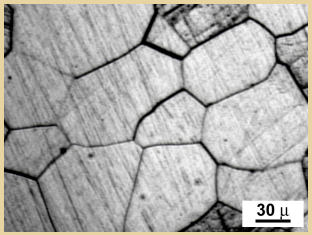
ML of Image Data

Luna Huang, Ph.D.
Materials Science and Engineering, UW

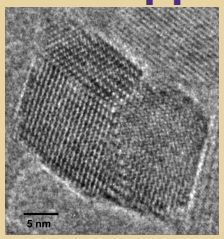
Why talk about Image Data for Materials Informatics?

- Imaging techniques are one of the main categories of characterization methods for materials study, more application can be used

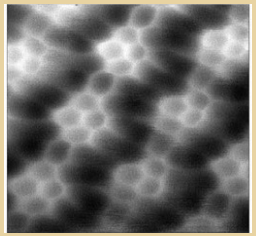


Microscope


SEM



TEM


AFM

- One of the most important applications of ML is image recognition and classification
- ML application to image data is quite unique also advanced
- The application of image ML on additive manufacturing (3D printing) is promising and important.

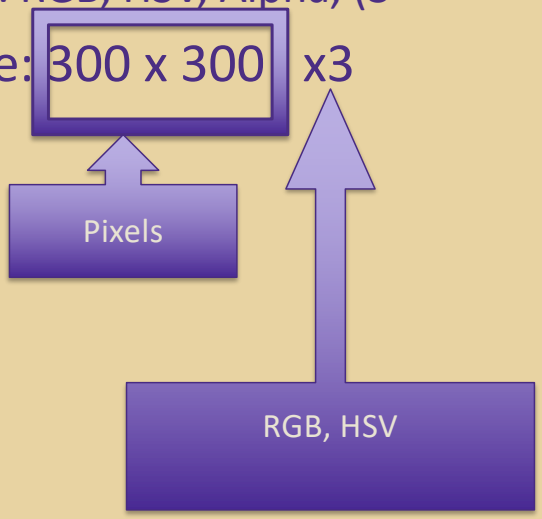
Outline

- Simplified introduction of Convolutional Neural Networks (CNN)
- Concepts that are important for image ML
  - Input image augmentation
  - Convolution
  - Pooling
- Example of using CNN for crystal study: using CNN to assist in evaluating protein crystallization—expediting the process and increasing accuracy of sorting protein crystallization images for biological study (disease, DNA, RNA…. Etc)

Image data

- All digital images are multi-dimensional arrays.
  - First two dimensions: (x,y) **position** on the image
  - Third dimension: describe the **channels** for the image, eg. RGB, HSV, Alpha, (3 or 4 channels)

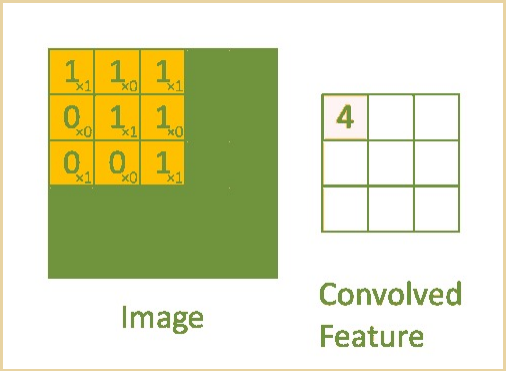
Data Size of an image: 300 x 300 x3



RGB: Red, Green, Blue
HSV: Hue, Saturation, Value
Alpha: Transparency, may be added as the 4th channel

A simple introduction of CNN (Convolutional Neural Networks)

- Convolutional Neural Networks are very similar to ordinary Neural Networks from the previous chapter: they are made up of neurons that have learnable weights and biases.
  - They apply Filters defined by their Convolution Kernel
  - Filters help extract features from the images: edges, corners, blobs, etc…



10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0
10	10	10	0	0	0

*

1	0	-1
1	0	-1
1	0	-1

=

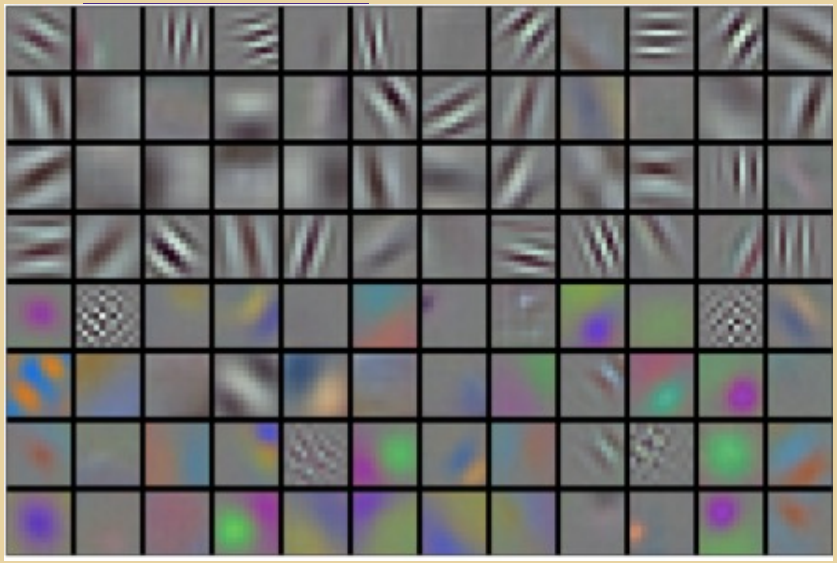
0	30	30	0
0	30	30	0
0	30	30	0
0	30	30	0

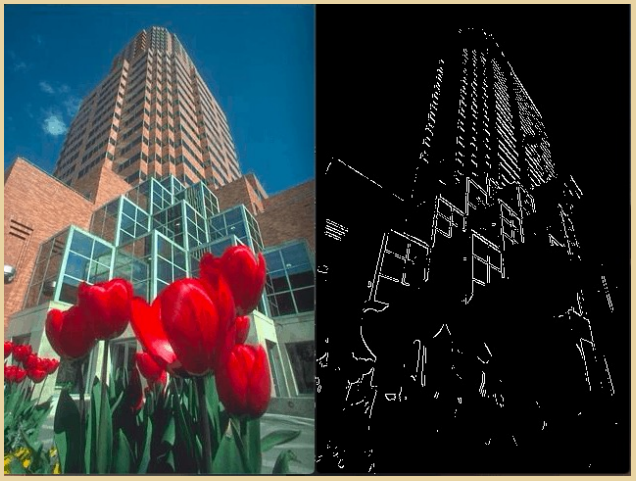
3*3 matrix times 3*3 matrix

UNIVERSITY of WASHINGTON

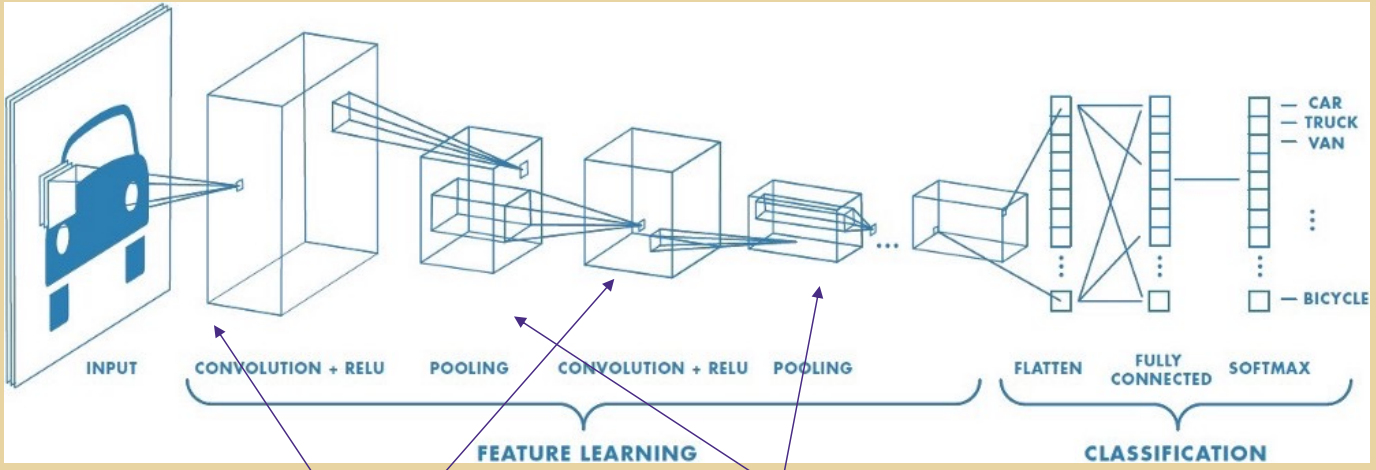
CNN Filters, and why use the filters

Filters and the outcome of the filter

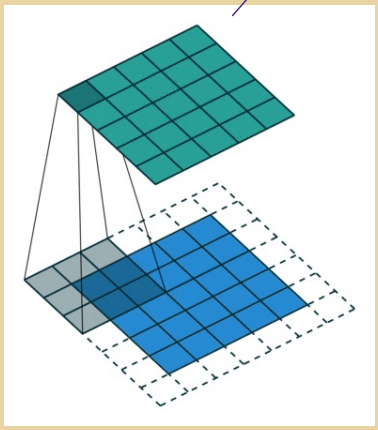




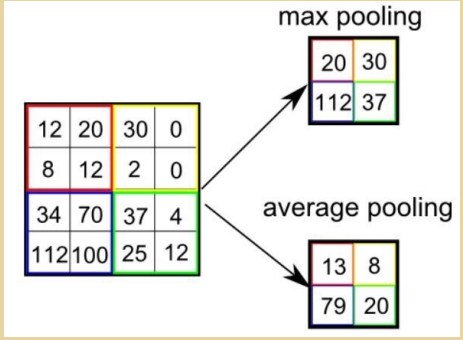
- **Why normal Neural Networks won't work? Because images size (datasize) is really big, so other NN not applicable, 800x800 will need 6400000 weights,**
- **CNN, filter is shared, each filter only have one set of weights (if filter is 3x3 , has 9 weights) no matter how big the image is, weights number = size of filters X9**



Each filter processing the image will generate one channel on the next layer, increasing the channels of the data at that layer



Pooling decreases the dimension of each channel.



As a result, the image data which has high dimension, thin depth turns into a data with low dimension, but high depth.

Important points of Image ML implementation

- **Transfer Learning: Pre-trained neural networks with validated models, such as Densenet-121, VGG19, Resnet50...etc. (reduce the training cost, a good initialization)**
- **Augmentation of input. (why and how)**
- **Methods to avoid over-fitting (drop-out layers, dataset balancing, etc, batch size and epoch number selection)**
- **Definition:**
  - **Epoch:** One Epoch is when an ENTIRE dataset is passed forward and backward through the neural network only ONCE.
  - **batch size:** how many images present in a single batch,
  - **drop-out layers,** De-active certain percentage of randomly selected neurons to avoid overfitting. Drop out of certain amounts of neurons, and use other neurons to represent the features
  - **Dataset Balancing**, make sure the training sets represent the future candidates for input

Example

Using CNN to classify protein crystal images to expedite protein molecule research

- **Image ML application in MSE is not very common, main reasons are lack of centralized database and awareness: not lack of Data!**
- **Alphafold! Big news in AI and Biology study on Nov.30th, 2020: https://www.nature.com/articles/d41586-020-03348-4**


nature

View all Nature Research

Explore our content · Journal information · Subscribe

nature › news › article

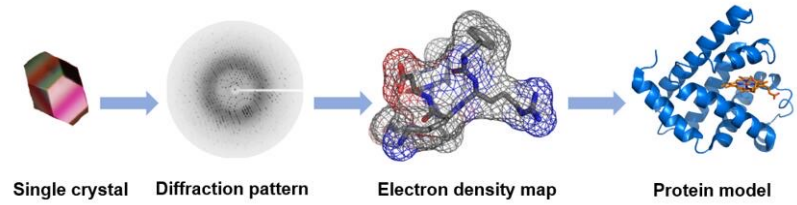
NEWS · 30 NOVEMBER 2020

'It will change everything': DeepMind's AI makes gigantic leap in solving protein structures

Google's deep-learning program for determining the 3D shapes of proteins stands to transform biology, say scientists.

AstraZeneca

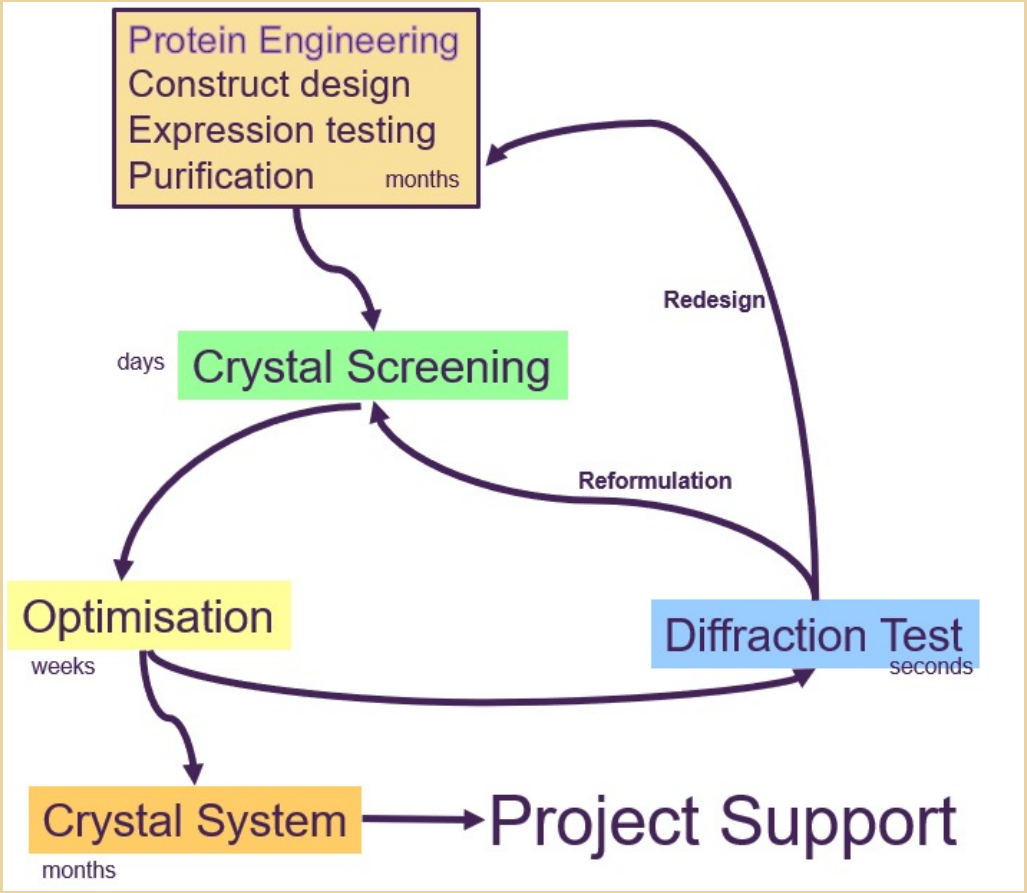
What is Protein Crystallization



- **Protein crystallization** is the process of formation of a regular array of individual protein molecules stabilized by crystal contacts. If the crystal is sufficiently ordered, it will diffract
- **Hard to form**, proteins evolved to be recalcitrant to crystallization: they only carry out their biological function when non-crystallized
- Crystallization is essential for structure characterization of proteins: XRD Diffraction requires crystallization
  - About 90% of all known protein structures have been obtained through x-ray crystallography, which successfully crystalized

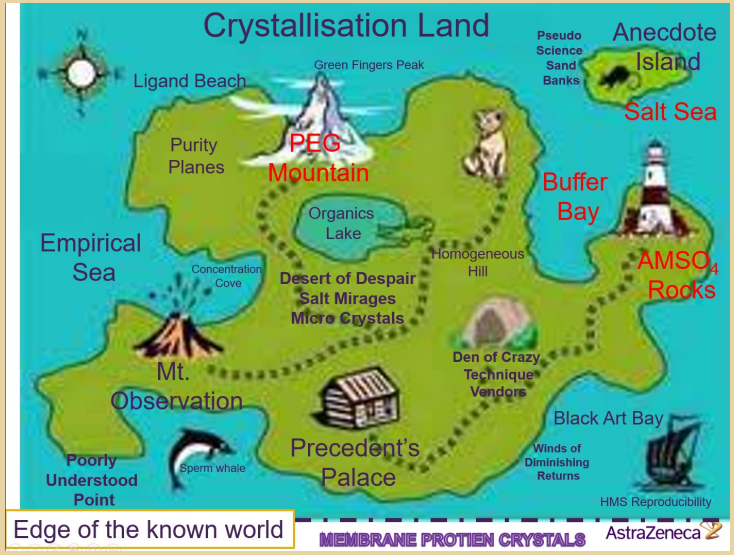
Batch Cocktail experiment → Try to grow single crystal for XRD diffraction → XRD diffraction results → obtain structure of the crystal through diffraction pattern → Protein Model → Understand the function of the protein → Cure or vaccine for the diseases

The Industrial Crystallisation Process



Thousands of different experiments that cost up to 2 months before a single (if any) crystal forms.

Crystallisation Screening

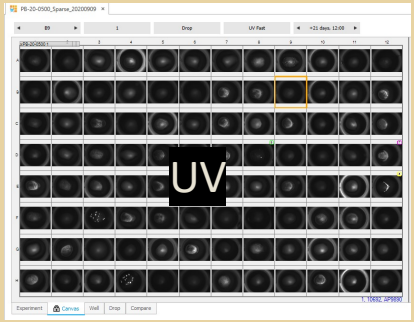


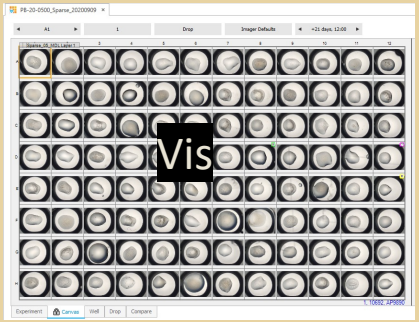
All we have is:

- The sequence of the protein in solution and associated calculable & measured values

- A list of chemicals used in the experiments ~30 chemicals in 96 combinations

- An image of the outcome





96 drop experiment:

- 9 visible light inspections
- 6 UV light inspections
- 1440 outcomes to judge

Classification of crystallization results can be subjective - crystallographers have been shown to score their own results higher than those of others.

The table shows percentage agreement with the average score for 16 individuals* :

Score		Crystals (51)	Needles (87)	Micro-crystals (78)	Phase separation (113)	Precipitate (356)	Denatured protein (87)	Clear (435)
6	Crystals	84.7	13.7	1.2	0.0	0.0	0.4	0.0
5	Needles	11.2	68.5	17.7	1.5	0.1	0.3	0.4
4	Micro-crystals	3.3	29.0	50.4	12.6	1.8	1.3	1.1
3	Phase separation	0.6	2.0	18.9	46.7	23.5	6.5	1.4
2	Precipitate	0.0	0.1	1.9	20.6	59.3	14.4	3.5
1	Denatured protein	0.1	0.0	8.5	18.1	39.5	32.5	1.3
0	Empty	0.0	0.0	0.1	0.9	1.8	5.4	91.6

*crystallographers in the York Structural Biology Laboratory

FORMULATRIX®

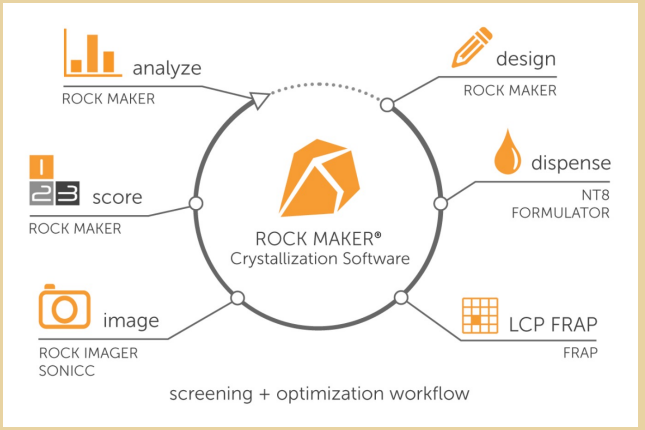
Defining **reliability** as the percentage of images scored exactly the same by two human classifiers and
consistency as the percentage of images scored exactly the same twice by the same human classifier,

Tests performed by researchers at Formulatrix, on 50,000 images showed **87%** reliability and just **74%** consistency.

Image scoring by eye is monotonous and time-consuming …

…and therefore prone to human

AI enabled automated image analysis can be faster, more consistent and reliable.



MARCO dataset

- MARCO, Machine Recognition of Crystallization Outcomes, is a dataset made available by the university at Buffalo that has 493,214 protein crystallization images collected from several well-known institutions.
  - you can consider this as the ImageNet (http://www.image-net.org/) of protein crystallization image database.
  - Provide a training and validating dataset to improve the performance of ML methods in protein crystallization classification.
  - Images are taken by different institutes with varying quality and standard---a lot data cleaning work needed.
- Example and rundown of Image data ML using MARCO dataset: