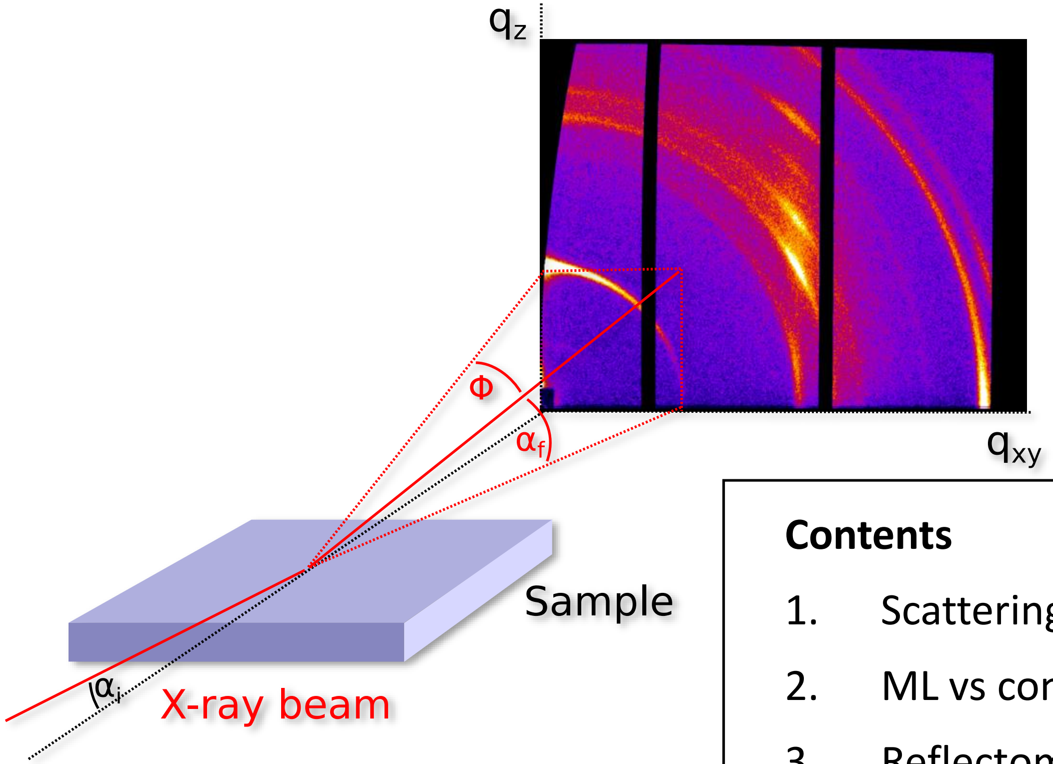
Machine learning for reflectometry: Concepts, applications, and challenges

Vladimir Starostin and Frank Schreiber
http://www.soft-matter.uni-tuebingen.de
V. Munteanu, C. Völter, M. Romodin, D. Lapkin, V. Herbst, M. Hylinski, D. Baláž, A. Greco, L. Pithan, A. Gerlach, A. Hinderhofer

special thanks to
Kowarik group (Graz)
Murphy group (Kiel)

synchrotrons
neutron facilities
ML excellence cluster

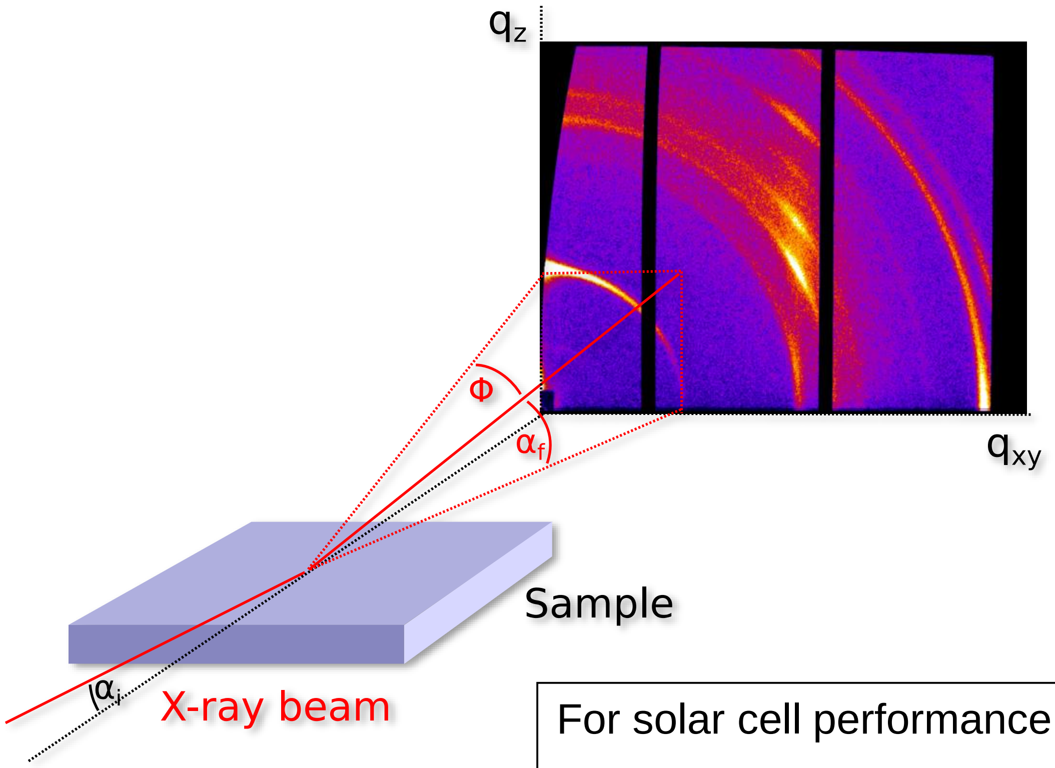


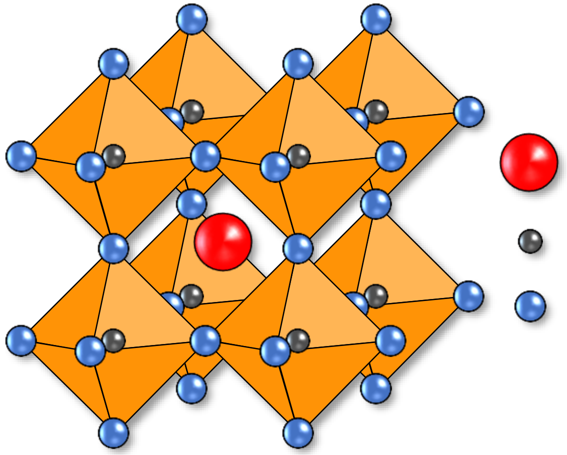
Contents

1. Scattering (X-rays and neutrons) and data acquisition rates
2. ML vs conventional data analysis
3. Reflectometry (XRR/NR) and specific challenges (“1D”)
4. Grazing-incidence diffraction and specific challenges (“2D”)
5. ML packages mlreflect and gixi

overall ambition: connect ML to physical understanding of scattering

Scattering from soft and hybrid materials



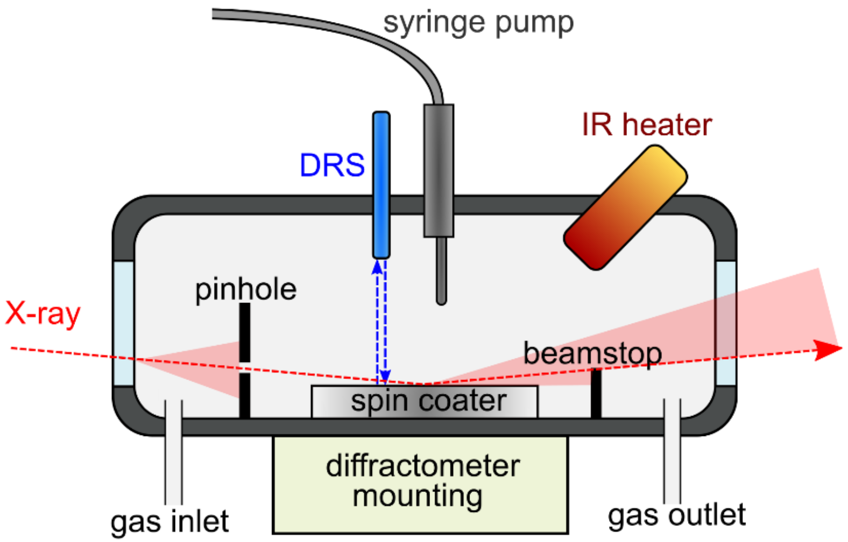


ABX$_3$

A = MA, FA

B = Pb^{2+}

X = Cl$^-$, Br$^-$, I$^-$



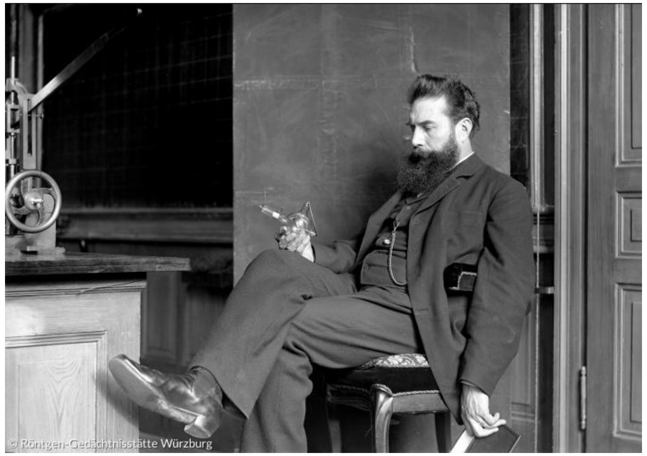
For solar cell performance (molecular or hybrid), the exact structure is key!

Applies to many other soft / bio-related systems!

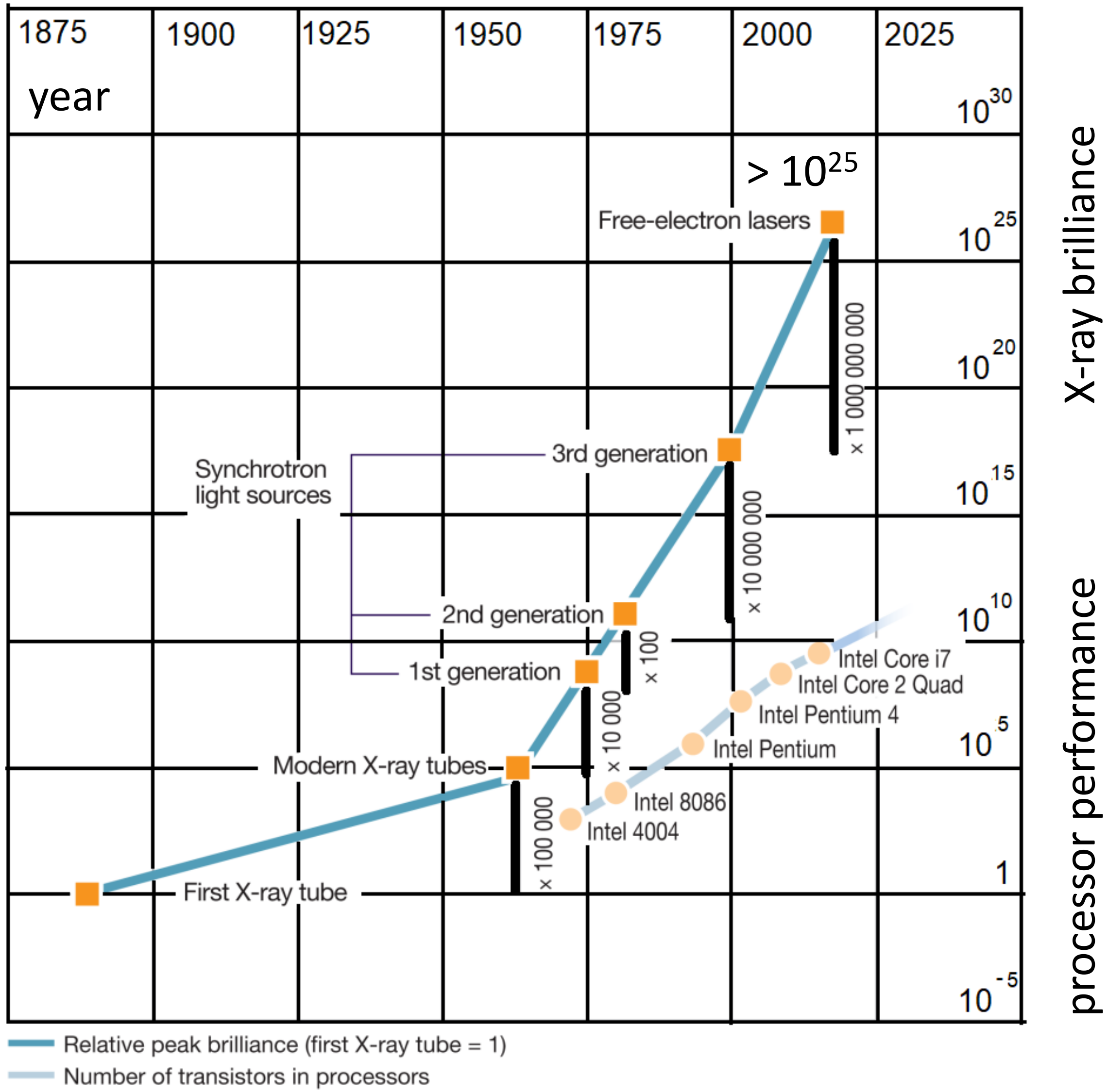
Do real-time scattering during crystallization!

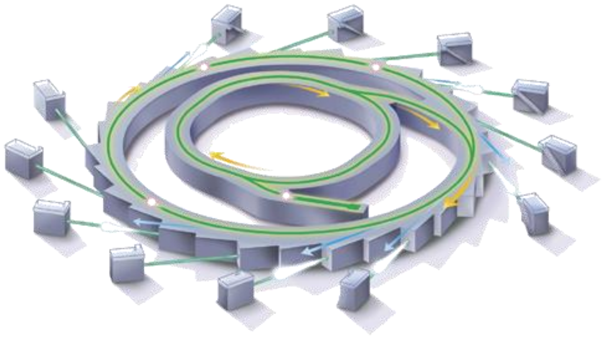
Arora et al. Science 358 (2017) 768
Greco et al. J. Phys. Chem. Lett. 9 (2018) 6750
Brinkmann et al. Nature 604 (2022) 280

X-ray technology is outpacing Moore’s law



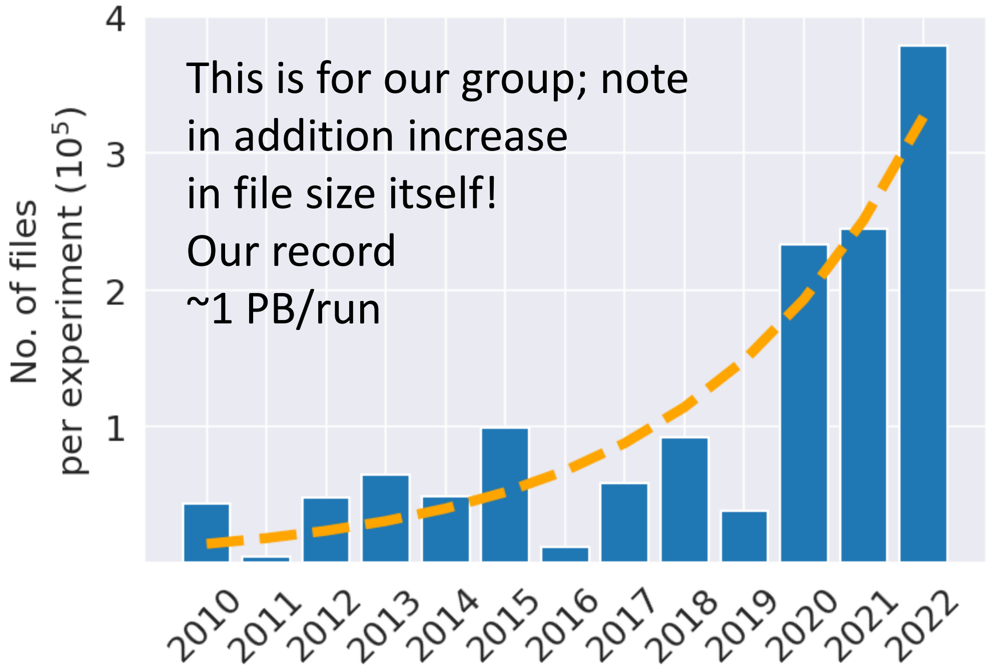
© Röntgen-Gedächtnisstätte Würzburg



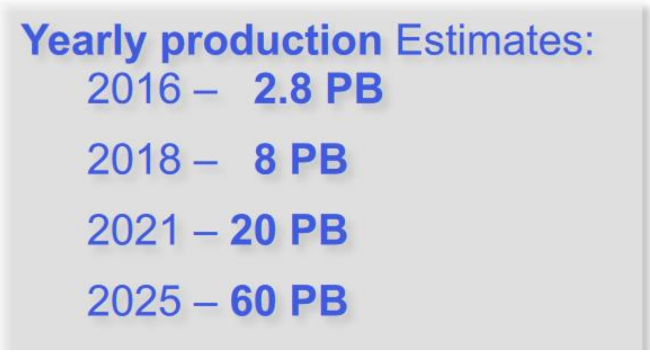


DESY, Hamburg
XFEL, Hamburg
ESRF, Grenoble
DLS, Oxfordshire
Soleil, Paris
APS, Chicago
ALS, Berkeley
…

Data acquisition



This is for our group; note in addition increase in file size itself!
Our record
~1 PB/run



Yearly production Estimates:

2016 – **2.8 PB**

2018 – **8 PB**

2021 – **20 PB**

2025 – **60 PB**

This is for the ESRF; other facilities similar; storage need:
10 years raw data
after 3 years public

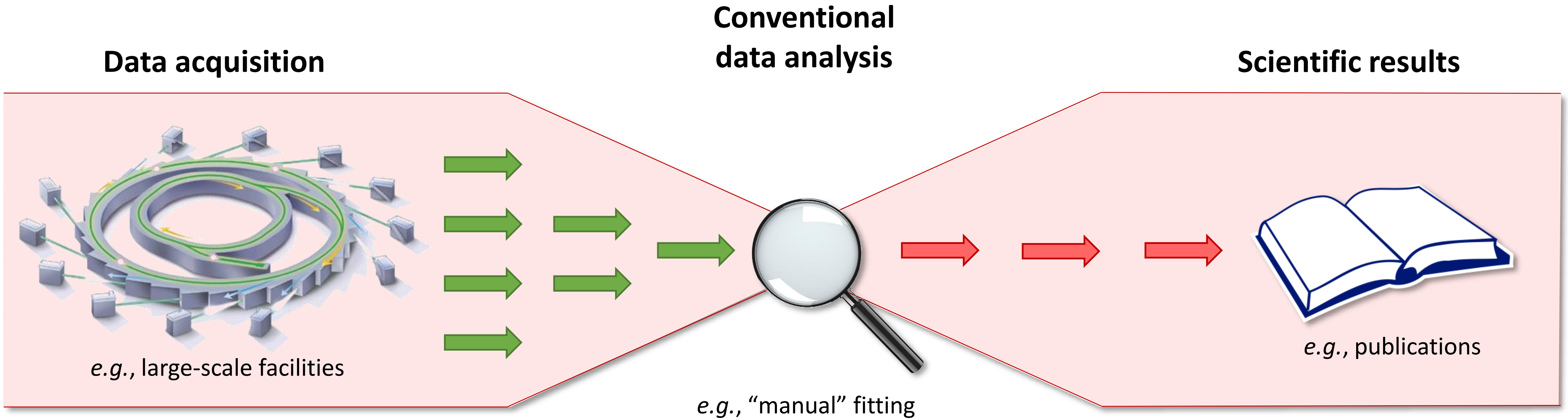
Improvements in measurements

- Better sources (brilliance, coherence)
- Better detectors (area detectors with high resolution/framerate (>100 MB/s))
- New/advanced experimental setups

New initiatives for handling data

- National Science Data Infrastructure (NFDI) (includes KFS and KFN)
- Backed by about 5000 PhDs + students
- DAPHNE4NFDI consortium

Data acquisition is outpacing data analysis



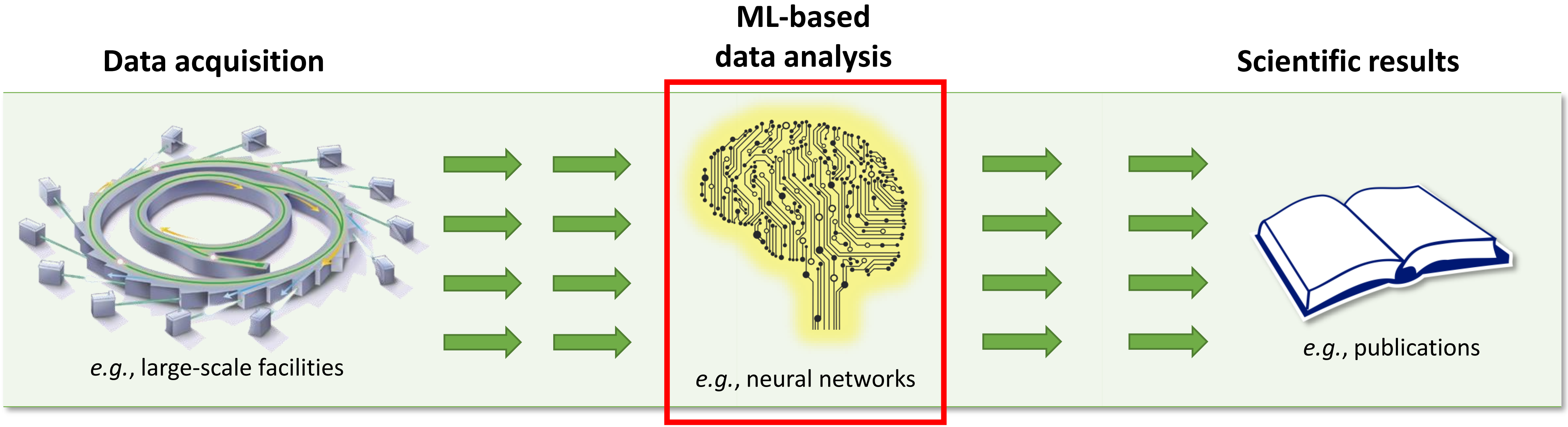
Improvements in measurements

- Better sources (brilliance, coherence)
- Better detectors (area detectors with high resolution/framerate (>100 MB/s))
- New/advanced experimental setups

New initiatives for handling data

- National Science Data Infrastructure (NFDI) (includes KFS and KFN)
- Backed by about 5000 PhDs + students
- DAPHNE4NFDI consortium

ML-based methods can help avoid bottlenecks



Data acquisition

e.g., large-scale facilities

ML-based data analysis

e.g., neural networks

Scientific results

e.g., publications

Improvements in measurements

- Better sources (brilliance, coherence)
- Better detectors (area detectors with high resolution/framerate (>100 MB/s))
- New/advanced experimental setups

New initiatives for handling data

- National Science Data Infrastructure (NFDI) (includes KFS and KFN)
- Backed by about 5000 PhDs + students
- DAPHNE4NFDI consortium

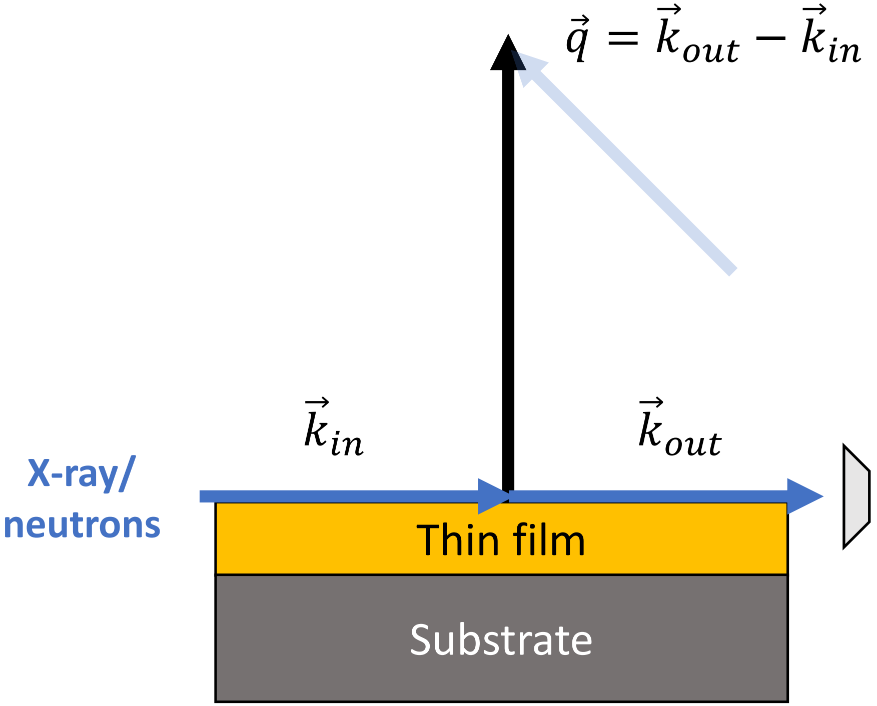
X-ray and neutron reflectivity (XRR/NR)

Greco et al. J. Appl. Crystallogr., 2019, 52, 1342

X-ray and neutron reflectivity (XRR/NR)

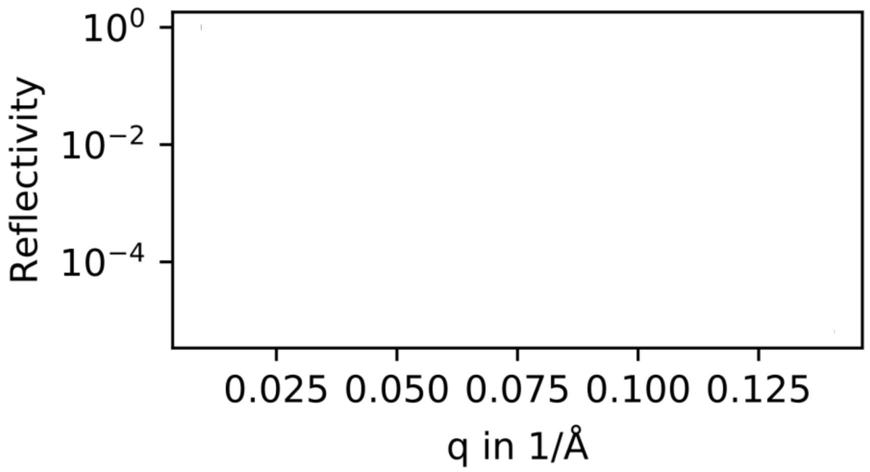
Experiment

X-ray beam at certain discrete angles



Data

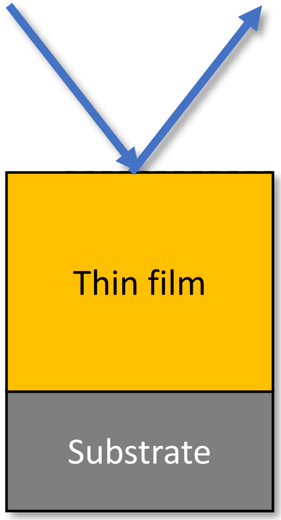
Reflected beam intensity for each angle



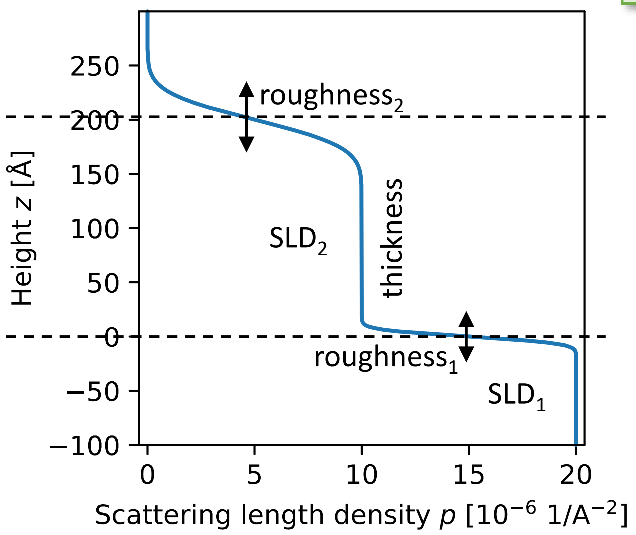
Shape of reflectivity curve provides information about thin film properties

Greco et al. J. Appl. Crystallogr., 2019, 52, 1342

Characterizing samples with XRR/NR

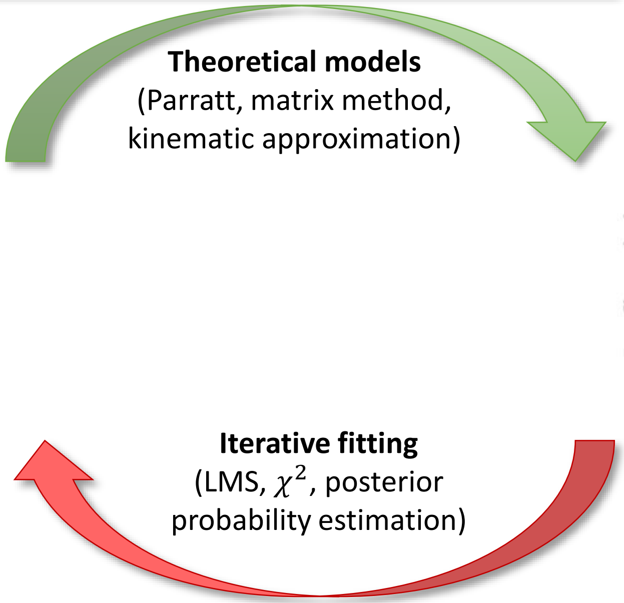


SLD profile $\rho(z)$



$$R(q_z) \propto q_z^{-4} \left| \int \frac{d\rho(z)}{dz} e^{iq_z z} dz \right|^2$$

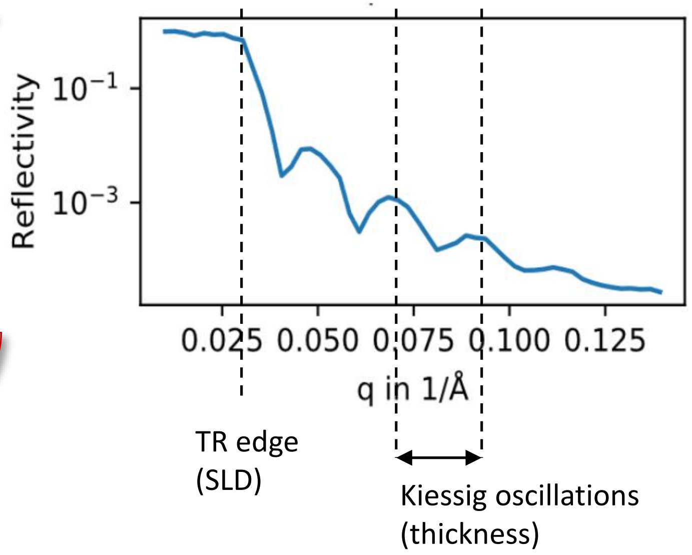
Fourier transform
with phase loss!



Theoretical models
(Parratt, matrix method,
kinematic approximation)

Iterative fitting
(LMS, χ^2, posterior
probability estimation)

Measured reflectivity curves $\boldsymbol{R}(\boldsymbol{q};\boldsymbol{p})$



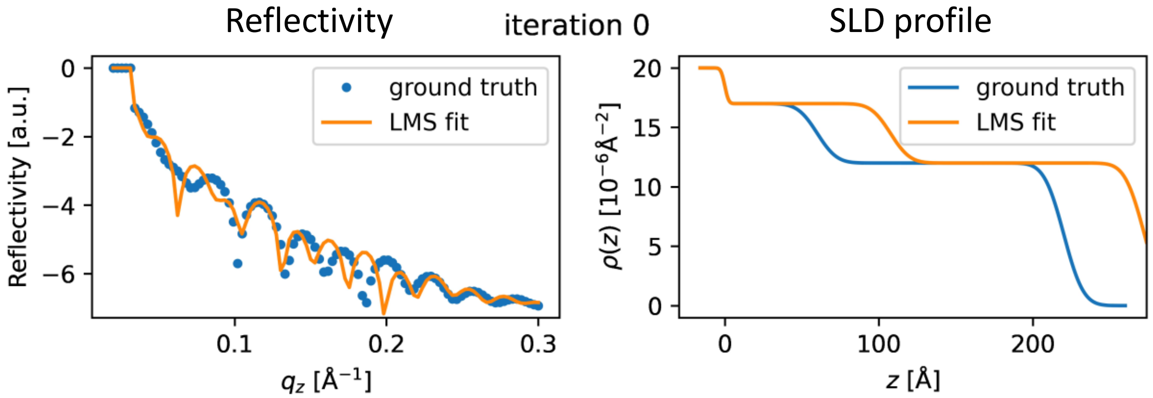
No analytical
“back-transformation”!

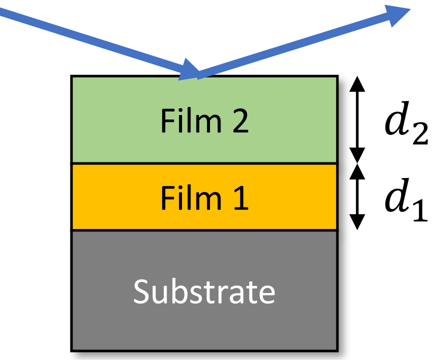
SLD profile usually parameterized, *i.e.*
$\rho(z) = \rho(z; \boldsymbol{p})$, with $\boldsymbol{p}$ = (thickness, roughness, SLD)

Greco et al. J. Appl. Crystallogr., 2019, 52, 1342

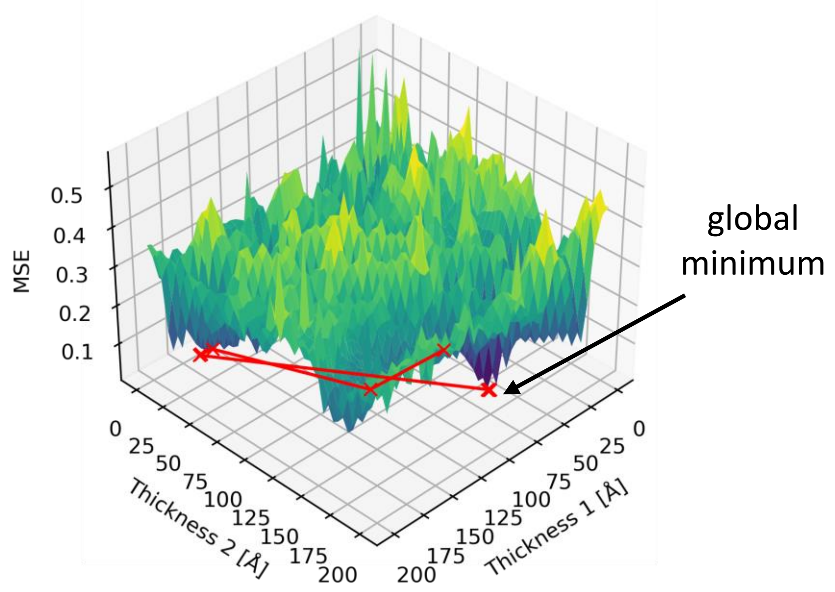
Conventional LMS fit

Example: Least mean squares fit with (only) two open parameters





Multi-dimensional mean squared error surface

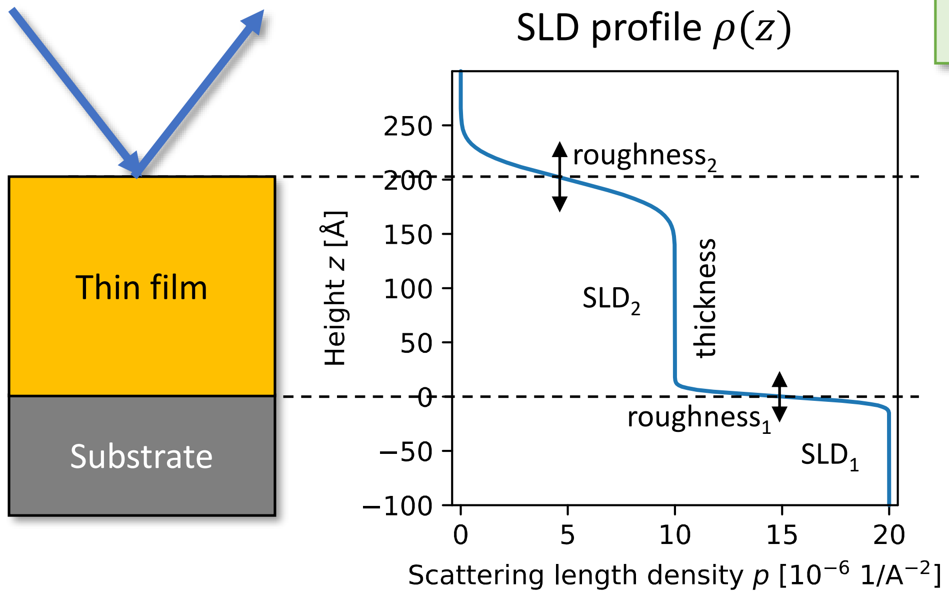


- Conventional approach: iterative fitting algorithms
- Stochastic algorithms (e.g. differential evolution) usually find a good minimum
- However, fitting boundaries must often be adjusted manually!

Iterative fitting is often slow and requires human expertise!

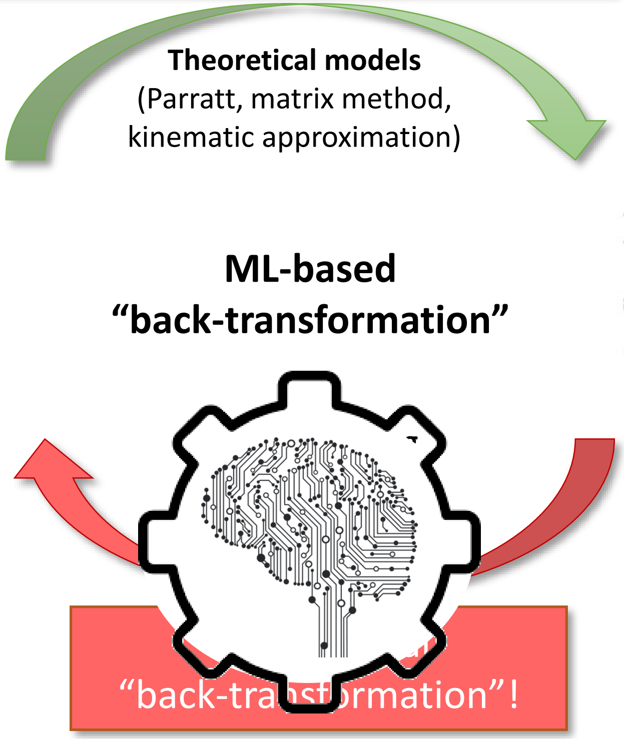
Greco et al. J. Appl. Crystallogr., 2019, 52, 1342

ML: Modeling of the “back-transformation”



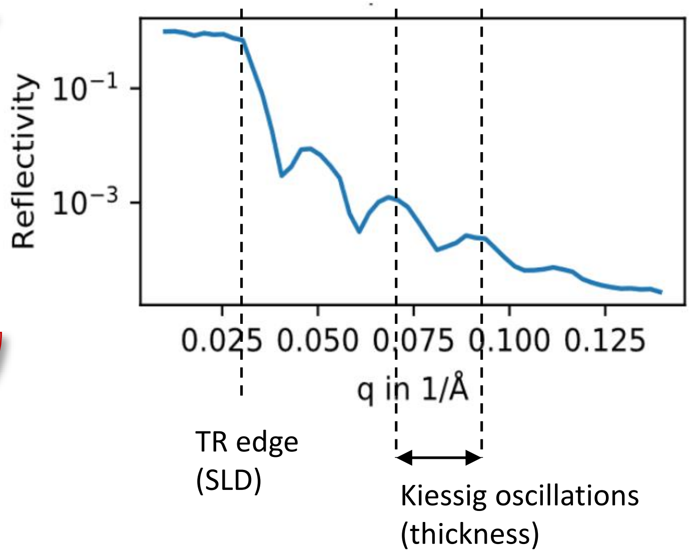
$$R(q_z) \propto q_z^{-4} \left| \int \frac{d\rho(z)}{dz} e^{iq_z z} dz \right|^2$$

Fourier transform with phase loss!


Theoretical models
(Parratt, matrix method, kinematic approximation)

ML-based “back-transformation”

Measured reflectivity curves $\boldsymbol{R}(\boldsymbol{q};\boldsymbol{p})$



SLD profile usually parameterized, *i.e.*
$\rho(z) = \rho(z;\boldsymbol{p})$, with $\boldsymbol{p}$ = (thickness, roughness, SLD)

Greco et al. J. Appl. Crystallogr., 2019, 52, 1342

Different approaches to inverse problem with ML

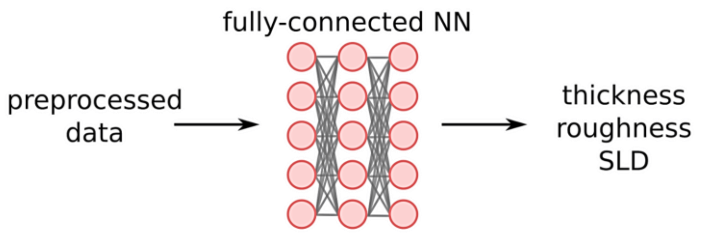
In general, inverse problem in reflectometry is ill-posed due to *the phase problem* → *possible multimodal solutions*

***Point estimators**. To avoid ambiguity, **the task should be narrowed down to specific cases** (e.g., silicon + silicon oxide + organic layer)*

***Probability density estimators.** Resolves the ambiguity issue*

e.g.:

Regression via NNs (MLP, CNN, …)


fully-connected NN

preprocessed data

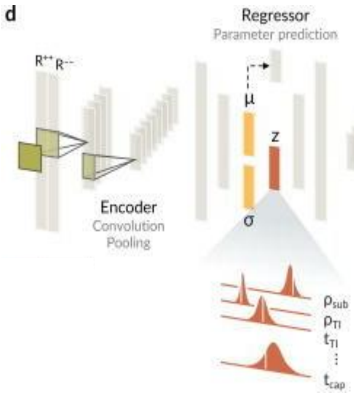
thickness
roughness
SLD

Greco et al. J. Appl. Cryst. 2022, 55, 362
Greco et al. Mach. Learn.: Sci. Technol. 2021, 2, 045003
Greco et al. J. Appl. Cryst. 2019, 52, 1342

- Simple implementation
- Fast inference

- Fails on multimodal cases
- Does not account for parameter distribution
- Does not provide error bars / uncertainty estimation
- Should be retrained for different use cases

Variational Autoencoders (VAE)



Andrejevic et al. Appl. Phys. Rev. 2022, 9, 011421
Timmermann et al. J. Appl. Cryst. 55 (2022) 751 (XPCS)

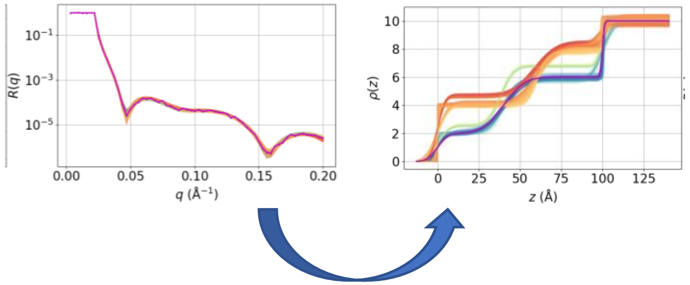
- Reduces data dimensionality

Normalizing Flows

neural posterior estimation

Experimental curve

Sampled SLD profiles



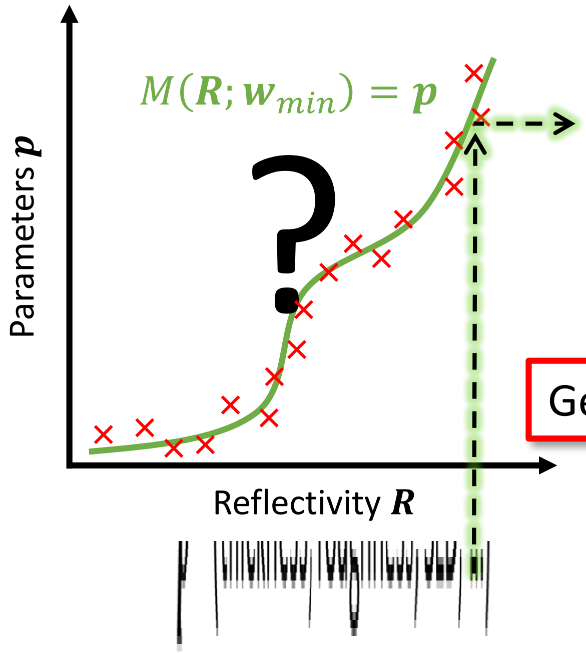
Sample profiles via conditional inverse Normalizing Flows transformation

Starostin et al., in preparation

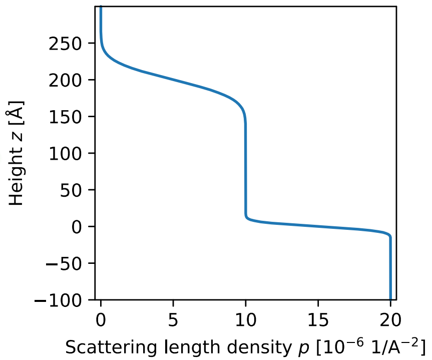
- Accelerated Bayesian analysis
- Resolves ambiguity problem
- Provides error bars
- No retraining required
- More difficult to implement

Neural networks can approximate “back-transform”

How do we find a heuristic model for the “back-transform”?



$M(\boldsymbol{R}; \boldsymbol{w}_{min}) = \boldsymbol{p}$
?
Parameters $\boldsymbol{p}$
Reflectivity $\boldsymbol{R}$

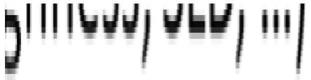


Get $\boldsymbol{p}$ for any $\boldsymbol{R}$ within ~1ms!

$\boldsymbol{p}$ = (thickness, roughness, SLD, …)

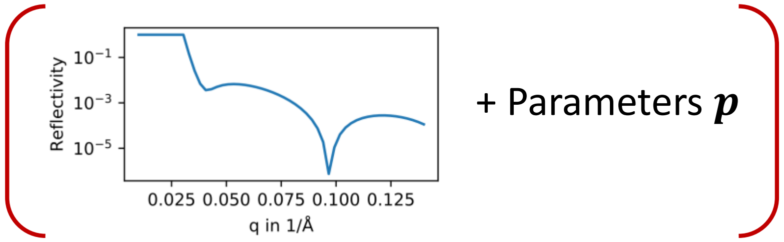
1. Define a neural network architecture $M(\boldsymbol{R}; \boldsymbol{w}) = \boldsymbol{p}$

e.g. Multi-layer perceptron



2. Generate training data, *i.e.* regression targets

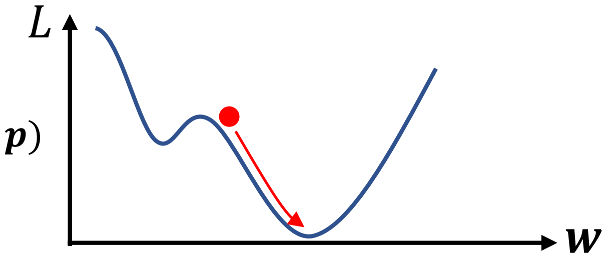
$N \times$



+ Parameters $\boldsymbol{p}$

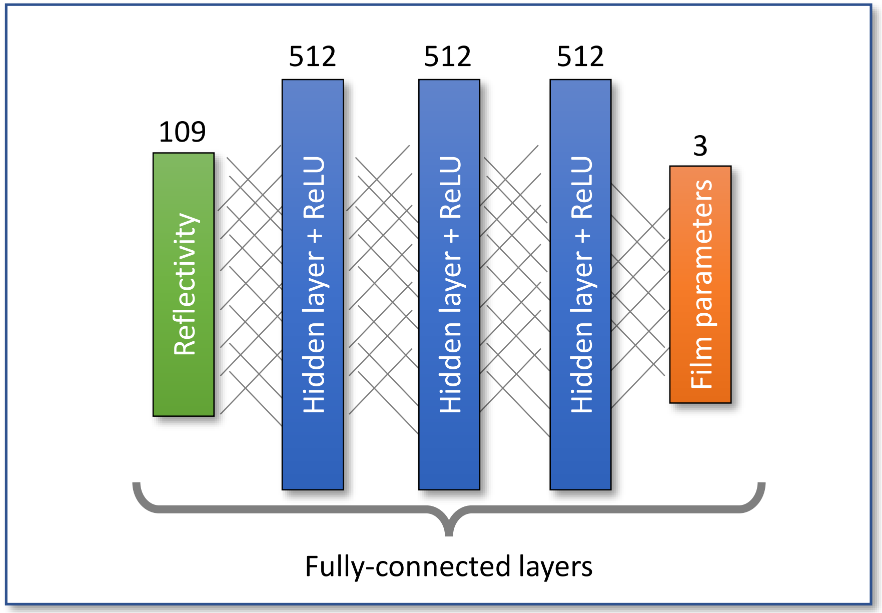
3. Perform training, *i.e.* non-linear regression

Training “loss”:
$L = \mathrm{MSE}(M(\boldsymbol{R}; \boldsymbol{w}), \boldsymbol{p})$



Choice of hyperparameters

Number and size of layers



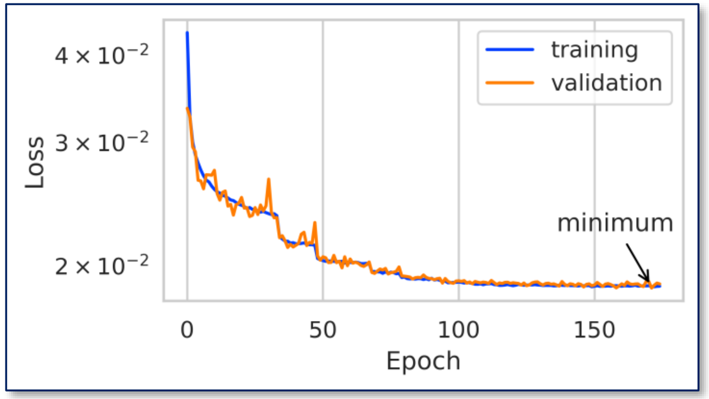
≙ about 560.000 trainable parameters $\boldsymbol{w}$

Hyperparameters

Training set size	300k, 100k, 3M, …
Mini-batch size	512, 256, 1024, …
Activation functions	ReLU, sigmoid, tanh, …
Initialization	Glorot uniform, normal, …
Optimizer	Adam, RMSprop, …
Initial learning rate	10^{-3}, 10^{-2}, 10^{-4}, …
LR schedules	reduce on plateau

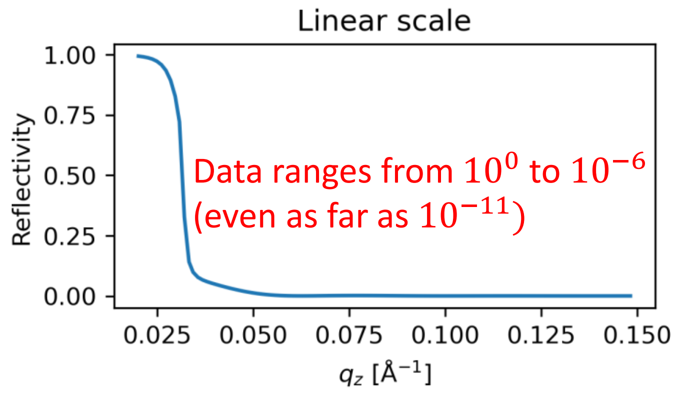
- Network size was reduced until loss was affected (larger models performed similarly)
- Hyperparameters were chosen empirically based on the lowest achieved validation loss

Loss was strongly affected by the input data!



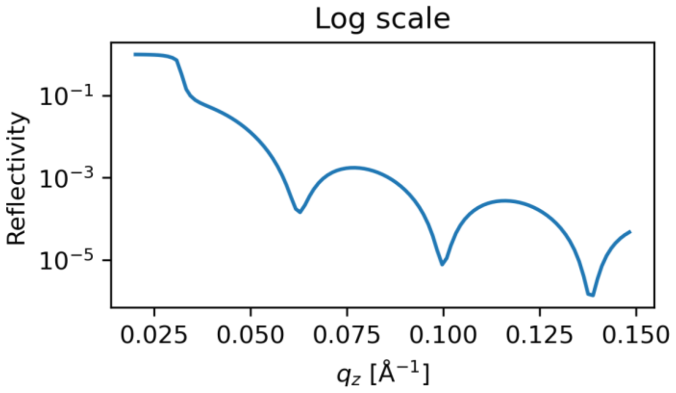
Challenges for ML specific to reflectometry

1. High dynamic range

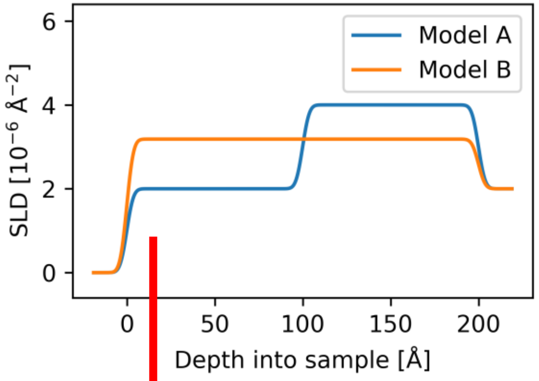


Data ranges from 10^0 to 10^{-6} (even as far as 10^{-11})

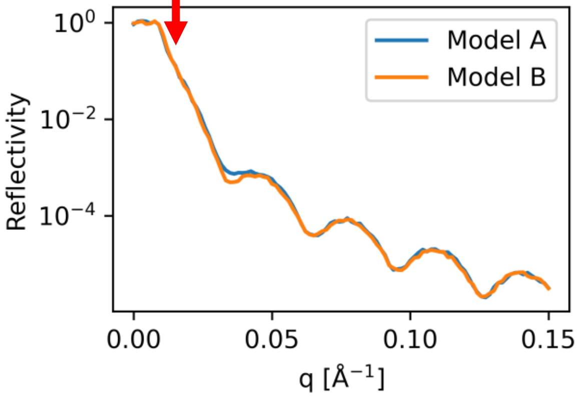
Log scale can help, but inputs are still not equally distributed



2. Phase problem/ambiguity

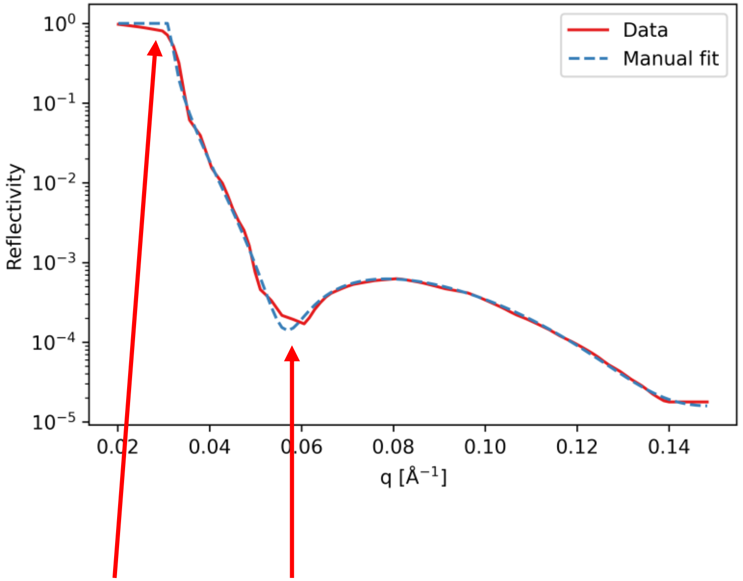


Different samples create the same reflectivity signal!



3. Experimental artifacts

The neural network is trained with simulated data, but meant to be used with experimental data!



Experiment and theory do not follow the same distribution!

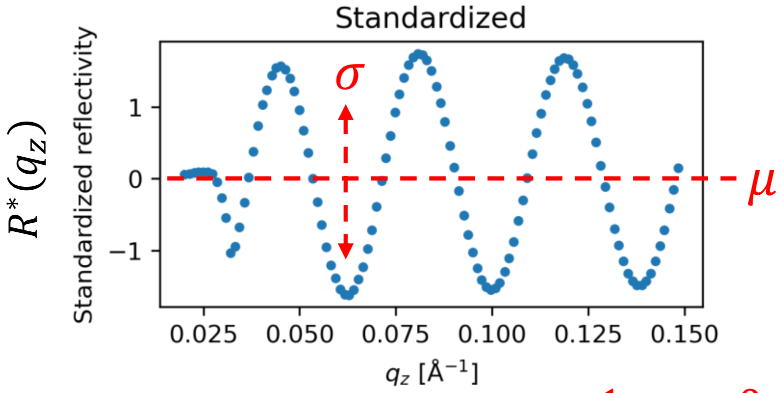
Solutions for reflectometry-related challenges

1. High dynamic range

Standardize input

$$R^*(q_z) = \frac{R(q_z) - \bar{R}(q_z)}{\hat{R}(q_z)}$$

- $\bar{R}(q_z)$: mean
- $\hat{R}(q_z)$: standard deviation
- derived from training set with artificial noise

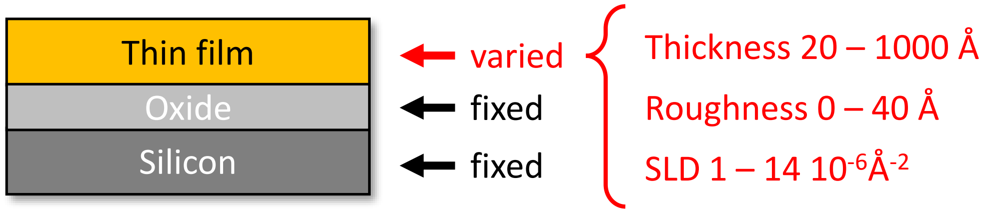


$\sigma = 1, \mu = 0$

2. Phase problem/ambiguity

Reduce solution space

E.g., 3 thin film parameters with a certain range



Thin film ← varied

Oxide ← fixed

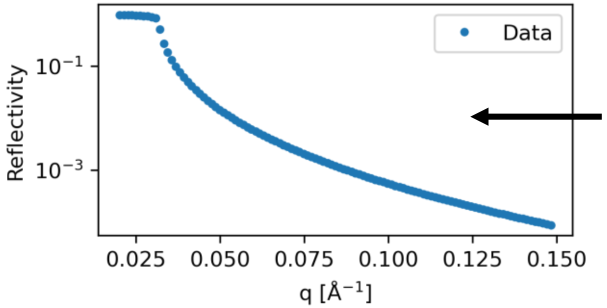
Silicon ← fixed

Thickness 20 – 1000 Å

Roughness 0 – 40 Å

SLD 1 – 14 10^{-6}Å^{-2}

Remove “featureless” curves



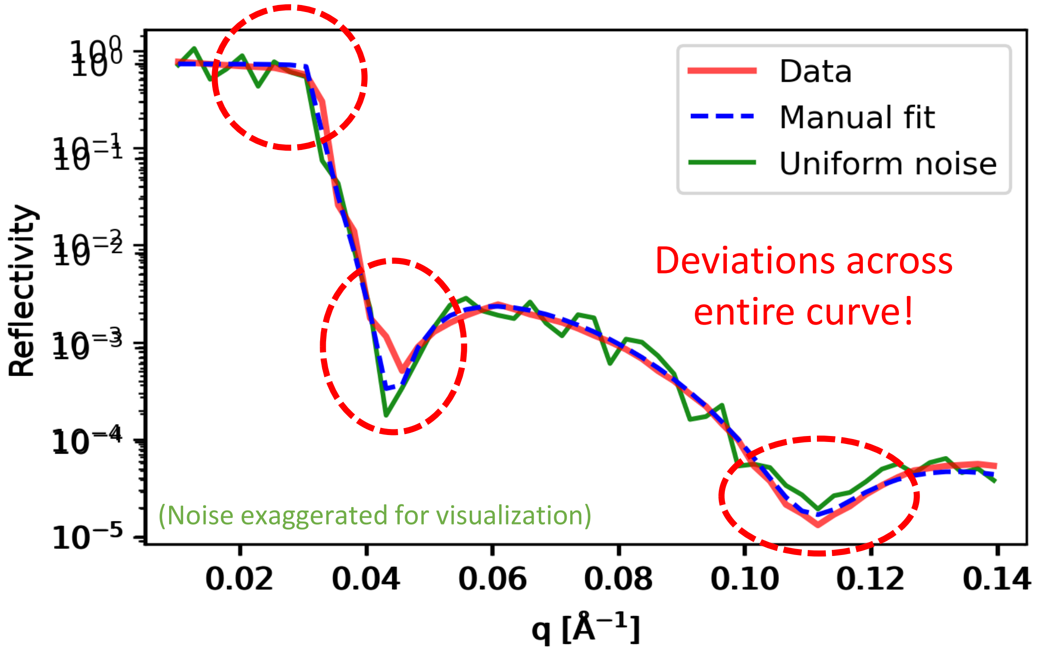
Exclude from training:

- Low thickness: < 20 Å
- Low contrast: < 10^{-6} Å^{-2}
- High roughness: > 40 Å

Greco *et al. Mach. Learn.: Sci. Technol.*, 2021, **2**, 045003

Optimizing the noise of the training data

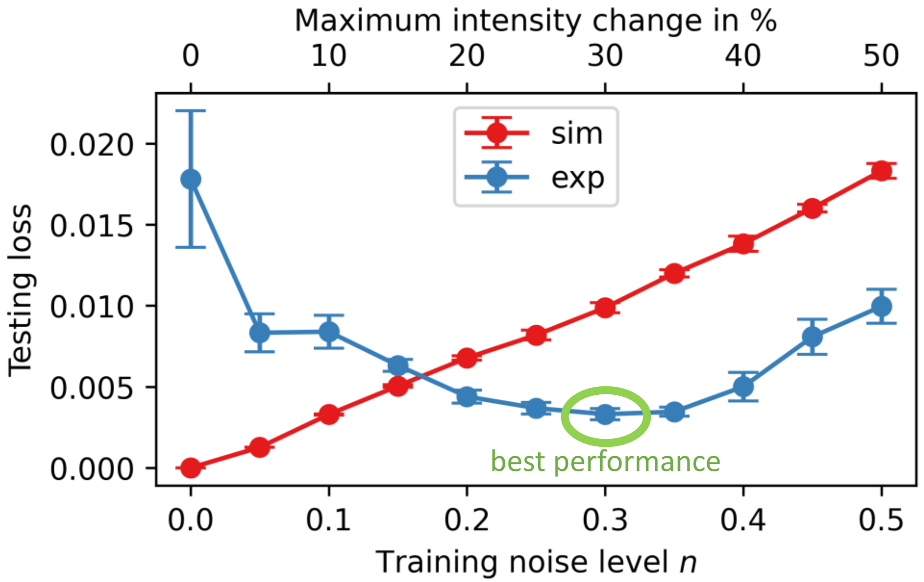
3. Experimental artifacts



Add noise that acts independently of q_z!

E.g. uniform noise:
each input multiplied with a value between $1 - n$ and $1 + n$

Fitting performance vs. training noise



- Trained 11 different neural networks with increasing noise level n on training data
- Applying noise decreases loss by a factor of 3!

Greco *et al. J. Appl. Crystallogr.*, 2019, **52**, 1342-1347

Improving performance through input resampling

- Systematic errors (*e.g.* misalignment, footprint correction, slit convolution) can impact prediction quality
- Resampling the data can help minimize this

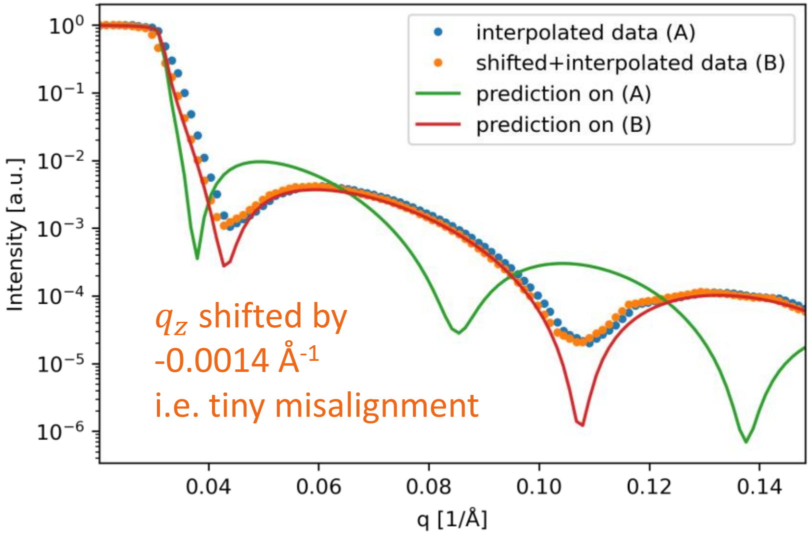
Resampling using q_z shifts:

1. Interpolate the data for many small q_z shifts
2. Predict parameters using neural network
3. Calculate MSE between data and prediction
4. Pick parameters/shift with the lowest MSE

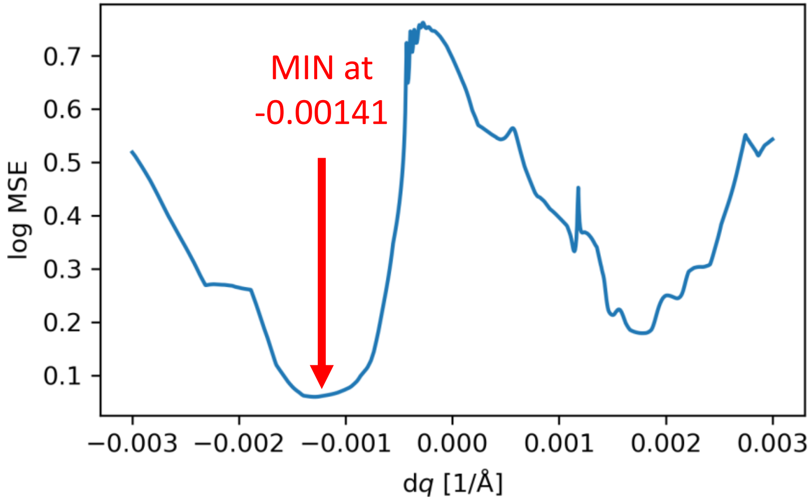
Neural network speed can be exploited to evaluate 1000 different q_z shifts within less than a second!

Resampling method implemented in *mlreflect*!

PEN on Si/SiO$_x$



q_z shifted by -0.0014 Å^{-1}
i.e. tiny misalignment



MIN at
-0.00141

The *mlreflect* package

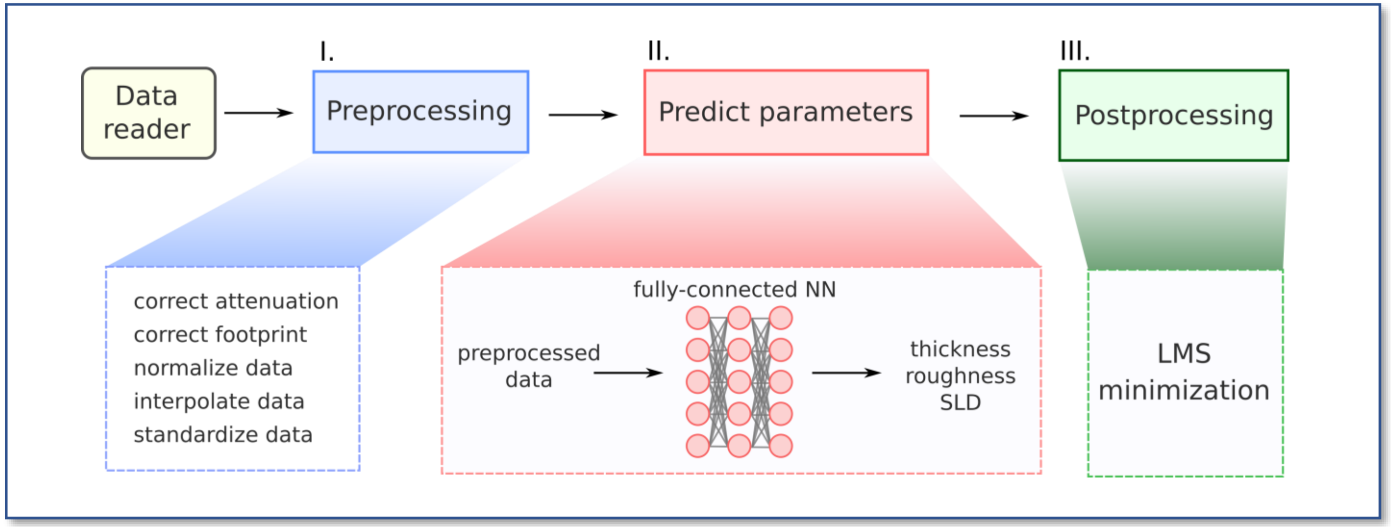
GitHub repo

Bundesministerium
für Bildung
und Forschung

Python package *mlreflect* was developed for a BMBF project

- Installed on Maxwell Custer at DESY (P08/PETRA III)
- Available on GitHub
- Installable via PyPI
- Online documentation available on Read the Docs
- Can be used with Jupyter notebooks as GUI

The *mlreflect* pipeline



Data reader → I. Preprocessing → II. Predict parameters → III. Postprocessing

correct attenuation
correct footprint
normalize data
interpolate data
standardize data

fully-connected NN

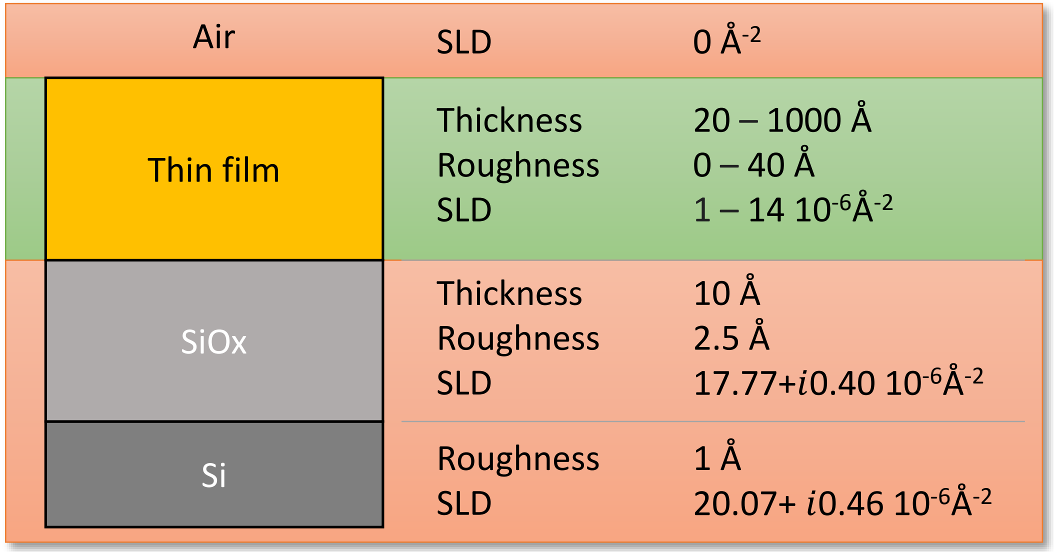
preprocessed data → → thickness roughness SLD

LMS minimization

Greco *et al. J. Appl. Crystallogr.*, 2022, **55**, 362-369

Example: organic thin film on Si/SiOx

Thin film model for training:



Layer	Parameter	Value
Air	SLD	0 Å^{-2}
Thin film	Thickness	20 – 1000 Å
	Roughness	0 – 40 Å
	SLD	1 – 14 10^{-6}Å^{-2}
SiOx	Thickness	10 Å
	Roughness	2.5 Å
	SLD	17.77+i0.40 10^{-6}Å^{-2}
Si	Roughness	1 Å
	SLD	20.07+ i0.46 10^{-6}Å^{-2}

🟩 open 🟥 fixed

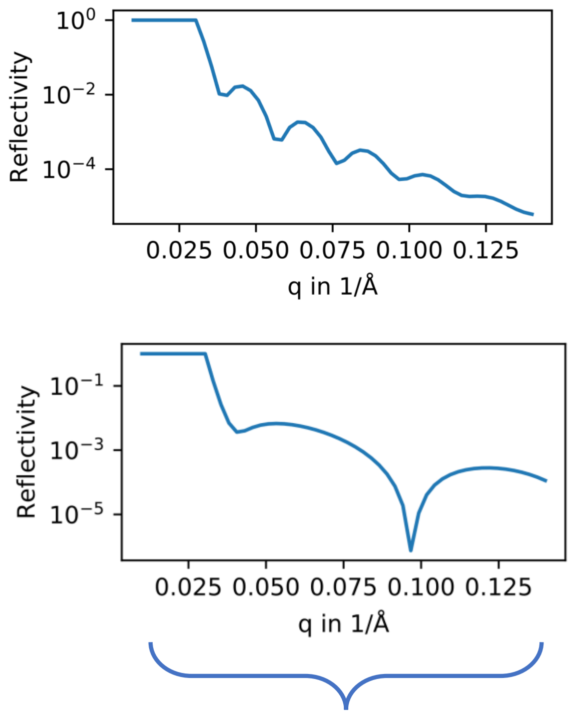
(trained model included in *mlreflect*)

Generate random parameter sets and simulate curves

→

+ uniform noise

Training set



q_z up to 0.15 Å^{-1}
109 points

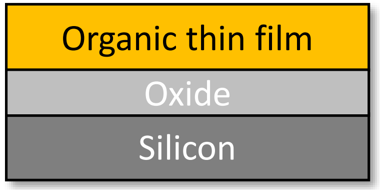
Greco *et al. J. Appl. Crystallogr.*, 2022, **55**, 362-369

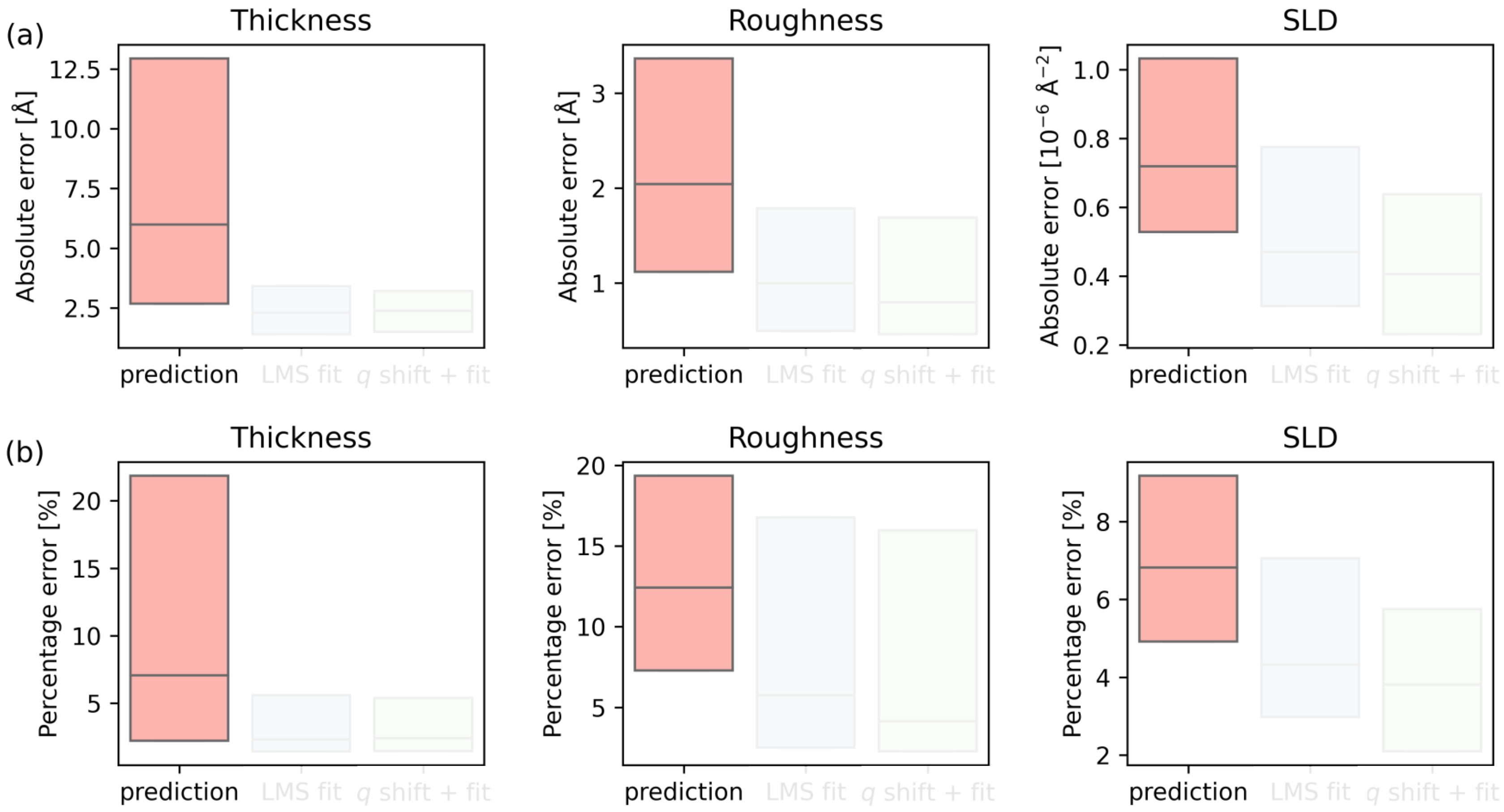
Prediction error distribution of *mlreflect*

Test neural network on a test dataset of **242 curves**

Dataset contains thin films of:

- Diindenoperylene
- Pentacene
- Perylene diimides
- Molecular mixtures

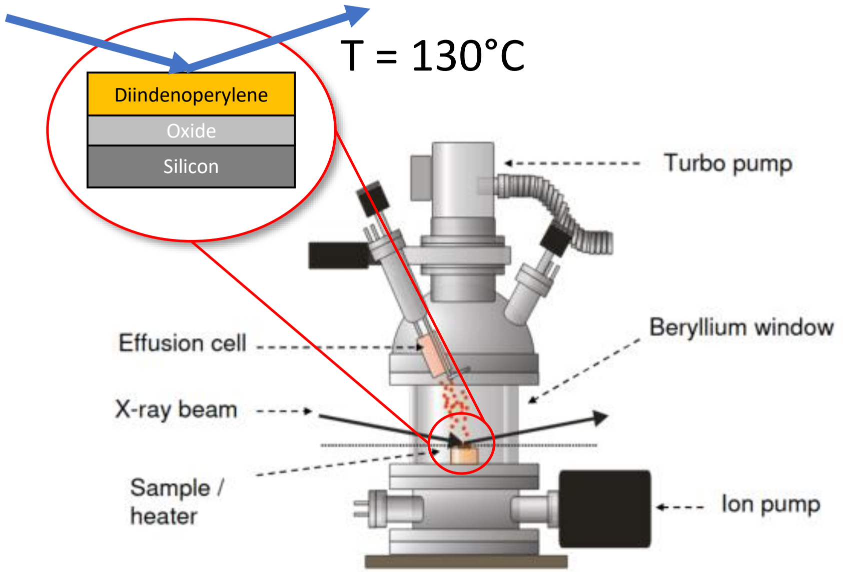




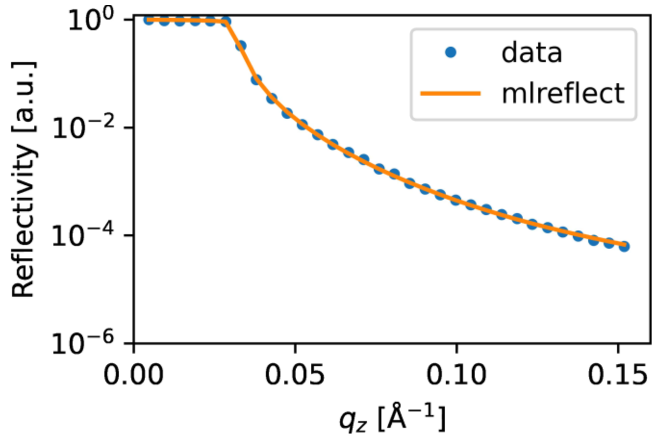
Greco *et al. J. Appl. Crystallogr.*, 2022, **55**, 362-369

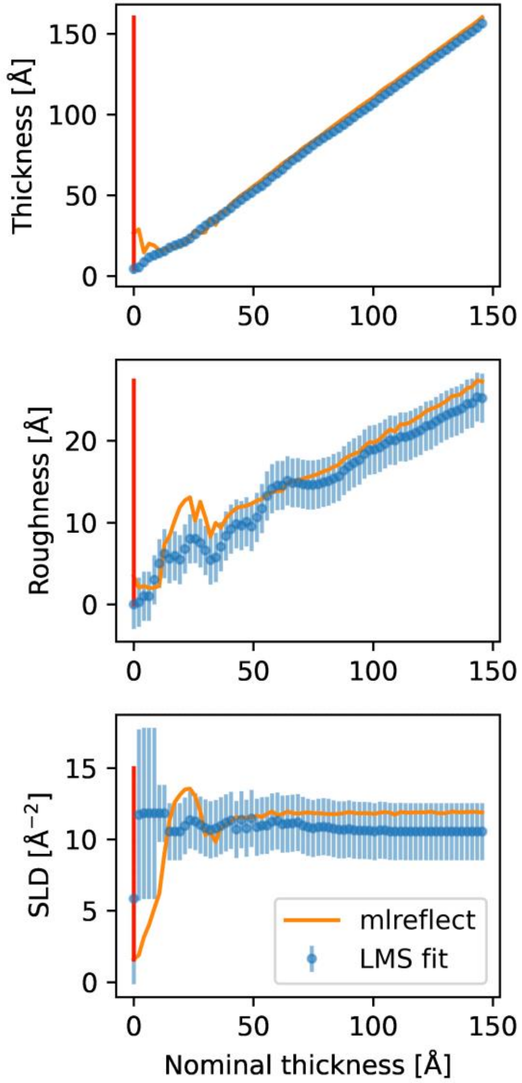
In situ applications of *mlreflect*

In situ XRR during film deposition



- Real-time parameter prediction is useful for in situ experiments
- After training, no human input is necessary
- Results are obtained within <1s per curve
- Ideal for monitoring and feedback loops





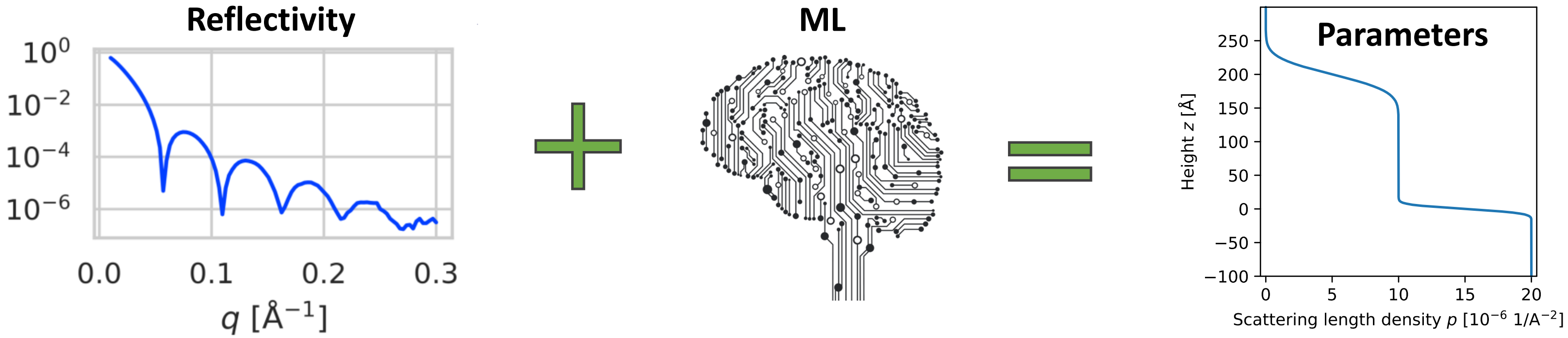
Hinderhofer *et al. Europhys. Lett.*, 2010, **91**, 56002
Kowarik *et al. Phys. Rev. Lett.*, 2006, **96**, 125504
Bommel *et al. Nat. Comm.*, 2014, **5**, 5388

Greco *et al. J. Appl. Crystallogr.*, 2022, **55**, 362-369

mlreflect: Summary / Successes

Features of *mlreflect*

- Once a neural network model is trained, predictions are obtained within <1ms
- Predictions can be refined via LMS fit
- Fast prediction time can be exploited for input resampling
- Final result is obtained within <1s/curve
- Everything is provided in a Python package



Bundesministerium
für Bildung
und Forschung

Greco *et al. J. Appl. Crystallogr.*, 2022, **55**, 362-369

mlreflect: Remaining Challenges

- Full analysis pipeline / *mlreflect* package
- Can be used for fitting in real time / for in-situ experiments
- Even if ML does not give final result, it provides starting parameters for faster conventional fit

- High dynamic range in XRR / NR is a challenge
- Adding FFT of reflectivity curve as input did not affect performance
- Correction of q_z scale important (even small misalignment of 10^{-3} degrees!)
- Pathological cases addressed (e.g., no edge for some NR curves)
- Co-refinement of larger data sets / XRR & NR / contrast variation NR yet to be addressed

- Closed-loop experiment-ML-experimental control-experiment (demonstrated; see Pithan et al.)
- More than 1 layer / more complex layered structures (Starostin)
- Handling phase problem / fitting ambiguity (Starostin)

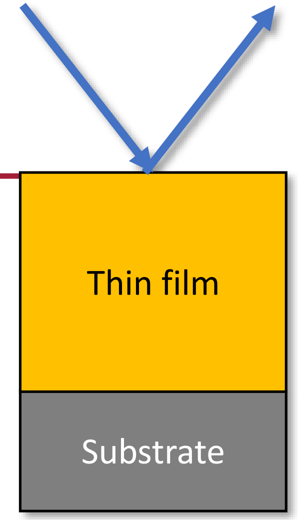
- XRR / NR data base for proper ML training; please do contribute! (All)
- Support efforts for a coherent data infrastructure; e.g., DAPHNE (All)

Bundesministerium
für Bildung
und Forschung

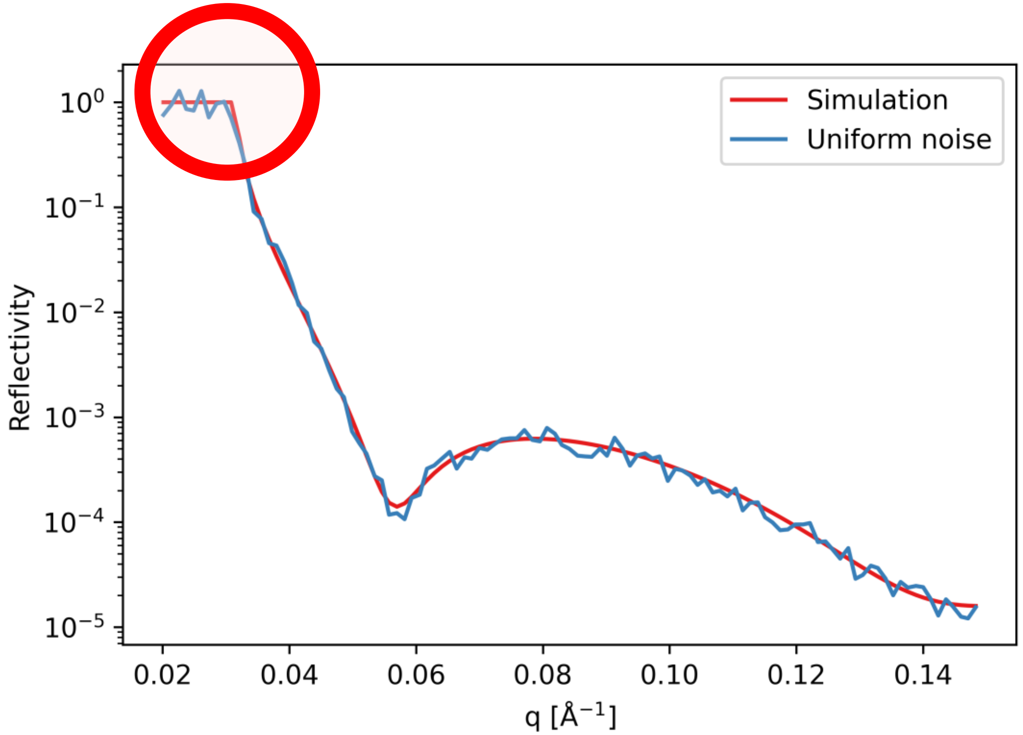
Greco *et al. J. Appl. Crystallogr.*, 2022, **55**, 362-369

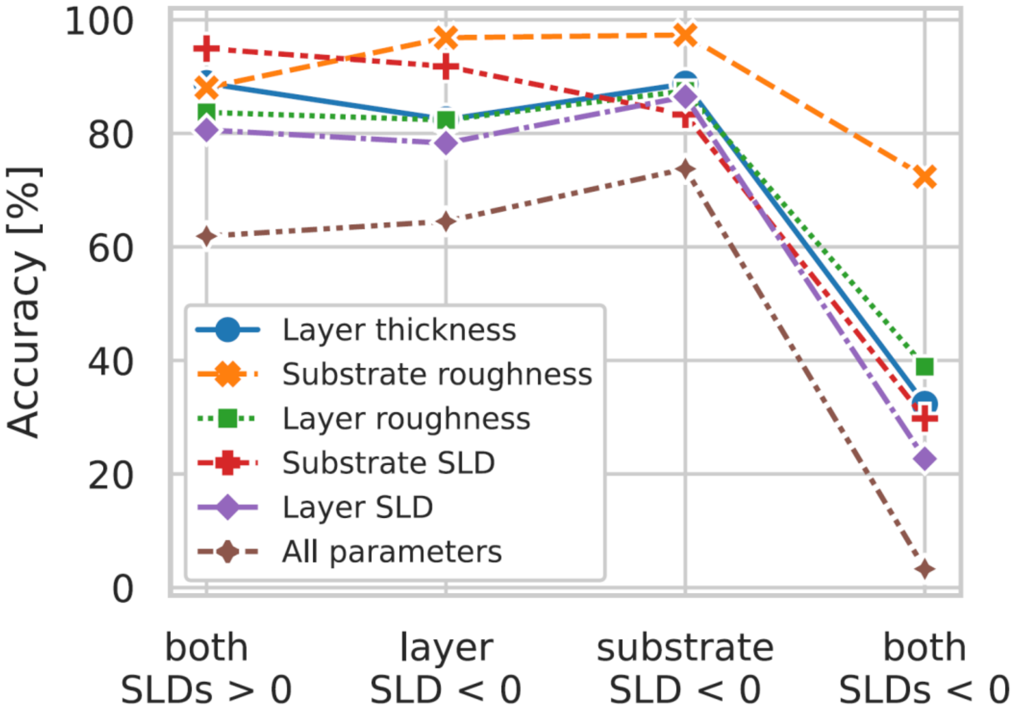
Challenges in XRR/NR

Challenge: Specifics of neutrons (NR)



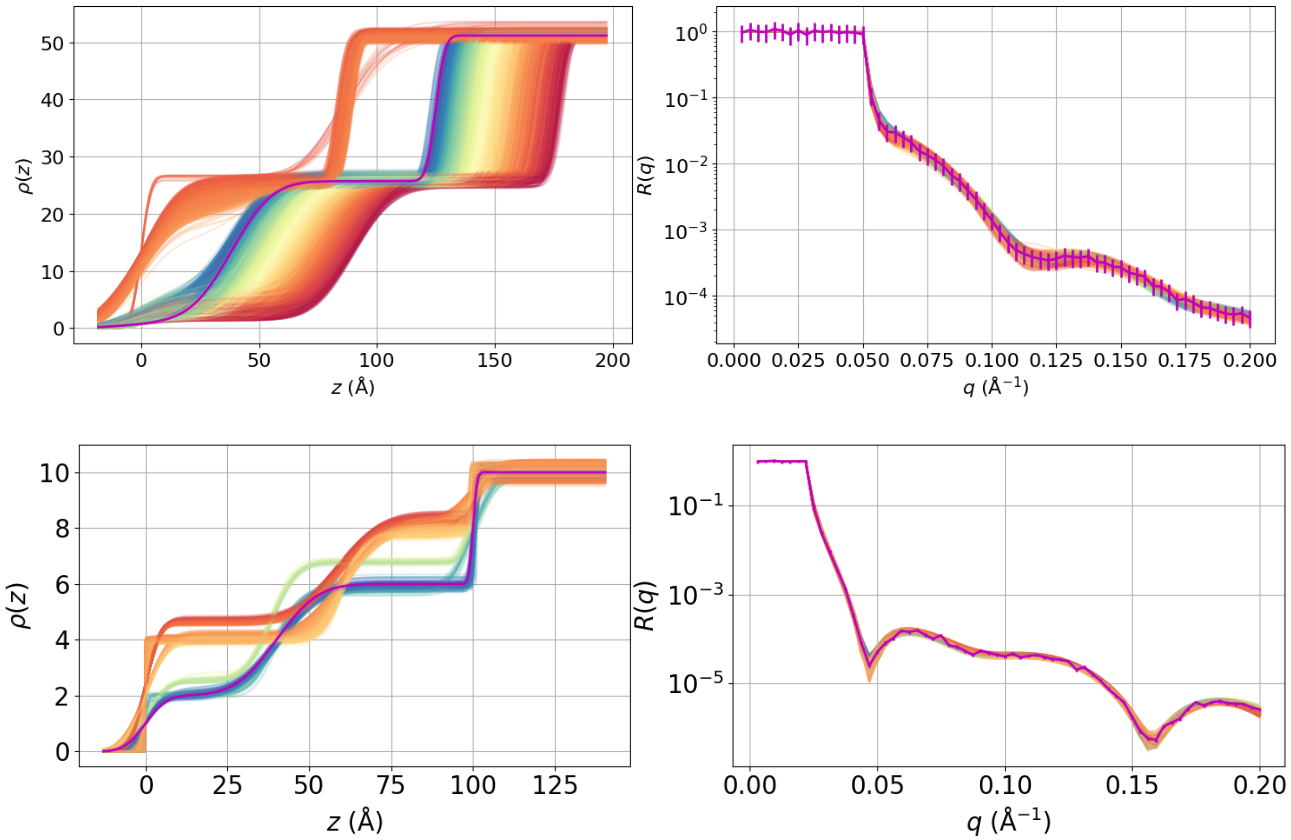
- Typically lower counts and statistics (but isotope variation and other opportunities)
- Cross section for some isotopes can lead to SLD < 0 for NR (no edge)
- Lack of edge complicates ML analysis (if training is with edge!)





Greco *et al.*, Mach. Learn.: Sci. Technol. 2 (2021) 045003

Challenge: Ambiguity & phase problem



Electron density profiles $\rho(z)$

Simulated reflectivity curves

Color is used to distinguish between different profiles & connect to the corresponding curves

Multiple solutions occur even in the simplest case of two-layer structures

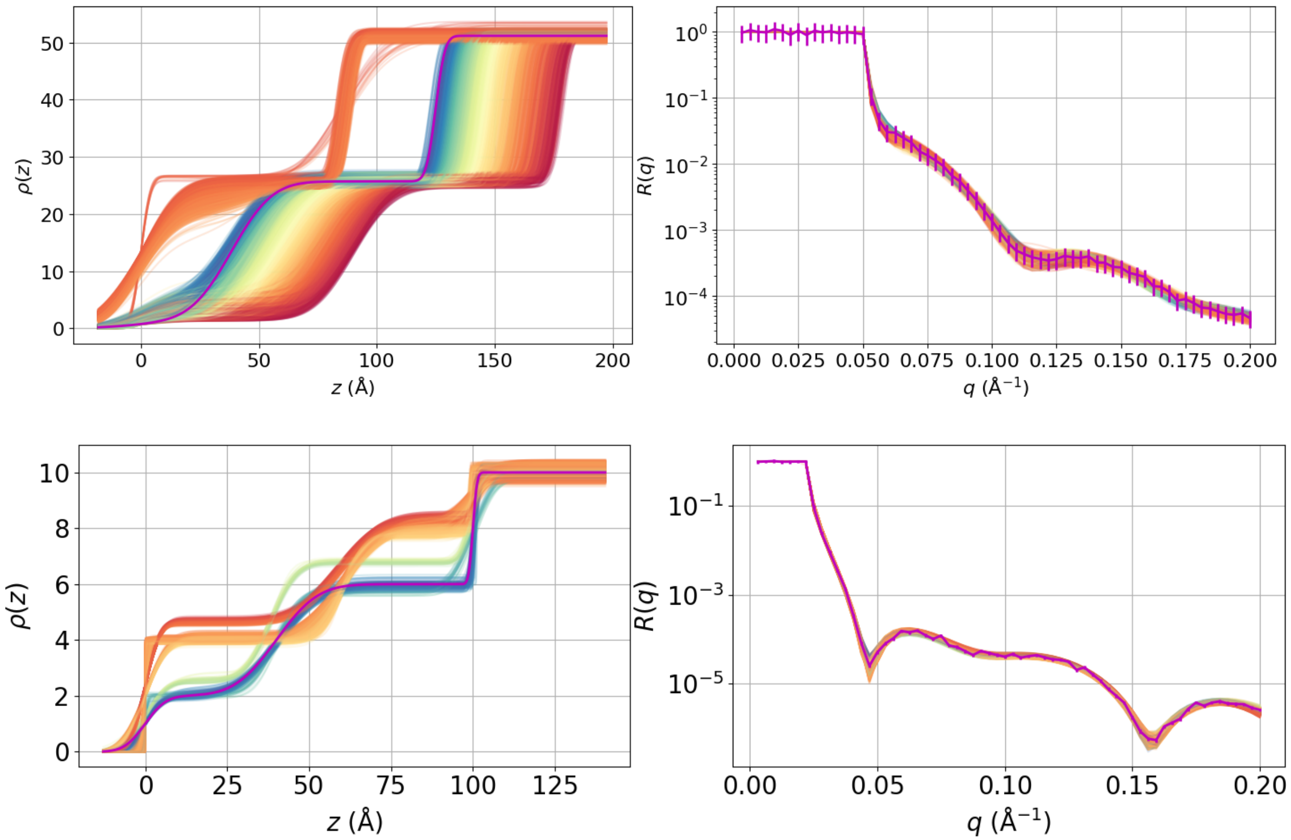
- Theoretical ambiguity (phase loss)
- Counting statistics
- Finite q range & q resolution
- Deviations from the box model
- Experimental artefacts
- …

More parameters → more possible solutions

Simple regression approach does not work in this case and requires modifications

Starostin *et al.*, in preparation

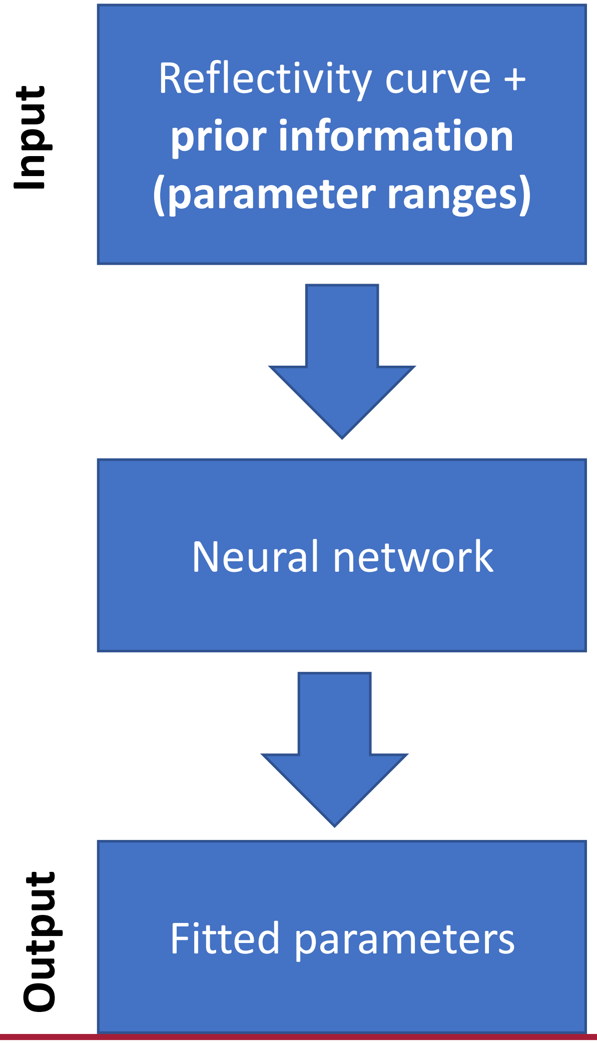
Challenge: Ambiguity & phase problem


Electron density profiles $\rho(z)$

Simulated reflectivity curves

Color is used to distinguish between different profiles & connect to the corresponding curves

The simplest solution:


Reflectivity curve + **prior information (parameter ranges)**

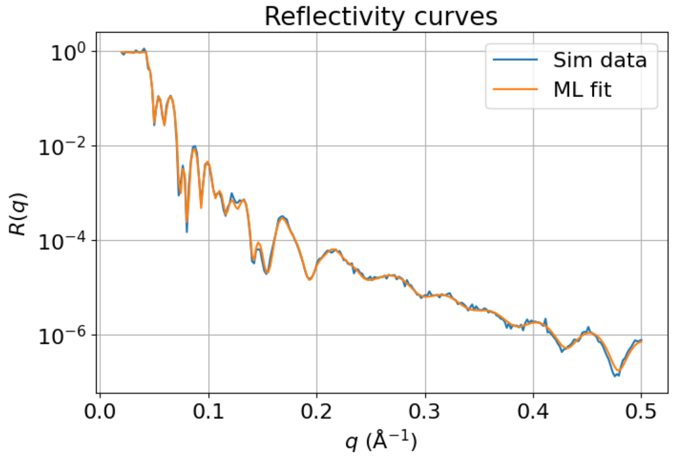
Neural network

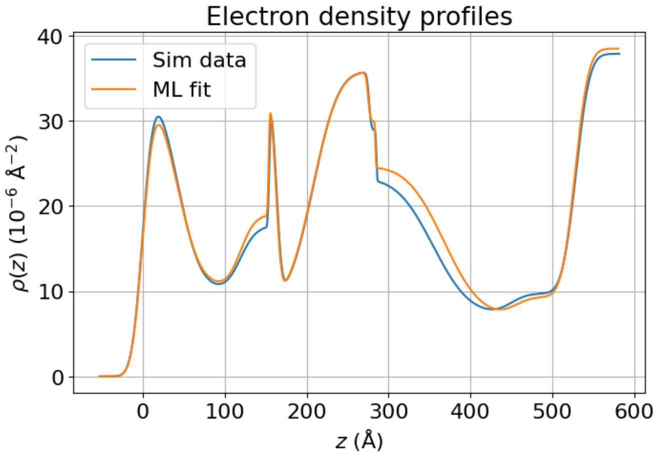
Fitted parameters

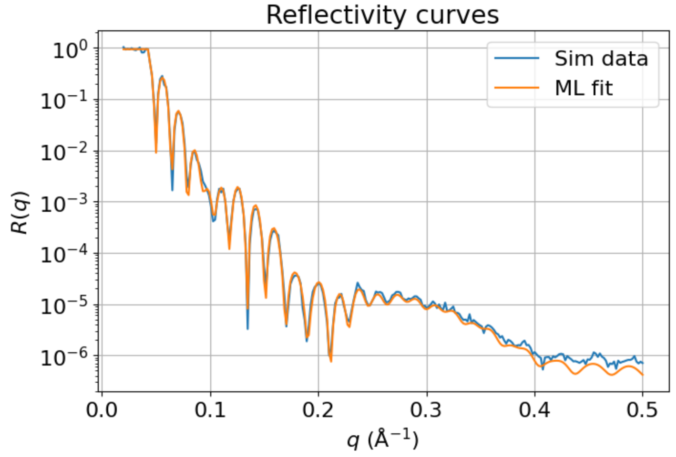
Starostin *et al.*, in preparation

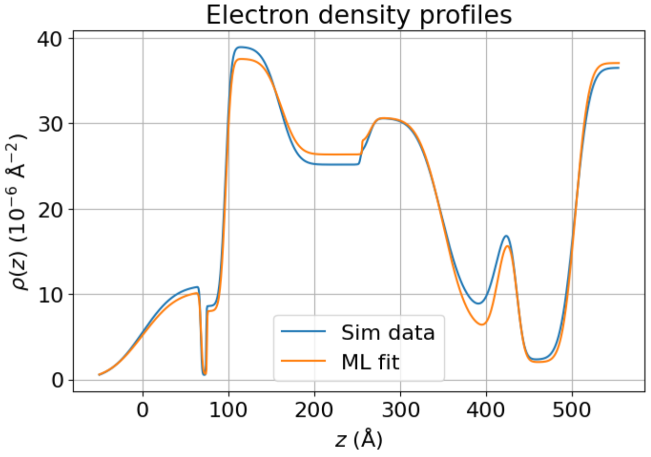
Challenge: Complex layers & prior information

- Model with up to 10 independent layers (34 parameters)

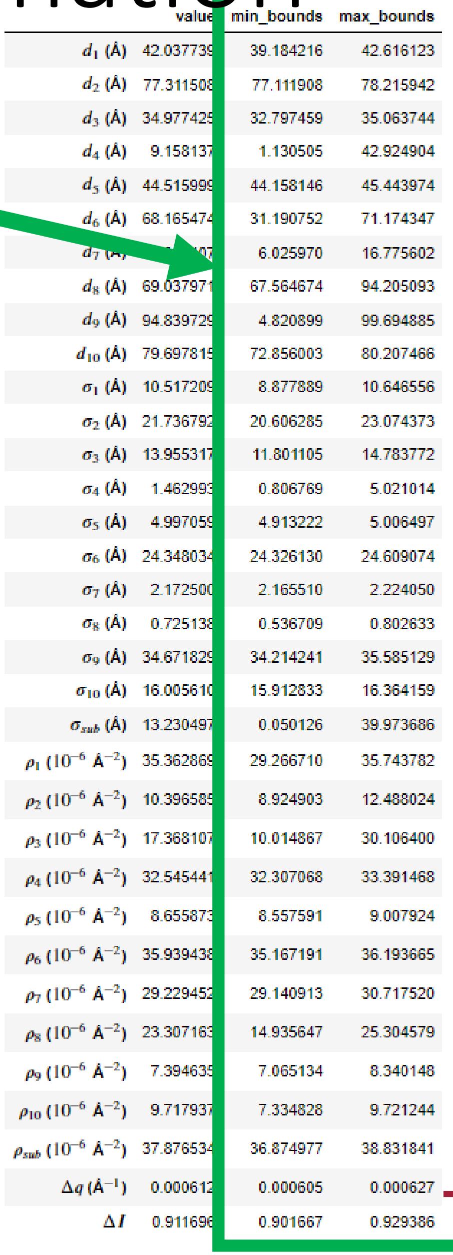

Reflectivity curves


Electron density profiles


Reflectivity curves


Electron density profiles

Additional input to the NN: prior information (parameter bounds)

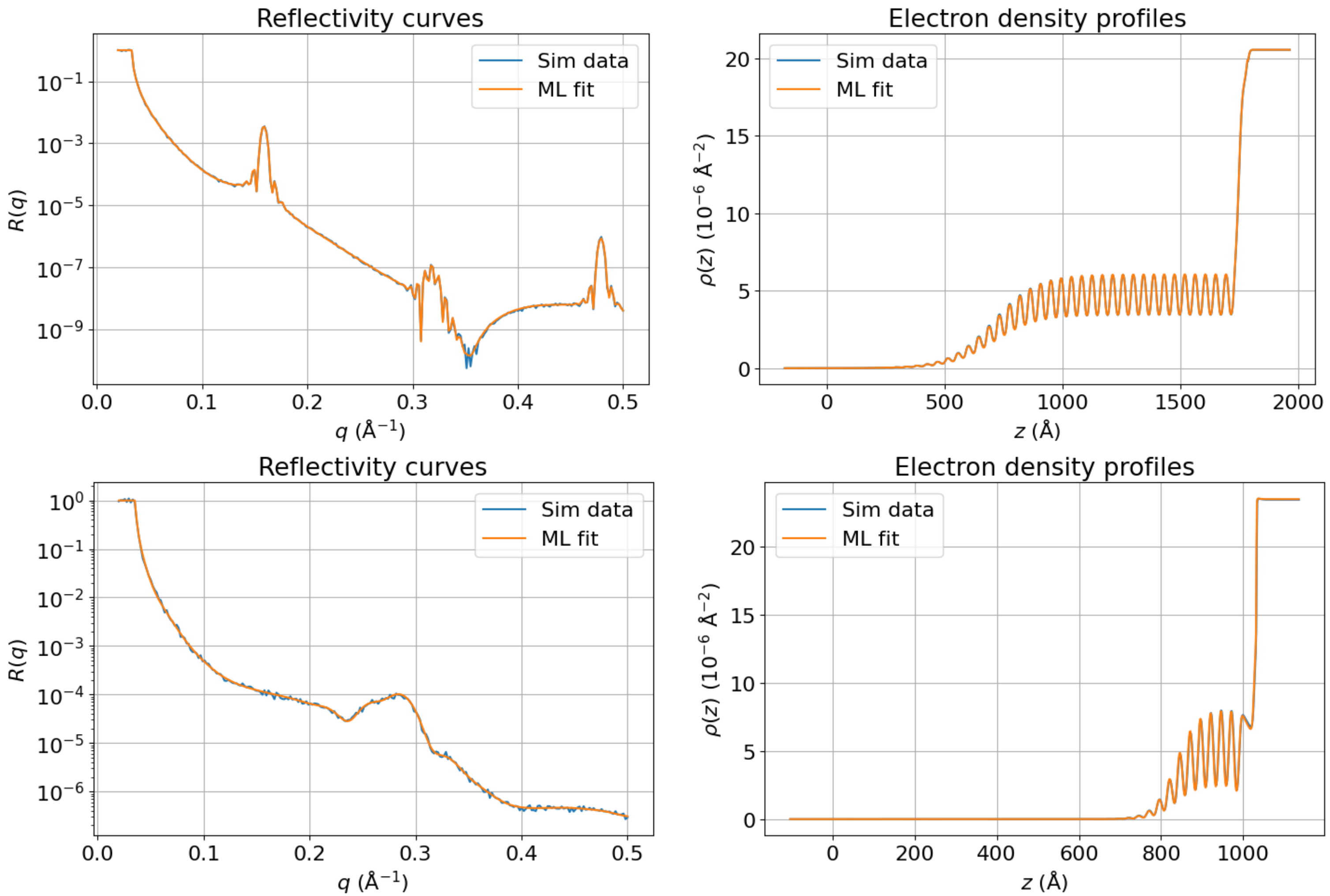


	value	min_bounds	max_bounds
d_1 (Å)	42.037739	39.184216	42.616123
d_2 (Å)	77.311508	77.111908	78.215942
d_3 (Å)	34.977425	32.797459	35.063744
d_4 (Å)	9.158137	1.130505	42.924904
d_5 (Å)	44.515999	44.158146	45.443974
d_6 (Å)	68.165474	31.190752	71.174347
d_7 (Å)	[illegible]	6.025970	16.775602
d_8 (Å)	69.037971	67.564674	94.205093
d_9 (Å)	94.839729	4.820899	99.694885
d_{10} (Å)	79.697815	72.856003	80.207466
σ_1 (Å)	10.517209	8.877889	10.646556
σ_2 (Å)	21.736792	20.606285	23.074373
σ_3 (Å)	13.955317	11.801105	14.783772
σ_4 (Å)	1.462993	0.806769	5.021014
σ_5 (Å)	4.997059	4.913222	5.006497
σ_6 (Å)	24.348034	24.326130	24.609074
σ_7 (Å)	2.172500	2.165510	2.224050
σ_8 (Å)	0.725138	0.536709	0.802633
σ_9 (Å)	34.671829	34.214241	35.585129
σ_{10} (Å)	16.005610	15.912833	16.364159
σ_{sub} (Å)	13.230497	0.050126	39.973686
ρ_1 (10^{-6} Å^{-2})	35.362869	29.266710	35.743782
ρ_2 (10^{-6} Å^{-2})	10.396585	8.924903	12.488024
ρ_3 (10^{-6} Å^{-2})	17.368107	10.014867	30.106400
ρ_4 (10^{-6} Å^{-2})	32.545441	32.307068	33.391468
ρ_5 (10^{-6} Å^{-2})	8.655873	8.557591	9.007924
ρ_6 (10^{-6} Å^{-2})	35.939438	35.167191	36.193665
ρ_7 (10^{-6} Å^{-2})	29.229452	29.140913	30.717520
ρ_8 (10^{-6} Å^{-2})	23.307163	14.935647	25.304579
ρ_9 (10^{-6} Å^{-2})	7.394635	7.065134	8.340148
ρ_{10} (10^{-6} Å^{-2})	9.717937	7.334828	9.721244
ρ_{sub} (10^{-6} Å^{-2})	37.876534	36.874977	38.831841
Δq (Å^{-1})	0.000612	0.000605	0.000627
ΔI	0.911696	0.901667	0.929386

Starostin *et al.*, in preparation

Challenge: Complex layers & prior information

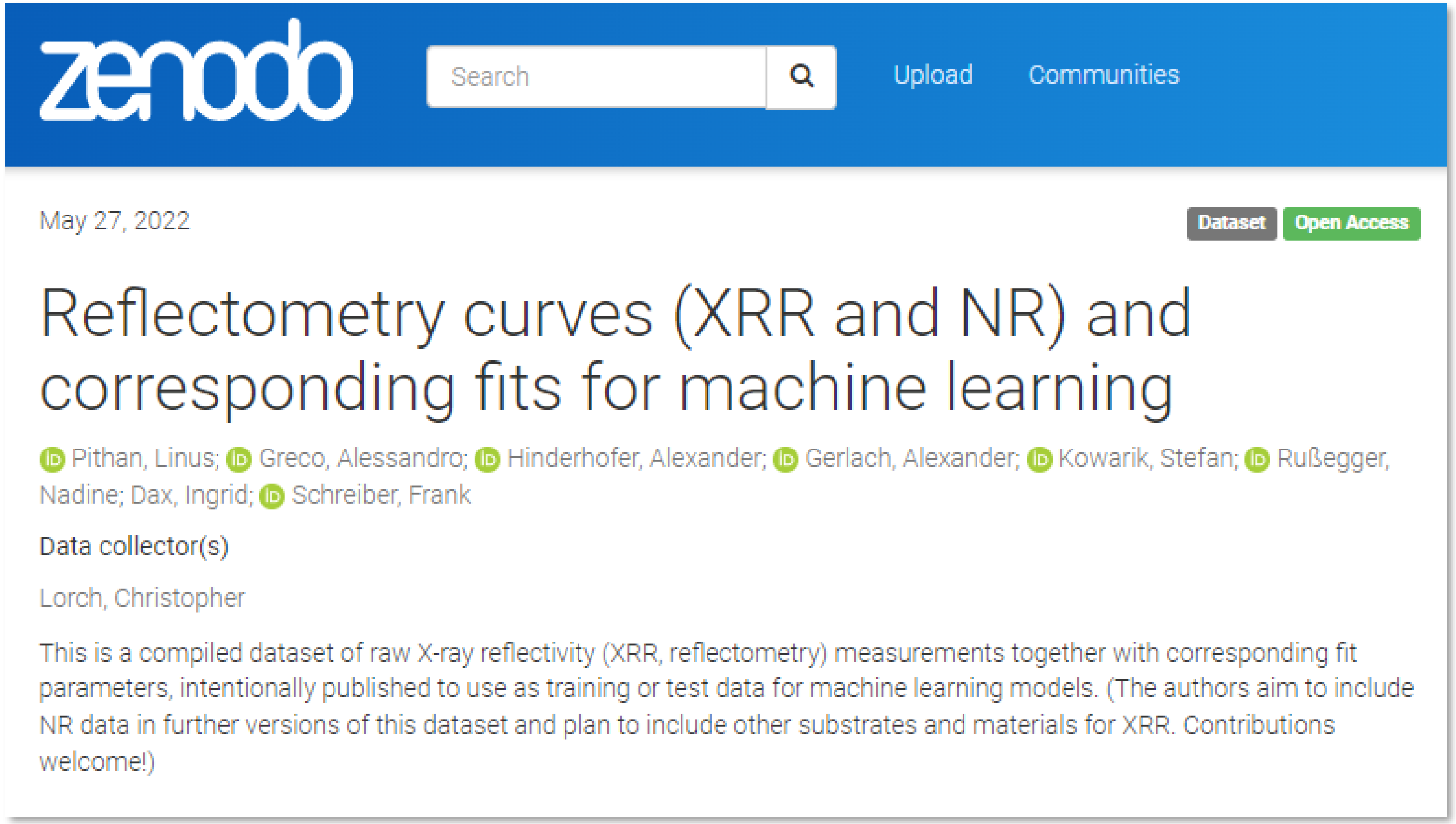
- Multilayer model with Bragg peaks (19 parameters)



- We can provide different parametrization to analyze various cases such as multilayer structure
- As before, parameter ranges are used as an additional input to the model

Starostin *et al.*, in preparation

Challenge: XRR/NR datasets for ML



May 27, 2022

Reflectometry curves (XRR and NR) and corresponding fits for machine learning

Pithan, Linus; Greco, Alessandro; Hinderhofer, Alexander; Gerlach, Alexander; Kowarik, Stefan; Rußegger, Nadine; Dax, Ingrid; Schreiber, Frank

Data collector(s)

Lorch, Christopher

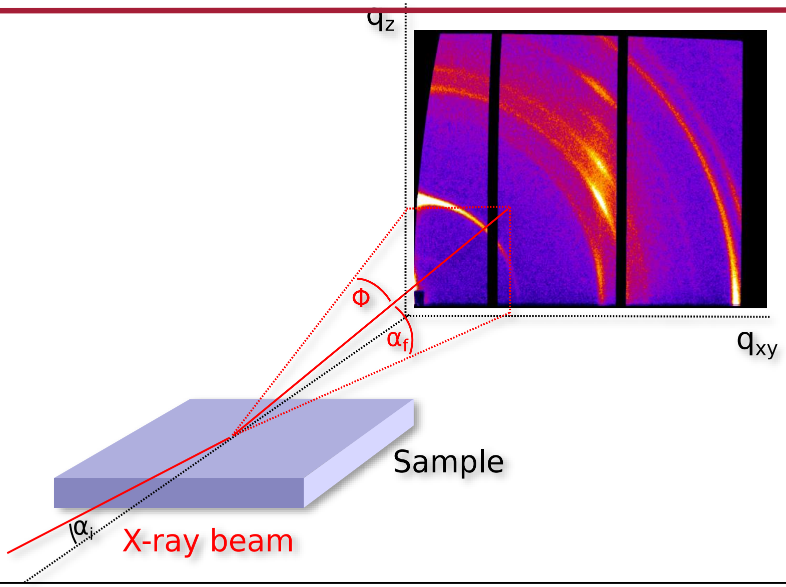
This is a compiled dataset of raw X-ray reflectivity (XRR, reflectometry) measurements together with corresponding fit parameters, intentionally published to use as training or test data for machine learning models. (The authors aim to include NR data in further versions of this dataset and plan to include other substrates and materials for XRR. Contributions welcome!)

https://doi.org/10.5281/zenodo.6497437

- Repository for NR/XRR data for machine learning
- Currently contains 242 XRR curves from Schreiber group
- Contains raw data + layer parameters
- Plan to convert to ORSO formats
  - HDF5
  - Nexus
  - ORSO model language
- Currently maintained by Linus Pithan (linus.pithan@uni-tuebingen.de)

Please contribute if possible (especially NR data)!

Machine Learning for Surface Scattering



- ML for scattering data works, is fast, and is needed
- Further need comes from ever-improving sources
- Surface scattering geometry has specific challenges
- Two working packages established: mlreflect and gixi
- XRR/NR (“1D”) … ML works, but trickier than thought
- GIXD (“2D”) … feature recognition etc works with ML
- Full 2D-structure determination yet to be addressed

A. Hinderhofer et al., Machine learning for scattering data: Strategies, perspectives, and applications to surface scattering
J. Appl. Cryst. 56 (2023) 3
V. Starostin et al., End-to-end deep learning pipeline for real-time processing of surface scattering data at synchrotron facilities
Synchrotron Radiation News 35 (2022) 21
V. Starostin et al., Tracking perovskite crystallization via deep learning-based feature detection on 2D X-ray scattering data
npj Comput Mater – Nature 8 (2022) 101
S. Timmermann et al., Automated matching of two-time X-ray photon correlation maps from protein dynamics … using autoencoder networks
J. Appl. Cryst. 55 (2022) 751
A. Greco et al., Neural network analysis of neutron and X-ray reflectivity data: automated analysis using mlreflect, experimental errors and feature engineering
J. Appl. Cryst. 55 (2022) 362
A. Greco et al., Neural network analysis of neutron and X-ray reflectivity data: Pathological cases, performance and perspectives
Mach. Learn.: Sci. Technol. 2 (2021) 045003
A. Greco et al., Fast fitting of reflectivity data of growing thin films using neural networks
J. Appl. Cryst. 52 (2019) 1342

ML

DFG
Deutsche Forschungsgemeinschaft

Bundesministerium für Bildung und Forschung

DAPHNE 4NFDI

SciCat