形式机制可解释性：具备可证明保证的自动化电路发现

ItamarHadad[1], **GuyKatz**[1], **Shahaf Bassan**[1]耶路撒冷希伯来大学计算机科学与工程学院[1]itamar.hadad@mail.huji.ac.il, g.katz@mail.huji.ac.il, shahaf.bassan@mail.huji.ac.il

arXiv:2602.16823v1 [cs.LG] 18 Feb 2026

摘要

自动化电路发现是机制可解释性中的核心工具，用于识别神经网络中负责特定行为的内部组件。尽管以往的方法已取得显著进展，但它们通常依赖启发式方法或近似，且无法为所得电路在连续输入域上提供可证明保证。在本研究中，我们利用神经网络验证领域的最新进展，提出一套自动化算法，可生成具备可证明保证的电路。我们重点关注三类保证：(i) 输入域鲁棒性，确保电路在连续输入区域内与模型保持一致；(ii) 鲁棒补丁，认证电路对齐在连续补丁扰动下仍成立；(3) 最小性，对广泛多样的简洁性概念进行形式化并加以刻画。有趣的是，我们揭示了这三类保证之间一系列多样且新颖的理论联系，这些联系对算法的收敛性具有关键影响。最后，我们使用最先进的验证器，在多种视觉模型上开展实验，结果表明，与标准电路发现方法相比，我们的算法生成的电路具有显著更强的鲁棒性保证——这为可证明电路发现奠定了原则性基础。

1 引言

Transformer 等变革性架构推动了神经网络的迅速发展，重塑了理论与应用领域。伴随这场变革，可解释性 已成为核心研究方向（Zhang 等人，2021；Räuker 等人，2023）；近年来，研究者的关注点进一步转向机制可解释性 （MI），旨在对神经网络进行逆向工程，将其分解为人类能够理解的组件和功能模块（Olah 等人，2020；Olah，2022；Zhao 等人，2024）。MI 提供细粒度的可解释性，可用于多种目的，包括提升透明度、可信度、安全性，以及支持其他应用（Bereska & Gavves，2024；Zhou 等人，2025）。

MI 面临的一项核心开放挑战是电路发现 （Olah 等人，2020），其目标是在神经网络中识别称为电路的子图，这些子图驱动模型的特定行为。近期研究提出了多种方法（Wang 等人，2023；Conmy 等人，2023；Rajaram 等人，2024），在定义域（文本与视觉）、补丁方法（零值、均值、采样）以及人工步骤与自动化步骤的权衡方面各不相同。然而，尽管取得了显著进展，目前大多数电路发现算法仍依赖启发式方法或近似方法，缺乏对电路忠实度的严格保证，尤其是在连续扰动定义域下（Adolfi 等，2025；Miller 等人，2024；Méloux 等人，2025）。这一局限令人担忧：即使是微小扰动也可能破坏电路忠实度；而电路发现又与安全性 考量密切相关（Bereska & Gavves，2024），因此此类保证至关重要。

我们的贡献。 为了解决这些问题，我们提出了一种新颖的算法框架，该框架建立在新兴领域神经网络验证近期令人振奋的进展之上（Wang 等人，2021；Zhou 等人，2024；Brix 等人，2024；Kotha 等人，2023；Ferrari 等人，2022），从而能够推导出在所关注的连续域上具有可证明保证的电路。

1.1 理论贡献

- 我们对一系列适用于整个连续域的新颖可证明保证进行了形式化定义，用于电路发现。这些保证包括：(i) 输入域鲁棒性，确保电路在连续输入区域内保持忠实；(ii) 修补域鲁棒性，回应对基于采样的消融的批评；以及 (iii) 一大类极小性保证，将早期概念拓展至准-、局部-、子集-和基数-极小性。
- 我们提出了新的理论证明，揭示了这三类保证之间的紧密联系。核心是电路单调性这一性质，它是优化算法极小性保证的基础，并阐明了这些保证成立的条件。我们还确立了电路与小型“阻断”子图之间至关重要的对偶性，从而能够高效地发现具有强得多的极小性保证的电路。

1.2 实证贡献

- 我们提出了一种框架，用于对神经网络及其电路中的输入鲁棒性和修补鲁棒性保证进行编码。该框架采用一种技术性的李生编码，将网络与其关联的电路或修补定义域一同编码，从而能够认证所需性质。
- 我们提出了一组新颖的自动化算法，这些算法能够保持鲁棒性保证的不变量，并证明每种算法都会收敛到满足我们各种最小性标准的电路。这些算法使计算成本与所得电路达到的最小性程度之间能够进行权衡。
- 我们使用神经网络验证领域的最先进方法 α–β-CROWN开展了大量实验，以推导出具备输入、修补和最小性保证的电路。我们在神经网络验证文献中常用的标准神经视觉模型上对这些电路进行了评估。与基于采样的方法相比，我们的框架能够认证鲁棒性，而即使是无穷小的扰动也会破坏基于采样的电路的忠实性。

总而言之，我们认为，这些贡献标志着在建立具备可证明保证的电路发现的理论与实证基础方面迈出了重要一步，并为广泛的未来研究方向铺平了道路。

2 预备知识

2.1 符号

令 $f_G\colon \mathbb{R}^n \to \mathbb{R}^d$ 表示一个神经网络，其中 $G := \langle V,\ E\rangle$ 表示其计算图。 G 的具体结构——也就是说，每个节点和边分别对应什么（例如神经元、注意力头、位置嵌入、卷积滤波器）——由网络架构和用户选择的粒度共同决定。电路 C 定义为一个子图 $C \subseteq G$，由被假设为驱动模型在某项任务上行为的组件构成。每个这样的电路自然会导出一个函数 $f_C\colon \mathbb{R}^n \to \mathbb{R}^d$，该函数通过将 f_G 限制到 C中的组件而得到。

在电路发现中，补集 $C := G \setminus C$ 通常被固定为常量激活值，这种做法称为修补，其变体包括零值修补和均值修补。设 f_G 有 $L \in \mathbb{N}$层，其激活空间为 V_i （ $i \in [L]$），并设 $\mathcal{X} \subseteq \mathbb{R}^n$ 为输入域。对于$\mathbf{x} \in \mathcal{X}$，将第 i 层的激活值记为 $h_i(\mathbf{x}) \in V_i$，将可达激活空间记为 $\mathcal{H}_G(\mathcal{X}) = \{(h_1(\mathbf{x}), \ldots, h_L(\mathbf{x}))\colon \mathbf{x} \in \mathcal{X}\} \subseteq V_1 \times \cdots \times V_L$。我们用 $\mathcal{H}_C(\mathcal{X})$ 表示 C 上的部分可达激活空间。对于部分激活赋值 $\alpha \in \mathcal{H}_C(\mathcal{X})$，我们用 $f_C(\mathbf{x} \mid C = \alpha)$ 表示通过 $f_C(\mathbf{x})$ 进行推理；该推理由 C 的组件构成，同时将 C 的激活值固定为 α 中的值。

2.2 神经网络验证

考虑一个具有任意逐元素非线性激活函数的通用神经网络 f_G。目前已有许多工具可用于形式化验证此类网络的属性，其中对抗鲁棒性是研究最广泛的属性（Brix 等人，2024）。形式上，神经网络验证问题可表述如下：

神经网络验证（问题陈述）：
输入：一个神经网络模型 f_G，其中 $\mathbf{y} := f_G(\mathbf{x})$，给定一个输入规范 $\psi_{\text{in}}(\mathbf{x})$，以及一个不安全的输出规范 $\psi_{\text{out}}(\mathbf{y})$。**输出**：如果存在某个 $\mathbf{x} \in \mathbb{R}^n$，使得 $\psi_{\text{in}}(\mathbf{x})$ 和 $\psi_{\text{out}}(\mathbf{y})$ 同时成立，则为否；否则为是。

目前已有多种现成的神经网络验证器（Brix 等人，2024）。当输入约束 $\psi_{in}(\mathbf{x})$、输出约束 $\psi_{out}(\mathbf{y})$和模型 f_G 均为分段线性（例如，采用 ReLU 激活）时，可以精确求解验证问题（Katz 等人，2017）。在实践中，为提高效率，通常会对问题进行松弛处理，并将输出限制在考虑了近似误差的范围内。

3 电路发现的可证明保证

3.1 输入定义域保证

在本小节中，我们介绍算法旨在满足的第一组保证——具体而言，是针对输入扰动的可证明鲁棒性 。评估电路忠实性的一个核心挑战在于，即使它在某一点或有限多个点上与模型相符，微小的输入扰动也可能迅速破坏这种一致性。为克服这一局限，我们的第一个定义考虑的电路不仅在离散点集上忠实于模型，而且在整个无限的连续输入集 上也具有可证明鲁棒性 。

定义 1（输入鲁棒电路）。 给定一个神经网络 f_G 以及连续域的并集$\mathcal{Z} \subseteq \mathbb{R}^n$（例如，围绕一组离散点 $\{\boldsymbol{x}_j\}_{j=1}^k$、半径为 ϵ_p 的 ℓ_p-球的并集），当且仅当以下条件成立时，我们称子图 $C \subseteq G$ 是关于 $\langle f_G$、$\mathcal{Z}\rangle$、施加于补集 $C: = G \setminus C$的固定修补向量 α 以及容差水平 $\delta \in \mathbb{R}^+$的输入鲁棒电路：

$$\forall \boldsymbol{z} \in \mathcal{Z} := \bigcup_{j=1}^{k} \mathcal{B}_{\epsilon_p}^p(\boldsymbol{x}_j). \quad \|f_C(\boldsymbol{z} \mid \overline{C} = \alpha) - f_G(\boldsymbol{z}))\|_p \leq \delta, \\ s.t.\ \mathcal{B}_{\epsilon_p}^p(\boldsymbol{x}_j) := \{\boldsymbol{z}_j \in \mathbb{R}^n \mid \|\boldsymbol{z}_j - \boldsymbol{x}_j\|_p \leq \epsilon_p\} \tag{1}$$

通过验证认证电路输入鲁棒性。 神经 l 网络 k verificati的性质 i是典型的ically 编码在一个单一的模型 f_G上，而电路输入鲁棒性性质（定义 1）要求评估模型图 G 和电路 $C \subseteq G$。为了解决这一问题，我们提出了一种新方法来认证定义 1 中的性质，采用孪生编码网络 f_G。具体来说，我们复制电路 $C' := C$，并将其与 G “堆叠”起来，形成一个组合模型 $G \sqcup C'$，并使用一个共享的输入层。这会产生一个函数 $f_{(G \sqcup C')}: \mathbb{R}^n \to \mathbb{R}^{2d}$。激活值 — 复制体中的非电路组件 $C' :=$ $G \setminus C'$ 被固定为常数 α，因此对于任意$\mathbf{z} \in \mathbb{R}^n$ 推理为 $f_{(G \sqcup C')}(\mathbf{z} \mid C' = \alpha)$，从而能够对组合 模型直接进行定义 1 的认证。输入约束 $\psi_{in}(\mathbf{x})$ 将$\mathbf{x}$ 限制在 $\mathcal{Z}$ 内，而输出约束 $\psi_{out}(\mathbf{y})$ 则限制 C' 与 G 的逻辑值之间的距离度量。该编码的更多详细信息见附录 E。

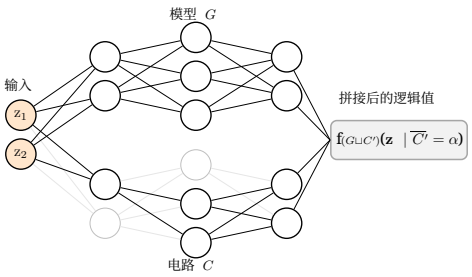


图1：用于证明定义 1 中保证的孪生编码示意图。

3.2 修补定义域保证

电路发现中的一个核心挑战，在于决定如何为电路的互补激活赋值——这一过程称为修补。修补的目标是替换这些值，以分离出电路的贡献。以往研究考察了几种方法，包括零值修补；由于这些值若在训练期间从未出现过，就可能属于分布外，因此这种方法受到任意性的批评（Conmy 等人，2023；Wang 等人，2023）。其他策略包括均值修补（Wang 等人，2023），以及从离散输入分布中采样（Conmy 等人，2023）。然而，这些方法仍依赖于在一组离散 采样输入上评估互补激活，因此可能无法推广至连续 定义域：即使只是对

修补方案进行微小扰动，也可能损害电路的忠实性。类比于输入鲁棒性，我们提出修补鲁棒性：要求电路在连续输入域中，面对互补激活所受的一整个可证明范围内的可行扰动时，始终保持其忠实性。

定义 2（修补鲁棒回路）。 给定一个神经网络 f_G、一个连续输入域$\mathcal{Z} \subseteq \mathbb{R}^n$和一个参考输入集 $\{\boldsymbol{x}_j\}_{j=1}^k \subseteq \mathcal{X}$，若对于每个属于 $\{\boldsymbol{x}_j\}_{j=1}^k$ 的$\boldsymbol{x}^j$，满足以下条件，则称 $C \subseteq G$ 是相对于 $\langle f_G、\ \mathcal{Z}\rangle$ 和容差水平 $\delta \in \mathbb{R}^+$ 的修补鲁棒回路：

$$\forall \alpha \in \mathcal{H}_{\overline{C}}(\mathcal{Z}).\quad \|f_C(\boldsymbol{x}_j \mid \overline{C} = \alpha) - f_G(\boldsymbol{x}_j)\|_p \leq \delta \tag{2}$$

通过验证认证电路补丁鲁棒性。 与输入鲁棒性类似，我们在此提出一种新方法，通过对 f_G 使用孪生编码来认证定义 2 中的性质。具体而言，我们复制电路 $C' := C$，并将其与 G“堆叠”起来，但此时 G 和 C' 的输入域互不相交，从而得到 $f_{(G \sqcup C')}$： $\mathbb{R}^{2n} \to \mathbb{R}^d$。我们通过固定 C' 和 G 的激活值，使其等于评估 f_G**(z)** 时 C' 在 $\mathcal{H}_C$**(z)** 中取得的激活值。因此，**(x,z)** $\in \mathbb{R}^{2n}$ 的推理结果为 $f_{(G \sqcup C')}$**(x** $\mid C' =$ $\mathcal{H}_C$**(z))**。随后，我们设置输入约束，将**z** 限制在 $\mathcal{Z}$ 内，并设置输出约束，以限制 $\|\cdot\|_p$ 与 C' 和 G 的逻辑值之间的距离。详细信息见附录 F。我们指出，输入鲁棒性和修补鲁棒性也可以同时进行认证

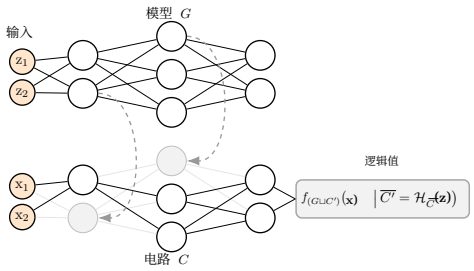


图 2：用于认证定义 2 中保证的孪生编码示意图。灰色神经元表示非电路单元，其激活值被替换为完整模型中的对应激活值。

o可在单次验证查询中实现这一点，方法是扩展孪生编码（见附录 G）。

4 从电路到最小电路

文献中有一种常见惯例：较小的 电路（即电路大小较小）通常被认为比更大的电路更具可解释性（Mueller 等人，2025；Adolfi 等，2025；Chowdhary 等人，2025；Wang 等人，2023；Bhaskar 等人，2024；Shi 等人，2024；Conmy 等人，2023）。因此，最小性 是一项重要的附加保证（Adolfi 等，2025；Chowdhary 等人，2025；Shi 等人，2024；Mueller 等人，2025）。尽管许多研究都致力于寻找最小电路，近期研究指出，“最小性”本身有不同形式（Adolfi 等，2025），从较弱的准最小性概念到较强的基数最小性概念。在本研究中，我们将这一谱系扩展为四种主要形式，并提供严谨的证明，将它们与不同的优化

算法。

4.1 极小性保证

在本小节中，我们介绍四种核心的极小性概念。由于必须相对于何为“有效”或忠实电路来定义极小性，我们首先指定一个通用忠实性谓词， $\Phi(C, G)$。给定模型 f_G 的计算图 G 中的一个电路 $C \subseteq G$，$\Phi(C, G)$ 会在 C 在某个关注条件下忠实时返回 *True*，否则返回*False*。 Φ 的实例可以包括电路发现中使用的标准基于采样的度量，例如要求 C 与 G 之间的均方误差或 KL 散度低于某个阈值 τ（Conmy 等人，2023）。或者， Φ 也可以体现我们在连续域上成立的可证明度量（第 3 节），例如，定义$\Phi(C, G)$以要求输入鲁棒性和/或修补鲁棒性。

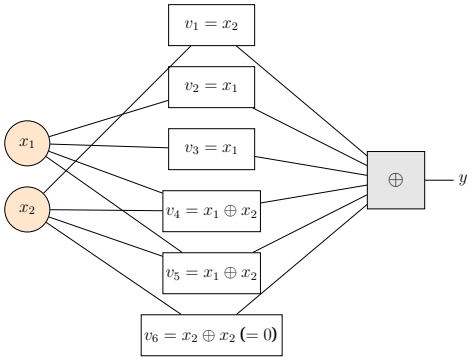


图 3：一个布尔玩具电路。

例如，考虑图3中展示的一个布尔玩具电路，作为贯穿全文的示例。对于 simplicity，我们假设电路中的每个节点对应于模型的一个组件。

该网络接收输入（x_1, x_2） $\in \{0,1\}^2$，其输出由异或聚合，得到f_G（x_1, x_2） $= v_1 \oplus v_2 \oplus \cdots \oplus v_6 = x_2$，因为

$$f_G(x_1,\ x_2) := x_2 \oplus x_1 \oplus x_1 \oplus (x_1 \oplus x_2)\ \oplus (x_1 \oplus x_2)\ \oplus (x_2 \oplus x_2)$$
$$= x_2 \oplus \underbrace{(x_1 \oplus x_1)}_{=0} \oplus \underbrace{[(x_1 \oplus x_2)\ \oplus (x_1 \oplus x_2)]}_{=0} \oplus \underbrace{(x_2 \oplus x_2)}_{=0} = x_2.$$

为简单起见，我们将忠实性谓词设定为考虑一种非常强的条件：对于每个布尔输入，电路与模型严格相等。换言之，对于所有 x_1 和 x_2 属于$\{0,1\}^2$的情况，均有 $\Phi(C,\ G) := f_C(x_1,\ x_2) = f_G(x_1,\ x_2) = x_2$。附录 C.3 对此示例进行了更详细的说明。

我们从最弱的最小性形式开始，逐步推进，直到达到最强形式。第一个概念是由（Adolfi 等，2025）提出的准最小性，它将 C 定义为一个包含“某个破坏点”的子集：移除该组件后，电路的忠实性会被破坏：

定义 3（准最小电路）。给定 G，关于 Φ 的准最小电路 $C \subseteq G$ 是满足 Φ（C、G）的电路（“C 具有忠实性”），并且存在某个元素（即节点/边） i 属于C，使得 Φ（$C \setminus \{i\}$、G）不成立。

在我们的示例中，$C = \{v_1, \ldots, v_5\}$ 是准最小的：它满足 $v_1 \oplus \cdots \oplus v_5 = f_G$，但移除某个关键节点（例如 v_1）会破坏这一等式。然而，我们注意到，Adolfi 等，2025 年提出的准最小性概念可以得到加强：不再要求存在单一破坏点，而是可以要求每个组件都充当一个破坏点。遵循优化文献中的惯例，我们将这一更强的概念称为局部最小性。

定义 4（局部最小电路）。给定 G，关于 Φ的局部最小 电路 $C \subseteq G$ 是这样一个电路：Φ（C，G）成立，并且对于 C 中的任意元素 i，Φ（$C \setminus \{i\}$，G）为假。

在我们的示例中，$C = \{v_1, v_2, v_3\}$ 是局部最小的，因为该谓词对 $\{v_1, v_2, v_3\}$ 成立，但对 $\{v_1, v_2\}$、$\{v_1, v_2\}$ 和 $\{v_2, v_3\}$ 不成立。然而，尽管局部最小性强于准极小性，它仍有一个局限。虽然从电路 C 中移除任意单个 组件都会破坏它，但移除多个 组件仍可能使 C 保持有效，从而给人一种误导性的最小性印象。例如，在我们的局部最小电路 $C = \{v_1, v_2, v_3\}$ 中可以看到，虽然移除任意单个组件（即 v_1、v_2 或 v_3）会破坏忠实性，但移除一对 $\{v_2, v_3\}$（即仅保留 v_1）仍会得到一个忠实电路。为解决这一问题，我们定义一个更强的概念：子集极小性，它要求组件的每一个子集 都是破坏点。

定义 5（子集最小电路）。给定 G，关于 Φ 的子集最小电路 $C \subseteq G$ 是一个使 $\Phi(C, G)$ 成立的电路，并且对于任意子图 $S \subsetneq C$，$\Phi(C \setminus S, G)$ 为假。

在贯穿示例中，$C = \{v_2, v_4\}$ 是一个子集最小电路，因为 $v_2 \oplus v_4 = x_1 \oplus (x_1 \oplus x_2) = x_2$，这是由于每个真子集（$\{v_2\}, \{v_4\}$）都无法计算 f_G。我们注意到，即便这一显著更强的子集极小性概念，也并不必然得到基数绝对最低的子集。为解决这一局限，最后一个概念引入了最强形式：基数极小电路，它对应于最小性的全局最优解。

定义 6（基数极小电路）。给定 G，关于Φ 的基数极小电路 $C \subseteq G$ 是一个使 $\Phi(C, G)$ 为真、且具有最低基数 $|C|$ 的电路（即不存在电路$C' \subseteq G$ 使得 $\Phi(C', G)$ 为真且 $|C'| < |C|$）。

在这里，$\{v_1\}$ 是基数最小的，因为 $v_1 = x_2$ 与 f_G 函数等价，且不存在更小的忠实电路。

4.2 局部极小电路和准极小电路算法

在本小节中，我们介绍用于发现具有可证明保证的电路的优化算法。在以往电路发现框架的基础上，我们展示了修改或验证优化目标如何改变所得保证。我们还建立了基于连续鲁棒性保证的目标与不同最小性概念之间的理论联系。我们从一个标准的贪心算法（算法 1）开始：

算法 1 贪心电路发现迭代搜索

1: **输入**模型 f_G，电路忠实性谓词 Φ2: $C \leftarrow G$ 遵循某一给定的元素排序（例如，反向拓扑排序）3: **对所有** $i \in C$ **执行**4: **如果** $\Phi(C \setminus \{i\}, G)$ **则**5: $C \leftarrow C \setminus \{i\}$6: **结束如果**7: **结束循环**8: **返回** C

算法 1 描述了一种标准的贪心过程：从完整模型图 G 出发，只要忠实性谓词成立，就迭代地移除组件。尽管许多电路发现方法都以这一结构为基础，我们还考虑了一种穷举变体，即附录 C 中的算法 3；它反复遍历剩余组件并尝试进一步移除，仅当每一个单个组件都是关键组件时才停止。这保证了以下结论：

命题1. 给定任意模型 f_G 与忠实性谓词 Φ，算法 *1* 会访问一个准最小电路，而算法 *3* 收敛到一个局部最小电路 C，二者均相对于 Φ。

我们指出，对 $\Phi(C \setminus \{i\}, G)$ 的每次评估都可能代价高昂，这取决于谓词 Φ（例如，认证输入或修补鲁棒性）。为缓解这一问题，可以采用一种更轻量的最小性概念——（Adolfi 等，2025）的准最小电路——它只需要对数次而非线性次调用。为完整起见，我们正式给出并扩展了他们的二分搜索过程，即附录 C 中的算法 4。算法 4 遵循与算法 1 类似的过程，但采用二分搜索而非迭代搜索。因此，它得到的是一种更弱的最小电路概念，同时所需查询更少。由此我们可以建立如下命题：

命题 2. 虽然算法 *4* 仅需 $\mathcal{O}(\log |G|)$ 次评估即可收敛到准最小电路 $\Phi(C, G)$（*Adolfi* 等，*2025*），但算法 *1* 需要 $\mathcal{O}(|G|)$ 次评估才能访问准最小电路，而算法 *3* 则需 $\mathcal{O}(|G|^2)$ 次评估才能收敛到局部最小电路。

最后，我们指出，算法 1、3 和 4 收敛到的都只是相对“较弱”的最小性形式。即便算法 3 提供了更强的局部最小性保证，也可能仍不够：虽然移除任一单个元素 $C \setminus \{i\}$ 都会破坏忠实性，但同时移除两个元素 $C \setminus \{i, j\}$仍可能得到忠实电路。这削弱了 C的“最小性”，并表明这两种算法都无法保证更强的子集最小性。

命题 3. 存在无穷多个 f_G 和 Φ 的配置，使得算法 *1*、*3* 和 *4*无法收敛到关于 Φ 的子集最小电路 C。

4.3 电路单调性性质及其对最小性的影响

为了解决算法收敛到“坏”局部极小值的问题，我们指出忠实性谓词 Φ 的一个关键性质，它对更强的极小子集有重要影响——单调性：

定义 7. 我们称电路忠实性谓词 Φ 是单调的，当且仅当对于任意 $C \subseteq C' \subseteq G$，若 $\Phi(C, G)$ 为真，则 $\Phi(C', G)$ 为真。

直观而言，单调性意味着，一旦电路 C 满足 $\Phi(C, G)$（即它是“忠实的”），随着元素不断添加，该条件仍会成立。换言之，扩大电路永远不会破坏其忠实性。这一性质对算法 1 至关重要，因为它是更强最小性保证的基础：

命题 4. 若 Φ 是单调的，则对于任意模型 f_G，算法 *1* 都会收敛到关于 Φ 的子集最小电路 C。

单调性条件。有趣的是，我们建立了一个新的联系：第 3 节中概述的针对输入和修补域的保证，与 Φ 的单调行为之间存在联系：

命题 5. 令 $\Phi(C, G)$ 表示验证 C 关于 $\langle f_G, \mathcal{Z} \rangle$（定义 *1*）是否为输入鲁棒，并同时关于 $\langle f_G, \mathcal{Z}' \rangle$（定义 *2*）是否为修补鲁棒。那么，若 $\mathcal{Z} \subseteq \mathcal{Z}'$ 且 $\mathcal{H}_G(\mathcal{Z}')$ 在拼接下封闭，则 Φ是单调的。

直观而言，命题 5 表明：如果修补域 $\mathcal{Z}'$ 包含输入域 $\mathcal{Z}$，且激活空间 $\mathcal{H}_G(\mathcal{Z}')$ 在拼接下封闭，即将任意两个部分激活拼接后仍位于 $\mathcal{H}_G(\mathcal{Z}')$ 之内——则忠实性谓词 Φ 是单调的。这引入了一个新的评估条件类别，这些条件在构造上是单调的，从而得到更强的最小电路。

命题 6. 若将条件 $\Phi(C, G)$ 设为验证 C 相对于 $\langle f_G, \mathcal{Z}\rangle$ 是否具有输入鲁棒性（定义 *1*），并且相对于 $\langle f_G, \mathcal{Z}'\rangle$ 是否具有修补鲁棒性（定义 *2*），那么如果 $\mathcal{Z} \subseteq \mathcal{Z}'$ 和 $\mathcal{H}_G(\mathcal{Z}')$ 在拼接下封闭，算法 *1* 将收敛到一个子集最小电路。

4.4 从包含关系极小电路到基数极小电路

尽管 Φ 的单调性提供了更强的子集极小性保证，但仍无法确保收敛到全局最小电路（即基数最小电路）。一种朴素的做法是枚举所有 $C \subseteq G$，验证 $\Phi(C, G)$，然后选择基数 $|C|$ 最小的那个，但即使对于规模适中的图，这种方法也会很快变得难以处理。

利用电路阻断集对偶性高效近似基数极小电路。在本小节中，我们利用如下思想：神经网络往往包含较小的"电路阻塞集"——即子图 $C' \subseteq G$，其改变后的激活会破坏任何排除它们的电路 C 的忠实性。我们证明了电路（在单调忠实性谓词下）与这些阻塞集之间的对偶性，从而得到一种新的算法构造，它能够近似——有时还能精确恢复——基数最小电路，且效率远高于朴素搜索。形式上，对于 f_G 和 Φ，电路阻塞集 是任意 $C' \subseteq G$，使得对所有 $C \subseteq G$，$\Phi(C \setminus C', G)$均不成立。由此得到一种植根于电路与阻塞集之间最小击中集（*MHS*）关系的对偶性：

命题 7. 给定某个模型 f_G和一个单调谓词 Φ，所有涉及 Φ 的电路阻塞集的最小击中集是一个基数最小电路 C，使得 $\Phi(C, G)$ 为真。此外，所有使 $\Phi(C, G)$ 为真的电路 $C \subseteq G$ 的最小击中集，是关于 Φ的基数最小阻塞集。

最小击中集这一经典的 NP 完全问题的定义见附录 B.7。尽管最小击中集对偶性涉及 NP 完全问题，但最小击中集在实践中往往可以借助混合整数线性规划或最大可满足性等现代求解器高效求解，因此这一对偶性十分强大（Ignatiev 等，2019a）。因此，类似的对偶性已成为形式化推理和可证明的可解释性方法的核心（Bacchus 与 Katsirelos，2015；Ignatiev 等，2019c；Bassan 与 Katz，2023；Liffiton 等，2016）。有了这一对偶定理，我们便可以设计一种算法，通常能够计算（或近似）基数最小电路：

算法 2使用最小击中集对偶性近似基数最小电路

1: **输入** 模型 f_G，忠实性谓词 Φ，$t_{\max} \in [|G|]$2: 阻塞集 $\leftarrow \emptyset$3: **对于** $t \leftarrow 1$**至** $t_{\max}$ **执行**4: $\mathcal{C}_t \leftarrow \{ S \subseteq G, \forall U \subseteq$ 阻塞集 : $|S| = t$， $U \not\subseteq S\}$5: **对于所有** $S \in \mathcal{C}_t$ **执行** ▷ 并行化

6: **如果** $\neg\Phi(G \setminus S, G)$ **则**7: 阻塞集 $\leftarrow$ 阻塞集 $\cup S$8: **结束 if**9: **结束 for**10: $C \leftarrow$最小击中集(阻塞集)11: **如果** $\Phi(C, G)$ **则**

返回 C12: **结束 if**13: **结束 for**

算法 2 利用命题 7，通过并行 遍历阻塞集，并计算每个集合的最小击中集，以获得一个电路 C。这为基数最小电路建立了下界，并通过连续精化收敛到最小电路。虽然在最坏情况下，阻塞集子集的数量可能过多，但在实践中，其大小通常是可处理的（见第 5 节），从而产生一个较低的 $t_{\max}$，并能够更高效地计算基数最小电路，或近似计算基数最小电路。以下命题对此进行了形式化表述：

命题 8. 给定一个模型 f_G 和一个单调谓词 Φ，算法 *2* 计算出一个子集 C，其大小是基数最小电路的下界，且对于该电路，$\Phi(C, G)$ 为真。当 t_{max} 的值足够大时，该算法会恰好 收敛到基数最小电路。

5 实验评估

实验设置。我们在神经网络验证文献中的标准基准测试上评估我们的方法（Brix 等人，2024；2023）：(i) MNIST，(ii) CIFAR-10，(iii) GTSRB，以及 (iv) TaxiNet（一个真实世界的数据集，曾用于基于验证的输入可解释性研究（Wu et al., 2023a；Bassan et al., 2025a））。对于神经网络验证，我们使用最先进的 α，β-CROWN 验证器（Bak 等人，2021；Müller 等人，2022a；Brix 等人，2024）；对于最小击中集，我们使用 RC2（Ignatiev 等，2019a）。为确保一致性，我们在先前验证研究中的模型架构上评估结果，尤其是神经网络验证竞赛（VNN-COMP）中的模型架构（Brix 等人，2024）；完整的架构详细信息见附录 D。为平衡电路发现的难度与人类可解释性，我们将 MNIST 的电路粒度设为神经元，将 CIFAR-10、GTSRB 和 TaxiNet 的电路粒度设为卷积滤波器。可证明变体和基于采样的变体均采用标准的*logit*差 度量（Conmy 等人，2023）。更多详细信息见附录 E、F 和 G。

5.1 具有输入鲁棒性保证的电路发现

我们首先评估方法的连续输入鲁棒性保证。我们比较算法 1 的两个变体：(i) 标准的基于采样的方法，在该方法中，通过在采样输入上应用容差为 δ 的 logit 差值谓词来评估忠实性；以及 (ii) 一种可证明的方法，该方法通过第 3.1 节的孪生编码，认证在整个连续输入域内 logit 差值始终低于 δ。对于这两种方法，我们报告跨越连续输入邻域的电路大小和鲁棒性，统计 100 个批次的结果（每个批次对应一个电路）。我们使用与 VNN-COMP（Brix 等人，2024）一致的 ϵ_p 个值来设置输入邻域（附录 E 中进行了变体设置），并在两种设置中均采用零值修补。结果（表 1）表明，可证明方法由于需要求解认证查询而速度较慢，但在电路规模相当的情况下实现了 100% 的鲁棒性，而基于采样的基线的鲁棒性则显著较低。图 4 展示了一个说明性示例。

表 1：算法 1 得出的电路结果，其中 Φ 通过对输入采样时的 logit差异进行界定来定义，或通过 验证孪生编码下的界限来定义。

数据集	基于采样的电路发现			可证明输入鲁棒的电路发现		
	时间（秒）	大小（$\lvert C\rvert$）	鲁棒性（%）	时间（秒）	大小（$\lvert C\rvert$）	鲁棒性（%）
CIFAR-10	0.23 ±0.52	16.47 ±9.08	**46.5** ±5.0	2970.85 ±874.23	19.18 ±10.16	**100.0** ±0.0
MNIST	0.31 ±0.89	12.56 ±2.30	**19.2** ±4.0	611.93 ±97.14	15.84 ±2.33	**100.0** ±0.0
GTSRB	0.11 ±0.33	28.91 ±4.69	**27.6** ±4.0	991.08 ±162.91	29.59 ±4.45	**100.0** ±0.0
TaxiNet	0.01 ±0.00	5.77 ±0.80	**9.5** ±3.0	180.00 ±40.39	6.82 ±0.46	**100.0** ±0.0

5.2 具有修补鲁棒性保证的电路发现

表 2：算法 1 得出的电路结果，其中 Φ 通过在零修补、均值修补下界定 logit 差异，或通过孪生编码验证该界限来定义。

数据集	零修补			均值修补			可证明修补鲁棒的修补		
	时间（秒）	大小 ($\lvert C\rvert$)	鲁棒性（%）	时间（秒）	大小 ($\lvert C\rvert$)	鲁棒性（%）	时间（秒）	大小 ($\lvert C\rvert$)	鲁棒性（%）
CIFAR-10	0.1 ± 0.3	65.1 ±3.0	**46.4** ±6.0	0.0 ±0.0	64.1 ±3.6	**33.3** ±5.7	5408.5 ±1091.0	65.6 ±1.6	**100.0**
MNIST	0.1 ±0.3	20.0 ±1.5	**58.0** ±5.3	0.0 ±0.0	19.2 ±1.8	**55.7** ±5.3	714.9 ±207.1	17.0 ±2.3	**100.0**
GTSRB	0.3 ±0.9	32.6 ±4.2	**38.0** ±4.4	0.0 ±0.0	33.4 ±4.2	**40.5** ±4.5	2907.2 ±721.7	34.3 ±4.1	**100.0**
TaxiNet	0.0 ±0.1	5.8 ±0.8	**57.1** ±5.0	0.0 ±0.1	5.4 ±0.7	**63.3** ±4.9	175.7 ±52.7	5.4 ±0.6	**100.0**

为评估修补鲁棒性，我们研究算法 1 的三种变体，它们在不同修补方案下强制有界的logit差值：(i) 零值修补、(ii) 均值修补，以及 (iii) 一种经过认证的

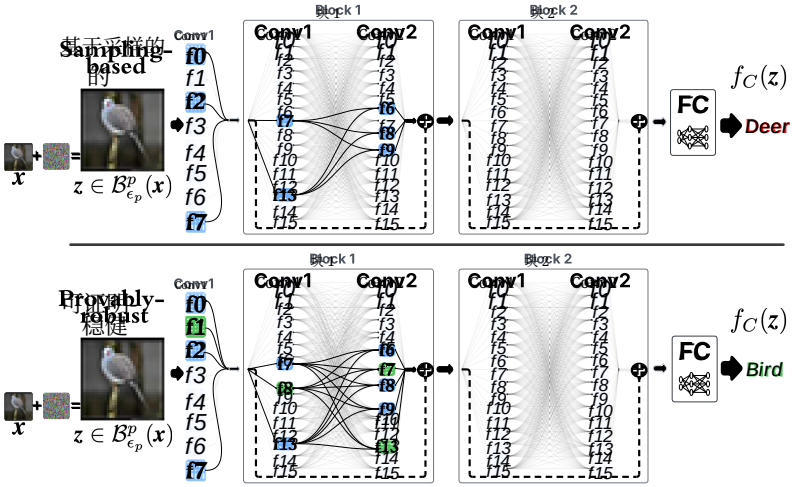


图 4：CIFAR-10 上 ResNet 在滤波器层级的电路示例。每层中的滤波器均已编号，非电路滤波器以灰色显示，残差连接以虚线表示。我们比较基于采样的发现方法与可证明鲁棒变体所发现的电路，并用绿色标出可证明回路独有的组件。

变体；该变体使用孪生编码（第 3.2 节），在连续修补域上一致地验证该界。随后，在与认证设定相同的连续域判据下，评估通过零值或均值修补找到的电路。结果见表 2。在标准修补（零值/均值）下找到的电路对修补域的变化敏感，鲁棒性较低；而经过验证的方法能够认证这一性质，并达到 100% 的鲁棒性。尽管计算成本更高（因为依赖验证），可证明方法仍能在相当的电路规模下提供强得多的鲁棒性。

5.3 探索电路发现的不同极小性保证

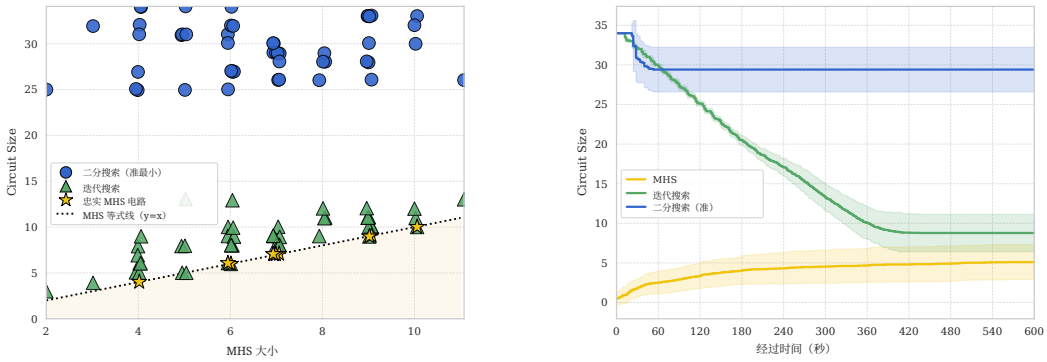
我们对第 4 节中的最小性保证及其与第 3 节中鲁棒性和修补保证的联系进行了实验。对于 Φ 谓词，我们使用双孪生编码（第 3.2 节）认证输入鲁棒性和修补鲁棒性，环境为 $\mathcal{Z} \subseteq \mathcal{Z}'$，并运行算法 1、2、4。算法 2 使用 $t_{\max} = 3$ 运行，将对比阻塞集枚举限制为大小不超过三的集合。实验表明，MHS 大小始终是电路大小的下界，没有任何电路的大小低于其 MHS。在一些运行中，算法 1 得到的电路恰好达到该界；一些 MHS 电路也被认证为忠实（即同时满足输入鲁棒性和修补鲁棒性），如图 5a 所示。

我们还注意到图 5b 中的效率–电路大小权衡：准最小（算法 4）过程终止最快，但在较大规模处趋于平稳，最小性较弱；迭代搜索（算法 1）以更高的运行时间获得更小的规模（更强的最小性）；而最小击中集（算法 2）循环最慢，却逐步逼近基数最小（最优）的电路大小。

6 相关工作

电路发现。我们的工作加入了机制可解释性中电路发现的近期努力（Olah 等人，2020；Elhage 等人，2021；Dunefsky 等人，2024），尤其是那些推进自动化算法的工作（Conmy 等人，2023；Hsu 等人，2024）。其他相关方向包括电路忠实度的指标（Marks 等人，2025；Hanna 等人，2024）、不同的修补策略（Jafari 等人，2025；Syed 等人，2024；Haklay 等人，2025；Miller 等人，2024；Zhang 与 Nanda，2024）、最小性标准（Chowdhary 等人，2025；Wang 等人，2023）以及应用（Yao 等人，2024；Sharkey 等人，2025）。

机制可解释性的理论研究。近期若干方向考察了机制可解释性的其他理论方面，这些方向往往与电路发现相关联。其中包括将机制可解释性置于更广泛的因果抽象框架中（Geiger 等人，2025；2024），并将其与分布式对齐搜索（DAS）联系起来（Wu



(a) 电路大小与 MHS 大小的关系　　(b) 电路大小随时间的变化

Fig图 5：(a) 电路大小与阻塞集大小的 MHS 大小对比，虚线表示相等线 $y{=}x$ as th e lower 界。(b) 电路大小在前 10 分钟内的收敛；阴影区域表示偏差 .

等人，2023b；Sun 等人，2025）；通过来自程序分析的抽象解释来分析习得的电路逻辑（Palumbo 等人，2025）、证明论（Miller 等人，2024；Wu 等人，2025）、统计识别（Méloux 等人，2025）以及复杂性理论（Adolfi 等，2025）。

神经网络验证与形式化可解释性。 我们对鲁棒性保证（输入与修补）的认证，建立在神经网络验证的快速进展之上（Brix 等人，2024；Wang 等人，2021；Zhou 等人，2024；Chiu 等人，2025；Müller 等人，2021；Singh 等人，2019年）。这些进展也被用于为基于输入的可解释性概念认证可证明保证（Wu et al., 2023a；Bassan et al., 2025a；Izza 等人，2024；Audemard 等人，2022b；La Malfa 等人，2021）（通常称为形式化可解释人工智能（Marques-Silva & Ignatiev，2022））。我们的工作首次将基于神经网络验证的策略用于机制可解释性中的电路发现。

7 局限性与未来工作

我们的框架有一项局限性，而所有在连续域上提供鲁棒性保证的方法也都存在这一局限性：依赖神经网络验证查询。尽管当前的验证技术在处理最先进模型时仍有局限，但其可扩展性正在迅速提升（Brix 等人，2024；Wang 等人，2021；Zhou 等人，2024）。我们的框架将这类工具创新性地整合到机制可解释性中，从而实现具有可证明保证的电路发现。因此，随着验证方法不断扩展至更大规模，我们的方法也将适用得更广泛；我们的广泛实验基于当前领先的验证器 α-β-CROWN，并在神经网络验证竞赛年度赛事中的标准基准测试上进行了评估。此外，我们提出的新理论结果涵盖了输入域、修补域和最小性的保证，为未来有关可证明电路发现的研究奠定了坚实基础，其中包括概率性和统计性形式的保证。

8 结论

我们提出了一个用于发现具备可证明保证的电路的框架，涵盖（i）连续输入定义域鲁棒性、（ii）连续修补定义域鲁棒性，以及（iii）多种形式的最小性。我们的方法核心是电路单调性的概念，它揭示了输入、修补与最小性保证之间深刻的理论联系，并为电路发现算法的收敛性奠定基础。我们利用神经网络验证领域的最新进展开展实验，结果证实，与电路发现中常用的标准基于采样的方法相比，我们的框架能够提供显著更强的保证。通过将电路发现与神经网络验证相结合，本研究朝着设计更安全、更可靠的电路迈出了创新的一步，并为未来可证明电路发现研究奠定了新的理论与算法基础。

致谢

本研究部分由欧盟资助（ERC，VeriDeL，101112713）。然而，文中表达的观点和意见仅代表作者本人，并不一定反映欧盟或欧洲研究委员会执行局的观点。欧盟和资助机构均不对这些观点和意见承担责任。本研究还获得了以色列科学基金会的资助（资助编号：558/24）。

参考文献

Federico Adolfi、Martina Vilas 和 Todd Wareham。《内部可解释性的电路发现的计算复杂性》。载于第*13*届国际学习表征大会（*ICLR*）论文集，2025。

Guy Amir、Shahaf Bassan 和 Guy Katz。《难以解释：分布内模型解释的计算难度》。载于第*27*届欧洲人工智能大会（*ECAI*）论文集，2024。

Alaa Anani、Tobias Lorenz、Mario Fritz 和 Bernt Schiele。《通过随机平滑实现像素级认证解释》。载于第*42*届国际机器学习大会（*ICML*）论文集，2025。

M. Arenas、P. Barceló、M. Romero 和 B. Subercaseaux。《决策树概率解释的计算》。发表于第 *36* 届神经信息处理系统会议（*NeurIPS*）论文集，第 28695–28707 页，2022。

Marcelo Arenas、Daniel Baez、Pablo Barceló、Jorge Pérez 和 Bernardo Subercaseaux。《模型可解释性符号语言基础》。发表于第 *35* 届神经信息处理系统会议（*NeurIPS*）论文集，第 11690–11701 页，2021。

G. Audemard、S. Bellart、L. Bounia、F. Koriche、J. Lagniez 和 P. Marquis。《布尔分类器的计算可理解性》。发表于第 *18* 届知识表示与推理原理国际会议（*KR*）论文集，第 74–86 页，2021。

Gilles Audemard、Steve Bellart、Louenas Bounia、Frédéric Koriche、Jean-Marie Lagniez 和 Pierre Marquis。《决策树与随机森林的优选溯因解释》。载于第*31*届国际联合人工智能大会（*IJCAI*）论文集，第643–650页，2022a。

Gilles Audemard、Steve Bellart、Louenas Bounia、Frédéric Koriche、Jean-Marie Lagniez 和 Pierre Marquis。《以复杂度换取随机森林解释的稀疏性》。载于第*36*届*AAAI*人工智能大会论文集，第5号，第5461–5469页，2022b。

Gilles Audemard、Jean-Marie Lagniez、Pierre Marquis 和 Nicolas Szczepanski。《计算提升树的溯因解释》。载于第*26*届人工智能与统计国际会议（*AISTATS*）论文集，第4699–4711页，2023。

Steve Azzolin、Sagar Malhotra、Andrea Passerini 和 Stefano Teso。《超越拓扑自解释 GNN：形式化可解释性视角》。载于第*42*届国际机器学习大会（*ICML*）论文集，2025。

Fahiem Bacchus 和 George Katsirelos。利用最小修正集更高效地计算最小不可满足集。载于第*27*届国际计算机辅助验证大会（*CAV*）论文集，第70–86页。Springer，2015。

Stanley Bak、Changliu Liu 和 Taylor Johnson。《第二届国际神经网络验证竞赛（VNN-COMP 2021）：总结与结果》。2021。技术报告。https://arXiv:2109.00498。

P. Barceló、M. Monet、J. Pérez 和 B. Subercaseaux。《计算复杂性视角下的模型可解释性》。载于第*34*届神经信息处理系统大会论文集（*NeurIPS*），第15487–15498页，2020。

S. Bassan 和 G. Katz。《迈向形式化可解释人工智能：神经网络最小解释的形式化近似》。载于第*29*届系统构造与分析工具和算法国际会议论文集（*TACAS*），第187–207页，2023。S. Bassan、G. Amir 和 G. Katz。《局部与全局可解释性：计算复杂性的视角》。载于第*41*届国际机器学习大会论文集（*ICML*），2024。S. Bassan、Yizhak Yisrael Elboher、Tobias Ladner、Matthias Althoff 和 Guy Katz。《解释有快有慢：可证明解释的抽象与精化》。载于第*42*届国际机器学习大会论文集（*ICML*），2025a。Shahaf Bassan、Guy Amir、Davide Corsi、Idan Refaeli 和 Guy Katz。《在反应式系统中形式化解释神经网络》。载于第*23*届计算机辅助设计形式化方法国际会议论文集（*FMCAD*），第1–13页，2023。Shahaf Bassan、Guy Amir、Meirav Zehavi 和 Guy Katz。《集成模型为何（不）可解释？》。载于第*42*届国际机器学习大会论文集（*ICML*），2025b。Shahaf Bassan、Ron Eliav 和 Shlomit Gur。《简要解释自己！具备简洁充分理由的自解释神经网络》。载于第*13*届学习表征国际会议论文集（*ICLR*），2025c。Shahaf Bassan、Shlomit Gur、Sergey Zeltyn、Konstantinos Mavrogiorgos、Ron Eliav 和 Dimosthenis Kyriazis。《用于业务流程监控的自解释神经网络》。2025d。技术报告。https://arXiv:2503.18067。Shahaf Bassan、Michal Moshkovitz 和 Guy Katz。《加性模型解释：一种计算复杂性方法》。载于第*39*届神经信息处理系统大会论文集（*NeurIPS*），2025e。Shahaf Bassan、Yizhak Yisrael Elboher、Tobias Ladner、Volkan ̧Sahin、Jan Kretinsky、Matthias Althoff 和 Guy Katz。《神经加性模型的可证明解释》。载于第*14*届学习表征国际会议论文集（*ICLR*），2026a。Shahaf Bassan、Xuanxiang Huang 和 Guy Katz。《统一形式化解释：复杂性理论视角》。载于第*14*届学习表征国际会议论文集（*ICLR*），2026b。Leonard Bereska 和 Efstratios Gavves。《人工智能安全的机制可解释性——综述》。机器学习研究汇刊（*TMLR*），2024。Adithya Bhaskar、Alexander Wettig、Dan Friedman 和 Danqi Chen。《通过边剪枝寻找Transformer电路》。载于第*38*届神经信息处理系统大会论文集（*NeurIPS*），第18506–18534页，2024。Guy Blanc、Jane Lange 和 Li-Yang Tan。《可证明高效、简洁且精确的解释》。载于第*35*届神经信息处理系统大会论文集（*NeurIPS*），第6129–6141页，2021。Guy Blanc、Caleb Koch、Jane Lange 和 Li-Yang Tan。《寻找反事实的查询最优算法》。载于第*39*届国际机器学习大会论文集（*ICML*），第2075–2090页，2022。Ryma Boumazouza、Fahima Cheikh-Alili、Bertrand Mazure 和 Karim Tabia。《ASTERYX：一种模型无关的基于SAT的方法，用于符号解释和基于评分的解释》。载于第*30*届*ACM*信息与知识管理国际会议论文集（*CIKM*），第120–129页，2021。Ryma Boumazouza、Raya Elsaleh、Melanie Ducoffe、Shahaf Bassan 和 Guy Katz。《FAME：神经网络的形式化抽象最小解释》。载于第*14*届学习表征国际会议论文集（*ICLR*），2026。Louenas Bounia。《以简洁可靠的概率解释提升布尔决策树的可理解性》。载于第*20*届基于知识系统中不确定性的信息处理与管理国际会议论文集（*IPMU*），第205–218页，2024。Louenas Bounia。《利用子模优化近似基于树的模型的最小大小溯因路径解释》。载于第*41*届人工智能不确定性会议论文集（*UAI*），2025。

Christopher Brix、Mark Niklas Müller、Stanley Bak、Taylor T Johnson 和 Changliu Liu。《国际神经网络验证竞赛（VNN-COMP）前三年》。软件工具与技术转移国际期刊（*STTT*），25(3):329–339，2023。Christopher Brix、Stanley Bak、Taylor T Johnson 和 Haoze Wu。《第五届国际神经网络验证竞赛（VNN-COMP 2024）：总结与结果》。2024。技术报告。https://arXiv:2412.19985。Hong-Ming Chiu、Hao Chen、Huan Zhang 和 Richard Y Zhang。《SDP-CROWN：利用半定规划的紧致性，实现高效的神经网络验证界传播》。载于第*42*届国际机器学习大会（*ICML*）论文集，2025。Pratim Chowdhary、Peter Chin 和 Deepernab Chakrabarty。《K-MSHC：通过句法分类任务实验，揭示大语言模型中的最小充分头部电路》。2025。技术报告。https://arxiv.org/abs/2505.12268。Arthur Conmy、Augustine Mavor-Parker、Aengus Lynch、Stefan Heimersheim 和 Adrià Garriga- Alonso。《迈向机制可解释性的自动化电路发现》。载于第*37*届神经信息处理系统大会（*NeurIPS*）论文集，第16318–16352页，2023。A. Darwiche 和 A. Hirth。《决策背后的原因》。载于第*24*届欧洲人工智能大会（*ECAI*）论文集，第712–720页，2020。Adnan Darwiche。《可解释人工智能的逻辑》。载于第*38*届*ACM/IEEE*计算机科学逻辑研讨会（*LICS*）论文集，第1–11页，2023。Adnan Darwiche 和 Chunxi Ji。《必要且充分解释的计算》。载于第*36*届*AAAI*人工智能大会论文集，第5582–5591页，2022。Alessandro De Palma、Greta Dolcetti 和 Caterina Urban。《更快速的神经网络经验证解释》。2025。技术报告。https://arXiv:2512.00164。Jacob Dunefsky、Philippe Chlenski 和 Neel Nanda。《转码器发现可解释的大语言模型特征电路》。第*38*届神经信息处理系统大会（*NeurIPS*）论文集，第24375–24410页，2024。Ruediger Ehlers。《分段线性前馈神经网络的形式化验证》。载于第*15*届自动化验证与分析技术国际研讨会（*ATVA*）论文集，第269–286页。Springer，2017。Nelson Elhage、Neel Nanda、Catherine Olsson、Tom Henighan、Nicholas Joseph、Ben Mann、Amanda Askell、Yuntao Bai、Anna Chen、Tom Conerly 等。《Transformer电路的数学框架》。*Transformer*电路专题，1(1):12，2021。Thomas Fel、Mélanie Ducoffe、David Vigouroux、Rémi Cadène、Mikael Capelle、Claire Nicodème 和 Thomas Serre。《别骗我！通过经验证的扰动分析实现稳健高效的可解释性》。载于*IEEE/CVF*计算机视觉与模式识别大会（*CVPR*）论文集，第16153–16163页，2023。Claudio Ferrari、Mark Niklas Mueller、Nikola Jovanovi´c 和 Martin Vechev。《通过多神经元松弛引导的分支定界实现完备验证》。载于第*10*届国际学习表征大会（*ICLR*）论文集，2022。Timon Gehr、Matthew Mirman、Dana Drachsler-Cohen、Petar Tsankov、Swarat Chaudhuri 和 Martin Vechev。《AI2：利用抽象解释对神经网络进行安全性与鲁棒性认证》。载于第*39*届*IEEE*安全与隐私研讨会（*S&P*）论文集，第3–18页，2018。Atticus Geiger、Zhengxuan Wu、Christopher Potts、Thomas Icard 和 Noah Goodman。《寻找可解释的因果变量与分布式神经表征之间的对齐关系》。载于第*3*届因果学习与推理大会（*CleaR*）论文集，第160–187页，2024。Atticus Geiger、Duligur Ibeling、Amir Zur、Maheep Chaudhary、Sonakshi Chauhan、Jing Huang、Aryaman Arora、Zhengxuan Wu、Noah Goodman、Christopher Potts 等。《因果抽象：机制可解释性的理论基础》。机器学习研究期刊（*JMLR*），26(83):1–64，2025。

Chuqin Geng、Nham Le、Xiaojie Xu、Zhaoyue Wang、Arie Gurfinkel 和 Xujie Si。《迈向可靠的神经规范》。载于第*40*届国际机器学习大会（*ICML*）论文集，第11196–11212页，2023。Chuqin Geng、Zhaoyue Wang、Haolin Ye 和 Xujie Si。《学习最小神经规范》。2024。技术报告。https://arXiv:2404.04662。Tal Haklay、Hadas Orgad、David Bau、Aaron Mueller 和 Yonatan Belinkov。《位置感知自动电路发现》。载于计算语言学协会第*63*届年会（*ACL*）论文集，第2792–2817页，2025。Michael Hanna、Sandro Pezzelle 和 Yonatan Belinkov。《相信忠实性：发现模型机制时超越电路重叠》。载于第*1*届语言建模大会（*COLM*）论文集，2024。Ouns El Harzli、Bernardo Cuenca Grau 和 Ian Horrocks。《单调神经网络的基数最小解释》。载于第*32*届国际人工智能联合大会（*IJCAI*）论文集，第3677–3685页，2023。Aliyah R Hsu、Georgia Zhou、Yeshwanth Cherapanamjeri、Yaxuan Huang、Anobel Odisho、Peter R Carroll 和 Bin Yu。《利用上下文分解高效地自动发现Transformer中的电路》。载于第*13*届国际学习表征大会（*ICLR*）论文集，2024。Alexey Ignatiev、António Morgado 和 Joao Marques-Silva。《RC2：一种高效的最大可满足性求解器》。*Journal on Satisfiability, Boolean Modelling and Computation*（*JSAT*），11(1):53–64，2019a。Alexey Ignatiev、Nina Narodytska 和 Joao Marques-Silva。《机器学习模型的基于溯因的解释》。载于第*33*届*AAAI*人工智能大会论文集，第1511–1519页，2019b。Alexey Ignatiev、Nina Narodytska 和 Joao Marques-Silva。《论解释与对抗样本之间的关系》。第*33*届神经信息处理系统大会（*NeurIPS*）论文集，2019c。Alexey Ignatiev、Yacine Izza、Peter J Stuckey 和 Joao Marques-Silva。《使用最大可满足性实现树集成的高效解释》。载于第*36*届*AAAI*人工智能大会论文集，第3776–3785页，2022。Yacine Izza、Xuanxiang Huang、Alexey Ignatiev、Nina Narodytska、Martin Cooper 和 Joao Marques-Silva。《概率性溯因解释的计算》。近似推理国际期刊（*IJAR*），159:108939，2023。Yacine Izza、Xuanxiang Huang、Antonio Morgado、Jordi Planes、Alexey Ignatiev 和 Joao Marques-Silva。《距离受限解释：理论基础与高效实现》。载于第*21*届知识表示与推理原理国际大会（*KR*）论文集，第475–486页，2024。Farnoush Rezaei Jafari、Oliver Eberle、Ashkan Khakzar 和 Neel Nanda。《RelP：通过相关性修补实现忠实且高效的电路发现》。2025。技术报告。https://arXiv: 2508.21258。Helen Jin、Anton Xue、Weiqiu You、Surbhi Goel 和 Eric Wong。《特征归因的概率稳定性保证》。2025。技术报告。https://arXiv:2504.13787。Kyle D Julian、Ritchie Lee 和 Mykel J Kochenderfer。《通过自适应压力测试验证基于图像的神经网络控制器》。载于第*23*届*IEEE*智能交通系统国际会议（*ITSC*）论文集，第1–7页，2020。Guy Katz、Clark Barrett、David L Dill、Kyle Julian 和 Mykel J Kochenderfer。《Reluplex：用于验证深度神经网络的高效SMT求解器》。载于第*29*届计算机辅助验证大会（*CAV*）论文集，第97–117页，2017。

Guy Katz, Derek A Huang, Duligur Ibeling, Kyle Julian, Christopher Lazarus, Rachel Lim, Parth Shah, Shantanu Thakoor, Haoze Wu, Aleksandar Zelji´c 等。《用于深度神经网络验证与分析的 Marabou 框架》。载于第*31*届计算机辅助验证国际会议（*CAV*）论文集，第443–452页，2019年。Suhas Kotha、Christopher Brix、J Zico Kolter、Krishnamurthy Dvijotham 和 Huan Zhang。《神经网络原像的可证明界定》。载于第*37*届神经信息处理系统大会（*NeurIPS*）论文集，第80270–80290页，2023年。Alex Krizhevsky 和 Geoffrey Hinton。《从微型图像中学习多层特征》，2009年。技术报告。https://www.cs.toronto.edu/~kriz/ learning-features-2009-TR.pdf。E manuele La Malfa、Agnieszka Zbrzezny、Rhiannon Michelmore、Nicola Paoletti 和 Marta Kwiatkowska。《关于自然语言处理模型的有保证最优鲁棒解释》。载于国际人工智能联合会议（*IJCAI*）论文集，第2658–2665页，2021年。Faezeh Labbaf、Tomáš Kolárik、Martin Blicha、Grigory Fedyukovich、Michael Wand 和 Natasha Sharygina。《神经网络分类的空间解释》。载于第*37*届计算机辅助验证国际会议（*CAV*）论文集，第287–303页，2025年。Y. Lecun、L. Bottou、Y. Bengio 和 P. Haffner。《应用于文档识别的基于梯度的学习》。*IEEE* 论文集，86(11):2278–2324，1998年。Mark H Liffiton、Alessandro Previti、Ammar Malik 和 Joao Marques-Silva。《快速、灵活的 MUS 枚举》。*Constraints*，21(2):223–250，2016年。Emanuele La Malfa、Agnieszka Zbrzezny、Rhiannon Michelmore、Nicola Paoletti 和 Marta Kwiatkowska。《关于自然语言处理模型的有保证最优鲁棒解释》。载于国际人工智能联合会议（*IJCAI*）论文集，2021年。Samuel Marks、Can Rager、Eric J Michaud、Yonatan Belinkov、David Bau 和 Aaron Mueller。《稀疏特征电路：发现并编辑语言模型中的可解释因果图》。载于第*13*届国际学习表征会议（*ICLR*）论文集，2025年。J. Marques-Silva、T. Gerspacher、M. Cooper、A. Ignatiev 和 N. Narodytska。《以多项式时间和延迟解释朴素贝叶斯及其他线性分类器》。载于第*34*届神经信息处理系统大会（*NeurIPS*）论文集，第20590–20600页，2020年。Joao Marques-Silva 和 Alexey Ignatiev。《通过形式化可解释人工智能实现可信人工智能》。载于第*36*届*AAAI*人工智能大会论文集，第12342–12350页，2022年。Joao Marques-Silva 和 Alexey Ignatiev。《没有灵丹妙药：可解释机器学习模型必须得到解释》。*Frontiers in Artificial Intelligence*，6:1128212，2023年。Joao Marques-Silva、Thomas Gerspacher、Martin Cooper、Alexey Ignatiev 和 Nina Narodytska。《单调分类器的解释》。载于国际机器学习会议（*ICML*）论文集，第7469–7479页，2021年。Reda Marzouk 和 Colin De La Higuera。《马尔可夫分布下 SHAP 解释的可处理性》。载于第*41*届国际机器学习会议（*ICML*）论文集，第34961–34986页，2024年。Reda Marzouk、Shahaf Bassan 和 Guy Katz。《SHAP 遇上张量网络：具有并行性的可证明易处理解释》。载于第*39*届神经信息处理系统大会（*NeurIPS*）论文集，2025a。Reda Marzouk、Shahaf Bassan、Guy Katz 和 De la Higuera。《关于（众多）Shapley 值的计算可处理性》。载于第*28*届人工智能与统计国际会议（*AISTATS*）论文集，第3691–3699页。PMLR，2025b。Maxime Méloux、Silviu Maniu、François Portet 和 Maxime Peyrard。《万物、处处、同时：机制可解释性是否可识别？》。载于第*13*届国际学习表征会议（*ICLR*）论文集，2025年。

Joseph Miller、Bilal Chughtai 和 William Saunders。《Transformer 电路评估指标并不鲁棒》。载于第*1*届语言建模会议（*COLM*）论文集，2024年。

Aaron Mueller、Atticus Geiger、Sarah Wiegreffe、Dana Arad、Iván Arcuschin、Adam Belfki、Yik Siu Chan、Jaden Fried Fiotto-Kaufman、Tal Haklay、Michael Hanna 等。MIB：机制可解释性基准。在 第*42*届国际机器学习大会论文集（*ICML*），2025 年。

Christoph Müller、François Serre、Gagandeep Singh、Markus Püschel 和 Martin Vechev。基于图形处理器扩展多面体神经网络验证。在 第*4*届机器学习与系统大会论文集（*MLSys*），第 733–746 页，2021 年。

Mark Niklas Müller、Christopher Brix、Stanley Bak、Changliu Liu 和 Taylor T Johnson。第三届国际神经网络验证竞赛（VNN-COMP 2022）：总结与结果。2022a。技术报告。https://arXiv.org/abs/2212.10376。

Mark Niklas Müller、Gleb Makarchuk、Gagandeep Singh、Markus Püschel 和 Martin Vechev。PRIMA：通过可扩展凸包近似实现通用且精确的神经网络认证。第*6*届 *ACM* 编程语言大会（*POPL*）论文集，第 1–33 页，2022b。

Neel Nanda、Lawrence Chan、Tom Lieberum、Jess Smith 和 Jacob Steinhardt。通过机制可解释性衡量 Grokking 的进展。在 第*11*届国际学习表征大会（*ICLR*）论文集，2023。

Chris Olah。机制可解释性、变量与可解释基底的重要性。*TransformerCircuitsThread*，2022。

Chris Olah、Nick Cammarata、Ludwig Schubert、Gabriel Goh、Michael Petrov 和 Shan Carter。深入观察：电路简介。*Distill*，2020。

Nils Palumbo、Ravi Mangal、Zifan Wang、Saranya Vijayakumar、Corina S. Pasareanu 和 Somesh Jha。验证机制解释：一种公理化方法。在第*42*届国际机器学习大会（*ICML*）论文集，2025。

Achyuta Rajaram、Neil Chowdhury、Antonio Torralba、Jacob Andreas 和 Sarah Schwettmann。《视觉电路的自动发现》。2024。技术报告。https://arxiv.org/abs/ 2404.14349。

Tilman Räuker、Anson Ho、Stephen Casper 和 Dylan Hadfield-Menell。《迈向透明人工智能：深度神经网络内部结构解读综述》。发表于第 *1* 届 *IEEE* 安全与可信机器学习国际会议（*SaTML*）论文集，第 464–483 页，2023。

Ramprasaath R Selvaraju、Michael Cogswell、Abhishek Das、Ramakrishna Vedantam、Devi Parikh 和 Dhruv Batra。《Grad-CAM：通过基于梯度的定位对深度网络进行可视化解释》。发表于第 *16* 届 *IEEE* 国际计算机视觉会议（*ICCV*）论文集，第 618–626 页，2017。

Lee Sharkey、Bilal Chughtai、Joshua Batson、Jack Lindsey、Jeff Wu、Lucius Bushnaq、Nicholas Goldowsky-Dill、Stefan Heimersheim、Alejandro Ortega、Joseph Bloom 等。机制可解释性中的开放问题。机器学习研究汇刊（*TMLR*），2025。

Claudia Shi、Nicolas Beltran Velez、Achille Nazaret、Carolina Zheng、Adrià Garriga-Alonso、Andrew Jesson、Maggie Makar 与 David Blei。对大语言模型中电路假设的假设检验。载于第*38*届神经信息处理系统会议论文集（*NeurIPS*），第94539–94567页，2024。

Zhouxing Shi、Qirui Jin、Zico Kolter、Suman Jana、Cho-Jui Hsieh 与 Huan Zhang。针对一般非线性函数的分支定界神经网络验证。载于第*31*届国际系统构造与分析工具和算法会议论文集（*TACAS*），第315–335页，2025。

A. Shih、A. Choi 和 A. Darwiche。《一种解释贝叶斯网络分类器的符号方法》。载于第 *27* 届国际人工智能联合会议（*IJCAI*）论文集，第 5103–5111 页，2018年。Gagandeep Singh、Timon Gehr、Markus Püschel 和 Martin Vechev。《用于认证神经网络的抽象域》。载于第 *3* 届 *ACM* 编程语言大会（*POPL*）论文集，第 1–30 页，2019年。

Jules Soria、Zakaria Chihani、Julien Girard-Satabin、Alban Grastien、Romain Xu-Darme 和 Daniela Cancila。《基于原型的网络的形式化溯因潜在解释》。2025年。技术报告。https://arXiv.org/abs/2511.16588。Johannes Stallkamp、Marc Sc hlipsing、Jan Salmen 和 Christian Igel。《德国交通标志识别基准：一项多类别分类竞赛》。载于第 *24* 届国际神经网络联合会议（*IJCNN*）论文集，第 1453–1460 页，2011。Jiuding Sun、Jing Huang、Sidharth Baskaran、Karel D'Oosterlinck、Christopher Potts、Michael Sklar 和 Atticus Geiger。《HyperDAS：利用超网络实现机制可解释性自动化》。载于第 *13* 届国际学习表征会议（*ICLR*）论文集，2025。Aaquib Syed、Can Rager 和 Arthur Conmy。《归因修补优于自动化电路发现》。载于第 *7* 届 *BlackboxNLP* 研讨会：分析和解释自然语言处理神经网络，第 407–416 页，2024。Vincent Tjeng、Kai Y Xiao 和 Russ Tedrake。《使用混合整数规划评估神经网络的鲁棒性》。载于第 *5* 届国际学习表征会议（*ICLR*）论文集，2017。Kevin Ro Wang、Alexandre Variengien、Arthur Conmy、Buck Shlegeris 和 Jacob Steinhardt。《野外可解释性：GPT-2 Small 中用于间接宾语识别的电路》。载于第 *11* 届国际学习表征会议（*ICLR*）论文集，2023。Shiqi Wang、Huan Zhang、Kaidi Xu、Xue Lin、Suman Jana、Cho-Jui Hsieh 和 J Zico Kolter。《Beta-crown：利用逐神经元拆分约束进行高效界传播，以验证神经网络鲁棒性》。载于第 *35* 届神经信息处理系统大会（*NeurIPS*）论文集，第 29909–29921 页，2021。

Haoze Wu、Omri Isac、Aleksandar Zelji´c、Teruhiro Tagomori、Matthew Daggitt、Wen Kokke、Idan Refaeli、Guy Amir、Kyle Julian、Shahaf Bassan 等。《Marabou 2.0：一种通用的神经网络形式化分析器》。载于第 *36* 届计算机辅助验证国际会议论文集（*CAV*），第 249–264 页，2024 年。

Min Wu、Haoze Wu 和 Clark Barrett。《VeriX：迈向深度神经网络的可验证可解释性》。载于第*37*届神经信息处理系统大会（*NeurIPS*）论文集，第22247–22268页，2023a。Wilson Wu、Louis Jaburi、Jason Gross 等。《迈向对群运算网络的统一且可验证的理解》。载于第*13*届国际学习表征会议（*ICLR*）论文集，2025。Zhengxuan Wu、Atticus Geiger、Thomas Icard、Christopher Potts 和 Noah Goodman。《规模化可解释性：识别 Alpaca 中的因果机制》。载于第*37*届神经信息处理系统大会（*NeurIPS*）论文集，第78205–78226页，2023b。Anton Xue、Rajeev Alur 和 Eric Wong。《乘法平滑下特征归因的稳定性保证》。载于第*37*届神经信息处理系统大会（*NeurIPS*）论文集，第62388–62413页，2023。Yunzhi Yao、Ningyu Zhang、Zekun Xi、Mengru Wang、Ziwen Xu、Shumin Deng 和 Huajun Chen。《预训练变换器中的知识电路》。载于第*38*届神经信息处理系统大会（*NeurIPS*）论文集，第118571–118602页，2024。Ji nqiang Yu、Alexey Ignatiev、Peter Stuckey、Nina Narodytska 和 Joao Marques-Silva。《排除不可能之事，剩下的无论是什么都必定为真：形式化解释背景下背景知识的提取与应用》。载于第*37*届 *AAAI* 人工智能大会论文集，第4123–4131页，2023。Fred Zhang 和 Neel Nanda。《迈向语言模型激活修补的最佳实践：指标与方法》。载于第*12*届国际学习表征会议（*ICLR*）论文集，2024。Huan Zhang、Tsui-Wei Weng、Pin-Yu Chen、Cho-Jui Hsieh 和 Luca Daniel。《利用一般激活函数高效认证神经网络鲁棒性》。载于第*32*届神经信息处理系统大会（*NeurIPS*）论文集，2018。

Yu Zhang、Peter Tiˇno、Aleš Leonardis 和 Ke Tang。《神经网络可解释性综述》。*IEEE*《计算智能新兴主题汇刊》，5(5)：726–742，2021。

Haiyan Zhao、Hanjie Chen、Fan Yang、Ninghao Liu、Huiqi Deng、Hengyi Cai、Shuaiqiang Wang、Dawei Yin 和 Mengnan Du。《大语言模型的可解释性：综述》。*ACM*《智能系统与技术汇刊》，15(2):1–38，2024。

Duo Zhou、Christopher Brix、Grani A Hanasusanto 和 Huan Zhang。使用分支定界推断割平面实现可扩展的神经网络验证。第 *38* 届神经信息处理系统会议论文集（*NeurIPS*），37:29324–29353，2024。

Zhenhong Zhou、Haiyang Yu、Xinghua Zhang、Rongwu Xu、Fei Huang、Kun Wang、Yang Liu、Junfeng Fang 和 Yongbin Li。《注意力头在大语言模型安全性中的作用》。载于第 *13* 届国际学习表征会议（*ICLR*）论文集，2025。

附录

附录汇集了支持正文的证明、模型规格说明和补充实验结果。

附录 A 介绍神经网络验证、电路发现和修补的背景知识。

附录 B 包含所有命题的完整证明。
附录 C 提供贪心准最小性算法的伪代码，并对用于阐释最小性概念的玩具示例作了更详细的说明。

附录 D 介绍所用数据集和架构的具体说明。
附录 E介绍了输入鲁棒性实验的方法、验证设置和评估的更多细节。

附录 F介绍了修补鲁棒性实验的方法、验证设置和评估的更多细节。

附录 G 介绍了最小性评估实验的方法、验证设置和评估的更多细节。

附录 H 介绍了大语言模型的使用披露。

A 神经网络验证、电路发现与形式化可解释人工智能的补充背景

神经网络验证。 神经网络验证为神经网络 f_G 在连续输入区域内的行为提供形式化保证。经典的基于 SMT/MILP 的方法（ Katz 等人，2017； Wu 等人，2024； Katz 等人，2019； Tjeng 等人，2017； Ehlers，2017）将 ReLU 网络和规范编码为逻辑约束或混合整数约束，并提供精确保证，但只能扩展到中小型模型。抽象解释方法（Singh 等人，2019；Gehr 等人，2018；Ferrari 等人，2022；Müller 等人，2022b）逐层传播过近似，从而快速生成但并不完备的鲁棒性证书。基于线性松弛的界传播带来了重大进展，其中尤以 CROWN（Zhang 等人，2018）及其后续工作（Wang 等人，2021；Chiu 等人，2025；Zhou 等人，2024；Shi 等人，2025）为代表。这些方法计算紧致的基于对偶的线性界，可用作可扩展但不完备的验证器，也可作为精确搜索中的强松弛。现代分支定界（ BaB）框架利用这些松弛实现完备验证，并能扩展到大规模问题； α-β-CROWN 及相关变体如今在 VNN-COMP（Brix 等人，2024）中占据主导地位，并可处理百万参数模型。近期进展包括更紧致的松弛（例如半定规划混合方法（Chiu 等人，2025））、割平面（Zhou 等人，2024），以及对非 ReLU 非线性函数的支持（Shi 等人，2025）。如今，验证技术已能常规地为中大型卷积神经网络和残差网络认证鲁棒性，但对于 Transformer、复杂架构以及更丰富的时序或关系规范，仍面临重大挑战。

电路发现与修补。 电路发现中的一个重要步骤是修补，其旨在通过对电路外部的激活值进行干预，分离所假设电路的计算作用。在典型设置中，模型在基础输入上运行，并将选定非电路节点处的激活值加以替换——要么替换为固定基线值（例如零值或平均激活值），要么替换为取自反事实输入的激活值。若模型的输出保持不变，则认为该电路足以支撑该行为；若输出发生变化，则表明其对被修补的组件存在依赖。目前已提出多种补丁干预方案，包括激活值替换、路径补丁干预以及注意力／头部干预（Jafari et al., 2025; Syed et al., 2024; Haklay et al., 2025; Miller et al., 2024; Zhang & Nanda, 2024; Nanda et al., 2023），其目标均为识别对于目标计算而言行为必要或充分的模型组件。据我们所知，我们的方法首次能够可证明地认证电路在此类修补干预族下的稳定性。

形式化可解释人工智能。 我们的工作也与一条名为形式化可解释人工智能（formal XAI）的研究脉络相关（ Marques-Silva & Ignatiev, 2022），该领域研究具有形式化保证的基于输入的解释（Yu et al., 2023; Darwiche & Ji, 2022; Darwiche, 2023; Shih et al., 2018; Azzolin et al., 2025; Audemard et al., 2021）。由于生成具有形式化保证的解释通常计算上难以处理（Adolfi 等人, 2025; Barceló et al., 2020; Arenaset al., 2021; Bassan etal., 2026b; Marzouk et al., 2025a;b; Marzouk & De La Higuera, 2024; Amir et al., 2024; Bassan et al.,

2024; Blanc et al., 2021; 2022），这方面的大部分文献都聚焦于较简单的模型类别（Marques- Silva & Ignatiev, 2023），包括决策树（Bounia, 2025; Arenas et al., 2022; Bounia, 2024）、线性模型（Marques-Silva et al., 2020）、加性模型（Bassan et al., 2025e; 2026a）、单调模型（Marques-Silva et al., 2021; Harzli et al., 2023）和树集成（Audemard et al., 2023; 2022a; Ignatiev 等人, 2022; Bassan et al., 2025b; Boumazouza et al., 2021）。与我们的研究设置更接近的是，已有若干研究将这些保证扩展到了神经网络（Malfa et al., 2021; Bassan et al., 2023; Wu et al., 2023a; Bassan et al., 2025a; Boumazouza et al., 2026; Labbaf et al., 2025; De Palma et al., 2025; Fel et al., 2023; Soria et al., 2025）。

形式化可解释性中的一个核心概念是充分理由 （Darwiche & Hirth，2020；Bassan 等人，2025c；d；Barceló 等人，2020），也称为溯因解释 （Ignatiev 等人，2019b；Izza 等人，2023）。这些概念与基于稳定性的特征归因框架密切相关（Xue 等人，2023；Jin 等人，2025；Anani 等人，2025）。非正式地说，充分理由类似于电路：固定它即可保持模型的预测。研究者在此背景下探讨了多种最小性概念，包括子集极小解释（Ignatiev 等人，2019b），以及通过类似于本文所研究的极小击中集（MHS）对偶性计算的基数极小解释（Ignatiev 等人，2019b；Bassan 与 Katz，2023）。然而，我们的工作解决的是一个本质上不同的任务：发现电路，而非输入子集。从特征子集转向模型子组件会带来若干新挑战，包括为认证而对模型进行对偶编码、处理修补操作的需求，以及在某些修补方案下单调性的丧失，这使得收敛到最小电路变得更加复杂。据我们所知，这是首次将形式化可解释性中的思想应用于机制可解释性，尤其是用于电路发现任务。

最后，另一项部分相关的研究将神经网络验证应用于激活模式规范（NAPs）（Geng 等人，2023），这类规范对神经元的激活/未激活状态进行编码，并由此确定超出局部扰动球范围的输入区域。近期研究（Geng 等人，2024）进一步尝试通过移除冗余的神经元状态，同时保持正确性，来学习最小 NAP。这些研究也不同于我们的电路：我们的电路计算的是神经网络的实际子图。

B 主要结果的证明

本附录给出了主论文中所述主要命题的证明。

B.1 命题 1 的证明

命题 1. *对于任意模型 f_G 和忠实性谓词 Φ，算法 1 会访问一个准最小电路，而算法 3 会收敛到一个局部最小电路 C，二者均相对于 Φ 而言。*

证明。 我们需要确立三点：（i）忠实性： 算法 1 和 3 返回的最终电路 C 确实是忠实的，即谓词 $\Phi(C, G)$ 成立。（ii）算法 1 在执行过程中会访问一个准最小电路，即存在某次迭代 t，使得相应的电路 C_t 包含一个组件 i，且 $\Phi(C_t \setminus \{i\}, G)$ 不成立。（iii）算法 3 实现局部最小性：在返回的电路 C中，移除任意一个组件 $i \in C$ 都会破坏忠实性；也就是说，对于每个 $i \in C$，$\Phi(C \setminus \{i\}, G)$ 均不成立。

在证明的第一部分，算法 1 和 3 都以 $C \leftarrow G$ 开始。根据定义，完整模型图G 是忠实的，因此 $\Phi(G, G)$ 成立。在每一步中，算法都会针对当前组件 i 检查 $\Phi(C \setminus \{i\}, G)$，并且只有在该谓词仍然成立时才更新 $C \leftarrow C \setminus \{i\}$。因此，不变量在整个执行过程中得以保持，最终电路 C 仍满足 $\Phi(C, G)$。

对于第二部分，我们证明算法 1 在执行过程中会访问一个准最小电路。我们假设空电路是不忠实的，即 $\neg\Phi(\emptyset, G)$。设 C_t 表示算法在第 t 次迭代时维护的电路。由于算法维护不变量 $\Phi(C_t, G)$，且仅在忠实性得以保持时才移除组件，因此它无法到达空电路。由此，在保持 Φ 的同时，不可能移除所有组件。因此，存在第一次迭代 t：此时算法考察某个组件 $i \in C_t$，但因为 $\Phi(C_t \setminus \{i\}, G)$ 为假而不移除它。因此，C_t 包含一个移除后会破坏忠实性的组件；根据定义，该电路相对于 Φ 是准最小的。这证明算法 1 在执行过程中会访问一个准最小电路。

对于证明的第三部分，请注意，算法 3 会持续遍历电路组件，直到完整遍历一轮且电路 C 不再发生变化。这个终止条件意味着，对于任何剩余的 $i \in C$，在最后一次迭代期间，谓词 $\Phi(C \setminus \{i\},\ G)$ 必定为假。否则，如果对于任何 $i \in C$，$\Phi(C \setminus \{i\},\ G)$ 为真，该组件就会被移除，算法也会继续进行下一次迭代。因此，算法终止时，每个 $i \in C$ 都是维持忠实性谓词所必需的，这证明 C 相对于 Φ 是局部最小的。证明至此完成。

□

B.2 命题 2 的证明

命题 2。 虽然算法 *4* 收敛到一个准最小电路，并执行 $\mathcal{O}$(对数 $|G|$)次 $\Phi(C, G)$ 求值（*Adolfi* 等，*2025*），但算法 *1* 会访问一个准最小电路，并执行$\mathcal{O}(|G|)$ 次此类求值；算法 *3* 则收敛到一个局部最小电路，并执行 $\mathcal{O}(|G|^2)$次求值。

证明。 我们分别给出每种算法的证明：

*(*算法 *1)*。根据命题 1，算法 1 在执行过程中会访问一个准最小电路。注意，算法 1 会逐个处理 G 中的每个组件，并且每个组件恰好处理一次。由于它会在移除这些 $|G|$ 组件中的每一个后测试 Φ，因此该过程总共执行 $\mathcal{O}(|G|)$次谓词求值。

*(*算法 *3)*。根据命题 1，算法 3 会收敛到一个局部最小电路。在运行时间分析中，注意算法 3 会处理 G 中的每个组件，并持续执行该操作直至收敛。在每一轮中，它都会在移除每个剩余组件后测试谓词 Φ。在最坏情况下，每一轮仅移除一个组件，此时谓词求值的总次数为 $|G| + (|G| - 1) + (|G| - 2) + \cdots + 1 = \mathcal{O}(|G|^2)$。因此，该过程总共执行 $\mathcal{O}(|G|^2)$ 次 $\Phi(C,\ G)$ 求值。

*(*算法 *4)*。准最小性算法 4 及其运行时间由 Adolfi 等人（2025）确立。为完整起见，我们重述关于求值次数的论证：该算法每一步都将候选索引范围减半，因此在终止前至多需要 $\lceil \log |G| \rceil$ 次谓词求值。由此，其复杂度为 $\mathcal{O}(\log |G|)$。

□

B.3 命题 3 的证明

命题 3。 存在无穷多个配置 f_G 和 Φ，使得算法 *1*、*3* 和 *4*不会收敛到关于 Φ 的子集最小电路 C。

证明。 考虑一个小型非线性反例网络 $f_G : \mathbb{R} \to \mathbb{R}$，其底层结构为 G，节点集为 $V_G := \{v_1, v_2, v_3, v_4\}$。该网络定义在一维输入 x 上，包含三个隐藏节点 v_1、v_2、v_3 以及一个输出节点 v_4。

考虑一维输入 x、三个隐藏单元 v_1，v_2，v_3 和一个输出神经元 v_4：

$$v_1(x) := \mathrm{ReLU}(x), \qquad v_2(x) := \mathrm{ReLU}(x), \qquad v_3(x) := x, \qquad v_4(u, v, w) := u - v + w \tag{3}$$

因此，以下结论成立：

$$f_G(x) \ := \ v_4\big(v_1(x), v_2(x), v_3(x)\big) \ = \ v_1(x) - v_2(x) + v_3(x) \ = \ x \tag{4}$$

对于任意子集 $C \subseteq \{v_1, v_2, v_3\}$，通过对 C 之外的所有隐藏单元应用零值修补来定义 f_C。我们将忠实性谓词取为强相等，即对于由 ℓ_∞ 范数在定义域 $[0, 1]$上给出的 $\delta = 0$ 和 ϵ_p 而言。具体来说：

$$\Phi(C, G) \quad \Longleftrightarrow \quad \forall\, x \in [0, 1],\ f_C(x) = f_G(x) \tag{5}$$

迭代式（算法 *1* 和 *3*）。我们按照从末尾到开头的标准遍历顺序考察这些组件（即逆拓扑序）：(v_4, v_3, v_2, v_1)。我们注意到，移除 v_4 （即对其进行零值修补）会直接破坏忠实性。此外，移除任何其他单个组件都无法保持 Φ：移除任意隐藏单元都会破坏相等性。

$$\begin{aligned} f_{G\setminus\{v_3\}}(x) &= \mathrm{ReLU}(x) - \mathrm{ReLU}(x) = 0 \neq x \text{ for } x > 0, \\ f_{G\setminus\{v_2\}}(x) &= \mathrm{ReLU}(x) + x \neq x \text{ for } x > 0, \\ f_{G\setminus\{v_1\}}(x) &= -\mathrm{ReLU}(x) + x \neq x \text{ for } x > 0 \end{aligned} \tag{6}$$

因此，算法 1 和算法 3 在 $C = G$ （局部最小）处停止。然而，存在一个忠实的严格子集：移除这一对 $\{v_1, v_2\}$ 后得到 $f_{G\setminus\{v_1,v_2\}}$（$x$）$= v_3$（$x$）$= x$，因此 Φ（$G \setminus \{v_1, v_2\}$，$G$）成立。因此，$G$ 不是子集最小的。

准（算法 *4*）。给定一个有效的顺序（v_1，v_3，v_2，v_4），算法 4 首先测试前缀移除 $\{v_1, v_3\}$，得到 f_G（x）$=$ −ReLU（x）$\neq x$；随后测试 $\{v_1\}$，得到 f_G（x）$=$ −ReLU（x）$+ x \neq x$。它从未考虑这一对 $\{v_1, v_2\}$，因此返回 $C = G$。正如我们先前在公式 6 中所示，该结果不是子集最小的。

由于这对每个 $m \in \mathbb{N}$ 都成立，我们得到无限多个配置，在这些配置中，两个算法都无法返回子集最小电路。

无限族。 对于任意 $m \geq 1$，添加满足 $p_i(x) = q_i(x) = \mathrm{ReLU}(x)$ 的配对 (p_i, q_i)，并设置

$$g_v(x) \;=\; v_3(x) + \sum_{i=1}^{m} \big(p_i(x) - q_i(x)\big).$$

那么， $f_G(x)$ 在 $[0, 1]$，上移除任何单个元素（ v_3、某个 p_j 或某个 q_j）都会破坏等式，但移除整个配对 $\{p_j, q_j\}$ 则会保留等式。按照顺序 $(p_1, \ldots, p_m, v_3, q_1, \ldots, q_m, v_4)$，迭代贪心算法 1 和 3 以及准最小算法 4 都会返回 G。因此，这样的配置有无限多个。

□

B.4 命题 4 的证明

Pro命题 4。 如果 Φ 是单调的，那么对于任意模型 f_G，算法 *1* 会收敛到一个子集最小 l *circ*关于 Φ的 *uit* C 。

证明。 根据定义，要证明电路 C 具有子集极小性，需要证明：(i) C 是忠实的，即 $\Phi(C, G)$成立；(ii) 不存在任何真子图 $C' \subseteq C$ 也满足 $\Phi(C', G)$。
对于证明的第一部分，我们沿用命题 1 证明中使用的忠实性保持论证。算法 1 从完整模型 G开始，该模型最初忠实于自身（$\Phi(G, G)$成立）。随后，它按照固定顺序（例如逆拓扑序）遍历各个组件。在每一步中，算法都会维护当前电路 C_t；只有当移除组件 i 能够保持忠实性时，才会将其移除，即 $\Phi(C_t \setminus \{i\}, G)$成立。因此，整个执行过程中，忠实性始终作为循环不变式得以保持。特别是，算法终止时，得到的电路 C满足 $\Phi(C, G)$。

对于证明的第二部分，反设存在某个 $C' \subsetneq C$，使得 $\Phi(C',\ G)$为真。现在，任取一个 $c^\star \in C \setminus C'$。令 C_t 为第 t次迭代时的电路，该次迭代标记了算法评估是否将 $c^\star$ 加入电路的步骤。由于 $c^\star$ 存在于返回的最终电路 C中，根据归纳法，它在被考察的第 t 步没有被移除，因此 $\Phi(C_t \setminus \{c^\star\},\ G)$必为假。另一方面，我们有以下包含关系：

$$C' \;\subseteq\; C \setminus \{c^\star\} \;\subseteq\; C_t \setminus \{c^\star\}. \tag{7}$$

因此，根据 C 关于 Φ 的单调性的定义，我们得到：

$$\Phi(C', G) \quad\Longrightarrow\quad \Phi(C_t \setminus \{c^\star\}, G) \tag{8}$$

这与我们根据算法的运行过程推导出 $\Phi(C_t \setminus \{c^\star\},\ G)$ 必为假的事实相矛盾。因此，我们得到，对于任意子集 $C' \subsetneq C$，$\Phi(C',\ G)$ 都为假，进而 C 是关于 Φ 的子集最小集合。

□

B.5 命题 5 的证明

命题 5. 设 $\Phi(C, G)$ 表示验证 C 关于 $\langle f_G, \mathcal{Z}\rangle$ *(定义 1)* 是否具有输入鲁棒性，并同时验证其关于 $\langle f_G, \mathcal{Z}'\rangle$ *(定义 2)* 是否具有修补鲁棒性。那么，如果 $\mathcal{Z} \subseteq \mathcal{Z}'$，且 $\mathcal{H}_G(\mathcal{Z}')$ 在拼接下封闭，则 Φ是单调的。

证明。 我们先正式陈述激活空间在拼接下封闭的条件：

定义 1. 对于某个模型 f_G 及定义域 $\mathcal{Z} \subseteq \mathbb{R}^n$上的激活空间 $\mathcal{H}_G(\mathcal{Z})$，若对于任意两个部分激活值 α, α'，其中 $\alpha \in \mathcal{H}_C(\mathcal{Z})$ 是电路 $C \subseteq G$上的一个激活值，且 $\alpha' \in \mathcal{H}_{C'}(\mathcal{Z})$ 是电路 $C' \subseteq G$上的一个激活值，都有 $\alpha \cup \alpha' \in \mathcal{H}_{C\cup C'}(\mathcal{Z})$，则称该激活空间在拼接下封闭。

现在，设存在某个 $C \subseteq C' \subseteq G$，使得 $C' := C \sqcup \{c'_1, \ldots, c'_t\}$。我们假设谓词 Φ 被定义为：验证某个电路 C 是否关于 $\langle f_G, \mathcal{Z}\rangle$ 是输入鲁棒的，并且关于 $\langle f_G, \mathcal{Z}'\rangle$ 也是修补鲁棒的。换言之，这意味着 $\Phi(C, G)$ 成立，当且仅当：

$$\forall \mathbf{z} \in \mathcal{Z} := \bigcup_{j=1}^{k} \mathcal{B}^p_{\epsilon_p}(\mathbf{x}_j), \quad \forall \alpha \in \mathcal{H}_{\overline{C}}(\mathcal{Z}') : \quad \left\| f_C(\mathbf{z} \mid \overline{C} = \alpha) - f_G(\mathbf{z}) \right\|_p \leq \delta \tag{9}$$

我们注意到，以下符号与以下内容等价：

$$\begin{aligned} \max_{\mathbf{z}\in\mathcal{Z}, \alpha\in\mathcal{H}_{\overline{C}}(\mathcal{Z}')} \quad & \left\| f_C(\mathbf{z} \mid \overline{C} = \alpha) - f_G(\mathbf{z}) \right\|_p \leq \delta \iff \\ \max_{\mathbf{z}\in\mathcal{Z}, \mathbf{z}'\in\mathcal{Z}'} \quad & \left\| f_C(\mathbf{z} \mid \overline{C} = \mathcal{H}_{\overline{C}}(\mathbf{z}')) - f_G(\mathbf{z}) \right\|_p \leq \delta \end{aligned} \tag{10}$$

其中，我们用符号 $\mathcal{H}_C(\mathbf{z}')$ 表示在对某个 $\mathbf{z}' \in \mathbb{R}^n$ 计算 $f_G(\mathbf{z}')$ 时，C上的特定激活函数。

由于我们的目标是证明 Φ 是单调的，因此只需证明 C' 也满足上述条件。一种等价表述是对任意 $C' := C \sqcup \{c'_i\}$ 验证该条件，即针对与 C 仅相差一个元素的超集，而非任意子集。这是成立的，因为以归纳方式逐个添加元素，等价于一次性添加整个子集。因此，我们在此表述下继续，并采用反证法假设这些条件对 C'不成立。换言之，我们假设如下：

$$\begin{aligned} \exists \mathbf{z} \in \mathcal{Z}, \quad \exists \alpha \in \mathcal{H}_{\overline{C'}}(\mathcal{Z}') : \quad & \left\| f_{C'}(\mathbf{z} \mid \overline{C'} = \alpha) - f_G(\mathbf{z}) \right\|_p > \delta \iff \\ \max_{\mathbf{z}\in\mathcal{Z}, \alpha\in\mathcal{H}_{\overline{C'}}(\mathcal{Z}')} \quad & \left\| f_{C'}(\mathbf{z} \mid \overline{C'} = \alpha) - f_G(\mathbf{z}) \right\|_p > \delta \iff \\ \max_{\mathbf{z}\in\mathcal{Z}, \mathbf{z}'\in\mathcal{Z}'} \quad & \left\| f_{C'}(\mathbf{z} \mid \overline{C'} = \mathcal{H}_{\overline{C'}}(\mathbf{z}')) - f_G(\mathbf{z}) \right\|_p > \delta \end{aligned} \tag{11}$$

这也等价于以下表述：

$$\max_{\mathbf{z}\in\mathcal{Z}, \mathbf{z}'\in\mathcal{Z}'} \quad \left\| f_{C\sqcup\{c'_i\}}(\mathbf{z} \mid \overline{C \sqcup \{c'_i\}} = \mathcal{H}_{\overline{C\sqcup\{c'_i\}}}(\mathbf{z}')) - f_G(\mathbf{z}) \right\|_p > \delta \tag{12}$$

我们用 $\mathbb{S} \subseteq \mathbb{R}$ 表示通过 $f_C(\mathbf{z} \mid C = \mathcal{H}_C(\mathbf{z}'))$可以获得的所有可行值的集合，并用 $\mathbb{S} \Subset \mathbb{R}$ 表示通过 $f_C(\mathbf{z} \mid C' = \mathcal{H}_{C'}(\mathbf{z}'))$可行的所有值。更精确地说：

$$\mathbb{S} := \{f_C(\mathbf{z} \mid \overline{C} = \mathcal{H}_{\overline{C}}(\mathbf{z}')) : \mathbf{z} \in \mathcal{Z}, \mathbf{z}' \in \mathcal{Z}'\}, \wedge \mathbb{S}' := \{f_{C'}(\mathbf{z} \mid \overline{C'} = \mathcal{H}_{\overline{C'}}(\mathbf{z}')) : \mathbf{z} \in \mathcal{Z}, \mathbf{z}' \in \mathcal{Z}'\} \tag{13}$$

为完成该命题的证明，我们将使用如下引理:

引理 1. 给定预定义的 f_G、电路 $C \subseteq C' \subseteq G$，以及前述关于$\mathbb{S}$ 和 $\mathbb{S}'$的符号，则有 $\mathbb{S}' \subseteq \mathbb{S}$。

证明。 首先，根据定义:

$$\begin{gathered}\mathbb{S}' := \{f_{C'}(\mathbf{z} \mid \overline{C'} = \mathcal{H}_{\overline{C'}}(\mathbf{z}')) : \mathbf{z} \in \mathcal{Z}, \mathbf{z}' \in \mathcal{Z}'\} = \\ \{f_{C \sqcup \{c_i'\}}(\mathbf{z} \mid \overline{C \sqcup \{c_i'\}} = \mathcal{H}_{\overline{C \sqcup \{c_i'\}}}(\mathbf{z}')) : \mathbf{z} \in \mathcal{Z}, \mathbf{z}' \in \mathcal{Z}'\}\end{gathered} \tag{14}$$

我们还注意到，对于任意 $\mathbf{z} \in \mathbb{R}^n$，根据定义，有:

$$f_C(\mathbf{z} \mid \overline{C} = \mathcal{H}_{\overline{C}}(\mathbf{z}')) = f_\emptyset(\mathbf{z} \mid \overline{C} = \mathcal{H}_{\overline{C}}(\mathbf{z}'), C = \mathcal{H}_C(\mathbf{z})) \tag{15}$$

符号 $f_\emptyset(\mathbf{z} \mid C = \mathcal{H}_{\overline{C}}(\mathbf{z}') \; C = \mathcal{H}_C(\mathbf{z}))$ 表示将 C 的激活值固定为$\mathcal{H}_C(\mathbf{z})$，并将 C 的激活值固定为 $\mathcal{H}_{\overline{C}}(\mathbf{z}')$。按照同样的符号处理方式，可得:

$$\begin{gathered}f_{C \sqcup \{c_i'\}}(\mathbf{z} \mid \overline{C \sqcup \{c_i'\}} = \mathcal{H}_{\overline{C \sqcup \{c_i'\}}}(\mathbf{z}')) = \\ f_\emptyset(\mathbf{z} \mid \overline{C \sqcup \{c_i'\}} = \mathcal{H}_{\overline{C \sqcup \{c_i'\}}}(\mathbf{z}'), C \sqcup \{i\} = \mathcal{H}_{C \sqcup \{c_i'\}}(\mathbf{z})) = \\ f_\emptyset(\mathbf{z} \mid \overline{C \sqcup \{c_i'\}} = \mathcal{H}_{\overline{C \sqcup \{c_i'\}}}(\mathbf{z}'), C = \mathcal{H}_C(\mathbf{z}), \{c_i'\} = \mathcal{H}_{c_i'}(\mathbf{z})) = \\ f_\emptyset(\mathbf{z} \mid \overline{C} \setminus \{c_i'\} = \mathcal{H}_{\overline{C} \setminus \{c_i'\}}(\mathbf{z}'), C = \mathcal{H}_C(\mathbf{z}), \{c_i'\} = \mathcal{H}_{c_i'}(\mathbf{z}))\end{gathered} \tag{16}$$

为证明 $\mathbb{S}' \subseteq \mathbb{S}$，取某个 $\mathbf{z}_0, \mathbf{z}_0' \in \mathcal{Z}$。现在证明，对于任意此类选择的$\mathbf{z}_0, \mathbf{z}_0'$，以下结论成立:

$$f_{C'}(\mathbf{z}_0 \mid \overline{C'} = \mathcal{H}_{\overline{C'}}(\mathbf{z}_0')) \subseteq \mathbb{S} \tag{17}$$

由于$\mathbf{z}_0' \in \mathcal{Z}'$、$\mathbf{z}_0 \in \mathcal{Z} \subseteq \mathcal{Z}'$和 $\mathcal{H}_{\overline{G}}(\mathcal{Z}')$在拼接下封闭，因此根据定义，固定 $C \setminus \{c_i'\}$ 的激活值为 $\mathcal{H}_{C \setminus \{c_i'\}}(\mathbf{z}_0') \in \mathcal{H}_{C \setminus \{c_i'\}}(\mathcal{Z}')$，并固定 $\{c_i'\}$ 的激活值为$\mathcal{H}_{\{c_i'\}}(\mathbf{z}_0) \in \mathcal{H}_{\{c_i'\}}(\mathcal{Z}')$，会得到一个激活值 $\alpha \in \mathcal{H}_{C \setminus \{c_i'\} \cup \{c_i'\}}(\mathcal{Z}') = \mathcal{H}_C(\mathcal{Z}')$。
因此，我们得到:

$$\begin{gathered}f_{C'}(\mathbf{z}_0 \mid \overline{C'} = \mathcal{H}_{\overline{C'}}(\mathbf{z}_0')) = \\ f_{C \sqcup \{c_i'\}}(\mathbf{z}_0 \mid \overline{C \sqcup \{c_i'\}} = \mathcal{H}_{\overline{C \sqcup \{c_i'\}}}(\mathbf{z}_0')) = \\ f_\emptyset(\mathbf{z}_0 \mid \overline{C} \setminus \{c_i'\} = \mathcal{H}_{\overline{C} \setminus \{c_i'\}}(\mathbf{z}_0'), C = \mathcal{H}_C(\mathbf{z}_0), \{c_i'\} = \mathcal{H}_{c_i'}(\mathbf{z}_0)) = \\ f_\emptyset(\mathbf{z}_0 \mid \overline{C} = \alpha, C = \mathcal{H}_C(\mathbf{z}_0))\end{gathered} \tag{18}$$

由于我们已经证明了$\alpha \in \mathcal{H}_C(\mathcal{Z})$，且 $\mathcal{H}_C(\mathbf{z}_0) \in \mathcal{H}_C(\mathcal{Z})$，因此可得:

$$\begin{gathered}f_{C'}(\mathbf{z}_0 \mid \overline{C'} = \mathcal{H}_{\overline{C'}}(\mathbf{z}_0')) = \\ f_\emptyset(\mathbf{z}_0 \mid \overline{C} = \alpha, C = \mathcal{H}_C(\mathbf{z}_0)) \in \\ \{f_C(\mathbf{z} \mid \overline{C} = \mathcal{H}_{\overline{C}}(\mathbf{z}')) : \mathbf{z} \in \mathcal{Z}, \mathbf{z}' \in \mathcal{Z}'\} = \mathbb{S}\end{gathered} \tag{19}$$

这证明了 $\mathbb{S}' \subseteq \mathbb{S}$，从而完成了该引理的证明。

□

现在，为了完成该命题的证明，我们回顾一下，我们已经证明以下结论成立（并且现在可以根据我们对 $\mathbb{S}$的新定义改写这一表达式）:

$$\begin{aligned} \max_{\mathbf{z}\in\mathcal{Z},\mathbf{z}'\in\mathcal{Z}'} \quad & \left\| f_C(\mathbf{z} \mid \overline{C} = \mathcal{H}_{\overline{C}}(\mathbf{z}')) - f_G(\mathbf{z}) \right\|_p \leq \delta \iff \\ \max_{\mathbf{z}\in\mathcal{Z},\mathbf{y}\in\mathbb{S}} \quad & \| \mathbf{y} - f_G(\mathbf{z}) \|_p \leq \delta \end{aligned} \tag{20}$$

我们也已为导出矛盾而假设以下结论成立，并且同样可以根据我们对 $\mathbb{S}'$的新定义进一步改写这一项：

$$\begin{aligned} \max_{\mathbf{z}\in\mathcal{Z},\mathbf{z}'\in\mathcal{Z}'} \quad & \left\| f_{C'}(\mathbf{z} \mid \overline{C'} = \mathcal{H}_{\overline{C'}}(\mathbf{z}')) - f_G(\mathbf{z}) \right\|_p > \delta \iff \\ \max_{\mathbf{z}\in\mathcal{Z},\mathbf{y}\in\mathbb{S}'} \quad & \| \mathbf{y} - f_G(\mathbf{z}) \|_p > \delta \end{aligned} \tag{21}$$

然而，由于我们已在引理 1 中证明了 $\mathbb{S}' \subseteq \mathbb{S}$，因此我们知道：

$$\max_{\mathbf{z}\in\mathcal{Z},\mathbf{y}\in\mathbb{S}} \quad \| \mathbf{y} - f_G(\mathbf{z}) \|_p \geq \max_{\mathbf{z}\in\mathcal{Z},\mathbf{y}\in\mathbb{S}'} \quad \| \mathbf{y} - f_G(\mathbf{z}) \|_p \tag{22}$$

这与公式 20 和 21 相矛盾，从而意味着 Φ 的单调性，并完成该命题的证明。

□

B.6 命题 6 的证明

命题 6。 如果将条件 $\Phi(C, G)$ 设定为验证 C 是否关于$\langle f_G、\ \mathcal{Z}\rangle$（定义 *1*）具有输入鲁棒性，并且关于 $\langle f_G、\ \mathcal{Z}'\rangle$（定义 *2*）具有修补鲁棒性，那么如果 $\mathcal{Z} \subseteq \mathcal{Z}'$ 和 $\mathcal{H}_G(\mathcal{Z}')$ 在拼接下封闭，则算法 *1* 将收敛到一个子集最小电路。

证明。 该结论直接由命题 4 和命题 5 得出。由于 $\mathcal{Z} \subseteq \mathcal{Z}'$，命题 5 蕴含 $\Phi(C,\ G)$ 是单调的。根据命题 4，可知算法 1 将收敛到关于 Φ 的子集最小电路 C。

□

B.7 命题 7 的证明

命题 7。 给定某个模型 f_G和一个单调谓词 Φ，关于 Φ 的所有电路阻断集的最小击中集，是一个基数最小电路 C，使得 $\Phi(C, G)$ 为真。此外，所有使 $\Phi(C, G)$ 为真的电路 $C \subseteq G$ 的最小击中集，是关于 Φ的基数最小阻断集。

证明。 在证明命题 7（它确立了最小击中集（MHS）与基数最小性之间的联系）之前，我们先回顾最小击中集的定义：
定义 2（最小击中集（MHS））。 给定全集 U上的一组集合 $\mathcal{S}$，击中集 $H \subseteq U$ 是关于 $\mathcal{S}$ 的一个集合，满足

$$\forall S \in \mathcal{S}, \quad H \cap S \neq \emptyset.$$

如果击中集 H 的任何子集都不是击中集，则称其为极小；如果其基数在所有击中集中最小，则称其为最小。

我们还证明了关于阻塞集的以下等价关系。
引理 2. 在单调谓词 Φ 下，相对于 G 的电路阻断集恰好与移除后会破坏完整模型忠实性的集合相同，即 $C' \subseteq G$ 是阻塞集，当且仅当 $\neg\Phi(G \setminus C',\ G)$。

证明。 我们证明两个方向，先从 $\Rightarrow$ 开始。设 C' 是一个阻塞集。根据阻塞集的定义（第 4.4 节），对于所有 $C \subseteq G$，$\Phi(C \setminus C',\ G)$ 均不成立，其中自然也包括 $C = G$。因此，$\neg\Phi(G \setminus C',\ G)$。对于反方向 $\Leftarrow$，考虑一个集合 $C' \subseteq G$，从 G 中移除该集合会破坏完整模型 G 的忠实性，即 $\neg\Phi(G \setminus C',\ G)$。为导出矛盾，假设 C'

不是一个阻塞集。那么，存在某个 $C \subseteq G$，使得 $\Phi(C \setminus C',\ G)$ 成立。根据 Φ 的单调性，由于 $C \subseteq G$，我们有 $\Phi(C \setminus C',\ G) \Rightarrow \Phi(G \setminus C',\ G)$，这与我们的假设矛盾。

This 完成证明，证明在单调性条件下，阻塞集恰好就是那些移除后会破坏完整模型 G 忠实性的集合。

□

现在我们来证明，阻塞集的最小击中集是基数最小忠实电路。设 C 是阻塞集集合的一个最小击中集（MHS）$\mathcal{B}$。

现在我们来证明命题 7，即阻塞集的最小击中集（MHS）是基数最小忠实电路。令 $\mathcal{B}$ 表示阻塞集的集合；根据引理 2，在单调性假设下，该集合与那些移除后会破坏完整模型 G 忠实性的集合一致。设 C 是 $\mathcal{B}$ 的一个最小击中集。

反设 $\neg\Phi(C,\ G)$。令 $B^\star := G \setminus C$。那么 $\Phi(G \setminus B^\star,\ G) = \Phi(C,\ G)$ 为假，因此 $B^\star \in \mathcal{B}$。然而，根据定义 $C \cap B^\star = \emptyset$，这与 C 命中 $\mathcal{B}$ 中的每个集合相矛盾。因此，$\Phi(C,\ G)$ 成立。

现在继续证明最小性。假设存在 $C' \subseteq G$，使得 $\Phi(C',\ G)$ 且 $|C'| < |C|$。我们断言 C' 也是 $\mathcal{B}$ 的击中集，这与 C 是一个最小击中集（ MHS ）的最小性相矛盾。事实上，若存在某个 $B \in \mathcal{B}$ 满足 $C' \cap B = \emptyset$，那么 $C' \subseteq G \setminus B$；根据 Φ 的单调性，我们有 $\Phi(G \setminus B,\ G)$，这与 $B \in \mathcal{B}$ 相矛盾。因此，C' 命中 $\mathcal{B}$ 中的所有集合，这与 C 是一个 MHS 相矛盾。

对于证明的第二部分，令 $\mathcal{C} := \{\, C \subseteq G :\ \Phi(C,\ G) \,\}$，并令 B 为 $\mathcal{C}$的一个最小击中集。反设它不是最小的，即 $\Phi(G \setminus B,\ G)$ 成立。于是 $C^\star := G \setminus B$ 是一个忠实电路，这意味着 $C^\star \in \mathcal{C}$。然而，根据定义，$C^\star \cap B = \emptyset$，这与 B 击中 $\mathcal{C}$中的每个集合矛盾。

最后，反设存在一个阻塞集 B'，满足 $|B'| \le |B|$。令 $C \in \mathcal{C}$。若 $C \cap B' = \emptyset$，则 $C \subseteq G \setminus B'$；由 Φ的单调性可得 $\Phi(G \setminus B',\ G)$ 成立，这与 B' 是一个阻塞集矛盾。因此，$C \cap B' \neq \emptyset$，所以 B' 是 $\mathcal{C}$的一个击中集。但由于$|B'| \le |B|$，这与 B 作为最小击中集的最小性矛盾。因此，B 在基数意义上是最小的阻塞集。

□

B.8 命题 8 的证明

命题 8. 给定模型 f_G 和单调谓词 Φ，算法 *2* 计算出子集 C，其大小是基数最小电路的下界，且 $\Phi(C, G)$ 为真。当 t_{max} 值足够大时，该算法会精确地 收敛到基数最小电路。

证明. 我们先证明关于基数意义上的最小电路规模下界的部分。设 C 是算法 2 对某个 $t_{\max}$ 的输出。根据定义，C 是算法累积的阻塞集的最小击中集，记为 $\mathcal{B}_{t_{\max}}$。

反设 $|C|$ 不是基数最小电路大小的下界。这意味着存在一个忠实电路 C'，满足 $\Phi(C', G)$ 且 $|C'| \le |C|$。由 C 作为击中集的最小性可知，C' 不是击中集。因此，存在某个 $B \in \mathcal{B}_{t_{\max}}$ 使得 $C' \cap B = \emptyset$。这意味着 $C' \subseteq G \setminus B$，并且根据 Φ 的单调性，我们得到 $\Phi(G \setminus B, G)$，这与 B 是阻塞集矛盾。

现在继续证明的第二部分，即当 $t_{\max}$ 足够大时，算法会收敛到基数最小电路。对于 $t_{\max} = |G|$，算法会遍历所有可能的阻塞集。因此，得到的输出 C 是所有电路阻塞集的最小击中集，根据命题 7，我们可以得出 C 是一个基数最小电路。

□

C 极小性保证：算法与示例

C.1 贪心穷举迭代电路发现

这里我们给出一种算法，用于穷尽式地反复遍历电路组件，并移除删除后不会违反忠实性的任何组件，直至无法再移除更多组件（算法 3）。

算法 3 贪心 y 穷尽式电路发现

y1: **输入**模型 f_G，电路忠实性谓词 Φ2: $C \leftarrow G$ 按给定的某种元素排序（例如，反向拓扑排序）3: 已更改 $\leftarrow$ **真**4: **当**已更改 **为真时**5: 已更改 $\leftarrow$ **假**6: **对所有** $i \in C$ **执行**7: **如果** $\Phi(C \setminus \{i\}, G)$ **则**8: $C \leftarrow C \setminus \{i\}$9: 已更改 $\leftarrow$ **真**10: **结束如果**11: **结束循环**12: **结束当**13: **返回** C

C.2 贪心电路发现：准最小电路的二分搜索

S

我们在算法中形式化了 (Adolfi 等人，2025) 引入的二分搜索过程 4.

算法 4贪心电路发现二分搜索

1：输入：模型 f_G**，电路忠实性谓词** Φ**，其中** $\Phi(G, G)$ $\wedge$ $\neg\Phi(\emptyset, G)$ **2：** $C \leftarrow G$**，下界** $\leftarrow 0$**，上界** $\leftarrow |G|$**3：当 high –low** > 1 **时 4：mid** $\leftarrow \lfloor$**(low +high)**$/2\rfloor$**5：** $C_{\text{mid}} \leftarrow G \setminus G[1$ **: mid**]**6：如果** $\Phi(C_{\text{mid}}, G)$**，则 7：low** $\leftarrow$ **mid;** $C \leftarrow C_{\text{mid}}$**8：否则 9：high** $\leftarrow$ **mid 10：结束 if 11：结束 while 12：返回** C

C.3 玩具示例：最小性概念

为说明定义 3、4、5、6（第 4 节）中引入的四种最小性概念之间的区别，我们构造了一个简单的布尔玩具网络。

为简化说明，我们使用带有异或门的布尔电路来展示不同的最小性概念。尽管这种抽象让示例更容易理解，但这并不失一般性，因为布尔门都可以用 ReLU 激活等价地表示。

（具体而言，异或门满足

$$x_1 \oplus x_2 \;=\; \mathrm{ReLU}(x_1 - x_2) + \mathrm{ReLU}(x_2 - x_1),$$

对于 $x_1,\ x_2 \in \{0, 1\}$ 当 $x_1 = x_2$ 时，两项都为零；而当 $x_1 \neq x_2$ 时，恰好有一项等于 1。）

我们强调，这种编码并不属于计算图，计算图可以独立定义。因此，我们的布尔玩具电路（图 6）可以视为一个小型前馈 ReLU

网络。尽管规模很小，这个网络仍清晰地区分了基数、子集、局部和准最小电路的概念。

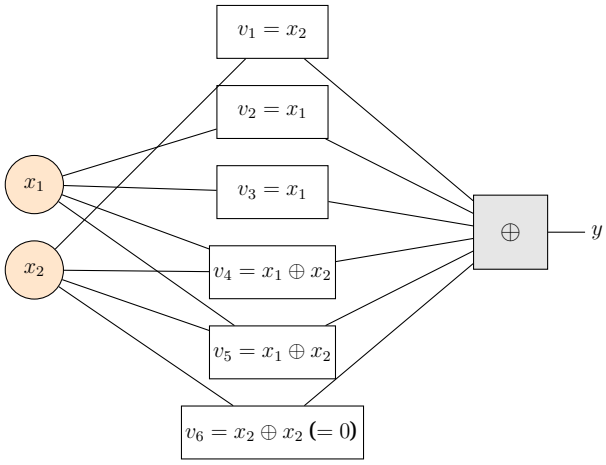


图 6：采用异或聚合的布尔玩具网络：$y = v_1 \oplus v_2 \oplus v_3 \oplus v_4 \oplus v_5$。借助 $v_1 = x_2$、$v_2 = v_3 = x_1$、$v_4 = v_5 = x_1 \oplus x_2$ 和 $v_6 = x_2 \oplus x_2 = 0$，完整模型计算 $f_G(x_1, x_2) = x_2$。

说明。为清晰起见并简化表述，我们定义一个布尔网络 G，其输入为 $(x_1, x_2) \in \{0,1\}^2$。该网络 G 由六个组件组成，其顶点集记为$V_G = \{v_1, v_2, v_3, v_4, v_5, v_6\}$，并通过异或进行聚合（图 6）：

$y = v_1 \oplus v_2 \oplus v_3 \oplus v_4 \oplus v_5 \oplus v_6$，其中 $v_1 = x_2$、$v_2 = v_3 = x_1$、$v_4 = v_5 = x_1 \oplus x_2$、$v_6 = x_2 \oplus x_2$

完整网络计算 $f_G(x_1, x_2) = x_2$，因为

$$\begin{aligned} f_G(x_1, x_2) &= x_2 \oplus x_1 \oplus x_1 \oplus (x_1 \oplus x_2) \oplus (x_1 \oplus x_2) \oplus (x_2 \oplus x_2) \\ &= x_2 \oplus \underbrace{(x_1 \oplus x_1)}_{=0} \oplus \underbrace{[(x_1 \oplus x_2) \oplus (x_1 \oplus x_2)]}_{=0} \oplus \underbrace{(x_2 \oplus x_2)}_{=0} = x_2. \end{aligned}$$

我们将忠实性谓词设为

$$\Phi(C, G) := f_C(x_1, x_2) = f_G(x_1, x_2) = x_2 \quad \forall (x_1, x_2),$$

也就是说，若 C 对所有输入都计算出与 G 相同的输出，则它是忠实的。

这一构造清晰地区分了四种最小性概念：

对于每个电路 C，我们验证它满足 $f_C(x_1, x_2) = x_2$，并说明它为何满足（或不满足）相应的最小性条件。

- **基数最小：** $C_{\text{基数}} = \{v_1\}$。它计算 $f_{C_{\text{基数}}} = v_1 = x_2 = f_G$。不存在组件更少且忠实的电路。
- **子集最小：** $C_{\text{子集}} = \{v_2, v_4\}$。其计算结果为
  $$v_2 \oplus v_4 = x_1 \oplus (x_1 \oplus x_2) = x_2.$$
  移除任何组件都会导致正确性失效：$\{v_2\} = x_1 \neq x_2$ 和 $\{v_4\} = x_1 \oplus x_2 \neq x_2$。
- **局部最小：** $C_{\text{局部}} = \{v_1, v_2, v_3\}$。其计算结果为
  $$v_1 \oplus v_2 \oplus v_3 = x_2 \oplus x_1 \oplus x_1 = x_2.$$
  移除任何单个组件都会导致正确性失效：$\{v_1, v_2\} = x_2 \oplus x_1 \neq x_2$、$\{v_1, v_3\} = x_2 \oplus x_1 \neq x_2$、$\{v_2, v_3\} = x_1 \oplus x_1 = 0 \neq x_2$。然而，移除两个组件后仍可能留下一个正确的单元素集合（例如 $\{v_1\}$），因此它不是子集最小。
- **准最小：** $C_{\text{准}} = \{v_1, v_2, v_3, v_4, v_5\}$。其计算结果为
  $$v_1 \oplus v_2 \oplus v_3 \oplus v_4 \oplus v_5 = x_2.$$
  该电路包含一个必要组件（v_1），而其余组件可以通过多种组合移除，且不影响输出（例如，$v_2 \oplus v_3 = 0$ 和$v_4 \oplus v_5 = 0$）。因此，它是忠实的，但不满足任何更严格概念下的最小性。

D 基准测试、模型与架构规范

我们在神经网络验证领域的四个标准基准测试上评估了我们的方法：三个分类基准测试（CIFAR-10 (Krizhevsky & Hinton, 2009)、GTSRB (Stallkamp et al., 2011) 和 MNIST (Lecun et al., 1998)），以及一个回归基准测试 TaxiNet (Julian et al., 2020)。
对于每个基准测试，我们按照模型的自然粒度执行电路发现：卷积神经网络中的卷积滤波器（或 通道）以及全连接网络中的神经元。每种粒度下的组件数量汇总于表 3。

表 3：各基准测试模型中用于电路发现的粒度和组件数量。

数据集	**模型**	**考察粒度**	**# 组件**
MNIST	FC	神经元	31
GTSRB	CNN	滤波器	48
CIFAR-10	ResNet	滤波器	72
TaxiNet	CNN	滤波器	8

数据选择。 对于输入和修补鲁棒性实验（附录 E、F），我们从测试集抽样，仅使用预测正确的输入（回归情况下则使用误差较小的输入），为每个基准测试构建了至少 100 个批次。在分类任务中，每个批次包含来自同一类别的 $k = 3$ 个样本，并在各类别间均匀分布。在回归任务中，批次$k = 3$ 取自绝对误差低于 0.2 的输入，并排除相对于模型性能偏差较大的样本。具体而言，我们为**CIFAR-10** 和**MNIST**各抽样了 100 个批次（每类 10 个），为**GTSRB** 抽样了 129 个批次（43 个类别，每类 3 个），并为回归基准测试**TaxiNet**抽样了 100 个批次。

对于极小性保证实验（第 5.3 小节，附录 G），我们使用了 50 个单元素集合批次（$k = 1$），具体做法是从上述每个 MNIST 批次中选择一个样本，从而保持类别分布均衡。

D.1 CIFAR-10

对于 CIFAR-10 基准测试（Krizhevsky & Hinton，2009），我们使用 ResNet2b 模型，该模型源自 VNN-COMP 神经网络验证竞赛（ Bak 等人，2021 ）。该残差网络由一个初始卷积层、两个残差块和一个输出 10 个类别的全连接分类头组成。该模型总计包含 72 个滤波器（也称为通道）。

层 / 块	输出维度	详细信息
输入	$32 \times 32 \times 3$	CIFAR-10图像
Conv1	$16 \times 16 \times 8$	3×3, 步长 2
残差块 *1*		
Conv1	$8 \times 8 \times 16$	3×3, 步长 2
ReLU	$8 \times 8 \times 16$	非线性
Conv2	$8 \times 8 \times 16$	3×3, 步幅1
跳跃连接	$8 \times 8 \times 16$	恒等映射/投影
输出	$8 \times 8 \times 16$	相加 +ReLU
残差块*2*		
Conv1	$8 \times 8 \times 16$	3×3, 步幅1
ReLU	$8 \times 8 \times 16$	非线性
Conv2	$8 \times 8 \times 16$	3×3, 步幅1
跳跃连接	$8 \times 8 \times 16$	恒等映射
输出	$8 \times 8 \times 16$	相加 +ReLU
展平	1×2048	–
Linear1	$2048 \to 100$	ReLU
Linear2	$100 \to 10$	输出逻辑值

表 4：Bak 等（2021）在 CIFAR-10 基准测试中使用的 ResNet2b 模型完整架构

D.2 GTSRB

德国交通标志识别基准（GTSRB）Stallkamp 等（2011）是一个大规模图像分类数据集，包含超过 50,000 张交通标志图像，涵盖 43 个类别，图像采集时的光照和天气条件各不相同。

我们采用了近期可解释性研究中使用的 GTSRB-CNN 模型（Bassan et al., 2025a）。该架构是一个卷积网络，包含两个使用 ReLU 激活和平均池化的卷积层，后接两个全连接层。它输出 43 个交通标志类别的逻辑值。
GTSRB-CNN 的两个卷积层总共包含 48 个滤波器（16 + 32） .

层	输出维度	详细信息
输入	$32 \times 32 \times 3$	GTSRB 图像
Conv1	$32 \times 32 \times 16$	3×3，填充 1，ReLU
平均池化层1	$16 \times 16 \times 16$	2×2
Conv2	$16 \times 16 \times 32$	3×3，填充 1，ReLU
平均池化层2	$8 \times 8 \times 32$	2×2
展平	1×2048	–
FC1	$2048 \to 128$	ReLU
FC2	$128 \to 43$	输出逻辑值

表5：Bassan et al., 2025a 中使用的 GTSRB-CNN 模型架构。

D.3 MNIST

我们使用了一个简单经典的全连接前馈网络进行 MNIST 分类，并针对该任务的简易性对其进行了训练。该模型在测试集上的准确率达到 95.20%。它由两个采用 ReLU 激活的隐藏层组成，大小分别为 13 和 11，后接一个线性输出层；总计包含 **31** 个非输入神经元和 10,479 个可训练参数。

层	维度	激活函数
输入	$28 \times 28 = 784$	–
全连接层（fc1）	$784 \to 13$	ReLU
全连接层（fc2）	$13 \to 11$	ReLU
全连接层（fc3）	$11 \to 10$	–

表 6：全连接 MNIST 网络的架构。该模型总共有 31 个隐藏神经元，我们将其作为电路发现的粒度。

D.4 TAXINET

TaxiNet 数据集（Julian 等人，2020）由美国国家航空航天局开发，用于基于视觉的飞机滑行，由合成跑道图像与连续控制目标配对组成。与分类基准不同，TaxiNet 是一项回归任务：该模型预测与飞行控制变量相对应的实值输出。

在我们的实验中，我们采用了 VeriX 框架（Wu et al., 2023a）中提出、并随后用于其他可解释性研究（Bassan et al., 2025a）的 TaxiNet CNN 回归模型。该卷积网络包含 8 个滤波器，均方误差（MSE）为 0.848244，均方根误差（RMSE）为 0.921。

层	输出维度	详细信息
输入	$27 \times 54 \times 1$	TaxiNet 图像
Conv1	$27 \times 54 \times 4$	3×3，填充 1，ReLU
Conv2	$27 \times 54 \times 4$	3×3，填充 1，ReLU
展平	1×5832	–
FC1	$5832 \to 20$	ReLU
FC2	$20 \to 10$	ReLU
FC3	$10 \to 1$	回归输出

表 7：用于 TaxiNet 的卷积神经网络回归模型架构，遵循 VeriX 框架（Wuetal., 2023a）。

实验细节

输入鲁棒性认证

在本实验中，我们评估发现的电路在连续输入邻域 $\mathcal{B}_\epsilon^p(\mathbf{x})$ 上的鲁棒性，如第 3.1 节所述。我们比较算法 1 中迭代电路发现过程的两个变体，它们的消除准则各不相同：

1. **基于采样的电路发现**：在每一步直接评估输入批次上的度量。 2. **可证明输入鲁棒的电路发现**：认证该度量在整个输入邻域内均成立（定义 1）。

该流程依次遍历网络组件，并在每一步决定保留还是移除某个组件。正如 Conmy等（2023）所指出的，遍历顺序会影响最终得到的电路。按照他们的方法，我们从较后的全连接层或卷积层开始，向较早的层推进，并按字典序排列每层中的神经元或滤波器。为保持一致，我们将所有非电路组件的修补方案固定为零值修补。

E.1 方法

E.1.1 用于验证的孪生网络

为将电路发现与形式化验证相结合，我们构建了一个孪生网络，将完整模型与候选电路配对，并输出两个网络的逻辑值**拼接** 结果。
这一孪生网络形式为以下两种设置中使用的神经网络验证提供了接口：

- **可证明输入鲁棒的电路发现：** 在每次消除步骤中进行认证，确保候选电路在连续输入邻域内满足该度量，从而保证鲁棒性得以保留。
- **评估：** 在发现电路后（通过采样方法或基于可证明性的方法）进行验证，确保所得电路在相同邻域内具有鲁棒性。

输出度量。 为保持一致性，我们所有实验在基于采样的方法和具有可证明输入鲁棒性的发现方法中，都采用相同的输出度量。我们测量逻辑值之差，并要求该差保持在容差 δ范围内。

对于输入 z上的分类任务，我们关注黄金类别的 logit，其索引为 k，并要求电路 C 的预测与完整模型 G的预测之差至多为 δ：

$$|f_G(z)[k] - f_C(z)[k]| \leq \delta,$$

其中，绝对值表示对应于第 k个条目的单维向量的 ℓ_p-范数。在基于采样的方法中，该条件在采样得到的批次上进行评估；而在可证明（孪生）设置中，则针对孪生编码拼接后的逻辑值对该条件进行验证。例如，在一个 10 类分类任务中，孪生网络输出的验证约束为：

$$|\,\text{logits}_{[:10]}[k] - \text{logits}_{[10:]}[k]\,| \leq \delta.$$

其中，前 10 个条目对应完整模型 G 的逻辑值，后面的条目对应电路 C的逻辑值。

在回归设置中（例如 TaxiNet），同一原则适用于完整输出（标量值预测），衡量模型与电路之间的绝对差值。在这两种情况下，该度量直接实例化了鲁棒性定义（定义 1、2）中使用的范数度量。

输入邻域。 在我们的设置中，邻域 $\mathcal{B}_{\epsilon_p}^{\infty}$**(x)** 在输入空间中相对于 L^∞ 范数定义。对于全连接模型（例如 MNIST），输入被展平为向量，扰动球在该表示上定义。对于卷积模型（例如 CIFAR-10、GTSRB、TaxiNet），输入为多通道张量，邻域独立地应用于每个通道和空间位置。

E.1.2 验证与实验设置（输入鲁棒性）

由于针对线性性质对分段线性激活函数网络进行可靠且完备的验证是 NP难的 Katz 等人（2017），部分查询可能无法在规定时间内完成；在此类情况下，结果被报告为 未知。在实践中，使用 α, β-CROWN 验证器时，我们将每个查询的分支定界时间限制为**45**秒。

为确保公平，我们仅针对基于采样的方法已得出鲁棒性检查结论（鲁棒或非鲁棒，不包括超时）的批次报告鲁棒性统计数据。在主论文的结果中，MNIST、GTSRB、CIFAR-10 和 TaxiNet 的实例超时率分别为 1%、1.6%、1% 和 5%。在邻域大小变化研究 E.2.1 中，也观察到了相近的比率（MNIST 平均为 0.5%，CIFAR-10 为 3%，TaxiNet 为 2.7%）；在容差水平变化 δ E.2.2 中亦如此（TaxiNet 平均为 8.6%，MNIST 为 3.6%）。
在 MNIST、GTSRB 和 CIFAR-10 上进行的实验均采用统一的硬件配置：搭载 48GB NVIDIA L40S GPU 和 2 核、16 GB CPU。对于 TaxiNet 模型，我们仅使用了一台配备 2 核、36 GB CPU 的机器。

E.2 鲁棒性评估与参数变化

我们使用上述的 李生网络 进行形式化验证，评估两种方法发现的电路在邻域 $\mathcal{B}^{\infty}_{\epsilon_p}(\mathbf{x})$ 内的鲁棒性。参数 ϵ_p 控制邻域大小：若该值过小，得到的电路会很简单；若过大，扰动则会变得不切实际且属于分布外。我们选择 ϵ_p 取值范围时，参考了以往验证工作中常用的范围，或根据经验选取，以平衡电路大小和鲁棒性。此外，我们还改变参数 ϵ_p 和 δ，分析它们的影响。

E.2.1 输入邻域大小的变化 ϵ_p

我们固定容差 δ，并改变输入邻域大小 ϵ_p。对于 CIFAR-10 和 MNIST，我们使用$\delta = 2.0$; 对于 GTSRB，使用 $\delta = 5.0$。对于 TaxiNet 回归模型，我们设定 $\delta = 0.92$ （模型的均方根误差，RMSE），以反映其典型预测尺度（更大的偏差会使电路相较于完整模型偏离真实值更多）。结果见表 8。灰色高亮的行对应于主论文中选用的结果。

数据集	ϵ_p	基于采样的电路发现			可证明输入鲁棒的电路发现		
		时间（秒）	大小（$\lvert C\rvert$）	鲁棒性（%）	时间（秒）	大小（$\lvert C\rvert$）	鲁棒性 (%)
MNIST (δ=2.0)							
	0.005	0.015 ±0.003	12.57 ±2.29	**46.0** ±5.0	677.36 ±183.65	14.51 ±2.41	**100.0** ±0.0
	0.009	0.016 ±0.013	12.56 ±2.30	**25.3** ±4.4	663.60 ±181.13	15.76 ±2.25	**100.0** ±0.0
	0.010	0.309 ±0.889	12.56 ±2.30	**19.2** ±4.0	611.93 ±97.14	15.84 ±2.33	**100.0** ±0.0
	0.050	0.027 ±0.065	12.57 ±2.29	**0.0** ±0.0	1700.05 ±562.36	28.75 ±6.67	**100.0** ±0.0
TaxiNet (δ=0.92)							
	0.001	0.040 ±0.146	5.76 ±0.77	**62.2** ±4.9	201.81 ±34.16	6.07 ±0.71	**100.0** ±0.0
	0.005	0.010 ±0.002	5.77 ±0.80	**9.5** ±3.0	180.00 ±40.39	6.82 ±0.46	**100.0** ±0.0
	0.010	0.010 ±0.002	5.78 ±0.79	**2.0** ±1.4	271.03 ±54.23	7.91 ±0.32	**100.0** ±0.0
CIFAR-10 (δ=2.0)							
	0.007	0.035 ±0.001	16.70 ±9.48	**73.7** ±4.4	2104.13 ±118.95	17.96 ±9.90	**100.0** ±0.0
	0.012	0.116 ±0.367	16.91 ±9.12	**58.1** ±5.1	2226.34 ±103.71	18.88 ±9.21	**100.0** ±0.0
	0.015	0.228 ±0.517	16.47 ±9.08	**46.5** ±5.0	2970.85 ±874.23	19.18 ±10.16	**100.0** ±0.0
GTSRB（δ=5.0）							
	0.001	0.111 ±0.329	28.91 ±4.69	**27.6** ±4.0	991.08 ±162.91	29.59 ±4.45	**100.0** ±0.0

表 8：在固定容差 δ 下改变输入邻域大小 ϵ_p 的效果。报告的数值为均值和标准差。对于鲁棒性（二元变量），我们报告标准误（SE）。粗体数值表示鲁棒性百分比。灰色高亮行对应于主论文中选取的结果。

E.2.2 容差水平的变化 δ

我们改变容差 δ，同时将输入邻域大小 ϵ_p 固定为主论文中使用的数据集特定值（MNIST: ϵ_p=0.01，TaxiNet： ϵ_p=0.005）。结果见表 9。与主论文结果对应的行以灰色突出显示。

数据集	δ	基于采样的电路发现			可证明输入鲁棒的电路发现		
		时间（秒）	大小（$\lvert C\rvert$）	鲁棒性（%）	时间（秒）	大小（$\lvert C\rvert$）	鲁棒性（%）
MNIST（ϵ_p=0.01）							
	0.50	0.083 ± 0.329	19.72 ± 1.50	**15.6** ± 3.8	460.14 ± 36.59	22.02 ± 2.52	**100.0** ± 0.0
	2.00	0.309 ± 0.889	12.56 ± 2.29	**19.2** ± 4.0	611.93 ± 97.14	15.84 ± 2.33	**100.0** ± 0.0
	3.00	0.013 ± 0.001	9.44 ± 1.92	**34.0** ± 4.7	577.41 ± 35.84	12.27 ± 2.52	**100.0** ± 0.0
TaxiNet (ϵ_p=0.005)							
	0.50	0.038 ± 0.115	6.84 ± 0.86	**32.1** ± 5.2	93.08 ± 22.75	7.75 ± 0.46	**100.0** ± 0.0
	0.70	0.008 ± 0.001	6.15 ± 0.81	**14.8** ± 3.8	114.51 ± 23.95	7.31 ± 0.53	**100.0** ± 0.0
	0.92	0.010 ± 0.002	5.77 ± 0.80	**9.5** ± 3.0	180.00 ± 40.39	6.82 ± 0.46	**100.0** ± 0.0
	1.00	0.009 ± 0.002	5.57 ± 0.82	**9.5** ± 3.0	142.62 ± 24.77	6.66 ± 0.52	**100.0** ± 0.0
	1.20	0.009 ± 0.001	5.43 ± 0.96	**6.1** ± 2.4	155.03 ± 26.34	6.32 ± 0.59	**100.0** ± 0.0

表 9：容差水平 δ的变化，输入邻域大小 ϵ 固定为主要实验中使用的数据集特定值。报告值为均值及标准差。对于鲁棒性（一个二元变量），我们报告标准误（标准误）。以灰色高亮的行对应于主论文中选定的结果。

E.2.3 替代输出指标下的评估

为进一步评估我们框架的通用性，我们使用默认*Logit*差 准则之外的替代输出指标，重复鲁棒性评估。具体而言，我们考虑两种额外的表述：

- **一致获胜类：** 给定某个目标类 $t \in [d]$，确保获胜类在指定区域内保持一致。该度量直接针对预测类在整个输入域中的保持情况，广泛用于鲁棒性验证研究。该判据要求：

$$\forall \mathbf{z} \in \mathcal{Z}, \quad \operatorname*{argmax}_j (f_G(\mathbf{z}))_{(j)} = \operatorname*{argmax}_j (f_C(\mathbf{x} \mid \overline{C} = \alpha)(\mathbf{z}))_{(j)} = t.$$

  为提高灵活性，我们放宽这一要求，允许预测类在任何变化不超过次优类以上容差 $\delta \in \mathbb{R}^+$ 的情况下保持一致。当$\delta = 0$时，这一要求便恢复为原始定义。为使该阈值更有意义且易于解释，我们将 δ 设为可配置比例 $\alpha \in$ （即模型在未扰动输入上的原始获胜-次优差距的 $0, 1]$ ）。在我们的实验中，为求简便，我们仅要求获胜类与次优类之间满足一致性条件。

- **绝对值最大值：** 将所有输出维度上的最大绝对偏差限制在指定阈值内。该判据不保证类别不变性，但会约束整体输出漂移。形式上，我们要求：

$$\forall \mathbf{z} \in \mathcal{Z}, \quad \|f_G(\mathbf{z}) - f_C(\mathbf{x} \mid \overline{C} = \alpha)\|_\infty \leq \delta,$$

  确保每个 logit 的差异都不超过 δ。

我们使用与主要输入鲁棒实验相同的发现配置和 ϵ_p 来评估这两种指标。对于 logits差值指标，我们使用了与主要实验（表 1）相同的 δ。在获胜-次优设定中，我们设置了 $\alpha = 0.5$（保留原始间隔的一半）；对于绝对最大准则，我们使用了 $\delta = 4.0$。结果见表 10，该表比较了每种指标下的基于采样的 发现方法和可证明的 发现方法。

数据集	指标	基于采样的电路发现			可证明输入鲁棒的电路发现		
		时间（秒）	大小（$\lvert C \rvert$）	鲁棒性（%）	时间（秒）	大小（$\lvert C \rvert$）	鲁棒性（%）
MNIST	Logit差	0.31 ±0.89	12.56 ±2.30	**19.2** ±4.0	611.93 ±97.14	15.84 ±2.33	**100.0** ±0.0
	赢家-运行者	0.13 ± 0.60	5.18 ± 1.05	**73.0** ±4.4	638.57 ± 20.94	12.04 ± 5.79	**100.0** ±0.0
	绝对值最大值	0.014 ± 0.012	25.55 ± 2.90	**6.0** ±2.4	362.61 ± 55.73	28.11 ± 2.37	**100.0** ±0.0

表 10：在替代输出指标下对比电路发现方法。报告的数值为均值及标准差。对于鲁棒性（二元变量），我们报告标准误（SE）。所有方法均使用与主要实验相同的配置。

在所有指标下，都观察到了相同的整体趋势：可证明方法始终接近 100% 的鲁棒性，同时保持与基于采样的基线相当的电路大小，而后者的鲁棒性则显著较低。不同指标下的一致性表明，可证明方法的鲁棒性并非依赖于某个特定的输出度量，而是反映了该方法的一项稳定特征。

E.2.4 可证明鲁棒与基于采样的电路的覆盖分析

S

为更好地理解可证明鲁棒的过程所识别的电路与基于采样的方法所生成的电路之间的关系，我们在 MNIST 基准测试上针对多个鲁棒性半径 $\epsilon \in \{0.005, 0.007, 0.01, 0.02, 0.03, 0.04, 0.05\}$ 进行了明确的覆盖分析。所有其他设置，包括容差和指标定义，均与主要实验中使用的设置一致（如 E.1 节所述）。我们在一台配备 16 GB 随机存取存储器的双核 CPU 计算机上进行了这些实验。

对于每个扰动半径 ϵ_p，我们考察针对该半径导出的可证明输入鲁棒的电路 C_p 以及基于采样的电路 C_s，二者均在 100 个不同输入上获得（如我们的实验设置所示），从而对每个 ϵ_p 得到 100 个回路对。

对于这两个电路，我们计算：**(i)** 交集的大小 $|C_p \cap C_s|$，**(ii)** 可证明回路所独有的组件（仅可证明部分） $|C_p \setminus C_s|$，以及 **(iii)** 基于采样的电路所独有的组件（仅采样部分） $|C_s \setminus C_p|$。

我们对 100 对回路的这些量取平均，并报告其均值和标准差。为概括 C_p 与 C_s 在 ϵ_p上的整体相似性／差异性，我们还计算了标准集合相似度度量：交并比（IoU）、Dice系数以及两种非对称覆盖指标（可证明对采样覆盖和采样对可证明覆盖）。这些汇总趋势如图 7 所示。为进一步突出未重叠的组件，表 11 报告了其数量及其占全网络规模的百分比。

由于基于采样的方法不强制施加鲁棒性条件，其电路大小在 ϵ_p期间保持不变；而基于可证明的方法构造的电路则会随着 ϵ_p 增加而自然扩展，以保证认证鲁棒性。我们确实观察到，随着所需鲁棒性的提高，可证明鲁棒电路会纳入认证所必需的额外组件。

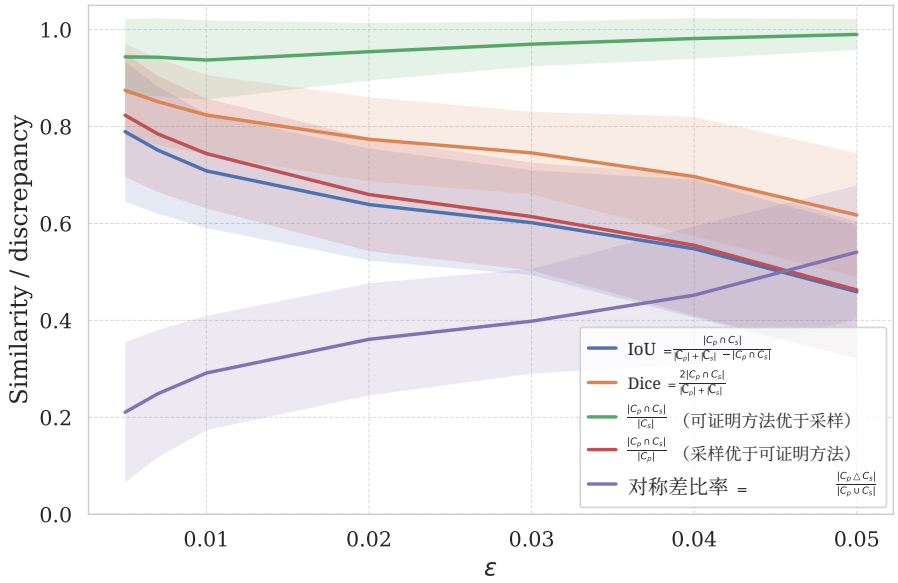


图 7：可证明鲁棒电路与基于采样的电路之间相似度和覆盖指标的比较。我们报告了对称度量（IoU、Dice系数）、对称差比率以及非对称覆盖比率（可证明对采样覆盖、采样对可证明覆盖），以展示重叠情况和方向性差异。

如图 7 所示，在所有 ϵ_p 取值下，可证明鲁棒电路都能稳定地找回基于采样的方法识别出的绝大多数单元；当扰动半径较小时（例如， $\epsilon_p = 0.005$处的 IoU 和 Dice ≈ 0.9 ），两者的一致性尤其高。随着 ϵ_p 增大，两个电路之间的重叠度逐渐降低（IoU 降至接近 0.5），这表明在较大的扰动下，基于采样的电路捕获到的可证明鲁棒电路部分比例较小。

数据集	ϵ_p	$\lvert C_p \setminus C_s \rvert$	占完整网络的百分比	$\lvert C_s \setminus C_p \rvert$	占完整网络的百分比
MNIST	0.005	2.64	7.76%	0.70	2.06%
	0.007	3.32	9.76%	0.74	2.18%
	0.010	4.10	12.06%	0.81	2.38%
	0.020	6.24	18.35%	0.58	1.71%
	0.030	7.73	22.74%	0.39	1.15%
	0.040	10.94	32.18%	0.24	0.71%
	0.050	16.11	47.38%	0.13	0.38%

表 11：可证明鲁棒电路 C_p 与基于采样的电路 C_s 之间的集合差。对于每个 ϵ_p，我们报告：(i) 仅出现在可证明鲁棒电路 $|C_p \setminus C_s|$ 中的单元数量；(ii) 仅出现在基于采样的电路 $|C_s \setminus C_p|$ 中的单元数量；以及 (iii) 相对于该数据集全网络规模的对应百分比。

这反映了可证明鲁棒电路会扩展，以满足更严格的鲁棒性要求。这一趋势在表 11 中也很明显：差异 $C_p \setminus C_s$ 随着 ϵ_p稳步增大，而$C_s \setminus C_p$ 在所有设置下都保持较小，并且在扰动半径较大时进一步减小——这表明基于采样的方法只会引入少量认证解并不需要的组件。

E.2.5 不同输入邻域大小下的运行时间权衡

我们旨在进一步分析可证明鲁棒方法的运行时间权衡。对于每个输入鲁棒半径 ϵ_p，我们运行该方法以获得相应的可证明鲁棒电路，并报告由增大 ϵ_p 所形成的不同输入邻域对应的这些电路的电路规模随时间变化曲线均值（标准差以阴影区域表示）。我们在 MNIST 基准测试上进行此分析，采用与附录 E.2.4 覆盖分析相同的扰动半径和实验设置。

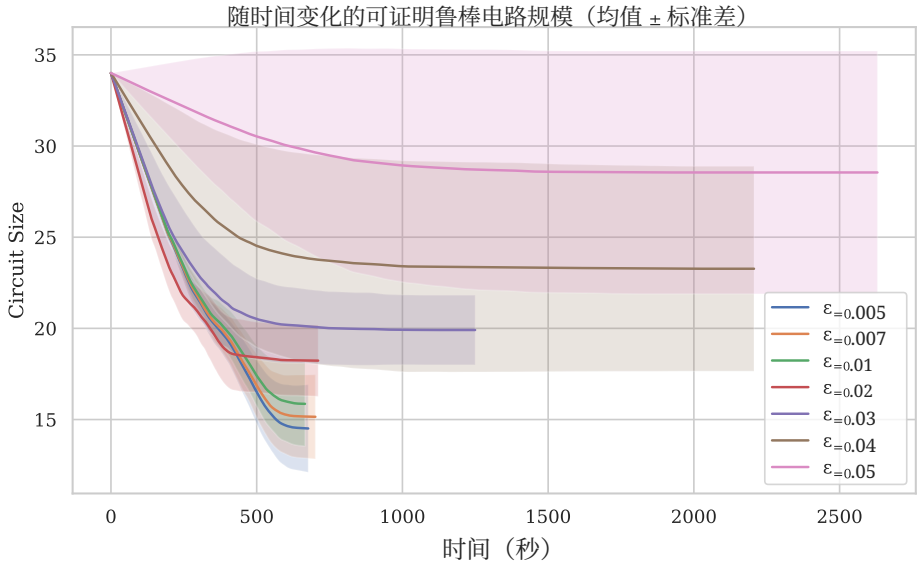


图 8：不同邻域半径下 MNIST 上可证明鲁棒电路大小随时间的变化 ϵ_p。阴影区域表示标准差。

在所有情况下，曲线都从时间 0 时的全网络规模开始，随后随着组件被剪枝而单调下降，直至程序终止。对于较小且彼此接近的 ϵ_p值，这些曲线呈现出非常相似的轨迹：在最初的 $\sim$ 300 秒内几乎线性下降，随后在约 $\sim$ 500 秒时进入稳定阶段。这些曲线的标准差

带也相近。正如预期，较大的邻域会带来较大的最终电路规模。

对于大得多的邻域（例如， $\epsilon_p = 0.03$-0.05），行为会发生变化：下降速度变慢，较晚才趋于稳定，且波动性（标准差）明显更高。此外，对于这些更大且间距更宽的 ϵ_p 值，我们观察到总运行时间明显增加（如曲线终止的位置所示），这反映了在更广泛的扰动区域内发现鲁棒认证电路所增加的复杂性。

E.2.6 发现的电路的定性观察

虽然我们的方法以形式化保证为核心，因此评估侧重于鲁棒性和最小性，但我们也对可证明鲁棒过程和基于采样的基线所发现的电路进行了简要的探索性考察。这项考察属于定性分析，仅旨在非正式地直观展示这两种电路的行为。

在这项分析中，我们考察了输入鲁棒性实验中生成的若干 GTSRB 通道级电路。回顾一下，对于每个批次（由同一类别的样本组成），我们都执行了可证明鲁棒发现（针对给定的 $\epsilon = 0.001$邻域）和基于采样的发现，从而得到两个电路。在以下示例中，我们选取大小相近的一对电路，其中基于采样的电路在经验上不鲁棒，而可证明鲁棒电路则经认证具有鲁棒性。随后，我们分析它们在该批次中一个具有代表性的干净输入以及其对应的对抗样本（一个受 $\epsilon = 0.001$ 限制的对抗扰动）上的行为。

为获得粗略的语义信号，我们将 Grad-CAM（Selvaraju 等人，2017）应用于以下模型的最后一个卷积层：(i) 完整模型，(ii) 可证明鲁棒电路，以及 (iii) 基于采样的电路。Grad-CAM 根据目标类别 logit 的全局平均梯度对空间激活进行加权，从而生成特定类别的重要性图。形式上，对于类别 c,

$$\alpha_k^{(c)} := \frac{1}{HW} \sum_{i,j} \frac{\partial y_c}{\partial A_{ij}^k}, \qquad \mathrm{CAM}_c(i,j) := \mathrm{ReLU}\left(\sum_k \alpha_k^{(c)} A_{ij}^k \right).$$

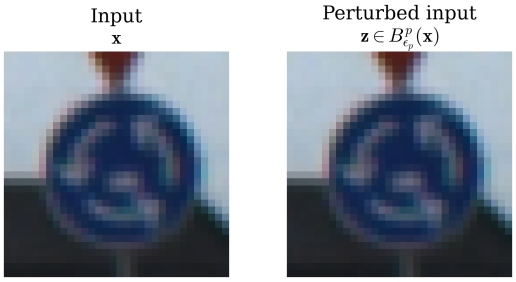
这里， y_c 表示目标类别 c （在我们的案例中为真实标签）的 logit， A^k 表示空间大小为 $H \times W$ 的第 k 个激活图（即滤波器 k 的输出）， $\alpha_k^{(c)}$ 表示通过对梯度 $\partial y_c / \partial A_{ij}^k$ 进行空间平均而得到的 Grad-CAM 权重。与 CAM$_c$ 中的做法一样，将这些权重与对应的激活图相乘并对通道求和，可以突出显示模型预测类别 c 时最依赖的空间区域。

我们利用这一机制比较发现的电路与完整模型的行为。按照视觉模型中的常见做法，我们将 Grad-CAM 应用于 GTSRB 网络的最后一个卷积层。为便于可视化，我们计算、归一化并上采样所得的 Grad-CAM 图。图 9 展示了一个具有代表性的 GTSRB 样本对应的这些图，该样本描绘了一个环岛标志。

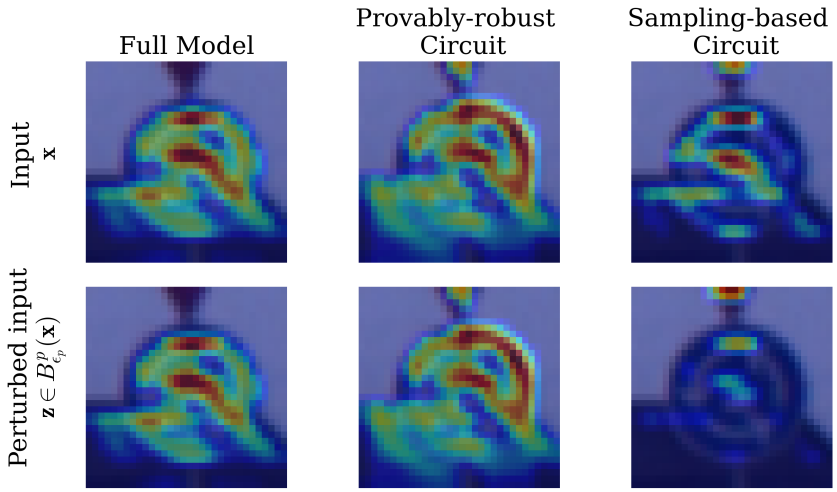
虽然基于采样的电路比可证明鲁棒电路更大（包含 35 个卷积通道，而后者包含 26 个），但在最后一个卷积层中，后者与完整模型更为接近。如图 9b 所示，完整模型在该样本上的热力图与可证明鲁棒电路的热力图吻合良好，而基于采样的电路所呈现的激活模式则不那么吻合，对标志内部区域的关注也有所减弱。

此外，尽管扰动半径非常小（使干净图像与受扰动图像在视觉上几乎无法区分；见图 9a），图 9b 显示，基于采样的电路会在两个输入之间转移注意力，而可证明鲁棒电路则基本没有变化。这或许表明，可证明鲁棒电路在受到扰动时能更好地保持关注焦点。

这些观察虽然并非我们评估的核心，但为理解所发现电路的运作方式提供了额外的定性视角。



(a) GTSRB 数据集中的原始输入和经过扰动的输入。



(b) 最后一个卷积层的 Grad-CAM 计算

图 9：GTSRB 输入及其对抗样本在三个模型（完整模型、可证明鲁棒电路和基于采样的电路）的最后一个卷积层上的 Grad-CAM 热力图对比。

E.2.7 组件级行为的补充定性解释

定性分析的另一个可能方向是为内部组件和子图赋予语义解释。这可能包括考察它们的行为，并推断它们所参与的因果路径。

在图 4 所示的示例中，我们选取了 CIFAR-10 中一张鸟类样本及其对抗扰动（也显示于图 10），并比较从 ResNet 模型中提取的两种电路变体：基于采样的电路和可证明鲁棒电路。如图所示，可证明鲁棒电路保留了若干滤波器级组件，使其能够在存在扰动的情况下满足鲁棒性标准。

为了分析额外的组件，我们重点关注第一卷积层，如图4所示。虽然后续层的交互也可能提供有价值的见解，但为求简洁明了，我们将注意力限定在应用于受扰动鸟类图像的第一层滤波器上。如图所示，该层包含两个电路共有的三个滤波器，以及一个仅存在于可证明鲁棒电路中的额外滤波器。

接下来，我们考察干净图像和对抗图像，以及它们对应的归一化差异热力图；相关内容见主论文中的图10。尽管 $\epsilon = 0.015$ 处的扰动在视觉上几乎与干净输入无法区分，但将两者的差异可视化后可以发现，扰动的大部分集中在图像下部、鸟的轮廓下方。

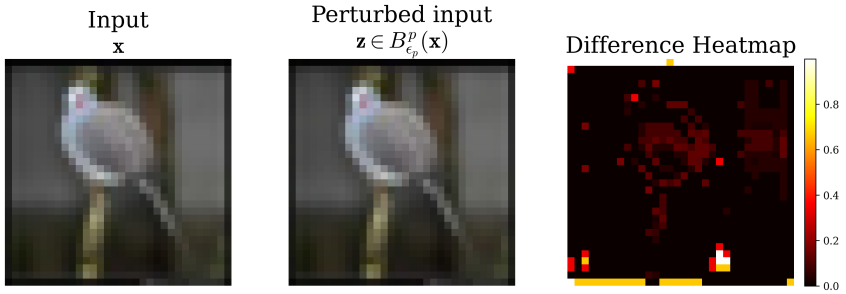
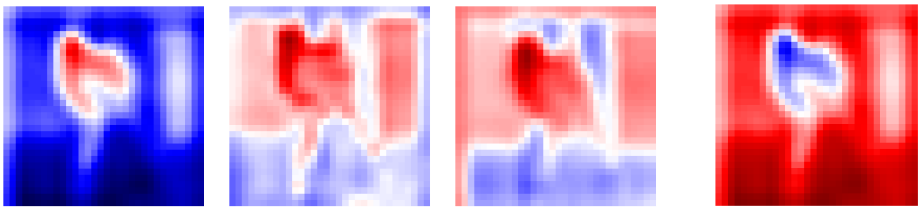


图10：干净输入$\mathbf{x}$、其扰动版本$\mathbf{z} \in B^p_{\epsilon_p}(\mathbf{x})$（含有 $\epsilon_p = 0.015$），以及相应的差异热图。



(a) 两个电路共享的滤波器

(b) 仅保留的滤波器在可证明鲁棒的电路中

图11：（a）可证明鲁棒电路和基于采样的电路中共有的第一层滤波器的激活图。（b）仅包含在可证明鲁棒电路中的额外第一层滤波器的激活图。蓝色表示负激活；红色表示正激活。白色表示中性值或接近零的值。

为了进一步了解情况，我们检查了这些滤波器在受扰动输入上的激活图（图11）。我们考察了两个电路共有的三个滤波器，以及可证明鲁棒电路独有的额外滤波器。对于每个滤波器，我们提取其带符号的激活值（因为该层不应用 ReLU），对其进行归一化，并放大以便观察。所得的符号归一化图显示，负激活为蓝色，正激活为红色。

观察图11a中三个共享滤波器的表现可以发现，尽管它们会对鸟的轮廓做出反应并捕捉其部分特征，但它们都在噪声集中的下方区域赋予负值。滤波器0在该区域输出强烈的负值，而滤波器2和7产生介于负值与中性值之间的数值，仅在该区域传递微弱的信号。

相比之下，额外的滤波器仅包含在可证明鲁棒电路中（图11b），并且恰好在受噪声影响的下方区域产生强烈的正激活。它是这一层中唯一如此响应的滤波器。这表明，将其与其他滤波器一同纳入，可能有助于丰富并稳定输入受扰区域的信号，从而可能提升电路在这种扰动下经认证的鲁棒性。这种示意性观察表明，保留的组件与电路的鲁棒性之间可能存在关联。

F 补丁鲁棒性认证

在本实验中，我们评估发现的电路的鲁棒性：根据第 3.2 节的定义，非电路组件使用从连续输入范围内抽取的可行激活进行修补，而非使用固定常数。具体而言，对于任意电路 C，我们测试在输入$\mathbf{z} \in \mathcal{B}^{\infty}_{\epsilon_p}(\mathbf{x})$所诱导的激活值范围内扰动非电路组件，是否会导致某个度量 $\|\cdot\|_p$ 违反容差 δ。如果不存在此类违规，我们就称 C 具有修补鲁棒2。

我们在迭代发现框架（算法 1）中比较三种修补方案：

1. **零修补：** 将所有非电路组件置为零值。2. **均值修补：** 用非电路组件的经验均值替换它们，这些均值根据随机选取的 100 个训练样本估算得出。3. **可证明鲁棒修补：** 验证由连续输入域（定义 2）所诱导的全部可行激活范围内的鲁棒性。

F.1 方法

F.1.1 用于验证的修补孪生网络

作为验证修补鲁棒性的接口，我们采用一个包含两个分支的修补孪生网络：**(i)** 完整网络修补分支，用于捕获待修补的非电路激活值；以及 **(ii)** 限定于 C 的电路分支，其中每个非电路组件都由修补分支中对应组件的激活值替换。此替换通过专用连线实现，将所需激活值从第一个分支复制到第二个分支。

我们考察由$\mathbf{x}$邻域内输入诱导的激活值： $\mathcal{Z} = \mathcal{B}^{\infty}_{\epsilon_p}(\mathbf{x})$。如第 3.2 节所述，孪生网络接收拼接后的输入（$\mathbf{x}$，$\mathbf{z}$）（对于 MNIST，沿特征轴拼接；对于卷积神经网络，沿通道轴拼接）。其中，$\mathbf{x}$ 被路由至电路分支，而$\mathbf{z}$ 被路由至修补分支。验证域仅施加于$\mathbf{z}$，而$\mathbf{x}$ 保持固定。此设置模拟电路在$\mathbf{x}$上运行，同时其非电路激活值由$\mathbf{z} \in \mathcal{Z}$所诱导的激活值替换。我们指出，为保证验证器一侧的数值稳定性，电路分支的输入被包含在一个可忽略的10^{-5} L^{∞} 球内。

对于输出准则，我们使用容差 δ，根据 logit 差值度量，将得到的电路逻辑值与完整模型的逻辑值$f_G(\mathbf{x})$进行验证（该值在孪生网络构造之外独立预先计算）。这保证了修补鲁棒性性质（定义 2）成立。

输入邻域。 与 E.1.2 节中一样，我们使用ℓ_∞ 范数定义邻域 $\mathcal{Z} = \mathcal{B}^{\infty}_{\epsilon_p}(\mathbf{x})$。

F.2 鲁棒性评估与参数变化

发现后（使用零值、均值或可证明鲁棒修补），我们使用修补孪生网络，在相同的 $\mathcal{B}^{\infty}_{\epsilon_p}(\mathbf{x})$ 上验证所得电路，并报告电路大小、运行时间和修补鲁棒性。由于文献中的典型 ϵ_p 值针对输入 扰动，我们在修补域中采用更大的值，以反映内部激活值更广泛的变化，同时避免分布外区域。下文报告不同(ϵ_p, δ)取值下的结果，以评估其对鲁棒性和大小的影响。

F.2.1 验证与实验设置（修补鲁棒性）

我们使用与输入鲁棒性研究 E.1.2 中相同的硬件配置。与输入实验一样，我们为 MNIST、GTSRB 和 TaxiNet 设置了分支定界超时，时长为 **45** 秒。对于 CIFAR-10，可证明鲁棒方法中的迭代发现查询限制为 **45**秒，而已发现的电路鲁棒性评估最多允许 **120** 秒。在这些限制时间内未完成的查询将报告为未知。

与输入鲁棒性实验 E 一样，为了公平起见，我们从报告的鲁棒性统计数据中排除零值修补或均值修补的鲁棒性检查超时的情况。在主要结果中，MNIST、TaxiNet、GTSRB 和 CIFAR-10 的超时率分别为 12%、2%、6.2% 和 31%；而在 ϵ_p (F.2.2) 的变化实验中，TaxiNet 和 MNIST 的超时率平均为 0.5% 和 8.8%。在 δ 变化实验中（F.2.3），TaxiNet 和 MNIST 的平均超时率分别为 1.5% 和 14%。

F.2.2 修补邻域大小的变化 ϵ_p

我们固定容差水平 δ，并改变修补邻域大小 ϵ_p。对于 CIFAR-10，我们使用 δ=0.1；对于 MNIST，使用 δ=0.5；对于 TaxiNet，使用 δ=0.92；对于 GTSRB，使用 δ=2.0。表 12 在主要结果的基础上，补充了 MNIST 和 TaxiNet 基准测试的其他 ϵ_p 变化实验。

数据集	ϵ_p	零修补			均值修补			可证明修补鲁棒的修补		
		时间（秒）	大小（$\lvert C\rvert$）	鲁棒性（%）	时间（秒）	大小（$\lvert C\rvert$）	鲁棒性（%）	时间（秒）	大小（$\lvert C\rvert$）	鲁棒性（%）
MNIST (δ=0.5)										
	0.005	0.054 ± 0.182	19.87 ± 1.55	**87.0** ±3.4	0.013 ± 0.001	19.19 ± 1.87	**93.0** ±2.6	671.08 ±36.86	11.32 ± 2.56	**100.0** ±0.0
	0.009	0.013 ±0.000	19.92 ±1.51	**65.9** ±5.0	0.013 ±0.001	19.13 ± 1.88	**64.8** ±5.0	583.59 ±44.99	16.75 ± 2.26	**100.0** ±0.0
	0.010	0.060 ±0.322	19.96 ±1.50	**58.0** ±5.3	0.016 ±0.003	19.16 ±1.84	**55.7** ±5.3	714.87 ±207.08	17.03 ±2.30	**100.0** ±0.0
	0.050	0.015 ±0.005	19.76 ±1.50	**1.2** ±1.2	0.015 ±0.003	19.09 ±1.81	**1.2** ±1.2	598.41 ±156.96	23.20 ±1.17	**100.0** ±0.0
TaxiNet (δ=0.92)										
	0.005	0.061 ±0.183	5.78 ±0.79	**93.0** ±2.6	0.022 ±0.061	5.38 ±0.65	**100.0** ±0.0	220.67 ±57.23	4.60 ±0.74	**100.0** ±0.0
	0.008	0.026 ±0.097	5.78 ±0.79	**62.0** ±4.9	0.008 ±0.001	5.38 ±0.65	**77.0** ±4.2	168.99 ±44.83	5.26 ±0.54	**100.0** ±0.0
	0.010	0.024 ±0.059	5.78 ±0.78	**57.1** ±5.0	0.025 ±0.068	5.39 ±0.65	**63.3** ±4.9	175.73 ±52.71	5.41 ±0.59	**100.0** ±0.0
	0.030	0.009 ±0.002	5.78 ±0.79	**27.0** ±4.4	0.008 ±0.001	5.38 ±0.65	**40.0** ±4.9	95.31 ±15.54	6.04 ±0.20	**100.0** ±0.0
	0.050	0.024 ±0.058	5.77 ±0.78	**0.0** ±0.0	0.012 ±0.032	5.37 ±0.65	**0.0** ±0.0	89.37 ±17.58	7.07 ±0.26	**100.0** ±0.0
CIFAR-10（δ=0.1）										
	0.030	0.109 ± 0.321	65.07 ± 3.00	**46.4** ± 6.0	0.046 ± 0.003	64.07 ± 3.60	**33.3** ± 5.7	5408.51 ± 1091.05	65.55 ± 1.64	**100.0** ± 0.0
GTSRB (δ=2.0)										
	0.005	0.284 ± 0.951	32.65 ± 4.24	**38.0** ± 4.4	0.041 ± 0.009	33.40 ± 4.16	**40.5** ± 4.5	2907.17 ± 721.67	34.34 ± 4.07	**100.0** ± 0.0

表 12：固定容差 δ下修补邻域大小 ϵ_p 的变化。报告值为均值和标准差（格式为 {mean ± std}）。对于鲁棒性（二元结果），我们报告鲁棒性的均值及其标准误（SE），并以**粗体**显示鲁棒性数值。灰色高亮的行对应主论文中选取的结果。

F.2.3 容差的变化 容差 δ

我们固定 ϵ_p 并考察不同的容差值。与主论文结果一样，我们对 MNIST 和 TaxiNet 使用 ϵ_p=0.01，并改变容差 δ。结果报告于表 13

数据集	δ	零修补			均值修补			可证明鲁棒修补		
		时间（秒）	大小（$\lvert C\rvert$）	鲁棒性（%）	时间（秒）	大小（$\lvert C\rvert$）	鲁棒性（%）	时间（秒）	大小（$\lvert C\rvert$）	鲁棒性（%）
MNIST (ϵ_p=0.01)										
	0.25	0.142 ± 0.476	21.36 ± 1.87	**48.6** ± 5.9	0.013 ± 0.000	21.22 ± 1.74	**52.8** ± 5.9	523.67 ± 41.31	19.42 ± 2.01	**100.0** ± 0.0
	0.50	0.060 ± 0.322	19.96 ± 1.50	**58.0** ± 5.3	0.016 ± 0.003	19.16 ± 1.84	**55.7** ± 5.3	714.87 ± 207.08	17.03 ± 2.30	**100.0** ± 0.0
	1.00	0.013 ± 0.000	16.88 ± 1.85	**61.2** ± 4.9	0.013 ± 0.001	16.11 ± 1.93	**66.3** ± 4.8	690.42 ± 44.38	11.37 ± 2.64	**100.0** ± 0.0
TaxiNet (ϵ_p=0.01)										
	0.50	0.011 ± 0.003	6.86 ± 0.78	**74.0** ± 4.4	0.010 ± 0.001	5.83 ± 0.40	**80.0** ± 4.0	157.67 ± 33.68	6.00 ± 0.00	**100.0** ± 0.0
	0.80	0.080 ± 0.210	6.04 ± 0.77	**57.1** ± 5.0	0.010 ± 0.001	5.49 ± 0.54	**61.2** ± 4.9	204.38 ± 40.04	5.68 ± 0.55	**100.0** ± 0.0
	0.92	0.024 ±0.059	5.78 ±0.78	**57.1** ±5.0	0.025 ±0.068	5.39 ±0.65	**63.3** ±4.9	175.73 ±52.71	5.41 ±0.59	**100.0** ±0.0
	1.20	0.011 ± 0.001	5.44 ± 0.96	**46.9** ± 5.0	0.010 ± 0.003	5.20 ± 0.79	**58.2** ± 5.0	246.14 ± 50.82	5.15 ± 0.51	**100.0** ± 0.0

表 13：固定修补邻域大小 $\epsilon_p = 0.01$时，容差水平 δ 的变化。报告值为均值及标准差（格式为均值 ± 标准差）。对于鲁棒性（二元结果），我们报告鲁棒性的均值及其标准误（SE），并以**粗体**显示鲁棒性均值。灰色高亮的行对应于主论文中选取的结果。

G 探索电路最小性保证

在本实验中，我们考察必须同时满足输入鲁棒性（定义 1）和修补鲁棒性（定义 2）的电路，相关内容见第 3.2 节。具体而言，我们将 Φ 定义为：当非电路组件使用从修补邻域 $\mathcal{Z}' = \mathcal{B}^{\infty}_{\epsilon_{\mathrm{p}}}(\mathbf{x})$ 中抽取的值进行修补时，电路在输入邻域 $\mathcal{Z} = \mathcal{B}^{\infty}_{\epsilon_{\mathrm{in}}}(\mathbf{x})$ 内仍保持鲁棒。

因此，这里涉及两个定义域：一个用于输入，另一个用于获取修补时使用的激活值。此处，ϵ_{in} 和 ϵ_{p} 表示相应 L^{∞}球的半径。在我们的设置中，我们使用 $\epsilon_{\mathrm{in}} = 0.01$、$\epsilon_{\mathrm{p}} = 0.012$和 $\delta = 2.0$。

G.1 使用三重暹罗网络验证同时满足输入鲁棒性和修补鲁棒性

我们使用一个三重孪生网络来认证同时满足输入鲁棒性和修补鲁棒性。该网络包含三个分支，每个分支都在其指定的定义域上进行评估：(i) 一个完整网络修补分支，它

处理输入$\mathbf{z}' \in \mathcal{B}^{\infty}_{\epsilon_{\mathrm{p}}}(\mathbf{x})$，以捕获激活值并将其用作修补值；**(ii)** 完整网络，它处理输入$\mathbf{z} \in \mathcal{B}^{\infty}_{\epsilon_{\mathrm{in}}}(\mathbf{x})$，以提供参考逻辑值 $f_G(\mathbf{z})$；以及 **(iii)** 电路分支，同样在$\mathbf{z} \in \mathcal{B}^{\infty}_{\epsilon_{\mathrm{in}}}(\mathbf{x})$上进行评估，但仅限于 C，其中非电路组件被屏蔽，并由修补分支移植而来的激活值取代。

这种特殊的连线方式能够直接传递非电路激活值，使验证器可以证明，在由 $\mathcal{B}^{\infty}_{\epsilon_{\mathrm{p}}}(\mathbf{x})$ 诱导的修补值下，电路逻辑值 $f_C(\mathbf{z})$ 在输入邻域 $\mathcal{B}^{\infty}_{\epsilon_{\mathrm{in}}}(\mathbf{x})$ 内始终忠实于 $f_G(\mathbf{z})$，从而确立输入与修补鲁棒性性质。

We 在此设置下评估三种发现策略，所有策略均采用组合鲁棒性 s
pre上文定义的 dicate Φ：

1. **迭代发现：** 算法 1。2. **准最小搜索：** 算法 4。3. **阻塞集 MHS 方法：** 算法 2，利用电路阻塞集对偶性来近似基数最小电路。

G.2 阻塞集 MHS 方法：分析与实验详细信息

实验设置。 我们在 MNIST 网络（34 个隐藏神经元）上进行实验。由于对比子集按大小递增的顺序枚举，验证查询的数量会组合式增长。即使将子集大小限制为 $t \in \{1, 2, 3\}$，每个批次仍需要数千次验证调用。为使计算可行，我们评估单元素集合（每批 $k = 1$ 个），并为每次查询设置 30 秒的超时。在一个罕见情况下，由于在特定环境和度量的选择下，Φ 平凡成立，消除了所有组件，因此该过程产生了一个空电路 （大小为 0）。此情况未纳入报告的结果。

并行性。 与消除步骤存在顺序依赖的迭代发现不同，对比子集的验证彼此独立。这种独立性使得完全并行化成为可能：我们将验证查询分配给 14 个工作进程，运行时间几乎随工作进程数量线性扩展。

性质。 在单调 Φ 条件下，MHS 方法会得出以下结果之一：**(i)** 任意基数最小电路大小的下界；或者 **(ii)** 当击中集本身满足 Φ 时，得出经认证的基数最小电路。尽管 MHS 的计算成本高于迭代发现，但它提供了严格更强的保证：如果击中集有效，则结果可被证明是基数最小的；否则，其大小给出了一个紧确的下界，揭示迭代发现是否达到了基数最小性，并量化其中的差距。

H 披露：大语言模型的使用

大语言模型仅被用作写作助手，用于纠正语法、修正拼写错误并提高表述清晰度。它未参与提出研究思路、设计研究、分析数据或解释结果；上述所有工作均由作者独立完成。