Selected topics from physics aware ML

Mateusz Ploskon, LBL

A Landscape of AI in Science

Data Analytics

- Classification
- Regression
- Clustering
- Dimensionality Reduction

Inverse problems

- Model reconstruction
- Parameter estimation
- Denoising

Surrogate models

- Approximate expensive simulations
- Approximate experiments
- Fill in missing models in simulations

Design and control

- Optimize design of experiments
- Control instruments
- Navigate state spaces
- Learn from sparse rewards

From AI4Sci / DOE

Physics-awareness – what is it?

- Think: **ML obeying conservation laws** (symmetries, invariances) – **incorporate physics into ML setup objective (e.g. loss function)**
- Benefits:
  - Trainable with simulated data;
  - Improving understanding of uncertainties and/or short comings of the modelling;
  - Expose non-physical components / noise
  - Build tools to improve simulators by working with *real* data (unsupervised learning) => if done in a well controlled way allows to gain knowledge => impact on analytic solutions / theory / modelling;
  - Towards explainable ML: eventually learn physics from the data alone
  - …
  - *note: in general, this is more than just apply "standard" ML techniques*

(selected) ML areas in physical sciences

HEP BIASED

Some challenges (and thus opportunities!) – national lab driven:

1. The biggest scientific datasets with complex, high-dimensional phase space
    - Challenging pattern recognition; benefits from **physics-aware learning**
    - Strict requirements on **uncertainty quantification / interpretability**
2. High-fidelity, first-principles simulations / theory
    - Can scaffold / exploit simulations with NNs for precise **likelihood-free inference**
    - Often too slow and need to be accelerated with **generative models**
3. Simulation-based inference is complemented by **data-driven learning**
    - **Anomaly detection** is becoming a key across the area
    - Often require fast inference (trigger), feedback (accelerator control), and/or environment awareness (hazards/safety)

Compilation Ben Nachman, MP

(selected) ML areas in physical sciences

HEP BIASED

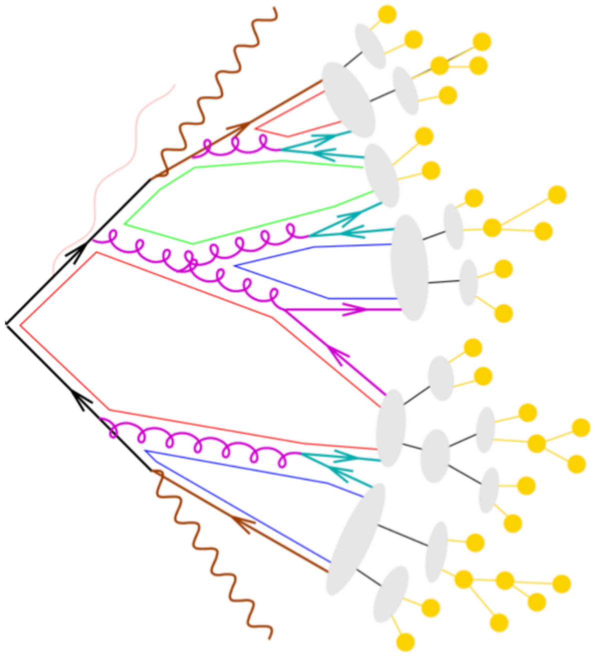
1. The biggest scientific datasets with complex, high-dimensional phase space
    - Particle tracking, noise mitigation, calibration, particle/jet/event classification, ...
    - Uncertainty quantification, interpretable observables/learning, robust learning...
2. High-fidelity, first-principles simulations / theory
    - Unfolding (deconvolution), parameter estimation, strong force dynamics, ...
    - Generative models for calorimeter emulation and cosmology; accelerator simulation
3. Simulation-based inference is complemented by data-driven learning
    - Searching for new particles and forces, mixed data/simulation labels, ...
    - Accelerator stability / control, learning in the presence of radioactive hazards, ...

arXiv: 2003.11603,2007.00149,1910.06286, trackML Kaggle, 1707.08600, ATL-PHYS-PUB-2018-013, 1910.03773, ATL-PHYS-PUB-2020-001, 1511.05190, ATL-PHYS-PUB-2017-017, 1807.10768, 2009.05930, ATLAS b/c-jet tagging (@Cori), top quarks, 1806.05667, 1909.03081, 1910.08606, 1910.10046, 1810.00835, 1902.07180, 1906.06429, 2010.02926, 2010.09745, 2007.14400, 1911.09107, 2010.03569, 1907.08209, 1612.04262, 1910.11530, 1906.06562, 2012.06582 2009.03796, 1712.10321, 1705.02355, 1701.05927, 1706.02390, 1711.08813, Snowmass LOI, 2005.02983, 2009.02205, 2001.05001, 2001.04990, 1902.02634, 1805.02664, LHC Olympics, slides, 1708.02949, 1702.00414, 1801.10158, paper, ...

Compilation Ben Nachman, MP

Recent example of **explainable**, physics aware ML

Explainable machine learning of the underlying physics of high-energy particle collisions
https://arxiv.org/abs/2012.06582
… proof-of-concept of our White Box AI approach using a Generative Adversarial Network (GAN) which learns from a DGLAP-based parton shower Monte Carlo event generator.
From "final state" particles learn internal workings of QCD – e.g. Altarelli-Parisi splitting function



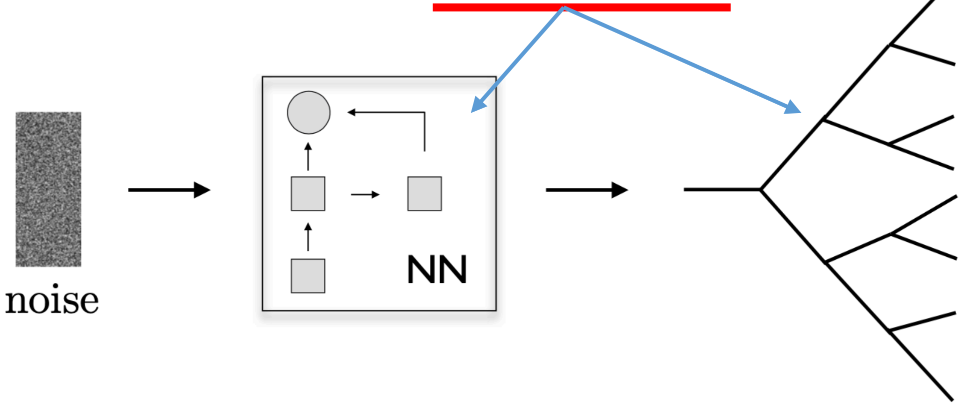
Learning from final state distribution the internal splitting probability density **f**

Methodology extendable to other areas –> e.g. modelling of nuclear reactions

Extract physics from the inner elements of the network – the "white box" to look at

Generic GAN $\mathbf{F}(\mathbf{p}^{\mathbf{in}}) \rightarrow \{\mathbf{p_i}\}$

White Box $\mathbf{F}(\mathbf{p}^{\mathbf{in}} \otimes \mathbf{f}^{\mathbf{n}}(\mathbf{z}^{\mathbf{k}}, \theta^{\mathbf{k}}, \phi^{\mathbf{k}})) \rightarrow \{\mathbf{p_i}\}$



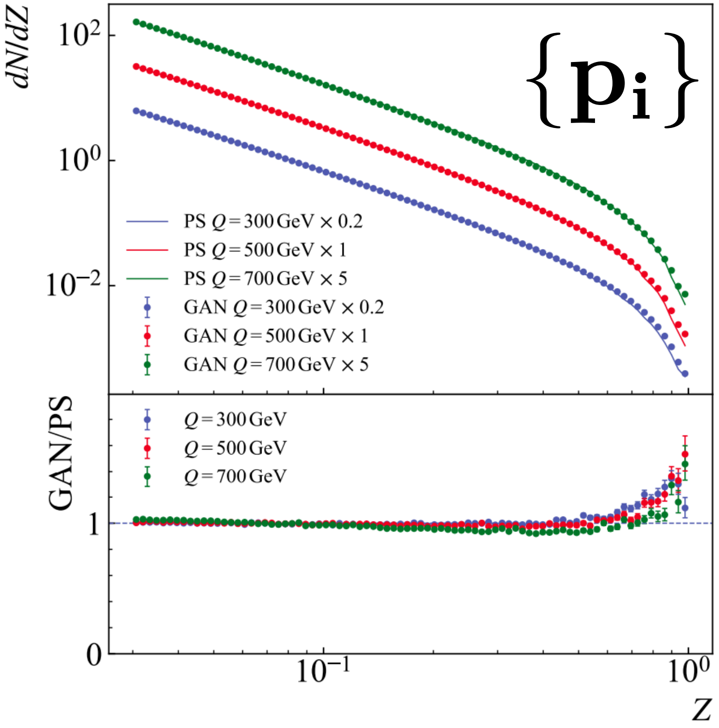
z = fraction of momentum carried by a particle in a 1->2 split

Recent example of **explainable**, physics aware ML

Explainable machine learning of the underlying physics of high-energy particle collisions
https://arxiv.org/abs/2012.06582

… proof-of-concept of our White Box AI approach using a Generative Adversarial Network (GAN) which learns from a DGLAP-based parton shower Monte Carlo event generator.
From “final state” particles learn internal workings of QCD – e.g. Altarelli-Parisi splitting function

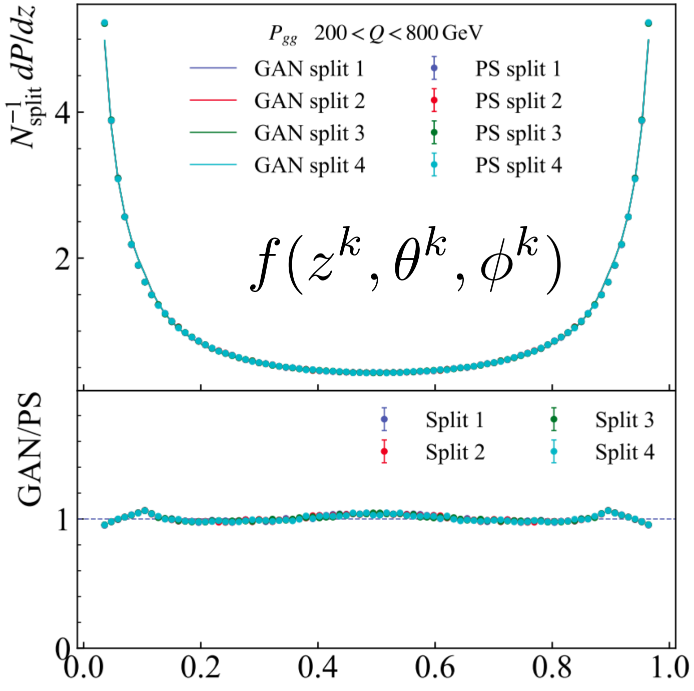
Z = fraction of momentum of the original parton carried by a final state particle



Learning from final state distribution the internal splitting probability density

Methodology extendable to other areas –> e. g. modelling of nuclear reactions

Extracted from inner elements of the network – the ”white box” to look at

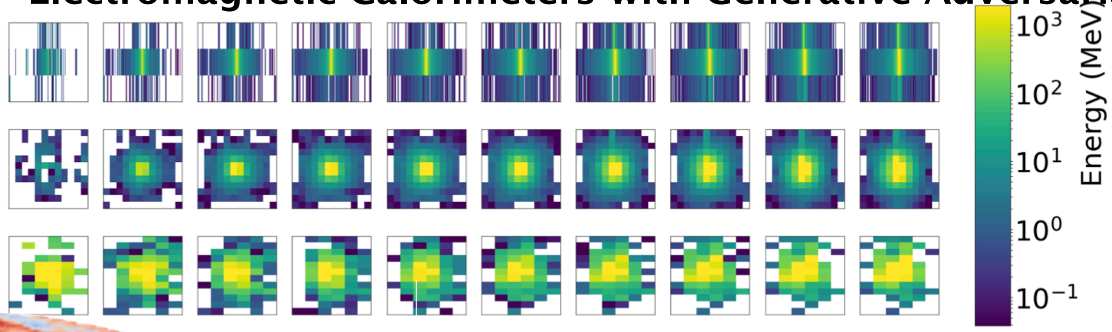


z = fraction of momentum carried by a particle in a 1->2 split

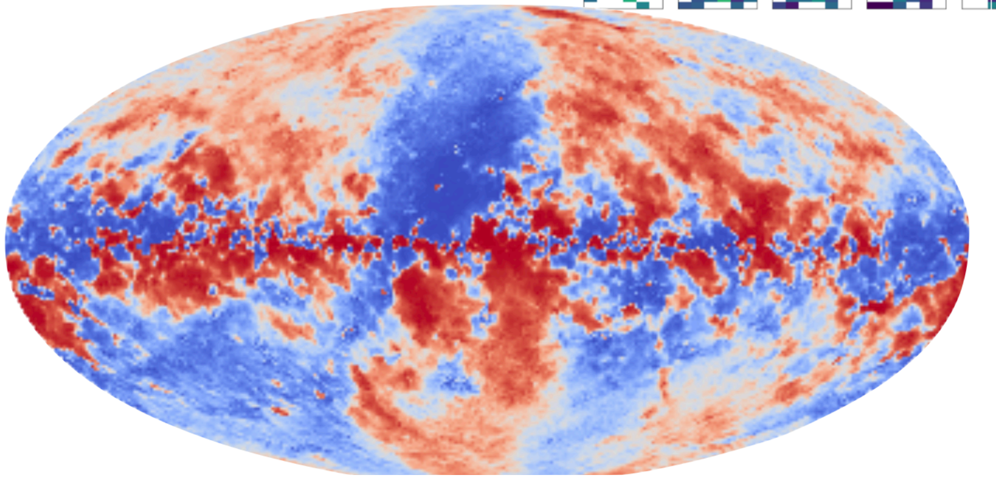
Other GAN applications

All of these pictures are fake!

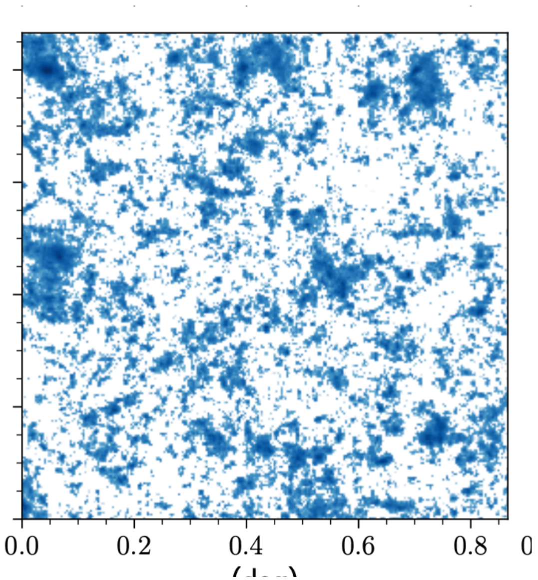
CaloGAN: Simulating 3D High Energy Particle Showers in Multi-Layer Electromagnetic Calorimeters with Generative Adversarial Networks



https://arxiv.org/abs/1712.10321



ForSE: a GAN based algorithm for extending CMB foreground models to sub-degree angular scales
https://arxiv.org/abs/2011.02221



CosmoGAN: creating high-fidelity weak lensing convergence maps using Generative Adversarial Networks
https://arxiv.org/abs/1706.02390

Selection from slide by B. Nachman, LBL

New horizons - challenges

- Explainable AI / interpretable ML
- Automated discovery
  - Physics laws inference or event of interest detection directly from data
  - Interesting discussion: https://www.frontiersin.org/articles/10.3389/frai.2020.00025/full
- Uncertainties evaluation
  - e.g.: **Uncertainty as a Form of Transparency: Measuring, Communicating, and Using Uncertainty -** https://arxiv.org/abs/2011.07586
- Learning on ensemble basis not only ‘event-based’
  - e.g.: **E Pluribus Unum Ex Machina: Learning from Many Collider Events at Once** - https://arxiv.org/abs/2101.07263
- Graph based ML – relational mapping
  - e.g.: **A Comprehensive Survey on Graph Neural Networks** - https://arxiv.org/abs/1901.00596
  - **…** there is an increasing number of applications where data are generated from non-Euclidean domains and are represented as graphs with complex relationships and interdependency between objects…

Selected problems in physics aware ML

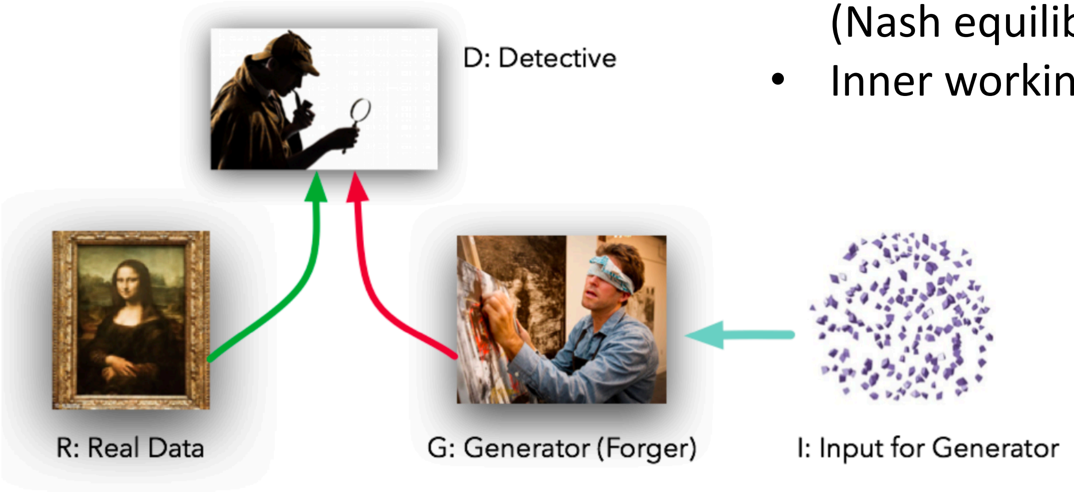
- Availability of models (& simulators)
  - Need for supervised vs. unsupervised (or partially supervised) learning
- Availability of data
  - ‘simple’ statistics – precision of measurements
    - Small variance – ‘mode collapse’ problem (also could be present in simulated sets)
- Quality of data
  - Purely labelled? not-labelled? Noisy or inconsistent?
  - Sufficient knowledge of uncertainties?

Extra Slides

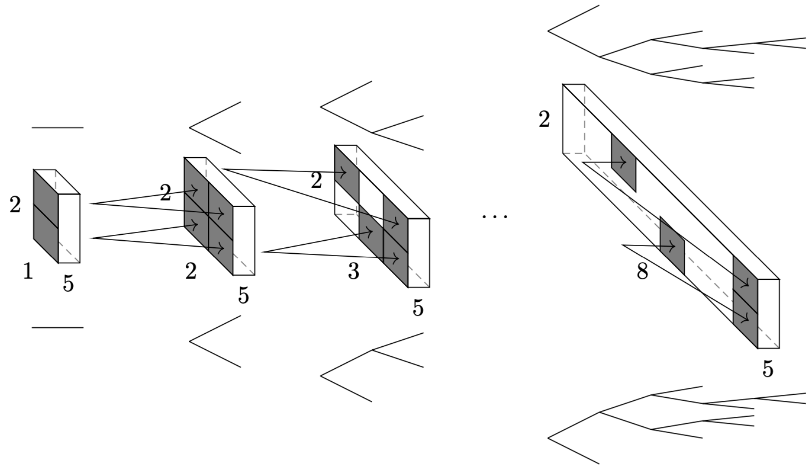
Recent example of **explainable**, physics aware ML

Explainable machine learning of the underlying physics of high-energy particle collisions
https://arxiv.org/abs/2012.06582
… proof-of-concept of our White Box AI approach using a **Generative Adversarial Network (GAN)** which learns from a DGLAP-based parton shower Monte Carlo event generator.
From "final state" particles learn internal workings of QCD - Altarelli-Parisi splitting function, the ordering variable of the shower, and the scaling behavior.

- Two neural networks: the generator ("forger") and discriminator ("detective")
- Simultaneously optimize both, causing both to be in competition with each other (Nash equilibrium)
- Inner workings (splitting kernels) of the NNs forced to produce physical splittings



https://medium.com/@devnag/generative-adversarial-networks-gans-in-50-lines-of-code-pytorch-e81b79659e3f



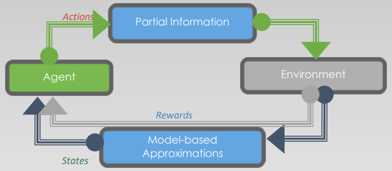
General scope ML4Sci(ence)

From AI4Sci / DOE

Data

- Experimental design
- Data curation and validation
- Compressed sensing
- Facilities operation and control

Learning

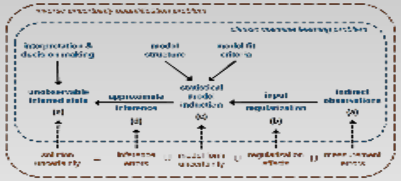


- Physics informed
- Reinforcement learning
- Adversarial networks
- Representation learning and multi-modal data
- "Foundational math" of learning

Scalability

- Algorithms, complexity and convergence
- Levels of parallelization
- Mixed precision arithmetic
- Communication
- Implementation on accelerated-node hardware

Assurance



- Uncertainty quantification
- Explainability and interpretability
- Validation and verification
- Causal inference

Workflow

- Edge computing
- Compression
- Online learning
- Federated learning
- Infrastructure
- Augmented intelligence
- Human-computer interface

From AI4Sci / DOE

AI Science Applications: One per Planet

AI Enabled

Design Workflows

(tell me what to make)

→ …materials, polymers, organisms…

AI Enabled

Experimental Workflows

(tell me how to make it)

→ …self-driving labs, synthesis search…

AI Enabled

Scientific Comprehension

(tell me what it means)

→

- data Sets
- literature
- science "news"
- strategy

→

Cleaned
Updated
Annotated
Aggregated
Interpreted

→ Insight?

Ai4Sci Subtopics

From AI4Sci / DOE

Science and Engineering

High Energy Physics	Nuclear Physics	Fusion	Energy Sciences	Earth Sciences	Biological Sciences	Energy Technologies	Computing Sciences
Cosmology and Astrophysics	Particle Identification	Feature Tracking	Materials and Chemistry Modeling	Climate & Carbon	Environmental Biology	Engineering and Manufacturing	AI for hardware / software
Particle Physics	Particle Tracking	Steering	Photon/Neutron Science	Subsurface	Imaging	Smart Grid/Urban	AI for networks / computing facilities
Accelerator Design and Control	Control	Design and Control	Electron Microscopy	Water	Syn Biology	Wind and Solar	
					Health	Mobility	

Crosscuts

- Foundations
- Data Lifecycle
- Software
- Hardware
- Computing Facilities

AI Science Services Building Blocks (examples)

Design
- Biology
- Materials
- Chemistry
- Batteries
- Devices
- Drugs

Control
- Accelerators
- Experiments
- Reactors
- Simulation
- Mobility
- Energy

Augmented Simulations
- Landscape Search
- Surrogates
- Optimize

Science and Math Comprehension
- Physics
- Mathematics
- Biochemistry

Generative Models
- Cosmology
- Detector Simulations
- Biodesign

Inverse Problems
- Spectra2 Structures
- Image2Phase
- Waveform2 Source

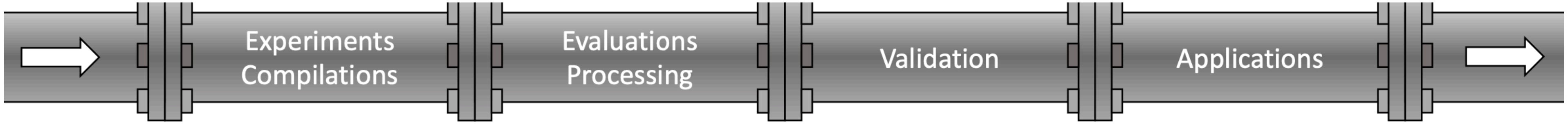
Multimodal Learning
- Text
- Waveforms
- Images
- Graphs
- Structured
- Time-series

Decision-Making
- Research Priorities
- Risk Assessment
- The Next Problem

From AI4Sci / DOE

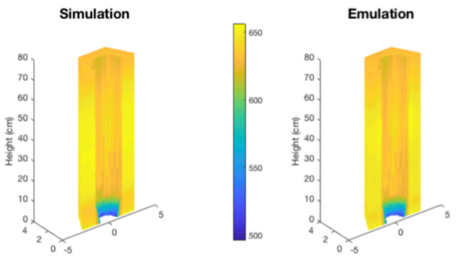
ML at WANDA

- 2020 report – this talk was not to repeat it…

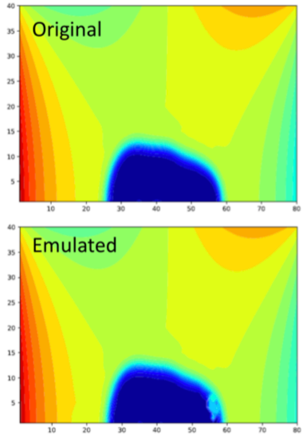


We strongly encourages entire ND community to embrace the advances that AI/ML tools can have for your work!

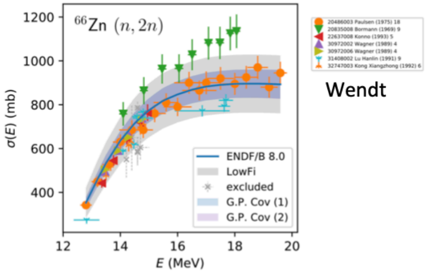
Emulation of expensive theory and application models for fast inference



Sobes



Schunck

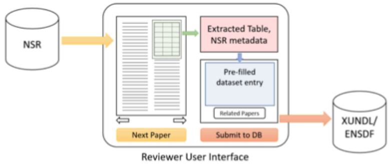


Wendt

Estimation of missing covariance information through AI/ML

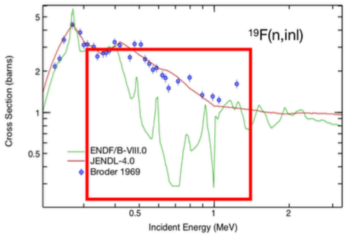
Yoo

NLP for processing of published articles in nuclear data



Neudecker

AI/ML-augmented Identification of regions to improve nuclear data



Neural Networks | Gaussian Processes | Supervised Learning | Generative Modeling | Reinforcement Learning | Deep Q Learning | Bayesian Optimization

Novel efforts – strength in community

[Submitted on 20 Jan 2021]

https://arxiv.org/abs/2101.08320

The LHC Olympics 2020: A Community Challenge for Anomaly Detection in High Energy Physics

A new paradigm for data-driven, model-agnostic new physics searches at colliders is emerging, and aims to leverage recent breakthroughs in anomaly detection and machine learning. In order to develop and benchmark new anomaly detection methods within this framework, it is essential to have standard datasets. To this end, we have created the LHC Olympics 2020, a community challenge accompanied by a set of simulated collider events. Participants in these Olympics have developed their methods using an R&D dataset and then tested them on black boxes: datasets with an unknown anomaly (or not). This paper will review the LHC Olympics 2020 challenge, including an overview of the competition, a description of methods deployed in the competition, lessons learned from the experience, and implications for data analyses with future datasets as well as future colliders.

Comments: 108 pages, 53 figures, 3 tables

Food for a thought:
Automated data mining

- Leverage NLP? – an example from other field…

Letter | Published: 03 July 2019

Unsupervised word embeddings capture latent knowledge from materials science literature

Vahe Tshitoyan, John Dagdelen, Leigh Weston, Alexander Dunn, Ziqin Rong, Olga Kononova, Kristin A. Persson, Gerbrand Ceder & Anubhav Jain

Nature **571**, 95–98(2019) | Cite this article

https://www.nature.com/articles/s41586-019-1335-8

Named Entity Recognition and Normalization Applied to Large-Scale Information Extraction from the Materials Science Literature | Journal of Chemical Information and Modeling

https://pubs.acs.org/doi/10.1021/acs.jcim.9b00470

https://github.com/materialsintelligence/matscholar

MATSCHOLAR

`matscholar` (Materials Scholar) is a Python library for materials-focused natural language processing (NLP). It is maintained by a team of researchers at UC Berkeley and Lawrence Berkeley National Laboratory as part of a project funded by the Toyota Research Institute.

This library provides a Python interface for interacting with the Materials Scholar API, performing basic NLP tasks on scientific text, and example notebooks on using these tools for materials discovery and design.

Documentation for the API can be found in this readme, as well as in the jupyter notebook: docs/demo.ipynb. If the notebook fails to render on github, paste the link into nbviewer: https://nbviewer.jupyter.org.

You can find our official support forum here, under the "Matscholar" category: https://dicuss.matsci.org