

基于机器学习算法的食品污染物 神经毒性预测模型建立

周悦¹, 李潇岚¹, 程薇¹, 冯艳¹, 王艳^{1,2*}

(1. 上海交通大学医学院公共卫生学院, 上海 200025)(2. 上海交通大学医学院附属第九人民医院, 上海 200011)

摘要: 基于机器学习算法建立分类预测模型, 研究常见食品中化学性污染物的理化结构与其神经毒性间关联。通过查阅文献建立化合物数据库, 纳入包含影响神经分化成熟、影响神经元迁移/空间定向等各类神经毒性机制化合物 57 种, 无神经毒性化合物 50 种。运用 R、SPSS 软件, 使用随机森林 (Random Forests, RF)、类神经网络 (Artificial Neural Network, ANN)、支持向量机 (Support Vector Machine, SVM) 等机器学习算法筛选分子描述符并构建分类模型, 预测化合物神经毒性。结果显示随机森林算法模型综合表现最佳, 十折交叉验证准确率 70.24%, 训练集、测试集预测准确率分别达 95.51% 和 83.33%, 曲线下面积分别达 0.99 和 0.85, 是个较为理想的算法。本研究基于机器学习算法建立的分类模型可通过化合物的分子描述符准确预测化合物的神经毒性。在多种机器学习算法中, 基于随机森林算法建立的预测模型表现最优。分子描述符重要性结果显示, 化合物神经毒性主要与其质量加权 Burden 矩阵最大特征值有关。

关键词: 食品污染物; 神经毒性; 机器学习; 毒性预测

文章编号: 1673-9078(2022)04-216-223

DOI: 10.13982/j.mfst.1673-9078.2022.4.0727

Neurotoxicity Prediction Model of Food Contaminants Based on Machine Learning

ZHOU Yue¹, LI Xiaolan¹, CHENG Wei¹, FENG Yan¹, WANG Yan^{1,2*}

(1. School of Public Health, Shanghai Jiao Tong University School of Medicine, Shanghai 200025, China)

(2. The Ninth People's Hospital of Shanghai Jiao Tong University School of Medicine, Shanghai 200011, China)

Abstract: Prediction models based on machine learning algorithms are established to predict the neurotoxicity of chemical pollutants in food. Database which includes fifty-seven neurotoxic compounds and fifty non-neurotoxic compounds was established through the published paper. By utilizing the R and SPSS software, the random forest (Random Forests, RF), neural network (Artificial Neural Network, ANN), support vector machine (Support Vector Machine, SVM) and other algorithms were used to build the classification models applying the molecular descriptors. The random forest algorithm represented the best performance in aspects of total accuracy and feasibility, illustrating total accuracy of training set and test set was 95.51% and 83.33%, respectively. The area under the curves of training set and test set were 0.99 and 0.85, respectively. The accuracy of 10-fold cross-validation was 70.24%. In this study, the prediction models established on the basis of machine learning algorithms and chemical informatics can accurately distinguish the neurotoxicity compounds from non-neurotoxicity compounds. Our result suggested that among models, the one constructed with the random forest algorithm performs better and the highest eigenvalue from Burden matrix as a molecular descriptor contributes dominantly to the classification of chemicals with neurotoxic potential.

Key words: food contaminant; neurotoxicity; machine learning algorithms; toxicity prediction

引文格式:

周悦, 李潇岚, 程薇, 等. 基于机器学习算法的食品污染物神经毒性预测模型建立[J]. 现代食品科技, 2022, 38(4): 216-223, +306

ZHOU Yue, LI Xiaolan, CHENG Wei, et al. Neurotoxicity prediction model of food contaminants based on machine learning [J]. Modern Food Science and Technology, 2022, 38(4): 216-223, +306

收稿日期: 2021-07-13

基金项目: 国家重点研发计划重点专项(2018YFC1602405); 上海市科学技术委员会科研计划项目(19140901200)

作者简介: 周悦(1997-), 女, 硕士研究生, 研究方向: 毒理学替代法、替代毒理学, E-mail: yjszhouyue@sjtu.edu.cn

通讯作者: 王艳(1966-), 女, 博士研究生, 研究员, 研究方向: 预测毒理学、环境污染的健康风险评估, E-mail: wangyan@shsmu.edu.cn

神经毒性是外源性化合物引起的神经系统结构或功能的损害,而经口摄入是众多可导致神经毒性中毒症状的化学品重要的暴露方式之一^[1,2]。如今随着食品工业的发展,食品种类增多,食品中添加剂使用的增加,许多食品化学污染物如重金属、食品添加剂、农药兽药残留物等,均可诱导发生氧化应激、神经细胞发育受损、神经递质含量变化等各种神经系统损伤,严重威胁人体健康^[3-5]。因此,神经毒性评估成为了食品污染物暴露毒性评价必不可少的环节。传统的神经毒性评估主要通过动物实验测试进行。但动物实验周期长且成本高昂,在高通量筛选和人类中枢神经系统的外推性方面存在一定的局限性^[6]。因此,迫切需要开发替代方法和工具,来评估化学品的神经毒性。

定量构效关系(Quantitative Structure Activity Relationship, QSAR)是目前常用的毒性预测模型之一,利用化合物的物理化学、几何等结构信息,预测其可能的生物活性^[7]。分子描述符分为2D和3D描述符。虽然3D描述符在模拟分子受体配体对接,处理空间化学方面可能有更大的优势,但2D描述符具有无需估算生物活性构象,计算成本低,数据处理快的优点^[8],同时预测结果依然优异^[9,10]。QSAR已被成功运用于经口毒性判断和内分泌干扰物筛选^[11],但针对神经毒性,相关研究多聚焦于同源化合物,纳入同属数据集建立,如多氯联苯类(Polychlorinated Biphenyls, PCBs)^[12]、芳基氨基脲类化合物等^[13],通过同源化合物间的相似结构构建线性模型,分析特定取代基或分子结构与其神经毒性间关联^[14]。因食品中的化学性污染物来源广泛、种类繁多,化合物的结构特征多样,神经毒性症状各异且毒性作用靶点各不相同,化合物可通过影响神经分化成熟、神经元迁移/空间定向、突触形成/可塑性、神经递质、神经营养功能、神经免疫系统、神经内分泌、离子稳态干扰等功能诱发神经毒性^[15]。因此,传统QSAR线性模型不适用于食品化学污染物分子结构与神经毒性关联的建立。机器学习算法中的监督学习,在学习框架的基础上半自动化创建分析模型,并通过交叉验证和迭代改进以细化和优化预测模型的准确性,具有“多指标、多关联、多交叉”的特点,尤其适用于大规模样本的归纳研究^[16]。

考虑到该模型的适用范围和高通量筛选的需要,本研究将选则2D描述符用于神经毒性预测模型的建立。通过支持向量机、随机森林、神经网络等机器学习算法,构建食品化学污染物神经毒性分类预测模型。据此阐明影响化合物神经毒性的主要分子结构因素。进一步了解化合物性质,为后续化合物的毒性预警及

风险评估开展提供参考依据。

1 材料与方法

1.1 数据库的建立与神经毒性资料获取

从现有文献^[17,18]、政府通告的违法食品添加剂或新检出物、生活常见药物和食品安全污染物中选择了107种化合物组建数据库。化合物组成包括20种杀虫剂(18.69%)、14种阻燃剂(13.08%)、16种多环芳烃(14.95%)、19种食品包装化工原材料(17.76%)、6种食品添加剂或污染物(5.61%)及32种药物(29.91%)。选择现有人类、非人灵长类动物、哺乳动物实验以及体外细胞实验中,有两篇及以上出版物表明其具有神经毒性的化合物做阳性判定。选择现有文献中无神经毒性相关实验,实验结果矛盾或仅单细胞系体外实验阳性的化合物做阴性判定^[19]。根据以上判定原则,最终获得包含各类神经毒性机制,如阻碍神经元发育成熟、干扰神经递质传递、影响突触形成等化合物57种,及明确无神经毒性化合物50种。数据库于2020年5月完成建立,于2021年5月对化合物文献检索结果进行更新。

1.2 数据集的建立

按8:2比例随机划分训练集及测试集,训练集用于模型构建,测试集用于模型验证^[20,21]。

1.3 描述符的获取与特征选择

分子描述符是化合物物理化学性质的量化表征,其将分子中的结构化学信息,转换成特定的指标或某些标准化实验的结果,其选择一定程度上决定了模型的优劣^[22,23]。2D分子描述符一般包括:(1)组成描述符(如分子量、卤素类型);(2)拓扑描述符(如Wiener指数、Randic指数);(3)电子描述符(如极性指数、Hammett常数);(4)几何参数描述符(如分子体积、溶剂接触面积)等^[24,25]。本研究从pubchem网站获取化合物的2D分子结构信息,导入MOLD2(2.0)软件后计算得到777个分子描述符。考虑到疏水性在经口毒性预测、药物筛选方面的重要作用,使用Episuite(version 4.1 US EPA 2012)及VEGA(core version 1.2.1 IRFMN)软件,计算常见于相关毒性预测模型中的正辛醇-水分配系数(Octanol-Water Partition Coefficient, Kow)纳入研究^[26,27],并针对以上778个计算变量开展特征选择。

特征选择主要由以下五个步骤组成。(1)删去恒定、半恒定(相同取值比例>0.85)描述符后剩余472

个计算变量。(2) 二元 logistics 回归完成描述符的初步筛选。(3) 结合已有文献经验, 补充纳入现有研究中广泛运用于预测药物吸收、代谢、毒性的相关描述符, 包括: 分子量、柔性可旋转键数、Kow、拓扑极性表面积等^[24,28]。(4) 将 472 个描述符和训练集 89 种化合物带入五种机器学习模型并分析计算变量重要性, 纳入重要性排名前五位者。(5) 删去重复后将以上合计 52 种计算变量纳入随机森林模型进行拟合, 采用递归特征消除法进行重复构建和计算变量筛选, 每次构建删除其中最不重要的变量^[29]。最终纳入如表 1 所示 5 种描述符, 主要覆盖化合物可旋转键数、电拓扑、分子极性相关性质, 包括 1 个原子数和键数描述符、2 个理化性质描述符、2 个拓扑描述符。根据“Rule-of-Thumb”规则, 2D-QSAR 模型中训练集(N)与描述符(M)的比值(N/M)应不小于 5, 本研究所得最优模型的 N/M 为 17.80>5, 满足此项规则, 表明描述符的选择数量适当^[30]。

表 1 构建 QSAR 模型所使用描述符

Table 1 The descriptors used in QSAR model

描述符编号	描述符类型	描述符名称
D015	键数描述符	可旋转键数比例
D142	Balaban 系数	Balaban 均方顶点距离指数
D152	理化性质描述符	碳-sp3 的平均原子极化率
D175	Wiener 系数	Wiener 最大路径指数
D564	理化性质描述符	质量加权 Burden 矩阵最大特征值

注: 该描述符类型及名称信息来源于文献^[24]。

1.4 模型的构建

共采用 5 种机器学习算法^[25,28,31]。

由于数据集的各个特征取值范围存在较大差异, 对描述符取值进行标准化处理:

$$Z = \frac{X - \mu}{\sigma} \quad (1)$$

式中:

X——化合物描述符参数数值;

μ ——参数平均值;

σ ——参数标准差。

1.4.1 贝叶斯网络 (Bayesian Network, BN)

贝叶斯算法适用于基于定量研究的不确定性推理, 利用概率统计对未知样品进行分类。变量间的关系强度由两个概率组成, 一为条件概率, 可被父节点个数和局部概率影响, 二为节点固定先验概率。贝叶

斯网络要求各个变量在除父节点以外节点间条件独立, 现有研究中可见于方剂配伍规律等药性研究^[32]。使用 Rstudio (version 4.0.2 下同) “e1071”软件包中的“naiveBayes”函数实现。

1.4.2 随机森林 (Random Forests, RF)

随机森林算法因其易于理解、执行力高等优势常见于药物安全性评价等药物分析领域^[32]。研究中使用随机森林算法的基分类器为 cart 算法, 通过重采样方法生成随机自助训练集, 未被纳入训练集样本称为袋外样本。构建多颗树后选择最佳模型输出。使用 Rstudio 软件“randomforest”软件包进行模型构建和优化, “pROC”软件包进行图形绘制, “varImpPlot”函数进行模型参数重要性可视化。最终构建模型节点中二叉树变量个数 (mtry) 取值为 3。构建基分类器模型数量 (ntree) 取值为 500, 最小子节点 (nodesize) 取值为 1, 由于特征参数较少, 不对最大子节点数加以限制。

1.4.3 类神经网络 (Artificial Neural Network, ANN)

类神经网络具有强大的拟合能力和自我学习能力, 用于非线性映射回归、多分类预测类研究^[33]。从信息处理角度对人脑神经网络进行抽象, 包括接受外部信号数据的输入单元、完成系统结果对外传递的输出单元, 和不能由系统外部观察得到的隐藏单元。本研究采用 ANN 中的多层感知机反向传播算法 (BP) 的三层人工神经网络进行拟合。使用 Rstudio 软件“nnet”软件包进行模型构建。同时为进一步增加模型稳定性, 避免过度拟合, 使用 SPSS MODELER (version 18.0 IBM) 软件中的 bagging 算法模块对原 ANN 模型进行优化, 合并基分类器规则设定为投票。bagging 组件模型数为 10。

1.4.4 支持向量机 (Support Vector Machine, SVM)

支持向量机相较于类神经网络, 更适用于高纬度、小样本回归或二分类研究^[34], 在现有研究中常根据待测化合物特征加以选择, 构建毒性预测模型。作为一种常见的, 对数据进行二元分类的广义线性算法, 其决策边界是对学习样本求解的最大边距超平面。当线性不可分时, 将松弛变量引入超平面, 并在模型中产生另一个惩罚系数。使用 Rstudio 软件“e1071”软件包中的“svm”函数进行模型的构建, 核函数为分类非线性 SVM 中使用最多的高斯径向基核 (RBF), 运用“tune”函数试错法将 γ 范围设定为 10^{-6} [-1], 惩罚因子 C 范围设定为 10^4 [1,4]进行模型优化, 最终选择最佳参数 γ 取值 0.1, C 取值 1000。

1.4.5 K 最近邻值算法 (K-NearestNeighbor, KNN)

KNN 算法依靠距离划分未知样本类别,通过寻找多个离测试样本最邻近的训练样本,以该样本的多数类别作为最终类别,操作简单结果容易解读,可见于部分药物分析模型^[35]。但这一计算方法不适用于大样本数据,且样本不平衡时,模型预测偏差较大^[36]。本研究通过 SPSS.MODELER 软件 KNN 模块进行内部十折交叉验证选择最合适 K 值,距离计算方法为 Euclidean 法。最终模型选定 K 值为 6 时具有最低错误率。

1.5 模型的评价与验证

本研究采用十折交叉验证进行数据集内部鲁棒性验证,即将原始数据随机平均分为 10 个子集,每个子集做测试集的同时,其余 9 个子集合并作为训练集,进行 10 次训练预测^[37]。十次十折交叉验证由 Rstudio 软件完成。测试集化合物分类进行外部验证,使用 Rstudio 软件“confusionMatrix”函数输出混淆矩阵,计算其中的真阳性(True Positive, TP)、真阴性(True Negative, TN)、假阳性(False Positive, FP)、假阴性(False Negative, FN),分别计算灵敏度(Sensitivity, SE)、特异度(Specificity, SP)、假阳性率(False Positive Rate, FPR)、假阴性率(False Negative Rate, FNR)、总准确性(Global Accuracy, GA)。通过“pROC”软件包绘制受试者工作曲线(Receiver Operating Characteristic, ROC)并计算曲线下面积(The Area Under The ROC Curve, AUC)进行模型评价。

$$SE = \frac{TP}{TP + FN} \quad (2)$$

$$SP = \frac{TN}{TN + FP} \quad (3)$$

$$GA = \frac{TP + TN}{TP + FN + TN + FP} \quad (4)$$

2 结果与讨论

2.1 描述符共线性分析

5 种描述符间相关性分析结果如表 2 所示,可见任意两个描述符间相关系数<0.75。描述符之间不存在明显的共线性问题。

2.2 模型质量评价

在进行数据分割后,训练集和测试集分别有 89 和 18 种化合物,且类别分布较为均匀无需进行数据不

均衡处理。具体数据见图 1。

表 2 5 个分子描述符的相关矩阵分析

Table 2 Correlation matrix analysis of five molecular descriptors

描述符	D015	D142	D152	D175	D564
D015	1.00				
D142	0.54	1.00			
D152	-0.39	-0.24	1.00		
D175	-0.16	-0.30	-0.01	1.00	
D564	0.37	0.36	0.16	-0.12	1.00

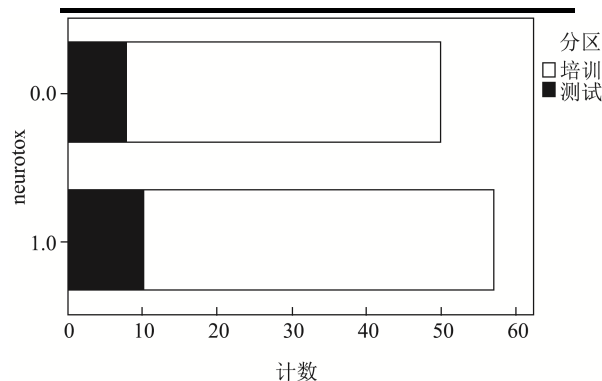


图 1 不同数据集中神经毒性类型分布

Fig.1 Distribution of neurotoxicity in different data sets

五种机器学习的表现各不相同,根据外部验证的结果,五种模型中随机森林模型综合表现最优。训练集中灵敏度 91.49%、特异度 100.00%, GA 和 AUC 均超过 90%。在测试集中表现也较为稳定,灵敏度、特异度分别为 80.00%、87.50%,同时 GA 和 AUC 均超过 80%。具体数据见表 3。

模型鲁棒性检验结果显示,在五种模型中,随机森林模型表现最为稳定,测试集平均准确率达到 70.24%。其次为 ANN 和 KNN 模型,准确率分别为 67.34%和 64.50%。支持向量机和贝叶斯模型的鲁棒性稍差,分别为 61.28%和 62.91%。

模型的预测准确性方面,随机森林的模型表现也更优,训练集准确性为 95.51%,测试集准确性为 83.33%,该模型给出的训练集、测试集 ROC 曲线如图 2 所示,曲线下面积可达到 0.99 和 0.85,是个较为理想的分类模型。

目前已有少量研究利用机器学习算法建立了的化合物毒性预测模型。如一项基于概率神经网络算法(PNN),纳入 47 种有机溶剂作为研究对象,预测其神经毒性的分类模型。训练集、测试集中 GA 分别达 0.91、0.93,曲线下面积分别达 0.92、0.94^[38]。一项基于 RF 算法,纳入 61 种农药作为研究对象,预测其熊蜂急性接触毒性的分类模型。训练集、测试集中 GA 分别达 0.83、0.87,曲线下面积分别达 0.91、0.92^[39]。

相较于前两种模型较为局限的化合物类别，本研究纳入的食品污染物种属各异，脱离了传统的线性模型和化合物类型的限制。在扩大模型适用范围的同时，保持了较高的总预测准确度和 AUC 值，训练集、测试集 AUC 分别达 0.99、0.85。而相较于现有的大样本中药的神经保护作用预测模型(训练集 SE:0.67、SP:0.99；测试集 SE:0.69、SP:0.98)^[40]，极大的提高了模型灵敏度，训练集、测试集灵敏度分别达 0.91、0.8，适用于大样本数据库化合物神经毒性的初步筛查。

2.3 分类错误案例分析

如上文所述，随机森林算法对数据库中化合物具有较好的分类准确率，但预测模型中仍有一些化合物不能被正确的分类，分类错误的 7 个化合物的结构、错分类型和描述符数值如表 4、表 5 所示。

表 4、表 5 数据表明，该随机森林模型在特异度上表现良好，但预测化合物时存在一定的假阴性现象。同时，描述符 D564（质量加权 Burden 矩阵最大特征值）取值可能在一定程度影响模型对于化合物神

经毒性的判定，D564 取值大的化合物更倾向于被判定为阳性，而 D564 取值小的化合物则相反。

化合物分类错误的原因可能由活性悬崖（化合物具有类似结构但活性不同）和骨架跃迁（化合物具有不同结构但活性相似）导致，如苯并(a)芘和苯并(b)芘。而化合物中质量加权 Burden 矩阵最大特征值在化合物神经毒性推断方面可能有比较重要的作用。为了进一步验证该结论，我们选取了两种已知神经毒性，且 D564 描述符相差较大的数据库外化合物作为参考带入模型，具体化合物和描述符取值见表 5。结果支持这一推论。本研究模型相较于现有研究中基于 SVM 算法的周围神经病变预测模型（训练集、测试集 GA 均为 0.89，未见 AUC 指标）^[41]，具有更好的模型解读能力，强调了质量加权 Burden 矩阵最大特征值在化合物神经毒性预测中的作用。未来研究中我们将进一步增加数据库中化合物数量，突破单一模型限制，构建集成模型以提高化合物神经毒性预测准确性和普适性，有望用于指导高风险化合物的毒性测试筛选策略的制定。

表 3 基于 5 个描述符对于神经毒性预测训练集及测试集的统计结果

Table 3 Statistical results of neurotoxicity prediction training set and test set based on 5 descriptors

模型	数据集	SE/%	SP/%	FPR/%	FNR/%	GA/%	AUC
RF	训练集	91.49	100.0	0.00	8.51	95.51	0.99
	测试集	80.00	87.50	12.50	20.00	83.33	0.85
KNN	训练集	80.85	64.29	35.71	19.15	73.03	0.76
	测试集	80.00	75.00	25.00	20.00	77.78	0.90
ANN	训练集	87.23	90.48	9.52	12.77	88.76	0.92
	测试集	60.00	100.00	0.00	40.00	77.78	0.74
SVM	训练集	65.96	83.33	16.67	34.04	74.16	0.81
	测试集	70.00	87.50	12.50	30.00	77.78	0.84
BN	训练集	57.45	83.33	16.67	42.55	69.66	0.79
	测试集	70.00	100.00	0.00	30.00	83.33	0.89

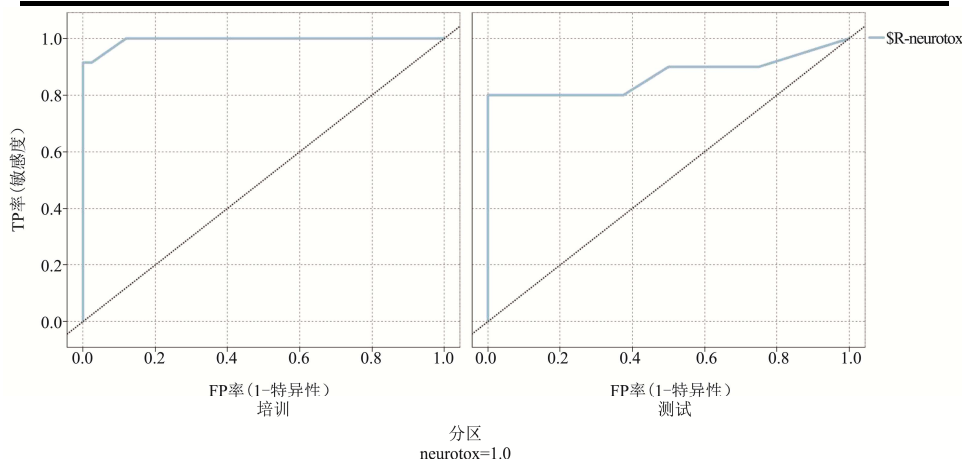


图 2 随机森林模型对训练集、测试集的预测结果 ROC 曲线

Fig.2 ROC curve of prediction results of random forest model on training set and test set

表4 随机森林算法中分类错误化合物

Table 4 Misclassified compounds in random forest algorithm

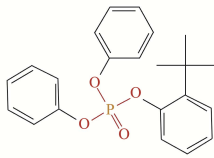
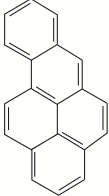
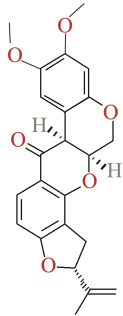
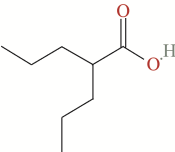

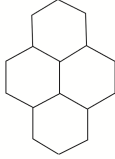
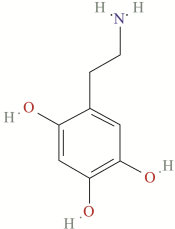
Cas 号	2D 结构	错分类型	Cas 号	2D 结构	错分类型
56803-37-3		FP	50-32-8		FN
83-79-4		FN	99-66-1		FN
110-54-3		FN	129-00-0		FN
28094-15-7		FN			

表5 错误分类化合物及库外参考化合物描述符取值

Table 5 Misclassified compounds and values of off-library reference compound descriptors

Cas	分类	D015	D142	D152	D175	D564
50-32-8	FN	0.00	0.29	0.95	27470.00	9.15
83-79-4	FN	0.18	0.25	0.77	67965.00	9.29
99-66-1	FN	0.89	0.34	0.58	300.00	9.11
110-54-3	FN	1.00	0.48	0.55	70.00	7.80
129-00-0	FN	0.00	0.22	0.94	10900.00	9.10
28094-15-7	FN	0.50	0.29	0.68	1098.00	9.05
56803-37-3	FP	0.35	0.22	0.76	16631.00	15.66
81-07-2*	FP	0.00	0.25	0.92	1987.00	16.66
1199-18-4*	FN	0.68	0.28	0.62	6254.00	9.85

注: *为库外化合物。

3 结论

3.1 本研究收集整理了具有影响神经分化成熟、神经元迁移/空间定向、突触形成/可塑性、神经递质、神经营养功能、神经免疫系统、神经内分泌、离子稳态

干扰等诱发神经毒性机制的食品污染物 57 种,无神经毒性污染物 50 种共 107 种食品污染物,运用 5 种机器学习算法,建立了食品化学污染物神经毒性预测模型。3.2 运用 R 和 SPSS 软件对 778 个计算变量进行了筛选,最终确定将 5 个分子描述符纳入模型,即 D015 可旋转键数比例、D142 Balaban 均方顶点距离指数、D152 碳-sp3 的平均原子极化率、D175 Wiener 最大路径指数和 D564 质量加权 Burden 矩阵最大特征。3.3 5 种机器学习算法中,基于随机森林算法建立的分类模型具有最好的预测效果,外部验证中训练集灵敏度 91.49%、特异度 100.00%,总准确性 95.51%,测试集灵敏度 80.00%、特异度 87.50%,总准确性 83.33%,曲线下面积分别达 0.99 和 0.85。内部验证中模型鲁棒性达 70.24%。模型预测结果中共有 7 种化合物不能被正确分类,造成这一结果的原因可能是活性悬崖或骨架跃迁现象的存在。在五种分子描述符中,质量加权 Burden 矩阵最大特征值对预警化合物高风险结构特征贡献最佳。

参考文献

- [1] Rezaei N J, Bazzazi A M, Naseri A S. Neurotoxicity of the antibiotics: a comprehensive study [J]. *Neurol India*, 2018, 66(6): 1732-1740
- [2] Richardson J R, Fitsanakis V, Westerink R, et al. Neurotoxicity of pesticides [J]. *Acta Neuropathol*, 2019, 138(3): 343-362
- [3] 王雪婷,沈艳花,付小龙.含 Mn~(2+)Pb~(2+)和 Al~(3+)化合物联合作用对神经细胞株 PC12 细胞毒性及机制[J].*公共卫生与预防医学*,2019,30(4):21-24
WANG Xueting, SHEN Yanhua, FU Xiaolong. Addictive toxicity and mechanism of Mn²⁺ Pb²⁺ and Al³⁺ compounds in PC12 cells [J]. *J of Pub Health and Prev Med*, 2019, 30(4): 21-24
- [4] 高瑜悦,欧仕益,张延杰,等.黄酮类化合物对食品中丙烯酰胺的抑制和毒性干预作用[J].*轻工科技*,2021,37(1):9-14
GAO Yuyue, OU Shiyi, ZHANG Yanjie, et al. Inhibitory and toxic effects of flavonoids on acrylamide in food [J]. *Light Industry Science and Technology*, 2021, 37(1): 9-14
- [5] Ortega García J A, Ferris I Tortajada J, Berbel Tornero O, et al. Environmental neurotoxins (IV). Tobacco, alcohol, solvents, fluoride, food additives: adverse effects on the fetal and postnatal nervous system. Preventive measures [J]. *Acta Pediatrica Espanola*, 2006, 64(10): 493-502
- [6] 田康,黄芝瑛,王雪,等.药物神经毒性评价体外模型的研究进展[J].*药物评价研究*,2020,43(7):1433-1438
TIAN Kang, HUANG Zhiying, WANG Xue, et al. Research progress on in vitro models for evaluating drug-induced neurotoxicity [J]. *Drug Evaluation Research*, 2020, 43(7): 1433-1438
- [7] Rahmani N, Abbasi-Radmoghaddam Z, Riahi S, et al. Predictive QSAR models for the anti-cancer activity of topoisomerase II α catalytic inhibitors against breast cancer cell line HCT15: GA-MLR and LS-SVM modeling [J]. *Structural Chemistry*, 2020, 31(2): 2129-2145
- [8] Huixiao Hong, Minjun Chen, Hui Wen Ng, et al. In Silico Methods for Predicting Drug Toxicity [M]. New York: Humana Press, 2016, 431-459
- [9] Zhao M, Wang L, Zheng L, et al. 2D-QSAR and 3D-QSAR analyses for EGFR inhibitors [J]. *Biomed Research International*, 2017, 2017: 4649191
- [10] Roy K, Das R N. A review on principles, theory and practices of 2D-QSAR [J]. *Current Drug Metabolism*, 2014, 15(4): 346-379
- [11] Antanasijević D, Antanasijević J, Trišović N, et al. From classification to regression multitasking QSAR modeling using a novel modular neural network: simultaneous prediction of anticonvulsant activity and neurotoxicity of succinimides [J]. *Molecular Pharmaceutics*, 2017, 14(12): 4476-4484
- [12] 金飙,柴欢,金俏.基于不同建模方法的硝基苯急性毒性 QSAR 比较研究[J].*沈阳化工大学学报*,2016,30(2):124-129
JIN Biao, CHAI Huan, JIN Qiao. Corrosion inhibition properties of phenylalanine and tyrosine [J]. *Journal of Shenyang University of Chemical Technology*, 2016, 30(2): 124-129
- [13] Mozaffari S, Ghasemi S, Baher H, et al. Synthesis and evaluation of some novel methylene-bridged aryl semicarbazones as potential anticonvulsant agents [J]. *Medicinal Chemistry Research*, 2012, 21(11): 3797-3808
- [14] Estrada E, Molina E, Uriarte E. Quantitative structure-toxicity relationships using tops-mode. 2. Neurotoxicity of a non-congeneric series of solvents [J]. *SAR and QSAR in Environmental Research*, 2001, 12(5): 445-459
- [15] Bal-Price A, Lein P J, Keil K P, et al. Developing and applying the adverse outcome pathway concept for understanding and predicting neurotoxicity [J]. *Neurotoxicology*, 2017, 59: 240-255
- [16] Handelman G S, Kok H K, Chandra R V, et al. eDoctor: machine learning and the future of medicine [J]. *Journal of Internal Medicine*, 2018, 284(6): 603-619
- [17] Sirenko O, Parham F, Dea S, et al. Functional and mechanistic neurotoxicity profiling using human iPSC-derived neural 3D cultures [J]. *Toxicological Sciences*, 2019, 167(1): 58-76
- [18] Tong Z, Hogberg H, Kuo D, et al. Characterization of three human cell line models for high-throughput neuronal cytotoxicity screening [J]. *Journal of Applied Toxicology*, 2017, 37(2): 167-180
- [19] Ryan K R, Sirenko O, Parham F, et al. Neurite outgrowth in human induced pluripotent stem cell-derived neurons as a high-throughput screen for developmental neurotoxicity or neurotoxicity [J]. *Neuro Toxicology*, 2016, 53: 271-281
- [20] Seth A, Ojha P K, Roy K. QSAR modeling with ETA indices for cytotoxicity and enzymatic activity of diverse chemicals [J]. *Journal of Hazardous Materials*, 2020, 394: 122498
- [21] 李钦玲,杨玉良.人工神经网络预测取代芳烃类化合物对黑头鱼的毒性[J].*计算机与应用化学*,2017,34(5):395-400
LI Qinling, YANG Yuliang. Predicting the toxicity of

- substituted aromatic compounds to fathead minnows using artificial neural network [J]. *Computers and Applied Chemistry*, 2017, 34(5): 395-400
- [22] 王晓静,刘爱梅,徐小庆,等.定量构效关系建模及其在联合毒性中应用研究进展[J].*动物医学进展*,2020,41(6):86-90
WANG Xiaojing, LIU Aimei, XU Xiaoqing, et al. Progress on QSAR modeling and its application in joint toxicity [J]. *Progress in Veterinary Medicine*, 2020, 41(6): 86-90
- [23] Chen B, Zhang T, Bond T, et al. Development of quantitative structure activity relationship (QSAR) model for disinfection byproduct (DBP) research: a review of methods and resources [J]. *Journal of Hazardous Materials*, 2015, 299: 260-279
- [24] Hong H, Xie Q, Ge W, et al. Mold2, molecular descriptors from 2D structures for chemoinformatics and toxicoinformatics [J]. *Journal of Chemical Information and Modeling*, 2008, 48(7): 1337-1344
- [25] 张钊.抗氧化剂分子数据库建立及其 QSAR 分析[D].哈尔滨:东北农业大学,2019
ZHANG Po, A curated database of antioxidant molecular and its QSAR analysis [D]. Harbin: Northeast Agricultural University, 2019
- [26] Bickerton G R, Paolini G V, Besnard J, et al. Quantifying the chemical beauty of drugs [J]. *Nature Chemistry*, 2012, 4(2): 90-98
- [27] Lambrinidis G, Tsantili-Kakoulidou A. Challenges with multi-objective QSAR in drug discovery [J]. *Expert Opinion on Drug Discovery*, 2018, 13(9): 851-859
- [28] 雷太龙.基于机器学习方法的药物毒性的理论预测研究[D].杭州:浙江大学,2017
LEI Tailong, Theoretical prediction of drug toxicity based on machine learning approaches [D]. Hangzhou: Zhejiang University, 2017
- [29] Darst B F, Malecki K C, Engelman C D. Using recursive feature elimination in random forest to account for correlated variables in high dimensional data [J]. *BMC Genetics*, 2018, 19(Suppl 1): 65
- [30] 康家雄,朱江,李爱秀,等.MKC-442 及其类似物的二维定量构效关系[J].*中国新药杂志*,2019,28(23):2885-2892
KANG Jiexiong, ZHU Jiang, LI Aixiu, et al. 2-Dimensional quantitative structure-activity relationship of MKC-442 and its analogues [J]. *Chinese Journal of New Drugs*, 2019, 28(23): 2885-2892
- [31] Carnesecchi E, Toma C, Roncaglioni A, et al. Integrating QSAR models predicting acute contact toxicity and mode of action profiling in honey bees (*A. mellifera*): data curation using open source databases, performance testing and validation [J]. *Science of the Total Environment*, 2020, 735: 139243
- [32] 吴地尧,章新友,张玉娇,等.分类算法在中药研究中的应用及其进展[J].*科学技术与工程*,2019,19(35):1-9
WU Diyao, ZHANG Xinyou, ZHANG Yujiao, et al. Application and progress of classification algorithm in traditional Chinese medicine research [J]. *Science Technology and Engineering*, 2019, 19(35): 1-9
- [33] 汪奇,刘尚全.2 型糖尿病患者亚临床动脉粥样硬化的多层人工神经网络分类预测模型的构建[J].*中国全科医学*,2021, 24(36):4612-4617
WANG Qi, LIU Shangquan. Construction of a multi-layer artificial neural network classification model for predicting subclinical atherosclerosis in type 2 diabetic patients [J]. *Chinese General Practice*, 2021, 24(36): 4612-4617
- [34] 陈慧,谭超,莫琳,夏成云.基于支持向量机建立取代苯类化合物毒性预测模型[J].*毒理学杂志*,2011,25(3):229-231
CHEN Hui, TAN Chao, MO Lin, et al. The toxicity prediction model of substituted benzene compounds was established based on support vector machine [J]. *Toxicol*, 2011, 25(3): 229-231
- [35] Bi Q, Goodman K E, Kaminsky J, et al. What is machine learning? A primer for the epidemiologist [J]. *American Journal of Epidemiology*, 2019, 188(12): 2222-2239
- [36] Salvador-Meneses J, Ruiz-Chavez Z, Garcia-Rodriguez J. Compressed kNN: K-nearest neighbors with data compression [J]. *Entropy (Basel)*, 2019, 21(3): 234
- [37] 梁子超,李智炜,赖铿,等.10 折交叉验证用于预测模型泛化能力评价及其 R 软件实现[J].*中国医院统计*,2020,27(4): 289-292
LIANG Zichao, LI Zhiwei, LAI Keng, et al. Application of 10-fold cross-validation in the evaluation of generalization ability of prediction models and the realization in R [J]. *Chinese Journal of Hospital Statistics*, 2020, 27(4): 289-292
- [38] Basant N, Guptas, Singh K P. Predicting the acute neurotoxicity of diverse organic solvents using probabilistic neural networks based QSTR modeling approaches [J]. *Neurotoxicology*, 2016, 53: 45-52

(下转第 306 页)