chemistry proceedings

MDPI

Proceeding Paper

LIFE.PTML Model Development Targeting Calmodulin Pathway Proteins †

Maider Baltasar Marchueta [1], Naia López [1], Sonia Arrasate [1], Matthew M. Montemore [2] and Humberto González-Díaz [1,3,4,*]

[1] Department of Organic and Inorganic Chemistry, University of the Basque Country UPV/EHU, 48940 Leioa, Spain; email1@email.com (M.B.M.); email2@email.com (N.L.); email3@email.com (S.A.)
[2] Department of Chemical and Biomolecular Engineering, Tulane University, 6823 St Charles Avenue, New Orleans, LA 70118, USA; email4@email.com (M.M.M.)
[3] Biofisika Institute, CSIC-UPV/EHU, 48940 Leioa, Spain
[4] IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Spain
* Correspondence: email5@email.com (H.G.-D.)
† Presented at the 29th International Electronic Conference on Synthetic Organic Chemistry (ECSOC-29); Available online: https://sciforum.net/event/ecsoc-29.

Abstract

Developing predictive models for drug efficacy is challenged by the complexity and heterogeneity of bioassay data. Here, we present LIFE.PTML, a methodology integrating drug Lifecycle (L), Information Fusion (IF), Encoding (E), Perturbation Theory (PT), and Machine Learning (ML), to predict compound activity across diverse experimental conditions. Using a dataset of 3748 molecule-assay combinations targeting calmodulin (CaM) and related proteins, LIFE.PTML combines chemical and protein descriptors, quantifies experimental variability via perturbation operators, and trains non-linear classifiers, including XGBoost and Gradient Boosting. XGBoost achieved the best performance, with 88.9% test accuracy and ROC AUC of 0.959, while feature importance analysis highlighted contributions from both drug- and protein-level descriptors. The results demonstrate that LIFE.PTML provides a robust, flexible, and interpretable framework for predictive chemoinformatics, facilitating the integration of multi-source data for drug discovery applications.

Keywords: drug discovery; calmodulin; chemoinformatics; machine learning; LIFE.PTML

Academic Editor(s): Name

Published: date

Citation: Marchueta, M.B.; López, N.; Arrasate, S.; Montemore, M.M.; González-Díaz, H. LIFE.PTML Model Development Targeting Calmodulin Pathway Proteins. *Chem. Proc.* **2025**, *volume number*, x. https://doi.org/10.3390/xxxxx

1. Introduction

While developing new drug formulations holds promise for improved treatments, the process is long and expensive [1]. For this reason, improving efficiency and accuracy in predicting the efficacy of potential new drugs could save both time and resources, ultimately accelerating the development of effective therapies [2]. Computational tools, including chemoinformatics models, have transformed drug discovery, enabling more systematic exploration of chemical space and prediction of bioactivity [3,4]. However, many traditional chemoinformatic approaches struggle to handle the complexity, heterogeneity, and volume of modern biomedical data, limiting their predictive power.

To overcome these limitations, advanced machine learning (ML) techniques have been increasingly employed to extract patterns from large and diverse datasets. Our group has developed the LIFE.PTML methodology, which integrates Life cycle (L) of

drugs, Information Fusion (IF), Encoding processes (E), Perturbation Theory (PT) and AI/ML techniques. LIFE.PTML allows the combination of multiple data sources—including chemical descriptors, protein features, and assay conditions—while PT operators quantify experimental variability and perturbations, providing normalized inputs for ML models. Previous applications of LIFE.PTML have shown its effectiveness in predicting drug–target interactions and other medicinal chemistry problems [5–7].

These computational approaches are particularly valuable for studying protein targets with critical biological functions [8]. For example, calmodulin (CaM), a key mediator of calcium signaling, is of great interest due to its structural complexity, central role in cellular function [9], and involvement in various diseases [10–12]. By leveraging LIFE.PTML models, it is possible to predict compound activity across diverse assay conditions, including variations in target proteins, experimental concentrations, and assay types. In this context, LIFE.PTML provides a framework for predicting the likelihood that a compound will be active under specific conditions (i.e., $f(v_{ij}) = 1$), integrating information from both the chemical and protein sides of the system.

The main objective of this work is to develop and validate LIFE.PTML-based chemoinformatic models capable of predicting assay efficacy with high accuracy. These models are trained using both linear and non-linear ML algorithms, including Random Forest, SVM, Decision Tree, K-Nearest Neighbors, Gradient Boosting, and XGBoost. By combining multi-source data, perturbation quantification, and advanced non-linear modeling, the LIFE.PTML approach aims to improve predictive accuracy, generalizability, and interpretability of chemoinformatic models in complex biological systems.

2. Materials and Methods

LIFE.PTML analysis involved four phases: the IF process, E part, PT variability quantification, and AI/ML algorithm training, validation, and use. In the initial IF phase, data gathering, data curation and data pre-processing tasks were carried out. In fact, system conceptualization is conceptual decomposition of the system in different sub-systems that are easy to study. In this case, the system was theoretically divided into two subsystems: drug information related to assays and protein data related to assays. Taking this into account, databases were examined and processed. Continuing the LIFE.PTML process, in the PT phase, the reference function and perturbation theory operators (PTO) or moving averages (MA) were calculated, which are used to quantify all the perturbations/variability on the input variables for all subsystems of the query system with respect to conditions or labels for the systems of reference. Lastly, the ML-Phase involved the training and validation of different ML models [6–8,13,14]. The general procedure followed in this part can be seen schematically in the Figure 1.

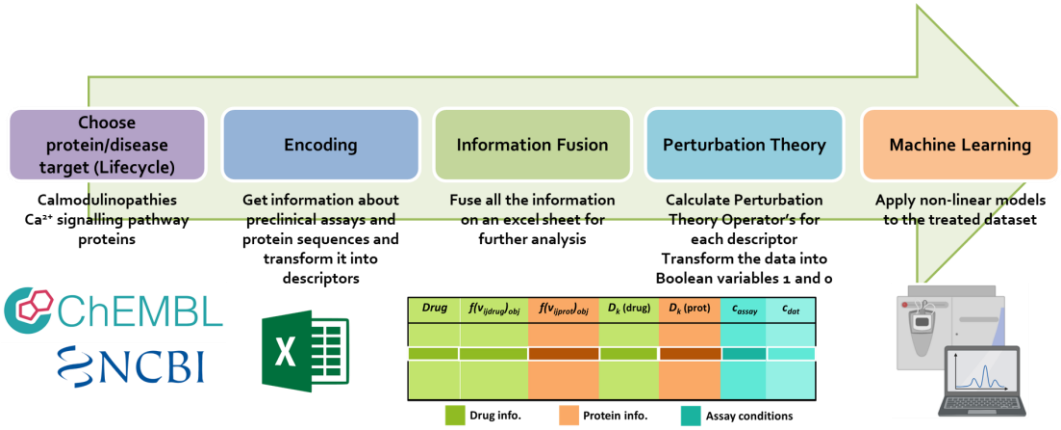

Figure 1. Workflow of the LIFE.PTML model development.

2.1. Information Fusion

In the first stage, the Information Fusion (IF) phase, the dataset was constructed using the ChEMBL database by retrieving assays involving proteins associated with calmodulin (CaM) within the Ca^{2+} signaling pathway, as defined in KEGG [15–17]. This procedure provided approximately 4000 assays, including responses of CaM- and riluzole-related compounds, where the activity values varied according to the compound tested and the experimental conditions. To complement this information, protein data were collected from NCBI resources [18,19], using GenBank and Protein BLAST to obtain the FASTA sequences and the three biologically active domains of each protein. Once gathered, descriptors were calculated and classified into drug-related and protein-related categories. Drug descriptors included molecular weight, Lipinski's Rule of Five, LogP [20–22], interatomic electronegativity, van der Waals surface area, and assay-specific variables such as inhibitor and substrate concentrations [23]. Protein descriptors were obtained through the MARCH-INSIDE 2.0® program, which encodes physicochemical properties such as electronegativity, polarity, and acidity along the amino acid sequence, propagating them through the residue network of each domain to generate quantitative values used as model inputs.

In the end, the dataset included ChEMBL compounds, biological activity data, and various assay types. Specifically, ChEMBL compounds included approximately 1000 chemical entities, many of which are FDA-approved drugs such as Gefitinib and Tamoxifen, along with investigational compounds. Biological activities were tested across 13 different measures, including parameters like IC_{50} (nM), K_i (nM), inhibition percentages, and potency (nM). The assays covered various target proteins, with key examples including Calmodulin (CaM), Myosin light chain kinase, and Epidermal growth factor receptor (EGFR). Furthermore, different assay cell types were used, such as HEK293 (human embryonic kidney cells), CHO (Chinese hamster ovary cells), and RAW264.7 (mouse macrophage cell line). The assays also spanned diverse tissues and organisms, including human (*Homo sapiens*), mouse (*Mus musculus*), and the pathogen *Plasmodium falciparum* (causing malaria). The total number of data points and different assays included in the final dataset amounts to 3748.

2.2. Perturbation Theory Operator's Calculation

In the second stage, the Perturbation Theory (PT) phase, Perturbation Theory Operators (PTOs) were calculated using Box–Jenkins moving average (MA) method [24]. The model considered two types of boundary conditions: those related to the assays themselves, such as target or assay type ($c_{assay} = (c_1, c_2, c_3, c_4, c_5)$), and those related to the dataset, including factors like organism or buffer ($c_{dat} = (c_6, c_7, c_8, c_9, c_{10})$). For each condition, the mean descriptor values were calculated (MA), and deviations from these averages were expressed as delta values ($\Delta Dk(c_j)$; Equation 1). In this way, the perturbation operators quantified the extent to which experimental variability influenced descriptor behavior.

$$\Delta D_k(\boldsymbol{c_j}) = D_k - < D_k(\boldsymbol{c_j}) > \quad (1)$$

2.3. Objective and Reference Function Calculation

The third stage focused on the definition of the output variable and the reference function. Since assays reported different measures of activity, such as IC_{50}, K_i, or inhibition percentages, a desirability function was introduced to standardize the classification. Variables expected to increase, such as inhibition or residual activity, were assigned a desirability of +1, while those expected to decrease, including IC_{50} or K_i, were assigned a value of –1. Cut-off criteria were then applied to distinguish active from inactive compounds, using 100 nM for concentration-based assays, 70% for inhibition percentages, and the

dataset mean when no conventional threshold was available. These cut-offs allowed the transformation of activity data into Boolean variables (objective function), identifying compounds as active ($f(v_{ij})_{obj} = 1$) or inactive ($f(v_{ij})_{obj} = 0$). At the same time, a reference function ($f(v_{ij})_{ref}$; Equation (2)) was calculated to estimate the probability of a compound being active under specific boundary conditions, providing a baseline measure for subsequent predictions.

$$f(v_{ij})_{ref} = p(f(v_{ij}/\boldsymbol{c}_j)_{obj} = 1))_{ref} = \frac{n(f(v_{ij}/\boldsymbol{c}_j)_{obs} = 1)}{n(f(v_{ij}/\boldsymbol{c}_j)_{obs})} \quad (2)$$

2.4. Machine Learning Model Development

Finally, in the Machine Learning (ML) phase, several non-linear models were implemented to predict whether a compound will be successful or not under some boundary conditions (i.e., objective function). Non-linear algorithms included Random Forest, Support Vector Machine with RBF kernel, Decision Tree, K-Nearest Neighbors, Gradient Boosting, and XGBoost. All models were developed in Python using scikit-learn and pandas libraries [25]. The dataset was divided into 80% for training and 20% for testing, with a performance threshold of 70% accuracy and sensitivity as the minimum requirement [26]. Hyperparameters were optimized through Grid Search and cross-validation, and model performance was evaluated using accuracy, sensitivity, specificity, AUC-ROC, and confusion matrices.

3. Results and Discussion

3.1. Dataset Construction

The information fusion stage resulted in a final dataset of 3748 molecule–assay combinations, corresponding to 1052 unique compounds. The distribution of the target variable was well-balanced, with 51.2% active cases and 48.8% inactive cases, providing favorable conditions for supervised learning without the need for additional resampling techniques. In terms of biological diversity, the dataset included 101 protein targets associated with Ca^{2+}-dependent signaling pathways, such as calmodulin, myosin light chain kinase, and epidermal growth factor receptor (EGFR). This diversity confers both biological relevance and complexity to the dataset. Additionally, the presence of 19 species, 11 cell types, and multiple assay conditions captured a wide spectrum of experimental variability, enhancing the representativeness of the model. The definition of the objective function and reference function enabled the integration of heterogeneous bioactivity measures (IC_{50}, K_i, inhibition %) into a binary classification scheme. A threshold of 100 nM for IC_{50}/K_i and 70% for inhibition assays was applied, resulting in a clear separation between active and inactive cases.

3.2. LIFE.PTML Non-Linear Models

Several non-linear machine learning models were trained to predict compound activity under specific assay conditions. These included ensemble methods (XGBoost, Gradient Boosting, Random Forest), support vector machines (SVM with RBF kernel), decision trees, and k-nearest neighbors (KNN). Each model was optimized via Grid Search or Randomized Search and evaluated using an 80/20 training/test split. ROC curves were calculated for both training and test sets to assess discriminative performance across models (Figure 2).

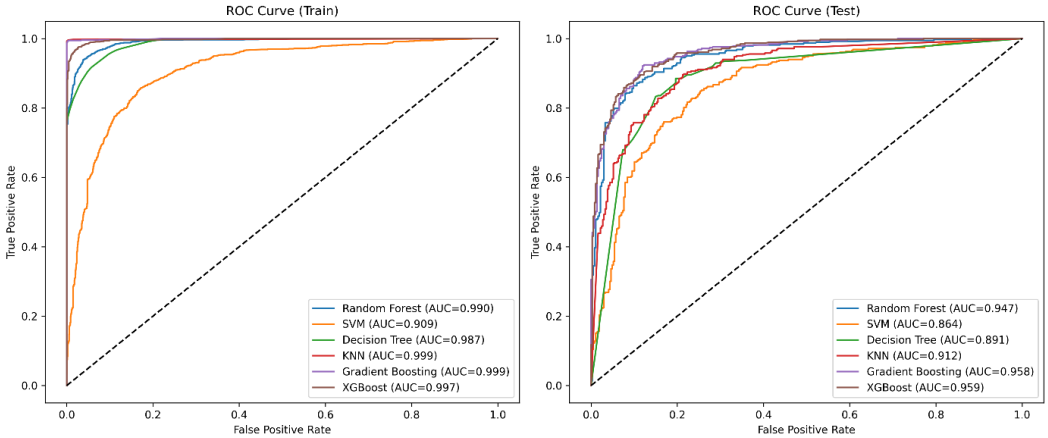

Figure 2. ROC Curves for nonlinear models developed. ROC curves for training and test sets for XGBoost and Gradient Boosting classifiers, illustrating discriminative performance of LIFE.PTML models.

A detailed comparison of the optimized hyperparameters and performance statistics for all non-linear models is presented in Table 1. The comparison of non-linear classifiers highlighted substantial differences in predictive behavior. Ensemble methods, particularly XGBoost and Gradient Boosting, consistently outperformed other approaches across all evaluation metrics. XGBoost achieved the most balanced performance, with a test accuracy of 88.9% and ROC AUC of 0.959, indicating excellent discriminative power. Training metrics were markedly higher (accuracy 97.3%, AUC 0.999), suggesting a degree of overfitting; however, the test set performance confirmed good generalization capacity. Gradient Boosting displayed very similar performance (test accuracy 89.5%, ROC AUC 0.958), with slightly better calibration between sensitivity and specificity. In contrast, models such as Random Forest also reached strong AUC values (0.947), but with greater variability in sensitivity and precision. KNN severely overfit the training data (accuracy 99.6%) and underperformed in the test set (accuracy 83.3%), underscoring its limited applicability in high-dimensional descriptor spaces. SVM also lagged behind, with reduced ROC AUC (0.864), consistent with the challenges of capturing nonlinear relationships without extensive kernel optimization (Table 1).

Table 1. Summary of hyperparameters and performance metrics for LIFE.PTML non-linear models.

Model	Best Hyperparameters (Randomized/ Grid Search)	Train Accuracy	Test Accuracy	Precision	Recall	F1-Score	ROC AUC
Random Forest	n_estimators = 610; max_depth = 26; criterion = entropy; max_features = sqrt; min_samples_split = 7; min_samples_leaf = 2	0.947	0.877	0.872	0.890	0.881	0.947
SVM (RBF)	C = 10; gamma = scale; kernel = rbf; degree = 3; probability = True	0.827	0.788	0.751	0.875	0.808	0.864
Decision Tree	criterion = entropy; max_depth = 32; min_samples_split = 7; min_samples_leaf = 5; max_features = sqrt	0.930	0.837	0.844	0.836	0.840	0.891
KNN	n_neighbors = 15; weights = distance; *p* = 2 (Euclidean)	0.996	0.833	0.827	0.851	0.839	0.912
Gradient Boosting	n_estimators = 310; learning_rate = 0.14; max_depth = 5; subsample = 0.75; max_features = sqrt	0.995	0.895	0.892	0.903	0.898	0.958

XGBoost	n_estimators = 160; learning_rate = 0.27; max_depth = 12; subsample = 1.0; colsample_bytree = 0.55; gamma = 1.7; reg_alpha = 0.3; reg_lambda = 1.5	0.973	0.889	0.891	0.893	0.892	0.959

Confusion matrices for XGBoost and Gradient Boosting (Figure 3) show high classification accuracy. XGBoost correctly classified 97.5% of training samples and 89.4% of test samples, while Gradient Boosting achieved 99.0% on training and 90.1% on test sets. Misclassifications remained low in both models, confirming robust generalization and reliable prediction of compound activity across assay conditions.

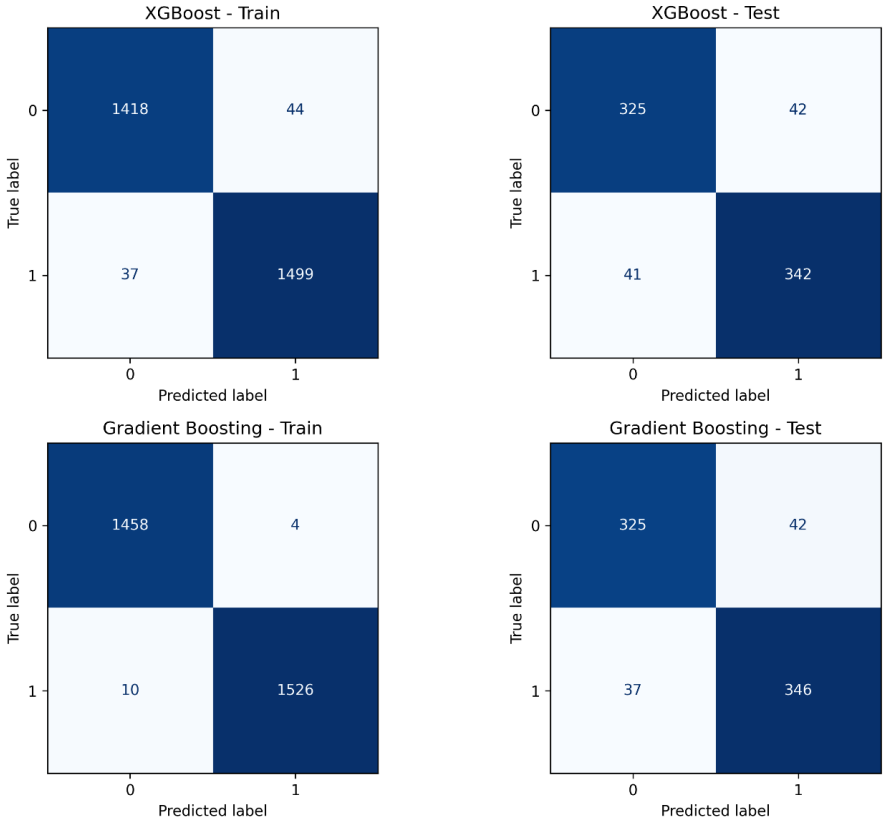

Figure 3. Confusion Matrices for Boosting-Based IFPTML Models.

The feature importance profile obtained from the XGBoost model (Figure 4) highlights the integration of both drug- and protein-derived descriptors within the LIFE.PTML framework. The most influential feature was the reference function $f(v_{ij})_{ref}$, which represents the baseline activity probability under specific boundary conditions. This indicates that the perturbation-based normalization step provides a strong prior for classification. Among the experimental conditions, the inhibitor ($\Delta V(drug, assay)_2$) and substrate concentrations ($\Delta V(drug, assay)_1$) emerged as key contributors, underscoring the role of assay setup in determining compound performance. Several drug-level perturbation operators linked to global physicochemical properties, particularly electronegativity ($\Delta D(drug, dat)_{1-2}$, $\Delta D(drug, dat)_{013}$, $\Delta D(drug, dat)_{053-054}$, $\Delta D(drug, dat)_{101}$) and van der Waals terms ($\Delta V(drug, dat)_1$), also ranked highly. This aligns with the chemical intuition that electronic distribution and steric interactions strongly influence ligand–protein binding within Ca^{2+}-dependent pathways. Importantly, protein-derived features were also represented, notably the second-domain electronegativity ($\Delta D(prot, dom_{II}, dat)_1$), reflecting the contribution of protein structural context to predictive power. Taken together, the feature ranking demonstrates that the LIFE.PTML methodology effectively captures multi-source information,

balancing chemical, biological, and experimental descriptors to provide mechanistically plausible predictions.

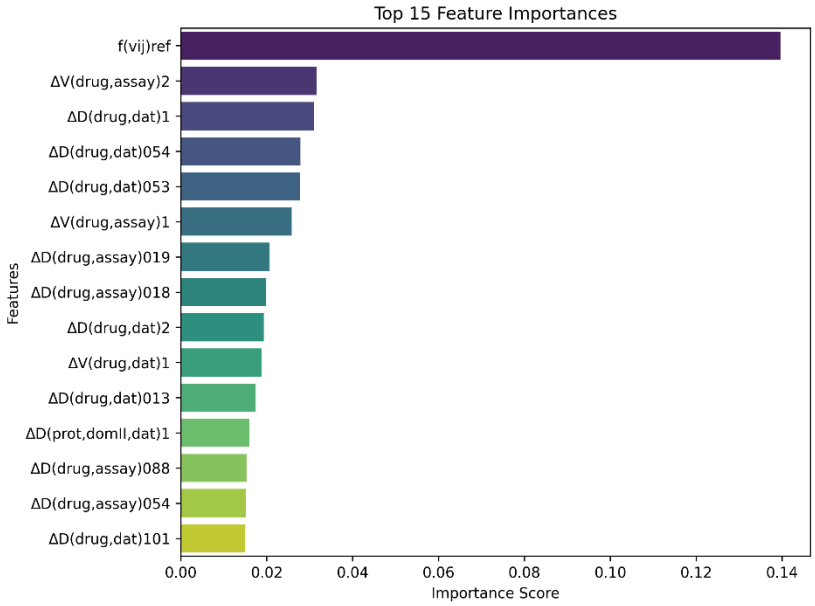

Figure 4. Top 15 Feature Importances from XGBoost Model.

Taken together, these findings confirm that boosting-based classifiers, particularly XGBoost, represent the most reliable LIFE.PTML models in this study. While some overfitting is present, the balance between sensitivity, specificity, and ROC AUC supports their suitability for predictive applications in complex biological systems.

4. Conclusions

In this study, we have developed and systematically applied the LIFE.PTML methodology, a combined framework of drug Lifecycle (L), Information Fusion (IF), Perturbation Theory (PT), and Machine Learning (ML), to handle complex, heterogeneous datasets arising from bioassays involving calmodulin (CaM) and related protein targets. The methodology provides a structured and reproducible pipeline for integrating chemical and protein descriptors, quantifying experimental variability, and predicting assay outcomes under diverse experimental conditions.

By combining these elements, the LIFE.PTML approach supports the development of non-linear predictive models capable of integrating multidimensional data while controlling for experimental variability. The framework has been designed to be flexible and generalizable, allowing the inclusion of diverse assay types, experimental conditions, and descriptor classes. Overall, the LIFE.PTML methodology represents a comprehensive and modular approach to predictive modeling in cheminformatics. It establishes a robust workflow that can be applied to future drug discovery efforts, offering a structured means to incorporate chemical, biological, and experimental variability into predictive ML frameworks.

Author Contributions:

Funding: This work was supported by MCIN/AEI/10.13039/501100011033 (PID2022-137365NB-100, 2023–2026) and by IT1558-22. It also received support from the Basque Government through the predoctoral grant *Programa Predoctoral de Formación de Personal Investigador* (reference PRE_2024_1_0387).

Institutional Review Board Statement:

Informed Consent Statement:

Data Availability Statement: All data and code used in this study are publicly available at the GitHub repository: https://github.com/maiderbaltasar/ECSOC29.

Acknowledgments: The authors thank the Basque Government for their support through the pre-doctoral fellowship and acknowledge the institutional support of UPV/EHU.

Conflicts of Interest: The authors declare no conflicts of interest.

Abbreviations

The following abbreviations are used in this manuscript:

CaM	Calmodulin
E	Encoding processes
EGFR	Epidermal Growth Factor Receptor
f(vij)obj	Objective function: binary label indicating compound activity under specific assay conditions (1 = active, 0 = inactive)
f(vij)ref	Reference function: probability of compound being active under specific bounary conditions (baseline for predictions)
IC50	Half maximal inhibitory concentration
Ki	Inhibition constant
KNN	K-Nearest Neighbors
LIFEPTML	Life cycle (L) + Information Fusion (IF) + Encoding processes (E) + Perturbation Theory (PT) + Machine Learning (ML)
IF	Information Fusion
ML	Machine Learning
MA	Moving Average
PT	Perturbation Theory
PTO	Perturbation Theory Operator
ROC AUC	Area under the Receiver Operating Characteristic curve
SVM	Support Vector Machine

References

1. DiMasi, J.A.; Grabowski, H.G.; Hansen, R.W. Innovation in the Pharmaceutical Industry: New Estimates of R&D Costs. *J. Health Econ.* **2016**, *47*, 20–33. https://doi.org/10.1016/j.jhealeco.2016.01.012.
2. Sadybekov, A.V.; Katritch, V. Computational Approaches Streamlining Drug Discovery. *Nature* **2023**, *616*, 673–685. https://doi.org/10.1038/s41586-023-05905-z.
3. García, I.; Fall, Y.; Gómez, G.; González-Díaz, H. First Computational Chemistry Multi-Target Model for Anti-Alzheimer, Anti-Parasitic, Anti-Fungi, and Anti-Bacterial Activity of GSK-3 Inhibitors In Vitro, In Vivo, and in Different Cellular Lines. *Mol. Divers.* **2011**, *15*, 561–567. https://doi.org/10.1007/s11030-010-9280-3.
4. Ni, D.; Liu, D.; Zhang, J.; Lu, S. Computational Insights into the Interactions between Calmodulin and the c/nSH2 Domains of P85α Regulatory Subunit of PI3Kα: Implication for PI3Kα Activation by Calmodulin. *Int. J. Mol. Sci.* **2018**, *19*, 151. https://doi.org/10.3390/ijms19010151.
5. Ferreira da Costa, J.; Silva, D.; Caamaño, O.; Brea, J.M.; Loza, M.I.; Munteanu, C.R.; Pazos, A.; García-Mera, X.; González-Díaz, H. Perturbation Theory/Machine Learning Model of ChEMBL Data for Dopamine Targets: Docking, Synthesis, and Assay of New l -Prolyl- l -Leucyl-Glycinamide Peptidomimetics. *ACS Chem. Neurosci. J.* **2018**, *9*, 2572–2587. https://doi.org/10.1021/acschemneuro.8b00083.
6. He, S.; Nader, K.; Abarrategi, J.S.; Bediaga, H.; Nocedo-Mena, D.; Ascencio, E.; Casanola-Martin, G.M.; Castellanos-Rubio, I.; Insausti, M.; Rasulev, B.; et al. NANO.PTML Model for Read-across Prediction of Nanosystems in Neurosciences. Computational Model and Experimental Case of Study. *J. Nanobiotechnol.* **2024**, *22*, 435. https://doi.org/10.1186/s12951-024-02660-9.

7. Baltasar-Marchueta, M.; Llona, L.; M-Alicante, S.; Barbolla, I.; Ibarluzea, M.G.; Ramis, R.; Salomon, A.M.; Fundora, B.; Araujo, A.; Muguruza-Montero, A.; et al. Identification of Riluzole Derivatives as Novel Calmodulin Inhibitors with Neuroprotective Activity by a Joint Synthesis, Biosensor, and Computational Guided Strategy. *Biomed. Pharmacother.* **2024**, *174*, 116602. https://doi.org/10.1016/j.biopha.2024.116602.
8. Martínez-Arzate, S.G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Díaz-Albiter, H.M.; Vázquez-Chagoyán, J.C.; González-Díaz, H. PTML Model for Proteome Mining of B-Cell Epitopes and Theoretical–Experimental Study of Bm86 Protein Sequences from Colima, Mexico. *J. Proteome Res.* **2017**, *16*, 4093–4103. https://doi.org/10.1021/acs.jproteome.7b00477.
9. Zhang, M.; Abrams, C.; Wang, L.; Gizzi, A.; He, L.; Lin, R.; Chen, Y.; Loll, P.J.; Pascal, J.M.; Zhang, J. Structural Basis for Calmodulin as a Dynamic Calcium Sensor. *Structure* **2012**, *20*, 911–923. https://doi.org/10.1016/j.str.2012.03.019.
10. Kotta, M.-C.; Sala, L.; Ghidoni, A.; Badone, B.; Ronchi, C.; Parati, G.; Zaza, A.; Crotti, L. Calmodulinopathy: A Novel, Life-Threatening Clinical Entity Affecting the Young. *Front. Cardiovasc. Med.* **2018**, *5*, 175. https://doi.org/10.3389/fcvm.2018.00175.
11. O'Day, D.H. Calmodulin Binding Proteins and Alzheimer's Disease: Biomarkers, Regulatory Enzymes and Receptors That Are Regulated by Calmodulin. *Int. J. Mol. Sci.* **2020**, *21*, 7344. https://doi.org/10.3390/ijms21197344.
12. Mustaly-Kalimi, S.; Gallegos, W.; Marr, R.A.; Gilman-Sachs, A.; Peterson, D.A.; Sekler, I.; Stutzmann, G.E. Protein Mishandling and Impaired Lysosomal Proteolysis Generated through Calcium Dysregulation in Alzheimer's Disease. *Proc. Natl. Acad. Sci. USA* **2022**, *119*, e2211999119. https://doi.org/10.1073/pnas.2211999119.
13. Blay, V.; Yokoi, T.; González-Díaz, H. Perturbation Theory–Machine Learning Study of Zeolite Materials Desilication. *J. Chem. Inf. Model.* **2018**, *58*, 2414–2419. https://doi.org/10.1021/acs.jcim.8b00383.
14. Nocedo-Mena, D.; Cornelio, C.; Camacho-Corona, M. del R.; Garza-González, E.; Waksman de Torres, N.; Arrasate, S.; Sotomayor, N.; Lete, E.; González-Díaz, H. Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. *J. Chem. Inf. Model.* **2019**, *59*, 1109–1120. https://doi.org/10.1021/acs.jcim.9b00034.
15. Kanehisa-Laboratories. KEGG Kyoto Encyclopedia of Genes and Genomes. Retrieved from Https://www.Genome.Jp/Kegg/. 1995.
16. Bento, A.P.; Gaulton, A.; Hersey, A.; Bellis, L.J.; Chambers, J.; Davies, M.; Krüger, F.A.; Light, Y.; Mak, L.; McGlinchey, S.; et al. The ChEMBL Bioactivity Database: An Update. *Nucleic Acids Res.* **2014**, *42*, D1083–D1090. https://doi.org/10.1093/nar/gkt1031.
17. Papadatos, G.; Overington, J.P. The ChEMBL Database: A Taster for Medicinal Chemists. *Future Med. Chem.* **2014**, *6*, 361–364. https://doi.org/10.4155/fmc.14.8.
18. Martí-Carreras, J.; Gener, A.; Miller, S.; Brito, A.; Camacho, C.; Connor, R.; Deboutte, W.; Glickman, C.; Kristensen, D.; Meyer, W.; et al. NCBI's Virus Discovery Codeathon: Building "FIVE"—The Federated Index of Viral Experiments API Index. *Viruses* **2020**, *12*, 1424. https://doi.org/10.3390/v12121424.
19. Morales, J.; Pujar, S.; Loveland, J.E.; Astashyn, A.; Bennett, R.; Berry, A.; Cox, E.; Davidson, C.; Ermolaeva, O.; Farrell, C.M.; et al. A Joint NCBI and EMBL-EBI Transcript Set for Clinical Genomics and Research. *Nature* **2022**, *604*, 310–315. https://doi.org/10.1038/s41586-022-04558-8.
20. Lipinski, C.A.; Lombardo, F.; Dominy, B.W.; Feeney, P.J. Experimental and Computational Approaches to Estimate Solubility and Permeability in Drug Discovery and Development Settings. *Adv. Drug Deliv. Rev.* **2001**, *46*, 3–26. https://doi.org/10.1016/S0169-409X(00)00129-0.
21. Veber, D.F.; Johnson, S.R.; Cheng, H.-Y.; Smith, B.R.; Ward, K.W.; Kopple, K.D. Molecular Properties That Influence the Oral Bioavailability of Drug Candidates. *J. Med. Chem.* **2002**, *45*, 2615–2623. https://doi.org/10.1021/jm020017n.
22. Kim, T.; Park, H. Computational Prediction of Octanol–Water Partition Coefficient Based on the Extended Solvent-Contact Model. *J. Mol. Graph. Model.* **2015**, *60*, 108–117. https://doi.org/10.1016/j.jmgm.2015.06.004.
23. Danishuddin; Khan, A.U. Descriptors and Their Selection Methods in QSAR Analysis: Paradigm for Drug Design. *Drug Discov. Today* **2016**, *21*, 1291–1302. https://doi.org/10.1016/j.drudis.2016.06.013.
24. Chou, Y.-L. *Statistical Analysis: With Business and Economic Applications*; Holt, Rinehart and Winston: Fort Worth, TX, USA, 1975.

25. Pedregosa, F.; Varoquaux, G.; Gramfort, A.; Michel, V.; Thirion, B.; Grisel, O.; Blondel, M.; Prettenhofer, P.; Weiss, R.; Dubourg, V.; et al. Scikit-Learn: Machine Learning in Python. *J. Mach. Learn. Res.* **2011**, *12*, 2825–2830.
26. Bediaga, H.; Arrasate, S.; González-Díaz, H. PTML Combinatorial Model of ChEMBL Compounds Assays for Multiple Types of Cancer. *ACS Comb. Sci. J.* **2018**, *20*, 621–632. https://doi.org/10.1021/acscombsci.8b00090.