·科技评述·

# *MIT Technology Review* 2022 年“全球十大突破性技术”解读

［编者按］ 自 2001 年起，*MIT Technology Review* 每年都会评选出年度“全球十大突破性技术”，不少在当年崭露头角的技术，如今已经深刻地改变了我们的生活，推动了人类社会的进步。2022 年 2 月 23 日，*MIT Technology Review* 评选出的年度“全球十大突破性技术”包括：终结口令、新冠变异追踪、长时电网储能电池、新冠口服药和除碳工厂等。这些上榜的科学技术突破代表了当前时代科技的发展前沿和未来方向，为了让广大读者深入了解这些突破性技术的科学价值及其背后的科学故事，《中国科学基金》编辑部特邀请各领域著名科学家对“十大突破性技术”中的九项分别进行深入解读，推进科技资源科普化，构建科学普及与科技创新“两翼齐飞”新格局。

## 1 终结口令（The end of passwords）

20 世纪 60 年代，口令（Password，坊间称为“密码”）最早被图灵奖得主费尔南多·科尔巴托教授用于大型机的本地文件访问控制。20 世纪 90 年代，互联网开始进入千家万户，口令也在互联网世界得到广泛应用。随着用户网络账号的增多，用户为了方便记忆，倾向使用流行口令、在口令中使用个人信息、在多个账号重用口令，导致严重的安全隐患。自 2000 年以来，数以百计的新型身份认证方案陆续被提出。其中，无口令（Passwordless）方案近年来受到企业的青睐，比如谷歌、苹果、微软等公司，都为用户提供了无需输入口令就能登录应用和服务的身份认证方案。在无口令身份认证方案中，要么用户拥有一部带摄像头或指纹识别器的移动设备，并安装相应的身份认证应用程序；要么用户拥有专门的硬件设备（如 U 盾），以存储身份认证所需的密钥及算法参数。当前无口令身份认证方案仍在初级阶段，面临可扩展性低、部署成本高和隐私泄漏等挑战，这些问题亟待解决。在可预见的未来，口令将仍是最主要的身份认证方法，无口令方案可能会使普通用户对口令的直接接触变少，但口令仍在幕后保护着我们的网络与信息安全。

### 专家点评：

**汪定** 南开大学网络空间安全学院教授、密码科学与技术系主任，天津市网络与数据安全技术重点实验室副主任，研究方向为数字身份安全。以第一作者（或通讯作者）在 *IEEE Symposium on Security and Privacy*、*IEEE Transactions on Dependable and Secure Computing* 等刊物发表论文 80 余篇。研究工作引起美国身份认证标准 NIST SP800-63-3 的修改。获教育部自然科学奖一等奖、中国计算机学会（CCF）优秀博士学位论文奖、ACM 中国优秀博士论文奖、中国密码学会优秀青年奖。

**陈晓峰** 西安电子科技大学网络与信息安全学院教授，国家高层次人才，互联网基金会网络安全优秀教师。主要研究领域为密码学和云计算安全，已在重要国际会议和期刊发表学术论文 200 余篇。担任 *IEEE Transactions on Dependable and Secure Computing*、*IEEE Transactions on Knowledge and Data Engineering* 等著名国际期刊的编辑，*Asia Conference on Computer and Communications Security 2016*、*International Conference on Network and System Security 2014* 等多个国际会议的大会主席。获 2019 年度教育部自然科学奖二等奖、2016 年中国密码学会密码创新奖。

**马建峰** 西安电子科技大学网络与信息安全学院教授，国家高层次人才，互联网基金会网络安全优秀人才，“网络与信息安全”教育部创新团队带头人，陕西省“网络与信息安全”三秦学者创新团队带头人。历任西安电子科技大学计算机学院院长、网络空间安全学部主任。担任国务院学位委员会“网络空间安全”学科评议组成员、陕西省网络安全与信息化专家咨询委员会副主任。曾以第一完成人身份获得国家技术发明奖二等奖两次。

身份认证是保障网络安全的第一道防线，口令（Password，坊间称为“密码”）是最常用的身份认证方法。近年来频频发生的大规模口令泄露事件，为黑客和不法分子破解用户的账号口令提供了源源不断的素材，引起人们对口令安全性的担忧。在这一背景下，美国 Okta 和 Duo 等面向企业用户的公司，微软和谷歌等面向个人用户的公司，都为用户提供了无需输入口令就能登录应用和服务的身份认证方案，引起社会广泛关注，并入选此次 *MIT*

*Technology Review* “十大突破性技术”。

实际上，这是“终结口令”的第二次浪潮。口令最早在 20 世纪 60 年代开始在大型机上使用[1]，设计初衷是用来控制大型计算机上本地文件的访问权限，避免分时操作系统的时间片滥用问题。20 世纪 90 年代以来，互联网服务（如电子邮件、电子商务、社交网络）蓬勃发展，口令成为互联网世界里保护用户信息安全的最主要手段之一。随着用户的口令账号越来越多，为方便记忆，用户倾向使用流行口令、在口令中使用个人信息（如姓名、生日）、在多个账号间直接重用或简单修改后重用口令，导致严重的安全隐患[2,3]。另一方面，攻击者的计算能力不断增强。在这一背景下，自 2000 年开始，数以百计的新型身份认证方案陆续被提出。

早在 2004 年，时任微软董事长的比尔·盖茨就对外宣称微软将不再使用口令[4]，掀起了“终结口令”的第一次浪潮。微软与当时世界最大的安全公司 RSA 合作开发了一种名为 SecurID 的技术，这种技术本质上是一种“硬件设备＋验证码”的双因子认证。与此同时，微软也开发了一种名为“tamper-resistant”的生物 ID 卡识别技术，本质是一种“生物特征＋硬件设备”的双因素认证。随后，学术界也陆续指出了“安全的口令记不住，能记住的口令不安全”等问题，提出了数以百计的各类新型身份认证方法，如基于各类基于生物特征、行为特征的认证，基于图形口令的认证和单点登录[5]等。

出乎意料的是，始于 2004 年的这波“终结口令”的浪潮，到 2009 年左右逐渐悄无声息地消失了，口令的地位不仅没有被撼动，反而得到了更广泛的应用。用户平均拥有的账号口令数量，也从 2007 年的 25 个增长到 2020 年的 80 个左右。这引起了学术界的反思。在数字世界里，信任不会凭空产生，而身份认证是构建信任的主要环节。身份认证方法有成百上千种，但基本手段可分为以下三类[5,6]：（1）基于用户所知，如口令；（2）基于用户所有，如 U 盾；（3）基于用户所是，如生物特征。这些尝试替代口令的新方法，有的在安全性方面优于口令，有的在可用性方面见长，但几乎都在可部署性上比口令差，并且在安全性、可用性、隐私保护方面几乎都难以做到均衡。因此，学术界从 2012 年开始逐渐形成一个共识[5-7]：口令在可预见的未来仍将无可替代。

2015 年至今，学术界逐渐认识到：除了用户因素，导致口令安全问题的另一原因在于服务运营商的安全保障缺失。长期以来，运营商把保护口令的责任推给用户，在最基本的口令策略设置、口令强度评价和口令存储安全等方面都是穿着“皇帝的新衣”[7]。最近，微软、谷歌和苹果等公司加强了口令安全防护措施，并即将为用户提供无需输入口令就能登录应用和服务的身份认证方案[8]。在这些无口令方案中，要么用户拥有一部带摄像头或指纹识别器的移动设备（如智能手机、平板电脑等），并安装相应的身份认证应用程序（如微软的 Authenticator App）；要么用户拥有专门的硬件设备（如支持 FIDO2 标准且能识别指纹的 U 盾），以存储身份认证所需的密钥及算法参数。此外，这些方案仍把口令（或 PIN 码）作为生物特征识别失效时的应急选项。

由此可以看出，当前无口令身份认证方案仍处于初级阶段，存在明显的缺陷：一方面，仅在大型公司的少数平台和设备上应用（如 iOS 16 上或 Win 10 以上），未考虑旧版本的系统和不使用智能手机的人群；另一方面，由于需要特定版本的系统或平台导致可扩展性低，涉及硬件导致部署成本高，由于生物特征的不可更改性导致存在隐私泄漏风险。此外，无口令认证方案降低了用户对身份的控制权，52％的被调研用户表示不接受把信任链条传递到手机等设备。截至 2022 年 2 月，78％的微软云服务企业用户仍仅使用账号名和口令登录，只有 22％启用了基于口令的多因素认证或无口令方案[9]。

综上所述，在可预见的未来，口令仍将是最主要的身份认证方法之一，基于口令的认证技术仍不可替代。未来，随着无口令方案的不断成熟，在一些场景下（如使用智能手机）用户对口令的直接接触可能会变少，但口令不会消失，仍是应急认证手段，将在幕后保护着我们的网络与信息安全。

**图 1　“终结口令”技术入选此次 *MIT Technology Review* 2022“全球十大突破性技术”**
**（图片来源：*MIT Technology Review* 官网）**

## 2 新冠变异追踪(COVID variant tracking)

2019 新型冠状病毒(SARS-CoV-2)仍在全球传播,这场全球疫情使得病毒基因组测序受到了前所未有的资金青睐,并极大地扩大了全球对此类病毒监测与预警的能力。2021 年 11 月,南非一家实验室的测序人员发现一个有 50 多个突变的病毒基因组,并首次发出警示信号,几乎在瞬间,西雅图、波士顿和伦敦的计算机都在利用这些数据做出预测:这种被命名为 Omicron 的新冠病毒变异体是个麻烦,它是一种可能逃避抗体的病毒突变体。科学家们借助于基因测序、分析技术,可绘制出 SARS-CoV-2 的基因组图谱,可监测病毒传播过程中基因组发生的变化,并可进一步地快速发现并警告新的病毒变异体,如阿尔法(Alpha)、德尔塔(Delta),以及最近出现的奥密克戎(Omicron)。其中,Omicron 被认为是迄今为止变异程度最高的病毒变种。这一项史无前例的努力,使 SARS-CoV-2 成为历史上接受基因测序最多的生物体,超越了流感病毒、人类免疫缺陷病毒(Human Immunodeficiency Virus,HIV)甚至人类基因组,极大地提高了全球对此类病毒的监测、传播跟踪与预警能力。

**专家点评:**

**陆剑** 北京大学生命科学学院教授、博士生导师,教育部长江学者特聘教授,国家重点研发计划重点专项首席科学家。目前担任 *Science Bulletin* 和 *Molecular Biology and Evolution* 的副主编、中华预防医学会生物信息学分会委员和北京市生物信息学会理事。研究方向为分子进化和基因组学,长期致力于群体遗传学、进化基因组学和基因表达调控等领域的研究。参加中国—世界卫生组织新冠病毒溯源联合研究,获得全国科技系统抗击新冠肺炎疫情先进个人称号。

**钱朝晖** 中国医学科学院/协和医学院病原生物学研究所研究员,博士生导师,国家病原微生物实验室生物安全专家委员会委员。长期从事冠状病毒入侵、复制以及致病机制研究。

**吴爱平** 中国医学科学院系统医学研究院、苏州系统医学研究所研究员,北京协和医学院博士生导师。获得全国科技系统抗击新冠肺炎疫情先进个人称号、中华医学科技奖二等奖和江苏省“双创人才”等奖励。研究方向为传染病生物信息学,专注于开发新型计算方法,建立病毒性传染病的生物信息分析框架,系统进行新发突发病毒的发现溯源、变异进化和免疫评估等。主持或参与了国家重点研发计划、国防科技创新特区和国家自然科学基金等多项国家级科研项目。

新型冠状病毒传播引发的疫情给全球经济和公共卫生带来了极大的破坏。作为一种 RNA 病毒,新型冠状病毒在流行过程中必然会不停地发生变异,导致新的变异株不断涌现。世界卫生组织已经定义过 5 个密切关注变异株(Variant of Concern,VOC),分别是阿尔法(Alpha)、贝塔(Beta)、伽玛(Gamma)、德尔塔(Delta)和奥密克戎(Omicron)。快速积累的大量病毒基因组,为大流行期间病毒的持续演化和流行病学研究提供了宝贵的数据基础。

对新冠病毒不同变异株进行科学的谱系划分和演化动态追踪,不仅有助于流行病学的调查和疫情精准防控政策的制定,对病原体检测,临床诊断,疫苗和治疗药物的研发以及有效性评估也具有不可估量的重要意义。在新冠疫情暴发早期,新冠病毒基因组序列还非常有限的情况下,我国科学家就开展了新冠病毒基因组分型和谱系划分的系统研究。例如,我国科学家准确地将新冠病毒分为 L 和 S 两个主要谱系,推测 S 谱系较为古老,而 L 由 S 谱系进化而来。进一步的研究发现早期病例中 S 谱系病毒感染者中危重症比例显著高于 L 谱系病毒感染者。为了便于追踪不同谱系病毒演化流行过程和特征,进一步构建了新冠病毒分层次谱系划分系统,绘制了完整的反映各个谱系之间亲缘关系的单倍型网络图,揭示谱系演化关系,并建立新冠病毒谱系时空动态分布的可视化平台(www.covid19evolution.net)。当前世界卫生组织定义的 VOC 变异株均是 L 谱系的分支谱系。S 和 L 谱系分别对应 Pango Lineage 分型系统的 A 型和 B 型。S 和 L 谱系划分已被科学界广泛接受和认可,并被全球禽流感基因共享数据库(GISAID)数据库、国家基因组科学数据中心、中国疾病预防控制中心以及《中国—世界卫生组织新冠病毒溯源联合研究》中英文报告所采用。

我国科学家还发现,新冠病毒感染人数的不断增加会加速新冠病毒变异的适应性演化,从而形成正反馈循环。庞大的全球感染人群,为新冠病毒的位点突变、片段插入/删除以及基因重组等基因组结构变异等提供了巨大空间。在可预期的一段时间内,新冠病毒将会与人类共存。因此,对病毒变异规律的及时解析和谱系演化动态的及时追踪仍然非常重要。如何科学地预测病毒的变异趋势,对可能造

成大流行的高风险株做到有效的先期预警尤为重要。早期新冠病毒变异的功能选择主要表现为传播力、受体结合能力以及病毒复制能力的增强。但是，在奥密克戎变异株高传播力的背景下，突破性感染不停发生，感染人群比例不断升高，病毒多样性持续扩大，免疫逃逸已经成为新冠病毒变异的主要驱动力。因此，如何在疫苗接种和突破感染形成的复杂免疫选择压力下，预测新冠病毒变异趋势和流行动态将会是一个充满挑战但又亟需解决的重要科学问题。GISAID 已经收录了超过 1 000 万条新冠病毒全基因组序列及部分样本的采集信息，基因组序列的超复杂性也为监测和分析新冠病毒演化趋势提出了巨大挑战。开展病原学、免疫学、结构生物学、群体遗传学、分子演化以及计算生物学等多学科的合作，结合人工智能和机器学习等新兴技术可能是解决这一问题的有效途径。

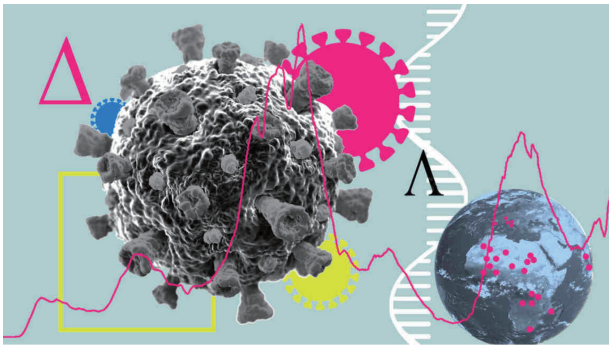


**图 2　SARS-CoV-2 病毒是地球上被测序最多的生物体，极大地增强了全球对此类病毒的变异追踪与预警能力**
**（图片来源：*MIT Technology Review* 官网）**

## 3　长时电网储能电池（Long-lasting grid battery）

2021 年 4 月，可再生能源打破了加利福尼亚州主电网的纪录，提供的电力足以满足 94.5% 的需求，这一时刻被誉为低碳化道路上的一个里程碑。我们使用的可再生能源比以往任何时候都多。然而，可再生能源带来的波动式电力需用一种廉价且长时（数小时甚至数天）的储能电池保存，以备日后使用。新型的铁基电池有望胜任这一任务。总部位于俄勒冈州的 ESS 公司，其电池可实现 4 至 12 小时的储能，并在 2021 年推出了其第一个电网规模的项目。总部位于马萨诸塞州的 Form Energy 公司称其电池可储存电能长达 100 小时，他们在 2021 年筹集了 2.4 亿美元，将在明尼苏达州安装一兆瓦级别的储能工厂，预计 2023 年完成。这两家公司都选择使用铁基电池，而铁是地球上最丰富的材料之一。这意味着他们的产品最终可能比锂离子电池和钒系液流电池等其他储能电池更便宜。Form Energy 公司表示，其电池最终的成本可能仅为 20 美元/千瓦时，甚至低于未来几十年对锂离子电池成本的乐观预测。但铁基电池也存在一些技术挑战，如它们的效率通常较低，这意味着投入其中的相当一部分能量无法被回收；此外，副反应也会随着时间的推移而使电池退化。但如果铁基电池能以足够低的成本被广泛安装使用，便可以为更多人提供来自可再生能源的电能。

**专家点评：**

**张新波**　研究员，中国科学院长春应用化学研究所稀土资源利用国家重点实验室主任，国家杰出青年科学基金获得者。致力于能源存储与转化研究，目前主要聚焦于金属—空气电池、新型离子电池与能源电催化方面的关键材料设计和高性能器件研制，开发了具有完全自主知识产权的锂空气电池器件。在 *Nature Chemistry*、*Nature Energy* 等国际权威期刊上发表论文 200 余篇，主编国际专著 1 部。授权发明专利 20 件。2019 年获吉林省自然科学奖一等奖。

未来在以可再生能源为主体的新型电力系统中，可再生能源的比例将超过 50%，这必然会要求储能设施具备十几个小时乃至几天的储能时长，以满足吉瓦（Gigawatt，GW）级别的再生能源并网和长时间削峰填谷的需求。然而，在目前的储能电池技术水平下，锂离子电池储能时长以 2 小时居多，部分已经提升至 3 到 4 小时，但要达到 6 小时及以上的储能时长则会面临成本与产品安全等方面的诸多挑战。因此，低成本、长时储能电池的发展将成为电力系统转型的关键。

此次入选 2022 年 *MIT Technology Review*“全球十大突破性技术”的水系铁基电池是基于廉价和储量丰富的铁元素构筑的，其具有高安全性和环境友好等特征。其中，美国俄勒冈州 ESS 公司的铁基液流电池以氯化亚铁为正负极电解液，通过电解液中铁离子的氧化还原实现电能的储存和释放，可实现长达 20 000 次的稳定循环。此外，该液流电池的储能活性物质与电极完全分开，功率和容量设计互相独立，便于模块组合设计和电池结构放置，其电网规模的储能模块可以实现 4 至 12 小时的能量储存。

不同于液流电池，Form Energy 公司的铁—空气电池是一种静态电池，其基本原理是基于铁的可逆氧化（生锈），可持续多达 10 000 次的循环。相比于铁基液流电池，铁—空气电池的储能容量更大，其可储存电能长达 100 小时（约可为电网提供超过 4 天的电力），这种电池将使具有成本效益的“多日储能”成为可能。上述两种铁基电池在大规模储能方面均具有明显的优势：超长循环寿命、高安全稳定性、可扩展性、低成本和绿色环保，可平衡可再生能源发电的波动式变化，实现低碳长时电网储能。

铁基长时电网储能电池的发展，可以弥补锂离子电池的一些不足，以科技创新的方式将电力系统从化石燃料发电转变为可再生能源发电，有利于在全球范围内减少碳排放，实现低碳电网碳的发展和碳中和的终极目标。然而，除了长时电网储能电池外，还有一些其他可以提供稳定电力服务的能源组合（核能、化石能源＋碳捕捉与封存技术、氢能等）与之竞争，这些技术未来的发展，也会在一定程度上左右长时储能电池在电网中的占比。此外，与其他储能技术的发展一样，长时储能电池从研发、示范、落地到规模化，一路上必将面临产能、供应链、建设、运营等多方面的挑战，必须严格控制每一环的风险，才能实现既定的成本目标。

我国的长时电网储能技术以全钒液流电池为主，其已经过十多年的示范考核，并且其大规模储能的工程效果已得到了充分的验证，产业配套成熟，可支撑起百兆瓦级储能项目的设计与开发。此外，全钒液流电池系统的单瓦时成本已可控制在 2～3 元的水平，具备了商业化应用的条件。2018 年以来，我国液流电池的装机量呈现爆发式增长。其中，2020 年规划的液流电池装机量超过 6 GW，容量超过 20 GWh。与此同时，单个项目的规模也在不断提升，如 200 MW/800 MWh 的全钒液流电池示范项目。整体而言，我国液流电池的产业研究和技术工艺处于国际领先水平，特别是国内液流电池的龙头企业，大连融科在海外市场的拓展也在如火如荼地进行。然而，全钒液流电池的低能量密度和钒高昂的价格，需要我们开发更具价格和能量密度优势的新型长时电网储能技术。

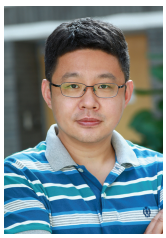
储能作为“双碳”背景下构建低碳电网的关键组成部分，跨天、跨月乃至跨季节的长时电网储能系统的发展迫在眉睫。目前长时储能技术仍处于百家争鸣的中早期研发示范阶段，孰胜孰劣尚未揭晓。电化学储能由于动力电池产业的推动，不受地理环境的制约，暂时处于比较有利的竞争地位。未来电网储能系统的发展需要以模型数据开源、学术产业结合等方式集思广益，甄选出最具经济可靠性的电源储能配置方案，形成多能互补的，新能源＋储能的电力系统，为实现“双碳”目标提供强有力的支撑。

**图 3 廉价、储能持久的铁基电池有望分摊可再生能源的供应压力，并扩大清洁能源的使用范围**
**（图片来源：*MIT Technology Review* 官网）**

## 4 AI 蛋白质折叠（Artificial intelligence for protein folding）

作为生命体最重要的功能载体之一，蛋白质在众多生命活动中发挥着关键的作用。蛋白质在行使功能时往往需要折叠成特定的三维结构，因此对蛋白质结构的测定和解析不仅能帮助人们在分子层面上理解大多数生命活动的机理，而且可以有效辅助基于结构的药物开发以及相关疾病的诊治。目前通过实验手段解析蛋白质结构费时费力，远远无法满足现实需求。2020 年底，谷歌重组后的“伞形公司”Alphabet 旗下名为 DeepMind 的人工智能实验室采用多种深度学习技术，开发出了一款名为 AlphaFold2 的软件，能根据蛋白质的氨基酸序列准确预测其三维结构。该软件使用一种称为深度学习的人工智能技术，可以预测蛋白质的形状，甚至精确到原子。由于大多数蛋白质的氨基酸序列已知，该软件可以在数个小时内提供目标蛋白质原子分辨率的结构信息，而且其预测的结构模型准确度很高，在很多蛋白上可以与实验解析的真实结果媲美。世界各地的科学团队已经开始使用它来研究癌症、抗生素抗性和新冠病毒。2022 年，该技术被 *MIT Technology Review* 评选为“全球十大突破技术”之一。

**专家点评：**



**龚海鹏**　清华大学生命学院副教授，博士生导师，生物信息学教育部重点实验室副主任。2009 年加入清华大学生命学院，主要从事蛋白质结构相关的计算方法研究。近年来的研究兴趣主要集中在结合人工智能技术发展蛋白质结构预测算法和分子模拟的采样方法，以通讯作者身份在 *Nature Machine Intelligence*、*Advanced Science*、*Bioinformatics*、*PLoS Computational Biology*、*Journal of Chemical Theory and Computation* 等计算生物学主流期刊上发表多篇论文。近 5 年主持国家自然科学基金项目 3 项。

生命体中的主要生命活动都通过蛋白质分子完成，因此理解单个蛋白质分子的工作机理至关重要。蛋白质的多肽链是由氨基酸顺序连接而成的线性分子，它往往折叠成特定的三维结构来行使功能。换言之，蛋白质的序列决定结构，而结构又决定功能。自上世纪五六十年代起，蛋白质序列、结构与功能间的关系就一直是生命科学的核心问题。作为这一信息链条的中心点，蛋白质结构既可以帮助人们理解生命活动的分子机理，也能有效地辅助蛋白质设计和基于结构的药物设计，因而结构解析已经成为生物物理领域最重要的研究方向之一。过去二三十年来，结构生物学取得了长足进展，包括蛋白质晶体学和冷冻电镜等技术的快速发展，使得人们可以较为快速地测定生物大分子的三维结构。目前蛋白质结构数据库（Protein Data Bank，PDB）中已经积累了超过 18 万个分子的结构。但是，总体而言，蛋白质结构的实验测定仍然较为耗时，往往至少要耗费数月时间。此外，由于新一代测序技术的发展，蛋白质序列的积累速度远远大于结构解析的速度。目前蛋白质序列库中的蛋白质数目已经超过结构数据库 3～4 个数量级，这一差距无法通过实验方法弥补。

根据安芬森法则（Anfinsen's dogma），大多数球状蛋白的三维结构由氨基酸序列唯一决定。自 20 世纪八九十年代起，人们就开始发展计算机算法，通过研究序列和结构间的关系，根据氨基酸序列预测蛋白质的三维结构。1994 年，约翰·莫尔特（John Moult）等人组织了第一届国际蛋白质结构预测评估竞赛（Critical Assessment of protein Structure Prediction，CASP），用于系统评测各种计算方法的预测准确性。该竞赛每两年举办一次，组委会收集未发表的结构数据，对参赛者发布其序列信息，然后收集其预测结果进行双盲评估。CASP 竞赛极大地促进了蛋白质结构预测领域的发展。在早期 CASP 竞赛中，发展的基于模板的建模方法 Modeller 以及基于统计和物理模型的建模方法 Rosetta 和 I-TASSER 等程序，结合物理知识和对结构数据库的统计分析，可以对某些特定蛋白提供较为准确的预测模型。但是，随着实验解析蛋白质数目的快速积累，这些方法的预测性能并未显示出相应提升，反而达到了瓶颈。2015 年，克里斯·桑德斯（Chris Sanders）等人提出可以从多重序列比对中获得氨基酸残基间的共进化关系，从而为结构预测提供额外信息。2016 年的 CASP12 竞赛中，许锦波等人提出的 RaptorX 程序，首次使用深度卷积模型，根据多重序列比对预测氨基酸残基间接触，再根据预测结果折叠蛋白，从而显著提升了结构预测的平均准确率。其后，人工智能方法开始广泛介入蛋白质结构预测领域。2018 年的 CASP13 竞赛中，DeepMind 发展的 AlphaFold 采用了类似的方案预测残基间距离并根据预测距离折叠蛋白。其后的大多数方法也主要沿这一思路进行。2020 年 CASP14 竞赛前，人们发现这类方法的性能并不能随模型参数量增加而继续提高，而且这类深度学习模型的预测准确性离现实需求尚有一段难以跨越的距离。但是，在 2020 年底 CASP14 结果公布时，DeepMind 提出的 AlphaFold2 算法远远超越了其他深度学习模型，对绝大多数目标蛋白都可以提供高度精准的预测模型。对有些蛋白质而言，AlphaFold2 预测的结果与实验解析的模型高度相似，甚至仅根据实验数据都很难区分孰优孰劣。这一结果也震惊了整个科学界。

后续的报告和论文显示，DeepMind 在设计 AlphaFold2 时完全摒弃了第一代 AlphaFold 的架构，规避了残基间接触或距离的预测，采用了一种全新的端对端模型直接根据序列预测结构。这一设计不仅能加快预测速度，而且可以有效抑制中间过程中的误差积累。此外，DeepMind 采用了近年来自然语言处理领域较流行的 Transformer 架构。与以前常用的卷积架构不同，这种基于注意力机制的模型允许所有氨基酸残基在每一步操作中发生信息交互，能更好地模拟蛋白质折叠过程中的残基间相互作用。最后，AlphaFold2 还使用了重循环和自蒸馏等工程技术，进一步有效提升了预测准确率。总之，在 AlphaFold2 中，DeepMind 通过具有高度创新性的深度学习网络架构设计，初步解决了蛋白质结构预测（或折叠）这一困扰人类 50 年之久的科学难题，也因此入选 *MIT Technology Review* 评选的 2022 年“全球十大突破性技术”。

AlphaFold2 的提出显然对生命科学有巨大的

促进作用。一方面，它能够快速准确地根据氨基酸序列预测蛋白质的三维结构，因此可以有效弥补现有结构生物学技术的缺陷。另一方面，通过对AlphaFold2的进一步研究，人们可以更好地理解蛋白质序列和结构间的映射关系。目前已有很多研究组开始把AlphaFold2应用于药物开发和蛋白质设计领域。比如，清华大学的张林琦和彭健合作，通过进一步优化的AlphaFold2模型预测抗原和抗体的结合能，并根据预测结果优化抗体设计，最终开发出我国首款新冠特效药。2021年，DeepMind与EBI合作，建立了基于AlphaFold2预测结果的数据库AlphaFold DB。该数据库中已经储备了近一百万蛋白质的预测结构，为生命科学各个领域的科学家们提供重要的蛋白质结构信息。这一行为很可能会改变很多领域的科研范式，促进分子层面的研究从以序列为基础转变为基于序列和结构的研究，从而加快生命科学定量化的步伐。

当然，AlphaFold2还有一定的局限性。首先，它在很多蛋白上的预测精度还有待进一步提高，目前还不能完全满足药物开发等领域需要的结构精度，特别是对蛋白质复合体的预测精度较低。因此，AlphaFold2还不能完全替代结构生物学研究。其次，AlphaFold2模拟的是从多重序列比对到三维结构的映射关系，并没有解决从单一序列到三维结构的映射关系，因此蛋白质折叠问题还没有完美解决。最后，也是最重要的一点，针对一个特定的目标蛋白，AlphaFold2仅提供有限的结构模型，不能揭示其结构的动态变化。而动态结构才是真正决定功能的基础。

无论如何，AlphaFold2的提出展示了人工智能技术对生命科学研究的巨大促进作用。在蛋白质结构预测领域，预计人工智能技术将继续引领后续的进展，解决目前AlphaFold2的局限性：(1) 提高蛋白质复合体的结构预测精度；(2) 发展根据单一序列预测蛋白质结构的人工智能算法；(3) 根据氨基酸序列预测蛋白质的动态结构。

我国在蛋白质结构预测领域的基础整体上还比较薄弱。近年来虽然有多个学术研究组在残基间距离预测、能量函数构建和模型质量评估等子领域做出了原创性的工作，但是由于学术研究组的规模和资金有限，还没有形成完整的自主研发的程序算法，能达到与AlphaFold2持平的程度。但是，目前已有企业(如深势科技、华深智药、腾讯、百度、华为等)牵头的研发小组以AlphaFold2为模板进行二次开发，并取得了一定的成绩。考虑到人工智能技术的飞速发展，我国在这一领域还有极大的潜力。希望在不久的将来，我国能通过多个学术研究组的联合攻关，或企业和高校的合作研发模式，开发出原创性的、具有完全自主知识产权的蛋白质结构精准预测算法。

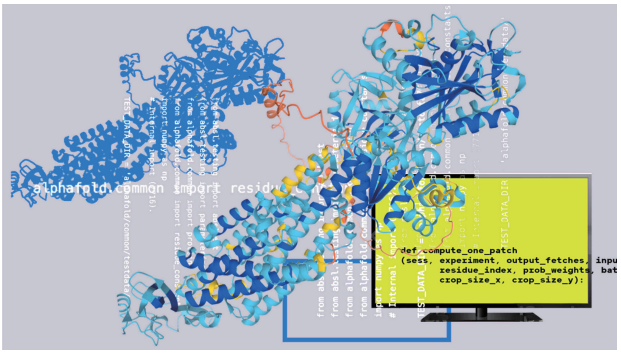


**图4 人工智能实验室DeepMind开发的AlphaFold2软件初步解决了蛋白质结构预测这一科学难题**
**(图片来源：*MIT Technology Review* 官网)**

## 5 疟疾疫苗(Malaria vaccine)

寄生虫是复杂的多细胞生物，其基因组比大多数病毒和细菌中的基因组大500～1 000倍。使其能够通过无数种方式的基因突变来躲避人体免疫系统的监视。疟疾是疟原虫所引起的严重危害人类生命健康的寄生虫病。其主要集中在撒哈拉以南的非洲地区，该地区的病例约占全世界的95%。每年有60多万人死于疟疾，其中大部分是5岁以下的儿童。2021年10月，世界卫生组织批准了世界上第一种对抗由蚊子传播的致命疾病的疫苗—疟疾疫苗。然而，这款由葛兰素史克公司研发的疟疾疫苗，RTS，S或Mosquirix，被认为不是一种特别有效的疫苗。它需要在5至17个月大的儿童中接种三剂，并在12至15个月后接种第四剂。此外，在肯尼亚、马拉维和加纳的80多万名儿童中，这种疫苗在第一年对严重疟疾的有效率约为50%，而且随着时间的推移，其疗效急剧下降。即便如此，公共卫生官员仍将这种自1987年就开始测试的疫苗誉为非洲的“游戏改变者”，主要原因是当其与其他疟疾控制措施(包括驱虫蚊帐和在雨季使用的预防药物)结合使用时，有望将疟疾死亡人数减少多达70%。Mosquirix作为第一个被批准用于寄生虫病的疫苗，旨在敲响免疫系统的警钟，保护潜在的宿主免受感染，对鼓励创新以及下一代疟疾疫苗的开发具有重大意义。

**专家点评：**

**江陆斌**　研究员，中国科学院上海巴斯德研究所副所长，上海科技大学特聘教授，国家杰出青年科学基金获得者，国家重点研发计划项目首席科学家，美国国立卫生研究院（National Institute of Health，NIH）R01 项目首席。曾获湖北省科技进步奖二等奖、上海市科技系统先进个人、中国科学院优秀教师“朱李月华”奖等奖励和荣誉。长期致力于恶性疟原虫致病的表观遗传学机制研究，首创了恶性疟原虫表观遗传基因编辑技术，揭示了恶性疟原虫免疫逃逸的调控网络，阐明了线粒体功能抑制的表观遗传机制，鉴定到一批具有药物开发潜力的表观遗传靶点，其中一种小分子候选药物已进入临床前研究。

RTS，S/AS01（RTS，S）是全球首款获得世界卫生组织（World Health Organization，WHO）批准的疟疾疫苗。它是恶性疟原虫环子孢子蛋白 CSP 的 C-末端序列（包括 NANP 抗原重复序列和 T 细胞表位序列）与乙型肝炎病毒表面抗原（HBsAg）融合、组装成病毒样颗粒结构的亚单位疫苗，并通过新型脂质体免疫佐剂 AS01 增强疫苗的免疫原性。自 2021 年 10 月起，RTS，S/AS01 获批在非洲疟疾传播的中、高风险地区 5 月龄以上儿童中使用。

疟疾是严重危害人类健康的全球三大传染病之一。随着青蒿素等各类抗疟药的临床耐药性问题日益加剧，目前全世界仍有近一半人口面临疟疾感染风险。致死性最强的恶性疟疾原虫每年造成 2 亿～3 亿的感染病例和近 60 万的死亡病例，是实现“人类卫生健康共同体”目标的关键阻碍之一。21 世纪以来，全球每年约有 10 项疟疾疫苗项目获批开展临床试验，约 150 项已完成或提前终止临床试验。其中，RTS，S/AS01 在非洲地区的多中心Ⅲ期临床试验数据显示，5～17 月龄儿童接种 4 剂疫苗后，临床发病的平均保护效率为 36.3%，部分地区可实现约 50%的临床保护效率。迄今为止，RTS，S/AS01 是唯一被证明可降低疟疾患儿临床发病率和死亡率的疫苗。需要指出的是，RTS，S/AS01 仅在接种 4 剂后的 1 年内对 5～17 月龄儿童具有较高的保护效率。随后，其免疫保护效率快速下降，接种 1 年半后平均保护效率已低于 30%。作为疟疾疫苗研究领域零的突破，RTS，S/AS01 具有重大的现实意义，WHO 预期它在未来每年可以挽救数万名 5 岁以下非洲儿童的生命。

不可否认，RTS，S/AS01 并没有达到疟疾疫苗的 WHO 官方标准（保护率＞50%，保护时间＞1 年）。因此，如何有效遏制疟疾在热带、亚热带等国家和地区的流行与传播，依然是全球疟疾研究人员亟需解决的科学问题。虽然在几代疾控工作者的不懈努力下，我国已于 2021 年正式获得由 WHO 颁布的消除疟疾认证，但输入性疟疾在华中、华南和西南省份呈上升趋势。而且，在我国云南、东南亚以及非洲等地已出现了具有青蒿素潜在抗性的恶性疟原虫。因此，研制新型疟疾疫苗刻不容缓，并具有重大的社会和经济意义。

与疟疾作为国际传染病学研究热点极不协调的是，相关寄生虫学研究在国内普遍不被重视，疟疾疫苗研究也多为靶向疟原虫单一抗原的亚单位疫苗策略。由于疟原虫生活史包括肝（细胞）内期、红（细胞）内期和蚊期等复杂的生长时期，恶性疟原虫具有高度变异的抗原蛋白和多变的免疫逃逸策略，这既限制了国内外疟疾疫苗的研发，同时也是导致 RTS，S/AS01 并不完美的主要原因。近年来，随着多种新型基因编辑技术在恶性疟原虫关键生物标志物功能鉴定中的广泛应用，使研究人员针对恶性疟原虫不同生长时期设计多价疫苗成为可能。同时，与传统疫苗相比，新兴的信使核糖核酸（Messenger Ribonucleic Acid，mRNA）疫苗技术、疫苗佐剂和抗原递送系统的技术革新也将为疟疾疫苗研究提供更多的潜在方案，使得新一代高效疟疾疫苗的研发有望在未来 5～10 年内取得关键性突破。

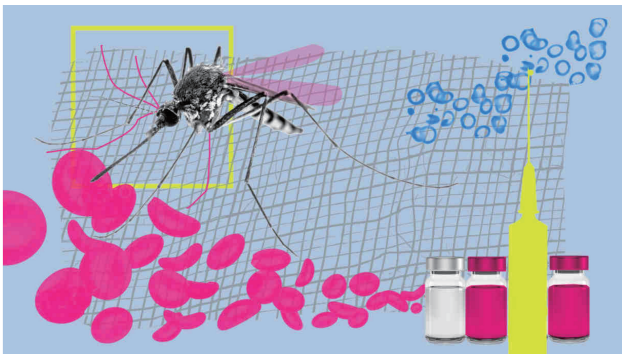


**图 5　世界上第一种对抗由蚊子传播的致命疾病的疫苗—疟疾疫苗获批。**
**（图片来源：*MIT Technology Review* 官网）**

## 6　权益证明（Proof of Stake，PoS）

共识算法是区块链技术的核心，通过算力竞争的“挖矿”工作量证明机制消耗了全球太多的电力和计算资源，一直为人诟病。而 PoS 权益证明机制则有望彻底摆脱这一困境。PoS 算法的突出特点

是引入了币龄的概念，币龄越长，权力越大，挖矿难度越低，所获奖励越多。这样节点不需要消耗过多的外部算力和资源去竞争记账权，一定程度上还缩短了达成共识的时间，提升了系统运行性能。2022 年 2 月，*MIT Technology Review* 发布了 2022 年“全球十大突破性技术”，“PoS 权益证明”与“新冠口服药”“实用型聚变反应堆”“终结口令”“AI 蛋白质折叠”等技术一起名列其中。

**专家点评：**

**张小松** 教育部“长江学者”特聘教授，电子科技大学网络空间安全研究院院长，博士生导师，中国电子学会区块链分会副主任委员。长期从事计算机网络与系统安全技术的研究。以第一完成人先后获国家科技进步奖一等奖、二等奖各 1 项，省部级科技进步奖一等奖 3 项、发明奖 2 项。

2008 年题名为“Bitcoin：a peer-to-peer electronic cash system”的论文发表至今，基于分布式账本技术的区块链在全球产生了巨大深远的影响，而实现分布式系统强一致性及最终一致性达成的共识算法无疑是区块链技术体系的核心，其本质是要解决在分布式网络环境下，如何让所有的节点对窗口内发生事务的顺序和内容正确性达成共识，确保系统内同一个事务处理的可靠和可信，为实现区块链去中心化、开放自治提供机制的支撑和保障。

对于严格维护去中心化机制的“公有链”（Public Blockchain）系统，工作量证明（Proof of Work，PoW）毫无争议是目前最具认可度的共识算法，在全球影响力最大公有链比特币和以太坊系统中均予以采用。PoW 的原理是区块链中各个节点通过算力计算哈希（Hash）难题，其中最先解决难题的节点将获得区块记账权，从而以算力竞争的方式保证数据的一致性，这一过程又俗称“挖矿”。PoW 机制可以表达为：H(param||nonce)＜target，其中，H 表示哈希函数，param 是区块相关的数据，nonce 是随机值，target 是由当前计算难度值决定的目标值。显然，要找到符合条件的 nonce，只能通过穷举的方法来实现，然而，公链节点规模的扩大和挖矿难度的不断增大，PoW 共识机制越来越暴露出无法克服的问题：

（1）能源浪费巨大。截至目前，采用 PoW 共识算法“挖矿”的比特币系统，产生一枚比特币的耗电量大约在 20 万度到 30 万度之间，导致全球范围内的比特币挖矿能源消耗非常巨大，剑桥大学替代金融研究中心数据显示，仅比特币挖矿年度消耗的电量高达 1 300 多亿度电，比很多国家的年度用电总量都要高。

（2）业务性能很低。PoW 共识算法要求每笔交易及其区块都要获得所有节点的确认，才会被记录到账本中，而随着网络规模的扩大，共识的耗时必然提升，目前比特币和以太坊系统的共识速度平均仅有 5 笔/秒左右。虽然有试图以增加区块大小和降低出块时间间隔来提升交易速度的其他衍生公链系统，但是它们仍无法规避出现分叉概率上升的风险和交易效率降低等问题。

（3）算力集中风险。在巨大的利益驱动下，越来越多的专业挖矿算力节点加入到比特币和以太坊系统，甚至出现多个节点联合挖矿形成了的几大矿池占据多数算力的局面，明显违背了区块链去中心化基本原则和设计初衷。

权益证明算法正是为弥补 PoW 不足应运而生。PoS 算法由 PeerCoin 创始人 Sunny King 和 Scott Nadal 提出并实现，其突出特点是引入了币龄的概念，将消耗币龄（代币数量与时间的乘积）与计算 hash 散列的工作量一起作为记账权分配的准则，从而等比例的降低 hash 运算的难度。PoS 机制可以表达为：工作量证明 hash()＜总目标值，而总目标值＝币龄 X 目标值 target。因此节点不再是仅依靠算力去竞争记账权，而是通过长期持有或者获得更多的币去增加币龄。与 PoW 算法相比，PoS 算法是在一个有限的空间里进行共识，不需要消耗过多的外部算力和资源，可以有效地弥补 PoW 的劣势，并且能够在一定程度上缩短达成共识的时间，提升系统运行性能。

股权授权证明（Delegated Proof of Stake，DPoS）基于 PoS 演化而来，由 Block.one 公司开发的企业操作系统（Enterprise Operating System，EOS）是第一个采用 DPoS 的公链项目。DPoS 在完成共识的过程中不需要消耗大量的算力，大大提高了区块的生成速度和交易确认效率，同时不会出现 PoS 机制中富有节点长期支配记账权的情况。

以太坊由于其率先实现了图灵完备的智能合约子系统，目前已经是全世界应用生态发展最好的公有链系统，为解决以太坊面临的网络拥堵、运行节点的算力要求门槛高、PoW 机制能耗巨大等困境，从 2015 年以来以太坊开发团队就一直致力于共识机制的切换研发：（1）利用分片链来减轻节点验证者

的工作量，解决可扩展性问题；(2) 利用信标链随机分配验证者降低作恶概率，保证安全；(3) 利用 PoS 机制降低节点门槛并保障生态的可持续发展，并最大程度上实现去中心化。

以太坊信标链已于 2020 年底上线。2022 年 4 月 11 日，以太坊完成了网络的第一个影子分叉 (Mainnet Shadow Fork)，启动了一个从 PoW 过渡到 PoS 的合并测试网。预计 2022 年以太坊将完成由 PoW 到 PoS 的切换，并由此形成世界范围内节点数最多，应用生态最大的公有链系统，并将进一步推动区块链技术发展。

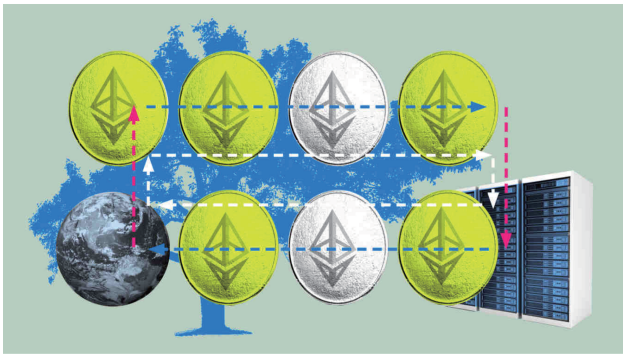


**图 6　一种确保数字货币安全的替代方法可以结束加密货币的能源消耗困境**
**(图片来源：*MIT Technology Review* 官网)**

## 7　新冠口服药(A pill for COVID)

吞下一粒药丸就能使新型冠状病毒消失，这是人们的愿望。现在，这个愿望变成了现实。感染新型冠状病毒几天的病人服用辉瑞公司的一种抗病毒药物后，可将住院的几率降低 89%。美国政府已经订购了价值 100 亿美元的这种名为 Paxlovid 的新药。这款新药的成功研制并不只是黑暗中一次幸运的尝试。针对一种能够调控新冠病毒进行威胁性复制的关键蛋白酶，化学家们设计了这款药物，用于阻断病毒的自我复制能力。事实上，其他类型的冠状病毒中也存在类似的蛋白酶，这也就意味着辉瑞公司的药物有望抵御下一次冠状病毒流行病。抗病毒新药的研发周期比病毒疫苗的设计、合成和测试时间更长，以前从未有一种全新的战胜疾病的分子能如此迅速地从化学家的实验室进入志愿者的口中，并获得美国食品和药物管理局的批准。该药物将防止许多人死于新型冠状病毒肺炎 (Corona Virus Disease 2019，COVID-19)，包括免疫系统较弱而疫苗对其无效的人。而且如果出现了能够打败疫苗的新变种，抗病毒药物可能是我们的最后手段。

**专家点评：**

**李岩**　华中科技大学同济医学院教授、博士生导师。入选国家高层次青年人才项目及湖北省公共卫生青年拔尖人才。主要从事重要传染病的致病机制及新药研究工作。在 *Science*、*Nature Communications*、*Journal of Virology*、*Journal of Infection* 等期刊发表 SCI 论文 50 余篇。

自 2019 年新冠肺炎疫情爆发以来，国内外已有多种新冠肺炎治疗药物和疫苗陆续被开发出来。由于专业医疗资源在新冠肺炎疫情中的紧缺性，许多生物制药研究机构将疗效好、副作用低、给药条件要求较低的新冠口服药作为新型冠状病毒药物开发的重点方向。

近期，由辉瑞公司开发的新冠口服药 Paxlovid 受到了广泛关注。2022 年，发表在 *The New England Journal of Medicine* 杂志上的临床 2/3 期双盲随机对照试验结果表明，蛋白酶抑制剂奈玛特韦 (Nirmatrelvir/PF-07321332) 和利托那韦 (Ritonavir) 联用，可导致进展为严重 COVID-19 的风险比安慰剂低 89%，并且无明显的安全性问题[10]。Paxlovid 实质上是两种药物的联合包装，即蛋白酶抑制剂奈玛特韦 (Nirmatrelvir/PF-07321332) 和能够改善奈玛特韦药代动力学行为的利托那韦 (Ritonavir)。新型冠状病毒 SARS-CoV-2 依赖一种蛋白酶 $M^{pro}$ 来切割蛋白前体，而奈玛特韦是一种针对 $M^{pro}$ 蛋白酶的小分子抑制剂，能够通过竞争结合 $M^{pro}$ 来抑制 SARS-CoV-2 的复制。一方面，奈玛特韦对重组 $M^{pro}$ 的抑制常数 (Ki) 以及对 SARS-CoV-2 抗病毒指标半最大效应浓度 (Concentration for 50% of Maximal Effect，$EC_{50}$) 均达到了纳摩尔每升的水平，同时其在小鼠适应的 SARS-CoV-2 模型中证明了口服活性，并在临床 I 期试验中达到了超过体外抗病毒细胞效力的口服血浆浓度；另一方面，奈玛特韦具备了可接受的溶解度、改进过的大规模合成潜力、与简单制剂载体的兼容性等特点，这些因素构成了奈玛特韦作为新冠口服药组分的分子基础[11]。利托那韦是一种酶抑制剂，其本身对 SARS-CoV-2 无明显活性，但它能抑制负责代谢奈玛特韦的酶 CYP3A4 的活性，从而提高奈玛特韦的血清浓度和半衰期，辅助奈玛特韦发挥功能[12]。

值得注意的是，自 2019 年以来，SARS-CoV-2 已发展出数种比原始株具有更强传播力的突变株。因此，开发抗新冠药物时，其对突变株和潜在新突变株的效力留存水平是必须考虑的问题。奈玛特韦的靶点 $M^{pro}$ 是 SARS-CoV-2 复制过程必需的重要蛋白酶，这种酶依赖一些高度保守的位点组成的口袋行使催化功能[13]。理论上发生在 $M^{pro}$ 上的突变有可能对 SARS-CoV-2 的复制能力造成直接的损害，从而使这种突变株难以获得遗传优势。但目前尚无明确证据表明 Paxlovid 不易引起 SARS-CoV-2 的耐药性。此外，尽管体外研究和动物实验结果提示，奈玛特韦对奥密克戎突变株仍具有抗病毒活性[14, 15]，Paxlovid 对奥密克戎及将来可能出现的新毒株引发的新冠肺炎感染是否仍有较好的临床疗效仍需进一步验证。同时，我们也注意到已有研究表明，在其他药物存在的情况下，利托那韦对奈玛特韦的药代动力学助推效应可能会引入有害的药物-药物相互作用，因此对具有特定用药史的轻度～中度新冠患者开具 Paxlovid 处方可能需要高度谨慎[16]。

总的来说，如其他重要的新冠药物一般，新冠口服药 Paxlovid 的开发和上市也为新冠防治事业打入了一针强心剂。然而，随着新冠口服药的深入研发，是否能有进一步新的突破？让我们拭目以待。

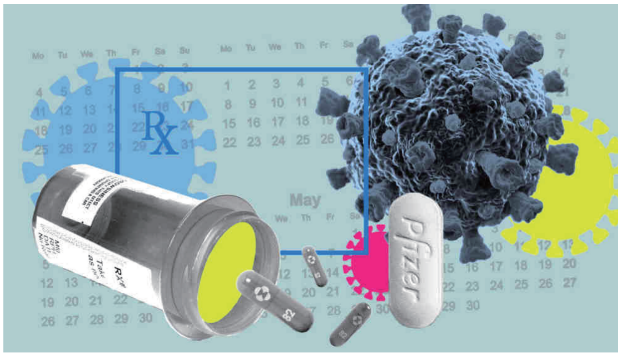



**图 7　易于服用的治疗严重的 COVID-19 的药片也可能对下一次大流行病起作用（图片来源：*MIT Technology Review* 官网）**

## 8　人工智能合成数据（Synthetic data for AI）

训练人工智能模型需要大量的数据。2021 年，尼日利亚数据科学公司的研究人员注意到，旨在训练计算机视觉算法的工程师可以选用大量以西方服装为特色的数据集，但却没有非洲服装的数据集。于是，该团队通过人工智能算法人为生成由非洲时尚服装的图像组成的数据来解决这一不平衡问题。这种通过算法人为合成出的符合真实世界情况的数据，具有与真实数据相似的统计学特征，且在数据饥渴的机器学习领域的应用越来越普遍。在真实数据稀缺或过于敏感的领域，如医疗记录或个人财务数据，这些“合成数据”可用于训练人工智能模型。实际上，合成数据的想法并不新鲜，例如，无人驾驶汽车已经在虚拟街道上进行了许多训练。2021 年，“合成数据”技术已经变得很普遍，许多初创公司和大学都在提供这种服务。例如，Datagen 和 Synthesis AI 可根据需要提供数字人脸，其他公司可为金融和保险业提供合成数据。

特别地，2021 年麻省理工学院发布了名为“Synthetic Data Vault”的开源工具，支持便捷生成不同领域、不同模态的数据。*MIT Technology Review* 关注到了数据合成方向的技术动态，并鉴于数据对智能算法的源头作用，将其列入 2022“全球十大突破性技术”。

**专家点评：**

**程学旗**　中国科学院计算技术研究所研究员、博士生导师，国家杰出青年科学基金获得者。主要研究方向为数据科学基础理论，大数据分析技术与系统，网络与社会治理大数据应用等。在国内外学术期刊与会议上发表论文 200 余篇，授权发明专利 80 余项，谷歌学术引用 20 000 余次。在数据表征学习、异构大数据广谱关联、信息检索与排序、群体分析与群智众包系统等方面取得突出成果，5 次获得本领域国际学术会议最佳论文奖。获国家科技进步奖二等奖 3 次、国家技术发明奖二等奖 1 次。

**陈薇**　中国科学院计算技术研究所研究员，博士生导师。主要研究领域为机器学习理论与算法，可信机器学习技术及其在智能算法安全中的应用。在 *International Conference on Machine Learning*、*Conference on Neural Information Processing Systems*、*International Conference on Learning Representations* 等机器学习和人工智能国际会议/期刊发表学术论文 50 余篇。2021 年入选福布斯“中国科技女性榜”。

人工智能技术已经在百姓生活和社会管理中广泛应用，例如日常购物娱乐和网络社交中的智能算法推荐、生活工作中的智能穿戴和智能算法助手、以及帮助规划调度城市高效运转的城市大脑。人工智能技术浸润着现代社会的每一个角落，已然成为世

界科技与社会发展的一大支柱。

2022 年 *MIT Technology Review* 评选出“全球十大突破性技术”，“人工智能合成数据（Synthetic Data for AI）”入选其中。如果说以深度学习为代表的智能算法是人工智能技术应用和发展的“引擎”，那么数据就是用于驱动“引擎”的“燃料”。虽然人工智能与机器学习领域的专家吴恩达认为，未来技术落地的重点将会转向数据，形成以“数据为中心的人工智能”[17]，但过去几年研究人员还是主要聚焦在模型、训练算法、或者是算力的改进上，对数据本身的关注相对较少。

有观点认为，在大数据时代，数据本身是廉价的，富有价值的是从数据中挖掘到的知识。这个观点并不完全正确。知识是宝贵的，但数据却并非廉价。人工智能模型的效果很大程度上取决于数据质量，“无效输入（Garbage In）”往往会导致“无效输出（Garbage Out）”[18, 19]。为了得到高质量的数据，需要对数据进行预处理，包括处理缺失数据和异常数据等。此外，为了提高模型训练的效果，还需要邀请领域专家人工为每一份数据附上标签，这大大地提高了数据的获取成本并制约了数据集的规模。除去获取成本高昂以外，特定领域的数据集还受限于用户隐私，极难采集。以医学影像领域为例，患者的医学影像（如 X 光片）被医院保管，医院无权泄露。这很好地保障了患者的隐私，但同时增添了领域研究者获取数据的难度。

因此，如何高效、廉价并在不侵犯隐私的情况下获取大量数据，是人工智能领域的关键问题之一。为了实现这一目标，研究人员提出了“合成数据（Synthetic Data）”的方法，即通过算法人为生成出符合真实世界情况的数据集[20-22]。合成得到的数据集可以用于人工智能模型的训练，且具有获取成本低、质量高、避免侵犯隐私等优点，有望解决目前模型训练中数据缺乏这一瓶颈问题。综上，笔者认为，*MIT Technology Review* 关注到了数据生成方向的技术动态，并鉴于数据对智能算法的源头作用，将其列入“全球十大突破性技术”。

国际上，“合成数据”技术研究的价值已经正在得到广泛认可，许多知名研究机构及科研院校都正在开展关于合成数据的项目。特别地，2021 年麻省理工学院发布了名为“Synthetic Data Vault”的开源工具，支持便捷生成不同领域、不同模态的数据[23]。此外，国际资本市场也提早预期到了“合成数据”技术的潜在价值，催生出了一批初创公司，如 AI. Reverie、Sky Engine、Datagen 等。其中，AI. Reverie 在 2021 年被 Meta 公司收购，用于支持元宇宙的开发；Datagen 在 2022 年 3 月获得 5 000 万美元的 B 轮融资。成功的商业模式正在表明“合成数据”这项技术并非只能用于实验室场景，在实际场景中也能够发挥重要作用。高纳德咨询公司在 2021 年 6 月的报告中甚至预测，到 2030 年，绝大部分用于训练人工智能模型的数据将是合成数据[24]。

我国的科研院所及商业公司也在“合成数据”领域积极进行研究探索，并取得了优秀的成果。例如，中国科学院的研究人员提出了对偶生成模型（Dual Variational Generation，DVG），该模型能够高效地生成大量现实中不存在的人脸虚拟图像，从而有效缓解异质人脸识别任务中缺乏数据及数据采集成本过高的问题[25]。商业公司也正在该领域进行有效探索，例如，支付宝公司基于实物建模技术与渲染技术提出了一套用于合成三维数据的方案，有效降低了模型训练中的数据成本，并且避免了人工标注数据带来的不可靠性[26]。相对而言，我国关于“合成数据”的研究主要着眼于服务下游任务，对“合成数据”技术本身的研究仍有待开拓。

“合成数据”领域的技术发展趋势迅猛，正在被期待对人工智能产生“再次点火”的作用。本次入选 *MIT Technology Review*“全球十大突破性技术”榜单，也将使其受到社会各界的更多关注。然而，笔者认为我们仍然需要重点关注以下几个问题：

（1）“合成数据”的评估问题。研究者们逐渐意识到，高质量的合成数据集不仅仅可以作为真实数据集的补充，更可以作为训练人工智能模型的主要数据来源。但在全面应用合成数据集之前，需要充分研究合成数据集与真实数据集的差异，从而避免应用合成数据集带来的偏差。如何评估合成数据集与真实数据集的差异仍是一个亟待解决的问题。

（2）“合成数据”仍存在“非自然数据”的问题。目前大多合成数据技术是基于统计机器学习方法的，由于经典统计学只关注了数据中蕴含的相关性，而忽视了因果性，因此有可能会生成不合逻辑的数据。例如，合成图像中可能会出现具有异常背景的图像，这类数据被称为“非自然数据”[27]。“非自然数据”对智能算法的影响目前仍然未知，尤其对算法的鲁棒性和可靠性。刻画影响的边界并提早思考应对办法将会是“合成数据”能否进入风险敏感领域的关键。

（3）“合成数据”的“隐式隐私”泄露问题。虽然“合成数据”并不由某个用户产生，但是目前的“合成数据”仍然需要借用数据来训练用于合成数据的模型，比如生成对抗网络。由于生成对抗网络结构的复杂度较高，因此在模型训练的过程中，存在记忆原始训练样本分布的可能。已经有最新研究结果表明，可以通过合成的数据反向推断出原始训练样本[28]。所以，数据合成技术存在上述“隐式隐私”泄露问题，如何更严密地保护隐私仍是有待探究的问题。

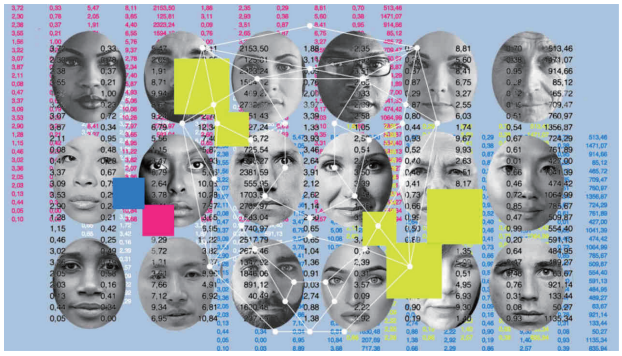


**图 8 人工智能的好处主要集中在数据资源丰富的领域，而“合成数据”有望填补领域空白。（图片来源：*MIT Technology Review* 官网）**

## 9 除碳工厂（Carbon removal factory）

减少碳排放是缓解气候变化的关键步骤，但据联合国称，这还不够。为了避免未来发生灾难性的气候变暖，我们还应采取一定的措施清除空气中的二氧化碳。2021 年 9 月，瑞士科技公司 Climeworks 开启了迄今为止最大的二氧化碳捕获工厂 Orca 的开关。该设施位于冰岛雷克雅未克的郊外，每年可捕获 4 000 吨的二氧化碳。该“除碳工厂”工作流程为：大型风扇将空气吸过一个过滤器，在那里碳捕获材料与二氧化碳分子结合；然后，该公司的合作伙伴 Carbfix，将二氧化碳与水混合，并将其泵入地下，进而与玄武岩反应，最终变成石头。该设施完全依靠无碳电力运行，电力主要来自于附近的地热发电厂。可以肯定的是，4 000 吨的年处理量并不是那么多，比 900 辆汽车的年排放量还要少。实际上，更大的“除碳”设施也在计划建设中。位于加拿大不列颠哥伦比亚省斯夸米什（Squamish）的碳工程公司，计划今年在美国西南部开始建设一个二氧化碳年处理量可达 100 万吨的工厂。此外，该公司与合作伙伴一起，也启动了苏格兰和挪威除碳工厂的工程设计工作，这些工厂将每年捕获 50 万～100 万吨二氧化碳。“除碳”企业也希望通过更多更大的“除碳工厂”建设、运行调试和操作优化，进一步降低运行成本，并实现规模经济效益。Climeworks 公司估计，到 21 世纪 30 年代末，捕集每吨碳的成本将从现阶段的 600～800 美元之间降低至约 100～150 美元。现如今，越来越多的个人及公司，包括微软、Stripe 和 Square，已经在支付高额费用来吸走空气中的二氧化碳，以努力抵消他们所产生的碳排放。而这些资金为“除碳工厂”提供了关键的早期收入。

**专家点评：**

**单文坡** 中国科学院城市环境研究所研究员，博士生导师。主要从事环境催化与大气污染控制研究，在国内外学术期刊发表论文 100 余篇。国家自然科学基金优秀青年科学基金和浙江省“万人计划”青年拔尖人才项目获得者。2019 年，以第三完成人身份获国家自然科学奖二等奖。

工业革命以来，人类活动大量排放二氧化碳（Carbon Dioxide，$CO_2$）等温室气体，使得温室效应持续加强，导致全球平均气温不断升高。2022 年 4 月 4 日，联合国政府间气候变化专门委员会（Intergovernmental Panel on Climate Change，IPCC）发布了题为《气候变化 2022：减缓气候变化》的第三工作组报告，指出 2010—2019 年全球温室气体年均排放量处于人类历史最高水平，排放量增速虽然放缓，但上升趋势并未改变；全球碳排放量必须在 2025 年达到顶峰，并在 2030 年之前削减 43%，才有机会将全球气温上升幅度控制在 1.5 ℃（与工业革命之前相比）之内。实际上，即使全世界达到了碳中和，由于工业革命以来人类已经排放了超过万亿吨的 $CO_2$，如果仅仅依靠自然过程，大气 $CO_2$ 浓度降低至工业革命前的水平也将是一个非常缓慢的过程。

作为一项利用工程系统从大气中去除 $CO_2$ 的技术，直接空气碳捕获（Direct Air Capture，DAC）技术的大规模应用对于有效降低大气中 $CO_2$ 浓度，遏制气候变化具有重要意义。该技术主要利用引风机将空气抽入，通过吸附、吸收或膜分离装置捕集 $CO_2$，并将贫 $CO_2$ 的空气排回大气，而捕获的 $CO_2$ 可以进行封存或利用，整个过程可以理解为一种工业“光合作用”。不同于针对工业固定源的 $CO_2$ 捕

获技术，DAC 可以部署在世界上任何有电力供应的地方，选址更灵活，且可以模块化建设。自 1999 年被提出以来，DAC 技术经过 20 余年的发展，已经初具实际应用的可能性。2021 年 9 月，瑞士 Climeworks 公司在冰岛启动了名为 Orca 的除碳工厂，以地热发电为主要能量来源，利用目前最大的 DAC 装置，每年可捕获 4 000 吨 $CO_2$。此次除碳工厂能够入选 *MIT Technology Review* 2022 年“全球十大突破性技术”，充分说明 DAC 技术工业化实践的重要意义。

DAC 在除碳方面具有明显的技术优势，对 Climeworks 公司 DAC 工艺的全生命周期分析也证实了其负碳排放效果[29]，但目前高昂的运行成本仍是限制 DAC 大规模应用的关键因素。近期，加州大学伯克利分校的研究人员对 DAC 技术的发展前景进行了展望，并提出了适于该技术发展的政策路线图，他们认为 DAC 的全球推广不能依赖市场杠杆效应，而应通过持续的“财政激励＋强制部署”政策推进其大规模部署[30]。另一方面，从技术角度来看，DAC 发展的关键在于高效低成本的碳捕集材料与工艺系统的研发，其商业化应用仍然需要依靠技术进步来大幅降低运行成本。

近年来，欧美发达国家已陆续开展 DAC 技术的研发与应用，通过材料与技术的进步不断降低运行成本，2021 年 8 月美国能源部宣布拨款 2 400 万美元支持 DAC 技术，一些比 Orca 更大型的除碳工厂也正在建设之中。这些先行工作可能使得发达国家更早掌握 DAC 前沿技术和核心知识产权，并为未来获取经济效益抢得先机。2020 年 9 月，在第 75 届联合国大会上，我国提出 $CO_2$ 排放力争在 2030 年前达峰，努力争取 2060 年实现碳中和的“双碳”目标，这也将我国绿色发展之路提升到了新的高度，为低碳、零碳、负碳技术的发展提供了重大机遇。当前，从实际国情出发，我国主要以产业结构和能源结构低碳转型来推动绿色发展，对 DAC 等负碳排放技术的创新和储备还相对不足。目前我国在碳捕集材料研发方面有着较为丰富的研究成果，但严重缺乏类似除碳工厂的工业化实践，以及以 DAC 为核心技术的商业化公司。为确保“双碳”目标的有序推进，我国应进一步鼓励和推动 DAC 等负碳排放技术的科技创新与应用实践，加强技术储备，抢占技术前沿，更好地参与引领全球气候治理。

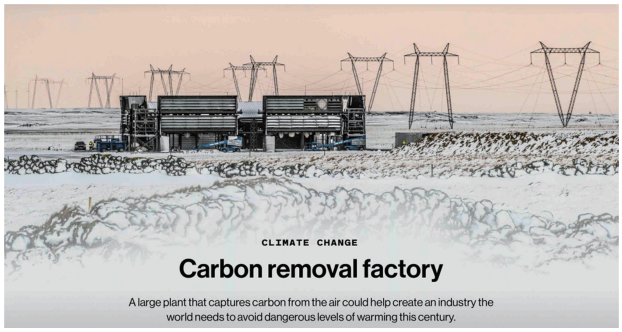



**图 9　一个从空气中捕获 $CO_2$ 的大型工厂将有助于创建一个世界需要的产业，以规避本世纪气候变暖的风险**
**（图片来源：*MIT Technology Review* 官网）**

## 参　考　文　献

[1] Buecker A, Chakrabarty B, Dymoke-Bradshaw B, et al. Reduce risk and improve security on IBM mainframes: volume 1 architecture and platform security. (2014-12-09)/[2022-06-14]. http://www.redbooks.ibm.com/redbooks/pdfs/sg247803.pdf.

[2] Wang D, Zhang ZJ, Wang P, et al. Targeted online password guessing: an underestimated threat. (2016-10-24)/[2022-06-14]. https://dl.acm.org/doi/10.1145/2976749.2978339.

[3] Lyons K. Hackers reportedly used a compromised password in Colonial Pipeline cyberattack. (2021-06-05)/[2022-06-14]. https://www.theverge.com/2021/6/5/22520297/compromised-password-reportedly-allowed-hackers-colonial-pipeline-cyberattack.

[4] Kotadia M. Gates predicts death of the password. (2004-02-25)/[2022-06-14]. https://www.cnet.com/news/privacy/gates-predicts-death-of-the-password/.

[5] Bonneau J, Herley C, van Oorschot PC, et al. The quest to replace passwords: a framework for comparative evaluation of web authentication schemes//2012 IEEE Symposium on Security and Privacy. San Francisco: IEEE, 2012: 553—567.

[6] Bonneau J, Herley C, van Oorschot PC, et al. Passwords and the evolution of imperfect authentication. Communications of the ACM, 2015, 58(7): 78—87.

[7] 汪定. 口令安全关键问题研究. 北京：北京大学，2017.

[8] Vijayan J. Apple, Microsoft are pushing passwordless; here's a reality check. (2022-02-15)/[2022-06-14]. https://techbeacon.com/security/apple-microsoft-are-pushing-passwordless-heres-reality-check.

[9] Micorsoft. Identity is the new battleground. (2022-02)/[2022-06-14]. https://news.microsoft.com/wp-content/uploads/prod/sites/626/2022/02/Cyber-Signals-E-1.pdf.

[10] Hammond J, Leister-Tebbe H, Gardner A, et al. Oral nirmatrelvir for high-risk, nonhospitalized adults with covid-19. The New England Journal of Medicine, 2022, 386(15): 1397—1408.

[11] Owen DR, Allerton CMN, Anderson AS, et al. An oral SARS-CoV-2 $M^{pro}$ inhibitor clinical candidate for the treatment of COVID-19. Science, 2021, 374(6575): 1586—1593.

[12] McDonald EG, Lee TC. Nirmatrelvir-ritonavir for COVID-19. Canadian Medical Association Journal, 2022, 194(6): E218.

[13] Hegyi A, Ziebuhr J. Conservation of substrate specificities among coronavirus main proteases. The Journal of General Virology, 2002, 83(Pt 3): 595—599.

[14] Abdelnabi R, Foo CS, Jochmans D, et al. The oral protease inhibitor (PF-07321332) protects Syrian hamsters against infection with SARS-CoV-2 variants of concern. Nature Communications, 2022, 13: 719.

[15] Li PF, Wang YN, Lavrijsen M, et al. SARS-CoV-2 *Omicron* variant is highly sensitive to molnupiravir, nirmatrelvir, and the combination. Cell Research, 2022, 32(3): 322—324.

[16] Girardin F, Manuel O, Marzolini C, et al. Evaluating the risk of drug-drug interactions with pharmacokinetic boosters: the case of ritonavir-enhanced nirmatrelvir to prevent severe COVID-19. Clinical Microbiology and Infection, 2022, doi: 10.1016/j.cmi.2022.03.030.

[17] Strickland E. Andrew ng, AI minimalist: the machine-learning pioneer says small is the new big. IEEE Spectrum, 2022, 59(4): 22—50.

[18] Rose LT, Fischer KW. Garbage in, garbage out: having useful data is everything. Measurement: Interdisciplinary Research & Perspective, 2011, 9(4): 222—226.

[19] Kilkenny MF, Robinson KM. Data quality: "garbage in-garbage out". Health Information Management Journal, 2018, 47(3): 103—105.

[20] Gupta A, Vedaldi A, Zisserman A. Synthetic data for text localisation in natural images//2016 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Las Vegas: IEEE, 2016: 2315—2324.

[21] Jaderberg M, Simonyan K, Vedaldi A, et al. Synthetic data and artificial neural networks for natural scene text recognition. (2014-12-09)/[2022-06-15]. https://arxiv.org/abs/1406.2227.

[22] Frid-Adar M, Klang E, Amitai M, et al. Synthetic data augmentation using GAN for improved liver lesion classification// 2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). Washington, DC: IEEE, 2018: 289—293.

[23] Patki N, Wedge R, Veeramachaneni K. The synthetic data vault// 2016 IEEE International Conference on Data Science and Advanced Analytics. Montreal, QC: IEEE, 2016: 399—410.

[24] Gartner. Maverick* Research: Forget About Your Real Data—Synthetic Data Is the Future of AI. (2021-06-24)/[2022-04-16]. https://www.gartner.com/en/documents/4002912.

[25] Fu CY, Wu X, Hu YB, et al. DVG-face: dual variational generation for heterogeneous face recognition. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2022, 44(6): 2938—2952.

[26] 阿里云开发者社区阿里技术. AI 训练数据不够用？支付宝 3D 合成数据方案揭秘. (2020-03-25)/[2022-04-16]. https://developer.aliyun.com/article/751561.

[27] Varga T, Bunke H. Perturbation models for generating synthetic training data in handwriting recognition machine learning in document analysis and recognition//Marinai S, Fujisawa H, eds. Machine Learning in Document Analysis and Recognition and Berlin, Heidelberg: Springer, 2008. 333—360.

[28] Hitaj B, Ateniese G, Perez-Cruz F. Deep models under the GAN: information leakage from collaborative deep learning//CCS '17: Proceedings of the 2017 ACM SIGSAC Conference on Computer and Communications Security. New York: Association for Computing Machinery Digital Library, 2017: 603—618.

[29] Deutz S, Bardow A. Life-cycle assessment of an industrial direct air capture process based on temperature—vacuum swing adsorption. Nature Energy, 2021, 6(2): 203—213.

[30] Meckling J, Biber E. A policy roadmap for negative emissions using direct air capture. Nature Communications, 2021, 12: 2051.

## Interpretation of 2022 *MIT Technology Review*'s Top10 Breakthrough Technologies

（责任编辑 魏鹏飞）