

基于RAY的融合计算

在生命科学领域的应用

演讲人 百图生科AI应用负责人 饶星

主题

目录

01

从2024年诺贝尔化学奖谈起

授予三位蛋白质研究科学家

02

加速蛋白质结构预测性能

基于RAY Workflows和异构调度深度优化性能

03

加速蛋白质生成设计性能

基于RAY data和streaming调度深度优化性能

04

融合计算架构

在RAY 基础组件上封装抽象，解决性能和效率问题

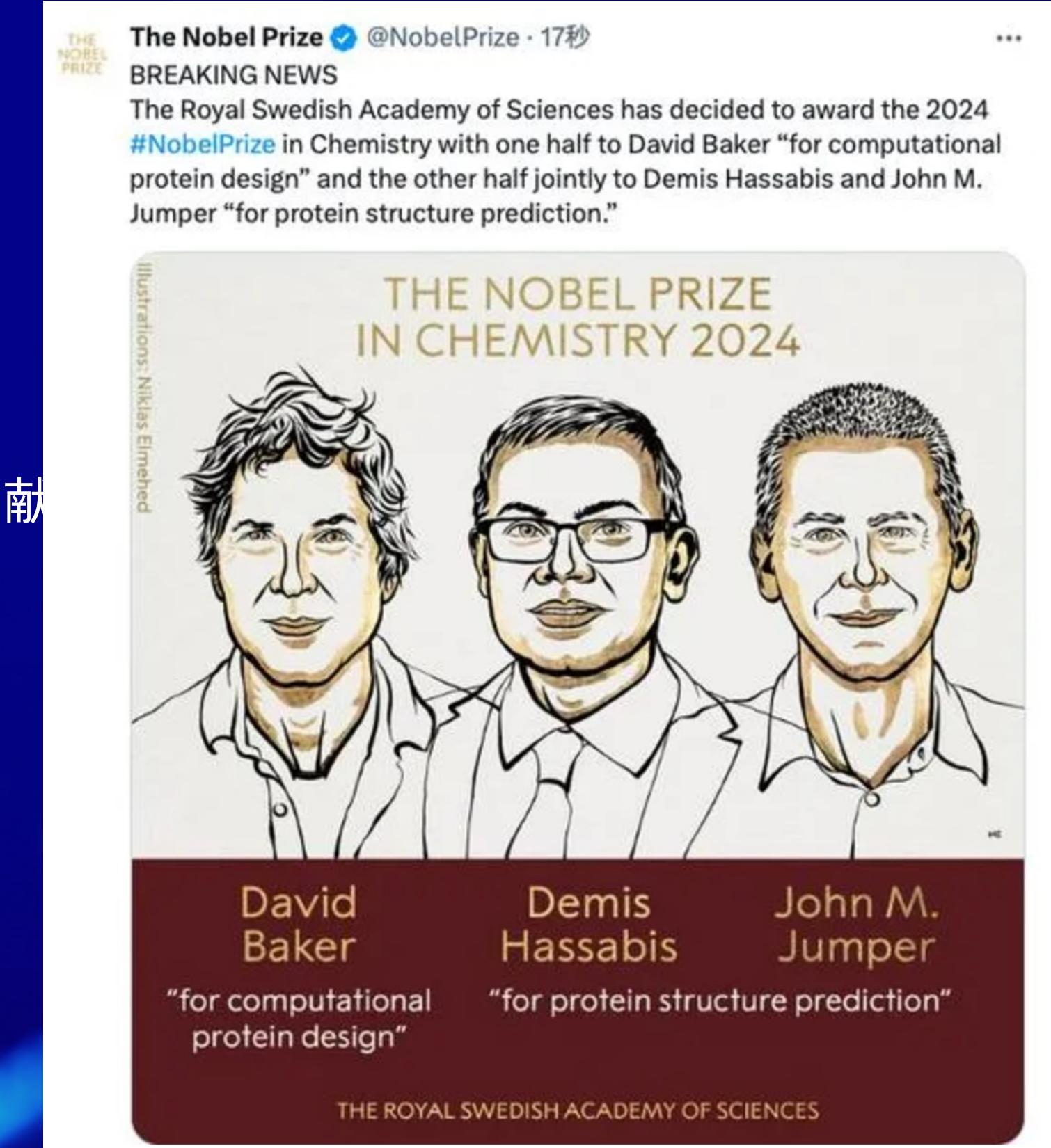
2024年诺贝尔化学奖揭晓

授予三位蛋白质研究科学家

获奖理由

授予David Baker，以表彰其在计算蛋白质设计方面的贡献；
一半则共同授予Demis Hassabis和John M.Jumper，
以表彰其在蛋白质结构预测方面(AlphaFold v1/v2/v3)的贡献。

蛋白质业务价值大，但资源开销也大，性能差。
开源的蛋白质结构模型(AlphaFold v2)预测一个1000AA序列的
复合物，在1卡A100上通常需要**30分钟**；
设计1000个蛋白质序列/结构，在1卡A100上更是需要**数小时**。



RAY FORWARD 2024

驱动AI 连接未来

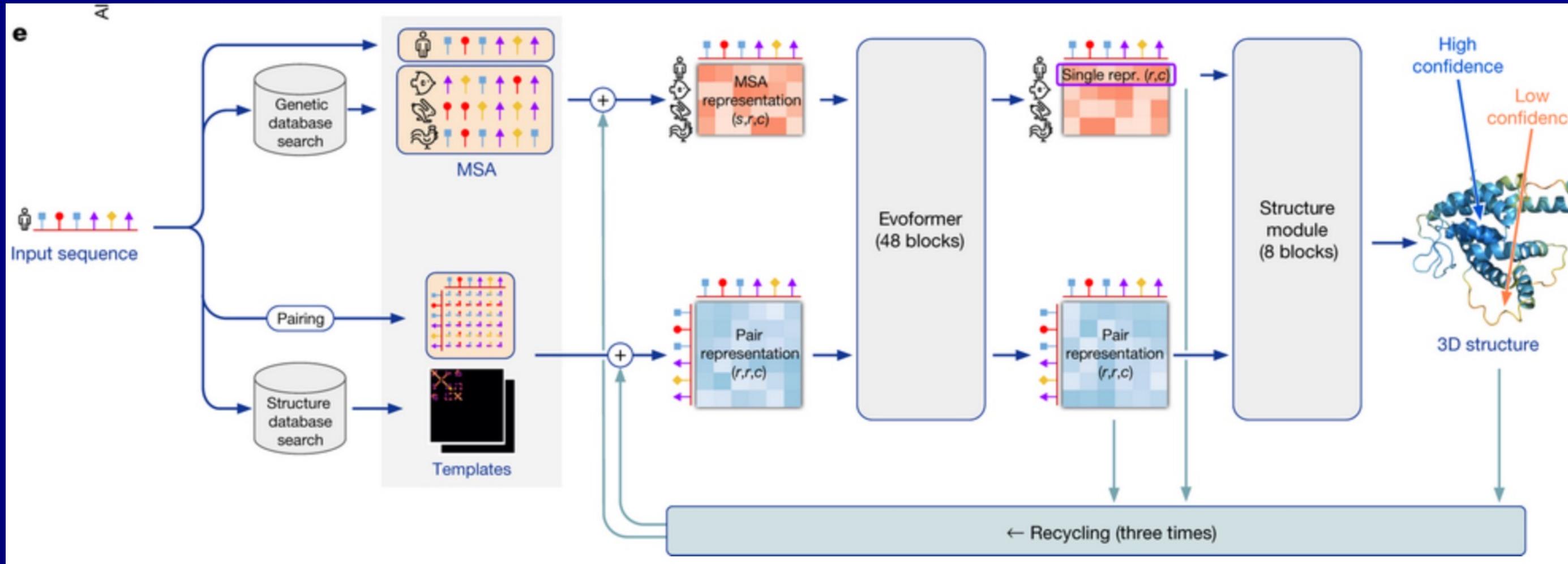
基于RAY Workflows和高效异构调度

加速蛋白质结构预测性能

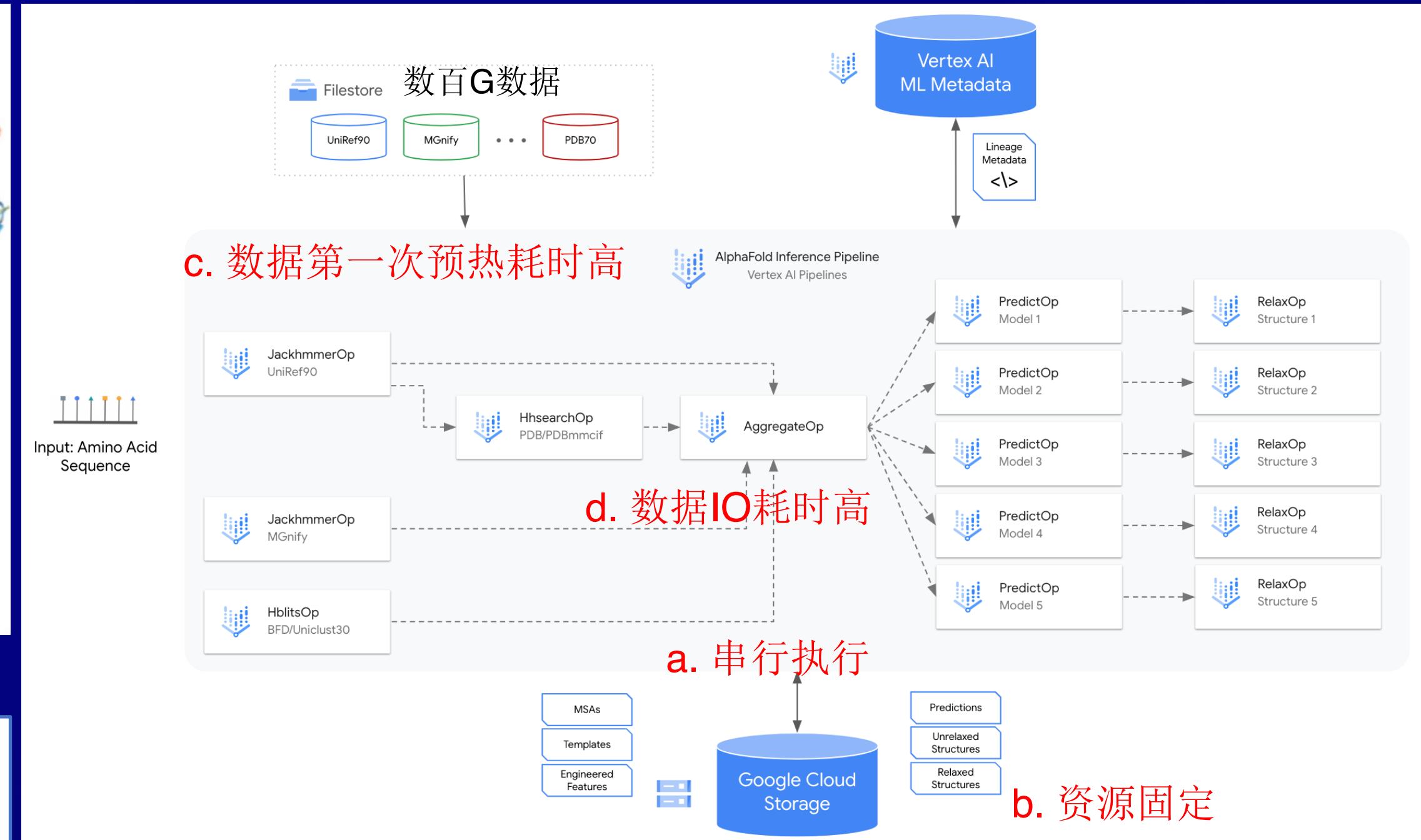
AlphaFold batch inference

AlphaFoldv2.3模型架构介绍和业界一般解决方案

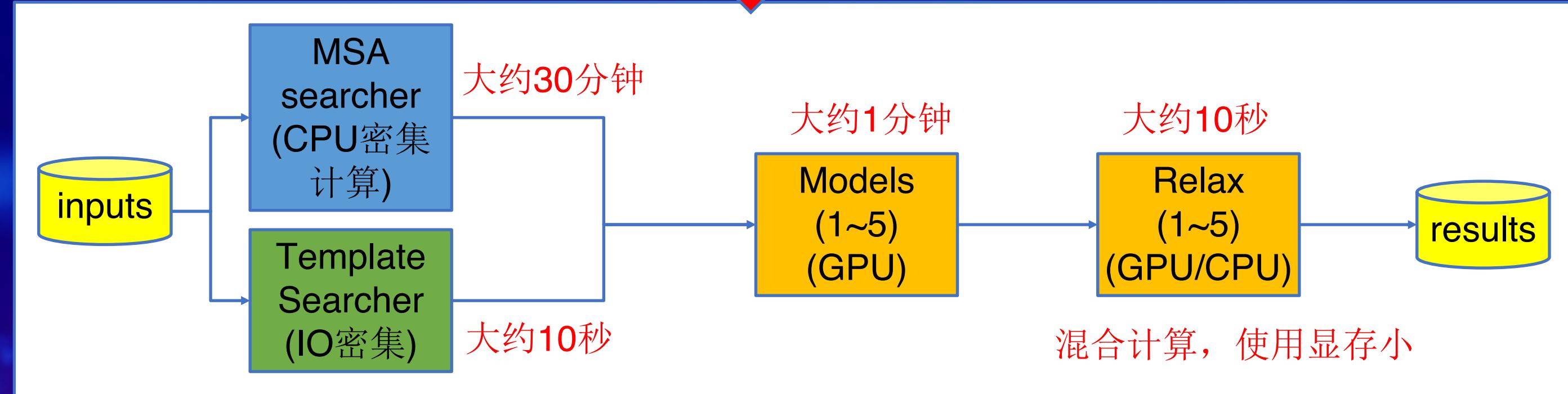
1. AlphaFoldv2.3模型架构：



3. Google云平台解决方案：基于Kubeflow Pipelines



2. 精简后的工程架构：

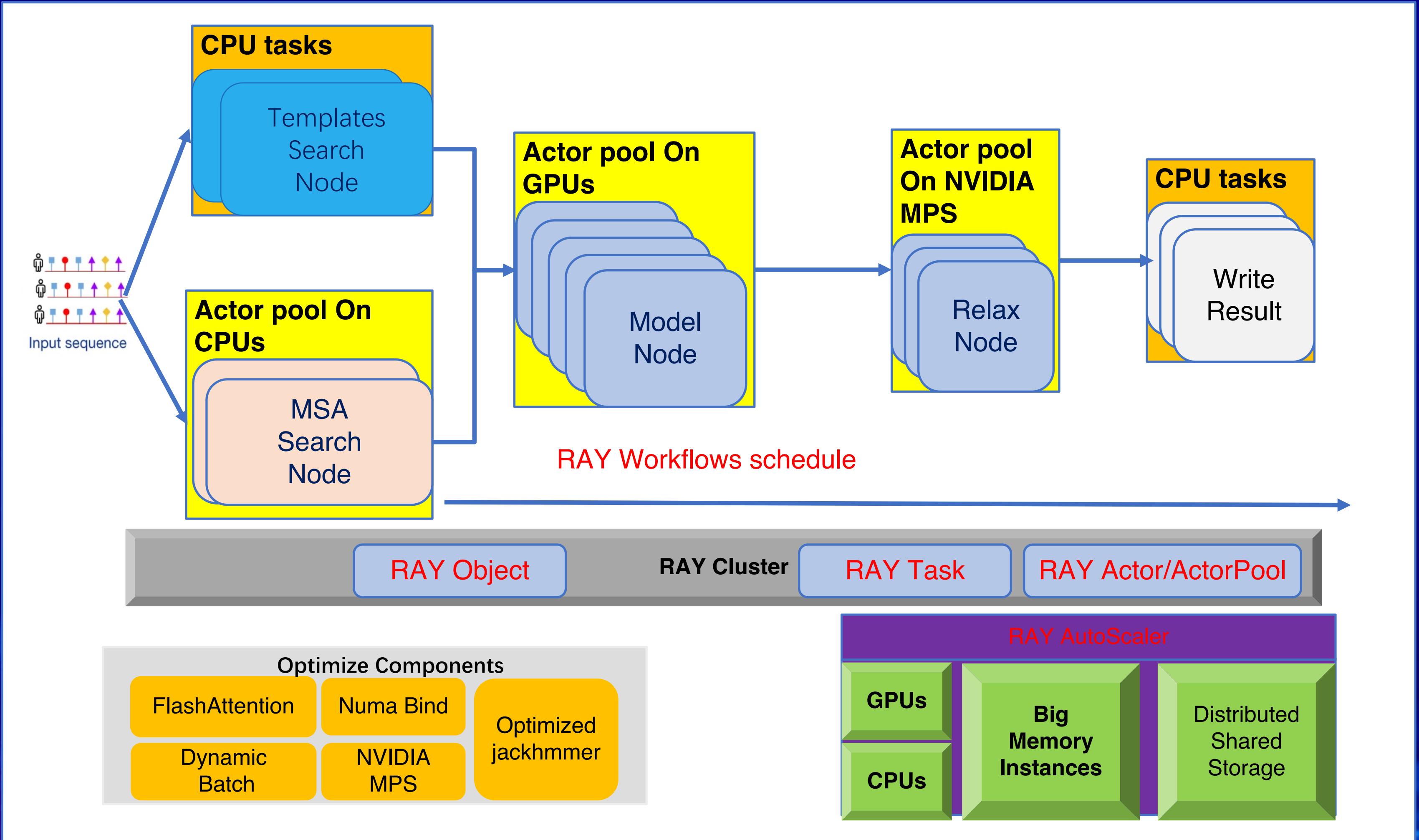


4. 问题：

- 速度慢、吞吐低
- 无弹性能力, 不能充分利用丰富的云资源
- Batch场景：串行执行，batch之间无overlap
- IO负载高, 资源利用率低

AlphaFold batch inference

基于RAY Workflows的方案



核心设计：

- ✓ 采用RAY Workflows高效构图，流式调度
- ✓ 引入RAY ActorPool，不同Node，不同实例数目
- ✓ GPU/CPU node资源动态扩缩容
- ✓ 基于RAY cluster部署，避免集群初始化开销
- ✓ MSA Node提前常驻，避免初始化开销
- ✓ 通过RAY共享object传递数据
- ✓ 集成算子深度优化和MSA优化

收益：

- ✓ 高吞吐，资源利用高
- 不同类型的Node之间，多个Batch之间，充分overlap
- ✓ 运维成本少，弹性扩缩容
- ✓ Python环境友好(接入算子优化成本低)
- ✓ 架构是通用，很容易扩展到自定义/变种模型

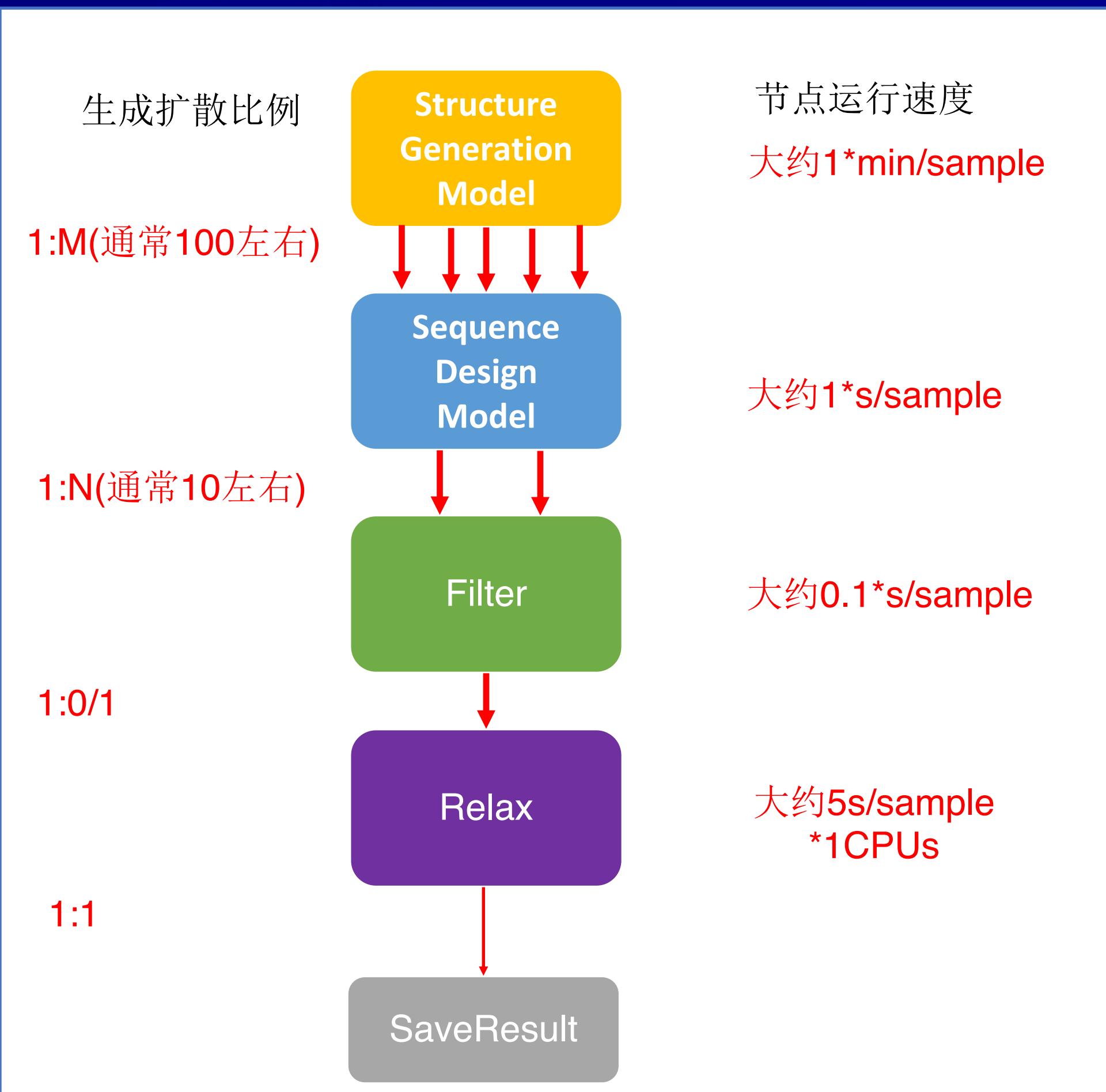
基于RAY data和streaming调度

加速蛋白质生成设计性能

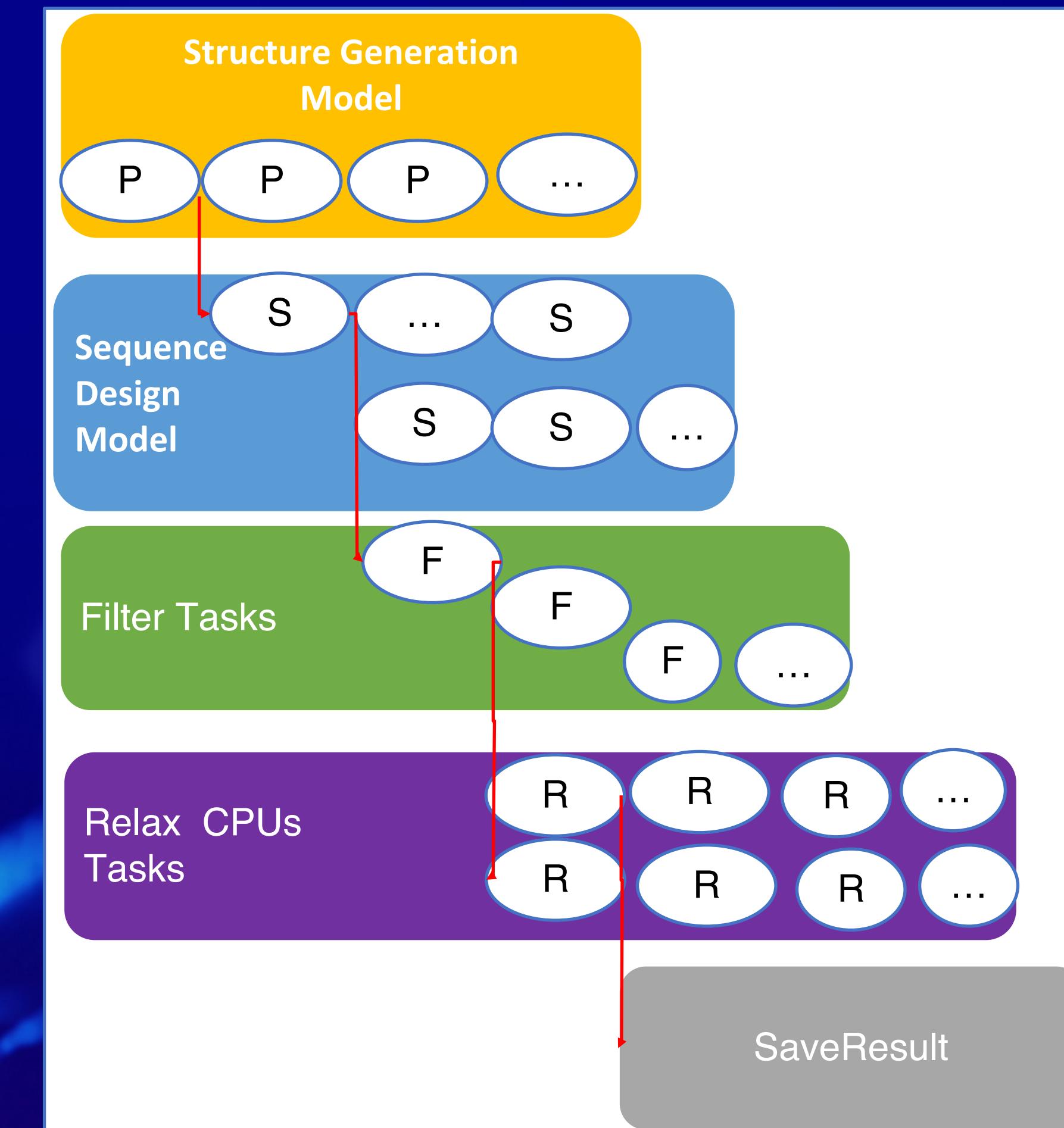
常用的蛋白质设计

基于RAY data简化代码，基于RAY Streaming 调度提高执行效率

1. 常用蛋白质生成的流程图



2. 期望理想中的调度执行



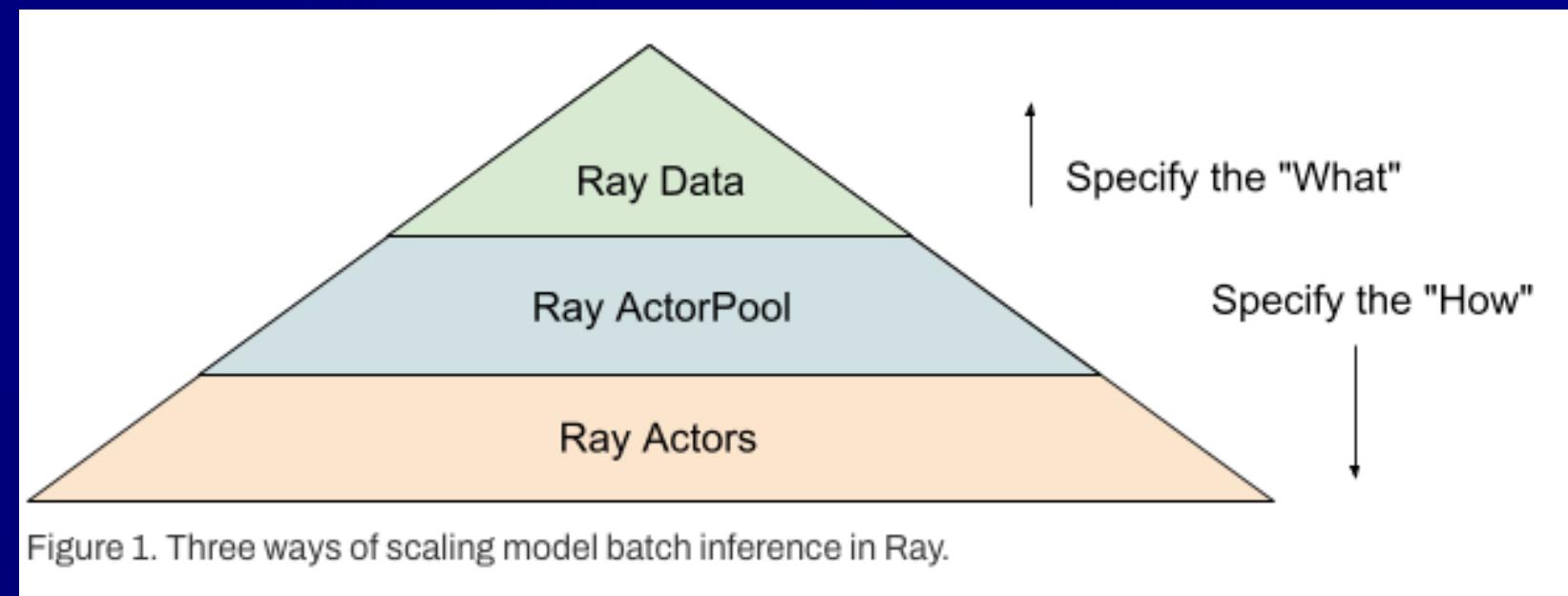
3. 传统自建实现方法，费时费力
- 关注大量分布式通信和调度
 - 考虑各种异常处理
 - 效率问题

常用的蛋白质设计

基于RAY data简化代码，基于RAY Streaming 调度提高执行效率

1. 基于RAY data的示例代码：简单少量代码就能实现复杂的逻辑

RAY data是不错的、简单又高效的执行引擎



```
class StructureGenerationModelPredictor:
    def __call__(self, input):
        pdbs = []
        for pdb in self.model.generate():
            pdbs.append(pdb)
        return pdbs
predictor =
StructureGenerationModelPredictor.options(
    num_gpus=1).remote()
for pdb in predictor(design_pdb):
    ds=ray.data.from_items([pdb])
    .map(SeqDesignModelPredictor)
    .map(filter_task)
    .map(relax_task)
    .map(save_result_task)
    ...
```

2. 结构生成模型流式生成输出

```
class StructureGenerationModelPredictor:
    def __call__(self, input):
        for pdb in self.model.generate():
            # streaming outputs
            yield {
                "item": pdb
            }
```

3. 进一步使用ActorPool提高吞吐

```
.map(SeqDesignModelPredictor,
num_gpus=0.25,
concurrency=(1,10),)
```

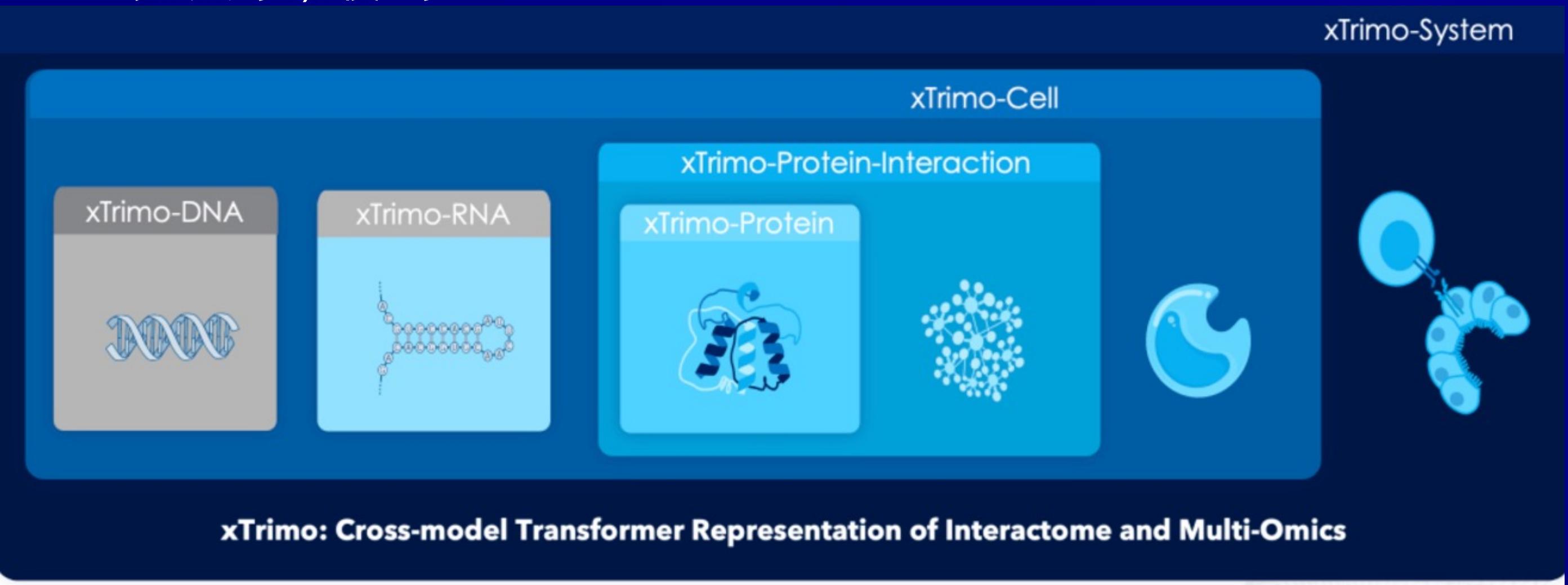
4. 流式分批save结果，避免内存OOM

```
.write_parquet(...)
```

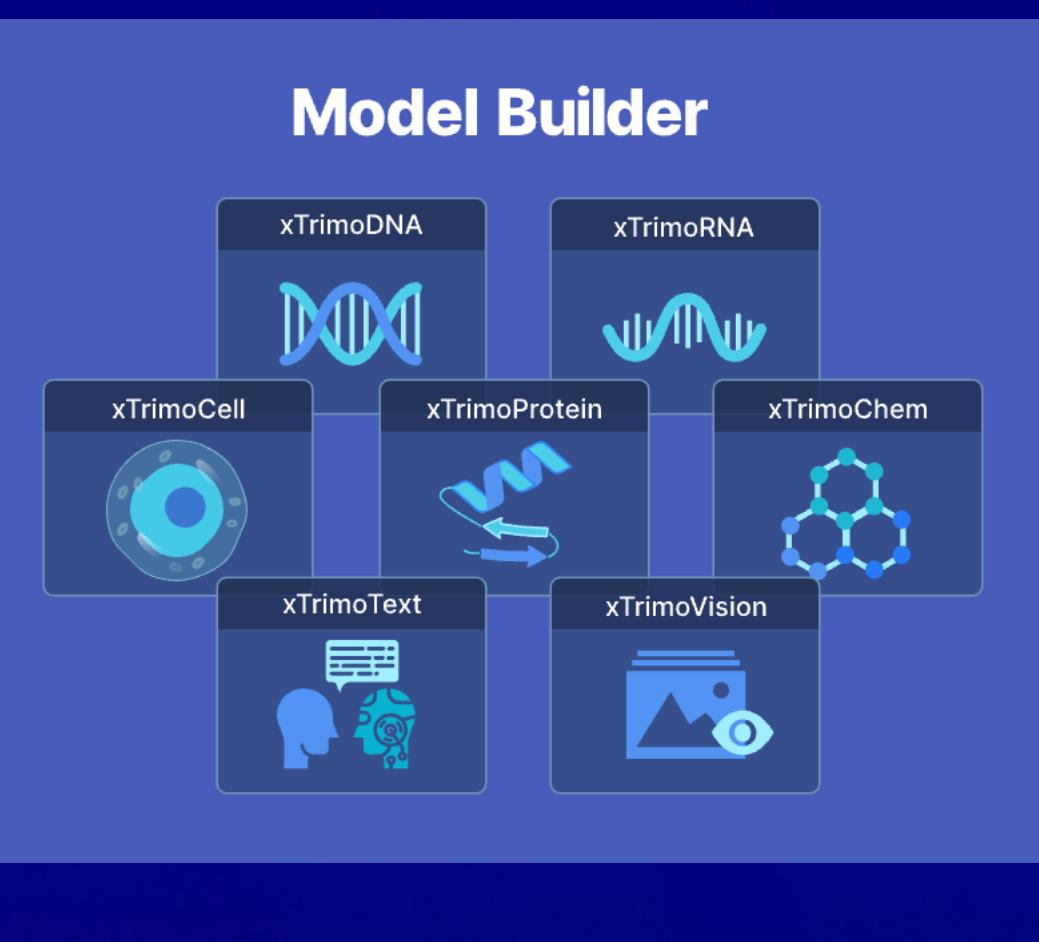
实际生命科学领域业务场景更复杂

多模型ensemble, 需要不同资源类型, 结合传统生信工具

1. 业务场景多, 模型多



2. 有在线服务, 也有批量任务 3. 自定义finetune模型



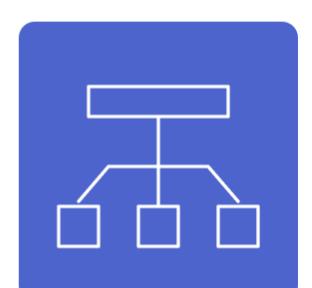
4. 极其复杂的端-端 蛋白设计, 不同场景带来个性化的pipeline

一键生成全新的蛋白序列

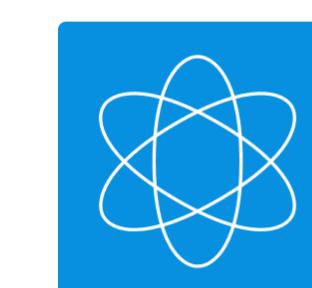
精准设计——利用AI生成具有目标效果的蛋白质



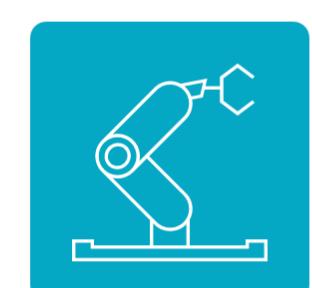
设计目标



生成式平台生成候选蛋白



多目标优化引擎对生成的蛋白质进行评估, 并选择最佳候选分子进行湿实验室验证



通过BioMap超级蛋白质打印机(高通量自动化蛋白质生产平台)或您的湿实验室生产和验证



生物治疗
从头发现治疗候选物、预测抗体和VHH与其靶蛋白的结构、计算亲和力成熟等



合成生物学
增强酶的功能、设计新的催化活动等

带来的挑战：人少事多，还要性能好

✓ 效率问题：人员不足&重复建设

✓ 性能问题：低延迟和高图吞

RAY FORWARD 2024

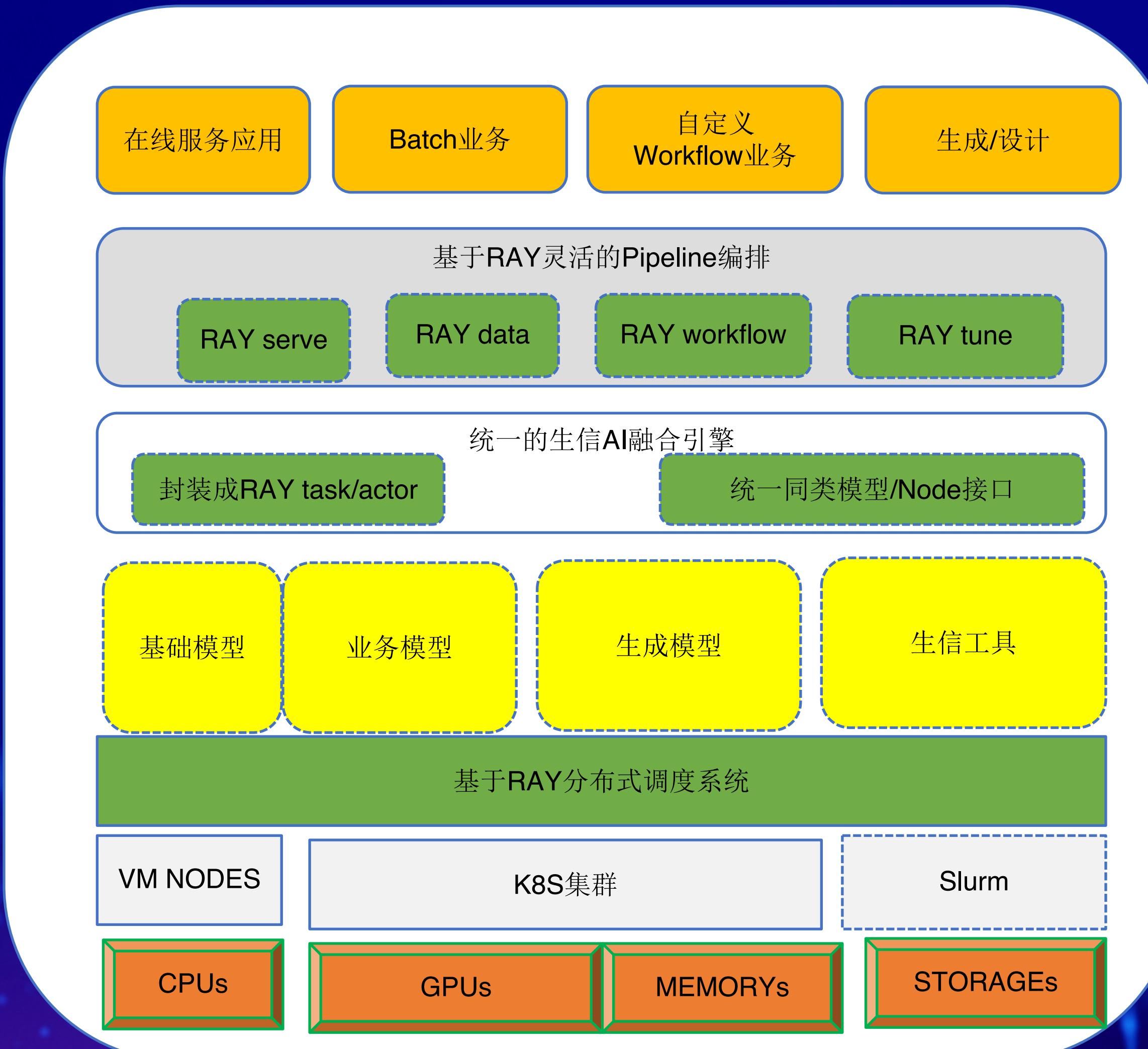
驱动AI 连接未来

基于RAY在生命科学领域场景

融合计算架构

统一Model接口，基于RAY积木化组装模型应用

高效构建，高性能执行，低成本运维



高效构建：

统一编程语言

分门别类，统一接口

统一调度，减少构建pipeline成本

高性能执行：

弹性自动扩缩容

streaming overlap执行

融合单节点、单模型优化

低成本运维：

私有化和集群化方案一致

云原生

关于百图生科



百图生科（**BioMap**）是全球生命科学基础大模型领域的领航者，由李彦宏和刘维在 **2020** 年创立。公司致力于研发和应用前沿 **AI** 与生物技术，助推生命科学领域创新进程；

BioMap 所构建的千亿参数跨模态生物语言大模型 **xTrimo** 平台，在百余个生命科学任务中取得了领先成绩，为行业提供强大的 **AI** 模型构建能力。基于此 **AI** 模型，公司打造的 **AIGP**（**AI Generated Protein**）生成式蛋白质设计平台，已在全球多个前沿药物和合成生物学领域，成功实现全新非天然蛋白质的从头设计。

目前，**BioMap** 已为全球超 **200** 家用户提供服务，涵盖国际药企、龙头 **CDMO**、创新药、合成生物学、绿色科技等领域的知名企业和研究机构。更多信息，敬请访问：<https://www.biomap.com>

Career at BioMap

NLP预训练大模型工程师 查看详情	蛋白结构算法工程师 北京-海淀区 查看详情	大模型训练和应用工程师 查看详情	测试开发工程师-AI方向 北京-海淀区 查看详情
创新战略（高级）经理 北京-海淀区 查看详情	市场实习生 查看详情	生物数据开发工程师（... 查看详情	生物视觉算法工程师 北京-海淀区 查看详情

RAY FORWARD 2024

驱动AI 连接未来

谢谢

Ray 中文社区 × 蚂蚁开源
ANT OPEN SOURCE