



中国科学院高能物理研究所
Institute of High Energy Physics
Chinese Academy of Sciences

AI for HEP

with a focus on AI at BESIII

李科 (like@ihep.ac.cn)

Institute of High Energy Physics

AI at lepton colliders

More than 10B events

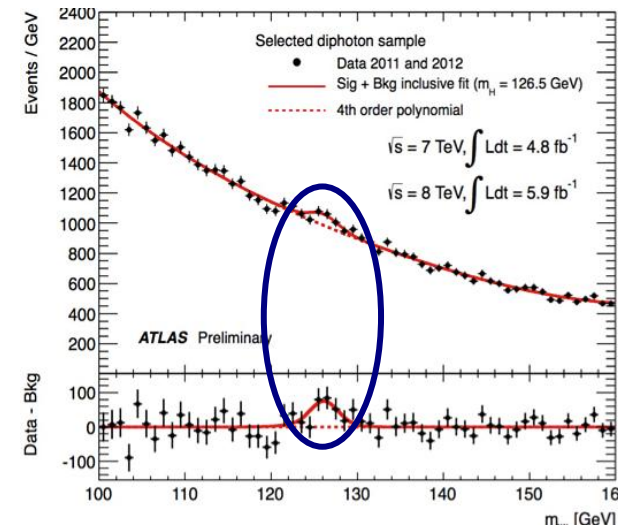


Significance

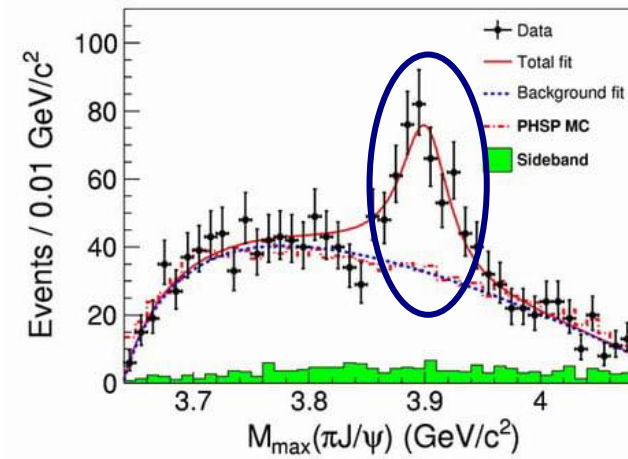
$$Z \approx \frac{S}{\sqrt{B}}$$

AI4S 3 main approaches:

- **Improve S**
- Reduce B (more important at LHC)
- **Accelerate analysis**



Higgs at ATLAS



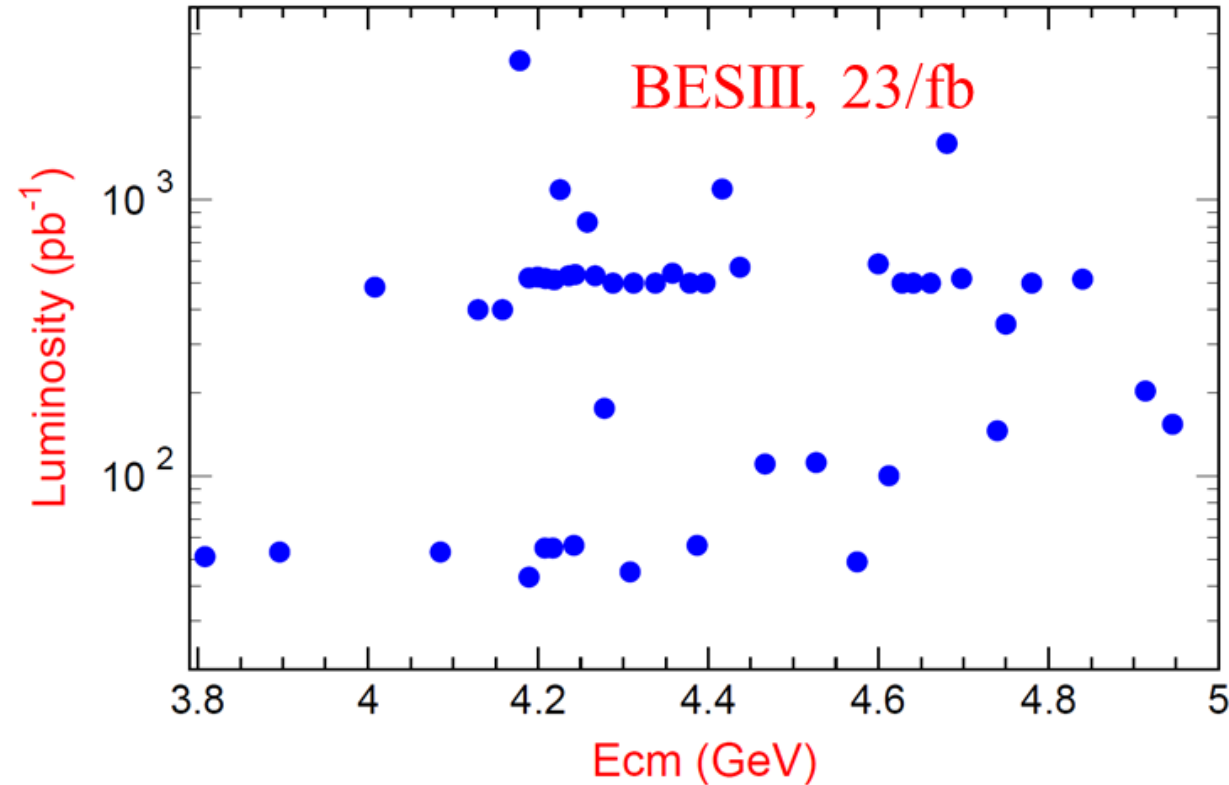
Zc(3900) at BESIII

Outline

- New ways to search for new particles
 - Accelerate physics analysis: DrSai
 - The systematic search by automating the full analysis
 - Improve S: Foundation model and resonance tagger
 - The unified event representation model to cover all known physics channels
- Summary and outlook

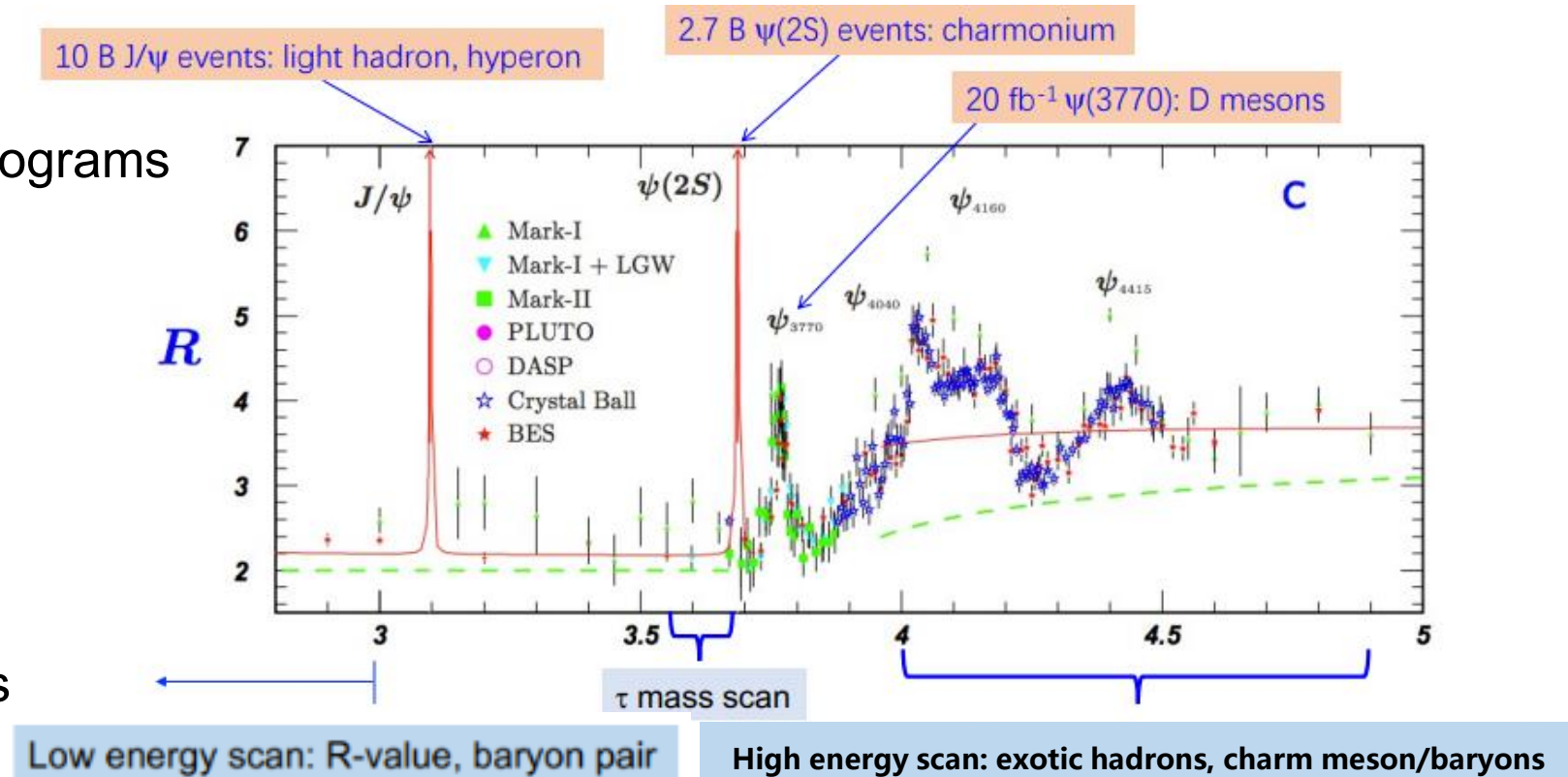
Personal Notes

- Many NP theories predict new particle/phenomena
- Typically, most searches are based on a specific production or decay process
 - Limited by resources
- Maybe two new approaches at lepton collider?
 - Systematic search via automated tools of all possible processes
 - Dr.Sai
 - Dataset level anomaly detection
 - Assuming the signal is not universal across all samples



Introduction of BESIII - physics program

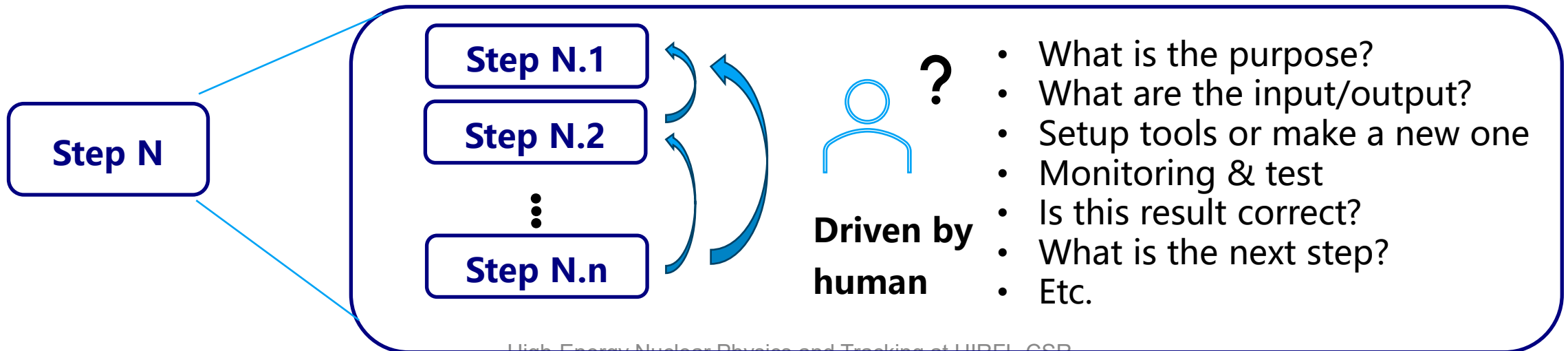
- >700 scientists and engineers
- Tau-charm factory, rich physics programs
 - Light hadrons
 - Charm meson/baryons
 - Charmonium
 - Precise test of SM
 - Search for new physics
- Hundreds of physics results
 - Discovered >30 new hadrons
 - First tetraquark: $Z_c(3900)$



- **Clear and relative simpler workflow for effective modelling**
- **Different from LHC, unique challenges and advantages in lepton collider experiments**

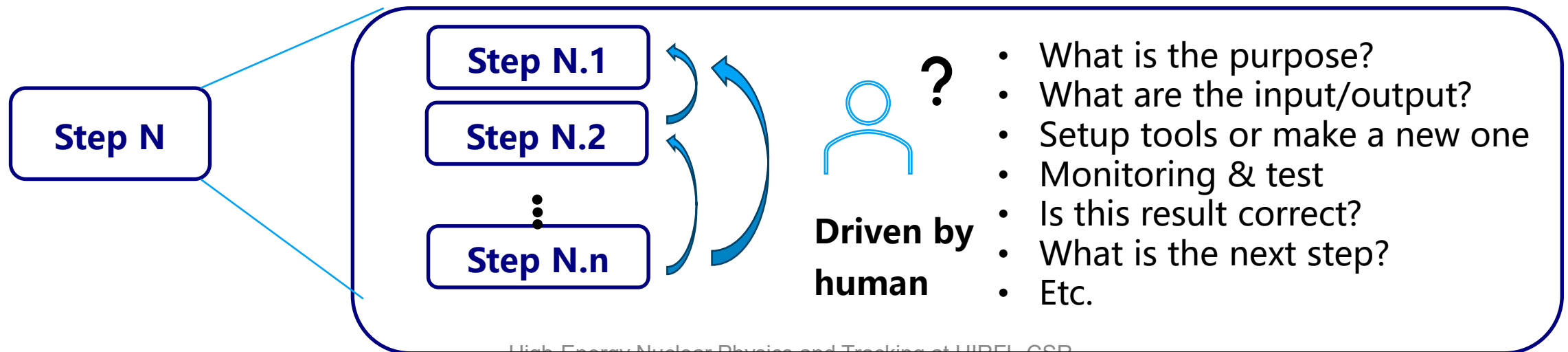
Agentic AI for HEP

Traditional approach



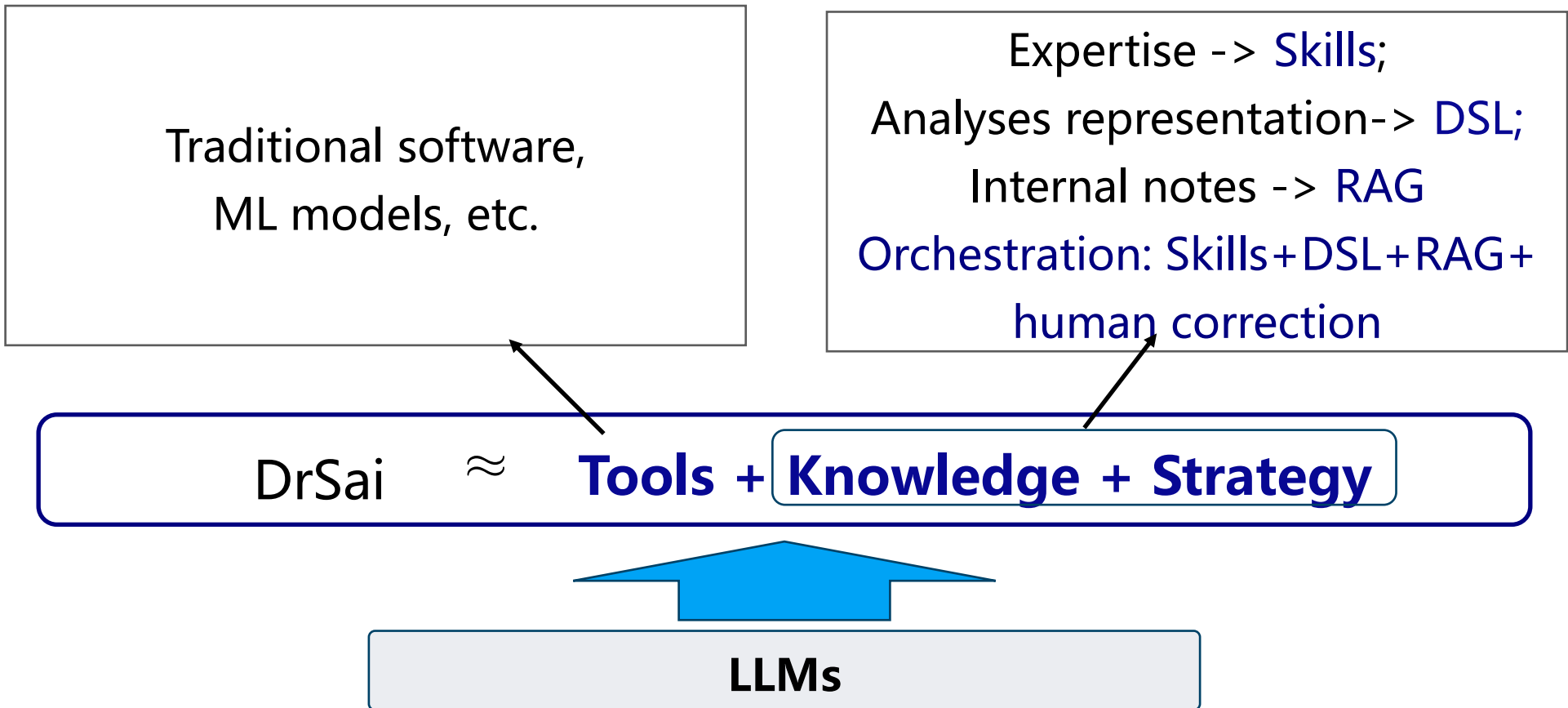
Agentic AI for HEP

- In traditional approaches, human involvement is required at all times. Expertise is often unstructured, though some can be represented by natural language.
- Today, we can leverage LLMs to build an AI partner for scientists.
 - **The AI partner designs and conducts the entire workflow based on human commands and feedback.**
- In the future, with a well-defined representation of knowledge, we can further develop an AI scientist.
 - **Humans would only need to propose a goal — and leave everything else to the AI.**

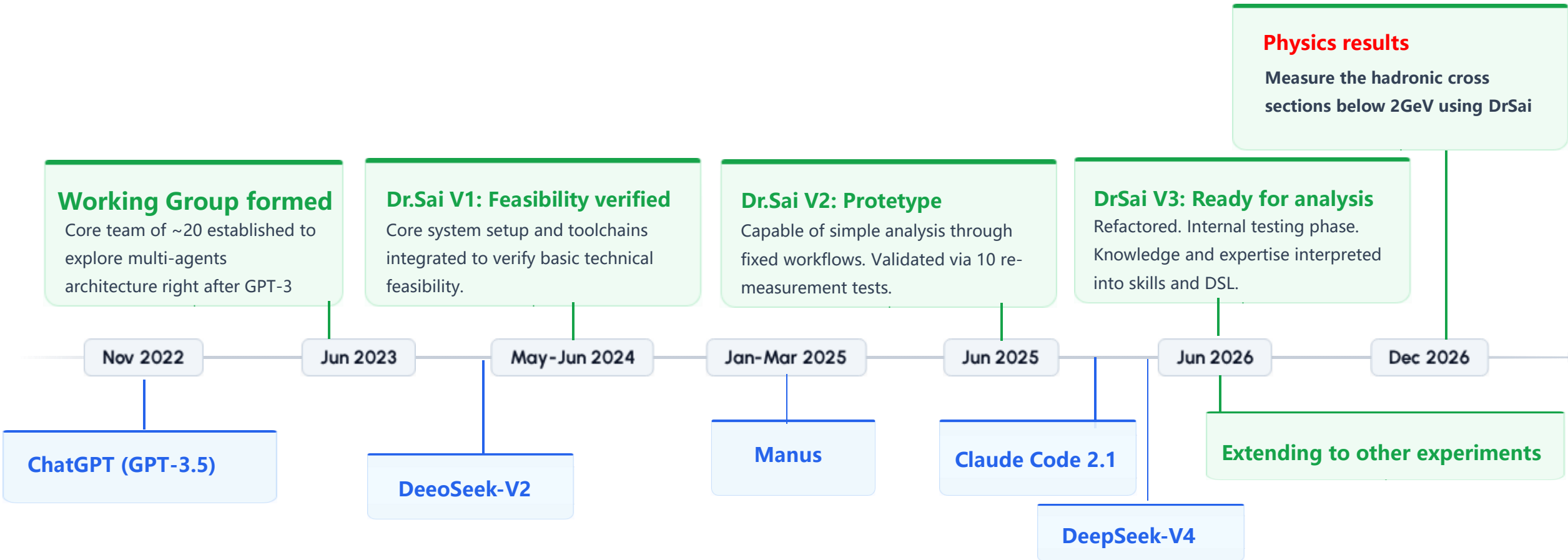


The DrSai project

Purpose: Pioneering autonomous physics discovery at **BESIII (and beyond)** through Agentic AI

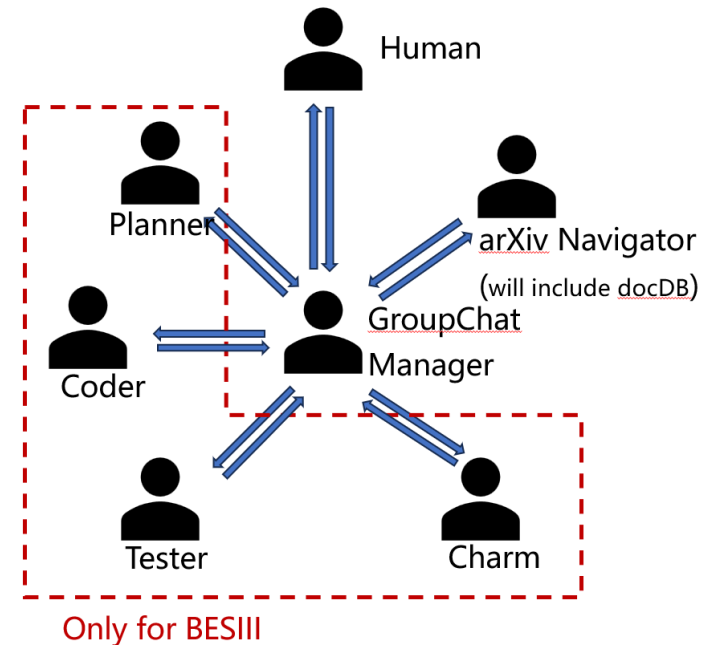
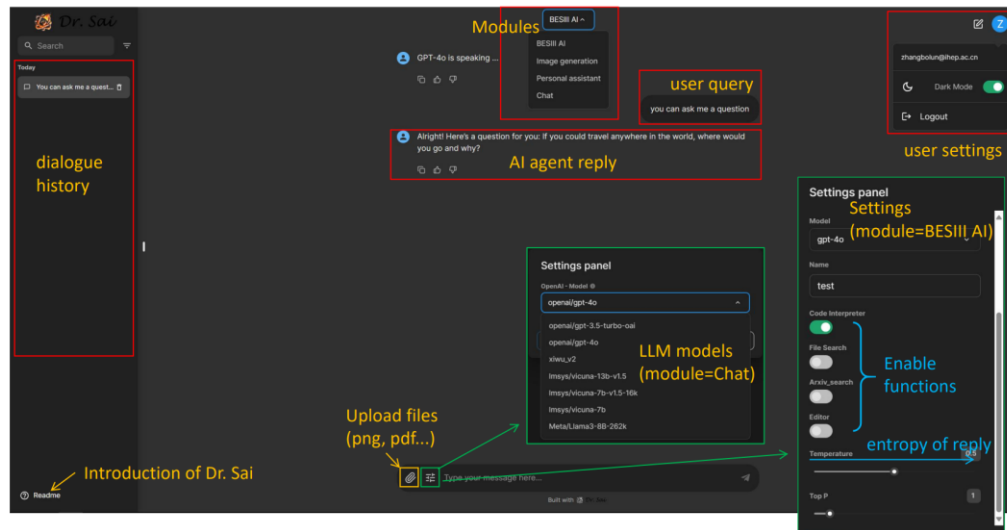


DrSai Evolution Milestones



DrSai V1

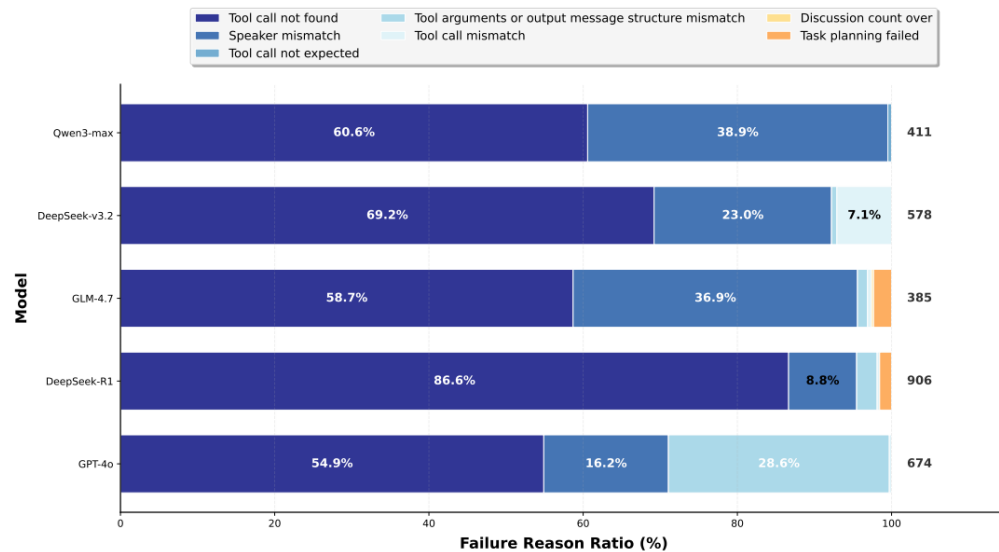
- Developed based on open-source multi-agents framework: AutoGen.
- Deploy workers to wrap-up the tools
- Tested LLMs on intent understanding, orchestration, tool invocation, etc.
- **Verified feasibility**



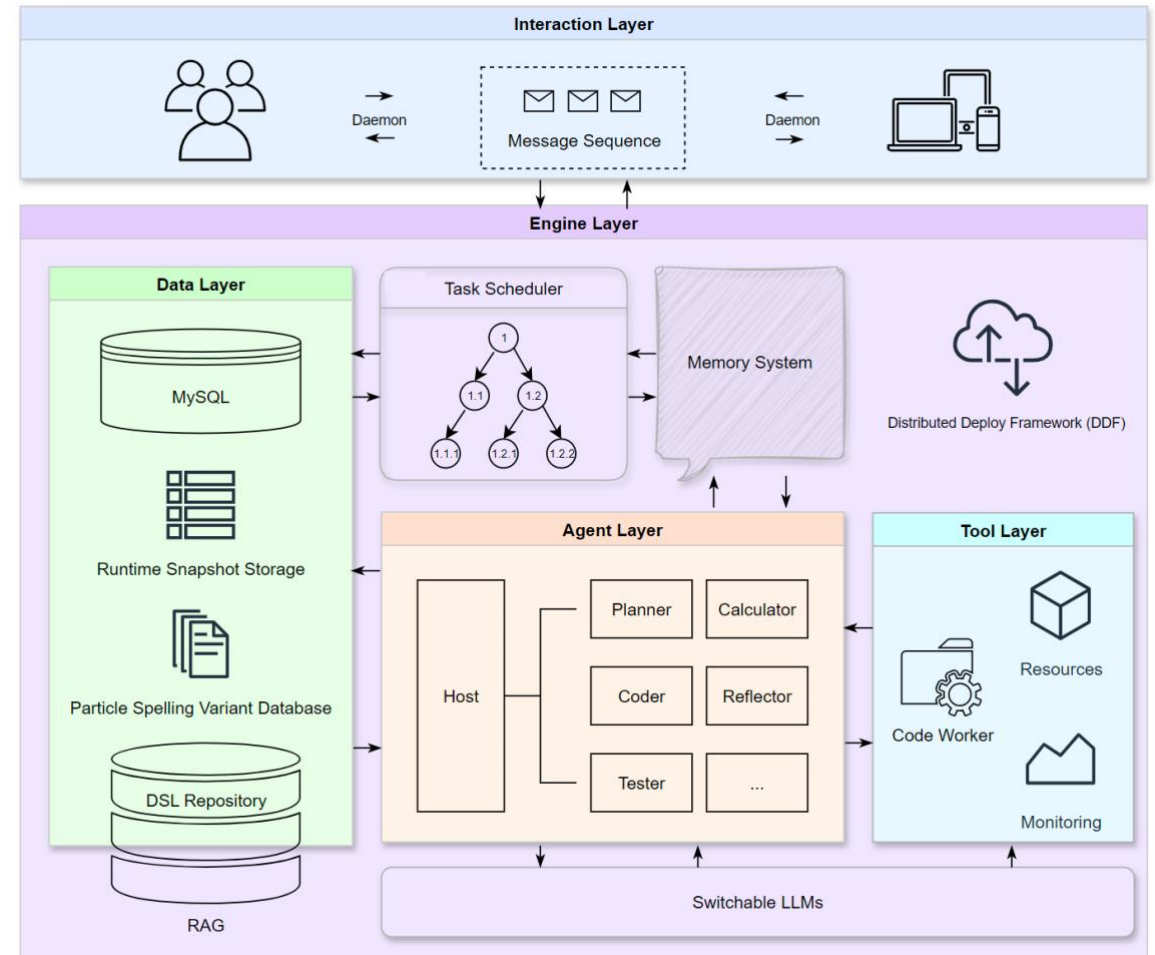
More details in [the talk at ICHEP2024](#)

DrSai V2

- Make BESIII analysis agent-based
- Analysis represented via DSL (preliminary)
- Built a benchmark for BESIII analysis

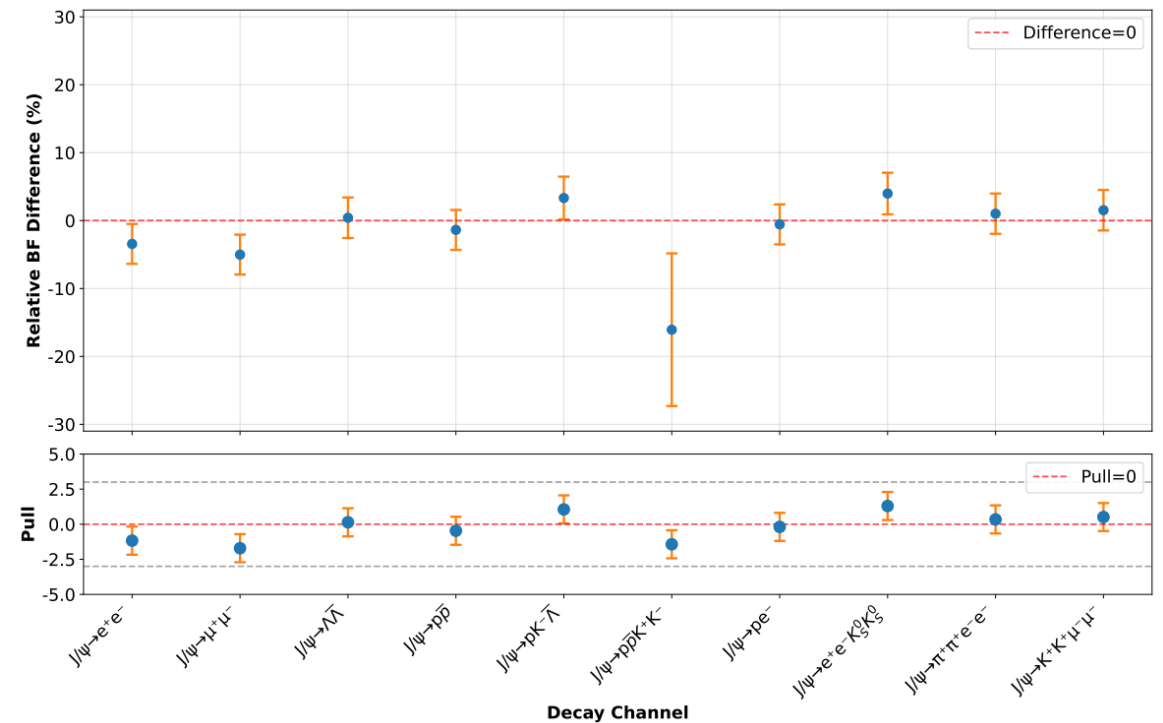


Tested several LLMs



DrSai V2

- The whole system was **validated by re-measuring** known branching fractions (BFs) of the J/ψ decays !
- Built a sample using MC with all known decays
- Use $\psi(3686) \rightarrow \pi^+ \pi^- J/\psi$ process to measure the BF of J/ψ
- DrSai V2 performed the whole analysis
 - From simulation to preliminary systematics study
- All the measured BF showed **good consistency** with injected values
- First full chain validation of DrSai on a realistic physics analysis pipeline



Accepted by Science Bulletin

Identified Core Challenges

Knowledge & Self-Evolution

Often unstructured (experience, intuition). How to make it computable, reasonable, and continuously growing?

Solution

Domain-Specific-Language (DSL) represent analysis with an on-the-fly extraction and iterative refinement from expert feedback

Tools/codes & Execution

Complex software system, and must align with physics goals, as well as auto-checking

Solution

Error-Aware GraphRAG – stores long connections. **DSL/Skill** align with code blocks or tools, enabling multiple checkpoints (e.g., injection tests).

Reasoning & Task planning

High level goals need complex and hallucinations-free decomposition + environment-aware reflections

Solution

Expert guides the system to orchestrate workflows for multiple analyses → distills into Skills → next workflow iteratively refined

Infrastructures & Safety

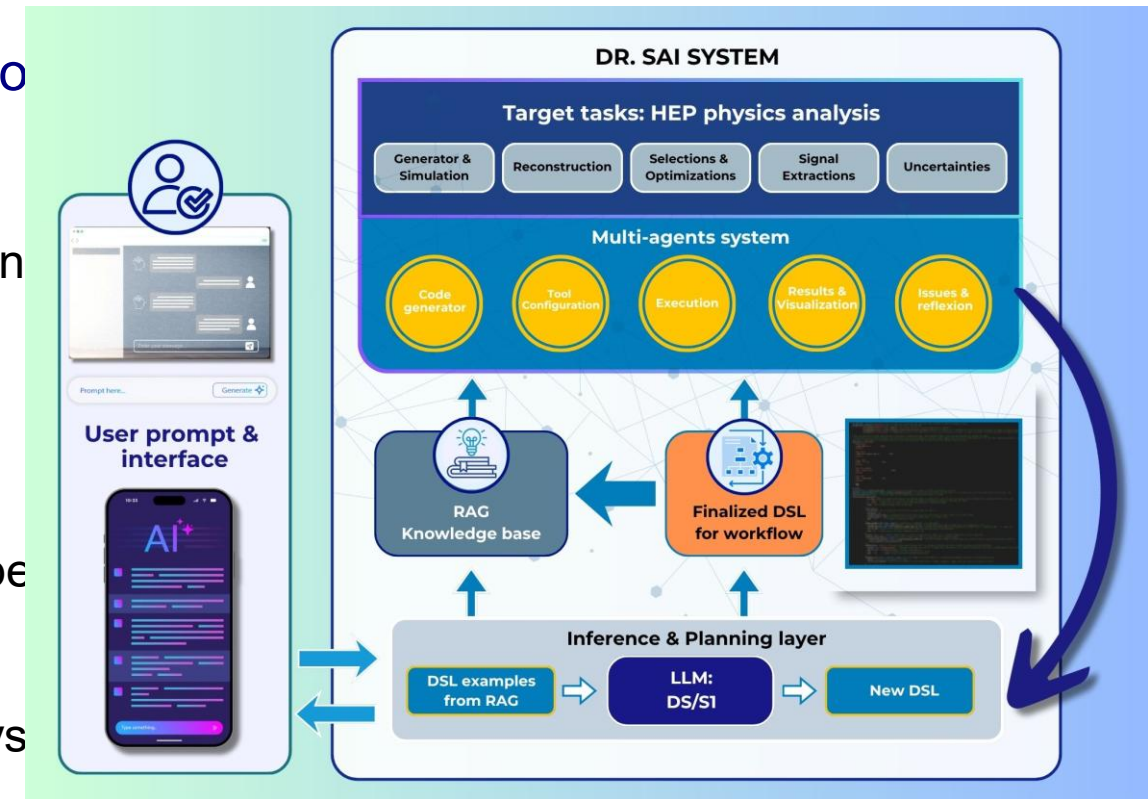
Scaling bottlenecks (GPU costs) vs. data privacy and autonomous safety risks

Solution

Local deployment of open-source LLM (DeepSeek) at IHEP + Cloud support.
Only allow predefined operations.

DrSai V3

- With DrSai V2, we have identified four core challenges and developed corresponding solutions
- Major updates
 - AutoGen-based framework -> Claude Code agent runtime
 - Also DeepSeek-TUI, Cline, etc.
 - A set of analysis Skills – built and refined
 - Hundreds of DSLs extracted from published papers
 - Error-aware GraphRAG for code generation
 - Agentic RAG for internal notes of previous analyses



[Front. Phys., 2026, 21\(9\): 096401](#)

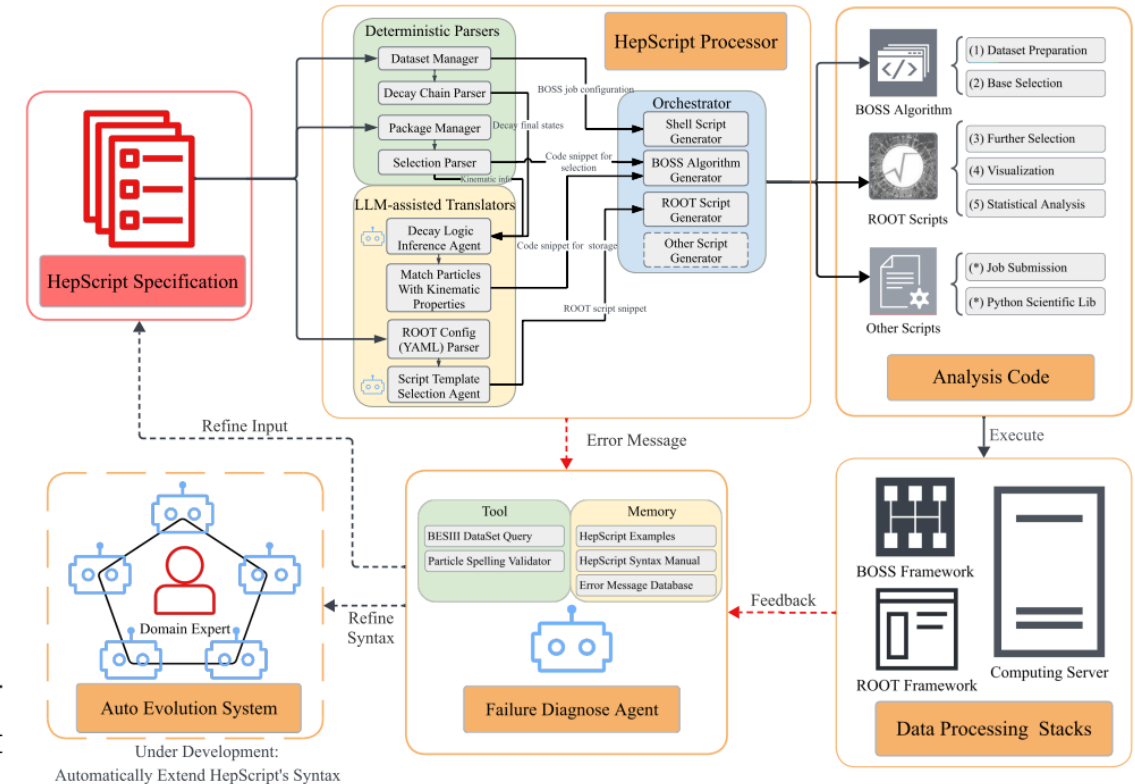
Analysis representation: HepScript (a DSL)

- Represent BESIII analysis
 - Pseudocode in Ruby/YAML format
- Bridges high-level intent with code/tools
- Defines a constrained action space
- The majority are mapped into code blocks
 - Only a small part need conversion by LLM
- **Scalable:** automatically extracted from published paper or generated on-the-fly for new analysis

HepScript	LLM	N_S	N_F	SR (%)	One Retry SR (%)	Three Retries SR (%)
LLM-generated (72 Packages)	Deepseek-R1	34	38	47.3 ± 5.8	87.8 ± 3.8	94.6 ± 2.6
	GLM-4.7	31	41	43.2 ± 5.7	90.5 ± 3.4	95.9 ± 2.3

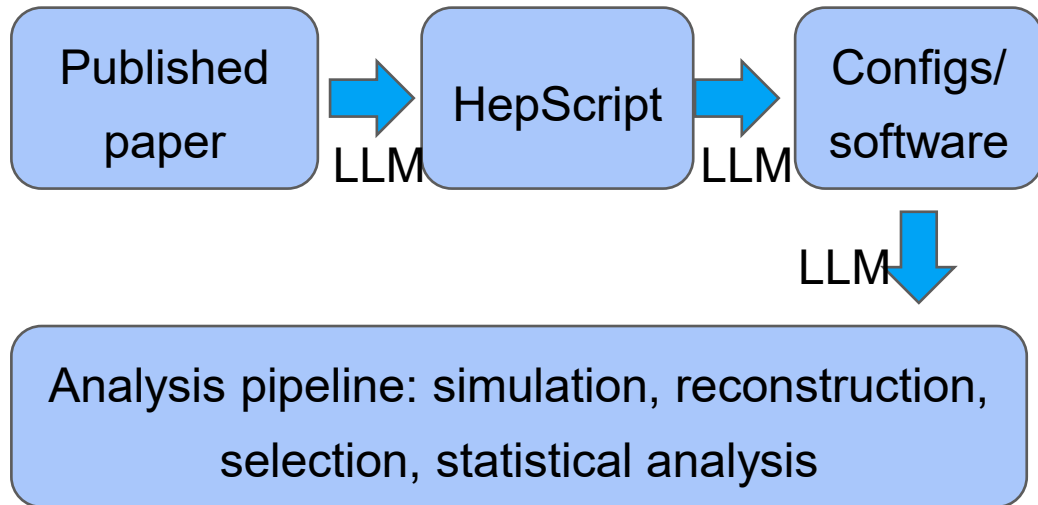
Good accuracy for generation
of analysis algorithms

arXiv:[2605.01423](https://arxiv.org/abs/2605.01423)

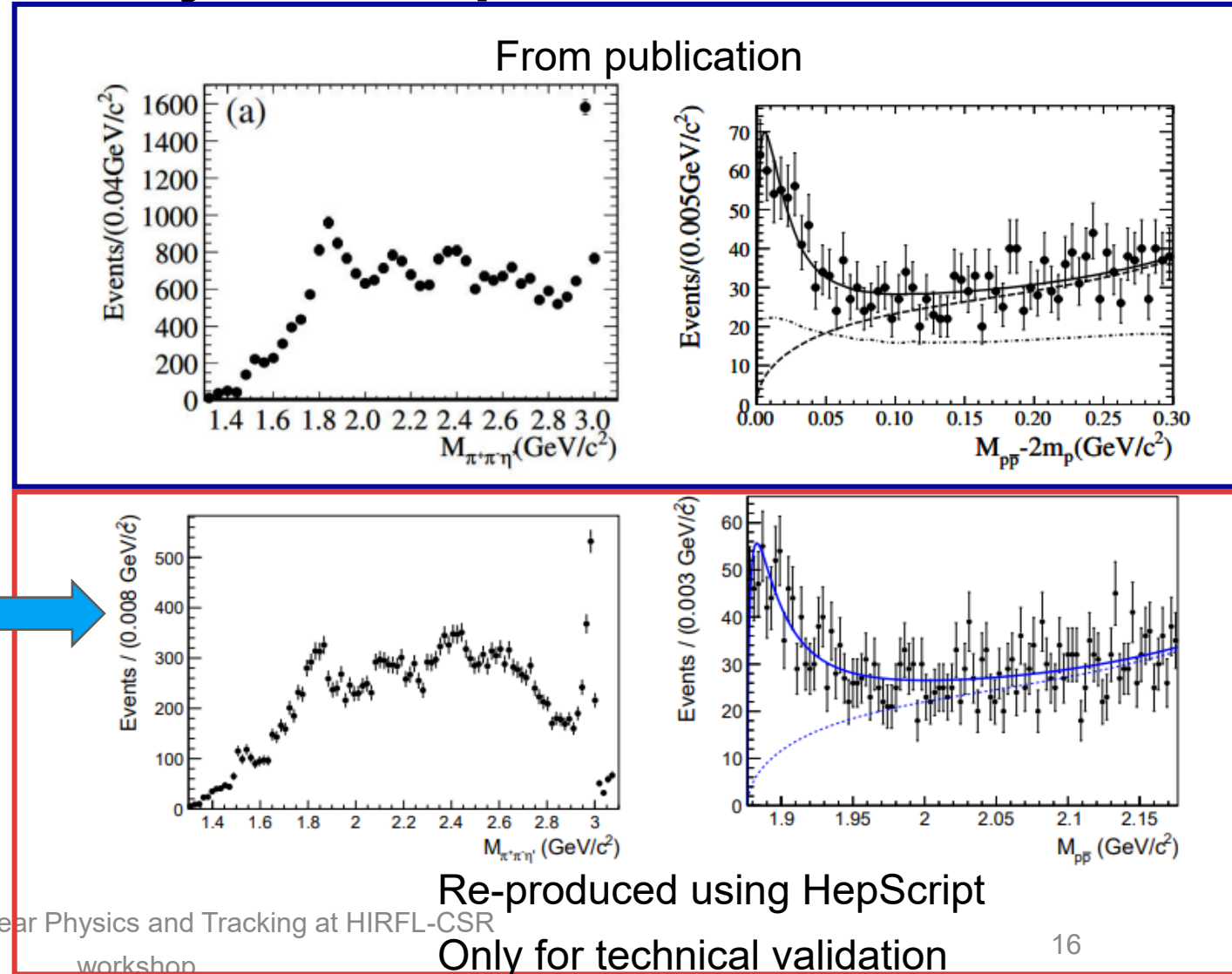


HepScript-driven analysis reproduction

- Necessary for real-world complex analysis
- A showcase: re-produce the previous analysis

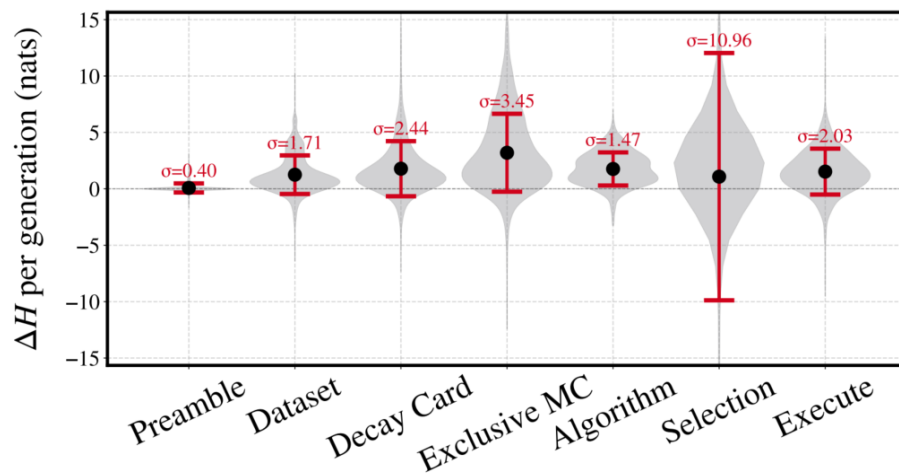


- Used in the DrSai V3 for intent understanding, information completion, and code generation



Knowledge base

- RAG + LLM-wiki, Content in knowledge base
 - Internal docs, e.g. memo and question-answer pairs from internal review
 - Chunked by analysis steps
 - Internal website and introductions
 - HepScript for published papers
- Introduce “systematic uncertainty” when building knowledge base
 - Define weights by “statistical” and “systematic” uncertainties



Algorithm 1 Generic HepScript template.

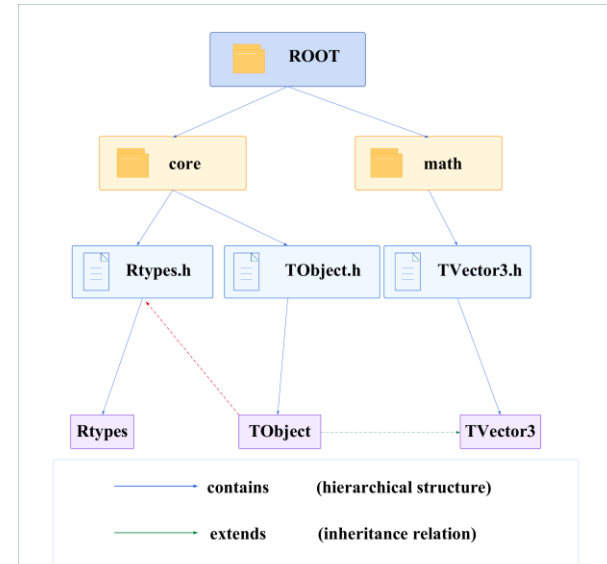
```

1: Step 1 – Datasets
2:    $D_{\text{data}} \leftarrow \text{DATASETMANAGER.REAL\_DATA.FIND}("703\_4260")$   $\triangleright$  At 4.260 GeV
3:    $D_{\text{incMC}} \leftarrow \text{DATASETMANAGER.INCLUSIVE\_MC.FIND}("703\_4260")$ 
4: Step 2 – Decay card  $C_{\text{sig}}$ 
5:    $C_{\text{sig}} \leftarrow \text{Decay } Y(4230): \pi^+ \pi^- J/\psi \text{ (PHSP)}$   $\triangleright e^+ e^- \rightarrow \pi^+ \pi^- J/\psi$ 
6: Step 3 – Exclusive MC  $D_{\text{exMC}}$ 
7:    $D_{\text{exMC}} \leftarrow \text{CREATE\_EXCLUSIVE\_MC}(\text{related} = D_{\text{data}}, \text{events} = 10^5, C_{\text{sig}})$ 
8: Step 4 – Algorithm config
9:    $\mathcal{A} \leftarrow \text{ANALYSIS}(\text{name} = "pipiJpsi", \text{output} = \text{root})$ 
10: Step 5 – Selection Pipeline  $\triangleright$  Analysis-specific
11:   TRACK SELECTION
12:   PHOTON SELECTION
13:   PARTICLE IDENTIFICATION
14:   KINEMATIC FIT
15: Step 6 – Note (un-expressible cut)
16:   NOTES IN CONTEXT  $\triangleright$  free-text physics constraints
17: Step 7 – Bind & Execute
18:    $\mathcal{A}.\text{BIND}(D_{\text{data}}, D_{\text{incMC}}, D_{\text{exMC}}, C_{\text{sig}})$ 
19:   ROOT Files  $\leftarrow \mathcal{A}.\text{EXECUTE\_ON}([D_{\text{data}}, D_{\text{incMC}}, D_{\text{exMC}}])$ 
20: return ROOT Files
  
```

Perturbation	Strict (%)			Struct (%)			Phys-Acc (%)		
	NR	Rel	Conf	NR	Rel	Conf	NR	Rel	Conf
$k = 3$	71.8	<u>76.2</u>	76.7	89.3	<u>91.0</u>	91.0	57.6	<u>62.4</u>	63.4
$k = 10$	71.8	<u>75.7</u>	76.3	89.3	<u>90.6</u>	90.8	57.6	<u>62.6</u>	63.3
paraphrase	71.8	<u>75.2</u>	76.0	89.3	<u>90.3</u>	90.7	57.6	<u>61.8</u>	63.9
ORACLE	100			100			100		

Domain code/script generation

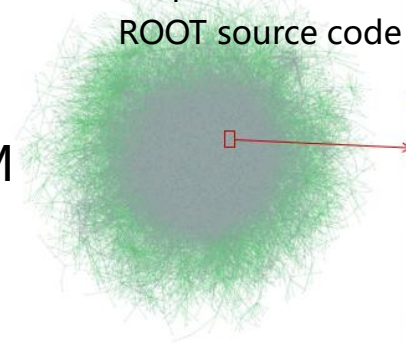
- Complex code dependencies at HEP
 - LLMs struggle to reason over large-scale, intricate dependency correctly and efficiently
- We start with open-source software - ROOT
 - **Massive interconnections -> a large graph**
 - Nodes : class, package, etc.
 - Edges : contains, extends, depends
 - Obvious improvement on accuracy
 - Further improvement from **Error-Aware RAG**
 - First find the related subGraph, then feed LLM
- This method can be easily applied to HEP software development



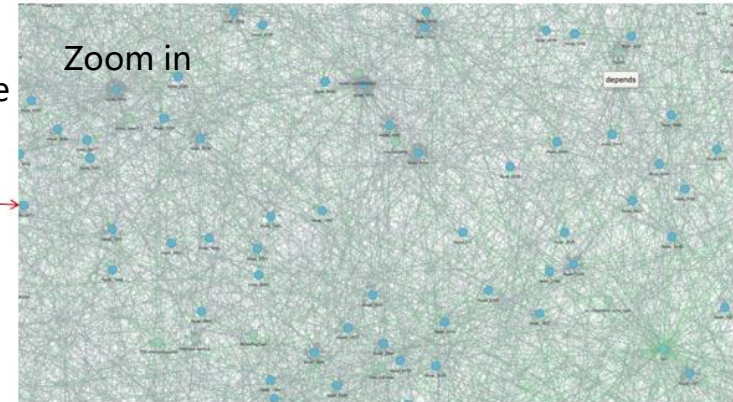
- Total nodes: 47851
- Total edges: 163795

- REPO nodes: 1 | - PACKAGE nodes: 2609 | - FILE nodes: 19347
- TEXTFILE nodes: 5255 | - CLASS nodes: 20639

Graph for whole
ROOT source code



Zoom in



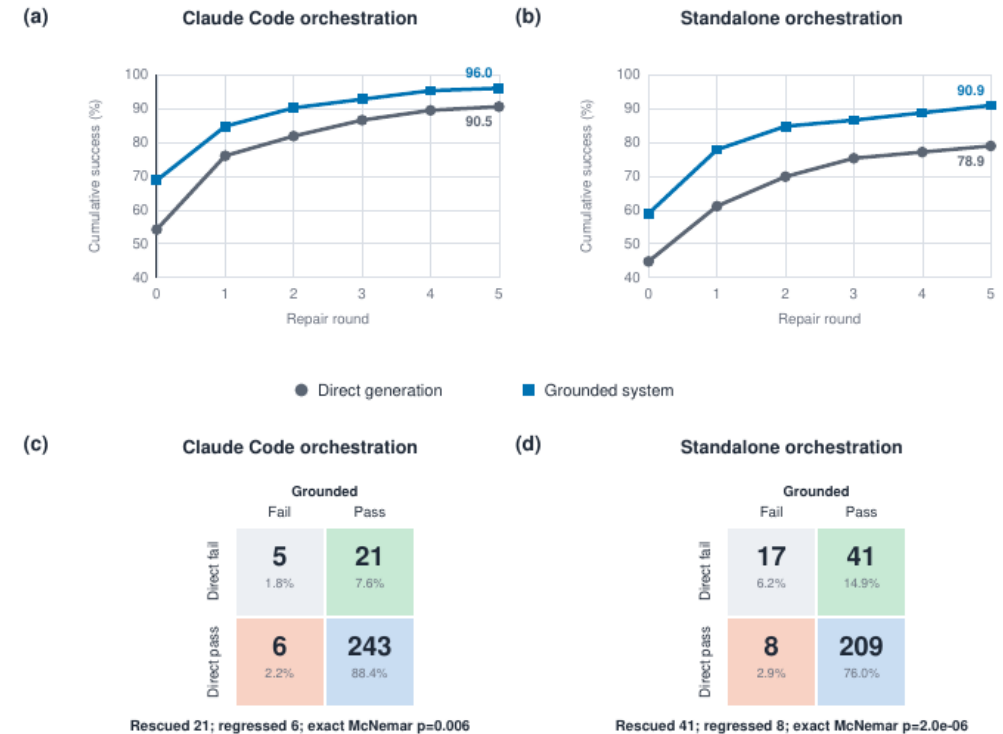
Github Repo:

<https://github.com/yues36067-cmd/ROOTER>

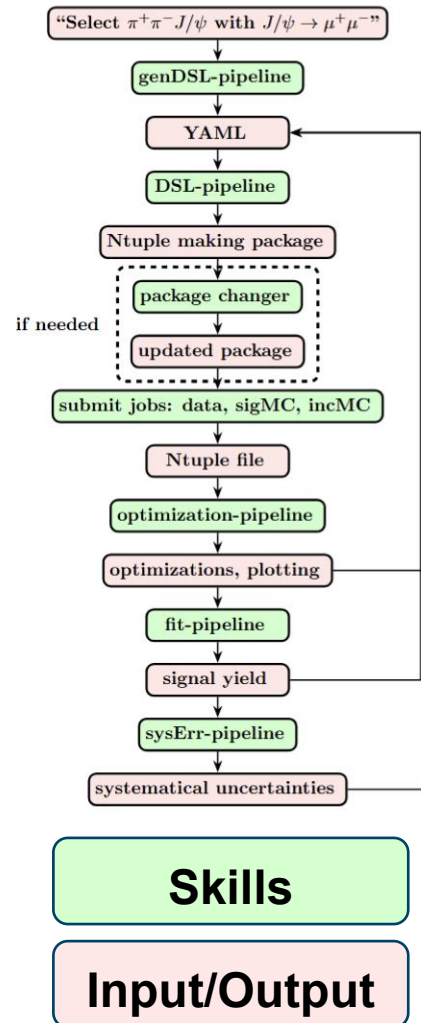
arXiv: [2609.01095](https://arxiv.org/abs/2609.01095)

Domain code/script generation

- Two benchmarks and evaluation methods have been developed
 - Good improvement on accuracy
 - LLM: deepseek V4 pro
- Success rate improved by ~10%
 - Still helpful for SOTA coding agent, **claude code** (~5%)
- Integrated to DrSai, works well



A recording for Zc(3900) re-discovery



```

(drSai) [liutong2016@beslogin005 DSL]$
(drSai) [liutong2016@beslogin005 DSL]$
(drSai) [liutong2016@beslogin005 DSL]$
(drSai) [liutong2016@beslogin005 DSL]$echo $PATH
/hpcfs/bes/gpupwa/liutong2016/conda/miniconda3/envs/drSai/bin:/hpcfs/bes/gpupwa/liutong2016/conda/miniconda3/condabi
n:/cvmfs/bes3.ihep.ac.cn/bes3sw/Boss/7.1.3/InstallArea/share/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/ExternalLib/Gaudi/Gau
di-v27r1-g4107-python2.7/InstallArea/x86_64-centos7-gcc49-opt/scripts:/cvmfs/bes3.ihep.ac.cn/bes3sw/Boss/7.1.3/Insta
llArea/x86_64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/ExternalLib/Gaudi-v27r1-g4107-python2.7/Insta
llArea/x86_64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/ExternalLib/lcg/LCG_84/python2.7/x86_
64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/ExternalLib/lcg/LCG_84/sqlite/3.11.0.0/x86_64-centos7-gcc49-op
t/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/ExternalLib/lcg/LCG_84/mysql/5.5.38/x86_64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihe
p.ac.cn/bes3sw/ExternalLib/lcg/LCG_84/XercesC/3.1.1p1/x86_64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Ext
ernallib/lcg/LCG_84/castor/2.1.13-6/x86_64-centos7-gcc49-opt/usr/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/ExternalLib/lcg/L
CG_84/ROOT/6.20.02/x86_64-centos7-gcc49-opt-python2.7/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/ExternalLib/lcg/LCG_84/xroot
d/4.2.3/x86_64-centos7-gcc49-opt/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/ExternalLib/lcg/LCG_84/Python/2.7.10/x86_64-centos7-gcc49-opt/bin
:/cvmfs/bes3.ihep.ac.cn/bes3sw/ExternalLib/lcg/LCG_84/gcc/4.9.3/x86_64-centos7/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/Ext
ernallib/lcg/LCG_84/CMT/v1r25/Linux-x86_64:/cvmfs/bes3.ihep.ac.cn/bes3sw/ExternalLib/contrib/gdb/7.6.1/bin:/hpcfs/be
s/gpupwa/liutong2016/conda/miniconda3/envs/drSai/share/rubygems/bin:/hpcfs/bes/gpupwa/liutong2016/conda/miniconda3/b
in:/afs/ihep.ac.cn/users/l/liutong2016/topoana-02-05-03/bin:/cvmfs/container.ihep.ac.cn/bin:/afs/ihep.ac.cn/soft/com
mon/sysgroup/hep_job/bin:/usr/local/bin:/usr/bin:/usr/local/sbin:/usr/sbin:/cvmfs/common.ihep.ac.cn/software/cctools
:/opt/puppetlabs/bin:/cvmfs/bes3.ihep.ac.cn/bes3sw/ExternalLib/lcg/LCG_84/ernlib/2006a/x86_64-centos7-gcc49-opt/bin
(drSai) [liutong2016@beslogin005 DSL]$C1
Claude Claude2
(drSai) [liutong2016@beslogin005 DSL]$Claude2
Claude Claude v2.1.143
glm-5.1[1m] · API Usage Billing
/besfs5/groups/tauqed/liut/R/DSL

> 做pipiPsi分析。配置在config_pipiPsi.yaml

• Explore(Find config_pipiPsi.yaml)
  | Done (3 tool uses · 13.9k tokens · 22s)
  | (ctrl-c to expand)

Read 1 file (ctrl-c to expand)

• 配置已读取。启动 full-pipeline.

• Skill(/full-pipeline)
  | Successfully loaded skill

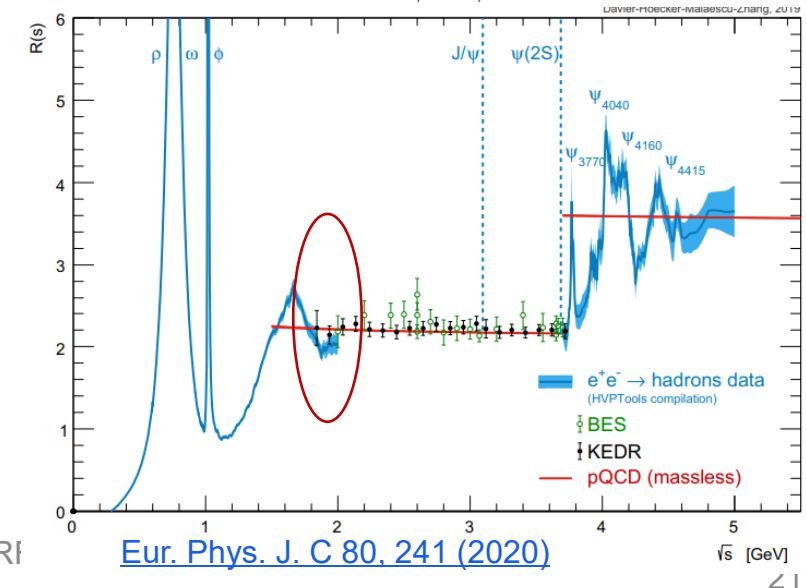
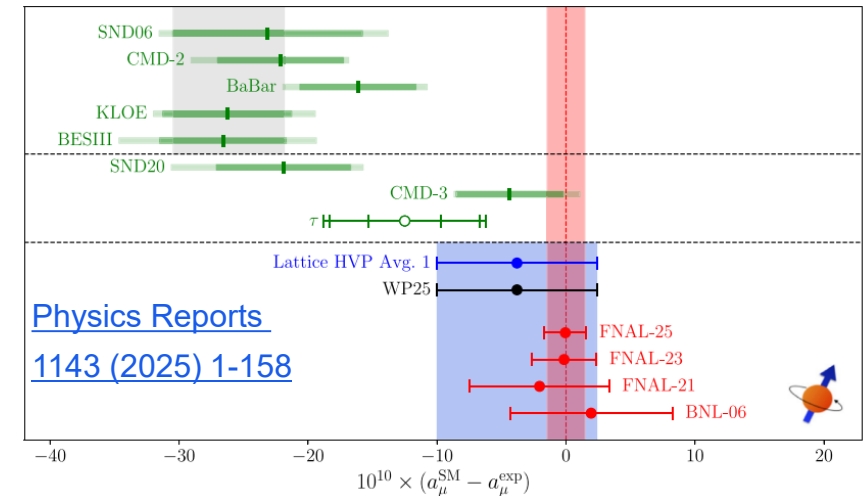
+ Creating... (47s · 1654 tokens · almost done thinking)
  | Tip: Use /btw to ask a quick side question without interrupting Claude's current work

>
  >> auto mode on (shift+tab to cycle) · esc to interrupt
  
```

begin

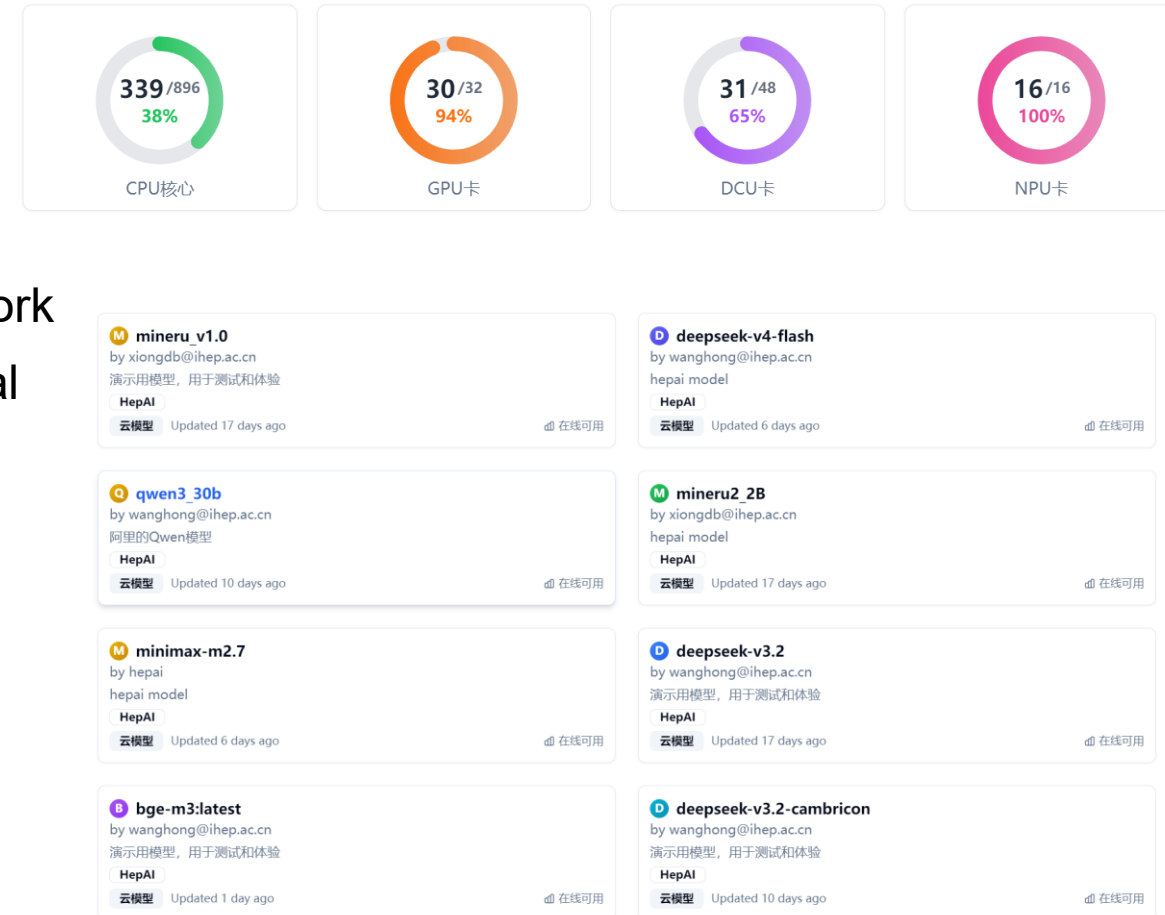
Analysis using DrSai at BESIII

- DrSai V3: **ready for BESIII physics analysis !**
 - **Vibe research – no coding needed now**
- **Now analyzing real data at BESIII:**
 - Measure all the hadronic cross section exclusively below 2 GeV, **a painful analysis for human thus long delayed**
 - A key input for HVP in muon g-2 calculation
 - Dispersive evaluation differ from LQCD
 - Experimental measurements of $e^+e^- \rightarrow \text{hadrons}$ are particularly important.
 - **Starting from 1.9 GeV:**
 - Experts steer the setup and input essential parameters, e.g. generators
 - DrSai design, execute and verify the entire analysis under expert supervision
 - Experts will **cross-check** all the measurements



AI Infrastructure at IHEP

- New dedicated AI platform:
 - 2240 CPU cores, 40 GPU, 48 DCU, 16 NPU
 - More (2X) are coming
- LLMs access and deployment
 - Unified API via Distributed Deployment Framework
 - **Hybrid solution:** local deployment + commercial cloud
 - Open-source LLMs (e.g. DeepSeek)
 - High concurrency
 - Ensure data security
 - Closed-source LLMs accessible [via API](#)
 - Cache-hit techniques reduce the cost to 2-10%



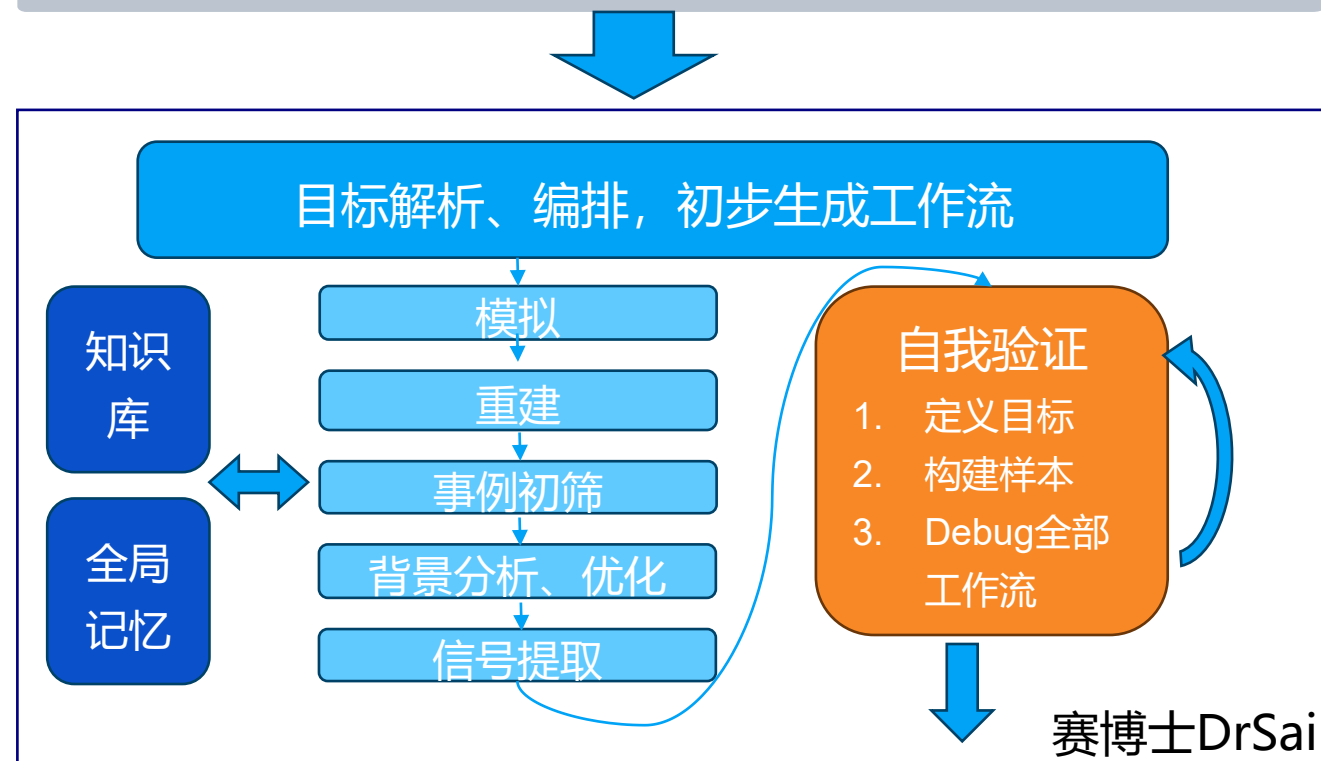
DrSai Collaboration

- DrSai is already yielding physics results
 - Physicists interact with DrSai using natural language, only a few times with **zero coding!**
- Now extending to other experiments:
 - Jiangmen Underground Neutrino observatory (JUNO)
 - Large High Altitude Air Shower Observatory (LHAASO)
 - Neutron/Light source, etc.
- We are forming a DrSai collaboration
 - Solving Agentic AI challenges together
 - Shared AI infrastructure & computing resources in collaboration
 - Welcome to join!

On-going efforts

- Migrating the DrSai to DeepSeek Harness
 - Open-source, good maintenance
 - Consist physics results with Claude-Code
- Developing a new module: Truth-Anchored Recursive Self-Improvement
 - Use injection test to help the Agent to improve the whole workflow
- Many cutting-edge techniques
 - Welcome to collaborate

提示词：请用质心系能量为4.26GeV的数据分析
 $e+e \rightarrow \pi+\pi^- J/\psi$ 过程，并寻找 $\pi+J/\psi$ 上新的粒子



显著提升稳定性，目前可并行处理>30个物理过程

AI at lepton colliders

More than 10B events

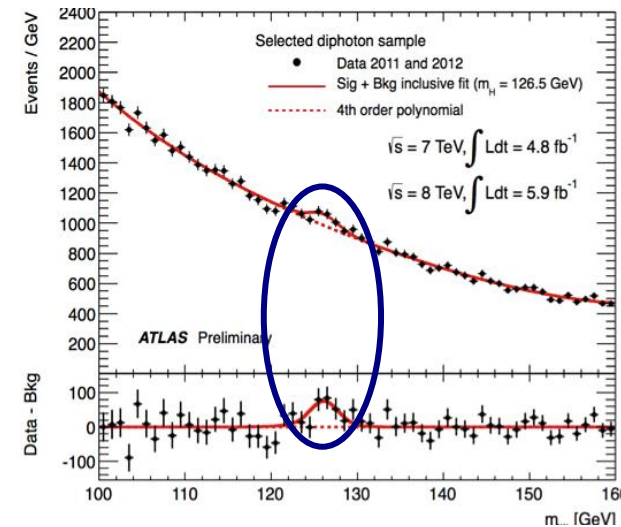


Significance

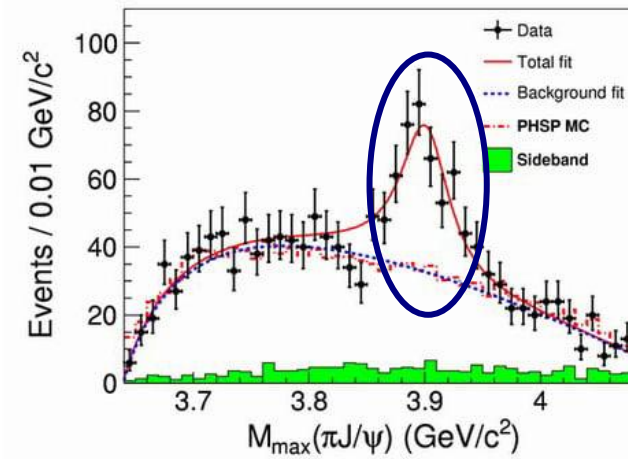
$$Z \approx \frac{S}{\sqrt{B}}$$

AI4S 3 main approaches:

- **Improve S**
- Reduce B (more important at LHC)
- Accelerate analysis



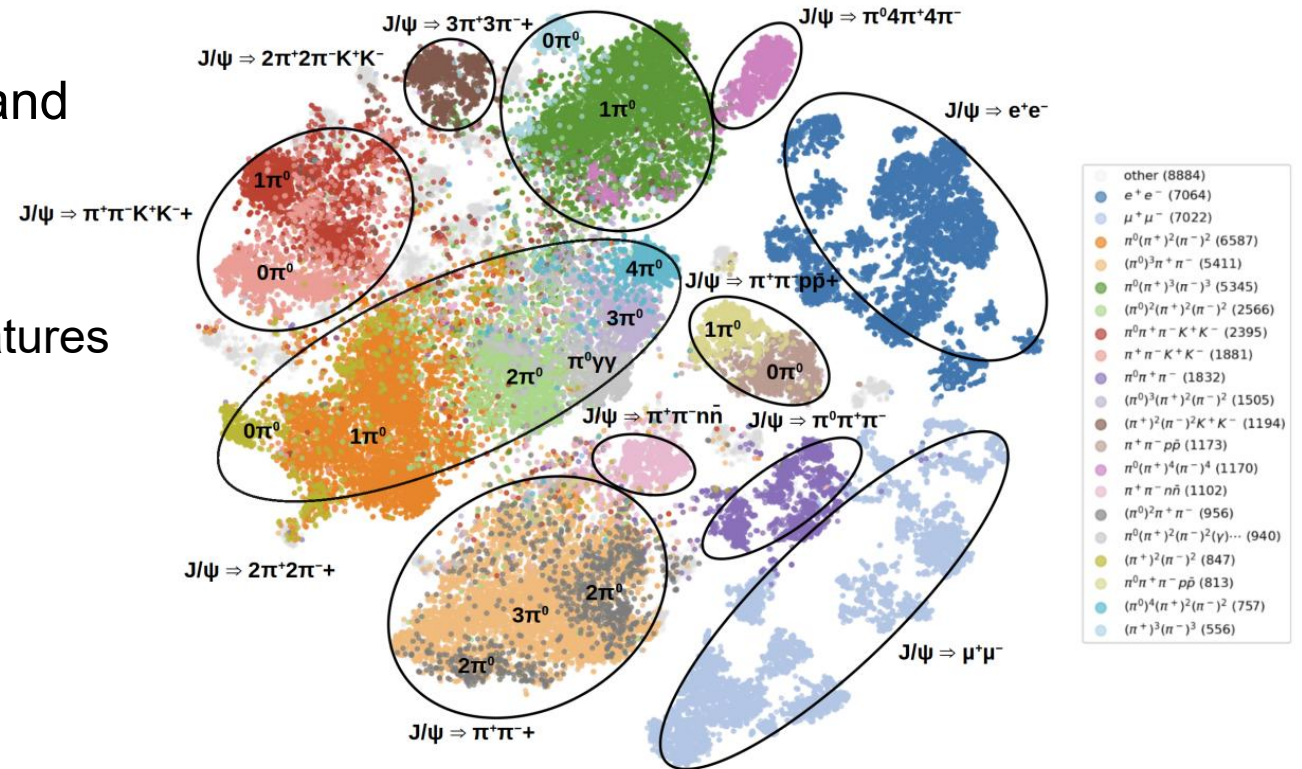
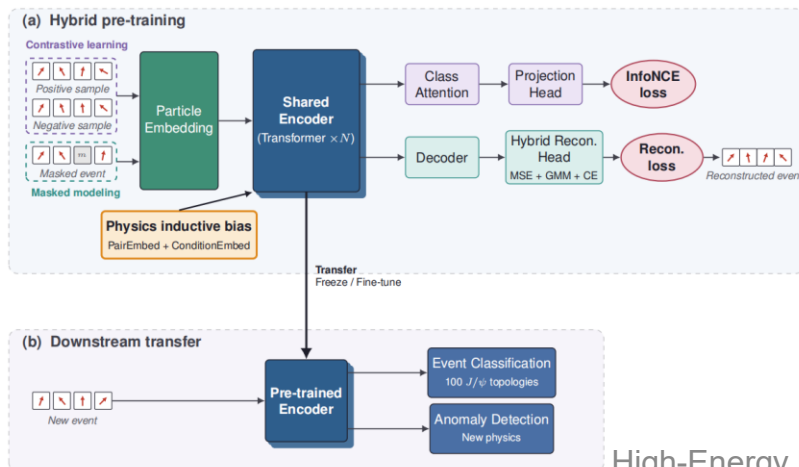
Higgs at ATLAS



Zc(3900) at BESIII

Foundation model to classify all the events

- Goal: classify all the channels and support all downstream analyses
- Event representation using self-supervised and Contrastive Learning
 - Mask one particle then predict
 - Gate attention conditioned on event level features
 - Supervised fine-tune with a small dataset

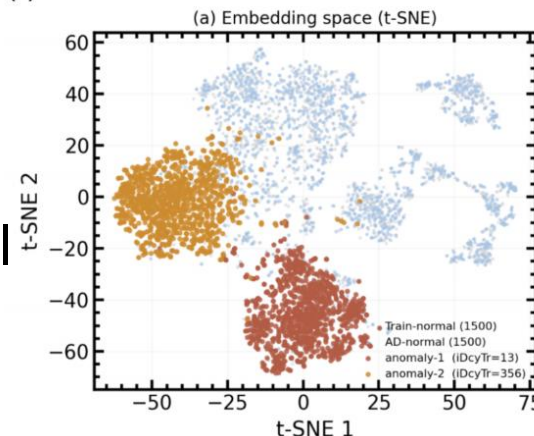


Foundation Model

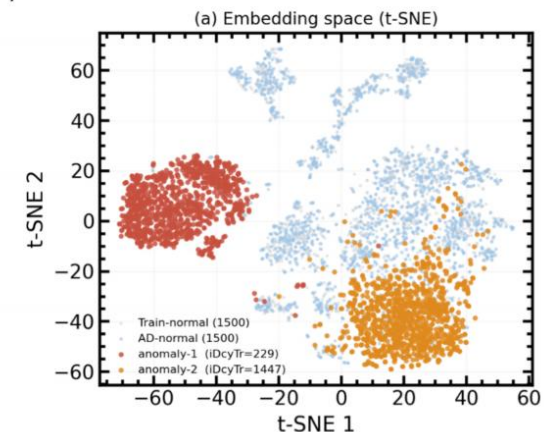
- Compared to cut-based analysis
 - Higher signal efficiency, cover all the decay channels
 - Lower background, all the bkg can be classified as well
 - Support almost all downstream analysis tasks
- Compared to ML-based analysis
 - No need to retrain the model
- For the new particle or channel
 - Auto-clusterization as anomaly
 - Different new particles or channels can be clusterized to different “shower”
- Now only use 10M events for training, will expand to 1B

Final state	Method	Precision	Recall	F_1
$\pi^+ \pi^- p\bar{p}$	Cut1	6.3%	56.5%	11.3%
	Cut2	99.5%	50.9%	67.3%
	CLEAR ($\tau = 0.5$)	82.9%	92.5%	87.4%
	CLEAR (matched)	—	—	—
$\pi^+ \pi^- K^+ K^-$	Cut1	31.9%	55.8%	40.6%
	Cut2	99.3%	46.0%	62.9%
	CLEAR ($\tau = 0.5$)	92.8%	78.0%	84.7%
	CLEAR (matched)	99.3%	59.8%	74.6%
$\pi^+ \pi^- \pi^0$	Cut1	79.0%	51.8%	62.6%
	Cut2	97.7%	37.2%	53.9%
	CLEAR ($\tau = 0.5$)	90.3%	90.9%	90.6%
	CLEAR (matched)	97.7%	46.0%	62.6%

(1). Near-OOD Detection

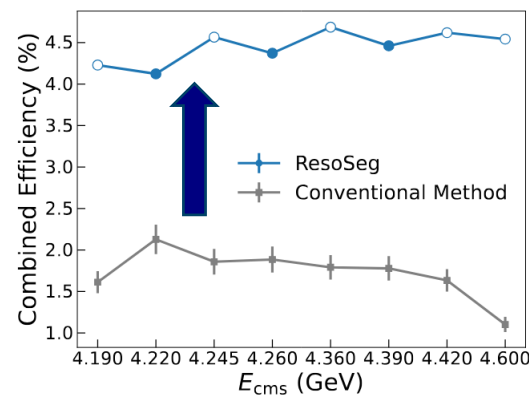
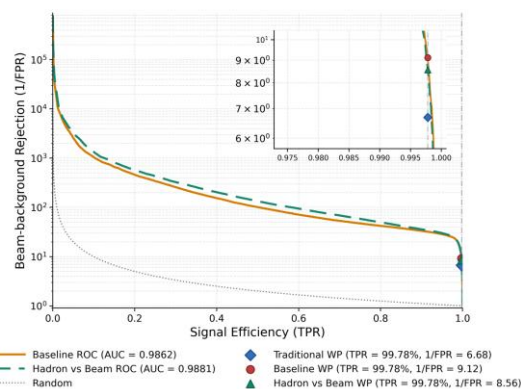
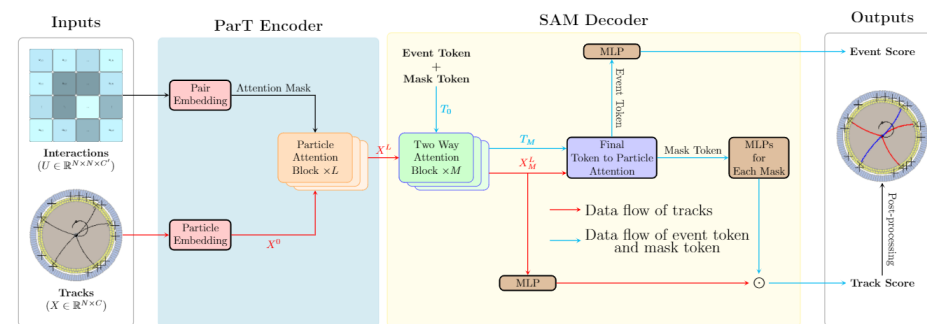


(2). Far-OOD Detection



Other studies

- 在FPGA上部署轻型NN用于快速去除背景
- 基于真实数据训练
- 束流背景排除提升20-50%**
- 延迟仅35ns
- 对BEPCLII升级和下一代对撞机尤其重要



通用强子标记器

- 利用Transformer训练表征空间
- 利用SAM模型分割末态粒子
- 效率相比于传统方法提升2.6倍**

Summary and outlook

- Possible two new approaches for AI4S studies at lepton collider
 - Unique scenario w.r.t. hadron collider
 - Systematically analyze all possible process automatically
 - Autonomous analysis using DrSai - analyzing real data now
 - Use experimental data to build a large model - finalizing prototype
 - Classify all the events with foundation model
 - Tag the intermediate resonance by segmentation
- Welcome to join ! We are forming a DrSai collaboration now
 - Extending to: JUNO/TAO data production, LHAASO analysis, ATLAS DQM, CMS ntuple analysis, neutron/light source analyses, etc.

backup