Statistical Evaluation of The Predictive Toxicology Challenge 2000–2001

Hannu Toivonen, Ashwin Srinivasan, Ross D. King, Stefan Kramer and Christoph Helma

November 11, 2002

Abstract

Motivation The development of *in silico* models to predict chemical carcinogenesis from molecular structure would help greatly to prevent environmentally caused cancers. The Predictive Toxicology Challenge (PTC) competition was organized to test the state-of-the-art in applying machine learning to form such predictive models.

Results Fourteen machine learning groups generated 111 models. The use of Receiver Operating Characteristic (ROC) space allowed the models to be uniformly compared regardless of the error cost function. We developed a statistical method to test if a model performs significantly better than random in ROC space. Using this test as criteria five models performed better than random guessing at a significance level p of 0.05. Statistically the best predictor was the Viniti model for female mice, with p value below 0.002. The toxicologically most interesting models were Leuven2 for male mice, and Kwansei for female rats. These models performed well in the statistical analysis and they are in the middle of ROC space, i.e., distant from extreme cost assumptions. These predictive models were also independently judged by domain experts to be among the three most interesting, and are believed to include a small but significant amount of empirically learned toxicological knowledge.

Availability PTC details and data can be found at: `http://www.predictive-toxicology.org/ptc/`

Contact `htoivone@cs.helsinki.fi`

1 Introduction

The *Predictive Toxicology Challenge (PTC)* was initiated to stimulate the development of advanced techniques for predictive toxicology models. The goal was to provide carcinogenicity predictions for a set of compounds with unknown classification, using information derived from their chemical structure alone. The challenge was taken up by 16 groups who produced 111 predictive models. The previous article (Helma & Kramer, 2002) presents the general layout of the competition, the datasets and the rules for participation. The focus of this paper is a statistical analysis of the significance of the submitted predictions.

2 Methods

2.1 ROC Curves

Ideally an *in silico* predictive carcinogenicity method should correctly identify all carcinogenic compounds while not incorrectly identifying any non-carcinogenic compounds. However, in practice, predictive results may overlap and errors be made. Two types of error are possible: errors of commission, and errors of omission. In an error of commission a compound is predicted to be carcinogenic when it is not: in an error of omission, a compound is not predicted to be carcinogenic when it is. The costs associated with these two types of error are not generally equal; they depend on the relative cost to society of people contracting cancer compared with the usefulness of the compound. Such costs are difficult to calculate and subject to considerable debate. This means that the assessment of predictive models should be robust to differing cost assumptions. This is not true for the standard approaches to measuring predictive success: use of accuracy, or maximization of coverage for a fixed false positive rate.

A better approach is to plot the predictive results in Receiver Operating Characteristic (ROC) space. ROC graphs were first developed for signal detection (Van Trees, 1971) and later extended to machine learning, e.g. by (Bradley, 1995). In a ROC plot true positive rate (or sensitivity) is plotted against the false positive rate (1- specificity). Sensitivity is the probability that a compound is predicted to be carcinogenic when it is actually carcinogenic. Specificity is the probability that a compound is predicted to be non-carcinogenic when the compound is actually non-carcinogenic. Both measures are expressed in the range 0–1. This produces a square space ranging from 0 to 1 along the two axes, or unit square. This space is called the ROC space. An ideal classifier, resulting from a perfect discrimination between carcinogenic and non-carcinogenic compounds, would be presented in ROC space as a point in the top left-hand corner.

If a range of predictive results for different models are plotted in ROC space, the convex-hull made from points to the extreme top and left (see Figure 1) defines the best available predictors. The closer the curve follows the left-hand border and the top border of the ROC space, the more accurate are the predictions made. In general a ROC curve indicates the trade-off between sensitivity and specificity, as an increase in sensitivity is accompanied by an decrease of specificity. If a predictor is not on the ROC convex-hull it is sub-optimal compared to the ones on the hull, regardless of the particular costs associated with errors of commission and omission (assuming cost is a linear function of the errors) (Provost & Fawcett, 1997).

2.2 Statistical Significance in ROC space

One problem with the use of convex-hulls in ROC space is that they do not include any measure of statistical significance. In this paper we adapt a standard statistical test to judge if a result in ROC space is significantly better than random predictions. The rationale for our p values in the ROC space is as follows. For a classifier C with N_C predicted positives, the null hypothesis is that the selection of the N_C examples was statistically independent of their true class. The p value of classifier C then expresses the probability with which random selection of N_C predicted positives would give at least as good a result as the one obtained by C.

Figure 2.A gives, as an example, the distribution of points in the ROC space for a classifier that makes 27 positive predictions (such as the Baurin_en model) for the male mice test set (156 negatives and 29 positives). The points are along a slanted line, with the most likely points closest to the diagonal. The p value of a classifier on this line is obtained by summing the probabilities above and at the same point with the classifier. In statistical terms, this corresponds to the one-tailed statistical significance of the 2×2 contingency table (confusion matrix) describing the classifier performance in terms of the number of true and false positives and true and false negatives. (Chi-square is a well known tool for this; we obtained the exact p values using Fisher's exact test.) By computing the p values for all N_C from 0 to 185 (the number of examples in the test set), one obtains p values over the whole ROC space, i.e., for all possible classifiers.

The p value isolines can be overlaid with the ROC space, to give an overview of how the p value behaves in relation to ROC. Figure 2.B shows the p value isolines for the male mice dataset. Since the p values also depend on the size and the class distribution of the data set, p values are different for different data sets and also for different amounts of unclassified cases. (Some submitted predictors left some test cases unclassified.)

The p value isolines are not symmetrical since the class distribution is skewed, cf. Figure 2.A. Further, the isolines have been drawn at the points where the p value is at most the one given in the label. For instance, 0.5 p value isoline is drawn above the diagonal since points in the diagonal tend to have p values larger than 0.5.

3 Results

We plotted each of the 111 predictive models in ROC space, see Figure 1. In these figures the convex-hull of the best predictive models is displayed. We tested all the PTC submissions to see if they were significantly better than random. For this the statistical significance, expressed as a p value, was computed for the submissions in ROC space.

Figure 3 shows the predictions again in the ROC space, this time overlaid with p value isolines. All classifiers are shown, but those that did not make predictions for all cases are in parenthesis. Their p values are more conservative than is apparent from the graphs, due to the smaller number of cases predicted.

Table 2 lists the statistically most significant results. Five submissions have p values below 0.05, three of them are below 0.005[1] Statistically the most significant predictors are clearly Viniti on female mice, and Baurin_en and Viniti on male mice. They obviously perform significantly better than random guessing. This is arguably the first conclusive proof that it is possible to empirically learn *in silico* models which can predict chemical carcinogenicity based *purely on chemical structure*.

The majority of the significant models cluster near the bottom left corner. Such models only make few positive predictions, but they are accurate in their predictions Among the five statistically best predictors, Leuven2 on male mice and Kwansei on female rats are particularly interesting as they occur in the middle of ROC space and distant from a corner. This means that these methods are optimal for non-extreme cost assumptions; the predictors are both quite accurate and applicable to a number of examples.

Given that 111 submissions were received in total, it is interesting to look at their overall performance. How significant, overall, are the results? Could the best results have been obtained by chance alone? The Bonferroni adjustment could be used to test this if the classifiers were independent. However, they are clearly

[1] In the PTE1 3 from 15 models had p-values lower than 0.05 and one model was below 0.005, in the PTE2 3 from 18 had p-values lower than 0.05, none of them was lower than 0.05. In both cases it was allowed to take advantage of biological information (e.g. short term tests).

not independent, since they had the same features and training examples available to them, some of them used similar methods, and they were tested on the same test cases. The Bonferroni adjusted p value for the statistically most significant result (Viniti, female mice) would be 0.206, but due to the contradicted independence assumption the value serves merely as a very conservative upper bound for the adjusted p value of Viniti.

A visual overview of the statistical significance of the whole set of 111 submissions is obtained by looking at the distribution of their (unadjusted) p values. By definition, under the null hypothesis p values are uniformly distributed in $[0, 1]$. Figure 4 illustrates the cumulative distribution of p values of the 111 submissions. The distribution is very close to the uniform distribution, indicating that "on average" the submitted classifiers did not make informed predictions but rather collectively perform as well as random guesses. This is clearly disappointing, and requires an explanation.

4 Discussion

4.1 Structural Similarity of the Training and Test sets

One of the fundamental assumptions which statistical and inductive learning methods are usually based on, is that the examples which a predictive model are tested on come from the same distribution as those on which the model was trained. This has two aspects:

1. The distribution of positive and negative examples (see Table 1).
2. The distribution of structural features among the training and the test set (i.e. if structural features of the test set are contained in the training set).

Although the test set contains more negative examples than the training set, the first point is of minor concern, as we have used ROC analysis for the evaluations of the results. Also note that Machine Learning algorithms outputting probabilities along with classifications are tunable for test sets with a changed class distribution (Elkan, 2001).

To estimate the structural similarity between the training and test set we have used the Molecular Feature Miner (MOLFEA) (Kramer *et al.*, 2001). This technique is based on the concept of *molecular fragments*. A molecular fragment is defined as a sequence of linearly connected heavy atoms. Figure 5 shows how to decompose an example molecule into its fragments, more examples and applications can be found in (Kramer *et al.*, 2001).

When predicting the biological activity of an untested compound the ideal case is that all structural features of this compound are already contained in the training set. Otherwise, the predictions should be labeled as unreliable, because unknown structural elements may contribute to biological activity. The Chemical Inductive Database Language of MOLFEA allows us to perform such a comparison. For a given molecule, we have to find all fragments that are present in this molecule, but not in the training set.

Of course not every unknown fragment has equal relevance. Long fragments can be decomposed into smaller ones that may provide enough information to provide a reliable estimation of biological activity. But if a short fragment, or even an element is unknown to the training set, the reliability of predictions will be much lower, if the missing fragment is relevant for a toxic mechanism. Thus there are three factors that have to be considered for the reliability of predictions:

1. The number of unknown fragments.
2. The length of unknown fragments.
3. The chemical "meaning" of unknown fragments.

These parameters can also be used to compare the information content of two datasets. To compare the structural diversity of the PTC training set (NTP) with the test set (FDA), we have used the following procedure: For each FDA compound we determined the number and size of fragments that do not occur in the NTP learning set. To obtain a reference, the same procedure was performed on the NTP dataset, using a leave-one-out procedure: Sequentially, one compound was removed from the dataset and compared with the rest of the compounds to obtain the unknown fragments for this compound.

Figure 6 shows the distribution of the number of unknown fragments among the FDA test set and the NTP training set. It is obvious that the test set contains more unknown fragments than the training set. A closer inspection reveals that the NTP dataset has a high proportion of compounds with completely known fragments (211 from 417 compounds). In the FDA data there are only 23 from 285 compounds without unknown fragments. It is tempting to see this as an indication that both datasets are structurally dissimilar. Indeed the NTP dataset contains a very diverse selection of compounds (industrial chemicals, pharmaceuticals, environmental contaminants, ...), whereas the FDA dataset is focused towards pharmaceuticals. This fact is also reflected in the distribution of the molecular sizes (measured as the number of heavy, i.e. non-hydrogen atoms per molecule) within both datasets (Figure 7). The median size of molecules in the NTP dataset is 13 heavy atoms/molecule, whereas the FDA dataset contains mostly medium-sized molecules with a median of 20 heavy atoms/molecule. Since the probability to encounter an unknown fragment increases with the size of a molecule, this is a possible explanation for the high ratio of compounds with completely known fragments in the training set. Comparing the sizes of the smallest unknown fragment within a molecule (Figure 8) gives a similar, although less pronounced result: In this case, the median size is equal in both datasets, but the FDA data is slightly skewed towards smaller (and potentially more relevant) fragments.

Summing up, we know that the test set contains structural features that are not present in the training set. But it is still unclear if the unknown features are toxicologically relevant and if they had any impact on the predictions in the challenge. To investigate this point, we have defined “easily predictable” and “poorly predictable” compounds for each sex/species group. Compounds that were correctly predicted by all optimal models (as determined by ROC analysis) were classified as ”easily predictable”, those with consistently false classifications were classified as ”poorly predictable”. Using the same procedure as above, we determined for each molecule in these groups the length and size of fragments that are not present in the NTP training set. The results are summarized in Figures 9 and 10. These figures and the associated statistics (Wilcoxon rank sum tests were used for all comparisons) clearly indicate that the distribution of unknown fragments does not differ between ”easily” and ”poorly predictable” compounds. A similar picture is obtained, by plotting the rate of correct predictions (i.e. *correct predictions/all predictions*) against the number of unknown fragments and the length of the smallest unknown fragment. Figures 11 and 12 present a summary of this data for all sex/species groups. They show clearly that there is no correlation between the number (size) of unknown fragments and the predictive accuracy.

We may therefore conclude that the structural differences between the training and the test set are not the main reason for the poor performance of the submitted models. This assumption is further substantiated by the cross-validation results on the NTP training set that are generally lower than 70% (Blinova *et al.*, 2002; Okada, 2002).

4.2 Toxicological Knowledge and Model Reliability

To investigate if the predictive models involve some (known or unknown) domain knowledge, we have asked toxicological experts to judge the value (and interpretability) of the submitted models.

This judgment resulted in the following ranking:

1. Kwansei
2. Wai
3. Leuven

It is noteworthy that two of these models are within the top five statistically strongest models. This reinforces our confidence that these models are probably the most interesting. We next review these models briefly.

The Leuven2 model for male mice considers a compound to be active if: (i) it contains bromine, or (ii) it does not contain oxygen, or (iii) it contains oxygen but no sulphur and no hydrogen bound to oxygen (i.e., if the oxygen is not alcoholic or phenolic or part of the carboxylic acid group); otherwise the compound is considered to be inactive.

This model has the benefit of being concise and simple. The first part about bromine is in general agreement with accepted toxicological knowledge. The second part has very broad coverage coupled with relatively low accuracy. The third part of the rule covers a diverse set of compounds, including ketones, aldehydes, ethers, epoxides and nitro-organic compounds.

The Leuven rules are probably too general to be of much current value to toxicologists. We believe this reflects the bias of decision tree methods, which favors short general rules. It is interesting to note that a learner which took into account the maximum number of features would probably have produced a more toxicologically relevant set of rules.

The Kwansei model was generated by the group with probably the most experience in toxicology, and the Kwansei methodology was favored by the toxicology judge as its approach resembled that of a human toxicologist. The Kwansei model is much more complicated than the Leuven2 model. We therefore only describe a couple of its rules which are judged to contain toxicological evidence. Rule Pos1 for carcinogenicity from the Kwansei model is:

```
[C-N: n] & [C-c:c:c:c: y] & [N: y]
```

This rule requires nitrogen, but not an aliphatic carbon nitrogen bond, and a set of aromatic carbon bonds (an aromatic ring) connected to a aliphatic carbon. It covers aromatic amines, nitroaromatics and azo compounds, all of them well known classes of carcinogens. Rule Neg2 for non-carcinogenicity is:

```
[S: y] & [HBD=0] & [O: y]
```

This rule requires sulphur and oxygen, but no hydrogen-bond donors (e.g. alcohols, amines, . . .).

Despite this success in generating toxicological knowledge from data, most models generated in the PTC contained little or no toxicological knowledge, which is consistent with them performing close to random on the test data.

4.3 Why is Carcinogenicity Hard to Predict?

From 111 models submitted to the Predictive Toxicology Challenge, only 5 submissions performed significantly better ($p \leq 0.05$) than random guessing. Before discussing the possible reason for this—at a first glance discouraging—result we should put it into perspective:

- We are unaware of a biological alternative to rodent carcinogenicity tests that perform better than the models submitted to the PTC. (Despite decades of efforts the *Salmonella* microsome assay seems to be still the most predictive short term test (Zeiger, 1998). The concordance with rodent carcinogenicity data is 60–70%, depending on the evaluation data.)
- The same or similar techniques as in the PTC have been successfully applied to predict other toxic effects (for a review see (Helma *et al.*, 2000)).

We have identified several possible causes that may lead to the poor predictability of carcinogenic effects:

1. The biochemical mechanisms involved in chemical carcinogenicity are too complex to be modeled by machine learning or statistical techniques.
2. The descriptors for chemical structures and properties are inadequate for predicting carcinogenicity.
3. *Structure Activity Relationship (SAR)* models ignore the importance of biological variables.
4. Rodent carcinogenicity classifications are too inaccurate to learn accurate models from them.
5. Predictions are impossible, because compounds in the test set contain too many structural features that are unknown to the training set.

As we have discussed the last point already in the previous section, we will focus on the first arguments. As it is presently impossible to decide between the alternatives, we will present arguments for and against each point, present ideas for future exploration and indicate our opinions.

Most experimental toxicologists will favor the first argument. But we have to keep in mind that it is not necessary (and impossible) to model the whole biological system (transport, metabolism, . . .), in order to make toxicity predictions. In *Structure Activity Relationship (SAR)* studies we are concerned only with the influence of chemical structure on a certain biological effect. SARs are in fact black box models that relate chemical structures to a biological outcome. (It is nevertheless possible to draw mechanistic conclusions from SAR studies.) Theoretically, machine learning techniques can learn arbitrarily complex relationships, but they will need more data for complex models than for simple ones. With this fact in mind it is possible (but beyond the limits of this article) to perform an experiment to test the complexity hypothesis, by comparing the predictivity of models based on different numbers of training examples. (For mutagenicity (Liu *et al.*, 1996) determined an optimal data base size/unit cost of ∼350 compounds). This should clarify if more than the presently publicly available ∼500 compounds are needed for accurate carcinogenicity classifications.

It is theoretically possible to generate an almost unlimited number of descriptors for chemical structures and their properties. To select a subset for a SAR study is more or less a trial and error process, especially when the underlying biological processes are to a large extend unknown. One of the goals of the PTC was to provide a variety of different descriptors in Stage 1, and let participants in Stage 2 choose the best parameters. Unfortunately none of the participants reported systematic parameter selection experiments. So the optimal representation of chemical structures and the choice of parameters for predicting carcinogenicity is still an open question and has room for improvements.

It is generally desirable to include biological information in SAR studies, e.g. to improve cross-species predictions or to identify susceptible subpopulations. For the sake of clarity we have to differentiate between three types of biological data:

1. information about the target organism
2. information about the experimental procedure
3. information about other biological effects of the same compound

In the NTP program the majority of studies were performed with genetically identical animals (F344/N rats and B6C3F1 mice), nonstandardized studies were removed from the training set during the data cleaning step (Helma & Kramer, 2002). This design ensures genetically determined variations do not influence the outcome of carcinogenicity assays, therefore it does not make sense to consider biological variables apart from sex and species.

The experimental detail that has probably the highest impact on the outcome of carcinogenicity experiments is the administration route. As the majority of compounds were administered by feed or gavage it was a deliberate and debatable decision not to include this information in the PTC. Further studies are certainly needed to clarify the impact of the administration route on the outcome of carcinogenicity assays.

Another detail that was omitted in the dataset is the tumorigenic dose TD_{50}. The consideration of dose information might help in distinguishing between weak and strong carcinogens, but it adds considerable complexity to the prediction problem, because we have to combine a classification task (discrimination between carcinogens and noncarcinogens) with a regression task (prediction of TD_{50}'s for carcinogens).

Information about biological effects on other organisms or other endpoints than the one to be predicted can be useful especially if there is a mechanistic relationship (e.g. between mutagenicity and carcinogenicity). We omitted such information from the PTC for two reasons: (i) Information of that kind is available only for a fraction of the compounds in the training and the test set. (ii) We intended to test the ability to predict biological activities purely *in silico*.

It is presently impossible to estimate the reproducibility of rodent carcinogenicity assays. Due to the high costs of these assays, we have presently no replicate experiments under standardized conditions. The 2-year rodent carcinogenicity assay is a very complex experiment that involves uncountable experimental, measurement and interpretation tasks that are later condensed into a simple carcinogen/noncarcinogen classification. As errors and mistakes accumulate, it is conceivable that complex experiments like this have a lower accuracy than comparatively simple toxicity assays (e.g. the Ames test for mutagenicity). (Gottmann *et al.*, 2001) compared data from the general literature with NTP results and found a very low concordance of the results (<60%). The NTP and FDA datasets, however have a high concordance of >80% (Helma & Kramer, 2002), but we were unable to ascertain that the replicate experiments have been performed really independently. It should also be noted that NTP priorities have shifted towards more problematic compounds (e.g. nongenotoxic carcinogens). The interpretation of these is subject of many expert discussions and classifications are presumably less reliable than those of strong, direct acting carcinogens. Thus, the possibility that the poor predictability of rodent carcinogenicity is due to uncertainties in the experimental results cannot be ruled out and requires further examination.

5 Conclusion

The aim of the PTC challenge was to test if it is possible for statistical/machine learning methods to learn models for the rodent carcinogenicity of chemical compounds.

Such a task is clearly scientifically challenging:

- Chemical carcinogenicity is a very complex process. It involves a complex network of biochemical mechanisms (transport, metabolism, DNA damage, …) that may differ from organism to organism.
- There is only a limited amount of training example (a few hundred).
- The training data is not randomly distributed.
- The training data has an unknown amount of noise.

Given these difficulties it was not surprising to see that the majority of contributions did not perform better than random guessing. It was however encouraging to observe a limited number of predictive models that performed significantly better than random (see Table 2), and were judged to have empirically learned a small but significant amount of toxicological knowledge.

Acknowledgments

We would like to thank Douglas Bristol and Yin-tak Woo for their toxicological expertise.

References

Blinova, V., Dobrynin, D., Finn, V., Kuznetsov, S. & Pankratova, E. (2002) Toxicology analysis by means of the JSM-method. *Bioinformatics,* **this volume**.

Bradley, A. P. (1995) The use of area under roc curve in the evaluation of learning algorithms. *Pattern Recognition,* **30**, 1145–1159.

Elkan, C. (2001) The foundations of cost-sensitive learning. In *Proceedings of the Seventeenth International Joint Conference on Artificial Intelligence (IJCAI'01)* Morgan Kaufmann Publishers, San Francisco, USA.

Gottmann, E., Kramer, S., Pfahringer, B. & Helma, C. (2001) Data quality in predictive toxicology: reproducibility of rodent carcinogenicity experiments. *Environ. Health Perspect.,* **109**, 509–514.

Helma, C., Gottmann, E. & Kramer, S. (2000) Knowledge discovery and data mining in toxicology. *Stat Methods Med Res.,* **9**, 329–358.

Helma, C. & Kramer, S. (2002) A survey of the predictive toxicology challenge. *Bioinformatics,* **this volume**.

Kramer, S., De Raedt, L. & Helma, C. (2001) Molecular feature mining in HIV data. In *Proceedings of the Seventh ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD-01)* pp. 136–143.

Liu, M., Sussman, N., Klopman, G. & Rosenkranz, H. (1996) Estimation of the optimal data base size for structure-activity analyses: the Salmonella mutagenicity data base. *Mutation Res.,* **358**, 63–72.

Okada, T. (2002) Characteristic substructures and properties in chemical carcinogens studied by the cascade model. *Bioinformatics,* **this volume**.

Provost, F. & Fawcett, T. (1997) Analysis and visualization of classifier performance: comparison under imprecise class and cost distribution. In *Proceedings of Knowledge Discovery in Databases 1997*.

Van Trees, H. L. (1971) *Detection, estimation, and modulation theory.* Wiley, New York.

Zeiger, E. (1998) Identification of rodent carcinogens and noncarcinogens using genetic toxicity tests: premises, promises, and performance. *Regulatory Toxicology and Pharmacology,* **28**, 85–95.

Tables

Training set (NTP)

	Positive examples	Negative examples
male rats	166	305
female rats	135	371
male mice	140	348
female mice	155	334

Test set (FDA)

	Positive examples	Negative examples
male rats	52	133
female rats	36	149
male mice	29	156
female mice	35	150

Table 1: Distribution of positive and negative examples among the training and test sets

	Sex/species		Negative	Positive	False	True
p	group	Model	examples	examples	pos. rate	pos. rate
0.0019	female mice	Viniti	84	21	0.036	0.286
0.0027	male mice	Baurin_en	156	29	0.109	0.345
0.0046	male mice	Viniti	97	14	0.031	0.286
0.0433	female rats	Kwansei	73	17	0.274	0.529
0.0488	male mice	Leuven2	153	29	0.366	0.552
0.0643	female mice	Animaths2v	148	34	0.115	0.235
0.0864	male rats	Gonzales	133	52	0.150	0.250
0.0916	female mice	Animaths1v	148	34	0.081	0.176
0.1186	female mice	Smuc1	150	35	0.280	0.400
0.1417	female rats	Viniti	69	17	0.101	0.235

Table 2: The 10 statistically most significant submissions. Negative and positive examples are the true class distributions of the test examples. If a model did not provide predictions for all compounds, the numbers are different from Table 1.

Figures

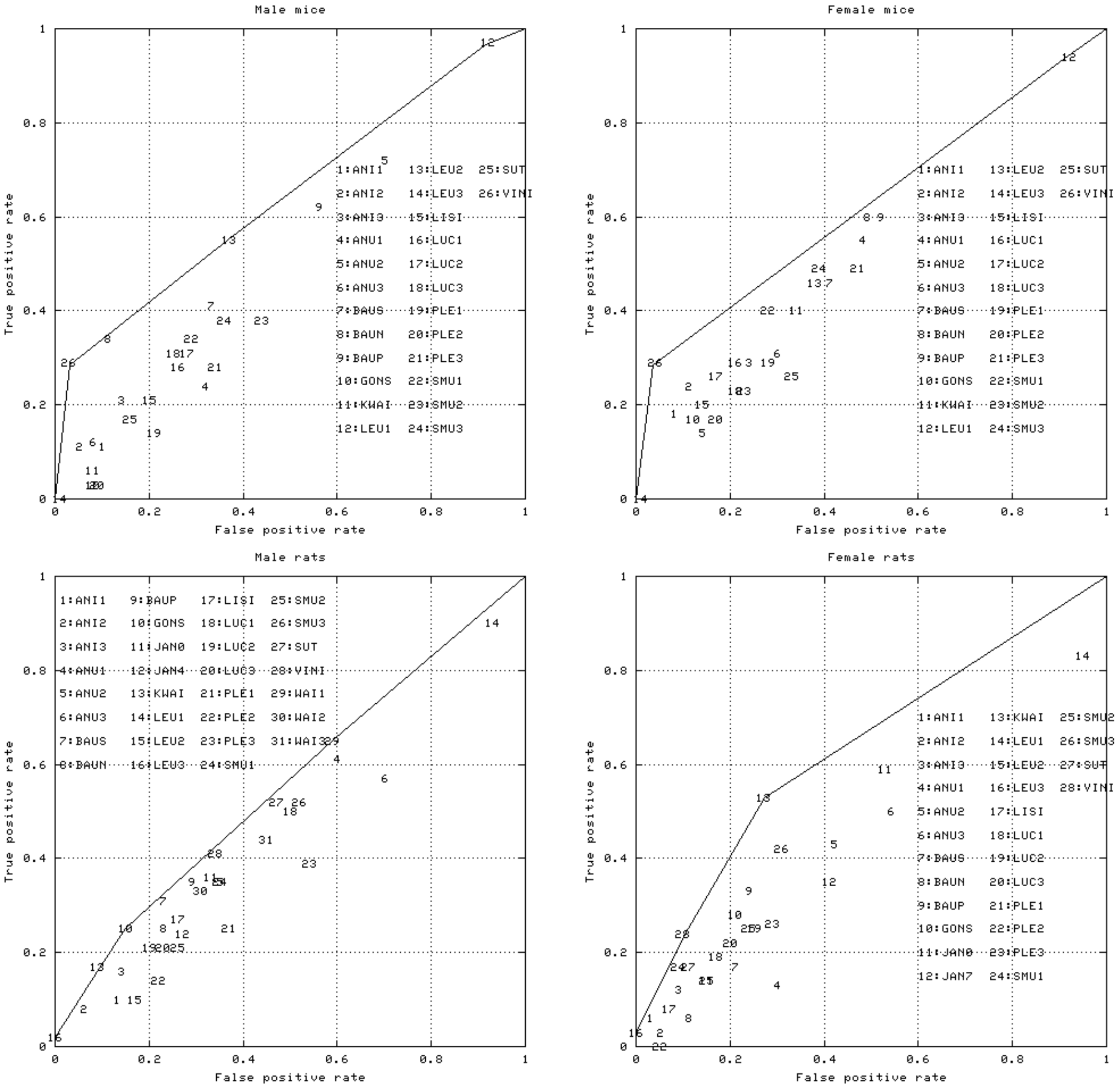


Figure 1: ROC points and convex hulls of the submissions.

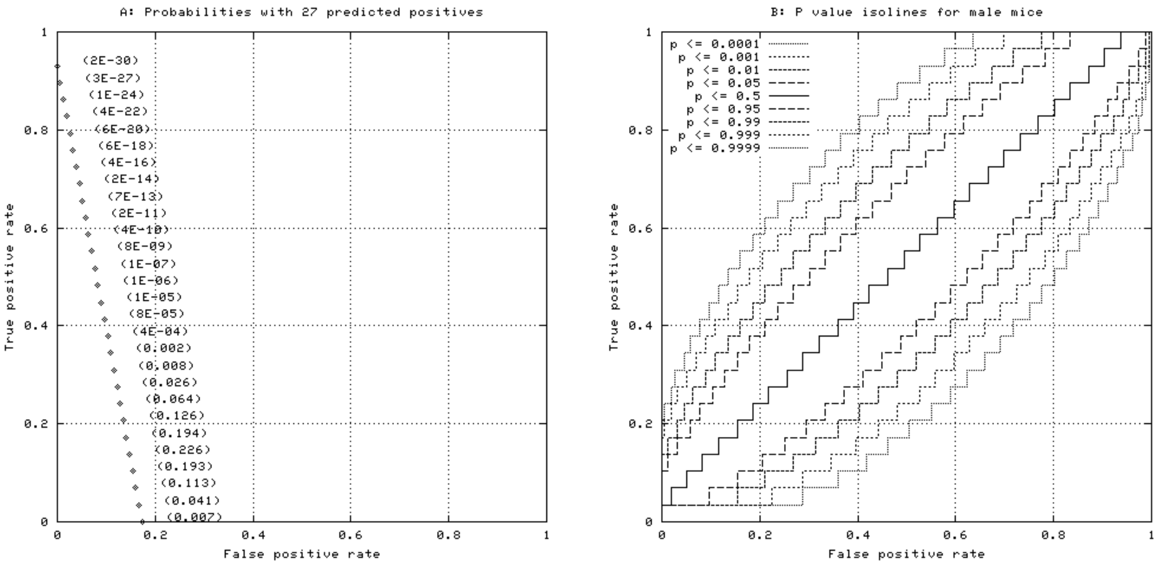


Figure 2: A: Example p values. B: Example p value isolines.

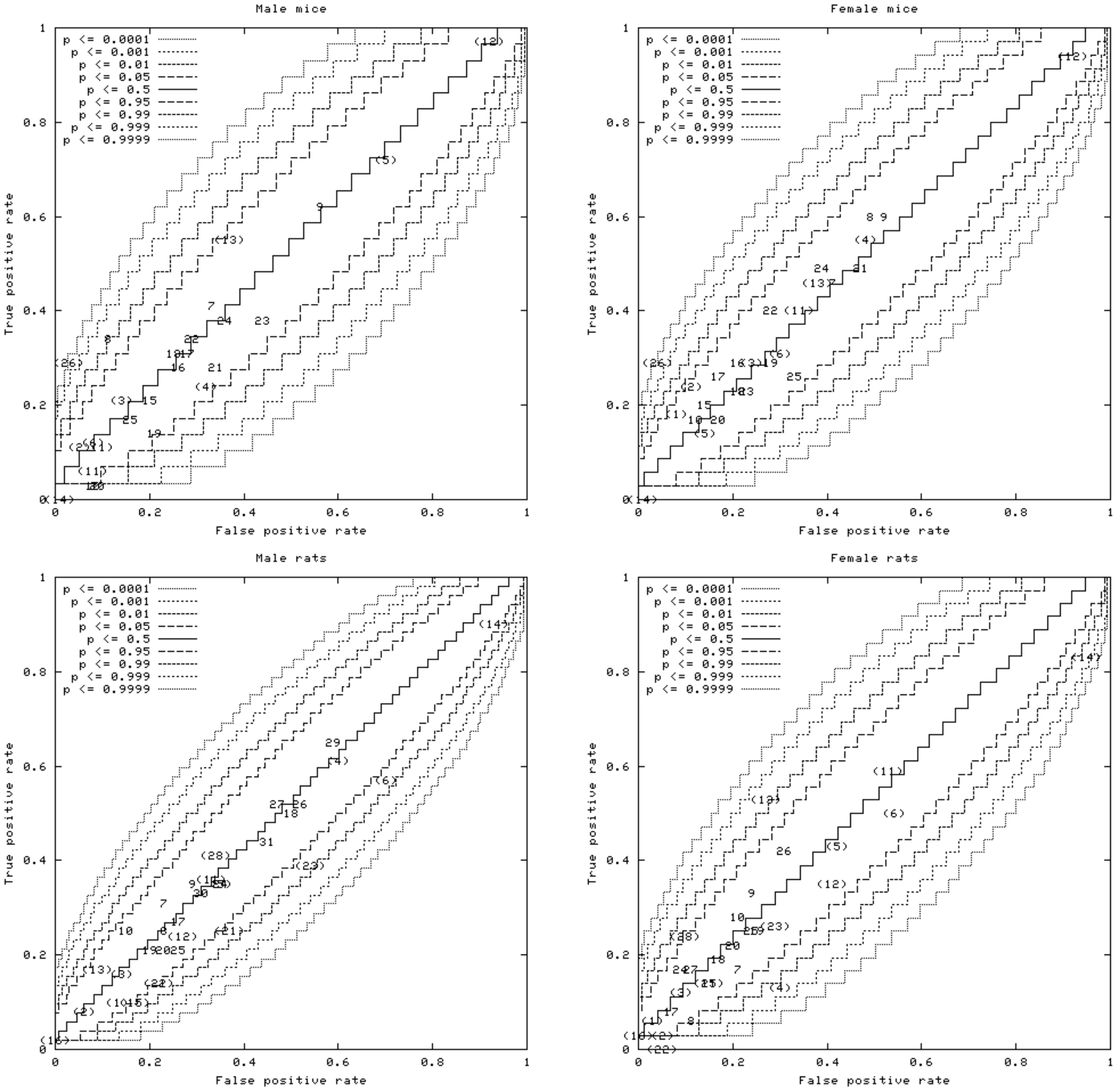


Figure 3: *P* value isolines and PTC submissions. Labels are the same as in Figure 1. (Parenthesis indicate classifiers that did not make predictions for all examples and for which the true *p* values are thus more conservative than appears from the figure.)

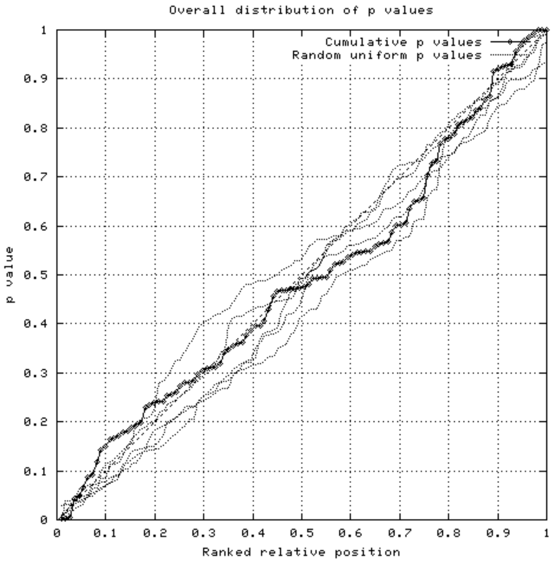
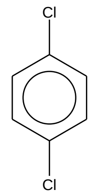


Figure 4: Cumulative p value distribution of all 111 submissions. The dotted lines, shown for comparison, are each based on 111 i.i.d. uniformly distributed random values in $[0, 1]$.

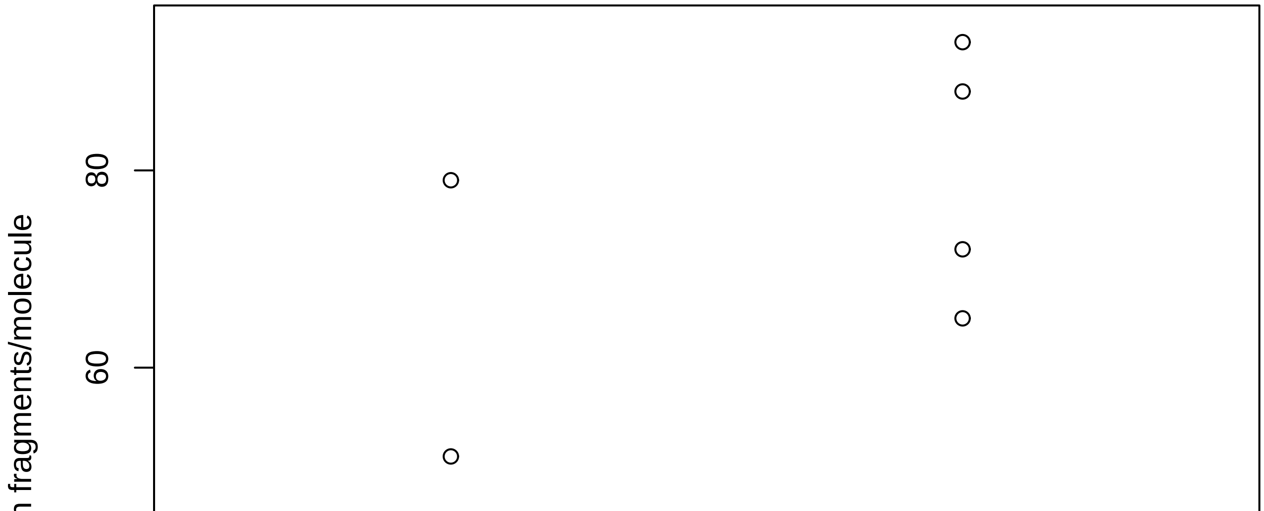
Structure



Fragments

```
Cl              c
Cl-c            c-c
Cl-c-c          c-c-c
Cl-c-c-c        c-c-c-c
Cl-c-c-c-c      c-c-c-c-c
Cl-c-c-c-c-Cl   Cl-c-c-c-c-c-c
```

Figure 5: Linear fragments of 1,4-dichlorobenzene (CAS 106-46-7)



p=1.93e−32

Figure 6: Number of unknown fragments in the training and the test set

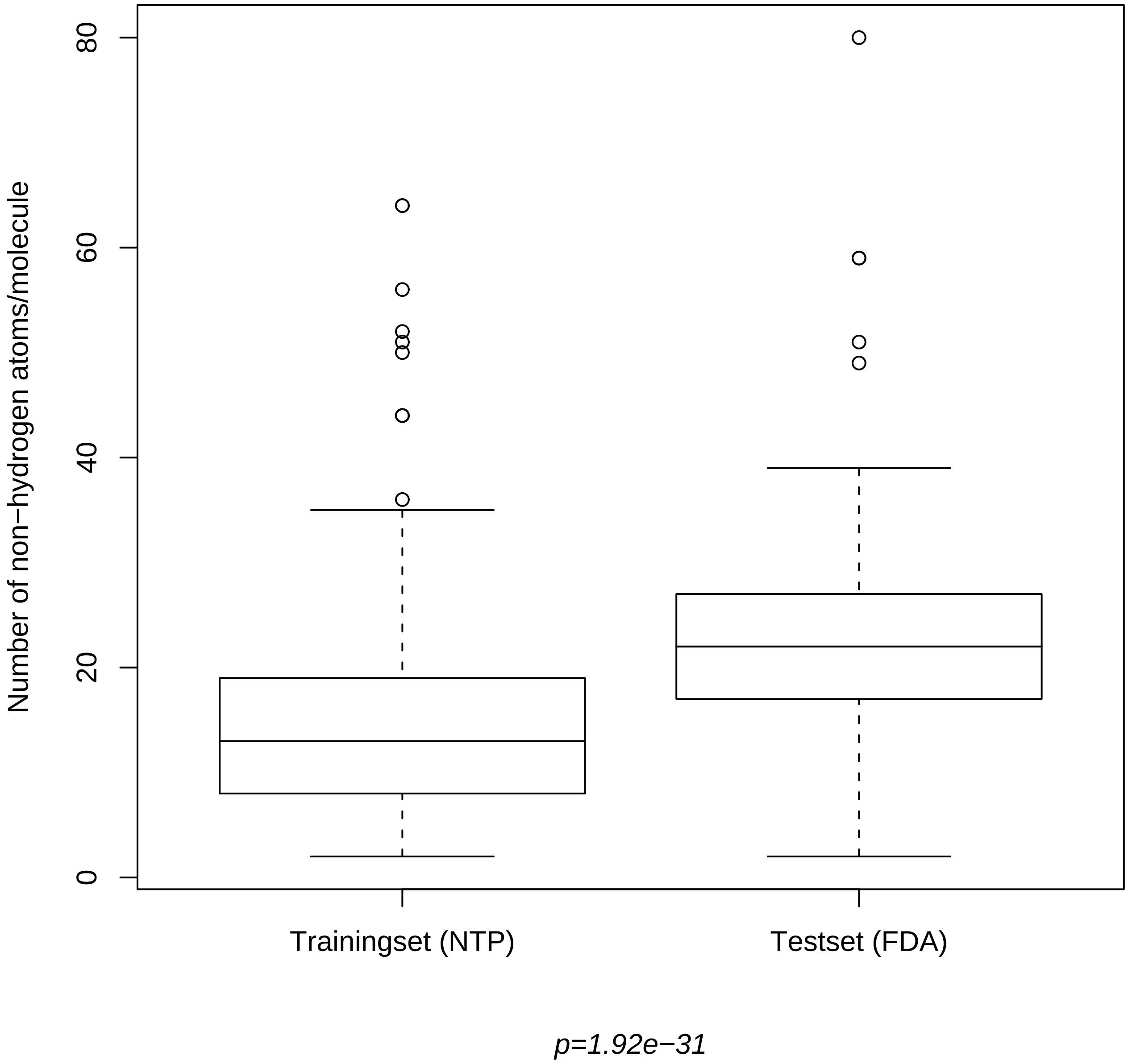


Figure 7: Distribution of molecule sizes in the training and the test set

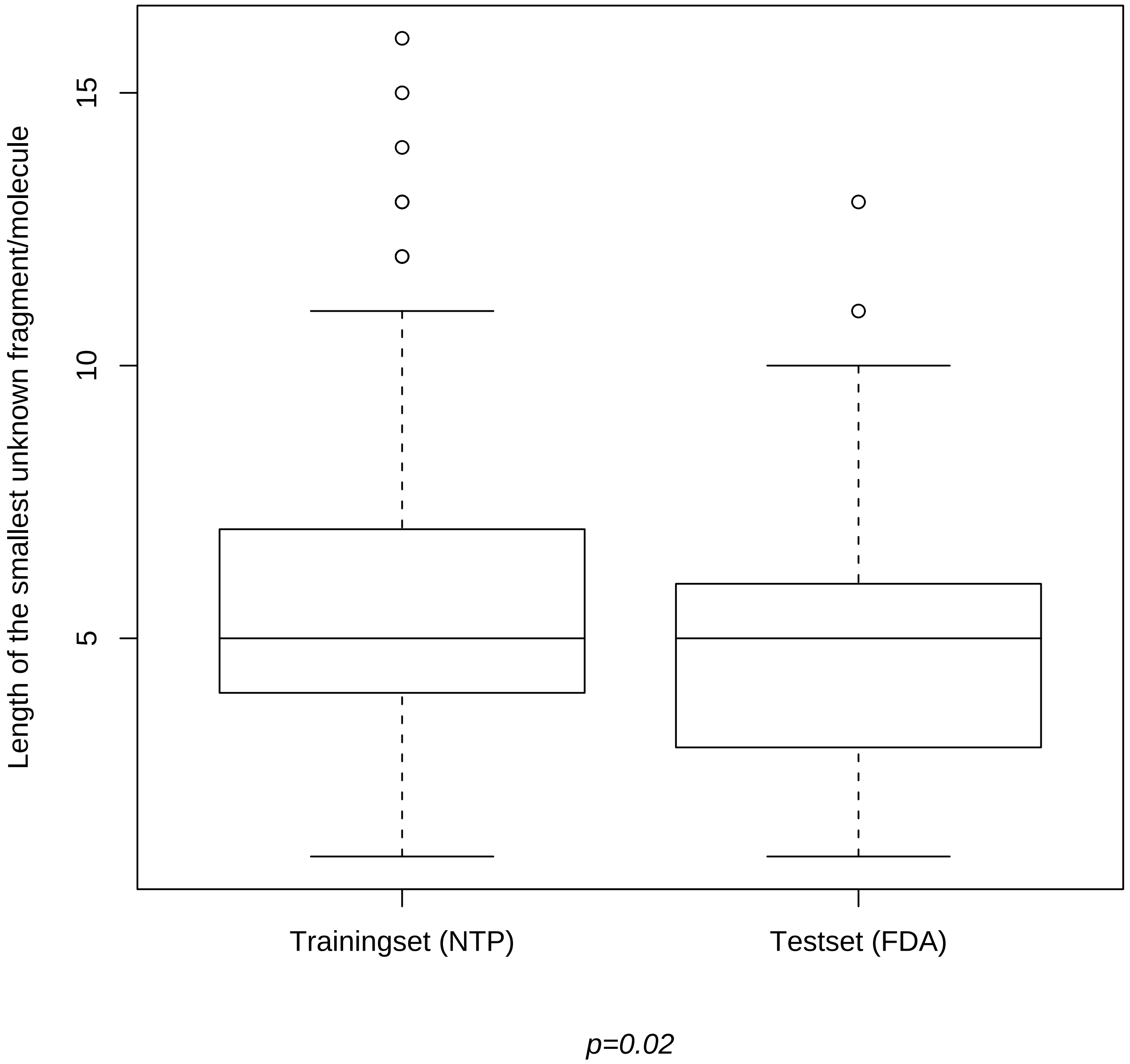


Figure 8: Sizes of the smallest unknown fragments in the training and the test set (compounds with completely known fragments are exluded)

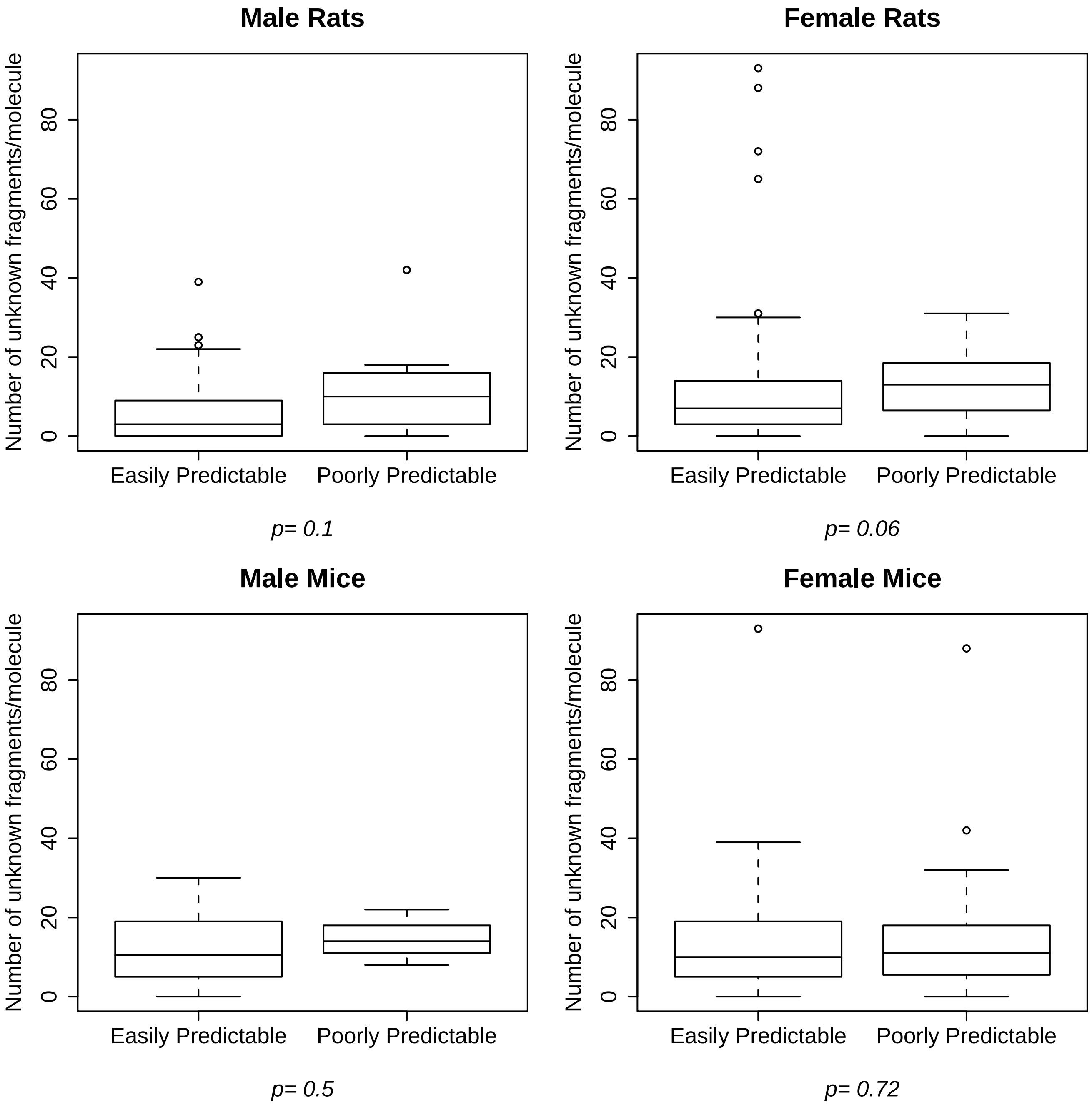


Figure 9: Number of unknown fragments in easily and poorly predictable compounds.

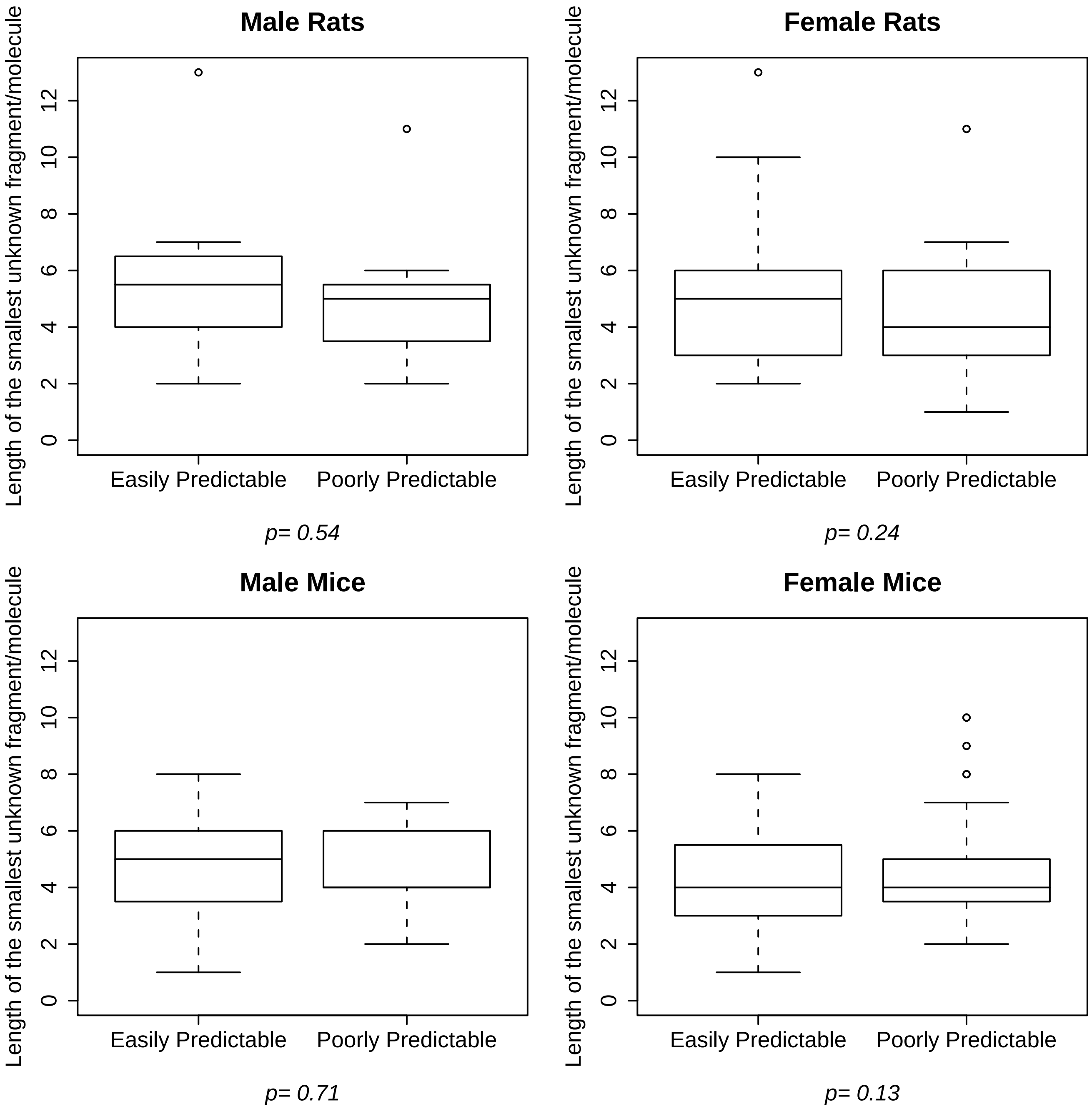


Figure 10: Sizes of the smallest unknown fragments in easily and poorly predictable compounds

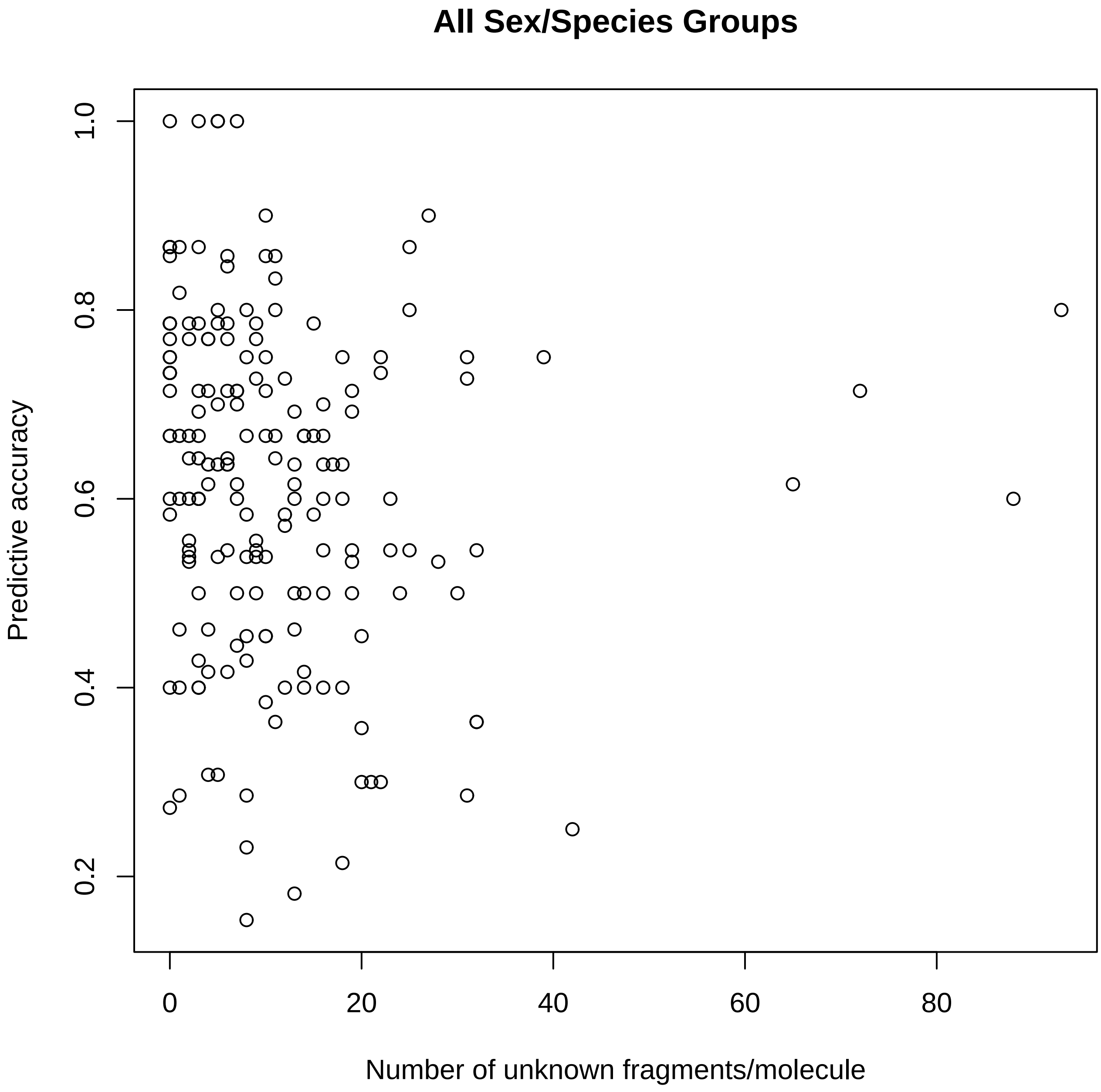


Figure 11: Number of unknown fragments vs. true prediction rate for all sex/species groups

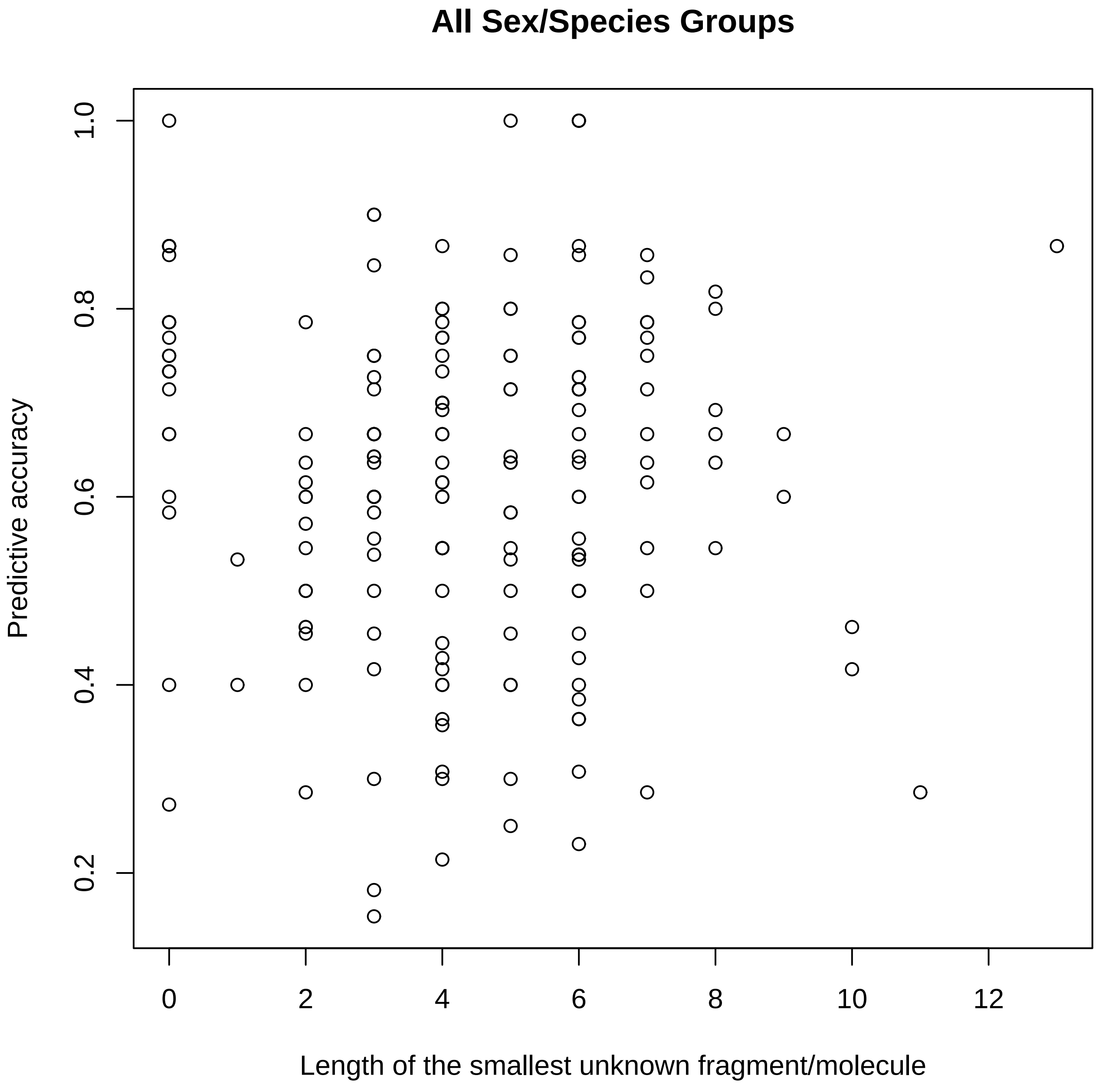


Figure 12: Sizes of the smallest unknown fragments vs. true prediction rate for all sex/species groups