LJMU Research Online

Cassano, A, Marchese Robinson, RL, Palczewska, A, Puzyn, T, Gajewicz, A, Tran, L, Manganelli, S and Cronin, MTD

 Comparing the CORAL and random forest approaches for modelling the in vitro cytotoxicity of silica nanomaterials

http://researchonline.ljmu.ac.uk/id/eprint/5267/

Article

1 **Comparing the CORAL and Random Forest approaches for modelling the *in vitro***
2 **cytotoxicity of silica nanomaterials**

3 Antonio Cassano[a], Richard L. Marchese Robinson[a], Anna Palczewska[b], Tomasz Puzyn[c], Agnieszka Gajewicz[c],
4 Lang Tran[d], Serena Manganelli[e], Mark T. D. Cronin[*a]

5 *[a]School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University, Byrom Street, Liverpool, L3*
6 *3AF, England.*

7 *[b]University of Leeds, School of Geography, Leeds, England.*

8 *[c]Laboratory of Environmental Chemistry, University of Gdansk, Wita Stwosza 63, 80-308 Gdansk, Poland.*

9 *[d]Institute of Occupational Medicine, Edinburgh, Midlothian, Scotland.*

10 *[e]IRCSS-Istituto di Ricerche Farmacologiche Mario Negri, Via Giuseppe La Masa, 19, 20156, Milan, Italy.*

11

12 **Summary**

13 Nanotechnology is one of the most important technological developments of the twenty-first century. In silico

14 methods such as quantitative structure-activity relationships (QSARs) to predict toxicity promote the safe-by-

15 design approach for the development of new materials, including nanomaterials. In this study, a set of cytotoxicity

16 experimental data corresponding to 19 data points for silica nanomaterials was investigated to compare the widely

17 employed CORAL and Random Forest approaches in terms of their usefulness for developing so-called "nano-

18 QSAR" models. "External" leave-one-out cross-validation (LOO) analysis was performed to validate the two

19 different approaches. An analysis of variable importance measures and signed feature contributions for both

20 algorithms was undertaken in order to interpret the models developed. CORAL showed a more pronounced

21 difference between the average coefficient of determination (R^2) between training and LOO (0.83 and 0.65 for

22 training and LOO respectively) compared to Random Forest (0.87 and 0.78 without bootstrap sampling, 0.90 and

23 0.78 with bootstrap sampling), which may be due to overfitting. The aspect ratio and zeta potential from amongst

24 the nanomaterials' physico-chemical properties were found to be the two most important variables for the Random

25 Forest and the average feature contributions calculated for the corresponding descriptors were consistent with the

26 clear trends observed in the dataset: less negative zeta potential values and lower aspect ratio values were

27 associated with higher cytotoxicity. In contrast, CORAL failed to capture these trends.

28

29 *Keywords:* pseudo-SMILES; nano-QSAR; CORAL software; Random Forest; silica nanoparticle; variable
30 importance; feature contribution

31 [*]Corresponding author at: School of Pharmacy and Biomolecular Sciences, Liverpool John Moores University,
32 Byrom Street, Liverpool, L3 3AF, England. *Email address: M.T.Cronin@ljmu.ac.uk*

1. Introduction

Nanotechnology, which may be defined as the technological application of engineered nanomaterials [1], is considered to be one of the most important technological developments of the 21st century [2, 3] so much so that the term "nano-revolution" has been used to describe the growth of this industry [4]. Nanotechnology is able to produce engineered nanomaterials having new or enhanced physico-chemical properties compared to the bulk material. However, some of these properties, e.g. high surface area to volume ratio, are potentially dangerous to humans [5-9]. *In silico* methods (e.g. (Q)SAR, grouping and read-across) promote the safe-by-design approach for the development of new nanomaterials by studying the relationship between the nanomaterials' "structures" and their biological effects [10, 11]. Since nanomaterials are complex [12], typically polydisperse, particulate materials, the concept of a "structure" in this context should not be confused with a single molecular structure but rather a description of the nanomaterial in terms of its measurable physico-chemical characteristics [9, 13] such as the composition of different components, aspect ratio etc. In this regard, the development of nanomaterial quantitative structure-activity relationships ("nano-QSAR") may offer an effective alternative to experimental testing, since they may enable the prediction of (eco)toxicological effects of nanomaterials based on a knowledge of their chemical composition and, where necessary, other physico-chemical properties [14-16]. QSAR models can be classified as linear or non-linear depending on whether they were developed using a linear method, such as a multiple linear regression [17, 18], or a non-linear methods, such as support vector machines in combination with a non-linear kernel function [19, 23] or Random Forest [24, 25]. The aim of this study was to evaluate different approaches to build nano-QSAR models for a dataset comprising 19 cytotoxicity experimental data points for silica nanomaterials. We focused on silica nanomaterials mainly because of the availability of a novel experimental dataset for nanomaterials with a silica core and due to the widespread use of silica based nanomaterials in consumer products (http://www.nanotechproject.org/cpi/browse/nanomaterials/silicon-dioxide/). In this work, a comparison was made between two commonly used approaches to develop QSAR and nano-QSAR models: the linear approach implemented in the CORelation And Logic (CORAL) program, which optimises a (linear) regression model using a Monte Carlo search procedure [26], and Breiman's non-linear Random Forest algorithm [24, 25], implemented in the R randomForest package [27]. Our motivation to focus on these two modelling approaches reflects the fact that these have been used to build QSAR/QSPRs (and nano-QSAR/QSPRs) for a variety of different datasets, as illustrated in the number of publications summarised in the Supplementary Information (SI); for instance, 28 and 21 articles describing QSAR/QSPRs and

64   nano-QSAR/QSPRs studies using the COARL and Random Forest approaches respectively, were

65   published in 2015. (Quantitative structure-property relationships, or QSPRs, are analogous to QSARs,

66   but aim to predict non-biological properties.) However, to the best of our knowledge, these algorithms

67   have never previously been compared. Indeed, Random Forest has only twice before been used to model

68   nanomaterial effects [28, 29]. Hence, this investigation serves as a timely comparison of two widely

69   employed QSAR modelling approaches on a suitable dataset. In addition to comparing their predictive

70   performance, we performed a comparison in terms of model interpretability between a linear (CORAL)

71   and a non-linear (Random Forest) approach. In other words, the ability of the two selected approaches to

72   describe the toxicological trends of this dataset was evaluated.

73   **2.   Materials and Methods**

74   **2.1. Experimental data**

75   The experimental data used to develop the models correspond to a subset extracted from the dataset

76   generated during the MODENA COST Initiative (MODENA TD1204 COST ACTION dataset,

77   http://www.modena-cost.eu/Home.aspx). This dataset is provided in Table 1 and it is available

78   electronically in the SI. The dataset consists of 19 *in vitro* WST-1 cytotoxicity experimental data points

79   for uncoated silica nanomaterials. Briefly, WST-1 is a colorimetric assay for assessing cell metabolic

80   activity which is similar to the MTT assay, but which offers certain experimental advantages [30, 31].

81   The changes in metabolic activity measured using the WST-1 assay are considered a proxy for changes

82   in cell viability [32]. The data used in this work consist of 19 values for the negative logarithm of the

83   EC_{25} i.e. the concentration level which induces 25% of maximum response above the baseline after a

84   given treatment time. For modelling, nanomaterial concentrations, hence the corresponding EC_{25} values,

85   were expressed as surface area of nanomaterial per millilitre (i.e. mm^2/ml), in keeping with guidance from

86   the Organisation for Economic Co-operation and Development [33]. Cytotoxicity data range from -1.299

87   to 0.483, with no values between -0.822 and -0.394 i.e. the data cluster at low and high activities as shown

88   in Figure 1. Furthermore, from the original dataset we selected five variables based on our expert

89   judgement expected to explain variability in these activities: treatment time and cell type are related to

90   the experimental conditions adopted in the assay protocol, whereas average size, aspect ratio and zeta

91   potential are measured physico-chemical properties of silica nanomaterials. Specifically, since CORAL

92   is only able to handle a maximum of five variables, less significant descriptors were discarded. The full

93   list of descriptors can be found in the SI.

2.2. Evaluation approach

We adopted an "external" leave-one-out (LOO) cross-validation technique as a method to validate the considered modelling approaches. In brief, LOO is a special case of cross-validation [34-36], where the number of folds equals the number of instances in the data set. In other words, the learning algorithm is applied once for each instance, using all other instances as a training set and using the selected instance as a single-item test set. To this respect, for the dataset used in this work which comprises 19 instances, both CORAL and Random Forest algorithms were applied 19 times over all the instances in the dataset, each time considering 18 instances as training set and the remaining one as a test set in order to generate a given set of LOO results. (For both methods, five sets of LOO results were obtained as explained below.) By "external" LOO, we mean that all model development – including selection of descriptors and algorithm parameters or "hyperparameters" – was carried out exclusively using each LOO training set in turn i.e. the biological activity of the correspond test instance was not considered, to remove this potential source of optimistic bias from the results [37-39]. The coefficient of determination (R^2) and the root mean square error (RMSE) were here used as statistics for comparing the two approaches, according to the equations (1) and (2) [36], based on two n value vectors $y_1...y_n$ and $f_1...f_n$ which are associated with the experimental and predicted values respectively. N.B. (a) As the dataset comprised 19 instances, $n = 18$ for training sets whereas $n = 19$ for LOO. (b) In equations (1) and (2), and all subsequent equations in this manuscript, the "$^-$" character indicates the arithmetic mean (or "average") value over all the elements of a vector.

$$\text{Coefficient of determination} = R^2 = 1 - \frac{\sum_{i=1}^{n}(y_i - f_i)^2}{\sum_{i=1}^{n}(y_i - \bar{y})^2} \tag{1}$$

$$\text{Root mean square error} = RMSE = \sqrt{\frac{\sum_{i=1}^{n}(y_i - f_i)^2}{n}} \tag{2}$$

We performed the LOO validation technique five times, since the CORAL and Random Forest algorithms employ random selections during the model building phase, in order to obtain a more robust estimate of the performance of these methods. We selected five different seeds for each repetition with the Random Forest algorithm. We further selected five different dataset partitions for each repetition with CORAL. (Each time the CORAL software is run, it automatically generates a random seed that cannot be set by

123 the end user.) Each dataset partition corresponds to partitioning of a given LOO training set, following

124 removal of a single test set instance for "external" LOO validation, to yield an internal "test set" for

125 hyperparameter selection. Repeating modelling with CORAL five times in this fashion is broadly in

126 keeping with the recommended procedure for CORAL model optimisation and robustness evaluation [40-

127 42]. Further discussion of the CORAL hyperparameters which were optimised is presented under

128 "CORAL modelling". The average ($\bar{a}$), standard deviation (s) and standard error of the mean (SE) for the

129 LOO R^2 and RMSE statistics across the five different repetitions were calculated as shown in equations

130 (3), (4) and (5), considering the three general formulas based on a vector of m values i.e. a_1, a_2, …, a_m.

131 N.B. Here, m = 19 for training set results averaged for a single seed (or CORAL training set split), or 95

132 (19×5) for "global" training set results, whereas m = 5 for LOO results.

133

134 **Average** $= \bar{a} = \frac{1}{m}\sum_{i=1}^{m} a_i$ **(3)**

135

136 **Standard deviation** $= s = \sqrt{\frac{\sum(a-\bar{a})^2}{(m-1)}}$ **(4)**

137

138 **Standard error of the mean** $= SE = \frac{s}{\sqrt{m}}$ **(5)**

139

140 **2.3. Descriptor calculations**
141 As is further explained under "CORAL modelling", continuous numeric properties, such as zeta potential,

142 were converted into binary descriptors corresponding to labels applied to specific ranges: each descriptor

143 took a value of 1 (or 0) if the corresponding property value for a given instance was inside (or outside) of

144 this range. In the case of the discrete qualitative variable "Cell Type", each value was converted into a

145 binary descriptor: the descriptor took the value 1 (or 0) if the "Cell Type" for a given instance matched

146 (or did not match) the value associated with that descriptor. This was necessary since, as is explained

147 under "CORAL modelling", the CORAL algorithm can only work with binary descriptors. These binary

148 descriptors were used for both CORAL and Random Forest modelling. For the CORAL software, these

149 descriptors were represented implicitly i.e. the presence of a corresponding label in a pseudo-SMILES

150 (see "CORAL modelling") denotes a descriptor value of 1. For the implementation of Random Forest

151   used in the current work (see "Random Forest modelling"), these binary descriptors were represented as

152   an explicit bit-string.

153   **2.4. Correlated descriptors**

154   We tested the influence of correlated descriptors on model results for both CORAL and Random Forest

155   approaches by generating two versions of the original dataset, as shown in Table 2. In one case, after the

156   binary splitting was applied to each continuous numeric variable, a label was assigned to each of the

157   generated value ranges, which translates into two perfectly correlated descriptors for a given continuous

158   numeric variable. For instance, by splitting the "Treatment Time" variable into 24 and 48 hours, we

159   generated two labels, namely "A" and "B", which refer to the 24 and 48 hours' exposure respectively, in

160   the *in vitro* model. This results in the generation of two perfectly correlated descriptors, since they are

161   mutually exclusive. Specifically, when the "A" label is applicable (i.e. the "A" descriptor value is 1), the

162   "B" label must not be applicable (i.e. the "B" descriptor value is 0), and vice-versa, according to the fact

163   that, a single experimental result can only be associated with a single "Treatment Time" value. In the

164   second case, after the splitting of the continuous numeric variables, only one of the ranges was assigned

165   a label and, hence, a corresponding binary descriptor. As a result, perfectly correlated descriptors were

166   removed. For the sake of brevity, even though the second approach does use correlated descriptors for the

167   cell line variable, throughout this paper results obtained "with correlated descriptors" refer to the first

168   approach i.e. two labels for each continuous numeric variable, whereas results obtained "without

169   correlated descriptors" refer to the second approach. In the main text of this paper, only results obtained

170   without correlated descriptors were presented, with results obtained with correlated descriptors presented

171   in the SI for comparison.

172

173   **2.5. CORAL modelling**

174   In this work, the Monte Carlo algorithm implemented in the CORAL software (version: December 17,

175   2014 for Microsoft Windows, available at http://www.insilico.eu/coral/) was used as a tool for developing

176   linear nano-QSAR models, taking into account both the information derived from the nanomaterials'

177   physico-chemical properties (e.g. zeta potential) and the experimental conditions (e.g. cell type).

178   Specifically, after we downloaded the zipped file from the aforementioned website containing the binary

179   executable files, we executed the CORAL.exe binary file included in the folder

180   "CORALSEA\MyCORALSEA\REGRESSION" to perform the modelling. In keeping with earlier work,

181   we generated a "pseudo-SMILES" string for each instance which represented both information related to

182 particular experimental conditions and nanomaterial properties [40]. In more detail, with this particular

183 approach, all the eclectic information is used for modelling, with the endpoint of interest being a function

184 of both the nanomaterials' physico-chemical properties and experimental conditions. Pseudo-SMILES

185 character strings were derived as shown in Table 2. When used to build linear models, as in the current

186 work, the CORAL algorithm effectively treats each character (or label) in the pseudo-SMILES strings as

187 a binary descriptor which takes a value of 1 (or 0) if the character is present (or absent) for a given instance

188 [40, 43]. The manner in which predictions are obtained, based on the values of these descriptors for a

189 given instance, is further explained when discussing "Variable importance" below (see equations 6 and

190 7). Table 2 shows the selected labelling approaches with and without correlated descriptors, but we

191 reported in the main text only results obtained without correlated descriptors (leave-one-out results

192 obtained with correlated descriptors can be found in the SI). For the current dataset, after removing

193 correlated descriptors, pseudo-SMILES labels were generated as follows. The information on the cell

194 type was coded with the 'C', 'D', 'E', 'F' and 'G' characters for the 16HBE, A549, HaCaT, NRK-52E

195 and THP-1 cell types, respectively. For all numeric descriptors in the dataset, a "binary split" was

196 performed i.e. numeric values beyond some threshold were assigned a label and values before that

197 threshold were not, thus avoiding incorporating perfectly correlated descriptors. Specifically, for the

198 treatment time descriptor, a label 'A' was assigned if the exposure time was 24 hours, whereas no label

199 was assigned if the exposure time was 48 hours. For each of the three properties related to the

200 nanomaterial physico-chemical properties, namely average size, aspect ratio and zeta potential, a binary

201 split of the values was applied based on the median value for the dataset, with the rationale of having a

202 similar number of instances in a given range for each property. (N.B. The odd number of instances – i.e.

203 19 – in the dataset meant that the number of instances in each range, for each binary split, could not be

204 perfectly equal and the ranges are expressed in terms of the values just beyond the median for aspect ratio

205 and zeta potential.) The thresholds used for the splits were 27.5 (no label for values below or equal to the

206 threshold, 'I' for values above it), 1.0 (no label for values equal to the threshold, 'K' for values above it)

207 and -32.0 (no label for values below the threshold, 'L' for values equal to or greater than it) for the average

208 size, aspect ratio and zeta potential, respectively. In earlier work with CORAL [40-42], the authors

209 developed five different splits of the same dataset in order to check whether the developed models were

210 obtained by chance. According to the recommended CORAL optimisation strategy, we selected the best

211 hyperparameter values (i.e. N = number of epochs, T = threshold) using the model performance on a

212   subset of the dataset, which is called a "test set" in the CORAL software documentation, and then we

213   predicted the single item "external" test set in a separate step after the model was built. For each LOO

214   training set, modelling was repeated five times, via splitting the training set to yield an internal "test set"

215   for hyperparameter selection, five times. More details on the application of the CORAL software to this

216   dataset are reported in the SI, including full details of the five different LOO training set partitions used

217   for hyperparameter selections. (See "Details on the CORAL software settings and optimisation" in the

218   SI.)

219

220   **2.6. Random Forest modelling**

221   Random Forest is an ensemble learning method for both classification and regression which operates by

222   building a multitude of decision trees, providing as output the class which represents the majority

223   prediction, for classification problems, or the average prediction, for regression problems, of the

224   individual trees [24, 25]. Each decision tree is grown using an independent random sample of the instances

225   in the training set, with the descriptors considered for splitting each node being independently sampled

226   from the total. In the current work, both bootstrap sampling of the training set, i.e. sampling of N from N

227   with replacement, and sampling without replacement were considered. The results presented in the main

228   text were obtained without bootstrap sampling, with results obtained with bootstrap sampling being

229   reported in SI. Whilst bootstrap sampling is typically used [25, 27] it is not currently possible to calculate

230   feature contributions (see the "Feature contribution analysis" section) with the available software [44] if

231   bootstrap sampling is used. The results in the SI show that, for this dataset, the model performance and

232   standard variable importance measures (see the "Variable importance" section) are very similar with both

233   types of sampling. In this work, we used the Random Forest algorithm implemented in the randomForest

234   R package (version 4.6-12) [27], with the default values for the algorithm "hyperparameters" i.e. number

235   of trees to grow (ntree) equal to 500 and the number of descriptors randomly sampled at each split (mtry)

236   equal to the total number of descriptors in the dataset divided by three (for regression problems) as

237   explained in the randomForest package documentation. The experimental data used for Random Forest

238   modelling were the same as for the CORAL software. The binary descriptors implicitly encoded in the

239   pseudo-SMILES strings created for the CORAL software were explicitly represented for modelling using

240   the randomForest package i.e. a "1" value was assigned each time a specific label was present whereas a

241   "0" value was assigned each time the label was absent in the considered pseudo-SMILES. Using this

242   procedure, an explicit bit string was built for each pseudo-SMILES, as shown in Table 3. As per modelling

8

243   with CORAL, with the Random Forest algorithm 95 models were developed with a given sampling

244   protocol i.e. 19 models for each LOO training set and all modelling on a given training set was repeated

245   five times to take account of the random sampling inherent to building models with Random Forest or

246   CORAL. This process was repeated twice with two different sampling protocols: with simple sampling,

247   without replacement, or bootstrap sampling i.e. the replace argument of the randomForest() function was

248   set to FALSE and TRUE respectively. Hence, 190 Random Forest models were built in total with or

249   without correlated descriptors. (It should be reiterated that only results "without correlated descriptors",

250   meaning without perfectly correlated descriptors, without bootstrap sampling are presented in the main

251   text.)

252

253   **2.7. Variable importance**

254   **2.7.1.   CORAL**

255   In the current work, we selected the additive scheme of the CORAL software which computes a so-called

256   "optimal descriptor" (DCW) as the sum of correlation weights associated with the labels present in the

257   pseudo-SMILES strings [40, 43], according to equation (6).

258   $$\mathbf{DCW(Threshold, \ N_{epoch})_i = \sum_{k=1}^{5} Cw_k \times SA_{k,i}} \qquad \textbf{(6)}$$

259   N.B. In equation (6), $SA_{k,i}$ takes the value 1 (or 0) if the corresponding pseudo-SMILES label is present

260   (or absent) in an instance (i) i.e. the correlation weights (Cw_k) are summed over all labels present in a

261   given instance. In order to understand the relationship between the correlation weights and the final

262   predicted value, it is important to note that the so-called "optimal descriptor" is used to calculate the

263   predicted value for the endpoint using a one variable linear equation, as shown in equation (7).

264   $$\mathbf{Prediction_i = C_0 + C_1 * DCW(Threshold, \ N_{epoch})_i} \qquad \textbf{(7)}$$

265   Hence, it can be seen that the correlation weights are essentially scaled values of (i.e. are directly

266   proportional to) the coefficients of the binary descriptors in the final linear model developed using

267   CORAL. In order to make a comparison between CORAL and the standard Random Forest methods for

268   variable importance, we calculated the absolute values of the correlation weights for each descriptor. This

269   is because the Random Forest standard variable importance measures do not take account of the sign of

270   the contribution a given descriptor value makes towards the prediction.

271

272

2.7.2. **Random Forest**

274 The Random Forest algorithm implemented in the randomForest R package which was used in this work

275 provided information on variable importance using two approaches, by setting the "importance" option

276 of the randomForest function to TRUE. The first method [25] calculates the percentage increase of the

277 mean squared error ("%IncMSE") on the out-of-bag (OOB) subset – i.e. the subset of training set

278 instances not used to build a given tree - after the permutation of descriptors' values. In greater detail, for

279 each tree in the forest, the prediction error on the OOB portion of the data, expressed by the mean square

280 error (MSE) is recorded (for regression problems). The MSE value is then calculated again after

281 permuting each predictor variable one at a time. The differences between the two calculated MSEs for

282 the original and shuffled datasets are averaged over all trees and then normalised by the standard deviation

283 of the differences. The second method ("IncNodePurity") calculates the total decrease in node

284 "impurities" from splitting on a given descriptor, averaged over all the generated trees. For regression,

285 "impurity" is measured by the residual sum of squares (RSS) metric for a given node [27].

286

287 **2.8. Summarising Variable Importance Values**

288 The different variable importance approaches employed with Random Forest and CORAL are applicable

289 for a single model, hence – in order to derive general conclusions – it was necessary to summarise these,

290 for a given combination of modelling approach and variable importance approach, over all 95 (19 LOO

291 training sets × 5 repetitions) models. Furthermore, it was necessary to take account of the fact that the

292 different approaches could vary in scale – which would confound comparisons. Hence, the raw values (v)

293 – for a given combination of modelling approach and variable importance approach – were scaled (v_{scaled})

294 between 0 and 1 as per equation (8), where the minimum (v_{min}) and maximum (v_{max}) values were obtained

295 across all 95 models and all 5 descriptors. Subsequently, the values were summarised in terms of the

296 arithmetic mean and the corresponding standard error of the mean.

297

298 $$\mathbf{v_{scaled}} = \frac{(\mathbf{v} - \mathbf{v_{min}})}{(\mathbf{v_{max}} - \mathbf{v_{min}})} \qquad\qquad\qquad \textbf{(8)}$$

299

300 **2.9. Feature contribution analysis**

301 By "feature contribution analysis", we refer to estimates of both the sign and magnitude of the influence

302 a given descriptor has on the prediction made by a given model, in contrast to "variable importance"

303   measures which only estimate the magnitude of the influence. As far as the CORAL software is

304   concerned, we calculated feature contributions based on the signed values of the correlation weights.

305   Indeed, as equations (6) and (7) show (see the "Variable importance" section), for each single model

306   which is obtained by selecting the additive method, the signed values of the correlation weights allow

307   understanding of whether a certain descriptor is contributing "positively", i.e. it contributes to increased

308   toxicity, or "negatively", i.e. it contributes to decreased toxicity. For Random Forest, a feature

309   contribution analysis was carried out using the technique developed by Kuz'min and colleagues [45] and

310   implemented in the rfFC R package [44] which is designed to work with the randomForest

311   implementation of Random Forest. (Specifically, version 1.0 of rfFC, as obtained via the

312   "install.packages("rfFC",repos="http://R-Forge.R-project.org")" command, was used in the current

313   work). This feature contribution method is a measure of the influence, in terms of the magnitude and sign,

314   of each variable on the model prediction for a single instance. In principle, the feature contribution

315   associated with the value of a given descriptor could vary between instances with the same value for that

316   given descriptor, due to the fact that Random Forest models are non-linear. In contrast, the feature

317   contribution associated with a single descriptor as calculated for CORAL is either equal to the value of

318   the corresponding correlation weight (if the descriptor value is 1) or 0 (if the descriptor value is 0). Hence,

319   to enable a comparison between the average influence of a given descriptor value being 1 for both CORAL

320   and Random Forest, pseudo-coefficients were derived from the Random Forest feature contributions.

321   These pseudo-coefficients were calculated by computing, for each descriptor, the difference between the

322   arithmetic mean average values calculated over the feature contribution values for the pseudo-SMILES

323   strings having a value of 1 for that specific descriptor (here called FC(1)) and pseudo-SMILES strings

324   having a value of 0 for the same descriptor (here called FC(0)), according to the equation (9).

325

326   **Pseudo − coefficient = $\overline{FC(1)}$ − $\overline{FC(0)}$**   **(9)**

327   **3.   Results**
328   **3.1. LOO results**
329   LOO results, in terms of R^2 and RMSE for both CORAL and Random Forest algorithms are reported in

330   Table 4. As far as the global results on the corresponding training sets are concerned, the average and

331   standard error of the mean, over the 95 developed models, of the R^2 and RMSE statistics were calculated

332   for both CORAL and Random Forest models. N.B. In contrast to the results shown in Table 4, results

with perfectly correlated descriptors (for both CORAL and Random Forest) and bootstrap sampling (for Random Forest) are presented in the SI. Considering the LOO results reported in Table 4, it is clear that CORAL's LOO test set performance was substantially worse than its performance on the corresponding training sets. With respect to CORAL, the global average value of the R^2 on training sets was 0.8285 whereas the average RMSE was 0.2347. Results from LOO (i.e. testing) for CORAL showed a decrease for the average R^2 to 0.6486, whereas the average value of the RMSE increased to 0.3456. However, the corresponding results for Random Forest showed a smaller reduction in estimated model performance upon going between training and LOO test global results. The average values of R^2 were 0.8723 and 0.7807 for training and test set respectively and the average values for RMSE were 0.2011 and 0.2604 for training and test set respectively. If one considers only results from the LOO test sets in Table 4, it can be stated that Random Forest performed better that CORAL and the smaller reduction in average model performance upon going from the training to the test sets indicates Random Forest did not overfit as much. As far as single run results are concerned, as shown in Table 4 for CORAL software, the average R^2 values on LOO training sets, for different splits of the same LOO training sets to yield internal "test sets" for hyperparameter selection, ranged between 0.7876 and 0.8570. (Here, it should be remembered that – for a given split of the data to yield internal "tests sets" for each LOO training set – the results were averaged across all 19 LOO training sets.) Corresponding LOO R^2 test set values ranged between 0.6143 and 0.7082. Average RMSE values for different splits of the CORAL input dataset ranged from 0.2119 and 0.2675 on training sets whereas RMSE values on test sets ranged between 0.3010 and 0.3712. The Random Forest approach showed less variability, in terms of both R^2 and RMSE, among the five runs of the software with different seeds. Indeed, average training set R^2 values ranged between 0.8711 and 0.8736, whereas LOO test set R^2 values ranged between 0.7707 and 0.7899. Moreover, according to Table 4, Random Forest average RMSE values ranged between 0.1995 and 0.2022 on training sets whereas on LOO test sets RMSE values ranged between 0.2544 and 0.2665. It is important to note that, among the five runs of LOO for the CORAL software, the largest difference in the R^2 values between training and test sets is 0.2427 (split 3) whereas, for Random Forest, the largest difference is 0.1004. Taking into account the reference value for the difference of R^2 between training and test sets reported in the article of Eriksson and colleagues [46], the average results obtained with CORAL are closer than Random Forest to the 0.3 threshold for which a model could be considered to overfit. Additional results obtained with CORAL and Random Forest under different scenarios are presented in Table S1 in the SI. Firstly, it can

363 be observed that no significant training/test set performance gap exists for Random Forest if the training

364 set is predicted using only out-of-bag samples. Furthermore, the comparison of the results obtained with

365 and without correlated descriptors for the dataset used in this work showed that Random Forest is, as

366 expected [25], less affected than CORAL by the presence of correlated descriptors (see Table S1 in SI).

367 Specifically, the split number 2 of the CORAL input dataset generated an outlier only when perfectly

368 correlated descriptors were used. For Random Forest, a very small difference in terms of R^2 and RMSE

369 global average values was observed for results with and without bootstrap sampling.

370

371 **3.2. Variable importance results**

372 Figure 2 shows the average and standard error of the mean (as error bars) of the scaled variable importance

373 values for each descriptor and each variable importance measure for CORAL and Random Forest. N.B.

374 In contrast to the results shown in Figure 2, results with perfectly correlated descriptors (for both CORAL

375 and Random Forest) and bootstrap sampling (for Random Forest) are presented in the SI (Figures S1 and

376 S2). As far as CORAL is concerned, the average values ranged between 0.0525 and 0.8941 for the K and

377 L descriptors, respectively, which are related to the nanoparticle aspect ratio (i.e. aspect ratio > 1) and

378 zeta potential (zeta potential $\geq$ -32.0 mV) nanomaterial physico-chemical properties respectively. Hence,

379 according to the CORAL variable importance measure, the nanoparticle aspect ratio and zeta potential

380 were respectively the least and most important variables related to cytotoxicity. With respect to the

381 Random Forest %IncMSE method, the average values ranged between 0.0424 and 0.8393 for the G and

382 L descriptors which are related to the THP-1 cell line and zeta potential respectively. On the other hand,

383 Random Forest IncNodePurity method average values ranged between 0.0124 and 0.8181 for the E and

384 L descriptors which refer to the HaCaT cell line and zeta potential respectively. In spite of small

385 differences depending upon the specific method used, the descriptors (K and L) corresponding to aspect

386 ratio and zeta potential are (on average) by far the most important according to both the Random Forest

387 variable importance measures. Conversely, even if CORAL also identified the descriptor corresponding

388 to zeta potential as the most important, descriptors D and G corresponding to cell lines A549 and THP-1,

389 respectively are the second and third most important variables. Furthermore, Random Forest variable

390 importance results with perfectly correlated descriptors also support the conclusion that aspect ratio and

391 zeta potential are the most toxicologically relevant variables, confirming that the Random Forest approach

392 is not significantly affected by correlated descriptors (see Figure S2 in SI). Similarly, Random Forest

393 variable importance results obtained without bootstrap sampling were largely consistent with those

394 obtained with bootstrap sampling, for both %IncMSE and IncNodePurity methods (see Figure S1 and

395 Figure S2 in SI). CORAL variable importance results with correlated descriptors showed that the two

396 most important variables were the J and L descriptors, corresponding to the aspect ratio and zeta potential.

397 However, it is important to note that, unlike for Random Forest, the other descriptors corresponding to

398 aspect ratio and zeta potential are not similarly important. It is also important to note that, for CORAL,

399 the variable importance values calculated with or without perfectly correlated descriptors were not as

400 consistent as compared to Random Forest.

401 **3.3. Feature contribution results**

402 Figure 3 shows the average values, for both CORAL correlation weights and Random Forest feature

403 contribution pseudo-coefficients, calculated across all the 95 models generated on the LOO training sets,

404 without perfectly correlated descriptors and without bootstrap sampling for Random Forest. (Results with

405 perfectly correlated descriptors are presented in SI Figure S3.) Broadly in keeping with what was observed

406 for the variable importance analysis (Figure 2), for Random Forest aspect ratio and zeta potential

407 nanoparticles' physico-chemical properties were the two most important variables whereas, for the

408 CORAL approach, zeta potential and the variable related to the A549 cell line appear most important.

409 Hence, as expected, feature contribution results are consistent with those obtained for variable importance

410 for both approaches. It is important to note that, for CORAL approach, feature contribution average values

411 were all positive. Conversely, Random Forest feature contribution results presented both positive and

412 negative values. Specifically, for the CORAL approach, the correlation weight associated with the zeta

413 potential feature had a magnitude that is more than double of the A549 cell line magnitude; whereas for

414 Random Forest the two highest feature contribution values have a similar magnitude. In addition, in

415 contrast to CORAL, for Random Forest there is a considerable difference between the average influence

416 of the two most important descriptors (relating to aspect ratio and zeta potential) and the others. These

417 observations regarding the importance of different variables according to the feature contributions

418 calculations (Figure 3) are broadly in keeping with those observed when perfectly correlated descriptors

419 are not excluded (Figure S3). Results with correlated descriptors in the SI (Figure S3) showed that once

420 again for Random Forest approach zeta potential and aspect ratio were the two most important properties.

421 Specifically, considering the two correlated descriptors for aspect ratio and zeta potential, namely the J

422 and K labels for aspect ratio and the L and M labels for zeta potential, it is worth noting that, for Random

14

423   Forest, the magnitudes of their average feature contribution values were not only very similar (roughly

424   0.23) but also much greater than the magnitudes for the other descriptors. Conversely, for CORAL, we

425   obtained average feature contributions of significantly different magnitude for the two correlated

426   descriptors related to the same variable, both for aspect ratio and zeta potential properties. When the

427   signed values are considered (Figure 3 or Figure S3), it is worth noting that for Random Forest high values

428   of zeta potential are associated with an increase in cytotoxicity, whereas high aspect ratio values are

429   associated with a decrease of toxicity since the average pseudo-coefficient value is negative for the

430   corresponding descriptors. It is important to note that these findings are consistent with the preliminary

431   analysis of the dataset reported in Figure 1. Conversely, the CORAL approach seems to only be able to

432   partially recognise the trend in the data for the zeta potential. The descriptor associated with higher zeta

433   potential values has a positive average feature contribution value, regardless of whether perfectly

434   correlated descriptors were removed (Figure 3) or not (Figure S3). However, when perfectly correlated

435   descriptors are not removed, the average feature contribution value for the descriptor corresponding to

436   lower zeta potential values is still positive, even if less so (Figure S3). Whether perfectly correlated

437   descriptors were removed (Figure 3) or not (Figure S3), the average feature contribution value for both

438   descriptors corresponding to aspect ratio was positive.

439   **4. Discussion**

440   Taking into account the results obtained in this comparison work, both in terms of their predictive

441   performance estimated via "external" LOO validation and their ability to be interpreted to reveal trends

442   in the data, the non-linear Random Forest approach performed better than the linear CORAL approach

443   for the specific dataset used in this paper. With respect to Random Forest, the difference for both R^2 and

444   RMSE average values between training and test sets was smaller and it had better results on the test set

445   compared to CORAL (Table 4). In addition, for Random Forest both average R^2 and RMSE values for

446   the OOB and LOO predictions methods were very similar, regardless of whether modelling was

447   performed with or without bootstrap sampling and with or without correlated descriptors (Table S1). This

448   is interesting since it suggests that, even for these small datasets, as is typical for nano-QSAR studies [47],

449   there may be no need to cross-validate Random Forest models as opposed to simply reporting their OOB

450   performance. (Of course, for comparing to other methods, cross-validation would still be required for a

451   fair, like-for-like comparison). However, it must be noted that this finding may not hold in general, e.g.

452   Ballester and Mitchell found the OOB predictions only converged to the test set performance as the

453   training set got larger [48]. Currently, there is an on-going discussion on the importance of so-called

454   intrinsic and extrinsic properties as well as composition of nanoparticles for toxicological studies [10, 13,

455   49]. In our work, we incorporated various intrinsic (e.g. average primary particle size) and extrinsic (e.g.

456   zeta potential) properties as descriptors for the modelled toxicity endpoint. We further sought to take

457   account of variability in the endpoint values due to the different experimental conditions, by treating the

458   varied experimental conditions as additional descriptors, as per the so-called "eclectic" approach

459   previously proposed in the literature [41, 42, 50]. The variable importance analysis performed in this work

460   showed that the aspect ratio and zeta potential nanoparticles' physico-chemical properties were the most

461   important variables for the Random Forest approach under all modelling scenarios with or without

462   perfectly correlated descriptors and with or without bootstrap sampling (Figure 2, Figure S1 and Figure

463   S2). This was not observed for CORAL. For example, when modelling was carried out without perfectly

464   correlate descriptors (Figure 2), the two most important descriptors related to zeta potential and the A549

465   cell line. In contrast to the results obtained with Random Forest, for which the most important descriptors

466   - associated with zeta potential and aspect ratio - were comparably important, zeta potential was more

467   important for CORAL than the A549 cell line, which had a comparable importance to the THP-1 cell line

468   (Figure 2). However, it must be noted that descriptors related to cell line appear relatively less important

469   when perfectly correlate descriptors are not removed from CORAL modelling (Figure S2). Regarding the

470   observations concerning the importance of descriptors related to cell lines, Kim and colleagues [51]

471   recently reported that cell type more than other factors like nanoparticles' size and dose level can influence

472   cytotoxicity and, in addition, in the same work they stated that identical nanoparticles' preparations yield

473   different outcomes depending on the selected cell lines even if they belong to the same cell type. Whilst

474   our findings are not directly comparable, they still suggest that cell line is at least as important an

475   experimental variable as average size, with the exact significance varying depending upon the specific

476   cell line, the specific variable importance approach and modelling scenario (Figure 2, Figure S1 and

477   Figure S2). As far as nanoparticles' size is concerned, the work of Rong and colleagues [52] showed a

478   potential important role of silica particles' sizes in increasing toxicity towards endothelial cells. In another

479   more relevant study, Tokgun and colleagues [53] reported results which showed that cytotoxicity towards

480   A549 cell line depends on silica nanoparticles' size. We found that the average size of silica nanoparticles

481   was not typically (Figure 2, Figure S1 and Figure S2) amongst the most important variables but it did

482   appear more significant when CORAL modelling was carried out including perfectly correlated

483 descriptors (Figure S2). Consider the clear trend observed in the dataset concerning the relationship

484 between cytotoxicity and both zeta potential and aspect ratio (Figure 1) which was also reflected in the

485 Random Forest variable importance (Figure 2, Figure S1 and Figure S2) and feature contributions (Figure

486 3 and Figure S3) calculations. The clear correspondence between both the average Random Forest

487 variable importance and feature contributions and the clear trends observed in the dataset makes it clear

488 that our findings are not a result of an artefact of modelling but rather a consequence of the experimental

489 dataset used in this work. However, these clear trends observed in the dataset appear to be at odds with

490 the literature. Various publications have previously considered the relationship between aspect ratio and

491 zeta potential nanoparticles' physico-chemical properties and cytotoxicity. Regarding the toxicological

492 significance of particle shape (as quantified via the aspect ratio), studies for both carbon nanotubes and

493 silica nanoparticles (as per the current work) either reported that aspect ratio had no relationship to toxicity

494 or that high aspect ratio particles are more toxic [54, 55]. In contrast, if we look at the specific dataset

495 used in this work, as shown in Figure 1, high aspect ratio silica nanoparticles are clustered at the low

496 toxicity side of the graph. This finding is also reflected in the average Random Forest pseudo-coefficients

497 presented in Figure 3. This discrepancy may be due to several reasons, such as differences in other

498 characteristics of nanomaterials or in the cytotoxicity protocol used or in the cell line adopted as well as

499 the concentrations selected for the test. To this respect, the review of Fruijtier-Pölloth and colleagues [56]

500 has shown that it is difficult to compare studies that are based on different experimental conditions and

501 nanomaterials since they could yield contradictory results, which might be due to diverse toxicological

502 mechanisms involved. As far as the relationship between zeta potential and toxicity is concerned, Cho et

503 al. [57] found that, for a set of metal/metal oxide/silica nanoparticles high positive zeta potential resulted

504 in more cytotoxicity and Karunakaran et al. [58] also suggested that cytotoxicity of alumina and silica

505 particles, both micro-sized and nanoparticles, increases as a result of positive zeta potential. In the current

506 work, both the feature contribution analysis results, as shown in Figure 3, and the preliminary analysis of

507 the data shown in Figure 1, revealed that less negative zeta potential values were associated with higher

508 cytotoxicity and that this trend was clearly captured by Random Forest and, to a lesser extent, CORAL

509 (see Figure 3 and Figure S3). Whilst this might be considered consistent with earlier indications that

510 increasing zeta potential leads to higher cytotoxicity [57, 58], it must be stressed that these earlier studies

511 indicated that it was specifically positive zeta potential values that led to higher cytotoxicity and all zeta

512 potential values reported in the dataset used for the current work were negative. One possible confounding

513   factor here could be that zeta potential is highly dependent upon the composition of the medium in which

514   it was measured [57] and the experimentalists who provided the data modelled in the current work

515   indicated that zeta potential values were measured in water rather than the exposure medium used for

516   cytotoxicity testing. Hence, the actual zeta potential values of the nanoparticles when they were exposed

517   to the cells could differ from those reported in our dataset. Arguably, better mechanistic insight would be

518   obtained if zeta potential values had been measured under biologically relevant conditions [49, 57]. It is

519   also the case that future studies might build upon our work via incorporating additional descriptors into

520   the models. Firstly, as shown in the electronic version of the dataset used in this work in the SI the original

521   dataset from which this was derived included other nanomaterial characteristics and experimental

522   variables that were not considered as descriptors in our current work e.g. serum concentration or

523   dispersion protocol. Indeed, prior to modelling analyses, we selected only five variables to model,

524   according to our expert judgement since serum concentration and, supposing stirring and vortexing

525   protocols were comparable, dispersion protocol experimental values, for this specific dataset, were the

526   same for 17 out of 19 instances of the original dataset. (The assumption that the stirring and vortexing

527   protocols were comparable was based on guidance from the MODENA COST team responsible for this

528   dataset.) Secondly, none of the parameters provided in the original dataset may be considered to capture

529   the surface reactivity or dissolution of the studied silica nanoparticles. One way of partially addressing

530   this in future work, other than making additional experimental measurements [13], would be to perform

531   additional quantum-mechanical calculations to obtain new different descriptors, i.e. independent variables

532   reflecting structural and chemical properties of the nanoparticles [14, 59]. Such variables could further

533   enhance our understanding of the possible mechanism of toxicity of the studied nanoparticles.

534   **5.   Conclusions**

535   In this work a comparison between the CORAL and Random Forest methods in predicting silica

536   nanoparticles' cytotoxicity, based upon physico-chemical characteristics and experimental conditions

537   encoded into pseudo-SMILES strings, was performed. It was demonstrated that the pseudo-SMILES

538   encoding proposed for CORAL could be translated into descriptors which can be used with other

539   modelling approaches, such as Random Forest. LOO was used to externally validate the results obtained

540   from the modelling task. The predictive performance estimated from LOO was significantly higher with

541   Random Forest and substantially less overfitting was observed. Different approaches were employed to

542 analyse the significance of different descriptors within both kinds of models, including the derivation of

543 pseudo-coefficients for Random Forest models that, in contrast to standard variable importance measures,

544 reflect the signed contribution of descriptors towards the modelled endpoint. Whilst differences were

545 observed with the different approaches to interpreting the models, the Random Forest approach, more

546 than CORAL, reflected the toxicological significance of zeta potential and aspect ratio observed from

547 preliminary analysis of the dataset. Interestingly, whilst these properties have previously been reported as

548 significant for nanomaterial toxic effects, the relationships observed here were not in complete agreement

549 with some previous studies – which could reflect different mechanisms. In summary, the results obtained

550 suggest the Random Forest modelling approach is readily applicable to modelling the cytotoxicity of

551 nanoparticles and can be used to develop models which offer reasonable predictive power and which can

552 be interpreted in terms of physico-chemical-toxicity relationships.

562 **References**

563 1. Lövestam, G., Rauscher, H., Roebben, G., Klüttgen, B. S., Gibson, N., Putaud, J.-P., & Stamm, H.
564 (2010). Considerations on a definition of nanomaterial for regulatory purposes: Publications Office.

565 2. Bolt, H., Marchan, R., & Hengstler, J. (2013). Recent developments in nanotoxicology. Archives of
566 toxicology, 1-2.

567 3. Kumar, A., & Dhawan, A. (2013). Genotoxic and carcinogenic potential of engineered nanoparticles:
568 an update. Archives of toxicology, 87(11), 1883-1900.

569 4. Gebel, T., Foth, H., Damm, G., Freyberger, A., Kramer, P.-J., Lilienblum, W., Röhl, C., Schupp, T.,
570 Weiss, C., Wollin, K.-M., Hengstler, J.G. (2014). Manufactured nanomaterials: categorization and
571 approaches to hazard assessment. Archives of toxicology, 88(12), 2191-2211.

572 5. Huo, L., Chen, R., Shi, X., Bai, R., Wang, P., Chang, Y., & Chen, C. (2015). High-Content Screening
573 for Assessing Nanomaterial Toxicity. Journal of nanoscience and nanotechnology, 15(2), 1143-1149.

574 6. Muthuraman, P., Ramkumar, K., & Kim, D. H. (2014). Analysis of dose-dependent effect of zinc oxide
575 nanoparticles on the oxidative stress and antioxidant enzyme activity in adipocytes. Applied biochemistry
576 and biotechnology, 174(8), 2851-2863.

577 7. Sre, P. R., Reka, M., Poovazhagi, R., Kumar, M. A., & Murugesan, K. (2015). Antibacterial and
578 cytotoxic effect of biologically synthesized silver nanoparticles using aqueous root extract of Erythrina
579 indica lam. Spectrochimica Acta Part A: Molecular and Biomolecular Spectroscopy, 135, 1137-1144.

580 8. El Mahdy, M. M., Eldin, T. A. S., Aly, H. S., Mohammed, F. F., & Shaalan, M. I. (2015). Evaluation
581 of hepatotoxic and genotoxic potential of silver nanoparticles in albino rats. Experimental and
582 Toxicologic Pathology, 67(1), 21-29.

583 9. Donaldson, K., & Poland, C. A. (2013). Nanotoxicity: challenging the myth of nano-specific toxicity.
584 Current opinion in biotechnology, 24(4), 724-734.

585 10. Lynch, I., Weiss, C., & Valsami-Jones, E. (2014). A strategy for grouping of nanomaterials based on
586 key physico-chemical descriptors as a basis for safer-by-design NMs. Nano Today, 9(3), 266-270.

587 11. OECD. (2014). Guidance on grouping of chemicals. Series on testing and assessment No. 194 (second
588 ed.).

589 12. Miller, J. B., & Hobbie, E. K. (2013). Nanoparticles as macromolecules. Journal of Polymer Science
590 Part B: Polymer Physics, 51(16), 1195-1208.

591 13. Stefaniak, A. B., Hackley, V. A., Roebben, G., Ehara, K., Hankin, S., Postek, M. T., Lynch, I., Fu,
592 W.-E., Linsinger, T. P. J., & Thünemann, A. F. (2013). Nanoscale reference materials for environmental,
593 health and safety measurements: needs, gaps and opportunities. Nanotoxicology, 7(8), 1325-1337.

594 14. Puzyn, T., Rasulev, B., Gajewicz, A., Hu, X., Dasari, T. P., Michalkova, A., Hwang, H.-M., Toropov,
595 A. A., Leszczynska, D., & Leszczynski, J. (2011). Using nano-QSAR to predict the cytotoxicity of
596 metal oxide nanoparticles. Nature nanotechnology, 6(3), 175-178.

597 15. Ying, J., Zhang, T., & Tang, M. (2015). Metal Oxide Nanomaterial QNAR Models: Available
598 Structural Descriptors and Understanding of Toxicity Mechanisms. Nanomaterials, 5(4), 1620-1637.

599 16. Winkler, D. A. (2015). Recent advances, and unresolved issues, in the application of computational
600 modelling to the prediction of the biological effects of nanomaterials. Toxicology and applied
601 pharmacology.

602 17. Gramatica, P., Chirico, N., Papa, E., Cassani, S., & Kovarich, S. (2013). QSARINS: A new software
603 for the development, analysis, and validation of QSAR MLR models. Journal of Computational
604 Chemistry, 34(24), 2121-2132.

605 18. Bigdeli, A., Hormozi-Nezhad, M. R., & Parastar, H. (2015). Using nano-QSAR to determine the most
606 responsible factor (s) in gold nanoparticle exocytosis. RSC Advances, 5(70), 57030-57037.

607 19. Müller, K.-R., Mika, S., Rätsch, G., Tsuda, K., & Schölkopf, B. (2001). An introduction to kernel-
608 based learning algorithms. Neural Networks, IEEE Transactions on, 12(2), 181-201.

609 20. Hsu, C.-W., Chang, C.-C., & Lin, C.-J. (2003). A practical guide to support vector classification.

610 21. Panaye, A., Fan, B., Doucet, J., Yao, X., Zhang, R., Liu, M., & Hu, Z. (2006). Quantitative structure-
611 toxicity relationships (QSTRs): A comparative study of various non linear methods. General regression
612 neural network, radial basis function neural network and support vector machine in predicting toxicity of
613 nitro-and cyano-aromatics to Tetrahymena pyriformis §. SAR and QSAR in Environmental Research,
614 17(1), 75-91.

615 22. Liu, R., Rallo, R., Weissleder, R., Tassa, C., Shaw, S., & Cohen, Y. (2013). Nano-SAR development
616 for bioactivity of nanoparticles with considerations of decision boundaries. Small, 9(9-10), 1842-1852.

617 23. Liu, R., Rallo, R., Bilal, M., & Cohen, Y. (2015). Quantitative structure-activity relationships for
618 cellular uptake of surface-modified nanoparticles. Combinatorial chemistry & high throughput screening,
619 18(4), 365-375.

620 24. Breiman, L. (2001). Random forests. Machine learning, 45(1), 5-32.

621 25. Svetnik, V., Liaw, A., Tong, C., Culberson, J. C., Sheridan, R. P., & Feuston, B. P. (2003). Random
622 forest: a classification and regression tool for compound classification and QSAR modeling. Journal of
623 chemical information and computer sciences, 43(6), 1947-1958.

624 26. Toropov, A. A., Toropova, A. P., Mukhamedzhanova, D. V., & Gutman, I. (2005). Simplified
625 molecular input line entry system (SMILES) as an alternative for constructing quantitative structure-
626 property relationships (QSPR). INDIAN JOURNAL OF CHEMISTRY SECTION A, 44(8), 1545.

627 27. Liaw, A., & Wiener, M. (2002). Classification and regression by randomForest. R news, 2(3), 18-22.

628 28. Sizochenko, N., Jagiello, K., Leszczynski, J., & Puzyn, T. (2015). How the "Liquid Drop" Approach
629 Could Be Efficiently Applied for Quantitative Structure–Property Relationship Modeling of Nanofluids.
630 The Journal of Physical Chemistry C, 119(45), 25542-25547.

631 29. Goldberg, E., Scheringer, M., Bucheli, T. D., & Hungerbühler, K. (2015). Prediction of nanoparticle
632 transport behavior from physicochemical properties: machine learning provides insights to guide the next
633 generation of transport models. Environmental Science: Nano, 2(4), 352-360.

634 30. Ngamwongsatit, P., Banada, P. P., Panbangred, W., & Bhunia, A. K. (2008). WST-1-based cell
635 cytotoxicity assay as a substitute for MTT-based assay for rapid detection of toxigenic Bacillus species
636 using CHO cell line. Journal of Microbiological Methods, 73(3), 211-215.

637 31. Mosmann, T. (1983). Rapid colorimetric assay for cellular growth and survival: application to
638 proliferation and cytotoxicity assays. Journal of immunological methods, 65(1-2), 55-63.

639 32. Domey, J., Haslauer, L., Grau, I., Strobel, C., Kettering, M., & Hilger, I. (2013). Probing the
640 cytotoxicity of nanoparticles: experimental pitfalls and artifacts.

641 33. OECD. (2012). Guidance on sample preparation and dosimetry for the safety testing on manufactured
642 nanomaterials (version of 20 June 2012).

643 34. Martin, J. K., & Hirschberg, D. S. (1996). Small sample statistics for classification error rates I: Error
644 rate measurements: Information and Computer Science, University of California, Irvine.

645 35. Hawkins, D. M., Basak, S. C., & Mills, D. (2003). Assessing model fit by cross-validation. Journal of
646 chemical information and computer sciences, 43(2), 579-586.

647 36. Alexander, D., Tropsha, A., & Winkler, D. A. (2015). Beware of R 2: Simple, Unambiguous
648 Assessment of the Prediction Accuracy of QSAR and QSPR Models. Journal of chemical information
649 and modeling, 55(7), 1316-1322.

650 37. Hawkins, D. M. (2004). The problem of overfitting. Journal of chemical information and computer
651 sciences, 44(1), 1-12.

652 38. Low, Y., Uehara, T., Minowa, Y., Yamada, H., Ohno, Y., Urushidani, T., Alexander Sedykh, A.,
653 Muratov, E., Kuz'min, V., Fourches, D., Zhu, H., Rusyn, I., & Tropsha, A. (2011). Predicting drug-
654 induced hepatotoxicity using QSAR and toxicogenomics approaches. Chemical research in toxicology,
655 24(8), 1251-1262.

656 39. Marchese Robinson, R. L., Glen, R. C., & Mitchell, J. B. O. (2011). Development and comparison of
657 hERG blocker classifiers: Assessment on different datasets yields markedly different results. Molecular
658 Informatics, 30(5), 443-458.

659 40. Manganelli, S., Leone, C., manganelli, A. A., Toropova, A. P., & Benfenati, E. (2016). QSAR model
660 for predicting cell viability of human embryonic kidney cells exposed to SiO 2 nanoparticles.
661 Chemosphere, 144, 995-1001.

662 41. Toropov, A. A., Toropova, A. P., Veselinovic, A. M., Veselinovic, J. B., Nesmerak, K., Raska, J.,
663 Duchowicz, P. A., Castro, E. O., Kudyshkin, V., Leszczynska, D., Leszczynski, J., Leszczynska, D.
664 (2015). The Monte Carlo Method Based on Eclectic Data as an Efficient Tool for Predictions of Endpoints
665 for Nanomaterials-Two Examples of Application. Combinatorial chemistry & high throughput screening,
666 18(4), 376-386.

667 42. Toropov, A. A., & Toropova, A. P. (2014). Optimal descriptor as a translator of eclectic data into
668 endpoint prediction: Mutagenicity of fullerene as a mathematical function of conditions. Chemosphere,
669 104, 262-264.

670 43. Toropov, A. A., Toropova, A. P., Benfenati, E., Gini, G., Puzyn, T., Leszczynska, D., & Leszczynski,
671 J. (2012). Novel application of the CORAL software to model cytotoxicity of metal oxide nanoparticles
672 to bacteria Escherichia coli. Chemosphere, 89(9), 1098-1102.

673 44. Palczewska, A., Palczewski, J., Marchese Robinson, R. L., & Neagu, D. (2014). Interpreting random
674 forest classification models using a feature contribution method Integration of Reusable Systems (pp.
675 193-218): Springer.

676 45. Kuz'min, V. E., Polishchuk, P. G., Artemenko, A. G., & Andronati, S. A. (2011). Interpretation of
677 QSAR models based on random forest methods. Molecular Informatics, 30(6-7), 593-603.

678 46. Eriksson, L., Jaworska, J., Worth, A. P., Cronin, M. T., McDowell, R. M., & Gramatica, P. (2003).
679 Methods for reliability and uncertainty assessment and for applicability evaluations of classification-and
680 regression-based QSARs. Environmental health perspectives, 111(10), 1361.

681 47. Oksel, C., Ma, C., & Wang, X. (2015). Current situation on the availability of nanostructure–
682 biological activity data. SAR and QSAR in Environmental Research, 26(2), 79-94.

683 48. Ballester, P. J., & Mitchell, J. B. (2010). A machine learning approach to predicting protein–ligand
684 binding affinity with applications to molecular docking. Bioinformatics, 26(9), 1169-1175.

685 49. Marchese Robinson, R. L., Lynch, I., Peijnenburg, W., Rumble, J., Klaessig, F., Marquardt, C.,
686 Rauscher, H., Puzyn, T., Purian, R., Åberg, C., Karcher, S., Vriens, H., Hoet, P., Hoover, M. D., Hendren,
687 C. O., & Harper., S. L. (2016). How should the completeness and quality of curated nanomaterial data be
688 evaluated? Nanoscale.

689 50. Toropova, A. P., Toropov, A. A., Manganelli, S., Leone, C., Baderna, D., Benfenati, E., & Fanelli, R.
690 (2016). Quasi-SMILES as a tool to utilize eclectic data for predicting the behavior of nanomaterials.
691 NanoImpact.

692 51. Kim, I.-Y., Joachim, E., Choi, H., & Kim, K. (2015). Toxicity of silica nanoparticles depends on size,
693 dose, and cell type. Nanomedicine: Nanotechnology, Biology and Medicine, 11(6), 1407-1416.

694 52. Rong, Y., Zhou, T., Cheng, W., Guo, J., Cui, X., Liu, Y., & Chen, W. (2013). Particle-size-dependent
695 cytokine responses and cell damage induced by silica particles and macrophages-derived mediators in
696 endothelial cell. Environmental toxicology and pharmacology, 36(3), 921-928.

697 53. Tokgun, O., Demiray, A., Kaya, B., Karagür, E. R., Demir, E., Burunkaya, E., & Akça, H. (2015).
698 SILICA NANOPARTICLES CAN INDUCE APOPTOSIS VIA DEAD RECEPTOR AND CASPASE 8
699 PATHWAY ON A549 CELLS.

700 54. Donaldson, K., Murphy, F. A., Duffin, R., & Poland, C. A. (2010). Asbestos, carbon nanotubes and
701 the pleural mesothelium: a review of the hypothesis regarding the role of long fibre retention in the parietal
702 pleura, inflammation and mesothelioma. Particle and fibre toxicology, 7(1), 1.

703 55. Yu, T., Malugin, A., & Ghandehari, H. (2011). Impact of silica nanoparticle design on cellular toxicity
704 and hemolytic activity. ACS nano, 5(7), 5717-5728.

705 56. Fruijtier-Pölloth, C. (2012). The toxicological mode of action and the safety of synthetic amorphous
706 silica—A nanostructured material. Toxicology, 294(2), 61-79.

707 57. Cho, W.-S., Duffin, R., Thielbeer, F., Bradley, M., Megson, I. L., MacNee, W., Poland, C. A., Tran,
708 L., & Donaldson, K. (2012). Zeta potential and solubility to toxic ions as mechanisms of lung
709 inflammation caused by metal/metal-oxide nanoparticles. Toxicological Sciences.

710 58. Karunakaran, G., Suriyaprabha, R., Rajendran, V., & Kannan, N. (2015). Effect of contact angle, zeta
711 potential and particles size on the in vitro studies of Al 2 O 3 and SiO 2 nanoparticles. Nanobiotechnology,
712 IET, 9(1), 27-34.

713 59. Gajewicz, A., Schaeublin, N., Rasulev, B., Hussain, S., Leszczynska, D., Puzyn, T., & Leszczynski,
714 J. (2015). Towards understanding mechanisms governing cytotoxicity of metal oxides nanoparticles:
715 Hints from nano-QSAR studies. Nanotoxicology, 9(3), 313-325.

716

717

718

719

720

721

722

723

724

725

726

727

728

729

730

731

732

733

734

735

736

737

738

739

740

741

742

743

744

745

746

747

748

749

750 **Tables**

751 Table 1: In vitro WST-1 cytotoxicity experimental data of silica nanomaterials used for modelling.

752 "Treatment time" and "Cell type" columns are related to the in vitro experimental conditions adopted

753 during the experimental test for the exposure duration and cell line model, respectively. "Average size",

754 "Aspect ratio" and "Zeta potential" columns are related to measured physico-chemical properties for size,

755 aspect ratio and zeta potential of each nanomaterial, respectively. The average size was calculated from

756 two primary size dimensions estimated by TEM or other measurements (see the ESI for more details).

757 The "pEC25" column is the modelled variable, namely the negative logarithm, to base 10, of the EC25

758 value expressed as surface area of nanomaterial per millilitre (i.e. mm2/ml). Units are reported in squared

759 brackets for numerical properties.

ID	Treatment time [h]	Cell type	Average size [nm]	Aspect ratio [adimensional]	Zeta potential [mV]	pEC_{25} [mm²/ml]
119	24	THP-1	20.0	1.4	-46.1	-1.299
104	24	16HBE	46.0	1.2	-40.0	-1.272
186	48	THP-1	18.0	1.6	-43.7	-1.165
105	48	16HBE	46.0	1.2	-40.0	-1.135
101	48	16HBE	27.5	1.2	-40.0	-1.105
100	24	16HBE	27.5	1.2	-40.0	-1.026
102	24	A549	27.5	1.2	-40.0	-0.920
103	48	A549	27.5	1.2	-40.0	-0.872
107	48	A549	46.0	1.2	-40.0	-0.844
106	24	A549	46.0	1.2	-40.0	-0.822
121	24	HaCaT	17.0	1.0	-28.1	-0.394
127	24	THP-1	100.0	1.0	-32.0	-0.281
120	24	A549	17.0	1.0	-28.1	-0.223
129	24	HaCaT	60.0	1.0	-30.6	-0.197
128	24	A549	60.0	1.0	-30.6	-0.147
122	24	NRK-52E	17.0	1.0	-28.1	-0.070
130	24	NRK-52E	60.0	1.0	-30.6	0.059
123	24	THP-1	17.0	1.0	-28.1	0.365
131	24	THP-1	60.0	1.0	-30.6	0.483

760

761

762

763

764

765

766

767

768

769 Table 2: Labelling approach adopted for building the pseudo-SMILES for CORAL modelling. Each label

770 is a character which maps a specific value or a range of values. N.B. For brevity, "Correlated descriptors"

771 refers to perfectly correlated descriptors, since the binary descriptors corresponding to the different "Cell

772 type" labels are partially correlated. Only results without correlated descriptors were presented in the

773 main text.

Descriptor	Experimental value	Correlated descriptors	
		Yes	No
Treatment Time [h]	24	A	A
	48	B	No label
Cell type	16HBE	C	C
	A549	D	D
	HaCaT	E	E
	NRK-52E	F	F
	THP-1	G	G
Average size [nm]	≤ 27.5	H	No label
	> 27.5	I	I
Aspect ratio [adimensional]	= 1.0	J	No label
	> 1.0	K	K
Zeta potential [mV]	≥ -32.0	L	L
	< -32.0	M	No label

774

775

776

777

778

779

780

781

782

783

784

785

786

787

788

789

790

791

792

793   Table 3: Descriptors used for CORAL and Random Forest software. Each descriptor was derived based

794   on the presence/absence of the relative single character in the original pseudo-SMILES used as input for

795   the CORAL software. Only the approach without correlated descriptors is described.

ID	CORAL: Pseudo-SMILES	Random Forest: Explicit representation of descriptors									pEC_{25}
		A	C	D	E	F	G	I	K	L	
119	AGK	1	0	0	0	0	1	0	1	0	-1.299
104	ACIK	1	1	0	0	0	0	1	1	0	-1.272
186	GK	0	0	0	0	0	1	0	1	0	-1.165
105	CIK	0	1	0	0	0	0	1	1	0	-1.135
101	CK	0	1	0	0	0	0	0	1	0	-1.105
100	ACK	1	1	0	0	0	0	0	1	0	-1.026
102	ADK	1	0	1	0	0	0	0	1	0	-0.920
103	DK	0	0	1	0	0	0	0	1	0	-0.872
107	DIK	0	0	1	0	0	0	1	1	0	-0.844
106	ADIK	1	0	1	0	0	0	1	1	0	-0.822
121	AEL	1	0	0	1	0	0	0	0	1	-0.394
127	AGIL	1	0	0	0	0	1	1	0	1	-0.281
120	ADL	1	0	1	0	0	0	0	0	1	-0.223
129	AEIL	1	0	0	1	0	0	1	0	1	-0.197
128	ADIL	1	0	1	0	0	0	1	0	1	-0.147
122	AFL	1	0	0	0	1	0	0	0	1	-0.070
130	AFIL	1	0	0	0	1	0	1	0	1	0.059
123	AGL	1	0	0	0	0	1	0	0	1	0.365
131	AGIL	1	0	0	0	0	1	1	0	1	0.483

796

797

798

799

800

801

802

803

804

805 Table 4: Coefficient of determination (R^2) and root-mean-square error (RMSE) statistics for the LOO training and

806 test sets for both CORAL and Random Forest approaches. LOO was performed five times using five different

807 training set splits, i.e. five different partitions of a given LOO training set to yield an internal "test set" for

808 hyperparameters' selection, for CORAL and five different seeds for Random Forest. R^2 and RMSE values were

809 computed on the predicted values for each training set. Average and standard error of the mean (here reported in

810 brackets) for training set results were calculated over the 19 models – one for each instance in the dataset – for

811 each run of the LOO procedure. These statistics were compared with those obtained from the LOO (i.e. test)

812 predictions. Global results for training sets were calculating by averaging over all the 95 models developed (i.e.

813 19 models $\times$ 5 LOO runs) whereas, for test sets, global results were obtained by averaging over the five statistics

814 resulting from LOO. N.B. For both methods, only results without perfectly correlated descriptors are presented.

815 For Random Forest, only results without bootstrap sampling are presented.

Software	Run	R^2		RMSE	
		Training set	Test set	Training set	Test set
CORAL	split 1	0.8031 (0.0166)	0.6529	0.2540 (0.0083)	0.3443
	split 2	0.8567 (0.0056)	0.6243	0.2176 (0.0085)	0.3675
	split 3	0.8570 (0.0029)	0.6143	0.2119 (0.0034)	0.3712
	split 4	0.7876 (0.0077)	0.6436	0.2675 (0.0064)	0.3441
	split 5	0.8383 (0.0108)	0.7082	0.2222 (0.0067)	0.3010
	global	0.8285 (0.0052)	0.6486 (0.0164)	0.2347 (0.0038)	0.3456 (0.0125)
Random Forest	seed 1	0.8717 (0.0023)	0.7741	0.2022 (0.0026)	0.2646
	seed 2	0.8736 (0.0020)	0.7899	0.1995 (0.0023)	0.2544
	seed 3	0.8725 (0.0022)	0.7822	0.2015 (0.0025)	0.2597
	seed 4	0.8711 (0.0023)	0.7707	0.2022 (0.0025)	0.2665
	seed 5	0.8726 (0.0022)	0.7864	0.2004 (0.0025)	0.2568
	global	0.8723 (0.0010)	0.7807 (0.0036)	0.2011 (0.0011)	0.2604 (0.0023)

816

817

Figures

819 Figure 1: Zeta potential versus in vitro cytotoxicity experimental values for each silica nanoparticle in the

820 dataset. The graph shows the gap of toxicological data between -0.822 (ID 106) and -0.394 (ID 121).

821 With the "I" symbols are indicated silica nanoparticles with aspect ratio greater than 1 whereas the "O"

822 symbols refer to nanoparticles with aspect ratio equal to 1. N.B. the labels refer to the descriptors used to

823 encode toxicologically relevant variables for modelling.

824

825

826

827

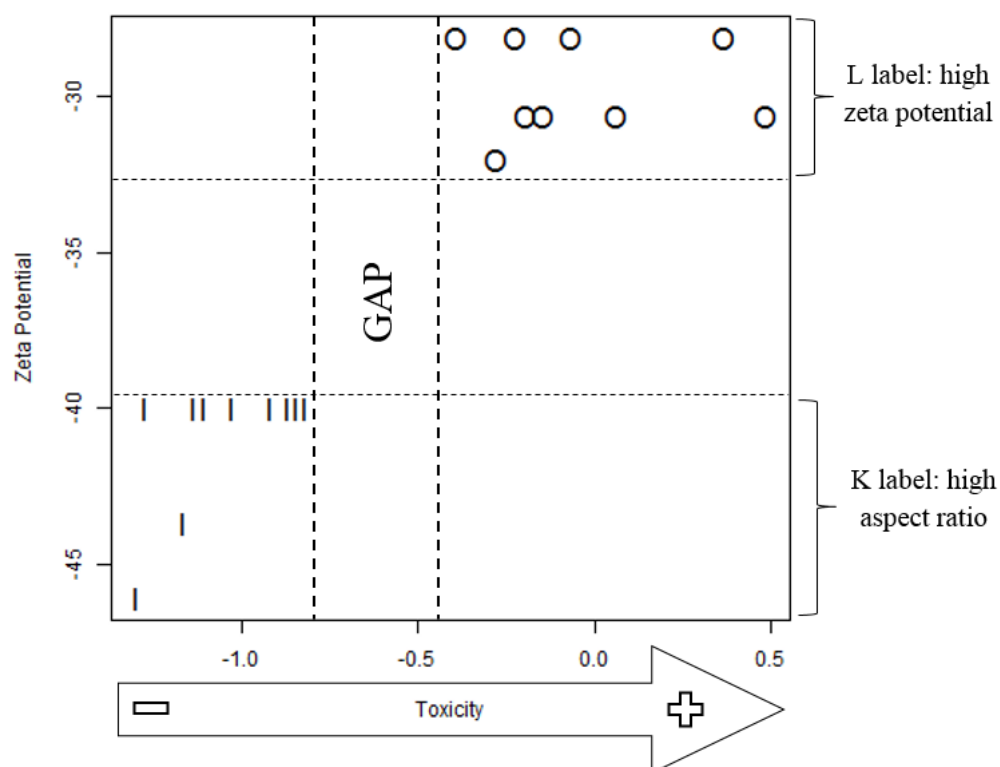828

829

830

831

832

833

834

835

836 Figure 2: Average of the scaled variable importance measures for CORAL and Random Forest (without bootstrap

837 sampling) methods. N.B. (1) All values were scaled to lie between 1 and 0 by dividing by the range (maximum –

838 minimum) of values for each variable importance method. (2) Each binary descriptor takes the value 1 or 0,

839 depending upon the value of the corresponding experimental condition or physico-chemical property. Results

840 were obtained without perfectly correlated descriptors. Error bars represent the standard error of the mean.
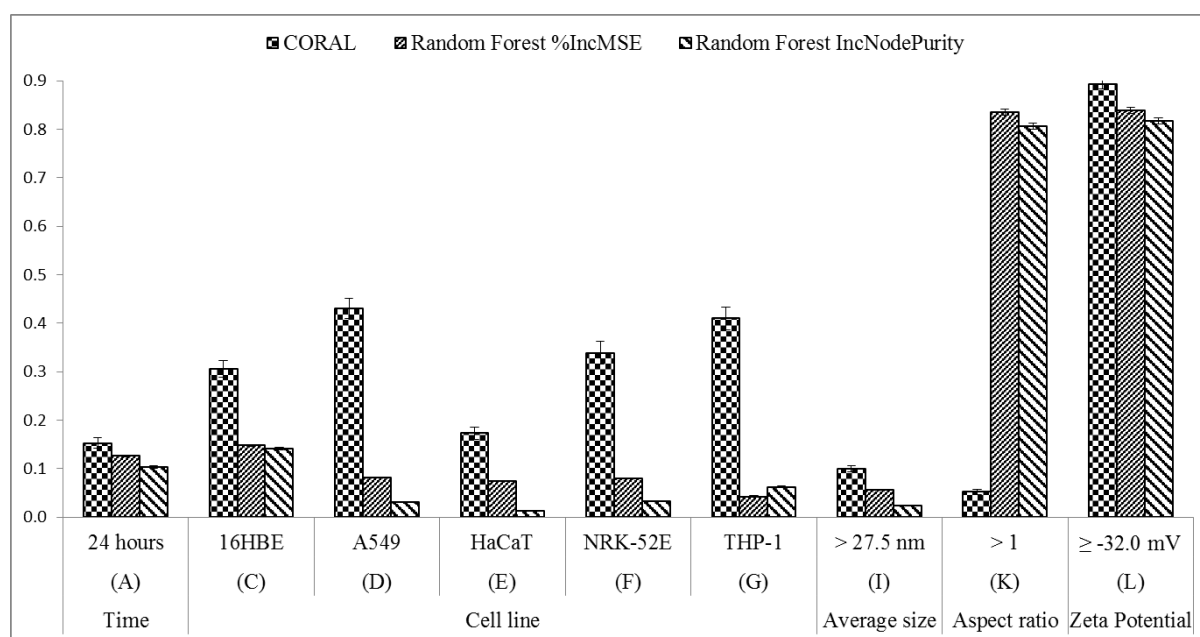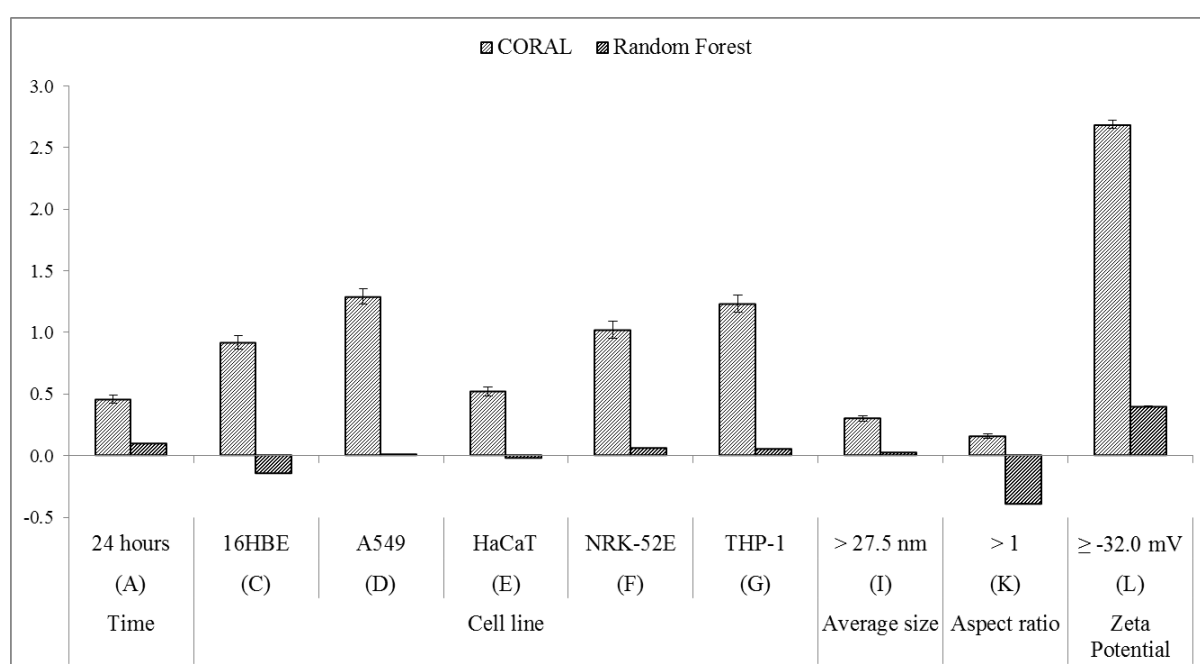

841
842
843
844
845
846
847
848
849
850
851
852
853
854
855

856   Figure 3: Comparison of the feature contribution results for CORAL and Random Forest methods. For

857   CORAL, the "feature contributions" are the correlation weights obtained for a given model built on a

858   given LOO training set. For Random Forest, feature contributions were summarised as pseudo-

859   coefficients for a given model built on a given LOO training set. The average value was calculated

860   across all the 95 models developed on LOO training sets. Error bars represent the standard error of

861   the mean. N.B. For both methods, only results without perfectly correlated descriptors are presented.

862   For Random Forest, only results without bootstrap sampling are presented.

863

864

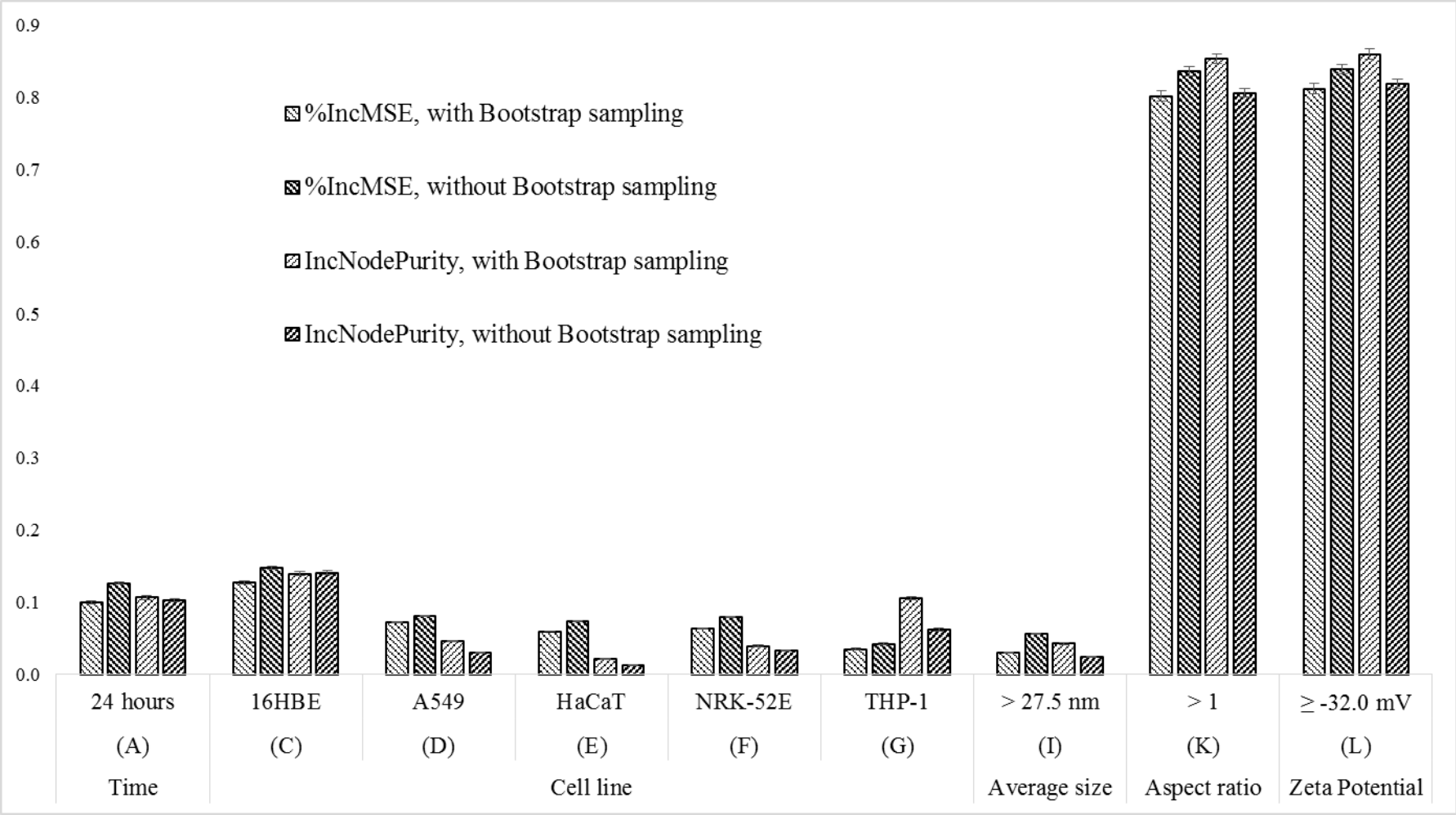865

866

867

868

869

870

871

872

873 **Supplementary Information**

874 Table S1: Summary of the comparison between CORAL and Random Forest approaches. The average and standard error of the mean (in parentheses) of the coefficient of
875 determination (R2) and the root-mean-square error (RMSE) were calculated across the 95 models developed in leave-one-out on different subsets. For Random Forest, the out-
876 of-bag (OOB) subset refers to the results obtained by predicting the training set based on out-of-bag samples and, otherwise, training set predictions were made via applying
877 all trees in the model to each training set instance. We considered the influence of correlated descriptors on both methods. N/A = not applicable. As explained in the main text,
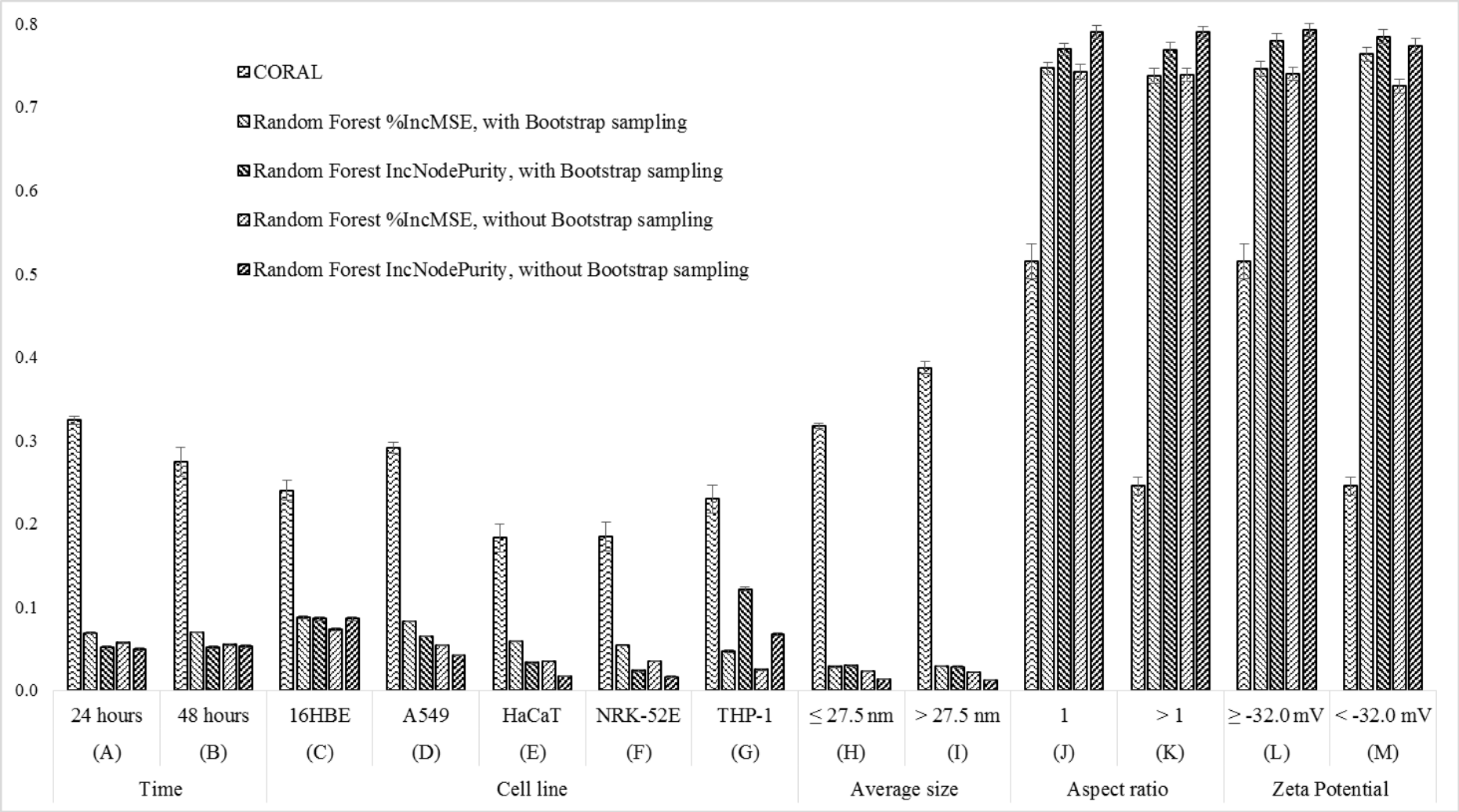878 "no" correlated descriptors refers to the absence of perfectly correlated descriptors.

Software	Correlated descriptors	Bootstrap sampling	LOO subset		Run 1		Run 2		Run 3		Run 4		Run 5		Global	
					R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE	R^2	RMSE
CORAL	Yes	N/A	Training		0.8144 (0.0074)	0.2519 (0.0057)	0.8200 (0.0311)	1.0537 (0.8322)	0.8528 (0.0037)	0.2157 (0.0034)	0.7838 (0.0085)	0.2688 (0.0066)	0.8444 (0.0053)	0.2229 (0.0060)	0.8231 (0.0071)	0.4026 (0.1663)
			Test		0.6053	0.3916	0.0293	7.8504	0.6465	0.3708	0.6508	0.3336	0.6303	0.3661	0.5124 (0.1210)	1.8625 (1.4970)
	No		Training		0.8031 (0.0166)	0.2540 (0.0083)	0.8567 (0.0056)	0.2176 (0.0085)	0.8570 (0.0029)	0.2119 (0.0034)	0.7876 (0.0077)	0.2675 (0.0064)	0.8383 (0.0108)	0.2222 (0.0067)	0.8285 (0.0052)	0.2347 (0.0038)
			Test		0.6529	0.3443	0.6243	0.3675	0.6143	0.3712	0.6436	0.3441	0.7082	0.3010	0.6486 (0.0164)	0.3456 (0.0125)
Random Forest	Yes	Yes	Training	OOB	0.7866 (0.0040)	0.2533 (0.0030)	0.7897 (0.0040)	0.2514 (0.0029)	0.7884 (0.0032)	0.2523 (0.0028)	0.7875 (0.0046)	0.2526 (0.0033)	0.7897 (0.0042)	0.2515 (0.0031)	0.7884 (0.0018)	0.2522 (0.0013)
				Predicted	0.8985 (0.0025)	0.1758 (0.0026)	0.9000 (0.0024)	0.1743 (0.0026)	0.8985 (0.0022)	0.1755 (0.0024)	0.8991 (0.0026)	0.1751 (0.0028)	0.9000 (0.0026)	0.1744 (0.0027)	0.8992 (0.0011)	0.1750 (0.0012)
			Test		0.7981	0.2468	0.7996	0.2459	0.7959	0.2481	0.7940	0.2494	0.7901	0.2519	0.7955 (0.0017)	0.2484 (0.0010)
		No	Training	OOB	0.7869 (0.0032)	0.2530 (0.0028)	0.7911 (0.0032)	0.2504 (0.0026)	0.7932 (0.0035)	0.2491 (0.0027)	0.7894 (0.0032)	0.2515 (0.0027)	0.7919 (0.0030)	0.2500 (0.0027)	0.7905 (0.0014)	0.2508 (0.0012)
				Predicted	0.8732 (0.0018)	0.1961 (0.0022)	0.8743 (0.0021)	0.1953 (0.0022)	0.8742 (0.0021)	0.1953 (0.0023)	0.8752 (0.0021)	0.1944 (0.0022)	0.8743 (0.0020)	0.1953 (0.0023)	0.8742 (0.0009)	0.1953 (0.0010)
			Test		0.7857	0.2542	0.7896	0.2519	0.7927	0.2500	0.7920	0.2505	0.7977	0.2470	0.7915 (0.0020)	0.2507 (0.0012)
	No	Yes	Training	OOB	0.7792 (0.0047)	0.2594 (0.0032)	0.7805 (0.0043)	0.2591 (0.0028)	0.7819 (0.0051)	0.2574 (0.0035)	0.7801 (0.0040)	0.2591 (0.0030)	0.7772 (0.0042)	0.2606 (0.0031)	0.7798 (0.0020)	0.2591 (0.0014)
				Predicted	0.8989 (0.0026)	0.1795 (0.0027)	0.8989 (0.0024)	0.1793 (0.0025)	0.8993 (0.0026)	0.1782 (0.0027)	0.8990 (0.0024)	0.1789 (0.0025)	0.8985 (0.0025)	0.1792 (0.0026)	0.8989 (0.0011)	0.1790 (0.0011)
			Test		0.7757	0.2617	0.7896	0.2544	0.7744	0.2622	0.7900	0.2538	0.7867	0.2559	0.7833 (0.0034)	0.2576 (0.0018)
		No	Training	OOB	0.7732 (0.0049)	0.2644 (0.0038)	0.7854 (0.0030)	0.2563 (0.0026)	0.7737 (0.0030)	0.2641 (0.0028)	0.7723 (0.0042)	0.2642 (0.0034)	0.7717 (0.0039)	0.2639 (0.0031)	0.7753 (0.0018)	0.2626 (0.0014)
				Predicted	0.8717 (0.0023)	0.2022 (0.0026)	0.8736 (0.0020)	0.1995 (0.0023)	0.8725 (0.0022)	0.2015 (0.0025)	0.8711 (0.0023)	0.2022 (0.0025)	0.8726 (0.0022)	0.2004 (0.0025)	0.8723 (0.0010)	0.2011 (0.0011)
			Test		0.7741	0.2646	0.7899	0.2544	0.7822	0.2597	0.7707	0.2665	0.7864	0.2568	0.7807 (0.0036)	0.2604 (0.0023)

879

880   Figure S1: Comparison between Random Forest %IncMSE and IncNodePurity scaled variable importance methods with and without bootstrap sampling. Average and standard

881   error of the mean (SEM) – here reported as error bar - were calculated across all the models developed by LOO. Perfectly correlated descriptors were deleted.
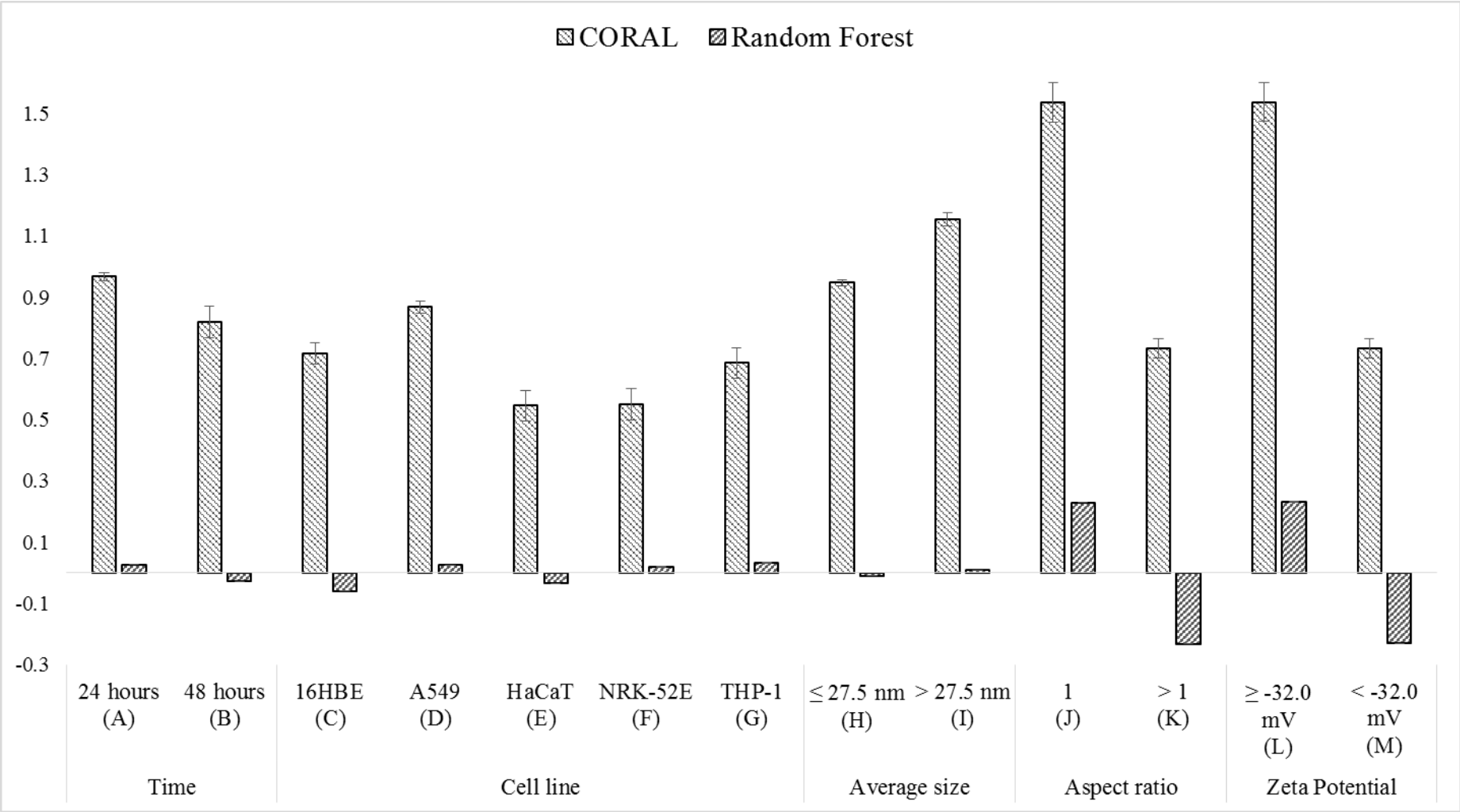

882

883   Figure S2: Comparison between scaled variable importance results for CORAL and Random Forest, including perfectly correlated descriptors. Average and standard error of

884   the mean (SEM) – here reported as error bars - were calculated across all the models developed by LOO.

885

886   Figure S3: Feature contribution analysis for CORAL and Random Forest methods with correlated descriptors. Error bars represent the standard error of the mean.

887
888

Details on the CORAL software settings and optimisation

As reported in the CORAL software documentation (version: December 17, 2014 for Microsoft Windows, available at http://www.insilico.eu/coral/), in order to use the software, specific text files must be prepared as input. The CORAL software requires the dataset used for model development to be split into three subsets, each of which is labelled with a different character in the input file, termed "sub-training" ("+"), "calibration" ("-") and "test" ("#") subsets. (Since the model hyperparameters may be optimised based upon performance on this "test" set, it may be considered an internal "test" set.) Each of these must contain a minimum number of 3 instances having a similar experimental range in order for the software to work properly. The exact CORAL settings we used in this work are shown in Figure S4. In this work we applied the "additive scheme" for which the optimal descriptor DCW is calculated by summing the correlation weights CWs of each single attribute S_k which is present in the input pseudo-SMILES (see main text). Moreover, we selected the "classic scheme" which doesn't use the "calibration" subset. The input files we used do contain instances with the "-" label for calibration subset but this label was automatically converted by the software into the "+" label for the sub-training subset (see input and output files in Supplementary Information). We applied the recommended approach [40, 42] of optimising the CORAL hyperparameters (i.e. the threshold and number of epochs, "T" and "N") on the internal test set i.e. instances labelled with "#" (see input files). According to the recommended approach, we prepared five splits of the same input dataset by shuffling the instances between training and test subsets with the rationale of having a similar experimental range among the subsets. Table S2 shows the five different splits we used in this work. For the LOO calculation, a Python script was written to create 19 input text files each containing 18 instances for CORAL modelling (see Supplementary Information). For each time the CORAL software was used for modelling, the single item external test set was predicted using the "Start of DCW and Endpoint Calculation for inserted SMILES" button as shown in Figure S5.

Computational resources used to carry out the calculations

We performed all the calculations, including with Random Forest, on a 32-bit Windows 7 machine with an Intel® Core™ i3-2120 CPU 3.30 GHz processor and 4 GB of installed memory (RAM).

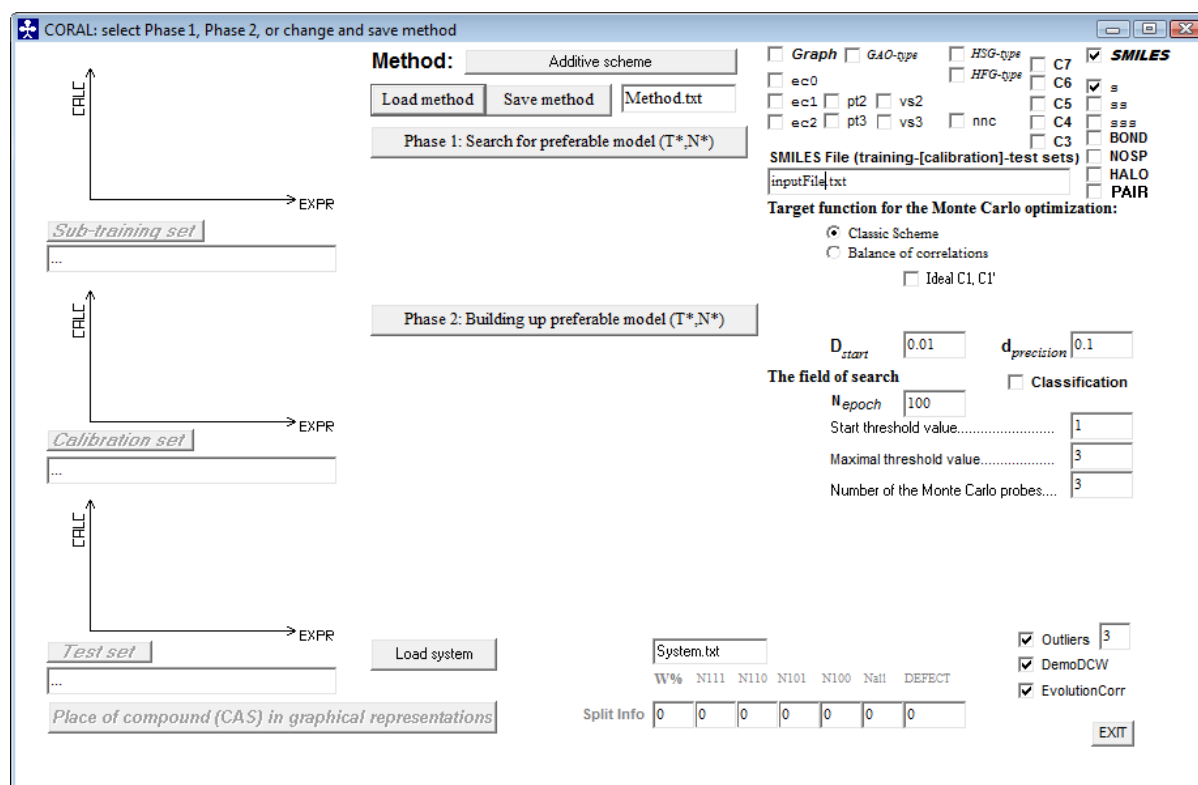919   Figure S4: CORAL graphical user interface settings used in this work.

920
921
922
923
924
925
926
927
928
929
930
931
932
933
934
935
936
937
938

939 Table S2: ID and experimental values (EXP column) for each split of the dataset used for CORAL modelling.

940 N.B. (1) The "+", "-" and "#" symbols stand for sub-training, calibration and test subsets, respectively. (2) For

941 "external" LOO validation, one instance was removed at a time and not used for model optimisation, which was

942 performed on the internal "test" subset ("#"), i.e. each of these splits of the dataset corresponds to a different split

943 of the corresponding LOO training set.

Subset	Split 1		Split 2		Split 3		Split 4		Split 5	
	ID	EXP	ID	EXP	ID	EXP	ID	EXP	ID	EXP
+	119	-1.299	119	-1.299	119	-1.299	119	-1.299	119	-1.299
+	105	-1.135	100	-1.026	105	-1.135	105	-1.135	105	-1.135
+	102	-0.920	102	-0.920	102	-0.920	102	-0.920	107	-0.844
+	106	-0.822	106	-0.822	127	-0.281	106	-0.822	106	-0.822
+	120	-0.223	128	-0.147	120	-0.223	120	-0.223	128	-0.147
+	123	0.365	123	0.365	123	0.365	128	-0.147	100	-1.026
+	131	0.483	131	0.483	131	0.483	131	0.483	131	0.483
-	104	-1.272	104	-1.272	104	-1.272	104	-1.272	104	-1.272
-	101	-1.105	127	-0.281	101	-1.105	101	-1.105	101	-1.105
-	103	-0.872	107	-0.844	100	-1.026	100	-1.026	103	-0.872
-	121	-0.394	121	-0.394	107	-0.844	121	-0.394	121	-0.394
-	129	-0.197	129	-0.197	129	-0.197	129	-0.197	129	-0.197
-	130	0.059	130	0.059	130	0.059	130	0.059	130	0.059
#	186	-1.165	186	-1.165	186	-1.165	186	-1.165	186	-1.165
#	100	-1.026	105	-1.135	103	-0.872	127	-0.281	123	0.365
#	107	-0.844	103	-0.872	121	-0.394	123	0.365	102	-0.920
#	127	-0.281	101	-1.105	106	-0.822	103	-0.872	127	-0.281
#	128	-0.147	120	-0.223	128	-0.147	107	-0.844	120	-0.223
#	122	-0.070	122	-0.070	122	-0.070	122	-0.070	122	-0.070

944

945

946

947

948

949

950

951

952   Figure S5: Screenshot of the CORAL graphical user interface showing an example of calculation of a single item

953   external test set for LOO.

954

955

956