ECSOC-29
2025
Conference

The 29th Intl Electronic Conference on Synthetic Organic Chemistry

14–28 November 2025 | Online

MDPI

LIFE.PTML MODEL DEVELOPMENT TARGETING CALMODULIN PATHWAY PROTEINS

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

[1] Department of Organic and Inorganic Chemistry, University of the Basque Country UPV/EHU, 48940, Leioa, Spain.

[2] Department of Chemical and Biomolecular Engineering, Tulane University, 6823 St Charles Avenue, New Orleans, Louisiana 70118, United States.

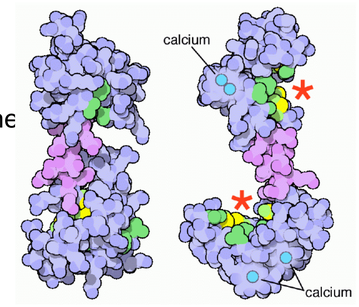
[3] Biofisika Institute, CSIC-UPV/EHU, 48940, Leioa, Spain.

[4] IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Spain.

INTRODUCTION

CALMODULIN (CaM)[1-3]

- CaM plays a crucial role in various cellular signaling processes.
- It acts as a mediator in processes such as synaptic transmission and plasticity, regulation of enzymatic activities, modulation of ionic channe functions, and control of gene expression.
- Diseases associated: cardiopathies, calmodulinopathies and neurodegenerative diseases.
- It is useful to use public information available to create models that allow predicting new drugs to treat those diseases associated with CaM.

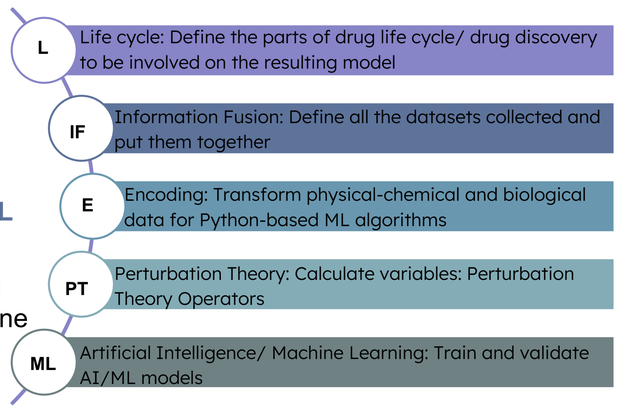


PROBLEM

- The analysis of large datasets of neuroprotective compounds that form multi-target complex networks is very difficult with classic Cheminformatics techniques.

ADDRESSING THE LIMITATIONS: LIFE.PTML MODELS

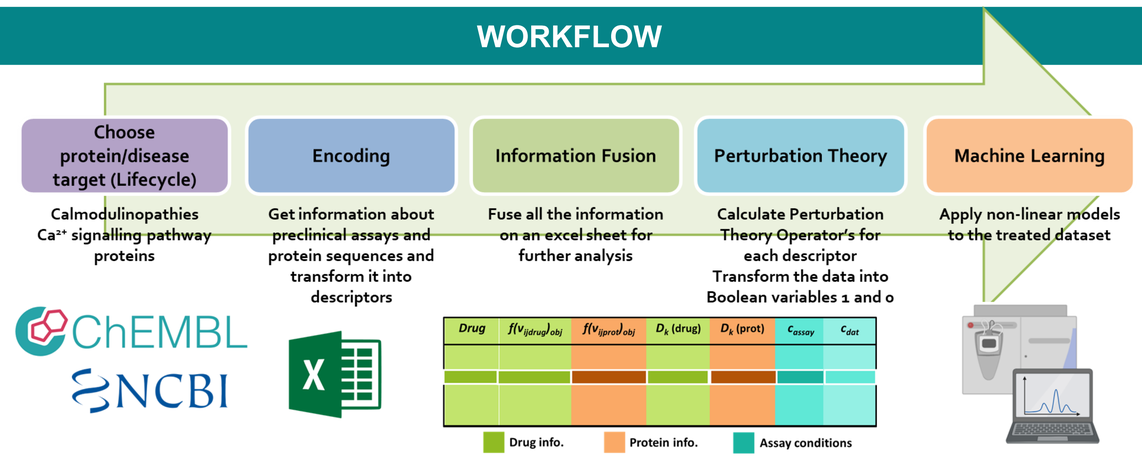
- The PT operators used are based on moving averages of multiple conditions, which combine different characteristics and simplify the difficulty of managing all the data.[4,5]



AIM

- We develop and apply a targeted model focusing specifically on diseases related to CaM and related proteins in the calcium signaling pathway.
- Develop a reproducible workflow (LIFEPTML) to integrate chemical, protein, and assay descriptors for predictive modeling.
- Train and validate non-linear machine learning models, including XGBoost and Gradient Boosting, to predict compound activity under diverse conditions.
- Identify key chemical and protein features contributing to assay outcomes, providing mechanistic insight for drug discovery.

WORKFLOW



METHODOLOGY: LIFE.PTML MODEL DESIGN

1.- INFORMATION FUSION

Name (Label)	Descriptor or variable information
Molecular weight (D_1(drug))	The molecular weight of the drug
Lipinski's Rule of Five (D_2(drug$_i$))	Lipinski's rule of five for the drug
^{n}octanol/water partition coefficient (D_3(drug$_i$))	AlogP value for the drug
Electronegativity (D_{001}(drug$_i$)-D_{035}(drug$_i$))	Average electronegativity difference between each atom of the drug and the adjacent atoms[a]
Van der Waals forces (D_{036}(drug$_i$)-D_{070}(drug$_i$))	Average of the Van der Waals forces between each atom of the drug and the adjacent atoms[a]
Contribution to the ^{n}octanol/water partition coefficient (D_{071}(drug$_i$)-D_{105}(drug$_i$))	Average of AlogP between each atom of the drug and the adjacent atoms
Substrate concentration (V_1)	The concentration of the substrate used on the assay (μM)
Inhibitor concentration (V_2)	The concentration of the inhibitor used on the assay (μM)
Electronegativity of first domain D_1(prot$_t$,dom$_I$)-D_5(prot$_t$,dom$_I$)	Average of first domain electronegativity for the five different levels[a]
Electronegativity of second domain D_1(prot$_t$,dom$_{II}$)-D_5(prot$_t$,dom$_{II}$)	Average of second domain electronegativity for the five different levels[a]
Electronegativity of third domain D_1(prot$_t$,dom$_{III}$)-D_5(prot$_t$,dom$_{III}$)	Average of third domain electronegativity for the five different levels*

2.- PERTURBATION THEORY

- Moving Average Box–Jenkins method.
- Boundary conditions:
- Assay-level: target, type, etc.
- Dataset-level: organism, buffer, etc.
- Perturbation = deviation from average

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

3.- OBJECTIVE & REFERENCE FUNCTION

- Standardization of outputs.
- Cut-offs: 100 nM (IC_{50}/Ki), 70% (% inhibition), dataset mean (others).
- Boolean classification: Active ($f(v_{ij})_{obj}$ = 1) vs Inactive ($f(v_{ij})_{obj}$ = 0).
- Reference function (baseline probability):

$$f(v_{ij})_{ref} = p\left(f(v_{ij}/\boldsymbol{c_j})_{obj} = 1\right) = \frac{n\left(f(v_{ij}/\boldsymbol{c_j})_{obs}=1\right)}{n\left(f(v_{ij}/\boldsymbol{c_j})_{obs}\right)}$$

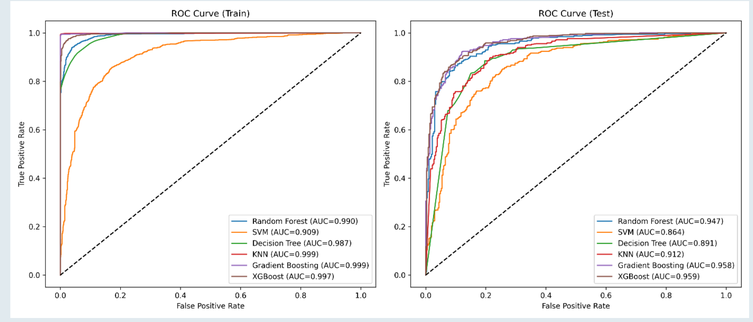
4.- MACHINE LEARNING

- Algorithms: RF, SVM (RBF), Decision Tree, KNN, Gradient Boosting, XGBoost.
- Dataset split: 80% training / 20% testing.
- Hyperparameter optimization: Grid Search + CV.
- Metrics: Accuracy, Sensitivity, Specificity, ROC-AUC, Confusion Matrix.

RESULTS

NON-LINEAR MODELS

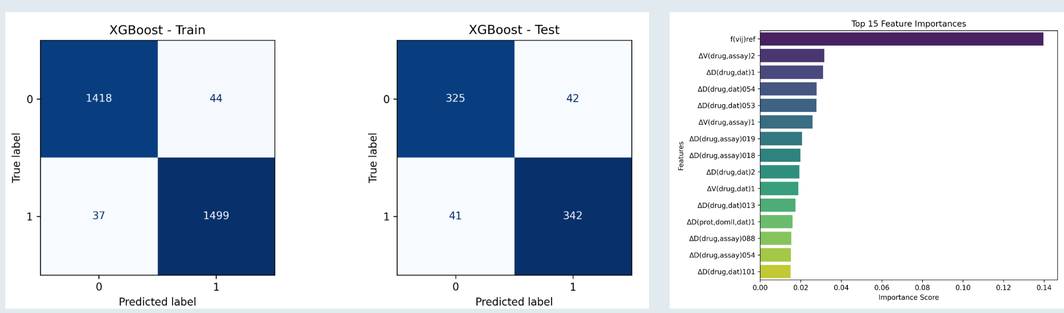
- Boosting methods (XGBoost, Gradient Boosting) gave the highest AUROC (>0.95).
- KNN showed strong overfitting (train ≈ 99%, test ≈ 83%).
- SVM achieved lower discriminative power (AUROC ≈ 0.86).
- Ensemble methods (RF, GB, XGB) proved more robust to dataset heterogeneity.



Model	Train Accuracy	Test Accuracy	Precision	Recall	F1-score	ROC AUC
Random Forest	0.947	0.877	0.872	0.890	0.881	0.947
SVM (RBF)	0.827	0.788	0.751	0.875	0.808	0.864
Decision Tree	0.930	0.837	0.844	0.836	0.840	0.891
KNN	0.996	0.833	0.827	0.851	0.839	0.912
Gradient Boosting	0.995	0.895	0.892	0.903	0.898	0.958
XGBoost	0.973	0.889	0.891	0.893	0.892	0.959

XGBOOST ANALYSIS

- Train accuracy: 97.3%; Test accuracy: 88.9%
- Confusion matrices: >97% correct train, ~89% correct test
- Slight overfitting, but good generalization
- Reference function (f(vij)ref) = most important feature
- Assay variables, such as substrate ($\Delta V(drug, assay)_1$) and inhibitor concentrations ($\Delta V(drug, assay)_2$), were strongly relevant.
- Drug descriptors, including 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 ($\Delta V(drug, dat)_1$), were key for activity prediction.
- Protein descriptors, particularly domain II electronegativity ($\Delta D(prot, domII, dat)_1$), also emerged as influential.



CONCLUSION

- LIFEPTML provides a framework for integrating chemical, protein, and assay data.
- Perturbation theory operators effectively capture experimental variability across heterogeneous assays.
- Non-linear models, particularly XGBoost and Gradient Boosting, deliver high predictive accuracy and generalization.
- The workflow can handle diverse assay types, experimental conditions, and multi-source datasets.
- LIFEPTML offers a flexible approach to accelerate drug discovery by guiding compound selection for experimental validation.

REFERENCES

1. 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, doi:10.1016/j.str.2012.03.019.
2. Walsh, M.P. Review Article Calmodulin and Its Roles in Skeletal Muscle Function. *Can Anaesth Soc J* **1983**, *30*, 390–398, doi:10.1007/BF03007862.
3. 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, doi:10.3389/fcvm.2018.00175.
4. 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, doi:10.1186/s12951-024-02660-9.
5. 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. *Biomedicine & Pharmacotherapy* **2024**, 174, 116602, doi:10.1016/j.biopha.2024.116602.