

识别纽结

(香港城市大学 张智毓、朱永健、代亮 编译自 Davide Michieletto. *Physics World*, 2024, (9): 29)

任何一个经验丰富的水手都知道，能否选择正确的纽结关乎他们的生死。无论是起帆还是定锚，纽结是绳子牢不牢固的决定因素。在微观世界里，对于许多维持我们生命活动的分子来说，情况也是如此。

蛋白质是所有生物体所必需的一类重要的微观分子链。这些由氨基酸组成的长链可以在三维空间中形成各种复杂的形状，使其他各种分子能够结合。长期以来，人们一直认为，虽然蛋白质可能有缠结，但在正常条件下不会形成纽结，因为“打结”可能会阻碍蛋白质折叠。然而，在20世纪70年代，研究人员陆续发现了各种纽结蛋白质(图1)。

事实证明，尽管蛋白质(或DNA)都是首尾不闭合的曲线，但依然可以形成纽结并且影响生物功能。这点也体现在蛋白质数据库(Protein Data Bank, PDB)中，其中约1%的蛋白质具有纽结结构。

然而，当试图将对纽结的认识应用于生物学和软物质中的问题时，我们会遇到一个困扰科学家一个多世纪的数学问题。

一段“纠缠”的历史

现代纽结理论的起源可以追溯到150多年前进行的一项著名实验，该实验不是用常见绳子，而是用一种意想不到的东西：烟雾。1867年，泰特(Peter Guthrie Tait)邀请他的朋友，也是同为物理学家的William Thomson(著名的开尔文勋爵)从格拉斯哥前往爱丁堡，来见证他的“烟雾环”实验。令开尔文勋爵惊讶的是，这些环相当地稳定：它们在房间里穿梭，甚至像橡胶一样碰到彼此时还会弹开。烟雾环是一种“涡旋环”，其中气溶胶和颗粒在小同心圆中旋转，这种运动使环可以保持稳定。

在那个时代，人们普遍认为空间中充满了“以太”这种物质，且引力和电磁辐射通过其传播。开尔文勋爵认为，原子可能是由这种以太中的稳定涡旋(就像是实验中的烟雾环)组成的。他甚至提出，打结的以太涡旋环可以用来解释不同的化学元素。

泰特对开尔文勋爵的假说很感兴趣。于是，在之后25年的时间里，

泰特编制了一份包含251个纽结的示意图，这些纽结的交叉数最多可达10个。虽然后来被证明原子涡旋理论是错误的，但纽结理论至今仍然是数学的一个重要分支。

识别纽结

对于和泰特一样的纽结理论学者来说，纽结分类是一项十分艰辛的工作。每当提出一个新的结时，他们都必须使用绘图和几何直觉来检查它的唯一性。泰特就曾写道：“尽管我已经将许多截然不同但等价的纽结分组在一起，但也不能绝对保证不同组的纽结是完全不同的”。他说的也不无道理，因为1974年Kenneth Perko证明，泰特示意图中的两个纽结实际上是同一个——它们现在被称为“Perko pair”。

如果你需要更多的说服力，可以尝试我的学生Djordje Mihajlovic开发的一个名为“识别纽结(Spot a Knot)”的小游戏。游戏的目标是从图片中找出等价的纽结。尽管研究了多年的纽结，我也经常出错。纽结具有唯一的拓扑结构，即无法在不断开的情况下平滑变形为任何其他已知的结。就像Perko pair和“识别纽结”小游戏所示，证明两个纽结拓扑上不同可不是一件“想当然”的事。

在纽结理论中，描述纽结拓扑的量被称为“不变量”。纽结理论家的梦想是找到这样一个类似原子质子数的“完备不变量”。这种完备不变量给每种不同的纽结赋予唯一的值，这个值如前文所述不会随着纽

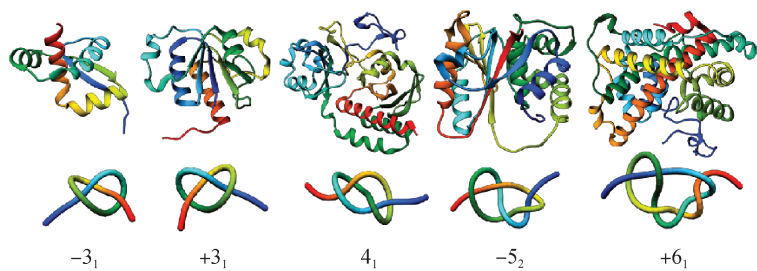


图1 纽结蛋白质(上)和对应纽结类型(下)。图中最左边的两个纽结即“三叶结”，正负号代表纽结的手性

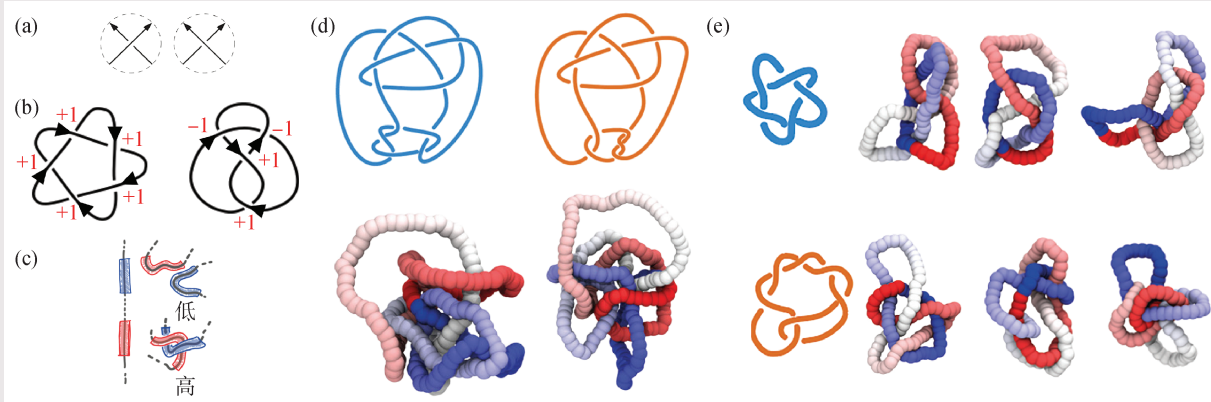


图2 纽结的绕数以及在研究中的应用 (a)左边代表上交叉,右边为下交叉;(b)五叶形和8字结形纽结:左边绕数为+5,右边为0;(c)局部绕数示意图。当局部线段“绕”过另一条线段,则局部绕数变高;(d,e)模拟实验中的纽结。其中,(d)图左右分别是康威和KT纽结,可以看到它们的结构十分类似,只是在最下方有些“突变”

结的平滑变形而改变。

“调制”这样一种拓扑不变量的一个“配方”可能如下。从二维平面上看,纽结是由交叉的线段组成的。这些交叉的线段有上下的关系(比如图2(a)左边的交叉是往右的箭头在上,往左的箭头在下)。随机选一个交叉点出发,沿着路径走,每经过一个交叉点就记为一步(第 n 点是第 n 步)。如果 n 是偶数,并且你当前所走的线段是从上面经过交叉点,那我们得到一个标记: $-n$ 。因为每个交叉点都会走过两次(一次从上面一次从下面),这样每个交叉点就会有俩个标记。把偶数标记拿出来就得到了Dowker—Thistlethwaite (DT)编码。DT编码首次于1983年提出。

虽然Dowker—Thistlethwaite编码可以分类许多简单的纽结,但它仍然不是一个完备的不变量。历史上第一种纽结不变量是由James W Alexander在1928年提出的,称为亚历山大多项式(Alexander Polynomial)。从那时起,纽结理论学家们提出了各种版本的不变量,但每种都至少存在一个无法进行唯一分类的情况。

沿着纽结走走

亚历山大多项式也是一种所谓的“代数不变量”。这种“代数不变量”是从纽结的二维投影中构建的。然而,如果像我这样的软物质物理学家想对蛋白质和DNA这样的分子纽结进行分类,这些分子是三维的,将这些分子简化为二维投影可能会抹去对其功能至关重要的空间特征。

表征分子的一个极具潜力的替代方法是所谓的“几何不变量”:通过遍历三维纽结并沿路线计算某些几何属性(例如曲率)来得到。

我非常喜欢的一种不变量是“绕数”(writhe)。这是由泰特引入的纽结几何特征,通过“上方”的交叉数减去“下方”交叉数来得到(图2(b))。

然而,绕数也可以作为几何量来计算。想象一下沿着三维纽结(例如一个纽结蛋白)行走,并在每一步可以通过你看到的交叉来记录绕数的估计值。当你遍历完这个纽结,这些数字的平均值将反映绕数的真实值。可惜的是,绕数也不是一个完备的不变量。事实上,就像代数不变量一样,截至今日没有任何几

何不变量被证明可以唯一地分类所有纽结。

2021年,谷歌DeepMind团队的AlphaFold人工智能程序解决了一个困扰科学家数十年的问题——如何根据氨基酸序列来预测蛋白质的结构。蛋白质的功能取决于其3D结构,因此AlphaFold被认为是药物发现和研究疾病的有力工具。

于是,我们问自己的问题是:人工智能是否也能用于研究纽结不变量?

扭转和绕数

使用人工智能对纽结进行分类已被先前的研究人员探索过,分别是香港城市大学的Olafs Vandans及其同事(2020年)和帕多瓦大学的Anna Braghetto及其团队(2023年)。在两项研究中,他们将不同的纽结视为一串珠子,并训练AI通过提供珠子的笛卡儿坐标作为特征来识别纽结。尽管取得了很高的准确性,但他们的结果仅限于5个最简单的纽结。我们希望将其扩展到更复杂的拓扑结构,同时简化AI架构并使用更小的训练数据集。

为此,我们从自然界中汲取了灵感。在生物体内,DNA中的纽结

是通过称为拓扑异构酶的特殊酶来解开的。我们猜想拓扑异构酶可以感知某些局部几何属性来定位纽结中最紧密的部分。于是试图通过使用各种量，包括密度和曲率，来证明这个猜想。最终，研究的结果都指向了起点——泰特和他的几何绕数。

除了全局计算之外，局部绕数可以表示片段 x 与附近片段 y 之间的缠绕程度(图 2(c))。我们发现，局部绕数是定位长环形分子中纽结片段的一种非常有效的方法。基于这一发现，我们决定将局部绕数提供给 AI。(译者注：原文作者研究中使用的是非柔性链，纽结相对平滑，绕数计算可以很好地表征纽结。但柔性链由于热效应十分杂乱，绕数不是一个很好的特征。)

当我们用笛卡儿坐标对 AI 模型进行训练时，模型正确率只有 80%，这与 Vandans 和 Bragetto 的发现相似。虽然这可能已经比我们大多数人在“识别纽结”游戏中取得的分数要好了，但仍然远非完美。

然而，当 AI 在局部绕数上进行训练时，结果令人震惊：它以超过 99.9% 的准确率正确地分类了纽结。

艰巨的挑战

尽管我对此感到惊讶，但识别 5 个最简单的纽结相对容易，可能一个优秀的“识别纽结”玩家也可以做到。

于是，我们决定给 AI 一个更具挑战性的任务。虽然这次它只需要分类 3 个纽结，但却是我们精心挑选的：康威纽结(Conway knot)、Kinoshita—Terasaka knot (KT 纽结) 和平凡纽结(unknot)——所有纽结中最简单的纽结(译者注：与圆环拓扑同胚)。前两个纽结有 11 个交叉且很多拓扑不变量相同(图 2(d))，而后者没有交叉。更重要的是，三者的某些纽结不变量是相同的。

我们发现，对于使用笛卡儿坐标进行训练的 AI 来说，康威纽结和 KT 纽结是不可区分的，但使用局部绕数训练能够以 99.9% 的正确率识别它们。

最终的测试是将 AI 应用于识别更多的纽结类型。我们模拟了 250 种纽结，交叉数最多可达 10 个(图 2(e))。当 AI 使用笛卡儿坐标进行训练时，它的正确率只有 20% 左右。相比之下，用局部绕数训练的 AI 可以在几秒钟内以 95% 的准确率对所有 250 个纽结进行分类，比任何其他算法或单个拓扑不变量都要好。

最后的“绕回”

我们的 AI 模型在不知道任何关于纽结或纽结理论的情况下，自己学会了做一件长期以来一直困扰人类直觉的事情，而我们仍在努力打开“黑匣子”并理解它究竟发现了什么。

我们发现，为了区分 5 个最简单的纽结，AI 会选择纽结上每对点，并将这两个点的绕数相乘。有趣的是，这个量等价于一个现有的不变量，称为“二阶瓦西里耶夫不变量(Vassiliev invariant of order two)”。

瓦西里耶夫不变量是通过将局部绕数矩阵成对、三对、四对，直到 n 对进行相乘来计算的。值得一提的是，二阶瓦西里耶夫不变量也是康威多项式二次项的系数。有人提出(尽管尚未证明)，可以作为用积分计算的完备瓦西里耶夫不变量集是长期以来一直在寻找的完备不变量。

因此，我们非常兴奋地发现，当 AI 面对更复杂的纽结时，它会通过计算更高阶的瓦西里耶夫不变量来适应。例如，为了唯一地分类前 5 个纽结，AI 只需要二阶瓦西里耶夫不变量。但对于包含 250 个纽结的数据集，它可能计算高达三阶或

四阶的瓦西里耶夫不变量。

几何不变量和代数不变量是非常不同的数学计算使用方法。因此人工智能能够发现它们之间的联系是一件值得兴奋的事情，而且这可能让我们更接近于发现一个完备的不变量。

纽结真的非常重要

仅仅在三年的时间里，AlphaFold 就生成了数百万种蛋白质，其中大多数尚未被彻底研究。在 2023 年，华沙大学的 Joanna Sulkowska 组预测，AlphaFold 生成的人类蛋白质中约有 2% 是打结的，其中最复杂的纽结有六个交叉数。在此前一年，美因茨约翰内斯古腾堡大学的 Peter Virnau 在 AlphaFold2 数据集发现了一个包含有 7 个交叉的蛋白质纽结。这种蛋白质纽结从未在实验中观察到，因此可能还有更复杂的纽结存在。

纽结不仅在生物学中出现，纽结拓扑结构还被发现会影响冰和水凝胶的热力学和材料特性。这意味着在未来，人们可以利用拓扑来设计新型材料。因此我们希望这个发现能够为构建一个强大的方法来识别分子和材料中的纽结添砖加瓦。我想强调一遍，纽结真的非常重要。

2004 年，加拿大三位研究人员利用计算集群扩展了由泰特编制的纽结表到 19 个交叉数，并发现了超过 60 亿个独特的结构，而泰特花了 25 年时间来创建他的纽结表。他可能会感到震惊，一个世纪后，一台机器能够在短短几天内将他多年的工作扩展超过 5 个数量级。

纽结理论中最大的未解决问题仍然是寻找难以捉摸的完备不变量。现在有了 AI 的助力，接下来的发展可能同样会让我们像泰特一样感到惊讶。