arXiv:2303.08060v1 [cs.AI] 14 Mar 2023

超越游戏：神经蒙特卡洛树搜索应用的系统综述

Marco Kemmerling[1*]、Daniel L¨utticke[1] 和 Robert H. Schmitt[1]

[1]机械工程信息管理（WZL-MQ/IMA），亚琛工业大学，德国亚琛。

*通讯作者。电子邮箱：marco.kemmerling@ima.rwth-aachen.de；

摘要

AlphaGo及其后继系统的出现，标志着利用人工智能进行博弈进入了一个新范式。这一成果是通过结合蒙特卡洛树搜索（一种规划程序）与深度学习实现的。尽管这对博弈领域的影响毋庸置疑，但类似方法在博弈以外的应用中有多大用处，以及它们需要如何从原有方法论调整而来，仍不甚明了。我们综述了129篇经同行评审的论文，这些论文详细介绍了神经蒙特卡洛树搜索方法在博弈以外领域的应用。我们的目标是系统评估这类方法在实践中的构造方式，以及它们的成功能否拓展到其他领域。我们发现，这些方法应用于各种领域，通过学习得到的策略函数和值函数引导树搜索的方式多种多样，训练方法也各不相同。本综述梳理了神经蒙特卡洛树搜索这一算法家族应用于实际问题时的当前面貌，这是迈向更有原则地针对特定问题及其需求设计此类算法的第一步。

关键词：蒙特卡洛树搜索，MCTS，神经蒙特卡洛树搜索，强化学习 ng, 基于模型的强化学习、决策时规划

1 引言

.....的结合蒙特卡洛树搜索（MCTS）与深度学习促成了一个具有历史意义的事件：计算机程序 AlphaGo 在围棋博弈中击败了人类冠军 [99]，而长期以来，人们一直认为计算方法无法胜任这一任务。此后，此类方法——在本综述中我们称之为神经蒙特卡洛树搜索——广受欢迎。它们已应用于许多其他博弈，并在通用游戏对弈领域取得了令人鼓舞的成果 [89, 100]。

尽管神经蒙特卡洛树搜索在博弈情境中的有效性已得到明确证实，但将此类方法迁移到非游戏应用领域仍是相当近期的发展，因此人们对其了解较少。

过去曾有一些关于蒙特卡洛树搜索方法的调查与综述发表，例如 Mandziuk [68] 对部分蒙特卡洛树搜索在非游戏问题中的应用进行了调查。然而，该文仅考察了两种不同的应用，且均未涉及任何神经引导。

更全面的综述见 [161]，其中还简要介绍了蒙特卡洛树搜索与深度学习的结合。不过，他们关注的是博弈，而非其他应用。

据我们所知，目前尚无针对神经蒙特卡洛树搜索方法在非游戏领域应用的全面综述。我们认为，这类综述有助于阐明此类方法向更实际问题迁移的程度，以及如何设计神经蒙特卡洛树搜索方法以满足不同应用场景的要求。为此，我们希望为更具原则性的神经蒙特卡洛树搜索方法设计指南贡献一份力量。

以下研究问题为本综述提供指引：

1. 神经蒙特卡洛树搜索应用于哪些学科、领域和应用方面？已观察到的应用有哪些共性和差异？2. 与博弈中的应用相比，神经蒙特卡洛树搜索方法的设计有哪些不同？3. 在树搜索过程中，神经引导可以用于哪些环节，以及如何使用？

为了解答这些问题，我们开展了系统性文献综述，并分析所得文献。我们从多个数据库中进行关键词检索，筛选出相关文章，随后进行额外的前向与后向检索，并从纳入综述的每篇文章中提取一组预先设定的数据项。详细的综述流程见第3节。接下来一节将简要介绍强化学习、蒙特卡洛树搜索和 AlphaZero 的相关概念。其余各节首先在第4节聚焦于所调查出版物中描述的问题，接着在第5节考察所采用的方法。最后，我们将在第6节简要讨论本综述。

2 强化学习与神经蒙特卡洛树搜索

2.1 强化学习

强化学习（RL）是一种机器学习范式，其中智能体从环境中收集的经验里学习。为此，智能体会观察环境的状态 s ，并根据该状态执行动作 a 。采取动作后，智能体会收到奖励 r 并观察新的状态 s'。遵循这种表述方式的问题称为马尔可夫决策过程（MDP），当新状态 s' 仅取决于紧邻其前的状态 s 和智能体的动作 a 时即是如此。在这种马尔可夫决策过程中，智能体的目标是通过学习恰当的策略 π来最大化回报，即长期累积奖励的期望值。策略是一种行为策略，它针对给定状态规定一个动作或动作的概率分布[108]。

这种策略可以直接从经验中学习，例如通过策略梯度方法；也可以从学习得到的动作价值函数中推导出来。动作价值函数 $Q^{\pi}(s,\ a)$ 估计的是在状态 s 下执行动作 a 的价值，即预期回报。从这样学习得到的动作价值函数中，可以通过贪心地选择最佳动作推导出确定性策略，也可以通过按照动作价值的比例进行采样，推导出随机策略。除了前面介绍的基于策略和基于价值的方法外，还常采用能够协同学习策略和价值函数的混合方法。这类方法称为演员-评论家方法，通常学习状态价值函数 $V^{\pi}(s)$，而非动作价值函数 $Q^{\pi}(s,a)$。前者计算遵循策略 π 时状态 s 的预期回报；后者则计算先执行动作 a 、随后在后续步骤中遵循 π 时状态 s 的预期回报。为使记法更简洁，上标 π 通常会省略。

与无模型方法不同，在基于模型的强化学习中，会使用环境模型进行规划。模型会精确或近似地模拟环境动态。规划指的是使用模型进行经验模拟，规划方法可分为后台规划和决策时规划。在前者中，收集自环境的真实经验所构成的训练数据，会通过模型生成的虚拟经验加以扩充。在后者中，给定时间步的动作选择取决于使用模型进行前瞻规划，即想象不同动作选择所产生的后果，以改进当前状态下的策略[73]。蒙特卡洛树搜索（下文将进一步解释）可视为一种决策时规划。

2.2 强化学习与蒙特卡洛树搜索之间的联系

蒙特卡洛树搜索最初是一种用于组合博弈的启发式搜索方法，它通过随机采样执行一种树搜索。在这类博弈中，玩家必须在给定状态下决定采取哪个动作，以便在博弈的 z终止状态中获得最大收益。尽管蒙特卡洛树搜索通常并不被视为一种强化学习，但这里描述的情境与前文给出的强化学习问题表述十分相似，一些作者也详细探讨了这种联系 [116]。为避免歧义，本文不会使用“强化学习”一词来指称蒙特卡洛树搜索。与强化学习类似，蒙特卡洛树搜索也会生成策略 π_{MCTS} 和价值估计v_{MCTS}。在蒙特卡洛树搜索中，会通过多步前瞻搜索为给定状态生成策略 s ，即考虑未来情境，并确定从 s出发时，哪一系列动作能够带来有利结果。每遇到一个状态，都会重新生成相应策略；也就是说，确定的策略无法泛化到当前遇到状态以外的其他状态。相比之下，传统强化学习通过从过去的经验中学习来生成策略，旨在泛化到未见过的情境。在决策时，不会执行前向搜索，而是根据从过去经验中学到的策略直接选择动作。从某种意义上说，蒙特卡洛树搜索着眼未来，而传统强化学习则回顾过去来确定动作。因此，强化学习需要成本高昂的前期训练，但决策时的计算成本可以忽略不计；而蒙特卡洛树搜索无需训练，但决策时需要进行计算成本高昂的规划。

2.3 蒙特卡洛树搜索

蒙特卡洛树搜索的总体思路是蒙特卡洛树搜索通过平衡对较少访问的树分支的探索与对已知且有前景的树分支的利用，逐步构建解空间的搜索树。该过程通过反复执行四个不同阶段来实现：选择、扩展、评估和反向传播。在选择阶段，从根节点出发，依次选择动作，直到到达叶节点 s_L 。随后在扩展阶段向该叶节点添加新的子节点，并在评估阶段估算它们的价值。最后，将新添加节点的价值沿树向上传播，以更新通往 s_L 的路径上各节点的价值。下文将更详细地介绍这些阶段。

选择

在选择阶段，从根节点出发，按照某种机制选取一个动作。这会到达一个新状态，并在该状态下再次应用选择机制。重复这一过程，直到遇到叶节点。

动作选择机制称为树策略。尽管存在不同的机制，UCB1 [4] 公式是一个常见的选择。当蒙特卡洛树搜索使用 UCB1 公式时，所得到的算法称为树的置信上限（UCT） [53]。在UCT中，动作选择定义如下：

$$a = \arg\max_{a} \frac{W(s,a)}{N(s,a)} + c\sqrt{\frac{ln\ N(s)}{N(s,a)}} \qquad (1)$$

其中， $W(s,\ a)$ 表示截至目前为止，在搜索过程中于状态 s 下选择动作 a 时获得的胜利次数； $N(s,\ a)$ 表示在 s 中选择 a 的次数； $N(s)$ 表示访问 s 的次数。总和的左侧部分鼓励利用已知能够带来良好结果的动作，其中分数$W\ (s,\ a)$ $N\ (s,\ a)$ 可视为 $Q(s,\ a)$ 的近似值。总和的右侧部分则通过提高相较于状态总访问次数而言访问较少的动作的权重来鼓励探索。探索与利用通过探索常数 c 进行平衡。对于游戏结果 $z \in [0,1]$, c 的最优选择是 $c = \frac{1}{\sqrt{2}}$ [54]；但对于超出此范围的奖励，可能需要调整 c [8]。

扩展

经过反复执行选择步骤后，搜索可能会到达一个尚有未扩展潜在子节点的节点。一旦出现这种情况，该节点的一个或多个子节点就会被扩展。此阶段可能存在一些不同做法。在某些情况下，遇到叶节点（即没有子节点的节点）时，会扩展所有可能的子节点。

在其他情况下，遇到可扩展节点（即拥有尚未扩展的 子节点的节点）时，只扩展一个子节点。立即扩展所有子节点可能会导致树的增长不尽如人意，具体取决于应用。

在一些文献中，可扩展节点也被称为叶节点。为清晰起见，本文中我们仅用“叶节点”指代没有任何子节点的真正叶节点。请注意，叶节点与终止节点并不相同：前者只是树枝当前的末端，后者则是表示博弈结束状态的节点（见图。1）。

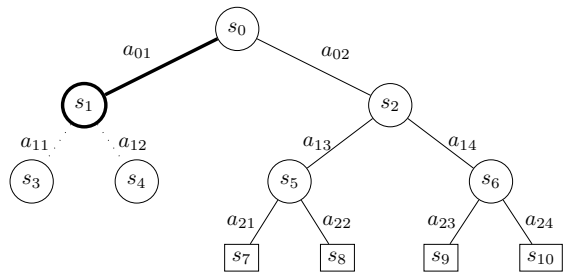


图 1 搜索过程中某一时刻的搜索树。当前节点以粗边框标示，当前搜索迭代中遍历过的边也以粗线标示。 s_1 当前是一个叶节点，其潜在子节点 s_3 和 s_4 正等待扩展，如虚线边所示。终端节点 s_7、 s_8、 s_9 和 s_10 以矩形节点表示。

评估
节点扩展后，会对其进行评估，以初始化 $W(s, a)$ 和 $N(s, a)$。这种评估有时也称为模拟、滚动模拟或对局模拟，其过程是从新扩展的节点开始进行博弈，直至遇到终止状态。终止状态下的结果 z 即为评估结果。博弈按照默认策略进行，该策略决定了新扩展节点与终止节点之间的动作序列。在最简单的情况下，默认策略会以均匀随机的方式采样动作 [54]。

除了评估新扩展的节点，也可以只评估叶节点，并将新扩展节点的评估留待搜索过程中的后续时刻进行，届时这些节点会再次作为叶节点被遇到。

反向传播
评估阶段的结果 z 沿树向上传播，以更新前序节点中的 $W(s, a)$。所有被选择节点的访问次数也会递增。

反向传播阶段结束后，流程会从选择阶段重新开始，直到达到预先设定的模拟预算。

2.4 AlphaZero

名为*AlphaGo* 的程序因结合深度学习和 [99] 蒙特卡洛树搜索，成为首个在完整棋盘的围棋对局中击败职业人类玩家的计算机程序，因而备受关注。AlphaGo 在强化学习之前，依赖于对人类专家棋步进行监督式预训练。其后继者 AlphaGo Zero 仅通过自我对弈进行强化学习训练。它还通过减少所用神经网络的数量，进一步简化了训练过程。AlphaGo Zero 仍是专为围棋开发的，并采用了一些针对该博弈的机制。相比之下，AlphaGo 系列的下一代 AlphaZero 更具通用性，可应用于多种棋盘游戏。

本小节介绍的算法都是神经网络蒙特卡洛树搜索的例子，即由神经网络引导的蒙特卡洛树搜索。尽管此处我们重点介绍广受欢迎的 AlphaZero 系列，但类似思想也曾由不同研究者独立提出，并称为专家迭代 [2]。下文将详细介绍 AlphaZero，作为用于博弈的神经网络蒙特卡洛树搜索方法的代表。

与常规的蒙特卡洛树搜索一样，AlphaZero 遵循选择、扩展、评估和反向传播这四个阶段。部分阶段由神经网络 f_θ辅助；该网络根据给定状态生成策略向量$\mathbf{p}$，即所有动作的概率分布，并估计 v 状态价值。

AlphaZero 的选择阶段采用了预测器 + UCT（PUCT）公式[83]:

$$a = \arg\max_a Q(s,a) + c\ P(s,a)\ \frac{\sqrt{N(s)}}{1+N(s,a)} \quad (2)$$

其中， $P(s, a)$ 表示在状态 s 下选择动作 a 的先验概率，由 f_θ [100] 给出。

当选择阶段到达叶节点 s_L 时，神经网络会对其进行评估（p v） $=f_\theta(s_L)$。随后，叶节点会被完全展开，其子节点会以 $N(s_L,a)=0, W(s_L,a) = 0, Q(s_L,a) = 0, P(s_L,a) = p_a$ 初始化。在反向传播步骤中，包含 s_L 及其之前的每个节点的统计数据都会更新为： $N(s_t,a_t) = N(s_t,a_t) + 1, W(s_t,a_t) = W(s_t,a_t) + v, Q(s_t,a_t) = {}^{W}(_{s_t,a_t)}\,{}_{N(s_t,a_t)}$ [100]。

请注意，此处的扩展与评估阶段在一定程度上相互交织，并未严格遵循标准蒙特卡洛树搜索中的阶段顺序。在本综述中，我们通常不太关注蒙特卡洛树搜索阶段是否严格按顺序构成一系列算法步骤，而更关注每个阶段在树搜索中发挥的作用。

搜索预算耗尽后，根据树中的访问次数 $N(s,a)$ 得出改进后的策略 π_{MCTS}，并提取相应的价值估计 v_{MCTS}。随后，将生成的策略 π_{MCTS}和价值估计 v_{MCTS}用作训练目标，以进一步改进 f_θ。由于 AlphaZero 进行的是双人游戏，因此需要某种机制来确定第二位玩家的动作。在所谓的自我对弈中，第二位玩家的动作由当前正在为第一位玩家 [100]训练的同一策略（或其某种版本）选出。

作为一种基于模型的强化学习算法，AlphaZero 需要一个环境模型来执行树搜索。这里假设该模型已给定，不过，MuZero[89] 等进一步扩展表明，可以在搜索过程中利用收集到的经验学习这种模型。

总而言之，AlphaZero 中的神经引导包括通过使用 f_θ 计算 v 和 p来评估节点，并在选择阶段使用这些评估结果。AlphaZero 中使用神经引导的方式并非唯一可能的形式。还存在其他引导搜索的方式，正如第 5节将会阐明的那样。

3 研究方法

我们按照 [117], 中概述的方法开展系统性序贯文献综述，即按照预先定义的一系列步骤依次进行：在多个数据库中进行关键词检索，通过筛选过程筛出相关文章，进行前向与后向检索，并提取数据从所有纳入的文章中提取数据，随后对结果进行分析与综合。

虽然我们力求保持中立，因此不希望基于任意理由限制所收集的文献，但由于所需时间过长，要通过全面的文献检索囊括所有相关文献并不可行。因此，我们希望在可行性与覆盖范围之间取得平衡：将检索范围限定为，在数量有限的数据库中，使用一组明确的关键词进行关键词检索，从而收集具有代表性的现有文献样本。扩大关键词集和数据库范围，都可以获得更全面的结果。

3.1 搜索查询与数据库

为查找相关出版物，我们归纳出三类关键词：

1. 基于神经蒙特卡洛树搜索是蒙特卡洛树搜索与神经网络或传统强化学习的结合：

- “强化学习” AND “蒙特卡洛树搜索”
- “神经蒙特卡洛树搜索”
- “神经蒙特卡洛树搜索”

2. 基于蒙特卡洛树搜索在基于模型的强化学习设定中提供执行决策时规划的能力：

- “决策时规划” AND “强化学习”

3. 根据 AlphaGo 系列中算法的名称：

- AlphaGo
- AlphaZero
- MuZero

上文项目符号后列出的每个部分搜索字符串都通过 OR 运算符连接起来，从而构成一个完整的搜索查询。

我们使用此搜索查询在科学引文索引、IEEE Xplore、Scopus、ScienceDirect 和 PubMed 数据库中搜索出版物。该搜索查询应用于摘要、标题和关键词。

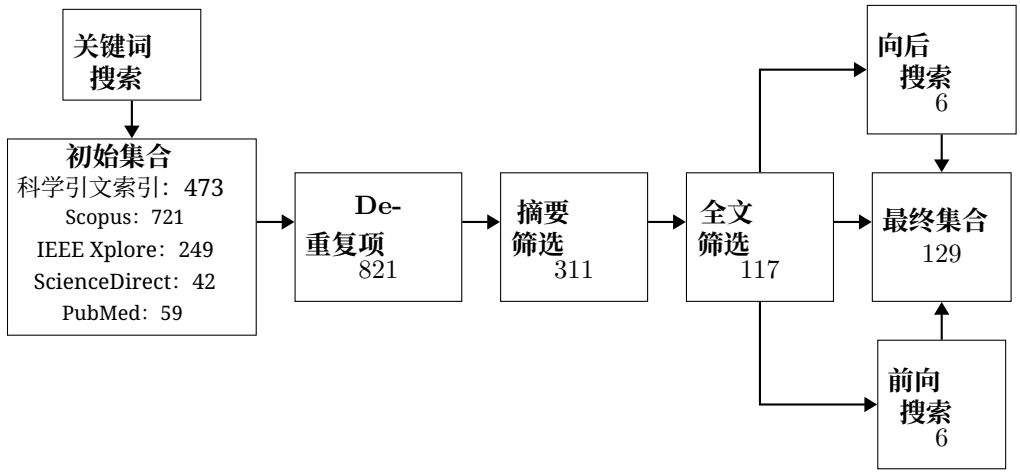


图 2 从一组关键词开始，经过文献检索和筛选，最终确定纳入综述的出版物集合。

3.2 纳入标准与筛选

要纳入综述，出版物必须满足一组预先设定的纳入标准：

1.**必须包含由神经网络引导的蒙特卡洛树搜索应用**。这不包括纯粹的综述或调查类出版物。这里所说的引导，是指在树搜索的至少一个阶段中使用学习策略或价值函数。当然，这些函数也可以通过神经网络以外的方法学习。我们在此明确只考虑基于神经网络的方法，因为初步搜索表明，若纳入其他方法，检索结果中不相关的内容会大幅增加，而额外价值却相对有限。

2.**出版物必须至少对所提出的方法进行一定程度的验证**。纯概念性文章不予考虑。

3.**神经蒙特卡洛树搜索所应用的问题不能是博弈**。尽管神经蒙特卡洛树搜索在游戏对弈中取得了许多令人瞩目的成果，但我们感兴趣的是确定这类方法能否同样迁移到其他应用中。我们也考虑那些通常不被视为博弈、但被建模为博弈以便应用神经蒙特卡洛树搜索的应用。

4. **出版物必须为英文**。

对于上一节所述关键词检索中检出的每篇出版物，我们均根据上述标准评估其是否符合条件。去重后，筛选过程分为两个阶段。第一阶段，我们只查看每篇文章的摘要；如果明确发现至少有一项上述标准未得到满足，就将其剔除。如有任何疑义，则在第二阶段重新审查该文章，并使用其全文重复这一过程。第二阶段未被剔除的文章将纳入我们的综述。

3.3 前向与后向检索

在完成上述摘要和全文筛选后，我们基于通过筛选过程的出版物集合进行前向与后向检索 [130]。也就是说，对于每篇文章，我们都会检查其参考文献，以寻找更多相关出版物；同时也会查找引用通过筛选的文章的出版物。

在这一步中，我们注意到，在额外识别出的文献中，许多文献只是引用了我们综述中已纳入的一种适用于类似应用的方法，却没有进行任何修改或提供额外细节。由于这些出版物对回答我们的研究问题没有任何帮助，因此我们不将其纳入综述。上述纳入标准同样适用于前向与后向检索的结果。

从关键词搜索到最终出版物集合的完整搜索过程如图 2 所示。

3.4 数据提取

对于上述搜索流程筛选出的每篇文章，我们都会根据一组预先定义的类别提取信息。这些类别列于表1中，其中 问题 指所研究问题的简要描述，时间 可以是连续的，也可以是离散的，时域 可以是有限的，也可以是无限的，状态转移 可以是确定性的，也可以是随机性的。

由于从术语各异、细节描述不一的不同学科的研究成果中提取信息可能并不容易，因此我们通常宁可提供不完整的信息，也不愿提供错误的信息。

部分收集到的信息并未带来值得关注的见解，因此本综述不会讨论这些内容。这包括激活函数、作者所属机构，以及神经 蒙特卡洛树搜索方法与无模型 强化学习 或蒙特卡洛树搜索（不含神经引导）之间的比较。在此类比较中，神经蒙特卡洛树搜索往往能找到质量更优的解，但这可能仅仅是正向结果发表偏倚所致。

4 神经蒙特卡洛树搜索的应用

4.1 应用领域

为确定神经蒙特卡洛树搜索在游戏之外的适用性，我们调查了神经蒙特卡洛树搜索已应用于哪些领域。第5节将分析各个方法的算法细节。此处，我们仅概述这类方法的应用领域。我们发现，神经蒙特卡洛树搜索广泛应用于化学、医学、生产、电气工程和计算机科学等领域。下文将每篇出版物归入特定的应用领域。请注意，这样做仅仅是为了概览当前的研究格局。许多文章可以归入多个类别，而类别本身也有多种划分方式。

化学

尤其是在化学文献中，神经蒙特卡洛树搜索 受到了广泛关注。它已被用于合成规划、从头分子设计、蛋白质折叠等任务（见表 2）。在许多情况下，状态由简化的分子输入线性表达系统（SMILES）字符串表示，这种表示法能够表征分子结构 [132]。随后，在蒙特卡洛树搜索过程中，可以迭代地构建此类字符串，直到找到可行分子或放弃尝试。

分子应用似乎是神经蒙特卡洛树搜索 研究中相对成熟的一个分支，证据包括作者们在彼此工作的基础上继续研究，以及已有标准化、广泛使用的实现，例如 ChemTS [145]。这只是特例，而非常态；大多数文献彼此关联较少，且其他大多数研究要么没有说明所用实现，要么使用自行开发的实现（详细的实现列表见附录A）。

材料科学

与之密切相关的是材料科学领域（见表3）。在某些情况下，这里也会使用SMILES 表示法，例如在金属有机框架的设计中 [110, 135], 金属离子和有机配体相结合，形成各种形状的结构。

在其他情况下，神经蒙特卡洛树搜索用于优化多层结构中两种材料交替层的厚度，以实现某些期望的特性 [24] ，并生成描述材料在各种情况下力学行为的模型[122]。

电子设计

在电子电路设计中，神经蒙特卡洛树搜索在多种情况下被用于解决布线问题[15, 41, 80] （见表4），其表现可能优于例如基于传统 A* 的方法 [41]。Thacker 等人 [111] 进一步探索了使用神经蒙特卡洛树搜索对存储芯片进行冗余分析。

能源系统

神经网络蒙特卡洛树搜索在能源系统运行中有许多应用，尤其是在旨在实现更可持续能源供应的创新电网概念中（见表5）。这包括优化住宅微电网的运行

表1 筛选后从每篇文章中提取的信息。

应用	神经蒙特卡洛树搜索	实现	其他
问题	选择详情	网络类型	国家
Time	扩展详情	激活函数	组织类型
奖励尺度	评估详情	策略损失	比较 蒙特卡洛树搜索
时域	自我对弈的使用	价值损失	对比无模型方法
状态转移	训练方法	MCTS 超参数	比较基线
动作空间	学习模型	代码可用性	所用硬件

表2 化学领域的应用。

来源	应用
[23]	蛋白质折叠
[25]	先导化合物发现（药物发现）
[67]	RNA 折叠
[105]	重构 分子结构 核磁共振谱
[109]	从头设计的稳定有机自由基的优化 用于水系氧化还原液流电池的
[36] [48] [90] [91] [112] [126] [148]	合成规划
[63] [65] [74] [78] [106] [145]	从头分子设计

表 3 材料科学中的应用。

来源	应用
[24]	深度渐变多层结构设计 用于X射线光学
[122]	本构模型生成
[152]	用于离子液体材料逆向设计的 二氧化碳捕集
[110] [135]	金属有机框架的设计

以在线方式 [96]进行，以及非侵入式负荷监测与识别 [50]。

此外，也有研究关注如何管理由可再生能源和电池系统组成的电网，其中电池系统可吸收前者

表 4 电子设计中的应用。

来源	应用
[15] [41] [80]	电子电路布线
[111]	存储芯片冗余分析

发电功率的波动。例如，Al-Saffar 等人 [1] 设计了一套系统，用于协调分布式能源网络中的电压调节，该网络在多个地点部署了电池系统；与此同时， [133] 他们使用神经网络蒙特卡洛树搜索来解决此类系统中的预测性维护问题。

表 5 能源系统中的应用。

来源	应用
[1]	分布式能源系统中的电压调节 系统
[50]	智能电网中的负荷监测与识别 电网格
[96]	住宅微电网调度
[107]	电气 输电 网络 自愈
[114]	人工调度智能控制系- 统在电力网络中
[133]	预测 维护 管理 电池储能系统

生产

生产系统中的应用主要涉及各种调度方法。在这些问题中，需要确定不同机器上的作业或工序的加工顺序，例如，最小化所有作业完成加工所需的总时间（见表6）。传统

强化学习方法也越来越多地用于研究这类问题 [34, 51, 86, 149]。深入考察传统强化学习方法与神经蒙特卡洛树搜索方法在调度方法中的优缺点，可能会成为未来研究的一个有趣方向。

生产领域的其他应用包括汽车制造中的生产线缓冲区规划 [37] ，以及协作式人机协作场景中的装配规划 [147]。

表6 生产系统中的应用。

来源	应用
[37]	生产线缓冲区规划
[38]	动态互联装配系-统调度
[58]	工业采矿生产调度
[81]	钣金生产调度
[121]	并行机车间调度
[147]	协同装配

组合优化

虽然上述调度问题属于组合优化领域，但作者从生产角度切入，并密切关注各自应用场景的具体细节。第二类组合优化应用见表 7。这里的问题更加抽象，并从计算机科学的视角进行研究。

组合优化问题与组合棋盘游戏有许多相似之处。在强化学习的情境下，这类问题通常通过从头开始逐步构建解来进行构造式求解，或通过改进法求解，即迭代改进某个已有解。两种情况下，问题都具有内在的离散时间步和离散动作空间。不同之处在于，这类问题没有明显的胜负概念，而是更关注相对性能。此外，组合博弈具有固定的棋盘和固定的一组棋子。例如，在旅行商问题中，与棋盘相对应的可以看作是一张由加权边连接不同城市的图。然而，这样的图会因具体问题实例而异，每个实例都包含需要经过的不同城市。在机器调度问题中，工序可以看作棋子。根据具体的问题表述，每道工序都需要在特定机器上处理特定时长。因此，工序可以描述为一种可根据时长等属性自由参数化的棋子，这与典型游戏中的棋子不同。

表 7 组合优化问题中的应用。

来源	应用
[46]	图着色
[52]	基于PBQP的寄存器分配
[75]	机器调度，车辆路径规划
[134]	装箱问题
[136] [137]	旅行商问题
[138] [139]	最高安全横档问题
[141]	量化布尔公式可满足性

云计算与边缘计算

与生产领域一样，云计算和边缘计算中的一项主要应用涉及调度问题（见表8）。同样，这些调度问题属于组合优化问题，但作者的研究兴趣源自云计算领域本身，所提出的问题也较为具体。

表 8 云计算与边缘计算中的应用 中。

来源	应用
[11] [12]	无人机边缘计算中的任务卸载 移动边缘计算
[19] [45]	有向无环图任务调度
[77]	云工作流调度

图导航

从给定节点导航至目标节点是许多场景中的一项相关任务，但近年来在知识图谱研究中尤其受到关注。在此，知识图谱补全是一项常见任务，涉及预测各个实体之间缺失的关系[128]。图导航是知识图谱补全的一项重要子任务 [95]„ 神经蒙特卡洛树搜索已针对该任务进行研究（见表9），且已证明其性能优于现有基线方法 [95]。

表 9 图上的应用。

来源	应用
[40]	知识图导航
[95]	（知识）图导航
[123]	图导航
[128]	知识图谱补全

网络与通信

网络与通信领域的应用涵盖网络功能虚拟化 [30, 57]，、网络拓扑优化 [69, 120, 144, 146, 159, 160]„ 以及采用多种无线接入技术的移动网络中的频谱共享[10, 142]。

一个值得注意的例子是，在软件定义网络场景中使用神经网络 蒙特卡洛树搜索 进行入侵防御 [32]。在此，防御问题实际上被建模为双人游戏；在所调查的文献中，大多数场景都是单人场景，因此这是一个例外。

自动驾驶与运动规划

自动驾驶应用构成了一个相对较大的类别（表 11），其中包括一般运动规划任务 [39, 61, 76, 131], 自动泊车场景中的运动规划任务 [104, 150], 以及多智能体场景中的运动规划任务 [82, 103]。此外，还研究了更专业的任务，例如车道保持 [56],超车[70]，以及自动驾驶过程中的高层决策 [44] 。

这类问题在根本上往往不同于组合博弈。例如，Weingertner 等人 [131] 考虑了一个运动规划问题，其中某物体的加速度

表10 网络与通信中的应用。

来源	应用
[30]	网络功能虚拟化中的服务功能链部署问
[32]	网络功能虚拟化中的问题 软件定义网络中的入侵防御- 正在运行
[57]	网络功能虚拟化映射 与调度
[118]	虚拟网络嵌入
[127]	开放式 RAN 网络中的资源分配 网络
[69] [120] [144] [159] [160] [146]	移动自组网布局优化 网络
[10]	长期演进和新空口的动态频谱共享
[87]	无线资源调度
[92]	无线通信调度
[142]	多无线接入技术接入
[71]	多输入多输出检测
[13]	多输入多输出系统的导频功率分配

车辆沿预先确定的路径行驶。在其自然表述中，这类问题需要在连续时间内选择连续动作。要应用神经网络蒙特卡洛树搜索，需要将时间和动作空间离散化。由此得到的解表现良好，优于 A*搜索、纯深度学习方法和模型预测控制。

表11 自动驾驶中的应用，以及路径与运动规划。

来源	应用
[104] [150]	自动泊车中的运动规划
[39] [61] [76] [131]	自动驾驶中的运动规划
[14]	驾驶操作预测
[44]	战术决策
[56]	车道保持任务
[70]	超车任务
[82]	多机器人运动与路径规划
[103]	多车辆运动与路径规划

自然语言处理

表12展示了对话代理的若干应用：智能体协助用户完成任务 [125]、与用户协商以分配给定的一组资源 [49]，，以及尝试通过不同方式措辞来使用户相信某种观点 [9]。虽然难以明确模拟人类作为对话伙伴的行为，但可以利用历史数据训练能够近似人类特定对话行为的模型，并将其用作树搜索的一部分[9, 125]。

自然语言处理本身就是一个多元化领域，其中包括情感分析等主题[20] 和命名实体识别 [59] 正借助神经蒙特卡洛树搜索进行研究。

表1 2 自然语言处理中的应用。

来源	应用
[9]	个性化消息
[49]	协商对话
[125]	任务完成对话
[20]	情感分析
[42]	文本匹配
[59]	命名实体识别

机器学习

蒙特卡洛树搜索 受机器学习模型引导，反过来也可用于某些机器学习任务（参见表 13）。例如 [47] 和 [129]分别应用神经 蒙特卡洛树搜索，通过网络蒸馏缩小前者中的神经网络规模，并在后者中进行 卷积神经网络（CNN）滤波器剪枝。

Lu等人 [64]进一步采用神经 蒙特卡洛树搜索 来处理符号回归任务。在这里，与通过调整例如线性函数或多项式函数的系数来解决回归任务不同，研究者会确定函数本身的各项（例如正弦函数、平方运算、常数），并通过加法和除法等数学运算符将它们连接起来。在 蒙特卡洛树搜索中，函数的完整表达式可以逐步构建。

计算机科学

计算机科学为神经蒙特卡洛树搜索(的应用提供了广泛的机会（参见

表13 机器学习中的应用。

来源	应用
[47]	神经网络蒸馏
[129]	卷积神经网络滤波器剪枝
[17]	高成本特征分类
[64]	符号回归
[158]	推荐系统

表14），其中许多应用都在单独的章节中介绍。其他应用由于其特定细分领域的出版物数量较少，不足以单独成章，但仍然颇具趣味。适用这一情况的出版物数量，恰恰体现了神经蒙特卡洛树搜索的广泛适用性。

一个值得注意的例子是 AlphaTensor [26]，其中神经网络蒙特卡洛树搜索被用于寻找高效的矩阵乘法算法。其他应用包括优化数据库查询 [151]，、恢复稀疏信号 [18, 155]，以及量子计算中的各种应用 [16, 22, 102]。

表14 计算机科学中的应用。

来源	应用
[26]	矩阵乘法算法发现
[94]	序列发现
[140]	模型检测
[151]	数据库查询优化
[153]	低密度奇偶校验码构造-tion
[18] [155]	稀疏信号恢复
[35] [156] [157]	自动定理证明
[16]	量子退火调度优化
[22]	量子动力学优化
[102]	量子比特路由

最后，表15展示了其他各类领域中的应用，这些应用不属于前面任何类别。所涉及的问题多种多样，包括用户界面的优化 [98, 113]，气动执行器的控制[55]，以及桁架的设计任务 [80] 和流体结构 [3]。正如这里的许多示例所示，类似的设计任务也正在被研究

表15 其他领域的应用。

来源	应用
[143]	主动空间多碎片清除
[154]	蜂窝卫星配置
[31]	认知雷达任务调度
[93]	
[5]	物体重排
[119]	场景布置规划
[98]	用户界面优化
[113]	
[84]	放射治疗射束方向选择
[85]	
[3]	流固拓扑优化
[80]	桁架设计
[21]	移动群智感知
[27]	电磁态势分析
[28]	机器人操作
[33]	野火蔓延预测
[55]	气动执行器控制
[62]	文档风格逆向工程
[124]	铁路时刻表重新调度

与传统强化学习 [29]。在未来的研究中，针对特定问题直接比较传统强化学习与神经蒙特卡洛树搜索方法，或许有助于确定在何种情况下哪种方法更可取。

总之，以上各节所述的应用源自多个不同学科，包括化学、医学、计算机科学、数学和电气工程。问题类型包括各种优化任务、控制问题、生成式设计任务以及许多其他类型。显然，神经蒙特卡洛树搜索不仅广泛适用于组合博弈，也适用于部分具备、部分不具备博弈特性的问题。

4.2 应用特征

与博弈类似，此处综述的问题可以表述为 马尔可夫决策过程。组合博弈涉及在有限时域设置下的离散时间步中选择离散动作，即博弈是具有明确定义终止条件的回合制过程。奖励通常较为稀疏，对应少数几种可能的游戏结果：输（-1）、平局（0）和赢（1）。虽然每位玩家各自的状态转移通常是确定性的，但第二位玩家的存在会使下一回合将遇到的状态变得不确定。尽管此处综述的许多应用具有上述诸多属性，神经网络 蒙特卡洛树搜索 也应用于在一个或多个方面与组合博弈不同的应用。

Time

许多场景并非具有回合制特性，而是允许在任意时间点执行动作，也就是说，时间往往具有连续特性。这似乎并未妨碍神经蒙特卡洛树搜索的应用，因为许多作者只是在问题表述中将时间维度离散化 [5, 28, 33, 39, 47, 55, 56, 59, 76, 104, 131, 154]。

有限与无限时域

与组合博弈类似，这里综述的大多数应用都包含有限时域问题，即问题在有限长度的回合中求解。在某些情况下，问题的自然表述具有无限时域。为了应用神经网络蒙特卡洛树搜索，可以通过设置最大步数来人为创建回合；达到该步数后，回合总是终止，例如 [58, 96, 113]。

状态转移

许多综述中的问题都具有完全确定性，这与组合博弈领域存在根本区别。在这类情况下，可以调整树搜索以利用确定性转移（详情请参见第5.3节）。

尽管如此，仍可观察到一些具有随机状态转移的问题表述 [32,38, 58, 69, 92, 113, 125, 142]。

奖励

典型问题设置中的奖励结构往往不像博弈中的奖励函数那么简单。奖励通常不是取自仅包含两三个不同奖励值的集合，而是对应于所得解质量的连续统。很多时候，奖励甚至没有明确的上下界，或

两侧（例如参见 [5, 45, 67, 90, 96, 125, 147,154]）。

在某些情况下，奖励会通过受自我对弈启发的机制进行转换。我们将在第 5.2 节中对此进行更详细的探讨。

尽管所调查的问题大多在回合结束时具有某种稀疏奖励，但在某些情况下，树搜索中还会纳入每个动作之后更细粒度的奖励 [26, 44, 105]。

动作空间

蒙特卡洛树搜索 天然适用于离散动作空间，例如组合棋盘游戏中的情况。尽管存在针对连续动作空间的（神经网络）蒙特卡洛树搜索改进方法 [72]，但此处调查的绝大多数应用都采用离散动作空间。值得注意的例外包括 Lei 等人提出的方法 [61] 以及 Paxton 等人提出的方法。 [76]

此外，Raina 等人 [80]采用了一种分层强化学习方法，其中神经网络蒙特卡洛树搜索用于一组高层离散动作，而后续的连续动作则由另一种机制确定。

最后，也可以将天然连续的动作空间离散化。虽然这会降低动作选择的精度，但某些应用仍然可以通过这种方式得到成功处理 [26, 96]。

状态空间

虽然状态空间的具体特征不仅取决于底层问题，也取决于问题的建模方式，但围棋等博弈具有定义明确且规则的棋盘，这有助于构建状态空间。围棋棋盘大小固定，由空间位置相互关联的格子组成，因此很适合使用卷积神经网络进行处理。本文综述的问题涵盖了多种状态空间，并由不同类型的神经网络进行处理。除卷积神经网络外，所用的神经网络还包括递归神经网络，例如长短期记忆网络 [5, 25, 28, 35, 40, 42, 46, 96, 113, 119] 和门控循环单元 [20, 33, 49, 59, 65, 106, 110, 122, 123, 141, 145, 158]，以及图神经网络 [48, 52, 61, 63, 75, 77, 102, 105, 109, 123, 136, 137, 141, 157]。较少见的类型包括 Transformer [26, 78, 87]和 DeepSet 架构 [146]。在许多情况下，简单的多层感知机就足够了 [10–13, 16, 22, 32, 36– 38,41, 45,55, 56,58,76, 81, 90, 91,97,112, 125,126, 131, 140, 150, 151, 155]。

所选架构及其深度在一定程度上决定了训练神经网络蒙特卡洛树搜索 方法所需的硬件类型。

4.3 硬件要求

AlphaZero 的训练使用了 5000 多个张量处理单元 [100]。因此，人们或许会质疑，对于无法获得如此规模资源的研究人员和从业者而言，神经网络蒙特卡洛树搜索是否是一种可行的选择。

我们综述的部分出版物确实报告了对大量资源的使用。例如，Huang 等人 [46] 在训练期间最多使用了 300 个 NVIDIA 1080Ti 和 2080Ti 图形处理器。不过，报告中大多数硬件的配置，对一般机构乃至个人而言都并非遥不可及。Genheden 等人 [36] 报告称，他们使用一颗 Intel Xeon CPU 和一块 NVIDIA 2080Ti 图形处理器，运行于配备 64 GB 内存的机器上。还有许多研究使用一颗高端消费级 CPU 和一块图形处理器 [57, 58, 84, 94, 155]。

在低端配置中，甚至有些研究人员会使用消费级笔记本电脑来训练神经网络蒙特卡洛树搜索方法 [45, 141, 143]。

显然，硬件要求因应用以及所用神经网络的复杂程度而异。虽然很难预测特定应用及其所需解的质量需要何种硬件配置，但可以确定的是，在许多应用中，中等性能的硬件也能得到有效利用。

5 神经蒙特卡洛树搜索方法

在上一节概述了神经蒙特卡洛树搜索可能应用的广泛范围之后，我们现在将注意力转向综述中遇到的神经蒙特卡洛树搜索 方法的设计。

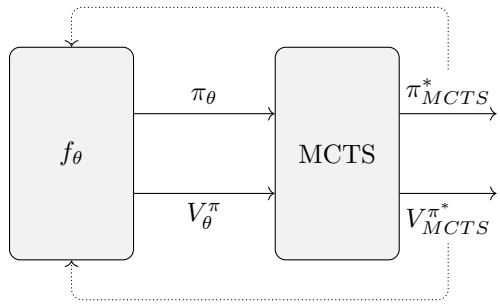


图 3 蒙特卡洛树搜索作为策略改进算子。学习策略和价值函数用于引导树搜索，进而生成改进后的策略 π^*_{MCTS}和价值估计 $V^{\pi^*}_{MCTS}$，适用于给定状态。正如虚线所示，π^*_{MCTS} 和 $V^{\pi^*}_{MCTS}$ 随后也可用作神经网络的训练目标。

5.1 引导网络训练

在深入探讨神经蒙特卡洛树搜索的内在机制之前，我们只需了解学习策略网络和价值网络会以某种方式引导树搜索。接下来，我们先着重介绍 AlphaZero [100],所采用的训练流程，然后再讨论综述中发现的其他方案。

通过蒙特卡洛树搜索进行策略改进

在 AlphaZero [100], 中，通过引导蒙特卡洛树搜索（MCTS）来迭代改进学习策略，再利用搜索结果进一步改进学习策略（见图 3）。我们将这一过程称为通过蒙特卡洛树搜索进行策略改进。具体而言，学习策略 π_θ 在一个或多个搜索阶段中引导树搜索；在经过给定次数的蒙特卡洛树搜索模拟后，便可得到针对当前状态的新策略 π_{MCTS} 。由于这一新策略通常强于最初学得的策略，因此可以将其作为策略网络的训练目标。更确切地说，对于给定状态，策略网络和蒙特卡洛树搜索分别生成策略概率向量 $\mathbf{p}_\theta$ 和 $\mathbf{p}_{MCTS}$ ，前者可视为实际预测，后者则是预测目标。随后可以使用交叉熵损失函数来训练策略网络：$L_{CE} = -\mathbf{p}^T_{MCTS} \log \mathbf{p}_\theta$。相应地，如果在学习策略的同时也学习价值函数，则可使用均方误差（MSE）作为损失函数，使其价值估计朝蒙特卡洛树搜索所得的价值估计方向调整：$L_{MSE} = (v_{MCTS} - v)^2$。两者通常会与正则化项结合，构成单一损失函数

$$L = (v_{MCTS} - v)^2 - \mathbf{p}^T_{MCTS} \log \mathbf{p}_\theta + c \|\theta\|^2 \quad (3)$$

我们综述的文章中，绝大多数通过蒙特卡洛树搜索改进学习策略时，使用的是式（3）给出的损失函数。我们发现有两种例外：其中一种使用 Kullback–Leibler散度而非交叉熵 [17]；另一种也使用 Kullback–Leibler散度而非交叉熵，但还额外使用分位数回归分布损失来代替 MSE [26]。值得研究这些损失函数对训练的影响，但在我们的综述中，交叉熵与MSE的组合显然是默认选择。

替代训练方法

我们在综述中发现，训练方法大致分为两类：（1）在使用网络引导树搜索之前进行训练；（2）在树搜索过程中进行训练，即迭代执行树搜索，并利用其结果进行训练。对于第（1）类，训练可以通过对标注样本进行监督学习，或通过策略型、价值型和演员-评论家类的经典强化学习算法来实现。对于第（2）类，主流方法是通过蒙特卡洛树搜索进行策略改进，但也存在一些变体。最后，可以将两类方法结合起来：先进行可视为预训练的训练，然后在树搜索过程中继续训练。这有时称为热启动。

图4 展示了所调查出版物中这些不同方法的分布，以及每种方法所采用的具体技术。许多作者在搜索过程中使用蒙特卡洛树搜索进行策略改进，以训练其网络。在某些情况下，搜索过程中的训练会采用其他方法，例如 Q学习 [1, 37, 95, 123]和最大熵强化学习 [157] ，作用于蒙特卡洛树搜索轨迹。除了通过蒙特卡洛树搜索改进策略（以及价值函数）外，也可以单独细化价值函数 [62, 137], ，而不使用任何学习策略。我们将这种方法称为通过蒙特卡洛树搜索细化价值函数。同样，虽然

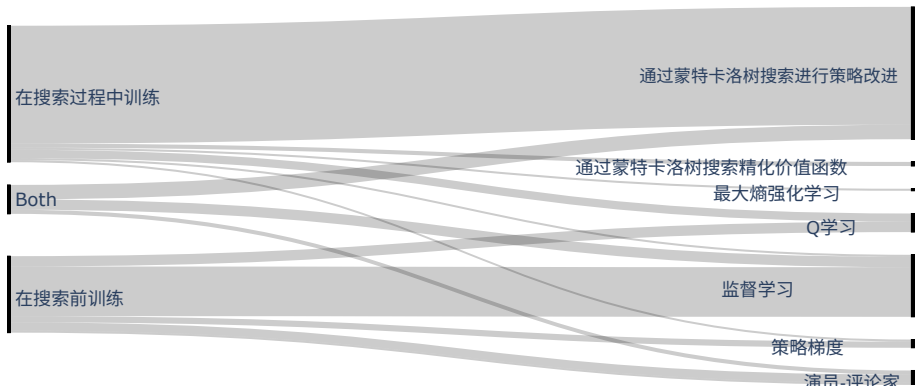


图 4 根据所采用的方法，策略网络和价值网络可以在搜索前、搜索过程中或两者兼有时进行训练。根据这一选择，会采用不同的训练方法。策略梯度和演员-评论家是涵盖多种具体算法的算法类别。

通过蒙特卡洛树搜索 进行策略改进通常会同时训练策略和价值函数，有时也会单独训练策略 [93]。

在树搜索之前进行训练时，网络通常通过对带标签的示范进行监督学习来训练。也会采用Q学习 [75, 76, 119,125, 131]，策略梯度 [24, 45, 55, 57, 64, 77]以及演员-评论家方法 [17, 41, 121, 129] 。

当两个训练阶段都进行时，最常见的方法是将监督式预训练与通过蒙特卡洛树搜索 [16, 81, 107, 126, 154]进行的策略改进相结合。

总体而言，所采用的训练方法有所不同，但主流策略是在搜索之前进行监督训练，以及通过蒙特卡洛树搜索进行策略改进；有时也会像 AlphaGo 的原始论文中那样，将二者结合在一种方法中 [99]。在组合博弈场景中，通过蒙特卡洛树搜索进行策略改进，需要某种机制来生成对手的走法。下文将讨论这种机制及其对博弈之外应用的意义。

5.2 博弈之外的自我对弈

促成 AlphaGo 成功的组件之一是自我对弈这一概念 [99]。在自我对弈的设置中，多人博弈中的对手由相同策略（或其某种变体）控制，也就是说，策略会与自身对弈 [43]。在这种情境下进行学习的优势在于，随着训练的推进，策略始终面对技能水平相当的对手，而对手也会随之不断进化。然而，由于许多非游戏应用本质上都是单人性质的，因此自我对弈在游戏之外的作用并不明显。

为了更清楚地了解自我对弈在这类情形中的适用性，我们调查了综述所收录出版物中的相关用法。在此过程中，我们发现许多作者都使用“自我对弈”一词，但其含义各不相同。这可能是因为尚缺乏公认的标准定义。下文将首先划分我们遇到的该术语的不同含义，然后报告综述中不同形式自我对弈的用法。

在双人设定中，“自我对弈”一词的含义直观易懂。按照最初的 AlphaGo 出版物中的用法 [99], ，自我对弈是指当前正在学习的策略，在双人轮流制设定下与自身的某个旧版本进行博弈。这意味着，当前策略的另一个版本会在博弈中每隔一回合行动一次，以此生成新的状态，并获取该博弈的最终奖励。

在单人设置中，也经常使用“自我对弈”一词，但其含义不那么明显。在单人设置中，上文所述的状态生成属性并不适用，因为状态转移不取决于另一位玩家的动作。生成新状态只需要当前状态和智能体的动作。然而，如果奖励函数经过相应设计，上文所述的奖励生成属性仍然适用。如果奖励并非仅取决于当前策略的性能，而是取决于相对性能的比较

将自我对弈的奖励生成属性从某个先前版本的策略转移到单人设置中。换言之，策略试图打破自身最高分的过程，与双人设置中的自我对弈概念有相似之处。虽然有时也将其称为自我对弈，但Mandhane等人 [66]提出了自我竞争 这一术语来指代这类方法。在本综述的其余部分，我们将采用这一术语，并将自我对弈 专用于多人设置，以避免混淆。尽管简单版本的自我竞争很容易实现，但 Laterre等人 [60]提出了一种更有依据的自我竞争形式，称为排名奖励，随后 Mandhane等人 [66] 和 Schmidt 等人 [88] 也提出了相关方法。

然而，许多作者声称采用了自我对弈，却没有明显应用上述任何概念（例如，参见 [80, 111]）。虽然很难确定这些情况下所指的含义，但我们怀疑，文献中有时还会用自我对弈来称呼另外两种做法。第一种做法是，通过将当前策略与先前的最佳策略进行评估，持续记录最佳策略。如果当前策略在某个预先定义的问题集合上表现优于先前的最佳策略，就用它替换当前保存的最佳策略。这样做的目的仅仅是在训练结束后能够获取最佳策略，因为训练并不一定会使策略性能单调提升。在这种情况下，评估结果不会被用作训练策略的奖励。因此，策略与自身的另一个版本对弈并不会带来学习；换言之，这并非一种改进当前策略的机制，而仅仅是对其进行评估。

最后一个概念，我们怀疑有时会将其称为自我对弈，指的是通过蒙特卡洛树搜索（如上所述）进行策略改进（例如参见 [12]）。

需要说明的是，我们只将“自我对弈”用于描述满足状态生成属性和奖励生成属性的多玩家情形，而“自我竞争”仅用于描述满足奖励生成属性的单玩家情形。

虽然理想情况下，我们希望报告综述所纳入的所有出版物中自我对弈和自我竞争的使用情况，但当这些术语的含义模糊不清时，我们不会这样做，而是仅报告一些具有代表性的例子，其中这些术语的含义已得到明确界定。

自我对弈

在非游戏领域文献中，真正的自我对弈似乎相当少见。我们只能将自我对弈的使用归于一项研究 [27],: 该研究将一个问题建模为双人游戏，并通过自我对弈学习策略。在其他一些情形中，问题同样被建模为双人游戏，但由此产生的游戏是不对称的，即玩家拥有不同的动作空间，因此需要不同的策略 [32, 138–141]。在这些情形中，两个不同的神经网络分别学习一种策略。

自我竞争

在自我竞争方面，我们观察到排名奖励的实例 [109, 123, 127] 以及一些朴素方法（见表16）。在一种朴素方法中，只需将当前策略在当前问题上的性能评估为某个分数，并与截至目前在同一问题上观察到的最佳策略分数进行比较。如果当前策略的分数更高，则赢得博弈（$r = 1$）；如果更低，则输掉博弈（$r = -1$）；如果相同，则为平局（$r = 0$） [46, 52]。另一种变体是不使用最佳策略，而是与一组已保存策略的平均分数进行比较 [122]。

在一个案例中，采用了上述朴素方法，但不是使用策略的过去版本，而是另外学习一个完全独立的策略，让两个策略持续相互竞争 [3]。

虽然可以利用自我竞争，根据策略的相对性能来生成奖励，但这并非必需，因为同样可以根据绝对性能计算奖励。自我竞争的一个好处或许仅仅是让奖励处于一个明确定义的范围内 $[-1, 1]$ 或类似范围，因为蒙特卡洛树搜索超参数的最优选择取决于该范围 [8]。然而，自我竞争似乎还有其他益处，因为已有研究表明，排名奖励方法的表现优于使用范围 $[0, 1]$ [60]内标准奖励训练的智能体。尚不清楚朴素的自我竞争方法是否也有这一优势。

表 16 自我竞争

Type	出版物
排名奖励	[109], [123], [127]
朴素（最佳）	[46], [52]
朴素（平均）	[122]

5.3 引导式选择

前文各节论证了蒙特卡洛树搜索具有策略改进算子的作用。现在我们来探讨这种策略改进的机制，即神经蒙特卡洛树搜索的内部工作原理。一次蒙特卡洛树搜索搜索迭代始于选择阶段：从根状态出发，逐步选择动作，直至遇到叶节点。如第2.3节所述，动作的选择由树策略决定，其通常形式如下

$$a = \arg\max_a Q(s,a) + U(s,a) \tag{4}$$

其中 $Q(s,a)$ 鼓励利用已知的高价值动作，而 $U(s,a)$ 则鼓励探索搜索树。具体形式既有对式（4）的不同表述，也有对方程中各项的不同确定方式，即通过学习策略和价值函数，或采用传统方法。以下将分别探讨这些方面。

树策略形式

综述中提到的树策略通常基于 UCT规则的某种版本或扩展，但规则的具体形式，尤其是探索部分，存在一些差异。

我们在表1 17中概述了综述期间发现的 $U(s,a)$的各种变体。编制该表时，我们对各篇出版物报告的具体公式进行了修改，以统一记号。为此，我们假定所有报告的对数均为自然对数，并且 $N(s) = \sum_b N(s,b)$，即 $N(s)$指状态s中所有子节点的访问次数，而 $N(s,a)$指状态 s中动作a的访问次数。在本综述中，探索常数有时记作 c_{uct}、c_{puct} 或类似形式，我们统一称为 c。$P(s,a)$表示在状态 s下选择动作 a 的某种先验概率，无论

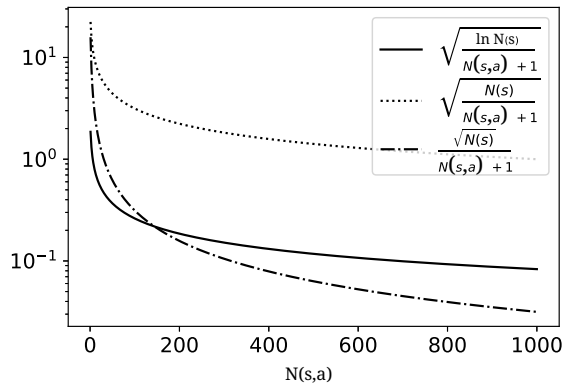


图 5 不同的 UCT 风格分式，$N($ $s =$ 1000 $)$ 固定。请注意，纵轴采用对数刻度，显示图例中的表达式在不同 $N(s,a)$ 下的取值。

它是由学习策略给出的，还是通过其他方式获得的。

在所调查的出版物中，仍有相当大一部分采用（某种变体的）UCT公式（见表 17），但 PUCT 在 AlphaZero 中的用法 [100]（PUCT 表 17中的变体 0）是使用最频繁的选择机制。还有一些较少使用的PUCT变体，主要涉及是否使用对数项、分子和分母中的常数因子，以及平方根的作用范围。这些差异会影响探索项的整体幅度，以及随着各个动作被访问次数增加，其衰减的方式（见图 5）。由于作者通常不会直接比较不同形式，因此很难判断它们对搜索的影响。在一个罕见的例子中，Xu 和 Lieberherr [138] 同时测试了 AlphaZero 的 PUCT 变体和表 17 中的 PUCT 变体 9，并报告称 AlphaZero 变体的表现要好得多，但未量化这一差异。

一个值得注意的 PUCT变体，即变体 3，引入了一个新常数 μ，该常数决定先验概率的影响程度，表示为 $P(s,a)^{\mu}$。这一变体似乎是独立提出的，见 [95, 104, 123]。

除了 UCT和 PUCT的变体外，MuZero [89] 选择公式或其变体被三位作者采用。我们还发现了两种典型选择公式的独特改动，无法归入其他任何类别：UCT_D和 $PUCT_B$。前者将在后文讨论。$PUCT_B$ 旨在利用

表17 基于UCT的树策略变体。两条水平线将不同组的选择公式分隔开，每组中的每个成员都有一个变体编号。每组的第一个成员始终是最常见的变体，但不一定是该组最初的公式。

频率	Name	变体	公式
27	UCT	0	$c\sqrt{\frac{2\ lnN(s)\ (,a)}{N\ s}}$
2		1	$c\sqrt{\frac{lnN(s)}{1+N}}$ (s,a)
1		2	$c\ \frac{\sqrt{N(s)}}{1+N(s,\ a)}$
39	$PUCT$	0	$c\ P(s,\ \ a)\frac{\sqrt{N(s)}}{1+N(s,\ a)}$
5		1	$c\ P(s,\ \ a)\sqrt{\frac{2\ lnN}{N\ s}(s)\ (,\ a)}$
3		2	$c\ P(s,\ \ a)\frac{\sqrt{N(s)}}{N(s,a)}$
3		3	$c\ P(s,a)^{\mu}\sqrt{\frac{N}{1+N}(s)\ (s,a)}$
2		4	$c\ \frac{P(s,a)}{1+N(s,a)}$
2		5	$c\ P(s,a)\ \frac{\sqrt{N(s)+1}}{N(s,\ a)+1}$
2		6	$c\ P(s,a)\ \frac{N(s)}{1+N(s,a)}$
1		7	$c\ P(s,a)\ \frac{\sqrt{N(s)+\epsilon}}{N(s,\ a)+1}$
1		8	$c\ P\ (s,\ \ a)\ \sqrt{\frac{lnN(s)+1}{1+N\ s\ a}(,)}$
1		9	$c\ P(s,a)\sqrt{\frac{N}{1+N}(s)\ (s,a)}$
1		10	$c\ P(s,a)\sqrt{\frac{lnN}{1+N}(s)\ (s,}$ $a)$
2	$MuZero$	0	$P(s,a)\ \frac{\sqrt{N(s)}}{1+N\ s\ a}(,)\left[c_1+ln\ \frac{N(s)+c_2+1}{c_2}\right]$
1		1	$P\ (s,\ \ a)\ \frac{\sqrt{N(s)}}{1+N\ s\ a}(,)\left[c_1+ln\ \frac{\sqrt{N(s)+c_2+1}}{c_2}\right]$
1	UCT_D	0	$c\ A(s,\ \ a)\sqrt{\frac{2\ ln\ N}{N\ s}(s)\ (,\ a)}$
1	$PUCT_B$	0	$c_1\ P(s,a)\ \frac{\sqrt{N(s)}}{1+N(s,\ a)}+c_2B(s,\ \ a)$

确定性单人场景的特点是，未来轨迹不受另一位玩家的选择影响。在这种情况下，与其仅仅考察平均状态价值，不如.....

在搜索过程中，跟踪遇到的最佳价值可能会更有优势。基于平均值做出决策存在问题，因为给定状态下的大多数动作可能都是糟糕的选择，而某一个特定的单独动作却可能是

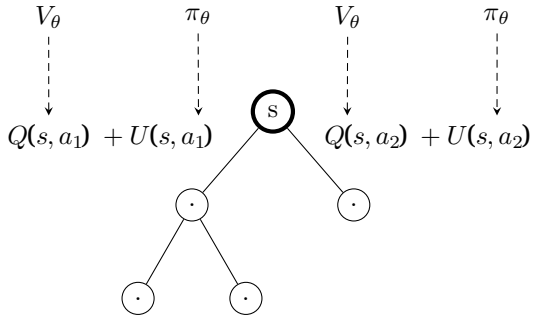


图 6 神经选择。对当前状态的每个子节点 s 进行考量，并选择使 $Q(s, a) + U(s, a)$ 最大的子节点。神经引导可能会以某种方式影响 $Q(s, a)$ 和 $U(s, a)$。

不错的选择。这样一来，节点的平均价值就会很低，尽管存在一个有前景的子节点。在确定性环境中，可以可靠地执行最佳路径，因此，根据节点预期的最佳价值而非平均价值来选择节点是合理的。Deng 等人 [23] 设计了一种利用这一事实的选择公式，我们在此将其称为 $PUCT_B$。在 $PUCT_B$中，动作的最佳价值只需乘以一个常数，再加到 PUCT 变体 0 上。

树策略中的神经引导

神经引导可用于公式（4）（见图 6）中利用和探索的部分。在利用部分使用时，神经引导通常用于估计 $Q(s, a)$。这不会改变选择机制的工作方式，只会影响相应价值的确定方式。由于价值估计是评估步骤的函数，这类神经引导将在第 5.5 节中探讨，此处不再进一步讨论。

在式（4）的探索部分，通常使用神经引导来确定 $P(s, a)$ 中的先验概率，并将其用于 PUCT 风格的公式中。如图 7 所示，在所有综述文章中，约 62% 报告以这种方式引导树搜索，不到 30% 报告了不使用神经引导的选择阶段。其余文章则完全没有说明如何执行选择阶段。虽然后一类情况很可能可以解释为不使用神经引导的选择，但我们尽可能避免作出推断，因此将标准选择和未报告的选择分别归类。



图 7 受调查文章中选择阶段各种选择的比例。 *Standard* 指不涉及使用学习函数的某种选择策略。

大多数采用神经引导选择阶段的方法都采用上述形式，但也有少数特殊情况。Zombori 等人 [157] 认为，即使学习策略网络的预测质量较低，它往往仍会给出置信度很高的预测，从而在搜索中造成强烈且缺乏依据的偏差。如果预测质量不够高，策略给出的预测置信度较低或许更理想。为此，他们采用最大熵强化学习，并使用由此得到的策略来计算选择阶段的先验概率。

有一种例外情况，神经引导并非通过提供先验概率来实现，如表 UCT_D 中的公式 17所示。这种方法因使用决斗网络而得名，该网络除了生成状态价值外，还会生成动作优势估计 $A(s, a)$。在 UCT_D中，动作优势取代先验概率 $P(s, a)$，用于搜索。作者的观点与 Zombori 等人的推理略有关联：他们认为，使用策略梯度方法训练的策略网络往往会专注于给定状态下的最佳动作，却不会按照其他动作预期效用的比例分配概率 [125]。换言之，熵过低的策略向量可能会在不理想的程度上使搜索产生偏差。相比之下，动作优势不会过度聚焦于最佳动作。

5.4 引导扩展

当在 s_L 蒙特卡洛树搜索的选择阶段遇到叶节点时，蒙特卡洛树搜索会执行扩展步骤，以创建 s_L 的子节点。此步骤也可以采用神经网络进行引导，从而偏向某些方向并加快搜索。为避免混淆，我们将先介绍扩展步骤的几种可能替代实现方式，然后再回到神经引导这一主题。

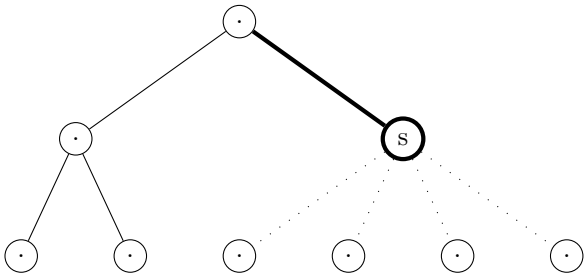


图 8 神经扩展。当遇到叶节点时，通过某种涉及已学习策略的机制，对叶节点状态下可能采取的动作进行采样。对于每个采样动作，都会创建一个新的子节点。

如第2.3节所述，通常在遇到可扩展节点 s_E 时逐个扩展节点；或者在遇到叶节点 s_L 时同时扩展该节点的所有子节点。

显然，实现蒙特卡洛树搜索 方法需要决定在给定时间扩展多少个子节点。不过，还需要做出另一个相关决定：哪些子节点以及多少个子节点会被纳入考虑 以供扩展？在最简单的情况下，搜索可以从状态 $A(s_L$ 中的所有可能动作集合 s_L 中自由选择任意动作。不过，也可以在扩展步骤中进行更有选择性的筛选。为了限制树的增长，可以随机选择有限数量的子节点，或根据某种规则或启发式方法选择子节点以供扩展。换言之，搜索可以被限制为仅从一个˜集合 $A(s_L) \subset A(s_L$ 中选择动作。这在连续情形下尤其重要，因为潜在子节点的数量是无限的，因此必须以某种方式加以限˜制。一旦定义了这样的集合 $A(s_L)$，在搜索过程中遇到叶节点时，可以一次性扩展相应的子节点；也可以在树搜索过程中每当遇到可扩展节点时，逐个扩展这些子节点。

在扩展步骤中，两种范式都可以采用神经引导：既可以在遇到真正的叶节点时进行扩展，也可以在遇到可扩展节点时进行扩展。前一种情况下，神经引导意味着使用˜学习策略来确定 $A(s_L)$；后一种情况下，神经引导意味着在 $A(s_E)$中选择一个动作，以创建新的子节点。理论上，可以结合这两种˜方法：首先确定并保存 $A(s_L)$



图 9 受调查文章中扩展阶段各项选择所占的比例。 *Standard* 指某种不涉及使用神经网络的扩展形式。

首次遇到叶节点时，但此时不扩展所有相应的节点。此后每次再次遇到该节点时，都可以通过选择某个动作 $a \in A(s_E)$，逐一扩展其子节点˜。不过，我们在收集的文献中没有发现这种组合方法。

我们确实在相当一部分出版物中观察到某种形式的神经引导扩展（见图 9），并在表 18中对其进行了分类。在其他一些示例中也使用了神经引导扩展，但未报告具体细节 [25, 36, 90]。

˜$A(s_L)$可以通过从学习策略中随机采样动作来确定，也可以通过枚举策略分布中的所有动作并选择排名靠前的 k 个动作来确定[48, 91, 112]。在 Thakkar 等人[112], 的方法中，选择累计策略概率达到 0.995 的排名靠前的 k 个动作，或最多 50 个动作。在这两类神经引导扩展中，当然也可以将学习得到的价值函数通过 softmax 算子转换为策略，而非使用学习策略，如 [119]中所采用的方法。

表18 神经引导扩展的类型。

Type	出版物
选择 $a \in A(s_E)$	[5] [24] [45] [119] [143] [64]
确定 $\tilde{A}(s_L)$	[26] [48] [61] [80] [82] [91] [112] [145] [148]

尽管神经引导扩展的确切影响会因应用而异，但[82] 报告称，采用神经扩展后，计算时间缩短了 20 倍，同时获得了质量更高的解。

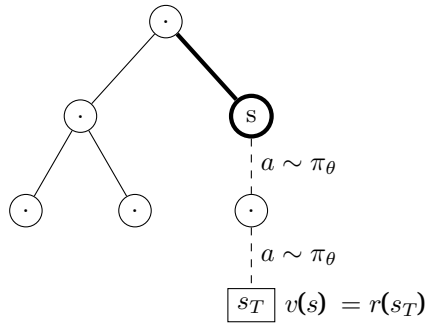


图10 通过学习策略滚动模拟进行评估：根据树策略到达状态为 s 的叶节点后，需要确定 s 的价值。此处使用学习策略 π_θ，通过迭代采样动作生成滚动模拟，直至到达终止状态 s_T。该终止状态的奖励可作为 s价值的估计。

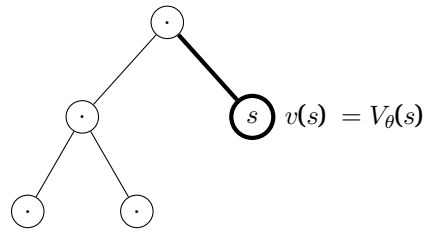


图11 通过学习得到的价值函数进行评估：遵循树策略到达状态为 s 的叶节点后，需要确定 s 的价值。此处使用学习得到的价值函数 V_θ 直接估计 s 的价值，无需进行滚动模拟。

5.5 引导式评估

蒙特卡洛树搜索中的评估步骤蒙特卡洛树搜索 用于估计树搜索过程中遇到的叶节点的（状态）价值。虽然这一步通常也称为滚动模拟步骤或模拟步骤，但其目的是估计叶节点的价值。滚动模拟或模拟只是用于生成价值估计的方法。这里，我们使用“评估”一词，因为神经蒙特卡洛树搜索文献中的评估方法并非都基于滚动模拟。蒙特卡洛树搜索 。

在评估阶段，使用学习策略和价值函数有两种显而易见的方式：使用学习策略进行滚动模拟，以及使用学习得到的价值函数进行直接预测。前者从遇到的叶节点开始，反复从学习策略中采样动作，直到到达终止节点（见图10）；后者则无需进行滚动模拟，直接由学习得到的价值函数预测叶节点的价值（见图11）。



图12 受调查文章中评估阶段各类选择的比例。 *Value function* 表示使用学习得到的价值函数进行评估， *policy* 表示使用学习策略进行滚动模拟， *standard* 表示随机模拟展开。

大多数作者使用上述学习策略或价值函数，其中提及学习得到的价值函数的情况有 54 次，提及学习策略函数的情况有 26 次（见图12）。其余出版物要么在评估阶段不使用神经引导，要么其方法并不明确。

在神经网络评估方法中，有些作者会根据训练的进展阶段采用不同的评估方法。Song等人 [104] 将两种方法结合起来：在训练早期，按照学习策略网络进行滚动模拟；在后期阶段，则使用学习价值网络进行估计。Zhang 等人 [150] 先通过随机模拟对策略网络进行预训练，再在后期阶段使用学习到的策略网络进行滚动模拟。

有些情况下，模拟并非简单地使用学习策略来执行，而是由学习函数引导更复杂的模拟展开程序。He 等人 [40] 使用价值网络来引导特定于问题的模拟展开程序，而 Xing 等人 [136] 则使用学习策略函数来引导束搜索。Kumar 等人 [58] 将神经网络预测的价值估计与领域特定的模拟展开策略相结合，原因是他们的奖励函数比棋盘游戏中通常见到的奖励函数更加细致。最后，Lu 等人 [64] 在符号回归任务中使用价值网络，估计叶节点是否值得通过优化方法进一步细化，但他们的方法高度依赖于具体问题。

Deng 等人 [23]会根据待评估节点的深度采用不同的评估方法。如果节点更接近树的根节点，他们会使用神经网络估计状态的价值；而对于树中更深层的节点，则会进行随机模拟展开。他们的实验表明，这种混合方法能够兼顾解的质量和计算时间。

由于神经网络推理会带来不可忽略的计算成本，用随机模拟替代它们可以缩短搜索时间，尤其是在较深层级，此时模拟展开的长度较短。

在 He 等人所描述的应用中 [41]，回合可能产生成功解或失败解。成功解的质量也可能有所不同，因此会获得相应的奖励。他们使用学习策略进行滚动模拟；但如果由此得到失败终止状态，就会回溯，直到找到成功解。他们认为，这样可以提高搜索效率，因为无需重复先前的轨迹。

其他阶段的设计选择蒙特卡洛树搜索也会影响评估阶段的要求。Kovari 等人 [56] 将 UCT 公式中的利用项（即估计值）替换为策略网络预测的采取某个动作的概率。因此，他们在评估步骤中不进行滚动模拟或直接预测价值，而是只预测这一概率。

5.6 多阶段中的引导

如上所述，神经引导可用于蒙特卡洛树搜索的选择、扩展和评估阶段。当然，不必将这种引导一次只限于一个阶段；不同类型的神经引导也可以结合在同一种方法中。

为了概览不同类型的神经引导通常如何结合使用，我们在图13中直观展示了它们的用法。

最常见的方法是引导选择步骤，执行标准的蒙特卡洛树搜索扩展，然后使用学习得到的价值函数进行评估。还存在许多其他组合，但它们的使用频率都不高。采用标准选择时，神经扩展的相对使用比例高于采用神经选择时的比例。不过，这或许可以解释为，相关作者只是希望突出神经扩展的效果，因为这通常是其相关出版物的主要关注点。

不常出现的组合可能表明需要进一步研究。

5.7 动力学模型的使用

神经网络 蒙特卡洛树搜索是一种基于模型的强化学习方法，也就是说，它需要访问环境的动力学模型才能进行规划。通常，这个动力学模型是给定的[99, 101]，并可直接用于树搜索。与例如 OpenAI Gym 标准中定义的常规强化学习环境相比，[7], 动力学模型允许根据任意初始状态和待执行动作计算下一状态和奖励。另一方面，环境在任意给定时刻都处于某个特定状态，该状态可以受到动作的影响，但无法基于任意状态进行动力学计算。换句话说，环境是有状态的，而动力学模型不是。

开发此类动力学模型的难度会因应用而异。无论如何，开发都需要付出一些额外的努力。为了避免这一点，也可以学习动力学模型，然后使用该学习模型在蒙特卡洛树搜索[89, 115]中进行规划。虽然这会增加训练过程的复杂度，但在无法轻易获得精确且高效的环境模型的场景中，这样做可能会有所帮助。

在综述过程中，我们发现绝大多数方法都使用现有的动力学模型，但学习模型在实践中也有一定应用。

例如，Chen 等人 [14] 研究了一项自动驾驶任务，其中需要一个模型来预测车辆状态。在这种情况下，车辆状态是一幅图像，这意味着模型需要根据输入图像（对应于初始状态）和一个动作生成输出图像。虽然手动实现这类模型并非易事，但 Chen 等人 [14] 成功训练了一个卷积神经网络来实现这一功能。

同样，Challita 等人 [10] 应用神经 蒙特卡洛树搜索实现长期演进和新空口系统之间的动态频谱共享，并指出这需要为长期演进和新空口分别设计独立调度器的模型，而这并非易事。相反，他们采用类似于 MuZero 所提出的方法来学习模型 [89]。也就是说，动态并非基于原始观测计算，而是基于隐藏表征计算，这些表征由学习得到的表征函数生成

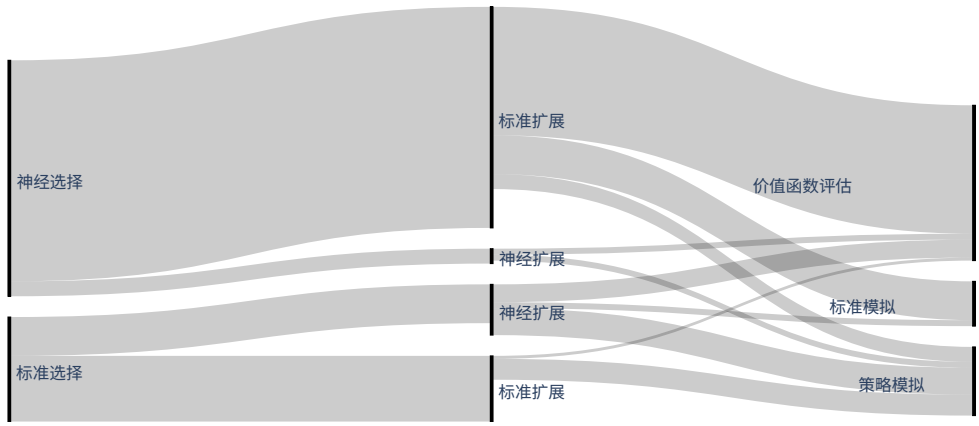


图 13 所调查的神经引导配置平行集示意图。每个垂直条表示 蒙特卡洛树搜索 各阶段中的一个选项：选择（左）、扩展（中）、评估（右）。在扩展步骤中，每个选项（标准扩展和神经扩展）均显示两次，以便更轻松地追踪图示配置。

这些表征由学习得到的表征函数根据观测生成。其他研究者也采用了这种方法[32, 96], ，但也有研究采用直接作用于观测的动力学模型 [21, 125]。

在许多情况下，神经蒙特卡洛树搜索 训练过程中并不会学习动力学模型，而是提前单独训练该模型，然后在树搜索期间直接用于推理 [9, 28, 39, 122]

如上所述，学习动力学模型的一个动机可能是手动创建模型很困难。另一个动机是，学习得到的模型可以快速评估 [109]。

在某些情况下，状态转移相对容易建模，而奖励的计算却很耗时。有些作者不训练完整的动力学模型，而是训练评分模型，以便快速评估给定解的质量。与学习得到的价值函数不同，后者可用于评估任意深度新扩展的节点；评分模型只对完整解（即终止节点）赋予分数。由此得到的分数随后可用作价值函数的训练目标[42, 106, 151]。

5.8 蒙特卡洛树搜索的改进

现在，我们将注意力转向综述过程中遇到的典型（神经）蒙特卡洛树搜索程序的若干改进。

平均值与最佳值

如第5.3节简要所述，确定性单人场景的要求与组合博弈不同

。基于节点平均价值的动作选择会导致次优结果，因为强子节点可能被弱兄弟节点包围。虽然 $PUCT_B$第5.3 节所述的机制是一种解决方案，但其他作者也发现了这一问题，并提出了各自的解决方案。

Deng 等人 [23]报告称，在实证实验中，最终的搜索结果通常不如滚动模拟过程中找到的最佳解。他们指出，树搜索结束后的最终动作选择是基于节点访问次数 $N(s)$进行的。为解决这一问题，他们为优质解引入了一种过采样机制。每当滚动模拟中找到一个优于此前所有已发现解的解时，后续的选择阶段就会在一段时间内优先考虑该解，因此该解会被访问得更频繁。

Peng 等人 [77]和 Xing 等人 [137]采用了更简单的方法来解决同一问题。这里，式（4）中的利用部分，即节点的平均价值，被直接替换为该节点观测到的最佳价值。Fawzi 等人 [26]也采用了类似的策略。

Zhang 等人 [154] 仅记录搜索过程中遇到的最大奖励及其对应的动作序列，并在搜索结束后将其返回。

价值归一化

虽然组合博弈很适合采用范围内的奖励函数形式

$[-1,1]$，但在其他应用中，奖励往往不那么规律，有时甚至完全没有界限。如前所述，探索常数 c 需要根据不同的奖励范围进行调整 [8]。此外，即使 c经过了完美调整，超出 $[-1,1]$ 或 $[0,1]$ 范围的奖励通常也不利于算法收敛 [97]。因此，许多作者建议根据截至当前时刻在树搜索中观察到的最小值和最大值对 Q值进行归一化 [78, 97, 137]。

神经网络推理的成本

神经网络 蒙特卡洛树搜索方法的核心思想是利用神经推断提高树搜索的效率。虽然单次神经推断的计算成本并不高，但大量执行时，所需的计算时间可能会累积到相当可观的程度。

Deng 等人 [23]通过在搜索过程中按一定比例的时间关闭神经引导，来改变神经推断的数量。他们发现，神经引导通常有助于搜索，但超过某个限度后，进一步增加神经引导只会提高计算成本，而不会带来额外收益。

设计机制，以有针对性的方式将神经引导限制在能发挥最大效益的环节，或许是一个有趣的研究方向。

5.9 MCTS超参数

神经 蒙特卡洛树搜索方法设计的最后一个方面，是选择恰当的超参数。尽管超参数的选择高度依赖于具体问题，但查看超参数的平均值仍可能有所帮助，可将其作为起点，并据此确定针对特定问题进行超参数优化时的合理范围。为此，我们在图中总结了综述期间发现的蒙特卡洛树搜索特定超参数的取值，见图 14。探索常数的取值差异很大；该常数 c 用于 UCT风格的选择公式，而蒙特卡洛树搜索预算 n_{MCTS}即在给定时间步内执行的模拟或对局模拟次数。

众所周知， c的最优选择取决于所遇到奖励的尺度

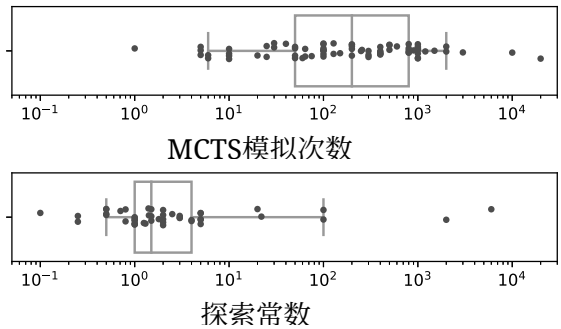


图 14报告的蒙特卡洛树搜索超参数分布：蒙特卡洛树搜索模拟次数（上）和探索常数 c （下）。请注意，两种情况下均使用对数坐标轴。纵轴上的变化没有实际意义，只是添加了随机抖动，以便更清楚地显示各个数据点。

[8]。虽然奖励尺度差异很大，而且我们调查的文献中并不总是报告这一信息，但最常报告的奖励尺度区间为 $[-1,1]$，其中选用了六种不同的 c 取值，范围从 0.5 到 5.0。换言之，奖励的尺度似乎并非作者选择超参数值的唯一依据。

有时，探索常数不采用固定值，而是根据具体情况动态调整 c ，并且蒙特卡洛树搜索预算 n_{MCTS} 也会随具体情况动态变化。

有时，随着训练的进行，探索常数 c 会逐渐减小 [16] [157], ；这或许是因为在后期训练迭代中，相较于进一步探索，完善当前策略更有益。Wang 等人[126] 根据当前观察到的最大 Q 值动态调整 c ，以便随着 Q 值估计的变化平衡探索与利用，并报告称观察到的结果有显著改善。

某些问题设置中的实例规模会有所不同，通常实例规模越大，难度也越高。Zhong 等人 [155] 会随着问题规模增大而提高 c ，这或许是因为要充分覆盖更大的状态空间，需要进行更多探索。出于类似原因，他们及其他研究者 [123, 136] 还会为更大的实例规模选择更大的 n_{MCTS} 。Wang 等人 [121]甚至明确地将n_{MCTS} 设为问题规模的函数。

在许多问题中，剩余搜索空间的大小会随着

树深度的增加而减小。Hu 等人 [45] 认为，搜索预算应取决于搜索起始节点的深度，并提出了一种会n_{MCTS} 随树深度增加而递减的机制。

Fawzi 等人 [26] 报告称，在经过一定数量的训练步数后， n_{MCTS}有所提升；这可能是因为后期的训练迭代能从更高的搜索预算中获益更多，因此应将总体训练时间预算中更大比例分配给后期迭代。

有些作者会减少蒙特卡洛树搜索在测试时的模拟次数。例如，Chen 等人[14] 将 n_{MCTS} 在测试时减少了十倍，与训练阶段相比。他们还限制了两个阶段的搜索深度，但在测试时限制得更多。

在某些情况下，蒙特卡洛树搜索仅用于训练策略网络，而该网络在测试时不再进行树搜索，直接应用 [9, 10, 18, 20, 155]。这可能是由于应用的特定要求所致，即某些应用需要在测试时快速推理，因此无法进行树搜索，但仍能从训练时使用蒙特卡洛树搜索中获益 [10]。在某些应用中，训练时使用蒙特卡洛树搜索后，测试时再使用它并不能显著提升性能 [20]。

6 讨论与结论

在关注神经蒙特卡洛树搜索在博弈之外的应用时，我们通过开展系统性文献综述，考察了应用的多样性、特征以及所采用的神经蒙特卡洛树搜索方法的设计。

关于引言中提出的**研究问题 1**，我们发现，神经蒙特卡洛树搜索应用于多种多样的领域，以解决不同的问题。尽管大多数问题都具有相似的特征，例如离散时间和动作、有限时域以及确定性状态转移，但许多作者也证明，神经蒙特卡洛树搜索能够应用于属性各异的问题。

综述中涉及的应用通常具有与组合博弈略有不同的需求和属性。这确实会影响解的设计方式，这是**研究问题 2**所探讨的一个方面。例如，自我对弈这一概念通常不适用于单人问题。在某些情况下，单人问题可以建模为多人问题，但这只是例外，而非常态。可以采用一种称为自我竞争的机制，它能够在单人场景中模拟自我对弈生成奖励的方式。

许多作者还指出，在单人确定性环境中，根据节点价值的平均值选择动作并不理想。对此，可以采用多种以追踪节点最大值为基础的机制，调整典型的蒙特卡洛树搜索机制。

与博弈领域的神经蒙特卡洛树搜索生态系统以及传统的强化学习生态系统相比，神经蒙特卡洛树搜索领域几乎完全缺乏标准化实现和组件。相反，几乎所有实现都是完全定制的。少数情况下，其他研究者会重复使用领域专用实现，但这些实现仅适用于范围非常狭窄的问题（例如，见 [145]）。对于更基础的研究而言，这种情况可以理解，因为实现本身往往处于不断变化之中；但我们认为，标准化组件能够显著推动应用研究领域的进展。

缺乏标准化组件的一个原因，可能是神经蒙特卡洛树搜索方法的设计差异很大，涵盖了训练方法、自我博弈相关机制的使用、神经引导形式，以及对传统蒙特卡洛树搜索设置的其他改进。广泛适用的神经蒙特卡洛树搜索框架必须具备高度可配置性，才能适应不同学科和问题设置。尽管这是一项艰巨的任务，但传统强化学习生态系统表明，标准 [7] 和公开可用的算法实现 [79]能够加快进展，而且适用于研究的灵活框架 [6] 是可以设计出来的。

针对**研究问题3**，可以得出结论：游戏文献中采用的神经引导方式，往往也用于其他应用。最常见的引导类型是神经选择和利用学习得到的价值函数进行评估，这与 AlphaZero 中的做法相同 [100]。文献中也能找到其他类型的神经引导及其组合。鉴于不同神经蒙特卡洛树搜索系统之间存在诸多差异，实践者面临的一个核心问题是如何搭建自己的系统

，具体取决于其应用的特征。理想情况下，我们能够将观察到的问题特征映射到观察到的神经蒙特卡洛树搜索配置，从而为他人提供参考指南。某些问题特征，例如动作空间的离散或连续特性，在审阅现有出版物时相对容易确定；另一些则不那么容易判断。例如，（完整）树的广度和深度很少有明确报告。在某些情况下，或许可以推断出这些信息，但在一篇涵盖众多不同学科的综述中，要可靠地做到这一点很困难。

尽管如此，我们仍可从收集到的文献中得出一些见解。无论是在博弈领域还是其他领域，显然，神经引导有助于提高树搜索的效率，但在某些情况下，也可能增加计算成本，却没有带来多少额外收益。因此，应仔细权衡何时以及如何采用神经引导。例如，在评估阶段，选择何种机制取决于整体树的深度，以及待评估节点的深度。如果滚动模拟的长度较短，那么相比于使用学习得到的价值函数进行估计（这会产生相应的推理成本），滚动模拟可能更可取。究竟在什么深度下其中一种方法优于另一种，取决于所用神经网络的规模——它会影响推理成本——以及其预测的质量。若要与高质量的学习得到的价值函数估计相媲美，可能需要进行多次模拟，因为单次滚动模拟对节点价值的估计方差较大。良好的初始估计有助于将搜索集中在解空间中有希望的区域，而糟糕的估计则可能使搜索偏离方向。

在树的宽度较大的应用中，神经网络引导在扩展阶段可能特别有用，因为它可以完全避免将搜索树中的某些路径纳入考虑。当然，这也有可能因此排除高质量的解。在这种情况下，神经网络预测的质量同样决定了这种方法是否可行。然而，尤其是在树的宽度极大，甚至涉及连续域的应用中，如果不对解空间进行一定程度的限制，搜索可能根本无法进行。

显然，仍有许多问题悬而未决。此外，纯粹回顾过去的综述往往总结过去已经做了什么，而不是本应做什么。因此，本综述主要呈现的是现有方法的图景，而非一套指导性知识。不过，本综述汇集的结果可以作为进一步实验研究的基础。针对具有不同属性的多种应用明确设计实验，并系统地控制本综述中识别出的各项因素，有助于更全面地理解神经蒙特卡洛树搜索方法的设计。基于这种理解，可以总结出指导性经验法则，帮助从业者根据应用的具体属性做出恰当的设计选择。

致谢

本研究由德国研究联合会（DFG）通过德国卓越战略项目“生产互联网”（EXC-2023，项目编号390621612）资助。

参考文献

[1] Al-Saffar, M., and Musilek, P. 基于强化学习的分布式电池储能系统管理：缓解高光伏渗透率系统中的过电压问题。*IEEE*智能电网汇刊 *11*,4（2020年）,2980–2994。

[2] Anthony, T., Tian, Z., and Barber, D. 用深度学习和树搜索进行快思考与慢思考。《神经信息处理系统进展》30 (2017)。

[3] Audrey, G., and Francesco, M. 用于流固拓扑优化的深度神经网络和蒙特卡洛树搜索。《科学报告》9，1（2019年4月

），15916。

[4] Auer, P., Cesa-Bianchi, N., and Fis-cher, P. 多臂老虎机问题的有限时间分析。《机器学习》47，2（2002年），235–256。出版商：施普林格。

[5] Bai, F., Meng, F., Liu, J., Wang, J.,

and Max Q.-H., M. 用于非抓取式多物体重排的深度强化学习分层策略。仿生智能与机器人学 *2*,3（2022年），100047。

[6] Bitter, C., Thun, T., and Meisen, T. Karolos：用于机器人任务环境的开源强化学习框架。arXiv 预印本 arXiv:2212.00906（2022年）。[7] Brockman, G., Cheung, V., Petters-son, L., Schneider, J., Schulman, J.,Tang, J., and Zaremba, W. OpenAI Gym。arXiv 预印本 arXiv:1606.01540（2016）。[8] Browne, C. B., Powley, E., White-house, D., Lucas, S. M., Cowling, P. I.,Rohlfshagen, P., Tavener, S., Perez,D., Samothrakis, S., and Colton, S. 蒙特卡洛树搜索方法综述。《IEEE计算智能与 AI 博弈汇刊》4，1（2012），1–43。出版商：IEEE。[9] Carfora, V., Di Massimo, F.,Rastelli,R., Catellani, P., andPiastra, M. 通过说服心理学和机器学习进行对话代理的对话管理。《多媒体工具与应用》79，47（2020年），35949–35971。出版商：施普林格。

[10] Challita, U., and Sandberg, D. 用于长期演进和新空口动态频谱共享的深度强化学习。载于*ICC 2021 - IEEE*国际通信会议 (2021)，第 1–6 页。

[11] Chen, J., Chen, S., Luo, S., Wang, Q., Cao, B., and Li, X. 一种用于无人机边缘计算网络的智能任务卸载算法（iTOA）。数字通信与网络 *6*，4（2020年），433–443。出版商：爱思唯尔。

[12] Chen, J., Chen, S., Wang, Q., Cao,B., Feng, G., and Hu, J. iRAF：一种用于协作式移动边缘计算物联网的深度强化学习方法网络。*IEEE*物联网期刊 *6*，4（2019年），7011–7024。出版方：IEEE。

[13] Chen, J., Luo, S., Zhang, L., Zhang,C., and Cao, B. iPAS：一种基于深度蒙特卡洛树搜索的海量 MIMO 系统智能导频功率分配方案。数字通信与网络 7，3（2021年），362–372。出版商：爱思唯尔。

[14] Chen, J., Zhang, C., Luo, J., Xie, J., and Wan, Y. 基于深度蒙特卡洛树搜索的自动驾驶控制中的驾驶操作预测。*IEEE*车辆技术汇刊 *69*，7（2020年），7146–7158。出版商：IEEE。

[15] Chen, P.-Y., Ke, B.-T., Lee, T.-C., Tsai, I.-C., Kung, T.-W., Lin, L.-Y.,Liu, E.-C., Chang, Y.-C., Li, Y.-L., andChao, M. C.-T. 一种用于避障直角斯坦纳树构造的强化学习智能体。载于《2022年国际物理设计研讨会》论文集（2022年），第107–115页。[16] Chen, Y.-Q., Chen, Y., Lee, C.-K., Zhang, S., and Hsieh, C.-Y. 利用蒙特卡洛树搜索优化量子退火调度，并通过神经网络增强。《自然·机器智能》4, 3（2022年），269–278。出版方：《自然》出版集团。[17] Chen, Z., Huang, J., Ahn, H., andNing, X. 使用蒙特卡洛树搜索进行高成本特征分类。载于《2021年国际神经网络联合会议》（IJCNN）（2021年），第1–8页。[18] Chen, Z., Zhong, S., Chen, J., andZhao, Y. DeepPursuit：结合经典智慧与深度强化学习，实现稀疏恢复。载于《2021年第55届阿西洛马会议：信号、系统与计算机》（2021年），第1361–1366页。[19] Cheng, Y., Wu, Z., Liu, K., Wu, Q.,and Wang, Y. 使用蒙特卡洛树搜索方法，在可信与不可信实体之间智能调度DAG任务。《可持续发展》11, 7

（2019年），1826。出版商：MDPI。

[20] Dai, Y., Wang, P., and Zhang, L. 面向目标的观点词提取的强化句法依存树推理。载于国际人工神经网络会议（2021年），施普林格，第531–543页。

[21] Dai, Z., Liu, C. H., Ye, Y., Han, R., Yuan, Y., Wang, G., and Tang, J.基于模型的图卷积强化学习实现AoI最小化的无人机众包感知。载于*IEEE INFOCOM 2022 - IEEE*计算机通信会议（2022年），第1029–1038页。

[22] Dalgaard, M., Motzoi, F., Sørensen, J. J., and Sherson, J. 利用AlphaZero深度探索实现量子动力学的全局优化。*npj*量子信息 *6*，1（2020年），1–9。出版商：自然出版集团。

[23] Deng, H., Yuan, X., Tian, Y., andHu, J. 用于HP模型蛋白质折叠的结合过采样的神经网络增强型两阶段蒙特卡洛树搜索。《电气与电子工程师学会电气与电子工程学报》*17*，5（2022年），685–694。出版商：Wiley在线图书馆。

[24] Dieb, S., Song, Z., Yin, W.-J., and Ishii, M. 利用机器学习优化用于X射线光学的深度渐变多层结构。《应用物理学杂志》*128*，7（2020年），074901。出版商：AIP出版有限责任公司。

[25] Erikawa, D., Yasuo, N., and Sekijima, M. MERMAID：一种基于深度强化学习的开源自动化先导化合物发现方法。《化学信息学杂志》*13*，1（2021年），1–10。出版商：施普林格。

[26] Fawzi, A., Balog, M., Huang, A., Hubert, T.,Romera-Paredes, B., Barekatain, M., Novikov, A., R Ruiz,F. J., Schrittwieser, J., Swirszcz, G.,Silver, D., Hassabis, D., and Kohli,P. 利用强化学习发现更快的矩阵乘法算法。《自然》*610*, 7 (2022年10月), 47–53.

[27] Feng, Y., Li, B., Zheng, Q., Wang, D., Xu, X., and Zhang, R. 基于深度学习的电磁态势分析与判断。《*IET*通信》*15*, 11 (2021年), 1455–1466。出版商：WileyOnline Library。

[28] Fong, J., Campolo, D., Acar, C., and Tee, K. P. 用于非抓取式操作规划的基于模型的强化学习与 LSTM 网络。载于*2021*年第*21*届国际控制、自动化与系统会议（*ICCAS*）（2021年），第 1152–1158 页。

[29] Fricke, C., Wolff, D., Kemmerling, M., and Elgeti, S. 针对二维型材挤压模具几何形状优化的强化学习研究。计算科学与工程进展（2023年），0–0。出版方：计算科学与工程进展。

[30] Fu, Z., Fan, Q., Zhang, X., Li, X., Wang, S., and Wang, Y. 用于智能服务功能链部署的策略网络辅助蒙特卡洛树搜索。载于*2021*年 *IEEE* 第*20*届计算与通信中的信任、安全与隐私国际会议（*Trust-Com*）（2021年），第 1161–1168 页。

[31] Gaafar, M., Shaghaghi, M., Adve, R. S., and Ding, Z. 用于认知雷达任务调度的强化学习。载于*2019*年第*53*届阿西洛马会议：信号、系统与计算（2019年），第1653–1657页。

[32] Gabirondo-López, J., Egaña, J., Miguel-Alonso, J., and Orduna Urru-tia, R. 利用基于 MuZero 的智能体实现 SDN 网络自主防御。*IEEE Access 9*（2021年），107184–107199。

[33] Ganapathi Subramanian, S., and Crowley, M. 结合蒙特卡洛树搜索与

A3C，用于预测森林野火场景中的空间扩散过程。载于加拿大人工智能会议（2018），施普林格，第 285–291 页。

[34] Gannouni, A., Samsonov, V., Behery, M.,Meisen, T., and Lakemeyer,G. 用于生产调度的神经组合优化：考虑序列相关的设置浪费。载于*2020*年 *IEEE* 国际系统、人与控制论会议（*SMC*）（2020），第 2640–2647 页。

[35] Gauthier, T. 用于在高阶逻辑中合成函数的深度强化学习。*EPiC*计算系列 *73*（2020），230–248。

[36] Genheden, S., Thakkar, A., Chadimová, V., Reymond, J.-L., Engkvist,O., and Bjerrum, E. AiZynthFinder：一种快速、稳健且灵活的开源逆合成规划软件。《化学信息学杂志》*12*，1（2020），1–9。出版商：BioMed Central。

[37] Gros, T. P., Gross, J., and Wolf, V. 使用深度强化学习进行汽车制造过程的实时决策。发表于*2020*年冬季仿真会议（*WSC*）（2020年），第3032–3044页。

[38] Göppert, A., Mohring, L., and Schmitt, R. H. 利用人工神经网络预测动态互联装配系统中基于人工智能的在线调度绩效指标。《生产工程》第*15*卷，第5期（2021年），619–633页。出版商：施普林格。

[39] Ha, T., Cho, K., Cha, G., Lee, K., and Oh, S. 基于预测模型的蒙特卡洛树搜索车辆控制。发表于*2020*年第*17*届普适机器人国际会议（*UR*）（2020年），第303–308页。

[40] He, L., Shao, B., Xiao, Y., Li, Y., Liu, T.-Y., Chen, E., and Xia, H. 知识图谱中的神经引导语义导航。《*IEEE*大数据汇刊》*8*，3（2018），607–615。出版方：IEEE。

[41] He, Y., Li, H., Jin, T., and Bao, F. S.使用蒙特卡洛树搜索和深度强化学习进行电路布线。发表于*2022*年超大规模集成电路设计、自动化与测试国际研讨会（*VLSI-DAT*）（2022年），第1–5页。

[42] He, Y., Tao, S., Xu, J., Guo, J., Lan,Y., and Cheng, X. 基于蒙特卡洛树搜索的文本匹配。发表于中国信息检索会议（2018年），施普林格，第41–52页。

[43] Hernandez, D., Denamganaï, K., Gao, Y., York, P., Devlin, S., Samothrakis,S., and Walker, J. A. 一种用于自我对弈训练的通用框架。载于*2019*年 *IEEE* 博弈会议（*CoG*）（2019年），IEEE，第 1–8 页。

[44] Hoel, C.-J., Driggs-Campbell, K., Wolff, K., Laine, L., and Kochender-fer, M. J. 将规划与深度强化学习相结合，用于自动驾驶中的战术决策。*IEEE* 智能车辆汇刊 *5*，第 2 期（2020年），294–305 页。

[45] Hu, Z., Tu, J., and Li, B. Spear：利用深度强化学习优化依赖感知型任务调度。载于*2019*年 *IEEE* 第 *39* 届分布式计算系统国际会议（*ICDCS*）（2019年），第 2037–2046 页。

[46] Huang, J., Patwary, M., and Diamos,G. 使用 AlphaGo Zero 为大型图着色。*arXiv* 预印本 *arXiv:1902.10162*（2019年）。

[47] Huang, Y., and Yu, Y. 使用强化学习蒸馏深度神经网络。在*2018*年 *IEEE* 信息与自动化国际会议（*ICIA*）（2018年），第 133–138 页。

[48] Ishida, S., Terayama, K., Kojima, R., Takasu, K., and Okuno, Y. 结合逆合成知识的 AI 驱动合成路线设计。《化学信息与建模杂志》*62*, 6（2022年），1357–1367。出版商：ACSPublica- tions。

[49] Jang, Y., Lee, J., and Kim, K.-E. 面向目标导向对话的贝叶斯自适应蒙特卡洛规划与学习。载于*AAAI* 人工智能会议论文集（2020年），第34卷，第7994–8001页。期号：05。

[50] Jiang, Y., Liu, M., Li, J., and Zhang, J.面向智能电网中基于认知绿色计算的非侵入式在线负荷识别的强化蒙特卡洛树搜索。*MathematicalBiosciences and Engineering 19*，第11期（2022年），11595–11627。

[51] Kemmerling, M., Samsonov, V., Lütticke, D., Schuh, G., Gützlaff,A., Schmidhuber, M., Janke, T., andMeisen, T. 面向生产就绪型强化学习调度智能体：一种基于离散事件仿真的混合两步训练方法。Simulation in Produktion und Logistik（2021年），325–336.[52] Kim, M., Park, J.-K., and Moon, S.-M. 使用深度强化学习解决基于 PBQP 的寄存器分配问题。载于《2022年 IEEE/ACM 代码生成与优化国际研讨会》（CGO）（2022年），第 1–12 页。

[53] Kocsis, L., and Szepesvári, C. 基于赌博机的蒙特卡洛规划。载于《欧洲机器学习会议》（2006年），施普林格，第 282–293 页。

[54] Kocsis, L., Szepesvári, C., and Willemson, J. 改进的蒙特卡洛搜索。塔尔图大学，爱沙尼亚，技术报告 1（2006年）。

[55] Kovari, B., Becsi, T., Szabo, A., and Aradi, S. 利用蒙特卡洛树搜索增强的气动执行器基于策略梯度的控制。载于《2020年第六届机电一体化与机器人工程国际会议》（IC MRE）（2020年），第 177–182 页。

[56] Kovári, B., Hegedüs, F., and Bécsi, T. 面向自动驾驶车辆的基于强化学习的车道保持规划智能体设计。《Applied Sciences》10卷20期（2020年），7171。出版商：MDPI。

[57] Kuai, Z., Wang, T., and Wang, S. 基于近端策略优化的公平虚拟网络功能映射与调度。*IEEE* 通信汇刊 *70*，第11期（2022年），7434–7445。

[58] Kumar, A., and Dimitrakopoulos, R. 利用树搜索和深度强化学习，对工业采矿联合体在接收新信息时进行生产调度。*Applied Soft Computing 110*（2021年），107644。出版商：爱思唯尔。

[59] Lao, Y., Xu, J., Gao, S., Guo, J., andWen, J.-R. 基于策略价值网络的命名实体识别。在第*42*届 *ACM SIGIR* 信息检索研究与开发国际会议论文集(2019年)，第1245–1248页。

[60] Laterre, A., Fu, Y., Jabri, M., Cohen,A., Kas, D., Hajjar, K., Dahl, T.,Kerkeni, A., and Beguir, K. 排名奖励：为组合优化启用自我对弈强化学习。神经信息处理系统进展第*31*卷（*NeurIPS 2018*）（2018年）。

[61] Lei, L., Luo, R., Zheng, R., Wang,J., Zhang, J., Qiu, C., Ma, L., Jin,L., Zhang, P., and Chen, J. KB- Tree：用于自动驾驶规划的可学习、连续蒙特卡洛树搜索。在*2021*年 *IEEE/RSJ* 智能机器人与系统国际会议（*IROS*）（2021年），第4493–4500页。

[62] Li, H., Hu, Y., Cao, Y., Zhou, G., and Luo, P. 通过强化学习恢复富文本文档样式。《计算机科学前沿》*15*，4（2021年），1–11。出版商：施普林格。[63] Li, Y., Pei, J., and Lai, L. 使用三维深度生成模型进行基于结构的从头药物设计。《化学科学》*12*，41（2021年），13664–13675。出版商：英国皇家化学会。[64] Lu, Q., Tao, F., Zhou, S., and Wang, Z.

将演员-评论家融入用于符号回归的蒙特卡洛树搜索。《神经计算与应用》*33*，14（2021年），8495–8511。出版商：施普林格。

[65] Ma, B., Terayama, K., Matsumoto, S., Isaka, Y., Sasakura, Y., Iwata, H., Araki, M., and Okuno, Y. 结合人工智能与分子对接模拟的基于结构的从头分子生成器。《化学信息与建模杂志》*61*, 7 (2021年), 3304–3313。出版商：ACS 出版社。

[66] Mandhane, A., Zhernov, A., Rauh, M.,Gu, C., Wang, M., Xue, F., Shang, W.,Pang, D., Claus, R., Chiang, C.-H.,and others。用于 VP9 视频压缩码率控制的 MuZero 自我竞争。*arXiv* 预印本 *arXiv:2202.06626* (2022年)。

[67] Mao, K., and Xiao, Y. 基于强化学习和蒙特卡洛树搜索学习最快的 RNA 折叠路径。《分子》*26*, 15 (2021年), 4420。出版商：MDPI。

[68] Mańdziuk, J. MCTS/UCT 在解决现实问题中的应用。载于《计算智能方法在数据分析中的进展》。施普林格，2018年，第277–292页。

[69] Meng, X., Inaltekin, H., and Krongold, B. 面向自组织无线传感器网络的深度强化学习拓扑优化。载于*2019 IEEE* 全球通信大会（*GLOBECOM*）（2019年），第1–6页。

[70] Mo, S., Pei, X., and Wu, C. 使用蒙特卡洛树搜索的自动驾驶车辆安全强化学习。《*IEEE* 智能交通系统汇刊》第*23*卷，第7期（2022年），6766–6773页。

[71] Mo, T.-W., Chang, R. Y., and Kan, T.-Y. DeepMCTS：用于 MIMO 检测的深度强化学习辅助蒙特卡洛树搜索。载于*2022*年 *IEEE* 第 *95* 届车辆技术大会：（*VTC2022*-春季）（2022年），第 1–6 页。

[72] Moerland, T. M., Broekens, J.,Plaat, A., and Jonker, C. M. A0c：连续动作空间中的 Alpha zero。*arXiv* 预印本 *arXiv:1805.09613*（2018年）。

[73] Moerland, T. M., Broekens, J., Plaat, A., and Jonker, C. M. 强化学习与规划的统一框架。《人工智能前沿》第*5*卷（2022年）。出版商：Frontiers Media SA。

[74] Motomichi, T., Yutaka, I., and Michio, K. 结合蒙特卡洛树搜索与循环神经网络，从头生成具有光学活性的小分子有机化合物。《计算化学杂志》42，3（2021年1月），136–143。

[75] Oren, J., Ross, C., Lefarov, M., Richter, F., Taitler, A., Feldman, Z., Di Castro, D., and Daniel, C. SOLO：在线搜索，离线学习，用于组合优化问题。载于国际组合搜索研讨会论文集（2021），第12卷，第97–105页。期号：1。

[76] Paxton, C., Raman, V., Hager, G. D., and Kobilarov, M. 结合神经网络与树搜索，在具有挑战性的环境中进行任务与运动规划。载于*2017*年 *IEEE/RSJ* 智能机器人与系统国际会议（*IROS*）（2017），第6059–6066页。

[77] Peng, H., Wu, C., Zhan, Y., and Xia, Y. Lore：一种基于学习的云环境工作流调度方法。发表于自适应与融合系统研究大会论文集（2022年），第 47–52 页。

[78] Qian, H., Lin, C., Zhao, D., Tu, S., and Xu, L. AlphaDrug：针对特定蛋白质靶标的从头分子生成。*PNAS Nexus1*，4（2022年），pgac227。出版商：牛津大学出版社。

[79] Raffin, A., Hill, A., Gleave, A., Kanervisto, A., Ernestus, M., and Dormann, N. Stable-Baselines3：可靠的强化学习实现。《机器学习研究杂志》22，268（2021年），1–8。[80] Raina, A., Cagan, J., and McComb,C. 无需先验数据的设计学习：利用深度学习和树搜索发现可泛化的设计策略。《机械设计杂志》145，3（2022年），031402。出版方：美国机械工程师学会。

[81] Rinciog, A., Mieth, C., Scheikl, P. M., and Meyer, A. 使用 AlphaGo Zero 进行钣金生产调度。载于《生产系统与物流会议论文集：CPSL 2020》（2020年）。[82] Riviere, B., Hönig, W., Anderson, M.,and Chung, S.-J. 用于非合作环境下多机器人规划的神经树扩展。《IEEE机器人与自动化快报》6，4（2021年），6868–6875。出版方：IEEE。

[83] Rosin, C. D. 具有回合上下文的多臂老虎机。《数学与人工智能年刊》61，3（2011年），203–230。出版方：施普林格。

[84] Sadeghnejad-Barkousaraie, A., Bohara, G., Jiang, S., and Nguyen, D.一种用于放射治疗射束方向选择的引导式蒙特卡洛树搜索算法的强化学习应用。《机器学习：科学与技术》2，3（2021年），035013。出版方：IOP出版社。

[85] Sadeghnejad Barkousaraie, A., Ogunmolu, O., Jiang, S., and Nguyen,D. 在放射治疗中使用监督学习和引导式蒙特卡洛树搜索进行射束方向优化。载于《放射治疗人工智能研讨会》（2019年），施普林格，第1–9页。

[86] Samsonov, V., Kemmerling, M., Paegert, M., Lütticke, D., Sauermann, F., Gützlaff, A., Schuh, G., and Meisen, T. 强化学习在作业车间环境中的制造控制。载于智能体与人工智能国际会议第*13*届会议论文集（2021年），第589–597页。

[87] Sandberg, D., Kvernvik, T., and Calabrese, F. D. 利用搜索与注意力进行鲁棒调度。载于*ICC 2022- IEEE* 国际通信会议（2022年），IEEE，第1549–1555页。

[88] Schmidt, D., Moran, N., Rosenfeld, J. S., Rosenthal, J., and Yedidia,J. 不使用奖励指标的自我对弈学习。*arXiv* 预印本 *arXiv:1912.07557*（2019年）。

[89] Schrittwieser, J., Antonoglou, I., Hubert, T., Simonyan, K., Sifre, L., Schmitt, S., Guez, A., Lockhart, E., Hassabis, D., Graepel, T., and oth-ers。通过基于学习模型的规划，在 Atari、围棋、国际象棋和将棋中取得卓越表现。《自然》*588*，7839（2020年），604–609。出版方：自然出版集团。

[90] Segler, M. H., Preuss, M., and Waller, M. P. 利用深度神经网络和符号人工智能规划化学合成。《自然》*555*，7698（2018），604–610。出版商：自然出版集团。

[91] Segler, M. H. S., Preuss, M., and Waller, M. P. 迈向“AlphaChem”：利用树搜索和深度神经网络策略进行化学合成规划。载于第五届国际学习表征会议，*ICLR 2017*，法国土伦，*2017*年*4*月*24*日至*26*日，研讨会分会场论文集 (2017)。[92] Shafieirad, H., and Adve, R. S. 利用强化学习满足最大时延约束。*IEEEAccess 10* (2022)，97897–97911。[93] Shaghaghi, M., Adve, R. S., and Ding,Z. 多功能多通道认知雷达的资源管理。载于

*2019*年 第*53*届阿西洛马会议信号、系统与计算机（*2019*），第*1550–1554*页。

[94] Shao, Y., Liew, S. C., and Wang, T. AlphaSeq：基于深度强化学习的序列发现。*IEEE* 神经网络与学习系统汇刊第*31*卷，第9期（2020年），3319–3333。

[95] Shen, Y., Chen, J., Huang, P.-S., Guo,Y., and Gao, J. M-walk：使用蒙特卡洛树搜索学习在图上行走。神经信息处理系统进展第*31*卷（2018年）。

[96] Shuai, H., and He, H. 基于蒙特卡洛树搜索和学习模型的住宅微电网在线调度。载于*2021*年 *IEEE* 电力与能源学会年度大会（*PESGM*）（2021年），第01–01页。

[97] Shuai, H., Li, F., She, B., Wang, X., and Zhao, J. 利用随机蒙特卡洛树搜索和深度神经网络进行配电网恢复的风暴后维修队伍调度。《国际电力与能源系统杂志》*144*（2023），108477。出版商：爱思唯尔。

[98] Silva, K., Abeyasekare, W., Dasanayake,D.,Nandisena, T., Kasthurirathna, D., and Kugath-asan, A. 基于深度强化学习的动态用户界面个性化。在*2021*年第三届计算进展国际会议（*ICAC*）（2021年），IEEE，第25–30页。

[99] Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L.,Van Den Driessche, G., Schrittwieser, J., Antonoglou, I., Panneershelvam, V., Lanctot, M., and others。利用深度神经网络和树搜索掌握围棋。《自然》*529*，7587（2016），484–489。出版商：自然出版集团。

[100] Silver, D., Hubert, T., Schrittwieser,J., Antonoglou, I., Lai, M., Guez, A.,Lanctot, M., Sifre, L., Kumaran, D.,Graepel, T., Lillicrap, T., Simonyan, K., and Hassabis, D. 一种通用的强化学习算法，通过自我对弈掌握国际象棋、将棋和围棋。《科学》*362*，6419（2018），1140–1144。

[101] Silver, D., Schrittwieser, J., Simonyan, K., Antonoglou, I., Huang,A., Guez, A., Hubert, T., Baker,L., Lai, M., Bolton, A., and oth-ers。无需人类知识即可掌握围棋。《自然》*nature 550*,7676 (2017), 354–359。出版方：自然出版集团。

[102] Sinha, A., Azad, U., and Singh, H. 借助图神经网络辅助的蒙特卡洛树搜索进行量子比特路由。载于*AAAI*人工智能会议论文集（2022），第36卷，第9935–9943页。期号：9。

[103] Skrynnik, A., Yakovleva, A., Davydov, V., Yakovlev, K., and Panov,A. I. 用于多智能体寻路的混合策略学习。*IEEE Access 9*（2021年），126034–126047。

[104] Song, S., Chen, H., Sun, H., and Liu, M. 用于自动泊车系统横向规划与控制一体化的数据高效强化学习。*Sensors 20*，24（2020年）。

[105] Sridharan, B., Mehta, S., Pathak, Y., and Priyakumar, U. D. 利用深度强化学习解决从核磁共振谱到分子结构的分子逆问题。《物理化学快报》*13*（2022年），4924–4933。出版商：ACS 出版社。

[106] Srinivasan, S., Batra, R., Chan, H., Kamath, G., Cherukara, M. J., and Sankaranarayanan,S. K. 以人工智能为指导、针对新冠肺炎的从头分子设计。*ACS omega6*，19（2021年），12557–12566。出版商：ACS 出版社。

[107] Sun, R., and Liu, Y. 考虑风电的输电网络自愈混合强化学习。

IEEE 神经网络与学习系统汇刊 (2021年), 1–11。

[108] Sutton, R. S., and Barto, A. G. 强化学习：导论。麻省理工学院出版社，2018年。

[109] SV, S. S., Law, J. N., Tripp, C. E., Duplyakin, D., Skordilis, E., Biagioni,D., Paton, R. S., and St John, P. C.面向多目标的目标导向优化：为水系氧化还原液流电池从头设计稳定有机自由基。《自然·机器智能》*4*, 8(2022年), 720–730。出版方：自然出版集团。

[110] Tang, C., Fu, S., and Liu, F. 用于氢燃料汽车储氢的金属有机框架生成系统的设计与实现。载于*2021*年 *IEEE* 国际人工智能与工业设计会议（*AIID*）(2021年)，IEEE，第549–553页。

[111] Thacker, H. K., Kumar, A., Barari, A., Damini, D., Gupta, A., Jagan-nathachar, K. K., and Yoon, D.AlphaRA：一种基于 AlphaZero 的冗余分析方法。载于 *2021*年 第*20*届*IEEE*机器学习及应用国际会议（*ICMLA*）（2021年），第477–483页。

[112] Thakkar, A., Kogej, T., Reymond, J.-L., Engkvist, O., and Bjerrum, E. J.数据集及其对制药领域计算机辅助合成规划工具开发的影响。《化学科学》*11*，1（2020年），154–168。出版方：英国皇家化学会。

[113] Todi, K., Bailly, G., Leiva, L., and Oulasvirta, A. 使用基于模型的强化学习调整用户界面。载于*2021*年计算机系统人为因素 *CHI*大会论文集（2021年），第1–13页。

[114] Tomin, N. 基于深度强化学习构建用于电力自动控制系统的人工调度智能系统的概念网络。《计算机与系统科学国际期刊》*59*，第6期（2020年），939–956页。出版商：施普林格。

[115] Van Eyck, J., Ramon, J., Guiza, F., Meyfroidt, G., Bruynooghe, M., and Van den Berghe, G. 用于学习环境中规划的引导式蒙特卡洛树搜索。载于亚洲机器学习会议 (2013)，PMLR，第 33–47 页。

[116] Vodopivec, T., Samothrakis, S., and Ster, B. 关于蒙特卡洛树搜索与强化学习。《人工智能研究杂志》*60* (2017)，881–936。

[117] Vom Brocke, J., Simons, A., Riemer, K., Niehaves, B., Plattfaut, R., and Cleven, A. 站在巨人的肩膀上：信息系统研究中文献检索的挑战与建议。信息系统协会通讯 *37*，1 (2015)，9。

[118] Wang, C., Chen, X., Luo, Y., and Zhang, G. 结合神经网络与蒙特卡洛树搜索，快速解决虚拟网络映射问题。载于*2021*年高级云计算与大数据第九届国际会议（*CBD*）(2022)，第19–24 页。

[119] Wang, H., Liang, W., and Yu, L.-F. Scene mover：基于深度强化学习的场景布置自动移动规划。《*ACM*图形学汇刊》*(TOG) 39*，6（2020年），1–15。出版商：ACM，美国纽约州纽约。

[120] Wang, H., Yang, R., Yin, C., Zou, X., and Wang, X. 基于深度强化学习的无线自组网络移动节点部署自我对弈难度研究。 无线通信与移动计算 *2021*年（2021年）。出版商：Hindawi。

[121] Wang, J. H., Cheng Luo, P., Xiong, H. Q., Zhang, B. W., and Peng,J. Y. 基于近端策略优化训练与蒙特卡洛树搜索集成的并行机器作业车间调度。在*2020*年中国自动化

大会 *(CAC)* (2020年), 第3277–3282页。

[122] Wang, K., and Sun, W. 通过深度强化学习推导理论一致、基于微观结构的牵引-分离定律的元建模博弈。《计算机方法在应用力学与工程》*346* (2019年),216–241。出版商：爱思唯尔。

[123] Wang, Q., Hao, Y., and Cao, J. 基于蒙特卡洛树搜索的自我博弈框架学习遍历图。《工程人工智能应用》105 (2021年), 104422。出版商：爱思唯尔。[124] Wang, R., Zhou, M., Li, Y., Zhang, Q.,and Dong, H. 基于蒙特卡洛树搜索的铁路列车时刻表重新编排方法。载于2019年IEEE智能交通系统会议（ITSC）(2019年), 第 3738–3743 页。[125] Wang, S., Zhou, K., Lai, K., and Shen,J. 通过结合对抗网络的蒙特卡洛树搜索学习任务完成对话策略。载于2020年自然语言处理经验方法大会（EMNLP）论文集 (2020年), 第3461–3471页。[126] Wang, X., Qian, Y., Gao, H., Coley,C. W., Mo, Y., Barzilay, R., andJensen, K. F. 利用蒙特卡洛树搜索和强化学习，高效发现绿色合成路径。《化学科学》11, 40 (2020年), 10959–10972。出版方：英国皇家化学会。[127] Wang, X., Thomas, J. D., Piechocki,R. J., Kapoor, S., Santos-Rodríguez,R., and Parekh, A. 用于开放式无线接入网资源分配的自我博弈学习策略。《计算机网络》206 (2022年), 108682。出版商：爱思唯尔。

[128] Wang, Y., Sun, M., Wang, H., and Sun, Y. 结合时间卷积网络与蒙特卡洛树搜索的知识图谱补全模型研究

搜索。《工程数学问题》2022（2022）。出版商：Hindawi。[129] Wang, Z., and Li, C. 通过前瞻搜索引导的强化学习进行通道剪枝。载于2022年 IEEE/CVF 冬季计算机视觉应用会议（WACV）（2022），第3513–3524页。

[130] Webster, J., and Watson, R. T. 分析过去以备未来：撰写文献综述。《MIS季刊》（2002），xiii–xxiii。出版商：JSTOR。

[131] Weingertner, P., Ho, M., Timofeev, A., Aubert, S., and Pita-Gil, G. 用于运动规划的强化学习蒙特卡洛树搜索。载于2020年IEEE 第23届智能交通系统国际会议（ITSC）（2020），第1–7页。[132] Weininger, D. SMILES，一种化学语言和信息系统。1. 方法论和编码规则简介。《化学信息与计算机科学杂志》28卷1期（1988），31–36。出版商：ACS 出版社。[133] Wu, Q., Feng, Q., Ren, Y., Xia, Q.,Wang, Z., and Cai, B. 一种基于强化学习的电池储能系统智能预防性维护方法。《IEEE工业信息学汇刊》17卷12期（2021），8254–8264。[134] Wu, Y., and Yao, L. 基于深度强化学习的不完全信息下三维装箱问题研究。载于2021年电子商务与电子管理国际会议（ICECEM）（2021），第38–42页。

[135] Xiangyu, Z., Kexin, Z., and Yongjin, L. 机器学习赋能的应用专用金属有机框架定制设计。《美国化学会应用材料与界面》12卷1期（2020年1月），734–743。[136] Xing, Z., and Tu, S. 图神经网络辅助的蒙特卡洛树搜索

旅行商问题求解方法。*IEEE Access 8*（2020年），108418–108428。

[137] Xing, Z., Tu, S., and Xu, L. 通过蒙特卡洛树搜索和深度神经网络求解旅行商问题。*arXiv*预印本 *arXiv:2005.06879*（2020年）。

[138] Xu, R., and Lieberherr, K. 为组合优化问题学习自我博弈智能体。在第*18*届自主智能体与多智能体系统国际会议论文集（2019年），第2276–2278页。

[139] Xu, R., and Lieberherr, K. 组合优化问题的自我对弈智能体学习。*The Knowledge EngineeringReview 35*（2020年）。出版商：剑桥大学出版社。

[140] Xu, R., and Lieberherr, K. 使用神经蒙特卡洛树搜索进行在线模型检测。在*NASA*形式化方法研讨会（2022年），施普林格，第557–575页。

[141] Xu, R., and Lieberherr, K. 利用深度学习和蒙特卡洛树搜索解决QSAT问题。发表于《科学与信息会议论文集》(2022年)，施普林格，第45–58页。[142] Yan, M., Feng, G., Zhou, J., and Qin,S. 基于多智能体强化学习的智能多无线接入技术接入。*IEEE*车辆技术汇刊 *67*，第5期（2018年），4539–4551页。

[143] Yang, J., Hou, X., Hu, Y. H., Liu, Y., and Pan, Q. 一种采用改进的上置信界树搜索的主动多碎片清除任务规划强化学习方案。*IEEEAccess 8*（2020年），108461–108473页。

[144] Yang, R., Zou, X., Nie, Z., and Yin, C.基于深度强化学习的通信节点车辆部署研究。发表于*2018*年第五届*IEEE*云计算与智能系统国际会议（*CCIS*）（2018），第484–487页。

[145] Yang, X., Zhang, J., Yoshizoe, K., Terayama, K., and Tsuda, K. ChemTS：用于从头生成分子的高效Python库。先进材料科学与技术*18*，第1期（2017），第972–976页。出版商：泰勒与弗朗西斯出版集团。

[146] Yin, C., Yang, R., Zhu, W., Zou, X., and Zhang, J. 利用深度强化学习对紧急通信网络进行最优规划。电子信息通信工程师学会通信汇刊 *104*，第1期（2021年），20–26页。出版方：电子信息通信工程师学会。

[147] Yu, T., Huang, J., and Chang, Q. 基于强化学习掌握人机协同装配中的工作序列。*IEEE Access 8*（2020年），163868–163877。

[148] Zhang, B., Zhang, X., Du, W., Song,Z., Zhang, G., Zhang, G., Wang,Y., Chen, X., Jiang, J., and Luo,Y. 化学信息引导的分子图作为机器学习逆合成规划的反应描述符。美国国家科学院院刊 *119*，41（2022年），e2212711119。出版方：美国国家科学院。

[149] Zhang, C., Song, W., Cao, Z., Zhang,J., Tan, P. S., and Chi, X. 通过深度强化学习学习作业车间调度中的派工。神经信息处理系统进展 *33*（2020年），1621–1632。

[150] Zhang, J., Chen, H., Song, S., and Hu, F. 基于强化学习的自动泊车系统运动规划。*IEEE Access 8*（2020年），154485–154501。

[151] Zhang, J., Zhou, K., and Schelter,S. Alphajoin：连接顺序选择，借鉴 AlphaGo。发表于*CEUR*研讨会论文集（2020年），第2652卷。

[152] Zhang, K., Wu, J., Yoo, H., and Lee, Y. 基于机器学习的离子液体定制化设计方法：应用于二氧化碳捕集。分离与

《净化技术》第*275*卷（2021年），119117。出版商：爱思唯尔。

[153] Zhang, M., Huang, Q., Wang, S., andWang, Z. 基于深度强化学习的 LDPC 码构造。载于*2018*年第*10*届国际无线通信与信号处理会议（*WCSP*）（2018年），第1–4页。

[154] Zhang, Y., Wang, W., Zhang, P., and Huang, P. 基于强化学习的蜂窝卫星自重构任务规划。《IEEE 航空航天与电子系统杂志》37卷第6期（2022年），38–47页。[155] Zhong, S., Zhao, Y., and Chen, J.学习恢复稀疏信号。载于2019年第57届奥尔顿通信、控制与计算年会（Allerton）（2019年），第995–1000页。[156] Zombori, Z., Urban, J., and Brown, C. E. Prolog 技术强化学习证明器。载于国际自动推理联合会议（2020年），施普林格，第489–507页。[157] Zombori, Z., Urban, J., and Olšák,M. 熵在指导连接证明器中的作用。载于国际自动推理分析表格与相关方法会议（2021年），施普林格，第218–235页。[158] Zou, L., Xia, L., Ding, Z., Yin, D.,Song, J., and Liu, W. 通过强化学习使前 n 项推荐结果多样化。载于先进应用数据库系统国际会议（2019年），施普林格，第104–120页。[159] Zou, X., Yang, R., and Yin, C. 不完整可视地形下基于 AlphaZero 技术的 MANET 节点布局模型优化研究。载于2019年国际人工智能与计算机科学会议论文集（2019年），第562–565页。

[160] Zou, X., Yang, R., Yin, C., Nie, Z., and Wang, H. 采用 AlphaZero 算法部署战术通信节点车辆。《IET通信》14，9（2020年），1392–1396。出版方：Wiley在线图书馆。[161] Świechowski, M., Godlewski, K., Sawicki, B., and Mańdziuk, J. 蒙特卡洛树搜索：近期改进与应用综述。人工智能综述 （2022年），1–66。出版方：施普林格。

附录 A 实现

表 A1 神经网络方法的公开可用实现 蒙特卡洛树搜索。

来源	实现链接
[5]	https://github.com/baifanxxx/NPMO-Rearrangement
[25]	https://github.com/sekijima-lab/mermaid
[32]	https://github.com/werner-duvaud/muzero-general (muzero general)
[35]	https://github.com/HOL-Theorem-Prover/HOL
[36]	https://github.com/MolecularAI/aizynthfinder
[37]	https://mgit.cs.uni-saarland.de/timopgros/carmanufacturin
[82]	https://github.com/bpriviere/决策 -
[94]	https://github.com/lynshao/AlphaSeq
[106]	参见 [145]
[109]	https://github.com/NREL/rlmolecule
[113]	https://userinterfaces.aalto.fi/adaptive/
[119]	https://github.com/HanqingWangAI/SceneMover
[131]	https://github.com/PhilippeW83440/MCTS-NNET
[135]	参见 [145]
[136]	https://github.com/CMACH508/2020年-GNN-MCTS-TSP
[145]	https://github.com/tsudalab/ChemTS
[148]	https://github.com/zbc0315/synprepy
[156]	https://github.com/zsoltzombori/plcopprolog
[157]	https://github.com/zsoltzombori/plcop
[158]	https://github.com/zoulixin93/FMCTS