Molecule activity studies on *Entamoeba histolytica* organism via QSAR modelling and molecular docking

**Abstract**

*Amebiasis* is a neglected tropical disease that’s caused by the protozoan parasite *Entamoeba histolytica*. This disease is one of the leading causes of diarrhea globally, affecting largely impoverished residents in developing countries. *Amebiasis* also remains one of the top causes of gastrointestinal diseases in returning international travelers. [1]. In this study, I use machine learning approaches (CatBoost, Forest Regressor, and Gradient Boost) to predict the pIC50 value of molecules from a dataset extracted from the ChEMBL and employed in quantitative structure-activity relationship (QSAR) study so as to gain insights on their origin bioactivity.

# Introduction

*Amebiasis* is a disease caused by the protozoan parasite *Entamoeba histolytica* and is a major public health crisis in developing countries [2]. The global burden lies in tropical and subtropical countries suffering from poor sanitation facilities [3]. As per WHO, *Entamoeba histolytica* infects approximately 50 million people worldwide and causes the death of around 100,000 people annually [4]. The parasite is among the top 15 causes of diarrhea in children under the age of two years and thus hampers their mental and physical growth [5].

For the research method, this study uses QSAR. Quantitative structure–activity relationship (QSAR) is a paradigm that enables the prediction of biological activities for compounds of interest as a function of their descriptors through the use of statistical or machine learning methods [6]. Aside from the ability to predict the activity, QSAR models have been instrumental in understanding the origin of these biological activities by interpreting the descriptors used in building such models.

Diagram

Description automatically generated

### Figure 1: Workflow of QSAR modelling and molecular docking for investigating AChE inhibitory activity.

DOI: [10.7717/peerj.2322/fig-1](https://doi.org/10.7717/peerj.2322/fig-1)

In this study, the data is extracted from the ChEMBL database with the IC50 as the Standard Type. Then explore the data using the Lipinski descriptor and process them using the PaDEL descriptor. The purpose of this study is to find a good prediction model to determine if a molecule would react to Entamoeba histolytica organism.

# Dataset

# Collecting the data

# The dataset of molecules against Entamoeba histolytica was compiled from the ChEMBL database which is comprised of a total number of 1,672 records. SMILES notations of the compounds were curated with ChemAxon’s Standardizer [7] using the same parameter settings as described in our previous study [8]. The initial data set was assembled from several bioactivity measurement units including (in order of decreasing data size) IC50, Ki, % activity, % inhibition, MIC, EC50, etc. IC50 was selected for further investigation as it constituted the largest subset with 1,672 compounds.

# Exploring the data with Lipinski descriptor

Christopher Lipinski, a scientist at Pfizer, came up with a set of rule-of-thumb for evaluating the drug-likeness of compounds. Such drug-likeness is based on Absorption, Distribution, Metabolism, and Excretion (ADME), also known as the pharmacokinetic profile. Lipinski analysed all orally active FDA-approved drugs in the formulation of what is to be known as the Rule-of-Five or Lipinski's Rule [9].

Lipinski's Rule stated the following:

* Molecular weight < 500 Dalton
* Octanol-water partition coefficient (LogP) < 5
* Hydrogen bond donors < 5
* Hydrogen bond acceptors < 10

| **MW** | **LogP** | **NumHDonors** | **NumHAcceptors** |
| --- | --- | --- | --- |
| 186.136 | 2.58170 | 1.0 | 1.0 |
| 220.581 | 3.23510 | 1.0 | 1.0 |
| 171.156 | 0.09202 | 1.0 | 5.0 |
| 234.608 | 3.24550 | 0.0 | 2.0 |
| 200.163 | 2.59210 | 0.0 | 2.0 |

### Figure 2: Dataset sample after processed by the Lipinski Descriptor.

# Convert to IC50 to pIC50

To allow IC50 data to be more uniformly distributed, we will convert IC50 to the

negative logarithmic scale which essentially is -log10(IC50).

The custom function pIC50() will accept a DataFrame as input and will:

* Take the IC50 values from the **standard\_value** column and converts them from nM to M by multiplying the value by 10^{-9}
* Take the molar value and apply -log10
* Delete the **standard\_value** column and create a new pIC50 column

# SMILES pre-processing

The Canonical SMILES were pre-processed by applying sequential filters to remove stereochemistry, salts, and molecules with undesirable atoms or groups[10]. SMILES strings >100 symbols in length were removed, as ∼97% of the dataset consists of SMILES strings with <100 symbols [11]. The dataset was then canonicalized to remove redundant small molecules. The RDKit library in Python was used for dataset pre-processing.



|  |  |
| --- | --- |
| **Canonical SMILES** | **Pre-processed Canonical SMILES** |
| CCC/N=C(\S)N/N=C/c1ccc([N+](=O)[O-])s1 | CCCN=C(S)NN=Cc1ccc([N+](=O)[O-])s1 |
| O=[N+]([O-])c1ccc(/C=N/N=C(\S)NC2CCCCC2)s1 | CCCCN=C(S)NN=Cc1ccc([N+](=O)[O-])s1 |
| CC(C)/N=C(\S)N/N=C/c1ccc([N+](=O)[O-])s1 | CCC(C)N(C(S)=NN=Cc1ccc([N+](=O)[O-])s1)C(C)CC |
| CCCC/N=C(\S)N/N=C/c1ccc([N+](=O)[O-])s1 | CCC(C)N=C(S)NN=Cc1ccc([N+](=O)[O-])s1 |

### Figure 4: Canonical SMILES comparison before and after the Salt has been removed.

# PaDEL Descriptor

PaDEL-Descriptor is a software used to calculate molecular descriptors and fingerprints. The software currently calculates 797 descriptors (663 1D, 2D descriptors, and 134 3D descriptors) and 10 types of fingerprints. These descriptors and fingerprints are calculated mainly using The Chemistry Development Kit [13].

# Feature Selection

The pre-processed Canonical SMILES are processed using PaDEL Descriptor to get the fingerprints which are later used for the modeling. The fingerprint data has 882 columns so to maximize the modeling, I removed the low variance using the Variance Threshold from the sklearn feature selection.

# Data Exploratory

Taking a look at pIC50 values, the actives and inactives displayed statistically significant differences, which is to be expected since threshold values (IC50 < 1,000 nM = Actives while IC50 > 10,000 nM = Inactives, corresponding to pIC50 > 6 = Actives and pIC50 < 5 = Inactives) were used to define actives and inactives.

Chart, scatter chart

Description automatically generated

### Figure 5: LogP and MW compared on scatter plot.

This analysis uses the Mann-Whitney U test, which is also known as the Wilcoxon rank sum test, tests for differences between two groups on a single, ordinal variable with no specific distribution (Mann & Whitney, 1947; Wilcoxon, 1945). In contrast, the independent samples t-test, which is also a test of two groups, requires the single variable to be measured at the interval or ratio level, rather than the ordinal level, and to be normally distributed. We accordingly refer to the Mann-Whitney U test as the nonparametric version of the parametric t-test. Both tests require two independently sampled groups and assess whether two groups differ on a single, continuous variable. The two tests, however, differ in the assumed distribution. A nonparametric test assumes no specific distribution, whereas a parametric test assumes a specific distribution [12].

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

Chart, box and whisker chart

Description automatically generatedChart, box and whisker chart

Description automatically generated

### Figure 6: Lipinski Descriptor’s output exploratory.

Out of the 4 Lipinski's descriptors (MW, LogP, NumHDonors and NumHAcceptors), only LogP exhibited no difference between the actives and inactives while the other 3 descriptors (MW, NumHDonors and NumHAcceptors) shows statistically significant difference between actives and inactives.

# Method

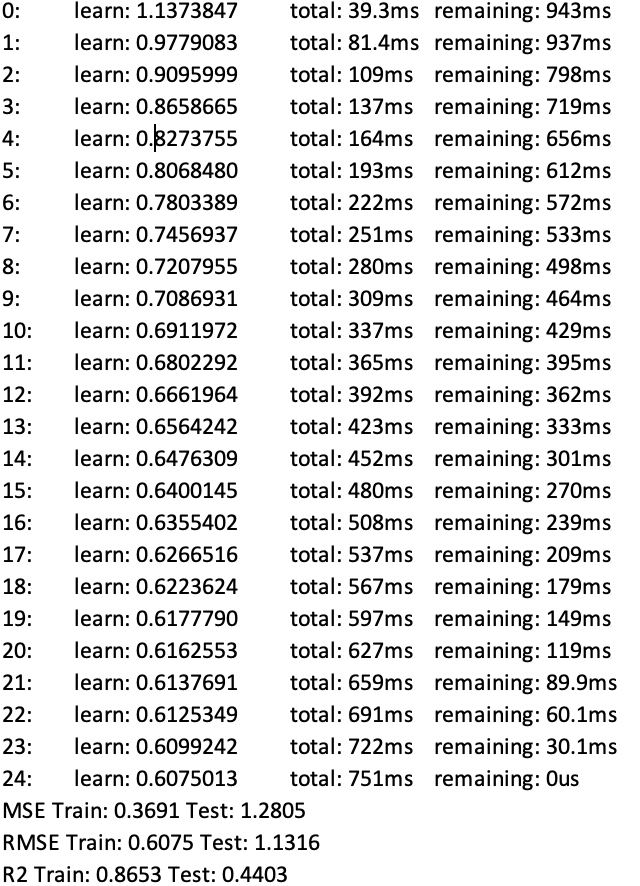
# GridCV

The hyperparameter optimization approaches Grid search were used to optimize the parameters of Random Forest Regressor and Gradient Boosting Regressor.

# CatBoost Regressor

CatBoost is an open-sourced gradient boosting library that handles categorical features and takes advantage of dealing with them during training as opposed to pre-processing time. Another advantage of the algorithm is that it uses a new schema for calculating leaf values when selecting the tree structure, which helps to reduce overfitting[14].

The model ran through 25 iterations, 1 learning rate, and 10 depth. From the said process, we get the following result:



### Figure 7: CatBoost process.

# Random Forest Regressor

Random forests are a scheme proposed by Leo Breiman in the 2000s for building a predictor ensemble with a set of decision trees that grow in randomly selected subspaces of data. Despite growing interest and practical use, there has been little exploration of the statistical properties of random forests, and little is known about the mathematical forces driving the algorithm [15,16].

# Gradient Boosting Regressor

Gradient boosting is a strategy of combining weak predictors into a strong predictors. The algorithm designer can select the base learner according to specific applications. Many researchers have tried to combine gradient boosting with common machine learning algorithms to solve their problems [17].

# Metrics

In this study, I use MSE to measure the average of the squares of the error as well as RMSE, and Root-mean-square (R2) to measure the differences between values (sample or population values) predicted by a model or an estimator and the values observed [18].

# Results

The result of various models is presented in Figure 7. The baseline of the model is the mean of the data, which is 5.43. The study uses various models to do better than the base model.

Although **CatBoost** is said to perform well on categorical features, the testing data result was far worst compared to the train data result.

**Forest Regressor** performs well on the training dataset. With the help of Grid CV to find the best parameter, the model has high performance (R2= 0.84). However, the result of the testing dataset is rather lower, although it surpasses the baseline model.

**Gradient Boost** has a similar performance as Forest Regressor, with a small difference in the R2. Both on the training dataset and the testing dataset.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CatBoost | Forest Regressor | Gradient Boost | Baseline model |
| MSE | 0.37 | 0.45 | 0.45 | 5.4 |
| RMSE | 061 | 0.67 | 0.67 |
| R2 | 8.7 | 8.386 | 8.388 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | CatBoost | Forest Regressor | Gradient Boost | Baseline model |
| MSE | 1.28 | 0.994 | 0.98 | 5.4 |
| RMSE | 1.13 | 0.997 | 0.993 |
| R2 | 4.40 | 5.651 | 5.684 |

### Figure 7: Various models results

Chart, scatter chart

Description automatically generatedChart, scatter chart

Description automatically generated

### Figure 8: Training and Testing data result of Gradient Boost on scatter plot.

# Conclusion

On the molecule dataset of 893 individuals and 175 features, we achieve a somewhat discouraging result, especially on the testing dataset. A larger dataset with more individuals and more variables included may improve this study. Here are some improvements that could be implemented in the future study:

* Exploring the data using non-Decision Tree based methods
* Using the Lipinski Descriptor’s dataset for the modelling

**References**

[1] Shrivastav MT, Malik Z and Somlata (2021) Revisiting Drug Development Against the Neglected Tropical Disease, Amebiasis. Front. Cell. Infect. Microbiol. 10:628257. doi: 10.3389/fcimb.2020.628257

[2] Jeelani, G., and Nozaki, T. (2014). Metabolomic analysis of Entamoeba: applications and implications. Curr. Opin. Microbiol. (2014) 20, 118–124. doi: 10.1016/j.mib.2014.05.016

[3] Shirley, D.-A. T., Farr, L., Watanabe, K., and Moonah, S. (2018). A Review of the Global Burden, New Diagnostics, and Current Therapeutics for Amebiasis. Open Forum Infect. Dis. 5 (7), ofy161. doi: 10.1093/ofid/ofy161

[4] Ben Ayed, S. (2015). Entamoeba histolytica. Global Water Pathogen Project (Michigan State University: UNESCO). Available at: https://www. waterpathogens.org/book/entamoeba-histolytica.

[5] Haque, R., Mondal, D., Duggal, P., Kabir, M., Roy, S., Farr, B. M., et al. (2006). Entamoeba histolytica Infection in Children and Protection from Subsequent Amebiasis. Infect. Immun. 74 (2), 904–909. doi: 10.1128/IAI.74.2.904-909.2006

[6] Isarankura-Na-Ayudhya, Chartchalerm. Naenna, Thanakorn. Nantasenamat, Chanin. Prachayasittikul, Virapong. A practical overview of quantitative structure-activity relationship. 2009-07-08T08:22:46Z

[7] ChemAxon Kft. 2015. Standardizer. Version 15.9.14.0

[8] Simeon S, Möller R, Almgren D, Li H, Phanus-umporn C, Prachayasittikul V, Bülow L, Nantasenamat C. 2016. Unraveling the origin of splice switching activity of hemoglobin β-globin gene modulators via QSAR modeling. Chemometrics and Intelligent Laboratory Systems 151:51-60

[9] Lipinski, C. A., Lombardo, F., Dominy, B. W., & Feeney, P. J. (2012). Experimental and computational approaches to estimate solubility and permeability in drug discovery and development settings. Advanced drug delivery reviews, 64, 4-17.

[10] Olivecrona M, Blaschke T, Engkvist O, Chen H. Molecular de novo design through deep reinforcement learning. J. Cheminformatics 9(1), 48 (2017).Crossref, Medline, Google Scholar

[11] Popova M, Isayev O, Tropsha A. Deep reinforcement learning for de novo drug design. Sci. Adv. 4(7), eaap7885 (2018). •• Recent developments in AI-based drug design.Crossref, Medline, CAS, Google Scholar

[12] McKnight, P.E. and Najab, J. (2010). Mann-Whitney U Test. In The Corsini Encyclopedia of Psychology (eds I.B. Weiner and W.E. Craighead). <https://doi.org/10.1002/9780470479216.corpsy0524>

[13] Yap, C.W. (2011), PaDEL-descriptor: An open source software to calculate molecular descriptors and fingerprints. J. Comput. Chem., 32: 1466-1474. <https://doi.org/10.1002/jcc.21707>

[14] Dorogush , Anna Veronika. Ershov, Vasily. Gulin, Andrey. Workshop on ML Systems at NIPS (2017)

[15] Pedregosa et al. Scikit-learn: Machine Learning in Python (2011). JMLR 12, pp. 2825-2830.

[16] Biau, Ge ́rard. Analysis of a Random Forests Model (2012). JMLR 12, 2012

[17] Li, Wen. Wang, Wei. Huo, Wenjun. RegBoost: a gradient boosted multivariate regression algorithm (2020)

[18] Wackerly, Dennis; Mendenhall, William; Scheaffer, Richard L. (2008). Mathematical Statistics with Applications (7 ed.). Belmont, CA, USA: Thomson Higher Education. ISBN 978-0-495-38508-0.