chemistry proceedings

Proceeding Paper

A Robust Regression-Based Modeling to Predict Antiplasmodial Activity of Thiazolyl-pyrimidine Hybrid Derivatives against *Plasmodium falciparum* †

Kevin S. Umoette *, Charles O. Nnadi * and Wilfred O. Obonga *

Department of Pharmaceutical and Medicinal Chemistry, Faculty of Pharmaceutical Sciences, University of Nigeria Nsukka, 410001 Enugu, Nigeria; umoettekevin@gmail.com (K.S.U.); charles.nnadi@unn.edu.ng (C.O.N.); wilfred.obonga@unn.edu.ng (W.O.O.); Tel.: +234-7038356262 (K.S.U.); +234-8064947734 (C.O.N.); +234-8033305022 (W.O.O.)

† Presented at the 27th International Electronic Conference on Synthetic Organic Chemistry (ECSOC-27), 15–30 November 2023; Available online: https://ecsoc-27.sciforum.net/.

Abstract: Thiazolyl-pyrimidine hybrid plays significant roles in the biological activities and SAR of thiazolylpyrimidines (Tzpd), thiazolopyrimidines and thienopyrimidines due the combination of the thiazole and pyrimidine pharmacophores. The study developed regression-based models for the prediction of antiplasmodial activity of 43 Tzpd hybrid obtained from the ChEMBL database. The molecular descriptors (145 features) were scaled down to 6 using the recursive feature elimination. The X- and Y-matrix were split into 34 train and 9 test sets using a split ratio of 0.20. Regression models were built using scikit-learn algorithms: multiple linear regression (MLR), k-Nearest Neighbours (kNN), Support Vector Regressor (SVR) and Random Forest Regressor (RFR) to predict the pIC_{50} of the test set. The models were evaluated using R^2, mean squared error (MSE), mean absolute error (MAE), root mean squared error (RMSE), *p*-values, *F*-statistic, and variance inflation factor (VIF). Of the 145 features calculated for the 43 Tzpd, 6 molecular features: FCASA-, MNDO_LUMO, E_str, vsurf_HB1, vsurf_G and vsurf_DD12 ($p < 0.05$; VIF < 5) were found to significantly influence the antiplasmodial activity. Five-fold cross-validation performance scores of MLR, kNN, SVR, and RFR showed that the performance metrics of MLR (MSE = 0.1453; R^2 = 0.680; MAE = 0.290; RMSE = 0.381; pIC_{50}(predicted) = 8.06 − 0.45vsurf_G + 0.37FCASA– − 0.42MNDO_LUMO − 0.20E_str + 0.30vsurf_HB1 − 0.38vsurf_DD12) outperformed other models. The study developed predictive models and provided insights into the chemical features necessary for the optimization of thiazolyl-pyrimidine to enhance antiplasmodial activity.

Keywords: machine learning; *Plasmodium falciparum*; QSAR; regression; thiazolylpyrimidines

Citation: Umoette, K.S.; Nnadi, C.O.; Obonga, W.O. A Robust Regression-Based Modeling to Predict Antiplasmodial Activity of Thiazolyl-pyrimidine Hybrid Derivatives against *Plasmodium falciparum*. **2023**, *14*, x. https://doi.org/10.3390/xxxxx

Academic Editor(s): Name

Published: 15 November 2023

1. Introduction

Malaria is a disease caused by the parasite of the genus *Plasmodium* and transmitted through the saliva of female anopheles mosquitoes [1]. Sub-Saharan Africa is currently overwhelmed by *P. falciparum*. Several heterocyclic compounds and their derivatives are important chemotherapeutic classes and are still useful singly and in combinations for the treatment of malaria [2]. Various structural modification of heterocycles with improved activities has been reported, and translated to useful drugs [3]. To date, artemisinin-based combination therapy has remained the most potent first-line treatment for *P. falciparum*. The emergence and rapid spread of artemisinin-resistant strains of *P. falciparum* are indications that a continuous search for a more efficacious remedy for malaria is imperative [2]. The combined safety, favourable physicochemical properties and cost-effectiveness of hybrid designs make it a good candidate for structural modifications to overcome resistance and declining efficacy

Different strategies have been put forth to design new chemical entities with optimum pharmacokinetic and pharmacodynamic properties [4]. The QSAR method uses computation modeling to unravel associations between the biological activities and physicochemical properties of chemical substances to create a robust statistical model to predict the biological activities of novel chemical entities [5]. Pyrimidines are important substances in the synthesis of various active molecules that are extensively used in the intermediate skeleton of antiplasmodial and have attracted more attention due to their extensive biological activities including antiviral, antibacterial, antifungal, and insecticidal activities [5]. For example, pyrimidine derivatives bearing a dithioacetal moiety as effective antiviral agents have been reported [6] Thiazolyl-pyrimidine hybrid plays significant roles in the biological activities and SAR of thiazolylpyrimidines (Tzpd), thiazolopyrimidines and thienopyrimidines due the combination of the thiazole and pyrimidine pharmacophores.

This study, therefore, developed a robust model using regression and classification such as knearest neighbours, kNN classifier, support vector classifier, (SVC) and Random Forest Regressor (RFR)) algorithms to; (i) develop a model to predict the pIC_{50} of any untested Tzpd analogues or similar derivatives against *P. falciparum* strains; and (ii) explain SARs of Tzpd derivatives against *P. falciparum* strains.

2. Methods

2.1. Chemical Data Set

The chemical data set comprises 43 derivatives of thiazolyl-pyrimidine hybrids obtained from the ChEMBL database of compounds with antimalarial activity against *Plasmodium falciparum*. The detailed chemical structures and pIC_{50} of the compounds used in this study are shown in the supplementary materials (Figure S1a,b).

2.2. Preparation of Data Set

The SMILES were initially converted to structures to form a molecular database and converted to 3D by energy minimization using the MMFF94x force field. The energy-minimized compounds were subjected to conformational search using LM dynamics [5]. The molecules were then subjected to further energy minimization using the Hamiltonian semi-empirical AM1 MOPAC modules and the resulting conformers were used for further studies.

2.3. Computation of Molecular Descriptors

The molecular fragments of the AM1 energy minimized Tzpd were subjected to both 2D and 3D molecular descriptor calculation using the default settings of the molecular operating environment (MOE) software [7]

2.4. Data Pretreatment

One hundred and forty-five chemical features/descriptors were computed for the compounds and the pIC_{50} was calculated from the negative decadic logarithm of the IC_{50}. The pIC_{50} column (the values to be predicted) formed the Y-matrix, while the rest of the dataset formed the X-matrix. Standardization of the X-matrix was done using the StandardScaler function [8]. It is important to standardize the variables so that they will all have a comparable scale.

2.5. Selection of Relevant Descriptors

Recursive feature elimination (RFE) was used to select significant features using the linear regression function from Skearn for RFE [8]. The number of features considered to build the model was placed at 25 using m > n2. where m is the number of molecules, and n is the number of features

2.6. Data Splitting

The X- and Y-matrix were split into the train (34 molecules) and test (9 molecules) sets using a split ratio of 0.2, where 80% is assigned to the train set and 20% is assigned to the test set. The size of the training dataset was denoted as X-train, Y-train, while the size of the test dataset was X-test, and Y-test. The training set was used to train the model using a fit method, while 9 molecules belonging to the test set were used to validate the models. The hyper-parameters of the models were adjusted on the test dataset to obtain the best hyper-parameter configuration using a random search because their hyper-parameters were continuous

2.7. Regression Modeling

The Statsmodel package of the Python software was used to get the detailed statistics and summary of the model [8,9]. The machine learning scikit-learn algorithms; multiple linear regressor (MLR), k-Nearest Neighbours (kNN), Support Vector Regressor (SVR) and Random Forest Regressor (RFR)) were deployed to predict the pIC_{50} values of the test set compounds. The goal was to discover the best algorithm capable of predicting the activity of untested compounds

2.8. Model Evaluation

Different evaluation metrics such as the coefficient of determination (R^2), mean squared error (MSE), mean absolute error (MAE) and root mean squared error (RMSE) were deployed to assess the performance of the models. The *p*-values, *F*-statistic, and variance inflation factor (VIF) were also used [10].

3. Results and Discussion

3.1. Chemical Data Set

The 43 congeners of the thiazolyl-pyrimidine hybrid (Figure 1) used for the study were obtained from the ChEMBL. They were selected based on pharmacophore (thiazolyl-pyrimidine skeleton), the diverse chemical substituents forming the congeners, the in vitro antiplasmodial activity (against *P. falciparum*) and the high negative decadic logarithm values (3.04 units for $5.73 < pIC_{50} < 8.77$).

Figure 1. Pharmacophore of thiazolyl-pyrimidine hybrid derivatives.

3.2. Selection of Significant Features

The number of significant features to be considered to build the model was fixed at a hypothetic value of 25 out of 145 using the RFE. To further eliminate the insignificant features, the RFE-selected features were further subjected to the Statsmodelling function to check the detailed statistics and summary of the model from the selected features. The result of the analysis showed that there were still features with *p*-values greater than 0.05 on assumptions that the covariance matrix of the standard errors (SEs) was correctly specified and that the smallest eigenvalue of 1.99×10^{-33} might indicate strong multicollinearity problems or that the design matrix was singular.

Then the VIF values for each feature of the model were calculated (Table 1). All the features with VIF > 5 and $p > 0.05$ were considered insignificant and as a result, dropped from the model. Since the *p*-values and VIF of FCASA-, vsurf_G, vsurf_HB1, E_str, MNDO_LUMO, vsurf_DD12 were in the desired range, that means they are significant features and will be used to build the machine learning models.

Table 1. Results of Statsmodel analysis.

Features	Coeff	SE	T	*p*-Value	0.025–0.875	VIF
const	8.0584	0.088	91.607	0.000	7.878–8.238	-
vsurf_EDmin3	0.3627	0.271	1.341	0.190	−0.191–0.916	39.46
vsurf_D7	−0.3807	0.266	−1.430	0.163	−0.925–0.164	9.16
vsurf_D8	0.1435	0.255	0.562	0.578	−0.379–0.665	8.42
vsurf_EDmin1	−0.3076	0.252	−1.222	0.232	−0.823–0.207	8.19
FCASA-	0.5262	0.159	3.305	0.003	0.201–0.852	3.28
vsurf_G	0.3665	0.147	2.495	0.019	0.066–0.667	2.79
vsurf_HB1	−0.3665	0.131	−2.808	0.009	−0.633–0.100	2.20
E_str	−0.3877	0.127	−3.044	0.005	−0.648–0.127	2.10
MNDO_LUMO	−0.3641	0.126	−2.880	0.007	−0.623–0.105	2.07
vsurf_IW1	−0.1351	0.123	−1.101	0.280	−0.386–0.116	1.94
vsurf_IW2	0.0463	0.117	0.396	0.695	−0.193–0.285	1.77
vsurf_DD12	−0.2716	0.108	−2.514	0.018	−0.493–0.051	1.51
vsurf_Wp6	0.0651	0.101	0.647	0.523	−0.141–0.271	1.31

The molecular features are: third lowest hydrophobic energy (vsurf_EDmin3); hydrophobic volume at −1.4 (vsurf_D7); hydrophobic volume at −1.6 (vsurf_D8); lowest hydrophobic energy (vsurf_EDmin1); fractional charge-weighted negative surface area (FCASA-); surface globularity (vsurf_G); H-bond donor capacity at −0.2 (vsurf_HB1); hydrophilic integy moment at −0.2 (vsurf_IW1); hydrophilic integy moment at −0.5 (vsurf_IW2); vsurf_EDmin1, vsurf_EDmin2 distance (vsurf_DD12); polar volume at −4.0 (vsurf_Wp6); LUMO energy, ev (MNDO_LUMO); bond stretch energy (E_str); lowest unoccupied molecular orbital (LUMO).

3.3. Residual Analysis of the Model

The residue analysis of the error terms was checked to ascertain their normal distribution and the error terms of the histogram were plotted (Figure 2). A normal distribution is one of the major assumptions of multiple linear regression and since the error terms are normally distributed, the model can be used to make predictions on the test dataset.

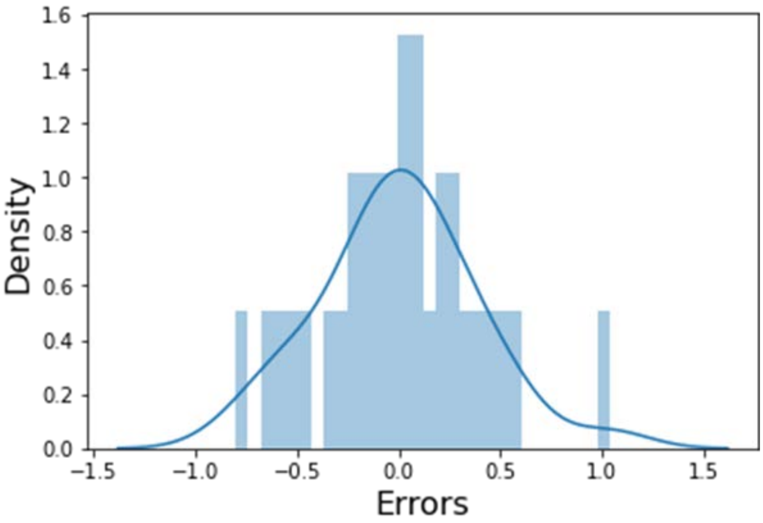

Figure 2. Histogram of error terms.

3.4. Model Building

Machine learning-based algorithms were built from the significant features to predict the pIC_{50} values of the test molecules. The predicted pIC_{50} values for the test compounds are shown in Table 2.

Table 2. Predicted pIC_{50} of the test molecules using MLR model.

Tzpd	Actual pIC_{50}	Predicted pIC_{50}
25	8.37	8.380461
8	8.64	8.667578
27	7.35	6.915064
11	8.64	8.122377
22	8.77	7.905053
14	8.64	7.908562
6	8.77	8.151523
2	7.28	7.760386
7	8.42	8.064295

The details of Tzpd used as test set can be found in the supplementary Figure S1a,b.

To prove further confidence in our predicted pIC_{50} values, the predicted pIC_{50} scores were plotted against the experimental pIC_{50} scores for both the train set and the test set, using different machine learning models (Figure 3). The closeness of the predicted pIC_{50} scores and the experimental scores for Figure 3A,C shows the robustness of the MLR and SVR models in predicting the antiplasmodial activity of Tzpd. This showed that the predictive powers of the models are competent. The correlations of the predicted and experimental pIC_{50} values are shown in Figure 3A–D. The R^2 indicates how closely the data resemble the regression line and how well the data fit the regression line.

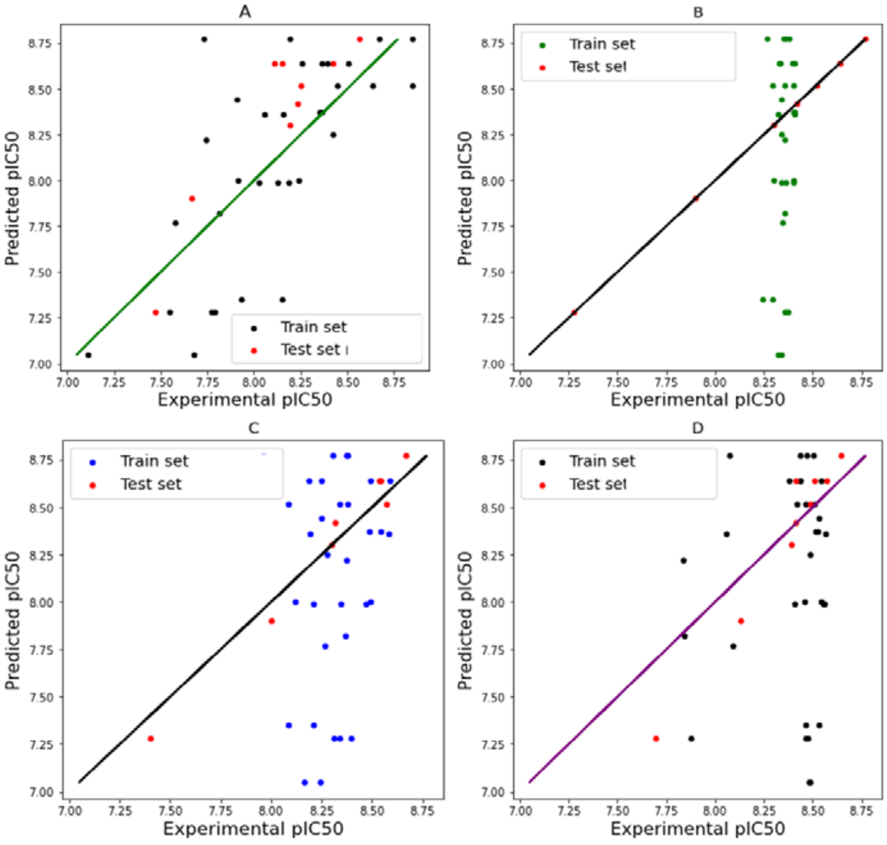

Figure 3. Regression plots of different models (A = MLR; B = KNN; C = SVR and D = RFR).

3.5. Model Evaluation and Comparison

The summary of the performance of the models is shown in Table 3.

Table 3. Model prediction statistics.

ML Algorithms	kNN	SVR	RFR	MLR
Test MSE	0.00	0.053	0.069	0.1453
5-fold cross-validation	0.59 ± 0.41	0.67 ± 0.45	0.75 ± 0.29	0.091 ± 0.010
Test R^2	1.00	0.61	0.36	0.68
5-fold cross-validation	0.36 ± 0.46	0.63 ± 0.62	0.59 ± 2.21	0.745 ± 0.281
Test MAE	0.00	0.174	0.209	0.290
5-fold cross-validation	0.55 ± 0.18	0.58 ± 0.20	0.60 ± 0.60	0.270 ± 0.101
Test RMSE	0.00	0.230	0.262	0.381
5-fold cross-validation	0.72 ± 0.27	0.77 ± 0.27	0.84 ± 0.18	0.302 ± 0.021

Five-fold cross-validation scores of MLR, kNN, SVR, and RFR were plotted on a box-plot and their performances were compared (Figure 4). The performance metrics for each model were plotted as a box. Using the 5-fold cross-validation approach, MLR and SVR outperforms the other models, as the median line was visibly higher in all the metrics used.

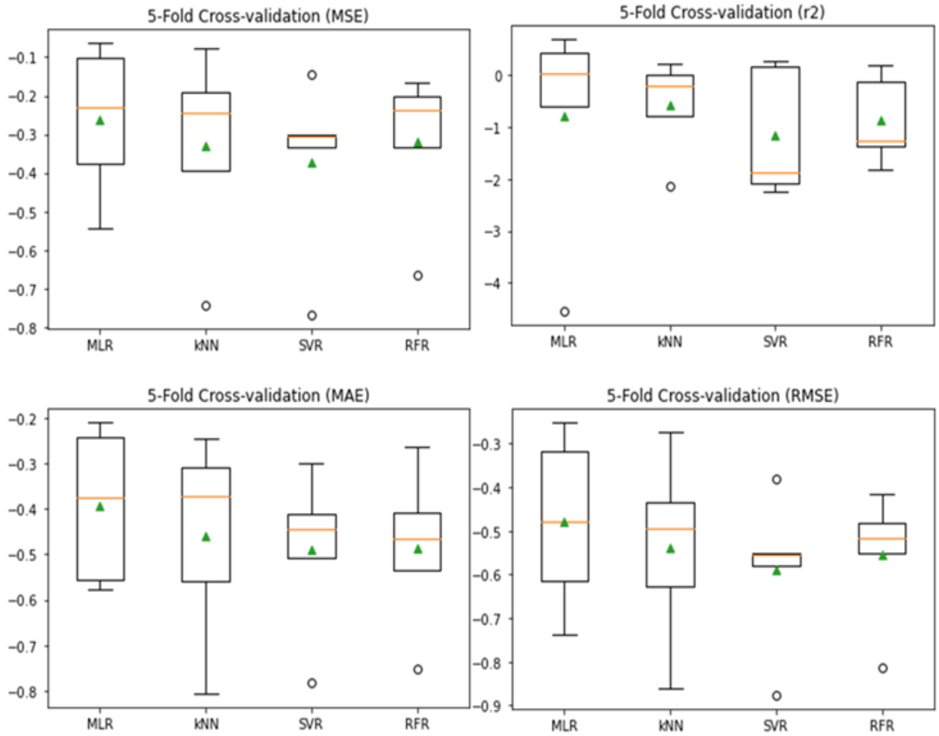

Figure 4. Boxplots of 5-fold CV scores.

4. Conclusions

The study demonstrated that MLR and SVR are powerful predictive supervised learning model with reproducible outcomes and the lowest model errors when compared to kNN and RFR. The multiple linear regression equation: pIC_{50}(predicted) = 8.06 – 0.45vsurf_G + 0.37FCASA– – 0.42MNDO_LUMO – 0.20E_str + 0.30vsurf_HB1 – 0.38vsurf_DD12) allows for the prediction of antiplasmodial activity which can be utilized in the design of new bioactive chemical entities using artificial intelligence qualities.

Supplementary Materials: The following supporting information can be downloaded at: www.mdpi.com/xxx/s1, Figure S1: Chemical data set.

Author Contributions: Conceptualization, design and supervision, C.O.N. and W.O.O.; software, C.O.N.; validation, W.O.O.; formal analysis, K.S.U.; investigation, K.S.U.; resources, and data curation, C.O.N. and K.S.U.; writing—original draft preparation, K.S.U.; writing—review and editing, C.O.N.; visualization, K.S.U. All authors have read and agreed to the published version of the manuscript.

Funding: This research received no external funding.

Institutional Review Board Statement: Not applicable.

Informed Consent Statement: Not applicable.

Data Availability Statement: Not applicable.

Acknowledgments: Authors are grateful to Thecla Ayoka for the software she provided.

Conflicts of Interest: The authors declare no conflict of interest.

References

1. Ikerionwu, C.; Ugwuishiwu, C.H.; Okpala, I.; James, I.; Okoronkwo, M.; Nnadi, C.; Orji, U.; Ebem, D.; Ike, A. Application of machine and deep learning algorithms in optical microscopic detection of *Plasmodium*: A malaria diagnostic tool for the future. *Photodiagn. Photodyn. Ther.* **2022**, *40*, 103198.
2. Oguike, E.; Ugwuishiwu, C.; Asogwa, C.; Nnadi, C.; Obonga, W.; Attama, A. A systematic review on the application of machine learning to quantitative structure-activity relationship modeling against *Plasmodium falciparum*. *Mol. Divers.* **2022**, *26*, 3447–3462.
3. Ikwuka, C.E.; Asogwa, C.M.; Ikwuka, O.J.; Ogbonna, J.E.; Onah, C.E.; Ohama, C.C.; Nnadi, C.O. Insights into the In-vivo Antiplasmodial Activity of Trisdimethylamino Pyrimidine Derivative in *Plasmodium berghei* Infected Mouse Model. *J. Pharm. Res. Int.* **2022**, *34*, 33–40.
4. Nnadi, C.O.; Ayoka, T.O.; Okorie, H.N. A ligand-based approach to lead optimization of N, N′-substituted diamines for leishmanicidal activity. *Biointerface Res. Appl. Chem.* **2022**, *12*, 7429–7437.
5. Nnadi, C.O.; Althaus, J.B.; Nwodo, N.J.; Schmidt, T.J. A 3D-QSAR study on the antitrypanosomal and cytotoxic activities of steroid alkaloids by comparative molecular field analysis. *Molecules* **2018**, *23*, 1113.
6. u, W.; Chen, M.; Fei, Q.; Ge, Y.; Zhu, Y.; Chen, H.; Yang, M.; Ouyang, G. Synthesis and Bioactivities Study of Novel Pyridylpyrazol Amide Derivatives Containing Pyrimidine *Motifs. Front. Chem.* **2020**, *8*, 522.
7. Chemical Computing Group. *Molecular Operating Environment (MOE) rel. 2011.10*; Chemical Computing Group Inc.: Montreal, QC, Canada, 2014.
8. Sydow, D.; Morger, A.; Driller, M.; Volkamer, A. TeachOpenCADD: A teaching platform for computer-aided drug design using open-source packages and data. *J. Cheminform.* **2019**, *11*, 29–36.
9. Du, Z.; Yang, H.; Lv, W.J.; Zhang, X.Y.; Zhai, H.L. prediction of the inhibitory concentrations of chloroquine derivatives using deep neural network models. *J. Biomol. Struct. Dyn.* **2021**, *39*, 672–680.
10. Afuwape, A.A.; Xu, Y.; Anajemba, J.H.; Srivasta, G. Performance evaluation of secured network traffic classification using machine learning approach. *Comput. Standards Interfaces* **2021**, *78*, 103545.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.