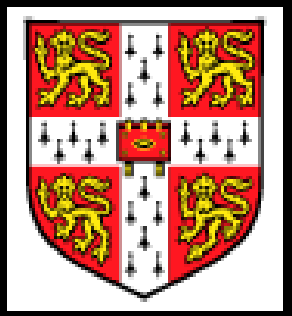
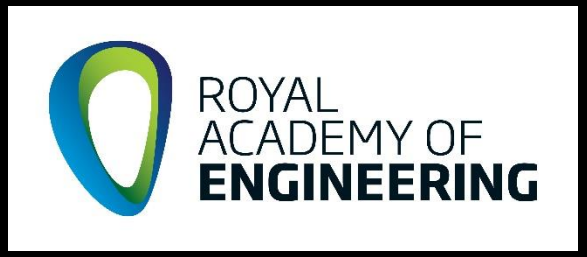
Next-generation Materials-Data Curation and Processing Methods for Machine-Learning Applications

Jacqueline M. Cole

Cavendish Laboratory, University of Cambridge, UK

ISIS Neutron and Muon Source, Rutherford Appleton Laboratory, UK

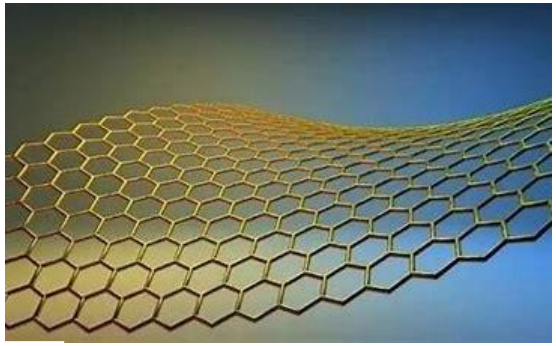




The Evolution of Structural Science

Single Compound

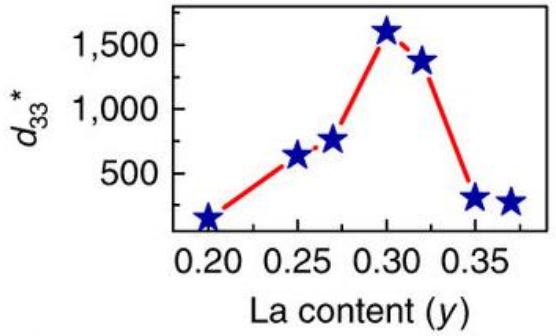
STRUCTURE; PROPERTY

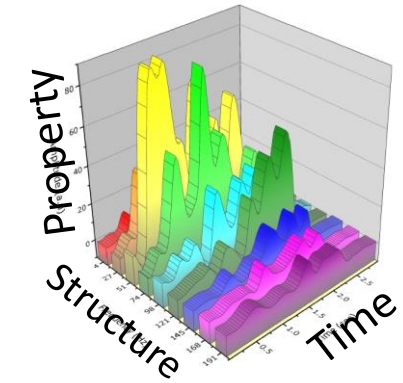


Series of Compounds

STRUCTURE versus PROPERTY

versus TIME

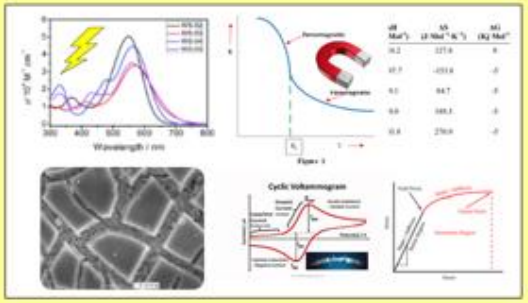




Systems Approach

STRUCTURE & PROPERTY SPACE

(& versus TIME)

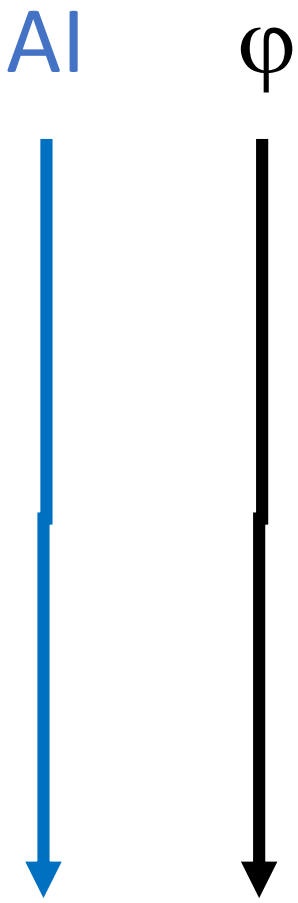


Data Sources

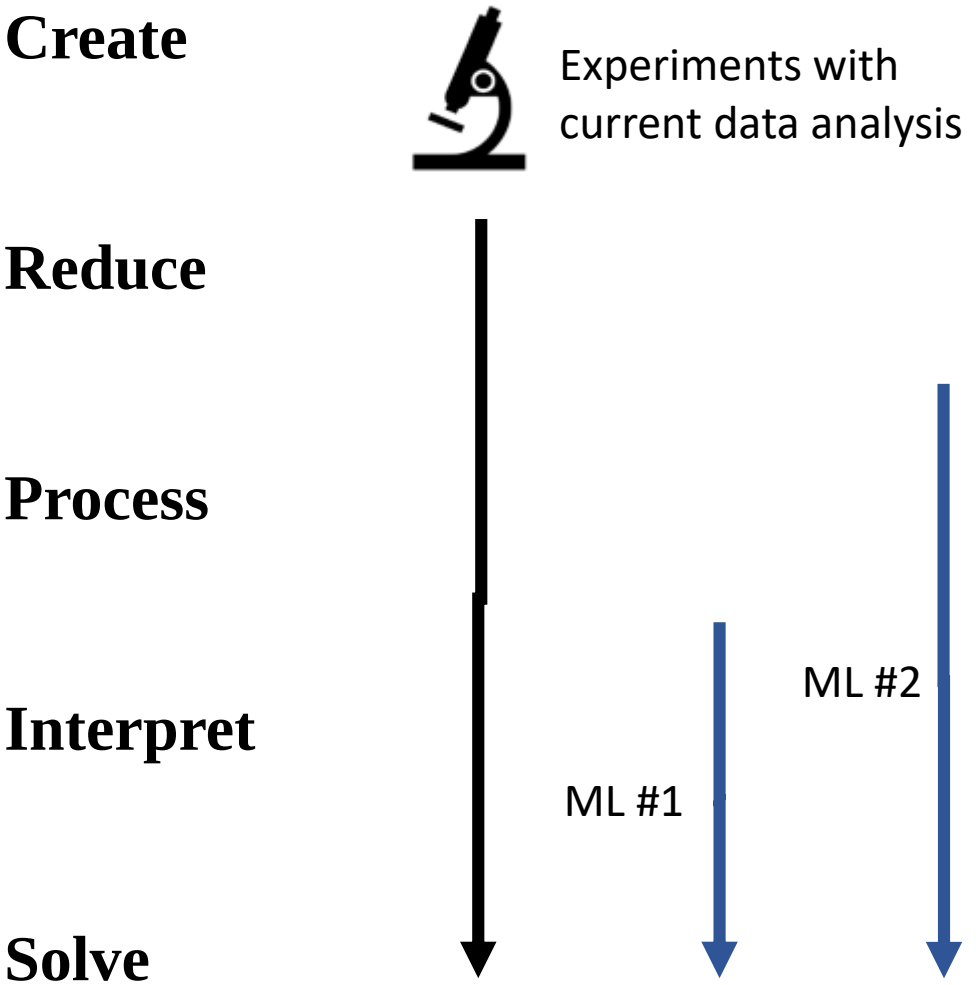
Chemical Space

Material Properties

Pipelines for Data-driven Science



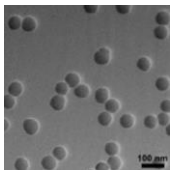
But needs experiments for validation (and training for AI)



Input: Processed Data

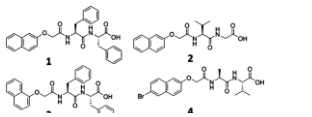
A Design-to-Device Approach

- Create software tools for auto-generating materials databases

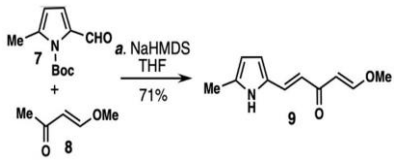


ImageDataExtractor

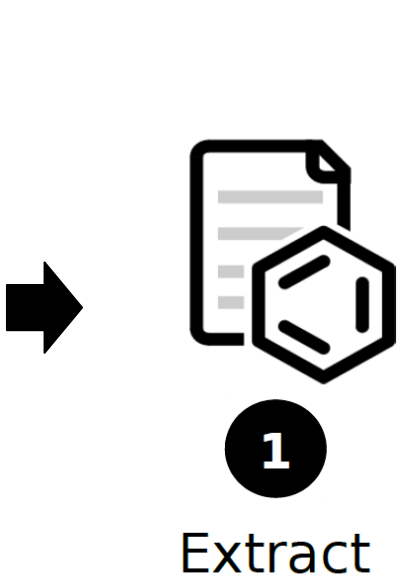
ChemDataExtractor



ChemSchematicResolver

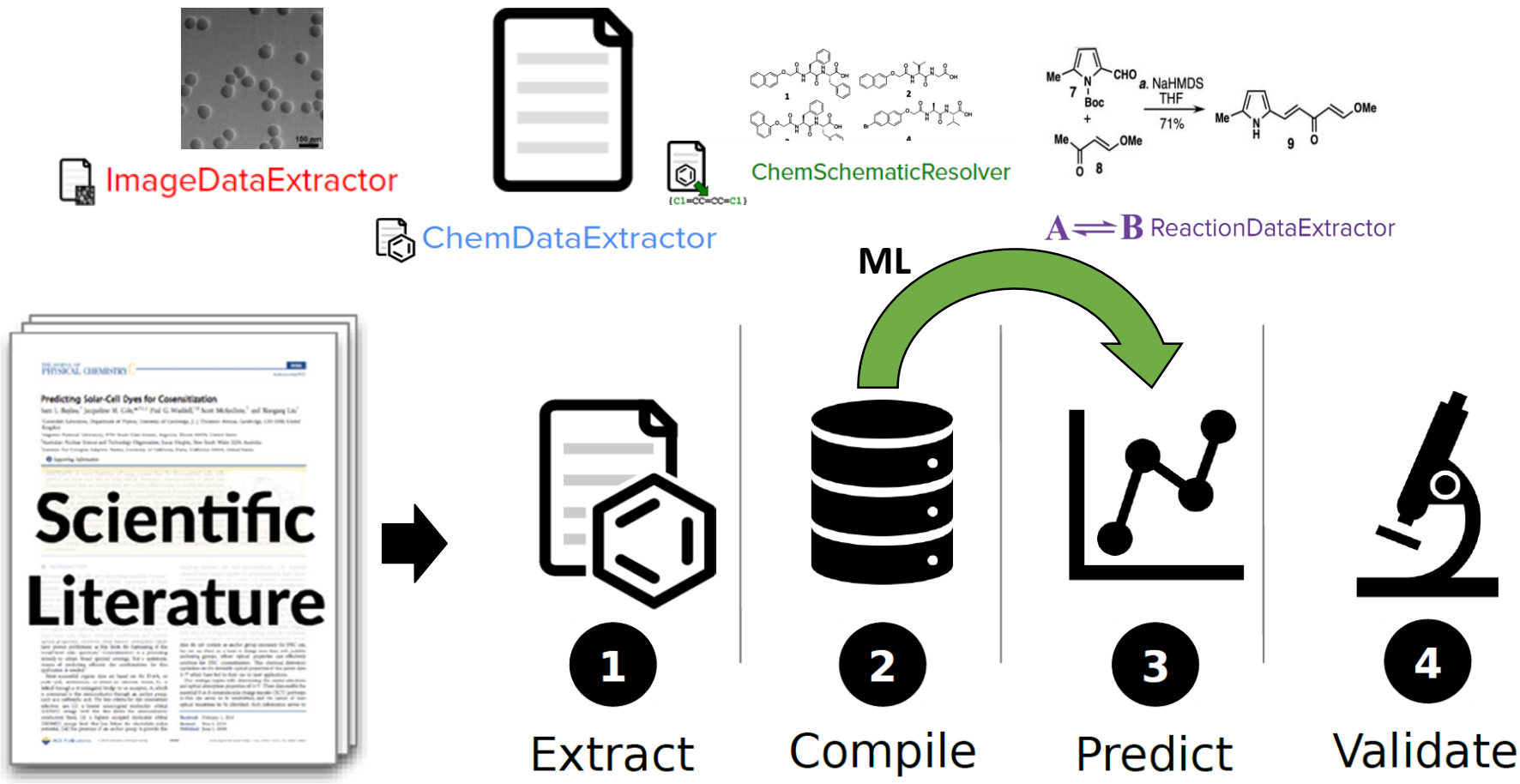


A⇌B ReactionDataExtractor



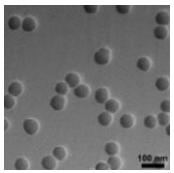
A Design-to-Device Approach

- Train machine-learning algorithms with the data to predict materials



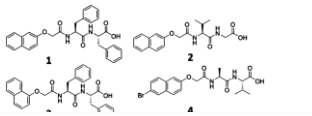
A Design-to-Device Approach

- Realise data-driven materials discovery to aid the energy sector

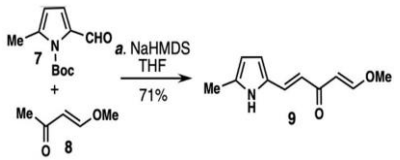


ImageDataExtractor

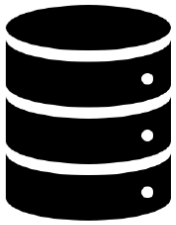
ChemDataExtractor

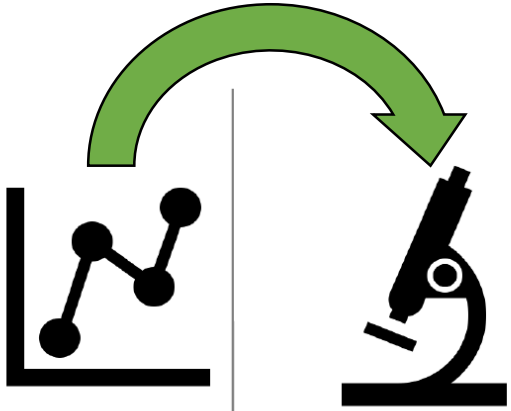


ChemSchematicResolver



A⇌B ReactionDataExtractor





1 Extract

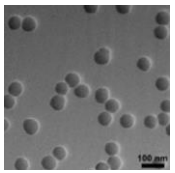
2 Compile

3 Predict

4 Validate

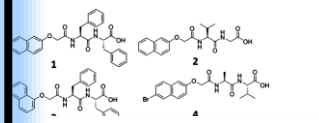
A Design-to-Device Approach

- Realise data-driven materials discovery to aid the energy sector

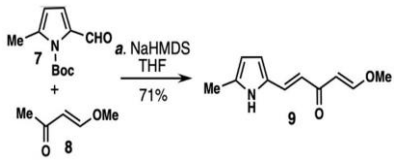


ImageDataExtractor

ChemDataExtractor



ChemSchematicResolver



A⇌B ReactionDataExtractor

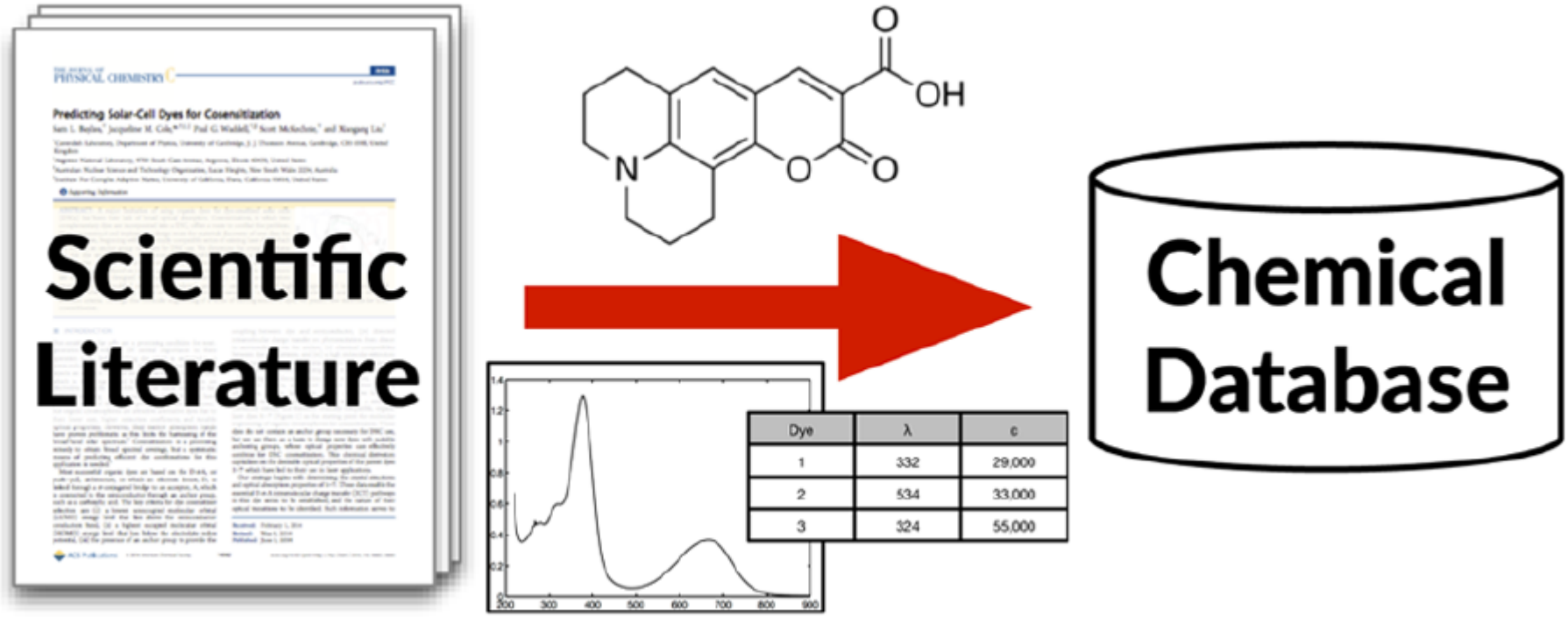
1 Extract

2 Compile

3 Predict

4 Validate

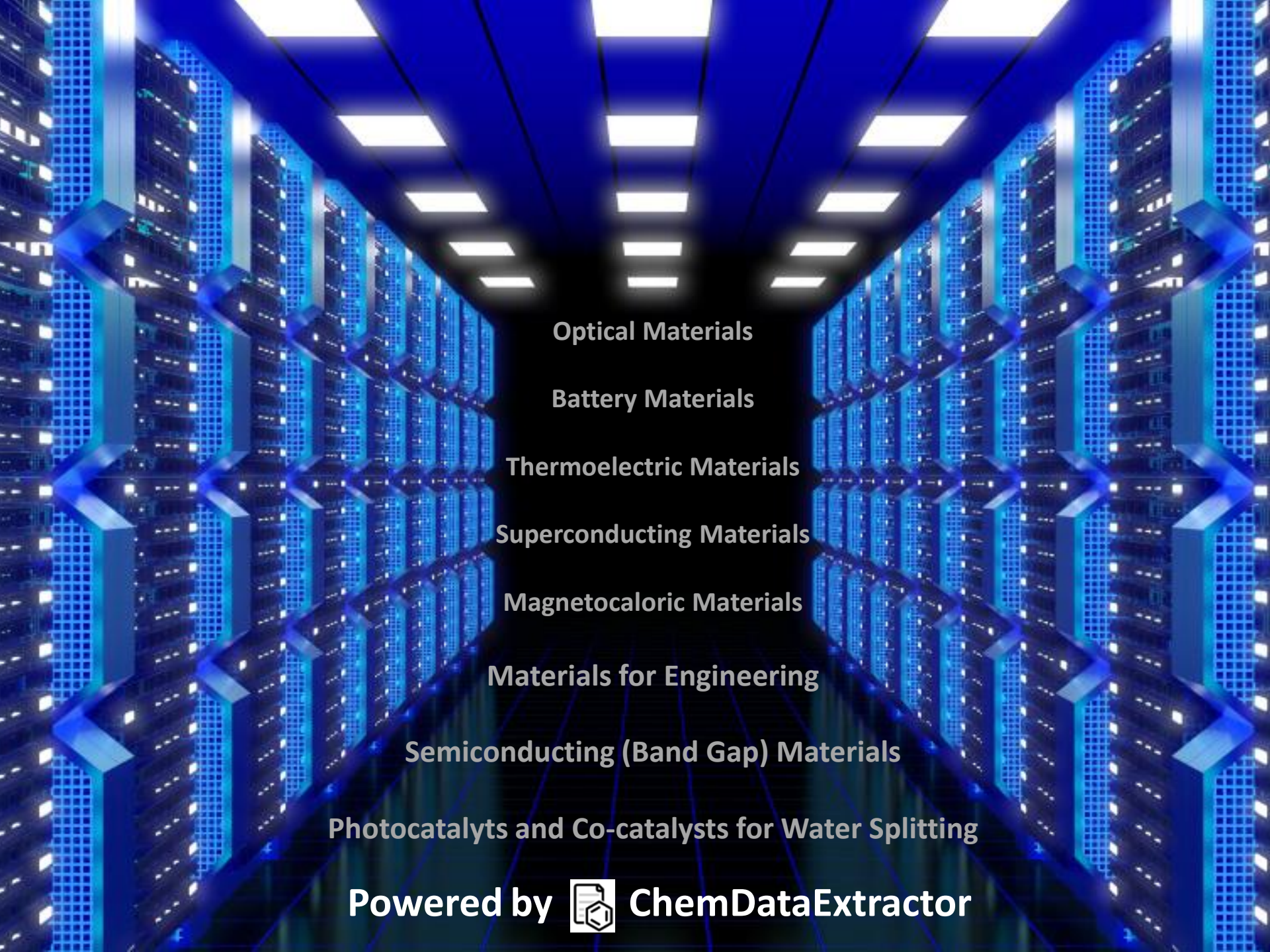
ChemDataExtractor



Input **Output**

www.chemdataextractor2.org

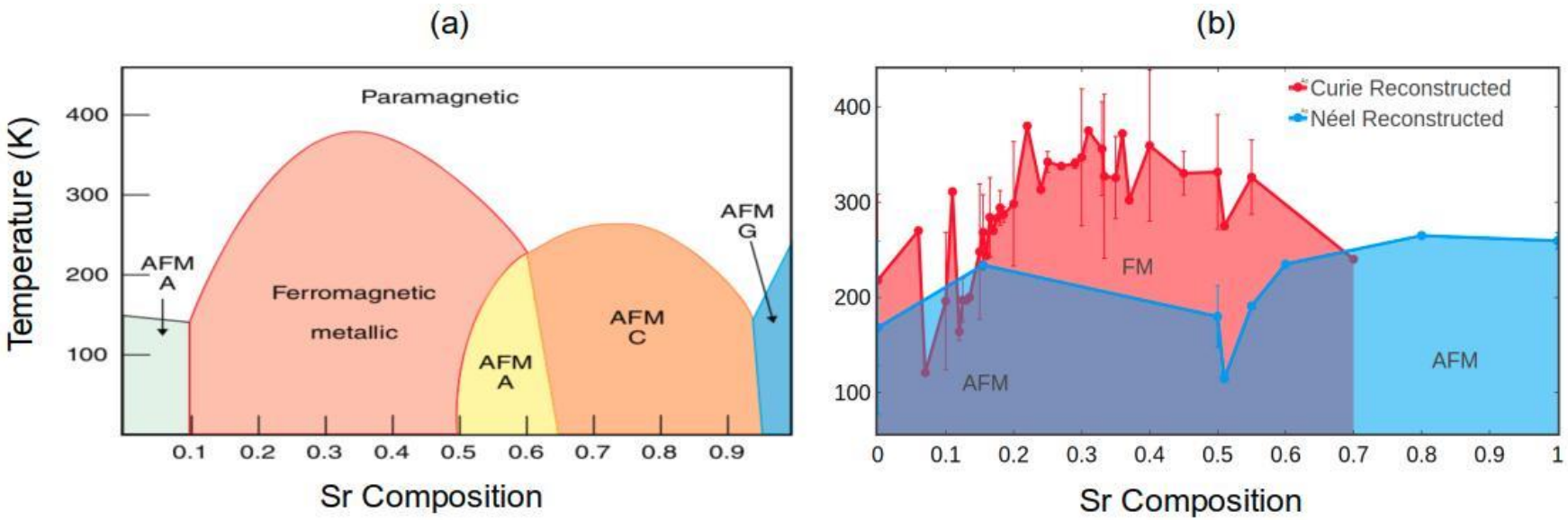
Applications


Optical Materials
Battery Materials
Thermoelectric Materials
Superconducting Materials
Magnetocaloric Materials
Materials for Engineering
Semiconducting (Band Gap) Materials
Photocatalyts and Co-catalysts for Water Splitting
Powered by ChemDataExtractor

Magnetism

Reconstruct Phase Diagrams

$La_{1-x}Sr_xMnO_3$ series



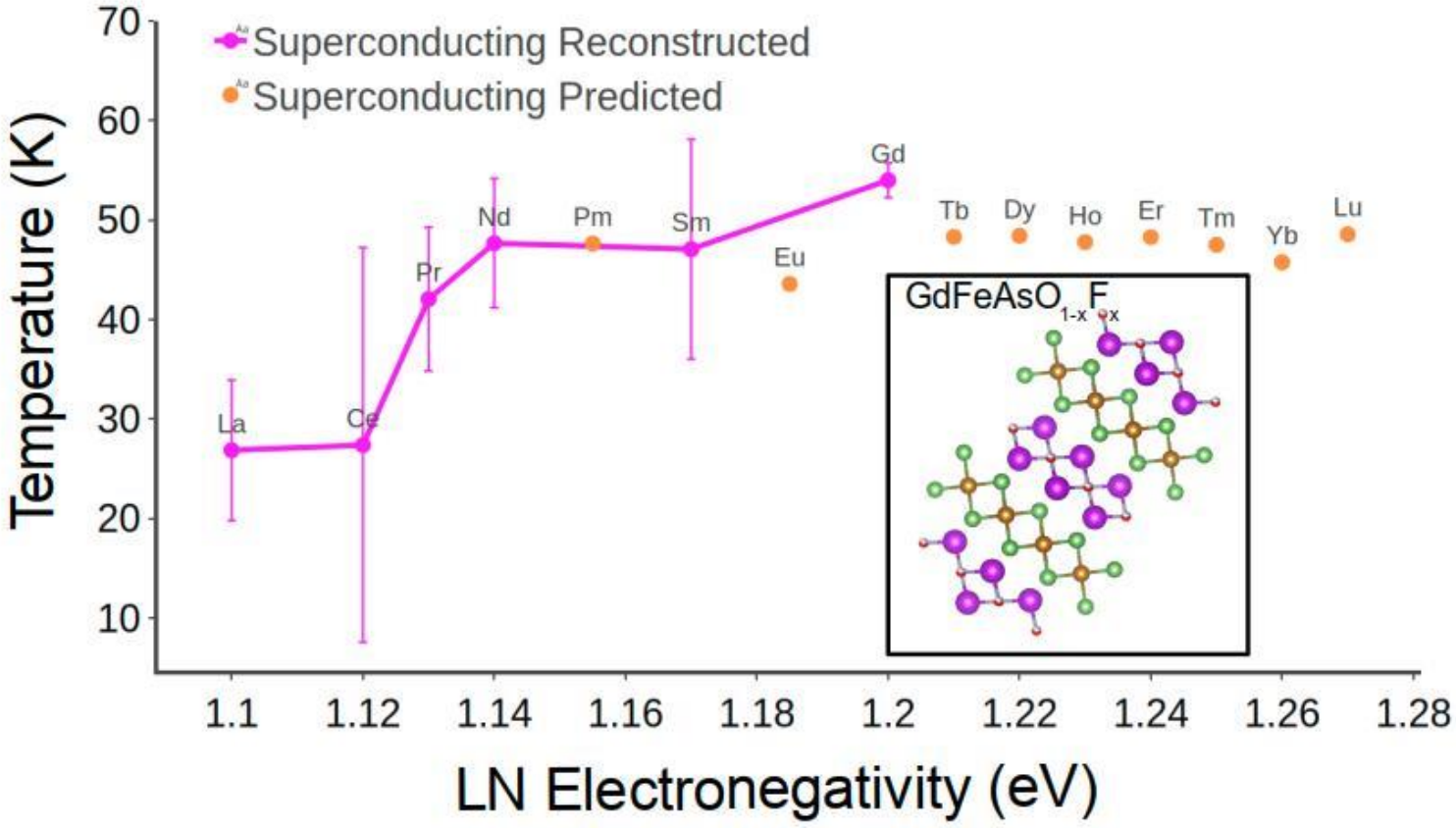
Holistic 'cartoon' that has built up over the years

Reconstructed using

ChemDataExtractor

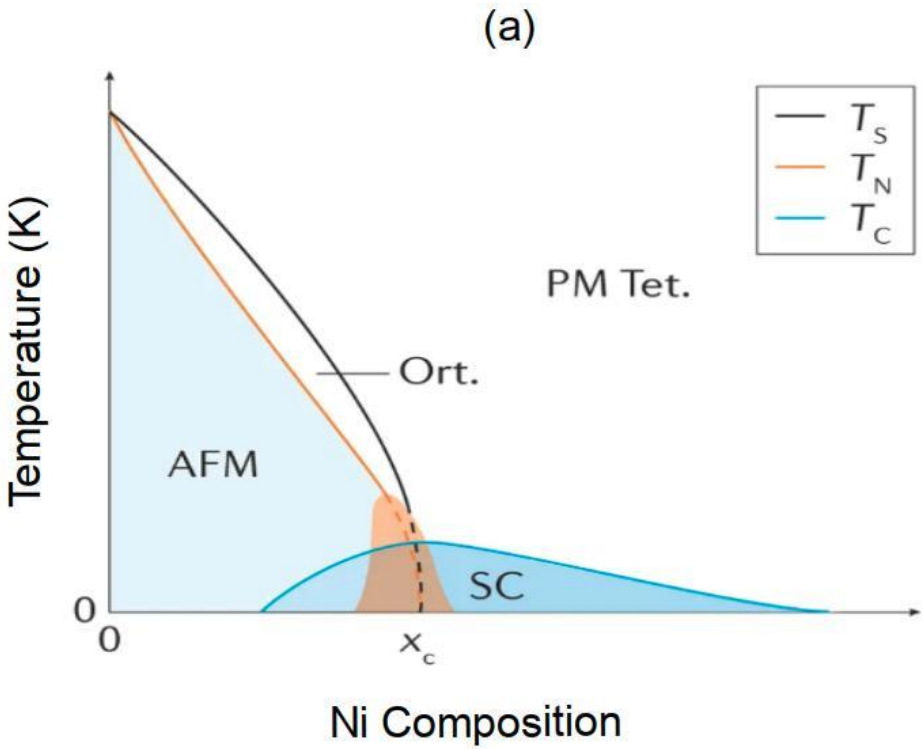
Superconductivity

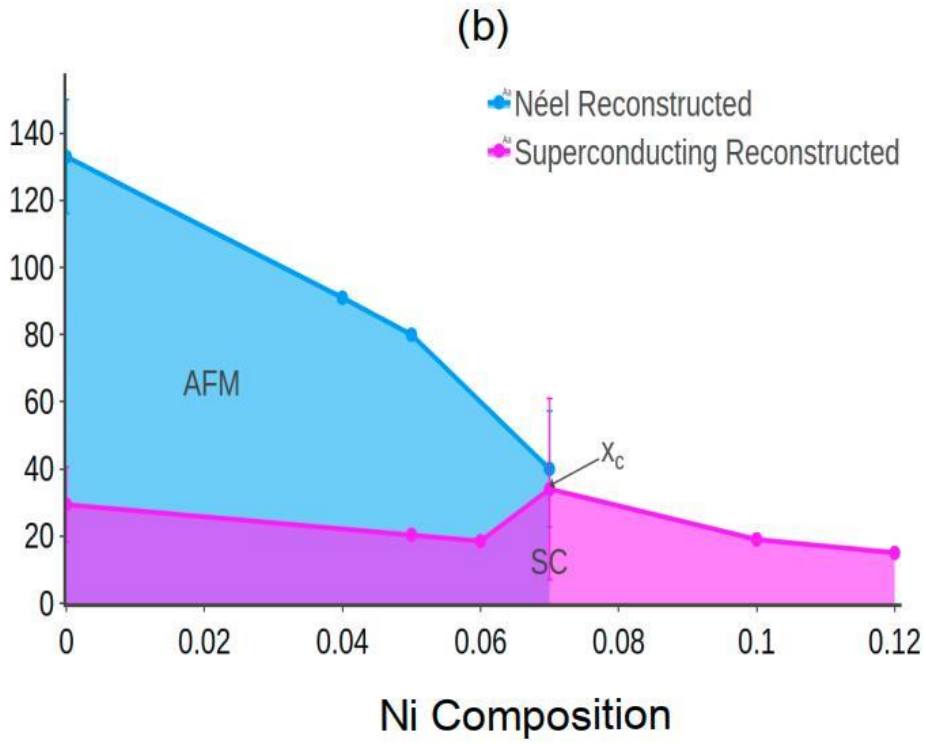
ML predictions of Tc: $LnFeAsO_{1-x}F_x$



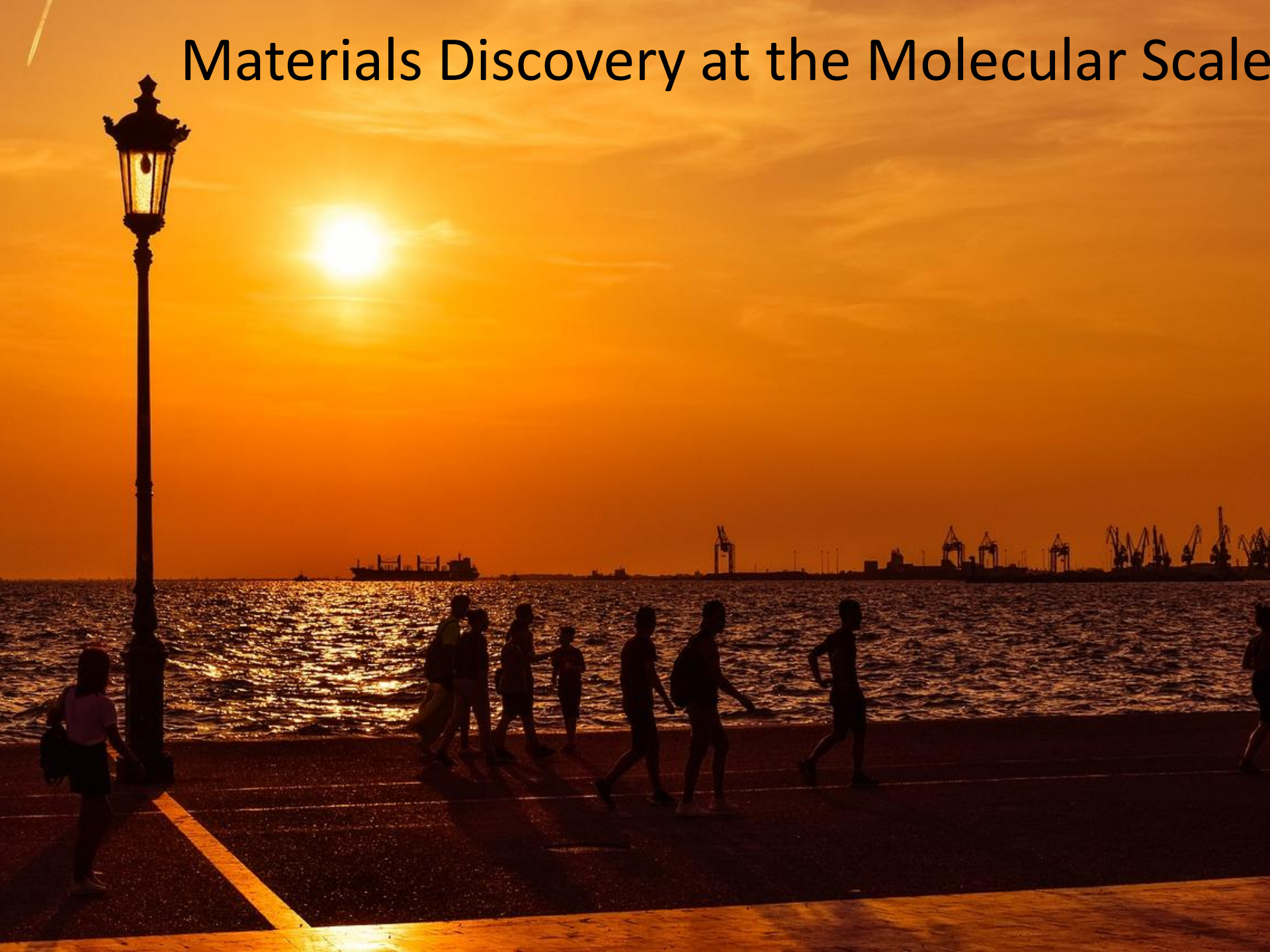
ML prediction: Random forest regression with K-best selection (shows Tc is dependent on ionic radii, atomic number, work function of Ln)

Phase diagram of $BaFe_{2-x}Ni_xAs_2$





Materials Discovery at the Molecular Scale



A Database from ChemDataExtractor

{chemical, λmax, ε}

9,431 dye candidates

B Remove small molecules, organometallic dyes, and chemicals not absorbing in the solar spectrum

3,053 dye candidates

C Select dyes with COOH anchors and molecular dipole moment > 5 Debye

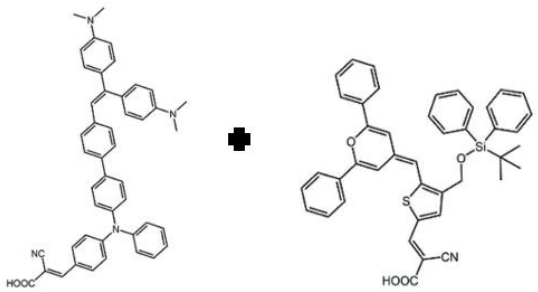
309 dye candidates

D Identify combinations of dyes with complementary optical absorption

33 dye candidates

E Check HOMO/LUMO Energy Levels

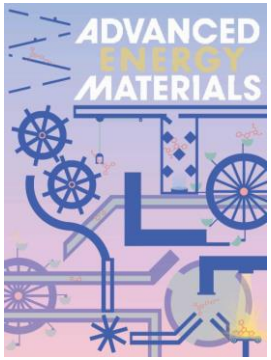
29 dye candidates



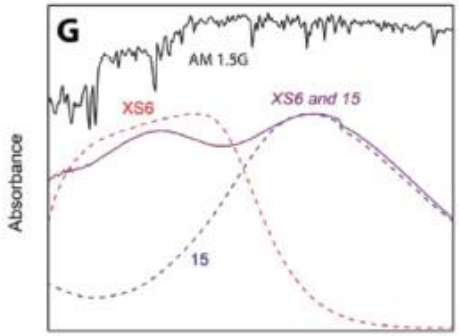
F Final Selection (5 dyes)

Power Conversion Efficiency = 92% Industry Standard (N719)

Experimental Validation



Adv. Energy Mater. 2019, 9, 1802820



Extract Data

Enrich data

Predict

Validate

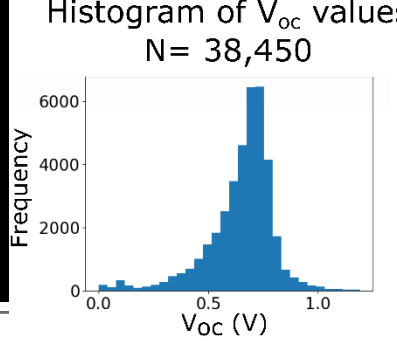
Photovoltaics Device Databases Auto-generated via

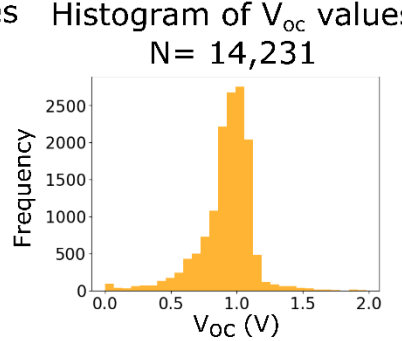
Manufacturing applications

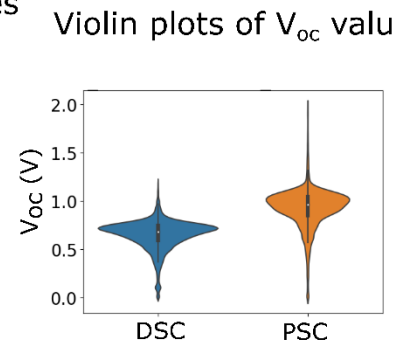
Device data:

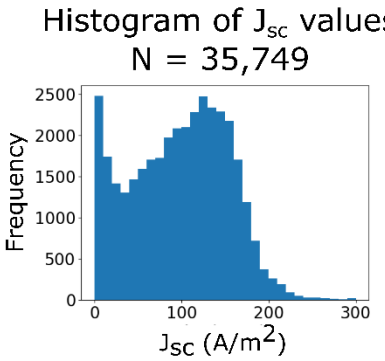
Dye-sensitized Solar Cells (left)

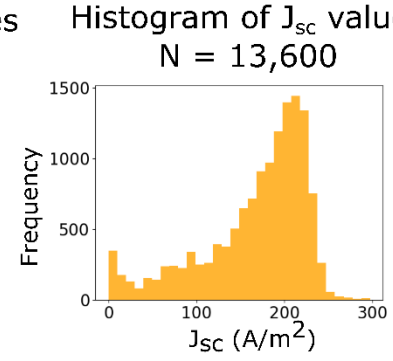
Perovskite Solar Cells (middle)

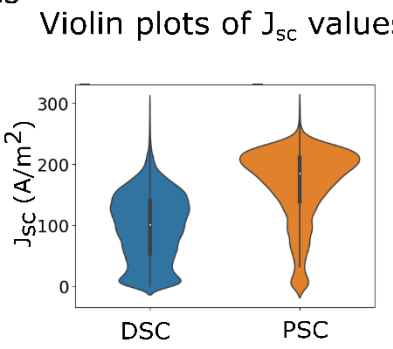


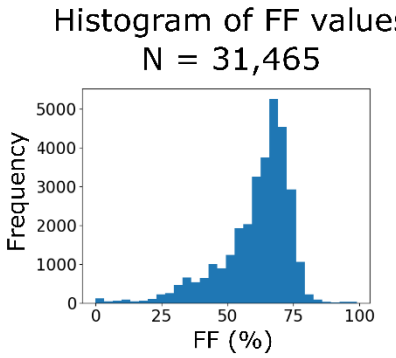


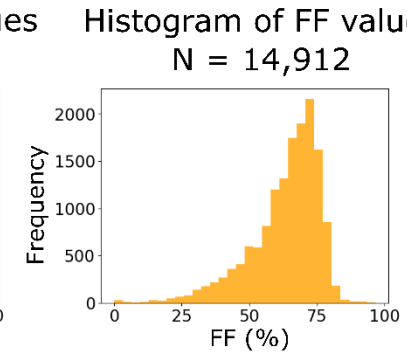


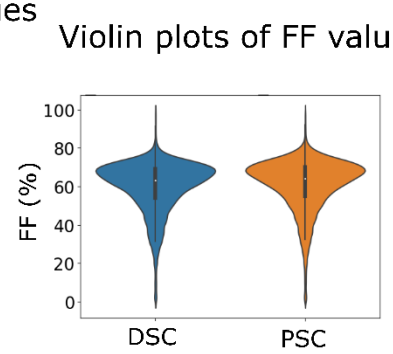


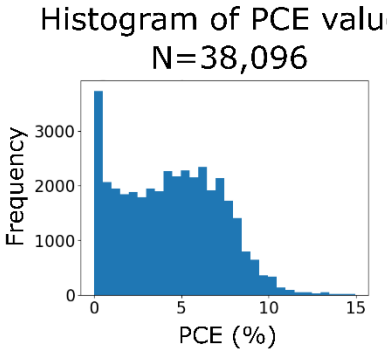


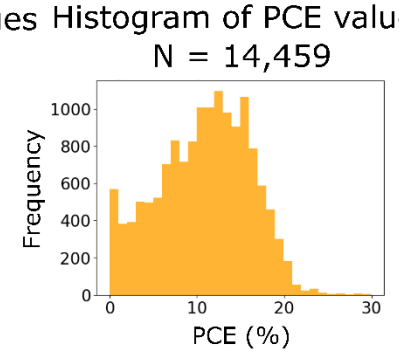


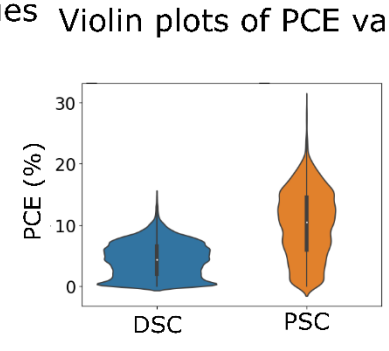












E. J. Beard, J. M. Cole, *Scientific Data* 9, 329 (2022).

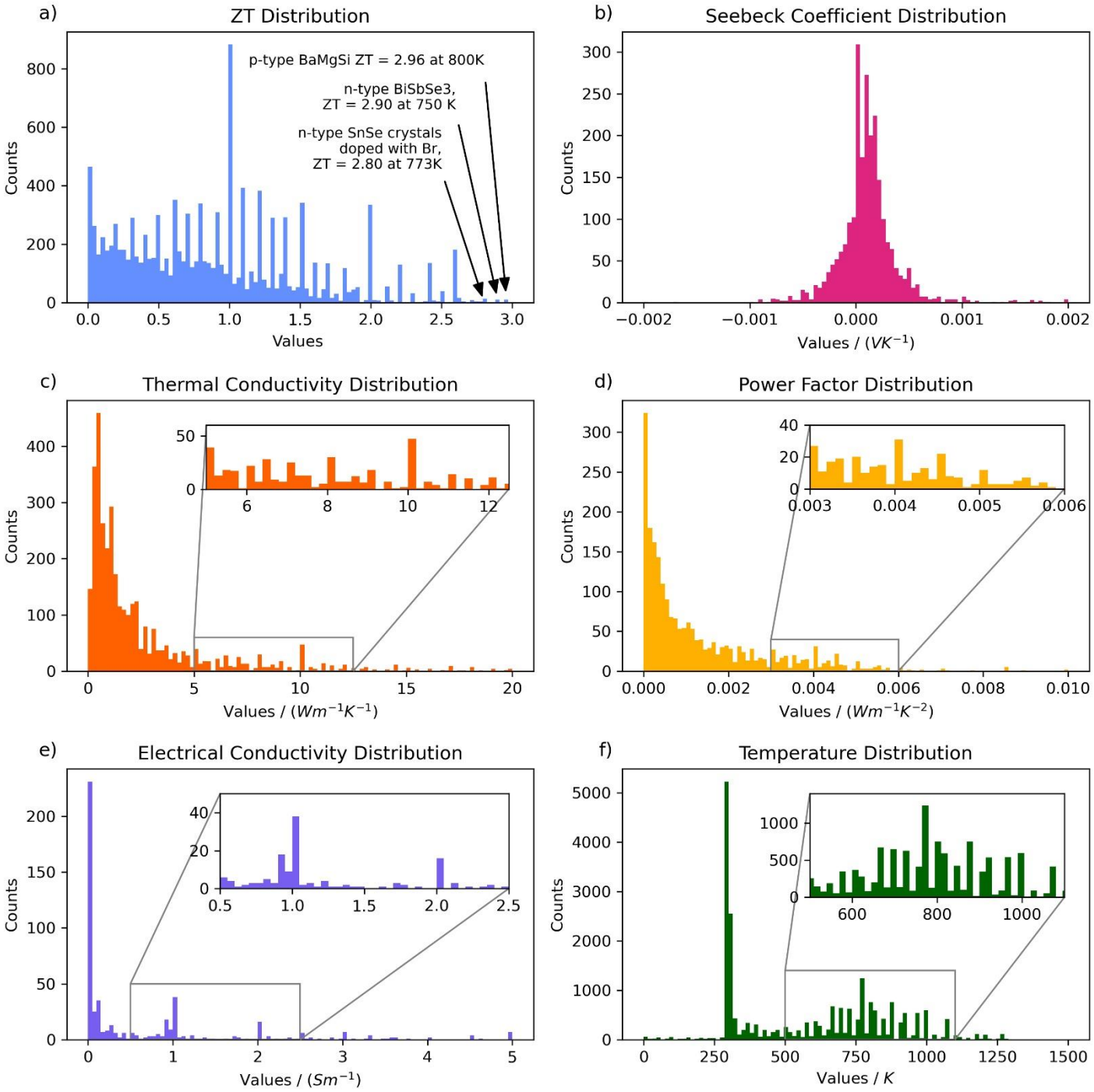
Device metrological attributes

Dye Sensitized Solar Cell Database	Perovskite Solar Cell Database
Solar simulator (irradiance)	Solar simulator (irradiance)
Substrate	Substrate
Active area	Active area
Semiconductor	Counter electrode
Semiconductor thickness	
Dye loading	
Redox couple	

E. J. Beard,
J. M. Cole,
Scientific Data
9, 329 (2022).

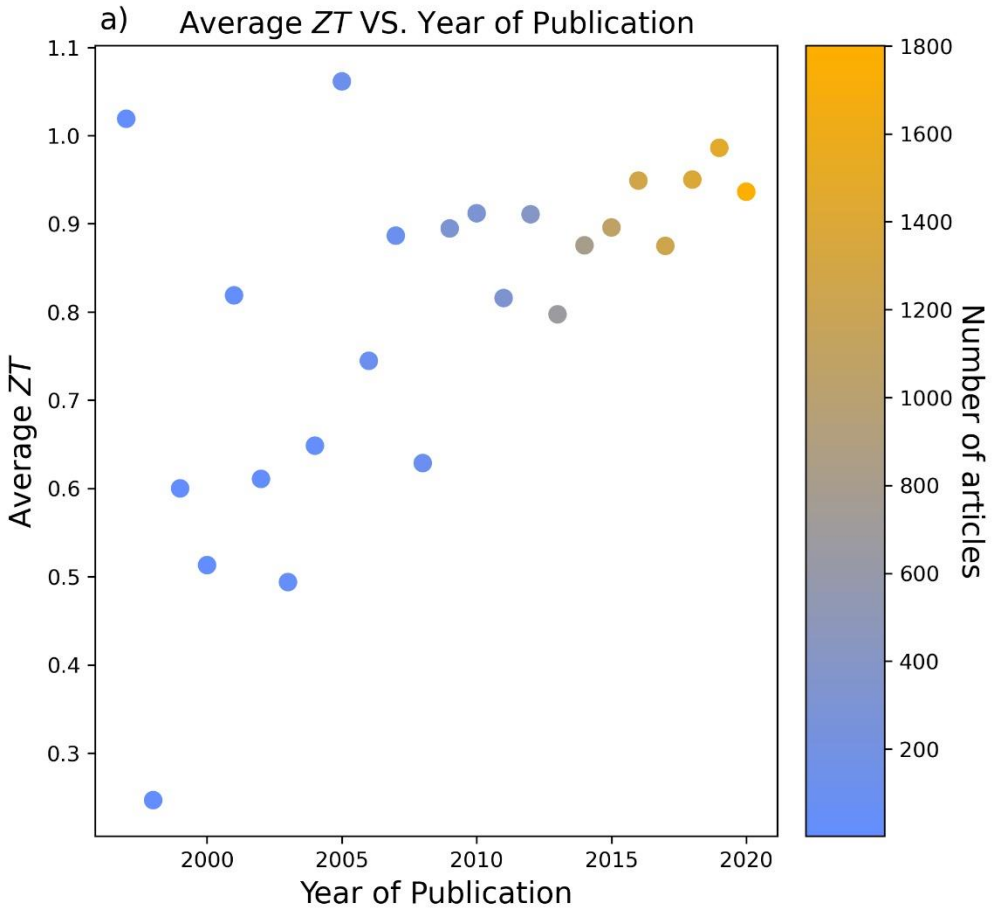
Thermoelectrics

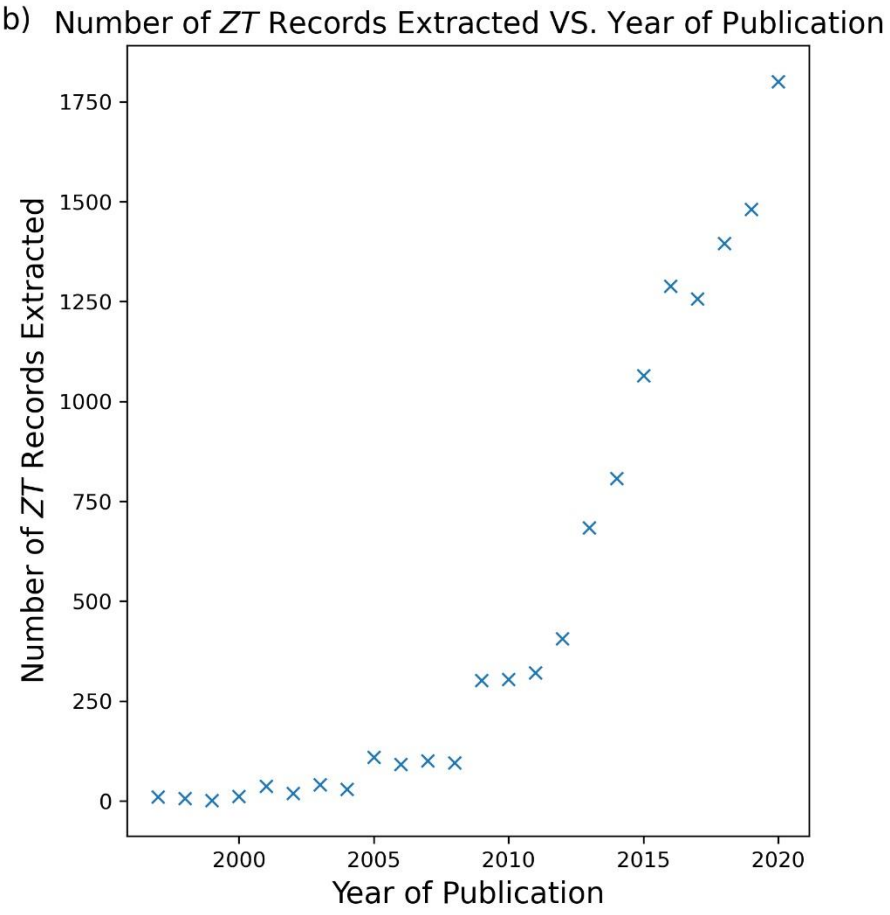
$$ZT = \frac{S^2\sigma}{\kappa}T$$



O. Sierepeklis, J. M. Cole, *Scientific Data* 9, 648 (2022).

Trend-setting data ....

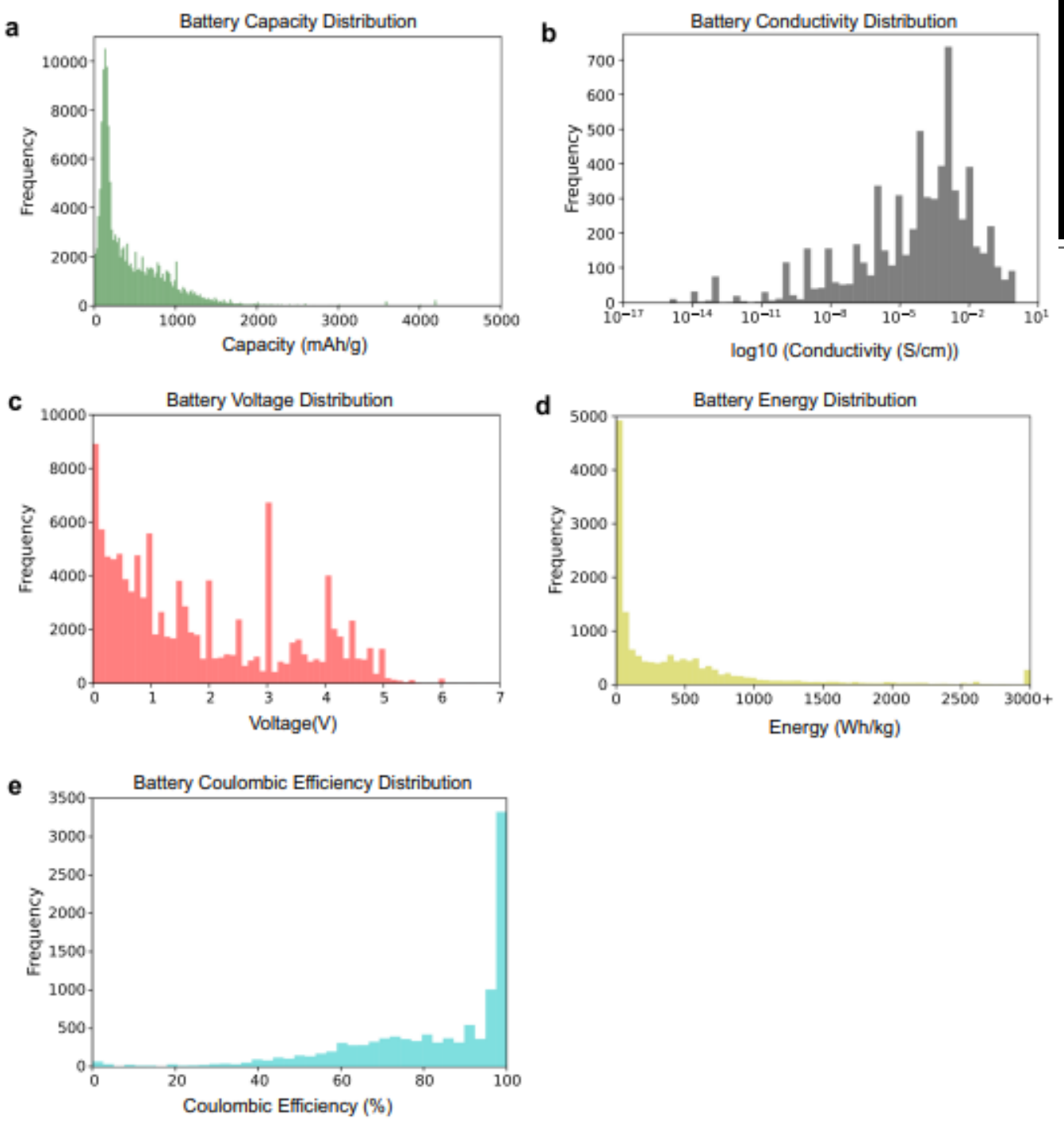




Battery Device Data

Huang and Cole, *Chemical Science* 13 (2022) 11487-11495.

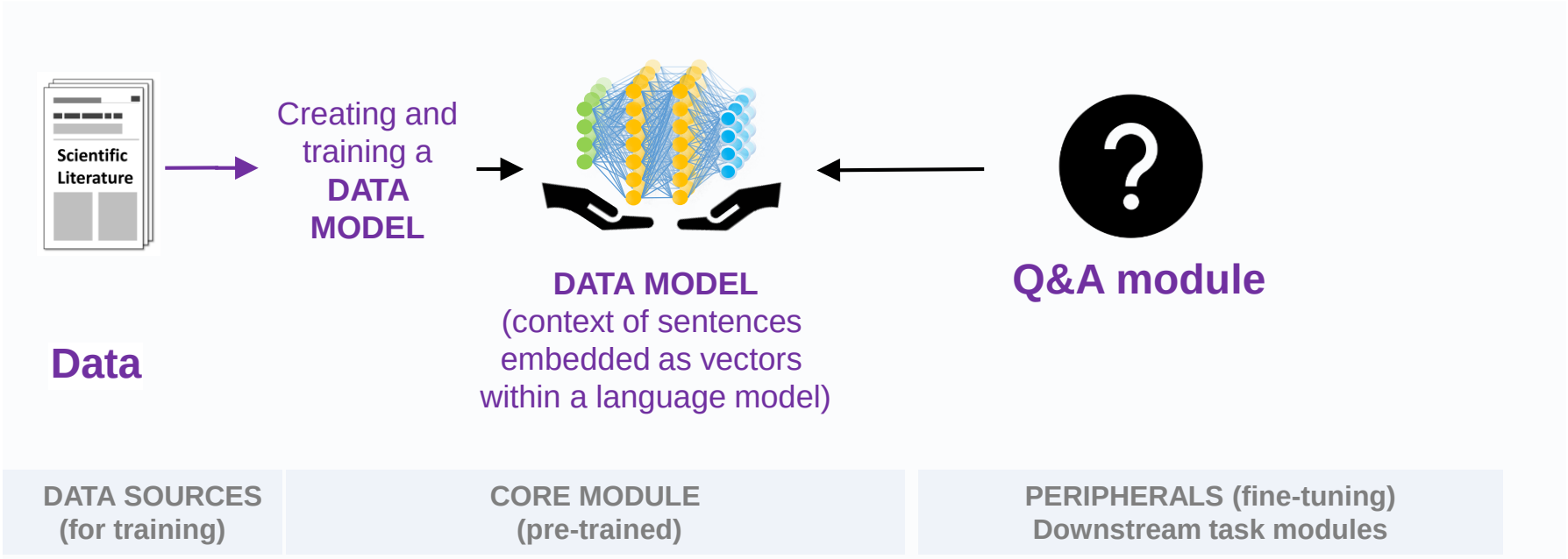
Image Credit: Shu Huang and Nan Tian, and embedded image by rawpixel.com on Freepik



292k data records in database

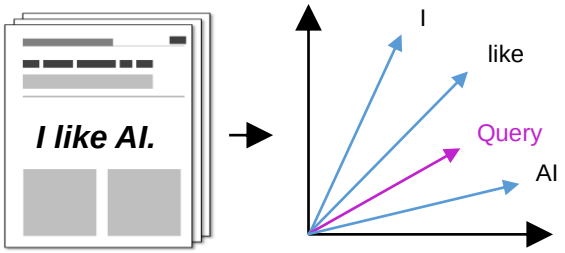
S. Huang, J. M. Cole, *Scientific Data* (2020) 7, 260.

Distinguishing device materials as anode, cathode, or electrolyte



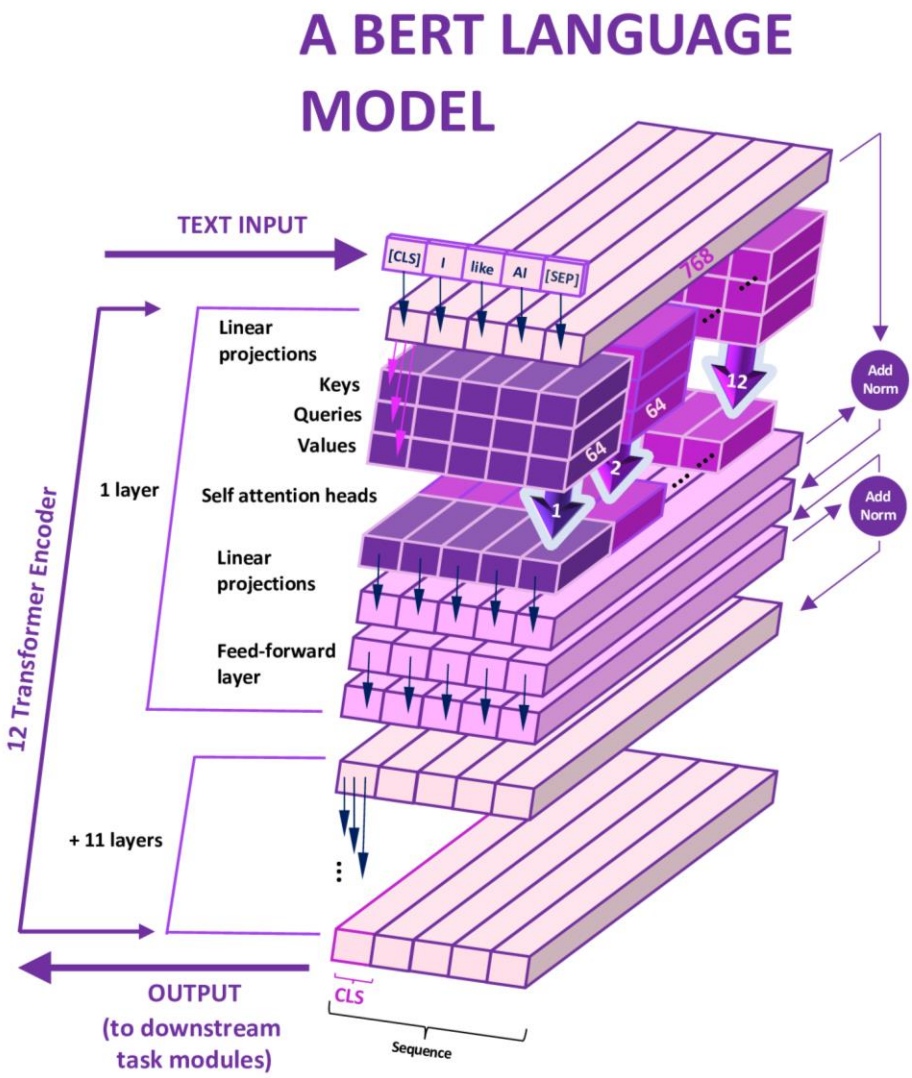
S. Huang, J. M. Cole, *J. Chem. Inf. Model*. 62 (2022) 6365-6377

A BERT Language Model



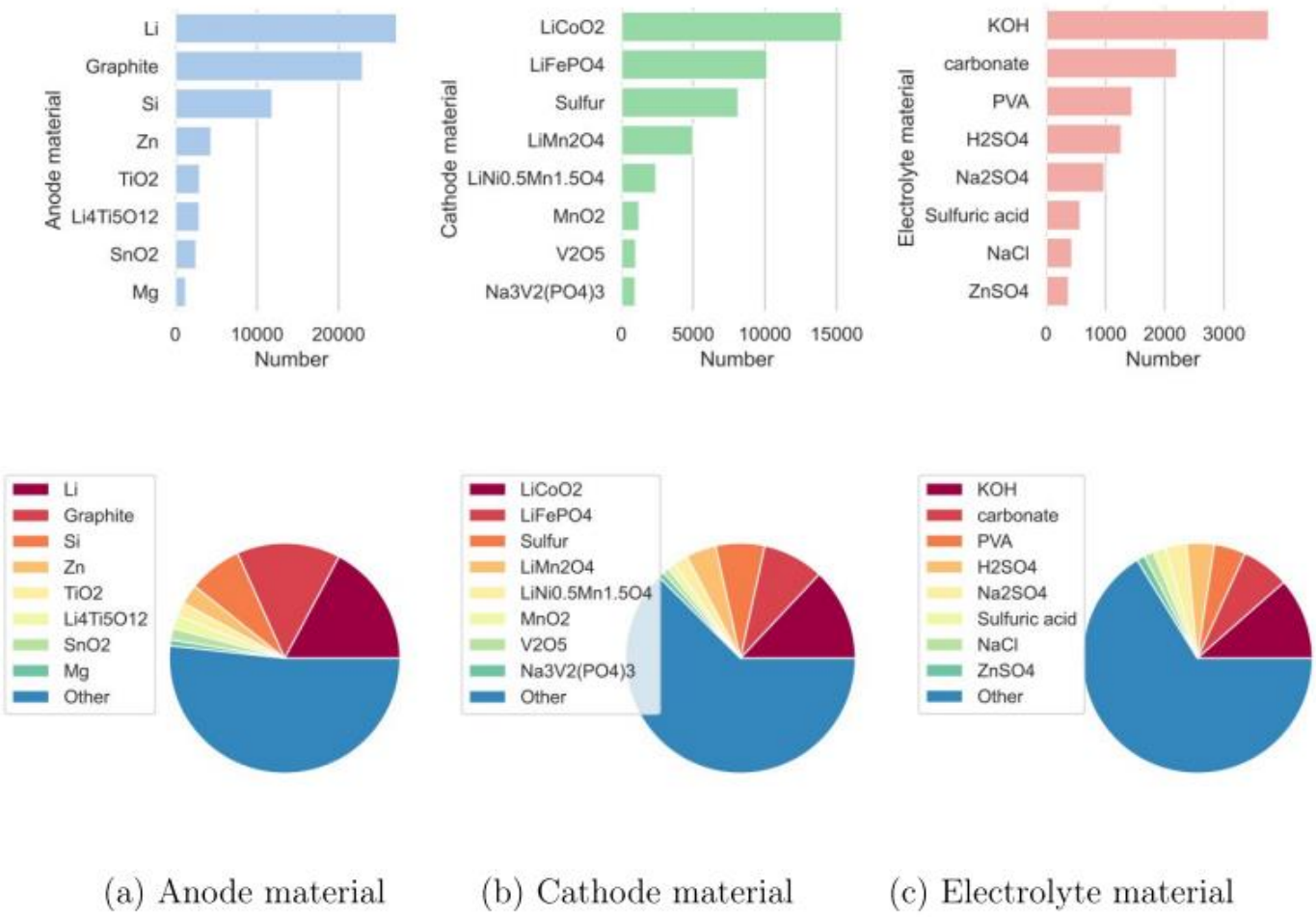
Input sentence

Embedding sentence Into vectorial representation



Distinguishing device materials as anode, cathode, or electrolyte

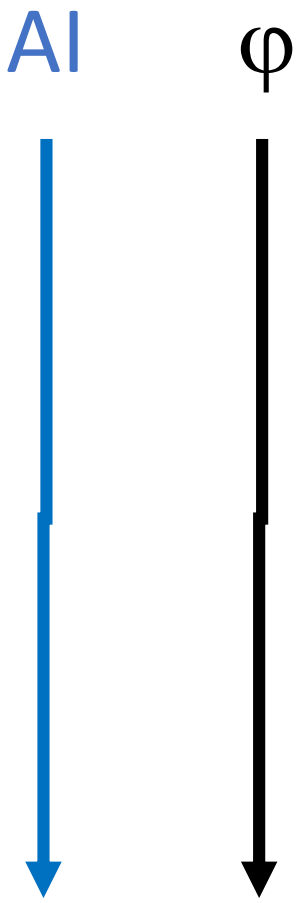
309k data records in database



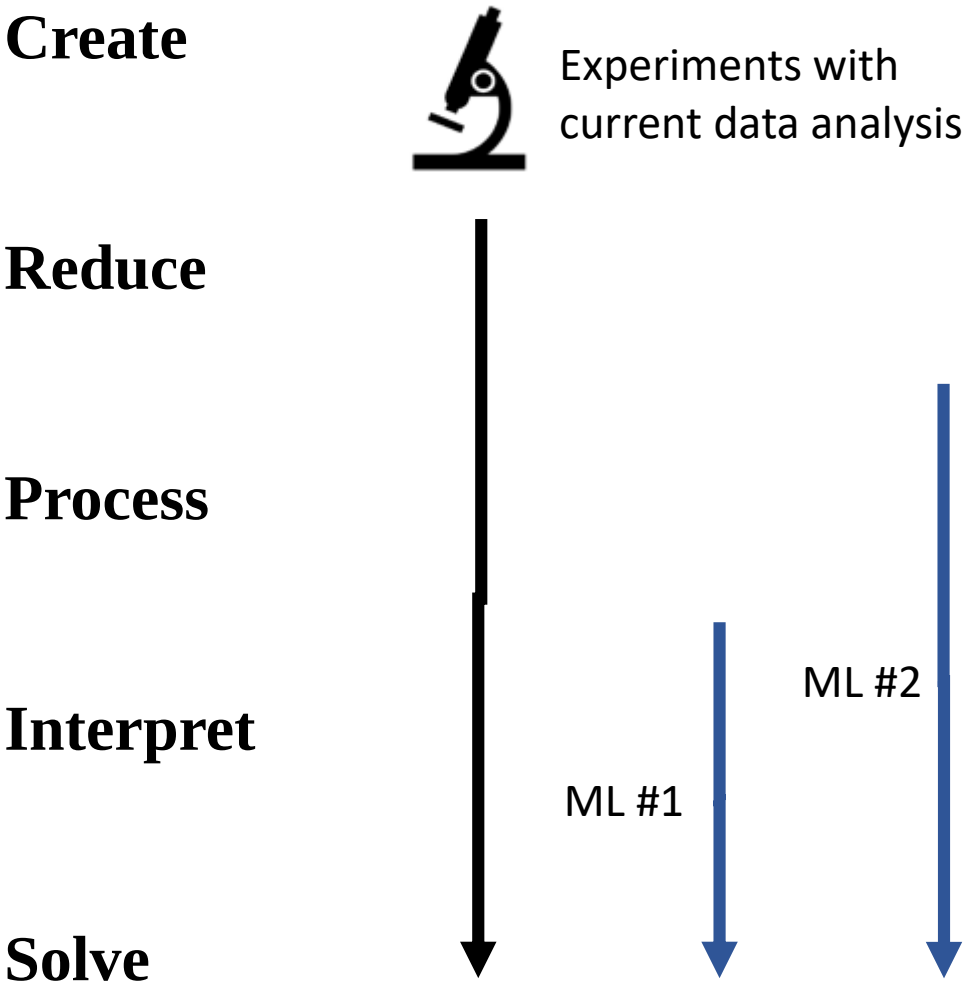
(a) Anode material (b) Cathode material (c) Electrolyte material

S. Huang, J. M. Cole, *J. Chem. Inf. Model*. 62 (2022) 6365-6377

Pipelines for Data-driven Science

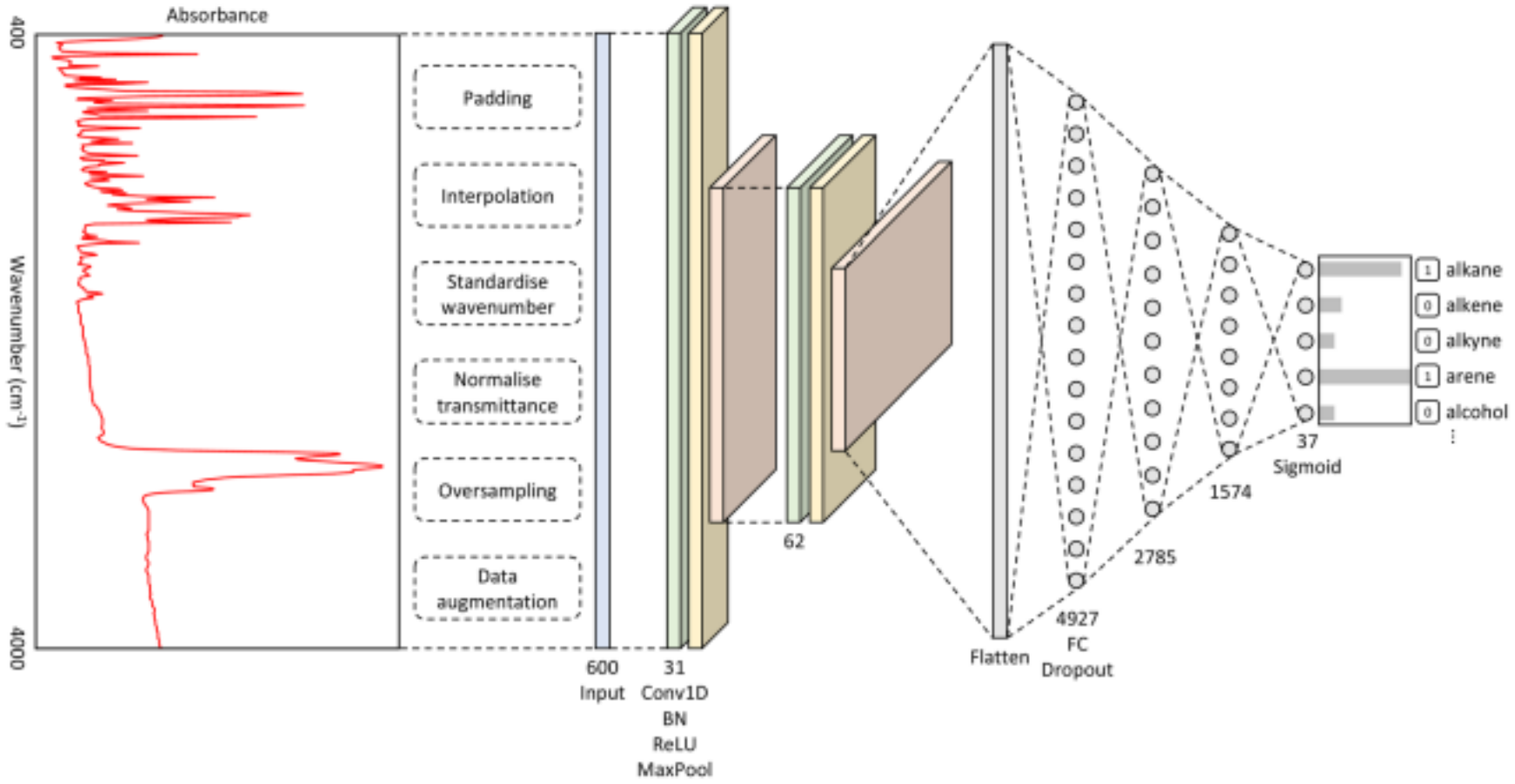


But needs experiments for validation (and training for AI)



Input: Reduced Data

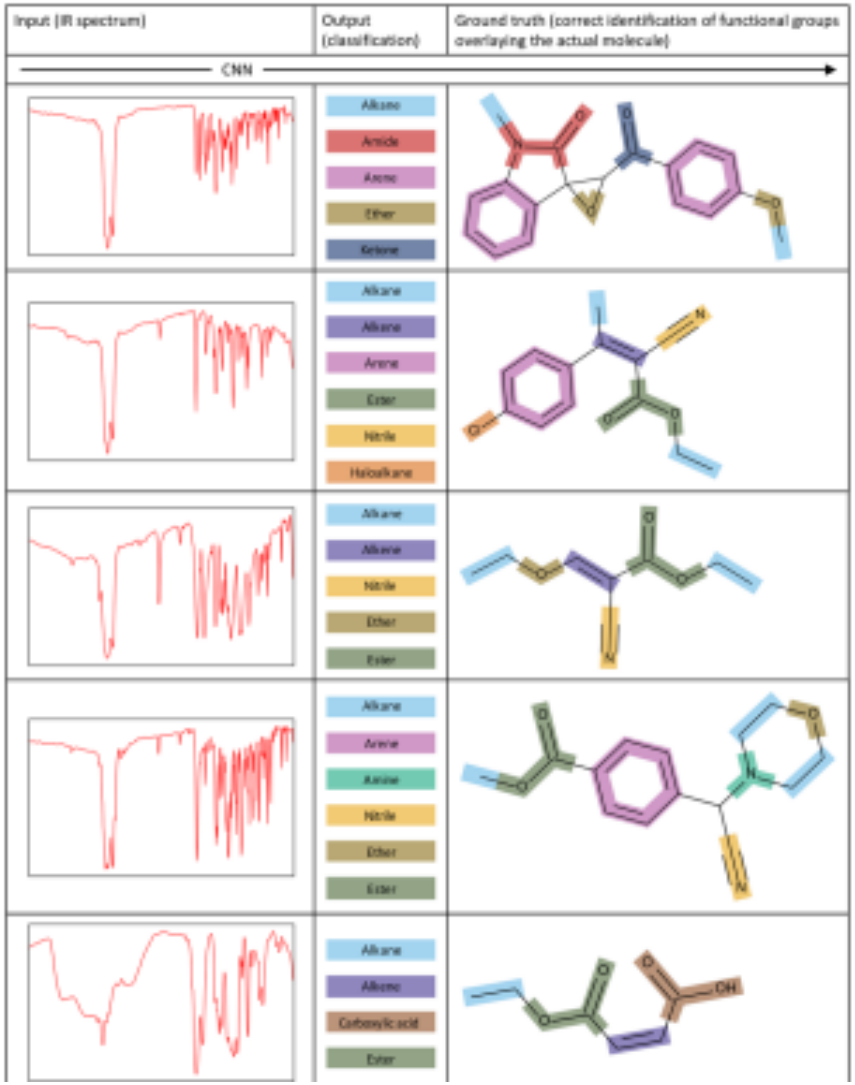
Spectral Image → Molecular Structure

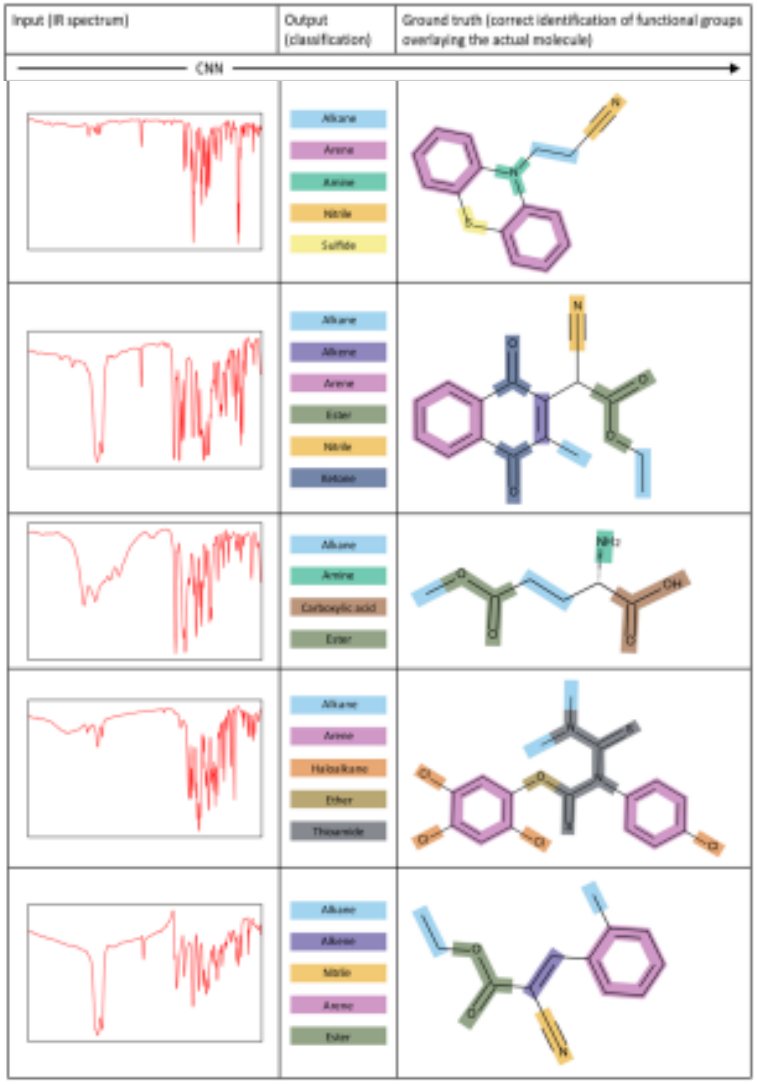


Collaboration with SCD and ISIS at Rutherford Appleton Laboratory

Jung, Jung, Cole, *Chem. Sci*. (2023) **4**, 3600–3609

Spectral Image → Molecular Structure



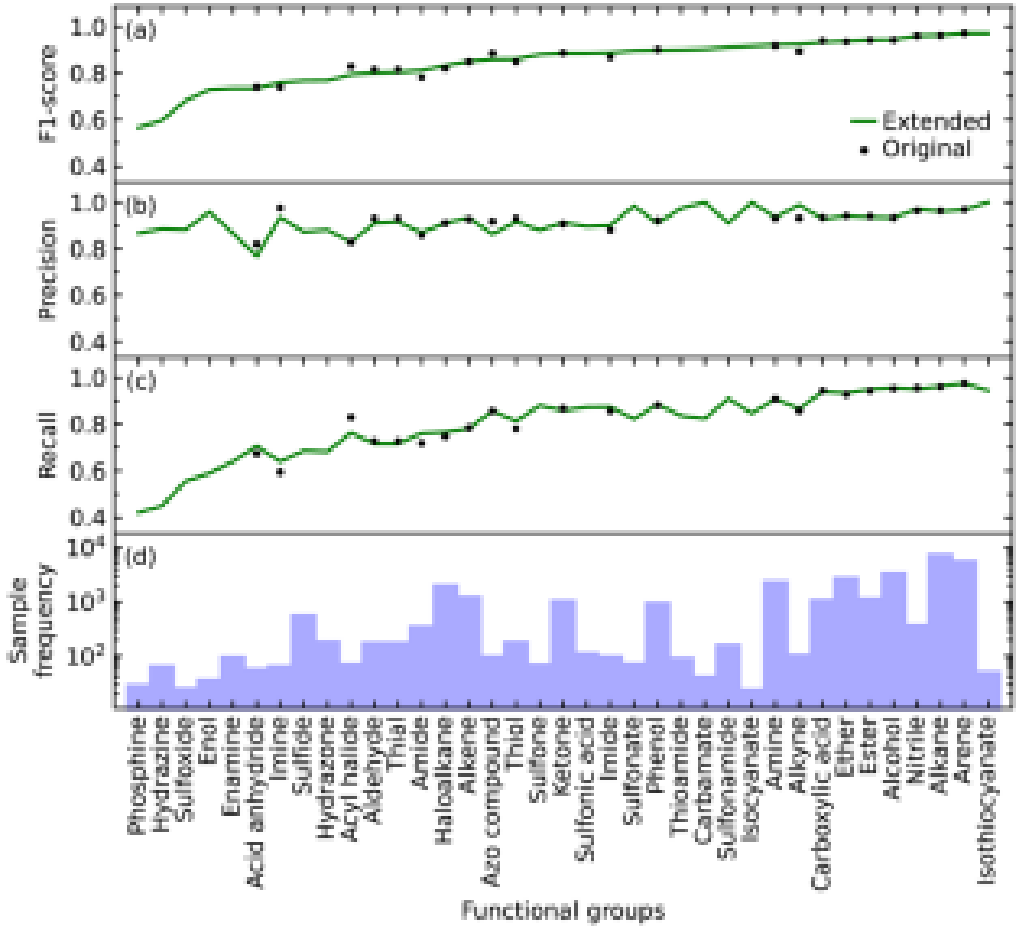


Spectral Image → Molecular Structure

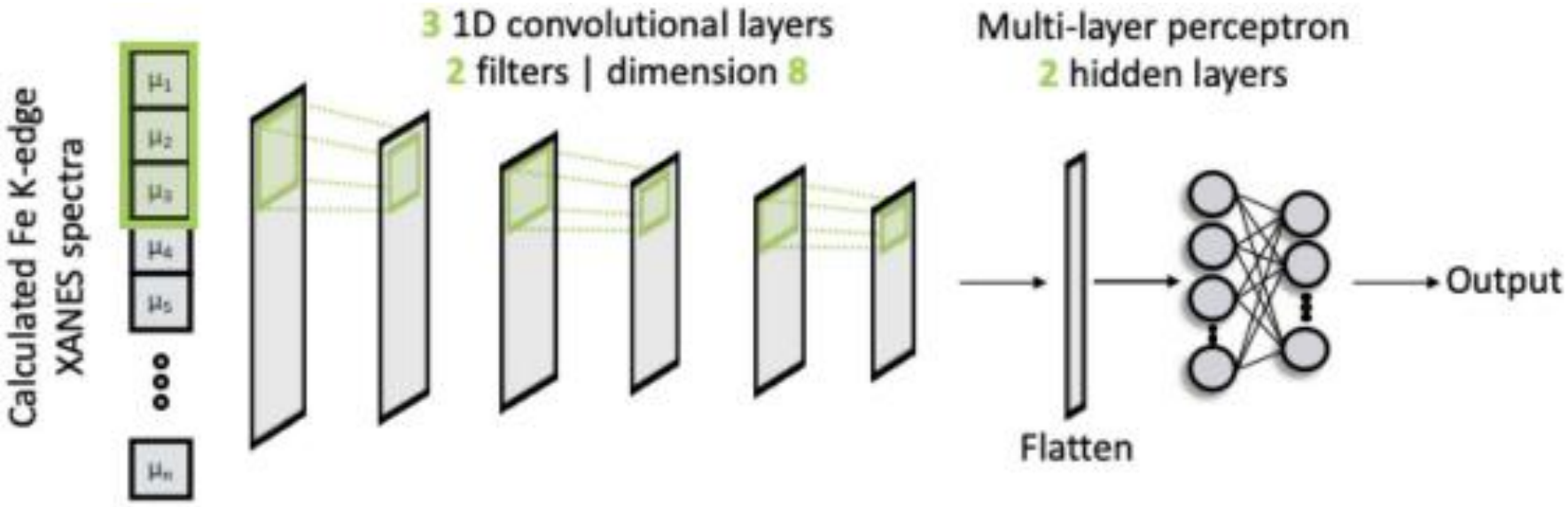
Training Data for CNN:

Experimental infra-red spectra dataset of 30,000 unique compounds

Technical validation: model can identify 37 functional groups of a molecule



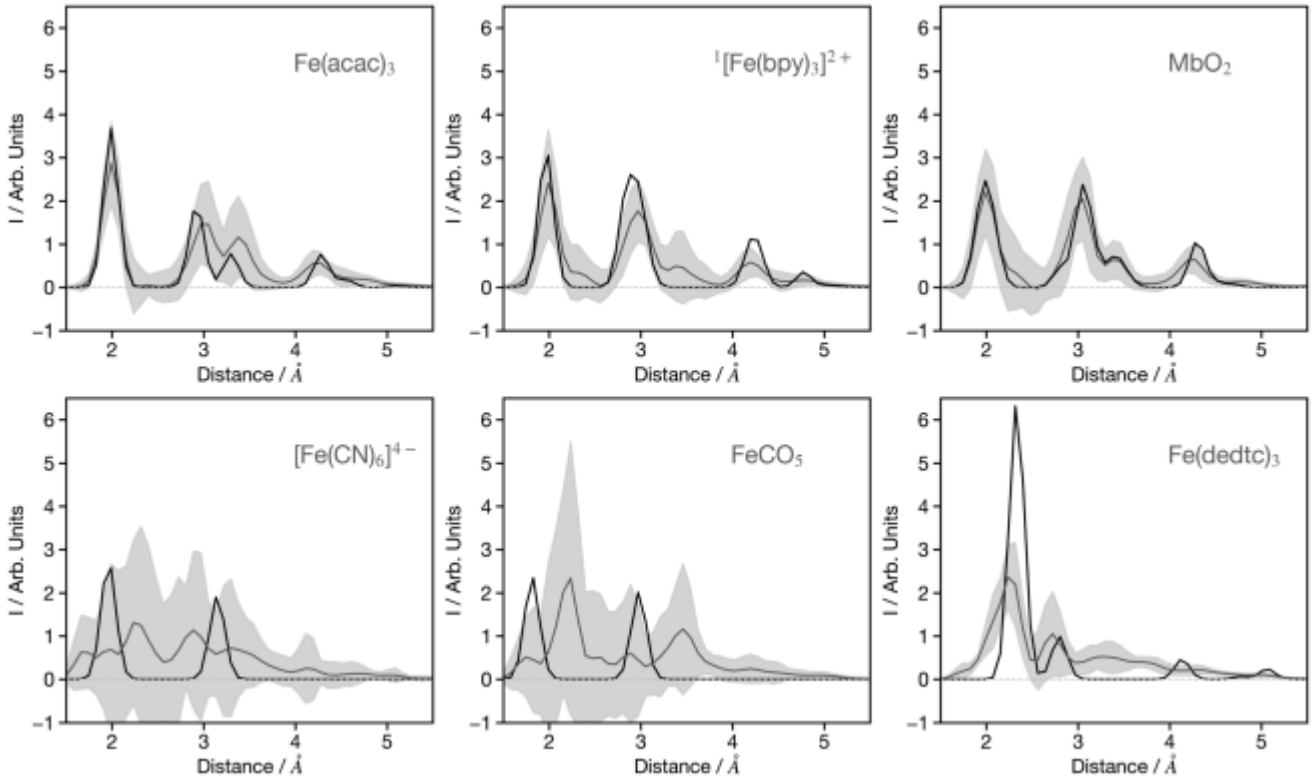
Spectral Image → Molecular Structure



CNN architecture trained on theoretically-generated X-ray absorption spectra

Penfold et al, Digital Discovery (2023) 2, 1461.

Spectral Image → Molecular Structure



Grey line = CNN Prediction from experimental spectra (but with CNN trained on theoretical data)

Black line = result from standard exptal approach to data analysis

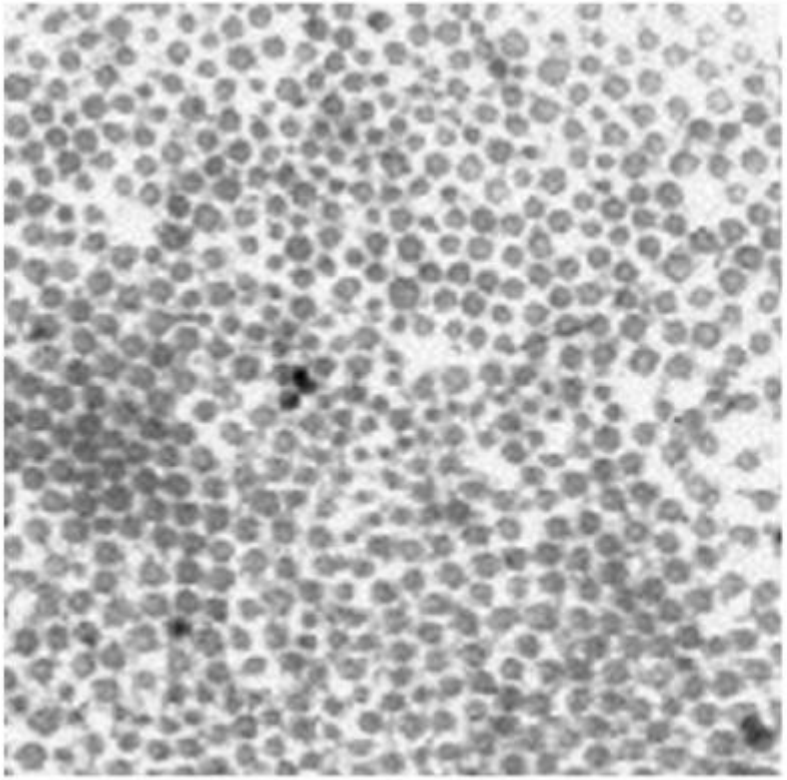
Penfold et al, Digital Discovery (2023) 2, 1461

Multi-modal Data

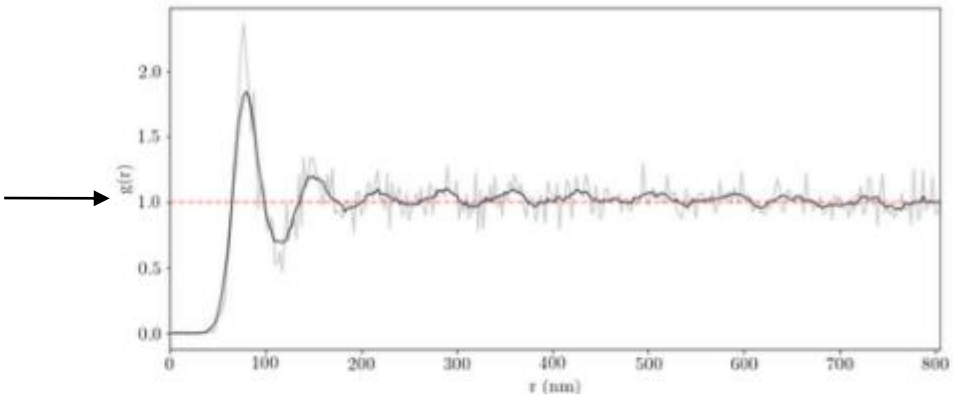
For ML training and for crossing materials-characterisation techniques

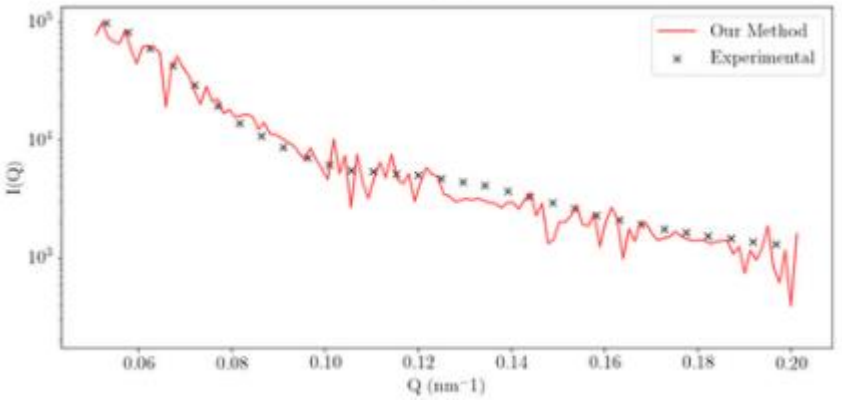
Yildirim, Washington & Cole (2020)

Image to RDF to SANS



Sheep spine collagen image





Conclusions

- Pipelines to curate high-quality databases on structure-property relationships
- AI architectures for auto-processing reduced data to afford structure
- Data analysis as we know it will become fully governed by AI
- The future data science?
  - will involve synonymous data scientists and ML experts
  - will employ multi-modal data sources as standard
  - will embrace language models for mainstream operations
  - may open or private (“data are the new oil”) – data quality is becoming a premium

What Facebook think…

And finally...

https://www.chemdatawriter.org

ChemDataWriter Home About Reader Finder Retriever Summariser Paraphraser Generator Documentation

ChemDataWriter

ChemDataWriter is a transformer-based toolkit for auto-generating books that summarise research

Learn more

Search, read, and retrieve

Summarise and paraphrase

Auto-generate research books

Source Code

Explore

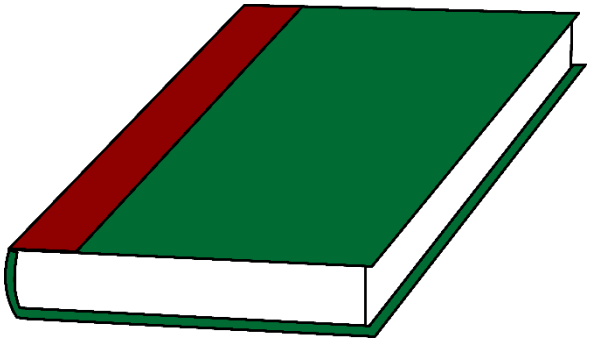
Documentation

Explore

ChemDataWriter

Auto-authored book

Fully-sectioned
Fully-indexed
Fully-referenced



Input

e.g. 152 papers on battery research

Output

e.g. 120 page book that reviews battery research

Acknowledgements

All the synthetic and materials characterisation chemists on the Advanced Energy Materials paper

Ganesh Sivaraman, Alvaro Vazquez-Mayaoitia, Venkat Vishwanath, Argonne National Laboratory.

Chris Cooper, Ed Beard, Matt Swain, Ed Beard, Shu Huang, Callum Court, Taketomo Isazawa, and all my research group @ Cambridge

ISIS Neutron & Muon Source

References:

Data-driven materials discovery:

Cole et al, *Adv. Ener. Mat* (2019) 9, 1802820

Cole, *Acc. Chem. Res*. 2020, 53, 3, 599–610

Databases:

Beard, Cole, *Sci. Data* (2022) 9, 329 (PV)

Beard et al, *Sci. Data* (2019) 6, 307 (UV/vis)

Huang, Cole, *Sci Data* (2020) 7, 260 (Battery)

Huang, Cole, *J. Chem. Inf. Model.* (2022) 62 (2022) 6365-6377 (BatteryBERT)

Text mining:

Swain, Cole, *J. Chem. Inf. Model* (2016)

ChemDataExtractor ImageDataExtractor ChemSchematicResolver ReactionDataExtractor