“Data-Centric-AI”助力材料科学

杜一

2023年4月

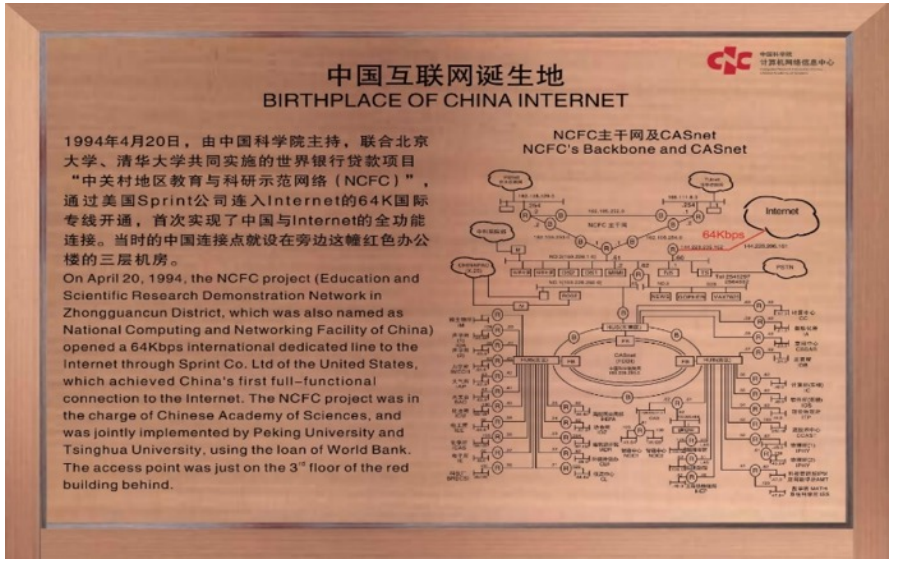
中国科学院计算机网络信息中心

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

计算机网络信息中心

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences

中国互联网诞生地，1994年4月20日，一条64K的国际专线从中心接入Internet，实现了中国与Internet的全功能连接，从此中国成为第77个拥有全功能Internet的国家。

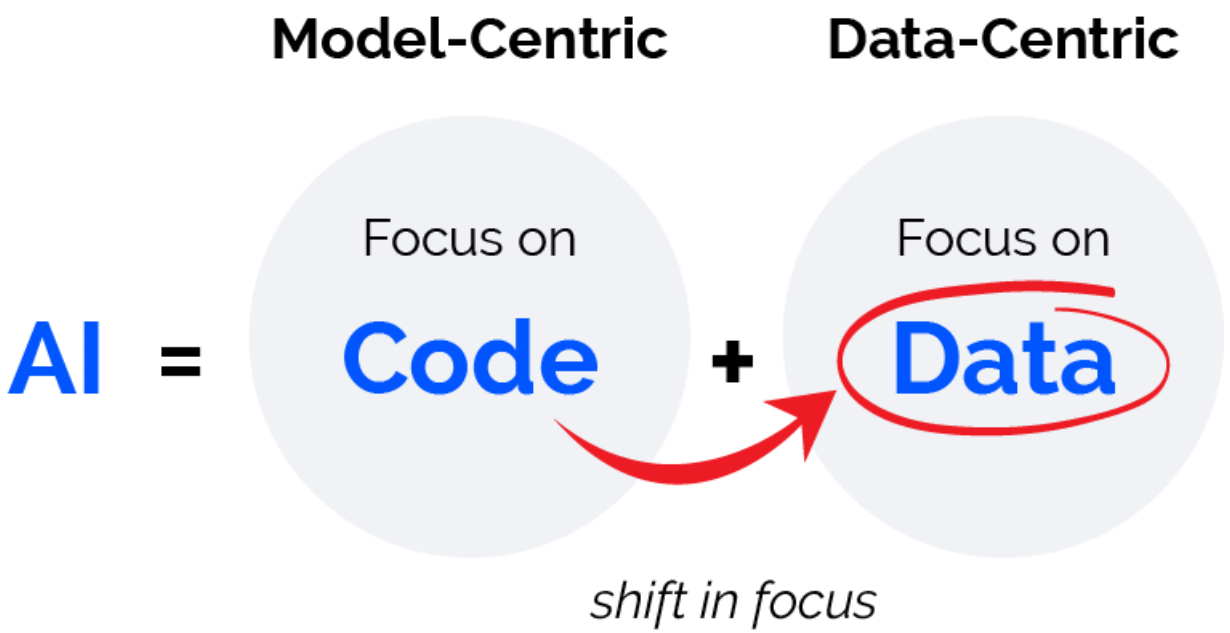


服务1244国家重点研发计划项目数据汇交

新能源汽车、纳米科技、战略性先进电子材料、材料技术提升与产业化、增材制造与激光制造等专项

数据为中心的人工智能(Data-Centric-AI)

- Data-Centric AI (DCAI) 是一个新兴研究方向，主要研究如何通过改进数据集(质量、数量等)来提升机器学习应用的效果

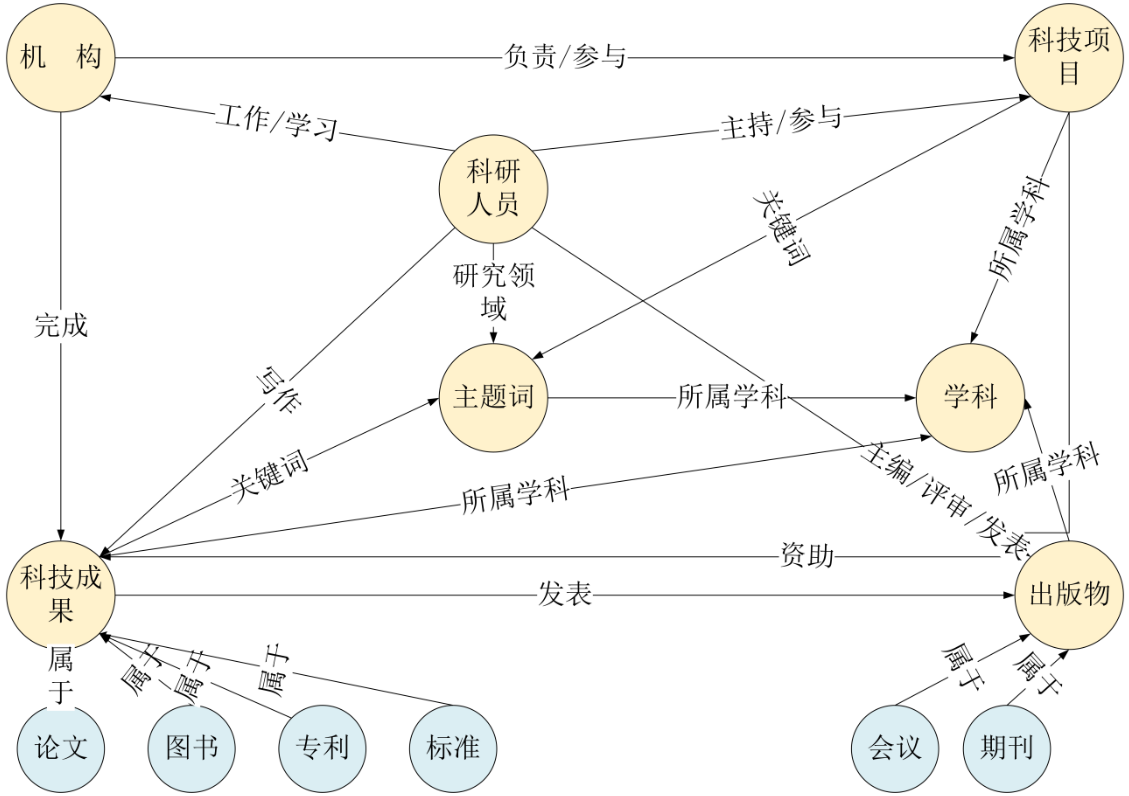


Computer vision task (steel sheet inspection)	Accuracy
Baseline	76.2%
Model-Centric	+0%
Data-Centric	+16.9% (93.1%)

通过将研究的重点从模型算法转向数据，可以大幅提升已有模型的效率

知识图谱-DCAI的一个很好的抓手

- 知识图谱（Knowledge Graph）：一种**大规模语义网络**，以实体语义为核心，能够提供从关系的角度分析问题的能力。



通过机器学习算法，利用科技领域知识图谱(项目为中心)，辅助发现具有交叉特征的项目/成果

Discipline #1 : C060703 (Biological Data Integration and Biological Big Data)

Discipline #2 : F0305 (Biological and Medical Information Systems and Technology)

Title: Genetic Variation Driven 遗传变异驱动的肿瘤耐药相关非编码RNA深度挖掘算法及在肺癌中的功能研究 (Drug Resistance; Encode; Deep Mining Algorithm; Lung Cancer)

Keywords: 人非编码RNA肿瘤化疗耐药深度挖掘算法生物大数据 (Encode; Deep Mining Algorithm; Bio Big Data)

Abstract: 肿瘤细胞对化疗药物产生耐药性是肿瘤治疗失败的重要原因长链非编码RNA lncRNA可以参与肿瘤耐药调控网络为挖掘关键治疗靶点和解析肿瘤耐药机制提供了新的机遇但利用多组学肿瘤大数据系统预测肿瘤耐药相关lncRNA遗传变异并将其用于肿瘤病人的化疗用药指导仍是挑战本项目在前期工作的基础上将大数据分析算法开发和实验验证相结合系统整合肿瘤测序数据基因组注释数据肿瘤组学特征和临床用药数据构建完整的肿瘤化疗耐药 (Mining; Algorithm; Genetic Variation)

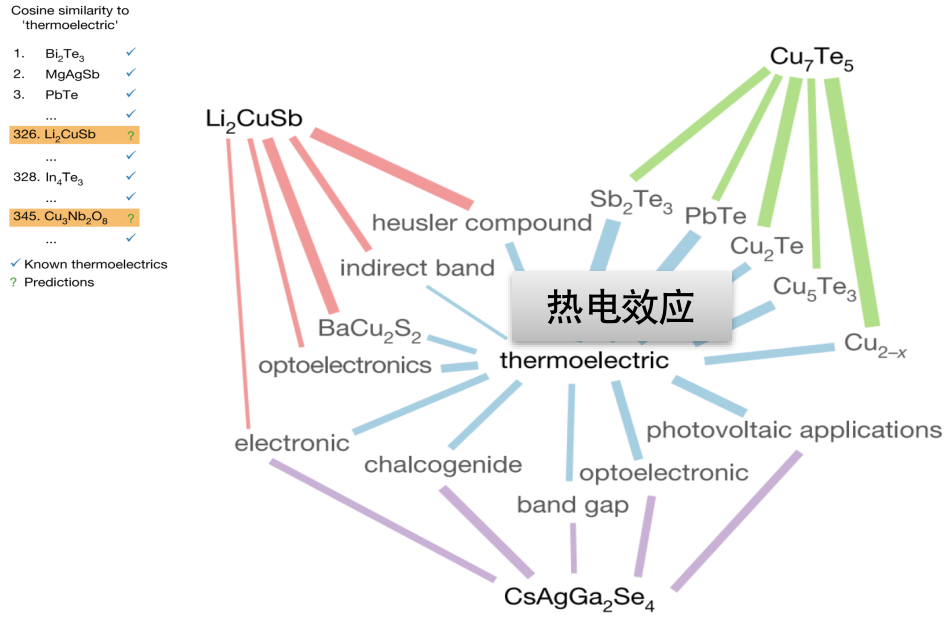
Research of Field: 生物数据与信息挖掘与共享 (Bio data; Information Mining)

1. Meng Xiao, Ziyue Qiao, Yanjie Fu, Yi Du*, Pengyang Wang, Yuanchun Zhou. Expert knowledge-guided length-variant hierarchical label generation for proposal classification. *IEEE International Conference on Data Mining (ICDM).2021*
2. Meng Xiao, Ziyue Qiao, Yanjie Fu, Hao Dong,Yi Du*, Pengyang Wang, Hui Xiong, and Yuanchun Zhou*. Hierarchical Interdisciplinary Topic Detection Model for Research Proposal Classification. *IEEE Transactions on Knowledge and Data Engineering 2023*

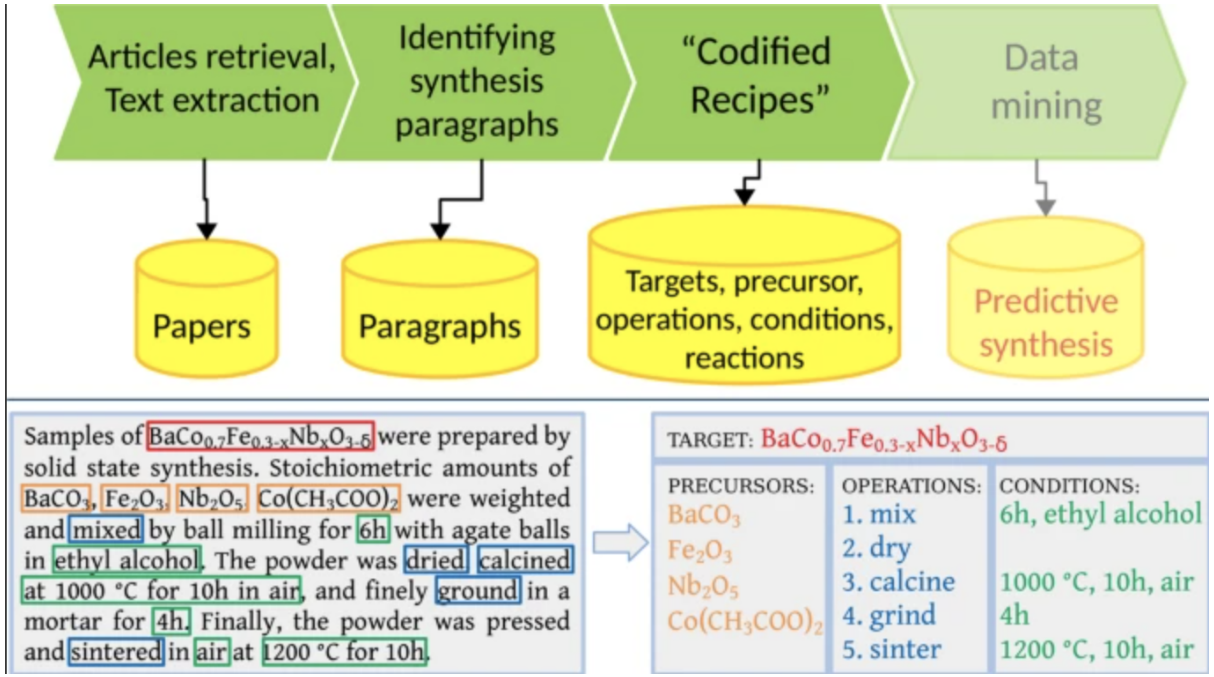
材料科学:源起

利用海量领域知识(1000本刊的330万摘要)构建知识网络，辅助新材料发现:

- underlying structure of the periodic table and structure– property relationships in materials.
- recommend materials for functional applications several years before discovery.



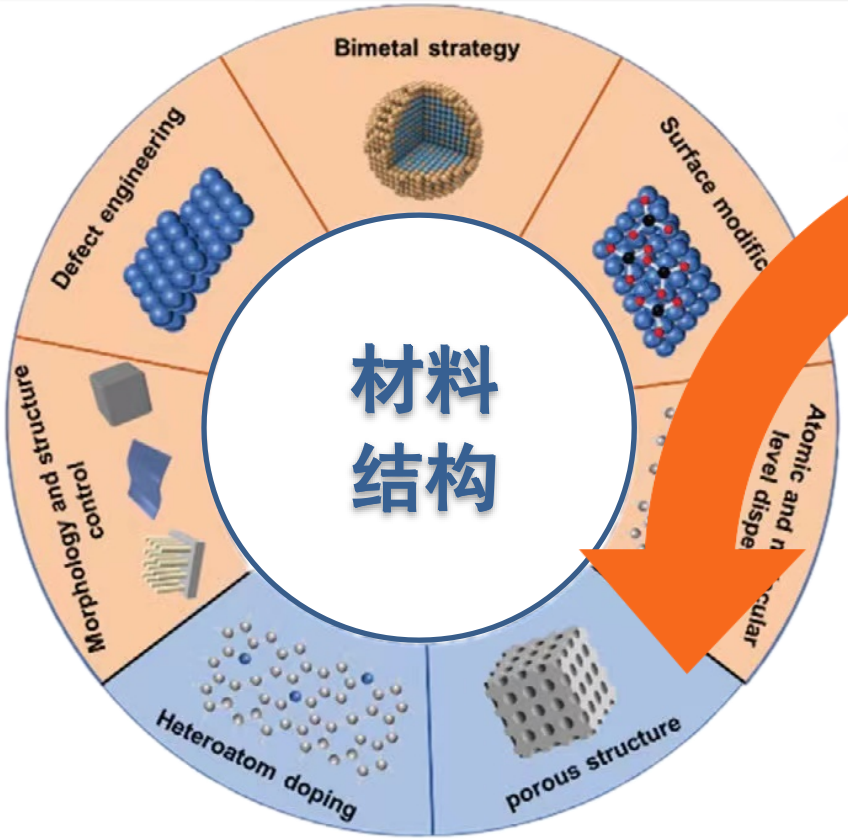
从论文中挖掘precursor materials(前体材料)，并辅助推荐新目标材料的前体材料



1. Tshitoyan, V., etc. (2019). Unsupervised word embeddings capture latent knowledge from materials science literature. Nature, 571(7763), 95–98.
2. He, T., Sun, W., etc. (2020). Similarity of Precursors in Solid-State Synthesis as Text-Mined from Scientific Literature. Chemistry of Materials, 32(18), 7861–7873.

契机:基于领域知识图谱的光电催化材料挖掘软件

中国科学院
计算机网络信息中心
Computer Network Information Center,
Chinese Academy of Sciences



计算/测量

科技文献数据
模拟计算数据

材料性质
价带结构、载流子浓度 … …

测量

材料功能
选择性、量子效率 … …

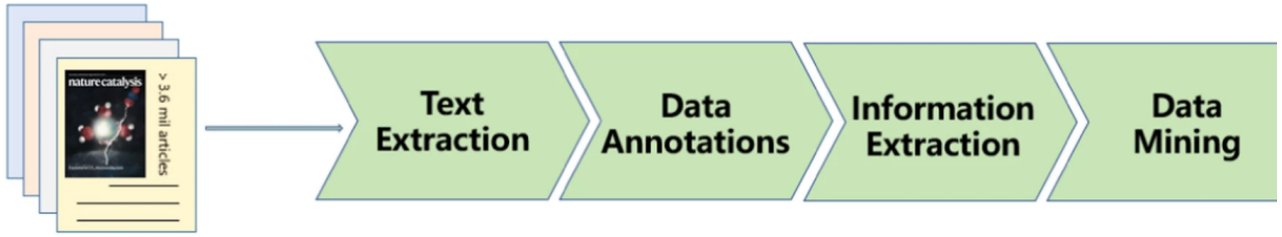
发现

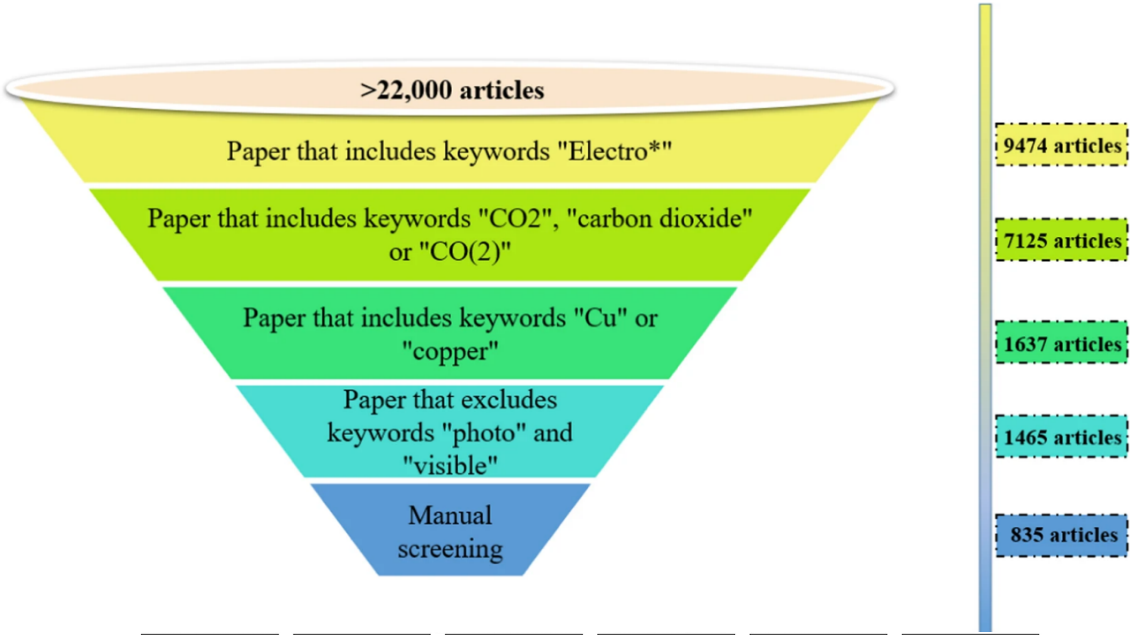
- 数据库：ICSD(无机晶体结构/德国)、MAGNDATA(磁结构/西班牙)、AFLOW(金属材料/美国)、Springer Material(通用/德国)、Material Project(电池材料/美国)
- 软件：VASP(商业/奥地利)、SM Search(商业/德国)、Materials Studio(商业/美国)、CP2K(德国)

根据材料功能直接预测材料的结构

近期进展1：Cu基CO2还原催化数据集构建

针对科学问题，设计高效的数据集制作方法，并结合“人工标注+自动抽取+人工校对”，形成可用于“下游任务”的高质量数据集



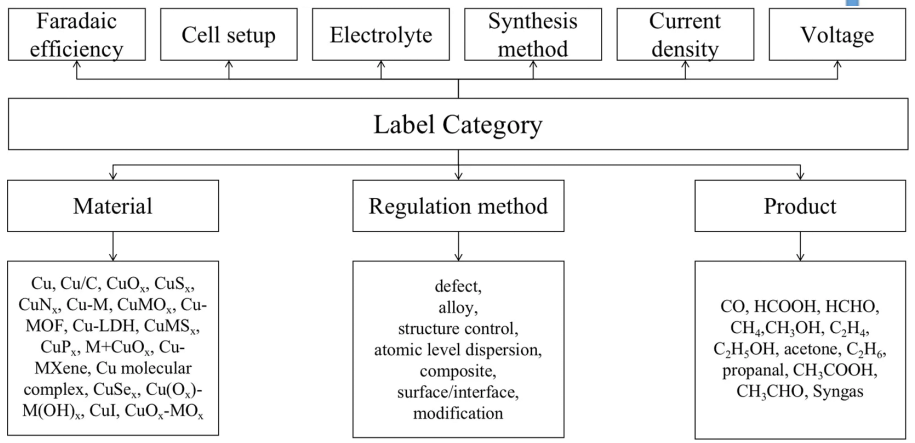


Herein, an elaborate Au-Cu catalyst where an alloyed AuCu shell caps on a Cu core is developed and evaluated for CO2-to-CO electrochemical conversion. Specific roles of Cu and Au for CO2RR are revealed in the alloyed core-shell structure, respectively, and a compositional-dependent volcano-plot is disclosed for the Cu@AuCu catalysts toward selective CO production. As a result, the Au-2-Cu-8 alloyed core-shell catalyst (only 17% Au content) achieves an FECO value as high as **94% and an MA(CO) of 439 mA/mg(Au) at -0.8 V (vs RHE)**, superior to the values for pure Au, reflecting its high noble metal utilization efficiency.

Entity: Au-Cu
Fine-grained Type: Cu-M
Coarse-grained Type: Material

Entity: alloyed
Fine-grained Type: alloy
Coarse-grained Type: Method

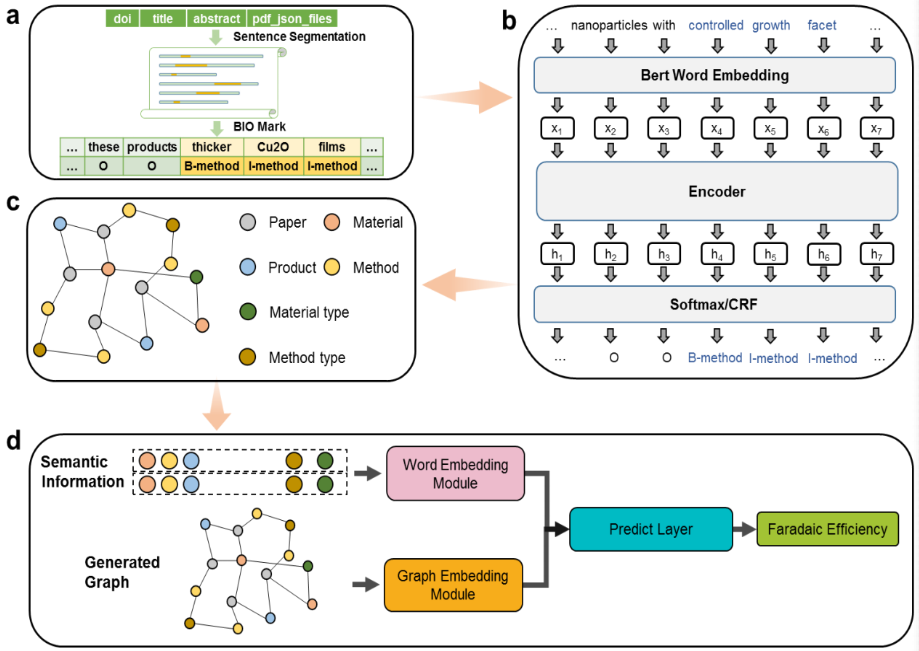
Entity: 94% ... -0.8 V (vs RHE)
Fine-grained Type: Faradaic Efficiency
Coarse-grained Type: Faradaic Efficiency

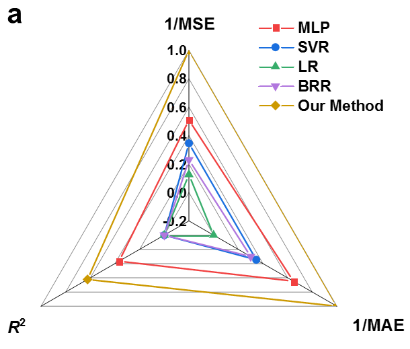


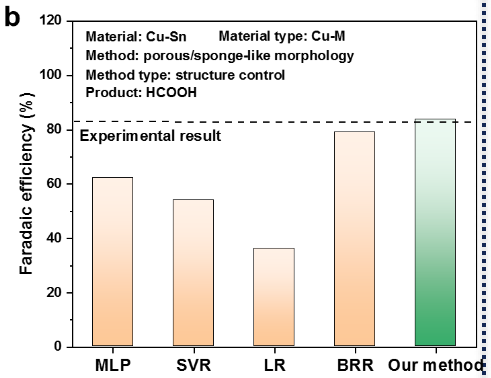
1. Wang, L., Gao, Y., Chen, X., Cui, W., Zhou, Y., Luo, X., Xu, S.,Du, Y.*, Wang, B.*. A corpus of CO2 electrocatalytic reduction process extracted from the scientific literature. Sci Data 10, 175 (2023).

近期进展2：Cu基CO2还原催化预测模型与算法

在Cu基CO2还原催化剂方面，已对小样本下的知识抽取方法进行了初步探索，知识抽取效果超出Bert、Sci-Bert等通用领域知识抽取方法，推荐出的材料法拉第效率预测精度与效果均高于已有推荐方法。

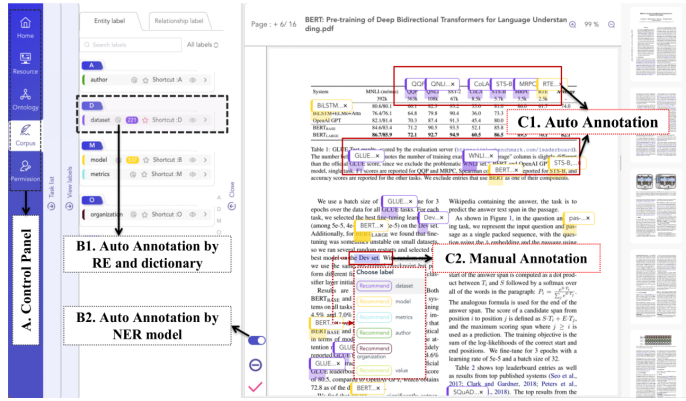






Methods	MSE	MAE	R^2
Multi-Layer Perceptron	0.040	0.171	0.367
Support Vector Regression	0.065	0.226	-0.033
Linear Regression	0.172	0.358	-1.733
Bayesian Ridge Regression	0.099	0.237	-0.570
Our Method	**0.023(-42.5%)**	**0.132(-22.8%)**	**0.623(+69.7%)**

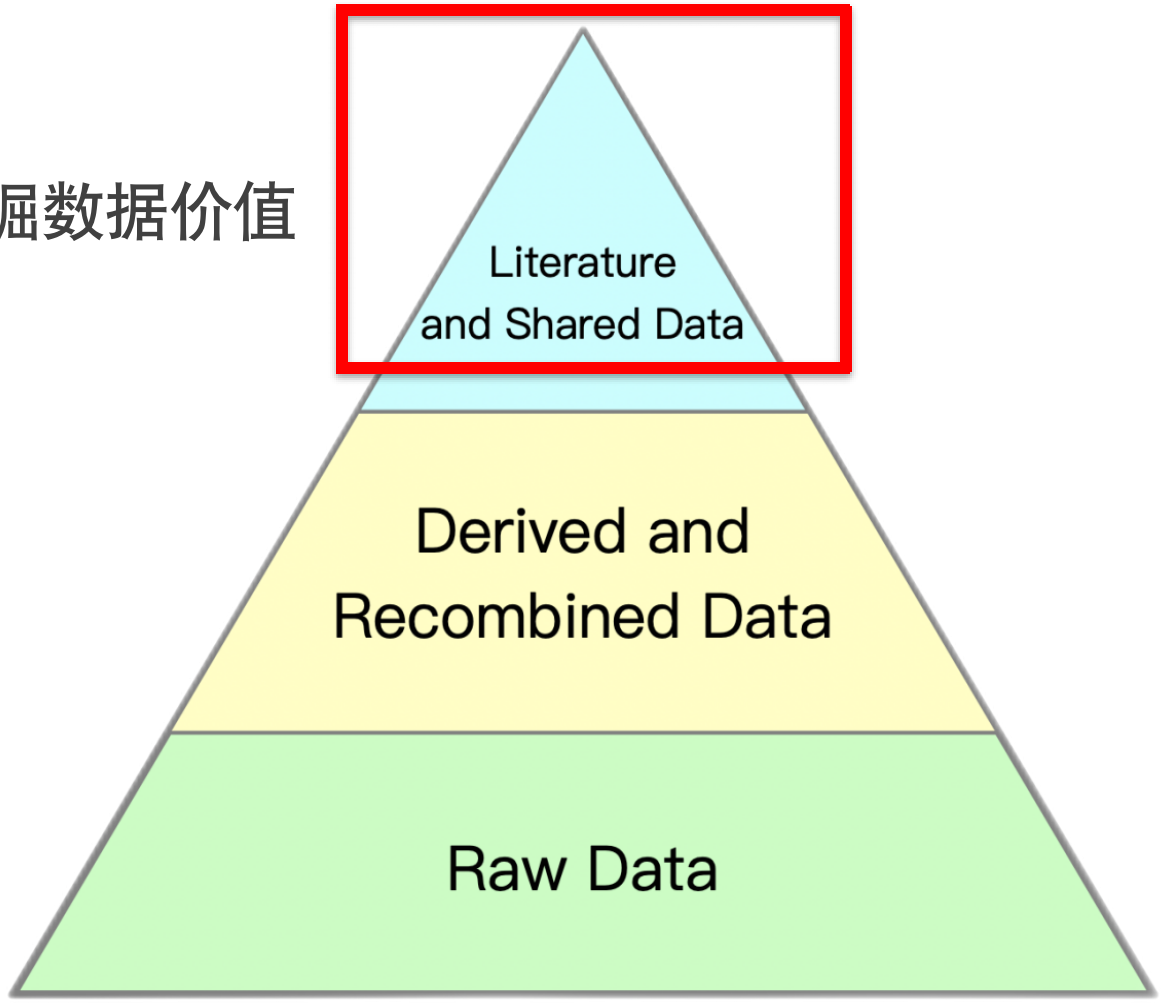
一款标注工具，与科研论文(资源池)深度融合:①PDF文件的一站式标注②根据词表、语法规则及ML算法自动标注③协同标注与管理



1. Accelerating electrocatalysts design by a knowledge graph for CO2 reduction. in submission
2. Autodive: An Integrated Onsite Scientific Literature Annotation Tool. in submission

总结

- LLM有可能颠覆很多现有科研模式
  - （高质量）数据-知识在过程中将发挥重要作用
- 有数据积累及工具研发能力，可以合作挖掘数据价值



请批评指正，谢谢！

duyi@cnic.cn