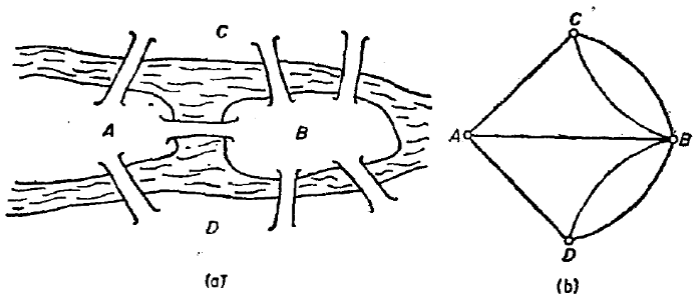
图神经网络架构简单介绍

第一部分：图的引入

图的由来

- 七桥问题



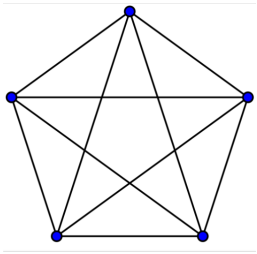
> 在达到目的地的前提下，是否可能已有且仅有一次的方式走过每一个桥？

欧拉于1735年证明符合该条件的走法并不存在，并将这类问题归结为一笔画问题

在现代图论中，我们会使用抽象的节点替换每个陆地，用抽象的边替换每一座桥（就相当于简笔画）。就好似物理中的"质点"一般，我们只关心实际对象的最关键的**连接**信息（"每对节点是否存在连接的边"），至于边的长短，边的形状我们并不关心。

图的概念

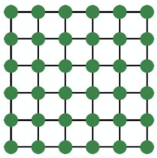
图，是一种描述一组对象之间相互关联的关系的结构。包含**节点**与**边**，其中节点与节点之间通过边相互连接。



为什么需要图?

现实世界的**文本，语音，图像，视频**这类可捕捉到有限维距离度量的数据被称为欧式数据（定义在欧几里得空间）。而其余类似于**知识图谱**，**社交网络**，**分子结构**这类节点往往无序（见下图）

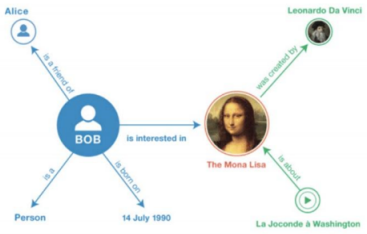
随着机器学习，深度学习的发展，语音，图像，自然语言处理逐渐取得了很大的突破。然而语音，图像以及自然语言都是很简单的序列或网格数据，是及其结构化的数据。其欧氏域数据表示的大量场景中取得了成功。（如：语义分割，医学影像分析）



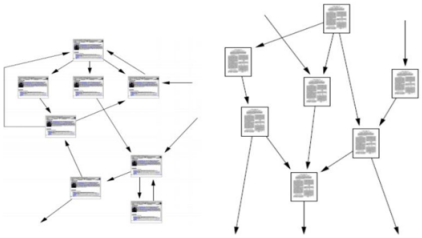
Images

但现实世界并不是所有东西都可以表示成一个序列或者一个网络。

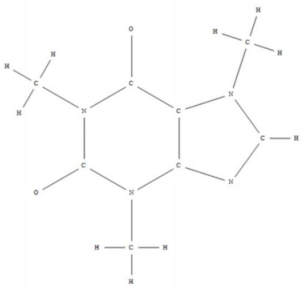
Social networks



Knowledge graphs



Complex Systems



Molecules

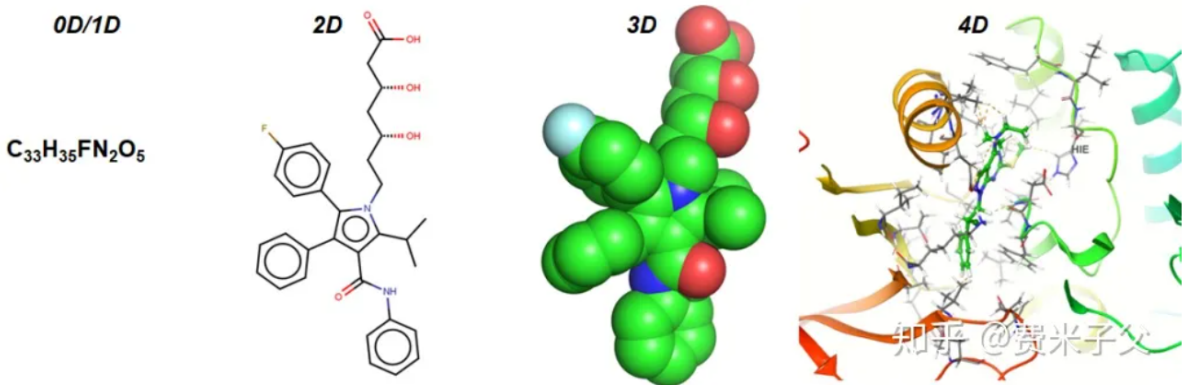
然而，非欧氏的数据在节点之间（甚至节点与子图，子图与社群）的关联信息十分庞大，对经典的机器学习方法学习无法**挖掘**到如此深度的**模式**，对其学习及其不友好。

而图作为一个强调对象之间的关联的结构有着得天独厚的优势。在处理无法结构化描述的非欧氏数据时，图不会受到空间的限制；能够极大程度地保留内在的结构信息，节点之间的关系可以被节点之间的距离捕获；

第二部分：图神经网络

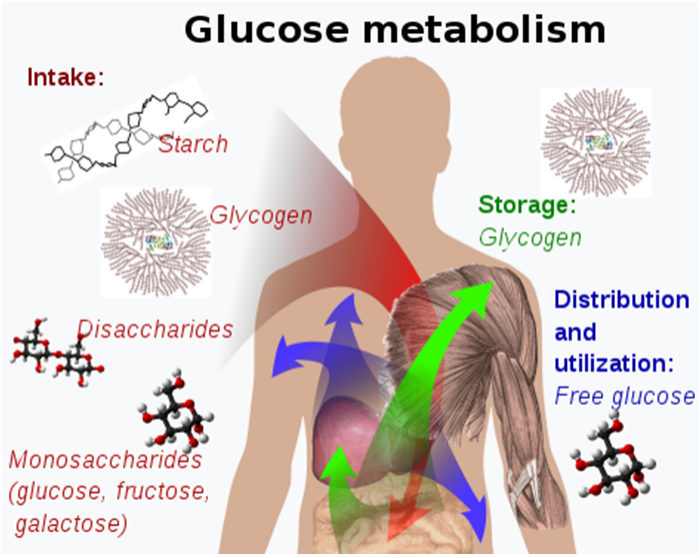
分子表征

要想描述一个分子，我们有哪几种方法？想一想分子在不同维度下所表现出来的特征

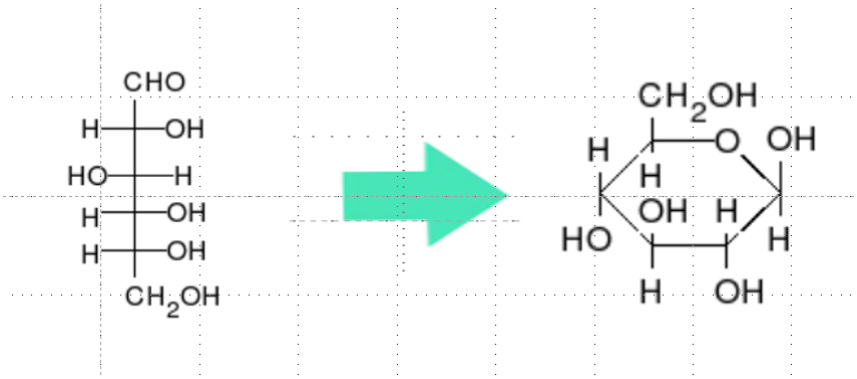


因此我们可以各类分子描述符，我们便可以从多个层面对分子的结构进行表达。

举个简单的例子



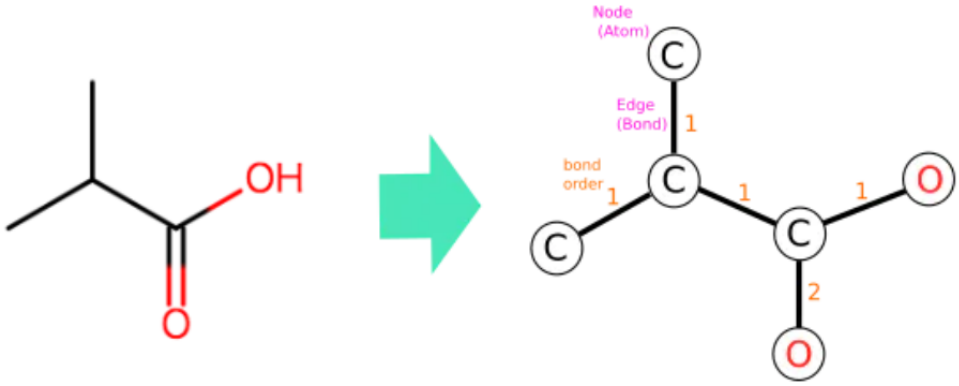
葡萄糖是生命体活动能量的来源之一，其中葡萄糖在生命体中通常以链状/环状化合物的形态下存在



如果只是利用序列去编码链状的葡萄糖，由于其只有特定部位羟基会与醛基发生亲核取代，其编码出的特征可能有失偏颇。

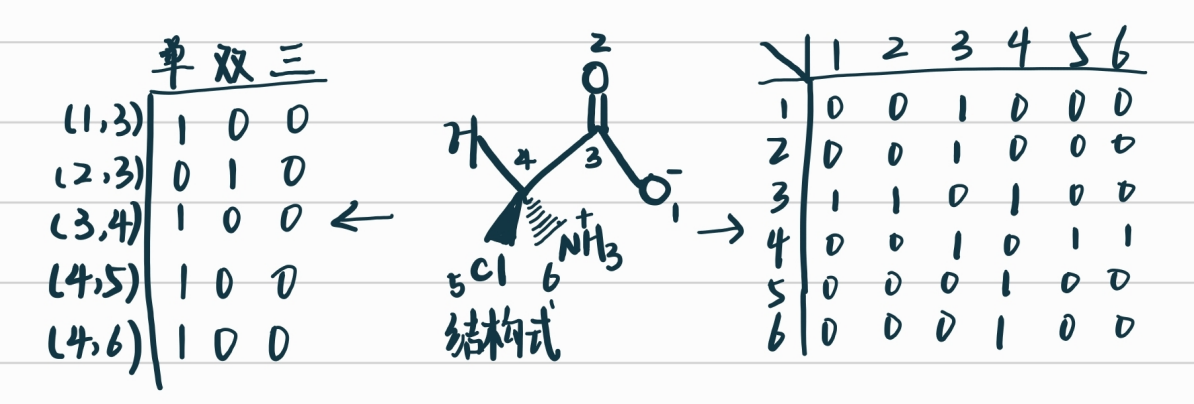
基于图的分子表征

计算机更善于处理数字而不是识图，从这一点来看，使用基于字符串（欧氏数据）的分子表征更有利于计算机处理，然而图能展现出比字符串更多的结构信息，**节点与节点之间的链接可以非结构化的距离度量**。数据更加完整，如果能利用图进行学习当然是不错的方法。换而言之，我们需要解决的是如何将分子图中的信息转变为计算机方便处理的信息格式。事实证明，矩阵提供了一个强大的工具，可以将图形转录成为计算机友好的数据集。



在上图示例中，我们将以**化合物的原子为节点，键为边**，在省略氢的情况下以图表示。 节点存储信息（标签），例如原子类型、电荷、多重性和质量，而边存储键合顺序。

但这类数据似乎并不好被计算机所理解，所以对一些我们关注的部分性质，我们可以用矩阵（非常适合于信息存储）的方式表达。（下图示例中 1表示**存在关系**，0表示**不存在关系**）



矩阵能够存储信息：左图的矩阵反映了边的特征；右图反应了节点与节点之间的联系；

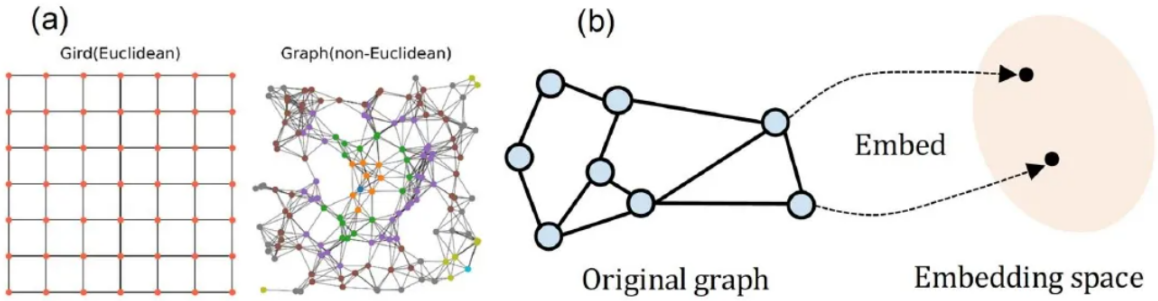
如上，图是描述分子的更自然的数据结构。**化合物可被翻译成矩阵**。可以设计不同的标准来整合分子中相邻原子的信息以及连接键的信息等。分子有许多特征，单节点或边能够代表的信息很少。因此**一个节点或边的特征还要由其相邻元素的加权求和来体现。**

至此我们知晓了化合物可以图的形式表达（原子可表示为节点，键则可表示为边），并可被矩阵形式数值描述

图神经网络的应用

既然我们已经解决了复杂的非欧式数据的表达的问题（实际上就是把**高维数据嵌入至易表示的低维空间**，努力地去保留图的拓扑结构和节点属性信息）；那接下来我们是不是就可以利用传统的机器学习/深度学习方法了？

是的，基于**嵌入方式的差异**，我们提出了各类基于图的神经网络模型。就像我们通常做的那样，将**欧式数据**或**非欧式数据**表示成可被计算机**识别**和**学习**的形式，然后对其建立一种可能的**映射**，从而实现对未知数据的分类。

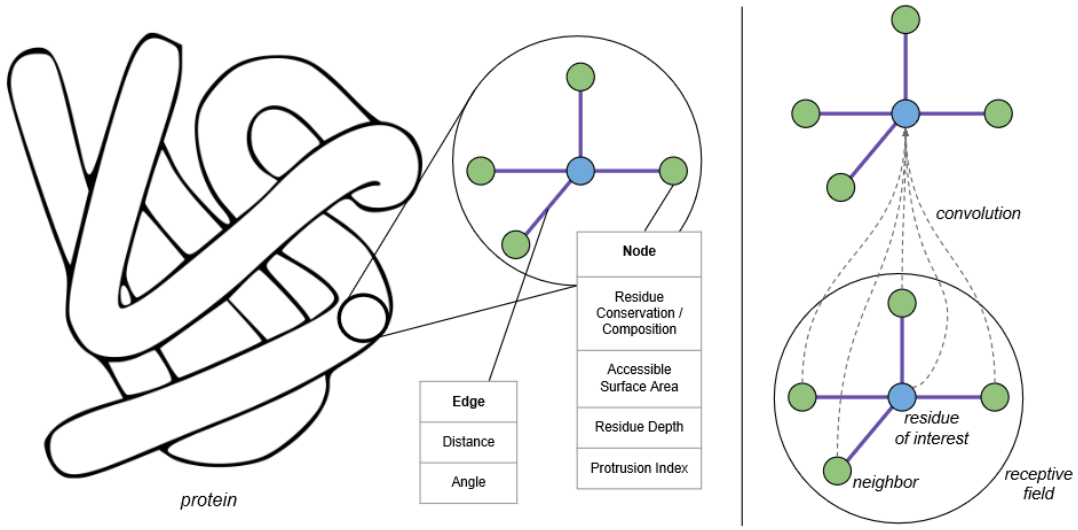


那么现在就让我们来看看图神经网络与化学，生物信息会产生怎样的碰撞

关注一篇2017年有关**蛋白质界面预测**的工作：[Protein Interface Prediction using Graph Convolutional Networks]

我们都知道，生命活动离不开蛋白质的参与。蛋白质之间的相互作用在许多生物学功能中起着至关重要的作用。而蛋白质与配体形成复合物时，便是通过蛋白质界面发挥作用的。

他们讨论了利用图卷积神经网络进行蛋白质界面预测的具体情况



通过对所关注的蛋白质的性质的合理编码（他们将蛋白质的结构表示为图数据，其中**节点**代表氨基酸残基序列信息及其周围的残基环境，**边**代表构型的几何特征），在聚合相邻节点的有关特征后。使得这个模型**能够学习**到**有效的潜在表示**，这些表示能够整合图中的信息，从而得出所感兴趣的结果。

> 其实重点就在于如何对非欧式数据进行合适的表示，有了合适的表示，就可以利用好模型来解决问题。

```
Title : A brief introduction to graph neural network architecture

AU : Hiragi

Time : 2024/1/26
```