21 全体大会

图灵研究所数学家 Terry Lyons：签名与数据流——数据科学中的新数学

整理：智源社区 罗丽

在 2020 北京智源大会第二次全体大会上，图灵研究所数学家 Terry Lyons 介绍了数据科学中新的研究领域“Signatures and Streamed Data”，即签名和流数据。他的演讲题目是：Mathematics of rough paths action recognition and health（粗糙路径的数学动作识别与健康）。

在演讲中 Terry Lyons 表示，“粗糙路径理论”是数学中一个新领域，而“路径签名”是一种新的数学工具，它为传统的机器学习的研究提供新思路，以更好地“理解”演化数据流。“粗糙路径理论”的目标是开发一个强大的数学框架，以系统地理解不断演化的多模数据中的模式，并对这些模式进行分类，用于构建 PyTorch、TensorFlow 等的工具，以便从多模流中学习。最后，他分享了数据流和签名在精神病学、阿尔兹海默病等智能治疗研究中的重要意义。

Terry Lyons：牛津大学 Wallis 讲习数学教授，英国牛津大学数学学院教授，Alan 图灵研究所研究员，国际知名数学家，曾任伦敦数学学会主席，英国皇家学会院士，国际数理统计学会会士，2000 年获得 Polya 奖。主要研究领域为：随机分析、粗糙路径，随机分析在金融大数据上的应用。

正文：

数学在我们的世界中具有重要影响，而数据科学为我们提供了了解数学和了解世界的窗口，它对人们看待问题的方式也会产生重要的影响。粗糙路径是用于处理复杂演化系统的数学语言。在我们所处的真实世界中，存在着很多演化现象，比如人类行为的演化。为什么演化意味着某种事物会随着时间而变化？实际上，理解如何能够自动地识别行为是非常重要的，而这种研究可能是人类行为的研究。

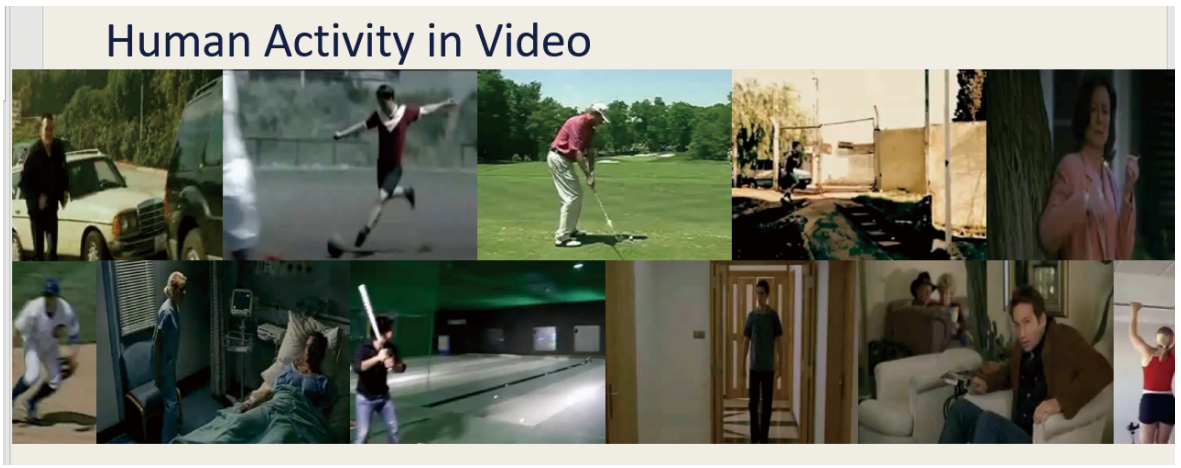


图 1：人类行为

在观看以上图像时，我们可能会想到两个完全不同的问题。一是识别图中事物，比如，识别图中所有的树、人等，这是一项非常成熟且非常重要的研究内容，而另一个让 Terry Lyons 真正感兴趣的研究，是了解图像中事物变化的方式，这是一个非常不同且具有挑战性的研究内容。即使是一个很小的群体，例如小于 9 的数字，当我们按顺序进行研究时，产生的可能性远远大于我们所能考虑到的对象的数量。

“粗糙路径理论”是一个新的数学研究领域，也是数学的一部分，它是我们理解高度复杂理论演化数据的框架，经过 20 年的发展，现在已经发展地相对成熟，且具有一定的影响力。粗糙路径理论在数据科学中的早期应用实际上是由 Facebook 的 Ben Graham 提出的，Ben Graham 在研究理解在线中文笔迹时，运用了粗糙路径理论中的一些技术来改进他的研究分析，并获得成功。

图 2：在线中文笔记 APP

之后，科学家们试图将其设计为图中所示的 APP(应用程序)，它能够翻译数十亿的字符，在经过一段时间的吸收发展之后，得到了良好的运用效果。

“路径签名”能够以通俗易懂的方式向我们解释：为什么它能告诉我们一些重要的事情；数据科学为什么能够改变我们处理某些问题能力。Terry Lyons 通过介绍几个可访问的应用程序，使我们对复杂顺序数据有所了解，他表示可以通过使用静态图像，成功地抽象出一个场景，并使用开放式或阿尔法式（例如地标位置）的地标，使我们能够理解一个人在做什么。

通过图像，所有人都可以理解火柴人的行为代表人类行为。

而火柴人是如何运动的呢？问题的关键是，它不再是一个人，而是一系列的位置。人的左手腕、右手腕、肘部、脚，以及所有的部位都被赋予了一个值，而实际上，这个值是它们在平面上的位置。因为是在图像中，所以图像中的参数也同样有效。所以，实际上我们将获得一系列标签，该标签具有附加的向量。在案例中，是在中等

高维空间中得到这些路径的，在这种情况下，可以得到每个标签的真实尺寸，因此，可能有 30 或 40 维路径，而目标是尝试从这 40 维路径中了解正在发生的事情。有人可能认为，可以通过深度学习理解，在某种程度上是可行的，但研究的挑战在于，需要通过相对较少的数据、样本，和具有一定解释性的方法来获取数据，结果证明数学可以做到这两方面。使用相对较少的样本进行工作的能力极大地增加了使用这种方法的用户范围。因为在在心理研究、社会科学和公共政策等的研究中，我们一般无法获得巨量的数据集。

那么，“流数据”和“路径签名”中真正的挑战是什么?

在日常生活中，流数据到处都有，可能是手指在手机屏幕上画出来的，可能是金融市场中的事件，可能是一本书、一段文字的顺序，也可能是医院的病历，到最后甚至是不断变化的人类情绪。人们通常的想法是，把流数据看作的一系列的值，但这不仅需要时间，而且可能会出现“这个 3 可能会被看作是那个 3”的情况。Terry Lyons 表示，流数据在某种意义上是一条曲线，很可能它一开始是以时间的形式显示的，但有一点很重要，就是这条曲线的本质和绘制这条曲线的速度无关。曲线的本质是它和形状有关。实际上，它是一种对称曲线，采样数据并不会改变它的本质，就像旋转某个人的脸，我们并没有改变他的脸，而是改变了代表它的数据。对称性对数据科学来说是一个非常不好的消息，对称意味着可以有很多种方法来表达数据所代表的内容，但就其本质而言，数据所表达的内容都是相同的。所以必须教机器去识别数据。如何来识别数据? 这就需要签名来迫使他们采用某种正常的方式来识别数据。比如，在所有地方查找某个词，机器可以一次又一次的辨认出这个词。Terry Lyons 表示，“实际上，这是可以克服的挑战，这是本次演讲的本质”。

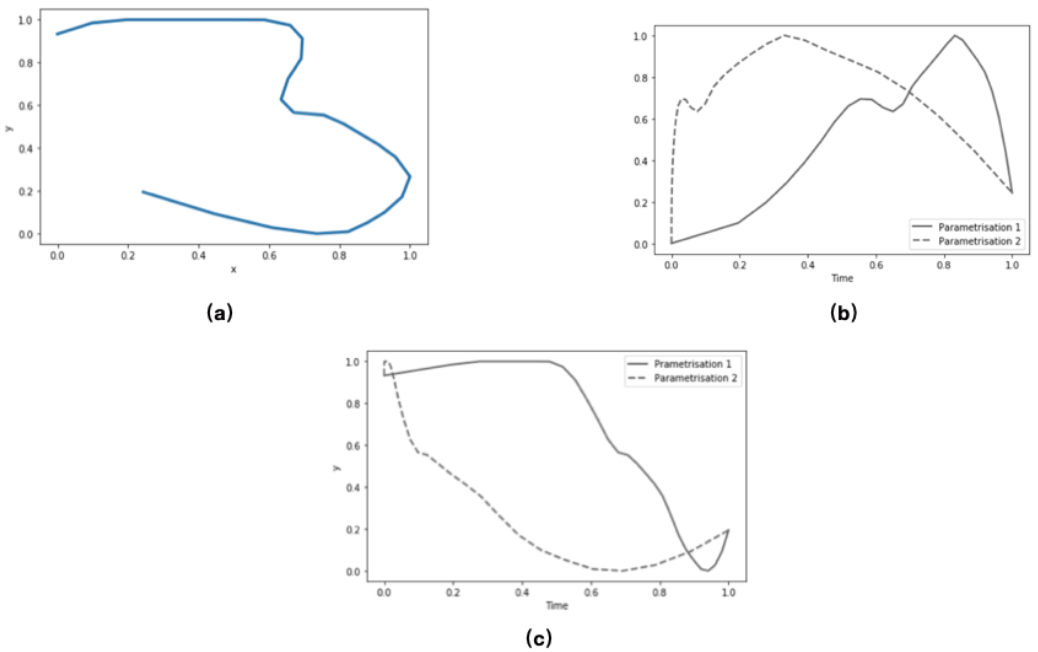


图 3：字母“3”的不同描述方式

图 (a) 是字母“3”从顶到底的绘制图，图 (b) 和图 (c) 是字母“3”的为 2 条数据记录路径，图 (b) 中的实线是演化的象征性符号相对于时间在不同速度下绘制的 x 坐标，图 (c) 中的实线是演化的象征性符号相对于时间

在不同速度下绘制的 y 坐标，虚线是经典的 ML 方法绘制的图像。虚线和实线分别对应不同版本的 3，即相同的 3 在不同速度下的参数表示。因此，真正的挑战是，尽管数据看起来完全不同，但有多种不同的方法可以来表达同一件事，这对于流数据来说是完全支持的，几乎所有的流数据都没有特别规范的权限，在同一个地方有各种不同的数据，有时候数据在不同的地方是有规律的，比如，人们会在不同的时间进入某个场所，这是流数据一个非常普遍的特性，也是一个坏消息，因为参数的空间是变化的，不是 3 维的，甚至不是 2 维的，例如旋转。因此，具有高维对称性的数据集将破坏数据的种类，也会使得数据难以理解。与所具有的观测数据的数量相比，高维对称性将导致数据的变量太多，出现的可能性也会很多。

只有一个平滑的数据也是处理数据必不可少的挑战。但如何从本质上解决这个问题？ Terry Lyons 表示应该提供一个更好的特征集，用更好方法来描述这些对象，这就是演化序列，演化序列是一个非常普遍、非常基础的数据描述方法。事实上，演化序列能够消除对采样速度的依赖，但这并不意味着我们可以忽略时间，时间可能也很重要，研究人员也可以将事件视为另一个维度，一组不断变化的变量，可以在图像中及时添加。添加时间后会出现两个不同的问题： 一是，在一定条件下数据的不变性，如每秒观看几帧视频并不会真正改变视频； 另一个是时间的变化可以改变一切。但它们是不同的概念，这里真正谈论的是重采样下的不变性。

那么，如何在数学上描述非参数化路径?

演讲中，Terry Lyons 表示，可以用多维路径对某些非线性系统的影响来描述该多维路径。用 dS=S ⊗ d γ的前几个术语描述γ。路径签名描述了未参数化的流$\gamma_{[u,v]}$，签名是非参数化路径的自上而下的描述，它是通过 S_u 对程式化非线性系统的影响描述路径段

$$dS_t = S_t \otimes d\gamma_t$$
$$S_u = 1$$

消除无穷维不变性，从而可以使用更小的学习集进行预测和分类，并给出与样本点无关的固定尺寸特征集（不会丢失数据，不会发生各种参数设置）。

签名也是描述非参数流的通用特征。流γ在 I=[s,t] 上的签名，定义为

$$\sum_{k=0}^{\infty} S_k \text{ where } S_0 = 1 \text{ and}$$
$$S_k(\gamma, I) := \int\!\!\cdots\!\!\int_{s<u_1<\cdots<u_k<t} d\gamma_{u_1} d\gamma_{u_2} \ldots d\gamma_{u_k}$$

这些“傅里叶式”系数准确地描述了未参数化的流。签名的精妙之处在于它们能够对正在发生的事情给出完整的描述，它们能将数据流转换为关于系统效果的完整描述。图为两个具有不同计算方式的手写“3”的例子，它们包含很多不同的签名，也包含了曲线的信息。

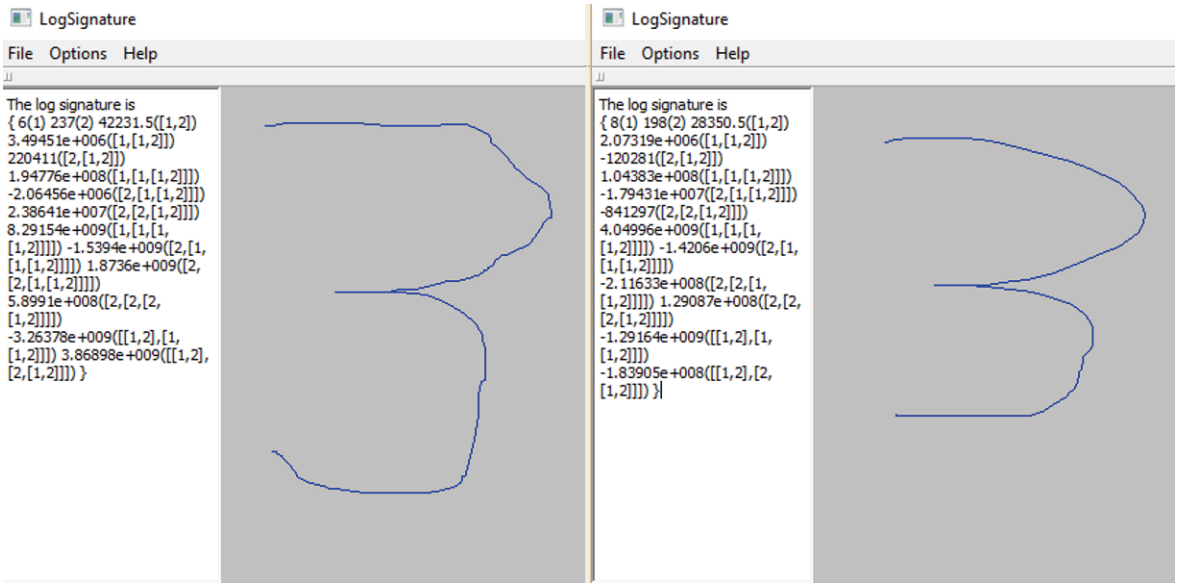


图 4：两种不同计算方式的手写“3”的例子

取两个“3”的不同参数，然后在整个过程使用傅立叶级数，因为傅立叶级数可以测量随时间间隔发生的事和发生的不同事情，而签名和日志签名不会改变，这是一种更有效的、完全不同的思考信息流的方式。实际上就像一个特征集本身不一定能出色地工作但它可以增强其他方法一样，所以，这种方式将依赖于深度学习、随机森林等其他工具。

EFFECT OF PATH SIGNATURE (PERCENT)

Path signatures	CR	AR	*Chinese*	*Symbol*	*Digit*	*Letter*
Sig0	90.18	89.24	91.64	80.21	81.18	47.94
Sig1	91.80	91.02	93.14	84.03	77.96	58.35
Sig2	92.25	91.57	93.50	83.80	84.63	54.24
Sig3	92.35	91.70	93.57	84.37	82.88	59.81
Sig0	92.59	91.86	93.91	83.22	86.16	78.61
Sig1	94.03	93.37	95.20	86.21	86.68	81.15
Sig2	94.37	93.82	95.44	86.62	90.15	81.28
Sig3	94.52	93.99	95.62	87.00	88.30	82.49

[1] The right columns list the CRs for different character types.
[2] Upper: Dataset-ICDAR; Lower: Dataset-CASIA.

图 5：路径签名在手写文字识别中的作用

粗略路径的流数据中流的种类很多，现在已经可以将视频缩小为地标和姿势，火柴棍所代表的男人和女人都是具有 30–75 维的数据流，但仍然需要获取有意义的数据，考虑时间的转移和速度变化。

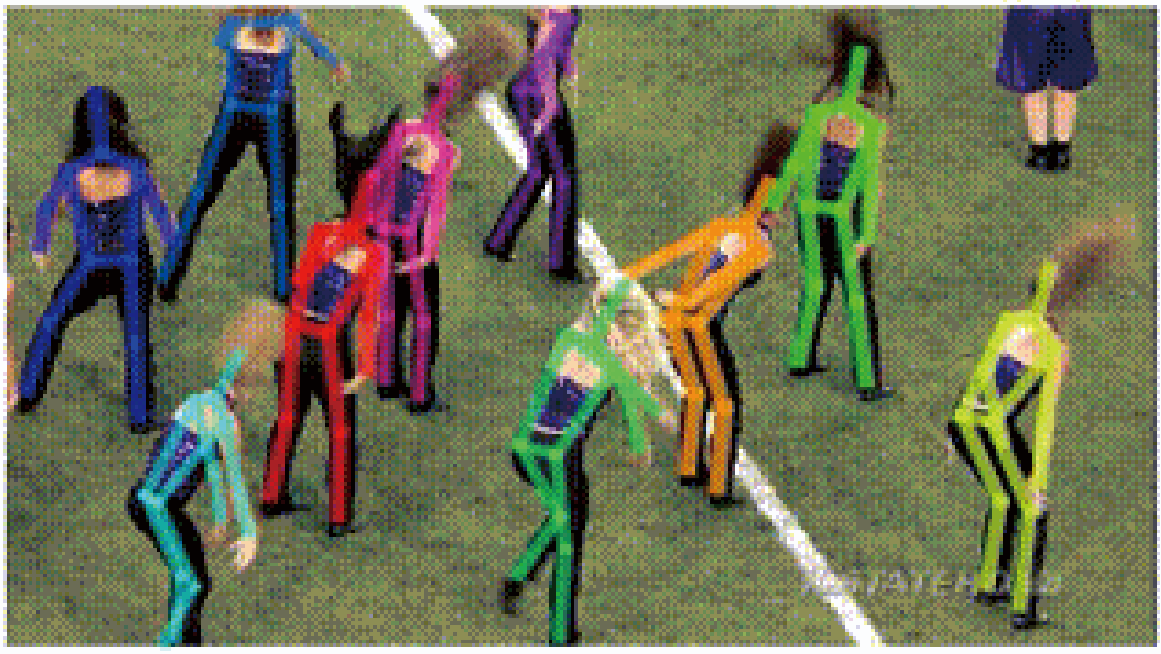


图 6：粗略路径流数据的应用

之前的研究，是尝试用小的数据集以及用户不需要参与的方式来了解人们在做什么，研究人员可以通过图像识别来研究图像并计算出身体的不同部位分别在什么地方。而真正让 Terry Lyons 感兴趣且具有广泛应用的是，机器如何理解人们在做什么。所以，实验的目标是，尝试从人的姿势中识别动作。

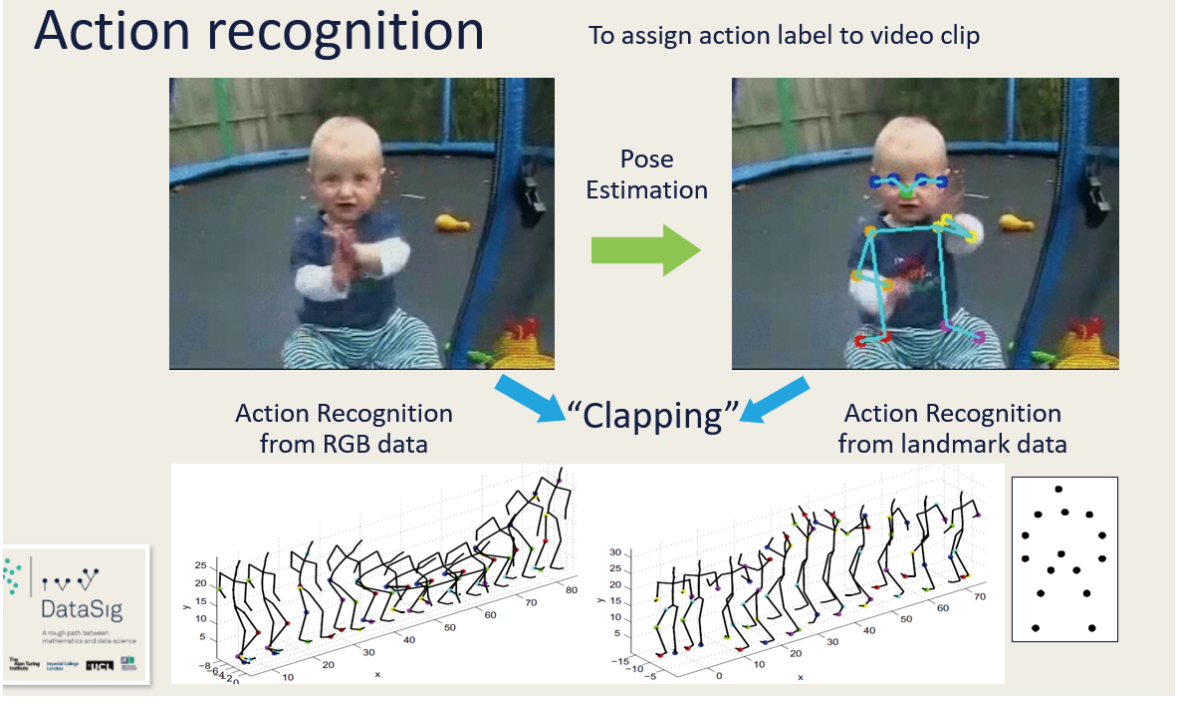


图 7：根据 RGB 数据和地标数据进行动作识别

图 8：“高尔夫运动”动作识别

图 8 的动作识别，实际是一个二维空间中的路径识别，路径的数量是地标数量的两倍，图中大概有 15 个地标，也就是一个 30 维的空间，机器可以识别这 15 个 2 维向量和一个 3 维空间所对应的值，这样就可以对抽象路标进行分析，同时能够很好的识别人类的动作。以下为“路径签名”与其他方法的精确度对比结果。

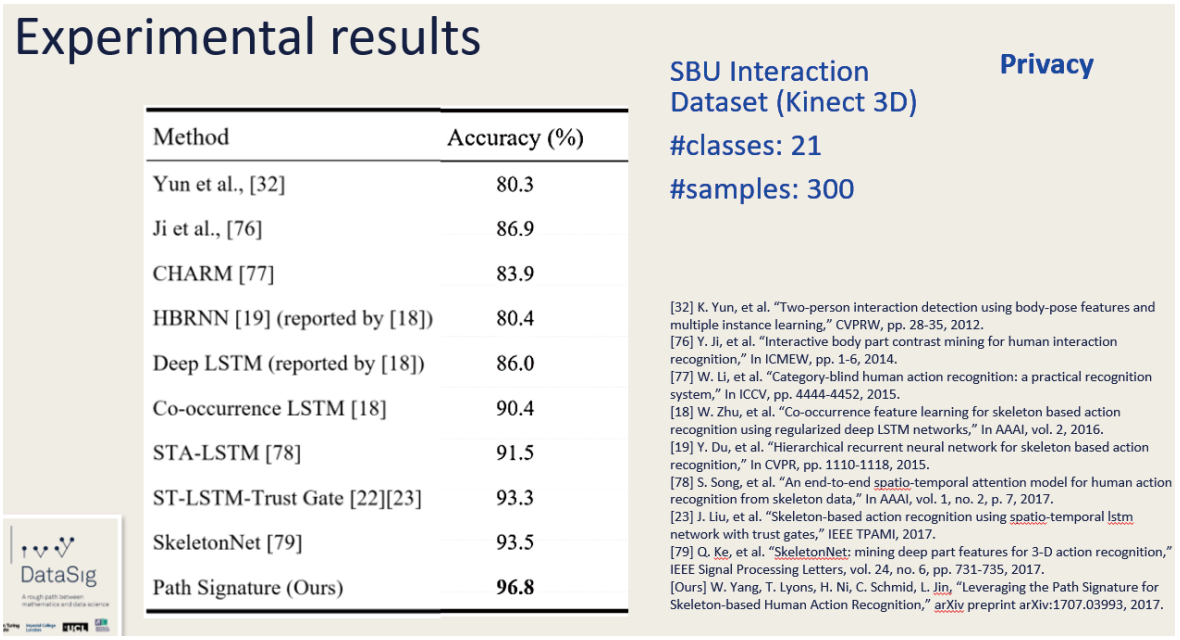


图 9：精确度结果对比实例

之后，Terry Lyons 介绍了关于社会数据的研究。在牛津精神病学临床实验中，研究人员研究了“Triaging BP BP & N on the basis of mood zoom”，即情绪缩放对 BP、BP 和 N 的分类，临床试验中，每天从三组人群（一年共 130 人）中获取具有不同确诊情绪的心情缩放数据，包括躁郁症、边缘型人格障碍或健康控制（数据有噪音

或丢失)。这些情绪数据被分为 20 个连续反应动作，在这些事件的训练中，使用具有二阶签名特征的随机森林分类器，实验时使用一次交叉验证能够在三组数据中心获得的很好分离结果。而二阶信息也很重要，对于给定样本量的情况下，签名对于控制尺寸至关重要。

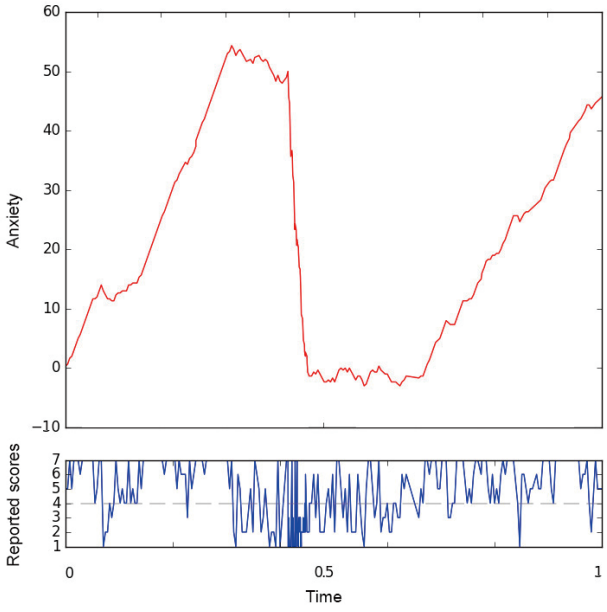


图 10：被诊断患有躁郁症的参与者的焦虑评分演变

使用低维特征捕获的高阶信息（抑郁之前的愤怒……）可以在频谱上进行分类。该项目是在 ATI 上展示并可以复制的三个项目之一。

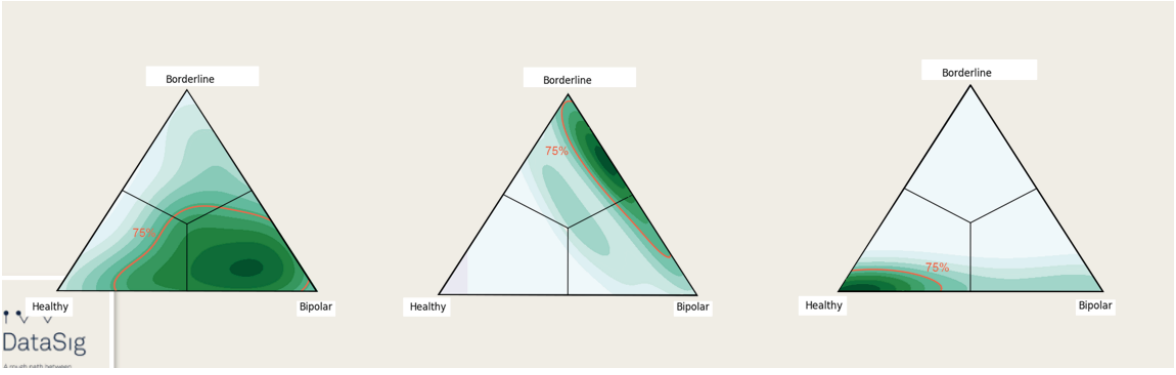


图 11：在 ATI 上展示的情绪演变

Terry Lyons 认为，在多次使用数据时，要了解数据出现的复杂状态演变方式，找到一个好特征值，以了解事情发生的顺序，并从数据中得到有用信息。他也表示在“中文手写”、“动作识别”、“阿尔兹海默病”、“复杂社交数据”等的研究中，利用数学具有重要意义，数据流还有很多理论方面的研究，比如，一些研究主题受到日志签名的影响，这些都是具有数学意义的。

MIT CSAIL 教授 Regina Barzilay：学习分子的表征

整理：智源社区　熊宇轩

在本届智源大会上，来自人工智能研究重镇 MIT CSAIL 的 Regina Barzilay 教授为听众带来了题为「学习化学结构」的主题演讲。Regina 教授高屋建瓴地从虚拟筛选和全新药物设计两个方面对机器学习在药物发现领域的应用进行了概述，并重点介绍了表征能力、泛化性能、不确定性估计、机制理解这四个关键问题。Regina 教授指出，目前该领域仍然存在巨大的研究空间，期待更多计算机科学家加入到这一方兴未艾的领域中来。

以下为智源社区整理的演讲全文:

本次演讲将介绍如何学习化学结构。我本人原本从事的是自然语言处理（NLP）领域的研究。大概五年前，我和 MIT 的另一名教授 Tommi Jaakkola 迁移到了对化学（分子）结构建模的新研究领域中。本次演讲向大家展示的内容涉及到 MIT 的一个大研究组的多项工作，我想特别强调的是，其中很多杰出的工作都来自于 Wengong Jin，他是我们研究组从事这方面研究的第一个学生。

一、基于人工智能的药物发现

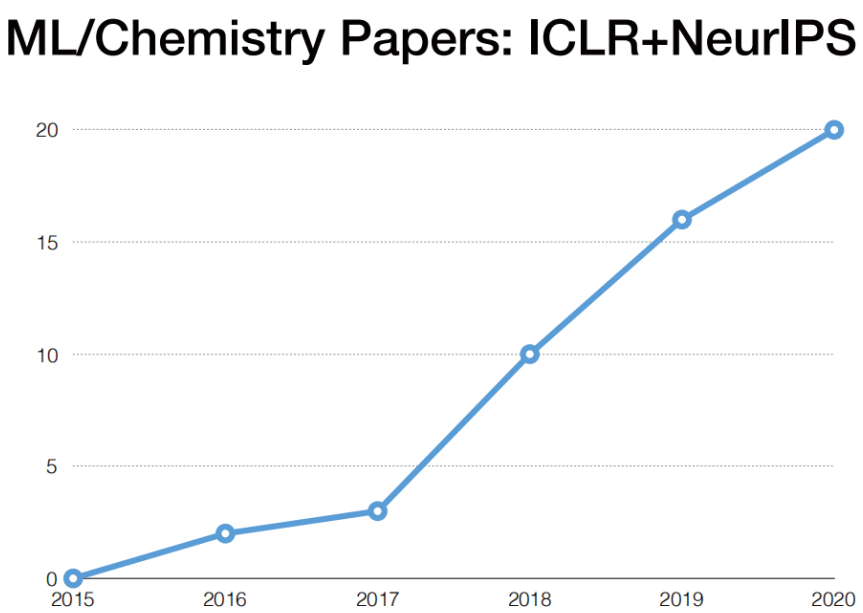


图 1：「机器学习 + 化学」领域在人工智能顶会上的论文发表趋势

接下来，我们将讨论如何将化学和机器学习相结合。很多人对机器学习在计算机视觉、自然语言处理等领域中的应用非常熟悉，但是化学是一个完全不同的领域，将机器学习应用到化学领域，仍然是一个新的研究课题。

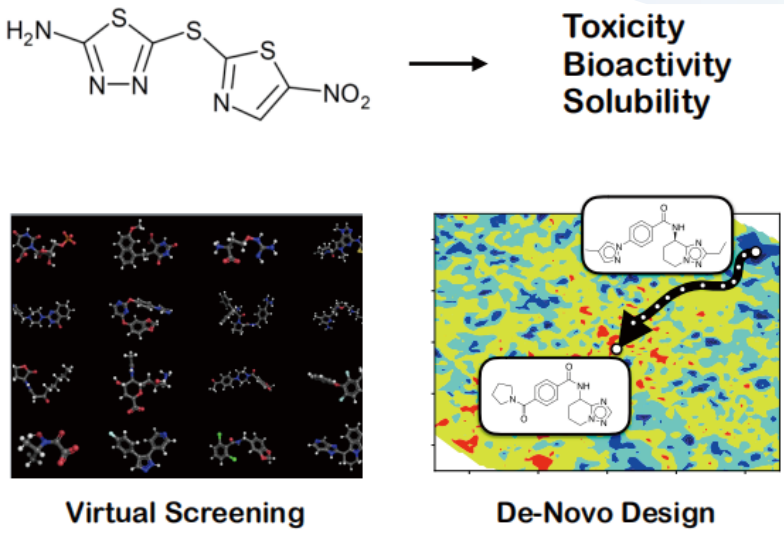


图 2：虚拟筛选与全新药物设计

分子建模对于制药、材料设计、化学领域的从业人员来说都是非常必要的。但直到今天，绝大多数的分子发现都是由实验驱动的，研究人员不断对数百万种分子进行不断实验，从而确定它们的性质，最终只能凭借偶然和直觉来发现某种特定性质的分子。即使这样，我们也只能探索其中极其小的一部分。

那么能否通过研发强大的机器学习模型，从中预测出符合要求的分子，而不是盲目地进行代价高昂的实验呢？答案是肯定的。通常，此类工作可以被分为两种路线：(1) 虚拟筛选。给定各种各样的分子，然后用模型预测出那些高概率会具备某种特性的分子；(2) 全新设计 (de-novo design)。这是一种「艺术」！尽管我们已有大量的分子可供选择，但我们要意识到，在这些分子之外，仍存在很多其它可能，所以我们在现有分子之外，也应当设计一些符合特定要求的全新的分子。本次演讲将围绕以上两个方向展开，大家也可以将虚拟筛选看做一种判别任务，而将全新药物设计看做一种生成任务。

图 3：MIT 基于虚拟筛选技术发现新型抗生素「Halicin」

首先介绍一下我们使用虚拟筛选技术，找到的一种名为「Halicin」的抗生素。相关成果已发表在国际顶尖生物学术期刊《Cell》上。这种分子的特别之处在于： 它可以治疗由多种对传统药物产生耐药性的病原体引起的病症，更值得注意的是，这种特殊的分子有着一种全新的生物作用机制。

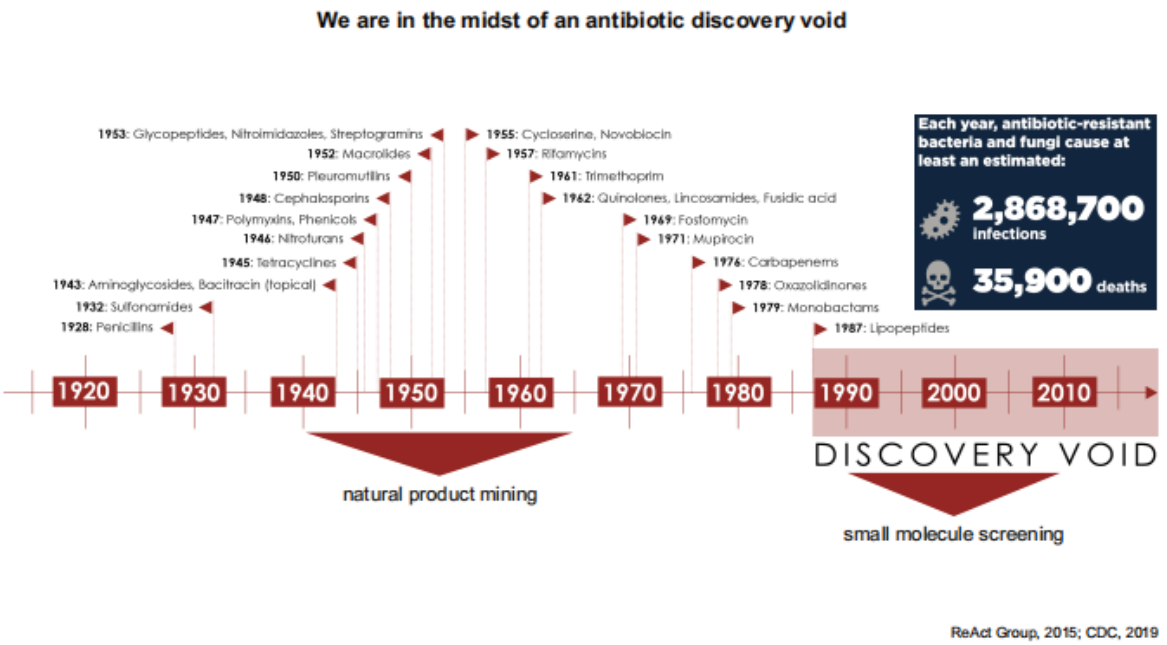


图 4：抗生素发现的研究历史

回顾抗生素发现的历史（如图 4 所示），尽管病原体有着越来越高的耐药性，很多人因为缺乏有效的抗生素而死亡，但在过去的 30 年间，人类发现的新抗生素越来越少。主要原因在于，研发成本越来越高。

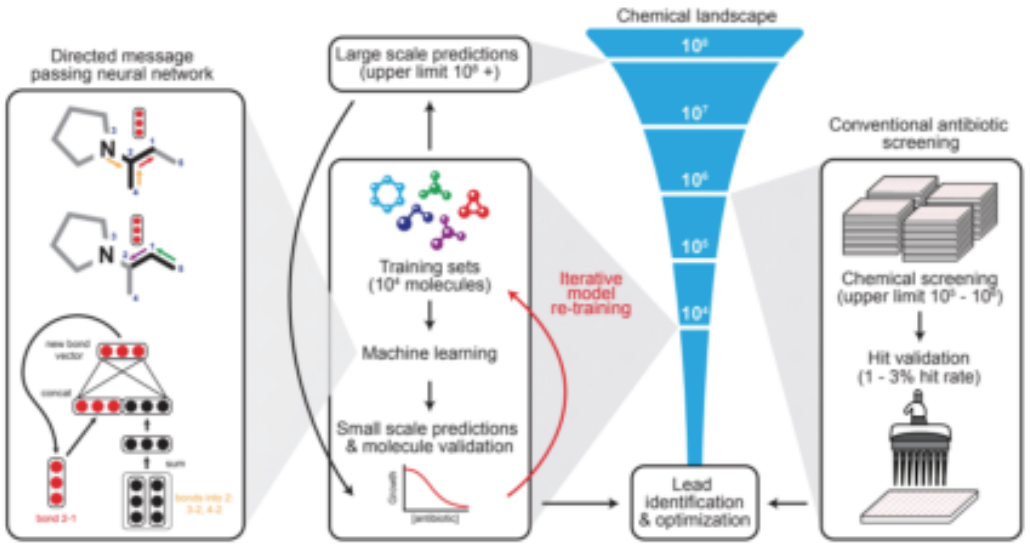


图 5：通过机器学习重新思考抗生素设计的动机

我们通过机器学习的方式先进行分子筛选，找到候选分子，然后再进一步测试它们的功效，这种方式则可以大大降低研发的成本。例如在 Halicin 的发现中，我们选取了某种模型，训练它，然后在实验室中对其进行活体动物测试。

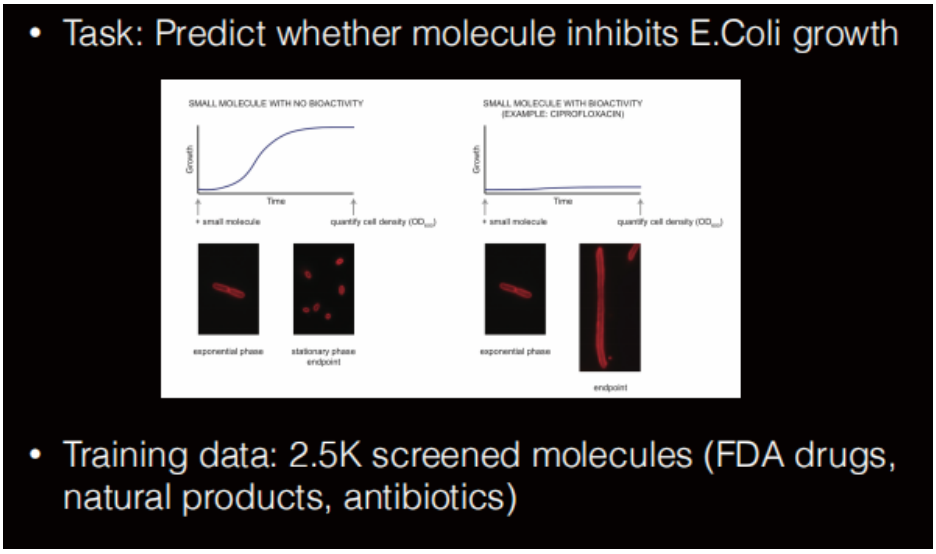


图 6：任务设定与数据集构建

具体而言，我们在该任务中生成训练数据的方式和我们在自然语言处理领域、计算机视觉领域中获得训练数据的方式是不同的。我们首先选取受到细菌感染的细胞，将某种分子与该细菌放到一起，查看它会不会抑制细菌的生长。因此，如果我们想要获得包含 2,500 个分子的训练数据集，他需要选取 2,500 个细胞，分别施加不同的分子，然后看看会得到怎样的实验结果。

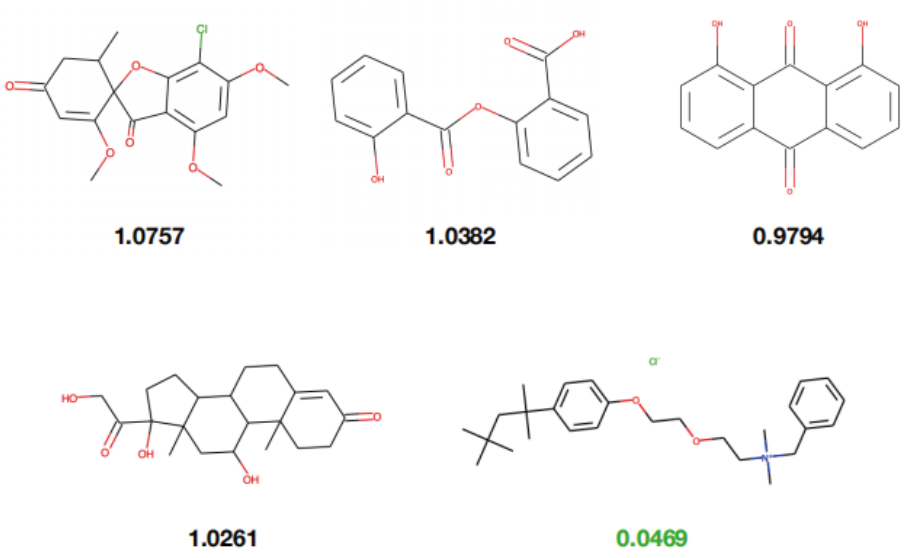


图 7：实验结果

最终，我们会得到一个表示分子结构的二维图，以及一个表示该分子对目标病原体杀灭作用大小的数字。

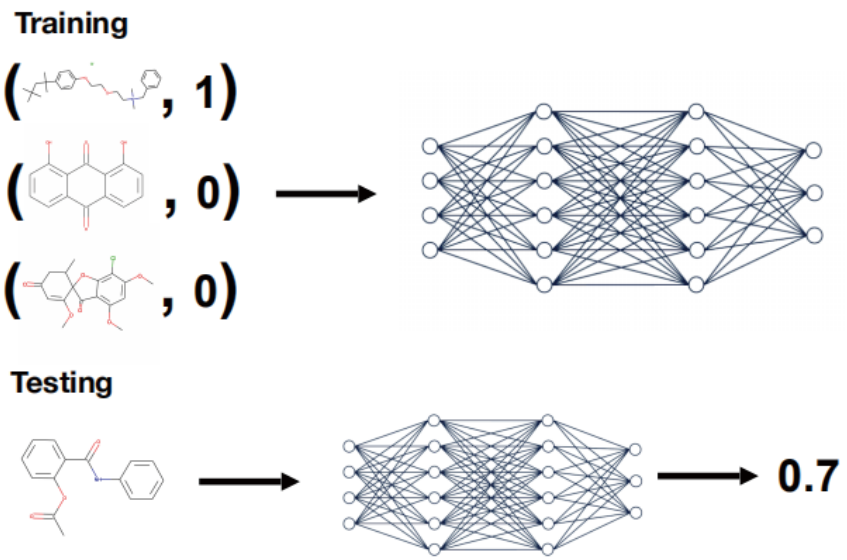


图 8：学习抑制病原体生长的情况

接下来，我们可以训练机器学习模型，在给定分子及其活动的情况下，预测分子对于病原体的抗菌活性。我们使用给定的数据集进行训练，该数据集包含分子的结构，以及它是否有抑菌作用的标签（1 代表有抑菌作用，0 代表没有抑菌作用）。当我们向训练好的模型输入一种新的分子时，模型可以预测出其抗菌活性有多大。

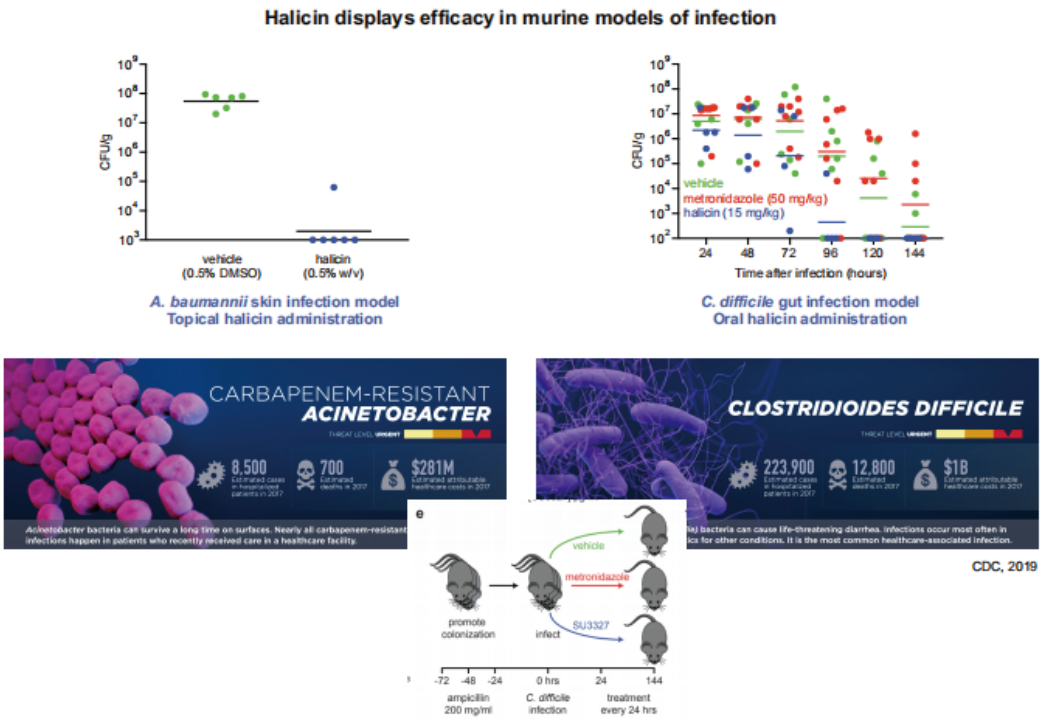


图 9：Hacilin 对两种之前无法治疗的病原体起作用

如上所述，我们所进行的是繁重的筛选工作，我们在数亿种分子上运行我们的模型，找出具有良好特性的分子，接着在实验室中的动物身上测试这些分子。实验结果证明，我们的模型找到的分子对两种目前尚无法治疗的病原体有很好的抑菌作用。

正如大家已经看到的，这些模型可以为药物发现任务带来很多的好处。那么，我们应该如何使用目前所掌握的方法让这一过程变得更好？也就是说，人工智能（AI）将如何改变这种「游戏」？

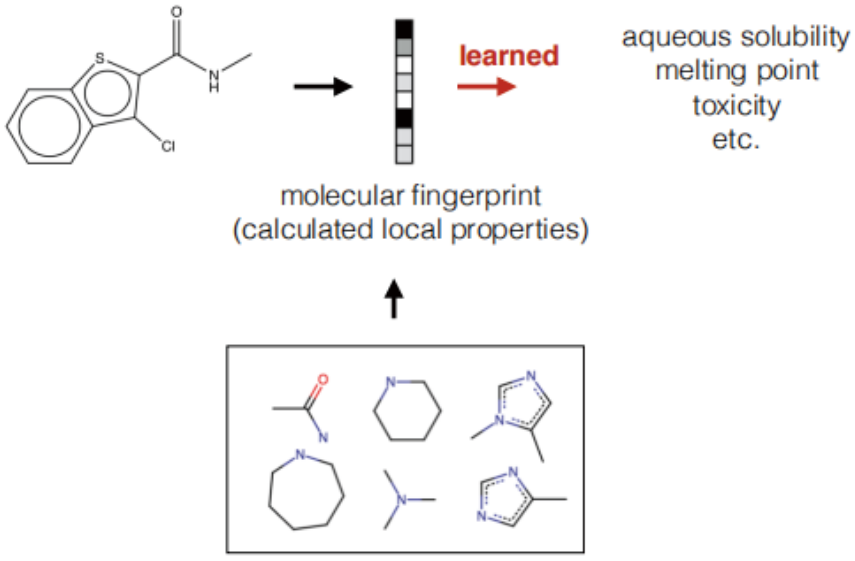


图 10：固定的人工分子表征

早在 1970 年代，人们就开始尝试在化学领域应用人工智能技术。给定分子的二维图，我们如何将其中的信息归纳到一个特征向量中，进而利用该特征向量执行我们的机器学习任务？「分子指纹」（molecular fingerprint）是实现上述目标的经典方法之一。给定一个分子的二维图，我们将这个图归纳为一个特征向量，向量的每一维坐标都代表一种特定的化学子结构（例如，环）。

那么，我们如何决定应该使用怎样的子结构？哪些子结构更加重要呢？这时，我们就需要使用一些化学的专业知识，确定分子中有哪些重要的子结构。很显然，这是十分困难的。因为，在我们考虑不同的特性（例如，活性、毒性）的时候，我们需要考虑不同的子结构的集合，而这方面有很多知识是尚不明确的。

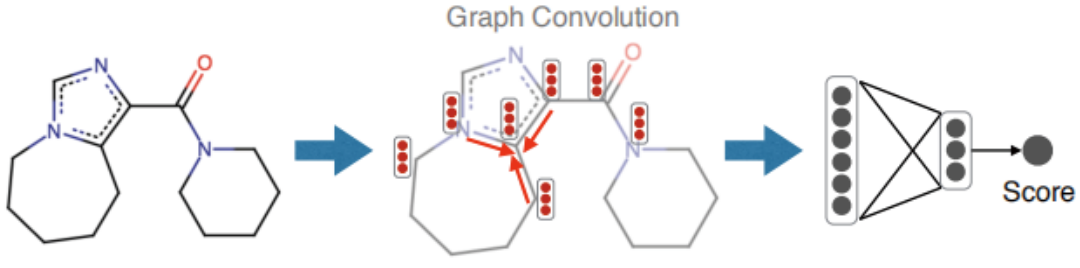


图 11：基于机器学习的分子表征

解决该问题的一种新的思路是：给定某种分子，使用神经网络学习将分子的二维图压缩到一个向量中，而不是使用固定的人工设计的表征，我们可以使用生成的向量预测分子的活性。从积极的一面看，我们可以根据期望预测的特性类型，将分子归纳到通过不同的向量中。但是从另一个消极的角度来看，我们损失了可解释性，我们无从知晓每一个坐标所对应的意义。但是我们并不关心这些坐标的意义，只需要使用该向量进行预测。在后面的演讲中，我们将通过实例向大家更详细地介绍构建这些表征的细节。在这里，我们将先为大家提供一个有关分子表征的高屋建瓴的概览。

我们的期望是，给定分子的二维图，我们将其抽象到一个高维空间中，这种高维空间需要拥有正确的几何性质(即拥有相似的溶解性等特性的分子在该空间中距离较近，而拥有不同特性的分子之间的距离则较远)。

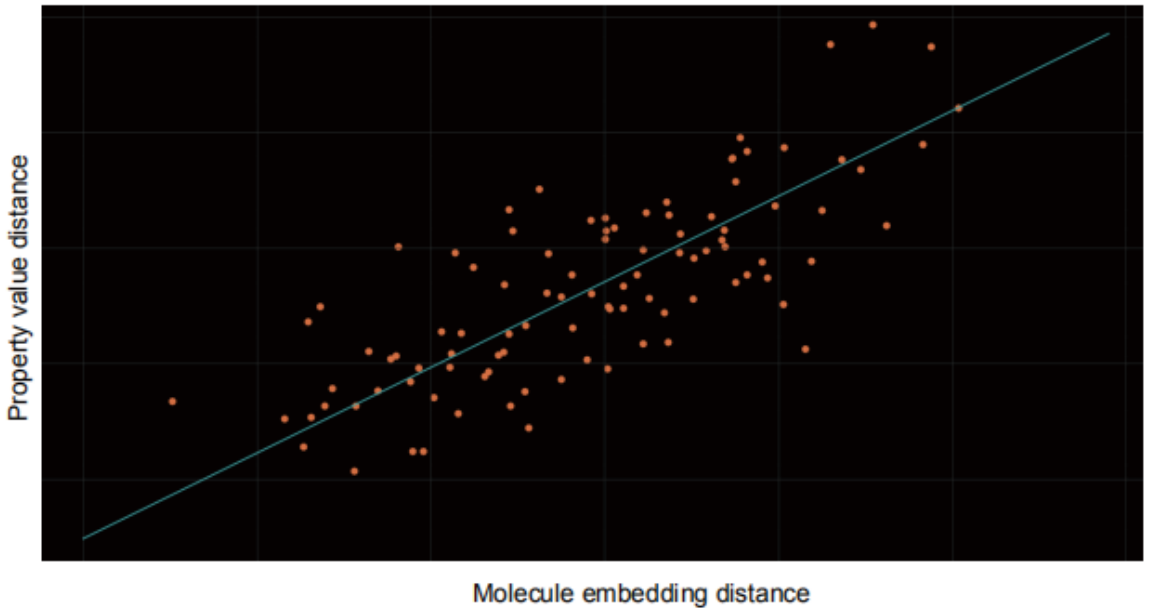


图 12：理想的平滑潜在嵌入空间

换而言之，我们假设分子在嵌入空间中的距离可以体现出其特性之间的差距。而关键在于，如何将分子抽象到这种嵌入空间中，使上述声明成立。

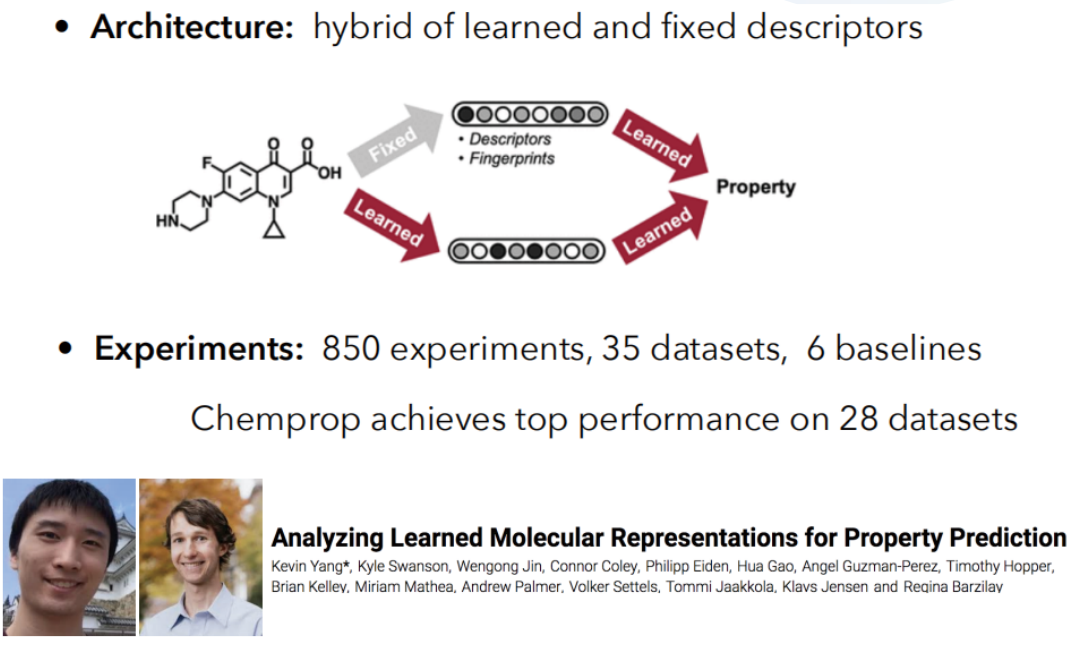


图 13：公开的 Chemprop 系统

我们构建了一个名为「Chemprop」的系统，它被制药行业以及许多其它的论文广泛使用，目前已经可以公开获取。去年，我们撰写了一篇名为「Analyzing Learned Molecular Representations for Property Prediction」的论文，针对化学研究社区面临的窘境，说明我们学习到的分子表征要优于人工设计的分子指纹。这是因为，每个公司都有自己的分子指纹，那些对使用这些指纹非常有经验的人可以做得更好。在这份工作中，Kevin Yang 和 Kyle Swanson 进行了 850 次实验，我们说明学习到的分子表征整体上表现得更好。我们也可以使用一种混合的架构，同时使用学习到的表征和分子指纹（尤其实在较小的数据集上）。实际上，在本次演讲之前，清华大学和腾讯公司在 Arxiv 上发表了它们最新的研究论文「Multi-View Graph Neural Networks for Molecular Property Prediction」，它们通过改变神经网络的架构取得了更好的性能。

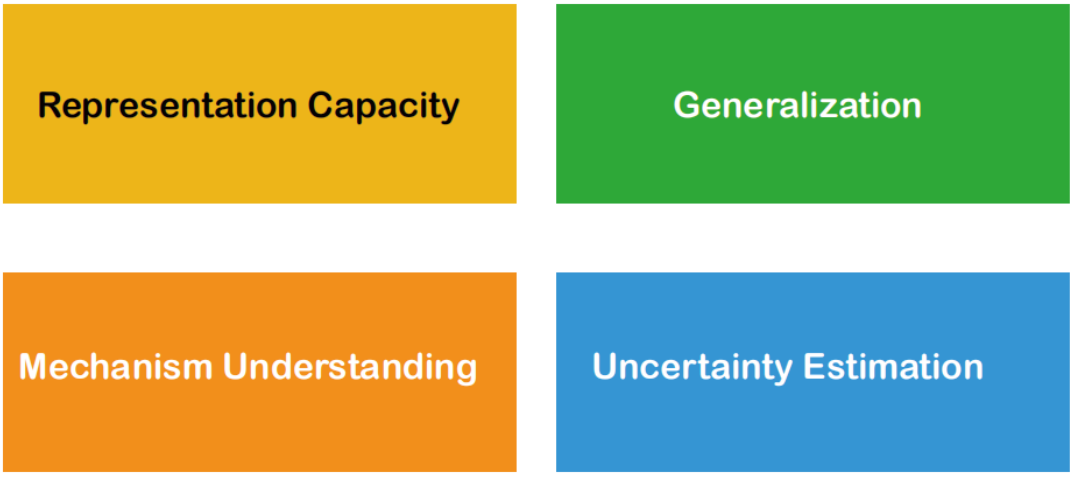


图 14：分子特性预测建模领域的开放性问题

尽管通过改变网络架构取得性能提升也是相当重要的，但是在本次演讲中，我将更多地介绍该领域中一些有待研究的问题，目前整个研究社区还很难提出有效的解决方案。首先，我们将讨论表征能力（Representation Capacity）。这些二维图实际上表征了分子的一些信息，如今很多研究人员使用图神经网络对其进行表征，那么这是否是正确的前进方向呢?

二、表征能力

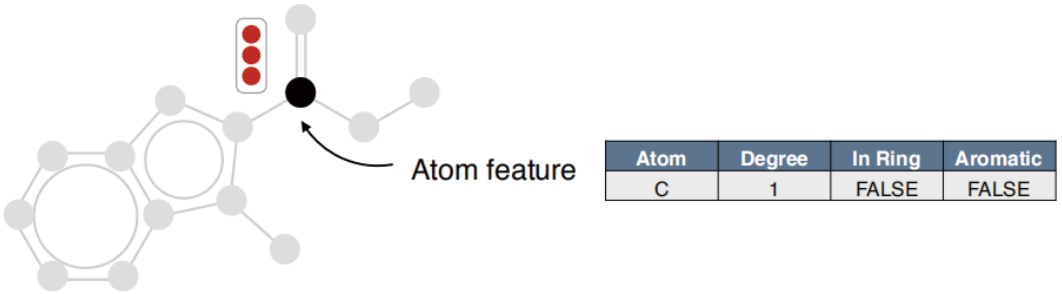


图 15：图卷积初始化

让我们看看研究人员可以如何使用图卷积技术，实际上所有此类模型都会以各种各样的方式使用图卷积的思想。首先，我们将分子视为一种原子的组合，并且将每个原子表征为一个向量。该向量中各个维度上的值是固定的，它们分别表示原子的类型、度、是否在环结构中，等等。以上特征是通过对原子进行简单的计算得到的。

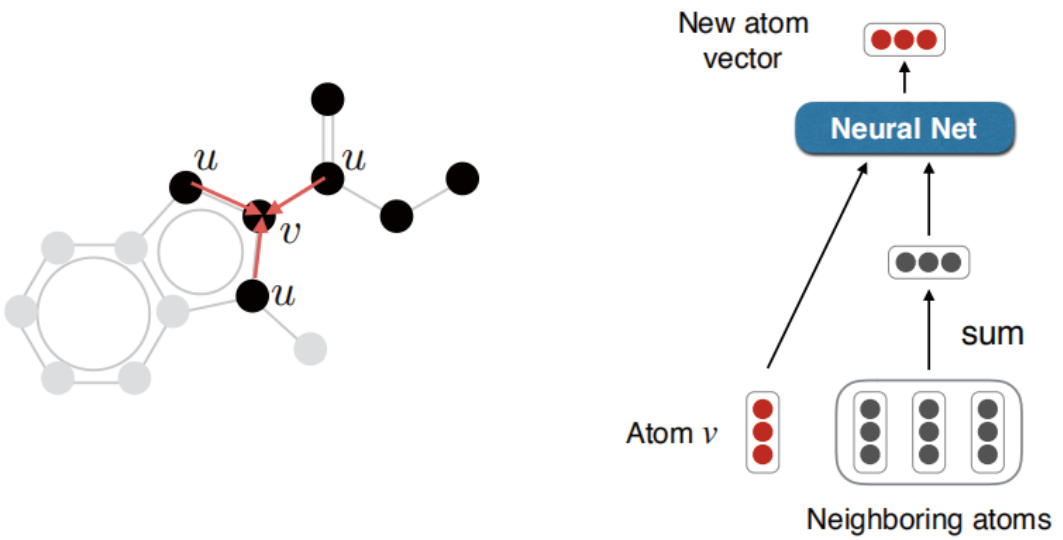
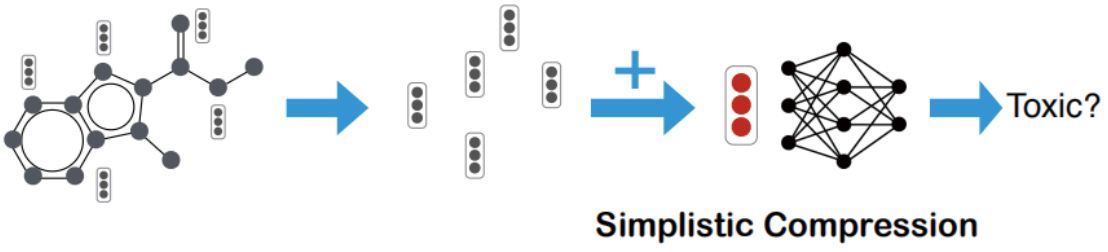


图 16：局部信息聚合

接下来，在图卷积网络中，我们往往会开始进行消息传递（ message passing）。给定某个原子及其邻居节点的向量，我们试图学习如何将它们的向量结合起来，从而优化我们最终的预测结果。在 1 跳邻域内执行这种消息传递机制后，我们不仅知道该原子本身的信息，也考虑了所有与其紧邻的节点的信息。如果我们继续执行这种

消息传递，由于所有的原子都会同时更新信息，最终 2 跳、3 跳甚至更多跳之内的信息都会传递给当前节点。在该过程结束时，所有的原子都会保留其周围局部的环境信息及其自身的特征。



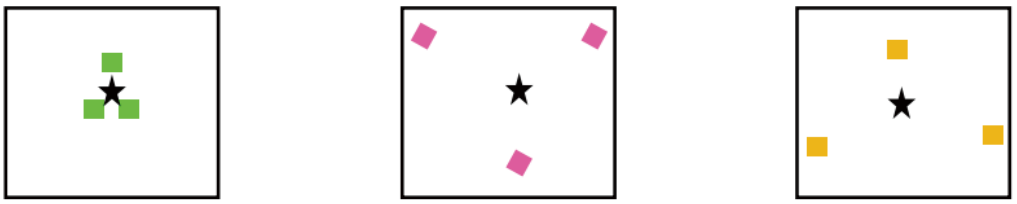
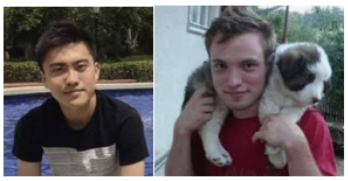


图 17：重新思考分子表征

在这里，有趣的事情发生了。每个原子都有其自身的特征向量，在经过图卷积后，我们如何利用这些原子的表征得到分子的表征？有趣的是，在本例中，我们直接将这些原子的表征相加，并将其作为分子的表征。这个步骤是必要的，因为我们必须以某种方式整合各个原子的信息。但问题是，这种整合方式是最佳的吗？如图 17 下方所示，三个差异很大的特征向量集合拥有相同的和。关键之处在于，当我们在最后执行这种过于简单的加和压缩过程时，我们会损失掉很多的信息。



“Optimal Transport Graph Neural Networks for Molecular Representations”

图 18：更好的信息聚合策略

解决该问题的一种方法是：使用更丰富的方式来整合各个节点的信息。在此，我将展示近期提交给 NeurIPS 的论文「Optimal Trasport Graph Neural Networks for Molecular Representation」(https://arxiv.org/pdf/2006.04804.pdf)，我们试图使用一种名为「Wasserstein 原型」的方式来整合各节点的信息。

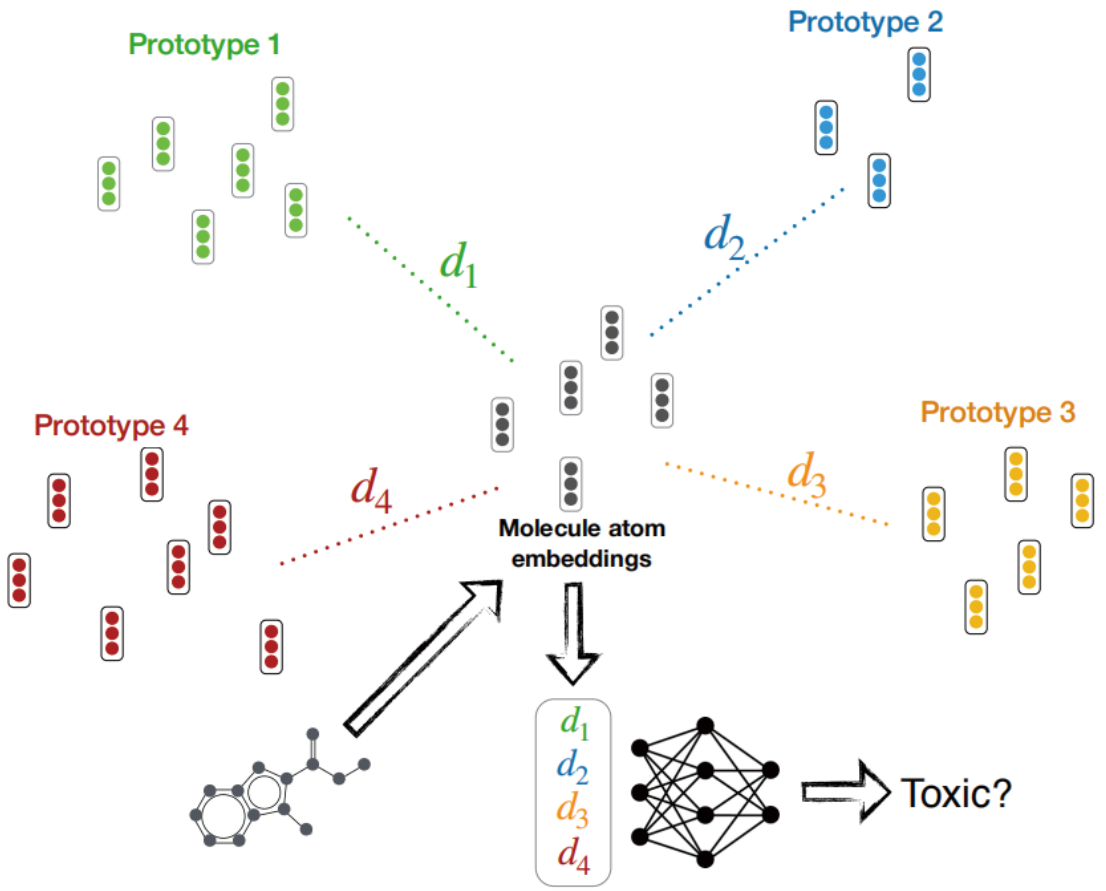


图 19：Wasserstein 原型

给定一个表征空间，该空间可以被表征为四种原型。为简单起见，每种原型都代表一组原子，这一组原子构成了某种分子（实际上，这种原型可以对表示任意学习到的集合）。当我们向嵌入模型输入一个新的分子（同样由一组原子组成）时，嵌入模型会通过计算该分子与四种原型的距离，将其转换为一个向量。这样一来，我们就可以通过一种精细得多的方式表征不同的分子结构。

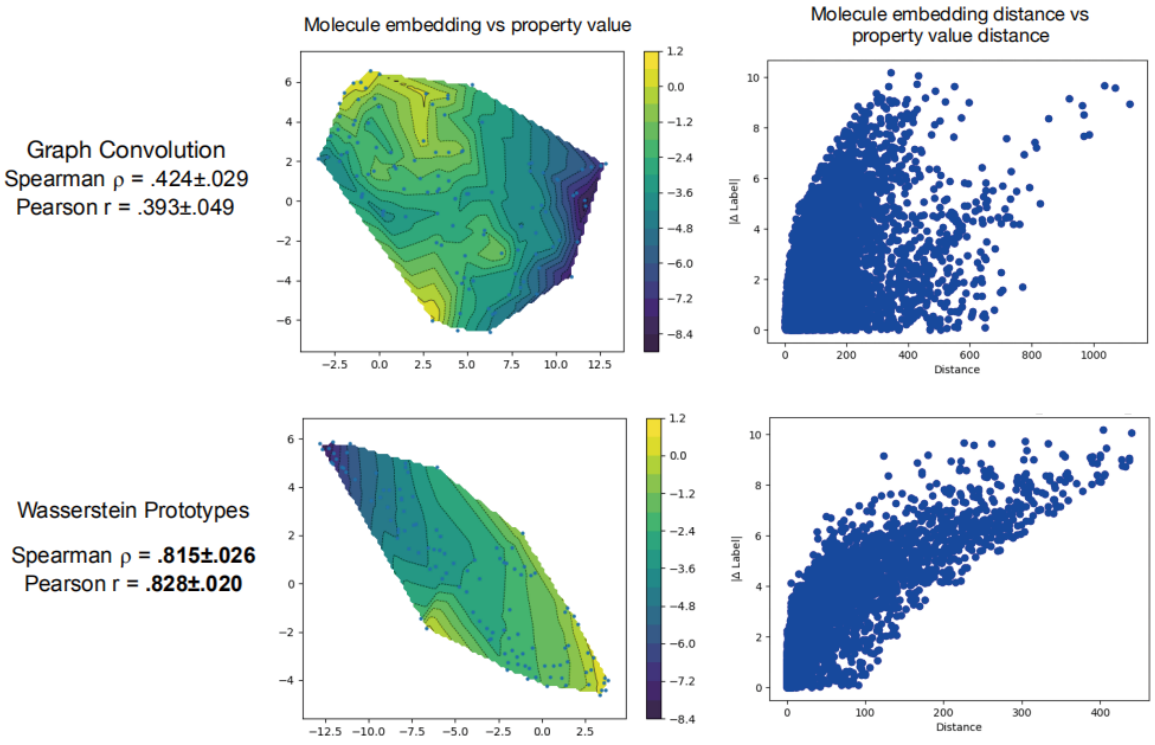


图 20：真实模型潜在空间中嵌入的可视化结果

图 20 显示了我们将潜在空间中的嵌入投影到 2 维空间中的情况，不同的颜色代表不同的特性。如前文所述，我们希望该潜在空间具有如下的几何性质：相近的区域拥有相近的特性的值，距离较远的区域拥有差距较大的特性的值。仔细观察通过 Wasserstein 原型构建的潜在空间，我们发现，从左上角到右下角，代表特性的颜色缓慢地从深绿色过渡为浅黄色；而当使用标准的图卷积网络时，我们发现在整幅图中，特性的值始终在非平滑地改变。

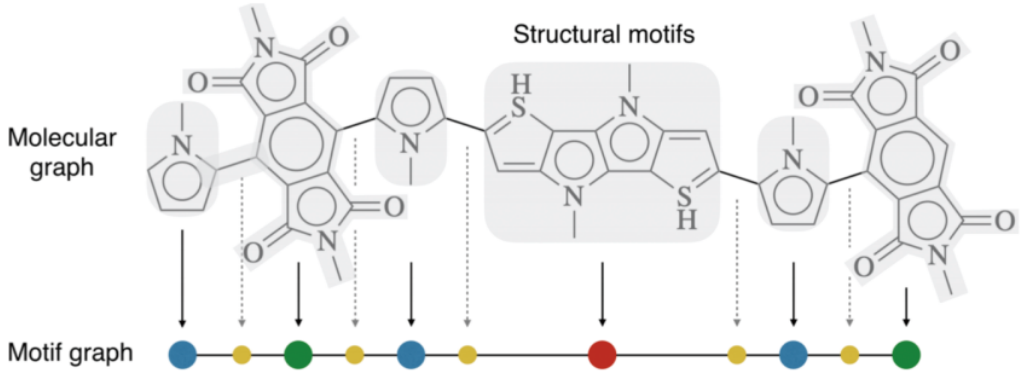


图 21：结构化的模体（motif）

Wengong Jin 的工作「结构化的模体」也与该的话题相关。当一名计算机科学家看到一个分子时，我们只能看出它是一张图。而当化学家看到一个分子时，他们会发现一些子结构。这就好比我们在观察一个英语句子或中

文句子时，我们可以将一些字符的组合看成具有特定意义的单词，这有助于我们解释句子的意思。同样地，化学家也将这些子结构（模体）看做构建分子的大型模块。

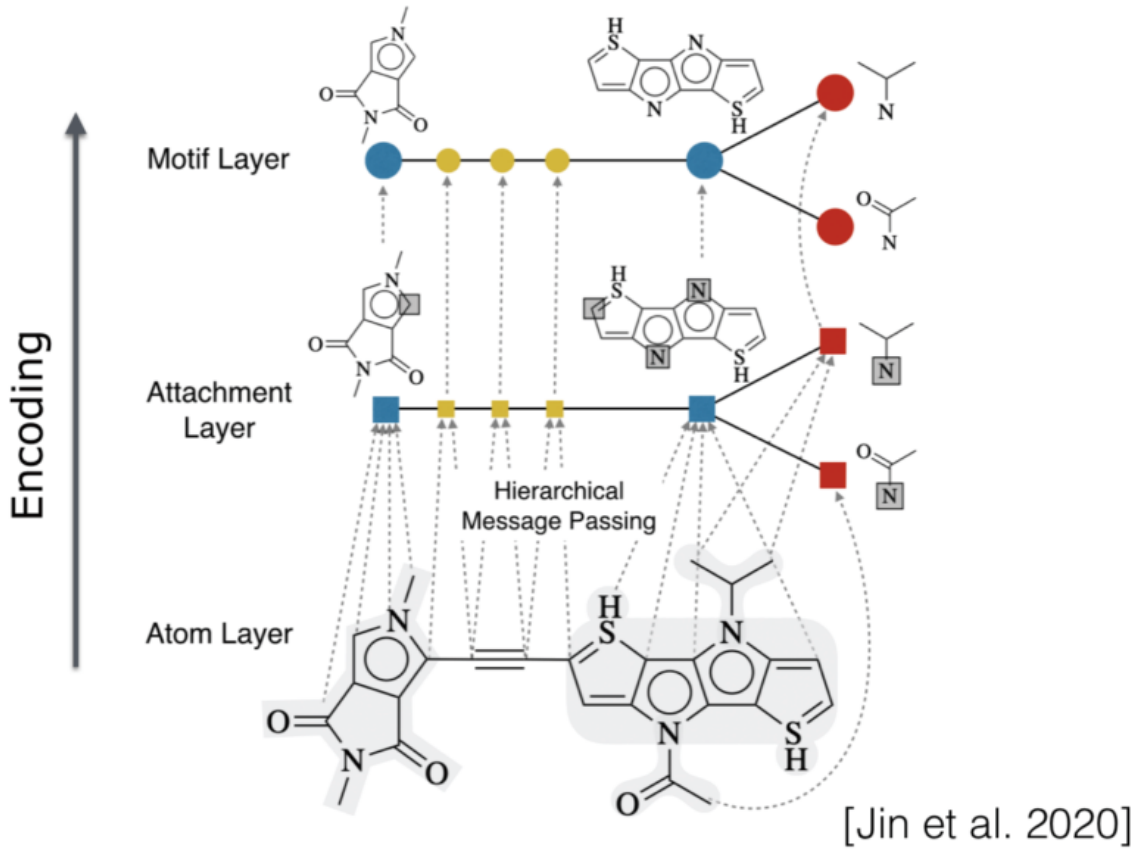


图 22：由细到粗的图编码

在 Wengong Jin 最新发表的 ICML 论文中，他展示了如何对图编码，从而学习到这些构建模块，并将他们组合起来，这种编码可以保留层次化的表征。

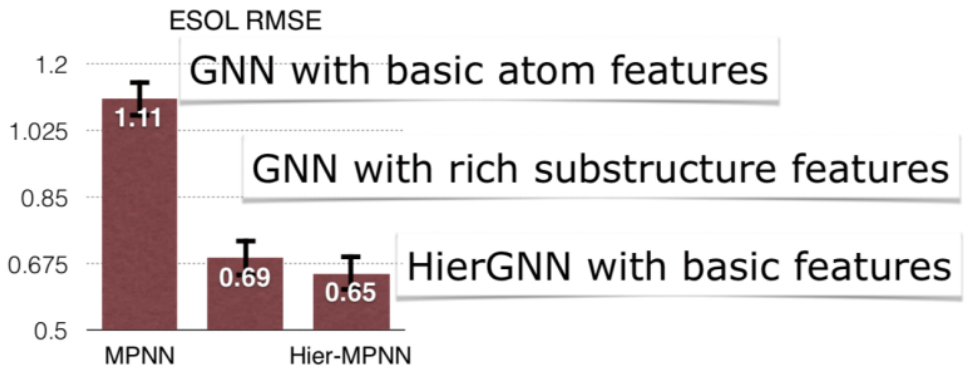


图 23：层次化推理实验结果

实验结果表明，以层次化的方式进行这种推理是十分重要的，可以切实提升模型的性能。因此，考虑大分子（如聚合物）也变得十分重要。可见，模仿化学家观察分子图的思维方式是十分有帮助的。在这个领域中，一个有待解决的问题是：我们应该如何处理三维的分子表征？因为分子实际上存在于三维空间中。当我们与化学家交流时，他们总是认为我们需要引入三维信息。但是据我所知，目前还没有工作表明，融合三维信息真的能提升二维模型的性能。我并不认为这个研究思路是错的，但我们仍然需要思考如何有效地达成这一目的。

三、泛化性

接下来，我们将讨论另一个极为重要的话题：泛化性。

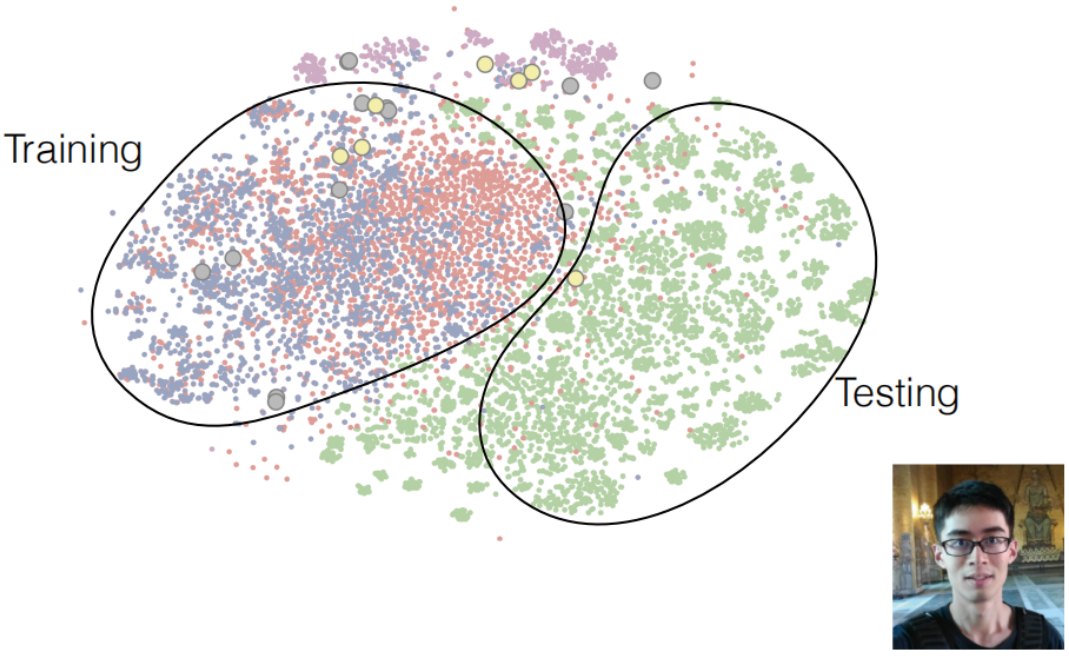


图 24：化学空间中的泛化性能

我经常与制药行业的人交流，当我首次涉足人工智能药物发现领域时，我在「如何划分训练集、验证集、测试集」这个问题上犯了难。根据我之前在自然语言处理领域的经验，我会将一个语料库划分为训练集、验证机、测试集，然后取得较好的结果。然而，制药行业的人关心的并非是这样的问题。他们想要做的是，将一个数据库划分为一些 scaffolds（组装的较长的基因序列），scaffold 就好比分子的骨架，我们期望测试数据与训练数据的差别较大。这是因为，也许他们出于某种纯粹的目的筛选出了某种分子，而当目标产生改变时，他们想看看在化学空间中的另一个部分会发生什么？

图 24 显示了用于抗生素发现的数据可视化结果。图中蓝色部分是 Wengong 手动收集的用于训练模型的数据。而我们最终将训练好的模型应用于绿色的无锡化学库。测试数据和训练数据的差异非常大，而当测试数据与训练数据的距离越远时，模型就会产生越大的误差。而在化学领域，能够在数据空间中的某一个部分上训练，而在另一个部分上测试，对于模型来说是非常重要的，因此我们对模型泛化性能的要求很高。

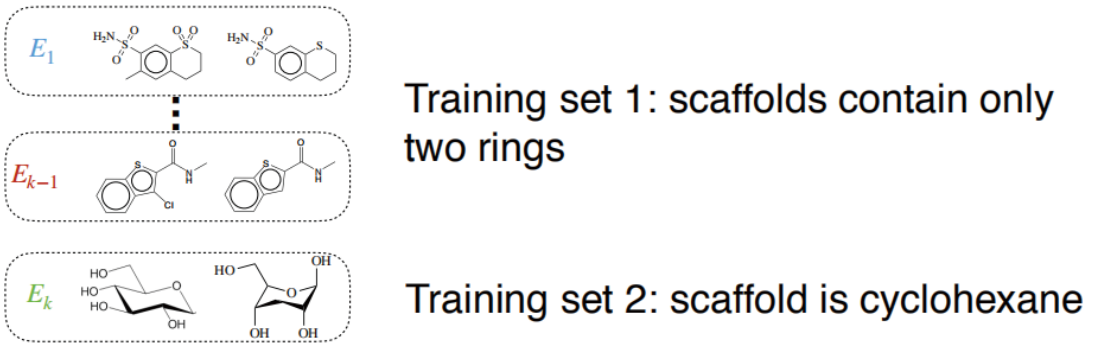


图 25：在 scaffold 之间进行泛化

Wengong Jin 在他的新论文「Domain Extrapolation via Regret Minimization」中试图通过扩展「不变性最小化」(invariance minimization) 框架来实现这一目标。这份工作的主要思路是，迫使算法通过创建人造的环境在数据上泛化，这些环境表示若干组包含不同 scaffold 的分子。你可以认为，我们通过某种方式将整个训练集分成了多个子集，其中每个子集与其它子集中的 scaffold 差异很大。如图 25 所示，训练集 1 中的 scaffold 仅仅包含两个环，而训练集 2 中的 scaffold 则是环己烷。

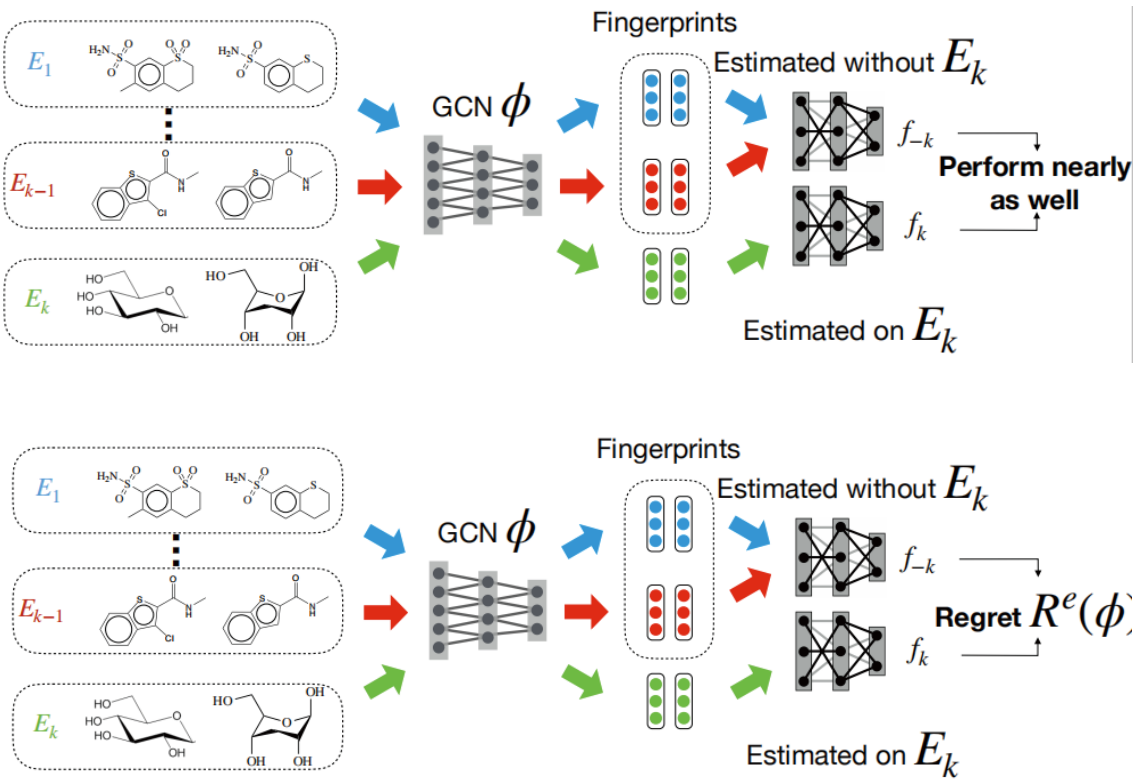


图 26：遗憾最小化算法的实现

下面，我们将介绍遗憾最小化算法 (Regret Minimization) 背后的思想。首先，我们训练某种分类器，它会学习用于预测的分子表征。我们想要确保，在不使用训练集 2 (oracle domain) 的情况下和仅仅使用训练集 2 时训

练出的模型性能相当，即遗憾被最小化。在本次演讲中，我们将跳过他论文中非常有趣的一部分，在这个部分中，他思考了如何划分不同的数据域。因为，在化学领域中，scaffold 有一个树形结构，因此这就变成了一个组合问题。Wengong Jin 设计了一种非常巧妙的方式，实现了对于空间扰动的动态不变性。

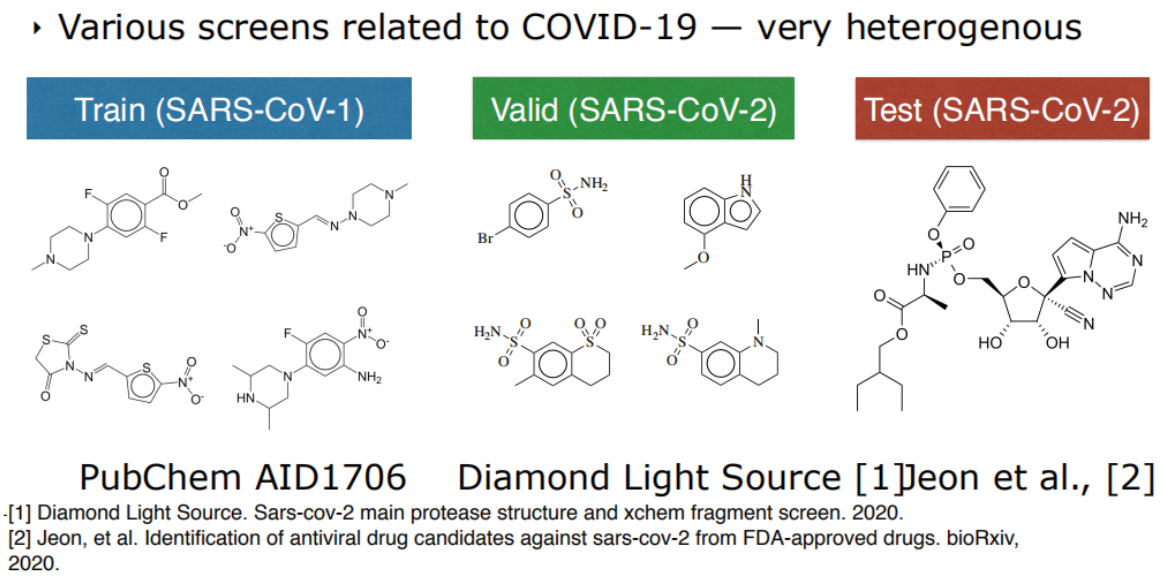


图 27：在异质数据（COVID–19）上的数据集划分

在图 27 中，我们向大家展示了这种方法在困难场景下取得的惊人的性能，作者基于分子的质量将数据集划分成了对于泛化性能非常具有挑战的形式（训练集的分子质量小于 400，验证集分子质量介于 400 到 500 之间，测试集的分子质量大于 500）。实验结果表明，采用本文提出的训练方法可以取得显著的性能提升。

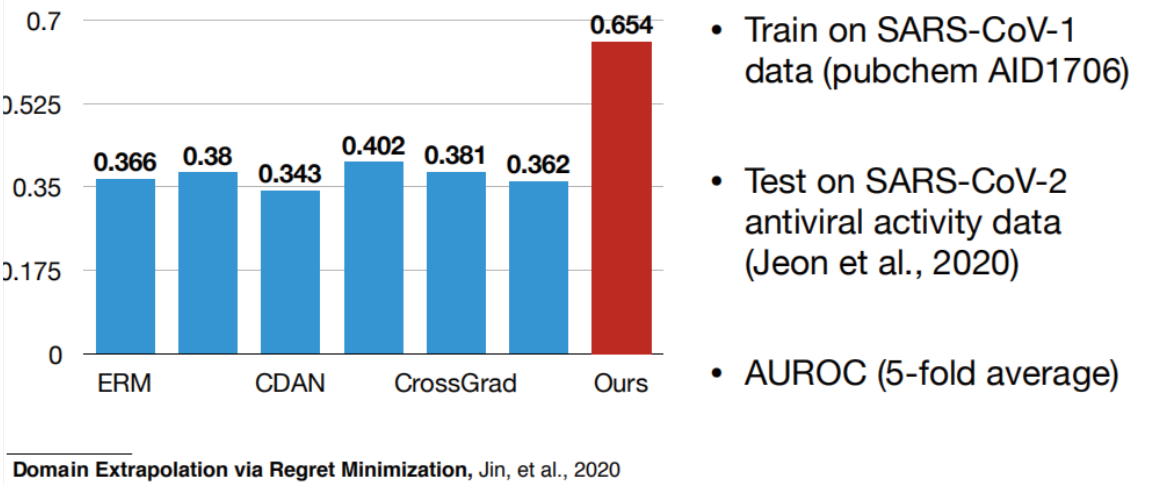


图 28：在新冠抗病毒数据集上的实验结果。

他还在新冠数据上进行了测试，使用 CoV–1 数据作为训练集，使用 CoV–2 数据作为测试集，取得了非常显著的性能提升。

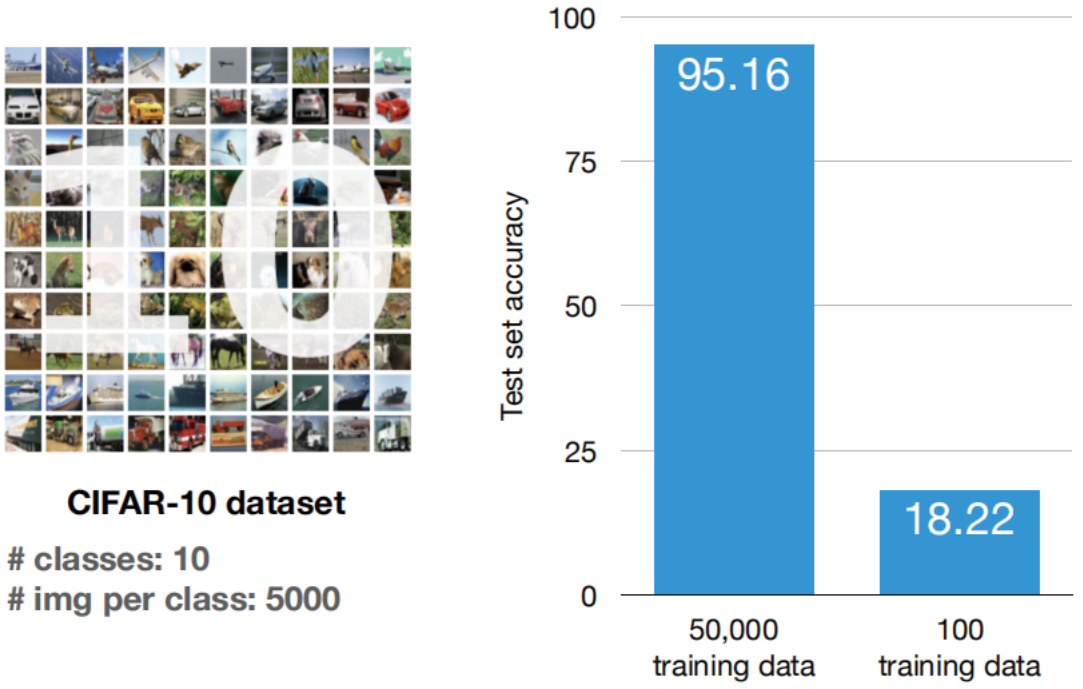


图 29：模型过度依赖于数据

下面，我希望各位读者能够帮我解决一个困扰我多年的问题，我并不认为人们已经找到了有关这一问题的解决方案，即使许多人声称他们做到了。如图 29 所示，在计算机视觉领域，当我们训练模型时，若训练数据量减小，则模型的性能会急速下降。

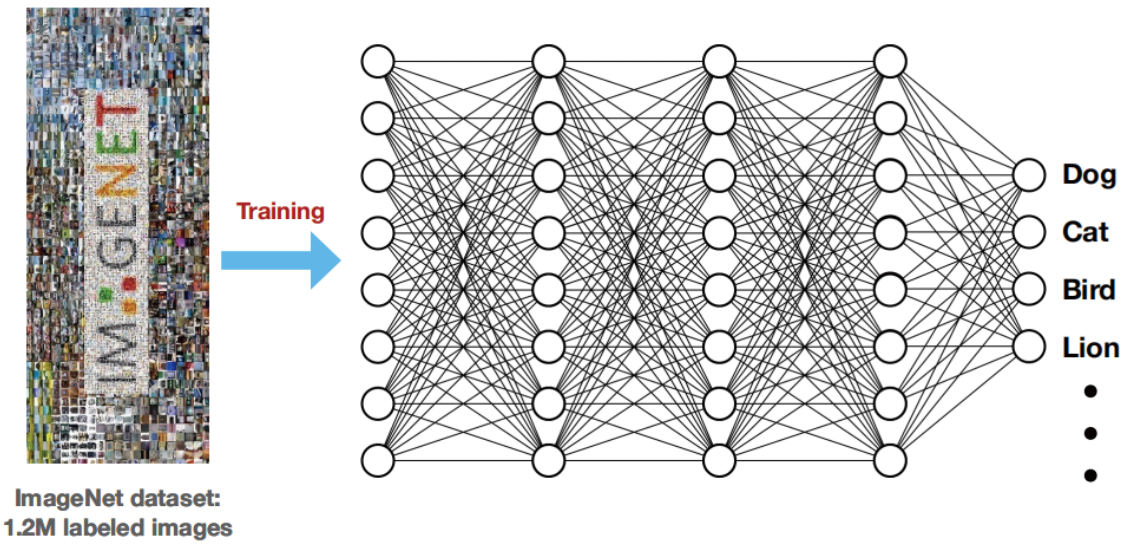


图 30：通过预训练进行初始化

在计算机视觉（CV）和自然语言处理领域（NLP），我们往往会采取预训练技术。此时，我们采用可以得到的大

型数据集对模型进行预训练，然后将预训练好的模型用于感兴趣的下游任务。

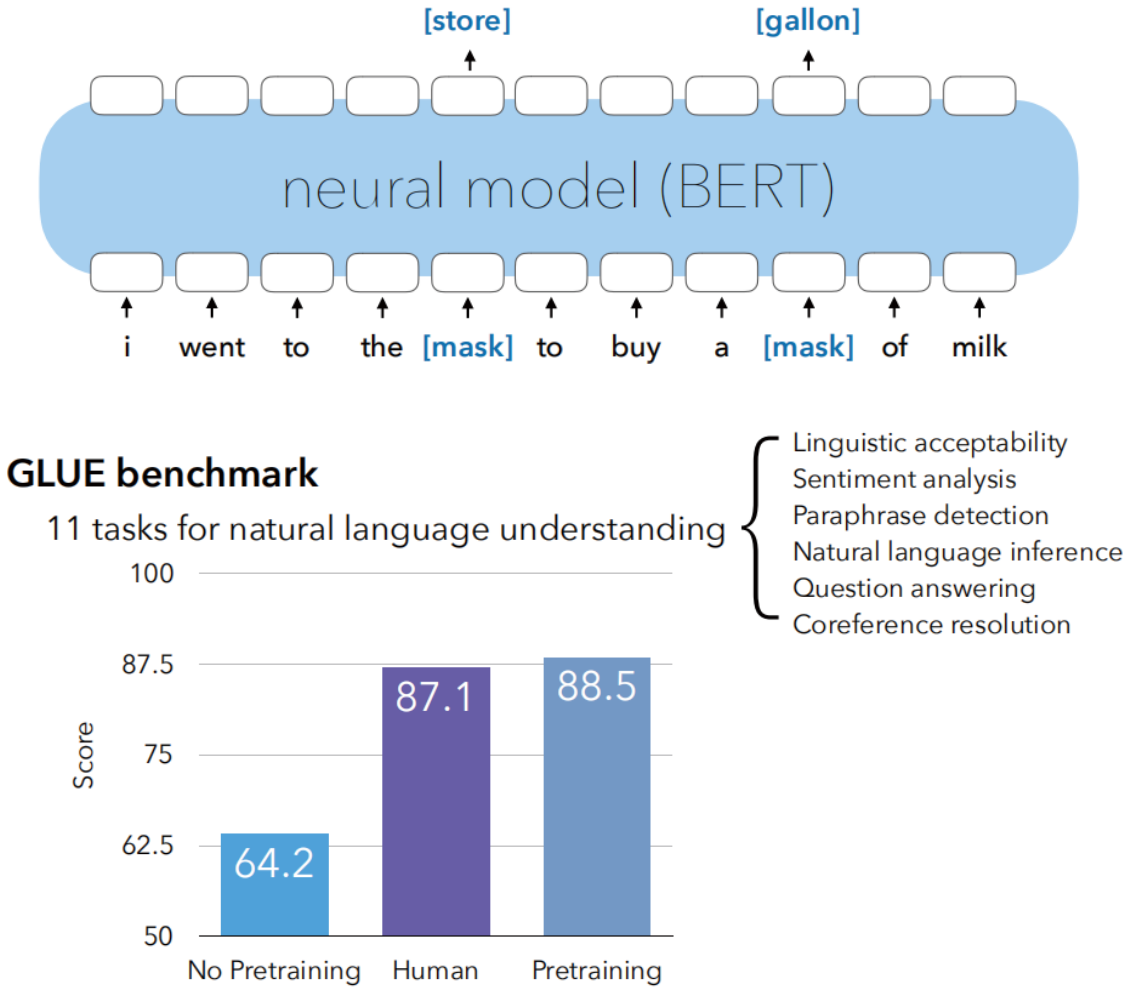


图 31：NLP 的预训练

许多读者也许都曾经看到过图 31 所示的 NLP 领域中的预训练示意图。在图 31 中，研究人员使用 11 种任务进行了实验。实验结果表明，相较于蓝色的对比基准，使用大量数据进行预训练的模型取得了超过 20% 的巨大性能提升。

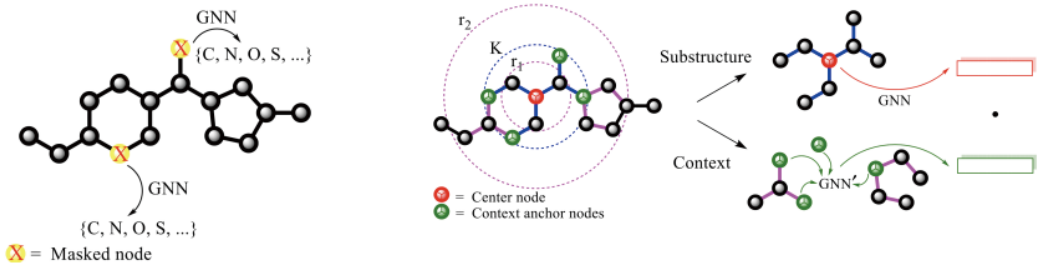


图 32：直接将 NLP 领域中的预训练技术迁移到化学领域行不通。

似乎这种预训练的思路可以很容易地被迁移到化学领域（化学领域中有数十亿的分子），我们可以直接借鉴自然语言处理领域的思路。例如，在 NLP 领域中预测句子中的单词可以类比为在化学领域中预测分子中的原子及其邻居。然而，有趣的是，这样做完全行不通。即使有人声称他们通过预训练取得了极其微小的性能提升，但这远远不及我们在 NLP 和 CV 领域看到的那样。尽管 MIT 的团队非常努力地实现这一目的，但是至今仍收效甚微，我们也不清楚这背后的原因。然而，该领域的研究对于提高数据的利用率是极为重要的。

四、不确定性估计

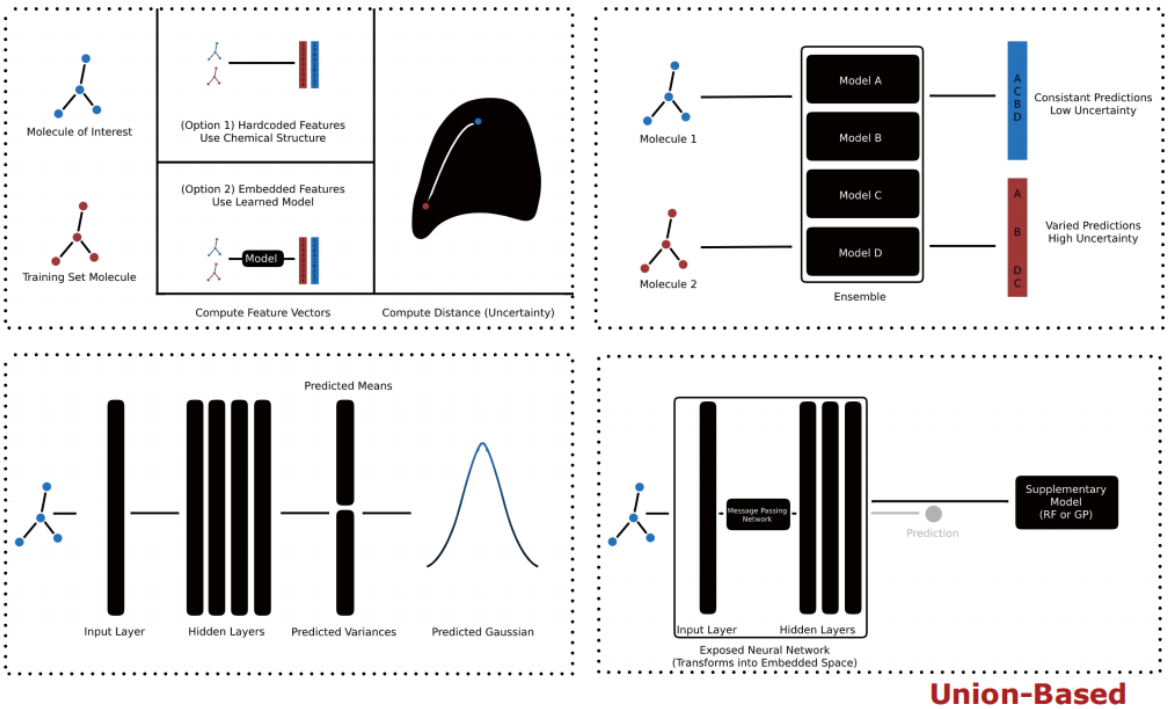


图 33：不确定性估计的各种方法

下面，我将简要介绍一个与化学领域非常相关的具体问题： 不确定性估计。这一问题在 NLP 和 CV 领域很少出现，然而化学领域的人却非常关心该问题。根据我在设计 Halicin 的过程中仅有的一点化学领域的经验，在设计好一个模型后，你需要使用数以亿计的化合物作为输入运行该模型。现在，我们找出了这些化合物中的一个子集，模型认为子集中的化合物活性很强。由于预算和时间有限（购买每个分子可能需要花费数千美元），我们需要决定最终应该购买哪些分子，以及我们能够在多大程度上相信模型输出的结果。不幸的是，直接使用预测器的概率效果并不好。实际上，化学领域的研究者们已经在这个方面开展了大量的工作，因为我们需要知道对于预测结果的置信度如何。例如，我们可以计算测试分子与训练分子在使用化学结构硬编码的特征空间中或嵌入空间中的距离。此外，我们还可以采用集成学习的方法，查看不同的模型得到的预测结果是否一致。还有一些方法，可以显式地预测出方差。

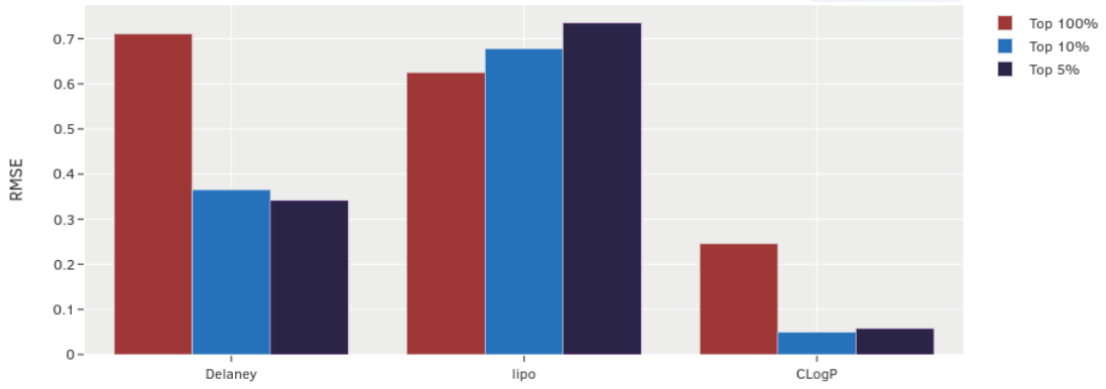


图 34：不确定性估计——朋友还是敌人？

在论文「Uncertainty Quantification in Molecular Property Prediction using Message Passing Networks」中，我们在多个数据集上采用了这种方法，实验结果让人非常忧虑。如图 34 所示，误差越小越好。在 Delaney 数据集上，红色的部分代表使用 100% 的数据得到的误差，而蓝色、紫色的部分说明采用挑选出来的子集可以有效减小误差，这是一种非常好的情况。然而，在数据集 Lipo 上，情况竟然完全反过来了，当我们选用置信度最高的一些数据训练时，模型性能反而下降了。在 NeurIPS 上，有工作旨在验证预测的结果（幻灯片中未列出）。我认为，在很多场合下，都需要进行有选择的定量分析。我们如何设计一种新的机器学习模型，它只在很有把握时才作出预测。

五、机制理解

如今，可解释性在 NLP 和 CV 领域中是非常火热的话题。通常，当我们考虑可解释性时，会高亮显示出数据（例如，医学影像）中呈阳性（正例）的部分。在这里，我们通过另一种完全不同的方式思考可解释性。

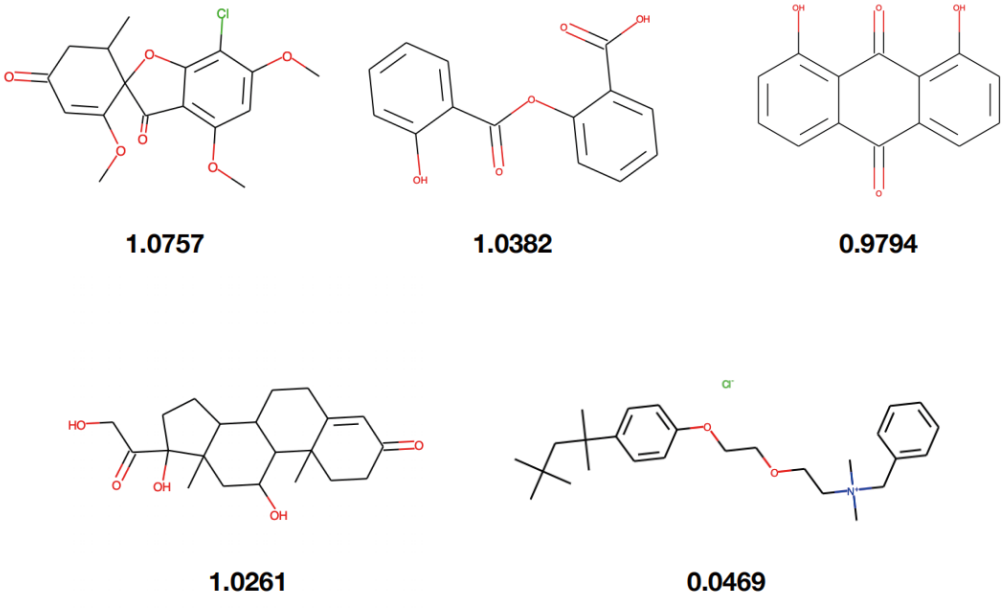


图 35：训练数据示意图

对于研究以上化学问题的计算机科学家来说，假设给定分子结构图和一个数字，而并不知道这些数据是怎么得来的，我们需要学习他们之间的关系。

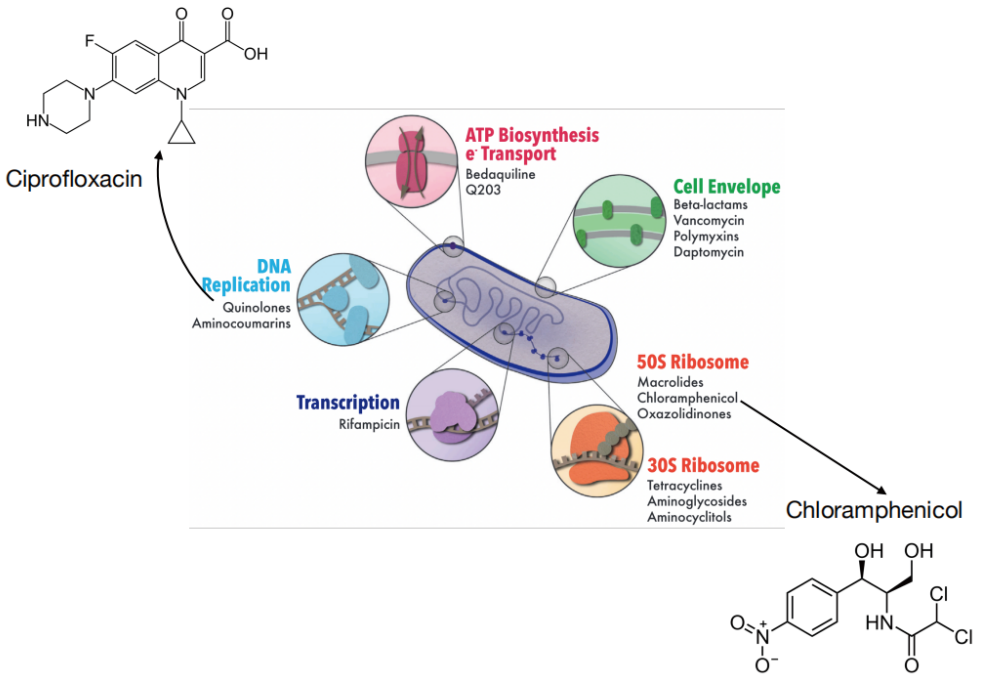


图 36：理解背后的机制。

但是实际上，这背后是有一套机制的。对于某些具有毒性的物质，生物学家会对其作用机制做出详细的解释。在完成我们发表在《Cell》上的论文的过程中，即使我们已经确定了分子，计算机科学家们还是花了非常长的时间找出这种分子杀灭病原体的机制。

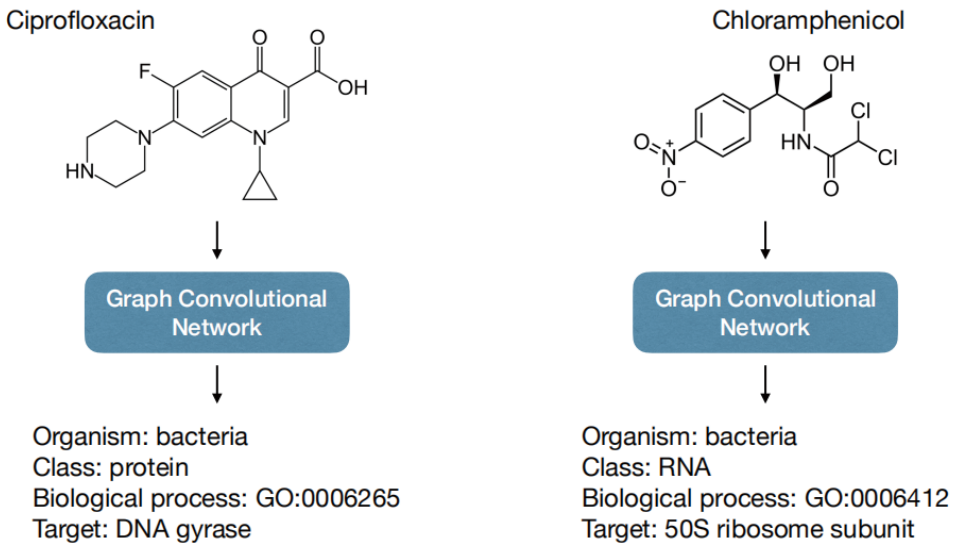


图 37：环丙沙星与氯霉素

我们希望为化学家设计出不但能够预测分子活性，还能够给出背后的生化机制的模型。目前尚不确定是否有人在这方面进行了研究，但是我们将在这个具有广阔前景的领域继续进行探索。

六、全新药物设计

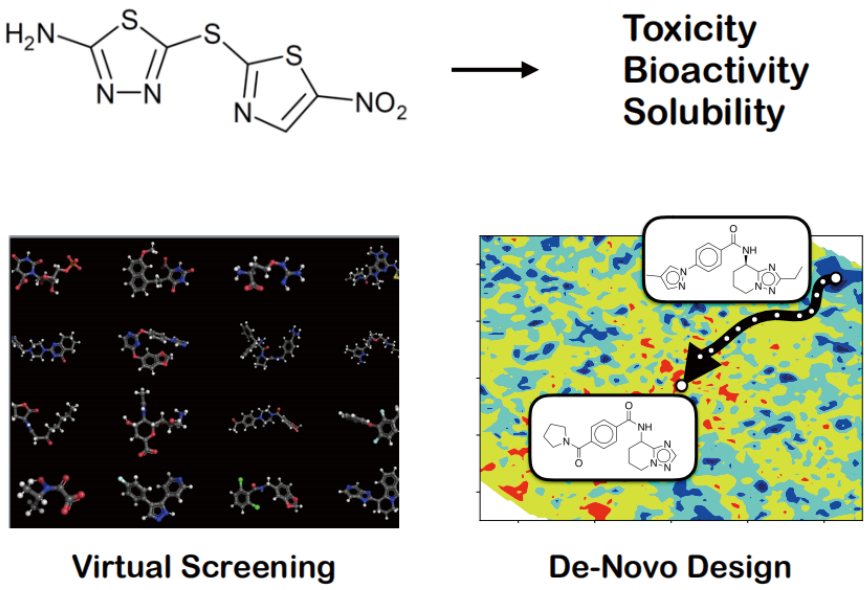


图 38：全新药物设计

接下来，我想继续讨论全新药物设计 (de-novo design)。在前文中，我们假设已经拥有了各种各样的分子，我们只需要从中挑选出一些符合要求的分子。但是，如果我们想要设计一种从来没有出现过的新分子怎么办呢？在这里，我们可能要面对一个分子结构优化的问题。给定具有某种功能的分子，其功效并不好，我们希望创建一种具有更好的特性的新分子。

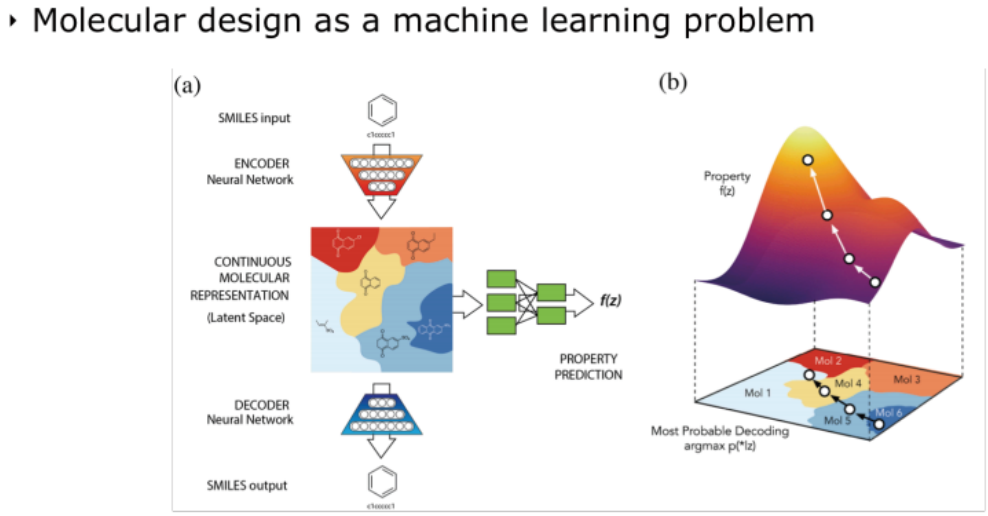


图 39：将全新药物分子设计作为一个机器学习问题。

在这里，我想再次向大家介绍将为不同领域开发的计算机科学技术组合应用会发生什么。Gomez–Bombarelli 等人于 2018 年发表的论文继承了一些传统计算机视觉领域的思路。给定一个分子，假设我们要预测其特性，我们首先将其编码到一个潜在空间中，通过梯度下降对其进行优化，然后得到一个更好的分子。但实际上，尽管他们是第一个想到该问题并提出该问题的人，但是上述方法的效果并不好。这是为什么呢？改进的空间又在哪里呢？

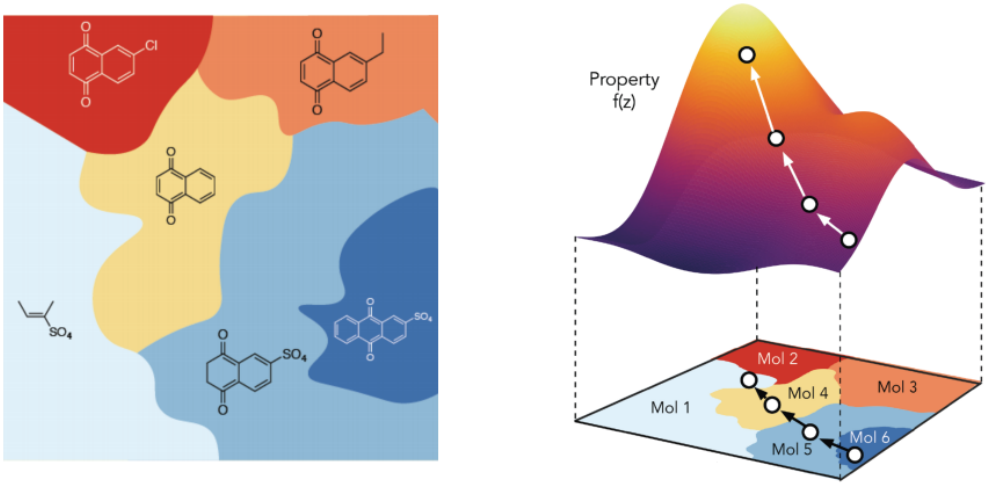


图 40：连续的表征

这是因为该领域的潜在空间并不平滑，在分子嵌入的等高线（contour）周围有很大的梯度偏移。那么，我们应该如何在这种极其不平滑的复杂空间中采取优化策略呢？

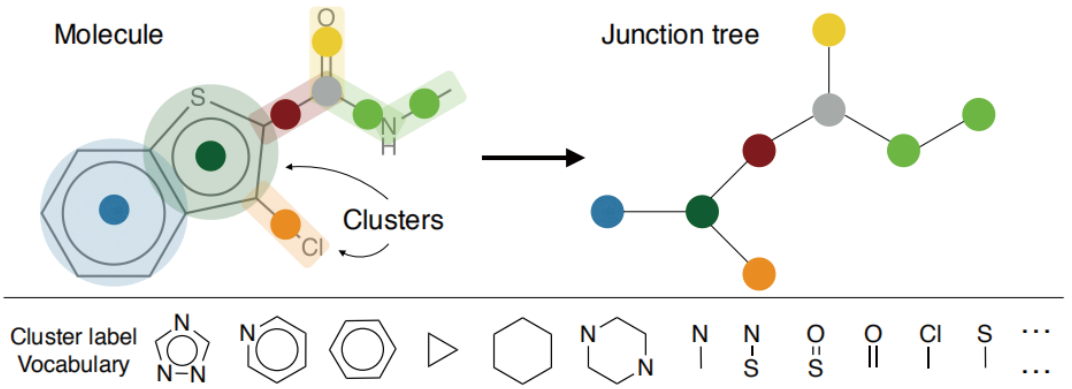


图 41：联结树

我们用到了许多重要的思路来实现这一目标。例如，我们可以使用一种联结树的方式进行编码，更多地进行层次化的编码是有所助益的。

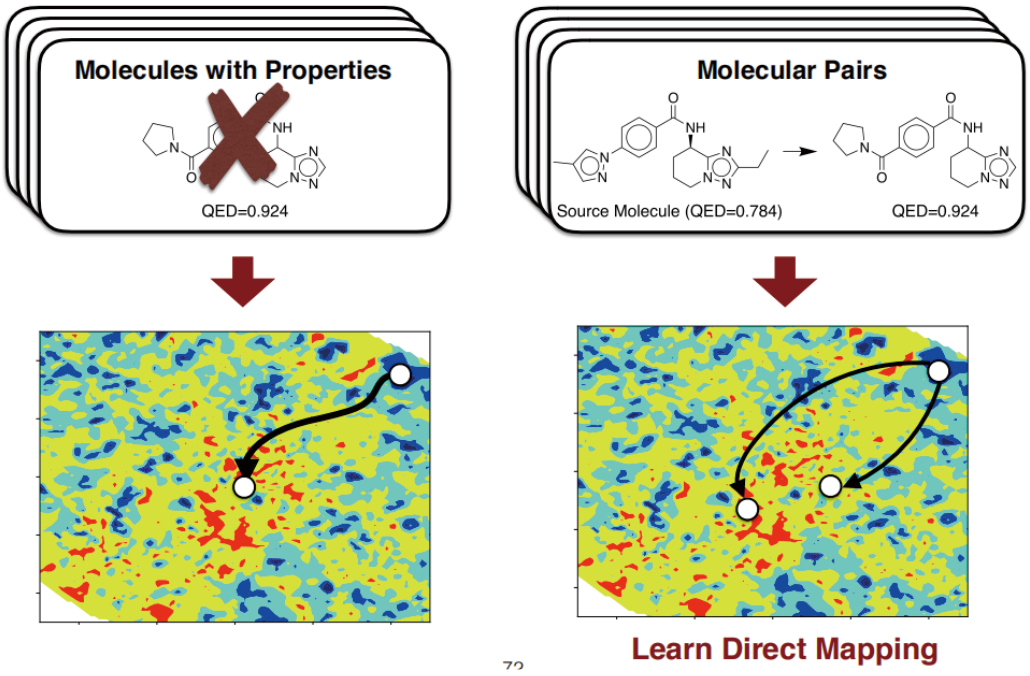


图 42：利用分子对进行学习

第二种方法用到了一些机器翻译的思想。该算法从某一个初始化的点出发，序列化地生成分子。这种方式缺乏约束，会在很多地方出现错误。

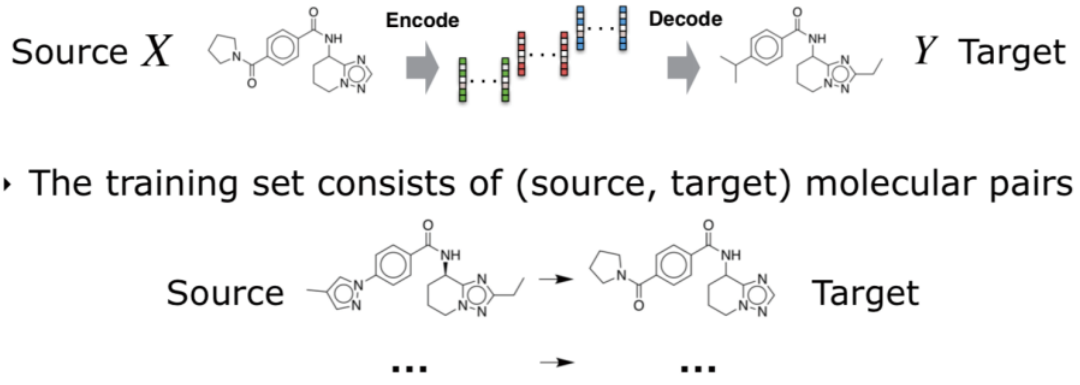


图 43：通过图变换的方式进行优化

假设我们拥有由许多分子及其特性组成的训练集，我们可以识别出具有不同距离、不同特性的 (source, target) 分子对 (而不是单个分子)。这样一来，我们就可以借鉴机器翻译任务中的思想，从一个原始的分子出发，然后生成一个有着更好的特性的邻居节点。

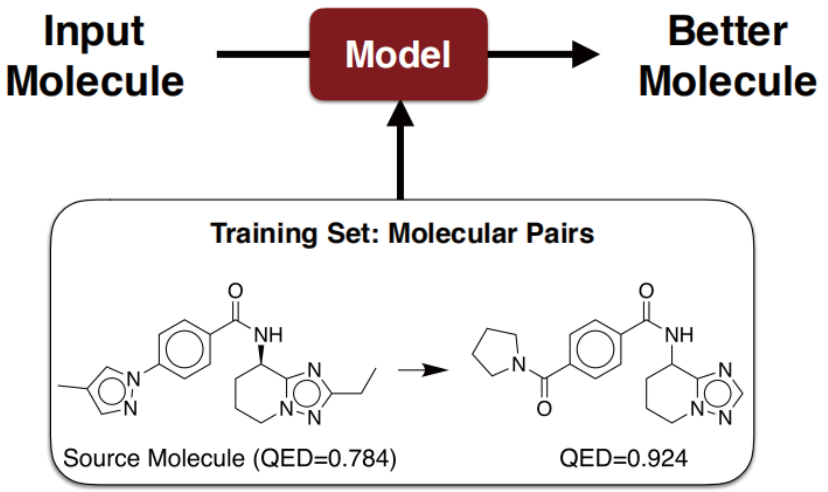


图 44：优化过程

我们可以通过一个标准的机器翻译工作流程对这个过程进行建模，在给定一个起始分子、一个终点分子的情况下，我们需要学习中间的「编码器 – 解码器」架构，这个架构似乎效果相当好。

七、多目标分子设计

最后，我将介绍 Wengong Jin 最新发表在 ICML 上的论文的核心思想。在前文介绍的类似于机器翻译的工作中，模型可以在构建的分子只包含一种特性时取得较好的性能。但是，如果我们希望创建的分子包含多达 20 种特性又该怎么办呢？实际上，这种任务在制药工业中非常常见，但是这方面的工作仍然还处于空白状态。

在这篇论文中，Wengong Jin 模拟了化学家设计分子时的思维。如果你想要得到具有某种特性的分子，可以设计某些子图代表的相应的官能团，当我们需要其它特性时，就可以进一步将其它的子图与之相结合，从而得到更好的分子。具体而言，我们的策略包含两个步骤：(1) 基本原理提取。我们需要训练模型去预测某些特性（例如，溶解性），同时学习究竟是哪种子图（官能团）导致分子具有这种特性。这样一来，对于每一种特性而言，我们都可以找出具有这种特性的子结构。(2) 多原理集成。在设计具有多种特性的分子时，我们可以将表征各种特性的官能团集成到该分子中。

Method	DRD2 + GSK3β + JNK3		
	Success	Novelty	Diversity
GVAE + RL	0%	0%	0.0
GCPN	0%	0%	0.0
REINVENT	48.3%	**100%**	0.166
Ours	**86.2%**	**100%**	**0.726**

图 45：实验结果

实验结果表明，我们的方法取得了显著的性能提升。在论文中，我们还介绍了如何创建一些精准的抗生素，它们能杀灭病菌，但是保留身体中好的组织，或者使某些特性仅仅针对某些特定的病原体生效。

八、结语

「ML+ 化学」是一个飞速发展的研究领域，该领域中最好的模型在算法上是具有创新的，并不仅仅是将为图像设计的卷积神经网络用到分子上这么简单。每当我们看到的新的算法，它们都会带来实实在在的性能提升。诚然，某些该领域的算法也是由通用机器学习算法发展而来，但是该领域仍然存在巨大的研究空间，这可以让计算机科学家能够产生更大的社会影响力。