An assessment of submissions made to the Predictive Toxicology Evaluation Challenge

A. Srinivasan
Oxford University Computing Lab
Wolfson Bldg Parks Road
Oxford, U.K.

R.D. King
Dept. of Computer Sc.
University of Wales
Aberystwyth, U.K.

D.W. Bristol
NIEHS
Lab. of Carcinogenesis & Mutagenesis
RTP, NC U.S.A.

Abstract

Constructing "good" models for chemical carcinogenesis was identified in IJCAI-97 as providing a substantial challenge to "knowledge discovery" programs. Attention was drawn to a comparative exercise which called for predictions on the outcome of 30 rodent carcinogenicity bioassays. This - the Predictive Toxicology Evaluation (or PTE) Challenge - was seen to provide AI programs with an opportunity to participate in an enterprise of scientific merit, and a yardstick for comparison against strong competition. Here we provide an assessment of the machine learning (ML) submissions made. Models submitted are assessed on: (1) their accuracy, in comparison to models developed with expert collaboration; and (2) their explanatory value for toxicology. The principal findings were: (a) using structural information available from a standard modelling package, layman-devised features, and outcomes of established biological tests, results from ML-derived models were at least as good as those with expert-derived techniques. This was surprising; (b) the combined use of structural and biological features by ML-derived models was unusual, and suggested new avenues for toxicology modelling. This was also unexpected; and (c) significant effort was required to interpret the output of even the most "symbolic" of ML-derived models. Much of this could have been alleviated with measures for converting the results into a more "toxicology-friendly" form. As it stands, their absence is sufficient to prevent a whole-hearted acceptance of these promising methods by toxicologists. This suggests that ML techniques have been able to respond - not fully, but nevertheless substantially - to the PTE Challenge.

1 Introduction

In his essay "Two conceptions of science" [Medawar, 1984], the distinguished biologist Peter Medawar describes the valuation of contributions to science thus:-

> Here then are some of the criteria used by scientists when judging their colleagues' discoveries and the interpretations put upon thern. Foremost is their *explanatory value* - their rank in the grand hierarchy of explanations and their power to establish new pedigrees of research and reasoning. A second is their clarifying power, the degree to which they resolve what has hitherto been perplexing. ..

Explanations that reach this stage of inspection are usually understood to have achieved an acceptable level of accuracy, however measured. With the emergence of ML programs capable of constructing empirical generalisations from scientific data, it is possible to examine the extent to which such machine-authored descriptions meet the criteria used to judge their human counterparts. This is of special interest if such programs are intended to act as genuine scientific assistants to experts.

One area for conducting such an examination was proposed in the form of the Predictive Toxicology Evaluation Challenge (in IJCAI-97, see [Srinivasan *et a*/., 1997]). The problem of predicting chemical carcinogenesis described there is particularly well-suited as a testbed for a number of reasons. Besides its undisputed humanitarian value, principal reasons are that: (1) there is an urgent need for low-cost, accurate toxicity models that can reduce a reliance on slow, expensive rodent bioassays [Bristol *et a*/., 1996]; (2) there is much to be learnt about the molecular mechanisms underlying carcinogenic activity; and (3) there is a well-established scientific programme within the U.S. National Toxicology Program (NTP) concerned with the comparative evaluation of toxicity models (which may be of human origin, see: dir.niehs.nih.gov/dirlecm/pte2.htm1). These provide machine-based "hypothesis constructors" the opportunity to construct accurate models, which may yield

[1] All Internet sites mentioned in this paper are to be prefixed with http:// unless otherwise indicated

new insights and subject to review much in the manner described by Medawar.

This paper reports on the machine learning (ML) submissions made to this IJCAI challenge from its inception in August, 1997 to December, 1998. The paper is organised as follows. Section 2 summarises the course of the challenge from 1997, and presents the models selected for further evaluation. Section 3 contains an assessment of the accuracies of the ML models in comparison to those developed under the guidance of expert toxicologists (this includes toxicology expert systems). Section 4 contains an appraisal of the explanatory value of the ML models[2]. Section 5 concludes this paper.

2 The IJCAI PTE Challenge: details and submissions

As part of the NTP, the National Institute of Environmental Health Sciences (NIEHS) organises the Predictive Toxicology Evaluation (or PTE) project. The project [Bristol *et* a/., 1996] is concerned with predicting the outcome of rodent bioassays measuring the cancerous activity of a pre-specified set of compounds. In its simplest setting, predictions are restricted to either "POS" to denote carcinogenic, or "NEG" if otherwise. There is no restriction on the type of method used to construct the toxicity model. The PTE project accepted predictions until late 1996 for 30 compounds (collectively known as PTE2) undergoing bioassays within the NTP - the last of these assays being completed by June, 1998.

The relevance of the PTE project to programs concerned with "knowledge discovery" directly led to the PTE Challenge in IJCAI-97. Here, it was proposed to collect submissions from AI techniques. Submissions were to be made at a prescribed Internet site (*www. comlab.ox. ac.uk/oucl/groups/machlearn/PTE)* and consisted of two parts: (1) *prediction:* POS and NEG classification for the PTE2 compounds; and (2) *description:* details of the materials and methods used, and results obtained with the technique. The former was needed to assess model accuracy, and the latter for replicabiiity of results and evaluations of model comprehensibilty.

The site accepted submissions from August 29, 1997 (one week after the challenge was announced at IJCAI-97). Submissions received up to November 15, 1998 were eligible for assessments of chemical comprehensibility. The challenge was regularly advertised at major AI conferences and in electronic newsgroups, and our records indicate that the data provided by the challenge site were retrieved over 100 times[3]. By November 15 1998, 9 legal submissions were received (by legal here we mean that both "prediction" and "description" parts of the submission were in order). These are summarised in Figure 1. Space restrictions prevent us from providing a description here of each entry the reader is directed to the Internet site under the "Description" column for complete details[4].

At this point, it is worth noting an important point of difference between the submissions made to the NTP's PTE project, and those in Figure 1. All predictions by the former were made before true classifications on any chemical in PTE2 were known. The timing and duration of the IJCAI challenge has precluded the possibility of such a truly blind trial. We rely on submissions to abide by challenge regulations that prevent the use of PTE2 classifications in any way to direct model formation or selection.

3 Assessment of predictive accuracy

At the time of writing this paper, the classification of 23 of the 30 compounds had become available. Figure 2 tabulates the predictive accuracies achieved by the models described in the submissions in Figure 1 (henceforth called "ML-derived models").

Benigni [Benigni, 1998] provides a tabulation of the predictions made by several toxicity prediction methods on a subset of the PTE2 compounds. We concentrate here on those techniques that involve substantial input from experts. These include models devised directly by toxicologists or those that rely on the application of compilations of such specialist knowledge (that is, expert systems). In [Benigni, 1998], there are 9 such "expert-derived" models due to: Huff *et al.* (HUF, [Huff *et al,* 1996]), OncoLogic (ONC, [Woo *et* a/., 1997]), Bootman (BOT\ [Bootman, 1996]), Tennant *et al.* (TEN, [R.W. Tennant, 1996]), Ashby (ASH, [Ashby, 1996J), Benigni *et al* (BEN, [R.Benigni *et al.,* 1996]), Purdy (PUR, [Purdy, 1996]), DEREK (DER, [Marchant, 1996]), and COMPACT/HAZARDEXPERT (COM, [Lewis *et* a/., 1996]). Excluding missing entries, predictions are available from these methods for 18 PTE2 compounds. A comparative tabulation on this subset against the ML-dcrived models is in Figure 3.

Comparisons based on predictive accuracy overlook an important practical concern, namely that the costs of different types of errors may be unequal. In toxicology modelling, the cost of false negatives is usually higher than those of false positives. Borrowing from terminology in signal-detection, "sensitivity" refers to the fraction of POS chemicals classified as POS by a model; and "specificity" refers to the fraction of NEG chemicals

[2] Performed by one of the authors (D.W.B.), who is a toxicologist.

[3] Clearly, only very limited conclusions can be drawn from this figure. Our records suggest that the data were extracted by groups with a wide range of research interests. However, the reader will note that the final submissions appear to be largely from those interested in Inductive Logic Programming (ILP). While it is possible that the emphasis on a descriptive component discouraged the use of methods like neural networks, we have no way of knowing why some ML researchers failed to respond to the challenge.

[4] The reader should note that the submission OU2 was from two of the authors here (A.S. and R.D.K.). As far as we are aware, none of the submissions appear to have involved a toxicologist during model-development.

Submission	Method	Description
LE1	ILP	*www.cs.kuleuven.ac.be/~hendrik/PTE/PTE1.html*
LE2	ILP	*www.cs.kuleuven.ac.be/~ldh/PTE/PTE2.html*
LE3	ILP	*www.cs.kuleuven.ac.be/~wimv/PTE/PTE2.html*
LRD	Stochastic voting technique	*www.lri.fr/~fabien/PTE/Distill/*
LRG	Stochastic rule construction	*www.lri.fr/~fabien/PTE/GloBo/*
OAI	Decision tree and Naive Bayes	*www.ai.univie.ac.at/~bernhard/pte2/pte2.html*
OU1	Decision tree and ILP	*www.comlab.ox.ac.uk/oucl/groups/machlearn/PTE/oucl1.html*
OU2	Decision tree and ILP	*www.comlab.ox.ac.uk/oucl/groups/machlearn/PTE/oucl2.html*
TA1	Genetic search	*ailab2.cs.nthu.edu.tw/pte*

Figure 1: Legal submissions to the PTE Challenge. Here "ILP" stands for Inductive Logic Programming.

Submission	Accuracy
LRD	0.87 (0.07)
LRG	0.78 (0.09)
OU2	0.78 (0.09)
OAI	0.74 (0.09)
LE3	0.70 (0.10)
LE2	0.65 (0.10)
OU1	0.57 (0.10)
TA1	0.52 (0.10)
LE1	0.48 (0.10)
DEF	0.74 (0.10)

Figure 2: Estimated accuracies of submissions made to the PTE Challenge. Here, accuracy refers to the fraction of PTE2 compounds correctly classified by the ML-derived model. The quantity in parentheses next to the accuracy figure is the estimated standard error. The classifications are based on the outcome of 23 of the 30 PTE2 bioassays. The classification of remaining 7 is yet to be decided. "DEF" refers to the simple rule that states that all compounds will be "POS". This was not an official submission to the challenge and is only included here for completeness.

classified as NEG by a model. Figure 4 is a scatter-diagram that shows the position of each model in this two-dimensional probability space.

Complementary to sensitivity and specificity are: the fraction of POS predictions that are actually POS, and the fraction of NEG predictions that are actually NEG. Termed here as "positive predictivity" and "negative predictivity", these measure the accuracy of each type of prediction. Good models should exhibit high predictivity values. Figure 5 shows the scatter-diagram of the models along these dimensions.

Keeping in mind the mandatory caution that must be exercised when interpreting figures derived from such small test-sets, Figures 2, 3, 4 and 5 appear to suggest that ML-derived models are able to at least match the performance attained by their expert counterparts. As is evident, 7 of the 9 ML-derived models achieve the predictive accuracy threshold set by the DEF model. This is in contrast to 3 of the 9 expert-derived models. 6 ML-derived models (LE2, LE3, LRD, LRG, OAI, OU2) achieve false-negative error rates of at most 0.25 with false-positive rates of no more than 0.50. This is matched by only 1 expert-derived model (ONC). 7 ML-derived models (the previous 6 and OU1) also achieve positive and negative predictive rates of at least 0.50 (Figure 5). This is in contrast to 4 expert-derived models. Elsewhere, we present a more detailed assessment, of these trends based on a cost-sensitive technique termed ROC-analysis [Srinivasan *et al.*, 1999]. Due to space restrictions, a summary has to suffice here. The analysis shows the ML-derived models to be extremely competitive, with LRG being the pick of the best across a range of reasonable error-costs and prior distributions over class values. LRG was obtained with a stochastic technique which resulted in rules that use, amongst others, attributes encoding the results from ILP methods.

4 Assessment of explanatory value

At the outset of this section, it is worth emphasising that as submitted, none of the ML-derived models would be considered toxicology acceptable. This comment extends even to the most transparent submission like OU2, which presents a relatively simple (by ML standards) decision-tree obtained from a well-known algorithm (C4.5). Much of this probably stems from a lack of toxicology expertise amongst the program users, and the lack of any "client specifications" in the statement of the challenge. We intend to rectify the latter in future experiments (PTE-3, see [Srinivasan *et al.*, 1999]). However, some attempt by all developers at improving clarity by including tables of

Model	Type	Accuracy
LRD	ML	0.89 (0.07)
LRG	ML	0.84 (0.09)
HUF	Expert	0.78 (0.10)
LE3	ML	0.78 (0.10)
OAI	ML	0.78 (0.10)
ONC	Expert	0.78 (0.10)
OU2	ML	0.78 (0.10)
LE2	ML	0.72 (0.11)
BEN	Expert	0.67 (0.11)
OU1	ML	0.67 (0.11)
TA1	ML	0.62 (0.11)
ASH	Expert	0.56 (0.12)
LE1	ML	0.56 (0.12)
TEN	Expert	0.56 (0.12)
BOT	Expert	0.50 (0.12)
COM	Expert	0.50 (0.12)
DER	Expert	0.50 (0.12)
PUR	Expert	0.28 (0.11)
DEF	-	0.67 (0.11)

Figure 3: Comparison of estimated accuracies of expert and ML-derived models. The figures are based on the classification of 18 of the 30 PTE2 compounds for which predictions are available from all models. As before, estimates of standard errors are in parentheses. Some expert-derived models include a third category of classification called "borderline carcinogen." These are simply taken as a POS classification here. As before DEF predicts all chemicals as POS.

names and structures identified by the rules, clear statements of their reliability etc., would have greatly assisted the evaluation exercise. The application of ML techniques to modelling toxicity endpoints is a relatively new research development in toxicology, and it is essential that descriptions of representations used and results obtained are as thorough as possible (see for example, [Bristol, 1995] for an appraisal of the requirements of models from both developer and prospective-client points of view). Nevertheless, the performance of the models have been sufficiently intriguing to foster further examination.

In performing an evaluation of the explanatory value provided by the ML-derived models, we have found it instructive to examine their contributions in the following categories:

A. Those that suggest any new lines of investigation for toxicology modelling;

B. Those that confirm, clarify or contribute to current ideas in toxicology; and

C. Those that are uninteresting or unlikely.

Our examination is restricted to the models that showed the most promise in the previous section, namely: LE2, LE3, LRG, OAI, OU2. Unfortunately, the most accurate model (LRD) could not be considered, as no explicit model was provided. Further, it is not our intention to single out any one model as being the "best" - rather, it is to provide an overall assessment of the value of using ML methods in toxicology.

Of most interest is the frequent use of combinations, in models like LRG, of chemical structure and biological tests. For some time, there has been vigourous debate on how classical structure-activity modelling can be applied to toxicity problems. This form of modelling relates chemical features to activity, and works well *in-vitro*. The extent to which these ideas transfer to toxicity modelling - which deals with the interaction of chemical factors with biological systems - is not evident. By using a combination of chemical features and biological test outcomes, the ML-derived models provide one possible method for dealing with the chemical effects in such "open" systems. If the accuracies obtained with such rules are borne out on larger datasets, then this would constitute a significant advance in structure-activity modelling for toxicology. This is certainly worth further investigation and falls in Category A.

A number of aspects of some of the models can be categorised in Category B. As an example, OU2 selects a combination of mouse lymphoma and Drosophilla tests as a strong indicator of carcinogenicity. Many toxicologists believe that relationships exist between genotoxicity and carcinogenicity. While the only accepted correlation involves the Salmonella assay, this rule suggests a different combination of short-term tests could be equally, or more effective. Similar comments could be made on a number of other fronts: the presence of methoxy groups, sulphur compounds, and biphenyl groups are all identified in various ways as being related to toxicity. These are in line with what is currently

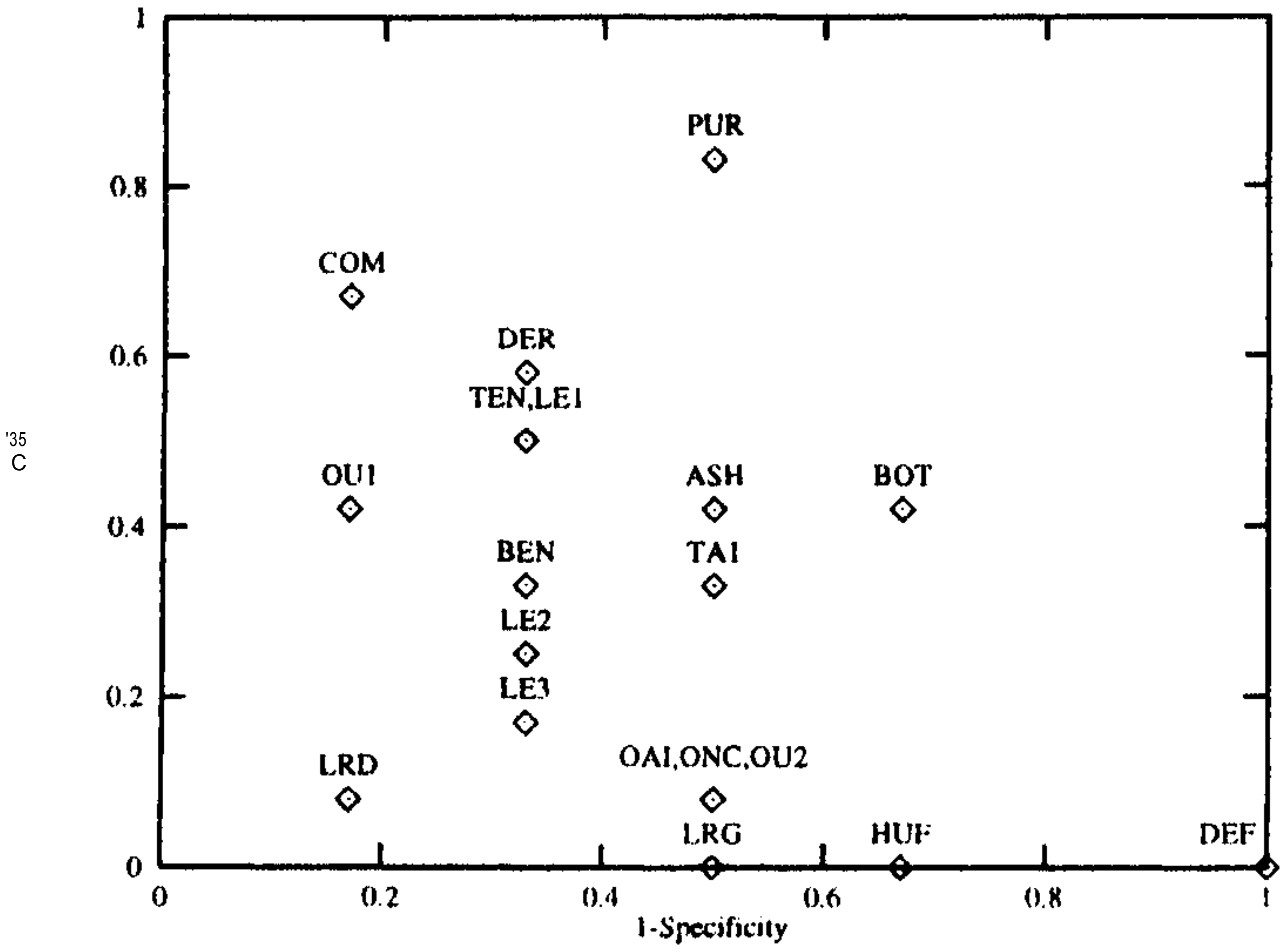

Figure 4: Scatter-diagram showing the performance of expert and ML-derived models based on their false positive (x-) and false negative *(y-)* error rates. For two models with the same x-value, the one with the smaller y-value is preferable.

known in toxicology.

Given the relative opaqueness of the output, it is hard to judge the extent to which the models have identified aspects in Category C. The general approach in toxicology is to be wary of "explanations" that only pertain to a few or uninteresting chemical structures. These do occur in the models submitted, and appeal to have been ignored (except in the case of OU2, where some editing attempt is undertaken). We do not enumerate examples of this here.

5 Conclusions

Toxicology is a young science that is primarily driven by intense health and industrial interests focused on specific chemical substances. A practicing toxicologist is regularly confronted with urgent requests to provide reliable information about the next substance of interest - whatever it might be. This situation demands that the toxicologist be able to call on, or develop, predictive models that are not only accurate, but also cover an extremely wide range of noncongeneric dissimilar chemicals. These range from pure organic and inorganic compounds to polymers and complex mixtures. Predictions also need to be generated for a variety of toxicity endpoints [Bristol, 1995]. Aspiring assistants - human or otherwise - seeking to aid an expert toxicologist in this model-building endeavour, must be capable of suggesting robust solutions that are accurate and understandable. This forms the crux of the PTE Challenge - do AT programs meet these requirements when constrained to the task of predicting chemical carcinogenesis? The short answer, for the submissions participating in the challenge, is: "not yet." The qualifier is important though, as they do show considerable potential to achieve this goal. This opinion is based on the evidence that (a) models developed by these programs are clearly competitive on accuracy terms with those derived with significant expert assistance; and (b) even with almost no effort made to render the output chemically understandable the models have still suggested unusual ways to proceed with toxicology modelling. Whether such programs can make the transition from promising apprentices to valuable assistants will depend on whether their developers recognise the paramount importance of ensuring that the models are phrased in terms familiar to a toxicologist, and on continued good results with larger datasets.

Acknowledgements

The authors would like to acknowledge the significant effort made by the Machine Learning community in responding to the PTE Challenge. In particular, sincere thanks are due to the groups whose models we have used in the analysis here, namely those at Leuven (Belgium), LR1 (France), OFAI (Austria), and the AI Lab, Taiwan. At Oxford, the model OUI was largely developed by Ngozi Dozie, a MSc student in Computation. A.S currently holds a Nuffield Trust Research Fellowship at Green College, Oxford. During the first six months of the PTE Challenge, he was supported by Smith-Kline Beecham. A.S. and R.D.K would also like to thank Donald Michie, Stephen Muggleton, and Michael Sternberg for interesting and useful discussions concerning the use of machine learning for predicting biological activity. We thank the support staff' of the Computing Laboratory, Oxford in particular, Ian Collier for his help with the Wet) pages related to the PTE Challenge.

References

[Ashby, 1996] J. Ashby. Predictions of rodent carcinogenicity for 30 compounds. *Environmental Health Perspectives,* pages 1101 1104, 1996.

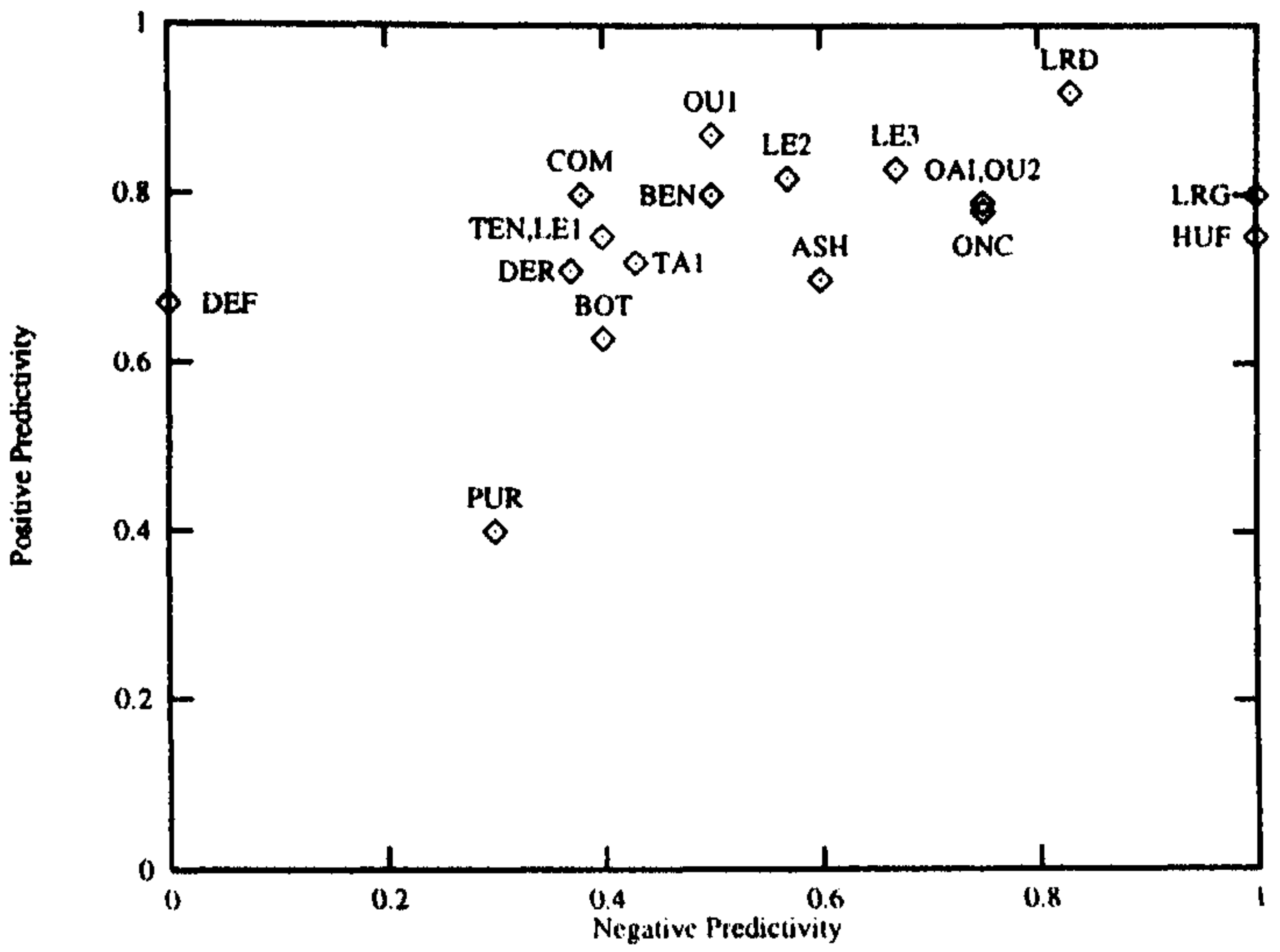

Figure 5: Scatter-diagram showing the performance of expert and ML-derived models based on their negative *(x-)* and positive (y-) predictivity. Models in the upper right are preferable to those in the lower left. The *x*-value for DEF has been taken as 0, as no chemicals are classified NEG by this rule.

[Benigni, 1998] R. Benigni. (Q)sar prediction of chemical carcinogenicity and the biological side of the structure activity relationship. In *Proceedings of The Eighth International Workshop on QSARs in the Environmental Sciences,* 1998. Held in Baltimore, May 16-20, 1998.

[Bootman, 1996] J. Bootman. Speculations on the carcinogenicity of 30 chemicals currently under review in rat and mouse bioassays organised by the us national toxicology program. *Mutagenesis,* 27:237-243, 1996.

[Bristol *et al,* 1996] D.W. Bristol, J.T. Wachsman, and A. Greenwell. The NIEHS Predictive-Toxicology Evaluation Project. *Environmental Health Perspectives,* pages 1001 1010, 1996. Supplement 3.

[Bristol, 1995] D.W. Bristol. Summary and Recommendations: Activity Classification and Structure-Activity Relationship Modeling for Human Health Risk Assessment of Toxic Substances. *Toxicology Letters,* pages 265-280, 1995.

[Huff *et al,* 1996] J. Huff, E. Weisburger, and V.A. Fung. Multicomponent criteria for predicting carcinogenicity: dataset of 30 ntp chemicals. *Environmental Health Perspectives,* 104:1105-1112,1996.

[Lewis *et al,* 1996] D.F.V. Lewis, C. Ioannides, and D.V. Parke. COMPACT and molecular structure in toxicity assessment: a prospective evaluation of 30 chemicals currently being tested for rodent carcinogenicity by the NCI/NTP. *Environmental Health Perspectives,* pages 1011-1016, 1996.

[Marchant, 1996] C.A. Marchant. Prediction of rodent carcinogencity using the DEREK system for 30 chemicals currently being tested by the National Toxicology Program. *Environmental Health Perspectives,* pages 1065 1074, 1996.

[Medawar, 1984] P.B. Medawar. *Pluto's Republic.* Oxford University Press, Oxford, 1984.

[Purdy, 1996] R. Purdy. A mechanism-mediated model for carcinogenicity: model content a prediction of the outcome of rodent carcinogencity bioassays currently being conducted on 25 organic chemicals. *Environmental Health Perspectives,* pages 1085 1094, 1996.

[R.Benigni *et al,* 1996] R.Benigni, C. Andreoli, and R.Zito. Prediction of the carcinogenicity of further 30 chemicals bioassayed by the US National Toxicology Program. *Environmental Health Perspectives,* pages 1041-1044, 1996.

[R.W. Tennant, 1996] J. Spalding R.W. Tennant. Predictions for the outcome of rodent carcinogenicity bioassays: identification of trans-species carcinogens and non-carcinogens. *Environmental Health Perspectives,* pages 1095 1100, 1996.

[Srinivasan *et al,* 1997] A. Srinivasan, R.D. King, S.H. Muggleton, and M.J.E. Sternberg. The Predictive Toxicology Evaluation Challenge. In *Proceedings of the Fifteenth International Conference on Artificial Intelligence (IJCAI-97).* Morgan Kaufmann, Los Angeles, CA, 1997.

[Srinivasan *et al,* 1999] A. Srinivasan, R.D. King, and D.W. Bristol. An assessment of ILP-assisted models for toxicity and the PTE-3 experiment. In *Proceedings of the Ninth International Workshop on Inductive Logic Programming,* LNAI, Berlin, 1999. Springer-Verlag. (to appear).

[Woo *et* a/., 1997] Y.T. Woo, D.Y. Lai, J.C. Arcos, M.F. Argus, M.C. Cimino, S. DeVito, and L. Keifer. Mechanism-based structure-activity relationship (SAR) analysis of 30 NTP test chemicals. *Environ. Carcino. Ecotox. Revs. C,* 15:139 160, 1997.