



This open access document is posted as a preprint in the Beilstein Archives at <https://doi.org/10.3762/bxiv.2024.10.v1> and is considered to be an early communication for feedback before peer review. Before citing this document, please check if a final, peer-reviewed version has been published.

This document is not formatted, has not undergone copyediting or typesetting, and may contain errors, unsubstantiated scientific claims or preliminary data.

Preprint Title On Additive Artificial Intelligence Discovery of Nanoparticle-Neurodegenerative Disease Drug Delivery Systems

Authors Shan He, Julen Segura Abarrategi, Harbil Bediaga, Sonia Arrasate and Humberto González-Díaz

Publication Date 15 Feb. 2024

Article Type Full Research Paper

Supporting Information File 1 SI01_Tables.xlsb; 34.1 KB

ORCID® IDs Shan He - <https://orcid.org/0000-0001-6965-6276>; Julen Segura Abarrategi - <https://orcid.org/0009-0006-3098-5543>; Sonia Arrasate - <https://orcid.org/0000-0003-2601-5959>



License and Terms: This document is copyright 2024 the Author(s); licensee Beilstein-Institut.

This is an open access work under the terms of the Creative Commons Attribution License (<https://creativecommons.org/licenses/by/4.0>). Please note that the reuse, redistribution and reproduction in particular requires that the author(s) and source are credited and that individual graphics may be subject to special legal provisions.

The license is subject to the Beilstein Archives terms and conditions: <https://www.beilstein-archives.org/xiv/terms>.

The definitive version of this work can be found at <https://doi.org/10.3762/bxiv.2024.10.v1>

On Additive Artificial Intelligence Discovery of Nanoparticle-Neurodegenerative Disease Drug Delivery Systems

Shan He^{1,2}, Julen Segura Abarrategi¹, Harbil Bediaga^{2,3}

Sonia Arrasate¹, and Humberto González-Díaz^{1,4*}

Address: ¹Department of Organic and Inorganic Chemistry, University of Basque Country UPV/EHU, 48940 Leioa, Spain, ² IKERDATA S.L., ZITEK, UPV/EHU, Rectorate Building, nº6, 48940 Leioa, Greater Bilbao, Basque Country, Spain, ³ Department of Physical Chemistry, University of Basque Country UPV/EHU, 48940 Leioa, Spain and ⁴ IKERBASQUE, Basque Foundation for Science, 48011 Bilbao, Biscay, Spain.

E-mail: Humberto González-Díaz- humberto.gonzalezdiaz@ehu.es

*Corresponding author

Abstract

Neurodegenerative diseases are characterized by slowly progressive neuronal death. Conventional treatment strategies often fail due to poor solubility, lower bioavailability, and the inability to effectively cross the Blood–Brain Barrier (BBB). Therefore, the development of new Neurodegenerative Disease Drugs (NDDs) requires immediate attention. Nanoparticle (NP) systems are increasingly of interest for transporting NDDs to the central nervous system. However, discovering effective Nanoparticle Neuronal Disease Drug Delivery Systems (N2D3S) is challenging due to the vast number of NP and NDDs compound combinations, as well as various assays involved. Artificial Intelligence/Machine Learning (AI/ML) algorithms have the potential to accelerate this process by predicting the most promising NDDs and NP candidates for assay. Nevertheless, the relatively limited amount of reported data on N2D3S activity compared to assayed NDDs makes AI/ML analysis challenging. In this work, the IFPTML technique, which combines Information Fusion (IF), Perturbation Theory (PT), and Machine Learning (ML), was employed to address this challenge. Initially, we

conducted fusion into a unified dataset comprising 4403 NDDS assays from ChEMBL and 260 cytotoxicity NP assays from journal articles. Through a resampling process, three new working datasets were generated, each containing 500,000 cases. We utilized Linear Discriminant Analysis (LDA) along with Artificial Neural Networks (ANN) algorithms like Multi-Layer Perceptron (MLP) and Deep Learning Networks (DLN) to construct linear and non-linear IFPTML models, respectively. The IFPTML-LDA models exhibited Sensitivity (S_n) and Specificity (S_p) values in the range of 70% to 73% (>375K training cases) and 70% to 80% (>125K validation cases), respectively. Conversely, the IFPTML-MLP and IFPTML-DLN achieved S_n and S_p values in the range of 85% to 86% for both training and validation series. Additionally, IFPTML-ANN models showed an Area Under the Receiver Operating Curve (AUROC) of approximately 0.93 to 0.95. These results indicate that the IFPTML models could serve as valuable tools in the design of drug delivery systems for neurosciences.

Keywords

Neurodegenerative disease; Nanoparticle; Machine Learning; LDA; ANN

Introduction

Over the years, there has been a drastic change in diet and living standard of people worldwide. The inadequate diet, food consumption patterns, long working hours along with inactive lifestyle have headed for an inclination. It has brought about widespread disease in elderly population related to chronic degenerative/lifestyle/human-made diseases. The degenerative diseases are a type of heterogeneous disorder that is by progressive degeneration of structure and function of system/organs.[1] In spite of the fact that the initiating causes heading to these diseases are unidentified, oxidative damage seems reflected to play a vital role in the gradually progressive neuronal death, specifically the proliferation production of reactive oxygen and nitrogen species.[2] Among these, Alzheimer's and Parkinson's diseases can be considered as the most acute incurable. Traditional treatment approaches, such as acetyl-cholinesterase inhibitor drugs, often fail due to their poor solubility, lower bioavailability, and ineffective ability to cross the Blood–Brain Barrier (BBB).[3] In this sense, the development of new Neurodegenerative Diseases Drugs (NDDs) call for immediate attention.[4] The major obstacle encountered by NDDs is the selectivity of the BBB, which importantly limits the number of therapeutic substances able to reach the brain in order to induce a positive effect. Recently, many efforts have been made to develop systems that facilitate the passage of NDDs through the BBB.

Interestingly, nano-particle (NP) systems are gaining increasing interest among the possible Nanomedicine strategies for NDDs transport to the central nervous system.[5] For simplicity, we are going to call them Nanoparticles Neuronal Diseases Drug Delivery Systems (N2D3S). N2D3S have the ability to protect NDDs from chemical and enzymatic degradation, direct the active compound towards the target site with a substantial reduction of toxicity for the adjacent tissues, and pass physiological barriers increasing bioavailability without resorting to high dosage forms.[6] Therefore, researchers are studying and developing a new treatment approach that uses N2D3S to diagnosis and treatment.[7-10]

On the other hand, over the last few years, AI/ML models have been applied successfully to solve problems in different disciplines, specially, in the interface of chemistry and NDDs research.[11-14] In this sense, we consider AI/ML to be helpful in N2D3S to select the most efficient combination of NP and drug, taking into account the ADMET properties and ND-biological activity respectively.[15] Nevertheless, there are a relatively limited availability of NP experimental data reported in the scientific

literature so far in comparison to drug, which increases the difficulty of designing the system based on AI/ML techniques.

An additional essential downside of developing N2D3S with AI/ML techniques is the great complexity of the data to be explored. As a result, N2D3S development by the additive approach requires an AI/ML technique to achieve multi-output and multi-label classification.[16-19] In addition, the AI/ML technique takes into account a pre-processing step to perform Information Fusion (IF) of the preclinical essays for NDDs and cytotoxicity NP datasets. Nevertheless, most of AI/ML methods have been reported to date only consider the structural/molecular descriptors of the NDDs or NP as an input. Therefore, these authors exclude completely non-structural parameters, specifically experimental conditions of the preclinical assay in order to list NDDs or NP labels, correspondingly. Consequently, the resulting model cannot predict multi-output properties and/or labels such as different organisms, cell lines and so on.[20-32] Sizochenko *et al.* reported a new methodology for NP safety estimation in different organisms.[33] Predicting NP safety instead of biological activity is the objective of other studies as well.[32-34]

As a new strategy to tackle this problem, González-Díaz *et al.* have developed IFPTML, a multi-output, and input-coded multi-label ML method, which stands for Information Fusion (IF)+Perturbation-Theory (PT) + Machine Learning (ML) algorithm.[35] In the recent investigation study, IFPTML model has shown to be a powerful tool in molecular sciences and in NDs research for big dataset analysis tasks, which includes both structural and non-structural parameters. For instance, mapping drug, target protein, coated NP drug release systems,[36-44] multi-target networks of neuroprotective compounds for theoretical study of new asymmetric 1,2-rasagiline carbamates,[45] TOPS-MODE model of multiplexing neuroprotective effects of drugs and experimental-theoretic study of new 1,3-rasagiline derivatives potentially useful in neurodegenerative diseases,[46] or QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer, and neurosciences and so on.[47] Furthermore, this new strategy also has been used for very similar system to this research work, such as NP systems taking into account NP structure and coating agents, NP experimental conditions of synthesis loaded drug structure, co-therapy loaded drugs, assay conditions, *etc.*[48-52] Here we developed the IFPTML model for N2D3S proposal, containing simultaneously multiple NDD and NP components.

Results and Discussion

In order to build the IFPTML models we carried out the following steps mentioned in **Figure 1**, step-by step which shows the general workflow of all computational procedures followed up in this research. Furthermore, for better understanding of all steps, we annotated them 2.1, 2.2., etc. in regard with their enumeration in the following materials and methods section.

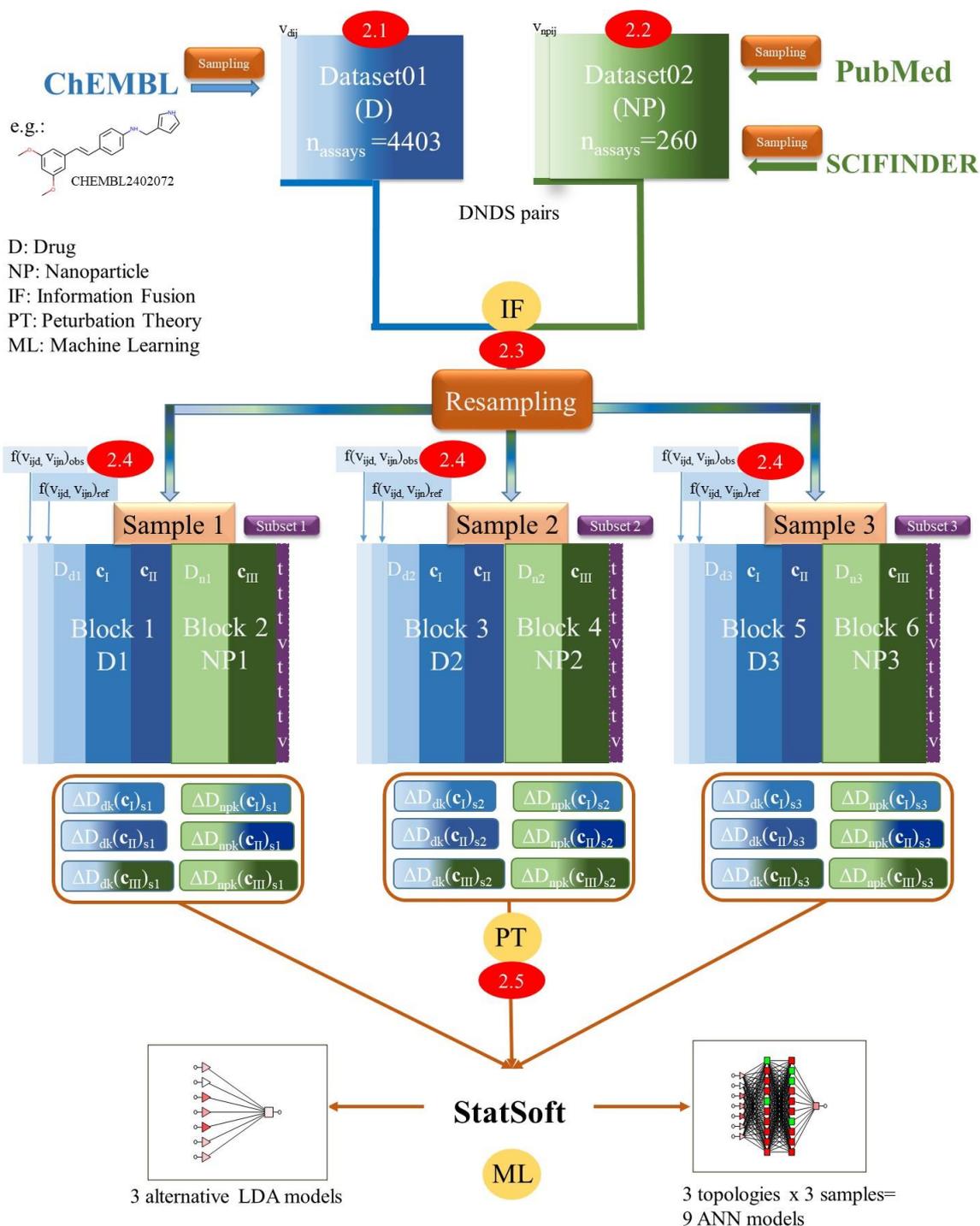


Figure 1: IFPTML detailed information-processing workflow. Step 2.1 and 2.2 Data collection (NDDs ChEMBL dataset and NP cytotoxicity dataset). Step 2.3 Data pre-processing and Information Fusion (NP and NDDS assay). Step 2.4 Objective and reference functions definition. Step 2.5 PTO calculation.

NDDs ChEMBL dataset.

Firstly, we collected the data of preclinical assay for NDDs from ChEMBL dataset (see step 2.1. in **Figure 1**).^[53-55] This dataset contained 4403 preclinical assays for 2566 NDDs (unique drugs) making approximately 1.71 assays for each drug. The information downloaded from ChEMBL included discrete variables c_{dj} used to specify the conditions/labels of each assay. These variables are c_{d0} = biological activity parameter, c_{d1} = target protein involved in NDs, c_{d2} = cell line for NDDs assays and c_{d3} = organism. Each one of these assays included one out of $n(c_{d0}) = 46$ possible biological activity parameters (EC_{50} , K_i (nM), etc.). They also involved some of the $n(c_{d1}) = 21$ target proteins, $n(c_{d2}) = 7$ cell lines (SH-SY5Y, CHO-K1, HEK293, PC-12, CHO, HEK-293T and HuT78), $n(c_{d3})=7$ organisms (*Homo sapiens*, *Rattus norvegicus*, *Mus musculus*, *Caviaporcellus*, *Canis lupus familiaris*, *Macacafas cicularis* and *Caenorhabditiselegans*). The information downloaded from ChEMBL also included another set of discrete variables used to codify the nature/quality of data. These variables are c_{d4} = type of target, c_{d5} = type of assay, c_{d6} = data curation, c_{d7} = confidence score and c_{d8} = target mapping. Specifically, the target types are $n(c_{d4}) = 6$ (single protein, organism, tissue, non-molecular and ADMET) and the assay types are $n(c_{d5}) = 3$ (Binding, Functional and ADMET). In addition, data curation has until $n(c_{d6}) = 3$ different values (auto-curation, expert and intermediate), confidences score $n(c_{d7}) = 4$ (9 = Direct single protein target assigned, 1 = Target assigned is non-molecular, 0 = Default value-Target assignment has yet to be curated and 8 = Homologous single protein target assigned) and target mapping $n(c_{d8}) = 3$ (protein, non-molecular and homologous protein). Furthermore, this database included the molecular descriptor $D_{dk} = [D_{d1}, D_{d2}, D_{d3}]$ in order to define the chemical structure of NDDS compound. Specifically, we used two types of molecular descriptor for the i^{th} compound: D_{d1} = Logarithm of the n-Octanol/Water Partition coefficient ($LOGP_i$) and D_{d2} = Topological Polar Surface Area (PSA_i). The detailed information of this dataset was released in Supporting Information (SI) file SI00.xlsx, datasheet ChEMBL.

NP cytotoxicity dataset

Simultaneously, we downloaded the data of preclinical assays for cytotoxicity NPs from different sources (see step 2.2. in **Figure 1**). Concretely, we selected 62 papers from scientific literature Pubmed and SciFinder. [56-58] This dataset included 260 preclinical assays for 31 unique NPs. Therefore, the number of essays for each NP is about 8.39. Moreover, the data covered a huge range of properties of NP such as morphology, physicochemical properties, coating agents, length, and time of assay. These properties were defined as discrete variables c_{nj} applied to identify the conditions/labels of each assay. Then, we enumerated all the particular conditions of each assay as a general vector $\mathbf{c}_{nj} = [c_{n1}, c_{n2}, c_{n3}, \dots, c_{nmax}]$. Precisely, these variables are c_{n0} = biological activity parameter, c_{n1} = cell line, c_{n2} = NP shape, c_{n3} = measurement condition and c_{n4} = coating agent. Each one of these assays involved at least one out of $n(c_{n0}) = 5$ possible biological activity parameters (CC_{50} , EC_{50} , IC_{50} , LC_{50} and TC_{50}). They also include $n(c_{n1}) = 53$ cell lines (A549 (H), RAW 264.7, Neuro-2A (M), etc.) and $n(c_{n2}) = 10$ NP shapes (spherical, irregular, slice-shaped, needle, rod, elliptical, pseudo-spherical, polyhedral, pyramidal and strip). In addition, they contain $n(c_{n3}) = 8$ NP measurement conditions (dry, H₂O, DMEM, RPMI, 1% Trion X-100/H₂O, H₂O/TMAOH, Egg/H₂O and H₂O/HMT) and $n(c_{n4}) = 16$ agents coating (UC, PEG-Si(OMe)₃, PVA, sodium citrate, 11-mercaptoundecanoic acid, PVP, propylammonium fragment, undecylazide fragment, CTAB, *N,N,N*-trimethyl-3(1-propene) ammonium fragment, potato starch, *N*-acetylcysteine, CMC-90, 2,3-dimercaptopropanesulfonate, 3-mercaptopropanesulfonate and thioglycolic acid). The full information of this dataset was shown in Supporting Information (SI) file SI00.xlsx, datasheet NP.

DNDS pairs resampling.

IF phase aimed at detected biological parameters.

Firstly, we described and acquired the objective value in order to design the IFPTML model for N2D3S. We defined the target function by applying the vectors of descriptors for all cases \mathbf{D}_k to use as the input variable in the ML model. The target function is commonly achieved by a mathematical conversion of the original theoretical or observed feature of the scheme under analysis. [59-61] In the recent IFPTML model, it includes two groups of observed values, specifically $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$. In addition, it contains two types of input vectors such as \mathbf{D}_{dk} and \mathbf{D}_{nk} for the preclinical assay NDDS and NP. Moreover, in this dataset was a large number of different biological parameters

c_{d0} and c_{n0} . For example, there are properties such as Half the maximum Inhibitory Concentration ($IC_{50}(nM)$), half the maximum Effective Concentration ($EC_{50}(nM)$) or the Lethal Concentration of a substance for an organism ($LC_{50}(nM)$) and so on. Another difficulty is that the majority of $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ values collected are numbers with decimals. Furthermore, in order to acquire the optimum N2D3S, we prioritize some properties and reduce others. In this context, we use the parameter desirability to tackle this problem.

The desirability value was established $d(c_{d0}) = 1$ or $d(c_{n0}) = 1$ if maximized the value of $v_{ij}(c_{d0})$ or $v_{nj}(c_{n0})$ was needed, otherwise $d(c_{d0}) = -1$ or $d(c_{n0}) = -1$. Due to the fact that the different NDDs and NP properties/characteristics possess large number of designations or labels c_{d0} and c_{n0} , increase the unreability of the data and make it more laborious to build up a regression model. For example, putting it into context with a specific case of biological activity parameters (c_{d0}) with $d(c_{d0}) = 1$ are B_{max} (fmol/mg)(the total number of receptors expressed in the same units), Activity (%), C_p (nM), *etc.* Whereas desirability parameters $d(c_{d0}) = -1$ that we want to minimize are for example $EC_{50}(nM)$, $IC_{50}(nM)$, I_{max} (%), *etc.* To address this problem, we used a cutoff value to divide AD or NP assays into favorable and non-favorable. It is worth mentioning that the usage of cutoff is a mainstream practice in drug discovery process. As a result, acquiring the final target function, the pre-process of all observed $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ values is crucial in order to remove or reduce imprecisions. Finally, the IF processing phase for these parameters $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ enable us to achieve a target function of the N2D3S.

On the other hand, we used the cutoff to rescale the parameters of $v_{ij}(c_{d0})$ and $v_{nj}(c_{n0})$ to achieve the Boolean (dummy) functions $f(v_{ij}(c_{d0}))_{obs}$ and $f(v_{nj}(c_{n0}))_{obs}$. These values were obtained as: $f(v_{ij}(c_{d0}))_{obs} = 1$, if $v_{ij}(c_{d0}) > cutoff$ and $d(c_{d0}) = 1$ or $v_{ij}(c_{d0}) < cutoff$ and desirability $d(c_{d0}) = -1$, $f(v_{ij}(c_{d0})) = 0$ otherwise. Similarly, $v_{nj}(c_{n0})$ was: $f(v_{nj}(c_{n0}))_{obs} = 1$ if $v_{nj}(c_{n0}) > cutoff$ and $d(c_{n0}) = 1$ or $v_{nj}(c_{n0}) < cutoff$ and $d(c_{n0}) = -1$, $f(v_{ij}(c_{d0}), v_{nj}(c_{n0})) = 0$ else. The values $f(v_{ij}(c_{d0}))_{obs} = 1$ and $f(v_{nj}(c_{n0}))_{obs} = 1$ means to have a positive desired effect of both NDDs and NP. As a result, the target function was described as $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs} = f(v_{ij}(c_{d0}))_{obs} \cdot f(v_{nj}(c_{n0}))_{obs}$. Therefore, the outcome of the IF-scaling $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs}$ is determined by the i^{th} NDDs compound, the n^{th} NP conditions, the rest of cases, $f(v_{ij}(c_{d0}), v_{nj}(c_{n0}))_{obs} = 0$, indicating that as minimum, one of the above-mentioned conditions fail.

Objective and reference functions definition

IF phase for combining the reference.

After we had the target function, the next step is to describe the input variables of the IFPTML model. As the input variables for this model is the reference function $f(v_{ij}(C_{d0}), v_{nj}(C_{n0}))_{ref}$. The $f(v_{ij}(C_{d0}), v_{nj}(C_{n0}))_{ref}$ plays an important role due to the fact that this function characterize the expected probability $f(v_{ij}(C_{d0}), v_{nj}(C_{n0}))_{ref} = p(f(v_{ij}(C_{d0}), v_{nj}(C_{n0}))_{ref} = 1)$ of achieving the interested level of activity for a specific property acquired from a well-known systems. IFPTML used as reference the value from well-known system or sub-set systems. Afterwards, this model includes the effect of different deviations (perturbations) of the query function from the reference function. Accordingly, $f(v_{ij}(C_{d0}), v_{nj}(C_{n0}))_{ref}$ can be considered a function related to observed (not predicted) outcomes. In the early section, we mentioned the step of IF-scaling taking into account to transform the original $v_{ij}(C_{d0})$ and $v_{nj}(C_{n0})$ values into $f(v_{ij}(C_{d0}))_{obs}$ and $f(v_{nj}(C_{n0}))_{obs}$ functions. When we acquire the $f(v_{ij}(C_{d0}))_{obs}$ and $f(v_{nj}(C_{n0}))_{obs}$ for all the cases in our dataset, the next steps is to quantify each of positive outcomes $n(f(v_{ij}(C_{d0})) = 1)$ and $n(f(v_{nj}(C_{n0})) = 1)$. Subsequently, in order to obtain the reference or expected functions, we divide the previous values by the entire number of cases for the NDDS and NP systems separately. We describe these functions as: $f(v_{ij}(C_{d0}))_{ref} = p(f(v_{ij}(C_{d0}))_{ref} = 1) = n(f(v_{ij}(C_{d0}))_{ref} = 1)/n(C_{n0})_j$ and $f(v_{nj}(C_{n0}))_{ref} = p(f(v_{nj}(C_{n0}))_{ref} = 1) = n(f(v_{nj}(C_{n0}))_{ref} = 1)/n(C_{n0})_j$. In this context, we can calculate the function of reference directly to recognize the probabilities product for both subsystems $f(v_{ij}(C_{d0}), v_{nj}(C_{n0}))_{ref} = p(f(v_{ij}(C_{d0}), v_{nj}(C_{n0}))_{ref} = 1) = p(f(v_{ij}(C_{d0}))_{ref} = 1) \cdot p(f(v_{nj}(C_{n0}))_{ref} = 1)$. It is worth mentioning that the usage of the function of reference at this point is another representation of the IF (combination) of NDDS and NP datasets.

PTO calculation

IFPTML N2D3S data analysis phases.

As we mentioned in the previous section, we acquired the results of many cytotoxicity preclinical assays of different NPs.[62-64] Complementarily, we obtained the data of preclinical assay for NDDs from ChEMBL database.[53-65-66] It included the calculation of the vectors D_{nk} and D_{dk} of structural descriptors for each NPs and NDDs. In addition, we constructed the vectors c_{nj} and c_{dj} in order to list each label and assay condition for all the pre-clinical assays of both the NPs and NDDs. Subsequently, we obtained the values $\Delta D_{dk}(c_{dj})$ and $\Delta D_{nk}(c_{nj})$ of the respective moving average deviation PTOs.

The NDDS vector lists each element: $\mathbf{D}_{dk} = [D_{d1}, D_{d2}]$. Precisely, these elements are the NDDS structural descriptors, which have allowed the development of various strategies to characterize and classify potential bioactive molecules structure, as in this work.[67] These structural descriptors are: $D_{d1} = \text{Logarithm of the n-Octanol/Water Partition coefficients (LOGP}_i)$ and $D_{d2} = \text{Topological Polar Surface Area (PSA}_i)$. On other hand, the cytotoxicity NP vector lists the elements as: $\mathbf{D}_{nk} = [D_{n1}, D_{n2}, D_{n3}, D_{n4}, D_{n5}, D_{n6}, D_{n7}, D_{n8}, D_{n9}, D_{n10}, D_{n11}, D_{n12}, D_{n13}, D_{n14}, D_{n15}, D_{n16}, D_{n17}, D_{n18}, D_{n19}, D_{n20}]$. Specifically, they are: $D_{n1} = \text{NMUn (number of monomer units)}$, $D_{n2} = \text{Lnp (NP length)}$, $D_{n3} = \text{Vnu (NP Volumen)}$, $D_{n4} = \text{Enu (NP Electronegativity)}$. They also contain $D_{n5} = \text{Pnu}$, $D_{n6} = \text{Uccoat (unsaturation count)}$, $D_{n7} = \text{Uicoat (unsaturation index)}$, $D_{n8} = \text{Hycoat (hydrophilic factor)}$, $D_{n9} = \text{AMR coat (Ghose–Crippen molar refractivity)}$, $D_{n10} = \text{TPSA(NO)coat (topological polar surface area using N,O polar contributions)}$, $D_{n11} = \text{TPSA(Tot)coat (topological polar surface area using N,O,S,P polar contributions)}$ and $D_{n12} = \text{ALOGPcoat}$. In this list also include $D_{n13} = \text{ALOGP2coat (squared Ghose–Crippen octanol/water partition coefficient (logP}^2)$, $D_{n14} = \text{SAtotcoat (total surface area from P_VSA-like descriptors)}$, $D_{n15} = \text{SAaccoat (surface area of acceptor atoms from P_VSA-like descriptors)}$, $D_{n16} = \text{SAdoncoat (surface area of donor atoms from P_VSA-like descriptors)}$, $D_{n17} = \text{Vxcoat (McGowan volume)}$, $D_{n18} = \text{VvdwMGcoat (van der Waals volume from McGowan volume)}$, $D_{n19} = \text{VvdwZAZcoat (van der Waals volume from Zhao–Abraham–Zissimos equation)}$ and $D_{n20} = \text{PDIcoat (packing density index)}$.

PT data preprocessing

Apart from the \mathbf{D}_{dk} and \mathbf{D}_{nk} vectors, the IFPTML study takes into account all vectors \mathbf{c}_{dj} and \mathbf{c}_{nj} as parts of the non-numerical experimental conditions and labels for both NDDS and NP preclinical assays as well. Later, we calculated the PTOs of the NDDS and NP preclinical assays including this extra information. We used (**Equation 1** and **Equation 2**) in order to obtain the NDDS and NP Moving Average (MA) PTOs. PT model begins with the expected value of a well-known activity and adds the effect of different perturbations/variations in the system. Consequently, the model includes two different of input variables: the reference or expected-value function $f(v_{ij})_{ref}$ and the PT operators $\Delta D_k(c_j)$. Specifically, they are applied for accounting NDDS and NP structural and assay information. In addition, the PTOs $\Delta D(D_{dk})$ and $\Delta D(D_{nk})$ label NDDS and NP

structural and/or physicochemical characteristics on the variables $\Delta D(D_{dk})$ and $\Delta D(D_{nk})$, correspondingly. Furthermore, the PTOs $\Delta D(D_{dk})$ and $\Delta D(D_{nk})$ classify NDDS and NP biological assay data with the variables $\langle D(D_{dk})_{c_{dj}} \rangle$ and $\langle D(D_{nk})_{c_{nj}} \rangle$, separately. The $\langle D(D_{dk}) \rangle$ and $\langle D(D_{nk}) \rangle$ are the representation the average operator for counting all cases with the equivalent subset of methodology conditions c_{dj} and c_{nj} , correspondingly. Accordingly, they ought to provide exact values for particular assay with minimum one altered element in methodology condition of the vector c_{dj} or c_{nj} . In this sense, they can specify which assay we are referring to.[48-52] Another kind of PTOs involved in this model is the NDDS-NP coat Moving Average Balance (MAB) PTO $\Delta\Delta D(D_{ca1}, D_{ca2}, D_{dk})$ (**Equation 3**). The MAB PTO takes into consideration the likenesses between the information of NDDS and the NP coating agent. Furthermore, PTOs centered straightly on MA and/or linear and non-linear conversions of MA have been applied for NDDS and NP development in the previous research work.[44-50-51] The MAS is the another way of expressing the combination between IF and PT cumulative procedure of NDDS and NP datasets.

$$\Delta D(D_{dk}) = \Delta D(D_{dk}) - \langle D(D_{dk})_{c_{dj}} \rangle \quad (1)$$

$$\Delta D(D_{nk}) = \Delta D(D_{nk}) - \langle D(D_{nk})_{c_{nj}} \rangle \quad (2)$$

$$\Delta\Delta D(D_{ca1}, D_{ca2}, D_{dk}) = \Delta D(D_{dk}) - [\Delta D(D_{ca1}) + \Delta D(D_{ca2})] \quad (3)$$

IF phase and proposal of training and validation series subsets

To develop model using the ML technique, the each of samples cases are designated as the training (subset = t) or validation (subset = v) series. The process of cases assignment ought to be haphazard, illustrative, and stratified.[68] Due to the nature of this combinatory system our sampling also have to take into account the IF-scaling procedure. Initially, we obtained the NDDS activity dataset from an open database ChEMBL, which has been arbitrarily abstracted from the primary published literature all over the world.[69] The cytotoxicity NP preclinical assays also were acquired randomly from journal articles. Afterwards, we prepared the each and every case as the following labels C_{d0} , C_{d1} , C_{d2} , C_{d3} , C_{d4} , C_{d5} , C_{d6} , C_{d7} , C_{d8} , C_{n0} , C_{n1} , C_{n2} , C_{n3} and C_{n4} . These cases were organized by ranking the labels alphabetically from A to Z (as we mentioned before, they are non-numeric variables in nature). The preference order of the labels on the procedure of ranking was $C_{d0} \Rightarrow C_{n0} \Rightarrow C_{d1} \Rightarrow C_{n1} \Rightarrow C_{d2} \Rightarrow C_{n2} \Rightarrow C_{d3} \Rightarrow C_{n3}$. In other words, we organized the cases firstly by C_{d0} , then by C_{n0} , and so forth.

This preference order considers the IF step by interchanging labels from AD and NP datasets. Afterwards, we assigned three of four equal parts cases to subset = t (training) and one-quarter subsets = v (validation) from all list. Subsequently, this random assignment improves the likelihood that nearly all the categories of individual label are denoted by subset = t and subset = v (stratified or proportional random sampling). In addition, this boost the possibility that practically all cases for each label are in a distribution of 3/4 in set = t and 1/4 in set = v, known as representative sampling. It is worth mentioning that the 75% and 25% proportion between training and validation is the most used one in the big data analysis. [68]

IFPTML-LDA model

IFPTML N2D3S model utilizes as input variables, the PTOs which are specified in the previous section to codify information of the putative N2D3S with its corresponding subsystems NDDS and NP. Combining objective function $f(v_{ij}, v_{nj})_{obs}$ and reference function $f(v_{ij}, v_{nj})_{ref}$, then by adding IF PTOs $\Delta\Delta D(D_{1c}, D_{2c}, D_{dk})$, we obtained the output function $f(v_{ij}, v_{nj})_{calc}$. This function carries out dataset cross-cut classification of NDDS and NP information. The generic equation for IFPTML linear model is the following (equation 4):

$$f(v_{ij}, v_{nj})_{calc} = a_0 + a_1 \cdot f(v_{ij}, v_{nj})_{ref} + \sum_{k=1, j=1}^{k=kmax, j=jmax} a_{k,j} \cdot \Delta D(D_{ki})_{cd_j} + \sum_{k=1, j=1}^{k=kmax, j=jmax} a_{k,j} \cdot \Delta D(D_{kn})_{cn_j} + \sum_{k=1, j=1}^{k=kmax, j=jmax} a_{k,j} \cdot \Delta\Delta D(D_{ki}, D_{kn})_{cd_j, cn_j} \quad (4)$$

Generalities for IFPTML models training and validation series

In many big data systems, the LDA model is the most used tool to seek the preliminary model due to the simplicity of this technique. In this sense, within this model we applied FSW process that can select automatically the most essential input variables for N2D3S in study. We obtained all the results by using the software STATISTICA 6.0.[68] Afterwards, we applied the Expert-Guided Selection (EGS) heuristic in order to retrain the LDA method using the most crucial parameters selected by FSW along with other missing aspects. All the IFPTML models sought were achieved by calculating the different statistical parameters, specifically Sensitivity (Sn), Specificity (Sp), Accuracy (Ac), Chi-square (χ^2), and the p -level.[70-71]

IFPTML-LDA vs. cross linear model

In the introduction section, we indicated the use of ML approaches as a promising strategy in order to tackle practical problems of Nanotechnology, such as reducing the number of experiments.[72-77] Specifically, in this paper the IFPTML method was used to combine NDDs with NP preclinical assays. In the recent year, Speck-Planche *et al.* described multiple IFPTML approaches of the toxicity activity and drug delivery of NPs with large number of species in a wide variety of experimental conditions. However, this study did not take into account the NDDs.[49-62-78] On the other hand, Nocado *et al.*, reviewed IFPTML method to explore the NDDs activity against numerous species and in different conditions of the assay, but this research they did not contemplate NP as part of the system.[79] Accordingly, in these models could not take into consideration both components (NDDs and NP) of the N2D3S system together. In our group, Dieguéz-Santana *et al.* for the first time, they applied successfully the IFPTML technique to study the combination of multiple antibacterial drug vs. cytotoxicity NP preclinical assay.[80] For this reason, in this paper, we used this new approach to develop a complex N2D3S systems, containing simultaneously both NDDs and NP components. In this system involved as we mentioned in the introduction, several NDDs assay, NP types alongside coating agents, NP morphology *etc.* To complete IF-scaling process, we calculated the objective function $f(v_{ij}, v_{nj})_{obs} = f(v_{ij})_{obs} \cdot f(v_{nj})_{obs}$. The main purpose of these functions is to increase the effect of certainty and maintain the homogeneity of scales. Once the PTOs were obtained, we applied the ML methods so as to fit this $f(v_{ij}, v_{nj})_{obs}$ function and achieve the IFPTML models. On the other hand, as indicated in the previous section, we classified NDDs preclinical assays \mathbf{c}_{dj} onto two different partitions (sub-sets) of variables \mathbf{c}_I and \mathbf{c}_{II} . The partition \mathbf{c}_I shows the biological characteristic, which contains c_{d0} = NDDs biological activity parameters (IC₅₀, K_i, Potency, Time *etc.*), c_{d1} = type of protein involves in NDs *etc.* However, the partition \mathbf{c}_{II} defines the data quality; which contains c_{d4} = type of target, c_{d5} = type of assay *etc.* For preclinical cytotoxicity NP essays \mathbf{c}_{nj} form only one partition \mathbf{c}_{III} , that describes its nature and involves multiple c_{n0} = NP biological activity parameters (CC₅₀, IC₅₀, LC₅₀, EC₅₀, *etc.*), c_{n1} = cell lines, c_{n2} = NP morphology and c_{n3} = NP synthesis conditions. In addition, we acquired two type of IFPTML-LDA models for designing the N2D3S systems. On the one side, we obtained the IFPTML-LDA by calculating the PTOs $\Delta D_k(\mathbf{c}_j)$ through the difference between the average value $\langle D_k(\mathbf{c}_j) \rangle$ and the partition \mathbf{c}_n within of their own set. As result, the best IFPTML-LDA model found is described below. (**Equation 5**)

$$f(v_{dij}, v_{nij})_{calc} = -4.46387 + 16.30655 \cdot f(v_{dij}, v_{nij})_{ref} + 0.00003 \cdot \Delta\text{DPSA}(\mathbf{c}_I)_{d_j} + 0.00450 \cdot \Delta\text{Dt}(\mathbf{c}_{III})_{n_j} + 0.00062 \cdot \Delta\text{DLnp}(\mathbf{c}_{III})_{n_j} + 0.00675 \cdot \Delta\text{DVnpu}(\mathbf{c}_{III})_{n_j} + 0.00431 \cdot \Delta\text{DVxcoat}(\mathbf{c}_{III})_{n_j} - 0.00537 \cdot \Delta\text{DVvdwMGcoat}(\mathbf{c}_{III})_{n_j} \quad (5)$$

$$N_{\text{train}} = 375000 \quad \chi^2 = 24273.63 \quad p\text{-level} < 0.05$$

On the other side, we tested the possibility to improve the results of statistical parameters for IFPTML-LDA algorithm, for this we calculated the PTOs $\Delta D_k(\mathbf{c}_j)$ by performing all the possible combinations among the average value $\langle D_k(\mathbf{c}_j) \rangle$ of both vectors \mathbf{D}_{nk} and \mathbf{D}_{dk} with each partition. As a result, we obtained 3 different combinations of crossing PTOs for each sample, one for NDDS ($\Delta D_{dk}(\mathbf{c}_{III})$) and two for NP ($\Delta D_{npk}(\mathbf{c}_I)$ and $\Delta D_{npk}(\mathbf{c}_{II})$). For simplicity, they are under the name of IFPTML-LDA with cross (see more details in **Figure 1**). The best IFPTML-LDA sought with cross model is the following. (**Equation 6**)

$$f(v_{dij}, v_{nij})_{calc} = -4.44505 + 14.28457 \cdot f(v_{dij}, v_{nij})_{ref} + 0.00216 \cdot \Delta\text{DPSA}(\mathbf{c}_I)_{cd_j} + 0.00241 \cdot \Delta\text{Dt}(\mathbf{c}_{III})_{cn_j} + 0.01201 \cdot \Delta\text{DLnp}(\mathbf{c}_{III})_{cn_j} + 0.16549 \cdot \Delta\text{DVnpu}(\mathbf{c}_{III})_{cn_j} - 0.02389 \cdot \Delta\text{DVxcoat}(\mathbf{c}_{III})_{cn_j} + 0.04902 \cdot \Delta\text{DVvdwMGcoat}(\mathbf{c}_{III})_{cn_j} + 2.040821 \cdot \Delta\text{DEnpu}(\mathbf{c}_{II})_{c_{dn}} + 0.03229 \cdot \Delta\text{DAMRcoat}(\mathbf{c}_{II})_{c_{dn}} \quad (6)$$

$$N_{\text{train}} = 375000 \quad \chi^2 = 43587.01 \quad p\text{-level} < 0.05$$

The output function is $f(v_{dij}, v_{nij})_{calc}$ which provide a real numeric value that will probably be applied to counting N2D3S systems. This function was acquired by calculating the objective function $f(v_{ij}(C_{d0}), v_{nj}(C_{n0}))_{obs}$ with the ML method making use of the PTOs. The characteristic of IFPTML models was defined by the statistical parameters such as Sensibility (Sn), Specificity (Sp), Accuracy (Ac), Chi-square test (χ^2), and the p -level.[68] These statistical parameters for each sample (standard IFPTML-LDA and with cross IFPTML-LDA) were collected in **Table 1**. The statistical parameters obtained for both methods were in the accurate range described in the scientific report for the classification model of ML algorithm.[70-71] In the standard IFPTML-LDA contains all the indispensable variables for defining the NDDS structure and the most significant parameters for NP such as morphology, size, assay conditions along with others. Nevertheless, with cross IFPTML-LDA system not only we include all the essential variables but also two crossing PTOs. These new PTOs are chosen by the FSW method, which can select the most influential variable in the system in study.

Table 1: IFPTML-LDA N2D3S model results summary.

| Data | Stat. | Param. | Without cross | Param. | Cross |
|------|-------|--------|---------------|--------|-------|
|------|-------|--------|---------------|--------|-------|

| Sample | Set | Sub-set | Param. | (%) | Sub Set Predicted | | (%) | Sub Set Predicted | |
|--------|-----|---------|--------|------|-------------------|--------|------|-------------------|-------|
| | | | | | 0 | 1 | | 0 | 1 |
| 1 | t | 0 | Sp | 73 | 255190 | 94292 | 72.2 | 252534 | 97042 |
| | | 1 | Sn | 71 | 7398 | 18120 | 74.4 | 6517 | 18907 |
| | v | 0 | Sp | 73.3 | 85369 | 31125 | 72.3 | 84183 | 32315 |
| | | 1 | Sn | 70.3 | 2522 | 5984 | 73.9 | 2218 | 6284 |
| 2 | t | 0 | Sp | 70 | 244548 | 105076 | 79.5 | 277907 | 71717 |
| | | 1 | Sn | 62.1 | 9528 | 15848 | 70.1 | 7584 | 17792 |
| | v | 0 | Sp | 70 | 81640 | 35009 | 79.7 | 92929 | 23720 |
| | | 1 | Sn | 63.1 | 3081 | 5270 | 70.7 | 2451 | 5900 |
| 3 | t | 0 | Sp | 70.6 | 246551 | 102809 | 79.6 | 277921 | 71439 |
| | | 1 | Sn | 62.3 | 11616 | 15974 | 70.1 | 7668 | 17972 |
| | v | 0 | Sp | 70.7 | 82370 | 34174 | 79.6 | 92726 | 23818 |
| | | 1 | Sn | 62.7 | 3828 | 5300 | 70.4 | 2500 | 5956 |
| Avg. | t | 0 | Sp | 71.2 | 248763 | 100726 | 77.1 | 269454 | 80066 |
| | | 1 | Sn | 65.1 | 9514 | 16647 | 71.5 | 7256 | 18224 |
| | v | 0 | Sp | 71.3 | 83126 | 33436 | 77.2 | 89946 | 26618 |
| | | 1 | Sn | 65.4 | 3144 | 5518 | 71.7 | 2390 | 6047 |

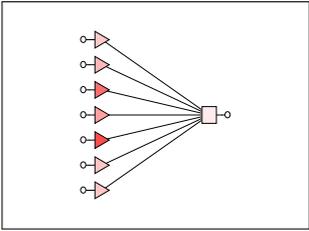
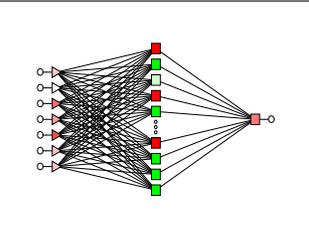
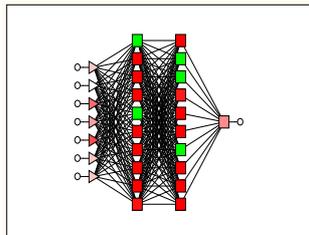
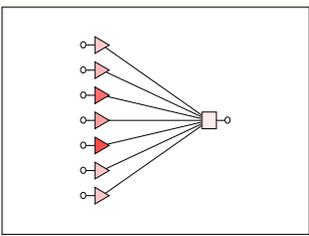
The result summary collected in **Table 1** contains the statistical parameters for IFPTML-LDA without and with cross for the best models found (equation 2). The IFPTML-LDA presented in this paper, had Sn and Sp \approx 70 – 73% values in both training and validation series. On the other hand, the IFPTML-LDA with cross showed significantly higher values of Sn and Sp \approx 70 – 80% in both series. In addition, by only adding two PTOs onto the standard model, the IFPTML-LDA could improve almost 7% of Sp value in training/validation series. However, the Sp and Sn values of with cross model are slightly unbalanced in comparison with the standard one but the Sp and Sn values remain approximately constant within the same training and validation series.

IFPTML linear vs. non-linear models

In order to obtain the ANN model, we used the same PTOs variables as in the LDA model. Furthermore, as an alternative of non-linear model we created the ANN by using the same software STATISTICA. The ANN can also be used as a new strategy to confirm and validate the linear hypothesis. Both are comparable due to the fact that the Linear Neural Networks (LNN) technique are analogous to LDA models and they are linear equations. Accordingly, the IFPTML-LNN model is a useful tool to assess the degree of strength of the linear relationship between PTOs and the N2D3S

objective function. IFPTML-LNN models showed in this work, presented remarkably lower Sn and Sp \approx 64 – 65% values in the training and validation series if we compare with the IFPTML-LDA models, see details in **Table 2**.

Table 2: The best result of IFPTML-ANN N2D3S systems models found.

| Sample | IFPTML-ANN Models ^a | Sub set | Stat . | Val. (%) | $f(V_{ij}(C_{d0}), V_{nj}(C_{n0}))$ | Observed | | AUROC |
|--|---|---------|--------|----------|-------------------------------------|----------|--------|-------|
| | | | | | Pred. | 1 | 0 | |
| 01 | LDA 7:7-1:1  FSTW + EGS | t | Sp | 0 | 73.0 | 94272 | 255178 | - |
| | | | Sn | 1 | 71.0 | 18057 | 7367 | |
| | | v | Sp | 0 | 73.3 | 31125 | 85319 | - |
| | | | Sn | 1 | 70.3 | 5980 | 2522 | |
| | MLP 7:7-11-1:1  BP96b | t | Sp | 0 | 86.1 | 300836 | 48740 | 0.943 |
| | | | Sn | 1 | 85.8 | 3610 | 2181 | |
| | | v | Sp | 0 | 86.1 | 100278 | 16220 | 0.934 |
| | | | Sn | 1 | 86.2 | 1173 | 7329 | |
| | DLN 7:7-10-10-1:1  BP100,CG20b | t | Sp | 0 | 85.8 | 299942 | 49634 | 0.945 |
| | | | Sn | 1 | 85.8 | 3621 | 21803 | |
| | | v | Sp | 0 | 85.9 | 100103 | 16395 | 0.933 |
| | | | Sn | 1 | 86.3 | 1168 | 7334 | |
| LNN 7:7-1:1  PI | t | Sp | 0 | 65.0 | 227184 | 122392 | 0.744 | |
| | | Sn | 1 | 64.7 | 8971 | 16453 | | |
| | v | Sp | 0 | 65.1 | 75788 | 40710 | 0.733 | |
| | | Sn | 1 | 64.1 | 3055 | 5447 | | |

Analogous to the IFPTML-LDA model, the values of statistical parameter (Sp and Sn) are considerably balanced and stay steady comparing training vs. validation series. On the other hand, we obtained two types of the non-linear models, the Multi-Layer Perceptron (MLP) and the Depth Learning Network (DLN). The MLP is made up by seven PTOs as input layer, a hidden layer with eleven neurons and an output layer. The most notably different is that the DLN involves two hidden layers and each one with ten neurons. Both MLP and DLN showed high Sp and Sn \approx 85 – 86% values in training and validation series. If we compare the IFPTML-ANN linear with non-linear models basing on the results of statistical parameters, we can confirm that N2D3S is a non-linear system. Another result obtained in the development of ANN is the Area Under Receiver Operating Characteristic (AUROC), see **Figure 2**. [68] The AUROC curve values are 0.93-0.94 for both MLP and DLN models in training and validation series. Precisely, AUROC values of non-linear models are remarkably deflected from random (RND) curve with AUROC = 0.5. [68]

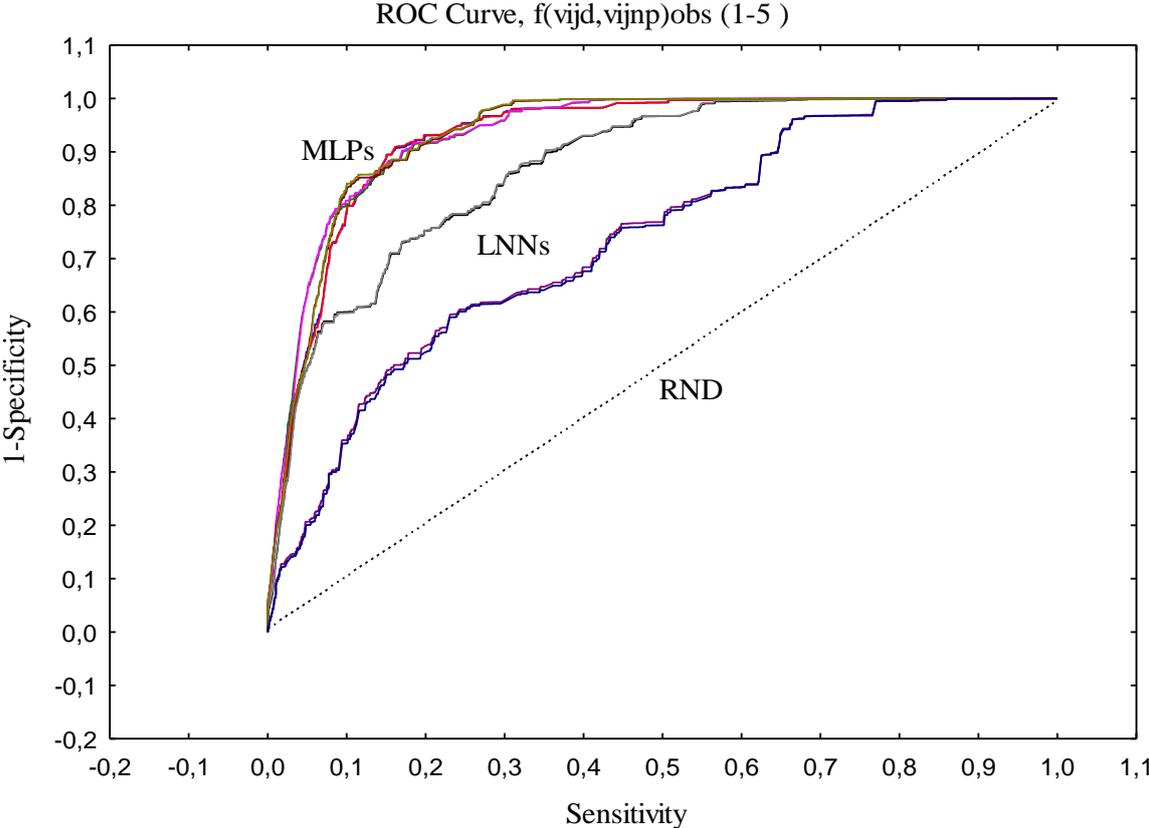


Figure 2: AUROC exploration of IFPTML-MLP and IFPTML-LNN models

IFPTML models robustness analysis.

The design of the N2D3S involve the combination of a large number of data, both NDDS and NP preclinical assays. Due to the nature of this big data system, we divided

the information fusion dataset onto 3 samples. In the previous section, we discussed the best model obtained for IFPTML-LDA, IFPTML-LDA with cross and IFPTML-ANN. Therefore, in this part for the 3 samples the robustness analysis is mentioned, see **Table 3**. In general sense, the number of cases (n) used in training and validation series for all models presented the lowest Standard DeViations (SDV), which indicated that most of the data in a sample tend to be clustered near its mean.[81] On the other hand, the high value of SDV such as DLN model specified that the data was distributed over a wide range of values. In addition, all models presented similar SDV values in the same training and validation series. Interestingly, the LDA model showed significantly lower value of SDV for (>1) Sp if we compare with (>4) Sn in training/validation series. However, the SDV values for LDA Cross model were contrary to LDA, with inferior SDV values for Sn and higher for Sp. It is worth mentioning that both MLP 1 and LNN models obtained statistical parameters close to its mean, in other words these models are robust. Furthermore, In IFPTML-ANN technique we obtained as results the AUROC values, after doing the robustness analysis we could confirmed that each ANN model the AUROC values are robust. In addition, the AUROC graphic (**Figure 2**) might be used to prove this evidence, due to the similarity between the curves shape.

Table 3: Result summary of N2D3S alongside average of 3 samples and standard deviation.

| | Model | t | | | v | | | AUROC (t/v) |
|-----|-----------|-------|-------|--------|-------|-------|--------|----------------|
| | | Sp | Sn | n | Sp | Sn | n | |
| AVG | LDA | 71.2 | 65.1 | 375000 | 71.3 | 65.4 | 125000 | - |
| | LDA Cross | 77.1 | 71.5 | 375000 | 77.2 | 71.7 | 125000 | - |
| | MPL 1 | 85.1 | 85.0 | 375000 | 85.1 | 85.1 | 125000 | 0.937/0.925 |
| | DNL | 79.2 | 79.0 | 375000 | 79.2 | 79.3 | 125000 | 0.893/0.879 |
| | LNN | 65.0 | 64.9 | 375000 | 65.1 | 64.9 | 125000 | 0.748/0.737 |
| SDV | Model | t | | | v | | | AUROC (t/v) |
| | | Sp | Sn | n | Sp | Sn | n | |
| | LDA | 1.587 | 5.082 | 0 | 1.739 | 4.277 | 0 | - |

| | | | | | | | |
|-----------|-------|-------|---|-------|-------|---|-------------|
| LDA Cross | 4.244 | 2.483 | 0 | 4.244 | 1.940 | 0 | - |
| MLP 1 | 1.266 | 1.217 | 0 | 1.940 | 1.102 | 0 | 0.010/0.010 |
| DLN | 8.489 | 8.568 | 0 | 8.584 | 8.727 | 0 | 0.069/0.071 |
| LNN | 0.100 | 0.153 | 0 | 0.153 | 0 | 0 | 0.005/0.003 |

The results reveal the strength of the linear hypothesis, nevertheless, the statistical parameters of this linear model obtained are not satisfactorily at all. As a result, in the IFPTML-LDA cross we enlarged the number of input variables from seven to nine. In this analysis, we did not obtain substantial change. Therefore, we tested more complex methods specifically, non-linear models so as to improve the Sp and Sn values. The IFPTML-MLP 7:7-11-1:1 model contains seven input variables in the input layer and eleven neurons in the hidden layer obtained the best statistical parameters of Sn and Sp values. Please, see the results of Sn and Sp values in **Table 3**. For more details, see results for every case in Supporting Information file SI00.xlsx. The IFPTML-DLN model which involves two hidden layers illustrate similar result as IFPTML-MLP 7:7-11-1:1.

Taking into account all the aforementioned results, we can consider both IFPTML-MLP and IFPTML-DLN as the best models with remarkably higher values of Sp and Sn \approx 85 – 86% and the AUROC values of 0.93-0.94. However, the more complex nature of DLN model and the non-significant improvement of statistical parameters in comparison with the MLP model, therefore we can confirm that the N2D3S requires the MLP model. In addition, this selection is reinforced by the principle of parsimony, prioritizing the simplest explanations among all the possible ones.[82] In **Table 4**, IFPTML-ANN model input variable sensitivity analysis for NDDS&NP, NDDS and NP subsystems were showed. The IFPMTL-LNN models involve almost all the significant parameters according to the EGS criteria. The majority of parameters provide a substantial influence on the Sensitivity \geq 1.[68] In many cases, the value of sensitivity analysis is slightly higher with a Sensitivity approximately 1.00 – 1.08. Nevertheless, the EGS perspective fails in the selection of Δ DPSA(**c**_I) and Δ Dt(**c**_{III}) variables. In this sense, the IFPTML-ANN model suggests that those variables do not affect in any model. On the other hand, the IFPTML-LNN obtains the lowest value of sensitivity

$\approx 1.00-1.13$ which would strengthen the need of a complex model in N2D3S. The DLN model involves the essential variables in accordance with the EGS proposal, however, they have remarkably higher sensitivity values with approx. 0.96 – 2.03. The MLP obtains the highest values of Sensitivity, between 1.13 and 2.57.

Table 4: IFPTML-ANN model input variable sensitivity analysis for different subsystems with their corresponding variables.

| Sub-systems | Variables | LNN | | MLP | | DLN | | | | | |
|-------------|---|------|------|------|------|------|------|------|------|------|------|
| | | t | v | t | v | t | v | t | v | t | v |
| NDDs&NP | $f(C_{d0}, C_{n0})_{ref}$ | 1.02 | 1.02 | 1.32 | 1.33 | 1.46 | 1.45 | 1.25 | 1.24 | 1.38 | 1.40 |
| NDDs | $\Delta DPSA(\mathbf{c}_i)$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| NP | $\Delta Dt(\mathbf{c}_{III})$ | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 | 0 |
| | $\Delta DL_{np}(\mathbf{c}_{III})$ | 1.00 | 1.00 | 1.14 | 1.13 | 1.08 | 1.08 | 1.08 | 1.08 | 1.60 | 1.59 |
| | $\Delta DV_{npu}(\mathbf{c}_{III})$ | 1.00 | 1.00 | 2.22 | 2.22 | 0.92 | 0.92 | 1.06 | 1.05 | 1.24 | 1.25 |
| | $\Delta DV_{xcoat}(\mathbf{c}_{III})$ | 1.00 | 1.00 | 1.96 | 1.98 | 1.45 | 1.47 | 1.45 | 1.48 | 1.99 | 2.03 |
| | $\Delta DV_{vdwMGcoat}(\mathbf{c}_{III})$ | 1.13 | 1.13 | 2.57 | 2.54 | 1.44 | 1.43 | 1.24 | 1.24 | 1.91 | 1.90 |

IFPTML-LDA for N2D3S simulation experiment.

In this section, we employed the IFPTML-LDA technique in order to calculate the probability values for some selected cases of N2D3S system. The linear model was chosen for its simplicity and the slight improvement of the non-linear model. The value of probability $p(N2D3S_{in})_{c_{dj}, c_{nj}}$ was obtained with N2D3S system created by the combination of i^{th} AD_i and the n^{th} NP_n which is highly likely to have a desired level of biological activity under both assay conditions \mathbf{c}_{dj} and \mathbf{c}_{nj} . This simulation experiment involved in total $N_{N2D3S} = 88$ systems-cell line vs. a total of $N_{NDDs} = 123$ drugs. Many of these drugs are NDDs with some known anti-neurodegenerative activity, generally for Alzheimer and Parkinson diseases. Some of these NDDs are approved by the FDA (Food and Drug Administration) agency while other have been showed to be active in several assays. In addition, it also contained cytotoxicity assays against multiple cell lines, the type of NP, its coat and the time of each assay. In this context, we performed a total $N_{tot} = N_{NDDs} \cdot N_{NP} = 22 \cdot 218 = 4796$ values of probability which were be able to predict successfully the putative N2D3S.

Figure 3 depicts the results in a 3-colour scale according to the value of probability: the green section indicates high probability (0.61-0.98), yellow low to middle probability (0.17-0.60) and red very low predicted probability (< 0.17). Those assays never reported before or very low represented in the original dataset or the combination between NDDS and NP are meaningless were illustrated in white color to avoid overestimation of results. The result of IFPTML-LDA model pointed out some of N2D3S systems as a promising combination for future additional assays. The resulting N2D3S systems shown in **Figure 3** involve twenty different NDDs. The first ten is 1= Clozapine, 2= Galantamine, 3 =Levodopa, 4=Apomorphine, 5=Fiduxosin, 6=Beagacestat, 7=Memoquin, 8=Mesodihydroguairetic Acid, 9 = Tarenflubil and 10 = Huperzine A. The other 10 NDDs are 11 = Guanidinonaltrindole, 12 = Semagacestat, 13 = Huprine X, 14= Carproctamide, 15= Tacrine, 16 = Tramiprosate, 17 = Preladenant, 18 = Piracetam, 19 = Istradefylline and 20 = Rivastigmine. These systems have the following coating agents: PEG = Polyethylene glycol, PVP = Polyvinylpyrrolidone, PPF = Propylammonium fragment, and UAF =Undecylazide fragment. The symbol UC = Uncoated represent non-coated N2D3S system. For more details, please see on Supporting Information file S001.xls. Interestingly, the high value of prediction involves PEG-Si(OMe)₃ as NP coat with the $p(N2D3S_{in})_{cdj.cnj} = 0.80-0.99$ for the majority of NDDs. Another important factor is the type of NP that may affect the value of probability. It appears that metal oxide compounds such as SiO₂ and TiO₂ along with PEG-Si(OMe)₃NP coated for almost all NDDs are likely to be promising for further assays. Nevertheless, double metal oxide compounds such as CoFe₂O₄ and ZnFe₂O₄ obtained an intermediate probability value $p(N2D3S_{in})_{cdj.cnj} = 0.17-0.70$ against TK6 (H) and WISH (H), respectively. In the general sense, the least propitious combination is the metal NP with all NDDs which gives low value of probability ($p=0.02-0.35$). It is worth mentioning that all predictions carried out by this method should be used with caution and required experimental corroboration. The potential utility of the IFPTML method is to speed up the experimental study and provides inexpensive preliminary results for a large database of N2D3S systems. This approach offers an efficient and powerful tool to direct the experimental research as an alternative of tedious trial and error tests.

| Syst. | NP | Coat | Cell line (Org.) | Time (h) | Avg | Drugs |
|-------|----------------------------------|--------------------------|------------------|----------|------|-------|
| 1 | Ag | PVP | HaCaT (H) | 168 | 0.10 | |
| 2 | Ag | PVP | HaCaT (H) | 72 | 0.08 | |
| 3 | Ag | PVP | HaCaT (H) | 48 | 0.07 | |
| 4 | Ag | CIT | RBE4 (R) | 24 | 0.26 | |
| 5 | Ag | UC | NR8383 (R) | 24 | 0.12 | |
| 6 | Al | UC | BRL 3A (R) | 24 | 0.11 | |
| 7 | Al ₂ O ₃ | UC | BJ (H) | 24 | 0.40 | |
| 8 | Al ₂ O ₃ | UC | L929 (M) | 24 | 0.39 | |
| 9 | Au | 11-MCUA | HepG2 (H) | 48 | 0.72 | |
| 10 | Au | 11-MCUA | HepG2 (H) | 72 | 0.71 | |
| 11 | Au | 11-MCUA | HepG2 (H) | 24 | 0.70 | |
| 12 | Au | CTAB | MDCK II (D) | 24 | 0.87 | |
| 13 | Au | CTAB | MDCK II (D) | 48 | 0.87 | |
| 14 | Au | CIT | HepG2 (H) | 72 | 0.58 | |
| 15 | Au | CIT | HepG2 (H) | 48 | 0.54 | |
| 16 | Au | CIT | HepG2 (H) | 24 | 0.54 | |
| 17 | Au | CIT | 3T3 (M) | 72 | 0.51 | |
| 18 | Au | CIT | 3T3 (M) | 24 | 0.49 | |
| 19 | Au | UC | PK-15 (P) | 72 | 0.33 | |
| 20 | Au | UC | 3T3 (M) | 72 | 0.33 | |
| 21 | Au | UC | Vero (Mon) | 72 | 0.30 | |
| 22 | Co | UC | 3T3 (M) | 72 | 0.12 | |
| 23 | Co | UC | 3T3 (M) | 24 | 0.11 | |
| 24 | Co | UC | 3T3 (M) | 4 | 0.10 | |
| 25 | CoFe ₂ O ₄ | PVA | L929 (M) | 48 | 0.50 | |
| 26 | CoFe ₂ O ₄ | UC | MDCK (H) | 72 | 0.43 | |
| 27 | CoFe ₂ O ₄ | UC | A549 (H) | 72 | 0.42 | |
| 28 | CoFe ₂ O ₄ | UC | HepG2 (H) | 72 | 0.42 | |
| 29 | CoFe ₂ O ₄ | UC | NCIH441 (H) | 72 | 0.42 | |
| 30 | CoFe ₂ O ₄ | UC | CaCo-2 (H) | 24 | 0.41 | |
| 31 | CoFe ₂ O ₄ | UC | TK6 (H) | 72 | 0.41 | |
| 32 | CoFe ₂ O ₄ | UC | CaCo-2 (H) | 72 | 0.40 | |
| 33 | CoFe ₂ O ₄ | UC | A549 (H) | 24 | 0.40 | |
| 34 | CoFe ₂ O ₄ | UC | L929 (M) | 48 | 0.39 | |
| 35 | CoFe ₂ O ₄ | UC | MDCK (H) | 24 | 0.39 | |
| 36 | CoFe ₂ O ₄ | UC | TK6 (H) | 24 | 0.37 | |
| 37 | CoFe ₂ O ₄ | UC | NCIH441 (H) | 24 | 0.37 | |
| 38 | Cr ₂ O ₃ | UC | A549 (H) | 24 | 0.50 | |
| 39 | Cu | UC | H4IIE (R) | 24 | 0.12 | |
| 40 | Cu | UC | HepG2 (H) | 24 | 0.12 | |
| 41 | CuO | UC | A549 (H) | 24 | 0.42 | |
| 42 | Fe ₂ O ₃ | UC | HepG2 (H) | 24 | 0.55 | |
| 43 | Fe ₂ O ₃ | UC | A549 (H) | 24 | 0.51 | |
| 44 | Fe ₃ O ₄ | UC | BRL 3A (R) | 24 | 0.34 | |
| 45 | Mn ₂ O ₃ | UC | A549 (H) | 24 | 0.52 | |
| 46 | MoO ₃ | UC | BRL 3A (R) | 24 | 0.54 | |
| 47 | Ni | UC | A549 (H) | 48 | 0.11 | |
| 48 | Ni | UC | A549 (H) | 24 | 0.10 | |
| 49 | NiO | UC | HepG2 (H) | 24 | 0.42 | |
| 50 | NiO | UC | A549 (H) | 24 | 0.32 | |
| 51 | Si | PAF | NR8383 (R) | 24 | 0.14 | |
| 52 | Si | PAF | CaCo-2 (H) | 24 | 0.13 | |
| 53 | Si | UDAF | CaCo-2 (H) | 24 | 0.07 | |
| 54 | Si | UDAF | NR8383 (R) | 24 | 0.07 | |
| 55 | SiO ₂ | PEG-Si(OMe) ₃ | HUVECs | 72 | 0.96 | |
| 56 | SiO ₂ | PEG-Si(OMe) ₃ | NCIH441 (H) | 72 | 0.95 | |
| 57 | SiO ₂ | PEG-Si(OMe) ₃ | BMSC (M) | 72 | 0.95 | |
| 58 | SiO ₂ | PEG-Si(OMe) ₃ | HEK293 (H) | 72 | 0.95 | |
| 59 | SiO ₂ | PEG-Si(OMe) ₃ | RAW 264.7 (M) | 72 | 0.95 | |
| 60 | SiO ₂ | PEG-Si(OMe) ₃ | BMSC (H) | 72 | 0.95 | |
| 61 | SiO ₂ | PEG-Si(OMe) ₃ | HepG2 (H) | 72 | 0.92 | |
| 62 | SiO ₂ | PEG-Si(OMe) ₃ | A549 (H) | 72 | 0.90 | |
| 63 | SiO ₂ | UC | HEK293 (H) | 72 | 0.78 | |
| 64 | SiO ₂ | UC | HUVECs | 72 | 0.78 | |
| 65 | SiO ₂ | UC | BMSC (M) | 72 | 0.77 | |
| 66 | SiO ₂ | UC | RAW 264.7 (M) | 72 | 0.76 | |
| 67 | SiO ₂ | UC | HaCaT (H) | 4 | 0.75 | |
| 68 | SiO ₂ | UC | BMSC (H) | 72 | 0.75 | |
| 69 | SiO ₂ | UC | NCIH441 (H) | 72 | 0.74 | |
| 70 | SiO ₂ | UC | HepG2 (H) | 72 | 0.68 | |
| 71 | SiO ₂ | UC | A549 (H) | 72 | 0.54 | |
| 72 | SiO ₂ | UC | A549 (H) | 72 | 0.54 | |
| 73 | SiO ₂ | UC | A549 (H) | 48 | 0.50 | |
| 74 | SiO ₂ | UC | HaCaT (H) | 24 | 0.49 | |
| 75 | SiO ₂ | UC | 3T3 (M) | 72 | 0.49 | |
| 76 | SiO ₂ | UC | A549 (H) | 24 | 0.47 | |
| 77 | TiO ₂ | UC | Neuro-2A (M) | 48 | 0.44 | |
| 78 | TiO ₂ | UC | A549 (H) | 24 | 0.43 | |
| 79 | Y ₂ O ₃ | UC | HEK293 (H) | 24 | 0.31 | |
| 80 | ZnFe ₂ O ₄ | UC | WISH (H) | 72 | 0.44 | |
| 81 | ZnFe ₂ O ₄ | UC | WISH (H) | 48 | 0.40 | |
| 82 | ZnFe ₂ O ₄ | UC | WISH (H) | 24 | 0.38 | |
| 83 | ZnO | UC | A549 (H) | 24 | 0.31 | |
| 84 | ZnO | UC | HeLa (H) | 24 | 0.31 | |
| 85 | ZnO | UC | HepG2 (H) | 24 | 0.30 | |
| 86 | ZnO | UC | HUVECs | 24 | 0.30 | |

Figure 3: IFPTML-LDA N2D3S systems experiment simulation.

In addition, the determination of the probability value distribution in a generic sense for the unique pairs of NP cytotoxicity assays and NDDs were carried out. For this, we depict the surface scatterplot of probability values against Histograms of NP length along with NDDs Hydrophobicity (see **Figure 4**). Generally, a third part of probability values remain in the dark green zone, which represents the promising N2D3S system for further assay. It is worth mentioning that most of the cases (white dots) are hydrophobic drugs (on the left of the graph). This feature is one of the most important physicochemical property for drug in order to cross the BBB.[83] High lipophilicity can contribute to excessive volumes of distribution, increased metabolic liability, and lower unbound drug concentration in the plasma and/or brain and may negatively affect pharmaceuticals properties, particularly solubility.[84] In this sense, most NDDs of this database are in the PSAdi range between 60-120 Å². The previous research work, Stephen *et al.* suggested that Central Neural System (CNS) drug should have a PSA value < 90Å² for a decent BBB permeability among other physicochemical characteristics such as number of hydrogen bond donors, molecular size and shape with lesser contributions from hydrogen bond acceptors.[83] Although this type of graphic is clearly a simplification of the whole database, it offers beneficial simple guidelines for the researcher concerned with designing NDDs compounds or libraries with improved probability of CNS penetration. On the other hand, the size of the vast majority of NPs for NDDs delivery in this database is in the range between 70-115 nm. Recently, Chithrani *et al.*[85] have demonstrated that the size, coating and surface charge of nanoparticles have a crucial impact on the intracellular uptake process. Similarly, Shilo *et al.* have investigated the influence of NP size on the chance to cross the BBB by using the endothelial brain cell method. Results indicated that intracellular uptake of NPs is strongly dependent on NP size. This feature affects directly the biomedical application. When NPs act as drug delivery carriers by drug encapsulation into the NPs, the highest NP size is required (70 nm). Nevertheless, when NPs act as drug delivery carriers by binding the drug molecules to NPs surface, the highest free surface area is required, therefore the appropriate size would be 20 nm.[86] This principle suggests that a high number of the NP of our database are proper drug delivery carriers by drug encapsulation.

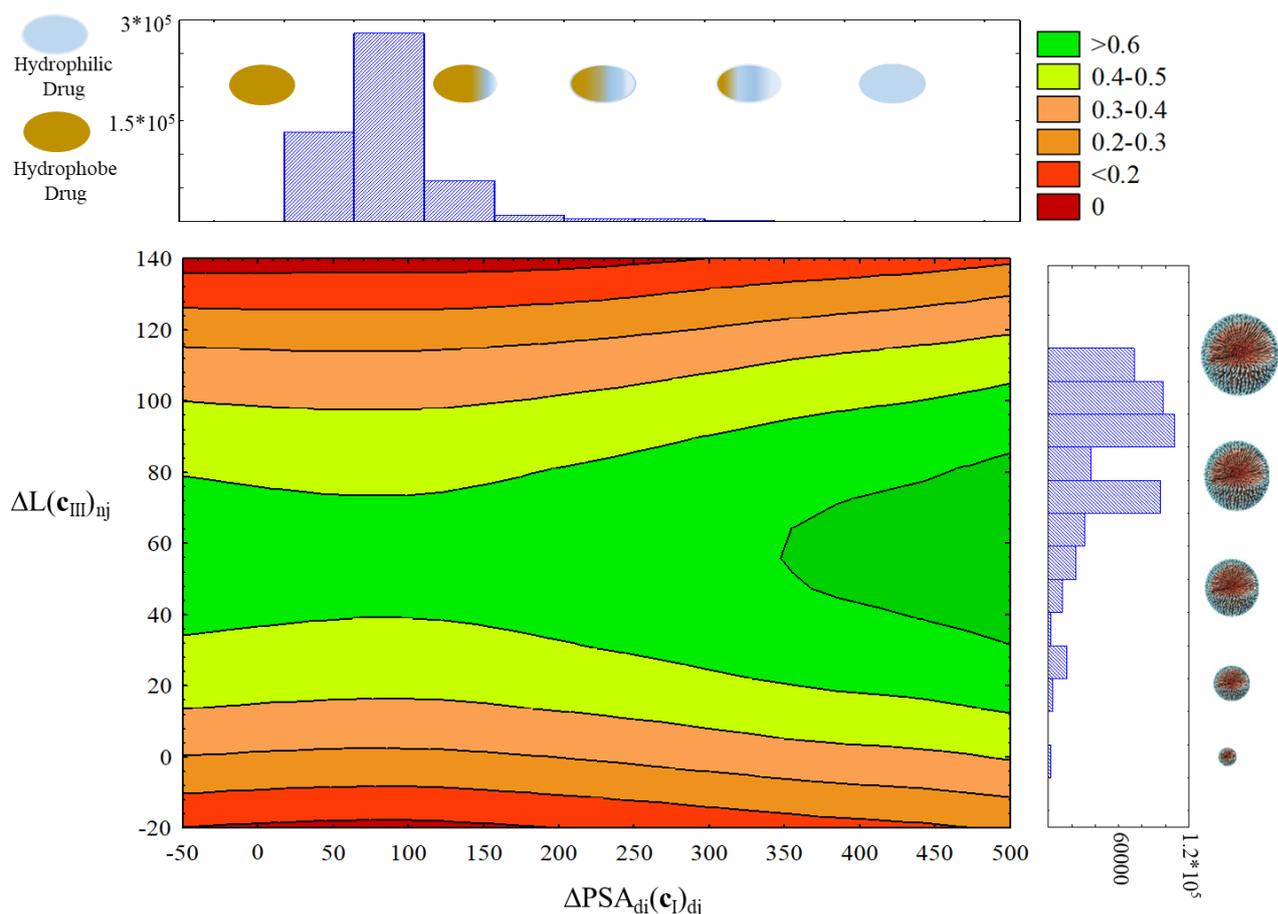


Figure 4: Probability surface scatter plot.

The design of the new N2D3S system containing multiple preclinical assay of cytotoxicity NP and NDDs has been carried out successfully. This database involves a high structural and biological diversity, which may support this additive approaching scheme to distinguish the active form from the non-active N2D3S system. Experimentally, the IFPTML-LDA method predicted with high probability $p(\text{N2D3S}_{\text{in}})_{\text{cdj, cnj}} > 0.81$ all the examples reported in **Table 5**. The result obtained supports our initial premise that the IFPTML additive approach is able to carry out an appropriate recognition of N2D3S system involving additive and synergic cases.

Table 5: IFPTML analysis of experimentally tested N2D3S compounds.

| Drug ^a | NP | $c_{d0} = \text{Activity}$ | $\Delta\text{DPSA}(\mathbf{c}_l)$ | Obs. ^b | Pred. ^c | p^d | $L(\text{nm})^e$ |
|-------------------|----|----------------------------|-----------------------------------|-------------------|--------------------|-------|------------------|
| Metal / N/A | | | | | | | |
| 2234684 | Ag | Time(h) | 0.57 | 1 | 1 | 0.88 | 12.50 |
| 2376472 | Ag | Time(h) | 4.30 | 1 | 1 | 0.88 | 12.50 |
| 2234683 | Ag | Time(h) | 0.57 | 1 | 1 | 0.88 | 12.50 |

| Metal oxide / n/a | | | | | | | |
|------------------------|--------------------------------|-----------------------|------|---|---|------|------|
| 3769671 | TiO ₂ | Cp(nm) | 0 | 1 | 1 | 0.94 | 56 |
| Levodopa | TiO ₂ | Time(h) | -3.5 | 1 | 1 | 0.93 | 56 |
| Sch-58261 | TiO ₂ | Time(h) | -1 | 1 | 1 | 0.93 | 56 |
| 2180030 | TiO ₂ | EC ₂₀ (nm) | 0 | 1 | 1 | 0.93 | 56 |
| Levodopa | TiO ₂ | Time(h) | -3.5 | 1 | 1 | 0.93 | 56 |
| Sch-58261 | TiO ₂ | Time(h) | -1 | 1 | 1 | 0.93 | 56 |
| 2234689 | TiO ₂ | Time(h) | 0.3 | 1 | 1 | 0.93 | 56 |
| Morin | TiO ₂ | Time(h) | 0 | 1 | 1 | 0.93 | 56 |
| Metal/ elliptical | | | | | | | |
| Datisctin | Ag | Time(h) | 0.3 | 1 | 1 | 0.81 | 36.8 |
| 2234993 | Ag | Time(h) | 0.4 | 1 | 1 | 0.81 | 36.8 |
| 1240582 | Ag | Time(h) | -1.7 | 1 | 1 | 0.81 | 36.8 |
| 1241456 | Ag | Time(h) | -2.1 | 1 | 1 | 0.81 | 36.8 |
| Metal oxide/elliptical | | | | | | | |
| 2180030 | Yb ₂ O ₃ | EC ₂₀ (nm) | 0 | 1 | 1 | 0.90 | 62.1 |
| Levodopa | Yb ₂ O ₃ | Time(h) | -3.5 | 1 | 1 | 0.90 | 62.1 |
| 3769671 | CeO ₂ | Cp(nm) | 0 | 1 | 1 | 0.90 | 44.8 |
| Metal oxide/needle | | | | | | | |
| 3747225 | La ₂ O ₃ | Time(h) | 2.8 | 1 | 1 | 0.89 | 65.8 |
| 3769671 | La ₂ O ₃ | Cp(nm) | 0 | 1 | 1 | 0.88 | 65.8 |
| Meta/rod | | | | | | | |
| 3218426 | Au | Activity(%) | -2.0 | 1 | 1 | 0.93 | 37.8 |
| Congo red | Au | Inhibition(%) | 3.6 | 1 | 1 | 0.93 | 37.8 |
| 3218189 | Au | Activity(%) | -2.0 | 1 | 1 | 0.93 | 37.8 |
| 3580774 | Au | Activity(nm) | 0 | 1 | 1 | 0.93 | 37.8 |
| Metal oxide/pyramidal | | | | | | | |
| PGA | TiO ₂ | Time(h) | -18 | 1 | 1 | 0.91 | 6.5 |
| Apomorphine | TiO ₂ | Time(h) | -17 | 1 | 1 | 0.91 | 50 |
| 1801682 | TiO ₂ | Time(h) | -20 | 1 | 1 | 0.91 | 50 |
| Metal oxide/irregular | | | | | | | |
| 3350757 | TiO ₂ | Time(h) | -5.3 | 1 | 1 | 0.93 | 21 |
| 3747225 | TiO ₂ | Time(h) | 2.8 | 1 | 1 | 0.93 | 21 |

| | | | | | | | |
|------------------------------|------------------|--------------------------|------|---|---|------|------|
| 1243007 | TiO ₂ | Time(h) | -0.7 | 1 | 1 | 0.92 | 21 |
| 3769671 | TiO ₂ | Cp(nm) | 0 | 1 | 1 | 0.92 | 21 |
| Levodopa | TiO ₂ | Time(h) | -3.5 | 1 | 1 | 0.92 | 21 |
| Metal Oxide/Pseudo-spherical | | | | | | | |
| 2376474 | CeO ₂ | Time(h) | 3.9 | 1 | 1 | 0.89 | 8 |
| 3747225 | CeO ₂ | Time(h) | 2.8 | 1 | 1 | 0.89 | 8 |
| 3769671 | CeO ₂ | Cp(nm) | 0 | 1 | 1 | 0.89 | 8 |
| Levodopa | CeO ₂ | Time(h) | -3.5 | 1 | 1 | 0.89 | 8 |
| Sch-58261 | CeO ₂ | Time(h) | -1.0 | 1 | 1 | 0.89 | 8 |
| Metal/spherical | | | | | | | |
| 2151181 | Au | ED ₅₀ (mg/kg) | -0.4 | 1 | 1 | 0.94 | 42.9 |
| 1222303 | Au | ED ₅₀ (mg/kg) | -0.4 | 1 | 1 | 0.94 | 42.9 |
| 2181911 | Au | Activity(%) | 1.6 | 1 | 1 | 0.90 | 42.9 |
| 3397881 | Au | Inhibition(%) | -1.1 | 1 | 1 | 0.90 | 42.9 |
| 3785241 | Au | Inhibition(%) | -1.5 | 1 | 1 | 0.90 | 42.9 |
| 3947919 | Au | Activity(%) | 1.0 | 1 | 1 | 0.90 | 42.9 |
| 3817925 | Au | Inhibition(%) | -0.7 | 1 | 1 | 0.90 | 42.9 |
| 3612821 | Au | Inhibition(%) | 0.3 | 1 | 1 | 0.90 | 42.9 |
| 2159510 | Au | Activity(%) | -0.8 | 1 | 1 | 0.90 | 42.9 |
| 2415095 | Au | Inhibition(%) | 0.5 | 1 | 1 | 0.90 | 42.9 |
| 436483 | Au | Inhibition(%) | 1.5 | 1 | 1 | 0.90 | 42.9 |
| 2159511 | Au | Activity(%) | -1.2 | 1 | 1 | 0.90 | 42.9 |
| 2349470 | Au | Activity(%) | -1.8 | 1 | 1 | 0.90 | 42.9 |
| 3127906 | Au | Activity(%) | 0.6 | 1 | 1 | 0.90 | 42.9 |
| Propidium | Au | Inhibition(%) | 0.4 | 1 | 1 | 0.90 | 42.9 |
| Metal oxide/spherical | | | | | | | |
| 3218188 | SiO ₂ | Activity(%) | 91 | 1 | 1 | 0.97 | 12.5 |
| 3087679 | SiO ₂ | Inhibition(%) | 69 | 1 | 1 | 0.97 | 60 |
| 3233831 | SiO ₂ | Inhibition(%) | 58 | 1 | 1 | 0.97 | 44 |
| 510384 | SiO ₂ | Ki(nm) | -30 | 1 | 1 | 0.97 | 47.5 |
| 81999 | SiO ₂ | Ki(nm) | -40 | 1 | 1 | 0.97 | 36.8 |
| 3218425 | SiO ₂ | Activity(%) | 91 | 1 | 1 | 0.97 | 70 |
| 55401 | SiO ₂ | Ki(nm) | -31 | 1 | 1 | 0.97 | 37 |
| 3233829 | SiO ₂ | Inhibition(%) | 58 | 1 | 1 | 0.97 | 36.8 |

| | | | | | | | |
|---------|------------------|---------------|----|---|---|------|------|
| 3087678 | SiO ₂ | Inhibition(%) | 69 | 1 | 1 | 0.97 | 3.4 |
| 3769671 | SiO ₂ | Cp(nm) | 0 | 1 | 1 | 0.99 | 5.5 |
| 2234689 | SiO ₂ | Time(h) | 37 | 1 | 1 | 0.99 | 36.8 |
| 2234690 | SiO ₂ | Time(h) | 37 | 1 | 1 | 0.99 | 16.4 |

^aChEMBL ID or Drug Name, the name of the drug is depicted if it is available, otherwise the ChEMBLID code of the drug is indicated, which can be easily consulted by accessing the ChEMBL website. ^bClass. Obs: $f(v_{ij}, v_{nj})_{\text{obs}}$. ^cClass. Pred: $f(v_{ij}, v_{nj})_{\text{pred}}$. ^dp: probability calculated as the following: $p(\text{N2D3S}_{\text{in}}/\mathbf{c}_{\text{dj}} \cdot \mathbf{c}_{\text{nj}})_{\text{pred}} = 1/(1+\text{Exp}(-f(v_{ij}, v_{nj}))_{\text{calc}})$. ^eL(nm):NP length.; PGA:Phloroglucin aldehyde

CONCLUSION

The N2D3S system presents a promising and plausible alternative for aiding conventional NDDS in crossing the BBB in the current situation. AI/ML algorithms can be instrumental in expediting this process. However, scientific literature lacks a sufficient number of real N2D3S experimental cases that characterize complex applications. In this context, the IFPTML model, encompassing both NDDS and NP models, could offer a practical solution. This approach has successfully addressed the challenges posed by the vast number of combinations of NP and NDDS compounds and the wide range of conditions to be tested in N2D3S discovery. The results of the IFPTML-LDA and IFPTML-ANN techniques showed satisfactory performance, achieving approximately 73-86.1% and $S_n \approx 70-86.2\%$ in the training and validation series, respectively, comprising 375K and 125K cases. Moreover, both models are easily accessible and provide logical solutions for predicting putative N2D3S. The most successful outcome was observed in the non-linear model, specifically, the IFPTML-MLP, which displayed S_n and S_p values around 85.8-86.2% and an AUROC of 0.94 in the training and validation series. Furthermore, the analysis of three N2D3S system samples yielded low SDV values, confirming the robustness of both IFPTML-LDA and IFPTML-ANN. In summary, the IFPTML models offer an initial solution in a rapid and less arduous manner for pre-screening putative N2D3S. This approach is widely utilized to minimize resource costs and optimize experimental time that would otherwise be spent on testing all possible combinations.

Supporting Information

Supporting Information File 1:

File Name: SI00_Dataset

File Format: xlsb

Supporting Information File 2:

File Name: SI02_Tables

File Format: xlsb

Funding

This work was funded by the grants AIMOFGIFT ELKARTEK project 2022 (KK-2022/00032) - 2022 – 2023 and grant (IT1045-16) - 2016 – 2021 of Basque Government and Grant IKERDATA 2022/IKER/000040 funded by NextGenerationEU funds of European Commission.

Data Availability Statement

The data generated and analyzed during this study is openly available in Figshare repository at DOI: <https://doi.org/10.6084/m9.figshare.25144544>.

ORCID[®]iDs

Shan He - <https://orcid.org/0000-0001-6965-6276>

Julen Segura Abarrategi - <https://orcid.org/0009-0006-3098-5543>

Harbil Bediaga Bañeres - <https://orcid.org/0000-0002-9055-0721>

Sonia Arrasate Gil - <https://orcid.org/0000-0003-2601-5959>

Humberto González Díaz - <https://orcid.org/0000-0002-9392-2797>

References

1. Murray, C.; Lopez, A. D. Global and regional cause-of-death patterns in 1990. *Bulletin of the World Health Organization* **1994**, *72* (3), 447.
2. Aslan, M.; Ozben, T. Reactive oxygen and nitrogen species in Alzheimer's disease. *Current Alzheimer research* **2004**, *1* (2), 111-9.
3. Contestabile, A.; Ciani, E.; Contestabile, A. The place of choline acetyltransferase activity measurement in the "cholinergic hypothesis" of neurodegenerative diseases. *Neurochemical research* **2008**, *33* (2), 318-27.
4. Agnihotri, T. G.; Jadhav, G. S.; Sahu, B.; Jain, A. Recent trends of bioconjugated nanomedicines through nose-to-brain delivery for neurological disorders. *Drug Deliv Transl Res* **2022**, *12* (12), 3104-3120.
5. Polchi, A.; Magini, A.; Mazuryk, J.; Tancini, B.; Gapiński, J.; Patkowski, A.; Giovagnoli, S.; Emiliani, C. Rapamycin Loaded Solid Lipid Nanoparticles as a New Tool to Deliver mTOR Inhibitors: Formulation and in Vitro Characterization. *Nanomaterials* **2016**, *6* (5), 87.
6. Cacciatore, I.; Ciulla, M.; Fornasari, E.; Marinelli, L.; Di Stefano, A. Solid lipid nanoparticles as a drug delivery system for the treatment of neurodegenerative diseases. *Expert opinion on drug delivery* **2016**, *13* (8), 1121-1131.
7. Asefy, Z.; Hoseinnejad, S.; Ceferov, Z. Nanoparticles approaches in neurodegenerative diseases diagnosis and treatment. *Neurol Sci* **2021**, *42* (7), 2653-2660.
8. Shayganfard, M. A Review on Chitosan in Drug Delivery for Treatment of Neurological and Psychiatric Disorders. *Curr Pharm Biotechnol* **2022**, *23* (4), 538-551.
9. Verma, R.; Sartaj, A.; Qizilbash, F. F.; Ghoneim, M. M.; Alshehri, S.; Imam, S. S.; Kala, C.; Alam, M. S.; Gilani, S. J.; Taleuzzaman, M. An Overview of the Neuropharmacological Potential of Thymoquinone and its Targeted Delivery Prospects for CNS Disorder. *Current drug metabolism* **2022**, *23* (6), 447-459.
10. Syed, A. A.; Reza, M. I.; Singh, P.; Thombre, G. K.; Gayen, J. R. Withania somnifera in Neurological Disorders: Ethnopharmacological Evidence, Mechanism of Action and its Progress in Delivery Systems. *Current drug metabolism* **2021**, *22* (7), 561-571.
11. Yu, Y.; O'Rourke, A.; Lin, Y. H.; Singh, H.; Eiguez, R. V.; Beyhan, S.; Nelson, K. E. Predictive Signatures of 19 Antibiotic-Induced Escherichia coli Proteomes. *ACS infectious diseases* **2020**, *6* (8), 2120-2129.
12. Pribut, N.; Kaiser, T. M.; Wilson, R. J.; Jecs, E.; Dentmon, Z. W.; Pelly, S. C.; Sharma, S.; Bartsch, P. W., 3rd; Burger, P. B.; Hwang, S. S.; Le, T.; Sourimant, J.; Yoon, J. J.; Plemper, R. K.; Liotta, D. C. Accelerated Discovery of Potent Fusion Inhibitors for Respiratory Syncytial Virus. *ACS infectious diseases* **2020**, *6* (5), 922-929.
13. Wang, X.; Perryman, A. L.; Li, S. G.; Paget, S. D.; Stratton, T. P.; Lemenze, A.; Olson, A. J.; Ekins, S.; Kumar, P.; Freundlich, J. S. Intrabacterial Metabolism Obscures the Successful Prediction of an InhA Inhibitor of Mycobacterium tuberculosis. *ACS infectious diseases* **2019**, *5* (12), 2148-2163.
14. Cooper, S. J.; Krishnamoorthy, G.; Wolloscheck, D.; Walker, J. K.; Rybenkov, V. V.; Parks, J. M.; Zgurskaya, H. I. Molecular Properties That Define the Activities of Antibiotics in Escherichia coli and Pseudomonas aeruginosa. *ACS infectious diseases* **2018**, *4* (8), 1223-1234.
15. Duncan, G. A.; Bevan, M. A. Computational design of nanoparticle drug delivery systems for selective targeting. *Nanoscale* **2015**, *7* (37), 15332-15340.

16. Zhou, H.; Cao, H.; Matyunina, L.; Shelby, M.; Cassels, L.; McDonald, J. F.; Skolnick, J. MEDICASCY: A Machine Learning Approach for Predicting Small-Molecule Drug Side Effects, Indications, Efficacy, and Modes of Action. *Molecular pharmaceutics* **2020**, *17* (5), 1558-1574.
17. Sun, L.; Yang, H.; Cai, Y.; Li, W.; Liu, G.; Tang, Y. In Silico Prediction of Endocrine Disrupting Chemicals Using Single-Label and Multilabel Models. *Journal of chemical information and modeling* **2019**, *59* (3), 973-982.
18. Kolesov, A.; Kamyshenkov, D.; Litovchenko, M.; Smekalova, E.; Golovizin, A.; Zhavoronkov, A. On multilabel classification methods of incompletely labeled biomedical text data. *Computational and mathematical methods in medicine* **2014**, *2014*, 781807.
19. Heider, D.; Senge, R.; Cheng, W.; Hullermeier, E. Multilabel classification for exploiting cross-resistance information in HIV-1 drug resistance prediction. *Bioinformatics* **2013**, *29* (16), 1946-52.
20. Manganelli, S.; Leone, C.; Toropov, A. A.; Toropova, A. P.; Benfenati, E. QSAR model for predicting cell viability of human embryonic kidney cells exposed to SiO₂ nanoparticles. *Chemosphere* **2016**, *144*, 995-1001.
21. Toropova, A. P.; Toropov, A. A.; Rallo, R.; Leszczynska, D.; Leszczynski, J. Optimal descriptor as a translator of eclectic data into prediction of cytotoxicity for metal oxide nanoparticles under different conditions. *Ecotoxicology and environmental safety* **2015**, *112*, 39-45.
22. Toropova, A. P.; Toropov, A. A.; Veselinovic, A. M.; Veselinovic, J. B.; Benfenati, E.; Leszczynska, D.; Leszczynski, J. Nano-QSAR: Model of mutagenicity of fullerene as a mathematical function of different conditions. *Ecotoxicol. Environ. Saf.* **2016**, *124*, 32-36.
23. Rybinska-Fryca, A.; Mikolajczyk, A.; Puzyn, T. Structure-activity prediction networks (SAPNets): a step beyond Nano-QSAR for effective implementation of the safe-by-design concept. *Nanoscale* **2020**.
24. Le, T. C.; Yin, H.; Chen, R.; Chen, Y.; Zhao, L.; Casey, P. S.; Chen, C.; Winkler, D. A. An Experimental and Computational Approach to the Development of ZnO Nanoparticles that are Safe by Design. *Small* **2016**, *12* (26), 3568-77.
25. Ahmadi, S.; Toropova, A. P.; Toropov, A. A. Correlation intensity index: mathematical modeling of cytotoxicity of metal oxide nanoparticles. *Nanotoxicology* **2020**, 1-9.
26. Ojha, P. K.; Kar, S.; Roy, K.; Leszczynski, J. Toward comprehension of multiple human cells uptake of engineered nano metal oxides: quantitative inter cell line uptake specificity (QICLUS) modeling. *Nanotoxicology* **2019**, *13* (1), 14-34.
27. Sizochenko, N.; Gajewicz, A.; Leszczynski, J.; Puzyn, T. Reply to the comment on "Causation or only correlation? Application of causal inference graphs for evaluating causality in nano-QSAR models" by D. A. Tasi, J. Csontos, B. Nagy, Z. Konya and G. Tasi, *Nanoscale*, 2018, 10, C8NR02377H. *Nanoscale* **2018**, *10* (44), 20867-20868.
28. Tasi, D. A.; Csontos, J.; Nagy, B.; Konya, Z.; Tasi, G. Comment on "Causation or only correlation? Application of causal inference graphs for evaluating causality in nano-QSAR models" by N. Sizochenko, A. Gajewicz, J. Leszczynski and T. Puzyn, *Nanoscale*, 2016, 8, 7203. *Nanoscale* **2018**, *10* (44), 20863-20866.
29. Villaverde, J. J.; Sevilla-Moran, B.; Lopez-Goti, C.; Alonso-Prados, J. L.; Sandin-Espana, P. Considerations of nano-QSAR/QSPR models for nanopesticide risk assessment within the European legislative framework. *Sci Total Environ* **2018**, *634*, 1530-1539.
30. Sizochenko, N.; Leszczynska, D.; Leszczynski, J. Modeling of Interactions between the Zebrafish Hatching Enzyme ZHE1 and A Series of Metal Oxide

Nanoparticles: Nano-QSAR and Causal Analysis of Inactivation Mechanisms. *Nanomaterials* **2017**, 7 (10).

31. Manganelli, S.; Benfenati, E. Nano-QSAR Model for Predicting Cell Viability of Human Embryonic Kidney Cells. *Methods in molecular biology* **2017**, 1601, 275-290.
32. Puzyn, T.; Rasulev, B.; Gajewicz, A.; Hu, X.; Dasari, T. P.; Michalkova, A.; Hwang, H. M.; Toropov, A.; Leszczynska, D.; Leszczynski, J. Using nano-QSAR to predict the cytotoxicity of metal oxide nanoparticles. *Nature nanotechnology* **2011**, 6 (3), 175-8.
33. Sizochenko, N.; Mikolajczyk, A.; Jagiello, K.; Puzyn, T.; Leszczynski, J.; Rasulev, B. How the toxicity of nanomaterials towards different species could be simultaneously evaluated: a novel multi-nano-read-across approach. *Nanoscale* **2018**, 10 (2), 582-591.
34. Toropov, A. A.; Toropova, A. P.; Benfenati, E.; Gini, G.; Puzyn, T.; Leszczynska, D.; Leszczynski, J. Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles to bacteria *Escherichia coli*. *Chemosphere* **2012**, 89 (9), 1098-1102.
35. Gonzalez-Diaz, H.; Arrasate, S.; Gomez-SanJuan, A.; Sotomayor, N.; Lete, E.; Besada-Porto, L.; Ruso, J. M. General theory for multiple input-output perturbations in complex molecular systems. 1. Linear QSPR electronegativity models in physical, organic, and medicinal chemistry. *Current topics in medicinal chemistry* **2013**, 13 (14), 1713-41.
36. Alonso, N.; Caamano, O.; Romero-Duran, F. J.; Luan, F.; MN, D. S. C.; 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), 1393-403.
37. Diez-Alarcia, R.; Yanez-Perez, V.; Muneta-Arrate, I.; Arrasate, S.; Lete, E.; Meana, J. J.; Gonzalez-Diaz, H. Big Data Challenges Targeting Proteins in GPCR Signaling Pathways; Combining PTML-ChEMBL Models and [(35)S]GTPgammaS Binding Assays. *ACS chemical neuroscience* **2019**, 10 (11), 4476-4491.
38. Gonzalez-Diaz, H.; Riera-Fernandez, P.; Pazos, A.; Munteanu, C. R. The Rucker-Markov invariants of complex Bio-Systems: applications in Parasitology and Neuroinformatics. *Bio Systems* **2013**, 111 (3), 199-207.
39. Gonzalez-Diaz, H.; Herrera-Ibata, D. M.; Duardo-Sanchez, A.; Munteanu, C. R.; Orbegozo-Medina, R. A.; Pazos, A. ANN multiscale model of anti-HIV drugs activity vs AIDS prevalence in the US at county level based on information indices of molecular graphs and social networks. *Journal of chemical information and modeling* **2014**, 54 (3), 744-55.
40. Gonzalez-Diaz, H.; Riera-Fernandez, P. New Markov-autocorrelation indices for re-evaluation of links in chemical and biological complex networks used in metabolomics, parasitology, neurosciences, and epidemiology. *Journal of chemical information and modeling* **2012**, 52 (12), 3331-40.
41. Concu, R.; MN, D. S. C.; Munteanu, C. R.; Gonzalez-Diaz, H. PTML Model of Enzyme Subclasses for Mining the Proteome of Biofuel Producing Microorganisms. *Journal of proteome research* **2019**, 18 (7), 2735-2746.
42. Martinez-Arzate, S. G.; Tenorio-Borroto, E.; Barbabosa Pliego, A.; Diaz-Albiter, H. M.; Vazquez-Chagoyan, J. C.; Gonzalez-Diaz, H. PTML Model for Proteome Mining of B-Cell Epitopes and Theoretical-Experimental Study of Bm86 Protein Sequences from Colima, Mexico. *Journal of proteome research* **2017**, 16 (11), 4093-4103.
43. Quevedo-Tumaili, V. F.; Ortega-Tenezaca, B.; Gonzalez-Diaz, H. Chromosome Gene Orientation Inversion Networks (GOINs) of Plasmodium Proteome. *Journal of proteome research* **2018**, 17 (3), 1258-1268.

44. Santana, R.; Zuluaga, R.; Ganan, P.; Arrasate, S.; Onieva, E.; Gonzalez-Diaz, H. Predicting coated-nanoparticle drug release systems with perturbation-theory machine learning (PTML) models. *Nanoscale* **2020**, *12* (25), 13471-13483.
45. Romero Duran, F. J.; Alonso, N.; Caamano, O.; Garcia-Mera, X.; Yanez, M.; Prado-Prado, F. J.; Gonzalez-Diaz, H. Prediction of multi-target networks of neuroprotective compounds with entropy indices and synthesis, assay, and theoretical study of new asymmetric 1,2-rasagiline carbamates. *International journal of molecular sciences* **2014**, *15* (9), 17035-64.
46. Luan, F.; Cordeiro, M. N.; Alonso, N.; Garcia-Mera, X.; Caamano, O.; Romero-Duran, F. J.; Yanez, M.; Gonzalez-Diaz, H. TOPS-MODE model of multiplexing neuroprotective effects of drugs and experimental-theoretic study of new 1,3-rasagiline derivatives potentially useful in neurodegenerative diseases. *Bioorganic & medicinal chemistry* **2013**, *21* (7), 1870-9.
47. Gonzalez-Diaz, H. QSAR and complex networks in pharmaceutical design, microbiology, parasitology, toxicology, cancer, and neurosciences. *Current pharmaceutical design* **2010**, *16* (24), 2598-600.
48. Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Speck-Planche, A.; Cordeiro, M. N. 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), 14686-94.
49. Luan, F.; Kleandrova, V. V.; González-Díaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N. D. S. Computer-aided nanotoxicology: Assessing cytotoxicity of nanoparticles under diverse experimental conditions by using a novel QSTR-perturbation approach. *Nanoscale* **2014**, *6* (18), 10623-10630.
50. Santana, R.; Zuluaga, R.; Ganan, P.; Arrasate, S.; Onieva, E.; Gonzalez-Diaz, H. Designing nanoparticle release systems for drug-vitamin cancer co-therapy with multiplicative perturbation-theory machine learning (PTML) models. *Nanoscale* **2019**, *11* (45), 21811-21823.
51. Santana, R.; Zuluaga, R.; Ganan, P.; Arrasate, S.; Onieva, E.; Montemore, M. M.; Gonzalez-Diaz, H. PTML Model for Selection of Nanoparticles, Anticancer Drugs, and Vitamins in the Design of Drug-Vitamin Nanoparticle Release Systems for Cancer Co-therapy. *Molecular pharmaceuticals* **2020**, *17* (7), 2612-2627.
52. Urista, D. V.; Carrue, D. B.; Otero, I.; Arrasate, S.; Quevedo-Tumaili, V. F.; Gestal, M.; Gonzalez-Diaz, H.; Munteanu, C. R. Prediction of Antimalarial Drug-Decorated Nanoparticle Delivery Systems with Random Forest Models. *Biology* **2020**, *9* (8).
53. Bento, A. P.; Gaulton, A.; Hersey, A.; Bellis, L. J.; Chambers, J.; Davies, M.; Kruger, F. A.; Light, Y.; Mak, L.; McGlinchey, S.; Nowotka, M.; Papadatos, G.; Santos, R.; Overington, J. P. The ChEMBL bioactivity database: an update. *Nucleic Acids Res* **2014**, *42* (Database issue), D1083-90.
54. Davies, M.; Nowotka, M.; Papadatos, G.; Dedman, N.; Gaulton, A.; Atkinson, F.; Bellis, L.; Overington, J. P. ChEMBL web services: streamlining access to drug discovery data and utilities. *Nucleic Acids Res* **2015**, *43* (W1), W612-20.
55. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B.; Overington, J. P. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic Acids Res.* **2012**, *40* (Database issue), D1100-7.
56. Gusenbauer, M.; Haddaway, N. R. Which academic search systems are suitable for systematic reviews or meta-analyses? Evaluating retrieval qualities of

- Google Scholar, PubMed, and 26 other resources. *Research synthesis methods* **2020**, *11* (2), 181-217.
57. Lu, Z. PubMed and beyond: a survey of web tools for searching biomedical literature. *Database* **2011**, 2011.
58. Islamaj Dogan, R.; Murray, G. C.; Névéol, A.; Lu, Z. Understanding PubMed® user search behavior through log analysis. *Database* **2009**, 2009.
59. Li, Y.; Li, H.; Pickard, F. C. t.; Narayanan, B.; Sen, F. G.; Chan, M. K. Y.; Sankaranarayanan, S.; Brooks, B. R.; Roux, B. Machine Learning Force Field Parameters from Ab Initio Data. *Journal of chemical theory and computation* **2017**, *13* (9), 4492-4503.
60. Xia, R.; Kais, S. Quantum machine learning for electronic structure calculations. *Nature communications* **2018**, *9* (1), 4195.
61. Na, G. S.; Chang, H.; Kim, H. W. Machine-guided representation for accurate graph-based molecular machine learning. *Physical chemistry chemical physics : PCCP* **2020**, *22* (33), 18526-18535.
62. Concu, R.; Kleandrova, V. V.; Speck-Planche, A.; Cordeiro, M. Probing the toxicity of nanoparticles: a unified in silico machine learning model based on perturbation theory. *Nanotoxicology* **2017**, *11* (7), 891-906.
63. 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.
64. Kleandrova, V. V.; Luan, F.; Gonzalez-Diaz, H.; Ruso, J. M.; Melo, A.; Speck-Planche, A.; Cordeiro, M. N. Computational ecotoxicology: simultaneous prediction of ecotoxic effects of nanoparticles under different experimental conditions. *Environment international* **2014**, *73*, 288-94.
65. Gaulton, A.; Hersey, A.; Nowotka, M.; Bento, A. P.; Chambers, J.; Mendez, D.; Mutowo, P.; Atkinson, F.; Bellis, L. J.; Cibrian-Uhalte, E.; Davies, M.; Dedman, N.; Karlsson, A.; Magarinos, M. P.; Overington, J. P.; Papadatos, G.; Smit, I.; Leach, A. R. The ChEMBL database in 2017. *Nucleic Acids Res* **2017**, *45* (D1), D945-D954.
66. Mendez, D.; Gaulton, A.; Bento, A. P.; Chambers, J.; De Veij, M.; Felix, E.; Magarinos, M. P.; Mosquera, J. F.; Mutowo, P.; Nowotka, M.; Gordillo-Maranon, M.; Hunter, F.; Junco, L.; Mugumbate, G.; Rodriguez-Lopez, M.; Atkinson, F.; Bosc, N.; Radoux, C. J.; Segura-Cabrera, A.; Hersey, A.; Leach, A. R. ChEMBL: towards direct deposition of bioassay data. *Nucleic Acids Res* **2019**, *47* (D1), D930-D940.
67. Moriwaki, H.; Tian, Y.-S.; Kawashita, N.; Takagi, T. Mordred: a molecular descriptor calculator. *Journal of cheminformatics* **2018**, *10* (1), 1-14.
68. Hill, T.; Lewicki, P. *Statistics: Methods and Applications*. 1st edition ed.; StatSoft, Inc.: 2005; p 800.
69. Gaulton, A.; Bellis, L. J.; Bento, A. P.; Chambers, J.; Davies, M.; Hersey, A.; Light, Y.; McGlinchey, S.; Michalovich, D.; Al-Lazikani, B. ChEMBL: a large-scale bioactivity database for drug discovery. *Nucleic acids research* **2012**, *40* (D1), D1100-D1107.
70. Huberty, C. J.; Olejnik, S. *Applied MANOVA and discriminant analysis*. 2nd ed.; John Wiley & Sons, Inc.: Hoboken, New Jersey, 2006.
71. Hanczar, B.; Hua, J.; Sima, C.; Weinstein, J.; Bittner, M.; Dougherty, E. R. Small-sample precision of ROC-related estimates. *Bioinformatics* **2010**, *26* (6), 822-830.
72. Bian, L.; Sorescu, D. C.; Chen, L.; White, D. L.; Burkert, S. C.; Khalifa, Y.; Zhang, Z.; Sejdic, E.; Star, A. Machine-Learning Identification of the Sensing Descriptors

Relevant in Molecular Interactions with Metal Nanoparticle-Decorated Nanotube Field-Effect Transistors. *ACS Appl Mater Interfaces* **2019**, *11* (1), 1219-1227.

73. Alafeef, M.; Srivastava, I.; Pan, D. Machine Learning for Precision Breast Cancer Diagnosis and Prediction of the Nanoparticle Cellular Internalization. *ACS sensors* **2020**, *5* (6), 1689-1698.

74. Sun, B.; Fernandez, M.; Barnard, A. S. Machine Learning for Silver Nanoparticle Electron Transfer Property Prediction. *Journal of chemical information and modeling* **2017**, *57* (10), 2413-2423.

75. Barnard, A. S.; Opletal, G. Predicting structure/property relationships in multi-dimensional nanoparticle data using t-distributed stochastic neighbour embedding and machine learning. *Nanoscale* **2019**, *11* (48), 23165-23172.

76. He, J.; He, C.; Zheng, C.; Wang, Q.; Ye, J. Plasmonic nanoparticle simulations and inverse design using machine learning. *Nanoscale* **2019**, *11* (37), 17444-17459.

77. Yan, T.; Sun, B.; Barnard, A. S. Predicting archetypal nanoparticle shapes using a combination of thermodynamic theory and machine learning. *Nanoscale* **2018**, *10* (46), 21818-21826.

78. Speck-Planche, A.; Kleandrova, V. V.; Luan, F.; Cordeiro, M. N. Computational modeling in nanomedicine: prediction of multiple antibacterial profiles of nanoparticles using a quantitative structure-activity relationship perturbation model. *Nanomedicine* **2015**, *10* (2), 193-204.

79. Nocado-Mena, D.; Cornelio, C.; Camacho-Corona, M. D. R.; Garza-Gonzalez, E.; Waksman de Torres, N.; Arrasate, S.; Sotomayor, N.; Lete, E.; Gonzalez-Diaz, H. Modeling Antibacterial Activity with Machine Learning and Fusion of Chemical Structure Information with Microorganism Metabolic Networks. *Journal of chemical information and modeling* **2019**, *59* (3), 1109-1120.

80. Dieguez-Santana, K.; Gonzalez-Diaz, H. Towards machine learning discovery of dual antibacterial drug-nanoparticle systems. *Nanoscale* **2021**, *13* (42), 17854-17870.

81. Leys, C.; Ley, C.; Klein, O.; Bernard, P.; Licata, L. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of experimental social psychology* **2013**, *49* (4), 764-766.

82. Arnedo, M. Cladismo: La reconstrucción filogenética basada en parsimonia. *Boletín de la Sociedad Entomológica Aragonesa* **1999**, *26*, 57-84.

83. Hitchcock, S. A.; Pennington, L. D. Structure- brain exposure relationships. *Journal of medicinal chemistry* **2006**, *49* (26), 7559-7583.

84. Doan, K. M. M.; Humphreys, J. E.; Webster, L. O.; Wring, S. A.; Shampine, L. J.; Serabjit-Singh, C. J.; Adkison, K. K.; Polli, J. W. Passive permeability and P-glycoprotein-mediated efflux differentiate central nervous system (CNS) and non-CNS marketed drugs. *Journal of Pharmacology and Experimental Therapeutics* **2002**, *303* (3), 1029-1037.

85. Chithrani, B. D.; Ghazani, A. A.; Chan, W. C. Determining the size and shape dependence of gold nanoparticle uptake into mammalian cells. *Nano letters* **2006**, *6* (4), 662-668.

86. Shilo, M.; Sharon, A.; Baranes, K.; Motiei, M.; Lellouche, J.-P. M.; Popovtzer, R. The effect of nanoparticle size on the probability to cross the blood-brain barrier: an in-vitro endothelial cell model. *Journal of nanobiotechnology* **2015**, *13* (1), 1-7.