frontiers | Frontiers in Bioinformatics

TYPE Original Research
PUBLISHED 30 September 2025
DOI 10.3389/fbinf.2025.1642039

OPEN ACCESS

EDITED BY
Nathan Brown,
Healx Ltd., United Kingdom

REVIEWED BY
Daniel Glossman-Mitnik,
Centro de Investigación de Materiales Avanzados (CIMAV), Mexico
Prosper Obed Chukwuemeka,
University of Pittsburgh, United States

*CORRESPONDENCE
José M. Alvarez-Suarez,
jalvarez@usfq.edu.ec
Eduardo Tejera,
eduardo.tejera@udla.edu.ec

RECEIVED 05 June 2025
ACCEPTED 18 September 2025
PUBLISHED 30 September 2025

CITATION
Villacrés M, Avila A, Jimenes-Vargas K, Machado A, Alvarez-Suarez JM and Tejera E (2025) Discovering molecules and plants with potential activity against gastric cancer: an *in silico* ensemble-based modeling analysis. *Front. Bioinform.* 5:1642039.
doi: 10.3389/fbinf.2025.1642039

Discovering molecules and plants with potential activity against gastric cancer: an *in silico* ensemble-based modeling analysis

Micaela Villacrés[1,2], Alec Avila[2], Karina Jimenes-Vargas[2,3], António Machado[4,5], José M. Alvarez-Suarez[6,7]* and Eduardo Tejera[2,8]*

[1]Clínica San Cayetano, Quito, Ecuador, [2]Bio-Cheminformatics Research Group, Universidad de Las Américas, Quito, Ecuador, [3]Departament of Computer Science and Information Technologies, Faculty of Computer Science, Universidade da Coruña, Campus Elviña s/n, A Coruña, Spain, [4]Departamento de Biologia, Centro de Biotecnologia dos Açores (CBA), Faculdade de Ciências e Tecnologia, Universidade dos Açores, Ponta Delgada, Portugal, [5]Laboratorio de Bacteriología, Colegio de Ciencias Biológicas y Ambientales COCIBA, Instituto de Microbiología, Universidad San Francisco de Quito USFQ, Quito, Ecuador, [6]Laboratorio de Investigación en Ingeniería en Alimentos (LabInAli), Departamento de Ingeniería en Alimentos, Colegio de Ciencias e Ingenierías, Universidad San Francisco de Quito (USFQ), Quito, Ecuador, [7]Laboratorio de Bioexploración, Colegio de Ciencias Biológicas y Ambientales, Universidad San Francisco de Quito (USFQ), Quito, Ecuador, [8]Facultad de Ingeniería y Ciencias Aplicadas, Universidad de Las Américas, Quito, Ecuador

Background: Gastric cancer (GC) remains a major global health burden despite advances in diagnosis and treatment. In recent years, natural products have gained increasing attention as promising sources of anticancer agents, including GC.

Methods: In this study, we applied an *in silico* ensemble-based modeling strategy to predict compounds with potential inhibitory effects against four GC-related cell lines: AGS, NCI-N87, BGC-823, and SNU-16. Individual predictive models were developed using several algorithms and further integrated into two consensus ensemble multi-objective models. A comprehensive database of over 100,000 natural compounds from 21,665 plant species, was screened for validation and to identify potential molecular candidates.

Results: The ensemble models demonstrated a 12–15-fold improvement in identifying active molecules compared to random selection. A total of 340 molecules were prioritized, many belonging to bioactive classes such as taxane diterpenoids, flavonoids, isoflavonoids, phloroglucinols, and tryptophan alkaloids. Known anticancer compounds, including paclitaxel, orsaponin (OSW-1), glycybenzofuran, and glyurallin A, were successfully retrieved, reinforcing the validity of the approach. Species from the genera *Taxus*, *Glycyrrhiza*, *Elaphoglossum*, and *Seseli* emerged as particularly relevant sources of bioactive candidates.

Conclusion: While some genera, such as *Taxus* and *Glycyrrhiza*, have well-documented anticancer properties, others, including *Elaphoglossum* and *Seseli*, require further experimental validation. These findings highlight the potential of

combining multi-objectives ensemble modeling with natural product databases to discover novel phytochemicals relevant to GC treatment.

KEYWORDS

gastric cancer prevention, plant-derived compounds, *in silico* screening, compound discovery, bioactive plant species, secondary metabolites

1 Background

Despite significant advances in medicine, gastric cancer remains a major global public health challenge, characterized by a dynamic historical evolution in its incidence, diagnosis, and treatment. Traditionally, it has ranked among the leading causes of cancer-related mortality, particularly in regions with a high prevalence of *Helicobacter pylori* infection and unhealthy dietary patterns. During the 20th century, the incidence of gastric cancer markedly declined in developed countries, primarily due to improvements in hygiene, widespread use of food refrigeration, and reduced consumption of salted and smoked foods (Kang et al., 2024). However, it remains a substantial cause of cancer-related deaths worldwide, with a heterogeneous geographical distribution. East Asia, Latin America, and Eastern Europe report the highest incidence rates, whereas North America and Western Europe have experienced a continuous downward trend (Mithany et al., 2024). In 2020, more than 1 million new cases of gastric cancer were diagnosed globally, accompanied by approximately 769,000 deaths, underscoring the persistent magnitude of this disease (Sung et al., 2021). Notably, the epidemiological profile of gastric cancer has undergone a significant shift in recent decades, with an increasing incidence observed among younger populations. This trend has prompted a reevaluation of preventive strategies and emphasizes the critical need for early-life interventions targeting modifiable risk factors (Kang et al., 2024). Several determinants of gastric cancer have been well-established. Infection with *H. pylori* remains the most prominent biological risk factor, often acting synergistically with behavioral and environmental influences (Poorolajal et al., 2020). Socioeconomic disparities further modulate the burden of disease. Lower educational attainment and limited access to healthcare services are associated with unhealthy lifestyles and delayed diagnosis, ultimately impacting survival outcomes (Alicandro et al., 2022). These complex, interrelated factors highlight the necessity for comprehensive, multidisciplinary approaches to the prevention, early detection, and management of gastric cancer globally.

In the treatment of gastric cancer, various chemotherapeutic agents have demonstrated significant efficacy in both *in vitro* cellular model systems and preclinical studies. Capecitabine, a prodrug of 5-fluorouracil (5-FU), has been shown to inhibit cell proliferation and angiogenesis in experimental models using BGC-823 cells, improving survival outcomes with low toxicity (Yuan et al., 2015). Similarly, docetaxel, an agent that disrupts microtubule polymerization, has proven effective by inducing G_2/M phase cell cycle arrest in AGS cells and exhibiting antiangiogenic and synergistic effects when combined with compounds such as gambogic acid in BGC-823 cells (Grabarska et al., 2023). Additional chemotherapeutic agents and alternative treatment strategies for gastric cancer have been extensively reviewed (Sexton et al., 2020; Guan et al., 2023). In parallel with the search for effective chemotherapeutic agents, increasing attention has been given to natural compounds. Curcumin, the principal bioactive component of turmeric, has demonstrated notable anti-inflammatory and antiproliferative properties relevant to the prevention and treatment of gastric cancer (Zhang et al., 2022). Various formulations of curcumin are currently being evaluated in clinical trials, such as NCT02782949, further supporting its potential application in gastric cancer management (Warias et al., 2024). Other natural molecules, including resveratrol, quercetin, and piceatannol, found in a variety of plant-derived products, have also shown the ability to modulate inflammatory processes and oncogenic pathways involved in gastric cancer progression (Zhao et al., 2023; Warias et al., 2024).

In the context of the discovery of novel natural products, *in silico* predictive modeling has emerged as a powerful methodology for identifying bioactive compounds. This approach employs computational tools to predict interactions between natural molecules and target proteins implicated in carcinogenesis, thereby accelerating the identification of novel therapeutic candidates (Liu et al., 2024a). Predictive modeling has been particularly instrumental in uncovering anticancer agents derived from food sources, notably polyphenols, which exhibit antioxidant and anti-inflammatory properties capable of reducing gastric cancer risk (Zheng et al., 2024). Computational strategies have been used to predict new drug targets, such as the epidermal growth factor receptor (EGFR) (Mashima et al., 2019), and to identify natural compounds, such as coumarin derivatives, capable of interacting with BCL2 and inducing apoptosis in gastric cancer cells (Perumalsamy et al., 2018). Moreover, the integration of network pharmacology approaches has enabled the identification of bioactive molecules like dehydroxy-isocalamendiol and spathulenol, which bind to critical cancer-related proteins, further highlighting their therapeutic potential (Pradhan et al., 2024). However, despite these advances, few studies have focused on large-scale screening of natural product libraries using phenotypic models, (Dai et al., 2016; Jin et al., 2020), as opposed to traditional target-centered modeling strategies (Jalali et al., 2023). Expanding the use of phenotypic screening could enhance the discovery of multifunctional compounds with broader mechanisms of action against gastric cancer. In this context, the present *in silico* study aims to identify effective molecules against gastric cancer (GC) cell lines, specifically AGS, NCI-N87, SNU-16, and BGC-823. Building upon our previous work (Perez-Castillo et al., 2018), we employed a consensus approach based on ensemble modeling, where individual predictive models were constructed and subsequently integrated to generate a final consensus probability for each compound. This methodology was applied to screen more than 100,000 molecules derived from 21,665 plant species. The ultimate objective is to pinpoint potential natural sources and plant species that could serve as promising candidates for further research in drug discovery and drug design collectively targeting GC cell lines.

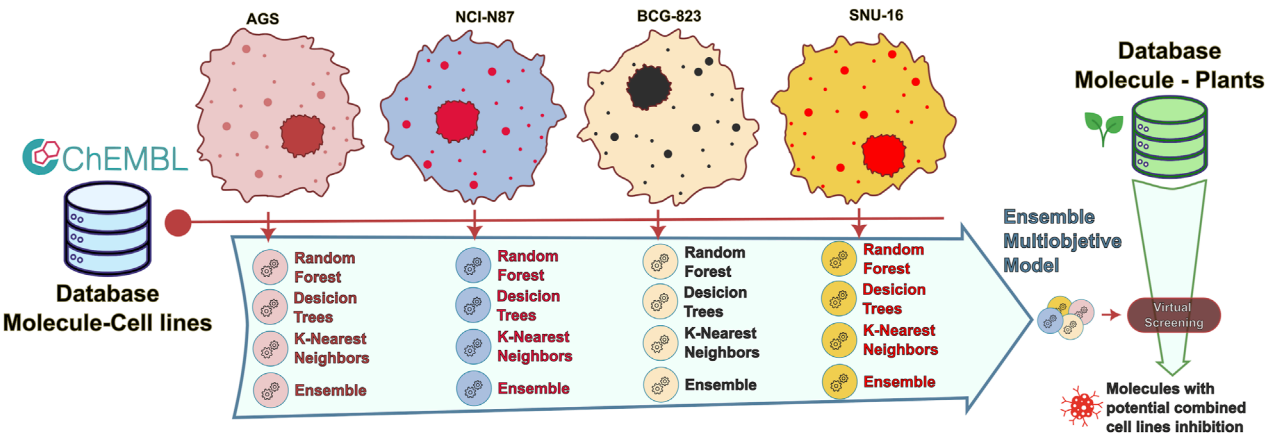

FIGURE 1
Flow diagram of the presented methodology.

2 Materials and methods

A full schematic representation of the methodology is presented in Figure 1. This representation will be described in detail across this section.

2.1 Database's description and curation

Four gastric cancer-related cell lines were selected for modeling: AGS (ChEMBL 3308078), NCI-N87 (ChEMBL 3307326), BGC-823 (ChEMBL 3307635), and SNU-16 (ChEMBL 3307273). All compounds with reported IC_{50} values were retrieved from the ChEMBL database, version 35 (Zdrazil et al., 2024). The data curation process followed a strategy similar to that described in previous studies (Perez-Castillo et al., 2018; Tejera et al., 2021). An IC_{50} threshold of 10 μM was employed to classify compounds as active (<10 μM) or inactive (>10 μM). When a molecule had multiple IC_{50} values reported for the same cell line across different studies, it was included only if all reports consistently classified it in the same activity class (i.e., all experiments agreed on its active or inactive status). If this criterion was not met, the compound–cell line pair was excluded. Additionally, compounds evaluated in more than two cell lines were excluded from the training sets and instead reserved for virtual screening purposes.

2.2 Modeling strategies and predictions

All molecules were described using ECFP4 fingerprints (1,024 bits) computed with RDKit (RDKit, 2018). Only ECFP4 description was used in this work. We had used this type of description previously in molecule-cell lines interaction across several cell lines (Tejera et al., 2019). However, we agree that it is not the only option available. Given that class imbalance is commonly observed, typically favoring either the inactive or active class, data balancing was performed through data reduction by applying a clustering algorithm to the majority class, following a strategy similar to that used in previous work (Tejera et al., 2021). Specifically, all molecules in the majority class (represented by their ECFP4 fingerprints) were clustered using the k-means algorithm (KMeans function from *sklearn. cluster*), incrementally increasing the number of clusters from 2 up to the number of compounds in the minority class. For each clustering step, the silhouette score was calculated (using the *silhouette*_score function from *sklearn. metrics*), and the number of clusters yielding the highest silhouette score was selected as optimal. Once the optimal number of clusters was determined, a proportional number of compounds was randomly selected from each cluster to match the size of the minority class. The final balanced datasets for each cell line are presented in Supplementary Material S1.

After balancing the data, a random split was performed for each cell line dataset into training, test, and external sets, following a 60%-20%-20% ratio. Prior to modeling, variable reduction was applied by removing all descriptors with a variance lower than 0.05 within the training subset. An important aspect of model evaluation (for both test and external sets) is the consideration of the applicability domain. To define this domain, a principal component analysis (PCA) was conducted on the training subset, extracting the principal components that together explained more than 90% of the cumulative variance. The maximum Euclidean distance between individual compounds and the centroid (computed using the selected principal components) was used to define the applicability domain. Any compound whose distance exceeded this maximum value was excluded from further analysis. The reason to use Euclidian distance is that the principal components are normalized numerically continuing description and it is fast, intuitive and simple to compute. In previous works we used the Tanimoto distance directly on the ECFP4 fingerprint (Jimenes-Vargas et al., 2024). This approach could be robust but computing the similarity matrix is computationally expensive over large datasets.

We evaluated four modeling strategies: random forest (RF), decision trees (DTREE), k-nearest neighbors (KNN), and an ensemble modeling approach combining models derived from RF, DTREE, and KNN. In the case of RF and DTREE we used Gini impurity to measure the quality of the split. Regarding the maximum depth of the tree, we initially explored several values

and decided to restrain to 100 for RF. In the case of DTREE all nodes are expanded until all leaves contain less than 2 samples. These parameters were not modified further in the analysis or optimization. Additionally, for variable selection in the RF, DTREE, and KNN models, a genetic algorithm was employed (Perez-Castillo et al., 2018; Tejera et al., 2021). The genetic algorithm was performed with an initial population of 1,000 individuals and was executed over 5,000 generations. To ensure model simplicity and prevent overfitting, the number of variables selected in each generated model was restricted to between 4 and 25. The balanced classification rate (BCR) was used as the fitness function for the genetic algorithm (Equation 1) (Perez-Castillo et al., 2018).

$$BCR = \frac{Se + Sp}{2}(1 - |Se - Sp|) \quad (1)$$

where *Se* and *Sp* are the sensitivity and specificity respectively. For each model, we computed: *BRC*, *Se*, *Sp*, F1-score, and accuracy for the test and external validation.

2.3 Ensemble modeling and virtual screening

For ensemble modeling, an initial population of 200 individual models was generated, each fulfilling the following criteria: (i) each model randomly used one of the RF, DTREE, or KNN algorithms; (ii) each model included between 4 and 25 variables; and (iii) each model achieved a BCR higher than 0.65 when evaluated on the test set. Based on these individual models, an initial ensemble population of 1,000 random combinations was created, with each ensemble comprising between 2 and 20 models. This initial ensemble population was further optimized using a genetic algorithm aimed at maximizing the BCR metric over 5,000 generations. In the final ensemble models, the mean probability of the individual models was used as the aggregation function to compute the final prediction.

For the virtual screening phase, several considerations must be addressed. The primary objective of the final models is to evaluate the probability that a compound exhibits anticancer activity, specifically against gastric cancer. Thus, the main challenge lies in accurately ranking compounds according to this criterion. However, defining anticancer activity based solely on cell line data presents difficulties, as (a) we are evaluating effects across four different gastric cancer cell lines, and (b) most compounds do not have activity data reported for all four cell lines, indeed, only two compounds in the ChEMBL database were found to have information across all four. To assess the performance of the combined cell line models in identifying potentially useful molecules, a virtual screening dataset was constructed. This dataset included: (i) 51 compounds from ChEMBL with reported activity against at least three of the four cell lines. Of these, 25 compounds were classified as active (active in at least two cell lines), while the remaining compounds were classified as inactive (active in only one or none of the cell lines). Additionally, (ii) compounds from the Genomics of Drug Sensitivity in Cancer (GDSC) database (Yang et al., 2012) were retrieved, specifically those evaluated across AGS, NCI-N87, and SNU-16 cell lines (BGC-823 data were not available in GDSC). A total of 125 compounds were identified and classified as active or inactive based on a z-score threshold of −1.5. Following a similar approach as with the ChEMBL dataset, 62 compounds were classified as active (active in at least two cell lines), and the remainder as inactive.

Finally, we also consulted the National Cancer Institute database (National Cancer Institute, 2025) and included four of the 22 small-molecule drugs currently used in the treatment of GC: capecitabine, docetaxel, doxorubicin, and mitomycin. Fluorouracil was not included, as it is already present in the GDSC database and was found to be inactive against the three gastric cancer cell lines evaluated. With this additional information, a total of 178 molecules were compiled for the virtual screening dataset, of which 86 were labeled as "active," indicating a higher likelihood of exhibiting anticancer activity.

One limitation of the constructed dataset is its relatively small size, which restricts the evaluation of enrichment metrics for virtual screening. For a robust calculation of virtual screening performance, a larger number of inactive (negative) molecules is required. As previously described (Perez-Castillo et al., 2018), decoy molecules were generated from the 86 active compounds. To this end, the DUD-E web service (Mysinger et al., 2012) was utilized, resulting in the generation of 4,357 decoy molecules, leading to a final dataset comprising 4,535 compounds. The early recognition metrics used to evaluate the models' performance in virtual screening are defined in Equations 2, 3 (Truchon and Bayly, 2007).

If $RIE_{min} = \frac{1-e^{-\alpha Ra}}{Ra(1-e^{\alpha})}$ and $RIE_{max} = \frac{1-e^{-\alpha Ra}}{Ra(1-e^{-\alpha})}$, then we can define BEDROC as:

$$BEDROC = \frac{RIE - RIE_{min}}{RIE_{max} - RIE_{min}} \quad (2)$$

In these equations, $Ra = n/N$, where n is the number of active compounds and N is the total number of molecules. The α-value corresponds to the portion of the ranked dataset where the recovery of active compounds is evaluated, representing early recognition performance. Mathematically, the α -value is associated with a fraction $0<\chi\leq1$, indicating the segment of the ranked list within which the active compounds are retrieved. This fraction is also necessary for computing the enrichment factor (EF), which is defined as:

$$EF = \frac{\sum_{i=1}^{n}\delta_i}{\chi^n}, \text{where } \delta_i = \begin{cases} 1 & r_i \leq \chi N \\ 0 & r_i > \chi N \end{cases} \quad (3)$$

Here, N represents the total number of compounds in the virtual screening list. Higher values of α correspond to smaller fractions of the ranked list used to retrieve active compounds, emphasizing early recognition. The computation of α can be adjusted depending on specific evaluation goals (Truchon and Bayly, 2007). For example, to evaluate enrichment within the top 1% of the ordered list while aiming for this fraction to contribute approximately 80% of the overall enrichment, an α-value of 160.9 is used. In our study, enrichment metrics were computed under different α-value and χ conditions to thoroughly assess model performance.

2.4 Compound-plants databases curation

To create a compound–plant database, information was integrated from FOODB (Foodb, 2025), COCONUT

(Sorokina et al., 2021), and LOTUS (Rutz et al., 2022) databases. COCONUT and LOTUS are specialized in natural products. In both the COCONUT and FOODB databases, full taxonomic identification of the source species is not always available. FOODB includes entries corresponding not to specific species but rather to processed or mixed foods (e.g., popcorn, cheese, milk), which were excluded from this analysis. To standardize species names, we used the National Center for Biotechnology Information (NCBI) taxonomy database (*fullnamelineage.dmp*), considering only plant species (Viridiplantae) that could be matched in NCBI records. Regarding compound curation, the RDKit package for Python was employed to perform the following operations: (1) removal of chiral information, (2) generation of InChIKeys, and (3) molecule sanitization. The removal of chiral information was based on two key considerations: (i) chiral descriptors are inconsistently reported across databases, either because the absolute configuration is unknown or because compounds exist as racemic mixtures, and (ii) the predictive models developed in this study do not account for chirality. Thus, treating enantiomers as distinct entities could artificially inflate the number of predicted active compounds. Consequently, enantiomers were considered duplicate entries when found within the same plant species. After filtering and cleaning the data, the final database included 21,665 plant species, 105,938 unique compounds, and 2,251,567 compound–species associations.

3 Results

Following the curation and balancing procedures, we characterized the datasets based on their active and inactive compound distributions. The final datasets (provided in Supplementary Material S1, SM1) comprised SNU-16 (n = 210; 100 active, 110 inactive), NCI-N87 (n = 247; 130 active, 117 inactive), BGC-823 (n = 1,565; 746 active, 819 inactive), and AGS (n = 791; 396 active, 395 inactive).

3.1 Predictive models and virtual screening

The performance of each applied algorithm is summarized in Table 1. We report both the best model identified within the final genetic algorithm populations and the mean performance metrics across the entire final population. Complete performance data for all models and ROC curves representations are provided in Supplementary Table S2.1 and Supplementary Figure S2.1 of Supplementary Material S2(SM2).

The selection of the best model from all models after genetic algorithm optimization is supervised. We consider a high BCR value in the test and external partitions. However, we also consider similar values between the mean BCR value from the genetic algorithm population and the BCR obtained in the external partition. This similarity is extended to other metrics like sensitivity, specificity and F1-score that can be consulted in SM2. We observe that, in some cases, ensemble models perform slightly better than the other models. For instance, this is the case for the AGS and BGC-827 cell lines. However, for the NCI-N87 cell line, decision tree models appear to outperform ensembles, while in the SNU-16 cell line, random forest and KNN models also seem to perform better. It is important to notice that we did not use cross-validation. The models are trained using the "training" partition and the fitness functions during model evolution are obtained by evaluation in the "test" group. The final models are evaluated in the "external" partition which is not used at any moment of the training or variable selection (Tropsha, 2010; Castillo-González et al., 2015). We can notice that the performance metrics are quite similar across the test group (presented as the average across the entire population), and the external partition which is a good indicator of the model's stability and generalization. All compounds used for virtual screening (n = 4,535) were evaluated with all models. To select the best combination of models providing the highest initial enrichment, we evaluated all possible model combinations (256 combinations) and computed the BEDROC score and enrichment factor (EF) across several initial enrichment fractions. The best combinations are presented in Figure 2 (individual model values are detailed in Supplementary Table S2.2; Supplementary Material S2).

Higher α-values assign greater weight to the BEDROC score in the early enrichment regions of the ranked dataset. Values around $\alpha = 160$ correspond approximately to the top 1% fraction (0.01 in Figure 2, Right) (Zhang et al., 2017). Notably, the EF results for the combinations E, DTREE, E, RF (dark blue line) and E, E, E, E (green line) are identical. In the screening procedure we desire models capable of correctly identifying (or retrieving) most of the active molecules (it is what BEDROC and EF quantify) in the minimal portion of the ranked list (it is computed by the fraction and/or the α -value). Analysis of the profiles in Figure 2 suggests that the two best-performing combinations are: C1) E, E, E, E (green line) and C2) E, E, E, RF (orange line). C2 exhibits a superior BEDROC score at $\alpha = 160$ and across higher α-values, while C1 achieves a higher EF. The enrichment factors (EF) of C1 and C2 in the top 0.5%–1% ($\chi = 0.005–0.01$) of the ranked list are between 12 and 16, suggesting suggestion 12 to 16 times better enrichment that what we should expect from random. In our case, we had a total of 86 active molecules in a total of 4,535, so in the top 1% (45 molecules), we ranked 14 and around 7 in the top 22 (close to 0.5%) ranked molecules. The maximum average probabilities obtained across the entire virtual screening dataset for C1 and C2 were 0.651 and 0.641, respectively.

3.2 Molecules and plants sources

After database curation as described, a total of 105,938 unique SMILES were obtained. However, after applying the applicability domain filters of the models comprising C1 and C2, 104,408 SMILES remained. Among these, 6,736 molecules (6.45%) in C1 and 5,512 (5.28%) in C2 exhibited a predicted probability greater than 0.5. The maximum predicted probabilities for C1 and C2 were 0.681 and 0.626, respectively, consistent with the peak values observed in the virtual screening dataset. Interestingly, only 91 molecules had a probability >0.6 in C1, compared to just 23 in C2, with only 8 compounds shared between the two models. This discrepancy suggests that the C2 model is more restrictive. The number of shared molecules between C1 and C2 at probability thresholds >0.5, >0.55, and >0.6 were 3,778, 340, and 8, respectively. No compounds were shared above the 0.65 threshold, as only C1 identified three molecules at that level. The 340 shared molecules at the 0.55

10.3389/fbinf.2025.1642039

TABLE 1 Performance metrics of the different modeling strategies, including ensemble approaches.

Cell Lines	Models	GA results		Best model				
		Test		Test		External		
		Mean ACC	Mean BCR	ACC	BCR	ACC	BCR	NV[a]
SNU-16	RF	0.825	0.806	0.857	0.776	0.810	0.793	19
	DTREE	0.796	0.778	0.786	0.748	0.857	0.825	26
	KNN	0.834	0.814	0.810	0.810	0.810	0.793	17
	ENSEMBLES	0.861	0.857	0.857	0.857	0.762	0.759	6
NCI-N87	RF	0.847	0.822	0.816	0.802	0.816	0.802	23
	DTREE	0.840	0.816	0.816	0.802	0.837	0.819	36
	KNN	0.812	0.787	0.816	0.802	0.837	0.819	25
	ENSEMBLES	0.929	0.915	0.918	0.909	0.857	0.838	11
BGC-827	RF	0.688	0.667	0.703	0.687	0.709	0.705	17
	DTREE	0.786	0.774	0.789	0.764	0.802	0.795	42
	KNN	0.788	0.775	0.792	0.782	0.805	0.782	33
	ENSEMBLES	0.858	0.851	0.863	0.853	0.815	0.809	15
AGS	RF	0.749	0.736	0.759	0.740	0.772	0.772	25
	DTREE	0.768	0.753	0.766	0.737	0.778	0.769	23
	KNN	0.764	0.749	0.753	0.744	0.778	0.769	36
	ENSEMBLES	0.856	0.845	0.880	0.869	0.823	0.823	8

[a]NV: Number of variables included in each model. For ensembles, this number indicates the number of models integrated into the final ensemble.

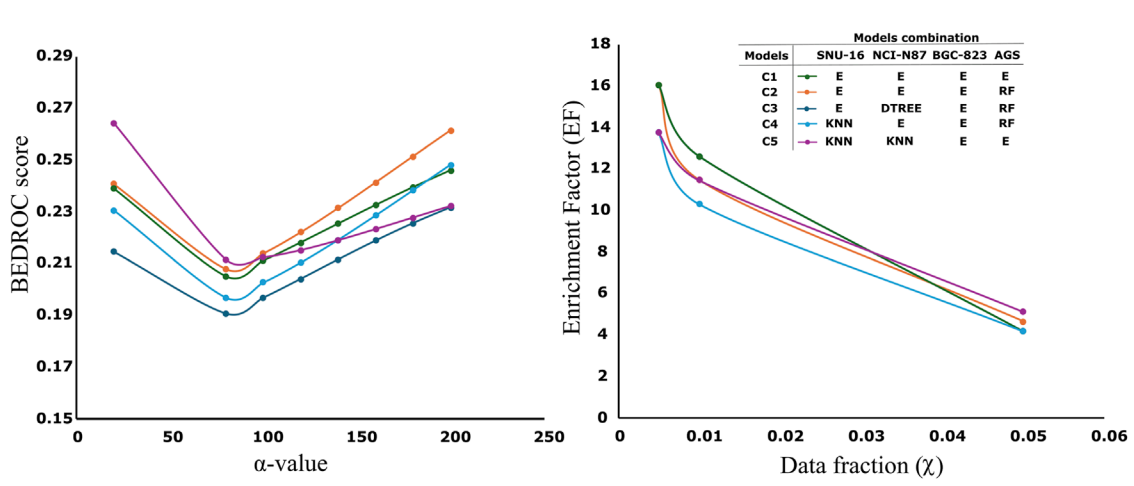

FIGURE 2
Left: Variation of the BEDROC values across different α-values for various model combinations. Right: Variation of the enrichment factor (EF) across different top fractions (χ) for the same model combinations. Each line represents a model combination from Table 1 for each cell line. The notation "E" denotes the ensemble model corresponding to each cell line, as described in Table 1.

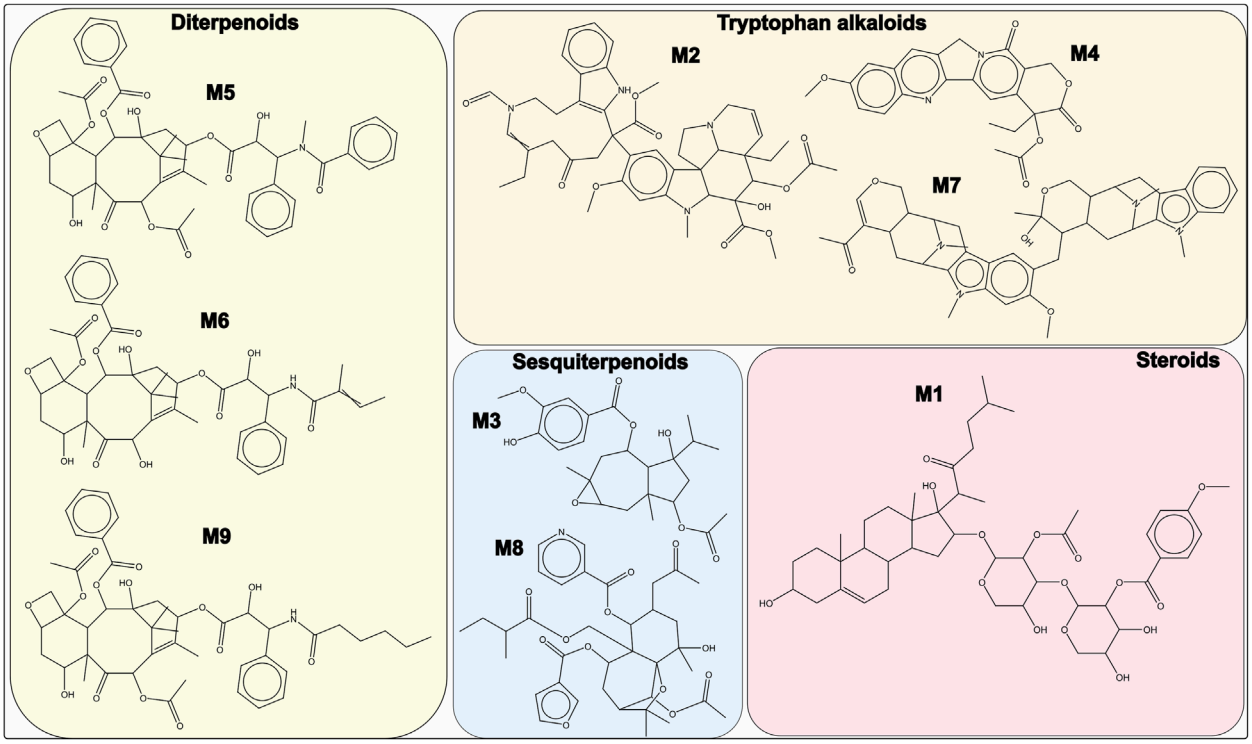

FIGURE 3
Molecules M1–M3 represent the top-ranked compounds in the C1 model (score >0.65). Molecules M1, M2, and M4–M9 are the eight compounds identified in both C1 and C2 models with scores >0.6. These represent the highest-scoring molecules across both models.

threshold were classified using NPClassifier (Kim et al., 2021), and their chemical classes, as well as the associated C1 and C2 probabilities, are provided in Supplementary Material S3 (SM3).

The top-ranked molecules (the top 3 from the C1 model and the 8 common molecules identified by both C1 and C2) are shown in Figure 3, along with their corresponding chemical classes. Molecule M3 is the only compound predicted by the C1 model with a score >0.65 that is not among the 8 shared molecules. Notably, M1 corresponds to orsaponin (PubChem CID: 72612554), M6 to 7-epi-10-Deacetyl Cephalomannine (PubChem CID: 72738999), and M9 to paclitaxel (PubChem CID: 23509308). The remaining molecules do not have common names or standardized notations in PubChem. It is also worth noting that M5 displays strong structural similarity to both M6 and M9. Flavonoids, isoflavonoids, di- and triterpenoids, sesquiterpenoids, and tryptophan-derived alkaloids represent the majority of the 340 molecules with combined probabilities greater than 0.55 (Figure 4A). Among the diterpenoids, taxane-type diterpenoids are the most abundant; this subclass includes compounds whose names are derived from the plant genus Taxus (to be discussed later). Within the flavonoid class, flavanones are the most represented subclass, followed by chalcones. In the isoflavonoid group, isoflavanones and pterocarpans are the most common. Pterocarpans are typically found in the Fabaceae family, while isoflavanones are broadly distributed, as are agarofuran and daucane (carotane-type) sesquiterpenoids.

The classes with the highest number of molecules, namely, flavonoids and diterpenoids, do not necessarily contain the compounds with the highest predicted probabilities (Figures 4B,C). The coefficient of determination (R^2) between the predicted probabilities of the C1 and C2 models across the entire set of 104,408 molecules is 0.743. However, this value drops substantially to R^2 = 0.108 when considering only the 340 shared molecules with a probability cutoff >0.55, suggesting that although the models are overall similar, they are specialized in distinct regions of chemical space. For instance, both models tend to assign higher probabilities to steroids and chromanes (Figure 4B), while phloroglucinols, coumarins, and diterpenoids are more highly ranked by the C2 model. In contrast, naphthalenes and linear polyketides are more strongly favored by the C1 model. Regarding maximum predicted probabilities, there is better consistency between C1 and C2 (Figure 4C), although phloroglucinols continue to be among the top-ranked classes in the C2 model. To identify the most relevant plant species, it is necessary to consider the number of potentially active compounds, their predicted probabilities, and the total number of compounds reported for each species. Some species are well-represented in the database, such as *Garcinia mangostana* and *Syzygium aromaticum*, while others are represented by only a few compounds, such as *Melicope durifolia* and *Turraea obtusifolia*. Out of a total of 21,665 species, only 1,045 (4.82%) contain at least one compound among the 340 molecules commonly identified by C1 and C2 models with a probability greater than 0.55. Moreover, only 215 species (0.99%) have two or more such compounds, and just 37 species (0.17%) have five or more. The average predicted probabilities from the C1 and C2

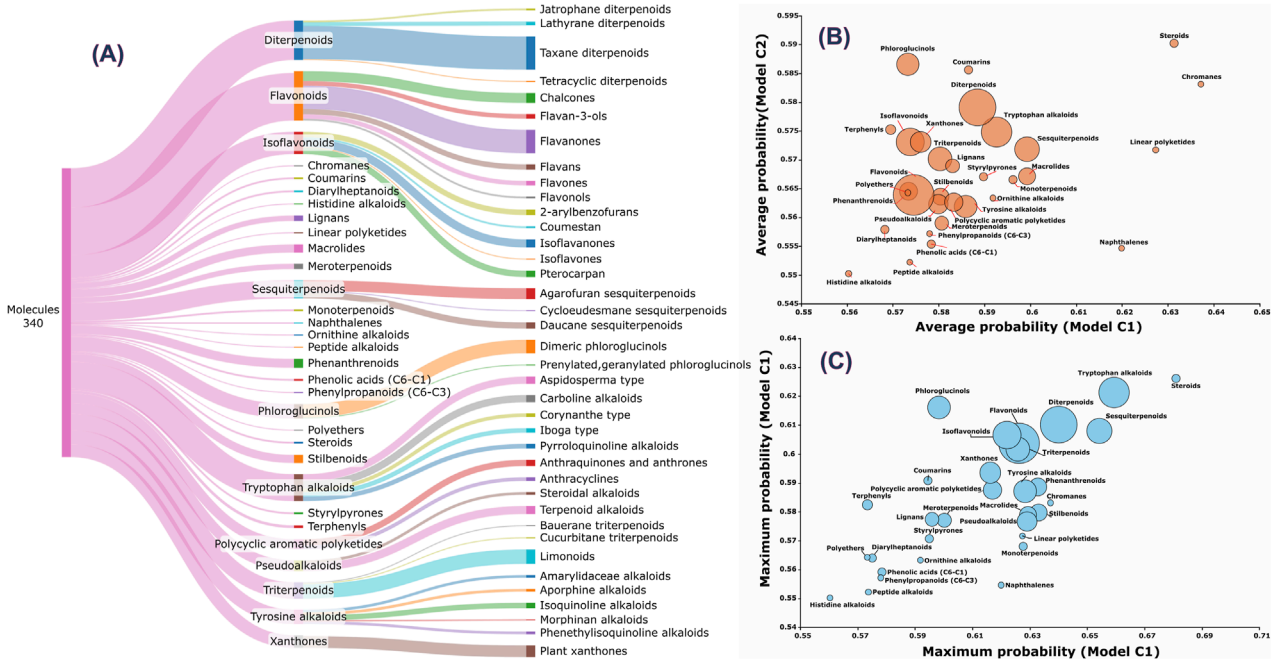

FIGURE 4
Analysis of molecular classes and subclasses. **(A)** Distribution of molecules by chemical class and subclass. **(B)** Average probabilities predicted by the C1 and C2 models for each chemical class. **(C)** Maximum probabilities predicted by the C1 and C2 models for each class. Bubble sizes in panels B and C represent are proportional to the number of compounds within each chemical class.

models, the number of identified active compounds, and the total number of compounds reported for each species are presented in SM3.

Several species from the genus *Taxus* (e.g., *Taxus baccata*, *Taxus cuspidata*) contain the highest number of predicted active compounds (Figure 5A), although the proportion of active compounds relative to the total number of molecules reported per species varies considerably (from 10% to 100%) (Figure 5B). In contrast, several species from the genus *Elaphoglossum* (*E. spatulatum*, *E. gayanum*, and *E. piloselloides*), despite having fewer compounds reported in the database (between 3 and 6), exhibit a high proportion (40–100%) of compounds with predicted probabilities >0.55.

Considering both absolute and relative representations, as well as the compound class results shown in Figure 4, the genera *Taxus*, *Elaphoglossum*, *Glycyrrhiza*, and *Seseli* appear particularly relevant for gastric cancer-related bioprospecting. In general, some chemical subclasses are distributed across various plant species, while others are more genus- or family-specific. For further analysis, all species from the aforementioned genera were grouped, resulting in 36, 20, 11, and 1 identified active compounds for *Taxus*, *Glycyrrhiza*, *Elaphoglossum*, and *Seseli*, respectively. Although several *Seseli* species appear in Figure 5B, they are all associated with a single compound (PubChem CID: 163026028), a tryptophan alkaloid. The dominant chemical classes identified within each genus were: 97.22% diterpenoids in *Taxus*, 85.00% isoflavonoids in *Glycyrrhiza*, 90.91% phloroglucinols in *Elaphoglossum*, and 100% tryptophan alkaloids in *Seseli*. This distribution highlights the presence of distinct bioactive scaffolds across different botanical lineages, underscoring their potential for targeted pharmacological exploration.

4 Discussion

In the individual models developed for each cell line, at least one modeling strategy achieved an accuracy greater than 0.8, although none surpassed 0.9. This trend was also observed in a previous study on the same cell lines (Perez-Castillo et al., 2018), except for one newly included cell line in the present work. In the previous work of (Perez-Castillo et al., 2018) we used a different fingerprint representation. Even though the objective is not focused on the analysis of the best chemical representation for this type of interactions, our results seem to indicate that ECFP4 description is slightly better than the description previously used. The ISIDA Fragmentor software (Varnek et al., 2008) was previously used to obtain 2D derived fingerprints descriptors that are quite different to the ECFP4. Even when the presented work is not directly comparable (i.e., different datasets) to our previous work, the performance metrics in the presented modelling are better. Our consensus models showed an accuracy range of 0.759–0.838 compared to 0.624–0.768 in the external dataset in our previous work (Perez-Castillo et al., 2018) while four and not three cell lines were used (we added the BGC-827). Several factors could explain this behavior including the increment of the compounds in the database. However, we should keep in mind that when modeling compound–cell line interactions, multiple mechanisms may underlie cell mortality and each mechanism could

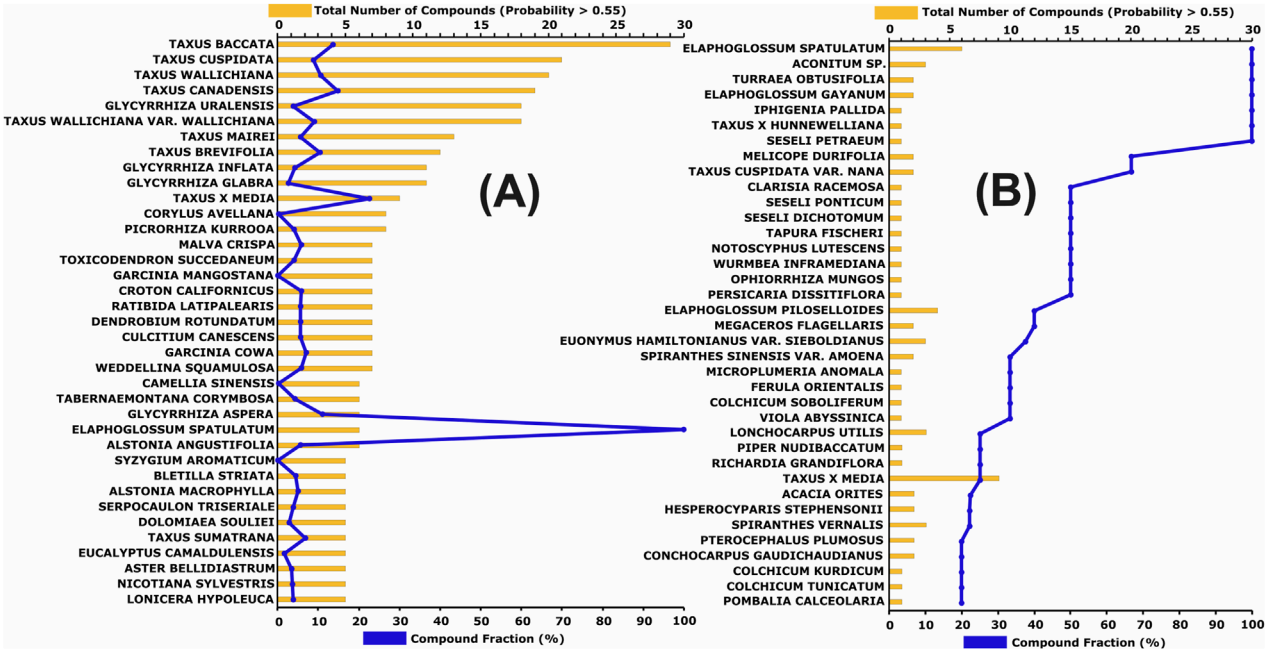

FIGURE 5
(A) Species with five or more active compounds, sorted by the total number of active compounds (orange bars) and the percentage of active compounds relative to the total number of molecules reported for the species (blue line). **(B)** Species sorted by the fraction of active compounds (blue line) and the corresponding number of active compounds (orange bars).

be potentially associated with distinct chemical spaces (i.e., targeting different molecular pathways or proteins). Moreover, the chemical space represented in the database for each cell line likely does not capture all these mechanisms uniformly. This hidden structure will increase the challenge for machine learning models to fully recognize and balance the internal chemical diversity. Notably, combining the models for different cell lines during virtual screening improved the chances of identifying active molecules by 12–15 times compared to random selection.

Despite the therapeutic advances achieved with chemotherapeutic agents such as capecitabine and docetaxel in the treatment of gastric cancer, significant clinical limitations persist, warranting the exploration of new pharmacological strategies. Treatment efficacy remains limited in patients with advanced-stage disease, and a high incidence of acquired resistance leads to early relapses and reduced overall survival (Liu et al., 2024b). Moreover, the cumulative toxicity associated with prolonged chemotherapy compromises quality of life, particularly in elderly patients or those with comorbidities (Rupp and Stengel, 2021). These challenges highlight the urgent need to investigate alternative or complementary therapies. Moreover, the research on natural products could also reveal new avenues on potential alternative mechanisms in gastric cancer inhibition or provide a synergic complement (Mao et al., 2020).

In the presented work, the goal of this virtual screening experiment was to evaluate the models' ability to retrieve compounds exhibiting inhibitory effects across more than two gastric cancer cell lines. Rather than relying on a single ensemble model, we employed the two best model combinations (C1 and C2) to enhance robustness. The observed differences in the average and maximum rankings of chemical classes between the two ensemble models (Figures 4B,C) suggest that the chemical space–activity relationship is represented differently in each model. This finding supports the strategy of combining both ensemble models during virtual screening to improve candidate molecule selection and decision-making processes.

The first noteworthy observation is that several of the molecules with the highest predicted probabilities in the C1 and C2 models are either well-known anticancer agents or structurally similar to such drugs. This result, independently, strengthens the reliability of the virtual screening approach but also indicates that some of the screened drugs could be acting as anticancer drugs but not necessarily specific to gastric cancer. For example, orsaponin and paclitaxel (Figure 3) have well-documented anticancer properties. Orsaponin (OSW-1) is a saponin isolated from *Ornithogalum saundersiae* that has demonstrated the ability to induce apoptosis in cancer cells in both *in vitro* and *in vivo* (xenograft) models (Zhang et al., 2017; Zhan et al., 2021). Although OSW-1 has been evaluated against several cancer types, such as colorectal and breast cancer, to our knowledge, it has not yet been studied in the context of gastric cancer or their cell lines. In contrast, paclitaxel is a broad-spectrum anticancer drug (Sharifi-Rad et al., 2021) that is also used in gastric cancer treatment, where it has been shown to improve overall survival (Fountzilas et al., 2024). In our database, paclitaxel was identified in *Corylus avellana*, *Taxus baccata*, *Taxus wallichiana*, *Taxus canadensis*, and other species of the *Taxus* genus. The detection of paclitaxel in *Corylus avellana* offers promising alternatives for biotechnological production through the cell culture of this species (Farhadi et al., 2020). Interestingly, *Corylus avellana* nuts are widely used in food products, and several studies, including

clinical trials, have suggested a positive effect of these products in reducing the risk of esophageal and gastric adenocarcinomas (Hashemian et al., 2017; Cao et al., 2023). Molecules M5 and M6 (7-epi-10-Deacetyl Cephalomannine) are structurally related to paclitaxel (as taxane diterpenoids), but no previous studies were found specifically addressing their bioactivity. A similar situation applies to the other top-ranked molecules selected by the C1 and C2 models and presented in Figure 3.

Among the 340 molecules with predicted probabilities greater than 0.55, the most abundant subclasses are taxane diterpenoids (diterpenoids), flavanones (flavonoids), isoflavanones (isoflavonoids), agarofuran and daucane sesquiterpenoids (sesquiterpenoids), dimeric phloroglucinols (phloroglucinols), various subclasses of tryptophan alkaloids, limonoids (triterpenoids), and plant xanthones (xanthones class) (Figure 4A). However, when considering the predicted probabilities, the dominant groups are steroids, tryptophan alkaloids, sesquiterpenoids, phloroglucinols, and isoflavonoids (Figures 4B,C). This trend is also reflected in the predominant genera identified in Figures 5A,B. Genera such as *Taxus*, *Elaphoglossum*, *Glycyrrhiza*, *Seseli*, along with specific species like *Turraea obtusifolia* and *Melicope durifolia*, appear highly relevant not only for phytotherapy but also for the discovery of new bioactive molecules targeting gastric cancer.

As previously mentioned, the genera *Taxus*, *Glycyrrhiza*, *Elaphoglossum*, and *Seseli* comprise groups of diterpenoids, isoflavonoids, phloroglucinols, and tryptophan alkaloids, many of which correspond to the chemical classes with the highest predicted probabilities. Among the isoflavonoids predicted at the top by the C2 model, we identified 2′,4′,7-trihydroxy-5-methoxy-3′,6-diprenylisoflavan, glycybenzofuran, glyurallin A, and 4-[4-hydroxy-6-methoxy-5-(3-methylbut-2-enyl)-1-benzofuran-2-yl]benzene-1,3-diol. All of these compounds are found within the Glycyrrhiza genus, and some have documented anticancer properties (Ito et al., 2020; Wu et al., 2022). Indeed, flavonoids (particularly flavanones) and isoflavonoids from various *Glycyrrhiza* species have been widely regarded as key contributors to the anticancer activity associated with these plants (Jain et al., 2022; Frattaruolo et al., 2024). In the phloroglucinol group, compounds with high predicted probabilities (especially from the C2 model) were predominantly identified in *Elaphoglossum*, including elaphopilosins A, B, and D, along with structurally related molecules. To our knowledge, these specific elaphopilosins have not yet been directly associated with anticancer activity; however, related phloroglucinols and extracts from the same genus have demonstrated inhibitory effects against several cancer cell lines (Arvizu-Espinosa et al., 2019). Taxane diterpenoids, as previously discussed, were strongly favored by both C1 and C2 models, and are abundant in *Taxus* species. In the case of *Seseli*, published evidence supports the anticancer activity of several species, although most studies have focused on essential oils (Cinar et al., 2020; Chen et al., 2024; Vaglica et al., 2024) or coumarins (Zengin et al., 2021; Onder et al., 2023), with no specific tryptophan alkaloids identified as responsible for the observed effects (Zengin et al., 2021). In our study, the compounds identified in *Seseli* belong to the tryptophan alkaloid class. Although tryptophan alkaloids are known to have pro-apoptotic effects in cancer cells (Guo et al., 2022), we did not find specific reports linking the particular molecule identified in Seseli species to anticancer activity.

Unfortunately, the molecules belonging to the steroid group (Figure 3, M1) and the tryptophan alkaloids (Figure 3, M2, M4, and M7) do not have any specific reports of biological activity. Among the plant groups identified, the genera *Elaphoglossum* and *Seseli* show the least available evidence regarding potential anticancer effects, both at the plant level and for the molecules identified (elaphopilosins and certain tryptophan alkaloids). A closer examination of our results also highlights additional plants and molecules that could serve as promising candidates for future experimental anticancer screening efforts.

5 Limitations and future perspectives

The integration of *in silico* ensemble-based modeling with natural product databases offers a powerful approach for accelerating the discovery of novel bioactive compounds against gastric cancer. However, several critical steps remain necessary to translate these computational predictions into therapeutic advances. First, we can´t be sure that these molecules will show specific activity to gastric cancer cell lines. It could be possible to display a wide anticarcinogenic effect not only focused on gastric cells and even normal vs. pathological cells. In the future, the inclusion of other cell lines could open the possibility to explore these specificities. Second, *in vitro* validation of the prioritized molecules, particularly those from underexplored genera (such as *Elaphoglossum* and *Seseli*), is essential to confirm their anticancer potential and to elucidate their mechanisms of action. Parallel assessment of cytotoxicity against normal gastric epithelial cells will be crucial to identify compounds with favorable therapeutic windows. Third, the incorporation of toxicity prediction models and absorption, distribution, metabolism, excretion, and toxicity (ADMET) profiling into future virtual screening pipelines will enhance the reliability and safety of selected candidates. Expanding the modeling framework to include additional gastric cancer subtypes and drug resistance models could further refine compound selection and clinical relevance. Moreover, future studies could explore the synergistic effects of compound mixtures, particularly from the same plant source, reflecting the complexity of natural extracts traditionally used in phytotherapy. Integration of multi-omics data (e.g., transcriptomics and proteomics) and systems biology approaches may also provide deeper insights into the multi-target potential of selected natural compounds. Ultimately, the combination of computational, experimental, and systems biology methodologies will be critical to fully exploit the therapeutic potential of plant-derived molecules, offering promising new strategies for gastric cancer prevention and treatment.

6 Conclusions

In this study, we developed and validated two ensemble-based predictive models targeting the inhibitory effects of natural compounds on four gastric cancer cell lines: AGS, NCI-N87, BGC-823, and SNU-16. The models achieved robust predictive performance and significantly enhanced the identification of bioactive molecules 12–15 times greater than random selection. Virtual screening of over 100,000 natural compounds from 21,665

plant species revealed both known anticancer agents (e.g., paclitaxel and orsaponin) and novel candidates belonging to underexplored chemical classes such as phloroglucinols and tryptophan alkaloids. The genera *Taxus*, *Glycyrrhiza*, *Elaphoglossum*, and *Seseli* emerged as promising botanical sources. While some have established pharmacological profiles, others represent untapped resources with limited or no prior evidence of anticancer activity. These findings highlight the potential of phenotypic *in silico* screening for uncovering multifunctional compounds of natural origin. Further experimental validation, including cytotoxicity assays and mechanistic studies, is essential. Moreover, the inclusion of predicting models related to ADMET and cells selectivity could improve these predictions and advance the discovery of safe and effective plant-derived agents for gastric cancer therapy.

Data availability statement

The original contributions presented in the study are included in the article/Supplementary Material, further inquiries can be directed to the corresponding authors.

Author contributions

MV: Data curation, Formal Analysis, Investigation, Methodology, Writing – original draft, Writing – review and editing. AA: Data curation, Formal Analysis, Investigation, Methodology, Software, Writing – original draft, Writing – review and editing. KJ-V: Data curation, Formal Analysis, Software, Writing – original draft. AM: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review and editing. JA-S: Conceptualization, Investigation, Methodology, Writing – original draft, Writing – review and editing. ET: Conceptualization, Formal Analysis, Funding acquisition, Investigation, Methodology, Software, Writing – original draft, Writing – review and editing.

Funding

The author(s) declare that financial support was received for the research and/or publication of this article. The research was funded by the Universidad de Las Américas, grant number: PRG.BIO.23.14.01.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

The author(s) declared that they were an editorial board member of Frontiers, at the time of submission. This had no impact on the peer review process and the final decision.

Generative AI statement

The author(s) declare that no Generative AI was used in the creation of this manuscript.

Any alternative text (alt text) provided alongside figures in this article has been generated by Frontiers with the support of artificial intelligence and reasonable efforts have been made to ensure accuracy, including review by the authors wherever possible. If you identify any issues, please contact us.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

Supplementary material

The Supplementary Material for this article can be found online at: https://www.frontiersin.org/articles/10.3389/fbinf.2025.1642039/full#supplementary-material

References

Alicandro, G., Bertuccio, P., Collatuzzo, G., Pelucchi, C., Bonzi, R., Liao, L. M., et al. (2022). The mediating role of combined lifestyle factors on the relationship between education and gastric cancer in the Stomach cancer Pooling (StoP) Project. *Br. J. Cancer* 127, 855–862. doi:10.1038/s41416-022-01857-9

Arvizu-Espinosa, M. G., von Poser, G. L., Henriques, A. T., Mendoza-Ruiz, A., Cardador-Martínez, A., Gesto-Borroto, R., et al. (2019). Bioactive dimeric acylphloroglucinols from the Mexican fern *Elaphoglossum paleaceum*. *J. Nat. Prod.* 82, 785–791. doi:10.1021/acs.jnatprod.8b00677

Cao, C., Gan, X., He, Y., Nong, S., Su, Y., Liu, Z., et al. (2023). Association between nut consumption and cancer risk: a meta-analysis. *Nutr. Cancer* 75, 82–94. doi:10.1080/01635581.2022.2104880

Castillo-González, D., Mergny, J.-L., De Rache, A., Pérez-Machado, G., Cabrera-Pérez, M. A., Nicolotti, O., et al. (2015). Harmonization of QSAR best practices and molecular docking provides an efficient virtual screening tool for discovering new G-quadruplex ligands. *J. Chem. Inf. Model* 55, 2094–2110. doi:10.1021/acs.jcim.5b00415

Chen, L., Ju, X., Wu, X., and Zuo, Z. (2024). Anti-melanoma cancer activity and chemical profile of the essential oil of *Seseli yunnanense* Franch. *Open Chem.* 22, 20240080. doi:10.1515/chem-2024-0080

Cinar, A. S., Bakar-Ates, F., and Onder, A. (2020). Seseli petraeum M. Bieb. (Apiaceae) significantly inhibited cellular growth of A549 lung cancer cells through G0/G1 cell cycle arrest. *An Acad Bras Cienc* 92, e20191533. doi:10.1590/0001-3765202020191533

Dai, S.-X., Li, W.-X., Han, F.-F., Guo, Y.-C., Zheng, J.-J., Liu, J.-Q., et al. (2016). *In silico* identification of anti-cancer compounds and plants from traditional Chinese medicine database. *Sci. Rep.* 6, 25462. doi:10.1038/srep25462

Farhadi, S., Salehi, M., Moieni, A., Safaie, N., and Sabet, M. S. (2020). Modeling of paclitaxel biosynthesis elicitation in Corylus avellana cell culture using adaptive neuro-fuzzy inference system-genetic algorithm (ANFIS-GA) and multiple regression methods. *PLoS One* 15, e0237478. doi:10.1371/journal.pone.0237478

FOODB (2025). *Foodb*. Available online at: https://foodb.ca/(Accessed April 23, 2025).

Fountzilas, E., Souglakos, J., Alafis, J., Dadouli, K., Koumarianou, A., Tsoukalas, N., et al. (2024). Real-world efficacy and toxicity data of paclitaxel and ramucirumab compared with other treatment regimens in patients with advanced gastric cancer. *ESMO Gastrointest. Oncol.* 5, 100073. doi:10.1016/j.esmogo.2024.100073

Frattaruolo, L., Lauria, G., Aiello, F., Carullo, G., Curcio, R., Fiorillo, M., et al. (2024). Exploiting Glycyrrhiza glabra L. (Licorice) flavanones: Licoflavanone's impact on breast cancer cell bioenergetics. *Int. J. Mol. Sci.* 25, 7907. doi:10.3390/ijms25147907

Grabarska, A., Luszczki, J. J., Gawel, K., Kukula-Koch, W., Juszczak, M., Slawinska-Brych, A., et al. (2023). Heterogeneous cellular response of primary and metastatic human gastric adenocarcinoma cell lines to magnoflorine and its additive interaction with docetaxel. *Int. J. Mol. Sci.* 24, 15511. doi:10.3390/ijms242115511

Guan, W.-L., He, Y., and Xu, R.-H. (2023). Gastric cancer treatment: recent progress and future perspectives. *J. Hematol. Oncol.* 16, 57. doi:10.1186/s13045-023-01451-3

Guo, M., Jin, J., Zhao, D., Rong, Z., Cao, L.-Q., Li, A.-H., et al. (2022). Research advances on anti-cancer natural products. *Front. Oncol.* 12, 866154. doi:10.3389/fonc.2022.866154

Hashemian, M., Murphy, G., Etemadi, A., Dawsey, S. M., Liao, L. M., and Abnet, C. C. (2017). Nut and peanut butter consumption and the risk of esophageal and gastric cancer subtypes. *Am. J. Clin. Nutr.* 106, 858–864. doi:10.3945/ajcn.117.159467

Ito, C., Matsui, T., Miyabe, K., Hasan, C. M., Rashid, M. A., Tokuda, H., et al. (2020). Three isoflavones from Derris scandens (Roxb.) Benth and their cancer chemopreventive activity and *in vitro* antiproliferative effects. *Phytochemistry* 175, 112376. doi:10.1016/j.phytochem.2020.112376

Jain, R., Hussein, M. A., Pierce, S., Martens, C., Shahagadkar, P., and Munirathinam, G. (2022). Oncopreventive and oncotherapeutic potential of licorice triterpenoid compound glycyrrhizin and its derivatives: molecular insights. *Pharmacol. Res.* 178, 106138. doi:10.1016/j.phrs.2022.106138

Jalali, Z., Nejad Ebrahimi, S., and Rezadoost, H. (2023). Identifying natural products for gastric cancer treatment through pharmacophore creation, 3D QSAR, virtual screening, and molecular dynamics studies. *Daru* 31, 243–258. doi:10.1007/s40199-023-00480-0

Jimenes-Vargas, K., Pazos, A., Munteanu, C. R., Perez-Castillo, Y., and Tejera, E. (2024). Prediction of compound-target interaction using several artificial intelligence algorithms and comparison with a consensus-based strategy. *J. Cheminform* 16, 27. doi:10.1186/s13321-024-00816-1

Jin, P., Ji, X., Kang, W., Li, Y., Liu, H., Ma, F., et al. (2020). Artificial intelligence in gastric cancer: a systematic review. *J. Cancer Res. Clin. Oncol.* 146, 2339–2350. doi:10.1007/s00432-020-03304-9

Kang, K., Bagaoisan, M. A., and Zhang, Y. (2024). Unveiling the younger face of gastric cancer: a comprehensive review of epidemiology, risk factors, and prevention strategies. *Cureus* 16, e62826. doi:10.7759/cureus.62826

Kim, H. W., Wang, M., Leber, C. A., Nothias, L.-F., Reher, R., Kang, K. B., et al. (2021). NPClassifier: a deep neural network-based structural classification tool for natural products. *J. Nat. Prod.* 84, 2795–2807. doi:10.1021/acs.jnatprod.1c00399

Liu, J., Xue, Y., Bai, K., Yan, F., Long, X., Guo, H., et al. (2024a). Experimental and computational study on anti-gastric cancer activity and mechanism of evodiamine derivatives. *Front. Pharmacol.* 15, 1380304. doi:10.3389/fphar.2024.1380304

Liu, J., Yuan, Q., Guo, H., Guan, H., Hong, Z., and Shang, D. (2024b). Deciphering drug resistance in gastric cancer: potential mechanisms and future perspectives. *Biomed. and Pharmacother.* 173, 116310. doi:10.1016/j.biopha.2024.116310

Mao, Q.-Q., Xu, X.-Y., Shang, A., Gan, R.-Y., Wu, D.-T., Atanasov, A. G., et al. (2020). Phytochemicals for the prevention and treatment of gastric cancer: effects and mechanisms. *Int. J. Mol. Sci.* 21, 570. doi:10.3390/ijms21020570

Mashima, T., Iwasaki, R., Kawata, N., Kawakami, R., Kumagai, K., Migita, T., et al. (2019). *In silico* chemical screening identifies epidermal growth factor receptor as a therapeutic target of drug-tolerant CD44v9-positive gastric cancer cells. *Br. J. Cancer* 121, 846–856. doi:10.1038/s41416-019-0600-9

Mithany, R. H., Shahid, M. H., Manasseh, M., Saeed, M. T., Aslam, S., Mohamed, M. S., et al. (2024). Gastric cancer: a comprehensive literature review. *Cureus* 16, e55902. doi:10.7759/cureus.55902

Mysinger, M. M., Carchia, M., Irwin, J. J., and Shoichet, B. K. (2012). Directory of useful decoys, enhanced (DUD-E): better ligands and decoys for better benchmarking. *J. Med. Chem.* 55, 6582–6594. doi:10.1021/jm300687e

National Cancer Institute (2025). Drugs approved for stomach (gastric) cancer - NCI. Available online at: https://www.cancer.gov/about-cancer/treatment/drugs/stomach (Accessed April 28, 2025).

Onder, A., Nahar, L., Cinar, A. S., and Sarker, S. D. (2023). The genus Seseli L.: a comprehensive review on traditional uses, phytochemistry, and pharmacological properties. *J. Herb. Med.* 38, 100625. doi:10.1016/j.hermed.2023.100625

Perez-Castillo, Y., Sánchez-Rodríguez, A., Tejera, E., Cruz-Monteagudo, M., Borges, F., Cordeiro, M. N. D. S., et al. (2018). A desirability-based multi objective approach for the virtual screening discovery of broad-spectrum anti-gastric cancer agents. *PLoS One* 13, e0192176. doi:10.1371/journal.pone.0192176

Perumalsamy, H., Sankarapandian, K., Veerappan, K., Natarajan, S., Kandaswamy, N., Thangavelu, L., et al. (2018). *In silico* and *in vitro* analysis of coumarin derivative induced anticancer effects by undergoing intrinsic pathway mediated apoptosis in human stomach cancer. *Phytomedicine* 46, 119–130. doi:10.1016/j.phymed.2018.04.021

Poorolajal, J., Moradi, L., Mohammadi, Y., Cheraghi, Z., and Gohari-Ensaf, F. (2020). Risk factors for stomach cancer: a systematic review and meta-analysis. *Epidemiol. Health* 42, e2020004. doi:10.4178/epih.e2020004

Pradhan, S. P., Gadnayak, A., Pradhan, S. K., and Epari, V. (2024). Integrating network pharmacology and *in silico* analysis to explore the bioactive compounds against gastric cancer treatment. *Cureus* 16, e75779. doi:10.7759/cureus.75779

RDKit (2018). *Open-source cheminformatics*. Available online at: http://www.rdkit.org.

Rupp, S. K., and Stengel, A. (2021). Influencing factors and effects of treatment on quality of life in patients with gastric cancer—a systematic review. *Front. Psychiatry* 12, 656929. doi:10.3389/fpsyt.2021.656929

Rutz, A., Sorokina, M., Galgonek, J., Mietchen, D., Willighagen, E., Gaudry, A., et al. (2022). The LOTUS initiative for open knowledge management in natural products research. *Elife* 11, e70780. doi:10.7554/eLife.70780

Sexton, R. E., Al Hallak, M. N., Diab, M., and Azmi, A. S. (2020). Gastric cancer: a comprehensive review of current and future treatment strategies. *Cancer Metastasis Rev.* 39, 1179–1203. doi:10.1007/s10555-020-09925-3

Sharifi-Rad, J., Quispe, C., Patra, J. K., Singh, Y. D., Panda, M. K., Das, G., et al. (2021). Paclitaxel: application in modern oncology and nanomedicine-based cancer therapy. *Oxid. Med. Cell Longev.* 2021, 3687700. doi:10.1155/2021/3687700

Sorokina, M., Merseburger, P., Rajan, K., Yirik, M. A., and Steinbeck, C. (2021). COCONUT online: Collection of open natural products database. *J. Cheminform* 13, 2. doi:10.1186/s13321-020-00478-9

Sung, H., Ferlay, J., Siegel, R. L., Laversanne, M., Soerjomataram, I., Jemal, A., et al. (2021). Global cancer Statistics 2020: GLOBOCAN Estimates of incidence and mortality worldwide for 36 cancers in 185 countries. *CA Cancer J. Clin.* 71, 209–249. doi:10.3322/caac.21660

Tejera, E., Carrera, I., Jimenes-Vargas, K., Armijos-Jaramillo, V., Sánchez-Rodríguez, A., Cruz-Monteagudo, M., et al. (2019). Cell fishing: a similarity based approach and machine learning strategy for multiple cell lines-compound sensitivity prediction. *PLoS One* 14, e0223276. doi:10.1371/journal.pone.0223276

Tejera, E., Pérez-Castillo, Y., Chamorro, A., Cabrera-Andrade, A., and Sanchez, M. E. (2021). A multi-objective approach for drug Repurposing in preeclampsia. *Molecules* 26, 777. doi:10.3390/molecules26040777

Tropsha, A. (2010). Best practices for QSAR model development, validation, and exploitation. *Mol. Inf.* 29, 476–488. doi:10.1002/minf.201000061

Truchon, J. F., and Bayly, C. I. (2007). Evaluating virtual screening methods: good and bad metrics for the "early recognition" problem. *J. Chem. Inf. Model* 47, 488–508. doi:10.1021/ci600426e

Vaglica, A., Maggio, A., Badalamenti, N., Bruno, M., Lauricella, M., Occhipinti, C., et al. (2024). Seseli tortuosum L. subsp. tortuosum essential oils and their principal constituents as anticancer agents. *Plants* 13, 678. doi:10.3390/plants13050678

Varnek, A., Fourches, D., Horvath, D., Klimchuk, O., Gaudin, C., Vayer, P., et al. (2008). ISIDA - platform for virtual screening based on fragment and pharmacophoric descriptors. *Curr. Comput. Aided-Drug Des.* 4, 191–198. doi:10.2174/157340908785747465

Warias, P., Plewa, P., and Poniewierska-Baran, A. (2024). Resveratrol, piceatannol, curcumin, and quercetin as therapeutic targets in gastric cancer—mechanisms and clinical implications for natural products. *Molecules* 30, 3. doi:10.3390/molecules30010003

Wu, Y., Wang, Z., Du, Q., Zhu, Z., Chen, T., Xue, Y., et al. (2022). Pharmacological effects and underlying mechanisms of licorice-derived flavonoids. *Evid. Based Complement. Altern. Med.* 2022, 9523071–25. doi:10.1155/2022/9523071

Yang, W., Soares, J., Greninger, P., Edelman, E. J., Lightfoot, H., Forbes, S., et al. (2012). Genomics of drug sensitivity in cancer (GDSC): a resource for therapeutic biomarker discovery in cancer cells. *Nucleic Acids Res.* 41, D955–D961. doi:10.1093/nar/gks1111

Yuan, F., Shi, H., Ji, J., Cai, Q., Chen, X., Yu, Y., et al. (2015). Capecitabine metronomic chemotherapy inhibits the proliferation of gastric cancer cells through anti-angiogenesis. *Oncol. Rep.* 33, 1753–1762. doi:10.3892/or.2015.3765

Zdrazil, B., Felix, E., Hunter, F., Manners, E. J., Blackshaw, J., Corbett, S., et al. (2024). The ChEMBL Database in 2023: a drug discovery platform spanning multiple bioactivity data types and time periods. *Nucleic Acids Res.* 52, D1180–D1192. doi:10.1093/nar/gkad1004

Zengin, G., Stojković, D., Mahomoodally, M. F., Jugreet, B. S., Paksoy, M. Y., Ivanov, M., et al. (2021). Comprehensive biological and chemical evaluation of two Seseli species (S. Gummiferum and S. Transcaucasicum). *Antioxidants (Basel)* 10, 1510. doi:10.3390/antiox10101510

Zhan, Z., Liu, Z., Lai, J., Zhang, C., Chen, Y., and Huang, H. (2021). Anticancer effects and mechanisms of OSW-1 isolated from Ornithogalum saundersiae: a review. *Front. Oncol.* 11, 747718. doi:10.3389/fonc.2021.747718

Zhang, Y., Fang, F., Fan, K., Zhang, Y., Zhang, J., Guo, H., et al. (2017). Effective cytotoxic activity of OSW-1 on colon cancer by inducing apoptosis *in vitro* and *in vivo*. *Oncol. Rep.* 37, 3509–3519. doi:10.3892/or.2017.5582

Zhang, W., Cui, N., Ye, J., Yang, B., Sun, Y., and Kuang, H. (2022). Curcumin's prevention of inflammation-driven early gastric cancer and its molecular mechanism. *Chin. Herb. Med.* 14, 244–253. doi:10.1016/j.chmed.2021.11.003

Zhao, N., Wang, W., Jiang, H., Qiao, Z., Sun, S., Wei, Y., et al. (2023). Natural products and gastric cancer: cellular mechanisms and effects to change cancer progression. *Anticancer Agents Med. Chem.* 23, 1506–1518. doi:10.2174/1871520623666230407082955

Zheng, Y., Ma, Y., Xiong, Q., Zhu, K., Weng, N., and Zhu, Q. (2024). The role of artificial intelligence in the development of anticancer therapeutics from natural polyphenols: current advances and future prospects. *Pharmacol. Res.* 208, 107381. doi:10.1016/j.phrs.2024.107381