Skoltech

LOMONOSOV MOSCOW STATE UNIVERSITY

AIRI

RSF
Russian Science Foundation

Crystallography and Crystal Chemistry
VIII International School-Conference of Young Scientists 2023

Data-driven solutions to model properties and accelerate the discovery of new materials

Dr. **Roman** A. **Eremin**

PhD in **Physics** and **Mathematics, Senior research scientist**

Artificial Intelligence Research Institute AIRI

Moscow, Russian Federation

November 11th, 2023

AGENDA

01 Data-driven solutions

02 Crystallographic tools

03 Modeling approaches

Goals

- Discovery of new materials
- Diversification of technology stacks
- Environmentally friendly production chains

Image «Зеленая энергетика» (style – detailed photo) generated by Kandinsky 2.1

01

Data-driven solutions

Materials discovery: the beginning of time

Stone Age

2.6 million years ago

400 000 years ago

Bronze Age

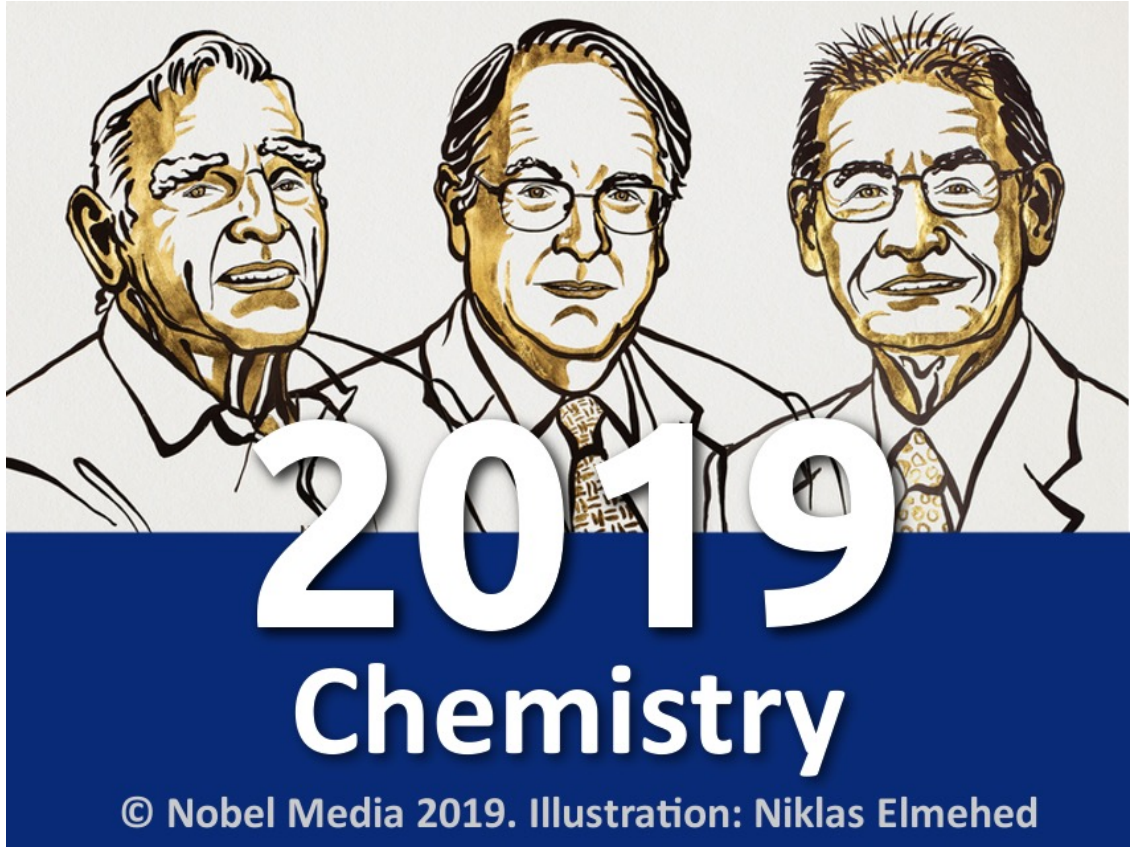
7 000 BC

Iron Age

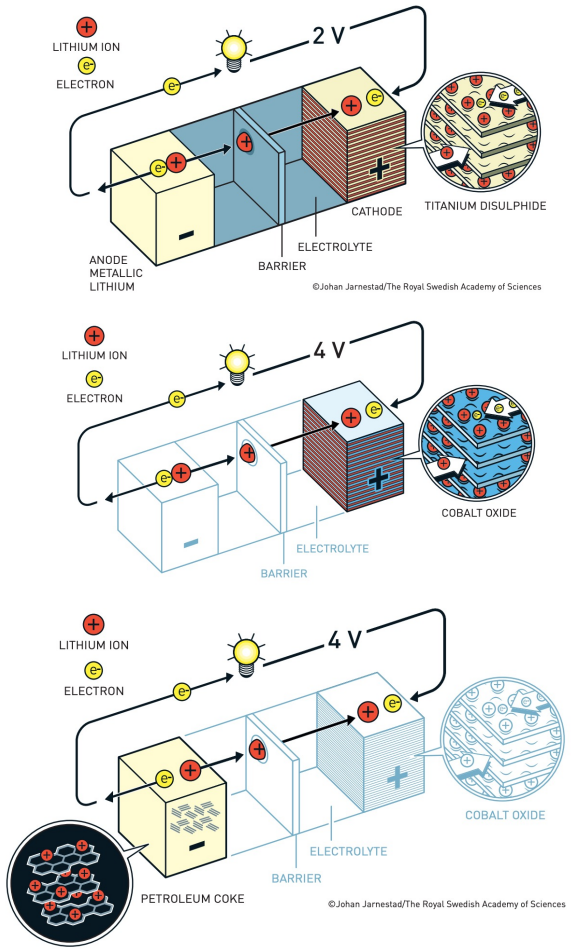
1 000 BC

Images generated by Kandinsky 2.1

From discovery to application



"for the development of lithium-ion batteries"



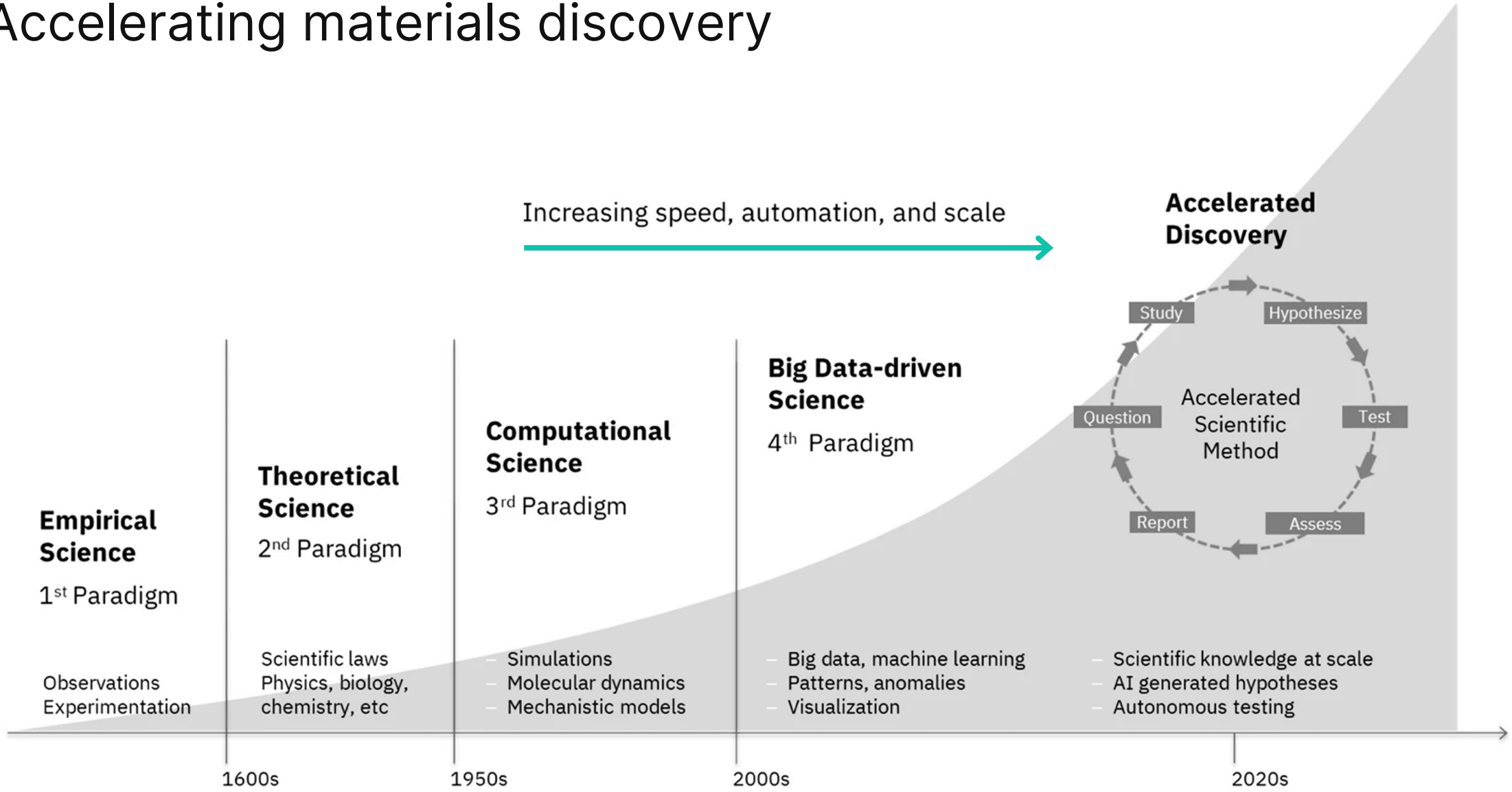
1973

1980

1985

1991

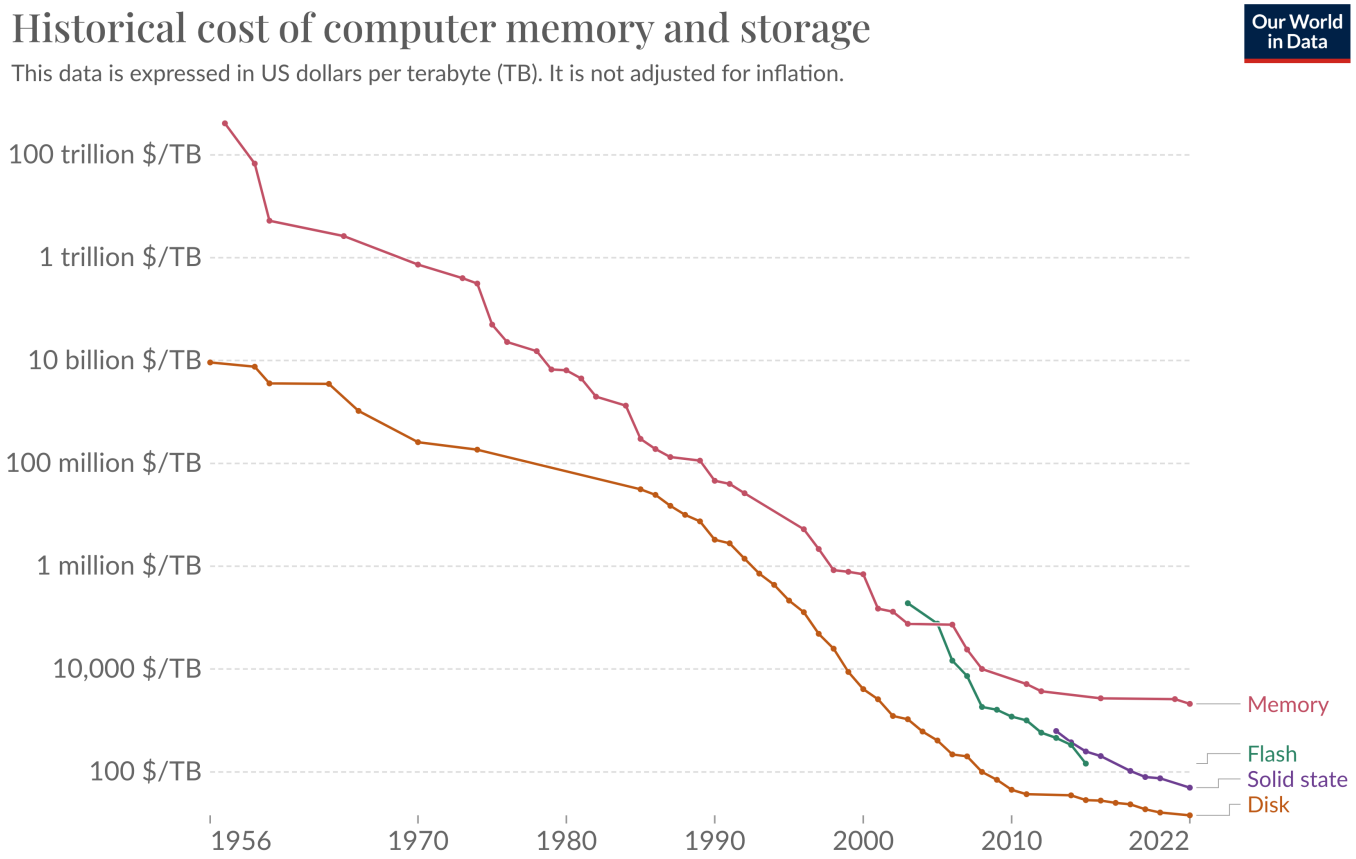
Accelerating materials discovery



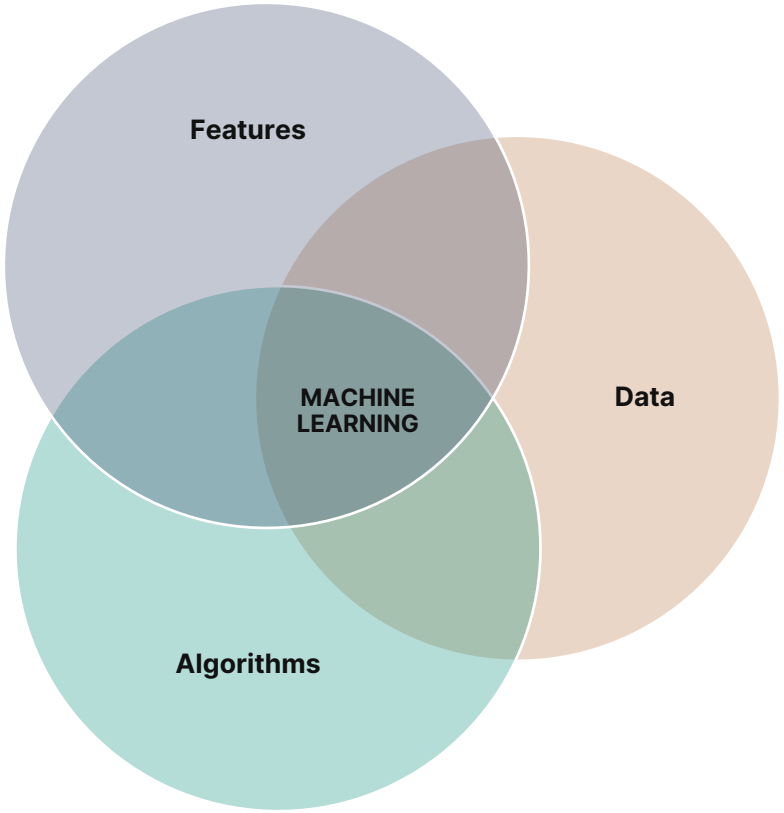
Pyzer-Knapp E. O. et al. (2022) Accelerating materials discovery using artificial intelligence, high performance computing and robotics. npj Computational Materials 8, 64.

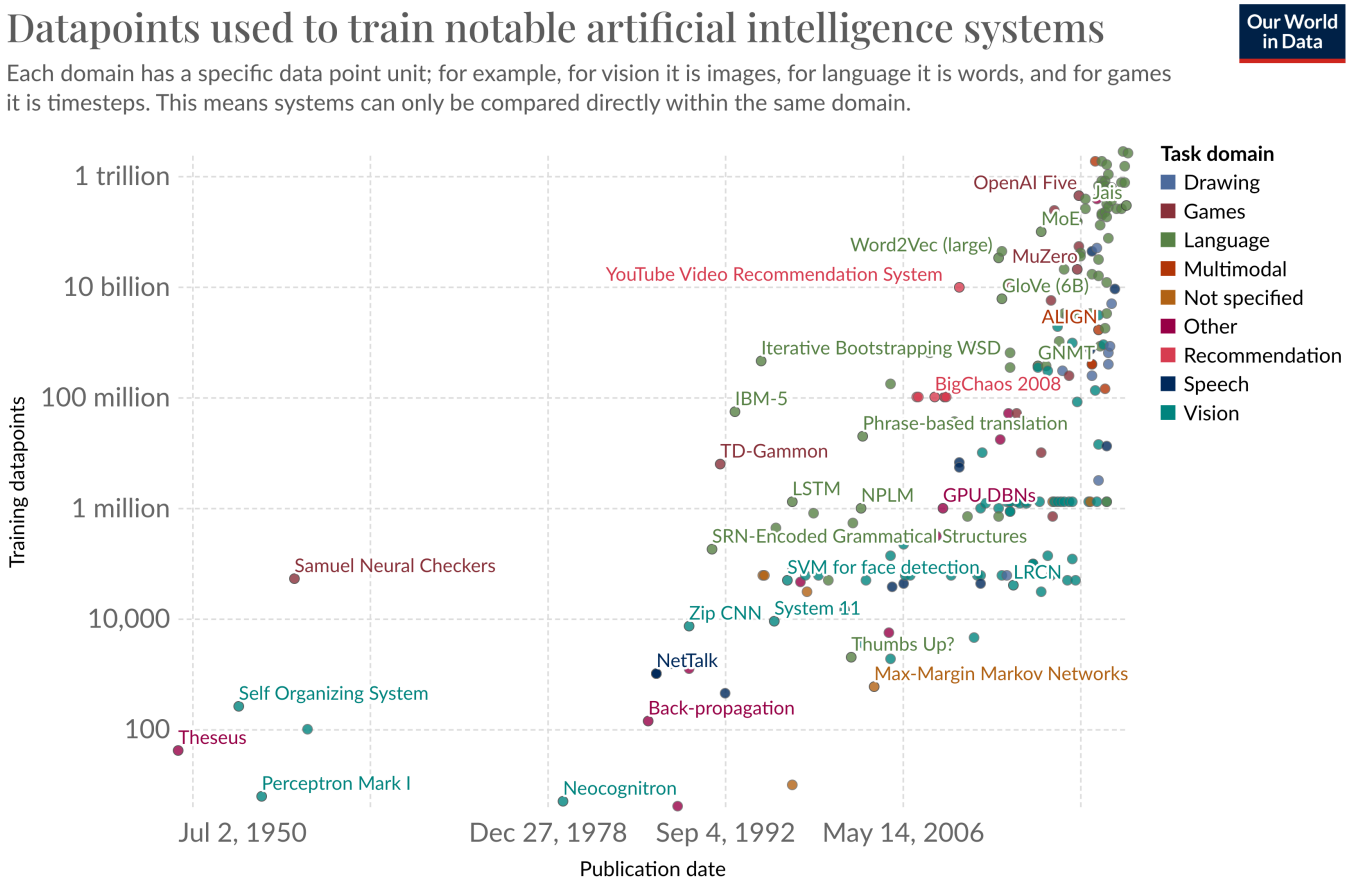
Is data really invaluable?

Image generated by Kandinsky 2.1:
«Компьютер создает и хранит новые знания»



Data in action





Data source: Epoch (2023)

OurWorldInData.org/artificial-intelligence | CC BY

Data on structures and thermodynamics of materials

Experimental data

ICSD

FIZ Karlsruhe – Leibniz Institute for Information Infrastructure

31.10.2023

ICSD now contains 291,382 crystal structures

The ICSD web version 5.1.0 is now online.

The Cambridge Structural Database (CSD)

The comprehensive repository of validated and curated small-molecule organic and metal-organic crystal structures.

Established in 1965 with historical structures dating back to the 1920s, the Cambridge Structural Database (CSD) now contains over 1.25M accurate 3D structures with data from X-ray and neutron diffraction analyses and additional curation from the CCDC. The database is used by researchers across the pharmaceutical, agrochemical, and fine chemicals industries to predict and guide future discoveries.

Computational datasets

The Materials Project

MATERIALS	154,718
INTERCALATION ELECTRODES	4,351
MOLECULES	172,874

AFLOW

Automatic - FLOW for Materials Discovery

3,479,057 form. enthalpies	366,988 band structures	172,488 Bader charges	5,650 elastic properties
5,664 thermal properties	1,738 binary systems	30,289 ternary systems	150,659 quaternary systems

AIRI

Search space

Group ▸ / Period ▾	1	2		3	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18
																			Noble gases
Nonmetals 1	1 H																		2 He
Metals 2	3 Li	4 Be												5 B	6 C	7 N	8 O	9 F	10 Ne
3	11 Na	12 Mg												13 Al	14 Si	15 P	16 S	17 Cl	18 Ar
4	19 K	20 Ca		21 Sc	22 Ti	23 V	24 Cr	25 Mn	26 Fe	27 Co	28 Ni	29 Cu	30 Zn	31 Ga	32 Ge	33 As	34 Se	35 Br	36 Kr
5	37 Rb	38 Sr		39 Y	40 Zr	41 Nb	42 Mo	43 Tc	44 Ru	45 Rh	46 Pd	47 Ag	48 Cd	49 In	50 Sn	51 Sb	52 Te	53 I	54 Xe
6	55 Cs	56 Ba	La to Yb	71 Lu	72 Hf	73 Ta	74 W	75 Re	76 Os	77 Ir	78 Pt	79 Au	80 Hg	81 Tl	82 Pb	83 Bi	84 Po	85 At	86 Rn
7	87 Fr	88 Ra	Ac to No	103 Lr	104 Rf	105 Db	106 Sg	107 Bh	108 Hs	109 Mt	110 Ds	111 Rg	112 Cn	113 Nh	114 Fl	115 Mc	116 Lv	117 Ts	118 Og
	s-block (incl. He)		f-block	d-block										p-block (excl. He)					

Transition metals (sometimes excl. group 12)

Some elements near the dashed staircase are sometimes called *metalloids*

Lanthanides	57 La	58 Ce	59 Pr	60 Nd	61 Pm	62 Sm	63 Eu	64 Gd	65 Tb	66 Dy	67 Ho	68 Er	69 Tm	70 Yb
Actinides	89 Ac	90 Th	91 Pa	92 U	93 Np	94 Pu	95 Am	96 Cm	97 Bk	98 Cf	99 Es	100 Fm	101 Md	102 No

more than 10^{100} combinations

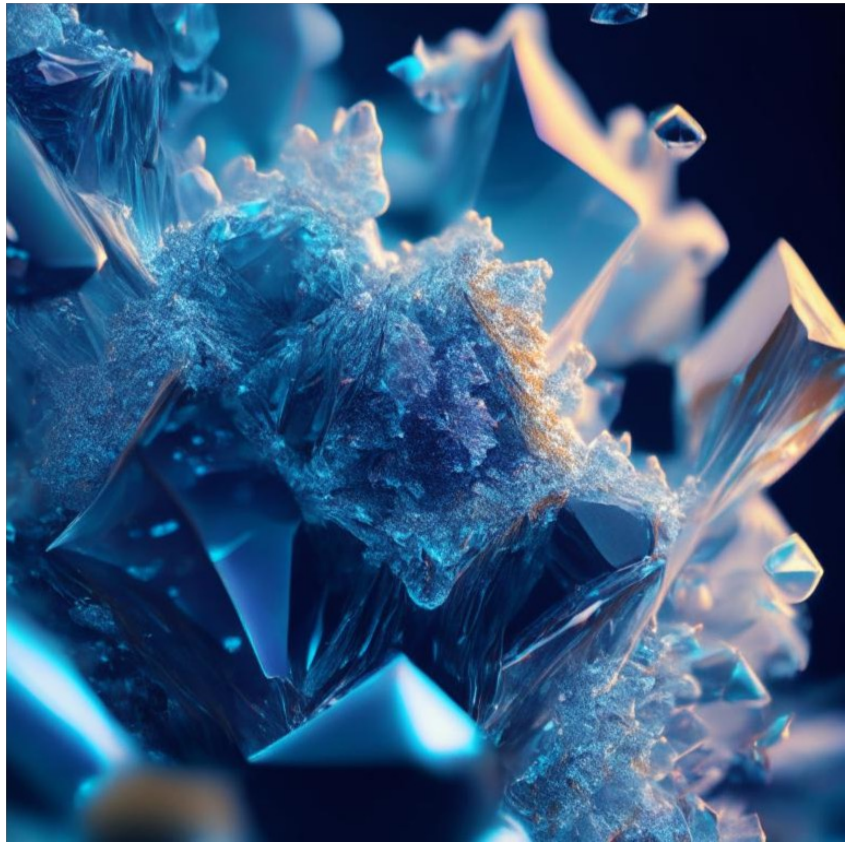


Image «Кристаллическая структура вещества (style – detailed photo) generated by Kandinsky 2.1

AI workflows in materials science

	Data	Algorithm	Output
1st generation Structure-property calculations	Structure	Local optimization	Property
2nd generation Crystal structure predictions	Composition	Global optimization	Structure + property
3rd generation Data-driven design of new materials	Chemical and physical data	Machine learning	Composition + structure + property

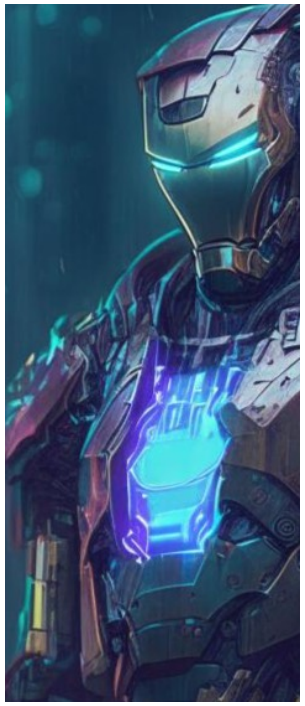


Image «J.A.R.V.I.S. Железный человек» (style – cyberpunk) generated by Kandinsky 2.1

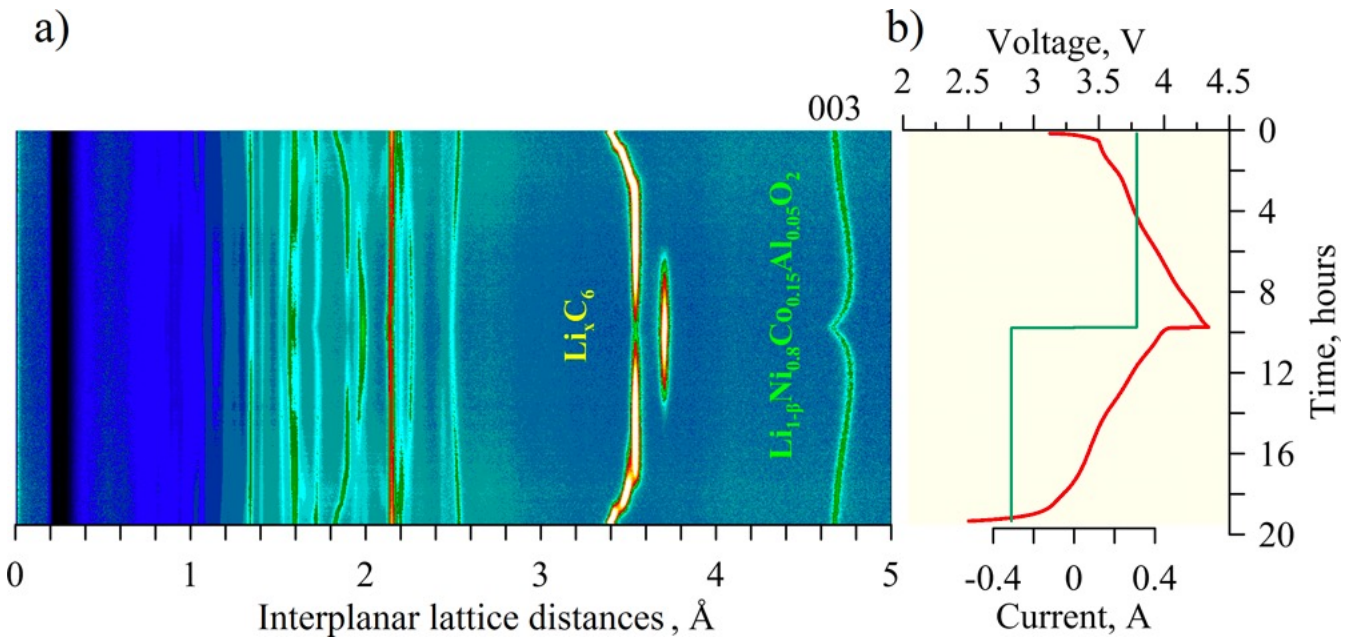
Butler, K. T., Davies, D. W., Cartwright, H., Isayev, O., & Walsh, A. (2018). Machine learning for molecular and materials science. *Nature*, *559*(7715), 547.

02

Crystallographic tools and modeling

Cathode materials

Operando studies of NCA

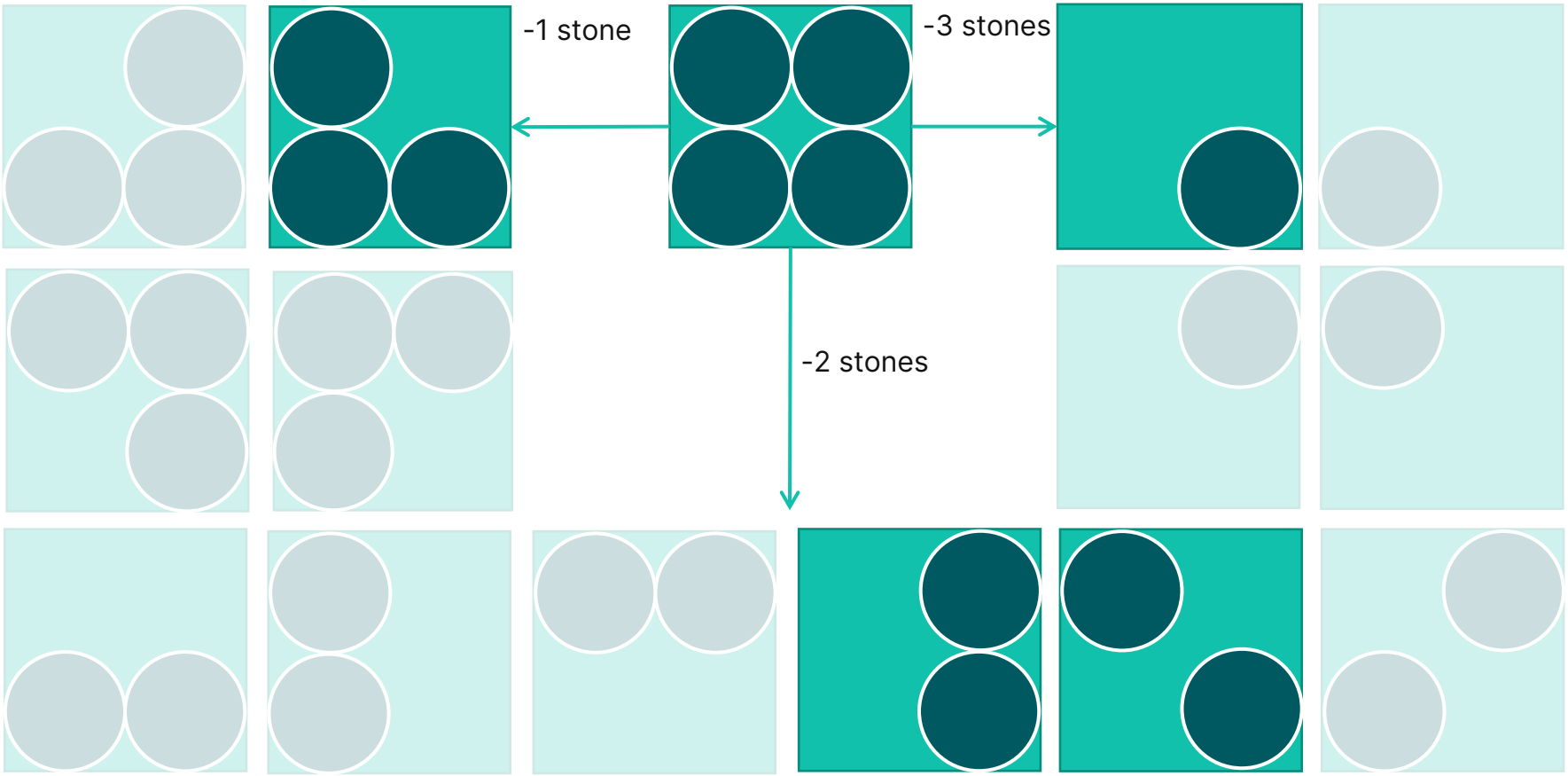


(a) Evolution of neutron diffraction patterns during a charge–discharge process. The intensive diffraction peaks at d = 3.4–3.7 Å correspond to Li_xC_6 phases, and that at d ≈ 4.7 Å is a 003 reflection of the NCA cathode material. (b) The corresponding changes of voltage and current.

J. Phys. Chem. C 2017, 121, 28293–28305

Cathode materials

Composition/configuration spaces of deintercalation



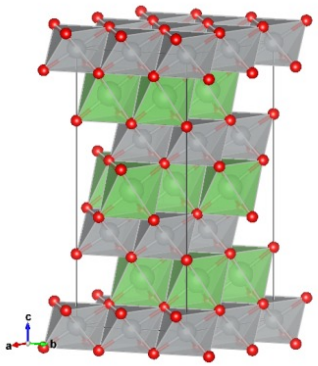
14 possible structures → **4** symmetrically inequivalent structures

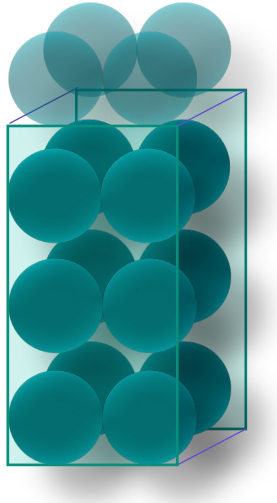
Cathode materials

Composition/configuration spaces of LNO and NCA

$LiNiO_2$

2x2x1 supercell



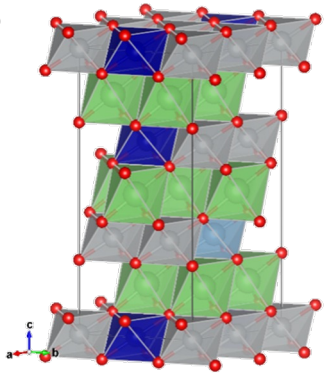


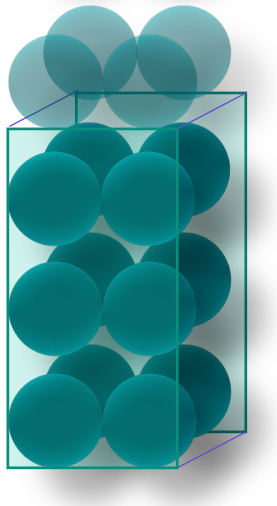
Remove Li ions

4096 / **87** structures

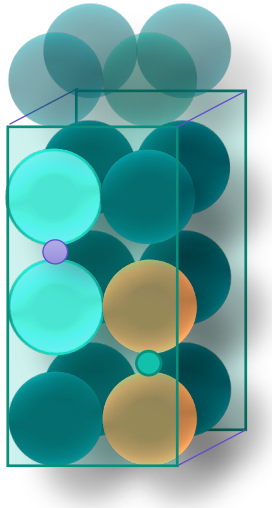
$LiNi_{0.8}Co_{0.15}Al_{0.05}O_2$

2x2x1 supercell





Add substituent



Remove Li ions

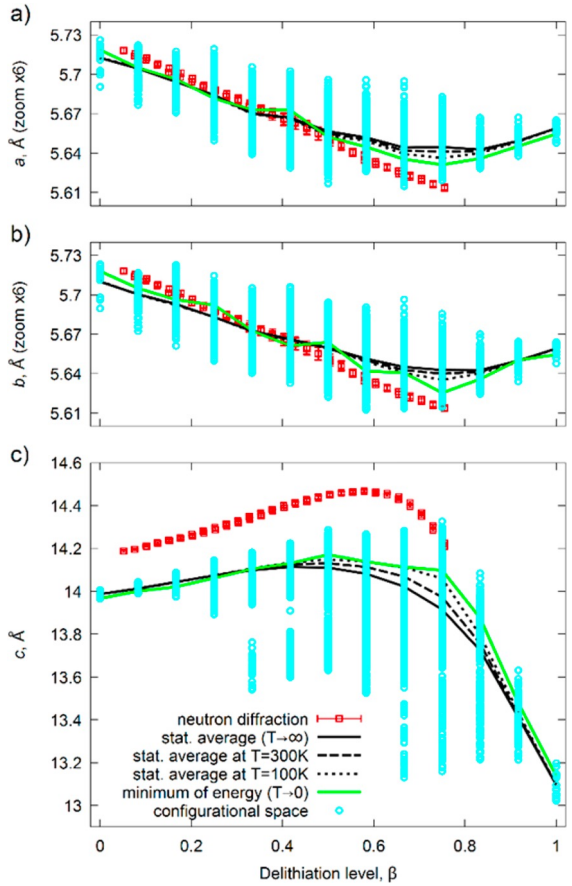
57420 / **20760** structures

J. Phys. Chem. C 2017, 121, 28293−28305

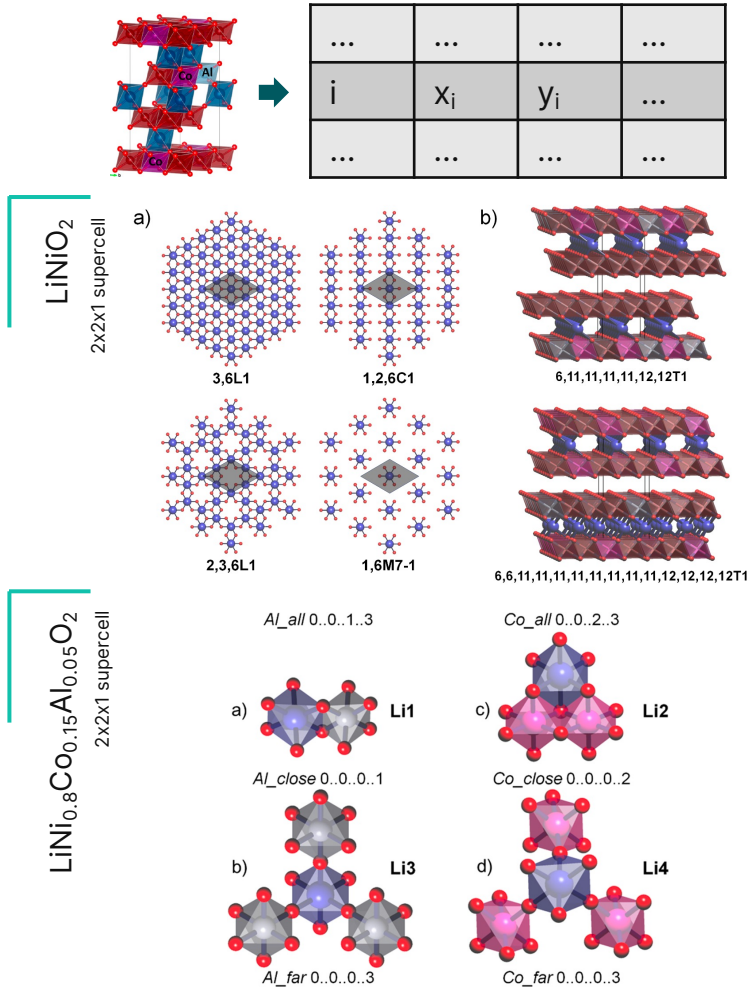
AIRI

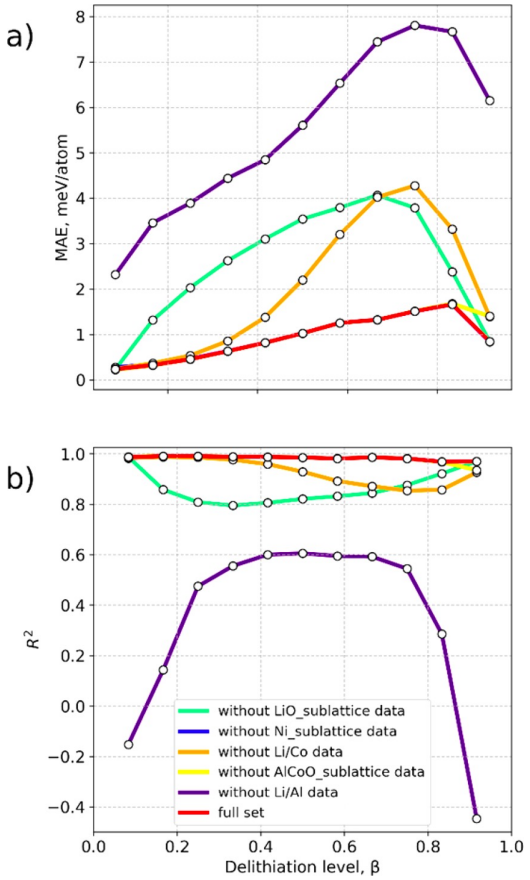
Cathode materials

DFT results for NCA and ML approaches



Comparison of the data of operando neutron diffraction (open squares) with calculated (a) a, (b) b, and (c) c lattice parameters of the NCA configurational space (open circles) within the "PBE-vdW" model.



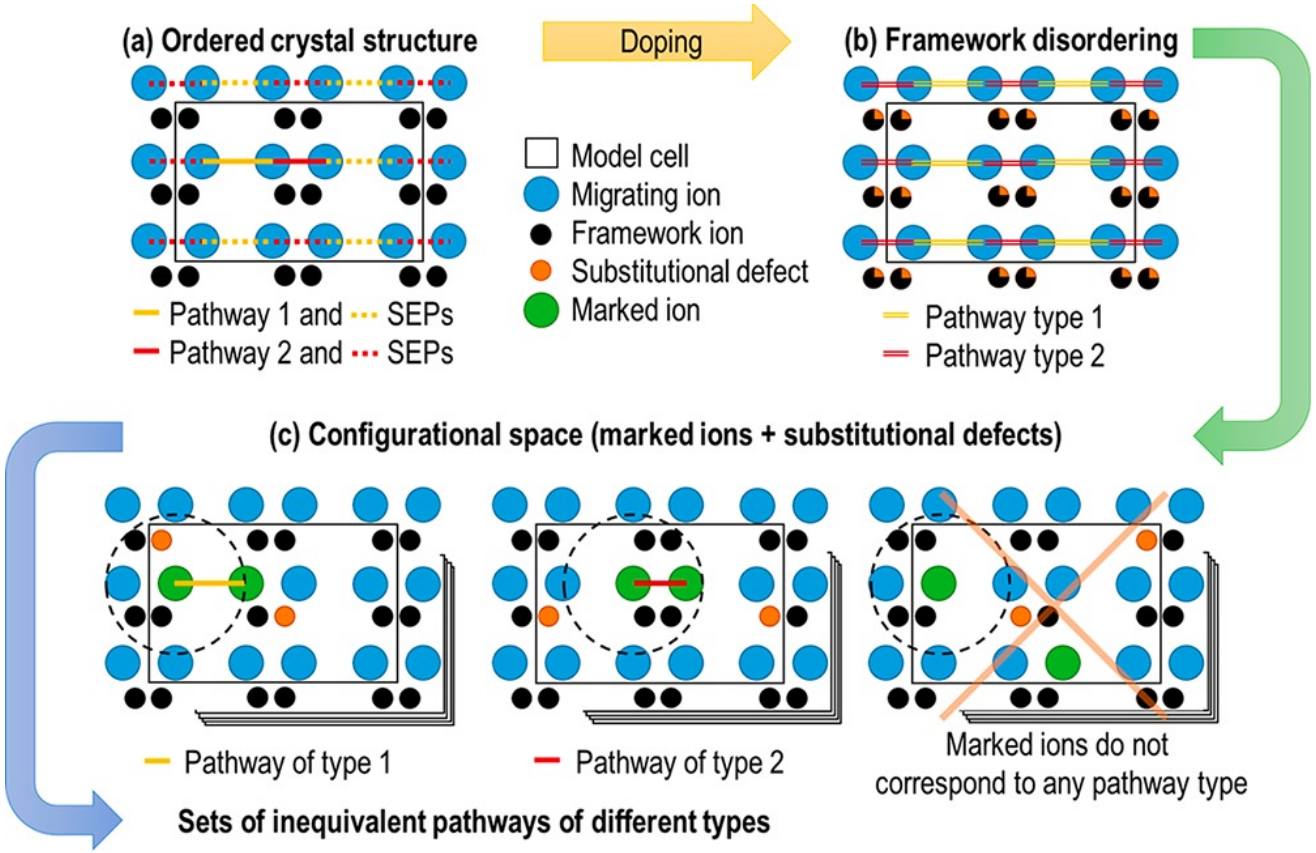


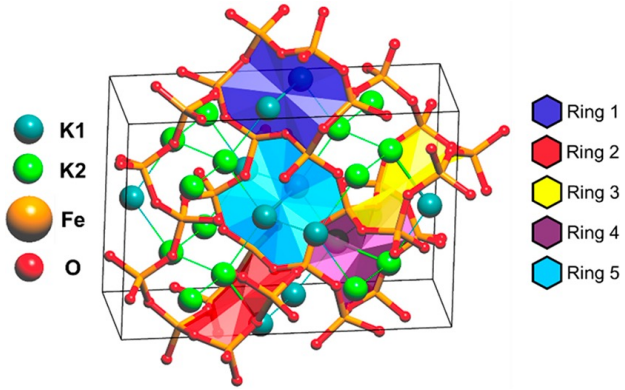
Dependencies of the estimators of energy prediction quality: (a) MAE and (b) R^2 determination coefficient versus delithiation for the ridge regression models trained in a sequentially reduced set of structural descriptors.

J. Phys. Chem. C 2017, 121, 28293–28305

Solid electrolytes

Symmetrically inequivalent K^+ pathways in Ti-doped $KFeO_2$





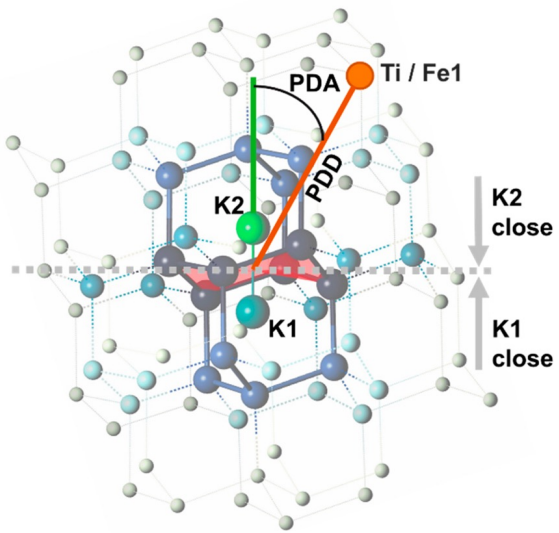
$KFeO_2$ crystal structure with five indicated $\{FeO\}_6$ rings corresponding to five inequivalent K^+ migration pathways.

Fe → Ti substitutions	Inequivalent structure realizations	Inequivalent K^+ pathways
1	64	128
2	15 552	59 520
3	1 537 600	8 630 400

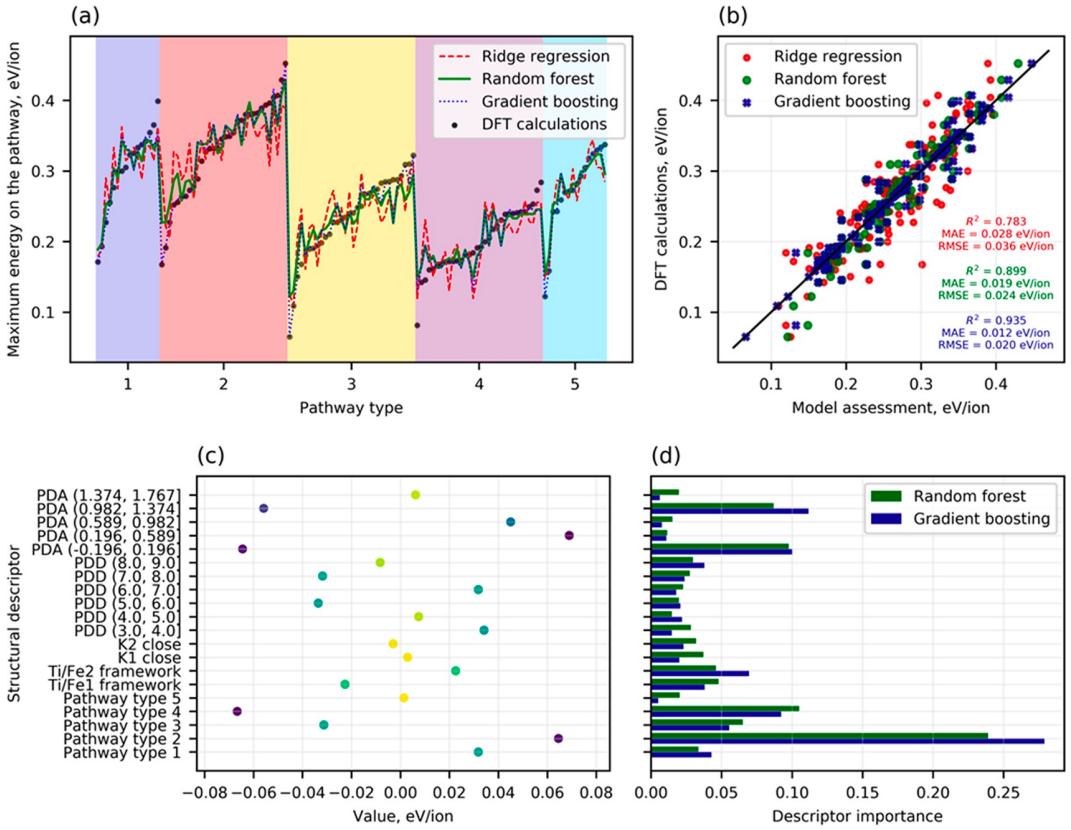
J. Phys. Chem. C 2019, 123, 29533–29542

Solid electrolytes

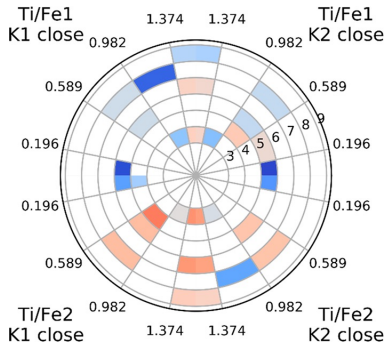
ML predictions of K^+ migration in Ti-doped $KFeO_2$



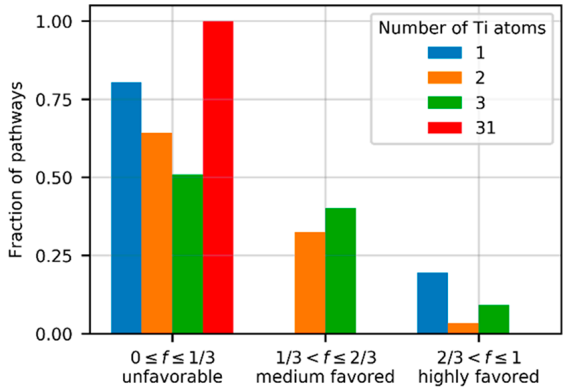
…	…	…	…
i	x_i	y_i	…
…	…	…	…



(a) Comparison of the DFT calculated maximum energies along the pathways (dots) with the different model predictions (lines) with respect to the pathway types. (b) Scatter plot of the DFT calculated maximum energies vs model assessments. While (c) summarizes the coefficients of the ridge regression model, (d) visualizes the results of the feature importance analysis within the random forest and gradient boosting regression models for 20 introduced categories of structural descriptors (PDA: intervals in radians; PDD: intervals in angstroms).



Ridge regression assessments of the maximum energies with respect to the values of the structural descriptors. PDD and PDA are the coordinates of the polar-type system.

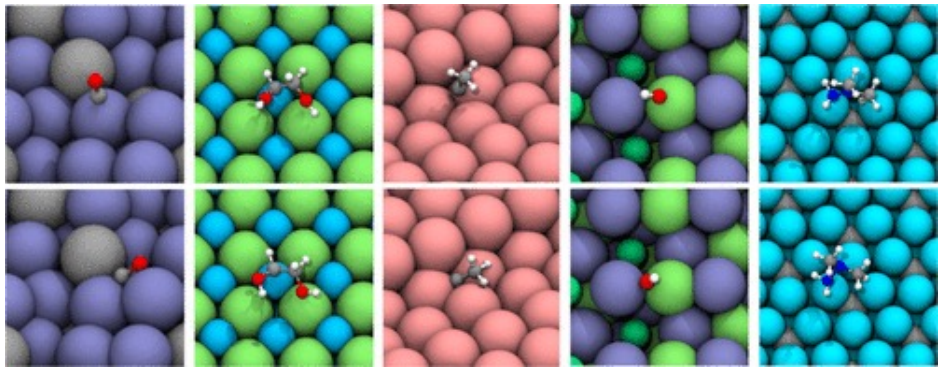


Pathway statistics with respect to their favorability f at the studied doping levels of the K1−xFe1−xTixO2 structure (1, 2, and 3 Ti atoms in the model cell).

J. Phys. Chem. C 2019, 123, 29533–29542

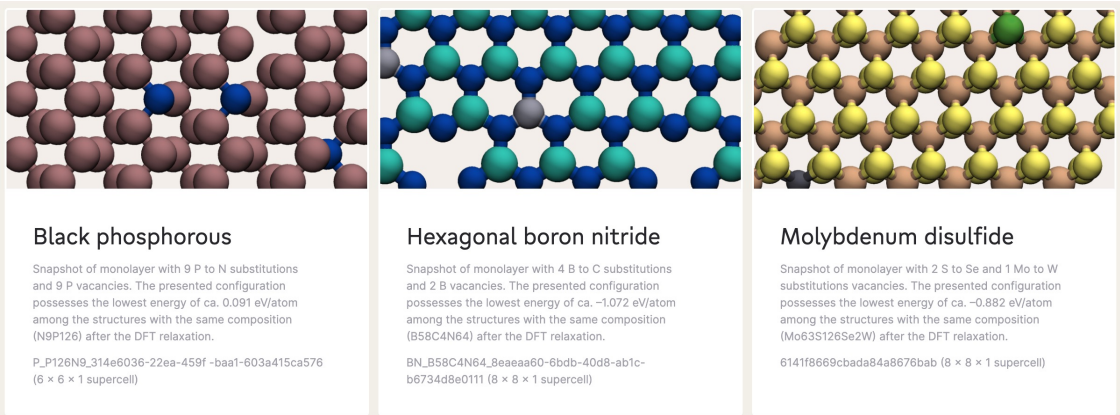
Electrocatalysis and optoelectronics

New materials discovery through the structure-to-property predictions



ACS Catalysis 2021, 11, 6059 – 6072

	Number	Inequivalent combinations
Adsorbate molecules	82	1 281 040 (~264 890 000)
Adsorbent materials	11 451	



	Number	Inequivalent combinations
Pristine 2D monolayers (bases)	6	Low defect contents: 11866 (MoS_2, WSe_2)
Point defects	vacancies and substitutions	High defect contents: 3000 (all bases)

npj 2D Materials and Applications 2023, 7, 6
https://2dmd.airi.net/

Crystal structure representations and AI models

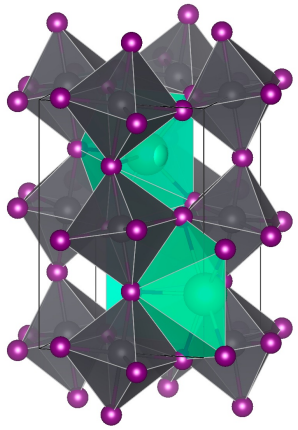
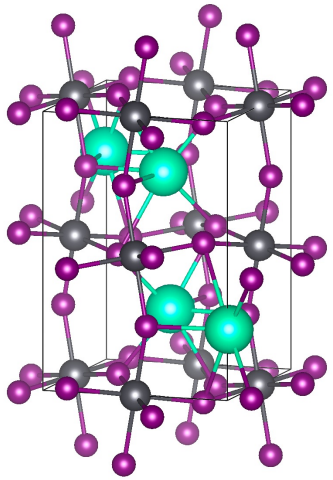
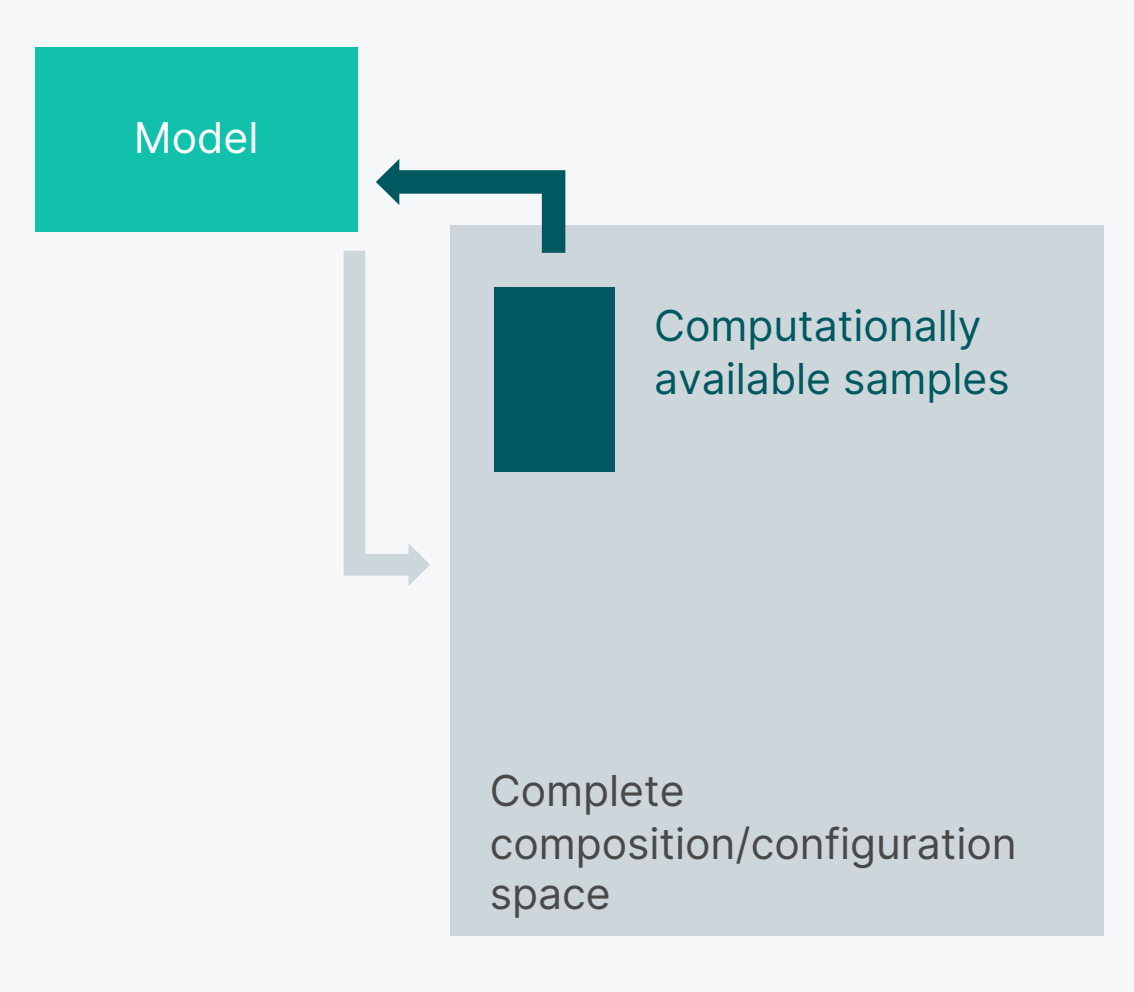


Table of descriptors

Graph representation

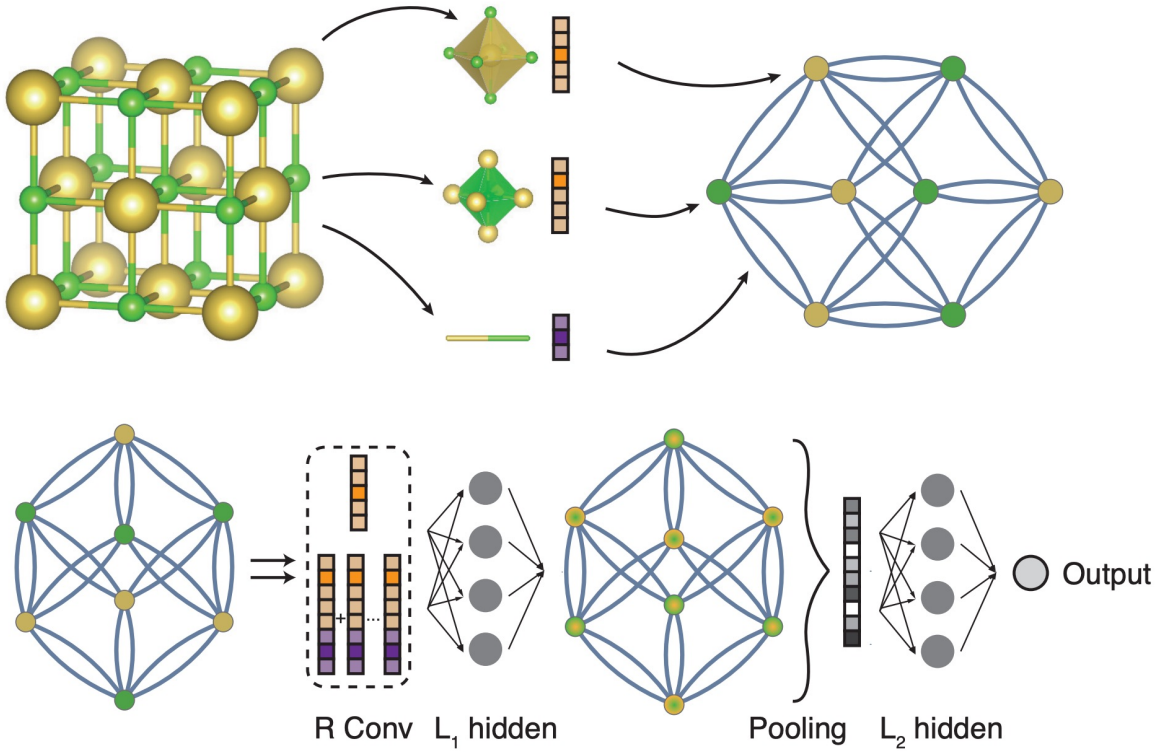
descriptor	value
composition	...
atomic coordinates	...
coordination numbers	...
defect content	...
structural motifs	...





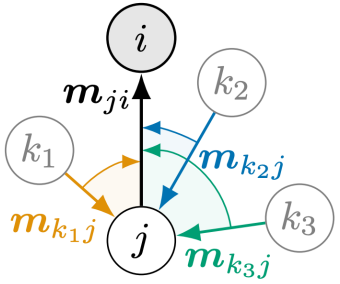
Graph neural networks

CGCNN



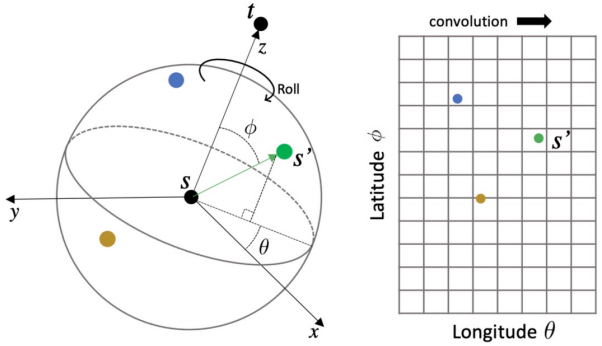
Phys. Rev. Lett. 2018, 120, 145301

DimeNet



*arXiv preprint arXiv:*2003.03123 (2020)

SpinConv

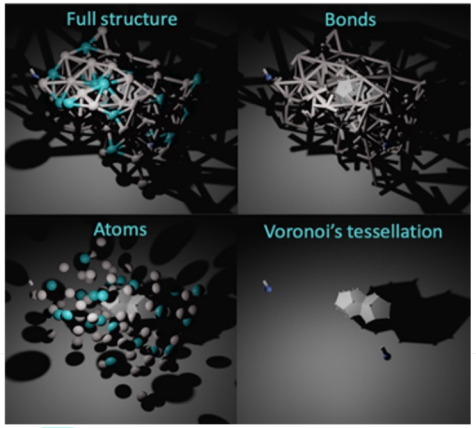


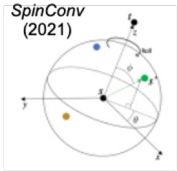
*arXiv preprint arXiv:*2106.09575 (2021)

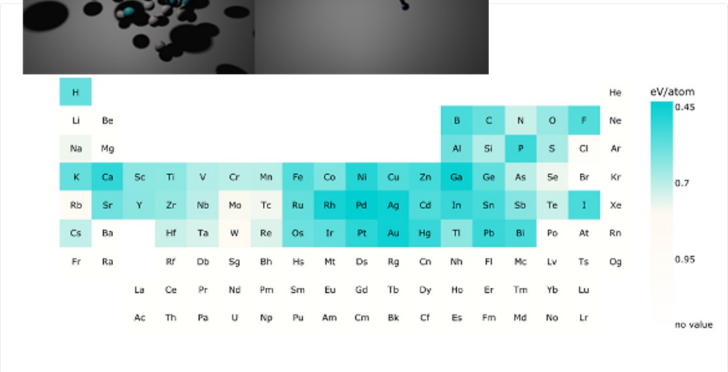
Discovering new materials

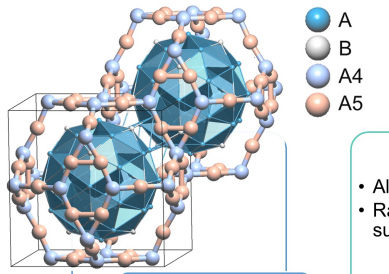
from structure-to-property predictions to synthesizability assessments

Physics-informed neural networks









DFT modeling and model training

Initial structure

Disorder introduction

Simplified CCS

- All types of defects
- Random subsampling

A4
A5

Enhanced CCS

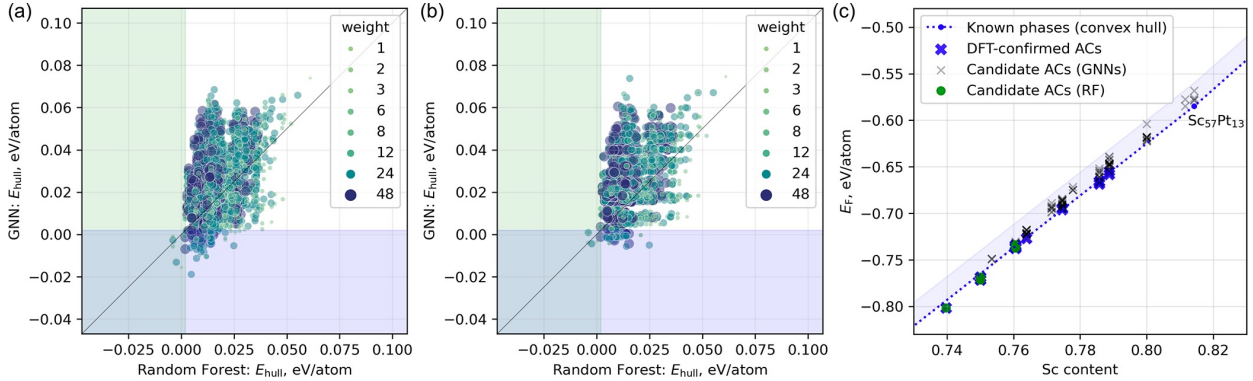
- Energetically favored types of defects
- All symmetrically inequivalent structures

A4

Model inference

Promising phases

- Data-driven estimation of thermodynamic properties
- Computational verification



For the Sc-Pt systems, comparison of EHs obtained by the RF regression and GNN-based models (a) without taking into account defect contents and (b) considering them by GNN-B. (c) Formation energies (gray crosses) calculated for the selected structures and the Sc2Pt–Sc57Pt13–Sc convex hull (blue dotted polyline). The structures with DFT-confirmed EHs below 2 meV/atom and those of them predicted by the RF regression are highlighted (blue crosses and green circles, respectively).

Cryst. Growth Des. 2022, 22, 4570–4581
Materials Today Chemistry 2023, 30, 101541

Discovering new materials

from structure-to-property predictions to structural stability

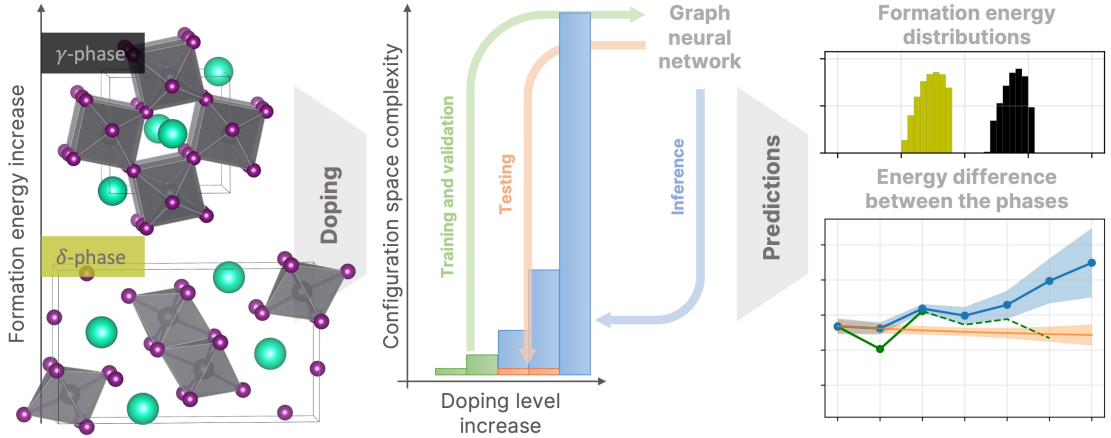
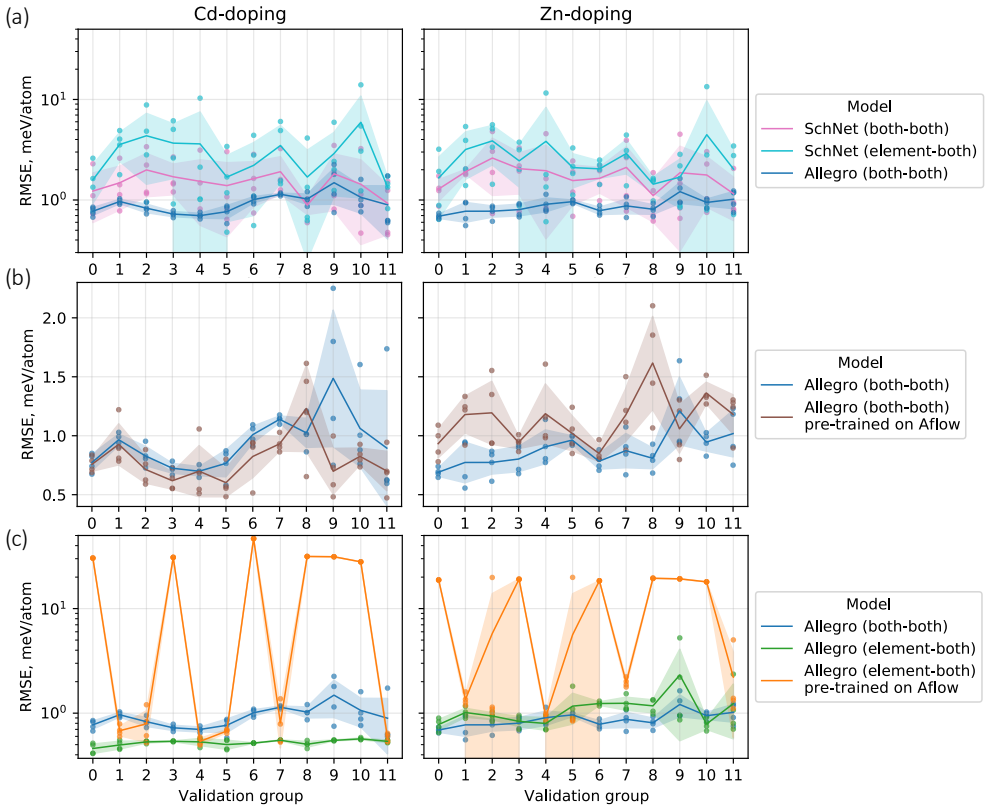


Table 1: Statistics of the obtained CCSs with mention of the corresponding assignment of data within the data-driven approach and corresponding calculation scenarios.

Pb atoms substituted	Number of inequivalent structures		Data for GNN		Calculation scenario
	γ (black) phase	δ (yellow) phase	train val	test	
0	1	1	all	0	DFT (full CCS)
1	1	1	all	0	
2	15	16	all	0	
3	87	87	18	5	DFT (subsample) +GNN
4	632	637	0	5	
5	3 225	3 225	0	5	
6	14 509	14 544	0	0	GNN

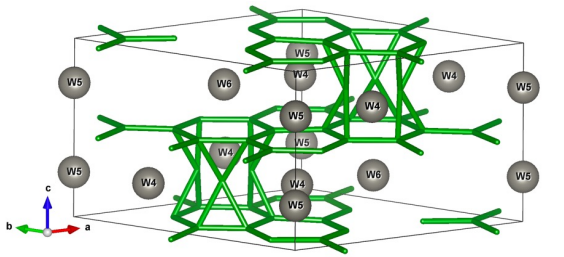


(a) For the SchNet and Allegro models with random initialization of their learnable parameters (without pre-training), the average test RMSE dependencies on the validation group used. Pre-training effects on the test scores for the (b) both-both and (c) element-both Allegro model. In all plots presented, the translucent areas correspond to one standard deviation of the test RMSE scores obtained using 4 random training/validation subsets.

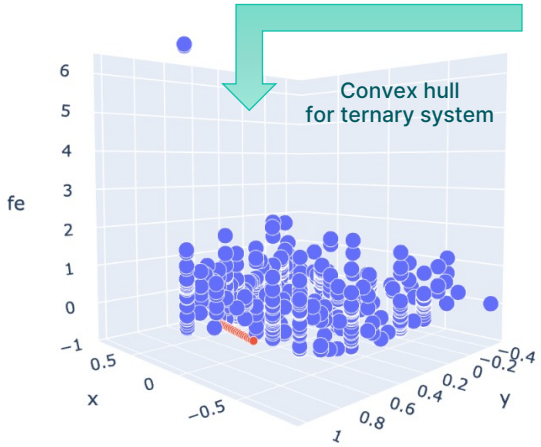
Submitted to Computational Materials Science

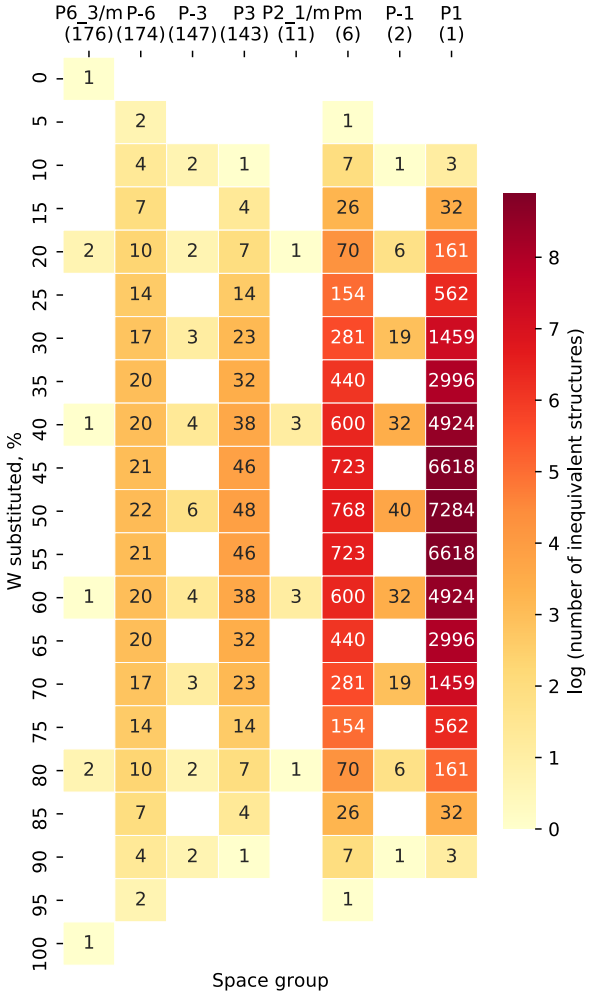
Discovering new materials

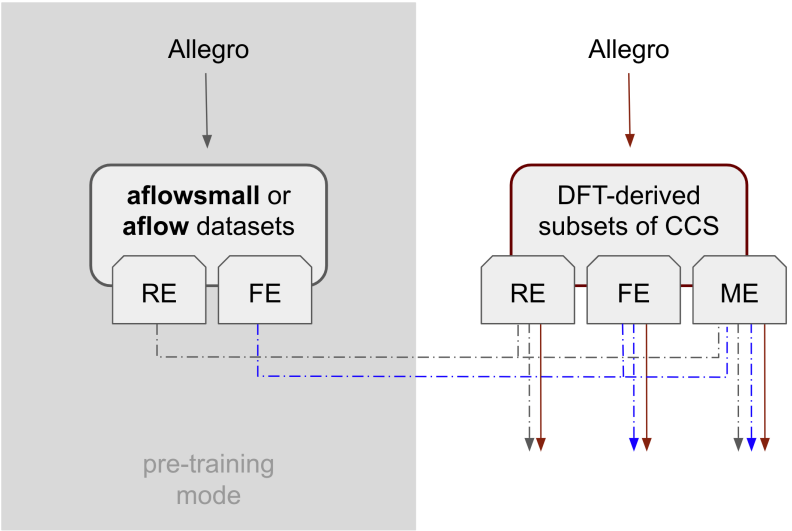
ML-based synthesizability assessments for higher borides

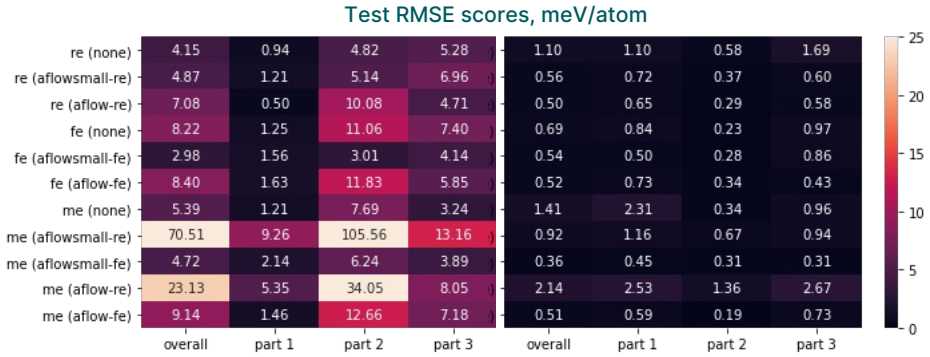


1 048 576 / **46 996**
structures









Discovering new materials

Task

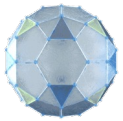
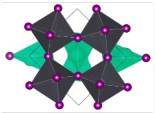
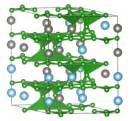
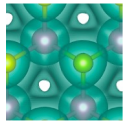
AI for predicting new materials and targeted modification of their properties

Methods

GNN (SpinConv, Graphormer, Allegro) with modification of graph properties

Optimal training and data sampling strategies

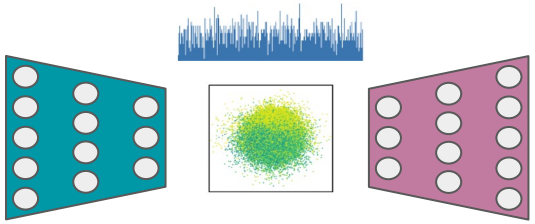
Physical modeling (DFT)

Material domain	**Quasicrystal approximants** 	**Perovskites** 	**Higher borides** 	**2D materials** 
Search space	16 000	74 000	376 000	100M+
DFT-derived samples	37% (6000)	0.3% (200)	0.07% (260)	< 0.006% (6000)

Discovering new materials

Task

Generation of crystal structures with optimization of composition/structure/properties



Methods

Generative neural network models (Variational autoencoder)

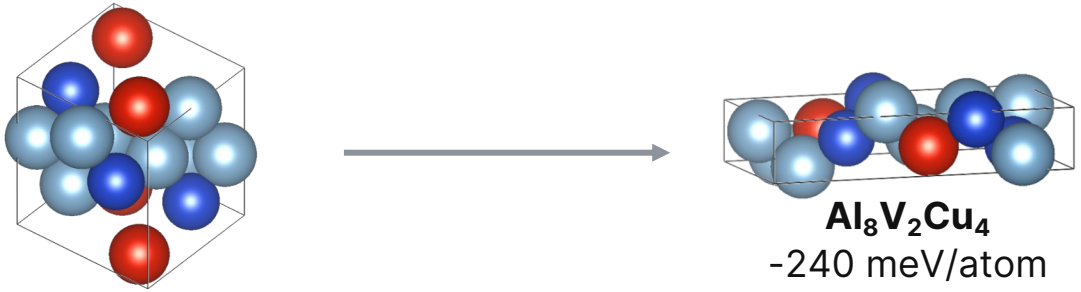
Modification of latent representation of crystal structures

Physical modeling (DFT)

Composition modification

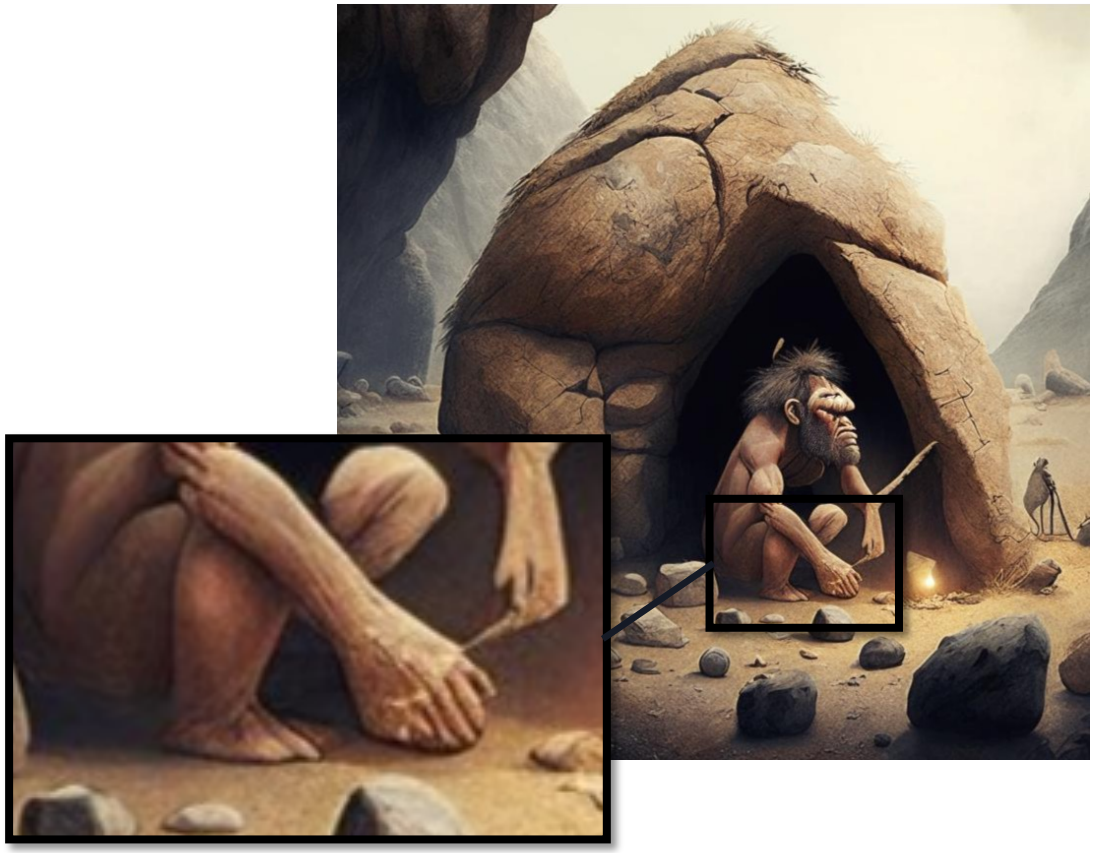
$LiCoO_2$ **$LiNiO_2$**

Structure/property modifications



$Al_8V_2Cu_4$
-240 meV/atom

Digital world *vs* Real world



1

The task of searching for new functional materials has almost unlimited complexity

2

The environmental and efficiency agenda determine the direction of the search for new materials

3

AI models are a universal approach allowing to predict properties of materials

New Materials Design Group @ AIRI

Semen Budenniy

PhD in Physics and Mathematics
Scientific Advisor

Aleksei Korovin

PhD in Chemistry
Senior Research Scientist

Roman Eremin

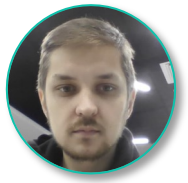
PhD in Physics and Mathematics
Senior Research Scientist

Innokentii Humonen

Junior Research Scientist

Mikhail Tiutiulnikov

Trainee Researcher



Artem Vasilev

Junior Research Scientist

Vladimir Lazarev

Junior Research Scientist

Alexey Boyko

Research scientist

Aliaksei Krautsou

Junior Research Scientist

Artificial Intelligence Research Institute

airi.net

airi_research_institute

AIRI Institute

AIRI Institute

AIRI_inst

artificial-intelligence-research-institute

Thx!