Machine Learning in Particle Physics

Matthew Schwartz
Harvard University

Aug 11, 2023

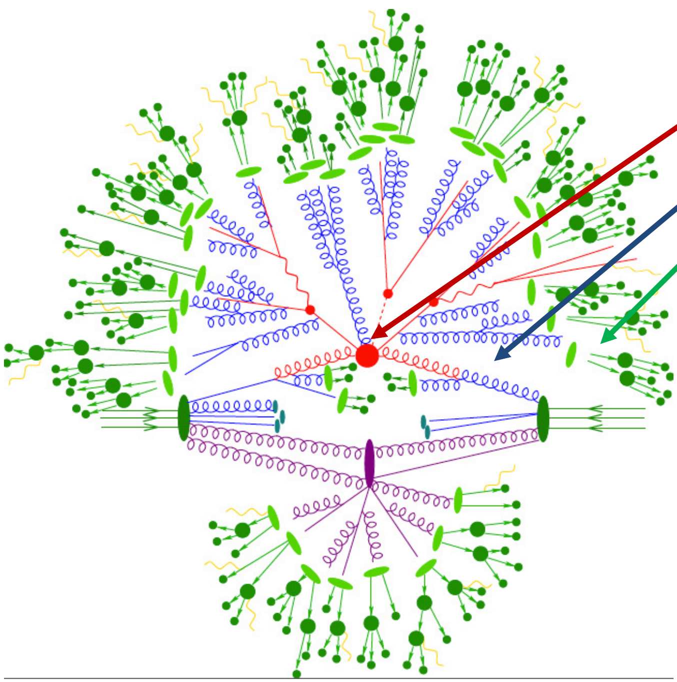
AI generated image by
Việt Vun Vút


Hà Ly Phúc Toàn
Harry Potter
ART
AIhug.ART
fb.com/AIhugART
ảnh báo cho người xem: Sản phẩm này do AI Art Generator tạo ra mang tính chất tham khảo để lên ý tưởng, dùng để học
i hoặc truyền cảm hứng, không dùng cho mục đích thương mại. Đọc kỹ hướng dẫn trước khi xem/ sử dụng.

How is ML used in particle physics?

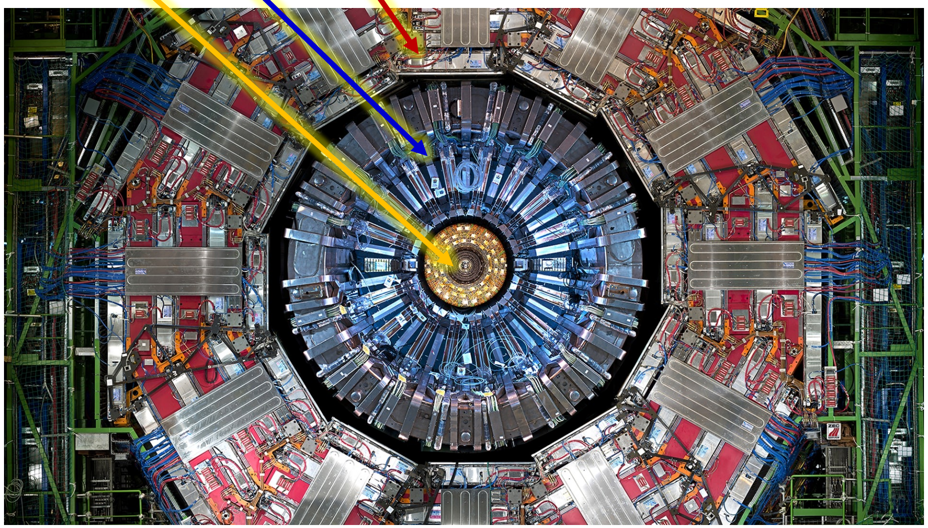
Main use so far is in **collider physics**

- Can generate billions of simulated events on a laptop
- Simulations accurate over **20 orders of magnitude**!
- Unheard of in any other area of science

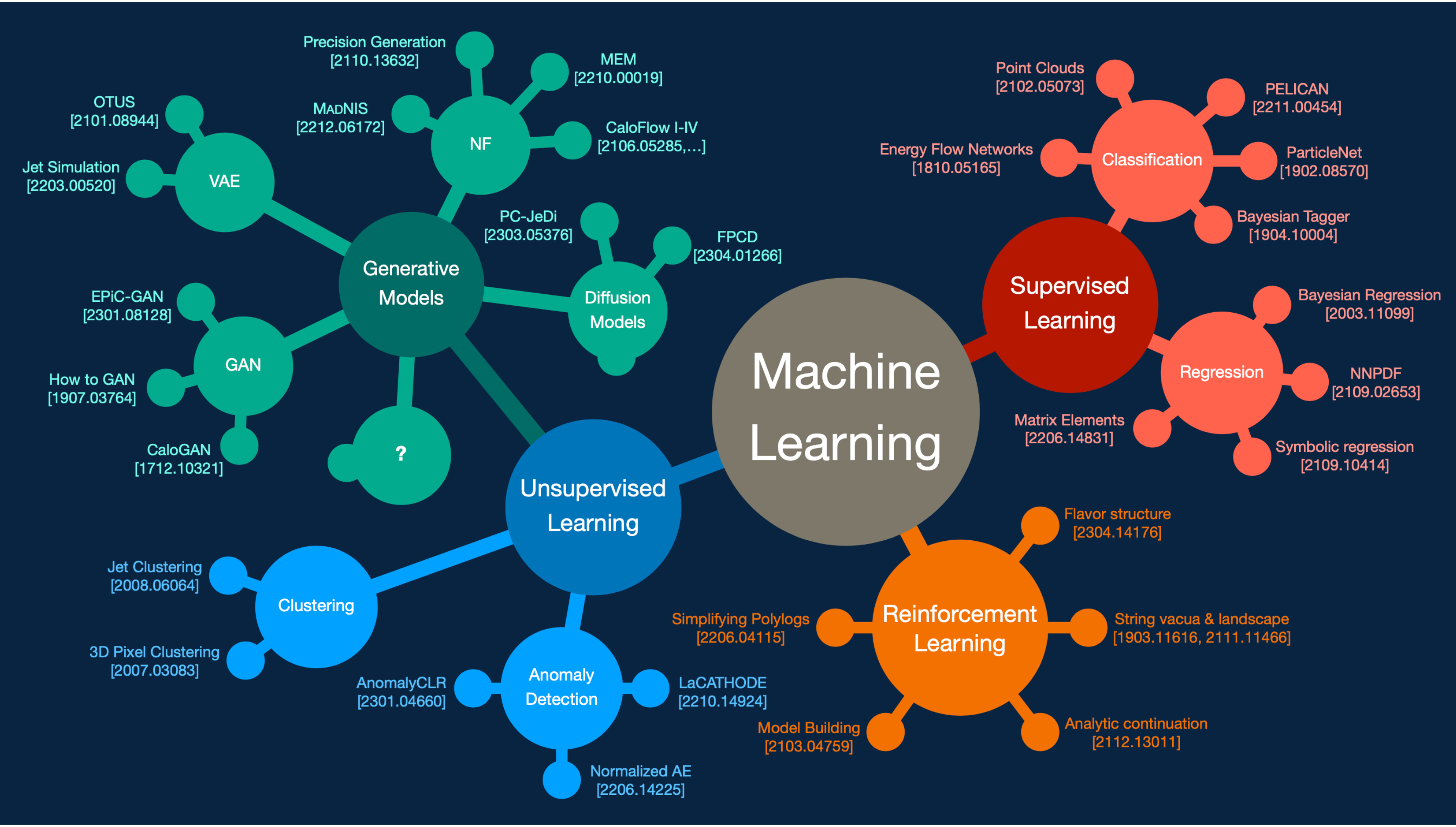


Simulation pipeline:

- Perturbative QCD: 10^{-19}m - 10^{-16} m
- Parton shower: 10^{-16} m – 10^{-14} m
- Hadronization/framentation 10^{-14}m – 10^{-12} m
- tracker/ecal/hcal simulation 10^{-12} m – 10^{2} m



Most results heavily use these simulations



Slide from R. Winteralder

Current areas of progress

1. Lattice QCD
    - Normalizing flows, Monte Carlo sampling, Spectral reconstruction, …

2. Simulation/unfolding
    - Learn to reproduce simulations with a neural network
    - Can speed up simulations by factors of 10^3 - 10^5
    - Can be used for unfolding: remove effects of simulation on data

3. Anomaly detection
    - Search for deviations from background
    - No signal hypothesis necessary (?)

4. Data representation
    - Can ML provide a better way to categorize and understand data?
    - e.g. optimal transport, graph networks, etc.

5. Classification
    - Top tagging, W tagging, Q v G discrimination, new physics searches

6. Symbolic regression
    - Large language models (chatGPT) ?

Anomaly detection

Anomaly detection

The Dream:

- ML sees something unusual in the data, new physics is found!
- dream is an **unsupervised** method: do not need a signal hypothesis
- Way to find “unknown unknowns”

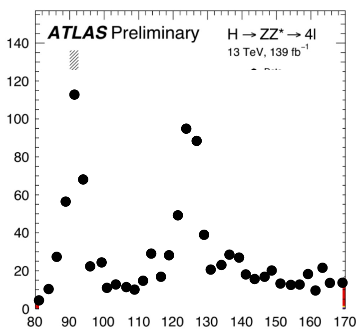
The main idea:

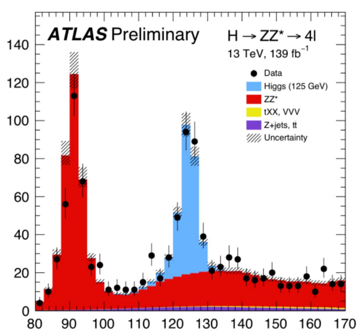
- Background is understood well enough by ML that statistical outliers are seen

Easy: outliers

$h \to e^- e^+ \mu^- \mu^+$

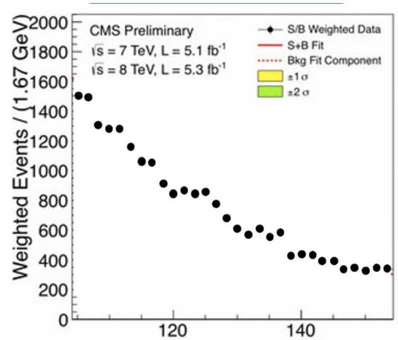
Hajer et al:1807.10261
Heimel et al: 1808.08979
Farina et al: 1808.08992
Cerri et al: 1811.10276
Roy + Vijay: 1903.02032,
Atkinson et al: 2105.07988
Carron et al: 2106.10164
Ngairangbam et al: 2112.04958
…

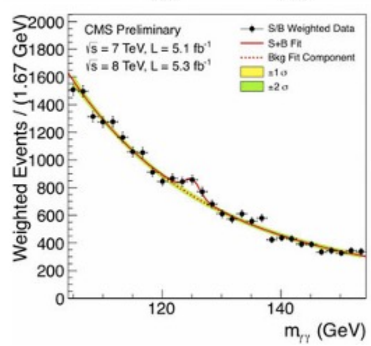




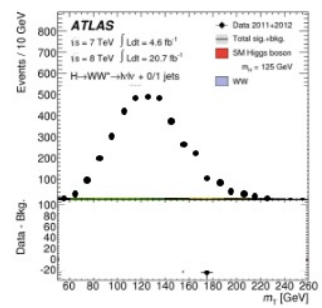
Hard: overdensities

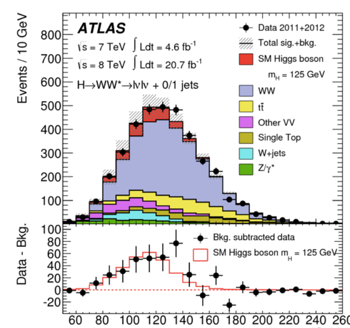
$h \to \gamma\gamma$





$h \to e^- \mu^+ \bar{\nu} \nu$



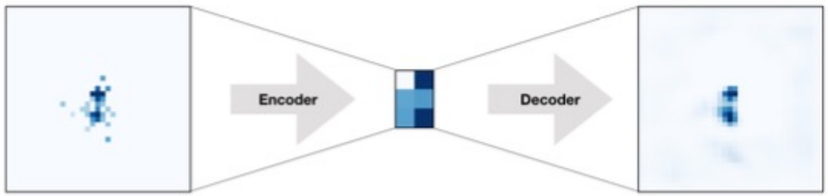


Collins et al: 1805.02664
D’Anglo and Wulzer: 1806.02350
Collins et al: 1902.02634
D’Anglo et al: 1912.12155
Nachman & Shih: 2001.04990
Stein et al: 2012.11638
Carron et al: 2106.10164
Hallin et al: 2109.00546
…

Autoencoders

Variational autoencoders: [Farina et al: 1808.08992]

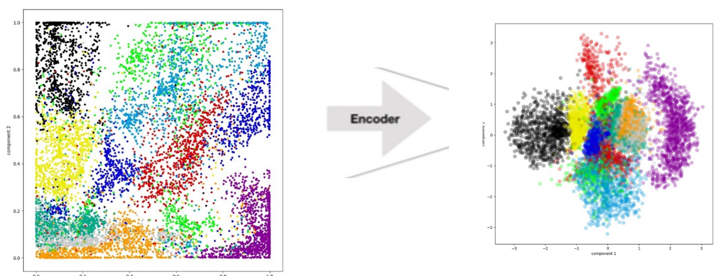
- Compress the background/data to a low-dimensional latent space



- Uncompress back to data space.
- Poorly reconstructed events are anomalies

Fraser, MDS, et al: 2110.06948

- Can look for anomalies directly in low-dimensional latent space



			Top jet		W jet	
Metric	Number of medoids	Method	AUC	$\epsilon_S(\epsilon_B = 0.1)$	AUC	$\epsilon_S(\epsilon_B = 0.1)$
Wass(1)	-	Avg	0.81	0.33	0.62	0.02
	1	Medoid	0.83	0.28	0.63	0.02
	3 (elbow)	Medoids (min)	0.85	0.43	0.67	0.04
	5	Medoids (min)	**0.87**	**0.54**	0.60	0.05
	7	Medoids (min)	0.87	0.54	0.61	0.05
Wass(5)	4 (elbow)	Medoids (min)	0.67	0.22	0.41	0.04
MAE	1	Medoid	0.82	0.40	**0.71**	0.07
	3 (elbow)	Medoids (min)	0.82	0.49	0.61	**0.08**

- Use k medoids or Wasserstein optimal transport metric
- Use event-to-ensemble distance for anomaly score
- Take home messages:
  - Performance depends on metric and sample
  - Cannot optimize in signal-independent manner

$$d^{(p)}_{Wass} = \left(min_f \sum_{i,j} f_{ij} (c_{ij})^p \right)^{1/p}$$

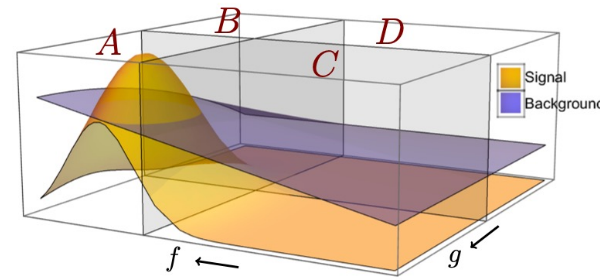
ABCDisCo: ML the ABCD method

ABCD method:

- Standard experimental sideband technique
- Estimate background in region A via $N_A = \frac{N_B N_C}{N_D}$
- Requires two features f and g to be uncorrelated
  - E.g. f = mass and g = rapidity

Distance Correlation (DisCo): alternative to adversarial networks

- Decorrelates observables, easy to train



Kasieczka and Shih [arXiv: 2001.05310]

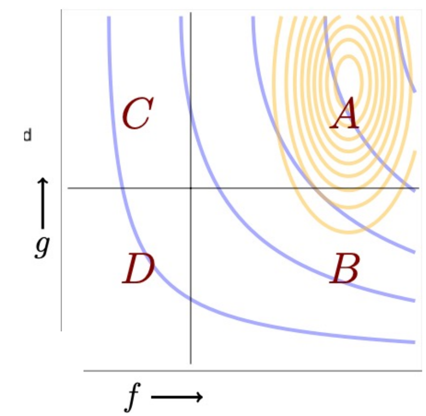
Kasieczka, Nachman, MDS, Shih [arXiv: 2007.14400]

- Single DisCo

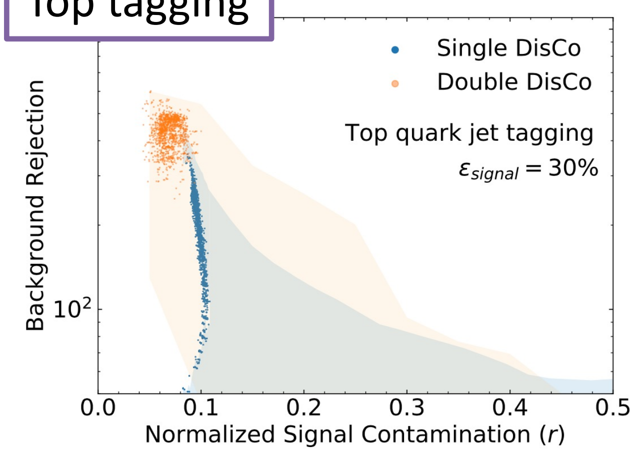
f is fixed (e.g. mass) g is learned

- Double DisCo

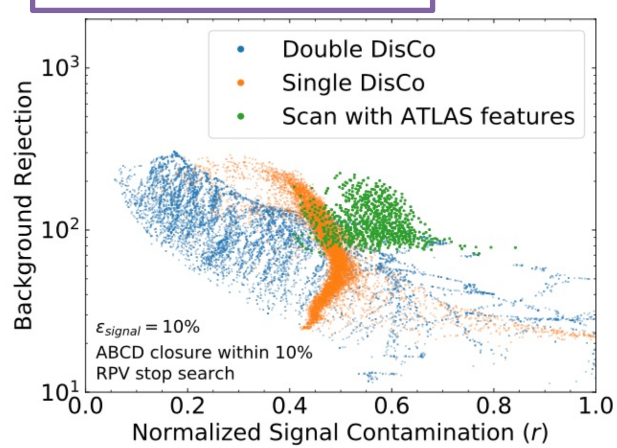
f and g are learned



Top tagging



RPV squark search



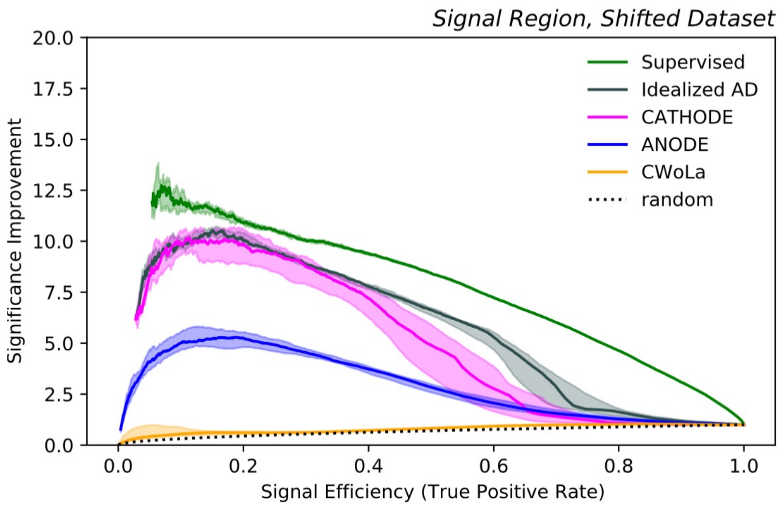
Challenges for Anomaly Detection

- Background regions are **signal dependent**
  - e.g. if looking for a dijet resonance, need a dijet background
  - No such thing as a signal-independent background

- **Looking under the lampost**
  - We only know how to look for resonances/new particles
  - Different signals are too varied to be pooled together

- Very **sensitive to metric** for what is anomalous
  - Tails of backgrounds are unique in their own way

- Supervised classifiers always do better



[Hallin et al. 2109.00546]

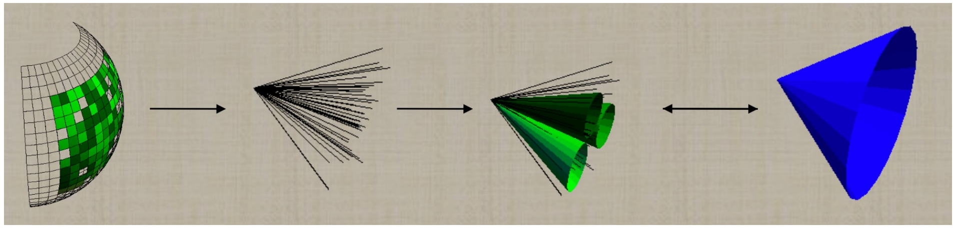
Classification

Classification: top tagging

e.g. top tagging

Jet substructure approach (2008-2017):

- Think about physics
- Deconstruct jet
- Look for W within top jet
- Look at helcity angle
- Hopkins Top Tagger (2008)



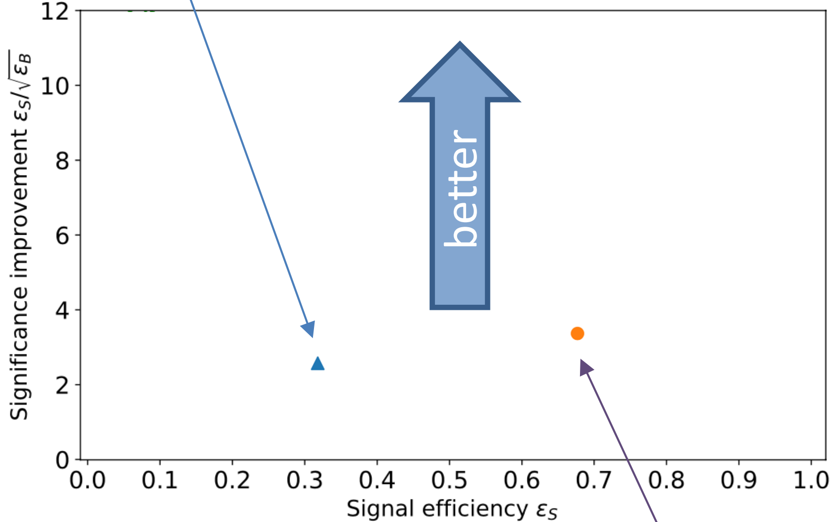


fig from 2103.12226

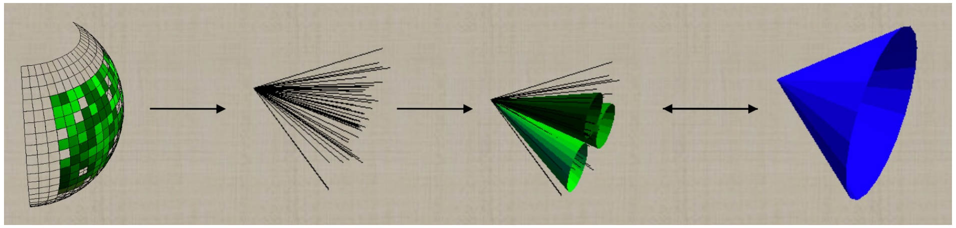
best jet subsubstructure algorihm (2014)

Classification: top tagging

e.g. top tagging

Jet substructure approach (2008-2017):

- Think about physics
- Deconstruct jet
- Look for W within top jet
- Look at helcity angle
- Hopkins Top Tagger (2008)



Machine learning methods are much better

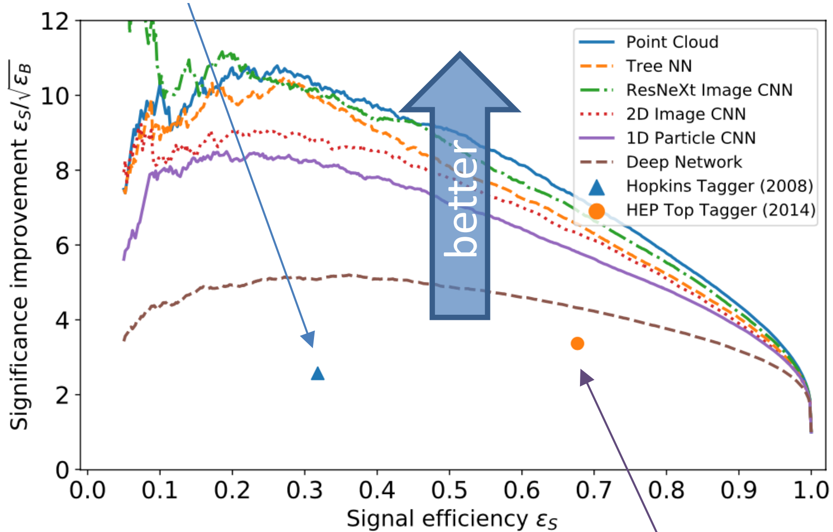
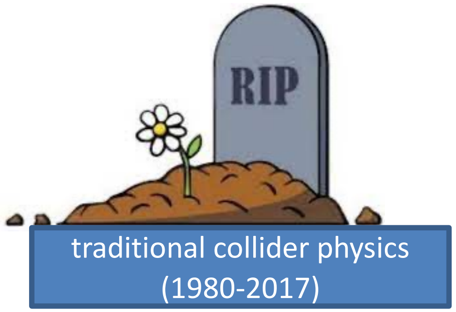


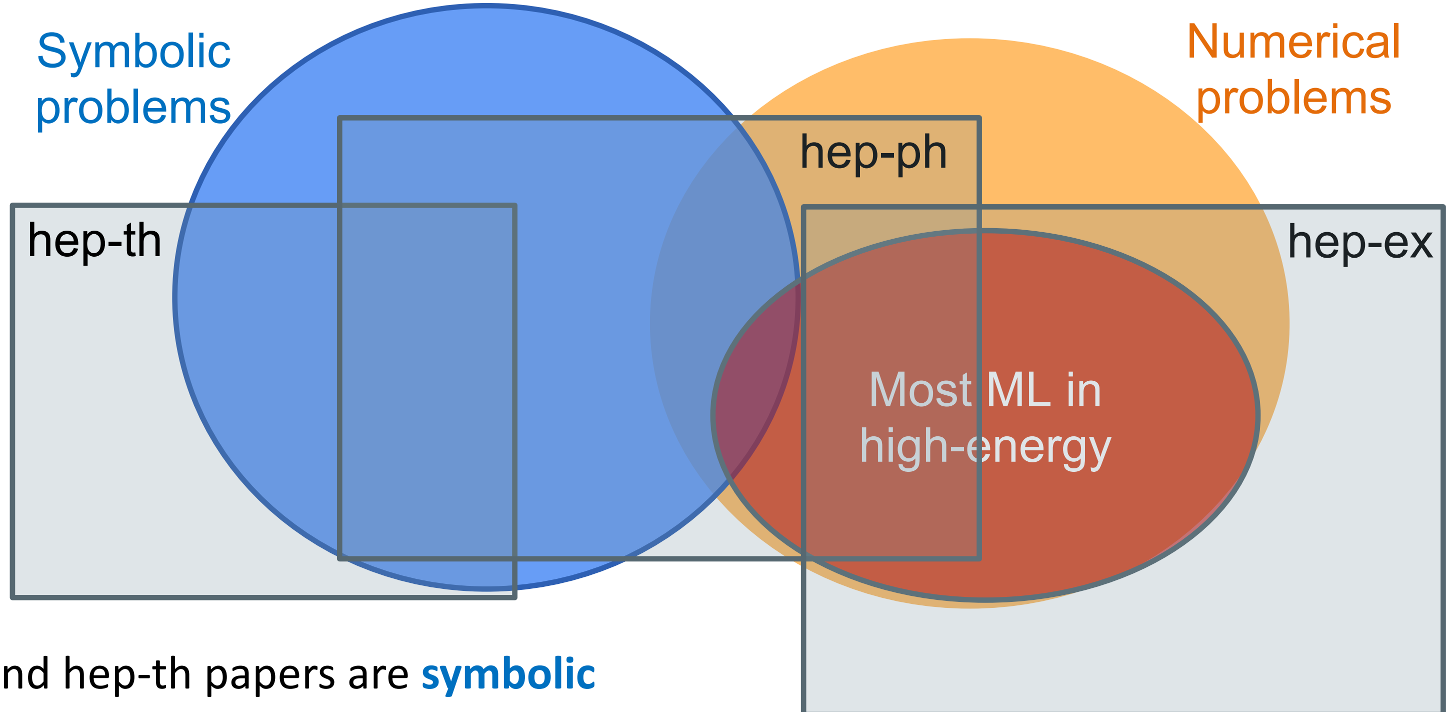
fig from 2103.12226

best jet subsubstructure algorihm (2014)



- ML requires less “thinking”
- Provdies less physical insight
- Better performance

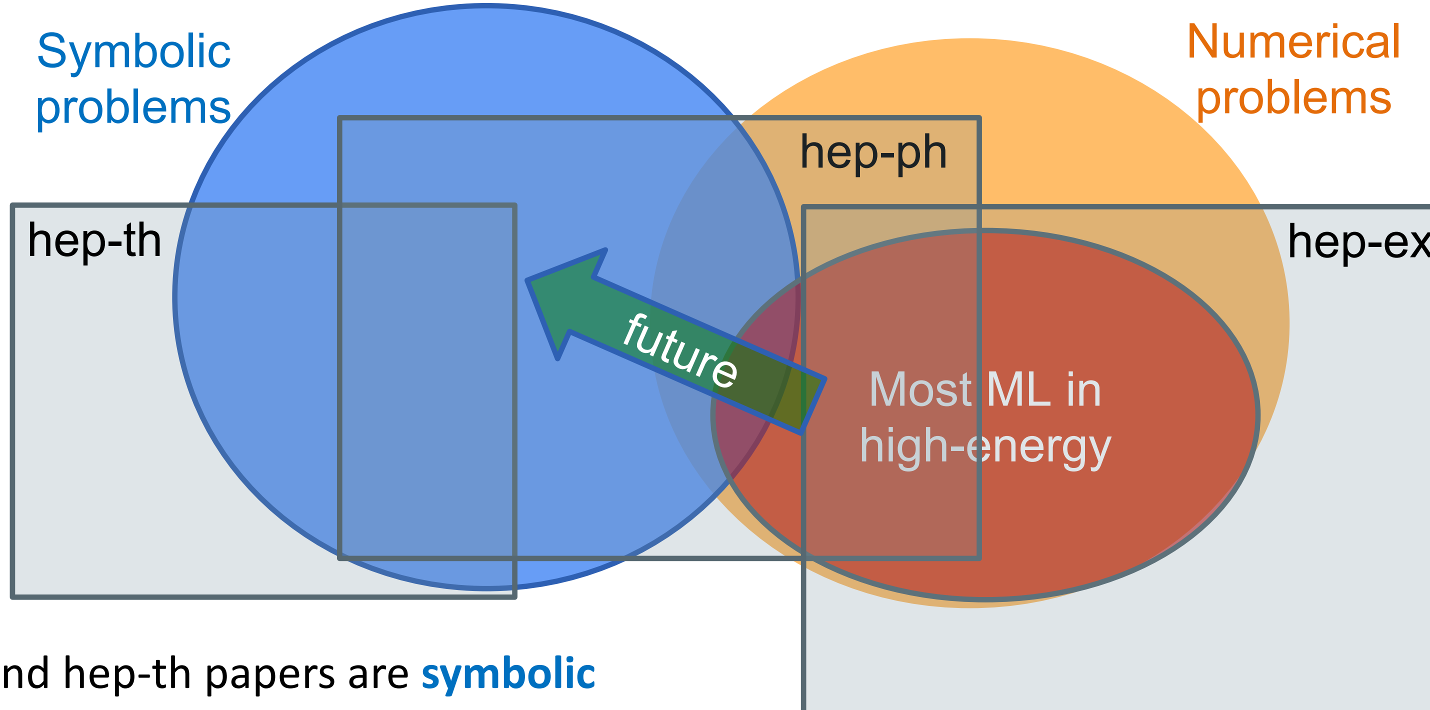
What subfield will ML **make obsolete** next?



Most hep-ph and hep-th papers are **symbolic**
- Qualitative understanding of some simplified model
- Approximate but exact solutions to some equation
- Analytic computations in some system

So far, most ML in physics is highly **numerical**
- Collider physics data is millions of numbers
- Approximate answers are ok

What subfield will ML **make obsolete** next?



Most hep-ph and hep-th papers are **symbolic**

- Qualitative understanding of some simplified model
- Approximate but exact solutions to some equation
- Analytic computations in some system

So far, most ML in physics is highly **numerical**

- Collider physics data is millions of numbers
- Approximate answers are ok

Symbolic ML methods will be **essential** for the future of High Energy Physics

- The world is changing because of symbolic large langauge models

Symbolic regression

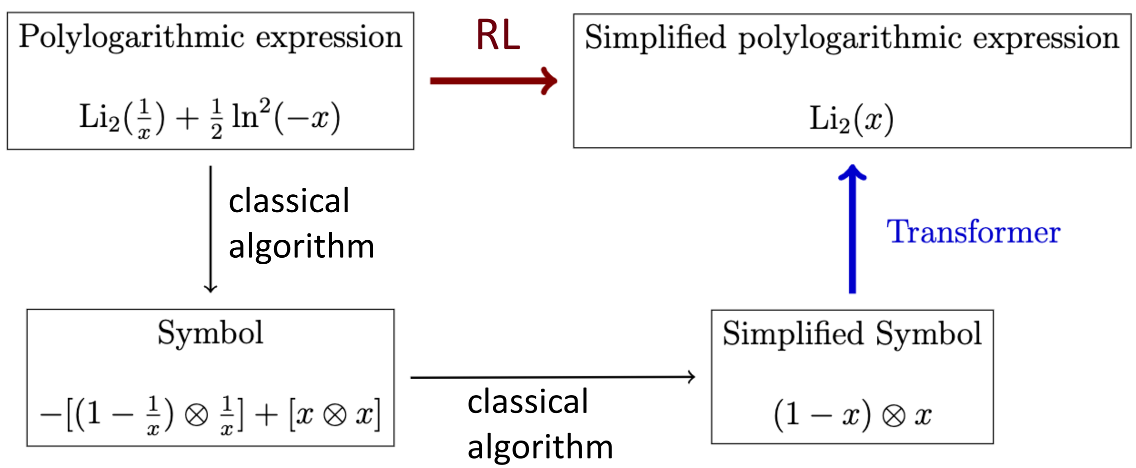
1. Simplifying polylogarithms

Given some polylogarithmic expression: Dersy, Schwartz, Zhang arXiv:2206.04115)

$$
\begin{aligned}
f(x) = 9\left(-\mathrm{Li}_3(x) - \mathrm{Li}_3\left(\frac{2ix}{-i+\sqrt{3}}\right) - \mathrm{Li}_3\left(-\frac{2ix}{i+\sqrt{3}}\right)\right) \\
+4\left(-\mathrm{Li}_3(x) + \mathrm{Li}_3\left(\frac{x}{x+1}\right) + \mathrm{Li}_3(x+1) - \mathrm{Li}_2(-x)\ln(x+1)\right) \\
-4\left(\mathrm{Li}_2(x+1)\ln(x+1) + \frac{1}{6}\ln^3(x+1) + \frac{1}{2}\ln(-x)\ln^2(x+1)\right)
\end{aligned}
$$

1. What is its simplest form?
2. Does it simplify to zero?
3. What identities do we apply in what order to simplify it?

Two approaches
1. Reinforcement learning
2. Transformer networks



Dersy, Schwartz, Zhang arXiv:2206.04115)

Example use case

1. Loop calculation gives some function of GPLs with complex arguments

$$f(x) = 4\zeta_3 + 9\left[G(0,0,1,x) + G\left(0,0,\frac{-1-\sqrt{3}i}{2},x\right) + G\left(0,0,\frac{-1+\sqrt{3}i}{2},x\right)\right]$$
$$+4\left[-G(-1,-1,-1,x) + G(-1,0,-1,x) + G(0,-1,-1,x) + G(0,0,1,x) - G\left(0,0,1,\frac{x}{x+1}\right)\right]$$

2. Express in terms of classical polylogs

$$f(x) = 9\left(-\mathrm{Li}_3(x) - \mathrm{Li}_3\left(\frac{2ix}{-i+\sqrt{3}}\right) - \mathrm{Li}_3\left(-\frac{2ix}{i+\sqrt{3}}\right)\right)$$
$$+4\left(-\mathrm{Li}_3(x) + \mathrm{Li}_3\left(\frac{x}{x+1}\right) + \mathrm{Li}_3(x+1) - \mathrm{Li}_2(-x)\ln(x+1)\right)$$
$$-4\left(\mathrm{Li}_2(x+1)\ln(x+1) + \frac{1}{6}\ln^3(x+1) + \frac{1}{2}\ln(-x)\ln^2(x+1)\right)$$

3. Compute the symbol and simplify

$$\mathcal{S}[f(x)] = 9\left(x^2+x+1\right)\otimes x\otimes x + 13(1-x)\otimes x\otimes x + 4(x+1)\otimes x\otimes x$$

4. Integate the symbol with a transformer network

$$f(x) = -\mathrm{Li}_3(x^3) - \mathrm{Li}_3(x^2) + 4\zeta_3 \quad \checkmark$$

highly non-trivial
powerful dedicated
neural network

2. Simplifying Spinor-helicity amplitudes

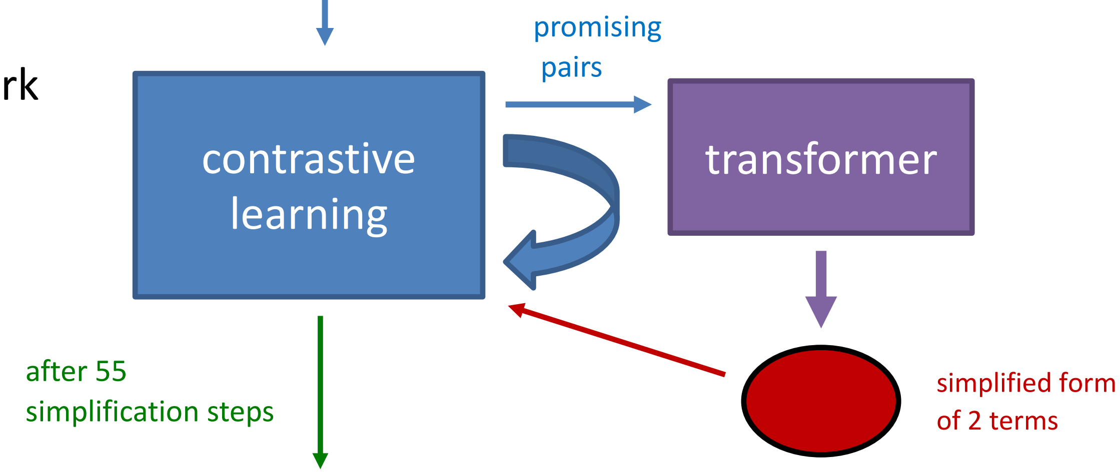
[Cheung, Dersy, MDS, in perparation]

1. Compute 5-point MHV amplitude with Feynman diagrams: 390 terms
2. Choose some smart reference vector to reduce to 79 terms

   (smarter choice can reduce to 17)

$$\frac{\langle 12\rangle^2\langle 15\rangle^2\langle 24\rangle\langle 34\rangle[12]\,[14]\,[15]\,[23]\,[25] + \langle 12\rangle^2\langle 15\rangle\langle 23\rangle\langle 34\rangle\langle 45\rangle[12]\,[15]\,[23]\,[25]\,[34] + \cdots 77\text{ terms}}{\langle 15\rangle^2\langle 23\rangle\langle 34\rangle^2\langle 45\rangle^2[12]^2\,[15]\,[23]\,[45]}$$

3. Feed to network



$$-\frac{\langle 12\rangle^3}{\langle 15\rangle\langle 23\rangle\langle 34\rangle\langle 45\rangle}$$

4. Output when simplification completes

3. S-matrix bootstrap

What is the S-matrix bootstrap?

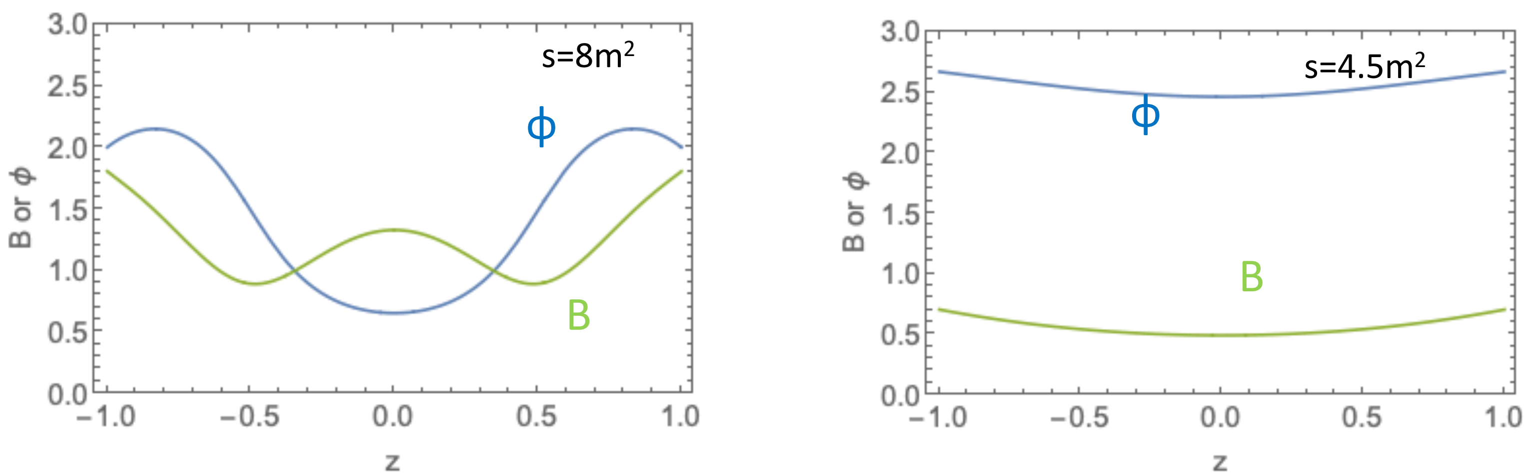
- Use analyticity, unitarity, crossing symmetries, etc. to completly fix S
- Stalled in 1960s: math too hard
- Recent revitalization: new insights from toy models, susy, numerical methods, etc.

Penedones et al 1708.06756
Fitzpatrick et al 2207.12448

Example ML application: Dersy, MDS, Zhiboedov, to appear

For a given cross section $\sigma \sim |F|^2$

- Does there always exist a phase ϕ so that $F = Be^{i\phi}$?



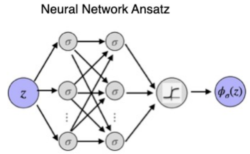
Open questions:

A. How do we determine ϕ from B?

B. Can there be many phases ϕ_1, ϕ_2, ... for the same B?

A. Can we find ϕ(z) given B(z)? ... Yes!

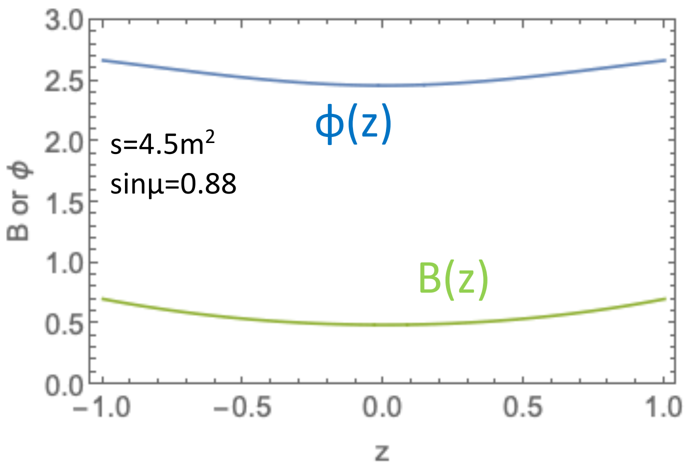
- Parametrize ϕ(z) as a neural network



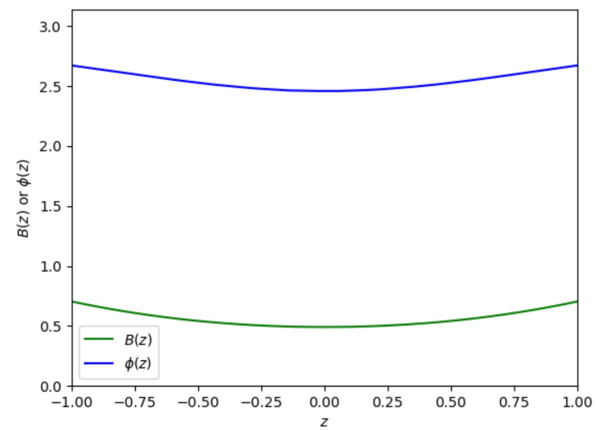
- Loss function is unitarity condition

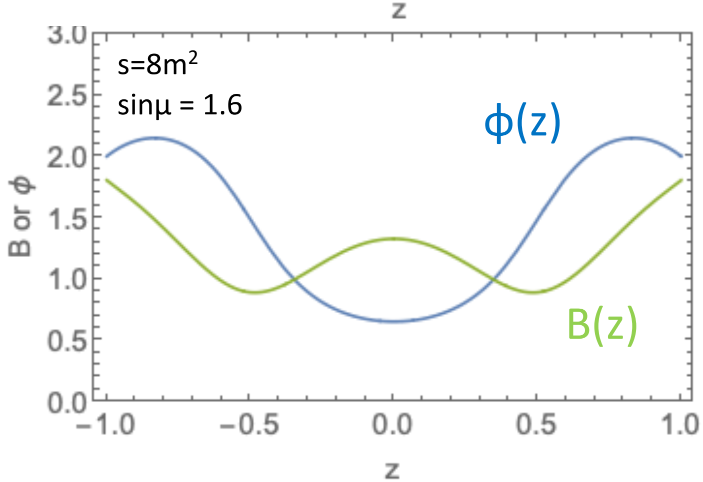
$$\mathscr{L} = \mathbb{E}\left|\left|B(z)\sin\phi(z) - \frac{1}{4\pi}\int_{-1}^{1} dz_1 \int_0^{2\pi} d\phi_1 B(z_1)B(z_2)\,\cos\left(\phi(z_1)-\phi(z_2)\right)\right|\right|^2$$

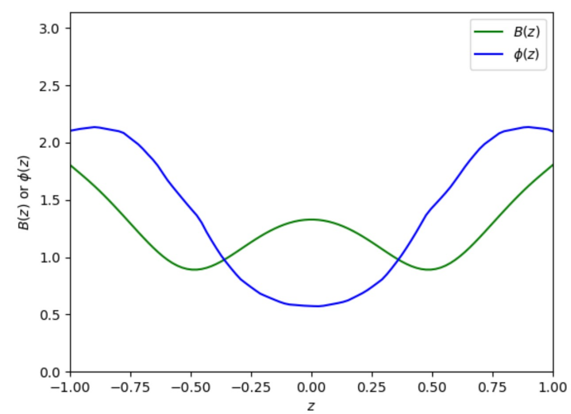
S-matrix bootstrap results



ϕ from B using ML

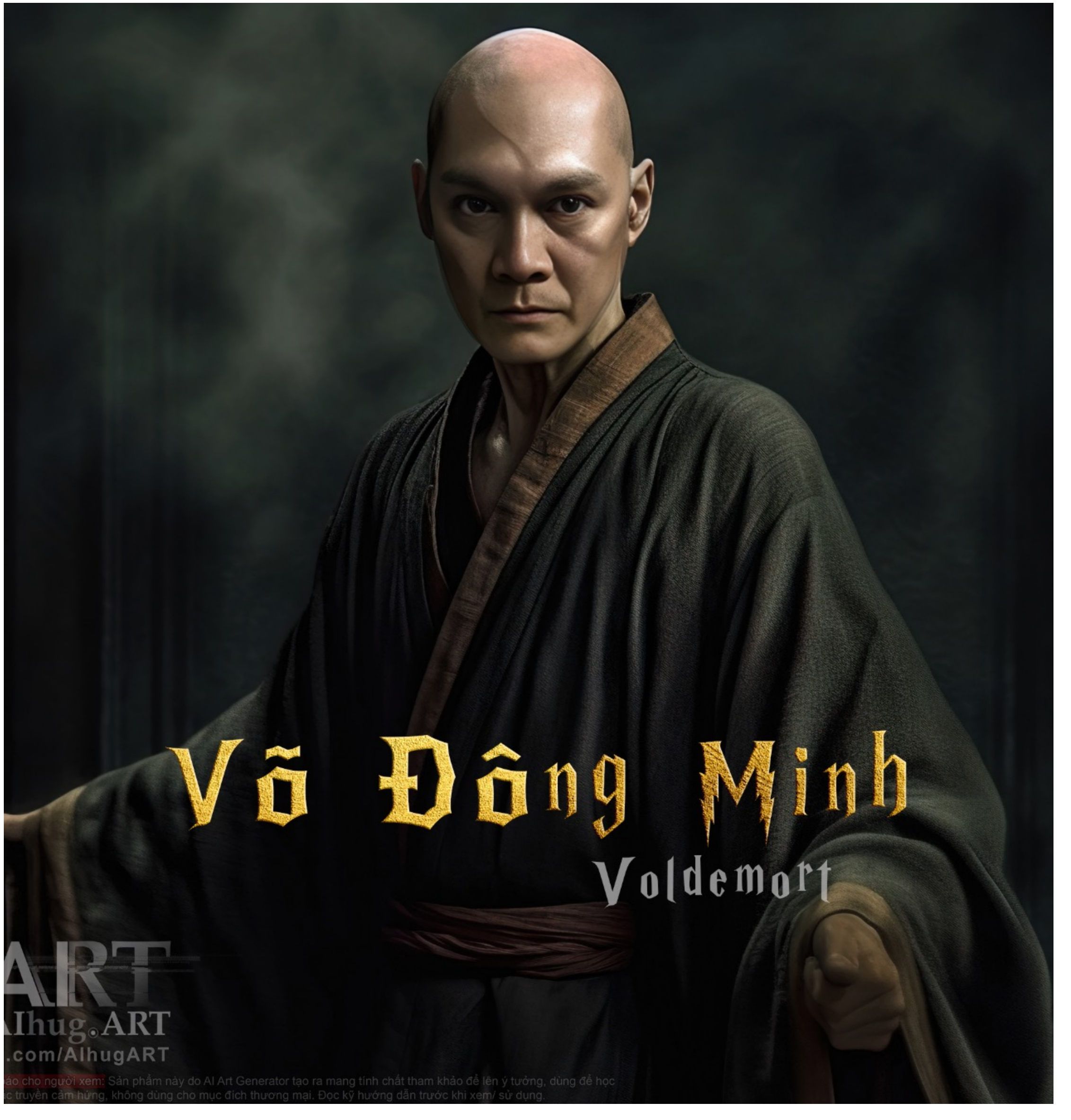




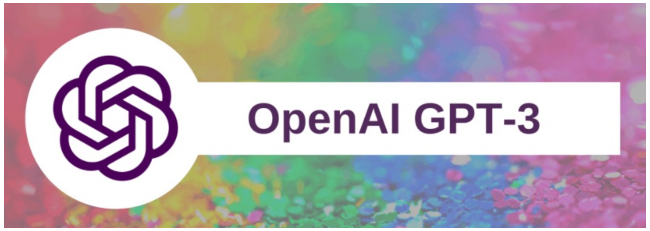


excellent agreement with known results

The future


Võ Đông Minh
Voldemort
ART
Ihug.ART
.com/AIhugART
áo cho người xem: Sản phẩm này do AI Art Generator tạo ra mang tính chất tham khảo để lên ý tưởng, dùng để học
c truyền cảm hứng, không dùng cho mục đích thương mại. Đọc kỹ hướng dẫn trước khi xem/ sử dụng.

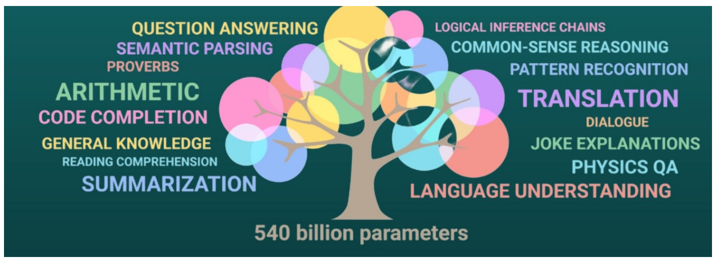
Future of AI



- 3rd generation model (2020)
- 175 billion parameters

Google : PaLM (2022)

540 billion parameters

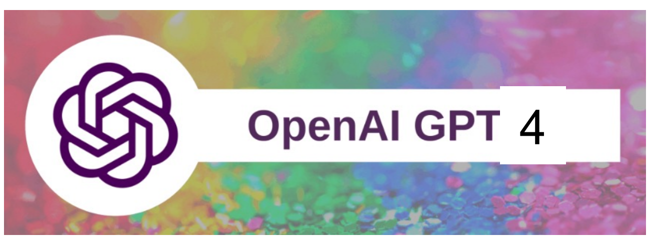


Human brain

80 billion neurons
150 trillion synapses

Cat brain

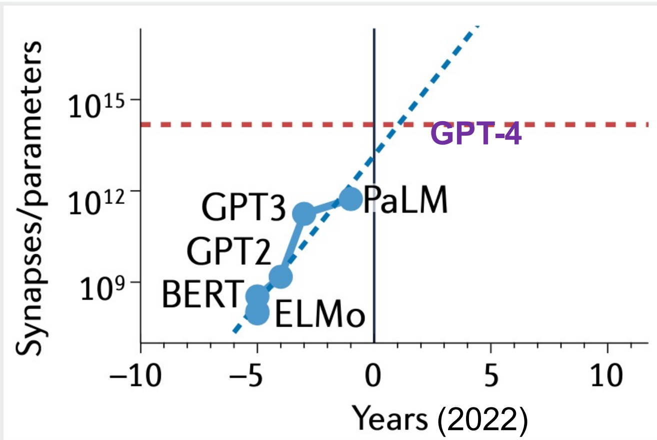
0.760 billion neurons
10 trillion synapses



170 trillion parameters!

2023 ... 2040... 2100? ...3000...?

Future of AI

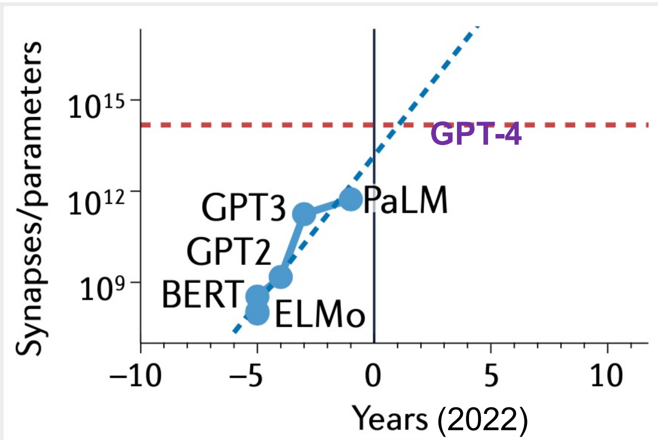


Should artificial intelligence be interpretable to humans? MDS, Nature reviews physics (2022)

- ELMo (94 million parameters, 2018)
- GPT2 (1.5 billion parameters, 2019)
- GPT3 (175 billion parameters, 2020)
- PALM (540 billion parameters, 2022)
- GPT4 (110 trillion parmaeters, 2023)

AI grows by factor of ~10/year

Future of AI



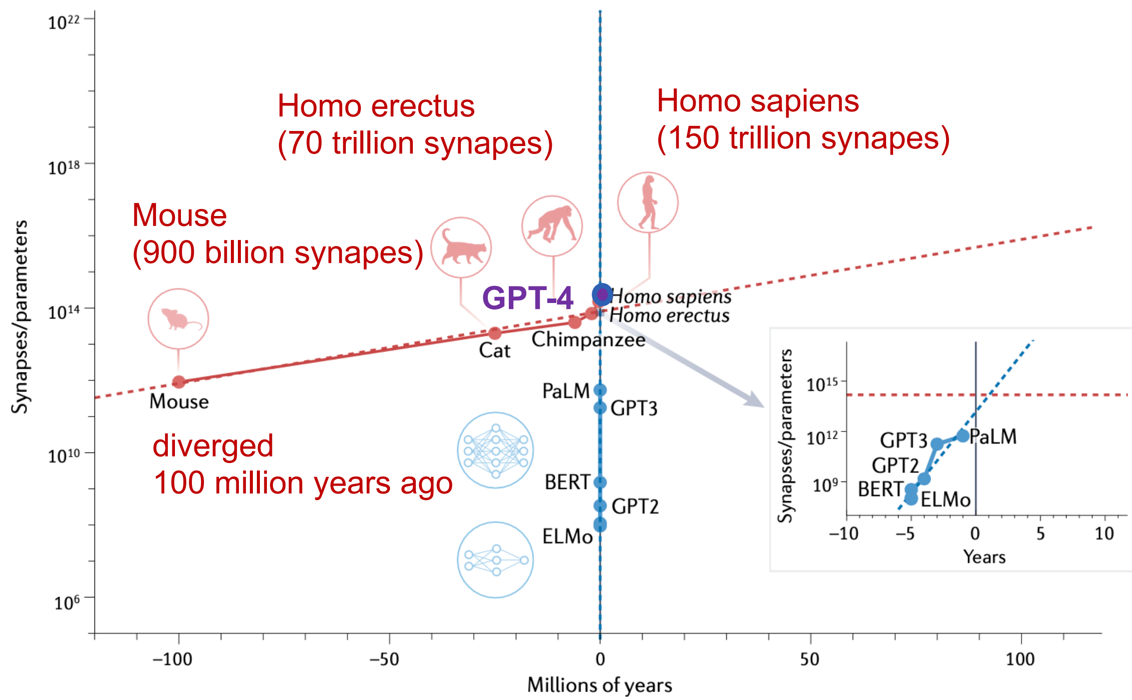
Should artificial intelligence be interpretable to humans? MDS, Nature reviews physics (2022)

- ELMo (94 million parameters, 2018)
- GPT2 (1.5 billion parameters, 2019)
- GPT3 (175 billion parameters, 2020)
- PALM (540 billion parameters, 2022)
- GPT4 (110 trillion parmaeters, 2023)

AI grows by factor of ~10/year

Biological intelligence grows by a factor of 2 in 10^6 years

- Both AI and biological intelligence grow exponentially
- Factor of 10^6 difference in exponent (!!)



Does AI look like biological intelligence?

Exam	GPT-4
SAT Math	700 / 800 (~89th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)
AP Environmental Science	5 (91st - 100th)
AP Macroeconomics	5 (84th - 100th)
AP Microeconomics	5 (82nd - 100th)
AP Physics 2	4 (66th - 84th)

- GPT4 can do freshman level college physics very well
- In ~ 1 year will do it perfectly
- In ~ 3 years it will have mastered college

MA Write me a good sophmore level physics problem about vietnam

Problem: Cable Car in Ba Na Hills

The Ba Na Hills cable car near Da Nang, Vietnam, holds the world record for the longest non-stop single track cable car at 5,801 meters in length. It also boasts a staggering height difference between its departure and arrival station, offering passengers an impressive view of the surrounding scenery.

Imagine a physics experiment conducted aboard this cable car. A student suspends a simple pendulum of length 0.5 meters from the ceiling of the cable car. The pendulum consists of a small weight at the end of a lightweight, inextensible string. As the cable car ascends the hill with a constant upward acceleration of $a = 0.2\ \mathrm{m/s}^2$, the student notices the pendulum no longer hangs vertically but makes an angle θ with the vertical line.

1. Determine the angle θ that the pendulum makes with the vertical due to the acceleration of the cable car.

Hint: Consider the effective gravitational acceleration in the frame of reference of the accelerating cable car.

Solution...

Thus, due to the acceleration of the cable car, the pendulum makes an angle of approximately 1.17° with the vertical.

Does AI look like biological intelligence?

Exam	GPT-4
SAT Math	700 / 800 (~89th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)
AP Environmental Science	5 (91st - 100th)
AP Macroeconomics	5 (84th - 100th)
AP Microeconomics	5 (82nd - 100th)
AP Physics 2	4 (66th - 84th)

- GPT4 can do freshman level college physics very well
- In ~ 1 year will do it perfectly
- In ~ 3 years it will have mastered college

Can the improvements continue?

- Already use all available training data (the internet)

Does AI look like biological intelligence?

Exam	GPT-4
SAT Math	700 / 800 (~89th)
Graduate Record Examination (GRE) Quantitative	163 / 170 (~80th)
Graduate Record Examination (GRE) Verbal	169 / 170 (~99th)
AP Environmental Science	5 (91st - 100th)
AP Macroeconomics	5 (84th - 100th)
AP Microeconomics	5 (82nd - 100th)
AP Physics 2	4 (66th - 84th)

- GPT4 can do freshman level college physics very well
- In ~ 1 year will do it perfectly
- In ~ 3 years it will have mastered college

Can the improvements continue?

- Already use all available training data (the internet)

Yes!

- It would be foolish to say this is the endpoint
- ML can learn just like we do
  - Data augmentation
  - Create and solve toy problems
  - Fewer sociological pressures than human beings have

Conclusions

Machine learning is having a huge impact on high energy physics

- QCD
- Anomaly detection
- Detector simulation/unfolding
- Jet substructure/classification

Future of ML in high energy will likely be more symbolic

- Begin with hybrid numerical/symbolic problems
    - (polylogarihms, spinor helicities, Feynman diagrams)
- Eventually exploit large language models to understand physical systems like we do

The future is bright!

ML provides hope at finally solving problems too dificult for human beings