



中国科学院高能物理研究所

Institute of High Energy Physics Chinese Academy of Sciences

面向高能物理的人工智能平台

张正德

2023年2月





白泽人工智能平台(Hakutaku AI, HAI)的总体目标是：

面向粒子物理、天体物理、同步辐射和中子科学提供共性技术支撑，涵盖数据标注、数据处理、算法研究、模型训练、模型评估和云边部署的ML开发全流程，提供AI软件框架、算法库接口、数据集接口、前沿神经网络构建方案、异构资源调度、智能高效开发环境、代码托管等功能，以推动AI赋能科研，解决具体科学问题，助力科学探索新发现，促进科研模式的转变。

关键词：

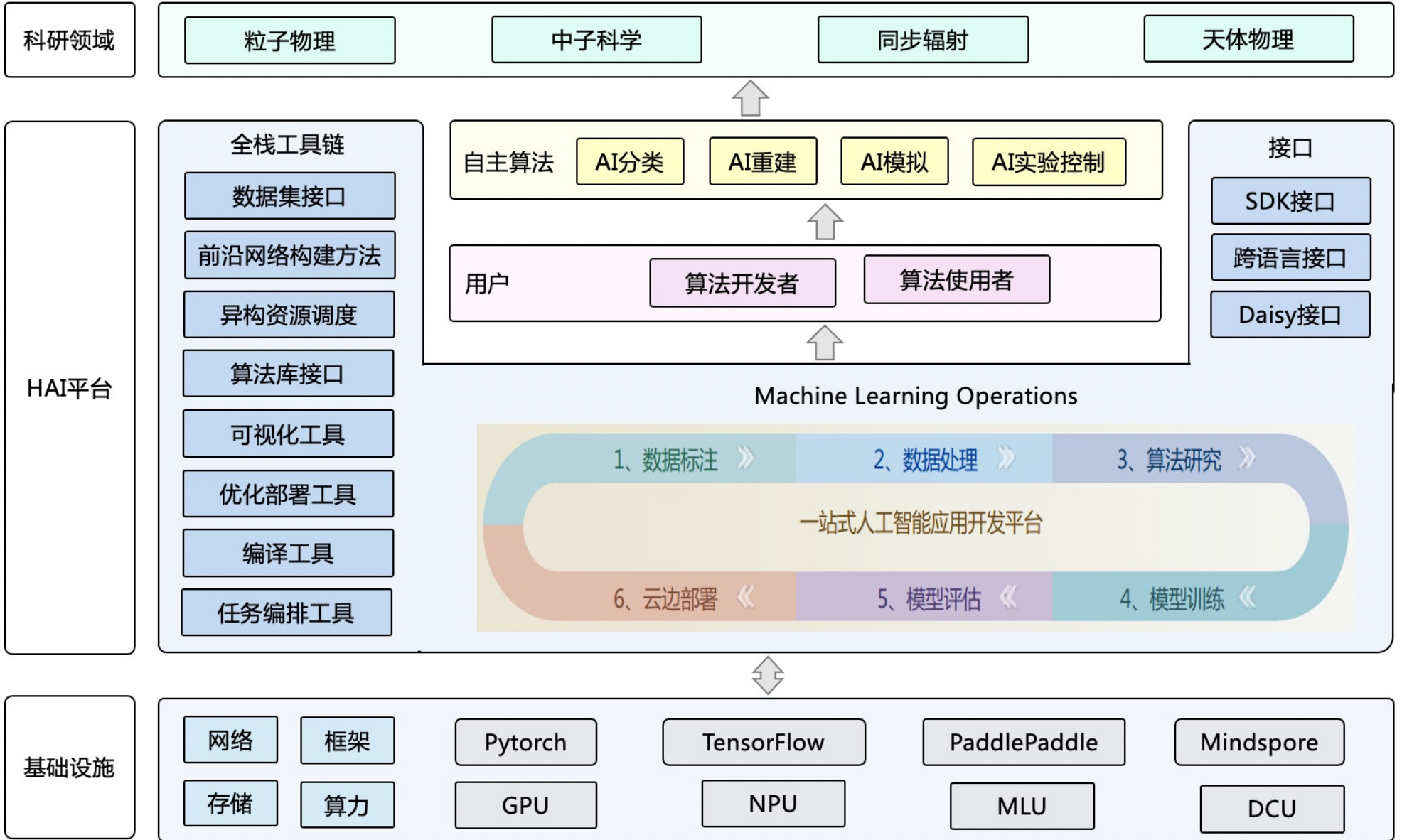
- 软硬件平台
- 工具集和基准
- 解决应用需求
- 推动交叉前沿



HAI平台架构图



- 提供数据标注、数据处理、算法研究、模型训练、模型评估和云边部署的**全栈工具链**
- 提供AI软件框架、算法库接口、数据集接口、前沿神经网络构建方案、异构资源调度、智能高效开发环境、代码托管等功能
- 推动AI赋能科研，解决具体科学问题，助力科学探索新发现，促进科研模式的转变。

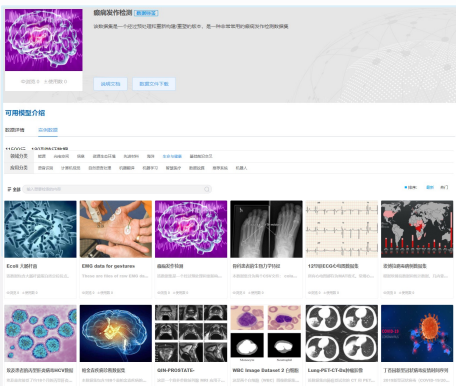




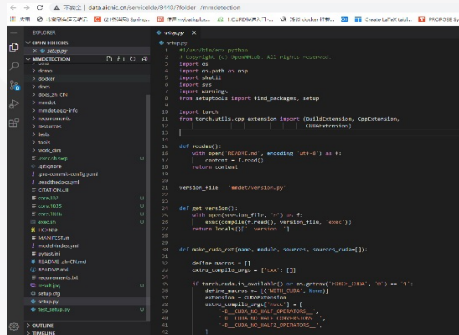
平台功能



领域数据集



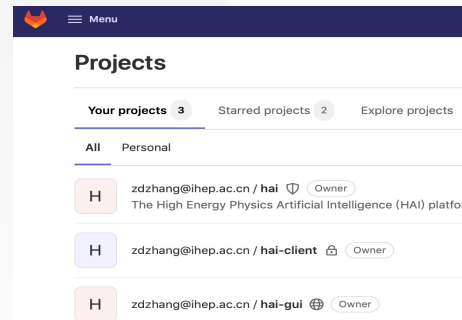
高效开发环境



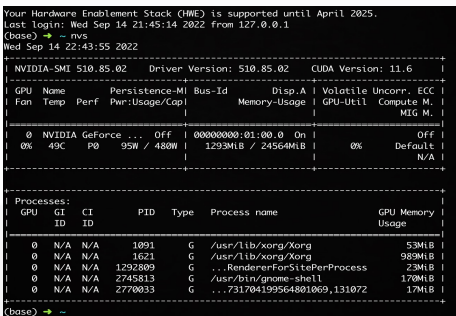
代码AI自动补全



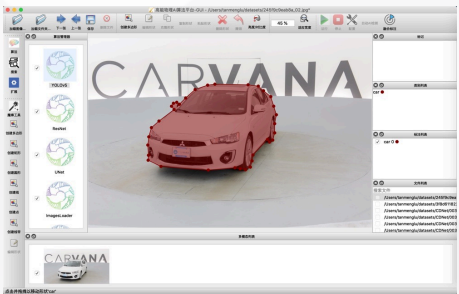
代码托管平台



弹性计算资源



AI标注工具



数据模拟软件

Tomopy	XDS
UFO*	cpp4
HEPSCT*	autoProc
PyFai	dials
PDFgetX3	Phenix
	CONUSS
	PHOENIX

学习网站





快速复现论文

```

1 import hai_client
2 import cv2
3 import numpy as np
4
5 ip = '127.0.0.1'
6 port = 9999
7 hai = hai_client.HAIClient(ip=ip, port=port)
8 modules = hai.hub.list(ret_fmt='json') # 列出所有模型
9 print(f'Modules: \n{modules} {type(modules)}')
10
11 model_name = 'UNet'
12 weights = hai.hub.list_weights(name=model_name) # 列出所有模型权重
13 model = hai.hub.load(model_name, # 根据模型名称加载云模型
14                      weights='hai/unet/unet_v1.1.pth', # 指定模型权重文件, 可选
15                      )
16 docs = hai.hub.docs(model_name) # 查看模型文档连接
17 config = model.config # 获取模型配置
18 # config.weights = "hai/unet/unet_v1.1.pth" # 方法2: 指定模型权重文件, 相同的方法可修改其他配置
19 config_source = '/Path/to/your/datasets' # 指定数据集路径
20 ret_url = model.train() # 云端训练模型
21 print(f'ret_url: {ret_url}') # 通过url查看训练过程和训练结果

```

获取源码、数据集

学习算法、编程、软件

学习前沿网络构建方法

- 空间注意力机制
- 通道注意力机制
- 跨阶段局部网络
- 空间金字塔
- 路径聚合网络
- 自注意力机制

AI辅助数据标注

使用开源算法训练自定义数据集

异构GPU资源调度训练

开源自研算法提升影响力

便捷地进行模型性能对比

神经网络剪枝优化压缩

模型硬件优化部署

算法国产化

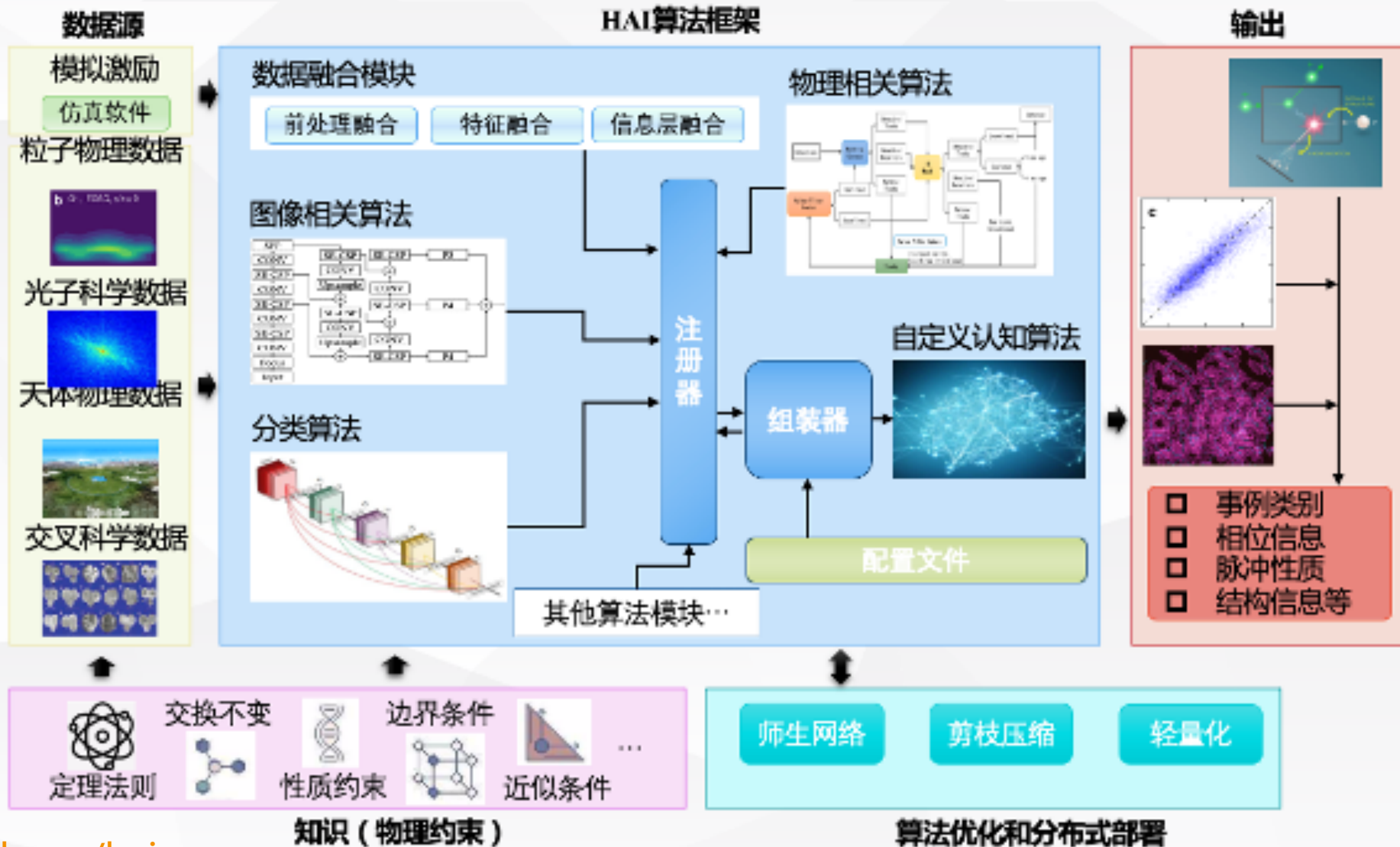
建设进展(2022.12)

- ◆ 整体进展20% <https://code.ihep.ac.cn/zdzhang/hai>
- ◆ 自研AI算法框架**HaiCore**，核心功能70%
 - ◆ 集成算法：2个通用(YOLOv5和ResNet)和4个粒子物理(ParNet, ParT, PCNN, PFN)
 - ◆ 集成开源数据集：3个(JetClass, TopLandscape, QuarkGluon)
- ◆ 自研AI软件界面框架**HaiGF**，核心功能60%
- ◆ 依托基础设施
 - ◆ 7万核CPU，400块GPU，高性能网络，国家高能物理科学数据中心✅
 - ◆ 代码托管平台✅
- ◆ 自有算力
 - ◆ 华为昇腾910 *8卡 NPU服务器✅



开发AI算法框架HaiCore

- ◆ 自研HAI算法框架为平台提供核心能力，包括模型训练、评估、推理、部署统一架构，算法库统一接口，跨语言、跨系统调度方案等，核心功能完成70%+。





全新自研**可扩展、轻量化**的AI软件界面框架(HAI GUI Framework, HaiGF)



①核心功能栏

- 资源浏览器
- 标注工具
- AI工具(远程)
- 更多功能

②主侧栏

- 核心功能的具体展开

③中央控件

- 可视化数据、交互
- 基于选项卡+页面的可扩展设计
- 基于分屏器的自动分屏

④辅助侧栏

- 详细属性、信息等

⑤面板栏

- 多选项卡输出面板

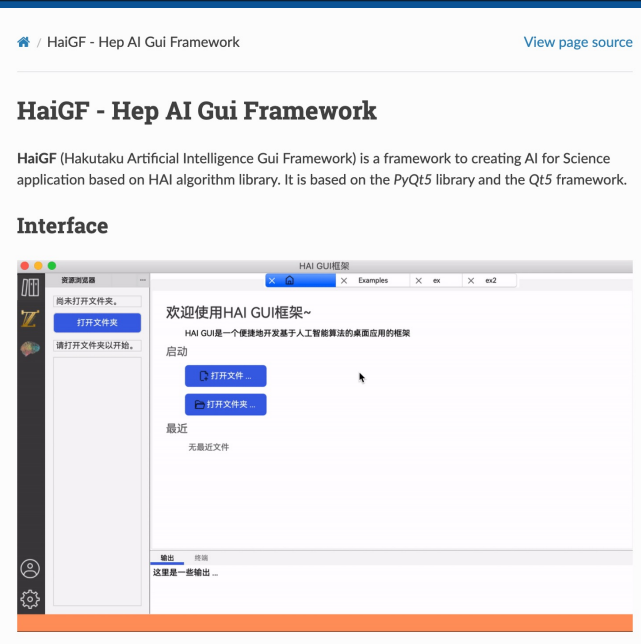
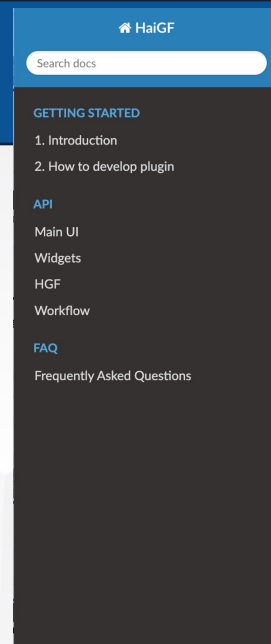


开发AI软件界面框架HaiGF

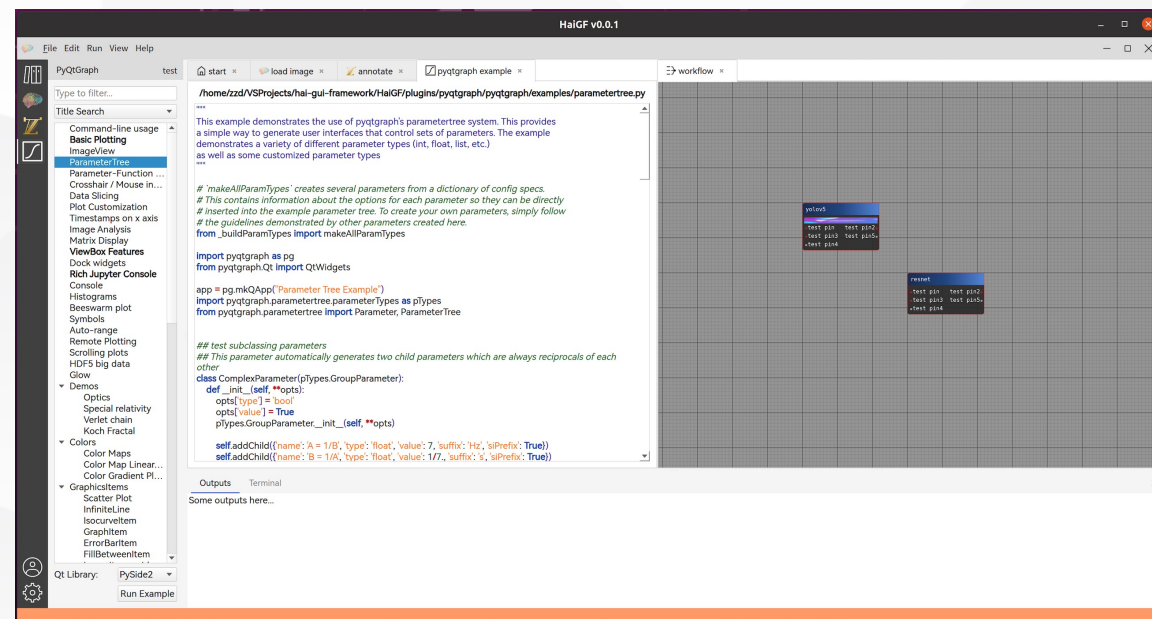
可扩展、轻量化的用于开发基于机器学习算法的界面程序的框架。

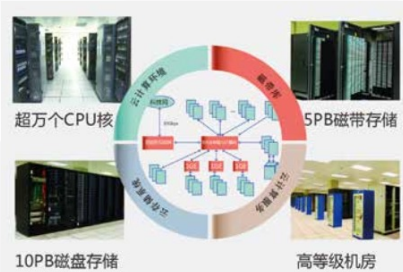
仓库：<https://code.ihep.ac.cn/zdzhang/hai-gui-framework>

- 分屏功能✅，多语言翻译✅。
- HAI算法库集成，完成20%，完成客户端与服务端通讯。
- 开发基于pyflow的算法 workflow，完成20%，完成中央控件的背景界面，算法节点及针脚显示，节点拖动，针脚接引。设计py脚本节点✅。
- 嵌入标注工具。TODO：图像显示，算法与图像联动。
- 撰写文档：如何开发插件、各组件的API、如何生成API_doc、如何翻译GUI界面
- 开发hai_tools插件，实现hai算法库中算法的远程调用
- 开发基于pyflow的工作流，实现算法模块的拖拽和连接
- 集成基于pyqtgraph的科研绘图案例



接口文档：<http://192.168.32.148:8000/>



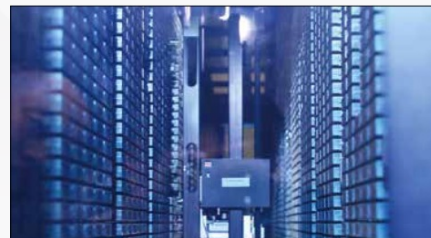
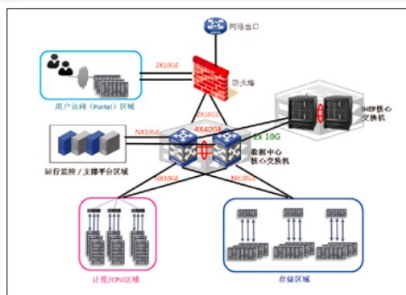


> 大规模高性能计算集群

万核级别高吞吐量计算集群
和混合异构高性能计算集群

> 高性能网络

提供高带宽、低延迟、稳定、安全的网络环境，骨干带宽160Gbps，4*10Gbps双栈互联网出口，到欧洲及北美的10Gbps级专用网络带宽



> 海量存储

建设了高等级的海量数据存储，磁盘70PB，聚合带宽40GB/s，磁带存储40PB

> 分布式计算

为全球高能物理提供超过800万CPU小时的计算服务
为BESIII、JUNO、CEPC等实验提供全球分布式、异构计算的数据处理和分析统一平台



算力：北京 1821节点，7W核，GPU (V100+A100) 400块+



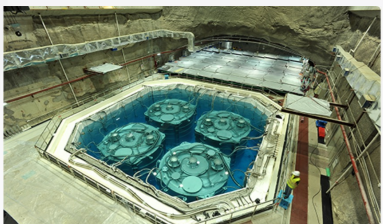
国家高能物理科学数据中心



分为北京数据中心和大湾区中心，实现数据资源、软件工具、数据分析等资源能力的汇交和共享。目前40PB，数万CPU核计算能力，万兆国际网络链路，信息支撑系统。

5个领域的的数据：

高能物理数据



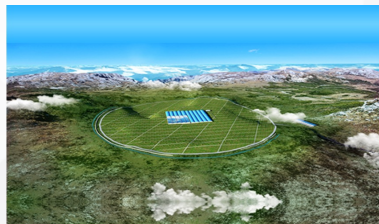
DYB 大亚湾
BESIII 北京谱仪
JUNO 江门
CMS ATLAS LHCb
强子对撞
L3c L3宇宙线

中子科学数据



CSNS
散裂中子源

天体物理数据



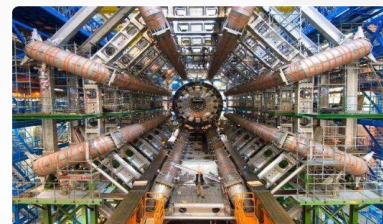
LHAASO 高海拔宇宙线
YBJ-As γ 羊八井
ARGO-YBJ 羊八井
HXMT 硬X射线调制望远镜

光子科学数据



BSRF 北京同步辐射
HEPS 高能同步辐射

交叉学科数据



钢轨在线激光选区
硅雪崩光电探测
碳碳熔合截面测量
试能谱数据
原子核质量数据
双TOF测试数据等



代码托管平台

https://code.ihep.ac.cn

The screenshot shows the GitLab profile page for user Zheng-De Zhang. The profile includes a circular avatar with a cat and the text '只想睡觉' (Just want to sleep). The user's name is 'Zheng-De Zhang', with the handle '@zdzhang' and user ID '1216'. The profile was created on July 22, 2022, at 10:08 AM. There is 1 follower and 0 users followed. The 'Overview' tab is selected, showing a heatmap of activity from March to February. The heatmap has columns for each month and rows for Monday, Wednesday, and Friday. Activity is indicated by blue squares of varying sizes, with the largest square representing issues, followed by merge requests, and then comments. A legend at the bottom shows the color coding for these activity types.

概览 活动 群组 参与贡献的项目 个人项目 星标项目 代码片段 关注者 正在关注的人

3月 4月 5月 6月 7月 8月 9月 10月 11月 12月 1月 2月

一
三
五

议题, 合并请求, 推送及评论。



华为Atlas800-9000 NPU训练服务器上架

配置表：

类型	描述
机框基础配置	Atlas 800 (Model 9000)(3*2.5"NVME SSD 风冷机箱,4*Kunpeng 920,8*Ascend 910 B)
CPU	192核 (Kunpeng 920 2.6GHz * 4)
内存	768G (32G 2933MHz * 24)
NPU卡	128G显存 (Ascend 910 B * 8)
硬盘2	NVME 1.92TB SSD
网卡1	板载GE电口
网卡2	TM272板载灵活网卡-100GE-2端口-QSFP28
电源	8000W (服务器白金2000W * 4)

安装位置：多学科机房

IP：192.168.68.22

系统：Centos 8.2

软件：Pytorch 1.8.1+ascend



方式一：

HaiChatGPT，基于OpenAI API的免费体验版（无需梯子）



Web界面



命令行界面

- 开源地址：<https://github.com/zhangzhengde0225/HaiChatGPT>
- 网页：ai.ihep.ac.cn(内网)，47.114.37.111(公网)

限制：

- 使用个人API_KEY，\$18耗尽后无法调用
- 使用基于GPT3的text-davinci-003模型，性能上不如官网基于GPT3.5的版本



方式二（推荐）：

ChatGPT官网：<http://ai.com/>

- 提供免费web访问，目前最强大的聊天机器人；
- 提供API调用接口，API每月\$18免费额度，按TOKEN计费，约1000次调用
- 不提供模型、预训练权重。
- 限制：
 - 注册需要科学上网，同时，OpenAI对出口限制、高度出口限制国家（中国、俄罗斯等）、受美国制裁国家不提供服务。
 - 免费版服务器爆满
 - Plus会员每月\$20，需国外信用卡和地址支付，Plus版与免费版性能一致，速度略有差别，Plus不排队

附：注册保姆级教程和临时科学上网梯子：

https://code.ihep.ac.cn/zdzhang/haichatgpt/-/blob/main/docs/reg_tutorial.md

- | | | |
|---|---|---|
| <ul style="list-style-type: none">• 科学上网卡密：<ul style="list-style-type: none">■ 服务器ip: 8.130.55.206■ 端口号port: 429■ 密码: bxx_ss_temp■ 加密方式: aes-256-cfb | <ul style="list-style-type: none">■ 服务器类型: ss■ 名称: bxx_ss (自定义)■ 带宽：1Mbps■ 到期时间：2023.03.16■ 客户端：shadowsocks | <p>注意事项：提供公共服务梯子不是政策支持，请低调使用；梯子为临时的、免费的，带宽很小，大家共用，仅用于注册，请勿用于其他用途。</p> |
|---|---|---|



目前所有公开的对 GPT-3 的复现都失败了，包括但不限于：GPT-3, PaLM, BLOOM, OPT, FLAN-T5/PaLM, HELM 等。“失败”是指训练得出模型有接近 GPT-3 或者更大的参数量，但仍无法与 GPT-3 原始文献中报告的性能所匹配。

主要原因是**训练昂贵、预训练数据和训练策略**等问题。

训练昂贵：一次训练就将需要在约 1000 个 80G A100 GPU 上花费至少 2 个月的时间（数据来自于 OPT 的原始文献）。

预训练数据问题

GPT-3 在共计 300B 的 token 上进行训练，其中 60% 来自经过筛选的 Common Crawl，其它则来自：webtext2（用于训练 GPT-2 的语料库），Books1，Books2、维基百科、Github Code。

每个部分的占比并不与原始数据集的大小成比例，相反的，具有更高质量的数据集被更加频繁地采样。

- 第一点是一个具有良好性能的用于**筛选低质量数据**的分类器。
 - 一些文章表示一个用更少但质量更高的数据集训练的预训练模型，可以在性能上超过另一个用更多的混合质量数据集训练的模型。当然，数据的多样性仍然是十分重要的，应当非常小心地处理在数据多样性和质量之间的权衡。
- 第二点是**预训练数据集的去重**。
 - 去重有助于避免预训练模型多次面对相同的数据后记住它们或者在其上过拟合，因此有助于提高模型的泛化能力
- 第三点是**预训练数据集的多样性**。
 - 包括领域多样性、格式多样性（例如：文本、代码和表格）和语言多样性。



训练策略问题

包括训练框架、训练持续时间、模型架构 / 训练设置、训练过程中的修改。在训练非常大的模型时，它们被用于获得更好的稳定性和收敛性。一般来说，由于未知的原因，预训练过程中广泛观察到损失尖峰 (loss spike) 和无法收敛的情况。

- 训练框架
 - 数据并行(分布式优化器)和模型并行(包括张量并行 (tensor parallel) 、流水线并行 (pipeline parallel) ，有时还包括序列并行 (sequence parallel))
 - bfloat16和float16问题，bfloat16 可以表示更大范围的浮点数，能够处理在损失尖峰时出现的大数值。
- 训练过程中的修改
 - 训练中调整，例如：中途调整并从最近的 checkpoint 重启训练，包括改变截断梯度范数 (clip gradient norm) 和学习率，切换到简单的 SGD 优化器然后回到 Adam，重置动态损失标量 (dynamic loss scalar)，切换到更新版本的 Megatron 等等。
- 训练架构/训练设置
 - 使用 Adafactor 的修改版本作为优化器，缩放在 softmax 之前的输出 logit，使用辅助损失来鼓励 softmax 归一化器接近 0，对词向量和其他层权重使用不同的初始化，在前馈层和层归一化中不使用偏差项，并且在预训练期间不使用 dropout。
- 训练过程：
 - 逐步增大的Batch Size
 - 激活函数的选择，ReLU、SwiGLU、GeLU等
 - 词向量建模方式，RoPR词向量、ALiBi词向量等



低成本复现方案 Colossal-AI

Colossal-AI快速跟进，**首个开源低成本复现ChatGPT完整流程**。作为当下最火热的开源AI大模型解决方案，Colossal-AI已收获开源社区**GitHub Star近万颗**，此次开源亮点包括：

- **开源完整基于PyTorch的ChatGPT复现流程**，涵盖全部3个阶段，可实现从预训练模型到ChatGPT的蜕变；
- 体验最小demo训练流程最低**仅需1.62GB显存**，任意单张消费级GPU即可满足，单卡模型容量最多**提升10.3倍**；
- 相比原生PyTorch，最高可**提升单机训练速度7.73倍**，单卡推理速度1.42倍，**一行代码即可使用**；
- 对于微调任务，可最多**提升单卡的微调模型容量3.7倍**，同时保持**高速运行**，仅需一行代码；
- **提供单卡、单机多卡、1750亿参数等多个版本**，支持从Hugging Face导入OPT，GPT-3，BLOOM等多种预训练大模型；
- 收敛验证正在进行中，该项目也在吸引合作者**共建生态**。

开源地址：

<https://github.com/hpcaitech/ColossalAI>

2月20日讲座中，作者表示LLM大语言模型175B没有完全跑通，后面的两个6B模型能跑通。



特斯拉前AI总监发布NanoGPT，已基于OpenWebText重现 GPT-2 (124M)，在单个8XA100 40GB节点上，训练时间为38小时。

开源地址：<https://github.com/karpathy/nanogpt>

Jaymody发布picoGPT，60行numpy代码实现复现GPT2✅

开源地址：<https://github.com/jaymody/picoGPT>

OpenAI的GPT-2: <https://github.com/openai/gpt-2>

复旦版ChatGPT: <https://txsun1997.github.io/blogs/moss.html>

2023年2月21日发布，冲上热搜，服务器挤爆，引起了很多争议。

Q: 模型是什么？怎么训练的？数据怎么获取的？

A: 暂时不方便透露，我们尽快开源，尽量三月份开源。



OpenAI官网：<https://openai.com/>

ChatGPT官网：<http://ai.com/>

HaiChatGPT：<https://github.com/zhangzhengde0225/HaiChatGPT>

HaiChatGPT网页：ai.ihep.ac.cn(内网)，47.114.37.111(公网)

官网账号注册教程：https://code.ihep.ac.cn/zdzhang/haichatgpt/-/blob/main/docs/reg_tutorial.md

扩展资源：

1. [解读 ChatGPT 背后的技术重点：RLHF、IFT、CoT、红蓝对抗](#)
2. [ChatGPT发展历程、原理、技术架构详解和产业未来（收录于先进AI技术深度解读）](#)
3. ChatGPT API（非官方）：<https://github.com/acheong08/ChatGPT>
4. HaiChatGPT：<https://github.com/zhangzhengde0225/HaiChatGPT>
5. ChatGPT中文调教指南：<https://github.com/PlexPt/awesome-chatgpt-prompts-zh>
6. 复旦版ChatGPT-MOSS: 体验链接：<https://moss.fastnlp.top/>
项目主页：<https://txsun1997.github.io/blogs/moss.html>
7. 首个开源低成本复现方案：<https://github.com/hpcaitech/ColossalAI>
8. NanoGPT：<https://github.com/karpathy/nanogpt>
9. PicoGPT：<https://github.com/jaymody/picoGPT>
10. GPT-2：<https://github.com/openai/gpt-2>
11. MOSS：<https://txsun1997.github.io/blogs/moss.html>



AI won't replace the scientist, but scientists who use AI will replace those who don't.

——Microsoft report "The Future Computed"

院里：发挥高能物理学科基础和优势，打造高水平的数据和AI驱动平台。

发展高能物理AI的优势：有海量数据、有高性能算力

劣势：刚刚起步，目前AI研究以课题组的形式进行，没有形成交叉合作模式

方案：多领域人才进行合作，实现AI4HEP的应用；探索共性技术，推动高能物理与人工智能交叉前沿，在知识-数据协同驱动的第三代人工智能弯道超车。

目标：实现高能物理领域研究和人工智能算法研究相互促进，进而推动社会、经济的发展。

目前任务：搭建高能物理AI平台，形成长效合作机制。



请批评指正！



Backup