Predicting Anti-Malarial Potency from Molecular Descriptors Using Gamma Regression

Ekadanta Soedarmadji

Los Angeles Unified School District, CA

Abstract

Malaria remains a global health challenge, demanding effective drug development. Key measures of drug efficacy include inhibition percentages and potency values. This study applies gamma regression to model these outcomes, utilizing its suitability for continuous, positively-skewed data.

Keywords: malaria, drug efficacy, inhibition percentage, potency value, gamma regression

1 Introduction

1.1 Background

Malaria remains one of the most persistent global health challenges, and the discovery of new antimalarial compounds continues to be a major priority in drug development. Quantitative Structure–Activity Relationship (QSAR) modeling provides a computational framework for linking the chemical features of molecules to their biological activity, enabling researchers to identify promising drug candidates more efficiently. Modern QSAR studies often rely on large, experimentally derived datasets to train statistical or machine learning models capable of predicting compound potency and guiding early-stage drug discovery.

In this study, we develop a QSAR model using a publicly available dataset from Kaggle that compiles experimental antimalarial activity data from the ChEMBL database [1].

ChEMBL is a manually curated repository of bioactive, drug-like molecules, widely used in computational chemistry and drug discovery. The dataset analyzed here focuses specifically on the Plasmodium falciparum 3D7 strain, a common reference strain used to evaluate the effectiveness of antimalarial compounds.

Two measures of biological activity are considered: inhibition percentage (PCT_IHB) and potency (pIC50). The PCT_IHB value indicates the percentage of malaria parasites inhibited by a compound at a given concentration, with higher values corresponding to stronger activity. The pIC50 value represents the negative logarithm of the IC50, the concentration required to inhibit 50% of the parasites; thus, higher pIC50 values indicate higher potency, as less compound is needed to achieve the same inhibitory effect.

1.2 Literature Review

Bellamy et al. address the challenge of noisy datasets by applying complex optimization methods [2]. The authors efficiently find active compounds despite noise in the dataset. Their use of the ChEMBL dataset and the task of predicting pIC50 values provides a strong baseline for comparison.

Bosc et al. apply molecular fingerprinting to accurately predict antimalarial activity [3]. This method is well-suited for machine learning applications, but results in less interpretable models compared to using raw descriptors. Their use of the ChEMBL database provides a strong baseline for comparison.

Mervin et al. address a core issue: models that do not account for experimental noise risk overfitting [4]. This is highly relevant because ChEMBL is an experimental database. By choosing gamma regression for this study, our methodology aligns with their findings, as gamma regression does not assume constant variance across the dataset. Accounting for experimental noise and uncertainty is one of the best ways to avoid overfitting.

Traditional QSAR approaches in antimalarial discovery often focus on rational design within a single chemical family. For example, Santos et al. successfully use multiple linear regression to optimize artemisinin derivatives, but their study is limited to a small set of 20 closely related compounds [5]. In contrast, the present work leverages large-scale public data from ChEMBL containing thousands of structurally diverse compounds, aiming to build a

more generalizable model for potency prediction across different chemical scaffolds.

Zhang et al. successfully use ChEMBL data for classification tasks [6]. Our work focuses on building interpretable regression models to predict continuous potency values, providing a different and complementary level of insight for drug discovery.

Borba et al. successfully apply explainable machine learning QSAR models to discover new multi-stage antimalarials [7]. Aligning with their focus on interpretability, our work adopts a similar principle but applies a simpler, statistically rigorous gamma regression to predict continuous potency specifically for the P. falciparum 3D7 strain.

Lin et al. experiment with a wide variety of prediction methods, establishing a strong baseline for antimalarial activity prediction [8]. While the authors explore complex deep learning architectures, their work shows that simpler, fingerprint-based machine learning models, such as Random Forest are more robust and achieve the best results. This study builds directly on this finding by applying a simple, interpretable regression approach (gamma regression) to a focused dataset for the 3D7 strain.

Roche-Lima et al. develop a tool to predict synergistic drug combinations, demonstrating that machine learning effectively predicts malaria data [9]. Notably, for the 3D7 strain, the authors find that a Random Forest model achieves the best performance using a dataset of only 1,540 samples. This result supports this study, which suggests that robust models can be built for the 3D7 strain. The larger dataset of 13,102 samples used in this study is expected to provide higher accuracy and improve predictions of antimalarial drugs.

Wang et al. argue that a model's high accuracy score does not necessarily indicate good molecule-ranking ability [10]. The authors propose training a bivariate model to rank drugs, which they show is superior for identifying better-performing candidates in drug discovery. This study uses a more standard regression approach, which can serve as a baseline for this new drug-ranking methodology.

Recent studies utilize more complex computational methods, combining active learning with QSAR and molecular docking, to identify new inhibitors against specific targets such as PfHsp90 [11]. While effective for hit identification, such approaches can be computationally intensive and yield models with limited interpretability.

2 Theoretical Framework for Gamma Regression

Suppose we aim to model the relationship between an outcome variable y and a set of descriptors $x_1, \ldots, x_k$, with observations available for n individuals. If the outcome y is continuous, strictly positive, and right-skewed (i.e., has a long right tail), a gamma regression model may be appropriate. In this model, y is assumed to follow a gamma distribution with density [12]

$$f(y) = \frac{y^{\alpha-1}}{\Gamma(\alpha)\,\mu^{\alpha}}\, e^{-y/\mu}, \quad \alpha > 0,\ \mu > 0,\ y > 0,$$

where the expected value of y is

$$\mathbb{E}[y] = \alpha\mu = \exp(\beta_0 + \beta_1 x_1 + \cdots + \beta_k x_k),$$

and α is a positive scalar known as the dispersion parameter. The model parameters $\beta_0, \ldots, \beta_k$ and α are estimated via maximum likelihood. The log-likelihood function for the sample $(x_{1i}, \ldots, x_{ki}, y_i),\ i = 1, \ldots, n$, is

$$\ln L(\beta_0, \ldots, \beta_k, \alpha) = \sum_{i=1}^{n} \Big[(\alpha - 1)\ln y_i - \frac{y_i}{\mu_i} - \ln\Gamma(\alpha) - \alpha \ln \mu_i\Big],$$

where

$$\mu_i = \exp(\beta_0 + \beta_1 x_{1i} + \cdots + \beta_k x_{ki}).$$

The fitted model can be expressed as

$$\widehat{\mathbb{E}}\,[y] \;=\; \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1 + \cdots + \widehat{\beta}_k x_k\},$$

and the estimated dispersion parameter $\hat{\alpha}$ should also be reported.

The regression coefficients in a gamma regression admit the following interpretation. If a descriptor x_1 is numeric, then $(\exp\{\widehat{\beta}_1\} - 1) \cdot 100\%$ represents the percent change in the estimated mean outcome for a one-unit increase in x_1, assuming all the other descriptors are unchanged.

The fitted model can be used for prediction as follows. For a new set of descriptors $x_1^0, \ldots, x_k^0$, the predicted outcome variable is given by:

$$y^0 = \exp\{\widehat{\beta}_0 + \widehat{\beta}_1 x_1^0 + \cdots + \widehat{\beta}_k x_k^0\}.$$

3 Applications and Results

3.1 Data Preprocessing for Analysis

The dataset initially contains 13,119 samples described by 23 molecular descriptors. Data cleaning removed four descriptors that exhibited either more than 30% missing values or non-numeric entries. Seventeen rows with missing values across the remaining descriptors were also removed, and for one descriptor with fewer than 10% missing entries, median imputation was applied. After preprocessing, the resulting dataset consists of 19 descriptors, 13,102 samples, and two target variables. Figures 1 and 2 show the distributions of the target variables before transformation.

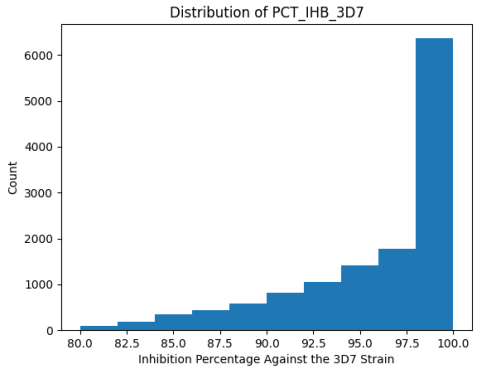


Figure 1: PCT_IHB_3D7 before transformation

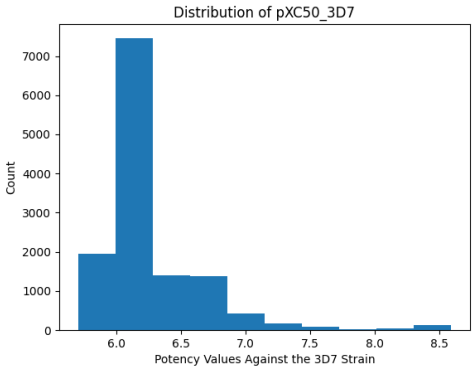


Figure 2: pIC50_3D7 before transformation

Basic linear transformations were applied to the outcome variables to ensure they were right-skewed and strictly positive, allowing for appropriate modeling with gamma regression. The transformations were as follows:

$$\texttt{PCT\_IHB\_3D7\_transformed} = \max(\texttt{PCT\_IHB\_3D7}) - \texttt{PCT\_IHB\_3D7} + 0.01,$$

and

$$\texttt{pIC50\_3D7\_transformed} = \texttt{pIC50\_3D7} - \min(\texttt{pIC50\_3D7}) + 0.01.$$

The histograms of the transformed outcome variables are given in Figures 3 and 4.

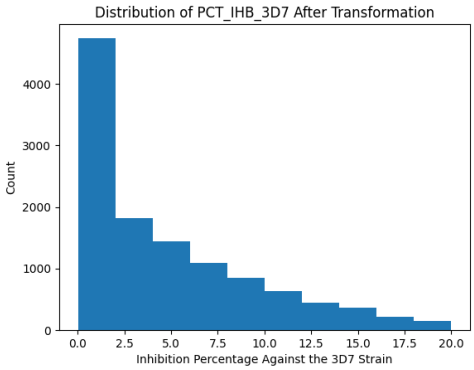


Figure 3: PCT_IHB_3D7 after transformation

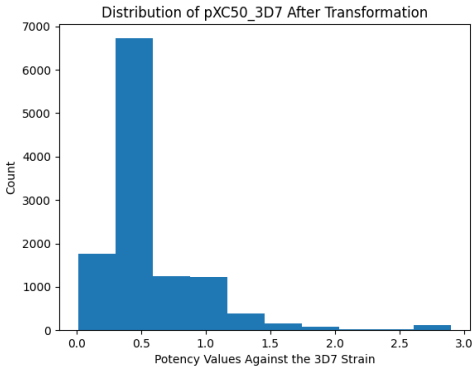


Figure 4: pIC50_3D7 after transformation

Details for all descriptors are provided in Table A in the Appendix.

3.2 Methodology

The training process consisted of three major steps: redundancy elimination, backward elimination, and gamma regression modeling. All regression analyses were conducted using the statsmodels [13] Python package's Generalized Linear Model framework [14]. Statistical significance tests for dispersion estimates were performed using SciPy's implementation of the Student's t-distribution [15].

Redundancy elimination was first applied to remove highly correlated descriptors using a Pearson correlation threshold of 0.8. For each pair exceeding this threshold, the descriptor with the higher variance was retained. This procedure reduced the descriptor set to 11 variables.

The dataset was then split into a 90–10 train–test partition, with 11,792 samples used for training and 1,310 reserved for testing. Descriptor selection was refined using backward elimination, reducing the set from 11 to 7 descriptors. Starting with all 11 descriptors, the least statistically significant descriptor (based on p-values from the gamma regression) was removed at each step until 7 remained.

Model performance was evaluated using percentage error, defined as the proportion of predictions falling within 5%, 10%, 15%, 20%, and 25% of the true target values. This metric offers an intuitive measure of predictive accuracy. As the error threshold increases, accuracy scores naturally increase, reaching 100% when all predictions fall within the specified range.

3.3 Results

Using the training set, the outcome variables `PCT_IHB_3D7_transformed` and `pIC50_3D7_transformed` were regressed on the full set of 19 descriptors using a gamma regression approach. Through backward elimination, each model was reduced to seven descriptors (not identical across the two regressions, though with considerable overlap). The reduced models were then applied to the testing set to generate predicted values. The results are presented in Tables 1 and 2 below. The fitted models achieved high accuracy across multiple tolerance levels: both produced over 90% of predictions within 10% of the true target values and reached accuracy scores of 99% at the 25% threshold.

Table 1: Prediction accuracy within ± x% for the `PCT_IHB_3D7_transformed` model

Threshold, x%	**Prediction Accuracy**
5%	80%
10%	93%
15%	98%
20%	99%
25%	100%

Table 2: Prediction accuracy within ± x% for the `pIC50_3D7_transformed` model

Threshold, x%	**Prediction Accuracy**
5%	72%
10%	94%
15%	98%
20%	98%
25%	99%

Next, we present and discuss the results of modeling the transformed inhibition percentage `PCT_IHB_3D7_transformed`. The estimated regression coefficients, along with the corresponding p-values for testing equality to zero, are provided in Table 3 below.

Table 3: Output in the `PCT_IHB_3D7_transformed` model

name	**coef**	**std err**	p**-value**
Intercept	1.2127	0.138	0.000
aromatic_rings	-0.1169	0.013	0.000
cx_logd	0.3988	0.080	0.000
cx_most_bpka	-0.2402	0.056	0.000
hbd	-0.1149	0.013	0.000
psa	0.5229	0.055	0.000
qed_weighted	0.0675	0.010	0.000
rtb	0.2427	0.048	0.000
Dispersion	1.3153	0.017	0.000

The fitted model is as follows:

$$\hat{\mathbb{E}}\big[\texttt{PCT\_IHB\_3D7\_transformed}\big] = 1.2127 - 0.1169 \cdot \texttt{aromatic\_rings} + 0.3988 \cdot \texttt{cx\_logd}$$

$$- 0.2402 \cdot \texttt{cx\_most\_bpka} - 0.1149 \cdot \texttt{hbd} + 0.5229 \cdot \texttt{psa} + 0.0675 \cdot \texttt{qed\_weighted} + 0.2427 \cdot \texttt{rtb}.$$

The estimated slopes can be interpreted through percent changes in the estimated mean transformed inhibition percentage per one-unit increase in each descriptor, computed as $(e^{\hat{\beta}} - 1) \cdot 100\%$. Specifically:

- `aromatic_rings`: $(e^{-0.1169} - 1) \cdot 100\% = -11.03\%$, a decrease of 11.03%,
- `cx_logd`: $(e^{0.3988} - 1) \cdot 100\% = 49.00\%$, an increase of 49.00%,
- `cx_most_bpka`: $(e^{-0.2402} - 1) \cdot 100\% = -21.35\%$, a decrease of 21.35%,
- `hbd`: $(e^{-0.1149} - 1) \cdot 100\% = -10.85\%$, a decrease of 10.85%,
- `psa`: $(e^{0.5229} - 1) \cdot 100\% = 68.69\%$, an increase of 68.69%,
- `qed_weighted`: $(e^{0.0675} - 1) \cdot 100\% = 6.98\%$, and increase of 6.98%,
- `rtb`: $(e^{0.2427} - 1) \cdot 100\% = 27.47\%$, an increase of 27.47%.

Lastly, we present the analysis of the transformed potency measure `pIC50_3D7_transformed`.

Table 4 reports the estimated regression coefficients together with their associated p-values for assessing whether each coefficient differs significantly from zero.

Table 4: Output in the `pIC50_3D7_transformed` model

name	**coef**	**std err**	p**-value**
Intercept	-0.2896	0.041	0.000
cx_logd	-0.4247	0.071	0.000
cx_logp	0.4645	0.091	0.000
cx_most_bpka	-0.1267	0.033	0.000
full_mwt	-0.0202	0.008	0.017
np_likeness_score	0.2164	0.010	0.000
psa	0.3976	0.028	0.000
rtb	-0.2906	0.027	0.000
Dispersion	0.3941	0.005	0.000

The fitted model is given by:

$$\hat{\mathbb{E}}\big[\texttt{pIC50\_3D7\_transformed}\big] = -0.2896 - 0.4247 \cdot \texttt{cx\_logd} + 0.4645 \cdot \texttt{cx\_logp}$$

$$- 0.1267 \cdot \texttt{cx\_most\_bpka} - 0.0202 \cdot \texttt{full\_mwt} + 0.2164 \cdot \texttt{np\_likeness\_score}$$

$$+ 0.3976 \cdot \texttt{psa} - 0.2906 \cdot \texttt{rtb}.$$

The estimated slopes can be interpreted as percent changes in the estimated mean potency per one-unit increase in each descriptor. Namely:

- `cx_logd`: $(e^{-0.4247} - 1) \cdot 100\% = -34.60\%$, a decrease of 34.60%,
- `cx_logp`: $(e^{0.4645} - 1) \cdot 100\% = 59.12\%$, an increase of 59.12%,
- `cx_most_bpka`: $(e^{-0.1267} - 1) \cdot 100\% = -11.90\%$, a decrease of 11.90%,
- `full_mwt`: $(e^{-0.0202} - 1) \cdot 100\% = -2.00\%$, a decrease of 2.00%,
- `np_likeness_score`: $(e^{0.2164} - 1) \cdot 100\% = 24.16\%$, an increase of 24.16%,
- `psa`: $(e^{0.3976} - 1) \cdot 100\% = 48.82\%$, an increase of 48.82%,

- `rtb`: $(e^{-0.2906} - 1) \cdot 100\% = -25.22\%$, a decrease of 25.22%.

4 Summary and Discussion

The goal of this study was to build a highly interpretable, robust, and accurate QSAR model for predicting antimalarial inhibition percentages and potency values. We achieved this goal with both prediction models— for inhibition percentage and potency value—successfully predicting over 90% of the samples within 10% of the actual values using only seven molecular descriptors.

Modern research studies [3] achieve highly accurate QSAR models by utilizing a method known as fingerprinting. This approach prioritizes prediction accuracy over model interpretability, which is essential in drug discovery. More sophisticated optimization algorithms [2] have also been employed to address the issue of noisy datasets. While these methods have their advantages, we showed that high predictive accuracy can be achieved without sacrificing interpretability. Our approach of using raw descriptors contrasts with the "black box" nature of many machine learning methods.

A major challenge when training QSAR models is noise from real-world datasets such as ChEMBL. Mervin et al. address this problem using probabilistic random forests [4]. Our choice of gamma regression tackles this by allowing the variance to vary across the data, rather than assuming it is constant. Additionally, whereas traditional QSAR models often performed well only within narrow chemical families [5], we trained on over 10,000 ChEMBL compounds, aiming to build a highly generalizable model across chemical scaffolds.

In conclusion, this study provides a model that not only yields high predictive accuracy, comparable to more complex methods [2, 3], but is also highly interpretable and resistant to noise in the dataset [4]. We went beyond predicting within a small scope of molecules [5] to achieve highly generalizable models. Furthermore, our study extends beyond classification tasks [6] by directly predicting the potency values of each compound. This work lays the foundation for using regression-based virtual screening of large compound databases to identify new antimalarial compounds.

4.1 Future Research Direction

While this study established a baseline using gamma regression, future work could explore more flexible machine learning approaches, such as random forests or gradient boosting, to capture potential non-linear relationships. Expanding the model to include *P. falciparum* strains beyond 3D7, including more drug-resistant variants, would enhance the predictive framework's relevance for identifying strategies to inhibit drug-resistant malaria parasites. The model could also be applied in real-world drug discovery efforts, helping to optimize resource allocation and improve efficiency.

Supplemental Materials

The following GitHub repository contains the raw data, processed data, environment setup code, and training/scoring scripts:

`https://github.com/dinodev24/antimalarial-potency-prediction`

It also provides code to fetch the full set of descriptors from the ChEMBL database; however, this step is optional since the repository already includes the processed dataset.

Acknowledgments

I would like to thank Dr. Olga Korosteleva, Professor of Statistics at California State University, Long Beach, for providing me with so much support and guidance throughout this project. I am also thankful to my high school math teacher, Ms. D Vu, for her support throughout my research experience. Additionally, I would like to thank Dr. Oleg Gleizer and all UCLA Math Circle instructors for teaching me advanced math topics, including statistics and probability, which were very helpful towards this research project. Finally, I would like to thank my family for providing support and guidance through this research journey.

References

[1] Zdrazil, B., Felix, E., Hunter, F., et al. The chembl database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Research*, 52(D1):D1180–D1192, 11 2023.

[2] Bellamy, H., Rehim, A. A., Orhobor, O. I., and King, R. Batched bayesian optimization for drug design in noisy environments. *Journal of Chemical Information and Modeling*, 62, 2022.

[3] Bosc, N., Felix, E., and Arcila, R. Maip: a web service for predicting blood-stage malaria inhibitors. *Journal of Cheminformatics*, 13, 2021.

[4] Mervin, L. H., Trapotsi, M., Afzal, A. M., Barrett, I. P., Bender, A., and Engkvist, O. Probabilistic random forest improves bioactivity predictions close to the classification threshold by taking into account experimental uncertainty. *Journal of Cheminformatics*, 13, 2021.

[5] Santos, C. B. R., Lobato, C. C., Braga, F. S., et al. Rational design of antimalarial drugs using molecular modeling and statistical analysis. *Current Pharmaceutical Design*, 21, 2015.

[6] Zhang, L., Fourches, D., Sedykh, A., et al. Discovery of novel antimalarial compounds enabled by qsar-based virtual screening. *Journal of Chemical Information and Modeling*, 53:475–492, 2013.

[7] Borba, J. V. B., Salazar-Alvarez, L. C., Ferreira, L. T., et al. Innovative multistage ml-qsar models for malaria: From data to discovery. *ACS Medicinal Chemistry Letters*, 15, 2024.

[8] Lin, M., Cai, J., Wei, Y., Peng, X., Luo, Q., Li, B., Chen, Y., and Wang, L. Malariaflow: A comprehensive deep learning platform for multistage phenotypic antimalarial drug discovery. *European Journal of Medicinal Chemistry*, 277:116776, 2024.

[9] Roche-Lima, A., Rosado-Quiñones, A. M., Feliu-Maldonado, R. A., et al. Antimalarial drug combination predictions using the machine learning synergy predictor (mlsypred©) tool. *Acta Parasitologica*, 69(1):415–425, 2024.

[10] Wang, Y. and King, R.D. Extrapolation is not the same as interpolation. *Machine Learning*, 113, 2024.

[11] Matlhodi, T., Makatsela, L. P., Dongola, T. H., Simelane, M. B. C., Shonhai, A., Gumede, N. J., et al. Auto qsar-based active learning docking for hit identification of potential inhibitors of Plasmodium falciparum Hsp90 as antimalarial agents. *PLoS ONE*, 19(11):e0308969, 2024.

[12] Korosteleva, O. Advanced regression models with sas and r. pp. 48–54, N.D.

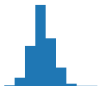
[13] Seabold, S. and Perktold, J. Statsmodels: Econometric and statistical modeling with python. *Proceedings of the 9th Python in Science Conference*, 2010, 01 2010.

[14] Nelder, J. A. and Wedderburn, R. W. M. Generalized linear models. *Journal of the Royal Statistical Society. Series A (General)*, 135(3):370–384, 1972.

[15] Virtanen, P., Gommers, R., Oliphant, T. E., et al. Scipy 1.0: fundamental algorithms for scientific computing in python. *Nature Methods*, 17:261–272, 2020.

Appendix

Table A. Information on descriptors selected for modeling `PCT_IHB_3D7_transformed` and `pIC50_3D7_transformed`

Name	Description	Histogram
Aromatic Rings, `aromatic_rings`	Number of flat, ring-shaped, stable parts of the molecule. Higher values may improve binding to malaria parasites through stronger molecular interactions.	
LogD (Dispersion coefficient, pH = 7.4), `cx_logd`	Measures the molecule's mobility in oil/water. Moderate values facilitate the compound reaching malaria parasites more effectively.	
LogP (Partition coefficient), `cx_logp`	Measures the compound's distribution between oil and water phases. Moderate values enhance transport to malaria parasites.	
Most basic pKa, `cx_most_bpka`	Indicates how basic the molecule is. Higher values may disrupt acidic environments necessary for malaria parasite survival.	
Full Molecular Weight, `full_mwt`	Total weight of the compound, including the molecule and attached salts. Moderate values balance binding ability and mobility.	
Number of Hydrogen Bond Donors, `hbd`	Number of hydrogen atoms capable of forming bonds. Moderate values improve binding to malaria while supporting absorption through cell membranes.	
Natural Product Likeness Score, `np_likeness_score`	Measures how "nature-like" the molecule is. Higher scores indicate potentially greater efficacy and lower side-effects.	
Polar Surface Area, `psa`	Total surface area capable of forming hydrogen bonds with biological targets. Larger areas can improve binding and inhibitory effects against malaria.	
Weighted Quantitative Estimate of Drug-likeness, `qed_weighted`	A score (0 to 1) estimating how "drug-like" a molecule is. Higher scores indicate safer and more effective compounds.	
Rotatable Bonds, `rtb`	Number of bonds in the molecule that can freely rotate. Fewer rotatable bonds usually improve cell membrane penetration and compound stability.