Agentic AI for Automated Foundation Model Analysis in HEP

Yue Xu[1], Owen Huang[1], Yulei Zhang[1], Yuan-Tang Chou[2], Shih-Chieh Hsu[1]

[1]University of Washington

[2]National Tsing Hua University

ML4Jets 2026

14-18 Sep, Vienna

Agentic AI in high-energy physics

The landscape — a fast-moving one

Recent studies have demonstrated agent-assisted execution of substantial parts of selected HEP analyses. (e.g. [1-4])

Automated analysis pipelines

Agent-assisted code generation, execution, and iterative correction in HEP analysis pipelines

End-to-end experimental analyses

Agents have performed event selection, background estimation, uncertainty quantification, statistical inference, and documentation

Demonstrated in experimental environments

Examples include open-data analyses and Dr.Sai's BESIII branching-fraction remeasurements using experiment-specific software and computing resources.

Where that leaves us

As agentic analyses become more capable, a natural next step is to make the workflow more structured, reproducible, and reusable across analyses.

[1] Gendreau-Distler et al., arXiv:2512.07785 (2025). [2] Moreno et al., arXiv:2603.20179 (2026). [3] He et al., Dr.Sai, arXiv:2604.22541 (2026). [4] Badea et al. arXiv:2603.05735 (2026).

Agentic AI + physics foundation models

What a physics foundation model (FM) gives the agent

- **Reusable physics capabilities**
  - Pretrained representations and task heads for reconstruction, assignment, and classification.
- **Less task-specific ML design**
  - Select and combine existing capabilities instead of building a new ML solution from scratch.

What the agent gives the physics FM

- **Physics-goal translation**
  - Maps a scientific goal to the appropriate tasks, targets, and model configuration.
- **Flexible analysis orchestration**
  - Connects data preparation, model execution, validation, and downstream physics observables.

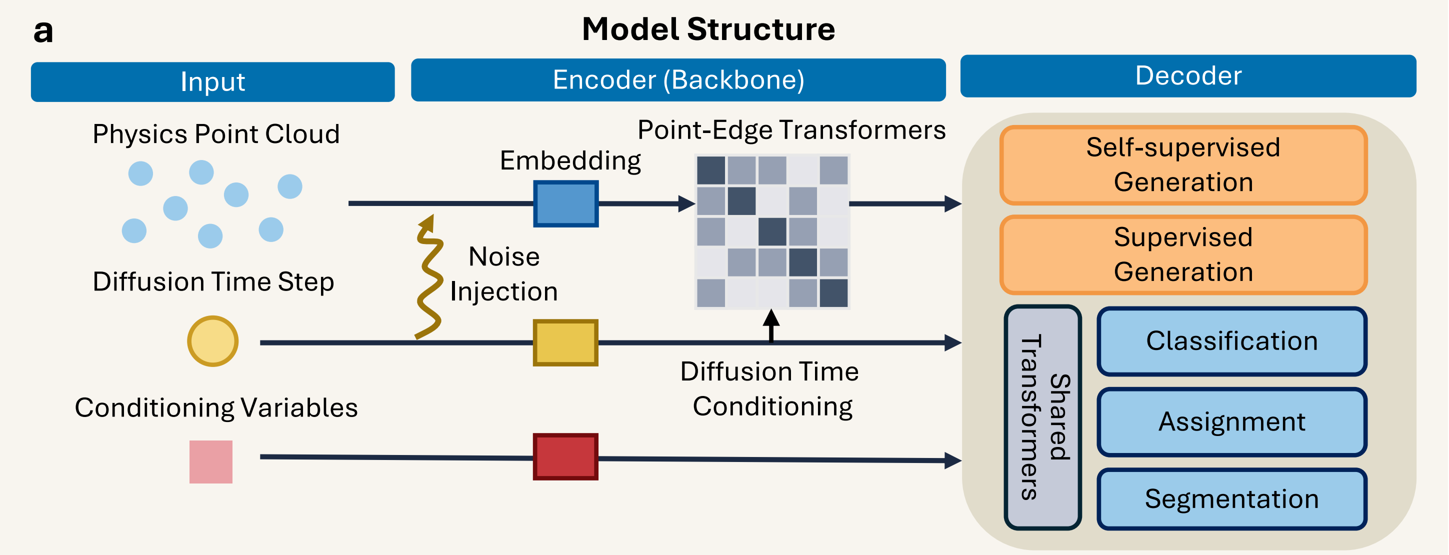
Agent flexibility + reusable physics capabilities

Toward controlled, reproducible end-to-end analyses

EveNet as a physics FM engine

- **EveNet** (arXiv:2601.17126):
  - Event-level foundation model for collider physics.
  - Point-edge transformer backbone, pretrained on 500M simulated SM events.
  - Reconstructed objects represented as a point cloud, with kinematic and identification features.
- It provides a shared pretrained backbone with multiple task-specific heads.

Head	Provides
Classification	process / category label
Assignment	object → parent-resonance pairing
Segmentation	grouping objects into final-state categories
Supervised generation	invisible-particle four-momenta, e.g. neutrinos
Self-supervised generation	masked reconstruction, background generation



Our goal: Develop a workflow for automated end-to-end physics analyses with foundation models, demonstrated through EveNet

See Ting-Hsiang's talk for EveNet details

Agent Workflow architecture

Physics prompt

“measure spin correlations in dileptonic ttbar …”

+ input data

Physics-planner

Inspects the actual data rather than assuming its structure. Proposes head(s), slot mapping, targets, split, training settings, observable.

read-only

LLM Wiki

Persistent knowledge

Shared across subagents

Physicist approval

Sees head choice, slot mapping, cuts, checkpoint, observable definition — then approves or revises.

hard gate · plan.md written on approval

approved

data-converter

ROOT/.pt/h5 → point-cloud parquet

fine-tuner

Configuration, Train + predict job submission and monitoring

predictor

Pred.-only submission when needed; output validation

data-reconverter

predictions → requested format

result-synthesizer

computes the observable

Persistent workflow state & recovery

Tracks execution and recovery history; enables evidence-guided diagnosis, minimal repair, and bounded retries.

Agent Workflow architecture

Physics prompt

“measure spin correlations in dileptonic ttbar …”

+ input data

Physics-planner

Inspects the actual data rather than assuming its structure. Proposes head(s), slot mapping, targets, split, training settings, observable.

read-only

LLM Wiki

Persistent knowledge

Shared across subagents

Explicit scientific intent

- Plan → physicist approval

Autonomous execution

- Specialized agents + validation

Persistent execution context

- Knowledge + state + recovery

Physicist approval

Sees head choice, slot mapping, cuts, checkpoint, observable definition — then approves or revises.

hard gate · plan.md written on approval

approved

data-converter

ROOT/.pt/h5 → point-cloud parquet

fine-tuner

Configuration, Train + predict job submission and monitoring

predictor

Pred.-only submission when needed; output validation

data-reconverter

predictions → requested format

result-synthesizer

computes the observable

Persistent workflow state & recovery

Tracks execution and recovery history; enables evidence-guided diagnosis, minimal repair.

Agent Workflow architecture

Physics prompt
“measure spin correlations in dileptonic ttbar …”
+ input data

Physics-planner
Inspects the actual data rather than assuming its structure. Proposes head(s), slot mapping, targets, split, training settings, observable.
read-only

LLM Wiki
Persistent knowledge
Shared across subagents

Explicit scientific intent
- Plan → physicist approval

Autonomous execution
- Specialized agents + validation

Persistent execution context
- Knowledge + state + recovery

Physicist approval
Sees head choice, slot mapping, cuts, checkpoint, observable definition — then approves or revises.
hard gate · plan.md written on approval

Physicist-controlled scientific intent, agent-driven execution

data-converter
ROOT/.pt/h5 → point-cloud parquet

fine-tuner
Configuration, Train + predict job submission and monitoring

predictor
Pred.-only submission when needed; output validation

data-reconverter
predictions → requested format

result-synthesizer
computes the observable

Persistent workflow state & recovery
Tracks execution and recovery history; enables evidence-guided diagnosis, minimal repair, and bounded retries.

Physics planner

WHAT THE PLANNER PROPOSES

- **Head selection**
  - Multiple heads when required by the physics task.
- **Slot mapping, verified against the data**
  - Opens the files and checks: tree/keys, branch names, per-event multiplicities, mass hypotheses per index, event weights
- **Targets, data splits, checkpoint, and training configuration.**
- **Observable definition and methodology**

Inspect the data, rather than assume its structure.
The planner verifies input conventions against actual files, records evidence for inferred mappings, and flags discrepancies with the user's description for physicist review.

Short report

Concise summary for physicist review

```
# Plan: ttbar dilepton spin correlation (neutrino regression + lepton-bjet pairing)

**Status**: Approved by user 2026-09-02.

## Short report

**Project / run name**: `tt2l_qc` / `tt2l_qc_run1`

**Head(s)**: `Assignment` + `TruthGeneration` — `Assignment` learns the lepton↔b-jet pairing (which lepton pairs with
which top's b-jet); `TruthGeneration` learns the two neutrino 4-momenta (px,py,pz via pt/eta/phi, massless). No other
head is needed: this is exactly the head pair `data-converter` supports for a resonance-topology + invisible-particle
problem, and matches the wiki's own record of the equivalent published benchmark for this exact physics case (per
[[evenet-tt2l-qc-study-code-2026]]: "Only two heads are enabled for this task ... `Assignment: include: true` and
`TruthGeneration: include: true`").
```

Technical report

Evidence, mapping and implementation details

```
## Technical report

**Data inspection**: Inspected via `shifter --image=docker.io/avencast1994/evenet:1.5 python3` against 4 full 10,000-
event files (2 from `TT2L`'s `train/`+`test/` subdirs, 2 from `TT2L_new`'s `run_0`/`run_349`), plus lighter checks on
additional files to confirm schema/file-count consistency. Both directories share an identical per-file HDF5 schema:

| Key | Shape | Decoded meaning (verified empirically) |
|---|---|---|
| `event` | (10000,1,3) | `[MET, MET_phi, <unidentified per-file constant ~22.7>]` |
| `jets` | (10000,10,7) | `[pt, eta, phi, mass, btag(0/1), <col5: 0-39, unidentified>, col6: parton-flavor code
∈{0,1,2,3,4,5,21}]` |
| `els` | (10000,4,5) | `[pt, eta, phi, mass, charge]` — mass column confirmed = 0.000511 GeV (electron PDG mass) |
| `mus` | (10000,4,5) | same layout; mass confirmed = 0.10566 GeV (muon PDG mass) |
| `phs` | (10000,2,4) | `[pt, eta, phi, mass≡0]` — photons; **excluded**, not part of the dilepton final state |
| `tas` | (10000,4,6) | `[pt, eta, phi, mass, charge, col5 0-22 unidentified]` — taus; **excluded**, see reasoning
below |
```

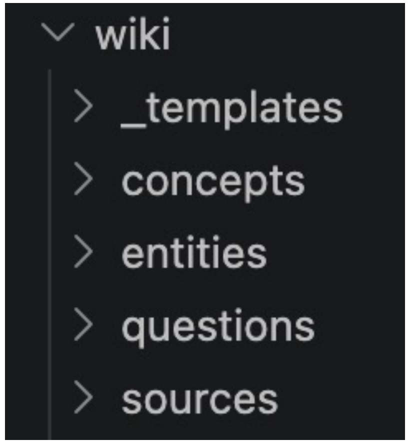
Data-grounded planning makes analysis choices explicit and reviewable.

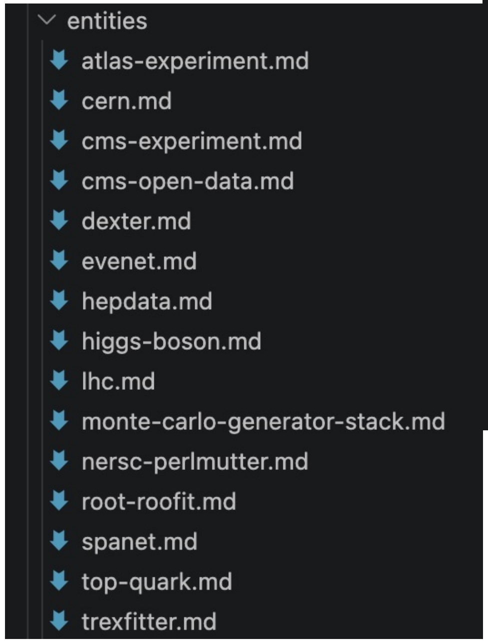
LLM Wiki — persistent scientific knowledge

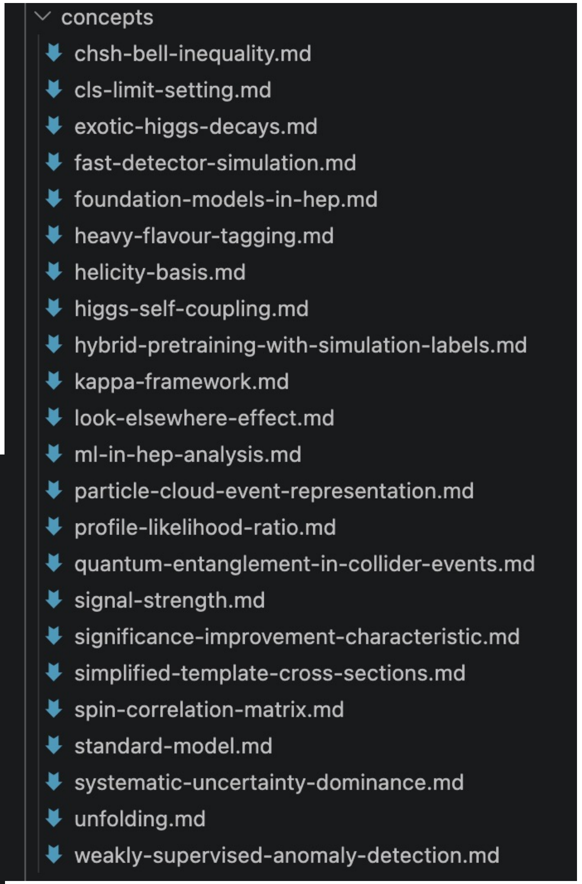
- Reusable knowledge is externalized from the agent conversation and available across workflow invocations.
- **LLM Wiki pattern**
  - A shared, LLM-maintained Markdown knowledge base.
  - Organizes reusable scientific, software/interface, and operational knowledge.
  - Accessible to task-specialized subagents.
- **Knowledge already captured**
  - Physics concepts: helicity basis, spin correlations, unfolding, significance, and Higgs analyses.
  - Entities and tools: EveNet, ROOT/RooFit, NERSC, experiments, and data resources.

Development: Expanding coverage and improving organization and quality control.

Current organization







[Andrej Karpathy, LLM Wiki (2026)]

Benchmark I — $H \to aa \to 4b$

"I want to search for the exotic Higgs boson decay $H \to aa \to 4b$, m_a=30 GeV. I want to improve the separation between signal and background, and determine the most likely pairing between reconstructed b-jets in the signal events..."

Approved analysis plan

1 Physics task & model

- **Classification**: signal vs combined QCD.
- **Assignment**: a_1/a_2 → bb pairing.

2 Data & analysis definition

- p_T-ordered jet slots; masked padding
- Assignment targets treat a1 and a2 as interchangeable
- Jet and b-jet event selections
- 2.06M selected events; fixed 50:50 holdout

3 Execution configuration

- Fine-tune EveNet-Full pretrained checkpoint
- 4 GPUs
- Event weights enabled for Classification training

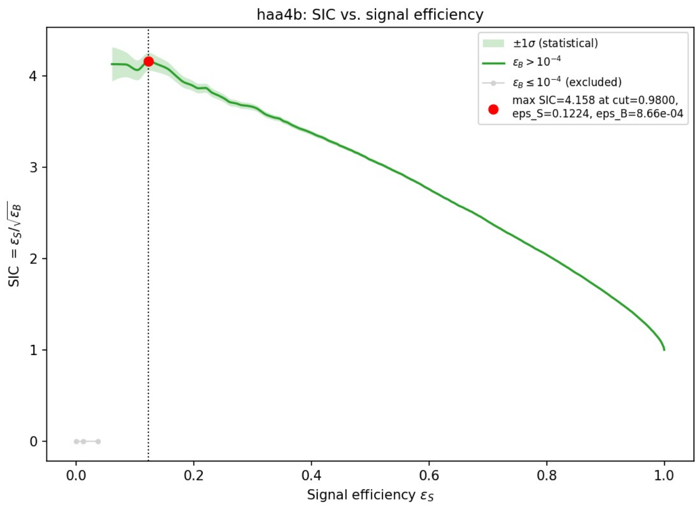
4 Physics outputs

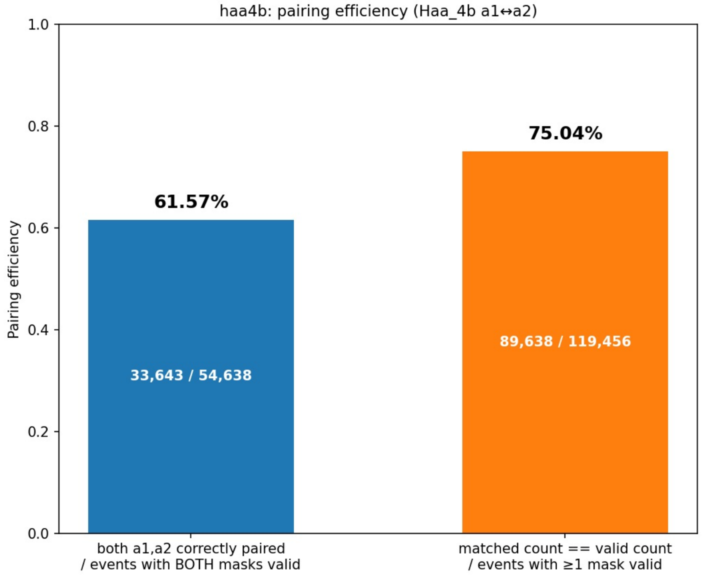
- ε_S vs. ε_B
- SIC = $\varepsilon_S/\sqrt{\varepsilon_B}$ (Significance Improvement Characteristic)
- Symmetry-aware pairing efficiency.

Specific physics request + data inspection → explicit, reviewable FM analysis plan

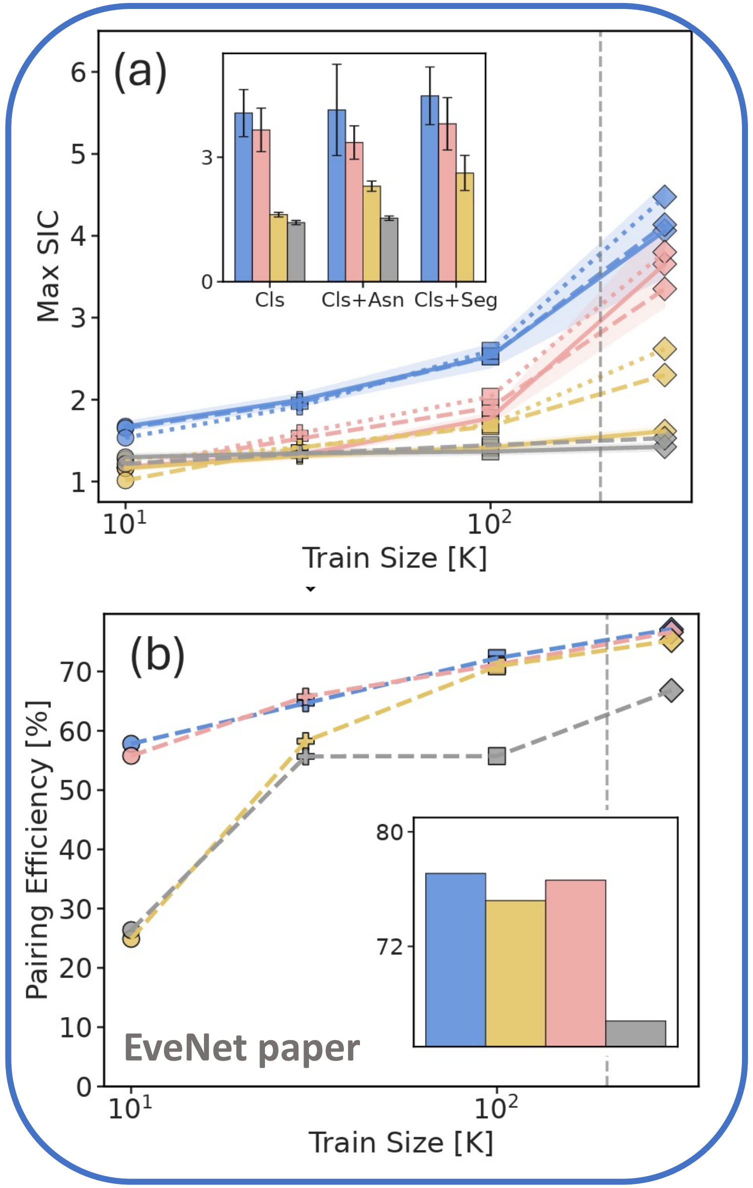
Benchmark I — H → aa → 4b

- **Max SIC** (Significance Improvement Characteristic)**: 4.2**
- **Two pairing-efficiency metrics**
  - Both resonances matched — 61.6%
  - All truth-valid resonances matched — **75.0%**
    - EveNet paper definition, explicitly provided to the agent
- **Comparable performance to the EveNet reference**





Workflow result — Claude Sonnet 5



Benchmark II — dileptonic $t\bar{t}$ quantum correlations

"I want to study $t\bar{t}$ quantum entanglement in the dilepton final state..."

Approved analysis plan

1 Physics task & model

- **Assignment**: lepton–b-jet pairing
- **TruthGeneration**: two massless neutrino momenta

2 Data & analysis definition

- 18 unified jet/lepton slots; masked padding
- Charge- and ΔR-matched assignment targets
- 3.45M train / 0.493M val / 8.2M test

3 Execution configuration

- Fine-tune EveNet-Full pretrained checkpoint
- 16 GPUs

4 Physics outputs

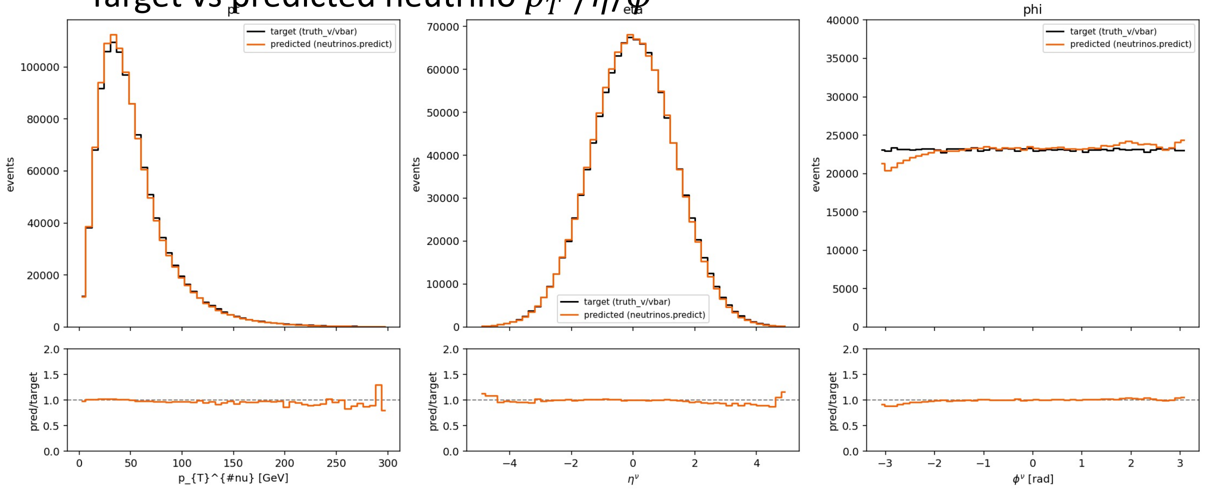
- Neutrino kinematics and pairing efficiency
- Unfolded spin-correlation matrix C_{ij}
- Entanglement observable Δ

High-level physics goal + Wiki + data inspection → composed FM capabilities and downstream analysis

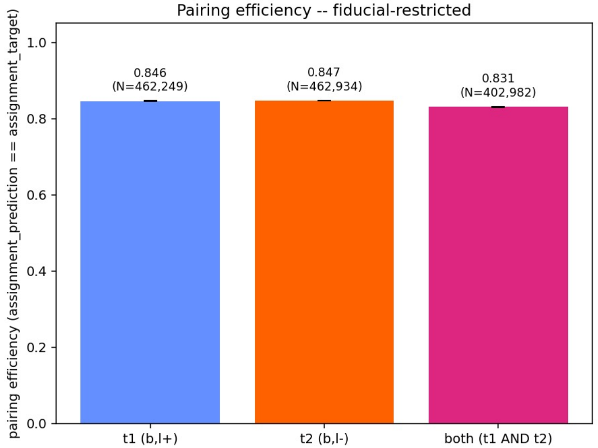
Benchmark II — dileptonic $t\bar{t}$ quantum correlations

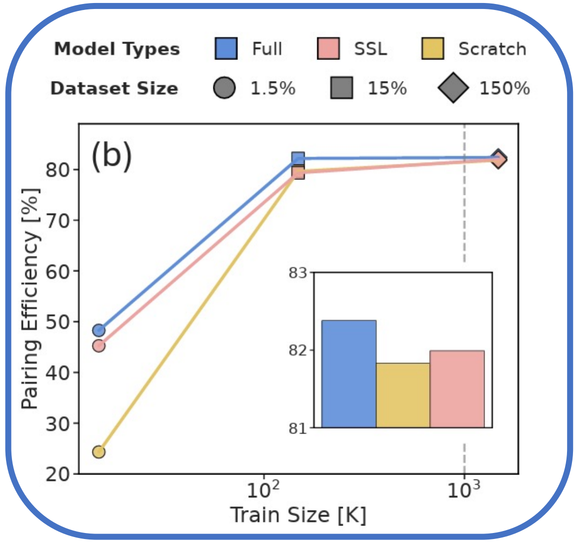
Neutrino reconstruction

Target vs predicted neutrino p_T / η/ϕ



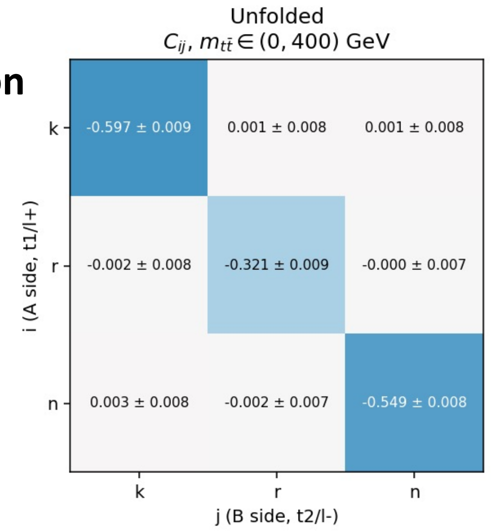
Lepton-b jet pairing





83.1% pairing efficiency comparable to EveNet

Quantum correlation analysis

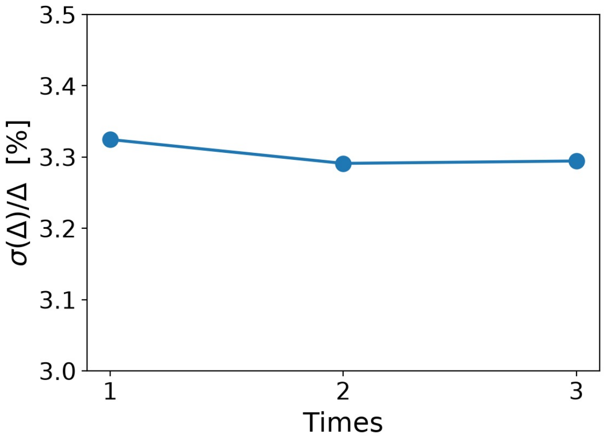


Construct Δ with unfolded C_{ij}:

- $\Delta \equiv |C_{kk} + C_{rr}| - C_{nn} - 1$
- $\frac{\sigma(\Delta)}{\Delta} = \mathbf{3.2\%}$

Comparable quantum-correlation precision to the EveNet reference.

Run-to-run reproducibility



Another three independent runs yield consistent results across Claude Sonnet/Opus 5

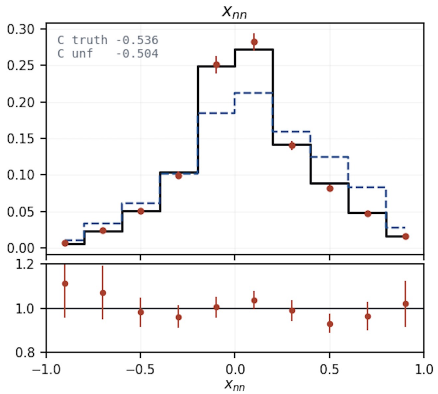
What does the structured workflow add?

Preliminary bare-agent test — dileptonic $t\bar{t}$

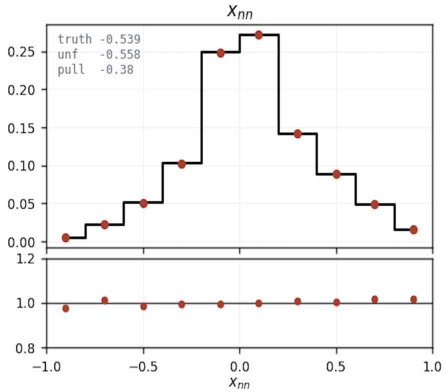
Bare Agent

Given:
- physics goal + input data
- EveNet repository + public docs
- EveNet paper + pretrained weights
- same computing environment

Not given:
- planner / LLM Wiki
- workflow-specific rules
- repair / workflow state



Human guide



Workflow: **Claude Sonnet/Opus 5** | Bare agent: **Claude Opus 5**

Same physics task and EveNet resources; no workflow-specific memory, approval or rules.

Unapproved analysis changes	2000/3280 training files; 300/820 test files; unrequested reco $m_{tt} < 400$ GeV cut.
Unapproved execution changes	Changed training from 50 → 20 epochs without approval; training stopped before the loss had converged.
Unfolding procedure / validation	Initial out-of-window fake subtraction; closure checks missed significant unfolding bias

Human-guided recovery reveals strong agent capability

Once the methodological issue was identified and the correction was provided, the bare agent successfully rebuilt the response and performed additional validation.

The workflow adds explicit intent, approval, persistent knowledge, validation, and structured recovery.

Conclusion

General-purpose agents + physics foundation models

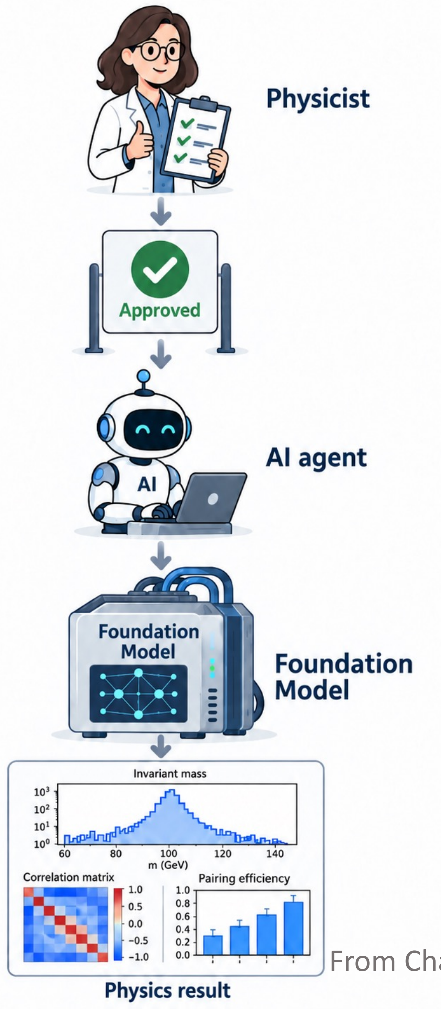
- Flexible translation from physics goals to reusable FM capabilities
- Demonstrated with two end-to-end EveNet analyses

Structured scientific execution

- Plan → Approve → Execute
- Explicit decisions, persistent knowledge, validation, and recovery

Evidence

- Comparable EveNet performance in both benchmarks
- Run-to-run reproducibility demonstrated for $t\bar{t}$
- Bare-agent test motivates explicit workflow control



From ChatGPT

Connecting general-purpose agents with domain-specific foundation models: toward reusable and reproducible physics analysis workflows.

Thanks!

BACKUP

Approved analysis plan — H → aa → 4b

```
# Plan: H→aa→4b exotic Higgs discrimination + pairing (m_a = 30 GeV) — revised, approved

**Approval record**: User approved this plan as presented on 2026-09-10. Original open questions (split mode,
event-column usage, duplicate-mask scope) were answered by the user; the revised plan below incorporates
those answers plus one additional judgment call (intra-resonance distinctness rule) that was explicitly
flagged to and approved by the user.

## Short report

**Project / run name**: `haa4b` / `haa4b_cls_asn_run1`

**Head(s)**: `Classification` — signal-vs-QCD discrimination (physics ask objective 1); `Assignment` — b-
jet↔a-boson pairing (objective 2). Trained together in one fine-tune, matching EveNet's own published
Benchmark 2 methodology for this exact channel (Cls+Asn jointly on H_SM→aa→4b) (per [[evenet-foundation-
model-2026]], [[exotic-higgs-decays]]).

**Data**:
`[redacted]
4b}_*.h5`, format `h5`. The two directory trees are used **as-is** as the train and test pools respectively
(see Split mode) — not pooled together.

**Slot mapping**: 18 slots: jets first (up to 10, pT-descending, from `jets`), then leptons (electrons+muons
merged, pT-descending, from `els`/`mus`, up to 8) — jets-first packing per the h5 default convention.
Photons/taus excluded (not part of EveNet's own point-cloud object set, per [[particle-cloud-event-
representation]]). No truncation ever observed (10+8=18 exactly fits 18 slots); if it ever occurs, leptons
truncate first (jets define the selection).

**Selection cut**: ≥4 jets (`jets.pt`>30 GeV, any b-tag) AND ≥3 b-tagged jets (`jets.btag`==1, `jets.pt`>20
GeV) AND 2nd-highest-pT b-tagged jet `pt`>30 GeV — applied identically to **train, val, and test** (per
explicit user instruction).

**Split mode**: `holdout` — confirmed by the user. `train/Combined_Balanced/` (all three glob patterns,
61,965,329 raw matching events before any cut) is the complete source for `train.npz`, with EveNet's internal
`Dataset.val_split` carving validation from it as usual; `test/Combined_Balanced/` (61,890,730 raw matching
events) is the complete source for `test.npz`. No `val.npz` is produced (standard for `holdout`); no pooling
or reshuffling of the two directories against each other.

**Training**: checkpoint `checkpoints.20M.a4.last.ckpt` (EveNet-Full); wall time `08:00:00` (increased from
the `04:00:00` default — two simultaneously-trained heads plus an estimated ~880K-event train pool); GPUs `4`
(single-node); batch size `1024` train / `2048` predict.
```

Approved analysis plan — dileptonic $t\bar{t}$

Plan: ttbar dilepton spin correlation (neutrino regression + lepton-bjet pairing)

Status: Approved by user 2026-09-02.

Short report

Project / run name: `tt2l_qc` / `tt2l_qc_run1`

Head(s): `Assignment` + `TruthGeneration` — `Assignment` learns the lepton↔b-jet pairing (which lepton pairs with which top's b-jet); `TruthGeneration` learns the two neutrino 4-momenta (px,py,pz via pt/eta/phi, massless). No other head is needed: this is exactly the head pair `data-converter` supports for a resonance-topology + invisible-particle problem, and matches the wiki's own record of the equivalent published benchmark for this exact physics case (per [[evenet-tt2l-qc-study-code-2026]]: "Only two heads are enabled for this task ... `Assignment: include: true` and `TruthGeneration: include: true`").

Data: two H5 directories, format `h5`:
- `[redacted]` — **6,000,000 events** (600 files × 10,000), laid out as pre-existing `train/`(6 groups)/`test/`(4 groups) subdirectories
- `[redacted]` — **35,000,000 events** (350 `run_N/` dirs × 10 files × 10,000), no train/test split
- **This contradicts what was stated in the original request** ("TT2L ~35M, TT2L_new ~6M") — measured directly from `attrs['total_entries']` and file counts, the sizes are reversed: **TT2L is the ~6M-event set, TT2L_new is the ~35M-event set.** Combined pool = **41,000,000 events**. The two directories share an identical per-file schema (verified across 4 files spanning both directories, both original train/test subdirs, and early/late run numbers), so they combine cleanly — see "Split mode" below for exactly how.

Slot mapping: 18 slots — one unified pool, not fixed per-type index ranges. Jets are packed first in priority (filling contiguously starting at slot 0, only as many slots as that event has real jets, up to 10, pT-descending), then leptons (electrons ∪ muons merged and re-sorted by combined pT) pack immediately after, contiguous — not at a fixed slot 10 boundary. Downstream heads distinguish jet vs. lepton via the per-slot `isLepton` flag, not slot position. Photons and taus excluded (not part of the dilepton final state; taus specifically excluded because the charged-lepton spin-analysing power α=±1 this whole measurement relies on is a prompt-e/μ fact, per [[top-quark]]).

Selection cut (two different cuts, exactly as specified, applied at the *raw-event-bucket* level after the 70:10:20 split — see Split mode):
- Train/val: ≥1 opposite-charge lepton pair (from the merged e/μ collection) with the subleading lepton's pT > 10 GeV, AND ≥2 b-tagged jets with pT > 20 GeV. **Measured on the full 41,000,000-event pool (not a sample): 12.02% pass rate** (3,450,621/28,700,000 train-bucket, 492,875/4,100,000 val-bucket).
- Test: truth-level ttbar → bW(ℓν) bW(ℓν) decay chain present. **Measured on the full 8,200,000-event test bucket: 100.00% pass rate** (8,200,000/8,200,000) — this h5 production is truth-level-dileptonic-exclusive, so this cut is a no-op filter for our own data (real for a general TT2L file, just not restrictive here).

Split mode: `standard` — **finalized 70% train : 10% val : 20% test** of the full combined raw pool, confirmed by the user. Because train/val and test use two different, non-complementary cuts, the raw pool is partitioned into buckets *before* either cut is applied, to prevent the same event landing in two splits (see Technical report for full mechanics). **Real post-cut counts: train 3,450,621 / val 492,875 / test 8,200,000.**

Training: checkpoint `checkpoints.20M.a4.last.ckpt` (EveNet-Full); wall time `12:00:00` (**set directly by the user**, overriding the calculated core estimate of ≈1.5–2.5h / proposed `04:00:00` — see Technical report for the original derivation, kept for reference); GPUs `16` (multi-node: 4 nodes × 4 GPUs/node); batch size `1024` train / `4096` predict

Output: `<run_dir>/output/predicted.pt` (h5 input → `.pt` output, per the fixed format-asymmetry rule) — `extra_save`: **8 fields** (see Technical report)

Downstream observable: `D = |C_kk + C_rr| - C_nn` (Peres entanglement marker), via the decay-angle method in the helicity basis, unfolded from a **reco-level fiducial region** (exactly 2 b-jets pT>25 GeV/|η|<2.5 + 2 oppositely-charged leptons pT>25 GeV/|η|<2.5, evaluated on the `Assignment` head's selected b-jet/lepton slots) to a **truth-level region** (`0 < m_tt_truth < 400 GeV`, no kinematic cuts) — two distinct, explicitly-named regions feeding RooUnfold's `Fill`/`Miss`. Target-vs-predicted neutrino kinematics (pt/eta/phi + ratio panel) and pairing-efficiency plots are each produced **twice**: inclusively (full test set) and restricted to the same fiducial region. Two headline numbers, each with its own uncertainty method: **`D ± D_err`** (quadrature sum of the three unfolded diagonal `C_kk`/`C_rr`/`C_nn` uncertainties) and **`Concurrence ± Concurrence_uncertainty`** (Wootters concurrence from the full unfolded 15-parameter density matrix, uncertainty from a coordinated ±1σ up/down shift across all 15 parameters) — see Technical report.