

## FRAMA 1.0: Framework for Moving Average Operators Calculation in Data Analysis

Bernabé Ortega-Tenezaca <sup>a,b</sup>, Viviana F. Quevedo-Tumaili <sup>a,b</sup>, and Humbert González-Díaz <sup>a, b,\*</sup>

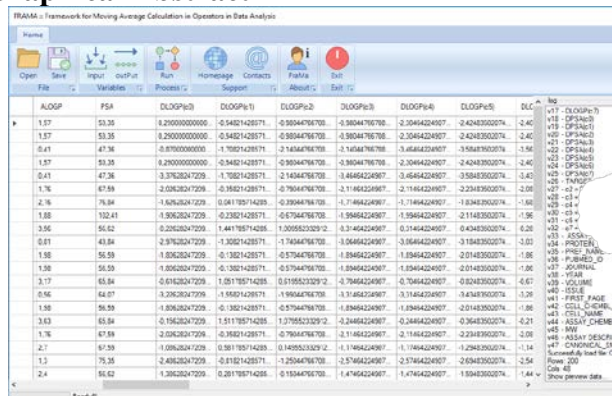
<sup>a</sup> RNASA-IMEDIR, Computer Science Faculty, University of A Coruña, 15071, A Coruña, Spain.

<sup>b</sup> Universidad Estatal Amazónica, Puyo, Pastaza, Ecuador

<sup>c</sup> Department of Organic Chemistry II, University of the Basque Country UPV/EHU, 48940, Leioa, Biscay, Spain

<sup>d</sup> IKERBASQUE, Basque Foundation for Science, 48011, Bilbao, Biscay, Spain

### Graphical Abstract



ALOGP	PSA	DLOGPe(0)	DLOGPe(1)	DLOGPe(2)	DLOGPe(3)	DLOGPe(4)	DLOGPe(5)	DLOGPe(6)
1.57	53.35	8.2500000000	-0.5482142857	-0.9004767676	-1.3004476768	-2.2044242407	-2.4243200274	-2.40
1.57	53.35	8.2500000000	-0.5482142857	-0.9004767676	-1.3004476768	-2.2044242407	-2.4243200274	-2.40
0.41	47.36	0.8700000000	-1.7921428571	-3.1456476768	-3.1626476768	-3.4564242407	-3.5843200274	-1.86
1.57	53.35	8.2500000000	-0.5482142857	-0.9004767676	-1.3004476768	-2.2044242407	-2.4243200274	-2.40
0.41	47.36	3.3762824729	-1.7092142857	-2.1434476768	-3.4664242407	-3.4664242407	-3.5843200274	-3.43
1.76	67.59	-2.032824729	-0.3821428571	-0.7904476768	-1.1464242407	-2.1164242407	-2.2343200274	-2.08
2.16	71.84	-1.62624729	0.9119214285	-0.3904476768	-1.7484242407	-1.7484242407	-1.8343200274	-1.58
1.88	102.41	1.9362824729	-0.2321428571	-0.5704476768	-1.8964242407	-1.8964242407	-2.1143200274	-1.98
3.56	95.62	0.2162824729	1.44178714285	1.2095232027	2.3744242407	0.7144242407	0.0424200274	6.20
0.61	43.84	2.9762824729	-1.3202142857	-1.7464476768	-1.9464242407	-3.8664242407	-3.1843200274	3.03
1.58	56.59	-1.8262824729	-0.1302142857	-0.5704476768	-1.8964242407	-1.8964242407	-2.0143200274	-1.86
1.58	56.59	-1.8262824729	-0.1302142857	-0.5704476768	-1.8964242407	-1.8964242407	-2.0143200274	-1.86
3.17	63.84	0.6162824729	1.05178714285	0.5195232027	-2.7964242407	0.7064242407	0.8243200274	4.67
0.64	64.07	3.2162824729	-1.9502142857	-1.9864476768	-3.1144242407	-3.1144242407	-3.4343200274	-1.29
1.88	96.58	-1.8262824729	-0.1302142857	-0.5704476768	-1.8964242407	-1.8964242407	-2.0143200274	-1.86
3.43	63.84	-0.1562824729	1.51178714285	1.0785232027	-2.2464242407	-0.2464242407	-0.3643200274	-0.21
1.76	67.59	-2.032824729	-0.3821428571	-0.7904476768	-1.1464242407	-2.1164242407	-2.2343200274	-2.08
2.1	77.59	-1.092824729	0.9119214285	0.1425232027	-1.1764242407	-1.1764242407	-1.2943200274	-1.14
2.1	77.59	-2.4362824729	-0.8182142857	-1.2504476768	-2.5764242407	-2.5764242407	-2.6943200274	-2.54
2.4	95.62	1.3862824729	-0.28178714285	-0.1804476768	-1.4764242407	-1.4764242407	-1.5943200274	-1.44

**Abstract.** Moving Average (MA) operators are used in Box-Jenkins's ARIMA models in time series analysis (1). We can use MA operators of structural descriptors or parameters in complex datasets in Omics, Medicinal Chemistry, Nanotechnology, etc. (2-7). Speck-Planche and Cordeiro have also used this kind of models in multiple problems (8-11). In this work, we develop a desktop application that allows applying mathematical and statistical calculations in batches, on input and output variables selected by the user. From the obtained result a percentage sample of data is taken with a random contrast on which Machine Learning algorithms are applied

### Introduction

In principle, we can calculate numerical parameters to quantify the structure of chemical compounds, peptides, and/or proteins. We can also use them as input variables for Machine Learning (ML) algorithms in order to predict the biological properties of these drugs, peptides, or proteins (13-29). On the other hand, Perturbation Theory (PT) models allow us to predict the solutions to a query problem (q) based on a previous known solution for a similar problem or problem of reference (r). In a recent work, we outlined a new type of ML method called PTML (PT + ML) based on both kind of models with applications in drug discovery and proteome research (25, 30). The PTML method uses different kind of PT operators to predict the properties of one system based on the properties of a system of reference. For instance, Moving Average (MA) operators used in Box-Jenkins's ARIMA models in time series analysis (31). We have used MA operators of structural descriptors are useful to quantify multiple conditions or parameters in complex datasets in Omics, Medicinal Chemistry,

Nanotechnology, *etc.* (32-37). Speck-Planche and Cordeiro have also used this kind of models in multiple problems (38-41).

### Discussion

González-Díaz *et al.* introduced a general-purpose PTML modeling technique useful to quantify the effect of perturbations in complex bio-molecular systems including DPINS and other networks (48, 49). Using PTML the model we can predict the values of the scoring function  $f(\varepsilon_{ij})_{\text{new}}$  for the DPI. The PTML model start using as input with the expected value of biological activity  $f(\varepsilon_{ij})_{\text{expt}}$  for one compound assayed in the conditions  $c_j$  and add the values of the PT operators  $\Delta D_k(m_i, c_j)$ . The expected value  $f(\varepsilon_{ij})_{\text{expt}} = \langle \varepsilon_{ij} \rangle$  is the average value of the biological activity parameter  $\varepsilon_{ij}$  for all cases in ChEMBL dataset with the same  $c_0 = \text{Activity parameter } \varepsilon_{ij}(\text{Units})$ . These PT operators added  $\Delta D_k(m_i, c_j) = D_k(m_j) - \langle D_k(c_j) \rangle$  are intended to account for the changes (perturbations) in the system with respect to the expected values. Specifically, perturbations on the value of the molecular descriptors of the drug  $D_k(m_j)$  with respect to the expected value  $\langle D_k(c_j) \rangle$  for a drug measured under the conditions of the experiment  $c_j$ . These PT operators resemble the Box-Jenkins MA operators (25, 30). We use both Linear Discriminant Analysis (LDA) and Artificial Neural Network (ANN) algorithms to seek alternative linear and non-linear models (50). At follow, we depict the compact and developed forms of a PTML linear model:

$$f(\varepsilon_{ij})_{\text{new}} = a_0 \cdot f(\varepsilon_{ij})_{\text{expt}} + \sum_{k=1}^{k_{\text{max}}} \sum_{j=0}^{j_{\text{max}}} a_{jk} \cdot \Delta D_k(m_i, c_j) + e_0 \quad (1)$$

$$f(\varepsilon_{ij})_{\text{new}} = a_0 \cdot f(\varepsilon_{ij})_{\text{expt}} + \sum_{k=1}^{k_{\text{max}}} \sum_{j=0}^{j_{\text{max}}} a_{jk} \cdot \left( D_k(m_i)_{\text{new}} - \langle D_k(c_j) \rangle_{\text{ref}} \right) + e_0 \quad (2)$$

### Results and Discussion

FRAMA, is a desktop application that supports different file formats, allows perform data preprocessing tasks on the selection of input and output variables, and its sub classification as grouping variables and continuous variables, where operations, operators and obtaining parametric values are applied, such as Mergin Data, Shannon Entropy, Z-Score, Moving Average, Euclidian Distance, among others. From the results obtained, a sample is selected for the application of Machine Learning algorithms on a sample of data

### References

1. Box, G. E. P.; Jenkins, G. M., *Time series analysis*. Holden-Day: 1970; p 553.
2. Blazquez-Barbadillo, C.; Aranzamendi, E.; Coya, E.; Lete, E.; Sotomayor, N.; Gonzalez-Diaz, H., Perturbation theory model of reactivity and enantioselectivity of palladium-catalyzed Heck-Heck cascade reactions. *Rsc Advances* **2016**, 6, (45), 38602-38610.
3. Casanola-Martin, G. M.; Le-Thi-Thu, H.; Perez-Gimenez, F.; Marrero-Ponce, Y.; Merino-Sanjuan, M.; Abad, C.; Gonzalez-Diaz, H., Multi-output Model with Box-Jenkins Operators of Quadratic Indices for Prediction of Malaria and Cancer Inhibitors Targeting Ubiquitin-Proteasome Pathway (UPP) Proteins. *Current Protein & Peptide Science* **2016**, 17, (3), 220-227.
4. Romero-Duran, F. J.; Alonso, N.; Yanez, M.; Caamano, O.; Garcia-Mera, X.; Gonzalez-Diaz, H., Brain-inspired cheminformatics of drug-target brain interactome, synthesis, and assay of TVP1022 derivatives. *Neuropharmacology* **2016**, 103, 270-278.
5. Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. D. S., Computational Tool for Risk Assessment of Nanomaterials: Novel QSTR-Perturbation Model for Simultaneous Prediction of Ecotoxicity and Cytotoxicity of Uncoated and Coated Nanoparticles under Multiple Experimental Conditions. *Environmental Science & Technology* **2014**, 48, (24), 14686-14694.
6. Luan, F.; Kleandrova, V. V.; Gonzalez-Diaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N., Computer-aided nanotoxicology: assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale* **2014**, 6, (18), 10623-30.
7. Alonso, N.; Caamano, O.; Romero-Duran, F. J.; Luan, F.; Cordeiro, M. N. D. S.; Yanez, M.; Gonzalez-Diaz, H.; Garcia-Mera, X., Model for High-Throughput Screening of Multitarget

- Drugs in Chemical Neurosciences: Synthesis, Assay, and Theoretic Study of Rasagiline Carbamates. *Acs Chemical Neuroscience* **2013**, 4, (10), 1393-1403.
8. Speck-Planche, A.; Dias Soeiro Cordeiro, M. N., Speeding up Early Drug Discovery in Antiviral Research: A Fragment-Based in Silico Approach for the Design of Virtual Anti-Hepatitis C Leads. *ACS Comb Sci* **2017**, 19, (8), 501-512.
  9. Kleandrova, V. V.; Ruso, J. M.; Speck-Planche, A.; Dias Soeiro Cordeiro, M. N., Enabling the Discovery and Virtual Screening of Potent and Safe Antimicrobial Peptides. Simultaneous Prediction of Antibacterial Activity and Cytotoxicity. *ACS Comb Sci* **2016**, 18, (8), 490-8.
  10. Speck-Planche, A.; Cordeiro, M. N., Computer-aided discovery in antimicrobial research: In silico model for virtual screening of potent and safe anti-pseudomonas agents. *Comb Chem High Throughput Screen* **2015**, 18, (3), 305-14.
  11. Speck-Planche, A.; Cordeiro, M. N., Simultaneous virtual prediction of anti-Escherichia coli activities and ADMET profiles: A chemoinformatic complementary approach for high-throughput screening. *ACS Comb Sci* **2014**, 16, (2), 78-84.