DEVELOPMENT OF HEPG2 CELL LINE INHIBITORS FROM FLAVONOIDS OF *HYLOCEREUS UNDATUS* USING *IN SILICO* METHOD

Received: 21-05-2025

Do Thi Kim Thanh, Nguyen Le Hoang Son, Vu Le Ha, Pham Van Tat*

Institute of Pharmaceutical Training and Research, Binh Duong University, 504 Binh Duong Avenue, Thu Dau Mot City, 82000, Vietnam

Email: vantat@gmail.com

TÓM TẮT

PHÁT TRIỂN CÁC CHẤT ỨC CHẾ DÒNG TẾ BÀO HEPG2 TỪ FLAVONOID CỦA THANH LONG (*HYLOCEREUS UNDATUS*) SỬ DỤNG PHƯƠNG PHÁP *IN SILICO*

Hylocereus undatus (thanh long ruột trắng) là loại trái cây nhiệt đới giàu dinh dưỡng, chứa nhiều vitamin, khoáng chất và đặc biệt là các hợp chất chống oxy hóa như flavonoid, acid phenolic, betacyanin, beta-carotene và lycopene. Các chất này có khả năng bảo vệ tế bào khỏi tổn thương do gốc tự do, góp phần ngăn ngừa ung thư và làm chậm quá trình lão hóa. Nghiên cứu này nhằm phát triển các hợp chất có hoạt tính sinh học từ H. undatus với khả năng ức chế dòng tế bào ung thư gan HEPG2 thông qua kết hợp mô hình hồi quy tuyến tính bội (MLR) và mô phỏng docking phân tử. Mô hình MLR được xây dựng cho thấy độ tương quan cao (R^2 = 0,897; R^2 hiệu chỉnh = 0,885; MSE = 0,024; F = 75,038) và xác định điện tích nguyên tử là thông số quan trọng ảnh hưởng đến hoạt tính. Kết quả docking cho thấy hợp chất dẫn xuất từ Isorhamnetin (Fla1) có ái lực liên kết mạnh nhất với protein đích 6THA (ES = –7,627 kcal/mol), tiếp theo là các dẫn xuất của Quercetin và Kaempferol. Tất cả các dẫn xuất mới (Fla1–Fla4) đều có ái lực liên kết cao hơn so với các chất ban đầu, cho thấy tiềm năng cải thiện hoạt tính sinh học. Những kết quả này khẳng định tiềm năng của các flavonoid được tối ưu hóa từ H. undatus như những ứng viên hứa hẹn cho quá trình phát triển thuốc điều trị ung thư gan dựa trên nguồn gốc tự nhiên.

***Từ khóa*:** *Thanh long ruột trắng*, Mô hình $QSAR_{GA-MLR}$, dòng tế bào HEPG2, phương pháp *in silico*.

1. INTRODUCTION

Cancer continues to be one of the leading causes of death worldwide. Among the various types, liver cancer (HEPG2) is particularly prevalent, with high incidence and mortality rates. Current cancer treatments face numerous challenges, including high costs, drug resistance, and severe side effects. As a result, the development of new, more effective, and safer compounds from natural resources has become a significant trend in pharmaceutical research [1].

H. undatus contains many compounds with potential anticancer properties; however, there are currently no specific studies directly evaluating the effects of white-fleshed *H. undatus* on cancer cell lines such as HEPG2 (liver cancer). Therefore, further scientific study needed to determine the specific effectiveness of *H. undatus* in inhibiting these types of cancer cells. White-fleshed *H. undatus* is a tropical fruit belonging to the cactus family (Cactaceae), originally from Central America but now widely cultivated in many countries, especially

Vietnam. With its refreshing taste, rich nutritional content, and abundance of important bioactive compounds, white-fleshed H. undatus was not only a popular food but also held great potential in pharmaceutical and cosmetic applications [2],[3]. White-fleshed *H. undatus* contained many potent bioactive compounds, including flavonoids (quercetin, kaempferol, and isorhamnetin) that had antioxidant, anti-inflammatory, cardioprotective effects and support cancer treatment; phenolic acids (gallic acid, caffeic acid, ferulic acid) which could neutralize free radicals, provided antioxidant benefits, and enhanced liver health; betacyanins (betanin, betanidin), natural pigments with anti-inflammatory properties that protected cells and helped manage chronic diseases; prebiotics (soluble fiber, oligosaccharides) that supported digestive health, promoted beneficial gut bacteria, and improved immune function; as well as vitamins and minerals (vitamin C, B vitamins, iron, magnesium, calcium) that helped strengthen immunity, improved skin health, and enhanced overall well-being [4],[5].

The integration of *in silico* screening and QSAR (Quantitative Structure–Activity Relationship) modeling provided a powerful approach for the prediction, optimization, and development of novel bioactive compounds. This combined strategy enhances the efficiency of compound selection by reducing both time and cost, while simultaneously improving the accuracy of identifying candidates with potent biological activity [6].

Bioactive compounds from Hylocereus undatus were identified and developed as potential inhibitors of the MCF-7 breast cancer cell line using in silico screening approaches. Virtual screening techniques, including molecular docking, were applied to evaluate compound–protein interactions and to rank candidates based on binding affinities. Structural descriptors were combined with regression analysis and artificial neural networks to construct Quantitative Structure–Activity Relationship (QSAR) models. Flavone-based derivatives were designed and optimized using molecular mechanics (MM+) methods to ensure accurate structural geometry for computational evaluation [6],[7]. 2D and 3D molecular descriptors were used to build multivariate models such as multiple linear regression (MLR) and artificial neural networks (ANN). The QSAR models were developed to identify the molecular descriptor parameters that influenced the anti-breast cancer activity, thereby guiding the design of molecules with enhanced activity [7].

This study aimed to develop compounds that inhibited HEPG2 cells through a combined approach of *in silico* screening and $QSAR_{GA\text{-}MLR}$ modeling, which would help optimize the selection and development of potential compounds, thereby contributing to new prospects for cancer treatment using natural resources.

2. DATABASE AND METHODS

2.1. Database

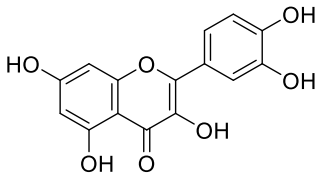
To ensure a reliable and comprehensive analysis, the molecular data used in this study were carefully collected and prepared from reputable sources. Molecular structures were standardized and curated to remove duplicates and incomplete entries. Physicochemical properties and relevant descriptors were calculated to support further modeling and screening processes. This rigorous data preparation laid the foundation for accurate and meaningful computational analysis. Molecular structure databases of

117 different flavonoid compounds with antioxidant and anticancer activities were screened from ChemBL database [8]. The structures of flavonoids were shown in Figure 1.

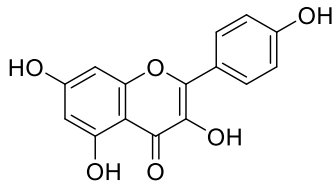
2D structure 3D structure

Figure 1. Structures of flavonoids

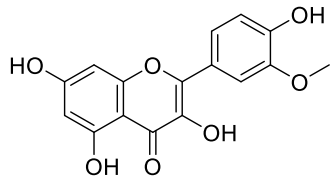
The dataset for this study was carefully compiled from multiple scientific sources to ensure comprehensive coverage of bioactive compounds. Molecular structures and biological activity information were curated and validated to support accurate screening and analysis. This systematic data preparation allowed for focused investigation of flavonoids with promising anticancer potential. According to various studies, flavonoids were a group of polyphenols known for their antioxidant and anticancer effects [1],[2],[3]. Some common flavonoids in white-fleshed *H. undatus* that exhibited biologically active properties and were of interest for cancer cell inhibition include [9]:



Quercetin

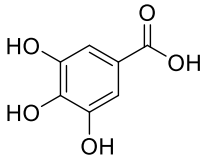


Kaempferol

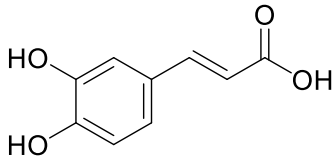


Isorhamnetin

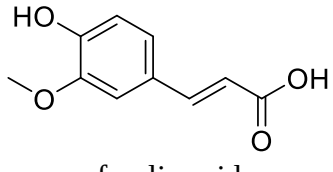
In addition to flavonoids, phenolic acids were another important group of bioactive compounds found in white-fleshed *Hylocereus undatus*. These compounds were well known for their strong cell-protective and anticancer properties. Understanding the specific phenolic acids present helped to further evaluate the fruit potential therapeutic effects. Some of the main phenolic acids identified in white-fleshed *H. undatus* include [10]:



gallic acid



caffeic acid



ferulic acid

Gallic acid, caffeic acid, and ferulic acid were well-studied phenolic compounds known for their potent antioxidant and anticancer activities. These acids helped neutralize harmful free radicals, reduced inflammation, and protected healthy cells from damage. Their ability to inhibit cancer cell growth, particularly in liver cancer, had been demonstrated in various studies. Understanding the effects of these compounds supports the development of new therapeutic agents. These compounds

all exhibited inhibitory effects on HepG2 cancer cell lines.

Although phenolic acids were present in *H. undatus*, their limited 3D structural data, lower molecular weights, and inconsistent bioactivity failed to meet Lipinski's criteria and $QSAR_{GA\text{-}MLR}$ requirements. Therefore, flavonoids were selected as the core scaffolds for designing novel bioactive derivatives with potential inhibitory activity against HEPG2 cells [9],[10]. White-fleshed *H. undatus* fruit contained a total phenolic content ranging from 25 to 55 mg gallic acid equivalent per 100 g, and a total flavonoid content of 15 to 35 mg equivalent per 100 g [11].

2.2. Computational methods

2.2.1. $QSAR_{GA\text{-}MLR}$ Method

QSAR (Quantitative Structure–Activity Relationship) models have emerged as powerful tools in computational drug design [11]. They enabled researchers to link chemical structure with biological activity using statistical or machine learning approaches. By analyzing molecular descriptors or fingerprints, QSAR models could provide insight into the key structural features responsible for a pharmacological effect. This facilitated virtual screening of large compound libraries and supported the design of more potent and selective drug candidates [7].

QSAR was a ligand-based method widely used to determine mathematical relationships between the biological activity of a molecule and its structural characteristics, expressed through molecular descriptors or fingerprints. Because it could reduce the time and cost associated with experimental testing, QSAR played an important role in identifying and optimizing bioactive compounds [12]. The $QSAR_{GA\text{-}MLR}$ approach combined the strengths of QSAR modeling, Genetic Algorithms (GA), and Multiple Linear Regression (MLR) to improve predictive accuracy in drug discovery [13]. GA was used as a feature selection method to identify the most relevant molecular descriptors from a large pool, while MLR helped build a robust mathematical model linking these descriptors to biological activity. This hybrid technique enhanced model interpretability and reliability, making it a powerful tool for screening and optimizing potential drug candidates efficiently.

The linear QSAR model represented a quantitative linear relationship between molecular structure and biological activity. Therefore, linear regression was an essential tool for constructing $QSAR_{GA\text{-}MLR}$ models [7],[11]. The $QSAR_{GA\text{-}MLR}$ model, in which biological activity depended on multiple molecular descriptor variables, was represented by the following equation [12]:

$$y = \sum_{i=1}^{n} a_i x_i + c \quad (1)$$

In the above equation, y represented the biological activity pIC_{50}; x_1, x_2, ..., x_n were the 2D and 3D molecular descriptors; a_1, a_2, ..., a_n were the corresponding regression coefficients; and a_0 is the intercept of the model. The quality of the $QSAR_{GA\text{-}MLR}$ model was evaluated based on statistical parameters calculated from both the training and test sets, including the coefficient of determination for the training set ($R^2_{training}$), model stability assessed via leave-one-out cross-validation (LOO), mean squared error (MSE) [11],[12]. These indices were calculated using the following formulas:

$$R^2_{training} = 1 - \frac{\sum_{i=1}^{n}(y_{i,\exp} - y_{i,\mathrm{cal}})^2}{\sum_{i=1}^{n}(y_{i,\exp} - \bar{y})^2} \quad (2)$$

$$MSE_{training} = \frac{\sum_{i=1}^{n}(y_{i,\exp} - y_{i,\mathrm{cal}})^2}{n} \quad (3)$$

Where $y_{\exp}$ experimental bioactivity value, y_{cal} bioactivity value calculated by the $QSAR_{GA\text{-}MLR}$ model on the training set and the test set, n: number of compounds used to build the model.

2.2.2. Docking calculation

Molecular docking using MOE 2024 followed a structured workflow designed to predict the binding affinity and orientation of ligands within a target protein's active site. The process began with importing the protein structure from the Protein Data Bank (PDB) [14], followed by cleaning steps such as removal of non-essential molecules (e.g., water), correction of protonation states, addition of hydrogen atoms, and energy minimization using force fields like AMBER10 or MMFF94x [15]. Ligands were then prepared by generating low-energy 3D conformations, optimizing geometry, and assigning correct charges. The binding site was identified either from co-crystallized ligands or predicted using MOE Site Finder tool [16]. Docking was carried out using the Triangle Matcher algorithm for pose generation, followed by rescoring using London dG and GBVI/WSA dG scoring functions to estimate the binding free energy ($\Delta G_{binding}$). To assess the accuracy of docking poses, the root-mean-square deviation (RMSD) between docked and reference ligand conformations was calculated. Ligands with low $\Delta G_{binding}$ values and acceptable RMSD (typically < 2.0 Å) were considered potential leads and are selected for further screening or experimental validation [17],[18].

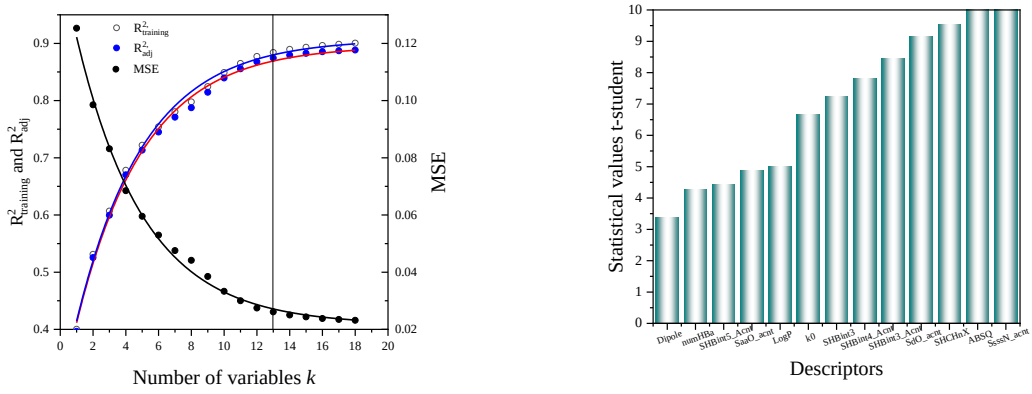
3. RESULTS AND DISCUSSION

3.1. Constructing $QSAR_{GA\text{-}MLR}$ Model

$QSAR_{GA\text{-}MLR}$ modeling often involved iterative optimization to select the most relevant molecular descriptors and improved predictive accuracy. This process helped in building robust models that could reliably predict biological activity for new compounds.

The $QSAR_{GA\text{-}MLR}$ model construction process was based on an evolutionary approach using genetic algorithms to find the best model. The structural and activity database of the flavonoid group was divided into two sets: a training set and a test set used for model building. The training set allowed for identifying predictive relationships. The predictive ability of the QSAR model was evaluated by comparing predicted values with the pIC_{50} activity of compounds in the test set. The linear $QSAR_{GA\text{-}MLR}$ models, which included 2D and 3D structural descriptors listed in Table 2, were calculated using the QSARIS system [19]. Among the $QSAR_{GA\text{-}MLR}$ models built from the training set (117 compounds), the predictive quality was assessed by comparing statistical values such as R^2_{train}, R^2_{pred}, and MSE. The variation in R^2 values, predicted correlation R^2_{pred}, MSE (standard error), and the 2D and 3D descriptor parameters in the $QSAR_{GA\text{-}MLR}$ models were clearly shown in Figures 2, 3, and Table 2. In order to evaluate the performance and significance of the $QSAR_{GA\text{-}MLR}$ model, statistical analysis was conducted based on different numbers of molecular descriptors. These analyses helped identify the optimal number of variables that provided the most reliable predictive power while maintaining model simplicity. The graph

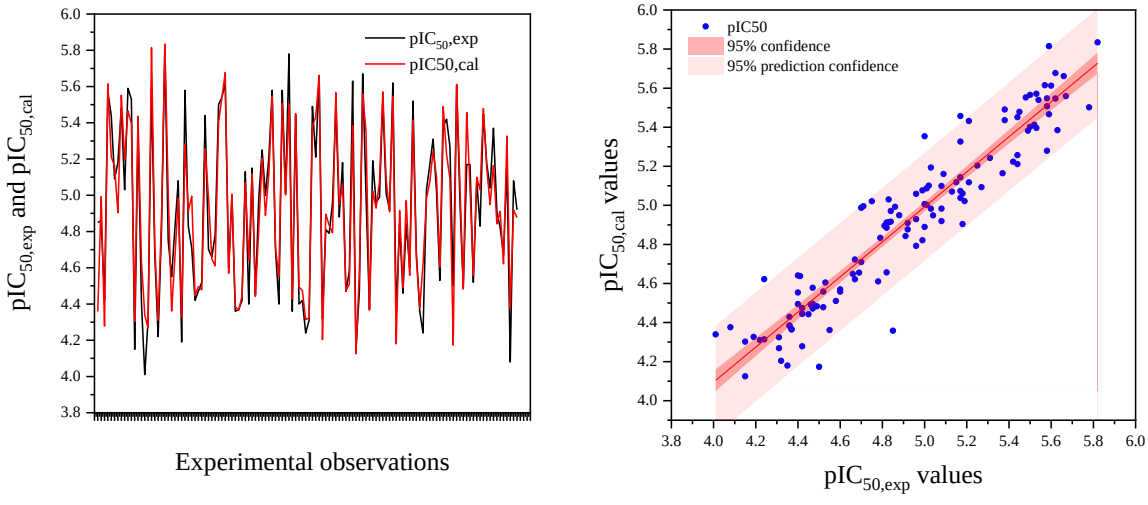
showing the variation in model statistics for 13 variables *k* in Figure 2a was based on the changes in $R^2_{training}$, R^2_{cal}, and the mean squared error (MSE). The graph indicated that the $QSAR_{GA\text{-}MLR}$ model with *k* = 13 achieved the highest $R^2_{training}$ value of 0.884 and a calculated value R^2_{cal} of 0.874. From Figure 2b, the contribution of each descriptor to biological activity was illustrated based on their t-student statistical values. The descriptors with the highest contribution were SHCHnX, ABSQ, and SsssN_acnt, while the descriptors with the lowest contribution was Dipole [19].



a) b)

Figure 2. Variation of statistical values with respect to the number of variables *k*; a) Statistical values of the model with 13 variables, b) t-Student statistical values indicating the contribution of molecular descriptors to biological activity in the training dataset.

To evaluate the reliability and predictive performance of the $QSAR_{GA\text{-}MLR}$ models, both visual and statistical analyses were carried out. These analyses focused on the relationship between predicted and experimental pIC_{50} values using multiple descriptors. They also helped determine the most suitable model based on statistical parameters such as $R^2_{training}$, R^2_{cal}, and MSE. From Figure 3, the correlation plots between predicted and experimental pIC_{50} values using 13 variables for the training set were shown.



a) b)

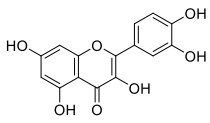
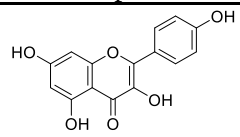
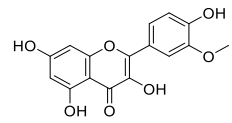
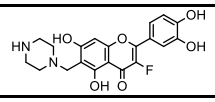
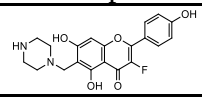
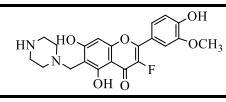
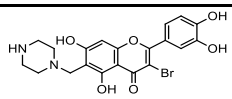
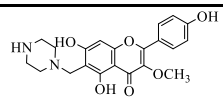
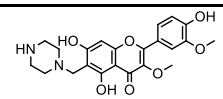
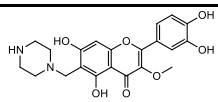
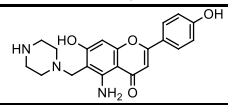
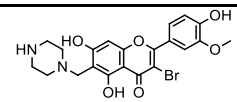
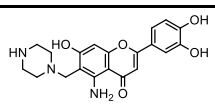
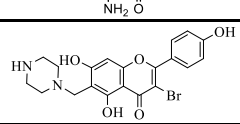
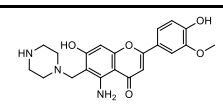
Figure 3. The training results of $QSAR_{GA\text{-}MLR}$ model; a) predictted results vs experimental cases; b) the correlation between predicted and experimental pIC_{50} values

Figure 3a illustrated the correlation between predicted and experimental pIC_{50} values, while Figure 3b showed that the data points closely align with the regression line, indicating a strong predictive performance of the model across all datasets. Among the 13 candidate models summarized in Table 1, the $QSAR_{GA\text{-}MLR}$ model with k = 13 exhibited the best statistical fit and was selected as the optimal model for further prediction.

Table 1. Statistical parameters and coefficients in the $QSAR_{GA\text{-}MLR}$ models with k = 1

Variables	Descriptors	Coefficients	SE	t_{stat} value	Pr > |t|
C	Intercept	4.831	0.082	58.554	< 0.0001
x_1	ABSQ	-0.383	0.038	-10.194	< 0.0001
x_2	Dipole	0.009	0.003	3.386	0,001
x_3	SsssN_acnt	0.252	0.025	10.240	< 0.0001
x_4	SdO_acnt	-0.264	0.029	-9.144	< 0.0001
x_5	SaaO_acnt	-0.283	0.058	-4.872	< 0.0001
x_6	k0	0.033	0.005	6.668	< 0.0001
x_7	SHCHnX	-0.882	0.093	-9.536	< 0.0001
x_8	SHBint3	-0.015	0.002	-7.239	< 0.0001
x_9	SHBint3_Acnt	0.268	0.032	8.466	< 0.0001
x_{10}	SHBint4_Acnt	-0.188	0.024	-7.802	< 0.0001
x_{11}	SHBint5_Acnt	-0.131	0.030	-4.431	< 0.0001
x_{12}	numHBa	0.090	0.021	4.277	< 0.0001
x_{13}	LogP	0.106	0.021	4.991	< 0.0001

Table 2. Structures of 4 newly designed flavonoid compounds and their predicted pIC_{50} values based on the $QSAR_{GA\text{-}MLR}$ model

No	Compound	pIC_{50}	Compound	pIC_{50}	Compound	pIC_{50}
lead	 Quercetin	4.679	 Kaempferol	4.650	 Isorhamnetin	4.631
Fla4		4.742		4.723		4.644
Fla3		5.431		5.365		5.437
Fla2		5.479		5.670		5.532
Fla1		5.632		5.672		5.647

The $QSAR_{GA\text{-}MLR}$ models with k ranging from 1 to 13 were listed in order, showing the variation of R^2, R^2_{cal}, and MSE values. Models with k between 11 and 13 had the highest R^2_{cal} values. To construct the $QSAR_{GA\text{-}MLR}$ model, statistical parameters

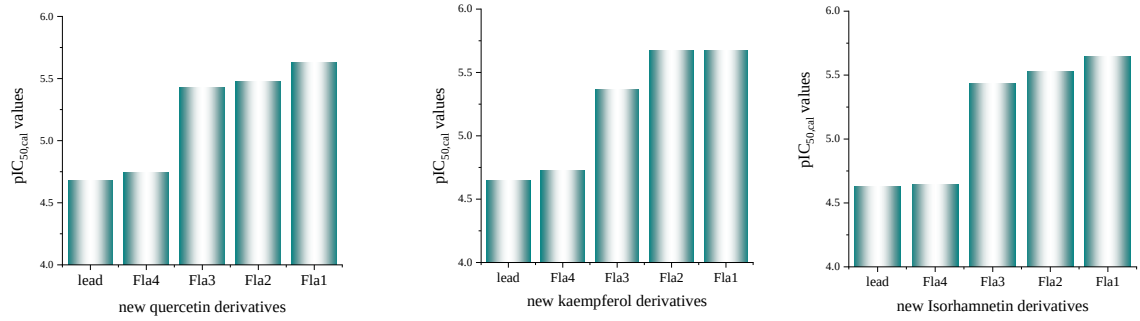
including coefficients, standard errors, t-values, and p-values play an essential role in developing the model using multiple linear regression. These statistical parameters and corresponding coefficients were presented in Table 3. The $QSAR_{GA-MLR}$ model was constructed using multiple linear regression combined with a genetic algorithm:

$$pIC_{50} = 4.831 - 0.383x_1 + 0.009x_2 + 0.252x_3 - 0.264x_4 - 0.283x_5 + 0.033x_6 - 0.882x_7 - 0.015x_8 + 0.268x_9 - 0.188x_{10} - 0.131x_{11} + 0.090x_{12} 0.106x_{13} \quad (4)$$

The statistical values used to evaluate the quality of the $QSAR_{GA-MLR}$ model were as follows:

$R^2 = 0.897$; $R^2_{Adj} = 0.885$; MSE = 0.024; giá trị F-stat = 75.038. The training dataset was well-described by the regression equation. The equation was statistically significant.

3.2. New molecular design



a) b) c)

Figure 4. The predicted results for four newly designed flavonoids from various lead compounds: a) the Quercetin; b) Kaempferol; c) Isorhamnetin.

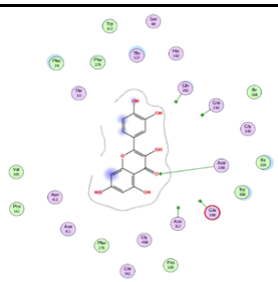
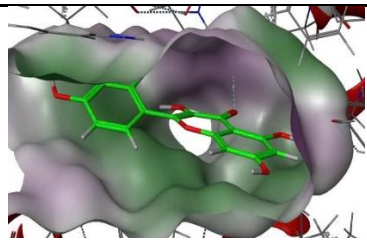
To explore the potential of new anticancer agents, molecular design was carried out by generating derivatives of key flavonoids found in *Hylocereus undatus*, namely Quercetin, Kaempferol, and Isorhamnetin. These new compounds were evaluated using previously developed QSAR models to predict their inhibitory activity against the HEPG2 liver cancer cell line. From Figure 1, new compounds were designed corresponding to each flavonoid. The QSAR models were developed and applied to predict the pIC_{50} activity against the HEPG2 cell line, as presented in Table 2. This representsed a key objective in new molecular design research, particularly in the development of compounds inhibiting cancer cell activity. The results also provided valuable insights into the practical applicability of the QSAR models. To rationally design new flavonoid derivatives with enhanced anticancer potential, we focused on optimizing key molecular descriptors identified in the $QSAR_{GA-MLR}$ model. Among these, SHCHnX (the sum of E-state values of sp^3-hybridized carbon atoms bonded to electronegative atoms), ABSQ (the absolute charge of the molecule), and SsssN_acnt (the count of tertiary nitrogen atoms) were shown to contribute significantly to inhibitory activity. Based on key molecular descriptors, a structure-based design approach was applied to generate four novel analogs from each of the flavonoids Quercetin, Kaempferol, and Isorhamnetin-compounds naturally present in *H. undatus*. These parent molecules were selected due to their reported anticancer potential. The predictive performance of the established $QSAR_{GA-MLR}$ model was employed to

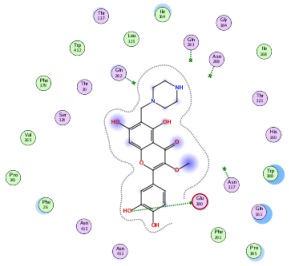
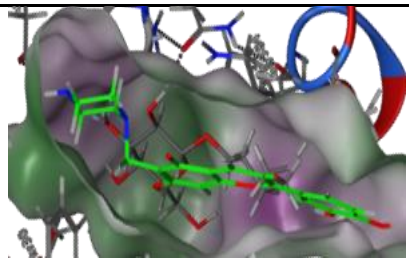
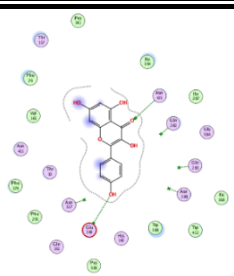
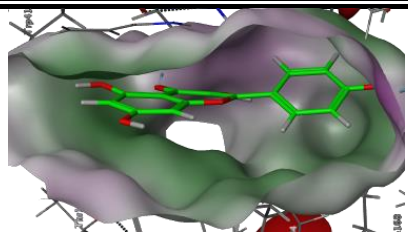
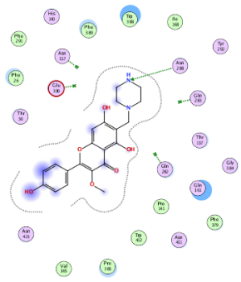
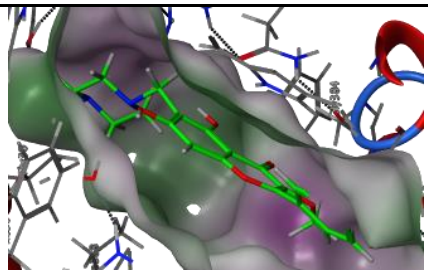
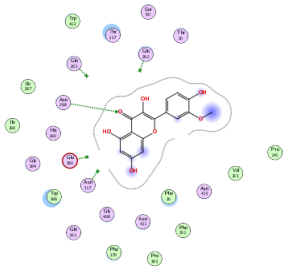
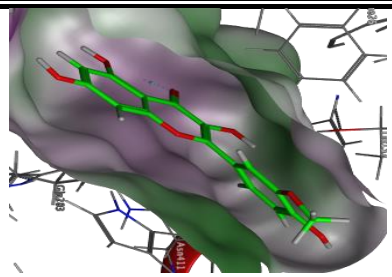
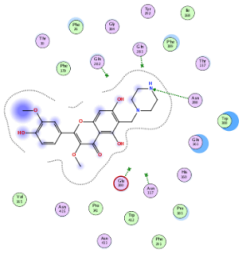
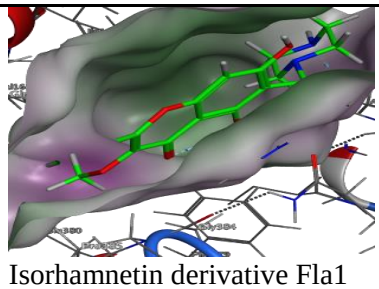
estimated the cytotoxic activity of the designed analogs against HEPG2 hepatocellular carcinoma cell lines. The predicted bioactivity data were presented in Table 2. The predicted pIC_{50} values from the $QSAR_{GA\text{-}MLR}$ model matched well with experimental data. These predicted activities would support further molecular docking studies. The newly designed flavonoid compounds exhibited higher predicted pIC_{50} values than their corresponding lead compounds, as showed in Figure 4.

3.3. Docking calculation

Before performing molecular docking in MOE, the crystal structure of the target protein 6THA was obtained from the RCSB Protein Data Bank (PDB) [14]. The structure was first loaded into MOE, and the Structure Preparation module was used to process the file. Water molecules, co-crystallized ligands, and irrelevant heteroatoms were removed unless they were involved in the active site. Hydrogen atoms were added using the Protonate 3D tool to assign proper protonation states and to optimize hydrogen-bonding interactions. The protonated structure was then subjected to partial energy minimization using the default Amber10:EHT force field, focusing on relieving steric clashes and correcting bond geometries, particularly around the binding pocket. After that, the binding site was defined either from the known ligand position in the crystal structure or using MOE Site Finder tool, which detected potential binding cavities. The processed protein structure was saved in MOE .moe or .mdb format for subsequent docking simulations [15],[16]. Prior to docking simulations in MOE, all ligand structures were first drawn or imported in 2D format using MOE Builder tool or imported from external databases such as ChEMBL [8]. The ligands were then converted into 3D format and their geometries were optimized using the Wash function, which ensured proper valency, tautomeric forms, and protonation states appropriate for physiological pH (typically ~7.4). Partial charges were assigned using the MMFF94x force field. Energy minimization was performed to relieve any strain and to generate a stable conformation using the Energy Minimize tool in MOE. Additionally, multiple low-energy conformations of each ligand were generated using the Conformation Import or Conformation Search tool to better simulate ligand flexibility [15]. The final prepared ligands were saved in MOE database (.mdb) format for use in the docking protocol. The docking results for the newly designed flavonoids with the highest predicted pIC50 values are presented in Table 5.

Table 5. Docking simulation results of the newly designed compounds

No	interaction 2D	interaction 3D	Docking Energy
Lead		 Quercetin	E_S = -5.943 kcal/mol RMSD = 1.153 Å

Fla1		 Quercetin derivative Fla1	E_S = -6.866 kcal/mol RMSD = 3.163Å
Lead		 Kaempferol	E_S = -5.699kcal/mol RMSD = 1.943 Å
Fla1		 Kaempferol derivative Fla1	E_S = -6.844 kcal/mol RMSD = 2.084Å
Lead		 Isorhamnetin	E_S = -6.0149 kcal/mol RMSD = 1.045Å
Fla1		 Isorhamnetin derivative Fla1	E_S = -7.627 kcal/mol RMSD = 2.705Å

Docking results indicated that compounds derived from white-flesh *H. undatus* possess potential inhibitory activity against HEPG2 liver cancer cells. The E-scores (binding energies) reflected the ability of the compounds to bind to the target protein 6THA. More negative E-scores indicated stronger and more stable interactions. Kaempferol showed the lowest binding energy among the parent compounds (E_S = –5.699 kcal/mol; RMSD = 1.943 Å), while the newly

designed Isorhamnetin derivative exhibited the strongest binding affinity (E_S = –7.627 kcal/mol; RMSD = 2.705 Å). These results suggest that the new Isorhamnetin compound could be a promising lead candidate for further research in liver cancer therapy. The other newly designed compounds also demonstrated strong potential in inhibiting the 6THA target protein involved in the growth of HEPG2 cells. As shown in Table 5, the newly designed compound Fla1 derived from Isorhamnetin exhibited the highest binding affinity, with an estimated binding energy (E_S) of −7.627 kcal/mol. This was followed by Fla1 derived from Quercetin (E_S = −6.866 kcal/mol) and that from Kaempferol (E_S = −6.844 kcal/mol). Notably, all Fla1 analogs outperformed their respective lead compounds in terms of predicted binding affinity, suggesting enhanced molecular interactions and potential for improved biological activity. These findings highlighted the promise of rationally optimized flavonoid derivatives as prospective candidates for further development in anticancer drug discovery pipelines. The newly designed compounds Fla1, derived from Quercetin, Kaempferol, and Isorhamnetin presented in Hylocereus undatus (dragon fruit), have demonstrated binding affinities toward the HEPG2-associated target protein with PDB ID: 6THA through multiple molecular docking interactions. These compounds exhibited potential inhibitory effects on HEPG2 cell proliferation by inducing cell cycle arrest at different phases (G0/G1 or G2/M). The binding mechanism involving the 6THA protein was associated with the regulation of cyclin expression (Cyclin D1, Cyclin B1) and cyclin-dependent kinases (CDK2, CDK4).

4. CONCLUSION

This study demonstrated the successful application of integrated in silico approaches, including $QSAR_{GA\text{-}MLR}$ modeling and molecular docking, to design and evaluate novel flavonoid derivatives from *H. undatus*. All designed analogs (Fla1–Fla4) exhibited enhanced binding affinities toward the 6THA target protein compared to their respective parent compounds, suggesting improved inhibitory potential against HEPG2 liver cancer cells. Notably, the Isorhamnetin-derived compound Fla1 showed the strongest interaction (E_S = −7.627 kcal/mol), highlighting its promise as a lead candidate. These findings underscore the potential of rationally optimized natural products in anticancer drug discovery and support further investigation of H. undatus-derived flavonoids for liver cancer therapy.

REFERENCES

[1] H. Nagai và Y. H. Kim, Cancer prevention from the perspective of global cancer burden patterns, *J. Thorac. Dis.*, Vol. 9(3), P. 448–451, (2017).

[2] Chen, Jy., Xie, Ff., Cui, Yz. *et al*. A chromosome-scale genome sequence of pitaya (Hylocereus undatus) provides novel insights into the genome evolution and regulation of betalain biosynthesis. *Hortic Res*, Vol. 8(1), P. 164, (2021). https://doi.org/10.1038/s41438-021-00612-0

[3] García-Cruz, L., Valle-Guadarrama, S., Salinas-Moreno, Y. et al. Physical, Chemical, and Antioxidant Activity Characterization of Pitaya (Stenocereus pruinosus) Fruits. *Plant Foods Hum Nutr*, Vol. 68, 403–410, (2013). https://doi.org/10.1007/s11130-013-0391-8

[4] Verona-Ruiz, A., Urcia-Cerna, J., & Paucar-Menacho, L. Pitahaya (Hylocereus spp.): Cultivo, características

fisicoquímicas, composición nutricional y compuestos bioactivos, Vol. 11(3), P. 439-453, (2020).

[5] F. F. Zhang, S. I. Barr, H. McNulty, D. Li, and J. B. Blumberg, Health effects of vitamin and mineral supplements, *BMJ*, Vol. 369, p. 2511, (2020), doi: 10.1136/bmj.m2511.

[6] P. V. Tat, B. T. P. Thuy, T. Duong, P. V. Trung, H. T. K. Dung, and P. N. N. Han, Prediction Of Anticancer Activities Of Kaempferol-3-O-Methylether And Kaempferol-3-O-(2,4-O-Diacetyl-Alpha-L-Rhamnopyranoside) Isolating From Plant Rhizome Zingiber Zerumbet Sm Using QSDAR Models From 13C-NMR and 15O-NMR Simulation Spectra Data, Org. Med. *Chem. Int. J*., Vol. 4(4), P. 82–84, (2017).

[7] P. Duchowicz, Linear Regression QSAR Models for Polo-Like Kinase-1 Inhibitors, *Cells*, Vol. 7(2), p. 13, (2018). DOI: 10.3390/cells7020013.

[8] A. Gaulton, A. Hersey, M. Nowotka, A. P. Bento, J. Chambers, D. Mendez, P. Mutowo, F. Atkinson, L. J. Bellis, E. Cibrián-Uhalte, M. Davies, N. Dedman, A. Karlsson, M. P. Magariños, and J. P. Overington, The ChEMBL database in 2017, *Nucleic Acids Research*, Vol. 45(D1), P. D945–D954, (2017). DOI: 10.1093/nar/gkw1074.

[9] D. M. Kopustinskiene, V. Jakstas, A. Savickas, and J. Bernatoniene, Flavonoids as Anticancer Agents, Nutrients, Vol. 12(2), Art.2, (2020). DOI: 10.3390/nu12020457.

[10]P. Esquivel, F. C. Stintzing, và R. Carle, Phenolic Compound Profiles and their Corresponding Antioxidant Capacity of Purple Pitaya (Hylocereus sp.) Genotypes, *Z. Für Naturforschung C*, Vol. 62, P. 636–644, (2007). DOI: 10.1515/znc-2007-9-1003.

[11] K. Roy, S. Kar, và R. N. Das, *Understanding the Basics of QSAR for Applications in Pharmaceutical Sciences and Risk Assessment*. Elsevier Inc., (2015). DOI: 10.1016/C2014-0-00286-9.

[12] H. Kubinyi, QSAR and 3D QSAR in drug design Part 1: methodology, Drug Discov. Today, Vol. 2(11), P. 457–467, (1997). DOI: 10.1016/S1359-6446(97)01079-9.

[13] A. K. Saxena and P. Prathipati, Comparison of MLR, PLS and GA-MLR in QSAR analysis, SAR QSAR *Environ. Res*., Vol. 14(5–6), P. 433–445, (2003). doi: 10.1080/10629360310001624015.

[14]H. M. Berman, J. Westbrook, Z. Feng, G. Gilliland, T. N. Bhat, H. Weissig, I. N. Shindyalov, and P. E. Bourne, The Protein Data Bank, Nucleic Acids Research, Vol. 28(1), P. 235–242, (2000). https://www.rcsb.org/structure/6THA

[15] S. Vilar, G. Cozza, and S. Moro, Medicinal Chemistry and the Molecular Operating Environment (MOE): Application of QSAR and Molecular Docking to Drug Discovery, Curr. Top. Med. Chem., vol. 8, no. 18, pp. 1555–1572, Dec. (2008).

[16]Chemical Computing Group Inc. *Molecular Operating Environment (MOE)*, 2023.09. Montreal, QC, Canada. https://www.chemcomp.com (accessed May 15, 2025).

[17] U. T. Tran, T. V. Pham, Design and screening of HepG2 cancer cell line inhibitors from Triterpenoid derivatives of Paramignya Trimera, *Sci. Technol. Dev. J.*, 2024, doi: 10.32508/stdj.v26iSI.4192.

[18] Sơn N. N., What is molecular docking? Scientific publications on molecular docking ?", Scholar Hub. Accessed: May 15, 2025. [Online]. Available at: https://scholarhub.vn/topic/docking phân tử

[19] Pham Van Tat, Development of Quantitative Structure-Activity Relationship And Quantitative Structure-Property Relationship, First reprint edition, Ha Noi, 2020.