Rapid Analysis of X-ray Images for Crystalline Materials Using Convolutional Neural Networks.

Eric Chan 19/5/2023

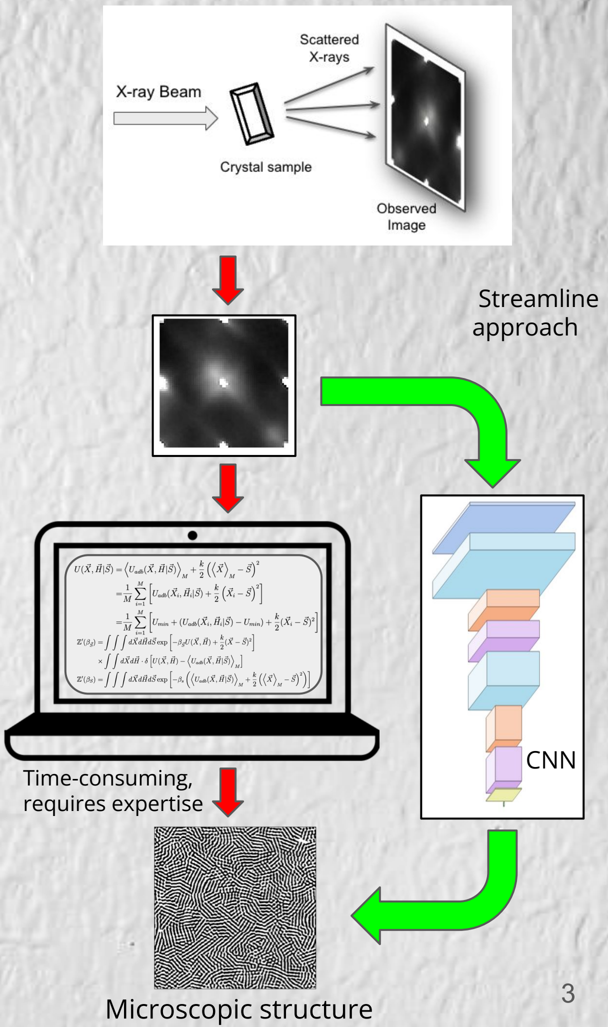
Job Seeker (Data Science, ML, AI)

Overview

- ❖ About me
- ❖ Project context
- ❖ Business perspective
- ❖ Design and workflow, technical aspects
- ❖ Model training and performance
- ❖ Conclusions and next steps

Project context

- ❖ **Business context:**
  - ➢ Crystal based Materials - develop, manufacture, quality control.
- ❖ **Background:**
  - ➢ Understand structure/property relationships
  - ➢ X-ray images provide detailed microscopic view of structure
- ❖ **Problem Statement:**
  - ➢ Interpretation of X-ray images is not-fully automated.
  - ➢ Involves significant physical modeling trial and error
  - ➢ Analysis requires a high level of expertise.
- ❖ **Goal:**
  - ➢ Enable rapid interpretation of X-ray images using CNN.



Convolutional Neural Network (CNN): Type of deep learning model specifically designed for processing and analyzing image data.

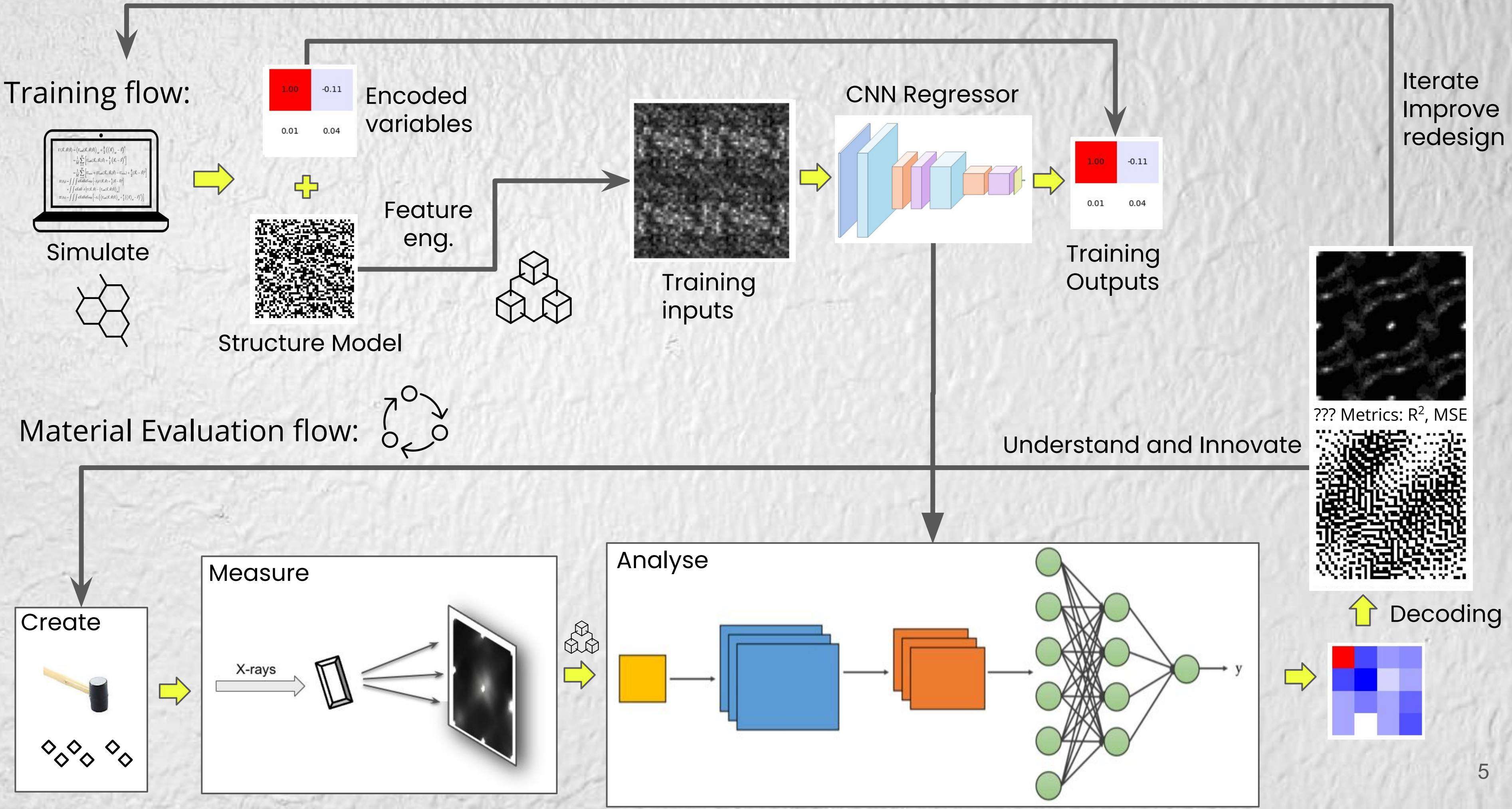
Stakeholders: Researchers, scientists, engineers, manufacturers.

***Domains*: [Companies]**

- *Organic semiconductors:* GE, Sony, Samsung, LG, Sigma-Aldrich.
- *Energy storage:* ABB LTD, Eos Energy Enterprises, BVSPC, Tesvolt.
- *Pharmaceuticals:* Pfizer, Merck, Eli Lilly.
- *Ceramics:* Kyocera, Corning Inc., Murata, CoorsTek.
- *Agrochemicals:* Bayer CropScience, Syngenta, BASF.
- *Thin film materials:* Vital Materials, Reynard Corporation, Kodak.

Business Perspective

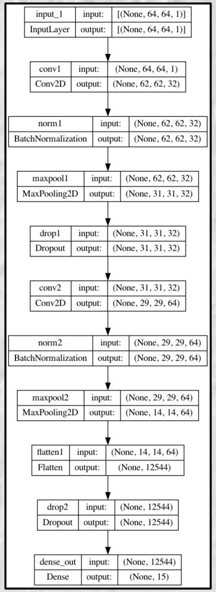
- **Business question:**
  - How to enhance the design space for profitable materials at lower cost?
- **Data science flow constraint:**
  - Shortage of available labeled X-ray data for training.
- **Cost effective Strategy:**
  - Leverage a physics-based model to simulate data of X-ray images for training/test.
- **Design and workflow:** *next slide .....*


Training flow:
Simulate
Encoded variables
1.00
-0.11
0.01
0.04
Structure Model
Feature eng.
Training inputs
CNN Regressor
Training Outputs
Iterate Improve redesign
??? Metrics: R^2, MSE
Understand and Innovate
Material Evaluation flow:
Create
Measure
X-rays
Analyse
y
Decoding

Technical Overview

- ❖ Model Stages:
  - ➢ 1. Decide and train.
  - ➢ 2. Evaluate on simulated X-ray images
  - ➢ 3. Evaluate on experimental X-ray images
- ❖ CNN:
  - ➢ Evaluated small and large architecture
    - ▪ Inputs: 64x64 pixel X-ray Images
    - ▪ Outputs: encoding variables, Dimensionalities = 3, 8, 15.
  - ➢ Different CNN for each resolution of training data

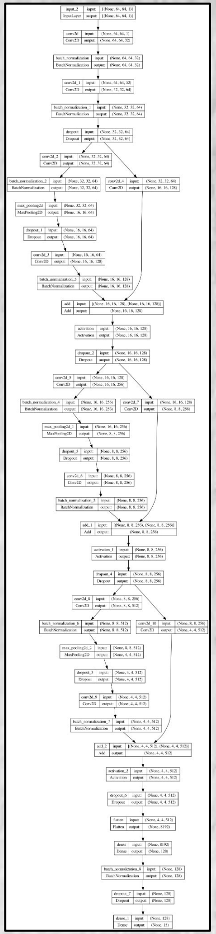
Small CNN



Total params: 207,375

Activations: ReLU (output is linear)

Large CNN

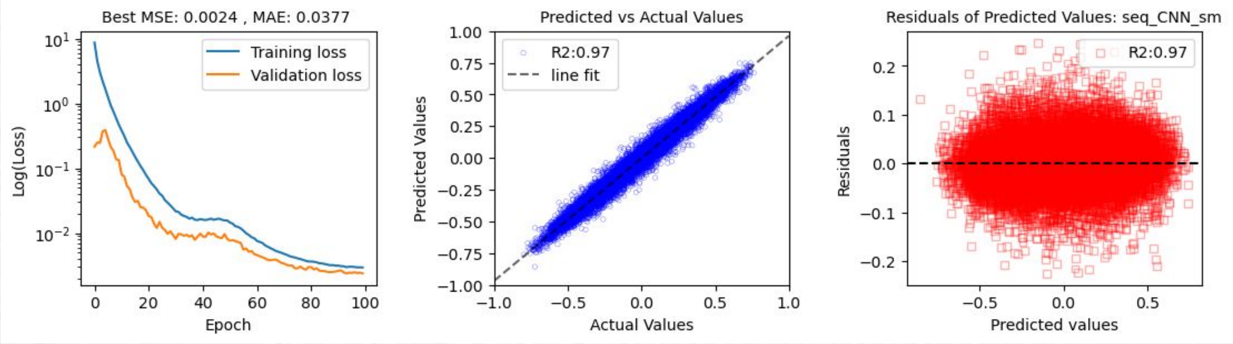


Total params: 7,162,447

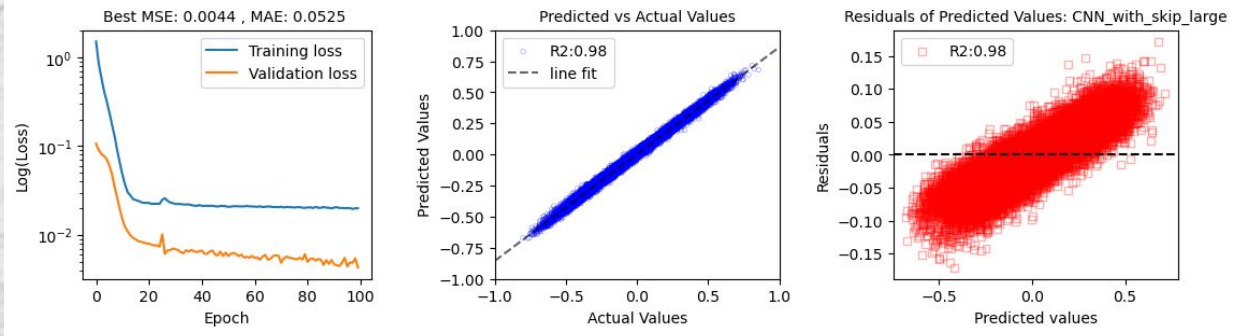
CNN Training and Validation.

- Datasets 5000x64x64 or 1000x64x64
- Batch size: 32, Epochs: 100
- Small CNN arch. performed better during evaluation with incremental training

Small CNN Arch.



Large CNN Arch.

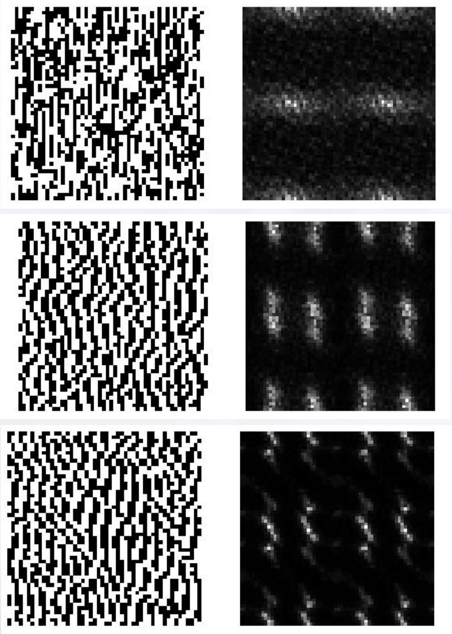


Validation metrics

Dim.	Arch.	Data points	Loss (MSE)	R^2
3	small	5000	6.5E-4	0.997
8	small	5000	1.9E-3	0.986
15	small	5000	2.4E-3	0.974
15	small	10000	3.0E-4	0.997
15	large	10000	2.2E-3	0.976

CNN performance: example with simulated holdout data.

Simulated(D:15) X-ray Image input to CNNs



D:3

D:8

D:15

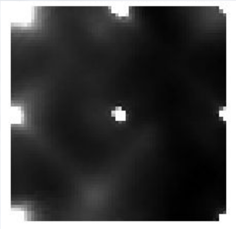
Decoded Outputs from CNN predictions.

CNN output Dimension	R^2 (Images)	MSE(Images)
3	-0.034	0.309
8	0.113	0.265
15	**0.675**	**0.097**

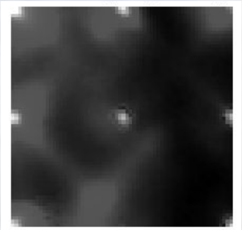
Performance at interpreting real X-ray data.

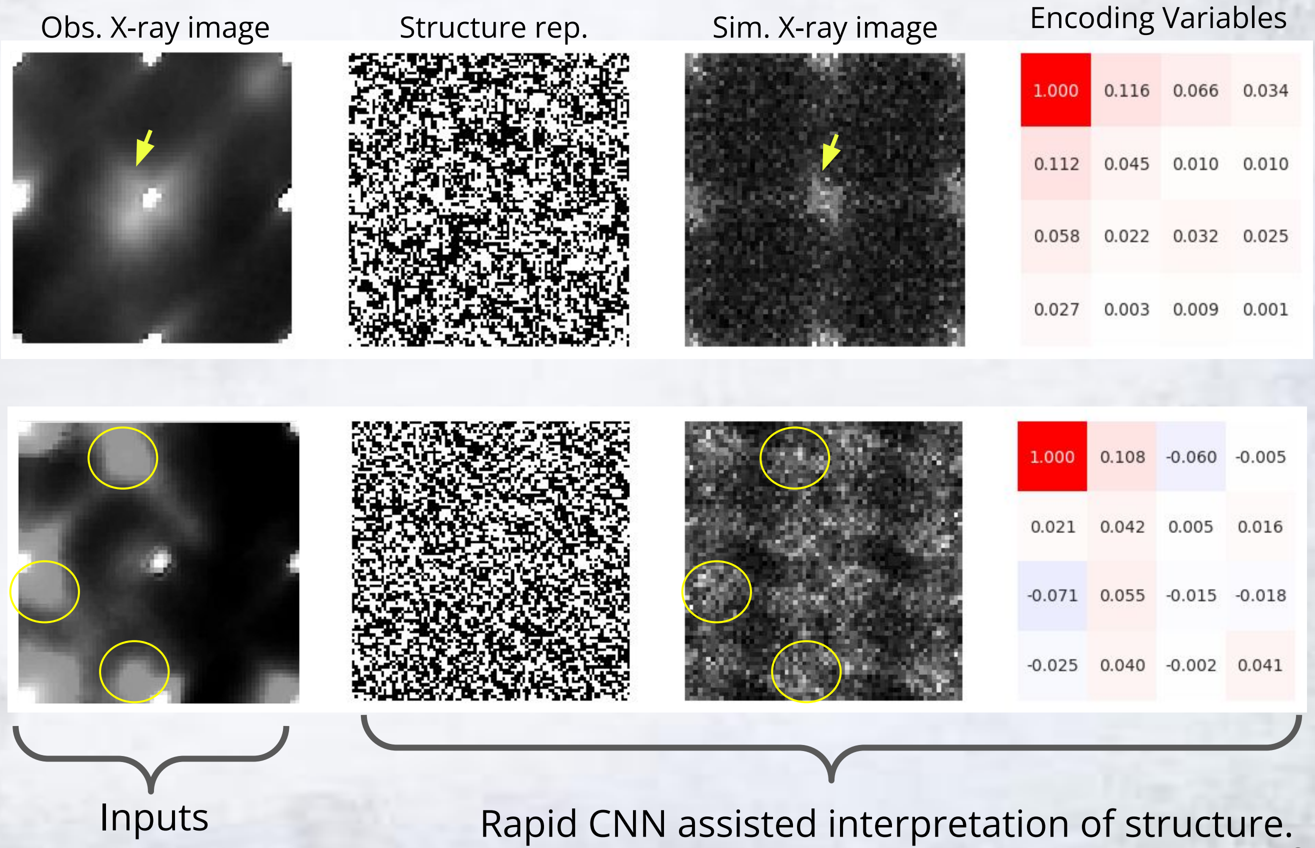
Image reconstruction, data compression, smoothing, normalization, re-scaling, (optional: Top-hat filter, image restoration in-painting).

Raw data



Corrected





1.000	0.116	0.066	0.034
0.112	0.045	0.010	0.010
0.058	0.022	0.032	0.025
0.027	0.003	0.009	0.001

1.000	0.108	-0.060	-0.005
0.021	0.042	0.005	0.016
-0.071	0.055	-0.015	-0.018
-0.025	0.040	-0.002	0.041

Summary, conclusions and next steps

- Summary:
  - X-ray imaging interpretation:
    - not automated
    - Requires high level analysis and expertise.
    - Lack of labeled data suitable for training.
  - **Goal:** Evaluate if CNN can be designed and integrated for streamlining.
- Conclusions:
  - Successful proof of concept.
    - CNN can be trained on simulated data to interpret real X-ray image.
  - Business Case Overview. **(***see next slide*)
- Next steps:
  - Future case study:
    - Enhance resolution
    - Incorporate molecular modeling.
  - Can an AGI system be implemented (VAE, cGANS, Deep Q learning)?

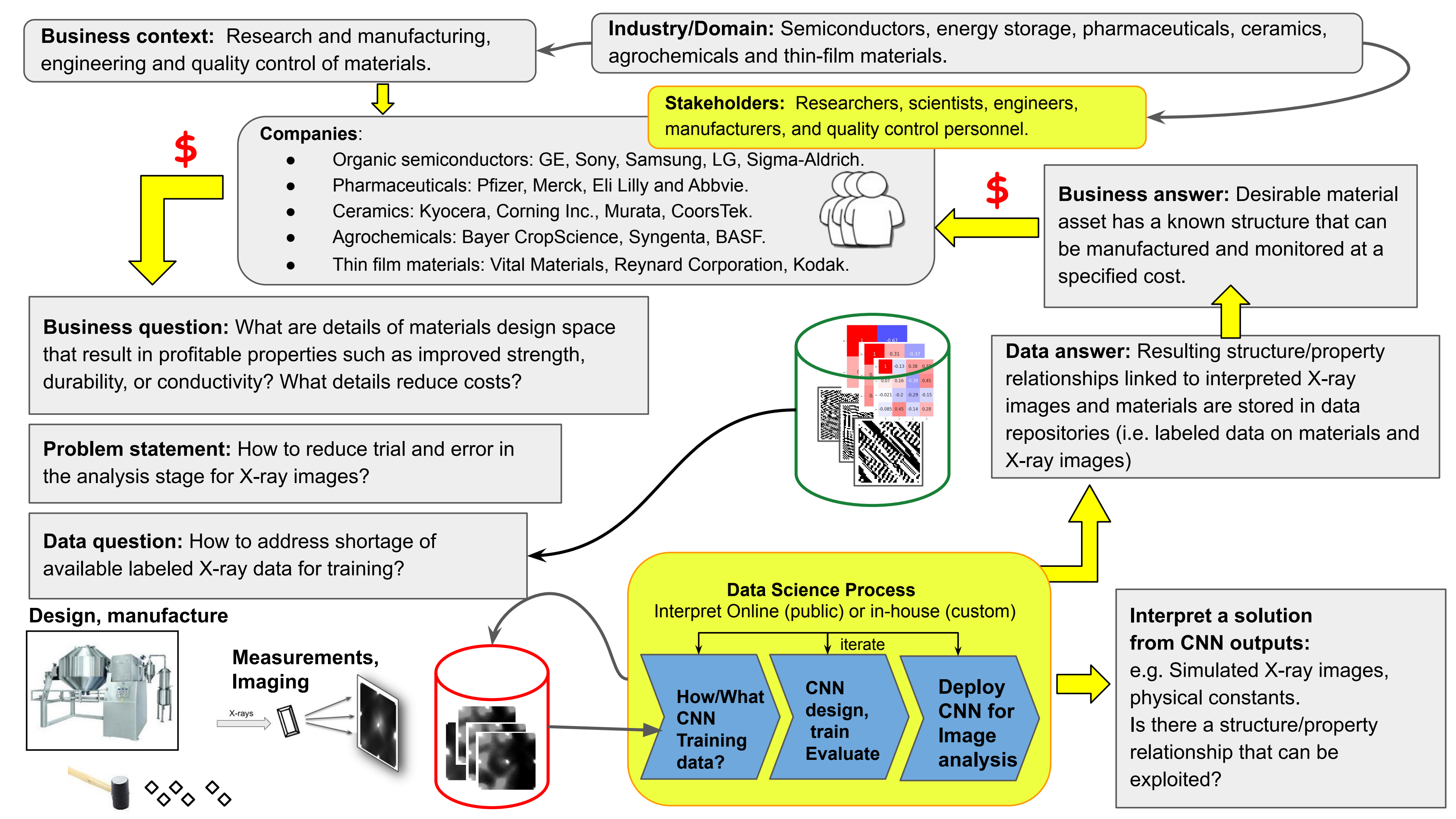
Business Case Overview

Resource	Current X-ray analysis pipeline			CNN enabled X-ray analysis pipeline		
	Units/year	Cost/profit estimate	Total	Units/year	Cost/profit estimate	Total
Materials Design Portfolio	100	$2,000,000	$200,000,000	150	$2,000,000	$300,000,000
Raw X-ray Data collection	40	-$20,000	-$800,000	200	-$20,000	-$4,000,000
Simulated Data collection	0	$0	$0	20000	$0	$2
X-ray Image analysis	5	$50,000	$250,000	105	$50,000	$5,250,000
SME workers	5	-$200,000	-$1,000,000	2	-$200,000	-$400,000
non-SME workers	10	-$60,000	-$600,000	15	-$60,000	-$900,000
Net Income			$197,850,000			$299,950,002

Assumptions:

- Instrument upkeep cost is unchanged because it is never fully utilized.
- Increasing X-ray image analysis by factor of 1 results in Materials design enhancement of 0.5
- Cost of simulated data is negligible wrt. Rest of the portfolio
- Tradeoff. Less SME and More non-SME

Estimated net profit increase is of the order of ~50%


Business context: Research and manufacturing, engineering and quality control of materials.
Industry/Domain: Semiconductors, energy storage, pharmaceuticals, ceramics, agrochemicals and thin-film materials.
Stakeholders: Researchers, scientists, engineers, manufacturers, and quality control personnel.
$
Companies:
Organic semiconductors: GE, Sony, Samsung, LG, Sigma-Aldrich.
Pharmaceuticals: Pfizer, Merck, Eli Lilly and Abbvie.
Ceramics: Kyocera, Corning Inc., Murata, CoorsTek.
Agrochemicals: Bayer CropScience, Syngenta, BASF.
Thin film materials: Vital Materials, Reynard Corporation, Kodak.
$
Business answer: Desirable material asset has a known structure that can be manufactured and monitored at a specified cost.
Business question: What are details of materials design space that result in profitable properties such as improved strength, durability, or conductivity? What details reduce costs?
Data answer: Resulting structure/property relationships linked to interpreted X-ray images and materials are stored in data repositories (i.e. labeled data on materials and X-ray images)
Problem statement: How to reduce trial and error in the analysis stage for X-ray images?
Data question: How to address shortage of available labeled X-ray data for training?
Design, manufacture
Measurements, Imaging
X-rays
Data Science Process
Interpret Online (public) or in-house (custom)
iterate
How/What CNN Training data?
CNN design, train Evaluate
Deploy CNN for Image analysis
Interpret a solution from CNN outputs:
e.g. Simulated X-ray images, physical constants.
Is there a structure/property relationship that can be exploited?

Thank you !....

Questions ???

Appendix

References

https://iopscience.iop.org/article/10.1088/2632-2153/acab4c

https://doi.org/10.1063/5.0013065

https://doi.org/10.1038/s41598-020-62484-z
https://doi.org/10.1063/5.0014725

Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett, “Deep Learning Techniques for Inverse Problems in Imaging.” IEEE JOURNAL ON SELECTED AREAS IN INFORMATION THEORY, VOL. 1, NO. 1, MAY 2020

https://doi.org/10.1038/s41524-021-00644-z

https://www.nature.com/articles/s41598-018-34525-1

Chan, E., On the use of molecular dynamics simulation to calculate X-ray thermal diffuse scattering from molecular crystals. Journal of Applied Crystallography 2015, 48 (5), 1420-1428.

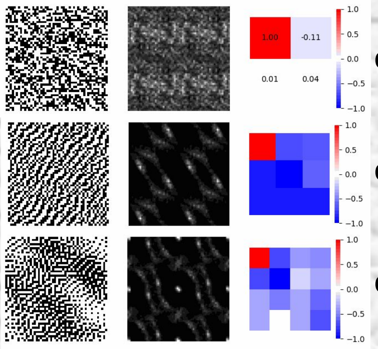
Chan, E. J.; Welberry, T. R.; Goossens, D. J.; Heerdegen, A. P., A Diffuse Scattering Study of Aspirin Forms I and II. Acta Crystallographica Section B 2010, 66, 696-707.

Chan, E. J.; Welberry, T. R.; Goossens, D. J.; Heerdegen, A. P., A refinement strategy for Monte Carlo modelling of diffuse scattering from molecular crystal systems. j. Appl. Cryst. 2010, (43), 913-915.

Heerdegen, A. P. (2000). Diffuse X-ray Scattering from an Optically Anomalous Material 1,5- dichloro-2,3-dinitrobenzene. Ph.D. thesis.

EDA

Outputs: encoding variables,
Dimensionalities = 3, 8, 15.
Correlation function span (cfs)= 2, 3, 4.

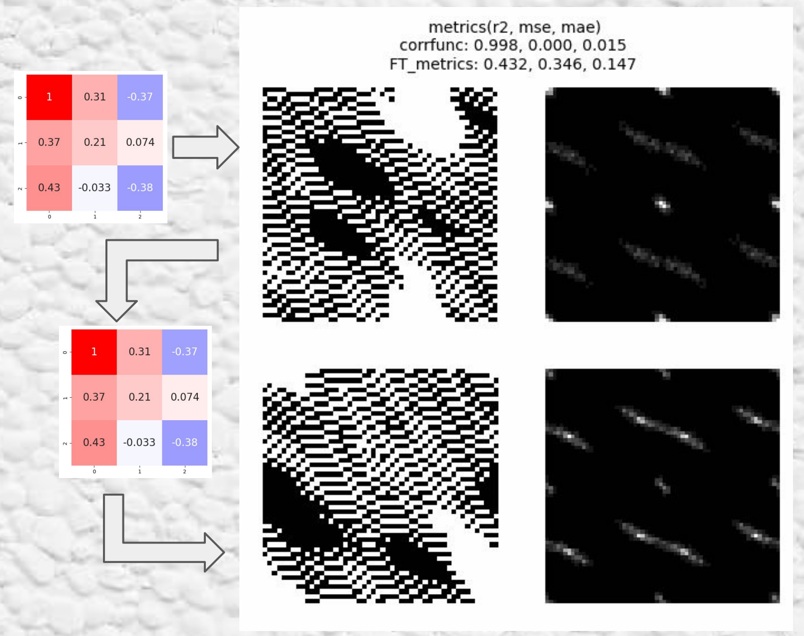


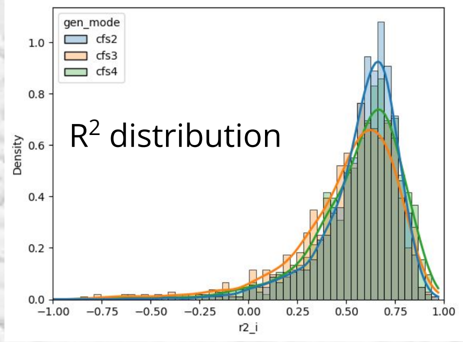
cfs:2

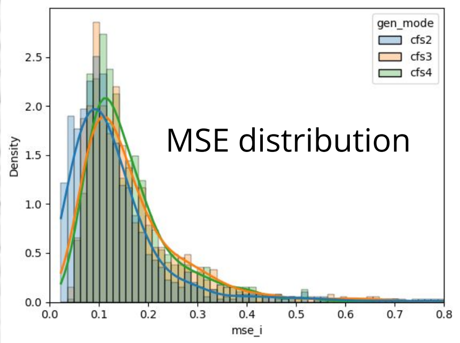
cfs:3

cfs:4

Encoding error due to statistical noise in simulated training data:

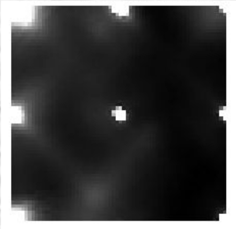




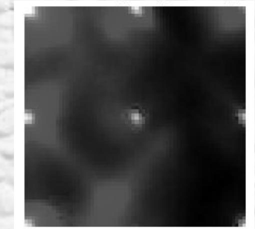


Raw X-ray data correction:

before



after



*cfs (Correlation Function Span)

Overview

- ❏ **Business use cases:** Materials science, manufacturing, Mining, medical imaging and diagnoses, Education.
- ❏ In most modern analytics for materials or biological research analytical data requires substantial modeling for practical interpretation.
- ❏ We are exploring the application of CNN for a type of these Inversion problems(eg. Deconvolution, halftoning, super-resolution).

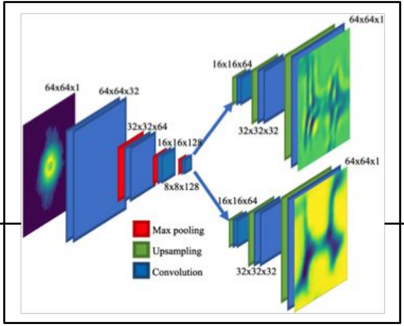
Diffraction imaging:

- ❏ It can be very difficult to identify the physical rules associated with how atoms decide to be structured in a crystal lattice. One way to do this is to study diffraction patterns. Often requires high level expertise (many trails) and costly technical/computing resources.
- ❏ We explore a basic formulation for interpreting micro-structure from diffuse X-ray diffraction of crystalline materials. I.e. (details of crystal growth and atomic ordering on lattices)
- ❏ In this simple exercise we are training CNNs on theoretical diffraction data and investigating how well it is able to interpret sections of observed data.

Final Objective: Deploy several CNNs available publicly online for a rapid general interpretation of diffraction image data for Microstructure. - *Very time saving option available to a wide audience (not just SMEs).*

Benefits:

- ❏ Rapid interpretation of possible microstructure without tedious modeling or preconditions for further models.
- ❏ Instantiate the benchmark for model complexity and CNN performance for this type of problem.
- ❏ Build and showcase skills for CNN utility for computer vision and imaging as well as solving types of inverse problems through accelerating pre-conditioning stages and reducing trial and error.



SCIENTIFIC REPORTS
natureresearch

OPEN **Multi-resolution convolutional neural networks for inverse problems**

Feng Wang[1,2*], Alberto Eljarrat[2], Johannes Müller[2], Trond R. Henninen[1], Rolf Erni[1] & Christoph T. Koch[2]

SCIENTIFIC REPORTS

OPEN **Real-time coherent diffraction inversion using deep generative networks**

Received: 12 June 2018
Accepted: 19 October 2018
Published online: 08 November 2018

Mathew J. Cherukara[1,3], Youssef S. G. Nashed[2] & Ross J. Harder[1]

Phase retrieval, or the process of recovering phase information in reciprocal space to reconstruct images from measured intensity alone, is the underlying basis to a variety of imaging applications including

IEEE JOURNAL ON SELECTED AREAS IN INFORMATION THEORY, VOL. 1, NO. 1, MAY 2020 39

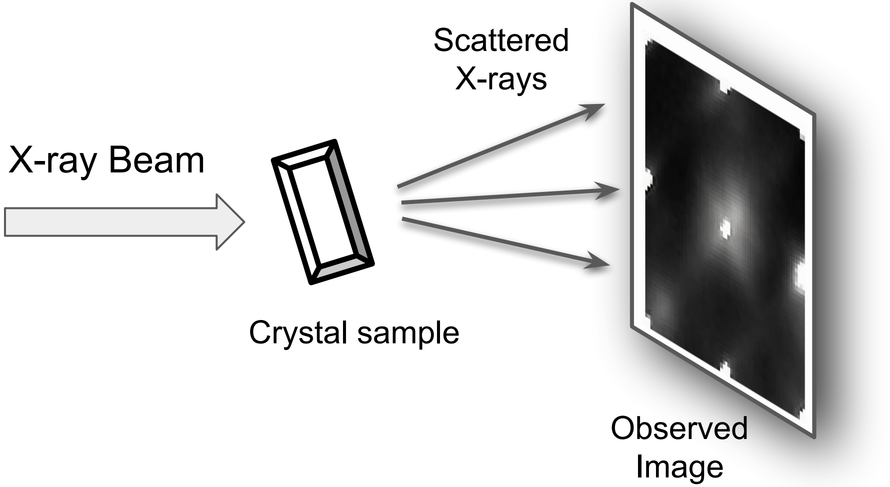
Deep Learning Techniques for Inverse Problems in Imaging

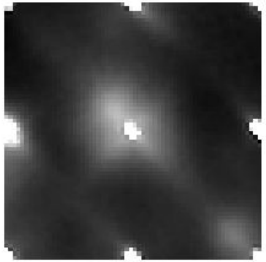
Gregory Ongie, Ajil Jalal, Christopher A. Metzler, Richard G. Baraniuk, Alexandros G. Dimakis, and Rebecca Willett

ARTICLE OPEN Check for updates

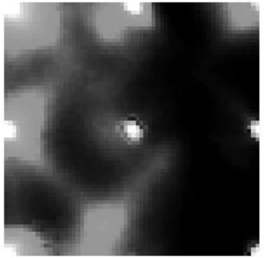
Three-dimensional coherent X-ray diffraction imaging via deep convolutional neural networks

Longlong Wu[1,2], Shinjae Yoo[1], Ana F. Suzana[2], Tadesse A. Assefa[2,3], Jiecheng Diao[4], Ross J. Harder[5], Wonsuk Cha[5] and Ian K. Robinson[2,4]





Crystal Structure w/o desirable properties



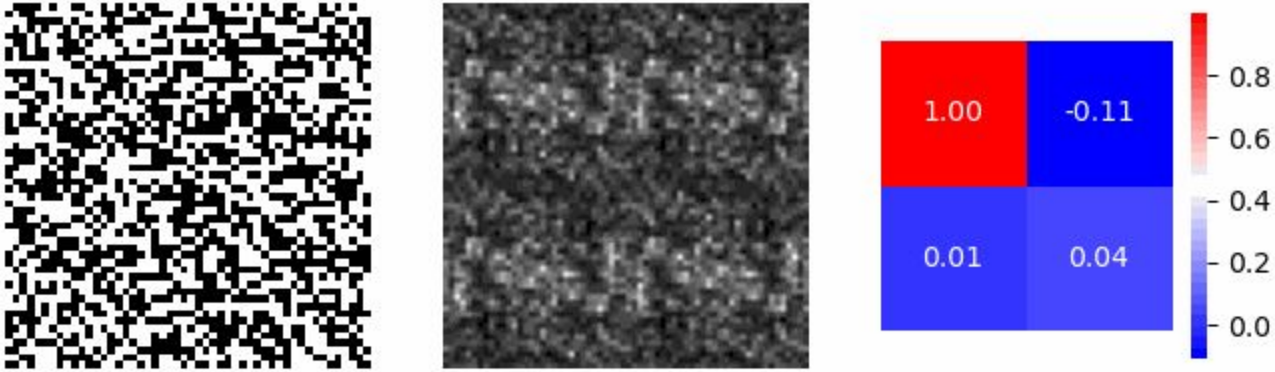
Crystal Structure has desirable properties

What are the details about the atomic or molecular structure that result in desirable properties?

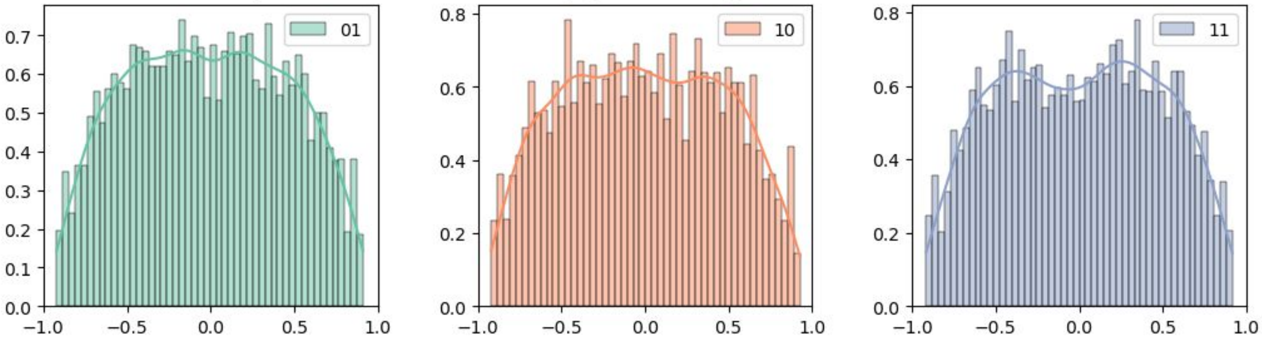
Understanding this relationship is important for developing new materials with desirable properties, such as improved strength, durability, or conductivity. This knowledge can be used to tailor the atomic or molecular structure of materials to achieve specific properties, which can be beneficial for various industries,

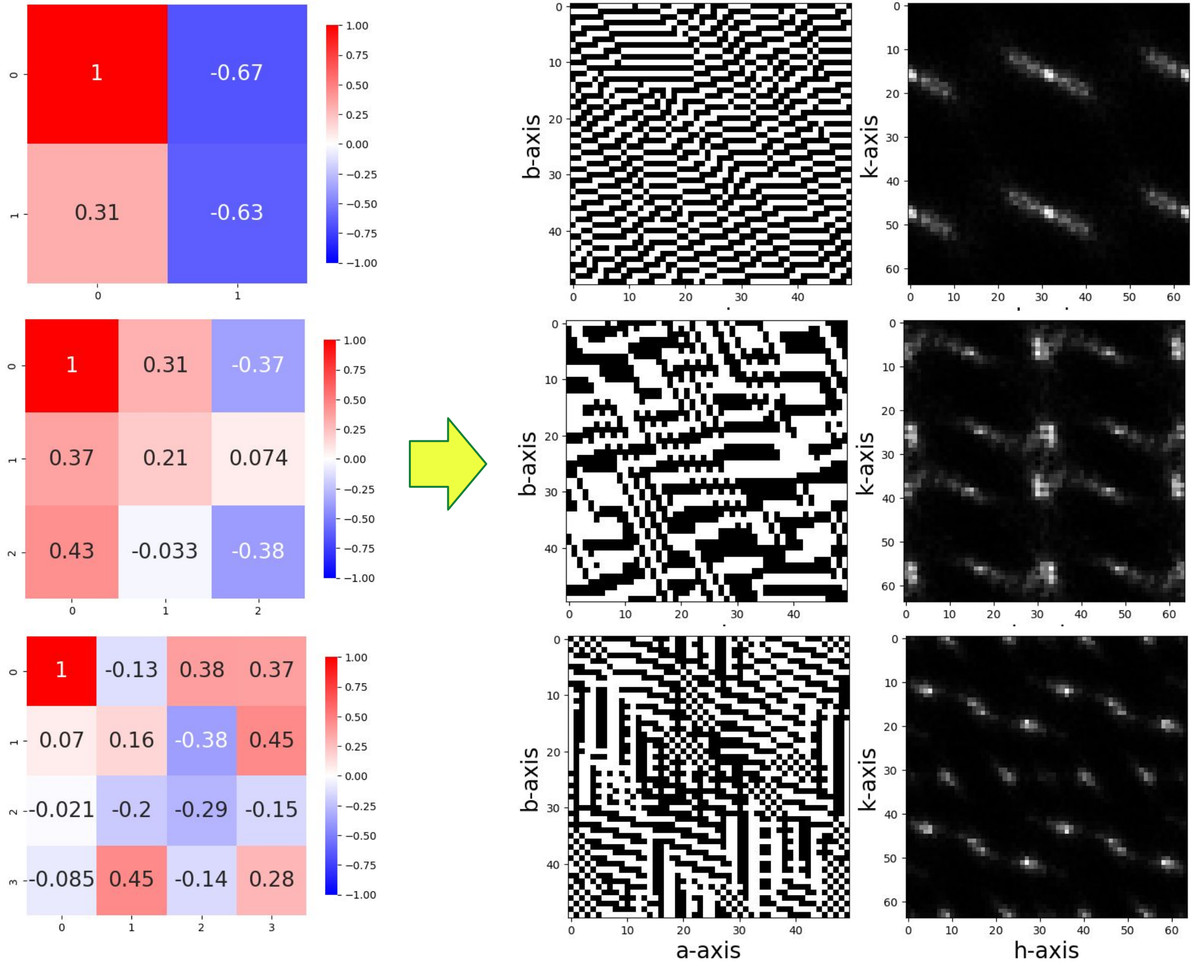
Desirable properties of a good organic semiconductor include low cost, light weight, mechanical flexibility, easy processing, and abundant availability compared to inorganic materials. Organic semiconductors are generally low cost and can be easily processed under a less controlled environment compared to inorganic semiconductors. They should also have good electrical conductivity, high charge carrier mobility, and high stability under ambient conditions. In addition, they should have good solubility in common solvents, high thermal stability, and good film-forming properties. These properties make organic semiconductors attractive for use in optoelectronic devices such as organic light-emitting diodes (OLEDs), organic solar cells (OSCs), and organic field-effect transistors (OFETs)

CFS2 vector L2 distance from origin 0.118



Density distributions of the randomly generated target CFS2 variables used for CNN fit.



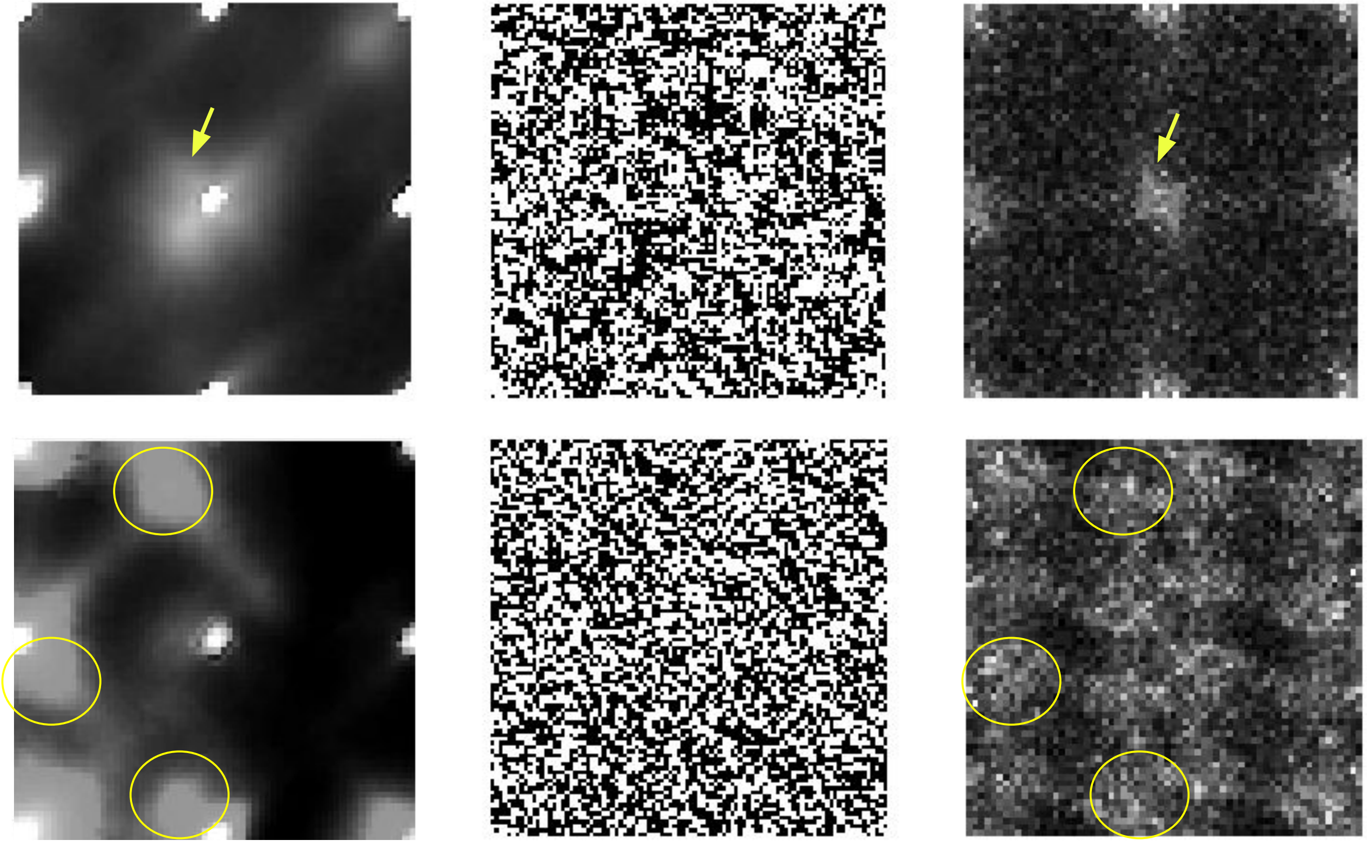


CFS Encoding

2D lattice representation

Feature Engineering
(Fourier Transform)

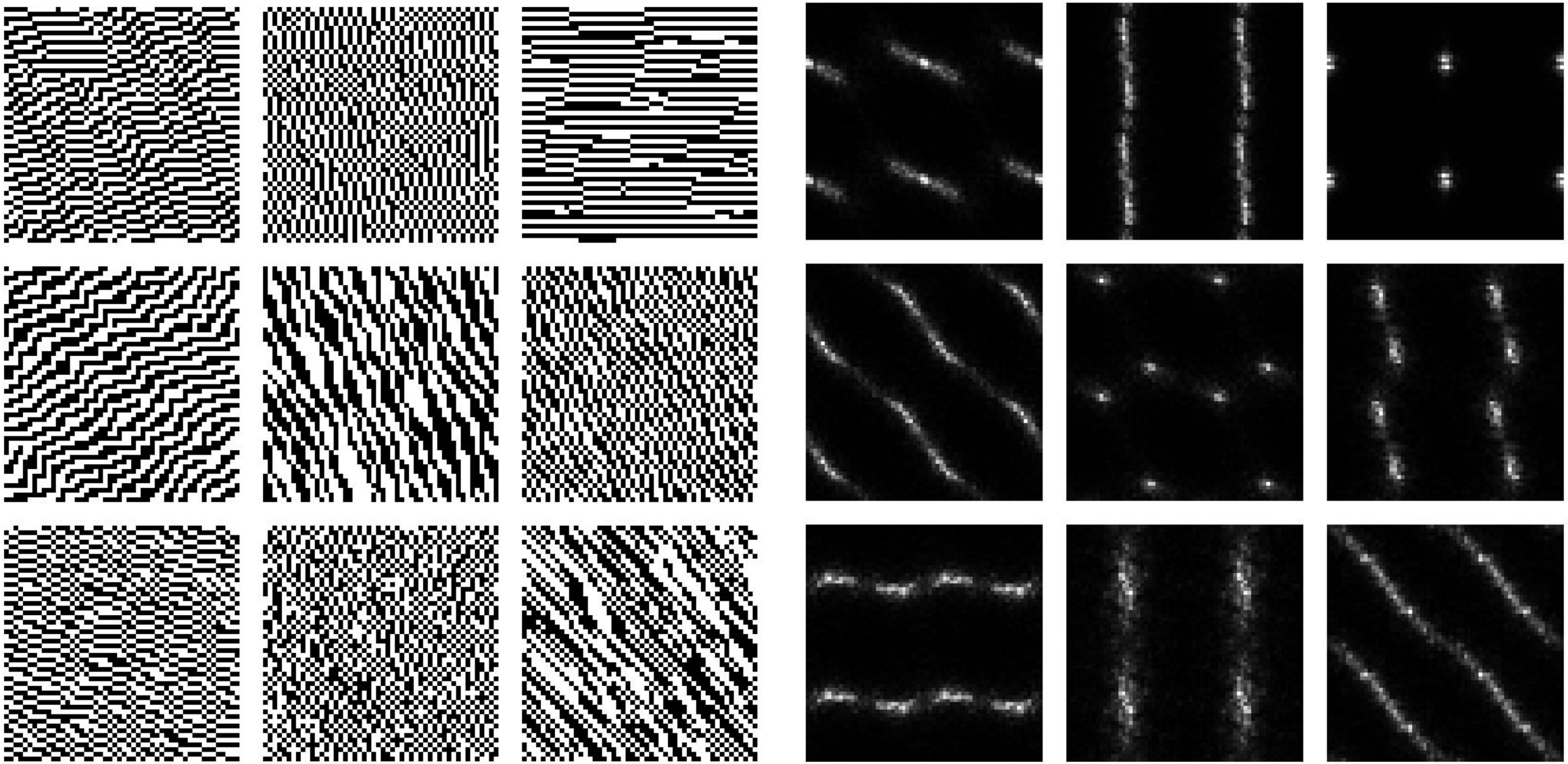
Different CNN are t
because the output
different orders of e
(CFS type)



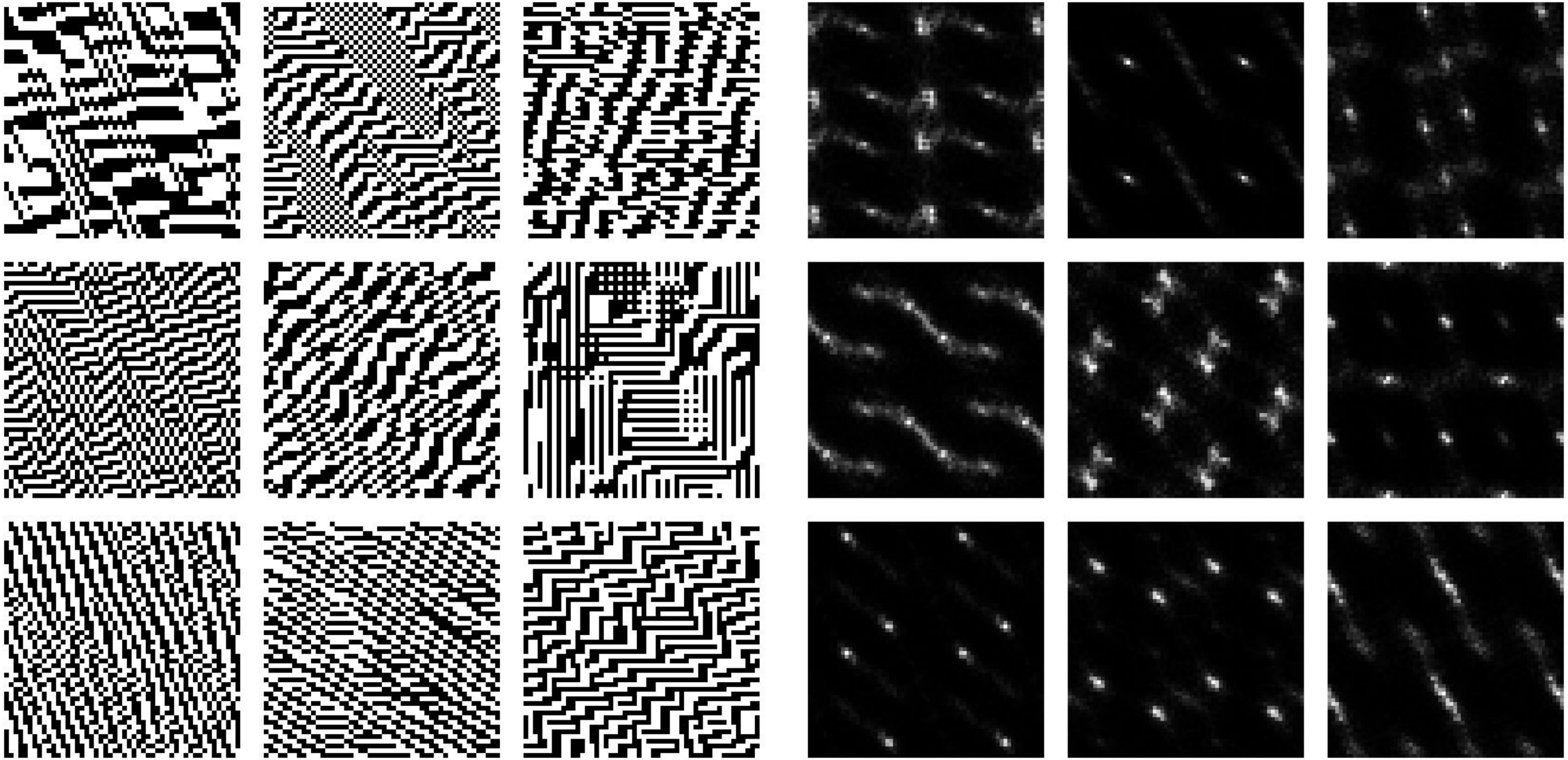
1.000	0.116	0.066	0.034
0.112	0.045	0.010	0.010
0.058	0.022	0.032	0.025
0.027	0.003	0.009	0.001

1.000	0.108	-0.060	-0.005
0.021	0.042	0.005	0.016
-0.071	0.055	-0.015	-0.018
-0.025	0.040	-0.002	0.041

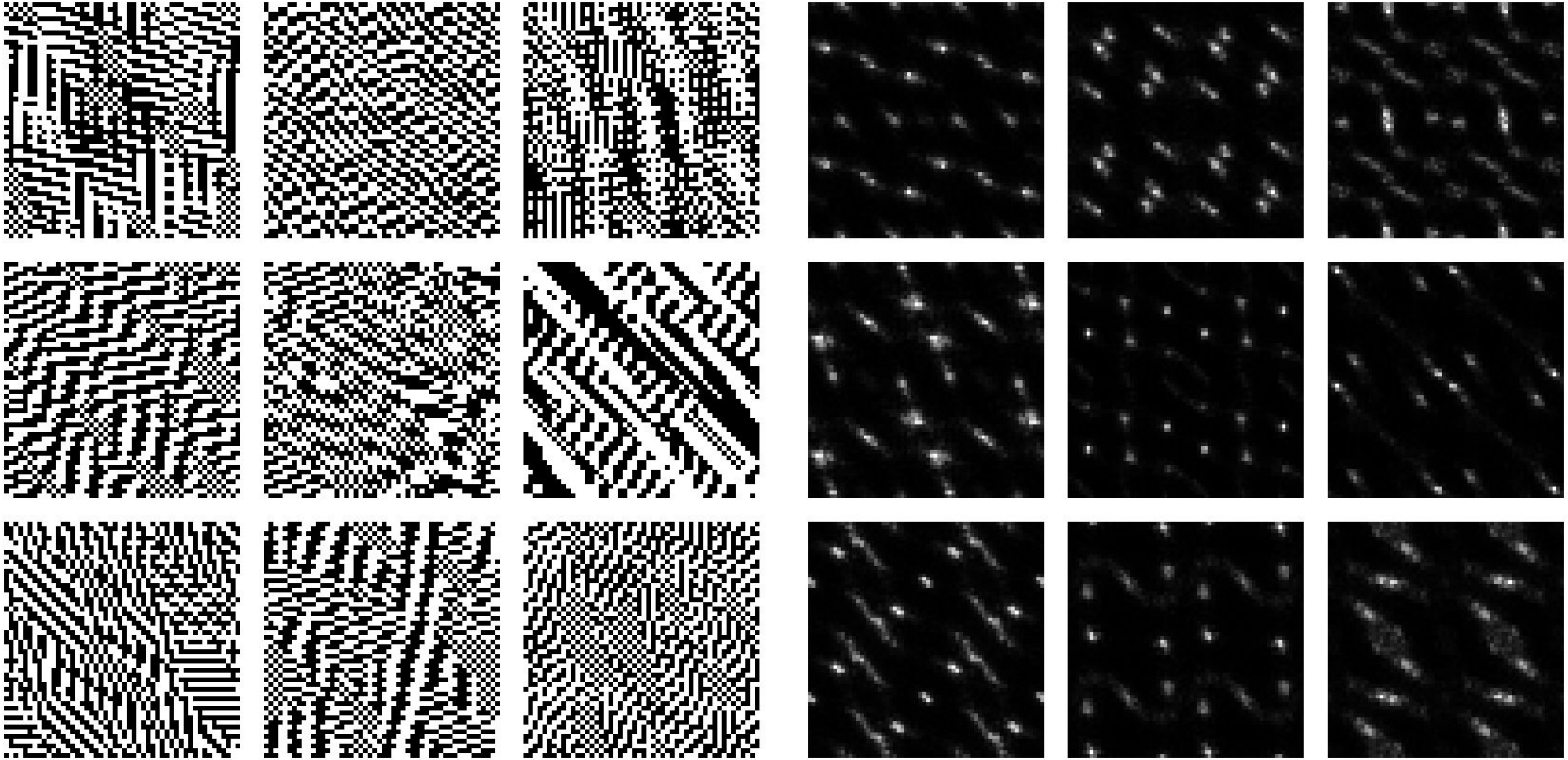
CFS2 type

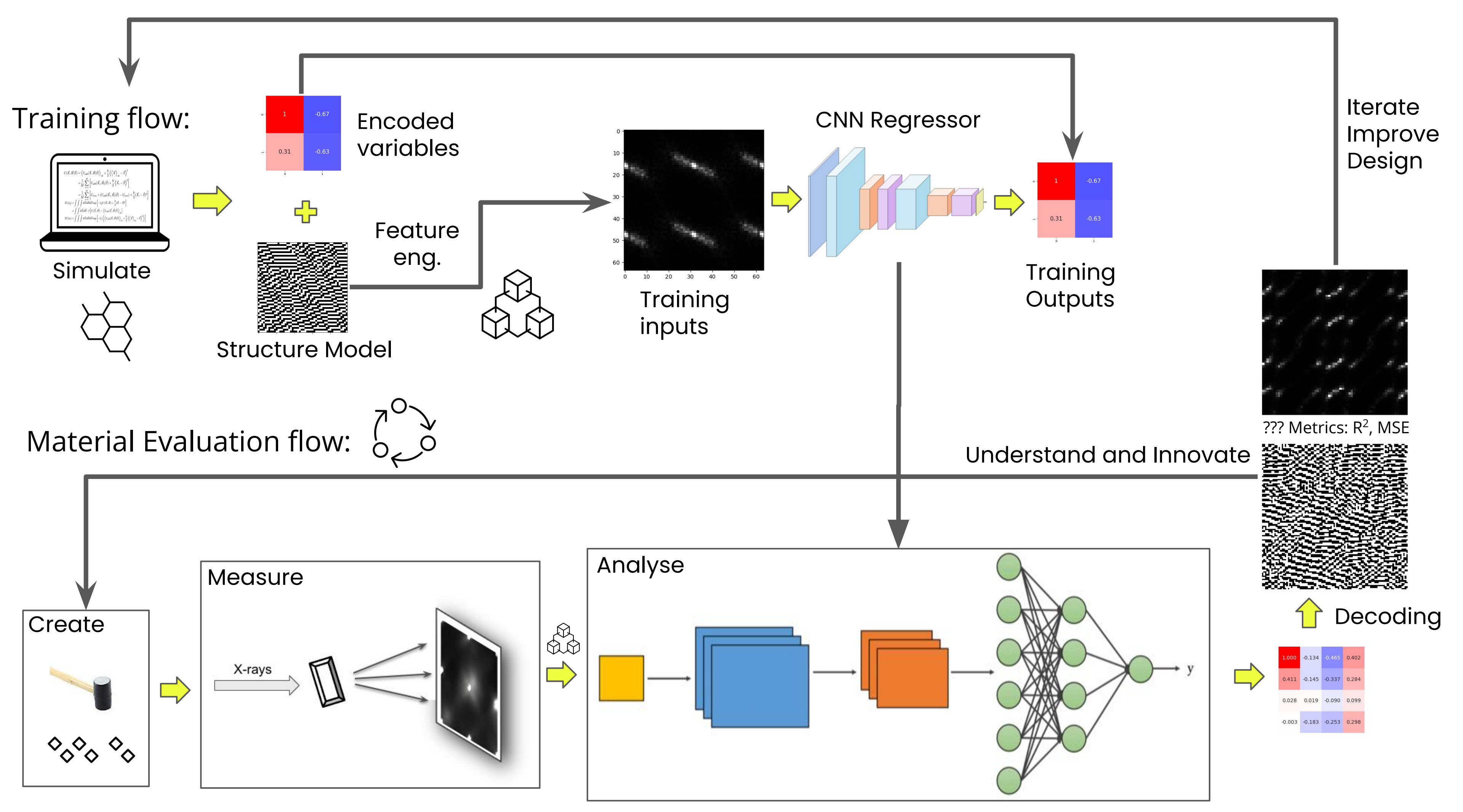


CFS3 type



CFS4 type

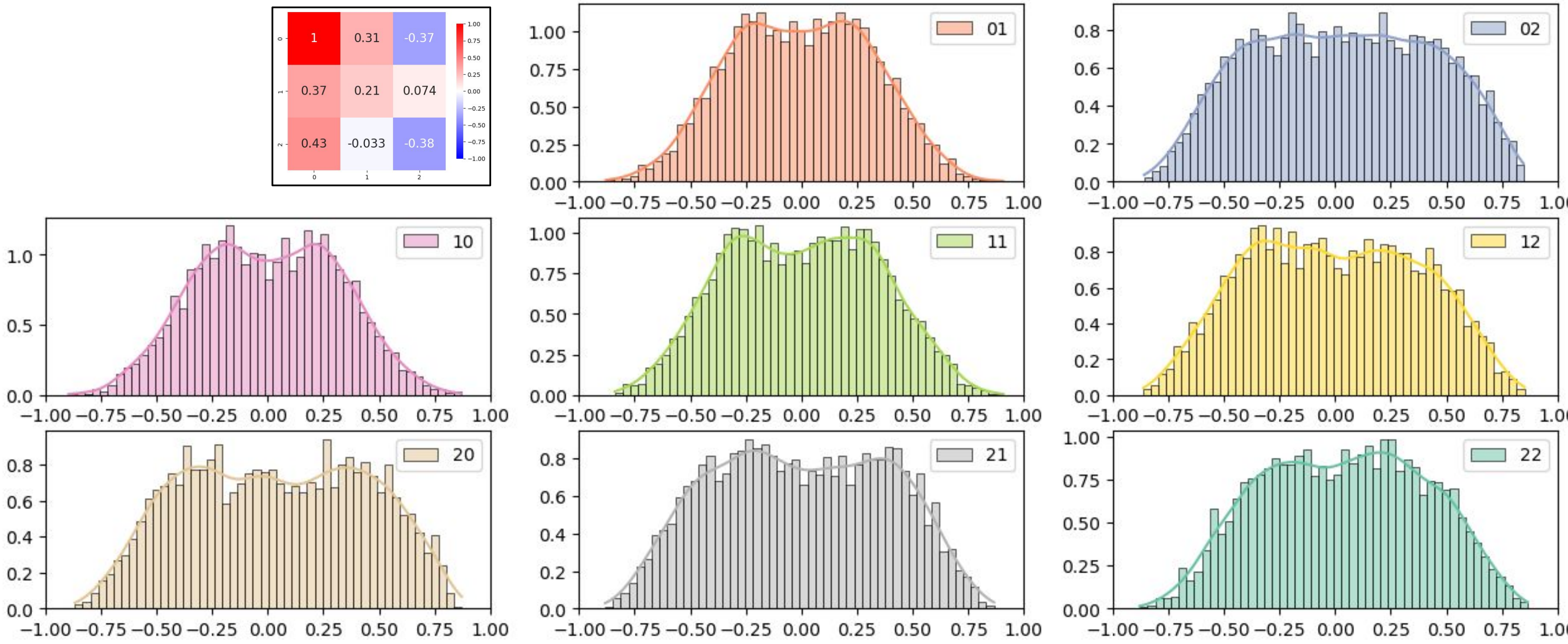




Training flow:
Simulate
Encoded variables
Structure Model
Feature eng.
Training inputs
CNN Regressor
Training Outputs
Iterate Improve Design
??? Metrics: R², MSE
Understand and Innovate
Decoding
Material Evaluation flow:
Create
Measure
X-rays
Analyse
y

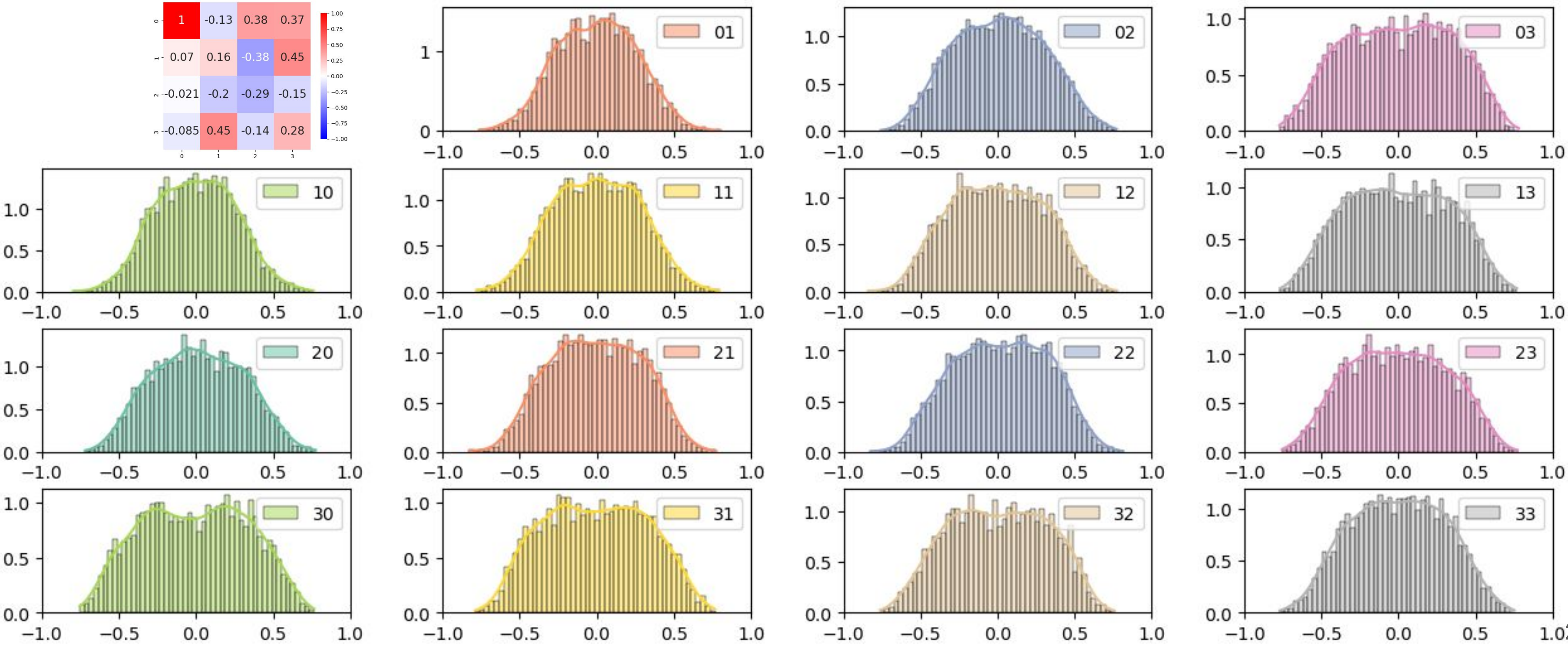
Univariate distributions of CFS3 target coordinates

Density distributions of the randomly generated target CFS3 variables used for CNN fit.



Univariate distributions of CFS4 target coordinates

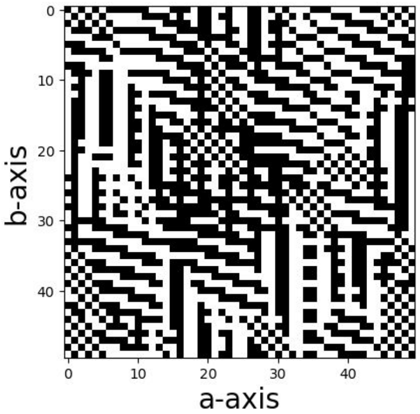
Density distributions of the randomly generated target CFS4 variables used for CNN fit.



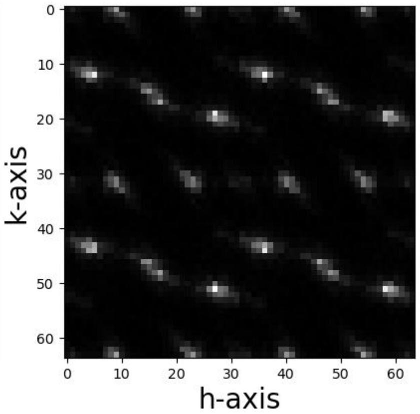
How does simulation size and sampling size affect the theoretical images ??

(keep in mind the Relative size in comparison to actual X-ray scattering from a real sample)

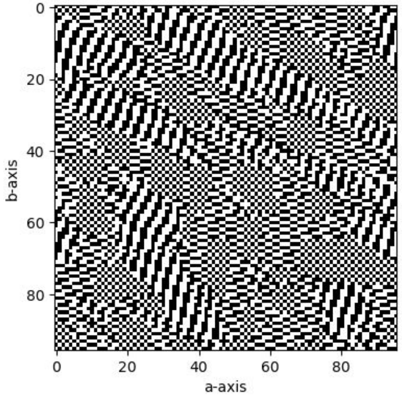
50x50 units



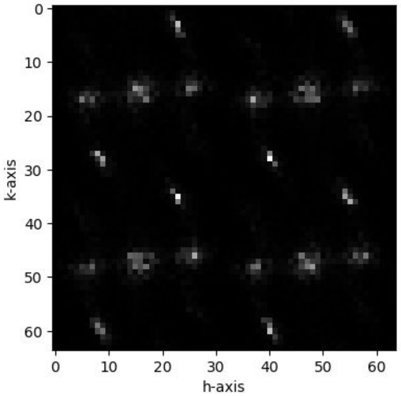
64x64 pixels
(DFT: 8x25x25 lots)

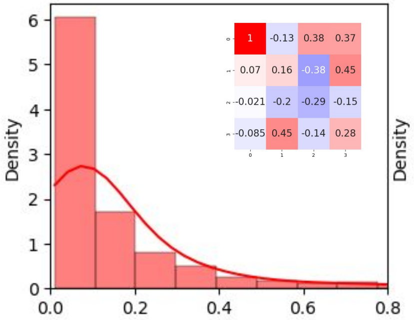


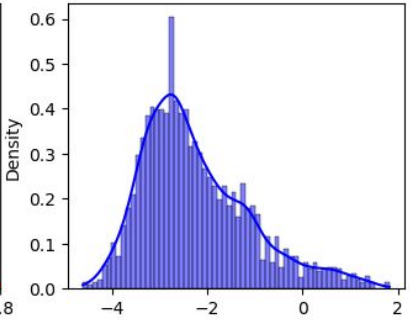
96x96 units



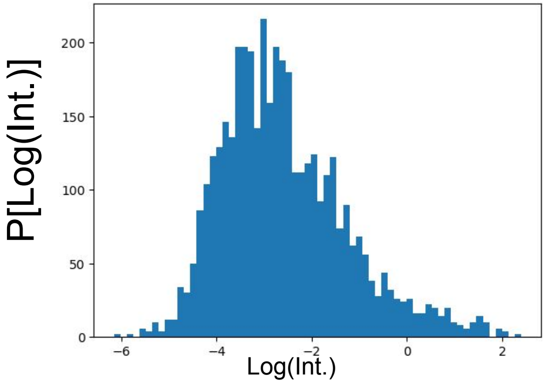
64x64 pixels
(DFT: 4x48x48 lots)

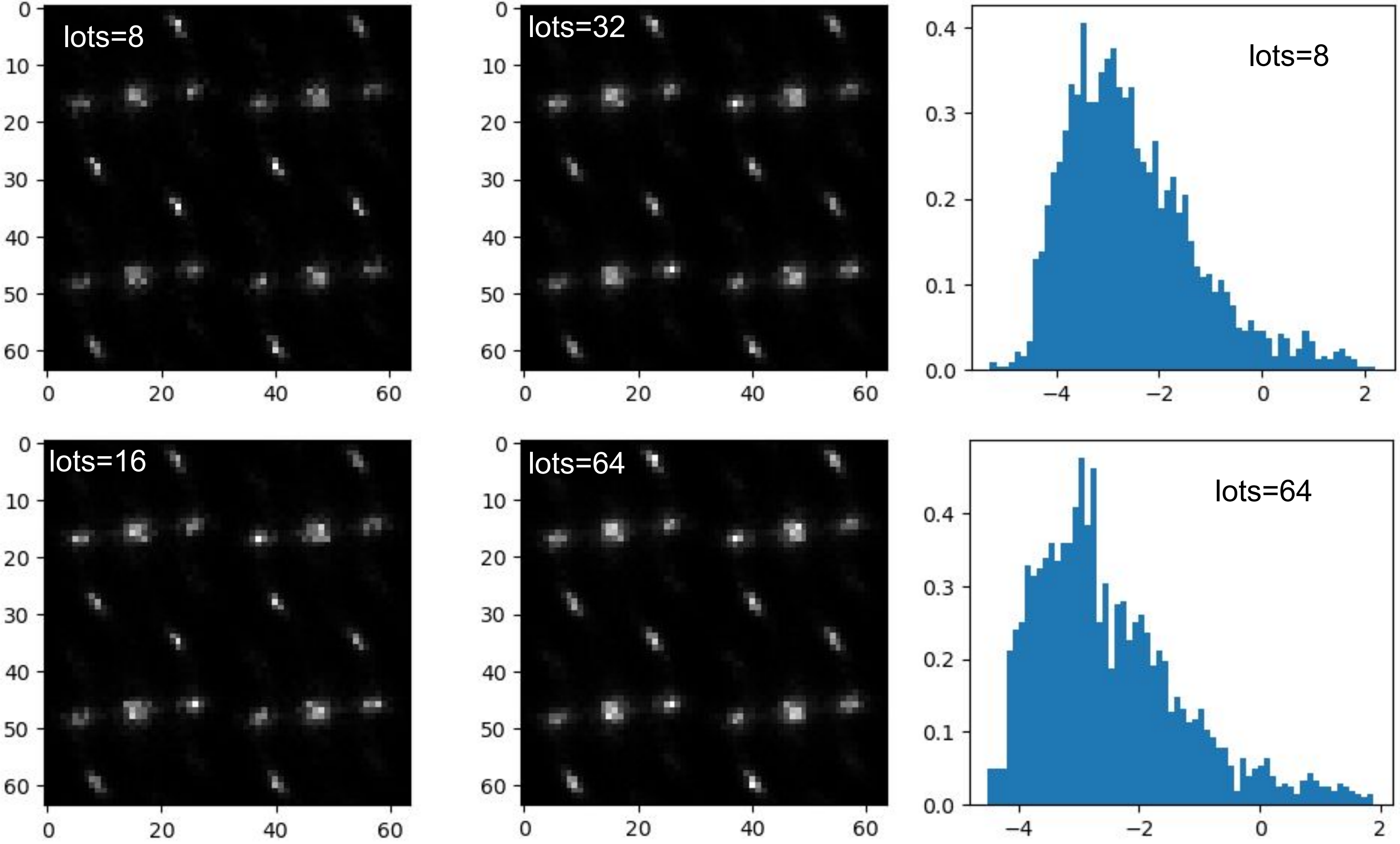






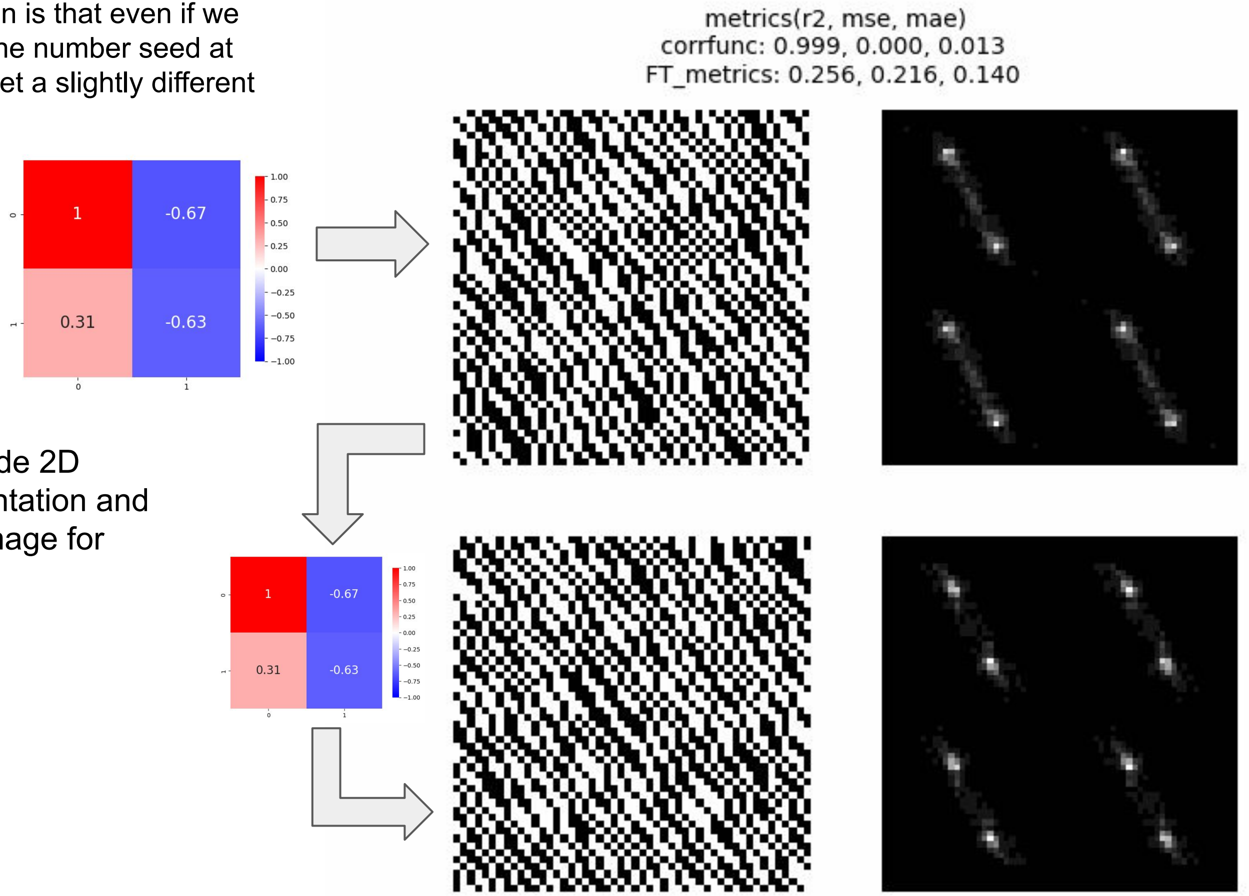
Intensity distribution above single image of CFS3




lots=8
lots=32
lots=8
lots=16
lots=64
lots=64

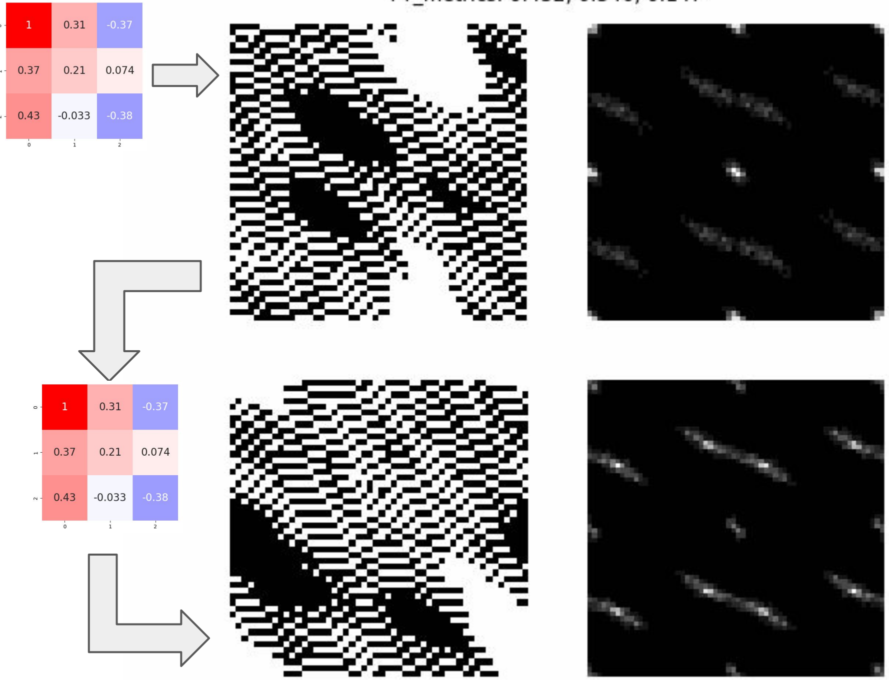
The nature of any MC simulation is that even if we use the same input, changing the number seed at various stage ensures we will get a slightly different answer.

Simulation Inputs encode 2D microstructure representation and calculated diffraction image for comparison.



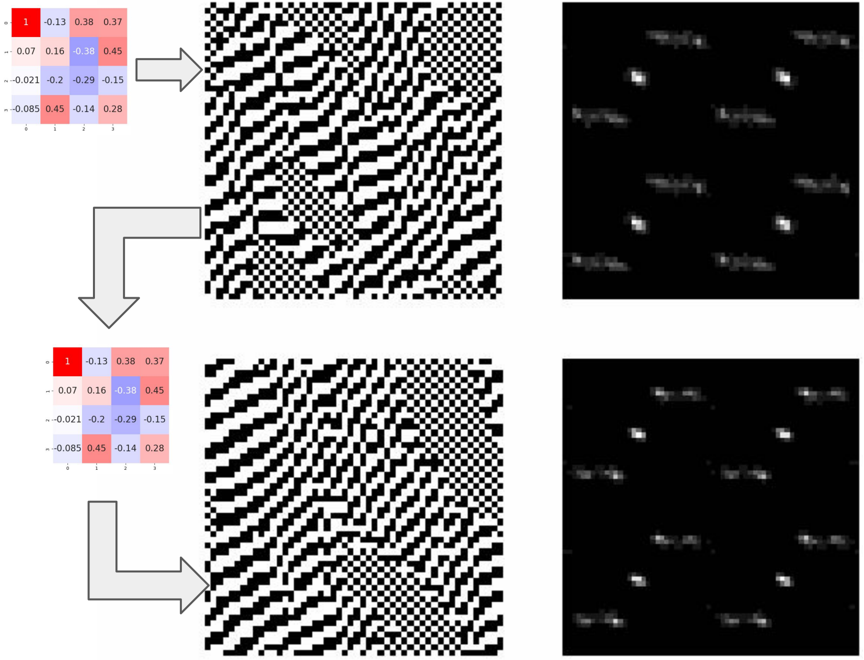
Gen-type: CFS3

metrics(r2, mse, mae)
corrfunc: 0.998, 0.000, 0.015
FT_metrics: 0.432, 0.346, 0.147

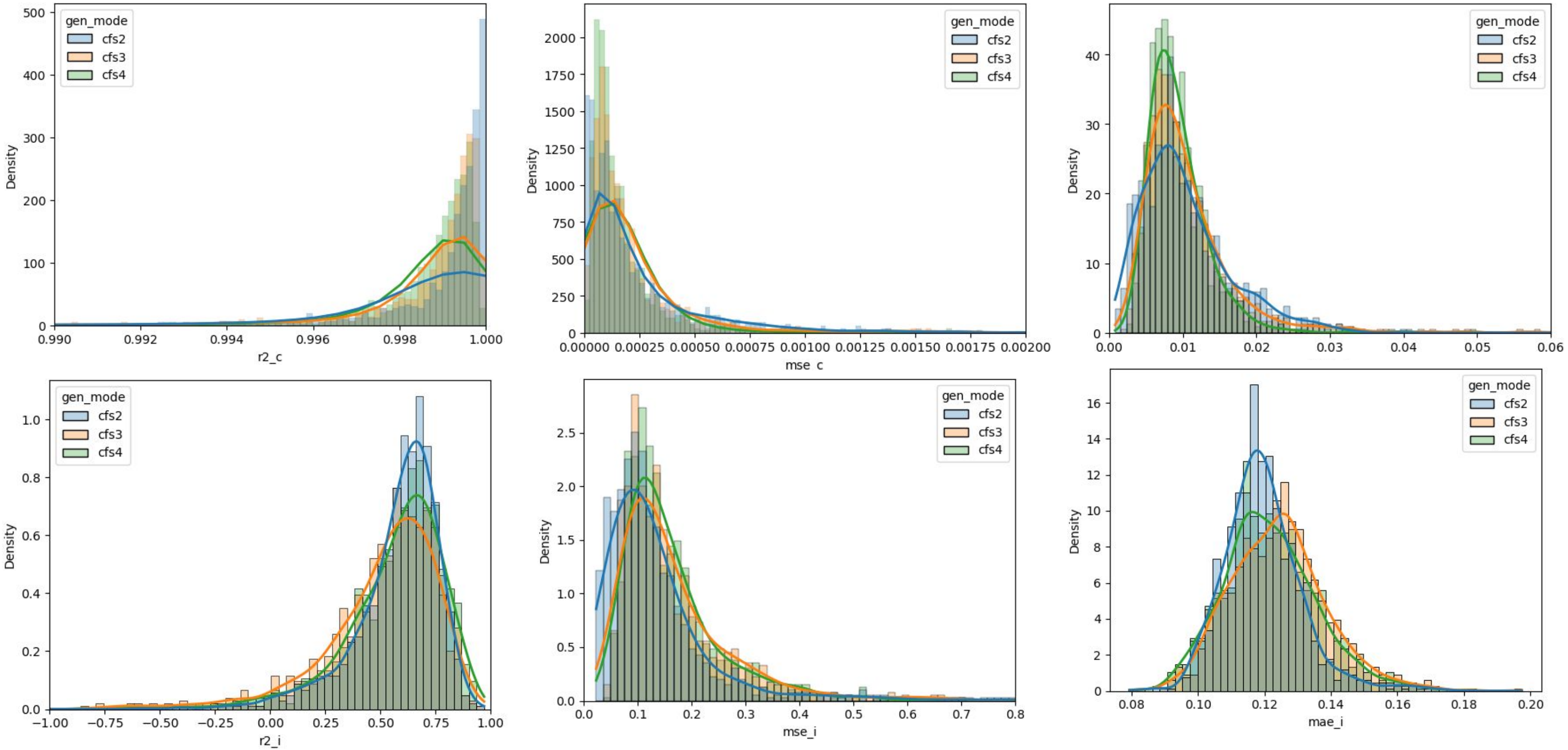


Gen-type: CFS4

metrics(r2, mse, mae)
corrfunc: 0.998, 0.000, 0.013
FT_metrics: 0.652, 0.131, 0.122

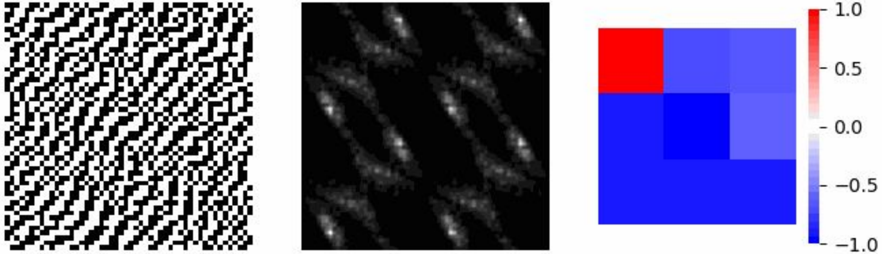


EDA: Random error in MC generation. 1000 random CFS vectors re-sampled from test/training data and compared with original sample.

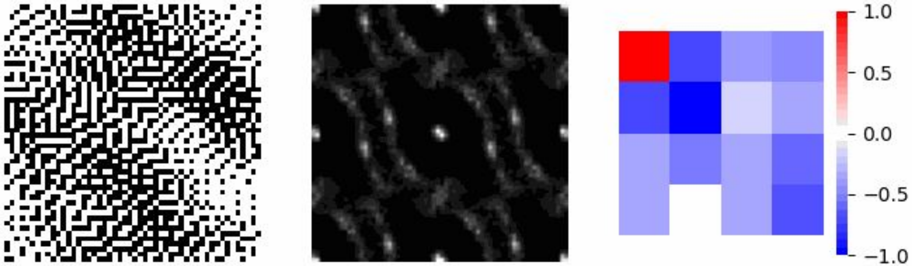


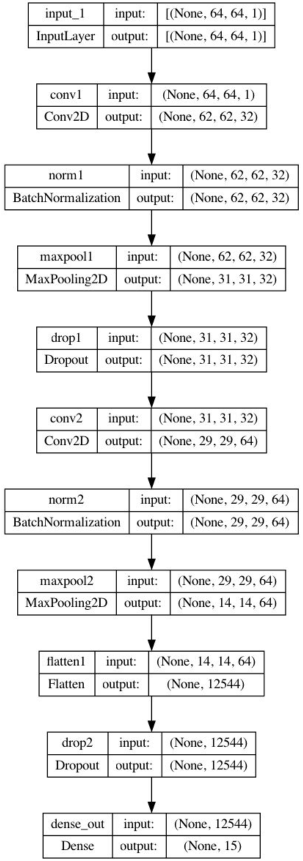
CFS2 vector L2 distance from origin 0.118

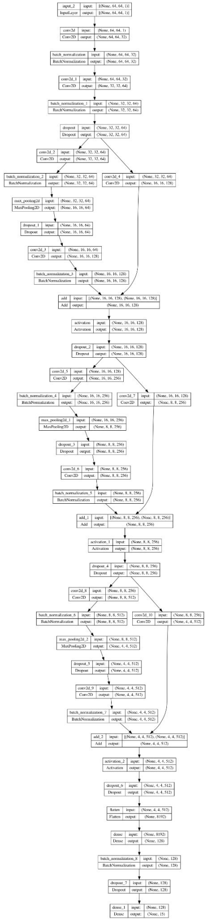
CFS3 vector L2 distance from origin 1.123



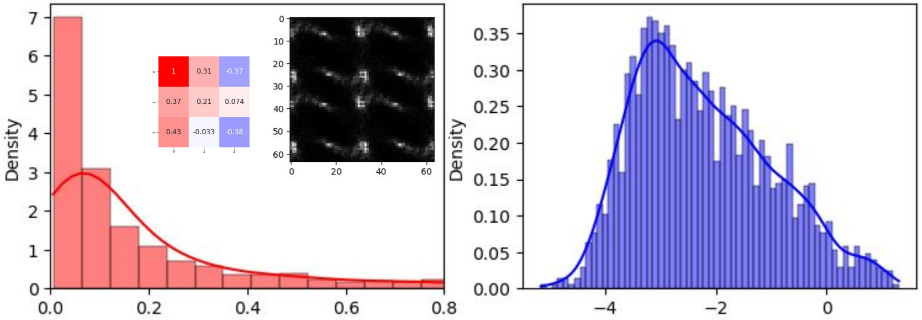
CFS4 vector L2 distance from origin 1.198



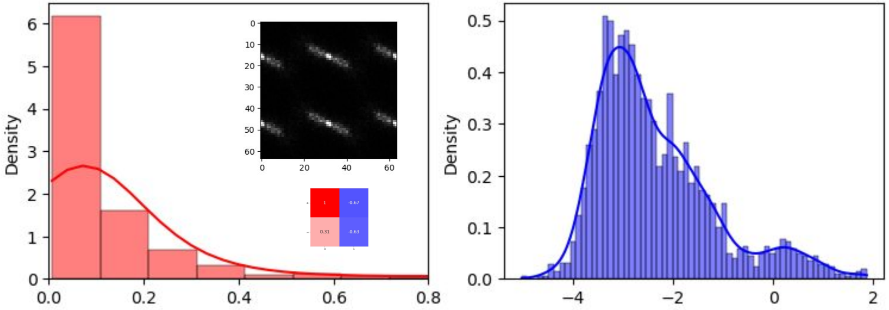




Intensity distribution single image of CFS3



Intensity distribution single image of CFS2



Intensity distributions in 5000 images of CFS2 theoretical test/train data for CNN_{CFS2}