SUPPORTING INFORMATION

Towards Rational Nanomaterial Design by Prediction of Drug-Nanoparticle Systems Interaction *vs*. Bacteria Metabolic Networks

*Karel Diéguez-Santana [1], Bakhtiyor Rasulev [2], and Humberto González-Díaz [1,3,4] **

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

[2] Department of Coatings and Polymeric Materials,
North Dakota State University, Fargo, ND, 58102, USA,

[3] BIOFISIKA, Basque Center for Biophysics CSIC-UPVEH, 48940 Leioa, Spain.

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

QSAR model report form (QMRF) for the Prediction of Drug-Nanoparticle Systems Interaction vs. Bacteria Metabolic Networks, following the OECD template

Model for prediction of the Drug-Nanoparticle Systems Interaction vs. Bacteria Metabolic Networks for Rational Nanomaterial Design, developed using IFPTML model (Linear and non linear)

In this section we rely on other QMRF reports conducted to demonstrate that the developed model is fully consistent with the OECD principles for validation of predictive models for regulatory purposes [1]we summarize here all available information on the development and evaluation of the IFPTML model in a concise manner. For this purpose, we follow the guidance of the JRC QSAR model database and the research by [2] that performed the necessary alterations for nanoinformatics data.[3]

Principle 1–A defined endpoint.

Species	Multiple. ChEMBL AD: >25 bacteria species with >90 strains. NP: 34 bacteria species/strains MN dataset >20 bacteria species
Endpoint	Antibacterial activity
Endpoint comments	Antibacterial activity (values of >300 parameters (MIC, IC50, etc.)). ChEMBL >160 000 biological assays of >50000 compounds NP dataset includes 1 out of 4 parameters of activity for 300 pre-clinical assays of NP vs AD
Endpoint units	Diverse (MIC: μg.mL^{-1} MBC μg ml^{-1}, IZ mm, etc)
Dependent variable	The scoring function f(vij, vnj, vsj)calc used to calculate the posterior probabilities which the DADNP is short listed for experimental biological assay

Experimental protocol	Full experimental description can be found in Hwang, et al. (2012). Journal of medical microbiology, 2012, 61, 1719-1726, DOI: 10.1099/jmm.0.047100-0, and Vazquez-Muñoz, et al. (2019), PLoS One, 14, e0224904-e0224904. DOI: 10.1371/journal.pone.0224904 Short description: The dataset consists of different cases of AD, NP, and MN We assigned all cases to either training (subset = t) or validation (subset = v) series. We selected the original data from the three datasets randomly to create triads. These triads are formed by one AD, one NP, and one MN cases (representing putative DADNP vs. MN interactions). However, we need to impose some constrains in some labels due to the IF process. The cases forming one triad have the same value of the labels c0d and c0n (same biological property) of AD and NP whenever it was possible. The cases of the triads have also the same cd1, cn1, and cs1 (bacteria specie) whenever it was possible. All triads have been ordered according to these main labels (stratified sampling). Subsequently, cases were assigned to set = t and set = v (representative sampling) in a proportion 75% vs. 25%.
Endpoint data quality and variability	The antibacterial biological activity was extracted from preclinical cases reported in the ChEMBL, NP, Ochem, and MNs databases of Barabasi's group (Jeong et al.) Nature, 2000, 407, 651-654.

Principle 2–An unambiguous algorithm.

Type of model	Perturbation-Theory Machine Learning Information Fusion (IFPTML), LDA, Artificial neural network, multilayer perceptron (MLP), k-nearest neighbour (kNN), Random Forest, boosting algorithms, etc
Explicit algorithm	Use kNN with k value equal to 1(LinearNNSearch with EuclideanDistance as type of nearest neighbor search algorithm), SVM non-linear with Radial Basis Function (RBF) kernel.
Descriptors in the model	Statistically significant descriptors used for prediction of the Drug-Nanoparticle Systems Interaction vs. Bacteria Metabolic Networks include all the important variables AD structure and assay conditions, NP properties, CA structure, NP assay conditions, MN structural parameters.
Descriptor selection	Number and type of descriptors initially screened: 18 descriptors: Reference Functions each systems and Reference Functions (AD-NP-MN) f(cd$_0$,cn$_0$ cd$_s$)$_{ref}$ AD: ΔShLOGP, Logarithm of the n-Octanol/Water Partition coefficient, ΔShPSA, Topological Polar Surface Area, ΔShNVLR, Number of Violations of Lipinski's Rule. NP: AMVn: Average of the molar volumes of the elements that form the nanoparticle. Aen Average of the electronegativities of the elements that form the nanoparticle. Apn Average of the polarizabilities of the elements that form the nanoparticle. APSn Average Size of the nanoparticle. Time: Time assays of NP-AD. Metabolic network: Number of substrate (N), number of links (Lins

	And Louts), number of individual reactions or temporary substrate-enzyme complexes (R), number of enzymes (E), the exponent γ and the diameter of the metabolic network (D). Systems: ΔΔSh(LOGPi,1c,2c): LOGP drugs, NP Coating agents 1 and Coating agents 2 (if applicable) and ΔΔSh(PSAi,1c,2c): PSA drugs, nanoparticle, NP Coating agents 1 and Coating agents 2 (if applicable). Method used to select the descriptors: Forward Step-Wise (FSW) feature selection strategy (Linear Discriminant Analysis (LDA)) Expert-Guided Selection (EGS) was used incorporated missing features.
Algorithm and descriptor generation	Experimental measurements: See Nocedo-Mena, et al. (2019). J Chem Inf Model, 59(3), 1109-1120. doi:10.1021/acs.jcim.9b00034.
Software name and version for descriptor generation	STATISTICA 6.0 software PTML Multi-Label Algorithms: Models, Software, and Applications (See Ortega-Tenezaca, et al. (2020). DOI: 10.2174/1568026620666200916122616)
DADNPs/Descriptors ratio	15545.75 (124366:8, number of data rows divided by the number of significant descriptors in the training set)

Principle 3–A defined domain of applicability.

Description of the applicability domain of the model	The Domain of Applicability domain (DoA) is defined by fixed boundaries (threshold). The threshold is calculated by considering Euclidean distances between all training set DADNPs vs MN Systems.
Method used to assess the applicability domain	Leverage Method
Software name and version for applicability domain assessment	STATISTICA 6.0 software was used.[4] Origin Pro, version 2019, OriginLab was used for graphics.
Limits of applicability	h* threshold: 0.0002 The leverage threshold was fixed at the critical hat value (h*=3(p+1)/n), where p is the number of descriptors of the model and n is the number of training compounds). Predictions outside this threshold are considered unreliable.

Principle 4–Appropriate measures of goodness-of-fit, robustness and predictivity.

Availability of the training set	Dataset AD-MN: Nocedo-Mena, et al. (2019). J Chem Inf Model, 59(3), 1109-1120. doi:10.1021/acs.jcim.9b00034 Dataset NP: Speck-Planche,(2015). Nanomedicine (Lond), 10(2), 193-204. doi:10.2217/nnm.14.96
Available information for the training set	Preclinical Antibacterial activity of the >160 000 biological assays of >50000. Preclinical antibacterial activity of the 300 pre-clinical assays of NP vs AD Analytical information on the experimental process can be found in Nocedo-Mena, et al. (2019). J Chem Inf Model, 59(3), 1109-1120.

	doi:10.1021/acs.jcim.9b00034 Wang et al. (2017). Int J Nanomedicine. doi:10.2147/IJN.S121956 Information on the molecular descriptors calculation can be found at: Ortega-Tenezaca, et al. (2020). DOI: 10.2174/1568026620666200916122616) and Nocedo-Mena, et al. (2019). J Chem Inf Model, 59(3), 1109-1120. doi:10.1021/acs.jcim.9b00034
Data for each descriptor variable for the training set	Yes
Data for the dependent variable (response) for the training set	Yes
Other information about the training set	Total of 124366 cases for the dependent and independent variables
Pre-processing of data before modelling	Shannon's information scaling of input variables
Statistics for goodness-of-fit	Chi square, validation through an external test set, Matthew's correlation coefficient (MCC),[5]**(Eq13),** F1 score, the random correlation model of classification proposed by Lucic et al.[6, 7] and the Y-randomization test
Ac > 0.6	0.97
$Ac_{vext} > 0.5$	0.97
MCC	0.798
F1score	0.808
AUROC	0.99
Robustness – Statistics obtained by Y-scrambling	Accuracy = 39.5 (24 iterations)
Robustness – Statistics obtained by other methods	Yes, see above
Availability of the external validation set	Yes
Available information for the external validation set	Yes
Data for each descriptor variable for the external validation set	Yes
Data for the dependent variable for the external validation set	Yes
Other information about the external validation set	Total of 41445 cases from DADNP for the dependent and independent variables
Experimental design of test set	Partitioning of the initial dataset using random, stratified sampling (75:25 training: test sets)
Predictivity – Statistics obtained by external validation	Ac> 0.6; Result: 0.97

Predictivity – Assessment of the external validation set	The external validation set is 25% of the initial dataset and all predictions for the validation set fall within the domain of applicability
Comments on the external validation of the model	N/A

Principle 5–A mechanistic interpretation.

Mechanistic basis of the model	The structural properties of the AD and assay conditions and the complexity of the reaction metabolic network of the bacterial species influence the inhibition of the biological activity of the DADNP system (e.g., the Octanol-water partition ratio property, that expressing the lipophilicity of drugs, influences multiple drug discovery studies).
A priori or a posteriori mechanistic interpretation	The NP properties such as NP size (lower inhibition of antibacterial activity), the molar volumes of the elements that form the nanoparticle and Time assays of NP-AD contribute to increase the biological activity and influence the activity of the dual AD NP system. Similarly, The Topological Polar Surface Area of Coating agents.
Other information about the mechanistic interpretation	No other information available.

REFERENCES (Supporting Information)

1. OECD Validation of (Q)SAR Models. https://www.oecd.org/chemicalsafety/risk-assessment/validationofqsarmodels.htm (december 15),
2. Papadiamantis, A. G.; Afantitis, A.; Tsoumanis, A.; Valsami-Jones, E.; Lynch, I.; Melagraki, G., Computational enrichment of physicochemical data for the development of a ζ-potential read-across predictive model with Isalos Analytics Platform. *NanoImpact* **2021**, 22, 100308.
3. Hub, E. S. JRC QSAR Model Database. https://ec.europa.eu/jrc/en/scientific-tool/jrc-qsar-model-database (december 13),
4. Hill, T.; Lewicki, P., *Statistics: Methods and Applications*. 1st edition ed.; StatSoft, Inc.: 2005; p 800.
5. Chicco, D.; Jurman, G., The advantages of the Matthews correlation coefficient (MCC) over F1 score and accuracy in binary classification evaluation. *BMC Genomics* **2020**, 21, 6.
6. Batista, J.; Vikić-Topić, D.; Lučić, B., The Difference Between the Accuracy of Real and the Corresponding Random Model is a Useful Parameter for Validation of Two-State Classification Model Quality. *Croatica Chemica Acta* **2016**, 89, 527-534.
7. Lučić, B.; Batista, J.; Bojović, V.; Lovrić, M.; Kržić, A. S.; Bešlo, D.; Nadramija, D.; Vikić-Topić, D., Estimation of random accuracy and its use in validation of predictive quality of classification models within predictive challenges. *Croatica Chemica Acta* **2019**, 92, 379-391.