BACHELOR'S THESIS · UNIVERSITY OF COPENHAGEN, FACULTY OF SCIENCE

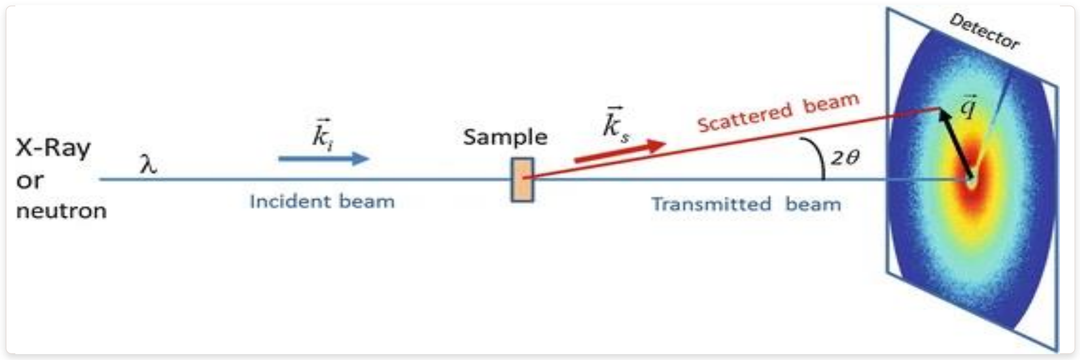
Generating Pair Distance Distributions from SAXS Data Using Machine Learning

Advisor: Andreas Haahr Larsen | June 2026 | github.com/Niklas03NS/BSc-Thesis

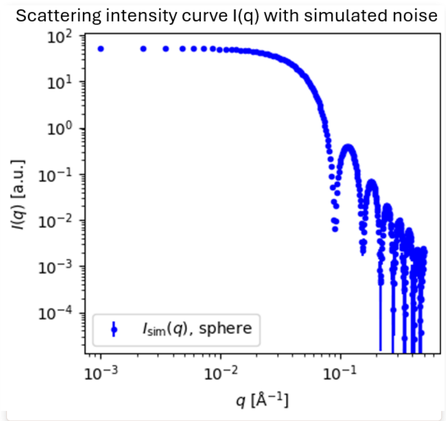
Niklas Sørensen

Ali Hussein

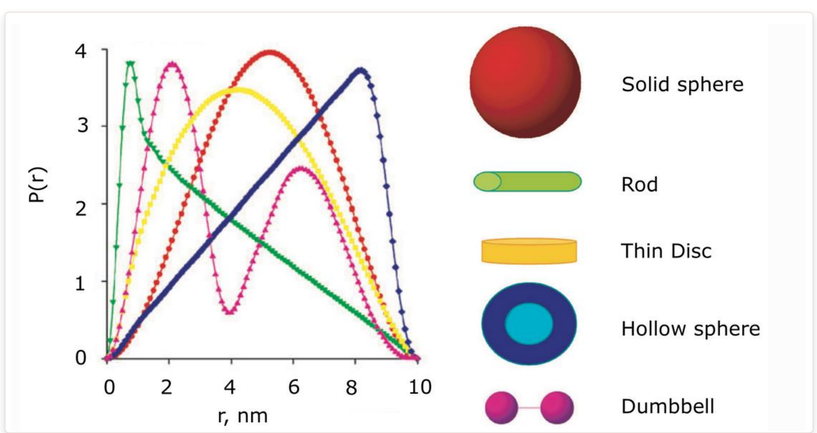
1 SAXS Experiment



- A collimated, monochromatic X-ray beam hits the sample; most photons pass straight through, while a small fraction scatter elastically off electron clouds in the sample.
- The scattered signal is recorded by a 2D detector, then azimuthally averaged into the 1D intensity curve I(q).
- Small-angle X-ray scattering (SAXS) probes the nanometer-scale structure of proteins, nanoparticles, andpolymers in solution.



2 Pair Distance Distribution



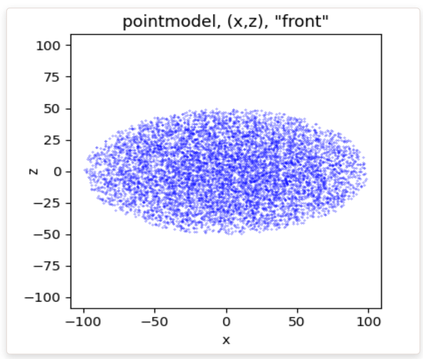
Characteristic p(r) fingerprints: solid sphere (smooth bell), rod (sharp peak + tail), thin disc, hollow sphere (steep rise/fall), dumbbell (double peak) — shape can often be read off the curve almost by eye.

- From intensity I(q) to pair distance distribution p(r): $I(q) = 4\pi \int p(r)\cdot\sin(qr)/(qr)\, dr$, integrated from r = 0 to Dmax
- The pair distance distribution p(r), recovered from the scattering curve I(q), encodes particle size, shape, and maximum dimension Dmax, radius of gyration
- Recovering p(r) normally requires an Indirect Fourier Transform (IFT): an ill-posed inversion that needs iterative regularization.
- A valid p(r) is smooth, positive, and falls to zero at r = 0 and r = Dmax
- Neural networks can already predict single structural numbers — Rg, Dmax, molecular weight — directly from I(q), near-instantly once trained.
- This project asks: can a network reconstruct the entire p(r) curve from I(q), and how does that depend on model choice and shape complexity?

3 Methods

Data generation

- Shape2SAS builds 3-D point-cloud particles for 12 geometric shapes (1–4 free shape parameters) and computes matching I(q) and ground-truth p(r).
- Stratified sampling covers each parameter range evenly; a noise model adds a realistic noise to I(q) here using the naming convention I_sim(q).



Preprocessing

- Fixed 512-point q-grid, max-intensity normalization, and a log transform to emphasize the weak high-q signal.

Models

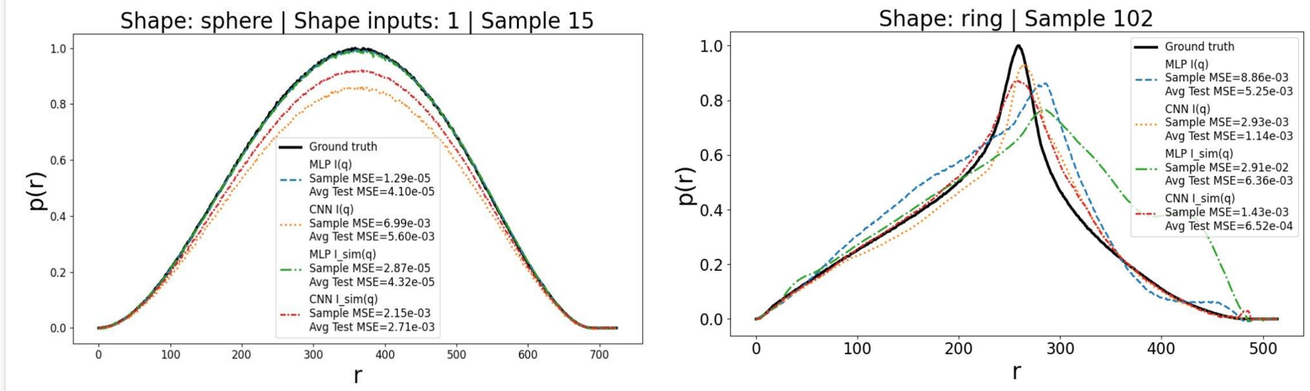
- MLP — up to 3 hidden layers × 1024 units, ReLU, dropout — treats all 512 q-points as independent features.
- 1D CNN — up to 3 conv. blocks (64→128→256 ch.), kernel 3 or 5, max-pooling — exploits locality along the curve.

Training & evaluation

- 70% / 15% / 15% train / validation / test split; MSE loss; Adam optimizer.
- Grid search over depth/width, dropout {0, 0.1}, learning rate {1e-3, 3e-4}, and weight decay {0, 1e-5} for both architectures.

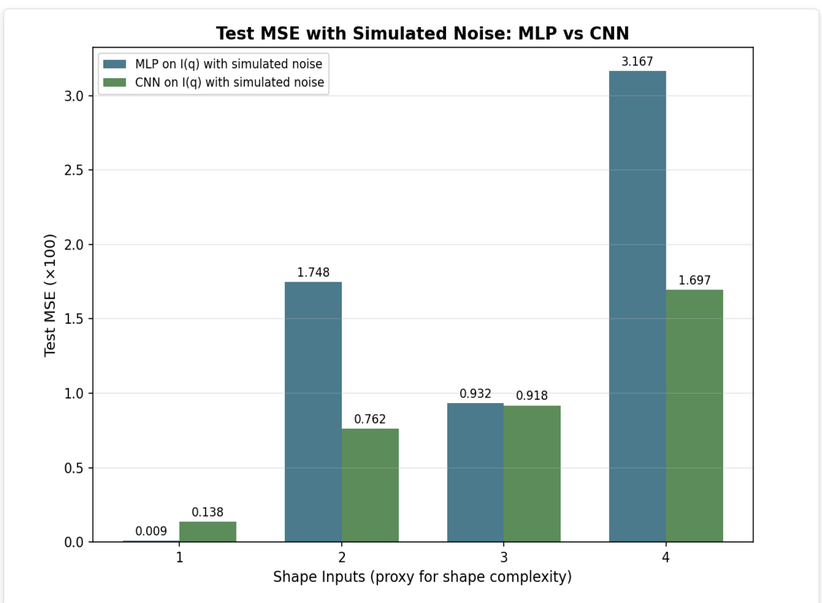
4 Results

4a · Prediction accuracy vs. shape complexity



Predicted vs. ground-truth p(r), the two figures show predictions of a sphere (left) with 1 degree of freedom, and an ellipsoid (right) with 3 degrees of freedom.

- As the number of shape parameters grows from 1 to 4, test-set error rises for both models, and predicted peaks become noticeably smoother than the ground truth.
- The MLP is most accurate on simple, smooth shapes such as spheres; the CNN better preserves sharp or asymmetric peaks on complex shapes such as the hollow sphere and superellipsoid.



4b · More training data helps the CNN

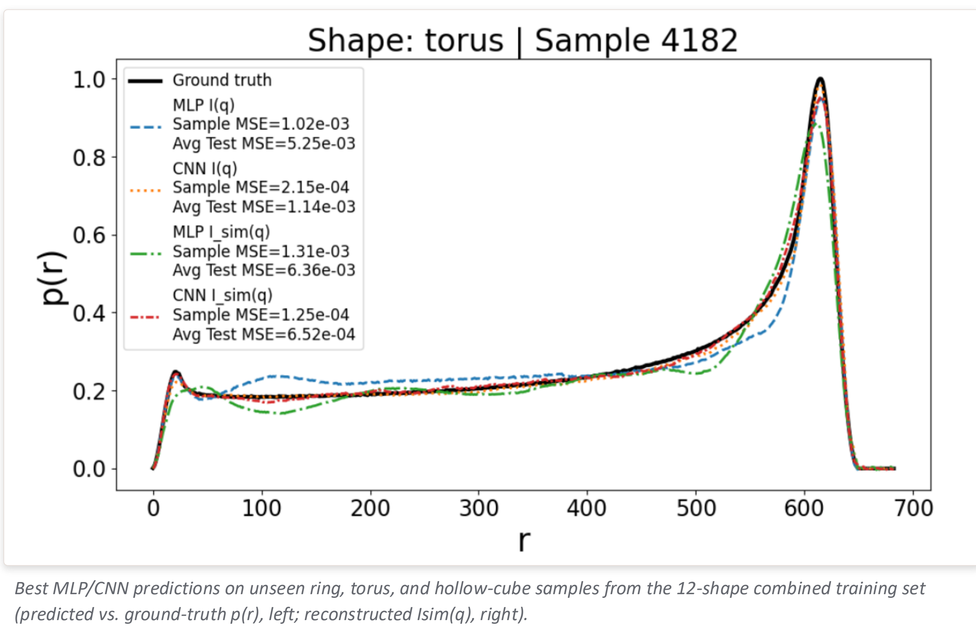
- Expanding the superellipsoid training set from 1,000 to 9,000 samples cut test-set MSE substantially for both architectures on this hardest, 4-parameter shape:

28–35% lower test MSE — MLP

77–81% lower test MSE — CNN

- The CNN's convolutional inductive bias appears to need more data to pay off fully — but scales far better once it has it.

4c · Best model on the full combined-shape dataset



Best MLP/CNN predictions on unseen ring, torus, and hollow-cube samples from the 12-shape combined training set (predicted vs. ground-truth p(r), left; reconstructed Isim(q), right).

- Grid search: the deepest / widest networks with no dropout won for both architectures — capacity, not overfitting, was the limiting factor.
- The CNN (64, 128, 256 channels, kernel 5) reached the lowest test MSE for both I(q) and Isim(q) — roughly 5–10× lower than the best MLP.

Model	Architecture	Best hyperparameters	Test MSE
MLP · Isim(q)	(1024,1024,1024)	dropout 0 · lr 1e-3 · wd 1e-5	0.00636
CNN · Isim(q)	(64,128,256), k=5	dropout 0 · lr 3e-4 · wd 0	0.00065

5 Discussion & Conclusion

What we have learned

- Both the MLP and 1D CNN learn a physically meaningful I(q)→p(r) mapping. After training, inference is near-instantaneous compared with iterative IFT fitting although more accurate results are still achieved that way.
- The CNN's local receptive fields and weight sharing preserve sharp, asymmetric features; the MLP — combined with ReLU, MSE loss, and Adam — tends to smooth and average, favoring simple, symmetric shapes.
- On the combined dataset, larger networks with little or no regularization performed best: the task was capacity-limited, not overfitting-limited.

Limitations & future work

- Only idealized geometric shapes were used, not real experimental data.
- No Dmax estimation: p(r) was scaled using the ground-truth Dmax, which makes the reported errors optimistic.
- The hyperparameter grid search was limited to a single GPU (≈ 6 h on a free Google Colab T4).
- Next steps: real SAXS data (SASBDB, Simple Scattering), joint Dmax prediction, larger architectures, and a shape-classifier that routes each curve to whichever model suits it best.

Machine learning can recover physically meaningful pair distance distributions directly from SAXS intensity curves — a fast, promising complement to conventional indirect Fourier transform analysis, with 1D CNNs offering the strongest balance of accuracy and robustness for this inverse problem.