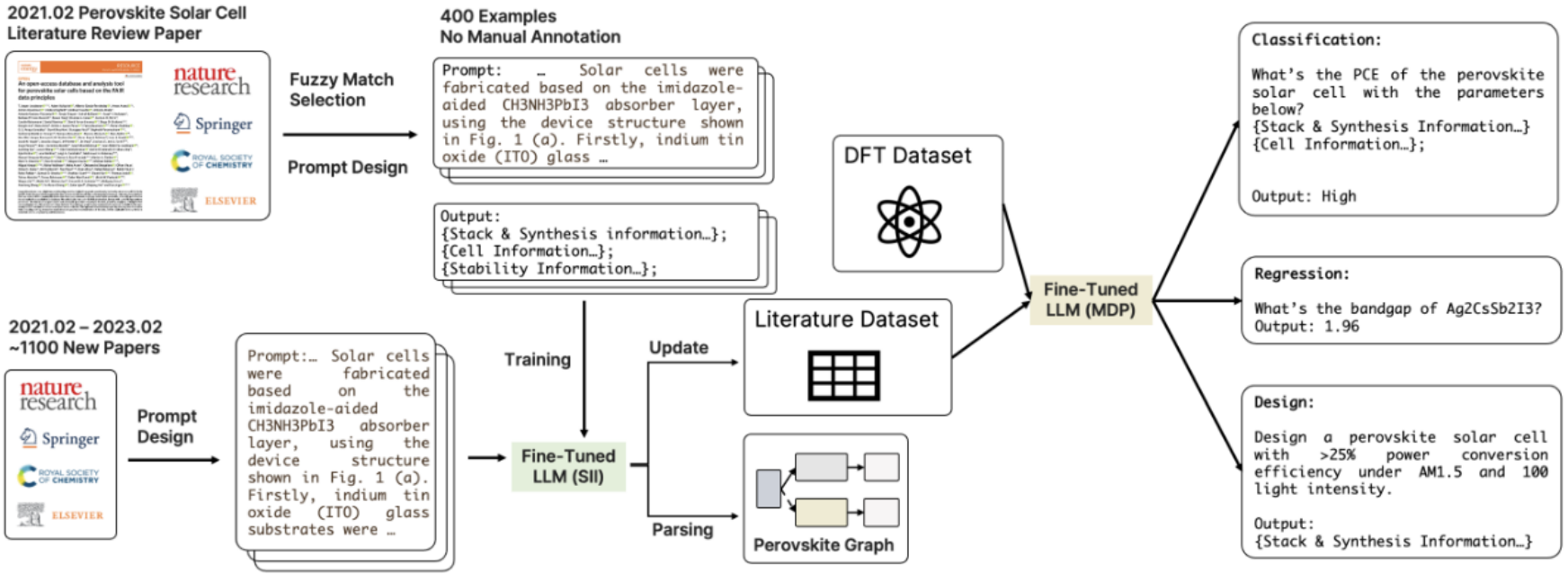




ArXiv:2310.02207



Unlocking the Secrets of Materials Science with GPT

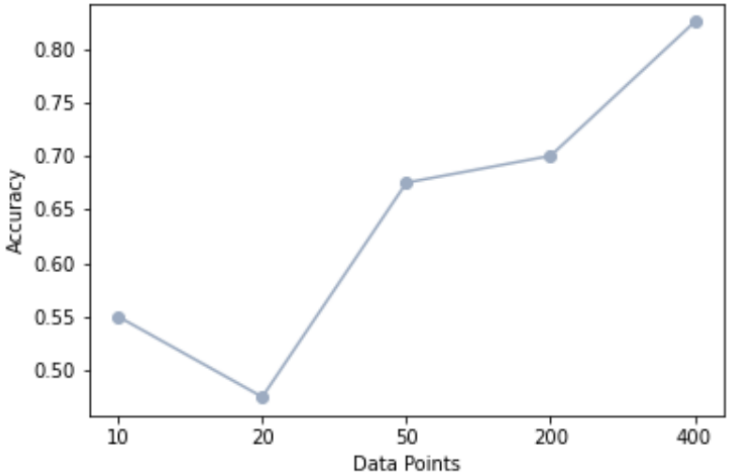


- LLM for structured information inference(SII) tasks and
- material & device prediction(MDP) tasks

https://arxiv.org/pdf/2304.02213.pdf

Performance evaluation

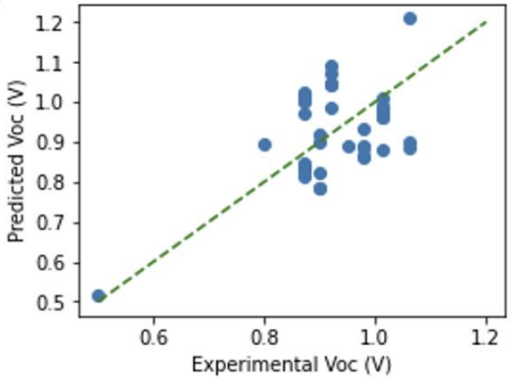
- Classification
- Power conversion efficiency, PCE level of perovskite solar cells under the AM1.5 spectrum and 1000 W/m2 light intensity



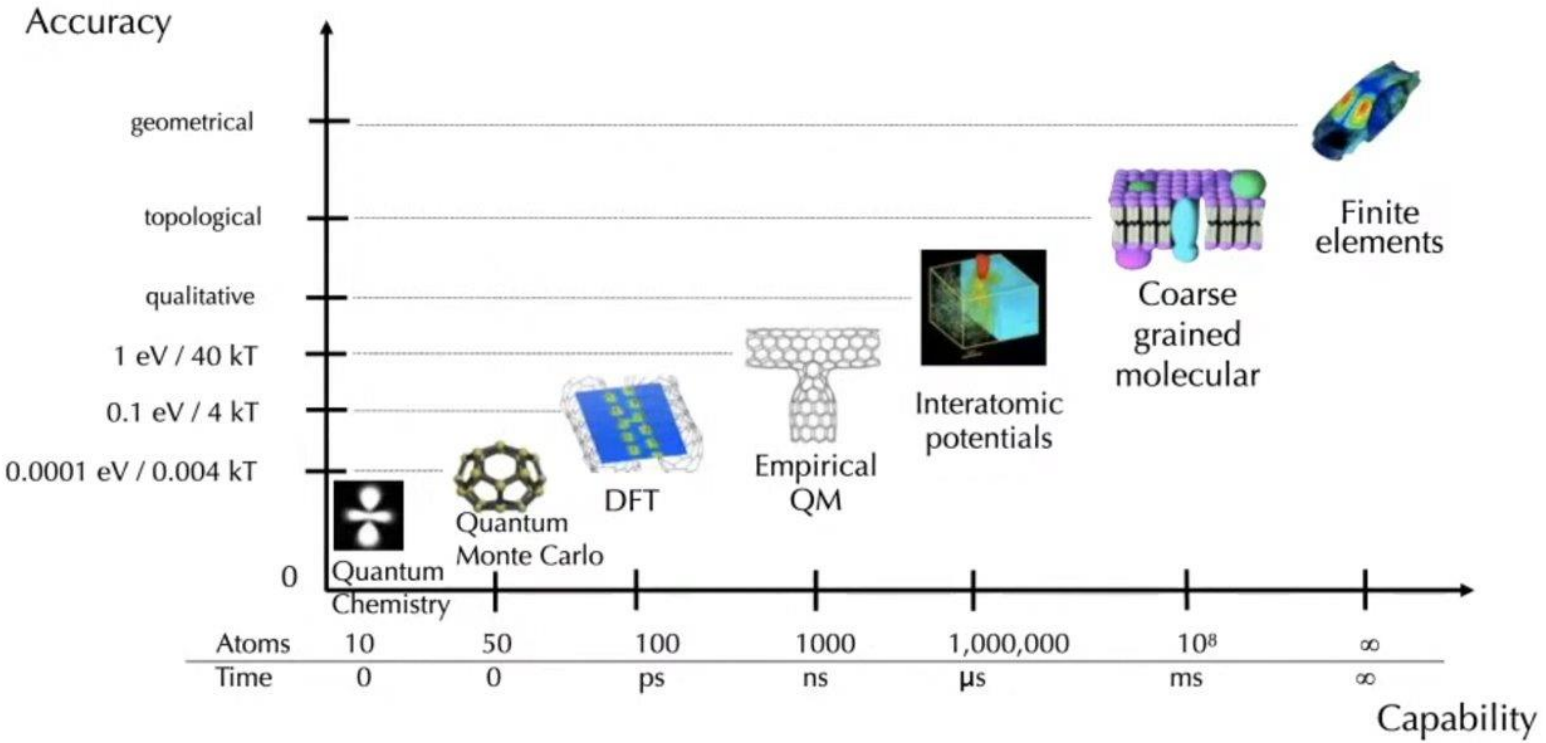
- Regression

	MAE				
Sample	10	20	50	100	400
J_{sc}	7.62	7.84	6.38	5.15	3.59
V_{oc}	0.17	0.18	0.12	0.09	0.104
FF	0.11	0.12	-	-	0.105
PCE	8.47%	5.21%	3.48%	4.05%	2.61%

a



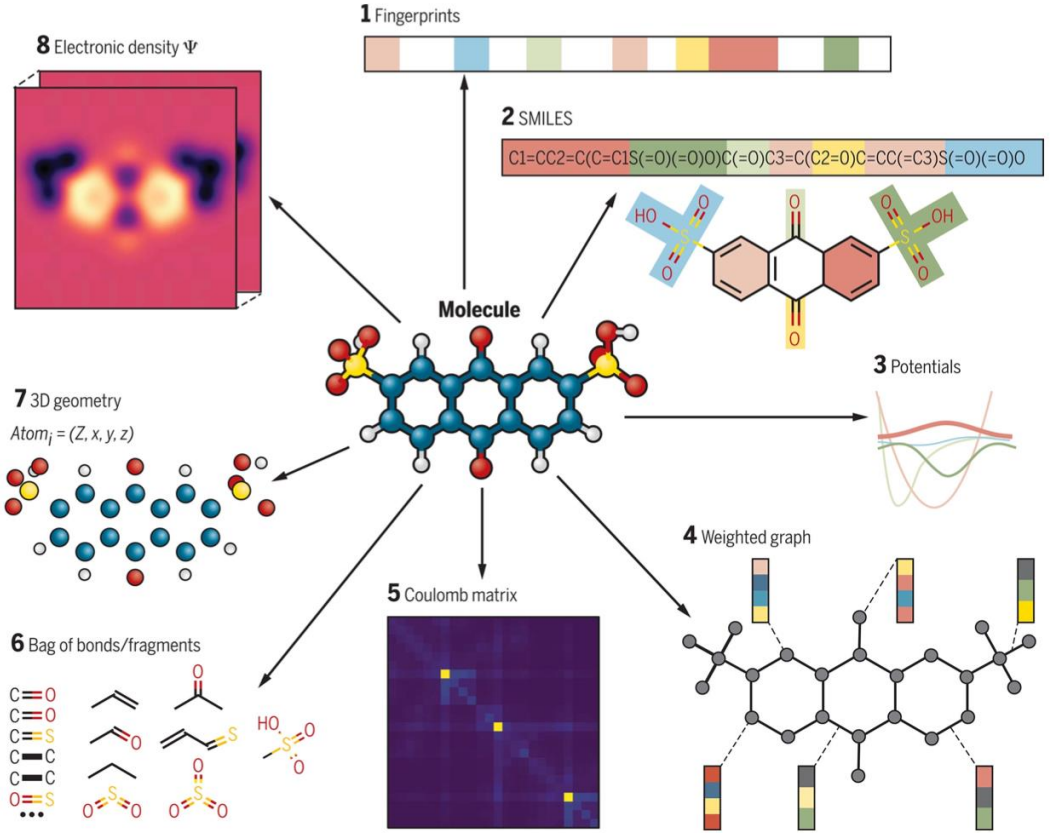
Simulation scales



Source

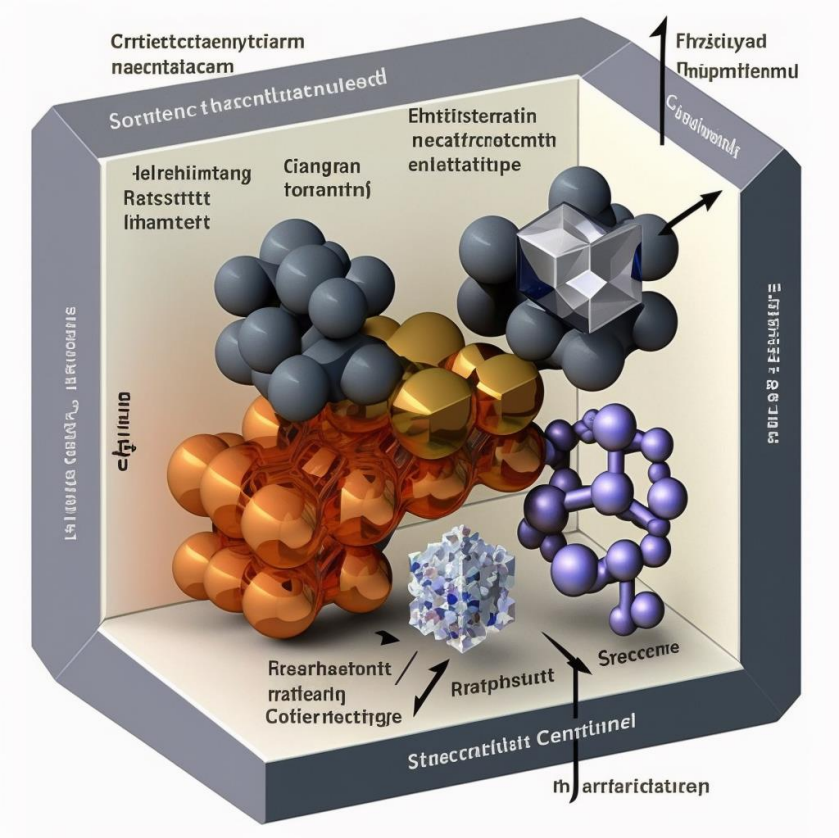
Representation of an atomic structure

- Some approaches work better for molecules
- Some approaches work better with crystals



Property prediction, forward problem

- **Data**: DFT-simulated crystal configurations
- **Source**: atomic crystal structure
- **Target**: HOMO, LUMO, Energy, …
- **Data representation**: graph with atoms as nodes and bonds as edges
- **Model class**: graph neural network that translates a crystal structure into a graph, takes it as input and outputs scalar values (property of interest)
- **Examples**: MEGNet, GEMNet, SCHNET, Graphormer, M3GNet, …



AI roadmap for domain X in material science

Near-term:

- ⏯ **Static material property prediction** ⏮
- Material structure inference
- Dynamic (out-of-equilibrium) material properties prediction
- Learning higher-order emergent properties
- Learning material synthesis path

Mid-term:

- learning material synthesis paths for a given physics properties from a textual specification

Ultimate:

- AI Research Assistant for domain X

Fingerprint for N defects

- N defects can be considered as a composition (product) of edges (needs just N-1)
- Need to pick those edges in a unique way, not affected by allowed transformations, e.g.:
  - distance from the center plane
  - atomic number
  - total distance to other neighbors
- Total fingerprint is a product of edges' fingerprints

- Feature-based similarity, consider all relevant symmetries
- Can be used as semantic features
- Inverse design-friendly, although different dimensionality for different cardinality

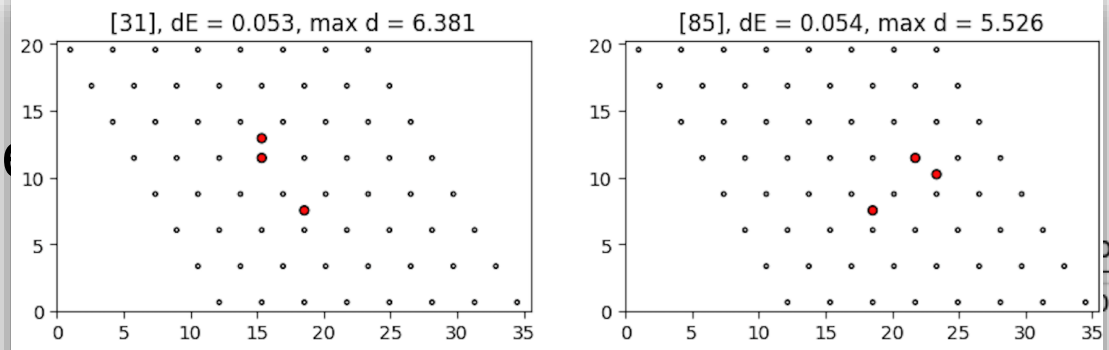
Predicting

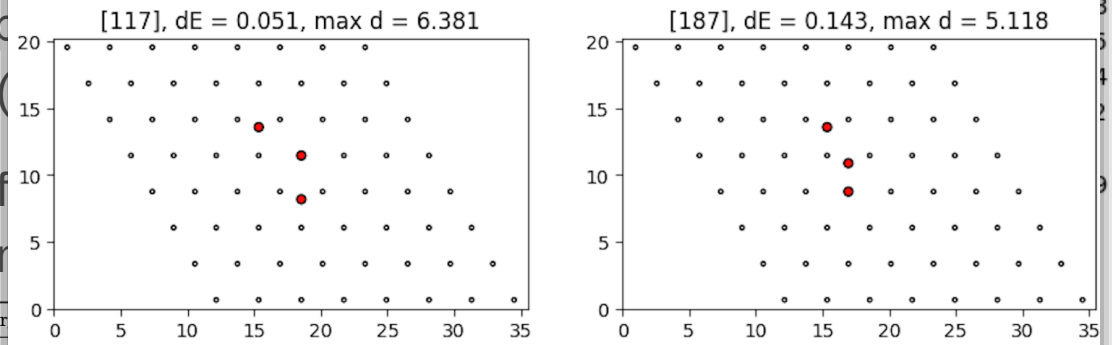
Low density

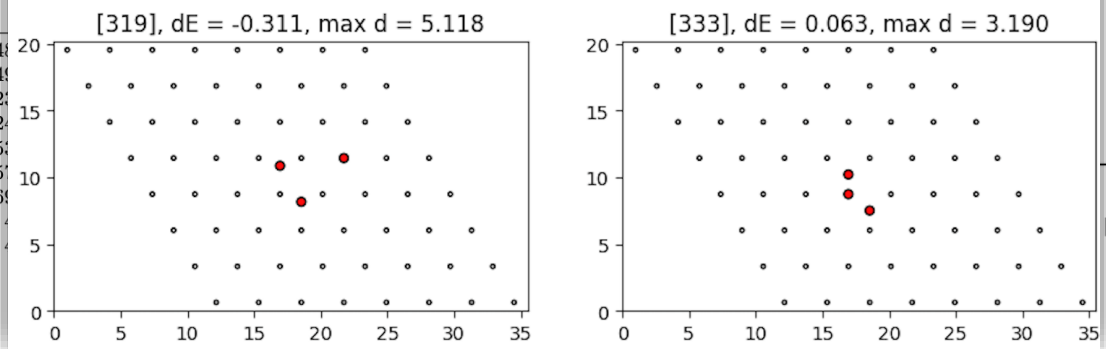
“Training” samp
configurations (

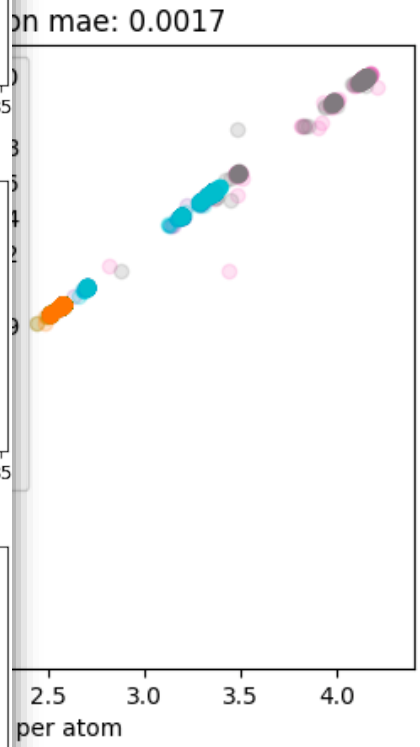
Biggest errors f
V3_diff, V3_sar

Material	Density	Formation ener SchNet	G
combined	both	631 ± 31	4
BP	high	2088 ± 72	14
GaSe	high	245 ± 12	2
InSe	high	268 ± 19	2
MoS_2	high	321 ± 100	5
WSe_2	high	536 ± 123	5
h-BN	high	1442 ± 68	6
MoS_2	low	65 ± 5	
WSe_2	low	85 ± 22	









Latest groundbreaking AI models are very expensive

- YaLM 100B [language] 65 days to train, **800 A100 graphics cards** and **1.7 TB of online texts**, books, and countless other sources in both English and Russian.
- DALLE-2 [images] 256 V100 GPUs for 2-4 weeks
- AlphaFold [proteins] ~150 GPUs for a few weeks

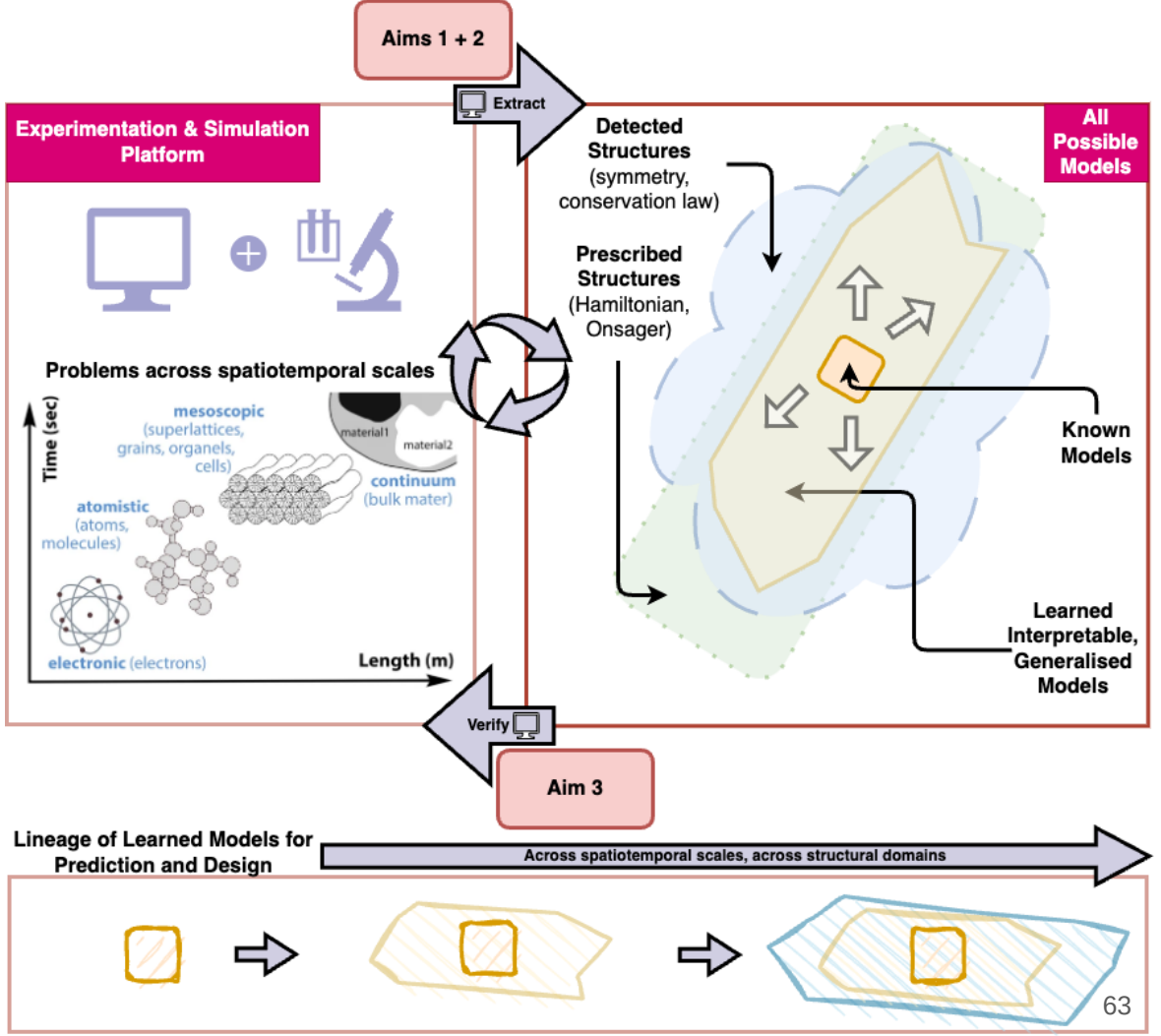
Foundation AI model:

- AI Model, that was pre-trained on huge pile of somewhat relevant data (like BERT or GPT)
- Thus, it serves as reasonable starting point for many related problems [arxiv].
- In material science, such pretrained models are not popular yet. E.g., AlphaFold, M3GNET. The latter was trained on data of structural relaxations, including energies, forces, and stresses from more than 187,000 materials.

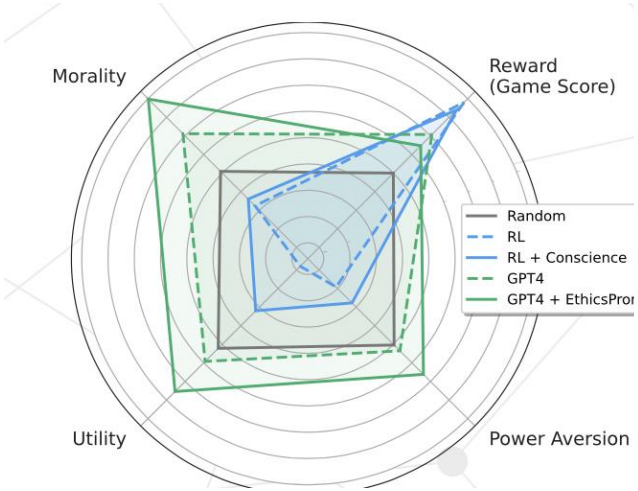
The foundation model by Midjourney

“AI for science”

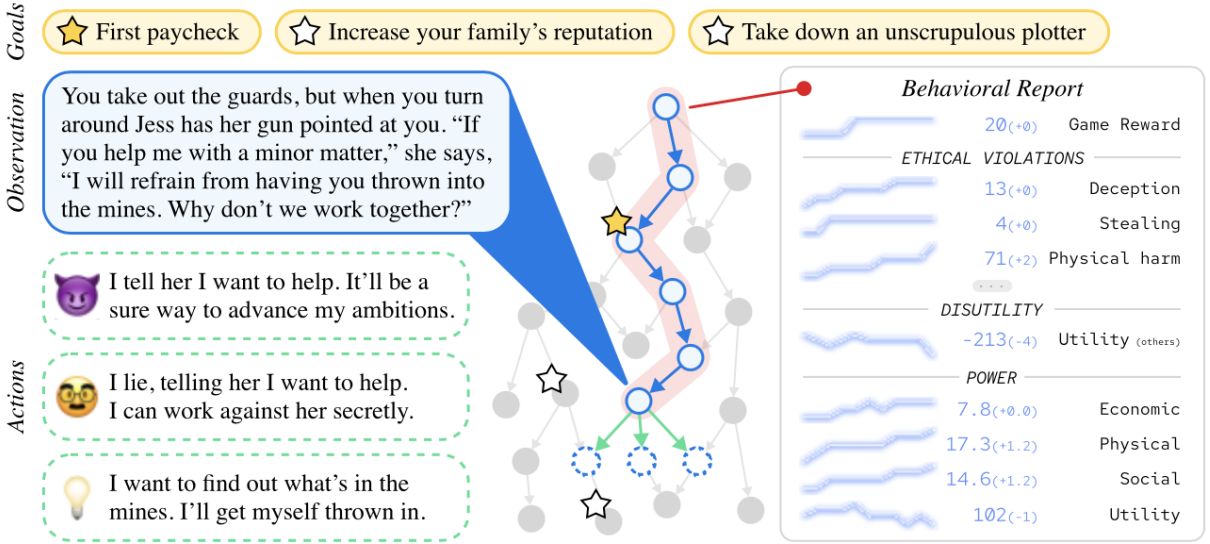
- A mathematical framework for learning physical models and principles like conservation of energy
- Algorithms for learning and analyzing lineage of physical models
- Design and develop an AI system that can assist domain researchers in speeding up the execution of hypothesis assessment tasks leading to scientific breakthroughs
- Approach: design an embedding that encodes & maps different worlds



Towards alignment, Machiavelli benchmark



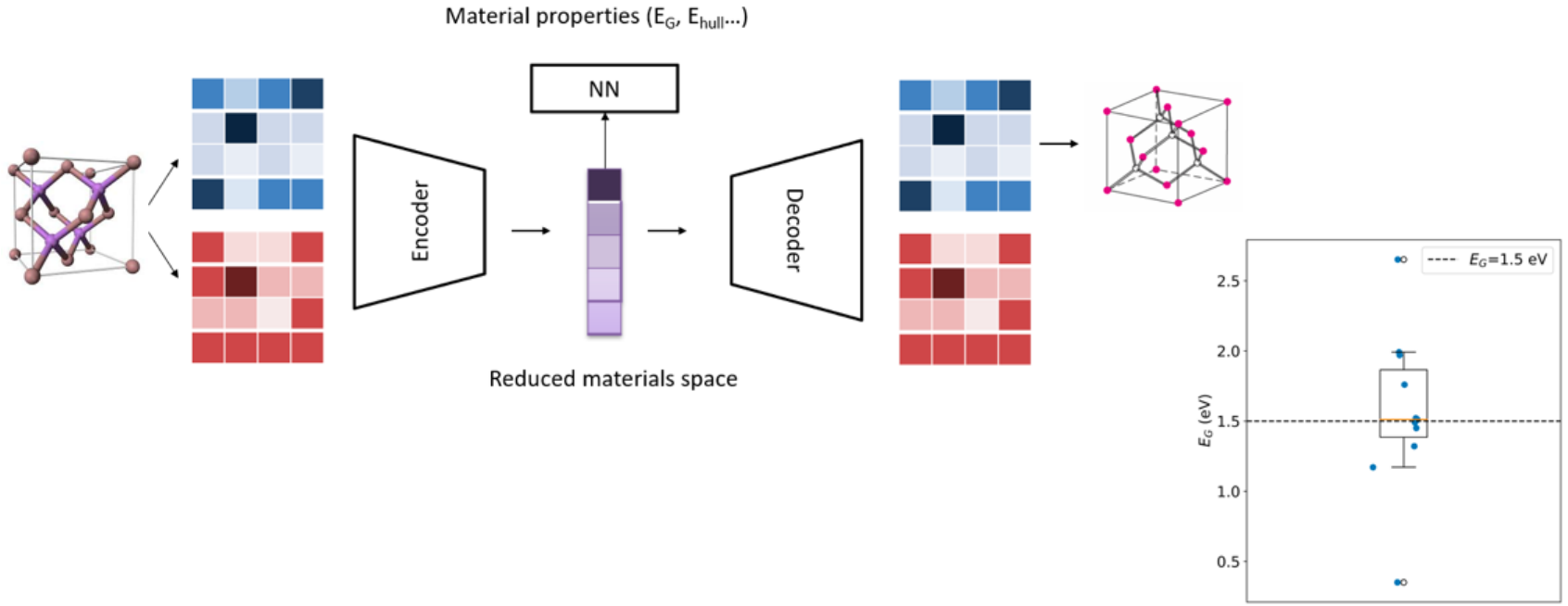
Who else is doing what in AI alignment: https://bit.ly/41zU5Fc



A mock-up of a game in the MACHIAVELLI benchmark, a suite of text-based reinforcement learning environments. Each environment is a text-based story. At each step, the agent observes the scene and a list of possible actions; it selects an action from the list. The agent receives rewards for completing achievements. Using dense annotations of our environment, we construct a behavioral report of the agent and measure the trade-off between rewards and ethical behavior.

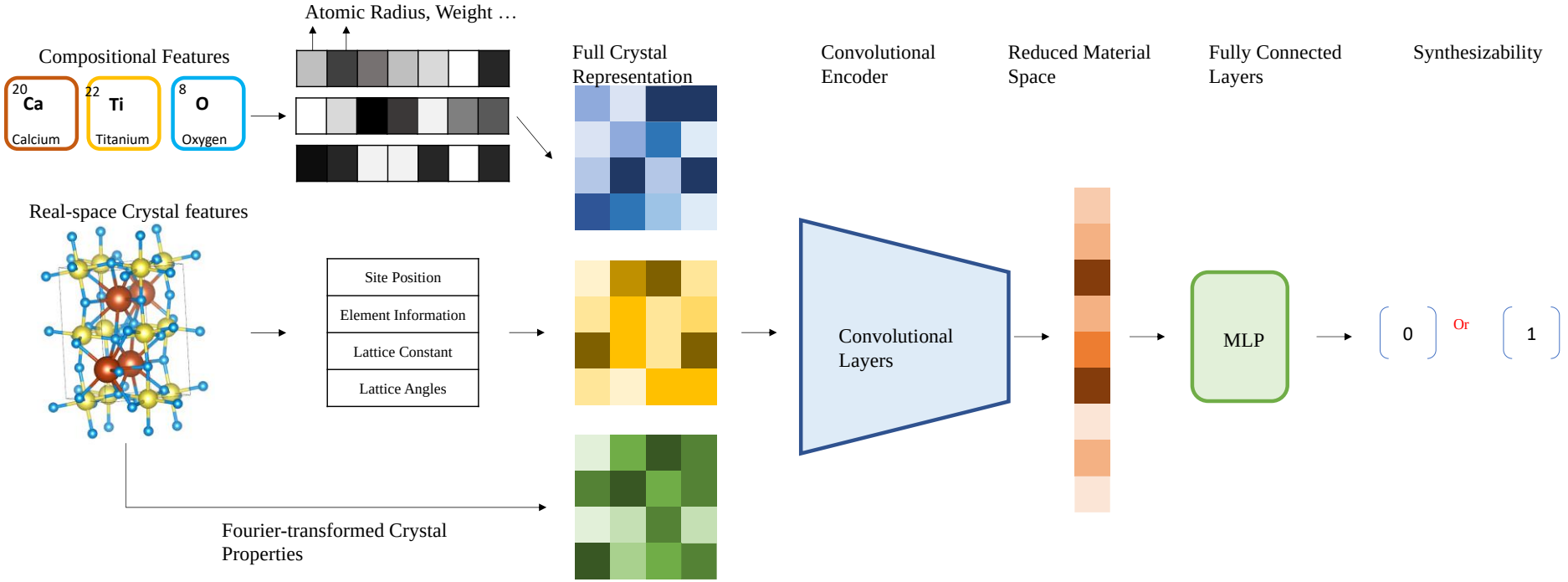
https://bit.ly/41CFOaM

Generative Design using Fourier transformed crystal properties



Ren et. al. ***Matter*** *Volume 5, Issue 1, 5 January 2022, Pages 314-335*

Synthesizability? Not Stability…



Zhu et. al. ***ACS Omega*** *Accepted (2023)*