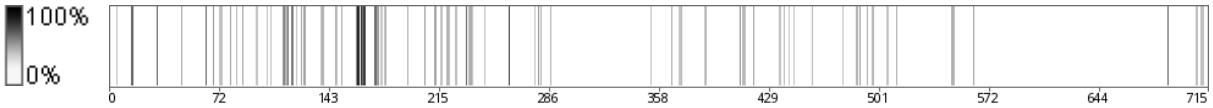
PaperPass旗舰版检测报告
简明打印版

比对结果(相似度):

总　体： 7% (总体相似度是指本地库、互联网的综合对比结果)
本地库： 7% (本地库相似度是指论文与学术期刊、学位论文、会议论文、图书数据库的对比结果)
期刊库： 4% (期刊库相似度是指论文与学术期刊库的比对结果)
学位库： 4% (学位库相似度是指论文与学位论文库的比对结果)
会议库： 1% (会议库相似度是指论文与会议论文库的比对结果)
图书库： 2% (图书库相似度是指论文与图书库的比对结果)
互联网： 1% (互联网相似度是指论文与互联网资源的比对结果)

报告编号：5CEB62D69D18EWZ82
检测版本：旗舰版
论文题目：化合物发育毒性预测模型构建
论文作者：肖力铭
论文字数：23861字符(不计空格)
段落个数：297
句子个数：715 句
提交时间：2019-5-27 12:08:54
比对范围：学术期刊、学位论文、会议论文、书籍数据 、互联网资源
查询真伪：http://www.paperpass.com/check

句子相似度分布图:



本地库相似资源列表(学术期刊、学位论文、会议论文、书籍数据):
暂无本地库相似资源

互联网相似资源列表：
1.相似度： 1% 标题： 《Benzodiazepine及核苷嘌呤衍生物活性...》
http://www.doc88.com/p-9189559365728.html

全文简明报告:

{70%：化合物发育毒性预测模型构建}

摘要

发育毒性是导致药物失效的一个重要原因，也是对药物进行安全性评估的一个重要测试项。传统的对药物发育毒性的评估方法主要是动物实验，具有开销大，耗材多，时间久的缺点。

鉴于越来越多的崭新化合物在不断被发现、创造，找到一个建立在大量实验数据基础与现存分析算法上的化合物虚拟评估方案成了当务之急。 {41%：由于发育毒性范围广泛，影响终点过多，难以一概而论。} 本文将以从上海有机所物质毒性数据库获取的化合物发育毒性实验数据为基础， 选取了能够导致新生儿骨骼系统发育异常的1455种化合物经筛选后与1367种无毒化合物组合作为数据集， 尝试建立了一个基于 TensorFlow机器学习平台与神经网络算法思想的 QSAR模型，用以预测化合物对上述终点的发育毒性， 并使用包括正则化约束、随机失活在内的多种算法来维护模型的稳健性与准确度， 最终对比使用不同算法的模型的包括准确度在内的多种模型评估指标。 希望能够对未来建立更加泛用，准确的化合物发育毒性预测模型有借鉴作用。

关键词： TensorFlow，机器学习，神经网络，发育毒性，QSAR

II

{70%：化合物发育毒性预测模型构建}

Abstract

Developmental toxicity is not only an important CAuse of drug failure, but also an important test item for drug safety assessment. The traditional assessment method of drug developmental toxicity is mainly animal experiment, which has the shortcomings of high cost, high consumption and long time. In view of the fact that more and more new compounds are being discovered and created, it is urgent to find a virtual assessment scheme for compounds based on a large number of experimental data and existing analytiCAl algorithms. Due to the wide range of developmental toxicity and excessive endpoints, it is difficult to unify into a single model. Based on the experimental data of1455 compounds with developmental toxicity and1367 non- toxic compounds, which obtained from the material toxicity database of Shanghai Institute of Organic Sciences, a QSAR model based on TensorFlow machine learning platform and neural network algorithm is built to predict the developmental toxicity of compounds to the above endpoints , and a variety of algorithms including regularization and dropout are used to maintain the robustness and accuracy of the model. It is hoped that it will be helpful for build a more general and accurate model for predicting developmental toxicity of compounds in the future.

Keywords: TensorFlow, Machine learning, Neural network, Developmental toxicity, QSAR

II

{70%：化合物发育毒性预测模型构建}

目录

{70%：化合物发育毒性预测模型构建}

1 前言

众所周知，药物在发现和开发过程中，会因为各种各样的原因被放弃，该比例曾经一度高达96%，甚至在可预见的未来，这一比例仍有上升趋势[1]。也由于其高风险，新药开发成为了一项昂贵的工作：{52%：研究一项新药的平均成本早已高达26亿美元。}

而造成药物高失效率的原因有很多，其中一个主要原因就是药品安全问题，每年因为药品安全问题而失效的药物基本都会占全部失效药物的30%[2]。而且，即使一种药物能够暂时的被批准进入市场，它也随时可能由于安全问题而被召回。{41%：因此，从长远利益来看，应尽早的对药品安全性进行广泛的评价。}

{42%：通常，要想评价药物的药品安全性主要有体外（in vitro）与体内（in vivo）实验两种方法。}近年来，开发出了很多体外模型以求降低实验成本[3-4]。然而，这些方法大多仍然是昂贵和耗时的。与实验方法相比，计算方法表现出很大的优势，因为它们是绿色的、快速的、廉价的、准确的，而且最重要的是它们可以在化合物被合成之前完成[5]。

到目前为止，已经建立了许多用于药物安全评估的计算模型，这些模型一般可分为三类：{50%：定性分类模型，定量回归模型与交叉参照模型。}作为药物安全性评价的第一步，我们只需知道一种化合物是有毒的还是无毒的，高毒性的或低毒性的，而不是其确切的毒性值，所以在本实验中，我们可以使用定性分类模型。

如今，许多用以评估药物安全的计算模型是基于机器学习算法的，而传统意义上的化合物结构表示难以作为机器学习算法的输入，{43%：需要提前将传统的化合物结构表示转化为数字化的化合物结构特征。}一般有两种主流的方法可以将化学结构表示为数字化的特征，而只有这些特征才可以通过机器学习方法来处理。其中一种方法是使用分子描述符，它可以根据化学结构、物理化学性质或拓扑性质来计算，其本身是量化的，相对于其他基于分类或标签的方法会具有更强的表现与描述能力。{50%：另一种方法是使用分子指纹，它将分子表示为二进制形式的字符串。}在分子指纹中，会预先定义一个子结构或其他类型模式的列表。如果一个指定的模式出现在一个分子中，二进制字符串中的对应位被设置为“1”，否则它将被设置为“0”。与分子描述符相比，这些二进制特征更易于解释，因为每个二进制位对应着特定的子结构。在本文中，会同时使用分子描述符与分子指纹作为输入，以求统合二者的优点，使模型有更优良的表现。

在原始数据的选取上，本实验将以从上海有机所物质毒性数据库获取的化合物发育毒性实验数据为基础，选取了能够导致新生儿骨骼系统发育异常的1455种化合物经筛选后与1367种无毒化合物组合作为数据集，尝试建立一个基于 TensorFlow机器学习平台与神经网络算法思想的 QSAR模型，用以预测化合物对上述终点的发育毒性，具体的工作流程图如图1-1所示。

{40%：图1-1 化合物发育毒性预测模型的构建、验证部分工作流程图}

1.1 发育毒性

发育毒性是导致药物失效的一个重要原因，也是对药物进行安全性评估的一个重要测试项。传统的对药物发育毒性的评估方法主要是动物实验，具有开销大，耗材多，时间久的缺点。而且，投入大量研发经费的新药因为药物安全问题而退市给制药公司带来的损失是不可估量的。

因此，为了降低药物失效的可能性，从而变相降低新药的开发成本，在新药开发早期对药物进行安全性评估（ 包括对毒性和药物对人体的不良效应的评估）成为了药物开发的重中之重。 {41%：而对化合物进行发育毒性评估则更是进行药物安全性评估的一个重要步骤。}

然而，经典的化合物毒性评估方案有着操作复杂、时间长、耗材大等缺点，以当前的新化合物发现速度， {42%：更是很难完全只使用经典的化合物毒性评估方法。} 相反的，尤其是在早期的化合物筛选中，适当的使用计算化学的思路、工具与方法， 利用化合物的分子结构可以在一定程度上决定其物质性质的特点，将化合物转换成计算机能够识别的， 能表示其结构性质的分子描述符，再利用机器学习方法建立化合物结构与其理化生性质的模型（也被称之为定量构效关系模型）。 从而可以实现对化合物进行高效的理化性质筛选。鉴于越来越多的有药理潜力的崭新化合物在不断被发现、创造，找到一个建立在大量实验数据基础上的化合物虚拟评估方案成了当务之急。

1.1.1 药物的发育毒性研究概述

具有发育毒性的化合物能够影响生物的发育过程，并且在发育过程中会表现出不良特性，具体表现可以大致包括： {54%：生长迟缓，即化合物导致的胚胎发育较正常发育过程迟缓；} {61%：致畸作用，即化合物导致的在幼仔出生时，出现的器官表现形态结构异常，} {45%：在理论上，由于致畸作用而导致的形态结构异常，可以在出生后立即被发现；} {66%：功能不全或功能异常，即化合物导致的幼仔在生化、生理、代谢、免疫、神经活动及行为水平的缺陷或异常，} 这一类表现通常具有隐蔽性，难以在出生后立即被发现，往往需要在出生后一定时间， 等某些功能发育完全后，才表现出来； {63%：胚胎或胎仔致死作用，即化合物在胚胎发育期间对胚胎具有的损害作用，对外具体表现为天然流产或死产、死胎率的增加。}

辨别发育性毒物是一项重要的科学工作。 在发育过程中表现出不良特性的化学物质会干扰发育，导致包括受胎体死亡、生长迟缓和功能性衰退在内的多种生物终点。 {40%：人们暴露在数千种没有发育毒性数据的化学品中。} 例如，高生产量（HPV）的工业化学品，即每年大量进口或生产的工业化学品（虽然通常此类化合物的发育毒性潜力很小）[6]。

对药物进行发育毒性研究，可以更好的保障药物在临床试验上的安全性，也可以为新药申请临床试验提供理论依据与数据支撑， {47%：从而能够在减少对应症状患者的用药风险方面发挥重要作用。}

1.1.2 发育毒性的传统评估方法

{51%：对化合物发育毒性的评估绝非易事。} 这是由于一种化学物质破坏胚胎发生并导致发育缺陷的能力是由许多因素决定的， 不仅仅包括化合物本身固有的化学性质，还会受到包括接触剂量和暴露时间、遗传易感性、 生物利用度、生物转化率和与生物系统的化学相互作用在内的影响。 进一步的，还有学者认为，发育过程中的不良变化可能还取决于管控形态生成的具体的细胞和分子进程， 细胞生长和分化，以及动态系统（发育中的胚胎）对局部或系统扰动的高阶响应（弹性）[7]。 鉴于这种复杂性，对化学物质诱发发育毒性潜力的评估主要依赖于观察哺乳动物体内研究中的顶端表型终点变化。 例如，在美国环境保护机构（ Environmental Protection Agency， EPA）健康效应测试指南OPPTS870与经济合作与 发展指导方针研究组织（ the Organisation for

Economic Co- operation and Development guideline studies）测试编号414中分别 都规定了一系列用老鼠和/或兔子产前发育的研究来评估化合物发育毒性的方法。毫无疑问的，这些测试涵盖了完整且宽广的剂量范围与丰富的物种跨度，然而由于发育毒性本身的特质， {41%：导致为发现待评估化合物可能具有的发育毒性潜力，发育毒性评估试验通常都需要观察一个完整的生命周期，}

{46%：根据 ICH中相关的毒性研究技术指导准则，一个完整的生命周期一般会被划分为六个子阶段，} 简单来讲，涵盖了从前代生物交配前到子代性成熟为止的全部时段[8]。

这一要求直接导致了，若将其作为发育毒性的评估方案考虑，将要需要大量劳动力，花费高昂，并会消耗大量的实验动物。 经查证，由美国国家毒理学项目（ National Toxicology Program）对一种化学物质进行的动物试验通常花费在200万到400万美元之间，而且还需要2到3年的时间才能完成。 此外，在有其他可靠的调查手段的情况下，无论是在全国还是在世界范围内，人们越来越不愿意使用动物研究来测试化学品。 一种替代动物实验的方法是使用定量构效关系（QSAR）推导，这是一种将化学试剂的生物活性与其化学结构联系起来的计算模型[9-12]。

1.2 建模思路与机器学习

{48%：建立计算模型本身是一项综合且复杂的工作，涉及到许多学科的关键领域。}{41%：目前在建立化合物相关的预测模型方面，主要基于的思路是化合物的定量构效关系。}旨在假设化合物化学结构可以在一定程度上决定化合物的相关理化生性质，即可以通过提取化合物的结构特征来预测化合物的实际性质。 而在提取化合物结构特征方面，使用机器学习相关算法是近年来发展较为完备的建模思路。

{48%：基于丰富的机器学习算法可以构建很多种类的计算模型。} 截至此时，已经许多研究者建立了各种可以用于药物安全评估的计算模型，一般可分为三类： 定性分类，定量回归与交叉参照。 然而作为药物安全性评价的第一步，本实验只需知道一种化合物是有毒的还是无毒的， 高毒性的或低毒性的，而不是其确切的毒性值，所以本实验可以选择构建分类模型。

1.2.1 定量构效关系

近年来，合成化学、组合化学、药物化学等学科高速发展，给全世界带来了数以万计的大量全新化合物， 如果还按照经典的化合物评估方案来评估，筛选化合物，将付出难以承受的巨量人力、物力（包括但不仅仅局限于大量的实验动物）以及时间。 因此，找到一个建立在大量实验数据基础上的化合物虚拟评估方案成了当务之急。 在这个背景上，化合物的定量构效关系（Quantitative Structure－Activity Relationship，QSAR）引起了人们的广泛关注。 {58%：化合物的定量构效关系研究指的是通过分析现有活性物质（比如一系列具有相同毒理作用的结构相似的化合物），} {100%：以化合物的理化参数或结构参数等为自变量，} 生物活性为因变量， {85%：用数理统计的方法建立起化合物的化学结构与物化生活性之间的定量关系，} {64%：并通过解释由于分子结构改变所引起化合物理化生参数或结构参数的改变，推测其可能的作用机理，} {82%：然后根据新化合物的结构数据预测其活性或改变现有化合物的结构以提高其活性[13-14]。}

QSAR不仅仅是新药设计和研究的一个重要思路，单就目前而言，QSAR已经成功的被运用于药物设计、风险评估和环境毒物预测等领域。 将其运用在预测化合物毒性上，即研究化合物的定量结构-毒性效应关系（ quantitative structure- toxicity relationship，QSTR）也是一个在新药研究早期根据化合物毒性筛选化合物的重要思路[15-17]。 尤其是在

实验数据相对缺乏的情况下，QSTR也被一些学者认为是在毒理学上最可靠的检测工具之一[18]。QSTR相对于其他毒性预测思想的优势在于，基于 QSTR建立的毒性预测模型不光是能够对未知化合物的毒性预测提供参考， 而且能够在一定程度上对能够影响化合物毒性的理化性质，子结构等加以揭示，为阐明中毒机理奠定理论基础， {68%：为进一步的理论研究提供现实支撑。}

{50%：1.2.2 基于QSAR的化合物发育毒性预测模型建模思路}

{52%：由于 QSAR模型是将化合物的生物活性与其化学结构联系起来的计算模型，因此，在}建模前期的准备工作中就需要收集一定数量化合物的生物活性与化学结构数据。 {40%：其中化学结构的度量通常是化合物的物理和化学属性或结构特征，具体而言会使用到化合物的分子指纹与计算分子描述符。} 而生物活性通常是在实验室实验中测得，就本实验而言则是在动物身上进行实验的数据，表明化合物是否具有生物毒性。

{41%：在建模过程中会有许多不同的技术方案可供选择，单就机器学习算法方面就有神经网络算法、} {47%：判别分析法、 Logistic回归法、偏最小二乘回归法与决策树等非常经典的方案。} 这些技术各有其优缺点， 而往往最佳的方法取决于数据中信号的强度、模型的基本假设的准确性以及需要预测结果输出的类型[19-22]， 实际工程中，当需要选择具体的机器学习算法时，往往会对多种机器学习算法提前进行预评估， 选择在预评估中表现最好的算法建模。

而在模型的具体实现上，也会有很多或免费或开放的 web工具、应用程序和开发工具包可以用来承担模型构造任务， 比较典型的就有 Orange应用程序[23]、 TensorFlow机器学习开发平台等。 本实验的前期模型预评估工作将选择在Orange应用程序上进行，而实际的建模与优化工作将选择在TensorFlow开发平台上使用自制的脚本进行。

1.2.3 预测模型构建平台的选择

上文已经提及到了 Orange应用程序与 TensorFlow机器学习开发平台是本实验所主要使用的两大建模工具， 本节将分别分析二者的特性与在本实验中承担的工作。

Orange本身带有图形界面，具有简单、方便、易上手的特点，然而也有其不足，为了具有简单、方便、易上手的优势， Orange选择性的对用户屏蔽了一些算法底层的可选项，导致其专业性有所下降。 简单的讲，其在某些指定算法上，定制、修改能力很差，以神经网络算法为例，在 Orange上基于神经网络算法的 {47%：模型只有包括隐含层节点数、隐含层数目、激活函数、迭代次数等寥寥无几的参数可供修改。} 不可质疑的，单就默认的模型参数而言， Orange本身已经给出了一个非常有参考价值的训练结果， 然而当需要用具体且更加复杂的算法去辅助优化模型、维护模型稳健或者降低过拟合程度时， Orange就会稍显后继乏力。 基于 Orange的上述特性，所以本实验选择了在建模初期使用 Orange来评估各算法表现（图1-2）， 从而具体的选择使用的算法，而在最终建模阶段则使用专业性更强的 TensorFlow开发工具包“定制”模型。

图1-2 Orange用户界面与使用Orange初步评估各算法表现

TensorFlow机器学习开发平台是谷歌机器学习团队开发的用于计算的代码库，他的特点是开源，免费，灵活。 在实际工程中， TensorFlow不会为请求的变量立即分配值，相反的，TensorFlow会选择将申请的变量视为数据流图（ data flow graph）中的节点， 节点之间拓扑有序，拓扑顺序的先后取决于计算的顺序，这一特性非常适合用于描述神经网络或其他类似算法的前向传播过程， 在运行时，也可以按照节点间的拓扑顺序来优化反向传播。{53%：而数据流图中的边就代表了节点间的数值运算，} TensorFlow这一名字就暗喻着代表

数据的张量（ Tensor）在数据流图中以运算的形式流动（ Flow）， 其自带的TensorBoard框架更是可以直接把对应工程的数据流图可视化输出（图1-3）。 TensorFlow本身为Python及其他多种编程语言提供了功能完备且强大的API，本实验选择以Python脚本的形式调用TensorFlow API从而实现机器学习算法。

图1-3 用TensorBoard输出的本实验的数据流图

1.2.4 神经网络算法

说到神经网络，无可避免的会联想到生物学中的概念，即生物学传统意义上的， {50%：指由生物的大脑神经元，细胞，触点等结构相互连接天然形成的，用于保证生物体对事物认知，} 反应，思考乃至学习的网络状结构。 而事实上，机器学习算法中非常经典的神经网络算法也是基于此联想的一个抽象实现。

概念上来讲，最简单的神经网络算法应该是基于一个神经元模型的分类器，类比于生物学上的同名概念， {42%：神经网络算法中的神经元模型指的是一个包含输入、输出与函数的简单模型。} 图2-2是一个简单的神经元模型，应该注意的是该模型输入中包含常数项1，其与对应的权值系数 b相乘代表截距， 由此保证了该神经元模型所划分的区间在坐标空间内不一定经过原点。

图1-4 单个神经元模型

{46%：实际上，稍加复杂的神经网络算法往往会使用到多个神经元模型，} {42%：较之单个神经元模型，同时使用多个神经元模型可以使得输出不再局限于单个的数，} 而可以复杂到向量，这保证了其作用于现实复杂问题的可行性，同时，多个神经元模型对分类区间也能具有更精确的划分。

图1-5 多个神经元模型

而显然的，多个神经元模型输出的结果本身也可以作为输入参与下一个神经网络模型的运算，从而得到更复杂更贴合实际的结果。 {41%：这一行为可以称之为为神经网络添加了一个隐层，特别的，对于含有多个隐层的，} {42%：特别复杂的神经网络，我们可以称其为深度学习网络，深度学习网络较之普通的神经网络由于具有更多的待训练变量，} 训练起来更为困难，耗时更大，但是对于复杂的函数的拟合能力也会更强。

图1-6 深度学习网络

1.2.5 模型评估方法

本实验旨在构建化合物发育毒性预测模型，由于单一化合物只会被划分为有毒与无毒两类，所以本实验所构建的模型在分类上属于传统的单标签二分类预测模型，而对于传统的单标签二分类或多分类模型， {66%：大多数性能指标是基于真阳性（ TP）、真阴性（ TN）、假阳性（ FP）、假阴性（ FN）的数目来计算的。} 比较典型的性能指标有准确性，可以用来表示预测模型在整体数据集内的预测能力； {44%：灵敏性，可以用来表示预测模型对阳性样本的预测精度；} {48%：特异性，可以用来表示预测模型对阴性样本的预测精度。} {40%：上述三种性能指标基于真阳性、真阴性、假阳性与假阴性的具体的计算公式见式（1-1）（1-2）（1-3）。}

（1-1）

（1-2）

（1-3）

1.3 化合物数据

建立化合物性质预测模型的一大关键难点就是如何获取高质量的化合物数据。 化合物数据的质量是决定预测模型性能好坏一个非常重要的因素。 在目前，如果想要一次性大批量的获取化合物数据，从化合物数据库中获取是一个非常好的选择。 {41%：本实验的化合物数据主要来源于上海有机所物质毒性数据库，由于其本身不带有批量下载的网络接口，} 于是选择使用自制的脚本从其网页中批量获取页面信息后，对页面信息进行处理得到。

1.3.1 化合物数据库

在互联网上，对于大量的有展示、查询、添加乃至修改需求的信息，建立对应数据库是一个非常好的解决方案。 在如今时代，随着化合物发现所产生的信息量甚至已经大大超过了过去几年，乃至几十年的总和。 因此，大量的数据库相关技术与工程被选择运用到了化学信息学领域，从而满足研究者对于大量化合物信息的处理需求。

数据库是计算机领域的概念，是独立在应用程序之外的有组织的数据集合，其理念在于要能够在保存尽可能多的数据的前提下， 支持对数据的快速增加、修改、删除与查询，提供支持多用户共享的程序接口，同时要尽可能的具有更小的数据冗余度。

而化合物数据库是从数据库中进一步引申出的概念，在实际研究中，其一般指的 是用于存储大量化合物理化性质、结构信息与数据来源资料的数据库的可视化网页接口。 目前，网上有大量定义明确的化合物数据库，比较典型的有 TOXNET毒性综合数据库， 它汇总了包括 ToxLine和 ChemIDplus在内的多个毒性数据库的化合物毒性数据[24]； OECD建立的eChemPortal，其主要提供化合物包括物理化学性质和毒性在内的化学信息，eChemPortal中也包含了许多化合物数据库，如ACToR和HSDB[25]； admetSAR是由我校开发的网络服务器，其中也包含着毒性数据。 这些数据库极大地便利了通过机器学习方法建立计算模型[26]。

1.3.2 上海有机所物质毒性数据库

{80%：上海有机所物质毒性数据库是上海有机所化学专业数据库系统的一部分。} 主要包含的是化合物的毒性数据（具体而言，包括化合物的毒性测试实验数据、毒性测试参考文献以及各种化合物毒性标准）以及结构数据[27]， 本实验主要使用的是结构数据，由于期望构建的是二分类预测模型，所以对于毒性数据只以有毒与无毒做简单划分， 而不会使用如接触途径、测试物种、测试类型等其他信息。

另外的由于在上海有机所物质毒性数据库中，具有发育毒性潜力的化合物数量多，范围广，影响终点过多， 表现形式多样（共有99种与化合物发育毒性相关的表现形式），显然是难以一概而论的。 故本实验主要采集的是上海有机所物质毒性数据库中表示会造成新生儿骨骼系统发育异常的具有发育毒性潜力的化合物与基本等量的无毒化合物。

1.3.3 数据描述工具

从上海有机所物质毒性数据库中获取的化合物信息经过筛选、比对后，得到以SMILES式形式存储的化合物结构信息。 然而计算机显然是无法识别以文本形式存储的SMILES式信息的，需要用现有的算法将其数字化以供计算机读取作为模型输入。 目前为止较主流的方法有两种，一是处理成能表征化合物化学结构、物理化学性质或拓扑性质的分子描述符， 二是根

据化合物所具有的子结构信息进行多标签分类，所得到的结构用一串二进制位表示，存在该标签对应的二进制位为“1”，不存在该标签对应的二进制位为“0”。上述处理步骤都可以在PaDEL-Descriptor中完成。

PaDEL-Descriptor是一个免费的用以计算分子指纹和分子描述符的开源软件，使用Java语言开发，同时具有图形化的用户界面与命令行接口，能在包括Windows、Linux、MacOS在内的多种主流操作系统下作业，支持超过90种不同的分子文件格式作为输入，当前最新版本共支持计算797种分子描述符（其中663种为一维或二维的{41%：分子描述符，另外134种为三维描述符）与10种分子指纹[28]。}

因为PaDEL-Descriptor简单、快捷、准确的特性，本实验选择它作为由SMILES式获取分子指纹与分子描述符的工具。

{55%：2 化合物发育毒性预测模型的构建与优化}

2.1 材料和方法

{46%：传统的建立预测模型的一般过程大致包括四个步骤：}数据收集、数据描述、模型构建和模型评估。每个步骤都有自己的要求，以求保证模型的可靠性和准确性，下文将按照上述的逻辑顺序叙述本次实验的基本过程与思路。

2.1.1 数据收集

本实验的毒性化合物数据全部来自于上海有机所物质毒性数据库。在数据收集阶段，笔者使用自制的脚本从目标网站爬取了合计1455种具有发育毒性潜力的化合物。{46%：原始数据包含丰富的化合物信息，具体包括中文名称、别名、英文名称、SRN号、CAS号、分子式、规范SMILES式、物质结构与文献信息等，}经过字符串正则处理后只以文本形式保留后续使用的规范SMILES式信息，舍去其余部分，由于部分化合物具有非有机物、为难以计算的盐、SMILES式不符合规范等客观问题而必须对所得原始数据进行进一步筛选，否则后续步骤难以进行。在筛选工作结束后，合计剩余1367种具有发育毒性潜力的化合物，尽量按照一比一的原则，从实验室留存的无毒化合物中随机选取了基本等量的无毒化合物，二者全部是以SMILES式文本形式保存的文件，作为数据描述步骤的输入等待进一步处理。

2.1.2 数据描述

本步骤的意义在于要将计算机不方便解析的化合物SMILES式信息，用特定的算法处理成计算机方便解析的分子指纹与分子描述符。主要基于的工具是PaDEL-Descriptor。

在利用收集数据计算分子指纹前，笔者参考了早期建立化合物对虹鳟的水生毒性预测模型时不同分子指纹与机器学习算法的表现（表2.1）

表2.1 在预测化合物水生毒性时所有分子指纹与机器学习算法对中表现最优秀的五组

指纹名

机器学习方法

AUC

CA

MACCSFingerprinter

Neural Network

0.87

0.846

MACCSFingerprinter

Random Forest

0.866

0.799

Fingerprinter

Neural Network

0.844

0.834

Fingerprinter

Random Forest

0.842

0.824

ExtendedFingerprinter

Neural Network

0.841

0.824

显然的MACCSFingerprinter指纹在所有分子指纹中表现最佳，其次为Fingerprinter指纹，而在所有的机器学习方法中神经网络算法的表现略优于包括随机森林在内的其他算法。 但是在进一步对分子指纹的评估中，发现在神经网络算法中若使用 MACCSFingerprinter作为输入，会具有输入分量过小的缺点， 换句话说 MACCSFingerprinter用以描绘化合物的二进制位在长度上显著的短于其他分子指纹， MACCSFingerprinter标准仅提供166个二进制位长度用以描述化合物（表2.2）， 也许是这一特点导致了在具有更少神经元的较简单的神经网络模型中，使用 MACCSFingerprinter描述数据会让模型具有更快的学习速率。 然而，当进一步使用稍显复杂的神经网络算法评估建模可行性时，输入分量过小的 缺点直接导致了使用 MACCSFingerprinter分子指纹作为输入的模型非常容易产生过拟合问题。 而在初步评估中表现仅稍次于MACCSFingerprinter分子指纹的Fingerprinter分子指纹因为其具有更多的数据分量，而在过拟合问题上的表现反而会优于MACCSFingerprinter分子指纹。 综合评估后，本文选择使用Fingerprinter标准来生成化合物数据的分子指纹。

表2.2 不同分子指纹的数据分量

指纹名	数据分量数
MACCSFingerprinter	166
Fingerprinter	1024
ExtendedFingerprinter	1024
PubchemFingerprinter	881
SubstructureFingerprinter	307

在利用收集数据计算分子描述符时，一个不可避免的问题是分子描述符的选取，最新版本的 PaDEL- Descriptor一共支持计算797种分子描述符， 要从中筛选出一系列最为有效的分子描述符无疑需要极高的专业素养与大量的时间，于是笔者退而求其次， 以是否有描述能力作为筛选标准。 具体而言，由于本实验的原始数据来源于上海有机所物质毒性数据库，而上海有机所物质毒性数据库的数据则是从不同的论文收集而来的， 这就导致了化合物的种类非常的复杂，很可能某一项分子描述符的计算标准并不适用于全部的化合物， 那么这一项分子描述符应该要被认为不适用于本次模型的构建，予以排除。 笔者使用自制的脚本读取了所有化合物从 PaDEL- Descriptor计算出的797种分子描述符， {41%：选择性删除了大部分化合物表现一致或部分化合物表现异常的分子描述符，} 最终合计删除了70项分子描述符，换句话说最后得到的数据在分子描述符上一共具有729个输入分量。 另一个必须要处理的问题是分子描述符的取值区间问题，因为同样作为输入的分子指纹在每一个输入分量上的取值只可能为0或1， 而计算得来的分子描述符的取值范围在理论上可能包含整个实数域， 这就需要在分子描述符的每一个分量上将数值映射到0到1区间， 即对数值进行归一化处理，从而保证在实际建模过程中权重参数的取值差距不会过大，这步处理使用的是自制的脚本， 参考如下公式进行的数值映射。

（2-1）

其中，表示该分量原始的取值，表示在所有化合物的该分量中数值最大的取值， 对应的表示在所有化合物的该分量中数值最小的取值，表示经过映射后得到的新值。

至此，化合物数据的描述工作已基本完成，每个化合物分别使用分子指纹与分子描述符描述，分子指纹使用 Fingerprinter标准， 具有1024个输入分量，分子描述符使用筛选得到的729个一维或二维描述符，具有729个输入分量。

2.1.3 损失函数的定义

众所周知，基于机器学习算法的模型的训练过程是基于损失函数的，良好的损失函数对一个模型的性能有重要意义。 {48%：本实验旨在预测化合物是否具有发育毒性潜力，} 故构建的模型应该为经典的二分类模型， 而针对分类模型普遍表现良好的损失函数有很大一部分是基于交叉熵的， 故本实验也将以交叉熵为基础构建损失函数。

交叉熵的计算是基于信息量的，通俗的讲，信息量指的是预测发生几率为的事件在实际发生时， {40%：会对模型产生的影响，换句话说，是模型从事件中获得的信息。} {44%：假若事件发生的预测概率极高，那么其实际发生后对模型的影响应该是微乎其微的，} 原因是模型已经准确的预测了其的发生，假若事件发生的预测概率极小，那么在其实际发生后模型应该进行修正， 原因是模型没有准确的预测其的发生。 计算事件信息量的公式如下。

（2-2）

其中为事件的信息量，为事件发生的概率。

而交叉熵则表示为所有事件信息量的加权平均和，公式如下。

（2-3）

显然，特别的，对于类似于本实验的二分类问题，由于化合物只有被划分为有毒与无毒两种可能， 所以，当令模型预测化合物有毒的概率为时，交叉熵可以化简为以下形式。

（2-4）

在建模的初期阶段，损失函数直接等于交叉熵，而当使用了其他的辅助优化算法后，损失函数会加上对应优化算法的补正。

（2-5）

2.1.4 初步评估与建模方法选择

在本实验进行初期，笔者只打算使用分子指纹来描述数据，在此基础上，针对使用Fingerprinter标准形式的分子指纹， {41%：使用了神经网络算法、 kNN、 SVM、随机森林与朴素贝叶斯五种机器学习方法构建了简单的模型，} {47%：并对模型使用Orange进行了初步评估，评估结果见表2-3。}

表2.3 使用Fingerprinter分子指纹描述的数据在不同机器学习算法下构建的模型的表现排名

机器学习方法

AUC

CA

Neural Network

0.744

0.673

Random Forest

0.737

0.667

kNN

0.728

0.675

Naive Bayes

0.647

0.622

SVM

0.399

0.411

综上，可以认为神经网络算法，随机森林，kNN算法在模型构建中的表现相对其他被评估的机器学习算法更加出色，因此本实验决定使用神经网络算法构建发育毒性预测模型，初步的模型示意图如图2-1。

{57%：图2-1 在建模初期使用的神经网络示意图}

从该神经网络开始，本实验开始使用TensorFlow机器学习平台，利用自己编写的脚本建模。{45%：该模型的输入层有1024个输入节点，对应1024个分子指纹分量，隐含层包含500个神经元，} {44%：输出层有2个节点，根据2个输出节点的相对大小决定预测结果与可信度。}该模型的表现较差，虽然损失函数的值在训练集上确实得到了收敛，且AUC值也随着迭代次数而增长，可以认为输入参数与输出相关，模型学习具有成果（图2-2），但是，在不使用其他辅助算法的情况下，迭代5000轮后AUC值也只有约0.65，收敛速度较慢，过拟合问题严重，以至于测试集数据在迭代后期呈发散趋势（图2-3）。

图2-2 待评估的神经网络模型确实具有学习性能（AUC值经对数处理）

{51%：图2-3 待评估的神经网络模型过拟合问题严重}

2.1.5 模型的假设与初步构建

由于上述模型实际运行效果不佳，笔者决定尝试性的在输入层加入化合物的分子描述符信息，经过测试，加入分子描述符信息作为输入的神经网络模型的性能（以AUC值为参考）普遍较只使用分子指纹信息的神经网络模型提高了2个百分点以上。

加入化合物分子描述符信息的建模方法有很多，其中非常典型的有以下两种（图2-4与图2-5）。

图2-4 可用方案一

图2-5 可用方案二

方案一与二的主要区别是代表分子描述符部分的输入节点位于神经网络的深度。 在实际评估时主要做了三个方面的考虑： 一是，在性能方面，容易计算得知二者在隐藏层节点数目不变的前提下所需要训练的参数量差别是十分微小的， 反应在工程上则表现为二者迭代相同代数所花费的时间是近乎相同的，在此基础上方案二在 AUC值上的表现略微强于方案一； 二是，考虑到分子指纹与分子描述符是两套完全不同的描述化合物性质的规则，将二者同时作为输入得到的化合物特征很可能是难以， 甚至于不能解释的，而拆分在两个层级则可以先进行单独的分析； 三是，分子指纹作为一套完整的规则，是难以筛选的，可能存在所有化合物在某一系列分量上的表现一致或近似相同， 然而分子描述符是经过自制脚本筛选的，无效或数值非法的分量已经被提前筛除， 其特征应该较分子指纹明显。

{41%：基于上述评估的基础，笔者在已有的神经网络模型基础上稍做了修改。} 使用了形如方案二的一个稍微不同于传统神经网络结构的模型。

该模型同时将化合物的分子指纹与分子描述符作为输入层，其中分子描述符不参与对第一层隐含层的运算， {44%：在利用分子指纹计算出第一层隐含层的所有节点后，将分子描述符部分输入与第一层隐含层拼接，} 参与下一层的运算，该模型前向传播过程的符号形式表述见下式。

（2-6）

{40%：其中， 表示隐含层节点矩阵， 表示激活函数， 表示权重矩阵， 表示截距矩阵， 表示分子指纹部分输入， 表示分子描述符部分输入。}

在这里，笔者假设了一个前提： 分子指纹单纯是基于多标签的分类手法描述分子结构性质的符号，而分子描述符是量化的能够更精确的对分子的理化性质加以描述的符号。{40%：由此，当同时将分子指纹与分子描述符作为输入时，分子描述符有比分子指纹更强的描述能力，} 反映在建模上，代表分子描述符输入的节点在网络中应该要比代表分子指纹输入的节点拥有更深的深度， 从而使其能够更直观的表达。

至此，化合物发育毒性的预测模型已初步构建完成，虽然构建出的模型仍有许多问题，但可以通过添加多种辅助优化算法提升模型的表现，在下一节中会分别介绍在本次建模过程中使用到的辅助优化算法与对应算法给模型带来的性能提升。

2.1.6 模型的优化

在本模型的优化过程中，根据在前期评估里模型表现出来的过拟合严重与迭代后期收敛不明显问题， 本实验分别使用了正则化方法（ regularization）与随机失活（ dropout）进行优化。

正则化是一个在处理过拟合问题上表现良好的优化算法，其基本思想是对过于复杂的拟合模型在计算损失函数时给予惩罚， 致使模型往更加合理、普适的方向学习，而非过度的去拟合训练集趋势，基本假设是在输入参数标准化的前提下， 更复杂的模型表现为会具有数值上更大的权重值。 本模型之所以可以使用正则化方法进行优化的一个大前提是在描述数据时对数据进行了归一化处理， {43%：导致对模型具有相似程度描述能力的节点权重大小相似，因此，可以用以下公式来刻画模型复杂度。}

（2-7）

其中为边上的权重，为自定的正则化参数。

如此，便可用正则化损失修正损失函数，从而提高学习性能，损失函数式（2-5）则可以改写为如下形式。

（2-8）

这样当神经网络后向传播时，不光会向使得结果更贴合训练集标签的方向学习，而且会考虑到整体模型的复杂程度，避免使用过于复杂的模型去拟合数据，乃至于会放弃拟合过分偏离整体趋势的数据。

图2-6 使用正则化优化前后模型的学习表现(右图因数据跨度过大，做了对数处理)

除此之外，对于模型在迭代后期收敛不明显缺点，笔者使用了随机失活方法优化。

随机失活的基本思想与假设基本同于正则化思想。在思想上，都是通过对过于复杂的拟合模型给予惩罚，致使模型往更加合理、普适的方向学习。在基本假设上，都是假设对标准化的输入数据，更复杂的拟合模型会具有更大的权重，从而可以通过约束权重值来对模型进行简化。与正则化约束的区别在于，随机失活的实现部分一般会发生在前向传播过程，而正则化约束实际起效会发生在后向传播过程，即对损失函数进行梯度下降时。

随机失活算法的基本算法是指先指定一个保持率（Keep Rate），然后在指定使用随机失活算法的层级遍历其所有节点，每个节点保存的概率为保持率，若没有被保存，则节点被视为删除，本次迭代先删除该节点与所有与其相连的边后再进行训练。

在本实验中，单纯的只对最后一层隐含层使用了随机失活算法，保持率为0.75。实际算法使用前后模型的具体表现如下图。

图2-7 使用随机失活优化前后模型的表现(左图因数据跨度过大，做了对数处理)

从图2-7中，可以明显的看出随机失活算法对于本次实验预测模型有着一定程度上的优化，主要体现在以下几个方面。

{40%：首先，最明显的是在使用了随机失活算法后模型的AUC值得到了约2个百分点的提升，}最终的五次随机划分测试集与训练集的实验后，得到的平均AUC值约为0.68。其次，模型的稳健度得到了一定程度的提升，体现在模型迭代训练的后期，使用随机失活算法的模型测试集的Loss值会具有更小幅度的波动，收敛效果更好。随后，另一个显而易见的优点是使用随机失活后模型的收敛速度得到了提升，明显的快于没有使用随机失活的模型。

2.1.7 模型评价

在建立了本实验所求的预测模型后，本实验将考虑使用准确性CA值（Cluster Accuracy）与Roc曲线下方面积大小AUC值（Area Under Curve）对其进行评估。

{45%：准确性CA值是基于真阳性（TP）、真阴性（TN）、假阳性（FP）、假阴性（FN）的数目来计算的。}其本质上是在衡量预测正确的化合物在全部化合物中所占据的比重。{52%：利用真阳性、真阴性、假阳性、假阴性来表示则形如式（1-1）。}是一种较简单的计算方法，在具体建模过程中，本实验将通过自制的脚本语句计算预测模型的CA值。

由于输入数据阴性与阳性分布的比值往往不是严格的1： 1，完全使用CA值来衡量模型表现优劣有时会出现偏差，稍微极端的讲，将一份阴阳数据比为7： 3的输入数据作为测试集输入一个预测输出永远为阴性的二分类分类器，得到的CA值应该也会近似为70%。 {45%：为了解决这一缺点，可以使用AUC值，即ROC曲线下方面积大小来辅助评估模型的性能。} 在TensorFlow机器学习平台中，针对给定模型计算其AUC值可以通过几条简单的语句实现，这也是TensorFlow平台的一大优点。 AUC值在统计学意义上是一项非常可靠的评估参数。{43%： AUC值越大，表明该分类器的分类效果越好，最大值为1.0，表明模型有近乎完美的分类性能，} {44%：特别的，当 AUC值为0.5时，表明模型的分类效果仅仅等同于随机分类器[29]。}

在以上理论依据下，具体模型评估方法为， 将无毒与有毒化合物的分子指纹与分子描述符数据分别按照7比3的比例划分出训练集与测试集， 重复5次得到5组待评估的实验数据，将这5组实验数据分别作为输入，用训练集数据训练得到神经网络模型， 在测试集数据上得到期望的评估数据（包括 AUC值与 CA值在内），最后结果取5次的平均值。

2.2 结果与讨论

2.2.1 化学空间多样性分析

一般而言，当我们想刻画化合物数据集之间的离散程度时，可以通过刻画化合物化学空间来实现。 具体方法为，首先对化合物组中的每一个化合物选择其一个或多个分子描述符，用选择的一系列分子描述符刻画化合物性质， {41%：并将这些分子描述符的具体数值对映射为多维空间内的一个个点，通过判断代表具体化合物的点在多维空间内距离的接近程度，}来判断化合物之间的相似性。 而上述的多维空间，也被称之为化学空间（ Chemical Space），如果所刻画的化合物数据集在化学空间中分布相对广泛， 我们就可以认为该化合物数据集在化学空间上具有多样性。

利用机器学习算法构建的模型的性能在很大程度上与数据集的化学空间多样性挂钩。一般的，如果选取的化学空间过小，会导致模型的学习被局限在所选的小空间内， {47%：对于总体化合物集的普适性较差，从而致使模型的运用受到各种限制。} 在本实验中，评估输入化合物数据集的化学空间分布主要通过化合物数据中的分子量(MW)与计算得到的脂水分配系数(ALogP)， 在评估时用自制的脚本读取输入化合物数据集，分开的对具有发育毒性的化合物与不具有发育毒性的化合物进行了评估。 具体结果如下图所示。

图2-8 具有发育毒性的化合物（红色三角形）与不具有发育毒性的化合物（蓝色圆形）利用分子量与脂水分配系数得到的化学空间分布

从图2-8中可以分析得知，无论是具有发育毒性的化合物还是不具有发育毒性的化合物，在利用分子量与脂水分配系数建立的化学空间中分布都较为广泛，且重叠度较高， 可以说明本实验的数据来源代表性较强，具有足够广的化学空间分布。

2.2.2 模型评价

根据上述2.1节中的模型优化与评估步骤，我们得到了一个化合物发育毒性预测模型的最佳实践， 即同时使用化合物分子指纹与分子描述符信息作为输入，其中化合物分子描述符信息不参与第一隐含层的计算， 加上对模型整体使用正则化约束修正，对第二隐含层使用随机失活算法修正。 对于模型的评估，则使用AUC值与CA值，按照上述章节中相关步骤进行。表2.4中展示了使用不同模型优化算法与否对模型性能造成的影响。 图2-9与图2-10则展示了最佳实践的预测模型性能随迭代训练次数增长的变化。

表2.4 不同的神经网络模型的评价结果

神经网络模型

AUC

CA

最佳实践

0.678

0.689

不使用正则化约束

0.665

0.682

不使用随机失活

0.666

0.683

图2-9 最佳实践的预测模型测试集与训练集Loss值与迭代训练次数的关系图（因数据跨度过大，做了对数处理）

图2-10 最佳实践的预测模型AUC值和训练集Loss值与迭代训练次数的关系图

从上述图表中可以看出，本实验建立的预测模型具有一定程度上的学习性能，基于最佳实践算法建立的模型在收敛速度与收敛效果上都有较好表现，过拟合问题得到明显改善。而在具体的预测性能上同样表现良好，可以在一定程度上正确预测化合物的发育毒性潜力。

3 小结

本实验主要工作旨在基于化合物的分子指纹信息与分子描述符信息来预测化合物的发育毒性潜力，即提取包含在化合物分子结构与理化性质中与化合物发育毒性潜力相关的统计学特征，{47%：从而对化合物具有的发育毒性潜力进行预测。} {45%：在本实验中，为了更好的去预测化合物的发育毒性潜力，} 笔者基于机器学习理论中的神经网络算法与TensorFlow机器学习开发平台构建了具有学习能力的化合物发育毒性预测模型，并辅以包括正则化约束、随机失活在内的辅助优化算法来优化模型性能。

首先，利用自制的爬虫脚本从上海有机所物质毒性数据库中批量爬取了大量具有发育毒性潜力化合物的数据资料，经过分析与处理后以 SMILES式的形式保存。 随后，选取了能够导致新生儿骨骼系统发育异常的1455种化合物经筛选后与1367种无毒化合物组合作为数据集。

其次，利用PaDEL-Descriptor软件对上述化合物数据集中的化合物结构数据进行描述与计算，分别生成对应化合物的分子指纹信息与分子描述符信息。 并基于 Orange应用程序使

用包括神经网络算法、随机森林算法、朴素贝叶斯、 kNN、 SVM在内的多种机器学习算法与多种分子指纹构建了多个分类模型， 并对进行预评估，选取了表现最好的神经网络算法与 FingerPrinter分子指纹进行后续进一步建模操作。

随后，基于 TensorFlow机器学习平台与 Python编程语言实现神经网络算法，并基于神经网络算法构建模型， 再基于模型的表现不断调整模型参数，同时使用包括正则化约束与随机失活在内的辅助优化算法尝试对模型的表现进行优化， 最终得到表现最佳的模型实现并对最佳实现的模型性能进行评估。

最佳实现的模型在五次随机划分测试集与训练集的实验中表现出的性能较好， ROC曲线下面积的平均值为0.678， {51%：准确度的平均值为0.689，具有一定的分类能力。} 无论在测试集还是训练集中损失函数的值都可以随迭代次数增长而快速下降，并迅速收敛， 表明模型具有很好的学习能力，且过拟合问题在使用了辅助优化算法后得到了显著改善。

综上所述，本实验为在临床前评估药物发育毒性潜力提供了一个行之有效的解决方案，对未来建立更加泛用，准确的化合物发育毒性预测模型有借鉴作用。 更长远的，还可以通过对训练好的模型进行持久化处理，分析其权重分配，来为进一步研究化合物发育毒性的机制奠定现实基础。

参考文献

[1] Paul S M， Mytelka D S， Dunwiddie C T， et al. How to improve R D productivity: the pharmaceutical industry’s grand challenge[J]. Nature Reviews Drug Discovery， 2010， 9(3): 203.

[2] Giri S， Bader A. A low-cost， high-quality new drug discovery process using patient-derived induced pluripotent stem cells[J]. Drug discovery today， 2015， 20(1): 37-49.

[3] Huh D， Hamilton G A， Ingber D E. From 3D cell culture to organs-on-chips[J]. Trends in cell biology， 2011， 21(12): 745-754.

[4] Huh D， Matthews B D， Mammoto A， et al. Reconstituting organ-level lung functions on a chip[J]. Science， 2010， 328(5986): 1662-1668.

[5] Segall M D， Barber C. Addressing toxicity risk when designing and selecting compounds in early drug discovery[J]. Drug discovery today， 2014， 19(5): 688-693.

[6] EDF T I. The Continuing Absence of Basic Health Testing for Top-Selling Chemicals in the United States[J]. Environmental Defense Fund， New York， 1997.

[7] National Research Council. Scientific frontiers in developmental toxicology and risk assessment[M]. National Academies Press， 2000.

[8] 周宗灿. 毒理学教程[M]. 北京大学医学出版社, 2006: 238-240.

[9] Bearden A P, Schultz T W. Structure‐activity relationships for Pimephales and Tetrahymena: A mechanism of action approach[J]. Environmental Toxicology and Chemistry, 1997, 16(6): 1311-1317.

[10] Bradbury S P. Quantitative structure-activity relationships and ecological risk assessment: an overview of predictive aquatic toxicology research[J]. Toxicology Letters, 1995, 79(1-3): 229-237.

[11] Hansch C, Fujita T. p-σ-π Analysis. A method for the correlation of biological activity and chemical structure[J]. Journal of the American Chemical Society, 1964, 86(8): 1616-1626.

[12] Kubinyi H. QSAR: Hansch analysis and related approaches[M]. VcH, 1993.

[13] 王连生, 韩朔睽等著.有机物定量结构-活性相关[M].北京: 中国环境科学出版社, 1993.

[14] Martin Y C. Quantitative drug design: a critical introduction[M]. CRC Press, 2010.

[15] Isayev O, Rasulev B, Gorb L, et al. Structure-toxicity relationships of nitroaromatic compounds[J]. Molecular Diversity, 2006, 10(2): 233-245.

[16] Roy K, Ghosh G. QSTR with extended topochemical atom (ETA) indices. VI. Acute toxicity of benzene derivatives to tadpoles (Rana japonica)[J]. Journal of Molecular Modeling, 2006, 12(3): 306-316.

[17] Zmuidinavicius D, Japertas P, Petrauskas A, et al. Progress in toxinformatics: the challenge of predicting acute toxicity[J]. Current Topics in Medicinal Chemistry, 2003, 3(11): 1301.

[18] 陈雅菲, 钟儒刚.QSAR建模方法及其应用于化学物质毒性预测的研究进展[J].轻工科技, 2017, 33(02): 29-31.

[19] Cronin M T D, Aptula A O, Dearden J C, et al. Structure-based classification of antibacterial activity[J]. Journal of Chemical Information and Computer Sciences, 2002, 42(4): 869-878.

[20] Devillers J, Chezeau A, Thybaud E. PLS-QSAR of the adult and developmental toxicity of chemicals to Hydra attenuata[J]. SAR and QSAR in Environmental Research, 2002,

13(7-8): 705-712.

[21] Devillers J, Chezeau A, Thybaud E, et al. QSAR modeling of the adult and developmental toxicity of glycols, glycol ethers and xylenes to hydra attenuata[J]. SAR and QSAR in Environmental Research, 2002, 13(5): 555-566.

[22] Kaiser K L E. The use of neural networks in QSARs for acute aquatic toxicological endpoints[J]. Journal of Molecular Structure: THEOCHEM, 2003, 622(1-2): 85-95.

[23] Demšar J, Curk T, Erjavec A, et al. Orange: data mining toolbox in Python[J]. The Journal of Machine Learning Research, 2013, 14(1): 2349-2353.

[24] Fowler S, Schnall J G. TOXNET: information on toxicology and environmental health[J]. AJN The American Journal of Nursing, 2014, 114(2): 61-63.

[25] Fonger G C, Hakkinen P, Jordan S, et al. The National Library of Medicine's (NLM) Hazardous Substances Data Bank (HSDB): background, recent enhancements and future plans[J]. Toxicology, 2014, 325: 209-216.

[26] Cheng F, Li W, Zhou Y, et al. admetSAR: a comprehensive source and free tool for assessment of chemical ADMET properties[J]. 2012.

[27] Shanghai Institute of Organic Chemistry of CAS. Chemistry Database[DB/OL]. http: //www.organchem.csdb.cn.[1978-2019]

[28] Yap C W. PaDEL - descriptor: An open source software to calculate molecular descriptors and fingerprints[J]. Journal of computational chemistry, 2011, 32(7): 1466-1474.

[29] Fawcett, T. An introduction to ROC analysis [J]. Pattern Recognition Letters, 2006, 27(8): 861 - 874.

II

{70%：化合物发育毒性预测模型构建}

致谢

本科四年生活很快就要结束了，早先就有前辈警示过我，大学四年说短不短，但说长也绝对不会长，要尽力做好每一件事，不要毕业时才追悔莫及。我很感激那位前辈，大学四年里，我不敢说我尽力做了每一件事，不敢说这四年没有任何遗憾，但是我敢说我这四年没有虚度，可能回首过去，有些事情我可以兼顾，有些时候我可以做得更好，而我却选择了任凭我的才能在别处生根发芽，但是这四年给我带来的财富，是前所未有的，不光是在知识、才干上，甚至更多的是在一些难以明说的地方。

我对大学前几年的记忆可能和我身边的人有着些许的不同，没有什么雪月风花的浪漫，没有什么团体活动的欢乐，　　我印象最深的还是奉贤夜里　CD教冷冽的风和回去时朗马两块一根热乎的烤肠。　　在那段日子里，我在一些地方应该算得上是一事无成，但在另外一些地方倒还勉强算是稍有成就，　　虽然很多时候由于自身的秉性和外界的因素，这些所谓的“成就”大多难以善终。　　感谢华理通项目的成员，虽然我们后面选择了分道扬镳，但是我仍向往着那些我们通宵探讨需求与实现的时光。　　感谢和我一起打过比赛的队友，做完课题伴着朝阳回宿舍的那个清晨会永远是我记忆深处的瑰宝。　　感谢不知名的wenzhe先生，我名义上的第一个老板和饭友，当日的豪言壮语我已兑现，也祝你的创业之路一路顺风，莫安平庸。上述诸位，我们之间有过分歧。　　有过摩擦，有过争吵，现在和往后可能都难再联络，但是与诸位的共事都很愉快，　　与诸位的相遇成就了如今的我，仅能于此，聊表谢意。

在本科的最后一年，我非常幸运的能在唐赟老师实验室进行毕业论文的相关工作，{42%：非常幸运的遇见了博学多识的王学长、楼学长和李学长，}　三位学长不仅在学术上对我光照有加，而且在人生规划上对我也有莫大的启迪，也间接的促使我决定在医药大数据方向去研读深造，　　感恩这半年多来在529实验室度过的时光，让我明白了我是如此的幸运。

{43%：除此之外，还要感谢一直在背后默默支持我的父母，感谢你们的养育之恩，感谢四年来教导过我的所有老师，}　{41%：感谢你们的悉心教导，感谢一直对我关照有加的舍友、同学，感谢你们对我的无私帮助。}

值此良机，得以略舒胸怀，不胜感激，顺颂诸君，万事顺意。

肖力铭

2019年5月25日

检测报告由PaperPass文献相似度检测系统生成
Copyright 2007-2019 PaperPass