

МИНИСТЕРСТВО ОБРАЗОВАНИЯ РЕСПУБЛИКИ БЕЛАРУСЬ

БЕЛОРУССКИЙ ГОСУДАРСТВЕННЫЙ УНИВЕРСИТЕТ

ХИМИЧЕСКИЙ ФАКУЛЬТЕТ

Кафедра общей химии и методики преподавания химии

ЛОБКО Роман Александрович

**СОЗДАНИЕ QSAR МОДЕЛИ ДЛЯ ИССЛЕДОВАНИЯ БИОЛОГИЧЕСКОЙ
АКТИВНОСТИ ПРОИЗВОДНЫХ 1,4-ДИЗАМЕЩЁННЫХ-1,2,3-ТРИАЗОЛОВ**

Дипломная работа

Научный руководитель:
старший преподаватель кафедры
общей химии и методики
преподавания химии
В. А. Будевич

Допущен к защите

«__»_____ 2025 г.

Зав. кафедрой общей химии и методики преподавания химии
кандидат химических наук, доцент С. М. Рабчинский

Минск, 2025

РЕФЕРАТ

Дипломная работа, 64 с., 9 рис., 3 табл., 32 источников.

Создание QSAR модели для исследования биологической активности производных 1,4-дизамещённых-1,2,3-триазолов

Ключевые слова: триазол; VCP; ингибитор; QSAR; машинное обучение; метод XGBoost; дескриптор; SMILES

Объектом исследования настоящей дипломной работы являются соотношения структура-активность 1,4-дизамещённых-1,2,3-триазолов. *Целью* данной дипломной работы является построение модели QSAR методом XGboost для прогноза ингибирующей способности 1,4-дизамещённых-1,2,3-триазолов на основе структурных данных. На основе информации, имеющейся в научных публикациях (база данных PubChem), создана база данных ингибиторов VCP для обучения и тестирования модели QSAR. В созданной базе данных содержалась структурная информация молекул и информация об их активности относительно белка, выраженная в виде IC_{50} . Был написан код на Python, выполняющий все этапы QSAR моделирования и предсказания значений IC_{50} . На основе SMILES структур молекул из базы данных рассчитаны дескрипторы молекул. Из рассчитанных были отобраны дескрипторы с наибольшим коэффициентом корреляции с IC_{50} , используя библиотеки `padelru` и `scikit-learn`. Построена регрессионная модель QSAR методом XGBoost. Для построенной модели методом RFECV были подобраны наиболее оптимальные параметры. Оценено качество модели путём расчёта R^2 , MSE, RMSE и Q^2 . С помощью построенной модели осуществлён прогноз свойств исследуемых производных 1,4-дизамещённых-1,2,3-триазолов. Настоящей работой представлен алгоритм, который может быть использован для виртуального скрининга новых лигандов VCP, а также для построения новых QSAR моделей.

РЭФЕРАТ

Дыпломная работа, 64 с., 9 мал., 3 табл., 32 крыніцы.

Стварэнне QSAR мадэлі для даследавання біялагічнай актыўнасці вытворных 1,4-дызамешчаных-1,2,3-трыазолаў

Ключевые слова: трыазол; VCP; інгібітар; QSAR; машыннае навучанне; метада XGBoost; дэсрыптар; SMILES. *Аб'ект даследавання* дадзенай дыпломнай працы — гэта суадносіны структура-актыўнасць 1,4-дызамешчаных-1,2,3-трыазолаў. Мэта працы заключана ў пабудове QSAR мадэлі метадам XGBoost для прагнозу інгібіруючай здольнасці 1,4-дызамешчаных-1,2,3-трыазолаў на аснове структурных дадзеных. На падставе інфармацыі, даступнай у навуковых публікацыях (база даных PubChem), створана база даных інгібітараў VCP для навучання і тэсціравання QSAR мадэлі. У гэтай базе дадзеных утрымлівалася структурная інфармацыя малекул і даныя аб іх актыўнасці адносна бялку, выказаныя ў выглядзе IC_{50} . Быў напісаны код на Python, які выконвае ўсе этапы QSAR мадэлявання і прагназавання значэнняў IC_{50} . На падставе SMILES-структур малекул з базы даных разлічаны дэсрыптары, з якіх адабраны тыя, што маюць найбольшы каэфіцыент карэляцыі з IC_{50} , з выкарыстаннем бібліятэк `padelpy` і `scikit-learn`. Пабудавана рэгрэсійная мадэль QSAR метадам XGBoost. Для гэтай мадэлі метадам RFECV падобраны найбольш аптымальныя параметры. Ацэнена якасць мадэлі шляхам разліку R^2 , MSE, RMSE і Q^2 . З дапамогай пабудаванай мадэлі здзейснены прагноз уласцівасцяў даследаваных вытворных 1,4-дызамешчаных-1,2,3-трыазолаў. Гэтая праца прадстаўляе алгарытм, які можа быць выкарыстаны для віртуальнага скрынінгу новых лігандаў VCP, а таксама для пабудовы новых QSAR мадэляў.

ABSTRACT

Graduate work, 64 p., 9 fig., 3 tab., 32 references.

Development of a QSAR Model for Investigating the Biological Activity of 1,4-Disubstituted-1,2,3-Triazoles

Keywords: triazole; VCP; inhibitor; QSAR; machine learning; XGBoost method; descriptor; SMILES. *The subject* of this thesis is the structure–activity relationship of 1,4-disubstituted 1,2,3-triazoles. The aim of this work is to construct a QSAR model using the XGBoost method to predict the inhibitory activity of 1,4-disubstituted 1,2,3-triazoles based on their structural data. Based on the information available in scientific publications (PubChem database), a database of VCP inhibitors was compiled for training and testing the QSAR model. This database contained both the molecular structural information and data on the molecules' activity against the protein, expressed in terms of IC_{50} . A Python code was developed that executes all stages of QSAR modeling and predicts IC_{50} values. Based on the SMILES strings of the molecules from the database, molecular descriptors were computed. From these calculated descriptors, those with the highest correlation coefficient with IC_{50} were selected using the padelpy and scikit-learn libraries. A QSAR regression model was then constructed using the XGBoost method. For the developed model, the most optimal parameters were selected using the RFECV method. The quality of the model was evaluated by calculating R^2 , MSE, RMSE, and Q^2 . Using the constructed model, the properties of the studied derivatives of 1,4-disubstituted 1,2,3-triazoles were predicted. This work presents an algorithm that can be used for the virtual screening of new VCP ligands as well as for the construction of new QSAR models.

