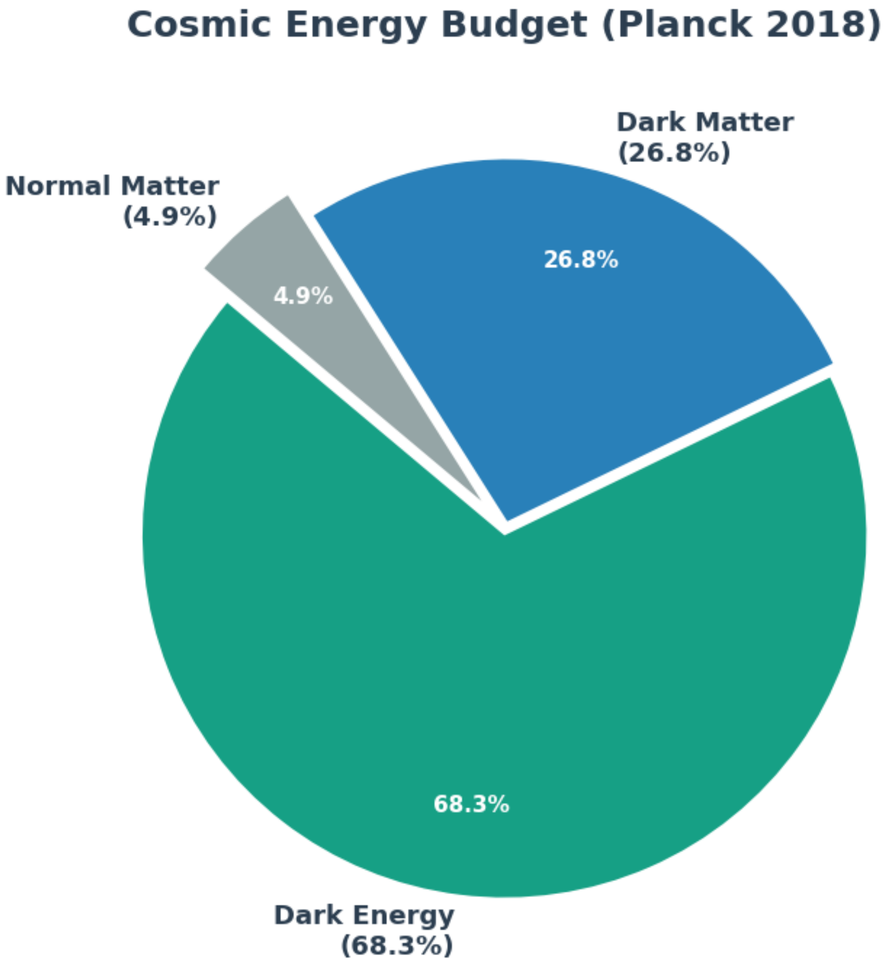
TIDMAD: A Sandbox for Physic-informed LLM Agents in Dark Matter Searches

Yue Ma, Ph.D.
Rare AI Lab (PI: Prof. Aobo Li), Halicioglu Data Science Institute
UC San Diego

03/16/2026@APS Global Physics Summit

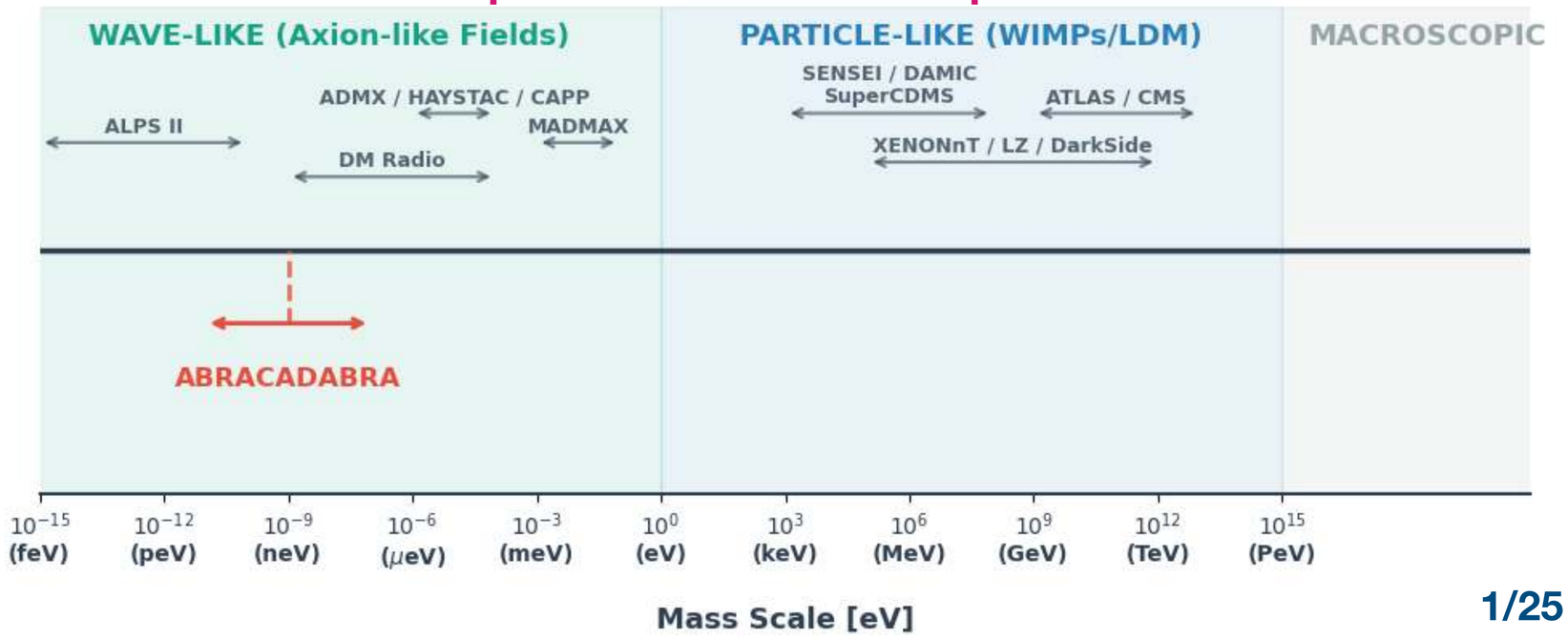
Dark Matter: the Multi-scale Mystery



Evidence for Dark Matter:

- **Rotation Curves** (Galactic scale)
- **Bullet Cluster** (Cluster scale)
- **CMB Anisotropies** (Cosmological scale)
- **Structure Formation** (Large scale)

Experimental Landscape



Axion Search: The ABRACADABRA Frontier

ABRACADABRA

Axion-Modified Maxwell's Equation

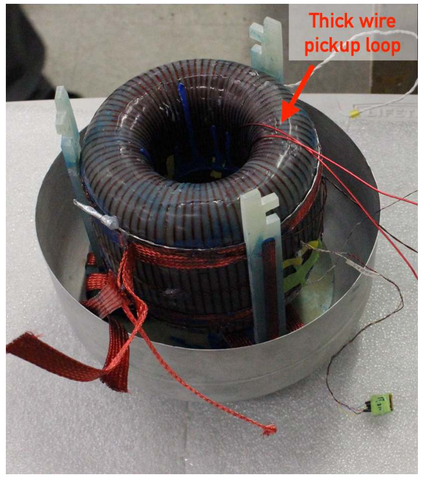
at $L \ll \lambda_a$ **:**

$$\nabla \times \mathbf{B} \approx \mathbf{J}_{eff}$$

$$J_{eff} = g_{a\gamma\gamma}\sqrt{2\rho_{DM}}cos(m_a t)B$$

Axion Search: The ABRACADABRA Frontier

ABRACADABRA



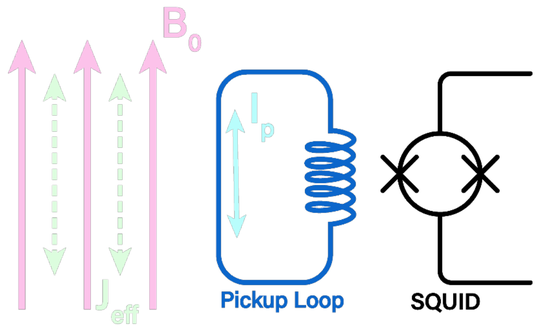
Axion-Modified Maxwell's Equation

at $L \ll \lambda_a$:

$$\nabla \times \mathbf{B} \approx \mathbf{J}_{eff}$$

$$J_{eff} = g_{a\gamma\gamma}\sqrt{2\rho_{DM}}cos(m_a t)B$$

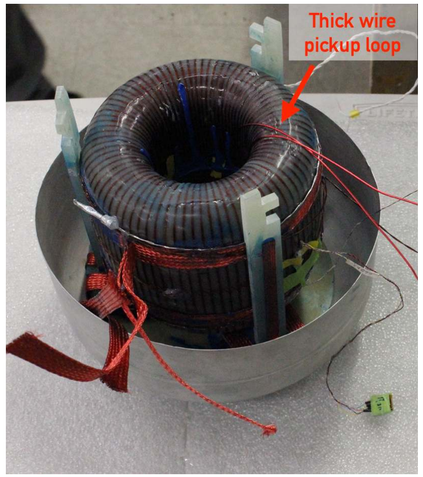
Magnetic Flux created by the effective current



Experimental Apparatus Constructed by Winslow Lab at MIT

Axion Search: The ABRACADABRA Frontier

ABRACADABRA



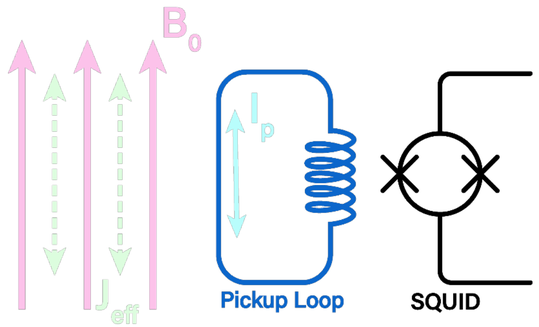
Axion-Modified Maxwell's Equation

at $L \ll \lambda_a$:

$$\nabla \times \mathbf{B} \approx \mathbf{J}_{eff}$$

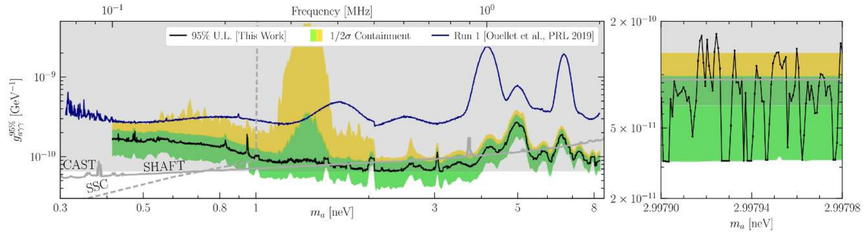
$$J_{eff} = g_{a\gamma\gamma}\sqrt{2\rho_{DM}}cos(m_a t)B$$

Magnetic Flux created by the effective current



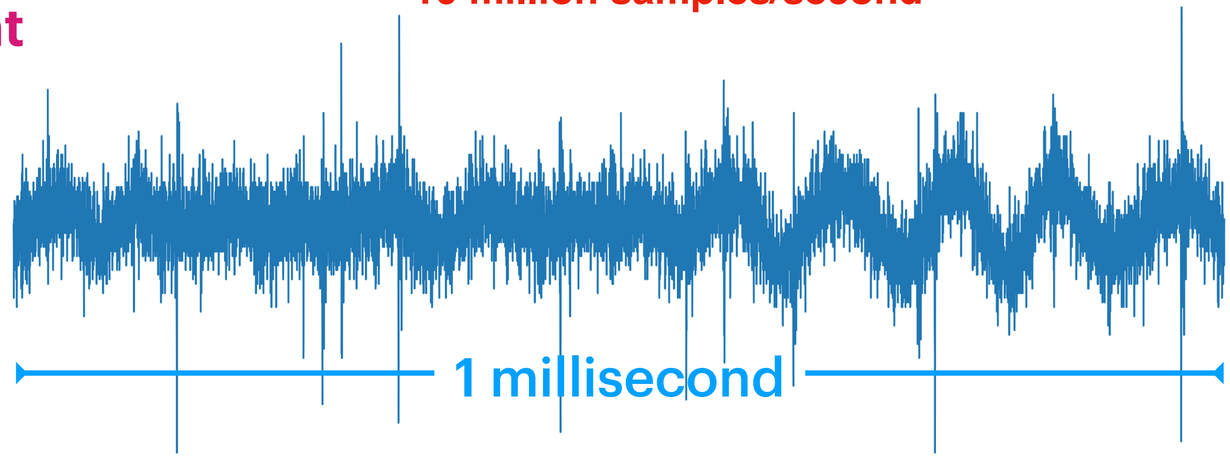
Frequency Spectrum

Broadband search for axion DM



Ultra-long Time Series

10 million samples/second



Experimental Apparatus Constructed by Winslow Lab at MIT

TIDMAD: Bridging AI and Axion Physics

ABRACADABRA

TIDMAD: Time Series Dataset for Discovering Dark Matter with AI Denoising

J. T. Fry[1] * jtfry@mit.edu
Aobo Li[2] liaobo77@ucsd.edu
Lindley Winslow[1] lwinslow@mit.edu
Xinyi Hope Fu[1] hopefu@mit.edu
Zhenghao Fu[1] fuzh@mit.edu
Kaliroe M. W. Pappas[1] kaliroe@mit.edu

[1]Department of Physics, Massachusetts Institute of Technology, Cambridge, MA 02139, USA
[2]Halıcıoğlu Data Science Institute, Department of Physics, UC San Diego, La Jolla, CA 92093, USA

J. T. Fry* , X. H. Fu, Z. Fu, K. M. W. Pappas, L. Winslow, A. Li*
arXiv:2406.04378
NeurIPS 2025 D&B Track Spotlight

Open Data

Release dark matter detector data for AI/ML algorithms

Easy Benchmarking

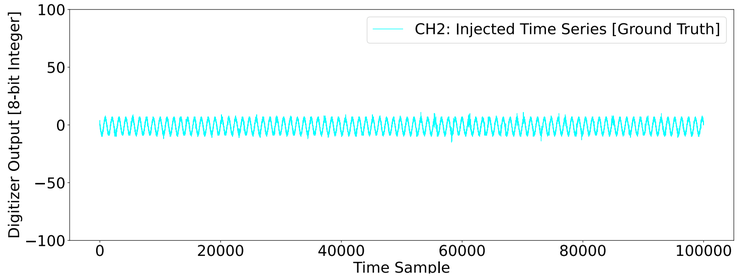
Design a quantitative benchmarking score to quantify the performance of different algorithms

AI for Science

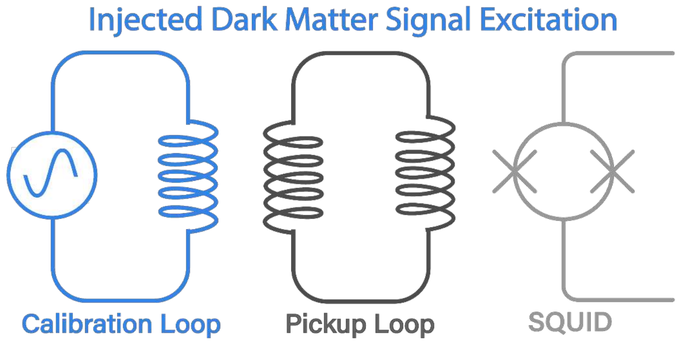
Enables core AI algorithms to extract the signal and produce real physics results thereby advancing fundamental science

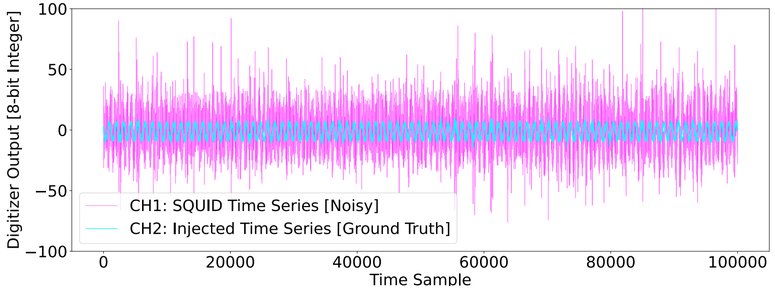
TIDMAD: The Principle

ABRACADABRA



CH2: Injected Time Series [Ground Truth]



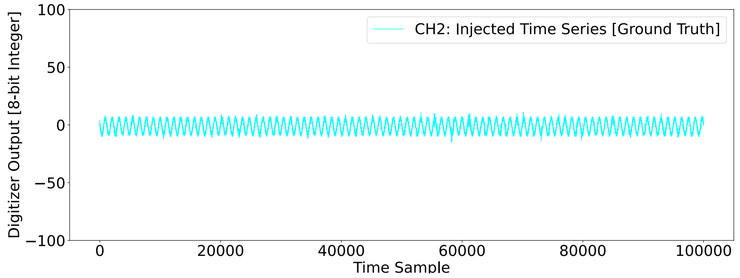


CH1: SQUID Time Series [Noisy]

Train AI denoising model to recover...

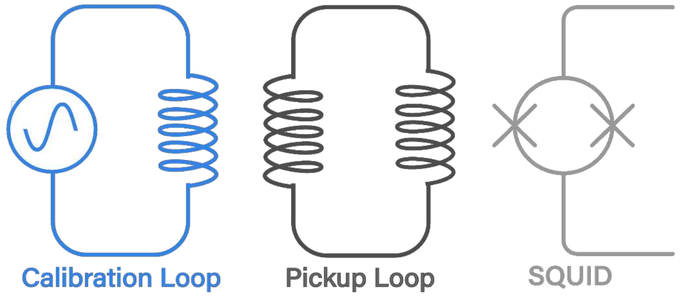
TIDMAD: The Principle

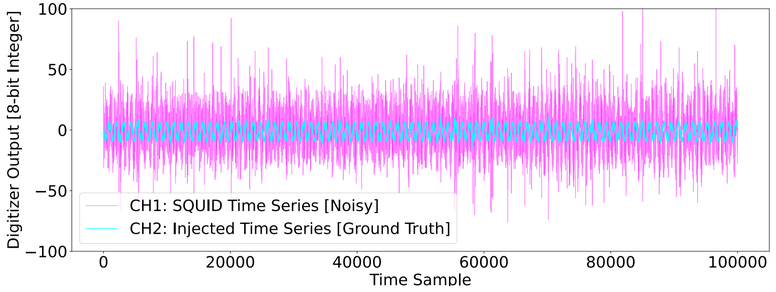
ABRACADABRA



CH2: Injected Time Series [Ground Truth]

Injected Dark Matter Signal Excitation



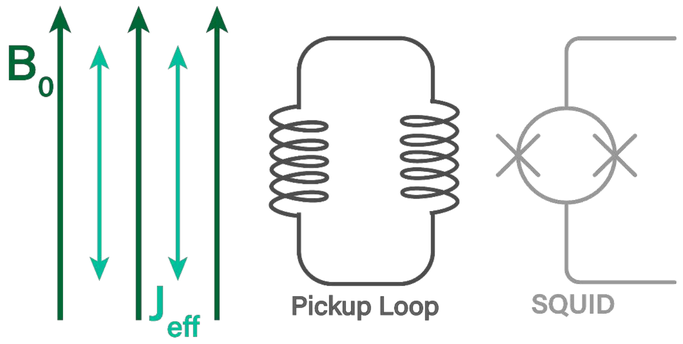


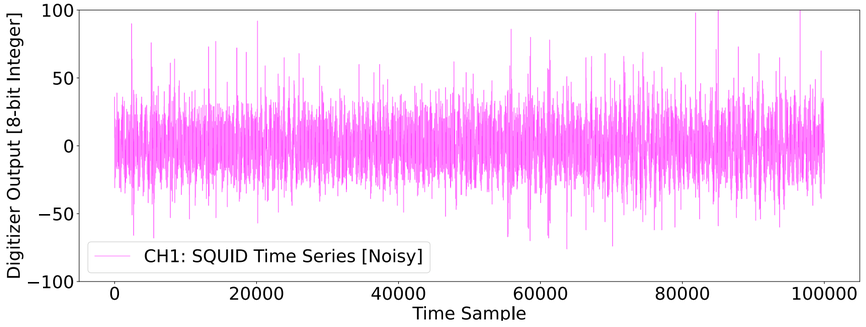
CH1: SQUID Time Series [Noisy]

Train AI denoising model to recover...

No Signal Injected

Real Dark Matter Signal Excitation



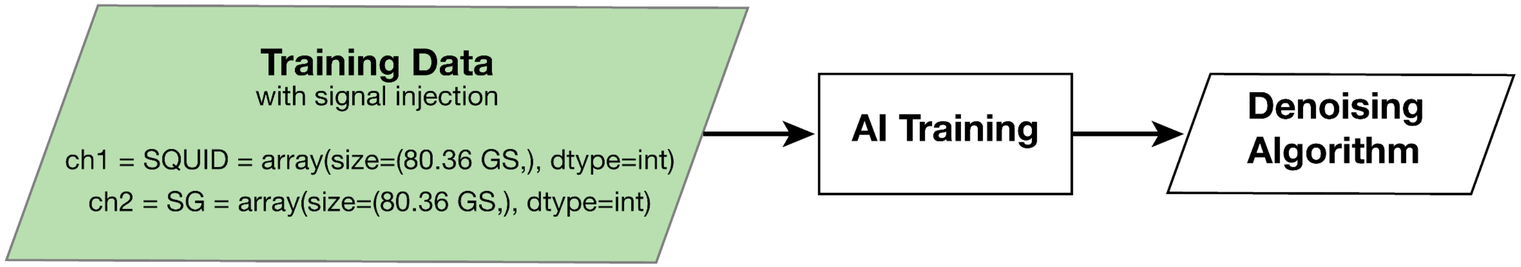


CH1: SQUID Time Series [Noisy]

Use trained AI model to denoise...

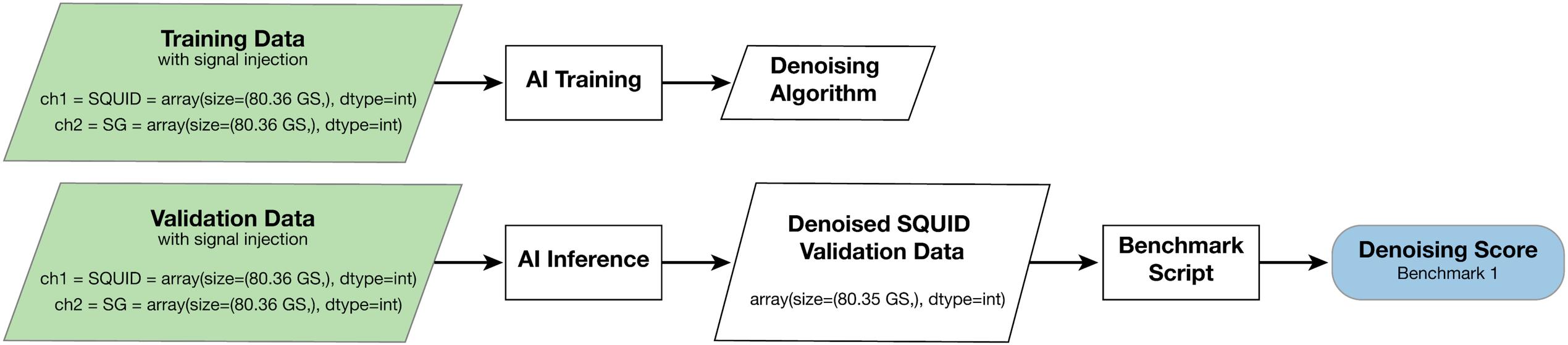
TIDMAD: Data Structure and Workflow

ABRACADABRA



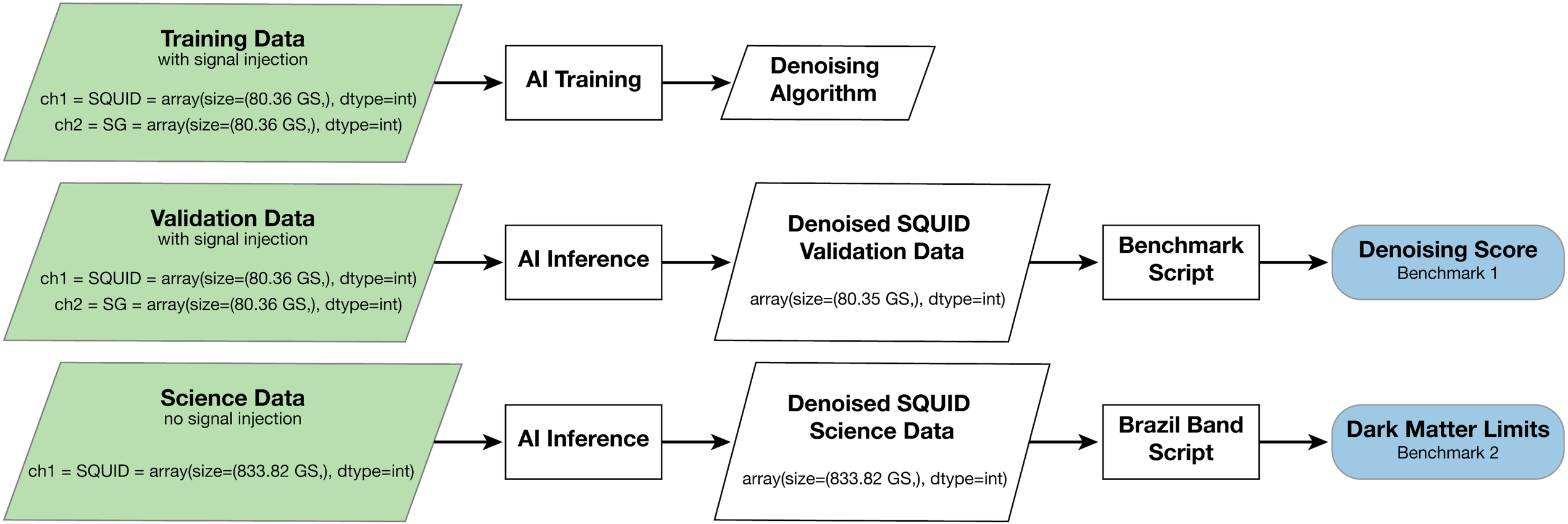
TIDMAD: Data Structure and Workflow

ABRACADABRA



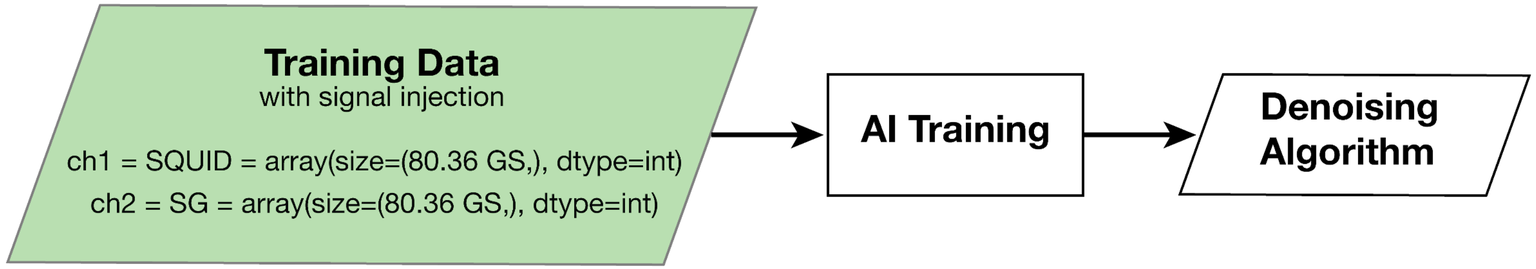
TIDMAD: Data Structure and Workflow

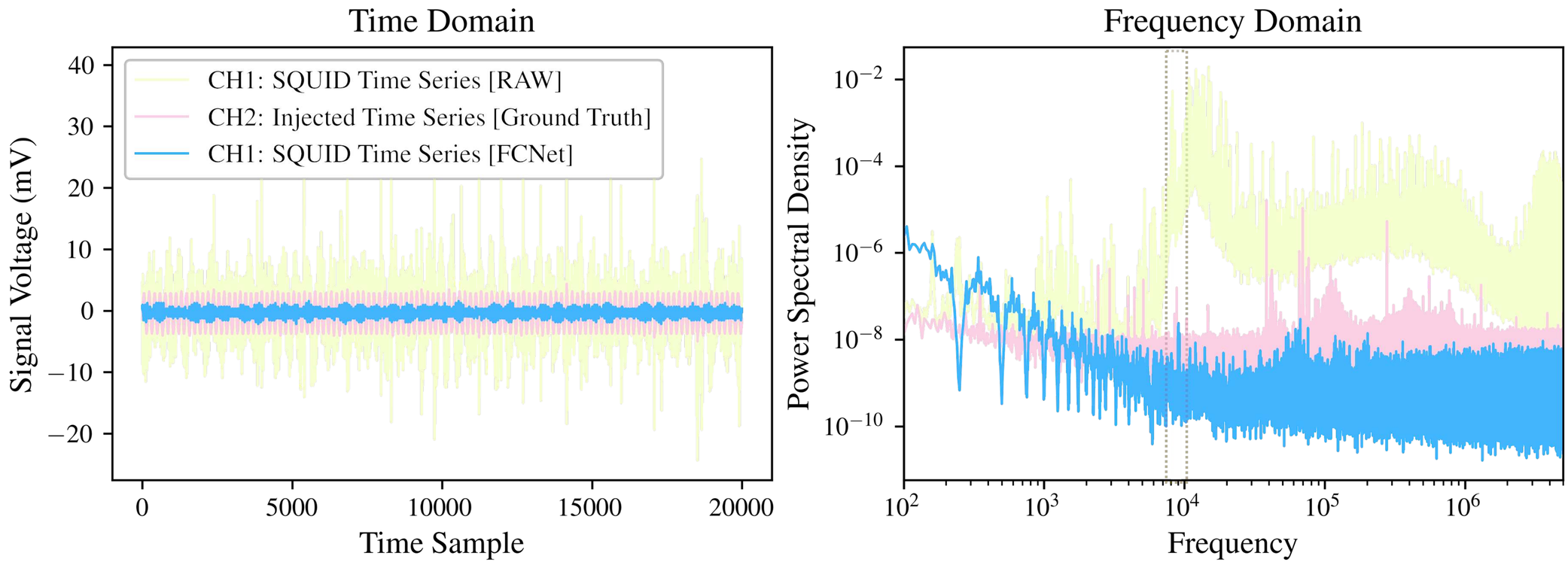
ABRACADABRA



TIDMAD: Example of Training Data

ABRACADABRA





TIDMAD: Benchmark 1: Denoising Score

ABRACADABRA

Validation Data
with signal injection

ch1 = SQUID = array(size=(80.36 GS,), dtype=int)
ch2 = SG = array(size=(80.36 GS,), dtype=int)

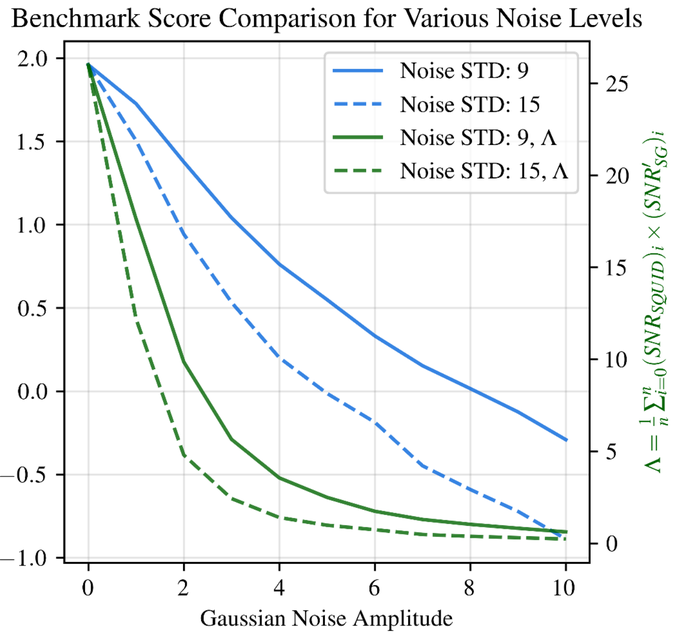
→ **AI Inference** →

Denoised SQUID Validation Data

array(size=(80.35 GS,), dtype=int)

→ **Benchmark Script** →

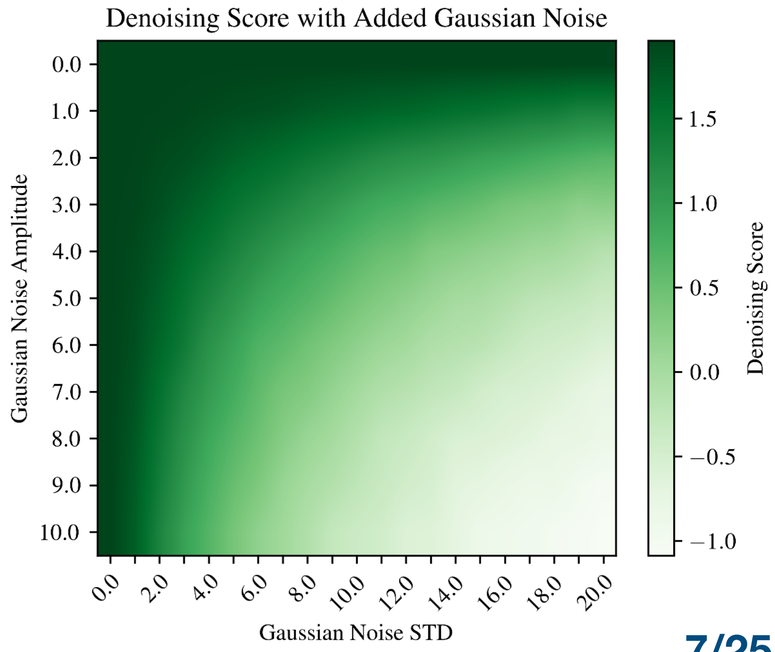
Denoising Score
Benchmark 1



$$\Lambda = \left(\frac{1}{n} \sum_{i=0}^{n} (SNR_{\text{SQUID}})_i \times (SNR'_{\text{Injected}})_i \right)$$

Denoising Score $= log_{5.27}\Lambda$

Test the denoising score by doping gaussian noise into Time Series



TIDMAD: Benchmark 1: Denoising Score

ABRACADABRA

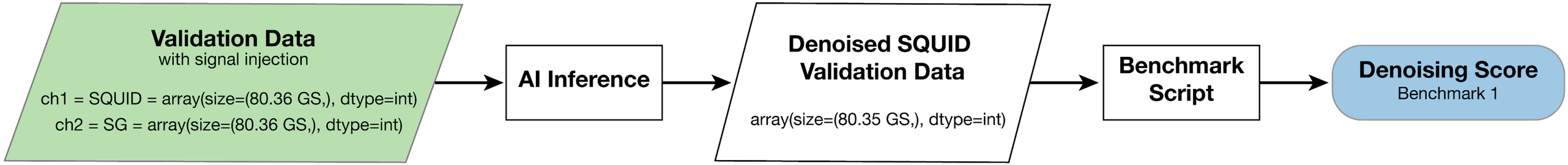
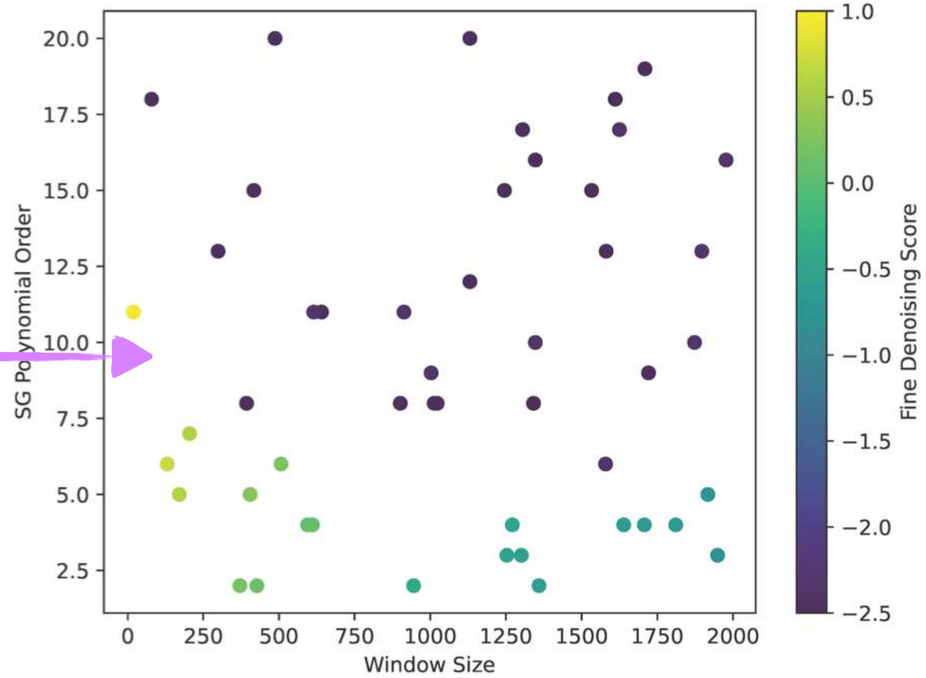


Table 1: Fine and coarse denoising score for raw data, traditional algorithms, and trained ML models. FS means frequency splitting, or training multiple versions of the same model to handle different frequency ranges. The detail of segment size and FS is discussed in Section B.

Algorithms	Segment Size	FS	Parameters	Fine Score	Coarse Score
None				1.00	1.10
Fourier Averaging	1×10^8	–	10-fold Average	0.24	0.26
Moving Average	1×10^6	–	window = 100	0.86	0.95
SG Filter	1×10^6	–	window = 19, order = 11	0.95	1.04
FC Net	4×10^4	Yes	See Appendix B	6.43	6.55
PU Net	4×10^4	Yes	See Appendix B	3.69	3.84
Transformer	2×10^4	Yes	See Appendix B	3.95	4.18
WaveNet	4×10^4	No	See Appendix B	4.99	5.16
RNN Seq2Seq	4×10^4	Yes	See Appendix B	3.38	3.79

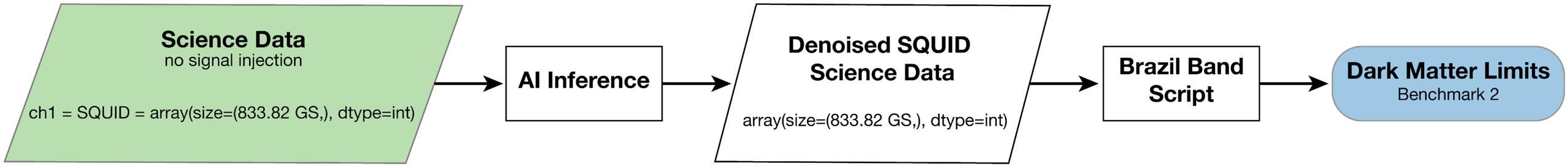
Traditional Denoising

AI Denoising



TIDMAD: Benchmark 2: Dark Matter Limits

ABRACADABRA



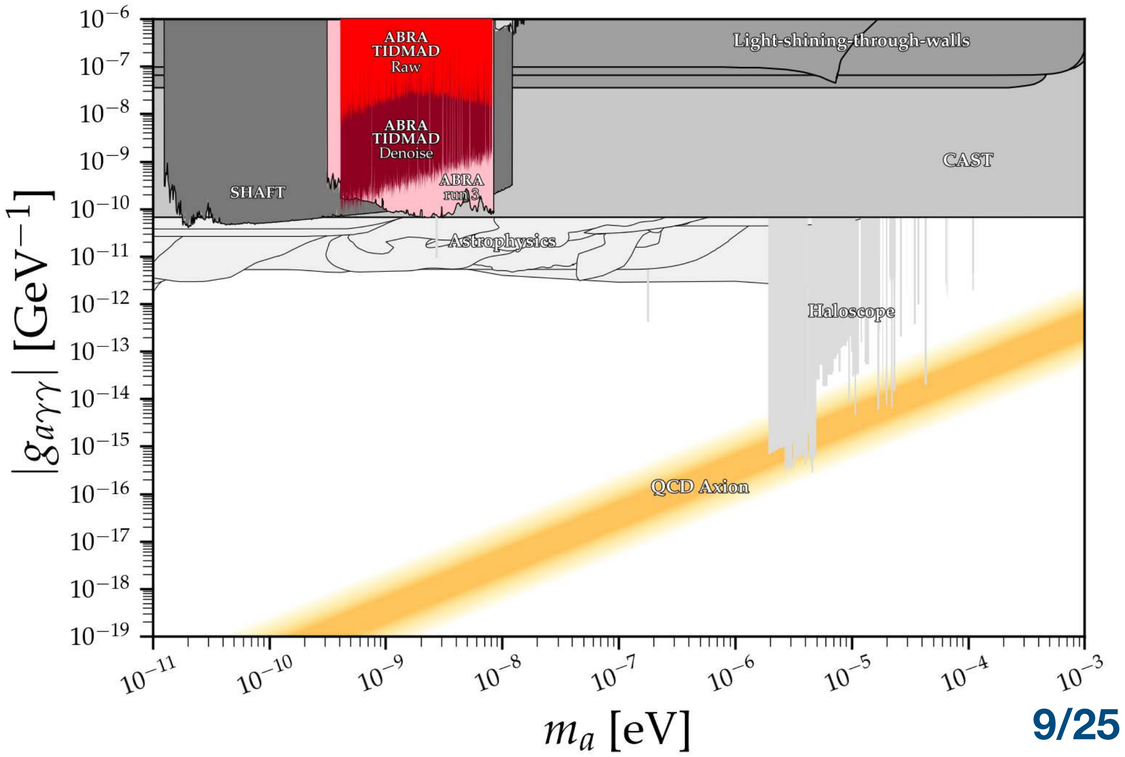
Axion Limit Boost

ABRA TIDMAD Raw: 24 hr data, no denoising

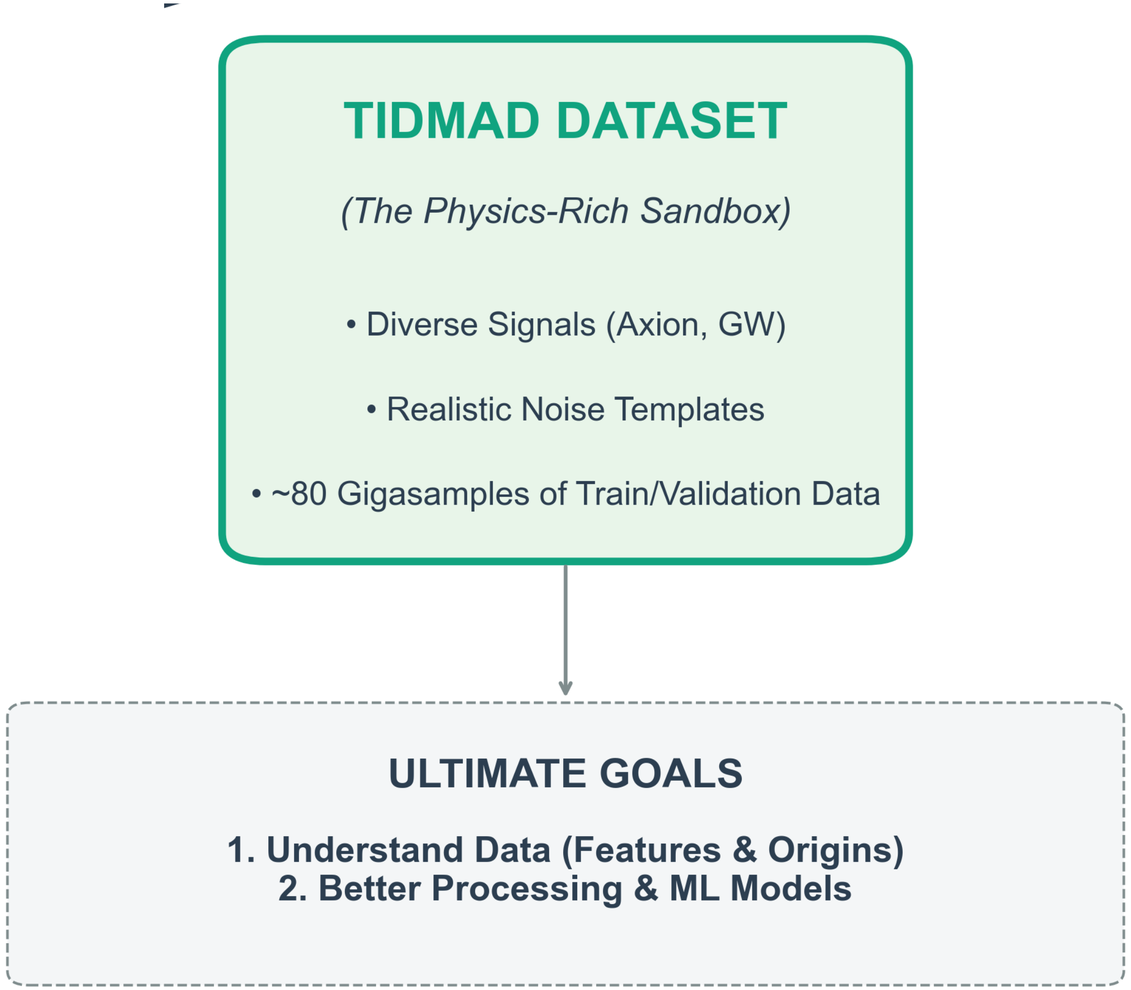
ABRA TIDMAD Denoised: 24 hr data with FCNet denoising

ABRA Run 3: 2,400 hr data, no denoising

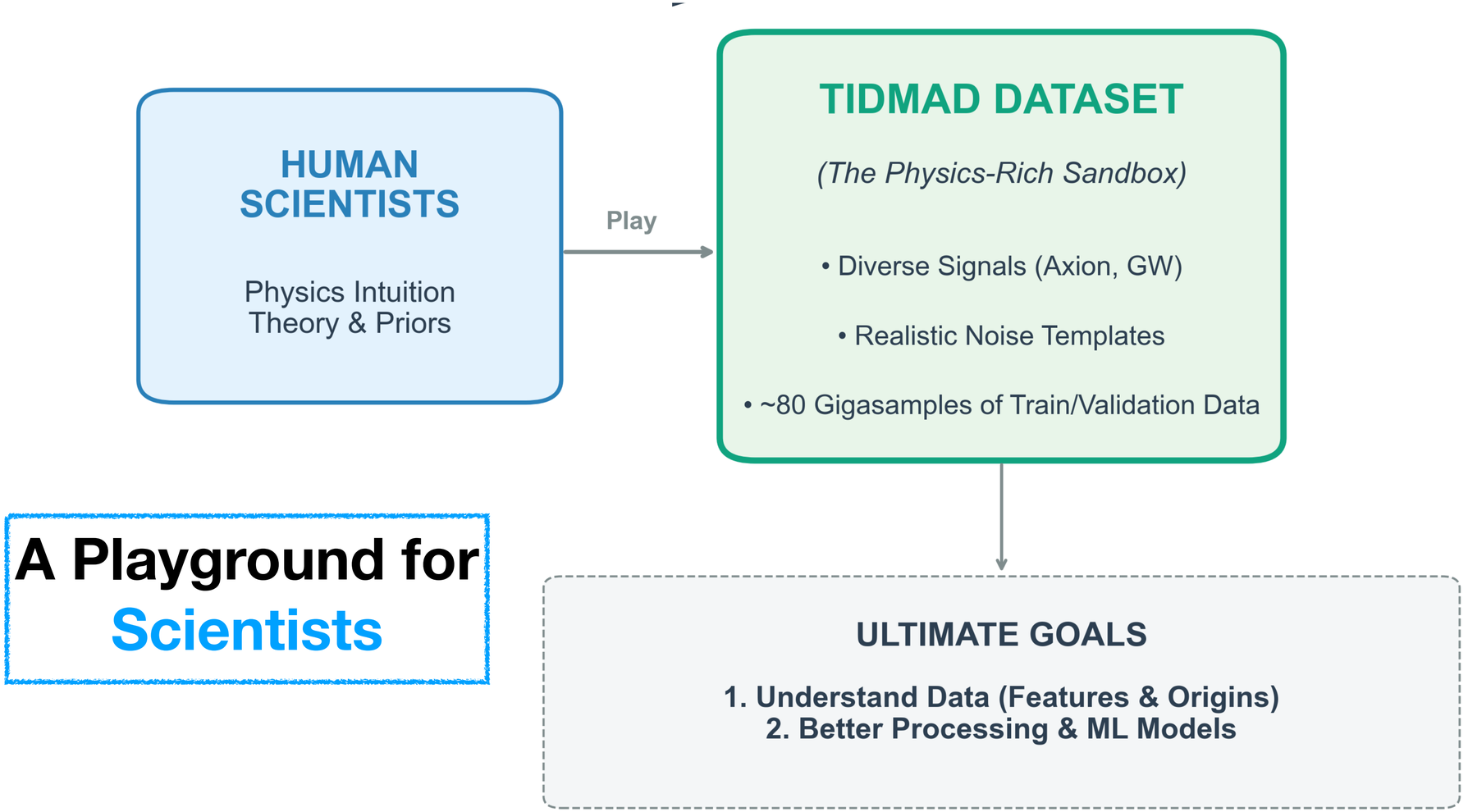
Efficient denoising algorithms increased Axion search limit by **1-2 orders of magnitude**, approaching the previous world-leading ABRA run 3 results with only **1%** of statistics



TIDMAD: From a Static Benchmark to a Physics-Rich Sandbox

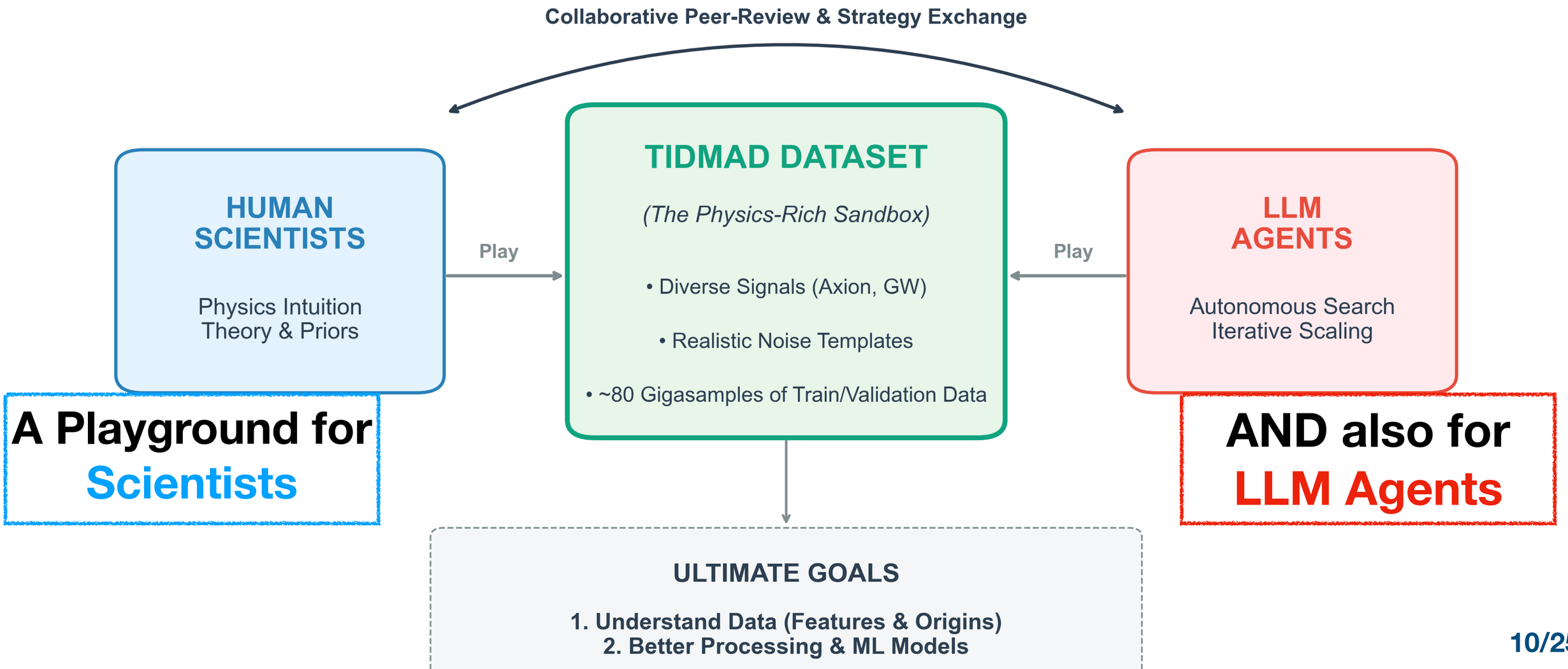


TIDMAD: From a Static Benchmark to a Physics-Rich Sandbox



A Playground for Scientists

TIDMAD: From a Static Benchmark to a Physics-Rich Sandbox



What LLM Agents Can Already Do

We are already at a new stage!

General Engineering

- Action-oriented execution
  Direct terminal & file access
  (Claude Code, OpenClaw)

- Repository-scale mastery
  Understanding 100k+ lines

- Autonomous self-healing
  Run → Error → Fix loop

What LLM Agents Can Already Do

We are already at a new stage!

General Engineering

• Action-oriented execution
Direct terminal & file access
(Claude Code, OpenClaw)

• Repository-scale mastery
Understanding 100k+ lines

• Autonomous self-healing
Run → Error → Fix loop

ML Research Automation

• Strategic reasoning
Dual-Memory (Semantic/Episodic)
prevents hallucination & retries

• End-to-end ML pipeline
From raw data to LaTeX paper
(MLZero, AI Scientist v2)

• Recursive refinement
Tree-search for idea discovery

What LLM Agents Can Already Do

We are already at a new stage!

General Engineering

• Action-oriented execution
Direct terminal & file access
(Claude Code, OpenClaw)

• Repository-scale mastery
Understanding 100k+ lines

• Autonomous self-healing
Run → Error → Fix loop

ML Research Automation

• Strategic reasoning
Dual-Memory (Semantic/Episodic)
prevents hallucination & retries

• End-to-end ML pipeline
From raw data to LaTeX paper
(MLZero, AI Scientist v2)

• Recursive refinement
Tree-search for idea discovery

Specialized Domains

• High-efficiency discovery
Years of R&D → Days
(GNoME, ChemCrow)

• Targeted tasks
Crystal search, Bio-correlation

• Deterministic workflows
Integration with specialized
scientific toolstacks

What Are the Current Limitations?

Yes, but...

General Engineering

• High operational cost
Extreme token consumption
during large-scale searches

• Explanability gap
Flexible tree or graph-like structure
But Lack of transparent
and well-defined logic chains

• Expert bottleneck
Requires massive manual 'skills'
Focus on breadth over depth

What Are the Current Limitations?

Yes, but…

General Engineering

• High operational cost
Extreme token consumption
during large-scale searches

• Explanability gap
Flexible tree or graph-like structure
But Lack of transparent
and well-defined logic chains

• Expert bottleneck
Requires massive manual 'skills'
Focus on breadth over depth

ML Research Automation

• Shallow task depth
Focus on metric optimization
rather than system discovery

• Structural data blindspot
Inadequate in 'deconstructing'
complex physics features

• Purely statistical logic
Treats physics as simple noise

What Are the Current Limitations?

Yes, but…

General Engineering

• High operational cost
Extreme token consumption
during large-scale searches

• Explanability gap
Flexible tree or graph-like structure
But Lack of transparent
and well-defined logic chains

• Expert bottleneck
Requires massive manual 'skills'
Focus on breadth over depth

ML Research Automation

• Shallow task depth
Focus on metric optimization
rather than system discovery

• Structural data blindspot
Inadequate in 'deconstructing'
complex physics features

• Purely statistical logic
Treats physics as simple noise

Specialized Domains

• Rigid pipelines
Highly domain-bound architecture
limiting cross-field expansion

• Low discovery depth
Restricted to well-characterized
or fixed-pipeline tasks

• Limited adaptability
to various input formats
-> Struggles with open-ended
environments

What Are the Current Limitations?

Yes, but…

General Engineering

• High operational cost
Extreme token consumption
during large-scale searches

• Explanability gap
Flexible tree or graph-like structure
But Lack of transparent
and well-defined logic chains

• Expert bottleneck
Requires massive manual 'skills'
Focus on breadth over depth

ML Research Automation

• Shallow task depth
Focus on metric optimization
rather than system discovery

• Structural data blindspot
Inadequate in 'deconstructing'
complex physics features

• Purely statistical logic
Treats physics as simple noise

Specialized Domains

• Rigid pipelines
Highly domain-bound architecture
limiting cross-field expansion

• Low discovery depth
Restricted to well-characterized
or fixed-pipeline tasks

• Limited adaptability
to various input formats
-> Struggles with open-ended
environments

Missing: A framework that embodies the Natural Science Paradigm—integrating Rigorous Logic with Adaptive Workflow to deconstruct physical systems

Introducing the SIDERIUS Program

Bridging the Gap between General AI Agents and Physics AI Agents

SIDERIUS: **S**cientific **I**nquiry, **D**esign, **E**xploration, and **R**easoning **I**ntelligence **U**sing Multi-Agent **S**ystems

HISTORY

- **Sidereus Nuncius (1610)**
  Galileo's 'Starry Messenger' marked the first use of the telescope for systematic cosmic observation.

- **The Paradigm Shift**
  Transition from Aristotelian dogmas to Evidence-driven discovery and mathematical reasoning.

"All truths are easy to understand once they are discovered; the point is to DISCOVER them."

— Galileo Galilei, 1632

...

"A Paradigm is an entire constellation of beliefs, values, and techniques."

— Thomas Kuhn, 1962
The Structure of Scientific Revolutions

COMPOSITION

1. THE MINDSET

Scientific Paradigm:
An adaptive framework for hypothesis and correction.

2. THE TOOLSET

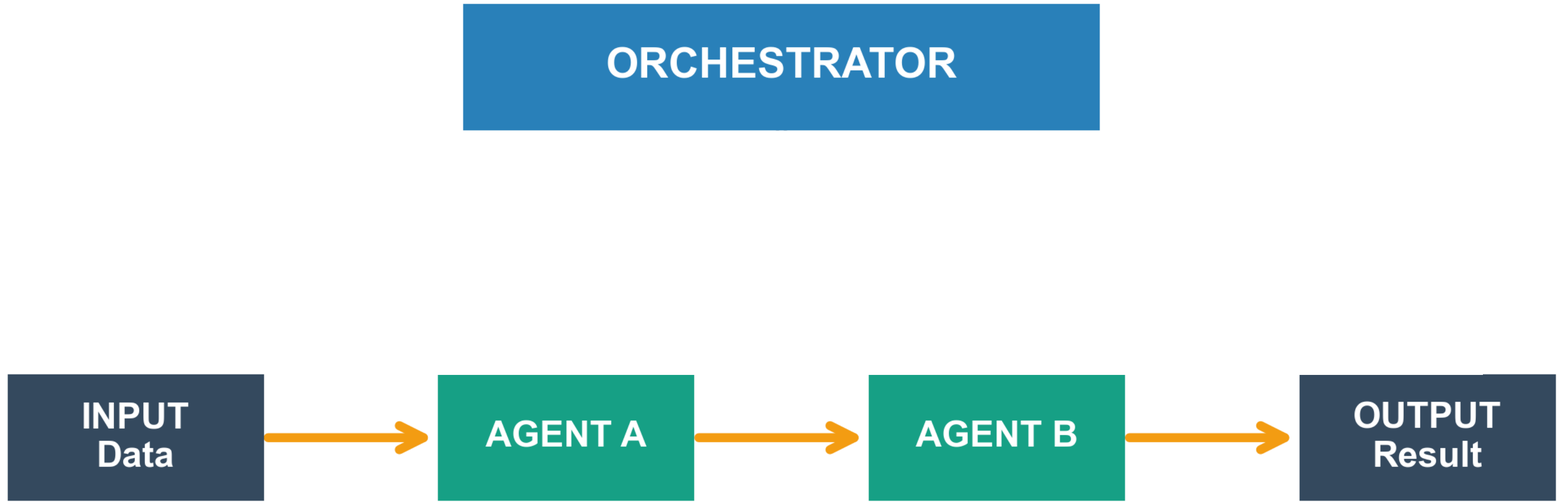
Rigorous Instruments:
- Statistical Tools
- Domain-Specific Methods
- Code Implementation
- etc.

Introducing the SIDERIUS Program

The High-level Design Principles

1. HIERARCHY

Managers plan the strategy;
Employees execute specialized skills.



Introducing the SIDERIUS Program

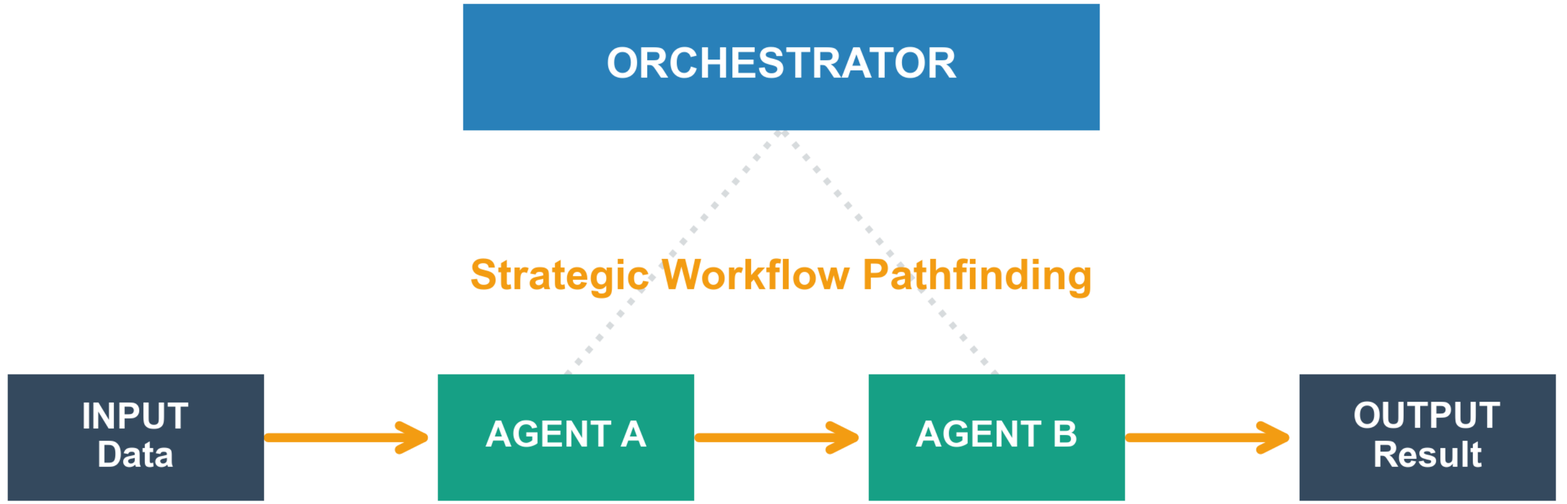
The High-level Design Principles

1. HIERARCHY

Managers plan the strategy;
Employees execute specialized skills.

2. PATHFINDING

Orchestrator finds the optimal I/O path
to construct end-to-end workflows.



Introducing the SIDERIUS Program

The High-level Design Principles

1. HIERARCHY

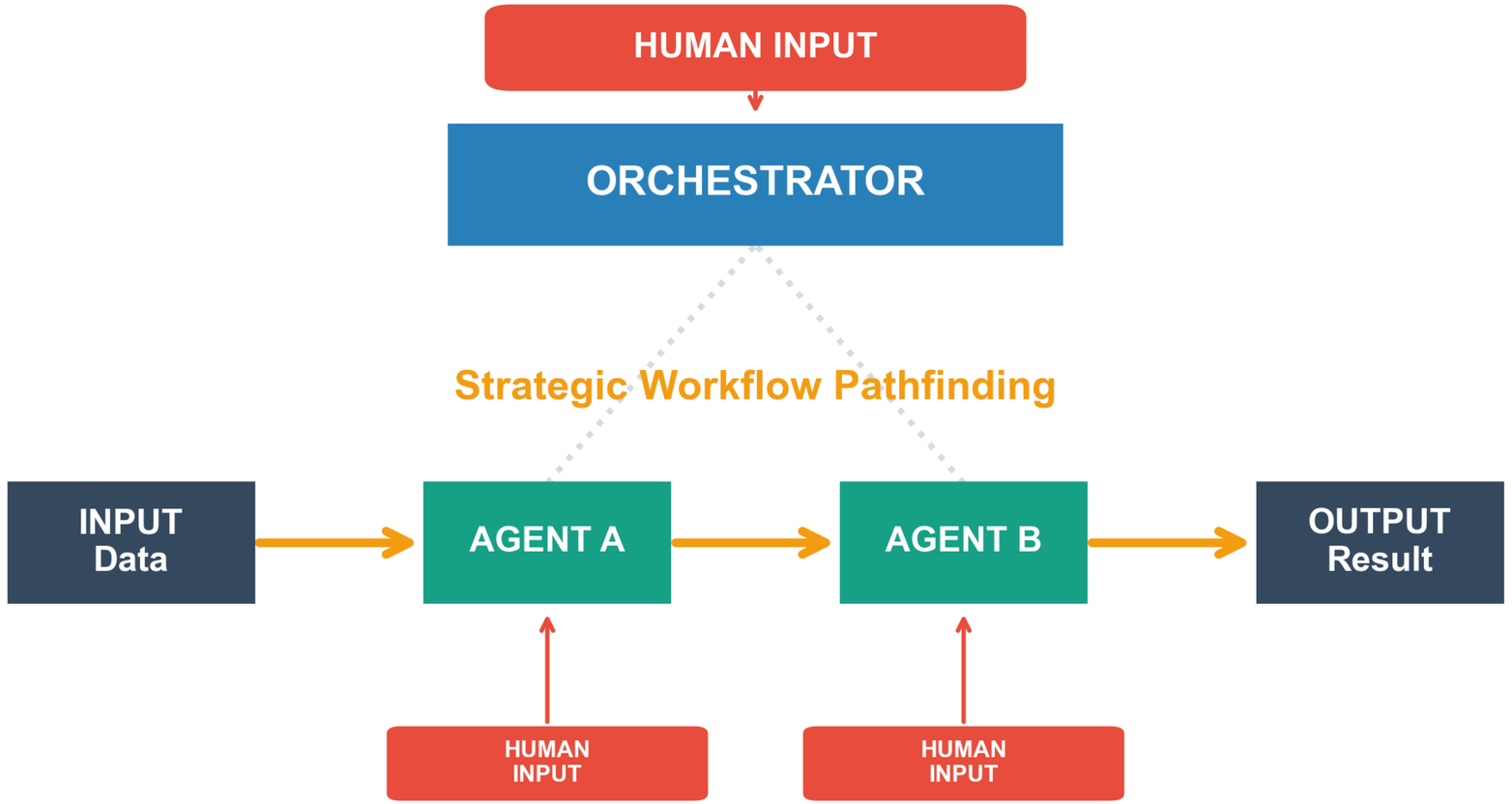
Managers plan the strategy;
Employees execute specialized skills.

2. PATHFINDING

Orchestrator finds the optimal I/O path to construct end-to-end workflows.

3. SURROGACY

Every module is pluggable and can be surrogated by a human expert.



Any module, at any stage, can be 'Surrogated' by a Human expert.

Introducing the SIDERIUS Program

The High-level Design Principles

1. HIERARCHY

Managers plan the strategy;
Employees execute specialized skills.

2. PATHFINDING

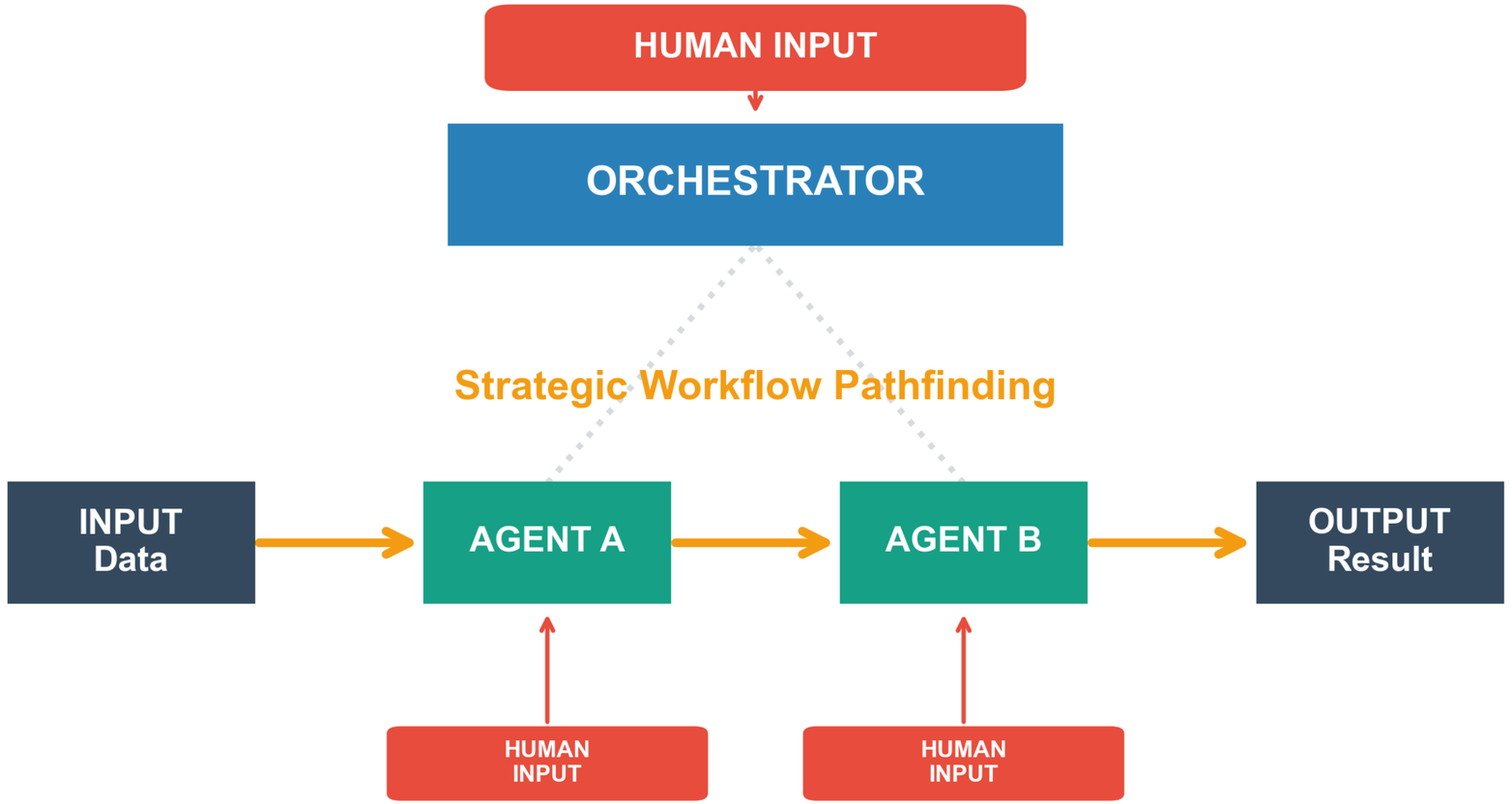
Orchestrator finds the optimal I/O path to construct end-to-end workflows.

3. SURROGACY

Every module is pluggable and can be surrogated by a human expert.

4. SCALABILITY

Test modules individually and keep adding new agents to the family.



Any module, at any stage, can be 'Surrogated' by a Human expert.

Dive into the Very First Agent: Hyperparameter Tuner

Baseline Models in the TIDMAD Dataset

1. FREQUENCY SPLITTING

20 Files with different frequency ranges;
Represents differenet axion masses

2. TRAINING BOTTLE-NECKS

Trade-offs between architectural complexity and performance.

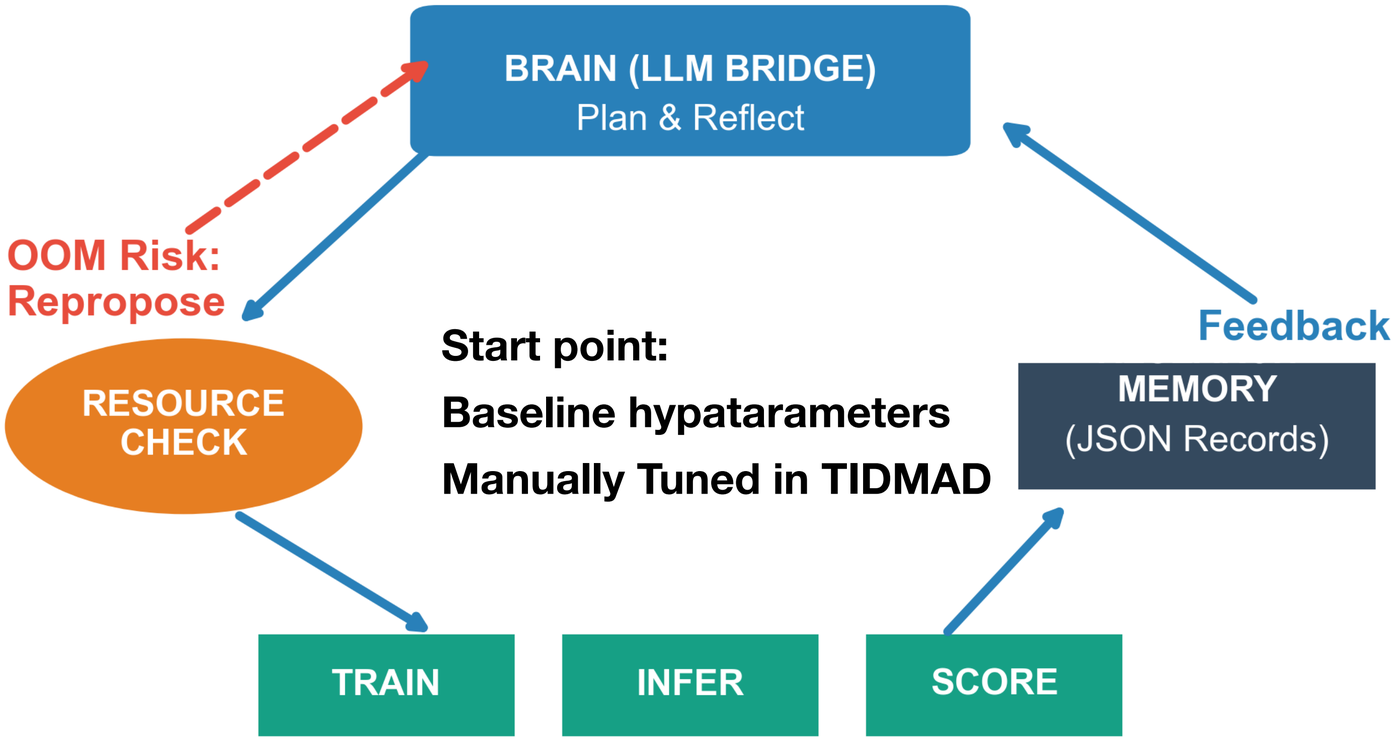
3. HARDWARE CONSTRAINTS

GPU Memory limits and parallelization efficiency.

FC NET Fine Score: 6.43 (Winner)	PRO	Simplest architecture; Highest score.
	CON	Large Number of Parameters.
WAVENET Fine Score: 4.99	PRO	Generalization accross various frequencies.
	CON	High computational density.
TRANSFORMER Fine Score: 3.95	PRO	Global context through self-attention.
	CON	O(N^2) memory scaling; Slow for high-rate data.
PU NET Fine Score: 3.69	PRO	Positional Encoder-Decoder framework.
	CON	May be over-complex for single-sine recovery.
RNN SEQ2SEQ Fine Score: 3.38	PRO	Classic temporal memory structure.
	CON	Sequential processing; Cannot parallelize.

Dive into the Very First Agent: Hyperparameter Tuner

The Working Loop



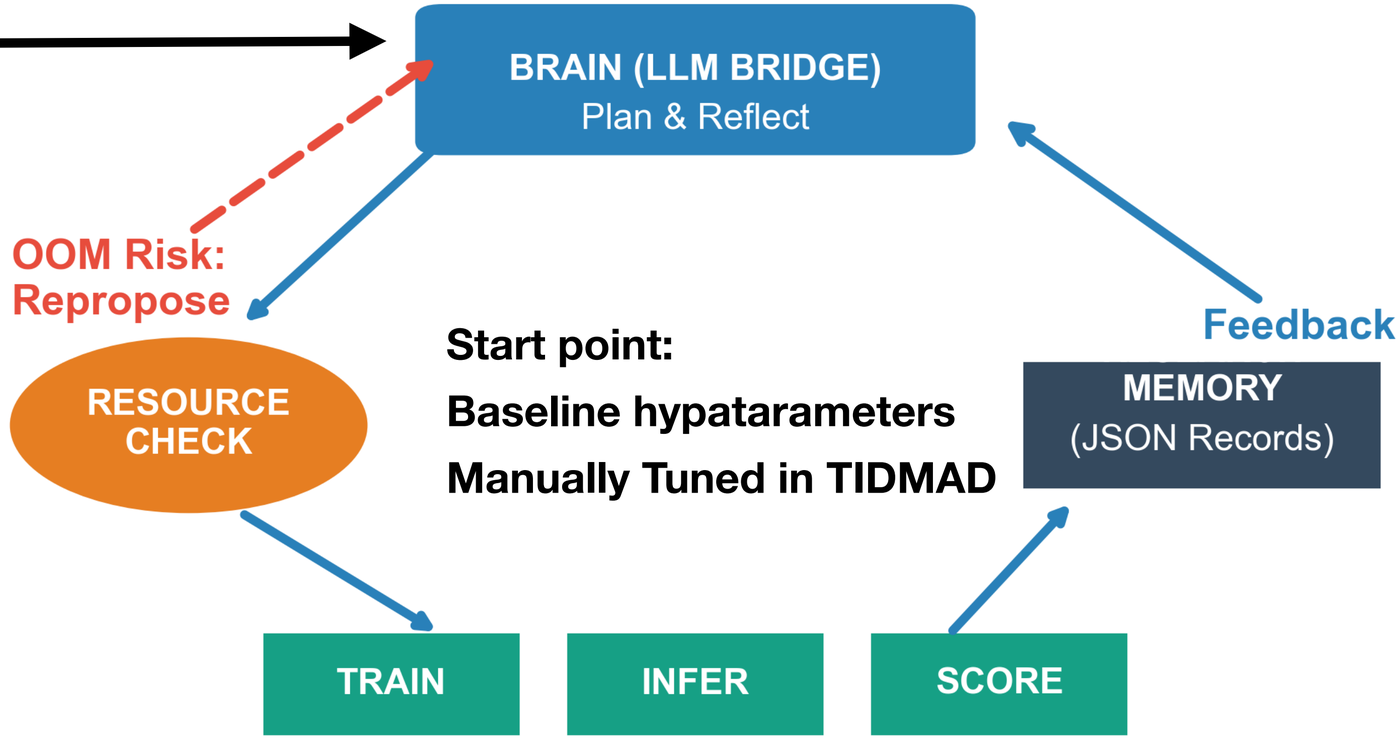
Dive into the Very First Agent: Hyperparameter Tuner

The Working Loop

Prompt:

You are a Research Analyst. Your job is to transform raw experiment results into Research Memory.

OBJECTIVES:
- Validate Hypothesis: Compare the initial hypothesis with the actual Denoising Score and Loss.
- Extract Discovery: Identify a specific pattern or rule learned from this run.
- Update Memory: Write a concise 'Memory Entry' that will guide the Planner in the next iteration.



The Hyper Parameter Tuning Loop

An example of a “research round”: choose loss for FCNET

"memory": {
"expert_advice_followed": None
"hypothesis": "Shifting from Smooth L1 to Focal Loss will force the AutoEncoder to prioritize reconstruction of complex, non-linear signal components, preventing the 'Mean-Output' behavior that leads to the -13.85 score.”,

* For fast prototyping, we only trained on one single file

The Hyper Parameter Tuning Loop

An example of a “research round”: choose loss for FCNET

"memory": {
"expert_advice_followed": None
"hypothesis": "Shifting from Smooth L1 to Focal Loss will force the AutoEncoder to prioritize reconstruction of complex, non-linear signal components, preventing the 'Mean-Output' behavior that leads to the -13.85 score.”,

"conclusion": "The hypothesis partially held true; while the 'Mean-Output' behavior was successfully mitigated, the transition to Focal Loss failed to yield a positive denoising score, resulting in a score of -1.18.”,

The Hyper Parameter Tuning Loop

An example of a “research round”: choose loss for FCNET

"memory": {
"expert_advice_followed": None
"hypothesis": "Shifting from Smooth L1 to Focal Loss will force the AutoEncoder to prioritize reconstruction of complex, non-linear signal components, preventing the 'Mean-Output' behavior that leads to the -13.85 score.”,

"conclusion": "The hypothesis partially held true; while the 'Mean-Output' behavior was successfully mitigated, the transition to Focal Loss failed to yield a positive denoising score, resulting in a score of -1.18.”,

"discovery": "While Focal Loss forces the model to attend to harder samples, it also introduced significant gradient instability in the later training stages, causing the loss to plateau and oscillate rather than converge toward an optimal reconstruction.”

The Hyper Parameter Tuning Loop

An example of a “research round”: choose loss for FCNET

"memory": {
"expert_advice_followed": None
"hypothesis": "Shifting from Smooth L1 to Focal Loss will force the AutoEncoder to prioritize reconstruction of complex, non-linear signal components, preventing the 'Mean-Output' behavior that leads to the -13.85 score.”,

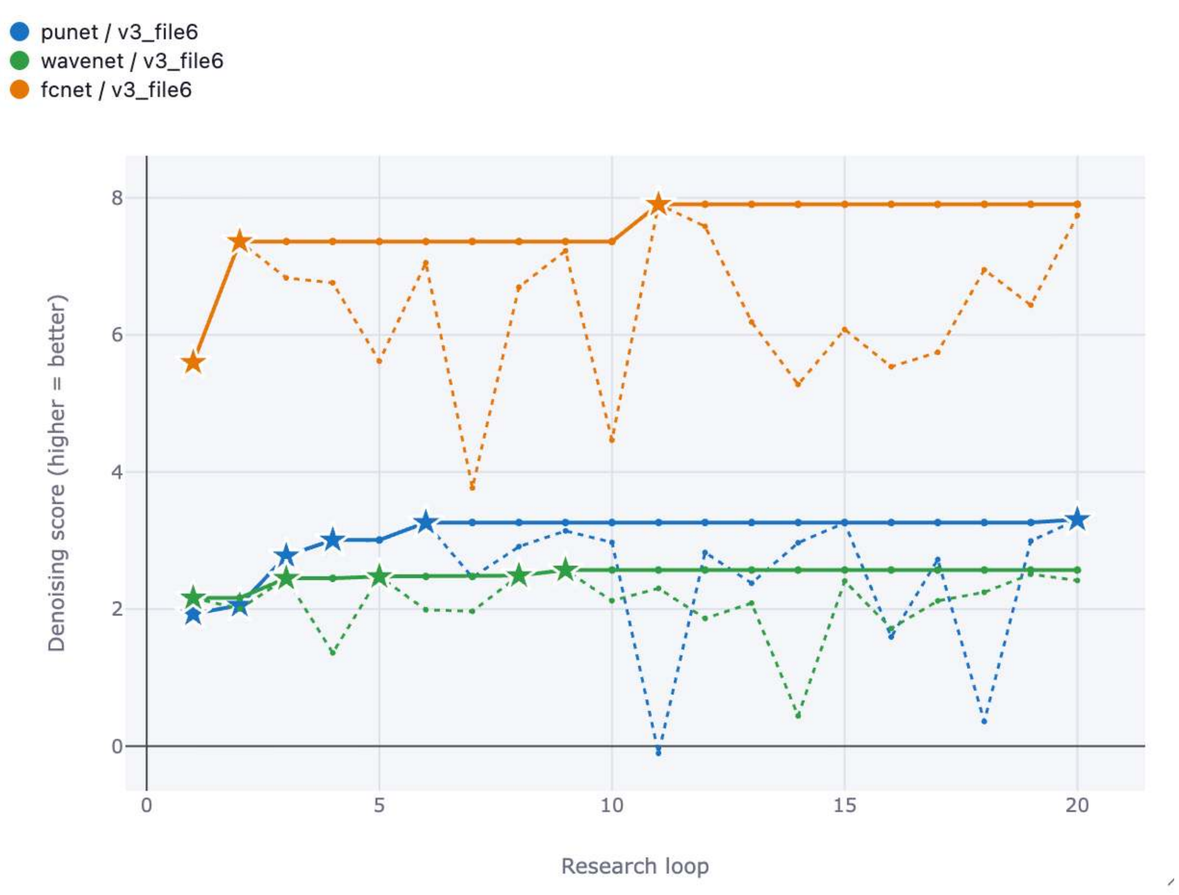
"conclusion": "The hypothesis partially held true; while the 'Mean-Output' behavior was successfully mitigated, the transition to Focal Loss failed to yield a positive denoising score, resulting in a score of -1.18.”,

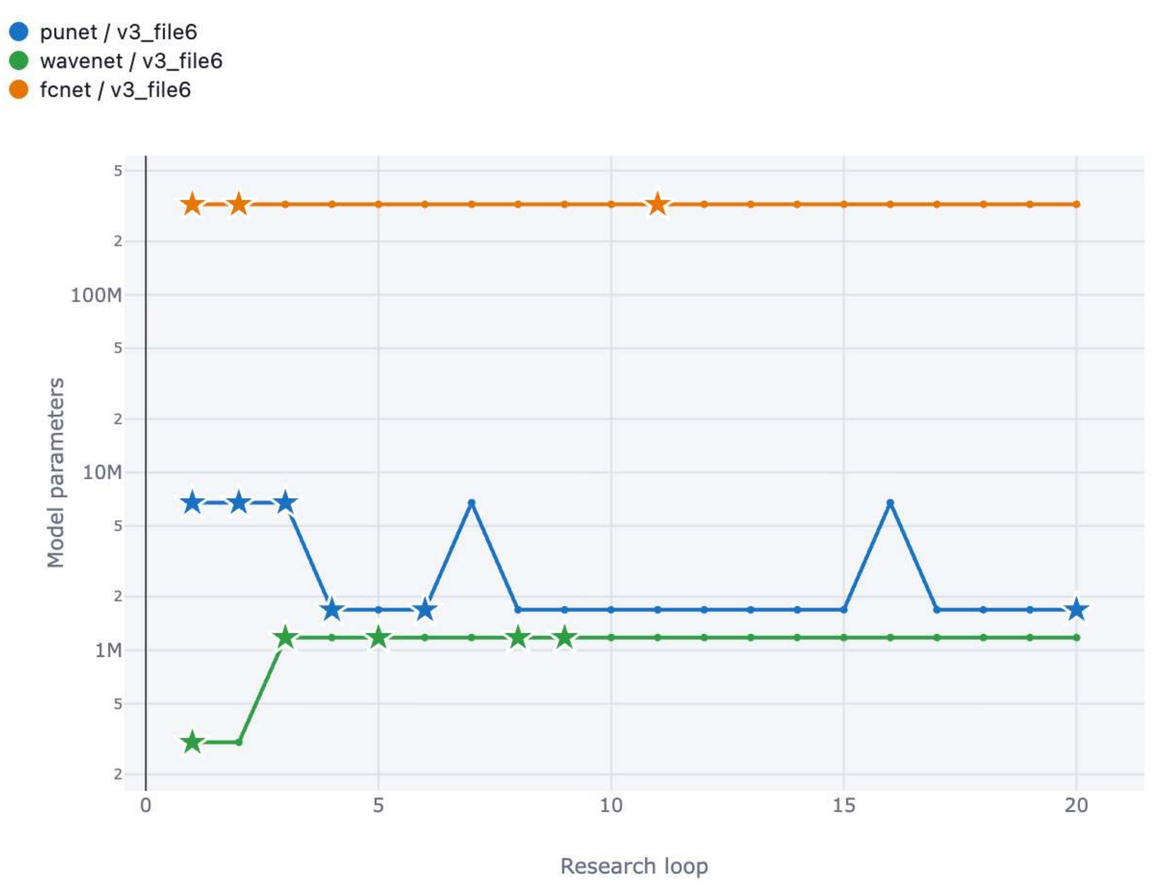
"discovery": "While Focal Loss forces the model to attend to harder samples, it also introduced significant gradient instability in the later training stages, causing the loss to plateau and oscillate rather than converge toward an optimal reconstruction.”

"memory_update": "The model is now sensitive to outliers rather than ignoring them. Future experiments should keep the Focal Loss structure but incorporate a learning rate scheduler or gradient clipping to stabilize the training phase and prevent the observed late-stage oscillation."
}

Milestone 1: Auto Hyperparameter Tuning

Clear Improvement within a few Research Loops





*** For fast prototyping, we only trained on one single file**

Connecting the Multi-agent System

The Logic Engine: Protocol-Driven Graph Architecture

1. UNIVERSAL SKILLS

Every module (LLM, Tool, Workflow)
adheres to a single {Input, Output} contract.

Connecting the Multi-agent System

The Logic Engine: Protocol-Driven Graph Architecture

1. UNIVERSAL SKILLS

Every module (LLM, Tool, Workflow)
adheres to a single {Input, Output} contract.

2. TYPED PROTOCOLS

Nodes never call each other directly.
Edges are explicitly defined by mapping functions.

Connecting the Multi-agent System

The Logic Engine: Protocol-Driven Graph Architecture

1. UNIVERSAL SKILLS

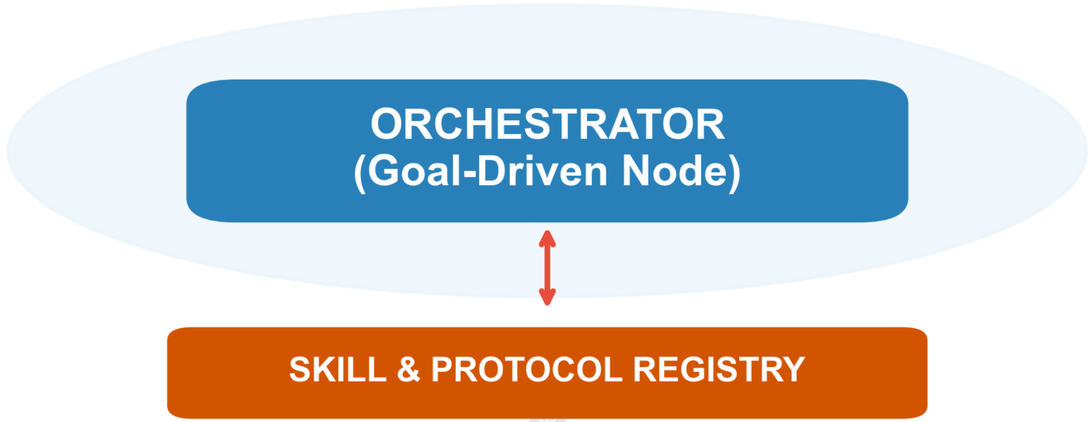
Every module (LLM, Tool, Workflow)
adheres to a single {Input, Output} contract.

2. TYPED PROTOCOLS

Nodes never call each other directly.
Edges are explicitly defined by mapping functions.

3. ORCHESTRATION AS COMPOSITION

An Orchestrator is itself a Skill.
It recursively composes sub-graphs into a larger workflow.



Connecting the Multi-agent System

The Logic Engine: Protocol-Driven Graph Architecture

1. UNIVERSAL SKILLS

Every module (LLM, Tool, Workflow)
adheres to a single {Input, Output} contract.

2. TYPED PROTOCOLS

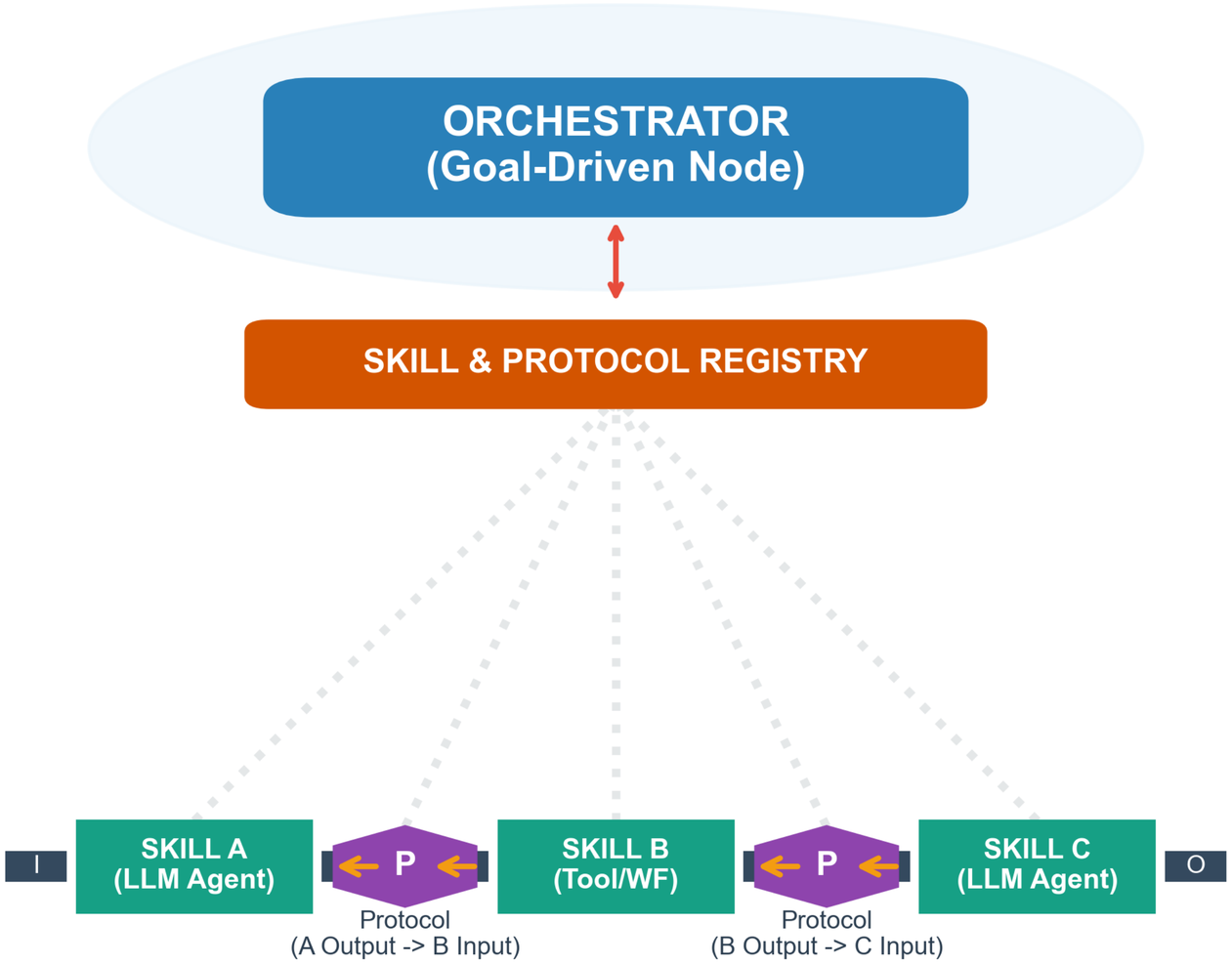
Nodes never call each other directly.
Edges are explicitly defined by mapping functions.

3. ORCHESTRATION AS COMPOSITION

An Orchestrator is itself a Skill.
It recursively composes sub-graphs into a larger workflow.

4. DYNAMIC PATHFINDING

An LLM-driven Orchestrator dynamically discovers
optimal paths via the Skill & Protocol Registry.



The Graph is fixed, but the 'Path' is dynamically designed.

Introducing the SIDERIUS Program

Current Achievement: Workflow to propose and train new ML models

Atomic Node Readiness

All 5 specialized agents are functional, stateless, and independently verified.

Standardized Protocols

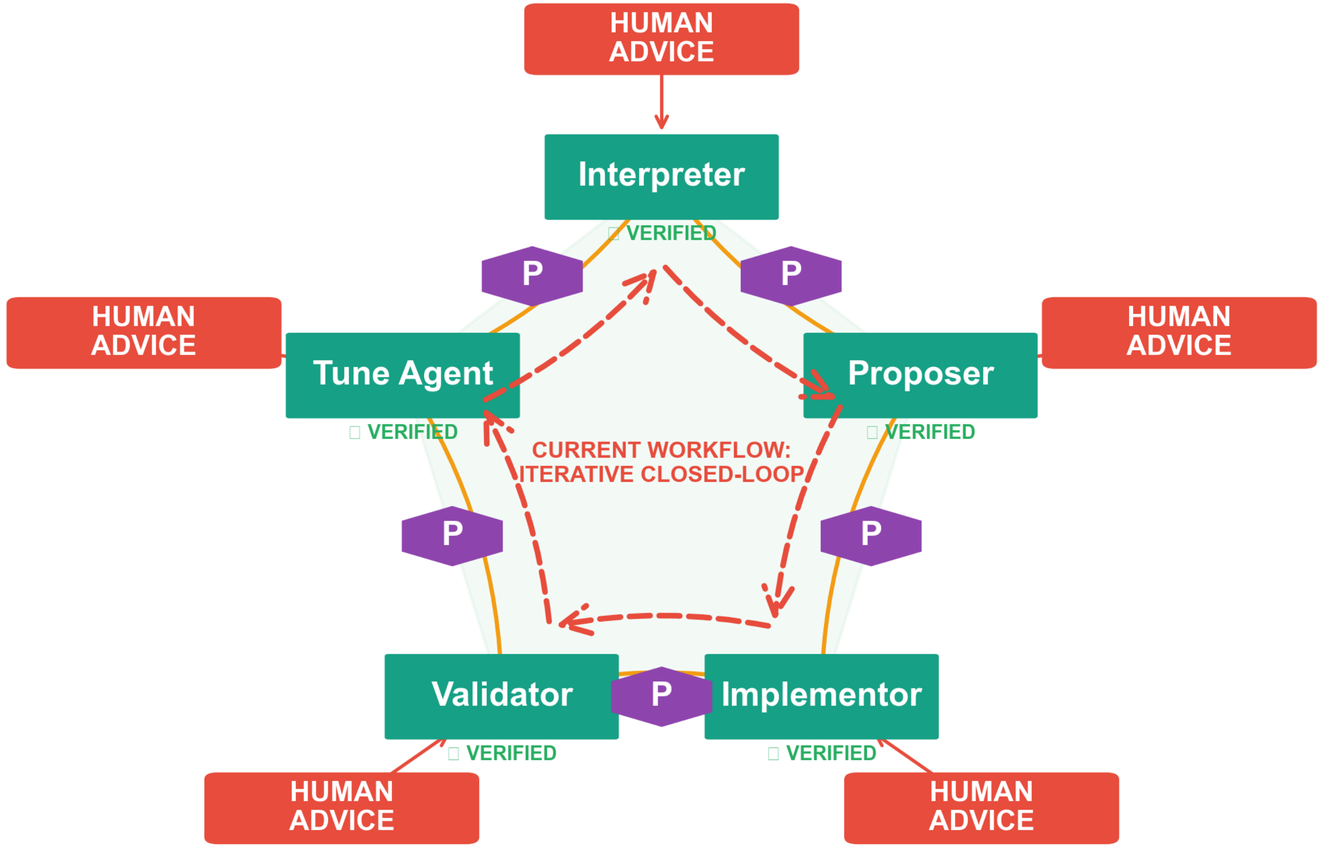
Typed communication interfaces ensure stable data exchange between any node.

Human-in-the-Loop

Every node accepts 'Human Advice' to steer or surrogate the LLM's decision.

Graph-Ready Infrastructure

The architecture is ready to transition from fixed loops to autonomous exploration.



Decoupled Knowledge Graph: Any module can be steered by Human experts.

Ready for the next phase: Transitioning to Autonomous Research Orchestration.

Milestone 2: Proposing New ML Models

Start the workflow by a few sentences

INE PARAMETERS

t Model Pool:	punet, wavenet, fcnet
Name:	v3_file6
Iterations:	50
jet Score:	10.0
ing Cycles:	20
l Engine:	Gemini-3.1-Flash

HUMAN ADVICE: PROPOSAL PHASE

DO NOT propose complex architectures (Attention, Transformers, Mamba).
1. Option A: Simple and straightforward commonly used models.
2. Option B: Variants of existing models (Dilated CNN, Residual CNN).
• Constraint: < 50K parameters, basic PyTorch modules only.
• Strategy: Build on the strengths of fcnet (Score: 7.9).

HUMAN ADVICE: TUNING PHASE

• Set batch_size = 1.
• Focus: Optimize hyperparameters within 20 rounds per candidate.

API Token Usage: < $0.3 for a full loop

Milestone 2: Proposing New ML Models

A demo of one complete loop: 1. Interpretation

Interpreter Synthesis: Initial 3 Models Review

KEY SCIENTIFIC FINDINGS

- **FC-Net (7.91) vs. PU-Net (3.31) / WaveNet (2.57).**
- Regression loss (SmoothL1) > Classification (Focal).
- Extreme sensitivity to LR (1e-4 to 4e-4 range).
- Performance plateaus at representational limits.

CRITICAL BOTTLENECKS

- Inductive Bias: Lack of adaptive temporal scaling.
- Optimization: High input dimensionality causes volatility.
- Stability: Narrow, unstable hyperparameter configs.
- Overfitting: Depth rapidly diminishes Val performance.

TAKE-HOME MESSAGE & DIRECTION

The performance plateau and hyperparameter fragility across all paradigms necessitate a transition toward models with ADAPTIVE TEMPORAL RECEPTIVE FIELDS and regularization that decouples feature depth from overfitting risk.

Milestone 2: Proposing New ML Models

A demo of one complete loop: 2. Proposal

Agent Proposal: Dilated Residual Bottleneck (DRB)

ARCHITECTURAL CONCEPT & MOTIVATION

"The DRB architecture replaces dense, parameter-heavy input layers with a series of hierarchical residual blocks... addressing receptive field misalignment while maintaining global context awareness."

"Eliminates the O(T x d) parameter explosion of FCNet and avoids training instabilities of hierarchical U-Nets by providing direct, linear-cost identity paths across the entire 40k temporal sequence."

MATHEMATICAL DEFINITION

$$y_{s+1} = y_s + f_s(y_s)$$

$$z = \mathrm{GELU}(\mathrm{Conv}_{1\times 1}(y_s)) \leftarrow \textit{Pointwise Compression}$$

$$\hat{z} = \mathrm{GELU}(\mathrm{DWConv}_d(z)) \leftarrow \textit{Dilated Context Aggregation}$$

$$f_s(y_s) = \mathrm{Conv}_{1\times 1}(\hat{z}) \leftarrow \textit{Pointwise Expansion}$$

where $d = \mathrm{base}^s$ enables receptive field $\approx O(T)$

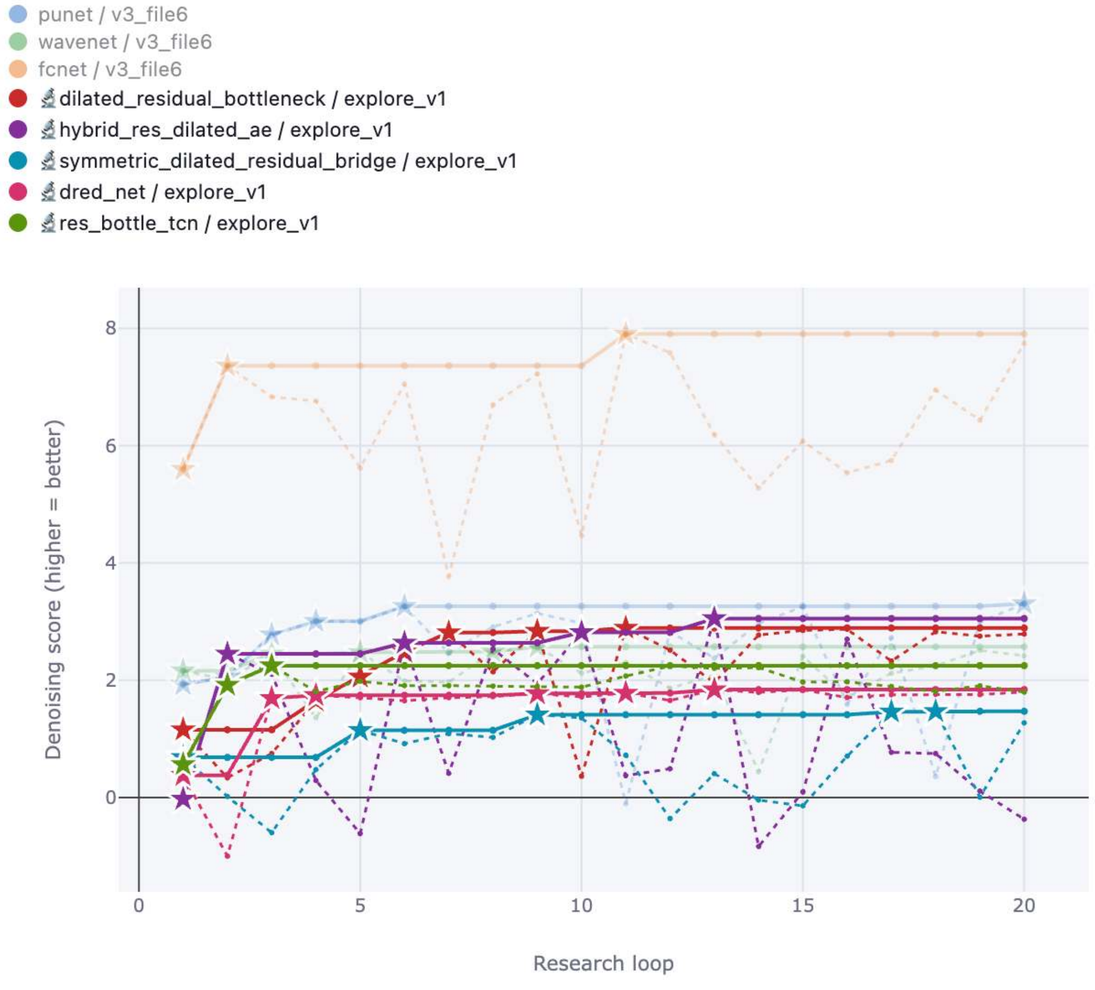
EXPERT CONSTRAINTS & BASELINE

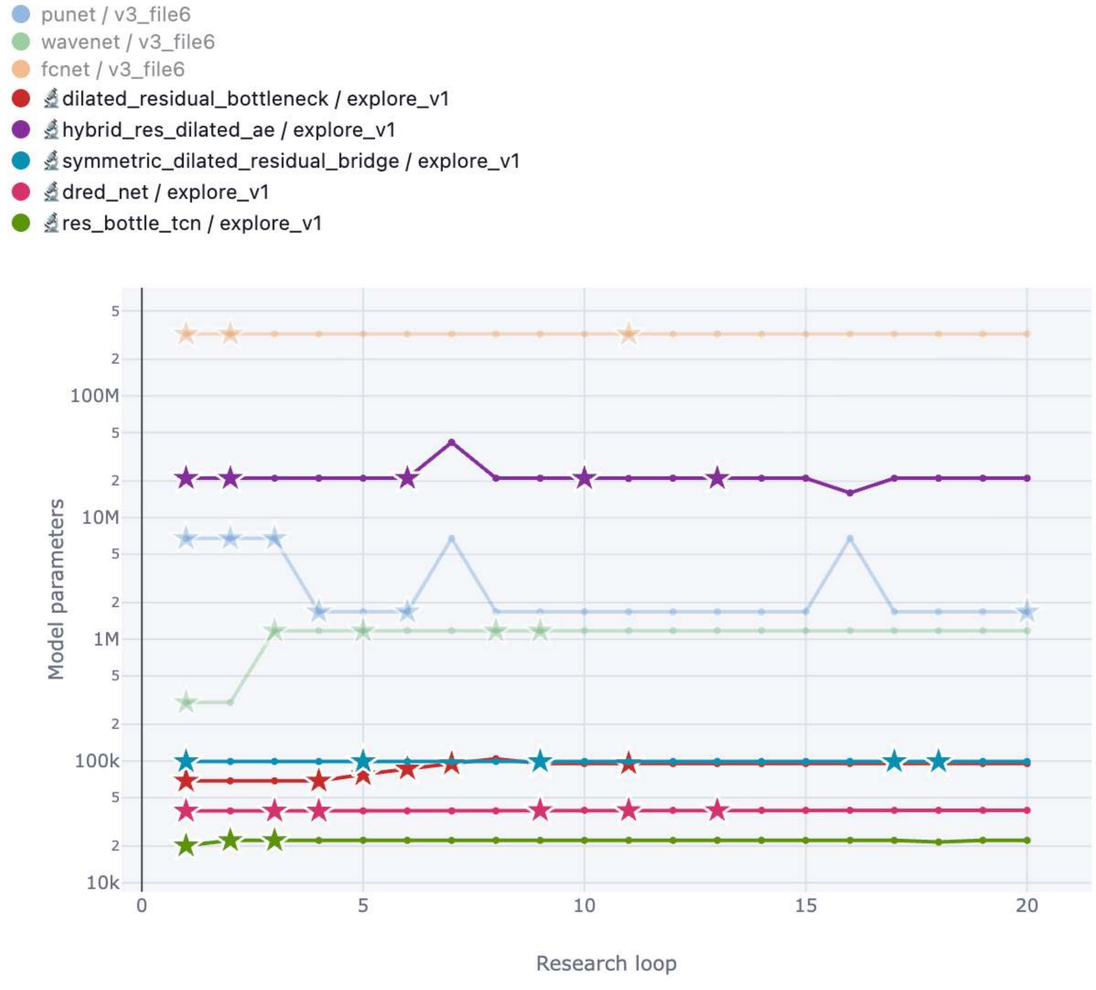
- **VRAM strictly capped at 8GB during training**
- **Parameter count < 1,000,000 (Strict limit)**
- Dilation schedule: Receptive field > 40,000
- Baseline: blocks=8, channels=64, base=2, lr=1e-4

"Rationale: Dilation covers long-range dependencies efficiently without the parameter cost of dense layers."

Milestone 2: Proposing New ML Models

Autonomous Loop with a few sentences of human Advice





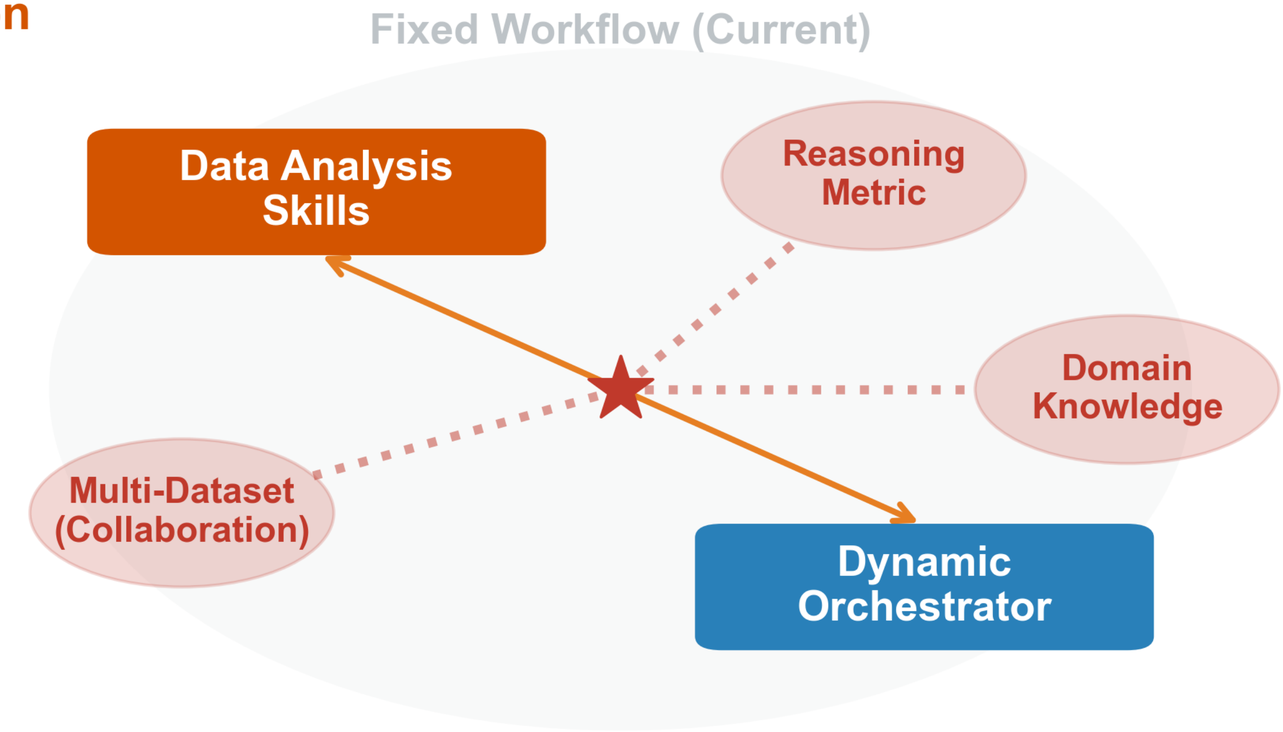
Proposing New Models: Reached decent denoising score with lightweight models

The Next Steps & Long-term Proposal

Expand the scope of auto ML task solver

Short-term: Data Insight & Logic Validation

- Step A: Deploy Data Analysis skills to improve denoising performance beyond blind search.
- Step B: Implement new metrics to verify if the agent's reasoning aligns with physics ground truth.



The Next Steps & Long-term Proposal

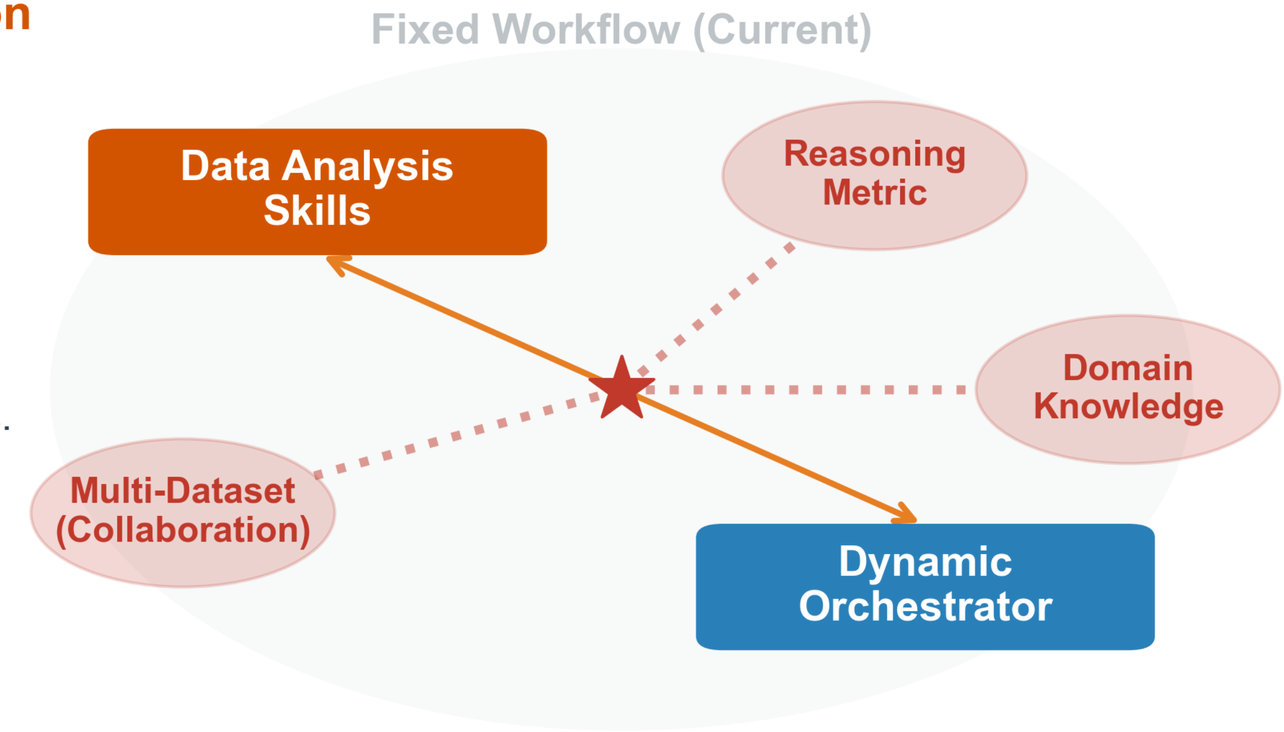
Expand the scope of auto ML task solver

Short-term: Data Insight & Logic Validation

- Step A: Deploy Data Analysis skills to improve denoising performance beyond blind search.
- Step B: Implement new metrics to verify if the agent's reasoning aligns with physics ground truth.

Long-term: Autonomous Orchestration

- Transition from Fixed Pipelines to Goal-Driven agents.
- Orchestrator dynamically queries Skill Registry.
- Pathfinding beyond iterative loops.



The Next Steps & Long-term Proposal

Expand the scope of auto ML task solver

Short-term: Data Insight & Logic Validation

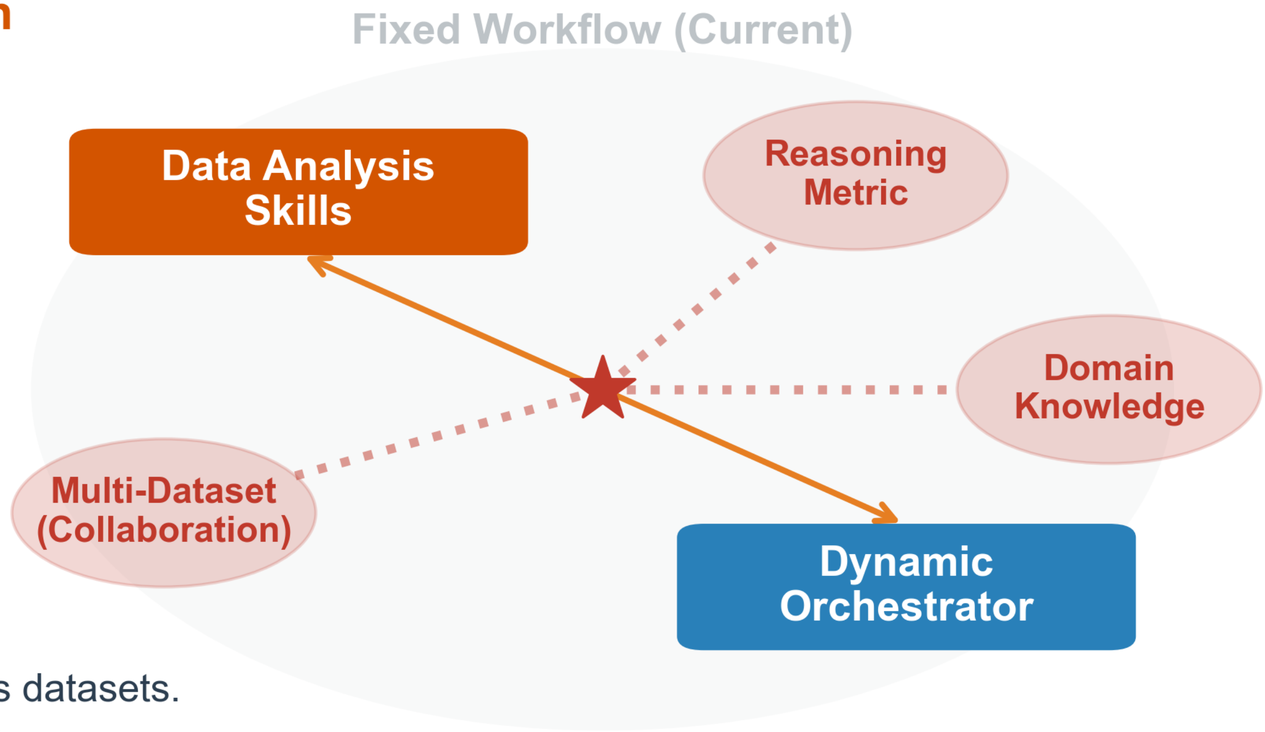
- Step A: Deploy Data Analysis skills to improve denoising performance beyond blind search.
- Step B: Implement new metrics to verify if the agent's reasoning aligns with physics ground truth.

Long-term: Autonomous Orchestration

- Transition from Fixed Pipelines to Goal-Driven agents.
- Orchestrator dynamically queries Skill Registry.
- Pathfinding beyond iterative loops.

Vision: Cross-Domain Scalability

- Generalize SIDERIUS from TIDMAD to diverse physics datasets.
- Build a collaborative ecosystem for AI-driven discovery.



Evolving from an ML Automation tool to an Autonomous AI (Co-)Scientist.

Backups

Human Advice to the Workflow:

This one Single Call

Initiate the Auto-model proposing loop

```
results = run_workflow(
    data_dir="/home/klz/Data/SIDEREIS_DATA",
    model_types=["punet", "wavenet", "fcnet"],
    source_run_name="v3_file6",
    workspace="/home/klz/Data/SIDEREIS_DATA/exploration",
    run_name="explore_v1",
    max_iterations=50,
    max_rounds=20,
    max_proposal_attempts=10,
    target_score=10.0,
    llm_config=llm_config,
    human_advice_propose=(
        "Do NOT propose complex architectures like attention, transformers, or mamba. "
        "Start simple: please use equal probability to choose from these two options: "
        "1. propose a easy and straightforward, commonly used model. "
        "2. try modification or variant of the existing models. "
        "For example, a WaveNet-style dilated CNN, a simple residual CNN, or a deeper "
        "version of the existing U-Net with minor changes. The model must be easy to "
        "implement correctly with basic PyTorch modules (Conv1d, ReLU, BatchNorm, "
        "residual connections). No custom attention mechanisms. Under 50K parameters. "
        "Note: fcnet (AutoEncoder) achieved the best score of 7.9 — consider building "
        "on its strengths."
    ),
    human_advice_tune=(
        "Use batch_size=1. Focus on finding the best hyperparameters within 20 rounds."
    ),
)
```

Success Rate in New Model Proposing

```
SIDERIUS Model Exploration Workflow — Status Report
============================================================

| Iter | Model                                    | Attempts | Rounds  | Best Score | Status   |
|------|------------------------------------------|----------|---------|------------|----------|
|    1 | dilated_residual_bottleneck              |        1 | 20/20   |     2.8918 | Complete |
|    2 | hybrid_res_dilated_ae                    |        2 | 20/20   |     3.0527 | Complete |
|    3 | symmetric_dilated_residual_bridge        |        6 | 20/20   |     1.4716 | Complete |
|    4 | dred_net                                 |        4 | 20/20   |     1.8406 | Complete |
|    5 | dilated_bottleneck_net                   |        5 | 20/20   |     1.7089 | Complete |
|    6 | res_bottle_tcn                           |        1 | 20/20   |     2.2490 | Complete |
|    7 | ?                                        |        1 | ~ 0/20  |          - | Running  |

Completed iterations : 6/7
Total proposals      : 20
Best agent score     : 3.0527 (hybrid_res_dilated_ae)
Target score         : 10.0

Input models (from v3_file6):
| Model         | Rounds | Best Score |
|---------------|--------|------------|
| punet         |     20 |     3.3063 |
| wavenet       |     20 |     2.5696 |
| fcnet         |     20 |     7.9052 |
```

Early Stop and Error Handling

"memory": {
"expert_advice_followed": "You should actively try different model configs, loss types, and train configs, while not exceeding the limit of GPU memory. The baseline result is already in your memory — your goal is to find configurations that outperform it.”,

"hypothesis": "Switching from 'ce' to 'focal' loss with a smaller learning rate will provide better gradients for the harder-to-classify samples, potentially improving the denoising score further without requiring additional model complexity.”,

"conclusion": "Skipped: estimated VRAM (23.941 GB) exceeds 80% safety limit (22.83 GB).”,

"discovery": "❌ OOM RISK — Estimated 23.94 GB vs 28.54 GB free (31.3 GB total, 2.80 GB already used). Safety limit: 22.83 GB (80%). Bottleneck: transformer_attn.”,

"memory_update": "Reduce segmentation_size (attention scales $O(T^2)$), or reduce nhead/num_layers."
}