Results in Chemistry 5 (2023) 100888

Contents lists available at ScienceDirect

Results in Chemistry

journal homepage: www.sciencedirect.com/journal/results-in-chemistry

Quantitative Structure-Activity Relationship (QSAR) modelling of the activity of anti-colorectal cancer agents featuring quantum chemical predictors and interaction terms

Han Ngoc Bao Nguyen [a], Michael Yudistira Patuwo [b,*]

[a] St. Andrew's Junior College, 5 Sorby Adams Drive, Singapore 357691, Singapore
[b] National University of Singapore, Department of Chemistry, Block S8 Level 3, 3 Science Drive 3, Singapore 117543, Singapore

ARTICLE INFO

Keywords:
Anti-colorectal cancer agents
Interaction terms
Logistic regression
QSAR
Quantum chemical predictors

ABSTRACT

A Quantitative Structure-Activity Relationship (QSAR) study was performed to 1) predict the activity of new compounds as anti-colorectal cancer agents and 2) select chemical predictors that accurately model relationships between molecular structures and bioactivity against colon cancer cell lines. Logistic regression was used to obtain coefficients for the QSAR classification model. Quantum chemical predictors were found to be significant predictors, such as the total electronic energy (E_T), charge of the most positive atom (Q_{max}) and electrophilicity (ω), thus validating the need to consider them besides those routinely used in anti-colorectal cancer agents QSAR. The model selected simple and significant predictors that can be obtained directly using information from gaseous-state Gaussian optimisation at HF/3-21G in a vacuum, with no external calculation software, providing an inexpensive yet robust QSAR model. In addition, two interaction terms were proven to be significant at 95 % confidence level, reiterating the importance of interaction between predictors that is not commonly seen in QSAR models.

Introduction

Colorectal cancer (CRC) is the third most common type of cancer worldwide [1], besetting more than a million people every year [2]. It is the second leading cause of death among various types of cancer, with 935 000 deaths in 2020 [3]. Surgery is only successful in CRC treatment in the early localised stage when it is limited within the wall of the colon [4]. Furthermore, CRC has little response to radiation [5], which makes chemotherapy crucial in the treatment of widespread malignancies of CRC [6]. Therefore, the discovery of new anticancer drugs that have higher efficacy and more manageable side effects is extremely vital [7].

However, drug discovery is a complex, expensive and time-consuming process, with only 11 % success rate in 2014 [8]. Thus, computational methodologies have been employed to minimise costs and shorten the research cycle, in which QSAR is a powerful approach to identify lead compound [9]. QSAR was first developed by Hansch [10] 40 years ago. It involves building a robust statistical algorithm that relates the molecular structures of a molecule to its activity, by selecting appropriate molecular descriptors [11]. As such, the potential of a newly discovered compound can be validated [12]. QSAR has helped to speed up the evaluation of drug compounds properties [13], thus accelerating the development of new drugs [14].

In binary classification problems, QSAR models have been proven to be very effective [15].There has been numerous literature on modelling anti-CRC activity, such as Kim BS et al. [16], Sagir Yusuf Ismail et al. [9], Verma et al. [17], Deokar et al. [18], Rybka et al. [19] and Cruz et al. [20]. These QSAR models used structural, geometrical and topological descriptors to predict the anti-CRC activity of compounds. However, very few of these QSAR models selected quantum chemical predictors. Quantum chemical predictors, such as highest occupied molecular orbital (HOMO), lowest unoccupied molecular orbital (LUMO) and charge predictors, have been shown to encapsulate information that characterises the reactivity, shape and binding properties [21]. Shruti et al. [22] showed that the best anti-cancer agents QSAR model required MiERIC, MiERIN, MaREHN quantum chemical predictors from the CODESSA programme [23]. Violeta Marković et al. [24] proved that quantum-chemical predictors were crucial in predicting anti-tumour activity of edaravone derivatives. However, they have hitherto gained little recognition in anti-CRC QSAR. Thus, this represents an opportunity to explore more into this area.

* Corresponding author.
E-mail address: chmmiy@nus.edu.sg (M.Y. Patuwo).

https://doi.org/10.1016/j.rechem.2023.100888
Received 29 September 2022; Accepted 7 March 2023
Available online 10 March 2023

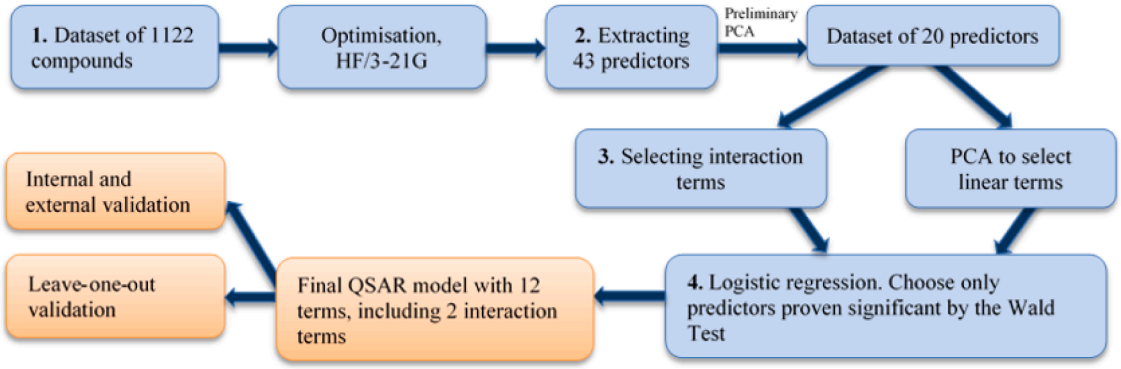

Fig. 1. Flowchart of methodology and validation.

Table 1
Snippet of the dataset showing 10 compounds and their activities.

Number	Molecule name	SMILES	Activity/ 1,-1
41	CHEMBL116776	COc1cc(OC)cc(\C=C/c2ccc(N)cc2)c1	1
42	CHEMBL102786	COc1cc2c(Nc3ccc(Sc4nccs4)c(Cl)c3)c(cnc2cc1OCCCN5CCOCC5)C#N	1
43	CHEMBL1090478	COc1ccc2c(c1)c(\C=C\3/Oc4cc(O)cc(O)c4C3=O)c(C)n2CCN(C)C	1
44	CHEMBL1242274	COc1cc(OC)cc(\C=C/c2ccc(cc2)C(F)(F)F)c1	1
45	CHEMBL593700	[O-][N+](=O)c1ccc2\C(=C/3\Nc4ccccc4C3=O)\C(=O)Nc2c1	1
46	Acetaminophen	Oc1ccc(NC(C)=O)cc1	−1
47	Amantadine	NC13CC2CC(CC(C1)C2)C3	−1
48	Amineptine	O=C(O)CCCCCCNC2c3ccccc3CCc1ccccc12	−1
49	Amlodipine	O=C(OC)C1 = C(C)NC(COCCN) = C(C1c2ccccc2Cl)C(=O)OCC	−1
50	Aspirin	O=C(C)Oc1ccccc1C(=O)O	−1

For the QSAR models which made use of quantum chemical predictors, they were noted to be computationally expensive, such as MiVH, MiNRC, MaVO, AVN, MiERC, ANRN, MaBON predictors calculated by CODESSA [23] (Bohari et al. [4]) or Spartan 14 Wave function software (Ismail et al. [9]). Other non-quantum chemical and structural predictors are also calculated by external software, such as Canvas of the Schrodinger Suite, Virtual Computational Chemistry Laboratory [25] and PaDEL-Descriptor [26] (Deokar et al. [18], Rybka et al. [19] and Cruz et al.'s [20]). Thus, there is potential in using simple predictors without external calculation software, to obtain a reliable model that is inexpensive and more convenient.

Lastly, most QSAR studies predicting activity against cancer cell lines use MLR models with single linear terms, such as in the works of Girgis et al., Zolnowska et al., Slawinski et al. and Gramatica et al. [27–31]. However, the structural and quantum chemical predictors may interact with one another, a phenomenon usually observed in machine learning problems [32]. Therefore, there is a need to address the lack of interaction terms in QSAR models.

Hence, this research aims to generate a statistically valid, robust and inexpensive model to predict the anti-CRC activity without the use of external calculation software. It also evaluates the significance of quantum chemical predictors and the importance of interaction between predictors in QSAR studies.

Materials and methods

The procedures used in this research are summarised in Fig. 1.

Dataset

The complete dataset has 563 compounds with anti-CRC activity against ten CRC cell lines from Alejandro Speck-Planche et al.'s work [33], and other 559 inactive drugs extracted from the ChemBL database [34]. These drugs' profiles did not include anti-CRC activity and thus were considered inactive. In total, there are 1122 compounds (See Table 1). There are approximately equal numbers of active and inactive compounds in the dataset, which generate higher accuracy and balanced detection rate [35].

Extracting predictors from Gaussian output files

All molecules were optimised in the gaseous state and in a vacuum at HF/3-21G (Hartree-Fock), using Gaussian 16 [36]. A Python code was used to extract a total of 43 predictors, 25 quantum chemical predictors and 18 structural predictors from the Gaussian output log files. A preliminary Principal Component Analysis (PCA) (*Appendix A*) was used for predictor screening to reduce 43 descriptors to 20 descriptors, by

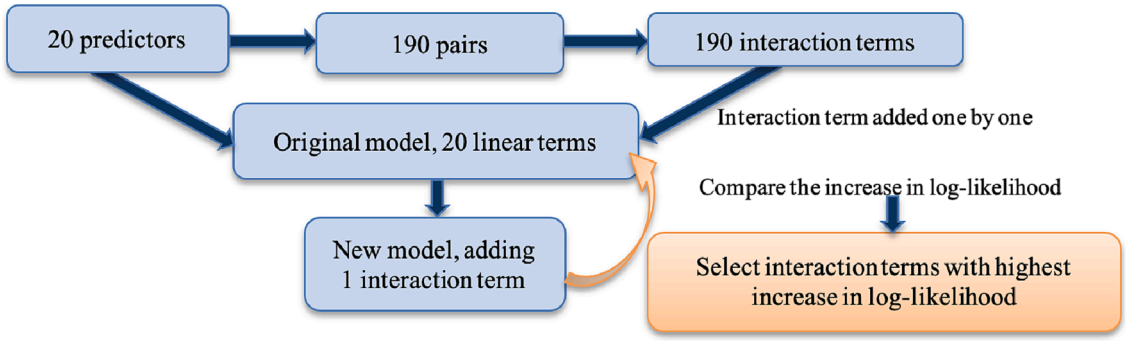

Fig. 2. Flowchart of interaction between predictors testing procedure.

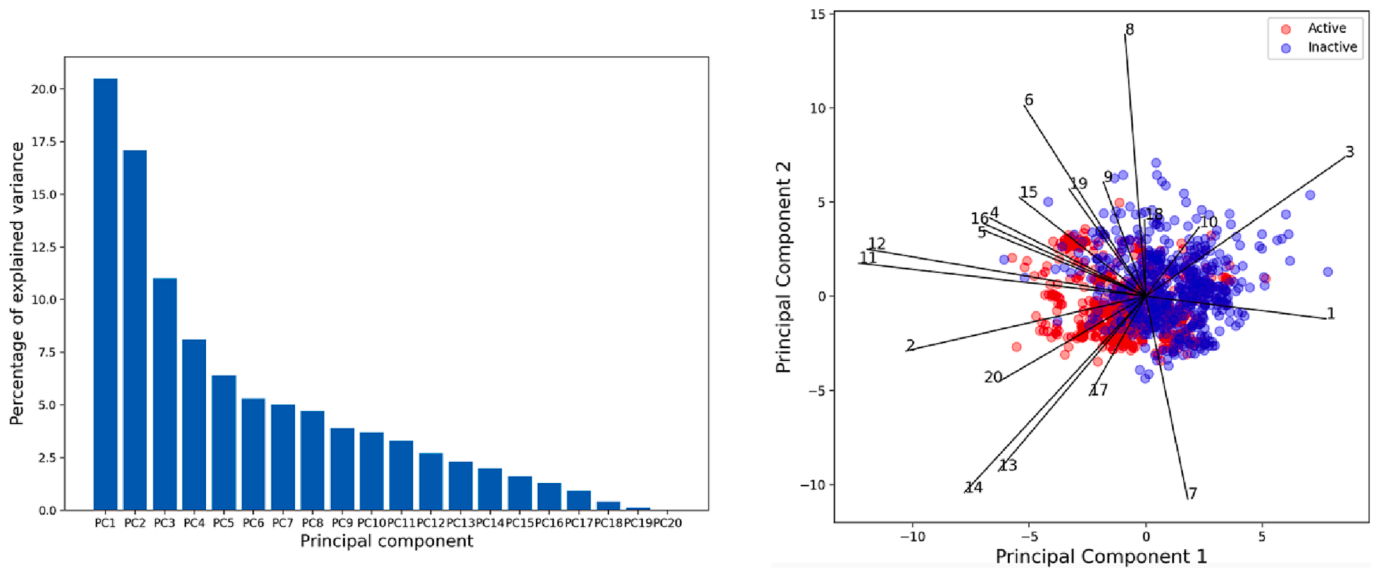

Fig. 3. (a) Scree plot of 20 principal component. (b). Biplot (PC1 and PC2).

Table 2
List of 20 predictors.

Predictor No	Predictor category	Abbreviation	Name
1		E_T	Total electronic energy
2	Quantum chemical (energy)	ω	Electrophilicity $\omega = \chi^2/2\eta$ (η = 0.5 (HOMO + LUMO)) [37]
3		χ	Electronegativity $\chi = -.5$ (HOMO-LUMO) [37]
4		S	Softness, $S = 1/\eta$ [37]
5	Quantum chemical (dipole)	μ	Total dipole
6		Q_{max}	Charge of the most positive atom
7	Quantum chemical (Natural Bond Orbital charge)	Q_{min}	Charge of the most negative atom
8		Q_m	Average of the absolute values of the charges on all atoms
9		Q_{SO2N}	Sum of absolute values of charges of SO2N groups
10		Q_{CONH2}	Sum of absolute values of charges of $CONH_2$ groups
11		$\sqrt{M}$	Square root of molecular mass
12		M	Molecular mass
13		$No_{unsub\ aromatic\ C}$	Number of unsubstituted aromatic carbons
14		$No_{aromatic\ C}$	Number of aromatic carbons
15	Structural	$No_{oxetane}$	Number of oxetane groups
16		$Presence_{CO}$	Presence or absence of CO groups
17		$No_{X\ aromatic}$	Number of halogens to aromatic rings
18		No_{COOH}	Number of COOH carboxylic groups
19		No_{OH}	Number of OH hydroxyl groups
20		No_{ArNHR}	Number of fragments type ArNHR

selecting descriptors with longest loading vectors and removing those that were highly correlated.

Selecting interaction terms with the highest increase in log-likelihood

From the dataset of 20 predictors, 190 interaction terms from 190 distinct pairs of predictors were generated. The procedure is seen in Fig. 2. The original logistic regression model assumes there is no interaction between predictors, thus consisting of 20 linear terms from 20 predictors. Subsequently, each interaction term was added to this model. The new model's log-likelihood (*Appendix B*) is then compared to of the original model. The interaction terms which gave the highest increase in log-likelihood were chosen to be further tested by the Wald test.

Logistic regression

See Appendix B for formula of logistic regression, McFadden's R^2 and the Wald test

The dataset was first divided into training and test set by randomly selecting 40 % of the compounds to be in the test set. Using linear terms chosen from PCA of 20 predictors and interaction terms chosen by the increase in log-likelihood, different logistic regression models were attempted. Only significant terms (determined by the Wald test at 95 % confidence level) were selected. The model which had the highest McFadden's R^2 and highest accuracy on the test set was chosen to be the final model.

Results and discussion

Principal component Analysis (PCA) of 20 predictors

The dataset with 20 predictors underwent PCA. The first three principal components' percentage of explained total variance is: PC1-20.5 %, PC2-17.1 %, PC3-11 % (Fig. 3a). PC1 accounts for the most variance, with 20.5 %. Several score plots were examined and the most informative one, which shows the separation of two classes most clearly, is the PC2 against PC1 plot (Fig. 3b – see Table 2 for predictor numbers from 1 to 20).

From the biplot, PC1 is responsible for the separation between the compounds with and without anti-CRC activity. Predictors with the highest loading of PC1 are selected. They are those that have the longest loading vectors along the PC1 axis, such as total electronic energy, electrophilicity, electronegativity, mass-related predictors, and aromatic carbon-related predictors. Therefore, PCA results initially show that quantum chemical predictors play a role in contributing to the variance of the response. Afterwards, the selected predictors undergo the Wald test of significance in the QSAR model, before being chosen for the final model.

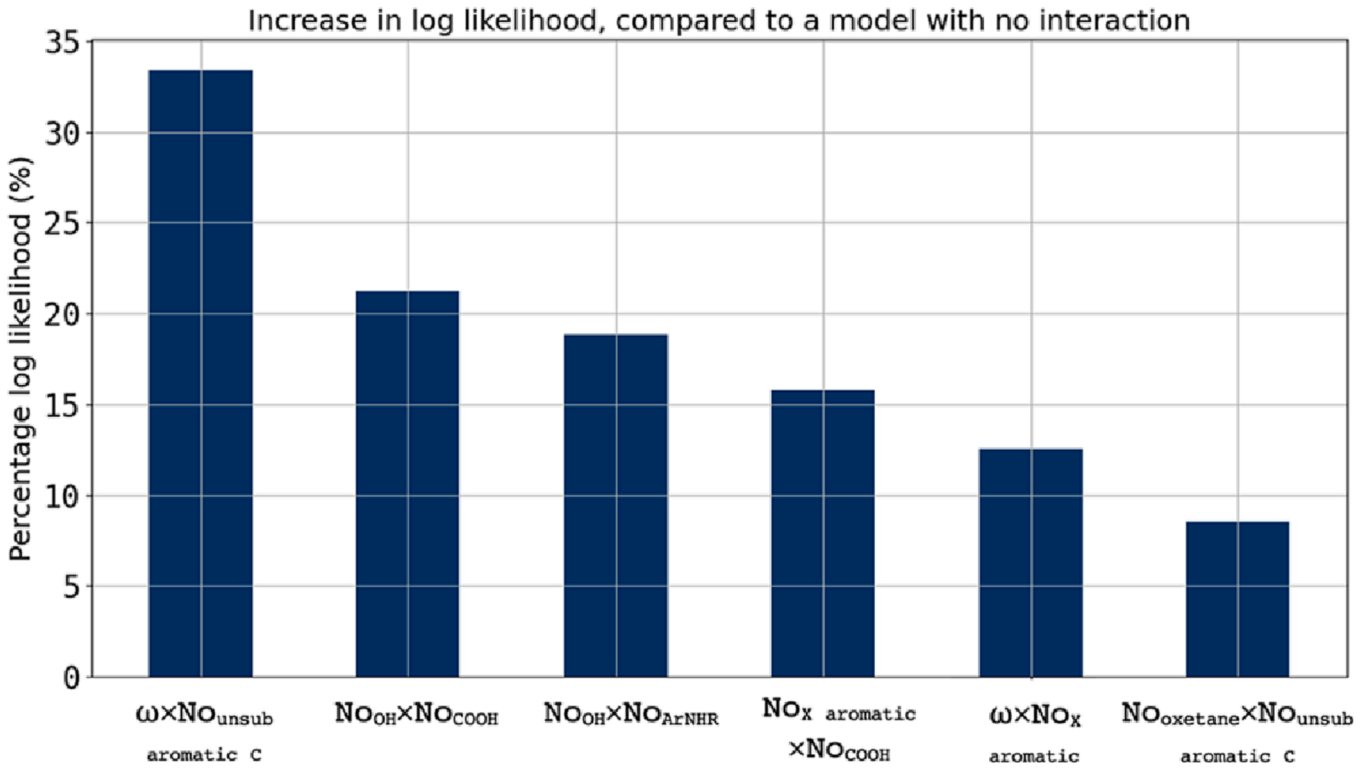

Fig. 4. Bar chart of increase in log-likelihood when each interaction term is added.

Table 3
Results of the logistic regression model.

Internal validation	(training set)	External validation	(test set)
Accuracy	McFadden's R^2	Accuracy	Sensitivity
0.86776	0.96555	0.85301	0.82540

Table 4
Results of leave-one-out cross-validation.

Number of correct predictions	Number of false predictions	Percentage of correct predictions
965	157	86.01 %

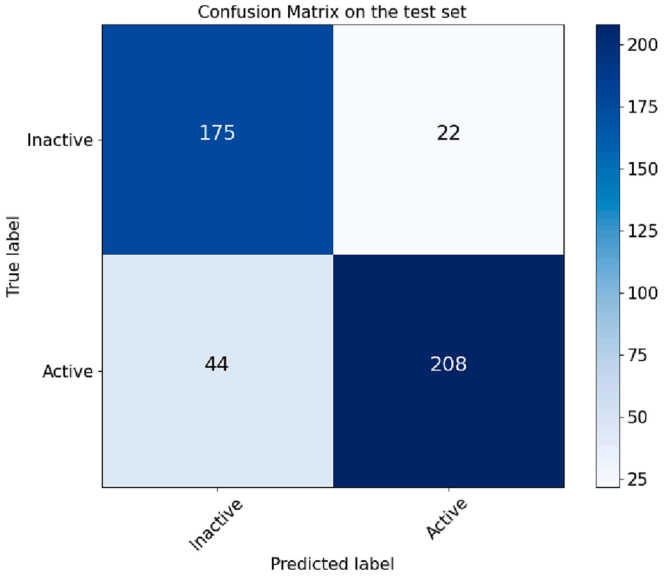

Fig. 5. Confusion matrix on the test set.

Table 5
Wald test results:

Abbreviation	Coefficients	Standard deviation	Z-score	P values
$\sqrt{M}$	0.136357	0.046812	2.912842	0.003582
E_T	−0.0002	8.32×10^{-5}	−2.40232	0.016291
Q_{max}	−0.91213	0.297275	−3.0683	0.002153
$No_{aromatic\ C}$	0.516346	0.05671	9.105063	8.62×10^{-20}
$No_{unsub\ aromatic\ C}$	−0.31479	0.115809	−2.7182	0.006564
$Presence_{CO}$	0.846033	0.323624	2.614245	0.008942
No_{ArNHR}	0.619019	0.259708	2.38352	0.017148
$No_{oxetane}$	2.477856	0.762761	3.248534	0.00116
$No_{X\ aromatic}$	−1.25122	0.20772	−6.0236	1.71×10^{-9}
No_{COOH}	−1.82476	0.47864	−3.81238	0.000138
No_{ArNHR}*No_{OH}	−0.58492	0.100188	−5.8382	5.28×10^{-9}
ω*$No_{unsub\ aromatic\ C}$	1.214862	0.381891	3.181174	0.001467

Selecting interaction terms

Fig. 4 shows the 5 interaction terms which gave the highest increase in log-likelihood.

Each of those 5 interaction terms is added subsequently to the QSAR logistic regression model to undergo the Wald test. Results show that 'Electrophilicity and Number of unsubstituted aromatic carbons' and 'Number of fragment type ArNHR and Number of OH groups' are significant by the Wald test. The inclusion of this combination gave the model with the highest accuracy and McFadden's R^2. Without these two terms, the model with only linear terms showed substantially lower accuracy and McFadden's R^2. Therefore, interaction between predictors plays a significant role in modelling the anti-CRC activity of a

compound. There should be more attention to test interaction terms in QSAR, especially in anti-CRC QSAR.

QSAR result

The final logistic regression model has 12 variables, including 2 interaction terms, all have null hypothesis rejected in the Wald test (*Appendix C*). Its equation is shown below, where β is the coefficients vector, $\mathbf{X}$ is the variables vector and π is the probability of the compound being active:

$$\pi(\mathbf{X}) = \frac{1}{1 + e^{-\beta \bullet \mathbf{X}}} \tag{1}$$

in which

$$\begin{aligned}\beta \bullet \mathbf{X} = &-5.13991 + 0.13636\sqrt{M} - 0.00020E_T - 0.91213Q_{max} + 0.51635No_{aromaticC} - 0.31479No_{unsubaromaticC} + 0.84603Presence_{CO} + 0.61902No_{ArNHR} \\ &+ 2.47786No_{oxetane} - 1.25122No_{Xaromatic} - 1.82476No_{COOH} - 0.58492No_{ArNHR} \times No_{OH} + 1.21486\omega \times No_{unsubaromaticC}\end{aligned} \tag{2}$$

Internal and external model validation

For internal validation, a fairly high McFadden's R^2 of 0.96555 ensures the developed QSAR model's goodness of fit, which implies the model's robustness (see Table 3). For external validation, a relatively high accuracy of 85.30 % on the test set also implies its strong predictive power for novel compounds. Therefore, it shows that the QSAR model for CRC is reliable to classify active and inactive compounds and predict the activity of novel agents. It also reiterates the possibility to generate a robust QSAR model using simple predictors, both structural and quantum chemical, without external calculation software. The sensitivity can be further explained by the confusion matrix in Fig. 5.

The number of True Positive (TP) predictions is 208, of True Negative (TN) predictions is 175, while of False Positive (FP) is 22 and of False Negative (FN) is 44. Sensitivity, which is the measure to check correctly positive predicted outcomes out of the total number of true positive outcomes,of the model is 208/(208 + 44) = 0.82540. This model had the highest sensitivity among all different QSAR logistic regression models generated. High sensitivity is particularly important in the case of identifying active anti-CRC agents when it is more detrimental to have FN than FP.

For FN, if there are wrong predictions of active agents as inactive, actual active compounds will be overlooked in preliminary screening, which reduces the chance of finding a novel anti-cancer agent. Meanwhile, for FP, an inactive compound predicted as active can still be eliminated upon further checking. Therefore, high sensitivity of 0.82540 implies the model's power to correctly predict observations in the positive class and minimise FN when predicting anti-CRC activity.

Additionally, the logistic regression explains how the structural key points correlate to differential behaviour in bioactivity against colon cancer cells. The magnitudes and signs of the coefficients have significant roles in deciding the overall biological activity of the molecule. In this study, the descriptors with positive coefficients are very significant by contributing towards increasing anti-CRC. $\sqrt{M}$, $No_{\text{aromatic C}}$, $Presence_{CO}$, No_{ArNHR}, $No_{oxetane}$, $\omega^*No_{\text{unsub aromatic C}}$ have positive coefficients. Therefore, molecules with high molecular mass, a large number of aromatic carbons, carbonyl groups, fragments of type ArNHR and oxetane rings; as well as high value of electrophilicity interacting with unsubstituted aromatic carbons tend to be more active. These characteristics are useful in designing new compounds with high anti-CRC activity.

Furthermore, even though E_T, Q_{max} and $\omega^*No_{\text{unsub aromatic C}}$ are the only three quantum chemical-related predictors in the final model, attempts made to remove those from the equation decreased the accuracy substantially. Thus, the contribution of those three significant quantum chemical predictors validates the need to consider quantum chemical predictors. They play a role in governing bioactivity against colorectal cancer cell lines and should not be undermined in future QSAR works.

Among many promising quantum chemical predictors after PCA, only E_T, Q_{max} and $\omega^*No_{\text{unsub aromatic C}}$ had the null hypothesis rejected by the Wald test. Note that even though E_T has a small coefficient, a smaller standard deviation (0.0000832019) gave rise to high Z-score that rejects the null hypothesis. After many attempts, these 3 predictors also gave better logistic regression models with higher accuracy compared to other quantum chemical predictors. This result shows that the three predictors are more essential in explaining anti-CRC activity.

Moreover, it is also interesting to note that compounds with high $\omega^*No_{\text{unsub aromatic}}$ are more active. While electrophilicity is not a significant predictor and was removed by the Wald test, its interaction term with unsubstituted aromatic carbons serves as a very significant predictor that increases anti-CRC activity. Therefore, the ability of the molecule to accept electron pairs interacting with unsubstituted aromatic carbons may control chemical reactions that determine anti-CRC activity.

Additionally, between the square root of molecular mass and molecular mass, only the square root of molecular mass is proven to be significant by the Wald test. The logistic regression model with the square root of molecular mass also gave better results. The insight that square root molecular mass serves better than molecular mass is also seen in past QSAR works [38], which is useful for future anti-CRC QSAR.

Leave-one-out cross-validation (LOO)

Besides internal and external validation, LOO was used to evaluate the model's predictive power. One compound from the set of 1122 compounds is extracted each time. The model is recalculated using a training set of 1121(n-1) compounds, so that each compound is predicted once. In classification problems, the statistical metric is the percentage of observations correctly classified during the n repeated model fittings [39].

86.01 % of 1122 predictions from 1121 different logistic regression models were correct (see Table 4). It reiterates the model's high accuracy to predict novel compounds, as well as robustness in predicting anti-CRC activity.

Conclusion

By considering both quantum chemical and structural predictors, the proposed QSAR model correctly classified 85.30 % of active and inactive compounds in the test set. It also has relatively high McFadden's R^2, suggesting good goodness of fit. The model selected 12 predictors, 9 structural and 3 quantum-chemical related predictors from PCA and the Wald test, that are sufficiently rich in electronic and quantum chemical information to encode molecular structures and predict anti-CRC profiles. Structural fragments with positive contribution to anti-CRC activity were reported to be used in the design of new active compounds. The model also proves the significance of interaction terms, and hence, the importance of interaction between predictors. In addition, the study

demonstrates the possibility of obtaining a robust model using simple predictors without the purchase of calculation software, hence making anti-CRC agents preliminary screening more accessible.

CRediT authorship contribution statement

Han Ngoc Bao Nguyen: Software, Validation, Formal analysis, Investigation, Data curation, Writing – original draft, Visualization. **Michael Yudistira Patuwo:** Conceptualization, Methodology, Resources, Writing – review & editing, Supervision, Project administration, Funding acquisition.

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data is provided in the submission package

Acknowledgement

This research did not receive any specific grant from funding agencies in the public, commercial, or not-for-profit sectors.

Appendix A

Principal component analysis (PCA)

PCA is an important tool in multivariate data analysis to reduce the dimensionality of the dataset and is achieved by a linear transformation of the original data set of variables into a smaller number of uncorrelated significant principal components (PCs). The maximum residual variance is given by the first principal component axis; the second principal component, orthogonal to the first one, has the second maximum variance and so on. In this way, maximum amounts of statistical information can be preserved and visualized [40,41]. Before applying PCA, the dataset was scaled to unit standard variation and zero mean so that the variables could be compared to each other on the same scale.

In this study, PCA was initially used to select 20 most potential predictors from the pool of 43 predictors. This is achieved by: first, predictors which have exceedingly small angles between their loading vectors, which suggests they are correlated [42], were removed to ensure no predictor carries the same information. Then, predictors with highest loading values of PC1, which suggested they were responsible for the variance of response, were selected. Consequently, the dataset with 20 predictors further underwent PCA to select linear terms for the model.

Appendix B

Logistic regression

1) Formula of logistic regression

Logistic regression is one of the most widely applied machine learning tools in binary classification problems [43]. The equation for multiple binary logistic regression model is as followed [44]. **X** is the predictor vector. π denotes the predicted probability of the compound being active, ranging from 0 (inactive) to 1 (active).

$$\pi(\mathbf{X}) = \frac{e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}}{1+e^{\beta_0+\beta_1 X_1+\cdots+\beta_k X_k}} = \frac{e^{\beta \bullet \mathbf{X}}}{1+e^{\beta \bullet \mathbf{X}}} = \frac{1}{1+e^{-\beta \bullet \mathbf{X}}}$$

Logistic regression uses Maximum Likelihood Estimation (MLE), in which parameters $\beta_0, \beta_1, \ldots, \beta_k$ are computed so that the final logistic regression model has maximum log-likelihood [45].

For a sample of size n, the likelihood for the logistic regression model is given by

$$L(\beta, y, X) = \prod_{i=1}^{n} \pi_i^{Y_i}(1-\pi_1)^{1-Y_i}$$

$$= \prod_{i=1}^{n} \left(\frac{e^{X_1\beta}}{1+e^{X_i\beta}}\right)^{Y_i} \left(\frac{1}{1+e^{X_i\beta}}\right)^{1-Y_i}$$

Thus, the log-likelihood is given by

$$l(\beta) = \sum_{i=1}^{n} [y_i \log(\pi_i) + (1-y_i)\log(1-\pi_i)]$$

$$= \sum_{i=1}^{n} \left[y_i X_i \beta - \log\left(1+e^{X_i\beta}\right)\right]$$

Log-likelihood of logistic regression plays a role analogous to the residual sum of squares in linear regression. The higher the value of the log-likelihood, the better a model fits a dataset.

2) The model's goodness of fit: McFadden's R^2

Among different R^2 for logistic regression, McFadden's R^2 (1974) [46] has been reported as one of the most used metric [47]. The formula for McFadden's R^2 is [48]:

$$R^2 = 1 - \frac{\ln L_{\text{full}}}{\ln L_0}$$

where L_{full} is the estimated likelihood of the full model, and L_0 is the estimated likelihood of the null model (model with only intercept).

3) Significance of predictors: the Wald test

Wald test is the test of significance for individual regression coefficients in logistic regression and is analogous to the *t*-test in linear regression. The Wald statistics where σ denotes standard deviation, is [44]:

$$Z_i = \frac{\widehat{\beta}_i}{\widehat{\sigma_{\beta_i}}}$$

The Wald statistics act as a Z-score. The standard normal curve is used to determine the p-value to test the null hypothesis $H_0 : \beta_i = 0$, which states the specified predictor is not significant in the model. In this study, the confidence level to reject the null hypothesis $H_0 : \beta_i = 0$ is 95%.

Calculating the standard deviation of coefficients in logistic regression

Where there are p predictors and n observations, the ways to calculate the standard deviation of regression coefficients in logistic regression is [49]:
Consider **X**, the design matrix

$$X = \begin{bmatrix} 1 & x_{1,1} & \cdots & x_{1,p} \\ 1 & x_{2,1} & \cdots & x_{2,p} \\ \cdots & \cdots & \cdots & \cdots \\ 1 & x_{n,1} & \cdots & x_{n,p} \end{bmatrix}$$

where $x_{i,j}$ is the value of the j^{th} predictor for the i^{th} observation
Consider matrix **V** with only values in the diagonal:

$$V = \begin{bmatrix} \widehat{\pi}_1(1\widehat{\pi}_1) & 0 & 0 & 0 \\ 0 & \widehat{\pi}_2(1-\widehat{\pi}_2) & \cdots & 0 \\ \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & \cdots & \widehat{\pi}_n(1-\widehat{\pi}_n) \end{bmatrix}$$

where $\widehat{\pi}_i$ represents the predicted probability for i^{th} observation. $\widehat{\pi}_i$ ranges from 0 to 1.

The covariance matrix can then be written as $(X^T V X)^{-1}$. The standard deviation of model coefficients is the square roots of the diagonal entries of this covariance matrix.

Appendix C

Wald test results from logistic regression

All predictors have the null hypothesis rejected in the Wald test at 95% confidence values with p-values lower than 0.05, as shown in Table 5.

Appendix D. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.rechem.2023.100888.

References

[1] Worldwide cancer data | World Cancer Research Fund International. (n.d.). WCRF International. Retrieved October 30, 2021, from https://www.wcrf.org/dietandcancer/worldwide-cancer-data/.
[2] Colorectal cancer statistics | World Cancer Research Fund International. (n.d.). WCRF International. Retrieved January 20, 2022, from https://www.wcrf.org/dietandcancer/colorectal-cancer-statistics/.
[3] Cancer. (n.d.). WHO | World Health Organization. Retrieved October 28, 2021, from https://www.who.int/news-room/fact-sheets/detail/cancer.
[4] H. Bohari, K. Srivastava, N. Sastry, Analogue-based approaches in anti-cancer compound modelling: the relevance of QSAR models, Org. Med. Chem. Lett. 1 (1) (2011 Jul 18) 3, https://doi.org/10.1186/2191-2858-1-3. PMID: 22373294; PMCID: PMC3279142.
[5] Radiation Therapy | CancerQuest. (n.d.). CancerQuest. Retrieved November 7, 2021, from https://www.cancerquest.org/patients/treatments/radiation-therapy.
[6] Chemotherapy as a Treatment for Colorectal Cancer. (2 C.E.). WebMD. https://www.webmd.com/colorectal-cancer/chemotherapy.
[7] A. Ibrahim, K.A. Oyebamiji, R. Oyewole, B. Semire, A DFT-based QSAR and molecular docking studies on potent anticolon cancer activity of pyrazole derivatives, Glob. J. Med. Res.: B Pharma, Drug Discovery, Toxicology & Medicine. (2018).
[8] J.A. DiMasi, H.G. Grabowski, R.W. Hansen, Innovation in the pharmaceutical industry: New estimates of R&D costs, J. Health Econ. 47 (2016) 20–33.
[9] S.Y. Ismail, A. Uzairu, In silico QSAR and molecular docking studies of sulfur containing shikonin oxime derivatives as anti-cancer agent for colon cancer. Radiol. Infect. Dis. 6 (3) (2019) https://doi.org/10.1016/j.jrid.2019.10.001.
[10] C. Hansch, T. Fujita, A method for the correlation of biological activity and chemical structure, J. Am. Chem. Soc. 86 (1964) 1616–1626.
[11] L. Wang, J. Ding, L. Pan, D. Cao, H. Jiang, X. Ding, Quantum chemical descriptors in quantitative structure–activity relationship models and their applications, Chemometr. Intell. Lab. Syst. (2021), https://doi.org/10.1016/j.chemolab.2021.104384.
[12] A. Cherkasov, N. Muratov, D. Fourches, et al., QSAR modeling: where have you been? Where are you going to? J. Med. Chem. 57 (12) (2014) 4977–5010, https://doi.org/10.1021/jm4004285.
[13] Zhao Y.S., Zhang X.P., Zhao J.H., Zhang H.Z., Kang X.J., Dong F., Research of QSPR/QSAR for ionic liquids. Prog. Chem., 24 (2012), pp. 1236-1244.
[14] T. Wang, X.S. Yuan, M.B. Wu, J.P. Lin, L.R. Yang, The advancement of multidimensional QSAR for novel drug discovery – where are we headed? Expet Opin. Drug Discov. 12 (2017) 769–784, https://doi.org/10.1080/17460441.2017.1336157.
[15] R. Guha, P.C. Jurs, Determining the validity of a QSAR model – a classification approach, J. Chem. Inf. Model. 45 (2005) 65–73.

[16] S. Kim, Y. Shin, S. Ahn, D. Koh, H. Lee, Y. Lim, Biological evaluation of 2-pyrazolinyl-1-carbothioamide derivatives against HCT116 human colorectal cancer cell lines and elucidation on QSAR and molecular binding modes, Bioorg Med Chem. 25 (20) (2017) 5423–5432, https://doi.org/10.1016/j.bmc.2017.07.062. Epub 2017 Aug 4 PMID: 28811071.
[17] P. Verma, C. Hansch, QSAR modeling of taxane analogues against colon cancer, Eur J Med Chem. 45 (4) (2010 Apr) 1470–1477, https://doi.org/10.1016/j.ejmech.2009.12.054. Epub 2010 Jan 13 PMID: 20116905.
[18] Deokar H., Deokar M., Wang W., Zhang R., Buolamwini K. QSAR Studies of New Pyrido[3,4-*b*]indole Derivatives as Inhibitors of Colon and Pancreatic Cancer Cell Proliferation. Med Chem Res. 2018 Dec;27(11-12):2466-2481. doi: 10.1007/s00044-018-2250-5. Epub 2018 Oct 3. PMID: 31360052; PMCID: PMC6662939.
[19] M. Rybka, A. Mercader, E. Castro, Predictive QSAR study of chalcone derivatives cytotoxicity activity against HT-29 human colon adenocarcinoma cell lines, Chemometr. Intell. Lab. Syst. 132 (2013), https://doi.org/10.1016/j.chemolab.2013.12.005.
[20] S. Cruz, E. Gomes, M. Borralho, P. Rodrigues, P. Gaudêncio, F. Pereira, In Silico HCT116 human colon cancer cell-based models en route to the discovery of lead-like anticancer drugs, Biomolecules 8 (3) (2018 Jul 17) 56, https://doi.org/10.3390/biom8030056. PMID: 30018273; PMCID: PMC6164384.
[21] Karelson M., Lobanov S., and Katritzky R. Quantum-Chemical Descriptors in QSAR/QSPR Studies. Chem. Rev. 1996, 96, 3, 1027–1044, Publication Date:May 9, 1996 https://doi.org/10.1021/cr950202r.
[22] Satbhaiya S. and Chourasia O. P. Scaffold and cell line based approaches for QSAR studies on anticancer agents, RSC Adv., 2015, 5, 103, 84810-84820, R. Soc. Chem., 10.1039/C5RA18295F, https://doi.org/10.1039/C5RA18295F.
[23] R. Katritzky, S. Lobanov, M. Karelson, CODESSA 2.0, comprehensive descriptors for structural and statistical analysis, University of Florida, 1994.
[24] Marković V., Erić S., Juranić D., Stanojković .T, Joksović L., Ranković B., Kosanić M., Joksović D. Synthesis, antitumor activity and QSAR studies of some 4-amino-methylidene derivatives of edaravone. Bioorg. Chem. 2011 Feb;39(1):18-27. doi: 10.1016/j.bioorg.2010.10.003. Epub 2010 Oct 28. PMID: 21078519.
[25] Tetko I.V., Gasteiger J., Todeschini R., Mauri A., Livingstone D., Ertl P., Palyulin V. A., Radchenko E.V., Zefirov N.S., Makarenko A.S., Tanchuk V.Y., Prokopenko V.V. Virtual computational chemistry laboratory — design and description, J. Comput. Aided Mol. Des. 19 (2005) 453–463] VCCLAB, Virtual Computational Chemistry Laboratory, 2005.
[26] C.W. Yap, PaDEL-descriptor: an open source software to calculate molecular descriptors and fingerprints, J Comput Chem 32 (7) (2011) 1466–1474.
[27] A.S. Girgis, J. Stawinski, N.S.M. Ismail, H. Farag, Synthesis and QSAR study of novel cytotoxic spiro[3H-indole-3,2′(1′H)-pyrrolo[3,4-c]pyrrole]-2,3′,5′(1H,2′aH,4′H)-triones, Eur. J. Med. Chem. 47 (2012) 312–322, https://doi.org/10.1016/j.ejmech.2011.10.058. - DOI - PubMed.
[28] A.S. Girgis, S.S. Panda, I.S.A. Farag, A.M. El-Shabiny, A.M. Moustafa, N.S. M. Ismail, G.G. Pillai, C.S. Panda, C.D. Hall, A.R. Katritzky, Synthesis, and QSAR analysis of anti-oncological active spiro-alkaloids, Org. Biomol. Chem. 13 (2015) 1741–1753, https://doi.org/10.1039/C4OB02149E. - DOI - PubMed.
[29] B. Zolnowska, J. Slawinski, M. Belka, T. Baczek, A. Kawiak, J. Chojnacki, A. Pogorzelska, K. Szafranski, Synthesis, molecular structure, metabolic stability and QSAR studies of a novel series of anticancer N-acylbenzenesulfonamides, Molecules 20 (2015) 19101–19129, https://doi.org/10.3390/molecules201019101.
[30] J. Slawinski, B. Zolnowska, Z. Brzozowski, A. Kawiak, M. Belka, T. Baczek, Synthesis and QSAR study of novel 6-chloro-3-(2-arylmethylene-1-methylhydrazino)-1,4,2-benzodithiazine 1,1-dioxide derivatives with anticancer activity, Molecules 20 (2015) 5754–5770, https://doi.org/10.3390/molecules20045754.
[31] P. Gramatica, E. Papa, M. Luini, E. Monti, M.B. Gariboldi, M. Ravera, E. Gabano, L. Gaviglio, D. Osella, Antiproliferative Pt (IV) complexes: Synthesis, biological activity, and quantitative structure–activity relationship modeling, J. Biol. Inorg. Chem. 15 (2010) 1157–1169, https://doi.org/10.1007/s00775-010-0676-4.
[32] Feature Interaction - PyCaret, 2020, https://pycaret.org/feature-interaction/.
[33] A. Speck-Planche, V. Kleandrova, F. Luan, M. Cordeiro, Rational drug design for anti-cancer chemotherapy: multi-target QSAR models for the in silico discovery of anti-colorectal cancer agents, Bioorg Med Chem. 20 (15) (2012 Aug 1) 4848–4855, https://doi.org/10.1016/j.bmc.2012.05.071. Epub 2012 Jun 15 PMID: 22750007.
[34] EBI-Team. ChEMBL Database. http://www.ebi.ac.uk/chembldb/, 2010.
[35] Amruthnath, N. (2020, June 24). Why balancing your data set is important? | R-bloggers. R-Bloggers; https://www.r-bloggers.com/2020/06/why-balancing-your-data-set-is-important/.
[36] Gaussian 16, Revision C.01, M. J. Frisch, G. W. Trucks, H. B. Schlegel, G. E. Scuseria, M. A. Robb, J. R. Cheeseman, G. Scalmani, V. Barone, G. A. Petersson, H. Nakatsuji, X. Li, M. Caricato, A. V. Marenich, J. Bloino, B. G. Janesko, R. Gomperts, B. Mennucci, H. P. Hratchian, J. V. Ortiz, A. F. Izmaylov, J. L. Sonnenberg, D. Williams-Young, F. Ding, F. Lipparini, F. Egidi, J. Goings, B. Peng, A. Petrone, T. Henderson, D. Ranasinghe, V. G. Zakrzewski, J. Gao, N. Rega, G. Zheng, W. Liang, M. Hada, M. Ehara, K. Toyota, R. Fukuda, J. Hasegawa, M. Ishida, T. Nakajima, Y. Honda, O. Kitao, H. Nakai, T. Vreven, K. Throssell, J. A. Montgomery, Jr., J. E. Peralta, F. Ogliaro, M. J. Bearpark, J. J. Heyd, E. N. Brothers, K. N. Kudin, V. N. Staroverov, T. A. Keith, R. Kobayashi, J. Normand, K. Raghavachari, A. P. Rendell, J. C. Burant, S. S. Iyengar, J. Tomasi, M. Cossi, J. M. Millam, M. Klene, C. Adamo, R. Cammi, J. W. Ochterski, R. L. Martin, K. Morokuma, O. Farkas, J. B. Foresman, and D. J. Fox, Gaussian, Inc., Wallingford CT, 2016.
[37] M. Khoshneviszadeh, N. Edraki, R. Miri, A. Foroumadi, B. Hemmateenejad, QSAR study of 4-aryl-4H-chromenes as a new series of apoptosis inducers using different chemometric tools, Chem. Biol. Drug Des. 79 (4) (2012 Apr) 442–458, https://doi.org/10.1111/j.1747-0285.2011.01284.x. Epub 2012 Jan 30 PMID: 22136072.
[38] Smeeks C., F., & C. Jurs, P. (1989). Prediction of boiling points of alcohols from molecular structure. Anal. Chim. Acta, 233 (1990) 111-119.
[39] Zach. (2020, November 3). A Quick Intro to Leave-One-Out Cross-Validation (LOOCV). Statology. https://www.statology.org/leave-one-out-cross-validation/.
[40] R.A. Johnson, D.W. Wichern, Applied Multivariate Statistical Analysis, Prentice-Hall, Englewood Cliffs, NJ, 1992.
[41] K.V. Mardia, J.T. Kent, J.M. Bibbly, Multivariate Analysis, Academic Press, New York, 1979.
[42] Jolliffe I., Cadima J. (2016). Principal component analysis: a review and recent developments. https://doi.org/10.1098/rsta.2015.0202.
[43] Enes M., Schmidt, Francis D. Review of Modern Logistic Regression Methods with Application to Small and Medium Sample Size Problems. AI 2010: Advances in Artificial Intelligence, 2011, Springer Berlin Heidelberg, Berlin, Heidelberg.
[44] 1 - Logistic Regression | STAT 462. (n.d.). Statistics Online | STAT ONLINE. Retrieved December 17, 2021, from https://online.stat.psu.edu/stat462/node/207/.
[45] A Gentle Introduction to Logistic Regression With Maximum Likelihood Estimation. (2019, October 27). Machine Learning Mastery; https://www.facebook.com/MachineLearningMastery/. https://machinelearningmastery.com/logistic-regression-with-maximum-likelihood-estimation/.
[46] D. McFadden, Conditional logit analysis of qualitative choice behaviour, University of Berkeley, California, 1974.
[47] Allison, P. (2013, February 13). What's the Best R-Squared for Logistic Regression | Statistical Horizons. Statistical Horizons | Statistics Training That Makes Sense. https://statisticalhorizons.com/r2logistic.
[48] Pseudo r-squared for logistic regression — Data Science Topics 0.0.1 documentation. (n.d.). Data Science Topics — Data Science Topics 0.0.1 Documentation. Retrieved December 17, 2021, from https://datascience.oneoffcoder.com/psuedo-r-squared-logistic-regression.html.
[49] Lecture 26 — Logistic regression. (2016). Stanford University. https://web.stanford.edu/class/archive/stats/stats200/stats200.1172/Lecture26.pdf.