Capstone Project

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

By
Eric Joseph Chan
Date: 19/05/2023

Table of Contents

Problem statement

The state-of-the-art in X-ray imaging of crystalline materials provides a highly detailed view of atomic or molecular scale structure that is still yet to be fully exploited by industry. Uncovering such details leads to understanding structure-property relationships which leads to optimising materials properties as well as reducing manufacturing costs. High throughput of correctly interpreted and labelled X-ray images is limited. Inherent challenges in interpretation are due to both a lack of automation and lack of expertise.
Successful interpretation of X-ray images still requires significant physical modelling, ongoing trial and error with expensive high performance computing resources and a need for this high level analysis to be performed by someone with significant expertise.

A general schematic for this part of the problem from the high level production development pipeline is depicted in **Figure 1**. By empowering data interpretation workflows through integration of a data science pipeline that can streamline the analysis process and provide valuable insights into the properties of crystals, it will be possible to efficiently and accurately analyse necessary details from large sets of raw X-ray images. It will further enable technologies towards understanding the physical nature of atomic or molecular structure that leads to desirable properties.

The underlying problem statement for the scope of a data science project is in how to apply feature engineering and machine learning engineering towards enabling this rapid interpretation of X-ray images. As we will endeavour to show later, this goal can be achieved through developing a data science pipeline that incorporates a method based on convolutional neural networks (CNNs). This approach is depicted in **Figure 2** as bypassing the traditional approach with the CNN, thus streamlining the analysis process and rapidly providing valuable insights into the properties of crystals without the needed for tedious trial and error or expert opinion.

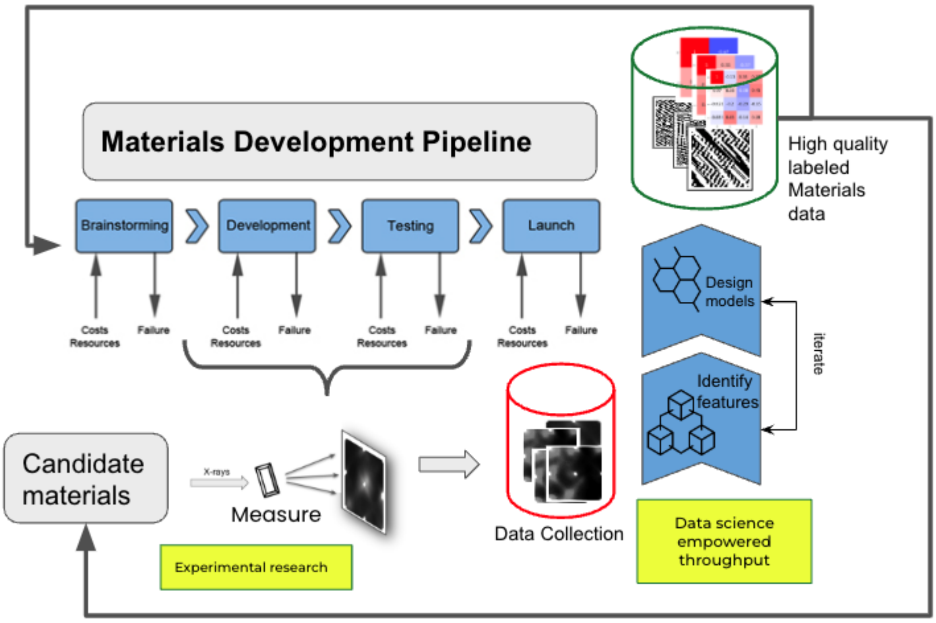


Figure 1. Modified portfolio development pipeline for the representative stakeholder companies who have internal or external materials research incorporated as part of the development process. Here, the objective of data science is to integrate into the workflow and empower the throughput of the experimental research such that an increase in high quality interpretations of the experimental research will give feedback not only into the development and testing of the candidate materials but also in enhancing the design of new products.

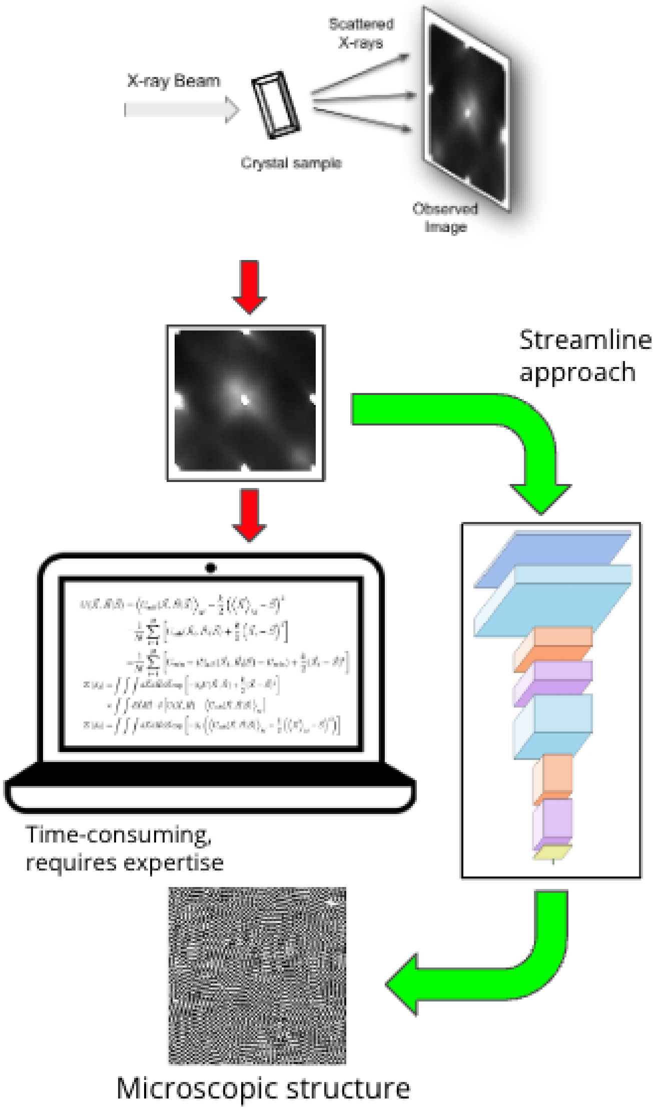


Figure 2. The traditional path towards obtaining a detailed view of microscopic structure from X-ray images is difficult and requires significant resources both from the standpoint of computing power and necessary trials but also workflow must be managed by workers with a high level of expertise. The idea of integrating AI into the workflow will streamline analysis by bypassing the traditional approach with something more automated.

Industry/ domain

Most domains that have workflows which intersect the field of materials science can benefit, but primarily research and manufacture of semiconductors, energy storage, pharmaceuticals, ceramics, agrochemicals and thin-film materials.

Stakeholders

Researchers, scientists, engineers, manufacturers, and quality control personnel who are involved in the development, production, and quality control of high-tech materials and devices. For example, they may have observed that adding traces of acetylsalicylic anhydride to aspirin during its production has the highly profitable effect of increasing the dissolution rate. They formulate a theory that has to do with the interaction of the impurity incorporated with and compromising the overall crystal structure. They want to be able to rapidly confirm that the theory is correct and thus it can be patented and exploited in their assets when required.

Business question

How can we enhance the design space for profitable materials with desirable properties at reduced manufacturing costs? This will ultimately lead to an increased throughput of profitable products (e.g., materials that go into making lighter yet more powerful laptop computers). This question drives our curiosity about understanding the details of material structure. For example, perhaps we want to design a new alloy with improved strength, durability, or conductivity. What details at the microscopic level can we exploit in a cost-effective manner? Perhaps we can replace traces of boron with carbon. The notion of a materials design space encompasses practical aspects of the materials analysis pipeline in that we must have a high throughput analysis scheme in place to screen through large quantities of experimental materials and want to enable non-SME (subject matter expert) employees to perform the analogous high-quality interpretations pioneered by SMEs. This also allows for a reduction in business overhead for SME's or the reallocation of SME resources.

Data question

There are several data-related questions that pertain to the problem statement of how to streamline the analysis stage for X-ray images. Primarily, the main data question is: How can we address the fact that there is a shortage of available X-ray image data for training that has been successfully labelled by SMEs? Perhaps we can supplement or fortify the available data.

Related to this question are the machine-learning engineering considerations. For instance, we might consider that one powerful artificial general intelligence system (AGI) can be trained to interpret any given X-ray image on a general scale, much like a generative pre-trained transformer model (GPT). On the other hand, for a highly functional workflow, CNNs may operate within a narrow AI scope, and multiple models need to be constructed on a case-by-case basis. We must also consider how new interpreted data can be fed back into the development pipeline.

Data

A few small samples of raw X-ray images from crystals of 1,5-Dichloro- 2,3-dinitrobenzene were made available upon request. The crystals were the topic of a Ph.D thesis (see references) and the image data that was received had been pre-processed. Further image processing was performed so that the images were suitable for evaluating the performance of CNN interpretation (described in next section).

To train the CNN a computational workflow for generating simulated X-ray image data was established and described herein and in references provided. How this works begins with a variable encoding in the form of a correlation space state vector which has a dimensionality (D) referred to as the correlation function span (CFS), whereby D=(CFS^2-1). This CFS vector encoding is then forwarded as input to a statistical physics simulation engine which outputs a 2D-lattice grid (representation of the crystal structure) where each point on the lattice represents a binary variable (1 or 0). A feature engineering step is then performed on the grid known as a Fourier transformation (FT) to prepare data for learning by the CNN as a simulated form of X-ray image (Feature engineering involves creating new features or representations from the existing data that can enhance the performance of machine learning models).

The procedure thus far decodes the information about a possible crystal structure which was originally encoded as the CFS vector (see **figure 3** and references). To generate a numerous variety of training data images as CNN inputs we use pseudo-random numbers to generate the CFS vector encodings which map as outputs for the CNN. Using this approach to 5000 simulated images, grids with corresponding encodings for three different output dimensionalities (D=3, 8 and 15) with examples shown as **figure 4** (the standard settings were grid: 50x50, block averaging FT over 8 lots, lots size: 25x25). To evaluate the effect of lattice simulation grid size and FT sampling we generated a further batch of 5000 image/encoding pairs with the increased resolution setting for the D=15 (grid: 96x96, block averaging FT over 64 lots, lots size: 48x48). Inputs to the CNNs are all essentially 16-bit grayscale images fixed at 64x64 pixels, however we stored the data in binary format or .h5 for convenience.

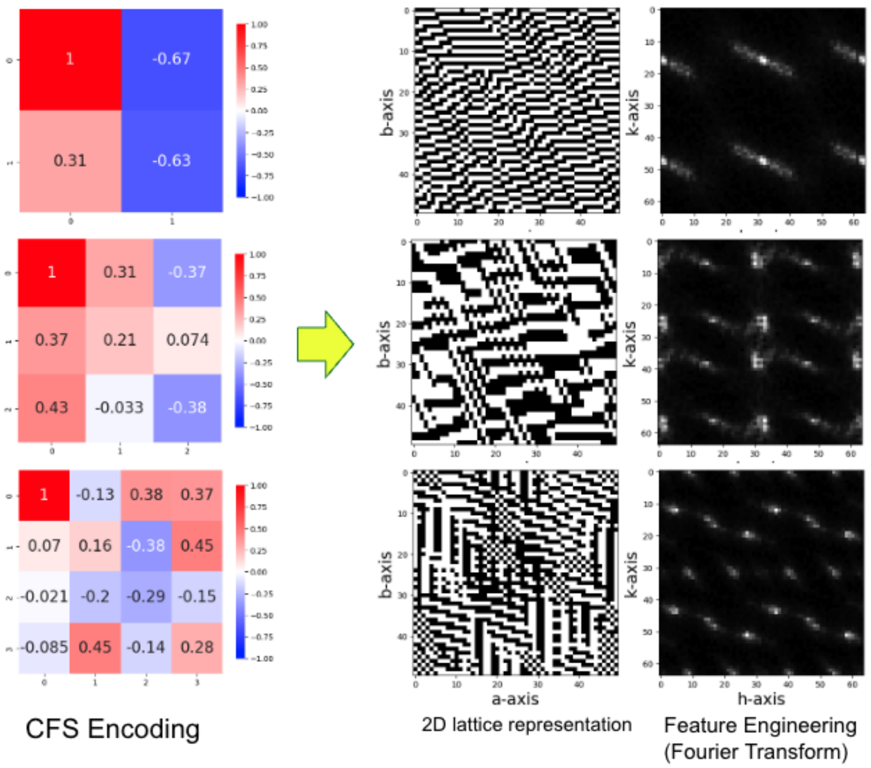


Figure 3. How the different CFS vector encodings are decoded by the physics simulation to produce a 2D-lattice grid representation of the image. (*rows top to bottom*) depict the respective D=3,8 and 15.

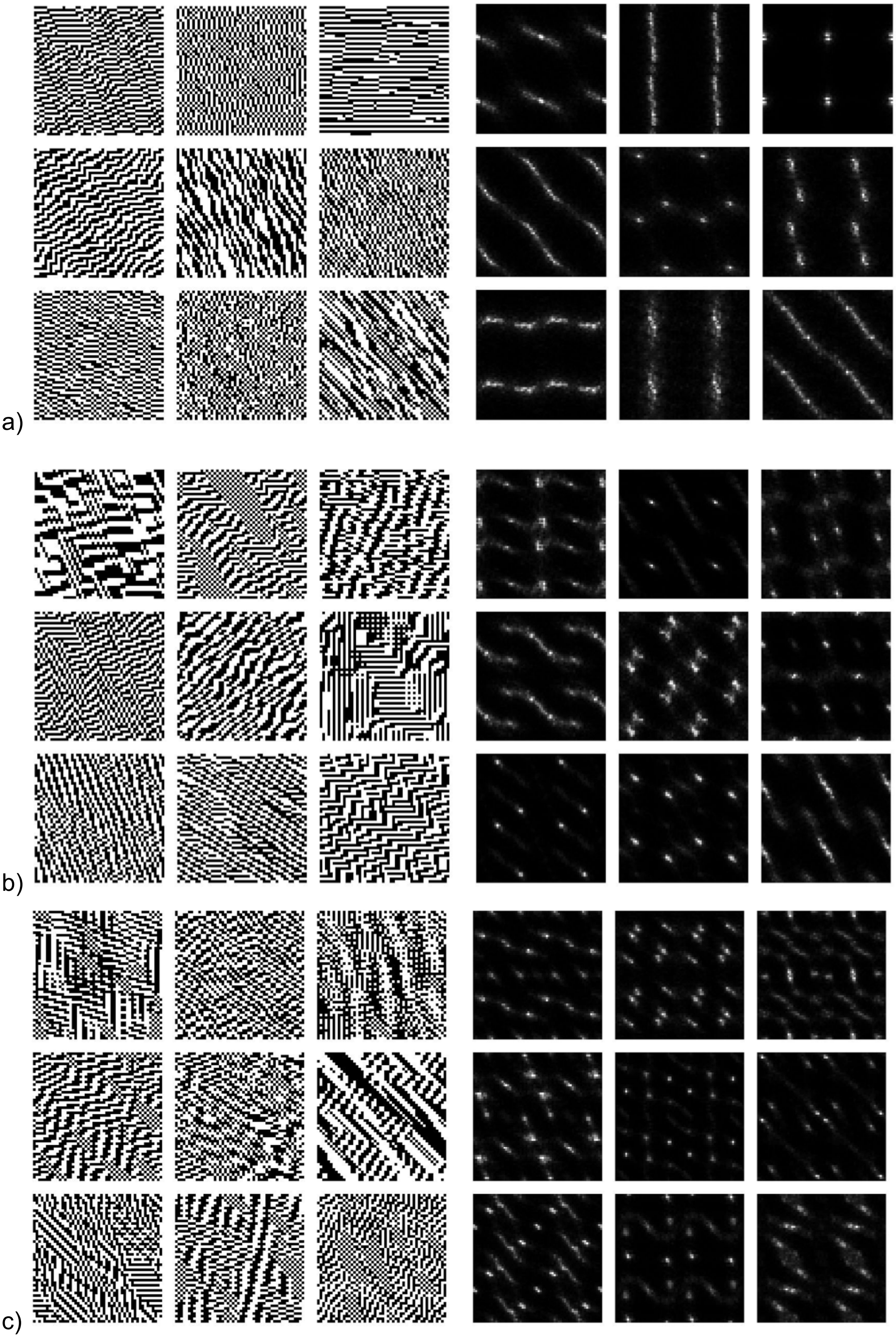


Figure 4. Examples of different randomly generated structures and simulated X-ray images for (a) CFS2, (b) CFS3 (c) CFS4.

Data science process

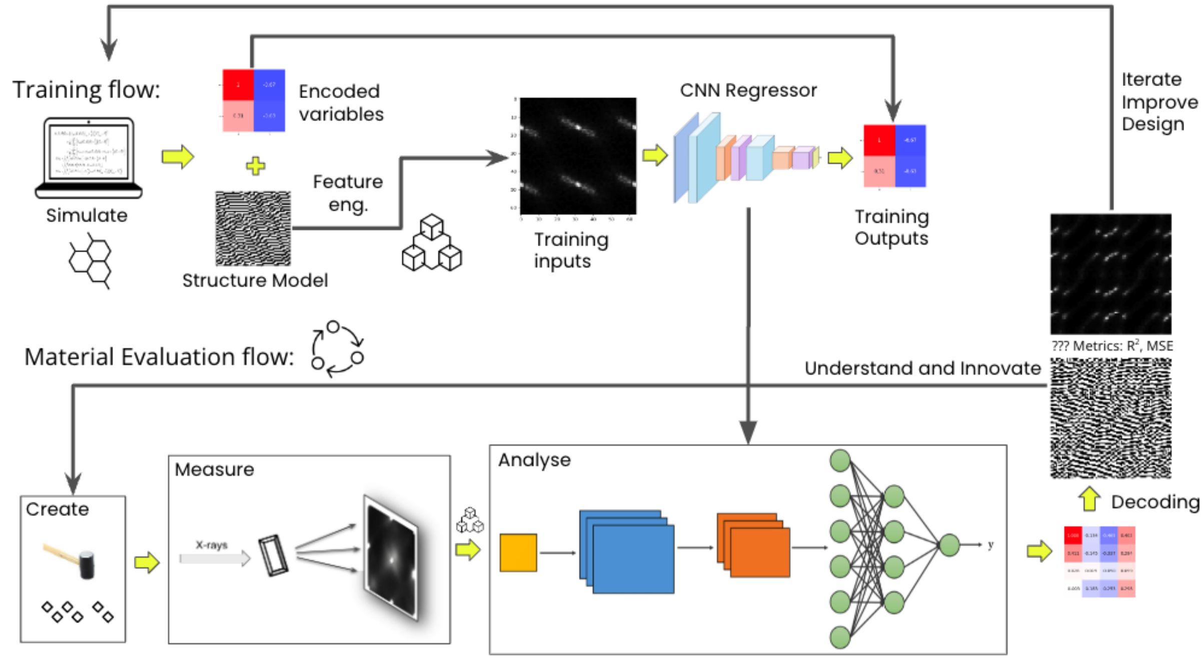


Figure 5. Schematic representation of the Data science pipeline which depicts simulation of training data and subsequent CNN regression to interpret X-ray images which is part of the original materials analytics workflow. As the pipeline builds up interpreted materials data the knowledge stream feeds back to improve the materials design process as well as the model training and interpretation cycles.

Figure 5. represents a schematic guide of the data science workflow. As depicted in the figure, the CNN is being trained in the context of a supervised learning regression model. It takes in images and outputs a CFS encoding. The encoding then provides a means of both understanding physically meaningful parameters for the material as well as a crystal lattice representation and comparative X-ray image generated via the same statistical physics engine that were used to generate the training data. For this investigation we evaluated several CNN which were trained (in the narrow AI sense) to output CFS of specific dimensionality (explained in further detail in the modelling section of this report). The computational workflow for generating the simulated X-ray images was setup both as jupyter notebook and standalone python scripts. The physics engine and fourier transform components require some external fortran routines and dependencies to run.

Data analysis

Exploratory data analysis (EDA) was performed on the collection of encodings for each CFS (CNN model output variables) in the form of histograms (**figure 6**). This EDA was for two reasons, firstly to make sure that nothing went wrong during the generation process of the simulated data, secondly, to get a look at the coverage of the variables within each distribution and make sure they were reasonably symmetrical. The example shown in **figure 6** shows this is indeed the case.

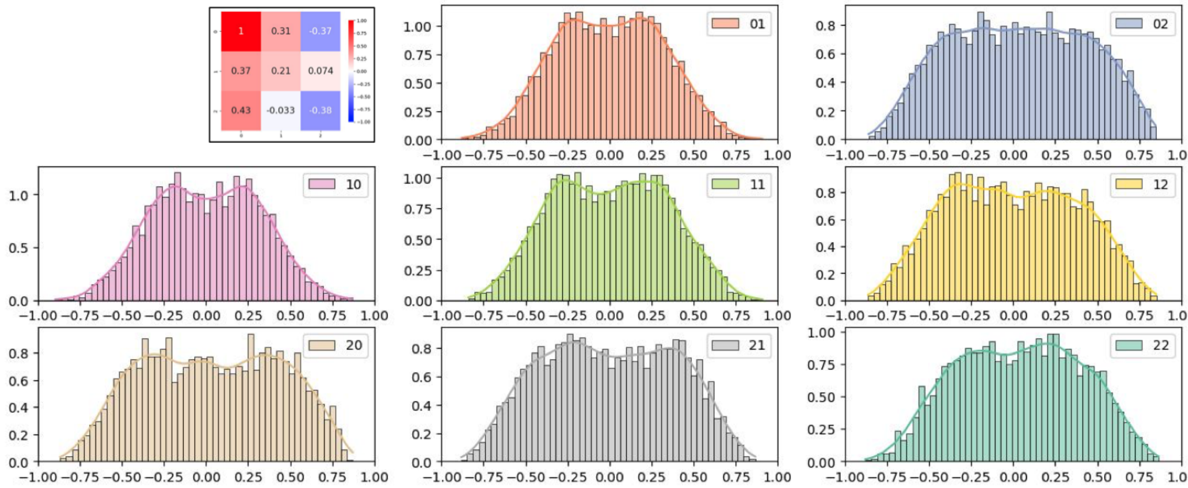


Figure 6. Density Histograms of the 8 CFS3 variables. each encoding (example shown in top left has 8 parameters)

The experimental images and the simulated image/encoding pairs were also subject to EDA prior to any training. The reason for the latter was primarily as a control action to ensure that the data was free of errors other than the expected statistical noise. In the case of the observed data we want to make sure that the distribution of intensities over all pixels is reasonably matched with the intensity distribution of the simulated data. The observed data required Image reconstruction, data compression, smoothing, normalisation, re-scaling (with optional: Top-hat filter, image restoration in-painting). The images and intensity histograms before and after correction for two 64x64 images are shown in **figure 7**. The reasoning for image corrections was to get the real X-ray images to have intensity distributions as close as possible with the training data, since areas of strong intensity (Bragg peaks) might limit the performance of the CNN. The image correction functions other than those of SKimage library are all custom made and made available in a jupyter notebook for this project.

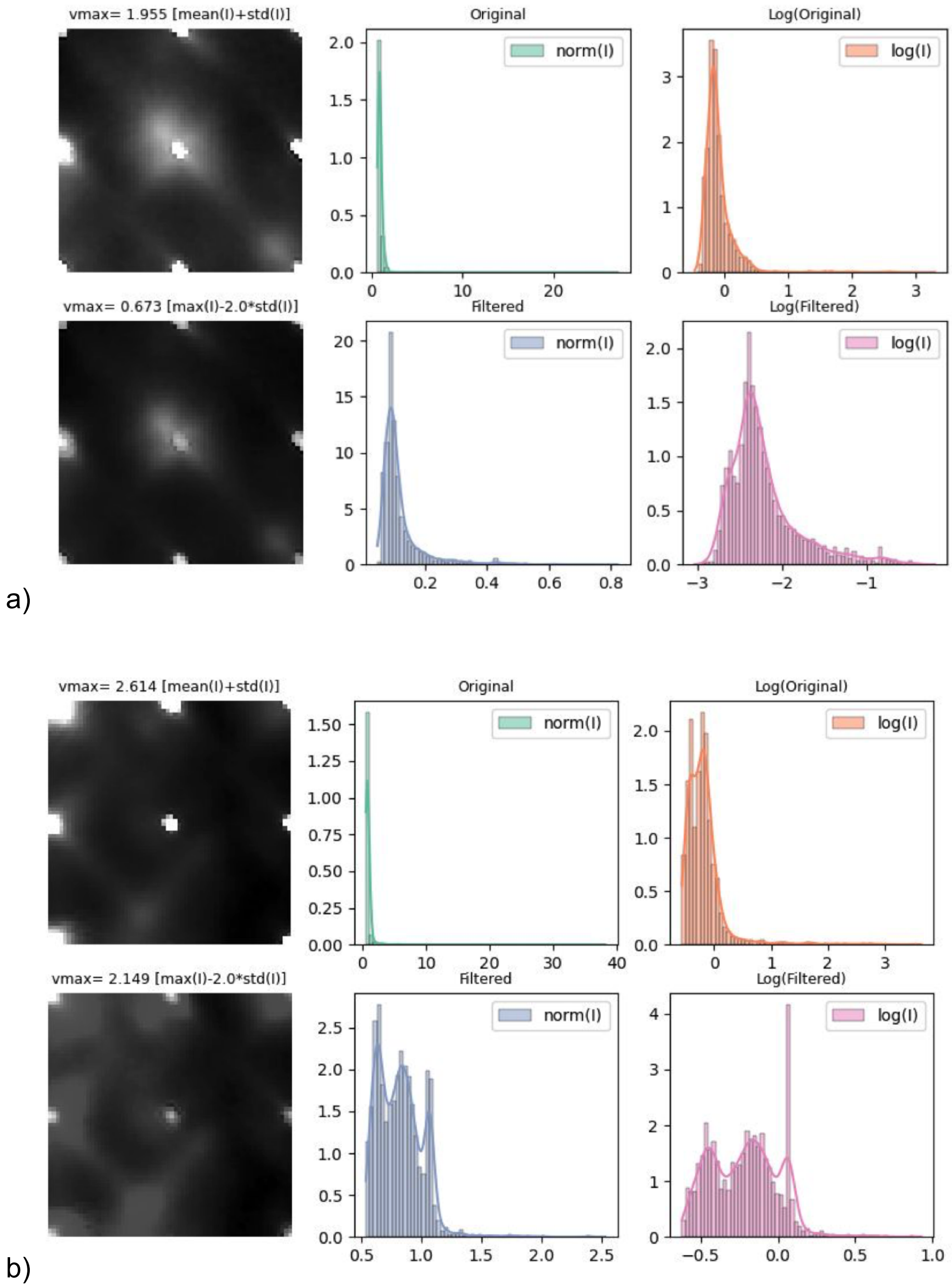


Figure 7. EDA for observed X-ray images (a,b) with comparison before(top row) and after (bottom row) corrections are made to the images. images are on the left column, with density histogram in the middle and density histogram of log(Intensity) on the right.

For EDA of the training images **Figure 8** depicts the corresponding intensity histograms of a single example image taken from the CFS3 training set. **Figure 9** shows density histograms for all pixel intensities of the CFS2 training set. The aim of the EDA is to make sure that the current state of the data was suitable for training. Most of the simulated data is weak and we made no further processing. it is difficult to tell at this stage if the modelling would benefit from further preprocessing and feature engineering the training data. This is difficult to discern because most fits with the current CNNs are already converged with very low loss functions. The simulation and image decoding settings for the training data are an important consideration of the data-science flow as depicted in **figure 10**. Further benchmarking at this level will be required to improve the interpretation performance of the CNN.

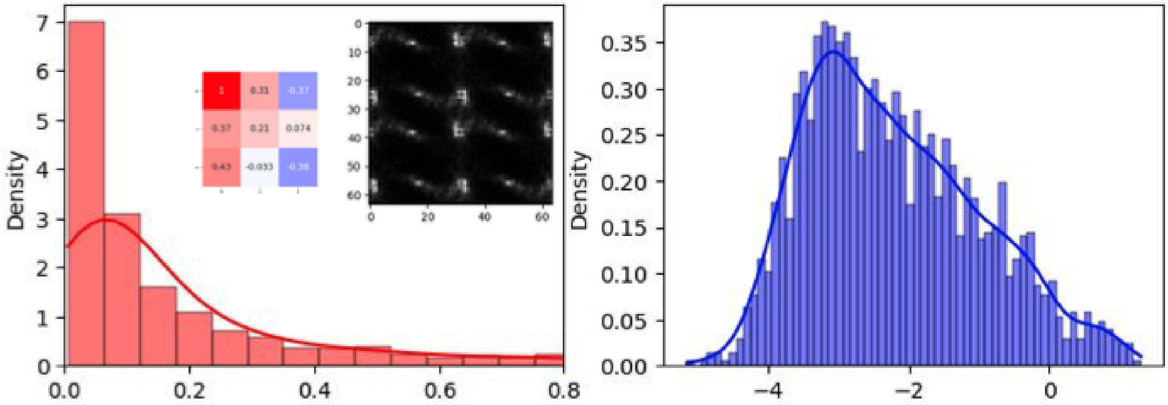


Figure 8. Density histograms for (left) intensity and (right) log[intensity] of an example training image. The log-normal distribution of intensity is less apparent.

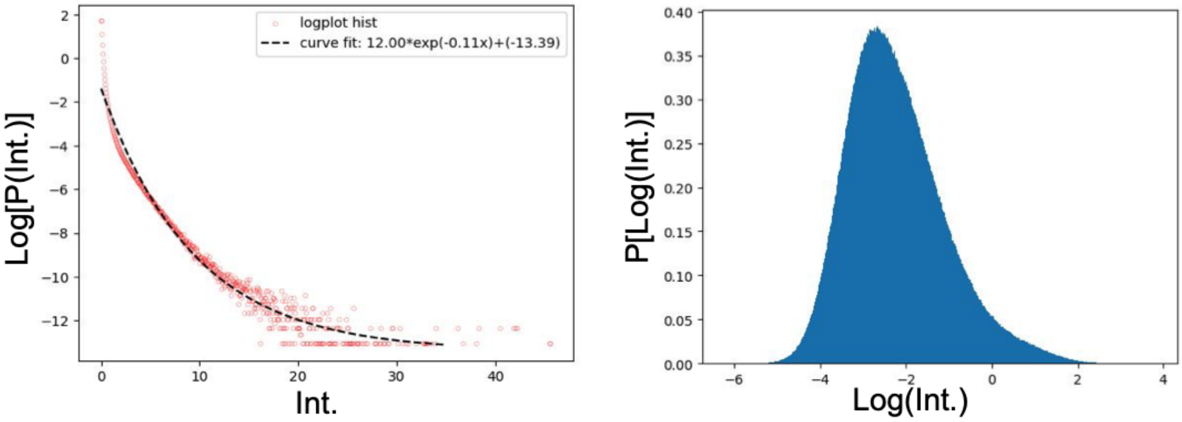


Figure 9. Density histograms for all pixel intensities of the CFS2 training set. The log-normal character of calculated intensities is apparent, however the distribution has some skewness and kurtosis.

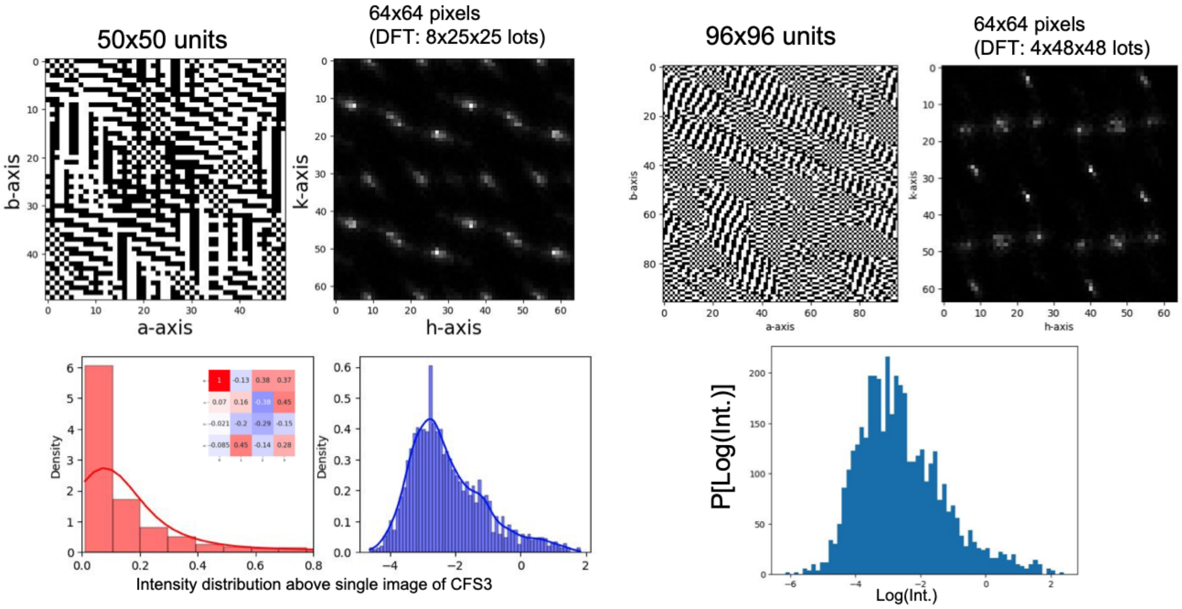


Figure 10. Comparison of training images generated with different simulation sizes and different averaging over the FT. The effect of these settings is noticeably different and very likely to affect the quality of the CNN. Consideration of these settings are a component of the design aspect of the data science process.

An important part of the EDA analysis is to determine the statistical error associated with the simulation due to the inherent randomness of the statistical physics engine. To obtain some measure of this we sampled 1000 image/encoding pairs at random from each training dataset and then re-calculated the lattice structure, image and new-encoding from this sample encoding. We were then able to quantitatively compare each image/encoding pair using MSE and R^2 as metrics. The procedure is shown in **figure 11** and is an important consideration because it sets a limit on the quality of the decoding aspect (via randomised statistical simulations) when using the CNN for X-ray image interpretation (by design) as a distribution of metrics (shown as histograms) to which it is able to reproduce the same simulated X-ray image. What this means is that sometimes given the same encoding, once decoded the condition of $R^2<0$ can occur between images but as we see in **figure 11(b)** this error is in the tail end of a distribution which is centred mostly $R^2 > 0.5$. This is also important to recognise because it places a limit on the degree to which an optimization algorithm can be used to fit an image with the CFS vector encodings.

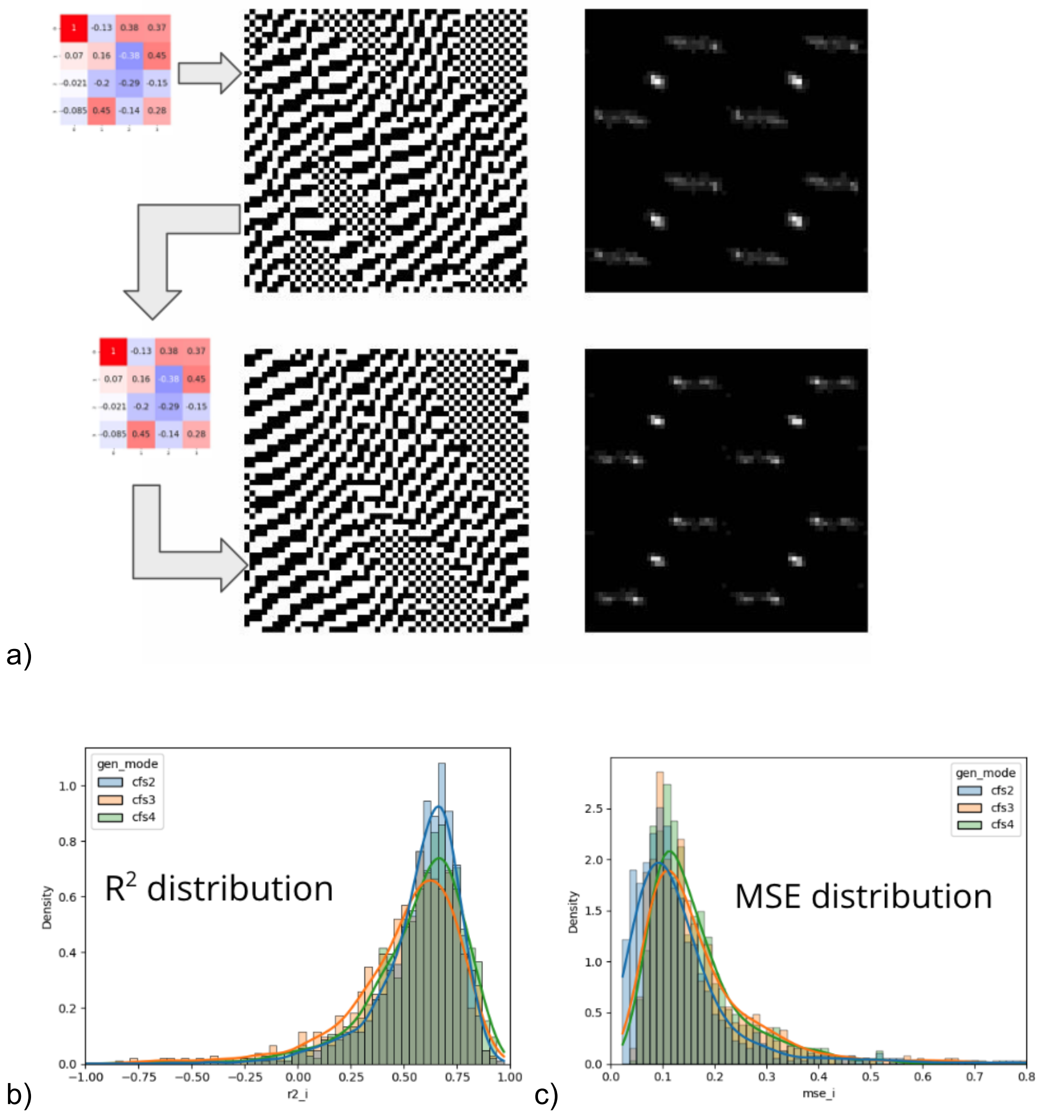


Figure 11. (a) Example of difference in decoded image and structure from a previously sampled encoding and the intrinsic error due to inherent randomness of the statistical physics engine. Notice how the structures are not the same but they look similar and in term of microscopic structure can be classified as equivalent (much like different handwritings of the same number). The effect on X-ray depiction is also similar. (b,c) This intrinsic error is best represented as distributions by histograming the error metrics associated with comparison of pixel intensities.

Modelling

There were two different CNN architectures investigated in this project which comprise different sizes and settings (**figure 12**). The smaller architecture uses the default settings, which means it has fewer layers and parameters. The larger architecture, on the other hand, has skip connections and uses the He_normal weight initialization technique, which can help to improve performance by mitigating the vanishing gradient problem often encountered in deep neural networks. Skip connections allow information to be passed forward more easily, helping to prevent the loss of data as it flows through the network.

To account for different CFS in the training data, different CNNs (irrespective of architecture) are used for each CFS by way of the fact that each CNN takes as inputs images of 64x64 pixels, and outputs the encoding variables with dimensionalities of 3, 8, or 15. The output encoding variables represent an underlying structure for the crystal. By using different CNNs for each resolution, the analysis pipeline can be optimised to detect structure at different scales, which may assist to improve the accuracy and reliability of the results. Overall, this approach allows for an optional coarse- or fine-grained analysis of the possible X-ray images.

To evaluate CNN performance, a set of performance metrics suitable for regression models is used. This includes R^2, MSE and MAE. To evaluate the performance on interpretation of real data visual inspection of the final output is better suited. These metrics are applied to both simulated and experimental X-ray images. All CNNs are trained with 100 epochs and batch size of 32 using the ‘adam’ optimizer for weight adjustment with the default settings available in keras.

The CNN training and cross validation results are summarised in **Table 1**. **Figure 13** shows training curves for the instances of small and large models using the totality of 10000 points of CFS4 encoded training samples with a cross-validation split of 0.2. It is important to note the residual plot (Residual=Y_{test}-Y_{pred}) for the large model indicates significant bias in contrast with the small model. Because of this we decided that the small model architecture as the most suitable architecture at this stage for deployment as further testing will be necessary to determine if the large model can supersede the performance of the smaller. It became more obvious that the large model was unsuitable because it performed inadequately upon attempts to interpret real X-ray data. The large model could be overfitting the data and might only be necessary for fitting to very high dimension CFS vectors.

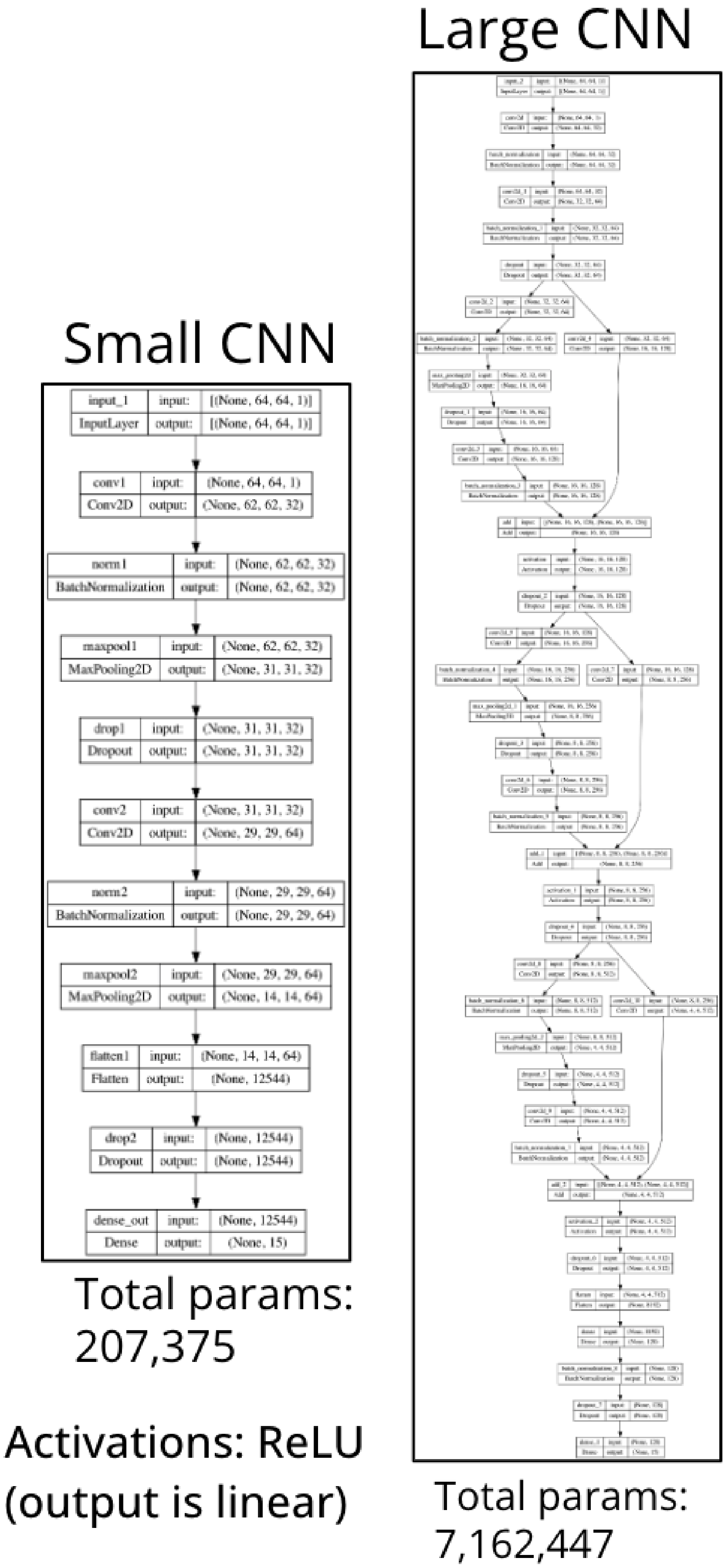

Large CNN
Small CNN
input_1 | input: | [(None, 64, 64, 1)]
InputLayer | output: | [(None, 64, 64, 1)]
conv1 | input: | (None, 64, 64, 1)
Conv2D | output: | (None, 62, 62, 32)
norm1 | input: | (None, 62, 62, 32)
BatchNormalization | output: | (None, 62, 62, 32)
maxpool1 | input: | (None, 62, 62, 32)
MaxPooling2D | output: | (None, 31, 31, 32)
drop1 | input: | (None, 31, 31, 32)
Dropout | output: | (None, 31, 31, 32)
conv2 | input: | (None, 31, 31, 32)
Conv2D | output: | (None, 29, 29, 64)
norm2 | input: | (None, 29, 29, 64)
BatchNormalization | output: | (None, 29, 29, 64)
maxpool2 | input: | (None, 29, 29, 64)
MaxPooling2D | output: | (None, 14, 14, 64)
flatten1 | input: | (None, 14, 14, 64)
Flatten | output: | (None, 12544)
drop2 | input: | (None, 12544)
Dropout | output: | (None, 12544)
dense_out | input: | (None, 12544)
Dense | output: | (None, 15)
Total params:
207,375
Activations: ReLU
(output is linear)
Total params:
7,162,447

Figure 12. Two different CNN architectures that were evaluated. (left) small (right) large

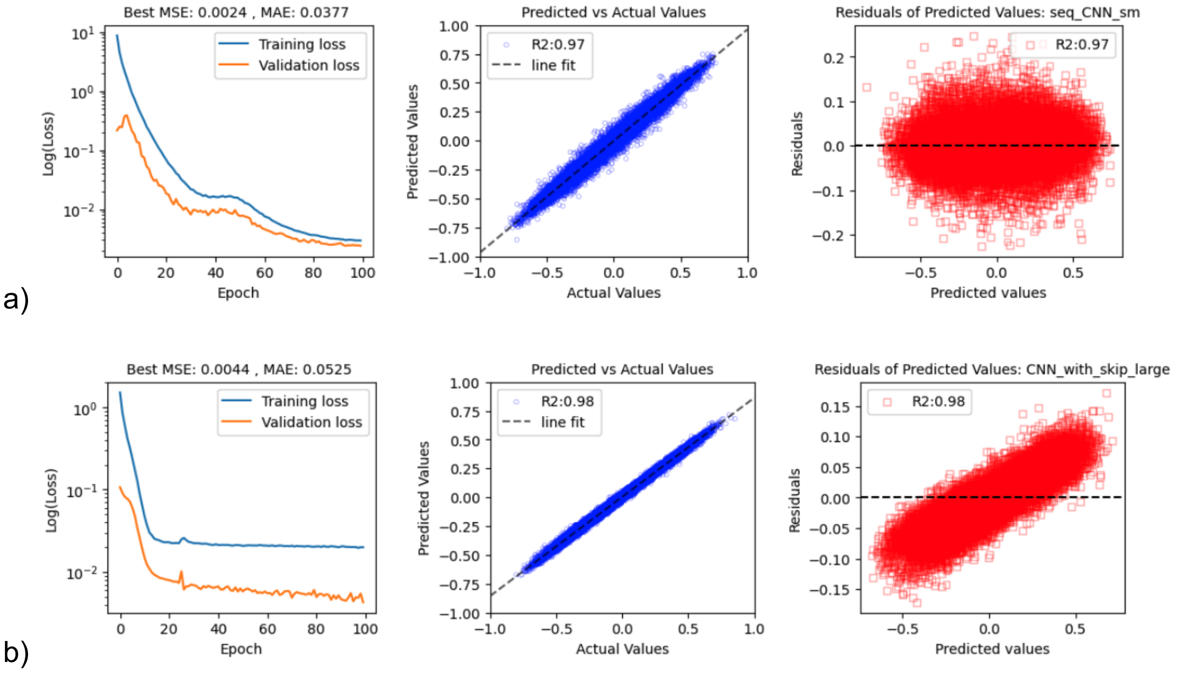


Figure 13. Training and validation results for (a) small and (b) large CNN architectures trained using the dataset of CFS4 image/encoding pairs losses (MSE). The middle scatterplot is for the actual vs. predicted values. (right) Residual plots

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

Table 1. Summary results for fitting CNNs with the different datasets of image/encoding pairs the third row and last row correspond with the results shown in **figure 13**.

Outcomes

To further gather appreciation of the ability for the CNN models to interpret the real X-ray images we initially make a more quantitative assessment using simulated hold-out test images as the benchmark data. Later, it will become apparent why it is not possible at this stage of development to make quantitative error metric based assessments based on the outputs and interpretations for real X-ray images. **Figure 14** depicts an example of performance of different CNN for a simulated X-ray image generated with the CFS4 encoding (D:15) which is the highest resolution used in this investigation. The capacity (metrics shown as **table 2**) to which the CNNs trained at lower resolution (D:3 and D:8) are able to interpret the D:15 image is remarkable. This is in support that CNN trained on simulated X-ray images should be able to provide some interpretation of a real image. It is expected that real X-ray images have several orders of magnitude resolution by way of structural information (after all, it is real data).

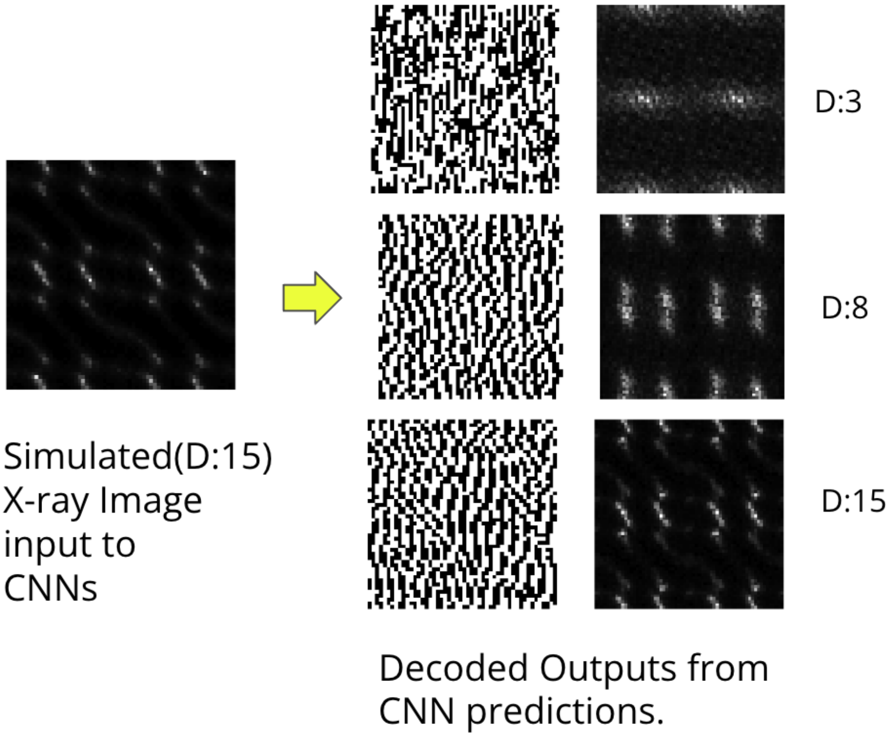

D:3
D:8
D:15
Simulated(D:15) X-ray Image input to CNNs
Decoded Outputs from CNN predictions.

Figure 14. CNN performance with the simulated holdout data. As the resolution of CNN used to interpret the data is increased we get a better reproduction of the input image via the decoding step.

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

Table 2. Error metrics between the simulated CFS4 image input into the trained CNN and decoded image which resulted from CNN interpretation. It is remarkable that details of the image with a finer-grained structure correspond with a better fit and is an indication that the CNN using this approach has some ability to generalise.

For evaluating the performance of the CNNs on real data, we demonstrate using the small arch. CNN trained on CFS4 image and encoding pairs (see **Table 1**) for reasons discussed earlier. The results for two real X-ray images are shown in **Figure 15**. It is remarkable how CNN, which has no concept of what real data looks like, is able to output an encoding that best describes its input. Once decoded, the simulated X-ray image resembles the observed data, and the CNN has effectively made what appears to be a successful interpretation of the real data. Especially for the image in the bottom row. It might have taken a lot of trial and error to just manually adjust the encoding vector with the correct balance of variables in order to get something that looked similar. Also, a global optimisation approach might have encountered problems if it was not preconditioned close enough to the correct solution due to getting stuck in local minima (of course, it is difficult to say exactly unless we try, and studies from the available literature indicate that this is often the case). Certainly, the encoding that is output from the CNN is a very good place to start prior to further parameter optimisation or other refinement approaches based on physical modelling or molecular simulation strategies. Further evaluation for other CNNs that were setup is made available in the jupyter notebooks provided.

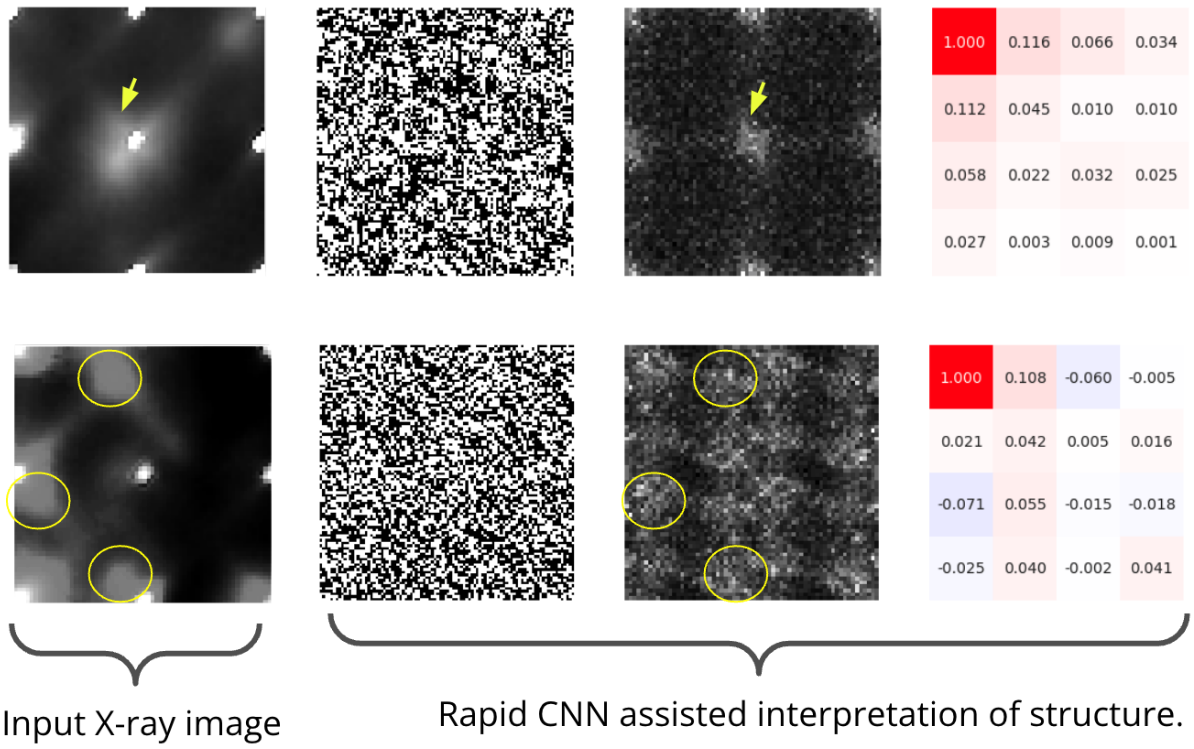


Figure 15. CNN performance at interpreting real X-ray data. Subplot rows correspond to CNN performance evaluation of two different X-ray images. Columns from left to right: observed X-ray image; simulated microscopic material structure abstracted via decoding; simulated X-ray image from the FT of the structure representation; CFS encoding directly output by the CNN as a matrix.

Implementation

We need to consider that the current CNNs evaluated were simplified and general in order to successfully establish a proof-of-concept. We can make a preliminary deployment of these CNNs online for public use and to receive further feedback. However, further consideration will be required to fill gaps and integrate the approach as shown in the data-science flow shown in **Figure 5** into a complete end-to-end business-to-data-science-to-business pipeline. It might be more important to perform a case study with a real industry compound that has X-ray images available that have not been correctly interpreted by a subject matter expert (SME). It must be demonstrated beyond reasonable doubt, perhaps requiring only a simple modification of the current scheme, that the CNN strategy can be impactful.

Other factors must also be considered, such as: “What will be the computing resource requirements for more sophisticated, higher-end use cases?" addressing implementation for better resolution (larger CNNs, better training data). Actual use cases will require testing and refinement with other known global optimisation methods incorporating molecular modelling considerations. There are also questions of how variational AE, cGANS, and deep Q learning might be implemented, and even consideration of if an AGI system can be implemented.

Data answer

The current model is successful at interpreting real observed data to a given resolution. It was difficult to predict in advance how it would perform on the X-ray signals from the real images based on only being trained to interpret simulated images. We have an answer concerning the lack of labelled data and where to begin with training the model with supplemented simulation data to improve the materials portfolio knowledge bucket. We have a remarkable proof of concept to build on for real-world case studies. The data question is solved, and we have a path forward for setting up an automated X-ray interpretation CNN-based device, thus creating more labelled data that can add value for our stakeholders.

Business answer

Implementing the rapid analysis data-science pipeline will result in knowledge of the structure details for specific desirable materials assets. We can begin to answer related questions such as because of this structure the material can be manufactured in a more profitable manner or monitored at a specified cost. Understanding how to enhance certain materials properties will result in high demands for the related products and increase net profits. Projected estimates are provided as **Table 3**.
The following assumptions are made:

- Instrument upkeep cost is unchanged because it is never fully utilised.
- Increasing X-ray image analysis by factor of 1 results in a materials design enhancement of 0.5.
- Cost of simulated data is negligible wrt. rest of the portfolio.
- Tradeoff. less SME and more non-SME staff.

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

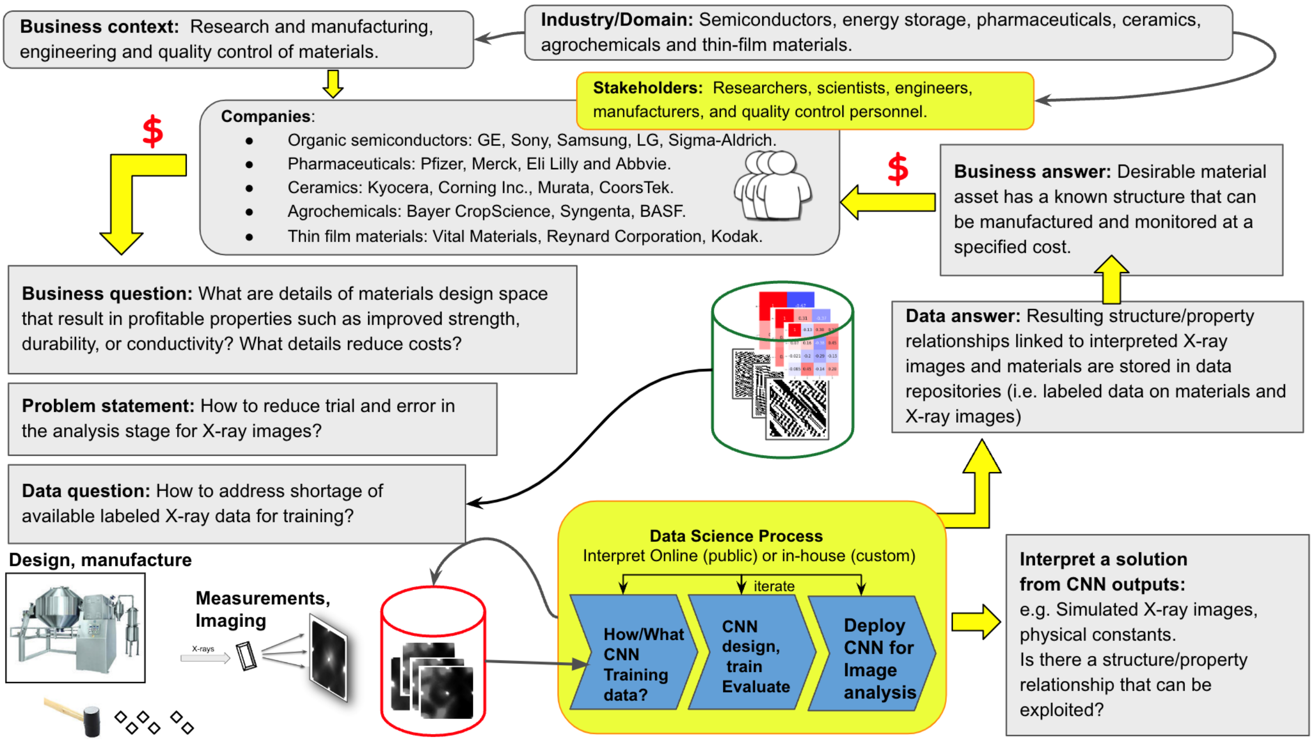
Table 3. Business case overview and estimated net profit.

Response to stakeholders

We have confirmed proof of concept in that CNN is capable of recognizing features of the images and output encodings that can simulate the structure.

End-to-end solution

A full end-to-end business→data science→business workflow is shown as **Figure 16**.



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?

Interpret a solution from CNN outputs: e.g. Simulated X-ray images, physical constants. Is there a structure/property relationship that can be exploited?

$
$
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

Figure 16. End-to-End Business Pipeline and Integrated Data Science Flow.

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.