# 对面向科学研究的人工智能技术的调研报告

刘翰文 522030910109

## 摘　要

新兴的人工智能为传统的科学研究提供了强大的工具，为科学研究带来了一个全新的研究范式。本调研报告列出了人工智能在当前科学研究中的应用方向。报告从数据的收集与生成、数据的表示、假设生成和实验模拟四个方面分析了人工智能的应用方向。同时，对于每个应用方向，我们给出了当前人工智能和科学研究进行融合的实例。最后，我们对人工智能技术在科学研究中所面临的挑战进行了讨论，包括数据质量、可解释性、分布外泛化等。本调研的目的在于为研究者和政策的制定提供参考，推动人工智能与科学研究的深度融合。

**关键词：**人工智能，科学研究，融合，应用

## 1　绪论

人工智能 (Artificial Intelligence, AI) 作为 21 世纪以来最具颠覆性的技术之一，自从它在 1956 年被首次提出以来，已经取得了显著的发展。早期的人工智能技术强调逻辑推理和专家系统，更像是按照人类既定的规则按部就班执行的机器。但是随着人工智能理论的不断涌现，计算资源的提升和数据资源的飞速增长，以数据驱动进行自主学习的新一代人工智能技术逐渐取代了传统的人工智能。人工智能在机器学习、自然语言处理、语音识别等领域的应用取得了巨大的进步。尤其是自 2010 年代深度学习的兴起以来，人工智能在大规模深度神经网络 (DNNs) [38] 上又掀起一阵热潮，从最初的卷积神经网络 (CNNs) [26] 和循环神经网络 (RNNs) [21] 到 Transformer [40]，再到近年来发展迅猛人工智能大模型，例如 GPTs [34]、BERT [14] 等。人工智能正在以惊人的速度革新人类社会的各个方面，从日常生活到工业生产，无处没有人工智能的身影。

在人工智能席卷各个领域的同时，人工智能对于自然科学研究领域而言也是一个重大的革新，它为传统的科学研究的数据、理论假设等方面提供

了强有力的工具。例如数据收集和分析作为科学研究的两个核心目标 [11]，长期以来提出了许多方法（从显微镜等物理设备到自助法等）来实现这一目标 [35]。而数据科学的兴起使得人工智能能够从海量数据集中找到其中隐藏的客观规律，为科学研究提供科学的指导。

发展面向科学研究的人工智能技术无疑是新时代人工智能技术发展的一项重要而艰巨的任务，这不仅依赖于人工智能算法、算力、数据的发展，更需要人工智能技术于科学研究的深度融合。数十年来，无数学者尝试将人工智能技术引入到科学研究，例如 Koscher 等人 [25] 使用预测模型和生成模型设计和合成染料分子，Deng 等人 [13] 提出了晶体汉密尔顿图神经网络 (CHGNet) 的原子建模方法预测未知原子电荷的晶体结构的特性。同时，也有许多学者对科学研究中的人工智能技术进行归纳总结，例如 Zhang 等人 [46] 阐述了人工智能在亚原子、原子、宏观系统等不同时空尺度的科学领域的现状，Wang 等人 [42] 为人工智能时代的科学探索做出了系统的概述。本调研基于 Wang 等人 [42] 提出的概述，在这基础上结合了近些年来人工智能技术在科学研究中的突破，为面向科学研究的人工智能技术做出了一个系统的调研。本调研旨在对过去的技术做出完整系统的总结同时为其他学者提供参考，助力人工智能技术与科学研究的深度融合。

# 2 融合方向

如图1所示，科学研究大体可以分为三个互相关联，逐层递进的阶段：观察现象、提出假设和设计进行实验。在这个过程中，人工智能技术在每个阶段都在一定程度上影响了传统的科学研究。尽管对于不同学科，人工智能技术与科学研究的融合形式有所不同，但从宏观的角度来看，可以把人工智能技术与科学研究的融合归纳在四个方面：数据的收集与生成、数据的表示、假设生成和实验模拟。在本节中，我们讨论了在这四个方面人工智能是怎么与科学研究进行融合。

## 2.1 数据的收集与生成

数据作为不管是人工智能还是科学研究的出发点，对它的收集和获取变得尤为重要。随着实验设备的不断迭代，实验平台收集的数据集的规模和复杂性呈指数型增长，传统的数据处理方法在面对当今海量的数据时往往显得乏力，导致科学研究越来越依赖人工智能和现代高性能计算机对数据

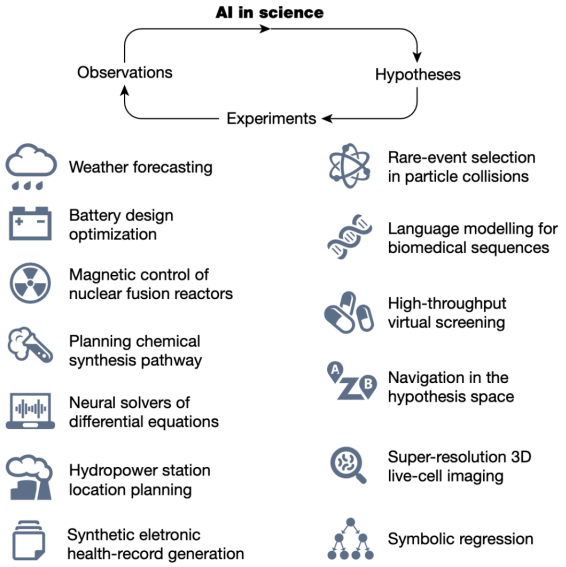


图 1: 科学研究的三个阶段：观察现象、提出假设和设计进行实验，以及人工智能技术在各学科的部分应用。

进行筛选和分析等操作。

人工智能首先可以做到对数据的智能选择。例如，在典型的粒子碰撞实验中，平均每秒大概会产生超过 100TB 的数据，但是其中超过 99.9% 的数据其实是研究者不需要的背景噪声 [23]。在对噪声数据的筛除过程中，深度学习中被用于检测异常的算法 (如 DNNs 的记忆效应 [6, 43]) 取代了传统基于先验知识剔除噪声的筛选器。IceCube [10] 提出了一种将最大似然估计与多重对称神经网络结合的方法，用于过滤中微子天文台观测数据中的噪声数据，创新性地提高了天文台的观测能力。

人工智能还能被用于生成科学研究所需的新数据。强化学习中的数据增强作为典型的方法，已经被广泛应用于数据集的扩充。近些年来，生成式人工智能 (AIGC) 的蓬勃发展更拓宽了数据生成的前景，变分自编码器 (VAEs) [24]、生成对抗网络 (GANs) [19] 和扩散模型 (Diffusion Model) [20] 等可以学习数据的分布并生成类似分布的新数据。为了解决生物医学中可用的观测数据量稀缺的问题，Marouf 等人 [29] 提出了条件单细胞生成对抗网络 (cscGAN) 来生成单细胞 RNA 序列数据。ATLAS [9] 利用变分自编码器和生成对抗网络生成符合客观规律的电磁簇射用以满足对大规模实验模拟结果的需求。

## 2.2 数据的表示

深度学习的出现使得在不同层次上提取科学数据的有意义的表述成为现实，Word2Vec [31]、卷积神经网络的卷积结果等已经被用于数据的表示在许多邻域发挥了重要的作用。这些表述是用来处理复杂数据的重要前提。高质量的表述应在保持简单紧凑的同时能够清楚的表示数据的特征，可以编码数据中的潜在规律。在科学研究中，提取复杂的高维数据的有意义的表述无疑可以帮助研究人员理解数据中隐含的联系，在药物研发、材料探索、生物结构研究中发挥重大的作用。

图2展示了学习数据表示的三种不同策略：几何深度学习、自监督学习和掩码语言建模。几何深度学习是深度学习的一个分支，专注于处理非欧几里得数据，几何深度学习使得算法能够发掘利用复杂数据中的结构和关系信息，图神经网络 (GNNs) [49] 作为其中的关键技术，已经在众多邻域发挥了重大作用。Merchant 等人 [30] 提出了用于材料探索的图神经网络 (GNoME) 用于筛选和改进模型预测的材料结构，Zhang 等人 [45] 提出了 Higashi，一种基于超图表示学习的算法，用于理解细胞中染色质的结构。

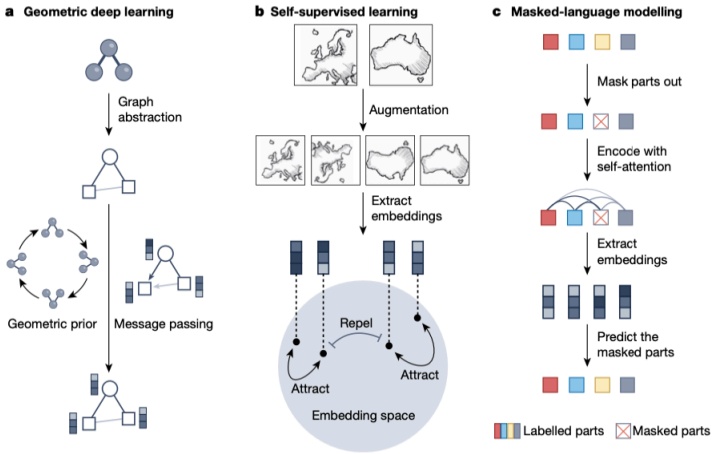


图 2: 学习数据表示的三种策略：**a**、几何深度学习，**b**、自监督学习，**c**、掩码语言建模。

自监督学习是一种能使模型不依赖于标签而学习数据集的一般特征的学习方法。这些自监督学习方法往往通过解决一些预先设定的任务 (如预测图像中的遮挡区域、预测视频中的随机帧等) 来学习数据集的特征。Ji 等人 [22] 参考 BERT 提出了 DNABERT，使用自监督学习中的预测任务，通过预测屏蔽的 DNA 序列进行模型预训练用于更好的理解生物中的 DNA 信

息。Moor 等人 [32] 提出了一种新的医疗 AI 范式 GMAI，通过对大型多样化的多模态医疗数据集进行自监督预训练，用于解释医疗多模态数据和执行医疗应用。

掩码语言建模是自然语言处理中自监督学习的一种经典方法，通过上下文信息预测被掩盖的词来学习语言的特征。掩码语言建模也是生物序列自监督学习的一种常用方法，将氨基酸序列或 DNA 序列视作句子，通过掩码预测学习的方法，在模型训练上已经产生了很好的效果。Rivers 等人 [37] 使用氨基酸序列的掩码语言建模，得到包含蛋白质特性表征的模型。Shepherd 等人 [15] 使用掩码语言建模自监督学习的语言模型学习复杂的分子分布结构。

## 2.3 假设生成

假设作为科学研究的一个关键的阶段，决定了科学研究探索的方向，后续实验均是围绕提出的假设构建。在传统的科学研究中，假设的提出往往离不开对学科理论的深刻理解和对现象的长期观测，这使得一个有意义的假设无法被轻易提出。人工智能技术的出现使得科学假设的大量涌现成为可能，图3展示了三种人工智能技术帮助科学研究生成假设的可能的方式：黑箱预测器、探索组合假设空间和优化可微假设空间。

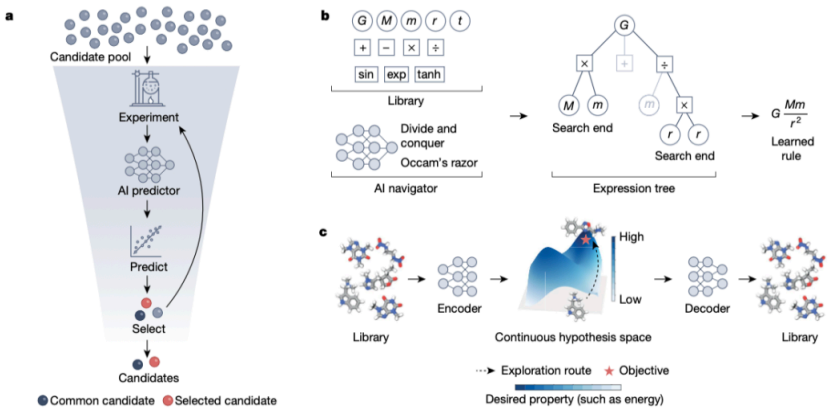


图 3: 假设生成的三种方式：**a**、黑箱预测，**b**、探索组合假设空间，**c**、优化可微假设空间。

在科学研究中，研究者往往需要分析大量复杂的数据来发现其中的规律来验证提出的假设，因此对所有待验证的假设全部进行实验验证是一件

费时费力的事。因此，黑箱预测器起到了过滤假设的作用，它具备在大多数情况下能够有效的筛选那些可以最大限度提高实验效率和收益的假设进行验证的能力。Liu 等人 [28] 使用神经网络作为黑箱预测器从大量假设分子中筛选出具有抑菌能力的新分子，Bombarelli 等人 [17] 从 160 万种假设分子中筛选出其中的 40 万种候选分子用于新型有机发光二极管的制造。

探索组合假设空间可以表述为一个优化问题，目标与黑箱类似，是找到一个有意义的假设。这通常是由强化学习训练如何在假设搜索空间中找到具有更高价值的搜索方向，聚焦在最有前景的假设上。探索组合假设空间可以被用于药物研发上，具体而言，对于一个分子的结构设计都可以看作一系列离散的选择：在何处添加什么原子。强化学习可以规避在搜索空间中漫无目的的搜寻，而是优先考虑最有希望的分支来有效地指导搜索。Zhavoronkov 等人 [48] 提出了一种生成张量强化学习模型 (GENTRL) 在假设空间中搜寻对 DDR1 有活性的新型化合物分子，Zhou 等人 [50] 提出了分子深度 Q-网络 (MolDQN) 使用强化学习进行分子优化。

优化可微假设空间将上文提到的离散的科学假设放到可微空间中，使得基于梯度的方法可以被用于找到局部最优假设。通常的方法使用 VAEs 等模型将离散候选假设映射到可微空间中或将原离散假设去除离散的约束进行条件的放宽使之在可微空间上能进行优化。Gabbard 等人 [16] 使用预训练的 VAEs 将引力波信号的先验参数空间进行映射获得描述后验分布的样本，Bombarelli 等人 [18] 使用 VAEs 将分子的 SMILES 字符串表示映射到连续的向量空间，允许使用贝叶斯优化技术进行优化。

## 2.4 实验模拟

在得到有意义的科学假设之后的一个重要环节是通过科学实验评估科学假设。但由于现实的种种原因，进行完整彻底的科学实验成本高昂且不切实际。随着科学技术的发展，计算机的模拟已经能够在一定程度上替代一部分简单的科学实验，但是计算机的模拟需要人工设置参数和许多先验知识，且无法保证速度和精度。而人工智能技术的崛起为这些难题带来了希望，通过深度学习的方法，可以进行识别和优化科学假设来进行有效的科学实验。具体而言，人工智能可以规划和指导实验测试的进行，它通过一种系统的方法来规划设计实验以提高实验效率，同时使用强化学习的方法通过学习先前的实验和结果动态地指导调整实验的进程，将实验过程引导到高价值的假设上。除此之外，人工智能还可以通过精准地拟合复杂系统的关键参数、

准确求解复杂系统的方程或建模复杂系统的状态等方法来提高计算机模拟的性能。

图4展示了人工智能技术和科学实验模拟融合的三个应用。除此之外，还有许多人工智能的应用，如 Bellemare 等人 [7] 使用强化学习构造了一个高性能飞行控制器，用于在环境的干扰下做出实时决策，D.Smith 等人 [39] 提出了 EikoNet 用于求解 Eikonal 方程，Noé 等人 [33] 提出一种使用正则化流模拟多体系统中平衡状态的有效采样算法，用来对复杂系统进行建模等。

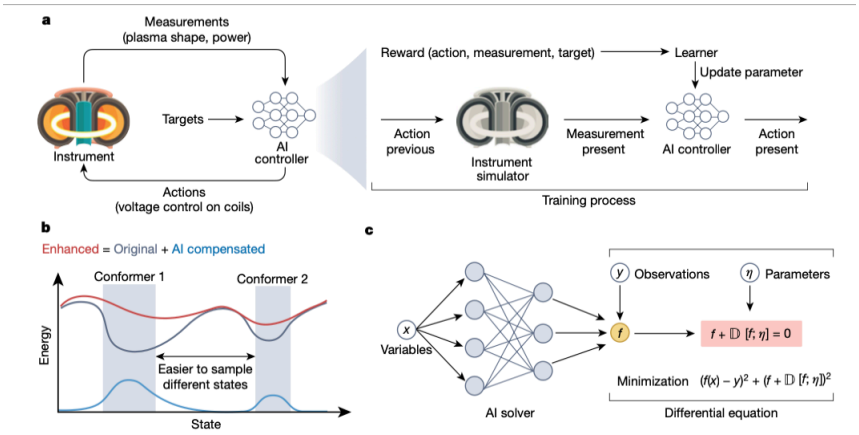


图 4: **a**、Degrave 等人 [12] 使用强化学习的方法动态控制复杂动态系统的核聚变。**b**、Wang 等人 [41] 使用神经网络估计不确定性的方法指导对势能进行补偿，从而提高模拟效率和准确性。**c**、使用物理信息神经网络 (PINN) [36] 求解偏微分方程的框架

## 3 挑战

人工智能技术在科学研究领域所展现出的光明前景无可否认，但在目前对人工智能技术的引用上任面临着一系列技术、理论和实践上的挑战。为了人工智能技术在科学研究领域上持续发展融合，必须提前明确这些挑战。本文在这里例举了三点目前面临的挑战：科学数据的问题、可解释性问题和分布外泛化问题，同时这些也是人工智能领域中亟待解决的问题。

在科学研究领域，往往需要高质量的数据作为支撑，但由于测量设备、测量方法等实际问题的影响，测量的数据往往存在噪声、缺失和偏差等问

题。另外，由于不同人工智能需要的数据格式不同，为了方便数据的转化，需要对数据进行统一的标准化。而且由于隐私和安全等问题，收集大型的数据集不是一件轻易的事。同时，有人工标注数据的稀缺也成为了制约人工智能技术在科学研究领域进一步发展重要限制之一。

人工智能模型尤其是神经网络模型，往往被看作一个黑箱，缺乏在科学研究领域中非常重视的推导过程部分，而往往只有输入和输出，缺乏直观的可解释性，这也是多数科研人员不完全信任人工智能技术的一个主要原因。而没有可解释性的模型往往无法在高精尖领域得以应用，制约了人工智能技术在各领域上的全面发展。尽管近年来，越来越多的研究者开始将目光投向模型的解释领域 [44, 47]，但在该领域仍尚未发展成熟。

在人工智能技术中，往往假设数据集符合独立同分布，无法在一些分布中 (如长尾分布等) 展现出良好的效果。而一旦进入科学研究领域，大量未知分布的数据和极端数据的存在制约了模型的实际性能，模型需要提高自身的泛化能力。分布外泛化在人工智能领域也是一个前沿问题，许多方法被提出 [8, 27]，但缺少一个通用框架用以解决泛化问题。

除了上述三个挑战，仍有许多挑战未被提及，如多模态数据的处理、先验科学知识的引入、与科学研究设备和硬件的相互依赖等，人工智能在科学研究领域上任重而道远。

# 4 总结

人工智能正以前所未有的速度推动科学研究的变革，它凭借其出色的能力在科学研究领域得到了应用，在数据收集与生成、数据的表示、假设生成和实验模拟等方面上表现出开阔的前景。为了使人工智能技术能够在科学研究领域领域持续发展，必须认识当下面临的一系列挑战，包括数据、可解释性、泛化能力等。我们相信在未来，人工智能有望进一步加速科学发现的进程，为解决全球性难题和推动人类社会进步提供坚实的技术支撑。

本文主要基于 Wang 等人的综述 [42]，图1-图4均从中截取，有少量部分参考了 GPT-4o 和 Kimi 生成的文本。参考的网站博客 [1–5]。GPTZero 检测结果如图5所示，自查重结果如图6所示。

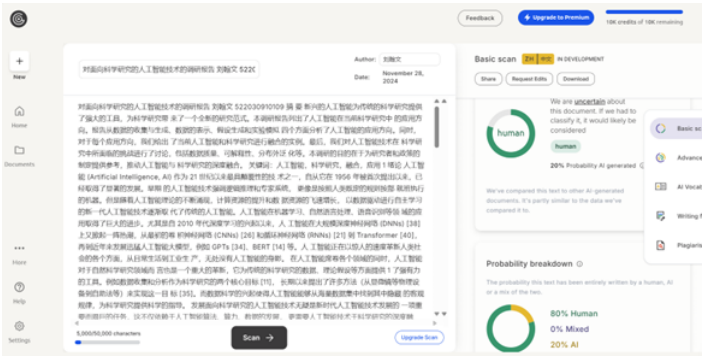


图 5: GPTZero 检测结果


图 6: 自查重结果

# 参考文献

[1] 2023 年人工智能在科学领域的应用：社区入门指南. https://medium.com/@AI_for_Science/ai-for-science-in-2023-a-community-primer-d2c2db37e9a7.

[2] Bengio 团队 nature 发文：四个维度讲 ai for science，还讨论了 ai 跨界核心挑战. https://www.aixinzhijie.com/article/6829598.

[3] Yoshua bengio 领衔跨学科团队，nature 刊文综述人工智能时代的科学发现. https://www.linkresearcher.com/theses/0c251423-b7e0-4be5-99e5-128c77226a2b.

[4] 什么是几何深度学习（geometric deep learning）？ https://blog.csdn.net/sinat_39434559/article/details/125996388.

[5] 多国 63 位学者发布“ai for science”综述. https://ecas.cas.cn/xxkw/kbcd/201115_143064/ml/xxhzlyzc/202308/t20230823_4965338.html.

[6] Devansh Arpit, Stanisław Jastrzębski, Nicolas Ballas, David Krueger, Emmanuel Bengio, Maxinder S Kanwal, Tegan Maharaj, Asja Fischer, Aaron Courville, Yoshua Bengio, et al. A closer look at memorization in deep networks. In *International conference on machine learning*, pages 233–242. PMLR, 2017.

[7] Marc G Bellemare, Salvatore Candido, Pablo Samuel Castro, Jun Gong, Marlos C Machado, Subhodeep Moitra, Sameera S Ponda, and Ziyu Wang. Autonomous navigation of stratospheric balloons using reinforcement learning. *Nature*, 588(7836):77–82, 2020.

[8] Kamalika Chaudhuri, Claire Monteleoni, and Anand D Sarwate. Differentially private empirical risk minimization. *Journal of Machine Learning Research*, 12(3), 2011.

[9] ATLAS collaboration et al. Deep generative models for fast photon shower simulation in atlas. *arXiv preprint arXiv:2210.06204*, 2022.

[10] IceCube Collaboration*†, R Abbasi, M Ackermann, J Adams, JA Aguilar, M Ahlers, M Ahrens, JM Alameddine, AA Alves Jr, NM Amin, et al. Observation of high-energy neutrinos from the galactic plane. *Science*, 380(6652):1338–1343, 2023.

[11] Henk W De Regt. Understanding, values, and the aims of science. *Philosophy of Science*, 87(5):921–932, 2020.

[12] Jonas Degrave, Federico Felici, Jonas Buchli, Michael Neunert, Brendan Tracey, Francesco Carpanese, Timo Ewalds, Roland Hafner, Abbas Abdolmaleki, Diego de Las Casas, et al. Magnetic control of tokamak plasmas through deep reinforcement learning. *Nature*, 602(7897):414–419, 2022.

[13] Bowen Deng, Peichen Zhong, KyuJung Jun, Janosh Riebesell, Kevin Han, Christopher J Bartel, and Gerbrand Ceder. Chgnet as a pretrained universal neural network potential for charge-informed atomistic modelling. *Nature Machine Intelligence*, 5(9):1031–1041, 2023.

[14] Jacob Devlin. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[15] Daniel Flam-Shepherd, Kevin Zhu, and Alán Aspuru-Guzik. Language models can learn complex molecular distributions. *Nature Communications*, 13(1):3293, 2022.

[16] Hunter Gabbard, Chris Messenger, Ik Siong Heng, Francesco Tonolini, and Roderick Murray-Smith. Bayesian parameter estimation using conditional variational autoencoders for gravitational-wave astronomy. *Nature Physics*, 18(1):112–117, 2022.

[17] Rafael Gómez-Bombarelli, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, David Duvenaud, Dougal Maclaurin, Martin A Blood-Forsythe, Hyun Sik Chae, Markus Einzinger, Dong-Gwang Ha, Tony Wu, et al. Design of efficient molecular organic light-emitting diodes by a high-throughput virtual screening and experimental approach. *Nature materials*, 15(10):1120–1127, 2016.

[18] Rafael Gómez-Bombarelli, Jennifer N Wei, David Duvenaud, José Miguel Hernández-Lobato, Benjamín Sánchez-Lengeling, Dennis Sheberla, Jorge Aguilera-Iparraguirre, Timothy D Hirzel, Ryan P Adams, and Alán Aspuru-Guzik. Automatic chemical design using a data-driven continuous representation of molecules. *ACS central science*, 4(2):268–276, 2018.

[19] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial networks. *Communications of the ACM*, 63(11):139–144, 2020.

[20] Jonathan Ho, Ajay Jain, and Pieter Abbeel. Denoising diffusion probabilistic models. *Advances in neural information processing systems*, 33:6840–6851, 2020.

[21] S Hochreiter. Long short-term memory. *Neural Computation MIT-Press*, 1997.

[22] Yanrong Ji, Zhihan Zhou, Han Liu, and Ramana V Davuluri. Dnabert: pre-trained bidirectional encoder representations from transformers model for dna-language in genome. *Bioinformatics*, 37(15):2112–2120, 2021.

[23] Georgia Karagiorgi, Gregor Kasieczka, Scott Kravitz, Benjamin Nachman, and David Shih. Machine learning in the search for new fundamental physics. *Nature Reviews Physics*, 4(6):399–412, 2022.

[24] Diederik P Kingma. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[25] Brent A. Koscher, Richard B. Canty, Matthew A. McDonald, Kevin P. Greenman, Charles J. McGill, Camille L. Bilodeau, Wengong Jin,

Haoyang Wu, Florence H. Vermeire, Brooke Jin, Travis Hart, Timothy Kulesza, Shih-Cheng Li, Tommi S. Jaakkola, Regina Barzilay, Rafael Gómez-Bombarelli, William H. Green, and Klavs F. Jensen. Autonomous, multiproperty-driven molecular discovery: From predictions to measurements and back. *Science*, 382(6677):eadi1407, 2023.

[26] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. *Advances in neural information processing systems*, 25, 2012.

[27] David Krueger, Ethan Caballero, Joern-Henrik Jacobsen, Amy Zhang, Jonathan Binas, Dinghuai Zhang, Remi Le Priol, and Aaron Courville. Out-of-distribution generalization via risk extrapolation (rex). In *International conference on machine learning*, pages 5815–5826. PMLR, 2021.

[28] Gary Liu, Denise B Catacutan, Khushi Rathod, Kyle Swanson, Wengong Jin, Jody C Mohammed, Anush Chiappino-Pepe, Saad A Syed, Meghan Fragis, Kenneth Rachwalski, et al. Deep learning-guided discovery of an antibiotic targeting acinetobacter baumannii. *Nature Chemical Biology*, 19(11):1342–1350, 2023.

[29] Mohamed Marouf, Pierre Machart, Vikas Bansal, Christoph Kilian, Daniel S Magruder, Christian F Krebs, and Stefan Bonn. Realistic in silico generation and augmentation of single-cell rna-seq data using generative adversarial networks. *Nature communications*, 11(1):166, 2020.

[30] Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85, 2023.

[31] Tomas Mikolov. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 3781, 2013.

[32] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[33] Frank Noé, Simon Olsson, Jonas Köhler, and Hao Wu. Boltzmann generators: Sampling equilibrium states of many-body systems with deep learning. *Science*, 365(6457):eaaw1147, 2019.

[34] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[35] John V Pickstone. *Ways of knowing: A new history of science, technology, and medicine*. University of Chicago Press, 2001.

[36] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

[37] Alexander Rives, Joshua Meier, Tom Sercu, Siddharth Goyal, Zeming Lin, Jason Liu, Demi Guo, Myle Ott, C Lawrence Zitnick, Jerry Ma, et al. Biological structure and function emerge from scaling unsupervised learning to 250 million protein sequences. *Proceedings of the National Academy of Sciences*, 118(15):e2016239118, 2021.

[38] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[39] Jonathan D. Smith, Kamyar Azizzadenesheli, and Zachary E. Ross. Eikonet: Solving the eikonal equation with deep neural networks. *IEEE Transactions on Geoscience and Remote Sensing*, 59(12):10685–10696, 2021.

[40] A Vaswani. Attention is all you need. *Advances in Neural Information Processing Systems*, 2017.

[41] Dongdong Wang, Yanze Wang, Junhan Chang, Linfeng Zhang, Han Wang, and Weinan E. Efficient sampling of high-dimensional free energy landscapes using adaptive reinforced dynamics. *Nature Computational Science*, 2(1):20–29, 2022.

[42] Hanchen Wang, Tianfan Fu, Yuanqi Du, Wenhao Gao, Kexin Huang, Ziming Liu, Payal Chandak, Shengchao Liu, Peter Van Katwyk, Andreea Deac, et al. Scientific discovery in the age of artificial intelligence. *Nature*, 620(7972):47–60, 2023.

[43] Xiaobo Xia, Tongliang Liu, Bo Han, Chen Gong, Nannan Wang, Zongyuan Ge, and Yi Chang. Robust early-learning: Hindering the memorization of noisy labels. In *International conference on learning representations*, 2020.

[44] Muhammad Rehman Zafar and Naimul Khan. Deterministic local interpretable model-agnostic explanations for stable explainability. *Machine Learning and Knowledge Extraction*, 3(3):525–541, 2021.

[45] Ruochi Zhang, Tianming Zhou, and Jian Ma. Multiscale and integrative single-cell hi-c analysis with higashi. *Nature biotechnology*, 40(2):254–261, 2022.

[46] Xuan Zhang, Limei Wang, Jacob Helwig, Youzhi Luo, Cong Fu, Yaochen Xie, Meng Liu, Yuchao Lin, Zhao Xu, Keqiang Yan, et al. Artificial intelligence for science in quantum, atomistic, and continuum systems. *arXiv preprint arXiv:2307.08423*, 2023.

[47] Haiyan Zhao, Hanjie Chen, Fan Yang, Ninghao Liu, Huiqi Deng, Hengyi Cai, Shuaiqiang Wang, Dawei Yin, and Mengnan Du. Explainability for large language models: A survey. *ACM Transactions on Intelligent Systems and Technology*, 15(2):1–38, 2024.

[48] Alex Zhavoronkov, Yan A Ivanenkov, Alex Aliper, Mark S Veselov, Vladimir A Aladinskiy, Anastasiya V Aladinskaya, Victor A Terentiev, Daniil A Polykovskiy, Maksim D Kuznetsov, Arip Asadulaev, et al. Deep learning enables rapid identification of potent ddr1 kinase inhibitors. *Nature biotechnology*, 37(9):1038–1040, 2019.

[49] Jie Zhou, Ganqu Cui, Shengding Hu, Zhengyan Zhang, Cheng Yang, Zhiyuan Liu, Lifeng Wang, Changcheng Li, and Maosong Sun. Graph neural networks: A review of methods and applications. *AI open*, 1:57–81, 2020.

[50] Zhenpeng Zhou, Steven Kearnes, Li Li, Richard N Zare, and Patrick Riley. Optimization of molecules via deep reinforcement learning. *Scientific reports*, 9(1):10752, 2019.