McStas meets Machine Learning

Petroula Karacosta

Master's student, Computational Physics

UNIVERSITY OF COPENHAGEN

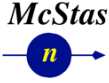
Master's Thesis:

Using McStas Union components to simulate a magnet sample environment

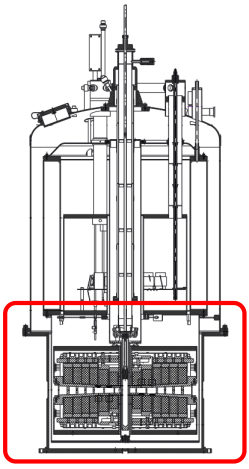
Supervisor: Kim Lefmann, KU
Co-supervisor: Mads Bertelsen, ESS DMSC

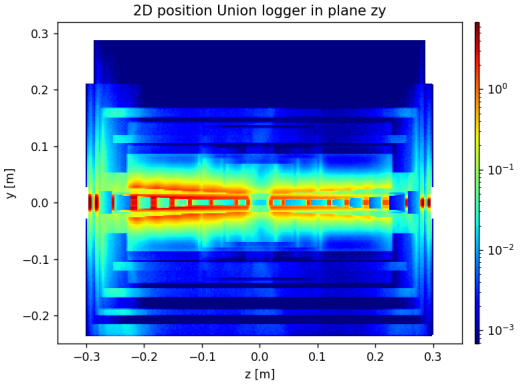
Two main focal points:

- Simulation of a 15 T Magnet: BIFROST Sample Environment
- Prediction of background signal with Machine Learning

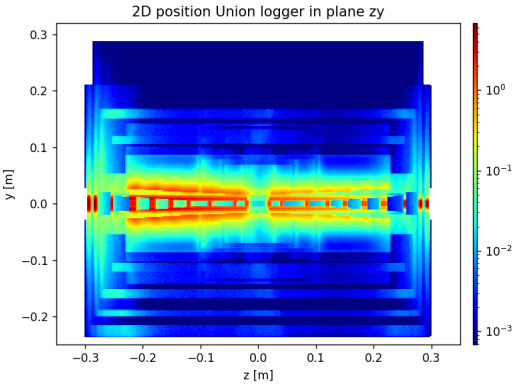


McStas Simulation: BIFROST Sample Environment, 15 T Magnet





Building the magnet with Union components



Union Components

72 cylinders: 48 solid & 24 vacuum

2 cones

1 box (beam path)

Union Materials

Most structures: Al & Vacuum

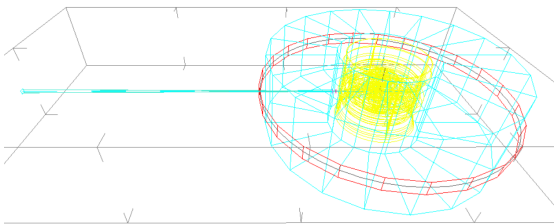
Magnet coils

Outer coils: NbTi

Inner coils: Nb3Sn

Simulation Layout

Powder diffraction using a monochromatic beam



Instrument components

- Simple monochromatic source
- Slit
- Simulated Sample Environment and Magnet
- Radial Collimator
- "Banana" Monitors

Where McStas meets Machine Learning

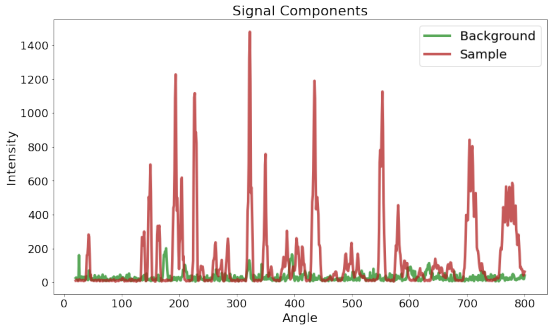
Simulations → Machine Learning model → Background Prediction

Structure of a Machine Learning project

- Framing the problem
- Data Collection and Cleaning
- Exploratory Data Analysis
- Pre-processing
- Model Selection
- Training and Evaluation

Classification or Regression?

Predict **intensity** for each angle increment



Classification
Signal or background?
Little information - Qualitative result

Regression
Prediction of background
as a continuous value
More information - Higher accuracy

Regression Models to try

Random Forest

Parallelised ensemble of weak learners
Bootstrap Sampling: Sampling part of the training data
Aggregation: Result decided by "majority vote", or the mean of all decision trees output

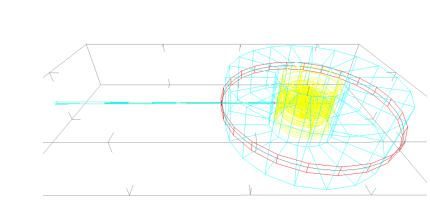
Gradient Boosting Algorithm

Sequential ensemble of weak learners
Combined weak learners fitted iteratively
Sequential training: "Bad" performance of a learner is given more importance by next learner

Neural Networks

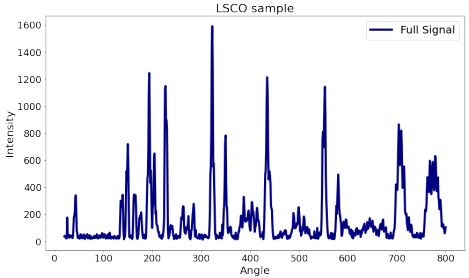
Shallow or deep networks. More versatile, customisable, higher complexity

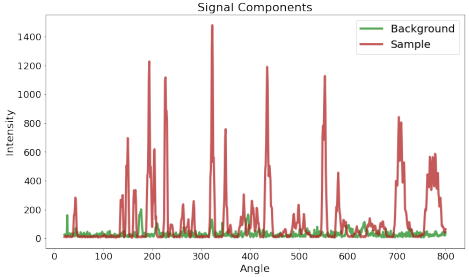
Making a Database of Simulations: Adjusting the instrument file



Three monitors used to produce
Features and **Target Values**

1. All scattering events
   Input Values: Total signal (intensity per angle)
2. Neutrons scattering only once and through the sample
3. All remaining scattering events
   Target Values: Background



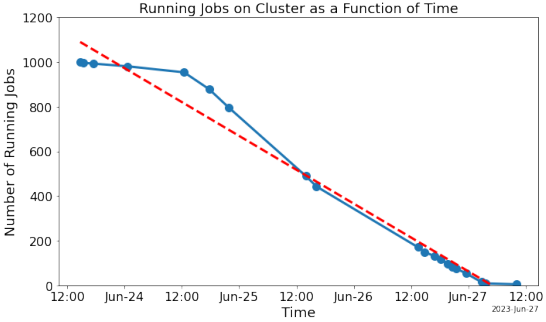


Data Collection: Making a Database of over 25000 Simulations

Parameter values were selected from random uniform distributions within ranges:

- λ: 1-9 Å
- dλ: 0.5-2.5 % of λ
- Beam divergence: 20-40 arcmin
- Detector radius: 0.897 - 2 m
- Cylindrical sample size:
  diameter: 2-10mm, height: 5-30mm
- Sample material: 11 choices

  $Na_2Ca_3Al_2F_{14}$, I_2, Al_2O_3, H_2O (ice),
  Y_2O_3, $Y_3Fe_5O_{12}$ (YIG), UO_2, Sn, B_4C,
  Isco 64, V, Vacuum



Cleaning and processing data

Collecting raw data into structured format
Encoding categorical values: Assign a value to each material choice

	wavelength	d_wavelength	beam_div	sample_radius	sample_height	detector_rad	material	signal_1	signal_2	signal_3	...	sample_795	sample_796	sample_797	sample_798	sample_799	sample_800
0	3.45088	0.045748	31.0377	0.004838	0.028443	1.17775	1	2093.574301	2974.172070	1874.180052	...	8177.138115	7330.162894	7383.390836	6330.836814	4713.540587	5039.678809
1	8.96147	0.156229	30.5311	0.001658	0.016362	1.08426	10	15.358362	10.168159	11.581088	...	0.658292	1.495205	0.746088	1.615491	1.169181	1.192733
2	7.39940	0.043200	37.9386	0.003891	0.017288	1.34551	5	70.475290	102.198785	114.620864	...	102.432599	105.990537	92.729147	79.835609	74.717216	83.668936
3	7.97298	0.126721	38.7282	0.003378	0.013200	1.11053	4	6.939370	8.437269	9.217048	...	26.710119	24.194719	15.866420	18.989273	24.970643	26.257607
4	3.86942	0.027450	24.0516	0.001217	0.010006	1.22373	0	169.282559	128.612371	85.224789	...	99.441613	70.400378	54.195730	82.083699	109.508592	94.387039
...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...	...
24014	6.11130	0.023404	20.2375	0.003351	0.009776	1.80551	5	10.582897	14.637855	17.614988	...	11.939805	16.118960	16.420973	18.307155	14.021256	11.434716
24015	8.14771	0.022988	34.6325	0.001056	0.014663	1.93028	3	1.175203	1.100478	1.087013	...	0.000000	0.000000	0.000000	0.000000	0.000000	0.000000
24016	7.95184	0.033603	37.5247	0.004623	0.006092	1.02142	10	3.290148	2.561215	2.895277	...	1.686605	2.455162	1.850315	1.935624	2.389665	1.745362
24017	4.65534	0.023013	33.3229	0.004397	0.007370	1.96536	4	52.520738	44.948128	56.360807	...	0.456724	0.414174	0.523308	0.365179	0.351744	0.382613
24018	6.66295	0.031765	22.4598	0.004973	0.007580	1.13144	2	15.253862	14.560504	12.649853	...	15.046782	10.913429	14.515927	17.319521	16.451860	13.440308

24019 rows × 2407 columns

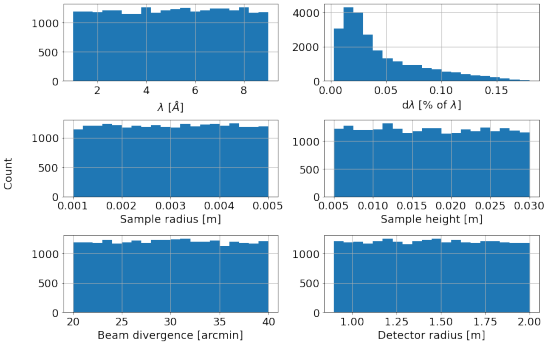
Features: Input to ML model
7 instrument parameters
800 intensity/angle values (total signal)

Target values: Output of ML model
800 intensity/angle values
(background signal)

Exploratory Data Analysis

No unexpected distributions, as we "engineered" the features

	wavelength	d_wavelength	beam_div	sample_radius	sample_height	detector_rad	material
count	24019.000000	24019.000000	24019.000000	24019.000000	24019.000000	24019.000000	24019.000000
mean	5.020123	0.044507	29.993715	0.003004	0.017404	1.447813	5.453474
std	2.303003	0.036424	5.746773	0.001152	0.007235	0.317502	3.449786
min	1.000110	0.002617	20.000300	0.001000	0.005000	0.897006	0.000000
25%	3.016495	0.017667	25.049550	0.002001	0.011161	1.173830	2.000000
50%	5.047200	0.031102	30.000300	0.003008	0.017367	1.448200	5.000000
75%	7.011210	0.062886	34.910250	0.003999	0.023723	1.721835	8.000000
max	8.999950	0.179152	39.999900	0.005000	0.029999	1.999980	11.000000



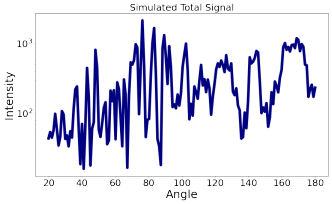
Training Using labeled data, allow the model to "learn" the algorithm that predicts the target value

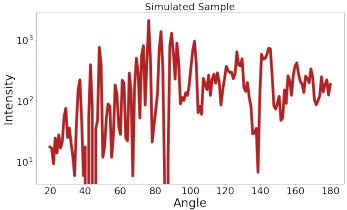
Tuning Adjusting the model's parameters to obtain optimal results

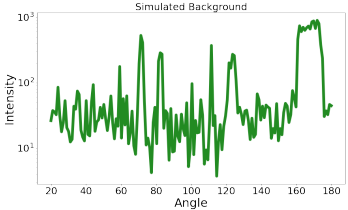
Number of estimators	Decision trees in the model	RF, GB
Maximum depth	Longest path from decision tree's root node to leaf node	RF, GB
Maximum features	Max number of features provided to each tree	RF, GB
Learning rate	Contribution of each tree to final prediction	GB

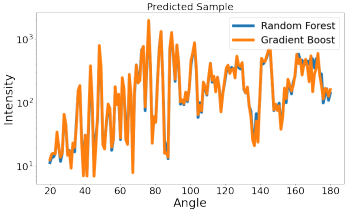
Evaluation Using metrics to assess model's performance:
Mean Absolute Error, Mean Squared Error, Root Mean Squared Error, R^2, etc.

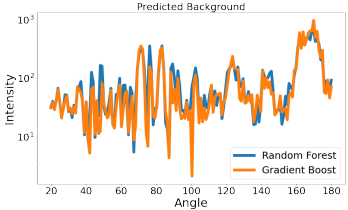
Some results: $Na_2Ca_3Al_2F_{14}$, λ=2.386 Å



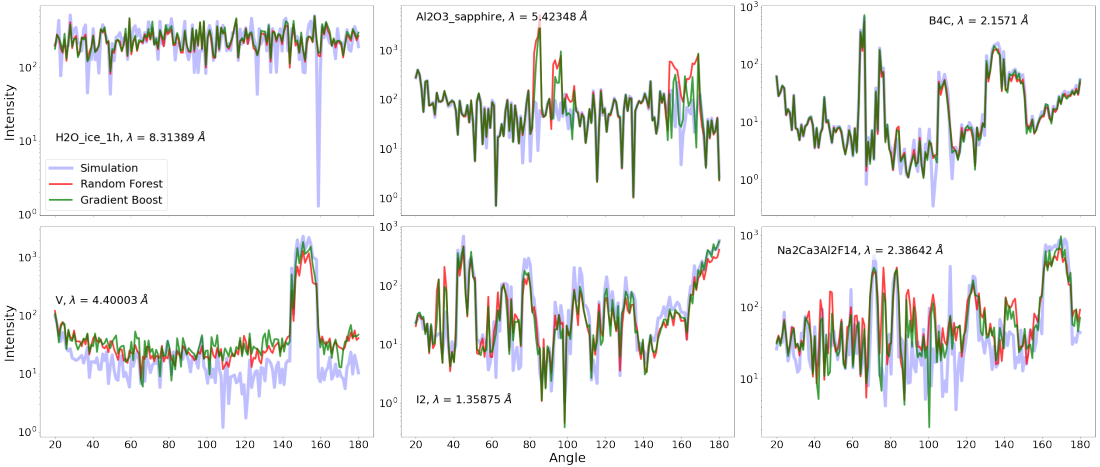








More results: Predicted Background



	Random Forest	Gradient Boost	Predicted values
Mean Squared Error	0.0017	0.0016	(0,1)

Thank you!

Using McStas Union components to simulate a magnet sample environment

Master's Thesis Defence: October 30th, Auditorium 7, HCØ, KU