

科学计算与AI4S基础设施 白皮书

上海仪电（集团）有限公司
上海思朗科技股份有限公司
上海市人工智能行业协会

2026年7月

前 言

当前，人工智能与科学研究深度融合，推动科研范式迈入以 AI4S（科学智能）为代表的第五发展阶段，成为全球科技竞争的核心赛道。全球多国相继出台国家级战略布局 AI4S 领域，依托人工智能、大数据与高性能计算重塑生物医药、新材料、高端制造等领域研发模式，大幅缩短研发周期、降低试错成本。在我国，“人工智能+”系列政策持续落地，明确提出推动 AI 与前沿科学、实体经济深度协同，加快培育新质生产力，为科学计算与 AI4S 产业发展指明方向。

在此背景下，上海仪电（集团）有限公司、上海思朗科技股份有限公司等单位联合编制本白皮书，系统梳理 AI4S 发展现状、技术体系、市场格局与落地实践。本白皮书旨在厘清科学计算与 AI4S 的内在关联，剖析行业现存痛点与技术瓶颈，解析专用超算等核心基础设施的技术路径与应用价值。

科学计算与 AI4S 不仅是前沿科研的创新工具，更是保障产业链供应链安全、实现科技自立自强的战略底座。本白皮书立足产业实际与全球竞争格局，总结技术经验、研判市场趋势、提出发展建议，希望为科研机构、产业企业及政策制定者提供参考，助力我国构建自主可控的 AI4S 产业生态，推动新一代科研范式落地生根，充分释放科学智能的创新动能。

摘要

本白皮书围绕科学计算与 AI4S 基础设施展开全面研究，梳理科研五大范式演进历程，剖析 AI4S 带来的颠覆性科研变革、全球战略布局、核心内涵与发展瓶颈，并结合关键技术、市场规模、典型案例形成系统性结论。报告提出，算力、模型、数据是 AI4S 发展三大核心要素，其中高质量“端到端”数据集是当前行业最大“卡脖子”瓶颈，算力需求与通用大模型相比相对可控，垂类模型研发难度适中，数据短板成为制约技术规模化落地的首要因素。

科学计算是生成高精度科研数据、补齐数据短板的核心手段，依靠数值仿真与微观机理模拟，可获取传统实验方式所无法观测的底层数据，构建 AI4S 所需的标准化数据集。面对大体系的生物仿真、新材料多尺度模拟带来的海量计算压力，传统通用超算存在效率低、通信延迟高、能耗大等局限性，专用超算成为支撑高端科学计算的核心底座。

同时，报告重点解析国产“天穹”专用超算的软硬件协同技术，并梳理全球及国内 AI4S、科学计算市场高速增长态势，结合制药、新能源材料、航空航天、半导体等领域典型应用案例，验证技术落地价值。最后报告提出，科学计算与 AI4S 深度融合将成为主流科研范式，建议从国家层面统筹布局科学数据中心、专用算力集群与产学研攻关体系，加快打造自主可控的 AI4S 全产业链生态，助力我国抢占全球科技发展制高点。

目 录

第一章 AI4S 发展的时代背景	1 -
1.1 科研第五范式演进——科学智能	1 -
1.2 AI4S 带来的颠覆性科研变革	4 -
1.3 全球各国纷纷制定 AI4S 重大战略	5 -
第二章 AI4S 的概念与内涵	8 -
2.1 科学智能的概念	8 -
2.1.1 科学智能闭环中的科学计算定位: 数据生成环节的核心能力	8 -
2.1.2 模型训练与训练工厂: 对高质量数据的需求与反馈机制	9 -
2.2 AI4S 对于科研价值和意义	9 -
2.3 AI4S 需要精准解决产业核心科学问题的模型	11 -
2.4 科学智能的最大挑战: 高质量数据集的缺乏	14 -
2.4.1 算力层面: 需求相对可控, 无需巨额算力硬件投入	14 -
2.4.2 模型层面: 研发难度适中, 核心依赖专业人才支撑	15 -
2.4.3 数据层面: 核心短板凸显, 高质量数据集“卡脖子”	16 -
第三章 构建服务 AI4S 的精准、端到端的高质量数据集	20 -
3.1 AI4S 的数据核心要求: 精准无错、“端到端”串联	20 -
3.2 “端到端”数据首先是对微观机理的精准描述	22 -
3.3 “端到端”数据集的构建思路: 转变传统思路, 逆向构	

建高质量科研数据“教材”	- 24 -
3.4 数据构建的核心属性：高成本、高创新性，需国家层面 牵头推进	- 29 -
第四章 科学计算与 AI4S 落地应用关键技术分析	- 31 -
4.1 科学计算是获得高质量数据集的关键手段	- 31 -
4.2 面临计算瓶颈：仿真模拟任务带来庞大的计算量	- 33 -
4.2.1 生物仿真计算领域的算力瓶颈	- 33 -
4.2.2 新材料研发中面临的计算卡点	- 35 -
4.3 传统超算面对复杂科学计算任务的局限性	- 36 -
4.4 专用超算应运而生	- 37 -
4.4.1 各国发展现状	- 37 -
4.4.2 应用领域现状	- 39 -
4.5“天穹”专用超算关键技术分析	- 40 -
4.5.1 核心策略：算法级全局优化计算	- 40 -
4.5.2 关键技术 1：专用芯片	- 43 -
4.5.3 关键技术 2：专用系统架构	- 45 -
4.5.4 关键技术 3：专用计算软件	- 46 -
第五章 科学计算与 AI4S 市场规模分析	- 48 -
5.1 全球及我国 AI4S 市场概况	- 48 -
5.2 国内科学计算市场规模分析	- 49 -
第六章 科学计算与 AI4S 典型应用案例	- 55 -
6.1 爱赛思 OpenAI4S 社区：开放式科学智能基础设施	- 55 -
6.2 谷歌从 AlphaFold 到 IsoDDE 推动 AI 制药落地	- 57 -
6.3 礼来携手英伟达共建 AI 联合实验室	- 58 -

6.4 思朗科技打造基于“天穹”的制药新引擎 sDDE.....	- 59 -
6.5BASF 引入 Quriosity 超算赋能新产品研发	- 59 -
6.6 字节跳动推出 BAMBOO 助力新型电解质设计....	- 61 -
6.7 思朗科技基于“天穹”打造电解液 AI 数据平台	- 62 -
6.8 美国波音公司以虚拟仿真计算加速新机型研发 ...	- 63 -
6.9 台积电采用英伟达计算光刻平台突破先进芯片制程-	64
-	
第七章 总结与展望	- 66 -
7.1 新一代科研范式将由科学计算与 AI4S 深度融合驱动， 成为引领科技创新的核心引擎	- 66 -
7.2AI4S 能力跃升将持续释放产业创新效能，为新质生产力 形成提供关键技术路径.....	- 66 -
7.3 科学计算与 AI4S 的市场潜力将持续释放，形成规模增 长与生态共赢的良性发展格局	- 67 -
7.4 构建自主可控的 AI4S 生态是实现大国科技自立自强的 战略必争方向	- 67 -
7.5 政策与行动建议.....	- 68 -
附录 缩略语中英文对照表	- 69 -

第一章 AI4S 发展的时代背景

1.1 科研第五范式演进——科学智能

人类科学研究在数百年的发展进程中，逐步形成了实验科学、理论科学、计算科学、数据密集型科学、AI 驱动的科学发现（即 AI4S, 科学智能）五大范式的演进，它们从经验感知到理论提炼，从仿真模拟到数据洞察，依次迭代、层层递进，共同构筑起现代科学研究的完整方法论体系，推动人类对世界的认知不断走向深入，也为科技创新与社会发展提供着持续不竭的动力。



图 1 科研第五范式的演进

第一范式实验科学——是科学研究的原始形态与根基，以观察自然、开展可控实验、总结实践经验为核心，旨在回答“事物是什么”的基本问题。从古代朴素的天象观测、技艺实践，到近代伽利略确立“实验+逻辑”的规范方法，实验科学通过可重复、可验证的操作，实现了从定性描述到初步定量分析的跨越，为科学发展积累了大量真实可靠的观测数据与事实依据。但其局限明显，难以涉足极端环境、微观领域与复杂系统，多停留在现象层面总结。

第二范式理论科学——在实验成果的积累之上应运而生，以

数学建模、逻辑推理、理论演绎为核心，致力于揭示“事物为何如此”，实现从“知其然”到“知其所以然”的跃升。科学家依托严谨的逻辑与数学工具构建理论体系，诞生了牛顿力学、电磁场理论、相对论等里程碑成果，不仅能统一解释已有现象，还可精准预测未知存在。理论科学具备高度抽象性与普适性，但在生物、社会等复杂系统中难以建立精准模型，且严格的前提假设易与现实产生偏差。

第三范式计算科学——伴随计算机技术兴起，成为应对复杂问题的关键手段。它以高性能算力为支撑，通过数字化建模与数值仿真，模拟传统实验无法开展、理论方程难以解析的场景，广泛应用于生命科学、材料科学、航空航天、气象预报、核科学研究等领域，大幅降低研发成本与安全风险。仿真结果的准确性高度依赖模型设计与参数精度，同时对硬件算力提出较高要求，成为实验与理论范式之外不可或缺的重要补充。

第四范式数据密集型科学——是大数据与云计算兴起的全新科研形态，以海量多维度数据为核心资源，通过数据挖掘、统计分析等直接从数据中总结规律，此举打破了“先假设、后验证”的传统路径，实现了“数据驱动知识发现”的全新路径。在精准医疗、天文探索、环境监测、智能推荐等领域，数据驱动模式高效捕捉隐藏关联，展现出强大的应用价值。该范式受数据质量、伪关联等问题制约，却标志着科学研究开始迈入智能化、高效化发展的新阶段。

第五范式 AI 驱动的科学发现——是近年来随着人工智能技术（尤其是深度学习、强化学习）的飞速发展，在第四范式基础上演化而来的新一代科研范式，核心是“将人工智能、大数据、高

性能计算与领域科学深度融合，实现全自动、智能化的科学发现”，能够加速从基础研究到产业应用的转化，甚至实现“从 0 到 1”的颠覆性创新。这一范式的产生，源于前四种范式的局限性——实验科学效率低、成本高，理论科学难以处理复杂系统，计算科学依赖人工建模，数据密集型科学难以实现自主发现，而 AI 技术能够自主学习、自主优化、自主决策，恰好弥补了这些不足。

AI4S 的核心逻辑是利用 AI 技术（如深度学习、强化学习、生成式 AI）自主处理海量多模态数据，自主构建模型、优化参数，自主设计实验方案、预测实验结果，甚至自主提出新的科学假设，实现“数据输入→AI 分析→规律发现→实验验证→知识生成”的全流程自动化。其典型应用场景在前沿科学领域取得了突破性成果：AlphaFold 利用深度学习技术，仅通过氨基酸序列，就能精准预测蛋白质的三维结构，解决了困扰生物学界半个多世纪的难题，极大加速了生物医药研发进程；AI 辅助新药研发中，AI 能够自主筛选药物分子、预测药物疗效和毒性，将传统新药研发的周期从 10 年以上缩短至数年，大幅降低研发成本；材料科学领域，AI 能够自主设计新型材料的成分和结构，预测材料的性能，为新能源、航空航天等领域提供新型材料支撑。

AI4S 的核心特征是“**智能化、自主性、跨尺度融合**”：它实现了科研全流程的自动化，AI 能够自主学习领域知识、处理复杂数据、优化研究方案，无需人工过多干预；同时，它融合了大数据、算力、AI 技术和领域科学，能够处理跨尺度、跨领域的复杂问题，实现多学科的交叉融合。但其局限性在于，AI 模型的“黑箱效应”导致研究过程难以解释，结论的可解释性较差；同时，AI 模型的训练需要海量高质量数据和强大的算力支撑，且对领域知

识的融合要求较高，目前仍处于快速发展阶段。作为新一代科研范式，AI4S 正在重塑科学研究的模式，加速科学创新的进程，代表了未来科研的主流发展方向。

1.2 AI4S 带来的颠覆性科研变革

实际上，AI4S 已经在诸多行业前沿领域已经展开了实质探索，并带来远超传统研发模式的惊人成效，尤其在生物医药、新材料等领域。

在生物医药领域，2018 年美国 AlphaFold 横空出世，其通过 AI 预测蛋白质结构彻底革新了生命科学研究范式，这一技术不仅解决了困扰科学界数十年的蛋白质折叠难题，还推动了生命科学研究进入“AI 驱动”的新纪元。AlphaFold 的成功让 AI4S 的理念深入人心，证明 AI 不仅是辅助工具，更能成为攻克基础科学难题的核心力量。它摒弃了传统依赖实验试错和经验推导的研究逻辑，通过深度学习从海量生物数据中挖掘规律，为生命科学提供了全新的研究思路。2025 年 10 月，英伟达和礼来合作共建全球首个制药专属“AI 超级工厂”，该系统由超过 1000 块 Blackwell Ultra GPU 组成，旨在通过管理从数据输入、模型训练到推理的完整 AI 生命周期，加速药物研发的全流程，其潜力在于探索仅靠人类难以发现的新分子，缩短早期药物发现周期、降低临床研发成本。2026 年 3 月，罗氏与 Nvidia 合作打造全球首个医疗领域的“AI 工厂”，这是一座大规模、高算力的数字基础设施，旨在重塑药物研发范式。最终通过尖端算力与制药原理的深度融合，为重塑药物研发范式和推动个性化精准医疗奠定了重要基础。

在新材料领域，新一轮材料研发范式变革的探索中，材料计

算正成为突破传统实验试错路线的核心驱动力，AI4S 正为新材料设计研发模式带来颠覆式变革：美国谷歌 Deepmind 仅 1 年内发现 38 万种稳定无机材料，远超此前人类已知无机材料总量；美国微软研究院开发的生成式 AI 材料设计工具仅用 80 小时便从 3200 万种材料中筛选出 18 种高潜固态电解质，并迅速验证其中 1 种具有光催化、介电、磁性等多功能潜力材料；日本东北大学和麻省理工学院联合开发了新型人工智能工具成功发现 246 种太阳能转换效率超过 32% 的材料，为加速能源和量子材料的发现提供可能性；宁德时代基于计算模拟和人工智能技术从 50 万种候选分子中高效筛选出 26 种电解液添加剂，大幅提升了新型电解液配方研发效率；中石油、中石化则都基于自身产业优势，分别承接开展“高性能合成橡胶材料研发”和“AI 辅助聚酰亚胺气体分离材料开发”，探究自主可控高性能算力赋能领域大模型；万华化学与华为合作利用人工智能技术使得研发新产品效率提升 30% 以上，并利用人工智能技术对 400 多种产品 1000 多种牌号产品进行生产管理和控制，实现产品质量完全一致性；中铝集团打造材料数字化研发平台，通过全流程数字化仿真、数据分析和机器学习、研发数据库及管理转变高端新铝合金材料研发模式和理念，推动高端新材料研发的原始性创新。

1.3 全球各国纷纷制定 AI4S 重大战略

当前，AI4S 已成为全球科技竞争的核心赛道，以算力为底座、数据为支撑、AI 为引擎、科学机理为内核，引领科研范式全方位变革。它不仅能加速重大科技原创突破，抢占未来科技制高点，还能支撑战略产业升级，依托自主可控的数据集与智能模型，筑牢国家科研数据安全与科技自主屏障，为国家科技创新战略落

地提供核心支撑，助力国家科技自立自强与人类科学进步。

从国外看，随着全球范围内的 AI4S 进入战略落地期，欧美等国家纷纷发布国家级 AI 战略计划，深入推动 AI4S 实施。2025 年 7 月，美国白宫发布《美国 AI 行动计划》，明确将“高质量数据”定位为“国家战略资产”，提出制定生物、材料科学、化学等领域科学数据用于人工智能训练数据标准，并构建世界一流的科学数据集，为训练下一代大模型提供“燃料”。2025 年 11 月，白宫再次发布国家级 AI 战略“创世纪计划”，其地位被定位为“AI 曼哈顿计划”，旨在通过人工智能彻底变革科研范式，加速科学发现进程，其核心举措是由美国能源部主导构建“美国科学与安全平台”（American Science and Security Platform），平台整合联邦数据、超级算力与 AI 模型资源，重点攻坚先进制造、生物技术等六大战略领域。2025 年 6 月，英国启动“Openbind”大科学计划，拟构建全球最大的蛋白小分子交互数据库，其为 AI 药物设计算法“量身定制”，将节省千亿英镑药物研发成本。此外，日本发布《综合创新战略 2025》，将 AI、量子技术和核聚变能源列为国家战略重点；德国新联合政府在执政协议中确立 AI 与量子技术为研究与创新资助的优先方向；澳大利亚推出国家人工智能计划，旨在推动 AI 在各经济领域的深度应用与人才培养。

从国内看，国务院于 2025 年 8 月印发《关于深入实施“人工智能+”行动的意见》，提出大力推进人工智能规模化商业化应用，强化算力、算法和数据供给，加强人工智能与生物制造、新材料等领域技术协同创新，推动人工智能驱动的技术研发、工程实现、产品落地一体化协同发展。2026 年 1 月，工业和信息化部、国家发展改革委、国家数据局等 8 部委发布《“人工智能+制造”专项

行动实施意见》，提出到 2027 年打造 100 个工业领域高质量数据集，推广 500 个典型应用场景，人工智能关键核心技术实现安全可靠供给，产业规模和赋能水平稳居世界前列。2026 年 3 月，国务院印发《中华人民共和国国民经济和社会发展第十五个五年规划纲要》，提出加快数智技术创新，深化拓展“人工智能+”，赋能经济社会发展和治理能力提升，促进生产方式深层次变革和生产力革命性跃迁；全面实施“人工智能+”行动，加强人工智能同科技创新、产业发展等相结合，抢占人工智能产业应用制高点，全方位赋能千行百业。2026 年 3 月，第十四届全国人民代表大会第四次会议审议了《2026 年政府工作报告》，报告提出深化拓展“人工智能+”，促进新一代智能终端和智能体加快推广，推动重点行业领域人工智能商业化规模化应用，培育智能原生新业态新模式；推动科技创新和产业创新深度融合，建设北京（京津冀）、上海（长三角）、粤港澳大湾区国际科技创新中心，打造世界级科技创新策源地，并强化企业创新主体地位，支持科技领军企业牵头组建创新联合体，提高承担国家重大科技项目比例。

第二章 AI4S 的概念与内涵

2.1 科学智能的概念

科学智能是面向科学研究领域的新兴方向，可概括为“数据生成—模型训练—生态反馈”三个环节。其中，科学计算定位于数据生成阶段，是高质量科研数据的重要来源和科研算力的核心支撑。其通过高性能计算（HPC）提供强大的算力基础，并结合高精度数值仿真手段，在计算机中重建真实科学过程得到大量的高质量数据集，从而为后续的科学智能模型训练提供可靠的“燃料”。

2.1.1 科学智能闭环中的科学计算定位：数据生成环节的核心能力

科学计算在科学智能闭环中承担“数据生成”的核心职能，能力本质由高性能算力与高精度数值仿真共同构成。在算力侧，科学计算提供稳定可扩展的算力底座，重点解决“算力供给与算效提升”，典型能力包括双精度（FP64）高精度计算、超大规模并行与调度，以及面向海量数据的高吞吐处理；在需求侧，随着效率、能耗比、可控性与交付效率约束增强，科学计算正由追求通用峰值转向面向特定负载的体系化优化，通过专用加速器、互联网络、并行软件栈与整机系统的协同创新提升任务算效。

高精度数值仿真是科学计算生成可靠数据的关键手段，其核心是以机理模型为驱动实现“可预测、可验证、可复现”的高保真模拟。数值仿真基于物理、化学与工程机理模型，通过将微分方程、守恒定律、传输方程等离散求解，在虚拟环境中重构复杂系统行为，关注“算什么、如何建模、是否符合真实机理”，从而批量生成贴近真实过程的高质量数据集，为后续数据驱动模型训练

提供可控、可信的数据基础。

科学计算产业的核心价值，在于持续供给高可信数据，并提炼能够驱动科学理论优化与创新的关键洞见。依托“仿真+算力”的组合，科学计算为科研与工程提供可规模化的“虚拟实验”能力，科研人员可以通过虚拟实验实现模型理论验证、优化设计与探索创新，同时持续沉淀高质量数据集。

2.1.2 模型训练与训练工厂：对高质量数据的需求与反馈机制

科学智能模型是建立在科学计算长期积累之上的能力延伸层，其核心是在科学数据与机理模型基础上引入机器学习/深度学习。围绕科学问题，将 AI 方法嵌入模型构建、参数寻优、仿真加速、结果反演与不确定性分析等关键环节，贯穿科学发现、模型求解、仿真计算、实验验证与迭代优化全过程，从而系统性提升科研效率与问题求解能力。

在这一机制中，科学计算提供“机理建模与高可信数据生成”的底座，而科学智能模型侧重于对计算过程与结果进行智能化利用并形成闭环反馈。通过联合学习仿真数据、实验数据与历史科研数据，科学计算从单一“模型驱动”演进为“模型—数据协同驱动”的新模式：AI 模型在物理一致性约束下可对复杂数值仿真进行近似替代或快速预测以降低成本、缩短周期；同时科学计算生成的高质量仿真数据为训练提供可信、可控来源，缓解科学问题中的数据稀缺与泛化风险。

2.2 AI4S 对于科研价值和意义

AI4S 依托智能模型、深度学习与科学计算融合能力，将彻底打破传统科研的边界桎梏，重塑科学发现、技术攻关、产业转

化的全链路逻辑，为基础科研深耕、硬核技术突破、交叉学科创新赋予全新底层动能，其科研价值与战略意义贯穿基础研究、应用攻关、生态变革三大核心维度。

在基础研究层面，AI4S 打破传统科研的探索壁垒，重塑科学发现的底层逻辑。传统基础研究受实验条件、理论建模局限，难以突破微观机理观测、极端工况复刻、海量变量分析等瓶颈，探索周期漫长且试错成本高昂。AI4S 依托科学计算沉淀的高保真数据集与物理约束智能算法，实现“以算代试、智能预判”，可快速拟合科学机理、筛选核心变量、挖掘隐性规律，同时打通理论推演与实验验证的断层，用仿真数据修正理论偏差，用实验数据校准仿真模型，大幅缩短基础研究探索周期，助力科研人员破解实验不可观测、机理难以建模的天然壁垒，拓展人类认知边界，推动基础科学理论的创新与突破。

在应用攻关层面，AI4S 赋能硬核领域突破，筑牢关键技术自主根基。针对量子科学、生命科学、高端材料、航空航天、新能源等战略领域的“卡脖子”难题，AI4S 融合高性能计算与智能建模能力，破解海量复杂变量耦合、隐性机理解析、极端工况仿真等技术瓶颈，助力核聚变仿真、蛋白质结构预测与新药设计、新型材料原子级设计、新型高能量密度电解质开发乃至新一代大涵道涡扇发动机等核心攻关。同时，它推动跨学科融合，打破物理、化学、生物、工程等学科壁垒，推动跨域数据融合与多机理协同建模，催生智能材料学等新兴交叉学科，激活应用攻关的创新动能，为关键核心技术自主可控、国家战略科技力量夯实提供有力支撑。

在生态变革层面，AI4S 革新科研生产模式，构建现代化智

能科研生态。传统科研存在资源集中、门槛较高、成果复用性差、产学研脱节等问题，大量科研数据闲置、重复研发现象突出。AI4S 依托标准化科学计算数据集、轻量化智能模型与普惠式算力调度，推动科研资源下沉，降低创新门槛，让中小科研机构、高校及企业等更多创新主体参与前沿攻关，盘活全域科研资源。同时，它将分散的科研数据、实验结论、机理模型等科研成果，转化为可积累、可复用、可迭代的标准化知识资产，形成“数据积累—模型优化—成果落地—数据再沉淀”的正向循环，在避免重复研发的同时更打通基础研究与产业落地的“最后一公里”，构建高效协同、开放共享的现代化智能科研生态。

2.3 AI4S 需要精准解决产业核心科学问题的模型

科学智能模型的核心目标，是精准解决产业实际问题，这一定位从根本上区别于通用大模型以内容生成、对话交互、多任务泛化适配为主的目标。通用大模型追求更强的创作能力、更广的知识覆盖和更自然的语言表达，擅长开放式问答、文本生成、逻辑推理等通用性任务，其输出偏向完整、流畅、具备普适性的内容；而科学智能模型以产业需求为导向，以产业实际问题为核心，聚焦问题求解、结果可信、答案可落地，不追求泛化表达，而是针对产业场景中明确、具体、可量化的科研与工程问题，给出高精确度、高专业性、高可靠性的解决方案，实现从提出问题到直接给出可用答案的闭环。

科学智能模型深度嵌入制药、新材料等高端制造场景，在实际产业场景中，不同领域的科研问题具有极强的指向性、专业性和约束性，无法依靠通用生成式内容满足需求。科学智能模型正是围绕这种高度聚焦的问题进行构建，通过深度融合领域知识、

实验数据、仿真规律与理论公式，形成垂直领域的专业推理与预测能力，直接回应产业研发中的关键卡点，为产业研发降本、提速、增效，从而成为支撑科学研究与产业创新深度融合的重要基础设施。

以制药领域为例，药物研发链条长、投入大、风险高，每一个环节都需要精准答案而非泛化信息，科学智能模型往往需围绕特定疾病靶点展开深度推理：可以精准预测与靶点空间结构相匹配的潜在成药物质，识别蛋白质结构中的隐藏位点与结合口袋；同时，甚至还能进一步预测候选分子的药代动力学性质，包括吸收、分布、代谢和排泄等关键参数，评估其靶器官毒性、急性与慢性毒性、遗传特殊毒性等毒理学特征，甚至可基于已有临床数据与分子特征，提前预判候选药物在临床一期、二期中可能出现的有效性、安全性与剂量反应关系。这种从分子设计到临床预判的全链条精准预测，将直接服务于药物快速筛选、降低实验成本、缩短研发周期，是通用大模型难以企及的专业价值。

在新材料研发领域，科学智能模型同样需体现出极强的问题导向与实用价值。新材料开发高度依赖配方设计、组分调控与工艺优化，传统模式依赖大量试错实验，周期长、成本高。科学智能模型则需针对电池电解液、特种钢、特种铝合金、特种高性能高分子材料等关键品类的研发需求，精准回答核心技术问题：例如在电解液开发中，预测添加何种添加剂、溶剂、锂盐配方可提升离子电导率、宽温域稳定性与阻燃性；在特种金属材料中，给出合金元素配比与制备工艺建议，以实现高强度、耐腐蚀、耐高温等目标性能；在特种高分子材料领域，通过结构-性能关系推理，确定单体种类、共聚比例、官能团修饰方式，从而获得满足介电

性能、力学强度、耐老化性等特定指标的新材料方案。这些答案直接指导实验设计，减少无效试错，加速新材料从实验室走向工业化应用。

专栏 1 科学智能模型的研发需聚焦产业关键研发卡点

科学智能模型的核心方向必须是聚焦产业真实存在的关键研发卡点，而非停留在浅层工具化应用。然而现实中行业普遍存在误区：大量模型研发趋于同质化，主要集中在文献检索、论文速读、信息整理、科研辅助写作等表层功能，虽然在一定程度上提升了科研效率，却并未真正切入产业研发链条中的核心科学问题。这类应用本质上只是对科研流程的效率优化，而非对科学难题的本质突破，与真正意义上支撑重大原始创新、解决“卡脖子”问题的 AI4S 的初衷目标相去甚远。

对标美国“创世纪计划”等面向科学重大突破的 AI 模型体系可以看出，真正的科学智能并非由算法团队闭门造车，而是以问题为源头、以应用为导向。一般而言，具备颠覆性价值的科学问题，必须由一线应用场景、产业龙头企业和资深科学家来提出。这类问题高度依赖产业深耕经验、长期研发积累和对技术瓶颈的深刻理解，涉及复杂机理、极端条件、多尺度耦合等深层规律，模型厂商仅靠理论研究和公开数据根本无法预判，更难以自主定义出真正有产业价值的攻坚方向。因此，科学智能的研发起点，理应是产业界，而非单纯的技术供给端。

从产业落地需求来看，科学智能的核心价值，正是解决企业自身难以攻克的 5%-10% 的瓶颈问题，而非通用型、共性基础技术问题。以国内龙头企业为例，锂电行业的宁德时代、钢铁行业的鞍钢、攀钢，生物医药行业的恒瑞医药等，在成熟工艺、常规研发、基础技术上已建立极强优势和完整体系，但在前沿方向仍存在难以突破的卡点。例如锂电领域新型高安全、高倍率、宽温域电解液开发；钢铁领域特殊工况下高强、耐高温、耐腐蚀特种钢材设计；生物医药领域新靶点发现、隐藏蛋白位点挖掘、高成药性小分子快速筛选等。这些问题往往周期长、试错成本高、机理复杂，传统实验与理论手段难以高效解决，正是 AI4S 模型需要精准发力的核心战场。

由此可见，科学智能模型必须确立“先明确产业关键问题，再针对性研发模型”的正确逻辑。目前行业大量模型仍停留在文献助手、

信息检索等基础层面，虽有实用价值，但远非 AI4S 的核心价值所在。真正具有战略意义和颠覆性的科学智能，应对标国际前沿科学 AI 项目，把技术能力真正注入到科学发现与研发突破中。而要实现这一目标，就必须坚持产业牵头，由企业提出真问题、难问题，AI 企业围绕问题构建专用模型、数据体系与推理能力，形成“产业出题、AI 解题、共同验证、落地迭代”的闭环研发模式。这不仅是 AI4S 健康发展的关键路径，也是其实现真正科学价值与产业价值的核心出发点。

2.4 科学智能的最大挑战：高质量数据集的缺乏

AI 模型的稳定运行与高效落地，始终依赖数据、算力、模型三大核心要素的协同支撑，三者相互关联、缺一不可：数据是模型训练的“燃料”，决定了模型的学习边界与应用精度；算力是模型运行的“引擎”，提供了数据处理与模型迭代的硬件支撑；模型是 AI 的“核心算法框架”，定义了数据转化为价值的路径与逻辑。但在 AI4S 这一聚焦生物医药、新材料等前沿领域的细分赛道中，三大要素的重要性与挑战难度呈现显著不均衡态势——数据已成为制约科学智能发展的最大核心瓶颈，其构建难度、资源消耗与落地阻碍，远远超出算力与模型层面的挑战，具体原因可从三大要素的行业特性展开详细分析。

2.4.1 模型训练算力：相对可控，门槛不高

与通用大模型追求“大参数量、广覆盖性”的发展路径不同，科学智能领域的 AI 模型更注重“精准性、场景化”，其算力需求处于可控范围，无需投入巨额资金搭建超大规模算力集群。实际上，直接训练一个 AI4S 模型的算力需求可控，但为这个模型生产“燃料”（即通过分子动力学、DFT 等仿真生成端到端数据集）的科学计算算力需求是极其庞大且特殊的，这是当前真正的算力“卡脖子”环节。从参数量来看，当前科学智能领域的最大模型

其参数量（多为百亿级以下）远远少于通用大模型（数千亿乃至万亿规模）——作为通用大模型领域的代表性产品，OpenAI 的 GPT-3 拥有 1750 亿参数，华为盘古 UltraMoE 拥有 7180 亿总参数，字节豆包 2.0 的总参数量达 1 万亿，DeepSeek-V4 的总参数量达 1.6 万亿，阿里 Qwen3-Max-Thinking 模型总参数量超 1 万亿，而生物领域的 AlphaFold2 的参数量约为 9300 万左右，目前最前沿的 AlphaFold3 的参数量仅为亿级（参数量约 3.7 亿），这也使得仅需少量 GPU 即可完成完整的 AI4S 模型训练与推理任务。例如，在生物医药领域的分子结构预测、新材料领域的性能仿真等场景中，单台高性能服务器搭配 4-8 块专业 GPU，即可满足模型训练的算力需求，无需像通用大模型那样依赖成百上千块 GPU 组成的算力集群。这种可控的算力需求，使得科学智能领域的算力成本相对较低，也降低了行业进入门槛，算力层面的挑战已不再是制约科学智能发展的核心问题。

2.4.2 模型算法研发：研发难度适中，核心依赖专业人才支撑

科学智能领域的模型研发，并未走通用大模型“从零研发、全面突破”的重资产投入路径，而是基于现有通用大模型（如 DeepSeek、千问等）进行微调优化，或针对性研发垂类小模型，其研发难度远低于 OpenAI 研发 ChatGPT、阿里研发千问等通用大模型的难度。通用大模型的研发需要投入数十亿甚至上百亿资金，组建数百人的跨领域研发团队，覆盖算法、工程、数据等多个环节，且面临技术突破、算力支撑、数据积累等多重挑战；而科学智能模型的研发核心，往往不在于资金与团队规模，而在于专业人才的技术深度——需要具备计算数学、理论物理等基础学

科能力，或对生物、材料等具体产业有深度理解的“天才型”专业人才。这类人才的核心价值，在于能够从零散的产业数据中挖掘背后隐藏的共性物理规律、化学原理或生物特性，将产业需求转化为模型训练的目标，让 AI 模型能够精准学习并提炼这些核心规律，从而实现科研场景的落地应用。

例如，在新药研发场景中，具备药物化学与计算数学背景的人才，能够从药物分子实验数据中提取结构-活性关系规律，指导 AI 模型优化分子设计；在新材料研发中，懂理论物理与材料科学的人才，能够基于材料性能测试数据，构建 AI 模型的训练逻辑，提升材料性能预测的精度。这种“轻规模、重专业”的研发模式，使得相对年轻的专业人才也能快速上手开展研发工作，无需依赖庞大的团队与巨额的算力资金投入，因此模型层面的挑战也远低于数据层面。

2.4.3 数据层面：核心短板凸显，高质量数据集“卡脖子”

在科学智能领域，高质量数据集是真正的“卡脖子”环节，其构建难度、资源消耗与落地阻碍，远超算力与模型投入，甚至远超出了当前行业的普遍预期。不同于通用大模型可通过互联网公开数据、用户交互数据等快速积累海量训练数据，科学智能领域的训练数据多来自产业一线的科研实验、生产实践，其获取、整理、整合与应用面临多重难以突破的困境，具体体现在以下方面：

（1）产业数据存量不足且难以整合，有效数据供给稀缺

科学智能的核心应用场景集中在生物医药、新材料、高端制造等领域，这些领域的药厂、化工厂等企业目前仍以传统实验研发范式为主，大部分研发过程的核心数据的记录、存储方式较为传统，研发数据多分散在实验报告、纸质记录、内部系统中，未

进行统一的数字化整理与标注，即便投入人力物力进行数据清洗，能够转化为模型训练可用数据的比例也极低，导致可直接用于 AI 模型训练的有效数据极少。

更关键的是，不同企业之间的数据存在严重的“信息孤岛”问题，数据互通性极差。由于科研数据涉及企业核心研发机密、技术专利等核心利益，即便行业内建立数据贡献的激励机制，引导企业共享数据，单家企业能够拿出的有效训练数据依然十分有限——多数企业仅愿意共享非核心的、零散的数据，而核心的实验数据、研发数据根本不会对外披露。这种“存量不足、整合困难、核心数据保密”的现状，导致科学智能模型训练所需的海量、高质量、标准化数据严重短缺，成为制约科学智能模型研发与落地的首要障碍。

专栏 2 AI 底层逻辑：数据决定模型能力，无规律则无创造

截至目前，AI 技术的底层基础理论仍未脱离机器学习的核心范畴，各类主流模型架构均围绕“数据驱动规律提取”展开核心运作。无论是应用于计算机视觉领域的卷积神经网络（CNN），还是支撑自然语言处理与通用大模型发展的 Transformer 架构，其核心能力均聚焦于从海量输入数据中挖掘、提炼已存在的关联规律，通过模型参数的迭代优化，将这些规律固化为可复用的算法逻辑，并以“黑盒子”模式实现规律的高效应用。需要明确的是，此类模型的核心局限在于“规律提取而非创造”，即无法突破数据本身的边界，生成数据中未包含的全新规律，其所有输出均是对现有数据中规律的重组与复用。

这一核心逻辑同样适用于通用大模型与科学智能模型的研发。通用大模型之所以能够实现生成式回答、多场景交互等功能，本质上是其训练数据集覆盖了互联网海量文本、图像等多元信息，囊括了各类问题的知识体系、逻辑关联与表达范式，模型在大规模训练中充分学习并掌握了这些隐含规律，进而能够基于规律生成符合场景需求的输出。科学智能模型作为聚焦生物医药、新材料等科研领域的垂类模型，其研发逻辑与通用大模型高度一致，不存在脱离数据规律的“自主创

新”，模型的能力边界、应用精度，始终由训练数据所包含的规律范围与质量水平所决定。

从技术本质来看，计算机的核心价值定位是高效执行，即按照预设指令或已学习的规律，快速完成数据处理、逻辑运算、结果输出等任务，其核心功能并非“创造性设计”。AI 技术作为计算机技术的延伸与升级，虽能在特定场景下呈现出一定的创造性，例如生成新颖的科研假设、优化分子设计方案等，但这种创造性始终建立在数据中存在可提取规律的基础之上。若训练数据本身缺乏核心规律、信息片面或质量低下，AI 模型将无法实现真正的创造性突破，仅能完成现有信息的低质拼接与无效重复，甚至可能输出偏离科研逻辑的错误结果。

基于上述逻辑，科学智能模型的研发核心可概括为“**什么样的数据，喂出什么样的模型**”，**端到端的高质量数据是科学智能模型具备核心能力的根本保障**。与通用大模型相比，科学智能模型的应用场景高度垂直、专业门槛极高，对训练数据的要求更为严苛，不仅需要保证数据的完整性、准确性与标准化，更需要构建数据间的逻辑关联，形成覆盖科研全流程的端到端数据链条。唯有夯实数据基础，确保模型能够从数据中提取到科研领域的核心规律，才能让科学智能模型真正赋能生物医药、新材料等领域的科研创新，实现研发效率与成果质量的双重提升。

（2）“平台+生态”思路也难以解决核心数据问题

当前，国内科学智能发展较快的地区，不少研发主体试图通过“搭建平台、构建生态”的思路，汇聚各类模型与数据，试图解决数据短缺的问题。这种思路仅能解决少量表层问题，但却难以支撑能真正解决产业核心应用问题的模型研发。**科学智能目前仍处于“重大科学挑战”的初级阶段**，其核心任务是突破关键科学问题、研发能够适配产业需求的核心模型，尚未发展到通用大模型那样的成熟生态阶段——通用大模型的“平台+生态”模式，是建立在核心模型已经成熟、数据积累已达到一定规模的基础上，通过汇聚第三方模型与数据，实现生态的迭代升级；而科学智能领

域，目前尚未突破核心科学问题，也未研发出类似 ChatGPT 那样的核心基础模型。在这种情况下，单纯堆砌模型和数据，无法实现核心技术突破，也难以通过自然融合的方式得到一个“科学智能的 ChatGPT”。

更值得注意的是，科学智能领域的数据缺失，并非“机制问题”，而是客观存在的产业现状。部分机构试图通过设计制度、规则、激励机制，鼓励企业贡献数据，但当前的核心矛盾恰恰是产业中本身就缺乏有效、可串联的端到端数据——即便企业愿意响应激励、贡献数据，现有数据的质量、标准化程度也无法满足模型训练的需求，更何况部分核心研发数据、商业机密数据，企业从根本上就不愿意对外披露。这种源自供给端本身数据短缺的困境，使得“平台+生态”的思路难以发挥实际作用，也进一步凸显了数据问题的核心挑战性。

综上，在科学智能领域，AI 模型的三大构成要素中，算力需求可控、模型研发难度适中，二者均不会成为行业发展的核心瓶颈；数据作为模型训练的核心基础，其存量不足、难以整合、核心供给短缺等问题，使其成为当前科学智能领域最大的核心挑战，其难度与资源消耗，远远高于算力与模型层面，也成为制约科学智能技术落地、产业升级的关键所在。

第三章 构建服务 AI4S 的精准、端到端的高质量数据集

AI4S 作为聚焦生物医药、新材料等前沿科研领域的核心技术方向，其发展核心的是“数据质量优先”，而非“数据数量优先”。AI4S 数据构建的核心逻辑，是摆脱通用大模型“海量堆砌”的思维定式，聚焦精准化、端到端的高质量数据，这一导向源于 AI4S 的技术特性与产业应用需求，也是其实现科研赋能的基本前提。唯有通过科学的主动构建方法，搭建适配科研场景的标准化数据体系，才能让 AI 模型精准提取科研领域的核心规律，真正实现对生物医药、新材料等领域科研工作的赋能，推动科研模式的智能化升级，助力重大科研成果的突破。

3.1 AI4S 的数据核心要求：精准无错、“端到端”串联

首先，支撑 AI4S 的数据集需要精准无错。不同于通用大模型依托海量数据覆盖广泛场景的训练逻辑，AI4S 的核心使命是解决具体科研领域的核心难题，如新药分子设计、新材料性能预测、生物机制解析等，这类场景对数据的核心需求是“精准匹配科研逻辑、完整覆盖研究链条”，而非数据的数量堆砌。通用大模型依赖海量互联网数据实现多场景适配，即便部分数据存在冗余、低质问题，也可通过规模优势稀释误差；但 AI4S 若盲目追求数据量，引入与科研需求无关的冗余数据、低质数据，不仅会增加模型训练的算力成本与时间成本，更会干扰模型对核心科研规律的提取，导致模型输出偏离科研实际需求，且模型一旦学习错误数据后续难以修正，进而无法实现对科研工作的有效赋能。因此，AI4S 对数据错误的容忍度极低，首要要求是“精准性”，即数据需紧密贴合具体科研场景，聚焦核心研究对象，能够精准反映科研过程中的关键规律、实验结果与逻辑关联，剔除无效冗余信息，

实现数据与科研需求的高度匹配。

其次，数据要做到“端到端”串联。所谓“端到端”是指数据需覆盖科研全流程，从研究设计、实验开展、数据采集，到结果分析、成果转化，形成完整的数据链条，确保数据之间的逻辑连贯性与关联性，避免出现数据碎片化、孤岛化的问题。“端到端”的高质量数据是 AI4S 数据构建的核心，其核心价值，在于打破科研各环节的数据壁垒，实现全流程数据的贯通与复用，为科学智能模型提供完整、连贯的规律学习基础，这也是区别于传统科研数据积累的关键特征。

例如，在生物医药领域，“端到端”数据的构建需紧密围绕药物研发的全生命周期，全面涵盖分子筛选、细胞实验、动物实验、临床研究等核心环节，最终实现“分子结构-活性测试-临床疗效”的全链条贯通，这是 AI 赋能药物研发的核心前提。分子筛选作为药物研发的起点，其核心数据包括候选分子的化学结构、理化性质、靶点结合能力等，这些数据是模型预测分子活性、筛选潜在药物分子的基础，直接决定了后续研发方向的准确性；细胞实验阶段，需重点采集分子对细胞的毒性、抑制率、作用机制等数据，建立分子活性与细胞层面反应的关联，为后续动物实验提供数据支撑；动物实验环节，需记录实验动物的种类、给药剂量、给药方式，以及动物体内的药物代谢、毒性反应、疗效表现等数据，验证分子在活体层面的有效性与安全性；临床研究阶段，则需采集受试者的基本信息、给药方案、疗效指标、不良反应等数据，最终实现从分子结构到临床疗效的全环节数据闭环。这种全链条贯通的数据，能够让 AI 模型清晰学习到“分子结构如何影响活性、活性如何转化为临床疗效”的核心规律，从而精准辅助新

药分子设计、疗效预测、不良反应预警等关键科研工作，大幅缩短研发周期、降低研发风险。

再如，新材料领域的“端到端”数据体系，需突破传统研发边界，聚焦材料设计、研发、生产与应用的全流程，全面覆盖材料结构、原料配方、制备工艺、性能测试、应用场景等核心环节，构建“微观结构-配方-工艺-性能-应用”的完整数据体系，这是 AI 深度赋能新材料研发与优化的核心支撑，也是破解新材料研发周期长、成本高、试错难度大的关键路径。其中，材料结构环节聚焦微观层面，核心包含晶体结构、相组成等关键信息，需通过精密检测手段采集，是 AI 理解“结构-性能”关联的基础，如高性能合金的晶体结构直接影响其强度与耐腐蚀性；原料配方作为研发前提，不仅涵盖原料种类、配比等信息，还需关联原料特性与微观结构的对应关系，如新型复合材料中碳纤维与树脂的特性搭配决定微观结合状态；制备工艺环节需同步记录关键工艺参数及其对微观结构的影响，如陶瓷材料的烧结参数影响晶粒尺寸与致密度；性能测试阶段需全面采集宏微观性能数据，关联微观结构与性能的对对应关系，记录结构缺陷对性能的影响；应用场景环节则聚焦材料实际使用表现，记录不同工况下的性能衰减与微观结构变化，如新能源汽车电极材料的充放电表现，反向指导各环节优化，这些数据共同构成“微观结构-配方-工艺-性能-应用”的完整体系，为 AI 辅助新材料研发提供可靠支撑。

3.2“端到端”数据首先是对微观机理的精准描述

“端到端”数据的核心前提，是对微观机理的精准描述，这是构建完整端到端数据体系的基础与起点，也是确保数据能够真实反映物质本质、支撑科学智能模型有效学习的核心前提。从当前

主流科学理论及科学计算实践来看，以牛顿运动规律、微观世界的原子运动规律为核心基础，已能够在很大程度上精准描述自然界的物质本真规律，也足以全面表征物质在自然状态下的诸多核心特征，这一底层逻辑广泛适用于各类科研领域，为微观机理的精准描述提供了理论支撑与实践路径。

在具体科研场景中，微观机理的精准描述需结合不同领域的研究特性，聚焦对应微观尺度开展精准刻画。在制药领域，需通过分子动力学（MD）等专业物理计算方法，精准描述药物分子在蛋白、溶液等复杂环境中的运动轨迹、相互作用模式，捕捉分子层面的核心特性，为后续药物研发提供底层数据支撑；在材料领域，需深入原子乃至电子层面，通过专业理论计算，精准刻画材料组份的微观运动规律、结合状态，厘清微观结构与材料特性的内在关联；而在催化反应、核反应等更为精密的科研领域，甚至需要深入到量子层面，精准描述物质的微观作用规律，确保数据能够完整呈现物质的核心本质，为科研探索提供精准依据。

“端到端”数据集的构建，第一步就是要搭建起这些最微观尺度层面的精准描述与表征体系，筑牢数据体系的底层基础。在此基础上，再逐步整合物质在合同研究组织（CRO）实验、临床研究、实验室验证、中试放大乃至产业化等各阶段的表现数据，形成“微观机理-实验验证-产业应用”的完整端到端数据闭环。这种以微观机理精准描述为起点的构建模式，能够确保模型的提问与训练数据的维度保持高度一致，让模型能够从底层层面学习到物质的核心规律，而非仅停留在表面数据的简单关联，为科学智能模型的高效训练、精准赋能提供坚实的底层数据保障。

3.3 “端到端”数据集的构建思路：转变传统思路，逆向构建高质量科研数据“教材”

AI4S 的数据体系搭建并非被动采集现有数据，而是通过科学的方法主动构建，这也是其实现高质量数据供给的关键路径。传统科研数据积累多依赖科研人员的零散记录、实验报告归档，存在数据标准化程度低、关联性差、有效数据占比低等问题，无法满足 AI 模型训练的需求。而 AI4S 则通过主动设计数据采集方案、建立标准化数据管理流程、搭建专业化数据平台，系统性推进数据体系建设：一方面，结合具体科研场景的需求，明确数据采集的范围、指标与标准，引导科研人员按照规范开展数据采集与记录，从源头保证数据的精准性与高质量；另一方面，通过数据清洗、标注、整合等技术手段，对采集到的原始数据进行优化处理，剔除无效信息、修正数据偏差，构建标准化的数据集；同时，建立数据共享与更新机制，推动不同科研机构、企业之间的合规数据共享，补充完善数据链条，实现数据体系的持续迭代升级。

专栏 3 “端到端”数据集的构建需要系统科学规划

当前 AI4S 的应用覆盖领域极为广泛，已深度渗透至生物医药、新材料、高端装备、先进半导体等多个战略新兴大领域，而每个大领域内部又包含诸多细分行业，不同领域、不同细分行业的科研范式、数据特性、核心需求存在显著差异，这使得建设面向这些行业的端到端数据集，成为一项复杂且庞大的系统工程，必须依托系统科学的规划设计，才能确保数据集的实用性、兼容性与可扩展性。

从领域差异来看，生物医药领域聚焦药物研发全生命周期，核心围绕分子、细胞、临床等层面的数据构建；新材料领域侧重材料设计、制备、应用全流程，聚焦微观结构、配方、工艺等数据；高端装备领域则聚焦装备研发、制造、运维等环节，核心是性能测试、工况运行

等数据，不同领域的核心数据维度、采集标准、关联逻辑截然不同，增加了数据集建设的复杂度。

从细分行业来看，即便是同一大领域内部，不同细分行业的需求也存在较大差异：以生物医药领域为例，涵盖新药研发、临床诊断、精准医疗等细分方向，新药研发需聚焦分子筛选、临床实验等数据，临床诊断则侧重影像数据、检测指标等，精准医疗则需要整合基因数据、临床数据、生活习惯数据等；新材料领域中，高性能合金、新型复合材料、半导体材料等细分行业，其数据采集的重点、标准也各不相同。

这种多领域、多细分行业的差异化需求，使得端到端数据集建设无法采用统一的模式，需针对不同领域、不同细分行业的特性，进行个性化的规划设计。同时，端到端数据集的建设并非单一环节的工作，还需兼顾数据的采集、清洗、标注、整合、存储、共享等全流程，涉及科研、技术、管理等多个层面，需要协调科研机构、企业、高校等多方资源，若缺乏系统科学的规划，极易出现数据标准不统一、数据孤岛、数据质量参差不齐等问题，导致数据集无法满足 AI4S 模型训练的需求，甚至造成资源浪费。

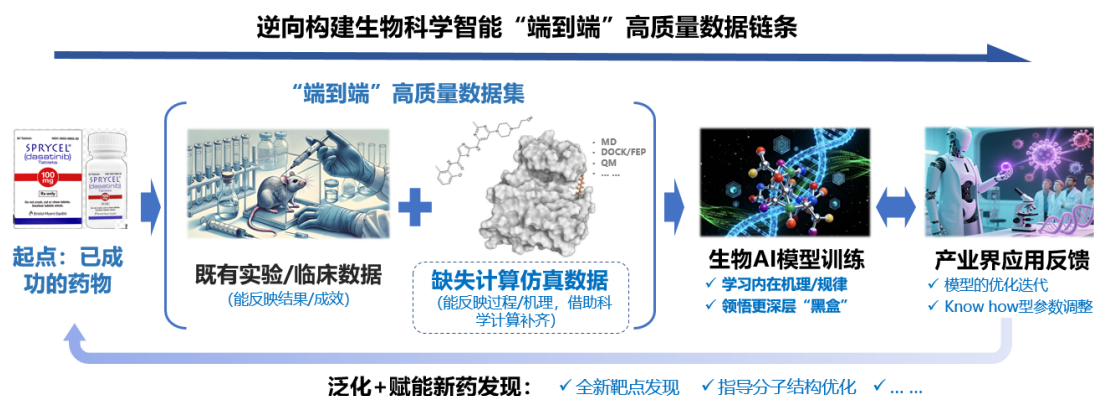
综上所述，面向 AI4S 的端到端数据集建设，其复杂性体现在领域的多样性、细分行业的差异性以及建设流程的系统性上。唯有进行系统科学的规划设计，明确各领域、各细分行业的核心数据需求，制定统一的行业数据标准，搭建规范化的建设流程，协调多方资源协同推进，才能构建出适配不同行业需求、高质量、可复用的端到端数据集，为 AI4S 技术在各领域的深度落地提供坚实的数据支撑，进而推动 AI4S 产业的规模化、高质量发展。

“端到端”数据集的构建思路：打破传统正向研发思路的局限，采用“逆向”构建模式，可大幅加快数据构建速度、提升数据质量，适配 AI4S 模型训练的迫切需求。

长期以来，生物医药领域企业构建端到端数据时，普遍采用正向研发思路，即药厂先通过仿真模拟计算潜在成药物质，一旦找到理论上可行的物质，便停止后续计算工作，转而推进实验验

证。这种思路存在两大核心问题：一是计算的物质共性过强，缺乏代表性与多样性，仅聚焦于理论上的“可行物质”，忽略了不同结构、不同特性物质的差异，构建出的数据无法满足 AI 模型训练对数据多样性的需求，导致模型学习的规律片面，难以适配复杂的科研场景；二是生物医药领域研发与数据构建目标相对分散、构建速度缓慢，若继续沿用这种正向研发的传统思路，仅完成 100 条、1000 条端到端数据的构建，就需要耗费大量的时间、资金与人力成本，不仅效率低下，更极易在全球 AI4S 研发竞争中落后于美国等发达国家的进度，错失发展机遇。

相较于传统正向数据构建路径，**基于既有上市药物开展端到端高质量数据集建设**，具备更强的科学合理性与实施效率。从数据基底属性来看，上市药物是医药行业长期研发迭代形成的优质存量成果，整体具备低毒性、结构稳定性强、靶点亲和性优等固有特征。聚焦既有上市药物为基底，系统性归集整理其相应 CRO 实验数据、临床研究数据等全生命周期资料，结合分子动力学等物理计算手段，补全研发链路中缺失的微观机理维度数据，可快速形成规模化、标准化的端到端高质量数据资源，有效压缩数据集建设周期，为 AI4S 模型训练提供高价值样本支撑。



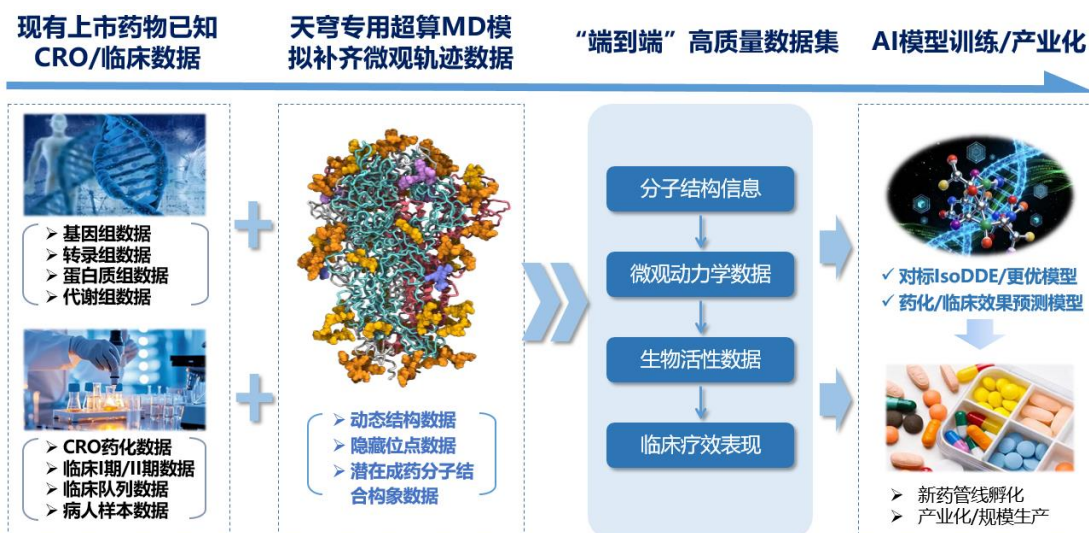


图 2 逆向构建科学智能“端到端”高质量数据集流程示意图

与此同时，同步纳入临床 I / II 期失败药物数据资源，对该类研发失败样本，同样采用物理计算方式补齐微观机理参数，形成完整闭环的端到端数据链路。过去医药研发历程中留存的临床失败样本数据，同样具备不可替代的科研与模型训练价值。医药领域垂类 AI 模型的体系化训练，不应仅依赖上市成功药物数据，还需系统性纳入各阶段临床失败药物的全量信息。临床失败数据能够为模型提供反向学习样本，助力 AI 归纳识别易引发生理毒性、药效不达标、靶点结合失效的分子结构特征与微观作用机理，明确药物研发中的负面边界与风险规律。通过正向成功样本与反向失败样本的双向训练，可使医药领域垂类模型建立完整的药效与毒性关联逻辑，从机理层面规避无效结构设计与高毒性分子筛选，减少后续研发环节的盲目试错，有效降低创新药物研发的时间成本与试验投入。

专栏 4 “两步走”构建端到端数据集

端到端数据集的高效、高质量构建，需遵循“两步走”的系统性策略，其核心逻辑的是“先验证、后批量，先精准、后全面”，既规避了盲目批量构建导致的数据质量参差不齐、逻辑断裂等问题，也解决了单条数据代表性不足的短板。通过第一步的闭环验证，确保数据构建思路科学可行、数据质量达标；通过第二步的批量构建，实现数据的多样性与代表性覆盖，二者相辅相成，共同构建出高质量、可复用的端到端数据集，为 AI4S 技术在生物医药、新材料等各领域的深度落地提供坚实的数据支撑。

第一步：打通单条端到端数据闭环，验证数据构建思路的可行性与数据的正确性，为后续批量构建奠定坚实基础。具体而言，需以单一物质为研究对象，先通过分子动力学、量子计算等专业物理计算方法，精准构建该物质的微观机理数据，完整刻画其原子、分子层面的运动规律、结合状态等核心信息，筑牢数据链条的底层基础；在此基础上，通过湿实验等实际科研手段，逐步验证该物质在研发、实验、应用等各阶段的表现，采集其在细胞实验、动物实验、中试放大、产业应用等环节的核心数据，最终形成从微观机理到实验验证、再到产业应用的完整端到端数据链。尤为关键的是，无论是实验成功还是失败的数据，均具有不可替代的价值——成功数据能够让 AI 模型学习到物质的最优规律、最优应用路径，为后续创新研发提供正向参考；失败数据则能让模型深入分析微观机理层面的失败原因，厘清微观特性与实验结果、应用效果之间的负面关联，避免模型重复学习错误规律，进一步提升模型的预测准确性与容错能力。通过这一步骤，可充分验证“微观机理-实验验证-产业结果”的数据构建思路的可行性，确保后续批量构建的数据符合 AI4S 模型训练的核心需求。

第二步：在打通单条端到端数据闭环、验证可行性的基础上，推进批量构建具有代表性的端到端数据，解决单条数据无法支撑模型训练的痛点。单条端到端数据仅能反映单一物质、单一场景下的规律，难以覆盖 AI4S 模型训练所需的多维度、多场景需求，若仅基于单条或少数几条数据训练模型，极易导致模型学习的规律片面，难以适配复杂的科研与产业场景。因此，批量构建阶段的核心的是挖掘共性、覆盖多样：一方面，由计算数学家、物理学家等专业人才牵头，深入分析单条数据背后的核心规律，挖掘不同物质、不同场景下的数据共

性，明确数据构建的核心标准与逻辑；另一方面，结合 AI4S 模型训练的实际需求，有针对性地构建覆盖不同场景、不同物质特征、不同环境条件的代表性数据，确保数据涵盖不同物质结构、不同工艺参数、不同应用工况，既包含成功案例数据，也涵盖失败案例数据，既覆盖常规场景，也兼顾特殊场景。通过这种批量构建模式，能够让 AI 模型学习到多维度、全方位的核心规律，摆脱单一共性的局限，提升模型的泛化能力与适配能力，真正实现对科研与产业的精准赋能。

3.4 数据构建的核心属性：高成本、高创新性，需国家层面牵头推进

端到端数据集的构建成本极高，这是制约其规模化建设的核心瓶颈之一，单家企业根本无法承担如此高额的研发投入。从实际研发成本来看，采用正向研发思路构建端到端数据，成本消耗极为惊人，仅构建 10 个 PCC（临床前候选化合物）和 IND（新药临床试验申请）相关的端到端数据，所需研发成本就高达数亿元。这一成本规模远超多数企业的承受能力，即便是投入几十亿元的大型研发项目，也仅能构建少量的端到端数据，无法形成规模化、多样化的数据集，更难以满足 AI4S 模型训练的需求。对于单家企业而言，无论是资金实力、研发资源还是人才储备，都无法支撑这种长期、高额的投入。因此，若仅依靠企业自身推进，端到端数据体系的构建将陷入停滞，进而制约整个 AI4S 产业的发展。

更为关键的是，端到端数据集构建并非简单的技术操作，而是一项具有高度创新性的科学工作，必须依靠国家层面的**统筹布局与多方协同**。端到端数据集无法通过被动整合产业现有数据实现，现有产业数据多存在碎片化、标准化程度低、逻辑不连贯等问题，无法满足 AI4S 模型训练对数据完整性、关联性的核心需

求。其构建过程需要由顶尖科学家牵头，结合 AI4S 各领域的科研需求，提出科学、系统的构建思路，通过主动设计实验方案、持续调整优化研发思路、投入大量科研资源开展研发，才能打造出符合要求的端到端数据。这种创新性、系统性的科学工作，绝非单家企业能够独立完成，必须依托国家层面的力量，获得国家对构建思路的认可与支持。通过政策号召、资源集中，整合产业界、科学界、企业、医院、国有实验室等多方力量，形成协同推进的工作格局。通过制定针对性的扶持政策，引导各方资源向端到端数据构建领域倾斜；通过搭建协同平台，推动科研机构、企业、医院等主体的深度合作，实现数据、人才、技术、资金等资源的高效整合；通过集中优势力量，攻克数据构建过程中的核心技术难题，降低构建成本、提升构建效率。

此外，从国家 AI4S 发展的战略层面来看，相较于将有限的资源投入到相对低价值、重复性的研发工作中，集中资源构建端到端的科学智能数据体系，对推动国家 AI4S 技术突破、提升科研创新能力、抢占全球科技竞争制高点具有更为重要的价值，能够为我国生物医药、新材料等战略新兴产业的高质量发展提供坚实支撑，助力实现科学智能领域的自主可控。

第四章 科学计算与 AI4S 落地应用关键技术分析

当前，高质量的科研数据生成环节正遭遇严峻的算力瓶颈，亟需更加强大的专用算力来破局。科学计算是 AI4S 体系中高质量数据生成的核心手段，生物医药、新材料等领域的仿真模拟任务复杂度高、计算量庞大，传统通用超算架构适配性不足，运算效率、通信能力与能耗表现均难以满足业务需求。为此，专用超算应运而生，以国产“天穹”专用超算为例，其全栈核心技术自主研发，成为打通数据生成链路、保障 AI4S 技术落地与产业发展的关键支撑。

4.1 科学计算是获得高质量数据集的关键手段

当前，仿真模拟计算日渐成为生成高质量科学数据集的关键手段。在生命科学、材料科学等 AI4S 核心应用领域，传统实验手段受限于技术条件与观测尺度，难以直接捕捉微观层面的物质运动规律与作用机制，成为制约科研深入推进与端到端数据构建的重要瓶颈；而仿真模拟计算凭借其独特的技术优势，能够有效突破这一局限，精准表征传统实验不可见的微观维度，揭示微观世界的底层本真规律，为 AI4S 提供了坚实支撑。

传统实验手段的尺度受限，本质上是由实验仪器的观测能力与操作边界决定的——无论是生命科学领域的分子相互作用、细胞内信号传导，还是材料科学领域的原子排列、晶体缺陷、界面反应，其微观尺度均处于纳米、埃米级别，远超传统实验仪器的直接观测范围。例如，在生命科学研究中，传统实验可通过显微镜观察细胞的宏观形态，却无法直接观察到药物分子与靶点蛋白的微观结合过程、分子在细胞内的运动轨迹，也难以精准捕捉基因表达调控的微观机制；在材料科学领域，传统实验可检测材料

的宏观力学性能、物理特性，却无法直接观测到原子层面的运动规律、材料制备过程中微观结构的演变，更难以厘清微观缺陷与宏观性能之间的内在关联。这种观测尺度的局限性，导致传统实验只能获取微观机理的间接线索，无法形成完整、精准的微观描述，不仅制约了科研人员对物质本质的认知，也使得端到端数据体系所需的微观机理数据难以通过传统实验手段高效获取。

与之相对，仿真模拟计算依托现代计算科学、量子力学、分子动力学等理论基础，结合高性能算力平台的支撑，**能够突破传统实验的尺度限制，精准表征实验不可见的微观维度，实现对微观世界底层规律的清晰揭示。**仿真模拟计算可通过构建精准的数学模型，将微观层面的物质运动、相互作用转化为可计算、可分析的数字信号，无需依赖直接观测，即可还原微观过程的完整轨迹与核心规律。在生命科学领域，通过分子动力学仿真计算，可精准模拟药物分子在蛋白结合口袋中的运动模式、相互作用强度，清晰呈现分子层面的作用机理，弥补传统实验无法直接观测的短板。在材料科学领域，借助第一性原理仿真计算，可精准表征原子、电子层面的结合状态与运动规律，清晰呈现材料微观结构的演变过程，厘清原料配方、制备工艺与微观结构之间的关联；通过多尺度仿真模拟，可实现从原子级到宏观尺度的跨尺度表征，完整还原材料性能形成的微观机制。这种能够突破观测尺度、精准表征微观维度的优势，使得仿真模拟计算成为端到端数据体系中微观机理数据构建的核心手段，不仅能够高效获取传统实验无法捕捉的微观数据，降低数据构建的成本与难度，更能为 AI4S 模型提供精准、完整的底层规律数据，助力模型更精准地学习物质核心特性，推动生命科学、材料科学等领域的科研创新与产业

赋能。

4.2 面临计算瓶颈：仿真模拟任务带来庞大的计算量

4.2.1 生物仿真计算领域的算力瓶颈

21 世纪是生物学的世纪，也是生命科学的时代，尽管我国生物医药产业近年来取得长足进步，但原始创新能力仍不足。深究造成我国原研药稀少的本质原因，在于生命科学基础研究薄弱，特别是对于新靶点的研究。靶是新药研发的基础，靶点作用机理和结构解析是创新药物开发的核心环节，我国创新药真要是有所突破，就要从靶点的发现做起。在传统医药开发模式下，新靶点的发现到成药往往需要投入巨大的时间和精力投入，往往动辄数十年、花费十亿美金。基于此，通过仿真模拟计算开展蛋白质分子动力学仿真模拟恰恰是对新靶点、新位点/机制的基础性工作，能够促进蛋白质-小分子新结合位点发现和蛋白质-小分子结合区域活动性表征，从而有望从根本上改变我国生物医药基础研究薄弱、只能跟随的困境，并有望带动巨大生物医药产业价值，推动我国向医药强国转变。

需特别提出的是，短时（纳秒级）的蛋白质 MD 仿真模拟难以看到真正有意义的生命现象，更不能指导新药设计现实需求；针对重要蛋白质生理现象的观察，都要在微秒乃至毫秒级的仿真模拟以上才有意义。尽管长时程的动态仿真模拟的理论方法早已趋成熟，能够高度还原实验和逼近真实环境，但这种模拟计算所需的计算量是庞大的天文数字，使得国内绝大多数算力设施均难以支撑。例如，要完成百万原子体系百微秒仿真计算量需要 10^{21} 次算术运算， 10^{21} 次非线性运算， 10^{23} bit 邻域内及全局数据通信，传统超级计算机需要数百年以上的超长时间方能完成计算，这显

然不具备现实意义。过去以来，全球范围内也仅有美国 ANTON 生物专用超算能够高效完成此类任务。遗憾的是 ANTON 对全球其他国家科学家严格禁用，我国的科学家和产业界根本用不到这类强大的工具。

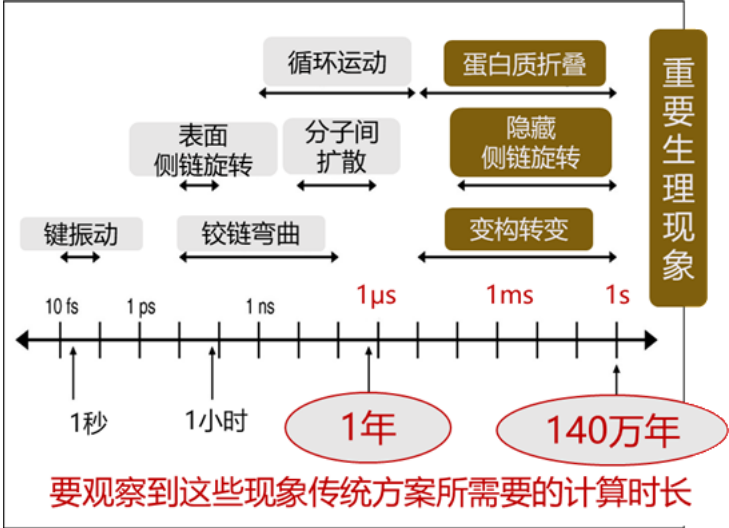


图 3 对重要生理现象的观察所需传统超算计算时长情况

专栏 5 美国安腾（ANTON）系列生物专用超算及应用

在生物领域，对于大体系和长时间尺度的分子动力学（MD）仿真模拟具有突出的科学意义。真实的生命功能往往由多个生物大分子构成的复合物（如核糖体、转运体、偶联受体等）来承载，这些体系的构建通常需要达到上百万个原子的量级。为了捕捉这些复合物关键的构象变化或重要的生理反应过程，模拟的时间尺度往往需要达到微秒乃至毫秒级别。因此，实现百万原子级大体系的长时间模拟是 MD 领域的一项核心挑战与追求。而针对这类大体系、长时长的仿真模拟计算对于算力需求具有极高的技术门槛，至少须达到微秒级以上（1 微秒/天）的仿真模拟计算性能方能满足产业及研发应用的实际需求。

在这一领域，美国的安腾（ANTON）系列生物专用超算能够实现比通用超算快数个数量级的模拟速度，ANTON2 计算效率可达 10 微秒/天，是全球范围内唯一实现百万原子级体系模拟超过 1 万微秒的技术平台，将 MD 模拟推入了“毫秒时代”。在全球范围内的新冠

肺炎疫情暴发初期，美国的计算化学家们便借助 ANTON2，对 COVID-19 病毒处于无配体结合、激活状态的主要蛋白酶实现了 100 微秒的分子动力学模拟，而这些关于 COVID-19 病毒蛋白的分子动力学模拟数据能够有效助力科学家们增进对病毒侵染分子机制更深刻的理解，为理解病毒的感染机制、揭示病毒的弱点，以及加快设计靶向药物提供了重要的信息。安腾系统的这一能力，使其成为观测蛋白质折叠、构象变化及药物结合过程的革命性工具，已被全球多家顶尖学术机构与制药企业用于重要靶点的机理研究与先导化合物优化，充分体现强大的计算硬件在攻克生物模拟关键瓶颈上的巨大价值。从国内看，基于“天穹”超级计算机（计算效率 3-5 微秒/天）完成了超过 10000 微秒的同类模拟，而除此之外，国内外其他单位目前尚难以完成超过 1 微秒的该类大体系计算任务。

与此同时，当前 AI4S 正深度渗透并重塑生物医药研发全流程，从靶点发现、药物分子设计到药效与毒性预测，人工智能已成为提升研发效率、降低试错成本的关键支撑。**生物医药领域的模型突破，长期受制于高质量、高维度、高可靠性的行业数据集极度稀缺**——实验数据往往碎片化、覆盖有限，难以支撑高精度模型对病理、药理本质规律的学习。而基于大规模科学计算开展的分子与蛋白质微观仿真模拟，能够在原子与分子尺度上精准捕捉蛋白质折叠、构象变化、配体结合、相互作用等核心过程，产出蕴含最本质病理药理规律的结构化、标准化数据。这类由仿真计算生成的微观数据，兼具完整性、可解释性与物理一致性，是训练新一代高精度生物医药 AI 模型不可或缺的“黄金燃料”，也将为 AI4S 走向精准设计奠定数据根基。

4.2.2 新材料研发中面临的计算卡点

在新材料产业革命中，材料研发效率已成为决定产业胜负的关键。传统材料开发依赖经验试错，单个材料研发周期常达数十

年甚至跨世纪，如钙钛矿材料从 1839 年发现至 2009 年光伏应用耗时 170 年；同时，现有实验表征技术在空间分辨率、动态追踪能力及样品损伤效应方面存在局限性，难以完整解析材料体系中的界面演变、缺陷动力学及多相耦合效应等微观机制，而这些机制是新材料理性设计的核心基础。在新一轮材料研发范式变革的探索中，材料计算——即基于计算机仿真的材料性能预测与设计技术，正成为突破传统实验试错路线的核心驱动力。

近年来超级计算机以及并行计算技术的发展，极大地提升了材料模拟计算的能力，但材料的固有构效关系导致面向材料筛选和设计的大规模计算具有极高的复杂度，且材料的模拟计算往往在时间和空间上是双跨尺度的，不同尺度模拟方法间的参数传递需要高效性，这对通用超算的内存、通信等提出了更高的要求。例如，核反应堆中结构材料的辐照效应模拟是典型跨尺度过程，时间上跨越中子-原子碰撞（ $<s$ ）到数十年的反应堆寿命（40~60 年），超过 21 个数量级；空间上跨越原子核（ $\sim fm$ ）到结构部件（ $>10m$ ），超过 15 个数量级。同时，由于反应堆工况下的材料性能降级非常复杂，材料辐照缺陷产生及演化的各阶段计算量急剧增加，甚至理论上需要 Z 级（每秒 10^{21} 次浮点运算）超算方能支撑，而如今最前沿的通用超算技术仅发展到 E 级（每秒 10^{18} 次浮点运算）超算，尚存在上千倍的性能代际鸿沟。

4.3 传统超算面对复杂科学计算任务的局限性

近年来，超级计算机以及并行计算技术的发展，极大地提升了仿真模拟计算的能力。我国已规划建设 14 个国家级超算中心，形成覆盖新材料、生物医药、气候模拟等多领域的算力网络，通过“计算-数据-智能”融合模式深度赋能产业创新。但面对这些生

物领域、材料领域复杂的大规模仿真计算任务的挑战，通用超算在追求峰值算力时，其架构固有的瓶颈日益凸显。主要包括：

1) 算法执行效率低下，通用 CPU/GPU 为兼容性而设计，在执行高度特化算法（如 MD 中的非键相互作用力计算、AI 中的大规模矩阵运算）时，实际效率远低于峰值性能。例如，GPU 在处理快速傅立叶变换（FFT）、稀疏矩阵计算问题时，效率仅能达到理论峰值的 35% 以下。

2) “内存墙”与“功耗墙”限制，处理器性能增长远超内存带宽增速，导致“内存墙”问题；同时，E 级超算功耗高达数十兆瓦，如美国“Frontier”系统约 21 兆瓦，日本“富岳”系统（Fugaku）约 29.8 兆瓦，巨大的能耗和运维成本限制了其应用规模与普适性。

3) 通信开销与并行效率瓶颈，典型的生物领域 MD 模拟、材料多尺度模拟等为通信密集型任务，在大规模并行计算中，节点间通信延迟和带宽成为制约整体性能的关键因素，致使通用超算实际性能表现远低于理论峰值。

为突破通用超算面向复杂计算任务的瓶颈，专用超算应运而生，作为针对特定领域或任务进行深度定制化设计的超算，其核心思路是通过“应用驱动、软硬协同”设计，为特定计算任务“量体裁衣”，这包括采用定制化处理器、优化存储层次与互连拓扑、开发专用算法库和编译器等，以此实现目标应用性能与能效比的数量级提升。

4.4 专用超算应运而生

4.4.1 各国发展现状

国际上，发达国家凭借其深厚的技术积累和前瞻性战略布局，在专用计算平台研发方面取得了显著成果，尤其在生物医药和 AI

领域。

美国：D. E. Shaw Research 研制的安腾 ANTON 系列是全球生物 MD 计算效率最快的专用超算，最新一代 ANTON3 在模拟速度上比通用超算快 3-4 个数量级，为揭示蛋白质折叠、药物作用机制等提供了前所未有的强大工具，其成果已催生多种药物进入临床阶段。AI 专用计算平台方面，Google 的 TPU 为加速其 TensorFlow 框架下的神经网络计算而设计，已发展至第七代（Ironwood），在 AI 训练和推理方面表现出卓越的性能功耗比。此外，美国能源部主导的 E 级超算项目（如 Frontier、Aurora、ElCapitan）虽为通用型，但也日益强调其对 AI 以及特定科学领域的深度优化。

日本：日本理化学研究所的分子动力学专用计算机 MDGRAPE 系列采用定制 MD 芯片，MDGRAPE-4A（2019 年）已成功应用于新冠药物研发。曾经的 TOP500 榜首，“富岳”（Fugaku）在设计上兼顾了高性能与高能效，并积极开展在 AI、药物研发、新材料、气候预测等领域的应用。

欧洲：EuroHPCJU 计划明确将 AI 优化型超算和“AI 工厂”作为战略发展重点；特定应用卓越中心（CoE）如 MaX（Materials design at the eXascale），专注于为 E 级计算移植、扩展和优化材料科学代码；意大利能源公司 Eni 的 HPC6 专为脱碳和能效研究设计，也是工业专用超算的代表。

中国：呈现出 AI 计算快速发展，面向特定科学领域的专用整机系统研制相对滞后的特点。**AI 计算发展迅速**，以各地智能计算中心为代表的 AI 专用算力平台大量涌现，主要采用国产 AI 芯片（如华为昇腾、寒武纪思元）和 GPU 集群。**科学计算专用**

超算探索起步，在生物计算领域，思朗科技等企业通过自研 MaPU（代数运算处理器）架构专用芯片及“天穹”超算，在生物 MD 领域带来百倍性能的加速；但国内目前依然缺乏针对新材料领域的大规模计算的专用系统，导致面对复杂材料体系计算模拟时，受制于算力瓶颈和国外软件垄断。



图 4 各国代表专用超算发展概况

4.4.2 应用领域现状

生物医药领域的成功应用：以美国 ANTON 系列在长时 MD 模拟和以 AlphaFold（基于大规模 GPU/TPU 集群的 AI 模型）在蛋白质结构预测方面的革命性突破为标志，专用计算已成为该领域不可或缺的核心研究工具。

材料科学领域的专用化需求与滞后现状：材料科学领域对高性能大规模计算的需求更为迫切且广泛，涉及从电子结构计算（如密度泛函理论 DFT）、原子尺度模拟（如 MD）到介观（相场、动力学蒙特卡洛 KMC）乃至宏观连续介质模拟的多尺度、多物理场耦合问题，以及基于 AI 的高通量材料筛选与逆向设计。然而，专门为材料科学整体优化的、成体系的专用超算平台在全球范围内仍相对缺乏。材料计算问题的复杂性和多样性（算法种类繁多、计算模式各异、数据精度要求不一、跨尺度耦合难度大）

对专用硬件的通用性和软件的灵活性提出了极高要求，这是其专用化进程滞后的主要原因之一。目前，材料大规模计算仍主要依赖通用超算或通用 AI 计算平台。国际上虽有针对特定材料计算环节，如美国劳伦斯伯克利国家实验室开展过针对 DFT 的加速器研究的项目，但尚未出现如 ANTON 之于生物 MD 具有统治力的材料科学专用超算系统和成熟生态。

4.5“天穹”专用超算关键技术分析

4.5.1 核心策略：算法级全局优化计算

“算法级全局优化计算”是一种革命性的计算范式，其核心思想是将算法的整体数据流特征系统地映射到硬件架构与软件执行流程上，从而实现计算资源的最优配置和性能的极致提升。该范式强调从算法全局视角出发，通过深度分析算法的计算特征、数据访问模式和并行性需求，对硬件架构和软件实现进行协同设计和优化，这种优化不仅涉及计算单元的重构，还包括存储层次、互连网络和任务调度等系统级优化，从而实现计算效能的数量级提升。

——在硬件层面，采用包括动态可重构特性、多粒度和离散并行存储结构、层次化互连网络等创新技术。硬件能够灵活高效地针对各类算法实时调整计算模式，动态适配算法的数据访问特征，降低访问延迟并提升有效带宽利用率，为高效计算筑牢硬件根基。

——在软件层面，结合硬件特性构建专用软件栈，对算法特征进行拆解和分析，深度优化数学库和算子库。运用先进代码优化技术，精简运算步骤，改进数据读写，确保软件在复杂场景下稳定、高效运行，实现计算资源的动态最优分配，充分发挥硬件

性能潜力。

专栏 6 科学与工程计算应用涉及的计算类型分类

要突破科学计算中的算力瓶颈，必须基于科学计算最本质的专业领域需求，从算法层面全局性地提炼各类算法在计算、通信和存储等方面的共性特征，并将其映射到科学计算平台的系统性设计中。美国伯克利大学学者 Colella 在 2004 年提出了科学计算领域著名的“七个小矮人”模型，将常见的算法/计算内核抽象归纳为 7 大类，后来加州大学伯克利分校对科学与工程计算应用的分类标准，对应扩展到了 13 种，应用分类如下：

1)稠密线性代数方程组，如 Linpack、大规模流固耦合和流声耦合计算、潜艇收发分置全向声散射特性等；

2)稀疏线性代数方程组，如高超声速飞行器数值模拟、大型客机失速特性模拟等；

3)谱方法，如基于 FFT 的湍流直接数值模拟、耦合气候系统模式（BNU_ESM）地球系统模式等；

4)多体问题，如微孔道扩散过程 MD 模拟等；

5)结构网格，如飞行器数值模拟、可压缩边界层湍流直接数值模拟、地球系统模式、地震模拟等；

6)非结构网格，如高超声速飞行器数值模拟、污染排放模拟等；

7)大规模数据集的并行运算模型（MapReduce），如蒙特卡罗模拟期权定价、基本局部比对搜索工具（BLAST）基因序列比对、托卡马克装置逃逸电子行为模拟等；

8)组合逻辑，如高级加密标准（AES）、信息摘要算法（MD5）等；

9)图的遍历，如社交网络分析等；

10)动态规划，如精确基因序列比对分析等；

11)回溯和分支限界，如可满足性问题（SAT）代数攻击等；

12)图的模型，如深度神经网络、隐马尔可夫模型等；

13)有限状态机，如网络协议分析等应用。

每一种“矮人”即代表了一类可捕获计算和数据移动模式的算法特征，科学计算的各类软件都可以被映射到这七种基本计算模式或其组合。例如，量子力学计算包含稠密线性代数（矩阵对角化）、谱方法

（平面波基 FFT）；分子动力学主要属于粒子方法，同时伴随少量稠密线性代数（求解约束）和 FFT（处理长程库仑相互作用）等。因此，这些最底层的算法模型为分析各类科学计算问题的计算模式组合提供了基础，是指导软硬件协同设计的重要思想工具。

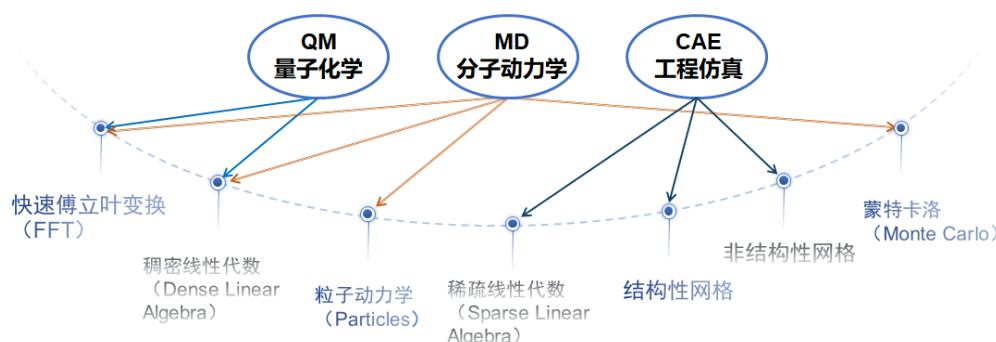


图 5 科学计算的相关软件都可以被映射到基本计算模式或其组合

4.5.2 关键技术 1：专用芯片

面向科学计算的核心算法定制专用芯片架构，支持多种算法特征下的高效调度与协同加速，突破算效瓶颈。

（1）内核架构。构建可重构的紧密交织网络架构，实现运算单元全动态互联与灵活资源调配，结合软流水线机制对指令执行顺序与全局流程进行深度优化，从而支撑不同科学计算算法的零时延动态切换与协同运算，开发科学计算专用指令集，减少无效计算和数据传输，满足科学计算中算法模式的多样性需求和计算效率提升。

面对一个复杂的大算法

以OPLS力场全原子模拟为例 $E_{OPLS} = E_{coul} + E_{vdwl} + E_{bond} + E_{angle} + E_{dihed}$

拆解成一系列
简单算术步骤

$c=a*b$

$d=c+e$

...

...

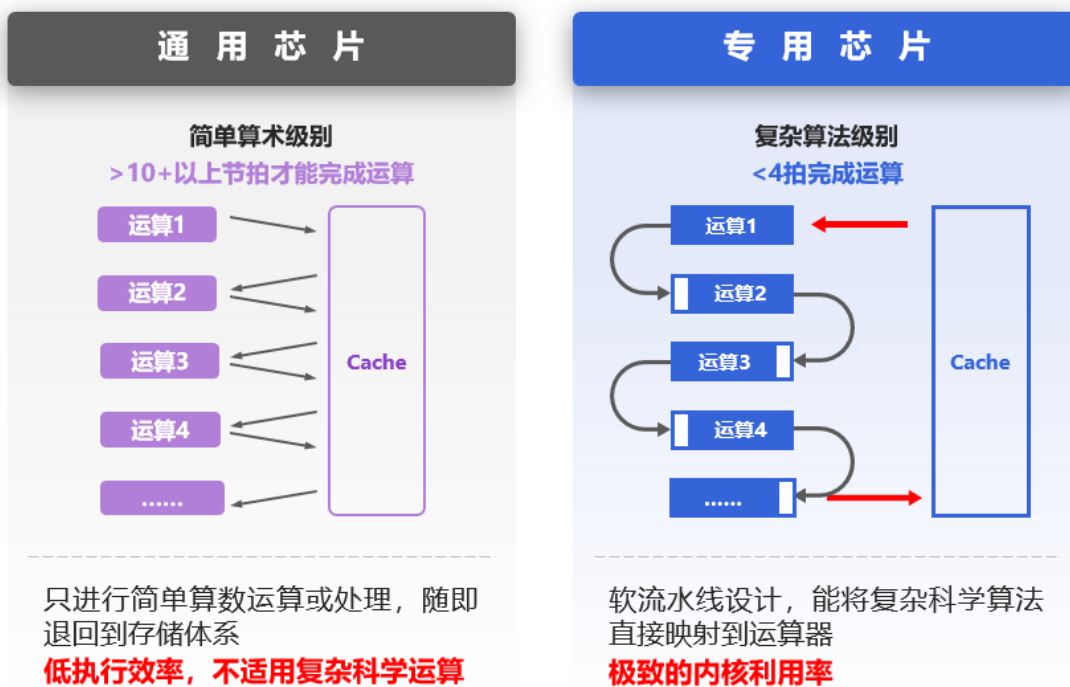


图 6 专用芯片 vs 通用芯片内核指令集层面区别

(2) **运算单元**。重点优化双精度浮点运算单元的性能与能效比，针对科学计算中超大规模稠密/稀疏矩阵运算与累加操作进行硬件级加速；集成高并发处理模块，突破传统通用处理器在仿真规模上的限制。

(3) **存储**。设计新型高速离散并行存储单元(2.5 维存储)，实现将数据按行、列、离散等方式灵活高效并行存储和访问。

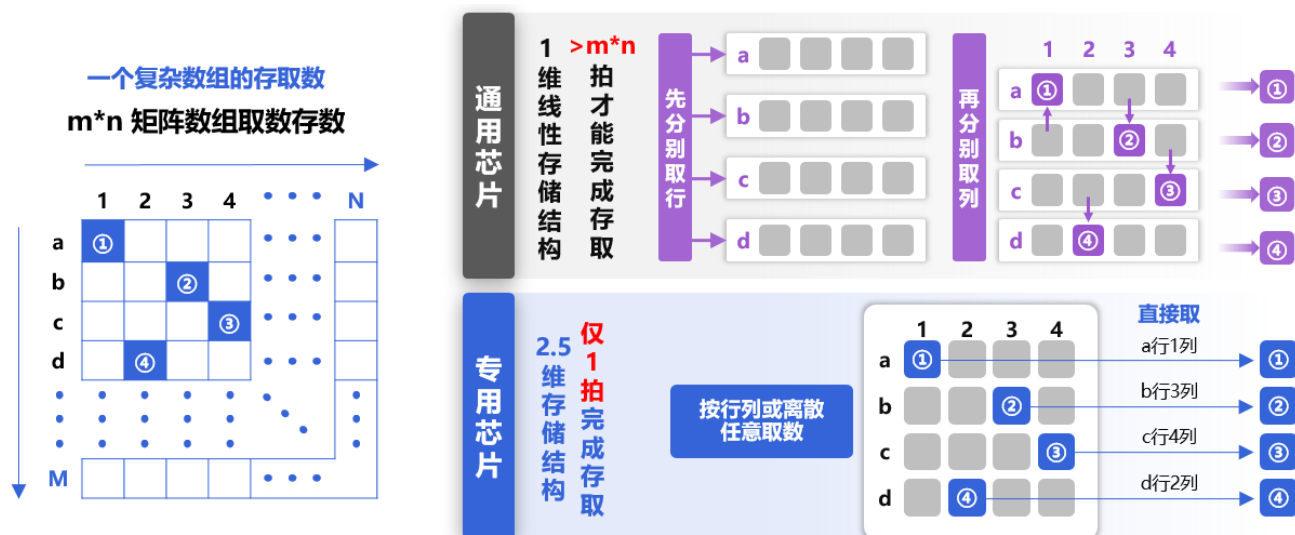


图 7 专用芯片 vs 通用芯片存储层面区别

(4) 通信。研发定制化核间互联总线（三维片上总线），构建新型传输模式，实现低延迟、高吞吐通信传输。

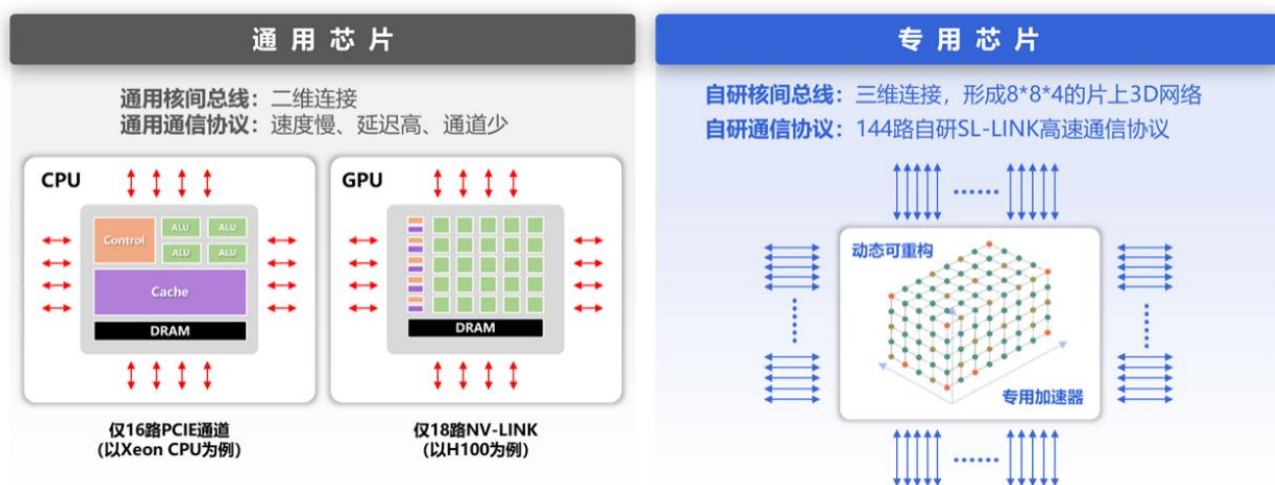


图 8 专用芯片 vs 通用芯片的芯片互连架构层面区别

4.5.3 关键技术 2：专用系统架构

设计适应复杂仿真模拟通信模式的新型互联网络与协议体系，提升跨节点并行效率，构建面向生命科学、材料科学等领域的高性能数学库、通信库和算子库，突破通信与拓扑瓶颈，提升系统整体效能。

(1) 整机系统设计。为了应对科学计算中原子空间模拟的低延时需求，开发专门的自主通信协议和基于 3D 环形网络（3D Torus）拓扑结构的组网架构。这一架构将实现芯片间点对点的高速互联，确保纳秒级延迟通信，提供更低的通信延迟和更高的通信效率，有效满足超大规模体系模拟中的全局同步需求。

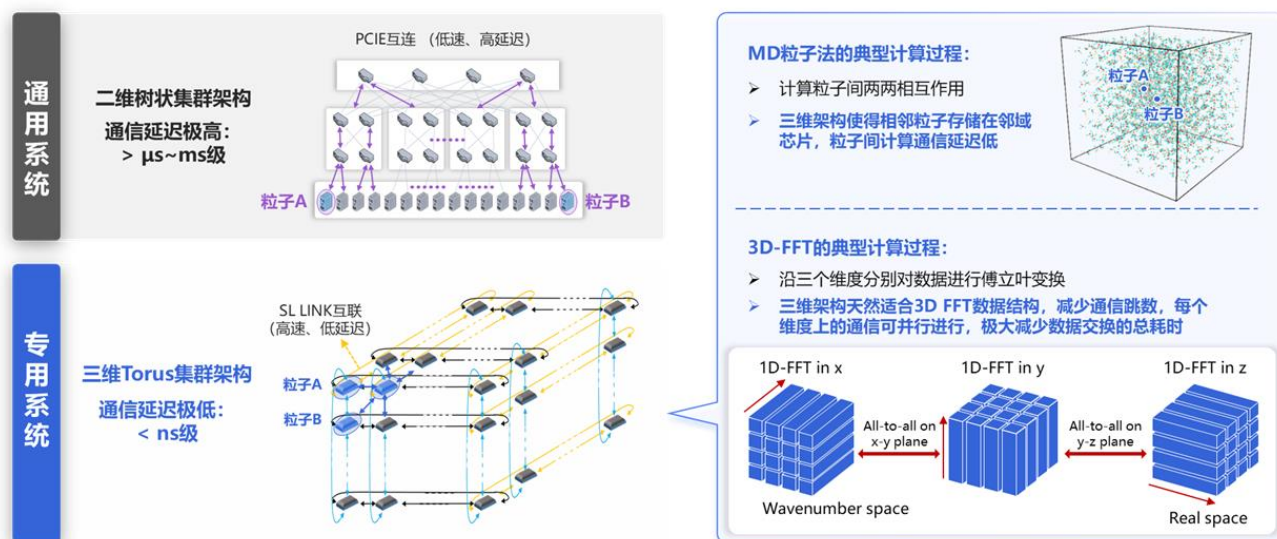


图 9 专用 3D 科学计算系统 vs 通用系统架构区别

(2) 基础库及软件开发。基于专用系统架构特性，设计和搭建开发运行框架和完整软件栈，集成开发环境、编译器及调试工具链，降低跨领域开发门槛。同时需要针对科学计算的共性算法，优化如线性代数、稀疏稠密矩阵、快速傅里叶变换等核心算子性能，优化底层通信库，提升计算密度。

4.5.4 关键技术 3: 专用计算软件

开发广泛的科学计算专用计算软件库，建立覆盖量子、原子、介观与宏观多尺度的科学计算软件体系，深度融合系统架构与算法特性，实现对硬件资源的精细调度与深层优化，形成“平台-算

法-软件”深度耦合的专用计算生态。

软件功能满足科学计算中所涉及到的物理、化学等多个学科领域、多尺度的常用需求，具备高效的计算能力和准确的数值结果，支持大规模复杂科学计算任务的执行；支持多种编程语言和接口，方便用户根据自己的需求选择合适的软件工具；支持并行计算和分布式计算，充分利用计算资源，提升计算效率。



图 10 专用软件 vs 通用软件

第五章 科学计算与 AI4S 市场规模分析

5.1 全球及我国 AI4S 市场概况

全球围绕科学研究流程的智能化工具、平台与服务投入持续增加，带动相关市场规模稳步扩张。据弗若斯特沙利文统计数据，全球科学智能市场预计将从 2020 年的 378 亿美元爆发式增长至 2035 年的近 5 万亿美元，15 年间增长超 130 倍，其增长轨迹呈现明显的“加速-稳定”节奏，中短期（2025-2030 年）复合增长率高达 60.5%，是市场扩张的关键期。从相关领域上看，生命科学与生物、高端制造两大领域将成为绝对双引擎，其中生命科学受益于 AI 驱动的药物与基因研发，而高端制造则与全球工业智能化浪潮深度绑定。整体而言，全球 AI4S 市场正从早期多点分散探索转向以生命科学、高端制造为核心的集中爆发阶段，构成支撑未来全球科技与产业竞争的关键基础设施。

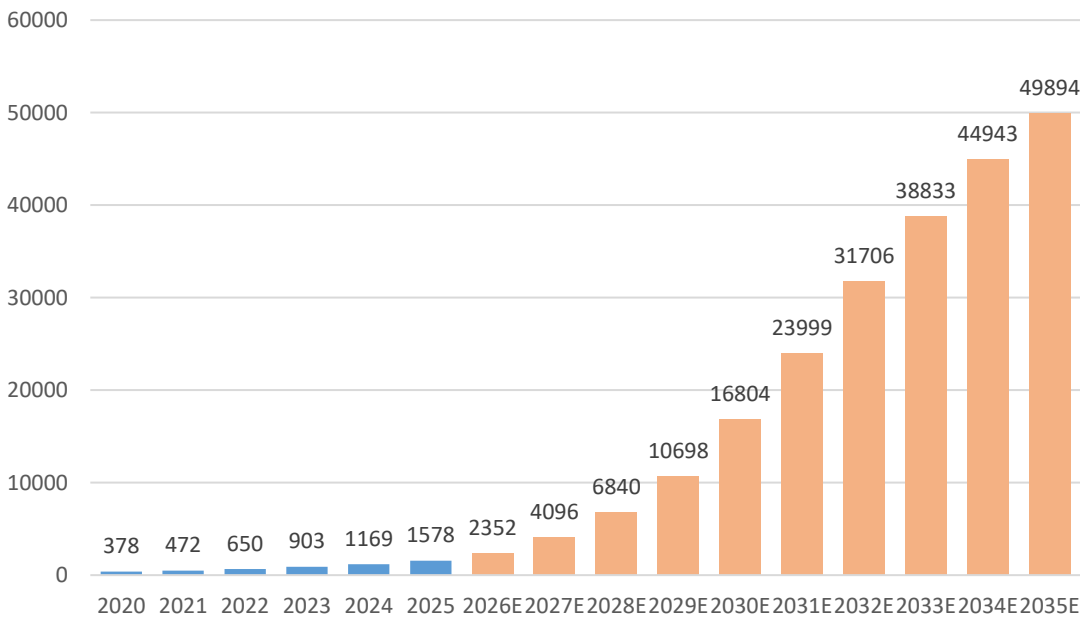


图 11 2020-2035 年全球 AI4S 市场规模及预测（单位：亿美元）

对于我国而言，中国科学智能市场在 2020 年至 2035 年间呈

现爆发式增长，预计规模将从 98 亿美元快速至约 1.8 万亿美元，增长近 160 倍。从细分领域看，生命科学与生物科学依然占据主导地位，与高端制造与新能源领域共同构成市场核心支柱，将催生成为万亿级规模的关键科技创新赛道。

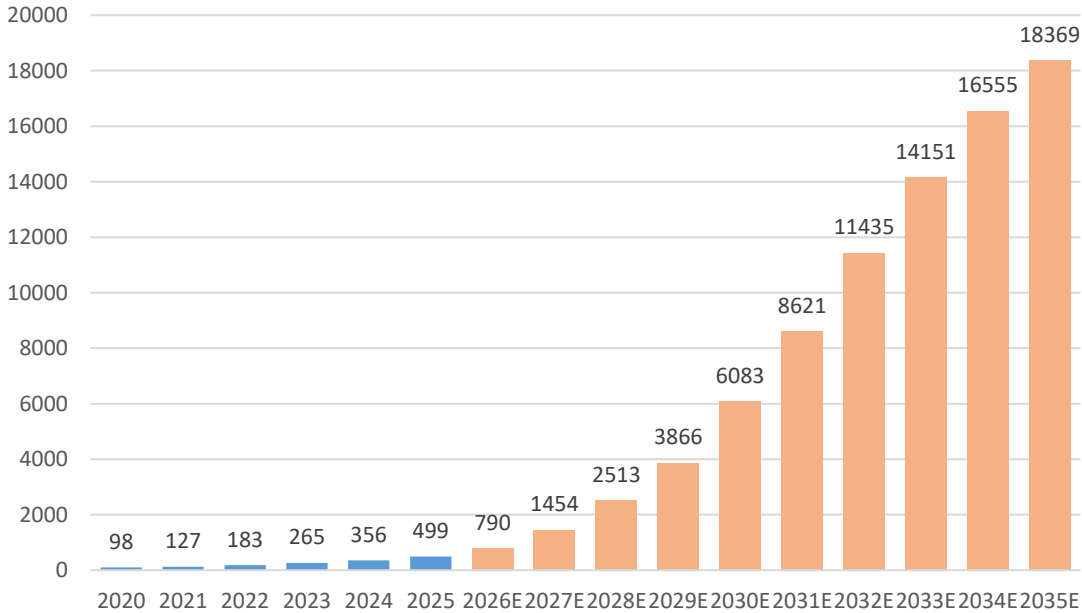


图 12 2020-2035 年我国 AI4S 市场规模及预测（单位：亿美元）

5.2 国内科学计算市场规模分析

（1）我国科学计算市场总体规模

科学计算作为科学智能闭环的关键底座支撑，在国家科技竞争与产业创新需求共同驱动下，正加速从基础设施建设走向规模化应用。按中国信通院测算口径，2025 年中国科学计算市场规模约 1347.0 亿元，预计 2029 年增至约 2682.8 亿元，2025—2029 年复合年均增长率约 18.8%。增长动因主要来自两端：一是供给侧算力性能与能效持续演进，二是企业数字化研发与科研创新需求快速释放，推动科学计算逐步成为支撑科技进步与产业升级的关键底座。

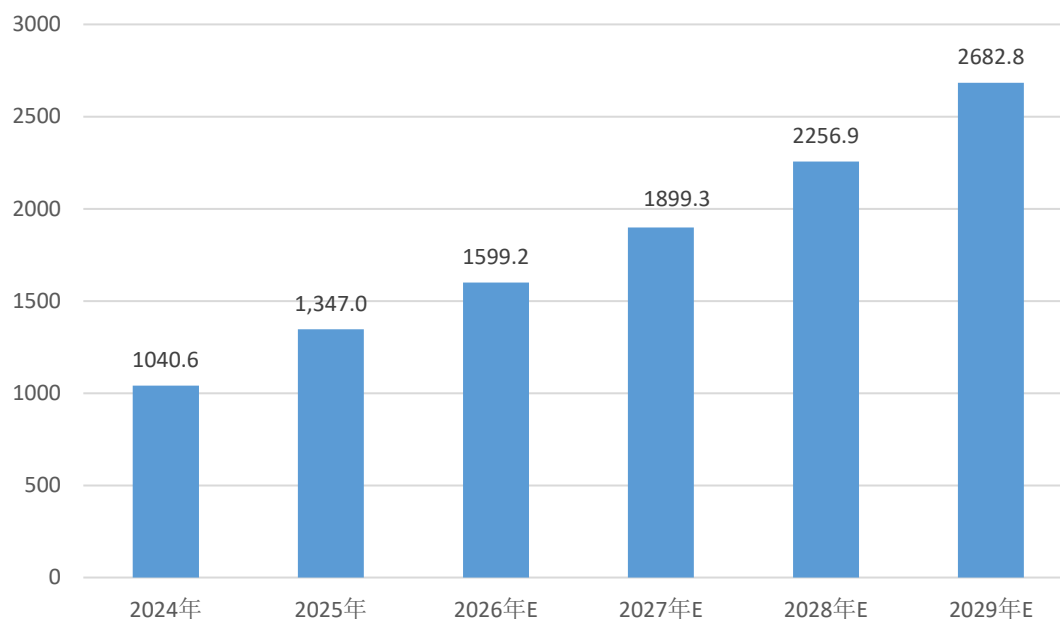


图 13 中国科学计算市场规模预测图（单位：亿元）

一方面，国家级战略持续锚定新型基础设施的投资方向。2025 年 11 月《中共中央关于制定国民经济和社会发展第十五个五年规划的建议》提出“适度超前建设新型基础设施，推进信息通信网络、全国一体化算力网、重大科技基础设施等建设和集约高效利用，推进传统基础设施更新和数智化改造”。在该顶层设计指引下，高性能算力设备与超大型算力中心作为新基建的重要组成部分，因其对前沿科技创新的支撑作用，将进一步吸引财政资金、社会资本与产业资源投入。另一方面，科研需求是市场扩容的核心动力：在高能物理、天文物理等对算力依赖度极高的领域，科学计算直接支撑重大科学发现与基础研究，未来高校与科研院所仍将是政府端科学计算的重要应用场景，并具备明确的扩展空间。

与此同时，随着高性能算力资源从传统互联网、科研机构向高端制造、能源勘探、未来产业等实体经济领域深度渗透，越来越

越多的企业将科学计算用于自身产品研发以提升市场竞争力。企业端市场需求高速扩张，其对算力资源需求从“可选项”向“必选项”转变，对高性能计算硬件采购、工程仿真软件和云上算力服务的投入大幅增加。在此背景下，诸多行业的科学计算投入呈现高增长态势，如汽车与航空航天、芯片设计、能源化工、生物医药等领域因研发密集、流程复杂、合规要求高，对高性能硬件、专业工程仿真软件（如 ANSYS、COMSOL、STAR-CCM+等）的需求持续攀升。这些行业不仅采购规模大，且对算力稳定性、软件适配性和数据安全性提出更高要求，从而带动整个产业链向高附加值环节升级。

（2）我国科学计算市场结构组成

据中国信通院测算数据，2025 年中国科学计算下游需求呈现“多支柱并立、先进制造与硬科技牵引”的结构特征，其中前五大领域为：高端装备 402.3 亿元，约占 29.9%；新材料 256.2 亿元，约占 19.0%；集成电路 215.2 亿元，约占 16.0%；医药制造 197.3 亿元，约占 14.6%；新能源 196.6 亿元，约占 14.6%，前五大领域合计约 94.1%，构成当前需求侧的核心增量来源；此外，合成生物 57.3 亿元约占 4.3%、地球科学 17.1 亿元约占 1.3%，体现出以新兴交叉学科与国家重大科学问题为牵引的稳步扩张趋势；低空经济现阶段占比虽不足 1%，但也已显现前瞻性需求萌芽。

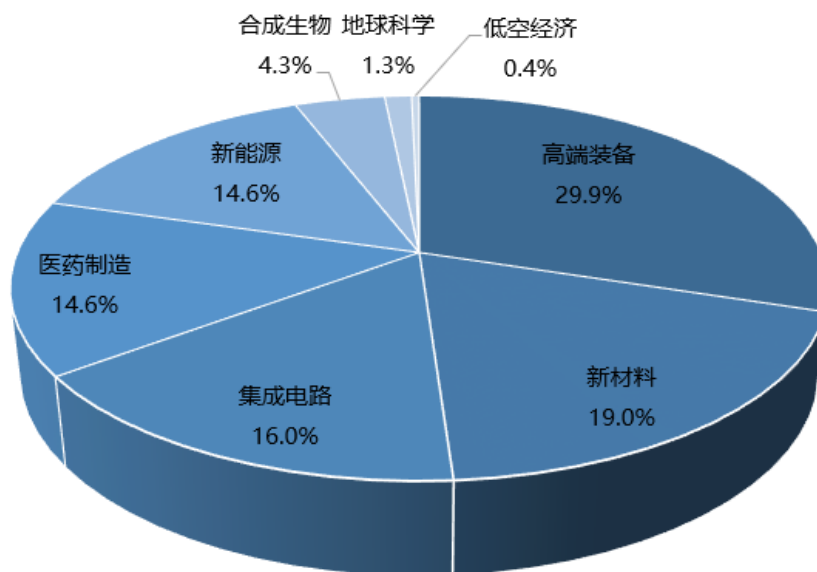


图 14 科学计算下游应用领域分布图（2025 年）

在全球科技竞争与产业升级背景下，科学计算正从“辅助验证”走向“主导设计”，其在研发全链条的渗透使得学科工具市场快速放大。按技术口径拆分，2025 年科学计算相关市场按量子力学计算（QM）、分子动力学（MD）、计算机辅助工程（CAE）等学科工具分类的情况为 QM 约 365.2 亿元、MD 约 306.5 亿元、CAE 约 452.3 亿元。其中，CAE 可视为工程仿真相关工具的统称，通常涵盖计算流体力学（CFD）、有限元分析（FEA/FEM）、多体动力学、热仿真以及电磁—结构等多物理场耦合仿真。其核心作用是在虚拟环境中完成结构设计、性能预测与可靠性验证，是航空航天、高端装备、汽车、能源等领域研发数字化的重要支撑。科学计算作为科学智能的“数据工厂”和“机理引擎”，通过 QM、MD、CAE 等工具在多尺度上开展高保真模拟，持续生成机理可信数据，并与 AI 方法结合推动仿真加速、参数寻优与设计迭代，由此形成“工具需求增长—算力投入提升—场景扩张”的正反馈。

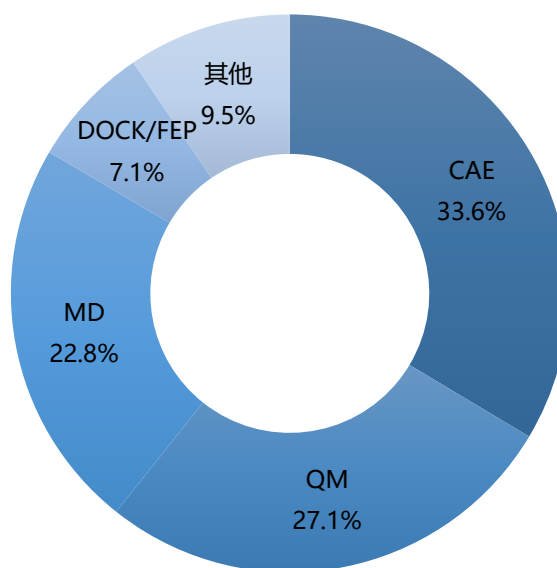


图 15 科学计算任务类型分布图（2025 年）

当前，中国科学计算市场正处于高速增长阶段，市场空间广阔。展望未来，在国家科技创新战略和数字经济政策的推动下，中国科学计算市场有望保持两位数的年复合增长率，成为全球增长最快的区域市场。预计到 2030 年，中国科学计算整体市场规模将达到数千亿元量级，在全球所占份额进一步提高。尤其是企业端需求的爆发式扩张，将成为市场增长的主要引擎：越来越多的制造业和新兴产业企业加大对超级计算和人工智能算力的投入，以支撑复杂模型仿真、海量数据分析和产品创新研发；政府端市场在“东数西算”等政策带动下也将稳步扩张，新一代国家超级计算中心布局和科研算力升级为市场提供长期动力。

总体看来，MD、QM、CAE 等专业仿真模拟工具市场规模持续扩容，产业增长逻辑本质来源于工业界对高精度、高可信度仿真数据的刚性需求。相较于传统物理实验，仿真软件能够在可控条件下完成多尺度、多物理场的精细化计算，生成一致性高、维度丰富、可追溯性强的标准化仿真数据，满足新材料、生物医

药、高端制造等领域的机理分析、性能预测与方案迭代需求。企业愿意为高端算力、商用仿真工具以及定制化计算服务持续投入成本，反映出行业已经形成共识：高质量仿真数据是工业研发流程中不可或缺的核心生产资料，算力与仿真工具正在成为企业数据资产建设的重要投入方向。与此同时，国内企业端科学计算需求的快速攀升，进一步标志着科学计算的应用属性发生结构性转变。长期以来，科学计算多服务于高校、科研院所等基础研究场景，偏向科研辅助工具；而当前工业企业大规模引入仿真计算、高精度数值模拟、多物理场耦合求解等技术，说明科学计算已深度下沉至工业研发流程，逐步从科研研究工具演变为支撑产品设计、工艺优化、风险验证、正向研发的产业刚需底座。

第六章 科学计算与 AI4S 典型应用案例

6.1 公共基础设施领域典型案例

6.1.1 爱赛思 OpenAI4S 社区：开放式科学智能基础设施

2026 年 3 月，在全球开发者先锋大会开幕式上，由上海市经济和信息化委员会指导，上海仪电牵头建设的“爱赛思 OpenAI4S 社区”正式发布。该社区定位为“向 AI 提问，为科学求解”的开放式科学智能基础设施，以“开启科研新范式，打造科研新支点”为使命，构建覆盖模型广场、数据集市、智能体库、工具链、实验资源的“科学智能要素即服务”基础设施，致力于打通“理论-计算-实验”全链路科研数智闭环。爱赛思 OpenAI4S 社区的主要功能如下：

1) 科研 workflow。提供“工具、数据、算力”三位一体的云端任务环境，科研团队可一键获取所需资源，实现“便捷建模、高速计算、打通实验资源”，按需匹配项目需求。已汇聚生命科学、新材料等多领域主流工具包，整合 30 余类高质量科学数据集，并完成 15000 卡算力调配，协同思朗科技、海光等构建多元异构、超智融合算力底座。

2) 智研飞轮引擎。新一代“科学求解”引擎——智研飞轮，可调用多智能体协同执行，自主拆解复杂科研任务，提升科学探索的精准度与效率。智研飞轮已接入“星河启智·科学宇宙”“上海交大 SciMaster”“浦江实验室 Intern-S1 Pro”“仪电科研助手”等多款科学智能体与模型，覆盖从问题定义到方案生成的全流程。

3) 科学智能要素资源池。社区构建统一标准体系，实现平台间工具链与数据集共建共享。截至目前，累计汇聚 262 款模型、125 个智能体、2139 款工具、10 余个干湿实验室，总计超过 2500

项要素资源，覆盖分子设计、仿真模拟、实验验证等科研环节。

4) 爱·空间。从“独思”走向“共创”。从“独思”走向“共创”。爱·空间以任务研讨为起点，支持组队撮合、问题征询、BBS 式交流等功能，帮助科研团队找伙伴、找话题、找资金，打破实验室围墙。用户可通过标准化任务发布模板(涵盖方案设计、数据智能、实验与模拟、协作对接等)清晰表达需求，确保真需求被精准识别、有效传递，践行“产业出题、科技答题”。同时，社区链接项目经理人与基金经理人，护航成果从孵化到转化。

5) 赛·百团百项。重点支撑“百团百项”专项工程，搭建线上申报与项目汇交直通渠道。科研团队可在线完成项目申报和成果提交，优秀项目将进入社区标杆案例库集中展示。聚焦生命科学与生物技术、物质科学与新材料、地球环境与科学、数学与量子、工程技术等前沿方向，以高价值项目孵化推动科学智能与产业创新深度融合，加速技术突破走向生产线。

6) 思·资讯。学习、政策、动态，一站直达。思·资讯模块整合三大核心内容，为科研工作者提供全方位信息支持。

“科研课程”上线 AI4S 专业课程与培训体系，从入门到进阶，助力科研团队快速掌握科学智能方法与工具。

“政策指南”深度解读各级政策文件，让政策看得懂、路径找得准、项目报得顺。

“资讯信息”实时更新行业动态、前沿进展与社区要闻，帮助科学家把握科学智能发展脉搏。

三大板块协同联动，让学习有门路、政策有路径、资讯有来源，真正做到信息赋能科研。

爱赛思 OpenAI4S 社区通过整合科学智能全要素资源，为

AI4S 从数据生成、模型训练到产业转化提供了可复用的基础设施底座，是构建自主可控、开放协同的科学智能生态的关键实践。

6.2 生物医药领域典型案例

6.2.1 谷歌从 AlphaFold 到 IsoDDE 推动 AI 制药落地

在生物医药领域，2018 年美国 AlphaFold 横空出世，其通过 AI 预测蛋白质结构彻底革新了生命科学的研究范式，这一技术解决了困扰科学界数十年的蛋白质折叠难题。AlphaFold 凭一己之力颠覆了结构生物学，让全球 300 多万科学家免费用上最顶尖的蛋白质预测工具，甚至助力团队拿下诺贝尔奖。它摒弃了传统依赖实验试错和经验推导的研究逻辑，通过深度学习从海量生物数据中挖掘规律，为生命科学提供了全新的研究思路。AlphaFold 的成功让 AI4S 的理念深入人心，证明 AI 不仅是辅助工具，更能成为攻克基础科学难题的核心力量。如今，免费开放的 AlphaFold 蛋白质数据库，已被全球 190 多个国家、超 300 万研究人员使用，推动了生命科学研究进入“AI 驱动”的新纪元。美中不足的是后续基准测试表明，对于与训练样本差异较大的结构，即便是最前沿的 AlphaFold3 预测精度也不尽如意，也就是说它难以泛化到生物分子空间中未被探索的区域，而这恰恰是药物发现中最富潜力、也最具挑战的未知领域。

2026 年 2 月，谷歌 DeepMind 的生物制药子公司 IsomorphicLabs 公布了一款更强大的 AI 药物发现引擎——IsoDDE，被科学家称为“AlphaFold4 级别的重大突破”。相较于 AlphaFold3 和 Boltz-2，IsoDDE 大幅提升了预测准确性，还能精准预测小分子结合亲和力，精度超越了传统基于物理的“金标准”，时间与成本却极低；更为重要的是，其仅凭氨基酸序列就能准确

识别靶点蛋白上的全新、可成药的结合口袋，这对于新药研发尤为重要。如今，IsomorphicLabs 与强生旗下的生物技术部门杨森制药达成合作协议，正将 AI 设计引擎与杨森在药物开发方面的专业知识相结合，共同进行跨模式、多靶点研究合作；同时，IsomorphicLabs 也与礼来和诺华签署了潜在价值近 30 亿美元的合作协议，合作运行了 17 条药物研发管线，预计首批由 AI 设计的药物将在 2026 年底进入临床试验。

6.2.2 礼来携手英伟达共建 AI 联合实验室

2025 年 10 月，英伟达和礼来计划合作共建全球首个制药专属“AI 超级工厂”，该系统由超过 1000 块英伟达 Blackwell Ultra GPU 组成，通过管理从数据输入、模型训练到推理的完整 AI 生命周期，加速药物研发的全流程，其潜力在于探索仅靠人类难以发现的新分子，计划将早期药物发现周期的 5-6 年实践缩短 40%，临床研发成本降低 30%。2026 年 1 月，礼来公司与英伟达宣布将投入高达 10 亿美元合作成立一家开创性的 AI 联合创新实验室，致力于运用 AI 技术解决制药行业长期面临的诸多挑战。该实验室将把礼来在药物发现、开发及制造领域位居世界领先的专业能力，与英伟达在 AI、加速计算及 AI 基础设施领域的领导地位相结合。依托行业领先的算力资源、高质量大规模数据生成能力，并依托 NVIDIA BioNeMo™这一加速药物研发的关键平台，双方团队将重点开发训练面向生物学和化学领域的下一代基础模型与前沿模型，旨在以前所未有的速度和精度完成新分子识别、优化与验证，从而加速新药研发。除药物发现外，英伟达与礼来还将探索在临床开发、生产制造及商业运营等领域应用 AI 的机会，整合多模态模型、代理式 AI、机器人技术及数字孪生。

6.2.3 思朗科技打造基于“天穹”的制药新引擎 sDDE

天穹新药引擎（sDDE, Smartlogic Drug Design Engine）是思朗科技原创的 AI+科学计算驱动的“端到端”全链条创新药设计引擎，以独立自主研发的 3D 科学计算机为基座，创新性实现**蛋白位点发现、药物结合构象预测、亲和力评估、分子胶理性设计与优化**全链条流程。在位点发现模型上实现了对传统位点、隐蔽位点、变构位点的高精度识别，在药物结合构象预测模型上支持多场景（小分子、抗体等）结合构象预测需求，在亲和力评估模型上精准量化分子与蛋白位点的结合强度，在分子胶理性设计与优化模型实现精准设计并筛选全新分子胶。分子胶理性设计与优化模型是天穹新药引擎的特色核心模块，支持分子胶全流程设计与优化需求，能够在 1-2 周内针对不同的靶蛋白完成分子胶的快速发现，进而能够进行分子胶理性设计并在短时间内进行验证的原创性开发引擎。

基于自研的天穹新药引擎，目前思朗科技已经形成融合“科学计算+AI”的小分子药物、分子胶药物、抗体药物、中药现代化、病毒机制研究等多个产业方向，并积极参与到生物药物研发的各个环节。例如，在小分子领域，思朗科技针对 JAK2-V617F 突变蛋白以及 TYK2、STAT6、PTP1B 等多个新靶点上，探索开发了全新的变构抑制剂；在分子胶领域，针对新兴靶点 VAV1 计算开发的一款临床候选化合物（PCC）初步显示出良好的生物学活性和安全性。

6.3 新材料/新能源领域典型案例

6.3.1 BASF 引入 Quriosity 超算赋能新产品研发

德国 BASF 作为一家世界级的化工龙头企业，也是最早将计

算仿真模拟应用于科研领域的企业之一。早在 2017 年 BASF 便引入了超算——Quriosity，其拥有 1.75PFLOPS 计算能力，在当年全球计算系统 500 强中该计算机列第 65 位。Quriosity 在 BASF 加快研发数字化的过程中发挥着关键性作用，其平均每天执行 20,000 余个计算任务，包括借助复杂模拟协助研究人员更精确地预测哪些化妆品成分可以进行优化协调，通过分子建模在上万种可能的方案中筛选出某些特定功效的化合物，或是仿真模拟反应器中的物质和温度分布实现反应过程的优化，乃至优化生产中操作单元的流体动力学等。鉴于 Quriosity 计算能力已饱和以及考虑到未来材料研究项目的愈发复杂，BASF 于 2023 年在路德维希港基地另行启动部署一台 3PFLOPS 计算能力的超算，借助更高性能的算力进一步缩短新材料开发所用的时间及成本。Quriosity 的引入大幅强化了 BASF 的研发实力，也为巩固 BASF 的行业地位做出了突出贡献。全新 BASF 超级计算机与它的前身一样被命名为 Quriosity，是世界上最大的用于工业化学研究的超级计算机之一，其具体应用案例如下：

1) 洗涤剂配方优化。餐具洗涤剂的起泡程度，乳霜涂抹在皮肤上的难易程度，这些都与表面活性剂有关。通过使用 Quriosity 进行模拟计算，模拟洗涤剂配方的分子层级，可以观察到 BASF 现有洗涤剂成分和新成分之间是如何工作的，可以进一步帮助 BASF 研究表面活性剂，从而优化开发新配方。

2) 农业农药评估。在农业生产过程中，农药残留可能会对土壤和地下水产生影响。在 Quriosity 的帮助下，可以通过数学方程式精确计算出植保农药产品与土壤之间的相互作用，及其渗入地下水的可能性，在短短几个小时内计算模拟超过 40 万种相关

复杂环境场景，而不是通过几年的时间来完成。

3) 潜力新材料开发。BASF 通过 Quriosity 建立了一个庞大的计算分子属性的数据库，而通过机器学习，这些分子属性可以与 BASF 产品的性能建立联系，以此帮助 BASF 寻找具有潜力的分子，进而赋能产品创新。

4) AI 反应器。化学反应产率最大化是化学家日常工作中一项典型且高度复杂的任务，传统做法通常是依次调整不同的反应参数，直至达到理想产率，这一过程往往耗时较长。巴斯夫 AI 反应器显著加速了这一复杂流程，可对化学实验进行规划、执行和分析。更为重要的是，该反应器能够学习并自主触发下一轮反应循环，使反应产率最大化。目前，巴斯夫的首批实验结果表明，该系统与人工操作相比速度提升了约 20 倍。

6.3.2 字节跳动推出 BAMBOO 助力新型电解质设计

液体电解质作为锂离子电池的“血液”，其性能直接决定电池能量密度、循环寿命与安全性能。当前商用碳酸盐基电解质普遍包含五种以上组分，其分子设计长期依赖实验试错法，不仅研发周期长达 3-5 年，且多组分体系的性能优化复杂度呈指数级增长。据相关统计，传统电解质开发平均需要筛选超过 10 万种配方组合，通过实验来探索和优化这些多组分体系，仅基础实验成本就高达数千万美元，很大程度上依赖研究人员的经验和直觉，这限制了新材料从实验室概念验证到市场产品的转化速度。原子模拟提供了一种替代方案，但面临两难：精确的量子力学模拟计算量巨大，而传统的经典力场方法又精度不足。

在此背景下，字节跳动的核心竞争力——AI 算法与大数据处理能力，恰好适配锂电研发的核心痛点。2025 年 4 月字节跳

动推出了一款名为 BAMBOO(ByteDance AI Molecular Simulation Booster) 的 AI 驱动分子模拟框架, 专门用于锂电池液体电解质的研发。模拟结果表明, BAMBOO 在预测各种溶剂和盐组合中的关键电解质特性, 尤其在预测各种化学物质的密度、粘度和离子电导率方面表现出了较高的准确性, 成为由分子结构工程驱动的电解质设计的宝贵工具。性能测试表明, BAMBOO 在单 GPU 上可实现每秒 10^6 原子步的模拟速度, 较传统 DFT 计算快 10 倍, 同时保持量子化学级别的精度。这种效率提升使得原本需要数月的电解质筛选过程可压缩至数天内完成, 极大加速了从实验室研究到工业应用的转化周期。在应用层面, BAMBOO 已成功指导开发出新型高电压电解质配方, 使 LiCoO_2 /石墨电池在 4.5V 循环下的容量保持率提升至 92% (1000 次循环)。随着模型的迭代优化, 字节跳动的研究团队计划将应用场景扩展至固态电解质界面 (SEI) 形成机制研究, 以及锂金属电池的枝晶生长预测。BAMBOO 正推动新能源材料研发从“经验驱动”向“数据智能驱动”的范式转变。

6.3.3 思朗科技基于“天穹”打造电解液 AI 数据平台

2025 年上海思朗科技股份有限公司联合长江 (孝感) 3D 科学计算中心、上海科学智能研究院, 共同打造基于天穹的电解液 AI 数据平台。基于“天穹”3D 科学计算机以及专为材料计算硬件架构设计的 LogicMD 软件, 对不低于 10000 组电解液体系进行长时间大尺度的分子动力学模拟, 其中单体系模拟时间至少 25ns, 模拟尺寸达到 10 万原子级别, 确保模拟计算数据结果具备统计稳定性与物理可靠性; 针对每组模拟轨迹, 提取并计算其密度、黏度、扩散系数、电导率、介电常数、结合能等关键物理化学性

质；针对上述得到的电解液性质数据，进行智能化标注管理，并构建多模态检索功能、元数据管理功能、数据资产管理功能，形成便捷的对外服务接口。同时，利用上述高质量电解液数据集可训练先进的机器学习模型，建立从电解质的分子结构（或配方组成）到其宏观物理化学性能的精准、快速的定量构效关系模型，实现接近实时预测新分子或新配方的性能，为研发人员的快速迭代和配方优化提供强大且可靠的智能预测工具。

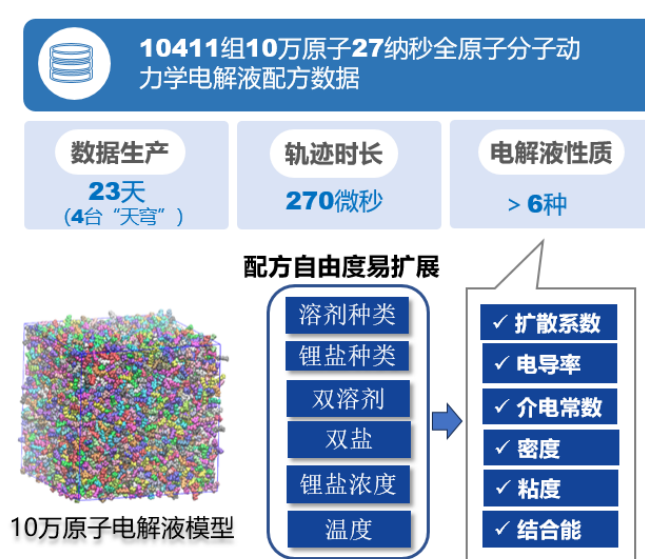


图 16 高精度锂电池电解液配方分子动力学数据库构建示意图

通过天穹 3D 科学计算机对众多电解液体系进行大体系、长时间的分子动力学仿真模拟计算，从而得到高精准性、覆盖足够广度的高质量电解液数据集。一方面，通过进一步揭示了电解液微观结构层面关键性“构效关系”信息，另一方面能够有效支撑电解液 AI 大模型的训练开发，为 AI4S 在电池材料领域的落地实施打通了道路，最终助力实现新型电解液材料的高效开发。

6.4 半导体/航空航天等领域典型案例

6.4.1 美国波音公司以虚拟仿真计算加速新机型研发

在现代航空航天产业中，仿真模拟计算已经成为设计、验证和优化过程中的关键环节：通过将物理世界的复杂现象转化为计算模型，能够查看结构受力、气动载荷、热分布、振动响应等各项指标的变化，从而在投入实际制造前发现问题、修正设计、降低成本和风险。仿真模拟计算并不是替代试验，而是试验前的“预演”，让工程师在虚拟环境中把更多被忽视的细节暴露出来，并为后续的验证提供清晰的重点。

新型飞机研发过程的物理测试成本高昂，如单次风洞实验的费用可能超过 1000 万美元，美国波音公司在波音 767 飞机研发项目中，总计进行了 77 种原型机翼的物理测试；而在波音 787 研制阶段，由于引入了大量的数字化设计工具与虚拟测试技术，大幅减少了对物理原型机的依赖，借助大规模的仿真模拟计算，仅进行了 11 种物理测试；对于下一代商用飞机，借助大量虚拟模拟，风洞实验次数有望进一步降至 3 至 4 次。波音 T-7ARed Hawk 喷气教练机更是基于建模和 3D 设计工具完成虚拟设计，研发周期仅 36 个月，较传统模式缩短约 80%，体现出计算对高端装备研发提效和降本的重大支撑作用。

6.4.2 台积电采用英伟达计算光刻平台突破先进芯片制程

在芯片制程从微米级迈向纳米级、甚至埃级的演进中，光刻技术始终是缩小晶体管尺寸的核心环节。然而，当制程节点进入 14nm、7nm、3nm 及以下后先进工艺，光学衍射效应、光刻胶化学反应偏差、晶圆表面缺陷等问题，会导致实际曝光的电路图形与设计目标严重偏离，甚至直接导致芯片功能失效。此时，计算光刻中的光学邻近效应修正（OPC）及反演光刻技术（ILT）等成为突破物理瓶颈、实现先进制程量产的刚需技术。

在 2024 年初，英伟达宣布台积电（TSMC）和新思科技（Synopsys）两大半导体行业巨头将使用其 cuLitho 计算光刻平台进行生产，加速制造并突破下一代先进半导体芯片的物理极限。台积电通过使用 cuLitho 计算光刻平台进行生产，将 GPU 加速运算整合到台积电的工作流程中，从而实现了性能的巨大飞跃，大幅提高了吞吐量，缩短了周期时间，使其能够加速下一代芯片技术的开发。一方面，英伟达的 cuLitho 平台通过 GPU 加速和并行计算，将光刻效率提升 40 倍；另一方面，英伟达开发了应用生成式人工智能的算法，cuLitho 平台结合生成式 AI 算法可进一步优化光刻流程中的模拟与校准环节，例如 AI 可自动修正芯片设计中的微小误差，减少人工干预，提升良品率。

第七章 总结与展望

7.1 新一代科研范式将由科学计算与 AI4S 深度融合驱动，成为引领科技创新的核心引擎

未来，科学计算与 AI4S 将深度融合成为新一代科研范式，以多尺度机理模拟、端到端高质量数据为核心支撑，全面赋能生物医药、新材料、先进制造等关键领域。依托高精度物理建模、大规模并行计算与生成式 AI 技术，传统依赖实验试错、周期长、成本高的研发模式将被彻底重构，在分子设计、材料开发、器件制造、催化反应模拟等关键环节实现预测前置、设计先行、验证闭环。通过持续突破微观机理表征、跨尺度数据贯通、专用算力优化等核心瓶颈，AI4S 将逐步成为支撑国家战略科技力量的重要底座，推动我国在关键核心技术领域从“跟跑”向“并跑、领跑”稳步跨越。

7.2 AI4S 能力跃升将持续释放产业创新效能，为新质生产力形成提供关键技术路径

随着算力、算法、数据三位一体体系不断完善，AI4S 将逐步实现从辅助分析到自主设计的能力跃升，为产业创新与新质生产力培育注入强劲动能。在高性能算力平台支撑下，AI 模型将更精准地学习物质底层规律，从数据拟合走向机理可解释、结果可信赖；“端到端”高质量数据体系的建成，将进一步提升模型在复杂科研场景中的泛化能力与落地效率。未来 AI4S 将深度嵌入研发全流程，大幅缩短新药、新材料、新器件、新装备的研发周期、降低试验成本与失败风险，真正实现“设计—模拟—验证—量产”的智能化提速，为突破关键核心技术瓶颈、构建现代化产业体系提供坚实支撑。

7.3 科学计算与 AI4S 的市场潜力将持续释放，形成规模增长与生态共赢的良性发展格局

随着技术落地加速与政策红利持续释放，科学计算与 AI4S 市场将迎来高速增长期，其下游覆盖的生物医药、新材料、半导体等领域市场空间极为广阔。未来，市场将逐步形成“算力供给—技术研发—场景应用—商业变现”的完整产业链，涌现出一批聚焦专用软件、算力服务、数据服务等细分领域的龙头企业与创新主体，形成梯度分明的产业格局。同时，随着国产化替代加速推进，全栈自研的科学计算软件、算力集群与数据体系将逐步占据市场主导地位，结合开源共享的生态理念，推动产学研用深度融合，实现技术创新、市场拓展与产业升级的协同发展，既为市场主体带来丰厚商业回报，也为科学计算与 AI4S 的持续迭代注入市场动力，助力我国在全球 AI4S 市场竞争中占据主动地位。

7.4 构建自主可控的 AI4S 生态是实现大国科技自立自强的战略必争方向

面向中长期科技发展战略，我国将持续加快构建自主可控、开放协同的科学计算与 AI4S 生态，以国家级统筹布局破解数据分散、算力不足、软件薄弱等瓶颈。通过整合高校、科研院所、企业、医院与重点实验室等多方力量，统一数据标准、共建共享高质量数据集、打造自主科学计算软件栈与专用算力基础设施，形成协同高效、安全可信的科学智能创新体系。此举将大幅提升我国在科学智能领域的核心竞争力，衔接技术创新、产业落地与市场发展，为应对全球性科技挑战、保障产业链供应链安全、实现高水平科技自立自强提供持久动力，助力我国抢占全球未来科技竞争与产业变革的战略制高点。

7.5 政策与行动建议

（1）在国家级科学数据中心层面：建议将端到端高质量科学数据集纳入国家科学数据体系重点建设范畴，设立专项扶持计划与配套资金支持，规范数据标准、归口汇聚与共享机制，系统性布局生物医药、新材料等重点领域端到端数据集建设，夯实我国 AI4S 发展的数据底座。

（2）算力基础设施层面：建议依托东数西算国家算力枢纽布局，在现有枢纽节点规划中，专项增设科学计算专用异构集群，适配 MD、QM、CAE 等多尺度仿真作业负载，统筹算力调度、存储带宽与网络适配能力，为科学计算规模化仿真、端到端数据生成提供稳定普惠的国家级算力支撑。

（3）产学研攻关机制层面：建议面向 AI4S 产业发展需求，设立 AI4S 科学数据构建揭榜挂帅专项项目，以实际应用痛点为牵引，鼓励行业龙头企业、超算平台、科研院所、高校组建创新联合体，围绕科学计算仿真补全、成败样本数据融合、数据集标准化与行业落地开展联合攻关，加快形成可复用、可产业化的 AI4S 数据体系与技术成果。

附录 缩略语中英文对照表

序号	英文缩写	中文全称
1	AI4S	科学智能
2	HPC	高性能计算
3	MD	分子动力学
4	QM	量子力学计算
5	CAE	计算机辅助工程
6	DFT	密度泛函理论
7	KMC	动力学蒙特卡洛
8	PCC	临床前候选化合物
9	IND	新药临床试验申请
10	CNN	卷积神经网络
11	FFT	快速傅立叶变换
12	3D Torus	3D 环形网络
13	CFD	计算流体力学
14	FEA/FEM	有限元分析
15	CRO	合同研究组织
16	PCC	临床候选化合物
17	sDDE	天穹新药引擎
18	SEI	固态电解质界面
19	OPC	光学邻近效应修正
20	ILT	反演光刻技术