Chemometrics and Intelligent Laboratory Systems 246 (2024) 105085

Contents lists available at ScienceDirect

Chemometrics and Intelligent Laboratory Systems

journal homepage: www.elsevier.com/locate/chemometrics

Kernel-based mapping of reliability in predictions for consensus modelling

Viviana Consonni, Roberto Todeschini, Marco Orlandi, Davide Ballabio [*]

Milano Chemometrics and QSAR Research Group, Department of Earth and Environmental Sciences, University of Milano - Bicocca, Milano, Italy

ABSTRACT

Approaches of high-level data fusion, also known as consensus, combine predictions of individual models to increase reliability and overcome limitations of single models. Consensus strategies are frequently applied in the framework of Quantitative Structure - Activity Relationships (QSARs) to reduce the uncertainties in the prediction of molecular activities and provide better accuracy of the model outcomes. However, specific regions of the chemical space may systematically be associated with low accuracy and even consensus modelling cannot improve prediction reliability through the multiple outcomes of individual models.

In this study, a new heuristic metric to assess the degree of accuracy of consensus predictions in the chemical space is proposed. This metric can assist the mapping of reliability in prediction and enhance the delineation of a safe zone, where consensus predictions are expected to have better accuracy. The new metric is calculated by kernel-based potential functions and it can be used in the framework of both classification and regression consensus modelling. Four case studies, including extensive datasets for consensus modelling, were used to test the proposed approach.

Results demonstrated that a potential can be associated with regions of the chemical space as a function of accuracy of consensus modelling and it can be used to enable the mapping of reliability in prediction and the definition of specific regions where predictions are expected to be more reliable.

1. Introduction

Chemometrics strategies for the integration of heterogeneous information from diverse data sources allow to increase the reliability of predictions, overcome limitations of single approaches and enable the extraction of relevant information content [1–3]. Information sources can be defined as data blocks in the case of low-level data fusion, features when dealing with mid-level strategies or predictions of individual models when dealing with high-level approaches [4,5].

In the framework of Quantitative Structure - Activity Relationships (QSARs), fusion strategies are commonly applied as valuable tools to reduce the uncertainty in the prediction of molecular activities [6,7]. The high-level approach (also known as decision level or consensus modelling) is the most common fusion method. For a given biological activity or molecular property, individual models are trained on a shared training pool of chemicals, but using different molecular descriptors and/or different learning algorithms. Then, the different modelling outcomes for the same chemical are combined to obtain the final property predictions.

The fundamental assumption of consensus modelling in QSAR is that the individual models reflect limited structure–activity information, which depends on the specific set of molecular descriptors in the model and the specific modelling algorithm. As a consequence, the combination of multiple QSAR predictions may provide better accuracy and reliability with the predictions, compared to single models [8]. In fact, the effects of contradictory outcomes can be reduced by merging the diverse model predictions: their integration can enhance the minimization of the overall uncertainty due to a compensation effect [9].

It is often required in research and regulatory applications of QSAR to estimate the reliability of predictions. When dealing with individual models, this is primarily addressed through the definition of the applicability domain (AD), that is, the chemical space where predictions can be considered as reliable [10]. AD is calculated on the basis of the structural features of the molecules used to train the model: compounds which are not similar enough to the training chemicals fall outside the AD and their predictions are considered as model extrapolations and most likely to be unreliable. When dealing with data fusion and consensus modelling, the prediction reliability can be estimated looking at the degree of agreement among the predictions provided by the models included in the consensus strategy. For example, these measures can be based on the posterior probability when using Bayesian consensus [11] or belief and plausibility with Dempster-Shafter theory of evidence [9].

It is noteworthy that in most of the real QSAR case studies, the distribution of the prediction reliability in the chemical space is not uniform. In fact, recent studies have highlighted that specific regions of the chemical space may be systematically associated with incorrect predictions [12–14]. To the best of our knowledge, the chemical space mapping on the basis of prediction reliability has never been proposed in the framework of consensus modelling. Thus, we propose a novel metric,

[*] Corresponding author.
E-mail address: davide.ballabio@unimib.it (D. Ballabio).

https://doi.org/10.1016/j.chemolab.2024.105085
Received 30 November 2023; Received in revised form 9 February 2024; Accepted 12 February 2024
Available online 13 February 2024

which is used to map regions of the chemical space where models are likely not accurate and could assist the evaluation of prediction reliability. More specifically, we propose a measure to assess the extent to how each point of the chemical space is associated with prediction reliability. This mapping also allows an easier delineation of a sort of *comfort zone in the chemical space*, where models are supposed to have a steady level of performance and thus operate with higher accuracy and consequently reduced uncertainty. The proposed measure of reliability is based on the calculation of kernel-based potentials in the chemical space, defined in terms of molecular fingerprints. We tested and validated the approach on four different case studies, which included extensive consensus datasets related to the modelling of both qualitative and quantitative properties.

2. Material and methods

2.1. Kernel-based reliability potential for consensus prediction

In this study, we use potential functions to calculate the degree of reliability in accurate prediction in each point of the chemical space and, thus, for each virtual compound located in that point: the higher the potential, the higher the belief the compound is located in a specific region of the chemical space where chemicals are predicted with higher accuracy.

Let **X** (n x p) be a training dataset, arranged in a p-dimensional matrix, where n is the number of chemicals (samples) and p the number of molecular descriptors (variables), and let a virtual compound t be associated to the point **t** (1 x p) of the p-dimensional chemical space, then the density estimate (reliability potential) P_t in the point **t** of the chemical space is defined on the basis of the following equation [15,16]:

$$P_t = \frac{1}{h^p} \cdot \sum_{i=1}^{n} w_i \cdot \prod_{j=1}^{p} k\left(t_j, x_{ij}\right) \tag{1}$$

where $k(t_j, x_{ij})$ is the kernel smoothing function (eq. (2)) that accounts for the dissimilarity between the virtual compound t and the i-th training sample considering the j-th variable (eq. (3)), h is the bandwidth of the kernel-smoothing function, w_i is the weight for the i-th training sample (defined later for both qualitative and quantitative modelling), x_{ij} and t_j are the value of the j-th variable for the i-th training sample and the virtual compound t, respectively.

We considered only the most common kernel smoothing functions and specifically Gaussian (a), Epanechnikov (b), and triangular (c) kernel smoothing functions have been taken into account [17]:

$$\begin{aligned} &a) \quad k\left(t_j, x_{ij}\right) = \frac{1}{\sqrt{2\pi}} e^{-\frac{1}{2} \cdot u\left(t_j, x_{ij}\right)^2} \\ &b) \quad k\left(t_j, x_{ij}\right) = \frac{3}{4} \cdot \frac{1}{\sqrt{5}} \left(1 - \frac{u\left(t_j, x_{ij}\right)^2}{5}\right) \quad if \;\; k\left(t_j, x_{ij}\right) > 0,\; 0 \; otherwise \\ &c) \quad k\left(t_j, x_{ij}\right) = 1 - \left|u\left(t_j, x_{ij}\right)\right| \quad if \;\; k\left(t_j, x_{ij}\right) > 0,\; 0 \; otherwise \end{aligned} \tag{2}$$

where $u(t_j, x_{ij})$ is the distance measure defined as

$$u\left(t_j, x_{ij}\right) = \frac{t_j - x_{ij}}{h} \tag{3}$$

In this study, the chemical space was built as the 2D dimensional space defined by the first two scores of MultiDimensional Scaling (MDS) calculated on the molecular fingerprints of the training molecules [18]. Thus, t_j and x_{ij} refer specifically to the j-th MDS score of the t-th point and the i-th training sample, respectively.

2.2. Weighting for qualitative and quantitative modelling

Different weightings w can be associated to training samples to define the entity of potential in order to tune and adapt the reliability measure on the basis of the model prediction accuracy. In particular, different weighting schemes have been defined for qualitative (classification) and quantitative modelling (regression).

In the case of an ensemble of M classification models to be used for consensus prediction, the weight w_i for the i-th training sample is defined in the following way:

$$w_i = \left(1 - \frac{n_{g(i)}}{n}\right) \cdot \left(2 \cdot \frac{\sum_{m=1}^{M} \delta_{im}}{M} - 1\right) \qquad \delta_{im} = \begin{cases} 1 & if \;\; \widehat{c}_{im} = c_i \\ 0 & if \;\; \widehat{c}_{im} \neq c_i \end{cases} \tag{4}$$

where $n_{g(i)}$ is the number of training samples in the *g-th* class to which the i-th training sample belongs, M is the number of predictions provided by the individual models and used for the consensus and δ_{im} is the classification score of the i-th training sample for the m-th prediction, which is defined as: $\delta_{im} = 1$ if the i-th sample is correctly predicted by the m-th individual model (the predicted class $\widehat{c}_{im}$ corresponds to the experimental class c_i) and $\delta_{im} = 0$ when the i-th sample is erroneously predicted. Predictions outside the applicability domain of the models are not considered in the score calculation to avoid the use of unreliable information in the weight calculation. The first term of Equation (4) allows compensating the weights of objects belonging for underrepresented classes (unbalanced cases). On the other side, in the case of binary balanced classes ($n_g/n = 0.5$), each weight w_i is therefore bounded between -0.5 (the training sample is misclassified by all models) and 0.5 (all models predict correctly the training sample).

In the case of M regression models, the weight w_i for the i-th training sample is defined in the following way:

$$w_i = 2 \cdot \frac{\sum_{m=1}^{M} e^{-\frac{\left|\hat{y}_{im} - y_i\right|}{RMSE_m}}}{M} - 1 \tag{5}$$

where y_i is the experimental response for the i-th training sample, $\widehat{y}_{im}$ is the response predicted by the m-th individual model and $RMSE_m$ is the root mean squared error of the m-th model. Weights are therefore bounded between -1 (high residuals) and 1 (no residuals).

Therefore, coming back to the definition of the reliability potential P_t (eq. (1)) in the point **t** of the chemical space, the training compounds that contribute more to the positive potential value are those with small prediction errors.

2.3. Univariate example of reliability potential

A graphical representation of the reliability potential in the framework of classification modelling is provided in Fig. 1. As an example, 10 training samples are taken into account in a one-dimensional space to enhance the graphical demonstration, where X1 represents any quantitative variable or score used as coordinate to map samples in the space. Each sample is coloured in a grey scale: the darkness is proportional to the number of wrong predictions provided by the models participating in the consensus. Thus, the black point represents a sample which was erroneously predicted by all models, the dark grey point a sample with few correct predictions, the light grey points denote samples with several correct predictions and finally white points represent samples which were correctly predicted by all models.

The grey curves represent the individual potentials, which were calculated with a Gaussian kernel in this example. Individual potentials are related to each sample and proportional to their classification score: white samples have high positive potential, black samples negative potentials, while grey samples have intermediate potential values. The blue line defines the cumulative potential, which is the sum of all the individual contributions (eq. (1)). The chemical space is therefore characterised by higher reliability for low values of X1, where the majority of training samples associated with high accuracy are located (white and light grey points). On the contrary, the reliability in

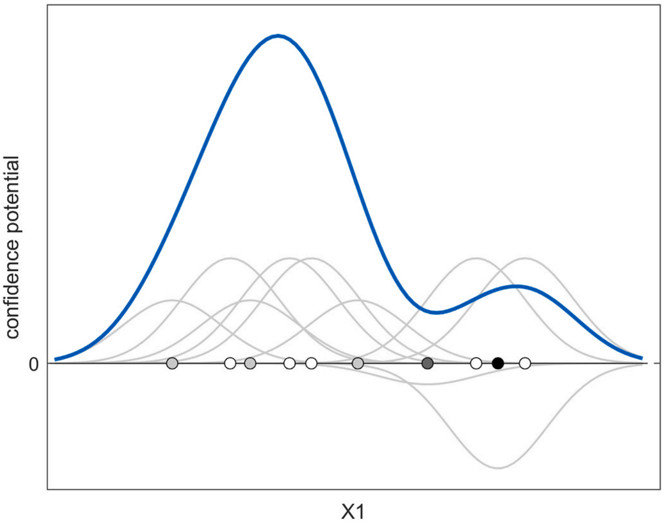

Fig. 1. Graphical representation of the reliability potential in the univariate space. Individual potentials are represented in grey, the cumulative potential in blue. Points are coloured on the basis of their classification score in a grey scale from black (low accuracy) to white (high accuracy). (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

prediction accuracy decreases as X1 increases, since in this region of the chemical space there are some training samples with not accurate predictions.

New samples can be projected in the chemical space (on the basis of their X1 coordinate) and according to the projection point the reliability degree associated with their prediction can be easily derived.

2.4. Case studies

Four case studies from collaborative projects based on consensus modelling were taken into account (Table 1) and briefly described below.

The CoMPARA dataset was distributed in the framework of the Collaborative Modelling Project of Androgen Receptor Activity [19], which was coordinated by the National Center of Computational Toxicology (U.S. Environmental Protection Agency) and aimed to provide *in-silico* predictions for the identification of potential androgen receptor modulators. The project coordinators provided a calibration set of molecules and the corresponding experimental annotations on binding activity (in the form of qualitative labels, active: positive/inactive: negative) to several research groups worldwide. The research groups were then asked to calibrate their own QSAR models and the predictive ability of each of the 34 final QSAR models was assessed by a blinded validation set including 3540 chemicals [20].

The CATMoS datasets were taken from the Collaborative Acute Toxicity Modelling Suite, which was again the outcome of a collaborative project [21]. The project coordinators collected and curated a database of rat acute oral toxicity, that is Lethal Dose 50 (LD50), which is the concentration needed to cause lethality in 50 % of the tested animals. Along with the acute oral toxicity in logarithmic scale (logLD50), two additional qualitative endpoints were taken into account: (a) very toxic, labelling molecules as positive (very toxic) if their experimental LD50 was lower than 50 mg/kg; (b) nontoxic, labelling molecules positive (nontoxic) if their experimental LD50 was greater than or equal to 2000 mg/kg. The datasets were distributed to the project participants for training of QSAR models, while the validation sets (including 2,894, 2,890, and 2195 chemicals for the very toxic, nontoxic and logLD50 endpoints, respectively) were used to carry out blinded validation of the QSAR models calibrated by the participants (30, 23 and 20 for the very toxic, nontoxic and logLD50 datasets, respectively).

The validation sets of both CoMPARA and CATMoS collaborative projects, including the predictions provided by all individual mdoels, were used in this study as the datasets for both defining and validating the kernel-based mapping.

When dealing with the CoMPARA case study, it is important to highlight the different nature between validation and training sets with respect to the type of assay which was used to define the classes (combination of multiple ToxCast in vitro assays from the same source for the training, different literature sources from single in vitro assays for the validation). However, literature data can be used to provide quality assessment, as done in the original CoMPARA study [19,20]. Moreover, since in this study all calculations have been performed on the validation set data, results can be considered as self-consistent for the set itself and therefore we can exclude that the difference in the assays between training and validation can affect the results of the analysis.

Finally, the validation sets provide comprehensive case studies to test the proposed method both in terms of the number of samples (higher than 2000 for all case studies) and the number of available predictions provided by the individual models and to be integrated in the consensus approach (ranging between 20 and 34).

For each case study, we randomly divided the molecules of these sets into two groups: training (70%) and test (30%), as reported in Table 1. Since the class proportion was unbalanced, the same proportion was maintained while splitting the data in the two sets. The density estimate (potential) for each test chemical was calculated with respect to the training molecules. This simulates the real application of the proposed uncertainty measure, which is supposed to be calculated for new chemicals, with unknown experimental class, once these are predicted through a number of classification models included in a consensus framework. Training molecules were also used to select the type of

Table 1
Characteristics of the datasets: modelling task, number of models participating in the consensus, number of chemicals, number of training and test molecules, molecule partition in the positive and negative class.

dataset	task	consensus models	set	molecules	positive	negative
CoMPARA binding	classification	34	dataset	3540	411	3129
			training	*2478*	*288*	*2190*
			test	*1062*	*123*	*939*
CATMoS very toxic	classification	30	dataset	2894	2651	243
			training	*2026*	*1856*	*170*
			test	*868*	*795*	*73*
CATMoS nontoxic	classification	23	dataset	2890	1655	1235
			training	*2022*	*1158*	*864*
			test	*868*	*497*	*371*
CATMoS logLD50	regression	20	dataset	2195	–	–
			training	*1537*	—	—
			test	*658*	—	—

kernel smoothing functions and to define the bandwidth, while test molecules were not used for optimisation. Table 1 collects the features of each case study.

2.5. Set up of the chemical space

The chemical space was encoded into the first two scores of Multidimensional Scaling (MDS) calculated on the molecular fingerprints of the training molecules. The molecular fingerprints (FPs) are binary or count descriptors encoding the structural features of chemicals; they are able to provide a holistic, comprehensive and unambiguous representation of molecular structures [22]. In particular, binary extended connectivity fingerprints (ECFPs) were calculated to represent the molecular structure of chemicals for CoMPARA and CATMoS very toxic and nontoxic datasets, and Path FingerPrints (PFP) for the CATMoS logLD50 dataset. For all these cases, molecules were described by binary vectors of 1024 bits. Moreover, additional trials for representing the chemical space were conducted by using the first two scores of Principal Component Analysis (PCA) calculated on specific molecular descriptors, which were specifically selected and proposed in previous studies for the prediction of CoMPARA and CATMoS endpoints in the OPERA models [19,21,23].

The Jaccard-Tanimoto metric [24] was selected to calculate the pairwise similarities between molecules, to be used as input for MDS [18], which reproduces similarity/diversity relationships through a projection in a low-dimensional space. Finally, the first two MDS scores were used to define the chemical space.

2.6. Selection of the kernel smoothing function and bandwidth

The types of kernel smoothing function and the bandwidth were selected by means of internal cross validation carried out with the training samples, which were divided in 5 groups on the basis of a venetian blind strategy. Thus, test samples were never used to tune the kernel smoothing function and the bandwidth. One group of the training samples at a time was used as a temporary set to evaluate the density estimate calculated with the other remaining groups. We considered the prediction accuracy over the different potential levels and then selected as optimal the smoothing function and bandwidth that allowed us to obtain increasing accuracy as the potential values increased. The bandwidth was varied from 0.2 to 1, with step of 0.1, while three different smoothing functions (Gaussian, Epanechnikov or triangular) were tested. As an example, the effect of the type of smoothing function and the bandwidth value on the potential levels is shown in Fig. 2 for the CoMPARA binding dataset.

2.7. Qualitative and quantitative figures of merit

The application of the Kernel-based reliability potential was assessed by different figures of merit when dealing with qualitative (classification) and quantitative (regression) modelling. In particular, when dealing with classification tasks, the class of each test compound was predicted through the majority voting criterion, which classifies a molecule as positive or negative on the basis of the most frequent class

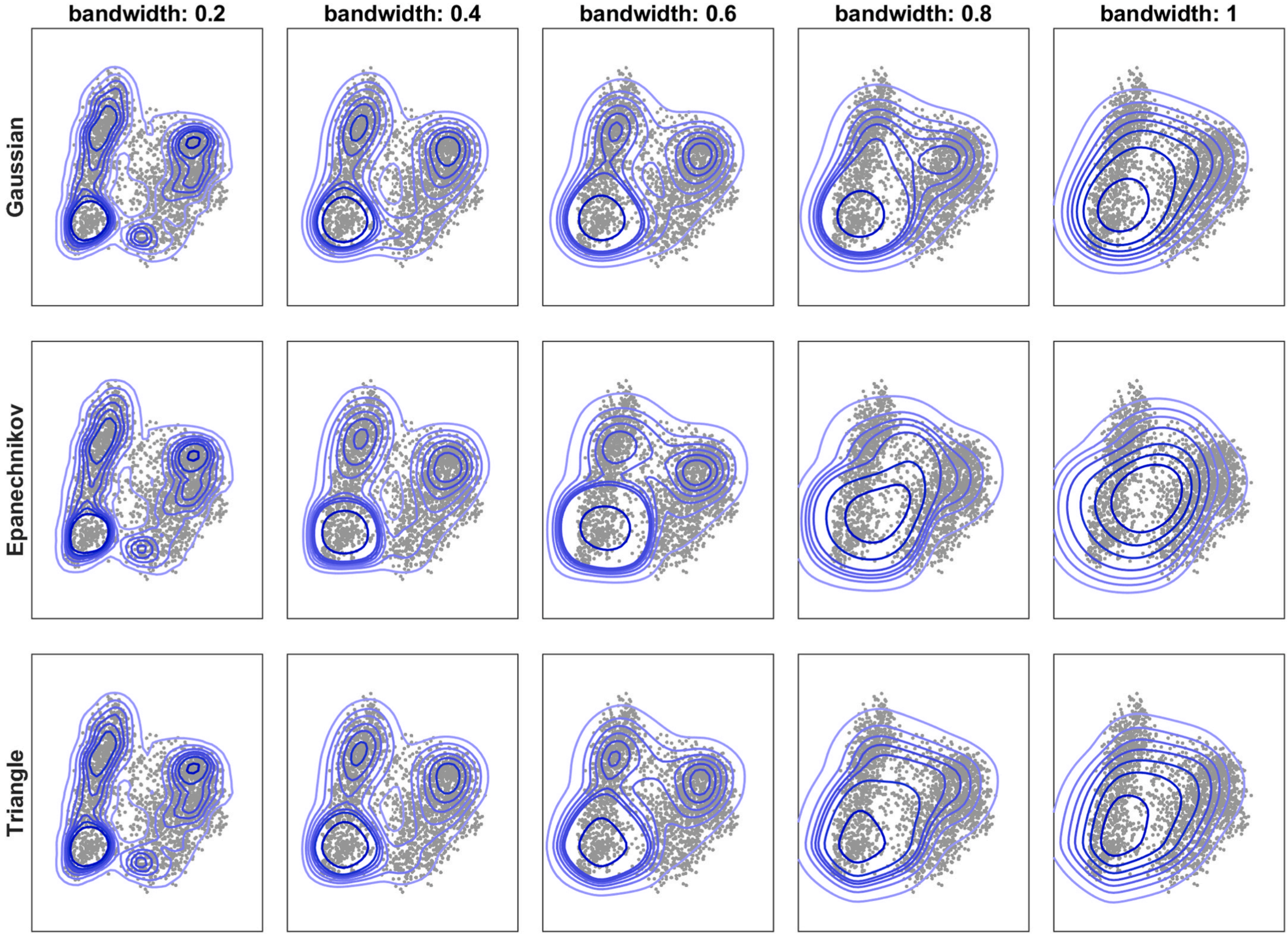

Fig. 2. Potential levels as a function of the kernel smoothing function and bandwidth for the CoMPARA binding dataset. The training chemicals are represented with grey dots in the first two MDS scores.

assignment. The classification performance was then assessed by means of sensitivity and specificity, which are the fraction of correctly predicted positive and negative chemicals, respectively; also their average, which is called Non Error rate (*NER*, also known as Balanced Accuracy), was calculated along with the class precision, which is defined as the ratio between the number of samples of the class correctly classified and the total number of samples predicted to that class [25].

On the other hand, in the CATMoS logLD50 case study consensus predictions were calculated by averaging the quantitative predictions of the individual regression models and the overall performance was evaluated by the root mean squared error in prediction (*RMSEP*) calculated on the test molecules:

$$RMSEP = \sqrt{\frac{\sum_{i=1}^{n_{test}} (y_i - \widehat{y}_i)^2}{n_{test}}} \quad (6)$$

where n_{test} is the number of molecules in the test set, y_i and $\widehat{y}_i$ are the experimental response and the consensus prediction for the *i*-th test molecule, respectively.

2.8. Software

Fingerprints were calculated with Dragon 7 software [26] and molecular descriptors with OPERA [23]. Kernel-based reliability potentials were calculated with ad-hoc MATLAB functions. Functions and data to reproduce the results obtained in this study are available at the Milano Chemometrics and QSAR Research Group website [27].

3. Results

3.1. Classification: CoMPARA binding

In this case study, Epanechnikov kernel function and bandwidth equal to 0.5 were selected by means of cross validation procedures. Fig. 3a shows the distribution of training compounds in the ECFPs fingerprints chemical space, defined through the first two MDS dimensions, along with the potential levels defining the mapping of reliability in prediction. Each molecule is coloured with a greyscale: the darkness is proportional to the number of wrong predictions provided by the single models (34 in this case study). Therefore, the darker the colour, the higher the number of misclassifications for a specific chemical, the lower its weight (eq. (4)) used to calculate the reliability potential. As a result, we then expect the potential being lower where several misclassifications are clustered (e.g., in the lower-right region for the CoMPARA binding chemical space) and higher where better classification accuracy is obtained, that is, where white and light grey points are mainly located. Path fingerprints and OPERA molecular descriptors provided an unsuitable representation of the potential distribution in the CoMPARA binding chemical space, as shown in Fig. S1 and Fig. S2 of the supplementary material, respectively.

At a first visual inspection, the potential levels provide a fairly good representation of the distribution of misclassified compounds in the chemical space. Moreover, consider that errors count proportionally in the potential calculation, thus wrong predictions for molecules experimentally labelled in the less represented class have higher weight (eq. (4)).

Afterwards, the test molecules were projected in the space and then associated with a reliability potential value according to the specific point in the chemical space they were located in. Results were evaluated looking at the classification performance over the test chemicals as a function of the different levels of reliability potential. In particular, in Fig. 3b the classification balanced accuracy (*NER*) for test chemicals is shown as a function of the potential levels. To get this plot, the potential was divided in 10 quantiles; the *NER* associated to the first potential level was therefore calculated taking into account all test chemicals with potential higher than the first potential quantile, that is, the test chemicals located inside the first potential level shown in Fig. 3a. Similarly, the NER associated to the second potential level was calculated considering only the test chemicals located inside the second potential level, and so on. As expected, *NER* increases when the reliability potential increases: the classification accuracy is higher if test chemicals are located in the inner part of the map, that is, where the reliability potential is higher. In particular, in the case of CoMPARA binding, *NER* significantly increases, from 0.74 to 0.93, when only test compounds in the most internal potential level of the map (highest potential) are

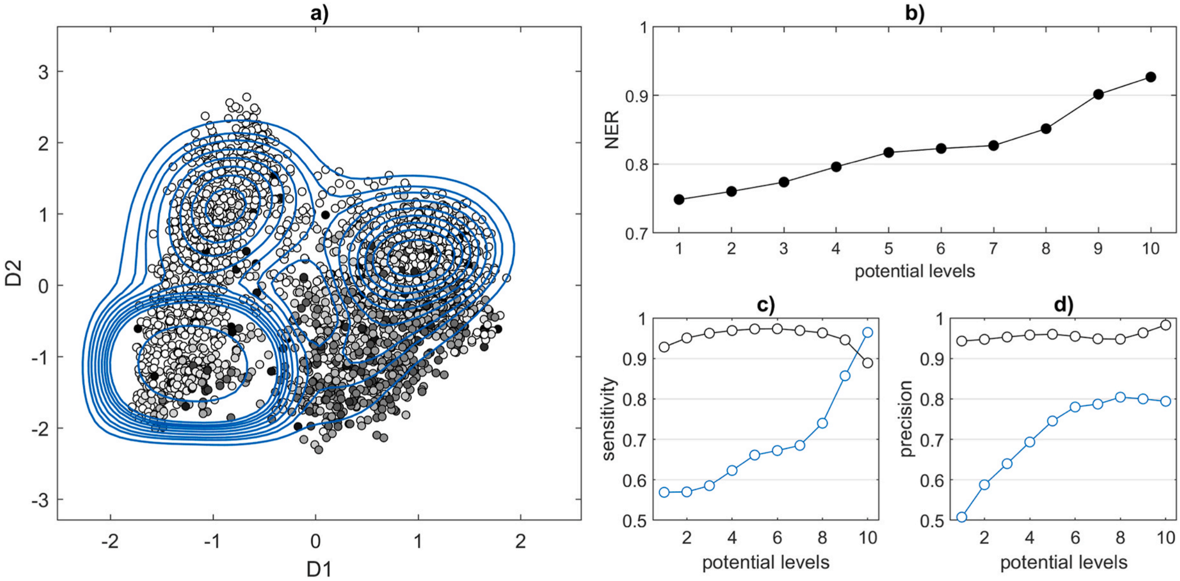

Fig. 3. Mapping of reliability in prediction for the CoMPARA binding case study: (a) scores of the training compounds in the first two MDS dimensions; molecules are coloured with a greyscale: the darker the colour, the higher the number of misclassifications; the levels of potential are coloured in blue; (b) plot of *NER* values of test chemicals as a function of potential levels; (c) plot of class sensitivity of test chemicals (blue: positive; black: negative) as a function of potential levels; (d) plot of class precision of test chemicals (blue: positive; black: negative) as a function of the potential levels. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

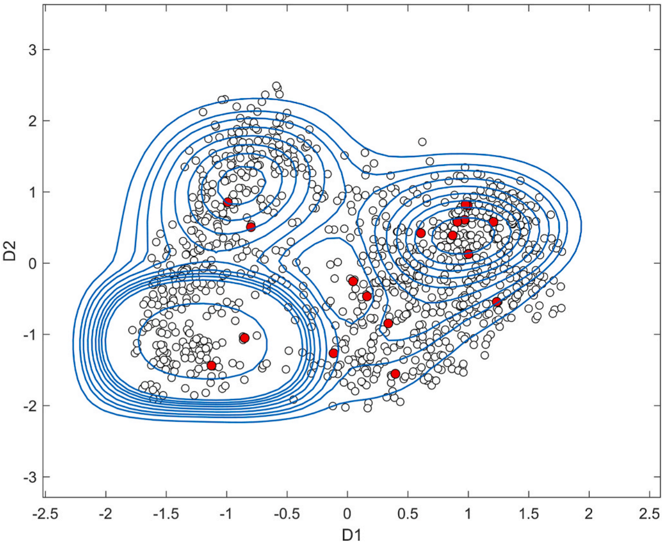

Fig. 4. Projection of the test compounds into the 2-dimensional MDS space for the CoMPARA binding case study; the levels of reliability potential calculated with training compounds are coloured in blue; test molecules correctly predicted are coloured in white; test molecules erroneously classified are coloured in dark grey; test molecules erroneously classified but with consensus prediction agreement higher than 80% are coloured in red. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

considered. In a similar way, Fig. 3c and d show how sensitivity and precision of the positive (blue) and negative (black) classes change as a function of the potential levels. As for *NER*, both sensitivity and precision increase while increasing the reliability potential, the sensitivity and precision for the positive class changing significantly from 0.56 to 0.96 and from 0.50 to 0.79, respectively.

To better explore the benefits for the use of reliability potential, we evaluated it in the case of wrong consensus predictions which however are supported by model agreement. Usually, a way to define the belief in the consensus framework is to evaluate the agreement among the combined predictions, the higher the agreement, the higher the belief. However, it may accidently happen to have wrong consensus predictions which are however based on high agreement, that is, the majority of the considered models predicts the chemical in the wrong class. The supplementary information provided by the mapping of reliability in prediction can help to identify these situations, as shown in Fig. 4, which represents the test molecules mapped in the chemical space within the potential levels previously calculated with the training samples and defining the reliability mapping (Fig. 3a). Test molecules are coloured in white if correctly predicted otherwise in dark grey, while test chemicals which (a) have agreement of consensus predictions higher than 80% (e. g. 80% of available models predicted the chemical in the same class) and (b) are however erroneously classified, are coloured in red. The belief based on the consensus agreement would be high for these cases, however it can be seen that only few of these chemicals are located in regions of the chemical space with high reliability potential. On the other hand, the majority of misclassified chemicals (grey points) are correctly located in regions with low reliability potential. Therefore, the proposed potential measure provides additional information to the estimate based on the consensus agreement and enhances the identification of those cases for which most of the classification models used for the consensus fail in the prediction.

3.2. Classification: CATMoS very toxic and nontoxic

The proposed approach was applied to both CATMoS classification datasets with the same strategy as in the previous case study. The potential for test molecules was calculated in the chemical space defined by the first two MDS dimensions of the training chemicals. Potentials were calculated by means of Epanechnikov kernel function with bandwidth equal to 0.50 and 0.80 for CATMoS very toxic and nontoxic datasets, respectively (Figs. 5 and 6). Results confirmed the potential being an effective measure to assist the reliability in predictions. For the CATMoS very toxic dataset, the increasing of potential causes an increase of *NER* of test chemicals from 0.76 to 0.90 (Fig. 5b), thus confirming that higher potential is associated to regions of the chemical space where prediction accuracy is higher. Also sensitivity and precision

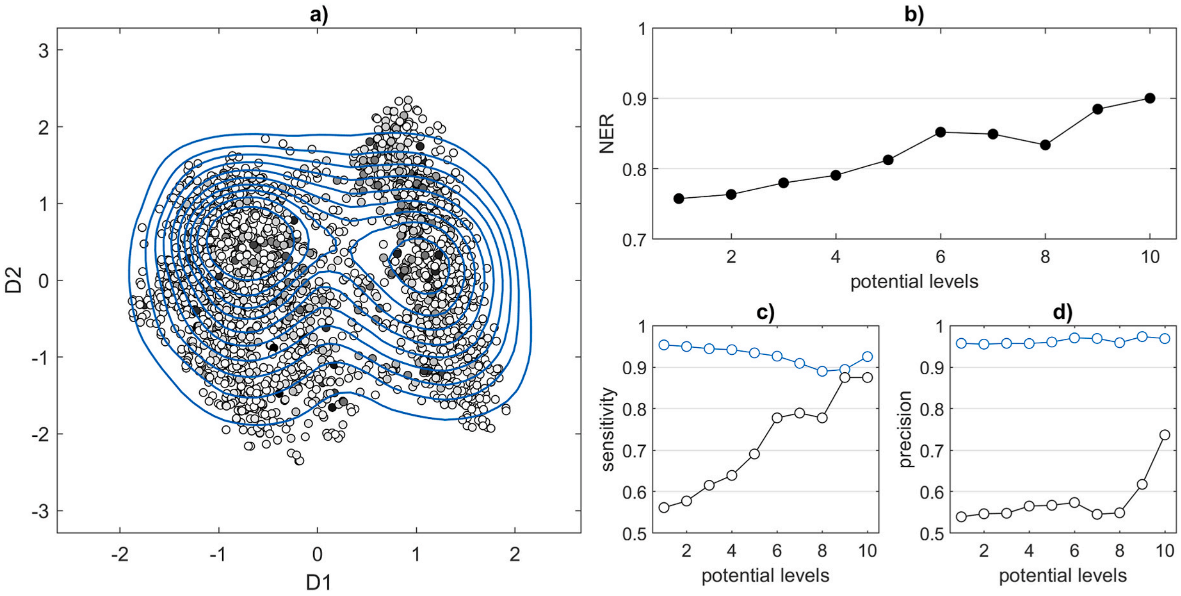

Fig. 5. Mapping of reliability in prediction for the CATMoS very toxic case study: (a) scores of the training compounds in the first two MDS dimensions; molecules are coloured with a greyscale: the darker the colour, the higher the number of misclassifications; the potential levels are coloured in blue; (b) plot of *NER* of test chemicals vs. potential levels; (c) plot of class sensitivity of test chemicals (blue: positive; black: negative) vs. potential levels; (d) plot of class precision of test chemicals (blue: positive; black: negative) vs. potential levels. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

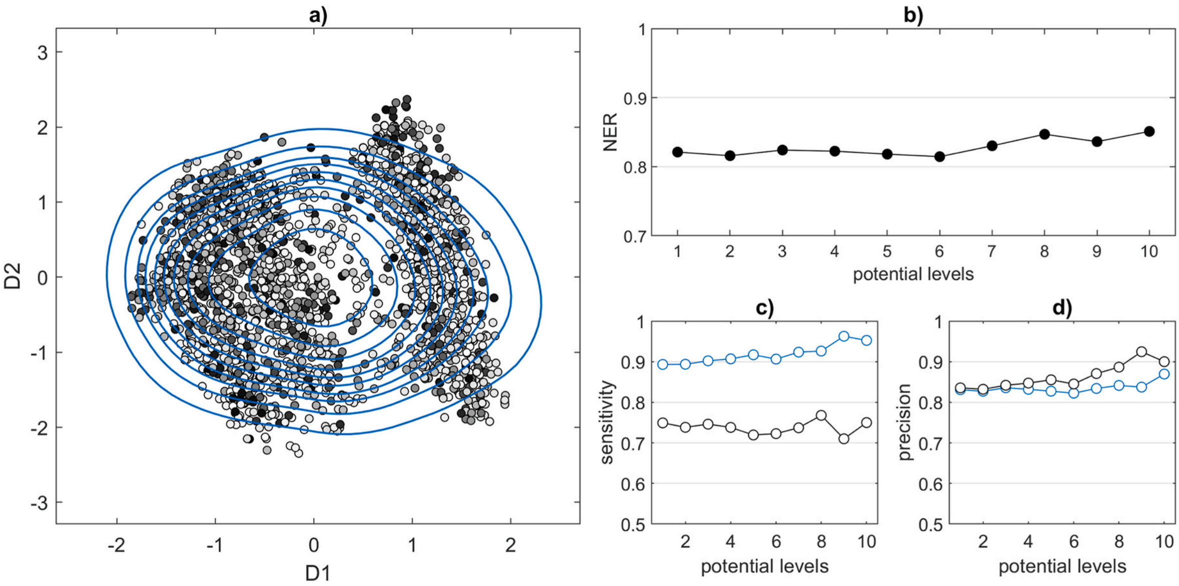

Fig. 6. Mapping of reliability in prediction for the CATMoS nontoxic case study: (a) scores of the training compounds in the first two MDS dimensions; molecules are coloured with a greyscale: the darker the colour, the higher the number of misclassifications; the levels of potential are coloured in blue; (b) plot of *NER* of test chemicals vs. potential levels; (c) plot of class sensitivity of test chemicals (blue: positive; black: negative) vs. potential levels; (d) plot of class precision of test chemicals (blue: positive; black: negative) vs. potential levels. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

(Fig. 5c and d) change significantly: when increasing the potential level, sensitivity of the negative class increases from 0.57 to 0.88, while the sensitivity of the positive class remains substantially constant and higher than 0.90. Precision of the negative class has a significant rise only when chemicals located in the inner levels are taken into account: it increases from 0.54 to 0.74 in the latest 3 levels.

When dealing with the CATMoS nontoxic case study, *NER*, sensitivities and precision do not increase when increasing the potential level (Fig. 6b–. c and d). The cause of this behaviour can be explained by looking at the chemical space (Fig. 6a), where training molecules associated with erroneous predictions (black and dark grey points) look to be spread out and do not cluster as in the previous case studies. Therefore, it is not possible to identify regions where consensus predictions are expected to have better accuracy.

For both case studies, chemicals spaces defined with Path fingerprints and OPERA molecular descriptors provided again unsuitable representations of the potential distribution, as shown in Figs. S3, S4, S5 and S6 of the supplementary material.

3.3. Extending the chemical space to more dimensions

As a consequence of results shown in Fig. 6, one option to get better representation of the mapping of reliability for the CATMoS nontoxic case study is to represent the chemical space taking into account more than two dimensions. This can enable a better definition of the chemical space, even if its graphical representation is not possible anymore. Fig. 7 shows the results when taking into account the first 5 MDS dimensions for the calculation of the potential, with Epanechnikov as kernel function and bandwidth equal to 0.4. With respect to the results obtained with 2 dimensions (Fig. 6), the extension of the chemical space dimensions allows *NER* of test chemicals to linearly increase from approximately 0.82 (in low potential region) to 0.90 (in high potential region). The same pattern can be observed when looking at sensitivity of the negative class (black line in Fig. 7b), which linearly increases when augmenting the potential levels from 0.75 to 0.91. The same trend can be noticed for the precision of the positive class (blue line in Fig. 7c). Therefore, augmenting the dimension for the definition of the chemical space allowed a better mapping of reliability in prediction for this case study, as demonstrated by the better classification performances associated to test chemicals with higher potential.

3.4. Regression: CATMoS logLD50

Beside the application of the kernel-based reliability mapping in the framework of classification, we tested it also with regression consensus modelling. The CATMoS logLD50 case study was used for this aim.

Even in this case, the potential for test molecules was calculated in the chemical space defined through the first two MDS dimensions. The best kernel function to define the mapping was again the Epanechnikov kernel, with bandwidth equal to 1. RMSEP was used as a figure of merit to evaluate the regression performance on test molecules as a function of their reliability potential (Fig. 8). RMSEP substantially decreases when increasing the potential level, even if fluctuations are observed when considering test molecules located in regions associated with lower values of potential. Nevertheless, RMSEP is equal to 0.525 when considering all test chemicals (first level of potential), while it decreases to 0.468 when considering only molecules with the highest potential (those included in the last level).

This pattern can be observed also looking at the experimental vs predicted response plots in Fig. 9. Plots were drawn taking into account only chemicals included in the potential levels number 2, 4, 6, 8 and 10, respectively, that is chemicals with reliability potential higher than that associated with these levels. Molecules are coloured with a grey scale indicating the residuals between predicted and experimental logLD50 (high residuals: black, low residuals: white). RMSEP is shown in the plot title and decreases when increasing the level (according to Fig. 8).

Molecules with highest residuals (coloured in black and associated with the highest prediction errors) have lower potential and, in fact, they are excluded step by step when increasing the potential levels. This can therefore be considered as a proof of the augmented belief associated with higher potential even for quantitative modelling.

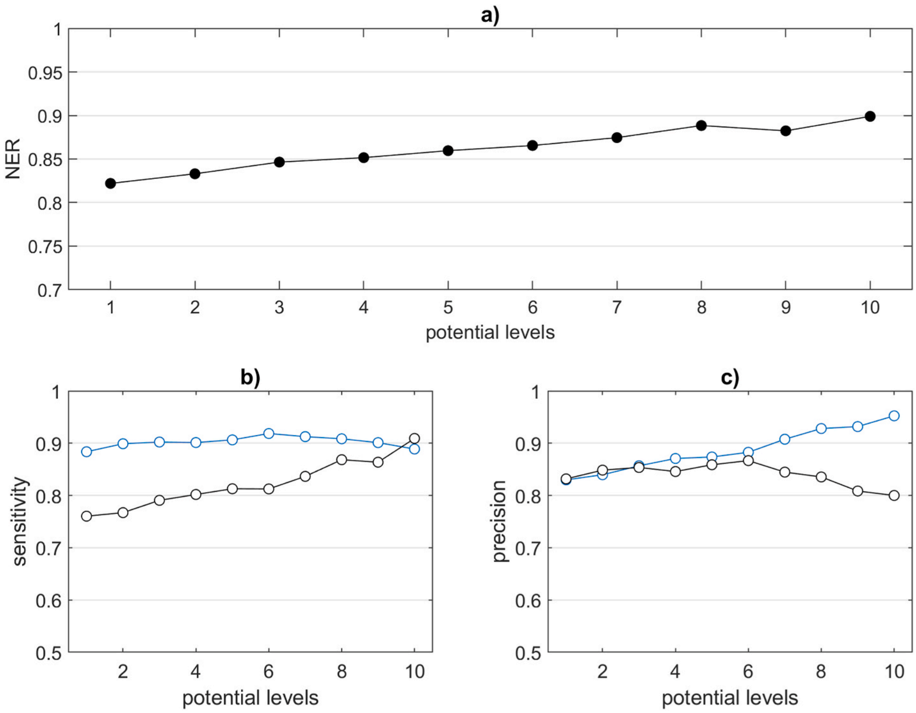

Fig. 7. Mapping of reliability in prediction for the CATMoS nontoxic case study with chemical space extended to 5 dimensions: (a) plot of *NER* values of test chemicals as a function of potential levels; (b) plot of class sensitivity of test chemicals (blue: positive; black: negative) as a function of potential levels; (c) plot of class precision of test chemicals (blue: positive; black: negative) as a function of potential levels. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

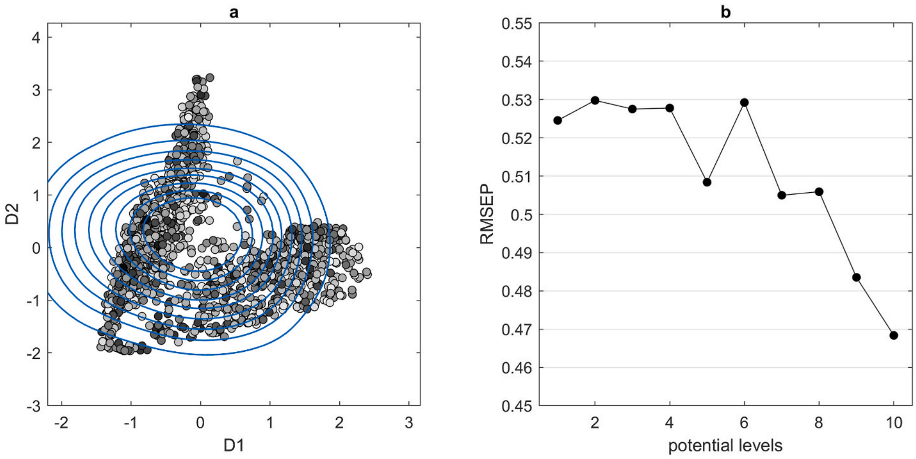

Fig. 8. Mapping of reliability in prediction for the CATMoS logLD50 case study: (a) scores of the training compounds in the first two MDS dimensions; molecules are coloured with a greyscale: the darker the colour, the higher the regression residuals; the levels of potential are coloured in blue; (b) plot of RMSEP as a function of potential levels. (For interpretation of the references to colour in this figure legend, the reader is referred to the Web version of this article.)

4. Conclusions

In this study we propose a new metric to assess the reliability in the degree of accuracy of consensus predictions associated with specific regions of the chemical space. This metric is calculated with kernel-based potentials and it can be used in the framework of both classification and regression consensus modelling. It can assist the prediction of untested compounds with reliability estimation: the higher the potential for a given compound, the higher the belief the compound is located in a chemical region where predictions are accurate.

The effectiveness of the proposed measure was evaluated with four case studies, including extensive consensus datasets. For each case study, chemicals were divided in training and test sets and mapped in the chemical space defined by their binary fingerprints; the training compounds were used to calculate the potential, while the goodness of the proposed measure was assessed taking into account the test chemicals.

Results demonstrated that higher reliability potential is associated with specific regions of the chemical space where consensus modelling performs better in terms of accuracy, both for qualitative and quantitative responses. Therefore, the kernel-based potential can be used to assess how much each point of the chemical space, and consequently the compounds located in that point, is recurrently associated with prediction accuracy of consensus modelling. This can enable the mapping of

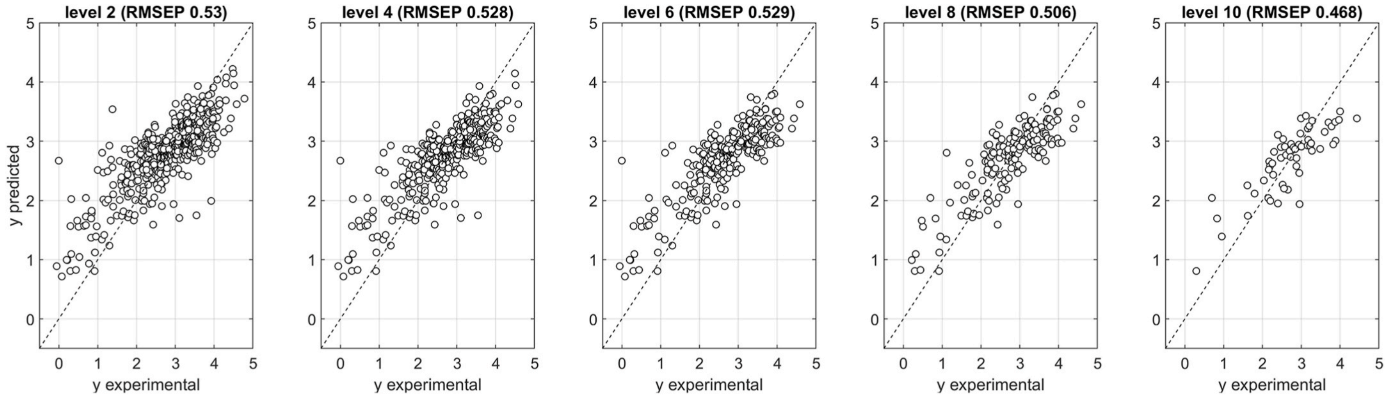

Fig. 9. Plot of experimental vs predicted logLD50 as a function of potential levels. RMSEP and potential level number are shown in the plot title; molecules are coloured with a grey scale indicating the residuals between predicted and experimental response, high residuals: black, low residuals: white.

reliability in predictions and enhance the definition of a comfort zone, where consensus predictions are expected to have a steady level of performance and better accuracy.

The definition of an appropriate chemical space can influence the results. Because it depends on the type of adopted descriptors, future studies will be devoted to testing the effect that different types of descriptors can have on the definition of the chemical space and therefore on the subsequent calculation of the reliability potential.

Finally, in addition to the application in the consensus framework, in future studies the proposed kernel-based mapping could be also tested and extended to assess reliability in predictions for individual models.

CRediT authorship contribution statement

Viviana Consonni: Writing – review & editing, Writing – original draft, Supervision, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Roberto Todeschini:** Writing – review & editing, Supervision, Methodology, Investigation, Formal analysis, Conceptualization. **Marco Orlandi:** Writing – review & editing, Supervision. **Davide Ballabio:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization.

Declaration of competing interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Data availability

Data and code are made available on a online repository

Acknowledgements

Although the topic of this paper is a little far from his main research topics, we would like to dedicate it to Svante Wold. We think he would have appreciated how we can turn new chemometric methods into novel QSAR tools.

Appendix A. Supplementary data

Supplementary data to this article can be found online at https://doi.org/10.1016/j.chemolab.2024.105085.

References

[1] R.G. Brereton, J. Jansen, J. Lopes, F. Marini, A. Pomerantsev, O. Rodionova, J. M. Roger, B. Walczak, R. Tauler, Chemometrics in analytical chemistry part I: history, experimental design and data analysis tools, Anal. Bioanal. Chem. 409 (2017) 5891–5899.

[2] B. Khaleghi, A. Khamis, F.O. Karray, S.N. Razavi, Multisensor data fusion: a review of the state-of-the-art, Inf. Fusion 14 (2013) 28–44.

[3] E. Borràs, J. Ferré, R. Boqué, M. Mestres, L. Aceña, O. Busto, Data fusion methodologies for food and beverage authentication and quality assessment - a review, Anal. Chim. Acta 891 (2015) 1–14.

[4] D.L. Hall, S.A.H. McMullen, Mathematical Techniques in Multisensor Data Fusion, Artech House, Inc ., Norwood, MA, 2004.

[5] F. Castanedo, A review of data fusion techniques, Sci. World J. 2013 (2013).

[6] M.B. Neumann, W. Gujer, Underestimation of uncertainty in statistical regression of environmental models: influence of model structure uncertainty, Environ. Sci. Technol. 42 (2008) 4037–4043.

[7] C.L. Weber, J.M. Vanbriesen, M.S. Small, A stochastic regression approach to analyzing thermodynamic uncertainty in chemical speciation modeling, Environ. Sci. Technol. 40 (2006) 3872–3878.

[8] C. Valsecchi, F. Grisoni, V. Consonni, D. Ballabio, Consensus versus individual QSARs in classification: comparison on a large-scale case study, J. Chem. Inf. Model. 60 (2020) 1215–1223.

[9] A. Fernández, R. Rallo, F. Giralt, Uncertainty reduction in environmental data with conflicting information, Environ. Sci. Technol. 43 (2009) 5001–5006.

[10] F. Sahigara, K. Mansouri, D. Ballabio, A. Mauri, V. Consonni, R. Todeschini, Comparison of different approaches to define the applicability domain of QSAR models, Molecules 17 (2012) 4791–4810.

[11] A. Fernández, A. Lombardo, R. Rallo, A. Roncaglioni, F. Giralt, E. Benfenati, Quantitative consensus of bioaccumulation models for integrated testing strategies, Environ. Int. 45 (2012) 51–58.

[12] D. Ballabio, F. Biganzoli, R. Todeschini, V. Consonni, Qualitative consensus of QSAR ready biodegradability predictions, Toxicol. Environ. Chem. 99 (2017) 1193–1216.

[13] V. Consonni, F. Gosetti, V. Termopoli, R. Todeschini, C. Valsecchi, D. Ballabio, Multi-task neural networks and molecular fingerprints to enhance compound identification from LC-MS/MS data, Molecules 27 (2022) 5827.

[14] C. Valsecchi, M. Collarile, F. Grisoni, R. Todeschini, D. Ballabio, V. Consonni, Predicting molecular activity on nuclear receptors by multitask neural networks, J. Chemometr. 36 (2022) e3325.

[15] D.W. Scott, Multivariate Density Estimation: Theory, Practice, and Visualization, John Wiley & Sons, 2015.

[16] A.W. Bowman, A. Azzalini, Applied Smoothing Techniques for Data Analysis, Oxford University Press Inc., New York, 1997.

[17] T. Hastie, R. Tibshirani, J.H. Friedman, The Elements of Statistical Learning: Data Mining, Inference, and Prediction, Springer Science, 2009.

[18] G.A.F. Seber, Multivariate Observations, John Wiley & Sons, Inc., Hoboken, NJ, USA, 2008.

[19] K. Mansouri, N. Kleinstreuer, A.M. Abdelaziz, D. Alberga, V.M. Alves, P. L. Andersson, C.H. Andrade, F. Bai, I. Balabin, D. Ballabio, E. Benfenati, B. Bhhatarai, S. Boyer, J. Chen, V. Consonni, S. Farag, D. Fourches, A.T. Garc+¡a-Sosa, P. Gramatica, F. Grisoni, C.M. Grulke, H. Hong, D. Horvath, X. Hu, R. Huang, N. Jeliazkova, J. Li, X. Li, H. Liu, S. Manganelli, G.F. Mangiatordi, U. Maran, G. Marcou, T. Martin, E. Muratov, D. Nguyen, O. Nicolotti, N.G. Nikolov, U. Norinder, E. Papa, M. Petitjean, G. Piir, P. Pogodin, V. Poroikov, X. Qiao, A. M. Richard, A. Roncaglioni, P. Ruiz, C. Rupakheti, S. Sakkiah, A. Sangion, K. Schramm, C. Selvaraj, I. Shah, S. Sild, L. Sun, O. Taboureau, Y. Tang, I. Tetko, V. R. Todeschini, W. Tong, D. Trisciuzzi, A. Tropsha, G. Van Den Driessche, A. Varnek, Z. Wang, E.B. Wedebye, A.J. Williams, H. Xie, A. Zakharov, V, Z. Zheng, R. S. Judson, CoMPARA: collaborative modeling project for androgen receptor activity, Environ. Health Perspect. 128 (2020) 027002.

[20] F. Grisoni, V. Consonni, D. Ballabio, Machine learning consensus to predict the binding to the androgen receptor within the CoMPARA project, J. Chem. Inf. Model. 59 (2019) 1839–1848.

[21] K. Mansouri, A. Karmaus, J. Fitzpatrick, G. Patlewicz, D. Alberga, N. Alepee, E. H. Allen, D. Allen, V.M. Alves, C.H. Andrade, T.R. Auernhammer, D. Ballabio, S. Bell, E. Benfenati, S. Bhattacharya, J.V. Bastos, S. Boyd, J.B. Brown, S.J. Capuzzi, Y. Chushak, H. Ciallella, A.M. Clark, V. Consonni, P.R. Daga, S. Ekins, S. Farag, M. Fedorov, D. Fourches, D. Gadaleta, F. Gao, J.M. Gearhart, G. Goh, J. M. Goodman, F. Grisoni, C.M. Grulke, T. Hartung, M. Hirn, P. Karpov, A. Korotcov, G.J. Lavado, M. Lawless, X. Li, T. Luechtefeld, F. Lunghini, G.F. Mangiatordi, G. Marcou, D. Marsh, T. Martin, A. Mauri, E.N. Muratov, G.J. Myatt, D. Nguyen, O. Nicolotti, R. Note, P. Pande, A.K. Parks, T. Peryea, A.H. Polash, R. Rallo, A. Roncaglioni, C. Rowlands, P. Ruiz, D.P. Russo, A. Sayed, R. Sayre, T. Sheils, C. Siegel, A.C. Silva, A. Simeonov, S. Sosnin, N. Southall, J. Strickland, Y. Tang, B. Teppen, I. Tetko, D. Thomas, V. Tkachenko, R. Todeschini, C. Toma, I. Tripodi, D. Trisciuzzi, A. Tropsha, A. Varnek, K. Vukovic, Z. Wang, L. Wang, K.M. Waters, A. J. Wedlake, S.J. Wijeyesakere, D. Wilson, Z. Xiao, H. Yang, G. Zahoranszky-Kohalmi, A.V. Zakharov, F.F. Zhang, Z. Zhang, T. Zhao, H. Zhu, K.M. Zorn, W. Casey, N.C. Kleinstreuer, CATMoS: collaborative acute toxicity modeling suite, Environ. Health Perspect. 129 (2021) 47013.

[22] R. Todeschini, V. Consonni, Molecular Descriptors for Chemoinformatics, Wiley-VCH, 2009.

[23] K. Mansouri, C.M. Grulke, R.S. Judson, A.J. Williams, OPERA models for predicting physicochemical properties and environmental fate endpoints, J. Cheminf. 10 (2018) 10.

[24] D.J. Rogers, T.T. Tanimoto, A computer program for classifying plants, Science (New York, N.Y.) 132 (1960) 1115–1118.

[25] D. Ballabio, F. Grisoni, R. Todeschini, Multivariate comparison of classification performance measures, Chemometr.Intel.Lab.Syst. 174 (2018) 33–44.

[26] A. Mauri, V. Consonni, M. Pavan, R. Todeschini, DRAGON software: an easy approach to molecular descriptor calculations, MATCH 56 (2006) 237–248.

[27] Michem website: https://michem.unimib.it/download/matlab-toolboxes/reliability-potential-toolbox-for-matlab/.