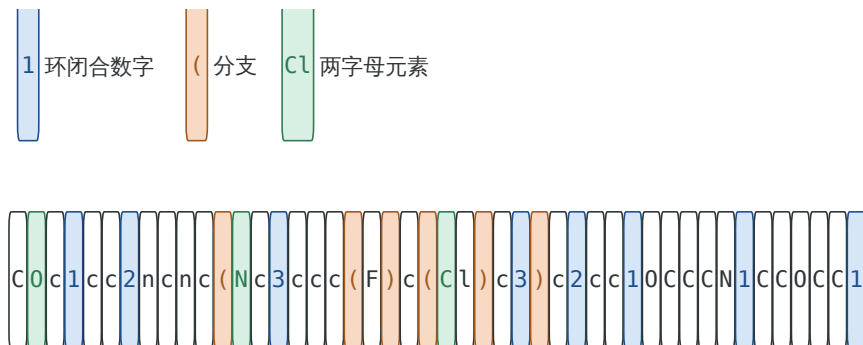


同一批例子 (gefitinib, ChEMBL939) : 四种分子表示

A

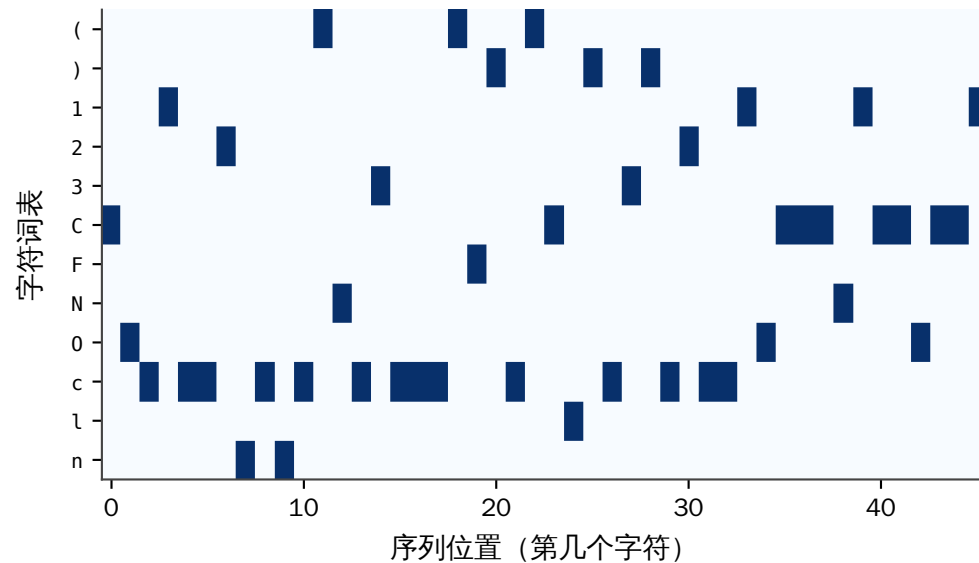
SMILES : 一串字符, 字符数 $\neq$ 原子数



gefitinib (ChEMBL939) : 46 个字符, 但只有 31 个重原子 (节点)

B

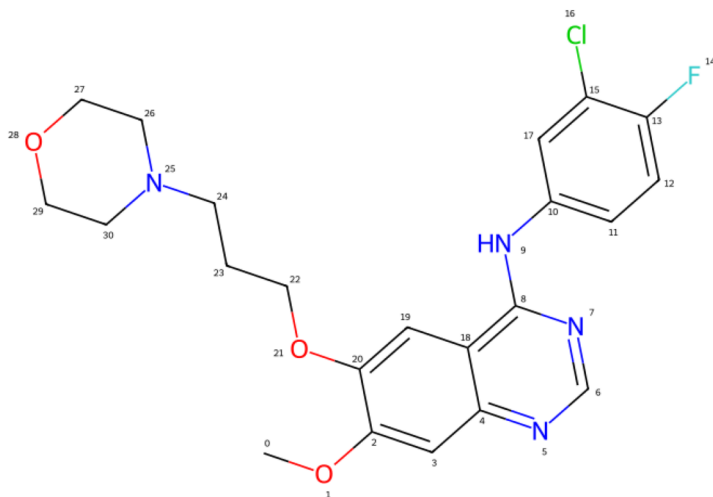
One-hot : 把字符变成数值向量 (字符级) 词表大小 = 12



字符级: 'Cl' 占两格、括号与环闭合数字各成类别 $\rightarrow$ 位置 $\neq$ 原子编号

C

分子图 : 原子=节点, 化学键=边



节点 31 个, 边 34 条; 图保留的是它所编码的属性 (元素/键级/电荷/手性), 不是三维构象

D

指纹 : 定长比特向量 (有损摘要)

两个指纹恰好都是 60 位为 1 (巧合, 含义并不相同)

MACCS
167 bit



60 位为 1

Morgan r2
2048 bit



60 位为 1

MACCS = 预定义结构键; Morgan = 哈希后的环形环境

横轴是位序、不是原子; 两行等长绘制仅为示意