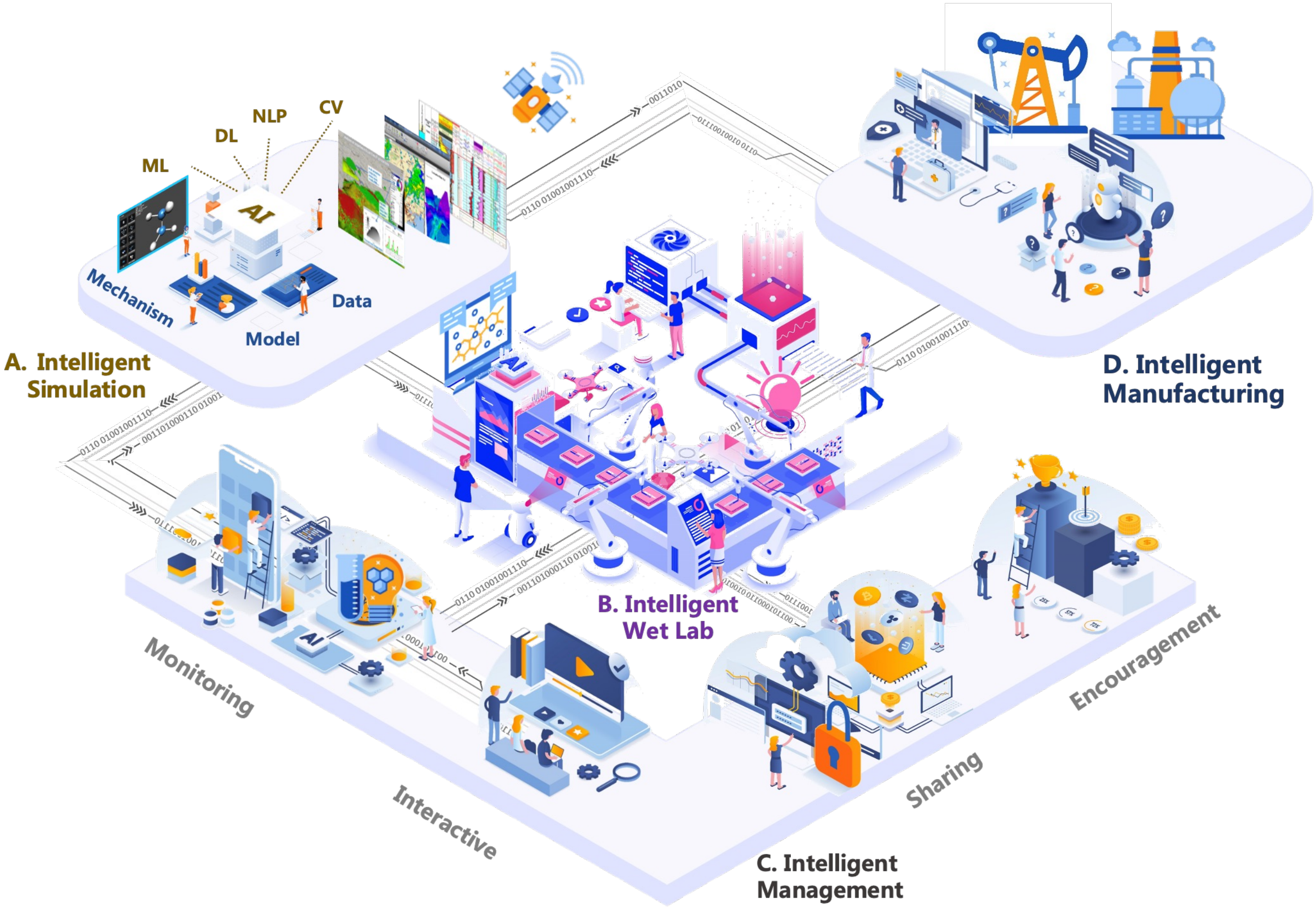


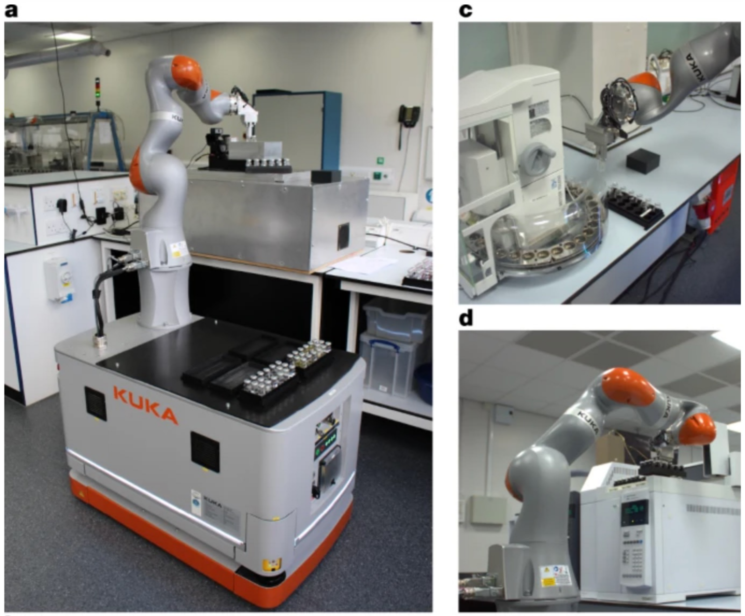
INTRODUCTION TO DEEP LEARNING BASED MATERIALS DISCOVERY AND INFORMATICS

Presenter: Li Jiali

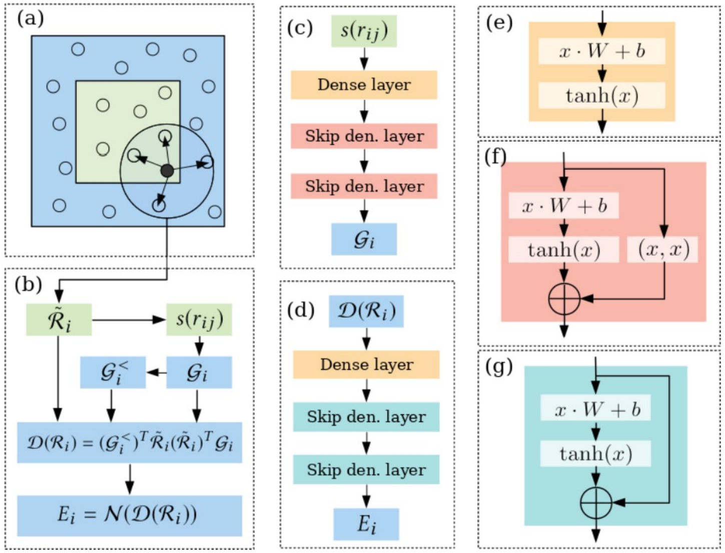
Outline

- Part 1 General Background
- Part 2 New Research Norm
- Part 3 Conclusions

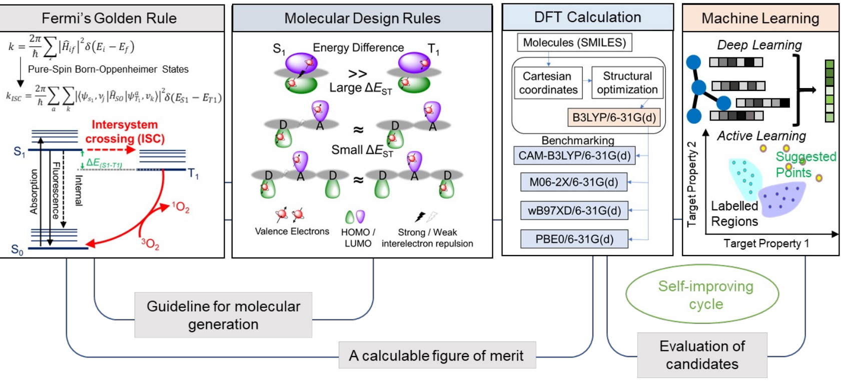
AI for Material and Chemistry (Quick Look)



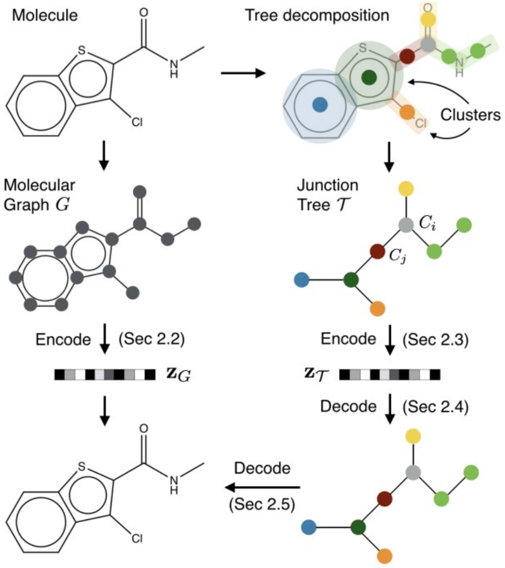
Mobile Chemist (In a wet lab)



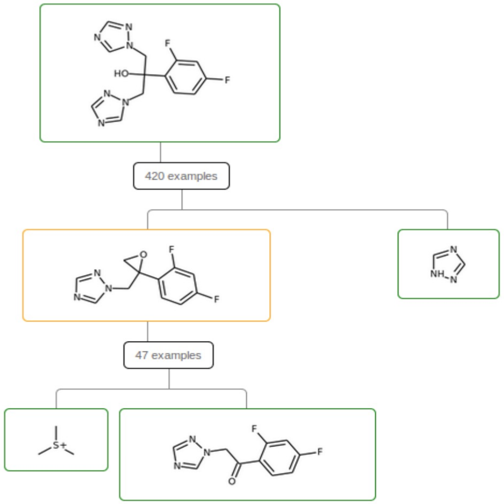
AI boosted simulation



AI for property prediction



AI for material generation



AI for Retrosynthsis

What is AI? In an intuitive way

$$Property = w_1 \times \ Molecular\ weight + w_2 \times \ temperature + w_3 \times time$$

A Simple function: $y = w_1\ x_1 + w_2\ x_2 + w_3\ x_3$

Written in Matrix Form: $y = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \cdot \begin{bmatrix} w_1 \\ w_2 \\ w_3 \end{bmatrix}$

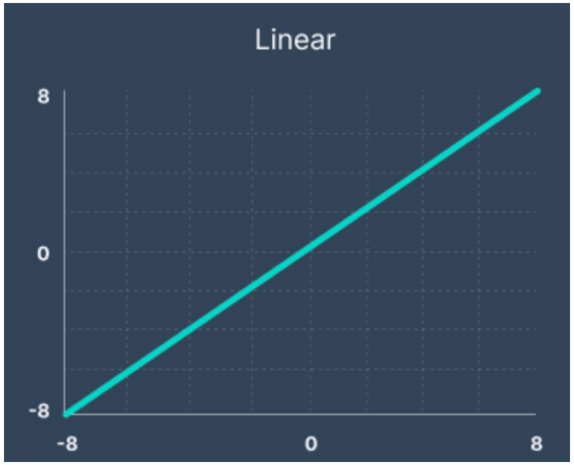
If y is a vector such as two properties of materials (one possible form of the function): $\begin{bmatrix} y_1 & y_2 \end{bmatrix} = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix} \times \boxed{\begin{bmatrix} w_1 & w_4 \\ w_2 & w_5 \\ w_3 & w_6 \end{bmatrix}}$

Matrix Shape $1 \times 2 = 1 \times 3\ mul\ 3 \times 2$

In terms of function: $\begin{bmatrix} y_1 & y_2 \end{bmatrix} = \begin{bmatrix} w_1\ x_1 + w_2\ x_2 + w_3\ x_3 & w_4\ x_1 + w_5\ x_2 + w_6\ x_3 \end{bmatrix}$

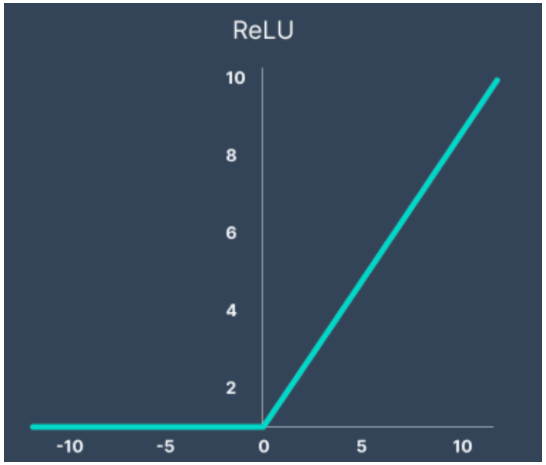
AI is to learn the **parameter of a function** from the **given data** with some **optimization methods**.

Linear and non-linear function



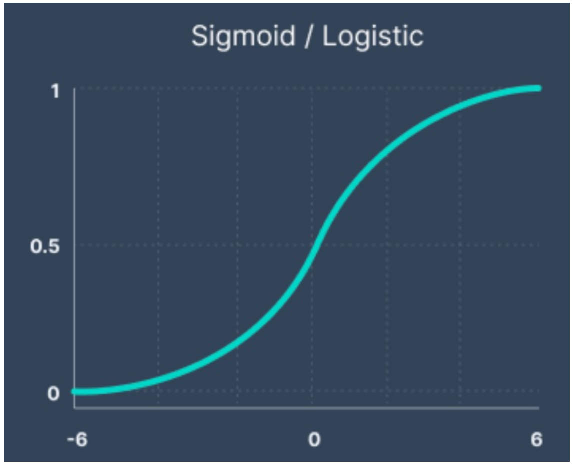
Linear Function

$$f(x) = ax + b$$



Non-Linear Function

$$f(x) = \max(0.0, x)$$



Non-Linear Function

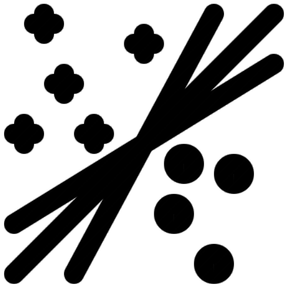
$$f(x) = \frac{1}{1 + e^{-x}}$$

A linear function is a function whose graph is a straight line, that is, a polynomial function of degree zero or one.

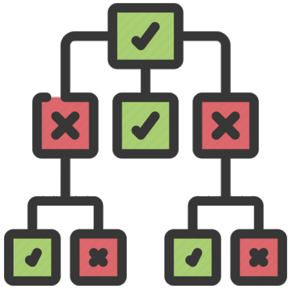
A nonlinear function is a function whose graph is NOT a line. A nonlinear system is a system in which the change of the output is not proportional to the change of the input. In reality, most of the science problems are non-linear systems.

Functions in different forms

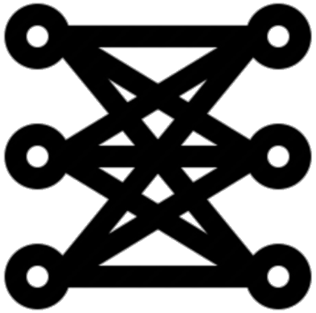
Three representative ML methods



Support Vector Machine



Decision Tree



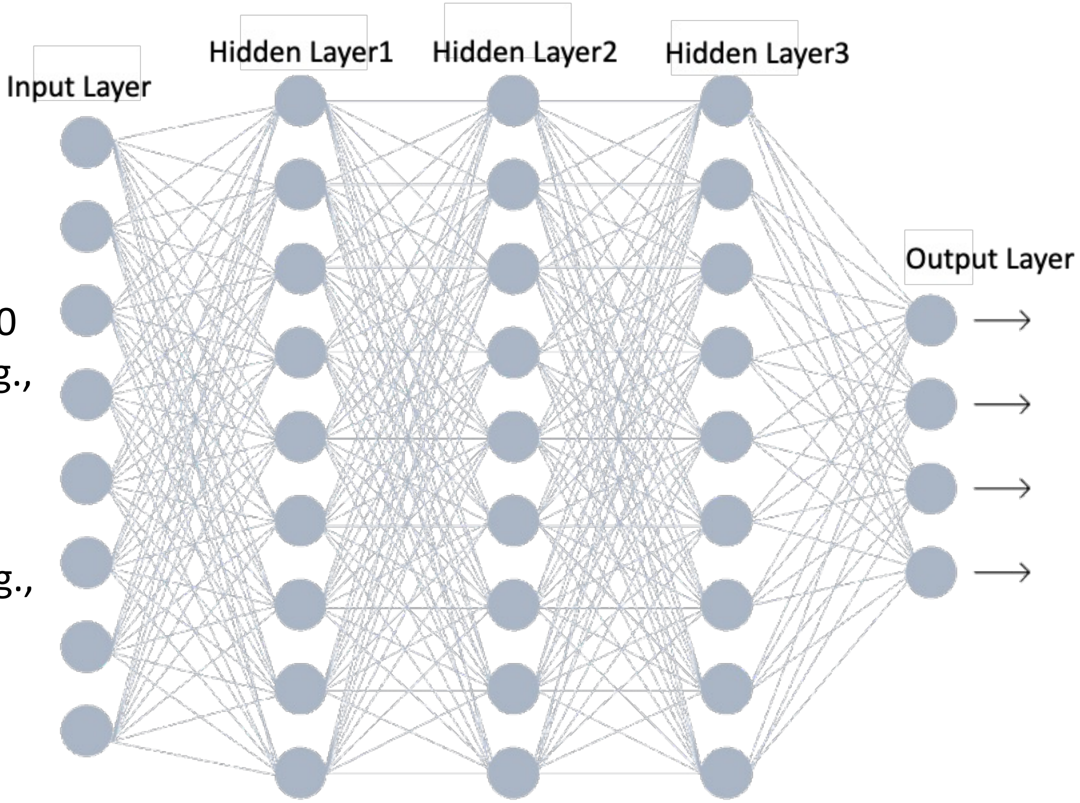
Neural Network

Why different? (a simple version) They are different in the data structures to store the trainable parameters.

SVM stores parameters in matrix form. Decision Tree stores parameters in tree form. Neural Network stores parameters with a series of matrix and with some non-linear operation.

Functions in different forms

Take neural network as an example, each layer of the neural network is a matrix.

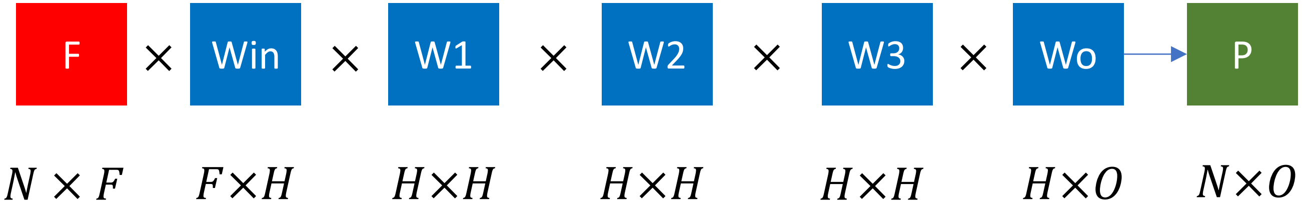


N: Number of input e.g., 50

F: Dimension of feature e.g., 10

H: Dimension of hidden vector e.g., 256

O: Dimension of output e.g., 4



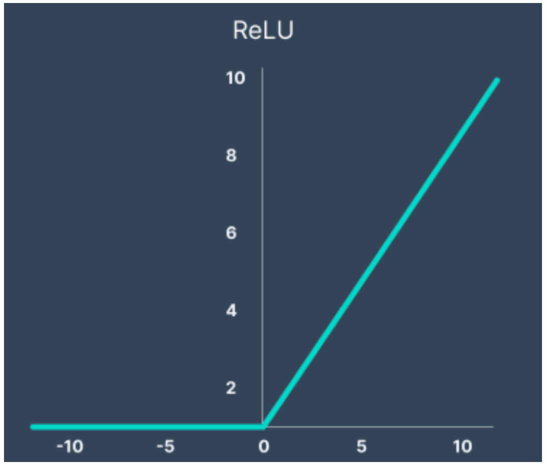
F × Win × W1 × W2 × W3 × Wo → P

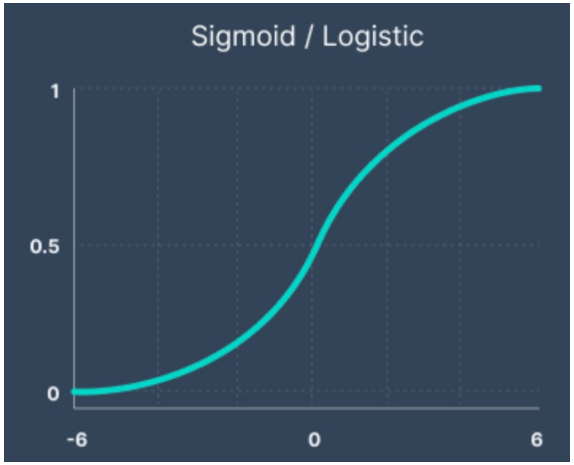
$N \times F$ $\quad F\times H$ $\quad H\times H$ $\quad H\times H$ $\quad H\times H$ $\quad H\times O$ $\quad N\times O$

The whole process is a **message transforming process**. The input message is what we know such as temperature, molecular structure and so on. With several transforming matrices, the input information can be linked to the output properties.

Functions in different forms

If only matrix directly multiple with each other. It is like doing linear transformation for multiple times. Then the whole function is still a linear function.





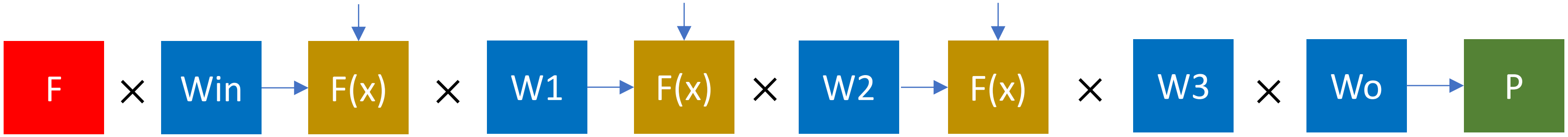
Non-Linear Function

$$f(x) = \max(0.0, x)$$

Non-Linear Function

$$f(x) = \frac{1}{1 + e^{-x}}$$

Add non-linear operation into the message transformation process



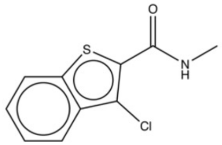
F × Win → F(x) × W1 → F(x) × W2 → F(x) × W3 × Wo → P

Data and learning schemes in different forms

Three major forms

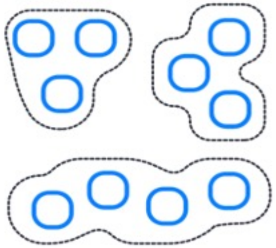
Supervised Learning:

	(X, Y)	X	Y
Simple output	$e.g.,$	$[Temperature \quad MW]$	$[Atomically\ Precise]$
Complex output			$[The\ molecule\ structure\ with\ desired\ property]$



Unsupervised Learning:

Only input (X) Find patterns in data features

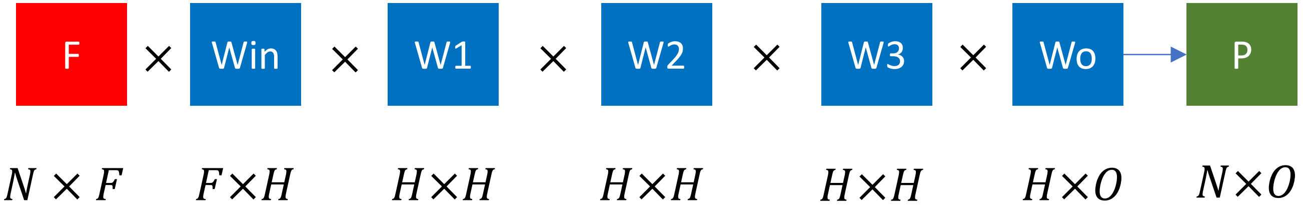


Reinforcement Learning: Have an environment which can give feedback to the models.

e.g., Control robots to do the experiments. Normally, when we don’t know how to achieve the target.

Optimization methods in different forms

The details of how the AI models are optimized on the given data or environment are quite complex. The details will be introduced in later lectures, this one is aiming to give some intuitive understanding.

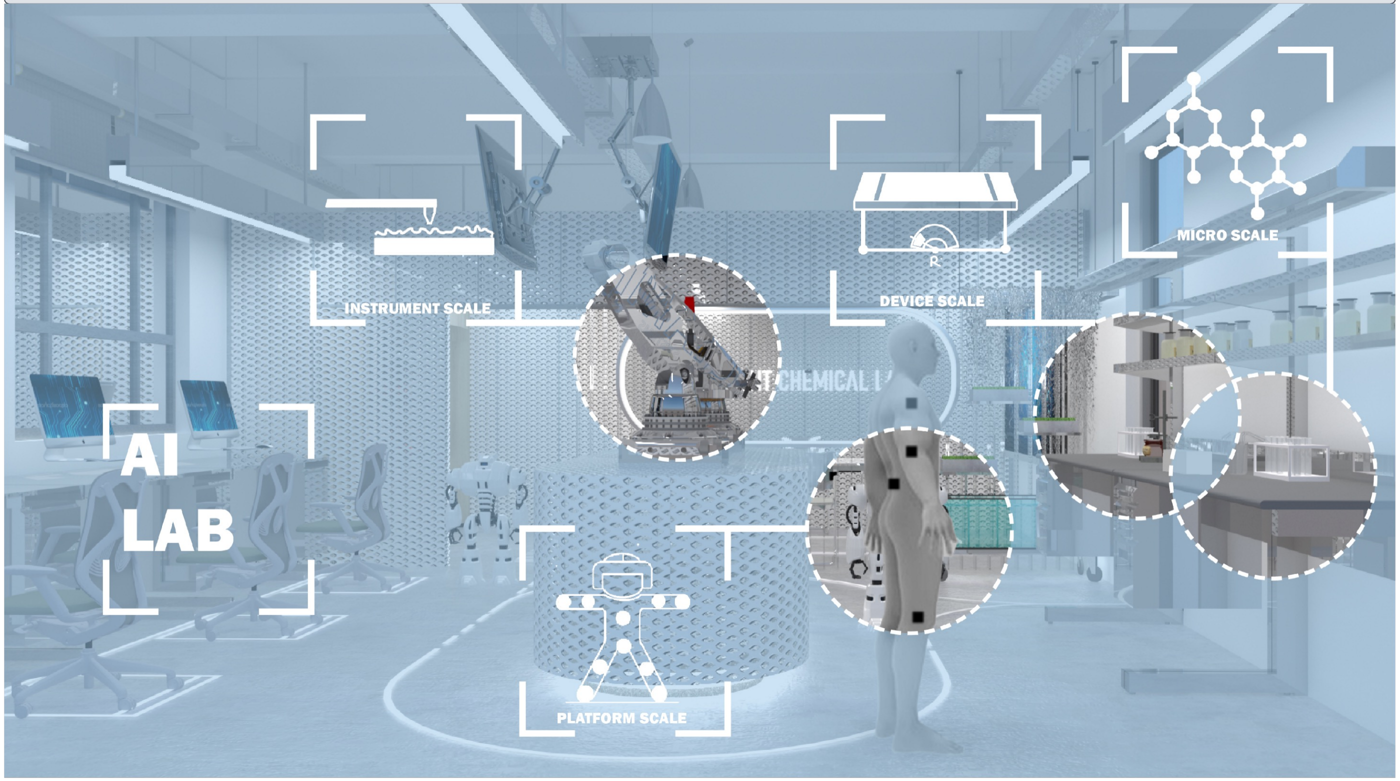


All methods have some trainable parameters as shown before. They will decide how the message transformation is done. At the beginning, the parameters are not optimized, after the transformation the predicted property will not be same as the ground truth.

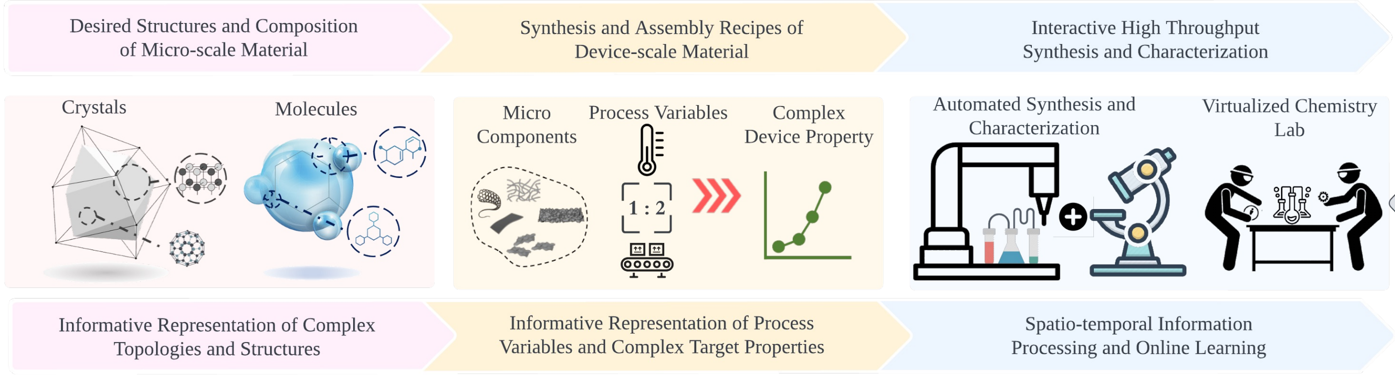
An optimization framework will be formed over the AI methods. The framework will optimize the trainable parameters. With the trained parameters the predicted properties will be much more accurate.

New Research Norm - Combine Data Science and Chemical Engineering to Develop Next Generation Research and Manufacturing Lab

New Research Norm - Combine Data Science and Chemical Engineering to Develop Next Generation Research and Manufacturing Lab



Vision



Desired Structures and Composition of Micro-scale Material

Synthesis and Assembly Recipes of Device-scale Material

Interactive High Throughput Synthesis and Characterization

Crystals

Molecules

Micro Components

Process Variables

Complex Device Property

Automated Synthesis and Characterization

Virtualized Chemistry Lab

Informative Representation of Complex Topologies and Structures

Informative Representation of Process Variables and Complex Target Properties

Spatio-temporal Information Processing and Online Learning

Multiscale Synergy

AI Next Generation Lab

The AI next generation lab can be divided into 4 different scales:

1. **Micro-scale**
2. **Device-scale**
3. **Instrument-scale**
4. **Lab-scale**

Take micro-scale as an example, it will be divided into 3 different parts with increasing complexities.

1. Representation

This part will focus on what information is contained (why them) and how they are represented in form of machine-readable language. In addition, what are the normal form of them into different ML models.

2. **Prediction**

This part will focus on commonly used and interesting forward prediction models for molecules and crystals. How the representations will be transformed will be discussed in detail.

3. **Inverse Design**

This part will focus on commonly used and interesting generative models for molecules and crystals. How a structured output can be generated will be discussed in detail.