# Comparison of multi-task convolutional neural network (MT-CNN) and a few other methods for toxicity prediction

# 摘要

毒性分析和预测对人类健康和环境保护极为重要。现有的计算方法由各种各样的描述符和回归器构成,这使得它们的性能分析变得困难。例如,深度神经网络（DNN）在很多场合都是成功的方法,就像一个黑匣子,并且没有概念上或物理上优雅的理解。目前的工作构建了一套基于电荷,表面积和自由能物理模型的微观描述符,以评估多任务卷积神经网络（MT-CNN）体系结构和其他一些方法的性能,其中包括随机森林RF）和梯度提升决策树（GBDT）。还比较了卷积神经网络（CNN）和非卷积深度神经网络（DNN）算法。四种基准毒性数据集用于评估各种方法。大量的数值研究表明,目前的MT-CNN架构能够胜过最先进的方法。

**关键词：毒性终点,多任务学习,深度神经网络。**

# 简介

毒性是衡量化学物质对生物体有不利影响的程度。这些不利影响称为毒性终点,可以通过其对特定目标的影响进行定量或定性测量。大多数毒性测试旨在抗衡人体免受化学物质引起的有害影响,并且传统上以体内或体外方式进行。尽管如此,这样的实验通常非常耗时且费用高昂,甚至在动物实验中引起道德问题。因此,计算机辅助方法或计算机模拟方法已被开发出来,以提高预测效率而不会牺牲太多的准确性。定量结构活性关系（QSAR）方法是最常用和最常用的方法之一。基本的QASR假设是类似的分子具有相似的活性。因此,通过研究化学结构与生物活性之间的关系,可以预测新分子的活性,而无需实际进行实验室实验。由于高通量筛选（HTS）技术的发展,研究人员现在能够快速和定期处理数千种化学物质,从而为药物发现以及培训和验证QSAR模型提供大量数据。

开发QSAR模型的主要关注点是选择适当的化学特征（描述符）。一方面,如果分子特征经过充分研究并具有化学或物理意义,QSAR模型易于解释。然而,使用过多或不足的描述符会使QSAR模型的解释变得非常困难。在统计上重要的特征可能不是生物可解释的情况下,一般特征选择过程可能导致问题。另一方面,准确的QSAR模型的开发需要足够多的培训化学品和合理的外部验证组合。如果有限的实验数据可用,QSAR模型可能会过度拟合数据并且不能给出令人满意的预测结果。另外,涉及大量描述符的QSAR模型可能经常遇到特殊感兴趣的化学品,这些化学品超出了QSAR的适用领域,因为不能构建一个或多个描述符。在这种情况下,必须开发新模型来做出准确的预测。

生成QSAR模型的算法有几种类型：基于线性回归和线性判别分析的线性模型1 ,包括最近邻算法的非线性模型2,3 支持向量机1,4,5和随机森林6。这些方法由于其统计特性而具有优点和缺点7。例如,线性模型忽略不同特征之间的相关性,而最近邻的方法很大程度上取决于描述符的选择。为了克服这些困难,已经引入了更精确和更先进的机器学习方法。多任务学习（MTL）8被部分地提出来处理上述数据稀疏性问题。MTL的想法是从相关任务中学习所谓的“归纳偏差”,以使用相同的表示来提高准确性。换句话说,MTL旨在从多个任务中学习共享和一般化的特征表示。因此,它很有希望学习跨不同任务的共享信息并提供更好的预测。事实上,MTL策略为生物信息学带来了新的见解,因为来自相关测定的化合物可以共享不同特征水平的特征,如果数据集很小,这非常有用。成功的应用包括序列生物学中的剪接位点和MHC-I结合预测9,基因表达分析和系统生物学10。

最近，深度学习（DL）11,12特别是卷积神经网络（CNN）已经成为在信号和信息处理领域呈现广泛最新结果的强大范例，如语音识别13,14和自然语言处理.15,16深度学习架构本质上是基于人工神经网络的。深度神经网络（DNN）模型和非DNN模型的主要区别在于，DNN模型由大量的图层和神经元组成，使得在每个图层中构建抽象特征并将更高级别的信息传递到下一层成为可能。例如，在人脸识别中，DNN会从原始未处理的图像中检测斑点和边缘，并进一步揭示眼睛和鼻子等物体。因此，DNN可以轻松应用于QSAR框架。给定分子特征表示，DNN能够从原始未处理数据中捕获抽象化学特征。 Dahl等人17使用人工神经网络来预测化合物同时进行多种测定的活性。默克的Ma等人18讨论了单个和多任务神经网络参数的优化，并提高了随机森林预测精度。 Mayr等[19]在Tox21数据挑战中完成了另一项相关工作，他们在包含2百万个相关数据点的1,280个生物目标数据集上训练了一个多任务网络。 Ramsundar等[20]进一步研究了数据量和任务数量之间的关系，以及多任务神经网络的预测能力和可转移性之间的关系。

然而，通常没有DNN结果的物理解释。 事实上，DNN方法提供的概念理解很少。 相比之下，基于物理的方法在第一原则方面提供概念上的优雅，即基本的物理定律。 此外，DNN对设计师和用户都表现神秘。 一旦创建，甚至设计师都无法解释神经网络在每一层学到了什么，以及最终如何预测感兴趣的属性。 因此，DNN的设计通常基于试验和错误，并且大部分受限于可访问的计算机能力。 此外，对于给定的方法，特征或描述符对预测准确度有重大影响。 DNN和非DNN方法的大多数早期比较不仅基于不同的方法，而且基于不同的特征。在相同的数据特征下比较不同方法的性能是十分有意义的。

这项工作的目的是基于已建立的物理模型设计一套通用的微观特征，以便DL模型包括多任务CNN（MT-CNN），单任务CNN（ST-CNN）和和两种非DL方法可以在平等的基础上比较。这些物理特征导致了我们早期工作中出色的蛋白质-配体结合亲和力预测能力21为了从原始数据中提取物理特征，我们首先为所有化合物建立了三维（3D）分子模型。 利用源于量子力学的合适的力场来获得原子电荷。 原子表面积是基于适当的几何模型计算的。 最后，通过泊松模型计算原子静电溶剂化自由能。 这些物理特征及其统计量使我们能够为毒性预测构建大量有效的DL和非DL模型。

本文结构如下。 第二节致力于方法和算法。 详细描述特征生成和基于元素的特征构造。 给出了经典集合方法的简要概述和深度学习架构的详细描述。 讨论了单任务和多任务深度神经网络体系结构。 在第三节中，我们给出了用于开发QSAR模型的数据集的描述，然后是模型参数化和评估标准。 第四节介绍了我们对四种基准毒性终点的预测结果，其次是第五节，其中强调了模型优化。 最后，本文在第六部分结束语中进行总结。

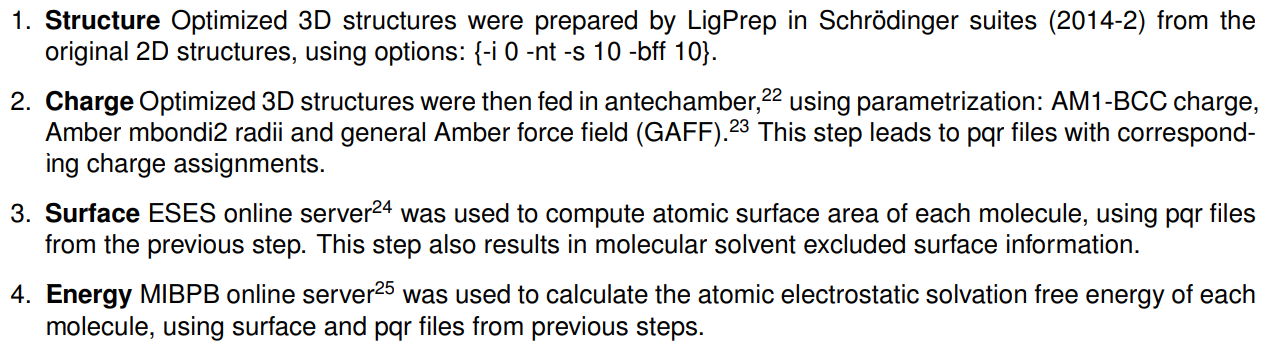
# 方法和算法

在本节中，我们首先提供特征生成的详细描述。 然后，给出了集成方法（随机森林和梯度增强决策树），深度神经网络，单任务学习和多任务学习的概述。 重点将放在定量毒性终点预测的多任务深度卷积神经网络的优点，以及如何为网络体系结构选择适当的参数。 还提供了我们的架构和培训程序的详细说明。

## 2.1物理特征生成

在这项工作中，我们有兴趣构建一套基于高级微观特征的物理模型来描述分子毒性。 这套功能应该适用于不同的机器学习方法，包括深度学习和非深度学习，以及单任务和多任务。 有人可能会争辩说，提供具有高级别物理特征给深层CNN是不公平的。 然而，值得注意的是，与图像分类问题不同，对于化学和生物数据集，基于原始图像的数据输入不适用于CNN，因为化学信息不能直接由原始图像表示。 因此，在将CNN应用于化学和生物数据时，总是需要一定程度的数据预处理

为了使我们的特征生成对于所有化合物都是可行和稳健的，我们考虑三种类型的基本物理信息，即“由量子力学或分子力场计算的原子电荷，针对溶剂排除表面定义计算的原子表面积，以及从泊松模型估计的原子静电溶剂化自由能 。 为了获得这些信息，我们首先为每个分子构建优化的三维结构。 然后计算上述的原子属性。 我们的特征生成过程可以分为几个步骤：



根据上述程序获得物理特征。具体而言，我们的特征来自步骤2，步骤3和步骤4.与其他传统的QSAR方法不同，该方法结合了2D宏观描述符，如ECFP4指纹或某些原子类型的数量，我们强调通过生成的原子特征来自物理模型。然而，原子特征与原子位置有关，不能直接在不同分子间进行比较。另外分子通常具有不同数量的原子。因此，对于大多数机器学习算法来说，原子级信息的直接比较是不可行的。我们构建元素要素，其中将相同元素类型的原子属性相加在一起。因此，我们有元素电荷，表面积和静电溶剂化自由能。此外，对于元素特征，我们进一步考虑它们的统计，即最大值，最小值，平均值和方差，这引起了另外四个特征。此外，从绝对值获得的原子电荷和原子静电溶剂化自由能的增强是重要的物理量。因此，在总和之前，取原子电荷和静电溶剂化自由能的绝对值，为每种类型的元素产生元素电荷强度特征和静电溶剂化强度特征。对于每个元素强度特征，其统计量，即最大值，最小值，平均值和方差都会导致另外四个特征。最后，为了原始三类物理量之间的对称性，我们还创建了五个元素表面的强度特征。这种对称性考虑对于我们的卷积算法是重要的，其冗余对于所有机器学习方法来说基本上是无害的。因此，每种类型的元素总共有30个元素级物理特征。为了使我们的物理模型适用于各种化学品，我们只考虑10种不同的常见元素类型，即H，C，N，O，F，P，S，Cl，Br和I我们的元素选择。

在这项工作中，我们也以类似的方式从原子电荷，表面区域和静电溶剂化自由能构建无坐标分子特征。 这一考虑产生了30个分子特征。 我们将这种元素特异性和分子特异性信息组织到11个通道中，这类似于RBG彩色图像表示中的RGB通道。 因此，给定分子的最终输入特征矩阵具有11×30的维数。为了在CNN中实现数值稳定性和更快的收敛（更大的学习速率），输入矩阵沿特征维度缩放。 即，对于每个特征，将其归一化为零均值和单位方差。 得到的输入矩阵很容易适用于CNN。

或者，我们将所有上述信息组织成具有330个组件的一维特征向量，这很容易适用于整体方法，如随机森林和梯度增强决策树，以及非卷积DNN。

## 2.2 集成方法

为了探索不同机器学习方法的优点和缺点，我们考虑了两种流行的ensamble方法，即随机森林（RF）和梯度提升决策树（GBDT）。这些方法已被广泛用于解决QSAR预测问题，以及溶剂化和蛋白质 - 配体结合自由能预测[21,26,27]。它们自然处理描述符之间的相关性，并且通常不需要复杂的特征选择程序。最重要的是，RF和GBDT对参数基本上不敏感。因此，我们选择这两种机器学习方法作为比较基准。

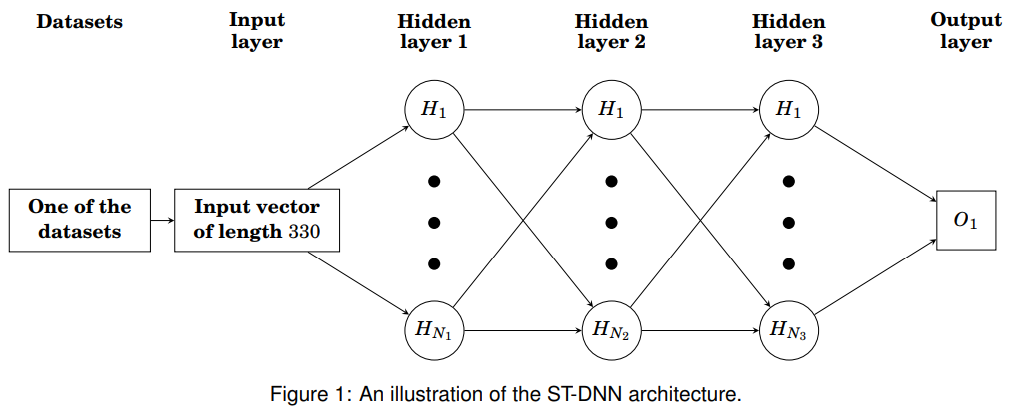
我们已经使用scikit-learn软件包（版本0.13.1）实施了这些回归器.28估计器的数量设置为2000，学习率针对GBDT方法进行了优化。对于每一组，50次运行（具有不同的随机状态）已经完成，并且平均结果在本工作中被报告。第2.1节中讨论的具有330个组件的1D特征向量用作RF和GBDT的输入数据。

## 单任务深度学习算法

神经网络充当将输入特征向量映射到输出向量的变换。 它本质上模拟了生物大脑解决许多由轴突连接的神经元单元的问题。 一个典型的浅层神经网络由几层与神经元组成，并使用反向传播来更新每层的权重。 然而，它不能构造分层特征，因此在揭示更多抽象特性方面存在缺陷，这使得难以建模复杂的非线性关系。

与浅层网络相比，单任务深度学习算法具有更广泛和更深的架构 它由每层中更多层和更多神经元组成，并揭示不同层次输入特征的各个方面。 针对每个单独的预测任务定义单任务深度学习算法，并且仅从特定任务中学习数据。

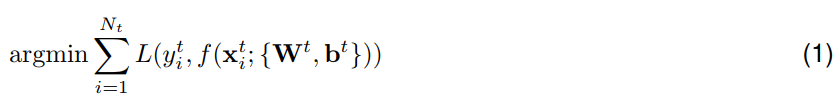
在这项工作中，我们构建了两种单任务（ST）深度学习算法，即ST-CNN和ST-DNN。 在第II.A节中讨论的具有330个分量的1D特征向量用作ST-DNN的输入数据，而尺寸为11×30的特征矩阵用于ST-CNN。 这种ST-DNN的表示可以在图1中找到，其中Ni（i = 1,2,3）表示第i个隐藏层上的神经元的数量



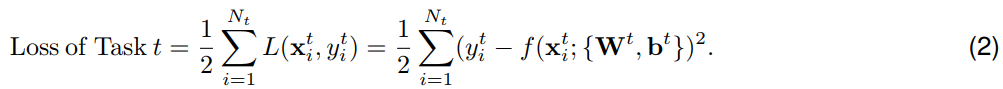
## 多任务学习算法

多任务学习是一种机器学习技术，在最近的Merck和Tox21预测挑战中取得了成功。 MT学习的主要优势在于同时学习多个任务，并利用共同点以及跨不同任务的差异。 MT学习的另一个优点是，具有不完全统计分布的小数据集以建立精确的预测模型通常可以从具有更完整统计分布的相对较大的数据集中显着受益

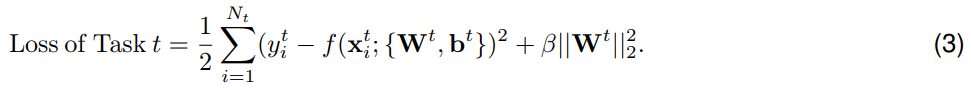
假设我们共有T个任务，第t个任务的训练数据记为（xti，yt i）Nt i = 1，其中t = 1，...，T，i = 1，...， Nt，Nt是第t个任务的样本数量，xti和yti分别是特征向量和目标值。 MTL的目标是同时最小化



其中f是由权重向量Wt和偏倚项bt参数化的函数，L是损失函数。 典型的回归成本函数是均方误差，因此第t个任务的损失可以定义为：



为了避免过度拟合问题，我们通过在权重向量上增加一个正则化项来定制上述损失函数（2），为第t个任务提供改进的损失函数：



|| ·|| 表示L2范数，β表示惩罚常数。

在毒性预测的背景下，MT学习将联合学习不同的毒性终点，并有可能提高多重回归任务的总体表现。 更具体地说，假定具有不同测量的毒性终点的不同小分子包含不同的物理或化学特征是合理的，而描述符例如某种化学结构的发生可导致类似的毒性特性

## 多任务卷积神经网络（MT-CNN）架构

为了利用MT学习，我们提出了两种MT深度学习架构 - 一种使用卷积滤波器（即MT-CNN），而另一种则不使用（即，MT-DNN）。 在这两种方法中，前几个共享层设计用于捕获不同毒性的共享表示，最后一层用于处理差异。 所提出的神经网络本质上是非线性回归器，其对输入数据执行重复的线性和非线性变换。

**我们的卷积体系结构的主要思想是通过化学元素构建不同的通道（类似于图像表示中的RGB通道），然后沿着由电荷，表面积和静电抗震动能量构成的物理特征维度对训练数据进行卷积。不同元素通道的特征提取阶段包含一个1D卷积层，一个最大池层和三个更完全连接的层。卷积层包含60个尺寸为3和步幅为1的滤波器，其权重不是空间共享的。这意味着在输入向量的每个位置都应用了一组不同的过滤器。选择定义为φ（x）= max（0，x）的校正函数φ（x）作为卷积层的激活函数。然后我们在上一层的顶部执行1D大小为2的非重叠最大池化操作。三个更加完全连接的各种宽度的密集层被进一步添加到池化层。请注意，以上所有图层都在不同的任务中共享。最后，为了捕捉不同任务之间的差异，我们将每个任务的单独预测器附加到前一层。对于回归任务，因为目标值是连续的，所以线性函数（f（x）= x）就足够了。具体而言，我们的MT-CNN体系结构可以表示为：**

1）输入层[11×30]具有30个分子通道和10个元素通道。

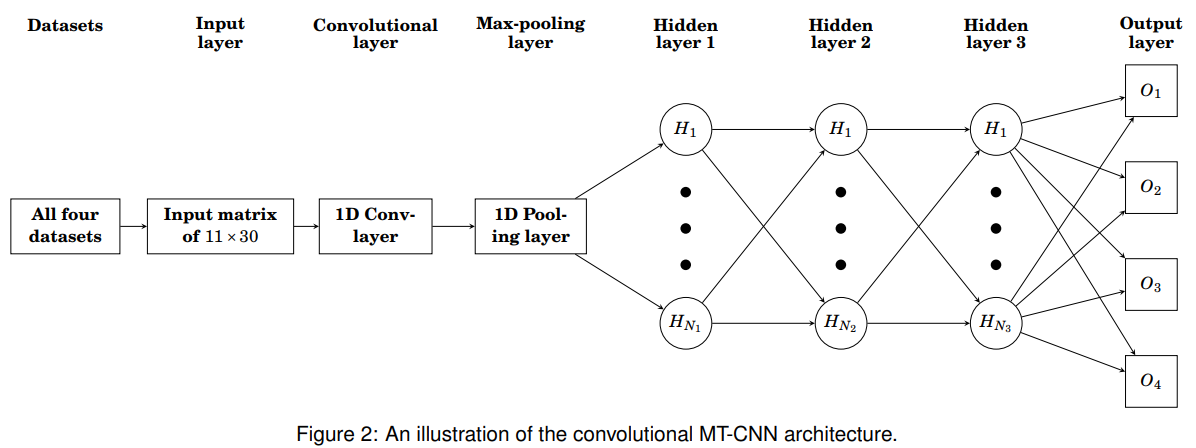
2）卷积层计算局部连接到输入特征向量的神经元的输出。 该操作以[11×28×60]的形式产生3D张量，其中28 = 30-3 + 1，其中3是每个滤波器的窗口大小，1是沿着滤波器移动滤波器的步长 特征维度。 这里“60”是可以优化的过滤器的数量。

3）整流层应用max（0，x）的元素式激活，同时保持值不变。

4）Maxpooling层沿特征维度执行下采样操作，导致[11×14×60]的维度的值。

5）完全连接的图层输出每个回归任务的预测。 在每一层中，每个神经元都连接到前一层中的所有神经元

6）图2显示了我们的卷积MT-CNN的详细体系结构。 这里Ni（i = 1,2,3）表示第i个隐藏层上的神经元数量。 卷积和非卷积体系结构之间的区别在于，非卷积版本具有一维特征向量，第2.1部分讨论了330个分量作为输入数据，它没有卷积层或最大池化层。



## 模型训练

训练过程基于使用动量小批量随机下降（SGD）的反向传播（BP）算法。 具体地说，设L是目标损失函数，是要更新的参数，v是速度函数，是当前最小批次的平均目标函数梯度，更新规则如下：



m是梯度下降动量，l是学习率。

（3）中使用均匀分布抽样的格洛特权重来初始化权重和偏差项。 为了减少当前模型的预测偏差，每个训练集中的分子随机洗牌，然后分成大小为200的小批量，然后用于更新参数。 当所有的小批量都被遍历时，一个训练“纪元”就完成了。

# 3.数据、参数和评估

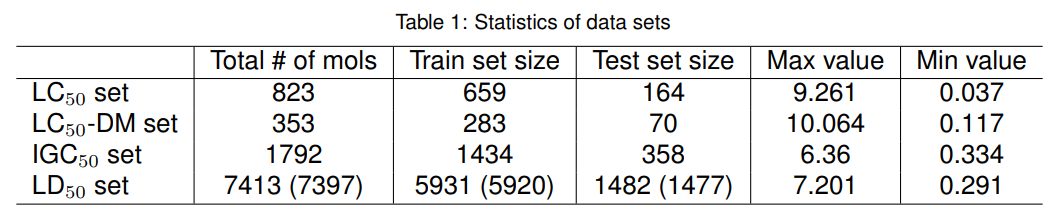
本部分致力于详细描述数据集，模型参数化和结果评估标准

## 3.1 数据集

（LC50集合），（LC50-DM集合），（IGC50集合）和（LD50集合）

本工作中使用的最终组合与那些经过预处理并用于开发（毒性评估软件工具（TEST））的那些相同.29 TEST的开发旨在使用各种QSAR方法评估化学毒性，使用非常方便，因为它不 需要任何外部程序，它遵循通用的QSAR工作流程 - 它首先计算797个二维分子描述符，然后利用这些预先计算的分子描述符预测给定目标的毒性

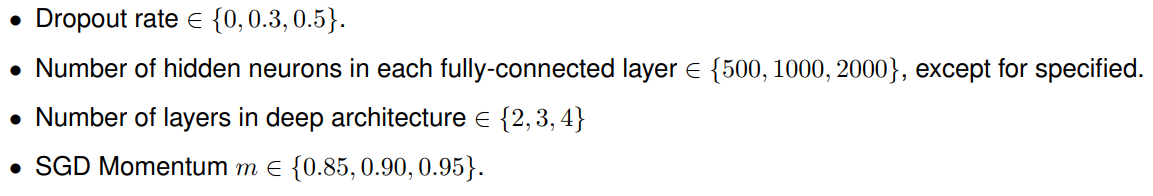
所有分子均为2D sdf格式或SMILE格式，其相应的毒性终点可在TEST网站上找到。 应该指出的是，我们对预测定量毒性终点特别感兴趣，所以不使用含有定性终点或物理性质的其他数据集。 而且，不同的毒性终点具有不同的单位。 LC50，LC50-DM，IGC50终点的单位为 - log10（T mol / L），其中T代表相应的终点。 对于LD50设置，单位是 - log10（LD50 mol / kg）。 虽然单位不完全相同，但应该指出，由于端点的大小顺序相同，因此不需要额外尝试重新缩放值。 这四个数据集的大小也不同，从数百到数千不等，这实质上挑战了我们方法的稳健性。 下面列出了四个数据集的详细统计表：

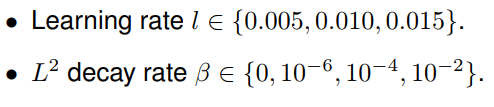


圆括号内的数字表示我们在本文中用于开发模型的分子的实际数量。 请注意，对于前三个数据集（即LC50，LC50-DM和IGC50组），适当包括所有分子。 然而，对于LD50设定，由于力场失效，涉及元素As的一些分子被剔除。 显然，TEST工具遇到了类似的问题，因为来自两个TEST模型的结果不可用，并且来自各种TEST模型的覆盖率（预测分子的分数）始终小于1。 我们模型的整体覆盖率总是高于TEST模型的覆盖率，这表明我们模型的更广泛的适用领域。

## 3.2 参数选择

由于大量的可调参数，优化所有可能的参数组合是非常耗时的。 因此，我们调整参数在一个合理的范围内，随后评估他们的表现。 下面列出了我们尝试优化的参数及其范围：





第五节给出了更详细的参数选择讨论

## 3.3 评价标准

Golbraikh等32提出了一个确定QSAR模型是否具有预测能力的协议。

