Electronic Supplementary Material (ESI) for Environmental Science: Nano.

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.

ChEMBL AD, NP, and MRN datasets. The ChEMBL dataset use here includes >160 000 outcomes of AD preclinical assays for 55 931 compounds. Each compound have the outcome of at least 1 out of >300 biological activity parameters (MIC, IC_{50}, *etc*.). Each compound was assayed against at least 1 out of >90 bacteria strains of >25 bacterial species. The chemical structure of each AD candidate compound was encoded into a vector of molecular descriptors $\mathbf{D}_{dk} = [D_{d1}, D_{d2}, D_{d3}]$. The elements of this vector are the molecular descriptors of the i^{th} compound: D_{d1} = Logarithm of the n-Octanol/Water Partition coefficient ($LOGP_i$), D_{d2} = Topological Polar Surface Area (PSA_i), D_{d3} = Number of Violations of Lipinski's Rule ($NVLR_i$). The specific labels or conditions of each assay were encoded into the vectors $\mathbf{c}_{dj} = [c_{d0}, c_{d1}, c_{d2}]$. The elements of these vectors are c_{d0} = name of the biological parameter (MIC, IC_{50}, *etc*.) c_{d1} = name of the bacteria specie, c_{d2} = label or code of the bacteria strain. Please do not confuse numeric value f the biological activity parameter v_{ij} with the name of the biological activity parameter c_{d0}. This dataset was obtained from a previous dataset reported before by our group after a new verification and pre-processing.[1] The NP dataset with the outcomes of N_n = 300 pre-clinical assays of metal, metal salt, and metal oxide NPs against different bacteria species (s).[2] The NP assays have multiple experimental variables conditioning the nature of the assay c_{nj}. We listed all the specific conditions of one assay as a vector $\mathbf{c}_{nj} = [c_{n1}, c_{n2}, c_{n3} \ldots\ldots c_{nmax}]$. It includes the report of 1 out of 4 possible NP action parameters for 34 possible bacteria/strains. The data also contains NP shapes, NP physicochemical properties, NP coating agents, and time of assay, and (see details on Supporting Information file SI00.doc). [2]

Nanoparticles Dataset (ABNP-set). We used a previously reported dataset with the outcomes of N_n = 300 pre-clinical assays of metal, metal salt, and metal oxide ABNPs against different bacteria species (s).[2] The metal ABNP have a core made of: gold (Au), silver (Ag), or copper (Cu). The metal salt ABNP cores are made of cadmium(II) sulfide (CdS) or copper(I) iodide. The metal oxide ABNP include: cadmium(II) oxide (CdO), zinc oxide (ZnO), copper(II) oxide (CuO), lanthanum(III) oxide (La_2O_3), aluminium oxide (Al_2O_3), iron(III) oxide

(Fe_2O_3), tin(IV) oxide (SnO_2), titanium(IV) oxide (TiO_2), iron(II, III) oxide (Fe_3O_4), and silicon dioxide (SiO_2). These assays of these 15 nanomaterials involved multiple experimental conditions c_{nj}. We listed all the specific conditions of one assay as a vector $\mathbf{c}_{nj} = [c_{nj}, c_{nj}, c_{nj} \ldots\ldots c_{nmax}]$. These conditions of assay include the measurement of 1 out of 4 possible antibacterial activity parameters, against 1 out of 34 possible bacteria species (different strains included). Other labels or experimental conditions considered are selecting at least 1 out of 3 ABNPs shape and running the experiment in 1 out of 4 possible intervals time during. The original data was downloaded from OCHEM database (https://ochem.eu/home/show.do)[3] and other sources.[4-15] The dataset also included information about physicochemical parameters of the ABNP and the coating agents used (see next sections). [2]

Bacterial MRNs dataset (MRN-set). The data were downloaded directly from Barabasi's group website (http://www.nd.edu/~networks/resources.htm) as gzipped ASCII files. In these files each number represents a substrate in the metabolic network. Data-format is: From → To (directed link). The information studied was previously obtained by Jeong *et al*. from the 'intermediate metabolism and bioenergetics' portions of the WIT database and used in order to try to understand the large-scale organization of metabolic networks.[16] According to the authors, the biochemical reactions described within the WIT database are composed of substrates and enzymes connected by directed links. For each reaction, educts and products were considered as nodes connected to the temporary educt-educt complexes and associated enzymes. Bidirectional reactions were considered separately. For a given organism with N substrates, E enzymes and R intermediate complexes the full stoichiometric interactions were compiled into an (N+E+R) X (N+E+R) matrix, generated separately for each of the different organisms. The names, abbreviations, and links for all the networks studied are: *Actinobacillus actinomycetemcomitans = AB; Bacillus subtilis = BS; Clostridium acetobutylicum = CA; Campylobacter jejuni = CJ; Chlamydia pneumoniae = CQ; Chlamydia trachomatis = CT; Deinococcus radiodurans = DR; Escherichia coli = EC; Enterococcus faecalis = EF; Haemophilus influenza =HI; Helicobacter pylori = HP; Mycobacterium bovis = MB; Mycoplasma genitalium = MG; Mycobacterium leprae = ML; Mycoplasma pneumonia = MP; Mycobacterium tuberculosis = MT; Neisseria gonorrhoeae = NG; Neisseria meningitidis = NM; Pseudomonas aeruginosa =PA; Porphyromonas gingivalis = PG; Streptococcus pneumonia = PN; Rhodobacter capsulatus = RC; Saccharomyces cerevisiae = SC; Streptococcus pyogenes = ST; Salmonella typhi =TY; Yersinia pestis =YP.*

Shannon's transform of input variables. This IFPTML model considers that the system under study (S) is composed of various subsystems ($S = S_d + S_n + S_s$) with S_d = AD, S_n = NP, S_s = MRN. The structure of each subsystem is encoded with the vectors of molecular/structural descriptors $\mathbf{D}_{dk}$, $\mathbf{D}_{nk}$, and $\mathbf{D}_{sk}$, respectively. The vectors of the subsystem S_d have the elements $\mathbf{D}_{dk} = [D_{d1}, D_{d2}, D_{d3}, D_{d4}]$. These elements are the descriptors the i^{th} AD. They are: D_{d1} = Logarithm of the n-Octanol/Water Partition coefficients ($LOGP_i$), D_{d2} = Topological Polar Surface Area (PSA_i), D_{d3} = Number of Violations to Lipinski's Rule ($NVLR_i$), and D_{d4} = Molecular Weight (Mw_i). The vectors of the subsystem S_n have the elements: $\mathbf{D}_{nk} = [D_{n1}, D_{n2}, D_{n3}, D_{n4}, D_{n5}, D_{n6}, D_{n7}, D_{n8}]$. They are the properties of the n^{th} NP. The two first are: D_{n1} = NP Molar Volume (AMV) and D_{n2} = Average Atomic Electronegativity (AAE). The two other are: D_{n3} = Average Atomic Polarizability (AAP) and D_{n4} = Average Particle Size (APS) of the NP core in nanometers (nm). The vector $\mathbf{D}_{nk}$ has also as elements the descriptors of the first (ca1) and second (ca2) CAs of the n^{th} NP: $D_{n5} = LOGP_{ca1}$, $D_{n6} = PSA_{ca1}$, $D_{n5} = LOGP_{ca2}$ and $D_{n6} = PSA_{ca2}$. Last, the vectors of the subsystem S_s have the elements: $\mathbf{D}_{sk} = [D_{s1}, D_{s2}, D_{s3},]$. They are the topological parameters of the structure of the s^{th} MRN_s: $D_{s1} = N_s$ MRNs Number of nodes (number of metabolites), $D_{s2} = L_{ins}$ Average in-degree (Average number of educts or substrates), $D_{s3} = L_{outs}$ Average out-degree (Average number of adducts or products). They have different units and scales making a necessity the re-scaling and/or standardization of all the information into the same scale towards the subsequent IF and ML processing. As one IF process is involved we selected the Shannon's entropy information measure as the scaling transformation. All the AD, NP, and NP coat variables have been transformed in $Sh(D_{nk})$ values using the following equations.

$$p(D_k) = \frac{1}{(1+\mathrm{Exp}\,(-D_k/1000)} \quad (S1)$$

$$Sh(D_k) = -p(D_k) \cdot \log(p(D_k)) \quad (S2)$$

Shannon-entropy scaling of NP structural information.

The original NP-set contains different experimental/theoretical physicochemical parameter to characterize the NP structure/composition details. These parameters were the Average Molar Volume (AMV), the Average Atomic Electronegativity (AAE), and the Average Atomic Polarizability (AAP). These physicochemical properties were retrieved from the website Chemicool Periodic Table (http://www.chemicool.com/elements).[17] The fourth parameter was the Average Particle Size (APS) expressed in nanometers (nm). However, in order to carry out the IF process making a fusion of the NP and AD on the same working dataset we decided to express all the information in the same scale. Consequently, the information of 2 datasets was transformed into a Shannon's entropy scale previously to fusion. The information about NP core and coating agents has been scaled using the same formulae to calculate the Shannon's entropy values. After that, we obtained the values of entropy $Sh_k(D_{kn})$. With the $Sh_k(D_{kn})$ values we can calculate the PTOs of the NP assays used as input for the PTMLIF model. The PTOs calculated here has the form of multi-condition MAs by analogy to previous reports. The formula of these PTOs is the following $\Delta Sh(D_{kn}) = Sh_{kn} - \langle Sh_{kn} \rangle_{cn}$. In **Table S1** we show selected examples of the average values $\langle Sh_{kn} \rangle_{cn}$ for different subsets of NP assay conditions $\mathbf{c}_n$ (Supporting Information file SI00.doc). The information about all the NPs, shape, type, and values of Sh_{kn} and $\langle Sh_{kn} \rangle_{cn}$ appear in the Supporting Information file SI01.xlsx, see NP sheet.

Table S1. Shannon's entropy information measures for NP (selected examples)

NP Type	NP	Shape	$Sh(MW_n)$	$Sh(AMV_n)$	$Sh(AAE_n)$	$Sh(AAP_n)$	$Sh(APS_n)$
Oxide	ZnO	Acicular	0.1476	0.1501	0.1504	0.1504	0.1497
	ZnO	N/A	0.1476	0.1501	0.1504	0.1504	0.1493
	CuO	N/A	0.1477	0.1502	0.1504	0.1504	0.1493
	La_2O_3	N/A	0.1371	0.1499	0.1504	0.1501	0.1493
	Al_2O_3	N/A	0.1468	0.1501	0.1504	0.1504	0.1493
	Fe_2O_3	N/A	0.1445	0.1501	0.1504	0.1504	0.1493
	SnO_2	N/A	0.1449	0.15	0.1504	0.1504	0.1493
	TiO_2	N/A	0.1477	0.1501	0.1504	0.1503	0.1493
	SiO_2	N/A	0.1484	0.1501	0.1504	0.1504	0.1493
	CdO	Spherical	0.1458	0.1501	0.1504	0.1504	0.1498
	Fe_3O_4	Spherical	0.1414	0.1501	0.1504	0.1504	0.1501
Metal	CuI	N/A	0.1432	0.15	0.1504	0.1503	0.1502
	CdS	Spherical	0.1452	0.15	0.1504	0.1503	0.1504
	Au	Spherical	0.143	0.1502	0.1504	0.1503	0.1505
	Ag	Spherical	0.1466	0.1502	0.1505	0.1503	0.1504

	Cu	Spherical	0.1483	0.1503	0.1504	0.1503	0.1502
NP	Org.	Strain	Shape	Average Values			
Type	cn_1	cn_2	cn_3	$\langle Sh(AMV_n)_{cnj} \rangle$	$\langle Sh(AAE_n)_{cnj} \rangle$	$\langle Sh(AAP_n)_{cnj} \rangle$	$\langle Sh(APS_n)_{cnj} \rangle$
All	EC	K-12	Spherical	0.14647	0.15013	0.15044	0.15032
	EC	MDR		0.14296	0.15017	0.15043	0.15031
	EC	ATCC 10536		0.1471	0.1502	0.1505	0.1503
	EF	VCM-R		0.1466	0.1502	0.1505	0.1503
	SA	ATCC 9144	Acicular	0.14763	0.15012	0.15043	0.15039
	EC	ATCC 10536		0.14763	0.15012	0.15043	0.15039
	PA	ATCC 9027		0.14763	0.15012	0.15043	0.15039

In **Table S2** we show the individual values of $Sh(D_{cak})$ and the average values $\langle Sh(D_{ack})_{cnj} \rangle$ for each descriptor D_{ack} of the coating agents. These MAs quantify the variability on the first coating agent, the second coating agent (if any), and the time of assay, respectively. However, the values of variance of these MAs were too low to be included in ML analysis. Consequently, we decided to encode all this information into a modified type of PTOs based on multiple Shannon's entropy information measures $\Delta Sh(D_{ca1}, D_{ca2}, D_{dk})$. The use of many different types of PTOs in PTMLIF analysis applied to Nanotechnology was discussed in the literature before.[18-20]

Table S2. Shannon's entropy information measures for NP coating agents

Coating systems			Coating systems numerical information			
N_{coat}	Poly.	Coating System	Coating Agent 01		Coating Agent 02	
			$Sh(LOGP_{ac1})$	$Sh(PSA_{ac1})$	$Sh(LOGP_{ac2})$	$Sh(PSA_{ac2})$
Double	Mono.	PDT/Mel	0.14776	0.15050	0.14627	0.15054
		PDT/ACh	0.14776	0.15050	0.14962	0.15055
		PDT/CQ	0.14776	0.15050	0.14956	0.15037
		PDT/DMB	0.14776	0.15050	0.14734	0.15049
		PDT/CPB	0.14776	0.15050	0.14700	0.15045
		PDT/G	0.14776	0.15050	0.14783	0.15051
Single		PDT	0.14776	0.15050	0.15051	0.15051
		Maltose	0.14328	0.15065	0.15051	0.15051
		Lactose	0.14328	0.15065	0.15051	0.15051
		Glutathione	0.14457	0.15058	0.15051	0.15051
		Glucose	0.14652	0.15059	0.15051	0.15051
		DMA	0.15051	0.15022	0.15051	0.15051
		Galactose	0.14652	0.15059	0.15051	0.15051
	Poly.	PVP	0.14983	0.15052	0.15051	0.15051
		PGA	0.14690	0.15055	0.15051	0.15051

None	None	None	0.15051	0.15051	0.15051	0.15051
Nc	Poly.	Coating	Coating Agent 01		Coating Agent 02	
cc_1	cc_2	Type	$<Sh(LOGP_{ac1})_{cnj}>$	$<Sh(PSA_{ac1})_{cnj}>$	$<Sh(LOGP_{ac2})_{cnj}>$	$<Sh(PSA_{ac2})_{cnj}>$
Double	Mono	I	0.148	0.150	0.148	0.150
Single	Mono.	II	0.146	0.151	0.151	0.151
Single	Poly.	III	0.148	0.151	0.151	0.151
None	None	IV	0.151	0.151	0.151	0.151

Shannon's entropy scaling of MRN local structural information.

As we mentioned before the same kind of operators $Sh_k(D_k)$ can be used for different subsystems. Firstly, we calculated the parameters N_{ms} number of metabolites (m), or $D_{ks} = <L_{ins}>$ average in-degree, $D_{ks} = <L_{outs}>$ average out-degree for all metabolites in the MRN of the s^{th} organism. The calculation of these parameters was carried out with the software MI-NODES[21] developed by our group and verified with the software CentBin.[22] Next, by using **Equation 1** we also applied the same probability operator $p(D_k)$ to the structural descriptors of the and MRNs (D_{ks}). After that we obtained the values of respective entropy $Sh(D_{ks})$ descriptors $Sh(N_{ms})$, $Sh(L_{ins})$ and $Sh(L_{outs})$ of MRN of the s^{th} organism by using **Equation 2**. It is important to note that N_{ms}, L_{ins}, and L_{outs} are local node centralities of the MRNs.[16] Consequently, the entropies obtained $Sh(N_{ms})$, $Sh(L_{ins})$, and $Sh(L_{outs})$ are also local descriptors.[21] In **Table 3,** you can see also the names of the organisms, two-letter codes, and their respective values of $Sh(N_{ms})$, $Sh(L_{ins})$, and $Sh(L_{outs})$ for all the MRNs studied. These values have been calculated in this work by the first time for this set of MRNs.

Table 3. Shannon entropy information measures of MRN_s studied in this work.

MRN Ns	Org. Code	MRNs Shannon Entropy Information Measures				
		$Sh_3(N_m)$	$Sh_4(L_{in})$	$Sh_5(L_{out})$	$Sh(\boldsymbol{\pi}_1)$	$Sh_2(\boldsymbol{\pi}_2)$
1	AB	0.134	0.088	0.090	0.015	0.014
2	BS	0.112	0.024	0.026	0.016	0.014
3	CA	0.128	0.065	0.068	0.007	0.009
4	CJ	0.134	0.091	0.093	0.01	0.012
5	CQ	0.143	0.133	0.134	0.038	0.038
6	CT	0.142	0.129	0.130	0.017	0.018
7	DR	0.110	0.023	0.024	0.008	0.007
8	EC	0.112	0.022	0.023	0.008	0.008
9	EF	0.134	0.085	0.087	0.008	0.011
10	HI	0.127	0.058	0.059	0.016	0.013
11	HP	0.135	0.089	0.091	0.015	0.017
12	MB	0.132	0.085	0.087	0.008	0.009
13	MG	0.142	0.126	0.127	0.016	0.017
14	ML	0.132	0.084	0.085	0.008	0.009
15	MP	0.144	0.130	0.130	0.019	0.02
16	MT	0.123	0.054	0.056	0.015	0.014
17	NG	0.133	0.082	0.084	0.008	0.011

18	NM	0.134	0.087	0.089	0.009	0.012
19	PA	0.115	0.033	0.035	0.019	0.016
20	PG	0.132	0.088	0.090	0.008	0.011
21	PN	0.133	0.081	0.082	0.008	0.011
22	RC	0.119	0.042	0.044	0.017	0.015
23	SC	0.125	0.051	0.053	0.01	0.011
24	ST	0.133	0.082	0.084	0.01	0.011
25	TY	0.110	0.020	0.021	0.007	0.007
26	YP	0.124	0.059	0.061	0.01	0.013

Markov-Shannon entropy scaling of MRN high-order structural information.

In any case, N_{ms}, L_{ins}, and L_{outs} are local topological descriptors that only account for information of the node (metabolite in question) and the nodes directly linked to it direct precursors (educts) for the case of $<L_{ins}>$ and direct products (adducts) for the case of L_{out}.[16] Consequently, we also used Shannon operators of the type $Sh(D_k) = -p(D_k)\cdot logp(D_k)$ to quantify higher order structural information of the MRNs. However, in this particular case, the operator is not applied to the local descriptor *per se*. In this case we apply the operator to the probabilities obtained from a Markov Chain calculation. In so doing, we calculated the values of entropy Sh_k of k^{th} order for the s^{th} species. The Sh_k values measure the connectivity information in the MRN of the s^{th} species for all metabolites and their neighbors (substrates or products) placed at a distance (number of reactions) $\leq k$. In order to calculate these indices we applied the $Sh_k(D_k) = -p(D_k)\cdot logp(D_k)$ operator directly to the absolute probabilities $D_k = p_k(m,s)$. These values are the absolute probabilities $p_k(m,s)$ with which the m^{th} metabolite transforms into another metabolite (catabolism) and/or is the product (anabolism) of the different metabolic reactions in the MRNs of the s^{th} organism. The Markov matrix ${}^1\mathbf{\Pi}_s$ was used to calculate $p_k(m,s)$ values by means of a Matrix-vector multiplication operation $\mathbf{M}^k\cdot\mathbf{v}$ involving the k^{th} natural powers $\mathbf{M}^k$ of the original matrix $\mathbf{M}$. In the case of a Markov matrix this product is $({}^1\mathbf{\Pi}_s)^k\cdot\boldsymbol{\pi}_0$ a component of Chapman-Kolgomorov equation. We calculated only the two first powers $({}^1\mathbf{\Pi}_s)^1$ and $({}^1\mathbf{\Pi}_s)^2$of the Markov matrix ${}^1\mathbf{\Pi}_s$ of each one of the s^{th} bacteria species. After that we made the products $\boldsymbol{\pi}_{1s} = ({}^1\mathbf{\Pi}_s)^1\cdot\boldsymbol{\pi}_0$ and $\boldsymbol{\pi}_{2s} = ({}^1\mathbf{\Pi}_s)^2\cdot\boldsymbol{\pi}_0$. The resulting vectors $\boldsymbol{\pi}_{1s}$ and $\boldsymbol{\pi}_{2s}$ containing as elements the absolute probabilities $p_1(m,s)$ and $p_2(m,s)$ for each metabolite of the network. The values $p_1(m,s)$ are the absolute probabilities with which the m^{th} metabolite comes directly from and/or transforms directly into another metabolite (k = 1). The values $p_2(m,s)$ are the absolute probabilities with which the m^{th} metabolite comes directly from and/or transforms directly into intermediate metabolites that in turn came from and/or transform into a second product (k = 2). Finally, $Sh_k(\boldsymbol{\pi}_1)$ and $Sh_k(\boldsymbol{\pi}_2)$ values are calculated with the operators $Sh_k(D_k) = -p(D_k)\cdot logp(D_k) = Sh_k(D_k) = -p(p_k(m,s))\cdot logp(p_k(m,s))$ as the sum of these values of entropy for each m^{th} node (metabolite) in the MRN_s, see **Equation 3**. In **Table 3,** you can see also the names of the organisms, two-letter codes, and values of $Sh_k(\boldsymbol{\pi}_1)$ and $Sh_k(\boldsymbol{\pi}_2)$ for all the MRNs studied. These values have been calculated in this work by the first time for this set of MRNs. The specific formula used to calculate these values of $Sh(\boldsymbol{\pi}_1)$ and $Sh_k(\boldsymbol{\pi}_2)$ of MRNs is the following, please see details on the literature:[48]

$$Sh(\boldsymbol{\pi}_k) = -\sum_{m=1}^{m=mmax} p_k(m,s)\cdot\log p_k(m,s) \qquad (S3)$$

IF step for observed biological parameters. The first step to obtain the IFPTML model for DADNP systems was to defining and obtaining the values of the objective function. The objective function is the function we want to fit with a ML model using as input the vectors of descriptors for each case $\mathbf{D}_k$. The objective function is obtained very often after a mathematical transformation of the original theoretical or observed property of the

system under study. [23-25] In the present IFPTML model we have two sets observed values (v_{ij} and v_{nj}) and two sets of input vectors ($\mathbf{D}_{dk}$ and $\mathbf{D}_{nk}$) for the AD and NP subsystems (S_d and S_n) respectively. In addition, we found many different biological parameters c_{d0} and c_{n0}. For instance, we find properties like Minimal Inhibitory Concentration (MIC ($\mu g \cdot mL^{-1}$)) or Minimal Bactericide Concentration (MBC ($\mu g \cdot mL^{-1}$)), *etc*. Do not help to solve the problem the fact that the v_{ij} and v_{nj} values compiled are not exact numbers in many cases. Many reports in both dataset are of the type MIC ($\mu g \cdot mL^{-1}$) < 100. In addition, we have to consider that in order to obtain optimal DADNP systems we want to maximize some properties and minimize other. We conceptualize this fact with the parameter desirability. In **Table 1** we depict the values of desirability, cutoff, and other parameters used for the different biological properties.

Table 1. Selected examples of reference function, cutoff, and other values for DADNP subsystems

c_{n0} = NP Activity (Units)[a]	d	Cutoff	n_j	$n(f(v_{nj})=1/c_{n0})$	f_{ref}
IC_{50}(μM)	-1	89.98	164	96	0.59
MIC(μM)	-1	125.15	123	55	0.45
MBC(μM)	-1	173.03	9	1	0.11
MBE	1	0.96	4	2	0.5
c_{d0} = AD Activity (Units)	d	Cutoff	nj	$n(f(v_{ij})=1/c_{d0})$	f_{ref}
IC_{50} nM	1	100.0	1590	1554	0.98
MIC nM	-1	2500.0	11099	2859	0.26
MIC ug.mL-1	-1	50.0	92583	67059	0.72
MBC ug ml-1	-1	117.2	1349	1004	0.74
MBC uM	-1	187.3	332	209	0.63

[a] IC_{50}(μM) = Concentration that is required for 50% inhibition of the growth of the bacteria. MIC(μM) = Minimum inhibitory concentration, i.e., the minimum concentration required to prevent the visible growth of the bacteria. MBC(μM) = Minimum bactericidal concentration, i.e., the minimum concentration required to complete kill the bacteria.

The parameter desirability was set d = 1 or d = 1 when we want to maximize the value v_{ij} or v_{nj} respectively. Remember, the different AD and NP parameter has names or labels c_{d0} and c_{n0}, respectively. Examples of biological activity parameters with d = 1 are the Selectivity ratio, Inhibition(%), *etc*. Conversely, negative desirability d = -1 parameters are for instance MIC($\mu g \cdot mL^{-1}$), IC_{50}($\mu g \cdot mL^{-1}$), *etc*. These facts increase the uncertainty of the data and difficult the development of regression model. To say all, it is a common practice in drug discovery to use a cutoff value to split AD or even NP assays as promising or not. Consequently, in order to obtaining our final objective function we need to pre-process all observed v_{ij} and v_{nj} values to eliminate or minimize all inaccuracies. In addition, we need to re-scale v_{ij} and v_{nj} values to obtaining a dimensionless variable not affected by scales. Last, IF processing step for both parameters v_{ij} and v_{nj} allows to obtaining an objective function of the putative DADNP system. In the **Figure 2** we depict a workflow summarizing all the steps of information flow (variable scaling, fusion, processing, *etc*.) of the IFPTML algorithm used here.

Firstly, we re-scaled the original parameters v_{ij} and v_{nj} to obtain the corresponding Boolean (dummy) functions $f(v_{ij})_{obs}$ and $f(v_{nj})_{obs}$. The scaling of v_{ij} was as follow: $f(v_{ij})_{obs} = 1$ when v_{ij} > cutoff and d = 1 or v_{ij} < cutoff and desirability d = -1, $f(v_{ij}) = 0$ otherwise. By analogy, v_{nj} scaling was: $f(v_{nj})_{obs} = 1$ when v_{nj} > cutoff and d = 1 or v_{nj} < cutoff and d = -1, $f(v_{ij}, v_{nj}) = 0$ otherwise. The values $f(v_{ij})_{obs} = 1$ and $f(v_{nj})_{obs} = 1$ points to an strong desired effect of both the AD and the NP over the target bacteria.[10] Accordingly, the objective function was defined as follow $f(v_{ij}, v_{nj})_{obs} = f(v_{ij})_{obs} \cdot f(v_{nj})_{obs}$. Then as result of the IF-scaling $f(v_{ij}, v_{nj})_{obs}$ depends on the i^{th} AD compound, the n^{th} NP system, the c^{th} CA used as coat, the s^{th} specie of assay, and the j^{th} sets of assay conditions. Otherwise, $f(v_{ij}, v_{nj})_{obs} = 0$, meaning that at least one of the previous conditions fail.

IF step for function of reference. Once we defined the objective function we must to define the input variables of the IFPTML model. The first and unique of his kind input variable of this model is the function of reference $f(v_{ij}, v_{nj})_{ref}$. In IFPTML models $f(v_{ij}, v_{nj})_{ref}$ place an special role because this function represent the expected probability $f(v_{ij}, v_{nj})_{ref} = p(f(v_{ij}, v_{nj})_{ref} = 1)$ of obtaining the desired level of activity for a property obtained from already known systems. The model starts with the value this function for an already known system or sub-set of systems used as reference. Later the IFPTML model adds the effect of deviations (perturbations) of the query system from the systems of reference (PT ideas, see next section). Consequently, $f(v_{ij}, v_{nj})_{ref}$ is also a function based on observed (not predicted) outcomes. In this work the reference function for putative DADNP systems was obtained by IF-scaling of the original v_{ij} and v_{nj} values as well. In the previous section we explained how to transform these values into the $f(v_{ij})_{obs}$ and $f(v_{nj})_{obs}$ functions. Once we get the values of these functions for all cases on the AD and NP datasets we are in position of counting the number of positive outcomes $n(f(v_{ij}) = 1)$ and $n(f(v_{nj}) = 1)$. Next we can divide these values by the total number of cases obtaining the functions of reference (expected probabilities) for the AD and NP systems alone. These values are $f(v_{ij})_{ref} = p(f(v_{ij})_{ref} = 1) = n(f(v_{ij})_{ref} = 1)/n_j$ and $f(v_{nj})_{ref} = p(f(v_{nj})_{ref} = 1) = n(f(v_{nj})_{ref} = 1)/n_j$. From this, the calculation of the function of reference is straightforward to realize as the product of the probabilities for each subsystem $f(v_{ij}, v_{nj})_{ref} = p(f(v_{ij}, v_{nj})_{ref} = 1) = p(f(v_{ij})_{ref} = 1) \cdot p(f(v_{nj})_{ref} = 1)$. The function of reference used here is then another expression of the IF step (union) of both AD and NP datasets.

$$f(v_{ij}, v_{nj}, v_{sj})_{ref} = f(v_{ij})_{ref} \cdot f(v_{nj})_{ref} \cdot f(v_{sj})_{ref} \quad (S4)$$

Table 2. Data pre-processing functions and cases distribution

Function	Value	Details
$f(v_{ij}, v_{nj}, v_{sj})_{obs}$	1	Plausible positive outcome in j^{th} assay *vs*. s^{th} species with MRNs for putative $DADNP_{in}$ formed by i^{th} AD and n^{th} NP
	0	Plausible negative outcome in j^{th} assay *vs*. s^{th} species with MRNs for putative $DADNP_{in}$ formed by i^{th} AD and n^{th} NP
$f(v_{ij}, v_{nj}, v_{sj})_{ref}$	0-1	Probability with which the systems of the same class of the system of reference have a positive outcome in j^{th} assay *vs*. s^{th} species with MRNs for putative $DADNP_{in}$ formed by i^{th} AD and n^{th} NP
$f(v_{ij})_{obs}$	1	Positive outcome for i^{th} AD in j^{th} assay
	0	Negative outcome for i^{th} AD in j^{th} assay
$f(v_{nj})_{obs}$	1	Positive outcome for n^{th} NP in j^{th} assay
	0	Negative outcome for n^{th} NP in j^{th} assay
$f(v_{sj})_{obs}$	1	s^{th} MRN belongs to a Human pathogen specie
	0	s^{th} MRN does not belongs to a Human pathogen specie
$f(v_{jns})_{obs}$	1	j^{th} AD and NP n^{th} specie = MRN s^{th} specie
$f(v_{jns})_{obs}$	0	j^{th} AD and NP n^{th} specie ≠ MRN s^{th} specie
$f(set)_{obs}$	1	Cases used to train the model (set = t)
	0	Cases used to validate the model (set = v)
Total	-	All cases in data set

PT data preprocessing. In addition to $\mathbf{D}_{dk}$ and $\mathbf{D}_{nk}$ vectors this IFPTML analysis also considers the vectors $\mathbf{c}_{dj}$ and $\mathbf{c}_{nj}$ having as components the non-numerical experimental conditions and/or labels for AD and NP assays. Using the $Sh(D_{dk})$ and $Sh(D_{nk})$ values explained before we can calculate the PTOs of the AD and NP assays in order to account for this additional information. We used here two kinds of PTOs. The first is the AD and NP MA PTOs (**Equation S3** and **Equation S4**). They are used to account for AD and NP structural and assay information. The PTOs $\Delta Sh(D_{dk})$ and $\Delta Sh(D_{nk})$ codify AD and NP structural and/or physichochemical information on the parameters $Sh(D_{dk})$ and $\Delta Sh(D_{nk})$, respectively. The PTOs $\Delta Sh(D_{dk})$ and $\Delta Sh(D_{nk})$ codify AD and NP biological assay information with the parameter $<Sh(D_{dk})_{\mathbf{cdj}}>$ and $<Sh(D_{nk})_{\mathbf{cnj}}>$, respectively. They are the values of the are average operator < > for $Sh(D_{dk})$ and $\Delta Sh(D_{nk})$ running over all cases with the same sub-set of experimental conditions $\mathbf{c}_{dj}$ and $\mathbf{c}_{nj}$, respectively. Consequently, they should give specific values for each assay with at least one different element (experimental condition) of the vector $\mathbf{c}_{dj}$ or $\mathbf{c}_{nj}$. In consequence, they can be used to indicate which assay we are using.[18, 20, 26-28] Please, see values of $<Sh(D_{dk})_{\mathbf{cdj}}>$ and $<Sh(D_{nk})_{\mathbf{cnj}}>$ values in **Table S1** of Supporting Information file SI00.doc. The second type of PTOs used is the AD-NP coat MA Balance (MAB) PTO $\Delta\Delta Sh(D_{ca1}, D_{ca2}, D_{dk})$ (**Equation S5**). The MAB PTO accounts for the similarities on the information of AD *vs*. the NP coating agent. PTOs based directly on MA and/or linear and non-linear transformations of MA have been used for AD and NP discovery before.[18-20] However, the MAB is reported here for the first time (see Results and Discussion). The MAS is another expression of the combined IF+PT additive processing of both AD and NP datasets.

$$\Delta Sh(D_{dk}) = \Delta Sh(D_{dk}) - \langle Sh(D_{dk})_{c_{dj}} \rangle \qquad (S3)$$

$$\Delta Sh(D_{nk}) = \Delta Sh(D_{nk}) - \langle Sh(D_{nk})_{c_{nj}} \rangle \qquad (S4)$$

$$\Delta\Delta Sh(D_{nk}) = \Delta Sh(D_{dk}) - [\Delta Sh(D_{ca1}) + \Delta Sh(D_{ca2})] \qquad (S5)$$

IF step and design of training and validation subsets

All the cases of a dataset should assigned to training (set = t) or validation (set = v) series. The procedure of cases sampling used should be random, representative, and stratified.[29] As an additional condition our sampling should take into consideration the IF-scaling process. Firstly, we downloaded the AD activity dataset from ChEMBL which has random uploads from many sources worldwide and from randomly selected journal papers dealing with NP antibacterial activity. Next, we organized all the cases based on the following labels c_{d0}, c_{d1}, c_{d2}, c_{n0}, c_{n1}, and c_{n2}. All cases were ordered by sorting the labels from A to Z (remember these are non-numeric variables in nature). The order of priority of the labels on the process of ordering was c_{d0} => c_{n0} => c_{d1} => c_{n1} => c_{d2} => c_{n2}. It means that firstly we ordered the cases by c_{d0}, next by c_{n0}, and so on. This priority order takes into account the IF process by alternating labels from both AD and NP datasets. After that 3 out of each 4 cases were assigned to set = t and 1 out of 4 set = v from top to down of the list. This increases the probability that almost all the levels of each label are represented in set = t and set = v (stratified sampling). This also increases the probability that almost all levels of each label are in a proportion 3/4 in set = t and 1/3 in set = v (representative sampling). The 75% vs. 25% proportion between set = t and set = v is not the only but is very commonly used. [29]

IFPTML-ANN model variable sensitivity analysis

All this confirms the strength of the linear hypothesis used here. However, the values of Sn and Sp obtained have still a margin from improvement. Consequently, we increased the number of variables in the PTML-LNN models from 9 to 10 and 11. In this study, no significant change was detected. As a result, we also considered the non-linear hypothesis here as a way to increase Sp and Sn values. In fact, the IFPTML-MLP 9:9-8-1:1 model with 9 neurons in input layer (input variables) and 8 neurons in the hided layer showed more balanced SN and Sp ≈ 88% values. See summary of results in **Table 7**. See detailed results for all cases in Supporting Information file SI01.xlsx. More complicated IFPTML-MLP2 models with two hidden layers do not show significant improvement.

Taking all the previous factors into consideration we were pivoting between IFPTML-LDA or IFPTML-LNN model and IFPTML-MLP model. One important point in favor of the IFPTML-MLP model is his notably higher value of AUROC = 0.94 and the notably better behavior (shape) of the ROC curve with respect to the IFPTML-LNN linear models and RND classifier behavior, see **Figure 3**. Once again Occam's razor comes to rescue herein by checking if minimal necessary features (no more no less) are being considered.[30] We carried out a feature sensitivity analysis on the input variables. In **Figure 4** we can see that the IFPMTL-LNN models include or important parameters EGS point of view. Almost all parameters have a significant contribution with Sensitivity ≥ 1.[29] However, in most cases it is only marginally higher with Sensitivity ≈ 1.00 – 1.08. On the other hand, the IFPTML-MLP model also includes the important parameters according to EGS criteria but they have notably higher values of Sensitivity ≈ 1.00 – 2.52. MLP2 has even higher values of feature Sensitivity ≈ 1.00 – 3.31 but as we mentioned before there is no gain on Sp and Sn values to justify the notably higher complexity of the model, see **Figure 4**.

IFPTML	MLP_I		MLP_{II}		MLP_{III}		RBF		LNN	
Variables	t	v	t	v	t	v	t	v	t	v
$f(\mathbf{c}_{d0},\mathbf{c}_{n0})_{ref}$	1.0	1.0	1.1	1.1	1.2	1.2	1.0	1.0	1.0	1.0
$\Delta Sh(ALOGP_i)_{cj}$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0
$\Delta Sh(Lins)_{cs}$	4.8	4.7	5.1	5.0	5.6	5.5	1.0	1.0	6.5	6.4
$\Delta Sh(Louts)_{cs}$	3.9	3.8	5.3	5.2	6.1	6.0	1.0	1.0	6.2	6.1
$\Delta Sh(AMVn)_{cn}$	3.1	3.0	5.5	5.4	5.6	5.5	1.0	1.0	1.0	1.0
$\Delta Sh(APSn)_{cn}$	3.8	3.8	2.7	2.6	3.7	3.6	1.1	1.1	1.1	1.1
$\Delta Sh(t)_{cn}$	1.3	1.3	1.3	1.3	1.3	1.3	1.0	1.0	1.0	1.0
$\Delta\Delta Sh(PSA_i,PSA_{CA1},PSA_{CA2})_{cj,cn}$	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0	1.0

Figure S1. IFPTML-ANN models sensitivity analysis

REFERENCES (Supporting Information)

1. Nocedo-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, 1109-1120.
2. 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, 193-204.
3. Sushko, I.; Novotarskyi, S.; Korner, R.; Pandey, A. K.; Rupp, M.; Teetz, W.; Brandmaier, S.; Abdelaziz, A.; Prokopenko, V. V.; Tanchuk, V. Y.; Todeschini, R.; Varnek, A.; Marcou, G.; Ertl, P.; Potemkin, V.; Grishina, M.; Gasteiger, J.;

Schwab, C.; Baskin, II; Palyulin, V. A.; Radchenko, E. V.; Welsh, W. J.; Kholodovych, V.; Chekmarev, D.; Cherkasov, A.; Aires-de-Sousa, J.; Zhang, Q. Y.; Bender, A.; Nigsch, F.; Patiny, L.; Williams, A.; Tkachenko, V.; Tetko, I. V., Online chemical modeling environment (OCHEM): web platform for data storage, model development and publishing of chemical information. *J Comput Aided Mol Des* **2011**, 25, 533-54.

4. Ruparelia, J. P.; Chatterjee, A. K.; Duttagupta, S. P.; Mukherji, S., Strain specificity in antimicrobial activity of silver and copper nanoparticles. *Acta Biomater* **2008**, 4, 707-16.

5. Pramanik, A.; Laha, D.; Bhattacharya, D.; Pramanik, P.; Karmakar, P., A novel study of antibacterial activity of copper iodide nanoparticle mediated by DNA and membrane damage. *Colloids and surfaces. B, Biointerfaces* **2012**, 96, 50-5.

6. Azam, A.; Ahmed, A. S.; Oves, M.; Khan, M. S.; Habib, S. S.; Memic, A., Antimicrobial activity of metal oxide nanoparticles against Gram-positive and Gram-negative bacteria: a comparative study. *International journal of nanomedicine* **2012**, 7, 6003-9.

7. Azam, A.; Ahmed, A. S.; Oves, M.; Khan, M. S.; Memic, A., Size-dependent antimicrobial properties of CuO nanoparticles against Gram-positive and -negative bacterial strains. *Int J Nanomedicine* **2012**, 7, 3527-35.

8. Hossain, S. T.; Mukherjee, S. K., Toxicity of cadmium sulfide (CdS) nanoparticles against Escherichia coli and HeLa cells. *Journal of hazardous materials* **2013**, 260, 1073-82.

9. Botequim, D.; Maia, J.; Lino, M. M.; Lopes, L. M.; Simoes, P. N.; Ilharco, L. M.; Ferreira, L., Nanoparticles and surfaces presenting antifungal, antibacterial and antiviral properties. *Langmuir* **2012**, 28, 7646-56.

10. Taglietti, A.; Diaz Fernandez, Y. A.; Amato, E.; Cucca, L.; Dacarro, G.; Grisoli, P.; Necchi, V.; Pallavicini, P.; Pasotti, L.; Patrini, M., Antibacterial activity of glutathione-coated silver nanoparticles against Gram positive and Gram negative bacteria. *Langmuir* **2012**, 28, 8140-8.

11. Hossain, S. T.; Mukherjee, S. K., CdO nanoparticle toxicity on growth, morphology, and cell division in Escherichia coli. *Langmuir* **2012**, 28, 16614-22.

12. Premanathan, M.; Karthikeyan, K.; Jeyasubramanian, K.; Manivannan, G., Selective toxicity of ZnO nanoparticles toward Gram-positive bacteria and cancer cells by apoptosis through lipid peroxidation. *Nanomedicine* **2011**, 7, 184-92.

13. Inbaraj, B. S.; Kao, T. H.; Tsai, T. Y.; Chiu, C. P.; Kumar, R.; Chen, B. H., The synthesis and characterization of poly(gamma-glutamic acid)-coated magnetite nanoparticles and their effects on antibacterial activity and cytotoxicity. *Nanotechnology* **2011**, 22, 075101.

14. Hu, X.; Cook, S.; Wang, P.; Hwang, H. M., In vitro evaluation of cytotoxicity of engineered metal oxide nanoparticles. *Sci Total Environ* **2009**, 407, 3070-2.

15. Zhao, Y.; Chen, Z.; Chen, Y.; Xu, J.; Li, J.; Jiang, X., Synergy of non-antibiotic drugs and pyrimidinethiol on gold nanoparticles against superbugs. *Journal of the American Chemical Society* **2013**, 135, 12940-3.

16. Jeong, H.; Tombor, B.; Albert, R.; Oltvai, Z. N.; Barabasi, A. L., The large-scale organization of metabolic networks. *Nature* **2000**, 407, 651-4.

17. Hsu, D. D. Chemicool Periodic Table. http://www.chemicool.com/ (October 4, 2013),

18. Santana, R.; Zuluaga, R.; Gañán, P.; Arrasate, S.; Onieva, E.; González-Díaz, H., Designing nanoparticle release systems for drug-vitamin cancer co-therapy with multiplicative perturbation-theory machine learning (PTML) models. *Nanoscale* **2019**, 11, 21811-21823.

19. 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, 13471-13483.

20. 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 Cotherapy. *Molecular pharmaceutics* **2020**, 17, 2612-2627.

21. Duardo-Sanchez, A.; Munteanu, C. R.; Riera-Fernandez, P.; Lopez-Diaz, A.; Pazos, A.; Gonzalez-Diaz, H., Modeling complex metabolic reactions, ecological systems, and financial and legal networks with MIANN models based on Markov-Wiener node descriptors. *Journal of chemical information and modeling* **2014**, 54, 16-29.

22. Junker, B. H.; Koschutzki, D.; Schreiber, F., Exploration of biological network centralities with CentiBiN. *BMC bioinformatics* **2006**, 7, 219.

23. 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, 4492-4503.
24. Xia, R.; Kais, S., Quantum machine learning for electronic structure calculations. *Nature communications* **2018**, 9, 4195.
25. 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, 18526-18535.
26. 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, 14686-94.
27. 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, 10623-30.
28. Urista, D. V.; Carrue, D. B.; Otero, I.; Arrasate, S.; Quevedo-Tumailli, 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.
29. Hill, T.; Lewicki, P., *Statistics: Methods and Applications*. 1st edition ed.; StatSoft, Inc.: 2005; p 800.
30. Van Den Berg, H. A., Occam's razor: from Ockham's via moderna to modern data science. *Science progress* **2018**, 101, 261-272.