JOURNAL OF CHEMICAL INFORMATION AND MODELING

Article

pubs.acs.org/jcim

Introducing Uncertainty in Predictive Modeling—Friend or Foe?

Ulf Norinder*[,†,‡,⊥] and Henrik Boström[§]

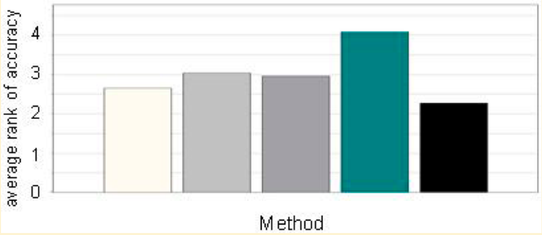
[†]AstraZeneca R&D Södertälje, Sweden

[‡]Department of Pharmacy, Uppsala University, Sweden

[§]Department of Computer and Systems Sciences, Stockholm University, Sweden

S Supporting Information

ABSTRACT: Uncertainty was introduced to chemical descriptors of 16 publicly available data sets to various degrees and in various ways in order to investigate the effect on the predictive performance of the state-of-the-art method decision tree ensembles. A number of strategies to handle uncertainty in decision tree ensembles were evaluated. The main conclusion of the study is that uncertainty to a large extent may be introduced in chemical descriptors without impairing the predictive performance of ensembles and without the predictive performance being significantly reduced from a practical point of view. The investigation further showed that even when distributions of uncertain values were provided, the ensembles method could generate equally effective models from single-point samples from these distributions. Hence, there seems to be no advantage in using more elaborate methods for handling uncertainty in chemical descriptors when using decision tree ensembles as a modeling method for the considered types of introduced uncertainty.



INTRODUCTION

The development of pharmaceutical drugs is today a costly and time-consuming procedure.[1] In order to streamline the drug development procedures, particularly in the lead identification and lead optimization phases, front loading of information has become an important component.[2] One aspect of front loading involves the utilization of modeling and simulation techniques, also commonly referred to as in silico modeling, in order to predict various biopharmaceutical properties, e.g., solubility and ADMET properties, of promising chemical entities (structures). In order for the derived models to be useful, a number of criteria need to be fulfilled. These include properties such as robustness, predictive performance, applicability of domain coverage as well as, in many cases, transparency and interpretability.[3]

Most often, a predictive model is found by employing some statistical or machine learning algorithm to find a mapping from input, e.g., a compound represented by a set of molecular descriptors, to output, e.g., toxicity, using available data. However, some of the employed descriptors, e.g., log P and other charged based variables, are not exact and come with an uncertainty, often expressed in terms of a standard deviation. The latter information is however typically not considered by current in silico predictive models, but instead each chemical descriptor is represented by a single (numerical or categorical) value, hence excluding any information on the uncertainty of the measurement or estimation. One obvious reason for this is that standard predictive modeling techniques cannot directly utilize this type of information.

Recently, there has however been an increasing interest in methods that are able to learn from uncertain data, where various standard machine learning methods have been adapted to deal with uncertain input features, including support-vector machines,[4] decision trees,[5,6] random forests,[7,8] artificial neural networks,[9] Bayesian classifiers,[10] and rule-based approaches.[11−13] In this work, we investigate the effect of introducing uncertainty in chemical descriptors when deriving predictive models. The aim of the study is to investigate how predictive performance is affected by increasing amounts of descriptor uncertainty. In this work, decision tree ensembles have been employed, since this type of method has frequently delivered models with state-of-the-art predictive performance.[14]

MATERIALS AND METHODS

Data Sets. Sixteen publicly available data sets have been used (see Table 1 for a list of names and end points as well as references and the Supporting Information for SMILES and activity classes). The structures have been described using an AstraZeneca in-house set of descriptors.[15] The 196 molecular descriptors are of physicochemical nature. These descriptors contain 1D and 2D as well as 3D descriptors, including properties such as various counts of atoms and bonds, charges, surfaces, and lipophilicity. The models in this work involve binary classification tasks with, in most cases, balanced classes, i.e., the two classes contain approximately the same number of examples.

Uncertainty. Uncertainty was introduced to the data sets to various degrees and in various ways. In contrast to a standard numeric feature value $X_i \in \mathbb{R}$, an uncertain numeric feature value is instead a cumulative distribution function (CDF) F_i: $x \rightarrow F_X(x) = P(X \leq x)$, where the right-hand side represents the probability that

Received: July 23, 2012
Published: October 6, 2012

dx.doi.org/10.1021/ci3003446 | J. Chem. Inf. Model. 2012, 52, 2815−2822

Table 1. Data Set Characteristics

data set	end point	no. compounds	ref
ace	angiotensin converting enzyme	114	22
ache	acetyl-cholinesterase	111	22
ai	steroid aromatase	69	21
amph1	amphiphysin-1 SH3 domain	130	21
ata	antituberculosis target	94	21
bzr	benzodiazepine receptor	163	22
comt	catechol-*O*-methyltransferase	92	21
cox2	cyclooxygenase-2	322	22
dhfr	dihydrofolate reductase	397	22
edc	estrogen receptor	119	21
gpb	glycogen phosphorylase B	66	22
hivpr	HIV protease	113	21
hivrt	HIV-1 reverse transciptase	101	21
hptp	human protein tyrosine phosphatase 1B	132	21
therm	thermolysin	76	22
thr	thrombin	88	22

the random variable X takes on a value less than or equal to x. The CDF may be defined in terms of a probability density function f as follows: $F_X(x) = \int_{-\infty}^{x} f(t)\ \mathrm{d}(t)$. In this study, we consider uncertain numeric feature values defined using the probability density function of the *uniform distribution* $f(t) = 1/(b - a)$ for $t \in [a, b]$ and $f(t) = 0$ otherwise, for some boundaries a and b, and of the *normal distribution* $f(t) = 1/[\sigma(2\pi)^{1/2}]\mathrm{e}^{-[(t-\mu)^2/2\sigma^2]}$, where μ is the mean and σ is the standard deviation. When considering uniform distributions, then for each descriptor to which uncertainty is introduced at a certain level L, each numeric value V for that descriptor is replaced by a uniform distribution with boundaries $[V - LV, V + LV]$. The levels considered in this study are 2.5%, 5%, 10%, and 20%, respectively. Thus, for a descriptor with a numeric value of 100, the introduction of uncertainty using a uniform distribution at the 10% level means that all values between 90 and 110 are equally likely for that descriptor value. When considering normal distributions, we instead consider the boundaries $[V - LV, V + LV]$ to represent a 95% confidence interval, from which the parameters for the normal distribution, μ and σ, can be derived. Since the original value V always resides in the middle of the interval $[V - LV, V + LV]$ and hence can be derived from the probability density functions, we have also considered the following more elaborated way of introducing uncertainty that prevents this derivation: from the interval $[V - LV, V + LV]$, a new midpoint V' is uniformly sampled. The new boundaries become $[V' - LV, V' + LV]$. In the example above, if we first sample the value 95 from the interval [90, 110], then this would result in a new interval [95 − 10, 95 + 10] = [85, 105].

We have performed three sets of experiments:

Set 1: In the first set, we have associated all descriptors with uncertainty.

Set 2: In reality not all descriptors may have uncertainties associated with them. For instance, descriptors related to properties such as counts of atoms and bonds have no uncertainty, since this would make no sense. In the second set of experiments, uncertainty is therefore introduced only for those descriptors for which uncertainty can be expected in reality, corresponding to about 55% of the original descriptors. In the two first sets of experiments, the midpoints of the probability density functions correspond to the original values in the data sets.

Set 3: In the final third set of experiments, the more elaborate way of introducing uncertainty into all descriptors has been chosen, where midpoints are sampled from the first set of intervals.

Learning Algorithms. The learning algorithm considered in this study is decision tree ensembles. The basic strategy that is employed when generating a single decision tree[16] from training examples is called recursive partitioning or divide-and-conquer. It works by partitioning the examples by choosing a set of mutually exclusive conditions on a feature (descriptor), e.g., the feature value is less than a particular threshold, or greater or equal to this threshold, and the choice is usually made such that the error on the dependent variable (or class variable) is minimized within each group. In this study, we consider only binary partitionings using numeric features, where the threshold values are determined from training data. The process continues recursively with each subgroup until certain conditions are met, such as that the error cannot be further reduced, e.g., all examples in a group have the same class label. The resulting decision tree is a graph that contains one node for each subgroup considered, where the node corresponding to the initial set of examples is called the root, and for all nodes there is an edge to each subgroup generated from it, labeled with the chosen condition for that subgroup. An example is classified by the tree by following a path from the root to a leaf node, such that all conditions along the path are fulfilled by the example. The estimated class probabilities at the reached leaf node are used to assign the most probable class to the example.

A decision tree ensemble consists of a set of decision trees, by which predictions are formed by voting. The idea behind this is that the error of the majority in the ensemble is often far less than the error of a single tree, as long as the trees in the ensemble perform better than random and to some extent make their errors independently. The random forest algorithm,[17] which perhaps is the most well-known strategy for generating decision tree ensembles, aims for the latter by incorporating randomness both in the selection of training examples and in the selection of features to consider when partitioning examples during tree construction. The former is done by employing bootstrap aggregating, or bagging, which works by randomly selecting n examples with replacement from the initial set of n training examples. Furthermore, when generating each tree in the forest, only a small randomly selected subset of all available input features are considered at each node in the tree.

Two main approaches to handling uncertain numeric features during model construction are considered.

(1) Sampling: Besides incorporating randomness through randomly selecting subsets of examples or features, this approach also select feature values randomly from the set of possible values, according to the specified probability density functions. Similarly to when sampling training examples, we here consider the sampling of feature values to be performed prior to the generation of each tree, i.e., each uncertain feature value in the training set is replaced by a specific, sampled, value for each tree that is to be generated. This means that from the point of view of the tree learning algorithm, no particular care has to be taken to uncertain numeric feature values, since these are not distinguished from certain ones during tree construction.

(2) Distributing fractions of examples: A common approach to handling missing values when partitioning examples according to the values on some feature during the

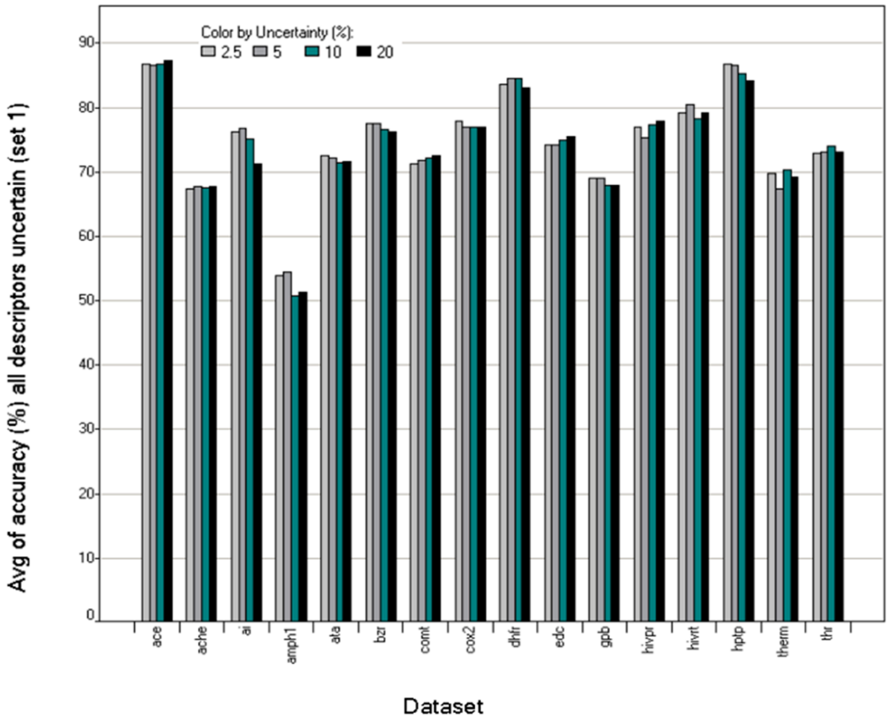


Figure 1. Average accuracy for all ensemble methods models with all descriptors uncertain (experiment set 1).

growth of a classification tree is to first partition all examples with known values on the feature, and then, distribute fractions of the remaining examples according to the estimated probabilities of belonging to the resulting partitions.[16] For example, if a set of examples is to be partitioned into two groups using the numeric feature F together with the threshold 9, all examples fulfilling the condition $F \leq 9$ fall into the first group and all examples fulfilling $F > 9$ will fall into the second. Furthermore, examples for which the value of feature F is missing will be distributed according to the probability of falling into the first and second group respectively, as estimated from those examples for which the feature values are not missing. To this end, each example is associated with a weight that is initially set to 1 and which is reduced whenever an example is distributed over multiple nodes. For example, if 80% of the examples fall into the first group, then each example e with weight w that has a missing value on F will result in that e with weight $0.8w$ appears in the first partition and e with weight $0.2w$ in the second. In a similar fashion, examples with uncertain feature values may be distributed over multiple nodes by using their associated distributions to calculate the probability of the random variable fulfilling the condition for each partition.

Given an example e with weight w having an uncertain feature value $x \rightarrow F_X(x)$ on feature F, and two groups formed by the conditions $F \leq t$ and $F > t$, respectively, for some threshold t, this would mean that e with weight $wF_X(t)$ would fall into the first group and e with weight $w(1 - F_X(t))$ would fall into the second. There is however a complicating issue: in the case the same feature has been used to partition examples in some ancestor node, i.e., existing conditions imply lower and upper bounds of the feature, the resulting weights have to be modified accordingly. Assume such a lower bound (l) and upper bound (u) for the feature F are given (possibly being $-\infty$ and ∞, respectively). Given an example e with weight w having an uncertain feature value $x \rightarrow F_X(x)$ on feature F and two groups formed by the conditions $F \leq t$ and $F > t$, respectively, for some threshold t, then e with weight $w[(F_X(t) - F_X(l))/(F_X(u) - F_X(l))]$ falls into the first group and e with weight $1 - w[(F_X(t) - F_X(l))/(F_X(u) - F_X(l))]$ falls into the second. It should be noted that in contrast to one of the previous approaches,[4] the method for distributing fractions of examples suggested here does not rely on sampling. Uncertain numeric feature values need not only be taken special care of when partitioning examples according to conditions on the corresponding feature, but also when finding a threshold for the feature to be used in the conditions. One common strategy for finding such thresholds for (certain) numeric feature values is to enumerate and evaluate all thresholds that separate examples belonging to different classes.[18] However, since uncertain numeric feature values are assumed to be represented by probability distributions, rather than specific values, we need to select some values from these distributions, in order to allow for the former discretization method to be employed. One candidate method is to select the expected value from each distribution, while another is to sample values from each distribution. The latter approach has been adopted in this study, where one value is sampled from each distribution. This contrasts to the previously proposed method[5] that samples sets of values, which leads to an increased computational cost due to that the number of threshold values to consider increases with a factor equal to the sample size.

The above strategies were implemented and tested within the rule discovery system (RDS).[19] For each strategy, an ensemble of 25 trees was generated, where each individual tree was grown until all leaf nodes contained examples of at most one class,

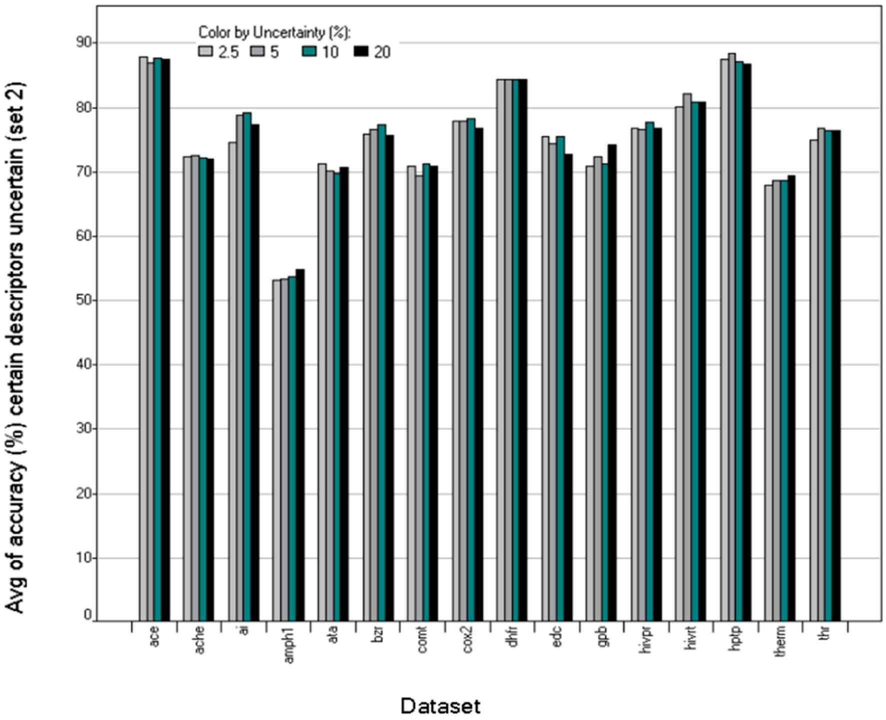


Figure 2. Average accuracy for all ensemble methods models with selected descriptors uncertain (experiment set 2).

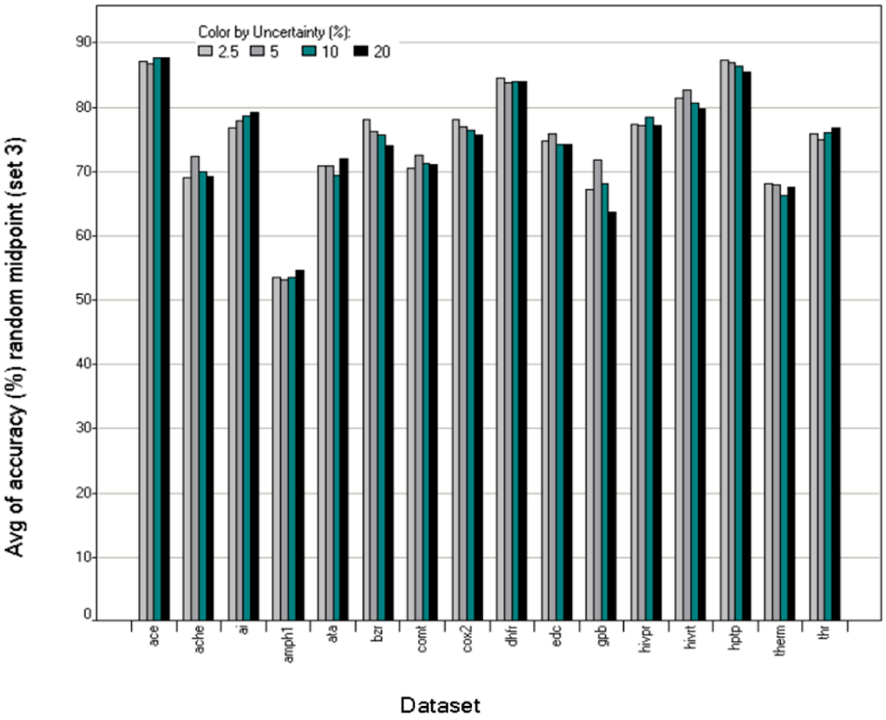


Figure 3. Average accuracy for all ensemble methods models with all descriptors uncertain and random midpoint (experiment set 3).

while ignoring classes for which the total frequency was less than one, or none of its examples had a weight over 0.5. No depth restrictions of the trees were employed.

Evaluation. Five approaches to handling information on uncertainty in decision tree ensembles were compared: (a) considering the expected value only (Ensemble-ev), i.e., effectively ignoring information on uncertainty; (b) sampling from uniform distributions (Ensemble-random-uniform), i.e., using the bounds of the confidence intervals; (c) sampling from normal distributions (Ensemble-random-normal), i.e., deriving means and standard deviations from the confidence intervals; distributing fractions of examples assuming (d) uniform distributions (Ensemble-range-uniform) and (e) normal distributions (Ensemble-range-normal), respectively. The predictive performance of the developed models was assessed by the average accuracy resulting from 10-fold cross-validation. In order to reduce variance, averages from repeating each 10-fold cross-validation 100 times with unique random seeds were calculated

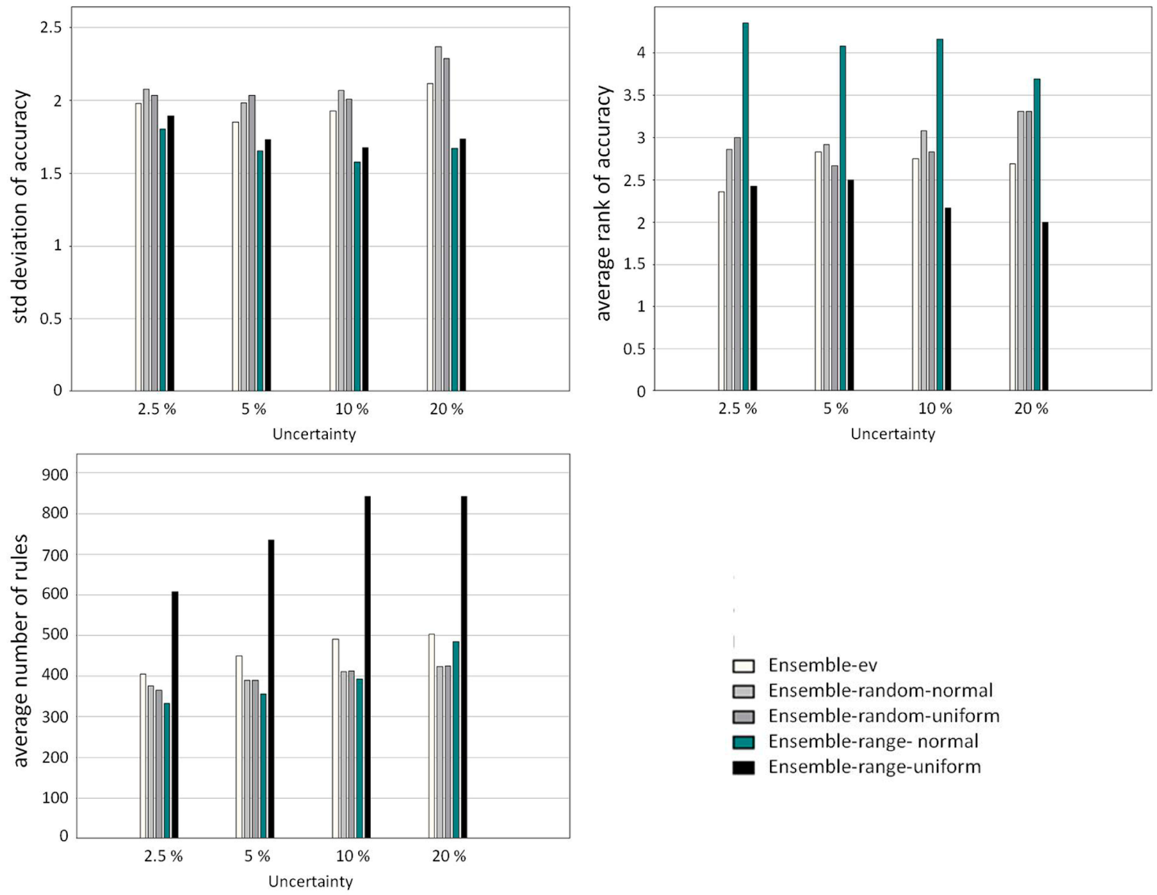


Figure 4. Average standard deviation (a), rank (b), and number of rules in predictive performance (accuracy) (c) from the repeated experiments with randomly assigned folds with respect to various levels of uncertainty for all ensemble methods and all data sets.

and reported for all of the data sets. To further validate the approach (as suggested by an anonymous reviewer), we set aside a randomly chosen external test sets consisting of 40% and 60%, respectively, of the largest data set (dhfr). The remaining portions were used for training. We repeated the two test set selections (40 and 60%) 100 times each.

■ RESULTS

The results in terms of overall accuracies for each data set and set of experiments (1, 2, 3) in the study when introducing various levels of uncertainty are reported in Table 2 (Supporting Information) and depicted in Figures 1−3. The average standard deviations in predictive performance (accuracy) and average predictive performance ranks, as well as the average number of rules from the 100 repeated experiments with randomly assigned folds among the ensembles at the various levels of uncertainty (set 1 experimental conditions) are depicted in Figures 4a−c. The overall average standard deviations in predictive performance (accuracy), average predictive performance ranks as well as the average number of rules from the 100 repeated experiments with randomly assigned folds for all uncertainty levels (set 1 experimental conditions) are depicted in Figures 5a−c. For further clarification, the Ensemble-ev method used the expected values only, i.e., without uncertainty, but with variations to the results caused by the random choices made when generating the ensembles, illustrating the levels of variation that normally occur for this type of approach. Furthermore, with respect to uncertain variables, the average percentage of such descriptors among the 20 most important descriptors in the derived models for experiment set 2 is approximately 50%, where the relative importance of each feature was calculated by dividing the error reduction obtained by choosing the feature to partition the examples when growing the trees by the total amount of error reduction. The results from 100 repeated experiments with external test sets at 40 and 60%, respectively, of the dhfr data set are presented in Table 3 (Supporting Information).

■ DISCUSSION

In this study, the outcome of in silico modeling using decision tree ensembles on 16 different biopharmaceutical data sets has been investigated, when introducing various levels of uncertainty to the chemical descriptors.

The overall results indicate that the derived models are remarkably stable (robust) to the introduction of various levels of uncertainty. For instance, the average standard deviation in accuracy for the 100 repeated 10-fold cross-validations for all ensemble methods and levels of uncertainty over all data sets is only 1.927. Also, the robustness to increasing amounts of noise with respect to accuracy is remarkable, as depicted in Figures 1−3, for the data sets, where only slight changes in accuracy can be noticed with increasing levels of uncertainty. Furthermore, the 100 repeated runs for the two external test sets of dhfr also reinforces this robustness where the average accuracies and standard deviations compare favorably with the 10-fold cross-validation results on the same data set (Table 3, Supporting Information). The average accuracy over all methods and uncertainty levels increases from 82.3% via 83.4−85.4% when

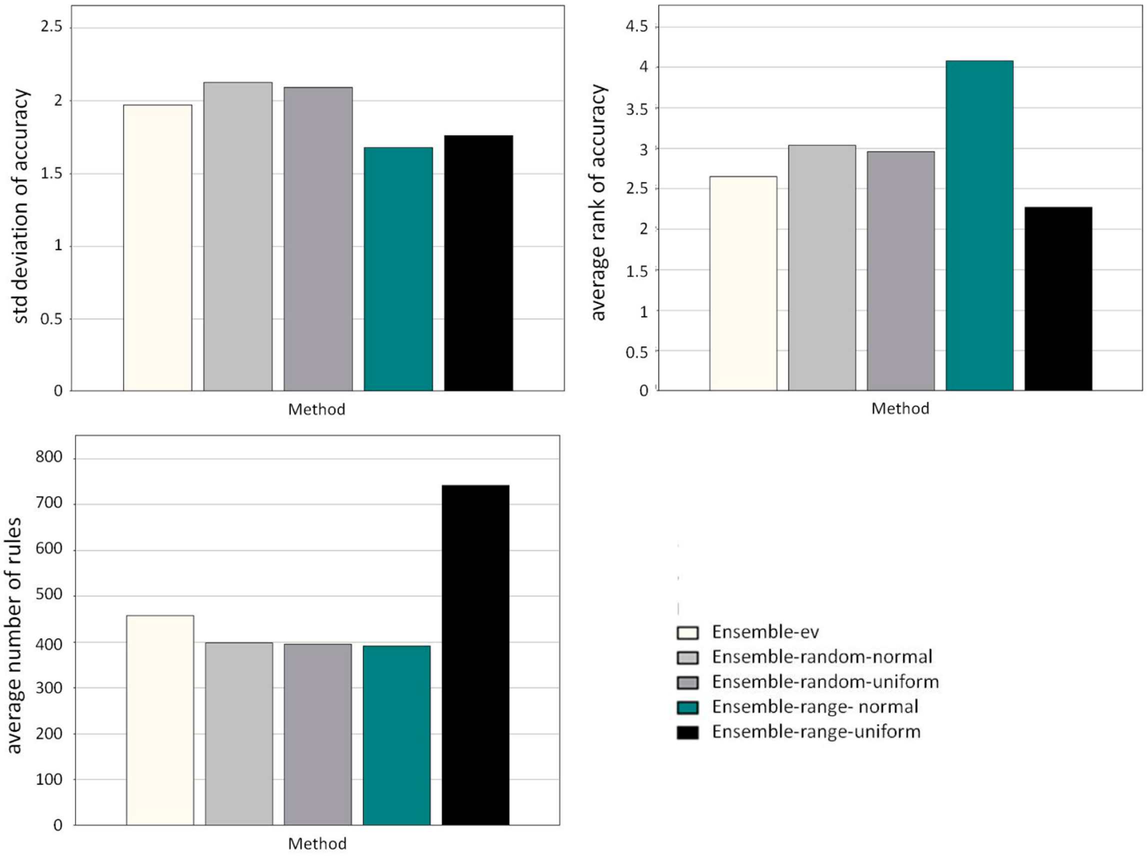


Figure 5. Average standard deviation in predictive performance (accuracy) (a), rank (b), and number of rules from the repeated experiments (c) with randomly assigned folds for all levels of uncertainty and all data sets for the five ensemble approaches.

using the external test sets of 60 and 40% and 10-fold cross-validation, respectively.

However, since not all descriptors have uncertainty associated with them, this naturally raises the question: Perhaps a majority of the most important variables in the resulting models are those without uncertainty? This would explain the observed stability and insensitivity to the increasing levels of uncertainty. However, an inspection of the 20 most important variables for the derived models of each data set reveals that this is not the case. On average ~50% of the 20 most important variables are associated with uncertainty among the 16 data sets. Thus, the robustness (stability) to uncertainty of the derived models in this study does not, to any large extent, originate from the models being dominated by variables without uncertainty. For instance, high quality models for the data sets ace, cox2, dhfr, and hivpr display a whole percentage spectrum of important variables with and without uncertainty. At one end, the models derived for the ace data set contain rather few uncertain descriptors (<40%, average 19%) while at the other end, models derived for the cox2 data set in all but one case contain a large number of uncertain descriptors (>55%, average 77%). Why then are the derived models so stable to increasing levels of uncertainty? One possible explanation stems from the important ensemble concept of variability (diversity) between the individual models (base classifiers) comprising the ensemble. It is well-known that in order for an ensemble to be accurate, the individual models must have sufficient diversity. By introducing uncertainty into the descriptors, the variance among the base classifiers is increased and, consequently, so is the diversity of the ensemble. At the same time, however, the introduction of uncertainty may also lead to a decrease in performance of the individual models. These two factors work in opposite directions; an increased diversity generally improves predictive performance of the ensemble, while a decreased base classifier accuracy generally degrades the performance.[20] For the results of the present study, we argue that it seems as if the former factor is not dominated by the latter and hence the performance of the ensemble models are relatively insensitive to increasing amounts of uncertainty. Furthermore, the differences in ensemble accuracies between models with partially uncertain descriptors and the corresponding ones where all variables are associated with uncertainty for each of the data sets are rather small. In fact the average absolute difference between the former and the latter ones across all data sets is only 2.6%.

Introducing randomly selected midpoints (experiment set 3) also shows that the derived models are remarkably stable to uncertainty and variation. There is a slight tendency by ensembles generated at the largest level of uncertainty (20%) for one data set, the smallest of all considered, gpb containing 66 compounds, to result in lower accuracy compared to the corresponding models for experiment set 1. However, the overall correlation between accuracies of models obtained from experiment sets 1 and 3 are very high with values of 0.944, 0.923, 0.914, 0.966, and 0.936 for the methods Ensemble- ev, Ensemble-random-normal, Ensemble-random-uniform, Ensemble-range-normal, and Ensemble-range-uniform, respectively. Also, the average absolute difference in accuracies between models from set 1 and set 3 is only 2.2%.

What other observations can be made from the results of this work? One clear indication is that generating models based on uniform distributions taking the descriptor range into account when deriving models seems to be the best approach in parity with models generated with no uncertainty (Figure 5b). Another observation is that the more advanced approaches that distribute fractions of examples, the range methods, to handle uncertain values do not seem offer any significant benefit over using sampled values, random methods, from the distribution, as demonstrated by experiment set 3 (Figure 5b).

An additional observation that can be made is that the number of rules generated for the different types of ensemble models varies considerably. The models derived using the ensemble range-uniform method has approximately twice as many rules as the models derived by the other four ensemble approaches (Figures 4c and 5c). Also, not surprisingly, the number of rules slightly increased as the level of uncertainty increased from 2.5% to 20% as depicted in Figure 4c. However, for most of the methods, the increase was not dramatic, but around 10% with the exceptions of the ensemble models derived without uncertainty and, again, the models derived using a uniform range approach where the increase was approximately 25 and 40%, respectively.

In a previous study,[7] we observed that utilizing information on uncertainty, when known, may outperform approaches ignoring uncertainty, i.e., Ensemble-ev in this work. The main difference between the earlier study and the current one is the source of the uncertainty. In the previous work, a single uncertain attribute of high importance was considered, i.e., log P, for which 95% confidence intervals were obtained from the software calculating the attribute value.[7] These intervals could hence be expected to somehow reflect the true uncertainty of the attribute values. This contrasts to the current study, where we instead consider artificially introduced uncertainty that does not necessarily relate to the actual uncertainty of the different attribute values. The former study thus investigated the case when additional information was brought into the representation of the compounds, while we in the current study provide no such additional information, but rather reduce the information by replacing point values by (arbitrary) intervals. The results indicate that although these intervals can be made very broad without hurting performance, there is not much to gain from introducing this artificial type of uncertainty, in contrast to when using nonartificial uncertainty as done in the previous study.

CONCLUSION

The main conclusion of the empirical investigation is that uncertainty may to a large extent be introduced in chemical descriptors without hurting predictive performance of decision tree ensembles. The investigation further shows that, even when distributions of uncertain values are provided, the ensemble method can generate equally effective models from single-point samples from these distributions. Hence, there seems to be no advantage in this case to use more elaborate methods for handling uncertainty compared to methods ignoring uncertainty.

ASSOCIATED CONTENT

Ⓢ Supporting Information

SMILES and activity classes for the 16 data sets presented in this work and Tables 2 and 3. This material is available free of charge via the Internet at http://pubs.acs.org.

AUTHOR INFORMATION

Corresponding Author

*E-mail: ulf.norinder@astrazeneca.com.

Present Address

⊥Department of Computational and Analytical Chemistry, H. Lundbeck A/S, 9 Ottiliavej, 2500 Valby, Denmark. E-mail: ulfn@lundbeck.com.

Notes

The authors declare no competing financial interest.

ACKNOWLEDGMENTS

This work was supported in part by the INFUSIS project (www.his.se/infusis) at the University of Skövde, Sweden, in partnership with AstraZeneca, Lexware Laboratories, and the Swedish Knowledge Foundation under grant 2008/0502 and in part by the Swedish Foundation for Strategic Research through the project *High-Performance Data Mining for Drug Effect Detection* (ref no. IIS11-0053) at Stockholm University, Sweden.

REFERENCES

(1) van de Waterbeemd, H.; Gifford, E. Admet in silico modelling: towards prediction paradise? *Nat. Rev. Drug. Discovery* **2003**, *2*, 192–204.

(2) Howe, T. J.; Mahieu, G.; Marichal, P.; Tabruyn, T.; Vugts, P. Data reduction and representation in drug discovery. *Drug Discovery Today* **2007**, *12*, 45–53.

(3) Johansson, U.; Sönströd, C.; Norinder, U.; Boström, H. The trade-off between accuracy and comprehensibility for predictive in silico modeling. *Fut. Med. Chem.* **2011**, *3*, 647–663.

(4) Bi, J.; Zhang, T. Support vector classification with input data uncertainty. In *Advances in Neural Information Processing Systems (NIPS'04)*, Vancouver, Canada, December 13–18, 2004; Saul, L. K., Weiss, Y., Bottou, L., Eds.; MIT Press: Cambridge, 2005; pp 161–168.

(5) Tsang, S.; Kao,B.; Yip, K. Y.; Ho, W.-S.; Lee, S. D. Decision trees for uncertain data. In *Proceedings of the 2009 IEEE International Conference on Data Engineering*, Shanghai, China, March 29 2009–April 2 2009; Golab, L., Johnson, T., Shkapenyuk, V., Eds.; IEEE Computer Society: WA, 2009; pp 441–444.

(6) Qin, B.; Xia, Y.; Li, F. Dtu: A decision tree for uncertain data. In *Proceedings of the 13th Pacific-Asia Conference on Advances in Knowledge Discovery and Data Mining (PAKDD)*, Bangkok, Thailand, April 27–30, 2009; Theeramunkong, T., Kijsirikul, B., Cercone, N., Ho, T. B., Eds.; Springer-Verlag: Heidelberg, 2009; pp 4–15.

(7) Boström, H.; Norinder, U. Utilizing information on uncertainty for in silico modeling using random forests. In *Proceedings of the 3rd Skövde Workshop on Information Fusion Topics (SWIFT 2009)*, Skövde, Sweden, October 12–13, 2009; Johansson, R., van Laere, J., Mellin, J., Eds.; University of Skövde: Skövde, 2009; pp 59–62.

(8) Dudas, C.; Boström, H. Using uncertain chemical and thermal data to predict product quality in a casting process. In *Proceedings of the First ACM SIGKDD Workshop on Knowledge Discovery from Uncertain Data*, Paris, France, June 28, 2009; Pei, J., Getoor, L., de Keijzer, A., Eds.; ACM New York: New York, 2009; pp 57–61.

(9) Ge, J.; Xia, Y.; Tu, Y. A discretization algorithm for uncertain data. In *Proceedings of the 21st international conference on Database and Expert Systems Applications (DEXA): Part II*, Bilbao, Spain, August 30–September 3, 2010; Bringas, P. G., Hameurlain, A., Quirchmayr, G., Eds.; Springer-Verlag: Heidelberg, 2010; pp 485–499.

(10) Qin, B.; Xia, Y.; Li, F. A Bayesian classifier for uncertain data. In *Proceedings of the 2010 ACM Symposium on Applied Computing*, Sierre, Switzerland, March 22–26, 2010; Shin, S. Y.; Ossowski, S.; Schumacher, M., Eds.; ACM New York: New York, 2010; pp 1010–1014.

(11) Qin, B.; Xia, Y.; Prabhakar, S.; Tu, Y. A rule-based classification algorithm for uncertain data. In *Proceedings of the 2009 IEEE International Conference on Data Engineering*, Shanghai, China, March

29 2009–April 2, 2009; Golab, L., Johnson, T., Shkapenyuk, V., Eds.; IEEE Computer Society: WA, 2009; pp 1633–1640.

(12) Gao, C.; Wang, J. Direct mining of discriminative patterns for classifying uncertain data. In *Proceedings of the 16th ACM SIGKDD international conference on Knowledge discovery and data mining (KDD)*, Washington, DC, USA, July 25–28, 2010; Rao, B., Krishnapuram, B., Tomkins, A., Yang, Q., Eds.; ACM New York: New York, 2010; pp 861–870.

(13) Qin, X.; Zhang, Y.; Li, X.; Wang, Y. Associative classifier for uncertain data. In *Proceedings of the 11th international conference on Web-age information management (WAIM)*, Jiuzhaigou, China, July 15–17, 2010; Chen, L., Tang, C., Yang, J., Gao, Y., Eds.; Springer-Verlag: Heidelberg, 2010; pp 692–703.

(14) Caruana, R.; Niculescu-Mizil, A. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd International Conference on Machine Learning*, Pittsburgh, Pennsylvania, June 25–29, 2006; Cohen, W., Moore, A., Eds.; ACM: New York: New York, 2006; pp 161–168.

(15) Bruneau, P. Search for predictive generic model of aqueous solubility using bayesian neural nets. *J. Chem. Inf. Comput. Sci.* **2001**, *41*, 1605–1616.

(16) Quinlan, J. R. *C4.5: Programs for Machine Learning*; Morgan Kauffman: San Francisco, 1993.

(17) Breiman, L. Random Forests. *Machine Learning* **2001**, *45*, 5–32.

(18) Fayyad, U.; Irani, K. On the Handling of Continuous-Valued Attributes in Decision Tree Generation. *Machine Learning* **1992**, *8*, 87–102.

(19) *Rule Discovery System (RDS)*, version 2.6.1.0 Modeling ed.; Compumine: Stockholm, Sweden, 2009.

(20) Krogh, A.; Vedelsby, J. Neural network ensembles, cross validation, and active learning. In *Advances in Neural Information Processing Systems*; Tesauro, G., Touretzky, D. S., Leen, T. K., Eds.; Morgan Kaufman MIT Press: Cambridge, MA, 1995; Vol. 7, pp 231–238.

(21) Mittal, R. R.; McKinnon, R. A.; Sorich, M. J. Comparison Data Sets for Benchmarking QSAR Methodologies in Lead Optimization. *J. Chem. Inf. Model.* **2009**, *49*, 1810–1820.

(22) Bruce, C. L.; Melville, J.l.; Pickett, S. D.; Hirst, J. D. Contemporary QSAR Classifiers Compared. *J. Chem. Inf. Model.* **2007**, *47*, 219–227.