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

Leveraging Agentic AI for Global and Systematic New Physics Search

李科（like@ihep.ac.cn）

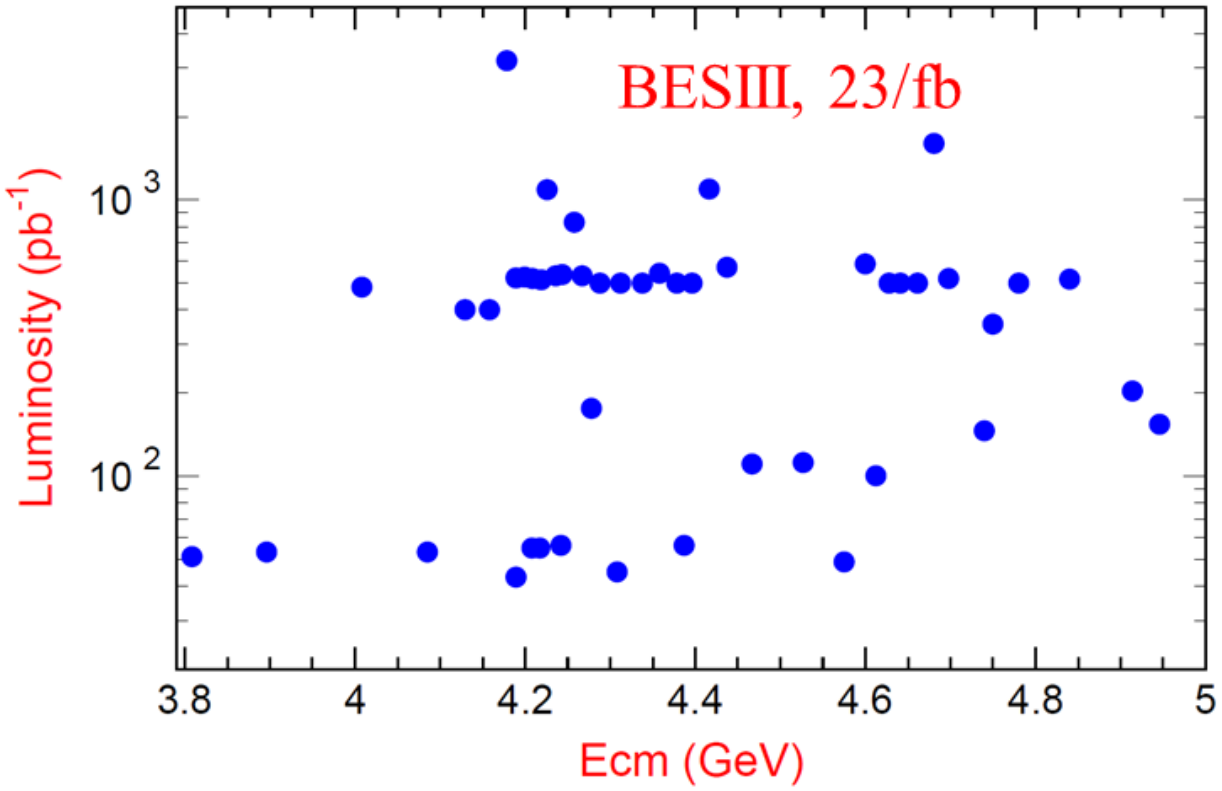
Institute of High Energy Physics

Outline

- New ways to search for NP
  - The systematic search with DrSai by automating the full analysis
  - The global search with cross-dataset anomaly detection
- Summary and outlook

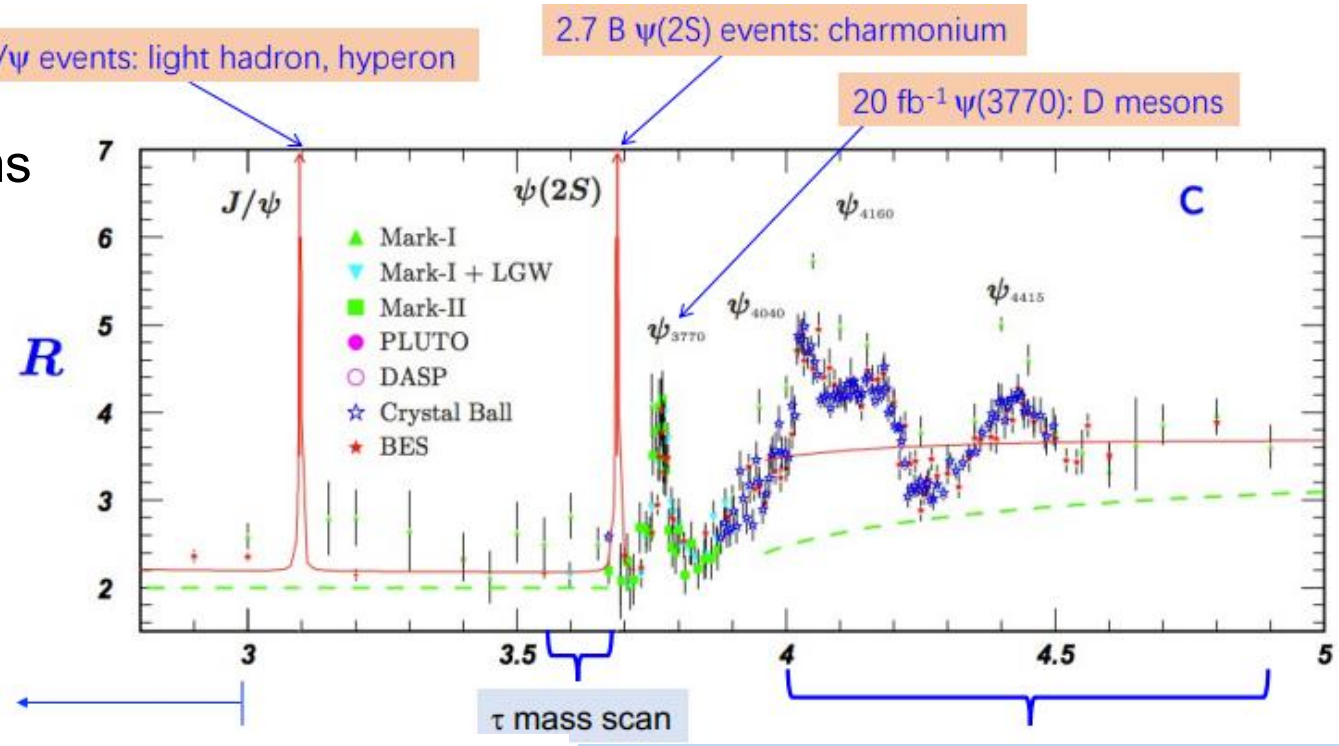
Personal Notes

- Many NP theories predict new particle/phenomena
- Typically, most searches are based on a specific production or decay process
  - Limitted 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



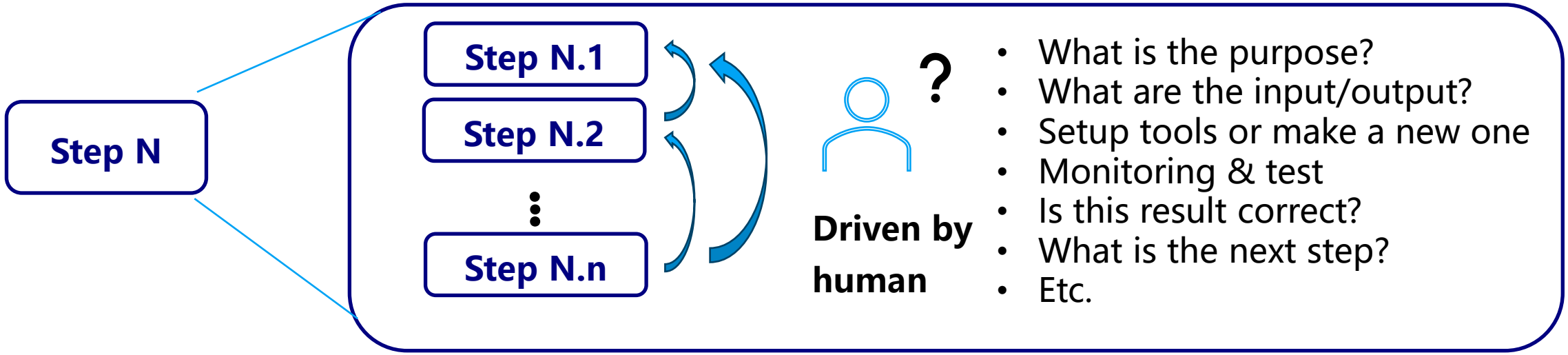
Introduction of BESIII - physics program

- >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 hardrons
    - First tetraquark: Zc(3900)
  - **Clear and relative simpler workflow for effective modelling**
  - **Different from LHC, unique challenges and advantages in lepton collider experiments**



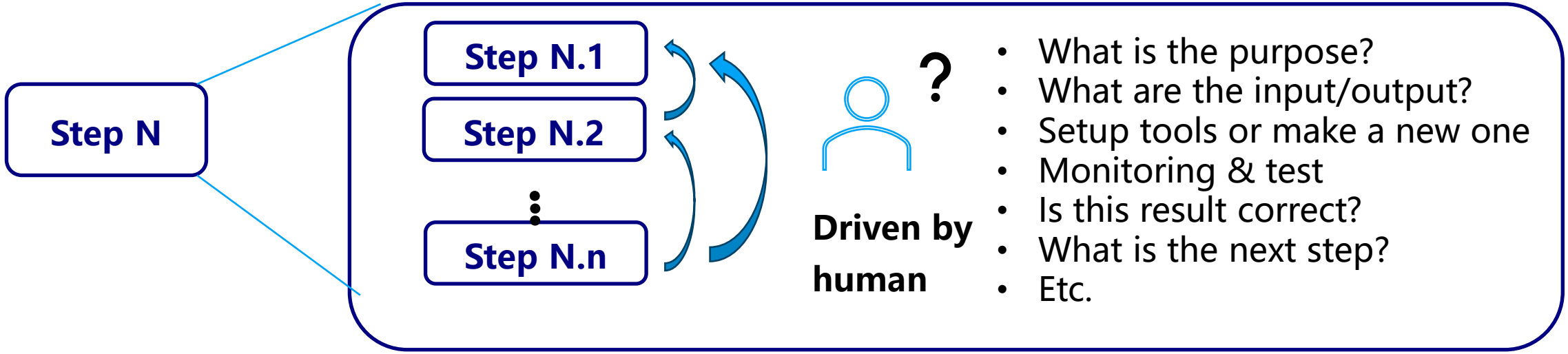
Agentic AI for HEP

Traditional approach



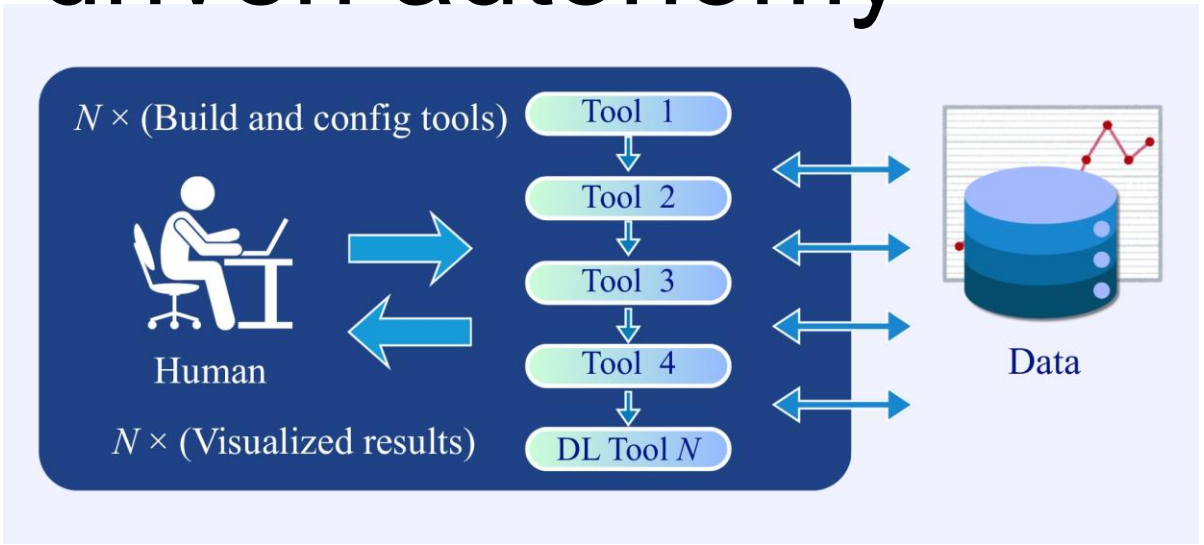
Agentic AI for HEP

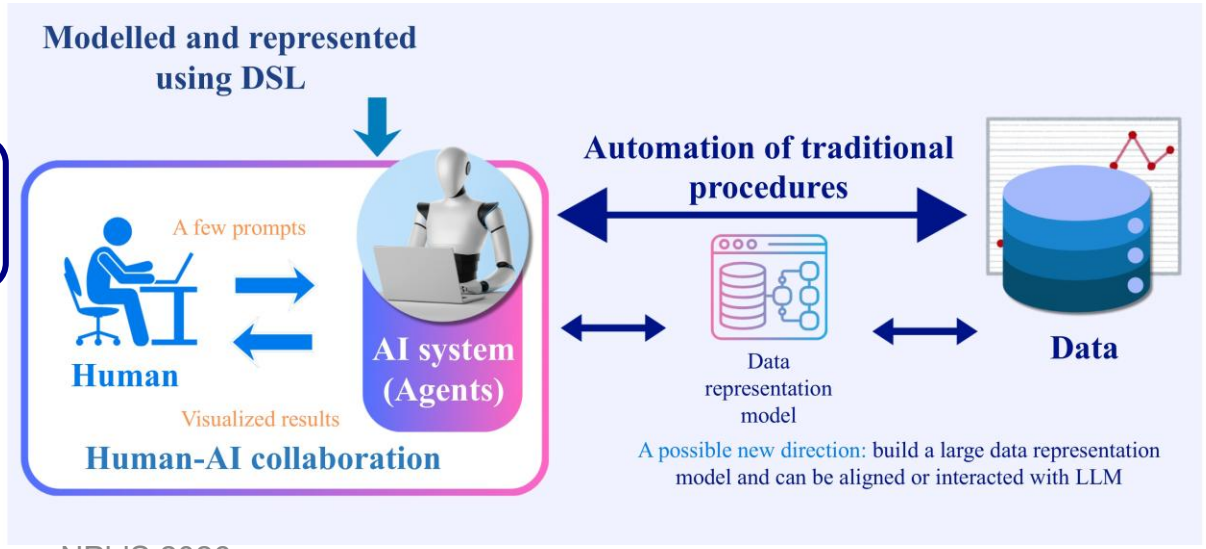
- In traditional approaches, human involvement is required at all times.
  Expertise is often unstructured, though some can be represented by natural language.
- Today, we can leverage LLMs to build an AI partner for scientists.
  - **The AI partner designs and conducts the entire workflow based on human 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.**



AI scientists: Goal-driven autonomy

- An AI system with structured knowledge representation and autonomous research capabilities
- Can perform the full research loop and continue revolution:
  - Goal → Design → Execute → Analyze → Generate new knowledge
- **What we can achieved in this year:**
  - AI Partner: relies on human commands and operate through "vibe research"
- Then target at AI Scientist: human only need to sets high-level goals





Challenges & Bottlenecks

- How to represent HEP-related knowledge?
  - Scientific knowledge is often unstructured (natural language, intuition). How to make it computable and reason-able?
- How to ensure reliable reasoning and decompose goals into tasks?
  - Avoid hallucinations or wrong assumptions without constant human feedback.
  - From a high-level goal (e.g., "find a low-toxicity anticancer molecule") to concrete experimental steps.
- How to efficiently and correctly write domain code/configure and call domain tools?
  - Ensure correct implementation and invocation of domain-specific software, instruments, or pipelines.
- How to enable self-validation and learning?
  - Automatically refine knowledge representations and hypothesis generation from experimental results.
- How to set safety boundaries?
  - Prevent uncontrollable risks (e.g., dangerous experiments, fake papers) by embedding human values and oversight.

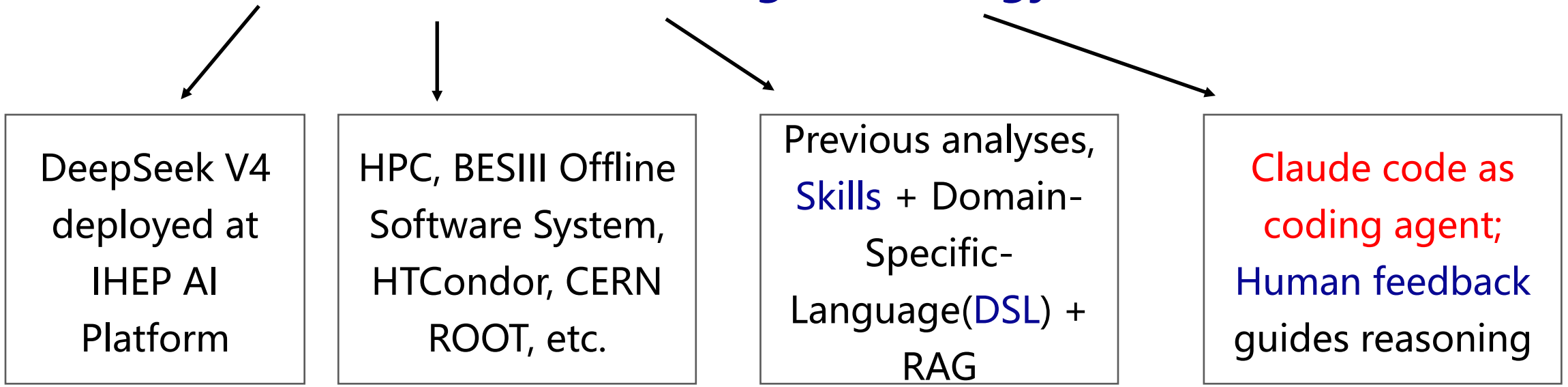
Possible solutions

- How to represent HEP-related knowledge?
  - Define a Domain-Specific-Language (DSL, HepScript) to represent data analysis.
  - Extract from published papers and align with domain code/tools.
- How to ensure reliable reasoning and decompose goals into tasks?
  - Use Skills and RAG to interpret and store previous analyses.
- How to efficiently and correctly write domain code / configure and call domain tools?
  - Use Error-aware GraphRAG and Skills.
- How to enable self-validation and learning?
  - Continuously update DSL and Skills based on human feedback.
- How to set safety boundaries?
  - Only allow predefined operations.

The DrSai project

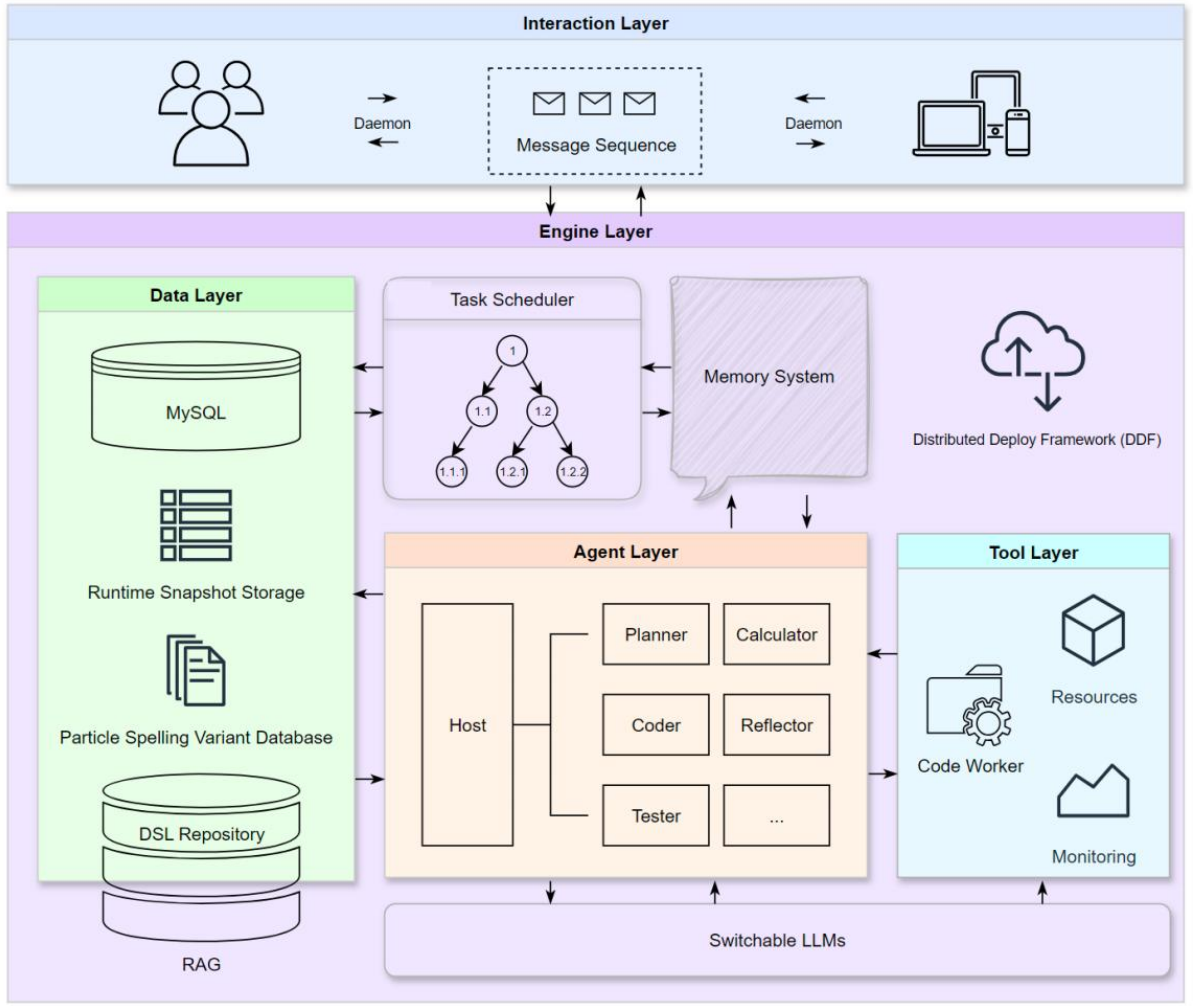
Pioneering autonomous physics discovery at BESIII (and beyond) through Agentic AI

Dr.Sai ≈ **LLM + Tools + Knowledge + Strategy**



The DrSai project

- Start from 2023.06
- Feasibility test completed at 2024.06, V1
- A prototype was conducted at 2025.06, V2
  - Integrated to BESIII Offline Software System, IHEP HPC, and HTCondor system
  - Re-organized analysis algorithms
  - Pre-defined analysis workflow
    - Capable of very simple analysis with fixed workflow
  - Validated in real-world environment
- Base LLM is switchable for agents
- Previously developed based on AutoGen framework, now using Claude code
  - **Almost finished (V3), expect to have first end-to-end experimental results at end of 2026 !**



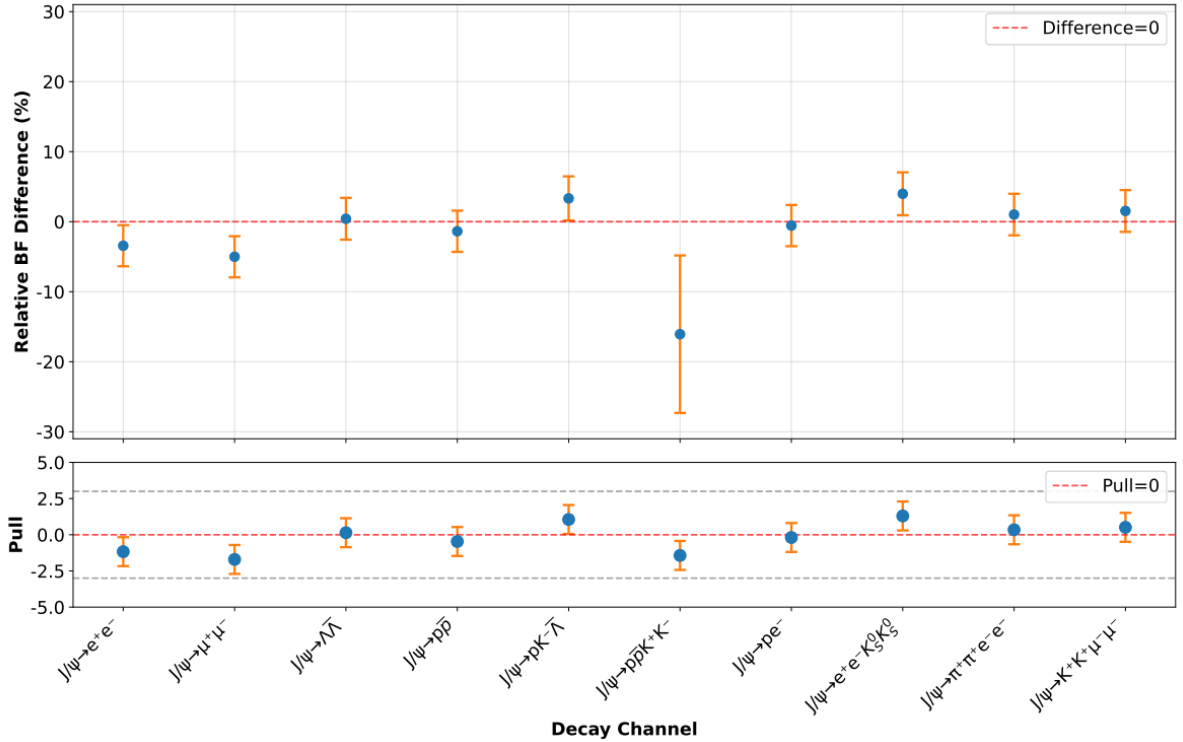
arXiv:2604.22541

Millestone developments

Standardize and integrate all the tools and computing environment, including BESIII Offline Software System(BOSS) and HTCondor	**Done at 2024.06**
Represent the knowledges on how to used tools and analysis procedures into a DSL, automatically extracted from published papers	**Done at 2025.06**
Distill human expertise into Skills on how to perform analysis, including workflow definition and tests (i.e. input/output test)	**Will finish at 2026.06**

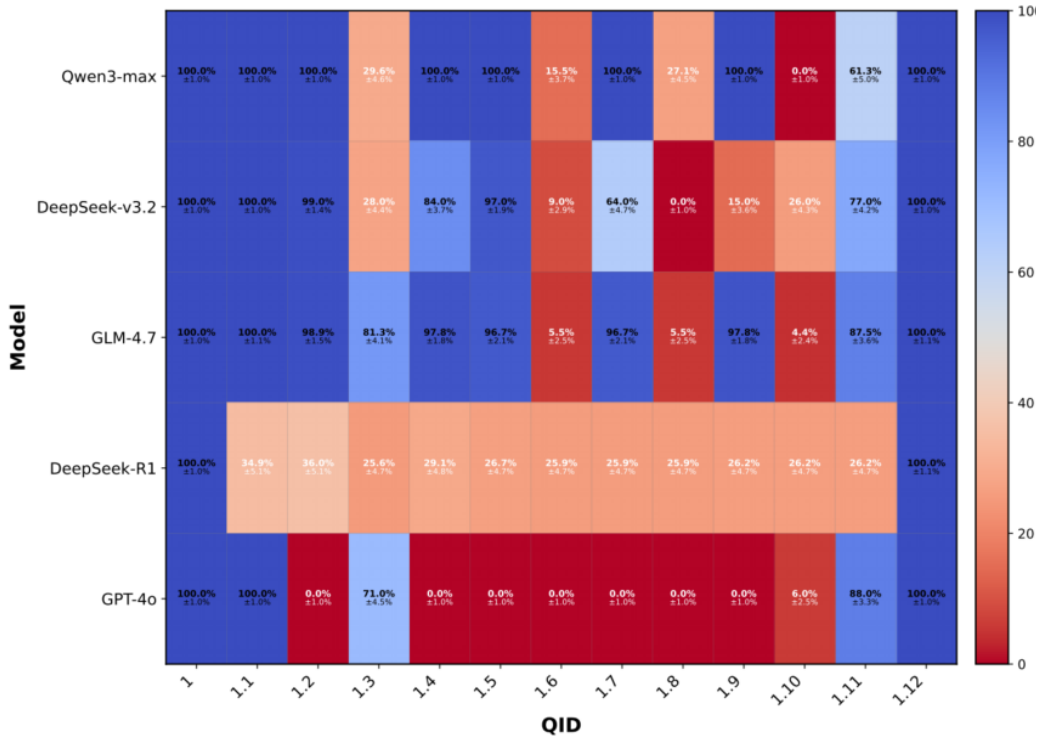
Demonstration of DrSai V2

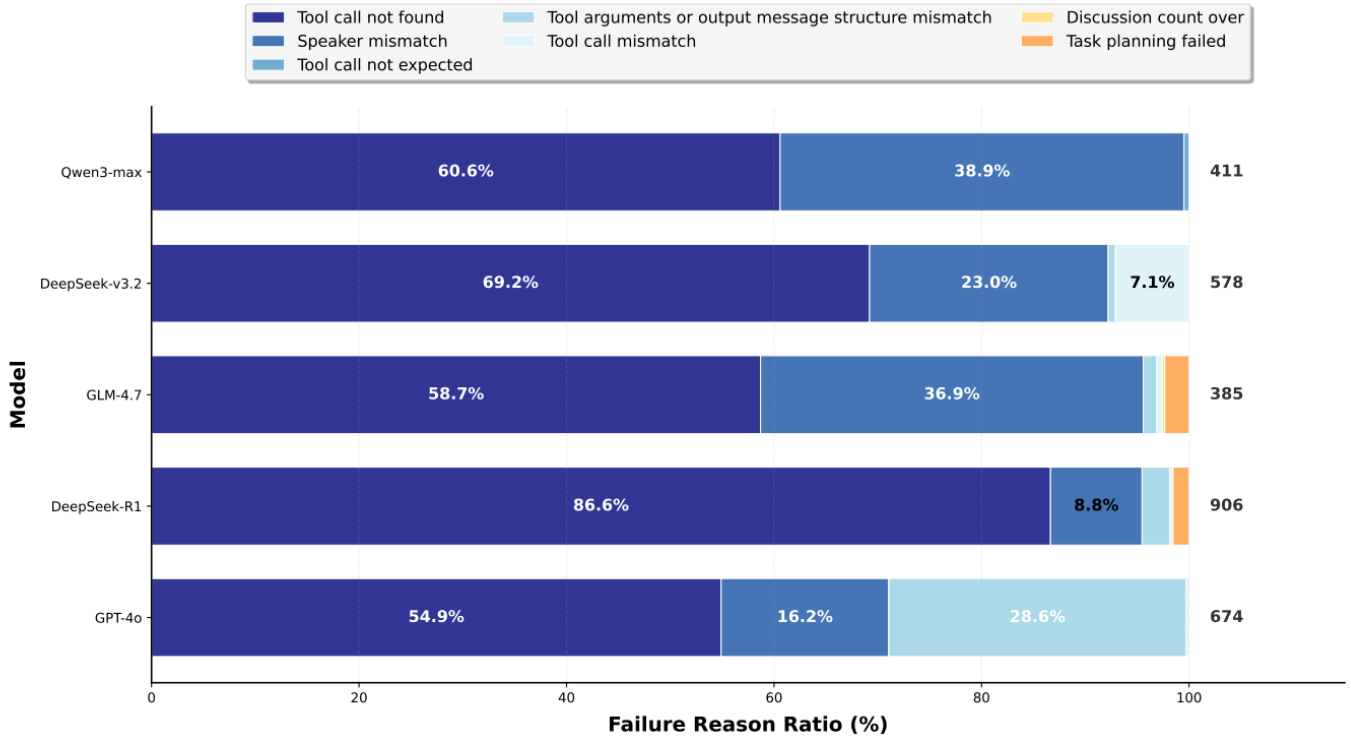
- We validated the whole system by re-measure branching fractions of J/psi decays !
  - Build an inclusive MC sample with all known J/psi decays + new physics signals
  - Use psi(3686)->pi+ pi- J/psi process to measure the BFs, good consistency with injected BFs



Demonstration of DrSai V2

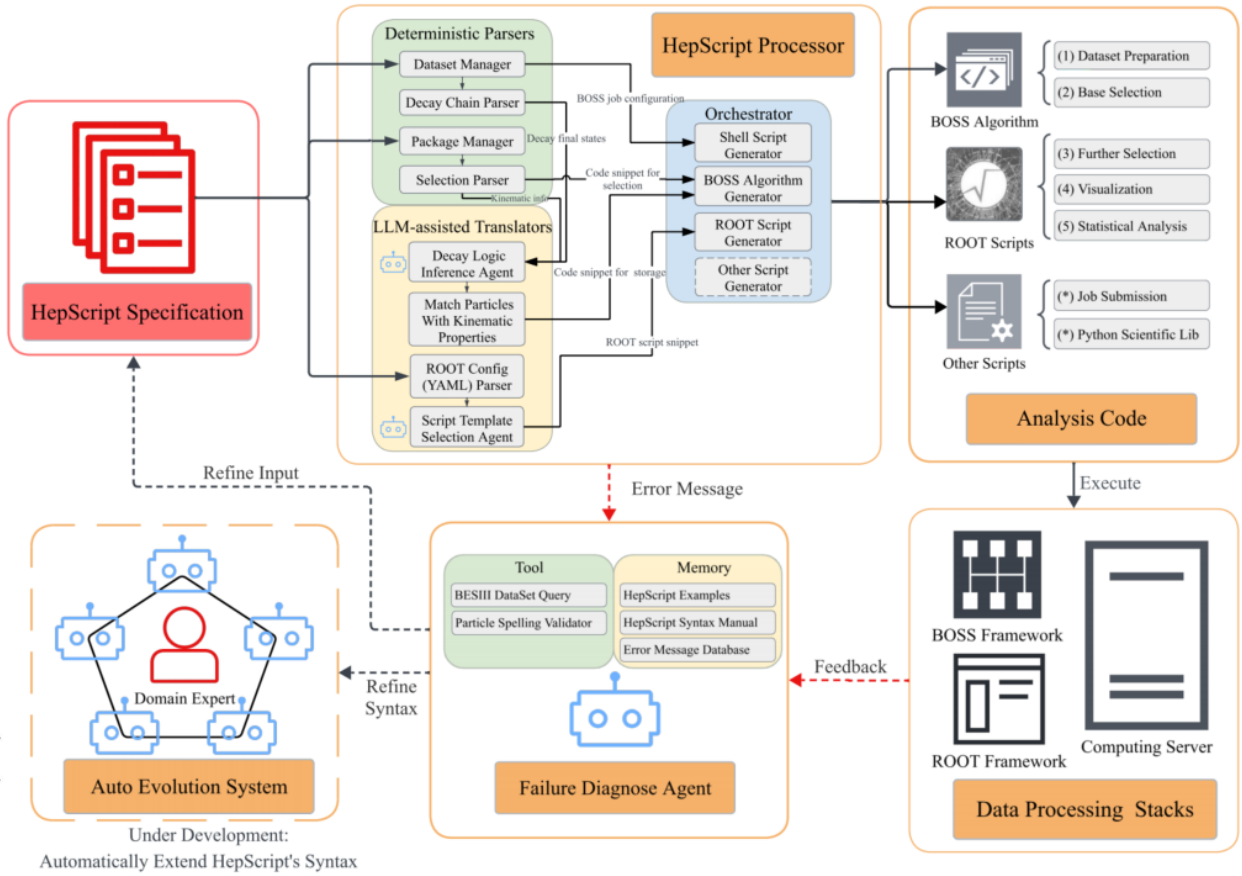
- Performance study and error analysis were conducted using LLMs at 2025
  - Tool call was the bottlenecks
- We are finalizing the DrSai V3 now, and will make a more comprehensive study





DSL (HepScript) for BESIII analysis

- A guide to agents/LLMs
  - Formalizing domain knowledge
  - Integrating with domain software/tools
  - Constraining action space
- Scalable and Automatically
  - Define preliminary DSL by human experts
  - Extract from published paper by LLM
  - Convert to codes/configs by LLM

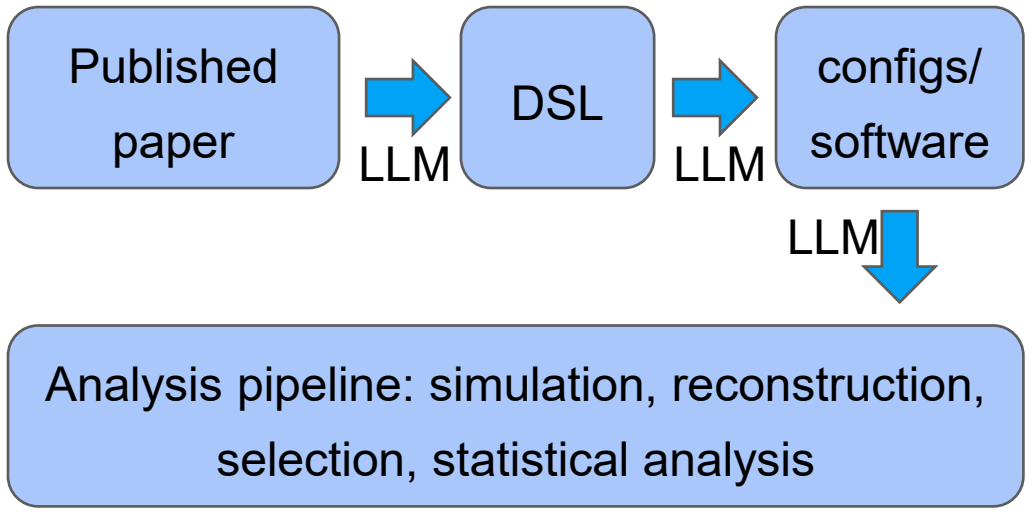


HepScript	LLM	N_S	N_F	SR (%)	One Retry SR (%)	Three Retries SR (%)
Human-written (63 Packages)	deepseek-v3.2	62	1	96.5 ± 2.1	$100^{+0.0}_{-1.5}$	-
	deepseek-r1	56	7	87.7 ± 4.0	$100^{+0.0}_{-1.5}$	-
	gpt-4o	62	1	96.5 ± 2.1	$100^{+0.0}_{-1.5}$	-
	glm-4.7	62	1	96.5 ± 2.1	$100^{+0.0}_{-1.5}$	-
	Qwen3-max	60	3	93.8 ± 3.0	$100^{+0.0}_{-1.5}$	-
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

arXiv:2605.01423

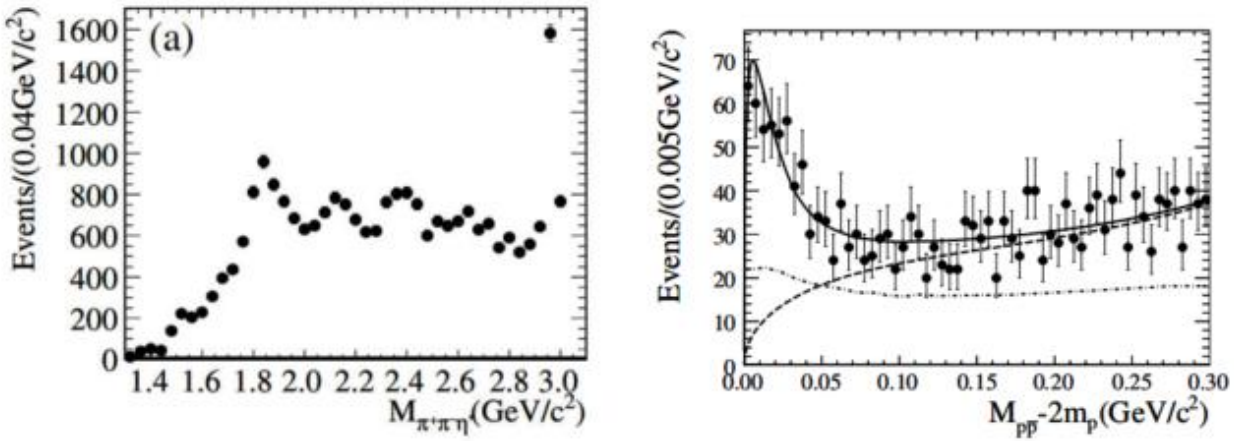
DSL (HepScript) for BESIII analysis

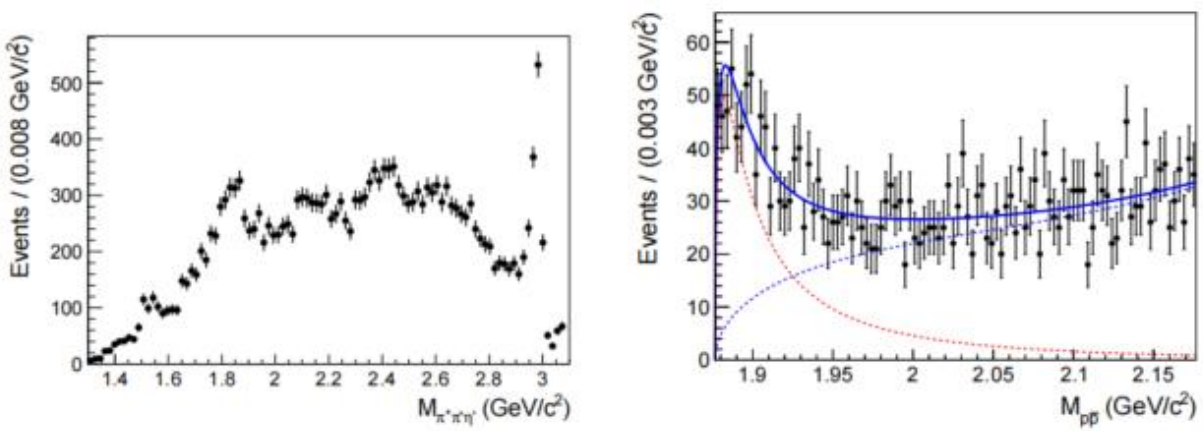
- Necessary for real-world analysis
- A showcase: repeat the previous analysis



- Just integrated to Dr.Sai
  - Segmented into skills descriptions and callable toolsets

From publication

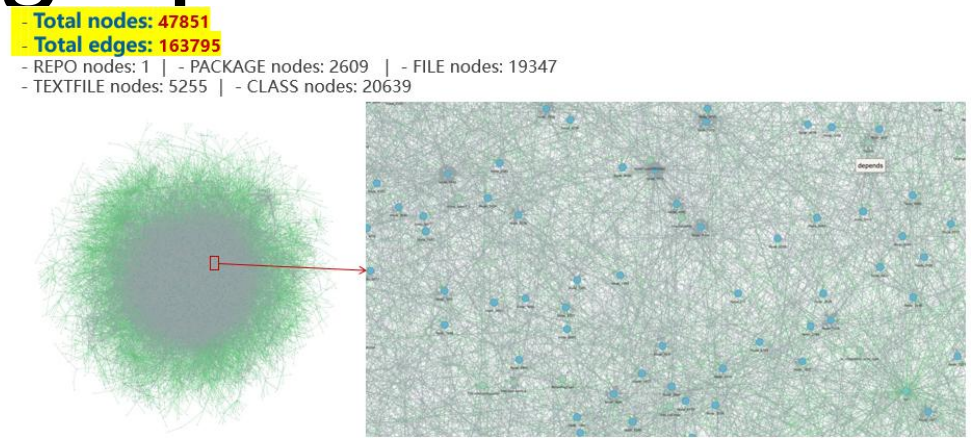


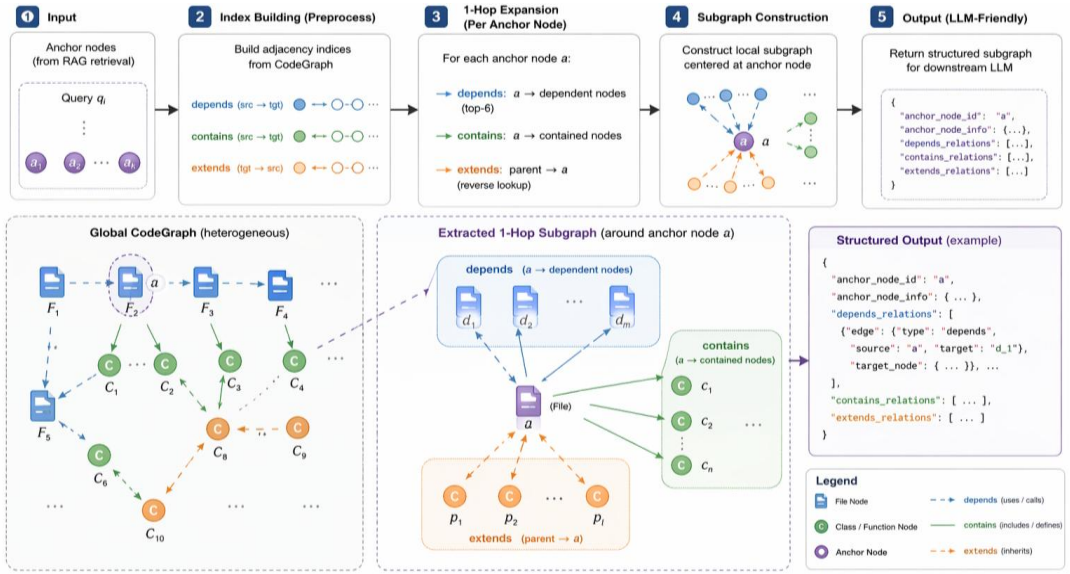


Re-produced using DSL

Code generation using graphRAG

- Another main challenge in application of agents/LLM
  - Domain code generation
  - Some softwares are not open-source and complex
- We start with open-source software - ROOT
  - Take each class/package as a node
  - Define 3 types of edge
    - contains, extends, depends
  - Construct a large graphRAG
    - Total nodes: **48813**
    - Total edges: **170619**
  - Obvious improvement on accuracy
- Now using skills + graphRAG

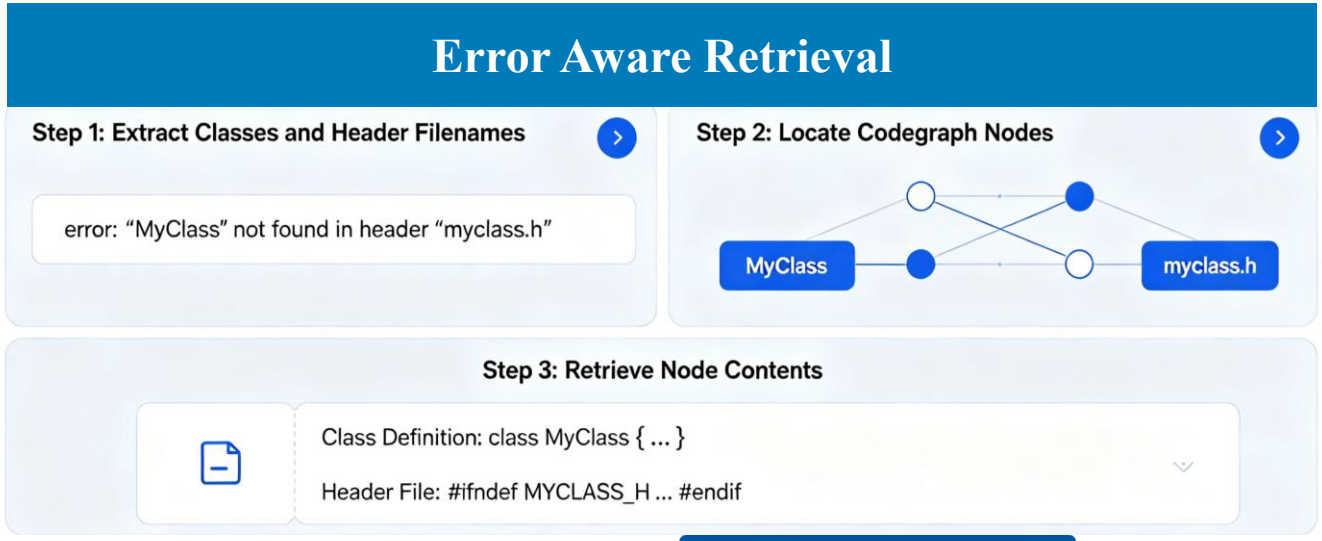


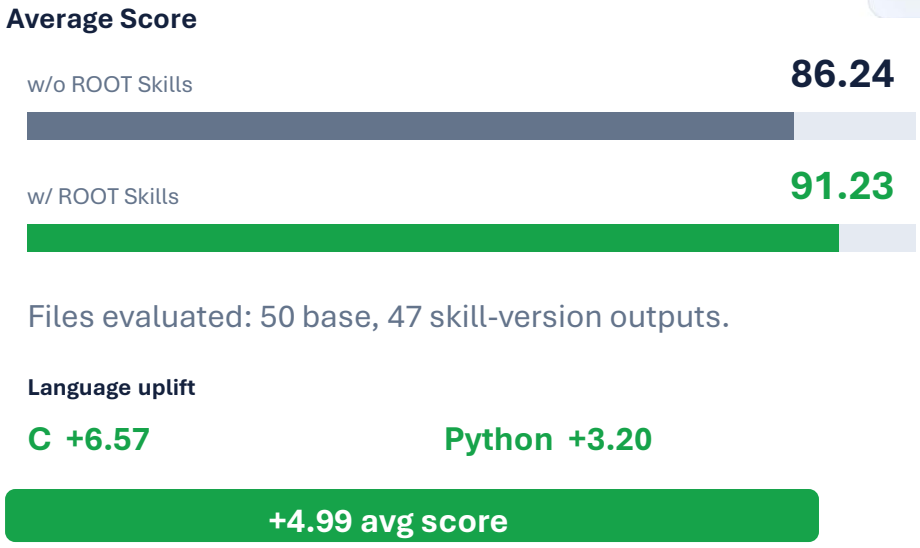


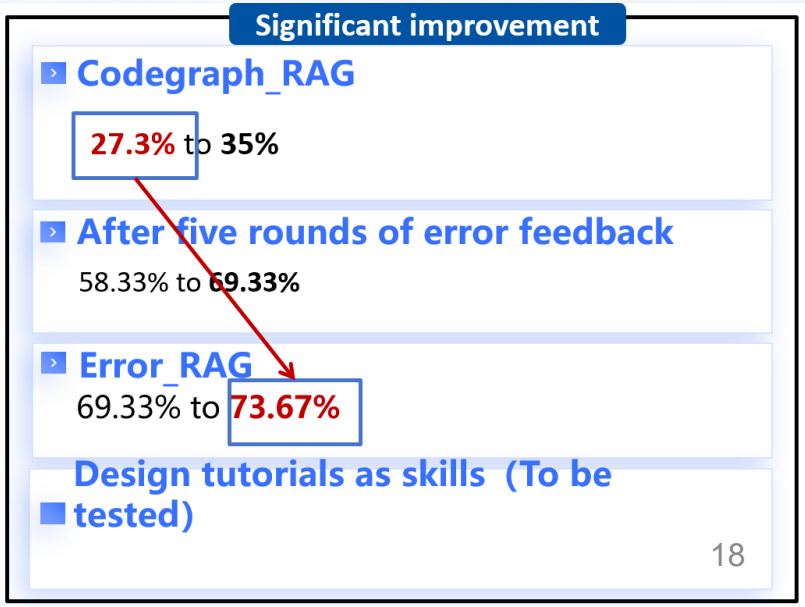
SubGraph extraction

Error-Aware Retrieval and Skills

- Another retrieval by identifying the sources of errors
- Improve the accuracy obviously
- Further improvement by introducing a set of Skills
- Two benchmarks and evaluation methods are provided

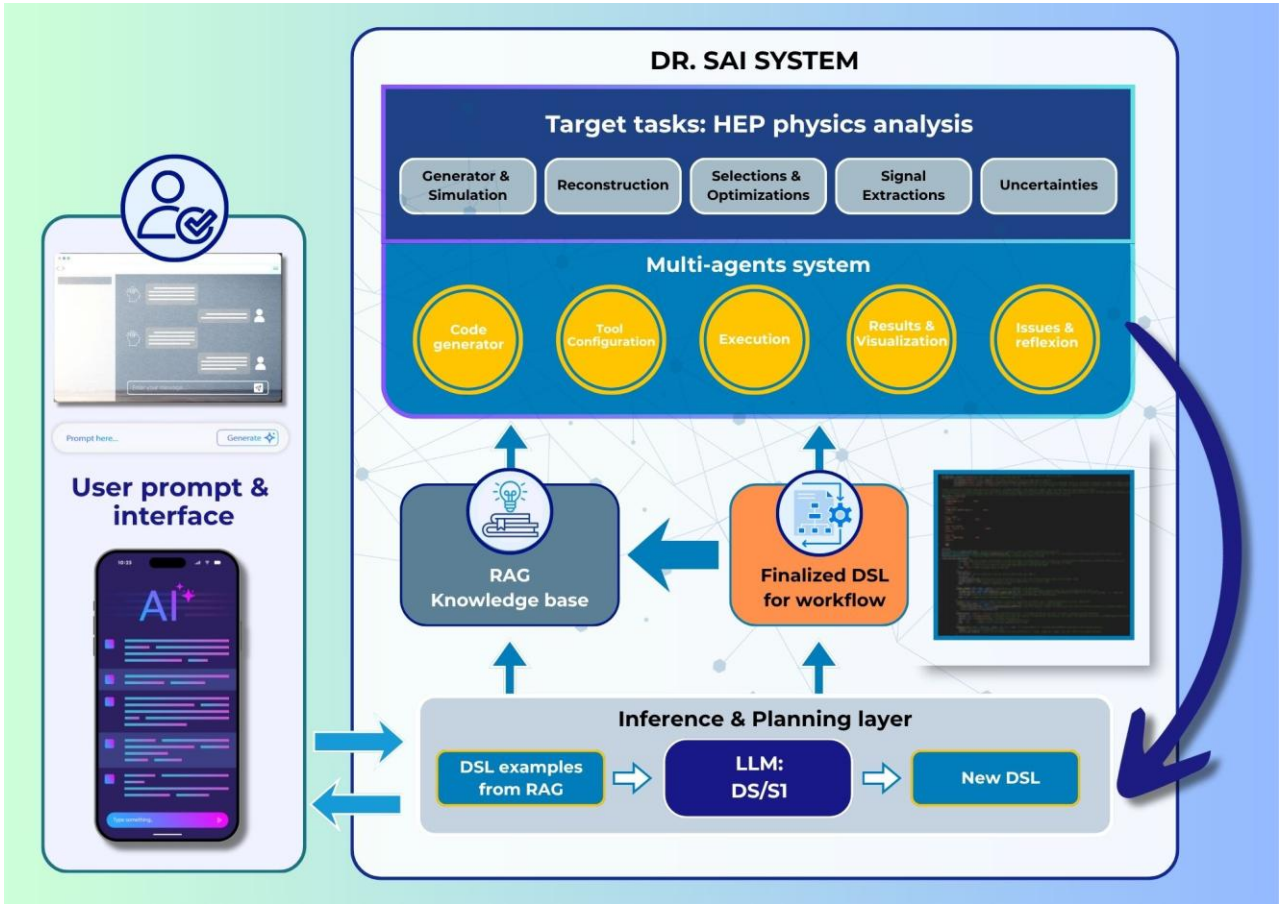






Status

- Physicists interact with Dr.Sai using natural language
- The system autonomously invokes tools/code, and executes simple workflows
  - Only need human to be involved 2-3 times
  - ~100 Yuan tokens for 1 analysis
- Scientific logic and domain expertise are represented through DSL/Skills
  - **Scalable:** Automated pipelines extract from published papers
  - **Continuous Evolution**: revised DSL/Skill based on expert feedback
- **Almost finished, first physics results will be released by the end of 2026 !**



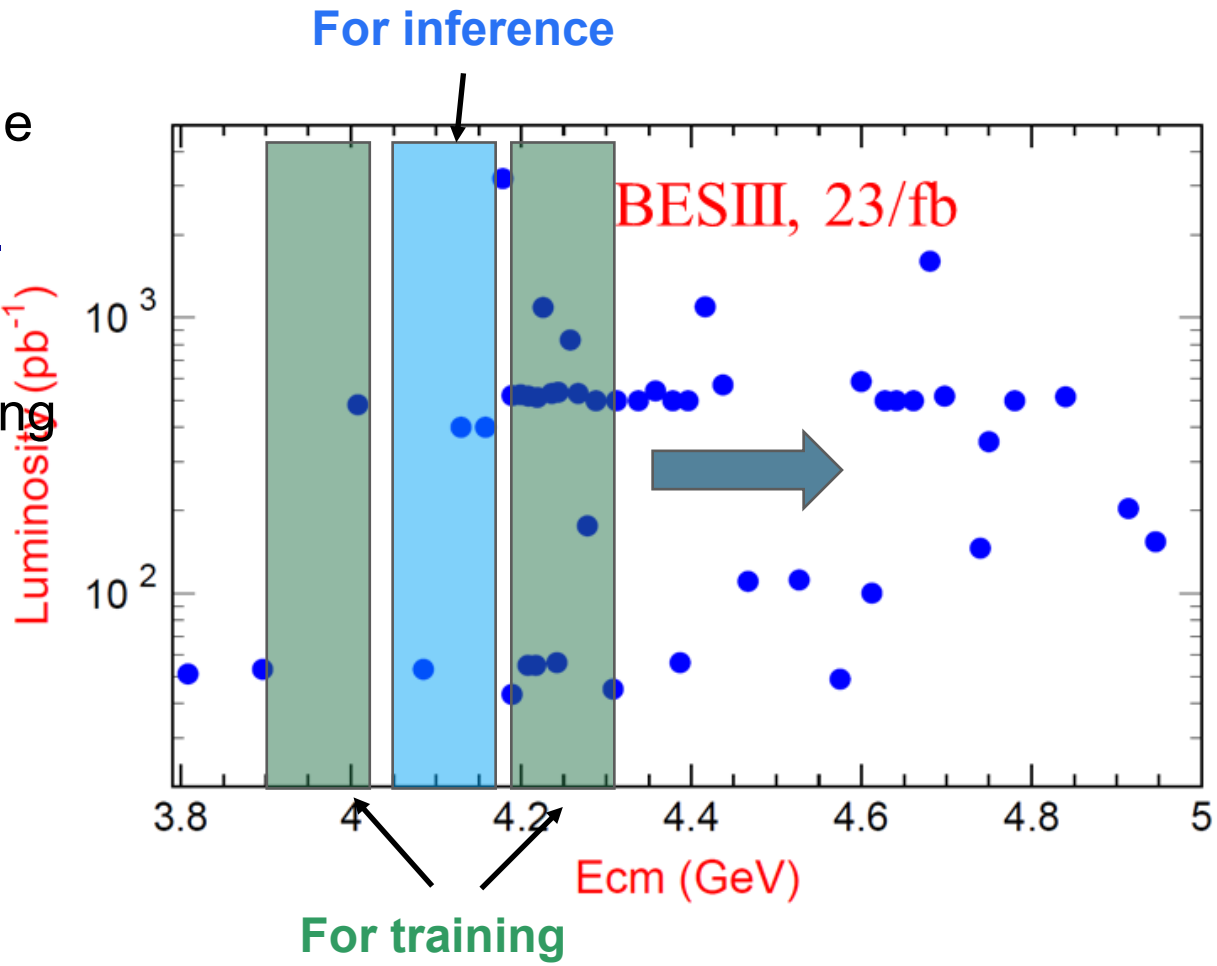
Front. Phys., 2026, 21(9): 096401

What can we do with DrSai

- Systematic search for new particles over all possible processes
- On-going study:
  - Test different framework and LLMs, e.g. Claude code, Tencent/Alibaba cloud
  - Measure the cross section exclusively for light hadron processes under 2 GeV
    - Important input for muon g-2
    - Human expert guide DrSai to complete one measurement, then DrSai transfer the Skills and workflow for other processes
- Please stay tune, we will have the first real experimental measurements done by AI soon.

Anomaly detection

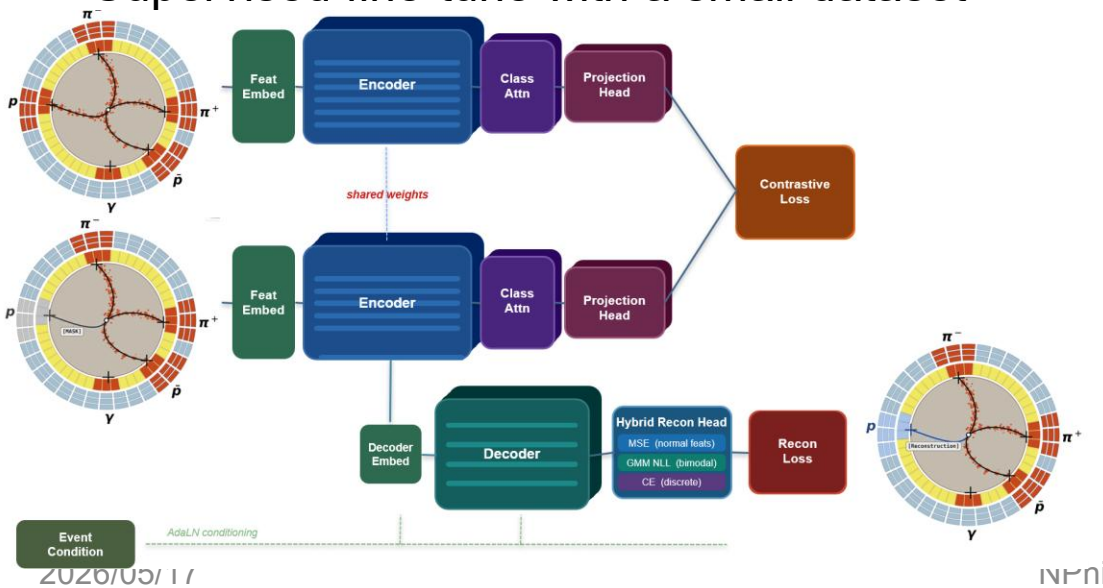
- At BESIII, we can do another type of anomaly detection to search for new particle
- We need an AI model robust against c.m.s. energy !
  - Data representation and transfer learning with Foundation Model (FM)
  - Cross-dataset Anomaly Detection (AD) for new hadron/particle

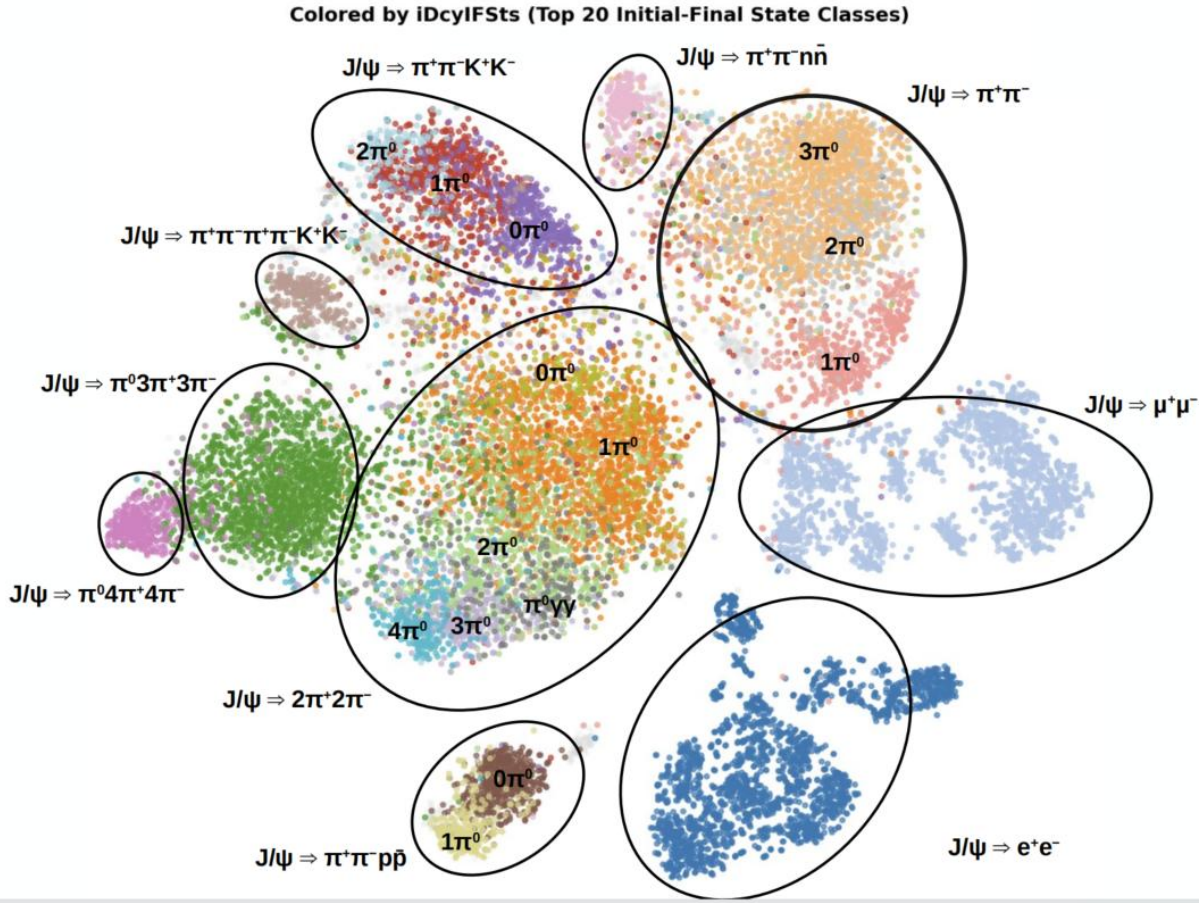


work in progress

Foundation model for analyses

- 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 for analyses

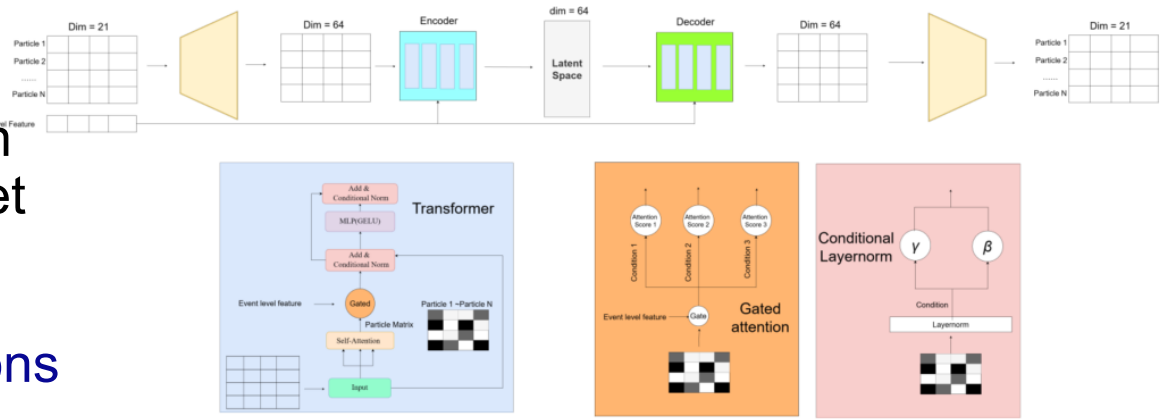
- A showcase: classification for 99 classes
- Better recall when precision are the same w.r.t. traditional method with kinematic constrains
- Next: add momentum conservation into the model for further improvement

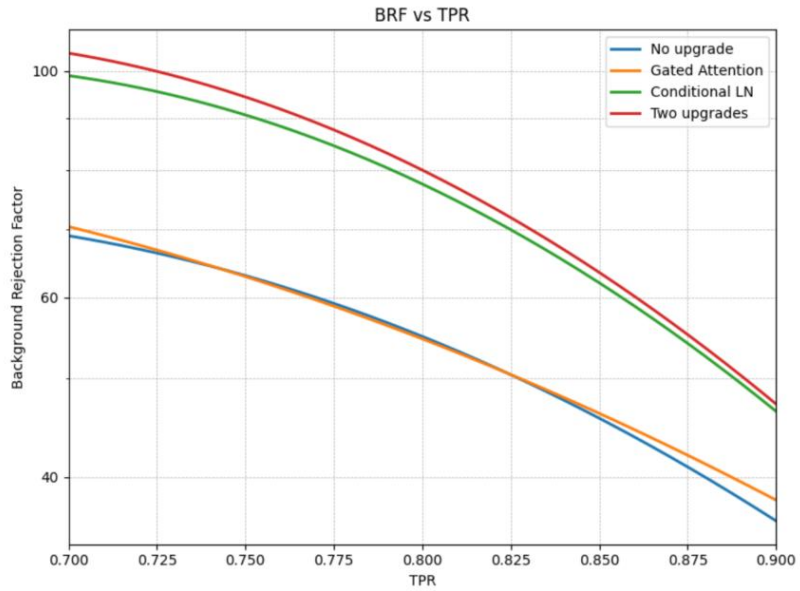
Table 1: Performance comparison between Foundation and Traditional methods

	Foundation			**Traditional**		
	precision	recall	F1-score	precision	recall	F1-score
$\pi^+\pi^-\pi^0$	0.74	**0.76**	0.74	0.74	0.70	0.71
$\pi^+\pi^-\pi^+\pi^-$	0.53	0.42	0.46	0.88	0.60	0.71
$\pi^+\pi^-K^+K^-$	0.78	**0.64**	0.70	0.96	0.50	0.65
$\pi^+\pi^-p\bar{p}$	**0.92**	**0.86**	0.89	0.86	0.52	0.64

Anomaly detection

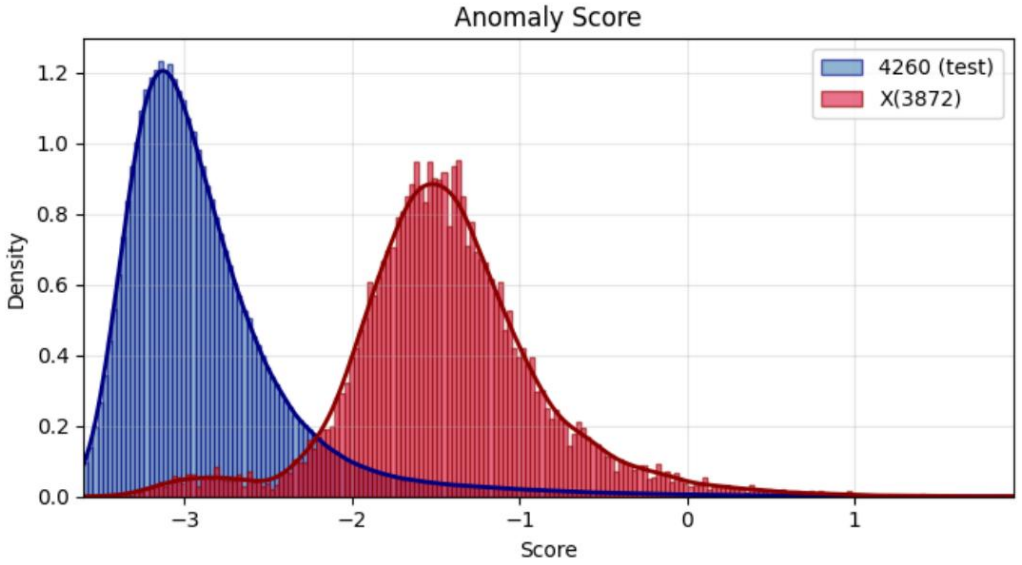
- Motivation:
  - Many datasets at different c.m.s. energy
  - We can use one dataset to train a model then identify the anomaly events in another dataset
- Model: Self-attention + AutoEncoder
  - Event-level features(energy, etc.) as conditions
- Methodology and Evaluation:
  - training sample: fake data + inclusive MC at 4.23 GeV + inclusive MC at 4.26 GeV
  - Inference on fake data + signal MC at 4.26 GeV
  - Preliminary study shows good robustness against c.m.s. energy !
    - Obvious improvement by introducing gated attention and conditional layernorm

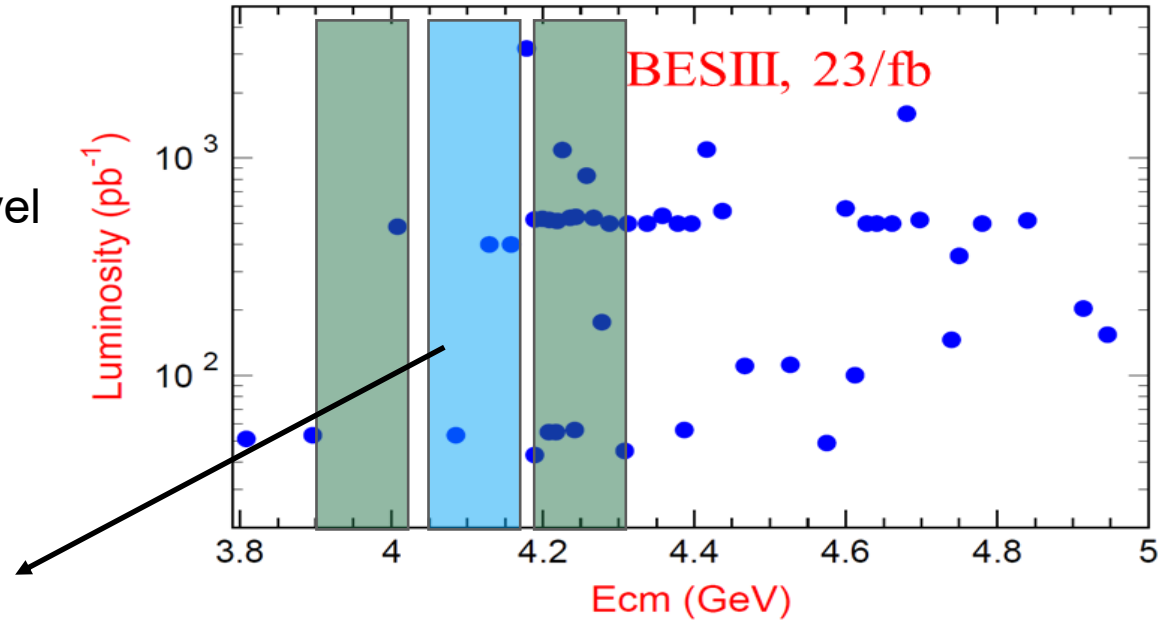




What can we do with FM+AD

- Global search for new particle at BESIII
  - Similar to direct comparison at dataset level
  - Sensitive to new hadrons with different mass/widths
- Next: Combine FM and AD





Why we need it?

- It trained on real data without further selections
  - Expected to be sensitive to most of the new particles
  - We can release the ML model weights to public, theorists can test NP model easily

Summary and outlook

- Possible two new approaches to search for new physics at lepton collider
  - Unique scenario w.r.t. hadron collider
  - Systematic scan of all possible process
    - Autonomous analysis using DrSai - analyzing real data now
  - Global search using cross dataset anomaly detection - finalizing prototype
    - A comprehensive representation of event from FM
    - Cross-dataset anomaly detection with robustness against c.m.s. energy
- Several studies are in the internal review, soon to be released
- 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.

We are organizing the 3rd Quantum computing and Machine Learning Workshop,

Please save the date: Aug 16-22

Sun Yat-sen University (Shenzhen).

There will be a Winter School on AI at end of 2026.

backup