An overview of the impact of machine learning in molecular sciences.

AI Training Series – Fall 2022
11/8/2022

Álvaro Vázquez Mayagoitia
Argonne Nat Lab – CPS Division

ACKNOWLEDGMENTS

ALCF Acknowledgment

This research used resources of the Argonne Leadership Computing Facility, which is a DOE Office of Science User Facility supported under Contract DE-AC02-06CH11357.

Collaborators:

Ganesh Sivaraman (DSL)
Murat Keceli (CPS/ALCF)

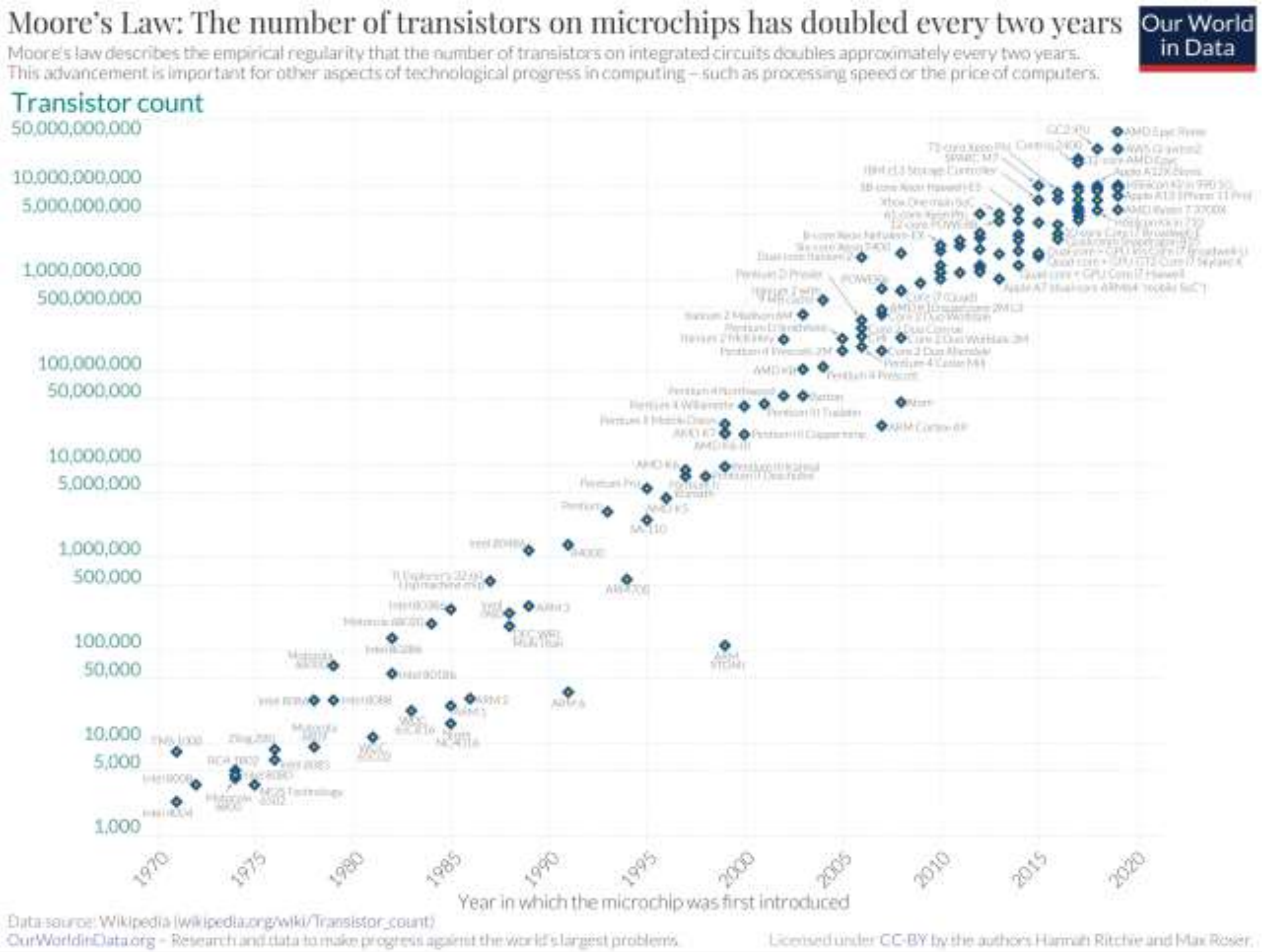
DoE HPC Roadmap: Exascale computing project (2021-2025)

Frontier AMD CPU,
AMD GPU

Perlmutter AMD CPU,
Nvidia GPU

Aurora Intel CPU,
Intel GPU

El Capitan AMD CPU,
AMD GPU



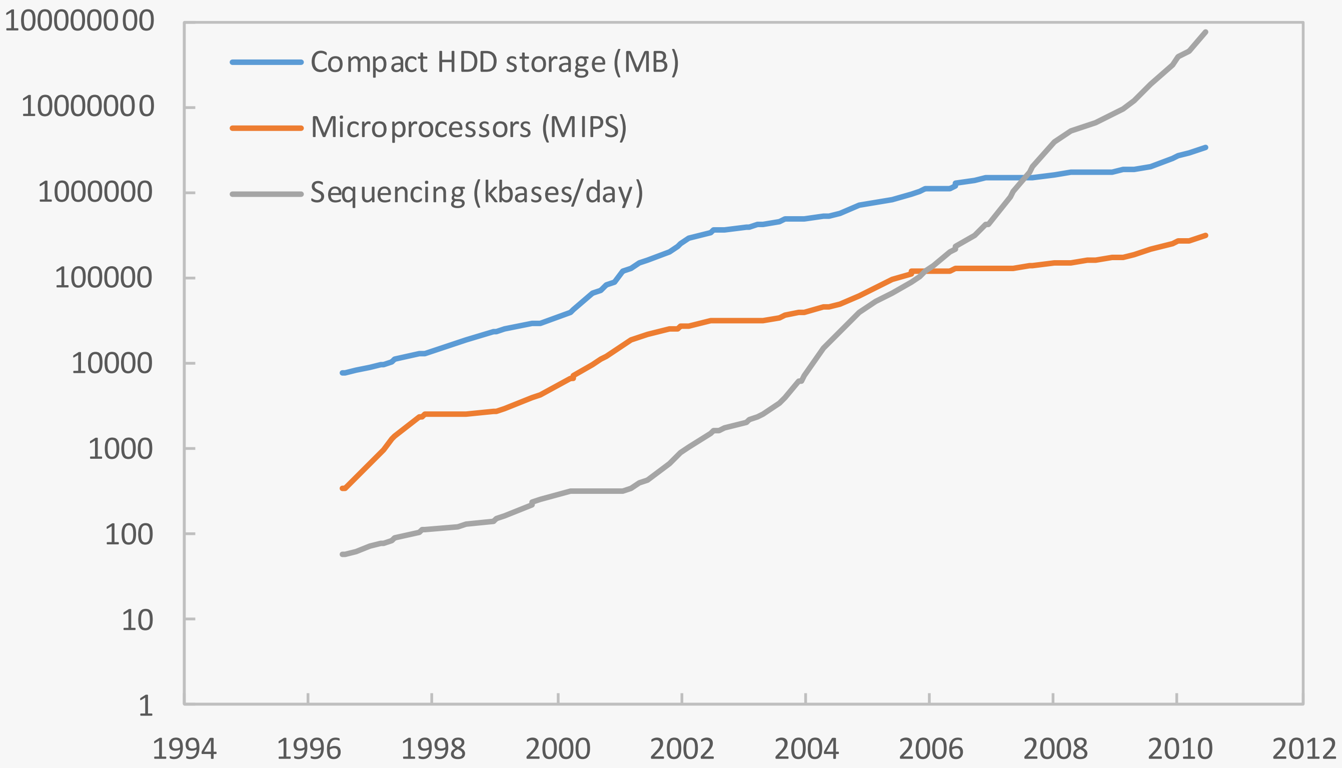
Moore's Law: The number of transistors on microchips has doubled every two years

Moore's law describes the empirical regularity that the number of transistors on integrated circuits doubles approximately every two years. This advancement is important for other aspects of technological progress in computing – such as processing speed or the price of computers.

Data source: Wikipedia (wikipedia.org/wiki/Transistor_count)

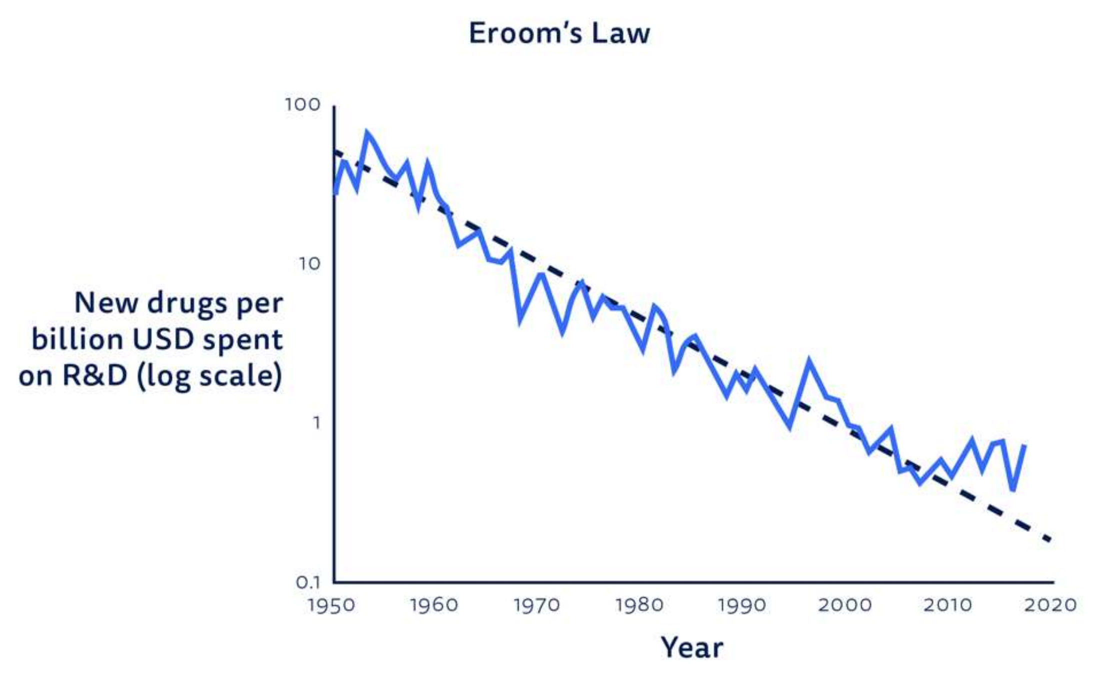
OurWorldinData.org – Research and data to make progress against the world's largest problems.

Evolution of DNA sequencing



In 2020’s: trillion bases a day

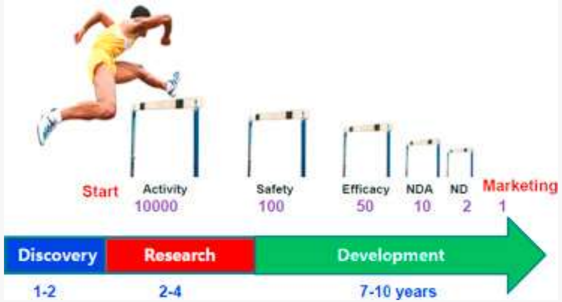
Decline in Pharmaceutical R&D efficiency

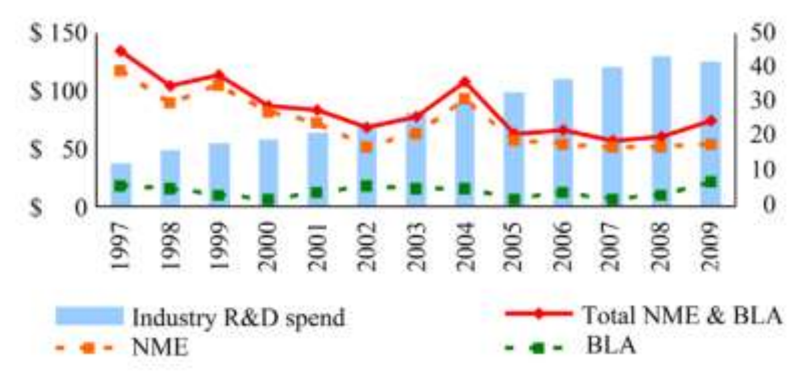


The **cost** of developing a new drug (~$2-3B) roughly **doubles every nine years**.

Scannell et al. Nature Reviews Drug Discovery, 2012, 11, 191-200
Olexandr Isayev http://olexandrisayev.com

R&D in Drug Discovery

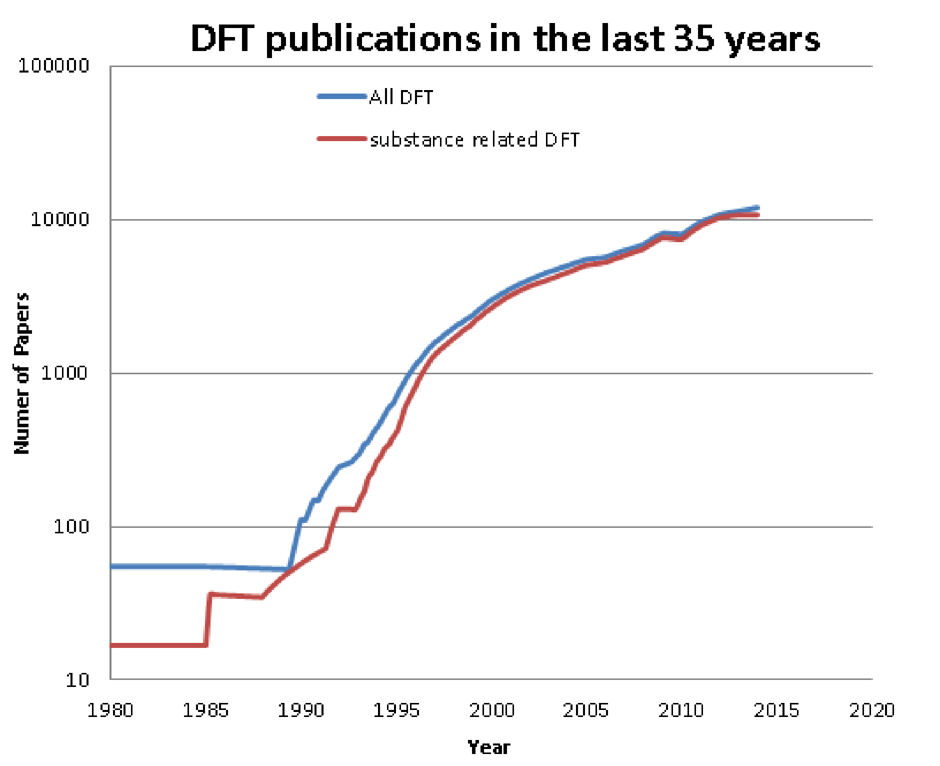




NME: New Molecular Entities
BLA: Biological App.

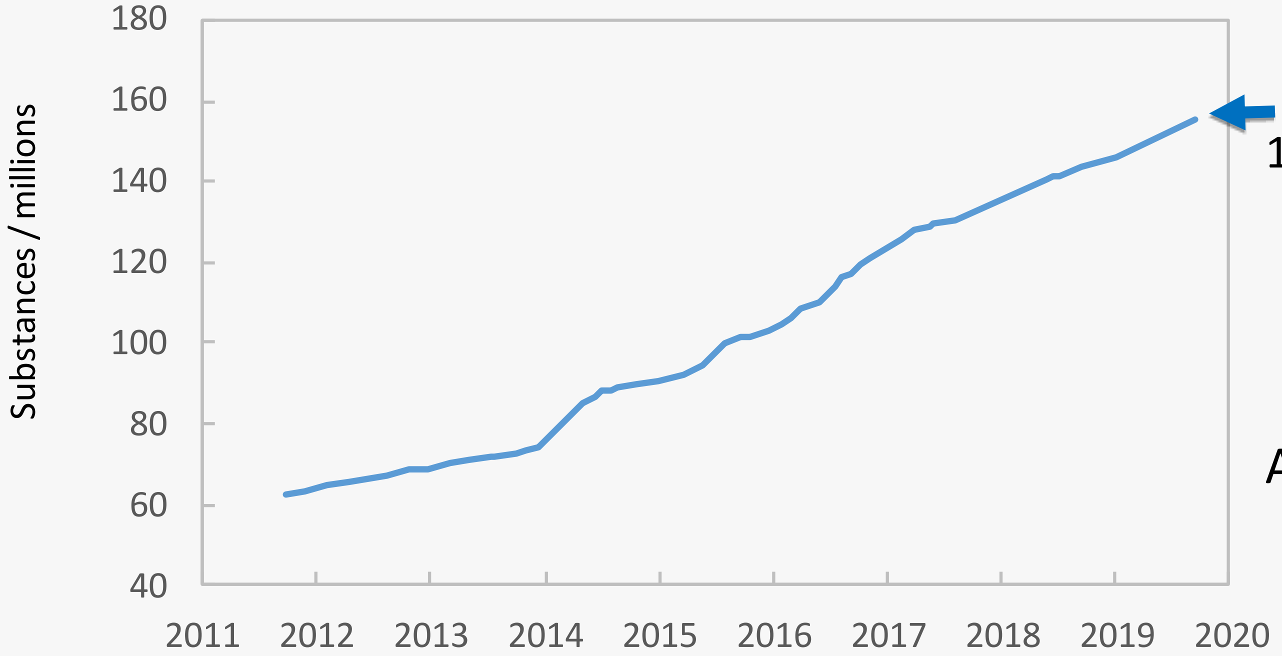
Acta Pharmaceutica Sinica B 2014;4(2):112–119

MATERIALS SCIENCE MODELING



*Web of science queries

Substances in CAS registry (ACS)

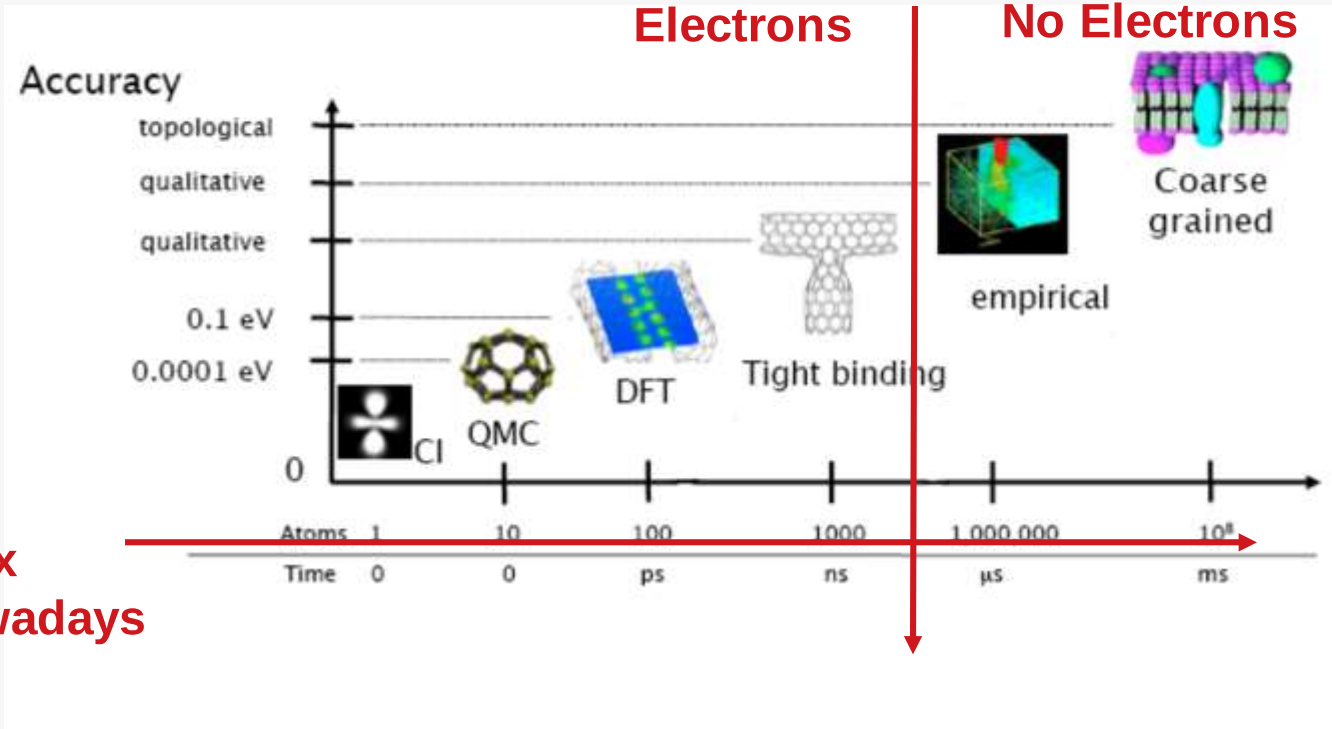


155M in Aug 2019

Are they any useful?

~~10M~~ 50M of new substances per year in average

250M in Abril 2021



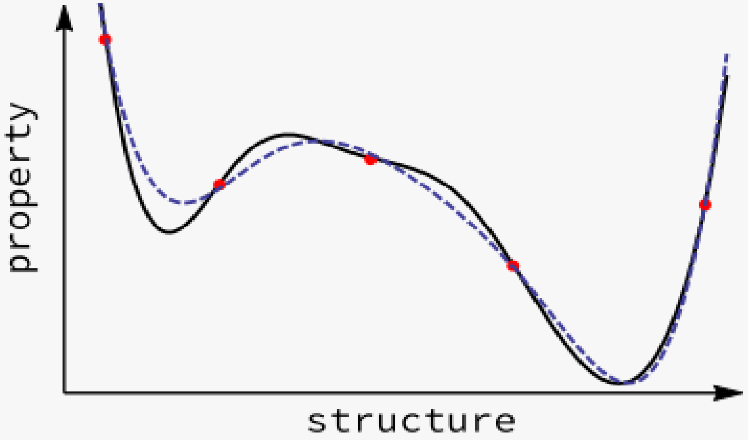
Electrons

No Electrons

100x nowadays

$$\hat{H}(\mathbf{R})\Psi(\mathbf{R},\mathbf{r}) = E\Psi(\mathbf{R},\mathbf{r})$$

Equation that describes the properties of an atom-scale system (time independent)



This is a hyper surface that can be approximated with empirical models.
ML models can remove bias and deal with very complex functions.

REVIEWS

Structure prediction drives materials discovery

Artem R. Oganov[1,2,3]*, Chris J. Pickard[4,5]*, Qiang Zhu[6] and Richard J. Needs[7]

Abstract | Progress in the discovery of new materials has been accelerated by the development of reliable quantum-mechanical approaches to crystal structure prediction. The properties of a material depend very sensitively on its structure; therefore, structure prediction is the key to computational materials discovery. Structure prediction was considered to be a formidable problem, but the development of new computational tools has allowed the structures of many new and increasingly complex materials to be anticipated. These widely applicable methods, based on global optimization and relying on little or no empirical knowledge, have been used to study crystalline structures, point defects, surfaces and interfaces. In this Review, we discuss structure prediction methods, examining their potential for the study of different materials systems, and present examples of computationally driven discoveries of new materials — including superhard materials, superconductors and organic materials — that will enable new technologies. Advances in first-principle structure predictions also lead to a better understanding of physical and chemical phenomena in materials.

[illegible]

New materials have historically been discovered by either trial and error processes or serendipity, both of which require labour intensive and challenging experiments. In the past decade, it has become possible to discover new materials systematically on a computer, and the path to this breakthrough has been paved by the development of crystal structure prediction (CSP) methods[1–3]. There are two largely complementary approaches: one based on existing knowledge and the contents of crystal structure databases (data mining) and the other based on powerful exploratory computer algorithms capable of making predictions with little or no pre-existing knowledge.

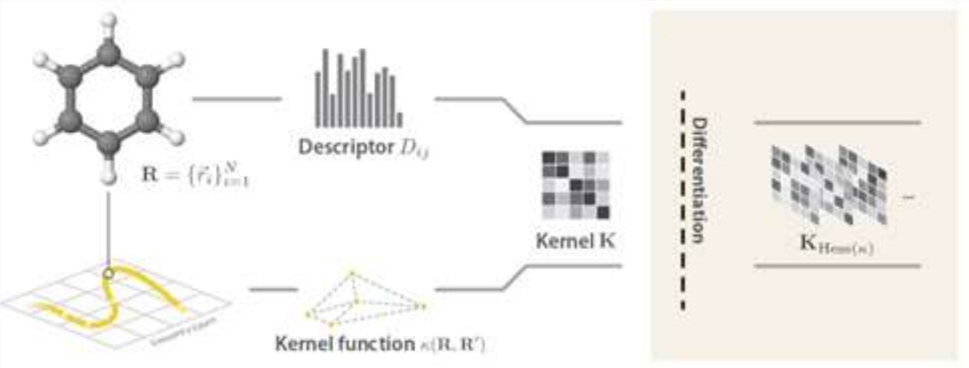
Databases such as the Inorganic Crystal Structure Database (ICSD)[4] and the Pauling File[5] are invaluable resources that report experimentally observed structures of inorganic materials. The ICSD currently contains approximately 200,000 peer reviewed data entries, and the Pauling File has some 310,000 entries; both databases are growing steadily. To date, approximately 120,000 ICSD entries have been assigned to a little over 9,000 distinct structure prototypes. However, the number of structures in the ICSD is small compared with the rapidly increasing number of structures that have been generated using first principles methods.

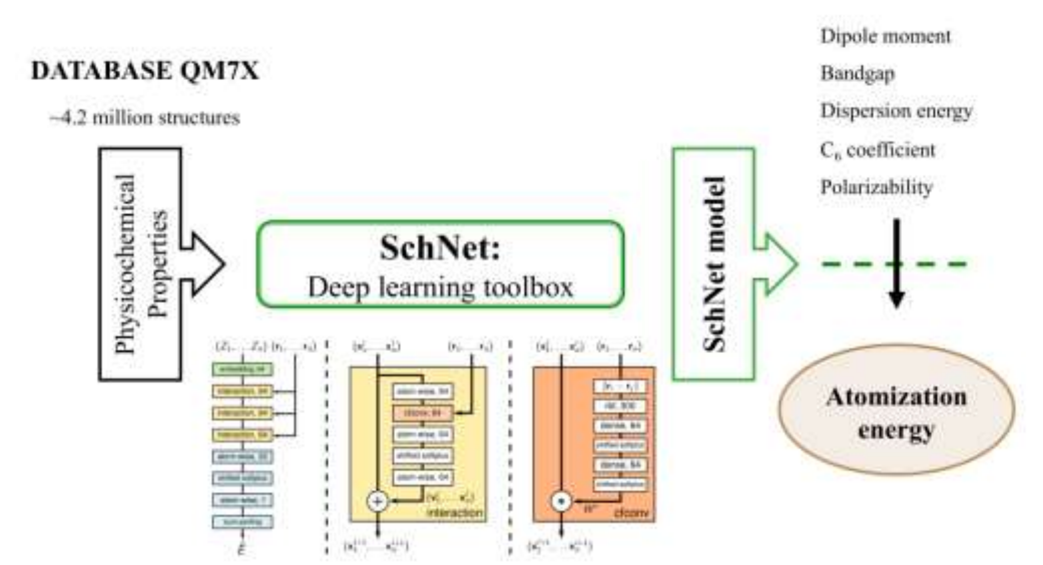
Data mining approaches have received much attention[6,7]. In this Review, we focus on the fundamental and reliable non-empirical methods based on powerful exploratory algorithms. The major advantage of such methods is their ability to generate completely new knowledge beyond existing databases and intuition. We discuss these methods, examining the basic concepts and the systems they can be applied to. We then highlight examples of recent discoveries of counterintuitive new materials and phenomena achieved through the use of these methods, ranging from superhard materials to electrides, organic materials and superconductors with the highest known critical temperatures[8–11]. We finally outline the future perspectives for the field, examining the challenges that will need to be overcome, which include working with large systems and taking into account disorder and temperature, predicting synthesizable metastable structures and predicting chemical properties.

Considerations for materials prediction

Structure. Understanding the structure of matter at the atomic level is central to modern materials science. Until recently, experiments offered the only reliable source of crystal structures, but computational methods have emerged as a complementary source. In particular, density functional theory (DFT) methods and modern computing power can be combined to relax thousands of structures to local minima in the Born–Oppenheimer energy surface, the potential energy surface calculated in the approximation that the nuclear and electronic motions can be treated separately[12]. CSP aims to determine the minimum energy (or, in a more general case, the global minimum or maximum of a property of interest) over all values of the relevant inputs while

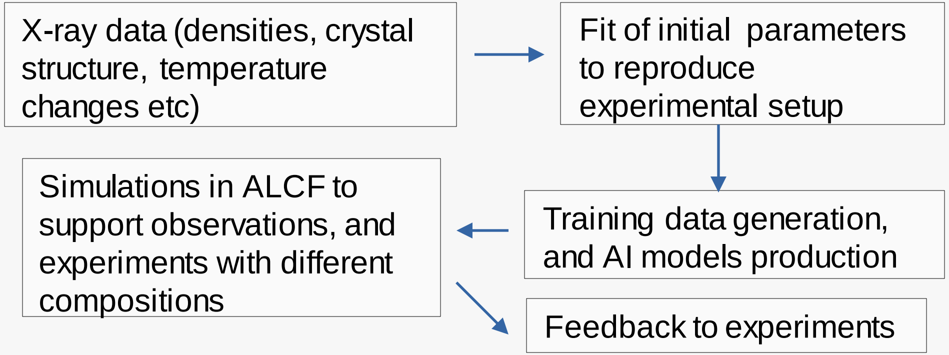
*Nature Reviews, Materials May 2019


$\mathbf{R} = \{\vec{r}_i\}_{i=1}^{N}$
Descriptor D_{ij}
Kernel K
Differentiation
$\mathbf{K}_{\mathrm{Hess}(\kappa)}$
...
Kernel function $\kappa(\mathbf{R}, \mathbf{R}')$


DATABASE QM7X
~4.2 million structures
Physicochemical Properties
SchNet:
Deep learning toolbox
SchNet model
Dipole moment
Bandgap
Dispersion energy
C_6 coefficient
Polarizability
Atomization energy
interaction
cfconv

Data driven AI/ML interatomic potentials for large scale multi physics simulations

- Currently developing AI/ML models, to predict with QM accuracy, energies and forces for in- and out- equilibrium, comparable to experimental observations.
- Our models will ultra fast infer (using GPU and FPGA accelerators) properties for MD and PIMD.



X-ray data (densities, crystal structure, temperature changes etc)

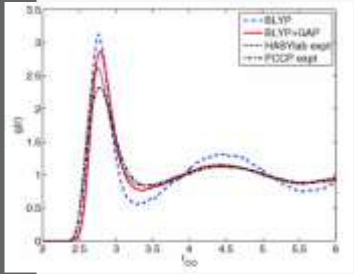
Fit of initial parameters to reproduce experimental setup

Training data generation, and AI models production

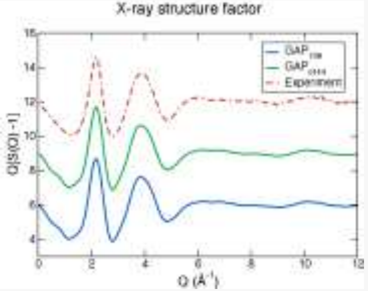
Simulations in ALCF to support observations, and experiments with different compositions

Feedback to experiments

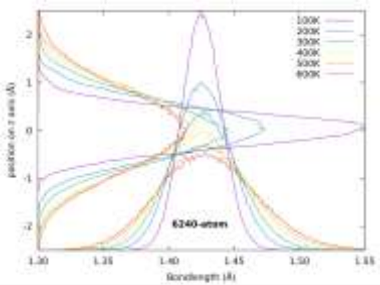
Water



Hafnia



Graphene



Perovskites

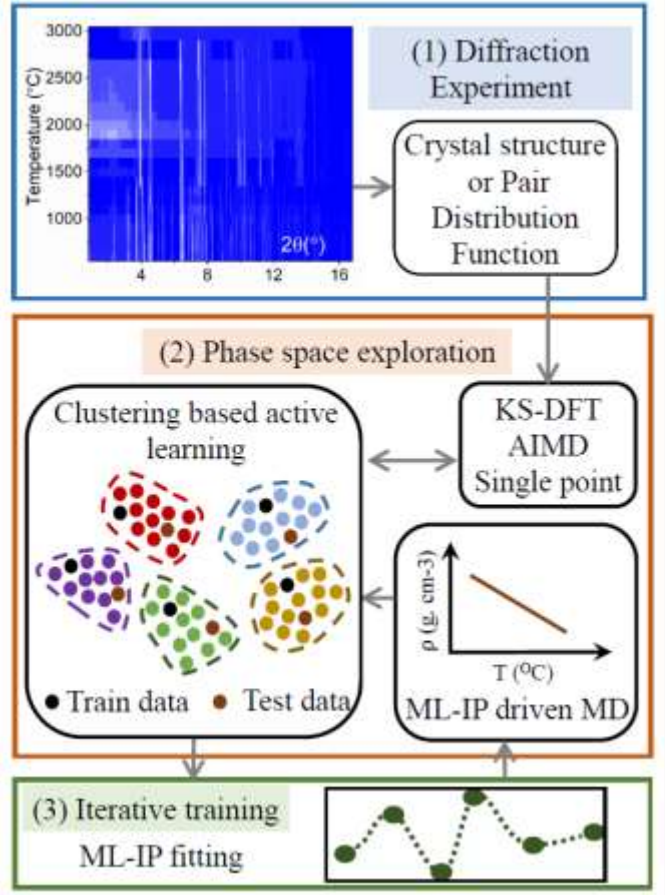
Molten salts

Silica

Single atom catalysis

...

Project: ML inter atomic potential for all hafnia phases.



(1) Diffraction Experiment

Crystal structure or Pair Distribution Function

(2) Phase space exploration

Clustering based active learning

● Train data ● Test data

KS-DFT AIMD Single point

ML-IP driven MD

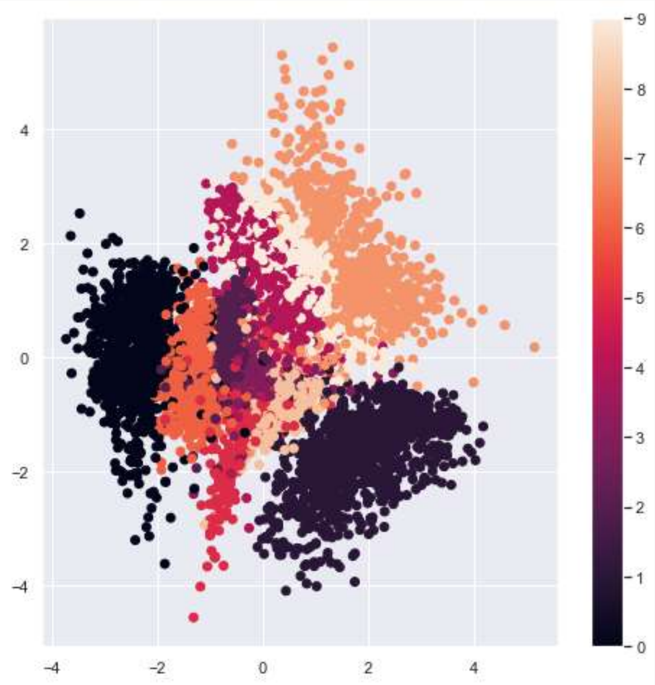
(3) Iterative training

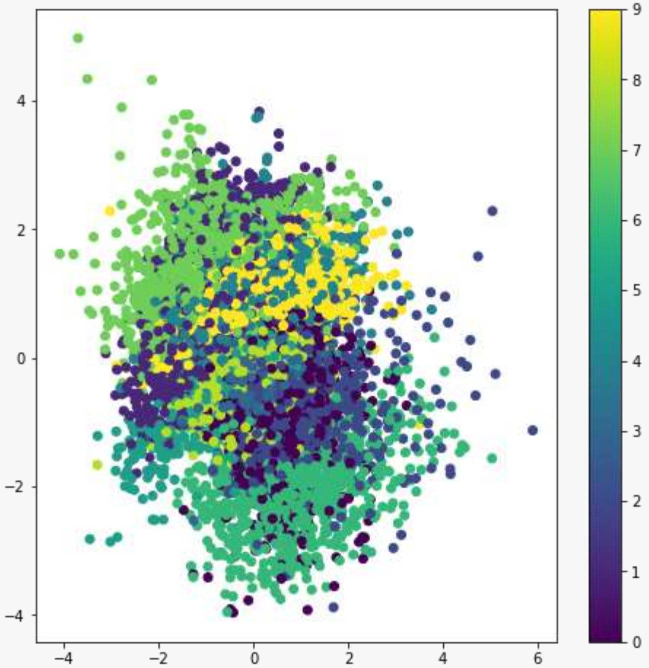
ML-IP fitting

PRL, 2021, Editor’s Pick

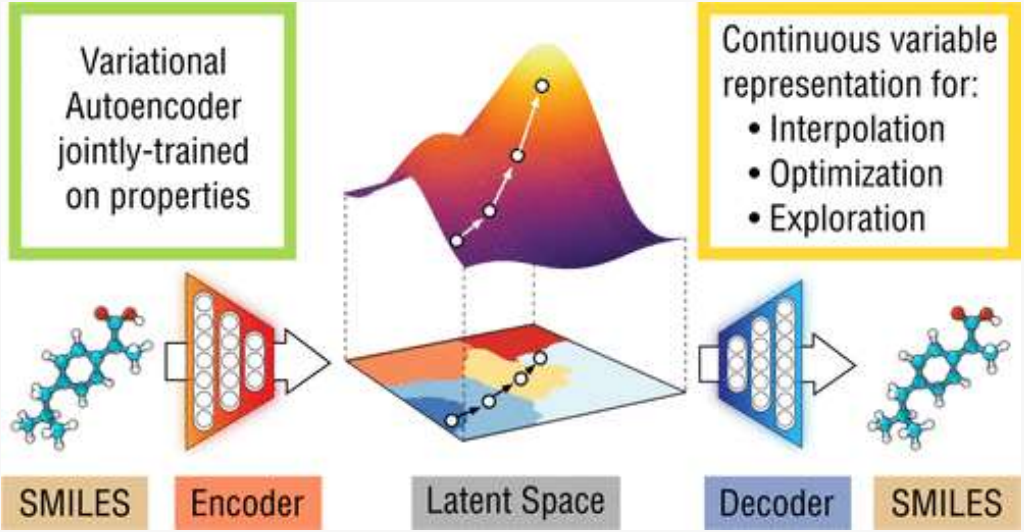
VAE-MNIST

CVAE-MNIST





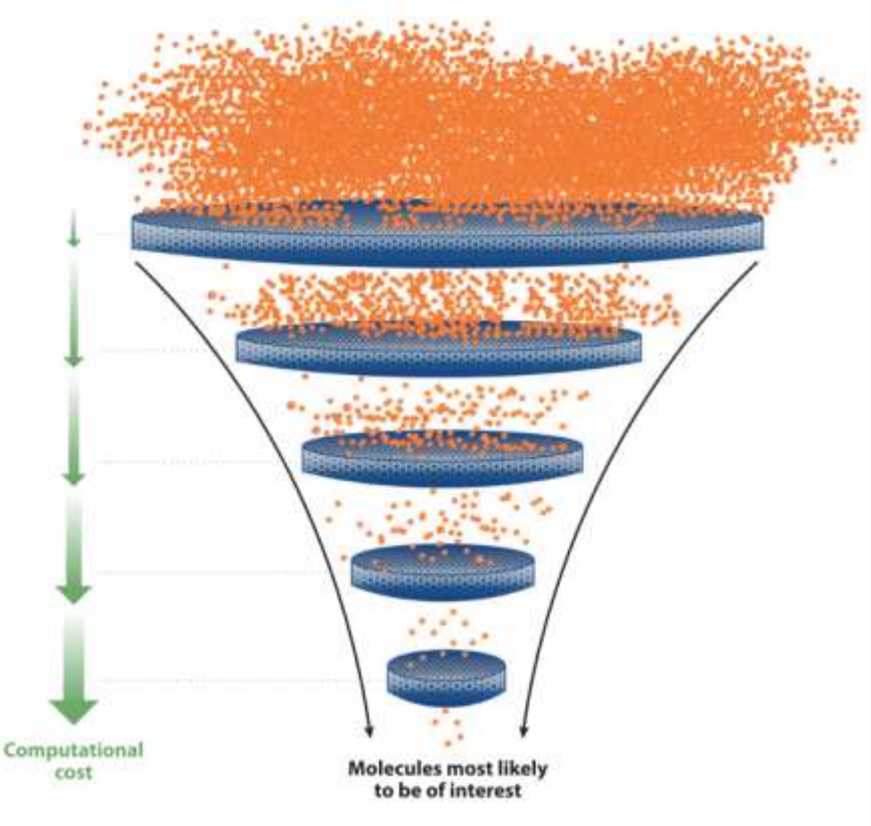
VAE jointly trained with a Regressor


Variational Autoencoder jointly-trained on properties
Continuous variable representation for:
• Interpolation
• Optimization
• Exploration
SMILES
Encoder
Latent Space
Decoder
SMILES

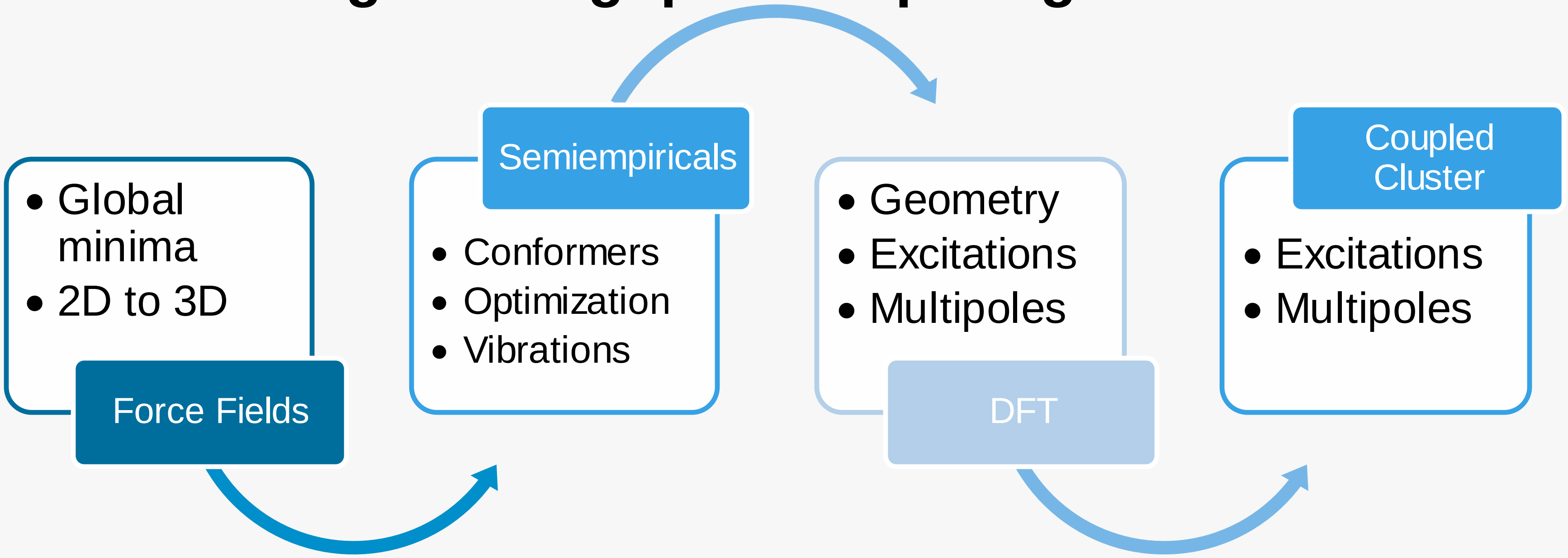
Bombarelli *et al.*,*ACS Cent. Sci.*, 2018, 4 (2), pp 268–276

Funnel approach

Virtual screening of the chemical space



Informed high throughput computing



Force Fields

- Global minima
- 2D to 3D

Semiempiricals

- Conformers
- Optimization
- Vibrations

DFT

- Geometry
- Excitations
- Multipoles

Coupled Cluster

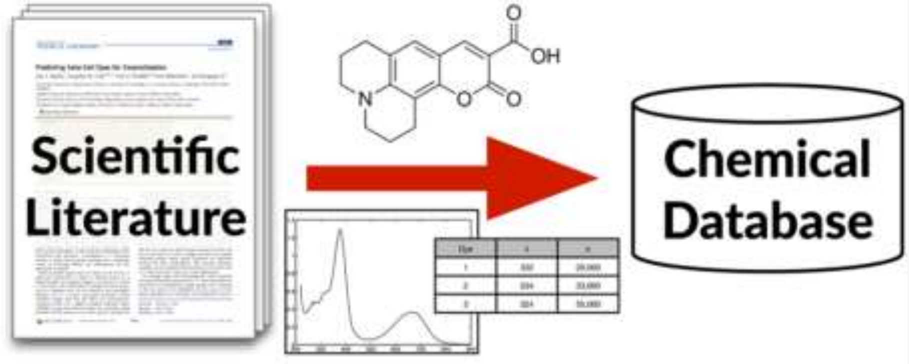
- Excitations
- Multipoles

Composite of codes:

- Babel
- Rdkit
- MOPAC
- ORCA
- NWChem

ChemDataExtractor

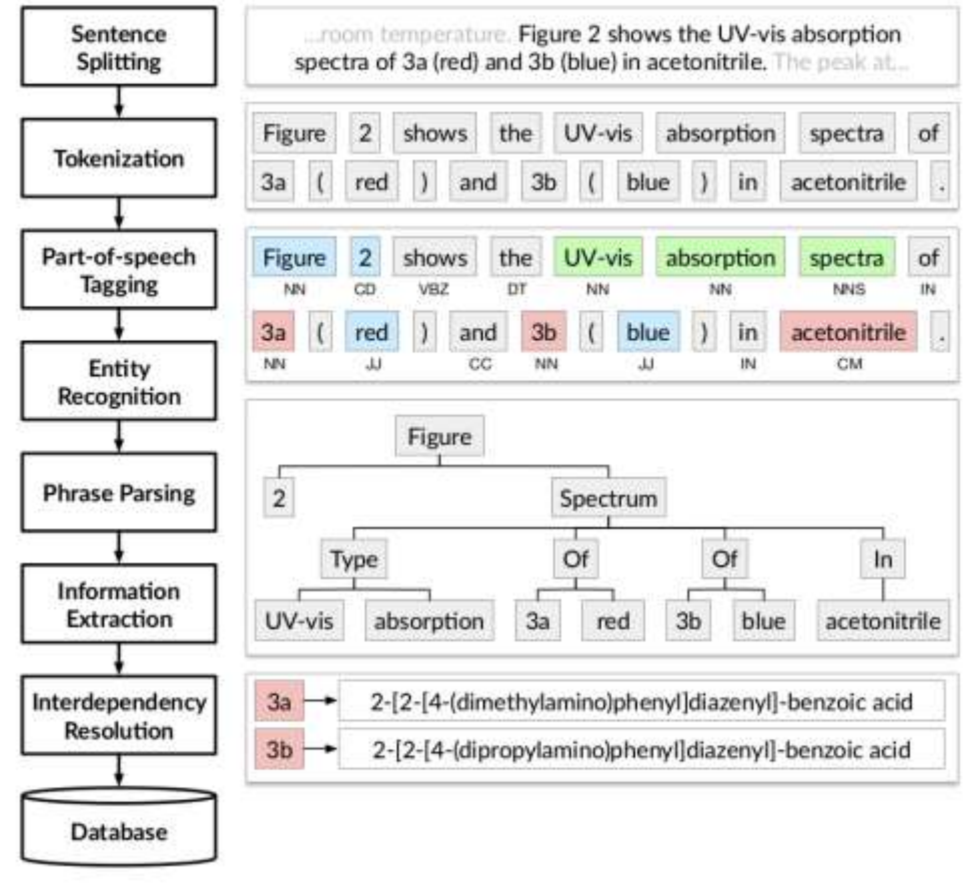
A Toolkit for Automated Extraction of Chemical Information from the Scientific Literature



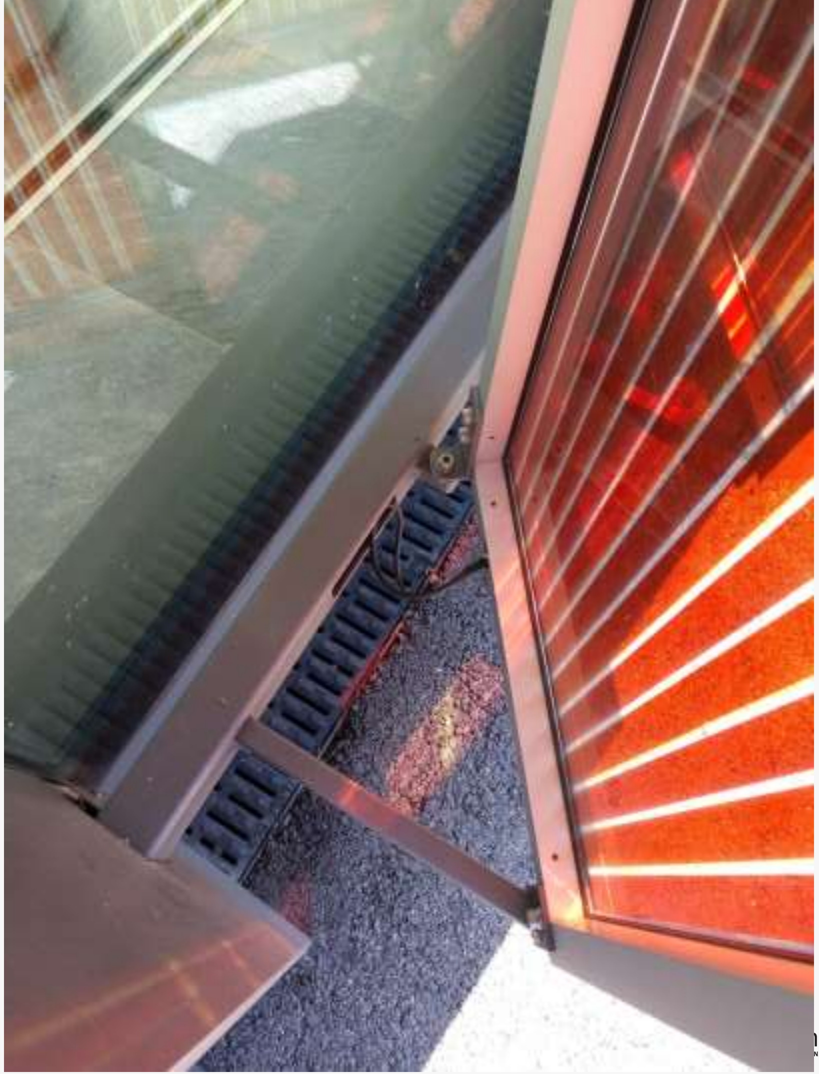
http://chemdataextractor.org

Swain, M. C., & Cole, J. M. J. Chem. Inf. Model. 2016

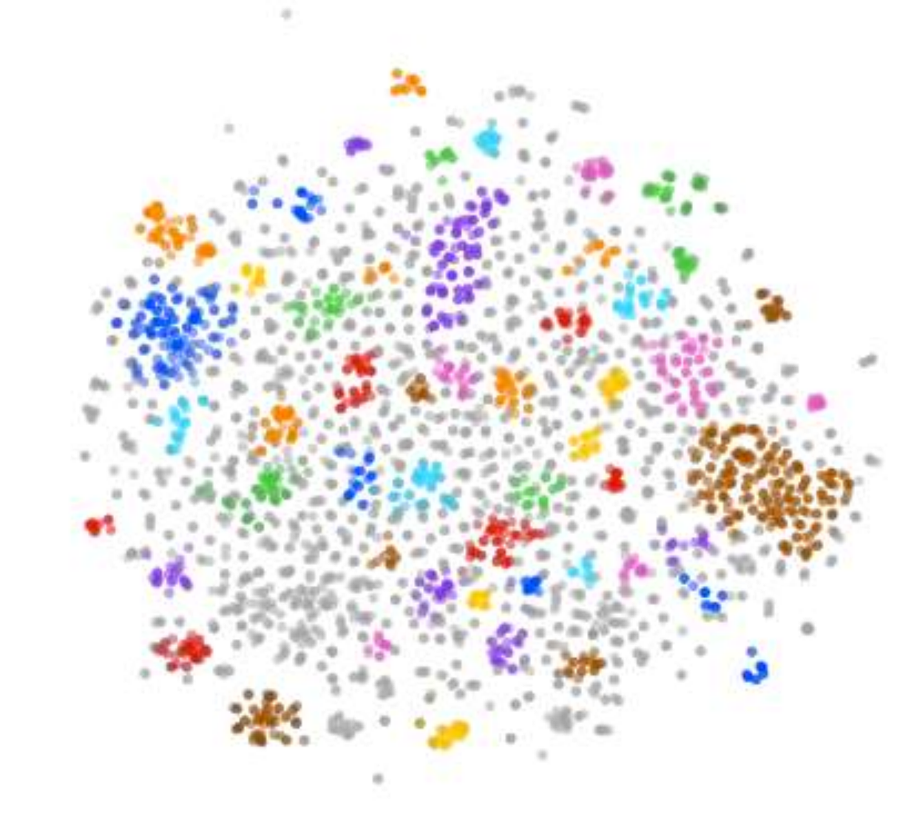
NATURAL LANGUAGE PROCESSING PIPELINE


Sentence Splitting
Tokenization
Part-of-speech Tagging
Entity Recognition
Phrase Parsing
Information Extraction
Interdependency Resolution
Database
...room temperature. Figure 2 shows the UV-vis absorption spectra of 3a (red) and 3b (blue) in acetonitrile. The peak at...
Figure 2 shows the UV-vis absorption spectra of
3a (red) and 3b (blue) in acetonitrile .
Figure NN 2 CD shows VBZ the DT UV-vis NN absorption NN spectra NNS of IN
3a NN (red JJ) and CC 3b NN (blue JJ) in IN acetonitrile CM .
Figure
2
Spectrum
Type
Of
Of
In
UV-vis
absorption
3a
red
3b
blue
acetonitrile
3a → 2-[2-[4-(dimethylamino)phenyl]diazenyl]-benzoic acid
3b → 2-[2-[4-(dipropylamino)phenyl]diazenyl]-benzoic acid

Swain, M. C., & Cole, J. M. J. Chem. Inf. Model. 2016



Molecular cartography - clusters

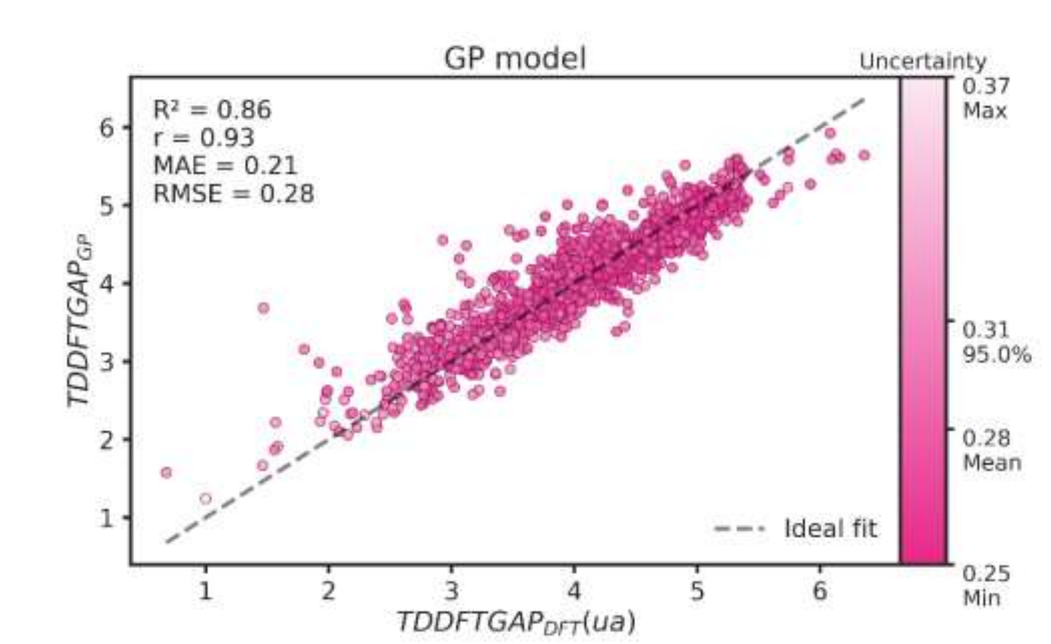


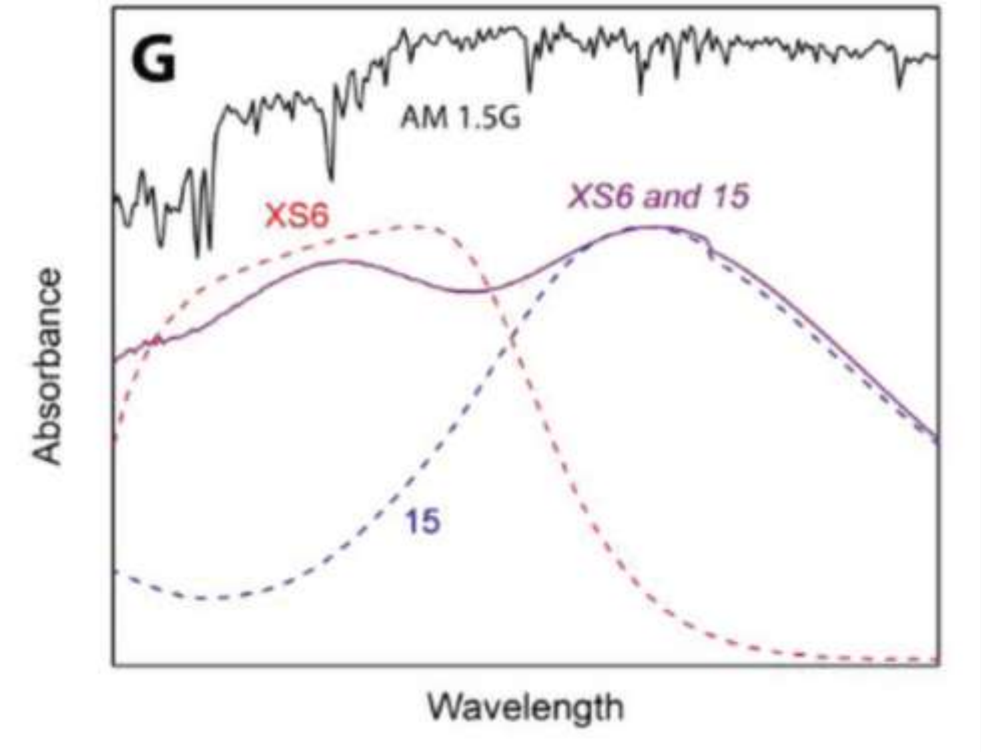
Dimension reduction with t-SNE
Clustering with HDBSCAN

Learn from data and feedback to experiments

Transition prediction

TDDFT gap prediction – We used Gaussian Process and Circular Morgan Fingerprints to predict the first transition of the a reduce scale TDDFT (sTDA//wB97X-D3/TZVP), we found that this value is predictable. Similar result found for HOMO-LUMO DFT gap.





pvinsights.com

Q&A