

Article

Development of a Predictive Model for the Biological Activity of Food and Microbial Metabolites Toward Estrogen Receptor Alpha (ER α) Using Machine Learning

Maksim Kuznetsov ^{1,*}, Olga Chernyavskaya ¹, Mikhail Kutuzov ^{2,3}, Daria Vilkova ², Olga Novichenko ^{2,4}, Alla Stolyarova ^{5,6}, Dmitry Mashin ⁶ and Igor Nikitin ^{1,2,*}

- ¹ Department of Food Technology and Bioengineering, Plekhanov Russian University of Economics, 36 Stremyanny per., 115054 Moscow, Russia; olga.chernyavskaya.00@inbox.ru
- ² Research Laboratory of Applied Biotechnology, Cherepovets State University, 5 Lunacharsky pr., 162602 Cherepovets, Russia; kutuzov35@gmail.com (M.K.); ddvilkova@chs.ru (D.V.); ollevi@bk.ru (O.N.)
- ³ Apq Control LLC, Office 11, 45 Pervomajskaya st., 162612 Cherepovets, Russia
- ⁴ Department of Biotechnology, Aquaculture, Soil Science and Land Management, Astrakhan Tatishchev State University, 20a, Tatishchev st., 414056 Astrakhan, Russia
- ⁵ Basic Department of Trade Policy, Plekhanov Russian University of Economics, 36 Stremyanny per., 115054 Moscow, Russia; stolyarova2011@mail.ru
- ⁶ Department of Management and Economics, State University of Humanities and Social Studies, 30 Zelenaya st., 140400 Moscow, Russia; dmashin@ya.ru
- * Correspondence: 5726219@gmail.com (M.K.); nikito.igor@gmail.com (I.N.); Tel.: +7-906-730-70-10 (M.K.)

Abstract: The interaction of estrogen receptor alpha (ER α) with various metabolites—both endogenous and exogenous, such as those present in food products, as well as gut microbiota-derived metabolites—plays a critical role in modulating the hormonal balance in the human body. In this study, we evaluated a suite of 27 machine learning models and, following systematic optimization and rigorous performance comparison, identified linear discriminant analysis (LDA) as the most effective predictive approach. A meticulously curated dataset comprising 75 molecular descriptors derived from compounds with known ER α activity was assembled, enabling the model to achieve an accuracy of 89.4% and an F1 score of 0.93, thereby demonstrating high predictive efficacy. Feature importance analysis revealed that both topological and physicochemical descriptors—most notably FractionCSP3 and AromaticProportion—play pivotal roles in the potential binding to ER α . Subsequently, the model was applied to chemicals commonly encountered in food products, such as indole and various phenolic compounds, indicating that approximately 70% of these substances exhibit activity toward ER α . Moreover, our findings suggest that food processing conditions, including fermentation, thermal treatment, and storage parameters, can significantly influence the formation of these active metabolites. These results underscore the promising potential of integrating predictive modeling into food technology and highlight the need for further experimental validation and model refinement to support innovative strategies for developing healthier and more sustainable food products.

Keywords: machine learning; linear discriminant analysis (LDA); metabolites; estrogen receptor alpha (ER α); food technology; functional food products



Academic Editors: Cristian Randieri, Giuseppe Tradigo, Riccardo Pecori and Jakub Mieczkowski

Received: 31 January 2025

Revised: 8 March 2025

Accepted: 28 March 2025

Published: 1 April 2025

Citation: Kuznetsov, M.; Chernyavskaya, O.; Kutuzov, M.; Vilkova, D.; Novichenko, O.; Stolyarova, A.; Mashin, D.; Nikitin, I. Development of a Predictive Model for the Biological Activity of Food and Microbial Metabolites Toward Estrogen Receptor Alpha (ER α) Using Machine Learning. *Big Data Cogn. Comput.* **2025**, *9*, 86.
<https://doi.org/10.3390/bdcc9040086>

Copyright: © 2025 by the authors. Licensee MDPI, Basel, Switzerland. This article is an open access article distributed under the terms and conditions of the Creative Commons Attribution (CC BY) license (<https://creativecommons.org/licenses/by/4.0/>).

1. Introduction

Modern biological and medical sciences face a formidable challenge: the need for the rapid and precise identification of novel bioactive compounds capable of interacting with key molecular targets, such as protein receptors that act as transcription factors. These receptors play a critical role in regulating gene expression [1,2] and in governing essential

biological processes [3–6], including cellular growth, differentiation, metabolism, and immune response. Dysregulation in these systems often underlies socially significant diseases such as cancer, diabetes, neurodegenerative disorders, and autoimmune conditions.

Traditional approaches for the discovery, identification, and testing of new bioactive compounds require enormous amounts of time and financial resources. Experimentally evaluating thousands—or even millions—of potential substances to determine their activity toward specific receptors is labor-intensive and costly. Thus, a promising direction in this field is the development and application of methodologies based on modern machine learning (ML) technologies, which are opening new horizons in drug discovery and the identification of key metabolite effectors [7].

Estrogen receptor alpha (ER α) is a pivotal nuclear receptor and transcription factor that regulates various physiological processes, including the development of the reproductive system, bone homeostasis, and cardiovascular function [8–10]. Moreover, its involvement in hormone-dependent cancers—especially breast cancer—underscores its significant biological importance [11–14]. Despite extensive research on ER α and its established ligands, the interactions between various metabolites and ER α remain relatively underexplored. Emerging evidence suggests that microbial byproducts can significantly impact human metabolism and health [15–21], implying that a deeper understanding of these interactions could unveil novel mechanisms underlying hormonal balance and metabolic regulation. Additionally, food-derived metabolites, such as xenoestrogens, may play an important role in ER α activation [22].

In light of these findings, our study is focused on predicting the potential biological activity of various metabolites toward ER α using machine learning [23,24]. Within this work, we developed 27 machine learning models optimized through cross-validation methods, enabling a comprehensive comparison of different algorithms. Our analysis demonstrated that the linear discriminant analysis (LDA) model provided the best performance across key metrics, which justified its selection for further application.

To construct the training dataset, we employed 75 molecular descriptors that encompass a broad range of physicochemical, topological, and electronic properties of the compounds, thereby ensuring a comprehensive characterization of the metabolites under study. A dataset was curated from the ChEMBL database [25], comprising compounds with quantitatively measured activity (IC₅₀ values). The quantitative activity data were transformed into a binary variable (Activity) using a threshold of 1000 nM—a value widely accepted in pharmacological research to distinguish active (IC₅₀ $\leq$ 1 μ M) from inactive compounds [26]. This threshold ensures consistency with established criteria and facilitates the reliable training of predictive models.

Among the various machine learning approaches, LDA was selected for its interpretability and efficiency in class discrimination [27,28]. The LDA model was developed and optimized using the selected dataset, achieving an accuracy of 89.4% and an F1 score of 0.93. Optimization was carried out using GridSearch, a method that exhaustively evaluates all combinations of specified hyperparameters to determine the optimal model configuration. This approach allows for an objective identification of the best settings within a relatively constrained parameter space, making it particularly suitable for such tasks [29]. Supporting literature confirms that LDA is effectively employed in similar contexts, offering clear interpretability while maintaining high predictive power [30,31]. Furthermore, the model coefficients provided valuable insights into feature importance, highlighting descriptors such as FractionCSP3 and AromaticProportion as critical for ER α binding [32,33].

Additionally, this study evaluates selected microbial metabolites, such as indole and various phenolic compounds, which are commonly found in food products. These metabo-

lites were chosen based on their frequent occurrence in food matrices and their reported biological activity [34]. Although the current model is predictive and does not account for the complexity of actual food systems—where the food matrix may influence biological activity—our approach provides a useful preliminary screening tool. It emphasizes the potential for optimizing food production methods by modulating processing conditions (e.g., fermentation, thermal treatment, and storage parameters) to promote the formation or preservation of beneficial metabolite activity.

This study is not merely a demonstration of a new model; it is part of a broader methodology that leverages the potential of genomics and machine learning in preserving health and preventing the development of socially significant diseases through the modulation of food systems and management of human microbiota composition. It lies at the intersection of personalized nutrition and personalized medicine and their impact on public health.

The objective of this work is to elucidate the relationships between the complex compositional data and properties of a vast array of bioactive food molecules and the practical knowledge required to understand various physiological processes in the human body—specifically, to predict the biological activity of food metabolites toward estrogen receptor alpha ($ER\alpha$). The extensive spectrum of metabolites involved in receptor activation and the transcriptional regulation of physiological processes remains largely unexplored.

Overall, this study not only broadens our understanding of the dynamics between receptors and metabolites but also lays the foundation for integrating predictive modeling into food technology. The results offer promising avenues for future experimental validation and further model development with the ultimate goal of creating healthier and more sustainable food products.

The remainder of this article is structured as follows. Section 2 details the materials and methods used in this work, including the set of molecular descriptors that reflect various physicochemical, topological, and electronic properties of the target molecules, as well as the factors for model construction and optimization. Section 3 describes the metrics for all optimized models and their interpretations. Section 4 presents the results and their implications for the food matrix and microbiome-related processes. Finally, Section 5 provides the conclusions and outlines directions for future research.

2. Materials and Methods

To develop a machine learning model capable of predicting the potential activity of gut microbiota metabolites against estrogen receptor alpha ($ER\alpha$), a specialized dataset was constructed. This dataset integrated the structural information of molecules and data on their biological activity, which was crucial for training an accurate and informative model. The creation of this dataset involved several sequential steps aimed at ensuring data quality and integrity for subsequent analysis.

Initially, data on the biological activity of compounds interacting with $ER\alpha$ (target identifier: ChEMBL206) were collected from the authoritative ChEMBL database [11] via programmatic access. Queries to the database were configured to retrieve records where the standard activity type was IC50, i.e., the concentration of a substance at which 50% of maximal receptor inhibition is achieved. To ensure the statistical significance of the model and sufficient data volume, the maximum number of records available for analysis was selected (up to 1000).

IC50 was chosen as the activity metric because of its wide recognition and use in pharmacological research. IC50 enables quantitative comparisons of the efficacy of various inhibitors and the assessment of their potential therapeutic value [15]. This standard-

ized parameter facilitates data comparability across different studies and simplifies the interpretation of results.

After collecting the biological activity data, preliminary data processing was carried out. At this stage, all records without an activity value (*standard_value*) were excluded from the dataset to avoid possible distortions in the results and to ensure data integrity. Furthermore, to simplify the dataset structure and focus on key variables, only the molecule identifiers (*molecule_chembl_id*) and their corresponding activity values were retained.

A binary variable, *Activity*, was introduced to transform the quantitative activity values into categorical labels. Compounds with $IC_{50} \leq 1000$ nM were labeled as active (*Activity* = 1), whereas those with $IC_{50} > 1000$ nM were labeled as inactive (*Activity* = 0). The 1000 nM threshold was chosen based on widely accepted pharmacological criteria, according to which compounds with $IC_{50} \leq 1$ μ M are considered sufficiently active to be potential drug candidates [15]. This threshold provides a balance between the model's sensitivity and specificity, allowing an effective distinction between active and inactive compounds.

The next important step involved retrieving the molecular structures for each unique molecule identifier. Using the ChEMBL database API, the structure of each molecule was obtained in the SMILES (Simplified Molecular Input Line Entry System) format [16]. SMILES provides a text-based representation of a molecule's structure and is widely used for data storage and exchange in computational chemistry and bioinformatics. When a SMILES string was missing for a particular molecule, that record was excluded from further analysis. This ensured that all molecules in the dataset had complete structural descriptions necessary for calculating molecular descriptors.

By combining the biological data with the retrieved structural information, a complete dataset for analysis was formed. The integration was performed based on the molecule identifier (*molecule_chembl_id*), ensuring the correct correspondence between each molecule's activity and structure. Records without structural information and duplicates were removed, leaving only the key variables—SMILES and *Activity*—in the final dataset. This approach improved the data quality, reduced the risk of errors in training and prediction, and ensured the statistical independence of the observations.

The resulting dataset was ready for use, containing text-based representations of compound structures (SMILES) and corresponding binary activity labels (*Activity*). This dataset served as the foundation for computing molecular descriptors, which were subsequently used as input features to train the machine learning model.

For reproducibility and convenience in further analyses, the compiled dataset was saved in CSV format. This enables other researchers to replicate the results, conduct additional analyses, or use the data for model validation. Storing data in a standardized format also simplifies integration with various data analysis tools and platforms.

The data collection and preparation process provided a reliable and high-quality foundation for training a machine learning model that can predict the activity of new metabolites with respect to ER α based on their structural characteristics. The use of authoritative data sources such as ChEMBL [11], combined with standardized preprocessing steps, ensured the model's high quality and credibility. Transforming quantitative activity values into binary labels based on a well-founded threshold aligned with practical pharmacological standards [15] and gave the classification task a clear definition. Focusing on structural information and guaranteeing the reproducibility of the study enables other researchers to repeat the experiment and verify the results obtained, thereby advancing the field and supporting the model's implementation in practical applications.

To quantitatively describe the molecular structures of metabolites and facilitate subsequent analysis, an extensive set of molecular descriptors was employed, reflecting various physicochemical, topological, and electronic properties of the molecules. The RDKit library

was used to convert the metabolites' SMILES strings into molecular objects and compute the corresponding descriptors [17]. This tool is widely used in computational chemistry and offers extensive functionality for analyzing molecular structures.

The calculated molecular descriptors were divided into several categories, each providing unique information about the molecules' properties [25]. This approach ensures a comprehensive molecular characterization, which is critically important for building an accurate and robust machine learning model. These categories include physicochemical properties, structural properties, surface and volume characteristics, topological indices, and counts of specific atoms and functional groups.

1. Physicochemical Properties

This category includes descriptors characterizing the basic physicochemical parameters that affect the behavior of molecules in biological systems:

- MolWt (molecular weight): the total mass of the molecule, influencing pharmacokinetic parameters such as distribution and excretion;
- ExactMolWt (exact molecular weight): a more precise mass value taking into account the isotopic composition of atoms, important for mass spectrometric analysis;
- MolLogP: the logarithm of the octanol/water partition coefficient, reflecting the lipophilicity of the molecule and its ability to penetrate biological membranes;
- MolMR (molar refractivity): a parameter related to the polarizability and volume of the molecule, influencing interactions with electromagnetic radiation;
- TPSA (topological polar surface area): the total area of polar atoms and functional groups, correlating with cell membrane permeability and bioavailability.

2. Structural Properties

These descriptors characterize structural features of the molecules:

- HeavyAtomCount: the number of atoms excluding hydrogen, reflecting the molecule's size and complexity;
- NumValenceElectrons: the total number of valence electrons across all atoms, influencing the chemical reactivity of the molecule;
- NumHAcceptors: the number of atoms (commonly oxygen and nitrogen) capable of accepting hydrogen bonds;
- NumHDonors: the number of groups (usually –OH and –NH) capable of donating hydrogen bonds;
- NumRotatableBonds: the number of single bonds not in rings, defining the molecule's flexibility;
- RingCount: the total number of cyclic structures in the molecule, influencing its rigidity and spatial configuration;
- FractionCSP3: the proportion of sp³-hybridized carbon atoms, indicating the degree of saturation and three-dimensional character of the molecule;
- AromaticProportion: the ratio of aromatic atoms to the total number of atoms, important for π – π interactions.

3. Surface and Volume Properties

These descriptors evaluate the molecular surface and volume:

- LabuteASA (Labute's approximate surface area): an estimate of the molecule's surface area accessible for solvent interaction, affecting solubility;
- PEOE_VSA1–PEOE_VSA10: descriptors of surface area weighted by partial atomic charges, reflecting the distribution of electron density on the molecular surface;
- SlogP_VSA1–SlogP_VSA10: descriptors weighted by the log of the partition coefficient (LogP), characterizing the molecule's lipophilic and hydrophilic regions;

- SMR_VSA1–SMR_VSA10: descriptors weighted by molar refractivity, reflecting the molecule's volumetric and polarizable regions;
 - EState_VSA1–EState_VSA10: descriptors weighted by electron-topological states of atoms, accounting for electronic effects and molecular topology.
4. Topological Indices
- Topological indices characterize the molecular structure in terms of shape, size, and complexity:
- Chi0 and Chi1: connectivity indices of the 0th and 1st orders, indicating the degree of branching;
 - Chi0n–Chi4n: normalized connectivity indices that account for variations in atomic properties and connectivity at different levels;
 - Chi0v–Chi4v: connectivity indices weighted by the valence of atoms, reflecting the molecule's electronic structure;
 - Kappa1, Kappa2, and Kappa3 (Kier shape indices): parameters describing the shape and flexibility of a molecule based on its graph structure;
 - BalabanJ (Balaban index): a topological index incorporating information on both cyclicity and branching [26];
 - BertzCT (Bertz complexity index): a quantitative measure of molecular structural complexity;
 - Ipc (information content index): an information-based metric derived from molecular symmetry and bond diversity;
 - HallKierAlpha: a parameter related to the correction of atomic properties affecting hydrophobicity and molecular flexibility.
5. Counts of Specific Atoms and Groups
- This category includes descriptors that account for the presence of certain atoms and functional groups:
- NHOHCount: the total number of hydroxyl (–OH) and amino (–NH) groups, influencing the molecule's hydrophilicity and reactivity;
 - NOCount: the total number of nitrogen and oxygen atoms, reflecting the potential for hydrogen bonding.
6. Target Variable
- Activity: A binary variable indicating the known biological activity of a compound with respect to estrogen receptor alpha (ER α). It serves as the target variable for training the machine learning model.

The chosen set of descriptors provides a comprehensive description of metabolite molecules from multiple perspectives—physicochemical, structural, electronic, and geometric [25]. Physicochemical descriptors, such as MolWt and MolLogP, help elucidate how molecules behave in biological systems, including absorption and distribution processes. Topological indices and connectivity indices (Chi, Kappa, and BalabanJ) enable the assessment of structural features that may influence receptor interactions.

Surface descriptors (PEOE_VSA, SlogP_VSA, SMR_VSA, and EState_VSA) account for the distribution of molecular properties across the surface, which is crucial for understanding binding mechanisms at the molecular level. Counting specific atoms and groups (NumHAcceptors, NumHDonors, NHOHCount, and NOCount) provides information on the potential for forming hydrogen bonds and other targeted interactions.

Using such a diverse set of descriptors allows the machine learning model to more accurately predict the potential activity of metabolites with respect to ER α . This is achieved by capturing complex relationships between a molecule's structure and its biological

function, thereby improving both the accuracy and the reliability of the model when forecasting the activity of new compounds [25].

To develop a machine learning model capable of predicting the activity of metabolites with respect to estrogen receptor alpha (ER α), a final dataset was compiled containing both calculated molecular descriptors and the activity labels of the compounds. The data were separated into the following components:

- Features (X): a set of molecular descriptors computed for each compound;
- Labels (y): a binary variable (*Activity*) indicating whether a compound is active (1) or inactive (0).

Separating features and labels is a standard procedure in machine learning, enabling algorithms to learn from input data (features) to predict target values (labels) [27].

To evaluate the model's ability to generalize to previously unseen data, the dataset was divided into training and test sets in an 80:20 ratio:

- Training set (X_{train}, y_{train}): used for training and simultaneously optimizing all models;
- Test set (X_{test}, y_{test}): used for the final assessment of the already-optimized models on "unseen" data.

This split enables an objective evaluation of algorithm performance and helps prevent overfitting, providing a more accurate assessment of the model's ability to predict the activity of new compounds [28].

In developing an effective predictive system, a wide range of machine learning algorithms was considered, including models designed to handle class imbalance. Class imbalance arises when the number of active and inactive compounds differs substantially, causing the model to potentially become biased toward the majority class [32].

The list of models evaluated included the following:

- Random forests: RandomForestClassifier (with and without class weighting), BalancedRandomForestClassifier;
- Ensemble methods: AdaBoostClassifier, GradientBoostingClassifier, BaggingClassifier, EasyEnsembleClassifier;
- Linear models: LogisticRegression (with and without class weighting), RidgeClassifier;
- Support vector methods: SVC, LinearSVC;
- Decision trees: DecisionTreeClassifier (with and without class weighting), ExtraTreeClassifier;
- Nearest neighbors: KNeighborsClassifier;
- Naive Bayes classifiers: GaussianNB, BernoulliNB;
- Discriminant analysis: LinearDiscriminantAnalysis, QuadraticDiscriminantAnalysis;
- Neural networks: MLPClassifier;
- Other models: SGDClassifier, GaussianProcessClassifier, VotingClassifier, CalibratedClassifierCV;
- Boosting methods: XGBoostClassifier, CatBoostClassifier (with and without class weighting).

Including such a diverse set of algorithms provides insight into which models perform best on the given classification task [33], with special attention paid to those that address the unequal ratio of active to inactive compounds [32].

All the listed algorithms were trained and optimized within a single process comprising the following main steps:

1. Hyperparameter Tuning

For each algorithm, a set of relevant parameters was specified (e.g., the number of trees for RandomForestClassifier, regularization coefficients for LogisticRegression, or LinearDiscriminantAnalysis, etc.);

2. k-Fold Cross-Validation

Cross-validation (commonly 5- or 10-fold) was conducted on the training set (X_{train} , y_{train}), providing performance estimates for each hyperparameter combination;

3. Search Methods

Depending on the specific model, GridSearchCV (exhaustive hyperparameter search) or RandomizedSearchCV (random search within specified ranges) was employed [31];

4. Class Balance Adjustment

For algorithms supporting the `class_weight` parameter, suitable settings (e.g., ‘balanced’) were chosen to improve sensitivity to the minority class;

5. Selecting Optimal Configurations

For each model, the hyperparameters yielding the best metric—Accuracy in this study—were identified via cross-validation.

After hyperparameter tuning, each model with its optimal configuration was retrained on the full training set (X_{train} , y_{train}). The performance of these retrained models was then evaluated on the test set (X_{test} , y_{test}), where the final metrics (e.g., Accuracy, Precision, Recall, and F1 Score) were calculated.

3. Results

Table 1 presents the final metrics for all the optimized models in predicting activity on the test set. Alongside Accuracy, metrics such as Precision, Recall, F1 Score, MCC, Cohen’s Kappa, Hamming Loss, and ROC AUC were also included, offering a comprehensive assessment of each algorithm’s strengths and weaknesses.

Table 1. Model Evaluation Results.

Model	Accuracy	Precision	Recall	F1 Score	MCC	Cohen’s Kappa	Hamming Loss	ROC AUC
<i>LinearDiscriminantAnalysis</i>	0.8944	0.9000	0.9612	0.9296	0.7255	0.7192	0.1056	0.8396
<i>RidgeClassifier</i>	0.8873	0.8991	0.9515	0.9245	0.7074	0.7030	0.1127	0.8347
<i>LogisticRegression</i>	0.8662	0.8500	0.9903	0.9148	0.6522	0.6116	0.1338	0.7644
<i>Bagging</i>	0.8451	0.8716	0.9223	0.8962	0.5953	0.5916	0.1549	0.7817
<i>Bagging (DecisionTree)</i>	0.8451	0.8716	0.9223	0.8962	0.5953	0.5916	0.1549	0.7817
<i>BalancedRandomForest</i>	0.8239	0.9535	0.7961	0.8677	0.6334	0.6108	0.1761	0.8468
<i>CalibratedClassifierCV (Logistic)</i>	0.8239	0.8250	0.9612	0.8879	0.5214	0.4889	0.1761	0.7114
<i>AdaBoost</i>	0.8169	0.8235	0.9515	0.8829	0.5003	0.4733	0.1831	0.7065
<i>GradientBoosting</i>	0.8169	0.8468	0.9126	0.8785	0.5151	0.5092	0.1831	0.7384
<i>CatBoost</i>	0.8169	0.8532	0.9029	0.8774	0.5206	0.5174	0.1831	0.7463
<i>XGBoost</i>	0.8099	0.8455	0.9029	0.8732	0.4989	0.4946	0.1901	0.7335
<i>RandomForest</i>	0.8028	0.8319	0.9126	0.8704	0.4710	0.4623	0.1972	0.7127
<i>QuadraticDiscriminantAnalysis</i>	0.8028	0.8866	0.8350	0.8600	0.5304	0.5277	0.1972	0.7765
<i>EasyEnsemble</i>	0.7817	0.9286	0.7573	0.8342	0.5479	0.5241	0.2183	0.8017
<i>Voting</i>	0.7817	0.7951	0.9417	0.8622	0.3859	0.3544	0.2183	0.6504
<i>ExtraTree</i>	0.7676	0.8070	0.8932	0.8479	0.3692	0.3607	0.2324	0.6646
<i>KNeighbors</i>	0.7606	0.8000	0.8932	0.8440	0.3452	0.3355	0.2394	0.6517
<i>Bagging (KNeighbors)</i>	0.7465	0.7680	0.9320	0.8421	0.2591	0.2285	0.2535	0.5942
<i>DecisionTree</i>	0.7324	0.7982	0.8447	0.8208	0.2965	0.2946	0.2676	0.6403
<i>SGDClassifier</i>	0.7254	0.7254	1.0000	0.8408	0.0000	0.0000	0.2746	0.5000
<i>LinearSVC</i>	0.7254	0.7254	1.0000	0.8408	0.0000	0.0000	0.2746	0.5000
<i>SVC</i>	0.7254	0.7254	1.0000	0.8408	0.0000	0.0000	0.2746	0.5000
<i>MLPClassifier</i>	0.7254	0.7254	1.0000	0.8408	0.0000	0.0000	0.2746	0.5000
<i>CalibratedClassifierCV (SVC)</i>	0.7254	0.7254	1.0000	0.8408	0.0000	0.0000	0.2746	0.5000
<i>BernoulliNB</i>	0.7183	0.7603	0.8932	0.8214	0.1881	0.1747	0.2817	0.5748
<i>GaussianNB</i>	0.6056	1.0000	0.4563	0.6267	0.4328	0.3155	0.3944	0.7282
<i>GaussianProcess</i>	0.3732	0.8500	0.1650	0.2764	0.1131	0.0530	0.6268	0.5441

Metric Interpretation:

- **Accuracy:** The proportion of correct predictions among all predictions. A value of 89.44% indicates high overall model accuracy;
- **Precision:** The proportion of true positives among all predicted positives. A high value (90.00%) signifies a low rate of false positives;

- Recall: The proportion of true positives among all actual positive instances. A value of 96.12% indicates that the model effectively detects active compounds;
- F1 score: The harmonic mean of Precision and Recall. A value of 92.96% reflects a good balance between them;
- MCC (Matthews correlation coefficient): The correlation between observed and predicted classes, accounting for true/false positives and negatives. A value of 0.73 points to a strong positive correlation;
- Cohen's kappa: A measure of agreement between the model's predictions and the actual labels, adjusted for chance. A value of 0.72 indicates good agreement;
- Hamming loss: The proportion of incorrect labels relative to the total number of labels. A low value of 0.11 indicates few classification errors;
- ROC AUC score: The area under the receiver operating characteristic curve, shows the model's ability to discriminate between classes. A value of 83.96% demonstrates a good discriminatory capacity.

Results Analysis:

- LinearDiscriminantAnalysis (LDA) achieved the highest accuracy (≈ 0.8944). It also demonstrated high Recall (≈ 0.9612) and balanced Precision (≈ 0.9000), leading to an F1 Score of ≈ 0.9296 . LDA further exhibited comparatively high MCC (≈ 0.7255) and Cohen's Kappa (≈ 0.7192);
- RidgeClassifier (Accuracy ≈ 0.8873) slightly lags behind LDA but also provides a high Recall (≈ 0.9515);
- LogisticRegression offers a high Recall (≈ 0.9903), though its overall Accuracy (≈ 0.8662) is lower than that of the top models;
- Ensemble methods (Bagging, BalancedRandomForest, GradientBoosting, etc.) generally yield Accuracy values in the 0.80–0.85 range, illustrating the benefits of combining multiple base models;
- Models with adjusted class weights (BalancedRandomForest, CatBoost with class weighting) indeed handle the minority class more effectively, as demonstrated by their Precision/Recall values [32];
- Certain algorithms (SGDClassifier, SVC, LinearSVC, MLPClassifier, etc.) converged on a "hard" boundary (Recall = 1.0 and Precision ≈ 0.725), indicating peculiarities in how they optimize given this dataset;
- GaussianProcessClassifier (Accuracy ≈ 0.3732) displayed the lowest results, likely due to the specific characteristics of this algorithm and the complexity of the data structure.

Comparing all the optimized algorithms by Accuracy (and additional metrics, including Precision, Recall, and F1 Score) reveals that LinearDiscriminantAnalysis delivers the best performance. The primary arguments in favor of choosing LDA include the following:

1. High Accuracy (≈ 0.8944)—the top result among all the models tested;
2. Balanced Metrics (Precision ≈ 0.900 , Recall ≈ 0.961)—LDA effectively predicts both active and inactive compounds;
3. High MCC (0.7255) and Cohen's kappa (0.7192)—indicate strong agreement between predictions and actual labels, along with a low probability of random alignment.

These findings are visually summarized in Figure 1, which illustrates the performance metrics across all evaluated models.

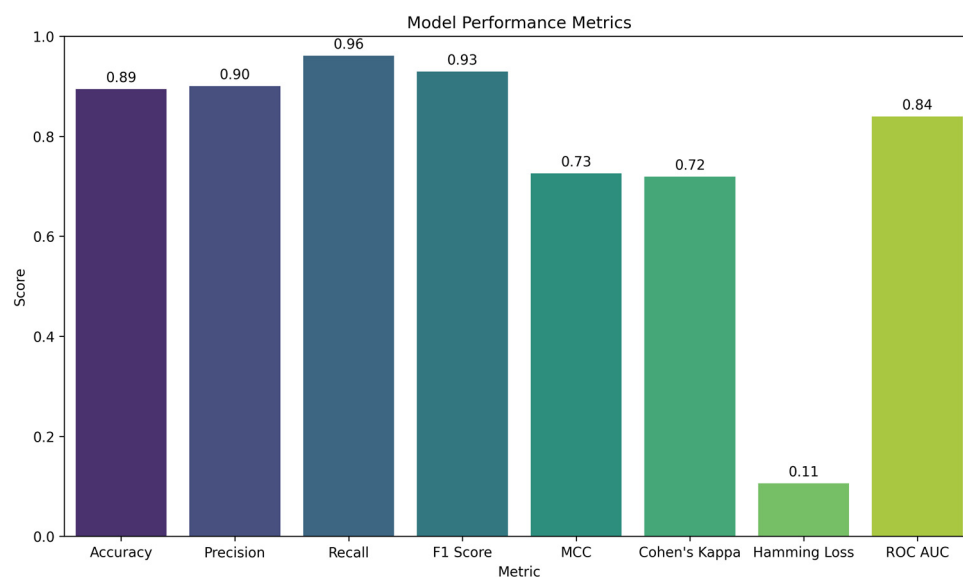


Figure 1. Model Performance Metrics.

Thus, Linear Discriminant Analysis (LDA) was identified as the most effective algorithm for predicting the activity of gut microbiota metabolites with respect to estrogen receptor alpha. Our conclusions regarding model construction and optimization are as follows:

- By simultaneously tuning hyperparameters across all the models using cross-validation, we ensured an objective and comprehensive comparison of various machine learning algorithms;
- LDA achieved the highest accuracy and exhibited balanced performance across additional metrics, confirming its suitability for this application;
- Models incorporating class weighting demonstrated enhanced sensitivity to the minority class—a critical consideration when dealing with small but biologically significant groups of compounds [32];
- The optimized LDA model is not only effective for screening novel metabolites but also holds promise as a tool for the early stages of drug development, where rapid and reliable evaluation of compound activity is essential.

Supporting literature has consistently highlighted the efficacy of LDA in similar predictive tasks, underscoring its robustness and interpretability [30,31].

After optimizing the linear discriminant analysis (LDA) model and evaluating its performance, a feature importance analysis was conducted to identify which molecular descriptors most significantly influence the prediction of a compound's activity with respect to estrogen receptor alpha (ER α).

LDA model coefficients reflect each feature's contribution to the linear discriminant function used to distinguish between classes. To evaluate feature importance, the absolute values of these coefficients were used, given that the sign indicates the direction of influence, whereas the magnitude indicates its extent [30].

These results are visually summarized in Figure 2.

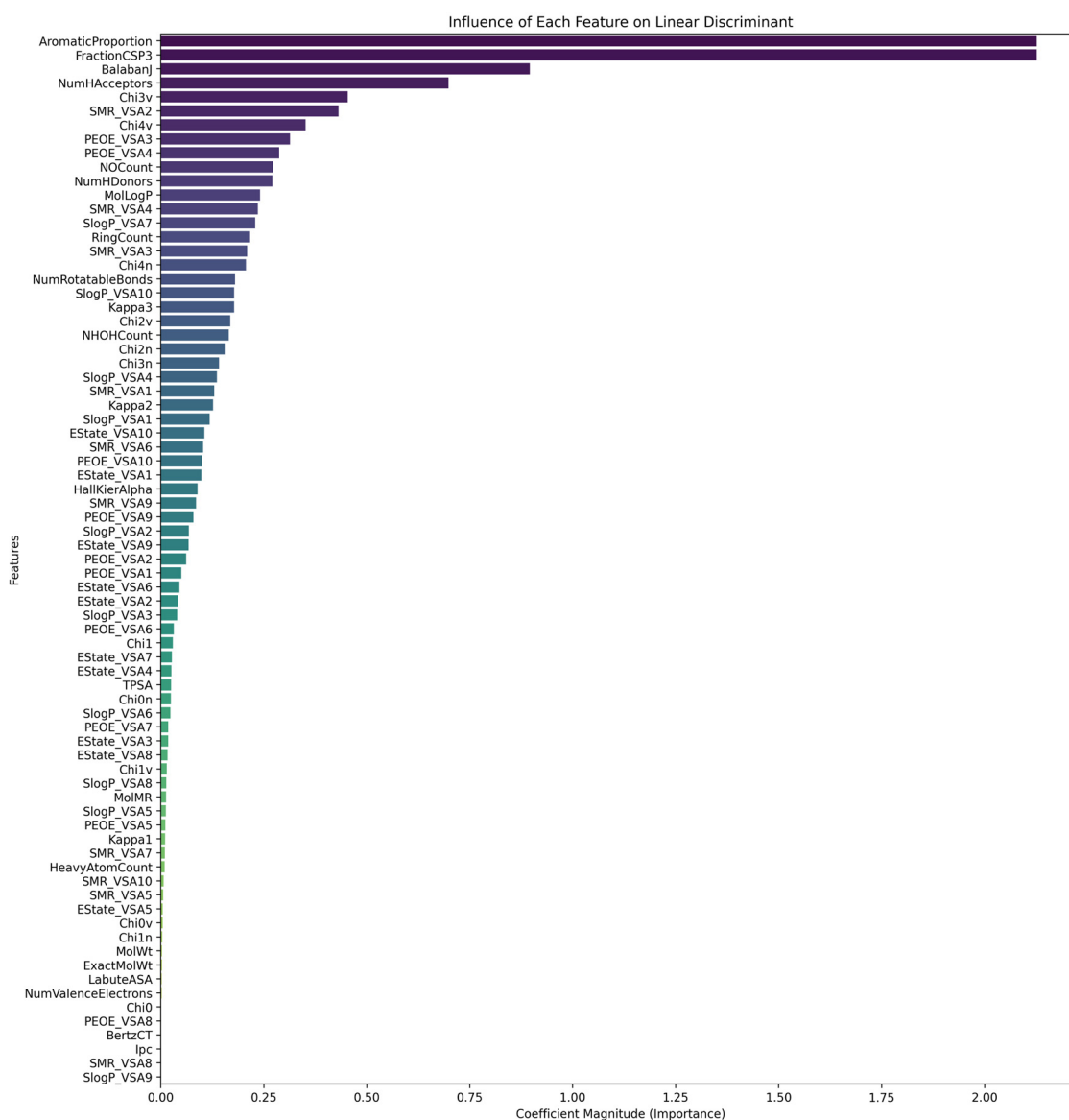


Figure 2. Influence of Each Feature on Linear Discriminant.

Table 2 presents ten molecular descriptors with the largest absolute coefficients in the LDA model:

Table 2. Top 10 Most Significant Features.

No.	Feature	Importance
1	FractionCSP3	2.126618
2	AromaticProportion	2.126618
3	BalabanJ	0.895716
4	NumHAcceptors	0.698623
5	Chi3v	0.453634
6	SMR_VSA2	0.431287
7	Chi4v	0.351189
8	PEOE_VSA3	0.313827
9	PEOE_VSA4	0.287226
10	NOCount	0.272061

Interpretation of Significant Features:

1. FractionCSP3 (Proportion of sp^3 -Hybridized Carbon Atoms)

Quantifies the fraction of sp^3 -hybridized carbon atoms relative to the total number of carbon atoms [35]. The high importance of this descriptor suggests that molecules with a larger proportion of sp^3 -hybridized carbons may exhibit greater ER α activity. This could be related to the impact of saturation on conformational flexibility and the three-dimensional molecular structure, which in turn affects receptor binding [36].

2. AromaticProportion (Proportion of Aromatic Atoms)

Represents the ratio of aromatic atoms to the total number of atoms in a molecule. Its importance highlights the role of aromatic systems in interacting with ER α . Aromatic rings are known to participate in π - π and hydrophobic interactions with receptor binding sites, contributing to the stability of the ligand–receptor complex [37].

3. BalabanJ (Balaban Index)

A topological index that integrates information about molecular cyclicity and branching [26]. High BalabanJ values indicate a more complex molecular topology, which can facilitate specific receptor interactions through unique spatial configurations.

4. NumHAcceptors (Number of Hydrogen Bond Acceptors)

The number of atoms (commonly oxygen and nitrogen) capable of accepting hydrogen bonds. The capacity to form hydrogen bonds with the amino acid residues of ER α can be critical for receptor binding and activation [38].

5. Chi3v and Chi4v (Valence-Weighted Connectivity Indices)

These indices reflect the complexity of the molecular structure while accounting for the valence of atoms and their connectivity at higher orders [39]. Their significance suggests that structural features—such as branching and cyclicity at the tertiary and quaternary levels—affect biological activity against ER α .

6. SMR_VSA2 (Surface Area Descriptor Weighted by Molar Refractivity)

Characterizes the surface area of the molecule associated with a specific range of molar refractivity values [40]. This descriptor indicates that polarizability and volumetric properties play a role in the molecule's interaction with the receptor.

7. PEOE_VSA3 and PEOE_VSA4 (Surface Area Descriptors Weighted by Partial Charges)

Represent the distribution of molecular surface area according to partial atomic charges calculated using the Partial Equalization of Orbital Electronegativities (PEOE) method [41]. The importance of these descriptors implies that electrostatic interactions significantly contribute to ligand binding with ER α .

8. NOCount (Total Number of Nitrogen and Oxygen Atoms)

The number of nitrogen and oxygen atoms in the molecule. These atoms can form hydrogen bonds and affect molecular polarity—factors that are crucial for interacting with hydrophilic receptor regions [39].

The feature importance analysis in the LDA model revealed that both topological and physicochemical descriptors substantially influence the prediction of a compound's ER α activity. Key conclusions include the following:

- Degree of saturation and aromaticity: the descriptors *FractionCSP3* and *AromaticProportion* indicate that both saturated and aromatic structures are important for ER α interactions;

- Topological complexity: the indices *BalabanJ*, *Chi3v*, and *Chi4v* underscore the importance of molecular complexity and unique structural configurations in receptor binding;
- Hydrogen-bonding capacity: *NumHAcceptors* and *NOCount* emphasize the role of hydrogen bonding in stabilizing the ligand–receptor complex;
- Electronic and surface properties: the *PEOE_VSA* and *SMR_VSA* descriptors highlight the effects of electron density distribution and surface polarizability on molecular interactions with ER α .

Identifying which molecular descriptors most significantly influence compound activity allows for the following:

1. Guidance in synthesizing new compounds—by focusing on structural features that enhance the likelihood of ER α binding, one can optimize the design of new ligands;
2. Improvement of predictive models—considering influential features during model training and potentially removing less informative descriptors can enhance predictive performance;
3. Advancement of mechanistic understanding—gaining deeper insight into the structural elements that promote biological activity supports the rational design of more effective therapeutic agents.

By identifying the key molecular descriptors that govern ER α activity, the LDA feature importance analysis underscores the need for a comprehensive approach—incorporating both topological and physicochemical properties—to successfully predict biological activity. Future studies can focus on a more detailed examination of these descriptors and on experimental validation of the predicted active compounds. Such efforts would not only improve predictive models but also contribute to the development of novel therapeutic agents targeting ER α .

To evaluate the ability of the developed machine learning model to predict the activity of new compounds with respect to ER α , specific gut microbiota metabolites were selected. These metabolites are of interest in the field of food technology and are present in various food products. The selection criteria included the following:

1. Presence in Food Products:

Metabolites are widely distributed in food or are produced during food processing and storage. Their presence can affect the quality, safety, and functional properties of food [8];

2. Influence on Organoleptic Properties:

Many of these compounds are responsible for the taste, aroma, and texture of products—crucial for consumer perception and market competitiveness [9];

3. Potential Biological Activity:

These metabolites can affect physiological processes upon consumption, which is important for developing functional foods and evaluating possible health risks [11].

For each selected metabolite, molecular descriptors analogous to those used during model training were calculated. The optimized LinearDiscriminantAnalysis model then predicted whether the metabolite is potentially active with respect to ER α . The predicted activities for the selected metabolites are summarized in Table 3. For additional details on the selected metabolites and their corresponding SMILES representations, please refer to Appendix A. The complete list is provided in Table A1.

These results demonstrate the potential applicability of the optimized LDA model for identifying compounds that may exhibit biological activity toward ER α . The predictions could serve as a foundation for further experimental validation and the design of novel compounds with desired properties.

Table 3. Predicted Activity of Metabolites.

Metabolite	Predicted Activity
Indole	Active
Skatole (3-methylindole)	Active
Trimethylamine (TMA)	Active
Indole-3-acetic acid (IAA)	Active
Indole-3-lactate	Inactive
4-Cresol (p-cresol)	Active
p-Cresol sulfate	Inactive
Phenylacetic acid	Inactive
Phenyllactic acid	Active
o-Cresol	Active
3-Methylphenol (m-cresol)	Active
Deoxycholic acid	Inactive
Lithocholic acid	Inactive
Urolithin A	Active
Urolithin B	Active
Acetoin	Inactive
2,3-Butanediol	Inactive
Cadaverine	Inactive
Putrescine	Inactive
Trimethylamine N-oxide (TMAO)	Active
Phenol	Active

4. Discussion

Predicting the activity of gut microbiota metabolites toward estrogen receptor alpha (ER α) is critical for food technology because ER α plays a key role in modulating metabolic and hormonal pathways that can influence both the nutritional quality and safety of food products. By understanding which active metabolites are present and controlling their formation during processes such as fermentation, thermal treatment, and storage, food technologists can optimize flavor profiles, enhance functional properties, and mitigate potential adverse health effects. This integrative approach not only supports the development of functional foods but also contributes to consumer safety by ensuring that food products meet high-quality standards [42].

Presence of Active Metabolites in Food Products:

1. Indole Compounds

- Indole and skatole (3-methylindole)

These compounds frequently occur in fermented products such as soy sauce, kimchi, and certain cheeses. They are also present in meat products, particularly those that have undergone prolonged storage or inadequate processing [43]. Indole and skatole contribute to the aroma and flavor of foods, imparting distinctive notes. Their activity toward ER α may be significant for the development of functional foods that support hormonal balance [44];

- Indole-3-acetic acid (IAA)

IAA is a plant hormone that may be found in fresh vegetables and fruits. During fermentation, gut microbiota can convert the amino acid tryptophan into IAA [45]. IAA can affect plant product ripening and storage processes, as well as microbiological processes during fermentation.

2. Phenolic Compounds

- 4-Cresol, o-cresol, 3-methylphenol (m-cresol), and phenol

These phenolic compounds are present in smoked foods, tea, coffee, and certain spices. They may form during thermal processing, such as smoking or roasting. Phenols contribute to food flavor and aroma while also exhibiting antioxidant properties. Their activity toward ER α makes them potentially important components of functional food products [46].

3. Urolithins

- Urolithin A and urolithin B

These are formed in the human gut from ellagitannins found in pomegranates, nuts, berries, and some other fruits. Although urolithins themselves are not directly found in foods, their precursors are widespread [47]. Urolithins possess antioxidant and anti-inflammatory properties, and their formation depends on gut microbiota composition, emphasizing the need to consume ellagitannin-rich foods [48].

4. Trimethylamine (TMA) and Trimethylamine N-oxide (TMAO)

- TMA and TMAO are found in seafood, especially fish and shellfish. TMAO acts as a natural osmolyte in marine organisms. TMA is responsible for the characteristic fishy odor, which is intensified in spoiled products due to the breakdown of TMAO into TMA. Their activity toward ER α may be relevant for food technologies involving seafood [49].

Controlling the formation of active metabolites is a key aspect of the food industry. First, processing methods and storage conditions play a major role in regulating the concentration of these compounds. Adjusting thermal processing parameters—such as temperature, duration, and humidity—allows for controlling the formation of phenolic and indole metabolites during smoking, frying, and other processes [50]. Moreover, some metabolites form under anaerobic conditions; thus, managing oxygen availability during fermentation and storage can influence their concentrations [51]. The formation of trimethylamine (TMA) is primarily driven by microbial activity, particularly the enzymatic reduction of trimethylamine N-oxide (TMAO) by spoilage bacteria. To control TMA formation, it is essential to limit the growth of these bacteria. This can be achieved by using antimicrobial agents and applying pasteurization or sterilization techniques, which significantly reduce the microbial load. Additionally, maintaining optimal storage conditions—such as low temperatures and controlled pH levels—can further inhibit the enzymatic conversion of TMAO to TMA, thereby enhancing food safety and quality [52].

Second, the use of starter cultures and probiotics is an effective method for influencing microbiota and metabolic processes. Employing specific starter cultures can promote the formation of desirable active metabolites (e.g., urolithins) while reducing unwanted compounds [53]. Introducing probiotics into products can modulate consumers' gut microbiota, fostering the generation of beneficial metabolites from dietary precursors [54].

Third, fortifying foods with precursors that promote the formation of beneficial metabolites represents another viable strategy. For instance, incorporating ingredients rich in polyphenols and ellagitannins—key precursors for the formation of urolithins (notably urolithin A)—can enhance their subsequent production in the human body, thereby increasing the functional value of food [55]. Additionally, including dietary fiber, which serves as a substrate for gut microbiota, may further modulate the metabolic pathways leading to the synthesis of these active compounds [56].

Finally, controlling pH and redox potential (Eh) can significantly influence microbial processes. Many microbial activities depend on acidity; thus, regulating pH can affect the enzymes and microorganisms responsible for metabolite formation. Managing redox

potential during processing and storage can either promote or inhibit the production of specific metabolites [57].

Practical Recommendations for the Food Industry:

1. Development of technical guidelines that account for processing parameters influencing the formation of active metabolites;
2. Use of predictive models, incorporating data on predicted metabolite activity in food quality and safety control systems;
3. Professional training of food technologists and quality specialists in microbiological process management and the impact of metabolites on food products.

5. Conclusions

This study demonstrates the potential of machine learning approaches for predicting the biological activity of food metabolites and gut microbiota-derived metabolites with respect to estrogen receptor alpha ($ER\alpha$). By employing an optimized linear discriminant analysis (LDA) model based on a curated dataset of molecular descriptors, we achieved a predictive accuracy of 89.44%. This high level of performance confirms the viability of our computational approach for screening compounds that may interact with $ER\alpha$.

A key finding of our research is the identification of critical molecular descriptors, such as FractionCSP3 and AromaticProportion, which significantly influence the interaction between metabolites and $ER\alpha$. These results not only validate the application of LDA in this context but also provide valuable insights into the molecular features underlying receptor binding. The clarity provided by feature importance analysis further enhances the interpretability of the model, offering a robust foundation for subsequent experimental investigations.

From a practical perspective, the successful prediction of active metabolites—such as indole, skatole, and various phenolic compounds—has important implications for food technology and healthy nutrition. It is well established that these compounds influence the organoleptic properties and overall quality of food products. By understanding and controlling the formation of these active metabolites through the adjustment of processing parameters (e.g., fermentation, thermal treatment, and storage conditions), food technologists can devise strategies to improve taste, aroma, and nutritional value while ensuring product safety.

This study also highlights the inherent limitations of predictive modeling when applied to complex food systems. Although the model provides a reliable screening tool, it does not account for the multifaceted interactions within the food matrix that may alter metabolite activity. Therefore, our results should be interpreted as a preliminary step, emphasizing the need for further experimental validation and in-depth investigation into how real processing conditions affect metabolite behavior.

Looking forward, future research should focus on expanding the dataset with new experimental data and exploring additional machine learning methods to enhance predictive efficiency. Integrating the predictive model into quality monitoring and control systems could further support its practical application in food production. Such efforts will not only refine the model but also contribute to the development of innovative strategies for creating healthier and more sustainable food products.

In conclusion, our study confirms that machine learning—specifically LDA—is a valuable tool for predicting the biological activity of gut microbiota-derived metabolites in relation to $ER\alpha$. By combining computational predictions with practical applications in food technology, we open new avenues for enhancing food quality and consumer health. This work lays the groundwork for future studies that will further integrate predictive modeling with experimental validation, ultimately advancing sustainable and innovative food production methods.

Author Contributions: Conceptualization, M.K. (Maksim Kuznetsov) and I.N.; methodology, M.K. (Maksim Kuznetsov); software, M.K. (Mikhail Kutuzov); validation, M.K. (Maksim Kuznetsov), O.C. and D.V.; formal analysis, M.K. (Mikhail Kutuzov), O.N. and D.V.; investigation, A.S. and D.M.; resources, D.V.; data curation, D.M.; writing—original draft preparation, M.K. (Maksim Kuznetsov) and M.K. (Mikhail Kutuzov); writing—review and editing, D.V., O.N., A.S. and D.M.; visualization, O.C. and A.S.; supervision, M.K. (Maksim Kuznetsov) and I.N.; project administration, I.N.; and funding acquisition, I.N. All authors have read and agreed to the published version of the manuscript.

Funding: This study was funded by a subsidy for the implementation of a state assignment from the Ministry of Science and Higher Education of the Russian Federation within the framework of research project No. FSSW-2025-0004.

Data Availability Statement: The dataset analyzed in this study is publicly available at https://github.com/4kumax/lda_model_era (accessed on 30 March 2025).

Conflicts of Interest: Author Mikhail Kutuzov was employed by the company Apq Control LLC. The remaining 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.

Abbreviations

The following abbreviations are used in this manuscript:

- LDA Linear discriminant analysis
- SVC Support vector classification
- ERα Estrogen receptor alpha
- TMA Trimethylamine
- TMAO Trimethylamine N-oxide

Appendix A

Table A1. List of Selected Metabolites and Their SMILES.

No.	Metabolites	Smiles
1	Indole	C1=CC=C2C(=C1)C=CN2
2	Skatole (3-methylindole)	CC1=CNC2=CC=CC=C12
3	Trimethylamine (TMA)	CN(C)C
4	Indole-3-acetic acid (IAA)	C1=CC=C2C(=C1)C(=CN2)CC(=O)O
5	Indole-3-lactate	C1=CC=C2C(=C1)C(=CN2)C(CO)C(=O)O
6	p-Cresol (4-cresol)	CC1=CC=C(C=C1)O
7	p-Cresol sulfate	CC1=CC=C(C=C1)OS(=O)(=O)O
8	Phenylacetic acid	C1=CC=C(C=C1)CC(=O)O
9	Phenyllactic acid	C1=CC=C(C=C1)C(CO)C(=O)O
10	o-Cresol	CC1=CC=CC=C1O
11	3-Methylphenol (m-cresol)	CC1=CC(=CC=C1)O
12	Deoxycholic acid	C[C@H](CCC(=O)O)C1CCC2C(C1)CCC3C2CC(C4C3(CCC(C4)O)C)O
13	Lithocholic acid	C[C@H](CCC(=O)O)C1CCC2C(C1)CCC3C2CC=C4C3(CCC(C4)O)C
14	Urolithin A	C1=CC(=C2C(=C1)C(=O)C3=CC=CC=C3C2=O)O
15	Urolithin B	C1=CC(=C2C(=C1)C(=O)C3=CC=CC=C3C2=O)O
16	Acetoin	CC(C=O)O
17	2,3-Butanediol	CC(CO)CO
18	Cadaverine	C(CN)CCCNC
19	Putrescine	C(CCN)CN
20	Trimethylamine N-oxide (TMAO)	C[N+](C)(C)[O-]
21	Phenol	C1=CC=C(C=C1)O

References

1. Ma, H.; Gollahon, L.S. ERα Mediates Estrogen-Induced Expression of the Breast Cancer Metastasis Suppressor Gene BRMS1. *Int. J. Mol. Sci.* **2016**, *17*, 158. [CrossRef] [PubMed]

2. Yang, M.; Lee, J.H.; Zhang, Z.; De La Rosa, R.; Bi, M.; Tan, Y.; Liao, Y.; Hong, J.; Du, B.; Wu, Y.; et al. Enhancer RNAs Mediate Estrogen-Induced Decommissioning of Selective Enhancers by Recruiting ER α and Its Cofactor. *Cell Rep.* **2020**, *31*, 107803. [CrossRef] [PubMed]
3. Tang, Z.-R.; Zhang, R.; Lian, Z.-X.; Deng, S.-L.; Yu, K. Estrogen-Receptor Expression and Function in Female Reproductive Disease. *Cells* **2019**, *8*, 1123. [CrossRef]
4. Della Torre, S.; Mitro, N.; Fontana, R.; Gomaraschi, M.; Favari, E.; Recordati, C.; Lolli, F.; Quagliarini, F.; Meda, C.; Ohlsson, C.; et al. An Essential Role for Liver ER α in Coupling Hepatic Metabolism to the Reproductive Cycle. *Cell Rep.* **2016**, *15*, 360–371. [CrossRef]
5. Fu, Q.; Li, T.; Zhang, C.; Ma, X.; Meng, L.; Liu, L.; Shao, K.; Wu, G.; Zhu, X.; Zhao, X. Butyrate mitigates metabolic dysfunctions via the ER α -AMPK pathway in muscle in OVX mice with diet-induced obesity. *Cell Commun. Signal.* **2023**, *21*, 95. [CrossRef]
6. Guo, M.; Cao, X.; Ji, D.; Xiong, H.; Zhang, T.; Wu, Y.; Suo, L.; Pan, M.; Brugger, D.; Chen, Y.; et al. Gut Microbiota and Acylcarnitine Metabolites Connect the Beneficial Association between Estrogen and Lipid Metabolism Disorders in Ovariectomized Mice. *Microbiol. Spectr.* **2023**, *11*, e00149-23. [CrossRef]
7. Qi, X.; Zhao, Y.; Qi, Z.; Hou, S.; Chen, J. Machine Learning Empowering Drug Discovery: Applications, Opportunities and Challenges. *Molecules* **2024**, *29*, 903. [CrossRef]
8. Björnström, L.; Sjöberg, M. Mechanisms of estrogen receptor signaling: Convergence of genomic and nongenomic actions on target genes. *Mol. Endocrinol.* **2005**, *19*, 833–842. [CrossRef]
9. Thomas, C.; Gustafsson, J.Å. The different roles of ER subtypes in cancer biology and therapy. *Nat. Rev. Cancer* **2011**, *11*, 597–608. [CrossRef]
10. Knowlton, A.; Lee, A. Estrogen and the cardiovascular system. *Pharmacol. Ther.* **2012**, *135*, 54–70. [CrossRef]
11. Gaulton, A.; Bellis, L.J.; Bento, A.P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; et al. ChEMBL: A large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40*, D1100–D1107. [CrossRef] [PubMed]
12. Huggins, R.J.; Greene, G.L. ER α /PR crosstalk is altered in the context of the ER α Y537S mutation and contributes to endocrine therapy-resistant tumor proliferation. *NPJ Breast Cancer* **2023**, *9*, 96. [CrossRef] [PubMed]
13. Scabia, V.; Ayyanan, A.; De Martino, F.; Agnoletto, A.; Battista, L.; Laszlo, C.; Treboux, A.; Zaman, K.; Stravodimou, A.; Jallut, D.; et al. Estrogen receptor positive breast cancers have patient specific hormone sensitivities and rely on progesterone receptor. *Nat. Commun.* **2022**, *13*, 3127. [CrossRef] [PubMed]
14. Heath, H.; Mogol, A.N.; Casiano, A.S.; Zuo, Q.; Madak-Erdogan, Z. Targeting Systemic and Gut Microbial Metabolism in ER+ Breast Cancer. *Trends Endocrinol. Metab.* **2024**, *35*, 321–330. [CrossRef]
15. Sebaugh, J.L. Guidelines for accurate EC50/IC50 estimation. *Pharm. Stat.* **2011**, *10*, 128–134. [CrossRef]
16. Weininger, D. SMILES, a chemical language and information system. 1. Introduction to methodology and encoding rules. *J. Chem. Inf. Comput. Sci.* **1988**, *28*, 31–36. [CrossRef]
17. RDKit: Open-Source Cheminformatics. Available online: <https://www.rdkit.org> (accessed on 1 October 2023).
18. Ortega, M.A.; Alvarez-Mon, M.A.; García-Montero, C.; Fraile-Martinez, O.; Guijarro, L.G.; Lahera, G.; Monserrat, J.; Valls, P.; Mora, F.; Rodríguez-Jiménez, R.; et al. Gut Microbiota Metabolites in Major Depressive Disorder—Deep Insights into Their Pathophysiological Role and Potential Translational Applications. *Metabolites* **2022**, *12*, 50. [CrossRef]
19. Ma, Z.; Zuo, T.; Frey, N.; Rangrez, A.Y. A systematic framework for understanding the microbiome in human health and disease: From basic principles to clinical translation. *Sig. Transduct. Target Ther.* **2024**, *9*, 237. [CrossRef]
20. Aggarwal, N.; Kitano, S.; Puah, G.R.Y.; Kittelmann, S.; Hwang, I.Y.; Chang, M.W. Microbiome and Human Health: Current Understanding, Engineering, and Enabling Technologies. *Chem. Rev.* **2023**, *123*, 31–72. [CrossRef]
21. Hou, K.; Wu, Z.X.; Chen, X.Y.; Wang, J.-Q.; Zhang, D.; Xiao, C.; Zhu, D.; Koya, J.B.; Wei, L.; Li, J.; et al. Microbiota in health and diseases. *Sig. Transduct. Target Ther.* **2022**, *7*, 135. [CrossRef]
22. Marino, M.; Pellegrini, M.; La Rosa, P.; Acconcia, F. Susceptibility of Estrogen Receptor Rapid Responses to Xenoestrogens: Physiological Outcomes. *Steroids* **2012**, *77*, 910–917. [CrossRef] [PubMed]
23. Zhao, S.; Zhang, B.; Yang, J.; Zhou, J.; Xu, Y. Linear discriminant analysis. *Nat. Rev. Methods Primers* **2024**, *4*, 70. [CrossRef]
24. Zeng, J.; Guo, Y.; Han, Y.; Li, Z.; Yang, Z.; Chai, Q.; Wang, W.; Zhang, Y.; Fu, C. A Review of the Discriminant Analysis Methods for Food Quality Based on Near-Infrared Spectroscopy and Pattern Recognition. *Molecules* **2021**, *26*, 749. [CrossRef]
25. Todeschini, R.; Consonni, V. *Handbook of Molecular Descriptors*; Wiley-VCH: Weinheim, Germany, 2000. [CrossRef]
26. Balaban, A.T. Highly discriminating distance-based topological index. *Chem. Phys. Lett.* **1982**, *89*, 399–404. [CrossRef]
27. Bishop, C.M. *Pattern Recognition and Machine Learning*; Springer: New York, NY, USA, 2006. [CrossRef]
28. Hastie, T.; Tibshirani, R.; Friedman, J. *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*; Springer: New York, NY, USA, 2009. [CrossRef]
29. Radzi, S.F.M.; Karim, M.K.A.; Saripan, M.I.; Rahman, M.A.A.; Isa, I.N.C.; Ibahim, M.J. Hyperparameter Tuning and Pipeline Optimization via Grid Search Method and Tree-Based AutoML in Breast Cancer Prediction. *J. Pers. Med.* **2021**, *11*, 978. [CrossRef]

30. McLachlan, G.J. *Discriminant Analysis and Statistical Pattern Recognition*; Wiley-Interscience: Hoboken, NJ, USA, 2004. [\[CrossRef\]](#)
31. Bergstra, J.; Bengio, Y. Random Search for Hyper-Parameter Optimization. *J. Mach. Learn. Res.* **2012**, *13*, 281–305.
32. He, H.; Garcia, E.A. Learning from Imbalanced Data. *IEEE Trans. Knowl. Data Eng.* **2009**, *21*, 1263–1284.
33. Kuhn, M.; Johnson, K. *Applied Predictive Modeling*; Springer: New York, NY, USA, 2013. [\[CrossRef\]](#)
34. Cohen, J. A Coefficient of Agreement for Nominal Scales. *Educ. Psychol. Meas.* **1960**, *20*, 37–46. [\[CrossRef\]](#)
35. Wang, M.; Yu, F.; Zhang, Y.; Chang, W.; Zhou, M. The Effects and Mechanisms of Flavonoids on Cancer Prevention and Therapy: Focus on Gut Microbiota. *Int. J. Biol. Sci.* **2022**, *18*, 1451–1475. [\[CrossRef\]](#)
36. Keserü, G.M.; Makara, G.M. The influence of lead discovery strategies on the properties of drug candidates. *Nat. Rev. Drug Discov.* **2009**, *8*, 203–212. [\[CrossRef\]](#)
37. Brylinski, M. Aromatic interactions at the ligand–protein interface: Implications for the development of docking scoring functions. *Chem. Biol. Drug Des.* **2018**, *91*, 380–390. [\[CrossRef\]](#) [\[PubMed\]](#)
38. Jeffrey, G.A. *An Introduction to Hydrogen Bonding*; Oxford University Press: New York, NY, USA, 1997.
39. Kier, L.B.; Hall, L.H. *Molecular Connectivity in Structure-Activity Analysis*; Research Studies Press: Letchworth, UK, 1986.
40. Labute, P. A widely applicable set of descriptors. *J. Mol. Graph. Model.* **2000**, *18*, 464–477. [\[CrossRef\]](#) [\[PubMed\]](#)
41. Gasteiger, J.; Marsili, M. Iterative partial equalization of orbital electronegativity—A rapid access to atomic charges. *Tetrahedron* **1980**, *36*, 3219–3228. [\[CrossRef\]](#)
42. Nicholson, J.K.; Wilson, I.D. Opinion: Understanding ‘global’ systems biology: Metabonomics and the continuum of metabolism. *Nat. Rev. Drug Discov.* **2003**, *2*, 668–676. [\[CrossRef\]](#)
43. Steinkraus, K.H. *Handbook of Indigenous Fermented Foods*; Marcel Dekker Inc.: New York, NY, USA, 1996; ISBN 9780824793524.
44. Negi, A.; Kesari, K.K.; Voisin-Chiret, A.S. Estrogen Receptor- α Targeting: PROTACs, SNIPERs, Peptide-PROTACs, Antibody Conjugated PROTACs and SNIPERs. *Pharmaceutics* **2022**, *14*, 2523. [\[CrossRef\]](#)
45. Spaepen, S.; Vanderleyden, J.; Remans, R. Indole-3-acetic acid in microbial and microorganism–plant signaling. *FEMS Microbiol. Rev.* **2007**, *31*, 425–448. [\[CrossRef\]](#)
46. Parr, A.J.; Bolwell, G.P. Phenols in the plant and in man. The potential for possible nutritional enhancement of the diet by modifying the phenols content or profile. *J. Sci. Food Agric.* **2000**, *80*, 985–1012. [\[CrossRef\]](#)
47. Piwowarski, J.P.; Stanisławska, I.; Granica, S.; Stefańska, J.; Kiss, A.K. Phase II conjugates of urolithins isolated from human urine and potential role of β -glucuronidases in their disposition. *Drug Metab. Dispos.* **2017**, *45*, 657–665. [\[CrossRef\]](#)
48. Selma, M.V.; Espín, J.C.; Tomás-Barberán, F.A. Interaction between phenolics and gut microbiota: Role in human health. *J. Agric. Food Chem.* **2009**, *57*, 6485–6501. [\[CrossRef\]](#)
49. Zhang, A.Q.; Mitchell, S.C.; Smith, R.L. Dietary precursors of trimethylamine in man: A pilot study. *Food Chem. Toxicol.* **1999**, *37*, 515–520. [\[CrossRef\]](#)
50. Toldrá, F. The role of muscle enzymes in dry-cured meat products with different drying conditions. *Trends Food Sci. Technol.* **2006**, *17*, 164–168. [\[CrossRef\]](#)
51. Jay, J.M.; Loessner, M.J.; Golden, D.A. *Modern Food Microbiology*; Springer Science & Business Media: New York, NY, USA, 2005; ISBN 9780387231808.
52. Davidson, P.M.; Sofos, J.N.; Branan, A.L. *Antimicrobials in Food*; CRC Press: Boca Raton, FL, USA, 2005; ISBN 9780824753337.
53. Leroy, F.; De Vuyst, L. Lactic acid bacteria as functional starter cultures for the food fermentation industry. *Trends Food Sci. Technol.* **2004**, *15*, 67–78. [\[CrossRef\]](#)
54. Hill, C.; Guarner, F.; Reid, G.; Gibson, G.R.; Merenstein, D.J.; Pot, B.; Morelli, L.; Canani, R.B.; Flint, H.J.; Salminen, S.; et al. Expert consensus document: The International Scientific Association for Probiotics and Prebiotics consensus statement on the scope and appropriate use of the term probiotic. *Nat. Rev. Gastroenterol. Hepatol.* **2014**, *11*, 506–514. [\[CrossRef\]](#) [\[PubMed\]](#)
55. Gill, S.K.; Panda, S.H.; Singh, A.; Meena, D.K. Polyphenols in human health and disease: Current knowledge and future directions. *J. Pharmacogn. Phytochem.* **2015**, *4*, 17–27.
56. Slavin, J. Fiber and prebiotics: Mechanisms and health benefits. *Nutrients* **2013**, *5*, 1417–1435. [\[CrossRef\]](#)
57. Adams, M.R.; Hall, C.J. Growth inhibition of food-borne pathogens by lactic and acetic acids and their mixtures. *Int. J. Food Sci. Technol.* **1988**, *23*, 287–292. [\[CrossRef\]](#)

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