The FERMIACC: Agents for Particle Theory

Amalia Madden
Kavli Institute for Theoretical Physics, UC Santa Barbara

Based on work with P. Agrawal, N. Craig, and I. Valenzuela Lombera

The Big Picture

LLMs are poised to make a huge impact on how we conduct scientific research.

We demonstrate a prototype of an agentic LLM system for automated BSM model building and analysis.

Built-in, hard-coded verification will ensure that outputs are reliable and scalable.

Our example will be based on collider physics, but the ideas are broadly applicable.

The “Bitter Lesson” of AI

- Deep Blue used a brute force search method that could explore millions of possible positions per turn + a hand-coded evaluation from human chess knowledge.
- “Search everything, evaluate with human knowledge”

1997, Deep Blue beats Kasparov

2006, AlphaGo beats Lee Sedol

- Go has an enormous branching factor, harder to brute force
- Solution: neural networks + tree search
- Improved via self-play (reinforcement learning)
- No human-designed evaluation function needed: “search with learned intuition”

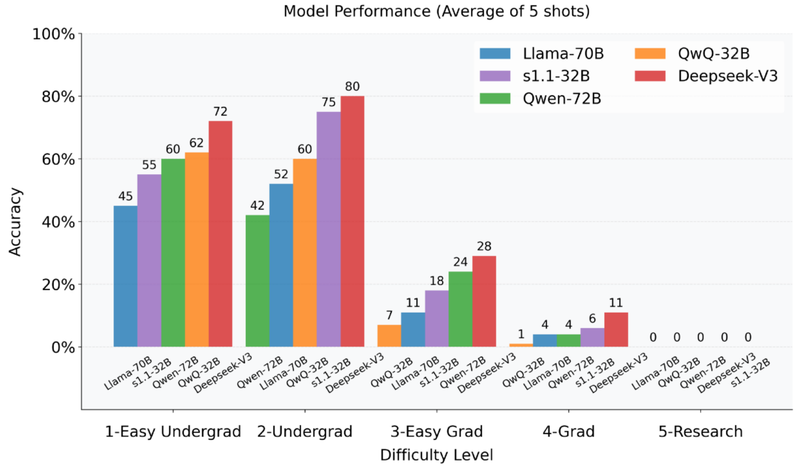
The “Bitter Lesson” of AI

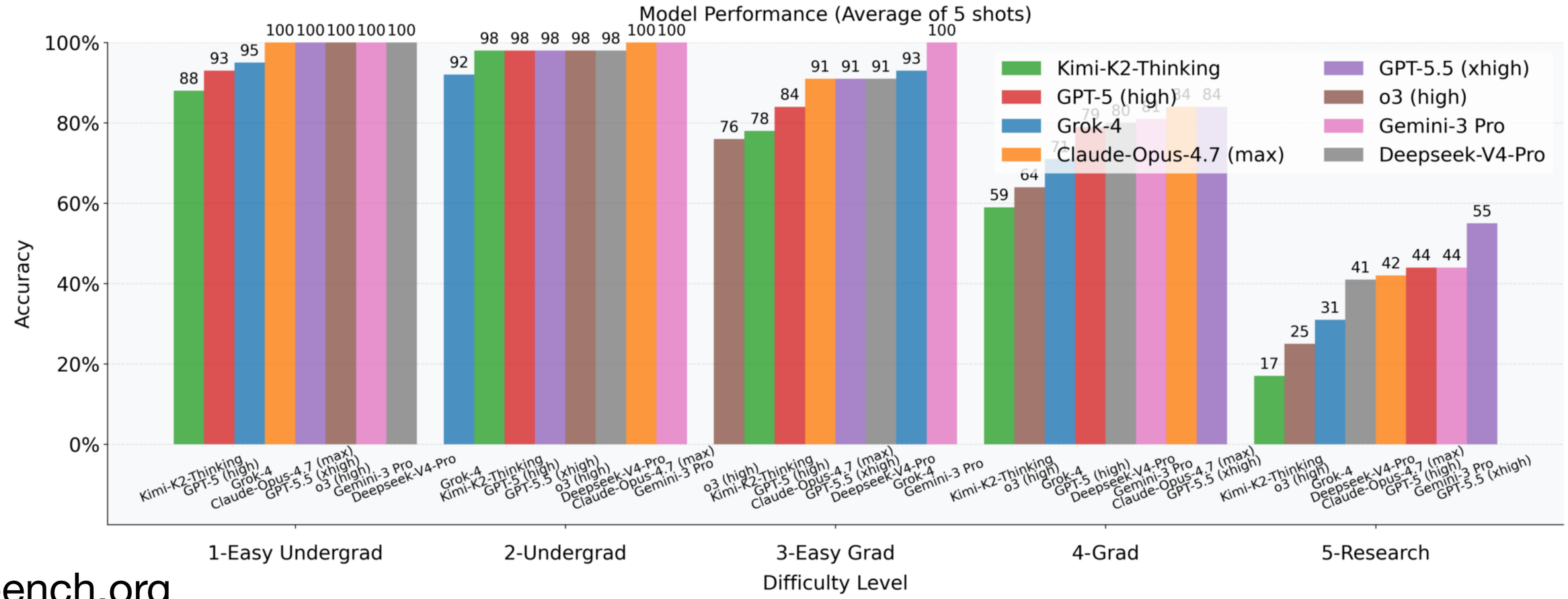
- Superhuman performance can arise from combining methods based on search, learning and scale.
- Methods that rely on built-in human knowledge may provide short term gains, but general methods tend to outperform over time.
- General methods that leverage compute have the advantage that they scale with Moore’s law (or similar).

Particle physics model building is a search problem, not a single-answer problem

- For some problems, data + consistency = unique answer
- In the case of model building, EFT logic tells us that heavy physics (scale M) affects low energy data via powers of $\left(\frac{p}{M}\right)^n$
- Data at scale p with precision Δ can only distinguish theories with M below $\sim \frac{p}{\Delta^{1/n}}$
- Many UV theories can agree with the same low-energy data. The challenge is to explore a huge space of possible theories and prune bad candidates efficiently.

2026: the year of physics utility

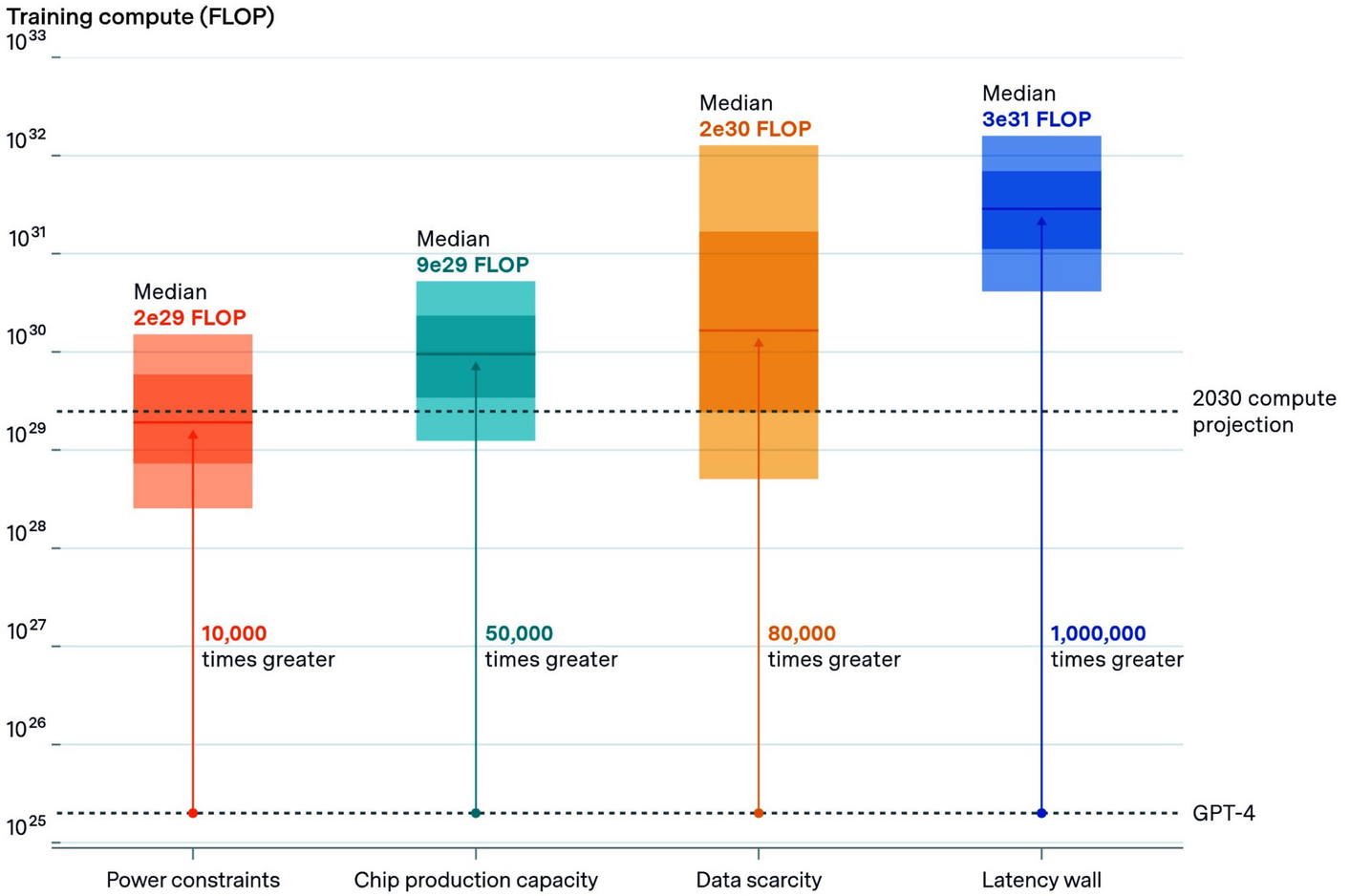




tpbench.org

Constraints to scaling training runs by 2030

EPOCH AI



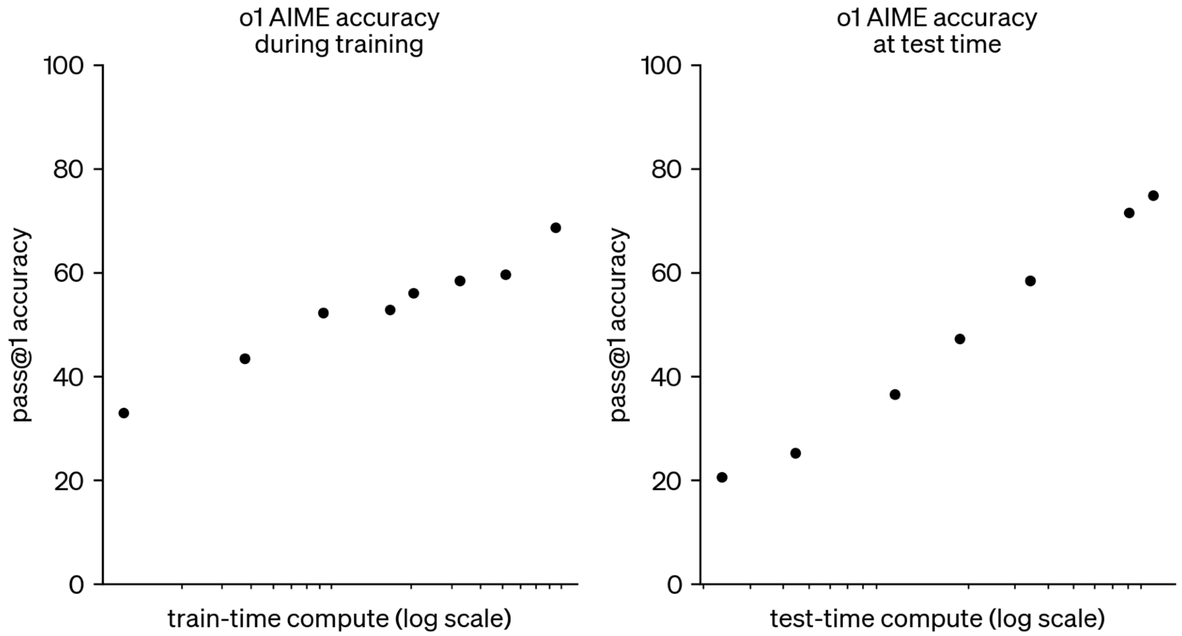
Estimates of the scale constraints imposed by the most important bottlenecks to scale. Each estimate is based on historical projections. The dark shaded box corresponds to an interquartile range and light shaded region to an 80% confidence interval.

New Axes of Model Improvement

- Scaling isn't just pretraining anymore
- Reinforcement learning
- Supervised fine tuning

} Requires internal weights + GPUs

- **Test-time compute/inference time**



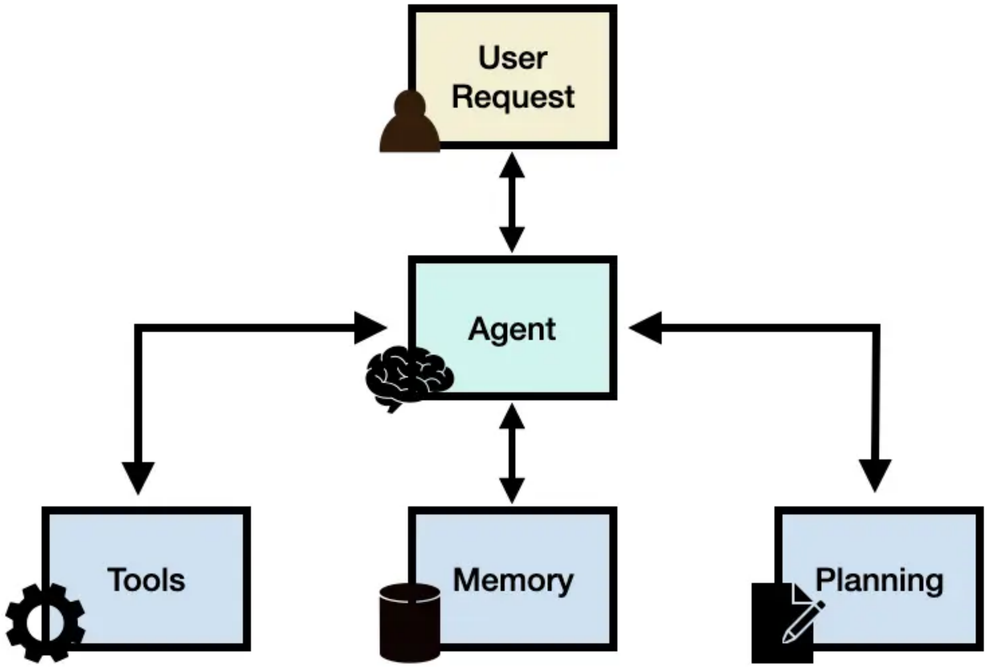
Why Raw LLMs Still Fail

- Hallucinations
- Probabilistic
- No verification / ground-truth
- No persistent memory or goals
- Struggle with layered, multi-step problems
- Alignment
- They have a “jagged edge”

Agents and Scaffolding

Agents turn chatbots into systems that can interact with their environment, test and refine ideas and manage complex workflows.

- Tools: e.g. a coding sandbox or specific software
- Schemas: structured inputs/outputs (not freeform text)
- Validators and Guardrails
- Loops, e.g. critique → refine
- Memory: track and store state/history

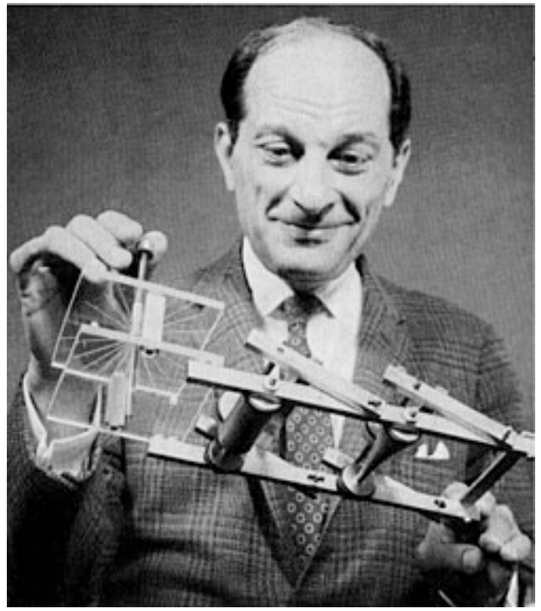


The FERMIACC

Fast **E**ngine for **R**einterpretation: a **M**achine **I**ntelligence **ACC**elerator

An automated system that proposes, simulates, and tests BSM models end-to-end

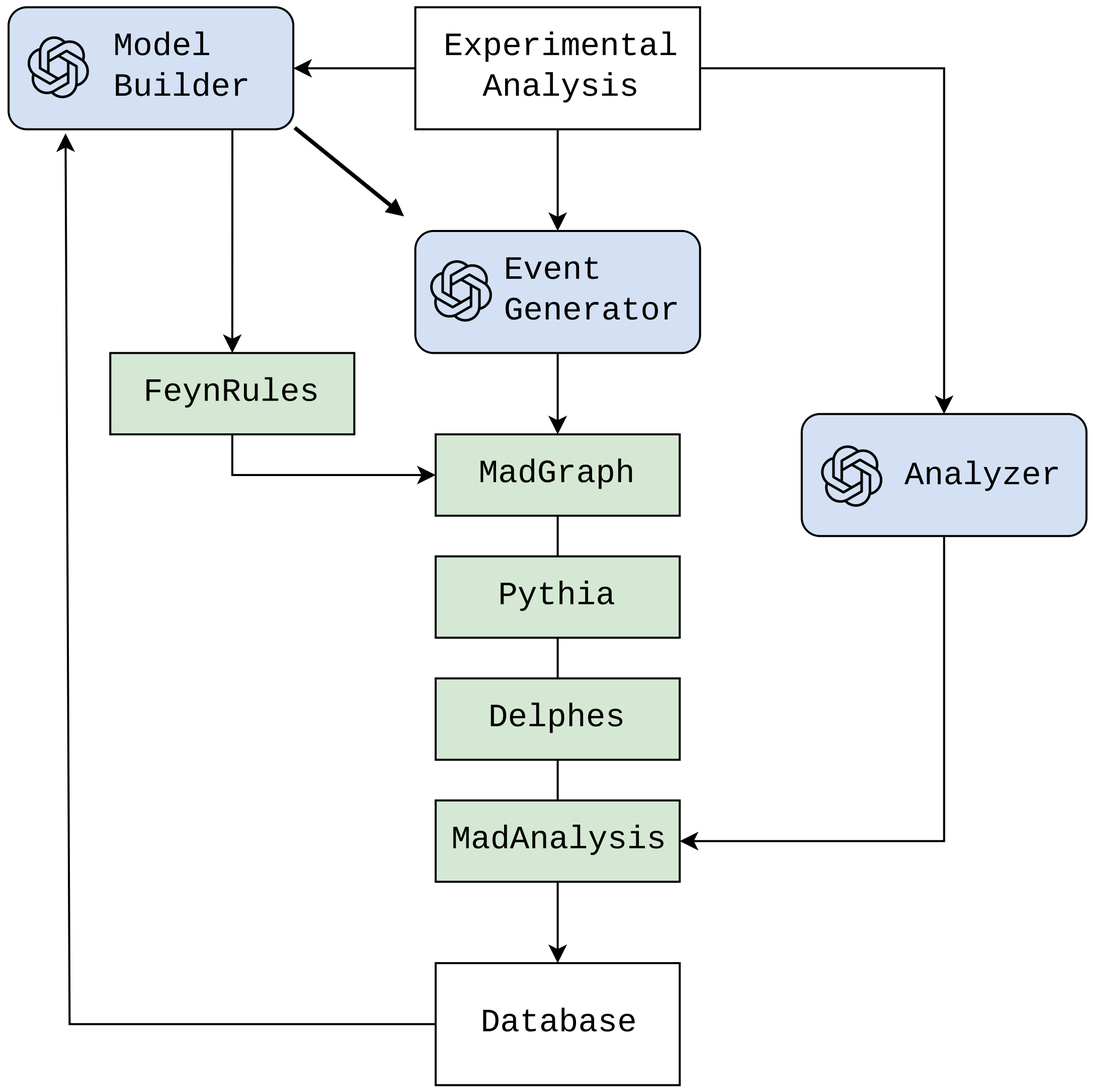
Built using the OpenAI Agents Software Development Kit (SDK) coupled to existing collider software



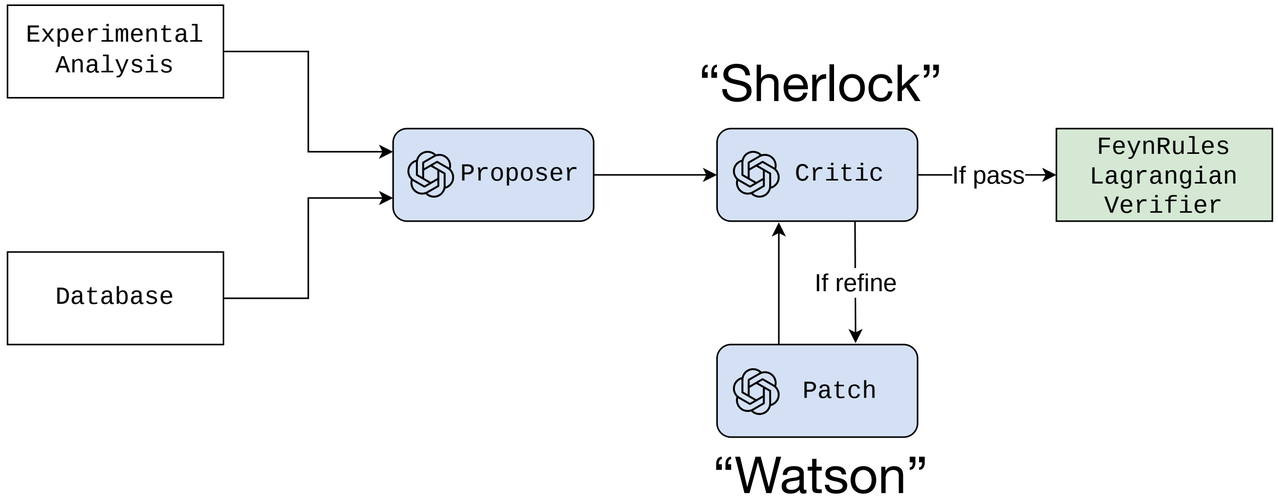
Ulam pictured holding the original FERMIAC (1947)

FERMIACC Pipeline

- Autonomous model proposal
- Deterministic verification
- Closed loop evaluation
- Prompts work across broad analyses
- Scalable exploration of model space
- Generation of high quality datasets

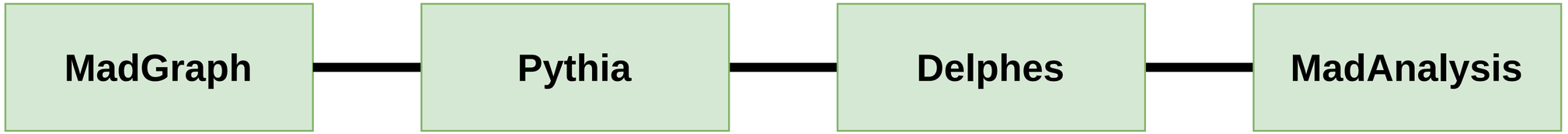


The adversarial critic loop



- Experimental signature matches analysis paper
- Novel
- Physically consistent
- Specific enough to explain experimental signature
- Correct parameter estimation

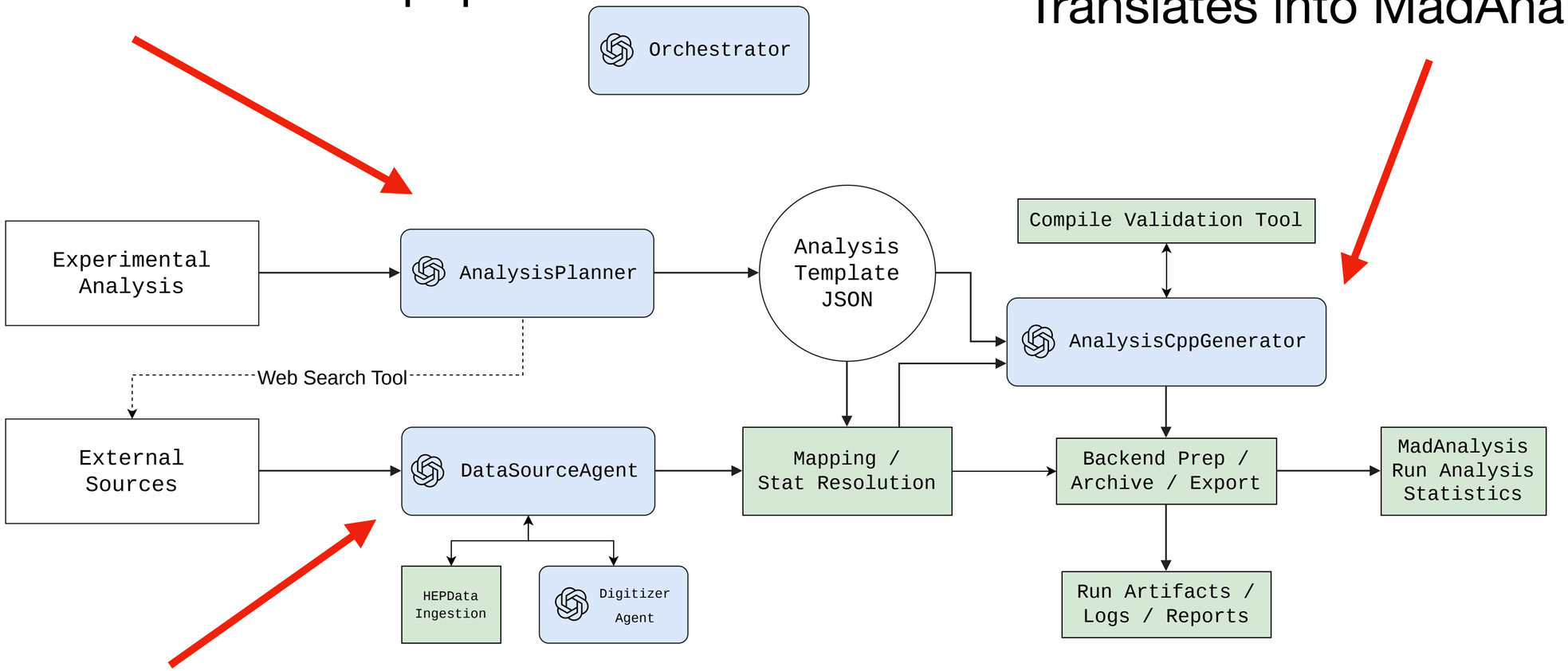
Collider Simulation Layer



- **MadGraph** – Generates the fundamental parton (quark/gluon) collision.
- **Pythia** – Simulates how partons shower and hadronize into jets.
- **Delphes** – Simulates detector response.
- **MadAnalysis** – Reproduces the experimental selection and histograms.

Analysis planner
Reconstructs objects and selections cuts from the paper

Code-generation agent
Translates into MadAnalysis 5



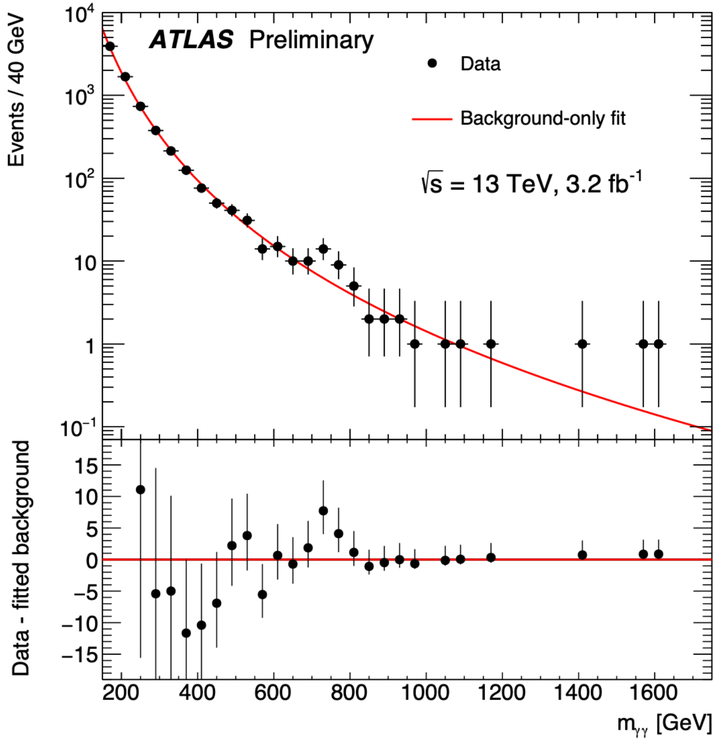
Data-source agent
Extracts yields, backgrounds, and uncertainties from HEPData, tables, or plots

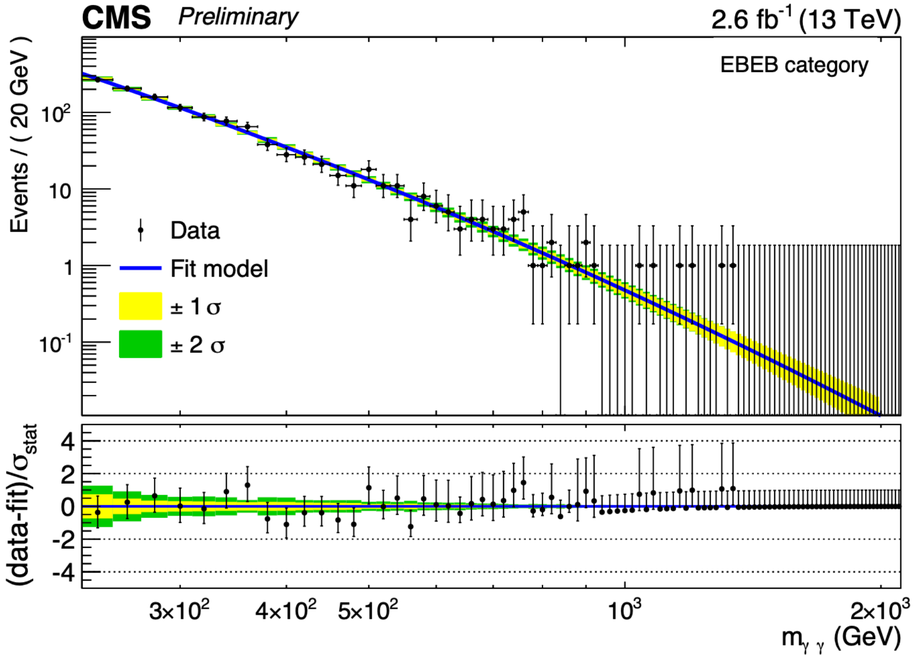
Examples of the FERMIACC in action!

Example 1: Ambulance chasing

The “750 GeV excess” (2015)

Compatible excesses (3.6 local ATLAS, 2.6 local CMS) in the $\gamma\gamma$ final state in early Run 2 data.
Same channel as the Higgs discovery.





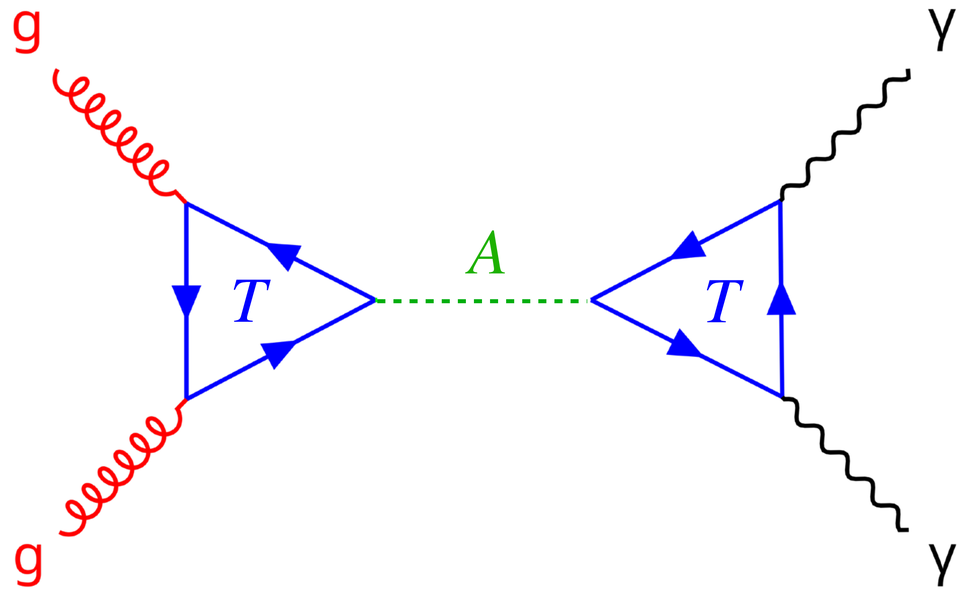
Over 500 papers written in the subsequent months

Experimental Signature: Two energetic (high pT) photons forming a 760 GeV resonance (CMS)

Proposal `544c3421` round 0, T=0.7, using GPT 5.4 and a very high number of tokens…

$A \sim (1,1,0),\ M \sim 750\,\mathrm{GeV}$, CP-odd scalar $T \sim (3,1,2/3),\ M \sim 1.2\,\mathrm{TeV}$, vectorlike fermion

$$\mathscr{L} \supset \frac{1}{2}(\partial_\mu A)(\partial^\mu A) - \frac{1}{2}m_A^2 A^2 + i\bar{T}\gamma^\mu D_\mu T - m_T \bar{T}T - iy_T A\bar{T}\gamma_5 T - y_H H\bar{T}q_{3L} + \text{h.c.}$$



```
new fields: [{
  name: "A",
  spin: "0",
  cp: "odd",
  sm reps: {"su3": "1", "su2": "1", "u1y":
  0.0},
  self conjugate: true,
  mass GeV: 750.0,
  mass scan: {"min value": 500.0, "max
  value": 2000.0, "spacing": "linear",
  "suggested points": 7}
  }, {
  name: "T",
  spin: "1/2",
  cp: "not applicable",
  sm reps: {"su3": "3", "su2": "1", "u1y":
  0.6666666667},
  self conjugate: false,
  mass GeV: 1200.0,
  mass scan: {"min value": 900.0, "max
  value": 2000.0, "spacing": "linear",
  "suggested points": 6}
  }]
```

```
parameter justifications: [{
  target kind: field mass,
  target name: A,
  oom value: 3,
  calculation steps: [ analysis inspects the
  diphoton mass spectrum well above 500 GeV,
  the paper notes the largest early-data
  fluctuation near 760 GeV for a narrow
  resonance hypothesis, choose a
  representative in-window benchmark mA =
  750 GeV, i.e. O(10^3 GeV), while keeping
  the model distinct from the paper's spin-2
  interpretation ]
  }, {
  target kind: field mass,
  target name: T,
  oom value: 3,
  calculation steps: [ to keep A -> T Tbar
  closed, require mT > mA/2 = 375 GeV, a
  heavier colored fermion also justifies
  matching to dimension-5 Agg and ABB
  operators, choose mT = 1.2e3 GeV so the
  EFT picture is reasonable while yT can
  still generate useful effective couplings
  without exceeding 4pi ] },
```

Critic Scorecard: The model fixes itself

Check	Result	Reason
LHC analysis grounding	**Pass**	
Novelty (vs paper and database)	**Pass**	
Physical consistency	**Refine**	One term not gauge invariant
UFO pipeline fit	**Refine**	Loop effects not directly implementable
Rates / parameter estimates	**Refine**	Need explicit EFT coupling estimates

$$y_H H T q_{3L} \quad \rightarrow \quad y_H \bar{q}_{3L} \tilde{H} T_R + \text{h.c.}$$

Implementable Benchmark

The agent must also decide how to implement the model in feynrules/madgraph

- No new gauge fields
- No new vevs beyond the SM Higgs vev
- No explicit broken-phase (post-EWSB) definitions
- No custom mixing matrices or mass diagonalization
- BSM benchmark processes at tree level only

Pass on round 10: UV Completion + EFT Benchmark

$$\mathscr{L}_{\text{UV}} \supset \frac{1}{2}(\partial_\mu A)(\partial^\mu A) - \frac{1}{2}m_A^2 A^2 + i\bar{T}\gamma^\mu D_\mu T - m_T \bar{T}T - iy_T A\bar{T}\gamma_5 T - y_H \bar{q}_{3L}\tilde{H}T_R + \text{h.c.}$$

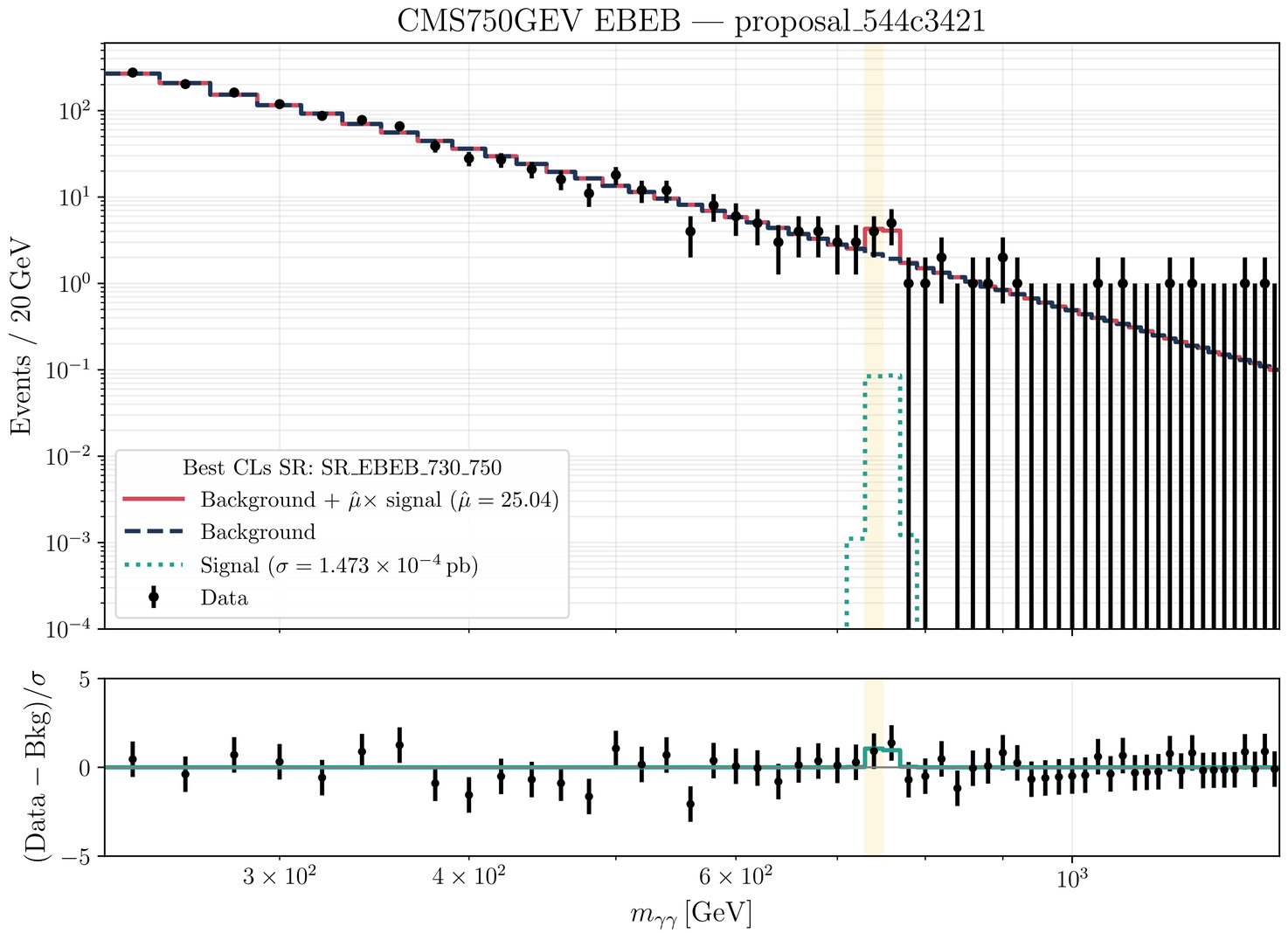
Matching (integrating out T):

$$c_{AG} \sim \frac{\alpha_s y_T}{8\pi m_T}, \qquad c_{AB} \sim \frac{\alpha_Y N_c Y^2 y_T}{4\pi m_T}$$

$$\mathscr{L}_{\text{eff}} \supset \frac{c_{AG}}{4} A\, G^a_{\mu\nu}\tilde{G}^{a\,\mu\nu} + \frac{c_{AB}}{4} A\, B_{\mu\nu}\tilde{B}^{\mu\nu}$$

```
new model reasoning: "The collider-facing content is intentionally not new: it
reuses the existing CP-odd singlet gg -> A -> aa EFT benchmark. The only novel layer
in this submission is the UV completion. Therefore this entry should be stored as a
UV-completion variant linked to that existing EFT benchmark, not as a distinct
standalone collider benchmark."
```

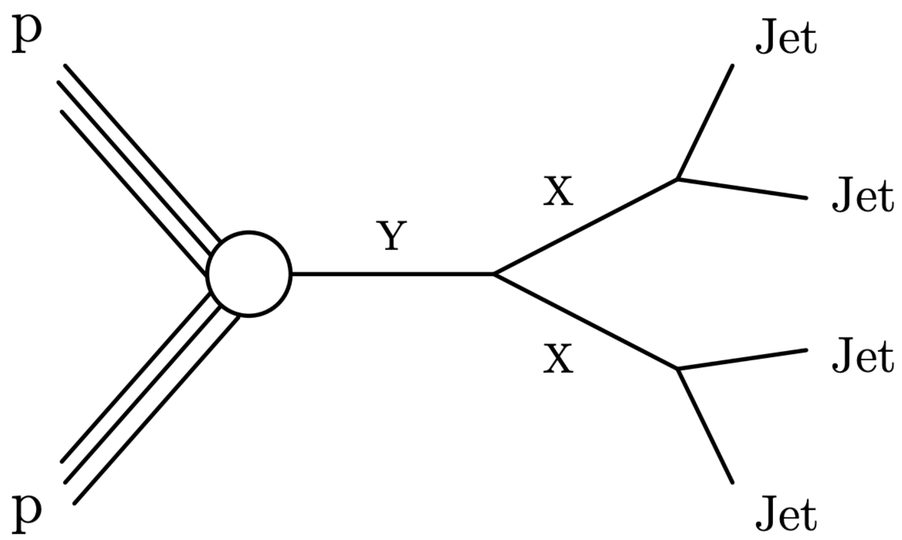
```
    {
      "target_kind": "coupling",
      "target_name": "yT",
      "oom_value": 1,
      "calculation_steps": [
        "target collider-level production needs cAG of a few x 10^-5 GeV^-1",
        "heavy-fermion matching gives cAG ~ alpha_s yT /(8 pi mT)",
        "with alpha_s ~ 0.09, mT = 1.2e3 GeV, and cAG = 3e-5 GeV^-1, solve yT ~ cAG (8 pi mT)/
alpha_s ~ 10",
        "10 is strong but still below 4pi, so use yT = 10"
      ]
    },
    {
      "target_kind": "coupling",
      "target_name": "yH",
      "oom_value": -1,
      "calculation_steps": [
        "the vectorlike quark should decay promptly through mixing with the third-generation
quark sector",
        "estimate Gamma(T -> SM) ~ yH^2 mT /(32 pi)",
        "with yH = 0.1 and mT = 1.2e3 GeV, GammaT ~ 1e-1 GeV order, implying a prompt decay
length",
        "choose a small mixing benchmark so T decays promptly without materially affecting the
diphoton resonance phenomenology"
      ]
    },
```


CMS750GEV EBEB — proposal_544c3421
Events / 20 GeV
Best CLs SR: SR_EBEB_730_750
Background + $\hat{\mu}\times$ signal ($\hat{\mu} = 25.04$)
Background
Signal ($\sigma = 1.473 \times 10^{-4}$ pb)
Data
$(\text{Data} - \text{Bkg})/\sigma$
$m_{\gamma\gamma}$ [GeV]

Example 2: CMS Paired Dijet Resonances (July 2025)

Experimental Signature Two local 3.9σ (global ~1.5σ) excesses in four-jet final states, showing resonance-like features in both the total four-jet mass m_{4j} and paired dijet masses m_{2j} at

$(m_{4j}, m_{2j}) = (8.6, 2.15)\,\mathrm{TeV}$ and $(3.6, 1.0)\,\mathrm{TeV}$



In the paper, the excesses are interpreted in terms of a color sextet scalar diquark S_{uu} or S_{dd} (charge 4/3 or –2/3) decaying to pairs of vector-like quarks, which in turn decay to quarks and gluons to give the dijets.

$$qq \;\rightarrow\; S_6 \;\rightarrow\; Q_{\mathrm{VL}}\, Q_{\mathrm{VL}} \;\rightarrow\; (qg)(qg)$$

FERMIACC proposed a model with 2 color-octet scalars for the second excess:

$$S_8 = (8,1,0), M \sim 3.6\,\text{TeV} \qquad P_8 = (8,1,0), M \sim 1\,\text{TeV}$$

Signal Topology: $gg \to S_8 \to P_8 P_8 \to (gg)(gg) \to 4j$

$$\left(\{T^a, T^b\} = \frac{1}{3}\delta^{ab}\mathbf{1} + d^{abc}T^c \right)$$

production + width

$$\mathscr{L} \supset k_{SP}\, d^{abc} S_8^a P_8^b P_8^c + c_{SGG}\, d^{abc} S_8^a G_{\mu\nu}^b G^{c\,\mu\nu} + c_{PGG}\, d^{abc} P_8^a G_{\mu\nu}^b G^{c\,\mu\nu}$$

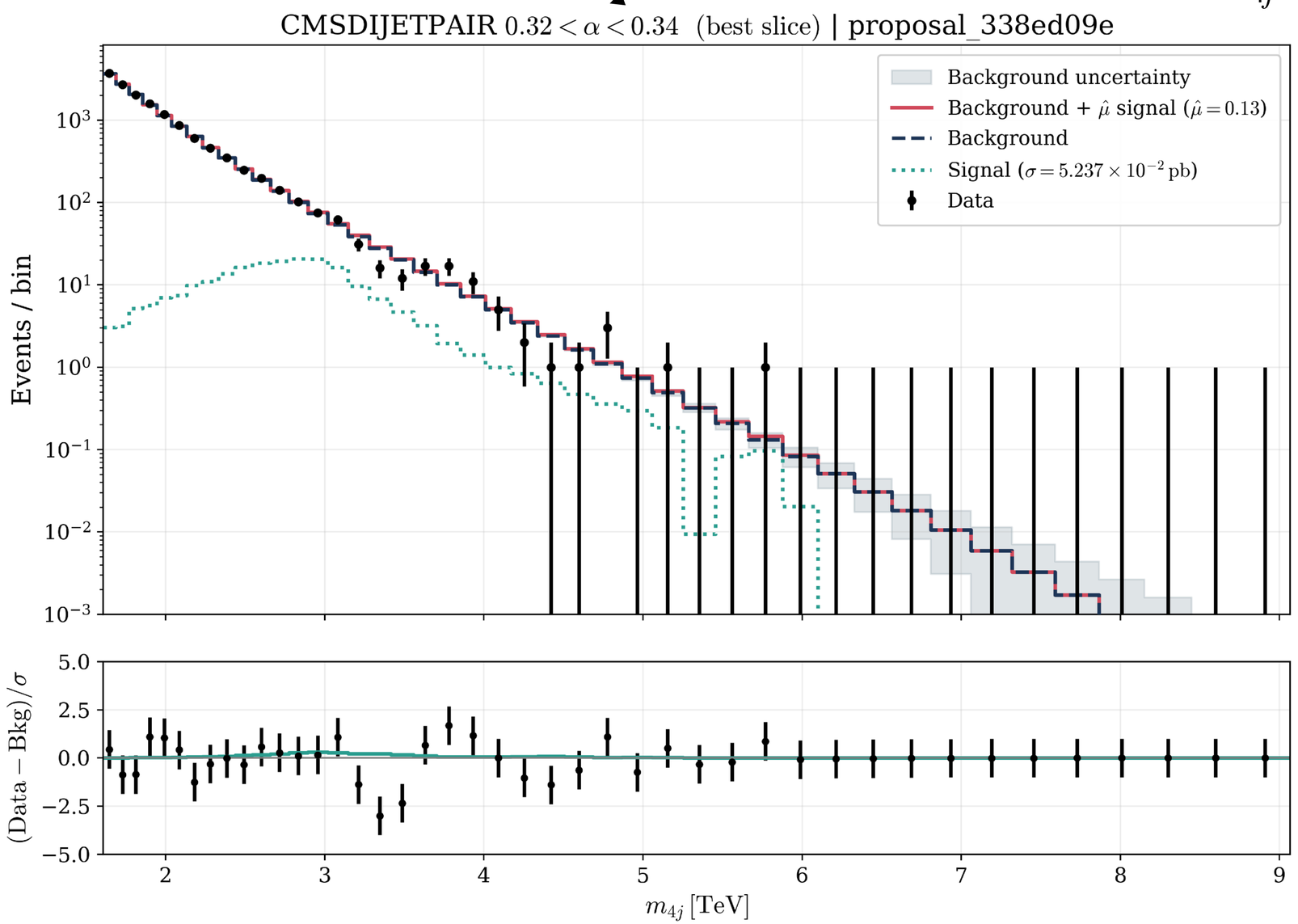
dominant width

narrow daughter width

As far as we can tell, this minimal model was *not* in the literature

* the agent also constrained subleading interactions with quarks “for completeness”.

$$\alpha = \frac{\bar{m}_{2j}}{m_{4j}}$$



The Verification Bottleneck

- Idea generation is fast
- Ideas that make it to the end of the pipeline are trustworthy
- However, many ideas fail for different reasons:
  - The LLM discovers its own idea is unphysical (hooray!)
  - Verification software does not yet support the proposal
  - A subset of good ideas fail because the LLM implements them in the pipeline incorrectly.

Next Steps

- Benchmarking
- Improved parameter optimisation
- Cross check limits against other LHC analyses
- Link with software to do UV matching verification (e.g. SARAH, SPHENO, NLOCT, FeynArts…)
- Apply to other types of data e.g. cosmological constraints
- Run at scale!

Summary

- The FERMIACC provides a general scaffolding for particle physics model building that can be used across many LHC analyses and will improve alongside underlying commercial LLMs.

- Particle physics is an ideal testbed for agents, since there is a wealth of collider papers online, as well as existing software for verification and experimental data.

- This could be extremely helpful if the LHC does not give us a 5σ bump, but instead many smaller excesses.

- The ideas here could be extended to wide ranging phenomenological fields within physics.

Database of past runs

The field content gets a discrete canonicalisation in the form:

$$\sigma(f) = (\text{spin}, \text{CP label}, \text{color rep}, \text{weak rep}, Y, \text{self-conjugacy})\,.$$

Coarse field-content search retrieves past-run summaries to guide future proposals.