Supplementary material:

Selection of Unique Molecules for Cancer Treatment by Distance-Based Method: Hypericin Effect on Respiratory Chain

Denis Horvath[1], Silvia Tomkova[2] and Veronika Huntosova[1]

[1] Center for Interdisciplinary Biosciences, Technology and innovation park, P.J. Šafárik University in Košice, Jesenná 5, 041 54 Košice, Slovakia

[2] Department of Biophysics, Institute of Physics, Faculty of Science, P.J. Šafárik University in Košice, Jesenná 5, 041 54 Košice, Slovakia

Summary of supplementary material:

Purpose: The main purpose of the material is to help understand the impact of the presence of deviations that we introduce using categorical variables. The specific context and data from cellular responses to selected molecules are important for the investigation and conclusions.

Methodology: Monte Carlo simulation in the analysis of the potential effects of uncertainty resulting from the use of categorical data. The output of the simulation is the mean value of sensitivity to perturbations. It is a measure obtained by random sampling with a large number of randomized experiments.

Findings: Range of possible changes in distances - simulation outputs in response to randomized simulated inputs.
Perspectives: Improved methodology towards a more appropriate categorization methods.

In the main text of this publication, the Minkowski variant and the Euclidean case of distance (as a special case) are used to compare the remoteness of vectors of categorical variables. There are certainly potential uncertainties and inaccuracies associated with categorical simplifications and comparisons. We comment on them in more detail in this supplementary material. Here we characterize the causes not only qualitatively but mainly quantitatively. As the shortcomings resulting from the categorical description may be of a relatively diverse nature, it seems useful to divide them into two main groups as follows:

(I) The first set of problems is caused by incorrect categorization. Here are two groups of problems of this type:

(Ia) For example, if a value is assigned to a cellular response 1 instead of 0, and so on. This may be due to shortcomings in specific experiments, altered experimental conditions or, to the extreme, shortcomings in the contextual knowledge of the experts. In comparison with the number of cases investigated by specialists as well as the knowledge gained by them, these are most likely only isolated cases, the impact of which is very difficult to quantify.

(Ib) The inadequate number of categorical values make it hard to understand and properly characterize biological reality. Some nuances of the original features not taken into consideration by the model are found in scientific developments in knowledge. Through natural data collection processes, the number of categorical values possibly increases as well.

Example: When we look at apoptosis in a broader biological context, we can see that it is a highly regulated process with early, intermediate, and late phases. As a result, a more detailed classification and understanding of the stages can aid in broadening and improving the overall categorization and understanding.

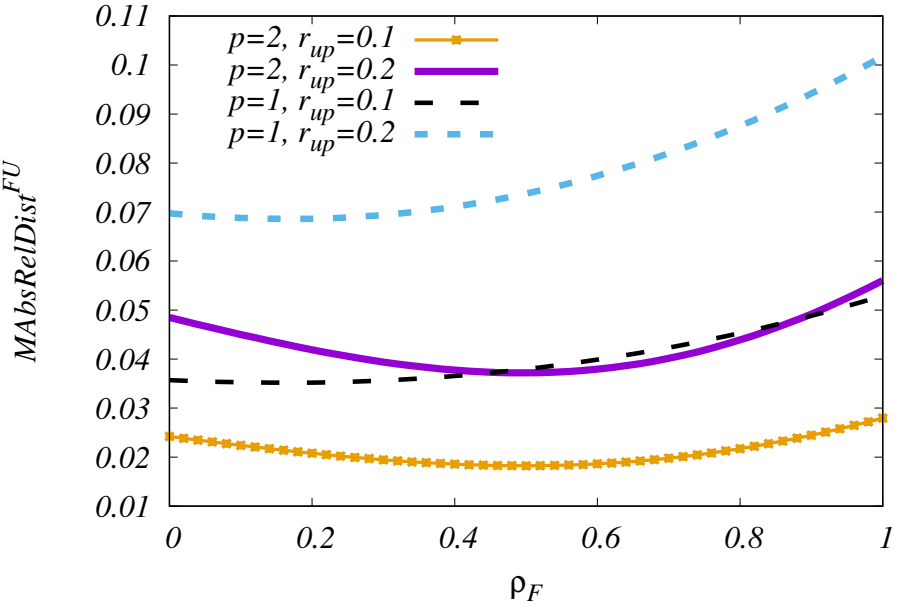

Figure S1: Four pairs of p, r_{up} were selected to obtain ρ_F dependencies. Monte Carlo simulations yielded the following results: (i) The presence of perturbations affects distances. (ii) Results depend non-monotonically on ρ_F. (iii) If $p = 2$ is used, then the curvature of $MAbsRelDist^{FU}(\rho_F)$ around the local minimum becomes more pronounced.

(II) Suppose we have a sufficiently reasonable integer categorization of cellular behavior at the beginning. For example, the one with values $\{0, 1, 2\}$. In addition, let there be at most corrections of the order of one between the integer description on one side and the three real improvements. This means the assumption of the existence of a modification towards a trio of real numbers, for example $\{0.11, 0.93, 2.24\}$. By characterizing differences or relatedness, distances are reasonable descriptors of multidimensional systems, including cellular responses. Improvements can be expected for modified real-value descriptions that reflect somewhat more thoroughly the role of selected molecules.

In the rest of the supplementary material, we use simulations to not only describe but also analyze the mentioned earlier point (**II**).

To better understand the impact of integer descriptions and descriptions using actual values in the context of available data, we performed some illustrative stochastic simulations for quantity $dist_{ij}^{FU}$ and data available (see the main text). We performed Monte Carlo-type simulations in which r_{jkl}^{U} and r_{jkl}^{F} are random perturbations from categorical data. Perturbations are represented by pseudorandom numbers. We performed simulations using numerical values drawn from the uniform distribution in the interval $[-r_{up}, r_{up}]$ for simplicity. In this case, the value $r_{up} \in [0, 1/2]$ expresses the degree of uncertainty. Then, to obtain the respective mean values, the pseudo random set $\{ r_{jkl}^{F}, r_{jkl}^{U} \}_{j=0,\, k=0}^{j=4,\, k=4}$ with the independent pair r_{jkl}^{F}, r_{jkl}^{U} for $l \in \{1, 2, \ldots, n_{rand}\}$ must be generated.

Since we want to evaluate the sensitivity of $dist_{ij}^{FU}$ due to random changes, we first introduce some auxiliary power forms of distance not only within the original data (see Eq.(2) of the main text)

$$D_{ij}^{F} = \frac{1}{7} \sum_{k=0}^{6} \left| \, Dat_{ik}^{F} - Dat_{jk}^{F} \, \right|^{p}, \tag{S.1}$$

$$D_{ij}^{U} = \frac{1}{7} \sum_{k=0}^{6} \left| \, Dat_{ik}^{U} - Dat_{jk}^{U} \, \right|^{p}$$

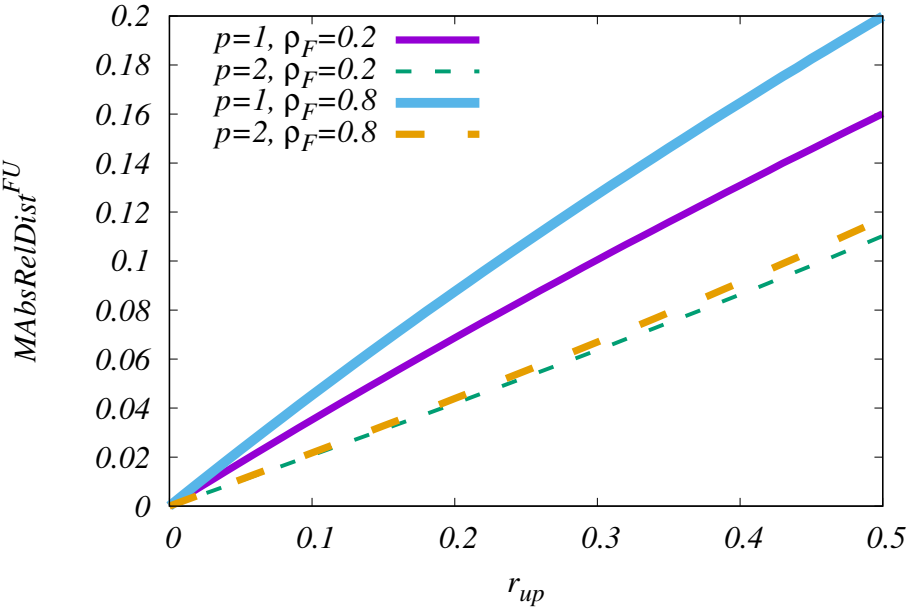

Figure S2: Results of Monte Carlo simulations; $MAbsRelDist^{FU}$ represents mean of the relative distance errors. These are mainly linearly dependent upon r_{up} (interval determination parameter). Four parametric examples illustrate how the obtained dependencies probably change when p and rho_F are altered.

but also for

$$\tilde{D}^F_{ijl} = \frac{1}{7}\sum_{k=0}^{6} \left| \, Dat^F_{ik} + r^F_{ikl} - Dat^F_{jk} - r^F_{jkl} \, \right|^p , \tag{S.2}$$

$$\tilde{D}^U_{ijl} = \frac{1}{7}\sum_{k=0}^{6} \left| \, Dat^U_{ik} + r^U_{ikl} - Dat^U_{jk} - r^U_{jkl} \, \right|^p$$

obtained [1] for real-valued entities $Dat^F_{ik} + r^F_{ikl}$, $Dat^F_{jk} + r^F_{jkl}$, $Dat^U_{ik} + r^U_{ikl}$, $Dat^U_{jk} + r^U_{jkl}$. Then exponentiating of D^F_{ij} and D^U_{ij} to $(1/p)$, as seen in Eq.(3) in the main text, gives us the true distance as follows:

$$dist^{FU}_{ij} = \left[\, \rho_F D^F_{ij} + (1-\rho_F) D^U_{ij} \, \right]^{1/p} . \tag{S.3}$$

Then the random perturbations are reflected in

$$\widetilde{dist}^{FU}_{ijl} = \left[\, \rho_F \tilde{D}^F_{ijl} + (1-\rho_F) \tilde{D}^U_{ijl} \, \right]^{1/p} .$$

To evaluate the impact of random effects, a specific metric

$$MAbsRelDist^{FU} = \frac{1}{10\, n_{rand}} \sum_{l=1}^{n_{rand}} \sum_{i=1}^{3} \sum_{j=i+1}^{4} \frac{\left| \, dist^{FU}_{ij} - \widetilde{dist}^{FU}_{ijl} \, \right|}{dist^{FU}_{ij} + \widetilde{dist}^{FU}_{ijl}} \tag{S.4}$$

is chosen we can call *mean relative distance error*. Its name reflects its computational structure [2] The metric from Eq.(S.4) still depends on ρ_F and p parameters. The result also includes the norm $1/(10\, n_{rand})$. Of course, at the theoretical limit $n_{rand} \to \infty$, the genuine mean value becomes attained. Two dependencies have been extracted by simulating random process for the experimental categorical

[1] In the above formulas, we use a tilde as a distinguishing mark for the perturbed cases.

[2] The first indicator, used to emphasize the role of mean is $M\ldots$; Next is $\ldots Abs\ldots$. This emphasizes the application of absolute value. When comparing perturbed and intact distances the relative change represented by $\ldots Rel\ldots$ was applied. Finally, there is the fact that we are dealing with matrix of distances $\ldots Dist^{FU}_{ij}$.

data. The results of numerical study are shown in Fig.S1 and Fig.S2. Important for stabilizing these results in terms of $MAbsRelDist^{FU}$ is that we used $n_{rand} = 500000$ steps within the Monte Carlo methodology. The significant finding of our numerical methodology, as shown in Fig.S1, is that the outputs, represented by $MAbsRelDist^{FU}$, are closely delimited to the range $[0.02, 0.10]$ for the four explored alternatives $\{p, r_{up}\} \in \{\{1, 0.1\}, \{1, 0.2\}, \{2, 0.1\}, \{2, 0.2\}\}$. Our numerical findings and the smallness of the recorded relative changes are relatively unique. This is especially due to the fact that all components are perturbed simultaneously, which means the application of global sensitivity analysis.

The increased range r_{up} has a nearly linear influence on the output, as seen in Fig.S2. The slopes are smaller than one in all four cases analyzed, indicating low sensitivity. Surprisingly, as the p parameter is increased from one to two, the sensitivity decreases, which is an unexpected result. Rather low sensitivity findings appear impressive due to the simultaneous perturbation of all data components. We are unable to answer the question of how to estimate the appropriate distribution function of the random effects in the particular circumstances in this material, or in general. As a result, we went with a one-parameter uniform distribution with the free parameter r_{up}.