Fuzzy Complex Number Aided Evaluation of Predictive Toxicology Models

Xin Fu*, Kim Travis†, Daniel Neagu‡, Mick Ridley‡ and Qiang Shen§
*School of Management, Xiamen University, Xiamen, 361005, China
Email: fu_xin@hotmail.co.uk
†Syngenta Ltd, Jealott's Hill International Research Centre, Bracknell, Berkshire, RG42 6EY, UK
Email: kim.travis@syngenta.com
‡School of Computing, Informatics and Media, University of Bradford, Richmond Road, Bradford, BD7 1DP, UK
Emails: Daniel Neagu - d.neagu@bradford.ac.uk and Mick Ridley - m.j.ridley@bradford.ac.uk
§ Department of Computer Science, Aberystwyth University, Aberystwyth, SY23 3DB, UK
Email: qqs@aber.ac.uk

***Abstract*—There is a growing interest in applying computational intelligence in the predictive toxicology (PT) domain, where a large number of predictive models are becoming available. Evaluation of such models is therefore considered to be a crucial part of their development and potential use, especially for regulatory purposes. The current evaluation approaches mainly focus on statistical measures of model performance, and few of them have taken data quality into consideration. However, it has been well recognised that datasets and models should not be considered in isolation. This paper proposes a new confidence index for evaluating PT models. A fuzzy complex number (FCN) framework is expanded in an effort to represent and evaluate dataset and regression-based model quality in a two-dimensional manner, thereby ensuring the linguistic evaluation is transparent and explainable. The utility and applicability of this research is illustrated by an experiment which evaluates 17 regression-based PT models. The experimental results have been compared and analysed against existing methods, and show that the FCN-based approach provides a consistent and interpretable means of model assessment. The proposed indexing mechanism can be used, together with customised statistical measures, in assisting PT model selection. This approach also helps to capture the relationships between datasets and models, and contributes to the development of data and model governance in PT.**

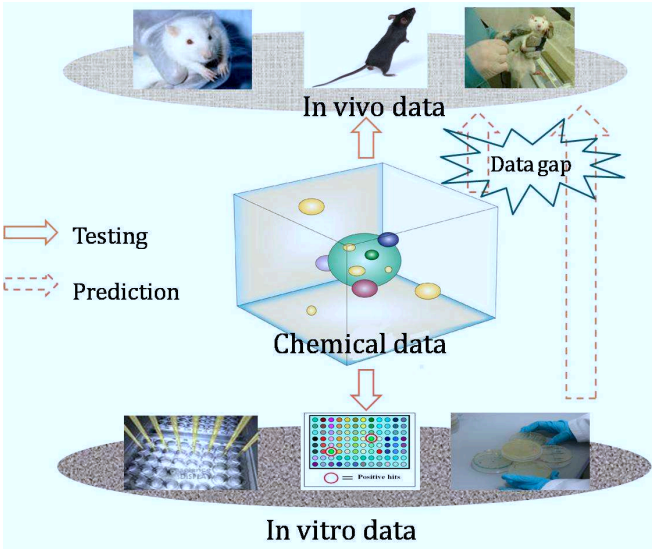

Fig. 1. Data domains for predictive toxicology

I. Introduction

Predictive toxicology (PT) aims to make use of existing chemical and biological/toxicological data, in conjunction with *in silico* modelling techniques to predict chemical toxicity. This will help to reduce the number of animal tests and speed up the chemical compound discovery process (see Fig. 1). The new REACH legislation [1] would require more animal testing to register a new chemical compound, if no alternative methods are used. This has resulted in the growing attention and importance of PT.

Collection of sufficient *in vivo* data (refers to data collected from studies done in live organisms) - *in vitro* data (refers to data collected from studies done in cell-based systems and at the molecular level) is a starting point to build PT models [2], [3]. Due to the diversity of species and endpoints, and the expense of animal testing, there is a huge data gap between *in vivo* data and chemical space. There is a need to use chemical properties, (in conjunction with *in vitro* data) to predict toxicity and to prioritise animal testing. A common technique is to build Quantitative Structure-Activity Relationships (QSAR) models that relate the chemical structures to their measured effect or activity, in an effort to predict toxicity [4].

Undoubtedly, data quality plays an important role in QSAR modelling and inherently affects model quality. As shown in Fig. 1, PT involves multidisciplinary data, so data quality may refer to not only the storage sense (e.g. data accuracy, consistency, completeness and integrity), but also the toxicological sense (e.g. the quality of experimental design and results). For example, whether the experiment was performed to Good Laboratory Practice standards (GLP), and whether the identity and purity of the tested chemical compounds were confirmed would affect the quality of experimental results. Currently, there is a lack of standard measures for quality assessment. While several data quality criteria are objective and defined by mathematical equations, lots of others need

more subjective evaluation and may vary depending on the manner or context in which they are assessed. As a result, it is quite difficult to quantify quality by using pure numerical values, and toxicologists prefer to employ linguistic terms (e.g. as provided by the most commonly used Klimisch criteria [5]: *reliable without restrictions*, *reliable with restrictions*, *not reliable* and *not assignable*) when performing data assessment.

In addition, with the increasing amount of varied toxicity data become publicly available, there is expected to be a boom in the number of QSAR models. A variety of measures have been proposed to describe model quality with respect to different criteria, ranging from accuracy, complexity, robustness, predictivity to applicability domain. In principle, quality measures can be qualitative or quantitative. Quantitative measures are naturally expressed by numerical values. However, using such seemingly precise measures to compare a number of models, their quality may turn out to be very close in value [6]. Given a chemical compound, selecting the most appropriate QSAR model amongst a number of choices whose performances are represented by close numerical values is a challenge faced by end users. It would be more appropriate and often desirable to describe the relative quality of models using linguistic terms, such as *good*, *average* and *bad*. This is because human beings appear to use qualitative reasoning when initially attempting to gain an understanding of a problem.

One of the major challenges in the PT domain is the understanding of how uncertainty inherent in the data may affect the usefulness and applicability of resulting QSAR models. Datasets with errors and missing information result in poor predictive performance and low statistical fit. However, if a QSAR model is built in a robust way upon a poor quality training dataset, its reliability and usefulness is still questionable. It has been well recognised that datasets and models should not be considered in isolation [7]. Due to the involvement of multi-modal uncertainties that arise from different components, the predictions derived from QSAR models need to be further assessed and validated. It is highly desirable to establish an index which represents the level of confidence of derived predictions [8].

The QSAR evaluation process is complex itself. Existing research mainly focuses on aggregating various statistical measures of model performance into a single index (e.g. [8], [9], [10]). Whilst the importance of data quality of PT has been greatly emphasised, very little existing work (e.g. [11]) has taken data quality into consideration when evaluating QSAR models. In terms of dealing with subjective assessments, existing methods are all framed by either simply ignoring them or using symbolic labels to simplify the calculation. Moreover, the sophistication of the evaluation process increases, as the interpretabiliy and transparency of QSAR evaluation/validation is also required by regulatory communities [12].

To alleviate the above difficulty, this paper extends a fuzzy complex number (FCN)-based approach [6], [13], [14] to represent QSAR data quality and model quality (with a particular focus on regression-based models) concurrently and explicitly, without necessarily integrating them into a single value. The derived FCNs can be used as an associated confidence index to assist QSAR model selection. Thanks to the properties of FCNs, the semantic meanings associated with data and models are well preserved. For this reason, the resulting confidence index would be more easily understable and trustworthy for end users. Furthermore, this work will be helpful in performing data and model quality assessment and achieving better management of available datasets and models.

The remainder of this paper is organised as follows. Section II introduces the notion of FCNs [6], which extends real-valued complex numbers to representing two-dimensional uncertainties. In Section III, this notion is utilised to develop a novel method to provide a confidence index for ranking regression-based QSAR models. Section IV shows the experiment carried out and discusses the results. The final section concludes the paper and identifies main directions for further work.

II. Fuzzy Complex Numbers

A. Prerequisites

1) Fuzzy numbers: Fuzzy numbers are special types of fuzzy sets which can be used to represent imprecise quantities such as *about 0.6*. Fuzzy numbers map real values from $\mathbb{R}$ on to a closed interval $[0, 1]$.

Definition 1: (Fuzzy numbers [15]) A fuzzy number, $\tilde{a}$, is defined as:

$$\tilde{a} = \{(x, \mu_{\tilde{a}}(x)) \mid \mu_{\tilde{a}}(x) \in [0, 1], x \in \mathbb{R}\},$$

and satisfies the following properties:

a) Continuity: $\mu_{\tilde{a}}(x)$ is a continuous function mapping from $\mathbb{R}$ to a closed interval $[0, 1]$.
b) Normality: i.e. $\exists x \in \mathbb{R}$ and $\mu_{\tilde{a}}(x) = 1$.
c) Convexity: i.e. $\forall x, y, z \in \mathbb{R}$, if $x \leq y \leq z$ then $\mu_{\tilde{a}}(y) \geq \min(\mu_{\tilde{a}}(x), \mu_{\tilde{a}}(z))$.
d) Boundness of support: i.e. $\exists S \in \mathbb{R}$ and $\forall x \in \mathbb{R}$, if $|x| \geq S$ then $\mu_{\tilde{a}}(x) = 0$.

2) Extension principle: The extension principle [16] provides a fundamental mechanism to translate conventional Boolean set-based concepts into their fuzzy-set counterparts. In this work, it forms the foundation to derive the arithmetic operations of the proposed FCNs.

Definition 2: Let f: $\mathbb{R}^n \longrightarrow \mathbb{R}$ be a function and $A_1, \ldots, A_n$ be fuzzy sets. Then $B = f(A_1, \ldots, A_n)$ is a fuzzy set with the following membership function:

$$\mu_B(y) = \bigvee_{y=f(x_1,\ldots,x_n)} (\mu_{A_1}(x_1) \wedge \cdots \wedge \mu_{A_n}(x_n)). \quad (1)$$

Note that the operators $\wedge$ and $\vee$ above denote a given t-norm and s-norm respectively. Throughout this paper, they are interpreted using the $\min$ and $\max$ operators.

B. Definition of FCNs

Inheriting from the real complex numbers, an FCN, $\tilde{z}$, is defined in the form of:

$$\tilde{z} = \tilde{a} + i\tilde{b}, \tag{2}$$

where both $\tilde{a}$ and $\tilde{b}$ are fuzzy numbers with membership functions $\mu_{\tilde{a}}(x)$ and $\mu_{\tilde{b}}(x)$, regarding a given domain variable x. $\tilde{a}$ is the real part of $\tilde{z}$ while $\tilde{b}$ represents the imaginary part, i.e. $Re(\tilde{z}) = \tilde{a}$ and $Im(\tilde{z}) = \tilde{b}$.

An FCN can be visually shown as in Figure. 2. Importantly, in general, for a given $\tilde{z}$, both $Re(\tilde{z})$ and $Im(\tilde{z})$ are fuzzy. If $\tilde{b}$ does not exist, $\tilde{z}$ degenerates to a fuzzy number. Further, if $\tilde{b}$ does not exist and $\tilde{a}$ itself degenerates to a real number, then $\tilde{z}$ degenerates to a real number.

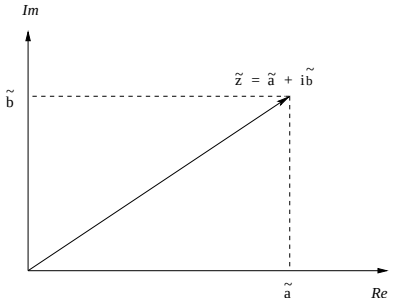

Fig. 2. A fuzzy complex number

C. Operations on FCNs

The operations on the proposed FCNs are a straightforward extension of those on real complex numbers. Let $\widetilde{z}_1 = \tilde{a} + i\tilde{b}$ and $\widetilde{z}_2 = \tilde{c} + i\tilde{d}$ be two FCNs, where $\tilde{a}$, $\tilde{b}$, $\tilde{c}$ and $\tilde{d}$ are fuzzy numbers with membership functions $\mu_{\tilde{a}}(x)$, $\mu_{\tilde{b}}(x)$, $\mu_{\tilde{c}}(x)$ and $\mu_{\tilde{d}}(x)$, respectively. The basic arithmetic operations on $\widetilde{z}_1$ and $\widetilde{z}_2$ are defined as follows:

- Addition

$$\widetilde{z}_1 + \widetilde{z}_2 = (\tilde{a} + \tilde{c}) + i(\tilde{b} + \tilde{d}), \tag{3}$$

where $\tilde{a} + \tilde{c}$ and $\tilde{b} + \tilde{d}$ are newly derived fuzzy numbers with the following membership functions:

$$\begin{aligned} \mu_{\tilde{a}+\tilde{c}}(y) &= \bigvee_{y=x_1+x_2} (\mu_{\tilde{a}}(x_1) \wedge \mu_{\tilde{c}}(x_2)), \\ \mu_{\tilde{b}+\tilde{d}}(y) &= \bigvee_{y=x_1+x_2} (\mu_{\tilde{b}}(x_1) \wedge \mu_{\tilde{d}}(x_2)). \end{aligned} \tag{4}$$

- Subtraction

$$\widetilde{z}_1 - \widetilde{z}_2 = (\tilde{a} - \tilde{c}) + i(\tilde{b} - \tilde{d}), \tag{5}$$

where $\tilde{a} - \tilde{c}$ and $\tilde{b} - \tilde{d}$ are newly derived fuzzy numbers with the following membership functions:

$$\begin{aligned} \mu_{\tilde{a}-\tilde{c}}(y) &= \bigvee_{y=x_1-x_2} (\mu_{\tilde{a}}(x_1) \wedge \mu_{\tilde{c}}(x_2)), \\ \mu_{\tilde{b}-\tilde{d}}(y) &= \bigvee_{y=x_1-x_2} (\mu_{\tilde{b}}(x_1) \wedge \mu_{\tilde{d}}(x_2)). \end{aligned} \tag{6}$$

- Multiplication

$$\widetilde{z}_1 \times \widetilde{z}_2 = (\tilde{a}\tilde{c} - \tilde{b}\tilde{d}) + i(\tilde{b}\tilde{c} + \tilde{a}\tilde{d}), \tag{7}$$

where $\tilde{a}\tilde{c} - \tilde{b}\tilde{d}$ and $\tilde{b}\tilde{c} + \tilde{a}\tilde{d}$ are newly derived fuzzy numbers with the following membership functions:

$$\begin{aligned} \mu_{\tilde{a}\tilde{c}-\tilde{b}\tilde{d}}(y) &= \bigvee_{y=x_1x_2-x_3x_4} (\mu_{\tilde{a}}(x_1) \wedge \mu_{\tilde{c}}(x_2) \wedge \mu_{\tilde{b}}(x_3) \wedge \mu_{\tilde{d}}(x_4)), \\ \mu_{\tilde{b}\tilde{c}+\tilde{a}\tilde{d}}(y) &= \bigvee_{y=x_1x_2+x_3x_4} (\mu_{\tilde{b}}(x_1) \wedge \mu_{\tilde{c}}(x_2) \wedge \mu_{\tilde{a}}(x_3) \wedge \mu_{\tilde{d}}(x_4)). \end{aligned} \tag{8}$$

- Division

$$\frac{\widetilde{z}_1}{\widetilde{z}_2} = \left(\frac{\tilde{a}\tilde{c} + \tilde{b}\tilde{d}}{\tilde{c}^2 + \tilde{d}^2}\right) + i\left(\frac{\tilde{b}\tilde{c} - \tilde{a}\tilde{d}}{\tilde{c}^2 + \tilde{d}^2}\right). \tag{9}$$

For notational simplicity, let $\tilde{t}_1 = \frac{\tilde{a}\tilde{c}+\tilde{b}\tilde{d}}{\tilde{c}^2+\tilde{d}^2}$ and $\tilde{t}_2 = \frac{\tilde{b}\tilde{c}-\tilde{a}\tilde{d}}{\tilde{c}^2+\tilde{d}^2}$, where $\tilde{t}_1$ and $\tilde{t}_2$ are newly derived fuzzy numbers with the following membership functions:

$$\begin{aligned} \mu_{\tilde{t}_1}(y) &= \bigvee_{y=\frac{x_1x_3+x_2x_4}{x_3^2+x_4^2}, x_3^2+x_4^2\neq 0} (\mu_{\tilde{a}}(x_1) \wedge \mu_{\tilde{b}}(x_2) \wedge \mu_{\tilde{c}}(x_3) \wedge \mu_{\tilde{d}}(x_4)), \\ \mu_{\tilde{t}_2}(y) &= \bigvee_{y=\frac{x_2x_3-x_1x_4}{x_3^2+x_4^2}, x_3^2+x_4^2\neq 0} (\mu_{\tilde{a}}(x_1) \wedge \mu_{\tilde{b}}(x_2) \wedge \mu_{\tilde{c}}(x_3) \wedge \mu_{\tilde{d}}(x_4)). \end{aligned} \tag{10}$$

- Modulus

Given $\tilde{z} = \tilde{a} + i\tilde{b}$, the modulus of $\tilde{z}$ is defined:

$$|\tilde{z}| = \sqrt{\tilde{a}^2 + \tilde{b}^2}. \tag{11}$$

It is obvious that $|\tilde{z}|$ is a newly derived fuzzy number with the following membership function:

$$\mu_{|\tilde{z}|}(y) = \bigvee_{y=\sqrt{x_1^2+x_2^2}} (\mu_{\tilde{a}}(x_1) \wedge \mu_{\tilde{b}}(x_2)). \tag{12}$$

The algebraic properties of the proposed FCNs are presented in [6]. These properties, including closure, associativity, commutativity and distributivity are important for further exploration and application of this framework. In particular, the associativity and commutativity of FCNs are used to derive the aggregation of components of an FCN.

D. Aggregation of components of an FCN

In general, there may be more than two components that are of interest in a given problem domain. As such, a hierarchical aggregation approach is necessary. This is possible because of the commutativity of FCNs. That is, any two (components) fuzzy numbers can be selected to construct a working FCN first. Then, the newly derived modulus of this FCN, together with a third fuzzy number can be used to construct another FCN. This process continues until all the involved fuzzy numbers are aggregated. For notation simplicity, arbitrary n components can be represented in one single FCN and each component is denoted as a fuzzy number.

Definition 3: Let $\tilde{a}_1, \ldots, \tilde{a}_n$ be n fuzzy numbers, an aggregation operator τ is defined as:

$$\tau(\tilde{a}_1, \ldots, \tilde{a}_n) = \sqrt{\tilde{a}_1^2 + \cdots + \tilde{a}_n^2}. \tag{13}$$

Note that this aggregation results in a new fuzzy number. Since the multiplication and addition on fuzzy numbers are

commutative (for detailed proof, see [6]), different components can be aggregated in a random order using this aggregation operator.

III. Evaluation of Regression-based QSAR Models

A. System Overview

The problem considered herein is that of QSAR models evaluation, with a particular focus on regression-based models. Different from most existing methods, the FCN-based approach aims to represent and assess both data quality and model quality conjunctively, without necessarily integrating them (as depicted in Fig. 3). A system implemented for this approach involves two main components: *QSAR Modelling* and *FCN-based Evaluation*, with each carrying out certain subtasks as outlined in Fig. 4.

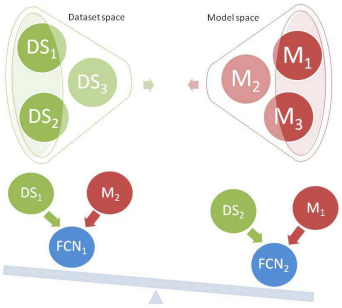

Fig. 3. FCN-based evaluation for QSAR models

The *QSAR Modelling* component covers the quality assessment (QA) and aggregation of individual data and models, involving multidisciplinary data. In general, before conducting the QA of biological/toxicological data, it is important to ensure that the chemical properties (i.e. chemical name, Chemical Abstracts Service Registry Number (CAS) and structure (e.g. SMILES code, mol file or INCHI key)) are consistent and correct.

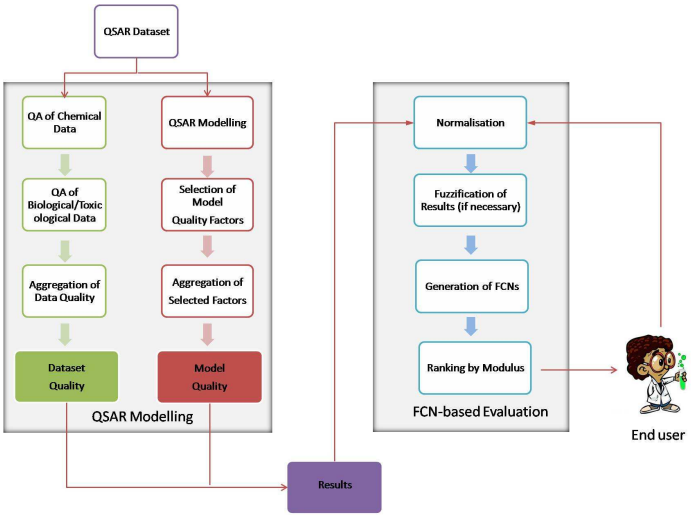

Fig. 4. System overview

QA of biological/toxicological data is not a simple task. The current evaluation is largely subjective and is dependent upon the judgment of assessors with regard to recognised assessment criteria (i.e. adequacy, reliability and relevance). Toxicologists tend to employ linguistic terms to capture the data quality, as is evident in [5], [17]. The next step is to aggregate the data quality from different criteria used by different groups of experts. In addition, the QA of biological/toxicological data may be assigned to individual data point or individual study. When grouping such data points into a dataset for QSAR modelling, an aggregation is required to derive a single quality label to represent the whole dataset. That is the output of the datasets QA and expected to be linguistic terms.

The curated datasets are then used to build QSAR models. There are two main types of QSAR models: classification and regression. Even just for regression QSAR models, which are of particular interest to this work, a variety of factors (e.g. R^2, Q^2, $RMSE$) are often taken into account in the literature [8], [9], [10], [18] to assess the model quality. Especially for regulatory purposes, validation factors such as goodness-of-fit, robustness and predictivity are required to be provided when submitting QSAR models [12]. Hence, selection and aggregation of such factors are essential steps for model quality assessment as highlighted in Fig. 4.

The resulting dataset and model qualities are recorded and fed to the *FCN-based Evaluation* component for further analysis. However, the absolute numerically valued regression results may not be easily understood by the end users. This is particularly true when the number of available models is large and whilst these numerical values are very close to each other. The relative performance may be more informative and humanly interpretable. A normalisation and fuzzification process, which converts the numeric values into fuzzy sets defined in certain universes of discourse, is often employed to accomplish this task. More technical details for implementing such processes can be found in [6].

To support the task of evaluating data and model qualities concurrently, both the real and imaginary parts of FCN need to be assigned with their embedding meanings. In this work, the real part of an FCN is utilised to represent the confidence index of dataset quality, $\widetilde{DS}_{CI}$, which is readily represented in linguistic terms/fuzzy sets, while the imaginary part, $\widetilde{M}_{CI}$, represents the fuzzified confidence index of regression-based model quality. As such, the modulus of the corresponding FCN offers an indication of the overall confidence degree of a given QSAR model:

$$\tilde{z}_{CI} = \widetilde{DS}_{CI} + i\widetilde{M}_{CI}, \tag{14}$$

The overall ranking of all available models is decided on the basis of the moduli of the derived FCNs. According to Equation (11), the modulus of an FCN is also a fuzzy number. Such derived fuzzy numbers can be ranked either by the conventional partial ordering relation holding amongst fuzzy numbers or via defuzzification. Because an absolute ordering

is more desirable in terms of ranking, defuzzification approach is applied for this purpose.

B. An Implemented Regression-based QSAR Models Evaluator

This subsection presents an implemented regression-based QSAR models evaluator which exemplifies the above general evaluation approach. In PT, the work of [11] offers one of very few existing methods which take data quality aspect into account when evaluating QSAR models. With an aim of validating the proposed FCN-based mechanism, the model confidence index (CI) proposed in [11] is employed for comparison. Nine determining factors (see Table I) which are associated with different weights are used to develop the CI as shown in Equation (15).

$$CI = \frac{((R^2) \times 6) \times ((Q^2) \times 6) \times \ln(N_c/10) \times \mathbf{S_{cf}^{0.5}}}{\ln((N_d)^2 + 2) \times \ln((N_m)^2 + 2) \times (((R^2) \times 6) - ((Q^2) \times 6) + 1) \times \mathbf{R_{cf}}} \quad (15)$$

TABLE I
FACTORS DETERMINING THE QUALITY OF QSAR MODELS IN [11]

Factor	Description
R^2	Goodness of fit of the model for training dataset
Q^2	Predictivity of the model for validation dataset
D_{fp}	Stability of the model: described as the difference between R^2 and Q^2
N_c	Number of chemical compounds used in the training dataset
N_d	Number of chemical descriptors used in the model
T_r	Range of toxicity values in training dataset
N_m	Number of mechanisms of toxic action covered by the training dataset
$\mathbf{R_{cf}}$	**Confidence degree of the repeatability and reliablity of biological/toxicological data**
$\mathbf{S_{cf}}$	**Confidence degree of the data source**

As pointed out in [11], the assessment of the repeatability and reliability of biological/toxicological data, $\mathbf{R_{cf}}$ and $\mathbf{S_{cf}}$, is probably the least objective component and may be susceptible to human biases. It is expected that high quality experimental data should be measured in a consistent manner (e.g. using a single protocol, from the same laboratory and preferably by the same worker, and following standardised guidelines), and will have fewer experimental errors associated. However, the work in [11] employs purely numerical values to capture such subjective assessment. For example, R_{cf} is quantified simply as an integer ranging from 1 to 3 (representing high to low confidence degree, respectively) and examples of assessing S_{cf} (ranging from 0 to 1) are given in Table II. Having realised this, relative low weights are assigned to these data quality factors in an effort to reduce their impact upon the overall model quality assessment.

In reality, the quality of the data used underpins the quality of the resulting QSAR models. However, there is no guidance indicates that data quality measures should receive lower weight than other model evaluation measures. Therefore, in this FCN-based approach, such subjective assessments are captured by linguistic terms and are assigned equal weights as model quality factors (as shown in Equation (14)). To facilitate this, two fuzzy quantity spaces are defined to reflect the linguistic terms adopted in [11]. In particular, $Q_{FN1} = \{Low, Medium, High\}$ (see Fig. 5), is used to describe R_{cf}, while $Q_{FN2} = \{Worst, VL, L, M, H, VH, Best\}$ (see Fig. 6) is used for S_{cf} and all other linguistic variables employed in the implemented evaluator.

TABLE II
FACTORS DETERMINING THE S_{cf} AND EXAMPLES [11]

Laboratory	Endpoint	Standardised Protocol	Other Factors	Example	S_{cf}
One	Same	Uniform	3/3	Duluth-Fathead Minnow	1.0
One	Same	Uniform	2/3	Tennessee-Tetratox	0.9
One	Same	Uniform	2/3	Utrecht-Guppy	0.9
> 1	Same	Uniform	2/3	Zebrafish mortality	0.7
> 1	Similar	Similar	1/3	Microbial bioluminescence	0.4
> 1	Varied	Similar	1/3	Daphnids	0.3
> 1	Ambiguous	Varied	0/3	RTECS	0.1

For those variables which are represented by numerical values (e.g. S_{cf}), a fuzzification process is needed, transforming a numerical value into a fuzzy set in Q_{FN}. This is achieved by: 1) normalising a crisp value x, $x \in D$ (the domain of the variable in question), 2) treating the normalised value $\bar{x}$, $\bar{x} \in [0, 1]$ as a special case of triangular fuzzy numbers i.e. $[\bar{x}, \bar{x}, \bar{x}]$, and 3) calculating the degree of similarity between such a specific fuzzy number and an element of Q_{FN}. The element with the highest similarity degree is then selected to represent x. In this work, the popular *Hausdorff* distance [19] is employed to measure the fuzzy set matching degrees.

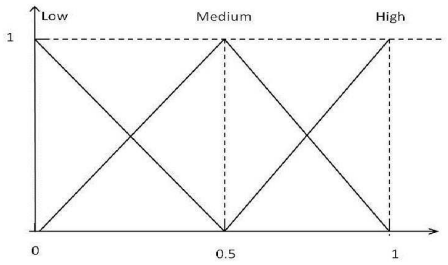

Fig. 5. Fuzzy quantity space, Q_{FN1}, for variable R_{cf}

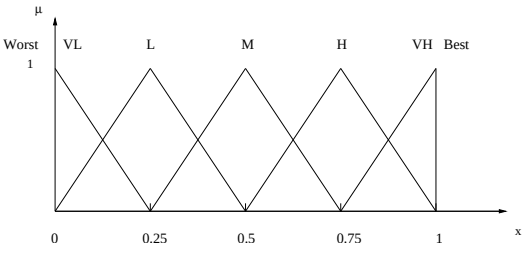

Fig. 6. Fuzzy quantity space, Q_{FN2}, for variables other than R_{cf}

In Equation (14), M_{CI} is defined as in Equation (16) which separates the data quality factors whilst preserving the expertise of assessing regression-based QSAR models as given in [11]. In addition, $\widetilde{DS}_{CI}$ is the joint evaluation of R_{cf} and S_{cf}. Since it involves more than two components in this FCN-based approach, it is intuitive to first select R_{cf} and S_{cf} to construct an FCN, $\widetilde{DS}_{CI}$, to represent the data quality aspect. Then, the newly derived $|\widetilde{DS}_{CI}|$, together with $\widetilde{M}_{CI}$

can be used to construct another FCN, $\widetilde{z}_{CI}$. As discussed in Section II-D, $|\tilde{z}_{CI}|$ can be represented as in Equation (17). To derive an absolute evaluation ranking, $|\tilde{z}_{CI}|$ is defuzzified using its representative value [20]. For computational simplicity, the representative value of a triangular membership function $A = [a, b, c]$ is defined as: $Rep(A) = \frac{a+b+c}{3}$.

$$M_{CI} = \frac{((R^2) \times 6) \times ((Q^2) \times 6) \times \ln(N_c/10)}{\ln((N_d)^2 + 2) \times \ln((N_m)^2 + 2) \times (((R^2) \times 6) - ((Q^2) \times 6) + 1)} \tag{16}$$

$$|\tilde{z}_{CI}| = \sqrt{(\tilde{R}_{cf})^2 + (\tilde{S}_{cf})^2 + (\widetilde{M}_{CI})^2} \tag{17}$$

IV. Experimental Results

A. Experimental data

To demonstrate the applicability and utility of the proposed method for regression-based QSAR model evaluation, a set of experiments are reported in this section, with a particular focus on comparing the results against those of [11]. Two scenarios, each of which contains a variety of combinations of factor values, are created to conduct the experiment. The realistic factor values are either borrowed from examples in [11] or selected from the recommended default values in the literature. The original R_{cf} terms can be readily mapped to Q_{FN1}, while a fuzzification process is performed when converting from the numerical domain of S_{cf} to Q_{FN2}. In addition, the normalised M_{CI} values are also mapped onto Q_{FN2} to derive the corresponding $\widetilde{M}_{CI}$.

B. Results Analysis

A set of factor values of ecotoxicity QSAR models which are used in [11], but with fuzzified S_{cf} are employed in Scenario 1. The results obtained by using the method of [11] and the FCN-based approach are reported in Tables III and IV, respectively. To demonstrate the impact of data quality, diverse values of R_{cf} and S_{cf} based on the same sets of M_{CI} are devised in Scenario 2. The results obtained are shown in Tables V and VI.

TABLE III
CONFIDENCE INDEX EVALUATION RESULTS OF ECOTOXICITY QSAR MODELS (SCENARIO 1) USING THE METHOD IN [11]

Model	R^2	Q^2	N_c	N_d	T_r	N_m	R_{cf}	S_{cf}	CI	Ranking(CI [11])
m_1	0.95	0.95	50	1	5	1	1 (High)	1	159.24	1
m_2	0.95	0.95	50	4	5	1	1 (High)	0.9	57.42	2
m_3	0.95	0.95	50	1	2	1	1 (High)	0.7	53.3	3
m_4	0.95	0.95	50	1	5	1	3 (Low)	0.5	37.53	5
m_5	0.95	0.95	50	1	5	10	1 (High)	0.4	23.93	6
m_6	0.85	0.85	50	1	5	1	1 (High)	0.4	41.36	4
m_7	0.85	0.85	50	4	5	1	1 (High)	0.1	7.86	14
m_8	0.85	0.85	50	1	2	1	1 (High)	0.7	21.89	7
m_9	0.85	0.85	50	1	5	1	3 (Low)	0.6	16.89	10
m_{10}	0.85	0.85	50	1	5	10	1 (High)	0.75	13.46	11
m_{11}	0.95	0.85	50	1	5	1	1 (High)	0.25	18.52	8
m_{12}	0.95	0.85	50	4	5	1	1 (High)	0.9	13.36	12
m_{13}	0.95	0.85	50	1	2	1	1 (High)	0.55	10.99	13
m_{14}	0.95	0.85	50	1	5	1	3 (Low)	0.35	7.31	16
m_{15}	0.95	0.85	50	1	5	10	1 (High)	0.2	3.94	17
m_{16}	0.85	0.75	50	1	5	1	1 (High)	1	17.75	9
m_{17}	0.95	0.75	50	1	5	1	1 (High)	0.25	7.76	15

TABLE V
CONFIDENCE INDEX EVALUATION RESULTS OF ECOTOXICITY QSAR MODELS (SCENARIO 2) USING THE METHOD IN [11]

Model	R^2	Q^2	N_c	N_d	T_r	N_m	R_{cf}	S_{cf}	CI	Ranking(CI [11])
m_1	0.9	0.85	50	1	4	10	1 (High)	1.0	8.66	1
m_2	0.9	0.85	50	1	4	10	1 (High)	0.8	7.74	2
m_3	0.9	0.85	50	1	4	10	2 (Medium)	0.6	3.35	7
m_4	0.9	0.85	50	1	4	10	2 (Medium)	0.3	2.37	10
m_5	0.9	0.85	50	1	4	10	3 (Low)	0.1	0.91	15
m_6	0.85	0.75	100	3	5	5	1 (High)	1.0	3.88	5
m_7	0.85	0.75	100	3	5	5	1 (High)	0.8	3.47	6
m_8	0.85	0.75	100	3	5	5	2 (Medium)	0.6	1.50	13
m_9	0.85	0.75	100	3	5	5	2 (Medium)	0.3	1.06	14
m_{10}	0.85	0.75	100	3	5	5	3 (Low)	0.1	0.41	17
m_{11}	0.95	0.75	100	5	5	1	1 (High)	1.0	7.40	3
m_{12}	0.95	0.75	100	5	5	1	1 (High)	0.8	6.62	4
m_{13}	0.95	0.75	100	5	5	1	2 (Medium)	0.6	2.86	8
m_{14}	0.95	0.75	100	5	5	1	2 (Medium)	0.3	2.03	12
m_{15}	0.95	0.75	100	5	5	1	3 (Low)	0.1	0.78	16
m_{16}	0.95	0.75	100	5	5	1	3 (Low)	1.0	2.47	9
m_{17}	0.95	0.75	100	5	5	1	3 (Low)	0.8	2.21	11

It can be seen from these results that the FCN-based approach is capable of selecting the same first two QSAR models as from [11] in both scenarios. Similar ranking results are obtained for the rest of cases also. The proposed work reflects well the intuition that based on the same M_{CI}, if the model is built upon more reliable data, it receives a higher confidence index (see Table VI for example). In terms of transparency and interpretability, the proposed method outperforms the work in [11]. Take the models m_1 and m_4 in Scenario 1 for example, Table III simply shows that the reason they are ranked as the 1^{st} and 5^{th} is that they receive CI 159.24 and 37.53, respectively. However, the FCN-based approach not only provides similar ranking scores (the 1^{st} and 7^{th} in Table IV), but also explains the underlying reason. Indeed, in terms of model quality, m_1 and m_4 perform equally the *Best* amongst all candidates. The reason for the different ranking scores is that m_1 is built upon *VH* quality training dataset, while the quality of dataset used in m_4 is only M (Medium) amongst all candidates. This makes the results obtained by the use of the present work easier to understand for the end users.

There exists a difference between the two ranking results. This might have been caused by the relatively low weight assigned to S_{cf}. However, as discussed in [11], the reason for assigning a low weight to S_{cf} is due to the lack of sufficient expertise to support such subjective evaluation. In the FCN-based approach, the impacts of data quality and other model performance are assumed to be equal, when no sound knowledge is available. A closer examination into the results reveals that, in addition to the inherent comprehensibility, the evaluation outcomes by the FCN-based approach are intuitively more reliable and consistent. For example, in Table III, m_7 and m_{17} receive very close CI scores (7.86 and 7.76, respectively). Due to the involvement of subjective opinions and noisy data, it appears rather difficult, even unfair, to say one is better than the other just based on such a minute numerical difference. Their relative confidence index to other QSARs may be of more interest to the end users. Similar examples include m_{10} and m_{12} in Scenario 1, and m_2 and

TABLE IV
FCNs-BASED EVALUATION RESULTS OF ECOTOXICITY QSAR MODELS (SCENARIO 1) USING FCNs-BASED APPROACH

Model	$\widetilde{DS}_{CI}$				$\widetilde{M}_{CI}$			$\tilde{z}_{CI}$	Rep($\lvert\tilde{z}_{CI}\rvert$)		Ranking (FCN-based)	Ranking(CI [11])
	R_{cf}		S_{cf}		Absolute	Normalised			Absolute	Normalised		
m_1	1	(High)	1.00	(Best)	159.24	1.0000	(Best)	VH + iBest	1.5814	1.0000	1	1
m_2	1	(High)	0.90	(VH)	60.53	0.3439	(L)	H/VH + iL	1.3129	0.7460	2	2
m_3	1	(High)	0.70	(H)	60.53	0.3439	(L)	H + iL	1.1227	0.5661	6	3
m_4	3	(Low)	0.50	(M)	159.24	1.0000	(Best)	M + iBest	1.1057	0.5500	7	5
m_5	1	(High)	0.40	(L)	37.83	0.1930	(L)	M + iL	0.9610	0.4131	11	6
m_6	1	(High)	0.40	(L)	65.40	0.3762	(M)	M + iM	1.0307	0.4790	9	4
m_7	1	(High)	0.10	(VL)	24.86	0.1068	(VL)	M + iVL	0.8776	0.3342	Joint 13	14
m_8	1	(High)	0.70	(H)	26.16	0.1154	(VL)	H + iVL	1.0881	0.5333	8	7
m_9	3	(Low)	0.60	(H)	65.40	0.3762	(M)	M + iM	0.8709	0.3279	15	10
m_{10}	1	(High)	0.75	(VH)	15.54	0.0048	(VL)	H/VH + iVL	1.1995	0.6387	Joint 4	11
m_{11}	1	(High)	0.25	(VL)	37.04	0.1877	(L)	M + iL	0.9203	0.3746	12	8
m_{12}	1	(High)	0.90	(VH)	14.08	0.0351	(VL)	H/VH + iVL	1.1995	0.6387	Joint 4	12
m_{13}	1	(High)	0.55	(M)	14.82	0.0400	(VL)	H + iVL	0.9928	0.4432	10	13
m_{14}	3	(Low)	0.35	(L)	37.05	0.1878	(L)	L+ iL	0.5243	0.0000	17	16
m_{15}	1	(High)	0.20	(VL)	8.80	0.0000	(Worst)	M + iWorst	0.8603	0.3178	16	17
m_{16}	1	(High)	1.00	(Best)	17.75	0.0595	(VL)	VH + iVL	1.2700	0.7054	3	9
m_{17}	1	(High)	0.25	(VL)	15.51	0.0446	(VL)	M + iVL	0.8776	0.3342	Joint 13	15

TABLE VI
FCNs-BASED EVALUATION RESULTS OF ECOTOXICITY QSAR MODELS (SCENARIO 2) USING FCNs-BASED APPROACH

Model	$\widetilde{DS}_{CI}$				$\widetilde{M}_{CI}$			$\tilde{z}_{CI}$	Rep($\lvert\tilde{z}_{CI}\rvert$)		Ranking (FCN-based)	Ranking(CI [11])
	R_{cf}		S_{cf}		Absolute	Normalised			Absolute	Normalised		
m_1	1	(High)	1.00	(Best)	8.66	1.0000	(Best)	VH + iBest	1.5814	1.0000	1	1
m_2	1	(High)	0.80	(H)	8.66	1.0000	(Best)	H + iBest	1.4326	0.8871	Joint 2	2
m_3	2	(Medium)	0.60	(M)	8.66	1.0000	(Best)	M + iBest	1.2285	0.7322	6	7
m_4	2	(Medium)	0.30	(L)	8.66	1.0000	(Best)	M + iBest	1.1721	0.6895	8	10
m_5	3	(Low)	0.10	(VL)	8.66	1.0000	(Best)	L + iBest	1.0045	0.5623	11	15
m_6	1	(High)	1.00	(Best)	3.88	0.0000	(Worst)	VH + iWorst	1.2700	0.7637	4	5
m_7	1	(High)	0.80	(H)	3.88	0.0000	(Worst)	H + iWorst	1.0737	0.6148	9	6
m_8	2	(Medium)	0.60	(M)	3.88	0.0000	(Worst)	M + iWorst	0.7512	0.3701	14	13
m_9	2	(Medium)	0.30	(L)	3.88	0.0000	(Worst)	M + iWorst	0.6326	0.2801	16	14
m_{10}	3	(Low)	0.10	(VL)	3.88	0.0000	(Worst)	L + iWorst	0.2634	0.0000	17	17
m_{11}	1	(High)	1.00	(Best)	7.40	0.74	(H)	VH + iH	1.4326	0.8871	Joint 2	3
m_{12}	1	(High)	0.80	(H)	7.40	0.74	(H)	H + iH	1.2619	0.7576	5	4
m_{13}	2	(Medium)	0.60	(M)	7.40	0.74	(H)	M + iH	1.0117	0.5678	10	8
m_{14}	2	(Medium)	0.30	(L)	7.40	0.74	(H)	M+ iH	0.9396	0.5131	13	12
m_{15}	3	(Low)	0.10	(VL)	7.40	0.74	(H)	L + iH	0.7381	0.3602	15	16
m_{16}	3	(Low)	1.00	(Best)	7.40	0.74	(H)	H/VH + iH	1.1893	0.7025	7	9
m_{17}	3	(Low)	0.80	(H)	7.40	0.74	(H)	M + iH	0.9767	0.5412	12	11

m_{11} in Scenario 2. The resulting CI scores of the two models are actually extremely close to each other for each of these two pairs. Again, it has a natural appeal to assign the same ranking score to such models, rather than distinguishing them. Hence, the results obtained by the FCN-based approach seem to be more reasonable.

V. CONCLUSION

This paper has proposed a new method of evaluating QSAR models in the PT domain. The initial work on FCNs of [6] has been extended, for the first time, to concurrently represent and evaluate both data and model quality in a unified manner. Also, different from the existing work where classification tasks are addressed, this research mainly focuses on the evaluation of regression-based models. The derived confidence index can assist the end users to make informed decisions when more QSAR models become available. The effectiveness and applicability of this approach is demonstrated by comparing the derived results to a representative approach [11] in the literature. The experimental results show that this new method is capable of providing a consistent and reasonable evaluation, and also outperforms the existing work in a number of aspects. In particular, the FCN-based indexing mechanism is capable of providing the following improved properties:

- *Interpretability and transparency:* As pointed out in [18], to increase regulatory acceptance and the use of QSARs, evaluation criteria that are not only reliable but also easily understandable are highly desirable. Existing methods aim to aggregate various factors into an overall score. In so doing, the underlying semantics associated with those individual factors is destroyed. This makes the derived evaluation outputs difficult to explain, thereby reducing the confidence of using QSAR predictions. To combat this, the FCN-based approach concurrently evaluates qualities in both datasets and models, without necessarily aggregating them. Importantly, the modulus of the derived FCN also has a semantic meaning embedded. This approach therefore facilitates an intuitive understanding of QSAR model evaluation process, providing more transparent and hence trustworthy outcomes for the users and regulatory communities.
- *Capability of handling uncertain information:* It is clear that the work of [11] is, to a great extent, based on the subjective expertise and intuition of the experts. Existing methods for QSAR evaluation either use numerical values to capture such knowledge or employ symbolic terms to simplify the calculation of linguistic terms. This may result in inaccuracy and loss of information. Fuzzy set theory has established itself as a leading tool to handle uncertain knowledge and data. The proposed FCN-based approach seems to be ideally suited to solving such problems.
- *Generality and flexibility:* Although the examples provided in this paper focus on the evaluation of regression-based QSAR models for ecotoxicological endpoints, the proposed FCN-based approach is mathematically generic and can be flexibly adapted to other customisable sets of data and model quality measures. Moreover, for evaluating classification-based QSAR models, the work of [6] can be readily applied to the PT domain.

Although the proposed method is promising, much may be done through future work. It would be interesting to investigate how this method may perform when applied to larger toxicological datasets (ideally with various data quality labels) and other model quality indices. In addition, it would be useful to exploit the proposed mechanism to determine on what combination of datasets and models may jointly perform the best, especially in conjunction with the use of data reliability measures [21], [22]. Also, an important piece of future work is to further validate the evaluation results obtained from this work. A way to implement this is to present results obtained from the FCN approach and other methods in a questionnaire. End users would be invited to choose their preferred evaluation results, ideally with reasons stated.

ACKNOWLEDGMENT

This work was conducted when the first author was with the University of Bradford, UK and Syngenta Ltd, UK on a collaborative project. The authors would like to thank Syngenta Ltd, TSB and BBSRC for sponsoring the Knowledge Transfer Partnership (KTP) Grant no. 7596.

REFERENCES

[1] "REACH," http://ec.europa.eu/environment/chemicals/reach/reach_intro.htm.
[2] C. Helma, Ed., *Predictive Toxicology*. Taylor & Francis Group, 2005.
[3] R. Judson, "Public databases supporting computational toxicology," *Journal of Toxicology and Environmental Health, Part B*, vol. 13, no. 2, pp. 218–231, 2010.
[4] J. Jaworska, N. Nikolova-Jelizkova, and T. Aldenberg, "QSAR applicability domain estimation by projection of the training set in descriptor space: A review," *ATLA. Alternatives to laboratory animals*, vol. 33, pp. 445–459, 2005.
[5] H.-J. Klimisch, M. Andreae, and U. Tillmann, "A systematic approach for evaluating the quality of experimental toxicological and ecotoxicological data," *Regulatory Toxicology and Pharmacology*, vol. 25, no. 1, pp. 1 – 5, 1997. [Online]. Available: http://www.sciencedirect.com/science/article/pii/S0273230096910764
[6] X. Fu and Q. Shen, "Fuzzy complex numbers and their application for classifiers performance evaluation," *Pattern Recognition*, vol. 44, no. 7, pp. 1403–1417, 2011.
[7] X. Fu, A. Wojak, D. Neagu, M. Ridley, and K. Travis, "Data governance in predictive toxicology: A review," *Journal of Cheminformatics*, vol. 3, no. 1, p. 24, 2011. [Online]. Available: http://www.jcheminf.com/content/3/1/24
[8] E. Kolossov and R. Stanforth, "The quality of QSAR models: problems and solutions," *SAR and QSAR in Environmental Research*, vol. 18, no. 1-2, pp. 89–100, 2007.
[9] C. Porcelli, A. Roncaglioni, A. Chana, and E. Benfenati, "A comparison of DEMETRA individual QSARs with an index for evaluation of uncertainty," *Chemosphere*, vol. 71, no. 10, pp. 1845 – 1852, 2008.
[10] A. Lombardo, A. Roncaglioni, E. Boriani, C. Milan, and E. Benfenati, "Assessment and validation of the CAESAR predictive model for bioconcentration factor (BCF) in fish," *Chemistry Central Journal*, vol. 4, no. Suppl 1:S1, pp. 1–11, 2010.
[11] T. Schultz, T. Netzeva, and M. Cronin, "Evaluation of QSARs for ecotoxicity: A method for assigning quality and confidence," *SAR and QSAR in Environmental Research*, vol. 15, no. 5-6, pp. 385–397, 2004.
[12] OECD, "OECD principles for the validation, for regulatory purposes, of QSAR models," http://www.oecd.org/dataoecd/33/37/37849783.pdf.
[13] X. Fu, T. Boongoen, and Q. Shen, "Evidence directed generation of plausible crime scenarios with identity resolution," *Applied Artificial Intelligence*, vol. 24, no. 4, pp. 253–276, 2010.
[14] X. Fu and Q. Shen, "Fuzzy compositional modelling," *IEEE Transactions on Fuzzy Systems*, vol. 18, no. 4, pp. 823–840, 2010.
[15] D. Dubois and H. Prade, *Fuzzy sets and systems: theory and applications*, 4th ed. Academic Press, 1980.
[16] L. A. Zadeh, "Fuzzy logic and approximate reasoning (in memory of Grigore Moisil)," *Synthese*, vol. 30, no. 3/4, pp. 407–428, 1975. [Online]. Available: http://www.jstor.org/stable/20115038
[17] K. Schneider, M. Schwarz, I. Burkholder, A. Kopp-Schneider, L. Edler, A. Kinsner-Ovaskainen, T. Hartung, and S. Hoffmann, "ToxRTool, a new tool to assess the reliability of toxicological data," *Toxicology Letters*, vol. 189, no. 2, pp. 138 – 144, 2009.
[18] L. Eriksson, J. Jaworska, A. P. Worth, M. T. Cronin, R. M. McDowell, and P. Gramatica, "Methods for reliability and uncertainty assessment and for applicability evaluations of classification- and regression-based QSARs," *Environ Health Perspect*, vol. 111, no. 10, pp. 1361–1375, 02 2003.
[19] D. Huttenlocher, G. Klanderman, and W. Rucklidge, "Comparing images using the hausdorff distance," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 15, no. 9, pp. 850–863, 1993.
[20] Z. Huang and Q. Shen, "Fuzzy interpolation and extrapolation: A practical approach," *Fuzzy Systems, IEEE Transactions on*, vol. 16, no. 1, pp. 13 –28, feb. 2008.
[21] T. Boongoen and Q. Shen, "Nearest-neighbor guided evaluation of data reliability and its applications," *IEEE Transactions on Systems, Man, and Cybernetics, Part B*, vol. 40, no. 6, pp. 1622–1633, 2010.
[22] T. Boongoen, C. Shang, N. Iam-On, and Q. Shen, "Extending data reliability measure to a filter approach for soft subspace clustering," *Systems, Man, and Cybernetics, Part B: Cybernetics, IEEE Transactions on*, vol. 41, no. 6, pp. 1705 –1714, dec. 2011.