

同行专家业内评价意见书编号: 20240860014

## 附件1

# 浙江工程师学院（浙江大学工程师学院） 同行专家业内评价意见书

姓名: \_\_\_\_\_ 杨金荣

学号: \_\_\_\_\_ 22160157

申报工程师职称专业类别（领域）: \_\_\_\_\_ 生物与医药

浙江工程师学院（浙江大学工程师学院）制

2024年03月18日



## 一、个人申报

**（一）基本情况【围绕《浙江工程师学院（浙江大学工程师学院）工程类专业学位研究生工程师职称评审参考指标》，结合该专业类别(领域)工程师职称评审相关标准，举例说明】**

### 1. 专业基础理论知识和专业技术知识掌握情况

对与人工智能药学交叉学科，需要既掌握AI算法和药学基础。首先我对对人工智能药学项目生物与医药专业基础理论知识掌握扎实，本专业的基础课程均取得不错的成绩，如生物大分子模拟88、新药发现理论与实践97；专业技术知识掌握熟练，如AI算法和制药技术课程也取得了不错的成绩，如人工智能算法与系统88，先进制药技术87，对工程专业知识把握扎实，具有工程师应用思维。

### 2. 工程实践经历

本人在中国科学院上海药物研究所进行为期一年的算法实习生专业实践，主要负责化合物多维结构提取信息从而建立QSXR预测模型

### 3. 实际工作中综合运用所学知识解决复杂工程问题的案例

近年来，我们项目通过校企合作项目，为国内外相关企业生产化学产品提供毒性和毒代预测和咨询，通过项目的实施，我们构建一系列毒性QSXR模型，例如：生殖发育毒性QSXR模型构建，建立的LOAEL-QSAR模型和NOAEL-

QSAR模型，分别以平均绝对误差（MAE），均方误差（MSE）和均方根误差（RMSE）3种误差类型和R2来评估模型拟合效果。结果显示构建的LOAEL-QSAR模型和NOAEL-QSAR模型都具有较低的误差值和较高的R2值，表明其具有良好的预测能力，可用于未知化合物生殖毒性LOAEL和NOAEL值估算。

根据团队分工，本人主要负责对化合物不同官能团进行统计学处理，从概率学角度出发，针对显著性水平进行分析，判断安全或危险阈值是否选取得当，显著性水平是假设检验中的一个概念，是指当原假设为正确时人们却把它拒绝了的概率或风险。它是公认的小概率事件的概率值，必须在每一次统计检验之前确定，通常取  $\alpha = 0.05$  或

$\alpha = 0.01$ 。这表明当作出接受原假设的决定时，其正确的可能性（概率）为95%或99%。根据各官能团统计学处理分析形成最后的分析报告，成为应用于药化学家合成过程中的辅助工具。

同时，本人根据数学算法公式辅助团队实现通过Python计算直接得到化合物的范德华表面和LeeRichards表面三维坐标点，通过Gaussian软件在B3LYP/6-

31(d, p)水平下对分子结构进行优化且对优化后的结构计算其原子的电荷信息。从小分子化合物的电子分布模型角度出发，根据电子分布模型的计算公式，赋予上述范德华表面和Lee-Richards表面三维坐标物理意义，构建全新的、可实用、且能可视化模型中分子重要点集分子描述符，并实现对化合物生殖毒性LOAEL和NOAEL值更好地预测。

在实践工作中，本人也负责化合物一、二维分子结构提取和机器学习算法应用，以及统计学数据分析。本人从化合物的一维结构信息主要是从化合物的SMILE式出发，采用自然语言处理如one-

hot和embedding等来提取化合物的信息。使用隐含层tanh函数激活，再经过池化层进行输出矩阵的展开，得到化合物的特征向量。将其与深度神经网络（DNN）、多层感知机（MLP）等前沿深度学习算法配对，实现化合物的定量毒性预测。同时也可根据自然语言处理方法提取蛋白质、DNA等受体的编码信息，推进小分子化合物与靶点结合作用力（CPI）的预测进程。从化合物的二维结构出发，深度挖掘化合物的图信息，并通过图信息开发新的二维物理模型的辅助描述符（如新的分子指纹）。将二维图结构和物理描述符与前沿的深度学习算法配对，如深度神经网络（DNN）、图卷积神经网络（GCN）和图神经网络（GNN）等，以推进其应用于定量毒性预测。同时针对毒性预测的靶点如蛋白质，也可通过图算法挖掘出其二维结构信息，将小分子化合物与靶点的信息配对，结合深度学习算法实现配体与受体的CPI指标（IC50、E



C50等) 预测。为了将化合物的毒性预测推广到实际药物制作过程中, 本人推出化合物合成过程中的官能团信号分析概念。通过统计学大数据处理分析, 计算在毒性预测过程中, 化合物的不同官能团出现概率大小。如果出现概率超出我们所设定的毒性危险阈值, 即给该官能团标注红色警示信号, 并给出相应的概率报告, 建议药化研究员在某种毒性测试下避免该官能团出现在化合物合成过程中, 反之, 如果出现概率大于安全阈值, 即给该官能团标注绿色安全信号, 推荐研究员合成化合物时使用该官能团。

**(二) 取得的业绩 (代表作) 【限填3项, 须提交证明原件 (包括发表的论文、出版的著作、专利证书、获奖证书、科技项目立项文件或合同、企业证明等) 供核实, 并提供复印件一份】**

**1. 公开成果代表作 【论文发表、专利成果、软件著作权、标准规范与行业工法制定、著作编写、科技成果获奖、学位论文等】**

成果名称	成果类别 [含论文、授权专利 (含发明专利申请)、软件著作权、标准、工法、著作、获奖、学位论文等]	发表时间/授权或申请时间等	刊物名称/专利授权或申请号等	本人排名/总人数	备注
论文发表	核心期刊	2022年06月01日	computer in biology and medicine	1/5	SCI期刊收录

**2. 其他代表作 【主持或参与的课题研究项目、科技成果应用转化推广、企业技术难题解决方案、自主研发设计的产品或样机、技术报告、设计图纸、软课题研究报告、可行性研究报告、规划设计方案、施工或调试报告、工程实验、技术培训教材、推动行业发展中发挥的作用及取得的经济社会效益等】**

<b>(三) 在校期间课程、专业实践训练及学位论文相关情况</b>	
课程成绩情况	按课程学分核算的平均成绩： 83 分
专业实践训练时间及考核情况(具有三年及以上工作经历的不作要求)	累计时间： 1 年(要求1年及以上) 考核成绩： 89 分(要求80分及以上)
<b>本人承诺</b>	
<p>个人声明：本人上述所填资料均为真实有效，如有虚假，愿承担一切责任，特此声明！</p> <p style="text-align: right;">申报人签名： 杨金荣</p>	





## 浙江大学研究生院

## 攻读硕士学位研究生成绩单

学号: 22160157	姓名: 杨金荣	性别: 男	学院: 工程师学院	专业: 生物与医药	学制: 2.5年						
毕业时最低应获: 24.0学分		已获得: 24.0学分		入学年月: 2021-09	毕业年月: 2024-03						
学位证书号: 1033532024602177			毕业证书号: 103351202402600403								
学习时间	课程名称	备注	学分	成绩	课程性质	学习时间	课程名称	备注	学分	成绩	课程性质
2021-2022学年秋季学期	研究生英语基础技能		1.0	65	公共学位课	2021-2022学年秋季学期	研究生论文写作指导		1.0	82	专业学位课
2021-2022学年秋季学期	人工智能算法与系统		2.0	88	专业学位课	2021-2022学年春季学期	数学建模		2.0	85	专业选修课
2021-2022学年冬季学期	新药发现理论与实践		2.0	97	专业学位课	2021-2022学年春季学期	科技创新案例探讨与实践		2.0	87	专业选修课
2021-2022学年冬季学期	先进制药技术		2.0	87	专业选修课	2021-2022学年春季学期	生物大分子模拟		2.0	88	专业选修课
2021-2022学年秋季学期	中国特色社会主义理论与实践研究		2.0	90	公共学位课	2021-2022学年春季学期	自然辩证法概论		1.0	78	公共学位课
2021-2022学年秋季学期	研究生英语		2.0	80	公共学位课	2021-2022学年夏季学期	药品创制工程实例		2.0	88	专业学位课
2021-2022学年冬季学期	工程伦理		2.0	77	公共学位课	2021-2022学年夏季学期	批判性思维与科学研究		1.0	91	专业选修课

说明: 1. 研究生课程按三种方法计分: 百分制, 两级制 (通过、不通过), 五级制 (优、良、中、

及格、不及格)。

2. 备注中“\*”表示重修课程。

学院成绩校核章:

成绩校核人: 张梦依

打印日期: 2024-04-02

《SCI-Expanded》收录证明

经检索《Web of Science》的《Science Citation Index Expanded (SCI-EXPANDED)》数据库，下述论文被《SCI-EXPANDED》收录。（检索时间：2023年12月20日）

第 1 条，共 1 条

标题:Reliable CA-(Q)SAR generation based on entropy weight optimized by grid search and correction factors

作者:Yang, JR(Yang, Jin-Rong);Chen, Q(Chen, Qiang);Wang, H(Wang, Hao);Hu, XY(Hu, Xu-Yang);Guo, YM(Guo, Ya-Min);Chen, JZ(Chen, Jian-Zhong);

来源出版物:COMPUTERS IN BIOLOGY AND MEDICINE 卷:146 文献号:105573 提前访问日期:MAY 2022 DOI:10.1016/j.combiomed.2022.105573 出版年:JUL 2022

入藏号:WOS:000799838700002

文献类型:Article

地址:

[Yang, Jin-Rong; Chen, Qiang; Wang, Hao; Hu, Xu-Yang; Guo, Ya-Min; Chen, Jian-Zhong] Zhejiang Univ, Coll Pharmaceut Sci, 866 Yuhangtang Rd, Hangzhou 310058, Zhejiang, Peoples R China.

[Yang, Jin-Rong] Zhejiang Univ, Polytech Inst, 269 Shixiang Rd, Hangzhou 310015, Zhejiang, Peoples R China.

通讯作者地址:

Chen, JZ (corresponding author), Zhejiang Univ, Coll Pharmaceut Sci, 866 Yuhangtang Rd, Hangzhou 310058, Zhejiang, Peoples R China.

电子邮件地址:chjz@zju.edu.cn

IDS 号:1M2YB

ISSN:0010-4825

eISSN:1879-0534

注:

1. 以上检索结果来自 CALIS 查收索引系统。
2. 以上检索结果均得到委托人及被检索作者的确认。







Contents lists available at ScienceDirect

## Computers in Biology and Medicine

journal homepage: [www.elsevier.com/locate/combiomed](http://www.elsevier.com/locate/combiomed)



### Reliable CA-(Q)SAR generation based on entropy weight optimized by grid search and correction factors

Jin-Rong Yang<sup>a,b</sup>, Qiang Chen<sup>a</sup>, Hao Wang<sup>a</sup>, Xu-Yang Hu<sup>a</sup>, Ya-Min Guo<sup>a</sup>, Jian-Zhong Chen<sup>a,\*</sup>

<sup>a</sup> College of Pharmaceutical Sciences, Zhejiang University, 866 Yuhangtang Rd, Hangzhou, Zhejiang, 310058, China

<sup>b</sup> Polytechnic Institute, Zhejiang University, 269 Shaoxing Rd, Hangzhou, Zhejiang, 310015, China

#### ARTICLE INFO

**Keywords:**  
QSAR  
Chromosome aberration  
Machine learning  
Connection index  
Similarity return search

#### ABSTRACT

Chromosome aberration (CA) is a serious genotoxicity of a compound, leading to carcinogenicity and developmental side effects. In the present manuscript, we developed a QSAR model for CA prediction using artificial intelligence methodologies. The reliable QSAR model was constructed based on an enlarged data set of 3208 compounds by optimizing machine learning and deep learning algorithms based on hyperparametric iterations and using multiple descriptors of molecular fingerprint in combination with drug-like molecular properties (MP) screened by entropy weight methodology on the open-source Python platform. Furthermore, molecular similarity for returning search and molecular connection index for additional descriptor were additionally introduced to differentiate the compounds with high similarity for correct CA prediction for QSAR model generation. The final generated CA-(Q)SAR model exhibited good prediction accuracy of 80.6%. The bias of the final model is about 0.9793. On the basis of generated QSAR model, data analyses were further performed to analyze the typical structure features in numerical intervals (MPI) of molecular properties MW, XlogP, and TPSA, respectively, for potential CA or non-CA toxicity with a normalized occurrence probability (NOP) more than 70%, which may provide useful clues for drug design of leads or candidate devoid of CA genotoxicity.

#### 1. Introduction

Chemical genotoxicity is defined for a substance to cause DNA damage or mutation in cell, resulting in carcinogenicity [1] and developmental side effects [2]. Genotoxicity studies of chemicals are particularly significant for drug R&D or environmental protection. *In vitro* genotoxicity test is to detect a chemical inducing gene mutation and chromosome damage, having the advantage of enabling the discovery of genotoxic compound to act on chromosomes or DNA without the necessity for experimental animals [3]. Chromosome aberrations (CA) include metaphase chromosome breakage, rearrangement, and quantitative changes. CA testing is a hazard evaluation for potential genotoxicity of a compound in the process of early drug candidate selection, product development, and regulatory review [4], and also for chemicals in toxicology research. However, the application of this testing remains controversial owing to the high rate of false-positive results generated under *in vitro* conditions [5], although standard experimental

approaches have been recommended for CA detections with OECD guidelines 473 [6], 475 [7], and 483 [8]. On the other hands, construction of prediction model based on reported CA data may provide a practicable and useful approach to estimate validly CA response of a compound and to understand structure-CA relationship, which is conducive to drug design and screening.

With rapid growth of biological data and development of computer science, the technology is becoming more and more mature for predicting lead biotoxicity using quantitative structure-toxicity relationship (QSAR) model established by artificial intelligence. A couple of (Q)SAR models [9–11] have also been developed for the prediction and discovery of compound CA genotoxicity. Typically, Hsu et al. [11] established CA-(Q)SAR models to predict the genotoxicity of compounds using Lead scope Enterprise (LSE) and Case ultra (CU) methods on commercial SAR platforms based on a CA dataset composed of 1770 compounds. Another two models [9,10] were generated based on a dataset including 129 compound and 383 compounds, respectively.

**Abbreviations:** CA, Chromosome Aberration; MP, Molecular Property; NOP, Normalized Occurrence Probability; QSAR, Quantitative Structure-Activity Relationship; LSE, Lead Scope Enterprise; CU, Case Ultra; FP, Fingerprint; MW, Molecular Weight; TPSA, Topological Polar Surface Area; MCI, Molecular Connectivity Index; HBD, H-bond Donor; HBA, H-bond Acceptor; RBC, Rotatable Bond; ACC, Prediction Accuracy; PAH, Polycyclic Aromatic Hydrocarbons.

\* Corresponding author.

E-mail address: [chjz@zju.edu.cn](mailto:chjz@zju.edu.cn) (J.-Z. Chen).

<http://doi.org/10.1016/j.combiomed.2022.105573>

Received 10 January 2022; Received in revised form 31 March 2022; Accepted 26 April 2022

Available online 30 April 2022

0010-4825/© 2022 Published by Elsevier Ltd.