Response to Mukwembi and Nyabadza

Addressing Concerns Over Regression Fit

E. Reyes and M. Riehl

Updated: 8 December 2025

Background

In “Predicting anti-cancer activity in flavonoids: a graph theoretic approach,” appearing in *Scientific Reports*, the authors consider a graph-theoretic model G for a flavonoid F. They propose combining the internal and external activities of the molecular graph to predict the anti-tyrosinase activity, the IC_{50}, for the flavonoid. The authors have data available from nine flavonoids, which they present in their Table 1. For each flavonoid, the external activity of the molecular graph $D(G)$, the internal activity of the molecular graph $\zeta(G)$, and the IC_{50} are provided. The authors propose using a “multivariable function [of $D(G)$ and $\zeta(G)$] to model the effects of the two invariants on the IC_{50} values.”

Using the data, the authors fit a model of the form

$$\begin{aligned} IC_{50}(G) &= f\left([D(G)],[\zeta(G)]\right) \\ &= \alpha_1 + \alpha_2[D(G)] + \alpha_3[\zeta(G)] + \alpha_4[D(G)]^2 + \alpha_5[D(G)][\zeta(G)] \\ &\quad + \alpha_6[\zeta(G)]^2 + \alpha_7[D(G)]^2[\zeta(G)] + \alpha_8[D(G)][\zeta(G)]^2 + \alpha_9[\zeta(G)]^3. \end{aligned} \tag{1}$$

The function $f(\cdot,\cdot)$ in Equation 1 specifies the average IC_{50} value given the internal and external activity of the molecular graph. The authors do not provide a theoretical justification for this particular functional form but state they “resorted to the multivariable polynomial since it gives the best goodness of fit value measured by R^2 or R-square.” The coefficient of determination, R^2, is a metric of model fit defined as

$$R^2 = 1 - \frac{SSE}{SST} = 1 - \frac{\sum_{i=1}^{n}\left((\text{Observed } IC_{50}(G))_i - (\text{Predicted } IC_{50}(G))_i\right)^2}{\sum_{i=1}^{n}\left((\text{Observed } IC_{50}(G))_i - (\text{Overall Average } IC_{50}(G))\right)^2} \tag{2}$$

where the observed $IC_{50}(G)$ values are those values used to fit the model and the predicted $IC_{50}(G)$ values are those predicted by the model fit to the original data; the sums are indexed over the observations ($i = 1, 2, \ldots, 9$ in this case). As the predictions become closer to the observed values R^2 increases toward 1. However, R^2 should primarily be used when the sample size n far exceeds the number of parameters p in the model. Specifically, note that Equation 2 shows that when considering different models on the same set of data, then R^2 is a function of the residual sum of squares SSE (since the total sum of squares SST will remain fixed). It can be shown that R^2 is then a non-decreasing function as more terms are added to a model (Rao 2008, pg 63). Further, if the model has the same number of parameters as the sample size, then $R^2 = 1$.

As the data collected by the authors has only $n = 9$ observations, the model in Equation 1 is a saturated model and will necessarily overfit the data. This leads to overconfidence in the model fit when using it to predict the IC_{50} for additional flavonoids. It also prohibits making any inference on the parameters of the model (as there are no observations left to estimate the variability in the parameter estimates).

We have argued here that Equation 1 is inappropriate from a statistical perspective. However, we should also reflect on the plausibility of the model from its biological use. The IC_{50} is the concentration of an inhibitor required to reduce a particular biological or biochemical activity to 50% of its control (or uninhibited) value (Cheng 1973). Reviewing the predicted IC_{50} for 26 compounds considered by the authors in their Table 2 (produced from Equation 1) should give substantial pause.

Several of the predicted IC_{50} values are negative. Values near 0 would mean that even trace amounts of the flavonoid were needed to inhibit the biochemical activity. A negative IC_{50} has no practical meaning: if a curve fitting procedure predicts a negative value, that generally signals trace amounts were needed or the inhibitor did not produce measurable inhibition in the tested range (or perhaps even acted as an activator!); the negative value is just a mathematical artifact of extrapolation (Montesinos López 2022). Extrapolation is a concern here as several of the flavonoids for which predictions were made have internal and external activities $D(G)$ and $\zeta(G)$ that are dissimilar from those flavonoids used to fit the model. The negative IC_{50} values suggest the model does not routinely provide biologically plausible predictions.

We also note the relative size of the IC_{50} values predicted. The data the model was based on indicated IC_{50} values between 96 and 624 μM. However, the predicted IC_{50} values reported in the authors' Table 2 are has high as 33,581 μM. What is a realistic dynamic range for IC_{50} values? Practical experimental

Table 1: Estimated parameters and corresponding standard errors when fitting Equation 3, a smaller model explaining the IC_{50} using the two invariants of the molecular graph.

Term	Estimate	Standard Error
β_0	6493.581	1380.2105
β_1	-616030.471	181971.7886
β_2	-3742.295	880.0634
β_3	16585103.828	5731099.4917

limits make anything above 10,000 μM effectively meaningless for biological assays (Sebaugh 2011). The extreme IC_{50} values predicted by the model, while not necessarily indicative of overfitting like the negative values, should perhaps simply be stated as "no inhibitory activity predicted."

Potential Improvements to the Model

Suppose, instead, we consider a smaller model of the form

$$IC_{50}(G) = \beta_0 + \beta_1[D(G)] + \beta_2[\zeta(G)] + \beta_3[D(G)]^2. \tag{3}$$

This model was developed by considering a model with only the linear terms and considering only second-order terms that were statistically significant. Table 1 provides estimates for the parameters and corresponding standard errors; fitting Equation 3 results in an $R^2 = 0.975$, though we again warn against over-interpreting R^2 in such a small sample.

Using the reduced model Equation 3, we predict the IC_{50} for the 26 compounds considered by the authors in their Table 2; Table 2 presents these comparisons. Confidence intervals were generated assuming the classical distributional assumptions; while these may not be the most appropriate, they do provide some sense of the variability in these predicted mean responses.

Reflecting on the biological plausibility of this reduced model, we note that even this reduced model produces some negative estimated IC_{50} values; however, the confidence interval for each of these includes positive values. Each of these instances involves predicting the IC_{50} for a flavonoid with a $D(G)$ and $\zeta(G)$ value beyond the range of that observed among the flavonoids within the original sample; that is, extrapolation is still a concern. We also note that the predicted IC_{50} for each flavonoid seems to be on an order of magnitude that is acceptable (nothing above 10,000 μM).

Table 2: Comparison of predicted IC_{50} values between the model proposed by the authors and the reduced model considered in Equation 3.

Flavonoid	D(G)	Z(G)	Predicted IC-50	Updated Predicted IC-50	95% CI
Kojic acid	0.0280000	0.2383330	-112089.01100	1356	(-840, 3551)
Chrysin	0.0040822	0.2022161	-40655.55419	3498	(1460, 5537)
Shikonin	0.0280747	0.1934996	-33061.25049	1547	(-637, 3731)
Baicalein	0.0062500	0.1983332	-26435.10902	2549	(1197, 3901)
Galangin	0.0072500	0.1800000	-14014.27835	2226	(1158, 3293)
Dihydromyricetin	0.0223556	0.1499684	-9665.55940	449	(-132, 1031)
Naphthazarin	0.0204082	0.1938776	-3748.19286	104	(-190, 398)
Xanthoxylin	0.0364431	0.1479592	-2455.07267	5517	(-674, 11707)
Tropolone	0.0082305	0.2222222	-1193.32990	1715	(842, 2589)
Morin	0.0178437	0.1701102	35.71991	145	(81, 210)
Quercetin	0.0184072	0.1773416	93.33756	110	(25, 195)
Quercetin-3-rutinoside	0.0184072	0.1773416	93.33764	110	(25, 195)
Taxfolin	0.0184072	0.1773415	93.33836	110	(25, 195)
Rhamnetin	0.0169310	0.1877757	105.59451	115	(73, 157)
Dihydroquercetin-4 -methylether	0.0168489	0.1871456	113.02278	122	(80, 164)
Tamarixetin	0.0169310	0.1937618	147.34330	93	(49, 136)
Dihydroquercetin-7,4 -dimethylether	0.0155527	0.1953125	162.70396	193	(147, 240)
Luteolin	0.0151172	0.1976568	259.61997	231	(184, 279)
Luteolin-7-methyl ether	0.0138993	0.2024793	352.58478	378	(303, 452)
5,7,3 ,5 -Tetrahydroxyflavanone	0.0146852	0.1557067	427.35326	441	(362, 520)
Blumeatin	0.0134298	0.1646006	617.41701	596	(521, 670)
Apigenin	0.0105000	0.2191667	4333.36622	1034	(593, 1474)
Fisetin	0.0142533	0.2120181	4757.00226	289	(209, 369)
Rosmarinic acid	0.0201980	0.2504930	6636.14232	-120	(-456, 215)
3,7,4 -Trihydroxyflavone	0.0102500	0.2350000	27875.46695	1042	(543, 1542)
Isoeugenol	0.0208333	0.2754630	33581.06508	-173	(-626, 281)

Measuring Quality of Model Fit

We have noted the concerns with R^2 as a metric for assessing the quality of a model fit. An alternative metric is predicted R-squared R^2_{PRESS} (Alcantara, 2023), which has also been referred to as Q^2 in the literature (Quan, 1988). This measure is defined as

$$R^2_{\text{PRESS}} = 1 - \frac{\sum_{i=1}^{n} \left((\text{Observed } IC_{50}(G))_i - (\text{Predicted } IC_{50}(G))_{(i)} \right)^2}{\sum_{i=1}^{n} \left((\text{Observed } IC_{50}(G))_i - (\text{Overall Average } IC_{50}(G))_{(i)} \right)^2}. \quad (4)$$

Note the similarity between Equation 2 and Equation 4. The (i) subscript is used to denote a statistic computed after leaving out the i-th observation in the dataset. First introduced by Allen (1971), the PRESS statistic is defined as

$$\text{PRESS} = \sum_{i=1}^{n} \left(Y_i - \widehat{Y}_{(i)} \right)^2$$

where the predicted value $\widehat{Y}_{(i)}$ for the i-th observation is computed by first removing the i-th observation from the dataset, fitting a model on the remaining data, and using the fitted model to predict the response for the i-th observation. This is also known as “leave one out cross validation.” R^2_{PRESS} is then analogous to R^2 where the original summations are replaced with their PRESS alternatives. The argument for PRESS is that it balances model fit with the future predictive ability of the model. As the authors are particularly interested in applying their model to additional observations, R^2_{PRESS} may be an improved metric for assessing quality of the model fit.

Unlike R^2, which is bounded between 0 and 1, R^2_{PRESS} can be negative. This occurs when the model gives extremely variable predictions for new observations. In such cases, it is typical to report $R^2_{\text{PRESS}} = 0$.

We note that R^2_{PRESS} cannot be computed for the model in Equation 1 as it has the same number of parameters as observations in the data. As a result, when an observation is dropped from the data, the model with 9 parameters cannot be fit with the remaining 8 observations.

We computed R^2_{PRESS} for a series of models (see Table 3).

Table 3: Comparison of predicted R-squared values for a series of models.

Model	Predicted R-squared
1 + D	0.619
1 + Z	0.029
1 + D + Z	0.827
1 + D + Z + DZ	0.654
1 + D + Z + D2	0.902
1 + D + Z + Z2	0.000
1 + D + Z + DZ + D2	0.913
1 + D + Z + DZ + Z2	0.000
1 + D + Z + D2 + Z2	0.000
1 + D + Z + DZ + D2 + Z2	0.234

We see that two models stand out with $R^2_{\text{PRESS}} > 0.9$. The first is the model with 4 parameters proposed in Equation 3. The other adds the second-order term $D(G)^2$ to the model in Equation 3. As the two R^2_{PRESS} statistics are similar, we chose the more parsimonious model.

We make no claims that the proposed reduced model Equation 3 is the "correct" model; however, both the model fit statistics and the biological plausibility of the predicted IC_{50} values suggest it is an improvement over the model proposed by the authors. We would advocate for developing a more comprehensive sample before developing a model that would be reliable broadly for prediction.

Acknowledgements

The authors would like to thank Emily Maxwell for her suggestions.

Citations

All analyses were performed using R version 4.5.2 (2025-10-31 ucrt).

- Ida Marie Alcantara, Joshua Naranjo, and Yanda Lang. 2023. "Model selection using PRESS statistic." *Computational Statistics*, 38(1), pp.285-298.
- David M. Allen. 1971. "The Prediction Sum of Squares as a Criterion for Selecting Predictor Variables." University of Kentucky Department of Statistics, *Technical Report 23*.

- Yung-Chi Cheng and William H. Prusoff. 1973. "Relationship between the inhibition constant (KI) and the concentration of inhibitor which causes 50 per cent inhibition (I50) of an enzymatic reaction." *Biochemical Pharmacology*, 22(23), pp.3099-3108.
- Montesinos López OA, Montesinos López A, Crossa J. Cham. 2022. *Multivariate Statistical Machine Learning Methods for Genomic Prediction [Internet].* Springer. https://www.ncbi.nlm.nih.gov/books/NBK583970/
- Nguyen T. Quan. 1988. "The Prediction Sum of Squares as a General Measure for Regression Diagnostics." *Jounral of Business & Economic Statistics*, 6(4), pp.501-504.
- R Core Team. 2025 *R: A Language and Environment for Statistical Computing.* R Foundation for Statistical Computing.
- C. Radhakrishna Rao, Helge Toutenburg, Shalabh, and Christian Heumann. 2007. *Linear Models and Generalizations: Least Squares and Alternatives (3rd. ed.).* Springer Publishing Company, Incorporated.
- JL Sebaugh. 2011. "Guidelines for accurate EC50/IC50 estimation." *Pharmaceutical Statistics*, 10(2), pp.128-134.