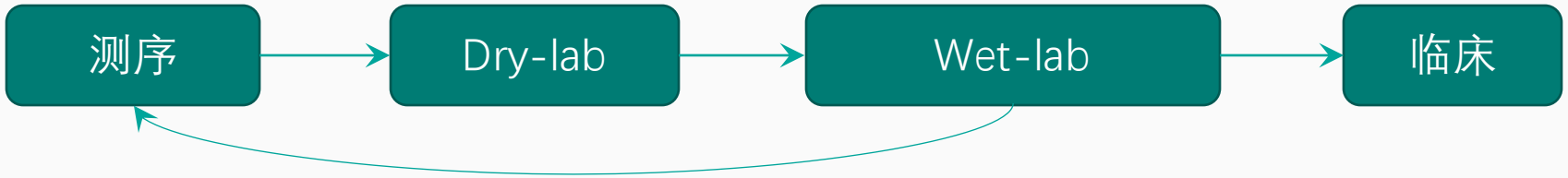
微生物大规模蛋白质功能注释

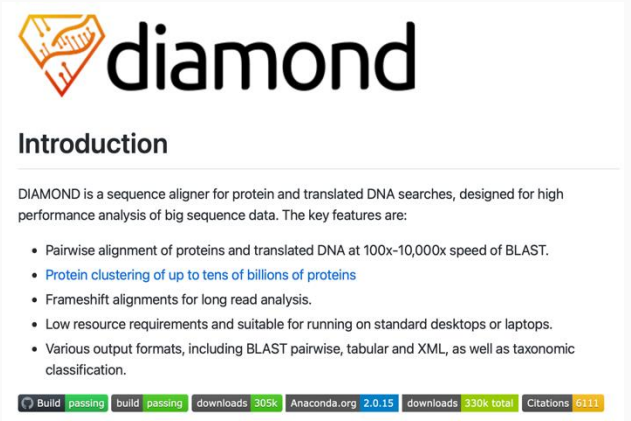
XBIOME | 未知君

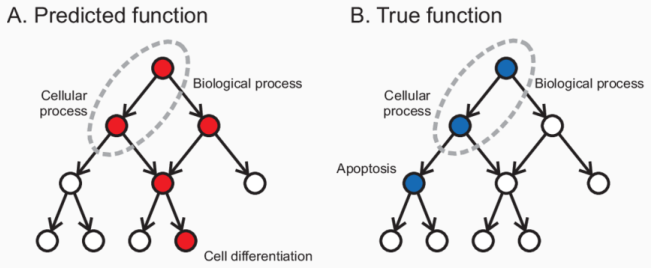
对微生物领域来说，测序是红利，分析是盈利

- 成本指数级下降的测序技术提供了微生物AI的大数据基础



- 在微生物分析中，蛋白功能注释是瓶颈之一
  - 在平均12h样本分析时间中占比20%-30%
  - 不能给出注释结果的序列较多，导致了对微生物理解的*黑洞*
  - 基准方法diamond其原理是基于参考数据库的序列比对
    - 算法复杂度O(m * n), m为样本中待分析的序列个数，n为参考数据库size
    - 阈值截断准则过于简单，导致误报率与漏报率的两难，一般牺牲后者
- 深度学习方案的机会与挑战
  - 机会：快、准、全
    - Reference-free使得算法复杂度降为O(m)，实测时间降低一个数量级
    - 对复杂分界面轻松应对，很好地兼顾了准确率和召回率
  - 挑战
    - 比蛋白功能注释国际比赛CAFA更难：有预测时间的要求，决定了不能使用结构预测或蛋白互作等额外信息
    - **extremely unbalanced hierarchical multi-label classification**





Ontology是生物领域普适现象，其建模方式类同chatGPT均可借鉴经典计算机理论

- 蛋白功能注释可采用gene ontology作为分类体系
  - 本质上是有向无环图（DAG）
  - 其建模思路可借鉴编译原理中的**偏序关系(Partial Order)**

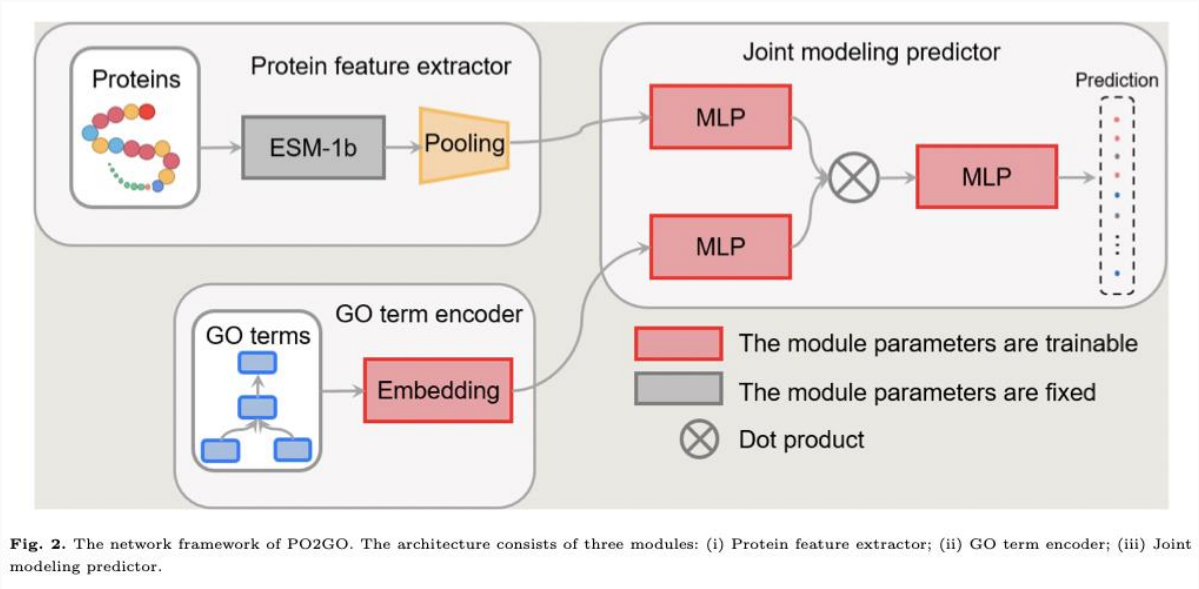


Fig. 2. The network framework of PO2GO. The architecture consists of three modules: (i) Protein feature extractor; (ii) GO term encoder; (iii) Joint modeling predictor.

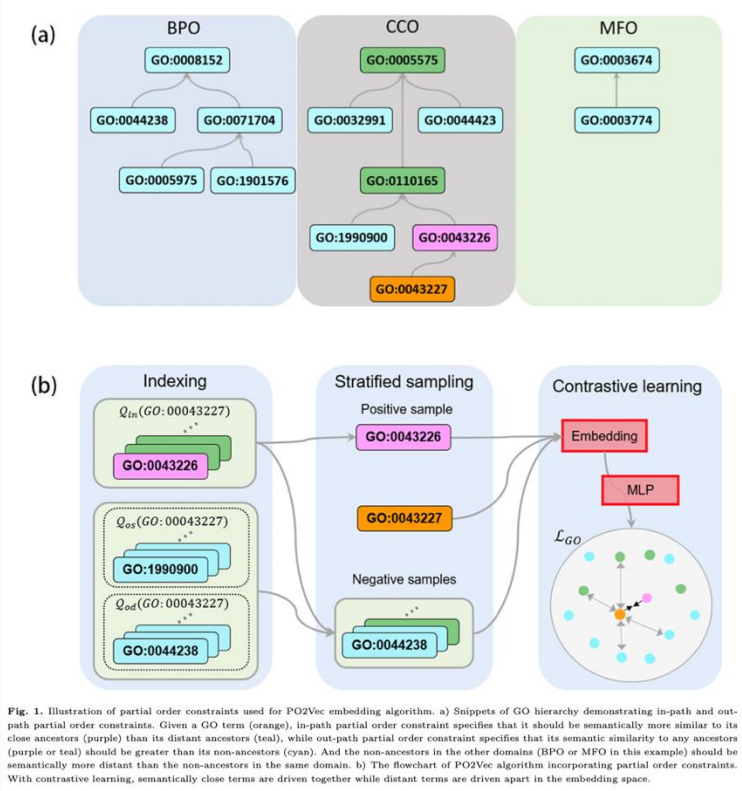


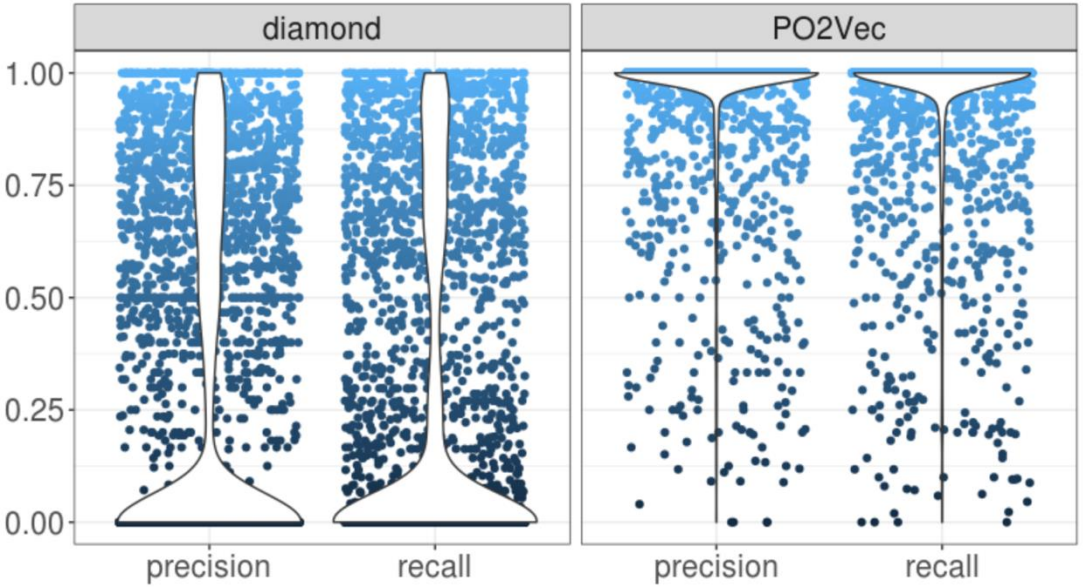
Fig. 1. Illustration of partial order constraints used for PO2Vec embedding algorithm. a) Snippets of GO hierarchy demonstrating in-path and out-path partial order constraints. Given a GO term (orange), in-path partial order constraint specifies that it should be semantically more similar to its close ancestors (purple) than its distant ancestors (teal), while out-path partial order constraint specifies that its semantic similarity to any ancestors (purple or teal) should be greater than its non-ancestors (cyan). And the non-ancestors in the other domains (BPO or MFO in this example) should be semantically more distant than the non-ancestors in the same domain. b) The flowchart of PO2Vec algorithm incorporating partial order constraints. With contrastive learning, semantically close terms are driven together while distant terms are driven apart in the embedding space.

- Ontology是生物领域普适现象，我们方法论具有高迁移性
  - genotype–phenotype association prediction
  - drug–target prediction
  - protein–protein interaction prediction
  - gene–disease association prediction

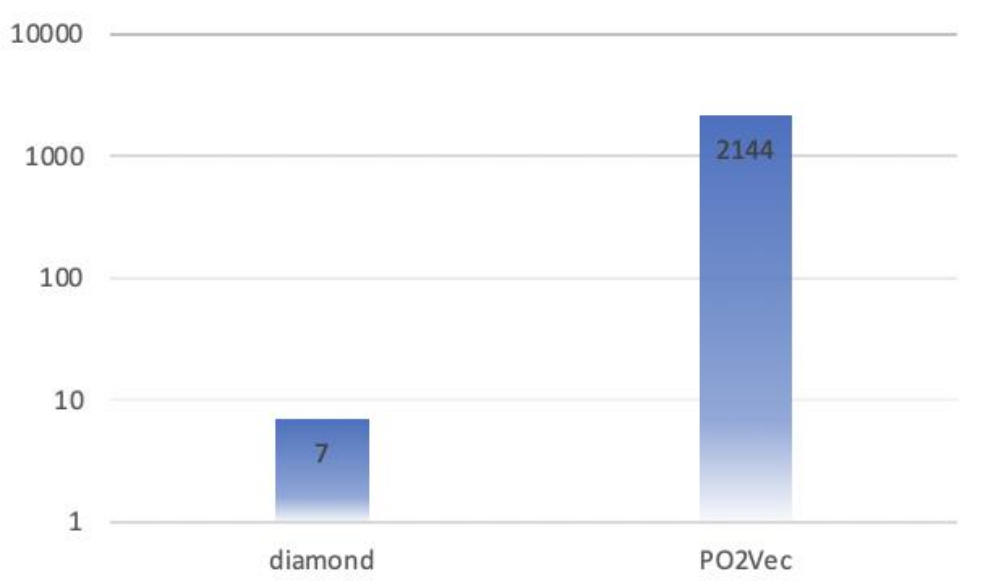
在测试数据集上我们方法全方位超越diamond和其他方法

Table 6. The performance evaluation of different methods for protein function prediction on the CAFA3 dataset

Methods	F_{max}			S_{min}			AUPR		
	BPO	MFO	CCO	BPO	MFO	CCO	BPO	MFO	CCO
Diamond	0.394	0.542	0.57	23.98	8.86	10.86	0.277	0.463	0.514
DeepGOPlus	0.451	0.439	0.584	23.57	9.941	11.2	0.354	0.454	0.603
ESM-1b&DeepGOA	0.431	0.537	0.625	23.33	9.057	10.56	0.379	0.543	0.665
ESM-1b&TALE	0.471	0.592	0.63	23.79	8.095	10.29	0.381	0.611	0.665
PO2GO	**0.523**	**0.614**	**0.646**	**21.32**	**7.739**	**9.895**	**0.44**	**0.63**	**0.693**



我们方法在准确率和召回率上均大幅度超过diamond



- 对单菌Megasphaera的2144个基因实现全覆盖，避免注释黑洞
- 针对实验上发现的吲哚类通路，实现4/6通路蛋白召回（diamond零召回）

潜在应用广泛：基因功能注释是面向肠道菌制药和合成生物学的基础模块

- 基因工程菌&肠道微生物biomarker诊断
  - 了解who does what是其先决条件
- 合成生物学
  - 酶工程需要精准的功能注释工具
- 提供优质的注释数据库有助于推动行业进展

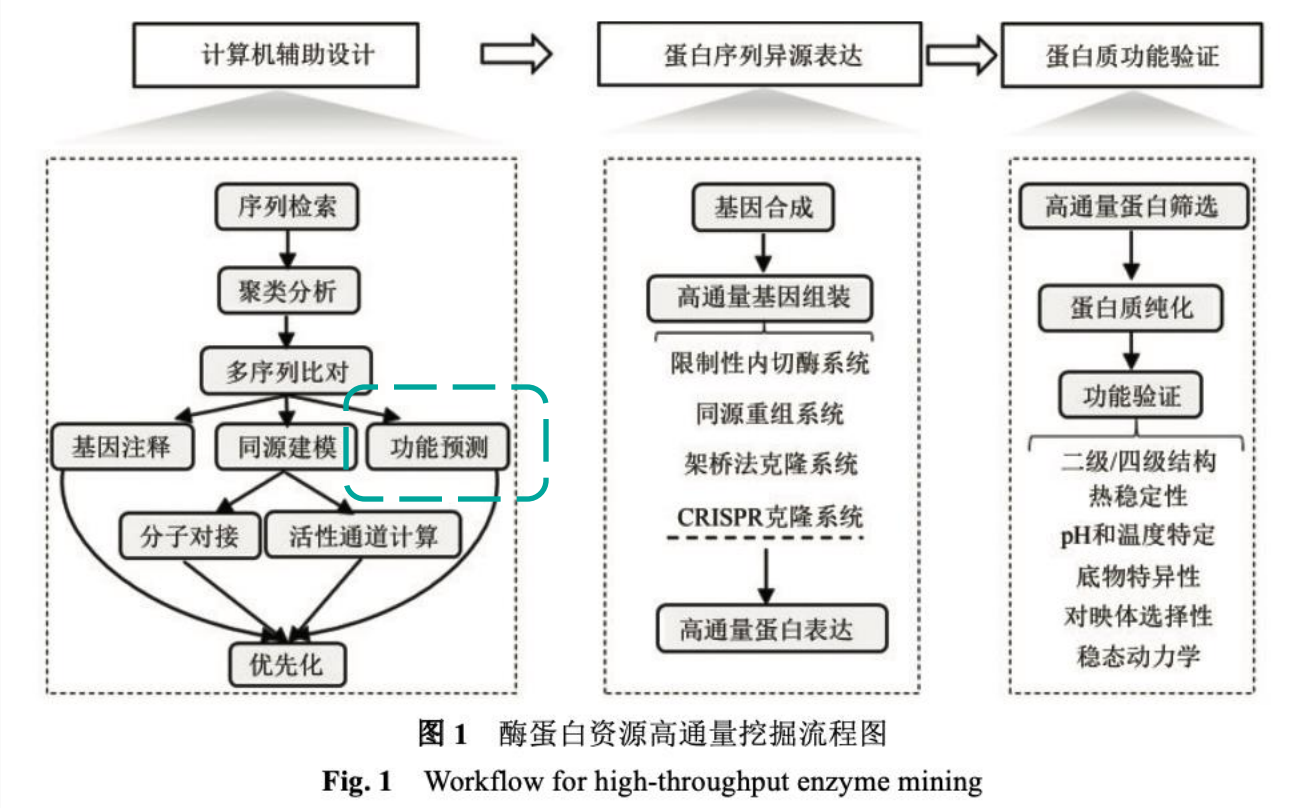


图 1 酶蛋白资源高通量挖掘流程图

Fig. 1 Workflow for high-throughput enzyme mining

司同等人 . 基于合成生物学策略的酶蛋白元件规模化挖掘[J]. 合成生物学，2020，1(3):319-336

XBIOME | 未知君

感 谢 聆 听

附录：Bold Goals for U.S. Biotechnology and Biomanufacturing (2023-03-23)

报告对于大健康领域主要包括，

- **生物制造供应链：**五年内，**至少四分之一**的小分子药物、抗生素等将实现美国本土化生产；五年内将10种常见治疗药物的生产速度**提高10倍**（具体药物没有披露）。
- **人类大健康：**扩大人工智能和机器学习在研发、生产等领域的应用，提高细胞疗法的制造规模，并在20年内将细胞基因疗法制造**成本降低10倍**。
- **推动跨领域发展：**五年内，对**100万种微生物**物种的基因组进行测序，并了解**至少80%**新发现基因的功能。


THE WHITE HOUSE

Administration Priorities The Record Briefing Room Español

MARCH 22, 2023

FACT SHEET: Biden-Harris Administration Announces New Bold Goals and Priorities to Advance American Biotechnology and Biomanufacturing

OSTP › BRIEFING ROOM › PRESS RELEASES

Today, the Biden-Harris Administration is announcing new bold goals and priorities that will catalyze action inside and outside of government to advance American biotechnology and biomanufacturing.

“美国已经意识到并且开始非常重视微生物资源，微生物资源是很重要的自然资源，在健康，医疗，农业上都会有大的应用，国家之间的竞争也会体现在对微生物资源利用能力的竞争上”

附录：赵惠民团队最新Science工作

主要技术特色

- 首次采用对比学习框架，替代传统的局部比对法
- 针对难解酶类的功能预测精度达到 86.7% 以上

与我们方法的异同

- 均使用对比学习技术，赵教授是针对序列做对比学习，我们是针对GO做对比学习
- 我们的方法是普适的，可以针对非酶蛋白进行功能预测，赵教授工作是针对酶蛋白进行注释