Application of conformal prediction in QSAR

Martin Eklund[1,2,⋆], Ulf Norinder[3], Scott Boyer[2], and Lars Carlsson[2]

[1] Department of Pharmaceutical Biosciences, Uppsala University, P.O. Box 591, SE-751 24 Uppsala, Sweden
[2] AstraZeneca Research and Development, SE-431 83 Mölndal, Sweden
[3] AstraZeneca Research and Development, SE-151 85 Södertälje, Sweden

Abstract. QSAR modeling is a method for predicting properties, e.g. the solubility or toxicity, of chemical compounds using statistical learning techniques. QSAR is in widespread use within the pharmaceutical industry to prioritize compounds for experimental testing or to alert for potential toxicity. However, predictions from a QSAR model are difficult to assess if their prediction intervals are unknown. In this paper we introduce conformal prediction into the QSAR field to address this issue. We apply support vector machine regression in combination with two nonconformity measures to five datasets of different sizes to demonstrate the usefulness of conformal prediction in QSAR modeling. One of the nonconformity measures provides prediction intervals with almost the same width as the size of the QSAR models' prediction errors, showing that the prediction intervals obtained by conformal prediction are efficient and useful.

1 Introduction

Elucidating the structural properties of chemical compounds required to elicit a desired pharmacological effect (cure or alleviate a disease) and to concomitantly avoid toxicity is a fundamental issue in drug development. Quantitative Structure-Activity Relationships (QSAR) is a framework for employing statistical learning methods to predict the pharmacological effect and the toxicity (as well as other desirable properties) from chemical compounds' structures.

For making informed decisions based on predictions from a QSAR model, we are interested in the confidence in the predictions. Substantial efforts have been devoted to research on this topic within the QSAR community over the last decade and a number of methods have been suggested for estimating the confidence of QSAR predictions (see e.g. [1–3] and references therein). These confidence estimates are typically based on the very loosely defined concept of a QSAR model's "applicability domain" (AD), which in [4] was described as *"the response and chemical structure space in which the model makes predictions with a given reliability"*. The assumption is that the further away a molecule is from a QSAR model's AD (according to some measure), the less reliable the prediction. The problem with the current approaches to estimating the prediction confidence

⋆ martin.eklund@farmbio.uu.se

in QSAR models is that their interpretation is cumbersome. For example, what does it mean that a chemical compound is outside the QSAR model's AD by a certain amount according to a certain measure? How does the plethora of different metrics used to define the AD relate to each other? What we ideally would like to know is in fact that a prediction is within a given prediction interval with a certain confidence (e.g. 80% or 95%).

In this paper we introduce conformal prediction [5, 6] into the QSAR field. Conformal prediction uses previous data to determine prediction regions for new predictions. Given a confidence level (of say 80% or 95%), it produces a valid prediction region (i.e. a region that contains the true value with a probability equal to or higher than the given confidence level) under the assumptions that the observed data is *i.i.d.* [5]. The conformal prediction framework provides a unified view of the different approaches to estimating a QSAR model's AD. Moreover, conformal prediction gives a natural and intuitive way of interpreting the AD estimates as prediction intervals with a given confidence.

The paper is organized as follows: We give a short introduction to QSAR modeling and AD estimation in Section 2. Continuing with Section 3 that contains a description of the methods (learner, nonconformity measure) used in the paper and how we apply them to QSAR modeling. The results are presented in Section 4 and a discussion of the results in Section 5.

2 QSAR

A molecule can be represented by an undirected labeled graph $G = (V, E, A)$ (possibly embedded in $\mathbb{R}^3$ if the three dimensional structure of the molecule is taken into account), where the vertices V are the atoms and the edges E are the bonds. The atoms of a molecular graph are labeled by A, a set of *atom types*, which for instance can be the set of elements of the periodic table, the set of physicochemical properties of atoms, or any set of atom types provided by a molecular force field (see e.g. [7]).

A *descriptor function* may be defined as a random variable

$$d : G \to \mathbb{R}^p. \tag{1}$$

A realization x_i of d applied to a molecular structure G_i is called the *descriptor vector* (or simply *descriptor*) of G_i (NB: we will in this paper use "descriptor" and "attribute" interchangeably to refer to an element in x_i). Given a training set $(x_1, y_1), ..., (x_n, y_n)$, where each y_i, $i = 1, ..., n$, is the known activity of the molecule with descriptor vector x_i, we can apply statistical learning methods to estimate a QSAR. A QSAR model fitted to $(x_1, y_1), ..., (x_n, y_n)$ permits us to predict the activity of new (possibly not yet synthesized) chemical compounds.

2.1 Applicability domain

It is difficult for researchers to use the predictions from a QSAR model if there is no information available on whether the predictions are reliable or not. A large

number of approaches have been suggested in the QSAR literature to assess whether a give prediction is reliable, typically by defining a QSAR's applicability domain. For example:

- Let $\hat{\mu}$ and $\hat{S}$ be the respective estimates of the mean vector and the covariance matrix of the distribution P from which $x_1, ..., x_n$ were sampled. A prediction of a new compound with descriptor vector x_{new} is regarded less reliable the larger the distance $D(x_{new}) = \sqrt{(x_{new} - \hat{\mu})^T \hat{S}^{-1} (x_{new} - \hat{\mu})}$ and may be defined as "unreliable" if $D(x_{new})$ exceeds some predefined limit c. The hypersphere $D(x) \leq c$ is thus the AD.
- Estimate the density of P. This is often done by assuming independence between descriptors, or by projecting $(x_1, y_1), ..., (x_n, y_n)$ to a lower dimensional space using e.g. principal component analysis (PCA). The density value at x_{new} is determined and the lower this value is, the less reliable the prediction is. If it is lower than a predefined limit, the prediction may be deemed "unreliable".
- A large number of methods are based on resampling (e.g. bootstrapping) from $(x_1, y_1), ..., (x_n, y_n)$ to estimate the variability of the prediction of a new compound x_{new} (see e.g. [8] for a survey of eight different resampling based methods). The AD is then implicitly defined as the region in descriptor space where the variability of predictions is lower than a given cutoff value.

A reader familiar with conformal prediction will realize that all these measures can be viewed as different nonconformity measures.

3 Methods

3.1 The signature descriptor

Let w be an atom of G, the signature of height h of w, σ_w^h, is a canonical representation of the subgraph of G containing all atoms that are at distance h bonds from w. It is usually encoded as a string, see Figure 1 and Table 1. The signatures can be viewed as letters of a finite alphabet, $\sigma_w^h \in \Sigma$ and the occurrence of each letter is used as a descriptor value (i.e. each letter in Σ is represented by a given position in a descriptor vector and the value on that position is the number of occurrences of the letter in a molecule). In this study Σ was defined by the set of signatures defined at training time of each particular QSAR model. Typically the number of letters in Σ is much greater than the number of examples in a training set and the maximum number of letters present for a molecule will be the number of heights times the number of atoms. Detailed explanations on the algorithm of the signature descriptor and the usage of the signature descriptor can be found in [9, 10].

The motivation for using the signature descriptor is that it has performed very well in benchmarking experiments on in-house AstraZeneca datasets. The limitation with the signature descriptor is that it only is a two-dimensional description of a molecule (as opposed to a potentially more information rich three

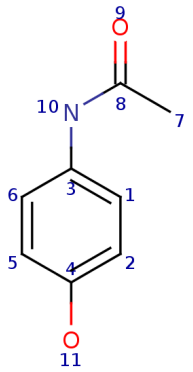

Fig. 1. Paracetamol is a drug molecule for treatment of pain and fever. The atom numbers are displayed in this figure.

w	$h=0$	$h=1$	$h=2$
1	[C]	[C]([C]=[C])	[C]([C](=[C])=[C]([C][N]))
2	[C]	[C]([C]=[C])	[C]([C](=[C])=[C]([C][O]))
3	[C]	[C]([C]=[C][N])	[C]([C](=[C])=[C]([C])[N]([C]))
4	[C]	[C]([C]=[C][O])	[C]([C](=[C])=[C]([C])[O])
5	[C]	[C]([C]=[C])	[C]([C](=[C][O])=[C]([C]))
6	[C]	[C]([C]=[C])	[C]([C](=[C][N])=[C]([C]))
7	[C]	[C]([C])	[C]([C]([N]=[O]))
8	[C]	[C]([C][N]=[O])	[C]([C][N]([C])=[O])
9	[O]	[O](=[C])	[O](=[C]([C][N]))
10	[N]	[N]([C][C])	[N]([C]([C]=[O])[C]([C]=[C]))
11	[O]	[O]([C])	[O]([C]([C]=[C]))

Table 1. The signatures of three different heights for paracetamol. w denotes the atom number and h the height.

dimensional description). However, valid three-dimensional representations are both costly to compute and there is also a great deal of uncertainty regarding what three-dimensional conformations a molecule would have for a particular problem, which introduces errors within a machine-learning based model.

3.2 Learning method

We used the support vector regression machine (SVM) [11] implemented in the Java library LIBSVM [12] with a Gaussian radial basis kernel function

$$K(x, x') = \exp\left(-\gamma \|x - x'\|^2\right). \tag{2}$$

γ and the cost parameter C were either chosen in a cross-validated grid search or preset such that $\gamma = 2^{-9}$ and $C = 100$. During the grid search γ could take on vales 2^{-s} with $s = 3, \ldots, 10$ and C values $10^{t/2}$ with $t = 0, \ldots, 5$.

3.3 Nonconformity measures

We will in this paper use inductive conformal prediction (ICP) due to its substantially smaller computational cost compared with transductive conformal prediction (TCP). We will therefore use nonconformity measures that are suitable for the inductive framework. These measures will essentially follow the nonconformity measures (31) and (32) defined in [13]. We will however make some changes compared to [13], which we point out below.

Let T be the training set $(x_1, y_1), ..., (x_n, y_n)$ and let $(x_{i_1}, y_{i_1}), \ldots, (x_{i_k}, y_{i_k})$ denote the k nearest neighbors of x_i in T. For an example x_i, we estimate the standard deviation of the labels of its k nearest neighbors as

$$s_i^k = \sqrt{\frac{1}{k}\sum_{j=1}^{k}(y_{i_j} - \bar{y}_{i_{1,\ldots,k}})^2}, \tag{3}$$

where

$$\bar{y}_{i_{1,\ldots,k}} = \frac{1}{k}\sum_{j=1}^{k} y_{i_j}. \tag{4}$$

The standard deviation of the labels is normalized with the median standard deviation of the k nearest neighbor labels of all training examples

$$\xi_i^k = \frac{s_i^k}{\text{median}(\{s_j^k : (x_j, y_j) \in T\})}. \tag{5}$$

Let

$$d_i^k = \sum_{j=1}^{k}\big(K(x_i, x_i) + K(x_{i_j}, x_{i_j}) - 2K(x_i, x_{i_j})\big) = 2\sum_{j=1}^{k}\big(1 - K(x_i, x_{i_j})\big) \tag{6}$$

be the sum of the distances in feature space between x_i and its k nearest neighbors, where $K(\cdot, \cdot)$ is the kernel function in Equation (2). This distance is different to the one used in [13] (where the Euclidean distance was used), however is natural to use in this work considering our choice of learning method. Similarly to (5), we divide (6) with the median of the distances of all training examples from their k nearest neighbors:

$$\lambda_i^k = \frac{d_i^k}{\text{median}(\{d_j^k : (x_j, y_j) \in T\})}. \tag{7}$$

We now introduce the two nonconformity measures that we will use:

$$\alpha_i = \left|\frac{y_i - \hat{y}_i}{\exp(\gamma\lambda_i^k) + \exp(\rho\xi_i^k)}\right| \tag{8}$$

and

$$\alpha_i = \left| \frac{y_i - \hat{y}_i}{\kappa + \eta\lambda_i^k + (1-\eta)\xi_i^k} \right|, \tag{9}$$

where $\hat{y}$ is the SVM prediction. The measure (8) is identical to measure (32) in [13] (apart from the difference in distance function). Measure (9) is similar to measure (31) in [13], however we have introduced the parameter η to control the relative influence of λ_i^k and ξ_i^k. In this work, we used $\kappa = \eta = 0.5$ in all tests. The rationale behind the denominator in measures (8) and (9) is done to normalize with the expected accuracy of the underlying method (see [13] for a more detailed discussion). We ask for 95% prediction intervals in all tests.

3.4 Data description

We used the signature descriptor of height zero to three to compute descriptor vectors for the chemical compounds in the datasets described below (two of the datasets are publicly available, whereas three datasets are proprietary in-house AstraZeneca datasets). We then applied SVM and ICP as outlined in Sections 3.2 and 3.3 to derive QSAR models with prediction intervals. All datasets were split into training-, calibration-, and test sets. The training sets were used for model fitting and parameter optimization, the calibration set for computing the prediction region, and the test set for evaluating the model fits and the prediction regions. The calibration examples were chosen by random sampling (repeated between 1 and 10 times for different datasets). The testset was chosen by selecting the last (most recent) q examples from the dataset. The rationale behind this method for selecting the testset is that when designing new compounds there is usually distinct changes to the composition of the chemical compounds, and it can thus be argued whether new examples are *i.i.d.*

Datasets The publicly available datasets contain the aqueous solubilities for a diverse set of 1297 organic compounds [14] and the US Food and Drug Administration (FDA) recommended daily doses for 1215 pharmaceuticals. These datasets are available at `http://cheminformatics.org/` and `http://www.epa.gov/ncct/dsstox/sdf_fdamdd.html`, respectively. We are not permitted to disclose the biopharmaceutical endpoints or the structures of the molecules in the AstraZeneca in-house datasets due to legal reasons, however summary statistics of these datasets are shown in Table 2 along with the same statistics for the public datasets.

4 Results

The main results are shown in Table 3 and Figure 2. Table 3 shows comparisons of the predictive region tightness for the different nonconformity measures outlined in Section 3.3 applied to the datasets described in Section 3.4, respectively. Figure 2 shows violin plots of the labels of the different datasets, the absolute

Dataset	Attributes	Training	Calibration	Test	Preset	Random
Solu.	6862	926	99	257	No	Yes
FDA	11295	901	99	215	No	Yes
AZ1	18313	3331	99	391	Yes	No
AZ2	52470	18320	499	871	Yes	No
AZ3	82432	38763	499	2862	Yes	No

Table 2. Description of the datasets. The table shows the number of attributes for each dataset, as well as the number of examples (molecules) used for training, calibration, and testing. Whether γ and cost parameter C were preset or estimated using cross-validation is also indicated in the "Preset" column, whereas the column "Random" indicates whether the calibration examples were chosen at random or if the q most recent ones were used.

value of the prediction errors and the widths of nonconformity measure (9) applied to them (a violin plot is a combination of a box plot and a rotated kernel density plot on each side of the box plot, see [15, 16]).

5 Discussion

Conformal prediction formalizes the AD-based approaches for estimating the reliability of QSAR predictions. Different AD measures can simply be regarded as special cases of nonconformity measures. In this paper we used the nonconformity measures (8) and (9) with the 3 and 10 nearest neighbors, respectively. Measure (9) using the 10 nearest neighbors appears to produce the best results (giving the overall tightest predictive regions) on the tested datasets. The reason for this is that the variance of the labels among the near neighbors can be very high for a subset of the molecules in the datasets (in fact, this is a well-known phenomenon in QSAR modeling and has been termed "the QSAR paradox" or the "the Kubinyi paradox" in the literature [17]). This leads to unnecessarily large prediction regions for measure (8) due to the exponentiation of the estimated variance among the near neighbors. Using 10 near neighbors instead of 3 alleviates this problem, since the effect on the variance of a problematic molecule is "diluted" to a larger extent. Further, using measure (9) avoids this problem and - for the large datasets where reasonably good predictions could be achieved (the AZ2 and AZ3 datasets) - generates informative prediction regions.

The improvement of measure (9) over measure (8) suggest that it may be interesting to optimize the parameters in nonconformity measures during the training of QSAR models (analogously to how e.g. γ and C in the SVM are chosen using cross-validation). It may also be interesting to try different AD measures suggested in the QSAR literature and evaluate their efficiency (tightness) compared to measures (8) and (9).

Dataset	Range	R_t^2	Measure	Number of runs	Median width	Mean width	Max width	% outside
Solu.	0 to 10.3	0.90	(8)/3NN	10	14	9.8×10^{15}	7.5×10^{18}	1.8
	0 to 10.3	0.90	(8)/10NN	10	140	3.8×10^{12}	2.2×10^{15}	2.1
	0 to 10.3	0.90	(9)/3NN	10	4.4	7.2	54	1.8
	0 to 10.3	0.90	(9)/10NN	10	2.0	2.4	12	3.9
FDA	0 to 6.7	0.45	(8)/3NN	10	2.8	1.6×10^{12}	8.5×10^{14}	1.1
	0 to 6.7	0.46	(8)/10NN	10	2.2	4.1×10^{8}	2.2×10^{11}	4.0
	0 to 6.7	0.44	(9)/3NN	10	3.1	4.8	81	1.2
	0 to 6.7	0.45	(9)/10NN	10	2.0	2.5	22	2.4
AZ1	0 to 3.5	-0.10	(8)/3NN	1	2.0	3.8	369	0.8
	0 to 3.5	-0.07	(9)/10NN	5	1.0	1.1	3.4	6.5
AZ2	0 to 5.4	0.75	(8)/3NN	1	1.0	4.5×10^{8}	2.0×10^{11}	4.0
	0 to 5.4	0.70	(9)/10NN	5	0.7	0.7	2.7	17
AZ3	0 to 5.4	0.74	(8)/3NN	1	0.7	4.8×10^{2}	1.2×10^{6}	5.7
	0 to 5.4	0.70	(9)/10NN	1	0.6	0.7	3.1	14

Table 3. The tightness and reliability results of nonconformity measure (8) and (9) on the test data. 3NN and 10NN denote the number of near neighbors used. Range is the range spanned by the labels in the different datasets (a constant was added to all labels for each dataset to make the range start at 0) and R_t^2 is the R^2 statistic of the SVM predictions on the test datasets. Measure indicates which nonconformity measure that was used and number of runs shows the number of repetitions used when selecting the calibration set. Median, mean, and max width show statistics of the widths of the obtained prediction intervals. % outside shows the number of predictions outside the 95% prediction interval.

The number of predictions outside the 95% prediction interval is too high on two of the tested datasets (AZ2 and AZ3, see Table 3). This indicates that the *i.i.d.* assumption is not valid for these datasets; a quantile-quantile plot of the α_i from the calibration set and the α_i from the test set (plot not shown due to space constraints) and testing for equality of the distribution of α_i from the calibration set and the α_i from the test set further strengthen this conclusion (Kolmogorov-Smirnov test; p-value $< 10^{-8}$).

QSAR models are often described as "global" or "local", where global models are derived on dataset containing structurally diverse compounds reflecting a range of different mechanistic actions, whereas local models are more focused and usually based on smaller sets of chemically similar compounds. In this paper we used ICP due to its computational efficiency, which is likely the only realistic approach for global QSAR modeling. However, for the smaller sets of compounds in local models, TCP makes more sense.

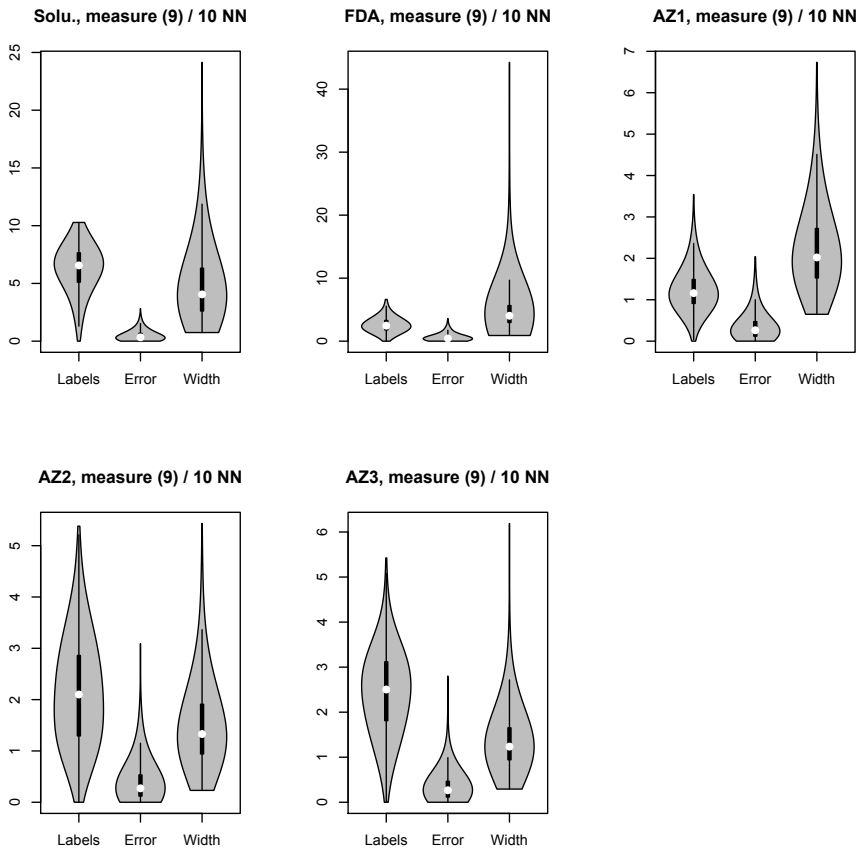

Fig. 2. Distributions of the labels, test set prediction errors, and prediction interval width for measure (9) applied to each dataset. The white dot shows the distribution mean and the black rectangle illustrates the 25th and the 75th percentile.

In conclusion, conformal prediction provides a unified view of previous attempts to estimate the reliability of QSAR prediction. By regarding AD estimates as nonconformity measures, conformal prediction permits an intuitive interpretation (p-values and prediction intervals) of AD measures, which are otherwise often very difficult to interpret.

References

1. Netzeva, T.I. *et al.*: Current status of methods for defining the applicability domain of (quantitative) structure-activity relationships. The report and recommendations

of ECVAM Workshop 52. Altern Lab Anim **33**(2) (Apr 2005) 155–73
2. Dragos, H., Gilles, M., Alexandre, V.: Predicting the predictability: a unified approach to the applicability domain problem of QSAR models. J Chem Inf Model **49**(7) (Jul 2009) 1762–76
3. Jaworska, J., Gabbert, S., Aldenberg, T.: Towards optimization of chemical testing under REACH: a Bayesian network approach to Integrated Testing Strategies. Regul Toxicol Pharmacol **57**(2-3) (2010) 157–67
4. Bassan, A., Worth, A.P.: Computational Tools for Regulatory Needs. In: Computational Toxicology: Risk Assessment for Pharmaceutical and Environmental Chemicals. John Wiley & Sons, Inc. (2007) 751–775
5. Vovk, V., Gammerman, A., Shafer, G.: Algorithmic Learning in a Random World. 1 edn. Springer (2005)
6. Shafer, G., Vovk, V.: A Tutorial on Conformal Prediction. Journal of Machine Learning Research **9** (2008) 371–421
7. Halgren, T.A.: Merck molecular force field. I. Basis, form, scope, parameterization, and performance of MMFF94. Journal of Computational Chemistry **17**(5-6) (1996) 490–519
8. Bosnić, Z., Kononenko, I.: Comparison of approaches for estimating reliability of individual regression predictions. Data Knowl. Eng. **67**(3) (December 2008) 504–516
9. Faulon, J.L., Visco, Jr, D.P., Pophale, R.S.: The signature molecular descriptor. 1. using extended valence sequences in QSAR and QSPR studies. J Chem Inf Comput Sci **43**(3) (2003) 707–20
10. Faulon, J.L., Collins, M.J., Carr, R.D.: The signature molecular descriptor. 4. Canonizing molecules using extended valence sequences. J Chem Inf Comput Sci **44**(2) (2004) 427–36
11. Vapnik, V.N.: Statistical learning theory. 1 edn. Wiley (1998)
12. Chang, C.C., Lin, C.J.: LIBSVM: A library for support vector machines. ACM Transactions on Intelligent Systems and Technology **2** (2011) 27:1–27:27 Software available at `http://www.csie.ntu.edu.tw/~cjlin/libsvm`.
13. Papadopoulos, H., Vovk, V., Gammerman, A.: Regression conformal prediction with nearest neighbours. J. Artif. Int. Res. **40**(1) (2011) 815–840
14. Huuskonen, J.: Estimation of aqueous solubility for a diverse set of organic compounds based on molecular topology. Journal of Chemical Information and Computer Sciences **40**(3) (2000) 773–777
15. Hintze, J.L., Nelson, R.D.: Violin plots: A box plot-density trace synergism. The American Statistician **52**(2) (1998) 181–84
16. Adler, D.: vioplot: Violin plot. (2005) R package version 0.2.
17. van Drie, J.H.: Pharmacophore discovery–lessons learned. Curr Pharm Des **9**(20) (2003) 1649–64